Máster en Tecnologías de Análisis de Datos Masivos: Big Data

TRABAJO DE FIN DE MÁSTER

Análisis de Comportamiento de Consumos de Clientes

Jean Deynis Valenzuela Najar

Junio 2018

Trabajo de Fin de Máster presentado en la Escuela Técnica Superior de Ingeniería de la Universidad de Santiago de Compostela para la obtención del Máster universitario en Tecnologías de Análisis de Datos Masivos: Big Data

Resumen

En el presente documento se muestra el resultado del proyecto realizado como Trabajo de Fin de Máster Interuniversitario en Tecnologías de Análisis de Datos Masivos: Big Data, de la Universidad de Santiago de Compostela.

El objetivo planteado es realizar el procesamiento masivo de la información de las ventas de los últimos tres años de Café Candelas, aplicando técnicas de modelado estadístico y machine learning.

En primer lugar, se revisaron, corrigieron y estandarizaron las bases de datos provistas por Café Candelas para posteriormente consolidarlas y almacenarlas en un archivo.

Posteriormente se realizaron análisis de clustering con la finalidad de poder determinar las agrupaciones de provincias con preferencias similares de consumo.

Finalmente se realizaron análisis de regresión lineal para poder determinar la proyección de ventas para los próximos años.

Índice

Índice de abreviaturas

CSS: Hojas de estilo en cascada (sigla en inglés de *Cascading Style Sheets*)

CSV: Tipo de documento en formato abierto sencillo para representar datos en forma de tabla, en las que las columnas se separan por comas (sigla en inglés *comma separated values*)

ERP: Sistema de planificación de recursos empresariales (sigla en inglés de *Enterprise Resource Planning)*. Estos programas se hacen cargo de distintas operaciones internas de una empresa que pueden ser: producción, distribución y/o recursos humanos.

HTML: lenguaje de marcas de hipertexto (sigla en inglés de *HyperText Markup Language*)

Mac OS: Sistema Operativo de Macintosh (siglas en inglés *Macintosh Operanting System*)

Markdown: formato que permite una fácil creación de documentos, presentaciones dinámicas y informes de R

Índice de tablas, gráficos u otras figuras

Introducción

Café Candelas comercializa una variedad de productos de alta calidad, los mismos que importa de distintas partes del mundo, entre los que se puede encontrar café e infusiones en distintas variedades y presentaciones, así como también, complementos para la elaboración y comercialización de bebidas.

Café Candelas tiene presencia a nivel nacional donde la mayoría de sus ventas se realizan en restaurantes, hoteles y catering[1], así mismo realiza ventas por internet, vending[2] y en supermercados; exporta y comercializa sus productos a los países de: Andorra, Portugal y Estados Unidos. En la actualidad, Café Candelas, cuenta con un promedio de quince mil clientes.

Para el desarrollo del siguiente proyecto se utilizará la información de las ventas de los años 2015, 2016 y 2017, esa gran cantidad de información será utilizada para analizar el comportamiento y evolución de los patrones de consumo por producto, canal, ámbitos geográficos a lo largo de los tres últimos ejercicios, se realizará una proyección para los próximos ejercicios. Que permita a Café Candelas tomar decisiones en cuanto a la mejora de sus servicios y productos.

[1] Se denomina catering al servicio de alimentación institucional o alimentación colectiva que provee una cantidad determinada de comida y bebida en fiestas, eventos y presentaciones de diversa índole.

[2] Vending es un neologismo en voz inglesa que se utiliza para denominar el sistema de ventas por medio de máquinas auto expendedoras accionadas por diversos medios de pago.

Se desarrollará un panel de control en Shiny de Rstudio, se usarán diferentes librerías para dar soporte a las tareas de aprendizaje no supervisado y de visualización de datos para la presentación de los resultados.

1 Objetivos

El objetivo del proyecto es realizar el procesamiento masivo de información aplicando técnicas de modelado estadístico y *machine learning*.

Con el fin de alcanzar este objetivo principal será necesario realizar los siguientes procedimientos específicos:

- Carga de los datos.
- Comprensión de los datos.
- Preparación de los datos.
- Implementación del agrupamiento utilizando librerías de *machine learning*
- Evaluación de los resultados obtenidos.
- Presentación visual de resultados con gráficas.

2 Herramientas

2.1 Herramientas de Software

RStudio: es un entorno de desarrollo integrado para el lenguaje de programación R, dedicado a la computación estadística y gráficos. Incluye una consola, editor de sintaxis que apoya la ejecución de código, así como herramientas para el trazado, depuración y la gestión del espacio de trabajo.

Rstudio está disponible para los sistemas operativos: Windows, Mac OS y Linux, así como también, para navegadores web conectados a RStudio Server y/o RStudio Server Pro.

2.2 Paquetes de R

Shiny: es un paquete de R que facilita la creación de aplicaciones web interactivas directamente desde R. Hace posible alojar aplicaciones independientes en una página web o insertarlas en documentos de R Markdown o crear cuadros de mando. También puede extender las aplicaciones con temas CSS, html widgets y acciones de JavaScript.

