Second project Coffee and Bean sales

Jeisson Steve Rojas Velasquez
May 2025.

Introduction

The purpose of this project is to analyze coffee consumption data in order to identify trends and segment customers based on their consumption patterns.

The database used in this project is intended for individuals, consumers, and distributors of coffee, with the purpose of understanding customer profiles by country, city, roast type, coffee type, and is considered valuable information that helps identify consumption trends based on location, coffee type, roast type, and even size.

The goal is to infer a lot of information regarding global coffee consumption, preferences for types and roasts, as well as profitability.

Specific Objectives

- Identify coffee consumption trends in customers related to coffee type and customer location.
- Identify the most profitable coffee types and provide insights to improve consumption.
- Identify trends in coffee product orders, including which products are ordered the most and which the least.
- Attempt to segment customers based on the information in the databases.

Database

The database used for this project was:

https://www.kaggle.com/datasets/saadharoon27/coffee-bean-sales-raw-dataset/data

This database is composed of:

Orders Worksheet:

Order ID: A unique identifier for each coffee order.
Order Date: The date when the order was placed.
Customer ID: An identifier linking the order to a specific customer.
Product ID: A unique identifier for each coffee product.
Quantity: The quantity of the coffee product ordered.

Customers Worksheet:

Customer ID: A unique identifier for each customer.
Customer Name: The name of the customer.
Email Address: Contact information for customers.
Phone Number: Another contact detail for customers.

And more: Explore a wide range of customer attributes for segmentation and analysis.

Products Worksheet:

Product ID: A unique identifier for each coffee product.
Coffee Type: The type or blend of coffee, such as Arabica or Robusta.
Roast Type: The roast level, including light, medium, or dark roast.
Size: Information about the product size.
Unit Price: The price of a single unit of the coffee product.
Price Per 100g: The price per 100 grams for detailed price comparisons.
Profit: Insights into the profitability of each coffee product.

Methodology

Methodology

The project was developed in three main phases: data preparation, including ETL, exploratory analysis, and result visualization.

1. **Data Preparation:**

In Excel, an initial cleaning and transformation of the data was performed. In the next steps

- ✓ Numeric formats (decimals, currency) were adjusted according to the type of variable, and new columns such as Profit Margin and Profit-to-Price Index were created using custom formulas. Additionally, calculated columns were created in pivot tables.

- ✓ Missing values were identified and handled: empty email fields were replaced with notiene@correo.com, and empty phone numbers were replaced with **0**.

- ✓ Power query were used to clean the data and change some errors, blanks values, null values and delete some unnecessary columns.

In SQL, .csv files were imported, tables were created, and cleaning tasks were carried out: normalization of formats, removal of inconsistencies, creation of foreign keys and indexes, and joins between tables. Several analysis queries were performed, including Total Price by Coffee Type and Most Profitable Coffee, among others. Stored procedures were also created to calculate total sales by month and by coffee type.

2. **Exploratory Data Analysis (EDA):**

    - In Excel, pivot tables were used, and descriptive statistical analysis techniques were applied to key variables such as Unit Price, Price per 100g, Profit, Profit Margin, and Quantity. This allowed a preliminary understanding of consumption patterns.

Descriptive statistics from unit price

Unit price	
Promedio unit price	$ 12,91
Max venta unit price	$ 36,46
Min venta	$ 2,69
Moda	$ 5,97
Mediana	$ 8,95
Rango	$ 33,77
Desviacion estandar	9,772455372
Varianza	95,500884
Asimetria	0,991699596
Curtosis	-0,3233117

Unit Price

The variable Unit Price shows a general behavior where the mean is $12.91, while the median is $8.95. This difference indicates that the distribution is asymmetric and skewed to the right. The mode is $5.97, which represents the most common value in the data.

The range (from a minimum of $2.69 to a maximum of $36.46) is $33.77, and the interquartile range (IQR) of $10.48 shows a high dispersion in the central prices. The standard deviation of $9.72 supports that the values are quite spread out around the mean.

The skewness of 0.99 means the distribution has a right tail, caused by moderately high unit prices. The kurtosis of -0.3 suggests the distribution is platykurtic, meaning it is flatter than a normal distribution with fewer extreme values.

No significant outliers were detected in this variable, so no additional adjustments were necessary.

