

# Online Coffee Shop

DianaHer

2024-09-10

## Online CoffeeShop

Este estudio se llevó a cabo mediante el análisis de datos de una tienda de café en línea que realiza envíos a tres países diferentes. Se utilizó el conjunto completo de variables relevantes del conjunto de datos para llevar a cabo el análisis. Finalmente, se realizaron dos tipos de análisis estadístico: una prueba t de Student para evaluar la significancia estadística de los datos y un estudio de regresión logística para determinar las probabilidades de recompra en función de ciertas variables. Además, se incluyó un estudio de correlación de Pearson para examinar la relación entre la demanda y el precio de los productos.

El estudio responde a diversas preguntas, tales como si el poseer tarjeta de lealtad influye en el número de compras y probabilidad de recompras, y qué enfoque debería tomar la tienda en línea si quisiera incrementar de manera más rápida sus ingresos según los productos más vendidos y los productos que generan mayor ganancia.

*El estudio se realizó utilizando RStudio.*

El set de datos puede obtenerse aquí (<https://data.world/dx76/coffee-data>)

### PASO 01. Instalar y Cargar paquetes

```
knitr:::opts_chunk$set(echo = TRUE)

## Instalar y cargar paquetes para el análisis.
```

```
options(repos = c(CRAN = "https://cloud.r-project.org/"))
install.packages("tidyverse")
```

```
## Installing package into 'C:/Users/Usuario/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)
```

```
## package 'tidyverse' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Usuario\AppData\Local\Temp\RtmpQthwws\downloaded_packages
```

```
library(tidyverse)
```

```
## — Attaching core tidyverse packages ————— tidyverse 2.0.0 —
## ✓ dplyr     1.1.4    ✓ readr     2.1.5
## ✓forcats   1.0.0    ✓ stringr   1.5.1
## ✓ ggplot2   3.5.0    ✓ tibble    3.2.1
## ✓ lubridate 1.9.3    ✓ tidyrr    1.3.1
## ✓ purrr    1.0.2
```

```
## — Conflicts ————— tidyverse_conflicts() —
## ✘ dplyr::filter() masks stats::filter()
## ✘ dplyr::lag()   masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

*## Activar readxl para Leer Los archivos de excel*

```
library(readxl)
```

*##Instalar las librerías para Limpiar Los datasets*

```
library(readr)
library(tidyverse)
library(skimr)
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

## Cargar y limpiar los sets de Datos

```
knitr::opts_chunk$set(echo = TRUE)

coffee_orders <- read_excel("C:/Users/Usuario/Desktop/Archivos/coffeeOrdersData_2024.xlsx", sheet = 1)
```

```
## New names:
## • ` ` -> `...16`
## • ` ` -> `...17`
```

```

coffee_customers <- read_excel("C:/Users/Usuario/Desktop/Archivos/coffeeOrdersData_2024.xlsx", sheet = 2)

coffee_products <- read_excel("C:/Users/Usuario/Desktop/Archivos/coffeeOrdersData_2024.xlsx", sheet = 3)

##Asegurar la integridad inicial de los datos mediante la limpieza de nombres

coffee_products <- clean_names(coffee_products)
coffee_customers <- clean_names(coffee_customers)
coffee_orders <- clean_names(coffee_orders)

##Revisar la limpieza de los nombres de las columnas

View(coffee_products)
View(coffee_customers)
View(coffee_orders)

```

## Revisar la integridad de los datos

```

knitr::opts_chunk$set(echo = TRUE)

##Buscar valores nulos
sum(is.na(coffee_products)) ##0 valores nulos

## [1] 0

sum(duplicated(coffee_products$product_id)) ##0 valores duplicados

## [1] 0

sum(is.na(coffee_orders)) ##2205 valores nulos

## [1] 2205

##Se cargaron dos columnas enteras con valores nulos, eliminarlas

coffee_orders <- coffee_orders %>%
  select(-x16, -x17)

##Existen valores nulos en el campo email, ya que este no es un dato de importancia para el análisis, simplemente se ignorará la columna

sum(duplicated(coffee_orders$order_id)) ##43 valores duplicados

## [1] 43

```

```
##Con esto verifico que la existencia de valores duplicados es debido a que
##el id de orden aparece por cada producto diferente aunque esté dentro de la misma compra

duplicated_coffee_orders <- coffee_orders %>%
  group_by(order_id) %>%
  filter(n() > 1)

View(duplicated_coffee_orders)

sum(duplicated(coffee_customers$customer_id))## 0 valores duplicados
```

```
## [1] 0
```

```
sum(is.na(coffee_customers))
```

```
## [1] 334
```

```
##334 valores nulos, Los campos con valores vacíos
##(e-mail y phone_number) no son relevantes para el análisis
```

## PASO 02. Obtener información de valor con los datos y realizar análisis estadísticos

```
knitr:::opts_chunk$set(echo = TRUE)

## Realizar algunas consultas exploratorias para conocer los datos

coffee_customers %>%
  group_by(country) %>%
  count(country)
```

```
## # A tibble: 3 × 2
## # Groups:   country [3]
##   country      n
##   <chr>     <int>
## 1 Ireland      150
## 2 United Kingdom    68
## 3 United States    782
```

```
## Los pedidos se dividen en tres países,
## existen 150 ordenes provenientes de Irlanda, 68 de Reino Unido y 782 de Estados Unidos
```

### Datos acerca de LoyaltyCard

```

knitr::opts_chunk$set(echo = TRUE)

##Cuantos clientes tienen Loyalty_card

loyaltycard_customers <- coffee_customers %>%
  filter(loyalty_card == "Yes") %>%
  count()

View(loyaltycard_customers) ##Son 487 clientes con Loyalty card

new_customers <- coffee_customers %>%
  filter(loyalty_card == "No") %>%
  count()

View(new_customers) ##Son 513 clientes nuevos o que no cuentan con Loyalty card

#Conteo de Loyaltycard para crear gráfico

loyalty_card <- coffee_customers %>%
  group_by(loyalty_card) %>%
  count()
View(loyalty_card)

```

## ¿Cuál es el café favorito de los clientes?

```

knitr::opts_chunk$set(echo = TRUE)

##Revisar cual es el producto más ordenado, el café favorito
##Asegurarse que el ID del producto está asociado a un solo producto, y que es igual para ambas tablas

productos_diferentes <- coffee_products %>%
  group_by(product_id) %>%
  count()

View(productos_diferentes)

favorite_coffee <- coffee_orders %>%
  group_by(product_id, coffee_type_name, roast_type_name, size) %>%
  summarise(total_orders = sum(quantity)) %>%
  arrange(desc(total_orders))

## `summarise()` has grouped output by 'product_id', 'coffee_type_name',
## 'roast_type_name'. You can override using the `.`groups` argument.

```

```

View(favorite_coffee)

top_5_coffee <- favorite_coffee %>%
  arrange(desc(total_orders)) %>%
  head(5)

View(top_5_coffee)

```

Existen 48 productos diferentes, el más ordenado corresponde a: *ID: R-L-0.2, Tipo Robusta, tostado Light, tamaño 0.2.*

### ¿Cuál País ordena más café?

```

knitr::opts_chunk$set(echo = TRUE)

##Revisar cual país y ciudad ordena más café

city_orders <- merge(coffee_customers, coffee_orders, by = "customer_id") %>%
  group_by(city, country.x) %>%
  summarise(total_quantity = sum(quantity)) %>%
  arrange(desc(total_quantity))

## `summarise()` has grouped output by 'city'. You can override using the
## `.groups` argument.