Highcharter: Highcharter es un contenedor de R para *Highcharts javascript library* y sus módulos. Permite realizar gráficos tales como: dispersión, burbuja, línea, series de tiempo, mapas de calor, mapa de árbol, gráficos de barras, redes, etc.

Cluster: Es un paquete de R que cuya utilidad es calcular la agrupación jerárquica aglomerativa de un conjunto de datos.

Gglot2: es un paquete de visualización de datos para el lenguaje de programación estadística R.

Factoextra: Proporciona algunas funciones fáciles de usar para extraer y visualizar el resultado de análisis de datos multivariantes, incluidos análisis de componentes principales (PCA), análisis de correspondencia (CA), análisis de correspondencia múltiple (MCA), análisis de factores de datos mixtos (FAMD), funciones de análisis de factores múltiples (MFA) y análisis de factores múltiples jerárquicos (HMFA) de diferentes paquetes R. Contiene también funciones para simplificar algunos pasos de análisis de clustering y proporciona visualización de datos elegante basada en ggplot2.

clValid: Es un paquete de R que permite realizar la validación estadística y biológica de los resultados de la agrupación.

Planificación

En esta sección se desglosan y explican las principales fases del proyecto, diferenciándolas mediante una descripción.

1- Revisión de las bases de datos entregadas por Café Candelas:

- Revisión de los tipos, formatos y escala de los datos.
- Realización del análisis exploratorio de las variables.

2- Identificación de las variables dependientes significativas.

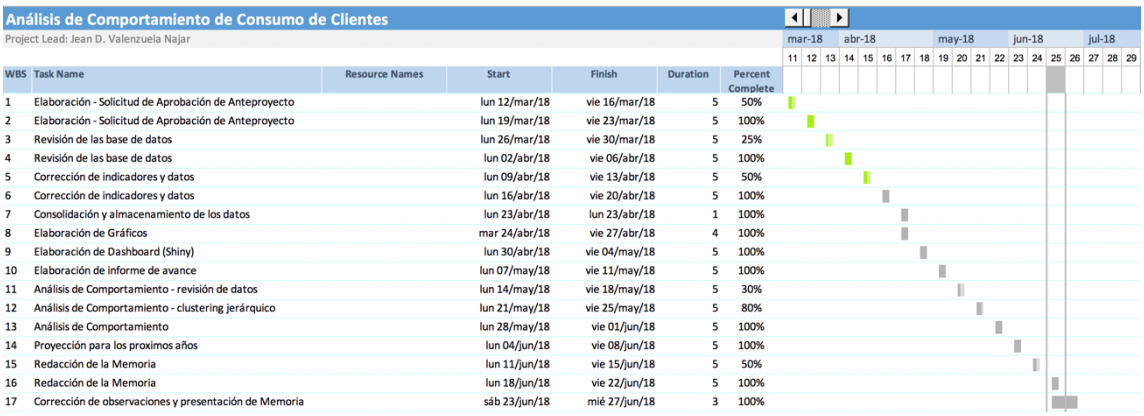
- Contraste de hipótesis de proporciones (análisis de estadística de significancia).
- Modelado del aprendizaje de maquina supervisado (machine learning).
- Comparaciones y conclusiones acerca de los modelos empleados.

3- Segmentación de perfiles:

- Descripción de las variables dependientes significativas
- Descripción de los perfiles de clientes.

4- Conclusiones

- Revisión de los tres apartados mencionados anteriormente.
- Ranking y descripción detallada de los perfiles de consumo identificados.



Análisis de Comportamiento de Consumo de Clientes

Project Lead: Jean D. Valenzuela Najar

WBS	Task Name	Resource Names	Start	Finish	Duration	Percent Complete
1	Elaboración - Solicitud de Aprobación de Anteproyecto		lun 12/mar/18	vie 16/mar/18	5	50%
2	Elaboración - Solicitud de Aprobación de Anteproyecto		lun 19/mar/18	vie 23/mar/18	5	100%
3	Revisión de las base de datos		lun 26/mar/18	vie 30/mar/18	5	25%
4	Revisión de las base de datos		lun 02/abr/18	vie 06/abr/18	5	100%
5	Corrección de indicadores y datos		lun 09/abr/18	vie 13/abr/18	5	50%
6	Corrección de indicadores y datos		lun 16/abr/18	vie 20/abr/18	5	100%
7	Consolidación y almacenamiento de los datos		lun 23/abr/18	lun 23/abr/18	1	100%
8	Elaboración de Gráficos		mar 24/abr/18	vie 27/abr/18	4	100%
9	Elaboración de Dashboard (Shiny)		lun 30/abr/18	vie 04/may/18	5	100%
10	Elaboración de informe de avance		lun 07/may/18	vie 11/may/18	5	100%
11	Análisis de Comportamiento - revisión de datos		lun 14/may/18	vie 18/may/18	5	30%
12	Análisis de Comportamiento - clustering jerárquico		lun 21/may/18	vie 25/may/18	5	80%
13	Análisis de Comportamiento		lun 28/may/18	vie 01/jun/18	5	100%
14	Proyección para los proximos años		lun 04/jun/18	vie 08/jun/18	5	100%
15	Redacción de la Memoria		lun 11/jun/18	vie 15/jun/18	5	50%
16	Redacción de la Memoria		lun 18/jun/18	vie 22/jun/18	5	100%
17	Corrección de observaciones y presentación de Memoria		sáb 23/jun/18	mié 27/jun/18	3	100%

Imagen 1: Diagrama de Gantt – Desarrollo del Trabajo

Desarrollo del trabajo

En la presente sección se comentan los diferentes procedimientos que se han realizado para el desarrollo del trabajo. Se hará mención de las librerías utilizadas, así como también, se presentará una breve descripción y justificación de la elección. Se llegará hasta un nivel de detalle que permita comprender en su totalidad.

