Análisis de relación de productos, segmentación de clientes y predicción de ventas de una tienda online

INTRODUCCIÓN

Guillermo Sanz Berjas

Master Data Science

Kschool

CONTENIDO

INDICE DE FIGURAS

INDICE DE TABLAS

Resumen

En este trabajo se va a intentar entender todos los parámetros que influyen en las decisiones de negocio de una tienda/supermecado online.

Para ello dividiremos el negocio en los distintos componente que lo conforman y intentaremos establecer y/o encontrar las relaciones que los unen.

Los tres principales componentes que analizaremos serán los clientes, los objetos a la venta y los ingresos.

Intentaremos establer relaciones entre;

- Los objetos entre sí (market basket analisys)
- Los clientes por su gasto, frecuencia y última compra (RFM analisys)
- Los clientes entre ellos, generando clusters según los objetos comprados (unsupervised machine learning)
- Establecer una precidión sobre los ingresos (SARIMA, TensorFlow)

Para ello, usaremos el lenguaje pyhton implementado sobre la plataforma Google Colaboratory.

Palabras clave: python, Google Colaboratory, market basket analisys, RFM analisys, unsupervised learning, SARIMA, TensorFlow

INTROUCCIÓN

En últimos años se ha incrementado el uso de nuevas tecnológias por parte de las diferentes compañias de ventas con el fin de optimizar, mejorear e incrementar tanto sus ingresos como su servicio al cliente. Usando la ciencia de datos, es posible analizar grandes cantidades de información y con ello conseguir dos objetivos fundamentales:

- Optimizar los modelos tradiccionales de relación, agrupamiento y clasificacón de prouctos y clientes.
- Establecer las relaciones y las tendencias escondidas en esa información.

PLANTEAMIENTO DEL PROBLEMA

Partiendo de una base de datos sobre las distintas transaciones que tuvieron lugar a lo largo de 14 meses en una tienda de venta online, se intentara extraer, aplicando sisitemas de minerias de datos, de aprendizaje no supervisado y de predicción en líneas temporales, estrategías para la mejora tanto del negocio, como del servicio, como de los ingresos.

Esto es importante porque este modelo será aplicable a multitud de negocios, implementando las peculiaridades propias en cada uno de ellos, además de acercar el punto de vista de personal propio de desarrollo de negocio con la nueva visión que nos da el trabajar con la eficacia del data science.

DATOS

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	2011-12-09 12:50:00	0.85	12680.0	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	2011-12-09 12:50:00	2.10	12680.0	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	2011-12-09 12:50:00	4.15	12680.0	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	2011-12-09 12:50:00	4.15	12680.0	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	2011-12-09 12:50:00	4.95	12680.0	France

541909 rows × 8 columns

Tabla1. Datos sin tratar

En la tabla anterior nos muesta las 6 primeras y las 6 últimas líneas del dataset original. Este dataset fue publicado por Dr Daqing Chen, de la escuela de ingenieros de la universidad London South Bank, y contiene la información de las transacciones realizadas en una tienda online desde el 01-12-2010 hasta el 31-12-2011.

El data set completo tiene cerca de 542000 líneas y 8 columnas que contienen:

- InvoiceNo: es el número de transacción
- StockCode: es un número asignado a cada producto. Este columna no será utilizada en ningún momento
- Description: Nombre del producto vendido
- Quantity: número del mismo objeto vendido
- InvoiceData: fecha en la que se hizo la tansacción
- UnitPrice: precio por unidad de producto
- CustomerID: número asignado a cada cliente
- Country: país donde se realizo la transacción, al considerarlo una tienda online, este parametro será ignorado.

METODOLOGÍA

Se han divido los diferentes pasos y métodos de analisis que se desarrollan a lo largo del proyecto en diferentes notebooks para mejor su entendimiento y para reducir la cantidad de procedo que hay que hacer para recorrer cada notebook.

1. LIMPIEZA DE DATOS

El modelo que se ha seguido en este caso es ir mirando columna a columna los diferentes problemas que pueden tener cada una de ellas por separado.

Empezando por la colunma InvoiceNO a los datos nos muestra que existen codigos empezados con la letra 'C' que llevan unidos valores negativos en la conlumna UnitPrice, lo que nos da a entender que tambíen se han registrado las devoluciones.. Asi que primeramente, elimanaremos todas esas filas ya que no nos interesan en nuestro análisis.

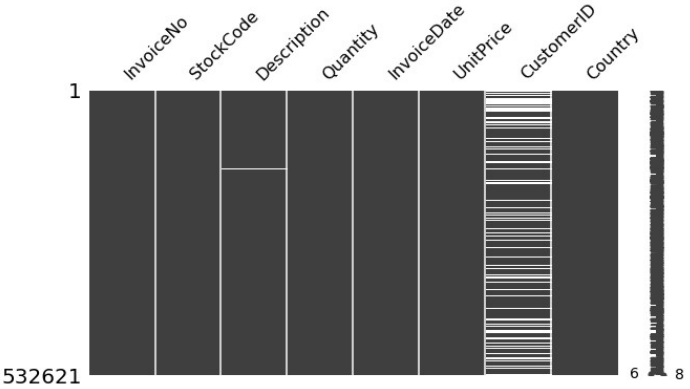
```
positive = df.UnitPrice > 0
positive.value_counts()
```

```
True     539392
False      2517
```

Tabla2. Número de valores negativos y positivos de las transacciones

Seguidamente, para la columna Description, nos aseguramos de eliminar todos los espacios que existen al principio y al final de string para evitar problemas de agrupamiento en el futuro.

Continuamos con la columna CustomerID donde vemos, utilizando la biblioteca missingno, que tenemos un gran número de datos en blanco. Como el número de cliente no se puede predecir o interpolar de otros datos, se decide eliminar esas filas.



Gráfica1. Número de celdas vacías en el dataset

Una vez comprobado que todas las columnas están correctas vamos a crear una nueva combinando el número de objets vendidos por su precio unidad para obtener el precio total por producto/transacíon, esto nos servirá en los proximos análisis.

Para desarrollar los futuros análisis, ya sea por una cuestión de capacidad de procesado para algunos de ellos, ya sea por reducción de variables para otros, sea ha decidio reducir la lista de productos, cercanos a 4000 referencias, a los 25 más frecuentes. Para ello establecemos un indice con los nombres de los 25 productos más frecuentes y transformamos los demás en un producto llamado 'other'

```
df.Description.value_counts()
```

```
other                                369810
WHITE HANGING HEART T-LIGHT HOLDER     2028
REGENCY CAKESTAND 3 TIER               1724
JUMBO BAG RED RETROSPOT                1618
ASSORTED COLOUR BIRD ORNAMENT          1408
PARTY BUNTING                          1397
LUNCH BAG RED RETROSPOT                1316
SET OF 3 CAKE TINS PANTRY DESIGN       1159
LUNCH BAG  BLACK SKULL.                1105
POSTAGE                                1099
PACK OF 72 RETROSPOT CAKE CASES        1068
PAPER CHAIN KIT 50'S CHRISTMAS         1019
SPOTTY BUNTING                         1017
LUNCH BAG SPACEBOY DESIGN              1008
LUNCH BAG CARS BLUE                     989
HEART OF WICKER SMALL                   985
NATURAL SLATE HEART CHALKBOARD          980
LUNCH BAG PINK POLKADOT                 957
REX CASH+CARRY JUMBO SHOPPER            952
LUNCH BAG SUKI DESIGN                   933
ALARM CLOCK BAKELIKE RED                899
LUNCH BAG APPLE DESIGN                  895
SET OF 4 PANTRY JELLY MOULDS            893
JUMBO BAG PINK POLKADOT                 890
JAM MAKING SET WITH JARS                888
WOODEN PICTURE FRAME WHITE FINISH       887
Name: Description, dtype: int64
```

Tabla3. Top25 productos más vendidos

Esta última acción la dividimos en diversos pasos por ser un dataset con demasiadas filas para hacerlo de una sola vez.

Incluso así, el programa avisa de posibles errores, pero en este caso, usando Google Colaboratory, gentiona el problema y da la solución esperada.

Finalmente concatenamos las tablas resultantes.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	PrecioTotal
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	15.30
1	536365	71053	other	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
2	536365	84406B	other	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	22.00
3	536365	84029G	other	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
4	536365	84029E	other	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
...	...	...	...	...	...	...	...	...	...
197919	581587	22613	other	12	2011-12-09 12:50:00	0.85	12680.0	France	10.20
197920	581587	22899	other	6	2011-12-09 12:50:00	2.10	12680.0	France	12.60
197921	581587	23254	other	4	2011-12-09 12:50:00	4.15	12680.0	France	16.60
197922	581587	23255	other	4	2011-12-09 12:50:00	4.15	12680.0	France	16.60
197923	581587	22138	other	3	2011-12-09 12:50:00	4.95	12680.0	France	14.85

397924 rows × 9 columns

Tabla4. Tabla resultante tras limpiar y ordenar los datos

2.ANÁLISIS CUALITATIVO DEL NEGOCIO

Partiendo de los datos anteriores, en este punto, intentaremos ver los datos económicos del negocio a un nivel cualitativo.

