Análisis de datos para La Redonda

Julio Vanegas

17 de abril de 2023

Análisis de datos para La Redonda

Paso 1

Primero hay que cargar las librerias.

```
library(tinytex) #Instalar y administrar fácilmente paquetes LaTeX en R.
library(dplyr) # Manipulación de datos
library(tidyverse) #Conjunto de paquetes de R diseñados para realizar análisis de datos
library(readxl) # readxl permite leer archivos de Microsoft Excel en R.
library(ggplot2) #paquete de visualización de datos
library(ggcorrplot) #permite visaulizar correlaciones
library(tools) #facilita la programación y el análisis de datos.
library(zoo) # herramientas para trabajar con series de tiempo y fechas.
library(lubridate) # herramientas para trabajar con fechas y horas en R.
library(stats) # paquete que permite hacer análisis estadisticos
library(scales) # controla la escala y los etiquetas en las visualizaciones de ggplot2.
library(openxlsx) #sirve para poder exportar las bases de datos
```

Ahora, hay que cargar la base de datos.

```
Data1 <- read_excel("~/Desktop/Informacion Ventas LaRedonda 20230324.xlsx", col_names = TRUE)
```

Paso 2

Hay que entender la estructura de nuestra base de datos.

```
str(Data1)
```

```
## tibble [24,682 x 13] (S3: tbl_df/tbl/data.frame)
##  $ Cliente           : chr [1:24682] "TIENDAS CHEDRAUI, SA DE CV" "TIENDAS CHEDRAUI, SA DE CV" "TIEN
##  $ FECHA             : POSIXct[1:24682], format: "2021-01-08" "2021-01-08" ...
##  $ CEDIS             : num [1:24682] 402 402 402 402 402 ...
##  $ DESCRIPCION       : chr [1:24682] "REDONDA BL RUBY 750 ML  12.5%" "LA REDONDA TINTO RUBY 750 ml 1
##  $ EAN               : num [1:24682] 7.5e+12 7.5e+12 7.5e+12 7.5e+12 7.5e+12 ...
##  $ CANTIDAD PEDIDA   : num [1:24682] 12 492 24 108 396 96 84 6 348 12 ...
##  $ CANTIDAD FACTURADA: num [1:24682] 12 492 24 108 396 96 84 6 348 12 ...
##  $ PRECIO UNITARIO   : num [1:24682] 81.8 81.8 81.8 92.1 81.8 ...
##  $ Importe           : num [1:24682] 981 40236 1963 9951 32385 ...
##  $ tipo de cliente   : chr [1:24682] "Grandes Superficies" "Grandes Superficies" "Grandes Superficie
##  $ Canal             : chr [1:24682] "Autoservicio" "Autoservicio" "Autoservicio" "Autoservicio" ...
##  $ Origen            : chr [1:24682] "NACIONAL" "NACIONAL" "NACIONAL" "NACIONAL" ...
```

```
##  $ Linea             : chr [1:24682] "La Redonda" "La Redonda" "La Redonda" "La Redonda" ...
```

Tenemos una base de datos de 24,682 observaciones por 13 variables. Entre las variables, hay 6 variables númericas, 6 categoricas y 1 variables que incluye fechas.

Hay que ver los primeros datos de la tabla.

```
head(Data1)
```

```
## # A tibble: 6 x 13
##   Cliente        FECHA               CEDIS DESCRIPCION     EAN `CANTIDAD PEDIDA`
##   <chr>          <dttm>              <dbl> <chr>         <dbl>             <dbl>
## 1 TIENDAS CHEDR~ 2021-01-08 00:00:00   402 REDONDA BL~ 7.50e12                12
## 2 TIENDAS CHEDR~ 2021-01-08 00:00:00   402 LA REDONDA~ 7.50e12               492
## 3 TIENDAS CHEDR~ 2021-01-08 00:00:00   402 REDONDA BL~ 7.50e12                24
## 4 TIENDAS CHEDR~ 2021-01-08 00:00:00   402 REDONDA TI~ 7.50e12               108
## 5 TIENDAS CHEDR~ 2021-01-08 00:00:00   402 REDONDA TI~ 7.50e12               396
## 6 TIENDAS CHEDR~ 2021-01-08 00:00:00   402 REDONDA TI~ 7.50e12                96
## # i 7 more variables: `CANTIDAD FACTURADA` <dbl>, `PRECIO UNITARIO` <dbl>,
## #   Importe <dbl>, `tipo de cliente` <chr>, Canal <chr>, Origen <chr>,
## #   Linea <chr>
```

Podemos observar que es una base de datos que almacena la información de la venta de productos alcoholicos. Es posible ver los clientes, la fecha en que hicieron sus pedidos, los productos, entre otras variables.

Paso 3

Para que nuestro programa de analisis pueda comprender esta información, es necesario homogeneizar los datos. Para esto, vamos a convertirlos todos a valores númericos.

```
col_names <- c("tipo de cliente", "Canal", "Origen", "Linea", "DESCRIPCION", "Cliente")
Data1[col_names] <- lapply(Data1[col_names], as.factor)
str(Data1)
```

```
## tibble [24,682 x 13] (S3: tbl_df/tbl/data.frame)
##  $ Cliente           : Factor w/ 300 levels "\" GRUPO ALFAJORES \", S DE RL DE CV",..: 267 267 267 2
##  $ FECHA             : POSIXct[1:24682], format: "2021-01-08" "2021-01-08" ...
##  $ CEDIS             : num [1:24682] 402 402 402 402 402 ...
##  $ DESCRIPCION       : Factor w/ 127 levels "A Sierra TI Cab Mer Ma 750 ml 13.5%",..: 76 44 79 92 11
##  $ EAN               : num [1:24682] 7.5e+12 7.5e+12 7.5e+12 7.5e+12 7.5e+12 ...
##  $ CANTIDAD PEDIDA   : num [1:24682] 12 492 24 108 396 96 84 6 348 12 ...
##  $ CANTIDAD FACTURADA: num [1:24682] 12 492 24 108 396 96 84 6 348 12 ...
##  $ PRECIO UNITARIO   : num [1:24682] 81.8 81.8 81.8 92.1 81.8 ...
##  $ Importe           : num [1:24682] 981 40236 1963 9951 32385 ...
##  $ tipo de cliente   : Factor w/ 9 levels "Cliente General",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Canal             : Factor w/ 6 levels "-","Autoservicio",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Origen            : Factor w/ 4 levels "Importado","IMPORTADO",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Linea             : Factor w/ 22 levels "Altiplano","Angove",..: 8 8 8 8 8 8 8 12 8 8 ...
```

Al tener las variables categoricas como factores, es más facil manipularlas. Podemos observar que hay algunos problemas en las variables.

```
levels(Data1$Canal)
```

```
## [1] "-"              "Autoservicio"    "club de precios" "conveniencia"
```

```
## [5] "Departamental"   "especializadas"
```

```
levels(Data1$Linea)
```

```
##  [1] "Altiplano"            "Angove"               "Aromo"
##  [4] "Borgo"                "Cyato"                "Fernandez de Pierola"
##  [7] "La redonda"           "La Redonda"           "labraz"
## [10] "mercadotecnia"        "Mueble"               "Orlandi"
## [13] "orlandi ti"           "Peñamayor"            "Portorojo"
## [16] "PortoRojo"            "Potorojo"             "Ruby"
## [19] "Sierra"               "sierra gorda"         "sierra luna"
## [22] "Traslascuestas"
```

```
levels(Data1$`tipo de cliente`)
```

```
## [1] "Cliente General"         "Distribuidor"
## [3] "Grandes Superficies"     "HORECAS"
## [5] "Mayorista"               "Otros"
## [7] "Tienda Especializada"    "Tiendas de Conveniencia"
## [9] "Vilared"
```

```
levels(Data1$Origen)
```

```
## [1] "Importado" "IMPORTADO" "Nacional"  "NACIONAL"
```

```
levels(Data1$Cliente)[1:20]
```

```
##  [1] "\" GRUPO ALFAJORES \", S DE RL DE CV"
##  [2] "3 CARBONES, S A DE C V"
##  [3] "ABARROTES EL DUERO, S.A. DE C.V."
##  [4] "ABARROTES VINOS Y LICORES LA BARATA, SA DE CV"
##  [5] "ACMEPARK, S.A. DE C.V."
##  [6] "ACSAPACK, SA DE CV"
##  [7] "ADMINISTRACION DE EMPRESAS AL MENUDEO, SA DE CV"
##  [8] "ADMINISTRADORA DE HOTELES GRT,"
##  [9] "ALEJANDRO HURTADO PADILLA"
## [10] "ALEJANDRO MORERA SILVA"
## [11] "ALELLA GRUP, SA DE CV"
## [12] "ALEXANDER ESTRADA ROSALES"
## [13] "ALICIA  PUERTAS PEREZ"
## [14] "ALIMENTICAZO, S.A. DE C.V."
## [15] "ALLENDE SUPPLY, SA DE CV"
## [16] "ALMA LUCIA LOPEZ CASTILLO"
## [17] "ALR CENTRO DEL DF,"
## [18] "ANA ALICIA MONTES VELAZQUEZ"
## [19] "ANA AUDREY MONTAÑO RICO"
## [20] "ANDRE PASCAL"
```

