
MINERIA DE DATOS

REGLAS DE ASOCIACION

UNIVERSIDAD TECNOLOGICA NACIONAL
FACULTAD REGIONAL ROSARIO

MARCO TEORICO Y RESOLUCION

DJEMDJEMIAN CID, EZEQUIEL



2020

Índice general

I	Introduccion	3
1.	Introduccion	4
1.1.	Introduccion	4
2.	Marco Teorico	5
2.1.	Marco Teorico	5
3.	Apriori	7
II	Aplicacion	8
4.	Aplicacion	9
4.1.	Introduccion	9
5.	Kaggle Market Basket	10
5.1.	Introduccion	10
5.2.	Explicacion de los Datos	10
5.3.	Analisis Exploratorio	11
5.4.	Generacion de Reglas	19
III	Clustering - Kaggle Market Basket	22
6.	Motivacion	23
6.1.	Introduccion	23
6.2.	Vista Minable	23
7.	Clustering Jerarquico	24
7.1.	Herramienta: IBM SPSS Statistics	24
7.2.	Resultados	25
8.	Clustering No Jerarquico - K-Means	32
8.1.	Herramienta: IBM SPSS Statistics	32
8.2.	Resultados	33
9.	Clustering - Seleccion de Modelos	39
9.1.	Conclusion	39

10.Clustering - Caracterizacion	40
10.1. N: Cantidad de Clientes por Cluster	40
10.2. P: Promedio Cantidad de Productos Vendidos por Mes	40
 IV Reglas de Asociacion Dirigidas	 42
11.Reglas Dirigidas	43
11.1. Reglas Generadas	43
12.Reglas Dirigidas sobre Categorias y Pasillos	47
12.1. Reglas por Categoria	48
12.2. Reglas por Pasillo	51
 V Conclusiones Finales	 55
13.Conclusiones Finales	56
13.1. Conclusiones	56
14.Versionado	57

Parte I

Introduccion

Capítulo 1

Introduccion

1.1. Introduccion

En este documento se introducira a una de las tareas de la Minería de Datos: *Reglas de Asociacion*.

El desarrollo estara dividido en etapas:

- Marco Teorico
 - Mettricas
 - Algoritmos
- Algoritmo Apriori
- Aplicacion
 - Analisis Exploratorio
 - Generacion de Reglas
 - Conclusiones

Capítulo 2

Marco Teorico

2.1. Marco Teorico

Las *Reglas de Asociacion* expresan patrones de comportamiento entre los datos, en funcion de la aparicion conjunta de valores. Por ejemplo, en un supermercado podemos conocer que productos suelen comprarse conjuntamente, para asi poder mejorar la distribucion de los productos, generar ofertas especificas, etc.

Una Regla de Asociacion es una proporcion probabilistica sobre la ocurrencia de ciertos estados.

Un ejemplo tipico de una Regla de Asociacion para un dominio de Ventas en un Supermercado seria de la forma:

SI pollo **Y** crema de leche **ENTONCES** champiñones

Donde *pollo* y *crema de leche* seria el **predecesor** de la regla, mientras que *champiñones* es el **consecuente**.

2.1.1. Metricas

Sea T el numero total de transacciones, entonces, se definen,

- Soporte de un item: Tambien denominado *cobertura*, es el numero de transacciones que contienen a ese item, sobre el total de transacciones. Por ejemplo:

$$S(pollo) = \frac{\sigma(pollo)}{T}$$

- Soporte de la Regla: Es el numero de transacciones en las que se cumple el predecesor y el consecuente, es decir, es la proporcion de casos en los que se aplica la regla.

Para el ejemplo anterior:

$$S(pollo, crema \rightarrow champiniones) = \frac{\sigma(pollo, crema, champiniones)}{T}$$

- **Confianza:** Tambien denominada *precision*, es el porcentaje de veces que la regla se cumple cuando se puede aplicar. Es decir, mide la cantidad de veces en las que se cumple el *consecuente*, cuando se cumple el *predecesor*.

$$C(\text{pollo, cema} \rightarrow \text{champiniones}) = \frac{\sigma(\text{pollo,crema,champiniones})}{\sigma(\text{pollo,crema})} = \frac{S(\text{pollo,crema})}{S(\text{champiniones})}$$

- **Lift:** El Lift de una Regla es la confianza de la regla, dividido por la confianza esperada, asumiendo que los items son independientes.

Sea $X = (\text{pollo, crema})$ e $Y = (\text{champiniones})$, entonces

$$Lift = P(X \rightarrow Y) = \frac{P(X, Y)}{P(X) \cdot P(Y)}$$

Análisis sobre el Lift:

- $Lift < 1$: La regla se cumple una cantidad de veces menor a lo esperado (bajo condiciones de independencia)
- $Lift = 1$: La regla se cumple una cantidad de veces igual a lo esperado (bajo condiciones de independencia)
- $Lift > 1$: La regla se cumple una cantidad de veces mayor a lo esperado (bajo condiciones de independencia) \rightarrow se puede intuir que existe una *relacion* que produzca que los productos ocurran conjuntamente.

2.1.2. Algoritmos

Entre los algoritmos mas relevantes se pueden destacar:

- Apriori
- Eclat

Capítulo 3

Apriori

Parte II

Aplicacion

Capítulo 4

Aplicacion

4.1. Introduccion

A lo largo de los siguientes capitulos se presentaran distintas soluciones, sobre distintos conjuntos de datos, utilizando Python como herramienta de programacion para la resolucion de estos problemas. Se detallara en cada caso, las bibliotecas y funciones utilizadas, como asi tambien, se facilitara el repositorio con los data set y el codigo pertinente.

Previo a comenzar con la generacion de reglas, se estudiara el conjunto de datos, realizando un **Analisis Exploratorio**, con el fin de determinar si se puede encontrar algun patron antes de comenzar con el proceso concreto de Minería.

Es importante destacar, que durante todo el proceso de Analisis se realizaran supuestos, los cuales valdran desde que son enunciados hasta el final del analisis, salvo que se exprese lo contrario.

Capítulo 5

Kaggle Market Basket

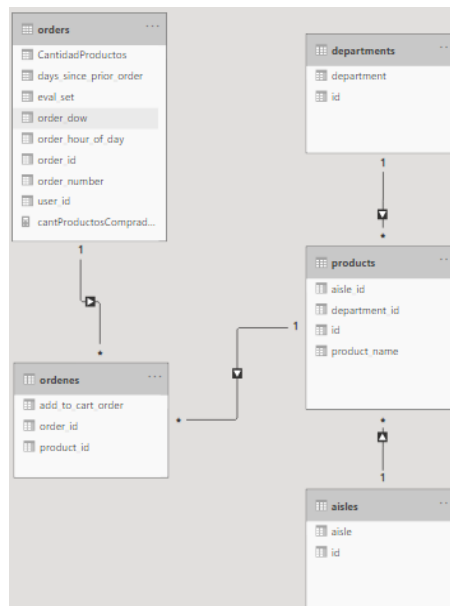
5.1. Introduccion

En esta seccion se utilizara el data set ofrecido por la plataforma Kaggle, el cual se puede visualizar [aqui](#).

Ademas, sobre el conjunto de datos se confecciono un informe, con el fin de facilitar la fase de *Analisis de Datos*. El mismo se desarrollo utilizando la tecnologia Power BI, y se puede descargar desde [aqui](#).

5.2. Explicacion de los Datos

El data set utilizado se encuentra estructurado de la siguiente forma:



5.3. Analisis Exploratorio

5.3.1. Tamano de Datos

- Ventas
 - #Ventas: 3.35 millones
 - #Productos: 33.82 millones
 - #Productos por Venta: 10.11 (Promedio)
- Productos
 - #Productos: 49.690
 - #Categorias: 21
 - #Pasillos: 134
- Clientes
 - #Clientes: 206.210

5.3.2. Variable Horario

De cada Venta realizada se conoce la hora de la misma.

Analisis

- Nombre: `hour_of_day`
- Type: `int`
- Missing Values: 0
- Dominio: {0,1, 2, ..., 23}

Un histograma entre la Cantidad de Ventas y su horario:



Estadisticos

- Media: 13,45 *horas*
- Mediana: 13 *horas*
- Desviacion: 4,23 *horas*
- Varianza: 17,86 *horas*²

Analisis

A priori el establecimiento presenta flujo de ventas durante las 24 horas del dia. Se desconoce si realmente el establecimiento permanece abierto 24 horas al dia, si los *horarios anormales* se deben a operaciones de e commerce o si se debe a dias festivos o excepcionales de atencion. Se analizara cuanto varia el flujo de ventas, si se consideraran las transacciones ocurridas en un *horario normal*.

- El establecimiento permanece abierto las 24 horas?
- Las ventas fueras del *horario normal* son casos relaes? (Promociones, fechas festivas, e commerce, etc)

Concentracion de Datos excluyendo colas horarias

Se analizara la cantidad de **Ventas** que se concentra dentro de los *horarios normales* de un establecimiento del tipo Hipermercado:

- De 07:00 a 20:00 → 91,60 %
- De 08:00 a 20:00 → 88,85 %
- De 08:00 a 19:00 → 85,83 %
- De 08:00 a 18:00 → 81,96 %

Conclusion sobre la variable Horario

Sin mucha mas informacion ni analisis se consideraran las transacciones ocurridas entre las 07 y 20 hs, concentrando al 91,60 % de las mismas. De esta forma, los estadisticos resultarian:



Como conclusion podemos establecer que:

- Mayor concentracion alrededor de la media $\overline{HoradeCompra}$.
- Al reducir las horas donde menos transacciones se registran, el numero de transacciones por hora aumenta.

Supuesto I

El negocio opera normalmente entre los horarios de 07:00 a 20:00, desestimandose las transacciones ocurridas fuera de dicho rango horario.

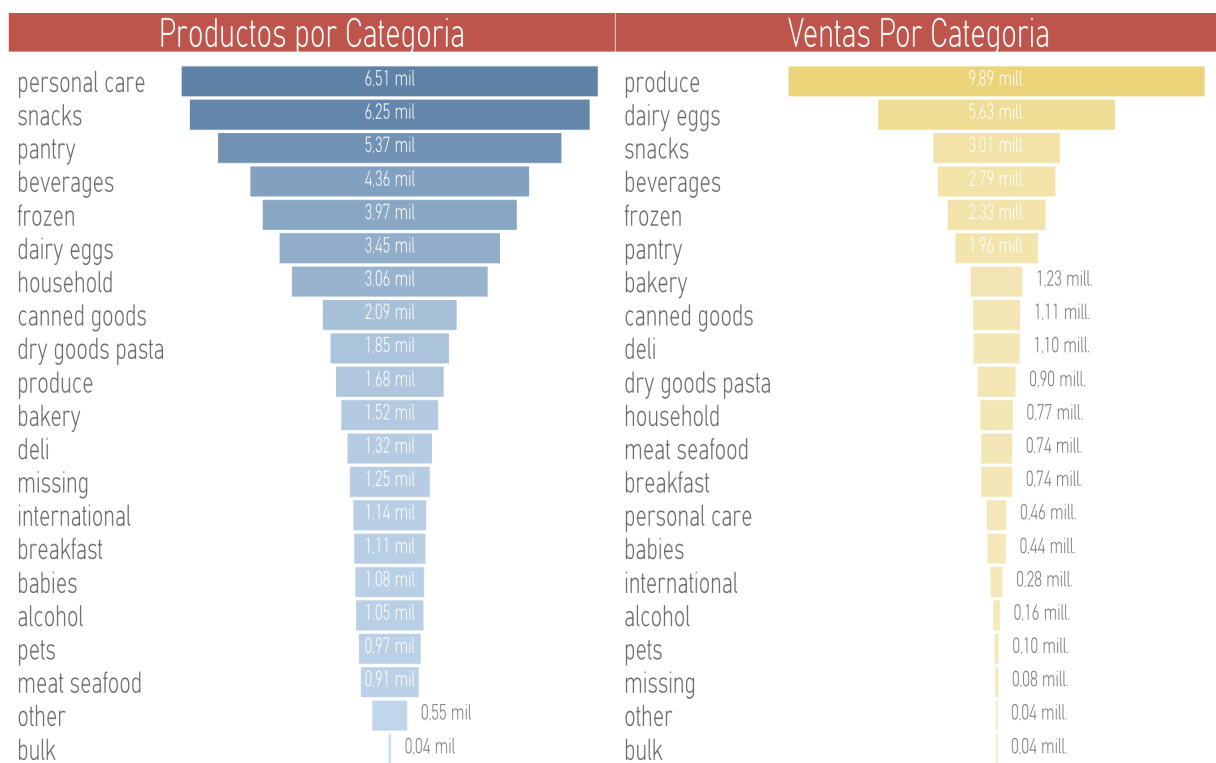
5.3.3. Variable Categoria

Cada producto pertenece a una y solo una categoria.

