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 de venta de productos futuro 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 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

0.1.2. Preparación de los datos (Preprocesado)

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

En primer lugar, procedemos a ver la estructura de éstos:

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

```
## tibble [140,025 x 10] (S3: tbl_df/tbl/data.frame)
##  $ ID_TICKET            : num [1:140025] 22549194 22549215 22549242 22549242 22549
##  $ LINEA_TICKET         : num [1:140025] 1 1 1 2 3 7 1 2 3 6 ...
##  $ FECHA                : Date[1:140025], format: "2020-08-01" "2020-08-01" ...
##  $ CODIGO               : chr [1:140025] "22336" "20445" "22336" "22336" ...
```

```
##  $ CANTIDAD             : num [1:140025] 1 6 6 6 1 1 6 1 1 5 ...
##  $ PRECIO               : num [1:140025] 1.35 1.26 1.35 1.35 1.35 1.35 1.35 1.35 1
##  $ PRECIO_CON_IMPUESTOS : num [1:140025] 1.49 1.39 1.49 1.49 1.49 1.49 1.49 1.49 1
##  $ DESCUENTO            : num [1:140025] 0 0 0 0 0 0 0 0 0 0 ...
##  $ IMPORTE              : num [1:140025] 1.35 7.58 8.13 8.13 1.35 1.35 8.13 1.35 1
##  $ IMPORTE_CON_IMPUESTOS: num [1:140025] 1.49 8.34 8.94 8.94 1.49 1.49 8.94 1.49 1
```

A continuación, procedemos a transformar las variables a un formato adecuado y volvemos a visualizar la estructura:

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

```
## '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 .
```

Hemos factorizado algunas variables y convertido otras de las numéricas.

0.1.2.1. Duplicados

Se comprueba la existencia de registros 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)
```

0.1.2.2. 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:

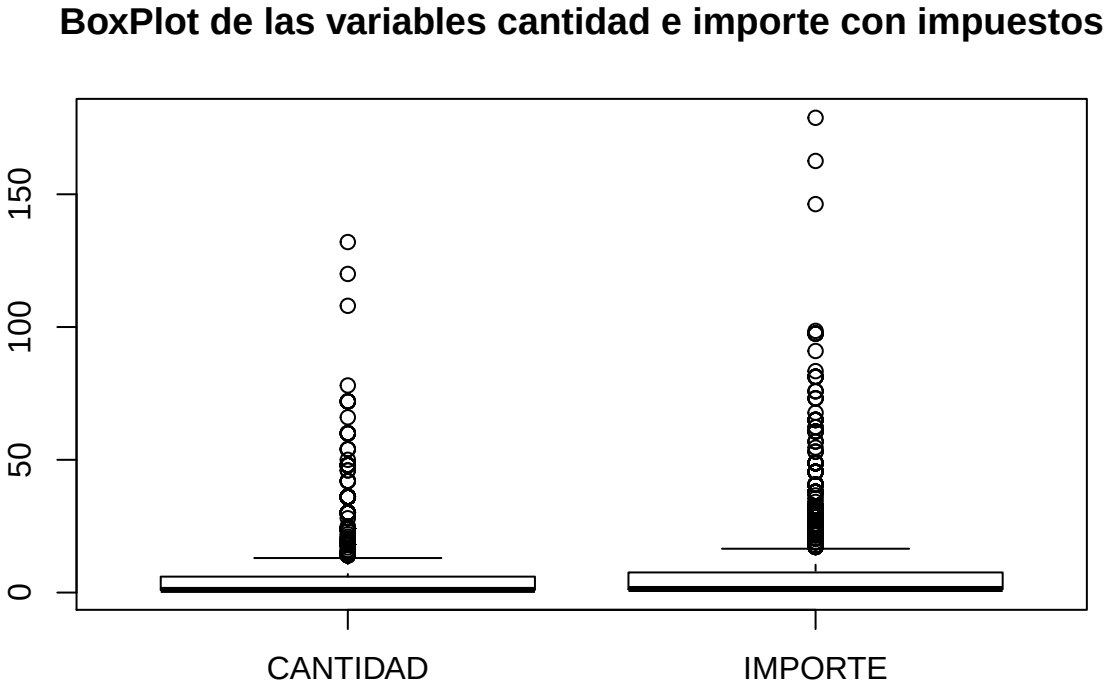
```
# Construcción de un cjto de datos con todas las fechas entre la primera fecha y la
FechasCompletas <- seq(min(dataset$FECHA), max(dataset$FECHA), by = "day")
FechasCompletas  <- data.frame(FECHA = FechasCompletas ) # Creo un DF con de fechas