Descriptive statistics from profit

Profit		
Promedio profit		$ 1,30
Max profit		$ 4,74
Min profit		$ 0,16
Moda		$ 0,61
Mediana		$ 0,98
Rango		$ 4,58
Desviacion estandar		1,128051908
Varianza		1,272501107
Asimetria		1,428279596
Curtosis		1,274058504

Profit

The variable Profit shows a general behavior where the mean is $1.11, while the median is $0.98. This small difference indicates a slightly right-skewed distribution. The mode is $0.61, representing the most frequent value in the data.

The range between the minimum ($0.25) and maximum ($3.89) is $3.64, and the IQR is $1.35, showing a moderate dispersion in the central values. The standard deviation of $0.79 suggests moderate variability around the mean.

The skewness of 0.67 shows the distribution has a right tail, caused by some relatively high values. The kurtosis of 0.5 suggests the distribution is slightly leptokurtic, meaning it has a bit more concentration of values around the mean.

Six outliers were excluded for this version of the analysis. These cases might represent special customers, errors, or specific promotions, so it is good to review them before making decisions based on this data.

Profit Margin		
Promedio		10%
Max profit		13%
Min profit		6%
Moda		13%
Mediana		9%
Rango		7%
Desviacion estandar		0,025591757
Varianza		0,000654938
Asimetria		-0,24413813
Curtosis		-1,2196318

Profit Margin

The variable Profit Margin has a mean of 10% and a median of 9%, which indicates a slightly left-skewed distribution. The mode is 13%, which is the most frequent value in the data.

The range (from 7% to 12%) is 5%, and the IQR of 3% shows low dispersion in profit margins. The standard deviation of 2.1% also supports that the margins are relatively consistent.

The skewness of -0.24 indicates a left tail, caused by slightly low values. The kurtosis of -1.2 suggests the distribution is platykurtic, meaning it is flatter than normal and has few extreme values.

No significant outliers were found in this variable, so no additional adjustments were needed.

Quantity		
Promedio Quantity		3,55
Max		6
Min		1
Moda		2
Mediana		4
Rango		5
Desviacion estandar		1,6817334
Varianza		2,828227227
Asimetria		0,010824673
Curtosis		-1,24787039

Quantity

The variable Quantity shows a mean of 3.55 and a median of 4, indicating a slightly left-skewed distribution. The mode is 2, which is the most frequent value in the data.

The range of values (from 1 to 6) is 5, and the IQR is 3, showing low dispersion in quantities sold. The standard deviation is 1.68, indicating relatively consistent margins.

With a skewness of 0.01, the distribution shows a slight right tail caused by very high values. The kurtosis of -1.24 tells us the distribution is platykurtic, also flatter than normal, with few extreme values.

No significant outliers were identified, so no further adjustments were necessary.

3. **Visualization and Insight Generation:**

Interactive dashboards were designed in **Excel** and **Power BI**, organized into three key areas:

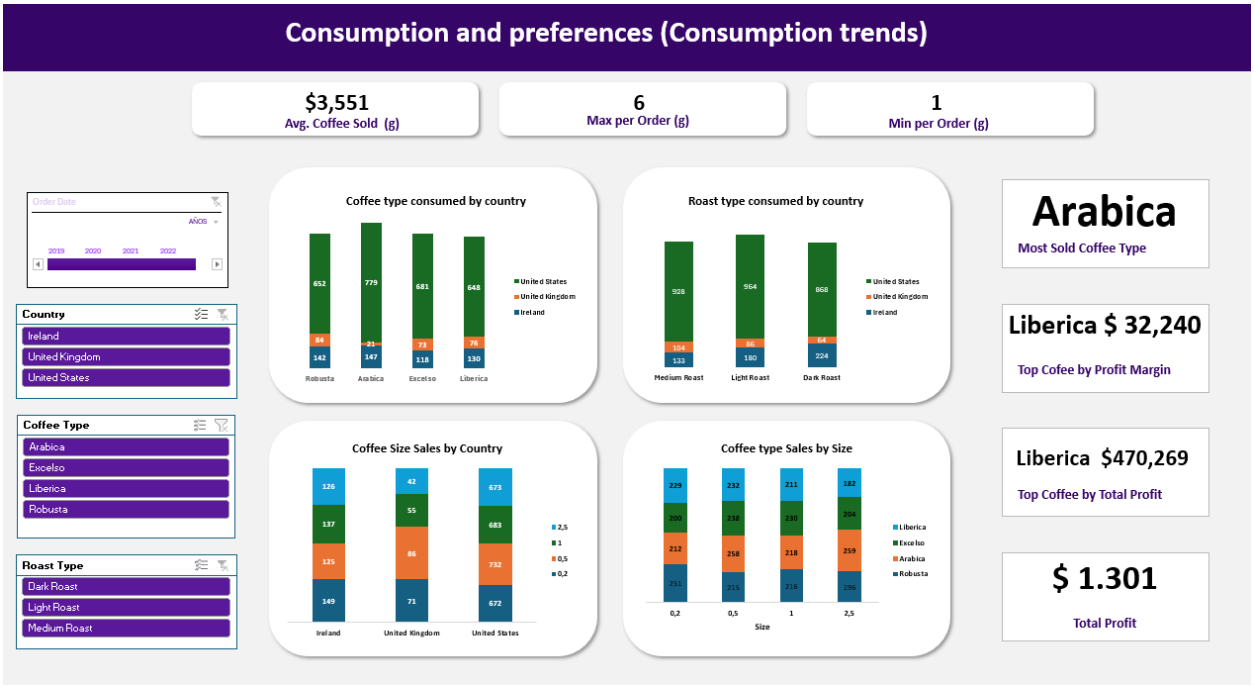
- Consumption Trends
- Prices and Profitability
- Customer Clusters