```

```

View(city_orders)

country_orders <- merge(coffee_customers, coffee_orders, by = "customer_id") %>%
  group_by(country.x) %>%
  summarise(total_quantity = sum(quantity)) %>%
  arrange(desc(total_quantity))

View(country_orders)

```

El país con mayor consumo de café, medido en piezas ordenadas: **Estados Unidos(2760), y la ciudad de Washginton(90)**

### Porcentaje de Ventas por País

```

knitr::opts_chunk$set(echo = TRUE)

##Aregar una columna calculada para asignar porcentajes

porcentaje_compras <- mutate (country_orders, porcentaje = round(total_quantity/sum(total_quantity)*100, 2))

View(porcentaje_compras)

```

Los porcentajes corresponden a EU:el 77.72%, Irleand: 15.12%, UK: 7.15

### Porcentajes de ventas que provienen de clientes con y sin loyalty card

```

knitr::opts_chunk$set(echo = TRUE)

##Crear un tibble o df solo con las columnas que se requieren para el cálculo

as_tibble <- loyaltocard.tb <- merge(coffee_customers, coffee_orders, by = "customer_id") %>%
  select(country.x, loyalty_card, quantity, sales)

View(loyaltocard.tb)

##Agrupar por lc y obtener los totales de cantidad y ventas

porcentajes_loyaltocard <- loyaltocard.tb %>%
  group_by(loyalty_card) %>%
  summarise(
    total_quantity = sum(quantity),
    total_sales = sum(sales))

View(porcentajes_loyaltocard)

##Obtener los totales en general de las cantidades y ventas

total_quantity_sum <- sum(porcentajes_loyaltocard$total_quantity)
print(total_quantity_sum)

```

```
## [1] 3551
```

```

total_sales_sum <- sum(porcentajes_loyaltocard$total_sales)
print(total_sales_sum)

```

```
## [1] 45134.25
```

El total de ventas es de \$45134.25, de los cuales \$24216.40 corresponde a las ventas de clientes nuevos o clientes que no poseen loyaltocard, mientras que \$20917.85 corresponden a clientes que cuentan con loyaltocard.

```

knitr::opts_chunk$set(echo = TRUE)

##Agregar las columnas calculadas para los porcentajes

porcentajes_loyaltocard <- porcentajes_loyaltocard %>%
  mutate(
    quantity_percentage = round((total_quantity/total_quantity_sum)*100, 2),
    sales_percentage = round((total_sales/total_sales_sum)*100, 2))
View(porcentajes_loyaltocard)

```

Los porcentajes para cada una varían por muy poco: Sin LC: Ventas 53.65%, Piezas ordenadas 53.11% Con LC: Ventas 46.34%, Piezas ordenadas: 46.88%

## ¿Los porcentajes obtenidos son significativos?

Ya que los porcentajes de totales difieren por muy poco es conveniente hacer una prueba de significancia estadística para definir si la diferencia es significativa.

*¿Por qué este estudio es importante? Mediante este resultado se puede definir si las estrategias de marketing deben ser enfocadas principalmente a los clientes nuevos o a obtener recompras por clientes afiliados*

```
knitr::opts_chunk$set(echo = TRUE)

##Crear dos variables para con y sin tarjeta de Lealtad

clientes_con_tarjeta <- loyaltycard.tb %>% filter(loyalty_card == "Yes")
clientes_sin_tarjeta <- loyaltycard.tb %>% filter(loyalty_card == "No")

##Aplicar la prueba t de student para calcular la significancia estadística de las dos variables

significancia_estadística <- t.test(clientes_con_tarjeta$sales, clientes_sin_tarjeta$sales)
print(significancia_estadística)
```

```
##
## Welch Two Sample t-test
##
## data: clientes_con_tarjeta$sales and clientes_sin_tarjeta$sales
## t = -1.051, df = 988.04, p-value = 0.2935
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -8.058753 2.437171
## sample estimates:
## mean of x mean of y
## 43.66983 46.48062
```

Según los resultados obtenidos no hay diferencia estadísticamente significativa entre ambos grupos. La media para el grupo con LC: 43.66, sin LC: 46.48 El valor p obtenido: 0.2935, muy lejos de 0.05 requerido para rechazar la hipótesis nula (hay diferencia significativa). Por lo que se acepta la hipótesis alterna: No ha diferencia estadísticamente significativa a un 95% de confianza. **Con este resultado se confirma que la medida de otorgar loyaltycard a los clientes no aumenta las probabilidades de recompra**

*Conocer la significancia estadística en las probabilidades de recompra puede ayudar a tomar la decisión a la hora del enfoque a las estrategias de marketing, si deben ser dirigidas a grupos de clientes específicos, y que tanto esto aumenta las probabilidades de generar recompras*

**El tamaño de café que más se pide y el que genera mayor ganancia**

```
knitr::opts_chunk$set(echo = TRUE)

## El tamaño de café que más se pide y el que más se vende

product_orders <- coffee_orders %>%
  group_by(size) %>%
  summarise(max_sales = sum(sales),
            max_quantity = sum(quantity)) %>%
  arrange(desc(max_sales))
View(product_orders)
print(max(product_orders$max_quantity))
```

```
## [1] 943
```

```
print(max(product_orders$max_sales))
```

```
## [1] 23785.56
```

Se manejan 4 tamaños: 0.2, 0.5, 1, 2.5. El tamaño de café que más se pide es **0.2**, con un total de **943 unidades** vendidas, y un total de ventas de **\$7029.99**. Sin embargo el café que genera más dinero en ventas es el de tamaño **2.5** con un total de **841 unidades**, y el total de ventas de **\$23785.56**

## Ganancias Netas

```
knitr::opts_chunk$set(echo = TRUE)

## Obtener la ganancia neta generada por cada producto

sales_per_product <- coffee_orders %>%
  group_by(product_id) %>%
  summarise(total_quantity = sum(quantity)) %>%
  arrange(desc(total_quantity))

View(sales_per_product)

## Combinar para obtener información de las dos tablas

product_profit <- merge(coffee_products, sales_per_product, by = "product_id") %>%
  mutate(total_profit = total_quantity * profit) %>%
  select(product_id, coffee_type, size, unit_price, profit, total_profit) %>%
  arrange(desc(total_profit))

View(product_profit)

## Calcular el total total de las sumas del profit

total_profit_sum <- sum(product_profit$total_profit)
print(total_profit_sum)
```

```
## [1] 4520.217
```

El total de ganancias netas es de **\$4520.217**

### Porcentaje de Ganancias netas para los 5 productos más vendidos

```
knitr::opts_chunk$set(echo = TRUE)

##Calcular el porcentaje de ganancias correspondiente para los 5 productos más vendidos

top_5_products <- product_profit %>%
  arrange(desc(total_profit)) %>%
  head(5)

top_5_percentage <- product_profit %>%
  mutate(percentage_profit =
    (total_profit/total_profit_sum)*100) %>%
  arrange(desc(percentage_profit)) %>%
  head(5)

print(top_5_percentage)
```

	product_id	coffee_type	size	unit_price	profit	total_profit	percentage_profit
## 1	L-D-2.5	Lib	2.5	29.785	3.87205	317.5081	7.024178
## 2	L-L-2.5	Lib	2.5	36.455	4.73915	279.6098	6.185761
## 3	E-L-2.5	Exc	2.5	34.155	3.75705	270.5076	5.984394
## 4	A-L-2.5	Ara	2.5	29.785	2.68065	230.5359	5.100106
## 5	E-M-2.5	Exc	2.5	31.625	3.47875	229.5975	5.079346

```
##Porcentaje total de ventas entre los 5 productos que generan mayor cantidad de ingresos

sum_percentage_top_5 <- top_5_percentage %>%
  summarise(representative_percentage = sum(percentage_profit))
print(sum_percentage_top_5)
```

```
##   representative_percentage
## 1                   29.37379
```

### El comportamiento de la demanda de cada producto según su precio

```
knitr::opts_chunk$set(echo = TRUE)

##Calcular el comportamiento de La demanda según el precio del producto

demanda_precio <- merge(favorite_coffee, coffee_products, by = "product_id") %>%
  select(product_id, coffee_type_name, size.x, total_orders, unit_price)
View(demanda_precio)

##Verificar la significancia de La correlación con el coeficiente de pearson

cor(demanda_precio$unit_price, demanda_precio$total_orders)
```

```
## [1] -0.1730607
```

El resultado de la correlación es -0.17 lo cual indica una correlación negativa bastante débil. Es decir que esta tendencia no es significativa. En otras palabras, **el precio puede no ser un factor determinante para la cantidad de pedidos** ya que la relación no es lineal. Una relación lineal es un valor positivo, un valor con una correlación lineal fuerte debe ser igual a 1.

