



TRABAJO PRACTICO

Retención en zonas competitivas de clientes, segmentando el ticket de compra y analizando la población por medio del uso de aprendizaje no supervisado de la canasta básica de una empresa de retail.

Alumno: Ricardo Báez Mulia.

Maestría: Ciencia de Datos e Inteligencia de Negocios.

Doctor: Emmanuel Olivera Pérez.



UPAEP – Secretaría General

Dirección General de Apoyos Académicos

Dirección del Centro de Recursos para el Aprendizaje y la Investigación.

Biblioteca Central - **Karol Wojtyla**

Tesis Digitales Restricciones de uso:

DERECHOS RESERVADOS ©

PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis está protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de textos, imágenes, gráficas, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente de donde la obtuvo mencionando el autor o autores involucrados en el documento.

Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

INDÍCE

Resumen	8
Introducción.....	8
Metodología CRISP-DM.....	8
Fase 1 de la metodología CRISP-DM	9
Fase 2 de la metodología CRISP-DM	9
Fase 3 de la metodología CRISP-DM	9
Fase 4 de la metodología CRISP-DM	9
Fase 5 de la metodología CRISP-DM	9
Fase 6 de la metodología CRISP-DM	10
FASE 1: Comprensión del Negocio	11
Antecedentes.....	12
Objetivos	14
Objetivos del negocio	14
Objetivos Específicos	15
Criterios de Éxito	15
Evaluación de la situación	15
Minería de datos.....	16
Objetivo de la minería de datos.....	16
Objetivos particulares	16
FASE 2: Comprensión de los Datos	22
Encuesta de seguridad alimentaria	23
Población y establecimientos económicos	30
Base interna de la compañía.....	36
Extracción y preparación del dataset.	37
Importar librerías.....	40
Carga y visualización de los datos	41
FASE 3: Preparación de los datos (proceso ETL)	48
Análisis exploratorio	49
Medidas de tendencia central.....	49
Series temporales	50
Histograma y distribución	58
Dispersión y correlación	60
FASE 4: Modelado.....	69

FASE 5: Evaluación.....	74
FASE 6: Implantación.....	82
Conclusión.....	95
Ilustraciones	
Ilustración 1. Modelos Data Driven, Cantón Croda 2020.....	10
Ilustración 2. Canasta básica alimenticia 2019 del gobierno de México, Gobierno de México 2019.....	13
Ilustración 3. Ubicación de los competidores, la tienda y su radio de 2 Km a la redonda, elaboración propia 2020.	16
Ilustración 4. Diagrama de flujo de la metodología CRISP-DM, elaboración propia 2020.	19
Ilustración 5. Diagrama de flujo del modelo de clasificación correspondiente a la fase 4, elaboración propia 2020.	19
Ilustración 6. Diagrama de flujo del algoritmo PCA, elaboración propia 2020.....	20
Ilustración 7. Diagrama de flujo del algoritmo de K-means 2020.	21
Ilustración 8. Pre visualización del resultado de la encuesta de seguridad alimentaria, ENSANUT 2018.	23
Ilustración 9. Base reducida, ENSANUT 2018.....	29
Ilustración 10. Gráfico de barra agrupadas por respuestas positivas y negativas de la encuesta, ENSANUT 2018.....	29
Ilustración 11. Gráfico de barra agrupadas por género y el tipo de respuestas de la encuesta, ENSANUT 2018.	30
Ilustración 12. Mapa: espacio y datos de México, INEGI 2015.....	30
Ilustración 13. Tool's box para extraer datos del área geográfica específica, INEGI 2015.	31
Ilustración 14. Pre visualización para la descarga exitosa de los datos del área geográfica, INEGI 2015.	31
Ilustración 15. Gráfico de barras de la edad de la población, INEGI 2015.	33
Ilustración 16. . listado superior del dataset establecimientos económicos, INEGI 2015.	33
Ilustración 17. listado inferior del dataset establecimientos económicos, INEGI 2015.	34
Ilustración 18. Treemap del top 10 los establecimientos económicos del área, elaboración propia 2020.	35
Ilustración 19. Diagrama de embudo del top 10 de los establecimientos económicos del área, elaboración propia 2020	35
Ilustración 20. Treemap del top 10 de los establecimientos económicos del área por establecimientos diferentes, elaboración propia 2020.	36
Ilustración 21. Mapa por puntos del top 10 de los establecimientos económicos del área, INEGI 2015.	36
Ilustración 22. Pirámide del conocimiento, Fayyad 1996.	37
Ilustración 23. Paso 1 para la transformación de datos, elaboración propia 2021.....	38
Ilustración 24. Paso 2 para la transformación de datos, elaboración propia 2021.....	38
Ilustración 25. Paso 3 para la transformación de datos, elaboración propia 2021.....	38
Ilustración 26. Paso 4 para la transformación de datos, elaboración propia 2021.....	39
Ilustración 27. Paso 5 para la transformación de datos, elaboración propia 2021.....	39
Ilustración 28. Paso 6 para la transformación de datos, elaboración propia 2021.....	40
Ilustración 29. Resultado final del proceso ETL de las unidades vendidas de cada categoría de producto, elaboración propia 2021.	40

Ilustración 30. listado de las librerías usadas en el Jupyter Notebook, elaboración propia 2021.....	41
Ilustración 31. librería usada para la carga de datos en Google Colab, elaboración propia 2021.....	41
Ilustración 32. Ejecución y carga de las librerías, elaboración propia 2021.	42
Ilustración 33. Carga de los datos "unidades vendidas" al entorno virtual, elaboración propia 2021.....	42
Ilustración 34. Éxito de la carga de datos, elaboración propia 2021.	43
Ilustración 35. Asignación del dataset en la variable "unidades" y uso de la función head () para obtener el top 5 registros, elaboración propia 2021.	43
Ilustración 36. top 5 registros con la función head (), elaboración propia.....	43
Ilustración 37. Uso de la función tail () para obtener últimos 5 registros, elaboración propia 2021.....	43
Ilustración 38. Resultado de la función tail (), elaboración propia 2021.	43
Ilustración 39. Uso de la función info() para conocer las características del dataset, elaboración propia 2021.	43
Ilustración 40. Copia para trabajar el dataframe, elaboración propia 2021.	44
Ilustración 41. Cambio de formato de la dimensión fecha, elaboración propia.....	44
Ilustración 42. Corroboración del cambio de formato de la fecha, elaboración propia 2021.	44
Ilustración 43. Función para pre visualizar el cambio de formato, elaboración propia 2021.	45
Ilustración 44. Pre visualización del cambio de formato, elaboración propia	45
Ilustración 45. Creación de la variable mes y año a partir de la variable fecha, elaboración propia 2021.	45
Ilustración 46. Nuevos dataframes para poder graficar individualmente las 6 categorías, elaboración propia 2021.	45
Ilustración 47. Función para desplegar del nuevo dataframe, elaboración propia 2021.	45
Ilustración 48. Visualización del nuevo dataframe, elaboración propia 2021.....	46
Ilustración 49. Transformación de la variable fecha en un index, elaboración propia 2021.	46
Ilustración 50. Eliminación de las variables repetidas, elaboración propia 2021.....	46
Ilustración 51. Comprobación de la existencia de valores nulos, elaboración propia 2021.	46
Ilustración 52. Impresión del dataframe con la función .shape, elaboración propia 2021.	47
Ilustración 53. Medidas de tendencia central con la función describe, elaboración propia 2021.....	49
Ilustración 54. Diagramas de caja del dataframe, elaboración propia 2021.	49
Ilustración 55. Serie temporal de las unidades vendidas de la panadería, elaboración propia 2021.....	50
Ilustración 56. Serie temporal de las unidades vendidas de cafés, leches y saborizantes, elaboración propia 2021.	50
Ilustración 57. Serie temporal de las unidades vendidas de cereales, galletas y pan, elaboración propia 2021.	51
Ilustración 58. Serie temporal de las unidades vendidas de comestibles en general y abarrotes, elaboración propia 2021.....	51
Ilustración 59. Serie temporal de las unidades vendidas de la dulces, chocolates y botanas, elaboración propia 2021.	52
Ilustración 60. Serie temporal de las unidades vendidas de aceites, granos y sopas, elaboración propia 2021.	52
Ilustración 61. Serie temporal de las unidades vendidas de todas las categorías desde el 2015 hasta 2020, elaboración propia 2021.....	53

Ilustración 62. Dataframes para comparar los productos de cada categoría por los 5 años mes a mes, elaboración propia 2021.	54
Ilustración 63. Resultado del Dataframe creado para comparar mes a mes, elaboración propia 2021.	54
Ilustración 64. Comparación mes por mes de las unidades vendidas de panadería, elaboración propia 2021.	55
Ilustración 65. Comparación mes por mes de las unidades vendidas de cafés, leches y saborizantes, elaboración propia 2021.	56
Ilustración 66. Comparación mes por mes de las unidades vendidas de cereales, galletas y pan, elaboración propia 2021.....	56
Ilustración 67. Comparación mes por mes de las unidades vendidas de comestibles en general y abarrotes, elaboración propia 2021.....	57
Ilustración 68. Comparación mes por mes de las unidades vendidas de dulces, chocolates y botanas, elaboración propia 2021.	57
Ilustración 69. Comparación mes por mes de las unidades vendidas de aceites, granos y sopas, elaboración propia 2021.....	58
Ilustración 70. Código para la creación de histogramas de las 6 categorías, elaboración propia 2021.....	58
Ilustración 71. Histograma de las 6 categorías, elaboración propia 2021.....	59
Ilustración 72. Gráficas de dispersión de las unidades vendidas de panadería, elaboración propia 2021.	60
Ilustración 73. Gráficas de dispersión de las unidades vendidas de cafés, leches y saborizantes, elaboración propia 2021.	61
Ilustración 74. Gráficas de dispersión de las unidades vendidas de cereales, galletas y pan, elaboración propia 2021.	61
Ilustración 75. Gráficas de dispersión de las unidades vendidas de comestibles en general y abarrotes, elaboración propia 2021.....	62
Ilustración 76. Gráficas de dispersión de las unidades vendidas de dulces, chocolates y botanas, elaboración propia 2021.	62
Ilustración 77. Gráficas de dispersión de las unidades vendidas de aceites, granos y sopas, elaboración propia 2021.	63
Ilustración 78. Nuevo dataframe para hacer transformaciones que permitan mostrar las correlaciones entre las variables, elaboración propia 2021	63
Ilustración 79. Código y visualización de las correlaciones entre todas las variables por mes en un gráfico de dispersión, elaboración propia 2021.	64
Ilustración 80. Creación de un dataframe para su normalización, elaboración propia 2021. .	65
Ilustración 81. Resultado del dataframe creado, elaboración propia 2021.	65
Ilustración 82. Código para normalizar los datos y mostrarlo en un gráfica de distribución, elaboración propia 2021.	65
Ilustración 83. Resultado de la distribución normal de cada producto, elaboración propia 2021.....	65
Ilustración 84. Código para crear las correlaciones y su mapa de calor, elaboración propia 2021.....	66
Ilustración 85. Mapa de calor de las correlaciones de cada categoría de productos, elaboración propia 2021	66
Ilustración 86. Código para crear un mapa de calor de los p-vaes de las correlaciones de las unidades de productos más vendidos.....	67

Ilustración 87. Mapa de calor de los p-vaes de las correlaciones de las unidades de productos más vendidos	67
Ilustración 88. Código para insertar una máscara de los p-values en los mapas de calor de las correlaciones de las unidades de productos más vendidos por categoría acorde a su significancia estadística.....	67
Ilustración 89. Resultado de aplicar la máscara de p-valúes a las correlaciones de cada producto vendido con base a la significancia estadística más importante.	68
Ilustración 90. Código y resultado del top 5 de correlaciones, elaboración propia 2021	68
Ilustración 91. Comprobación de la normalización de datos, elaboración propia 2021.	70
Ilustración 92. lista con el rango de clústeres probados, elaboración propia 2021.	70
Ilustración 93. Gráfico del codo para determinar el número óptimo de clústeres, elaboración propia 2021.	71
Ilustración 94. Código para obtener e imprimir el valor óptimo de clúster, elaboración propia 2021.....	71
Ilustración 95. Código para descomponer el dataframe con PCA, elaboración propia 2021...72	72
Ilustración 96. Valor del ratio de la varianza y valor acumulativo del ratio de varianza, elaboración propia 2021.	73
Ilustración 97. Valor del singular value o eigenvalues, elaboración propia 2021.....	73
Ilustración 98. Parte 1 del código para generar el análisis de la silueta y los clústeres del rango definido, elaboración propia 2021.	77
Ilustración 99. Parte 2 del código para generar el análisis de la silueta y los clústeres definido, elaboración propia 2021.	77
Ilustración 100. Resultado promedio del análisis de la silueta, elaboración propia 2021.	78
Ilustración 101. Análisis de la silueta para el número de clústeres 2, elaboración propia 2021	78
Ilustración 102. Análisis de la silueta para el número de clústeres 3, elaboración propia 2021.	78
Ilustración 103. Análisis de la silueta para el número de clústeres 4, elaboración propia 2021.	79
Ilustración 104. Análisis de la silueta para el número de clústeres 5, elaboración propia 2021.	79
Ilustración 105. Análisis de la silueta para el número de clústeres 6, elaboración propia 2021.	79
Ilustración 106. Análisis de la silueta para el número de clústeres 7, elaboración propia 2021.	80
Ilustración 107. Código y resultado para la creación optima de 4 clústeres, elaboración propia 2021.....	80
Ilustración 108. Transformación del dataset con PCA, elaboración propia 2021	81
Ilustración 109. Código para observar en un gráfico de dispersión los clústeres creados, elaboración propia 2021	81
Ilustración 110. Gráfico de dispersión de los 4 clústeres generados con los componentes principales, elaboración propia.	81
Ilustración 111. Resultados de la implantación en forma de informes	94

Tablas

Tabla 1 Planificador del proyecto: retención de clientes en zonas competitivas	17
Tabla 2 Descripción del Planificador de las actividades del planificador del proyecto.	18

Tabla 3. Shape de los datasets extraídos del Mapa de INEGI, INEGI 201531
 Tabla 4. Dataset de la población y vivienda, INEGI 2015.....32
 Tabla 5. Dataframe modificado con las edades de la población, INEGI 2015.....32

Ecuaciones

Ecuación 1. Varianza total presente en el set de datos, Amat 2021.....72
 Ecuación 2. Varianza explicada por la componente m, Amat 202172
 Ecuación 3. Proporción de varianza explicada por la componente m dada por el ratio, Amat 2021.....73
 Ecuación 4. Asignación del dato dentro del clúster, Kumar 2021.75
 Ecuación 5. Delimitación de la silueta respecto a su vecino cercano, Kumar 2021.75
 Ecuación 6. Parámetro que delimita al vecino, Kumar 2021.....75
 Ecuación 7. Análisis de la silueta con los parámetros que se incluyen dentro del *cluster*, Kumar 2021.....76

Tema de tesis

Retención de clientes en zonas competitivas, segmentando el ticket de compra y analizando la población por medio del uso de aprendizaje no supervisado (*PCA*) de la canasta básica de una empresa de *retail*.

Resumen

La caracterización de clientes para definir un público objetivo dentro una tienda ha sido una tarea sumamente importante para cualquier empresa del sector alimenticio que busca ser competitiva (Brenes, 2015), por tal motivo hoy en día se ha buscado entender el contexto pasado, presente y futuro de los clientes a través del análisis de datos provenientes de distintas fuentes como lo explica la metodología CRISP-DM (Chapman, 2021). Por consiguiente, en este trabajo se ha optado por usar esta metodología para extraer e interpretar la información más relevante; de igual forma se realizó un análisis descriptivo y algunos modelos de *machine learning* para identificar patrones de comportamiento de los posibles consumidores que visitan a la tienda en un rango de 2 km durante los últimos 5 años.

Palabras clave: K-means, Datos, clústeres, correlaciones, estadística, ETL.

Introducción

La minería de datos es un proceso usado en el descubrimiento de patrones desconocidos que aportan ventajas competitivas en la toma de decisiones. Dicho proceso está conformado de seleccionar, explorar, modificar, modelar y valorar grandes cantidades de datos ("La minería de datos: Definición, usos y funciones", 2021).

Los beneficios que acarrea la minería de datos son diversos, por lo que existen distintas estrategias con la capacidad de cumplir los puntos anteriormente mencionados. En la actualidad la metodología más se empleada por las empresas es: "*Cross-Industry Standard for Data Mining* (CRISP-DM) por su versatilidad y bajo coste (Chapman, 2021).

La metodología CRISP-DM se diseñó en 1999 mediante un consorcio de empresas europeas, estas buscaban no depender de especialistas costosos. Por lo que plantearon los siguientes objetivos que los distinguen dentro de su metodología:

- ▶ Fomentar la interoperabilidad de las herramientas durante todo el proceso de minería de datos, ya que anteriormente obligaban a las empresas a usar un software o herramienta específica.
- ▶ Eliminar la experiencia misteriosa y costosa de las tareas simples de la minería de datos, facilitando su implementación en las empresas.

Esta metodología al no tener propietario y ser adaptable para cualquier tipo de industria, se convirtió en la solución idónea para las empresas que quieren orientar su modelo de negocios a uno impulsado por datos (Data Driven). Además de que integra los problemas de negocio y el análisis técnico en la búsqueda de conocimiento.

Metodología CRISP-DM

La metodología CRISP-DM consta de 6 fases:

1. **Comprensión del negocio:** En esta primera fase, se alinean los objetivos del proyecto de la minería de datos con los objetivos del negocio. Al conseguirlo, evitamos embarcarnos en un proyecto de minería de datos que no produzca ningún efecto real que beneficie o aporte a la organización.
2. **Compresión de los datos:** en este paso se busca conocer la estructura, distribución y calidad de los datos.
3. **Preparación de los datos:** durante este paso se busca emplear los datos finales sobre los cuales se aplicarán los modelos.
4. **Modelado:** se busca construir un modelo que permita alcanzar los objetivos del proyecto. Esto a través de seleccionar la técnica más adecuada de modelado acorde a los datos disponibles y el cumplimiento de los objetivos establecidos en las fases previas; mediante esto se garantiza una estrategia de la calidad en la construcción del modelo al aplicar la técnica previamente seleccionada. Cabe destacar que se busca también ajustar el modelo de forma iterativa evaluando su fiabilidad e impacto en los objetivos establecidos.
5. **Evaluación:** se evalúa el grado de acercamiento del modelo a los objetivos del negocio.
6. **Despliegue:** se ofrece un despliegue de los resultados obtenidos de forma que sea propagado a los usuarios finales, así como el mantenimiento del mismo una vez el despliegue haya finalizado.

Fase 1 de la metodología CRISP-DM

Esta fase como lo muestra la ilustración 4 se conforma por 4 pasos:

- | | |
|---|---|
| ▶ Determinar los objetivos del negocio. | ▶ Determinar los objetivos de la minería. |
| ▶ Evaluación de la situación | ▶ Realizar el plan de proyecto |

Fase 2 de la metodología CRISP-DM

La fase llamada “Comprensión de los datos” consta de 4 pasos:

- | | |
|-----------------------------------|--------------------------------------|
| ▶ Recolectar los datos iniciales. | ▶ Exploración de los datos. |
| ▶ Descripción de los datos. | ▶ Verificar la calidad de los datos. |

Fase 3 de la metodología CRISP-DM

En la fase “Preparación de los datos (proceso ETL)” se integra por los siguientes 4 pasos:

- | | |
|--------------------------|------------------------|
| ▶ Seleccionar los datos. | ▶ Construir los datos. |
| ▶ Limpiar los datos. | ▶ Integrar los datos. |

Fase 4 de la metodología CRISP-DM

Esta fase como su nombre lo expresa se encarga del modelo y necesita cumplir los siguientes 4 pasos:

- | | |
|-----------------------------------|------------------------|
| ▶ Escoger la técnica de modelado. | ▶ Construir el modelo. |
| ▶ Generar el plan de prueba. | ▶ Evaluar el modelo. |

Fase 5 de la metodología CRISP-DM

La fase de la evaluación como lo muestra la ilustración 4 se conforma por 4 pasos:

- ▶ Evaluar los resultados.
- ▶ Revisar el proceso.
- ▶ Determinar los próximos pasos.

Fase 6 de la metodología CRISP-DM

Esta fase trata de la implantación y se conforma por 4 pasos:

- ▶ Planear la implantación.
- ▶ Planear la monitorización y mantenimiento.
- ▶ Producir el informe final.
- ▶ Revisar el proyecto

La facilidad con la que se adapta esta metodología a los distintos modelos de negocio orientado a datos como se observa en la ilustración siguiente, ha permitido tomar decisiones competitivas. Puesto que no todas las empresas presentan la misma cantidad de datos o desean realizar el mismo análisis.

Modelos de negocio impulsados por los datos

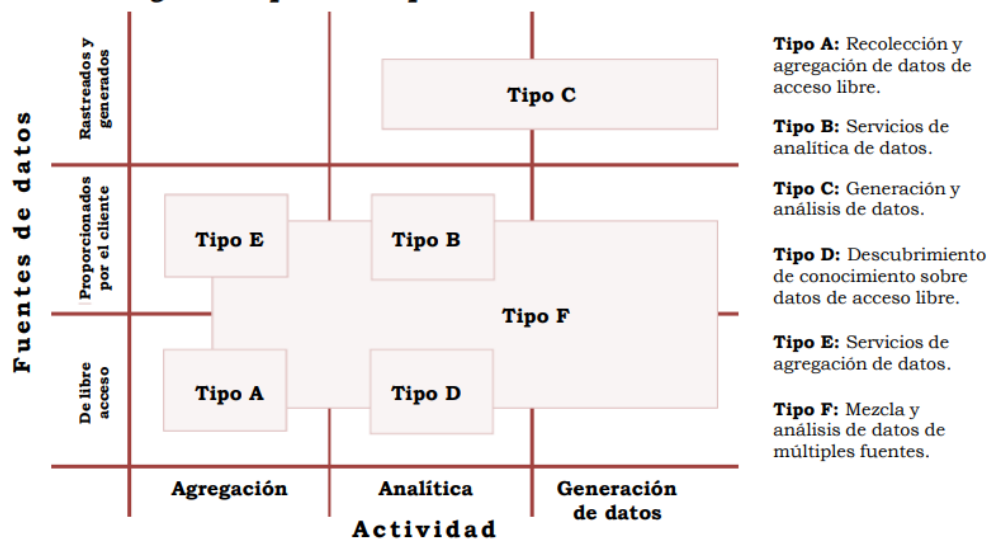


Ilustración 1. Modelos Data Driven, Cantón Croda 2020.

Lo anteriormente mencionado, nos permite reconocer el modelo con el que se abordará el trabajo para la tienda minorista. Dicho modelo es de tipo F puesto que se recopilarn fuentes de datos internas y de libre acceso, además, estas serán analizadas para incorporar los descubrimientos generados en los análisis cotidianos de la empresa.

Al conocer el modelo de negocio que utiliza la empresa, podemos con base a sus necesidades adecuar e implementar la metodología CRISP-DM. Esta recorrerá un total de 6 fases, la primera constará de entender el negocio, su situación actual, la de los datos y planificar del resto de las fases.

La segunda y tercera fase tratara de recopilar y adecuar los datos para su futuro uso en el análisis y creación del modelo. Una vez conseguido los datos óptimos se avanzará a la fase 4, la cual se encargará de crear un modelo de ciencia de datos que arroje posibles respuestas a las preguntas del negocio. Las soluciones de estas preguntas serán evaluadas en la fase 5 para así una vez conseguido los resultados óptimos pasar a la fase 6 de implantación.



FASE 1:

Comprensión del Negocio

| Antecedentes

Las grandes cadenas comerciales encargadas de suministrar productos al consumidor final son también conocidas como tiendas de retail. El acceso de la población hacia estas cadenas han permitido transformar en la actualidad sus hábitos de consumo, provocando la evolución de este sector tradicional a uno global, diverso y complejo (Alkint, 2018).

En México las empresas del sector minorista han existido desde 1800, ya que se ha buscado solventar las necesidades de la población en las zonas urbanas. Este mercado ha crecido tanto en el aspecto económico que se espera en el 2019, ganancias de 26.29 billones de dólares alrededor del mundo mientras que para el 2020 alrededor de 27.73 billones de dólares (Ramirez,2021).

La incorporación tiendas de *retail* focalizadas en dietas alimenticias de la población trae consigo un beneficio económico; ya que garantizan 100,000 empleos directos y muchos más de forma indirecta (Brenes, 2015). Por otra parte, Brenes también explica los beneficios sociales de proveer a los habitantes alimentos altos en estándares de calidad aptos para integrarse a los hábitos alimenticios de regiones puntuales sin interferir en sus dietas base.

El país cuenta actualmente con cadenas nacionales e internacionales en todo el territorio que buscan suministrar y solventar las necesidades básicas, para realizar esto en el ámbito alimenticio el gobierno mexicano en 1917 estipuló de manera constitucional una lista de servicios indispensables y necesarios para que los trabajadores y sus familias puedan satisfacer sus necesidades básicas (Ramirez, 2021).

En 1917 la constitución integró el concepto de la canasta básica, este originalmente fungió como una lista de productos alimenticios de primera necesidad accesibles para todos los mexicanos a partir de un sueldo mínimo; el objetivo de la lista es tener una serie de alimentos básicos en la dieta mexicana con un precio fijo que permitan su fácil acceso y consumo a la población (CONEVAL,2021).

En 2019 la canasta básica aumento la lista de 27 productos a 40 incluyendo productos de limpieza personal, baterías y una diversidad mayor de alimentos. Esto provocó que la dieta mexicana se volviera más diversa y nutritiva (CONEVAL,2021), además, permitió que los proveedores y las cadenas de *retail* tuvieran la necesidad de ofrecer estos productos en presentaciones y formatos más atractivos.

La lista de nuevos productos impulsada en 2019 es la siguiente:



Ilustración 2. Canasta básica alimenticia 2019 del gobierno de México, Gobierno de México 2019.

Los alimentos consumidos por la población están influenciados, pero no limitados a la canasta básica. La lista de la ilustración 2 está pensada para ser la opción mínima capaz de nutrir a la población, sin embargo, existen alimentos consumidos que no pertenecen a la lista como el pan dulce o productos derivados del nopal, ya sea por ser sustitutos veganos o por costumbres de la población ("CANASTA BÁSICA DE ALIMENTOS", 2021).

La emergencia sanitaria incentivó a los consumidores a realizar compras online en el mercado minorista y de bienes de consumo, este nuevo hábito se vio reflejado en la entrega masiva de alimentos y bebidas. Las empresas que supieron obtener beneficio de esta necesidad son aquellas las cuales contaban con una plataforma tecnológica y logística de los productos (Cálvillo, 2021).

La atención inmediata se convirtió en un requerimiento básico durante la pandemia para el éxito de los comercios minoristas. Por tal razón, se buscó lograrlo optimizando el catálogo de productos capaces de solucionar las necesidades del hogar, mejorando el tiempo y calidad de entregas; este último punto se le ha dado un peso importante ya que la conectividad de clientes entre sí ha permitido evaluar a las tiendas acorde al cumplimiento de estos objetivos (Cayuela, 2021).

Realizar todos los ajustes anteriormente mencionados en los hábitos de compra ha provocado que la variedad de opciones para el consumidor se diversificara y complicara, siendo la tarea de las empresas minoristas conseguir un grado de granularidad capaz de abarcar al mayor público posible de la manera más individual posible.

En la actualidad, el modelo de competencias económicos por el que se rigen las empresas de *retail* se orienta al oligopolio, lo que se traduce en un número limitado de

competidores fuertes que buscan la tracción completa y exclusiva del mercado. Esto se aborda en el modelo Stackelberg.

El modelo Stackelberg establece una estructura de mercado de tipo Oligopolio, las empresas son pocas y cada rival tiene la capacidad de influir en el nivel de producción de sus rivales. Dichas empresas actúan como un monopolio a fin de alcanzar éxito propio, aquellas que actúan primero se les confiere una ventaja crucial y un estatus de líder (Vela Meléndez et al., 2021).

Las expectativas son dinámicas para la empresa líder, mientras que son estáticas para las empresas seguidoras, dichas empresas establecen su función de reacción, además, la empresa líder incorpora las funciones de reacción de las seguidoras en función de maximizar sus beneficios.

En este modelo se aborda la existencia de dos empresas que conforman un duopolio; en donde ambas ofrecen productos similares. A una se le confiere la posición de empresa líder, la cual decide su producción en primer lugar. Por otra parte, la segunda empresa se le asigna la posición de seguidora la cual se encuentra a la espera de las decisiones de la empresa líder para elegir su producción.

A corto plazo el equilibrio de Stackelberg presenta un mejor beneficio en comparación a otros modelos de competencias económicas como lo es el equilibrio de Cournot; esto sucede ya que las utilidades en Stackelberg aumentan para la empresa líder al disminuir los precios de venta y aumentar la capacidad de producción. Cabe aclarar que existe una desventaja a largo plazo, debido a que cada vez habrá consumidores con ingresos más bajos; el precio irá disminuyendo hasta alcanzar el costo marginal lo cual resultará que obtendrá menos ganancias que en el equilibrio de Cournot (Vela Meléndez et al., 2021).

En relación con las distancias estratégicas que se habla en el modelo económico Stackelberg. Se ha documentado que para cubrir las necesidades alimentarias básicas con estrategias para las personas que se transportan a pie, los comercios y supermercados necesitan abarcar como mínimo un 1 km. Dicha distancia es capaz de satisfacer carencias alimentarias, aunque es necesario ajustar dicha distancia cuando existe presencia de personas que realizan sus compras a través de transporte público o su propio vehículo (Feng, 2018).

I Objetivos

El modelo de negocio orientado a data-driven es de tipo F, ya que se hace una mezcla de datos y análisis de múltiples fuentes. Por lo cual es importante definir objetivos con un enfoque al negocio y otro a la ciencia de datos, sin embargo, el objetivo principal que abarcará esta tesis es identificar los clientes con técnicas de *Machine Learning* para retenerlos en un ambiente competitivo.

Objetivos del negocio

Diseñar un modelo que permita retener clientes de Amalucan y aledaños a través de la segmentación de los distintos tipos de clientes acorde a sus tickets de compra y demografía, para así ofrecer revalorar la calidad y precios de productos de la canasta alimenticia básica de la empresa.

Objetivos Específicos

- ▶ Analizar los tickets promedio de productos alimenticios generados en el primer trimestre del 2021 para obtener las dimensiones de los hábitos de consumo de los clientes.
- ▶ Analizar las características de la población que habita en un radio de 2 km para entender la situación demográfica de la región.
- ▶ Identificar patrones en las características de los clientes que permitan asociarlos en grupos, para focalizar estrategias económicas y de marketing que permitan aumentar el incremento de ingresos.

Criterios de Éxito

A continuación, se describen los criterios de éxito:

- ▶ Seccionar en clústeres los clientes, acorde al consumo de productos.
- ▶ Asociar los clústeres de los clientes con relación a las características de la población que habita en un radio de 2 Km a la redonda de la sucursal.
- ▶ Presentar la información obtenida de los análisis en un dashboard e infografías que facilite la toma de decisiones.

I Evaluación de la situación

Esta empresa de *retail* maneja un enfoque horizontal en su diferenciación de producto, lo que significa que el negocio se enfoca en diversificar sus productos para maximizar la variedad de clientes. Sus operaciones empezaron en 2012 en una zona de plusvalía media baja, pero con proyecciones de crecimiento de la localidad de Puebla (*Zonificación Catastral y las Tablas de Valores Unitarios de Suelos Urbanos y Rústicos; así como los Valores Catastrales de Construcción por metro cuadrado, para el Municipio de Puebla, 2020*).

En este año su SMBD (sistema manejador de base de datos) SAP muestra una pérdida en promedio de 200 tickets diarios durante el mes de enero, esto junto con la disminución de clientes en el 2020 se tradujo en una pérdida de \$2 000 000.

Esta situación en la que se encuentra la empresa se debe a un conjunto de factores, uno de ellos es la pandemia que cambió los hábitos de consumo de toda la población a nivel nacional junto con la pérdida de empleos. Además, la reciente incorporación de 2 sucursales de una franquicia competidora se acentuó a las cercanías de la tienda lo que provoca una dispersión de clientes en la zona como se puede observar en la ilustración 3.