These dashboards allow the identification of consumption trends, customer segments, and the most profitable coffee types based on roast and geographic demand, as you can see below.

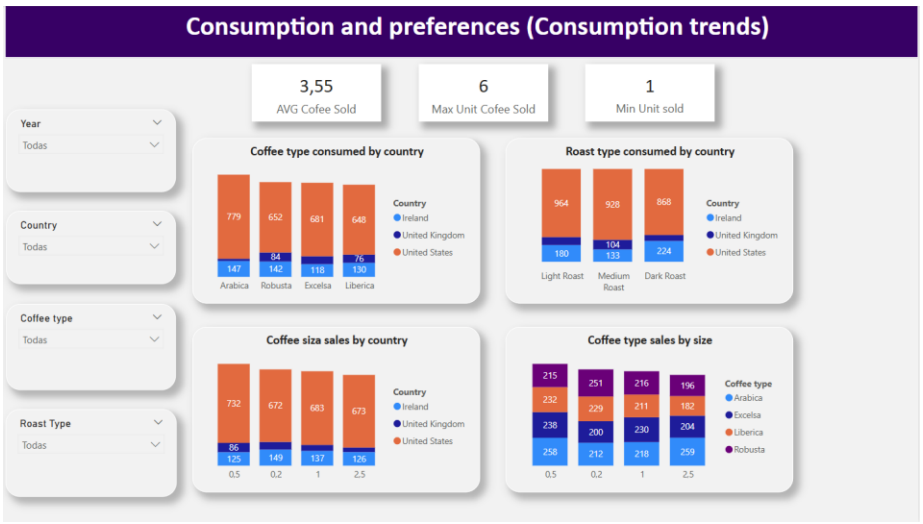
Dashboards

Consumption trends

Excel version

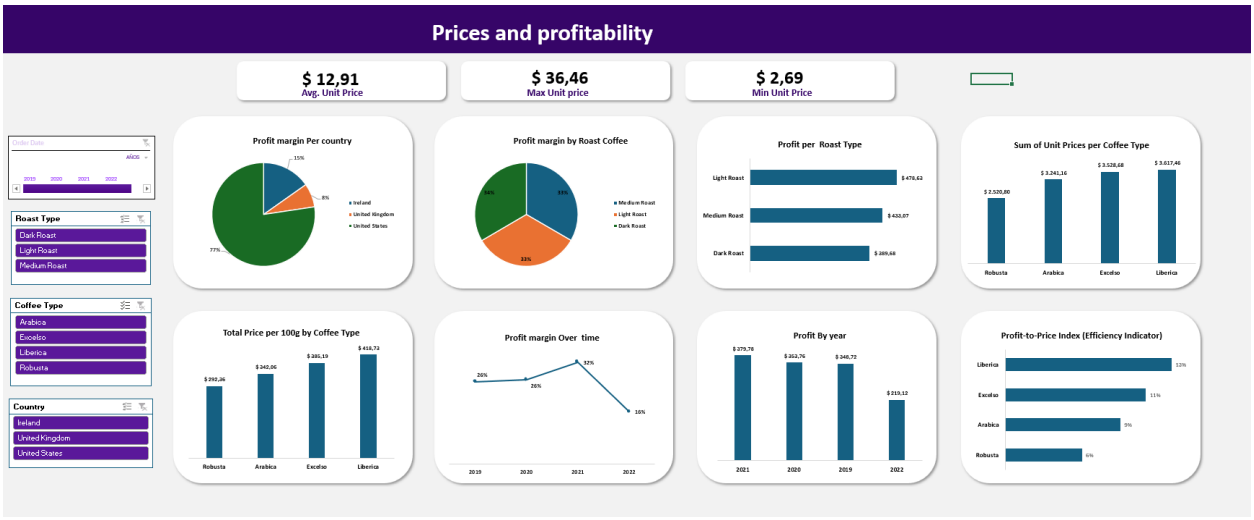


Power bi versión

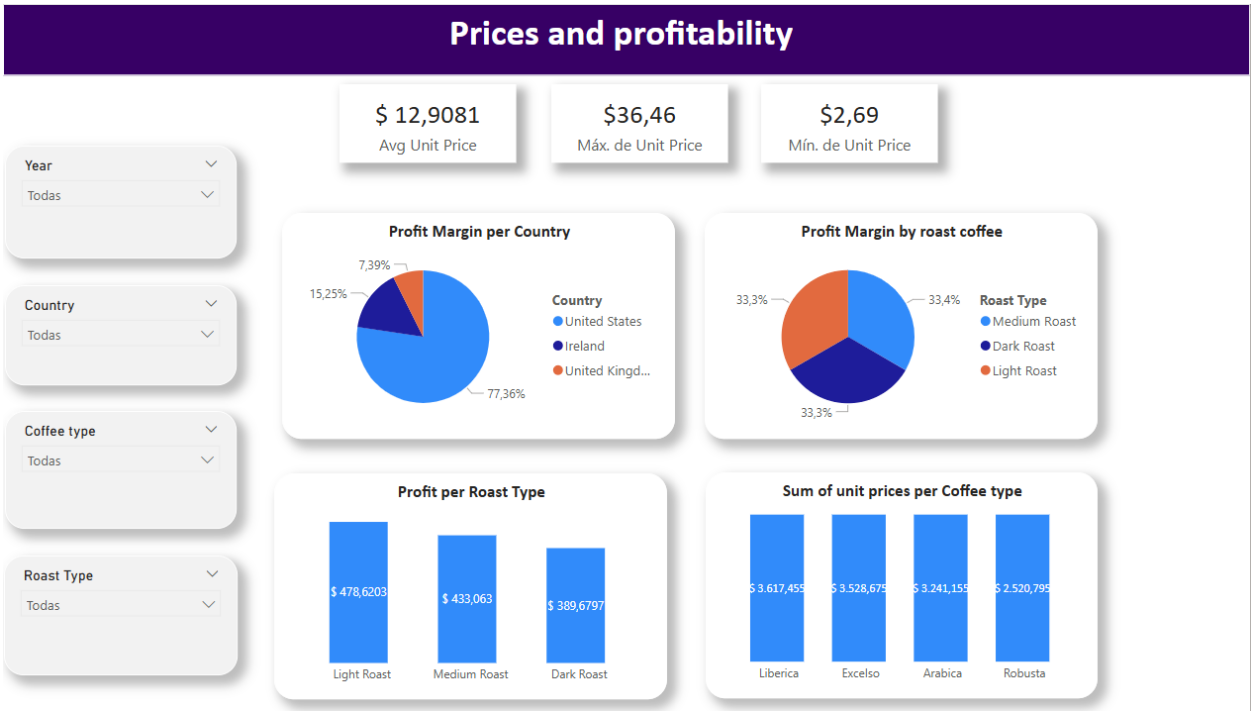


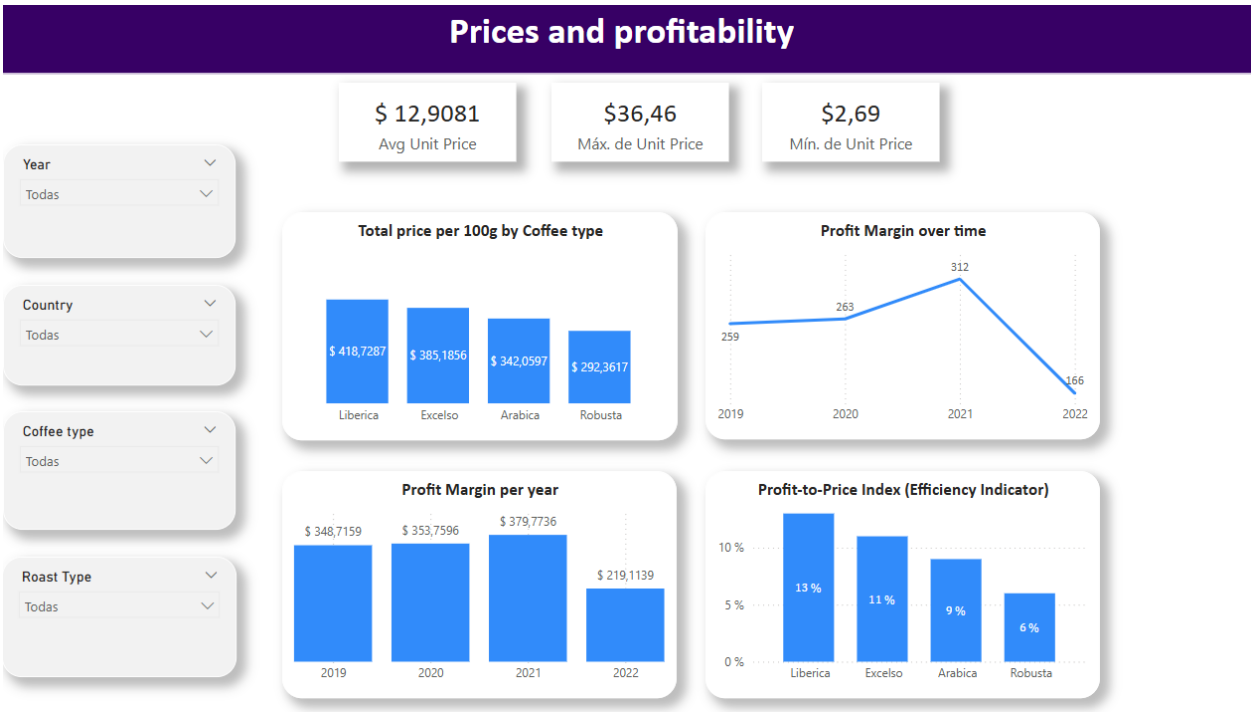