Los niveles te permiten observar problemas con la información. Por ejemplo, en Linea, vemos que hay tres niveles llamados “Portorojo”, “PortoRojo” y “Potorojo”. Es decir, que la base de datos identifica esta información como diferente cuando realmente es la misma. Esto suele ser error de dedo, un error humano.

Para la información de Clientes, hemos mantenido los nombres de los primeros 20 clientes para dar una idea, sin embargo en total hay 300.

ambien podemos observar este problema en Origen, donde hay un problema con la redacción; los niveles “Importado” e “IMPORTADO” se reconocen como dos niveles distintos. Lo mismo sucede con Nacional.

Para corregir esto, hay que homogeneizar la información.

```
levels(Data1$Linea) <- tolower(levels(Data1$Linea))
levels(Data1$Linea) <- gsub("potorojo", "portorojo", levels(Data1$Linea))
levels(Data1$Linea) <- toTitleCase(levels(Data1$Linea))

levels(Data1$Origen) <- tolower(levels(Data1$Origen))
levels(Data1$Origen) <- toTitleCase(levels(Data1$Origen))

levels(Data1$Linea)
```

```
##  [1] "Altiplano"            "Angove"               "Aromo"
##  [4] "Borgo"                "Cyato"                "Fernandez De Pierola"
##  [7] "La Redonda"           "Labraz"               "Mercadotecnia"
## [10] "Mueble"               "Orlandi"              "Orlandi Ti"
## [13] "Peñamayor"            "Portorojo"            "Ruby"
## [16] "Sierra"               "Sierra Gorda"         "Sierra Luna"
## [19] "Traslascuestas"
```

```
levels(Data1$Origen)
```

```
## [1] "Importado" "Nacional"
```

Podemos ver que esta empresa no solamente vende a otras empresas, sino que tambien a personas individuales. Tienen diferentes lineas de productos alcoholicos, tanto nacionales como importados, así como diferentes canales de venta.

Paso 4

Hay que revisar si hay información duplicada en la base de datos.

En la base original, hay 24682 observaciones, donde al eliminar los duplicados, nos queda una base con 24006 observaciones. Es decir, que hubo un total de 676 datos duplicados. Es importante eliminarlos ya que el tener valores duplicados pueden generar un sesgo en el analisis de la información.

```
Data2 <- distinct(Data1)
```

Habría que revisar si hay valores faltantes (NA) en nuestra información.

```
sum(is.na(Data2))
```

```
## [1] 48
```

Para entender en donde están distribuidos los valores nulos, habrá que ver en que columnas están distribuidos.

```
sapply(Data2, function(x) sum(is.na(x)))
```

```
##            Cliente              FECHA              CEDIS        DESCRIPCION
##                  0                  0                  0                  0
##                EAN    CANTIDAD PEDIDA CANTIDAD FACTURADA    PRECIO UNITARIO
##                 45                  0                  0                  0
##            Importe    tipo de cliente              Canal             Origen
##                  0                  0                  0                  3
##              Linea
##                  0
```

En las columnas de EAN hay 45 valores nulos y en Origen hay 3. Hay que entender la distribución de ambas columnas para decidir como tratarlos.

En el caso de EAN, podemos ver que es una especie de codigo de barras para marcar los productos. Realmente no resulta util para el analisis, por lo que hemos decidido eliminarlo.

```
Data3 <- Data2 %>% select(-EAN)
sapply(Data3, function(x) sum(is.na(x)))
```

```
##            Cliente              FECHA              CEDIS        DESCRIPCION
##                  0                  0                  0                  0
##    CANTIDAD PEDIDA CANTIDAD FACTURADA    PRECIO UNITARIO            Importe
##                  0                  0                  0                  0
##    tipo de cliente              Canal             Origen              Linea
##                  0                  0                  3                  0
```

Ahora solo nos queda tratar la información de Origen. Esta variable se distribuye en dos opciones categoricas, "Nacional" o "Importado". Al ser tres casos con NA, podemos analizarlos uno por uno.

```
Indices_NA <- which(is.na(Data3$Origen))
Indices_NA
```

```
## [1] 22725 23427 23928
```

Los indices de los productos donde no se especifica su origen son "22725", "23427"" y "23928".

```
Info_Data3 <- Data3 %>%
  slice(c(22725, 23427, 23928)) %>%
  select(DESCRIPCION,Importe)
Info_Data3
```

```
## # A tibble: 3 x 2
##   DESCRIPCION                   Importe
##   <fct>                           <dbl>
## 1 COPA IMPRESA PARA DEGUSTACION    0.36
## 2 COPA IMPRESA PARA DEGUSTACION    0.72
## 3 COPA IMPRESA PARA DEGUSTACION    0.36
```

Los indices de los productos donde no se especifica su origen son todos "COPA IMPRESA PARA DEGUSTACION". Buscando en la pagina de La Redonda, no hay información sobre venta de copas de degustación, por lo que no podemos saber con certeza del origen de estos productos. Para lidiar con esto, asumiremos que estos 3 productos son de origen nacional, ya que si fueran importados tendrían un costo mayor y resultaría poco probable que omitieran información en un producto de otro país.

```
Data3$Origen[is.na(Data3$Origen)] <- "Nacional"
sapply(Data3, function(x) sum(is.na(x)))
```

```
##            Cliente              FECHA              CEDIS        DESCRIPCION
##                  0                  0                  0                  0
##    CANTIDAD PEDIDA CANTIDAD FACTURADA    PRECIO UNITARIO            Importe
##                  0                  0                  0                  0
##    tipo de cliente              Canal             Origen              Linea
##                  0                  0                  0                  0
```

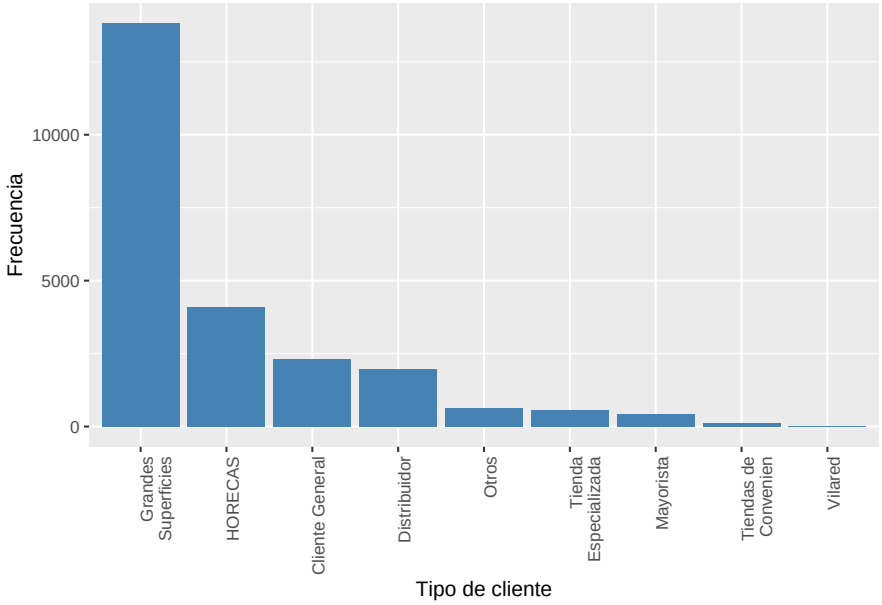
De esta forma, ya no tenemos valores nulos en nuestro dataframe.

