MATH36031 Project 3

Student ID: 11126411

Introduction

The global banana trade plays an important role in the agricultural economies for a lot of regions such as Ecuador, Philippines, Costa Rica, Guatemala and Netherlands (the top 5 largest banana exporters in the world) [1] Understanding the pricing patterns of bananas across all these regions provides lots of insights on market dynamics and economic interdependencies. This project uses MATLAB to analyse a larget dataset *bananas.csv* that contains the prices of bananas of different origin, answering several specific questions by computing average prices, boxplots, time series plots and correlation coefficients. Note that there is an entry with the country of origin as *'all_bananas'* which is an average of the collective data for that date.

Motivation

This report would be extremely important for stakeholders such as farmers, traders and policymakers as it highlights the pricing dynamics across regions, investigate seasonal trends that may affect pricing, and identify interdependencies between regions to support decision making.

Output and Analysis of Data

There are different prices of bananas on different dates with different prices, so there will be duplicated of origins. A list of 28 distinct origins is produced. Identifying these distinct origins will help to ensure that all geographic and economic factors influencing the banana prices are considered.

Banana prices fluctuate a lot during the year. After computing the average prices of bananas of different origins, the results are as followed:

Origin	Average Price (£)
acp_bananas	0.6413
all_bananas	0.7702
belize	0.7352
brazil	0.6306
cameroon	0.6698
colombia	0.7519
costa_rica	0.7633
dollar_bananas	0.7086
dominican_republic	0.6402
ecuador	0.7109
eu	0.8600
eu_bananas	0.5326
ghana	0.6399
guadeloupe	0.5616
guatemala	0.7263
honduras	0.7002
ivory_coast	0.6592
jamaica	0.6006
malaysia	0.5933
martinique	0.5736
mexico	0.6457
nicaragua	0.7223
panama	0.7811
somalia	0.7400
st_vincent	0.4415
surinam	0.5927
venezuela	0.5209
windward_isles	0.5460

Table 1: Average Prices of Bananas of different Origin

Average Prices

From the table of average price above, the origin with the lowest average price is St. Vincent with a price of £0.4415 and the origin with the highest average price is the EU with a price of £0.86. The low average may be due to high subsidies from the government or lower production costs due to effective strategies, while a higher price may be caused by imposed tariffs or transportation expenses.

Boxplot

Boxplots are produced for comparison of prices between selected regions: *'colombia'*, *'costa_rica'*, *'honduras'*, *'windward_isles'* and *'all_bananas'* using boxplots. Note that *'all_bananas'* is the average of the collective data for that date.

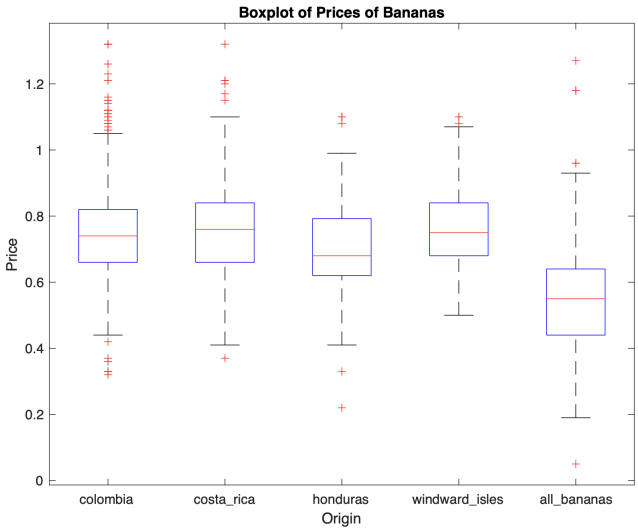

Figure 1: Boxplot of Prices of Bananas with Selected Countries of Origin

A boxplot consists of the median, spread(IQR), outliers, minimum and maximum of a data. [2] It can be used to identify trends and patterns when multiple boxplots plotted side by side. The four origins *'colombia'*, *'costa_rica'*, *'honduras'* and *'windward_isles'* have more consistent pricing patterns, with medians (red line) around 0.8, similar spread (blue box) sizes and have a slight skew towards higher prices, which is reflected by the upper whiskers that extended further than the lower whiskers.

Whereas the *'all_bananas'* origin have a lower median 0.6, which is expected since it reflects the aggregated banana prices and averages out the extreme values present in each of the origins. This is also backed up by the fewer outliers that are present as the averaging reduces the chance of the outliers influencing the distribution of the prices. Its smaller spread and shorter whiskers suggests that the price fluctuations across all bananas are more stable compared to the individual origins.

Overall, the *'all_bananas'* boxplot serves as a useful benchmark for stakeholders to assess the individual origins and identify which origins are contributing most to price stability or volatility. If any of the features of the individual boxplots differs significantly from *'all_bananas'*, it could indicate extreme or unreasonable pricing of banana for that origin on certain dates. For example, *'honduras'* and *'windward_isles'* show larger variability which indicates that markets in these origins may face more volatile conditions or diverse pricing factors. [3]

Time Series Plot

Next, the time series plot of the variation of the prices of bananas with countries of origin *'columbia'* is computed to look for seasonal trends or patterns.