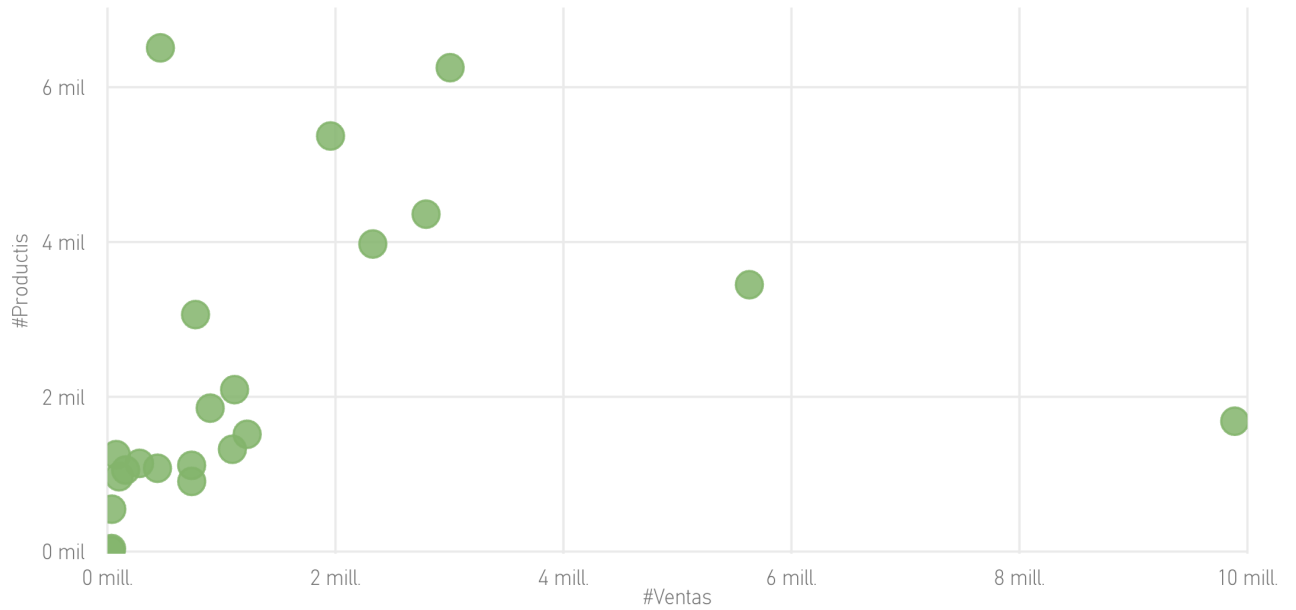
Analisis

- Nombre: department
- Type: *varchar*
- Missing Values: 0
- Dominio: {'fronen', 'other', 'bakery', ...}
- $|Dominio| = 21$
- Categoria 'Missing': Categoria que agrupa aquellos productos que no han sido Clasificados.

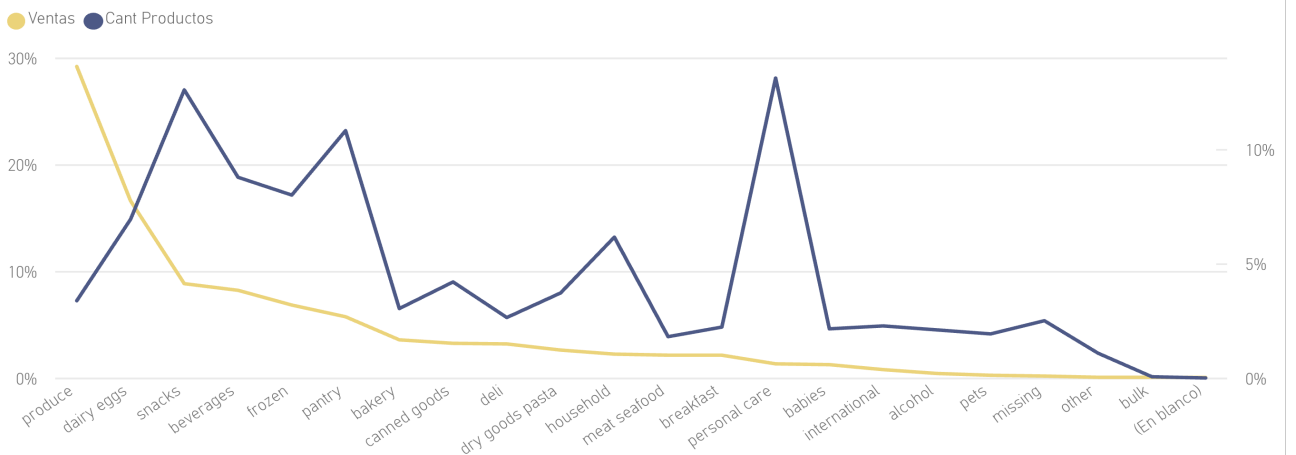
Se presentaran algunas graficas de interes, para luego analizar la informacion.



Productos y Ventas por Categoría



%Productos y %Ventas por Categoría



Conclusion sobre la variable Categoría

Teniendo en cuenta que el objetivo es establecer *Reglas* que permitan describir patrones, a priori, nos interesan aquellas categorías cuya relación $\frac{\#Ventas}{\#Productos}$ sea máxima.

- Analizando la figura *%Productos y %Ventas por Categoría*, podemos concluir que la Categoría **produce** es la que maximiza dicha relación, acompañada únicamente de **dairy eggs**.
- Lo contrario ocurre con **personal care**, que registra una gran cantidad de Productos, y muy pocas Ventas.

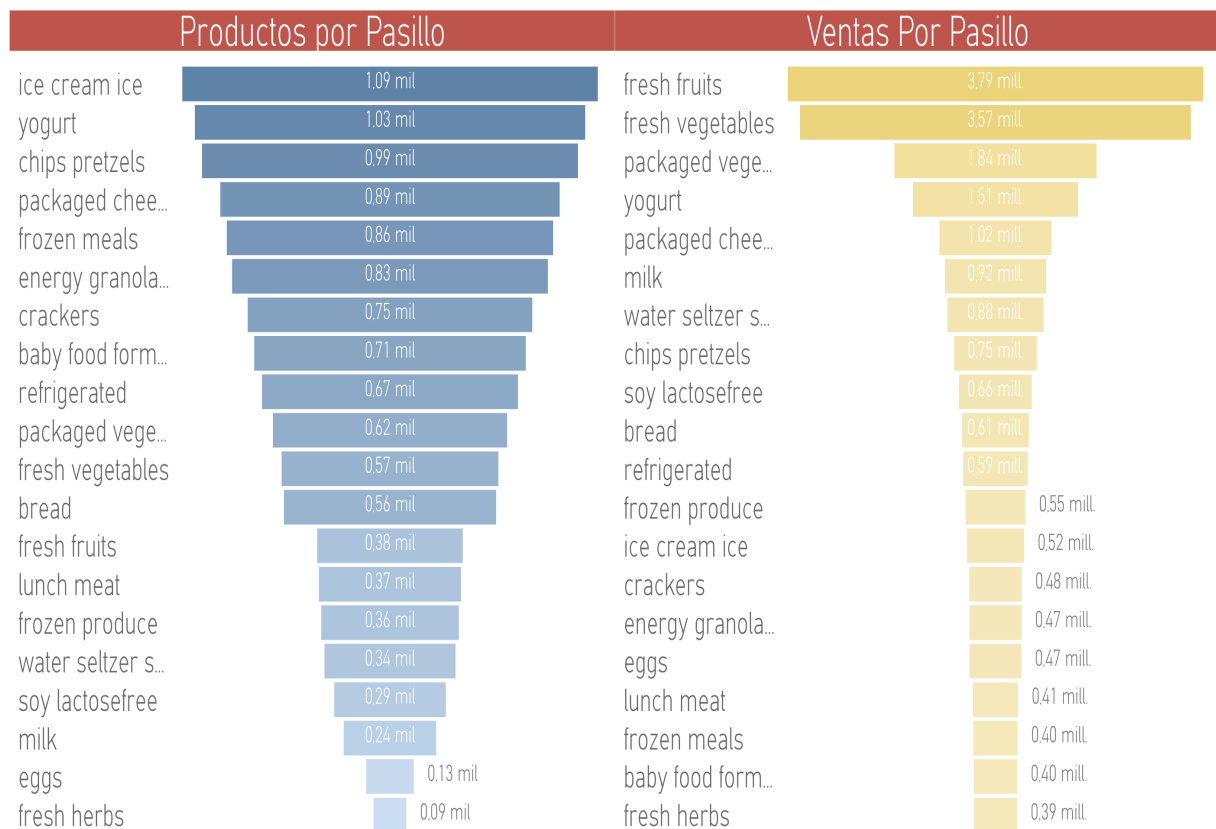
5.3.4. Variable Pasillo

Cada producto se encuentra en un unico pasillo.

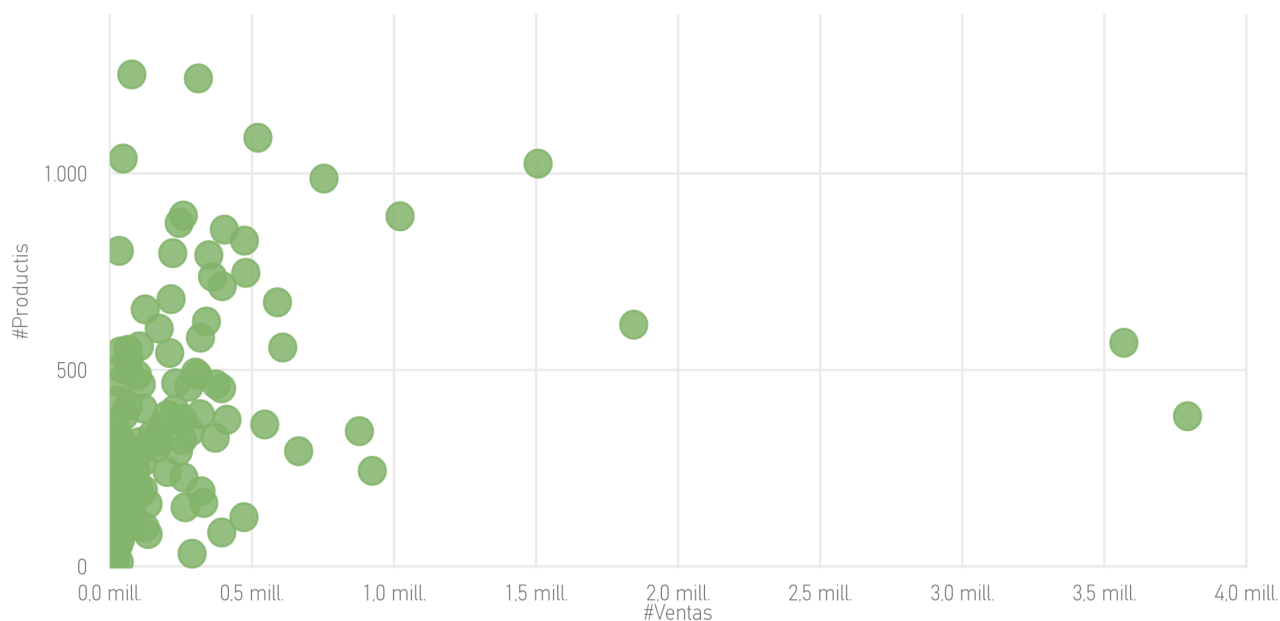
Analisis

- Nombre: aisle
- Type: *varchar*
- Missing Values: 0
- Dominio: { 'frozen meal', 'yogurt', 'bread', ... }
- $| Dominio | = 134$
- Categoria 'Missing': Categoria que agrupa aquellos productos que no han sido ubicados en ningun Pasillo.