Para ello intenremos ver si existe una estacionalidad en las compras, cual es la relación entre los productos más vendios, la direncia de ingresos de los dos años en relación con el número de productos vendidos…

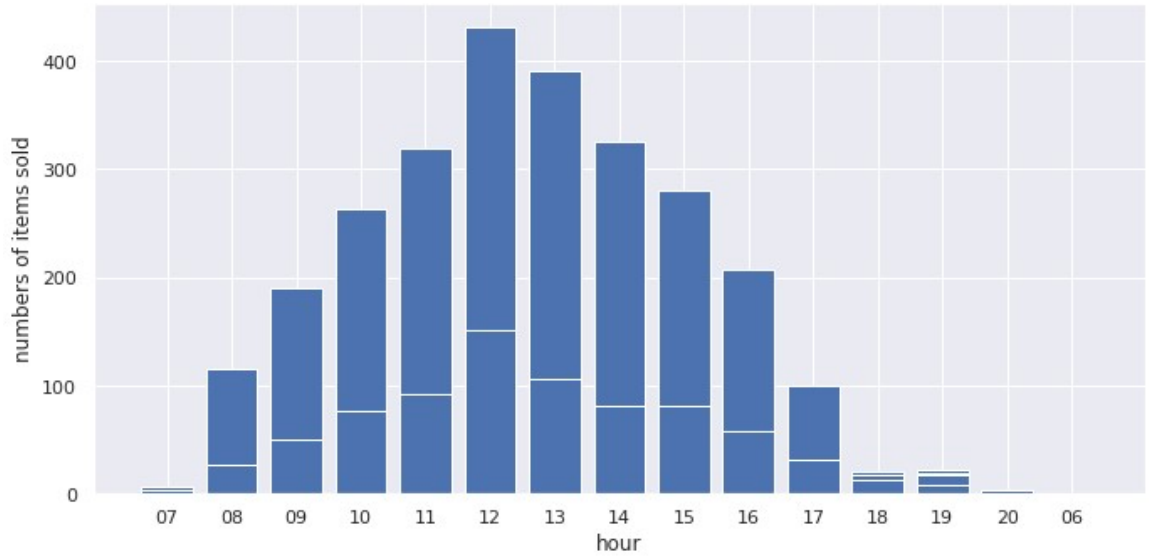
Todas estos análisis serán fundamentales para determinar el uso de los modelos de los análisis posteriores

Estacionalidad

Al ser un negocio online, existen dos parametros que nos interesan especialmente:

1. Las horas a las que se producen un mayor número de compras
2. Los meses donde se venden más productos

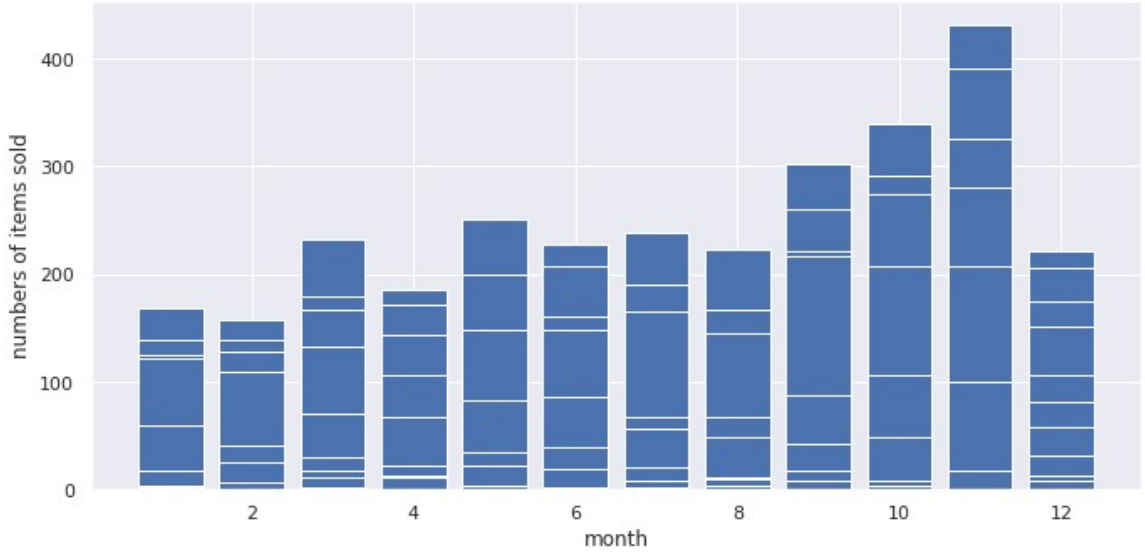
Número de compras por horas



Gráfica2. Número de transacciones por hora del día

En la gráfica anterior podemos comprobar que las distribución de las compras a lo largo del día sigue casi una distribución en campana de Gaus con el pico en las horas centrales del día. Esto nos permite establecer, entre otras cosas, los horarios optimos para los trabajadores de nuestra empresa o la capacidad máxima de peticiones que tiene que soportar nuesta web.

Número de compras por mes

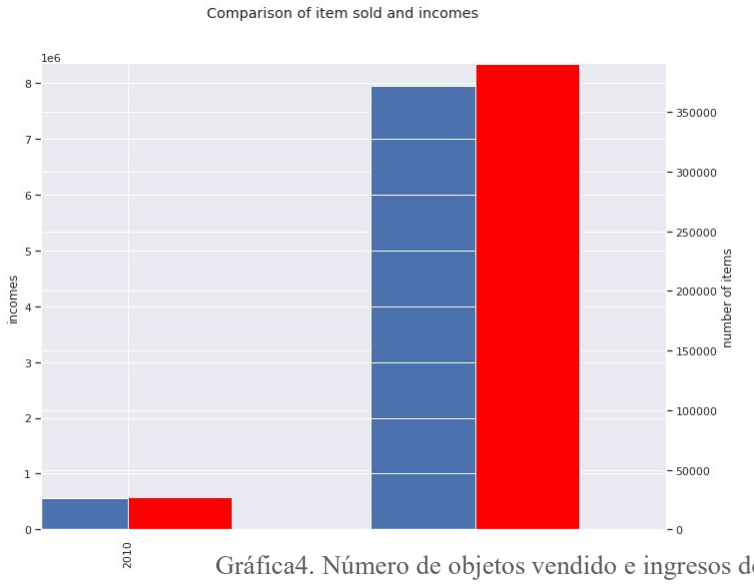


Gráfica3. Número de transacciones por mes

Los datos sugieren que hay una gran estacionalidad entorno a los últimos meses del año, pero las ventas caen precipitadamente en el mes de diciembre. Esto habria de ser analizado, ya que suele ser un mes muy fuerte de ventas.

Podía ser algún tipo de caida de la web, se han tomado menos datos ese mes que otros o algúna otra circunstancea.

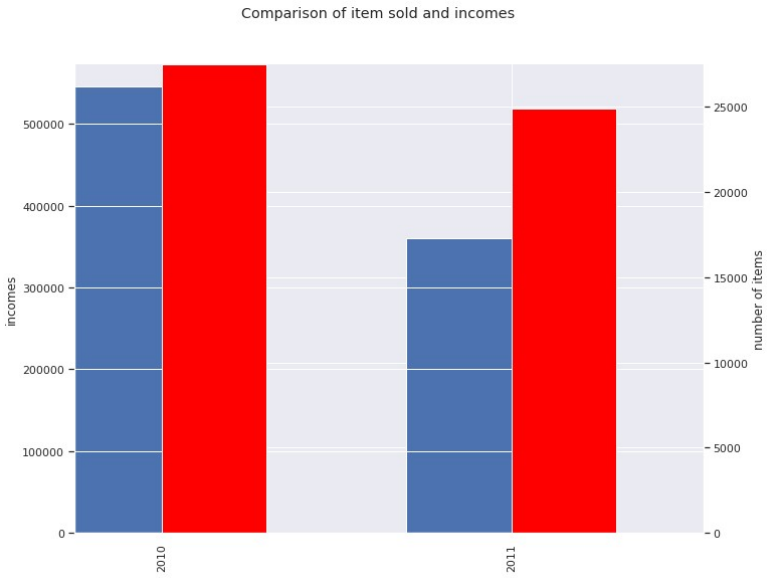
Datos de vendas comparativos por años



Gráfica4. Número de objetos vendido e ingresos de los años 2010 y 2011

No nos es posible comprar los datos de ventas de los dos años ya que no disponemos del año 2010 completo.

Al disponer solo del mes de diciembre del primer años vamos a proceder a comparar solo las ventas de ese mes.