3 Base de datos

Este procedimiento es muy importante debido a que, si se comete un error, en el tratamiento de la información, se podrían alterar los resultados de todos los procedimientos planificados. A continuación, se detallan los procedimientos que se realizaron.

Resulta necesario mencionar que Café Candelas cuenta con un ERP, desde el cual tuvo que exportar los datos a archivos de extensión XLS.

3.1 Revisión de los archivos recibidos

De la revisión de los archivos de extensión XLS proporcionadas, se identifica los siguientes indicadores:

- **Código Cliente:** Código numérico asignado a cada uno de sus clientes.

- **Código Postal:** consta de cinco dígitos. Los dos primeros hacen referencia a la provincia[3], los tres últimos corresponden a zonas postales y de reparto asignadas por la Sociedad Estatal Correos y Telégrafos de España (1).

- **Provincia:** Debido a que Café Candelas tiene presencia a nivel nacional e internacional, en la base de datos se puede encontrar las 52 provincias de España, en el caso de clientes que no corresponden a España se indica como *"*extranjero*"*.

- **Familia:** Los productos comercializados se encuentran agrupados en 5 familias:

  Azúcar y Edulcorantes.
  Chocolates, Galletas y otros.
  Infusiones.
  Café.
  Solubles

- **Gama**: Se encuentra información para productos relacionados al café, de acuerdo con la exploración se encuentran agrupados de acuerdo con las siguientes familias:

Café:	**Solubles:**
The special vending	The special vending
The premium coffee	The premium coffee
The organic coffee	The organic coffee
The gourmet coffee	The fair trade coffee
The fair trade coffee	The essential coffee
The essential coffee	
Café verde	
Café tostadero / cliente	
Café tostadero	

- **Jerarquía:** Se encuentra información desagregada de productos relacionados al café, de acuerdo con la exploración y para un mejor entendimiento se agrupará a nivel de familias.

[3] fueron asignados siguiendo un orden alfabético, del 01 al 50, y a las ciudades de Ceuta y Melilla, que, al estar fuera de la división provincial, se les asignaron el 51 y el 52 respectivamente

Café:	**Solubles:**
Café Bahía grano mezcla	Cápsula Selectum descafeinado
Café Bahía grano natural	Cápsula Selectum natural
Café Candelas grano descafeinado	Cápsula Profesional descafeinado
Café Candelas grano mezcla	Cápsula Profesional natural
Café Candelas grano natural	Soluble descafeinado
Café Candelas grano torrefacto	Soluble liofilizado
Café Candelas molido natural	Soluble natural
Café Candelas molido mezcla	Soluble natural sobres
Café Dakar grano natural	Soluble Candelas liofilizado
Café Doce grano natural	Otros Solubles
Café Giorno grano natural	

- **Composición:** Se refiere a las variedades de café que existen, el café Robusta tiene aproximadamente el doble de cafeína que el Arábica. Es un tipo de variedad originaria de África Central que, al crecer en zonas secas, es poco digestivo, tiene un gusto final amargo, con mucho cuerpo y poco perfumado. Su cultivo representa el 43% de la producción mundial.

  De acuerdo con lo observado en la base datos, Café Candelas comercializa productos cuya composición pueden variar entre:

  100% Arábica.
  100% Robusta.
  Mezcla A/R

- **Meses:** En estos campos (de enero a diciembre) se pueden visualizar las ventas mensuales de un determinado producto a un determinado cliente.

- **Total:** En ese campo se cuenta con la sumatoria de la venta anual de un producto a un determinado cliente.

3.2 Corrección de indicadores y datos (Estandarización)

Es comprensible que el personal de Café Candelas posterior al almacenamiento de los datos, en los archivos de extensión XLS, haya modificado los nombres de los campos quizás con la finalidad de facilitar el entendimiento de los datos, sin embargo, este no fue homogéneo.

Con la finalidad de estandarizar las variables y los datos se procedió a realizar la estandarización.

En la imagen siguiente podemos visualizar el procedimiento de consolidación y estandarización de los datos.

```
datos <- bind_rows(datos_2015, datos_2016, datos_2017)
datos <- datos %>% rename(Composicion = `Composición`)
datos <- datos %>% rename(Jerarquia = `Jerarquía`)
datos <- datos %>% mutate(Composicion = str_trim(Composicion))
datos <- datos %>% mutate(Provincia = str_trim(Provincia))

datos <- datos %>% mutate(Gama = recode(datos$Gama,
                          `CAFÉ TOSTADERO/CLIENTES` = "CAFÉ TOSTADERO/CLIENTE",
                          `CAFÉ TOSTADERO/ CLIENTES` = "CAFÉ TOSTADERO/CLIENTE",
                          `The Fair trade Coffee`   = "THE FAIR TRADE COFFEE" ))

datos <- datos %>% mutate(Jerarquia = recode(datos$Jerarquia,
                           `CAFÉ BAHIA GRANO NATURAL` = "CAFÉ BAHÍA GRANO NATURAL" ))

datos <- datos %>% select(-Total)
```

Imagen 2: Estandarización de indicadores y datos

3.3 Almacenamiento de la Información.

Luego de la revisión y estandarización cada uno de los archivos anteriormente mencionados se procedió a realizar el almacenamiento de los datos en un archivo de extensión CSV.

