

INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE OCCIDENTE

Departamento de Matemáticas y Física

Desarrollo tecnológico y generación de riqueza sustentable

PROYECTO DE APLICACIÓN PROFESIONAL (PAP)

Programa de Modelación Matemática para el desarrollo de planes y proyectos



**ITESO, Universidad
Jesuita de Guadalajara**

PAP4J09A PAP PROGRAMA DE MODELACION MATEMATICA PARA EL DESARROLLO DE
PLANES Y PROYECTOS DE NEGOCIO I

Segmentación de clientes para Herman Delikatessen

PRESENTAN

Programas educativos y Estudiantes

Espinosa García Lyha

Juárez Pineda Bryan Azhael

Ruíz Magaña Juan Pablo

Vazquez Vargas Ana Cristina

Profesor PAP: Benavides Herrera Pablo

Tlaquepaque, Jalisco, mayo de 2022

REPORTE PAP

Presentación Institucional de los Proyectos de Aplicación Profesional

Los Proyectos de Aplicación Profesional (PAP) son experiencias socio-profesionales de los alumnos que desde el currículo de su formación universitaria- enfrentan retos, resuelven problemas o innovan una necesidad sociotécnica del entorno, en vinculación (colaboración) (co-participación) con grupos, instituciones, organizaciones o comunidades, en escenarios reales donde comparten saberes.

El PAP, como espacio curricular de formación vinculada, ha logrado integrar el Servicio Social (acorde con las Orientaciones Fundamentales del ITESO), los requisitos de dar cuenta de los saberes y del saber aplicar los mismos al culminar la formación profesional (Opción Terminal), mediante la realización de proyectos profesionales de cara a las necesidades y retos del entorno (Aplicación Profesional).

El PAP es un proceso acotado en el tiempo en que los estudiantes, los beneficiarios externos y los profesores se asocian colaborativamente y en red, en un proyecto, e incursionan en un mundo social, como actores que enfrentan verdaderos problemas y desafíos traducibles en demandas pertinentes y socialmente relevantes. Frente a éstas transfieren experiencia de sus saberes profesionales y demuestran que saben hacer, innovar, co-crear o transformar en distintos campos sociales.

El PAP trata de sembrar en los estudiantes una disposición permanente de encargarse de la realidad con una actitud comprometida y ética frente a las disimetrías sociales. En otras palabras, se trata del reto de “saber y aprender a transformar”.

El Reporte PAP consta de tres componentes:

El primer componente refiere al ciclo participativo del PAP, en donde se documentan las diferentes fases del proyecto y las actividades que tuvieron lugar durante el desarrollo de este y la valoración de las incidencias en el entorno.

En caso de requerirse alguna adecuación al nombre de las fases propuestas para este componente, se puede realizar siempre y cuando sea complementario a lo ya establecido.

El segundo componente presenta los productos elaborados de acuerdo con su tipología.

El tercer componente es la reflexión crítica y ética de la experiencia, el reconocimiento de las competencias y los aprendizajes profesionales que el estudiante desarrolló en el transcurso de su labor.

Resumen

El propósito general del PAP *Segmentación de clientes para Herman Delikatessen* es como su nombre lo indica generar agrupamientos que permita reconocer patrones de compra entre sus clientes aplicando la ciencia de datos.

El objetivo general del proyecto es brindar a Herman Delikatessen un análisis operativo integral que apoye a la toma de decisiones sobre el negocio. Como objetivos específicos se tiene el generar un reporte exploratorio donde se entienda la analítica de negocio entorno a las ventas y compras de la tienda, proveer de un perfil de cliente que ayude a segmentar los tipos de compradores con los que cuenta la tienda y brindar una infraestructura que ayude a segmentar productos para la generación de potenciales promociones a futuro.

Para lograr cumplir con estos objetivos se realizó un análisis tanto estadístico como exploratorio de la información otorgada por la tienda en términos de compras y ventas. Posterior a esto se aplicaron algoritmos de clustering no supervisado (DBSCAN) para generar tipos de clientes a los cuales se le pueden dirigir ciertas promociones.

1. Ciclo participativo del Proyecto de Aplicación Profesional

El PAP es una experiencia de aprendizaje y de contribución social integrada por estudiantes, profesores, actores sociales y responsables de las organizaciones, que de manera colaborativa construir sus conocimientos para dar respuestas a problemáticas de un contexto específico y en un tiempo delimitado. Por tanto, la experiencia PAP supone un proceso en lógica de proyecto, así como de un estilo de trabajo participativo y recíproco entre los involucrados.

1.1. Entendimiento del ámbito y del contexto

El enfoque de este PAP gira entorno a Herman Delikatessen, una compañía dedicada a la distribución de productos gourmet de alta calidad. La empresa cuenta con más de treinta años en el mercado nacional y son distribuidores oficiales de Olio Fino: aceite y vinagre balsámico. Herman Delikatessen al ser una compañía con modelo de negocio por franquicias, nuestro caso de estudio se centrará en la sucursal con domicilio en: Av. Juan Palomar y Arias 123, Jardines Vallarta, 45020 Zapopan, Jal.

Dentro de la estructura de nuestra sucursal, la cantidad de productos que se ofertan en el inventario son bastantes, sin embargo, su giro principal se enfoca en la venta de carnes frías gourmet y productos derivados.

1.2. Caracterización de la organización o comunidad

Dentro de la sucursal se tienen a dos personas en el área de mostrador y armado de charolas de charcutería, la dueña del negocio está fuertemente involucrada con los inventarios, promociones a clientes frecuentes, y el funcionamiento del negocio en general. Siendo parte clave en el desarrollo y narrativa del negocio a lo largo de estos años.

Se cuenta también con una persona encargada de llevar las redes sociales de la empresa. La tienda al ser pequeña, no requiere una gran cantidad de personal, por lo que las actividades administrativas y técnicas de las ventas pueden realizarse por las dos trabajadoras actualmente contratadas.

El propósito de la tienda Herman Delikatessen sucursal Galerías es como se planteó en apartados anteriores la venta de carnes frías de primera calidad, así como otros productos gourmet que un amante de las carnes frías buscaría.

1.3. Identificación de la(s) problemáticas(s)

Al hablar con la dueña de la tienda, nos comentó que el método de promoción que tiene con sus clientes es mediante WhatsApp, y que si bien, le ha funcionado, no es la mejor manera de crear una recomendación para sus clientes, ya que se deja de lado el factor a la medida que puede atraer a los clientes a consumir nuevos productos.

Recientemente la compañía ha comenzado a capturar y registrar sus datos, no obstante, la mayoría de las ventas provienen de clientes no registrados. Es por la escasez de información que se propone encontrar una manera de crear clasificaciones que permitan recomendar a un cliente basándose en los productos que ya consume, un nuevo producto que complemente aquello ya adquirido. El clasificar estos productos de cierta manera ayuda a garantizar que el cliente disfrute de la recomendación y por lo tanto aumente su consumo en la tienda.

1.4. Planeación de alternativas(s)

Para resolver la problemática con la que se encuentra la empresa, se decide realizar un análisis exploratorio de las bases de datos proporcionadas para conocer a fondo el funcionamiento de la empresa, es decir, identificar el comportamiento de principalmente sus ventas y clientes, acompañado también de un análisis en términos de compras.

Una vez realizado este análisis, se buscará encontrar un perfil del cliente haciendo uso de la ciencia de datos. Por este medio, se pretende encontrar aquellas características que permitan generar grupos de productos ofertados por la empresa con los cuales se podrán generar recomendaciones puntuales al cliente sobre qué productos pueden ser de su agrado, generando no solo mayores ventas para la tienda sino aumentando la propuesta de valor de la tienda en términos de reconocimiento de clientela.

1.5.Desarrollo de la propuesta de mejora

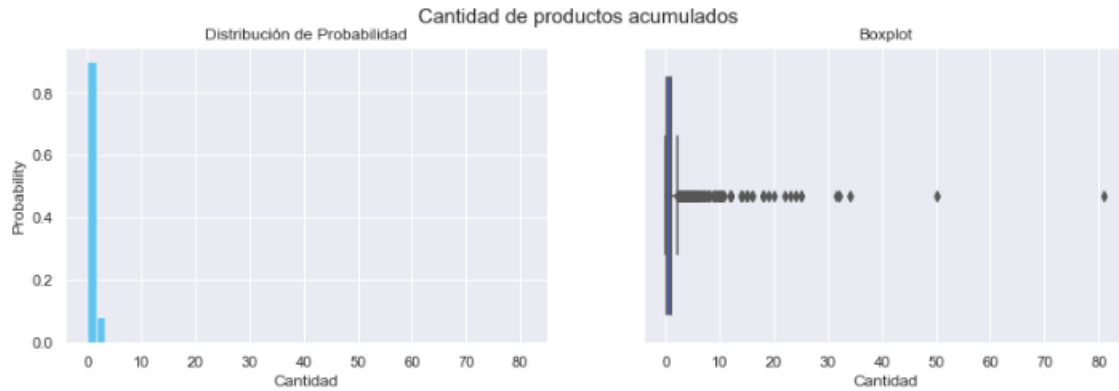
En este apartado se muestran los procesos que se trabajaron a lo largo de este proyecto dividido en 4 etapas:

- Análisis Estadístico
- Análisis de Compras
- Análisis de Ventas
- Clustering como herramienta para reconocer patrones de compra

En cada sección se describen tanto metodologías como resultados.

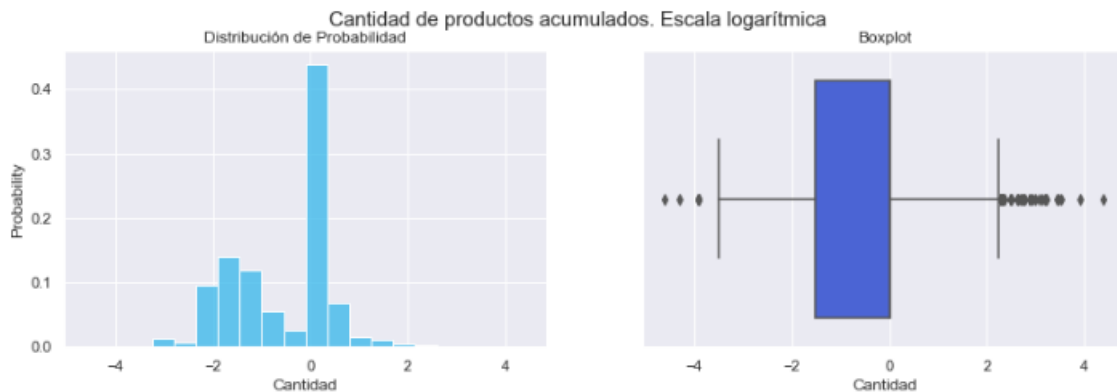
1.5.1. Estadística descriptiva

Con la intención de tener un panorama más completo de los datos otorgados por la tienda, se procede a realizar un análisis estadístico. Al observar la distribución correspondiente podemos darnos cuenta de que existen valores atípicos en la cantidad de productos adquiridos de manera acumulada en cada folio por parte de los clientes de Herman Delikatessen. Esto marca un foco de análisis interesante en la segmentación de los propios consumidores de la tienda.



Gráfica 1. Distribución de Probabilidad y Boxplot de la cantidad de productos vendidos

Con la intención de observar de una manera más escalada la distribución de la cantidad de productos adquiridos, realizaremos una transformación logarítmica a los datos y posterior a ello delimitaremos aquellos límites donde se encuentren concentrados la mayoría de los datos.



Gráfica 2. Distribución de Probabilidad y Boxplot de la cantidad de productos vendidos en escala logarítmica

Para los productos que se venden por pieza, la mayoría de estos se encuentran en el rango (1, 7) representando aproximadamente más del 60% de las observaciones en la muestra. Con esto en cuenta se infiere que aquellos clientes que hayan adquirido más de 7 productos por folio en Herman Delikatessen, se pueden clasificar como un tipo cliente preferencial, ya que ha demostrado que su patrón de compras (al menos en lo acumulado en cada ticket) es superior a la de la mayoría de los clientes. Incluso a este punto podríamos llegar a considerar si alguno de ellos revende productos propios de Herman Delikatessen.

En cuanto a los productos que se ofertan por gramaje, se presenta que la mayoría de estos siendo carnes frías y derivados, se encuentran en el siguiente rango (0.0498, 0.9991). Es decir que encontrar clientes que consuman menos de 4.98 gramos o más de 1,000 gramos de estos productos es sumamente raro, siendo el gramaje más popular para estos de 200 a 300 gramos. Con dichos puntos en consideración se sugieren las siguientes alternativas para guiar a un mejor procesamiento y un propio análisis mejor enfocado de las ventas de la empresa.

1. Clasificar a aquellos clientes que compren productos en escalas anormales a las descritas por la mayoría de los clientes.
2. Separar y redefinir las dos modalidades principales de venta de productos en Herman Delikatessen, que sería la modalidad por pieza y la modalidad por gramaje.

La cantidad de productos vendidos bajo la modalidad por piezas es de 12663 unidades mientras que la cantidad de productos vendidos bajo la modalidad por gramaje es de 13817 productos.

Al observar la muestra correspondiente a cada esquema de productos ofertados, se observa que se cuenta con una cantidad de datos lo suficientemente grande para poder realizar un análisis particular para cada uno de ellos. Se redefine el conjunto de datos en dos nuevos subconjuntos separados por su esquema de ventas, esto ayudará a realizar un análisis subsecuente más completo y enfocado a cada esquema de productos que puede aportar mayor información.