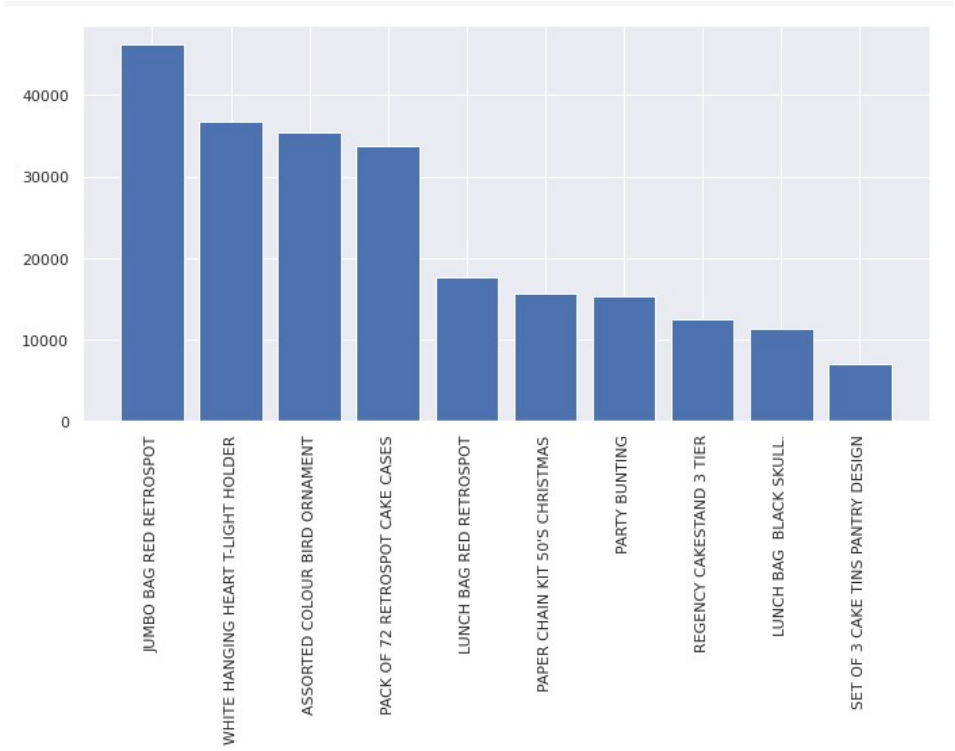


Gráfica5. Número de objetos vendido e ingresos en el mes de diciembre los años 2010 y 2011

Al comparar ambos meses, nos damos cuenta de que las ventas han disminuido el año 2011 y aunque la caída en el número de productos vendidos fue pequeña, la caída en los ingresos fue

muchisimo mayor y habrá que analizar los motivos de esta tendencia y establecer estrategias para revertirla.

Top10 Productos vendidos



Gráfica6. Top10 productos vendidos

Un simple vistazo a la gráfica anterior nos lleva a compobar que exiten 4 productos que se venden en grandes cantidad y sobrepasan a los demás en mucho.

Esto tendra utilidad cuando desarrollemos sistemas de clasificación y agrupamiento de clientes según los productos comprados

Conclusiones

El análisis cualitativo del negocio nos permite una primera aproximación a este, donde podemos establecer tanto los productos más vendidos como la estacionalidad de las ventas, lo que nos puede ayudar como primer paso para el desarrollo de modelos de analisis posteriores.

3. MARKET BASKET ANALYSIS

Preparación de datos

Este método de análisis de la cesta de la compra consiste en encontrar las relaciones existentes entre los distintos elementos o grupos de elementos, a la venta en un negocio, que se suelen comprar juntos.

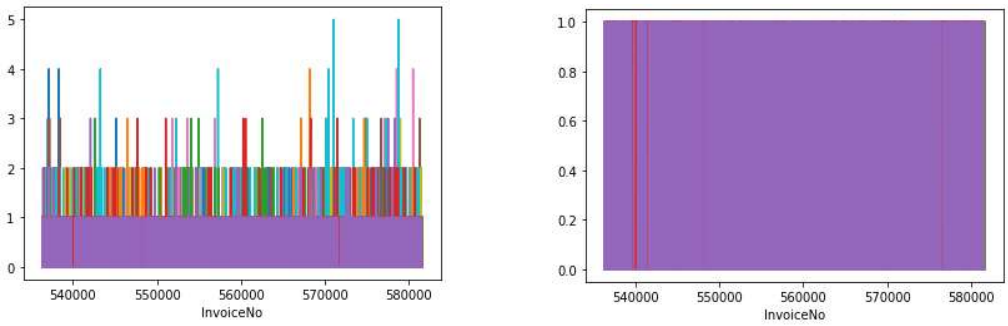
El método de cesta de la compra necesita saber los diferentes elementos que compra cada cliente en cada transacción. Para ello, utilizaremos el OneHotEncoder sobre los elementos a la venta de nuestro catálogo. Con este proceso convertimos una columna en múltiples, asignando como encabezado cada uno de los elementos en venta, y estableciendo que elementos ha comprado cada cliente en cada transacción.

	ALARM CLOCK BAKELIKE RED	ASSORTED COLOUR BIRD ORNAMENT	HEART OF WICKER SMALL	JAM MAKING SET WITH JARS	JUMBO BAG PINK POLKADOT	JUMBO BAG RED RETROSPOT	LUNCH BAG BLACK SKULL.	LUNCH BAG APPLE DESIGN	LUNCH BAG CARS BLUE	LUNCH BAG PINK POLKADOT	LUNCH BAG RED RETROSPOT	LUNCH BAG SPACEBOY DESIGN	LUNCH BAG SUKI DESIGN	NATURAL SLATE HEART CHALKBOARD	PACK OF 72 RETROSPOT CAKE CASES	PAPER CHAIN KIT 50'S CHRISTMAS	PARTY BUNTING	POSTAGE	REGENCY CAKESTAND 3 TIER	REX CASH+CARRY JUMBO SHOPPER	SET OF 3 CAKE TINS PANTRY DESIGN	SET OF 4 PANTRY JELLY MOULDS	SPOTTY BUNTING	WHITE HANGING HEART T-LIGHT HOLDER
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
197919	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
197920	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
197921	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
197922	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
197923	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Tabla5. Table OneHotEncoder

Al haber creado un elemento ‘other’ será necesario borrarlo para que no desvirtue las relaciones. Partiendo de aquí, agruparemos todos los datos por transacción y luego por cliente, para establecer la cesta de la compra de cada cliente en las distintas ocasiones a las que ha comprado.

Para aplicar este análisis no nos interesa cuantros productos de cada tipo a comprado el cliente por lo que habra que transformar todos las casillas de los productos vendidos a 0 y 1, ya que al agrupar por cliente estas se han sumado.



Gráfica7. Transformación de la gráfica del número de productos vendidos (por tipo de producto) a solo tipos de productos vendidos

Metodología

El método de la cesta de la compra se basa en reglas de asociación que tiene varias variables a tener en cuenta:

- support: esta variable establece la frecuencia con la que un elemento aparece en los datos
- antecedents: que es el elemento, o grupo de elementos, del que partimos para crear la asociación
- consequents: que es el elemento al que llegamos tras la asociación
- métrica lift: es la métrica que vamos a utilizar y tiene la ventaja de que los antecedents y los consequents son intercambiables

Empecemos creado la tabla support:

	support	itemsets
0	0.047313	(WHITE HANGING HEART T-LIGHT HOLDER)
1	0.074180	(REGENCY CAKESTAND 3 TIER)
2	0.051845	(JUMBO BAG RED RETROSPOT)
3	0.047421	(ASSORTED COLOUR BIRD ORNAMENT)
4	0.046990	(PARTY BUNTING)
...	...	...
99	0.011006	(SPOTTY BUNTING, LUNCH BAG SPACEBOY DESIGN, JA...
100	0.013541	(PACK OF 72 RETROSPOT CAKE CASES, SPOTTY BUNTI...
101	0.013595	(PACK OF 72 RETROSPOT CAKE CASES, LUNCH BAG SP...
102	0.012354	(SPOTTY BUNTING, LUNCH BAG SPACEBOY DESIGN, PA...
103	0.010412	(PACK OF 72 RETROSPOT CAKE CASES, JAM MAKING S...

Tabla6. Frecuencia en la que aparece cada producto medido según el modelo support

Al crear la regla de asociación establecemos un threshold minimo de 1, ya que tenemos una cantidad de datos limitada, al haber agrupado repetidamente los datos. El threshold igual a 1 significa que la posibilidad de que dos elementos se compren a la vez es igual al azar

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(REGENCY CAKESTAND 3 TIER)	(LUNCH BAG PINK POLKADOT)	0.074180	0.074450	0.010412	0.140364	1.885348	0.004889	1.076677
1	(LUNCH BAG PINK POLKADOT)	(REGENCY CAKESTAND 3 TIER)	0.074450	0.074180	0.010412	0.139855	1.885348	0.004889	1.076354
2	(REGENCY CAKESTAND 3 TIER)	(LUNCH BAG SUKI DESIGN)	0.074180	0.091929	0.012462	0.168000	1.827493	0.005643	1.091431
3	(LUNCH BAG SUKI DESIGN)	(REGENCY CAKESTAND 3 TIER)	0.091929	0.074180	0.012462	0.135563	1.827493	0.005643	1.071010
4	(REGENCY CAKESTAND 3 TIER)	(JAM MAKING SET WITH JARS)	0.074180	0.106334	0.014620	0.197091	1.853515	0.006732	1.113036
...	...	...	...	...	...	...	...	...	...
265	(PAPER CHAIN KIT 50'S CHRISTMAS, SET OF 3 CAKE...	(PACK OF 72 RETROSPOT CAKE CASES, JAM MAKING S...	0.027892	0.023036	0.010412	0.373308	16.205219	0.009770	1.558921
266	(PACK OF 72 RETROSPOT CAKE CASES)	(JAM MAKING SET PRINTED, PAPER CHAIN KIT 50'S ...	0.050227	0.014135	0.010412	0.207304	14.666361	0.009702	1.243686
267	(JAM MAKING SET PRINTED)	(PACK OF 72 RETROSPOT CAKE CASES, PAPER CHAIN ...	0.052115	0.016616	0.010412	0.199793	12.023904	0.009546	1.228912
268	(PAPER CHAIN KIT 50'S CHRISTMAS)	(PACK OF 72 RETROSPOT CAKE CASES, JAM MAKING S...	0.069486	0.014027	0.010412	0.149845	10.682776	0.009438	1.159757
269	(SET OF 3 CAKE TINS PANTRY DESIGN)	(PACK OF 72 RETROSPOT CAKE CASES, JAM MAKING S...	0.056754	0.015052	0.010412	0.183460	12.188588	0.009558	1.206246

Tabla7. Parámetros obtenidos después de aplicar las reglas de asociación del modelo de cesta de la compra

Para establecer cual son los elementos con relaciones de asociación más fuertes, filtramos nuestros datos según la métrica lift.