### Top 5 de clientes con más compras

```
knitr::opts_chunk$set(echo = TRUE)

##Top 5 de clientes con mayor cantidad de ordenes de compra

top_5_customers <- coffee_orders %>%
  group_by(customer_id) %>%
  summarise(num_orders = n_distinct(order_id)) %>%
  filter(num_orders > 1) %>%
  arrange(desc(num_orders)) %>%
  head(5)

View(top_5_customers)

top_5_customers_info <- left_join(coffee_customers, top_5_customers, by = "customer_id") %>%
  select(customer_id, num_orders, loyalty_card, city) %>%
  filter(!is.na(num_orders)) ## para filtrar solo Los que aparecen en num_orders

View(top_5_customers_info)
```

**¿Que tan probable es que los clientes hagan recompras?, ¿Tener Loyaltycard influye en la probabilidad de recompra?**

```

knitr::opts_chunk$set(echo = TRUE)

##Obtener datos para conocer la probabilidad de recompra.
##Calcular cuantas órdenes diferentes ha hecho cada cliente.

customer_orders_count <- coffee_orders %>%
  group_by(customer_id) %>%
  summarise(num_orders = n_distinct(order_id)) %>%
  arrange(desc(num_orders))

View(customer_orders_count)

customer_orders_total <- merge(customer_orders_count, coffee_customers, by = "customer_id") %>%
  select(customer_id, num_orders, loyalty_card)
View(customer_orders_total)

##Combinar la tabla con coffee_orders para saber que producto adquirió y cuanto gastó

customers_total_orders_info <- merge(customer_orders_total, coffee_orders, by = "customer_id") %>%
  select(customer_id, order_id, num_orders, loyalty_card, coffee_type_name, sales) %>%
  group_by(order_id)

View(customers_total_orders_info)

```

## Realizar el cálculo de probabilidad de recompra con un modelo de Regresión Logística, basado en tiene o no tiene loyaltycard

```

knitr::opts_chunk$set(echo = TRUE)

##Asignar un número binario a la variable Loyaltycard, 1=Yes, 0=No
##Y de acuerdo a cuantas veces ha comprado el cliente

customer_orders_total <- customer_orders_total %>%
  mutate(buyback = ifelse(num_orders > 1, 1, 0)) ## si num_orders es mayor a 1, se asigna 1, de lo contrario, 0
View(customer_orders_total)

##Una vez establecido el valor binomial de la variable, establecer el modelo de regresión Logística

modelo_logit <- glm(buyback ~ loyalty_card, data = customer_orders_total, family = "binomial")

summary(modelo_logit)

```

```

## 
## Call:
## glm(formula = buyback ~ loyalty_card, family = "binomial", data = customer_orders_total)
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.6420    0.2924 -12.454 <2e-16 ***
## loyalty_cardYes 0.1431    0.4059   0.353    0.724
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
## Null deviance: 229.20 on 912 degrees of freedom
## Residual deviance: 229.08 on 911 degrees of freedom
## AIC: 233.08
## 
## Number of Fisher Scoring iterations: 6

```

```

predicted_probabilities <- predict(modelo_logit, type = "response")

head(predicted_probabilities)

```

```

##      1       2       3       4       5       6
## 0.02553191 0.02553191 0.02934537 0.02934537 0.02934537 0.02553191

```

Según los resultados del modelo de predicción logística tenemos que: La razón de probabilidad cuando  $LC = No$  es de -3.64, lo que indica una probabilidad muy baja de recompra. La razón de probabilidad cuando  $LC = Yes$  es de 0.1431, lo que indica una probabilidad baja, ligeramente mayor a la anterior.

*Para que las probabilidades de recompra sean significativas el valor logarítmico debe ser mayor a 0 0 indica un 50% de probabilidad 1 indica un 73%, lo ideal es obtener un valor mayor a 1 2= 88%; =3= 95%*

El valor p para estos valores es igual a 0.74, lo que indica que la diferencia no es estadísticamente significativa.

**Por lo que las probabilidades de recompra no se ven influenciadas con el hecho de contar o no con una loyalty card.**

### Obtener datos de acuerdo a fechas

```

knitr:::opts_chunk$set(echo = TRUE)

##Asegurarse de que la columna tiene formato de fecha

coffee_orders$order_date <- as.Date(coffee_orders$order_date, format = "%Y-%m-%d")
str(coffee_orders$order_date)

## Date[1:1000], format: "2019-09-05" "2019-09-05" "2021-06-17" "2021-07-15" "2021-07-15" ...

```

```
View(coffee_orders)
str(coffee_orders)
```

```
## tibble [1,000 x 15] (S3:tbl_df/tbl/data.frame)
## $ order_id      : chr [1:1000] "QEV-37451-860" "QEV-37451-860" "FAA-43335-268" "KAC-83089-793" ...
## $ order_date    : Date[1:1000], format: "2019-09-05" "2019-09-05" ...
## $ customer_id   : chr [1:1000] "17670-51384-MA" "17670-51384-MA" "21125-22134-PX" "23806-46781-OU" ...
## $ product_id    : chr [1:1000] "R-M-1" "E-M-0.5" "A-L-1" "E-M-1" ...
## $ quantity      : num [1:1000] 2 5 1 2 2 3 3 1 3 1 ...
## $ customer_name : chr [1:1000] "Aloisia Allner" "Aloisia Allner" "Jami Redholes" "Christoffer O' Shea" ...
## $ email         : chr [1:1000] "aallner0@lulu.com" "aallner0@lulu.com" "jredholes2@tmall.com" NA ...
## $ country       : chr [1:1000] "Finland" "Finland" "Finland" "Ireland" ...
## $ coffee_type   : chr [1:1000] "Rob" "Exc" "Ara" "Exc" ...
## $ roast_type    : chr [1:1000] "M" "M" "L" "M" ...
## $ size          : num [1:1000] 1 0.5 1 1 2.5 1 0.5 0.2 0.5 0.5 ...
## $ unit_price    : num [1:1000] 9.95 8.25 12.95 13.75 27.48 ...
## $ sales         : num [1:1000] 19.9 41.2 12.9 27.5 55 ...
## $ coffee_type_name: chr [1:1000] "Robusta" "Excelsa" "Arabasta" "Excelsa" ...
## $ roast_type_name: chr [1:1000] "Medium" "Medium" "Light" "Medium" ...
```

*##verificar que no hay valores nulos*

```
sum(is.na(coffee_orders$order_date)) ## no hay valores nulos
```

```
## [1] 0
```

```
knitr::opts_chunk$set(echo = TRUE)
```

*##Agregar dos columnas nuevas que contengan por separado únicamente el mes, año*

```
coffee_orders_dates <- coffee_orders %>%
  mutate(year = format(order_date, "%Y"),
        month = format(order_date, "%m-%Y"))
```

```
View(coffee_orders_dates)
```

*##Calcular el total de ventas por año*

```
yearly_sales <- coffee_orders_dates %>%
  group_by(year) %>%
  summarise(sales_per_year = sum(sales, na.rm = TRUE))
```

```
print(yearly_sales)
```

```
## # A tibble: 4 × 2
##   year   sales_per_year
##   <chr>     <dbl>
## 1 2019     12187.
## 2 2020     12118.
## 3 2021     13766.
## 4 2022     7063.
```

```
monthly_sales <- coffee_orders_dates %>%
  group_by(month, year) %>%
  summarise(sales_per_month = sum(sales, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'month'. You can override using the
## `.groups` argument.
```

```
print(monthly_sales)
```

```
## # A tibble: 44 × 3
## # Groups:   month [44]
##   month   year   sales_per_month
##   <chr>   <chr>     <dbl>
## 1 01-2019 2019     829.
## 2 01-2020 2020     567.
## 3 01-2021 2021     838.
## 4 01-2022 2022    1269.
## 5 02-2019 2019     987.
## 6 02-2020 2020    1798.
## 7 02-2021 2021     959.
## 8 02-2022 2022     394.
## 9 03-2019 2019    1021.
## 10 03-2020 2020     915.
## # ... with 34 more rows
```

