Report for Project 3

December 14, 2022

1 Background

This report presents an analysis of banana data using MATLAB. The data includes information about the origin, date, price, and units of different bananas. A sample of the data is shown in Figure 1 below. The goal is to sort and transform this data into figures to investigate the relationship between seasonality and price, as well as the relationship between the two origins represented in the data. The analysis is divided into five questions, which will be introduced one by one in the following sections of the report.

	Origin	Date	Price	Units
1	'costa_rica'	2022-11-11	1	'£/kg'
2	'guatemala'	2022-11-11	0.9100	'£/kg'
3	'dollar_bananas'	2022-11-11	0.9600	'£/kg'
4	'all_bananas'	2022-11-11	0.9600	'£/kg'
5	'colombia'	2022-11-04	0.9700	'£/kg'

Figure 1: The part of the data shows in the table.

2 Questions and solutions

2.1 Question 1

Since the data is sorted by time, the first step is to determine the distinct origins and units represented in the data. This can be done by finding the unique elements in the ”Origin” and ”Units” columns of the data. Using the MATLAB function `unique` can return the required answer. The results are shown in figure 2(a) below.

2.2 Question 2

This question aims to find the mean price for each variety of bananas mentioned in the ”Origin” column of the data. Since prices fluctuate greatly, it is difficult to compare the different varieties directly. The idea is to find all the rows in the data that contain a particular variety, and then

calculate the mean price for that variety.
Due to the large size of the data, it would be time-consuming to search for each variety line by line. For example, the first loop is a vector with `'acp_bananas'`, and the next loop is a vector with `'all_bananas'` (according to figure 1), then the number of circulations needed is equal to the height of the table in question 1 which is several hundred thousand times.
The solution is to create a column vector as height as the data table with the same content on each row of the vector. For each loop, compare the created vector to the initial "Origin" column of the data, extract the rows where the comparison returns a logical one, calculate the mean price for those rows, and save the result to a new table named Q2. When the loop ends, the results for Question 2 will be shown in Figure 2(b).

There are 28 different origins, with unit £/kg, they are:

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'}

(a) Q1 distinct origins and the unit

The mean price for each origin is shown below:

Origin	Mean
{'acp_bananas'}	0.62421
{'all_bananas'}	0.73941
{'belize'}	0.70425
{'brazil'}	0.62292
{'cameroon'}	0.6698
{'colombia'}	0.73633
{'costa_rica'}	0.74284
{'dollar_bananas'}	0.68709
{'dominican_republic'}	0.63701
{'ecuador'}	0.69899
{'eu'}	0.86
{'eu_bananas'}	0.53265
{'ghana'}	0.6399
{'guadeloupe'}	0.56162
{'guatemala'}	0.66719
{'honduras'}	0.68532
{'ivory_coast'}	0.64778
{'jamaica'}	0.60061
{'malaysia'}	0.59333
{'martinique'}	0.57364
{'mexico'}	0.64567
{'nicaragua'}	0.60542
{'panama'}	0.76091
{'somalia'}	0.74
{'st_vincent'}	0.44149
{'surinam'}	0.59271
{'venezuela'}	0.52092
{'windward_isles'}	0.54596

(b) Q2 origins with mean price

Figure 2: distinct origins with mean price and distinct units

2.3 Question 3

The seven origins of our focus in this case are Colombia, Costa Rica, the Dominican Republic, Honduras, Jamaica, the Windward Isles, and Mexico. By utilizing the method mentioned in question 2, extract the price data for these seven origins, and draw the box plot for them to visualize the data and compare the differences. The results are shown in figure 3.