3.4 Provincias de España

Para poder trabajar con la información geográfica de España fue necesario elaborar un archivo que contiene: código de provincia, nombre de provincia, acrónimo, nombre de la comunidad autónoma a la que pertenece la provincia y código internacional.

La información contenida en el archivo mencionado es importante porque permite enlazar los datos de las ventas con otras bases de datos, como por ejemplo si fuese el caso en el que se desea elaborar un mapa donde se quiera reflejar las ventas a nivel de las provincias de España o a nivel mundial.

nombre_provincia <chr>	id_provincia <int>	acronimo <chr>	hckey <chr>	c_autonoma <chr>
alava	1	VI	es-vi	País Vasco
albacete	2	AB	es-ab	Castilla-La Mancha
alicante	3	A	es-a	Comunidad Valenciana
almería	4	AL	es-al	Andalucía
asturias	33	O	es-o	Principado de Asturias
avila	5	AV	es-av	Castilla y León
badajoz	6	BA	es-ba	Extremadura
baleares	7	PM	es-pm	Islas Baleares
barcelona	8	B	es-b	Cataluña
burgos	9	BU	es-bu	Castilla y León

1-10 of 52 rows Previous 1 2 3 4 5 6 Next

Imagen 3: Información de Provincias de España

3.5 Representación Gráfica de Resultados

Para un mejor entendimiento de la información, así como, para que Café Candelas pueda visualizar los datos cuando lo considere necesario. Con el uso de Rstudio, Shiny y Highcharter se ha diseñado una herramienta web donde se puede visualizar la distribución de los datos a nivel de provincias de España por año.

En la imagen siguiente, se están reflejando las ventas del año 2017 a nivel de la familia café, se están coloreando de colores más oscuros las provincias donde se han registraron mayores ventas y colores claros donde se registraron las menores, si posamos el cursor por encima de una provincia podemos visualizar el nombre de la provincia y total de ventas.

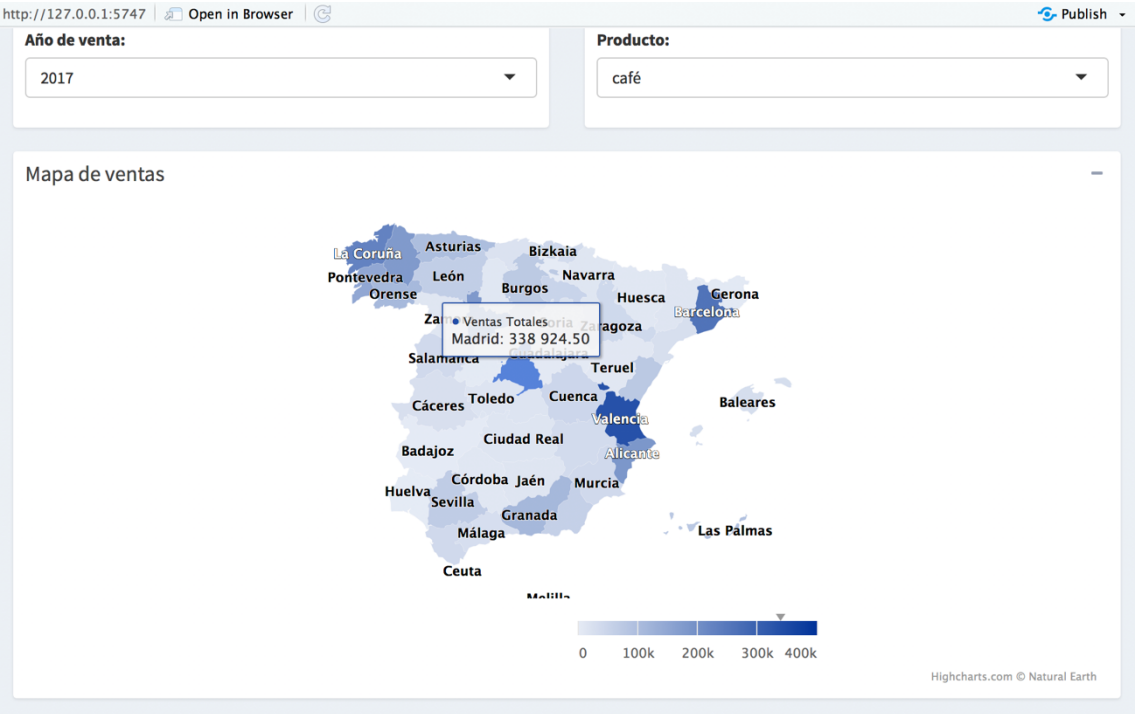


Imagen 4: Mapa de España – Ventas totales a nivel de familia por año

4 Análisis de Comportamiento y Evolución de Patrones de Consumo

Como primer aspecto, en el desarrollo del siguiente capítulo, se hizo uso de reconocimiento de patrones o aprendizaje automático no supervisado – se denomina no supervisado porque no nos guiamos por ideas a priori de qué variables o muestras pertenecen a qué clústeres. Aprendizaje porque el algoritmo de la máquina aprende a agrupar.