Este filtro nos dejaría una cantidad muy grande de datos por lo cual aplicamos un segundo filtro con la métrica confidence, que nos dice la probabilidad que hay de que al comprar un objeto se compre otro.

En esta probabilidad el orden no es intercambiable y podemos encontrar un segundo elemento con una alta probabilidad de compra si ya se ha comprado un primer objeto, pero si ese segundo elemento es el primero que se compra, no existir dicha probabilidad de compra del primero.

```
rules[ (rules['lift'] >= 6) &
       (rules['confidence'] >= 0.7) ]
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
244	(PACK OF 72 RETROSPOT CAKE CASES, LUNCH BAG SP...	(PAPER CHAIN KIT 50'S CHRISTMAS)	0.019314	0.069486	0.013595	0.703911	10.130192	0.012253	3.142678
257	(PACK OF 72 RETROSPOT CAKE CASES, JAM MAKING S...	(PAPER CHAIN KIT 50'S CHRISTMAS)	0.014027	0.069486	0.010412	0.742308	10.682776	0.009438	3.610948
259	(JAM MAKING SET PRINTED, PAPER CHAIN KIT 50'S ...	(PACK OF 72 RETROSPOT CAKE CASES)	0.014135	0.050227	0.010412	0.736641	14.666361	0.009702	3.606386

Tabla8. Tabla de objetos altamente relacionados

Estableciendo un lift mayor que 6 y una confidence mayor que 0.7 encontramos relaciones muy fuertes entres 3 grupos de objetos.

Conclusiones

- este método nos sirve para establecer relaciones entre objetos en venta
- este nos puede llevar a la modificación de la distribución de la tienda o del catálogo de nuestro negocio, poniendo elementos muy relacionados en los extremos de la tienda o del catálogo.
- también nos puede servir para desarrollar packs con elementos muy relacionados
- En el caso concreto que hemos analizado podemos observar que existe una relación muy estrecha entre los distintos objetos de fiesta

4.ANÁLISIS RFM

El método de análisis RFM es un sistema de análisis por clasificación de clientes a partir de tres variables que son:

- dinero gastado por el cliente
- la frecuencia de compra
- última vez que compró

Para desarrollar este método ignoraremos que objetos ha comprado cada cliente y nos centraremos únicamente en cuanto se gastó.

Tanto la frecuencia y el dinero gastado por un cliente dependen en muchas ocasiones de cuando fue la primera vez que el cliente realizo una compra.

Por ello tendremos que establecer los valores relativos de estas dos variables, ya que los absolutos conducirían a una agrupación defectuosa.

Los pasos para hacer esto serían establecer el último dato de compra como el día de referencia (ignoraremos las horas y nos quedaremos solo con los días ya que si no se generarían

demasiadas variables) y estableceremos los días totales que han pasado desde la primera compra y el día de referencia.

Por otra parte, sumamos todo el dinero gastado por cada cliente en total, posteriormente contaríamos las veces que cada cliente a usado nuestra página y por último se establecen el número de días que han pasado desde que cada cliente hizo su última compra. Esto daría lugar a la siguiente tabla:

	CustomerID	PrecioTotal	frecuency	NofDay	seniority
0	12346.0	77183.60	1	326	326
1	12347.0	4310.00	7	3	368
2	12348.0	1797.24	4	76	359
3	12349.0	1757.55	1	19	19
4	12350.0	334.40	1	311	311
...	...	...	...	...	...
4334	18280.0	180.60	1	278	278
4335	18281.0	80.82	1	181	181
4336	18282.0	178.05	2	8	127
4337	18283.0	2094.88	16	4	338
4338	18287.0	1837.28	3	43	202

Tabla9. Tabla de RFM en valores absolutos

La anterior tabla muestra los datos de los clientes en valor absoluto, pero como ya hemos dicho previamente, tanto la frecuencia como la cantidad de dinero hay que relativizarlas a partir de la antigüedad de cada cliente. Para ello hemos optado por dividir la frecuencia y el dinero por el número de días que lleva cada cliente usando nuestros servicios.

	CustomerID	PrecioDia	FrecuencyRel	NofDay
0	12346.0	236	0.003067	326
1	12347.0	11	0.019022	3
2	12348.0	5	0.011142	76
3	12349.0	92	0.052632	19
4	12350.0	1	0.003215	311
...	...	...	...	...
4334	18280.0	0	0.003597	278
4335	18281.0	0	0.005525	181
4336	18282.0	1	0.015748	8
4337	18283.0	6	0.047337	4
4338	18287.0	9	0.014851	43

Tabla10. Tabla de RFM en valores relativos

Estos son los valores que vamos a utilizar para aplicar el modelo.

Metodología

Este método se basa en dividir cada una de las tres variables en percentiles, en este caso usaremos cuantiles.

Este será nuestro siguiente paso.

	CustomerID	PrecioDia	FrecuencyRel	NofDay	NofDayScore	FrecuencyRelScore	PrecioDiaScore
0	12346.0	236.759509	0.003067	326	4	4	1
1	12347.0	11.711957	0.019022	3	1	2	1
2	12348.0	5.006240	0.011142	76	3	3	2
3	12349.0	92.502632	0.052632	19	2	1	1
4	12350.0	1.075241	0.003215	311	4	4	4
...	...	...	...	...	...	...	...
4334	18280.0	0.649640	0.003597	278	4	4	4
4335	18281.0	0.446519	0.005525	181	4	4	4
4336	18282.0	1.401969	0.015748	8	1	2	4
4337	18283.0	6.197870	0.047337	4	1	1	2
4338	18287.0	9.095446	0.014851	43	2	3	2

Tabla11. Tabla de RFM con percentiles