Prices and profitability

Excel version



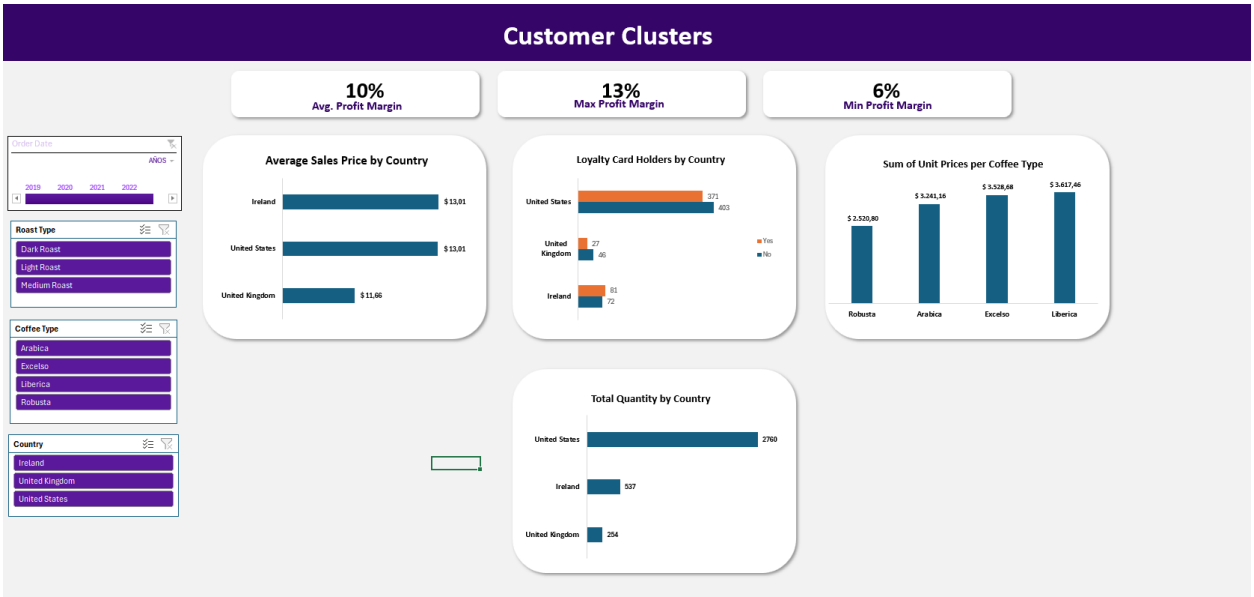
Power BI versión



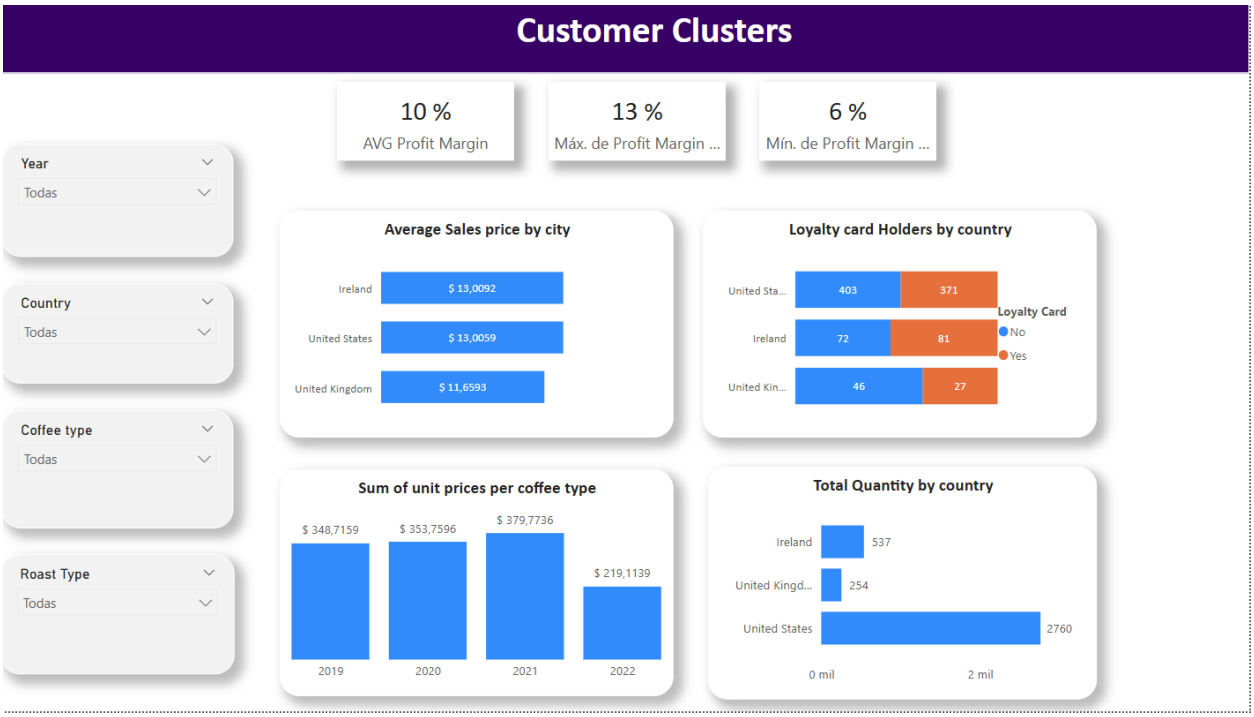


Customer Clusters

Excel Version



Power Bi versión



Findings:

Regarding consumption trends, it was found that, on average, customers buy 3.55 grams of coffee per order, with a median of 4g and a mode of 2g, which suggests a slight tendency toward smaller quantities. The distribution shows low dispersion, with a range of 5g, an IQR of 3g, and a standard deviation of 1.68, indicating that most sales are concentrated in these smaller amounts. The skewness close to zero (0.01) confirms a symmetric distribution, while the negative kurtosis (-1.24) shows a flatter shape than normal, with few extreme values and no significant outliers, which suggests that both small and large orders have a similar frequency.

The dashboard confirms these findings, showing that most orders are between 2 and 5 grams, which supports the consistency in customer behavior.

In terms of consumer preferences, the analysis shows that:

The best-selling coffee type is Arabica, with 947 units, and 779 of them were sold in the United States.