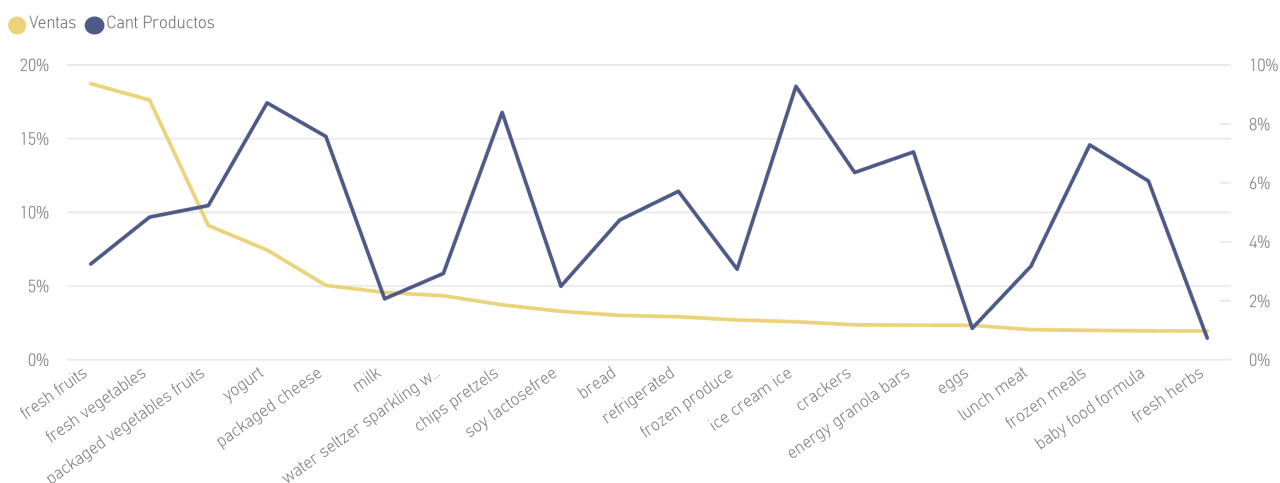
Se presentaran algunas graficas de interes, para luego analizar la informacion.



Productos y Ventas por Pasillo



%Productos y %Ventas por Pasillo



Conclusion sobre la variable Pasillo

Teniendo en cuenta que el objetivo es establecer *Reglas* que permitan describir patrones, a priori, nos interesan aquellas categorías cuya relación $\frac{\#Ventas}{\#Productos}$ sea máxima.

- Analizando la figura *%Productos y %Ventas por Pasillo*, podemos concluir que la Categoría **fresh fruits** es la que maximiza dicha relación, acompañada en una menor medida de **fresh vegetables**.
- Lo contrario ocurre con, por ejemplo, **ice cream ice** o **frozen meal**, entre otras, que registran una gran cantidad de Productos, y muy pocas Ventas.

5.3.5. Conclusiones Generales

- Dentro del horario 07:00 a 20:00 se concentra al 92 % de las transacciones.
- Cada cesta contiene, en promedio, 10 articulos.
- Cada cesta contiene, como mediana, 8 articulos.
- Top 3 de Categorías con mayor soporte:
 - produce
 - dairy eggs
 - snack
- Difícilmente existan productos un alto soporte dentro de las categorías:
 - personal care
 - household
 - pantry
- Top 3 de Pasillos con mayor soporte
 - fresh fruits
 - fresh vegetables
 - packaged vegetables
- Difícilmente existan productos un alto soporte dentro en los pasillos:
 - ice cream ice
 - frozen meals
 - energy granola bars

5.4. Generacion de Reglas

Para la Generacion de Reglas se utilizaron los algoritmos de *apriori* y *association-rules*, de la libreria **mlxtend**.

El repositorio para esta solucion → [aqui](#).

Vista Mitable

El conjunto de datos se genero con el siguiente query: [ver aqui](#). Las variables explicativas de dicho modelo son:

*Order_ID, User_ID, Order_Number, Order_Dow, Order_Hour_of_Day,
Days_since_prior_Order, Product_ID, Product_Name, Aisle_ID, Aisle, Department_ID,
Department, Quantity*

Parametros de Ejecucion

El algoritmo de generacion de reglas, requiere como entrada, el soporte minimo, que deben satisfacer los productos que conformen las Reglas. Por lo que, un soporte minimo alto, implica que mayor cantidad de productos sean candidatos a conformar la regla, pero como consecuencia, aumenta el tiempo computacional (evalua complejidad). Pruebas sobre esto se realizaran en apartados posteriores.

- Soporte Minimo = $0.015 = 1.5\%$
- Lift Minimo = 1

5.4.1. Reglas Generadas

Se generaron un total de 12 Reglas de Asociacion. Solo se enumeraran las tres mas relevantes.

- Regla I
 - Antecedente: Bag of Organic Bananas
 - Concecuyente: Organic Strawberries
 - Soporte Antecedente: 0.08213
 - Soporte Concecuente: 0.1169
 - Soporte regla: 0.0229
 - Confianza: 0.19647
 - Lift: 2.398
 - Leverage: 0.0131
 - Conviction: 1.222
- Regla II
 - Antecedente: Large Lemon
 - Concecuyente: Banana
 - Soporte Antecedente: 0.06164
 - Soporte Concecuente: 0.014288
 - Soporte regla: 0.01631
 - Confianza: 0.2645
 - Lift: 1.8525
 - Leverage: 0.00759
 - Conviction: 1.1655

- Regla III

Antecedente: Banana

Concecuente: Organic Avocado

Soporte Antecedente: 0.1428

Soporte Concecuente: 0.0566327

Soporte regla: 0.0168329

Confianza: 0.117859

Lift: 2.08111

Leverage: 0.0087

Conviction: 1.0694

5.4.2. Que nos dicen todas estas metricas?

Analisis sobre Regla I:

- Existe una relacion entre *Organic Strawberries Bag of Organic Bananas*.
- El 8.23 % de las cestas contienen *Organic Strawberries*.
- El 11.70 % de las cestas contiene *Bag of Organic Bananas*.
- El 2.28 % de las cestas contiene *Organic Strawberries* y *Bag of Organic Bananas* **simultaneamente**.
- Si una cesta contiene *Organic Strawberries* \Rightarrow contenera, tambien, *Bag of Organic Bananas* con una esperanza de 27.78 %.

5.4.3. Tiempos Computacionales

ID	Soporte	#Reglas	Lectura CSV	Limpieza	Agrupam	Items	Reglas	Total
1	1 %	32	14,231s	0,238s	676,194s	189,063s	0,0146	879,138s
2	1,5 %	14	14,288s	0,238s	631,036s	165,553s	0,018	811,921s
3	2,0 %	2	14,639s	0,249s	672,261s	153,677s	0,011	840,822s

Cuadro 5.1: Tiempos de computo

Parte III

Clustering - Kaggle Market Basket

Capítulo 6

Motivacion

6.1. Introduccion

Supongamos que el objetivo de negocio que intenta abordar este desarrollo es

Implementar una estrategia de marketing, que permita incentivar el consumo de nuestros clientes, para ello, propondremos emitir, para cada compra, ciertos tickets que ofrezcan algun beneficio/descuento, en ciertos productos, a priori determinados por el negocio, en cierto rango horario.

A continuacion se presentaran varios resultados de aplicar distintos algoritmos de Clustering, con distintas herramientas. El objetivo de este analisis, sera analizar las reglas de asociacion que se desprenden de cada Cluster resultante, con el objetivo de hacer una campana aun mas especifica y dirigida.

6.2. Vista Minalbe

El conjunto de datos se genero con el siguiente query: [ver aqui](#). La variables explicativas de dicho modelo son:

User_ID, AVG_Hour_of_Day, AVG_Days_since_prior_Order, Cantidad_Ventas, Cantidad_Productos, Cantidad_Babies, Cantidad_Pets, Cantidad_Produce, Cantidad_Pets, Cantidad_SnacksYAlcohol, Cantidad_DairyEggs

6.2.1. Tamano

El conjunto de datos cuenta con aproximadamente doscientas cinco mil clientes.

Capítulo 7

Clustering Jerarquico

Dado que todas las variables utilizadas son numericas, procederemos, en primer instancia, a realizar un Analisis de Clustering Jerarquico.

Como dicho algoritmo utiliza una matriz de distancias entre todas las instancias, es imposible ejecutar dicho algoritmo en un equipo hogareno, ya que deberiamos poder mantener en memoria una matriz, que si bien es simetrica, deberia contener $205,000^2$ celdas aproximadamente.

7.1. Herramienta: IBM SPSS Statistics

Para poder utilizar dicho algoritmo, se seleccionara el 5 % de datos de la muestra, lo que representan poco mas de *diez mil instancias*. Para esto, utilizaremos la herramienta de seleccion que ofrece la herramienta, asumiendo que dicho selector es *aleatoria, con distribucion uniforme*, dado que las instancias se encuentran ordenadas por numero de Cliente, cuyo ID es consecutivo creciente. Por lo que, aplicando una seleccion uniforme sobre estas, estariamos asegurandonos tomar casos desde el principio hasta el final, con la misma probabilidad.

7.1.1. Parametros

Rango de Soluciones

A un algoritmo de cluster jerarquico es necesario determinar cuantos clusters queres formar. En base a este parametro, la herramienta determinara en que iteracion cortar el dendograma.

Se realizaran analisis para dos, tres y cuatro clusters, analizando los estadisticos mas importantes para cada una de las variables explicativas, y finalmente se evaluaran los clusters.

Metodo de Agrupacion

Utilizaremos el *Metodo de Ward*, o tambien conocido como *Metodo de la varianza minima*.

Estandarizacion

Dado que el clustering jerarquico funciona con las distancias entre las instancias de las variables, se estandarizaran las variables, de forma tal que el modelo no se base mayoritariamente en las variables de mayor valor, es decir, en por ejemplo, *Cantidad_Productos*, desestimando otras cuyos valores tienden a cero, como por ejemplo *Cantidad_Babies*.

La estandarizacion que se utilizara, sera la que IBM denomina *Puntuacion Z*, que consiste en transformar todos los valores dentro del rango $[0,1]$, con la transformacion:

$$X'_i = \frac{X_i - \min(X)}{\max(X) - \min(X)}$$

Medida de Distancia

Utilizaremos la distancia euclideana al cuadrado:

$$D_{i,j} = \sqrt{\sum_{k=1}^n (X_{ki} - x_{kj})^2}$$

de esta forma, las distancias grandes seran aun mas grandes, y las distancias cortas seran aun mas cortas.

7.2. Resultados

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desviación estándar
AVG_Hour_Of_Day	10199	7,00000	20,00000	13,4209833	1,96543273
AVG_Days_Since_Prior_Order	10199	,0000000	30,0000000	12.99030332	5.901966241
Cantidad_Productos	10199	1	3459	143,16	189,821
Cantidad_Ventas	10199	1	99	14,23	15,404
Cantidad_Categoria_Babies	10199	,000000000	,666667000	.0090592806	.0370159222
Cantidad_Categoria_Pets	10199	,000000000	1,000000000	.0031193646	.0240312648
Cantidad_Categoria_SnacksYAlcohol	10199	,000000000	1,000000000	.0982374519	.1173719114
Cantidad_Categoria_Productos	10199	,000000000	1,000000000	.2803036807	.1857917879
Cantidad_Categoria_DairyEggs	10199	,000000000	1,000000000	.1557501636	.1099229479
N válido (por lista)	10199				

A continuacion, se desarrollara un analisis con graficas para cada cluster formado, haciendo especial enfasis en el agrupamiento de cuatro clusters.