Al haber ya eliminado los valores nulos, duplicados y limpiado la información restante en esta base de datos, podemos empezar a analizar la información.

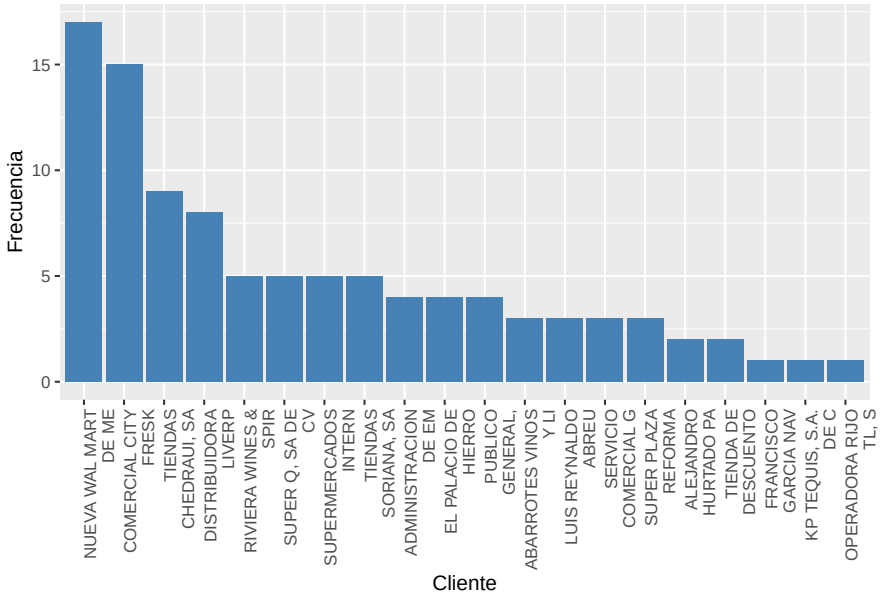
Paso 5

Análisis Univariado

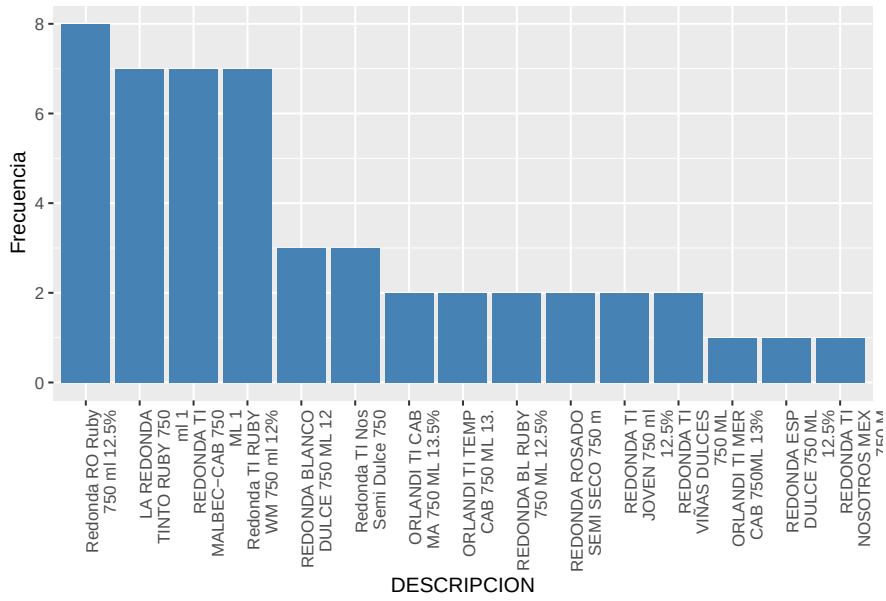
Análisis Categorico



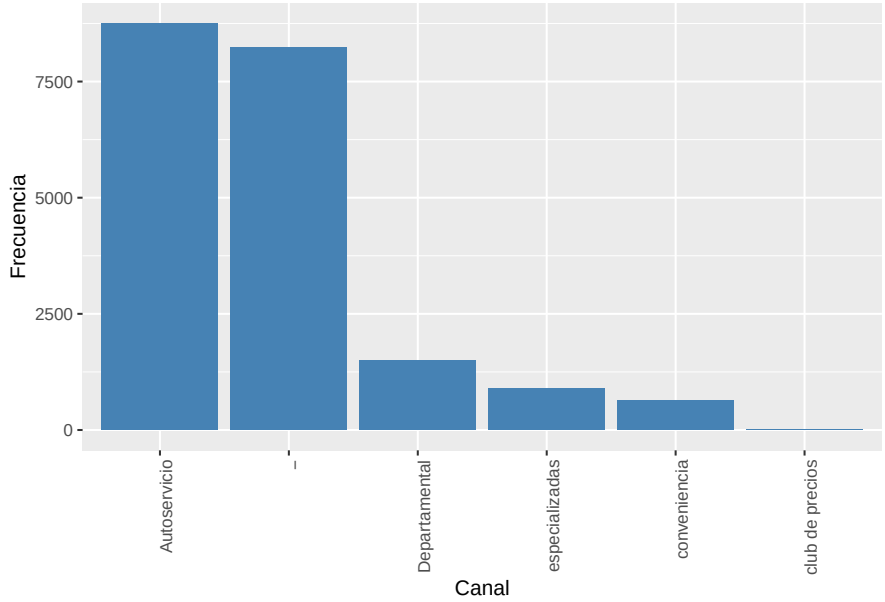
Podemos ver que el tipo de cliente que más se repite es el de "Grandes Superficies". La redonda debe de guardar especial atención a los clientes que entrén en esta categoría. En los analisis bivariados del proximo paso ahondaremos más en esto.



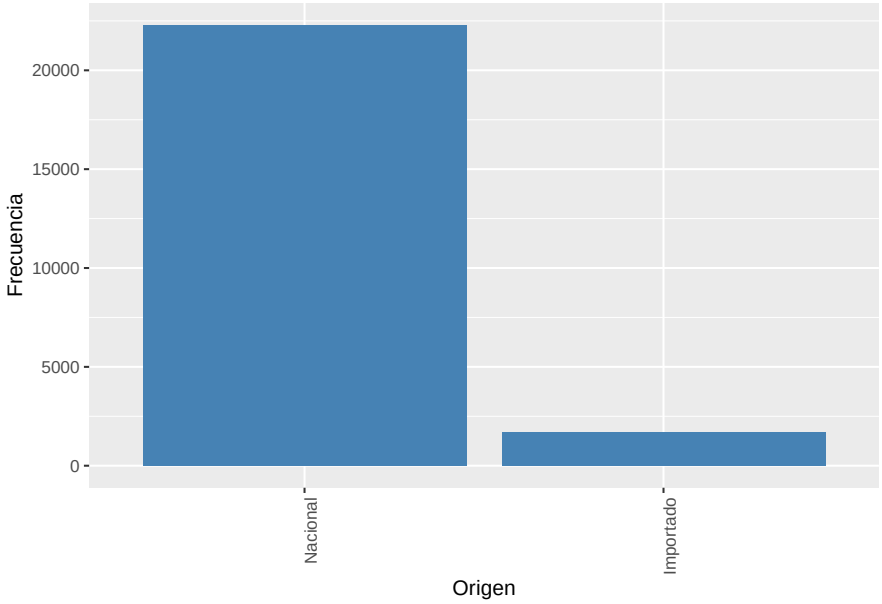
Los clientes que más aparecen son Walmart y City Fresko. Le siguen Chedraui y Liverpool, sin embargo ya no consumen tanto como los primeros dos.