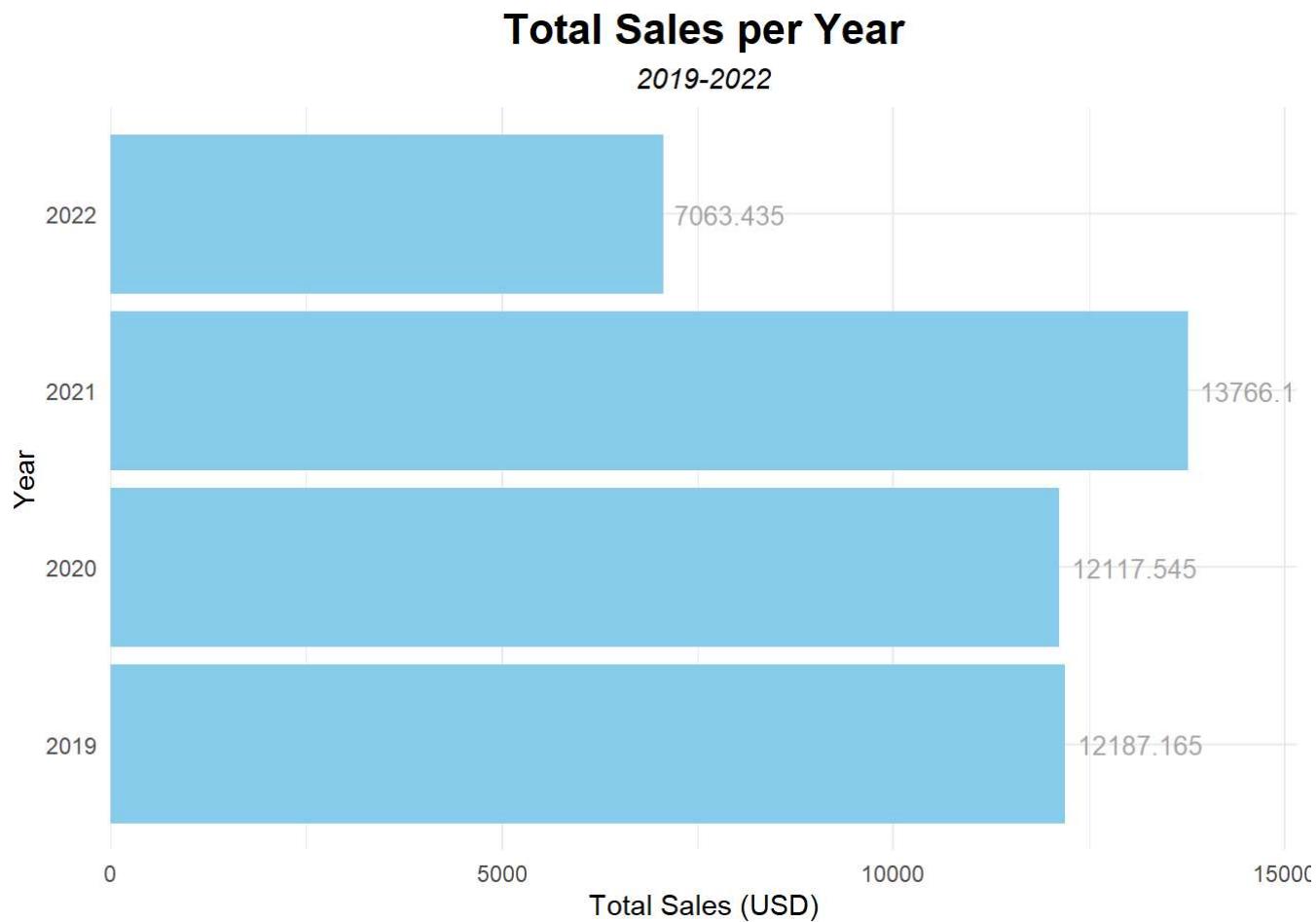
## PASO 03. Crear Gráficas e Insights de acuerdo a los resultados

### Resumen de Ventas por Año

```
knitr::opts_chunk$set(echo = TRUE)

##Ventas por año, GRÁFICO DE BARRAS

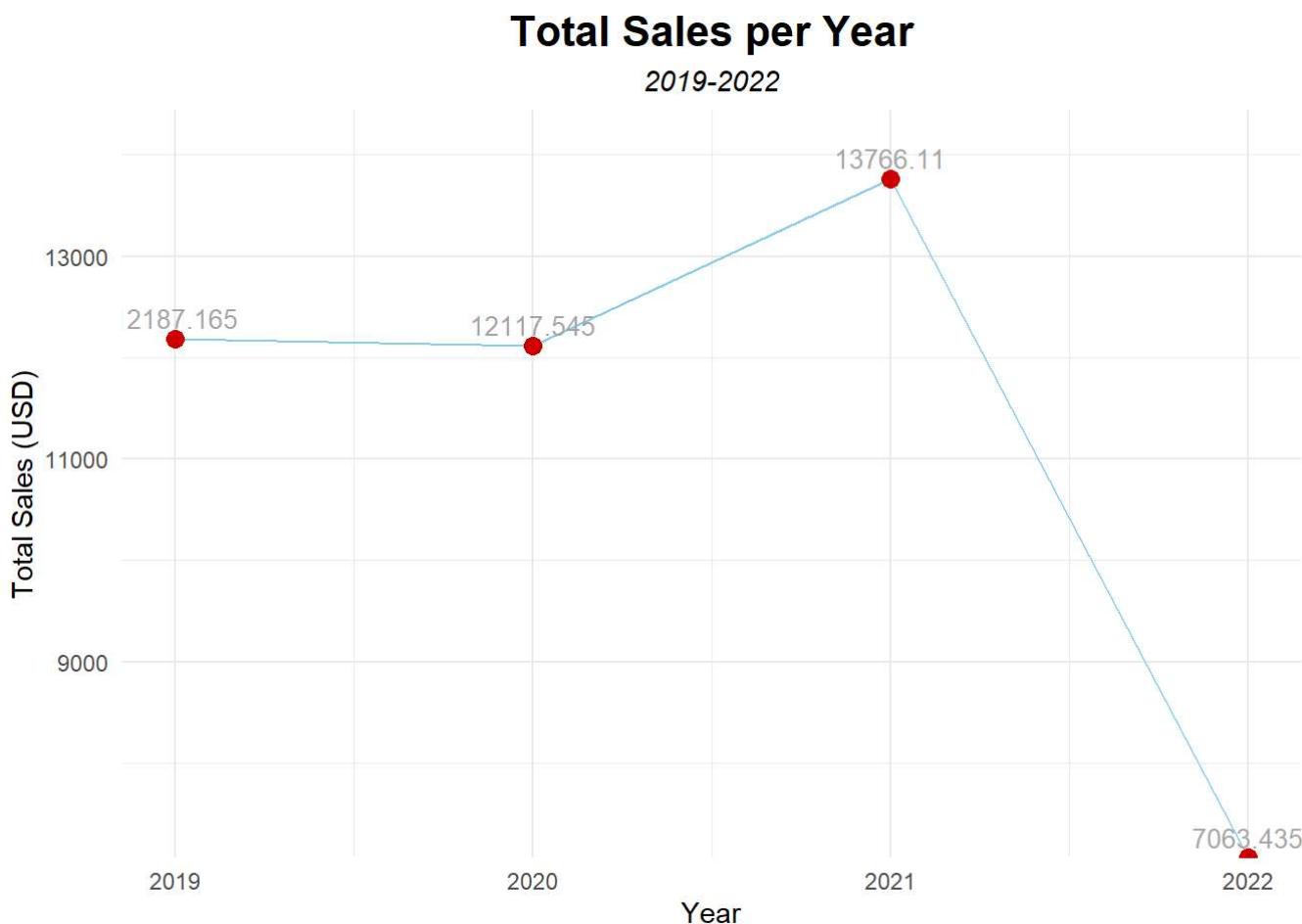
yearly_sales$year <- as.factor(yearly_sales$year)
ggplot(yearly_sales, aes(x = sales_per_year, y = year))+
  geom_bar(stat = "identity", fill = "skyblue")+
  geom_text(aes(label = sales_per_year), hjust = -0.1, color = "darkgray", size = 3.5)+
  labs(title = "Total Sales per Year", subtitle = "2019-2022", x = "Total Sales (USD)", y = "Year")+
  scale_x_continuous(expand = expansion(mult = c(0, 0.1)))+
  ##scale_y_discrete(labels = label_number(scale = 1e-2, suffix = "K"))+ reiniciar R y revisar que se haya aplicado
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
        plot.subtitle = element_text(hjust = 0.5, size = 11, face = "italic"))
```



```
knitr::opts_chunk$set(echo = TRUE)

##VENTAS POR AÑO, gráfico de Líneas

yearly_sales$year <- as.numeric(as.character(yearly_sales$year))
ggplot(yearly_sales, aes(x = year, y = sales_per_year)) +
  geom_line(color = "skyblue") +
  geom_point(color = "red3", size = 3) +
  geom_text(aes(label = sales_per_year), vjust = -0.5, color = "darkgray", size = 3.5) +
  labs(title = "Total Sales per Year", subtitle = "2019-2022", x = "Year", y = "Total Sales (USD)") +
  scale_y_continuous(expand = expansion(mult = c(0, 0.1))) +
  scale_x_continuous(breaks = yearly_sales$year) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5, size = 11, face = "italic"))
```