Existen dos métodos de analizar estos resultados:

- Damos mayor prioridad a una variable que a otra. Por ejemplo, nos importa más el dinero gastado por un cliente que cuando fue la última vez que compró. El método recomienda usar la variable recencia como la más importante, seguida de la frecuencia y finalmente el dinero.
- Otra forma seria sumar todos los valores de los cuantiles RFM sin dar prioridad a unos o a otros. Este método nos establecerá una puntuación, en este caso de 3 a 12, donde los clientes con números más bajos serán teóricamente mejores para nuestra empresa.

Nosotros añadiremos un tercero basado en ML:

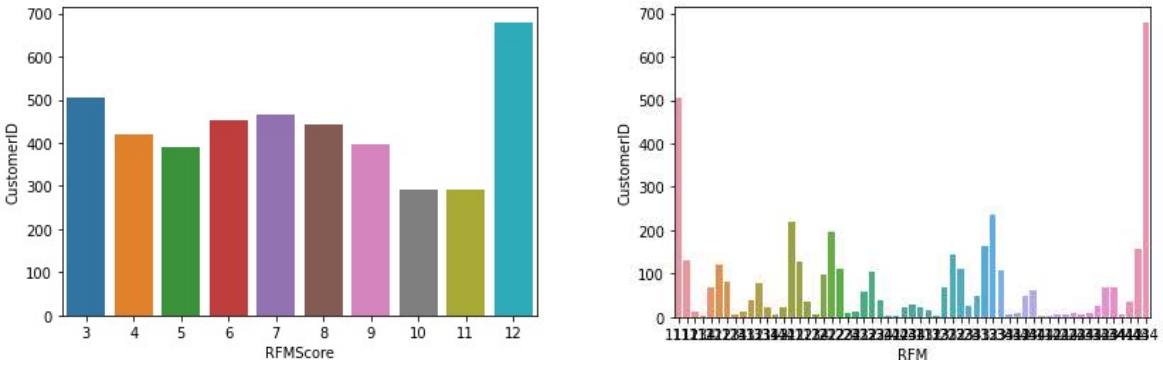
- O clasificamos a los clientes atendiendo a todas las características RFM relativas usando algún sistema de Unsupervised Machine Learning

Vamos a llevar a cabo primero los dos métodos propios del RFM, para ello estableceremos dos columnas, en la primera de tomarán los percentiles como string y se unirán, y en la segunda como int y se sumarán.

	CustomerID	PrecioDia	FrecuencyRel	NofDay	RFM	RFMScore
1118	13860.0	14.320690	0.057471	2	111	3
1502	14397.0	12.210093	0.088785	3	111	3
529	13040.0	19.911471	0.058824	9	111	3
2922	16333.0	85.892903	0.070968	8	111	3
3834	17595.0	10.231316	0.052632	13	111	3
...	...	...	...	...	...	...
826	13453.0	1.230966	0.006897	166	444	12
2607	15889.0	1.207757	0.005405	157	444	12
3566	17223.0	1.156612	0.005420	311	444	12
818	13439.0	1.108242	0.003906	256	444	12
1927	14981.0	0.413441	0.004049	247	444	12

Tabla12. Tabla de RFM final

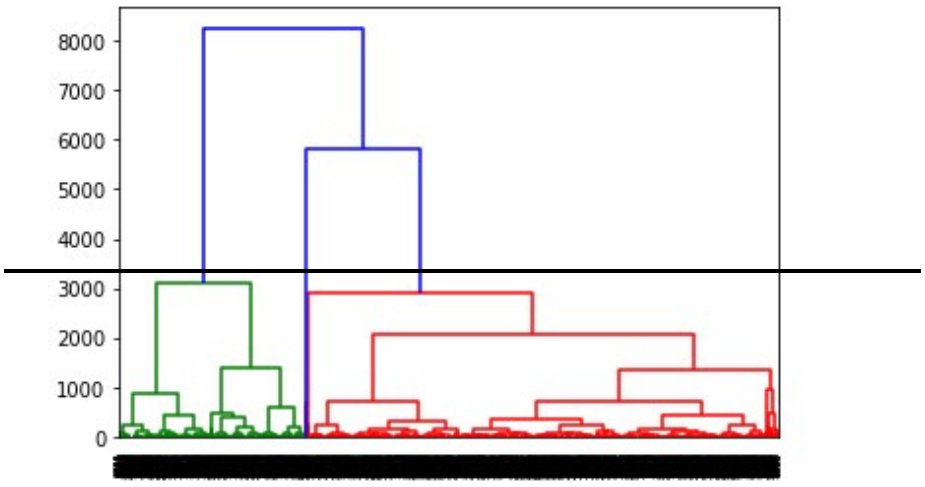
La primera variable nos establecería 222 subgrupos distintos de clientes, mientras que la segunda creará solamente 9:



Gráfica8. Número de clientes por subgrupo y método RFM

Por último, crearos un modelo de ML no supervisado utilizando los datos del RFM. Para ello usaremos una agrupación jerárquica, modelo que se explicará más detalladamente más adelante.

Para llevar a cabo este análisis usaremos los datos con las variables relativas, no los percentiles ni los grupos RFM.



Gráfica9. Dendograma aplicado al método RFM

Conclusiones:

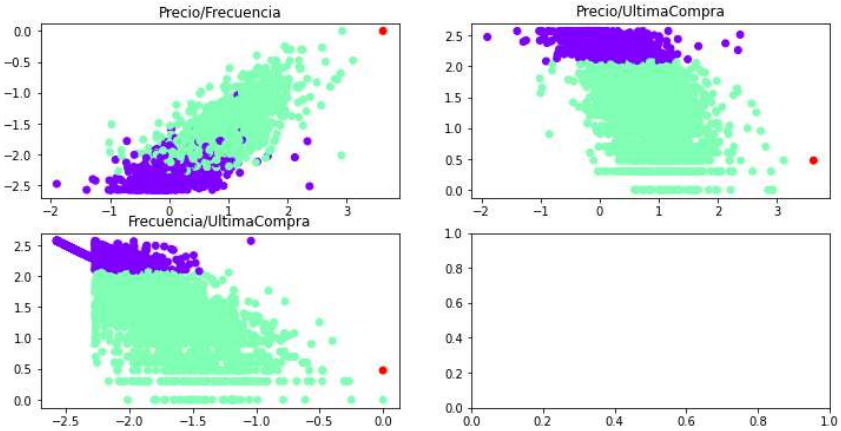
Método RFM:

- El mètodo de preferencia de las variables RFM nos da un modelo con demasiados subgrupos, lo que hace muy dificil de clasificar a los clientes.
- Además, los resultados de este método variarian según el modelo de negocio que queramos implementar, cambiando el orden de las variables RFM
- El mètodo de unión de las variables RFM nos genera una cantidad más de subgrupos más reducidad pero lleva a la simplificación excesiva de los mismos.

- La primer conclusión valida que generan ambos métodos es que tenemos un gran conjunto de clientes que no se gastan mucho dinero, compran con poca frecuencia y no lo hacen desde hace mucho tiempo
- Y la segunda, es que por el contrario, tenemos una cantidad bastante importante de clientes fidelizados, que gastan dinero, con frecuencia y hace poco tiempo

Método UL ML

Para obtener conclusiones de este método es necesario graficar los diferentes clusters. A partir de ellos se llegan a las siguientes conclusiones:



Gráfica10. Clusters producidos por el modelos de agrupación jerárquica

- Aunque podríamos elegir cualquier otro número de clusters, la grafica nos lleva a pensar que eligiendo 3 crearemos unos subconjuntos bien diferenciados.
- tenemos unos pocos clientes (rojo) que han comprado recientemente, se han gastado una cantidad grande de dinero y lo hacen con frecuencia. Aunque la relación entre frecuencia y ultima compra puede estar contaminada si el usuario solo ha realizado una compra.
- El grupo más popular (verde) responde a todos los que tienen las tres caracterísa o en niveles bajos o medios
- Y el último grupo (morado) tiene una gran frecuencia en la compra y han comprado en los últimos días, pero el dinero no les influye en gran medida

Comparación de los dos métodos:

Aunque el método RFM puede ser de gran utilidad en los extremos de los datos, genera mucha incertidumbre en los datos medios. Por otra parte, su fácil implementación y su sencillez hace que sea útil para un primer análisis.

Por otro lado, el método de agrupación jerárquica no lleva a datos más precisos, pero más difíciles de entender, y conllevaran, probablemente, la necesidad de un mayor entendimiento de negocio para su aplicación.

8.5 APRENDIZAJE NO SUPERVISADO ML

En este apartado vamos a establecer los métodos de agrupación basados en aprendizaje no supervisados, que son métodos propios del análisis de datos basados en machine learning.

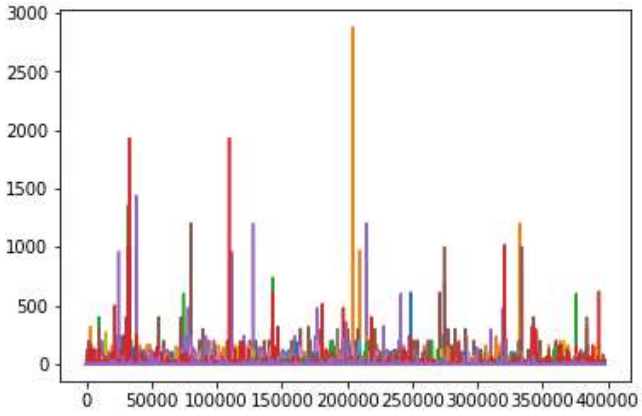
Desarrollaremos diversos modelos de aprendizaje no supervisado para intentar clasificar a nuestros clientes en diferentes grupos con características similares. Con este sistema podremos entender mejor como actúa el conjunto de nuestros clientes y desarrollar acciones más eficientes

