0.1. Proceso de ciencia de datos

Este apartado lo dedicaremos a realizar un proceso de ciencia de datos completo, teniendo en cuenta los siguientes objetivos:

- Analizar los datos proporcionados para conocer como varían las ventas de productos lácteos con el tiempo
- Demostrar que existe la posibilidad de construir buenos modelos para predecir el volumen futuro de venta de productos a partir de los datos.
- Desarrollo de modelos para predecir ventas.

0.1.1. Lectura y descripción de los datos

Los datos contienen información correspondiente a ventas de dos productos lácteos (con y sin calcio) 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.

En este conjunto de datos inicial encontramos las siguientes variables:

- **ID_TICKET**: Variable numérica que identifica unívocamente a cada ticket de venta.
- **LINEA_TICLET**: Variable numérica con la línea correspondiente del ticket.
- **FECHA**: Fecha en que se realizó la venta.
- **CODIGO**: Identificador del producto
- **CANTIDAD**: Número de items vendidos de un determinado producto
- **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.

0.1.2. Preparación de los datos (Preprocesado)

En este paso, vamos a llevar a cabo la limpieza de los datos para su posterior estudio, representación y modelado. En este punto del proceso, trataremos de encontrar, corregir o eliminar registros erróneos en los datos.

0.1.2.1. Transformación de los datos

Hay algunas variables que necesitan ser transformadas, en particular, la variable código, la cantidad de artículos vendidos y linea del ticket han sido transformadas para tenerlas en un formato adecuado. Vamos a visualizar la nueva estructura de los datos:

```
dataset %>% str() # Estructura de los datos trás reformateo
```

```
## 'data.frame':    140025 obs. of  10 variables:
##  $ ID_TICKET            : num  22549194 22549215 22549242 22549242 22549264 ...
##  $ LINEA_TICKET         : Factor w/ 89 levels "1","2","3","4",..: 1 1 1 2 3 7 1 2 
##  $ FECHA                : Date, format: "2020-08-01" "2020-08-01" ...
```

```
##  $ CODIGO               : Factor w/ 2 levels "20445","22336": 2 1 2 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 .
```

0.1.2.2. Duplicados

Se comprueba la existencia de registros duplicados y concluímos que no existen duplicados.

```
dataset[duplicated(dataset)==TRUE,]
```

```
##  [1] ID_TICKET             LINEA_TICKET          FECHA
##  [4] CODIGO                CANTIDAD              PRECIO
##  [7] PRECIO_CON_IMPUESTOS  DESCUENTO             IMPORTE
## [10] IMPORTE_CON_IMPUESTOS
## <0 rows> (or 0-length row.names)
```

0.1.2.3. Datos faltantes

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:

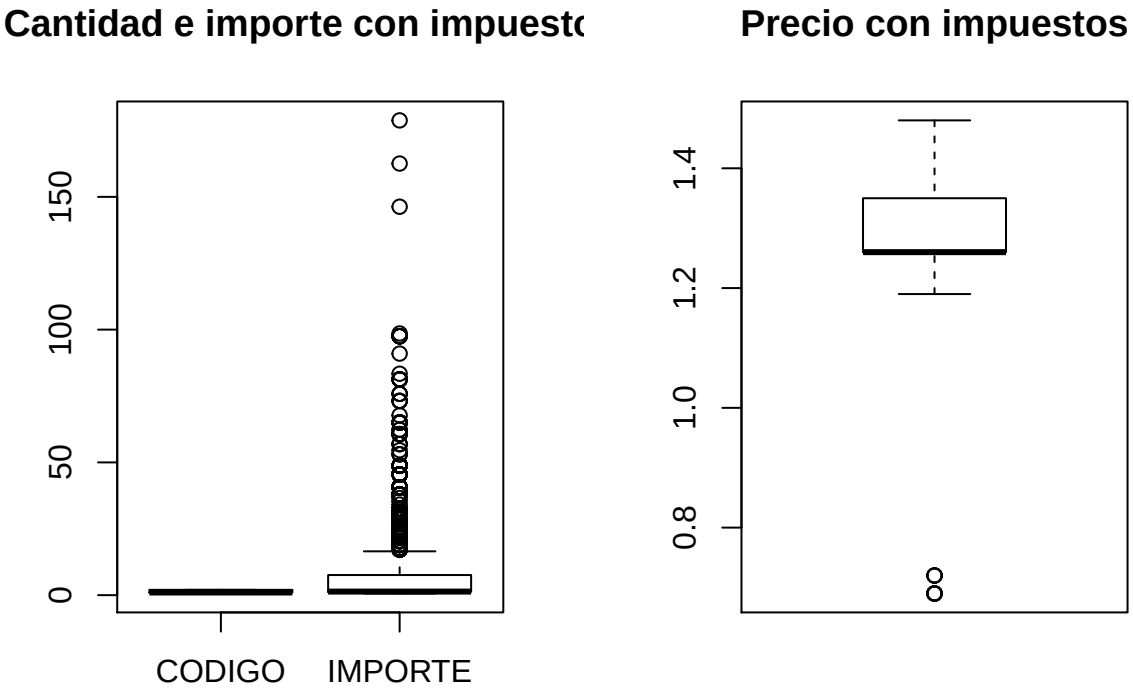
```
# Construcción de un cjto de datos con todas las fechas entre la primera fecha
#  y la última de los datos que tenemos
FechasCompletas <- seq(min(dataset$FECHA), max(dataset$FECHA), by = "day")

# Creo un DF de fechas
FechasCompletas  <- data.frame(FECHA = FechasCompletas )

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

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

0.1.2.4. Outliers

Estudiamos valores atípicos para las variables *precio*, *cantidad* e *importe con impuestos*, ya que estudiarlos para el resto no tiene mucho interés.



Para el precio de los productos (impuestos incluídos), un importe menor de 1.31 se podría considerar extremo, existiendo dos valores atípicos, con unos precios de vneta de 79 y 76 céntimos. Para la variable cantidad, se considerará un valor atípico la compra de 14 o más artículos. Para la variable importe con impuestos, una compra de más de 18.07€ podemos considerarla como una compra con un valor extremo.

0.1.2.5. Creación de variables

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.