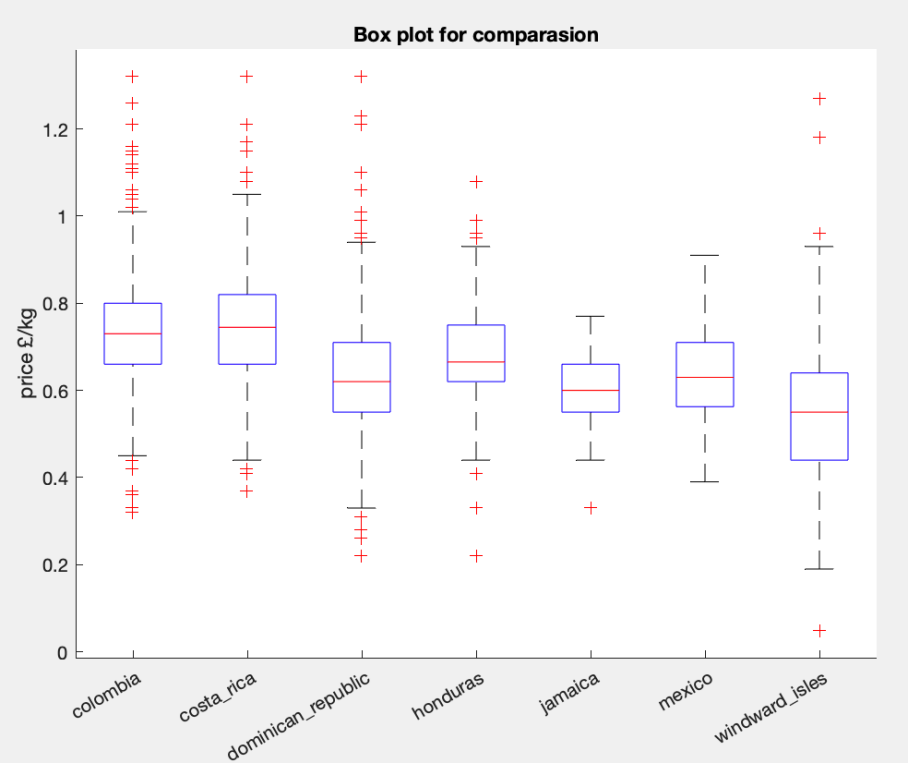


Figure 3: boxplot for seven varieties

To better understand the box plots, it's important to first understand what percentiles are. In statistics, a percentile is a score that falls below a given percentage of scores in a distribution. For example, the 25th percentile is a score below which 25% of the scores in the distribution fall. The box part of a box plot contains the data between the 25th and 75th percentiles, with the red line in the middle indicating the position of the median. Any data points that fall outside of the box plot are considered outliers. The outliers are shown separately using the "+" marker symbol, and the whiskers contain the data points that are not considered outliers[1]. The elements of the box plots are now clear, by comparing the box plots of the different origins, we can visualize and compare the differences in the price data for each origin.
Since the box contains fifty percent of the data, the position of the boxes implies that the Windward Isles have the lowest overall price level, while bananas from Colombia and Costa Rica have the highest overall price level. The thickness of the boxes indicates the concentration of the data, with thinner boxes indicating more stable prices. In this case, the bananas from Jamaica have relatively stable prices. The position of the median reflects the skewness of the data, with a left-skewed distribution indicating lower prices. The data for Honduras shows left-skewed prices. Outliers, which are data points that fall outside of the bounds of the box plot, are also an important consideration. The determination of outliers involves the calculation of percentile values. The upper bound is equal to 2.5 times the 75th percentile minus 1.5 times the 25th percentile, while the lower bound is equal to 2.5 times the 25th percentile minus 1.5 times the 75th percentile, i.e. the outliers are in the following range where Q3 denotes the 75th percentile of the data and Q1 denotes the 25th percentile of the data,

$$(-\infty, 2.5Q1 - 1.5Q3) \cup (2.5Q3 - 1.5Q1, +\infty).$$

Any data points that fall outside of these bounds are considered outliers. In the figure, the data for Colombia, Costa Rica, and the Dominican Republic have more outliers, which should be taken into account in further analyses.

2.4 Question 4

The aim of this question is to analyze seasonal trends in data from Colombia using Fast Fourier Transformation (FFT) [2] in MATLAB. The method used is based on the sunspot example [3]. FFT is a powerful tool for signal analysis that allows for the efficient calculation of the Discrete Fourier Transform (DFT) of a signal. It is particularly useful for analyzing signals with a large number of data points, as it is significantly faster than traditional DFT calculations. The FFT algorithm transforms a signal from the time domain to the frequency domain, providing valuable insights into the spectral properties and variations of the signal. This information can be used for a wide range of applications, including fault diagnosis, filtering, and image processing. Additionally, FFT can be used to analyze the variations in data, such as the seasonal trends being studied in this analysis.