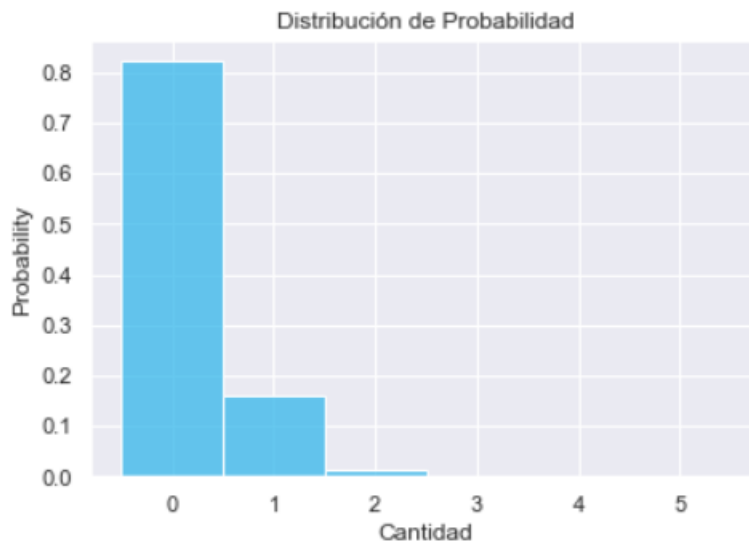
ipo	Cantidad	Producto	Clave_producto	Folio	Cliente	...	Total	Efectivo	Cheque	Crédito	Transferencia	Vales	Tarjeta	Anticipo	SICAR Pagos	Tipo producto
ZA	1.0	KOMBUCHITA JENGIBRE	[7500464022219]	219	Público en General	...	39.0	39.0	0	0.0	0.0	0	0.0	0	0	Venta unitaria
ZA	1.0	KOMBUCHITA JAMAICA	[7500464022226]	220	Público en General	...	39.0	0.0	0	0.0	0.0	0	39.0	0	0	Venta unitaria
ZA	1.0	MAIZ INFLADO	[GALBIS52766]	221	Público en General	...	30.0	0.0	0	0.0	0.0	0	30.0	0	0	Venta unitaria
ZA	1.0	QUESO NAVARRO MANCHEGO 1 KG	[7503004748204]	222	Público en General	...	422.0	422.0	0	0.0	0.0	0	0.0	0	0	Venta unitaria
ZA	1.0	ADEREZO DE MOSTAZA	[7500326039348]	222	Público en General	...	422.0	422.0	0	0.0	0.0	0	0.0	0	0	Venta unitaria

Ilustración 1. Visualización de la base de datos por tipo de producto

Productos ofertados por unidad

Con esta nueva separación en el tipo de producto ofertado, se observa que el rango definido de las cantidades acumuladas de productos con modalidad de venta unitaria, se encuentra prácticamente bajo el mismo espectro que el obtenido por el análisis anterior.

	count	mean	std	min	25%	50%	75%	max	median	mode
Cantidad	13817.0	1.341753	1.440638	1.0	1.0	1.0	1.0	81.0	1.0	1.0



Gráfica 3. Distribución de probabilidad por venta unitaria

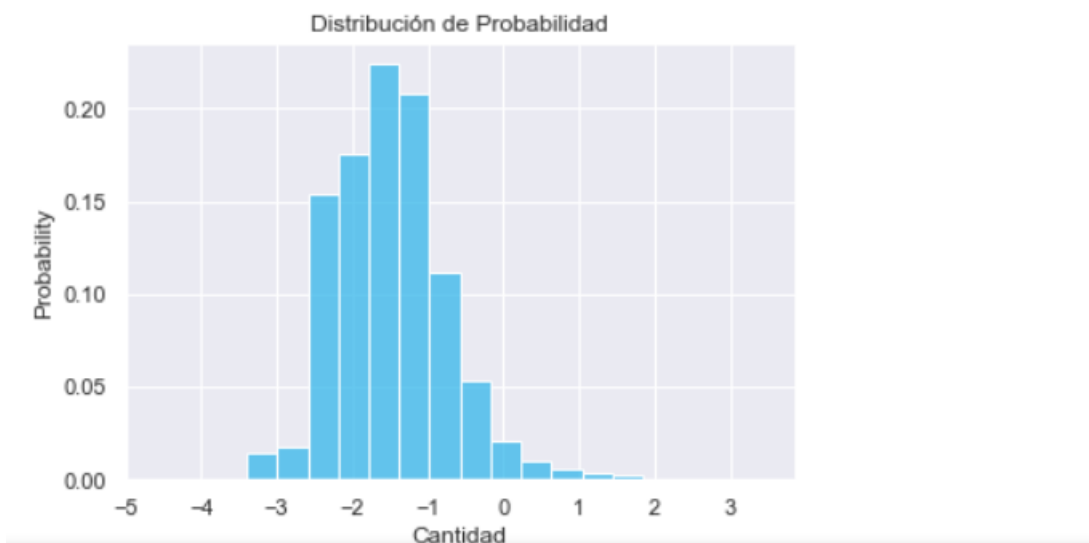
Se percibe que las probabilidades se definen de una manera diferente. Bajo este nuevo escenario propuesto, cerca del 95% de los casos de productos acumulados por ticket con esquema unitario se encuentran entre 1 y 4 productos por cliente, siendo así que la compra de más de 5 productos acumulados por folio comienza a mostrarse como poco común en la muestra.

Con este análisis se presume que con la agrupación anterior llegó a existir un traslape entre los dos tipos de ventas realizadas, es decir, que pueden llegar a existir ventas por gramaje que lleguen a superar más de 1 kg del producto.

Productos ofertados por gramaje

Caso similar a lo anterior, el análisis bajo esta modalidad arroja el mismo rango de los productos ofertados por gramaje.

	count	mean	std	min	25%	50%	75%	max	median	mode
Cantidad	12663.0	0.318956	0.550122	0.01	0.15	0.21	0.315	31.5	0.21	0.1



Grafica 4. Distribución de probabilidad por venta gramaje

Un dato interesante surge al momento de observar que sí existen ventas por gramaje que logran estar en términos de unidades de kilogramos, siendo 31.5 kg nuestro valor máximo para este tipo de productos. De no haber sido por la propia segmentación que se propuso, probablemente no se habría captado la tendencia en cada uno de los formatos de venta, ya sea por gramaje o por unidad.

Tipo de pago

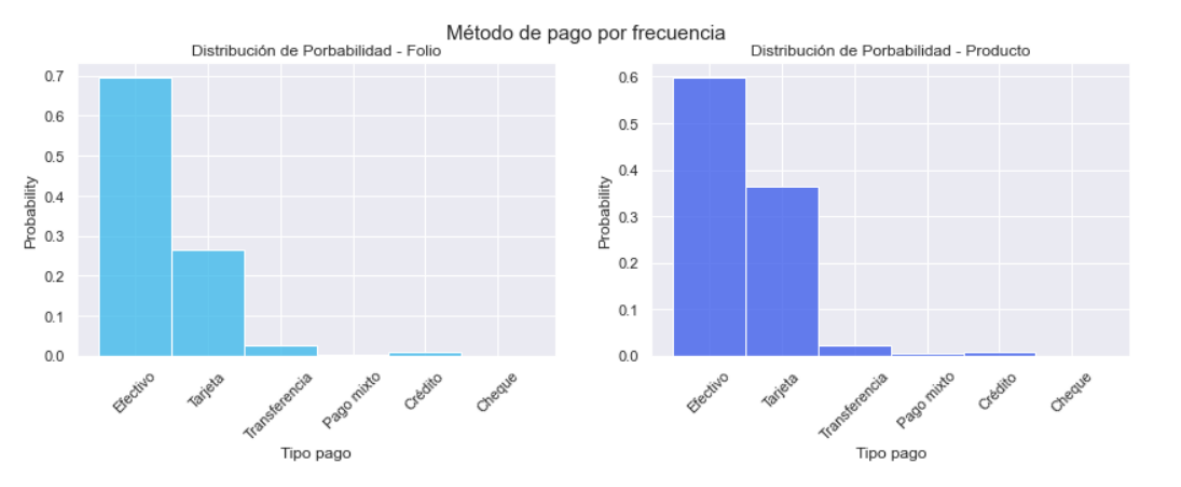
Análisis por frecuencia

Analizando la base de datos por su tipo de pago se observa que se encuentran registradas 26,480 transacciones, no obstante, éstas se encuentran concentradas en 9,326 folios.

Entonces para poder analizar el tipo de pago efectuado en cada folio registrado, se agrupó cada folio y basado en eso se determinó el tipo de pago que se realizó.

De la gráfica de distribución se tiene que los pagos en efectivo representan la mayoría de las transacciones, alrededor de un 70% de las ventas. La segunda categoría más popular refiere a los pagos con tarjeta con un 25% de la muestra total, mientras que el 5% de las observaciones restantes se categorizan en transferencia, crédito, cheque y una categoría que

creamos llamada pago mixto, la cual incluye cualquier método de pago que requiera de más de 1 de los tipos de pago previamente descritos.



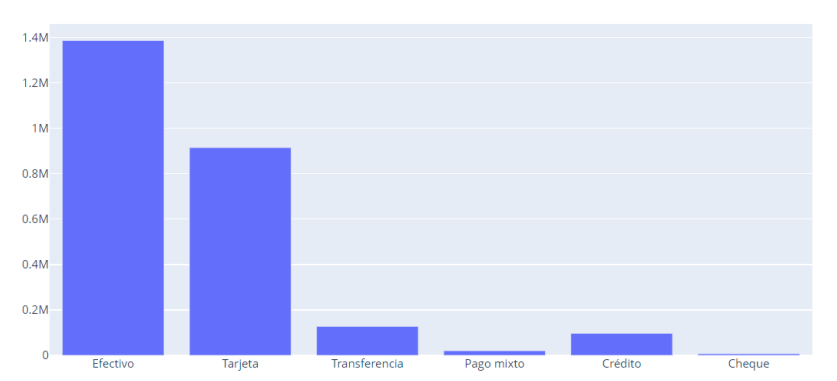
Grafica 5. Distribución de probabilidad de tipo de pago por folio y producto

Complementando el análisis anterior se grafica la distribución correspondiente a cuantos productos fueron adquiridos para cada tipo de pago, y con esto pudimos se observa que los pagos con efectivo continúan siendo la categoría más popular pero ahora solo representan el 60% de la muestra, mientras que los pagos con tarjeta aumentaron en un 10% respecto al análisis por folio.

Análisis por cantidad

Cantidad total en MXN	
Efectivo	1386797.22
Tarjeta	914999.88
Transferencia	127709.88
Pago mixto	20380.57
Crédito	96957.56
Cheque	7092.00

En términos monetarios, el tipo de pago preferido por los clientes es el efectivo, esto refuerza lo identificado anteriormente. El comportamiento del uso de los distintos tipos de pago es constante a través del análisis.



Grafica 6. Distribución de ventas por tipo de pago

Eventos

Dentro la base de datos se encontraron valores atípicos que fueron identificados como eventos cubiertos por la empresa. La magnitud de estos eventos sesga la información que se proporcionó por lo que para fines prácticos se omiten de la base de datos.

Los eventos encontrados responden a los siguientes conceptos:

- Anticipo Evento
- Liquidación Evento
- Evento Brochetas
- Dos Eventos

1.5.2. Análisis del comportamiento de las compras

Contamos con los datos de las compras desde el dos de noviembre del 2020, hasta el quince de marzo del 2022. La base de datos comprende un poco más de un año de información.

Reporte de calidad de datos

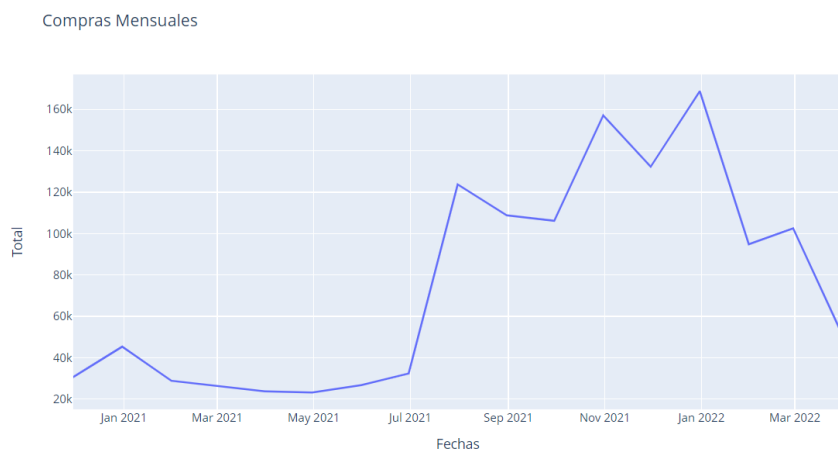
Para obtener generalidades de las variables que contiene la base de datos se realiza el siguiente reporte de calidad de datos que se presenta a continuación.

	Nombre	data_types	missing_values	present_values	unique_values	min	max
Fecha	Fecha	object	0	2093	407	01/02/2022	31/12/2021
Doc.	Doc.	object	0	2093	2	COM	GAS
Folio	Folio	object	2	2091	362	NaN	NaN
Proveedor	Proveedor	object	0	2093	142	ADELA GRANOLA ARTESANAL	YOGURT NATURY HEALTH
Caja	Caja	object	0	2093	1	Caja 1	Caja 1
Usuario	Usuario	object	0	2093	1	admin	admin
Est	Est	object	0	2093	2	V	X
Total	Total	float64	0	2093	1184	-9552	19820

Ilustración 2. Visualización reporte de calidad

Dado la tabla anterior se deduce que se cuenta con información de 2093 registros de compras para 407 días, durante este tiempo se tuvo relación con 142 proveedores. Dentro de la base de datos se registran también las compras concretadas y las compras canceladas con las banderas V y X respectivamente. Del mismo modo, se cuenta con una caja y un único usuario. Como se podemos ver la variable Total muestra un mínimo negativo debido a las cancelaciones.