```
dataset$ANO         <-  year(dataset$FECHA) # Extracción del año
dataset$MES         <- month(dataset$FECHA) # Extracción del mes
dataset$DIA         <-   day(dataset$FECHA) # Extracción del día
dataset$SEMANA_ANO <-   week(dataset$FECHA) # Extracción de la semana del año
dataset$DIA_SEMANA <-   wday(dataset$FECHA,week_start = 1 )
```

Además de las variables temporales, vamos a añadir una nueva variable: *TIPO*, que representa el tipo de producto lácteo, siendo esta una variable categórica que toma dos posibles valores: *Con calcio* o *sin calcio*. El producto que no lleva calcio es aquel con identificador 20445

0.1.2.6. Datasets de entrada de modelos

La variable objetivo es el volumen de ventas diario, por lo que necesitamos transformar el conjunto de datos inicial en uno que tenga una variable *ventas* con el volumen total de ventas diario. Obtendremos un datset con el volumen total, otro con el volumen de ventas diario del producto 20445 y un tercer conjunto de datos con el volumen de ventas del producto 22336, para así poder modelar el volumen de ventas en cada ocasión.

```
VolVentas_FECHA = dataset %>% group_by(FECHA) %>%
  dplyr::summarise(VENTAS = sum(CANTIDAD),
                   PRECIO_MEDIO_IMPUESTOS = mean(PRECIO_CON_IMPUESTOS),
                   DESCUENTO_MEDIO = mean(DESCUENTO),
                   IMPORTE_MEDIO_IMPUESTOS=mean(IMPORTE_CON_IMPUESTOS)
                   )
NUM_TRANSACCIONES = as.data.frame(datset %>% group_by(FECHA) %>% dplyr::summarise( u
VolVentas_FECHA$NUM_TRANSACCIONES =NUM_TRANSACCIONES
VolVentas_FECHA=VolVentas_FECHA[,c(1,2,6,3:5)]

VolVentas_CALCIO_FECHA = dataset %>% filter(CODIGO == 22336) %>% group_by(FECHA) %>%
  dplyr::summarise(VENTAS = sum(CANTIDAD),
                   PRECIO_MEDIO_IMPUESTOS = mean(PRECIO_CON_IMPUESTOS),
                   DESCUENTO_MEDIO = mean(DESCUENTO),
                   IMPORTE_MEDIO_IMPUESTOS=mean(IMPORTE_CON_IMPUESTOS) )
NUM_TRANSACCIONES_C = as.data.frame(datset %>%filter(CODIGO == 22336) %>% group_by(
VolVentas_CALCIO_FECHA$NUM_TRANSACCIONES =NUM_TRANSACCIONES_C
VolVentas_CALCIO_FECHA=VolVentas_CALCIO_FECHA[,c(1,2,6,3:5)]



VolVentas_SIN_CALCIO_FECHA = dataset %>% filter(CODIGO == 20445) %>% group_by(FECHA)
  dplyr::summarise(VENTAS = sum(CANTIDAD),
                   PRECIO_MEDIO_IMPUESTOS = mean(PRECIO_CON_IMPUESTOS),
                   DESCUENTO_MEDIO = mean(DESCUENTO),
                   IMPORTE_MEDIO_IMPUESTOS=mean(IMPORTE_CON_IMPUESTOS) )
NUM_TRANSACCIONES_SC = as.data.frame(datset %>%filter(CODIGO == 20445) %>% group_by
VolVentas_SIN_CALCIO_FECHA$NUM_TRANSACCIONES =NUM_TRANSACCIONES_SC
VolVentas_SIN_CALCIO_FECHA=VolVentas_SIN_CALCIO_FECHA[,c(1,2,6,3:5)]
```

Los conjuntos de datos para el entrenamiento de modelo constan de las siguientes variables: fecha, ventas totales, número de transacciones, precio medio con impuestos, descuento medio, importe medio con impuestos, año, mes, día, semana del año y día de la semana.

Una vez tenemos los cuatro conjuntos de datos, tres con los correspondientes volúmenes de ventas diarios y el dataset inicial con el pre procesado necesario, procedemos a guardarlos en formato RData para poder acceder a ellos en cualquier momento que sea necesario.