Lo primero que se hará se un OneHotEncoder sobre los datos de los productos que nos convertirá esta columna en otras con variables categóricas.

Para nuestro análisis solo tendremos en cuenta el gasto del cliente y los productos comprados

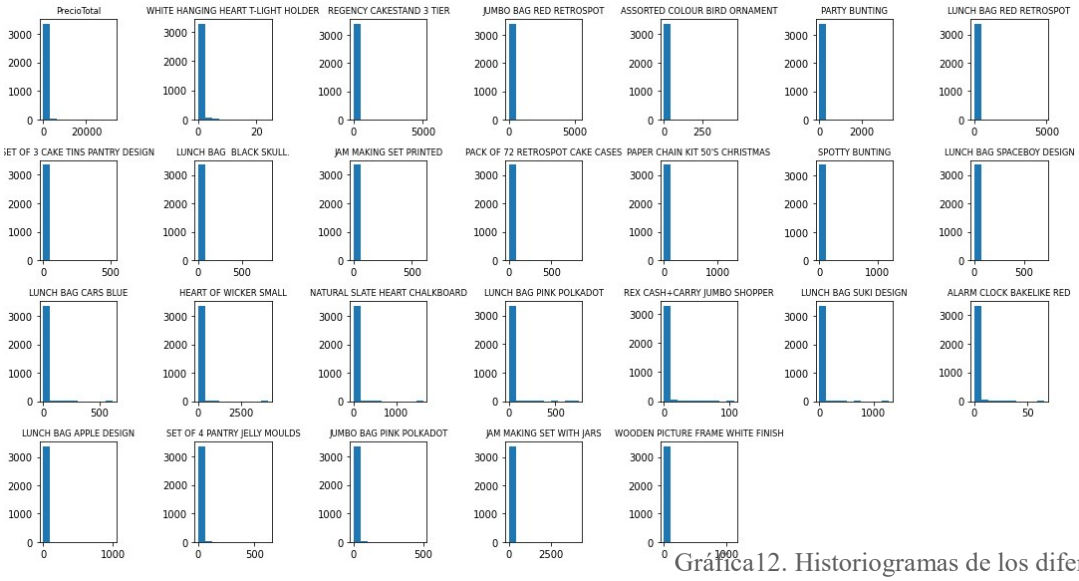
Al hacer un análisis visual de las variables comprobamos que hay un número pequeño de clientes que compran grandes cantidades, mientras que los demás compran poco,

Esto nos va a producir unas diferencias de peso muy importante que luego habrá que corregir.



Gráfica11. Número de objetos comprados por cliente

En el historiograma por columna podemos ver que existen en todos los casos valores muy extremos que nos van a generar clusters muy pequeños y alejados



Gráfica12. Historiogramas de los diferentes productos

Kmeans

El primer método a utilizar va a ser el de KMeans.

Es un método computacionalmente muy sencillo y que tiene una gran versatilidad, pudiéndose ser utilizado en prácticamente cualquier conjunto de datos. Por eso será nuestra primera opción.

Los dos mayores inconvenientes de este método son:

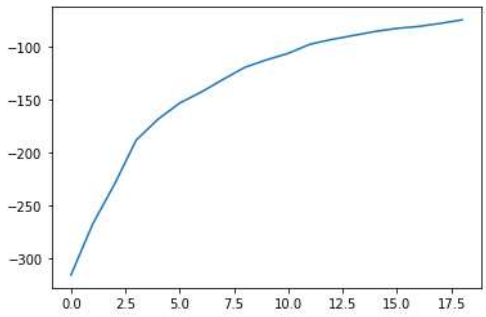
- Hay que elegir los clusteres antes de llevarlo a cabo
- No es un método estable, al repetir el proceso te pueden dar resultados diferentes

Este método se ha llevado a cabo con 3 dataset distintos para ver los diferentes resultados que producirían.

1. Con los datos sin modificar.
2. Eliminando datos extremos. Esta opción sería eliminando valores extremos en los datos, que desvirtúan los resultados. En este caso se podría eliminar a los grandes clientes y quedarnos solo con la mayoría que se mantendría cerca de la media, ya sea por arriba o por abajo
3. Usando PCA para disminuir la dimensionalidad.

1. Datos sin modificar

El primer paso sería crear una gráfica siguiendo el método del codo para ver cuantos clusters serían los idóneos para usar en nuestro calculo. Este método utiliza los valores de la inercia obtenidos tras aplicar el K-means a diferente número de Clusters (desde 1 a N Clusters), siendo la inercia la suma de las distancias al cuadrado de cada objeto del Cluster a su centroide. Al representar esto gráficamente:

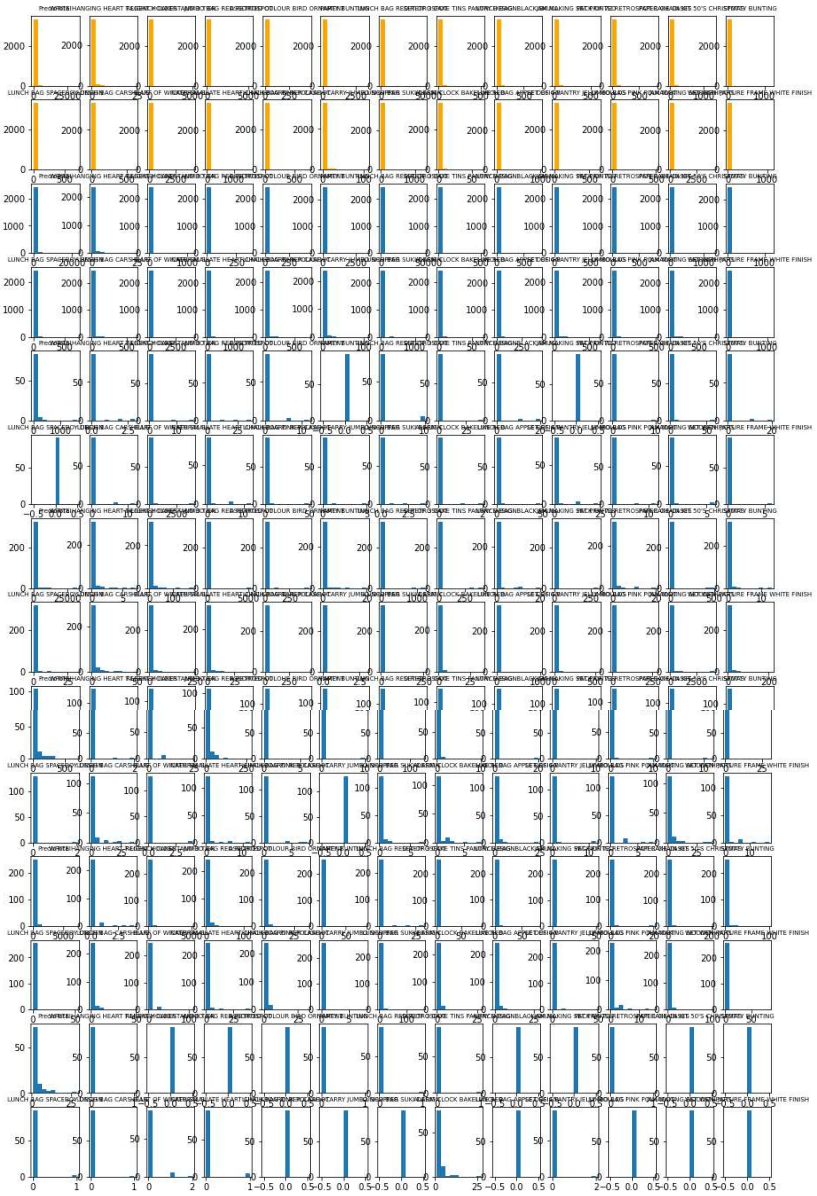


Gráfica13. Método del codo aplicado a los datos sin modificar

En este caso en concreto no hay un codo muy claro, cogeremos 6 clusters para hacer nuestro cálculo.

Entrenamos el modelo Kmeans con 6 clusters usando. Para posteriormente predecir en cual de esos clusters ira cada uno de nuestros clientes.

Al representar esto gráficamente obtenemos:



En esta figura se puede ver en amarillo los datos originales mientras que los siguiente en azul, cada dos lineas, son los nuevos clusters cleados.

Se pude deducir de esto que, exceptuando el último cluster que tiene valores muy altos, va a depender del número de objetos comprados en de 1 a 3 elementos, la pertenecia o no a un cluster.

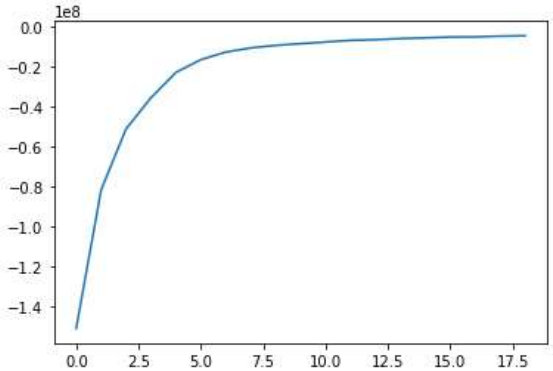
Podemos observar que el dinero no influye en ninguno de los casos.

Gráfica14. Historiogramas de los clusters Kmeans con datos originales

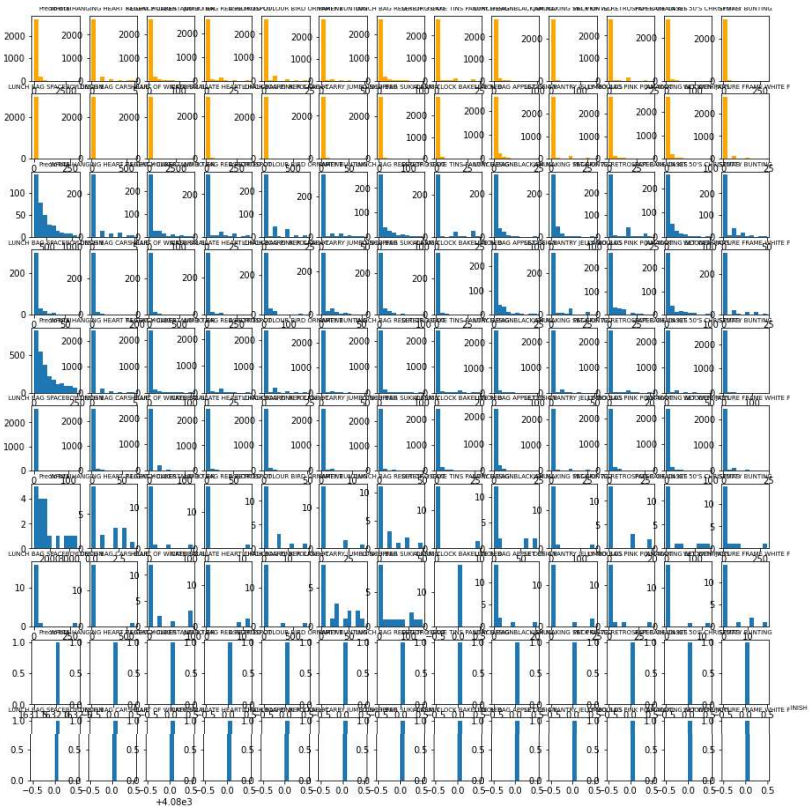
2. Eliminando valores extremos

Se repite el mismo método desarrollado anteriormente paso a paso, pero eliminando a los 400 clientes que tienen los datos más extremos,

Con el método del codo no damos cuenta que en esta ocasión 4 clusters sería la decisión más acertada:



Gráfica15. Método del codo aplicado a datos sin valores extremos



En esta ocasión nos encontramos historiogramas más complejos, donde la distribución entre las diferentes variables también es más compleja.

El último cluster, al igual que la última vez, contiene los valores más extremos

En este caso, el gasto de los clientes si que influye en su pertenencia a uno u otro cluster

Gráfica16. Historiogramas de los clusters Kmeans con datos sin valores extremos

3. Reducciòn de la dimensionalidad PCA

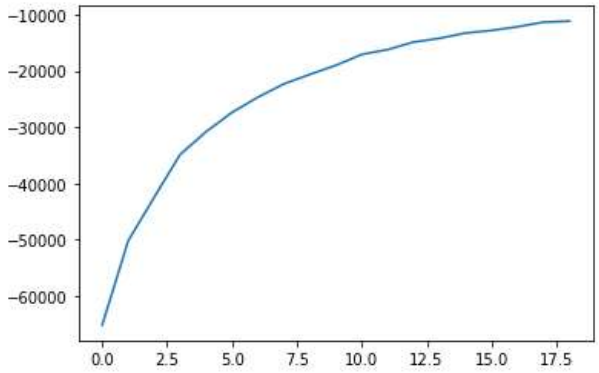
El tener tantas variables nos lleva generar problemas en los modelos. Una de las posibles soluciones es reducir la dimensinalidad de los datos estableciendo cuales tienen más peso en los modelos. Para eso se utiliza PCA.

En este caso aplicaremos PCA a los datos completos.

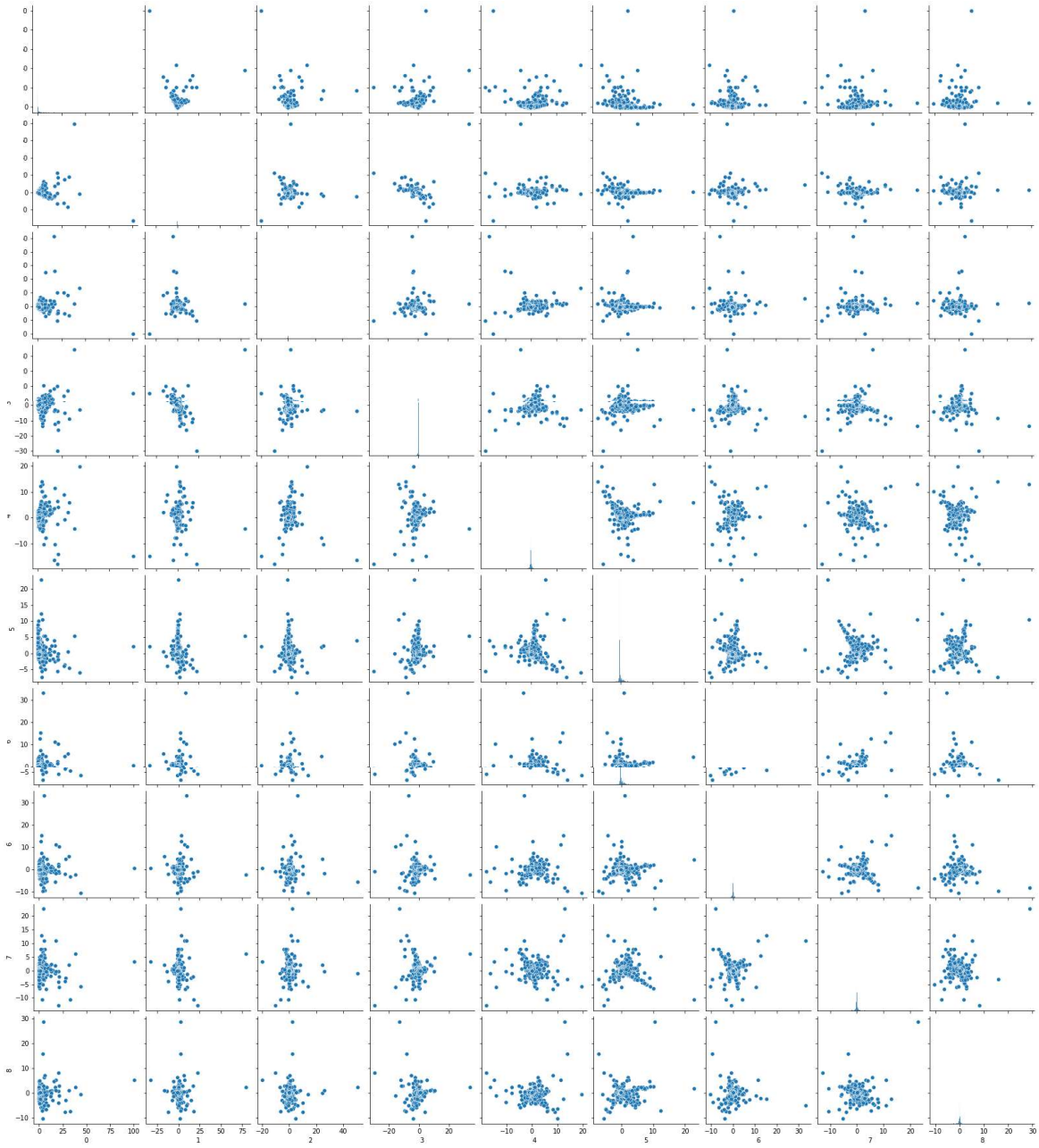
Con esta formula pordemos observar que si reducidos el número de variables a 9, abarcamos cerca del 80% de la varianza de nuestros datos, reduciendo cerca de dos tercios las variables.

Estas nuevas variables se pueden ver gráficamente en la figura Gráfica18

Deberemos volver a aplicar el metodo del codo para establecer cual es el número de cluster indóneo para este modelo. En este caso vuelve a no quedar muy claro pero eligiremos 6 clusters para nuesto modelo Kmeans.



Gráfica17. Método del codo aplicado a datos PCA



Gráfica18. Gráficas que relacionan las 9 variables PCA