Es importante destacar que estas graficas son unas pruebas estadisticas, ya que, representan las agrupaciones de los clusters en un Intervalo de Confianza, con un $\alpha = 0,95$

7.2.1. Resultados - Cuatro Clusters

A priori, parece una agrupación un tanto mala, ya que aproximadamente el 88% de las instancias fueron catalogadas dentro del Cluster C1.

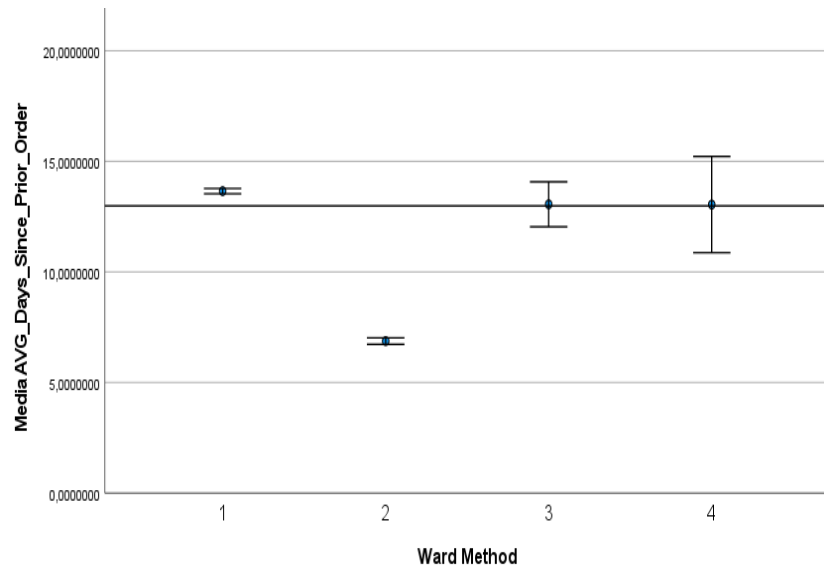
Ward Method		AVG_Hour_Of_Day	AVG_Days_Since_Prior_Order	Cantidad_Productos	Cantidad_Ventas	Cantidad_Categoría_Bebidas	Cantidad_Categoría_Pets	Cantidad_Categoría_Snack y Alcohol	Cantidad_Categoría_Productos	Cantidad_Categoría_Dairy Eggs
1	Media	13,4615394	13,65411624	97,84	10,44	.0045455802	.0019436324	.1003023017	.2784265077	.1543569603
	N	9044	9044	9044	9044	9044	9044	9044	9044	9044
	Desv. Desviación	2,00454490	5,783537529	94,238	8,597	.0179161799	.0107289072	.1218936117	.1896214254	.1128406314
2	Media	13,0028517	6,871584934	568,91	49,82	.0181306381	.0023543015	.0848930029	.3183205025	.1731871677
	N	983	983	983	983	983	983	983	983	983
	Desv. Desviación	1,43548965	2,362181307	299,225	19,042	.0380282386	.0119125072	.0683594224	.1425888318	.0785690183
3	Media	13,6649970	13,05871820	95,61	10,56	.2582897578	.0001974187	.0547938541	.1818380078	.1366387914
	N	128	128	128	128	128	128	128	128	128
	Desv. Desviación	2,20619299	5,804219503	96,134	9,932	.1044022652	.0013168585	.0535920040	.1330921362	.0942751898
4	Media	13,7164559	13,04528136	84,23	9,52	.0091333925	.2703782045	.0983254364	.1032633025	.1081549659
	N	44	44	44	44	44	44	44	44	44
	Desv. Desviación	2,24979327	7,154095037	140,277	10,840	.0341004341	.1899744733	.1293207750	.1161345697	.0993947706
Total	Media	13,4209833	12,99030332	143,16	14,23	.0090592806	.0031193646	.0982374519	.2803036807	.1557501636
	N	10199	10199	10199	10199	10199	10199	10199	10199	10199
	Desv. Desviación	1,96543273	5,901966241	189,821	15,404	.0370159222	.0240312648	.1173719114	.1857917879	.1099229479

Analisis por variable

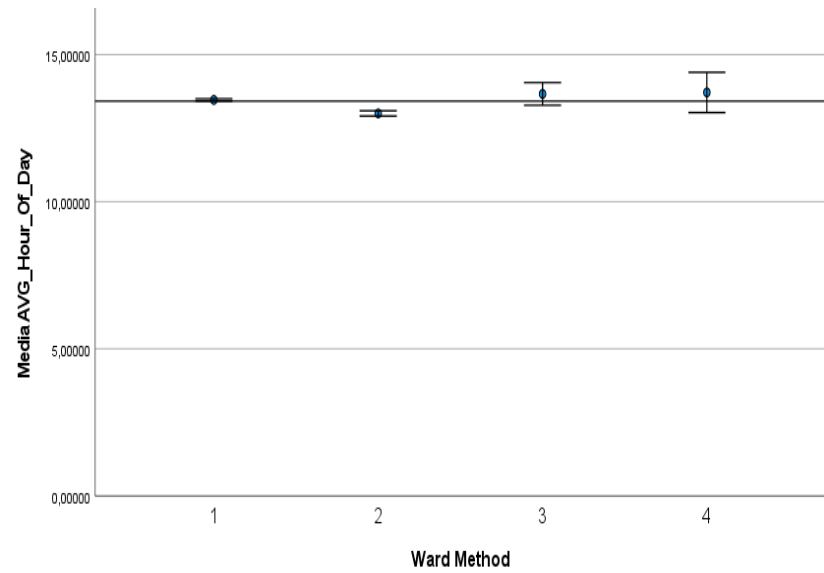
- AVG_Hour_Of_Day: Todos los clusters presentan una media muy cercana a la media poblacional.
- AVG_Days_Since_Prior_Order: C1, C3, C4 presentan medias muy similares. Se diferencia C2.
- Cantidad_Productos: C1 y C3 similares. C4 un tanto mas grande. C2 aun mas grande.
- Cantidad_Ventas: C1, C3, C4 presentan medias muy similares. Se diferencia C2.

A continuación se mostraran unas graficas, comparando variable por variable, con el fin de analizar dos aspectos relevantes de los clusters:

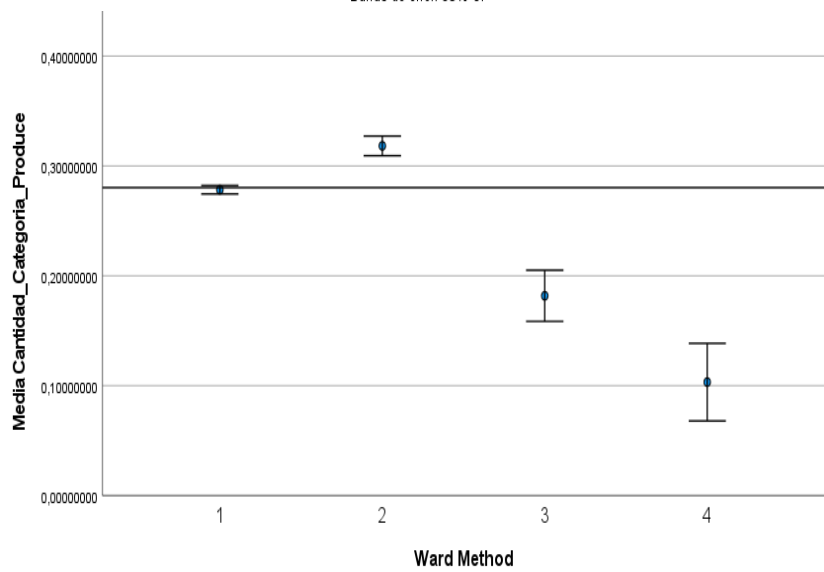
- Que tan diferentes son los clusters entre si
- Que tanto se diferencian entre clusters de la media poblacional



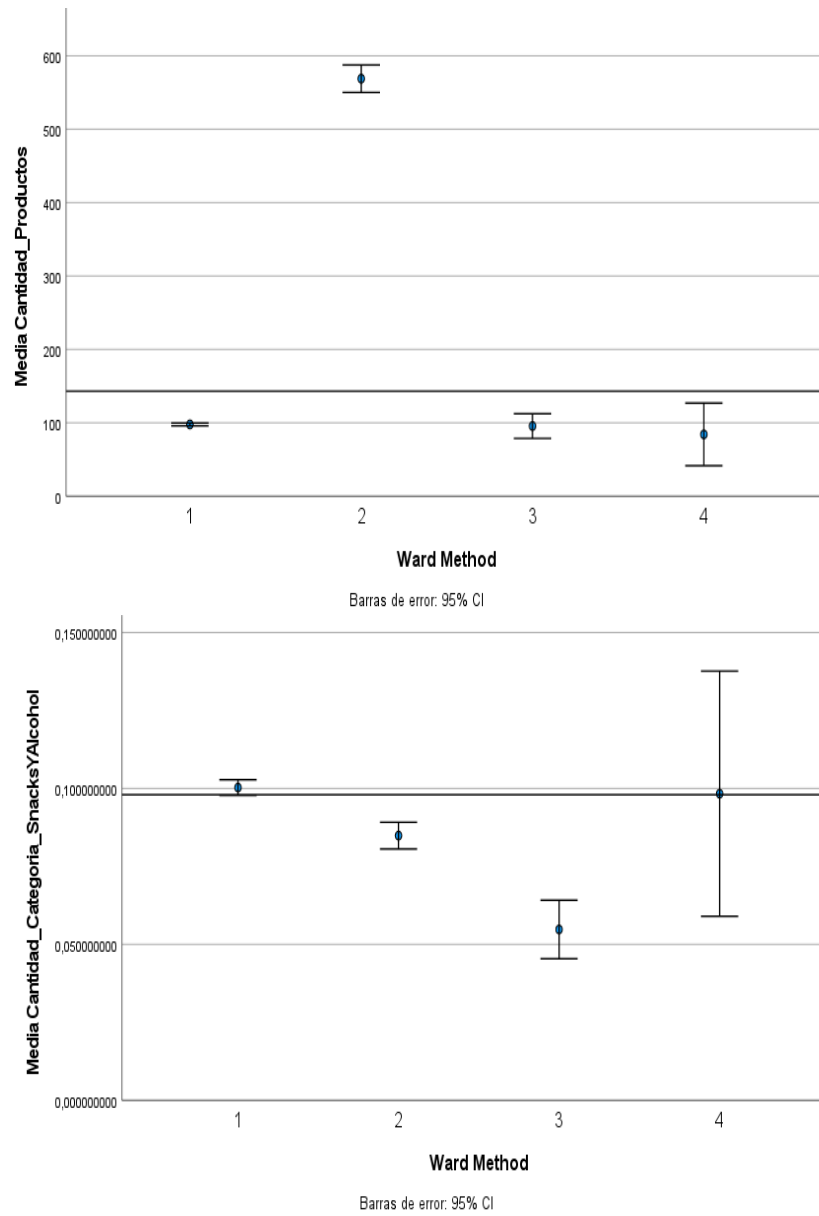
Barras de error: 95% CI



Barras de error: 95% CI



Barras de error: 95% CI



De todas estas graficas, ninguna cumple con los dos objetivos necesarios. Unicamente la variable *Cantidad_Productos* es en la que los C_i se diferencian de la media, ademas, la variable *Cantidad_Produce* es la unica en la que los C_i no se solapan entre si.

7.2.2. Resultados - Tres Clusters

A priori, parece una agrupacion un tanto mala, ya que aproximadamente el 89 % de las instancias fueron catalogadas dentro del Cluster C1. Da la impresion que de la clasificacion de Cuatro Clusters, C3 y C4 se combinaron con C1.

