Universidad del Valle de Guatemala
Data Science
Catedrática: Lynette Garcia Perez

Proyecto 2: Análisis Exploratorio

Grupo 5:

Maria Fernanda Rodas	17125
Pablo Viana	16091
Sergio Marchena	16387
Daniel Ixcoy	16748
José Martínez	15163
José Meneses	15140

Guatemala 28 de Agosto de 2019

Introducción

Lancasco es la primera compañía de la Industria Químico-Farmacéutica de Centroamérica, fundada en 1927. Se divide en dos unidades de negocios con la producción y comercialización de productos farmacéuticos de prescripción médica, venta libre (vitaminas, productos naturales y suplementos alimenticios) y genéricos, así como la producción y comercialización de productos del cuidado de la piel, salud y belleza que se comercializan por medio de venta directa. En este último renglón se encuentra Scentia. Esta empresa ha proporcionado sus datos para que se realicen análisis y se creen modelos de proyección de ventas.

Módulo a trabajar

Análisis de Ventas Directas en Centroamérica.

Situación actual de la Empresa

Scentia actualmente recopila todos los datos de los Catálogos de Ventas Directas por país. Además, guardan las unidades por sector en cada país y la paginación de sus productos. Con toda esta información histórica, Scentia cuenta actualmente con un modelo de predicción para poder saber la demanda de sus productos. Este modelo predictivo fue hecho hace aproximadamente 5 años, por lo que la precisión es muy baja, concluyendo que dicho modelo ya no es útil para la empresa. Scentia quiere hacer un nuevo modelo predictivo con una exactitud en las respuestas mayor a 45%. De este modo se podrá evitar el inventario que no es vendido, sino solamente genera un coste por almacenamiento.

Problema a resolver

La venta directa consta de una oferta y demanda muy dinámica, es decir, los compradores y vendedores que hoy están, mañana pueden ya no estarlo. Por lo tanto el manejo del inventario que se debe de tener mes a mes de cada artículo del catálogo es muy complicado. Por tanto, se intenta generar un modelo adecuado para poder predecir de la mejor forma la demanda que habrá cada mes tomando en cuenta estacionalidades de algunos de los productos que hay en el catálogo.

Objetivos

- Identificar las variables que resultan importantes para poder generar un modelo adecuado para la predicción de la demanda.
- Generar un modelo que haga predicciones con una exactitud mayor al 45% de la demanda mensual.

Estado de los datos

Los conjuntos de datos de Catálogos estaban separados por país, Guatemala, El Salvador, Nicaragua y Honduras; el primer paso fue unir estos datos para poder tener el dataset de Centroamérica. Al momento de leer los datos del archivo xlsx, R realizó una autocompletación de los nombres de las columnas que estaban equivocados, se corrigieron a los nombres correctos.

Los datos de todos los países estaban sucios. Había faltas de ortografía en todos los países y casi que en todas las variables. Por ejemplo:

- En Año Mes, había una observación de la siguiente manera: "210801" en vez de "201801"
- Había casillas en blanco, casillas con 0 y casillas con NULL.
- Se quitaron los caracteres "\r|\n" de la columna descripción.
- En la columna contingencia se encontraba "Farma" y "farma", se procedió a unificarlos.
- En la columna Pagina se encontraba "Pagina derecha" y "pagina derecha", además "Pagina izquierda" y "pagina izquierda". Se procedió a corregir ambos.
- En Precio de igual manera se encontraba escrito de distintas maneras "introducción" y "precio normal".
- En Atributo Neto, "No aplica" y "no aplica".
- En Energy Chart, "cierre" y "oferta esencial" se encontraban escritos de maneras diferentes, se unificó.
- Lo mismo sucede en las columnas de Promoción y Trébol extra; en ambas, se corrigió para tener una misma instancia de cada nivel del factor.
- Cabe agregar que en la categoría X se encuentran también varias incidencias del mismo texto, sin embargo, estos están precedidos por un número distinto, por lo que dicha categoría se deja como esta.
- En las descripciones de los productos, había faltas de ortografía y NAs, se corrigieron para que los códigos de productos coincidan con el nombre.

Exploración de los datos

---

a. Comience describiendo cuantas variables y observaciones tiene disponible, el tipo de cada una de las variables.

Cada Catálogo de Venta por País, cuenta con 19 variables principales, las cuáles son:

Nombre de Variable	Descripción	Tipo
Año Mes	año y mes del registro	cuantitativa discreta
Producto	código del producto	cuantitativa discreta
Conca	concatenación de año mes y producto	cuantitativa discreta
Código Catálogo	código dentro del catálogo específico	cuantitativa discreta
Descripción	nombre del producto y contenido	cualitativa
Categoría	Grupo de productos al que pertenece	cualitativa
Página	no. de página dentro del catálogo	cuantitativa discreta
Línea	Clasificación de línea de productos	cualitativa
Precio Catálogo	Precio de venta final	cuantitativa continua
Precio Vta s/iva	Precio de venta sin IVA	cuantitativa continua
Pronóstico	unidades proyectadas de venta	cuantitativa discreta
Unidades Vendidas	unidades vendidas en realidad	cuantitativa discreta
Venta Neta s/iva	Ingresos sin iva por producto	cuantitativa continua
Costo	Costo unitario de cada producto	cuantitativa continua
Utilidad	Utilidad de todas las unidades vendidas	cuantitativa continua
Margen	Razón de ingresos sobre	cuantitativa continua

	coste por unidad vendida	
Pedido Real	Unidades pedidas	cuantitativa discreta
Ratio	Razón de unidades realmente vendidas respecto a unidades pedidas	cuantitativa continua
Observaciones	Promociones con nombre y código	cualitativa