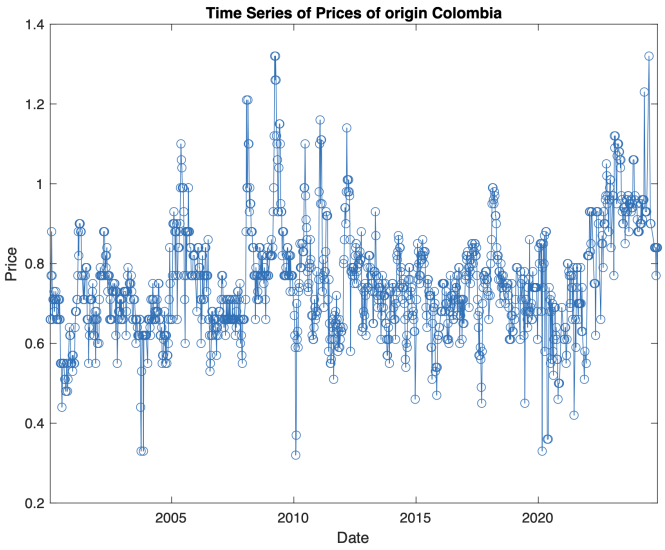

Figure 2: Time Series plot of the Prices of Bananas in *'columbia'*

A time series plot is a graph used to display data points over time to visualise trends and fluctuations in the data. Observed trends are useful and can be used to forecast future prices. [4]

From the plot above, patterns can be identified according to different time periods: During years 2005-2010, prices were relatively stable but showed occasional spikes; During years 2019-2020, there are more accented peaks with really high banana prices; After year 2020, the fluctuation in prices becomes more obvious.

There is also an obvious trend of high volatility in the short-term of some periods due to frequent spikes and drops. This reflects the price sensitivity to factors such as seasonal variations, market demand, or supply chain disruptions.

Overall, despite the frequent ups and downs in banana prices, it still looks like there is an upward trend in the average prices, which may be due to inflation and economic growth in Columbia, especially in recent years.

Correlation Coefficient

Lastly, the correlation between the fluctuation of *'columbia'* and *'ecuador'* prices is calculated using the **corrcoef** function in MATLAB, with a weak correlation of

0.1962. This highlights the independent pricing mechanisms between two regions, which is likely influenced by market demands and export volume. Columbia and Ecuador may operate in different markets, causing price changes to be driven by unique market conditions.

Also, Ecuador is one of the largest exporters of bananas globally, holding 29% of the world exports [1], hence its prices will be more influenced by global demand. Whereas Columbia only accounts for 1.83% of the world exports, so its prices may only affected by a few markets.

Method and Code

Listing 1

Firstly, the *bananas.csv* data is loaded into MATLAB by specifying the path of the file from my device (**Line 1**). In order to extract the distinct entries under the Origin header, the **unique** function is used to return the unique values in the header, removing duplicates (**Line 2**).

Listing 3

The for loop in Listing 3 iterates over all unique origins previously found. For each unique origin, it identify the rows that corresponds to each of them using the **strcomp** function. The function **strcomp** returns a logical array where **1**: the row matches the i-th origin; and **0**: the row does not match the i-th origin. The rows corresponding to **1** will then be filtered from the *banana.csv* dataset to calculate the average price using the **mean** function (**Line 1-4**). The lowest and highest average prices are then computed using **min()** and **max()** respectively (**Line 6-7**).

Listing 5

The origins to be investigated for the boxplot are *'colombia'*, *'costa_rica'*, *'honduras'*, *'windward_isles'* and *'all_bananas'*. These origins are filtered from the original *bananas.csv* dataset using the **ismember** function (**Line 2**), which checks each entry in the Origin column and returns **1** if the entries in the Origin column is one of the five origins to be investigated and **0** otherwise. The rows corresponding to the value **1** is then moved to a new dataset called *filteredbanana* (**Line 3**), which is then used to create a boxplot by plotting the prices against its origin (**Line 5-9**).

Listing 6

Similar to Listing 3, Listing 6 also uses **strcomp** function to filter the *banana.csv* data to obtain details of the origin *'columbia'* (**Line 1**). The date column in this filtered data is then converted to a **datetime format**, which allows the

date values to be properly handled as dates for accurate plotting (**Line 2**). The rest is then the plotting of the time series plot of the dates against the banana price (**Line 4-7**).

Listing 7

Lastly, in order to compare the fluctuation of prices of the origins Ecuador and Columbia, the filtered Ecuador data is computed the same way as the Columbia data (**Line 1-2**). Then we find the common dates between Columbia and Ecuador prices that are recorded in the dataset and find the corresponding banana prices of the two origins so that the fluctuations can be calculated accurately (**Line 4-7**). The MATLAB function **corrcoef** is used to return a correlation matrix of the fluctuation in banana prices, where the fluctuation is calculated by taking the change in price from one day to the next (**Line 9**). The (1,2) and (2,1) coefficient of the matrix reflects the correlation coefficient between the fluctuation of Columbia and Ecuador prices (**Line 10**).