El análisis de conglomerados es utilizado en muchos campos, incluyendo:

- En la investigación del cáncer para clasificar a los pacientes en subgrupos según su perfil de expresión genética. Esto puede ser útil para identificar el perfil molecular de pacientes con pronóstico bueno o malo, así como para comprender la enfermedad.
- En marketing para la segmentación del mercado mediante la identificación de subgrupos de clientes con perfiles similares y que puedan estar receptivos a una forma particular de publicidad.
- En la planificación de la ciudad para identificar grupos de casas de acuerdo con su tipo, valor y ubicación.

4.1 Carga y exploración de datos

4.1.1 Datos de Ventas

En este archivo se encuentra la información consolidada de los años 2015, 2016 y 2017, previamente al desarrollo de los clustering fue necesario realizar los siguientes procedimientos:

4.1.1.1 Conversión de datos

En las diferentes técnicas de *clustering* solo se pueden emplear variables de tipo numérico (*integer* o *numeric*), en el caso de variables de tipo cualitativo, tienen que ser binarizadas.

4.1.1.2 Normalización de las variables

Los métodos de *clustering* se ven influenciados por la escala en la que se miden las variables. Para evitar que las variables de mayor magnitud determinen la

agrupación, se procede a normalizar todas las variables para que estén en la misma escala.

4.1.1.3 Conversión del dataframe a matriz

Como la mayoría de los algoritmos de clustering necesitan los datos en forma de matriz, se convierte el dataframe a matriz.

4.1.2 Provincias de España

La tendencia de consumo de provincias que se encuentran próximas, algunas veces, tiende a ser similar. Entre los datos disponibles en el archivo de provincias, se encuentra la comunidad autónoma a la que pertenece cada provincia. Se utiliza esta información para colorear de un mismo tono los nombres de las provincias en función a la comunidad autónoma a la que pertenecen.

4.2 Clustering Jerárquico

Existen dos tipos de clustering jerárquico:

- Aglometarivo
- Divisivo

4.2.1 Aglomerativo:

donde cada observación se considera inicialmente como un cluster propio (hoja), luego los cluster con más similitud se fusionan sucesivamente hasta que solo hay un único gran cluster (raíz).

4.2.2 Divisivo:

comienza con la raíz, en la que todos los objetos se incluyen en un grupo. Luego, los cluster más heterogéneos se dividen sucesivamente hasta que todas las observaciones se encuentran en su propio cluster (hojas).

Los clusters aglomerativos son buenos para identificar pequeños grupos. Los cluster divisivos son buenos para identificar grandes agrupaciones (1). En la siguiente gráfica podemos entender mejor la diferencia entre una y otra.

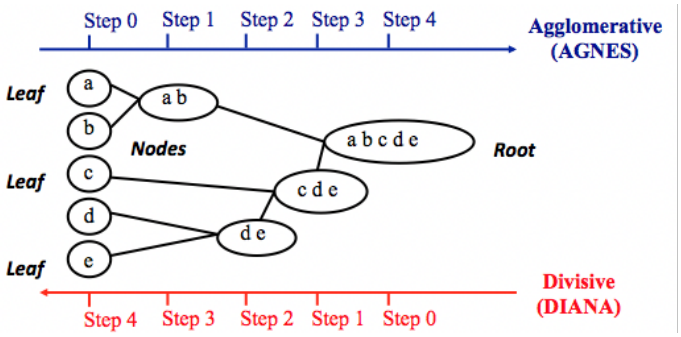


Imagen 5: Explicación de Clustering Aglomerativo y Divisivo

4.3 Pruebas de Clustering

Para el desarrollo del análisis se realizaron pruebas de clustering jerárquico aglomerativo y divisivo, con la finalidad de determinar cuál es el que mejor se ajusta. Las agrupaciones se encuentran pintadas de un color (rojo, verde, celeste y morado), así mismo, se están utilizando colores similares para las provincias que corresponden a una misma comunidad autónoma.

En las imágenes 6 y 7 podemos visualizar los resultados obtenidos a partir de los clustering jerárquico aglomerativo basado en distancia euclidiana y divisivo basado en distancia euclidiana respectivamente.

