Coffee Consumption Analysis

Yash Mittal

2025-11-24

1. Background and Problem Definition

Dataset Source: The dataset used in this project describes domestic coffee consumption by country from 1990 to 2020. It is a CSV file named `Coffee_domestic_consumption.csv`, that can be found on Kaggle.

Problem Statement: Coffee is one of the most traded commodities in the world. However, consumption patterns shift over time due to economic development, population growth, and changing cultural habits.

The goal of this project is to analyze the historical data to answer the following questions:
1. **Global Trends:** Is global domestic coffee consumption increasing, and if so, at what rate?
2. **Top Consumers:** Which countries are the largest consumers of domestic coffee in the modern era?
3. **Coffee Types:** Is there a distinction in consumption volume based on the type of coffee (Arabica vs. Robusta) produced/consumed in those regions?

To answer these questions, we will perform data wrangling to reshape the data, use visualization to identify trends, and apply a simple linear regression model to quantify growth.

2. Importing and Loading Libraries

We will use the `tidyverse` suite of packages for data manipulation and visualization.

```r
library(tidyverse)
library(knitr)
library(ggplot2)

# Load the dataset
# Note: Ensure the csv file is in the same directory as this Rmd file
coffee_raw <- read_csv("Coffee_domestic_consumption.csv")
```

3. Data Wrangling and Cleaning

First, let's inspect the raw data structure.

```r
# Quick overview of the first few rows and columns
head(coffee_raw[, 1:6])
```

```
## # A tibble: 6 x 6
##   Country                  ‘Coffee type‘ ‘1990/91‘ ‘1991/92‘ ‘1992/93‘ ‘1993/94‘
##   <chr>                    <chr>             <dbl>     <dbl>     <dbl>     <dbl>
## 1 Angola                   Robusta/Arab~   1200000   1800000   2100000   1200000
## 2 Bolivia (Plurinational ~ Arabica         1500000   1620000   1650000   1710000
## 3 Brazil                    Arabica/Robu~ 492000000 510000000 534000000 546000000
## 4 Burundi                   Arabica/Robu~    120000     96000    102000    114600
## 5 Ecuador                   Arabica/Robu~  21000000  21000000  21000000  21000000
## 6 Indonesia                 Robusta/Arab~  74520000  76800000  79140000  81540000
```

```r
dim(coffee_raw)
```

```
## [1] 55 33
```

Observation: The dataset is currently in a **"Wide" format**. The years (e.g., "1990/91", "1991/92") are spread across columns. This format is difficult to use for time-series analysis in R.

Steps for Cleaning:
1. **Pivot Longer:** We need to transform the data from wide to long format so that "Year" becomes a single variable.
2. **String Manipulation:** The years are formatted as "1990/91". We will extract the first four digits to convert this into a numeric `Year` column.
3. **Data Typing:** Ensure consumption numbers are numeric.

```r
# Reshape from Wide to Long format
coffee_clean <- coffee_raw %>%
  # We remove the pre-calculated 'Total_domestic_consumption' column
  # to avoid double counting during our own aggregation.
  select(-Total_domestic_consumption) %>%
  pivot_longer(
    cols = -c(Country, `Coffee type`),
    names_to = "Year_Raw",
    values_to = "Consumption"
  )

# Extract numeric year from the "Year_Raw" string (e.g., "1990/91" -> 1990)
coffee_clean <- coffee_clean %>%
  mutate(Year = as.numeric(substr(Year_Raw, 1, 4)))

# Check for missing values
sum(is.na(coffee_clean$Consumption))
```

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

```r
# Display the cleaned data structure
head(coffee_clean)
```

```
## # A tibble: 6 x 5
##   Country ‘Coffee type‘   Year_Raw Consumption  Year
##   <chr>   <chr>           <chr>          <dbl> <dbl>
## 1 Angola  Robusta/Arabica 1990/91      1200000  1990
## 2 Angola  Robusta/Arabica 1991/92      1800000  1991
## 3 Angola  Robusta/Arabica 1992/93      2100000  1992
## 4 Angola  Robusta/Arabica 1993/94      1200000  1993
## 5 Angola  Robusta/Arabica 1994/95      1500000  1994
## 6 Angola  Robusta/Arabica 1995/96       600000  1995
```

Now the data is tidy: every row represents a specific country's consumption in a specific year.