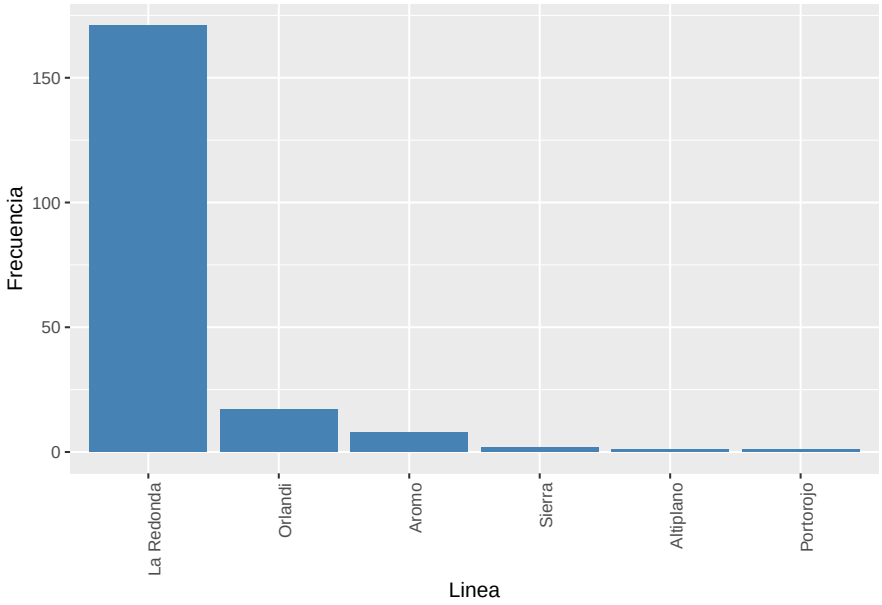
Los productos que más venden son “Redonda RO RUBY”, “LA REDONDA TINTO RUBY”, “REDONDA TI MALBEC-CAB” y “REDONDA TI RUBY”. Podemos ver que todos son Vinos, y tiene sentido al saber que son un viñedo. Se podría invertir en promoción en estos productos para aumentar aún más sus ventas.



Los canales que más aparecen son “Autoservicio” y “-”. Podemos ver que tenemos una variable que no está definida, ya que abarca todas las ventas que hace directamente La Redonda. Puede ser desde su pagina web o desde el viñedo.



Por mucho, los tipos de productos más vendidos son aquellos de origen Nacional. Se podrían considerar las utilidades de los productos importados para ver si realmente generan buena utilidad o solamente son un gasto innecesario.

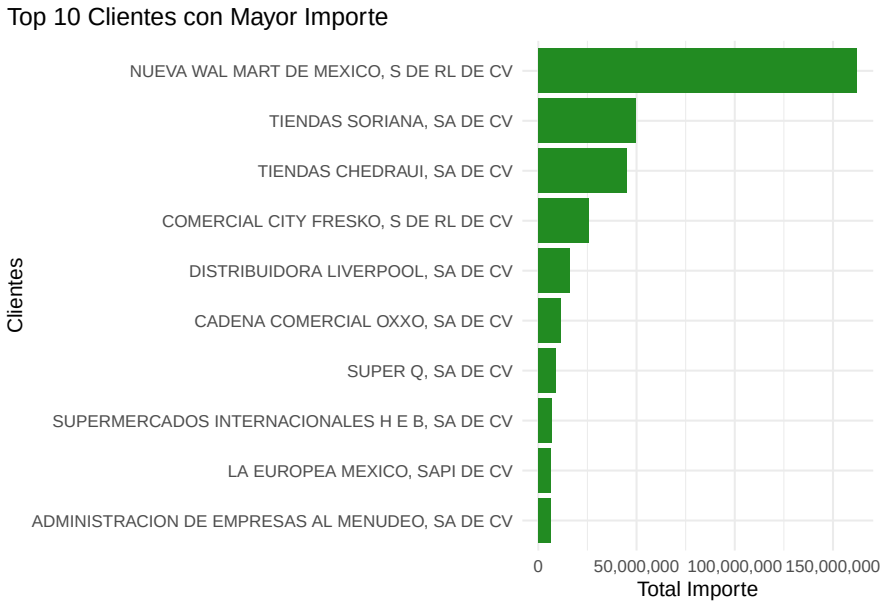


La linea de producto más vendida es La Redonda, la marca personal. Esto hace sentido al analizar el origen ya que este viñedo es mexicano.

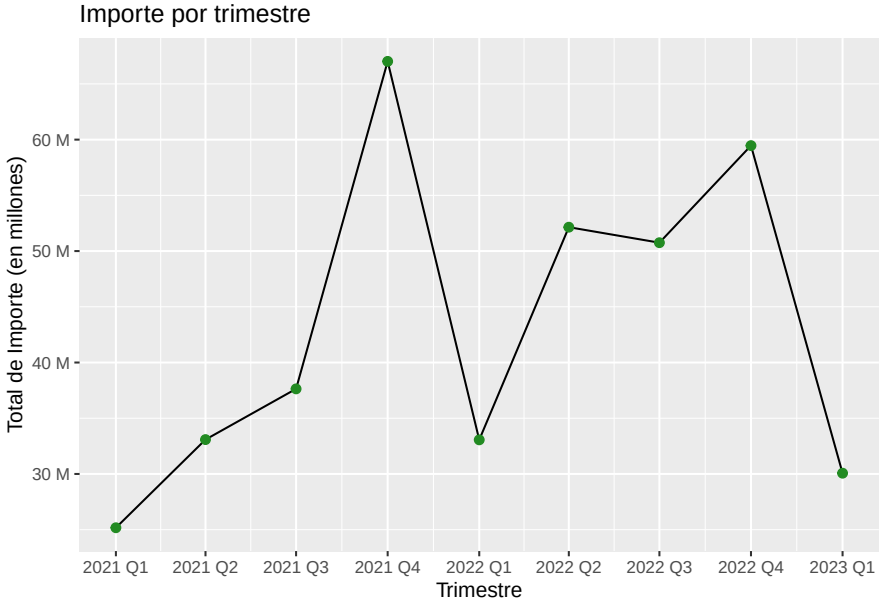
Paso 6

Analisis Bivariado

Para entender a mayor profunidad la relación entre diferentes variables, es necesario relacionarlas para ver su comportamiento.



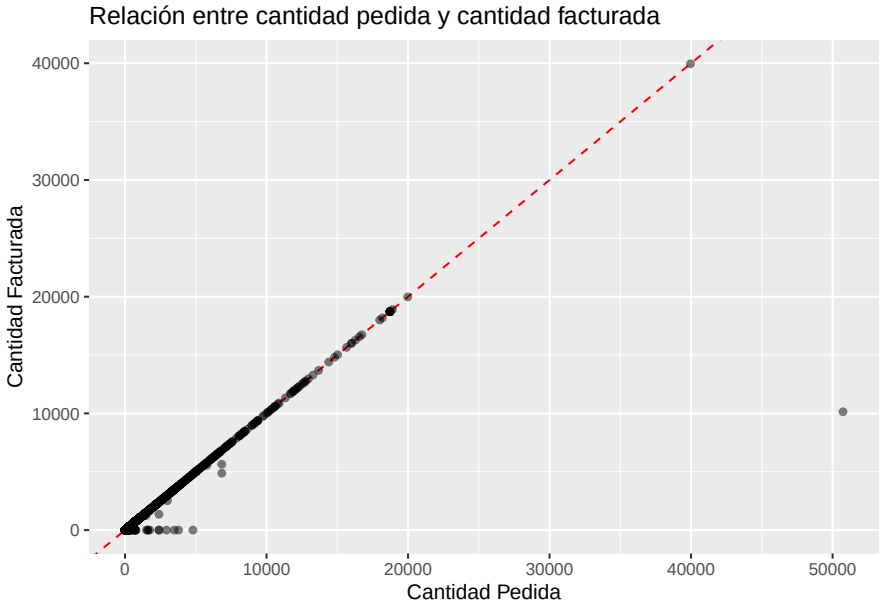
Podemos ver que el cliente que más importe tiene es Walmart. A su vez, no se encuentra "CLIENTE INTERNO PARA REQUERIMIENTOS DE PRODUCCION", que es el que más pedidos tiene. Esto quiere decir que no están facturando sus pedidos, y confirma que es una cuenta de control interna. Habrá que ponerla en otra base de datos, para no perder la información y disminuir la desviación estandar en las demás variables.



