

## 0.1. Modelos estadísticos clásicos

El objetivo principal de este apartado es aplicar los modelos estadísticos descritos en la parte teórica a un conjunto de datos correspondiente a ventas de dos productos lácteos de dos marcas distintas.

En primer lugar, se aplicarán distintos modelos estadísticos clásicos con el objetivo de entender como funcionan las ventas y predecir el volumen de ventas de ambos productos en el futuro. Con este objetivo, trataremos de identificar que variables son las mas adecuadas para la realización de las predicciones. También estudiaremos la existencia de variables que no aportan información adicional al volumen de ventas, y son por tanto irrelevantes para este estudio.

Para llevar a cabo este estudio haremos uso de la librería -

### 0.1.1. Lectura y descripción de los datos

Los datos contienen información correspondiente a ventas de productos lácteos durante un período de 5 meses, desde el 1 de Septiembre de 2020 hasta el 30 de Enero de 2021, obteniéndose un total de 140025 observaciones y se estructuran de la siguiente forma: Cada fila corresponde a la línea de un ticket y hace referencia a la venta de un artículo en particular. De esta forma, para una venta de un único producto únicamente encontraremos una fila en el conjunto de datos que identifique esa venta, sin embargo, para una venta de más de un artículo encontraremos tantas filas como productos se hayan vendido. Como únicamente tenemos las ventas para dos productos, como máximo, habrá dos filas en el conjunto de datos para una misma venta.

En este conjunto de datos inicial encontramos las siguientes variables:

- **ID\_TICKET**: Variable numérica que identifica unívocamente a cada ticket de compra
- **LINEA\_TICLET**: Variable numérica con la línea correspondiente del ticket
- **FECHA**: Fecha en que se realizó la transacción
- **CODIGO**: Id producto
- **CANTIDAD**: Número de artículos comprados
- **PRECIO**: Precio base del artículo libre de impuestos, euros
- **PRECIO CON IMPUESTOS**: Precio de venta del artículo, en euros
- **DESCUENTO**: Descuento aplicado
- **IMPORTE**: Importe de la compra libre de impuestos, en euros
- **IMPORTE CON IMPUESTOS**: Importe a pagar por el comprador, en euros

Se ha considerado oportuno la extracción de la siguiente información como nuevas variables temporales: día de la semana, semana del año, mes y año de cada instancia a partir de la variable *fecha* y haciendo uso de la librería lubridate. De esta forma, se podrá hacer un análisis del comportamiento de ventas teniendo en cuenta distintas granularidades tratando de entender cómo afecta la temporalidad a la venta de productos.

Al estar trabajando con fechas, es muy importante comprobar la uniformidad en los datos, para ello buscaremos la existencia de registros faltantes de la siguiente forma:

```
# Construcción de un cjto de datos con todas las fechas entre la primera fecha y la  
FechasCompletas <- seq(min(dataset$FECHA), max(dataset$FECHA), by = "day")
```

```

FechasCompletas <- data.frame(FECHA = FechasCompletas ) # Creo un DF con de fechas
# Merge al conjunto de fechas completas y al cjto inicial para añadir NA a aquellos
DatosCompletos <- merge(FechasCompletas, dataset, by = "FECHA", all.x = TRUE)

# Valores faltantes en el conjunto de datos completo
Miss_values <- which(is.na(DatosCompletos$ANO) == TRUE)

```

Con registros faltantes nos referimos a que falten las ventas correspondientes a algún día concreto dentro del período que estamos considerando, 1/09/2020-30/01/2022.

Existe un total de 2 valores faltantes, que corresponden a un 0.00143% del total de datos. Se trata de un porcentaje ínfimo del total. En otras condiciones, procederíamos a imputar estos valores, sin embargo, estos días no estaban contemplados en el conjunto de datos inicial debido a que corresponden a festivos: 2020-12-25, el día de Navidad y 2021-01-01, año nuevo. Por este motivo, podemos continuar con nuestro análisis haciendo uso del conjunto de datos inicial.