4. Exploratory Data Analysis (EDA)

4.1 Global Consumption Over Time

Let's aggregate the data to see the total world consumption per year.

```
global_yearly <- coffee_clean %>%
  group_by(Year) %>%
  summarise(Total_Consumption = sum(Consumption, na.rm = TRUE))

# Summary statistics of global yearly consumption
summary(global_yearly$Total_Consumption)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 1.171e+09 1.469e+09 1.937e+09 2.040e+09 2.657e+09 3.015e+09
```

4.2 Top Consuming Countries (2019)

We want to see who the current heavy hitters are. We will filter for the most recent complete data year (2019) and rank the countries.

```
top_consumers_2019 <- coffee_clean %>%
  filter(Year == 2019) %>%
  arrange(desc(Consumption)) %>%
  head(10)

kable(top_consumers_2019, caption = "Top 10 Coffee Consuming Countries in 2019")
```

Table 1: Top 10 Coffee Consuming Countries in 2019

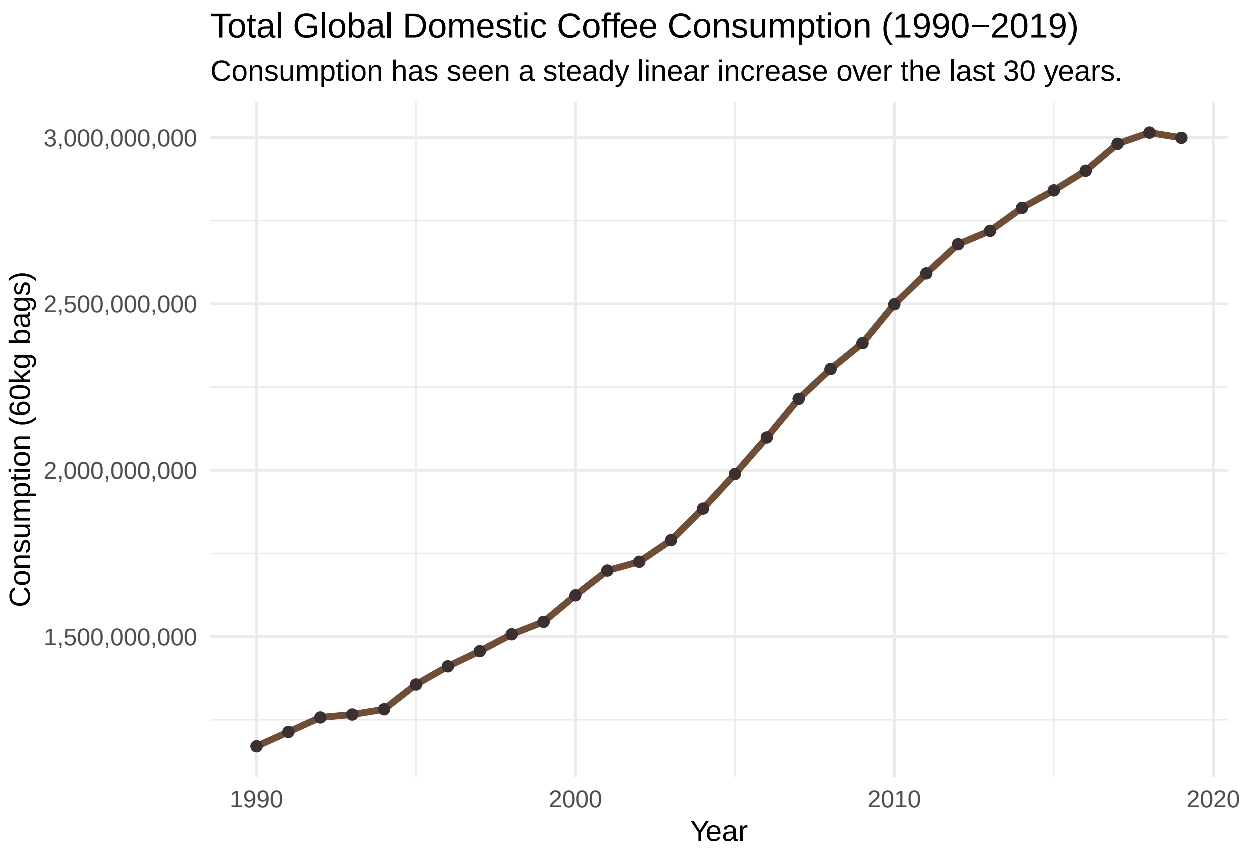
Country	Coffee type	Year_Raw	Consumption	Year
Brazil	Arabica/Robusta	2019/20	1320000000	2019
Indonesia	Robusta/Arabica	2019/20	288360000	2019
Ethiopia	Arabica	2019/20	226860000	2019
Philippines	Robusta/Arabica	2019/20	195000000	2019
Viet Nam	Robusta/Arabica	2019/20	159000000	2019
Mexico	Arabica/Robusta	2019/20	145500000	2019
Colombia	Arabica	2019/20	121486440	2019
India	Robusta/Arabica	2019/20	87000000	2019
Thailand	Robusta/Arabica	2019/20	84000000	2019
Venezuela	Arabica	2019/20	76500000	2019