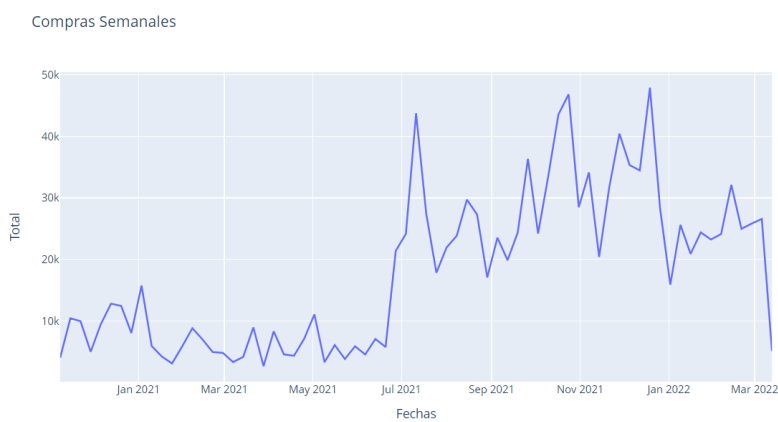
Compras frecuencia mensual



Grafica 7. Compras mensuales

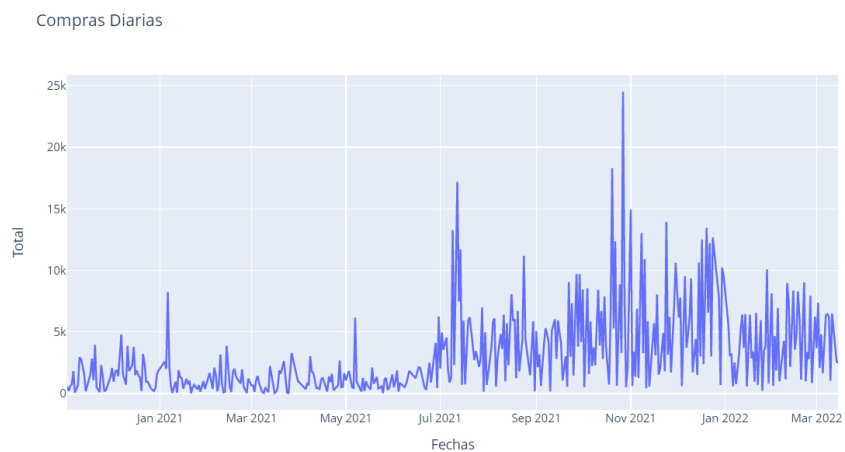
Se observa que a partir de Julio la actividad de compras se mantiene con una tendencia al alza hasta enero que es un mes en el cual las ventas tienden a resentir la cuesta del inicio de año, post fiestas decembrinas.

Compras semanales



Grafica 8. Compras semanales

Compras frecuencia diaria



Grafica 8. Compras diarias

Tanto para las compras con frecuencia semanal como frecuencia diaria, se identifica el mismo comportamiento. Al ir de lo general a lo específico se puede ver una fluctuación más detallada.

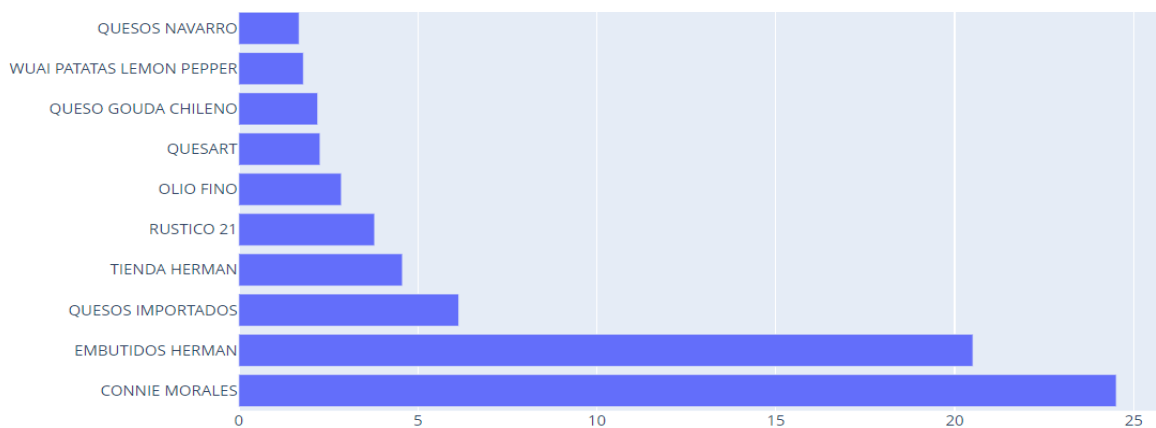
Proveedores

A continuación, se obtienen los 10 proveedores a los que la tienda les compra producto con mayor frecuencia. Es importante destacar que estos 10 proveedores representan aproximadamente el 70.37% de las compras en valor monetario.

	Proveedor	%Participacion
29	CONNIE MORALES	24.513155
43	EMBUTIDOS HERMAN	20.508418
103	QUESOS IMPORTADOS	6.145180
123	TIENDA HERMAN	4.571666
106	RUSTICO 21	3.795199
84	OLIO FINO	2.866561
101	QUESART	2.270834
102	QUESO GOUDA CHILENO	2.205712
138	WUAI PATATAS LEMON PEPPER	1.804673
104	QUESOS NAVARRO	1.685293

Ilustración 3. Principales proveedores

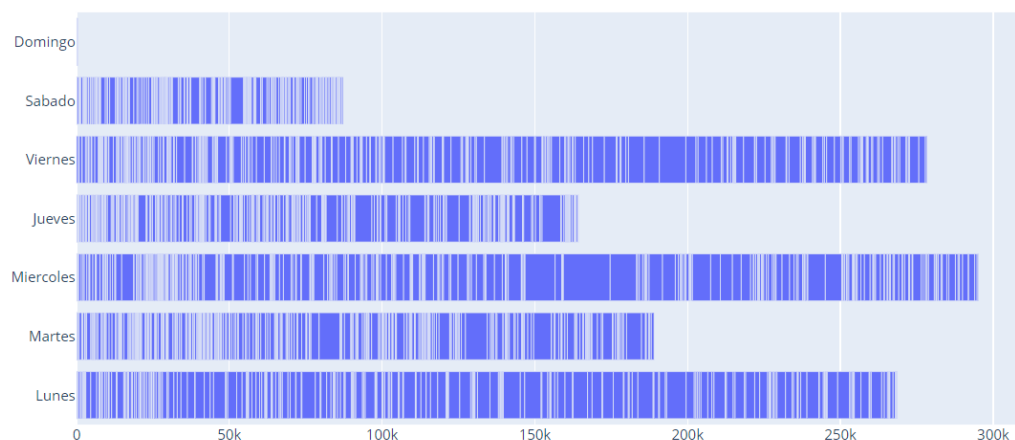
Por el nombre de algunos proveedores se puede inferir lo que la empresa adquiere como en el caso de Quesos Navarro, Olio Fino, etcétera. Sin embargo, para profundizar en el análisis un factor que ayudaría sería tener el desglose de cada folio por producto comprado. El proveedor al que más se le compra es Connie Morales quien es dueña de la tienda principal de Hermann Delicatessen quien también posee la fábrica de embutidos.



Grafica 9. Principales proveedores

Compras por día de la semana

Debido a que la empresa adquiere productos casi a diario, se busca encontrar en que días se adquiere más producto.



Grafica 10. Compras por día de la semana

Se observa que los días en que más se adquiere producto son lunes, miércoles y viernes. Mientras que fines de semana resultan ser los días con menor actividad. Esto podría deberse a dinámica de inventario o demanda de ciertos productos.

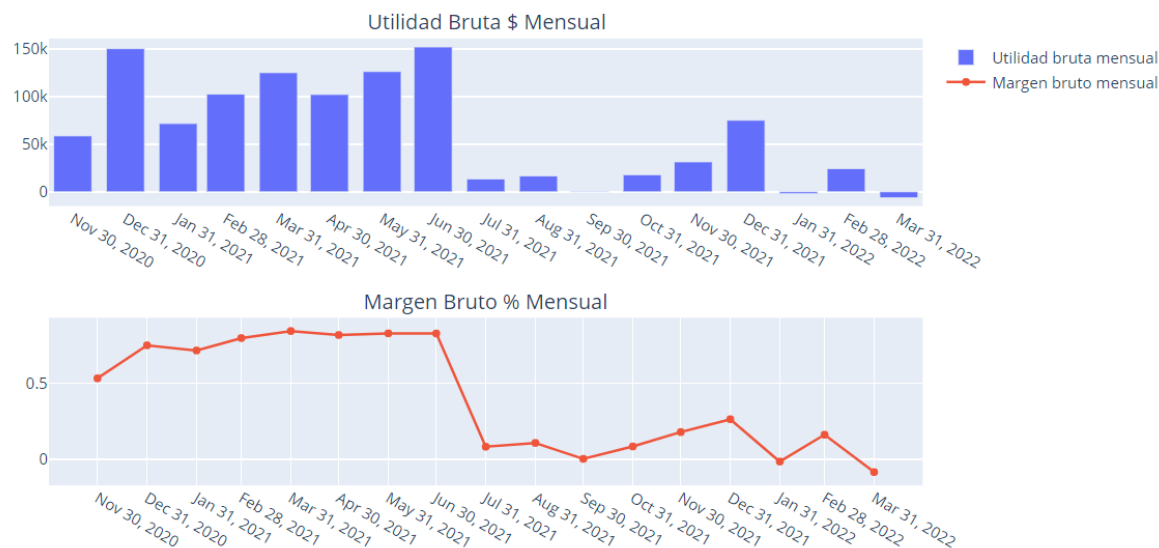
Análisis de margen bruto

Al enterarnos de que el negocio no tiene estados financieros, proponemos el cálculo del margen bruto a partir de las diferencias de las ventas y las compras. Tomando como nuestro costo de ventas el valor de las compras. Dicho análisis se hace para frecuencia mensual.

Rentabilidad bruta frecuencia mensual

Referente a la utilidad bruta se aprecia que los meses con mejor margen fueron diciembre de 2020 y junio de 2021. A partir del 20 de junio se comienza a notar un cambio de tendencia que hace sentido teniendo en cuenta que en estos meses se incrementa la actividad de compras fuertemente como lo que se observó en la serie mensual.

Además, durante el año 2021 se nota una tendencia secundaria alcista de octubre a diciembre. El peor mes es enero del 2022 registrando por primera vez una pérdida de 2,030.25 pesos. Por último, aunque la pérdida en el mes de marzo es de 6,267.55 no se considera el peor mes ya que solo se cuenta con la mitad de los días registrados de las operaciones de marzo.



Grafica 11. Utilidad y Margen Bruto mensual

Referente al margen bruto se aprecia igualmente que los márgenes más favorables se presentan en los meses con un menor volumen de compras, destacan los meses de febrero a junio donde el margen estuvo dentro de un intervalo de 79.44% y 82.43%. Posteriormente como se mencionaba, se da un cambio de tendencia en donde el margen se contrae hasta ser de 8.24% durante Julio, teniendo su peor registro durante enero del 2021 siendo la única ocasión donde se percibe un margen negativo del -1.14% el margen de marzo no se toma en cuenta debido a que solo se tiene datos de la mitad del mes.

1.5.3. Análisis del comportamiento de las ventas

Con el fin de reconocer patrones de compra entre los clientes de Herman Delikatessen, se procede a realizar un análisis puntual de las ventas mensuales, semanales y diarias.

Ventas Mensuales



Grafica 12. Ventas mensuales

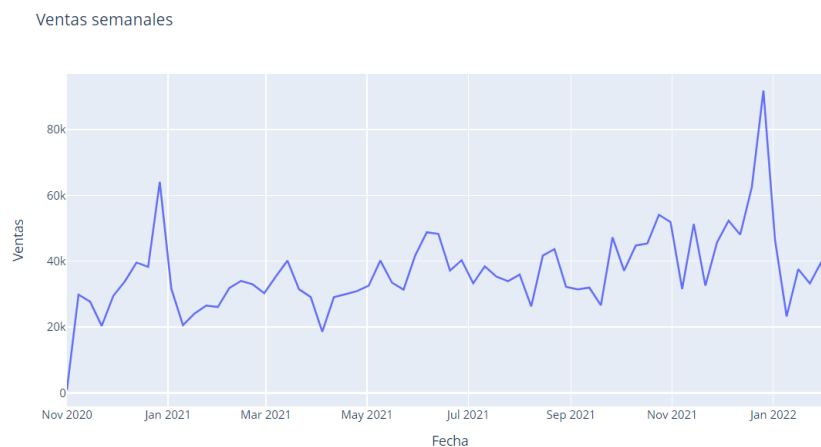
	Fecha	Ventas
13	2021-12-31	286291.57
11	2021-10-31	212227.46
1	2020-12-31	201472.20
7	2021-06-30	184657.04
12	2021-11-30	176067.40

Tabla 1. Top 5 días con más ventas

Con lo reportado en la tabla y el análisis visual del gráfico, se concluye que las ventas al final del año son las más altas en términos monetarios lo cual se debe a las festividades. Las ventas a partir de noviembre de 2021 presentan un alza notoria, y a partir de enero de cada año estas disminuyen, lo cual tiene sentido dado que enero es un mes tranquilo.

Dado los patrones ya identificados en el sector comercial, se puede puntualizar que durante el mes de diciembre se muestra cierta estacionalidad con respecto a sus ventas.

Ventas Semanales



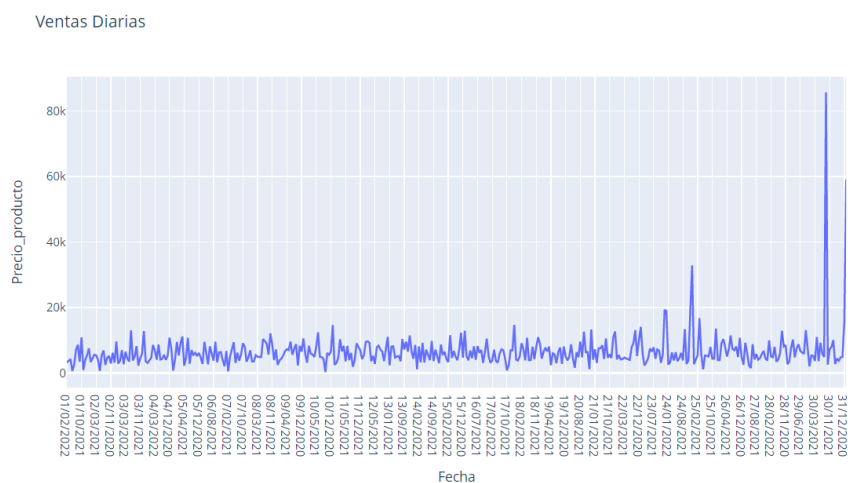
Grafica 13. Ventas semanales

	Fecha	Ventas
60	2021-12-26	91834.91
8	2020-12-27	64064.62
59	2021-12-19	62499.81
51	2021-10-24	54099.66
57	2021-12-05	52317.65

Tabla 2. Top 5 semanas con más ventas

Se reporta en la tabla las 5 semanas con las ventas más altas en términos monetarios. Se observa que cuatro de las cinco semanas reportadas pertenecen al mes de diciembre, tres de estas diciembre 2021. Claramente se confirma que el mes de diciembre es el mejor mes en ventas confirmando que el comportamiento de las ventas tiene cierta estacionalidad.