Ward Method		AVG_Hour_Of_Day	AVG_Days_Since_Prior_Order	Cantidad_Productos	Cantidad_Ventas	Cantidad_Categoria_Bebidas	Cantidad_Categoria_Pets	Cantidad_Categoria_SnacksAlcohol	Cantidad_Categoria_Productos	Cantidad_Categoria_DairyEggs
1	Media	13,4627736	13.65116853	97,77	10,44	.0045677923	.0032432716	.1002927307	.2775784464	.1541332711
	N	9088	9088	9088	9088	9088	9088	9088	9088	9088
	Desv. Desviación	2,00574386	5.790623357	94,508	8,609	.0180288453	.0251505443	.1219231946	.1897204521	.1128201626
	Desv. Desviación	2,00574386	5.790623357	94,508	8,609	.0180288453	.0251505443	.1219231946	.1897204521	.1128201626
2	Media	13,0028517	6,871584934	568,91	49,82	.0181306381	.0023543015	.0848930029	.3183205025	.1731871677
	N	983	983	983	983	983	983	983	983	983
	Desv. Desviación	1,43548965	2.362181307	299,225	19,042	.0380282386	.0119125072	.0683594224	.1425888318	.0785690183
	Desv. Desviación	1,43548965	2.362181307	299,225	19,042	.0380282386	.0119125072	.0683594224	.1425888318	.0785690183
3	Media	13,6649970	13.05871820	95,61	10,56	.2582897578	.0001974187	.0547938541	.1818380078	.1366387914
	N	128	128	128	128	128	128	128	128	128
	Desv. Desviación	2,20619299	5.804219503	96,134	9,932	.1044022652	.0013168585	.0535920040	.1330921362	.0942751898
	Desv. Desviación	2,20619299	5.804219503	96,134	9,932	.1044022652	.0013168585	.0535920040	.1330921362	.0942751898
Total	Media	13,4209833	12.99030332	143,16	14,23	.0090592806	.0031193646	.0982374519	.2803036807	.1557501636
	N	10199	10199	10199	10199	10199	10199	10199	10199	10199
	Desv. Desviación	1,96543273	5.901966241	189,821	15,404	.0370159222	.0240312648	.1173719114	.1857917879	.1099229479
	Desv. Desviación	1,96543273	5.901966241	189,821	15,404	.0370159222	.0240312648	.1173719114	.1857917879	.1099229479

7.2.3. Resultados - Dos Clusters

A priori, parece una agrupacion un tanto mala, ya que aproximadamente el 90 % de las instancias fueron catalogadas dentro del Cluster C1. Da la impresion que de la clasificacion de Cuatro Clusters, C4 se combino con C1.

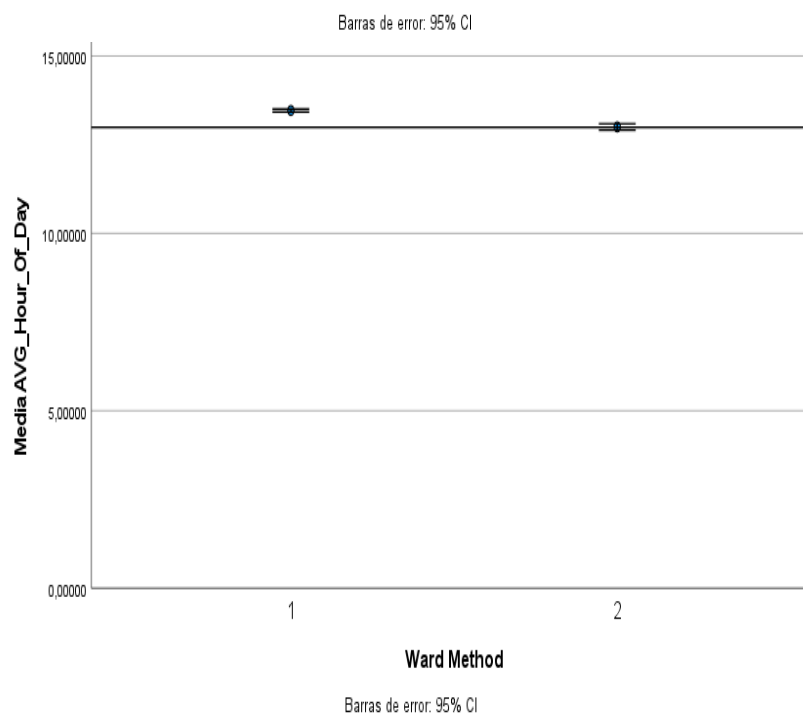
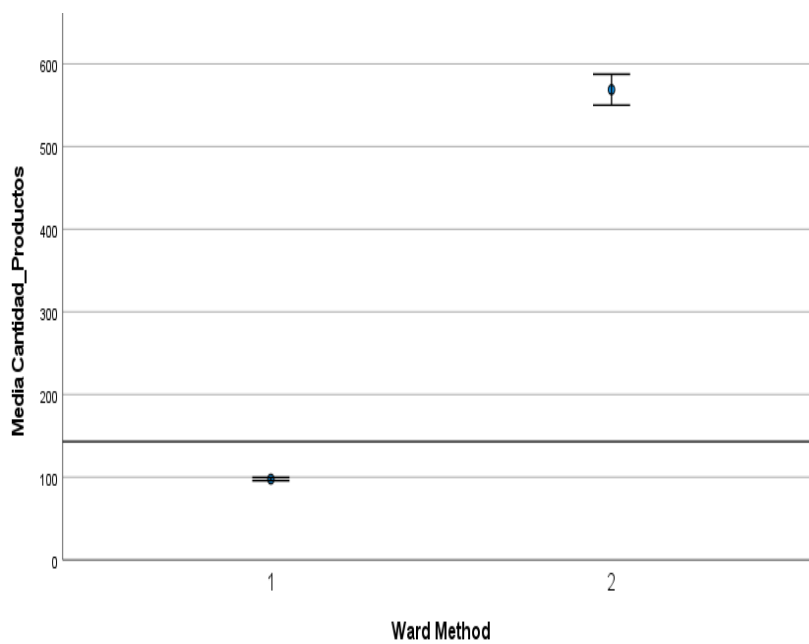
Ward Method		AVG_Hour_Of_Day	AVG_Days_Since_Prior_Order	Cantidad_Productos	Cantidad_Ventas	Cantidad_Categoria_Bebidas	Cantidad_Categoria_Pets	Cantidad_Categoria_SnacksAlcohol	Cantidad_Categoria_Productos	Cantidad_Categoria_DairyEggs
1	Media	13,4655822	13.64294006	97,74	10,44	.0080917085	.0032009681	.0996608018	.2762487181	.1538902922
	N	9216	9216	9216	9216	9216	9216	9216	9216	9216
	Desv. Desviación	2,00867314	5.790911890	94,526	8,628	.0367765841	.0249782802	.1213536968	.1893767696	.1125978226
	Desv. Desviación	2,00867314	5.790911890	94,526	8,628	.0367765841	.0249782802	.1213536968	.1893767696	.1125978226
2	Media	13,0028517	6,871584934	568,91	49,82	.0181306381	.0023543015	.0848930029	.3183205025	.1731871677
	N	983	983	983	983	983	983	983	983	983
	Desv. Desviación	1,43548965	2.362181307	299,225	19,042	.0380282386	.0119125072	.0683594224	.1425888318	.0785690183
	Desv. Desviación	1,43548965	2.362181307	299,225	19,042	.0380282386	.0119125072	.0683594224	.1425888318	.0785690183
Total	Media	13,4209833	12.99030332	143,16	14,23	.0090592806	.0031193646	.0982374519	.2803036807	.1557501636
	N	10199	10199	10199	10199	10199	10199	10199	10199	10199
	Desv. Desviación	1,96543273	5.901966241	189,821	15,404	.0370159222	.0240312648	.1173719114	.1857917879	.1099229479
	Desv. Desviación	1,96543273	5.901966241	189,821	15,404	.0370159222	.0240312648	.1173719114	.1857917879	.1099229479

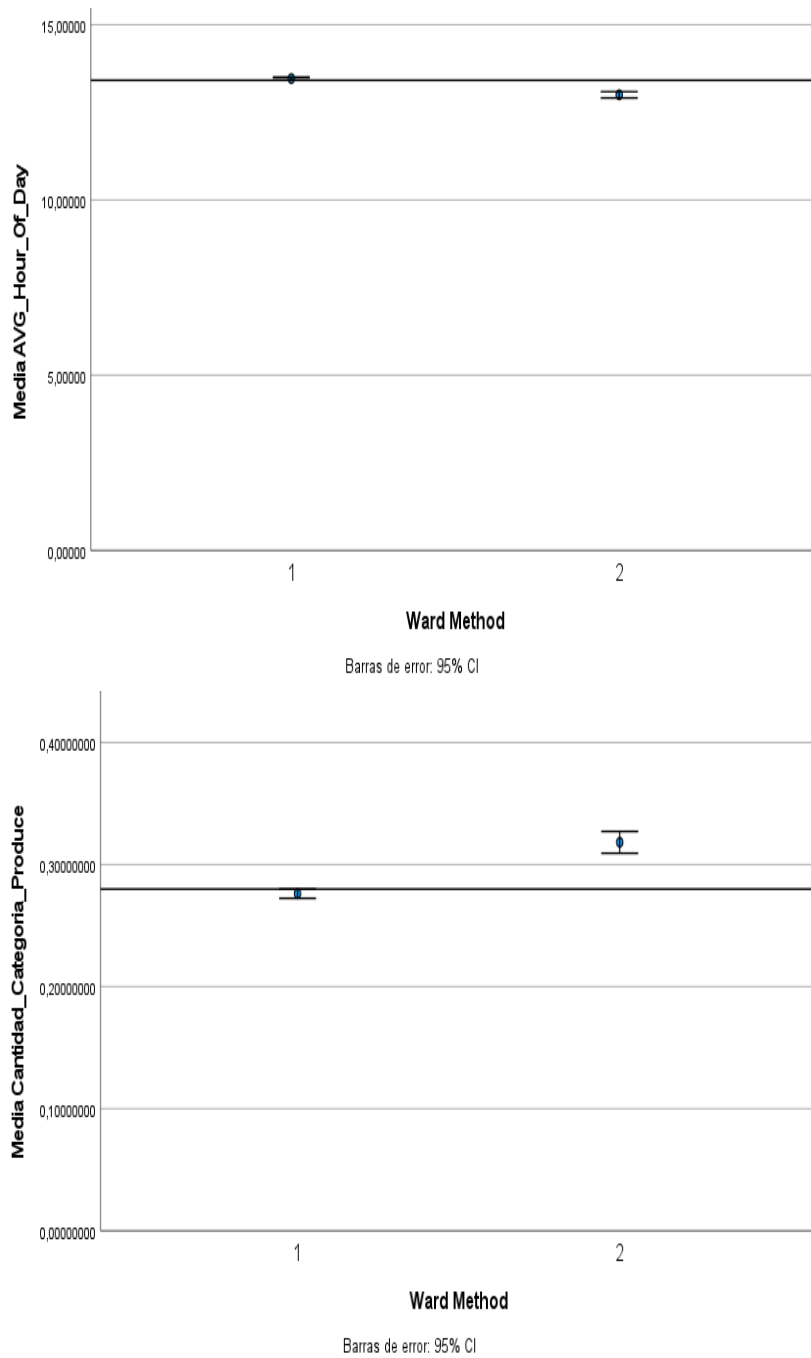
Analisis por variable

- AVG_Hour_Of_Day: Todos los clusters presentan una media muy cercana a la media poblacional.
- AVG_Days_Since_Prior_Order: C1 y C2 bien diferenciados.
- Cantidad_Productos: C1 y C2 bien diferenciados.
- Cantidad_Ventas: C1 y C2 bien diferenciados.

A continuacion se mostraran unas graficas, comparando variable por variable, con el fin de analizar dos aspectos relevantes de los clusters:

- Que tan diferentes son los clusters entre si
- Que tanto se diferencian entros clusters de la media poblacion





De las variables analizadas, unicamente el segundo cluster parece no diferenciarse de la media de *AVG_Hour_Of_Day*.

7.2.4. Resultados Finales

Asumiendo el rol de responsable de negocio, decidiremos quedarnos con el modelo de Cuatro Clusters, y ver si las reglas generadas en el Apartado II, muestran alguna variacion cuando las generamos para estos Clusters aislados.