Una vez hemos añadido las variables adecuadas y hemos estudiado la continuidad de los datos, vamos a realizar un breve resumen descriptivo de los datos para así conocer la estructura y el tipo de éstos y ver si es necesario la modificación de alguna variable.

```
dataset %>% str() # Estructura de los datos tras su transformación
```

```

## 'data.frame':    140025 obs. of  15 variables:
## $ ID_TICKET      : Factor w/ 97143 levels "22549194","22549215",...: 1 2 3 3
## $ LINEA_TICKET   : Factor w/ 89 levels "1","2","3","4",...: 1 1 1 2 3 7 1 2
## $ CODIGO         : Factor w/ 2 levels "20445","22336": 2 1 2 2 2 2 2 2 1
## $ CANTIDAD       : int  1 6 6 6 1 1 6 1 1 5 ...
## $ PRECIO         : num  1.35 1.26 1.35 1.35 1.35 1.35 1.35 1.35 1.35 1.26 .
## $ PRECIO_CON_IMPUESTOS : num  1.49 1.39 1.49 1.49 1.49 1.49 1.49 1.49 1.49 1.39 .
## $ DESCUENTO      : num  0 0 0 0 0 0 0 0 0 0 ...
## $ IMPORTE        : num  1.35 7.58 8.13 8.13 1.35 1.35 8.13 1.35 1.35 6.32 .
## $ IMPORTE_CON_IMPUESTOS: num  1.49 8.34 8.94 8.94 1.49 1.49 8.94 1.49 1.49 6.95 .
## $ FECHA          : Date, format: "2020-08-01" "2020-08-01" ...
## $ ANO            : int  2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 .
## $ MES            : int  8 8 8 8 8 8 8 8 8 8 ...
## $ DIA            : int  1 1 1 1 1 1 1 1 1 1 ...
## $ SEMANA_ANO     : int  31 31 31 31 31 31 31 31 31 31 ...
## $ DIA_SEMANA     : int  6 6 6 6 6 6 6 6 6 6 ...

```

```
dataset %>% summary() # Resumen de los datos
```

```

##      ID_TICKET      LINEA_TICKET      CODIGO      CANTIDAD
## 23776254:      22      1      :33883  20445:70951  Min.   : 1.000
## 23940599:      16      2      :26223  22336:69074  1st Qu.: 1.000
## 23596154:      14      3      :19016                Median : 1.000
## 24800697:      13      4      :13684                Mean   : 2.557
## 25037783:      13      5      : 9787                3rd Qu.: 6.000
## 22797480:      12      6      : 7310                Max.   :132.000
## (Other) :139935  (Other):30122