5. Data Visualization

5.1 Visualizing Global Growth

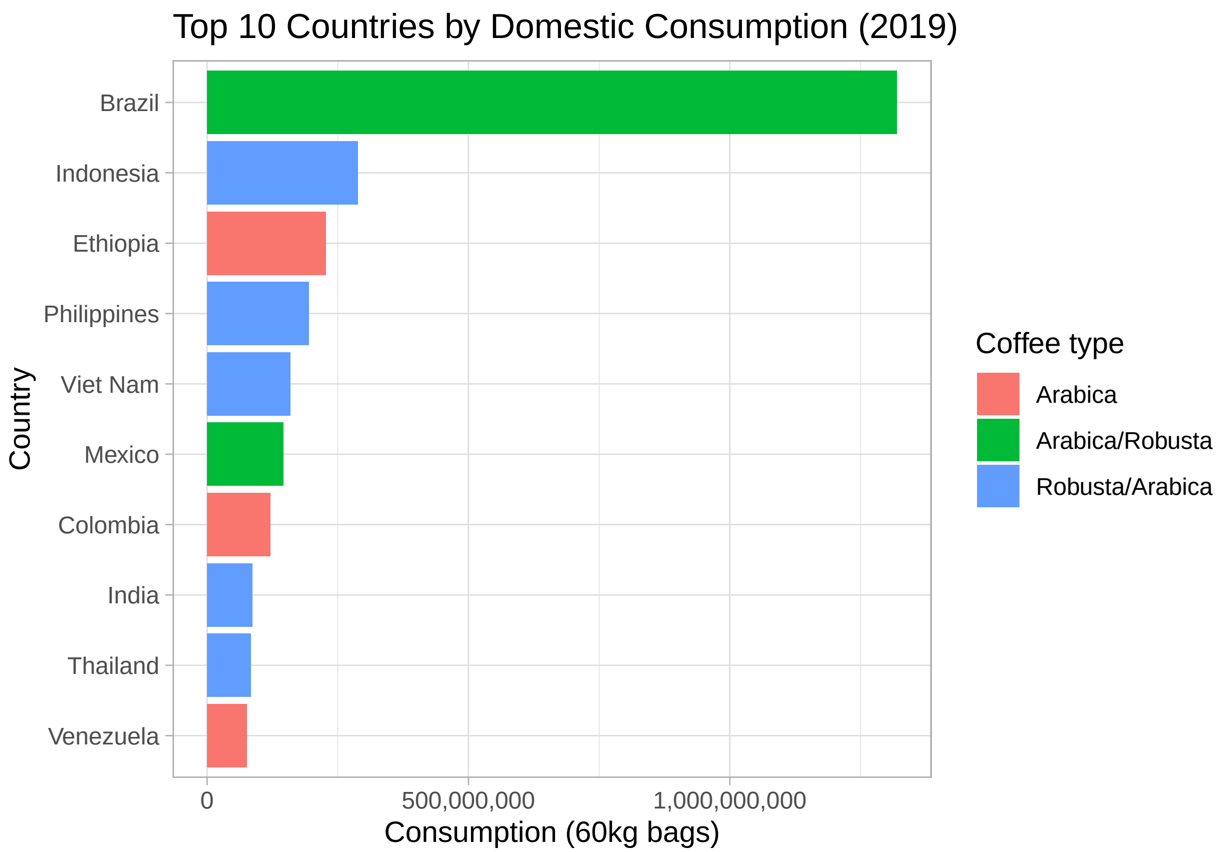
```
ggplot(global_yearly, aes(x = Year, y = Total_Consumption)) +
  geom_line(color = "#6f4e37", size = 1.2) + # Coffee brown color
  geom_point(color = "#3b2f2f") +
```

```r
  labs(
    title = "Total Global Domestic Coffee Consumption (1990-2019)",
    subtitle = "Consumption has seen a steady linear increase over the last 30 years.",
    y = "Consumption (60kg bags)",
    x = "Year"
  ) +
  theme_minimal() +
  scale_y_continuous(labels = scales::comma)
```