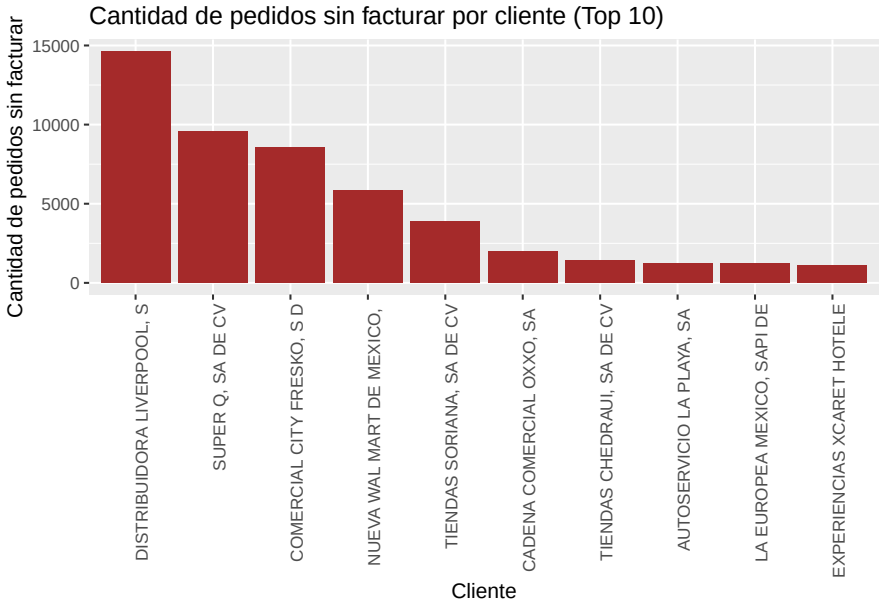
Podemos ver esta grafica que muestra los trimestres con más importes en los ultimos tres años. Podemos observar que, en estos ulimos años, suele haber un crecimiento considerable del tercer trimestre al cuarto. Esto puede deberse por la celebración de diferentes festividades, como lo es Navidad o el Día de Muertos. Los trimestres con menos ventas siempre son los primeros, pero de ahí suelen ir en aumento, a diferencia del 2022 Q3 que disminuyo un poco, sin embargo no fue una caida considerable.

Es importante tomar en cuenta esta información ya que nos permite entender en que fechas son las que se venden más productos, y su analisis permitirá generar estrategias para capitalizar en los patrones de consumo.

A continuación, se hará un analisis sobre las cantidades pedidas y las cantidades facturadas. Para evitar sesgos, se omitarán las cantidades pedidas de "CLIENTE INTERNO PARA REQUERIMIENTOS DE PRODUCCION", ya que son de uso interno.

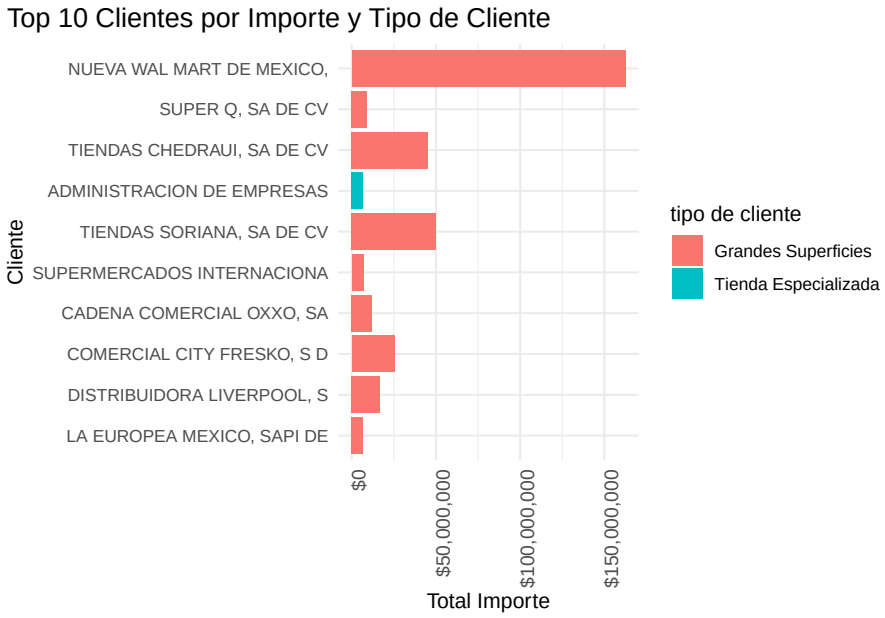


Esta es una grafica que muestra la relación entre las cantidad pedida y la cantidad facturada. Si hubiera una relación ideal, todos los valores (puntos) estarían sobre la linea roja. Sin embargo, podemos ver que hay valores fuera de esta linea, significando que hay pedidos que no logran facturarse. Esto hace que no haya un importe y que se generé, en esencia, un costo de oportunidad sobre posibles pedidos. Podemos ver que hay hasta 50,000 pedidos que no fueron facturados. Se pueden hacer una estimación fundamentada acerca de esto; los pedidos cancelados, problemas de inventario o errores en el proceso de facturación. Es necesario atender esto, ya que esos 50,000 pedidos no facturados pudieron haber significado un gran ingreso para el viñedo.

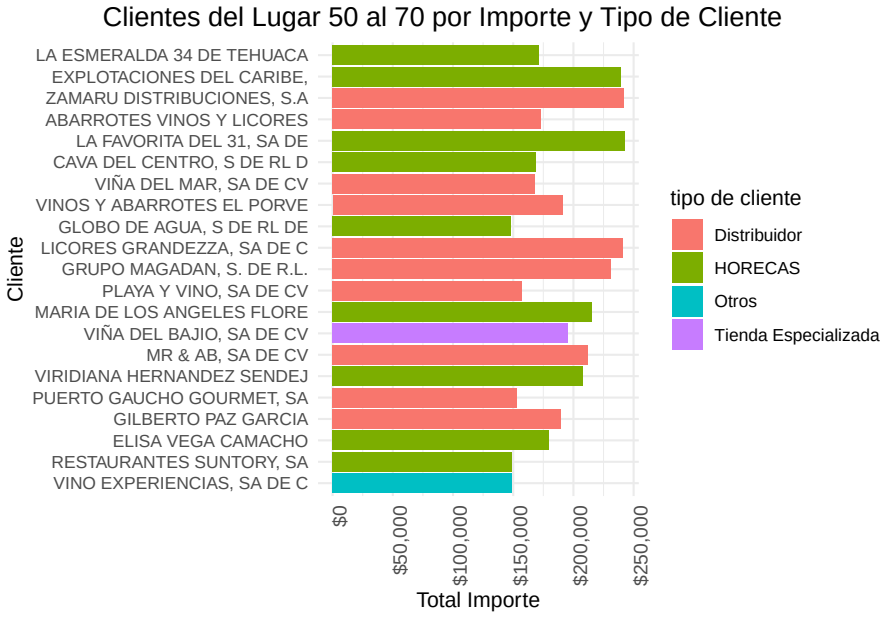


Es necesario verque está pasando con Liverpool, ya que es el cliente que más pedidos tiene que no lograron ser facturados. Le sigue Super Q y City Fresko. Es con estos clientes donde está el mayor costo de oportunidad