Ventas diarias



Grafica 14. Ventas diarias

	Fecha	Precio_producto
414	30/10/2021	85671.33
425	31/12/2021	59041.31
341	24/12/2021	32865.86
326	23/12/2020	19297.75
340	24/12/2020	18912.84

Tabla 3. Días con mas ventas

En la tabla se muestran los 5 días donde se registraron las ventas más altas en términos monetarios. Cuatro de las cinco fechas se encuentran cercanas a una festividad en este caso año nuevo y Navidad.

Agrupación del número de transacciones por día de la semana

Agrupando por día de la semana donde 1 corresponde a Domingo, se obtiene el total de dinero por grupo, el promedio y el total de transacciones (folios) dependiendo el día de la semana. Para este análisis usaremos el número de folios con el fin de observar cómo se comportan las ventas conforme el día de la semana.

Conforme a la gráfica siguiente, a lo largo de la semana las ventas aumentan, a partir del miércoles el aumento es más notorio continuando con el mismo comportamiento hasta llegar al viernes, que es el día con mayores ventas registradas.



Grafica 15. Transacciones por día de la semana

Las ventas del sábado son menores a las del viernes, esto puede deberse a que solo abren medio día, sin embargo, siguen siendo mayores a las del miércoles y el jueves.

La mayor cantidad de transacciones ocurren los fines de semana. Tomando en cuenta que el domingo es día de descanso, los lunes y los martes son los días con menor actividad, no obstante, esto no significa que sean días malos en cuestión de ventas puesto que por día superan los 1500 folios totales.

Agrupación por día del mes



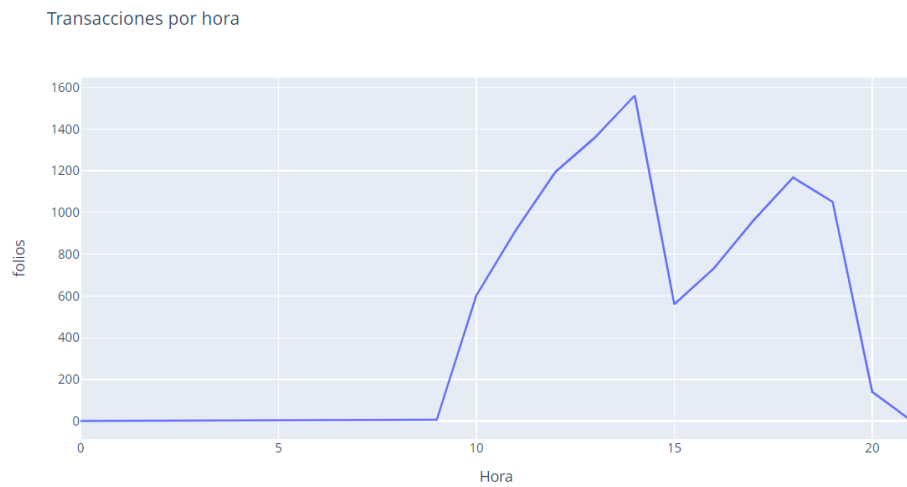
Grafica 15. Transacciones por día del mes

Unnamed: 0	Dia_mes	folios	numero_productos	total_dinero	Promedio_Total
29	30	30	305	908	476248.86
30	31	31	224	670	311344.20
28	29	29	288	855	391938.03
25	26	26	355	1016	465311.38
24	25	25	289	799	365667.06

Tabla 4. Días con más ventas

Observando la gráfica que muestra el análisis de las ventas agrupadas por día del mes a través del número de folios y la tabla anterior en donde se presenta el top 5 de los días con mayores ventas tomando el Promedio total, es claro que tres de los cinco días corresponden al fin de mes. Con esto se infiere que al final de mes existe una fuerte actividad de venta. Esto va de la mano con que coincide con el pago de salarios en la mayoría de los empleos.

Agrupación por hora



Grafica 16. Transacciones por hora

Unnamed: 0	Hora	folios	numero_productos	total_dinero	avg_Total
7	8	14	1560	4774	2230166.79
6	7	13	1360	4182	1795323.36
5	6	12	1195	3441	1377892.22
11	12	18	1168	3340	1314533.75
12	13	19	1051	3032	1134896.44

Tabla 5. Horas con más transacciones

Analizando la información en base al número de folios (transacciones) por hora, las horas con mayores transacciones son de 12 a 2 pm y de 6 a 7 pm.



Grafica 17. Cantidad de dinero vendida por hora

Unnamed: 0	Hora	folios	numero_productos	total_dinero	avg_Total
7	8	14	1560	4774	2230166.79
6	7	13	1360	4182	1795323.36
5	6	12	1195	3441	1377892.22
11	12	18	1168	3340	1314533.75
12	13	19	1051	3032	1134896.44

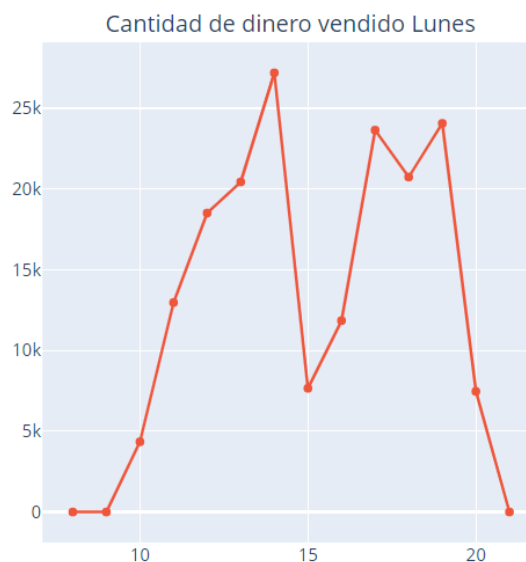
Tabla 6. Horas con mayor cantidad de dinero vendida

Al analizar por total de cantidad vendida por hora el top 5 de las mejores horas de venta coincide con el análisis de número de transacciones por hora.

Por tanto, se concluye que las mejores horas de venta son de 12 a 2 pm y de 6 a 7 pm que coinciden con los horarios de comida y/o salida de la mayoría de los trabajos.

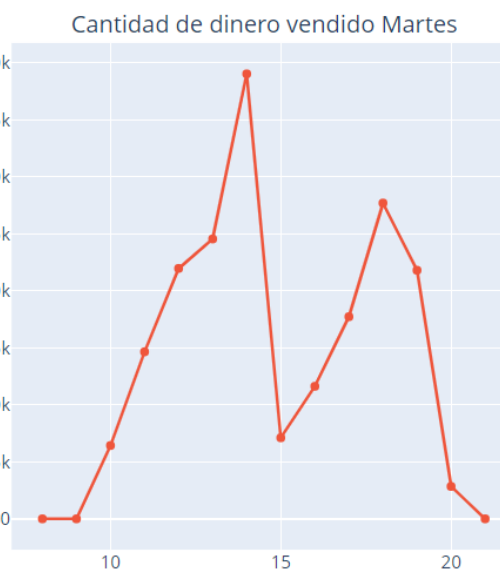
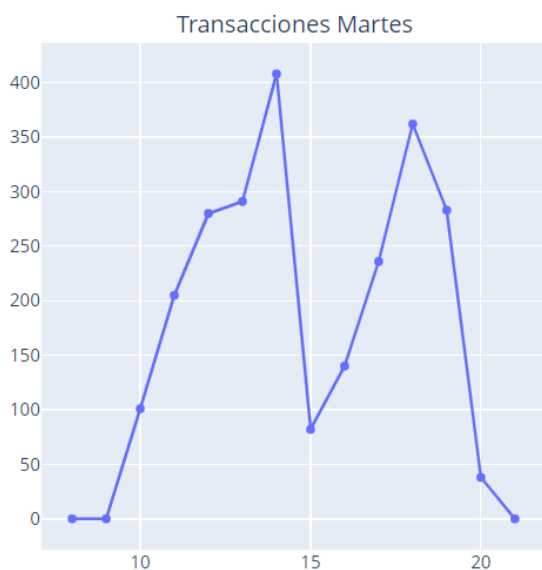
Ventas por día de la semana y hora

Lunes



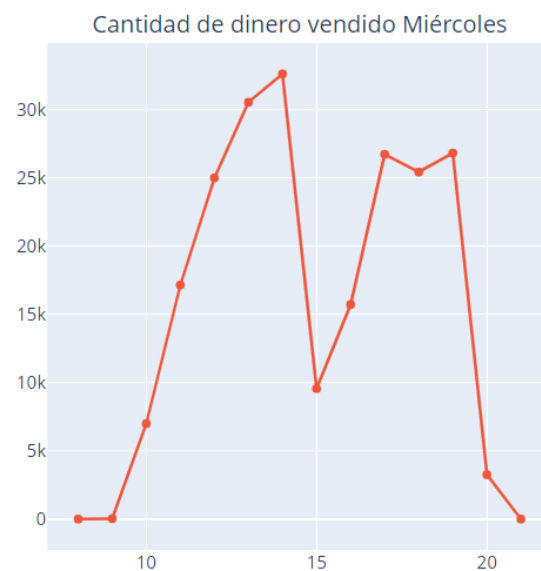
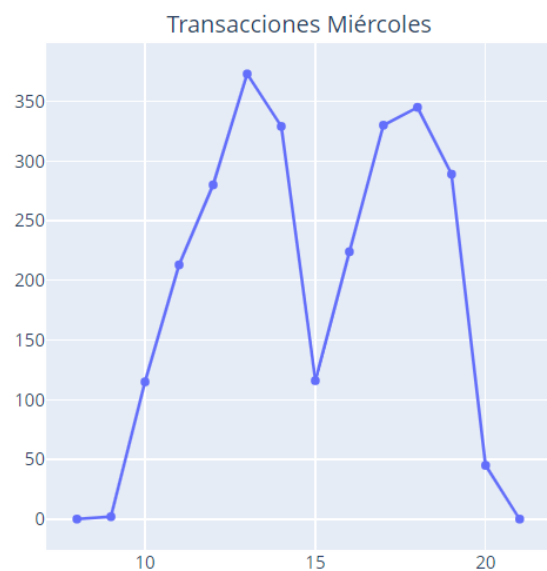
Grafica 18. Ventas lunes

Martes



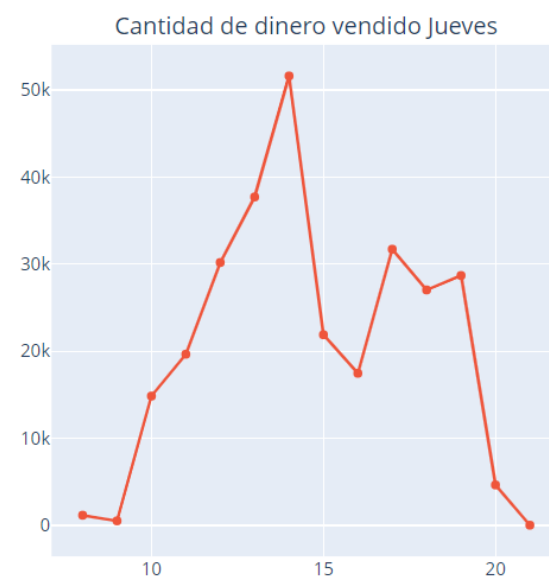
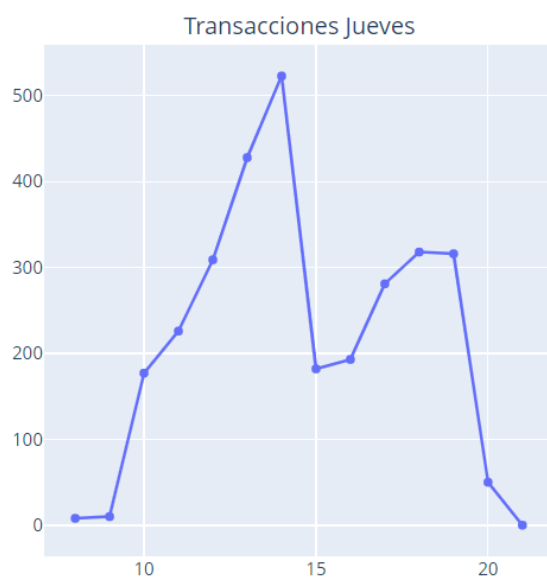
Grafica 19. Ventas martes

Miércoles



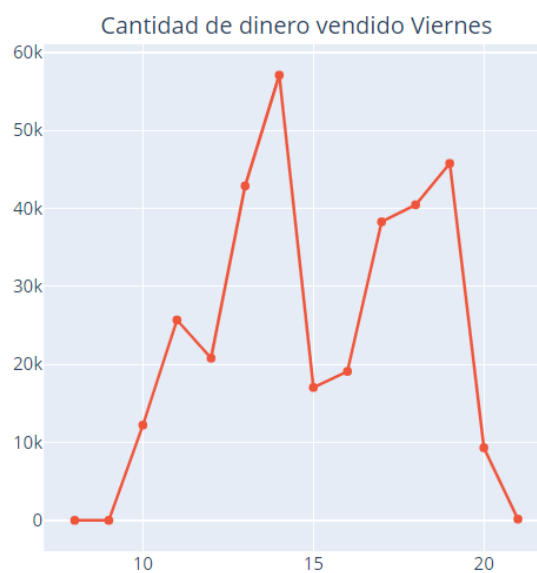
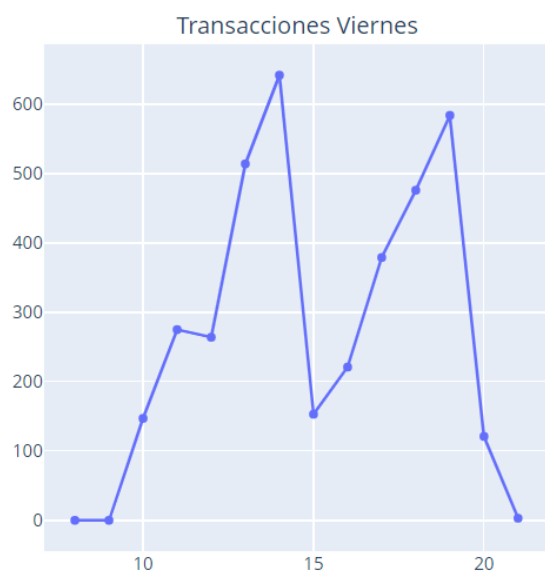
Grafica 20. Ventas miércoles