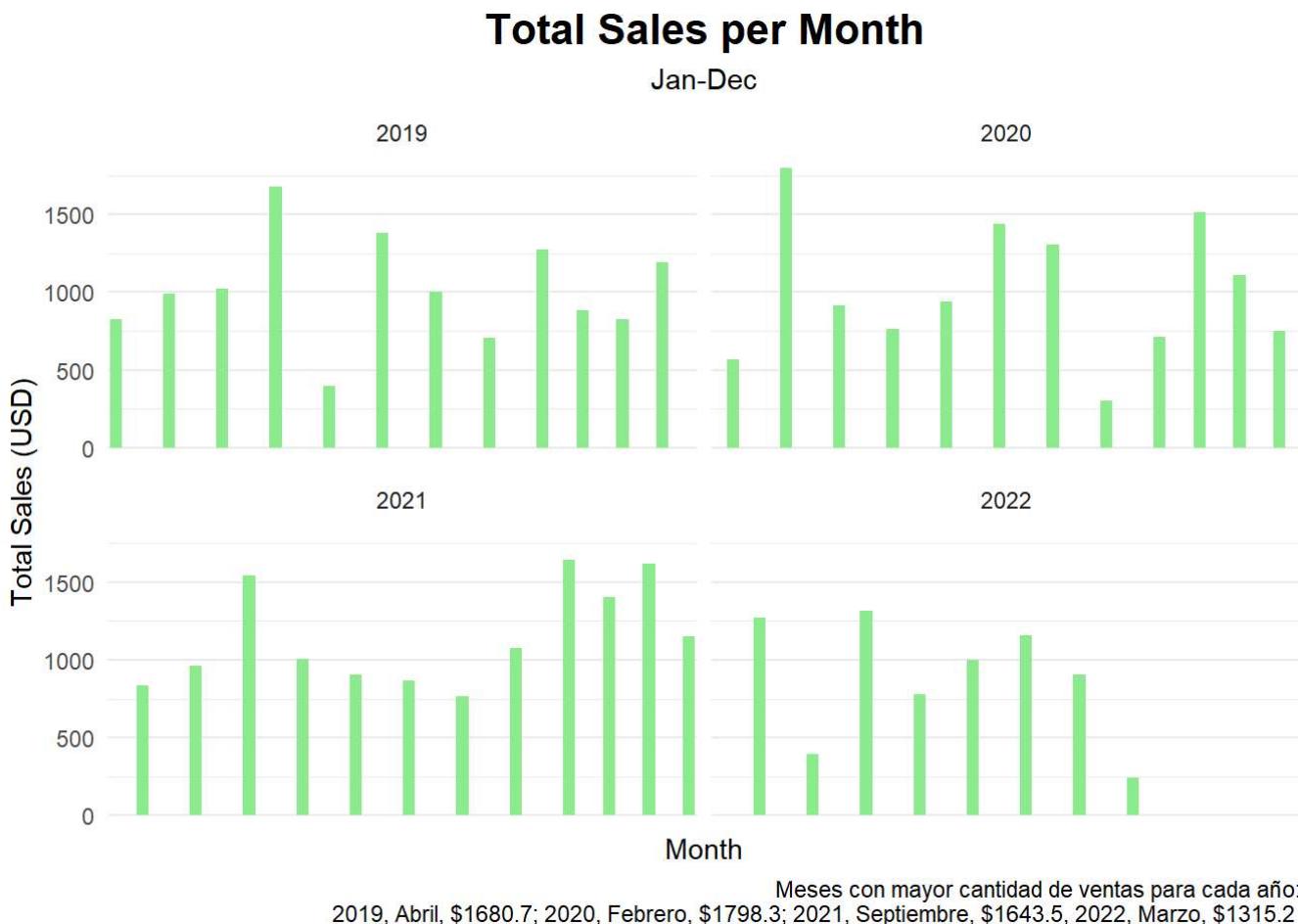


Mediante estas dos gráficas es posible observar que las ventas no llevan un comportamiento lineal ascendente. Es importante hacer énfasis en que el resumen de ventas para el año 2022 cuenta con la mitad de los datos (hasta el mes de julio), por lo que no es conveniente considerar el comportamiento del último año de ventas.

## Resumen de Ventas por mes

```
knitr::opts_chunk$set(echo = TRUE)

yearly_sales$year <- as.factor(yearly_sales$year)
ggplot(monthly_sales, aes(x = factor(month), y = sales_per_month)) +
  geom_bar(stat = "identity", fill = "lightgreen") +
  labs(title = "Total Sales per Month", subtitle = "Jan-Dec", x = "Month", y = "Total Sales (USD)", caption = "Meses con mayor cantidad de ventas para cada año:
  2019, Abril, $1680.7; 2020, Febrero, $1798.3; 2021, Septiembre, $1643.5, 2022, Marzo, $1315.2",
  caption.color = "darkgray") +
  facet_wrap(~ year, ncol = 2) +
  theme_minimal() +
  scale_x_discrete(breaks = seq(1, 12, by = 1)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5, size = 16, face = "bold"), plot.subtitle = element_text(hjust = 0.5, size = 11))
```

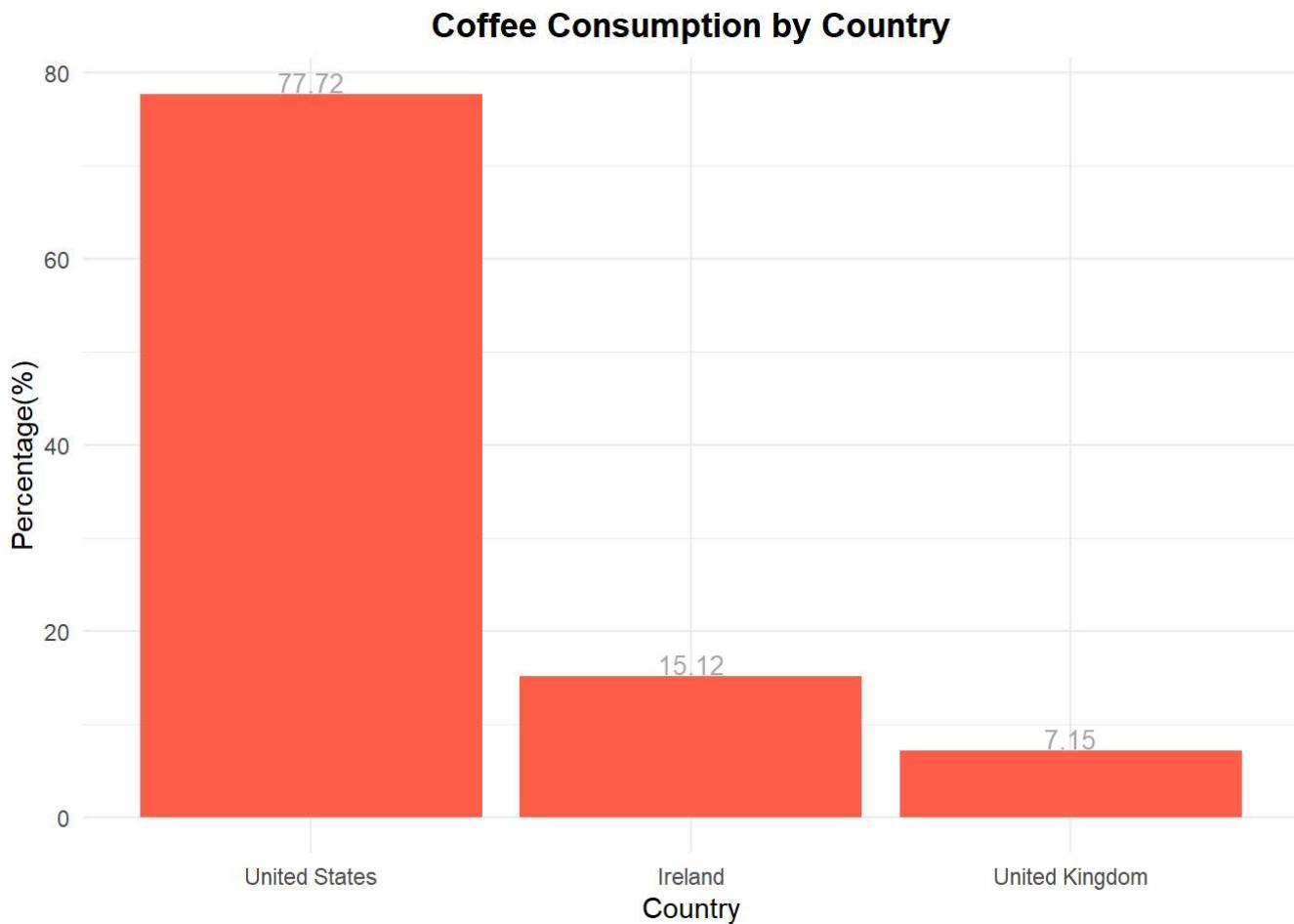


En este gráfico se pueden observar los meses con mayor cantidad de ventas, para ninguno de los años el mes con mayor número de ventas ha sido el mismo y las ventas tampoco muestran una tendencia similar, sino que el comportamiento ha sido diferente para cada año.

## País con mayor consumo de café

```
knitr::opts_chunk$set(echo = TRUE)

porcentaje_compras <- porcentaje_compras %>%
  mutate(country.x = reorder(country.x, -porcentaje))
ggplot(porcentaje_compras, aes(x = country.x, y = porcentaje)) +
  geom_bar(stat = "identity", fill = "tomato") +
  labs(title = "Coffee Consumption by Country", x = "Country", y = "Percentage(%)") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold")) +
  geom_text(aes(label = porcentaje), vjust = -0.1, color = "darkgray", size = 3.5)
```



En este gráfico puede verse el porcentaje de ventas correspondiente a cada uno de los tres países que conforman las ventas de la tienda en línea. Se aprecia que el 77.7% de las ventas se generan en Estados Unidos.