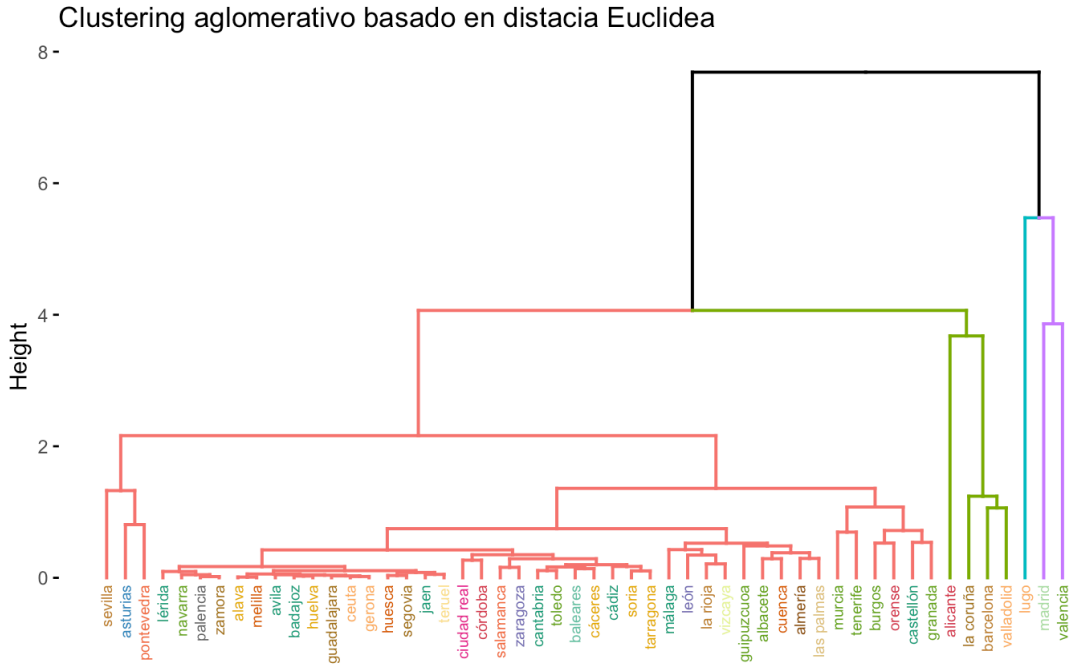


Imagen 6: Clustering Aglomerativo basado en distancia Euclidiana

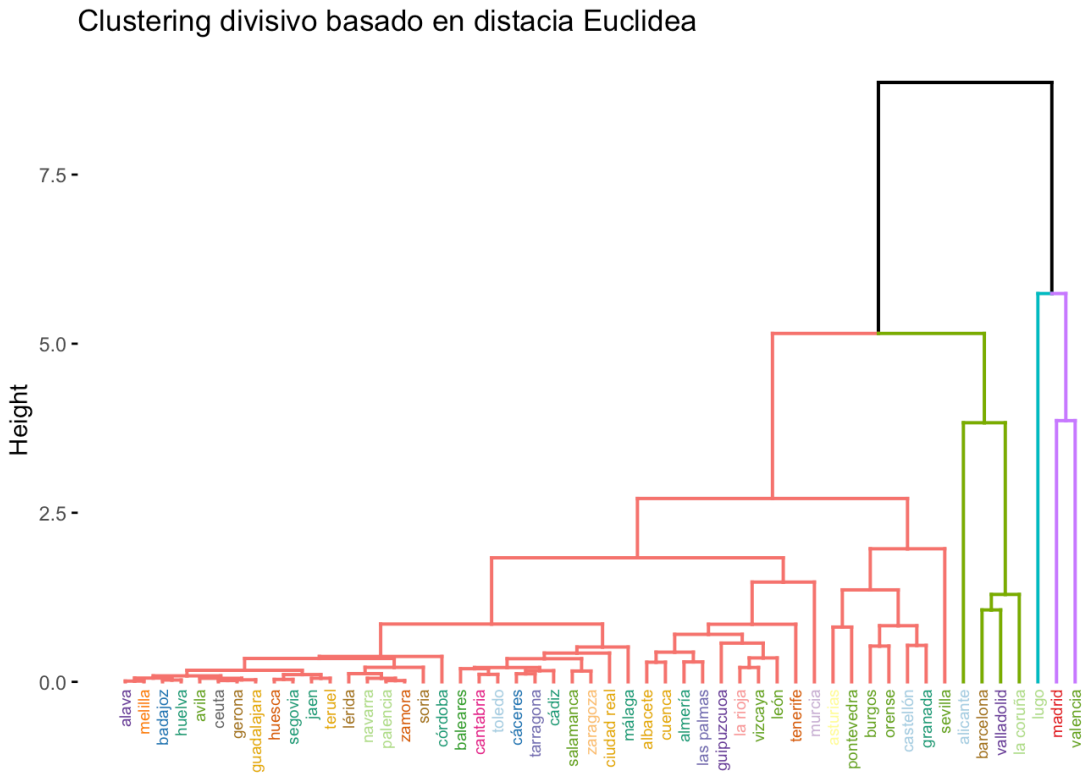


Imagen 7: Clustering Divisivo basado en distancia Euclidiana

5 Proyección para los proximos años

Para el desarrollo de la proyección de ventas para los próximos años se realizó el método de regresión lineal.

El objetivo del análisis de regresión como método causal es pronosticar la demanda a partir de una o más causas (variables independientes), las cuales pueden ser por ejemplo el tiempo, precios del producto o servicio, precios de la competencia, economía del país, acciones del gobierno o fomentos publicitarios.

Algunos aportes importantes sobre este método son:

Se pueden calcular series de tiempo y relaciones causales. En el primer caso, se ubica la demanda histórica de bien o servicio en análisis, para que cambie en función del tiempo.