5.2 Top 10 Consumers Bar Chart

```r
ggplot(top_consumers_2019, aes(x = reorder(Country, Consumption),
                               y = Consumption, fill = `Coffee type`)) +
  geom_bar(stat = "identity") +
  coord_flip() + # Make it horizontal for readability
  labs(
    title = "Top 10 Countries by Domestic Consumption (2019)",
    x = "Country",
    y = "Consumption (60kg bags)"
  ) +
  theme_light() +
  scale_y_continuous(labels = scales::comma)
```



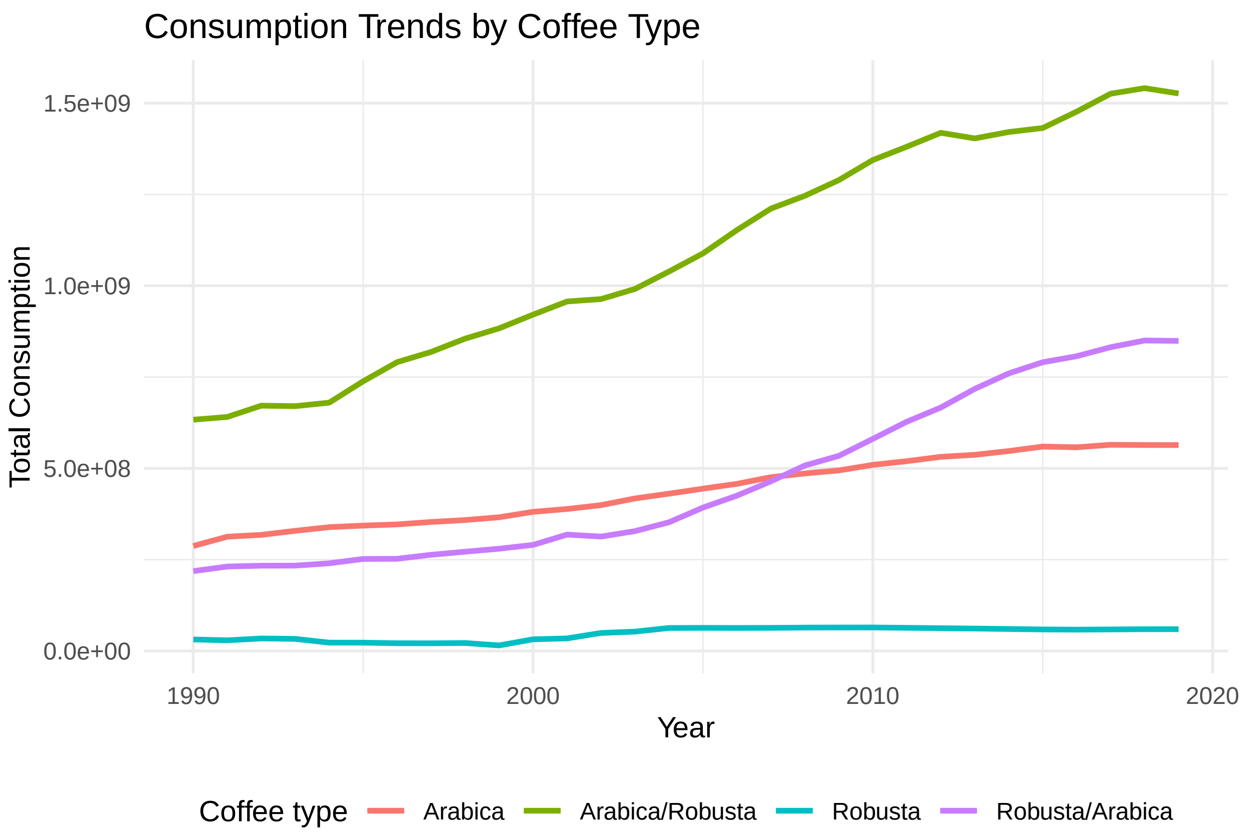
Analysis: Brazil is, by a massive margin, the largest domestic consumer of coffee in this dataset, followed by Indonesia and the USA (if included, though this dataset appears to focus on producing countries mostly, or specifically captures domestic consumption of produced stock).