Adicionalmente, cuenta con 12 variables que son dependientes por País y por Periodo, las cuáles son:

Nombre de Variable	Descripción	Tipo
Canal de Venta	medio por el cual se realiza la venta	cualitativa
Contingencia	Forma de canal de venta	cualitativa
Página	descripción de la página en la que se encuentra el producto	cualitativa
Tipo Precio	Precio normal o de oferta	cualitativa
%	Porcentaje del tipo de Precio	cuantitativa discreta
Tipo Comisión	Modalidad de comisiones	cualitativa
%	Porcentaje del tipo de comisión	cuantitativa discreta
Atributo Neto	Característica del atributo	cualitativa
Energy Chart	Lugar en el catálogo	cualitativa
Promociones	Tipo de promoción por producto	cualitativa
Recursos Especiales	Espacio especial dentro del catálogo o características extra	cualitativa
Tréboles Extra	Otro doblez u oferta extra	cualitativa

b. Haga un resumen de las variables numéricas y tablas de frecuencia para las variables categóricas, escriba lo que vaya encontrando.

Tablas de frecuencia de variables cualitativas

Se adjuntan primeras 5 tablas de frecuencia, las demás pueden ser encontradas en el código. En ellas se observa la cantidad de veces que un valor aparece en cada variable cualitativa.

Año Mes

x label	Freq	Percent	Valid Percent	Cumulative Percent
201801	1555	5.7	5.7	5.7
201802	1665	6.1	6.1	11.9
201803	1689	6.2	6.2	18.1
201804	1544	5.7	5.7	23.8
201805	1373	5.1	5.1	28.8
201806	1281	4.7	4.7	33.6
201807	1208	4.5	4.5	38.0
201808	1330	4.9	4.9	42.9
201809	1414	5.2	5.2	48.1
201810	1449	5.3	5.3	53.5
201811	1422	5.2	5.2	58.7
201812	1420	5.2	5.2	63.9
201901	1515	5.6	5.6	69.5
201902	1539	5.7	5.7	75.2
201903	1449	5.3	5.3	80.5
201904	1342	4.9	4.9	85.5
201905	1049	3.9	3.9	89.3
201906	1313	4.8	4.8	94.2
201907	1583	5.8	5.8	100.0
Total	27140	100.0	100.0	

Producto

		x label	Freq	Percent	Valid Percent	Cumulative Percent
8	Valid	4116260008	1	0.0	0.0	0.0
9		4117080142	4	0.0	0.0	0.0
10		4117080180	1	0.0	0.0	0.0
11		4117080226	12	0.0	0.0	0.1
12		4117080228	1	0.0	0.0	0.1
13		4117080239	1	0.0	0.0	0.1
14		4117080258	3	0.0	0.0	0.1
15		4117080280	4	0.0	0.0	0.1
16		4117080283	4	0.0	0.0	0.1
17		4117080289	3	0.0	0.0	0.1
18		4117080290	5	0.0	0.0	0.1
19		4117080292	4	0.0	0.0	0.2
20		4117080293	4	0.0	0.0	0.2
21		4117080294	7	0.0	0.0	0.2
22		4117080295	4	0.0	0.0	0.2
161		...	...	...	...	...
1800		4323811278	13	0.0	0.0	99.8
1801		4323811280	4	0.0	0.0	99.8
1802		4323811281	3	0.0	0.0	99.8
1803		4323811294	1	0.0	0.0	99.8
1804		4323811321	7	0.0	0.0	99.9
1805		4323811322	7	0.0	0.0	99.9
1806		4323811323	7	0.0	0.0	99.9
1807		4323811324	7	0.0	0.0	99.9
1808		4323811334	2	0.0	0.0	99.9
1809		4323811335	4	0.0	0.0	100.0
1810		4323811342	7	0.0	0.0	100.0
1811		4323811361	4	0.0	0.0	100.0
18111		Total	27139	100.0	100.0	
1	Missing	<blank>	0	0.0		
1812		<NA>	1	0.0		
1814		Total	27140	100.0		

CONCA

x	label	Freq	Percent	Valid Percent	Cumulative Percent
201801		1	0.0	0.0	0.0
2018014117080142		4	0.0	0.0	0.0
2018014117080180		1	0.0	0.0	0.0
2018014117080226		2	0.0	0.0	0.0
2018014117263475		1	0.0	0.0	0.0
2018014117263568		1	0.0	0.0	0.0
2018014117270738		1	0.0	0.0	0.0
2018014117270739		1	0.0	0.0	0.0
2018014123013019		4	0.0	0.0	0.1
2018014123040001		4	0.0	0.0	0.1
2018014123040501		4	0.0	0.0	0.1
2018014123040601		4	0.0	0.0	0.1
2018014123041001		4	0.0	0.0	0.1
2018014123042238		4	0.0	0.0	0.1
2018014123042239		4	0.0	0.0	0.1
...	...	...	...	...	...
2019074323811275		3	0.0	0.0	99.9
2019074323811277		4	0.0	0.0	99.9
2019074323811278		4	0.0	0.0	99.9
2019074323811321		3	0.0	0.0	99.9
2019074323811322		3	0.0	0.0	99.9
2019074323811323		3	0.0	0.0	99.9
2019074323811324		3	0.0	0.0	99.9
2019074323811334		2	0.0	0.0	100.0
2019074323811335		3	0.0	0.0	100.0
2019074323811342		4	0.0	0.0	100.0
2019074323811361		4	0.0	0.0	100.0
2108114123924039		1	0.0	0.0	100.0
Total		27140	100.0	100.0	