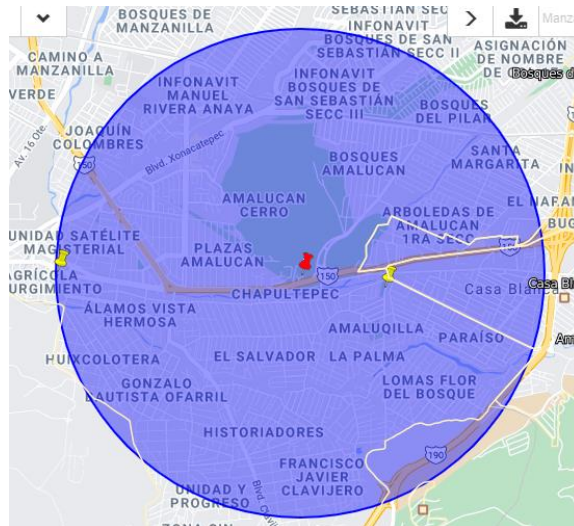


Ilustración 3. Ubicación de los competidores, la tienda y su radio de 2 Km a la redonda, elaboración propia 2020.

I Minería de datos

Objetivo de la minería de datos

El producto de ciencia de datos va ser (in->out), ya que las fuentes de datos usadas serán internas con repercusiones externas, aunque una de estas fuentes proviene de una fuente externa.

Utilizar distintas técnicas y tecnologías en ciencia de datos para encontrar patrones de comportamiento en los clientes que permitan agruparlos en clústeres acorde a su consumo de productos y las características de individuales de la poblacional asociada a dichos productos.

Objetivos particulares

- ▶ Utilizar las bases de datos de la empresa para visualizar su estado y entender el punto de partida del análisis.
- ▶ Implementar una metodología híbrida de ciencia de datos y de competencias económicas como CRIDP-DM y el modelo espacial stackelberg.
- ▶ Realizar un algoritmo de PCA para re-escalar las múltiples dimensiones de la base, con el objetivo de poder agrupar las características de los clientes.

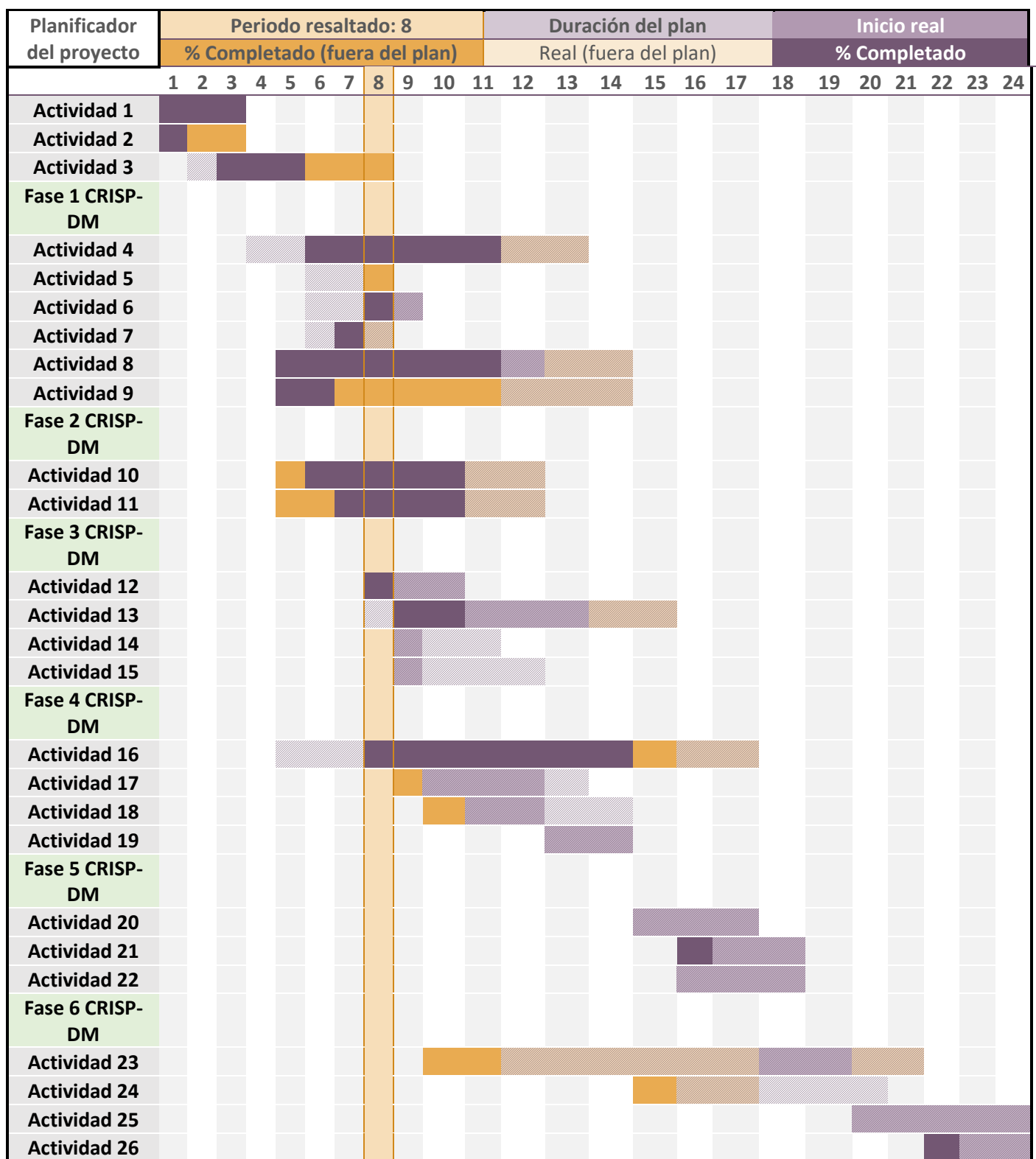


Tabla 1 Planificador del proyecto: retención de clientes en zonas competitivas, elaboración propia

No. actividad	Descripción
Actividad 1	Búsqueda de la empresa
Actividad 2	Contacto con la empresa
Actividad 3	Presentación de propuesta de tesis
Fase 1 CRISP-DM	Comprensión del negocio
Actividad 4	Investigar el ambiente y estado actual de la empresa
Actividad 5	Aceptación del proyecto
Actividad 6	Acta de Constitución del Proyecto
Actividad 7	Confirmación del proyecto
Actividad 8	Recolectar los datos iniciales
Actividad 9	Descripción de los datos
Fase 2 CRISP-DM	Comprensión de los datos
Actividad 10	Exploración de los datos
Actividad 11	Verificar la calidad de los datos
Fase 3 CRISP-DM	Preparación de los datos (proceso ETL)
Actividad 12	Seleccionar los datos
Actividad 13	Limpiar los datos
Actividad 14	Construir los datos
Actividad 15	Integrar los datos
Fase 4 CRISP-DM	Modelado
Actividad 16	Escoger la técnica de modelado
Actividad 17	Generar el plan de prueba
Actividad 18	Construir el modelo
Actividad 19	Evaluar el modelo
Fase 5 CRISP-DM	Evaluación
Actividad 20	Evaluar los resultados
Actividad 21	Revisar el proceso
Actividad 22	Presentación de reportes/conclusiones al cliente
Fase 6 CRISP-DM	Implantación
Actividad 23	Planear la implantación
Actividad 24	Planear la monitorización y mantenimiento
Actividad 25	Producir el informe final
Actividad 26	Revisar el proyecto

Tabla 2 Descripción del Planificador de las actividades del planificador del proyecto, elaboración propia.

Los objetivos y la evaluación de la situación fueron ya previamente descritos, mientras que la planeación del plan de proyecto se realizó en un diagrama de Gantt en Excel como lo muestra la tabla 1 y 2. La planeación consta de 22 semanas para las 26 actividades que constan el proyecto.

Las actividades descritas en la tabla 2 abordan la búsqueda y presentación de la empresa, además, de las 6 fases de CRISP-DM. En el diagrama podemos observar el tiempo planeado para cada actividad y el porcentaje completado de color morado, mientras que de color naranja se observa si ha existido alguna actividad que ha llevado menos o más tiempo del planeado.

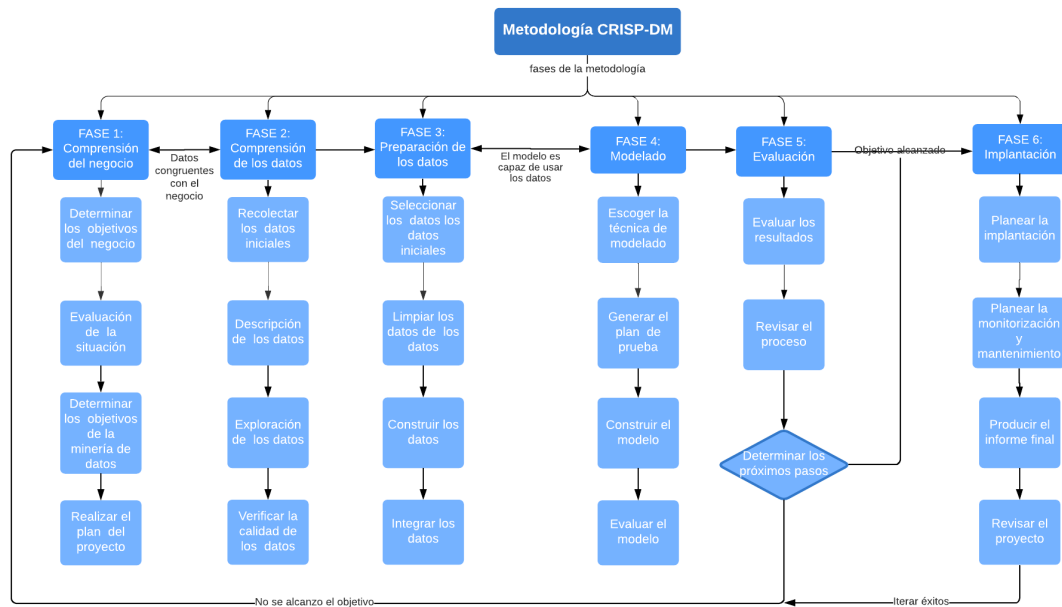


Ilustración 4. Diagrama de flujo de la metodología CRISP-DM, elaboración propia 2020.

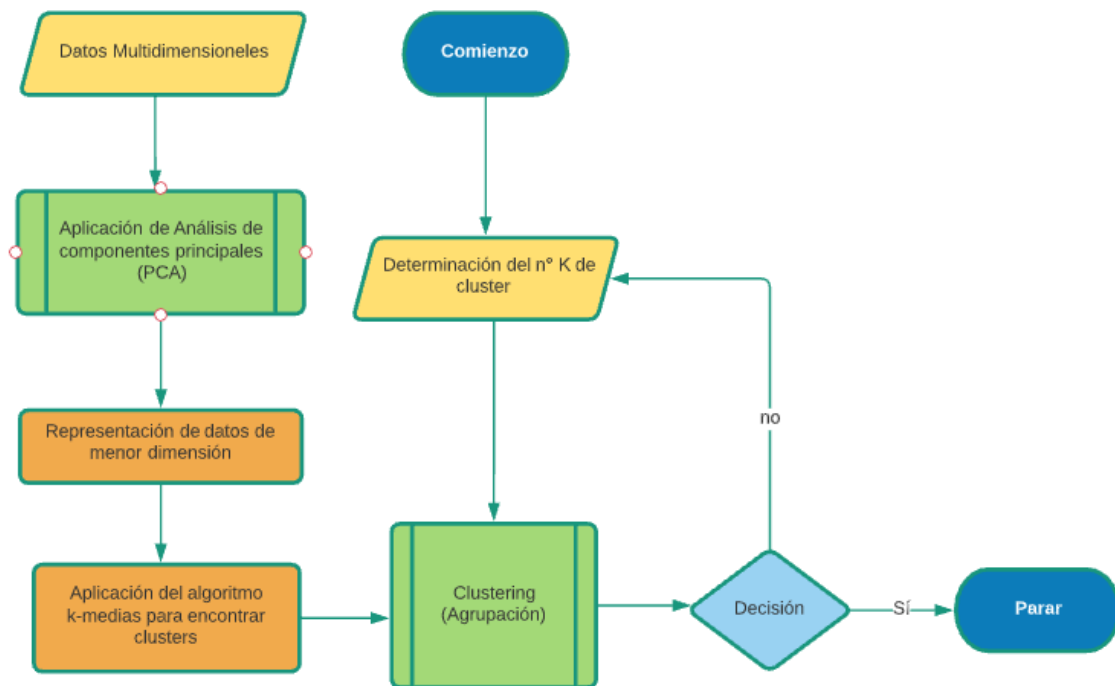


Ilustración 5. Diagrama de flujo del modelo de clasificación correspondiente a la fase 4, elaboración propia 2020.

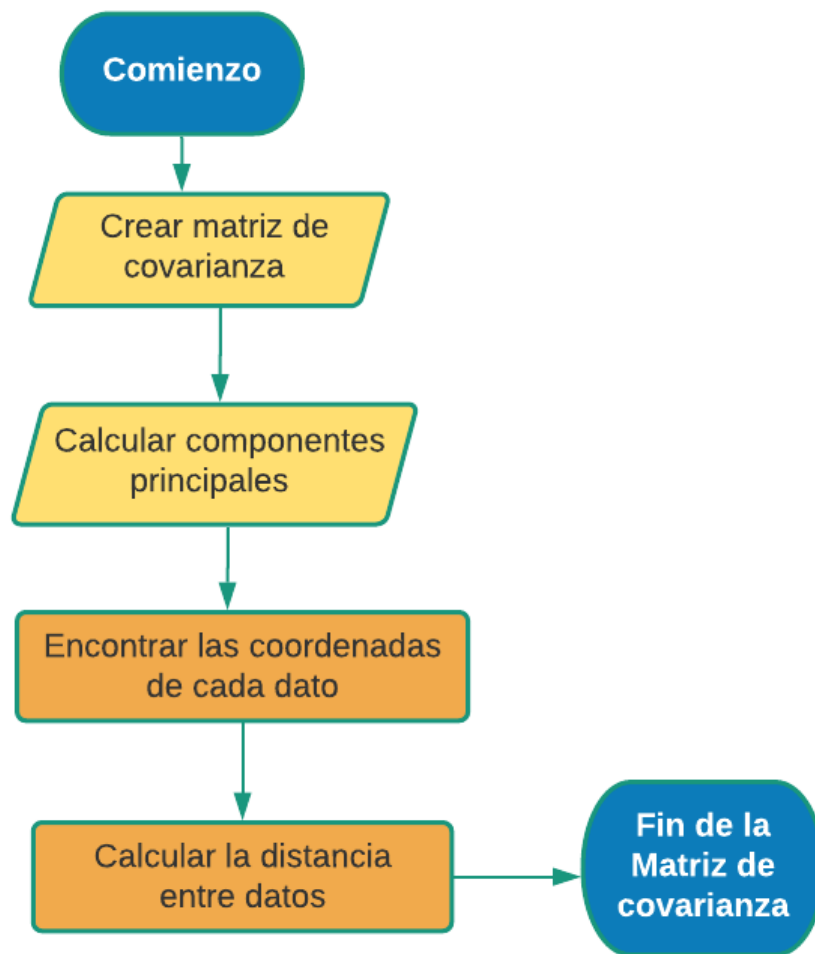


Ilustración 6. Diagrama de flujo del algoritmo PCA, elaboración propia 2020.

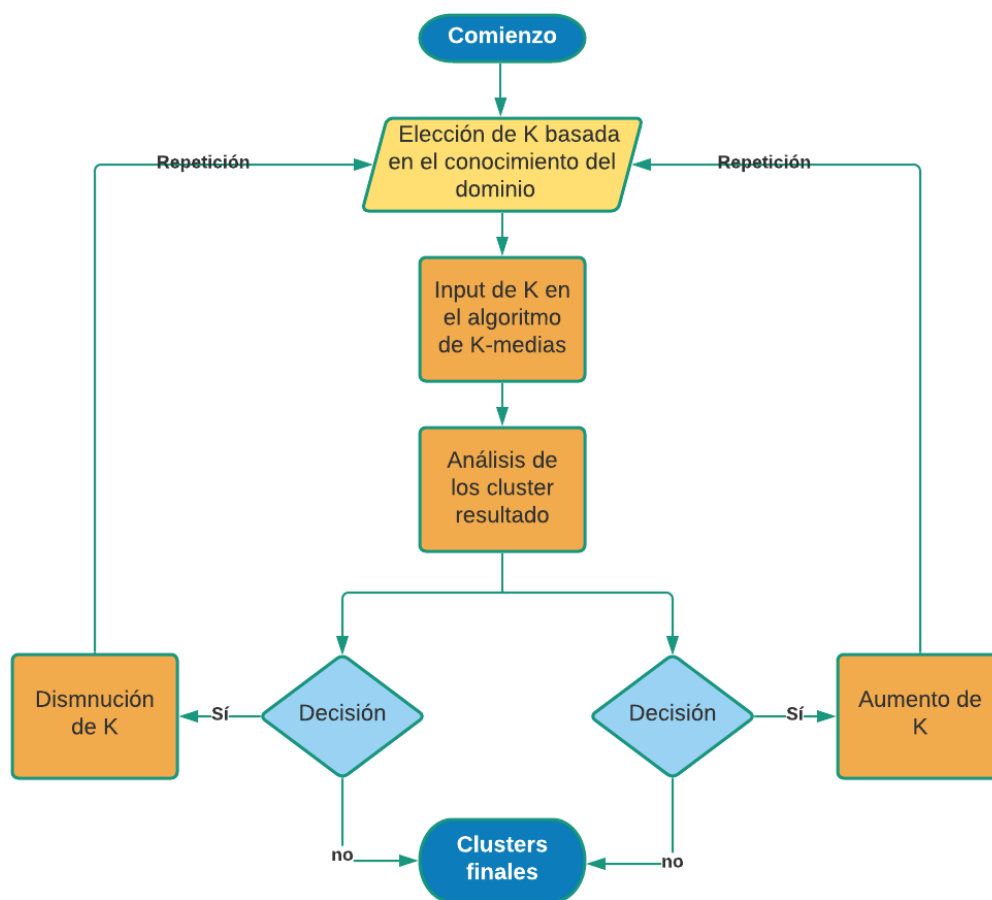
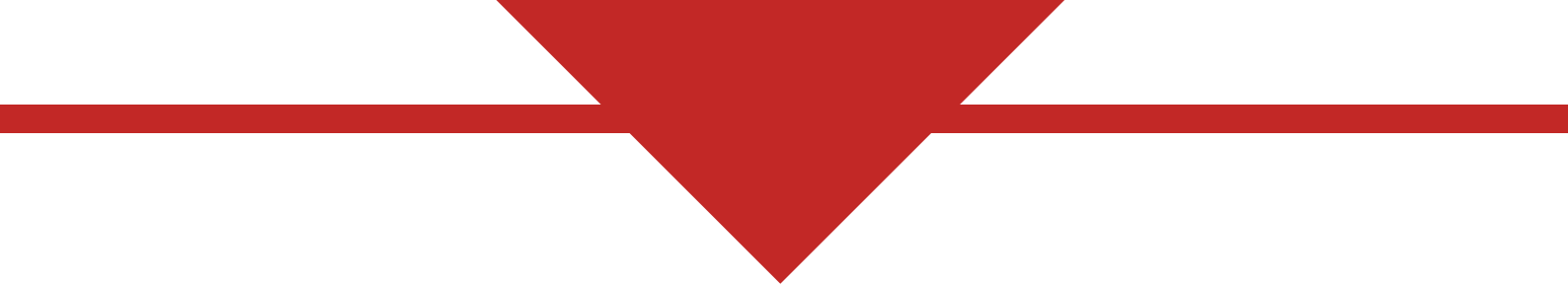


Ilustración 7. Diagrama de flujo del algoritmo de K-means 2020.

A su vez, como se observa en el cronograma de actividades se realizará la implementación del modelo “PCA” que muestra la ilustración 5 y 6 para poder ejecutar de manera eficiente el algoritmo de clasificación. Este nos permitirá respetar la dimensión de nuestras bases de datos, pero las adecuará en una matriz de distancias que podrá ser aprovechada por el modelo de *K means*.

Al haber conseguido preparar los datos con *PCA*, se pasará propiamente al algoritmo de clasificación *K means* que se observa en la ilustración 7. En este algoritmo se buscará un valor *k* que definirá el número particiones que segmentará los datos, se desea obtener un algoritmo eficiente que permita clasificar las compras realizadas mes a mes y así poder detectar patrones, de esta forma se podrá tener un número de tipos cliente que correspondan el hábito de consumo que se compró durante el año 2020. Al contar con esta información junto con la recopilada acerca de los datos demográficos de la región se podrá otorgar información valiosa para la empresa.



FASE 2:

Comprensión

de los Datos

La recolección y validación de los datos parte de 3 fuentes principalmente:

- Encuesta de seguridad alimentaria 2018, ENSANUT Centro de Investigación en Evaluación y Encuestas, Instituto Nacional de Salud Pública, Salud Secretaría de Salud.
- Población y establecimientos económicos fuente INEGI 2015.
- Base interna de la compañía.

La fuente de ENSANUT y del INEGI provienen de fuentes externas mientras que por otro lado la base interna de la compañía es de origen propio.

Encuesta de seguridad alimentaria

Los datos extraídos del ENSANUT representan a la seguridad alimentaria de México en 2018. Estos datos fueron recopilados por medio de encuestas realizadas y validadas por el Instituto Nacional de Salud Pública y Salud secretaria de salud; con el fin de entender por un lapso de 3 meses el tipo de alimentación de cada familia dentro de la república mexicana.

Los resultados de la encuesta conforman una base de:

- 44,574 filas.
- 29 columnas.

	UPM	VIV_SEL	HOGAR	NUMREN	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
1	128	5	1	2	2	2	2	2	2	2	2	2	2	1	2	2
2	129	1	1	2	1	2	2	2	2	2	2	2	2	1	2	2
3	129	2	1	1	2	2	2	2	2	2	2	2	2	2		
4	129	3	1	1	2	2	2	2	2	2	2	2	2	2		
5	129	4	1	2	2	2	2	2	2	2	2	2	2	1	2	2
6	129	6	1	1	1	2	2	1	1	2	2	2	2	2		
7	129	7	1	2	2	2	2	2	2	2	2	2	2	2		
8	129	8	1	1	2	2	2	2	2	2	2	2	2	2		
9	129	9	1	2	1	2	2	2	2	2	2	2	2	2		
10	129	10	1	1	1	2	2	2	2	2	2	2	2	1	2	2
11	129	11	1	2	2	2	2	2	2	2	2	2	2	1	2	2
12	129	12	1	2	2	2	2	2	2	2	2	2	2	1	2	2
13	129	13	1	3	2	2	2	2	2	2	2	2	2	2		
14	129	14	1	2	2	2	2	1	1	1	1	1	2	1	1	1
15	129	15	1	2	2	2	2	2	2	2	2	2	2	2		
16	129	16	1	1	2	2	2	2	1	2	2	2	2	2		
17	129	17	1	2	2	2	2	2	2	2	2	2	2	1	2	2
18	129	18	1	1	1	2	2	1	1	1	1	1	2	1	1	1
19	129	19	1	1	1	2	1	1	1	2	2	2	2	2		
20	129	20	1	2	1	2	2	1	2	2	2	2	2	2		
21	130	1	1	2	1	2	2	2	2	2	2	2	2	1	2	2
22	130	2	1	2	2	2	2	2	2	2	2	2	2	1	2	2
23	130	3	1	1	2	2	2	2	2	2	2	2	2	1	2	2
24	130	3	2	1	2	2	2	1	2	2	2	2	2	2		

Ilustración 8. Pre visualización del resultado de la encuesta de seguridad alimentaria, ENSANUT 2018.

La estructura del archivo la podemos ver en el siguiente catalogo:

(1)	Clave que identifica de manera única al campo dentro de la base de datos.
(2)	Corresponde al nombre de la variable en el instrumento de captación.
(3)	Pregunta textual en el instrumento de captación.

(4)	Tipo de caracteres aceptados en el campo
(5)	Es el número de caracteres que ocupa el campo.
(6)	Valores que acepta el campo.
(7)	Clases que se incluyen en la respuesta de cada pregunta.

Tabla 3. Estructura del archivo de la encuesta, ENSANUT 2018.

Control de la vivienda							
ID	Nemónico	Pregunta	Tipo	Long.	Códigos válidos	Clases	
						Valor	Etiqueta
1	UPM	Unidad Primaria de Muestreo	C	5	00001...06289		
2	VIV_SEL	Vivienda seleccionada	C	2	01...25		
3	HOGAR	Hogar	C	1	1...4		
4	NUMREN	Número de Renglón	C	2	01...13		

Tabla 4. Catálogo del control de la vivienda de la estructura del archivo.

Seguridad alimentaria							
ID	Nemónico	Pregunta	Tipo	Long.	Códigos válidos	Clases	
						Valor	Etiqueta
5	P1	En los últimos tres meses, por falta de dinero u otros recursos, ¿alguna vez usted se preocupó de que los alimentos se acabaran en su hogar?	C	1	1,2,8,9	1	Sí
						2	No
						8	No responde
						9	No sabe
6	P2	En los últimos tres meses, por falta de dinero u otros recursos, ¿alguna vez en su hogar se quedaron sin alimentos?	C	1	1,2,8,9	1	Sí
						2	No
						8	No responde
						9	No sabe
7	P3	En los últimos tres meses, por falta de dinero u otros recursos, ¿alguna vez en su hogar dejaron de tener una alimentación saludable, nutritiva, balanceada, equilibrada?	C	1	1,2,8,9	1	Sí
						2	No
						8	No responde
						9	No sabe
8	P4		C	1	1,2,8,9	1	Sí

		En los últimos tres meses, por falta de dinero u otros recursos, ¿alguna vez usted o algún adulto en su hogar tuvo una alimentación basada en poca variedad de alimentos?				2	No
						8	No responde
						9	No sabe
9	P5	En los últimos tres meses, por falta de dinero u otros recursos, ¿alguna vez usted o algún adulto en su hogar dejó de desayunar, comer, almorzar o cenar?	C	1	1,2,8,9	1	Sí
						2	No
						8	No responde
						9	No sabe
10	P6	En los últimos tres meses, por falta de dinero u otros recursos, ¿alguna vez usted o algún adulto en su hogar comió menos de lo que debía comer?	C	1	1,2,8,9	1	Sí
						2	No
						8	No responde
						9	No sabe
11	P7	En los últimos tres meses, por falta de dinero u otros recursos, ¿alguna vez usted o algún adulto en su hogar sintió hambre pero no comió?	C	1	1,2,8,9	1	Sí
						2	No
						8	No responde
						9	No sabe
12	P8	En los últimos tres meses, por falta de dinero u otros recursos, ¿alguna vez usted o algún adulto en su hogar solo comió una vez al día o dejó de comer todo un día?	C	1	1,2,8,9	1	Sí
						2	No
						8	No responde
						9	No sabe
13	P9		C	1	1,2	1	Sí

		¿En su hogar viven personas menores de 18 años?				2	No
14	P10	En los últimos tres meses, por falta de dinero u otros recursos, ¿alguna vez algún menor de 18 años en su hogar dejó de tener una alimentación saludable, nutritiva, balanceada, equilibrada?	C	1	1,2,8,9,b	1	Sí
						2	No
						8	No responde
						9	No sabe
15	P11	En los últimos tres meses, por falta de dinero u otros recursos, ¿alguna vez algún menor de 18 años en su hogar tuvo una alimentación basada en poca variedad de alimentos?	C	1	1,2,8,9,b	1	Sí
						2	No
						8	No responde
						9	No sabe
16	P12	En los últimos tres meses, por falta de dinero u otros recursos, ¿alguna vez algún menor de 18 años en su hogar dejó de desayunar, comer, almorzar o cenar?	C	1	1,2,8,b	1	Sí
						2	No
						8	No responde
						9	No sabe
17	P13	En los últimos tres meses, por falta de dinero u otros recursos, ¿alguna vez algún menor de 18 años en su hogar comió menos de lo que debía?	C	1	1,2,8,9,b	1	Sí
						2	No
						8	No responde
						9	No sabe
18	P14	En los últimos tres meses, por falta de dinero u otros recursos, ¿alguna vez tuvieron que	C	1	1,2,8,b	1	Sí
						2	No
						8	No responde

		disminuir la cantidad servida en las comidas a algún menor de 18 años en su hogar?					
19	P15	En los últimos tres meses, por falta de dinero u otros recursos, ¿alguna vez algún menor de 18 años en su hogar sintió hambre pero no comió?	C	1	1,2,8,9,b	1	Sí
						2	No
						8	No responde
						9	No sabe
20	P16	En los últimos tres meses, por falta de dinero u otros recursos, ¿alguna vez algún menor de 18 años en su hogar solo comió una vez al día o dejó de comer todo un día?	C	1	1,2,8,b	1	Sí
						2	No
						8	No responde

Tabla 5. Catálogo de las preguntas de seguridad alimentaria.

Variables de Sistema							
ID	Nemónico	Pregunta	Tipo	Long.	Códigos válidos	Clases	
						Valor	Etiqueta
21	EDAD	Edad	C	3	011...111		
22	SEXO	(NOMBRE) es hombre.	C	1	1,2	1	Hombre
		(NOMBRE) es mujer.				2	Mujer
23	ENT	Clave de Entidad Federativa	C	2	01...32		
24	DOMINIO	Dominio	C	1	1,2	1	Urbano
						2	Rural
25	REGION	Región	C	1	1...4	1	Norte
						2	Centro
						3	Ciudad de México
						4	Sur
26	EST_DIS	Estrato de diseño	C	3	001...328		
27	UPM_DIS	UPM de diseño	C	5	00001...06289		
28	ESTRATO	Estrato sociodemográfico	C	1	1...4		
29	F_SEGAL	Factor de seguridad alimentaria	C	4	72...6048		

Tabla 6. Variables del sistema de la encuesta, ENSANUT 2018.

Al conocer la estructura y la forma de la base, se decidió reconvertir los valores numéricos de los registros a su valor original para lograr una exploración mejor de los

datos. Una vez conseguido el cambio de formato se decidió solo tomar en cuenta aquellos registros que pertenecieran al estado de Puebla.

El resultado de la reducción como se observa en la ilustración 9 arrojó un total de 1437 registros y pese haberse limpiado los datos que no correspondían al estado de Puebla existen 559 vacíos en la sección de las preguntas 10 hasta la 16, pero se decidió conservarlas puesto que aun contiene información útil en el resto de las secciones. Tras haber hecho el filtrado de los datos se decidió hacer un conteo rápido para comparar las respuestas positivas negativas, este conteo se puede observar en el gráfico de la ilustración 10, donde los puntos más destacados son:

- Las preguntas con más respuestas positivas son:
 - ¿En su hogar viven personas menores de 18 años?
 - ¿Alguna vez usted se preocupó de que los alimentos se acabaran en su hogar?
 - ¿Alguna vez usted o algún adulto en su hogar tuvo una alimentación basada en poca variedad de alimentos?
 - ¿Alguna vez en su hogar dejaron de tener una alimentación saludable, nutritiva, balanceada, equilibrada?
- Las preguntas con más respuestas negativas son:
 - ¿Alguna vez algún menor de 18 años en su hogar en su hogar solo comió una vez al día o dejo de comer todo un día?
 - ¿Alguna vez algún menor de 18 años en su hogar en su hogar sintió hambre, pero no comió?
 - ¿Alguna vez algún menor de 18 años en su hogar en su hogar dejo de desayunar, comer, almorzar o cenar?

Estas respuestas nos indican que en Puebla el 60% está compuesta por menores de 18 años y su alimentación no es escaza, pero esto porque probablemente los jefes de familia priorizan su alimentación sobre la de los adultos mayores puesto que existe una preocupación real por surtir comida a todos los miembros de la familia ocasionando que en ocasiones se omita un plato o una variedad de alimentos.

Ensenut 1º analisis.xlsx - Excel (Error de activación de productos)																											
Inicio Insertar Diseño de página Fórmulas Datos Revisar Vista Desarrollador ¿Qué desea hacer?																											
Pegar Fuente Alineación Número Formato condicional Dar formato como tabla Estilos de celdas Insertar Eliminar Formato Ordenar y filtrar Buscar y seleccionar Modificar																											
C1																											
	HOGAR	NUMER	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22	P23	P24	P25
1	1	2	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
2	1	2	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
3	1	2	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
4	1	2	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
5	1	3	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI
6	1	2	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
7	1	3	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
8	1	2	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
9	1	2	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
10	1	1	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI
11	1	4	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
12	1	1	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
13	1	2	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI
14	1	1	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
15	1	1	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
16	1	1	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
17	1	1	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
18	1	3	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
19	1	1	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
20	1	1	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI
21	1	3	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
22	1	1	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
23	2	1	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
24	1	2	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
25	1	2	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
26	1	1	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
27	1	2	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI
28	1	1	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI
29	1	2	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI
30	1	3	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
31	1	2	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI
32	1	1	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI
33	1	2	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
34	1	1	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
35	1	1	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI
36	1	1	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI
37	1	1	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
38	2	1	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
39	1	1	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
40	1	1	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
41	1	1	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
42	1	4	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI	No	SI

Ilustración 9. Base reducida, ENSANUT 2018.