5.3 Consumption by Coffee Type

Let's see if Arabica or Robusta drives the market.

```r
type_trends <- coffee_clean %>%
  group_by(Year, `Coffee type`) %>%
  summarise(Total = sum(Consumption), .groups = 'drop')

ggplot(type_trends, aes(x = Year, y = Total, color = `Coffee type`)) +
  geom_line(size = 1) +
  labs(
    title = "Consumption Trends by Coffee Type",
    y = "Total Consumption",
    x = "Year"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```



6. Statistical Analysis: Simple Linear Regression

We observed a clear upward trend in the global consumption line chart. We will perform a simple linear regression to quantify this growth rate.

Model: $Consumption = \beta_0 + \beta_1(Year) + \epsilon$

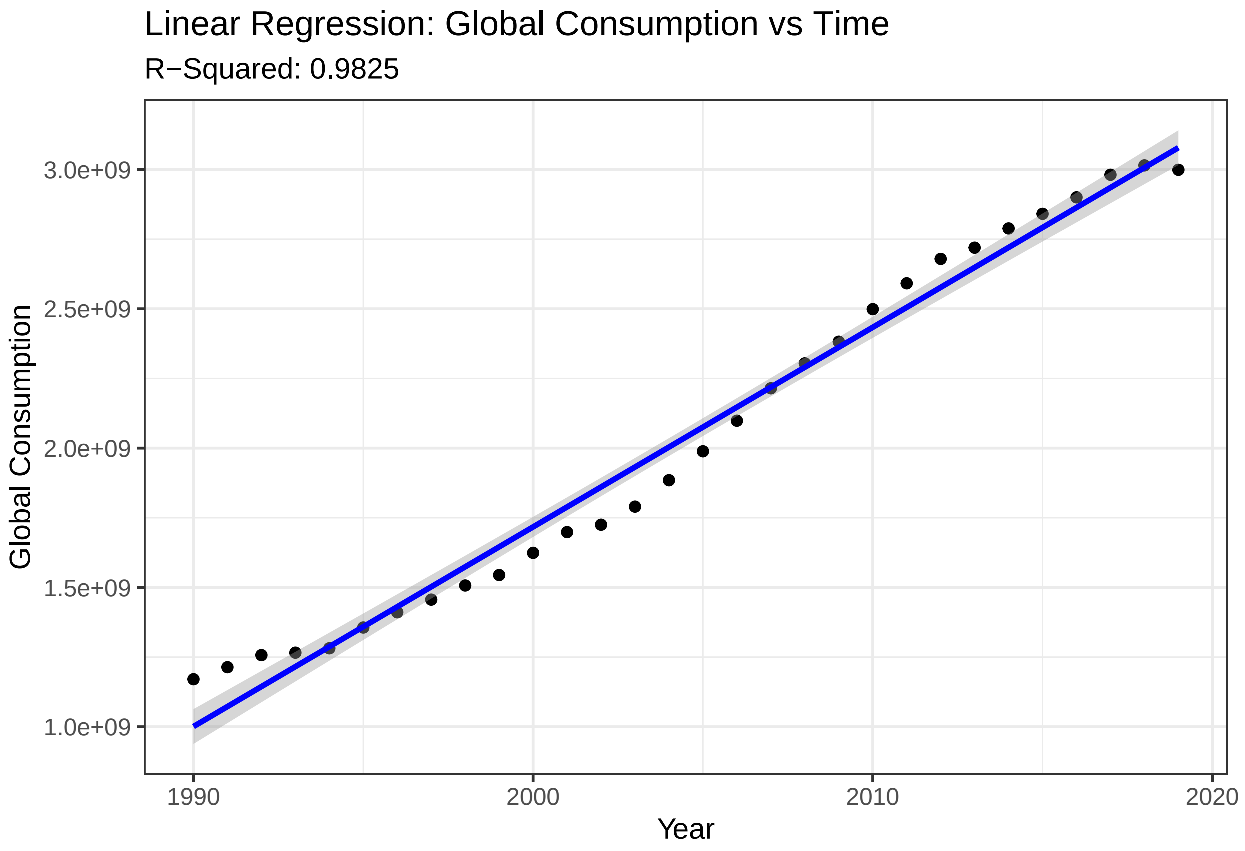
```
# Create linear model
lm_model <- lm(Total_Consumption ~ Year, data = global_yearly)