Código Catálogo

x	label	Freq	Percent	Valid Percent	Cumulative Percent
12		51	0.2	0.2	0.2
16		8	0.0	0.0	0.2
17		9	0.0	0.0	0.3
21		4	0.0	0.0	0.3
22		66	0.2	0.2	0.5
23		43	0.2	0.2	0.7
24		68	0.3	0.3	0.9
25		24	0.1	0.1	1.0
26		3	0.0	0.0	1.0
30		4	0.0	0.0	1.0
32		4	0.0	0.0	1.0
33		4	0.0	0.0	1.1
34		4	0.0	0.0	1.1
41		6	0.0	0.0	1.1
42		1	0.0	0.0	1.1
...	...	...	...	...	...
14969		4	0.0	0.0	99.9
14970		4	0.0	0.0	99.9
14971		4	0.0	0.0	99.9
14972		4	0.0	0.0	99.9
14973		4	0.0	0.0	100.0
14991		1	0.0	0.0	100.0
15031		1	0.0	0.0	100.0
15092		4	0.0	0.0	100.0
15093		1	0.0	0.0	100.0
15106		1	0.0	0.0	100.0
15107		1	0.0	0.0	100.0
15527		1	0.0	0.0	100.0
Total		27139	100.0	100.0	
<blank>		0	0.0		
<NA>		1	0.0		
Total		27140	100.0		

Descripción

x	label	Freq
2 STRONG BODY CREAM FRESA KIWI 200		2
2 STRONG BODY SPLASH FRESA KIWI 180ML		3
2STRONG ROLL ON VANILLA TWIST 80G		3
2STRONG SHAMPOO 2 EN 1 FRESA & KIW		2
2STRONG VAINIL TWIST BODY SPLASH 180ML		2
3 PACK BURBUJAS		4
8 PCS. FURNITURE PLAY SET ON CARD		4
10 PCS PIZZA PLAY SET ON CARD		4
10X7.5 BASKET BALL BOARD PLAY SET IN NET BAG		4
12 ESTUCHES 35% DE DESCUENTO 3 FRAGANCIAS GRATIS DICIEMBRE		4
25 ESTUCHES 35% DE DESCUENTO 6 FRAGANCIAS GRATIS 1/2 TREBOL DC		1
500 MILLAS EAU DE COLOGNE 100ML		13
"COMBO MASCARILLA CARBÓN ACTIVADO 990G. + ESPUMA FACIAL DE AZUFRE 240ML."		3
AC SOLUTION,POL COM ÁCI SAL BEIGE		1
ACETAMINOFEN ADULTOS 20 TABLETAS		66
...	...	...
WILD BABY SHAMPOO 1 LITRO		42
WILD BABYS COLONIA 240 ML		36
WILD BABYS CREMA 240G		36
WOOF LOCION PERFUMADA PARA PERROS 240ML		6
WOOF SHAMPOO PARA PERROS 470ML		9
WSC ENIGMA EAU DE TOILETTE 240 ML		36
WSC FLECHA DORADA EAU DE TOILETTE 240 ML		36
WSC MASCARA DE FUEGO EAU DE TOILETE 240 ML		44
ZAFIROS CREMA PERFUMADA 150 G		16
ZAFIROS DIJE		5
ZAFIROS EAU DE TOILETTE 100ML		60
ZAFIROS EAU DE TOILETTE ED ESPECIAL 50 ML		15
Total		27139
<blank>		0
<NA>		1
Total		27140

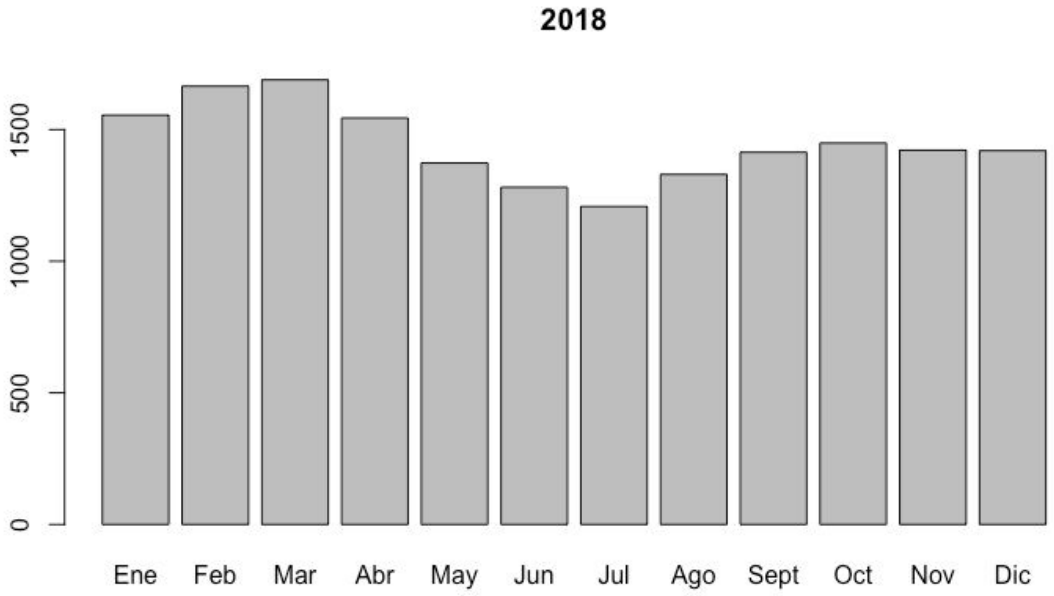
Estadísticas de las variables numéricas:

Precio Catalogo	Precio Vta s/iva	Pronostico	Unidades Vendidas	Venta Neta s/iva
Min. : 0.00	Min. : 0.000	Min. : 0.0	Min. :-1213.0	Min. : -69528
1st Qu.: 8.95	1st Qu.: 4.925	1st Qu.: 100.0	1st Qu.: 0.0	1st Qu.: 0
Median : 45.95	Median : 26.844	Median : 300.0	Median : 173.0	Median : 8941
Mean : 106.39	Mean : 63.097	Mean : 686.4	Mean : 553.9	Mean : 38876
3rd Qu.: 154.95	3rd Qu.: 91.274	3rd Qu.: 750.0	3rd Qu.: 552.0	3rd Qu.: 33507
Max. :6819.40	Max. :4150.939	Max. :42000.0	Max. :51957.0	Max. :2110540
NA's :1695	NA's :895			

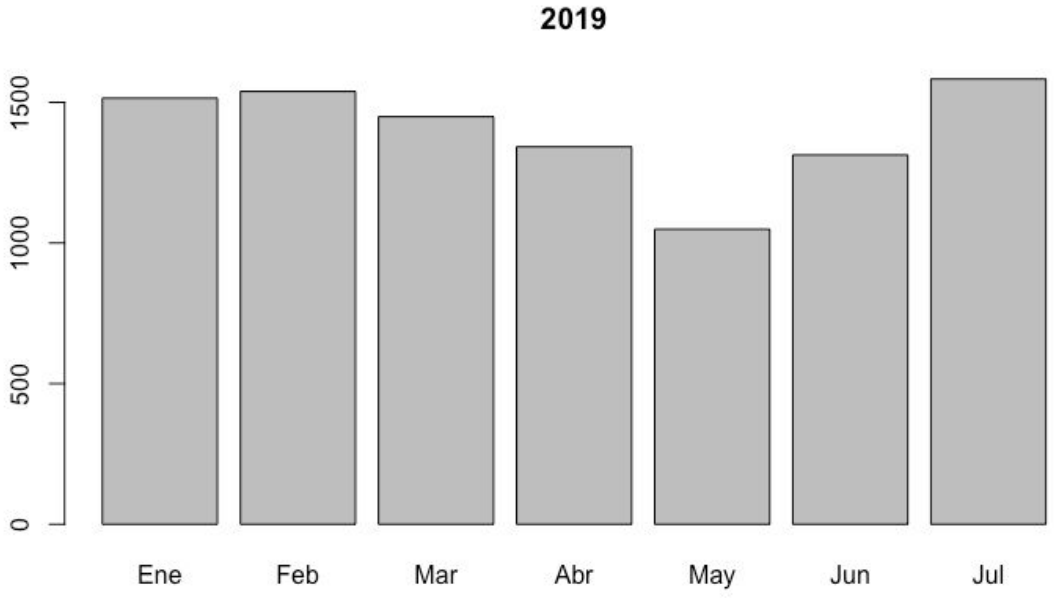
Costo	Utilidad	Margen	Pedido Real	Ratio
Min. : 0.00	Min. :-451731.8	Min. :-1997.0000	Min. : 0	Min. :-0.17413
1st Qu.: 0.65	1st Qu.: 279.8	1st Qu.: 0.1482	1st Qu.: 0	1st Qu.: 0.00000
Median : 5.92	Median : 6353.9	Median : 0.6001	Median : 7764	Median : 0.01358
Mean : 17.47	Mean : 30730.1	Mean : 0.5256	Mean : 6828	Mean : 0.05557
3rd Qu.: 27.69	3rd Qu.: 23978.8	3rd Qu.: 0.9768	3rd Qu.:10540	3rd Qu.: 0.05275
Max. :431.94	Max. :2110539.5	Max. : 754.0161	Max. :14527	Max. : 6.75959
NA's :1262	NA's :817		NA's :800	

Año Mes:

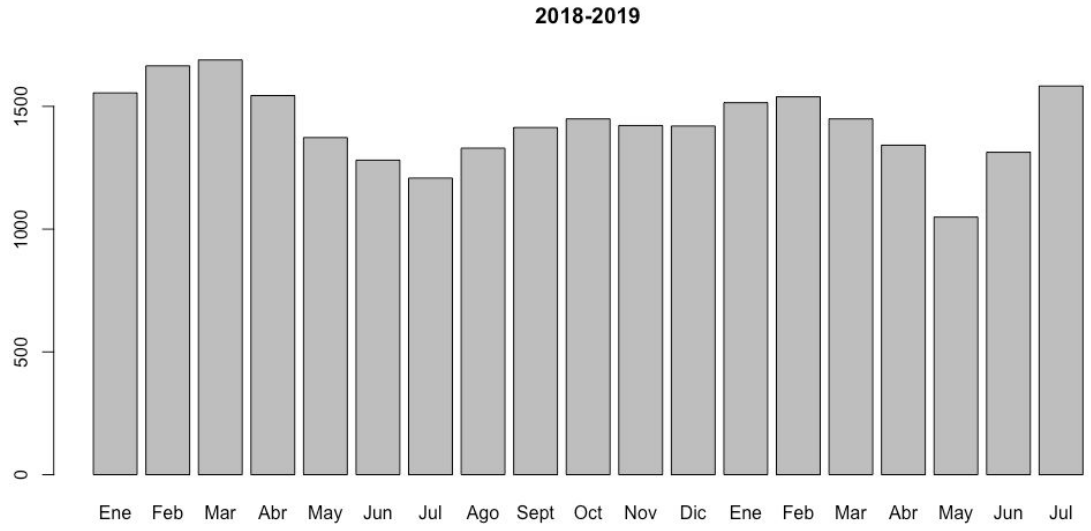
201801	201802	201803	201804	201805	201806	201807	201808	201809	201810	201811	201812
1555	1665	1689	1544	1373	1281	1208	1330	1414	1449	1422	1420
201901	201902	201903	201904	201905	201906	201907					
1515	1539	1449	1342	1049	1313	1583					



Se puede ver que en el 2018, el mes de Marzo fue el mes que tuvo más pedidos registrados, con 1689.