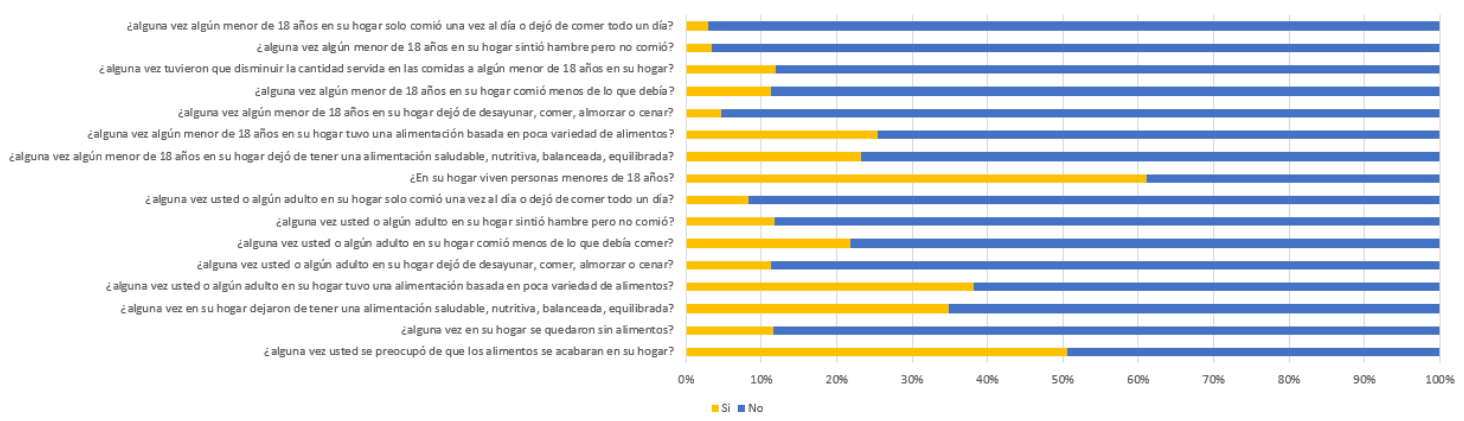


Ilustración 10. Gráfico de barra agrupadas por respuestas positivas y negativas de la encuesta, ENSANUT 2018.

Por otra parte, podemos cuantificar la respuesta por género para así poder medir cual es el sector de la sociedad que requiere de una mejor seguridad alimentaria. En el resultado destacan las preguntas 1,3,4,9 y 11 que son las siguientes:

- ¿Alguna vez usted se preocupó de que los alimentos se acabaran en su hogar?
- ¿Alguna vez en su hogar dejaron de tener una alimentación saludable, nutritiva, balanceada, equilibrada?
- ¿Alguna vez usted o algún adulto en su hogar tuvo una alimentación basada en poca variedad de alimentos?
- ¿En su hogar viven personas menores de 18 años?
- ¿Alguna vez algún menor de 18 años en su hogar dejó de tener una alimentación saludable, nutritiva, balanceada, equilibrada?
- ¿Alguna vez algún menor de 18 años en su hogar tuvo una alimentación basada en poca variedad de alimentos?

Estas preguntas y el gráfico de la ilustración 11 muestran que en las familias predominan las mujeres, los hombres realmente no presentan una restricción de alimentos significativa a comparación de las mujeres y estas al ser probablemente las que gestionan la distribución de alimentos presentan preocupación por estar limitados o privados de ellos.

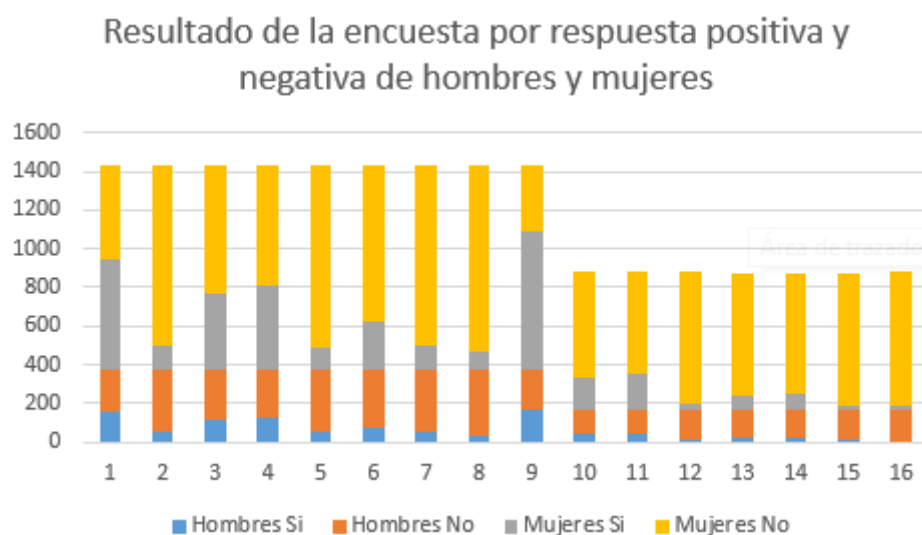


Ilustración 11. Gráfico de barra agrupadas por género y el tipo de respuestas de la encuesta, ENSANUT 2018.

Población y establecimientos económicos

Una manera interesante también para entender a la población y la competencia que rodea al establecimiento fue usando el mapa del INEGI: espacio y datos que permitió recopilar los datos en un radio aproximado de 2 km como lo muestra la ilustración 3.

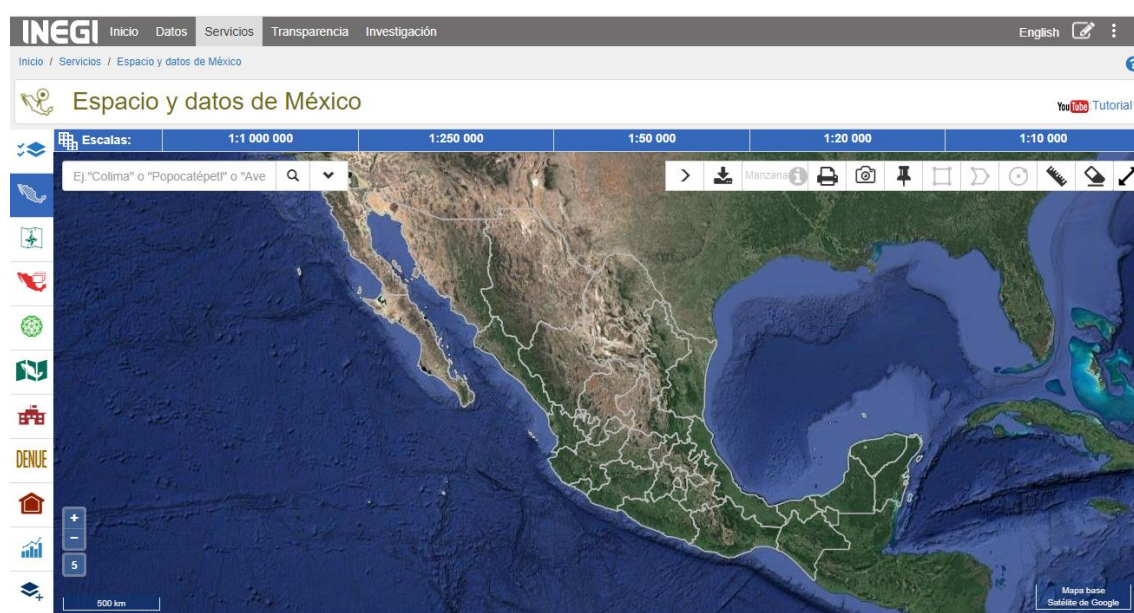


Ilustración 12. Mapa: espacio y datos de México, INEGI 2015.



Ilustración 13. Tool's box para extraer datos del área geográfica específica, INEGI 2015.

Una vez que ubicamos el área geográfica de la tienda, se marcó una circunferencia con la herramienta de la ilustración 12 y 13 para extraer los datos dentro de su radio del alcance. Al conseguirlo se desplegará una ventana para la descarga de los datos en un archivo “csv” como lo muestra la ilustración 14:



Ilustración 14. Pre visualización para la descarga exitosa de los datos del área geográfica, INEGI 2015.

Cuando se consiguió la descarga del archivo “csv” los datos de la población y establecimientos económicos se partió en 2 dataset con las siguientes características.

Nombre del dataset	Registros	Columnas
Población y vivienda	28	5
Establecimientos económicos	6650	7

Tabla 7. Estructura de los datasets extraídos del Mapa de INEGI, INEGI 2015.

En el dataset de población y vivienda solo nos interesa conocer la demografía más básica de la población por lo que se transformó en un dataframe como se ejemplifica en la tabla 7,8 y 9 para así observar en el gráfico de la ilustración 15 una distribución de la población por edades. A partir de esto se puede establecer que la suma de la población menor a 29 años supera al registrado por 30 a 59 años, significando que la población a la que atiende la empresa está conformada por familias jóvenes.

Características del entorno urbano

Nombre del indicador	En todas	En alguna	En ninguna	No especificado
Recubrimiento de la calle	330	602	143	26
Banqueta	352	634	91	24
Guarnición	329	657	91	24
Árboles o palmeras	128	791	158	24
Rampa para silla de ruedas	3	100	974	24
Alumbrado público	528	538	11	24

<i>Letrero con nombre de la calle</i>	237	663	176	25
<i>Teléfono público</i>	17	523	536	25
<i>Restricción del paso a peatones</i>	1065	20	no aplica	no aplica
<i>Restricción del paso a automóviles</i>	925	147	no aplica	no aplica
<i>Puesto semifijo</i>	0	62	1015	24
<i>Puesto ambulante</i>	0	133	944	24
<i>Vivienda</i>				
<i>Particulares</i>	39528			
<i>Habitadas</i>	34413			
<i>Particulares habitadas</i>	34354			
<i>Particulares no habitadas</i>	4698			
<i>Con recubrimiento en piso</i>	33190			
<i>Con energía eléctrica</i>	33748			
<i>Con agua entubada</i>	31598			
<i>Con drenaje</i>	33562			
<i>Con servicio sanitario</i>	33688			
<i>Con 3 o más ocupantes por cuarto</i>	1230			
<i>De 0 a 14 años</i>	37724			
<i>De 15 a 29 años</i>	36139			
<i>De 30 a 59 años</i>	50862			
<i>De 60 años y más</i>	10865			
<i>Con discapacidad</i>	3560			
<i>Fecha de actualización</i>	2015	2010		

Tabla 8. Dataset de la población y vivienda, INEGI 2015.

De 0 a 14 años	De 15 a 29 años	De 30 a 59 años	De 60 años y más	Con discapacidad	Total
37724	36139	50862	10865	3560	139150

Tabla 9. Dataframe modificado con las edades de la población, INEGI 2015.

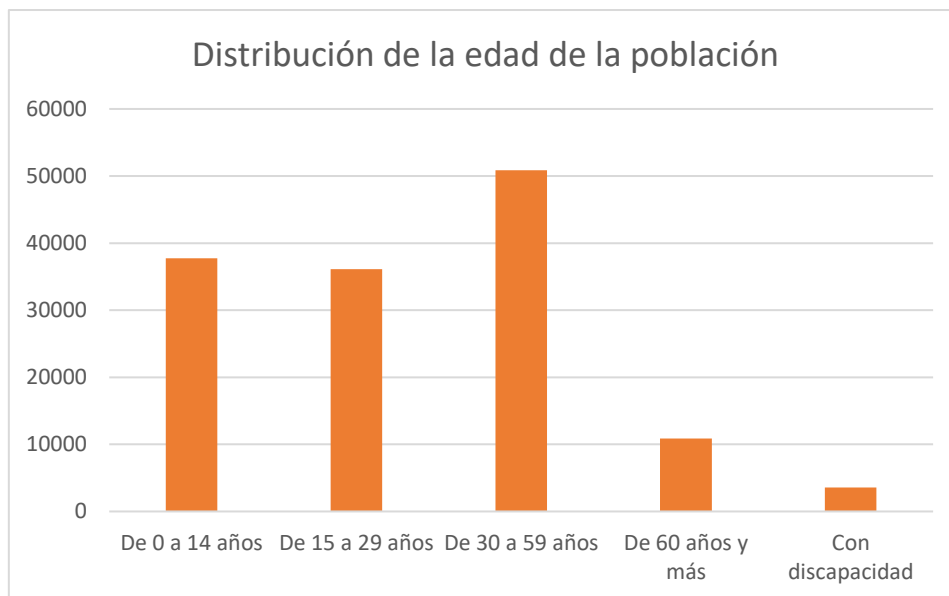


Ilustración 15. Gráfico de barras de la edad de la población, INEGI 2015.

El segundo dataset consta de 7 dimensiones y 6650 registros como se puede ver en la ilustración 16 y 17 y en la tabla 10.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Id	Nombre Esta	Id Sector	Descripción	Latitud	Longitud	Área Geográfica									
2	3437899	AGUA DE PUI	221311	Captación, tr	19.05320374	-98.14624672	Puebla, Puebla									
3	8704187	AGUA DE PUI	221311	Captación, tr	19.04502155	-98.12262751	Amozoc, Amozoc									
4	3358083	COMISION FI	221111	Generación	19.04086860	-98.14718356	Puebla, Puebla									
5	6514149	COMISION P	221311	Captación, tr	19.04063345	-98.15237997	Puebla, Puebla									
6	3356421	MACRO TAN	221311	Captación, tr	19.05268451	-98.14365845	Puebla, Puebla									
7	3357650	POZO DE AG	221312	Captación, tr	19.02538716	-98.13801296	Puebla, Puebla									
8	8628144	BALANCEO C	238190	Otros trabaj	19.03148868	-98.14881336	Puebla, Puebla									
9	6715367	BODEGA COF	236221	Edificación d	19.04452412	-98.12510485	Amozoc, Amozoc									
10	6149040	CHRONOS CA	236111	Edificación d	19.04359083	-98.14779712	Puebla, Puebla									
11	9289986	CONDADO C	236111	Edificación d	19.02920751	-98.13535315	Puebla, Puebla									
12	6149118	CONSTRULO	236111	Edificación d	19.05891374	-98.13796731	Puebla, Puebla									
13	6149216	CONSTRUCTO	237312	Construcción	19.04296042	-98.14875900	Puebla, Puebla									
14	3357840	DECORACIO	238330	Colocación d	19.04246416	-98.13637131	Puebla, Puebla									
15	6148966	EDIFICACIÓN	236221	Edificación d	19.05806185	-98.13220714	Puebla, Puebla									
16	8325321	HERCO SERV	238210	Instalacione	19.03966183	-98.14868034	Puebla, Puebla									
17	8324286	HERRERIA SI	238210	Instalacione	19.05334588	-98.12028595	Puebla, Puebla									
18	6707430	INGENIEROS	236221	Edificación d	19.04656035	-98.14554895	Puebla, Puebla									
19	9327183	INNOVACIO	237999	Otras constr	19.04306922	-98.14859153	Puebla, Puebla									
20	6809994	MG PROYECT	236111	Edificación d	19.04220718	-98.13910793	Puebla, Puebla									
21	6149004	OSVAR CON	236221	Edificación d	19.04274238	-98.14795366	Puebla, Puebla									
22	9328041	PROFESIONA	236111	Edificación d	19.04016996	-98.12644417	Puebla, Puebla									
23	6149051	R C CONSTR	236221	Edificación d	19.04457019	-98.15065470	Puebla, Puebla									
24	9278940	RENTA DE M	238910	Preparación	19.03875836	-98.12181075	Puebla, Puebla									
25	8324280	SERVICIO DE	238222	Instalacione	19.04637053	-98.12393854	Puebla, Puebla									

Ilustración 16. . listado superior del dataset establecimientos económicos, INEGI 2015.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
6626	3356389	VULCANIZAC	811191	Reparación r	19.05508041	-98.1339194C	Puebla, Puebla									
6627	8325564	VULCANIZAC	811191	Reparación r	19.03398845	-98.14677497	Puebla, Puebla									
6628	3438000	VULCANIZAC	811191	Reparación r	19.05663150	-98.13733307	Puebla, Puebla									
6629	3446284	VULCANIZAC	811191	Reparación r	19.04100701	-98.14811815	Puebla, Puebla									
6630	3344703	VULCANIZAC	811191	Reparación r	19.06076880	-98.14443435	Puebla, Puebla									
6631	3455242	VULCANIZAC	811191	Reparación r	19.03405803	-98.13076912	Puebla, Puebla									
6632	3357057	VULCANIZAC	811191	Reparación r	19.04468308	-98.14741365	Puebla, Puebla									
6633	8811829	XARENI	811499	Reparación y	19.03963311	-98.1480604C	Puebla, Puebla									
6634	8324913	XO NAILS & T	812110	Salones y cli	19.04456135	-98.14774028	Puebla, Puebla									
6635	8324223	ZARA NAILS	812110	Salones y cli	19.05309351	-98.1280449C	Puebla, Puebla									
6636	3437898	ADMINISTRA	931210	Administraci	19.05331670	-98.14600073	Puebla, Puebla									
6637	8811401	AYUNTAMIEI	931210	Administraci	19.04392079	-98.15407354	Puebla, Puebla									
6638	3356302	CASA DE SAL	931610	Actividades	19.05859213	-98.14684803	Puebla, Puebla									
6639	8811785	CASETA DE V	931410	Impartición	19.05347483	-98.14714805	Puebla, Puebla									
6640	8307884	CASTA DE PC	931410	Impartición	19.05422177	-98.15067492	Puebla, Puebla									
6641	3357153	CORRALON	931210	Administraci	19.04387084	-98.1537072C	Puebla, Puebla									
6642	3442605	DELEGACION	931210	Administraci	19.04919484	-98.1469967C	Puebla, Puebla									
6643	8811289	INSTITUTO N	931210	Administraci	19.04739343	-98.1525771C	Puebla, Puebla									
6644	3357773	LOCAL PARA	931810	Actividades	19.04294003	-98.14155062	Puebla, Puebla									
6645	3442618	MODULO DE	931410	Impartición	19.05042010	-98.1448585C	Puebla, Puebla									
6646	8692021	SUBCOMANC	931410	Impartición	19.05776061	-98.14152515	Puebla, Puebla									
6647	3357714	SUPERVISIO	931610	Actividades	19.04600550	-98.14499692	Puebla, Puebla									
6648	8544847	SUPERVISIO	931610	Actividades	19.05379770	-98.14492087	Puebla, Puebla									
6649	3357790	SUPERVISIO	931610	Actividades	19.04815934	-98.14698176	Puebla, Puebla									
6650	8811726	SUPERVISIO	931610	Actividades	19.05635778	-98.1428979C	Puebla, Puebla									

Ilustración 17. listado inferior del dataset establecimientos económicos, INEGI 2015.

Dimensiones		Tipo
ID		Int
Nombre Establecimiento		Object
IdSectorEconómico		Int
Descripción del Sector Económico		Object
Latitud		Float
Longitud		Float
Área Geográfica		Object

Tabla 10. Dimensiones del dataset de establecimientos económicos de la región, elaboración propia 2020.

Al contar con estos datos pudimos establecer una serie de visualizaciones básicas en Power BI para poder extrapolar el top 10 de sectores económicos de la zona, la cantidad de tienda que componen a cada uno y el número de franquicias o tiendas pertenecientes al mismo grupo. Con ayuda de estos gráficos se pudo llegar a las siguientes determinaciones:

- Los sectores económicos más importantes representan a la industria del alimento, artículos de papelería y a estéticas de belleza o peluquería.
- Al existir competidores especializados en productos de harina, maíz o alimentos fritos podemos estipular que hay demanda por estos productos.
- Las frutas, verduras y panificadoras, aunque en menor medida también son un ejemplo de producto demandados de interés para las tiendas de *retail*.

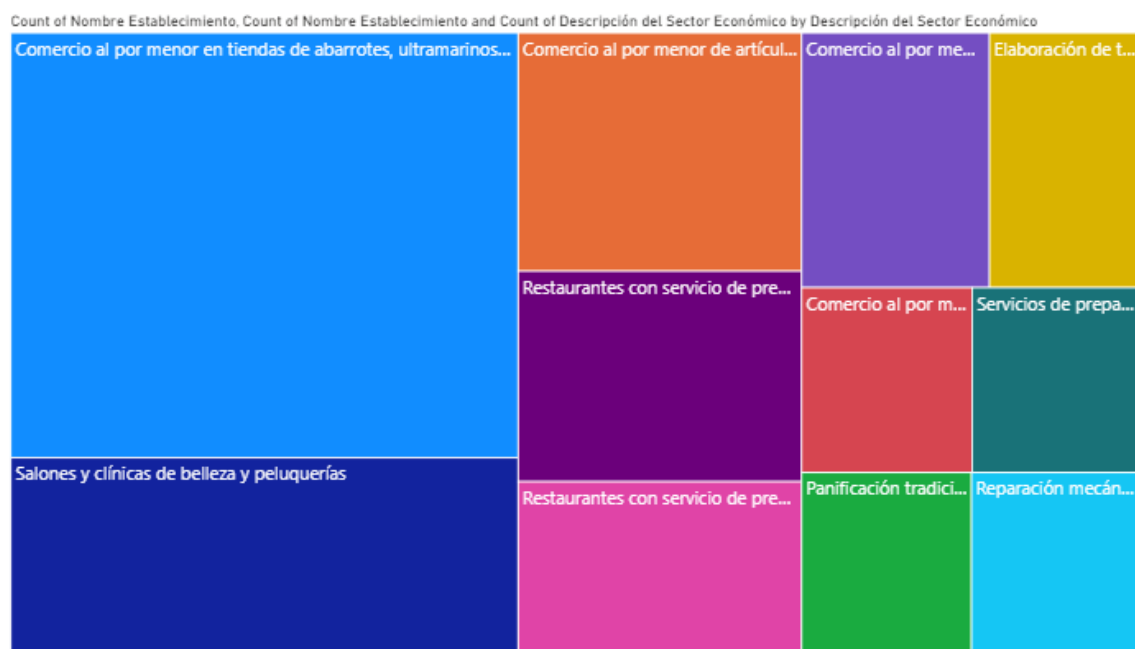


Ilustración 18. Treemap del top 10 los establecimientos económicos del área, elaboración propia 2020.



Ilustración 19. Diagrama de embudo del top 10 de los establecimientos económicos del área, elaboración propia 2020

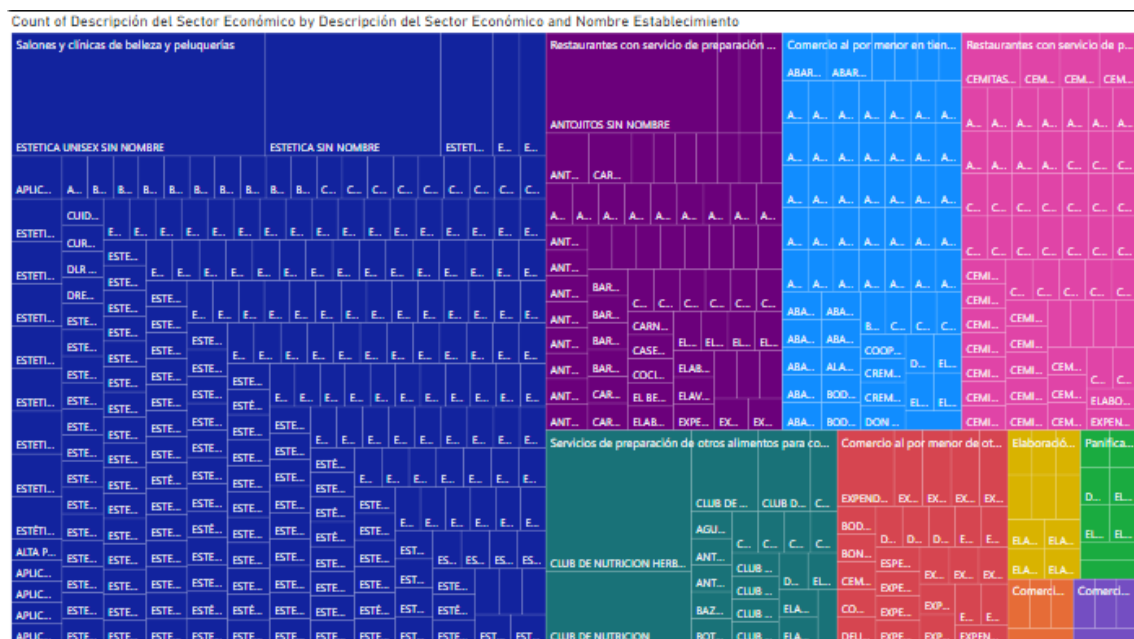


Ilustración 20. Treemap del top 10 de los establecimientos económicos del área por establecimientos diferentes, elaboración propia 2020.



Ilustración 21. Mapa por puntos del top 10 de los establecimientos económicos del área, INEGI 2015.

Base interna de la compañía

Al realizar la comprensión del negocio desde un enfoque de data science se necesita entender su contexto actual e histórico a través de los datos, lo cual se logra mediante proceso ETL de datos internos y externos. Al obtener los datos procesados y listo para su análisis podremos ubicar el tipo de dato, la dimensionalidad y el número de registros que contiene las base que previamente creamos, para así poder un análisis de calidad y con valor.

La finalidad de esto es generar un análisis exploratorio que responde a las primeras 3 bases de la pirámide del conocimiento:

- ¿Qué ocurrió?
- ¿Por qué ocurrió?
- ¿Qué está ocurriendo?

Lo cual, a su vez establece las bases necesarias para responder a los últimos escalones de la pirámide (Fayyad, Shapiro, 1996).

Pirámide del conocimiento (DIKW) (Fayyad, Shapiro, 1996)

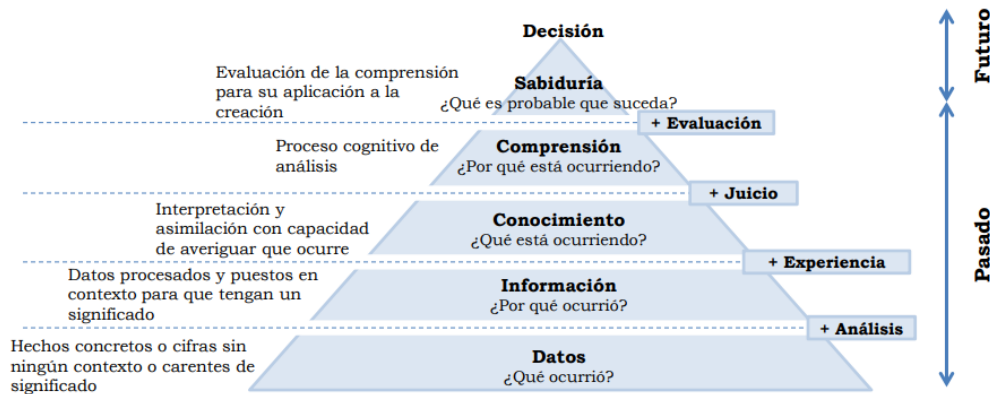


Ilustración 22. Pirámide del conocimiento, Fayyad 1996.

Extracción y preparación del dataset.

La manera en que encontraban los datos en el SMDB de la empresa implicaba que para cada producto de cada año se necesitaba genera una extracción de datos. Por lo cual, al ser 6 categorías por 5 años se generó un total de 30 sub bases que contenian la información de unidades vendidas, ventas reales y ventas esperadas a lo largo de los 12 meses del año en que se extrajo la información, sin embargo, es de nuestro interes tener el historial de las unidades vendidas de cada categoría desde el 2015 hasta el 2020 en un solo archivo.

Conseguir la unión de todos los datos con sus respectivos atributos requería de un proceso ETL en el que a través de Excel como se ve en las ilustraciones 23, 24, 25, 26 y 27 se pudo exportar los 30 archivos y así unirlos en uno solo.

La serie de pasos consistio en obtener los datos desde un archivo de texto csv, seleccionarlo e importarlo con ayuda del asistente, definir los parámetros por *default* y finalmente solo seleccionar el periodo y las unidades vendidas. Cabe destacar que se hizo una modificación de periodo para que sea más legible y una vez realizado este cambio se repitió estos pasos con cada una de las base para finalmente combinarlas como se puede ver en la ilustración 28.

Este proceso se repitió con las 6 categorías de productos:

- Panadería
- Cafés, leches y saborizantes
- Cereales, galletas y pan
- Comestibles en general y abarrotes
- Dulces, chocolates y botanas
- Aceites, granos y sopas

Una vez realizado este cambio en cada uno de los archivos, se procedió a unirlos en uno solo para tener en un solo documento las ventas de cada producto por unidad a lo largo de los últimos 5 años como se muestra en la ilustración 29.

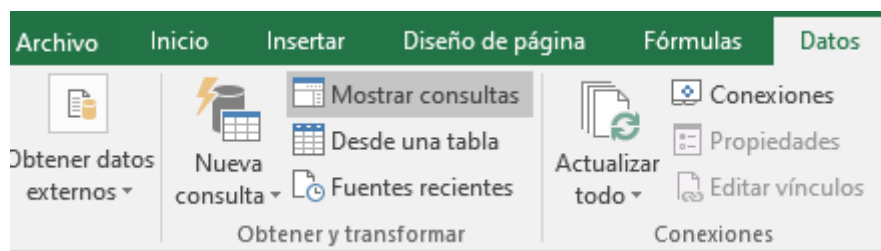


Ilustración 23. Paso 1 para la transformación de datos, elaboración propia 2021.

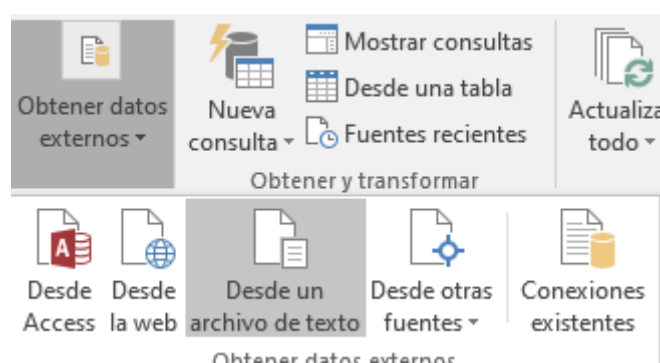


Ilustración 24. Paso 2 para la transformación de datos, elaboración propia 2021.

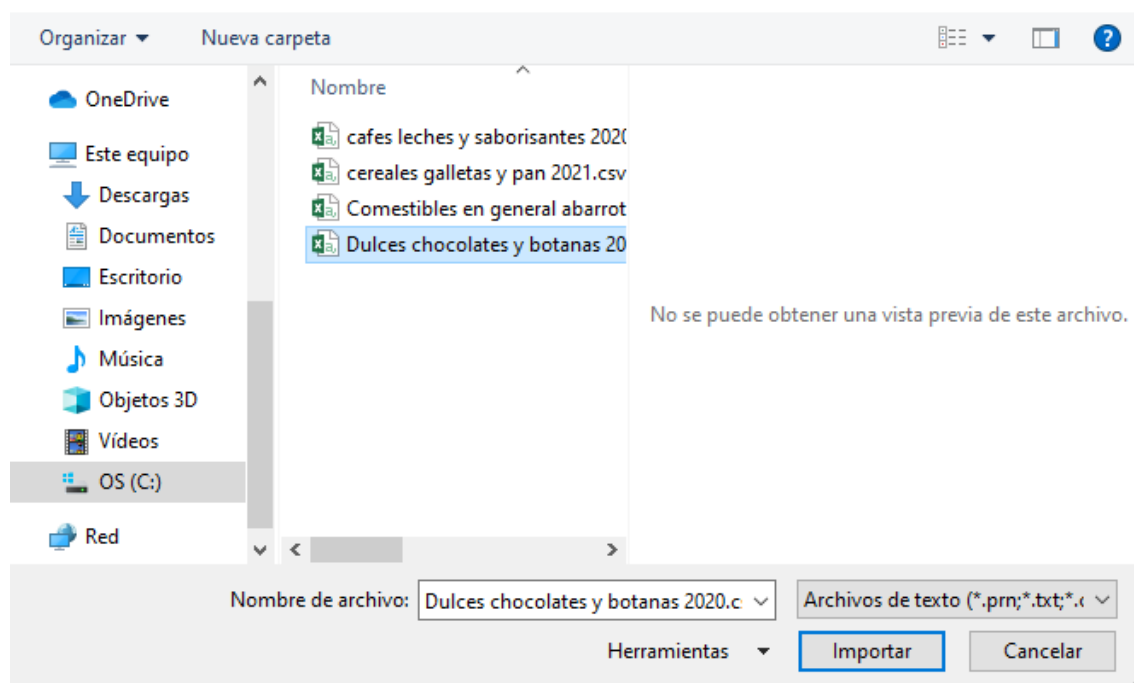


Ilustración 25. Paso 3 para la transformación de datos, elaboración propia 2021.