Jueves



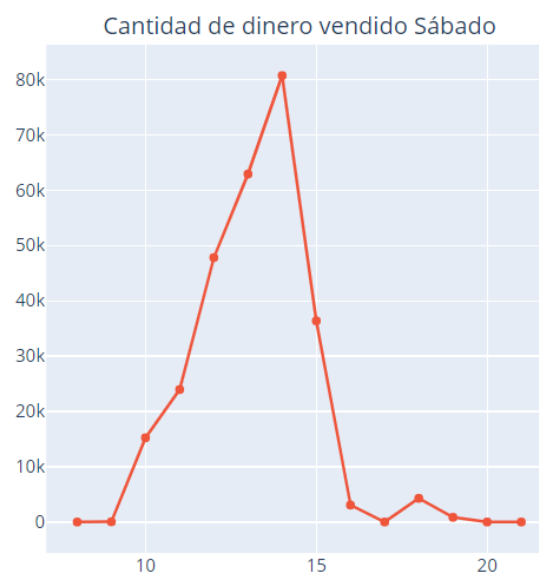
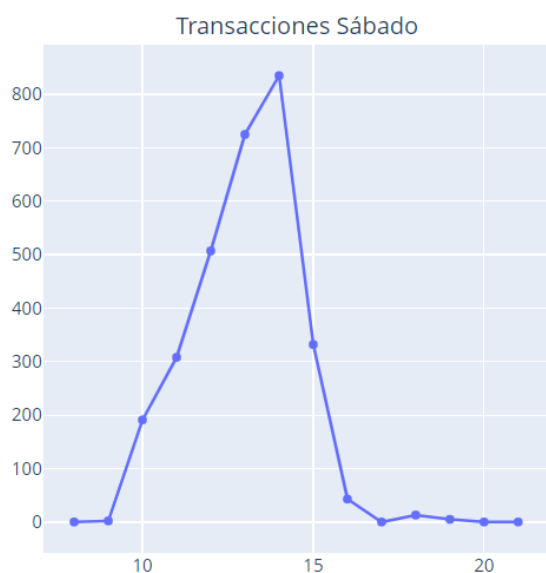
Grafica 21. Ventas jueves

Viernes



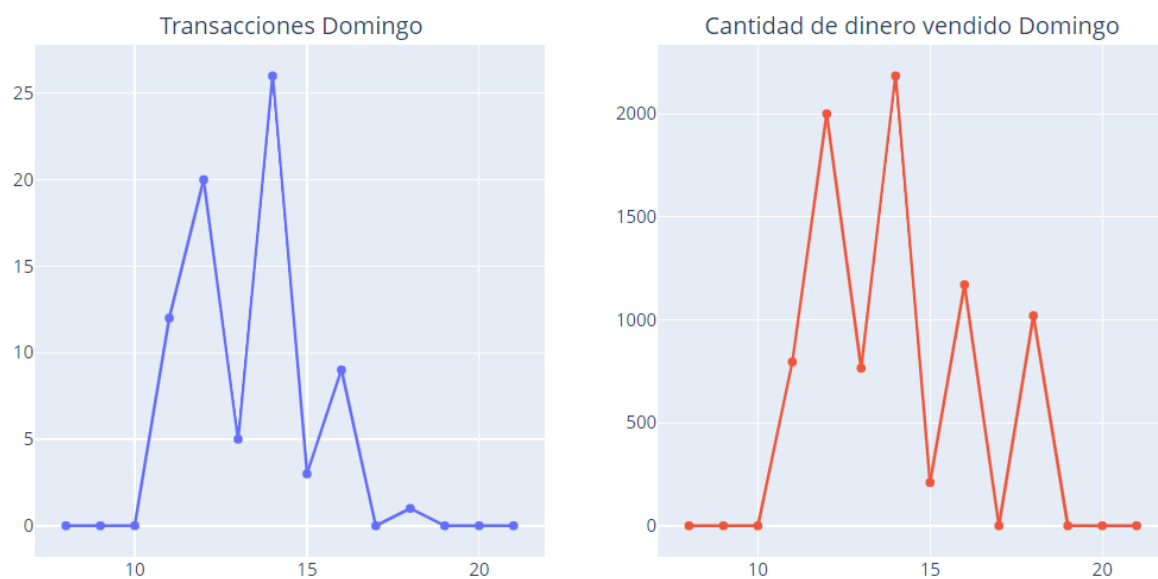
Grafica 22. Ventas viernes

Sábado



Grafica 23. Ventas sábado

Domingo



Grafica 24. Ventas domingo

Graficando la cantidad de transacciones como su valor en términos de dinero por hora y día de la semana hace más fácil identificar los horarios en los cuáles se encuentran las horas pico, es decir, cuándo la tienda tiene más trabajo. Dentro de la comparativa, se aprecia que los días siguen un comportamiento relativamente similar a excepción de sábado y domingo. Para el sábado, como ya se ha dicho anteriormente, se trabaja solo medio día, por lo que su pico de ventas está entre la 1 y las 3 de la tarde. Si bien tenemos datos fuera del horario laboral son pocos.

El domingo es un día no laboral, las transacciones que se presentan lo más probable es que sean cortes de caja o incluso transacciones que quedaron pendientes de días anteriores, por lo que los montos no se comparan a los que se tienen en días laborales.

Productos

Uno de los objetivos de este análisis operativo es identificar cuáles son los productos que más y menos se venden tanto en cantidad como en términos monetarios.

Análisis por cantidad

Productos más vendidos

Los productos que más se venden por cantidad son los baguettes, brownies, pan para tapas, coca cola light y el jamón york. Estos productos tienen en común que el rango de precios en los que se encuentran es bajo en comparación con otros de los productos que se venden en la tienda. Además, estos productos son de consumo popular por lo que el hecho de que en unidades sean de los productos más vendidos hace sentido.

De la misma manera, varios de estos productos van de la mano con los principales productos de las tiendas Herman, las carnes frías.

	Producto	Cantidad
119	BAGUETTE SUAVE MASA MADRE	1526.000
560	DELISCO BROWNIES MARIFER	598.000
885	PAN PARA TAPAS	509.000
506	COCA-COLA LIGHT 500 ML	444.000
676	JAMON YORK HD	404.558

Tabla 7. Productos más vendidos

Productos menos vendidos

Los productos menos vendidos son los presentados en la tabla siguiente. El hecho de que estos productos sean menos vendidos está relacionado con que no son productos populares entre los clientes. Igualmente, estos productos no son inventario usual de una tienda Herman.

	Producto	Cantidad
977	QUESO CON HABANERO	0.1400
683	JITOMATE	0.1800
1072	SALMÓN PORCIÓN SAN BLAS	0.1900
627	GALANTINA	0.2048
1011	QUESO PARMESANO RALLADO	0.2300

Tabla 8. Productos menos vendidos

Análisis Monetario

Se considera importante realizar, de la misma manera, un análisis monetario sobre los productos que se venden, ya que es relevante identificar los productos que lideran el ingreso de flujos de efectivo del negocio.

Productos más vendidos (por cantidad de dinero)

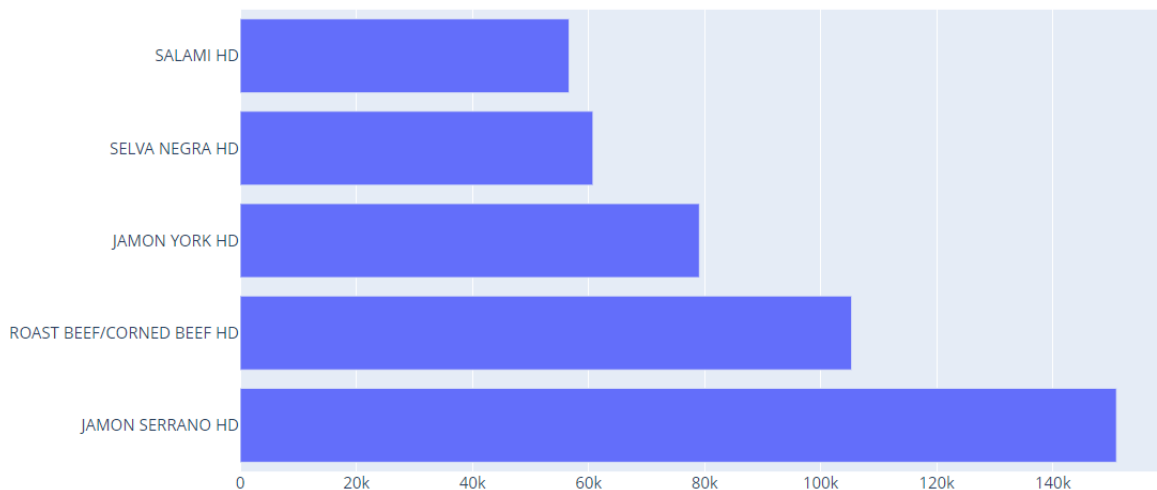
Los 5 productos que más se venden monetariamente conforman el 17.05% de las ventas y se muestran a continuación:

	Producto	Precio_producto
672	JAMON SERRANO HD	150955.65
1029	ROAST BEEF/CORNERED BEEF HD	105326.96
676	JAMON YORK HD	79095.98
1116	SELVA NEGRA HD	60763.94
1050	SALAMI HD	56649.84

Tabla 9. Productos más vendidos

Tal como se muestra en la gráfica siguiente, el producto que más se compra en términos monetarios es el jamón serrano, seguido muy cercanamente por el roast beef, siendo las carnes frías el giro del negocio, es de esperarse que sean los que más dinero atraen.

Del mismo modo, dentro del rango de precios, las carnes frías tienen a caracterizarse por su costo elevado, aunque estos productos no aparecen como los más vendidos en términos unitarios, si pueden ser lo que más ingresos generan.



Grafica 25. Productos más vendidos

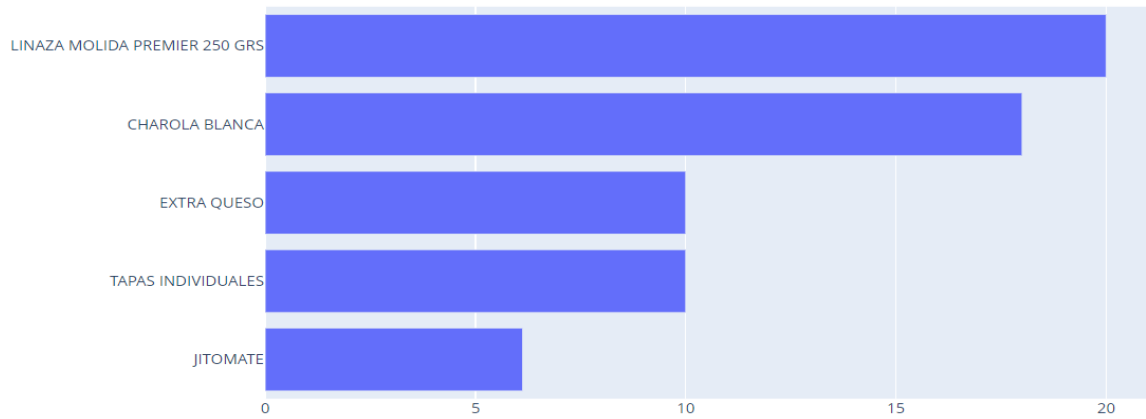
Productos menos vendidos (por cantidad de dinero)

Los 5 productos que menos se venden monetariamente se muestran en la siguiente tabla. Estos productos conforman el 0.0024% de las ventas.

	Producto	Precio_producto
683	JITOMATE	6.12
1141	TAPAS INDIVIDUALES	10.00
602	EXTRA QUESO	10.00
341	CHAROLA BLANCA	18.00
736	LINAZA MOLIDA PREMIER 250 GRS	20.00

Tabla 10. Productos menos vendidos

Complementando el análisis anterior, los productos que menos se venden no son propiamente el giro de la empresa, del mismo modo, el rango de precios en los que se ubican estos productos es bajo comparándolo con los productos que más dinero recaudan.



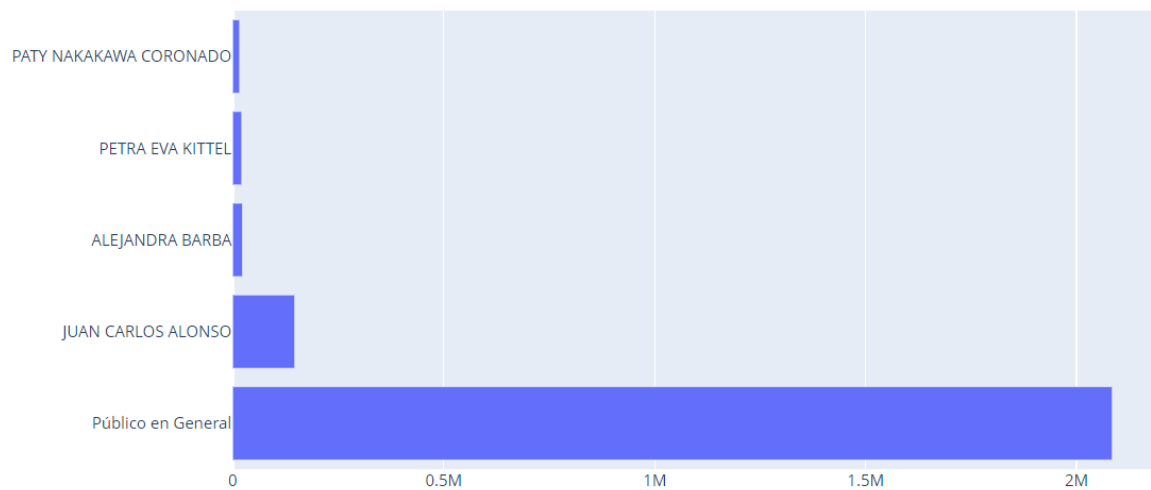
Grafica 26. Productos menos vendidos

Análisis por clientes

Cientes más activos

	Cliente	Precio_producto
231	Público en General	2084730.53
141	JUAN CARLOS ALONSO	146825.57
6	ALEJANDRA BARBA	23078.17
228	PETRA EVA KITTEL	21157.25
224	PATY NAKAKAWA CORONADO	16167.33

Tabla 11. Cientes más activos



Grafica 27. Cientes más activos

Al analizar la base de datos que concierne a los clientes se puede rescatar que la mayoría de los clientes que acuden a la tienda no están registrados, es decir, se consideran como ventas al público general. Si bien existen clientes de los que se tienen datos y sobre los cuales se puede llegar a varias conclusiones, el 78% de las ventas que genera la tienda están dentro de los clientes de los cuales no tenemos información por lo que esto da pie a la siguiente fase de trabajo en la cual estaremos aplicando clustering para crear los perfiles de cliente.