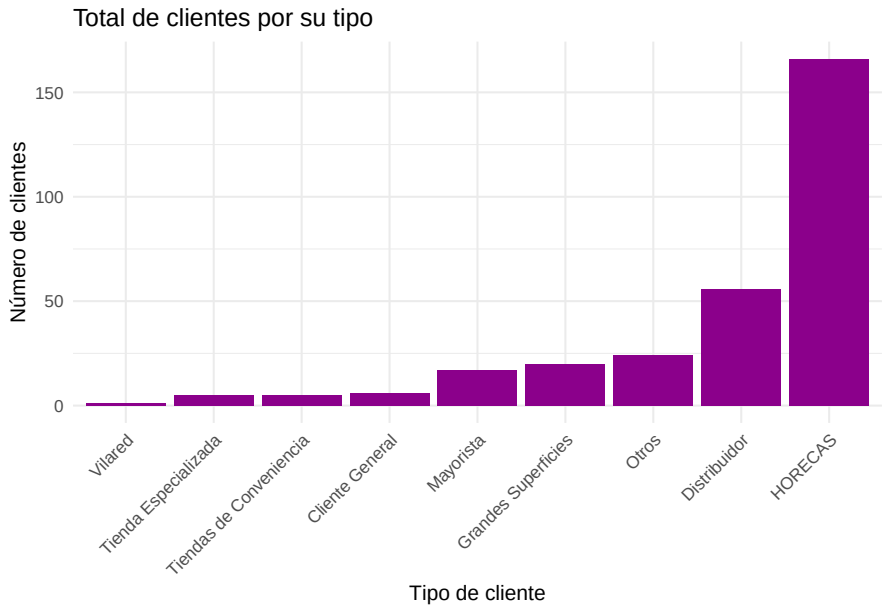
en los ingresos del viñedo. Estamos hablando de cantidades considerables, ya que asumiendo que Liverpool haya tenido 14,000 pedidos de un vino de 100 pesos, serían un millón cuatrocientos pesos de ganancias que no habrían sido procesados.



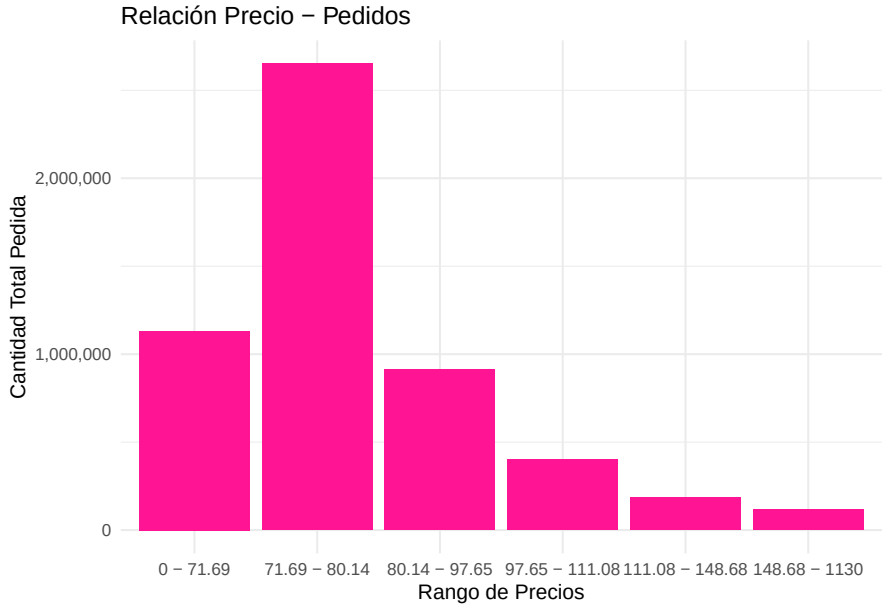
Los primeros diez tipos de clientes que más importe generan son de "Grandes Superficies". Es necesario tener en cuenta esto ya que ellos son los que generan más ingresos a la marca.



Sin embargo, podemos observar que empiezan a aparecer más tipos de clientes conformen vamos avanzando en la lista. El tipo de cliente HORECAS así como Distribuidor empiezan a estar más presentes. Así que, si bien los que más importe generan son las Grandes Superficies, no hay que dejar de lado los otros tipos de clientes.



De hecho, el tipo de cliente más común para La Redonda es HORECAS (Hoteles, Restaurantes y Cafeterias) y los distribuidores. Aunque no sean los que más importe generen, son los que más presencia tienen en la empresa. Un ejemplo de como capitalizar con este tipo de clientes son los eventos y degustaciones. El Organizar eventos y degustaciones en los establecimientos para que sus clientes finales puedan conocer y disfrutar de los productos vinos. Esto también puede generar publicidad y relaciones públicas positivas para la marca.



Podemos observar que los vinos que más se piden son los que se encuentran entre el rango de 71.69 a 80.14 pesos en su precio unitario. Esto quiere decir que la gente suele decantarse por vinos más baratos, entre los rangos de 0-98 pesos. Si bien la desviación estandar del precio unitario es alta (desvia 71.62 pesos del promedio), esto se debe a que si bien no hay tantos pedidos de valores más altos, estos siguen afectando el calculo.

Paso 7

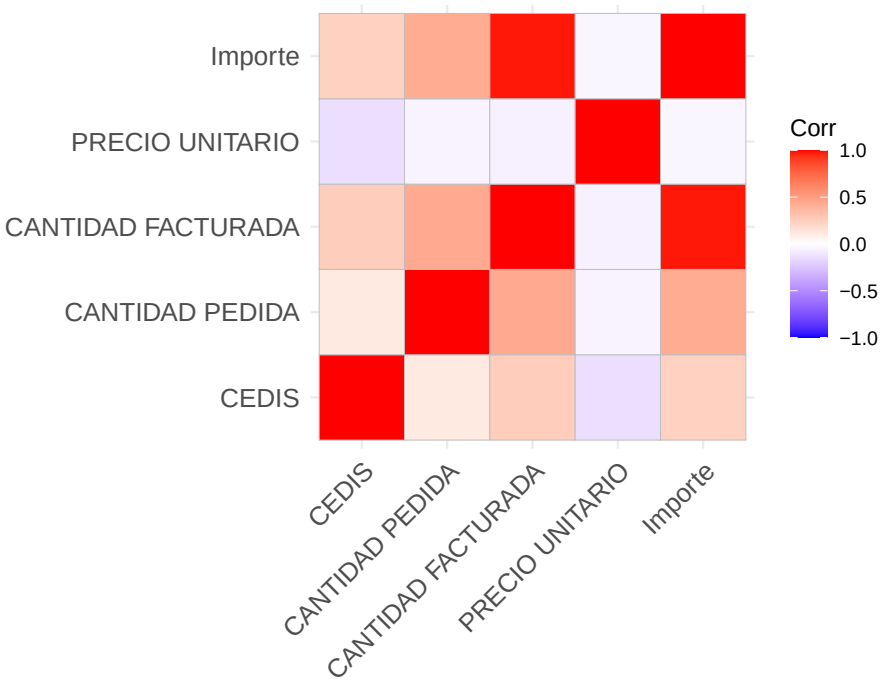
Analisis correlacional.

El análisis correlacional es una técnica estadística que se utiliza para medir la relación entre dos o más variables. El objetivo del análisis correlacional es determinar si existe una relación entre las variables y, de ser así, medir la fuerza y la dirección de esa relación. A continuación, vamos a buscar entender la correlación entre las variables de la base de datos que hemos estado analizando.