The most popular roast type is Light Roast, with 964 units, being the favorite in all cities analyzed, especially in the United States.

Regarding size, the 0.5 format is the most chosen, with 943 units, which matches the quantity distribution observed in the EDA (centered around 4g).

Also, the combination of coffee type and preferred size shows more specific consumption patterns:

Robusta is preferred in size 0.2.

Arabica in 2.5.

Excelso in 0.5.

Liberica also in 0.5.

These results suggest a segmentation of preferences, which may be influenced by customer profiles or the product formats offered in each region.

On the other hand, it is important to highlight that the most profitable coffee was Liberica, reaching a total global revenue of $470,269, showing its high perceived value, even though it was not the most sold by quantity. Additionally, the average profit per order was identified as $1.301, which is a significant value to consider when evaluating unit performance

Prices and profitability

In the second step, talking about Prices and profitability it is pertinent to say the average price was $12.91, with a minimum of $2.69 and a maximum of $36.46. The most frequent price was $5.97. This, along with the wide range of $33.77 and a standard deviation of $9.72, shows a high variation in market prices. This suggests a quite varied price structure.

Regarding absolute profit, the average value was $1.30, with a median of $0.98 and a mode of $0.61, which means that most profits were on the lower side. The maximum profit was $4.74 and the minimum was $0.16, making a range of $4.58. This high variability is also supported by a standard deviation of $1.13. The distribution shows a positive skew (1.43), meaning that most orders had low profits, but a few had very high profits. The kurtosis of 1.27 also shows this, indicating a leptokurtic shape, with more extreme values than in a normal distribution. In summary, most orders had low profit, but there were some outliers that increased the average.

In terms of profit margin, the average was 10%, with a median of 9% and a mode of 13%. The range was smaller (7%), from a minimum of 6% to a maximum of 13%, which means less variation compared to absolute profit. The standard deviation was 2.56%, and the slight negative skew (-0.24) shows a small tendency towards higher margins. The kurtosis of -1.22 indicates a platykurtic distribution, meaning there were fewer extreme values than normal. This

suggests that, even though the margins don't vary much, they tend to be concentrated around medium to high values.

Looking at profit margin by country, the United States leads with 77% of the total margin, followed by the United Kingdom (15%) and Ireland (8%). This is probably because of higher sales volumes and maybe higher unit prices in the U.S. market.

When looking at profit margin by roast type, the results are very balanced: Dark Roast has 34%, and Medium and Light Roast both have 33%. However, this percentage doesn't show the absolute profit. In that case, Light Roast stands out with a total profit of $478.63, higher than Medium Roast ($433.07) and Dark Roast ($389.68). This means that even if margins are similar, the sales volume or price favored Light Roast.

Looking at prices by coffee type, Liberica has the highest total unit price ($3,617.46) and the highest price per 100g ($418.73), followed by Excelsa, Arabica, and Robusta. This also translates to higher profitability.

From a time perspective, the most profitable year was 2021, with a 32% margin, while in 2022 there was a big drop to 16%. This matches the fall in absolute profit that year, which was $219.12, compared to $379.78 in 2021. This drop could be because of higher costs, fewer sales, or lower prices.

Finally, the Profit-to-Price Index (Efficiency Indicator) helps us understand how profitable each coffee type is for every dollar invested. In this index, Liberica also leads with 13%, followed by Excelsa (11%), Arabica (9%), and Robusta (6%). This confirms that Liberica not only has the highest prices but also the best relative profitability.

Customer Clusters

Regarding the customer cluster, we can observe the following key findings:

The average sales price is $13.01 in both the United States and Ireland, while the United Kingdom shows a lower average of $11.66.

When it comes to Loyalty Card ownership, the country with the highest number of cardholders is the United States with a total of 371, followed by Ireland (81) and the United Kingdom (27). This pattern suggests that customer loyalty is more established in the U.S. It is important to encourage the use or ownership of these cards through promotions or even offering them as gifts for coffee purchases.

By summing the unit prices by coffee type, we can see that Liberica has the highest total ($3,617.46), followed by Excelso, Arabica, and Robusta. This indicates that Liberica holds a

stronger position in the market and is preferred by customers who are willing to pay more to consume it.

Finally, in terms of Total Quantity Sold by Country, the United States clearly leads with 2,760 units sold, far surpassing Ireland (537) and the United Kingdom (254). This makes the U.S. the most important market not only in terms of sales but also in profitability.

