Project 3

December 16, 2022

1 General introduction

This project aimed to extract and analyze data from a given CSV table. The following sections will outline five problems related to the data and describe the methods used to solve them. The results of the analysis will also be discussed. The `Code Implementation` section will provide the code used in the analysis and explain them.

2 Question 1

A large CSV file contained 4 variables: `Origin`, `Date`, `Price`, and `Units`. To answer Question 1, which asked to list all entries of the variables `Origin` and `Units` with no repetition, the following steps were taken:
STEP 1 The data in the CSV file was loaded into MATLAB, enabling further processing.
STEP 2 A list of entries that appeared in the variable `Origin` was recorded, with no repetition.
STEP 3 The same process was applied to the variable `Units`.
Figure 1 shows there 28 variables in the variable `Origin`, and Figure 2 shows the only variable in the variable `Units`.

```
>> origin

origin =

  28×1 cell array

    {'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'   }
```

Figure 1: Items in the variable Origin

```
>> units

units =

  1×1 cell array

    {'£/kg'}
```

Figure 2: Items in the variable Units

3 Question 2

Question 2 required to determine the average price of all `Origin` listed in Question 1.
To solve this problem, the price data for each origin presented in figure 1 would be found and calculated the mean price and stored in a table along with the name of the origin.
Figure 3 shows the result of question 2.

```
>> G

G =

  28×3 table

        Origin            GroupCount    mean_Price
    ____________________    __________    __________

    {'acp_bananas'       }       1171       0.62421
    {'all_bananas'       }        646       0.73941
    {'belize'            }        386       0.70425
    {'brazil'            }        445       0.62292
    {'cameroon'          }        603         0.6698
    {'colombia'          }       1085       0.73633
    {'costa_rica'        }       1144       0.74284
    {'dollar_bananas'    }       1249       0.68709
    {'dominican_republic'}        763       0.63701
    {'ecuador'           }       1000       0.69899
    {'eu'                }          2          0.86
    {'eu_bananas'        }        378       0.53265
    {'ghana'             }        102        0.6399
    {'guadeloupe'        }        278       0.56162
    {'guatemala'         }        360       0.66719
    {'honduras'          }        190       0.68532
    {'ivory_coast'       }        406       0.64778
    {'jamaica'           }         99       0.60061
    {'malaysia'          }          3       0.59333
    {'martinique'        }        253       0.57364
    {'mexico'            }        127       0.64567
    {'nicaragua'         }         24       0.60542
    {'panama'            }        651       0.76091
    {'somalia'           }          2          0.74
    {'st_vincent'        }         47       0.44149
    {'surinam'           }        255       0.59271
    {'venezuela'         }        109       0.52092
    {'windward_isles'    }        673       0.54596
```

Figure 3: The mean price and corresponding Origin

3.1 Question 3

Question 3 required extracting the price data for a list of given Origins, including `Colombia`, `Costa Rica`, `Dominican Republic`, `Honduras`, `jamaica`, `windward isles` and `Mexico`, and commenting on the box plot of the extracted data. Steps used to solve the question:

STEP 1 Establish a cell containing the required origins.

STEP 2 Filter the dataset to include only the varieties of bananas with required origins listed in the question description and extract and store the dataset.

STEP 3 Use the box plot function is used to create a box plot visualization of the Price data, grouping the data by the Origin variable.

The boxplots and mean price (Figure 4) of banana prices for the different countries that are required by the question are shown below.

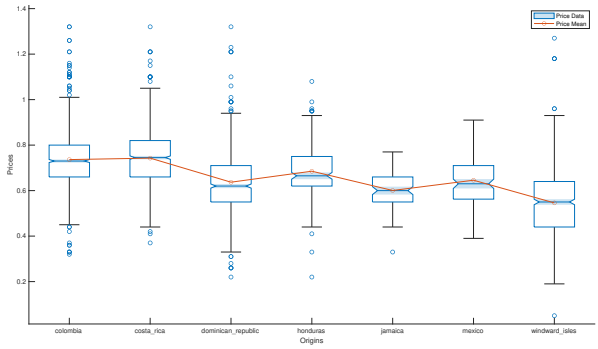

Figure 4: boxplots

Focusing on boxplots, it is possible to observe a box with a blue line in the middle, which represents the median or second quartile(Q2) of the data, larger than 50% of the data. The upper line of the box is the third quartile (Q3), which is larger than 75% of the data, and the lower line of the box is the first quartile (Q1), which is larger than 25% of the data. The box, therefore, contains 50% of the data in total.