El objetivo es entender más allá de la simple existencia de una relación y su fuerza, sino también cómo estas variables se relacionan entre sí y qué información se puede extraer de estas relaciones. Al realizar un análisis correlacional, es importante tener en cuenta que la correlación no implica necesariamente causalidad. Es decir, solo porque dos variables están correlacionadas entre sí, no significa que una variable cause la otra. Por lo tanto, es necesario tener en cuenta otras variables y factores que podrían influir en la relación entre las variables. ### Se observan las siguientes correlaciones entre las variables:

```
##                    CANTIDAD PEDIDA CANTIDAD FACTURADA PRECIO UNITARIO
## CANTIDAD PEDIDA         1.00000000         0.43660585     -0.04690244
## CANTIDAD FACTURADA      0.43660585         1.00000000     -0.06465433
## PRECIO UNITARIO        -0.04690244        -0.06465433      1.00000000
## Importe                 0.42858134         0.98469104     -0.04415081
##                        Importe
## CANTIDAD PEDIDA     0.42858134
## CANTIDAD FACTURADA  0.98469104
## PRECIO UNITARIO    -0.04415081
## Importe             1.00000000
```

```
CorRedonda <- round(cor(Data4),2)
ggcorrplot(CorRedonda)
```

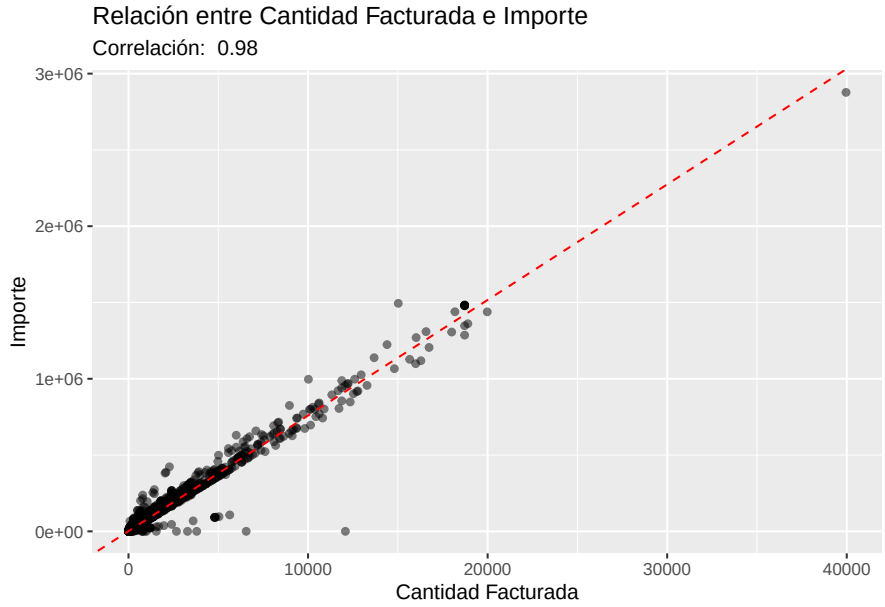


Correlación Facturas e Importe

```
FacturasEImporte <- lm(Importe ~ `CANTIDAD FACTURADA`, data = Data4)
pendiente <- coef(FacturasEImporte)[2]
```

```r
intercepto <- coef(FacturasEImporte)[1]

ggplot(Data4, aes(x = `CANTIDAD FACTURADA`, y = Importe)) +
  geom_point(alpha = 0.5) +
  geom_abline(intercept = intercepto, slope = pendiente, linetype = "dashed", color = "red") +
  labs(x = "Cantidad Facturada", y = "Importe",
       title = "Relación entre Cantidad Facturada e Importe",
       subtitle = paste("Correlación: ", round(cor(Data4$`CANTIDAD FACTURADA`, Data4$Importe), 2)))
```



- Hay una fuerte correlación positiva (0,984) entre "Cantidad Facturada" e "Importe", lo que tiene sentido ya que la cantidad facturada y el monto deben estar directamente relacionados. Podemos ver que no hay una relación perfecta ya que hay valores fuera de la linea roja, que representa la relación ideal. Por ejemplo, si una factura registra una cantidad facturada de 1000 unidades pero un importe de sólo 100 pesos, o una cantidad facturada de sólo 10 unidades pero un importe de 1000 dólares, sería un valor atípico en el gráfico de dispersión.

```r
modeloResiduosFacturaEImporte <- lm(Importe ~ `CANTIDAD FACTURADA`, data = Data4)
ResiduosFacturaEImporte <- resid(modeloResiduosFacturaEImporte)
LimSFactEImp <- mean(ResiduosFacturaEImporte) + 2 * sd(ResiduosFacturaEImporte)
LimIFacteIMP <- mean(ResiduosFacturaEImporte) - 2 * sd(ResiduosFacturaEImporte)

IndDesvFactEImp <- which(ResiduosFacturaEImporte > LimSFactEImp | ResiduosFacturaEImporte < LimIFacteIM
DataIndDesvFactEImp <- Data4[IndDesvFactEImp, ]
AtipicFactEImp<- subset(DataIndDesvFactEImp, select = c(`CANTIDAD FACTURADA`, Importe, `PRECIO UNITARIO
AtipicFactEImp
```

```
## # A tibble: 377 x 3
##    `CANTIDAD FACTURADA` Importe `PRECIO UNITARIO`
##                   <dbl>   <dbl>             <dbl>
##  1                 1200 120216              100.
##  2                 6300 453600               72
##  3                 1800 180324              100.
##  4                 8640 622080               72
##  5                  720  13738.              19.1
##  6                 1680  32054.              19.1
```

```
##  7                1200 120216             100.
##  8                1800 189000             105
##  9                6300 453600              72
## 10                9084 659226.             72.6
## # i 367 more rows
```

No es que haya un error en el calculo del importe, sino que, por ejemplo, la segunda fila tiene una Cantidad Facturada de 6300, pero un Importe de solo 453600, lo que sugiere un precio unitario muy bajo en comparación con otras observaciones con la misma cantidad facturada. Es por esto que la relación no es de 1, sin embargo solo hay 377 casos de esto.

Correlación Pedidos e Importe

En la correlación anterior (.49) podiamos observar una relación media entre los pedidos y el importe. Esto toma sentido al entender que no todos los pedidos son facturados, por tanto no generan un importe. Para tener un entendimiento más claro, es necesario hacer un analisis de aquellos pedidos que sí fueron facturados para entender la relación real ente estas variables.

```
PedidosFacturado <- Data4[Data4$`CANTIDAD FACTURADA` > 0, ]
cor(PedidosFacturado$`CANTIDAD PEDIDA`, PedidosFacturado$Importe)
```

```
## [1] 0.9465244
```

Podemos ver que la correlación es casi 1, lo cual indica que existe una relación fuerte entre ambas variables. La grafica sería casi igual que la vista anteriormente (Facturada o Importe)

```
PedidosFacturado <- Data4[Data4$`CANTIDAD FACTURADA` > 0, ]

modeloResiduosPedidoEImporte <- lm(Importe ~ `CANTIDAD PEDIDA`, data = PedidosFacturado)
ResiduosPedidoEImporte <- resid(modeloResiduosPedidoEImporte)

LimSPedidoEImp <- mean(ResiduosPedidoEImporte) + 2 * sd(ResiduosPedidoEImporte)
LimIPedidoEImp <- mean(ResiduosPedidoEImporte) - 2 * sd(ResiduosPedidoEImporte)

IndDesvPedidoEImp <- which(ResiduosPedidoEImporte > LimSPedidoEImp | ResiduosPedidoEImporte < LimIPedido

DataIndDesvPedidoEImp <- PedidosFacturado[IndDesvPedidoEImp, ]

AtipicPedidoEImp <- subset(DataIndDesvPedidoEImp, select = c(`CANTIDAD PEDIDA`, Importe, `PRECIO UNITARI
AtipicPedidoEImp
```

```
## # A tibble: 255 x 3
##    `CANTIDAD PEDIDA` Importe `PRECIO UNITARIO`
##                <dbl>   <dbl>             <dbl>
##  1              1800 180324              100.
##  2              1680  32054.              19.1
##  3              1800 189000              105
##  4              1200  22896               19.1
##  5              1200  22896               19.1
##  6              1200  22896               19.1
##  7              3480 322840.              92.8
##  8             10896 801837.              73.6
##  9              2400  45792               19.1
## 10              2928 271631.              92.8
## # i 245 more rows
```

Estos son los valores atipicos, y de igual forma podemos ver que se comportan de una manera similar que con

los pedidos facturados, ya que estás tres variables en realidad tienen una relación directa. Podemos ver que lo que hace que estos sean outliers es, en esencia, de nuevo el resultado del importe ya que, en promedio con otros pedidos de la misma cantidad, el importe o es muy bajo o es muy alto.