```
# Guardamos el conjunto de datos final
### save(dataset,file = "Datos/Dataset_Final.RData")
### save(VolVentas_SIN_CALCIO_FECHA,file = "Datos/VENTAS_Dia_SCALCIO.RData")
### save(VolVentas_CALCIO_FECHA,file = "Datos/VENTAS_Dia_CALCIO.RData")
```

```
### save(VolVentas_FECHA,file = "Datos/VENTAS_Dia_TOTAL.RData")
```

0.1.3. Análisis exploratorio de datos (EDA)

Una vez hemos realizado el preprocesamiento de los datos necesario, procedemos a la fase del análisis exploratorio.

Este apartado lo dedicaremos a hacer un análisis profundo de las ventas, añadiendo gráficos que muestren el comportamiento del consumidor.

Como tenemos datos correspondientes a ventas, trataremos de responder a las siguientes cuestiones:

- ¿Cuál es el patrón de venta de cada producto? ¿Se venden las mismas unidades, o destaca la venta de uno de ellos?
- ¿Cómo varían las ventas en función del tiempo?
- ¿Qué variables podrían influir más a la hora de vender un producto?

0.1.3.1. Resumen de los datos

```
VolVentas_FECHA[,c(1:6)] %>% summary()  # Resumen de los datos
```

```
##       FECHA                VENTAS     NUM_TRANSACCIONES PRECIO_MEDIO_IMPUESTOS
##  Min.   :2020-08-01   Min.   :  30   Min.   :  17.0    Min.   :1.418
##  1st Qu.:2020-09-15   1st Qu.:1811   1st Qu.: 472.0    1st Qu.:1.437
##  Median :2020-10-30   Median :2154   Median : 587.0    Median :1.439
##  Mean   :2020-10-30   Mean   :1978   Mean   : 536.7    Mean   :1.439
##  3rd Qu.:2020-12-14   3rd Qu.:2481   3rd Qu.: 680.0    3rd Qu.:1.441
##  Max.   :2021-01-30   Max.   :8011   Max.   :1440.0    Max.   :1.449
##  DESCUENTO_MEDIO    IMPORTE_MEDIO_IMPUESTOS
##  Min.   :0.000000   Min.   :1.716
##  1st Qu.:0.006196   1st Qu.:3.433
##  Median :0.031703   Median :3.601
##  Mean   :0.102185   Mean   :3.580
##  3rd Qu.:0.062889   3rd Qu.:3.752
##  Max.   :4.875000   Max.   :5.139
```

Observamos lo siguiente:

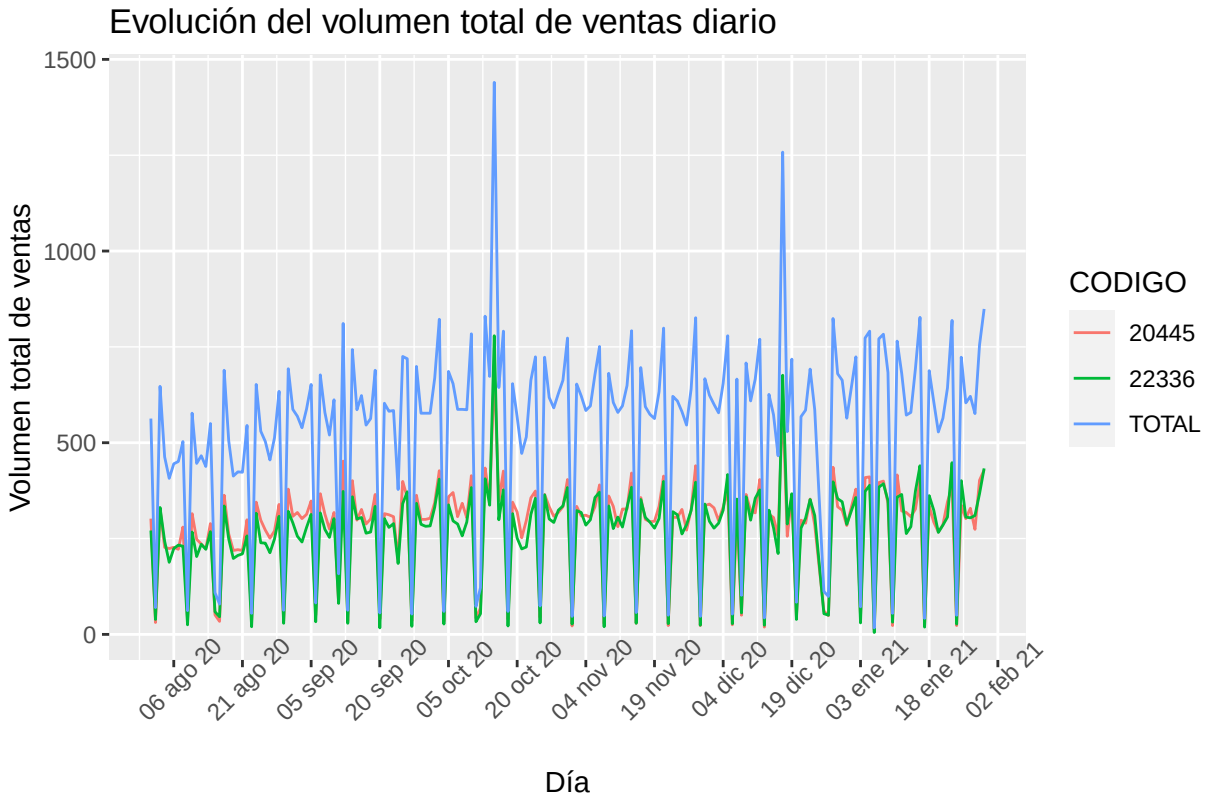
- La cantidad media diaria de productos vendidos es de 1978 unidades, con una media diaria de 537 transacciones
- El precio medio (con impuestos) de los productos es de 1.44€, siendo el importe medio diario de venta de 3.58€.

Analizando el número de items vendidos y el importe de venta, es interesante mencionar que de media, se han vendido cuatro items por transacción, con un precio medio por transacción de 4.25€. El 75 % de las ventas ha sido de seis items o menos, habiendo transacciones donde el número de productos ascendía hasta las 132 unidades.Con respecto al importe de venta (con impuestos), éste ha variando desde los 69 céntimos a los 196.680€. Sin embargo únicamente el 25 % de las transacciones han sido de más de 8.34€. El descuento medio no llega a los 10 céntimos de euro.

Las 140025 filas encontradas en el conjunto de datos corresponden a 97143 ventas diferentes.

0.1.3.2. Representaciones gráficas

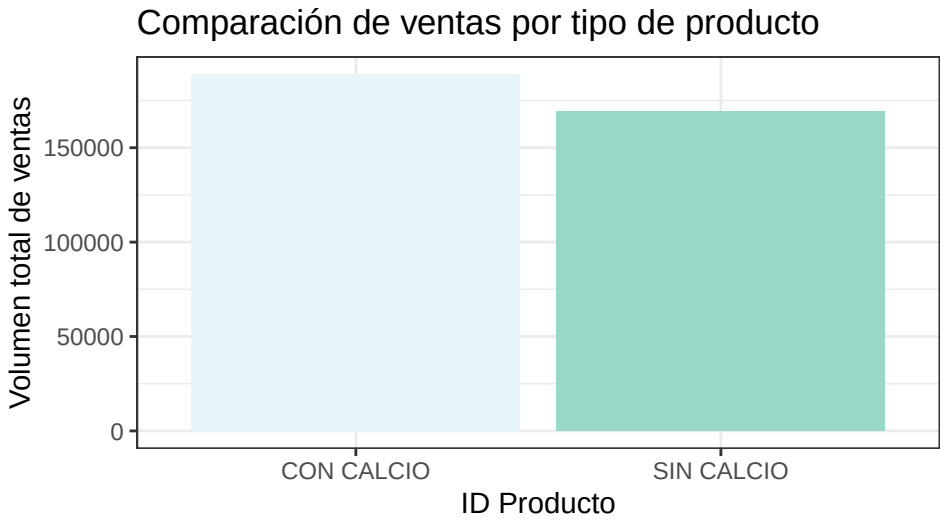
En primer lugar, podemos ver gráficamente la evolución del volumen total de ventas diarias:



Fuente: Elaboración propia con datos

En el gráfico podemos apreciar como el volumen de ventas diario fluctúa bastante en función del día, encontrándolas en un rango entre 17 y 1440 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 ambos productos, sin embargo, podemos afirmar que las ventas son ligeramente superiores para el producto 20445.

A continuación, vamos a hacer una comparación del volumen de ventas total por productos:



Fuente: Elaboración propia con datos de ventas

El volumen de ventas del producto con calcio ha sido ligeramente superior, con un volumen total de ventas de 188867 unidades frente a las 169196 unidades vendidas del producto que no lleva calcio.

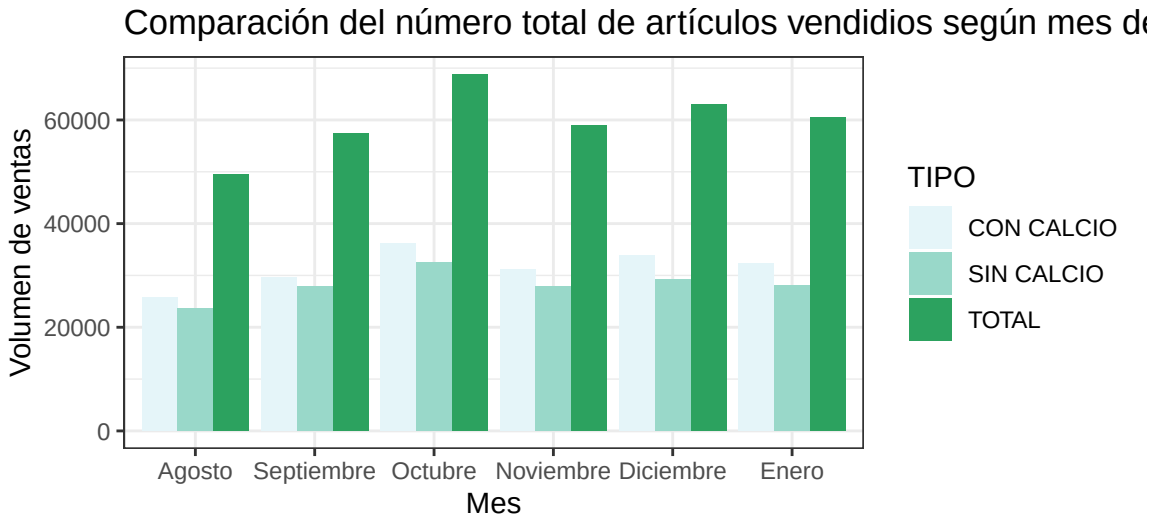
El importe de venta del producto con calcio es de 1.49€ (precio con impuestos) y el producto que no lleva calcio tiene un precio de venta de 1.39€. Esto nos conduce a que los usuarios no han optado siempre por la compra de la opción más económica, sino que han comprado en un mayor número de ocasiones el producto con calcio.

En la tabla que se muestra a continuación, vemos que en las transacciones donde aparece el producto con calcio, han tenido un importe medio de venta ligeramente superior que en las que aparece el producto sin calcio.

Tabla 1: Importe medio de venta y precio por producto.

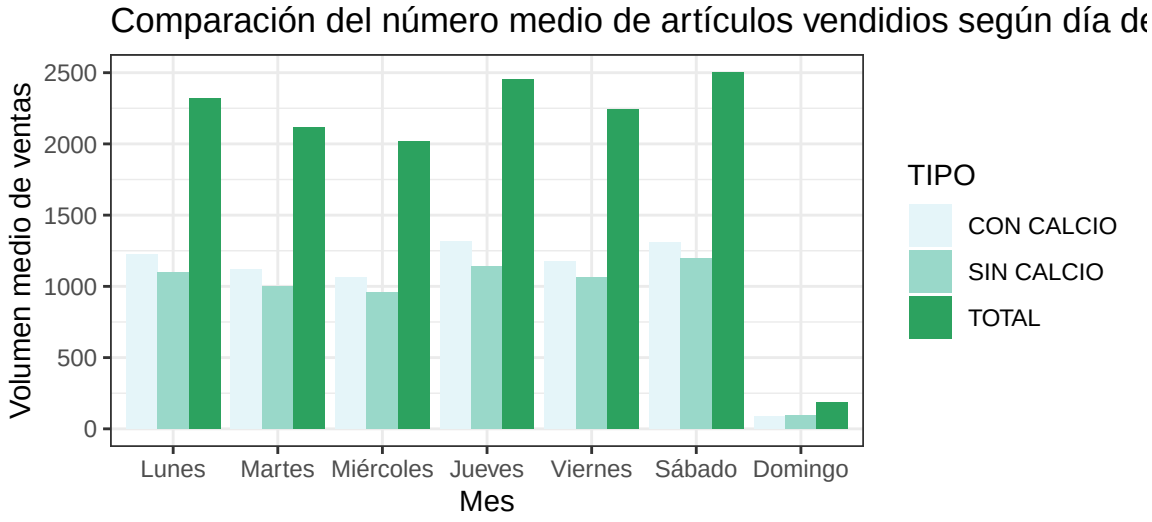
Tipo producto	Importe medio transacción	Precio medio producto
Con calcio	4.07	1.49