Capítulo 8

Clustering No Jerarquico - K-Means

El algoritmo *K-Means* es un algoritmo de agrupamiento, no jerarquico, por lo que es obligatorio definir el numero de clusters que se desean formar.

8.1. Herramienta: IBM SPSS Statistics

Se utilizara esta herramienta comercializada por IBM, para la utilizacion del algoritmo de agrupamiento mencionado.

8.1.1. Parametros

Iteraciones Maximias

Se ha establecido en 10 iteraciones maximas, o hasta que se cumpla la condicion de convergencia (establecida en 0).

Numero de Clusters

Se ha decidido formar 4 clusters, para comparar con los resultados obtenidos en el capitulo de Clustering Jerarquico.

8.2. Resultados

8.2.1. Resultados - Cuatro Clusters

Número de casos en cada clúster

Clúster	1	13639,000
	2	147328,000
	3	3105,000
	4	41940,000
Válidos		206012,000
Perdidos		,000

Centros de clústeres iniciales

	Clúster			
	1	2	3	4
AVG_Hour_Of_Day	13,03950	18,00000	12,74650	11,49430
AVG_Days_Since_Prior_Order	6,7082800	30,0000000	5,8991000	3,9034400
Cantidad_Productos	2379	1	3459	1046
Cantidad_Ventas	57	1	62	97
Cantidad_Categoria_Babies	,000000000	,000000000	,000000000	,000000000
Cantidad_Categoria_Pets	,000000000	,000000000	,000000000	,017208400
Cantidad_Categoria_SnacksYAlcohol	,216478000	,000000000	,326684000	,013384300
Cantidad_Categoria_Products	,12274100	1,000000000	,06302400	,14914000
Cantidad_Categoria_DairyEggs	,18663300	,000000000	,14512900	,28202700

Historial de iteraciones^a

	Cambiar en centros de clústeres			
Iteración	1	2	3	4
1	1417,865	84,453	1700,946	611,409
2	350,933	20,533	606,483	166,851
3	47,628	3,724	63,045	24,499
4	4,048	,282	5,602	1,892
5	,000	1,911E-6	,002	4,532E-5
6	2,225E-8	1,267E-11	6,163E-7	1,085E-9
7	1,155E-14	,000	2,065E-10	,000
8	,000	,000	3,469E-18	,000
9	,000	,000	,000	,000

a. Convergencia conseguida debido a que no hay ningún cambio en los centros de clústeres o un cambio pequeño. El cambio de la coordenada máxima absoluta para cualquier centro es ,000. La iteración actual es 9. La distancia mínimo entre los centros iniciales es 1049,745.

Centros de clústeres finales

	Clúster			
	1	2	3	4
AVG_Hour_Of_Day	13,05960	13,48822	12,68780	13,31395
AVG_Days_Since_Prior_Order	7,8569965	14,1412894	5,7694163	11,3095899
Cantidad_Productos	554	59	1076	243
Cantidad_Ventas	45	8	65	24
Cantidad_Categoria_Babies	,015931662	,007388394	,021261181	,011084060
Cantidad_Categoria_Pets	,002704708	,003349621	,002632174	,003175272
Cantidad_Categoria_SnacksYAlcohol	,092372846	,099468307	,100785694	,092933356
Cantidad_Categoria_Produce	,30162874	,27326202	,29385815	,29068335
Cantidad_Categoria_DairyEggs	,17433218	,15158523	,17896981	,16551484

Distancias entre centros de clústeres finales

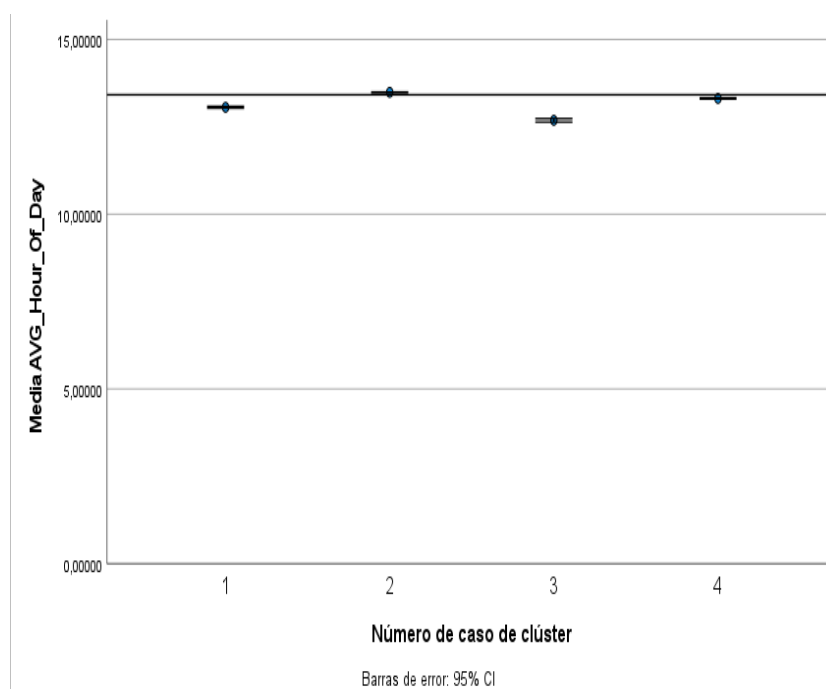
Clúster	1	2	3	4
1		497,266	521,634	312,226
2	497,266		1018,715	185,058
3	521,634	1018,715		833,762
4	312,226	185,058	833,762	

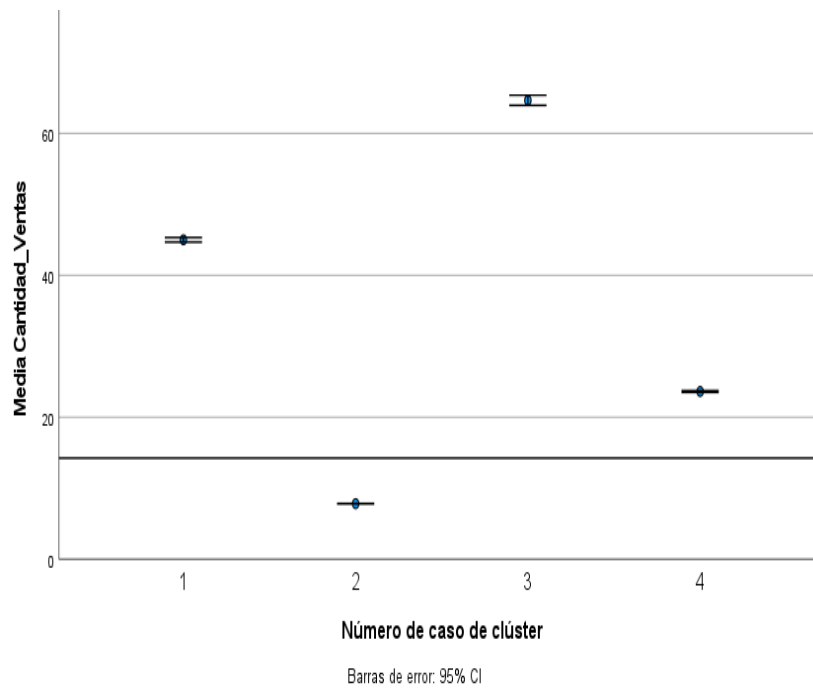
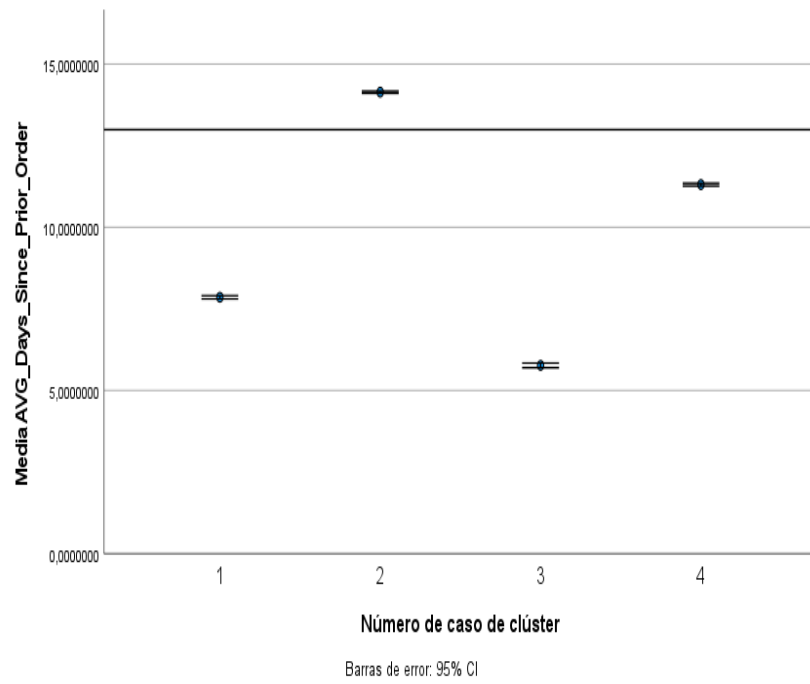
Analisis Variables

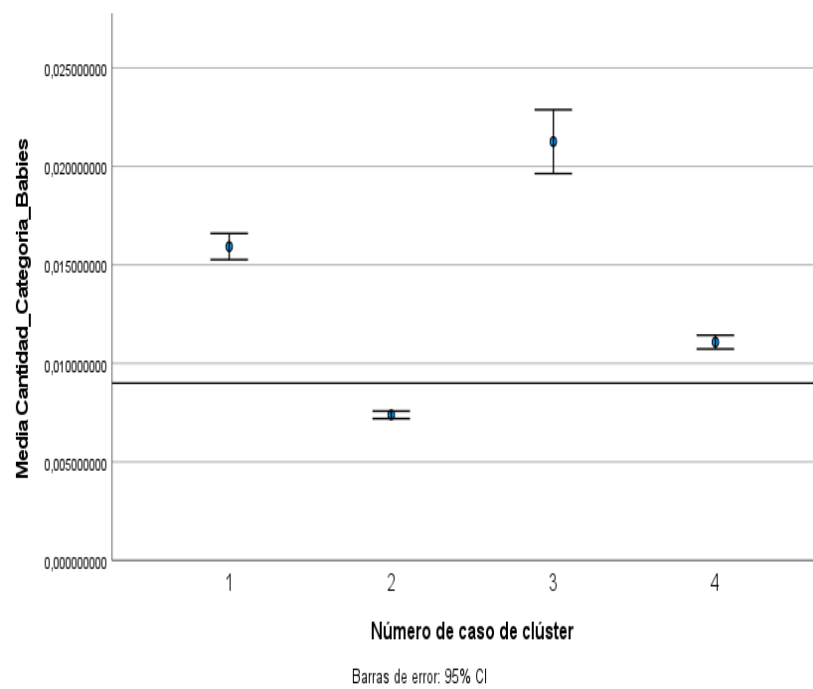
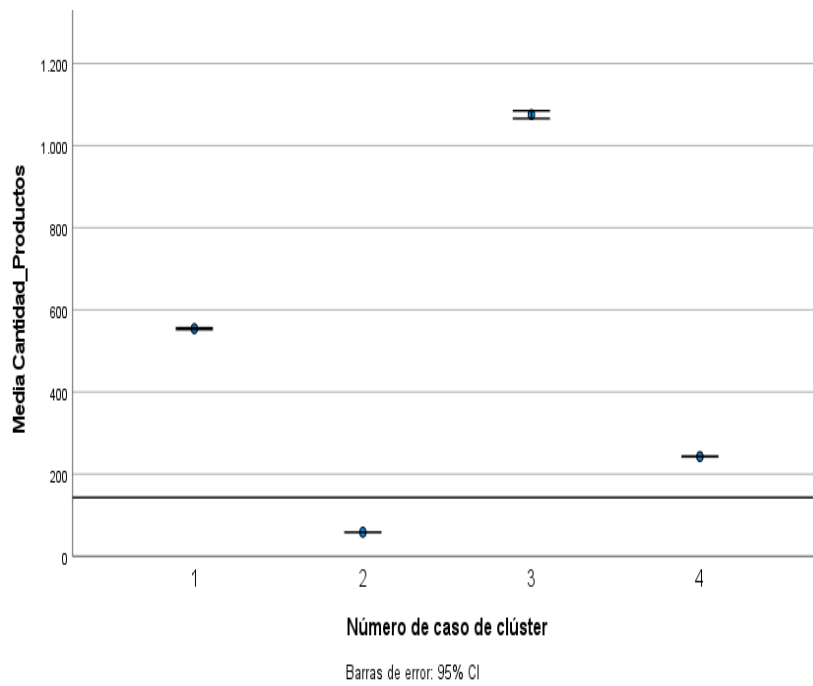
- AVG_Hour_Of_Day: Todos los clusters presentan valores muy cercanos a la media. Sin embargo, en la seccion posterior, veremos mediante un Intervalo de Confianza, que los todos los clusters se diferencian entre si, y ademas que, ninguno de ellos se solapa, en lo que respecta a la variable bajo analisis.
- AVG_Days_Since_Prior_Order: Todos los clusters se encuentran bien diferenciados, y alejados de la media poblacional.
- Cantidad_Productos: Tambien se encuentran bien diferenciados y lejos de la media poblacional. Su rango de medias esta comprendido entre [5, 77; 14,14] articulos por compra.
- Cantidad_Ventas: Igual que las demas variables, su distribucion es bien diferente. Su rango de medias esta comprendido entre [8; 65] ventas por cliente.