Sin embargo, este calculo no es real ya que a día de hoy siguen habiendo pedidos que no se facturan. Esto es algo que debería de disminuirse. En los pasos previos, se mencionan estrategias para tratar esta situación.

Correlación entre cantidad pedida y cantidad facturada

En el caso de la cantidad pedida y la cantidad facturada, la correlación inicial nos indicaba que no era una conexión muy fuerte (.43), pero esto se debe a que no todos los pedidos que se piden, se facturan. Hay un factor que tambien afecta, y es que el cliente de procesos internos de la redonda tiene pedidos muy grandes y donde no registran factura. Habrá que analizar la correlación sacando este factor.

```
FacturasSinPI <- subset(Data3, Cliente != "CLIENTE INTERNO PARA REQUERIMIENTOS DE PRODUCCION")
FacturasSinPI
```

```
## # A tibble: 23,962 x 12
##    Cliente               FECHA               CEDIS DESCRIPCION `CANTIDAD PEDIDA`
##    <fct>                 <dttm>              <dbl> <fct>                     <dbl>
##  1 TIENDAS CHEDRAUI, SA~ 2021-01-08 00:00:00   402 REDONDA BL~                12
##  2 TIENDAS CHEDRAUI, SA~ 2021-01-08 00:00:00   402 LA REDONDA~               492
##  3 TIENDAS CHEDRAUI, SA~ 2021-01-08 00:00:00   402 REDONDA BL~                24
##  4 TIENDAS CHEDRAUI, SA~ 2021-01-08 00:00:00   402 REDONDA TI~               108
##  5 TIENDAS CHEDRAUI, SA~ 2021-01-08 00:00:00   402 REDONDA TI~               396
##  6 TIENDAS CHEDRAUI, SA~ 2021-01-08 00:00:00   402 REDONDA TI~                96
##  7 TIENDAS CHEDRAUI, SA~ 2021-01-08 00:00:00   402 Redonda TI~                84
##  8 TIENDAS CHEDRAUI, SA~ 2021-01-08 00:00:00   402 ORLANDI TI~                 6
##  9 NUEVA WAL MART DE ME~ 2021-01-08 00:00:00  7464 Redonda TI~               348
## 10 NUEVA WAL MART DE ME~ 2021-01-08 00:00:00  7490 Redonda TI~                12
## # i 23,952 more rows
## # i 7 more variables: `CANTIDAD FACTURADA` <dbl>, `PRECIO UNITARIO` <dbl>,
## #   Importe <dbl>, `tipo de cliente` <fct>, Canal <fct>, Origen <fct>,
## #   Linea <fct>
```

```
cor(FacturasSinPI$`CANTIDAD PEDIDA`,FacturasSinPI$`CANTIDAD FACTURADA`)
```

```
## [1] 0.9618247
```

Al hacer esto, podemos observar que la correlación vuelve a ser alta (.96). La cuenta de producción interna aumentaba mucho la desviación, cosa que demostramos en el paso 5. Es por esto que la correlación no era propiamente alta. Entonces, si quitamos la cuenta de procesos internos, tenemos una correlación más apegada a la realidad.

Correlación Precio Unitario y Cantidad Pedida

La correlación es de -0.04690244, lo que indica una relación negativa muy débil. Esto sugiere que no hay una relación clara entre la cantidad pedida y el precio unitario del producto. Esto pasa ya que hay una gran variedad de productos con un gran rango de precios. Sin embargo, esto sucede porque se generaliza con todos los precios y todos los pedidos, ya que, si segmentaramos por rangos de precio, podríamos encontrar patrones de consumo. Por ejemplo, una relación entre el precio de los productos y las cantidades pedidas, como se vio en el analisis bivariado, demuestra que sí hay una clara decantación por productos más baratos. La correlación en este caso se ve afectada por que cada persona pide diferentes cantidades de vinos y esto sesga el ejercicio. Sin embargo, sí hay una relación.

Correlación entre el Precio Unitario y la Cantidad Facturada

Correlación de -0.06465433 , lo que indica una relación negativa muy débil. Esto sugiere que no hay una relación clara entre la cantidad facturada y el precio unitario del producto. Esto tambien puede no ser del todo cierto, ya que muchos pedidos que no se lograron facturar.

```
cor(PedidosFacturado[, -1])
```

```
##                     CANTIDAD PEDIDA CANTIDAD FACTURADA PRECIO UNITARIO
## CANTIDAD PEDIDA          1.00000000         0.96362277     -0.06388475
## CANTIDAD FACTURADA       0.96362277         1.00000000     -0.06648132
## PRECIO UNITARIO         -0.06388475        -0.06648132      1.00000000
## Importe                  0.94652438         0.98467170     -0.04553846
##                         Importe
## CANTIDAD PEDIDA      0.94652438
## CANTIDAD FACTURADA   0.98467170
## PRECIO UNITARIO     -0.04553846
## Importe              1.00000000
```

Aun cuando omitimos las cantidades que no fueron facturadas, vemos que no hay una relación aparente entre el precio unitario y la facturación. Es decir, que el precio de los productos no parece afectar la decisión de compra del consumidor. Sin embargo, esto está sesgado, ya que como demostramos en el analisis bivariado, sí hay una tendencia hacia los vinos más baratos.

Correlación entre el Precio Unitario e Importe

La correlación es de -0.04415081, lo que indica una relación negativa muy débil. Esto sugiere que no hay una relación clara entre el precio unitario y el importe. Esto está de igual forma sesgado, ya que sí bien el importe final va a estar determinado por las cantidades pedidas, tambien por el precio de venta. Sin embargo, como pueden haber diferentes ventas del mismo producto pero en diferentes cantidades, esto hace que no haya una relación aparante entre estas variables. A su vez, puede que entre más cantidades pedidas, en algunos pedidos se aplique un descuento, sesgando la relación.

Conclusión del análisis

Al haber limpiado la base de datos y entendido la estructura real de la información (el porqué de los valores anomalos, la distribución de los datos) es posible realizar diferentes propuestas acerca de cómo mejorar la situación actual de la redonda, aplicando las estrategias de negocios correspondientes. Pudimos ver que hay diferentes puntos a tratar, como la relación entre los pedidos y las facturas; hay un gran costo de oportunidad al no poder facturar todos los pedidos. A su vez, podemos crear diferentes propuestas de mercadotecnia para aprovechar las tendencias de venta en el ultimo trimestre del año, como para reforzar aquellos trimestres donde no se vende tanto (Q1). Es necesario realizar análisis constantes para entender a profundidad que es lo que se debe de hacer para capitalizar las oportunidades que tiene La Redonda.

Exportación de la base de datos limpia

```
#Data4 <- Data3
#Data5 <- cbind(Data4, Data2$EAN)
#write.xlsx(Data5, "~/Desktop/Mi Carpeta/RedondaConEAN.xlsx")
#write.xlsx(Data4, "~/Desktop/Mi Carpeta/RedondaSinEAN.xlsx")
```

Para exportar la data limpia (en este caso el dataframe lleva por nombre Data3), es necesario seguir el codigo antes mostrado. Entre parentesis, se ingresa la ruta de donde va a ser guardado el archivo. Así se podrá obtener el archivo ya limpio para seguir trabajandolo. Volvimos a agregar la columna de EAN pues, el tener

esa columna puede ser útil para la empresa. Tambien hemos exportado el archivo excel sin la columna de EAN, para futuros análisis.