Sin Calcio	3.31	1.39

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.



Fuente: Elaboración propia con datos de ventas diarias

Inicialmente, podemos ver una tendencia creciente del número de ventas total, donde se tiende a vender más cada mes.Sin embargo, en el mes de Octubre se alcanza un maximo de artículos vendidos, con un total de 68805 artículos. En los meses siguientes el comportamiento fué de una disminución de la venta, con posterior subida del número de artículos vendidos y luego otro descenso. La tendencia general de venta de artículos es creciente, la clientela tiende a comprar más productos cada vez. Respecto a las ventas de cada producto individual, todos los meses ha sido mayor el volumen de venta del artículo con calcio, alcanzándose en el mes de Octubre el mayor número de artículos vendidos de cada tipo, superando en ambos casos las 32 mil unidades vendidas.



Fuente: Elaboración propia con datos de ventas diarias

Observando el gráfico, vemos que el Sábado es el día de la semana donde el núnmero de artículos vendidos es mayor, con un total de 2323 items. Por el contrario, el Domingo no se superan ni los 250 artículos vendidos, en media. Cabe destacar que los Domingos se ha vendido, de media, mayor número de items sin calcio, comportamiento que no había ocurrido hasta el momento. Los Jueves es un día donde también se tiende a vender gran candidad de productos.

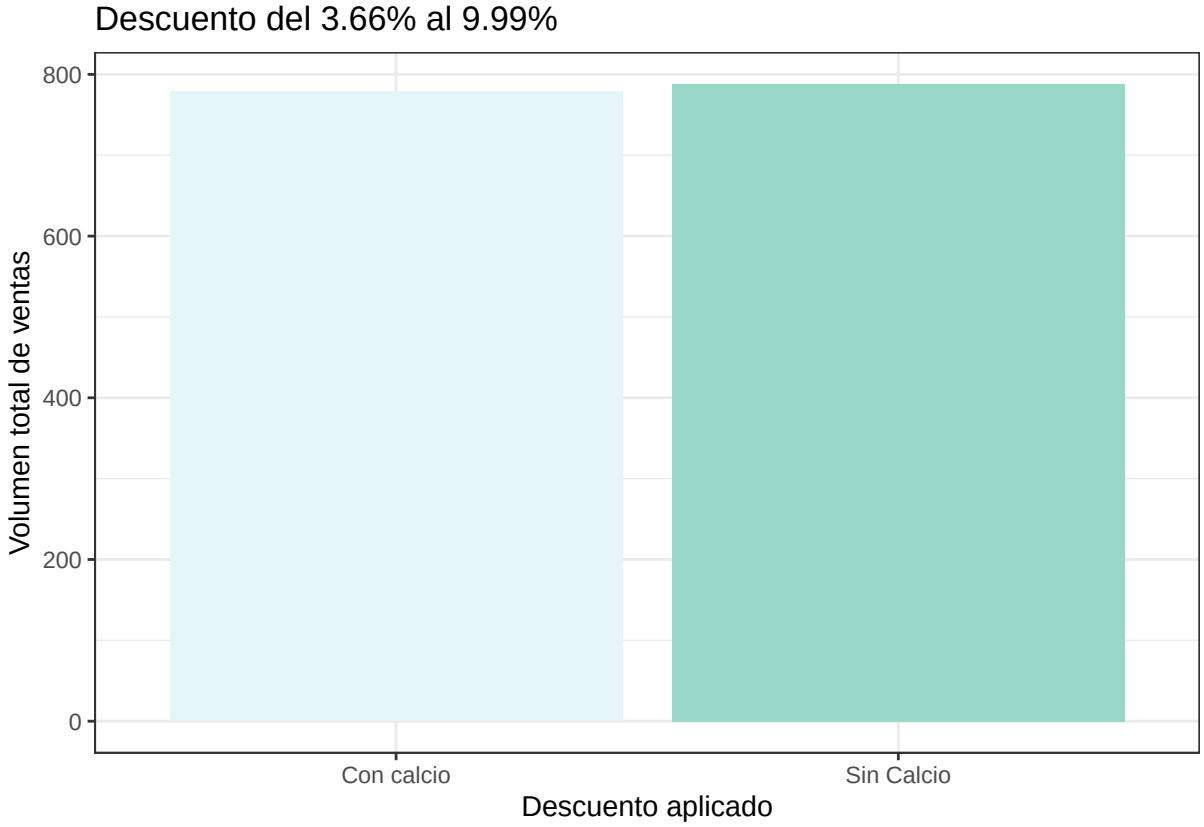
0.1.3.2.1. Variable precio