1.5.4. Segmentación de clientes

Tal como se mencionó en el apartado anterior, al no contar con registros suficiente que permitan generar clasificaciones de los clientes de la tienda, es necesario reconocer los patrones de compra de los anteriormente mencionados para así poder hacer recomendaciones para la empresa.

Este proceso se realizará aplicando un algoritmo de clustering no supervisado. A continuación, se muestra la metodología llevada a cabo.

Justificación del modelo

Para la segmentación de clientes de Herman Delikatessen, lo que se hará será definirla a partir de la detección de patrones en cuanto a las características propias de cada folio, específicamente a través de las cantidades adquiridas por cada segmento de producto que tiene la tienda en su esquema de ventas (ofertados en unidades, ofertados por gramaje) y su relación con el total a pagar asociado a cada ticket. Entonces, teniendo una perspectiva global de lo que se busca realizar, se define dicha obtención de segmentos de clientes a partir de algún modelo de clustering que sea capaz de ayudarnos a reconocer y clasificar a los tipos de compradores que atrae Herman Delikatessen.

En esta situación particular nosotros proponemos generar una clasificación que responda a las siguientes tres características:

- Aprendizaje no supervisado: Ya que a priori nosotros no podemos determinar el número de tipos de clientes ni la clasificación propia de estos (al final de cuentas es el objetivo final de la segmentación), por ello requerimos algún modelo de clustering que sea

capaz de clasificar estas estructuras de datos, sin necesidad de nosotros tener que parametrizarlos.

- Basado en densidad: Como bien se sabe, existe una amplia gama de modelos de clustering no supervisado entre ellos tenemos lo que se basan en densidad. A continuación se busca dar una explicación más formal de esto, sin embargo, rápidamente se comenta que nosotros buscamos agrupar a los clientes por su patrón de compra, entonces a medida que una mayor cantidad de ventas manejen una estructura similar en sus métricas esperamos sea un conjunto de datos denso, lo cual podemos traducir en similitud.
- Capacidad para lidiar valores extremos: Esto debido a que, si bien ya se eliminaron los eventos para términos del análisis propuesto, la realidad es que la diferencia en cuanto a magnitud de cada folio varía considerablemente, esto principalmente a que en términos monetarios no todos los productos tienen el mismo precio.

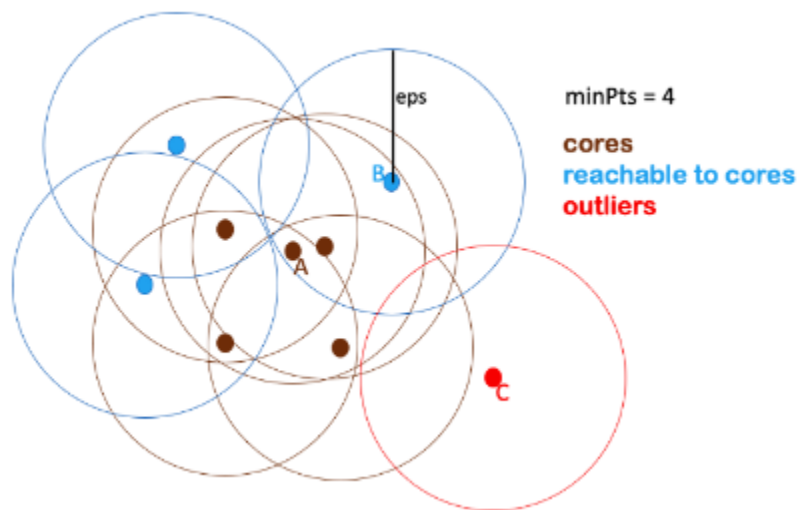
Con estas tres características principales ya definidas, el algoritmo que consideramos puede arrojar resultados más interesantes es el DBSCAN, el cual responde al nombre de Density-Based Spatial Clustering of Applications with Noise, el cual contiene se describe brevemente a continuación.

Como su propio nombre lo indica, el DBSCAN es un algoritmo de clustering que se basa en un agrupamiento por densidad y es justamente esta propiedad la que hace que el manejo que tiene para valores "ruidosos" sea muy bueno. Pero aquí es importante ir definiendo los conceptos que se van manejando, primero que nada ¿qué es densidad?, y para responder esta pregunta existe una respuesta muy sencilla, básicamente podemos describir como densidad a la cantidad de datos que se encuentran en una misma región o área, entonces a medida que tenemos más cantidad de datos en una misma región podremos decir que esa región en especial alberga una densidad de datos alta.

El concepto básico sobre el cual se mueve el DBSCAN es el término de densidad, sin embargo, aquí empiezan a surgir algunas otras dudas muy concretas, ¿cómo defino una región?, ¿a partir de que cantidad de datos puedo considerar una densidad alta? Son justamente estas dos interrogantes las que dan entrada a los hiperparámetros que maneja este algoritmo:

1. Epsilon (ϵ): El cual hace referencia a la distancia máxima que debe existir entre dos puntos para que puedan ser considerados de la misma región (generalmente es calculada a partir de la distancia euclidiana).
2. Minimal Points: Cantidad mínima de puntos que deben estar dentro de la distancia ϵ para que un punto se pueda considerar como un núcleo (core point). Es decir, que se defina como centroides que posteriormente irán conformando los clusters.

A partir de estos dos parámetros podemos ir definiendo entonces los diferentes puntos que se pueden encontrar en el desarrollo del DBSCAN. A continuación, se ilustra mediante una imagen del paper original dicha situación:



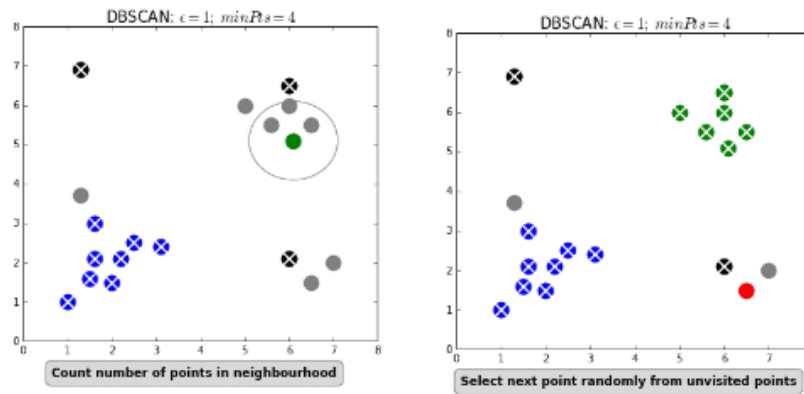
Si observamos con atención la imagen descrita podemos ir encontrando los siguientes puntos:

- **Cores (A):** Estos puntos corresponden a aquellos que cumplen con el requerimiento de contar con la cantidad mínima de puntos que se encuentren dentro del radio definido por ϵ .
- **Reacheable to cores:** También definidos como puntos de borde, estos hacen referencia a todos aquellos puntos que no cumplen con el requerimiento de la cantidad mínima de puntos, pero que sí se encuentran dentro del alcance de un punto núcleo, de manera que posteriormente pasarán a formar parte de su cluster.
- **Outliers:** O puntos atípicos hacen referencia a todo aquel punto que ni sea un punto de núcleo ni un punto de borde, de manera que estos no serán considerados para ningún cluster en particular.

Entonces ya con los elementos que definen el algoritmo del DBSCAN, podemos ahora dar una breve explicación de cómo es que se va desarrollando paso por paso:

1. Para cada punto en el espacio que cumpla con los requerimientos de distancia ϵ y cantidad mínima de "vecinos" se le definirá como puntos de núcleo.
2. Se comienzan a definir los grupos caracterizados por los núcleos y todos aquellos puntos de borde que sean alcanzados por estos pasar a formar parte dicho cluster.
3. En el caso de encontrarse frente a un punto que ni sea de núcleo ni de borde, se le categorizará como punto atípico, es decir, no contará con una clasificación.

A manera de resumen visual, se presenta el siguiente GIF el cual desglosa paso a paso el actuar del DBSCAN. Dicho ejemplo utiliza un $\epsilon=1$ y una cantidad mínima de puntos de 4:



Entonces a manera de resumen y entendiendo lo que caracteriza al funcionamiento del DBSCAN, podemos decir que a primera instancia pareciera ser un modelo de clustering óptimo para estar trabajando el conjunto de datos que se desea. A manera de contextualización a continuación se definirán las variables que pasar a formar parte de la clasificación:

- x : Cantidad de productos unitarios que adquiere un cliente. Desglosado por producto en cada folio.
- y : Cantidad de productos en gramaje que adquiere un cliente. Desglosado por producto en cada folio.
- z : Total pagado por cada folio.

Procesamiento de datos.

Dentro de la definición de la parametrización correspondiente a la agrupación de patrones de compra para Herman Delikatessen, una de las complicaciones más grandes que se enfrentaron fue la estructuración de la base de datos, ya que no se contaba con un conjunto de datos que estuviera directamente preparado para la aplicación del algoritmo.

Ante esta situación se propone el siguiente método de cruce de bases para poder obtener los atributos deseado:

1. Separar la base de datos original en dos subconjuntos, uno que incluya solo productos ofertados en esquema unitario y otro que corresponda a la estructura por gramaje.

2. Ya con esos subconjuntos generados nos encontramos ante la situación donde las longitudes para cada tipo de venta no corresponden, para arreglar esa situación se propone la generación de una nueva categoría llamada “sin producto” la cual corresponde a un valor de cero en las cantidades donde la diferencia en los esquemas de venta sea palpable.
3. Se extraen todos los atributos de la base de datos original que se considera pueden aportar más valor al momento de recuperar información para los grupos que potencialmente se generan.

Fecha	Dia_semana	Dia_mes	Folio	Cantidad_Unitaria	Producto_Unitario	Cantidad_Gramaje	Producto_Gramaje	Total	Tipo_pago
01/11/2020	1	1	219	1.0	KOMBUCHITA JENGIBRE	0.000	SIN PRODUCTO	39.00	Efectivo
01/11/2020	1	1	220	1.0	KOMBUCHITA JAMAICA	0.000	SIN PRODUCTO	39.00	Tarjeta
01/11/2020	1	1	221	1.0	MAIZ INFLADO	0.000	SIN PRODUCTO	30.00	Tarjeta
01/11/2020	1	1	222	1.0	QUESO NAVARRO MANCHEGO 1 KG	0.000	SIN PRODUCTO	422.00	Efectivo
01/11/2020	1	1	222	1.0	ADEREZO DE MOSTAZA	0.000	SIN PRODUCTO	422.00	Efectivo
...	...	...	...	...	...	...	...	...	...
01/02/2022	3	1	9603	0.0	SIN PRODUCTO	0.255	SELVA NEGRA HD	105.06	Efectivo
01/02/2022	3	1	9604	1.0	WUAI PATATAS LEMON PEPPER 500 GRS	0.000	SIN PRODUCTO	631.00	Tarjeta
01/02/2022	3	1	9604	1.0	BAGUETTE SUAVE MASA MADRE	0.000	SIN PRODUCTO	631.00	Tarjeta
01/02/2022	3	1	9604	1.0	PAN PARA TAPAS	0.000	SIN PRODUCTO	631.00	Tarjeta
01/02/2022	3	1	9604	1.0	CHAROLA GRANDE	0.000	SIN PRODUCTO	631.00	Tarjeta

En función de esta nueva base de datos, se seleccionan las variables de interés para la segmentación de clientes que son las siguientes:

Fecha	Cantidad_Unitaria	Cantidad_Gramaje	Total
14/05/2021	2.0	0.000	610.67
12/12/2020	1.0	0.540	146.90
20/02/2021	1.0	0.000	310.00
30/12/2021	0.0	0.100	164.16
19/02/2021	1.0	0.105	945.32
28/11/2020	1.0	1.005	268.09

Definición del modelo de clustering

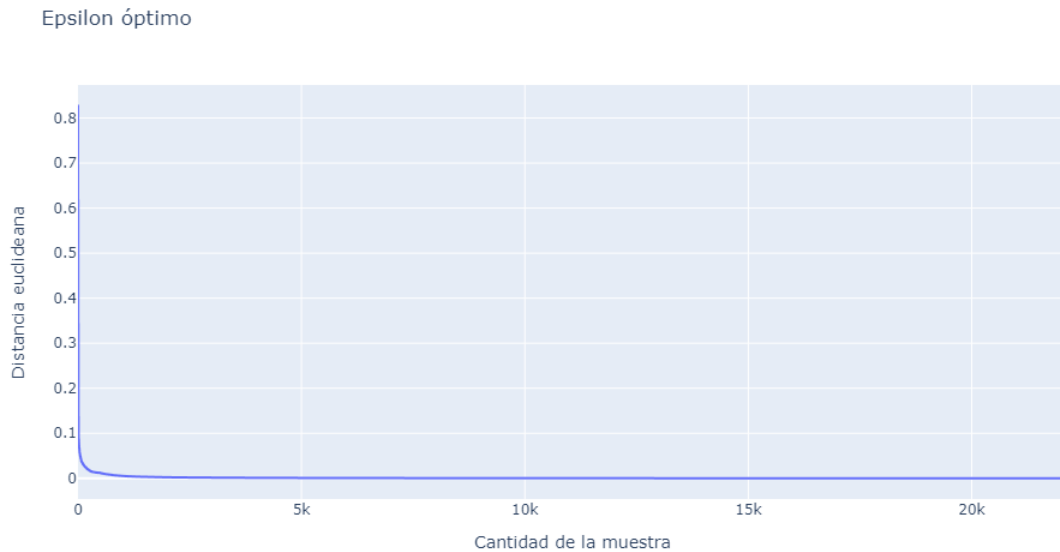
Para definir el modelo se opta por normalizar los datos para que sea más sencillo trabajar con ellos, ya que las magnitudes sobre las cuales se mueven los folios en términos monetarios, así como en cantidades puede ser muy dispares.