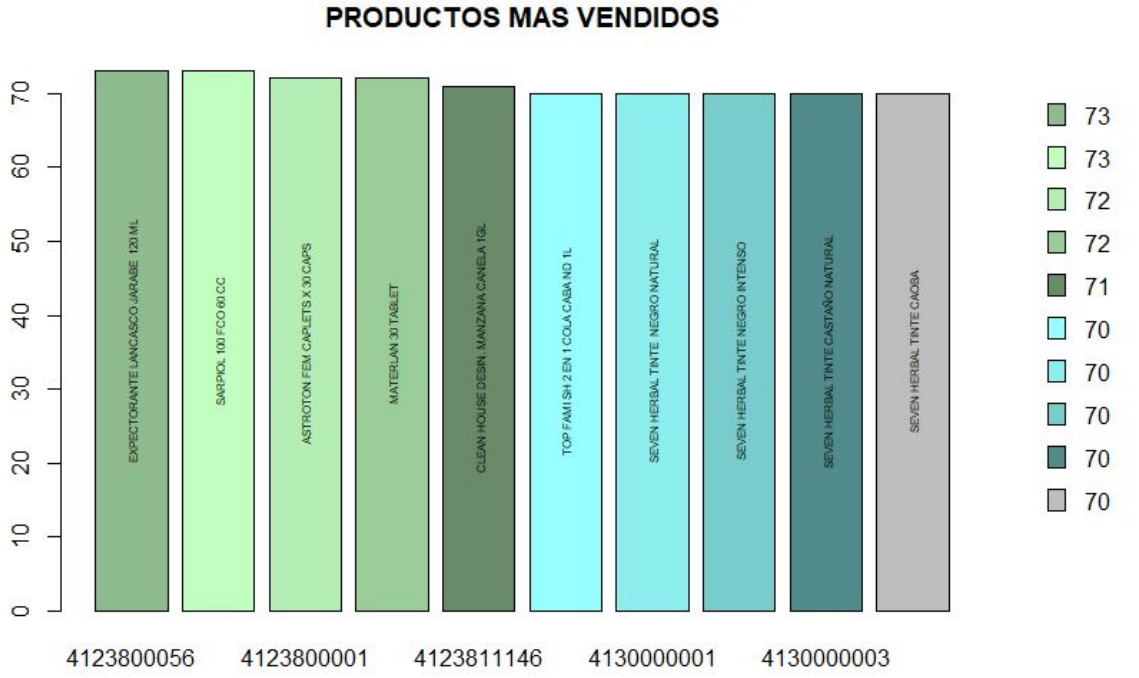
Se puede ver que en lo que va del 2019, el mes de Julio es el mes que tiene más pedidos registrados ha tenido, con 1583.



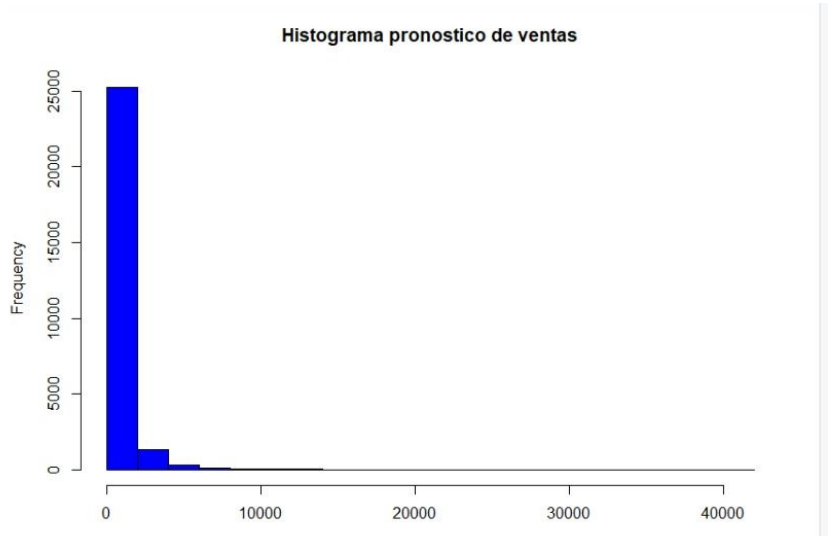
Se puede observar que de las registros totales, Marzo del año 2018 tiene el récord con 1689 pedidos registrados.

Producto:

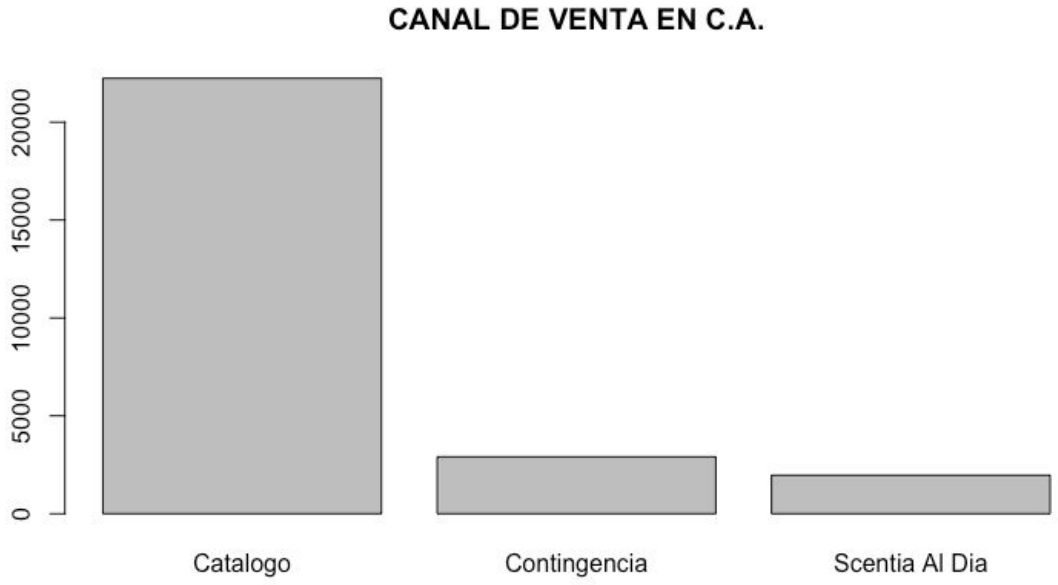
```
4123800056 4123800080 4123800001 4123800030 4123811146 4123660214 4130000001 4130000002 4130000003
        73         73         72         72         71         70         70         70         70
4130000004
        70
```



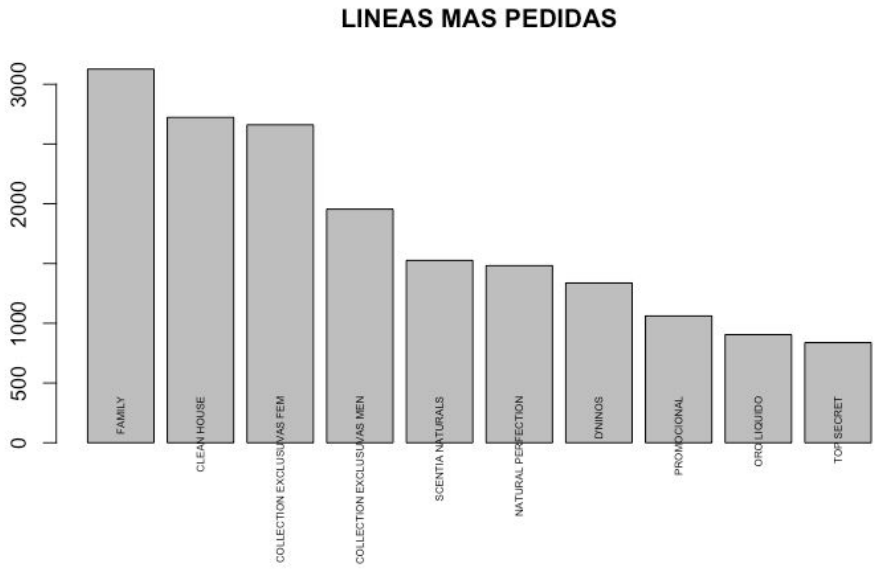
Se puede ver que los productos más vendido en Centroamérica por parte de Scentia en 2018-2019 son : “EXPECTORANTE LANCASCO JARABE 120 ML” y “SARPIOL 100 FCO 60 CC”



Respecto al pronóstico de unidades de venta, se puede observar una asimetría positiva. Esto indica que los datos no se encuentran bien distribuidos ni presentan normalidad; sino que los mismos se acumulan del lado izquierdo. La gran mayoría de predicciones son muy similares, de modo que no se tienen en cuenta factores estacionales a lo largo del año ni diferencias entre los países.



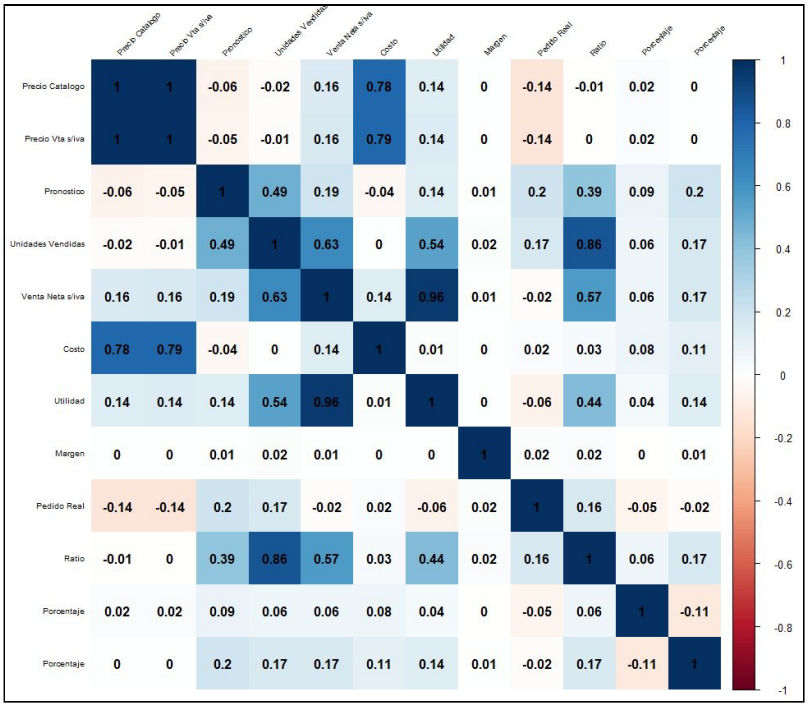
En los canales de venta que Scentia usa para vender sus productos, encontramos que el Catálogo es su fuerte.