From figure 4, it is easy to observe and compare the median price which can indicate the price of the banana at the most time. Costa Rica has the highest median banana price, while Windward Isles has the lowest median

price. Colombia has the second highest price, but the difference between it and Costa Rica is not significaant. The median price of Honduras is lower than that of Colombia, but it is significantly higher than the median prices of the remaining countries. The differences between the Dominican Republic, Mexico, and Jamaica are not significant, but the gap between them and Windward Isles is fairly large. One way to determine whether the differences are significant is by observing whether the notches of the box plots overlap.

In a box plot, the upper and lower whiskers are the vertical lines on either end of the box. The upper whisker represents the maximum value in the data, while the lower whisker represents the minimum value. The formula used to determine the upper and lower whisker is $Q_3 \pm 1.5(IQR)$, where Q_3 is the upper quartile, Q_1 is the lower quartile, and IQR is the interquartile range (the difference between the upper and lower quartiles).

Any data points that fall outside the whiskers are called outliers. Outliers may indicate extreme events or errors in recording, and they can impact the overall analysis if they are not accounted for. It is important to consider the impact of outliers when interpreting the results of an analysis. According to figure 1, Colombia has much more outliers than other countries, while Mexico and Jamaica have no outliers or only one outlier. Besides that, Windward isles and Honduras have a smaller amount of outliers than Costa Rica and Dominican Republic. Overall, having more outliers than others can indicate that a data set is more variable, has a higher level of dispersion, or is skewed in one direction.

The length of the box indicates the concentration of the data, with shorter boxes indicating higher levels of concentration. This can be used to rank the countries by their banana prices in terms of data concentration. From highest to lowest, the rank is Jamaica, Honduras, Colombia, Mexico, Costa Rica and the Dominican Republic (same rank), Windward Isles. Based on the fact that a higher level of data concentration indicates lower variance. the information regarding variance can be also determined.

The location of the median in the box and the relative location of the mean and median can also provide information about the skewness of the data. Jamaica can be regarded as not skewed, Costa Rica and Windward Isles are left-skewed, and the remaining countries have right-skewed banana prices.

3.2 Question 4

In this question, we are asked to perform a time series analysis of the origin of Colombia and to comment on the seasonal trends.

First, a time series is a sequence of observations that are collected at regular intervals over time. Time series analysis is a statistical method that decomposes a time series into its component, including trend, seasonal effect, and residual error. The seasonal effect is the periodic component of the time series.

Two tools used to analyze the periodic properties of a time series are the fast Fourier transform (FFT) and the `trenddecomp` function in Matlab. The FFT is an algorithm that calculates the discrete Fourier transform (DFT) of a time series, which is a mathematical operation that decomposes the time series into its individual frequency components. The FFT can be used to identify the dominant frequency in the time series data, which corresponds to the period of the time series. More information can be found in [1]. The `trenddecomp` function can provide pictures and data of the time series's component, but not the exact periods.

Before applying them, it is important to preprocess the data to ensure that the resulting period estimates are accurate and reliable. Here are the steps that can be followed to prepare the data for time series analysis:

3.2.1 Preparation

STEP 1 Reshape the data of price to have a consistent sampling period.

The given price data has inconsistent sampling intervals, which makes it unable to analyze as a time series. To fix this, we can resample the data to have a regular sampling time, in this question the interval is seven days. For the time points where no data is available, apply the Modified Akima cubic Hermite interpolation (MAKIMA) method to derive the price data. This method produces smooth and natural curves and is more stable and intuitive than other interpolation methods. The complete introduction can be found in the work of

Bonakdari and Zeynoddin (2022)[3].
STEP 2 Identify and process extreme data points, also known as outliers.
Outliers can impact the accuracy and reliability of time series analysis, so it is important to address them. We can use the Generalized extreme studentized deviate (ESDG) test to detect outliers in the data and use the value calculated by MAKIMA interpolation method to replace them. Also, detailed information can be found in the work of Bonakdari and Zeynoddin (2022)[3].
STEP 3 Remove the long-term trend.
Removing the long trend before using the fast Fourier transform (FFT) to find the seasonal trend can improve the accuracy and interpretability of the seasonality estimates, and can make the analysis more efficient and scalable. This is because removing the long trend allows the FFT to focus on the periodic components of the data, rather than being influenced by the overall trend. It can also reduce the size of the data, which can make the FFT calculation more efficient and reduce the amount of memory needed. This is particularly useful when working with large datasets.
STEP 4 Normalize the data.
The reason is that normalizing the data before performing the period analysis can improve the interpretability and accuracy of the results. Normalization scales the data so that it has a mean of 0 and a standard deviation of 1, which can help to stabilize the FFT algorithm and reduce noise in the data. Additionally, normalizing the data can make it easier to compare the results of the FFT across different datasets and to other types of analysis.