Asistente para importar texto - paso 1 de 3

El asistente estima que sus datos son Delimitados.

Si esto es correcto, elija Siguiente, o bien elija el tipo de datos que mejor los describa.

Tipo de los datos originales

Elija el tipo de archivo que describa los datos con mayor precisión:

☒ Delimitados - Caracteres como comas o tabulaciones separan campos.

☐ De ancho fijo - Los campos están alineados en columnas con espacios entre uno y otro.

Comenzar a importar en la fila: 1 Origen del archivo: MS-DOS (PC-8)

☐ Mis datos tienen encabezados.

Vista previa del archivo C:\Users\usuario\Documents\Maetría Upaep\Tesis\base_soriana\Dulces chocolates y botanas 2020.cs

1	Periodo	Unidades	Venta	ITI	Real	ITI	Nominal
2	96	5611	774	5601	726	973	
3	220016	647	128	675	130	362	
4	220027	501	159	877	143	849	
5	220037	797	138	156	122	24	

Cancelar < Atrás Siguiente > Finalizar

Ilustración 26. Paso 4 para la transformación de datos, elaboración propia 2021.

Asistente para importar texto - paso 2 de 3

Esta pantalla le permite establecer los separadores contenidos en los datos. Se puede ver cómo cambia el texto en la vista previa.

Separadores

☒ Tabulación

☐ Punto y coma

☐ Coma

☐ Espacio

☐ Otro:

☐ Considerar separadores consecutivos como uno solo

Calificador de texto: "

Vista previa de los datos

Periodo	Unidades	Venta						ITI	Real	ITI	Nominal
96	561			1	774	560		1	726	973	
22001	6	647			128	675					
22002	7		501	159	877	143	849				
22003	7		797	138	156	122	24				

Cancelar < Atrás Siguiente > Finalizar

Ilustración 27. Paso 5 para la transformación de datos, elaboración propia 2021.

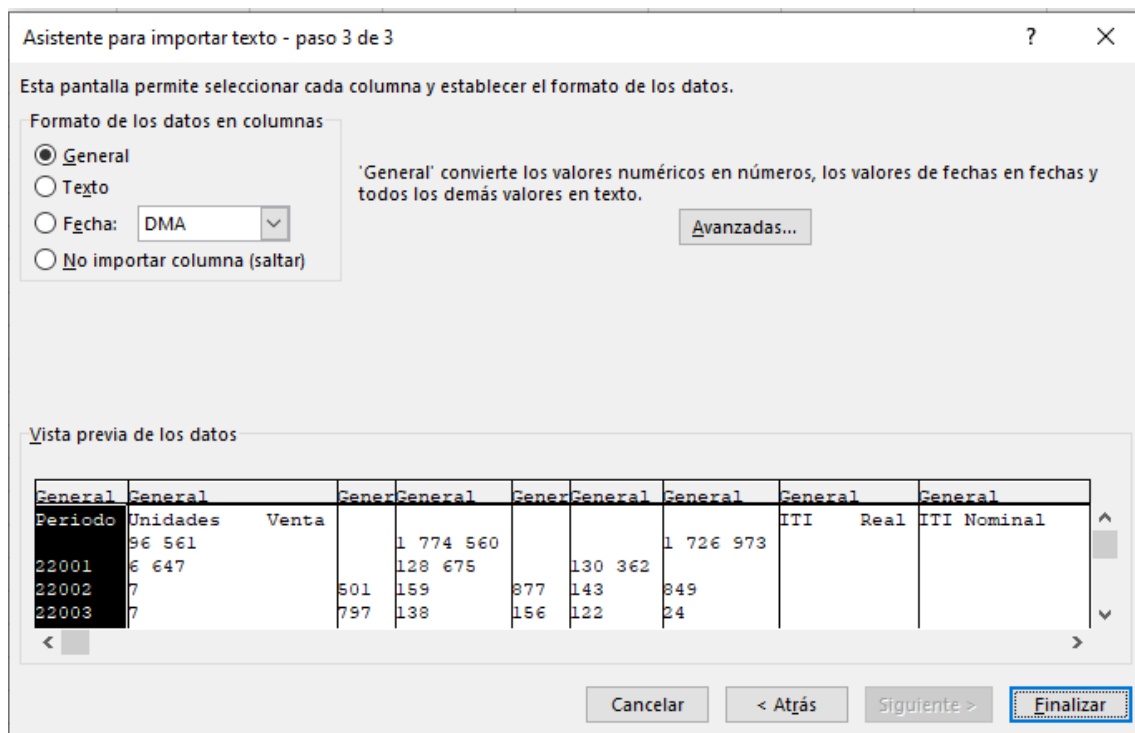


Ilustración 28. Paso 6 para la transformación de datos, elaboración propia 2021.

	A	B	C	D	E	F	G	H
1	Fecha	panaderia	cafes_leches_saborisantes	cereales_galletas_y_pan	comestibles_en_general_abarrotes	dulces_chocolates_botanas	aceites_granos_y_sopas	
2	2015-01		44,743	9,847	14,680	8,410	13,982	
3	2015-02		38,178	9,794	15,888	9,617	11,974	
4	2015-03		44,955	11,208	10,149	16,868	9,460	13,881
5	2015-04		42,125	8,137	8,081	14,657	8,917	12,463
6	2015-05		30,074	7,933	7,254	12,841	7,015	12,023
7	2015-06		28,512	7,748	7,303	12,805	6,880	11,432
8	2015-07		38,203	8,494	7,468	15,423	8,612	12,283
9	2015-08		42,308	9,554	9,071	13,098	8,660	17,489
10	2015-09		42,647	8,838	8,434	12,200	8,930	9,206
11	2015-10		36,431	10,277	7,864	13,827	9,144	13,477
12	2015-11		39,795	9,572	7,578	13,020	8,806	12,342
13	2015-12		74,433	10,778	7,106	15,896	11,141	14,561
14	2016-01		44,473	10,137	8,028	13,803	8,410	22,732
15	2016-02		35,324	9,950	9,032	13,872	9,617	12,474
16	2016-03		46,071	10,791	8,827	15,931	9,460	20,979
17	2016-04		35,228	10,752	8,106	13,970	8,917	12,962
18	2016-05		37,046	10,804	8,598	13,709	7,015	14,813
19	2016-06		35,920	12,382	14,462	19,979	6,880	24,547
20	2016-07		38,007	19,033	14,183	16,413	8,612	16,196
21	2016-08		36,305	11,007	9,528	16,249	8,660	15,283
22	2016-09		41,683	9,895	7,566	13,456	8,930	13,949
23	2016-10		36,520	10,783	9,260	14,733	9,144	16,429
24	2016-11		35,620	9,975	8,186	13,801	8,806	14,801
25	2016-12		65,183	11,986	8,066	17,462	11,141	17,094

Ilustración 29. Resultado final del proceso ETL de las unidades vendidas de cada categoría de producto, elaboración propia 2021.

Importar librerías

En nuestro análisis exploratorio se hizo la preparación y carga de los datos en un entorno de Jupyter Notebook en la nube con Google Colab, para esto se importó las librerías que permiten la carga de los datos con `pandas` y otra para su manejo de las visualizaciones como `seaborn` y `matplotlib`

A continuación, se muestra el listado total de las librerías usadas para el análisis exploratorio:

```
from google.colab import files
uploaded = files.upload()
import timeit
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import kn as sns
import datetime

from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
!pip install --upgrade kneed
from kneed import KneeLocator
from sklearn.metrics import silhouette_samples, silhouette_score
import matplotlib.cm as cm
%matplotlib inline
import io
```

Ilustración 30. listado de las librerías usadas en el Jupyter Notebook, elaboración propia 2021.

Al haber importado exitosamente las librerías realizamos propiamente la carga de la base de datos con la librería y el comando:

```
from google.colab import files
uploaded = files.upload()
```

Ilustración 31. librería usada para la carga de datos en Google Colab, elaboración propia 2021.

Carga y visualización de los datos

Este permite subir el archivo en la nube usando el entorno de nuestra máquina como se ve a continuación.

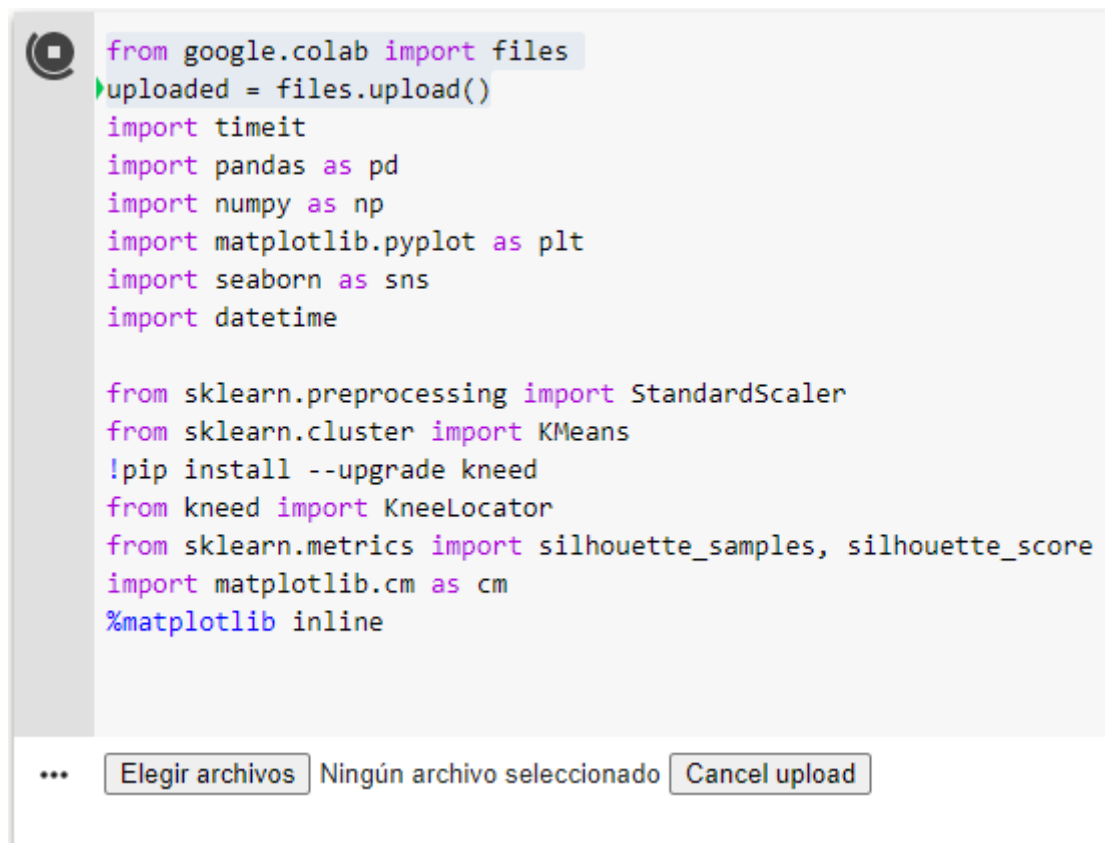


Ilustración 32. Ejecución y carga de las librerías, elaboración propia 2021.

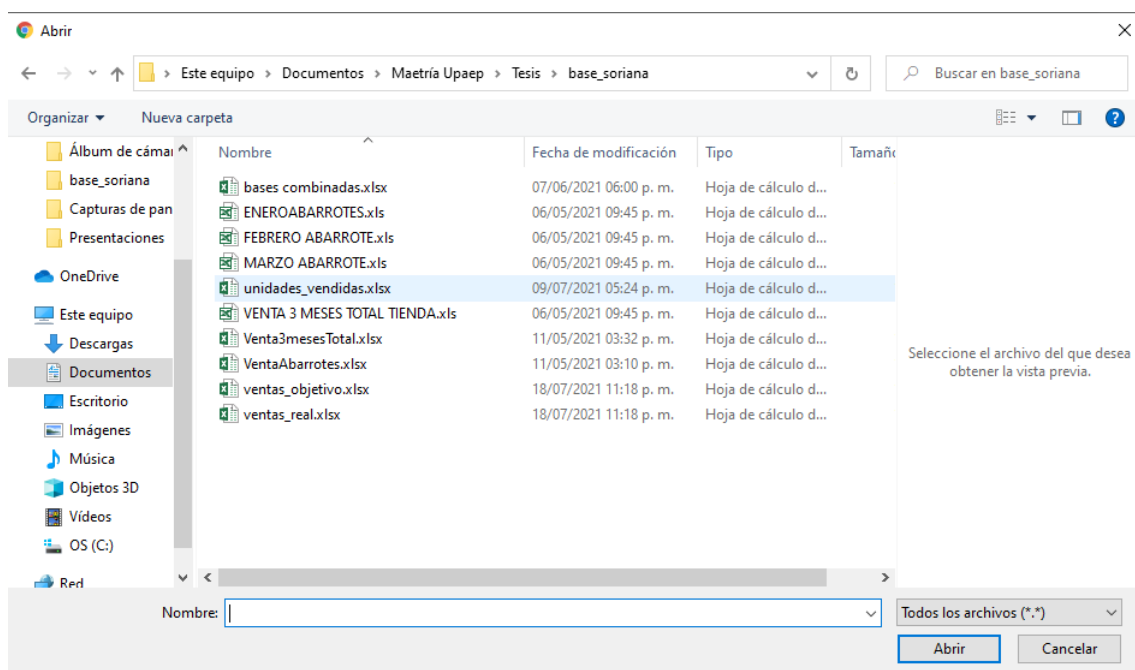


Ilustración 33. Carga de los datos "unidades vendidas" al entorno virtual, elaboración propia 2021.

De esta forma seleccionamos el archivo "unidades_vendidas.xlsx" el resultado exitoso de la sentencia es:

Elegir archivos unidades_vendidas.xlsx

- **unidades_vendidas.xlsx**(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 12591 bytes, last modified: 9/7/2021 - 100% done

Saving unidades_vendidas.xlsx to unidades_vendidas.xlsx

Collecting kneed

Downloading kneed-0.7.0-py2.py3-none-any.whl (9.4 kB)

Requirement already satisfied: matplotlib in /usr/local/lib/python3.7/dist-packages (from kneed) (3.2.2)

Requirement already satisfied: scipy in /usr/local/lib/python3.7/dist-packages (from kneed) (1.4.1)

Requirement already satisfied: numpy>=1.14.2 in /usr/local/lib/python3.7/dist-packages (from kneed) (1.19.5)

Requirement already satisfied: python-dateutil>=2.1 in /usr/local/lib/python3.7/dist-packages (from matplotlib->kneed) (2.8.1)

Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.7/dist-packages (from matplotlib->kneed) (0.10.0)

Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in /usr/local/lib/python3.7/dist-packages (from matplotlib->kneed) (2.4.7)

Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.7/dist-packages (from matplotlib->kneed) (1.3.1)

Requirement already satisfied: six in /usr/local/lib/python3.7/dist-packages (from cycler->0.10->matplotlib->kneed) (1.15.0)

Installing collected packages: kneed

Successfully installed kneed-0.7.0

Ilustración 34. Éxito de la carga de datos, elaboración propia 2021.

Aunque se cargó al entorno de la nube en Google Colab, requerimos también su carga en el notebook de Jupyter. Para realizarlo, se requiere ejecutar con pandas la siguiente línea de comandos para asignar el total de registros de la base a una variable que nos permitirá avanzar con el análisis.

```
# Carga de los documentos
import io
unidades = pd.read_excel(io.BytesIO(uploaded['unidades_vendidas.xlsx']))
# Importamos la base y mostramos los primeros y últimos 10 registros
unidades.head()
```

Ilustración 35. Asignación del dataset en la variable "unidades" y uso de la función head () para obtener el top 5 registros, elaboración propia 2021.

	Fecha	panaderia	cafes_leches_saborizantes	cereales_galletas_y_pan	comestibles_en_general_abarrotos	dulces_chocolates_botanas	aceites_granos_y_sopas
0	2015-01	44743	9847	9595	14680	8410	13982
1	2015-02	38178	9794	9867	15888	9617	11974
2	2015-03	44955	11208	10149	16868	9460	13881
3	2015-04	42125	8137	8081	14657	8917	12463
4	2015-05	30074	7933	7254	12841	7015	12023

Ilustración 36. top 5 registros con la función head (), elaboración propia

```
unidades.tail()
```

Ilustración 37. Uso de la función tail () para obtener últimos 5 registros, elaboración propia 2021.

	Fecha	panaderia	cafes_leches_saborizantes	cereales_galletas_y_pan	comestibles_en_general_abarrotos	dulces_chocolates_botanas	aceites_granos_y_sopas
67	2020-08	37517	7999	9426	10504	6119	13121
68	2020-09	42804	7182	7423	9725	5978	13539
69	2020-10	39077	7770	6581	10414	7980	14427
70	2020-11	42349	8656	7559	12291	7179	12629
71	2020-12	70582	9161	6462	12916	9300	13633

Ilustración 38. Resultado de la función tail (), elaboración propia 2021.

Se consiguió exitosamente la carga, se visualizó los primeros y últimos 5 registros de la base. En este primer vistazo se observa que el historial de los datos va mes a mes desde el 2015 hasta 2020; por otro lado, las 6 columnas restantes conforman a las categorías de los productos vendidos con base a producto desplazado.

Lo anterior permite tener una muestra de los datos, pero para poder dar inicio a un análisis estadístico requerimos de conocer el tipo de datos y la dimensión. Esto se llevará a cabo con el comando "unidades.info()", cuyo resultado es el siguiente:

```
[ ] unidades.info()
# Revisamos el tipo de datos y el tamaño de nuestra base
```

Ilustración 39. Uso de la función info() para conocer las características del dataset, elaboración propia 2021.

```

RangeIndex: 72 entries, 0 to 71
Data columns (total 7 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Fecha                                72 non-null     object
1   panaderia                           72 non-null     int64
2   cafes_leches_saborizantes           72 non-null     int64
3   cereales_galletas_y_pan              72 non-null     int64
4   comestibles_en_general_abarroses    72 non-null     int64
5   dulces_chocolates_botanas           72 non-null     int64
6   aceites_granos_y_sopas              72 non-null     int64
dtypes: int64(6), object(1)
memory usage: 4.1+ KB

```

Podemos observar que la base está compuesta por 72 registros y 7 dimensiones, todas son de tipo “int64” a excepción de la fecha la cual es de tipo “object”. Esto significa que la fecha esta guarda como texto y no como una variable de tipo “datetime”, además podemos empezar a intuir el tipo de análisis y procedimiento que tenemos que realizar al contar con variables de tipo numéricas discretas.

Para algunas visualizaciones convendrá mantener una copia de los datos por lo cual generaremos una nueva variable llamada “df” (abreviación de dataframe) para trabajar con ella después.

```
[ ] df = unidades
    df
```

Ilustración 40. Copia para trabajar el dataframe, elaboración propia 2021.

Una vez generado el respaldo, cambiamos el formato de la variable “fecha” a tipo DateTime con pandas con el código “pd.to_datetime(unidades['Fecha'], format='%Y-%m')”, por otra parte, este código permite manejar la fecha en un formato de años-mes.

```
[ ] #Cambiamos el formato de fecha a tipo datetime (año-mes-día) ya que aparece como tipo object
    unidades['Fecha'] = pd.to_datetime(unidades['Fecha'], format='%Y-%m')
```

Ilustración 41. Cambio de formato de la dimensión fecha, elaboración propia

Corroboramos y visualizamos el cambio repitiendo la línea de código .info() y .tail()

```
[ ] unidades.info() #Se corroboro el cambio de formato de la dimensión fecha
```

Ilustración 42. Corroboración del cambio de formato de la fecha, elaboración propia 2021.

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 72 entries, 0 to 71
Data columns (total 7 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Fecha                                72 non-null     datetime64[ns]
1   panaderia                           72 non-null     int64
2   cafes_leches_saborizantes           72 non-null     int64
3   cereales_galletas_y_pan              72 non-null     int64
4   comestibles_en_general_abarroses    72 non-null     int64
5   dulces_chocolates_botanas           72 non-null     int64

```

```
6 aceites_granos_y_sopas 72 non-null int64
dtypes: datetime64[ns](1), int64(6)
memory usage: 4.1 KB
```

```
[ ] unidades.tail()
```

Ilustración 43. Función para pre visualizar el cambio de formato, elaboración propia 2021.

	Fecha	panaderia	cafes_leches_saborizantes	cereales_galletas_y_pan	comestibles_en_general_abarrotos	dulces_chocolates_botanas	aceites_granos_y_sopas
67	2020-08-01	37517	7999	9426	10504	6119	13121
68	2020-09-01	42604	7182	7423	9725	5978	13539
69	2020-10-01	39077	7770	6581	10414	7980	14427
70	2020-11-01	42349	8656	7559	12291	7179	12629
71	2020-12-01	70582	9161	6462	12916	9300	13633

Ilustración 44. Pre visualización del cambio de formato, elaboración propia

Al transformar la variable a tipo “datetime” podremos separar la variable por mes y año. Lo cual será útil para su manejo de gráficas, esto se podrá hacer si creamos nuevas columnas que incluyan la separación de la fecha como se ve a continuación:

```
[ ] df['mes'] = df['Fecha'].dt.month
df['año'] = df['Fecha'].dt.year
```

Ilustración 45. Creación de la variable mes y año a partir de la variable fecha, elaboración propia 2021.

También generaremos nuevos dataframes para poder graficar individualmente las 6 categorías contrastándolos con los años y meses en un *stacked plot*. Los dataframes creados solo estarán compuestos por una categoría de producto y la fecha separada por mes y año, por lo que se borró la columna “fecha”.

```
df = df.drop(['Fecha'], axis=1)
df1= df.drop(['cafes_leches_saborizantes','cereales_galletas_y_pan','comestibles_en_general_abarrotos','dulces_chocolates_botanas',
'aceites_granos_y_sopas'], axis=1)
df2= df.drop(['panaderia','cereales_galletas_y_pan','comestibles_en_general_abarrotos','dulces_chocolates_botanas','aceites_granos_y_sopas'],
axis=1)
df3= df.drop(['panaderia','cafes_leches_saborizantes','comestibles_en_general_abarrotos','dulces_chocolates_botanas','aceites_granos_y_sopas'],
axis=1)
df4= df.drop(['panaderia','cafes_leches_saborizantes','cereales_galletas_y_pan','dulces_chocolates_botanas','aceites_granos_y_sopas'],
axis=1)
df5= df.drop(['panaderia','cafes_leches_saborizantes','cereales_galletas_y_pan','comestibles_en_general_abarrotos','aceites_granos_y_sopas'],
axis=1)
df6= df.drop(['panaderia','cafes_leches_saborizantes','cereales_galletas_y_pan','comestibles_en_general_abarrotos','dulces_chocolates_botanas'],
axis=1)
```

Ilustración 46. Nuevos dataframes para poder graficar individualmente las 6 categorías, elaboración propia 2021.

El resultado de esta transformación se observa en la ilustración 48.

```
[ ] df1
```

Ilustración 47. Función para desplegar del nuevo dataframe, elaboración propia 2021.

	panaderia	mes	año
0	44743	1	2015
1	38178	2	2015
2	44955	3	2015
3	42125	4	2015
4	30074	5	2015
...	...	...	...
67	37517	8	2020
68	42804	9	2020
69	39077	10	2020
70	42349	11	2020
71	70582	12	2020

72 rows × 3 columns

Ilustración 48. Visualización del nuevo dataframe, elaboración propia 2021.

Hay que recalcar que la separación y asignación de las fechas en las nuevas variable “mes” y “año” afectan tanto a “df” como “unidades”. Así que para el caso de unidades se necesita indexar la columna “fecha” y borrar la dimensión “mes” y “año”.

```
unidades.set_index('Fecha', inplace=True) #Indexamos las fechas para poder realizar agrupaciones en un futuro
```

Ilustración 49. Transformación de la variable fecha en un index, elaboración propia 2021.

```
[ ] unidades = unidades.drop(['mes', 'año'], axis= 1)
```

Ilustración 50. Eliminación de las variables repetidas, elaboración propia 2021.

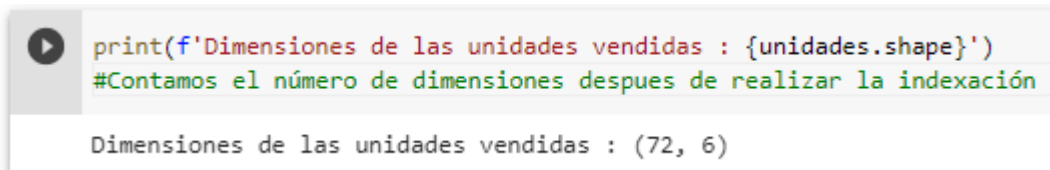
Después de haber ajustado la variable fecha, optamos por realizar un conteo de valores nulos en los registros, este peso será de utilidad para nuestro análisis descriptivo y predictivo. El conteo se realizó con la función `isnull().sum()` cuyo resultado mostró que no existen valores vacíos como se muestra en la ilustración 51 .

```
unidades.isnull().sum() #Comprobamos que no existan valores nulos
```

```
panaderia                0
cafes_leches_saborizantes 0
cereales_galletas_y_pan   0
comestibles_en_general_abarrote 0
dulces_chocolates_botanas 0
aceites_granos_y_sopas    0
dtype: int64
```

Ilustración 51. Comprobación de la existencia de valores nulos, elaboración propia 2021.

Aunque esta parte corresponde propiamente más a la limpieza de los datos que su análisis descriptivo, es fundamental para la correcta interpretación de los datos. Por tal motivo a manera de buenas prácticas imprimimos la dimensión de base una vez realizado los cambios finales, esto se observa en la ilustración 52.



```
print(f'Dimensiones de las unidades vendidas : {unidades.shape}')  
#Contamos el número de dimensiones despues de realizar la indexación
```

Dimensiones de las unidades vendidas : (72, 6)

Ilustración 52. Impresión del dataframe con la función .shape, elaboración propia 2021.



FASE 3:

Preparación de los datos (proceso ETL)

Análisis exploratorio

Medidas de tendencia central

Al tener la base lista, calcularemos las medidas de tendencia central con la función `.describe()`; esta nos arroja el total de registros, el promedio, desviación estándar, mediana, los cuartiles, el valor máximo y mínimo. Para un mejor análisis graficaremos en diagramas de caja las 6 categorías, de esta forma podremos una forma más visual las medidas de tendencia central.

```
unidades.describe()
```

	panaderia	cafes_leches_saborizantes	cereales_galletas_y_pan	comestibles_en_general_abarrotes	dulces_chocolates_botanas	aceites_granos_y_sopas
count	72.000000	72.000000	72.000000	72.000000	72.000000	72.000000
mean	40611.527778	11992.277778	10122.013889	16052.125000	8877.638889	19532.500000
std	10235.273993	7426.929123	3691.478587	5329.853963	1720.675983	11348.439782
min	28512.000000	6875.000000	6462.000000	9725.000000	5978.000000	9206.000000
25%	35435.750000	8783.250000	8077.250000	13194.000000	7653.500000	13759.750000
50%	37812.000000	9766.500000	9073.000000	14738.000000	8712.500000	16185.500000
75%	42318.250000	10797.250000	9918.750000	16380.000000	9337.250000	20500.000000
max	79996.000000	46291.000000	23119.000000	40108.000000	14708.000000	68014.000000

Ilustración 53. Medidas de tendencia central con la función `describe`, elaboración propia 2021.



Ilustración 54. Diagramas de caja del dataframe, elaboración propia 2021.

A partir de estos simples valores y de su representación gráfica podemos sacar las siguientes conjeturas:

- La panadería o panificadora interna de la tienda es el producto que más unidades mueve con rango normal de 49000 unidades aproximadamente y de 80000 en sus valores atípicos.
- Los dulces son el producto con el registro histórico más bajo en unidades desplazadas, aunque también es el más consistente, ya que no presenta outliers.
- Excluyendo la panadería; la categoría aceites, granos y sopas es el producto que más unidades mueve en promedio y con rango de 13759 a 20500 unidades.

- Las 3 categorías que en rango promedio no desplazan tantas unidades son 'cafes,leches y saborizantes', 'cereales, galletas y pan', 'dulces, chocolates y botanas'.
- No existen valores atípicos debajo del rango lo que nos indica que respecto al desplazamiento no ha habido merma.
- La alta presencia de valores atípicos u outliers puede deberse a los picos de venta obtenidos por las temporadas fuerte como diciembre o verano, pero para corroborarlo se requiere de un análisis de series temporales.
- El punto anterior podría indicar que los dulces, chocolates y botanas son un producto estable en el tiempo.

Series temporales

Debido a que contamos con registro histórico de los datos, se elaboró gráficas de series temporales para cada categoría de producto, es necesario mencionar que la indexación de la fecha permitió graficar de manera más eficiente. Además, se realizó un histograma el cual reforzará el análisis anterior.

```
unidades.panaderia.plot(title='# unidades vendidas de panaderia')
plt.show()
```

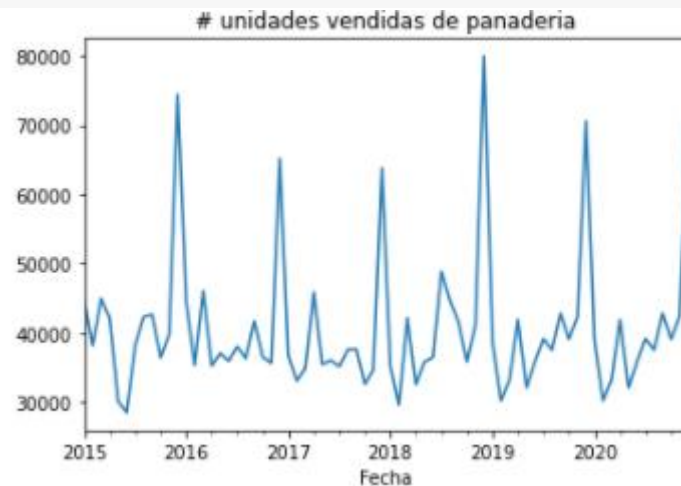


Ilustración 55. Serie temporal de las unidades vendidas de la panadería, elaboración propia 2021.

```
unidades.cafes_leches_saborizantes.plot(title='# unidades vendidas de cafes leches y saborizantes')
plt.show()
```

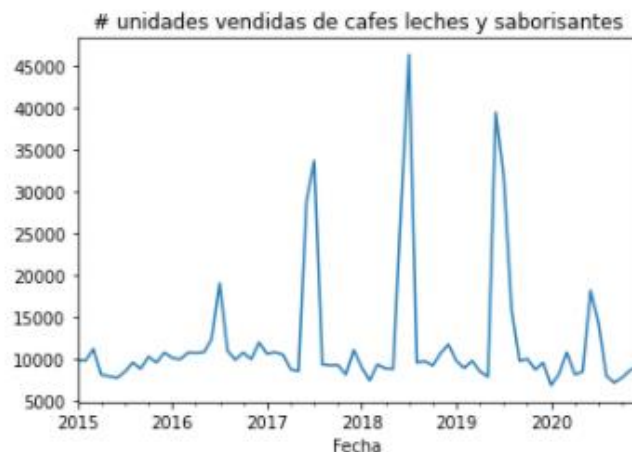


Ilustración 56. Serie temporal de las unidades vendidas de cafés, leches y saborizantes, elaboración propia 2021.

```

▶ unidades.cereales_galletas_y_pan.plot(title='# unidades vendidas de cereales galletas y pan')
plt.show()

```

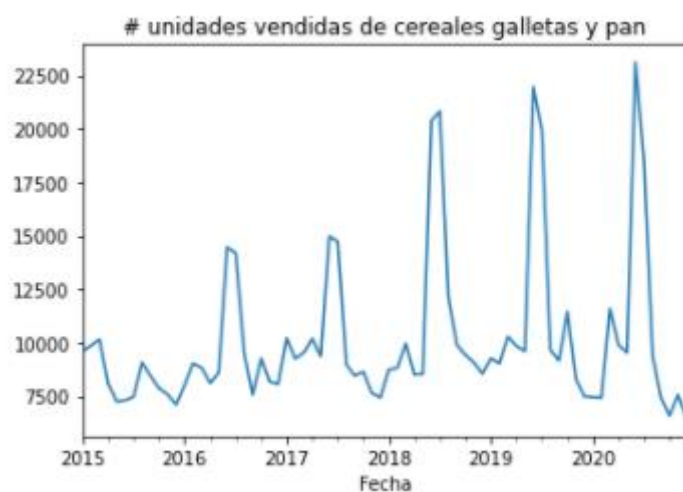


Ilustración 57. Serie temporal de las unidades vendidas de cereales, galletas y pan, elaboración propia 2021.