	Cantidad_Unitaria	Cantidad_Gramaje	Total
1416	0.000000	0.019206	0.076795
3898	0.012346	0.004921	0.010566
4242	0.012346	0.000000	0.018608
7439	0.024691	0.000000	0.019313
6880	0.024691	0.000000	0.029188
7666	0.000000	0.001746	0.019090

Una vez con los atributos necesarios y los datos escalados, se puede comenzar a definir los hiperpárametros que definirán el DBSCAN a utilizar. El primero de ellos conocido épsilon (ϵ) se define a través de una gráfica de distancias entre los puntos más cercanos.

Posterior a esto se define el hiperparámetro épsilon como un valor óptimo dada la gráfica siguiente, donde se muestra que este valor corresponde a 0.035.

Aplicando el algoritmo DBSCAN se reconocen 12 agrupaciones.



Parametrización de clientes.

Con la intención de parametrizar de una manera más eficiente a nuestros clientes, lo que se hará será definir una función generalizada para la obtención de cada tipo de cliente a partir del cluster generado

Los tipos de cliente obtenidos son los siguientes

Cliente tipo 1

Estos son los productos ofertados por el esquema de **ventas unitarias** con mayor frecuencia en el primer grupo generado por el *clustering*

	Dia_semana	Dia_mes	Folio	Cantidad_Unitaria	Cantidad_Gramaje	Producto_Gramaje	Total	Tipo_pago	Cluster
Producto_Unitario									
BAGUETTE SUAVE MASA MADRE	810	810	810	810	810	810	810	810	810
PAN PARA TAPAS	402	402	402	402	402	402	402	402	402
COCA-COLA LIGHT 500 ML	370	370	370	370	370	370	370	370	370
DELISCO BROWNIES MARIFER	318	318	318	318	318	318	318	318	318
CHIPS GDE COMALLI	231	231	231	231	231	231	231	231	231

Estos son los productos ofertados por el esquema de **ventas por gramaje** con mayor frecuencia en el primer grupo generado por el clustering

	Dia_semana	Dia_mes	Folio	Cantidad_Unitaria	Producto_Unitario	Cantidad_Gramaje	Total	Tipo_pago	Cluster
Producto_Gramaje									
JAMON SERRANO HD	1066	1066	1066	1066	1066	1066	1066	1066	1066
SELVA NEGRA HD	1052	1052	1052	1052	1052	1052	1052	1052	1052
JAMON YORK HD	939	939	939	939	939	939	939	939	939
LOMO CANADIENSE HD	714	714	714	714	714	714	714	714	714
ROAST BEEF/CORNE BEEF HD	679	679	679	679	679	679	679	679	679

Aquí se observa la parametrización correspondiente al primer tipo de cliente que encuentra en la tienda. De manera generalizada se puede mencionar que este primer grupo seleccionado corresponde a clientes que manejan un perfil tipo **retail**, esto debido a que en cuanto a cantidad consumida como a su gasto promedio se encuentran bajo esquemas de consumo relativamente bajos. Es importante mencionar que esta categoría corresponde es la que contiene una mayor cantidad de datos en su agrupación, con esto en cuenta podemos decir que este tipo de cliente es el que genera la mayoría de compras para **Herman Delikatessen**.

Nombre Producto	Rango Consumo	Cantidad Consumida	Gasto Promedio
BAGUETTE SUAVE MASA MADRE	1.0 - 8.0 PZAs	2.0000	339.8743
PAN PARA TAPAS	1.0 - 4.0 PZAs	1.0000	472.5244
COCA-COLA LIGHT 500 ML	1.0 - 2.0 PZAs	1.0000	85.9859
DELISCO BROWNIES MARIFER	1.0 - 10.0 PZAs	2.0000	132.7245
CHIPS GDE COMALLI	1.0 - 4.0 PZAs	1.0000	241.3711
JAMON SERRANO HD	0.02 - 4.02 Kg	0.2483	404.6302
SELVA NEGRA HD	0.02 - 2.27 Kg	0.2364	358.8191
JAMON YORK HD	0.055 - 2.55 Kg	0.3576	274.4390
LOMO CANADIENSE HD	0.02 - 2.005 Kg	0.2157	342.0273
ROAST BEEF/CORNE BEEF HD	0.05 - 4.515 Kg	0.3739	460.2887

Cliente tipo 2

Del tipo de cliente número 2 que se encuentra a partir de la segmentación se observa que la cantidad consumida aumenta para este caso, así como el gasto promedio del ticket. Por lo que podemos inferir que ya se trata de clientes más especializados en un producto y que probablemente no sean de consumo retail. También se puede notar por los rangos de consumo de los productos.

Nombre Producto	Rango Consumo	Cantidad Consumida	Gasto Promedio
6 PACK DEGUSTACION OLIO FINO	1.0 - 2.0 PZAs	2.0000	2952.0800
BALSAMICO BLANCO ARANDANO PERA OLIO	1.0 - 1.0 PZAs	1.0000	2952.0800
I NEED CHIPOTLE	1.0 - 1.0 PZAs	1.0000	3325.1050
BALSAMICO OSCURO MORAS NEGRAS Y	1.0 - 1.0 PZAs	1.0000	2952.0800
SALAMANCA HD	2.0 - 2.0 PZAs	2.0000	3350.8350
SELVA NEGRA HD	0.575 - 3.01 Kg	1.7912	3449.8575
EMBUCHADO HD	1.54 - 2.005 Kg	1.8500	3452.7467
PASTRAMI HD	2.98 - 3.015 Kg	3.0000	3420.9033
JAMON SERRANO HD	0.5 - 4.645 Kg	2.5725	3349.4050
SALAMANCA HD	1.505 - 1.505 Kg	1.5050	3440.3700

Cliente tipo 3

Las principales características del cliente 3 son consumo promedio elevado (arriba de 2000 pesos), presencia de aderezos y vinagres junto a embutidos ofertados por gramaje. De este último tipo de productos un consumo promedio alto (arriba del kg).

	Rango Consumo	Cantidad Consumida	Gasto Promedio
Nombre Producto			
SALAMANCA HD	2.0 - 2.0 PZAs	2.000	4188.77
EMBUCHADO HD	2.5 - 2.5 Kg	2.500	4188.77
JAMON SERRANO HD	3.01 - 3.01 Kg	3.010	4188.77
PASTRAMI HD	1.015 - 1.015 Kg	1.015	4188.77
ROAST BEEF/CORNEB BEEF HD	2.005 - 2.005 Kg	2.005	4188.77
SALAMI HD	2.505 - 2.505 Kg	2.505	4188.77

Cliente tipo 4

	Rango Consumo	Cantidad Consumida	Gasto Promedio
Nombre Producto			
BOCATA	15.0 - 15.0 PZAs	15.000	750.00
BOCATA CHICA	16.0 - 16.0 PZAs	16.000	880.00
BOCATA DE PECHUGA DE PAYO	15.0 - 15.0 PZAs	15.000	1050.00
BROCHETA	16.0 - 16.0 PZAs	16.000	778.32
CHAPATA	18.0 - 18.0 PZAs	18.000	809.18
CROISSANT CHICO	19.0 - 19.0 PZAs	19.000	845.07
QUESO GOUDA CHILENO	0.49 - 0.76 Kg	0.625	793.75
LOMO CANADIENSE	0.305 - 0.305 Kg	0.305	845.07

Cliente tipo 5

	Rango Consumo	Cantidad Consumida	Gasto Promedio
Nombre Producto			
CHAROLA DUSSELDORF CHICA	7.0 - 7.0 PZAs	7.00	2940.00
CHAROLA ESPECIAL	6.0 - 6.0 PZAs	6.00	2820.00
CHAROLA PERSONALIZADA	7.0 - 7.0 PZAs	7.00	2956.00
DISPENSADOR PARA BOTELLAS OLIO FINO	8.0 - 8.0 PZAs	8.00	3004.16
SELVA NEGRA HD	0.36 - 0.36 Kg	0.36	3004.16

Cliente tipo 6

	Rango Consumo	Cantidad Consumida	Gasto Promedio
Nombre Producto			
CAJA DE GALLETAS 1/2 BISCOTTI	3.0 - 3.0 PZAs	3.0	4085.0
CAJA DE GALLETAS 1/4 BISCOTTI	1.0 - 1.0 PZAs	1.0	4085.0
COYOTAS JAMONCILLO CAJETA	6.0 - 6.0 PZAs	6.0	4085.0
COYOTAS JAMONCILLO CON ALMENDRA	5.0 - 5.0 PZAs	5.0	4085.0
COYOTAS JAMONCILLO DE NUEZ	3.0 - 3.0 PZAs	3.0	4085.0
TORTILLA MAIZ MANO CHICA	2.0 - 2.0 PZAs	2.0	4085.0
WUAI PATATAS LEMON PEPPER 200 GRS	1.0 - 1.0 PZAs	1.0	4085.0

Cliente tipo 7

	Rango Consumo	Cantidad Consumida	Gasto Promedio
Nombre Producto			
SALCHICHA BLANCA HD	8.59 - 10.66 Kg	9.9133	1125.7278

Cliente tipo 8

	Rango Consumo	Cantidad Consumida	Gasto Promedio
Nombre Producto			
ROAST BEEF/CORNEB BEEF HD	4.505 - 5.985 Kg	5.3317	2527.91
JAMON SERRANO HD	5.5 - 5.5 Kg	5.5000	2558.87

Cliente tipo 9

	Rango Consumo	Cantidad Consumida	Gasto Promedio
Nombre Producto			
BAGUETTE 7 SEMILLAS MASA MADRE	14.0 - 14.0 PZAs	14.000	1197.34
BAGUETTE SUAVE MASA MADRE	14.0 - 14.0 PZAs	14.000	1197.34
BOLSA AL VACÍO	14.0 - 14.0 PZAs	14.000	1269.45
BOMBAS T&E	14.0 - 14.0 PZAs	14.000	1371.50
JAMON SERRANO HD	1.005 - 1.005 Kg	1.005	1197.34
QUESO MAASDAM HOLANDES T. KRON	0.35 - 0.35 Kg	0.350	1269.45

Ciente tipo 10

	Rango Consumo	Cantidad Consumida	Gasto Promedio
Nombre Producto			
BOCATA PANELA - 1/3	10.0 - 10.0 PZAs	10.000	2350.00
BOCATA PAVO 1/3	10.0 - 10.0 PZAs	10.000	2350.00
BOCATA ROAST BEEF - MEDIO	10.0 - 10.0 PZAs	10.000	2568.00
BOCATA ROAST BEEF Y SELVA NEGRA - 1/3	10.0 - 10.0 PZAs	10.000	2350.00
BOCATA SELVA NEGRA - MEDIO	10.0 - 10.0 PZAs	10.000	2568.00
BOCATA SERRANO - 1/3	10.0 - 10.0 PZAs	10.000	2350.00
BAGUETTE 7 SEMILLAS MASA MADRE	14.0 - 14.0 PZAs	14.000	1197.34
BAGUETTE SUAVE MASA MADRE	14.0 - 14.0 PZAs	14.000	1197.34
BOLSA AL VACÍO	14.0 - 14.0 PZAs	14.000	1269.45
BOMBAS T&E	14.0 - 14.0 PZAs	14.000	1371.50
JAMON SERRANO HD	1.005 - 1.005 Kg	1.005	1197.34
QUESO MAASDAM HOLANDES T. KRON	0.35 - 0.35 Kg	0.350	1269.45

Ciente tipo 11

	Rango Consumo	Cantidad Consumida	Gasto Promedio
Nombre Producto			
COCA-COLA REGULAR 500 ML	1.0 - 1.0 PZAs	1.000	4701.63
OBLEAS TIACHI BISCOTTI	1.0 - 1.0 PZAs	1.000	4701.63
SALAMANCA HD	3.0 - 3.0 PZAs	3.000	4701.63
SALAMI HD	3.0 - 3.0 PZAs	3.000	4701.63
CHORIZO PAMPLONA HD	2.01 - 2.01 Kg	2.010	4701.63
EMBUCHADO HD	2.03 - 2.03 Kg	2.030	4701.63
PASTRAMI HD	2.005 - 2.005 Kg	2.005	4701.63
ROAST BEEF/CORNEO BEEF HD	2.61 - 2.61 Kg	2.610	4701.63

Cliente tipo 12

Nombre Producto	Rango Consumo	Cantidad Consumida	Gasto Promedio
BOCATA PERSONALIZADA	22.0 - 32.0 PZAs	27.0000	1920.00
BOCATA	10.0 - 18.0 PZAs	14.0000	1960.00
BOCATA PECHUGA PAVO - MEDIO	81.0 - 81.0 PZAs	81.0000	5265.00
GALLETA LOTUS BISCOFF 250 GR	34.0 - 34.0 PZAs	34.0000	4085.00
ROSA DE CHOCOLATE	18.0 - 18.0 PZAs	18.0000	3367.00
ROAST BEEF/CORNEB BEEF HD	4.52 - 6.985 Kg	5.2700	2580.87
JAMON SERRANO HD	3.5 - 4.025 Kg	3.7625	2962.26
CHORIZO ESPAÑOL HD	2.04 - 2.04 Kg	2.0400	1895.16
JAMON VIRGINIA HD	7.435 - 7.435 Kg	7.4350	1524.18
LOMO CANADIENSE HD	0.305 - 0.305 Kg	0.3050	1168.19

Tipos de clientes finales

Tal como se mostró en las tablas anteriores, si bien el cluster reconoce 12 grupos distintos los grupos que en realidad nos interesan son:

- Cliente 1. Tipo retail
- Cliente 4. Enfocado a comida preparada
- Cliente 2. Retail sofisticado
- Cliente 3 y 5. Tipo de charolas
- Cliente 10. Tipo Mayorista

En vista lo mencionado anteriormente, se puede visualizar que la mayoría de los datos quedan en el grupo asignado al cliente 1 por lo que se realizará un nuevo clustering para el sector retail solamente, esto con el fin de identificar más patrones de compra.