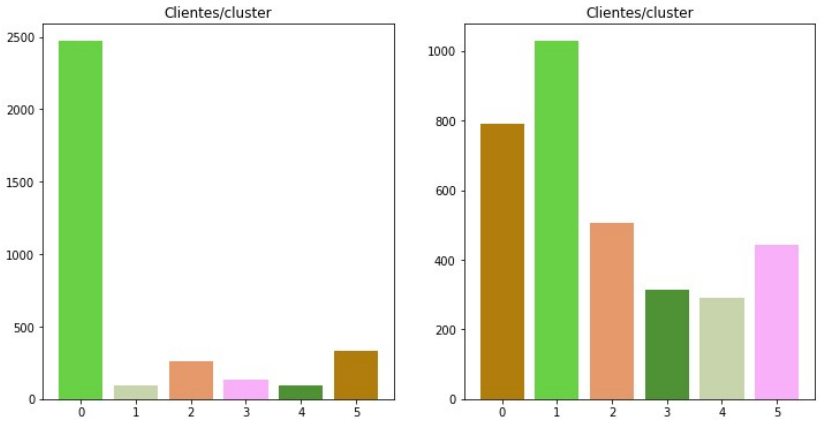
En este caso, el resultado de todo estre proceso nos dara una tabla que establecerá en que propoción pertenece cada uno de nuestros clientes a cada una de estas nuevas 9 variables.

HEART OF ICKER SMALL	NATURAL SLATE HEART CHALKBOARD	LUNCH BAG PINK POLKADOT	REX CASH+CARRY JUMBO SHOPPER	LUNCH BAG SUKI DESIGN	ALARM CLOCK BAKELIKE RED	LUNCH BAG APPLE DESIGN	SET OF 4 PANTRY JELLY MOULDS	JUMBO BAG PINK POLKADOT	JAM MAKING SET WITH JARS	WOODEN PICTURE FRAME WHITE FINISH	Componente1	Componente2	Componente3	Componente4	Componente5	Componente6	Componente7	Componente8	Componente9	SegmentosKmeansPCA
0.0	0.0	0.0	0.0	15.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.219557	0.199460	-0.318677	-1.214606	0.691730	2.156893	-0.703909	0.500812	-2.113574	0
0.0	0.0	0.0	9.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.393975	-0.111440	-0.207087	-0.079547	-0.711282	0.032387	-0.707616	0.822196	-0.380471	0
0.0	0.0	0.0	1.0	1.0	0.0	3.0	0.0	0.0	0.0	0.0	-0.427714	0.128851	-0.043761	0.101487	-0.285594	-0.246846	-0.061658	0.025679	0.044783	0
0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.607347	-0.108886	-0.035978	0.157077	-0.214023	-0.237193	-0.083292	0.049394	0.029901	0
0.0	0.0	0.0	7.0	5.0	0.0	6.0	12.0	3.0	0.0	0.0	-0.206418	0.232469	-0.234566	-0.134755	-0.676819	-0.075055	-0.577000	0.493100	-0.262432	0

Tabla13. Clientes asociados a las variables PCA

4.Conclusiones

Modelos sin PCA



Gráfica19. Número de clientes por cluster en Kmeans

En las tablas anteriores podemos ver que, en la primera, un único cluster agrupa a prácticamente todos los clientes, mientras que en la segunda, aun habiendo un cluster mucho más grande que el otro, la distribución es más homogénea.

Esto nos lleva a tomar la decisión de que el segundo modelo es mejor que el primero, aunque habría que entender el modelo de negocio que se quiere implementar ya que en el segundo caso al reducir el número de clientes estamos perdiendo información.

Modelo PCA

el modelo con PCA nos reduce la dimensionalidad de los datos iniciales y nos permite clasificar a los clientes siguiendo unas nuevas categorías generadas por el modelo. Esto nos permite reducir la contaminación en los resultados, pero en el caso que tratamos hace más complicado relacionar a un cliente en concreto con un objeto en concreto

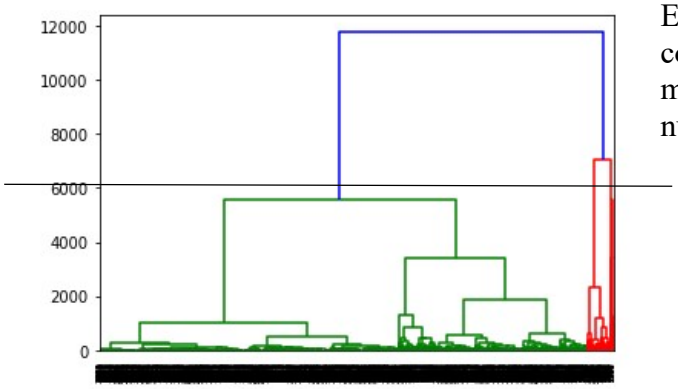
Agrupación jerárquica

En este modelo no hace falta determinar el número de clústeres. Simplemente va agrupando los distintos clientes sucesivamente según sus características hasta que todos pertenecen a un único grupo.

Vamos a utilizar un modelo aglomerativo para ello.

Los pasos que realizaremos serán los siguientes.

1. Entrenaremos dicho modelo aglomerativo con los datos (en este caso sin transformaciones) sin establecer un número de clústeres
2. Dibujaremos el dendograma con los datos predichos por el modelo anterior
3. Decidiremos el número de clúster que nos encajen según el dendograma anterior. Si no tenemos un objetivo previo, trazaremos una línea por la parte mas baja de la recta vertical más larga del dendograma, y el número de líneas que corte será nuestro número de clústeres.
4. Reentrenaremos el modelo inicial pero esta vez estableciendo notros en número de clúster



En nuesto caso, si establecemos la linea de corte por la parte más baja de la recta vertical más larga nos sale que tenemos que entrenar nuestro modelo con 3 clústeres.

Gráfica20. Dendograma del método de agrupación jerárquica

Conclusión

Este método, que nos permite establecer visualente el número de clusteres o elegir nosotros uno, es uno de los métodos más efectivos para agrupar estos datos.

Además, es un método estable que nos permite reproducir los dato en otro contexto

DBSCAM

DBSCAN es un algoritmo de agrupamiento basado en la densidad de puntos. Comienza en un punto al azar y cuenta todos los puntos que están dentro de un radio dado. Si el número de puntos es superior a un número predeterminado considera que esos puntos pertenecen al mismo cluster, Progresivamente el centro de la circunferencia se va moviendo de punto a punto hasta que todos los puntos han sido metidos en un cluster o se han considerado ruido si el número de puntos vecinos era inferior al número predeterminado

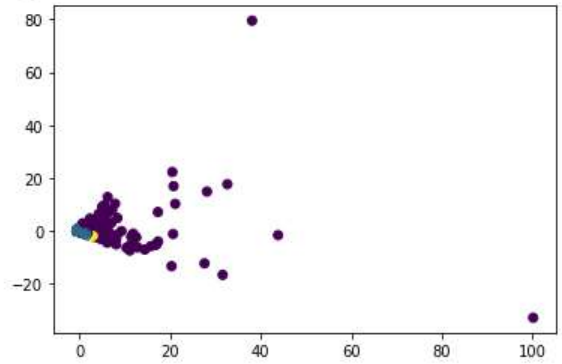
Siguiendo lo anterior, necesitaremos establecer tanto el radio elegido (épsilon) como el número mínimo de objetos.

Haciendo varias pruebas con nuestros datos hemos establecido que, con número mínimo de 10 objetos, el mayor ep=0.7. Si superamos ese número todos nuestros elementos se unen en un solo clúster.

Se puede apreciar en la gráfica que más de 450 clientes han sido etiquetados como ruido.

En este caso, podríamos utilizar los datos sin cliente extremos, normalizándolos, para intentar evitar a todos esos clientes-ruido