```

▶ unidades.comestibles_en_general_abarrotes.plot(title='# unidades vendidas de comestibles en general abarrotes')
plt.show()

```



Ilustración 58. Serie temporal de las unidades vendidas de comestibles en general y abarrotes, elaboración propia 2021.

```
[ ] unidades.dulces_chocolates_botanas.plot(title='# unidades vendidas de dulces chocolates y botanas')
plt.show()
```

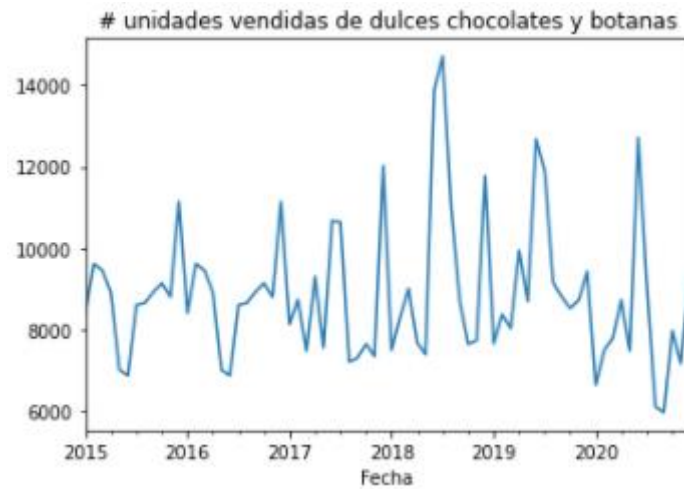


Ilustración 59. Serie temporal de las unidades vendidas de la dulces, chocolates y botanas, elaboración propia 2021.

```
unidades.aceites_granos_y_sopas.plot(title='# aceites granos y sopas')
plt.show()
```

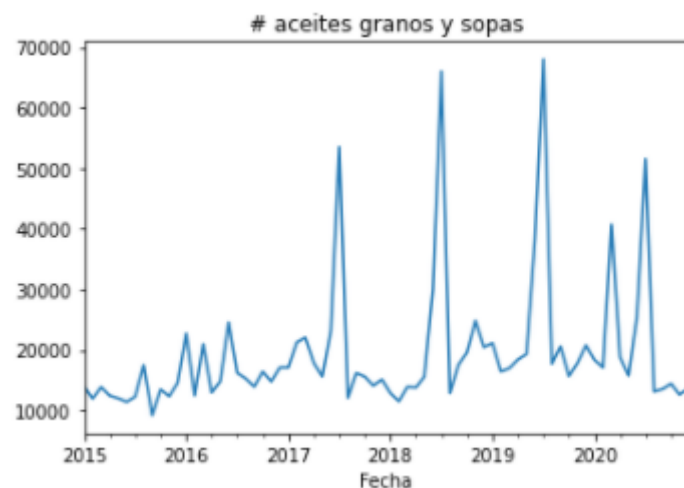


Ilustración 60. Serie temporal de las unidades vendidas de aceites, granos y sopas, elaboración propia 2021.

A partir de las series temporales podemos evaluar el comportamiento individual por temporadas de cada producto a lo largo del tiempo. En este análisis podemos observar:

- Todos los productos presentan picos en cada año y con frecuencia, lo cual indica que las ventas de temporada generan un movimiento fuerte de mercancía y este se mantiene año tras año.
- Los “cereales, galletas y pan” son los productos más desplazados proporcionalmente con respecto al inicio del 2015 hasta el 2020.
- “panadería” y “dulces, chocolates y botanas” son las categorías que han mantenido un ritmo estable con respecto a los 5 años transcurridos.
- “café, leche y saborizantes”, “comestibles en general y abarrotes”, “aceites, granos y sopas” tuvieron una caída de producto desplazado en el 2019 y 2020, posiblemente por la crisis mundial del Covid-19.

- “dulces, chocolates y botanas” son una categoría estable, pese a no presentar un patrón que corresponda a los picos de las temporadas como el resto de las categorías.
- “cafes leches y saborizantes”, “comestibles generales de abarrotes”, “aceites, granos y sopas” presentaron un bajo desplazamiento de producto en los primeros 3 años comparado con el 2018, 2019 y 2020.

En un análisis más completo de la evolución de las ventas en cada categoría de productos, realizamos el comparativo histórico en un gráfico de líneas, donde se muestra de cada categoría las unidades vendidas a lo largo del tiempo. El gráfico mencionado se crea y despliega como lo muestra la ilustración 61 respectivamente, con base a esta información podemos resaltar los siguientes puntos:

- Al observar picos recurrentes de ventas que representan las temporadas fuertes en un año y estas presentar una patron cíclico en cada año, podemos intuir que existe un hábito de consumo regular en la población.
- La panadería por lejos es el producto que mas unidades ha movido desde 2015 hasta 2020. Existe la excepción de los aceites, granos y sopas que desde el 2017 en cada mitad del año supera las unidades vendidas de pan.
- Excluyendo a la panaderia, las categorías que más venden son: “Aceites, granos y sopas”, “Cafes, leches y saborizantes” y “Comestibles en general y abarrotes”.
- Desde el 2017 las ventas reflejan un patrón de todas las categorías menos de la panadería, lo cual puede indicarnos que hasta ese año fue posible suplir la demanda y los hábitos de consumo del área.
- Los dulces, chocolates y botanas son la categoría que menor volumen de ventas tienen en los últimos 5 años.

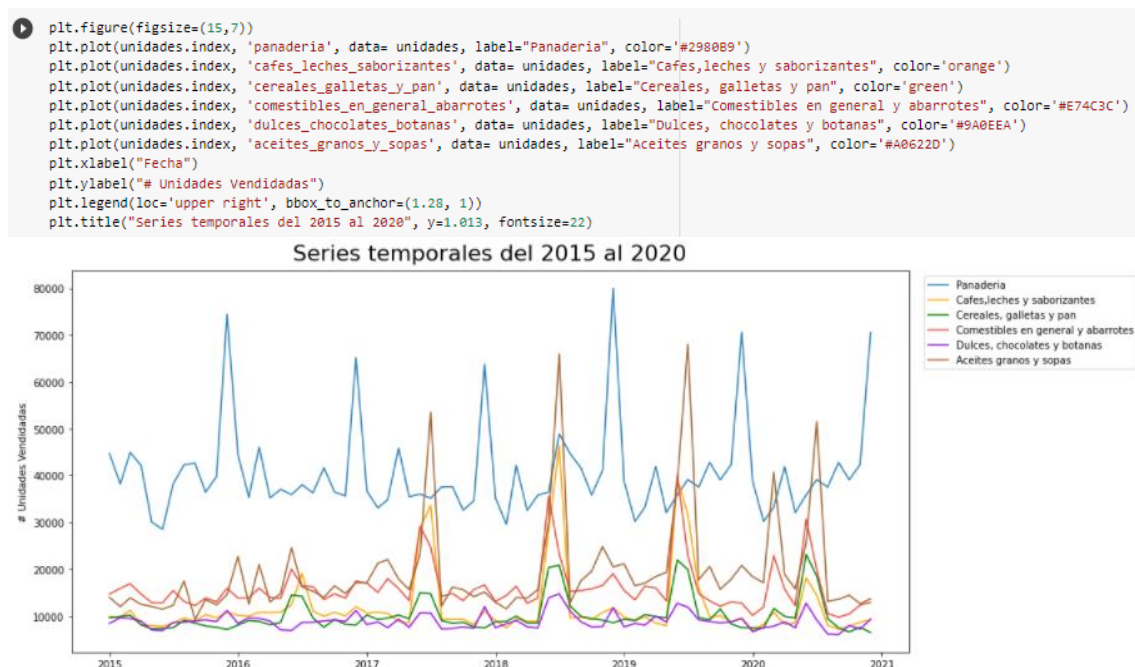


Ilustración 61. Serie temporal de las unidades vendidas de todas las categorías desde el 2015 hasta 2020, elaboración propia 2021.

Otro observación interesante a lo largo del tiempo es contrastar mes a mes los productos de cada categoría por los 5 años, para esto recurrimos a una *stacked plot* que mostrará los meses enumerados de cada año. Antes de remarcar los puntos importantes del

gráfico de barras agrupadas, se requiere cambiar el arreglo del dataframe donde se pivota la columna “año” e indexa la columna mes como se muestra en la ilustración 62.

```
df1 = df1.pivot(index="mes", columns="año", values="panaderia")
df2 = df2.pivot(index="mes", columns="año", values="cafes_leches_saborizantes")
df3 = df3.pivot(index="mes", columns="año", values="cereales_galletas_y_pan")
df4 = df4.pivot(index="mes", columns="año", values="comestibles_en_general_abarrotes")
df5 = df5.pivot(index="mes", columns="año", values="dulces_chocolates_botanas")
df6 = df6.pivot(index="mes", columns="año", values="aceites_granos_y_sopas")
```

Ilustración 62. Dataframes para comparar los productos de cada categoría por los 5 años mes a mes, elaboración propia 2021.

df1							
	año	2015	2016	2017	2018	2019	2020
mes							
1	44743	44473	36671	35215	38814	38814	
2	38178	35324	33069	29564	30198	30198	
3	44955	46071	34849	42153	33209	33209	
4	42125	35228	45875	32542	41898	41898	
5	30074	37046	35473	35786	32089	32089	
6	28512	35920	36029	36482	35848	35848	
7	38203	38007	35144	48886	39147	39147	
8	42308	36305	37574	44677	37517	37517	
9	42647	41683	37617	41574	42804	42804	
10	36431	36520	32599	35852	39077	39077	
11	39795	35620	34642	41095	42349	42349	
12	74433	65183	63818	79996	70582	70582	

Ilustración 63. Resultado del Dataframe creado para comparar mes a mes, elaboración propia 2021.

Una vez realizada la configuración de los dataframe mostrado en la ilustración 63, graficamos las comparaciones mes a mes de las distintas categorías. En cada una ellas se visualizó el contraste mes a mes por lo que se puede obtener las siguientes supuestos:

- Con estos gráficos podemos vislumbrar los picos de temporada alta, los cuales en promedio son marzo, junio, julio y diciembre.
- “panaderia” mueve su mayor cantidad de productos historicamente en diciembre. Además, cabe destacar que considerando al 2019 y 2020 como una epoca de ventas bajas por la pandemia los meses de abril, septiembre, octubre y noviembre no sufrieron una recaída drastica como se esperaba.
- “cafes leches y saborizantes” mueve su mayor cantidad de productos historicamente en junio y julio. En el resto de los meses mantuvo una venta regular pero baja de unidades con excepción de agosto 2019.

- “cereales galletas y pan” mueve su mayor cantidad de productos historicamente en junio y julio. Al igual que la panadería, si se considera los años de 2019 y 2020 como una época de ventas bajas los meses de abril y mayo no sufrieron de esta recaída, reflejando un buen manejo de la tienda. Además, en marzo en 2019 y octubre 2020 aumentaron sus ventas por unidad lo que demuestra una toma de decisiones acertada.
- “comestibles en general y abarrotes” mueve su mayor cantidad de productos historicamente en junio y julio. El resto de los meses la venta fue regular a excepción de marzo que tuvo un incremento notable en sus ventas de 2019.
- “dulces, chocolates y botanas” mueve la mayor cantidad de productos historicamente en junio, julio y diciembre. En los años 2019 y 2020 existió una baja de ventas en comparación a los años anteriores exceptuando los meses de mayo, abril y junio, pero por otro lado, las ventas más bajas de esos años e historicamente fueron en agosto, septiembre y diciembre.
- “aceites, granos y sopas” mueve la mayor cantidad de productos historicamente en marzo, junio y julio. Las ventas de marzo 2020 ha sido la mejor de todos los años con respecto a ese mes, cabe destacar que esta fecha en esta categoría ha sido la que mejores resultados presenta desde el inicio de la pandemia.

```
df1.plot(kind='bar', figsize=(17, 10), rot=0)
plt.title("Historical # unidades de panaderia: Mes por Mes Comparacion", y=1.013, fontsize=22)
plt.xlabel("Date [Mes]", labelpad=16)
plt.ylabel("Count [# unidades]", labelpad=16);
```

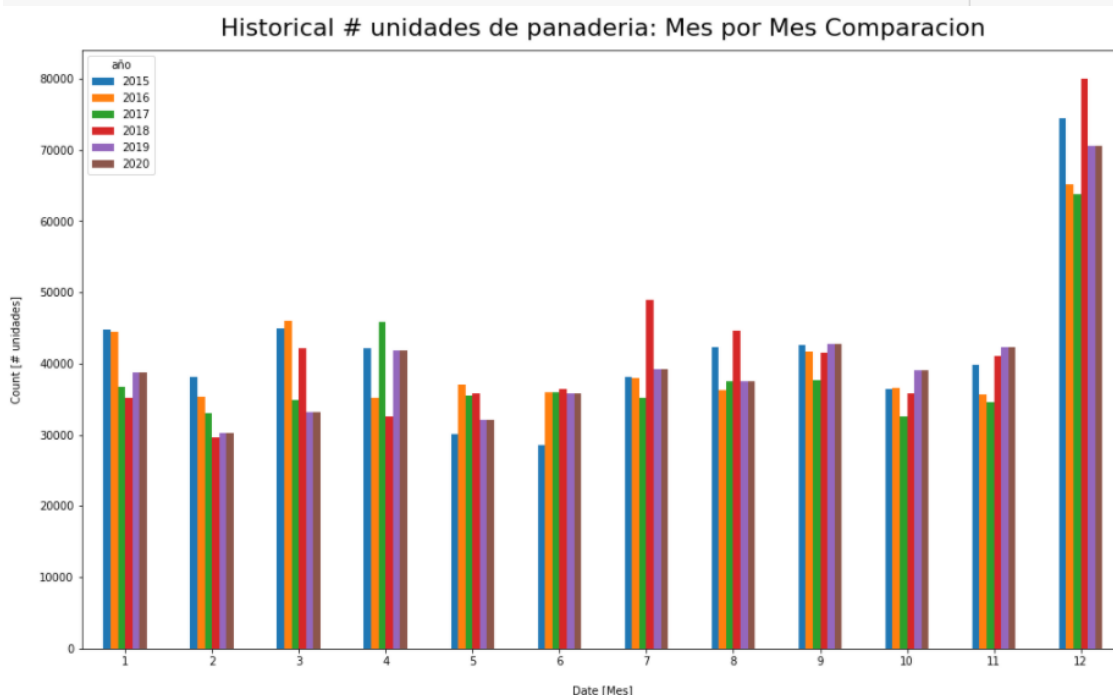


Ilustración 64. Comparación mes por mes de las unidades vendidas de panadería, elaboración propia 2021.

```
df2.plot(kind='bar', figsize=(17, 10), rot=0)
plt.title("Historical # unidades de cafes, leches y saborizantes: Mes por Mes comparacion", y=1.013, fontsize=22)
plt.xlabel("Date [Mes]", labelpad=16)
plt.ylabel("Count [# unidades]", labelpad=16);
```

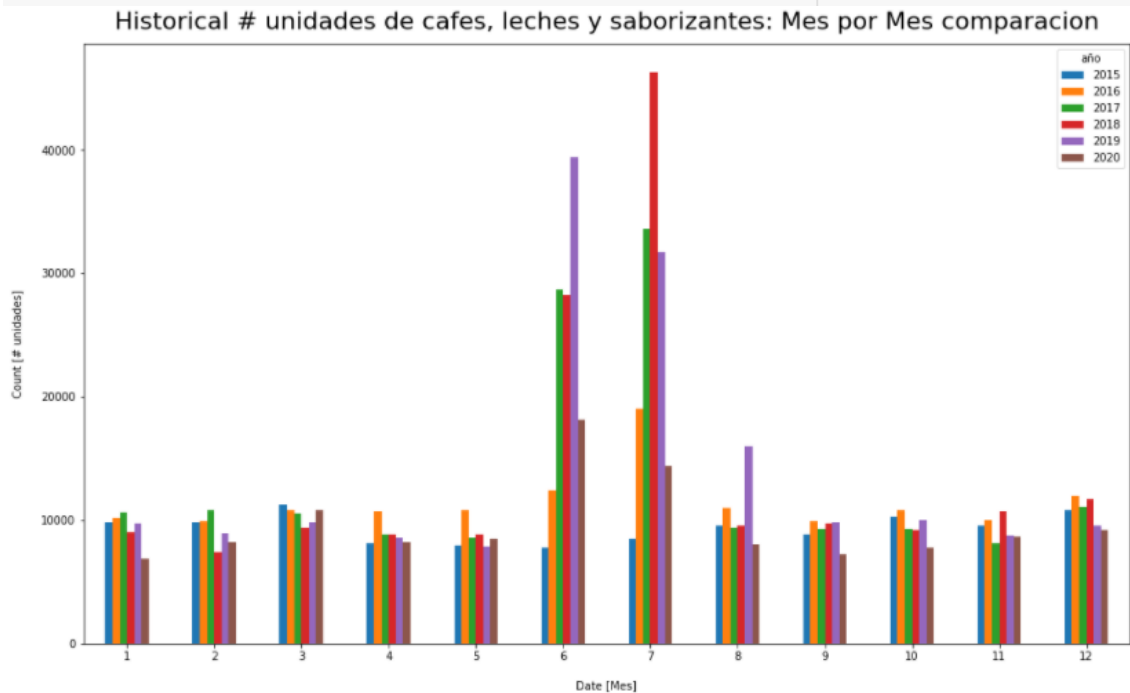


Ilustración 65. Comparación mes por mes de las unidades vendidas de cafés, leches y saborizantes, elaboración propia 2021.

```
df3.plot(kind='bar', figsize=(17, 10), rot=0)
plt.title("Historical # unidades de cereales, galletas y pan: Mes por Mes Comparacion", y=1.013, fontsize=22)
plt.xlabel("Date [Mes]", labelpad=16)
plt.ylabel("Count [# unidades]", labelpad=16);
```

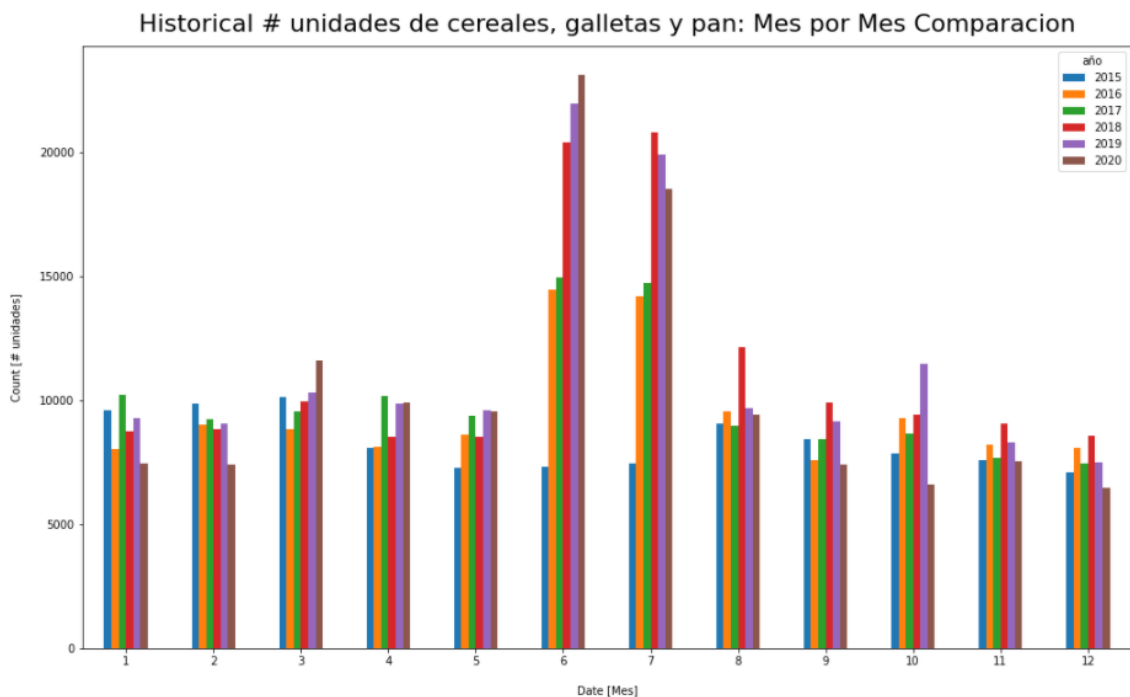


Ilustración 66. Comparación mes por mes de las unidades vendidas de cereales, galletas y pan, elaboración propia 2021.

```
[65] df4.plot(kind='bar', figsize=(17, 10), rot=0)
plt.title("Historical # unidades de comestibles en general abarrotes: Mes por Mes Comparasion", y=1.013, fontsize=22)
plt.xlabel("Date [Mes]", labelpad=16)
plt.ylabel("Count [# unidades]", labelpad=16);
```

Historical # unidades de comestibles en general abarrotes: Mes por Mes Comparasion

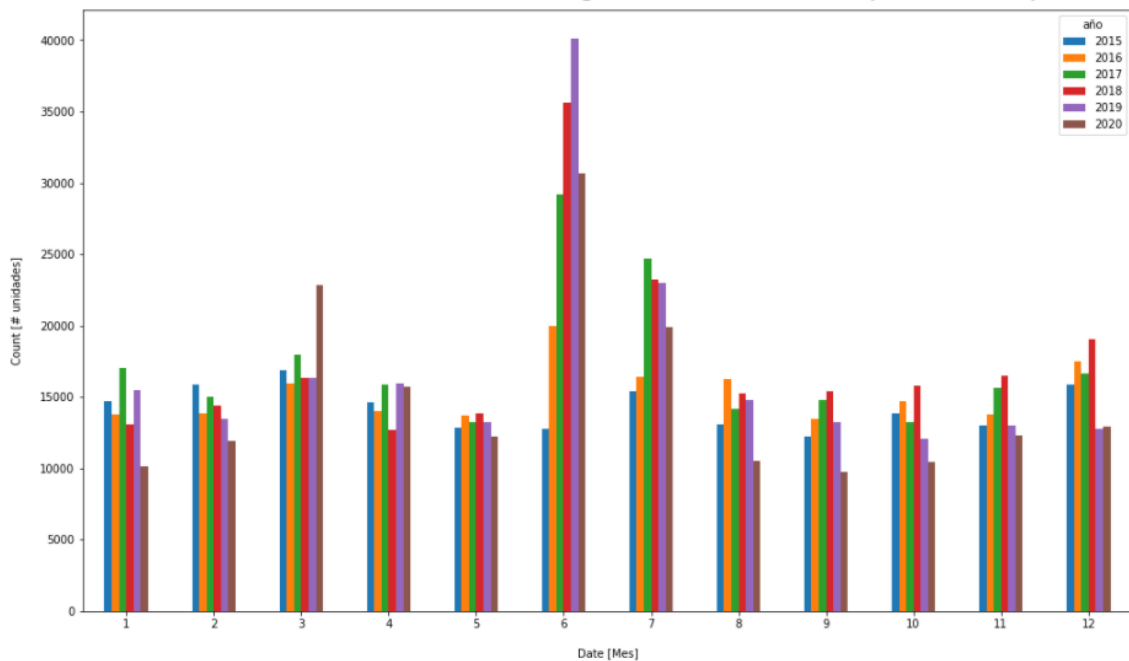


Ilustración 67. Comparación mes por mes de las unidades vendidas de comestibles en general y abarrotes, elaboración propia 2021.

```
[66] df5.plot(kind='bar', figsize=(17, 10), rot=0)
plt.title("Historical # unidades de dulces, chocolates y botanas: Mes por Mes Comparasion", y=1.013, fontsize=22)
plt.xlabel("Date [Mes]", labelpad=16)
plt.ylabel("Count [# unidades]", labelpad=16);
```

Historical # unidades de dulces, chocolates y botanas: Mes por Mes Comparasion

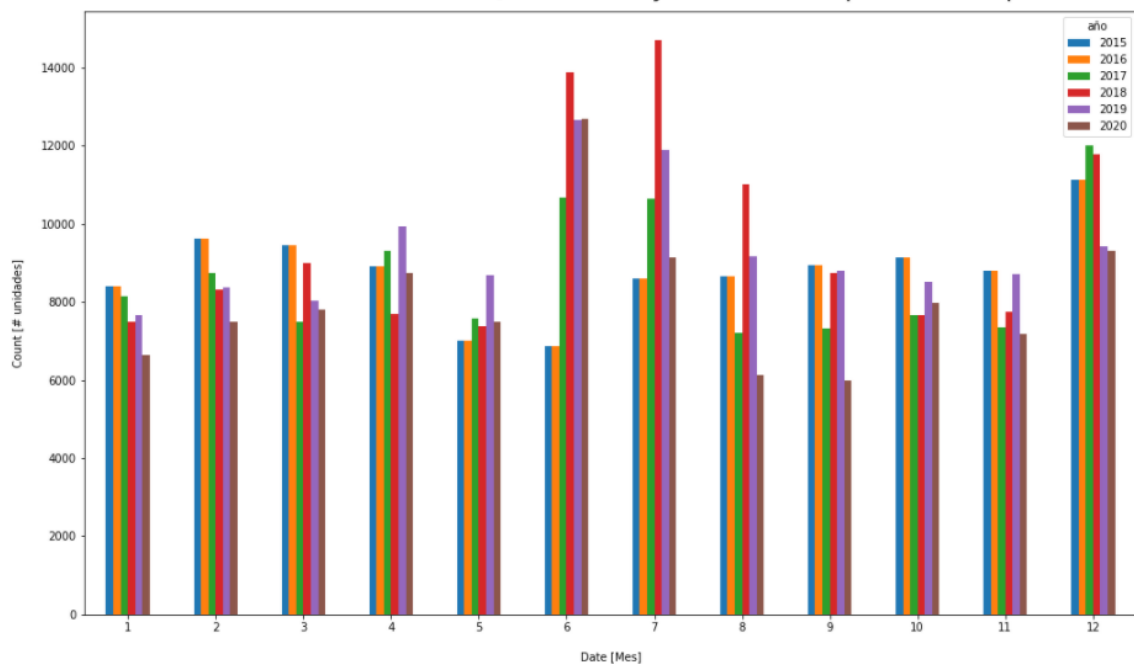


Ilustración 68. Comparación mes por mes de las unidades vendidas de dulces, chocolates y botanas, elaboración propia 2021.

```
df6.plot(kind='bar', figsize=(17, 10), rot=0)
plt.title("Historical # unidades de aceites, granos y sopas: Mes por Mes Comparasion", y=1.013, fontsize=22)
plt.xlabel("Date [Mes]", labelpad=16)
plt.ylabel("Count [# unidades]", labelpad=16);
```

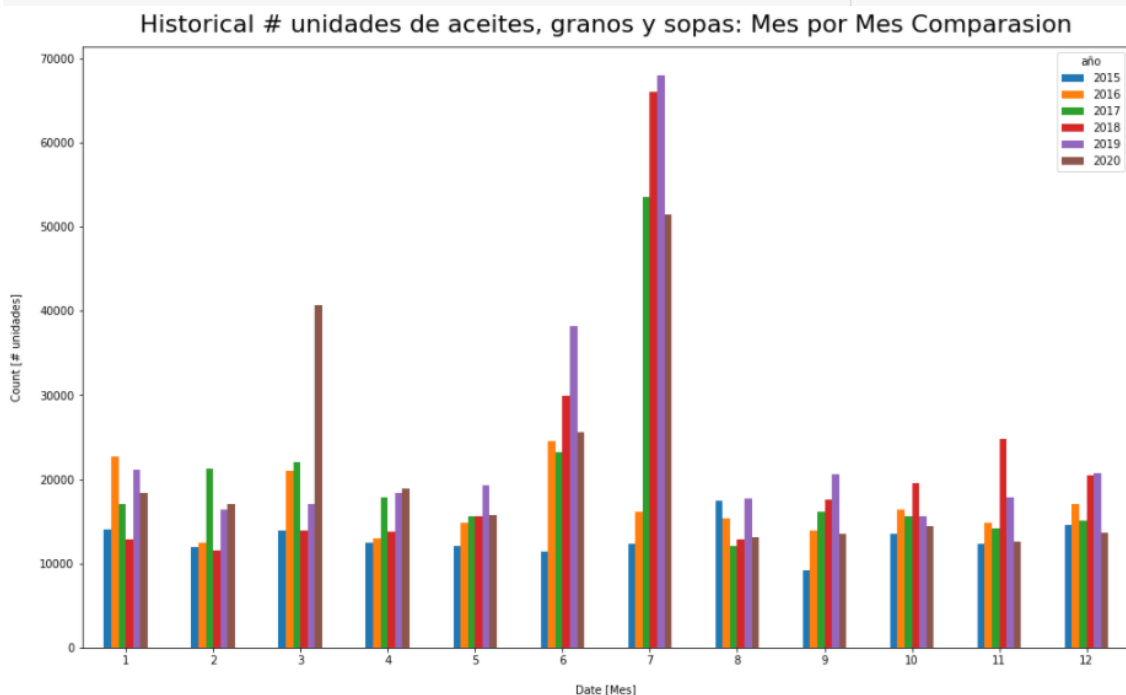


Ilustración 69. Comparación mes por mes de las unidades vendidas de aceites, granos y sopas, elaboración propia 2021.

Histograma y distribución

El código como se muestra en la ilustración 70 se generó con la función seaborn y con un *loop for* para así mostrar en una sola figura el histograma de las 6 categorías.