Insights and recommendations

Product diversification through key variable combinations

It is important to continue developing a wide variety of coffee products, considering combinations between coffee type, roast level, and package size. Strategic diversification across these three aspects can attract different customer profiles and improve the perceived value and personalization of the offer.

Promotions by volume and use of loyalty cards

It is recommended to launch promotions for purchases over 0.5 grams, encouraging the consumption of larger coffee sizes. Additionally, loyalty cards can be given with the purchase of small coffee sizes or used to offer free coffee, promoting repeated purchases and increasing the average ticket value.

Increase in Arabica with Light Roast offer

Since Arabica coffee and Light Roast are the most consumed types, it is suggested to expand the availability of this combination. This strategy responds directly to current demand and helps strengthen an offer that matches customer preferences.

Boost Liberica coffee consumption with new formats

It is advisable to promote Liberica coffee consumption by offering it in new formats or products, such as capsules, cold brew, or limited editions. This can help take advantage of its position as the coffee with the highest total unit price and a strong premium image.

Need for more consumer knowledge

It is a priority to carry out more market research to gather detailed information about customer profiles and their preferences in coffee formats. This will help to create more targeted and effective offers.

Cross-selling strategies with other products

It is suggested to develop cross-promotions that combine small-size coffee with other products (such as cookies, desserts, or other snacks). This will not only increase sales volume but also improve the overall customer experience.

Customized strategies by city and season

Campaigns should be tailored according to the city, coffee type, roast level, and time of year. Special focus should be given to Ireland and the United Kingdom, where buying behavior and customer loyalty differ from the U.S. market.

Customer loyalty through programs and benefits

It is essential to strengthen customer loyalty through programs like loyalty cards, exclusive promotions, free delivery, or rewards for buying a certain quantity of coffee. These actions help build long-term relationships and encourage repeat purchases.

Improving marketing in the United Kingdom

It is recommended to evaluate and redesign current marketing strategies in the UK, where loyalty card ownership and sales volume are lower. The brand presence should be reinforced, promotions improved, and a stronger emotional connection built with local customers.

Storytelling

Coffee as a Lifestyle… and as a Data Opportunity

Coffee is not just a drink. For many people, including myself, it's a daily ritual — a silent partner that marks the start of the day. It's more than just caffeine: it's purpose, clarity, and connection. Every morning, millions of people turn on their coffee machines not just to wake up, but to begin with intention.

And it was exactly that love for coffee that sparked my curiosity:
What is the most profitable type of coffee?
What roast do people prefer?
Which countries drink the most coffee?
And how can this data help us sell better?

To answer these questions, I analyzed a dataset from Kaggle with information about 1,000 coffee orders made between January 2019 and August 2022 in three countries: the United States, Ireland, and the United Kingdom. The dataset included details like:

- Coffee type: Arabica, Robusta, and Liberica
- Roast type: Light, Medium, and Dark
- Packaging format and size
- Unit price and profit
- Sales volume by country

The unit price ranged between $2.69 and $36.46, with an average of $12.41.
The average profit per unit was $1.30, with better results in lower-priced products.

The **profit-to-price index** showed that **Liberica** was the most efficient coffee: it gave a 13% return per dollar sold (854 units, $3,617.46 in total profit).

In comparison:

- **Arabica** was the most sold (947 units), but less profitable.
- **Robusta** was the cheapest, but gave only 6% return.

The United States showed more repeat customers, which suggests better customer loyalty.

Ireland and the UK had fewer sales and less loyalty — making them good markets to target with campaigns or loyalty strategies.

The most requested roast was **Light Roast**, with 1,230 units sold.
The most popular size was **small (0.5)** — likely for price or convenience.

The year **2021** had the highest profits, with $379.78 (32% of the total).
In **2022**, sales dropped, showing the need for new business strategies.

This analysis helped me learn key things:

- The most popular product is not always the most profitable — Liberica proved that.
- Small formats are preferred, so faster product turnover is a good strategy.
- Ireland and the UK show real potential to increase customer loyalty.
- We can explore bundles (like coffee + snacks), loyalty programs (like cards or rewards), and better pricing by type and region.

More than just a data project, this was a way to connect two passions: my love for coffee and my interest in data. Behind every cup, there are habits, choices, and opportunities waiting to be discovered.

And maybe, with more data — and more coffee — we'll keep learning not just what people drink, but *why* they drink it.