# 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 0 valores faltantes, que corresponden a un 0 % 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: NA, el día de Navidad y NA, año nuevo. Por este motivo, podemos continuar con nuestro análisis haciendo uso del conjunto de datos inicial.

0.1.2.3. Outliers

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



Para la variable cantidad, se considerará un valor atípico la compra de 18 o más artículos. Para la variable precio con impuesto, una compra de 25.02€ podemos considerarla como una compra con un valor extremo.

0.1.2.4. Creación de variables temporales

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 )
```

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

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

```
##    ID_TICKET          LINEA_TICKET     CODIGO          CANTIDAD
##  Min.   :22549194   1      :33883   20445:70951   Min.   :  1.000
##  1st Qu.:23248501   2      :26223   22336:69074   1st Qu.:  1.000
##  Median :24035461   3      :19016                 Median :  1.000
##  Mean   :24140551   4      :13684                 Mean   :  2.557
##  3rd Qu.:25029149   5      : 9787                 3rd Qu.:  6.000
##  Max.   :25912547   6      : 7310                 Max.   :132.000
##                     (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_ANO      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
```

##

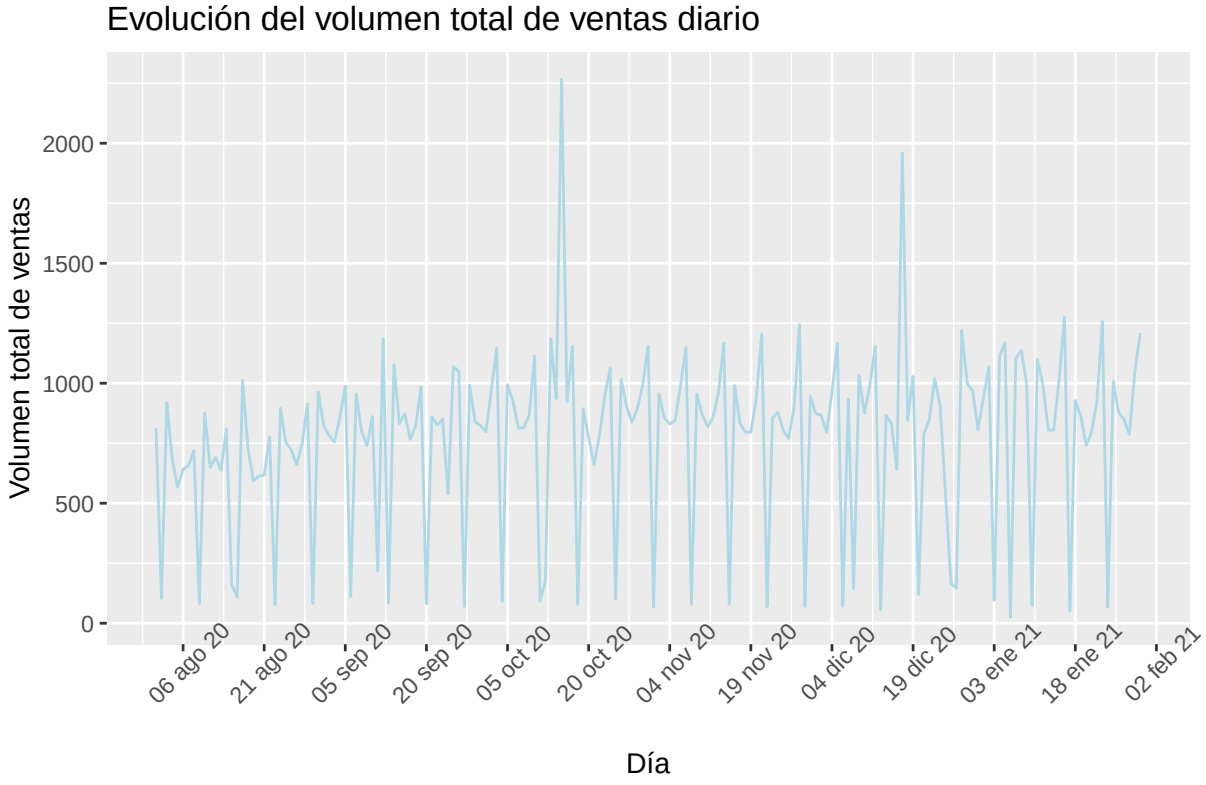
Observamos lo siguiente:

- La cantidad media de unidades vendidas es de 2.55, sin embargo, el número máximo de items comprado es de 132. Además, el 75 % de los usuarios ha comprado 6 unidades de producto o menos.
- El precio medio (con impuestos) de venta es de 1.44€, siendo el importe medio de venta de 3.68€, aunque encontramos que el importe máximo de venta es de 196.680€

Las 15 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:



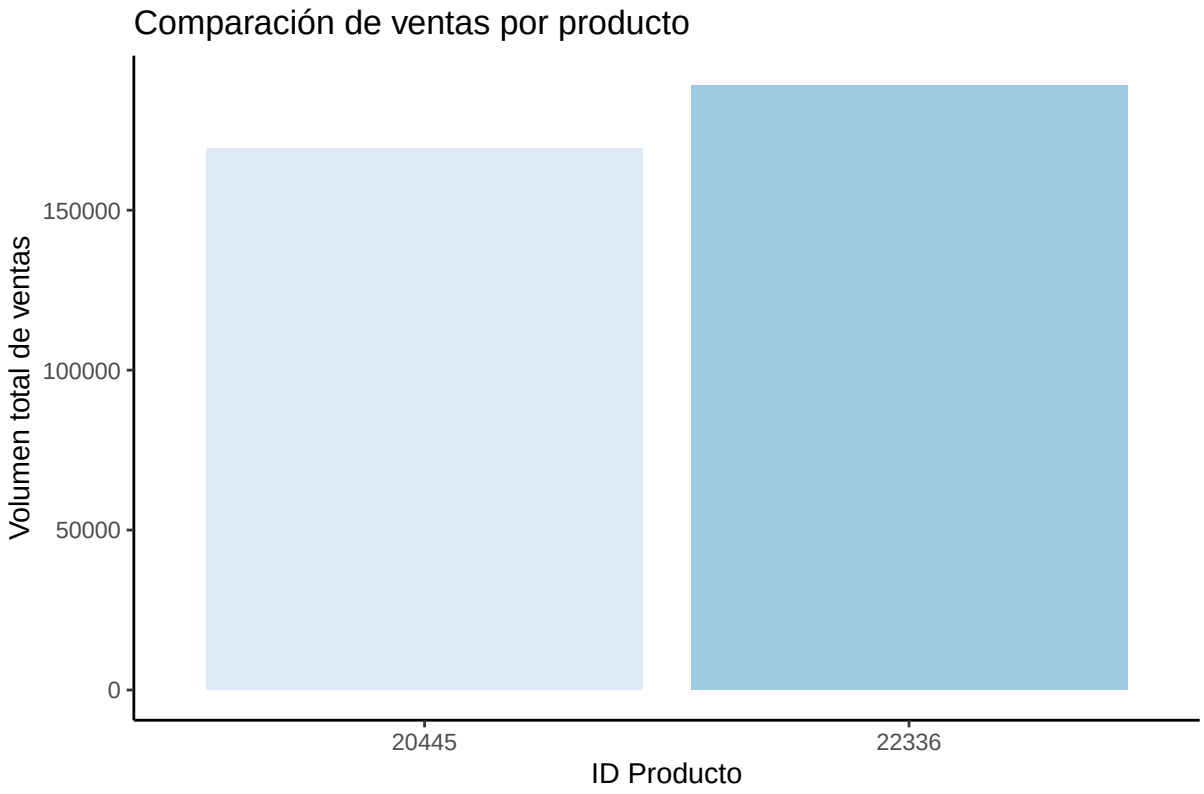
Fuente: Elaboración propia con datos