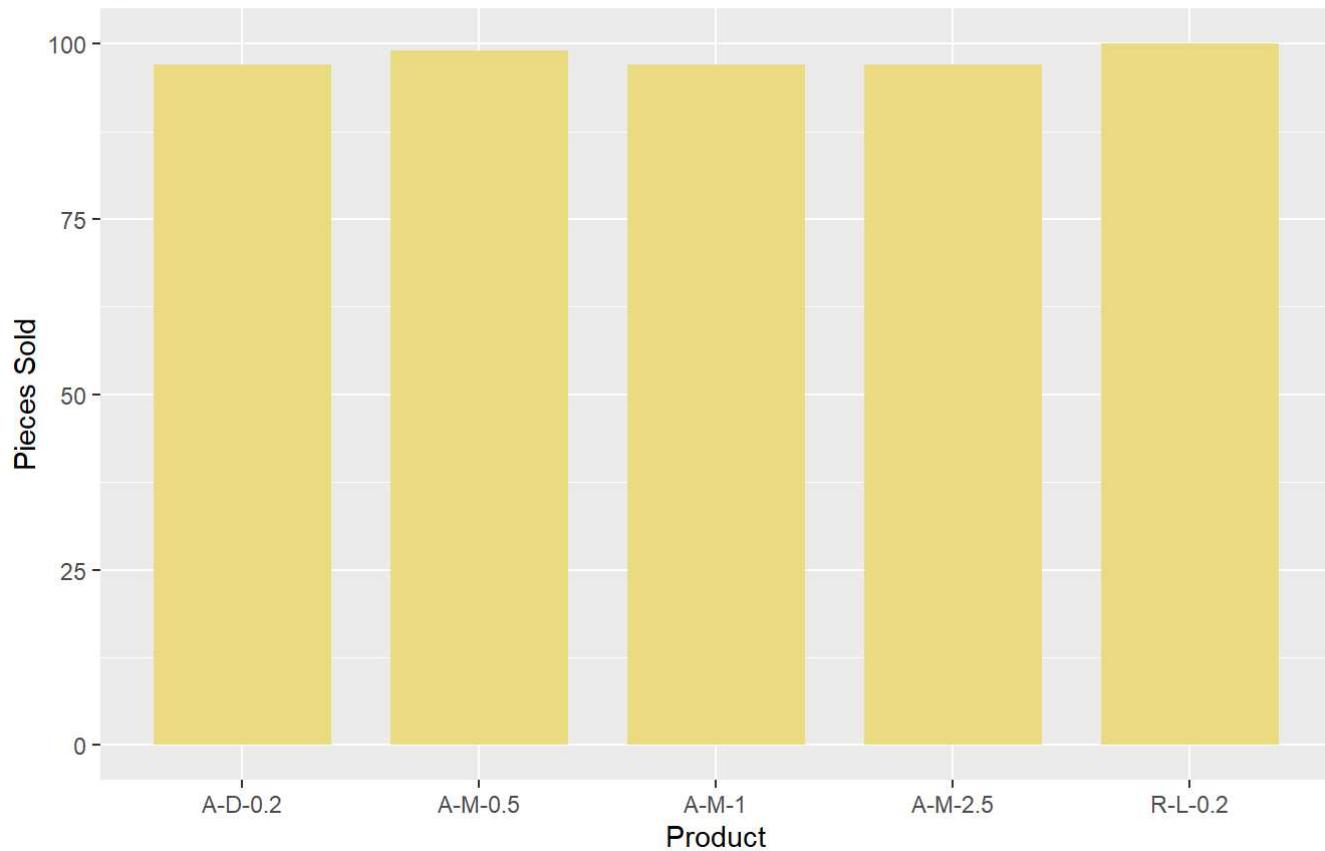
## Los 5 cafés más pedidos

```
knitr::opts_chunk$set(echo = TRUE)

ggplot(top_5_coffee, aes(x = product_id, y = total_orders)) +
  geom_bar(stat = "identity", fill = "lightgoldenrod", width = 0.75) +
  labs(title = "Favorite Coffees", subtitle = "Arabasta-Robusta", x = "Product", y = "Pieces Sold")
```

## Favorite Coffees

Arabasta-Robusta



Este gráfico ilustra cómo, entre los 48 productos diferentes disponibles en la tienda, los 5 productos más pedidos no muestran una diferencia significativa en la cantidad de órdenes. Esto sugiere que no hay un producto que se destaque de manera considerable en comparación con los demás.

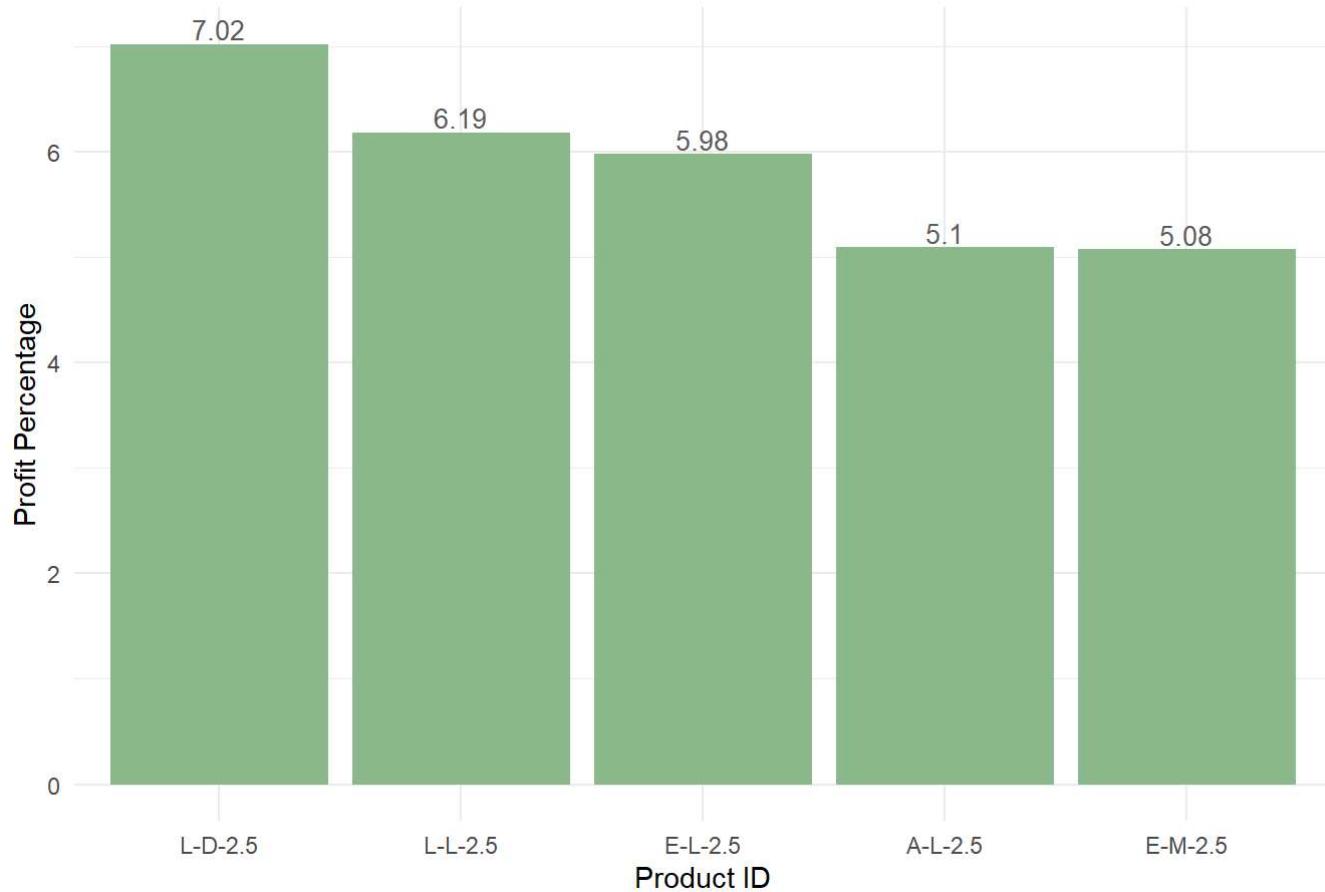
## Entonces... ¿Cuál café genera más ingresos?

```
knitr::opts_chunk$set(echo = TRUE)

##CREAR GRÁFICO CON EL ID DE LOS 5 MÁS RENTABLES, REPRESENTADOS EN PORCENTAJE

ggplot(top_5_percentage, aes(x = reorder(product_id, -percentage_profit), y = percentage_profit)) +
  geom_bar(stat = "identity", fill = "darkseagreen") +
  labs(title = "Most Profitable Products", y = "Profit Percentage", x = "Product ID") +
  geom_text(aes(label = round(percentage_profit, 2)), vjust = -0.2, hjust = 0.5, color = "gray3",
  size = 3.5) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```

## Most Profitable Products



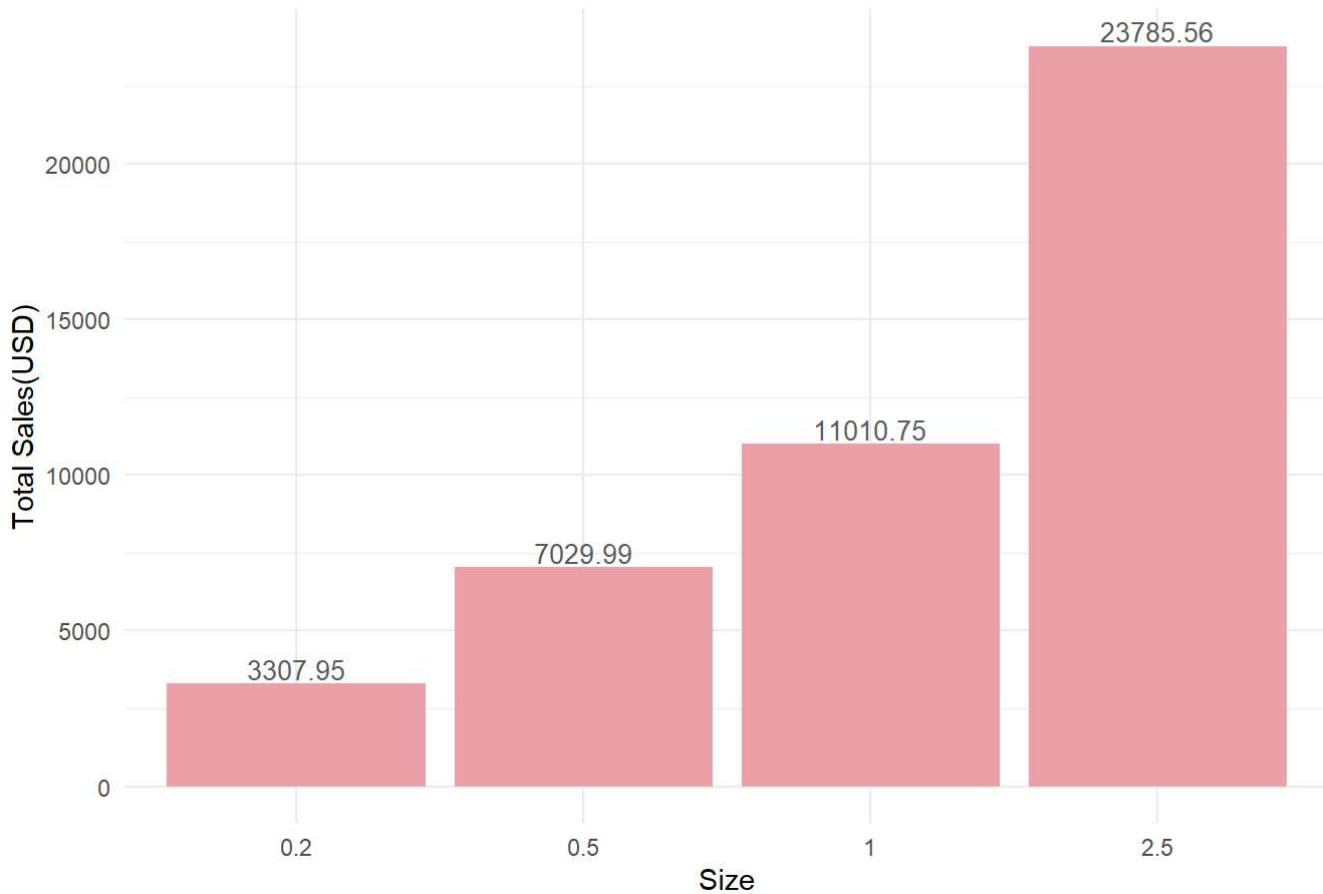
Como puede observarse, los 5 productos más vendidos *no representan un conjunto de porcentaje mayor al 30%*. Por lo cual, no es conveniente enfocarse en impulsar las ventas por producto, sino por tamaño, como es posible observar en las siguientes gráficas.

## El tamaño de café que genera más ingresos