```
# Gráfico de distribución para cada variable numérica
# =====
# Ajustar número de subplots en función del número de columnas
fig, axes = plt.subplots(nrows=3, ncols=2, figsize=(15, 9))
axes = axes.flat
columnas_numeric = unidades.select_dtypes(include=['int64']).columns

for i, column in enumerate(columnas_numeric):
    sns.histplot(
        data = unidades,
        x = column,
        stat = "count",
        kde = True,
        color = (list(plt.rcParams['axes.prop_cycle'])*2)[i]["color"],
        line_kws= {'linewidth': 2},
        alpha = 0.3,
        ax = axes[i]
    )
    axes[i].set_title(column, fontsize = 10, fontweight = "bold")
    axes[i].tick_params(labelsize = 12)
    axes[i].set_xlabel("")

fig.tight_layout()
plt.subplots_adjust(top = 0.9)
fig.suptitle('Distribución variables numéricas', fontsize = 10, fontweight = "bold");
```

Ilustración 70. Código para la creación de histogramas de las 6 categorías, elaboración propia 2021.

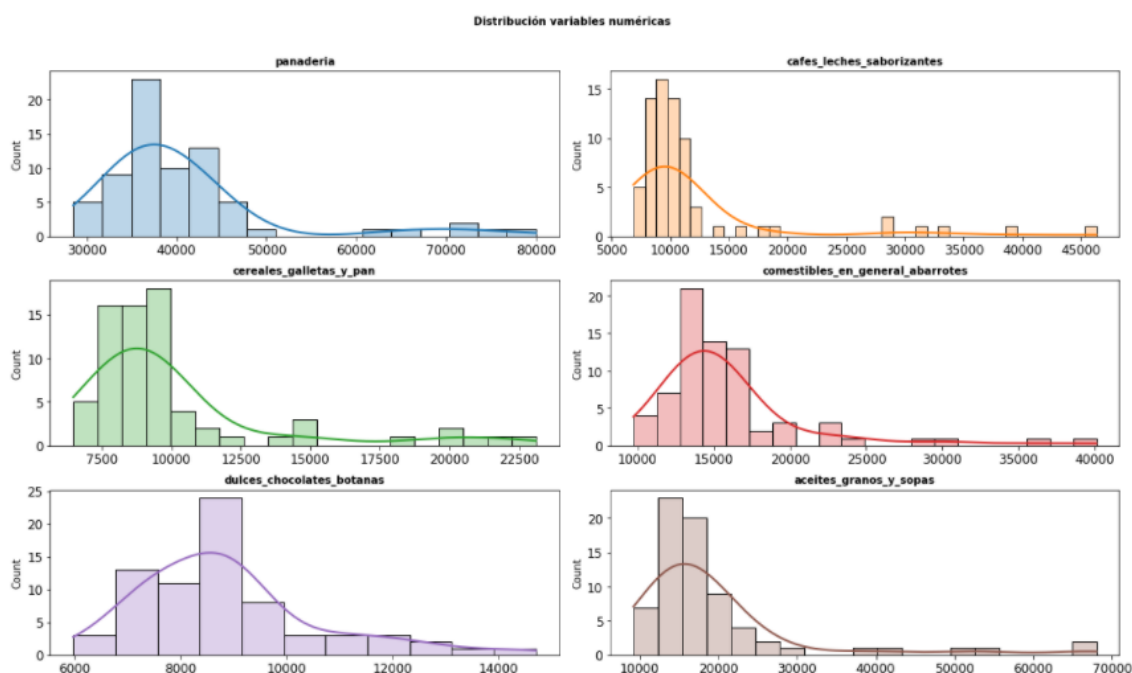


Ilustración 71. Histograma de las 6 categorías, elaboración propia 2021.

El histograma de cada producto presenta una distribución semi normal sesgada a la derecha, lo cual corresponde al análisis previo del *boxplot* que mostraba outliers positivos o superiores. Este nuevo gráfico permite obtener datos acerca de la frecuencia con la que se ha desplazado el producto con respecto a sus unidades, con base a lo anterior los puntos más interesantes podemos extrapolar son:

- Nos va a interesar siempre tener una distribución normal o sesgada a la izquierda en nuestras categorías, ya que esto significaría que hay una frecuencia alta de productos vendidos por unidad. Por otra parte, mientras más sesgado a la izquierda significará que mayor cantidad producto se ha desplazado, esto siempre y cuando haya existido temporadas donde se haya vendido menos porque si no la distribución no presentaría sesgo.
- El sesgo presente en cada categoría corresponde a los picos altos de venta por temporada.
- “dulces, chocolates y botanas” mueven con suma frecuencia cantidades de 7000 a 10000 unidades, además de presentar una mayor regularidad en este aspecto en comparación al resto de categorías.
- “panaderia” mueven con suma frecuencia cantidades de 35000 a 45000 unidades.
- “cafes leches y saborizantes” mueven con suma frecuencia cantidades de 7500 a 12500 unidades.
- “cereales galletas y pan” mueven con suma frecuencia cantidades de 3750 a 11000 unidades aproximadamente.
- “comestibles en general y abarrote”, mueven con suma frecuencia cantidades de 10000 a 17500 unidades.
- “aceites, granos y sopas”, mueven con suma frecuencia cantidades de 10000 a 20000 unidades.

Dispersión y correlación

En el análisis de las categorías es necesario contrastarlo en gráficos de dispersión para poder observar las tendencias y relaciones entre cada una de las dimensiones, brindando una herramienta que sirve para visualizar la dependencia en este caso de los productos y sus ventas. Además, se puede observar la tendencia de un producto a lo largo del año y estimar las ventas aunque en este trabajo practico no se realizara ese enfoque.

Se empezó por graficar cada producto con respecto a la fecha, la creación de las gráficas se pueden observar en las ilustraciones 72, 73, 74, 75, 76 y 77. En las gráficas de dispersion de cada producto con respecto al tiempo no se pudo encontrar ningun tipo de correlación.

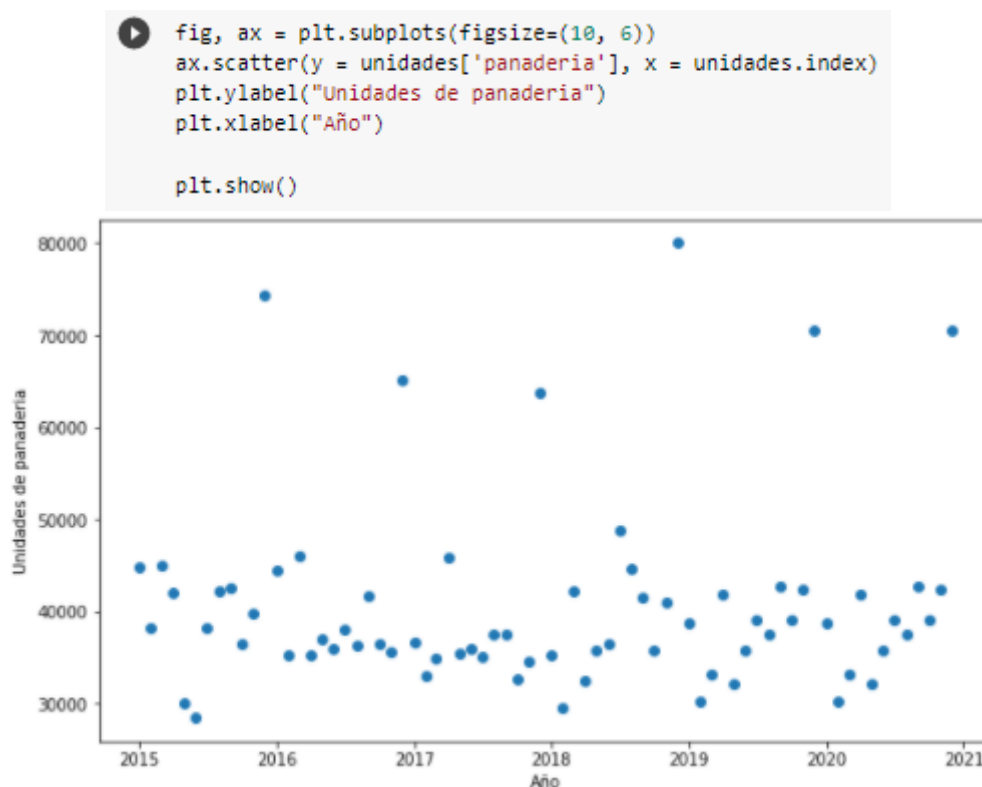


Ilustración 72. Gráficas de dispersión de las unidades vendidas de panadería, elaboración propia 2021.

```
[41] fig, ax = plt.subplots(figsize=(10, 6))
      ax.scatter(y = unidades['cafes_leches_saborisantes'], x = unidades.index)
      plt.ylabel("Unidades de cafes, leches y saborisantes")
      plt.xlabel("Año")
```

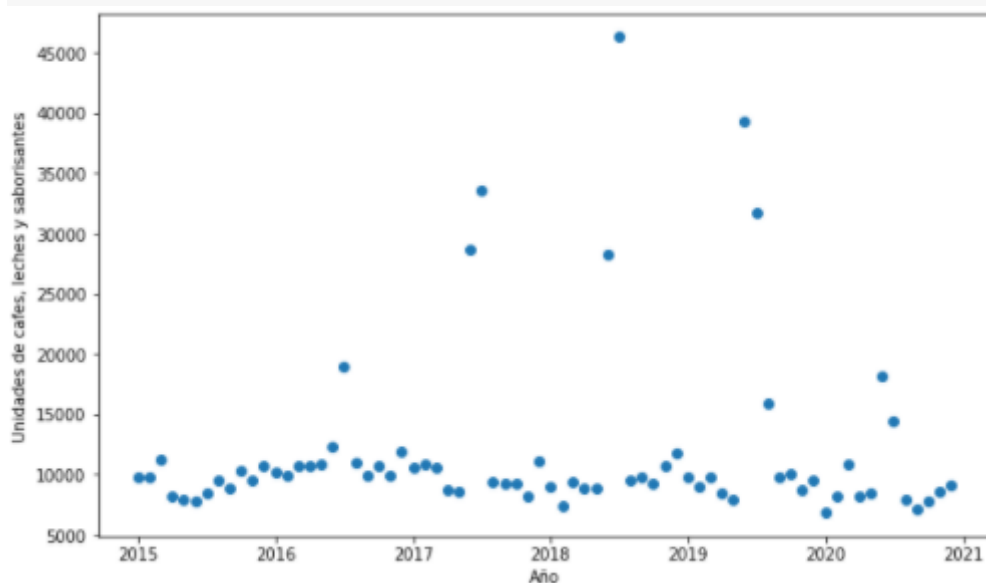


Ilustración 73. Gráficas de dispersión de las unidades vendidas de cafés, leches y saborizantes, elaboración propia 2021.

```
fig, ax = plt.subplots(figsize=(10, 6))
ax.scatter(y = unidades['cereales_galletas_y_pan'], x = unidades.index)
plt.ylabel("Unidades de cereales, galletas y pan")
plt.xlabel("Año")
```

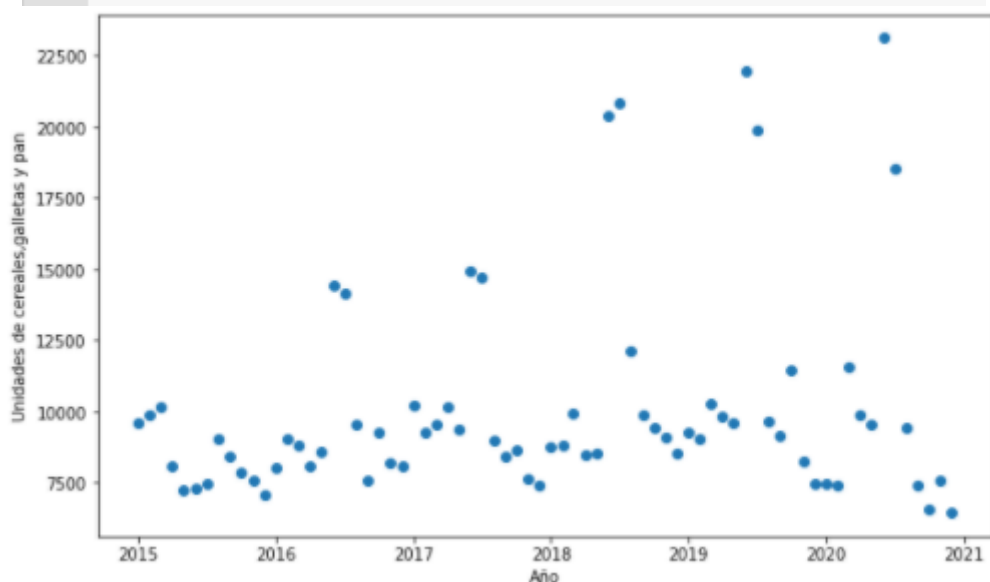


Ilustración 74. Gráficas de dispersión de las unidades vendidas de cereales, galletas y pan, elaboración propia 2021.

```
fig, ax = plt.subplots(figsize=(10, 6))
ax.scatter(y = unidades['comestibles_en_general_abarrotes'], x = unidades.index)
plt.ylabel("Unidades de comestibles en general y abarrotes")
plt.xlabel("Año")
```

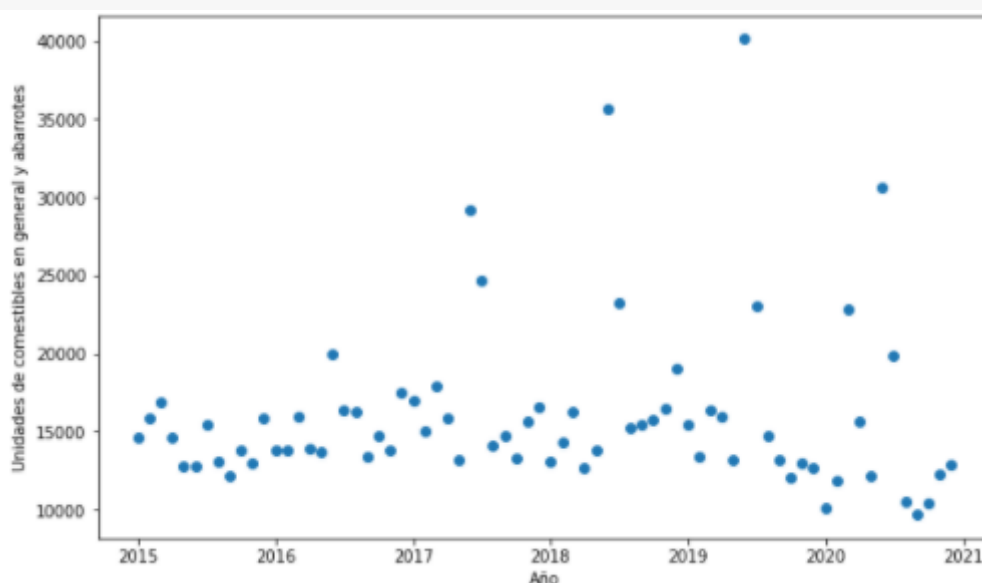


Ilustración 75. Gráficas de dispersión de las unidades vendidas de comestibles en general y abarrotes, elaboración propia 2021.

```
fig, ax = plt.subplots(figsize=(10, 6))
ax.scatter(y = unidades['dulces_chocolates_botanas'], x = unidades.index)
plt.ylabel("Unidades de dulces, chocolates y botanas")
plt.xlabel("Año")
```

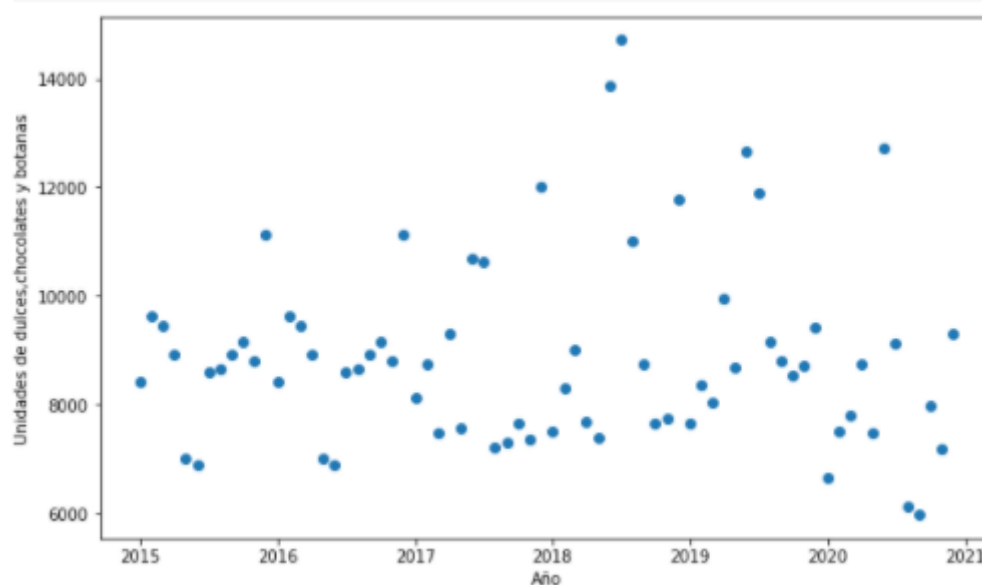


Ilustración 76. Gráficas de dispersión de las unidades vendidas de dulces, chocolates y botanas, elaboración propia 2021.

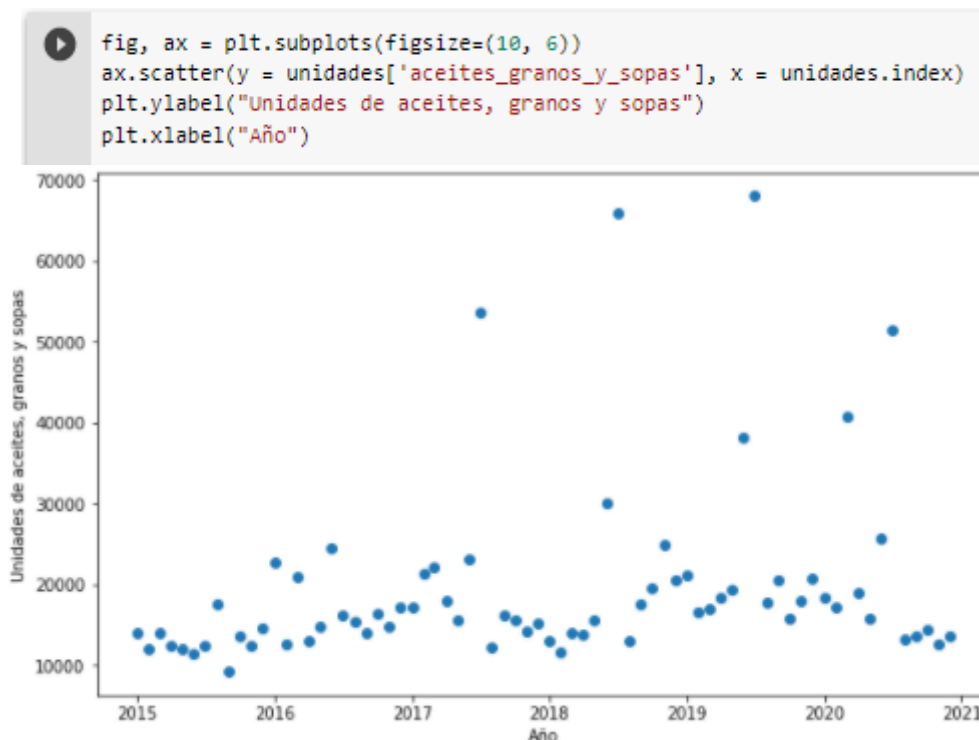


Ilustración 77. Gráficas de dispersión de las unidades vendidas de aceites, granos y sopas, elaboración propia 2021.

Ahora nos interesa saber si existe una correlación entre los productos, para lo cual usamos el dataframe “df” asignándolo en una variable nueva llamada “correlacion_df”. En la variable nueva se aplicó las transformaciones que nos permita observar las correlaciones como se puede ver en la ilustración 78 y 79.

```
[46] correlacion_df = df
correlacion_df
```

	panaderia	cafes_leches_saborizantes	cereales_galletas_y_pan	comestibles_en_general_abarrotes	dulces_chocolates_botanas	aceites_granos_y_sopas	mes	año
0	44743	9847	9595	14680	8410	13982	1	2015
1	38178	9794	9867	15888	9617	11974	2	2015
2	44955	11208	10149	16868	9460	13881	3	2015
3	42125	8137	8081	14657	8917	12463	4	2015
4	30074	7933	7254	12841	7015	12023	5	2015
...	...	...	...	...	...	...	...	...
67	37517	7999	9426	10504	6119	13121	8	2020
68	42804	7182	7423	9725	5978	13539	9	2020
69	39077	7770	6581	10414	7980	14427	10	2020
70	42349	8656	7559	12291	7179	12629	11	2020
71	70582	9161	6462	12916	9300	13633	12	2020

Ilustración 78. Nuevo dataframe para hacer transformaciones que permitan mostrar las correlaciones entre las variables, elaboración propia 2021

Al contar con la copia de trabajo del dataframe “correlacion_df” se realizó una relación cruzada de todas las variables con el código siguiente:

```
correlacion_df = correlacion_df.drop(['año'], axis=1)
sns.pairplot(correlacion_df, hue="mes", corner=True)

plt.savefig('correlaciones unidades de productos.png')
files.download("correlaciones unidades de productos.png")
```

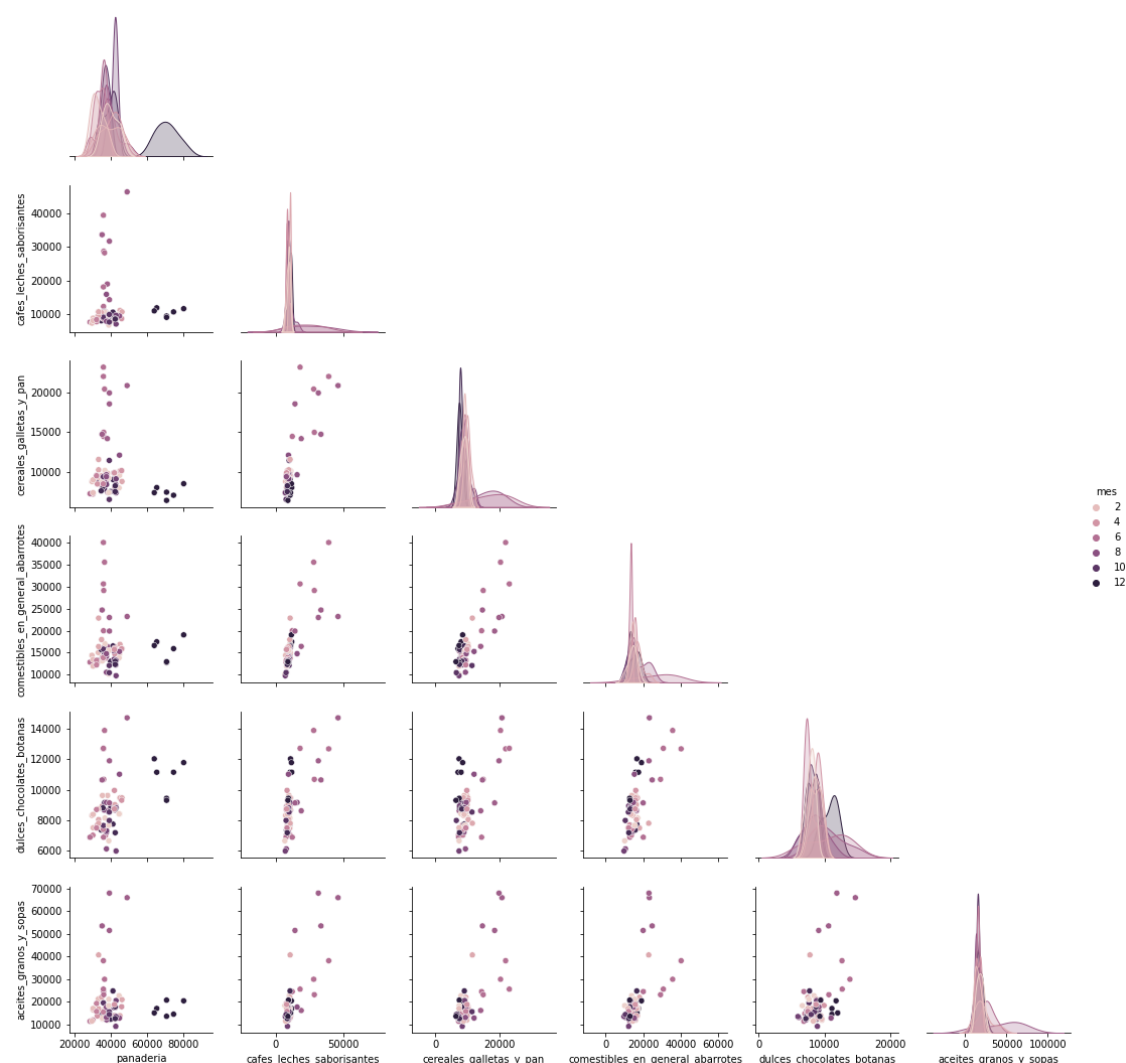


Ilustración 79. Código y visualización de las correlaciones entre todas las variables por mes en un gráfico de dispersión, elaboración propia 2021.

Ha diferencia de las gráficas anteriores se puede observar que existe una posible tendencia y correlación positiva, siendo estas tendencias indicadores de una dependendencia entre los productos respecto al incremento de ventas por unidad. Sin embargo, la dependendencia de los productos con base a la correlación requiere de estadandarizar los datos normalizandolos para así tener un distibución normal que no se vea afectada por la escala de los datos de las 6 categorías de productos.

La estandarización de datos se llevo acabo con el metodo de min-max en donde cada registro es rescalado en un rango de 0 a 1 a través de los valores extremos del rango original de cada columna. Esta modificación permite elaborar un mapa de calor por medio de correlaciones optimas que no se vea afectado por las magnitudes de cada categorías.

```
[48] df_normalizada = df
df_normalizada = df_normalizada.drop(['año'], axis=1)
#df_normalizada = df_normalizada.drop(['año','mes'], axis=1)
```

```
df_normalizada
```

Ilustración 80. Creación de un dataframe para su normalización, elaboración propia 2021.

	panaderia	cafes_leches_saborizantes	cereales_galletas_y_pan	comestibles_en_general_abarrotes	dulces_chocolates_botanas	aceites_granos_y_sopas	mes
0	44743	9847	9595	14680	8410	13982	1
1	38178	9794	9867	15888	9617	11974	2
2	44955	11208	10149	16868	9460	13881	3
3	42125	8137	8081	14657	8917	12463	4
4	30074	7933	7254	12841	7015	12023	5
...	...	...	...	...	...	...	...
67	37517	7999	9426	10504	6119	13121	8
68	42804	7182	7423	9725	5978	13539	9
69	39077	7770	6581	10414	7980	14427	10
70	42349	8656	7559	12291	7179	12629	11
71	70582	9161	6462	12916	9300	13633	12

72 rows x 7 columns

Ilustración 81. Resultado del dataframe creado, elaboración propia 2021.

```
#Normalización del dataset
#df.head()
df_normalizada = (df_normalizada-df_normalizada.min()/(df_normalizada.max()-df_normalizada.min()))
#scaler = StandardScaler()
"""df_normalizada[['panaderia','cafes_leches_saborizantes',
                    'cereales_galletas_y_pan','comestibles_en_general_abarrotes',
                    'dulces_chocolates_botanas','aceites_granos_y_sopas']] = scaler.fit_transform(
                    df[['panaderia','cafes_leches_saborizantes','cereales_galletas_y_pan',
                        'comestibles_en_general_abarrotes',
                        'dulces_chocolates_botanas',
                        'aceites_granos_y_sopas']])
df_normalizada"""

sns.kdeplot(df_normalizada['panaderia'], label='panaderia')
sns.kdeplot(df_normalizada['cafes_leches_saborizantes'], label='cafes_leches_saborizantes')
sns.kdeplot(df_normalizada['cereales_galletas_y_pan'], label='cereales_galletas_y_pan')
sns.kdeplot(df_normalizada['comestibles_en_general_abarrotes'], label='comestibles_en_general_abarrotes')
sns.kdeplot(df_normalizada['dulces_chocolates_botanas'], label='dulces_chocolates_botanas')
sns.kdeplot(df_normalizada['aceites_granos_y_sopas'], label='aceites_granos_y_sopas')
plt.title("Distribución normalizada de los productos", y=1.013, fontsize=22)
plt.legend(loc='upper right', bbox_to_anchor=(1.28, 1))
#ax.legend_.set_bbox_to_anchor((0.05, 0.95))
#ax.legend_.set_loc(2)
```

Ilustración 82. Código para normalizar los datos y mostrarlo en un gráfica de distribución, elaboración propia 2021.

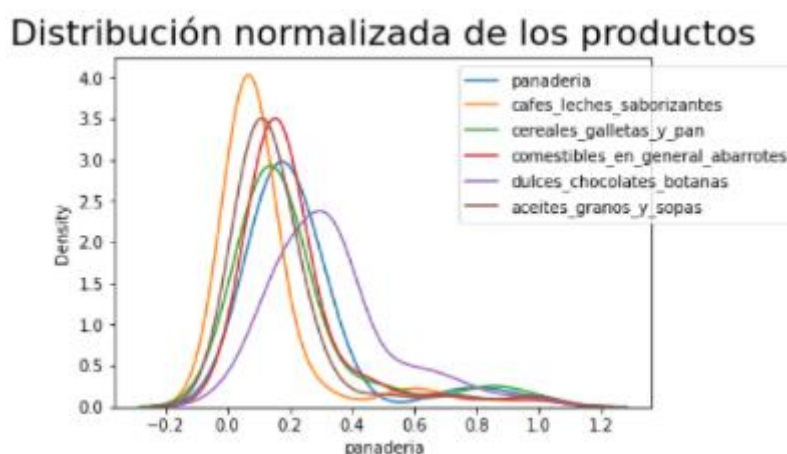


Ilustración 83. Resultado de la distribución normal de cada producto, elaboración propia 2021.

Una vez normalizada la base se elaboraron las correlaciones con el método “Pearson”, este usa un coeficiente con un rango de -1 a +1, en donde 0 indica que no existe relación.

Después de tener la correlación de cada categoría se despliega en un mapa de calor con la función de seaborn .heatmap().

El mapa de calor mostro la existencia correlaciones positivas entre las variables siendo las más importantes aquellas que presentan un color cercano al rojo claro cómo se puede apreciar en la ilustración 85, aunque esto no implica causalidad. Para demostrar dicha importancia y su causalidad, se optó por conseguir la significancia estadística de estas correlaciones por medio de un valor de $P < 0.05$ como se puede apreciar en el código de la ilustración 86, 87 y 89 de esta forma se conseguiría aquellos valores con alta correlación que sean altamente significativos.

Para listar el top 5 de las correlaciones más altas su uso el código de la ilustración 90. Y con base lo mostrado la dependencia más alta de productos son:

1. Comestibles en general y abarrotes con cereales galletas y pan: 0.846082
2. Cereales galletas y pan con cafés, leches y saborizantes: 0.816365
3. Comestibles en general y abarrotes con cafés leches y saborizantes: 0.790341
4. Aceites, granos y sopas con cafés leches y saborizantes: 0.789252
5. Cereales galletas y pan con aceites, granos y sopas: 0.743070

```
corr_df = df_normalizada.corr(method='pearson')

plt.figure(figsize=(20, 10))
sns.heatmap(corr_df, annot=True)
plt.show()
```

Ilustración 84. Código para crear las correlaciones y su mapa de calor, elaboración propia 2021.

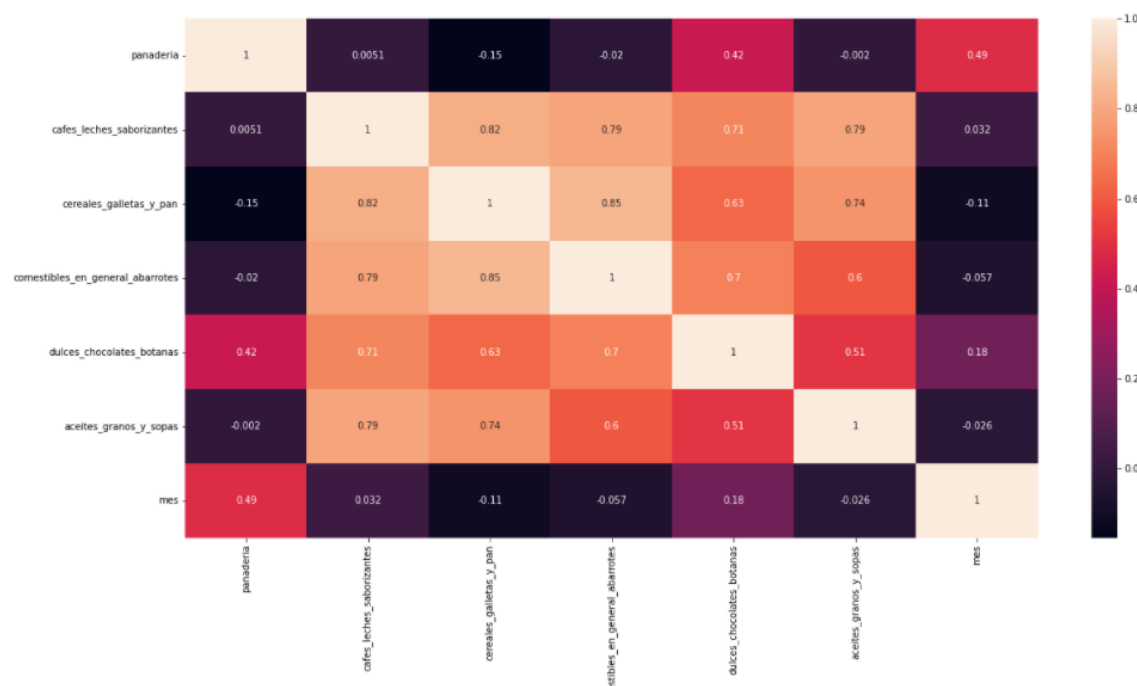


Ilustración 85. Mapa de calor de las correlaciones de cada categoría de productos, elaboración propia 2021

```
[ ] def corr_sig(df=None):
    p_matrix = np.zeros(shape=(df.shape[1],df.shape[1]))
    for col in df_normalizada.columns:
        for col2 in df_normalizada.drop(col,axis=1).columns:
            _, p = stats.pearsonr(df_normalizada[col],df_normalizada[col2])
            p_matrix[df_normalizada.columns.to_list().index(col),df_normalizada.columns.to_list().index(col2)] = p
    return p_matrix

p_values = corr_sig(df_normalizada)
mask = np.invert(np.tril(p_values<0.05))

[ ] plt.figure(figsize=(20, 10))
sns.heatmap(p_values, annot=True)

plt.savefig('Mapa de calor de las correlaciones unidades de productos respecto a su significancia estadística .png')
files.download("Mapa de calor de las correlaciones unidades de productos respecto a su significancia estadística.png")
plt.show()
```

Ilustración 86. Código para crear un mapa de calor de los p-vals de las correlaciones de las unidades de productos más vendidos

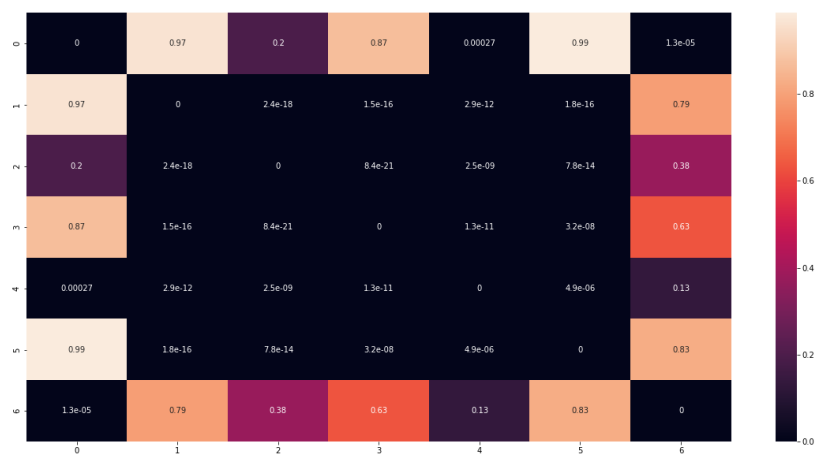


Ilustración 87. Mapa de calor de los p-vals de las correlaciones de las unidades de productos más vendidos

```
plt.figure(figsize=(20, 10))
sns.heatmap(corr_df, mask=mask,annot=True)

plt.savefig('Mapa de calor de las correlaciones unidades de productos respecto con la significancia estadística más importante.png')
files.download("Mapa de calor de las correlaciones unidades de productos respecto con la significancia estadística más importante.png")
plt.show()
```

Ilustración 88. Código para insertar una máscara de los p-values en los mapas de calor de las correlaciones de las unidades de productos más vendidos por categoría acorde a su significancia estadística

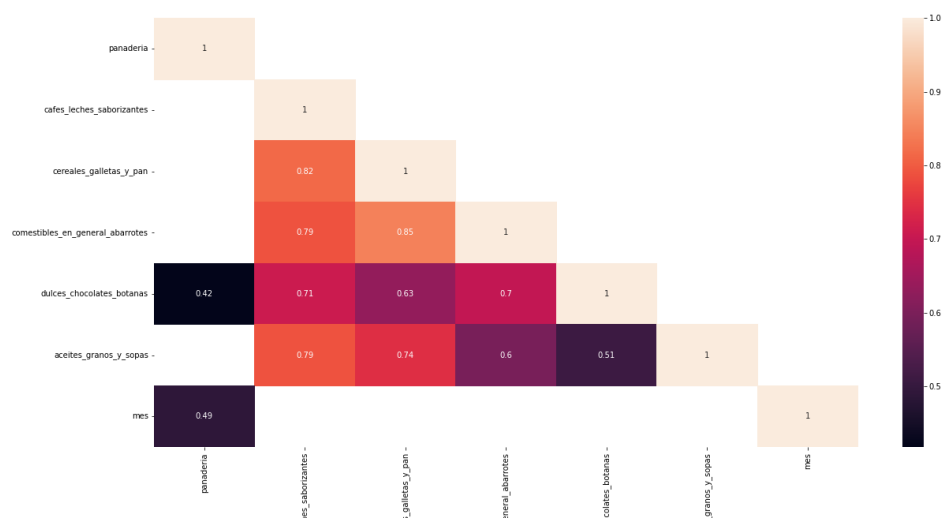


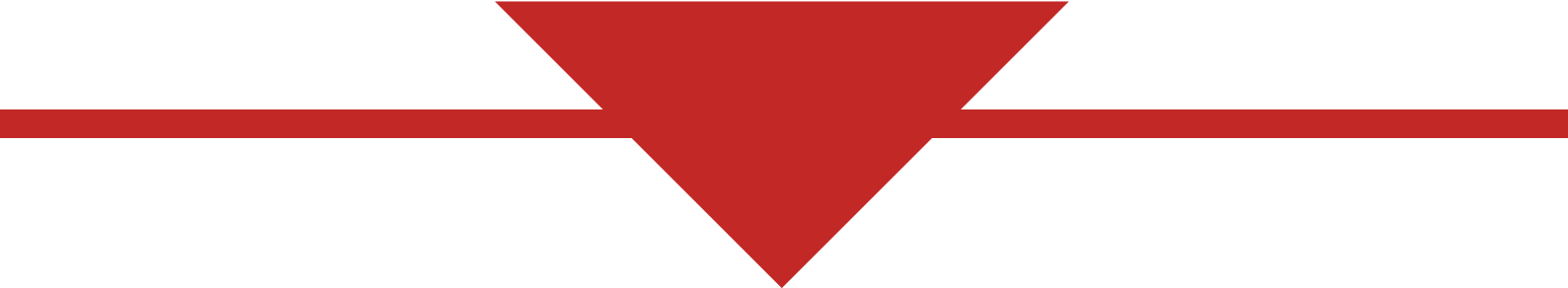
Ilustración 89. Resultado de aplicar la máscara de p-valúes a las correlaciones de cada producto vendido con base a la significancia estadística más importante.