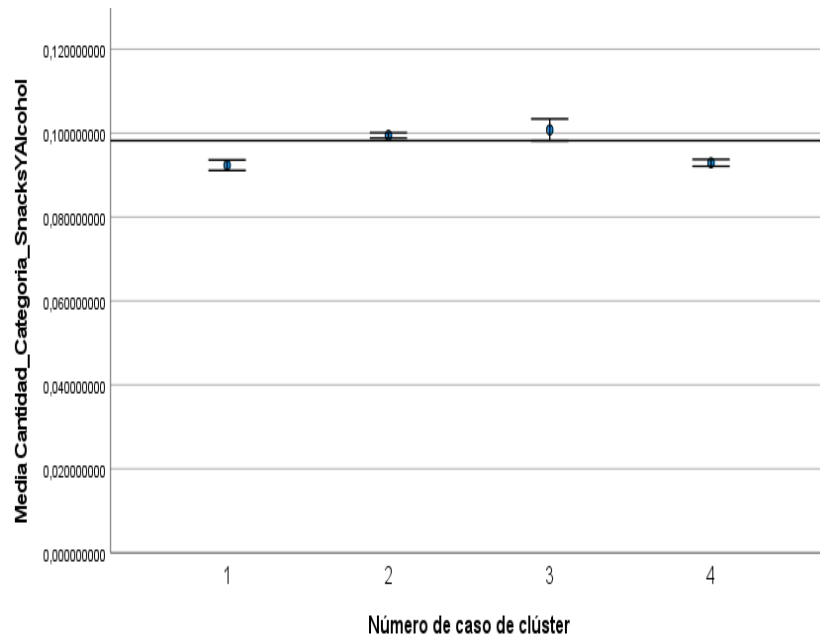
A continuacion se mostraran unas graficas, comparando variable por variable, con el fin de analizar dos aspectos relevantes de los clusters:

- Que tan diferentes son los clusters entre si
- Que tanto se diferencian entros clusters de la media poblacion

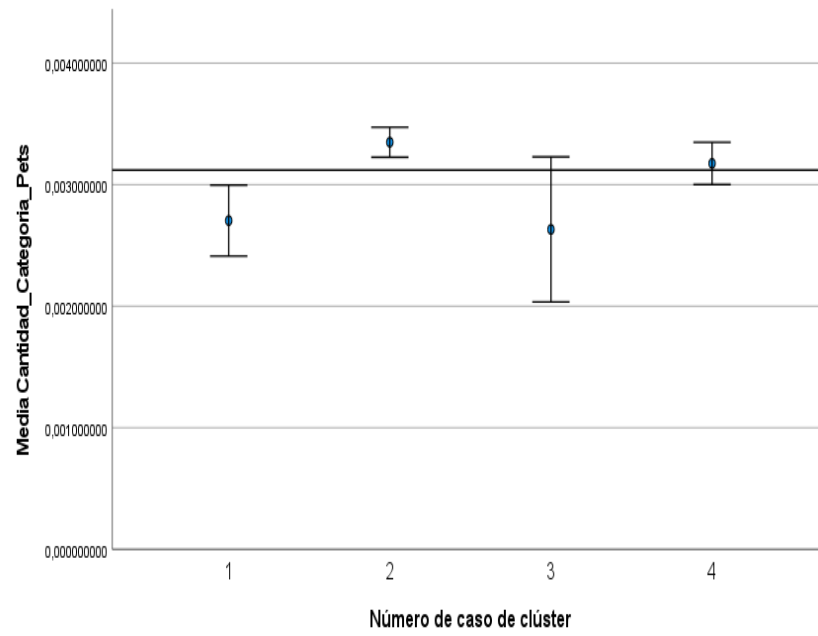








Barras de error: 95% CI



Barras de error: 95% CI

Capítulo 9

Clustering - Selección de Modelos

En este capítulo se realizará una evaluación y comparación entre los modelos de clustering generados por los algoritmos jerárquicos de Statistics, y el K-Means de la misma herramienta. Para esto, compararemos los resultados obtenidos con $K = 4$ (Cuatro Clusters).

Factores a considerar

- El modelo de Clustering Jerárquico contempla solamente el 5 % de los datos.
- En Clustering Jerárquico, ninguna de las variables analizadas logra cumplir con las dos condiciones (diferenciación de la media, no solapamiento).
- K-Means Contempla la totalidad de los datos.
- K-Means, al ser un método no jerárquico, está diseñado para trabajar con grandes volúmenes de datos.
- La mayor parte de las variables analizadas en K-Means, cumplen con las dos condiciones especificadas.

9.1. Conclusion

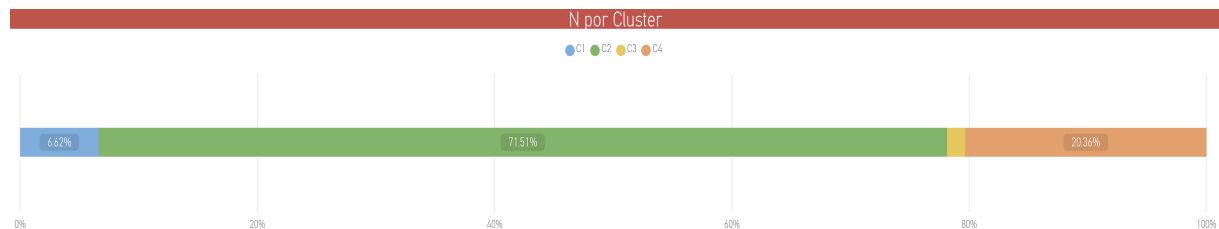
Se utilizarán los resultados obtenidos mediante la utilización del algoritmo K-Means, y se utilizarán esos subconjuntos de Datos para el estudio y generación de reglas de asociación.

Capítulo 10

Clustering - Caracterizacion

A continuacion se estimara un perfil, acorde a cada cluster resultante, con el fin de determinar características subjetivas que permitan describir a los mismos.

10.1. N: Cantidad de Clientes por Cluster



10.2. P: Promedio Cantidad de Productos Vendidos por Mes

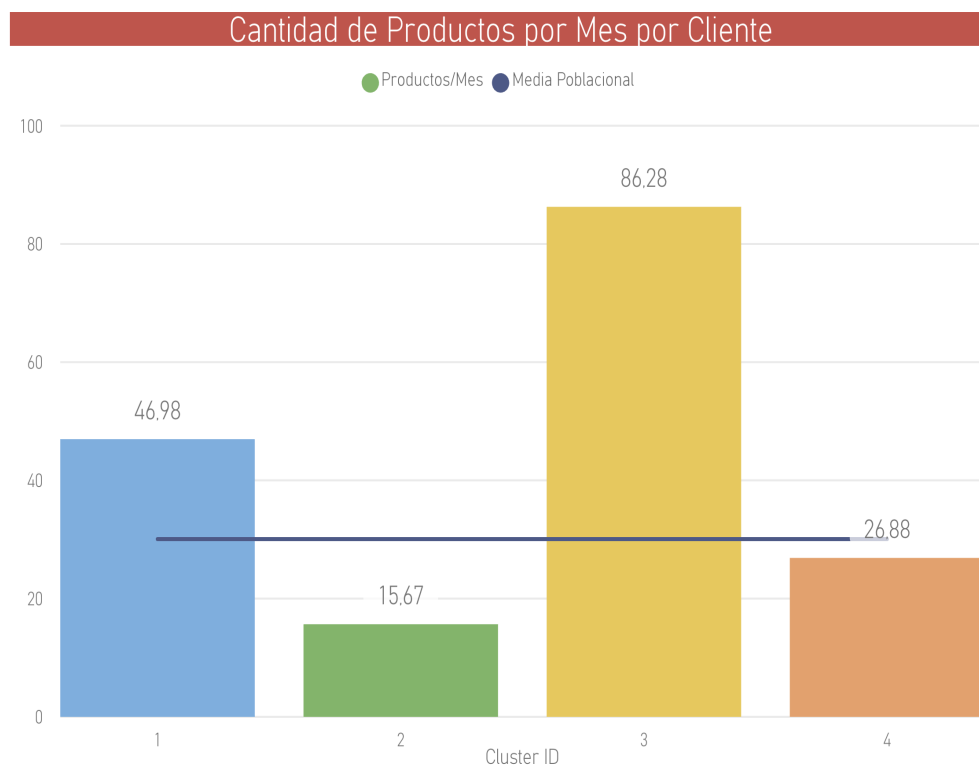
Tomando las variables explicativas:

AVG_Days_Since_Prior_Order, Cantidad_Productos, Cantidad_Ventas

se definira una metrica P , que explicara la cantidad de productos vendidos por mes. La metrica queda definida de la siguiente forma:

$$P = \frac{30 * Cantidad_Productos}{AVG_Days_Since_Prior_Order * Cant_Ventas}$$

] Con esta metrica, obtenemos los siguientes valores para cada cluster:



Parte IV

Reglas de Asociacion Dirigidas

Capítulo 11

Reglas Dirigidas

Nuevamente generamos las reglas de asociacion, pero esta vez, contemplando la segmentacion de clientes, y aumentando el umbral de soporte minimo a $0,040 = 4,0\%$. Mostramos los resultados mas importantes a continuacion:

11.1. Reglas Generadas

11.1.1. C_1

Se generaron un total de X Reglas, en un tiempo total de Y segundos. A continuacion se enumeran las reglas mas relevantes.

- Regla I

Antecedente: Organic Baby Spinach

Consecuente: Bag of Organic Bananas

Soporte Antecedente: 0.11087

Soporte Consecuente: 0.20731

Soporte regla: $0.0361 = 3,61\%$

Confianza: 0.3254

Lift: 1.570

Leverage: 0.0131

Conviction: 1.1751

■ Regla II

Antecedente: Bag of Organic Bananas

Concecuyente: Organic Strawberries

Soporte Antecedente: 0.207317

Soporte Concecuyente: 0.16314

Soporte regla: $0.0555 = 5,55 \%$

Confianza: 0.2677

Lift: 1.64095

Leverage: 0.02167

Conviction: 1.1427

■ Regla III

Antecedente: Organic Strawberries

Concecuyente: Banana

Soporte Antecedente: 0.16314

Soporte Concecuyente: 0.2190

Soporte regla: $0.0413 = 4,13 \%$

Confianza: 0.2532

Lift: 1.1562

Leverage: 0.00559

Conviction: 1.045

11.1.2. C_2

Se generaron un total de 8 Reglas, en un tiempo total de 487.97 segundos. A continuacion se enumeran las reglas mas relevantes.