1.6. Valoración de productos, resultados e impactos

Resegmentación de clientes tipo retail

Un problema que se identificó del algoritmo del clustering realizado es que la concentración en un solo perfil de clientes es considerablemente densa abarcando más del 95% del total de productos vendidos por folio por parte de Herman Delikatessen. Al observar las características propias de esta agrupación pudimos percatarnos que los patrones de compra en estos folios tenían un perfil retail o de consumo bajo.

Contrastando este resultado con los análisis exploratorios previamente realizados se muestra que en realidad esta es una situación bastante regular para la tienda, ya que en la mayoría de los folios los consumos tanto en cantidad como en términos monetarios no presentan un patrón de compra elevado. Otra cuestión que nos llamó bastante la atención es que dentro de este mismo grupo (al ser tan denso) cuenta con una gama de productos muy extensa tanto en el esquema unitario como en el esquema por gramaje, es por eso que consideramos pueden llegar a existir diversos patrones de compra incluso dentro de este mismo grupo.

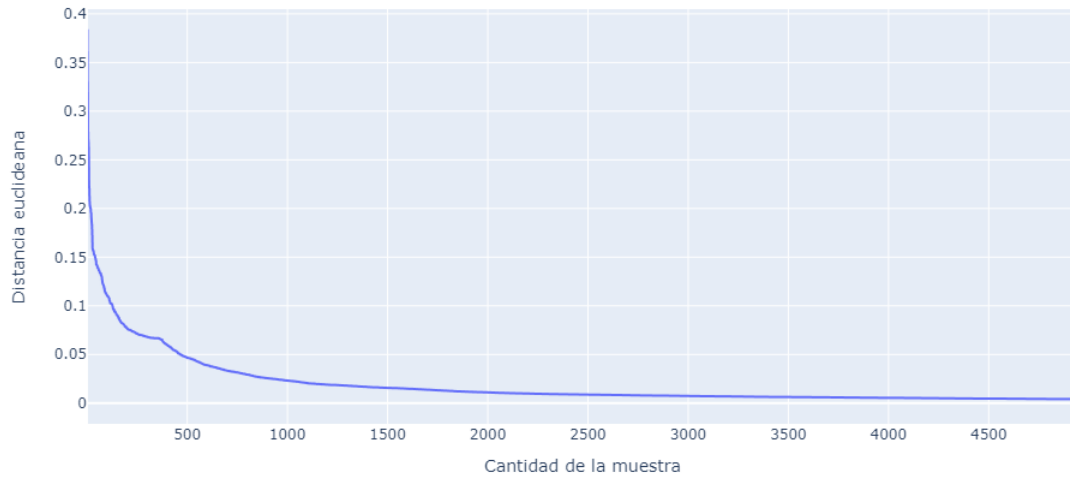
	Productos totales de Herman	Productos contenidos en cluster retail	Porcentaje representativo
Unitarios	1128	1123	0.995567
Gramaje	124	124	1.000000

De la gama de productos que encontramos en el cluster retail podemos observar que prácticamente todos los productos adquiridos en este tipo de compras corresponden al 100% de los productos que Herman Delikatessen contiene en su oferta total. Esta situación nos confirma que muy probablemente dentro de este grupo podremos encontrar nuevos patrones de compra, ya que es una muestra más que representativa de los clientes de la tienda. A continuación, se visualizará una gráfica de barras donde se mostrarán los productos más populares dentro del cluster retail, tanto para el esquema unitario como para el esquema por gramaje.



De las gráficas descritas anteriormente podemos concluir que los productos más populares para esta segmentación de clientes corresponden en general a los productos más populares en general para toda la tienda. Con esto en cuenta podemos concluir que en efecto la necesidad de hacer una nueva re-segmentación para este cluster es evidente, ya que dentro de aquí podremos definir nuevos patrones de consumo por parte de los clientes de Herman Delikatessen.

Epsilon óptimo para segmentación retail



Para esta nueva re-segmentación del sector de clientes *retail* de **Herman Delikatessen** podemos observar que el algoritmo del **DBSCAN** fue capaz de detectar 1212 nuevos grupos dentro del sector de consumo minorista de la tienda (el cual recordemos representa la mayoría de la estructura de ventas de la tienda). Si bien consideramos que tener 1212 tipos de clientes tipo *retail* puede sonar exagerado, a falta de un análisis particular a cada uno de ellos no podríamos afirmar si son muchos o no.

% of total data	
Cluster	
0	0.512211
3	0.371549
1	0.074401
2	0.015280
-1	0.012597
4	0.005958
5	0.002274
11	0.001683
6	0.001319
8	0.000682
12	0.000682
7	0.000455
9	0.000455
10	0.000455

Como podemos observar en el resumen obtenido en cuanto a la cantidad de compras que contiene cada cluster, se percibe que solamente tres agrupaciones generadas contienen alrededor del 95% de los datos acumulados, ante esta situación podemos concluir que en realidad solamente se logran percibir 3 agrupaciones que aportan la mayoría de la información recolectada en este sector de clientes retail. Es por esto que solo extraeremos y parametrizaremos la información obtenida en estas agrupaciones, las cuales consideramos son las más relevantes.

Parametrización de clientes retail

Como se mencionó previamente solo tomaremos las agrupaciones tomadas por los 3 clusters más populares.

Cliente retail 1

Nombre Producto	Rango Consumo	Cantidad Consumida	Gasto Promedio
BAGUETTE SUAVE MASA MADRE	1.0 - 1.0 PZAs	1.0000	274.0676
COCA-COLA LIGHT 500 ML	1.0 - 1.0 PZAs	1.0000	86.4688
PAN PARA TAPAS	1.0 - 1.0 PZAs	1.0000	421.0508
CHIPS GDE COMALLI	1.0 - 1.0 PZAs	1.0000	231.2879
DELISCO BROWNIES MARIFER	1.0 - 1.0 PZAs	1.0000	143.4594
COCA-COLA SIN AZÚCAR 500 ML	1.0 - 1.0 PZAs	1.0000	96.7040
JAMON SERRANO HD	0.035 - 1.5 Kg	0.2207	448.1517
QUESO GOUDA CHILENO	0.055 - 1.055 Kg	0.2784	444.5934
JAMON YORK HD	0.07 - 1.015 Kg	0.3066	350.4496
SELVA NEGRA HD	0.05 - 1.005 Kg	0.2038	455.2768
ROAST BEEF/CORNEED BEEF HD	0.065 - 1.095 Kg	0.2655	493.3161

Ciente retail 2

Nombre Producto	Rango Consumo	Cantidad Consumida	Gasto Promedio
SELVA NEGRA HD	0.02 - 2.27 Kg	0.2364	307.0447
JAMON SERRANO HD	0.02 - 2.01 Kg	0.2301	348.3741
JAMON YORK HD	0.055 - 2.55 Kg	0.3752	235.4185
LOMO CANADIENSE HD	0.02 - 2.005 Kg	0.2076	273.4431
SALAMI HD	0.02 - 1.665 Kg	0.1871	346.0753
ROAST BEEF/CORNEB BEEF HD	0.05 - 2.6 Kg	0.3066	358.0492
CHORIZO PAMPLONA HD	0.02 - 1.5 Kg	0.1658	335.6534
JAMON SERRANO	0.03 - 2.02 Kg	0.2214	330.6012
SELVA NEGRA	0.035 - 2.2 Kg	0.2060	305.4923
QUESO GOUDA CHILENO	0.04 - 1.5 Kg	0.2699	276.5913
JAMON YORK	0.1 - 1.265 Kg	0.3270	211.9231
JAMÓN PAVO (CUADRADO)	0.045 - 1.52 Kg	0.3495	273.0296
QUESO GOUDA HOLANDES	0.06 - 1.01 Kg	0.3345	279.8478
SALAMI	0.01 - 1.3794 Kg	0.1750	310.4181
LOMO CANADIENSE	0.055 - 0.67 Kg	0.1827	278.7171

Ciente retail 3

Nombre Producto	Rango Consumo	Cantidad Consumida	Gasto Promedio
BAGUETTE SUAVE MASA MADRE	2.0 - 2.0 PZAs	2.0000	387.4799
DELISCO BROWNIES MARIFER	2.0 - 2.0 PZAs	2.0000	97.6092
TORTILLAS MAIZ A MANO GRANDE	2.0 - 2.0 PZAs	2.0000	42.0308
PAN PARA TAPAS	2.0 - 2.0 PZAs	2.0000	679.7278
BAGUETTE SUAVE RÚSTICO 21	2.0 - 2.0 PZAs	2.0000	311.6820
OBLEAS TIACHI BISCOTTI	2.0 - 2.0 PZAs	2.0000	388.5979
PIZZAS CONGELADAS	2.0 - 2.0 PZAs	2.0000	587.8722
SELVA NEGRA HD	0.05 - 1.005 Kg	0.2243	485.0662
QUESO GOUDA CHILENO	0.105 - 0.8 Kg	0.2927	436.5119
ROAST BEEF/CORNEB BEEF HD	0.1 - 0.895 Kg	0.2587	658.2741
JAMON SERRANO HD	0.055 - 1.065 Kg	0.2673	576.3033
JAMON YORK HD	0.11 - 0.515 Kg	0.2450	333.3688

Conclusiones de la segmentación.

A manera de conclusión para la parte de segmentación de clientes de Herman Delikatessen podemos decir lo siguiente:

- Los resultados obtenidos por el DBSCAN son favorables y nos ayudaron a identificar patrones de compra dentro de la estructura de clientes de Herman Delikatessen.
- Se observa de manera sistemática que tanto los resultados obtenidos por el proceso del clustering como del análisis exploratorio arrojan que el target principal de Herman está orientado a la venta tipo retail con folios promedio que no superan los 500 pesos.
- Orientado al comentario anterior, dentro de la misma segmentación de clientes tipo retail, se pudo observar que existen diversos patrones para los mismos productos (siendo estos los más populares), ya que no todo tipo de cliente consume en cantidad promedio lo mismo.

Promoción: ¿Llevas la comida? Acompañala con postre o botana:

Clientes del grupo 1

En la compra de máximo 0-1kg de SELVA NEGRA HD, QUESO GOUDA CHILENO, JAMÓN SERRANO HD, ROAST BEEF/CORNED BEEF obtendrás un 15% de descuento en CHIPS GDE COMALI y DELISCO BROWNIES MAFER.

Clientes del grupo 2

(mayor cantidad consumida y más productos)

En la compra mayor a 1 kg de 'LOMO CANADIENSE', 'QUESO GOUDA HOLANDES', 'SELVA NEGRA', 'SELVA NEGRA HD', 'SALAMI', 'JAMON SERRANO', 'CHORIZO PAMPLONA HD', 'SALAMI HD', 'JAMON SERRANO HD', 'ROAST BEEF/CORNED BEEF HD' obtendrás un 30% de descuento en CHIPS GDE COMALI

1.7. Bibliografía y otros recursos

Garza García, J. d. I. (2013). Análisis estadístico multivariante: Un enfoque teórico-práctico.

México: McGraw-Hill.

Johnson, D. E. (2000). Métodos multivariados aplicados al análisis de datos. México:

Thomson.

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*,

6(60), 3021. <https://doi.org/10.21105/joss.03021>

2. Productos

Como entregable para esta fase del PAP se tiene una liga a repositorio en GitHub en el cual se encuentra la evidencia de todo lo planteado anteriormente. Está estructurado en 3 secciones, estadística descriptiva, análisis de compras y ventas y segmentación de clientes.

En términos de entregables para la empresa, en este periodo no se entregará ningún recurso, ya que se espera continuar trabajando con los resultados obtenidos para presentar un material útil para el personal de Herman Delikatessen.

Link del repositorio: <https://github.com/ITESO-PAP4J09-Ciencia-de-datos/herman-2022>

3. Reflexión crítica y ética de la experiencia

El RPAP tiene también como propósito documentar la reflexión sobre los aprendizajes en sus múltiples dimensiones, las implicaciones éticas y los aportes sociales del proyecto para compartir una comprensión crítica y amplia de las problemáticas en las que se intervino.

3.1. Sensibilización ante las realidades

Al trabajar en este proyecto con la tienda Herman Delikatessen, se logró contextualizar todo lo que conlleva el manejo y sustento de una franquicia. Si bien nuestro análisis se concentra en inferencias a partir de bases de datos, el interactuar con el personal de la tienda como un primer contacto ayudó a dirigir la línea temporal de este proyecto.

Su modelo de negocios se apega a lo que se considera una empresa ética ya que busca ofrecer lo mejor en productos y servicios a sus clientes con precios justos. Se sabe que llevar las riendas de un negocio no es nada sencillo, no obstante, Herman Delikatessen ha hecho un trabajo excepcional.

3.2. Aprendizajes logrados

Dentro de los aprendizajes obtenidos destacamos dos categorías principales, profesionales y social.

Para la categoría de aprendizaje profesional rescatamos lo siguiente:

- Se desarrollaron habilidades nuevas de programación en distintos lenguajes como serían Python y R.
- Se aplicaron técnicas de Inteligencia Artificial vistas en clases, las cuáles se reforzaron.
- Las habilidades de investigación y trabajo independiente mejoraron.
- El análisis se volvió más crítico por parte de los participantes.
- La capacidad de resolución de problemas se desenvolvió a lo largo del proyecto.
- Se encontró un aterrizaje entre los aprendizajes previos y la puesta en práctica de estos.

En términos de aprendizajes sociales se tiene que:

- La comunicación entre los participantes del equipo evolucionó hasta llegar a un método eficiente de concentrado de ideas.
- Se desarrolló la disposición a escuchar distintos puntos de vista sobre una misma problemática.
- Se propició un ambiente seguro que permitió generar ideas que llevaron a conclusiones del proyecto.