```
[53] correlacion_ordenada = corr_df.unstack().sort_values()
      print("Top 3 correlaciones:\n\n", correlacion_ordenada[-17:-7])
```

Top 5 correlaciones:

1. comestibles_en_general_abarrotes cereales_galletas_y_pan
0.846082
2. cereales_galletas_y_pan cafes_leches_saborizantes
0.816365
3. comestibles_en_general_abarrotes cafes_leches_saborizantes
0.790341
4. cafes_leches_saborizantes aceites_granos_y_sopas
0.789252
5. cereales_galletas_y_pan aceites_granos_y_sopas
0.743070

Ilustración 90. Código y resultado del top 5 de correlaciones, elaboración propia 2021



FASE 4:

Modelado

Los modelos de *machine learning* trabajan mejor con datos escalados en rangos de -1 a 1 o de 0 a 1 (Goedegebuure, 2021), por lo que necesitamos primero verificar la correcta transformación de los datos con la función `describe()` como se puede corroborar en la ilustración 91.

```
df_normalizada.describe()
```

	panaderia	cafes_leches_saborizantes	cereales_galletas_y_pan	comestibles_en_general_abarrotes	dulces_chocolates_botanas	aceites_granos_y_sopas	mes
count	72.000000	72.000000	72.000000	72.000000	72.000000	72.000000	72.000000
mean	0.235015	0.129827	0.219728	0.208246	0.332146	0.175597	0.500000
std	0.198805	0.188424	0.221617	0.175422	0.197099	0.192974	0.316025
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.134484	0.048413	0.096971	0.114176	0.191924	0.077434	0.250000
50%	0.180639	0.073359	0.156751	0.164994	0.313230	0.118683	0.500000
75%	0.268166	0.099509	0.207525	0.219037	0.384794	0.192049	0.750000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Ilustración 91. Comprobación de la normalización de datos, elaboración propia 2021.

Efectivamente se consiguió la normalización de los datos al tener valores mínimos y máximos de 0 y 1 respectivamente en cada una de las categorías. Ahora solo queda crear un vector llamado “Rango_de_clusters” que contenga los diferentes números *clusters* con los que se evaluará y verificará el valor óptimo en el algoritmo de *k-means*.

```
Rango_de_clusters =[2,3,4,5,6,7]
```

Ilustración 92. lista con el rango de clústeres probados, elaboración propia 2021.

Una vez realizados los preparativos podemos empezar con el despliegue del algoritmo a través del método *elbow* o codo en español; este nos permite generar un desplegado de los distintos números de clústeres en un rango del 1 al 11 junto con su gráfico como se puede observar en la ilustración 93.

Por medio de esta visualización, se busca observar en la gráfica el punto óptimo donde la cantidad de centroides refleje un cambio abrupto en la caída de los valores de *wcss*, ya sea por un crecimiento atípico en la bajada de la curva o por un decrecimiento plano, a lo que se traduce a una sobre clasificación innecesaria. De esta forma eligiendo correctamente la cantidad de centroides se consigue tener una buena agrupación de los miembros, reducción del *overfitting* y una asignación balanceada con respecto a las características dentro de los grupos.

Cabe destacar que, aunque el gráfico pueda mostrar el valor idóneo para la asignación del número de centroides para el algoritmo de *k-means*, se utilizó la función *KneeLocator* como se puede ver en la ilustración 93 para conocer con certeza el valor óptimo para nuestros datos.

El valor resultante de la técnica del codo nos indica que el mejor número de agrupaciones para nuestros datos es de 4 clústeres, sin embargo, existe ocasiones en la que la distribución de los datos presenta un patrón singular que no puede ser correctamente clasificado por el algoritmo de *k-means*. Por lo tanto, es necesario representar estos datos en un gráfico de dispersión para tomar la decisión de si es necesario cambiar el tipo de algoritmo para la agrupación.

Los datos multivariados de nuestra base son complicados de proyectar en una gráfica de dispersión, por lo cual recurrimos al algoritmo de *PCA* (*Principal Component Analysis*) para poder observar sus patrones y tendencias. Este método permite reducir la longitud de los datos en componentes más simples por medio *eigenvectors* y *eigenvalues* (Goedegebuure, 2021), además se conserva la información de sus dimensiones al trabajar con la matriz de covarianzas

resultante de combinar linealmente las varianzas de todas las dimensiones, en consecuencia, se generó 2 componentes principales útiles para representarlos en una gráfica de dispersión que respeta y mantiene la dimensionalidad de la base original.

```

wcss = []
for i in range(1,11):
    model = KMeans(n_clusters = i, init = "k-means++")
    model.fit(df_normalizada)
    wcss.append(model.inertia_)
plt.figure(figsize=(10,10))
plt.plot(range(1,11), wcss)
plt.xlabel('Numero de clústeres ')
plt.ylabel('WCSS')
plt.show()

```

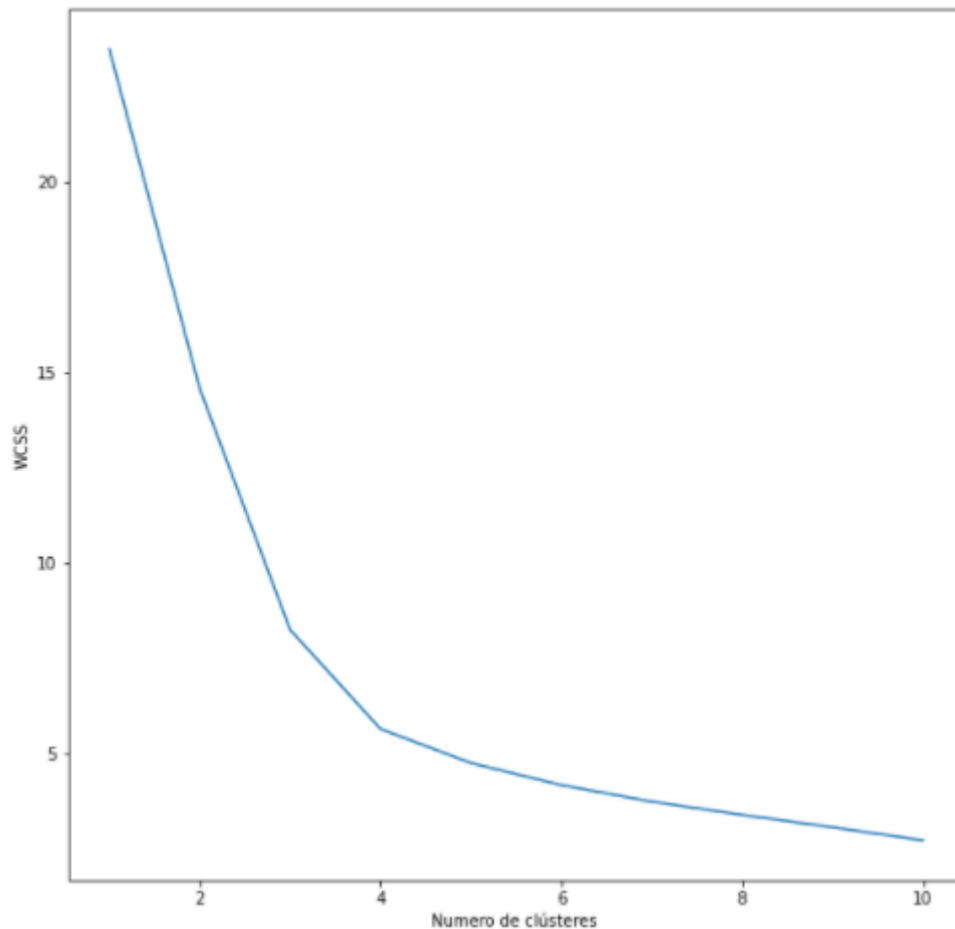


Ilustración 93. Gráfico del código para determinar el número óptimo de clústeres, elaboración propia 2021.

```

k1 = KneLocator(
    range(1, 11), wcss, curve="convex", direction="decreasing"
)

print('Cluster optimo con tecnica del codo:', k1.elbow)

```

Cluster optimo con tecnica del codo: 4

Ilustración 94. Código para obtener e imprimir el valor óptimo de clúster, elaboración propia 2021.

```

from sklearn.decomposition import PCA

pca = PCA(n_components= 2)
pca_df = pca.fit_transform(df_normalizada)
pca_df_2dimension = pd.DataFrame(data= pca_df, columns=['Componente_1', 'Componente_2'])

```

Ilustración 95. Código para descomponer el dataframe con PCA, elaboración propia 2021.

La forma en que PCA mantiene la complejidad de la base es primero utilizando variables normalizadas, de ahí parte del concepto previamente mencionado de *eigenvectors* y *eigenvalues* que vendría siendo la multiplicación de una matriz por un vector en donde su resultado presenta las siguientes propiedades (Rodrigo, 2021):

- Los *eigenvectors* solo existen para matrices cuadradas y no para todas. En el caso de que una matriz $n \times n$ tenga *eigenvectors*, el número de ellos es n .
- Al escalar un *eigenvector* antes de multiplicarlo por la matriz, se obtiene un múltiplo del mismo *eigenvector*. Puesto que, si se escala un vector multiplicándolo por cierta cantidad, solamente se consigue cambiar su longitud, pero la dirección es la misma.
- Todos los *eigenvectors* de una matriz son perpendiculares (ortogonales) entre ellos, independientemente de las dimensiones que tengan.

Lo anteriormente mencionado por Rodrigo, nos indica que solo se cambia su longitud, pero no las propiedades intrínsecas de los *eigenvectors* puesto que solo se está cambiando la dirección del plano.

El proceso interno que siguió el algoritmo con nuestros datos fue:

- Una resta de cada valor con la media de la variable a la que pertenece. Consiguiendo una media cero en todas las variables.
- Se maximiza la varianza encontrando la matriz de covarianzas con datos escalados para así evitar los crecimientos exponenciales de valores muy altos.

La varianza total que se calculó en los datos está representada como:

$$\sum_{j=1}^p Var(X_j) = \sum_{j=1}^p \frac{1}{n} \sum_{i=1}^n x_{ij}^2$$

Ecuación 1. Varianza total presente en el set de datos, Amat 2021

y la varianza explicada por la componente m (el nuevo componente principal creado) es:

$$\frac{1}{n} \sum_{i=1}^n z_{im}^2 = \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^p \phi_{jm} x_{ij} \right)^2$$

Ecuación 2. Varianza explicada por la componente m , Amat 2021

El resultado de la proporción de varianza acumulada por el componente m con base al *ratio* es de:

$$\frac{\sum_{i=1}^n \left(\sum_{j=1}^p \phi_{jm} x_{ij} \right)^2}{\sum_{j=1}^p \sum_{i=1}^n x_{ij}^2}$$

Ecuación 3. Proporción de varianza explicada por la componente m dada por el ratio, Amat 2021

Como en nuestro caso solo nos interesa generar 2 componentes principales para ser usados en un *scatter plot*, lo que se hizo fue definir a 2 el número de componentes principales de la función *PCA* de la ilustración 95. Al mantener toda la información pese a la reducción de la dimensión, es posible realizar un proceso reversible guardando previamente los *eigenvectors* o componentes principales para mostrar a que grupo pertenece la base de datos originales como se explica en la siguiente formula.

datos originales = (eigenvectors⁻¹) x (principal component scores) + medias originales

La manera en que el algoritmo de *PCA* nos puede decir con cuanta de la información se está manteniendo el estado original, mediante el valor de variance ratio como se observa en la ilustración 96. En donde podemos constatar que se está reteniendo el 80.14 de la varianza observada.

```
print(pca.explained_variance_ratio_)
print(pca.explained_variance_ratio_.cumsum())

[0.44906314 0.3524293 ]
[0.44906314 0.80149244]
```

Ilustración 96. Valor del ratio de la varianza y valor acumulativo del ratio de varianza, elaboración propia 2021.

Los valores de la ilustración 97 corresponden a los *eigenvectors* mientras que los valores de la ilustración 98 corresponden a los *eigenvalues*, así que ambos son los valores necesarios utilizados para realizar el proceso reversible.

```
print(pca.singular_values_)

[3.24798365 2.87737281]
```

Ilustración 97. Valor del singular value o eigenvalues, elaboración propia 2021.



FASE 5:

Evaluación

Una vez que se realizó la optimización del algoritmo y la “simplificación” de los datos con *PCA*, se decidió proyectar y evaluar cada uno de los clústeres asignados en el vector “Rango_de_clusters” mediante el uso de un ciclo *for* como se puede ver en la ilustración 98 y 99. Esto permitió repetir 6 veces el código para desplegar las gráficas de dispersión con su respectiva asignación de grupos para cada número de clústeres en la lista.

Siendo más específico, el ciclo *for* arrojó como resultado la agrupación de los datos para 2,3,4,5,6 y 7 grupos representados en una visualización de *scatter plot* como se observar desde la ilustración 101 hasta la 106. En cada una de ellas podemos ver la tendencia en la que están dispuestos los datos, por lo que con base a estos gráficos se puede llegar a la conjetura de que están ampliamente dispersos los datos y no presentan una configuración atípica que impida su clasificación con el algoritmo de *K-means*.

Por otra parte, durante los ciclos también se realizó la evaluación de los datos que partió desde observar la asignación de los grupos hasta ver la colindancia entre los valores extremos de cada clúster. Esta valoración de la pertenencia de los datos con respecto a su grupo sin este interferir con otro grupo vecino durante la clasificación se logró por medio del análisis de la silueta o *silhouette*, a través de este método se midió la distancia promedio interna del clúster y el promedio de la distancia del clúster más cercano para cada muestra ("*sklearn.metrics.silhouette_score* — *scikit-learn* 0.24.2 documentation", 2021).

El algoritmo lo primero que realiza es obtener el coeficiente de silueta para una muestra que es $\frac{(b - a)}{\max(a, b)}$ en donde, *b* es la distancia entre una muestra y el clúster más cercano distinto del clúster que forma parte. Otra manera de explicarlo es como:

- $|C_i|$ es el número de datos perteneciente al clúster *i*.
- $a(i)$ representa a los datos los datos *i* en el clúster C_i , siendo esta la medida para saber qué tan bien fue asignado el dato *i* a su clúster.

$$a(i) = \frac{1}{|C_i| - 1} \sum_{i \in C_i, i \neq j} d(i, j)$$

Ecuación 4. Asignación del dato dentro del clúster, Kumar 2021.

- La silueta que engloba los datos pertenecientes al *cluster* y los delimita con respecto su vecino cercano se especifica:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ si } |C_i| > 1$$

Ecuación 5. Delimitación de la silueta respecto a su vecino cercano, Kumar 2021.

Y

$$S(i) = 0, \text{ si } |C_i| = 1$$

Ecuación 6. Parámetro que delimita al vecino, Kumar 2021.

A lo que se puede representar como:

$$S(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, & \text{si } a(i) < b(i) \\ 0, & \text{si } a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1, & \text{si } a(i) > b(i) \end{cases}$$

Ecuación 7. Análisis de la silueta con los parámetros que se incluyen dentro del cluster, Kumar 2021.

- La silueta puede abarcar un rango posible de $-1 \leq S(i) \leq 1$
- El valor ideal de la silueta debe ser $S(i) \leq 0$ para tener una correcta asignación de los datos con su respectivo clúster.

En nuestro caso práctico al generar el código podemos ver desde ilustración 100 hasta la 106 el valor de la silueta con su respectiva gráfica. Siendo lo más importante el caso de $k = 2$ y $k = 4$ de la ilustración 101 y 103 al ser los únicos que no tienen ningún valor menor a 0. Esto es importante debido a que aquellos gráficos en donde se pueda encontrar un valor menor a 0 están sujetos a tener valores mal asignados en sus clústeres. Por otra parte, el score expresa de manera cuantitativa cuán densos y bien separados están los grupos, hecho que podemos ver con el grosor de “clúster labels” del eje Y.

Con respecto al tamaño y la densidad de los clústeres se busca un puntaje con valor promedio entre 0.5 y 1, ya que supera el umbral del 50% implicando que los datos están altamente separados y presentan buena densidad (Kumar, 2021); en nuestro caso el número de clústeres capaz de cumplir con esta condición fue el de $k = 4$ al tener un puntaje de 0.5748.

```

for n_clusters in Rango_de_clusters:
    # Creación de un subplot con 1 fila y 2 columnas
    fig, (ax1, ax2) = plt.subplots(1, 2)
    fig.set_size_inches(18, 7)

    # El 1er subplot es la grafica de silhouette
    # El coeficiente de silhouette usa un rango de -1 a 1 pero en este ejemplo
    # el rango sera de [-0.1, 1]
    ax1.set_xlim([-0.1, 1])
    # El (n_clusters+1)*10 es para insertar un espacio en blanco
    # entre las graficas silhouette de cada cluster para remarcarlos claramente.

    ax1.set_ylim([0, len(pca_df_2dimension) + (n_clusters + 1) * 10])

    # Inicializamos el cluster con el valor de n_clusters y un generador random
    # La semilla del generador fue 10 para la reproductibilidad.
    clusterer = KMeans(n_clusters=n_clusters, random_state=10)
    cluster_labels = clusterer.fit_predict(df_normalizada)

    # El puntaje de la silueta o silhouette_score da el valor promedio de todas las muestras
    # Esto representa la densidad y separación formada por los clusters

    silhouette_avg = silhouette_score(pca_df_2dimension, cluster_labels)
    print("Para cada n_clústeres =", n_clusters,
          "El valor promedio de la silueta es :", silhouette_avg)

    # Computa el puntaje de la silhouette para cada muestra
    sample_silhouette_values = silhouette_samples(pca_df_2dimension, cluster_labels)

```

Ilustración 98. Parte 1 del código para generar el análisis de la silueta y los clústeres del rango definido, elaboración propia 2021.

```

y_lower = 10
for i in range(n_clusters):
    # Agregamos el puntaje de la silhouette para la muestra perteneciente al
    # cluster i, y ordenarlos
    ith_cluster_silhouette_values = \
        sample_silhouette_values[cluster_labels == i]

    ith_cluster_silhouette_values.sort()

    size_cluster_i = ith_cluster_silhouette_values.shape[0]
    y_upper = y_lower + size_cluster_i

    color = cm.nipy_spectral(float(i) / n_clusters)
    ax1.fill_betweenx(np.arange(y_lower, y_upper),
                      0, ith_cluster_silhouette_values,
                      facecolor=color, edgecolor=color, alpha=0.7)

    # Etiqueta del grafico silhouette con su número de cluster en el medio
    ax1.text(-0.05, y_lower + 0.5 * size_cluster_i, str(i))

    # Computa el nuevo y_lower para la siguiente grafica
    y_lower = y_upper + 10 # 10 para las 0 muestras

ax1.set_title("Grafica de la silueta para varios clústeres .")
ax1.set_xlabel("Valor del coeficiente de la silueta ")
ax1.set_ylabel("Número del clúster")

# La linea vertical para el valor promedio de la silhouette de todos los valores
ax1.axvline(x=silhouette_avg, color="red", linestyle="--")

ax1.set_yticks([]) # Limpieza de la etiqueta yaxis / ticks
ax1.set_xticks([-0.1, 0, 0.2, 0.4, 0.6, 0.8, 1])

# 2nd Grafica muestra los clusteres actuales formadas
colors = cm.nipy_spectral(cluster_labels.astype(float) / n_clusters)
ax2.scatter(pca_df_2dimension.iloc[:, 0], pca_df_2dimension.iloc[:, 1], marker='.',
            s=400, lw=0, alpha=0.7,
            c=colors, edgecolor='k')

# Etiquetas de los clusters
centers = clusterer.cluster_centers_
# Dibuja circulos blancos en el centro de los cluster
ax2.scatter(centers[:, 0], centers[:, 1], marker='o',
            c="white", alpha=1, s=200, edgecolor='k')

for i, c in enumerate(centers):
    ax2.scatter(c[0], c[1], marker='%d$' % i, alpha=1,
                s=50, edgecolor='k')

ax2.set_title("Visualización de los datos agrupados .")
ax2.set_xlabel("Espacio entre las funciones para el 1º componente principal")
ax2.set_ylabel("Espacio entre las funciones para el 2º componente principal")

plt.suptitle(("Análisis de la silueta por agrupaciones de tipo K-Means de los datos "
              "with n_clusters = %d" % n_clusters),
             fontsize=14, fontweight='bold')

```

Ilustración 99. Parte 2 del código para generar el análisis de la silueta y los clústeres definido, elaboración propia 2021.

Para cada n_clústeres = 2 El valor promedio de la silueta es:
0.6099583301555795
Para cada n_clústeres = 3 El valor promedio de la silueta es:
0.5748323649791642
Para cada n_clústeres = 4 El valor promedio de la silueta es:
0.587998193683728
Para cada n_clústeres = 5 El valor promedio de la silueta es:
0.5018675132053909
Para cada n_clústeres = 6 El valor promedio de la silueta es:
0.33286653437195823
Para cada n_clústeres = 7 El valor promedio de la silueta es:
0.327832791756053

Ilustración 100. Resultado promedio del análisis de la silueta, elaboración propia 2021.

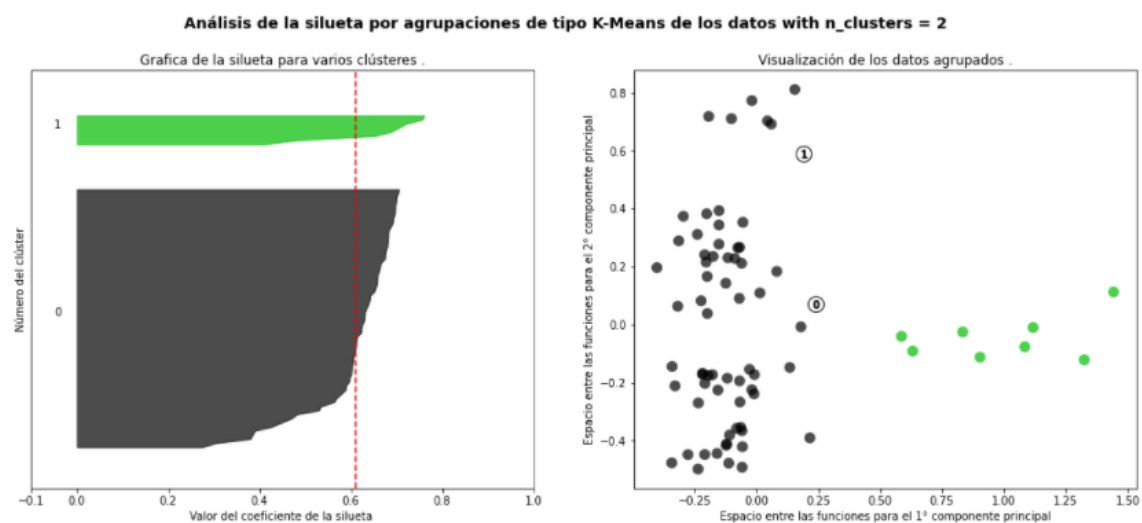


Ilustración 101. Análisis de la silueta para el número de clústeres 2, elaboración propia 2021

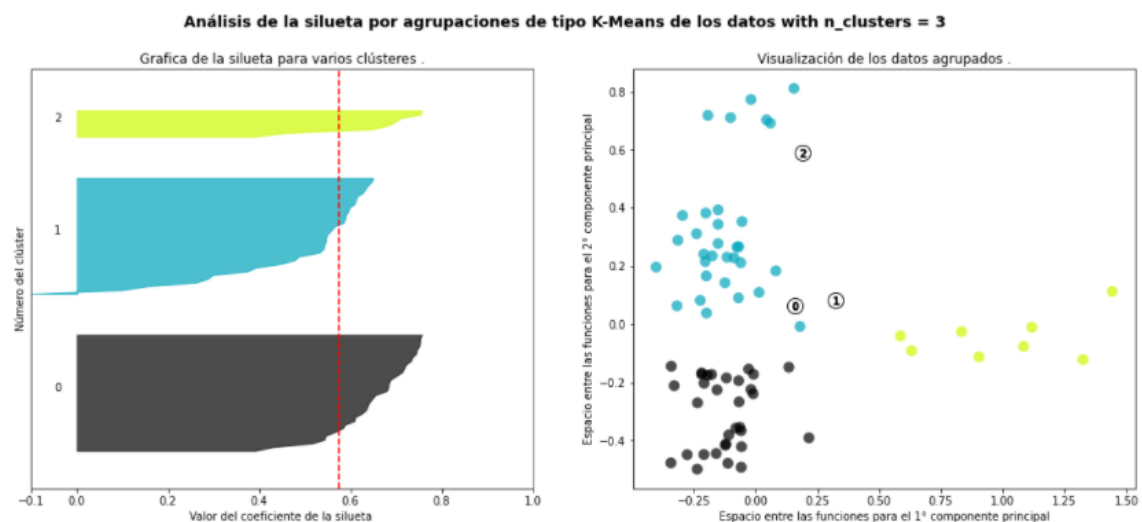


Ilustración 102. Análisis de la silueta para el número de clústeres 3, elaboración propia 2021.

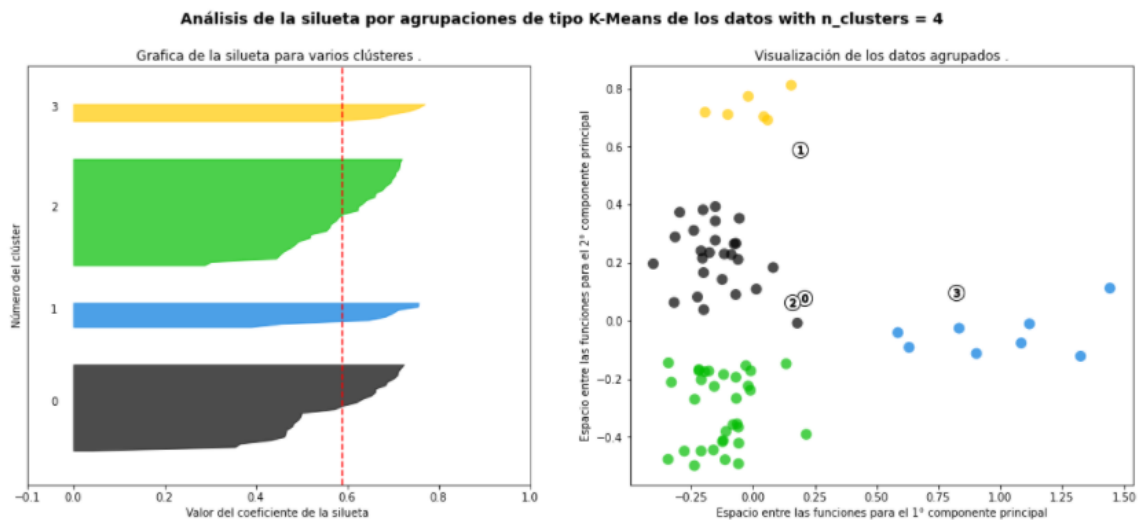


Ilustración 103. Análisis de la silueta para el número de clústeres 4, elaboración propia 2021.

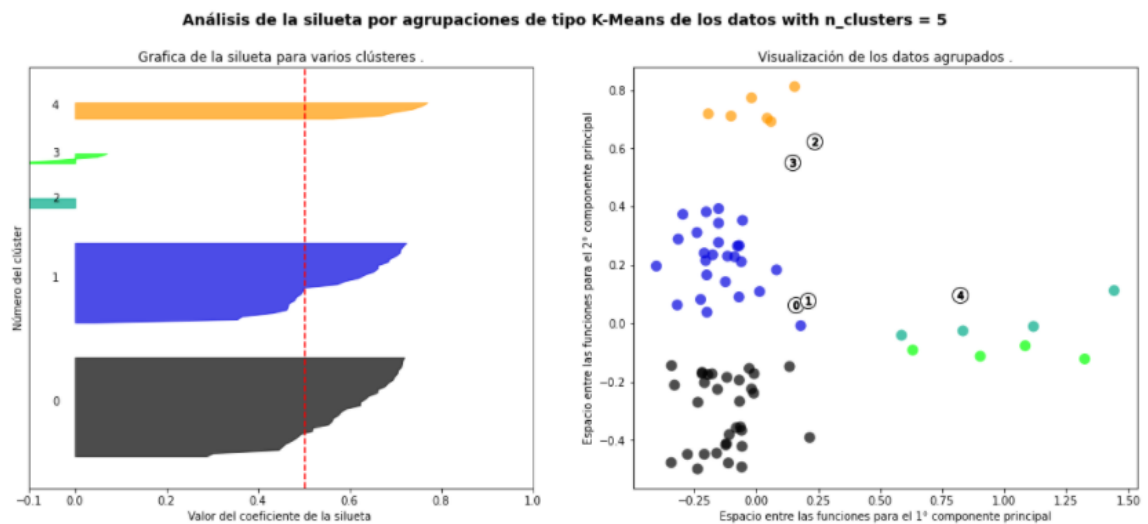


Ilustración 104. Análisis de la silueta para el número de clústeres 5, elaboración propia 2021.

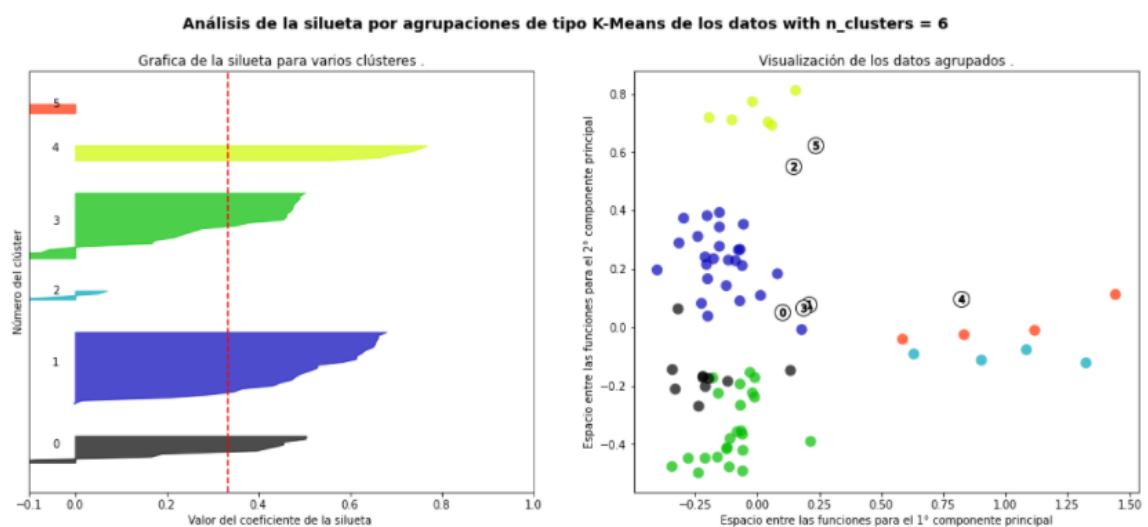


Ilustración 105. Análisis de la silueta para el número de clústeres 6, elaboración propia 2021.

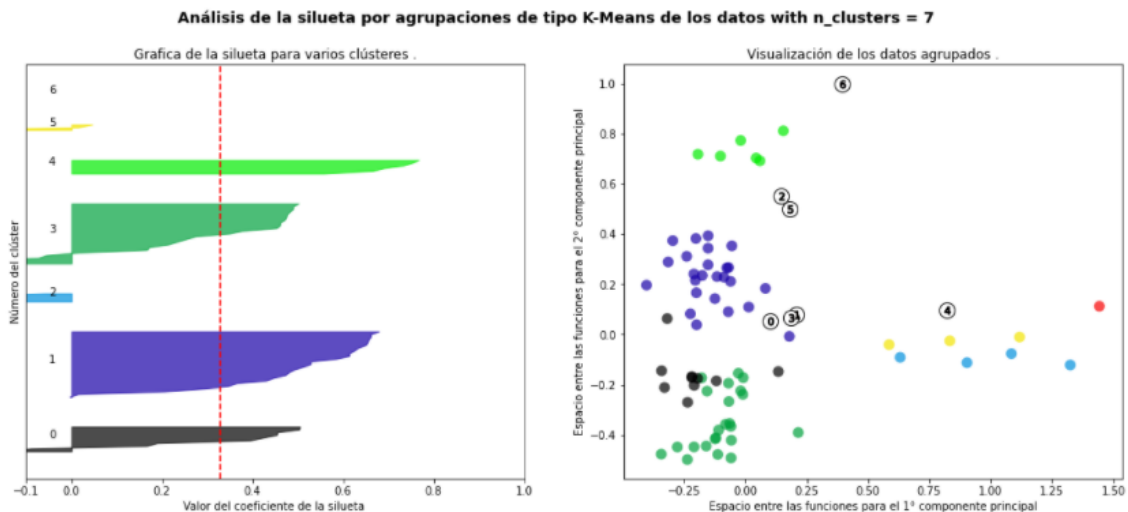


Ilustración 106. Análisis de la silueta para el número de clústeres 7, elaboración propia 2021.

El resultado del código anterior se muestra en las ilustraciones 101, 102, 103, 104, 105 y 106. En estas imágenes se observan 2 tipos de gráficas útiles que despliegan el resultado de implementar el algoritmo de *Kmeans* en un rango de 2 a 7 clústeres, en el primer gráfico de la izquierda se puede observar el número de clústeres, los cuales empiezan por el número 0 y concluyen en el número n-1, siendo n el número de clústeres elegidos; además de manera cualitativa se muestra el grosor del grupo formado por cada clúster y si este invade algún grupo vecino como lo es el caso de la ilustración 102, 104, 105 y 106 al tener valores negativos en el coeficiente de silueta. Por otra parte, el segundo gráfico despliega los datos simplificados con el algoritmo de *PCA* en una gráfica de dispersión para así obtener el resultado agrupación del algoritmo de aprendizaje no supervisada de *Kmeans*, en esta se aprecia de forma más granular la pertenencia de cada dato a su grupo y los formas naturales que estas presentan.

Una vez que se corroboró que el valor ideal para la signacion de nuestros clusteres es efectivamente de $k = 4$, se empezó con su despliegue individual del algoritmo en donde se realizo un total de 400 iteraciones a los datos normalizados guardados en la variable "df_normalizada", esto lo podemos observar en la ilustración 107.