■ Regla I

Antecedente: Bag of Organic Bananas

Concecuyente: Organic Baby Spinach

Soporte Antecedente: 0.0956

Soporte Concecuyente: 0.0633

Soporte regla: $0.0120 = 1,20 \%$

Confianza: 0.1256

Lift: 1.9819

Leverage: 0.00595

Conviction: 1.0711

■ Regla II

Antecedente: Bag of Organic Bananas

Concecuente: Organic Hass Avocado

Soporte Antecedente: 0.0956

Soporte Concecuente: 0.0956

Soporte regla: $0.01147 = 1,15 \%$

Confianza: 0.120

Lift: 2.8779

Leverage: 0.00749

Conviction: 1.08898

11.1.3. C_3

Se generaron un total de 3 Reglas, en un tiempo total de 16.54 segundos. A continuacion se enumeran las reglas mas relevantes.

■ Regla I

Antecedente: Organic Hass Avocado

Concecuente: Bag of Organic Bananas

Soporte Antecedente: 0.1540

Soporte Concecuente: 0.2659

Soporte regla: $0.0836 = 8,36 \%$

Confianza: 0.5431

Lift: 2.0425

Leverage: 0.0427

Conviction: 1.6068

■ Regla II

Antecedente: Organic Raspberries

Concecuente: Bag of Organic Bananas

Soporte Antecedente: 0.14127

Soporte Concecuente: 0.2659

Soporte regla: $0.0614 = 6,14 \%$

Confianza: 0.4352

Lift: 1.6368

Leverage: 0.0239

Conviction: 1.299

11.1.4. C_4

Se generaron un total de 1 Reglas, en un tiempo total de 47.38 segundos. A continuacion se enumeran las reglas mas relevantes.

- Regla I

Antecedente: Bag of Organic Bananas

Concecuyente: Organic Strawberries

Soporte Antecedente: 0.1502

Soporte Concecuente: 0.1081

Soporte regla: $0.0345 = 3,45 \%$

Confianza: 0.23008

Lift: 2.1270

Leverage: 0.0183

Conviction: 1.1583

11.1.5. Analisis

En lo que al tiempo computacional concierne, el clustering ayuda a disminuirlo notablemente. Incluso, para el cluster C_2 que concentra aproximadamente al 78 % de los datos, el tiempo de computo es aproximadamente la mitad. En los demas Grupos, el tiempo no supera a los 60 segundos de ejecucion.

En cuanto a las reglas generadas, las reglas sobre C_3 tienen presentan un soporte y lift considerablemente mayor que el resto de las reglas.

Capítulo 12

Reglas Dirigidas sobre Categorías y Pasillos

Tanto las reglas generadas sobre el conjunto entero de datos, como las generadas sobre los clusters definidos, dieron como resultado reglas bastante similares, en relación a los productos seleccionados.

En este capítulo se enunciarán nuevas reglas, pero construidas sobre las **Categorías** y **Pasillos** a los que pertenecen estos productos.

Estos resultados pueden ser útiles para, por ejemplo, elegir una mejor distribución de las gondolas, de forma tal de acercar o agrupar productos relacionados o no relacionados.

12.1. Reglas por Categoria

- Umbral Soporte minimo: $0,34 = 34\%$
- Umbral Lift Minimo: 1,1

12.1.1. C_1, C_2, C_3, C_4

Se generaron un total de 1 Regla, en un tiempo total de 14.78 segundos. A continuacion se enumeran las reglas mas relevantes.

- Regla I

Antecedente: produce

Concecuyente: dairy eggs

Soporte Antecedente: 0.7370

Soporte Concecuyente: 0.6658

Soporte regla: $0.5428 = 54,28\%$

Confianza: 0.7364

Lift: 1.1060

Leverage: 0.0520

Conviction: 1.2678

12.1.2. C_1

Se generaron un total de 1 Regla, en un tiempo total de 14.78 segundos. A continuacion se enumeran las reglas mas relevantes.

- Regla I

Antecedente: snacks

Concecuyente: beverages

Soporte Antecedente: 0.5668

Soporte Concecuyente: 0.5491

Soporte regla: $0.3525 = 35,25\%$

Confianza: 0.6219

Lift: 1.132

Leverage: 0.04126

Conviction: 1.1925

12.1.3. C_2

Se generaron un total de 1 Regla, en un tiempo total de 14.24 segundos. A continuacion se enumeran las reglas mas relevantes.

- Regla I

Antecedente: produce

Concecuyente: dairy eggs

Soporte Antecedente: 0.6941

Soporte Concecuyente: 0.6177

Soporte regla: $0.4828 = 48,28 \%$

Confianza: 0.695

Lift: 1.125

Leverage: 0.0539

Conviction: 1.2552

12.1.4. C_3

Se generaron un total de 24 Regla, en un tiempo total de 13.58 segundos. Se aumenta el soporte minimo a $0,40 = 40,0 \%$, y el umbral lift minimo a 1,15. A continuacion se enumeran las reglas mas relevantes.

- Regla I

Antecedente: beverages y dairy eggs

Concecuyente: snacks

Soporte Antecedente: 0.5374

Soporte Concecuyente: 0.6410

Soporte regla: $0.400 = 40,00 \%$

Confianza: 0.7443

Lift: 1.1612

Leverage: 0.0555

Conviction: 1.404

12.1.5. C_4

Se generaron un total de 24 Regla, en un tiempo total de 12.72 segundos. Se aumenta el soporte minimo a $0,40 = 40,0 \%$, y el umbral lift minimo a 1,15. A continuacion se enumeran las reglas mas relevantes.

- Regla I

Antecedente: produce y snacks

Consecuente: dairy eggs

Soporte Antecedente: 0.4300

Soporte Consecuente: 0.766

Soporte regla: $0.3644 = 36,44\%$

Confianza: 0.8473

Lift: 1.1052

Leverage: 0.0347

Conviction: 1.5288

12.1.6. Analisis

Un dato importante es que las categorías que conforman las reglas, varían cuando se itera sobre los clusters formados, lo que da a suponer que el agrupamiento realizado puede ser considerado útil.

Si bien, como era de esperar, el soporte de las reglas aplicadas sobre las Categorías es considerablemente mayor (x10 aproximadamente) al soporte de las mismas aplicadas sobre los productos, el *lift* $\rightarrow 1,1$, lo que demuestra que estas reglas son casi triviales (su aparición tiende a la esperanza matemática de las mismas). En conclusión, no aportan mucha información.

12.2. Reglas por Pasillo

- Umbral Soporte minimo: $0,25 = 25 \%$
- Umbral Lift Minimo: 1,4

12.2.1. C_1, C_2, C_3, C_4

Se generaron un total de 1 Regla, en un tiempo total de 18.82 segundos. A continuacion se enumeran las reglas mas relevantes.

- Regla I

Antecedente: fresh vegetables

Concecuente: packaged vegetables fruits

Soporte Antecedente: 0.4491

Soporte Concecuente: 0.3816

Soporte regla: $0.2506 = 25,06 \%$

Confianza: 0.5580

Lift: 1.4624

Leverage: 0.0792

Conviction: 1.399

12.2.2. C_1

Se generaron un total de 3 Regla, en un tiempo total de 13.16 segundos. A continuacion se enumeran las reglas mas relevantes.

- Regla I

Antecedente: packaged vegetables fruits

Concecuente: fresh fruits, fresh vegetables

Soporte Antecedente: 0.5441

Soporte Concecuente: 0.5022

Soporte regla: $0.3499 = 34,99 \%$

Confianza: 0.6430

Lift: 1.280

Leverage: 0.0766

Conviction: 1.3944

■ Regla II

Antecedente: packaged vegetables fruits

Concecuyente: fresh vegetables

Soporte Antecedente: 0.54417

Soporte Concecuyente: 0.60141

Soporte regla: $0.3947 = 39,47 \%$

Confianza: 0.7253

Lift: 1.2060

Leverage: 0.0674

Conviction: 1.4512

12.2.3. C_2

Se generaron un total de 1 Regla, en un tiempo total de 16.79 segundos. A continuacion se enumeran las reglas mas relevantes.

■ Regla I

Antecedente: fresh fruits

Concecuyente: fresh vegetables

Soporte Antecedente: 0.4923

Soporte Concecuyente: 0.4039

Soporte regla: $0.2742 = 27,42 \%$

Confianza: 0.5569

Lift: 1.3788

Leverage: 0.0753

Conviction: 1.3453

12.2.4. C_3

Se generaron un total de 24 Regla, en un tiempo total de 13.58 segundos. Se aumenta el soporte minimo a $0,40 = 40,0\%$, y el umbral lift minimo a 1,15. A continuacion se enumeran las reglas mas relevantes.

- Regla I

Antecedente: fresh fruits y packaged cheese

Concecuyente: fresh vegetables

Soporte Antecedente: 0.3728

Soporte Concecuyente: 0.614

Soporte regla: $0.2775 = 27,75\%$

Confianza: 0.744

Lift: 1.211

Leverage: 0.0484

Conviction: 1.508

- Regla II

Antecedente: fresh fruits y fresh vegetables

Concecuyente: packaged vegetables fruits

Soporte Antecedente: 0.54903

Soporte Concecuyente: 0.5950

Soporte regla: $0.4060 = 40,60\%$

Confianza: 0.7396

Lift: 1.2430

Leverage: 0.0794

Conviction: 1.555

12.2.5. C_4

Se generaron un total de 24 Regla, en un tiempo total de 13.58 segundos. Se aumenta el soporte minimo a $0,40 = 40,0\%$, y el umbral lift minimo a 1,15. A continuacion se enumeran las reglas mas relevantes.

- Regla I

Antecedente: fresh fruits

Concecuyente: packaged vegetables fruits

Soporte Antecedente: 0.6580

Soporte Concecuyente: 0.4708

Soporte regla: $0.3775 = 37,75\%$

Confianza: 0.57364

Lift: 1.2182

Leverage: 0.06762

Conviction: 1.2410

12.2.6. Analisis

Comportamiento similiar al obtenido cuando se contemplaron las Categorías. Se obtuvieron algunas reglas con un lift un tanto mas elevado que al obtenido con las Categorías. Se podria evaluar la generacion de reglas variando el soporte y el lift minimo (bajar el soporte y aumentar el lift).

Parte V























Conclusiones Finales

Capítulo 13

Conclusiones Finales

13.1. Conclusiones

A continuacion se presenta un resumen de las reglas mas destacadas

ID	Cluster	Antecedente	Consecuente	Soporte Regla	Confianza	Lift
1	All	Bag of Organic Bananas	Organic Strawberries	 2,29%	 19,65%	2,398
2	All	Large Lemon	Banana	 1,63%	 26,46%	1,8524
3	All	Banana	Organic Avocado	 1,68%	 11,79%	2,0811
4	All	Organic Baby Spinach	Bag of Organic Bananas	 1,69%	 22,77%	1,94876
5	C1	Organic Baby Spinach	Bag of Organic Bananas	 3,61%	 32,54%	1,5700
6	C1	Bag of Organic Bananas	Organic Strawberries	 5,55%	 26,77%	1,6410
7	C2	Bag of Organic Bananas	Organic Baby Spinach	 1,20%	 12,56%	1,9819
8	C2	Bag of Organic Bananas	Organic Hass Avocado	 1,15%	 12,00%	2,8779
9	C3	Organic Hass Avocado	Bag of Organic Bananas	 8,36%	 54,31%	2,0450
10	C3	Organic Raspberries	Bag of Organic Bananas	 6,14%	 43,52%	1,6368
11	C4	Bag of Organic Bananas	Organic Strawberries	 3,45%	 23,01%	2,1270

Como conclusion podemos decir que:

- Los productos que participan de las reglas generadas sobre el conjunto de datos completo, difieren de los productos generados sobre los clusters.
- Tanto el soporte como la confianza tienden a aumentar en las reglas generadas sobre los clusters.
- A simple vista no parecen mejorar los valores de Lift cuando se utilizan los clusters.

En definitiva, la utilizacion de clusters parece una buena idea, a priori, para generar reglas mas especificas.

Capítulo 14

Versionado

Version	Autor	Fecha	Descripcion
1.0	Djemdjemian, Ezequiel	2020-11-30	Primera Version

Cuadro 14.1: Tabla de Versiones