```
 0    2891
-1     452
 1      26
 2      11
dtype: int64
```



Gráfica21. Gráfica DBSCAM

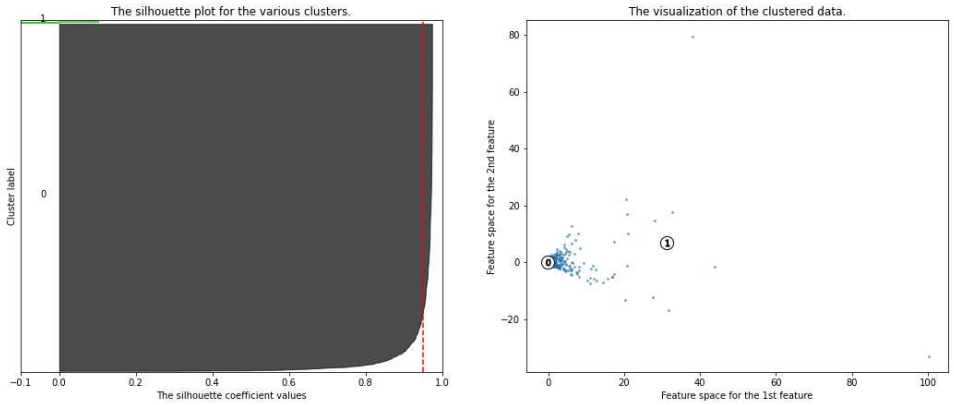
Conclusiones

Este método nos permite excluir objetos de nuestros datos que estén en los extremos. Esto puede ser bueno en ciertos contextos y malo en otro, Será decisión del individuo, dentro del conocimiento del negocio, si es aplicable o no a sus intereses.

Análisis silhouette

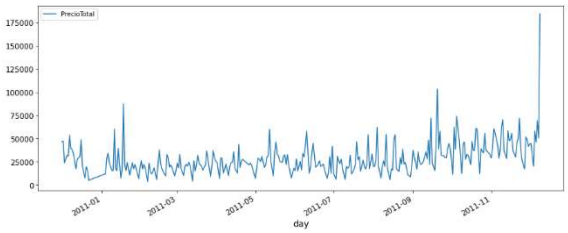
Este análisis puede ser usado para estudiar la distancia de separación de los clústeres que hemos establecido. Este análisis muestra una medida de lo cercano que está cada punto de un clúster de los puntos de los clústeres vecinos y proporciona una forma visual de evaluar parámetros como el número de clústeres. Las medidas de distancia se establecen entre el -1 y el 1. Siendo todo lo que esté por debajo del 0 considerado un valor atípico.

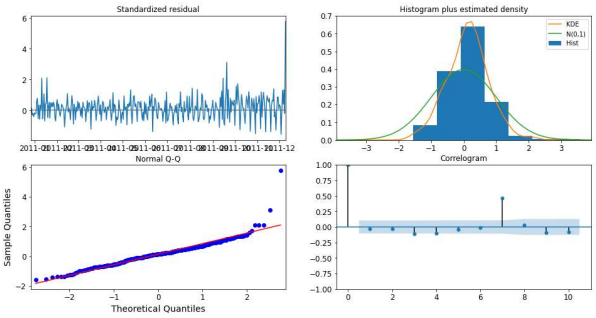
Si evaluamos nuestro kmeans, usando como dataset los datos con reducción de dimensionalidad PCA, obtendremos que tenemos una puntuación de silhouette por encima de 0.9 con 1, 2 o 3 cisteres

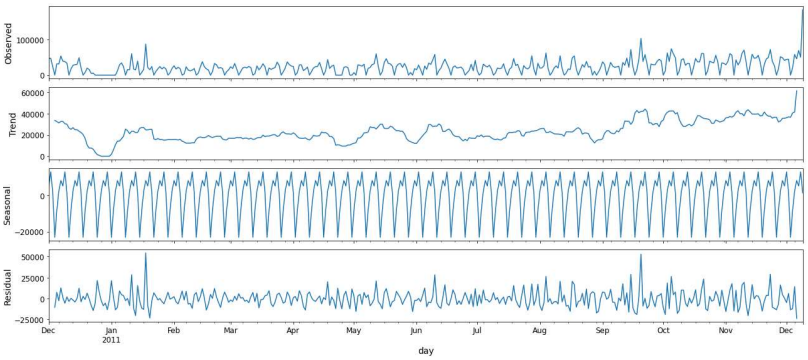


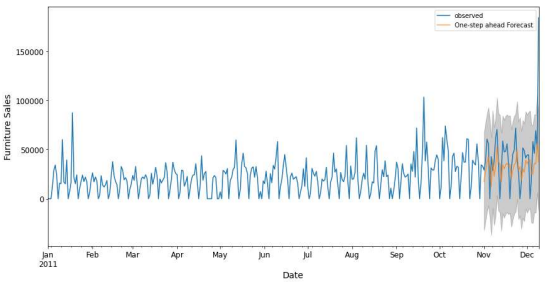
Gráfica22. Gráfica Silhoutte

8.6 PREDICIÓN DE INGRESOS. SERIE TEMPORAL. SARIMA


PrecioTotal
day


Standardized residual
Histogram plus estimated density
KDE
N(0,1)
Hist
Normal Q-Q
Correlogram
Sample Quantiles
Theoretical Quantiles


Observed
Trend
Seasonal
Residual
day


observed
One-step ahead Forecast
Furniture Sales
Date

Referencias

Apellidos, n. s. (Año). Título del artículo. *Título del diario*, Páginas desde - hasta.

Apellidos, n. s. (Año). *Título del libro.* Nombre de la ciudad: Nombre del editor.