Las líneas de productos más pedidas son de Familia, Clean House y Collection Exclusive FEM.

c. Cruce las variables que considere que son las más importantes para hallar los elementos clave que lo pueden llevar a comprender lo que está causando el problema encontrado.

Matriz de correlación



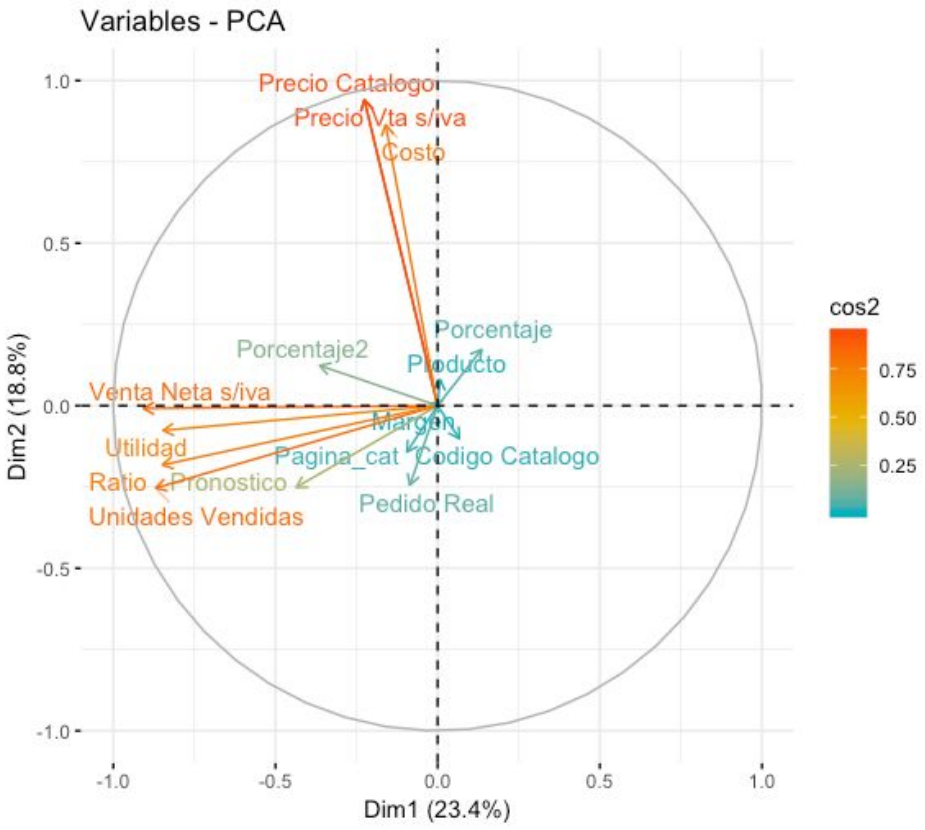
De la matriz de correlación se puede obtener las relaciones más importantes entre las variables. En color azul fuerte las correlaciones más altas, y en tonos rojos las más bajas.

Así, más adelante es posible que se elimine una de cada pareja de variables con coeficiente de correlación muy alto. En este caso se eliminará la que sea menos significativa para el modelo.

d. Haga gráficos exploratorios que le de ideas del estado de los datos.

PCA

Antes de hacer un análisis de factores principales se calculó el KMO que tuvo un valor de 0.67056 un valor aceptable que indica que sí vale la pena hacer este análisis, además de esto se hizo el test de esfericidad de Barlett obteniéndose un valor de chi-cuadrado muy alto y un valor p = 0. Esto es menor a 0.05, lo que indica que hacer PCA si podría funcionar.



Se puede ver que con 2 componentes podemos explicar aproximadamente el 41% de la variabilidad de los datos, lo cual es aceptable.

Las variables que presentan más relación entre sí son:

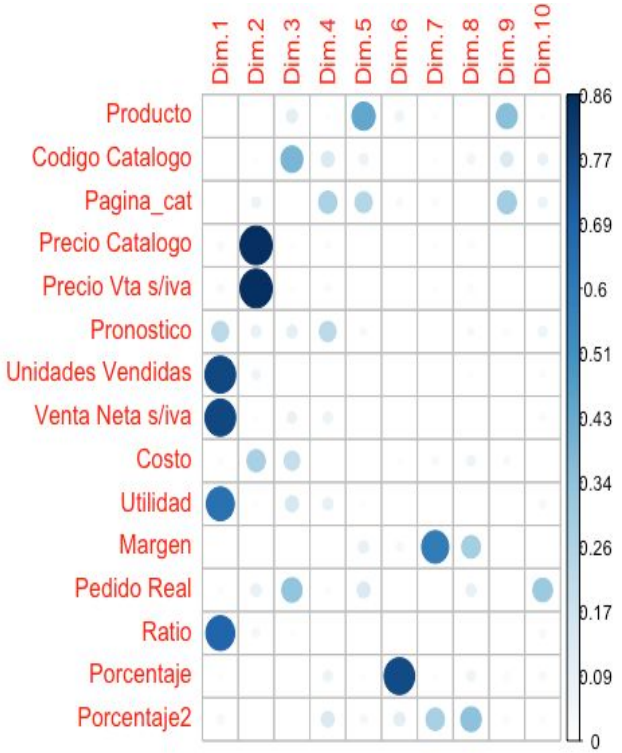
- Costo con precio Catálogo y Precio de venta sin IVA.
- Unidades vendidas con Ratio, Utilidad, Venta neta sin IVA y Pronóstico.

```
Importance of components:
                          PC1    PC2    PC3    PC4   PC5    PC6    PC7    PC8    PC9   PC10   PC11   PC12    PC13    PC14
Standard deviation     1.872 1.681 1.1704 1.0314 1.017 1.0030 0.9891 0.9132 0.8930 0.8242 0.6977 0.5212 0.33897 0.18504
Proportion of Variance 0.234 0.188 0.0913 0.0709 0.069 0.0671 0.0652 0.0556 0.0532 0.0453 0.0324 0.0181 0.00766 0.00228
Cumulative Proportion  0.234 0.422 0.5132 0.5842 0.653 0.7202 0.7854 0.8410 0.8942 0.9395 0.9719 0.9901 0.99772 1.00000
                          PC15
Standard deviation     0.00329
Proportion of Variance 0.00000
Cumulative Proportion  1.00000
```