```
knitr::opts_chunk$set(echo = TRUE)

ggplot(product_orders, aes(y = max_sales, x = factor(size))) +
  geom_bar(stat = "identity", fill = "lightpink2") +
  labs(title = "Sales by Size", y = "Total Sales(USD)", x = "Size") +
  geom_text(aes(label = round(max_sales, 2)), vjust = -0.2, hjust = 0.5, color = "gray36", size = 3.5) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```

### Sales by Size



A diferencia de los anteriores gráficos, es posible observar una evidente diferencia en la cantidad de ventas generadas por producto si nos enfocamos en el tamaño. Ya que no hay diferencia significativa en las ventas por producto más ordenado, es recomendable implementar el enfoque de marketing dirigido al tamaño de producto y no al tipo de café o al que más se pide, ya que los datos demuestran que el café que más se pide no es necesariamente el que genera más ingresos.

**¿El tamaño de café que genera más ingresos es el que más se vende?**

```
knitr::opts_chunk$set(echo = TRUE)

ggplot(product_orders, aes(y = max_quantity, x = factor(size))) +
  geom_bar(stat = "identity", fill = "thistle2", width = 0.7) +
  labs(title = "Orders by Size", y = "Total Orders", x = "Size") +
  geom_text(aes(label = round(max_sales, 2)), vjust = -0.2, hjust = 0.5, color = "gray36", size = 3.5) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```



Comparando los últimos dos gráficos se puede observar como el tamaño de café que más se pide es el de tamaño 0.5, sin embargo **no es necesariamente el que genera una mayor cantidad de ingresos**, por lo que se sugiere aumentar el foco de atención hacia el tamaño de café de 2.5

## Oferta y Demanda... ¿La demanda del producto está relacionada con su precio?

```
knitr::opts_chunk$set(echo = TRUE)

ggplot(demanda_precio)+
  geom_point(mapping = aes(x = unit_price, y = total_orders))+
```

```
  geom_smooth(method = lm, se = FALSE, aes(x = unit_price, y = total_orders))+
```

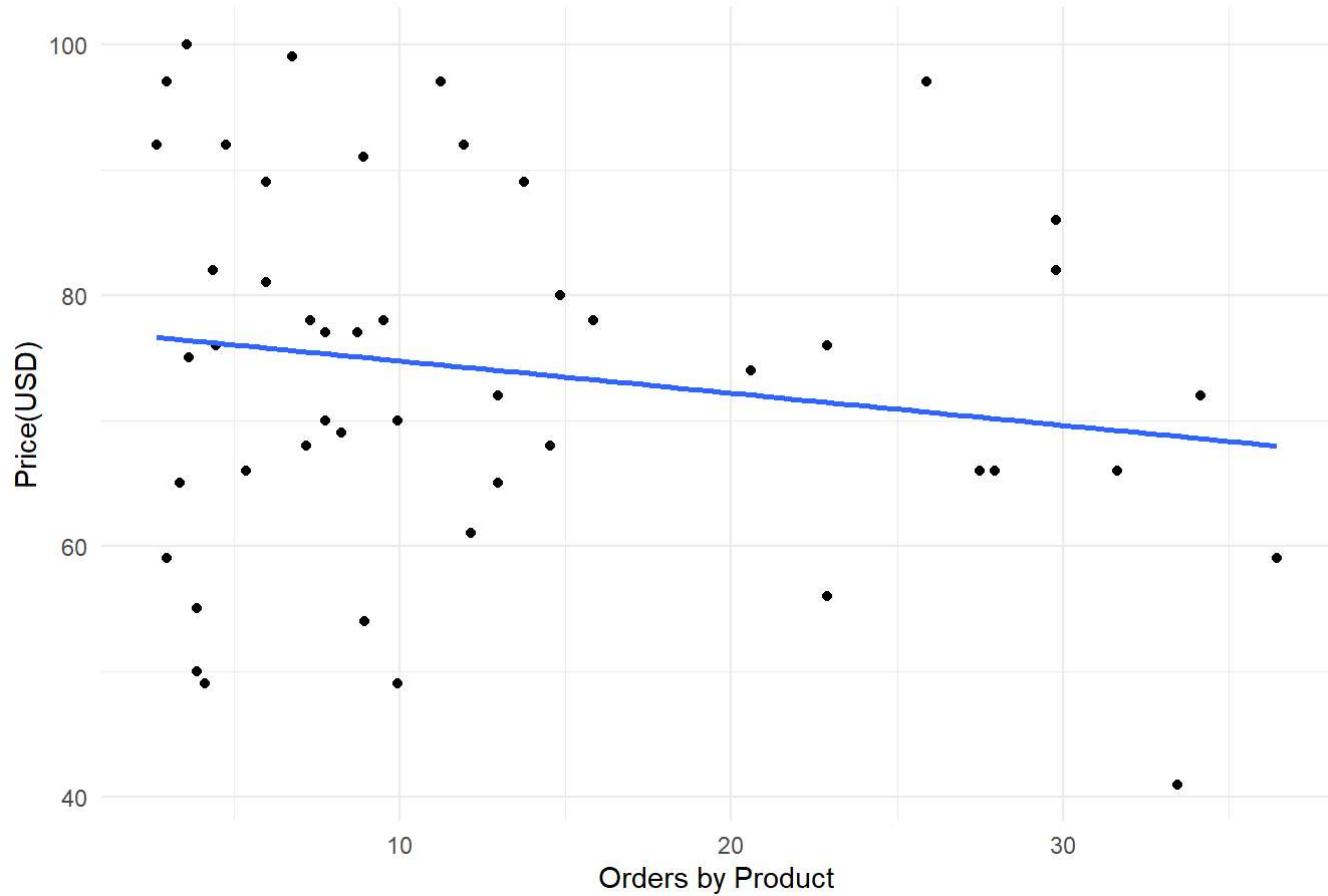
```
  labs(title = "Price vs Demand", x = "Orders by Product", y = "Price(USD)")+
```