3.3 Seasonal trend Analysis

After preprocessing the data, first, we will focus on the FFT.
As mentioned before, the FFT is a mathematical operation that decomposes a time series into its individual frequency components, allowing us to analyze the periodic patterns in the data. In Matlab, we can use the built-in `fft` function to apply the FFT to the data. After applying the FFT, we can extract the positive frequency components of the FFT result and double it to merge the negative part, then take the absolute value of these components. This will give us a sequence of amplitudes that correspond to the positive frequencies in the FFT. Then use the formula $f = (0 : N-1)\frac{fs}{N}$ to calculate the frequency range, where N is the length of the data and fs is the sampling frequency. By taking the reciprocal of the frequency range, a sequence of potential periods can be obtained. The peaks in an Amplitude-Period plot can indicate the presence of periodic components in a time series, but not all of these peaks may necessarily correspond to valid periods. This is because the FFT is sensitive to the length of the time series and the sampling rate of the data, which can affect the frequency resolution of the FFT. To verify the validity of the period estimates obtained from the FFT, we can use the Matlab built-in function `trenddecomp` to get the seasonal trends' data, calculate periods for them and compare them to the result figured by FFT.

3.4 Result

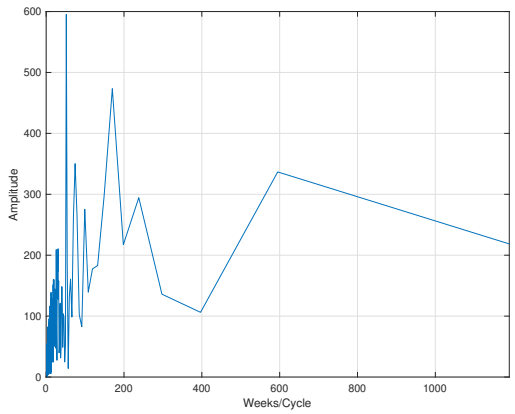

Figure 5: Amplitude-Period plot

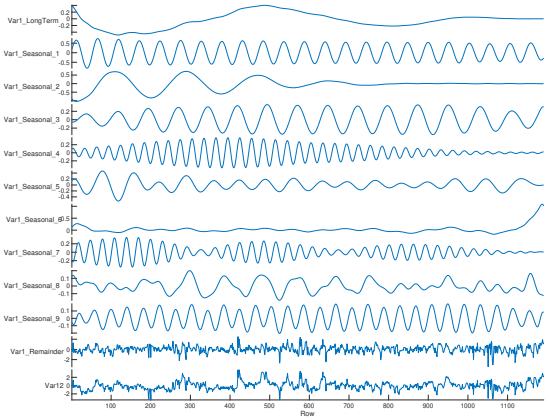

Figure 6: Component of time series

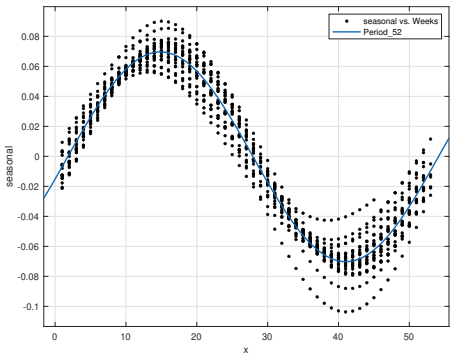

Figure 7: The most significant period

```
>> PE
PE =
  Columns 1 through 4
   51.7391  170.0000   74.3750   29.7500
  Columns 5 through 8
   70.0000  238.0000   31.3158   91.5385
  Column 9
   41.0345
>> k
k =
  logical
   1
```