Si usamos 10 Componentes Principales, podríamos llegar a explicar el 93% de la variabilidad de los datos. Estas 10 Componentes serían :

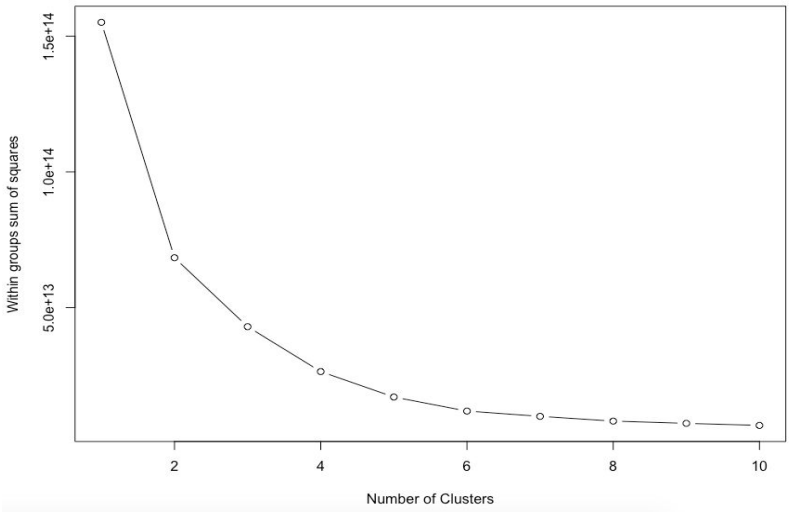
```
[1] "Producto"         "Codigo Catalogo"  "Pagina_cat"       "Precio Catalogo"  "Precio Vta s/iva"  "Pronostico"
[7] "Unidades Vendidas" "Venta Neta s/iva" "Costo"            "Utilidad"
```

Los componentes principales pueden explicar las variables de la siguiente manera.



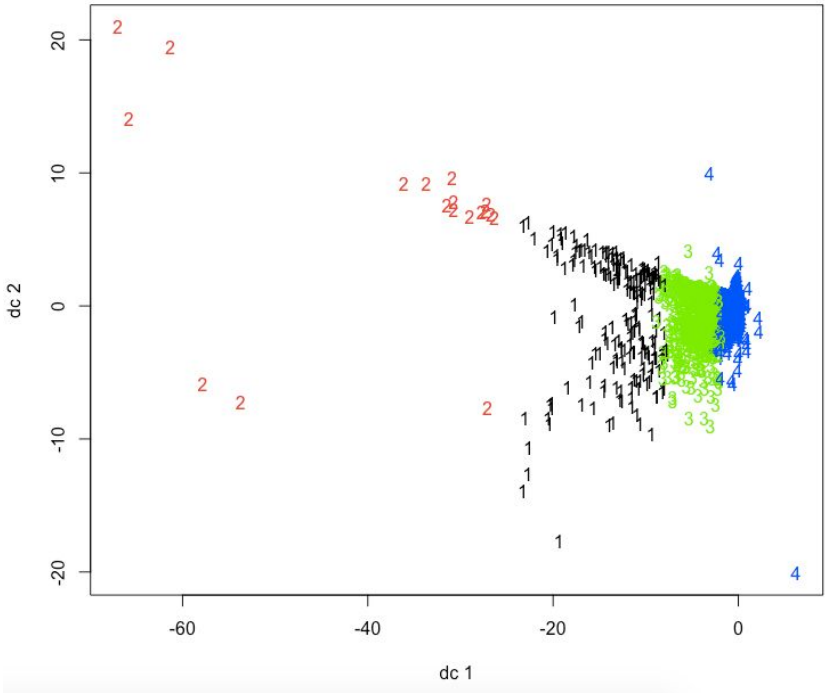
Clustering

Método de Ward



Según el gráfico obtenido del método de Ward para determinar el número adecuado de clusters para K-medias, se deben de crear 4 clusters para el conjunto de datos que se quiere analizar. Luego de esto se procedió a generar el gráfico donde se muestra la ubicación de cada uno de los clusters, se obtuvo el siguiente gráfico:

Ubicación de cada uno de los clusters



Finalmente se calculó la silueta y se obtuvo un valor de 0.76876 lo que indica que el agrupamiento hecho anteriormente es adecuado.

Reglas de asociación

Se realizó un análisis con reglas de asociación de variables, a modo de obtener las relaciones más importantes entre las variables. El soporte,se define como la proporción de transacciones en la base de datos que contiene dicho conjunto de items. Por otra parte la confianza es un estimador de probabilidad de que este suceso se de. Las reglas más importantes halladas fueron las siguientes:

Al ser el tipo de precio oferta, el canal de ventas será catálogo con un 38.61% de soporte, y un 85.88% de confianza.

Al ser el tipo de comisión normal, la promoción será en precio con un 30.11% de soporte, y un 66.27% de confianza.

Cuando la comisión es normal, el canal de venta será catálogo con 40.36% de soporte, y un 88.83% de confianza.