El segundo caso es cuando la variable que se pronostica cambia en función de otra (variable causal).

Lineal significa que los datos del periodo anterior y la proyección para el periodo futuro que se va a obtener se registran sobre una recta.

Si el análisis es sobre una sola variable independiente, es una regresión lineal simple, contrario a si son dos o más variables independientes, donde se estaría analizando una regresión lineal múltiple.

5.1 Manipulación de datos

Previamente al desarrollo de la regresión lineal es necesario trabajar con la base de datos con la finalidad de tener los datos ordenados y hacer así mucho más sencillo de realizar el análisis.

5.1.1 Conversión de variable mes a trimestre

En la variable mes encontramos la información por nombres y no por códigos, esto hace que ocupe mucho espacio además no permite realizar la regresión lineal por lo que se procedió a asignar un código de trimestre (entre 01, 02, 03 y 04), dependiendo al trimestre que le corresponde, posteriormente se procedió a crear la variable trimestre.

Mes	Trimestre	Mes	Trimestre
Enero	01	Julio	03
Febrero	01	Agosto	03
Marzo	01	Setiembre	03
Abril	02	Octubre	04
Mayo	02	Noviembre	04
Junio	02	Diciembre	04

Imagen 8: Asignación trimestre a variable mes

5.1.2 Concatenación de variable año con trimestre

Con la finalidad de tener la secuencia de tiempo se procedió a concatenar la variable año con trimestre, sin embargo, no era muy práctico por lo que se procedió a reemplazar esta información a número de trimestre, quedando de la manera siguiente:

Número	Variable Trimestre	
	Inicial	Final
01	201501	1
02	201502	2
03	201503	3
04	201504	4
05	201601	5
06	201602	6
07	201603	7
08	201604	8
09	201701	9
10	201702	10
11	201703	11
12	201704	12

Imagen 9: Modificación de la variable trimestre

Realizando este último cambio se continuó desarrollando el trabajo sin inconvenientes, lo siguiente fue realizar la sumatoria de la variable Total y agruparlas por la variable Trimestre.

5.1.3 Sumatoria y agrupación de variables.

Finalmente, se procedió a retirar los datos que no utilizaremos, sumar los totales y agrupar por la variable Trimestre.

```
datosc <- datosc %>%
  select(Total, Trimestre) %>%
  mutate(
    Trimestre = case_when (
      Trimestre =="201501" ~ 1,
      Trimestre =="201502" ~ 2,
      Trimestre =="201503" ~ 3,
      Trimestre =="201504" ~ 4,
      Trimestre =="201601" ~ 5,
      Trimestre =="201602" ~ 6,
      Trimestre =="201603" ~ 7,
      Trimestre =="201604" ~ 8,
      Trimestre =="201701" ~ 9,
      Trimestre =="201702" ~ 10,
      Trimestre =="201703" ~ 11,
      Trimestre =="201704" ~ 12)
    )
datosu <- aggregate(datosc$Total, by=list(Trimestre= datosc$Trimestre), FUN= sum)
```

Imagen 10: Sumatoria de totales y agrupación de trimestres

5.2 Cumplimiento del supuesto del modelo de regresión lineal

En la siguiente gráfica podemos observar el cumplimiento del supuesto del modelo de regresión lineal.

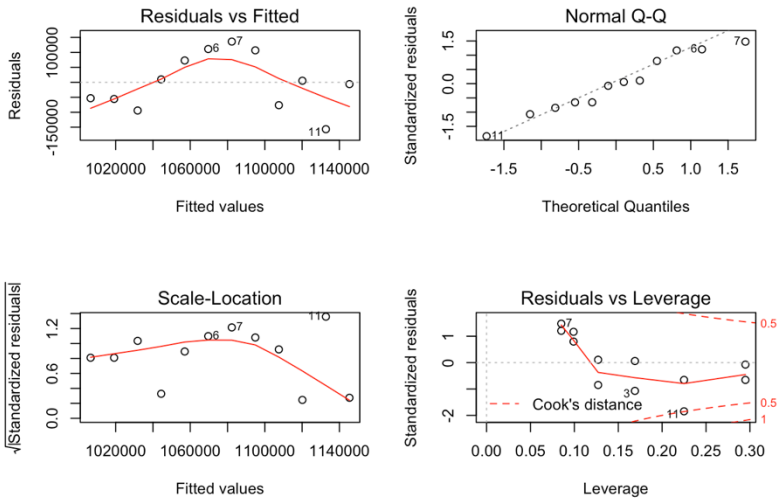


Imagen 11: Cumplimiento del supuesto del modelo de regresión lineal

5.3 Predicción

Se realizara una predicción para los trimestres 13, 14 , 15 y 16

```
nuevo.trimestre <- data.frame(Trimestre = seq(13,16))
nuevo.trimestre
```

Trimestre <int>
13
14
15
16

4 rows

```
predict(regresion, nuevo.trimestre)
```

```
##       1       2       3       4
## 1157999 1170620 1183242 1195863
```

Imagen 12: Resultados de la regresión lineal

Las predicciones obtenidas son las siguientes:

13: 1.157.999

14: 1.170.620

15: 1.183.242

16: 1.195.863

Conclusiones y aplicaciones

1- Entre los años 2015 y 2017 se registraron las mayores ventas en los productos relacionados a la familia Café, incluso se han incrementado.