Figure 8: The period of seasonal trends in figure 5

Figure 5 shows the plot of Amplitude-Period plot, with the peaks indicating possible periods. It is worth noting that the higher the peak, the higher the possibility. Figure 6 illustrates the composition of the time series determined by the Matlab function `trenddecomp`, and Figure 7 illustrates the dominant period (the first value in figure 8) in the time series fitted by the Fourier. The period of the seasonal trend in Figure 5 is listed in Figure 8, where `k=1` means that all periods in the variable `PE` are contained in the period of seasonal trend in figure 6, calculated by the FFT method. The unit of measurement for the period is weeks per cycle. By figure 8, we are able to know which peak in figure 5 indicates a valid period. For example, the last peak in figure 5 does not indicate a period. This implies that FFT is an insufficient tool for analysing non-dominant periods.
According to Figure 6, the price of bananas initially rises, then falls, and then rises again. The peak appears around the middle of March, while the lowest point appears around September. This may be related to the rainy and dry seasons in Colombia, where the rainy season occurs from April to November and the dry season occurs from December to March. Bananas are grown and harvested year-round in Colombia, with peak production occurring between the months of July and December. These patterns can help explain the low price in the figure. It is also worth mentioning that while bananas are generally quite resilient and can tolerate a wide range of growing conditions, they can be affected by excess rainfall. In the rainy season, the harvest of bananas may be affected, leading to low production of bananas and a corresponding increase in price due to the low supply. Additionally, black dots have a large variance from 40 to 52, which implies the price varies a lot from September to December. Relatively, the price from January to March is more stable.

Focusing on figure 6, the period which is less than one year may be affected by the low production of other countries. The period longer than one year may be related to the climate period or business cycle. It can be observed that some trends have eventually vanished. The disappearance of a period in a time series can be influenced by a variety of factors, including changes in the underlying drivers of the time series, shifts in the market or economic conditions, and changes in data collection or analysis methods.

3.5 Question 5

Question 5 asked to calculate the coefficients between the price under origin `Colombia` and `Costa Rica` by the **corrcoef** function in Matlab.
It is worth mentioning, the **corrcoef** function uses the Pearson correlation coefficient as the method to find the coefficient between two variables. This method is sensitive to outliers, and the datasets must contain data sampled at the same time points. Additionally, the data used for calculation should be normally distributed.
The first three steps are taken to determine the correlation coefficients are similar to the steps mentioned in the `preparation` section in question 4:
STEP 1 Take the union of the sampled date of two datasets, so they are sharing the same sampling time. The unknown data would be filled by the MAKIMA interpolation method mentioned before.
STEP 2 Find and substitute outliers in two datasets by the Generalized extreme studentized deviate test and replace the value of outliers with the result of the MAKIMA interpolation method.
STEP 3 Normalize two datasets.
After these steps, we can apply the `corrcoef` function on datasets which gives the result of the correlation coefficient.
The formula for Pearson coefficient is [2] $\rho(A,B) = \frac{1}{N-1}\sum_{i=1}^{N}(\frac{A_i-\mu_A}{\sigma_A})(\frac{B_i-\mu_B}{\sigma_B}) = \frac{cov(A,B)}{\sigma_A\sigma_B}$.
According to figure 9, the correlation coefficient determined by the **corrcoef** function in Matlab is 0.74, which indicates that the banana price of Colombia and Costa Rica is significantly correlated.

```
>> co

co =

    1.0000    0.7400
    0.7400    1.0000
```