# Display model summary
summary(lm_model)
```

```
##
## Call:
## lm(formula = Total_Consumption ~ Year, data = global_yearly)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -142138769  -76307588    2578759   61438813  169767908
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.416e+11  3.622e+09  -39.09   <2e-16 ***
## Year         7.164e+07  1.807e+06   39.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 85650000 on 28 degrees of freedom
## Multiple R-squared:  0.9825, Adjusted R-squared:  0.9819
## F-statistic:  1572 on 1 and 28 DF,  p-value: < 2.2e-16
```

Interpretation of Results: Looking at the `Year` coefficient from the summary above: 1. **Slope (Estimate):** The coefficient for Year tells us the average annual increase in coffee bags consumed. 2. **P-value:** If the value is `< 2e-16` (extremely small), it indicates the relationship between time and consumption is statistically significant. 3. **R-squared:** A high R-squared value indicates that the year explains a large portion of the variance in consumption.

```r
ggplot(global_yearly, aes(x = Year, y = Total_Consumption)) +
  geom_point() +
  geom_smooth(method = "lm", col = "blue") +
  labs(
    title = "Linear Regression: Global Consumption vs Time",
    subtitle = paste("R-Squared:", round(summary(lm_model)$r.squared, 4)),
    x = "Year",
    y = "Global Consumption"
  ) +
  theme_bw()
```



7. Conclusion

In this project, we analyzed domestic coffee consumption over a 30-year period.

Findings:
1. **Data Wrangling:** We successfully transformed the dataset from a wide format to a long format, enabling time-series analysis.
2. **Major Players: Brazil** is the dominant consumer in this dataset, consuming significantly more than the next closest countries (Indonesia and Ethiopia).
3. **Trends:** There is a strong, statistically significant positive correlation between time and coffee consumption. The global market has been growing consistently since 1990 without major downturns in this specific dataset.
4. **Types:** The “Arabica/Robusta” mixed category (dominated visually by Brazil’s classification) represents the highest volume, while pure Robusta and Arabica follow similar, though lower, growth curves.

This analysis suggests that the domestic markets of coffee-producing nations are robust and expanding, driven largely by Brazil’s massive internal market.