```
  theme_minimal()+
```

```
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

### Price vs Demand



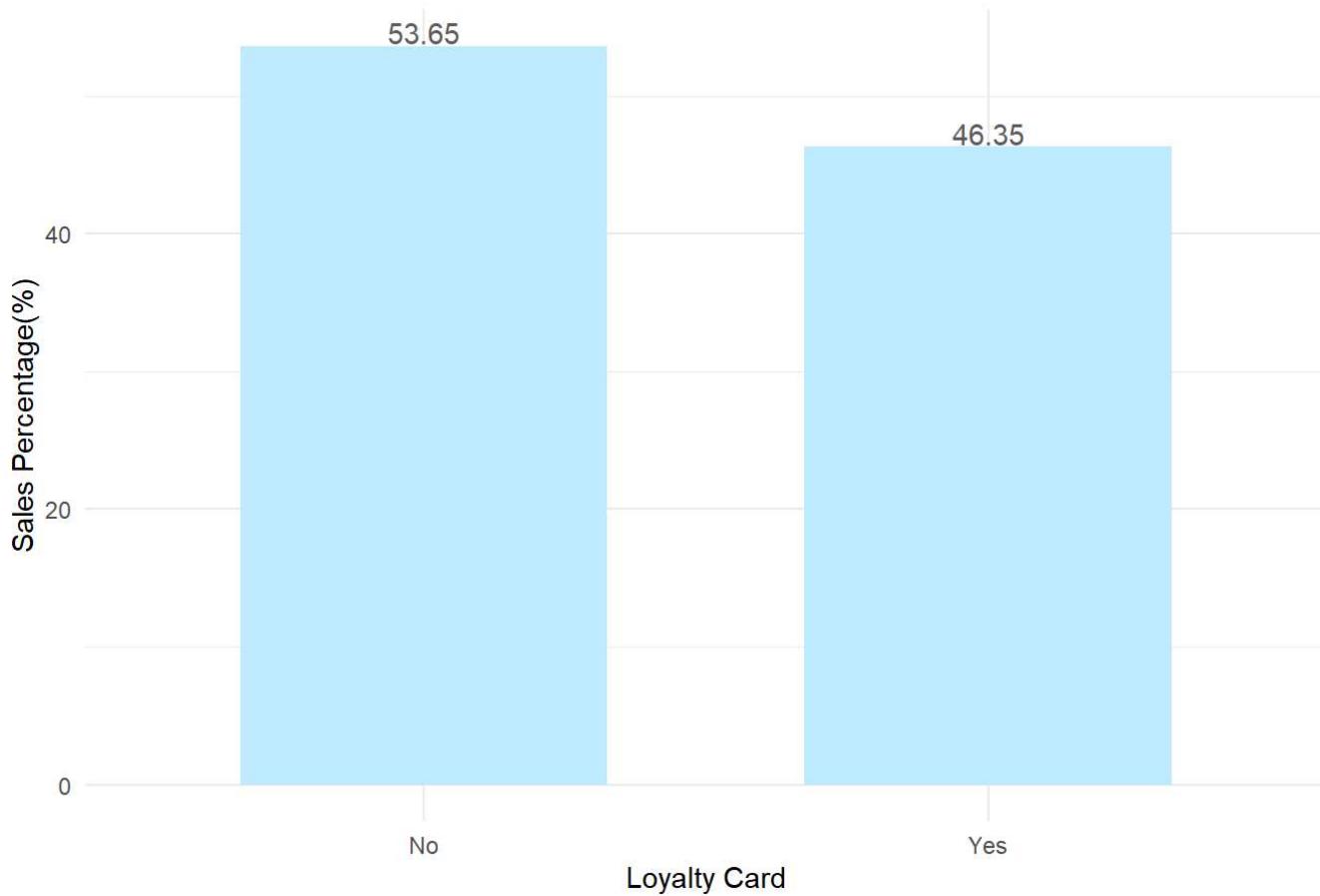
El gráfico de dispersión muestra una correlación negativa débil -0.17, la tendencia de compra se reduce de manera no lineal según el precio del producto. Es decir que esta tendencia no es significativa. En otras palabras, **el precio puede no ser un factor determinante para la cantidad de pedidos** ya que la relación no es lineal.

## ¿Tener una tarjeta de Lealtad marca una diferencia?

```
knitr::opts_chunk$set(echo = TRUE)

ggplot(porcentajes_loyaltycard, aes(x = loyalty_card, y = sales_percentage)) +
  geom_bar(stat = "identity", fill = "lightblue1", width = 0.65) +
  geom_text(aes(label = sales_percentage), color = "gray36", vjust = -0.1) +
  labs(title = "Distribution of Total Purchases",
       x = "Loyalty Card", y = "Sales Percentage(%)") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```

## Distribution of Total Purchases



Con este gráfico observamos que el porcentaje de ventas es mayor para los clientes sin tarjeta de lealtad, sin embargo el estudio de significancia estadística demostró que la diferencia no es significativa ( $p = 0.29$ ), así como tampoco lo es para la probabilidad de recompra basada en el mismo dato ( $p = 0.74$ ), por lo que la información se resume en que, **la probabilidad de compra y recompra no se ve influenciada por el hecho de poseer tarjeta de lealtad**

### **Resumen de datos obtenidos**

Las tendencias observadas en los datos no muestran un comportamiento lineal claro o predecible. Es posible que la distribución de las ventas en tres países diferentes sea un factor clave que contribuye a la dispersión no lineal en los datos. Basándose en la información obtenida se puede rescatar que, los productos más adquiridos corresponden a:

- R-L-02-Robusta-Light-0.2
- A-M-05-Arabasta-Medium-0.5
- A-D-0.2-Dark-0.2
- A-M-1-Arabasta-Medium-1.0
- A-M-2.5-Arabasta-Medium-2.5

Sin embargo no representan una diferencia estadísticamente significativa con el resto de los productos. El tamaño de café más ordenado es 0.2, por lo que se recomienda contar con inventario, sin embargo el tamaño de café que genera más ingresos es el 2.5.

El país con mayor consumo de café y que representa el mayor porcentaje de ventas es Estados Unidos.

La demanda de los productos no muestra una correlación que se encuentre ligada al precio del mismo.

El porcentaje de compra y probabilidad de recompra no se ve afectado por poseer o no una tarjeta de lealtad.

***Recomendaciones finales:***

- Enfocar las campañas de marketing hacia el tamaño y no al tipo de producto, ya que es donde se verá reflejado un aumento en las ganancias netas.
- Dirigir las campañas de publicidad hacia todos los grupos de clientes, sin importar que cuenten o no con loyaltycard.
- Ofrecer beneficios adicionales para clientes con LC si desean intentar aumentar las probabilidades de recompra basados en esta variable.
- Fortalecer las estrategias de ventas para intentar obtener mas clientes en los países con menos porcentajes de ventas.
- Tener una visión clara de las ventas por tamaño y por id de producto es importante para control de inventario.

***Sugerencias de Análisis Posteriore:***

- Enfocar la búsqueda de correlación entre la demanda del producto con un factor diferente al precio del mismo.
- Buscar la probabilidad de recompra considerando variables como el tamaño, el tipo de café que adquirió el cliente o cuanto gastó.

**¡Gracias por interesarte en mi estudio!**