```

▶ clustering = KMeans(n_clusters = 4, max_iter= 400)
  clustering.fit(df_normalizada)

```

```

KMeans(algorithm='auto', copy_x=True, init='k-means++',
max_iter=400,
      n_clusters=4, n_init=10, n_jobs=None,
precompute_distances='auto',
      random_state=None, tol=0.0001, verbose=0)

```

Ilustración 107. Código y resultado para la creación optima de 4 clústeres, elaboración propia 2021.

Al haber recreado exitosamente la agrupación de los 4 *clusters* se procedió a crear una nueva dimensión llamada "*KMeans_cluster*" que contiene el número del grupo al que pertenece, siendo posible encontrar valores de 0 al 4, este resultado se puede corroborar en la ilustración 107. Despues de juntar los datos se optó por graficarlos para observar como estan conformados los grupos por medio del codigo expuesto en la ilustración 108 y cuyo resultado se puede ver en la ilustración 109.

```
df['KMaens_cluster'] = clustering.labels_
pca_cluster_cada_df = pd.concat([pca_df_2dimension, df[['KMaens_cluster']]], axis= 1)

pca_cluster_cada_df
df
```

	panaderia	cafes_leches_saborizantes	cereales_galletas_y_pan	comestibles_en_general_abarrotes	dulces_chocolates_botanas	aceites_granos_y_sopas	mes	año	KMaens_cluster
0	44743	9847	9595	14680	8410	13982	1	2015	0
1	38178	9794	9867	15888	9617	11974	2	2015	0
2	44955	11208	10149	16868	9460	13881	3	2015	0
3	42125	8137	8081	14657	8917	12463	4	2015	0
4	30074	7933	7254	12841	7015	12023	5	2015	0
...	...	...	...	...	...	...	...	...	...
67	37517	7999	9426	10504	6119	13121	8	2020	3
68	42804	7182	7423	9725	5978	13539	9	2020	3
69	39077	7770	6581	10414	7980	14427	10	2020	3
70	42349	8656	7559	12291	7179	12629	11	2020	3
71	70582	9161	6462	12916	9300	13633	12	2020	2

Ilustración 108. Transformación del dataset con PCA, elaboración propia 2021

```
fig = plt.figure(figsize=(6,6))
ax = fig.add_subplot(1,1,1)
ax.set_xlabel('Componente 1', fontsize = 15)
ax.set_ylabel('Componente 2', fontsize = 15)
ax.set_title('Componentes principales', fontsize = 20)

color_theme = np.array(["blue","green","orange","gray"])
ax.scatter(x= pca_cluster_cada_df.Componente_1, y= pca_cluster_cada_df.Componente_2, c =color_theme[pca_cluster_cada_df.KMaens_cluster], s= 50 )
plt.show()
```

Ilustración 109. Código para observar en un gráfico de dispersión los clústeres creados, elaboración propia 2021

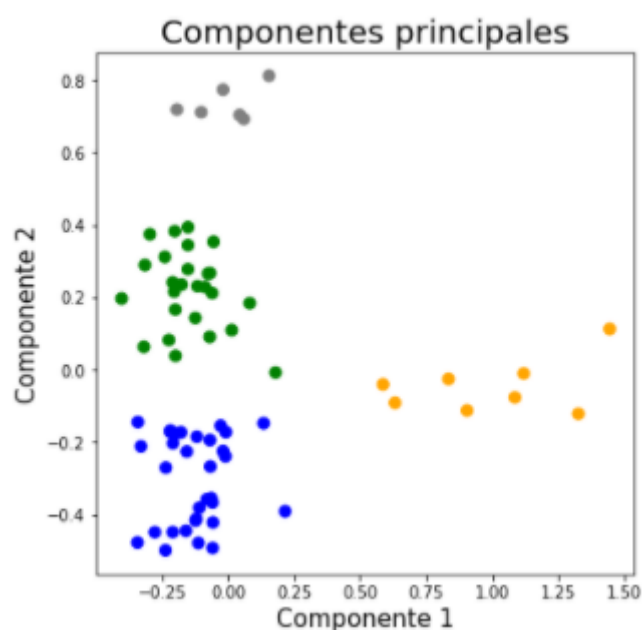


Ilustración 110. Gráfico de dispersión de los 4 clústeres generados con los componentes principales, elaboración propia.



FASE 6:

Implantación



LA DE TODOS LOS MEXICANOS

Autor
Ricardo Báez Mulia
ricardo.baezma@outlook.es

Retención de clientes identificando y segmentado al cliente de la región

Este es un análisis a partir de 3 fuentes de datos, se busca conocer el estado actual de la tienda y generar un modelo para segmentar a los clientes que consumen la canasta alimenticia básica .

Las fuentes de datos son: la interna de la empresa, INEGI 2015 y ENSANUT 2018 (Seguridad alimentaria)

Las categorías estudiadas de la base interna de la empresa que corresponden a la canasta básica son:

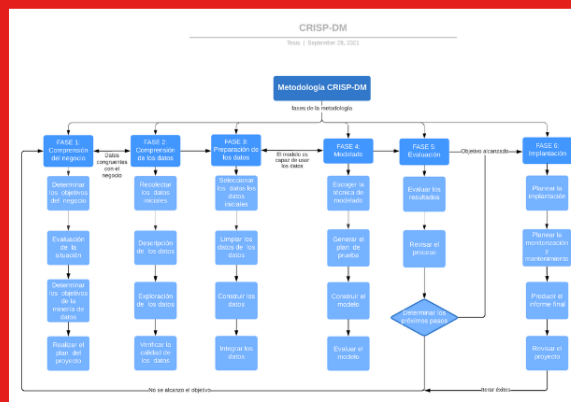


Metodología usada y formato de la infografía

Toda la investigación se llevó a cabo con la metodología CRIPS-DM, las cuales consta de 6 fases:



- Fase 1 de la metodología CRISP-DM
- Fase 2 de la metodología CRISP-DM
- Fase 3 de la metodología CRISP-DM
- Fase 4 de la metodología CRISP-DM
- Fase 5 de la metodología CRISP-DM
- Fase 6 de la metodología CRISP-DM



En esta infografía se plasma la fase 6, la cual consta de producir y difundir el informe final que comunica y aplica los resultados obtenidos a lo largo de las fases anteriores. Además, se divide en 4 segmentos:

- Resultados del análisis de los datos de la empresa
- Región
- Conocer a la competencia de los productos
- Buscando el número de tipos de clientes acorde a las compras realizadas en la tienda

Las primeras 3 fases hablan del pasado y presente la compañía (análisis descriptivo), en estas se utilizan los iconos de las categorías para saber de qué producto se habla e iconos de caras que indican si el punto es bueno, malo o interesante. Por otro lado, en el último segmento se muestra el análisis prescriptivo que resuelve la necesidad de conocer cuántos tipos de clientes existen para la empresa dentro de la región.



Resultados del análisis de los datos de la empresa

Con los datos de las 6 categorías de los últimos 5 años los puntos más destacados son:

La panadería y dulces, chocolates y botanas son las categorías que han mantenido un ritmo estable con respecto a los 5 años transcurridos.



Cafes, leches y saborizantes; comestibles en general y abarrotes; aceites, granos y sopas; tuvieron una caída de producto desplazado en el 2019 y 2020, posiblemente por a la crisis mundial del Covid-19.



Dulces, chocolates y botanas son una categoría estable, pese a no presentar un patrón que corresponda a los picos de las temporadas como el resto de las categorías.



Las 3 categorías que en rango promedio no desplazan tantas unidades son "cafes, leches y saborizantes"; "cereales, galletas y pan"; "dulces, chocolates y botanas".



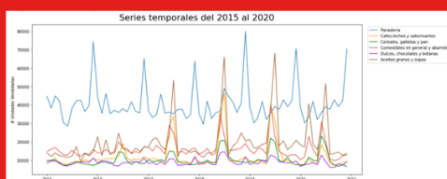
Existen ventas atípicas que representan a los picos de temporada como diciembre o verano



Los dulces, chocolates y botanas venden pocas unidades, pero de forma estable en el tiempo.



Todos los productos presentan picos en cada año y con frecuencia, lo cual indica que las ventas de temporada generan un movimiento fuerte de mercancía y este se mantiene año tras año.



Serie temporal de las unidades vendidas de todas las categorías vendidas desde el 2015 hasta 2020

	Cafes, leches y saborizantes; comestibles generales y abarroses; aceites, granos y sopas presentan un bajo desplazamiento de producto en los primeros 3 años comparado con el 2018, 2019 y 2020.
	Existe hábitos de consumo de la población reflejados en los picos de venta cíclicos de cada año.
	La panadería por lejos es el producto que mas unidades ha movido desde 2015 hasta 2020. También existe la excepción de los aceites, granos y sopas que desde el 2017 en cada mitad del año supera las unidades vendidas de pan.
	Excluyendo a la panaderia, las categorías que más venden son: "Aceites, granos y sopas", "Cafes, leches y saborizantes" y "Comestibles en general y abarroses".
	Desde el 2017 las ventas reflejan un patrón de todas las categorías menos de la panadería, lo cual puede indicarnos que hasta ese año fue posible suplir la demanda y los habitos de consumo del área.
 	Los dulces, chocolates y botanas son la categoría que menor volumen de ventas tienen en los últimos 5 años.
	Con los gráficos obtenidos podemos vislumbrar los picos de temporada alta, los cuales en promedio son marzo, junio, julio y diciembre.
 	La panaderia mueve la mayor cantidad de productos historicamente en diciembre. Además, cabe destacar que considerando al 2019 y 2020 como una epoca de ventas bajas por la pandemia los meses de abril, septiembre, octubre y noviembre no sufrierón una recaída drastica como se esperaba.

Cafes leches y saborizantes mueven la mayor cantidad de productos historicamente en junio y julio. En el resto de los meses mantuvo una venta regular pero baja de unidades con excepción de agosto 2019.



Cereales galletas y pan mueven su mayor cantidad de productos historicamente en junio y julio. Al igual que la panadería, si se considera los años de 2019 y 2020 como una epoca de ventas bajas los meses de abril y mayo no sufrieron de esta recaída, reflejando un buen manejo de la tienda. Además, en marzo en 2019 y octubre 2020 aumentaron sus ventas por unidad lo que demuestra una toma decisiones acertada.



Comestibles en general y abarrotes mueven su mayor cantidad de productos historicamente en junio y julio. El resto de los meses la venta fue regular a excepción de marzo que tuvo un incremento notable en sus ventas de 2019.



Dulces, chocolates y botanas mueven su mayor cantidad de productos historicamente en junio, julio y diciembre. En el 2019 y 2020 existio una baja de ventas en comparación a los años anteriores exceptuando los mese de mayo, abril y junio, pero por otro lado, las ventas más bajas de esos años e historicamente fueron en agosto, septiembre y diciembre.



Aceites, granos y sopas mueven la mayor cantidad de producto historicamente en marzo, junio y julio. Las ventas de marzo 2020 han sido la mejor de todos los años con respecto a ese mes, cabe destacar que esta fecha en esta categoria ha sido la que mejores resultados presenta desde el inicio de la pandemia.





La panadería interna de la tienda es el producto que más unidades mueve con rango normal de 49000 unidades aproximadamente y de 80000 en sus valores atípicos.



Dulces, chocolates y botanas" mueven con suma frecuencia cantidades de 7000 a 10000 unidades, además de presentar una mayor regularidad en este aspecto en comparación al resto de categorías.



La panadería mueven con suma frecuencia cantidades de 35000 a 45000 unidades.



Cafes leches y saborizantes" mueven con suma frecuencia cantidades de 7500 a 12500 unidades.



Cereales galletas y pan mueven con suma frecuencia cantidades de 3750 a 11000 unidades aproximadamente.



Comestibles en general y abarrotes, mueven con suma frecuencia cantidades de 10000 a 17500 unidades.



Excluyendo la panadería; la categoría aceites, granos y sopas es el producto que más unidades mueve en promedio y con rango de 13759 a 20500 unidades.



Los dulces son el producto con el registro histórico más bajo en unidades desplazadas, aunque también es el más consistente, ya que no presenta outliers.



Los cereales, galletas y pan (Bimbo) son los productos más desplazados proporcionalmente con respecto al inicio del 2015 hasta el 2020.

Aceites, granos y sopas, mueven con suma frecuencia cantidades de 10000 a 20000 unidades.



No existen valores atípicos debajo del rango lo que nos indica que respecto al desplazamiento no ha habido merma.



Mapa de calor de las correlaciones de cada categoría de productos

El mapa de calor mostro la existencia correlaciones positivas entre las variables siendo aquellas las que presentaban un color cercano al rojo claro, aunque esto no implica causalidad.

El top 5 de las correlaciones más altas y dependencia más alta de productos son:

- ✓ 1. Comestibles en general y abarrotes con cereales galletas y pan: 0.846082
- ✓ 2. Cereales galletas y pan con cafés, leches y saborizantes: 0.816365
- ✓ 3. Comestibles en general y abarrotes con cafés leches y saborizantes: 0.790341
- ✓ 4. Aceites, granos y sopas con cafés leches y saborizantes: 0.789252
- ✓ 5. Cereales galletas y pan con aceites, granos y sopas: 0.743070

1



2



3



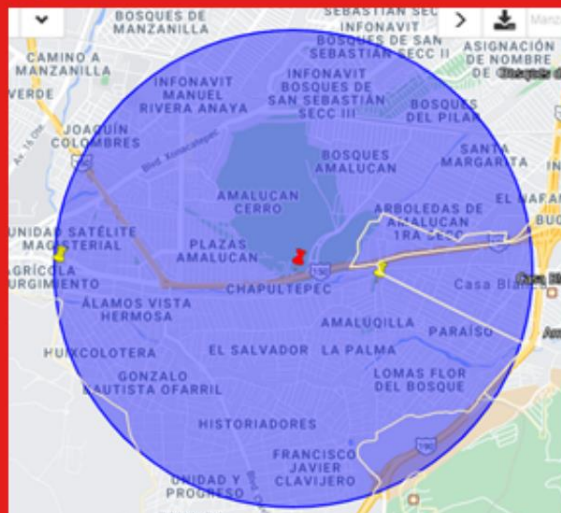
4



5



Región



Ubicación de los competidores, la tienda y su alcance en radio 2 Km a la redonda

Distribución de la edad de la población



En la región abundan familias jóvenes en donde se acumulan más miembros de 0 a 29 años.



Aproximadamente 1000 mujeres se preocupan de falta de alimentos, además de haber tenido poca, escasa o nula variedad de los alimentos.



De las personas mayores de 18 años no tienen escases de alimentos.



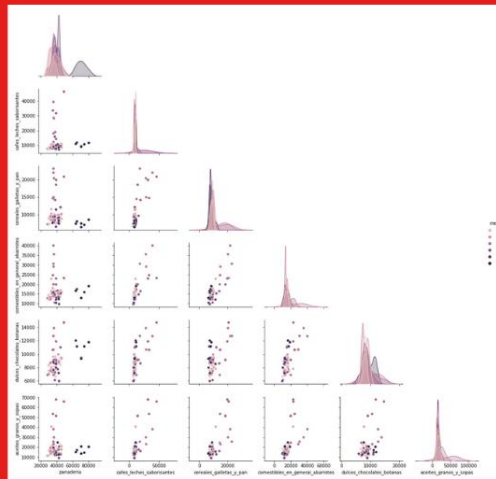
Conocer la competencia de los productos para ganar el corazón de los mexicanos

Los establecimientos económicos de la región nos indican cual es la demanda de productos de la zona

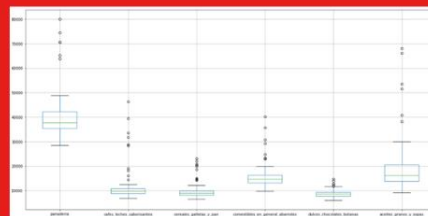
Top 10 del número de establecimientos económicos del área



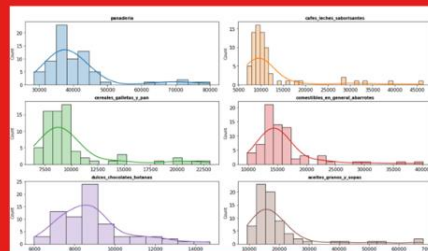
-  1. Comercio al por menor en tiendas de alimentos, ultramarinos y abarrotes
-  2. Salones y clínicas de belleza y peluquerías
-  3. Comercio al por menor de artículos de papelería
-  4. Restaurantes con servicio de preparación de antojitos
-  5. Restaurantes con servicio de preparación de tacos y tortas
-  6. Comercio al por menor de frutas y verduras frescas
-  7. Elaboración de tortillas de maíz y molienda de nixtamal
-  8. Comercio al por menor de otros alimentos
-  9. Servicios de preparación de otros alimentos para consumo inmediato
-  10. Panificación tradicional



Visualización de las correlaciones entre todas las categorías por mes en un gráfico de dispersión desde el 2015 hasta 2020

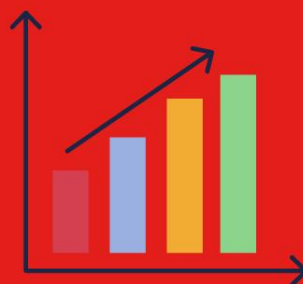


Diagramas de caja generadas a partir de las bases de datos internas de la empresa del 2015 al 2020



Histograma de las unidades vendidas de las 6 categorías

Nos va a interesar siempre tener una distribución normal o sesgada a la izquierda en nuestras categorías, ya que esto significaría que hay una frecuencia alta de productos vendidos por unidad. Por otra parte, mientras más sesgado a la izquierda significará que mayor cantidad producto se ha desplazado, esto siempre y cuando haya existido temporadas donde se haya vendido menos porque si no la distribución no presentaría sesgo.





Buscando el número de tipos de clientes acorde a las compras realizadas en la tienda

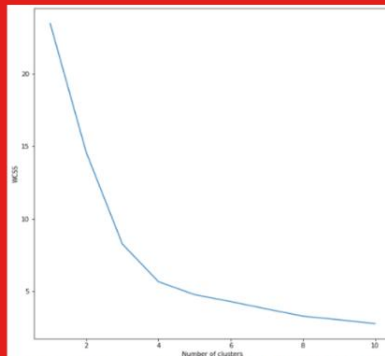
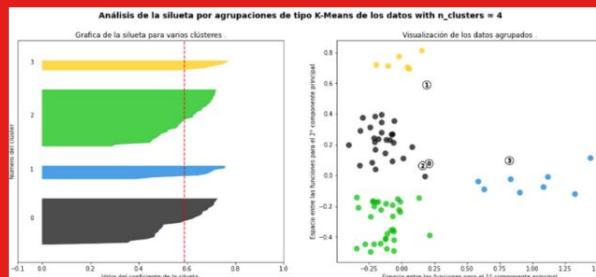


Grafico del codo para determinar el número óptimo de clústeres

Con el grafico del codo buscamos el número clústeres óptimo, en donde la cantidad de centroides refleje un cambio abrupto en la caída vs los valores de wcss, ya que al presentarse podemos saber que ese valor es el número de grupos correctos para nuestros datos sin que esté presente una sobre o falta clasificación.



Análisis de la silueta para el número de clústeres 4

El mejor número de agrupaciones para nuestros datos es de 4 clústeres, sin embargo, existe ocasiones en la que la distribución de los datos presenta un patrón singular que no puede ser correctamente clasificado por el algoritmo de k-means. Por lo tanto, es necesario representar estos datos en un gráfico de dispersión y en un análisis de la silueta para tomar la decisión de si es necesario cambiar el tipo de algoritmo para la agrupación.

Se puede observar que exitosamente la clasificación de las unidades vendidas en **4 grupos** perfectamente limitados en el gráfico de dispersión y silueta.





Ilustración 111. Resultados de la implantación en forma de informes

| Conclusión

En conclusión, se pudo realizar un proceso ETL a las 3 fuentes de datos de origen externo e interno; lo cual permitió realizar un análisis descriptivo de la población perteneciente a la región que a su vez ayudó a delimitar la edad, género y tipo de negocios que las persona tenían aledaños a la tienda. Este último aspecto sirvió para inferir los sectores económicos más importantes del supermercado y representarlos dentro de un dashboard elaborado en Power BI.

Finalmente, durante este trabajo se pudo desarrollar un informe completo que resume el estado pasado, presente y futuro de la tienda como parte de la fase final que establece la metodología CRISP-DM. También usando técnicas de *machine learning* como K-means y *PCA* se consiguió obtener distintas agrupaciones naturales formadas acordes al tipo de producto comprado a lo largo de los últimos 5 años, arrojando como resultado un total de 4 clústeres óptimos que representan posiblemente los tipos de clientes que visitan la tienda, este último apartado serviría como una posible línea de investigación al robustecer los 4 tipos de clientes con técnicas de investigación propias del “user persona”.

Bibliografía

1. Amat, J. (2021). Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE. Retrieved 6 November 2021, from https://www.cienciadedatos.net/documentos/35_principal_component_analysis
2. Atkin, D., Faber, B., & Gonzalez-Navarro, M. (2018). Retail Globalization and Household Welfare: Evidence from Mexico. *Journal Of Political Economy*, 126(1), 1-73. doi: 10.1086/695476
3. Brenes, E. R., Ciravegna, L., & Montoya, D. (2015). Super Selectos: Winning the war against multinational retail chains. *Journal of Business Research*, 68(2), 216–224. doi:10.1016/j.jbusres.2014.09.030
4. Calvillo, E. (2021). Efectos del Covid-19 en la Industria de Consumo | Consumo | Deloitte México. Retrieved 12 May 2021, from <https://www2.deloitte.com/mx/es/pages/consumer-business/articles/efectos-del-COVID19-en-la-industria-de-consumo.html>
5. CANASTA BÁSICA DE ALIMENTOS. (2021). Retrieved 12 May 2021, from <https://www.gob.mx/canastabasica>
6. Cantón Croda, R. (2021). Data Driven Roles [Ebook] (p. 18). HEROICA PUEBLA DE ZARAGOZA: UPAEP.
7. Cayuela, A. (2021). Nuevos hábitos de consumo impulsados por el Covid-19 y cómo se prepara tu negocio. Retrieved 12 May 2021, from <https://business.trustedshops.es/blog/nuevos-habitos-de-consumo/>
8. Chiu, K. S. Y., Luk, R. W. P., Chan, K. C. C., & Chung, K. F. L. (n.d.). Market-basket analysis with principal component analysis: an exploration. *IEEE International Conference on Systems, Man and Cybernetics*. doi:10.1109/icsmc.2002.1176057
9. CONEVAL. (2021). Evolución de la canasta alimentaria y del Índice la tendencia Laboral de la pobreza.; Retrieved 12 May 2021, from https://www.coneval.org.mx/Evaluacion/ECNCH/Documents/Resumen_ejecutivo_I_TLP_280715.pdf
10. Feng, X., Astell-Burt, T., Badland, H., Mavoa, S., & Giles-Corti, B. (2018). Modest ratios of fast food outlets to supermarkets and green grocers are associated with higher body mass index: Longitudinal analysis of a sample of 15,229 Australians aged 45 years and older in the Australian National Liveability Study. *Health & Place*, 49, 101–110. doi:10.1016/j.healthplace.2017.10.004
11. Gil Martínez, C. (2018). RPubs - Análisis de componentes principales (PCA). Retrieved 22 February 2021, from https://rpubs.com/Cristina_Gil/PCA
12. Gobierno del Estado de Puebla Secretaría de Gobernación Orden Jurídico Poblano. (2020). *Zonificación Catastral y las Tablas de Valores Unitarios de Suelos Urbanos y Rústicos; así como los Valores Catastrales de Construcción por metro cuadrado, para el Municipio de Puebla* [Ebook]. HEROICA PUEBLA DE ZARAGOZA. Retrieved from <https://ojp.puebla.gob.mx/index.php/zoo-items-landing/item/zonificacion-catastral-y-las-tablas-de-valores-unitarios-de-suelos-urbanos-y-rusticos-asi-como-los-valores-catastrales-de-construccion-por-metro-cuadrado-para-el-municipio-de-puebla-3>

13. Goedegebuure, R. (2021). Chapter 4 K-Means clustering | Market Segmentation & Clustering. Retrieved 27 July 2021, from https://bookdown.org/robert_statmind/mcms_03/k-means-clustering-1.html
14. González, A. (2021). Market Basket Analysis: Big Data y Machine Learning en Retail – Cleverdata. Retrieved 22 February 2021, from <https://cleverdata.io/basket-analysis-machine-learning/>
15. Griva, A., Bardaki, C., Pramataris, K., & Papakiriakopoulos, D. (2018). Retail business analytics: Customer visit segmentation using market basket data. *Expert Systems With Applications*, 100, 1-16. doi: 10.1016/j.eswa.2018.01.029
16. Kaur, M., & Kang, S. (2016). Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining. *Procedia Computer Science*, 85, 78–85. doi:10.1016/j.procs.2016.05.180
17. Kurniawan, F., Umayah, B., Hammad, J., Nugroho, S., & Hariadi, M. (2021). Market Basket Analysis to Identify Customer Behaviours by Way of Transaction Data. Retrieved 22 February 2021, from *La minería de datos: Definición, usos y funciones*. (2021). Retrieved 9 May 2021, from <https://www.tecnologias-informacion.com/mineria-de-datos.html>
18. Kumar, A. (2021). KMeans Silhouette Score With Python Examples - DZone AI. Retrieved 31 July 2021, from <https://dzone.com/articles/kmeans-silhouette-score-explained-with-python-exam>
19. Pérez-Villarreal, H. H., Martínez-Ruiz, M. P., & Izquierdo-Yusta, A. (2019). Testing Model of Purchase Intention for Fast Food in Mexico: How do Consumers React to Food Values, Positive Anticipated Emotions, Attitude toward the Brand, and Attitude toward Eating Hamburgers? *Foods*, 8(9), 369. doi:10.3390/foods8090369
20. Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". *Computational and Applied Mathematics* 20: 53-65.
21. Ramirez, A. (2021). Así ha sido la evolución histórica del retail en México | BELOW THE LINE, RETAIL | Revista InformaBTL |. Retrieved 22 February 2021, from <https://www.informabtl.com/asi-ha-sido-la-evolucion-historica-del-retail-en-mexico/>
22. Vela Meléndez, L., Chapoñan Coronado, A., Alexis I, G., Acha, D., Juárez, I., Loayza Mori, E., & Sales Campos, C. (2021). *TEORIA DE JUEGOS Y MODELOS DE OLIGOPOLIO* [Ebook]. Peru: UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO LAMBAYEQUE. Retrieved from <https://web.ua.es/es/giecryal/documentos/teoria-juegos.pdf>