En el gráfico podemos apreciar como el volumen de ventas diario fluctúa bastante en función del día, encontrándolas en un rango entre 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.

A continuación, vamos a hacer una comparación del volumen de ventas total según cada producto:

Tabla 1: Importe medio de venta según producto

CODIGO	IMPORTE_MEDIO
20445	3.310346
22336	4.069384



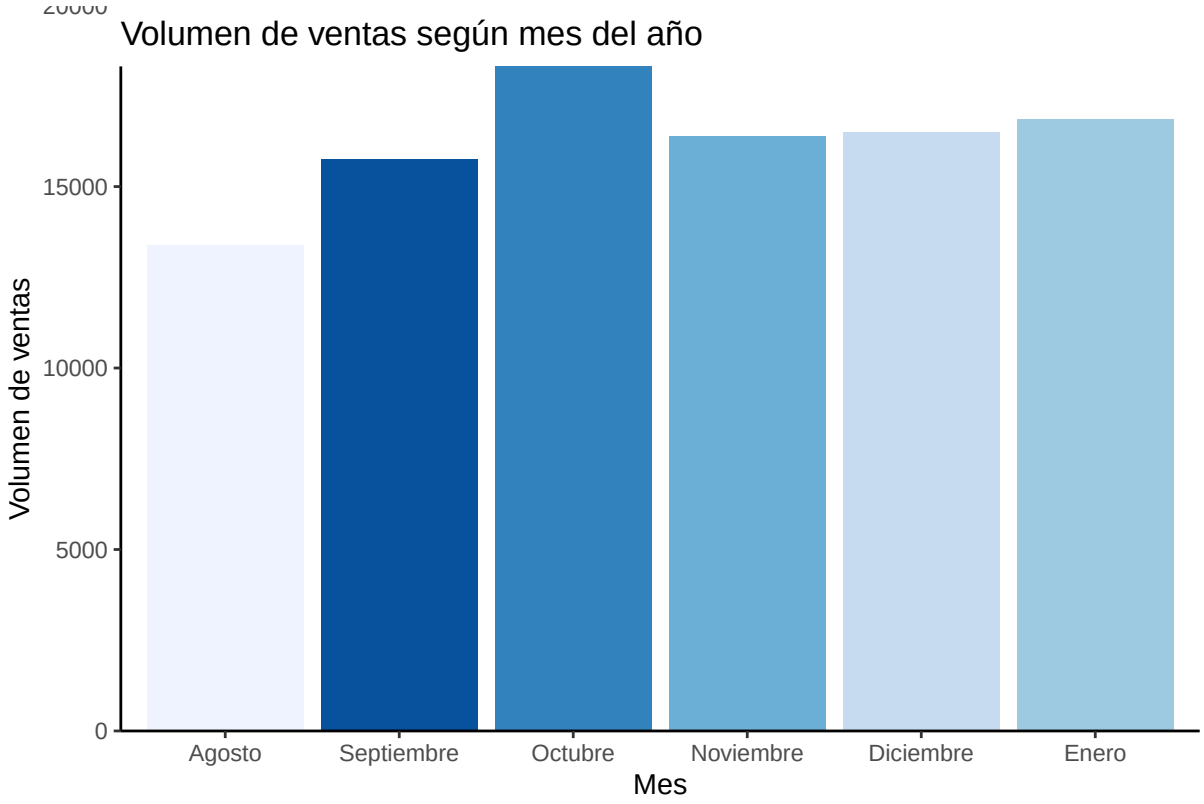
Fuente: Elaboración propia con datos de ventas

El volumen de ventas del producto 20445 ha sido ligeramente superior, con un volumen total de ventas de 169196 unidades, frente a las 188867 unidades vendidas del producto con identificador 22336. Estos datos han sido calculados teniendo en cuenta que se han podido vender más de una unidad en una misma transacción.

El importe del producto 20445 es de 1.49€ (precio con impuestos) y el producto 22336 tiene un precio de venta de 1.39€. Esto nos podría llevar a pensar que si se trata del mismo tipo de producto, los usuarios hayan optado por la opción más económica.

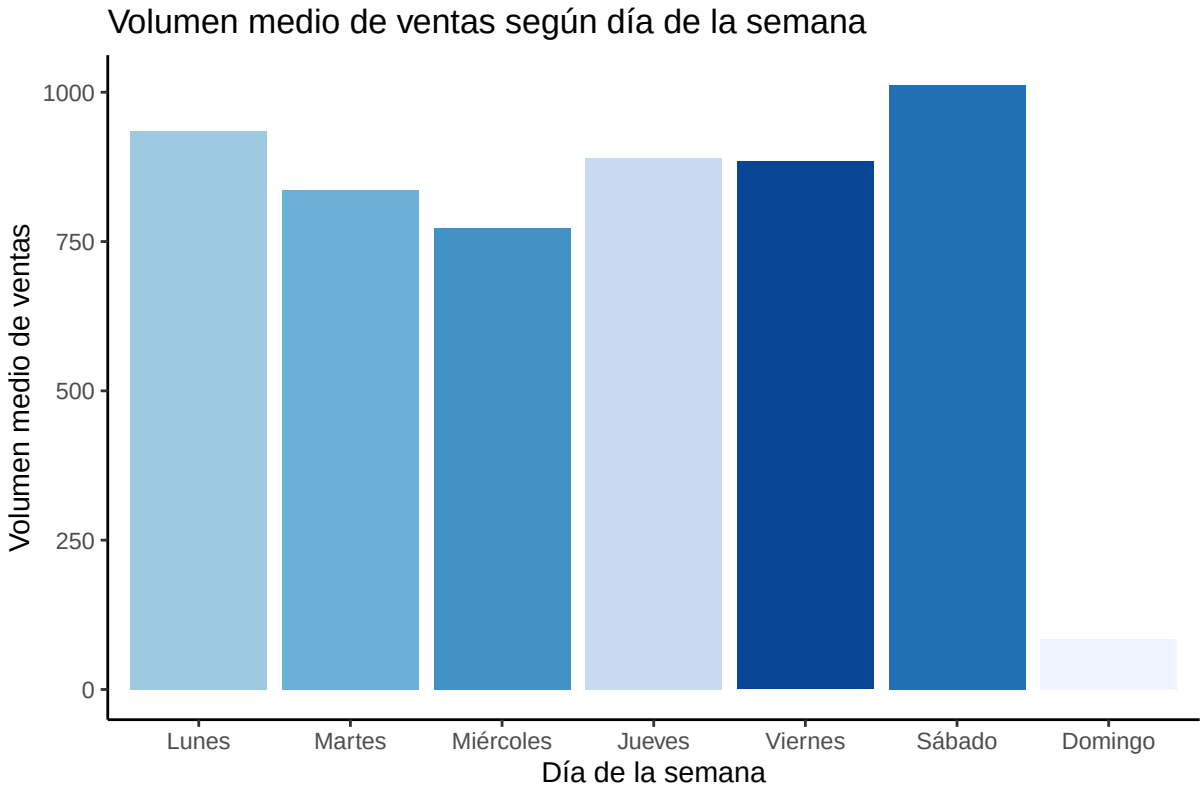
De nuevo, comprobamos que el importe medio de las ventas del producto 20445 tienen un importe medio de venta menor.

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

Podemos ver una tendencia creciente del número de ventas mensual, donde se tiende a consumir más cada mes. Sin embargo, encontramos que el mes en el que más ventas se han realizado ha sido en el mes de octubre, superando las 18 mil ventas.



Fuente: Elaboración propia con datos

Observando el gráfico, vemos que el Sábado es el día de la semana donde el número medio

de ventas es mayor, superando una media de mil ventas. Por el contrario, el Domingo es el día donde se registra un volumen medio de ventas diario considerablemente inferior.

0.1.3.2.1. Variable descuento

Si estudiamos la variable descuento, vemos que hay un descuento hasta en 1278 ventas, siendo 10.26 € el descuento medio aplicado.

Para el producto 20445, se ha aplicado descuento 638 veces frente a las 640 del producto 22336, por lo que el descuento se ha aplicado un número de veces similar.