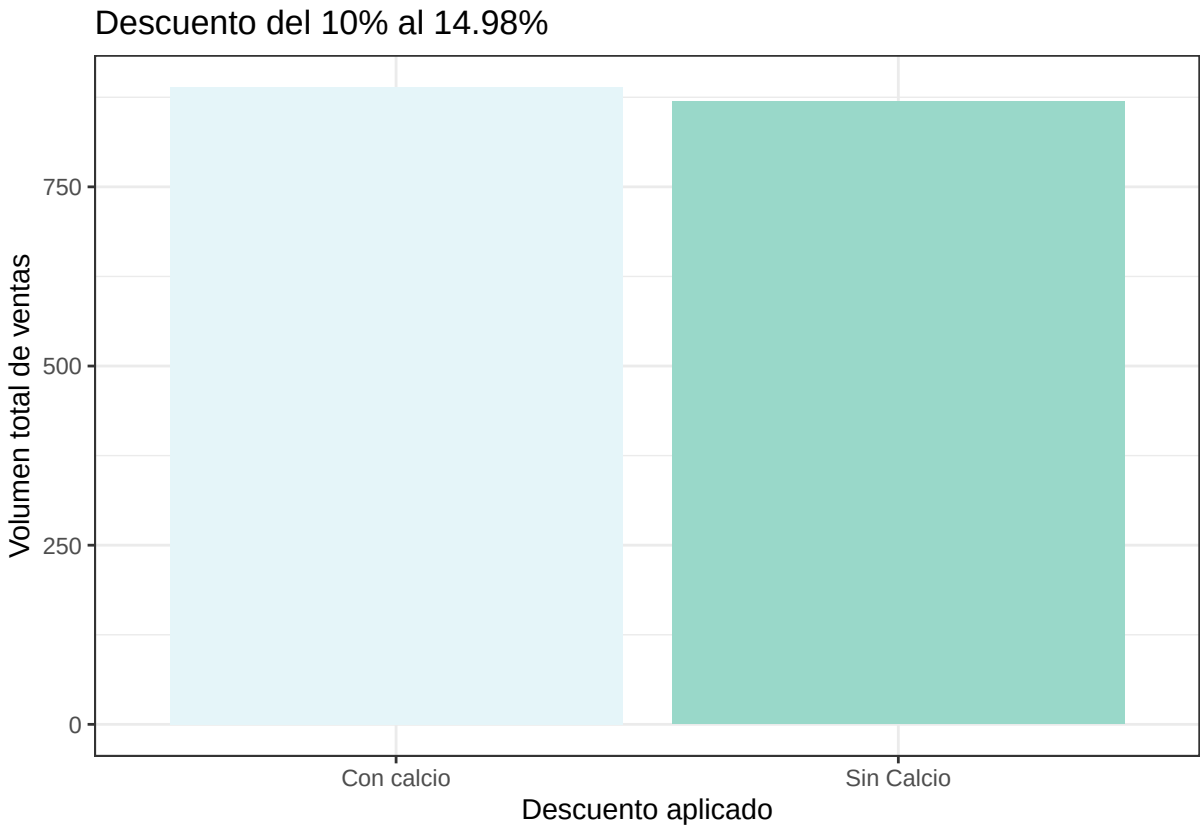
El precio medio de venta del que no lleva calcio es de 1.26€, mientras que para el artículo con calcio es algo superior: 1.35€. Sin embargo, el precio de éstos ha variado a lo largo del tiempo, y con el estudio de esta variable, vamos a tratar de averiguar si esta variación se debe a la época del año o al número de artículos comprado, ya que podría haber promociones para tratar de impulsar las ventas donde los precios pudieran verse afectados. También trataremos de averiguar si las ventas aumentan cuando los precios bajan.

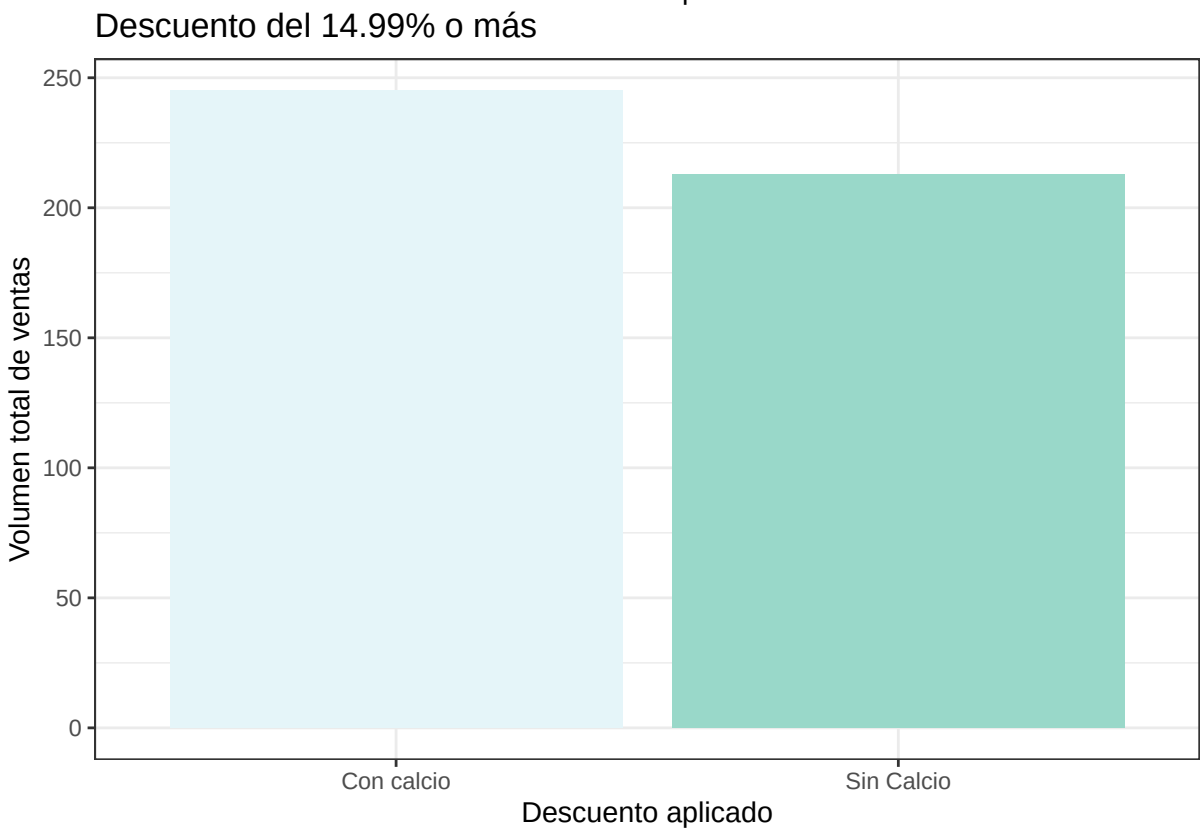
Los precios de los productos varían debido a descuento aplicados a todos los artículos de una misma compra en función del importe de la misma, por ejemplo, si se superan los 50€, se hará un 10 % de descuento, y por tanto, habremos pagado menos por cada artículo. Sin embargo, la variable descuento, se aplica a cada artículo de manera individual, por ejemplo, en la transacción con identificador 22551535 se vendieron 6 unidades del producto 22336 por un precio de 1.49€ (impuestos incluidos) cada unidad, lo que supone un total de 8.94€ por las 6 unidades, pero se ha aplicado un descuento del 5.03 %

Se ha aplicado descuento en un total de 1278 artículos, siendo el descuento medio aplicado del 10.26 % respecto del importe total de la venta. Estos descuentos han sido aplicados durante todo el período temporal estudiado. El mayor descuento aplicado ha sido del 50 %, aplicado en dos ocasiones a mediados de Diciembre. En el gráfico mostrado a continuación, se presenta un histograma comparativo del volumen de ventas total de cada producto en función del descuento habiendo dividido los datos en tres intervalos de tal manera que se ha tomado la distribución de los deciles de esta variable: 3.66, 5, 9.95, 10, 10, 10, 10, 10, 10, 14.99, 50. Lo que podemos observar es que donde se acentúa más la diferencia del volumen de ventas es al aplicar los mayores descuentos, en un 10 % del total de observaciones.

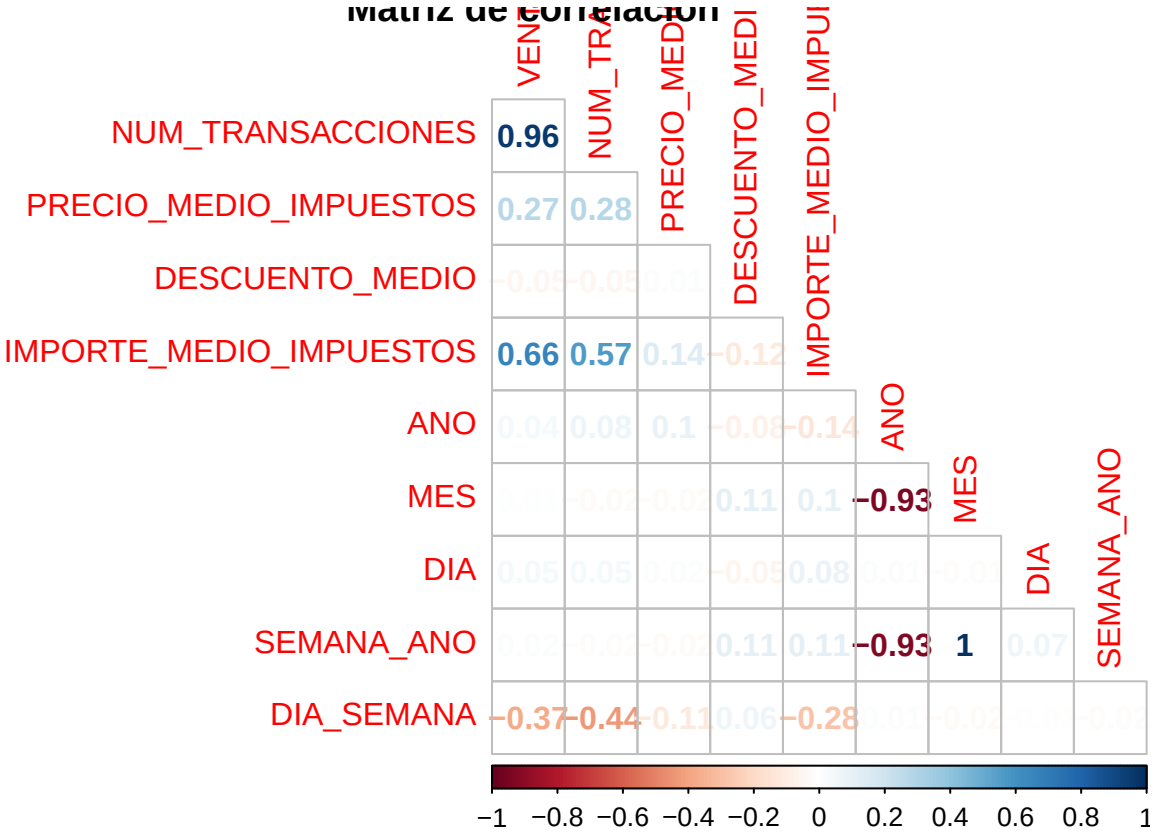
```
##    0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
##  3.66  5.00  9.95 10.00 10.00 10.00 10.00 10.00 10.00 14.99 50.00
```




Descuento del 10% al 14.98%
Volumen total de ventas
750
500
250
0
Con calcio
Sin Calcio
Descuento aplicado


Descuento del 14.99% o más
Volumen total de ventas
250
200
150
100
50
0
Con calcio
Sin Calcio
Descuento aplicado

0.1.3.3. Grado de asociación de las variables



Observamos como el grado de asociación positiva entre la variable número de transacciones diaria y la variable objetivo, ventas diarias (número de items) es muy elevado, indicando así que conforme más transacciones se hayan producido, mayor número de artículos se venderán. También tienen una correlación importante la variable volumen de ventas e importe medio con impuestos de la transacción. El comportamiento es similar en los conjuntos de datos individuales para cada producto.