Conclusion

This report analysed pricing patterns of bananas across 28 unique origins, identified regional trends and highlighted interdependencies. Information gained from this report such as St. Vincent with the lowest average banana price and the weak correlation of the price fluctuations between Ecuador and Columbia shows the complexity of the banana trade. For a more insightful research, external factors like trade policies, weather and transportation costs would be useful in terms of providing a more meaningful and accurate analysis. [3]

References

[1] TrendEconomy (2024), Bananas | Imports and Exports | 2023. Available: https://trendeconomy.com/data/commodity_h2/0803

[2] Newcastle University (date unknown), Reading a Box and Whisker Plot. Available: https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/data-presentation/box-and-whisker-plots.html

[3] OpenAI. (2024). ChatGPT (Dec 10 version) [Large language model]. https://chat.openai.com/chatAvailable:

[4] Minitab (2024), Interpret the key results for Time Series Plot Available: https://support.minitab.com/en-us/minitab/help-and-how-to/graphs/time-series-plot/interpret-the-results/key-results/

Appendix: Code Listings

```
banana = readtable("~/Downloads/bananas.csv"); %Load the csv file into MATLAB
distinct_origin = unique(banana.Origin)
```

Listing 1: List of the distinct Origin entries - Code

```
distinct_origin =

    {'acp_bananas'       }
    {'all_bananas'       }
    {'belize'            }
    {'brazil'            }
    {'cameroon'          }
    {'colombia'          }
    {'costa_rica'        }
    {'dollar_bananas'    }
    {'dominican_republic'}
    {'ecuador'           }
    {'eu'                }
    {'eu_bananas'        }
    {'ghana'             }
    {'guadeloupe'        }
    {'guatemala'         }
    {'honduras'          }
    {'ivory_coast'       }
    {'jamaica'           }
    {'malaysia'          }
    {'martinique'        }
    {'mexico'            }
    {'nicaragua'         }
    {'panama'            }
    {'somalia'           }
    {'st_vincent'        }
    {'surinam'           }
    {'venezuela'         }
    {'windward_isles'    }
```

Listing 2: List of the distinct Origin entries - Output

```
for i = 1:numel(distinct_origin)
    rows = strcmp(banana.Origin, origin{i}); % Filter rows by origin
    AveragePrice(i) = mean(banana.Price(rows));
end

[min_price, min_position] = min(AveragePrice);  % Min price and corresponding
     index
[max_price, max_position] = max(AveragePrice);  % Max price and corresponding
     index

fprintf('The origin with the lowest average price is %s with a price of %.4f\
    n', distinct_origin{min_position}, min_price);
fprintf('The origin with the highest average price is %s with a price of %.4f
    \n', distinct_origin{max_position}, max_price);
```

Listing 3: Minimum and Maximum Average Prices of Bananas - Code

```
The origin with the lowest average price is st_vincent with a price of 0.4415
The origin with the highest average price is eu with a price of 0.8600
```

Listing 4: Minimum and Maximum Average Prices of Bananas - Output

```
five_origins = {'colombia', 'costa_rica', 'honduras', 'windward_isles', '
    all_bananas'};
filter = ismember(banana.Origin, five_origins);
filteredbanana = banana(filter, :);

figure;
boxplot(filteredbanana.Price, filteredbanana.Origin, 'Labels', five_origins);
xlabel('Origin');
ylabel('Price');
title('Boxplot of Prices of Bananas');
```

Listing 5: Boxplot comparing the Variation of the Prices of the Bananas in certain Origins - Code

```
colombiaData = banana(strcmp(banana.Origin, 'colombia'), :);
colombiaData.Date = datetime(colombiaData.Date);

plot(colombiaData.Date, colombiaData.Price, '-o');
xlabel('Date');
ylabel('Price');
title('Time Series of Prices of origin Colombia');
```

Listing 6: Time Series plot showing the Variation of the Prices of the Bananas in 'columbia' - Code

```
ecuadorData = banana(strcmp(banana.Origin, 'ecuador'), :);
ecuadorData.Date = datetime(ecuadorData.Date);

[same_date, colombia_position, ecuador_position] = intersect(colombiaData.
    Date, ecuadorData.Date);

colombiaPrice = colombiaData.Price(colombia_position);
ecuadorPrice = ecuadorData.Price(ecuador_position);

matrix = corrcoef(diff(colombiaPrice), diff(ecuadorPrice));
corr_coef = matrix(1,2)
```

Listing 7: Correlation Coefficient between the Fluctuation of 'columbia' and 'ecuador' Prices - Code

```
corr_coef =

    0.1961
```

Listing 8: Correlation Coefficient between the Fluctuation of 'columbia' and 'ecuador' Prices - Output