```

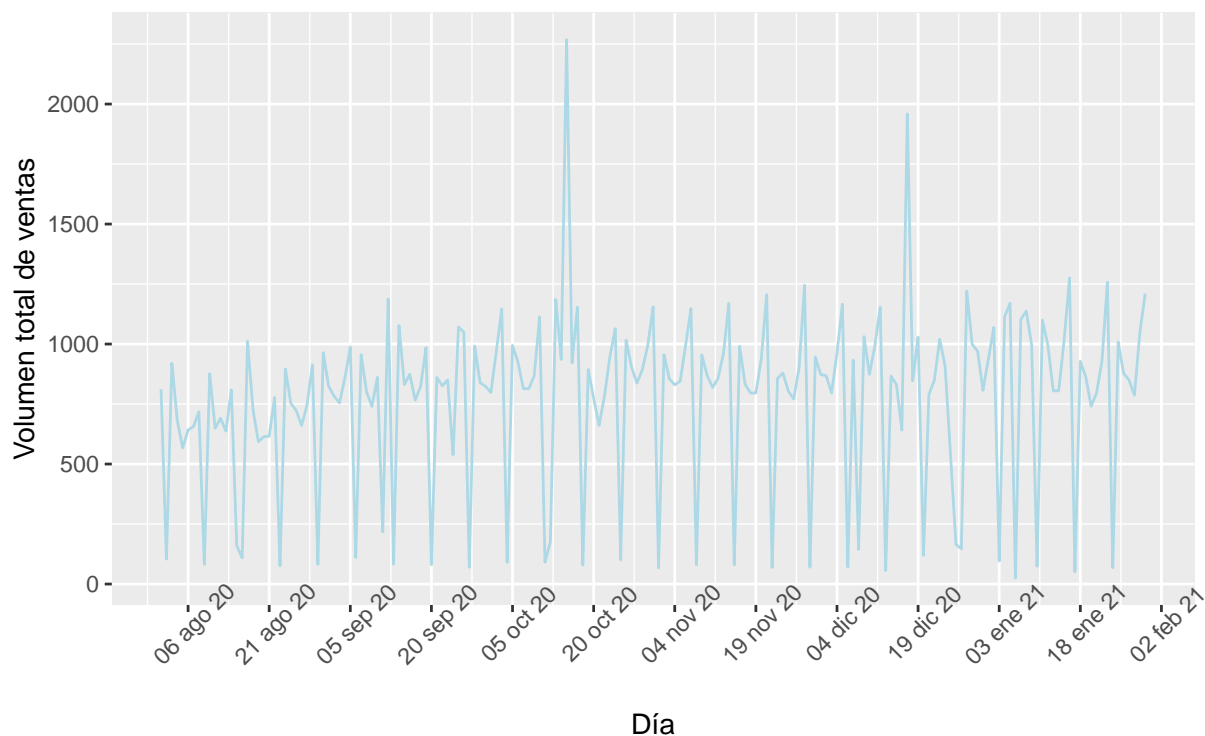
```

##          PRECIO          PRECIO_CON_IMPUESTOS          DESCUENTO          IMPORTE
## Min.      :0.690      Min.      :0.760      Min.      : 0.00000      Min.      : 0.630
## 1st Qu.:1.260      1st Qu.:1.390      1st Qu.: 0.00000      1st Qu.: 1.260
## Median :1.260      Median :1.390      Median : 0.00000      Median : 1.350
## Mean     :1.305      Mean     :1.439      Mean     : 0.09362      Mean     : 3.349
## 3rd Qu.:1.350      3rd Qu.:1.490      3rd Qu.: 0.00000      3rd Qu.: 7.580
## Max.     :1.480      Max.     :1.490      Max.     :50.00000      Max.     :178.800
##
## IMPORTE_CON_IMPUESTOS      FECHA          ANO          MES
## Min.      : 0.690      Min.      :2020-08-01      Min.      :2020      Min.      : 1.000
## 1st Qu.: 1.390      1st Qu.:2020-09-22      1st Qu.:2020      1st Qu.: 8.000
## Median : 1.490      Median :2020-11-03      Median :2020      Median :10.000
## Mean     : 3.685      Mean     :2020-11-03      Mean     :2020      Mean     : 8.521
## 3rd Qu.: 8.340      3rd Qu.:2020-12-17      3rd Qu.:2020      3rd Qu.:11.000
## Max.     :196.680      Max.     :2021-01-30      Max.     :2021      Max.     :12.000
##
##          DIA          SEMANA_AÑO          DIA_SEMANA
## Min.      : 1.00      Min.      : 1.00      Min.      :1.000
## 1st Qu.: 8.00      1st Qu.:33.00      1st Qu.:2.000
## Median :16.00      Median :40.00      Median :4.000
## Mean     :15.98      Mean     :35.47      Mean     :3.614
## 3rd Qu.:23.00      3rd Qu.:46.00      3rd Qu.:5.000
## Max.     :31.00      Max.     :53.00      Max.     :7.000
##

```

A continuación, podemos ver gráficamente la evolución del volumen total de ventas diarias:

**Evolución del volumen total de ventas diario**



Fuente: Elaboración propia con datos

En el gráfico podemos apreciar como el volumen de ventas diario fluctúa bastante en

Mes	Volumen Ventas	Precio Medio
Enero	24195	1.440358
Agosto	19270	1.438435
Septiembre	22523	1.438215
Octubre	26365	1.438526
Noviembre	23584	1.439412
Diciembre	24088	1.440203

Día	Volumen Ventas	Precio Medio
Lunes	24301	1.439475
Martes	21762	1.439216
Miércoles	20093	1.438421
Jueves	23132	1.439497
Viernes	21217	1.439355
Sábado	27312	1.439466
Domingo	2208	1.436332

función del día, encontrándolas en un rango entre 25 y 2268 ventas diarias. Hay dos momentos donde el volumen de ventas es considerablemente superior al resto, a mediados del mes de Octubre de 2020 y a mediados de Diciembre de este mismo año. Se observa un patrón muy marcado, con picos de muy pocas ventas y otros donde el volumen sube bastante.

Para tratar de entender mejor el comportamiento de venta, vamos a estudiar la evolución de los valores de ventas en función del día de la semana y mes del año.