Figure 9: Coefficient

4 Code implecation

Listing 1: code

```
%% Question 1
bananas = readtable('bananas.csv');%Read data
origin=unique(bananas.Origin);%create lists of the unique origins
units=unique(bananas.Units);%create lists of the unique units
%% Question 2
%Calculate the mean of the Price column of the bananas table, grouped by the Origin column.
G = groupsummary(bananas,{'Origin'},'mean','Price');
%% Question 3
ReqOri={'colombia', 'costa_rica', 'dominican_republic', 'honduras', 'jamaica', 'windward_isles
    ','mexico'};
```

```
idx = contains(bananas.Origin,ReqOri);%rows of the bananas table contain the origins listed in
     ReqOri would be 1.
reqOridata = bananas(idx,:);og=categorical(reqOridata.Origin);%get the desired origins and
    stores them
boxchart(og,reqOridata.Price,'Notch','on')
ax = gca; c = ax.TickLabelInterpreter;
ax.TickLabelInterpreter = 'None';
hold on
meanPrice = groupsummary(reqOridata.Price,og,'mean');
plot(meanPrice,'-o')%plots the meanPrice data
xlabel('Origins');ylabel('Prices');legend([Price Data,Price Mean])
%% Question 4
idx1 = contains(bananas.Origin,'colombia');
colombia = bananas(idx1,2:3);%Get the data regarding colombia
df=mode(hours(-diff(colombia.Date)));colombia=table2timetable(colombia);
if isregular(colombia)==0%Deal with missing data
    colombia=retime(colombia,'regular','makima','TimeStep',hours(df));
end
colombia.Price=filloutliers(colombia.Price,'makima','gesd');%Adjust the outliers
[trend, seasonal, residual] = trenddecomp(colombia.Price);
colombia.Price=seasonal+residual;%Contains detrended data
colombia.Price=zscore(colombia.Price);%Date normalization
PPP=table(colombia.Price);D=trenddecomp(PPP);
D=splitvars(D);D=addvars(D,colombia.Price);
figure
stackedplot(D)
P=fft(colombia.Price);N = length(P);%FFT
T=days(colombia.Date(end)-colombia.Date(1));%The time span
dt=T/N;Fs=1/dt;f=(0:1:(N/2)-1)*Fs/N;period=1./f;
PP=P(1:floor(N/2));% Extraction of the positive frequency
PP(2:end-1)=2*PP(2:end-1); %Negative frequencies merged into positive frequencies
A=abs(PP);%Calculate the amplitude
figure
plot(period/7,A)%Amplitude-Period diagram
grid on
xlabel('Weeks/Cycle')
ylabel('Amplitude')
xlim([0 T/7])
vars = D.Properties.VariableNames;PE=[];
for i=2:length(vars)-2
    P1=fft(D{:,i} );
    dt=T/N;Fs=1/dt;f=(0:1:(N/2)-1)*Fs/N;period=1./f;
    PP=P1(1:floor(N/2));% Extraction of the positive frequency
    PP(2:end-1)=2*PP(2:end-1); %Negative frequencies merged into positive frequencies
    A=abs(PP);%Calculate the amplitude
    y=findpeaks(A);t=find(A==max(y));PE=[PE,period(t)/7];
end
```

```
55 result = ismember(PE,period/7);k=any(result);
56 colombia.Week=week(colombia.Date);x = colombia.Week;cftool
57 %% Question 5
58 % Extract data for colombia and costa rica
59 A=contains(bananas.Origin,'colombia');B=contains(bananas.Origin,'costa_rica');
60 colombia = bananas(A,2:3);costa_rica=bananas(B,2:3);
61 %Let price data sampled at same time points
62 [CC,IC,ID]=union(colombia.Date,costa_rica.Date,'stable');
63 price1_interp = interp1(colombia.Date, colombia.Price, CC, 'makima');
64 price2_interp = interp1(costa_rica.Date, costa_rica.Price, CC, 'makima');
65 % Remove the outliers in Prices and delete the corresponding date
66 X=filloutliers(price1_interp,'makima','gesd');%Adjust the outliers
67 Y=filloutliers(price2_interp,'makima','gesd');%Adjust the outlier
68 CL=zscore(X);RI=zscore(Y);%Data normalization
69 co=corrcoef(CL,RI);%Find the coefficient
```

Now, I would like to interpret the codes In question 3 and 4. For question 1, 2, and 5, the codes have been explained in the in-line comment.

4.1 Question 3

Codes in line 8 create a list with seven required origins. Codes in line 9 mark rows of the `bananas` table contain the origins listed in `ReqOri` as 1(True in logical). Line 10 is used to organize the data. Codes in line 11 to 17 are used to complete drawing a boxplot with a line indicating the mean and set some properties of the boxplot.

4.2 Question 4

The codes in line 20 to line 29 are preprocessing the data, the step has been explained in the `Preparation` part. The codes in line 30 to line 33 are using `trenddecomp` function to find the component in the time series, and store them in a table. The codes in line 34 to 45 are applying FFT and plot the Amplitude-Period diagram. The codes in line 47 to 54 are using FFT to calculate periods of seasonal trends decomposed by the `trenddecomp` function, and store them in a variable PE. The code `cftool` in line 56, is a function in MATLAB that opens the Curve Fitting Tool. Setting fit type as `Fourier`, and select `x`(line 56) as `x`, `D.Var1_Seasonal_2` as `y`. A fit plot of the period equal to 51.7391(first in figure 8) would be presented.

References

[1] Korstanje, J. (2021). Fourier transform for time series.

[2] Liu, X. S. (2019). A probabilistic explanation of pearson's correlation. *Teaching Statistics*.

[3] (MOHAMMAD.), H. Z. B. (2022). *Stochastic Modeling : A Thorough Guide to Evaluate, Pre-Process, Model and Compare Time Series with MATLAB Software*. ELSEVIER-HEALTH SCIENCE.