Los productos de las familias infusiones; chocolates, galletas y otros se han ido incrementando, mientras que en solubles han ido decreciendo.

Caso extraño ocurre con los productos de la familia azúcar y edulcorante que durante el 2016 se incrementó, sin embargo, en el 2017 no se registran ventas.

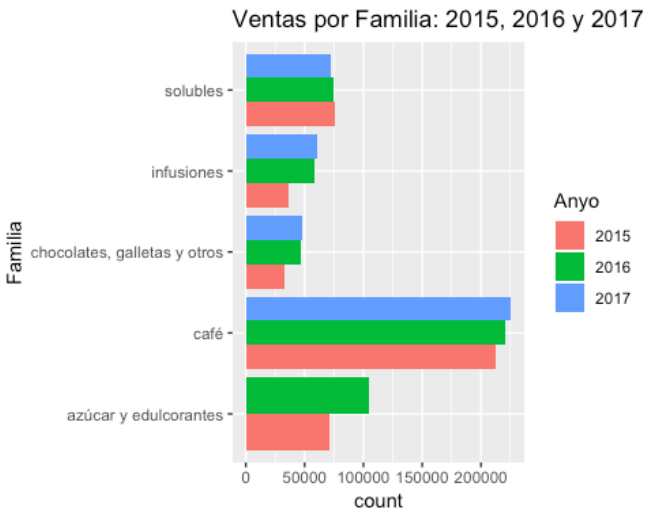


Imagen 13: Ventas a nivel de indicador Familia (años 2015, 2016 y 2017)

Filtrando la información de la familia Café, se puede visualizar que ***the essential coffee*** es el producto que registra las mayores ventas, sin embargo, esta ha ido decreciendo; mientras que el segundo producto significativo es ***the premium coffee*** y que además sus ventas se han incrementado.

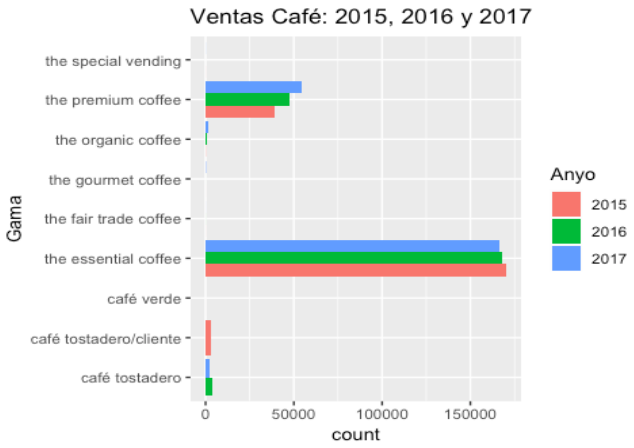


Imagen 14: Ventas de la familia Café (años 2015, 2016 y 2017)

2- Una de las principales aplicaciones del análisis clustering desarrollado es el poder identificar a los grupos de provincias de España con preferencias similares de consumo y poder orientar estrategias a cada uno de ellos para que la empresa pueda alcanzar los objetivos más fácilmente y de la manera óptima

   De acuerdo a las ventas registradas durante los años 2015, 2016 y 2017, se puede determinar que existen similares tendencias de consumo en las provincias ordenadas de acuerdo a las siguientes agrupaciones: Grupo 1 (Lugo, Madrid y Valencia), Grupo 2 (Alicante, Barcelona, La Coruña y Valladolid) y Grupo 3 (Provincias restantes de España). Por lo que Café Candelas debería de emplear similares técnicas de marketing para la segmentación del mercado, de estos grupos.

3- Para el desarrollo del análisis de la información Café Candelas exportó la información desde el ERP que utilizan, sin embargo, desde Rstudio se pudieron importar los datos directamente, por lo que se recomienda Café Candelas el uso de esta herramienta.

4- Para el desarrollo de este trabajo de fin de máster, se trabajó con la información agrupada a nivel de familia por que lo que se recomienda a Café Candelas que se realicen análisis por cada uno de los productos de la variable Jerarquía.

Bibliografía

1. Trevor Hastie, Roberts Tibshirani y Jerome Friedman. *The Elements os Statistical Learning Data Mining, Inference, and Prediction.* s.l. : Springer.

2. *Café Candelas.* [En línea] enero de 2018. https://www.cafescandelas.com.

3. James Gareth, y otros. *An Introduction to Statistical Learning with Applications in R.* s.l. : Springer.

4. *scikit-learn Machine Learning in Python.* [En línea] http://scikit-learn.org/stable/.

5. matplotlib. [En línea] https://matplotlib.org.

6. Wikipedia. [En línea] https://en.wikipedia.org/wiki/Feature_engineering.

7. Kassambara, Alboukadel. *Practical Guide To Cluster Analysis in R.* s.l. : STHDA, 2017.

8. Wikipedia. [En línea] 25 de julio de 2012. [Citado el: 15 de junio de 2018.] https://es.wikipedia.org/wiki/Código_postal.