The raw output of an FFT analysis is difficult to interpret directly, as it consists of complex magnitude values. However, the power of signal can be more easily understood and is often more useful for analysis. The power of a signal is simply the square of its magnitude. Since the magnitude values obtained from an FFT are symmetrical, only the power of half of the coefficients is needed to fully describe the signal. This power can then be used to create a Power Spectral Density (PSD) plot, which shows the distribution of spectral energy per unit of time (frequency). The PSD provides a visual representation of the frequency components of a signal and can be used to identify patterns and trends in the data.

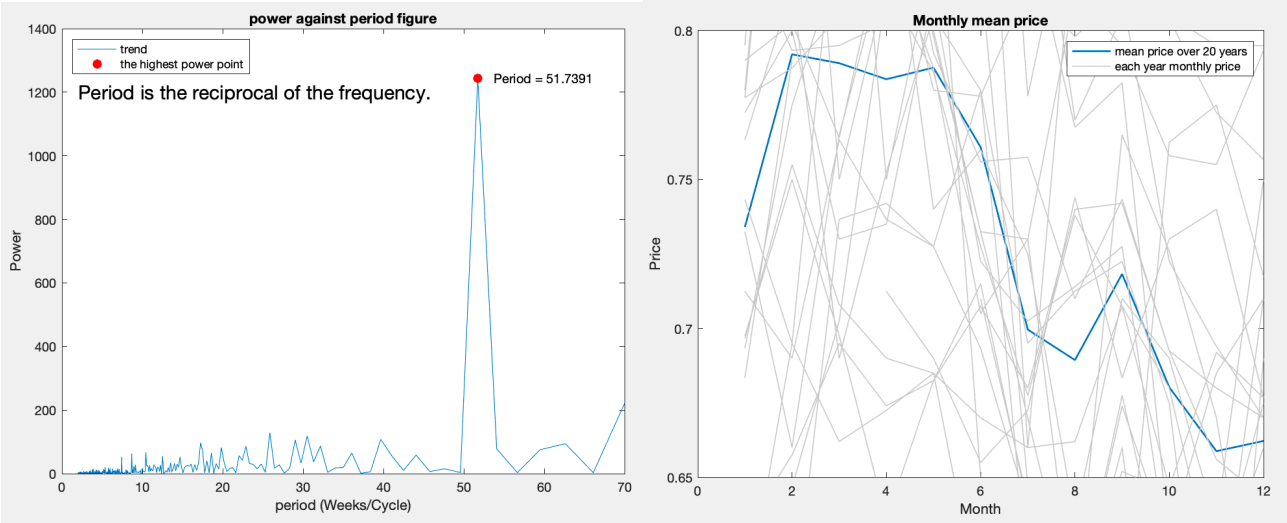
Before conducting the FFT analysis, it is necessary to prepare the data by removing any outliers and resampling the time series. Since the data points are not evenly spaced in time, resampling is required to create a regular time interval for analysis. In this case, it was determined that the most frequent time gap in the data was one week, so the resampling interval was set to one week. The values were then interpolated using the spline method, which creates a smooth curve through the data points using a series of polynomials. This spline interpolation ensures that the resampled data is both accurate and smooth, providing a suitable input for time series analysis using FFT.

To analyze the data using FFT, the algorithm is applied to the resampled time series, using the magnitude squared as a measure of power. The power values are then calculated for one half of the coefficients, as the other half is redundant due to the symmetry of the FFT output. The resulting power spectrum is plotted against period, rather than frequency, to more clearly show the cyclical patterns in the data.

The resulting plot (shown in Figure 4(a)) reveals a distinct peak at 51.7391 weeks, indicating that the price of bananas in Colombia exhibits a repeating pattern of peaks and troughs over approximately one year. This suggests that there may be seasonal factors that influence the price of bananas in the region.

The goal of this analysis is to understand the seasonal patterns in the price of bananas in Colombia over one year. To do this, the data is first plotted on a single figure, but the resulting plot is cluttered and difficult to interpret due to the large number of data points spanning over twenty years, and during this period currency appreciation and depreciation, inflation, all have an impact on prices. To simplify the data, the monthly average prices are calculated

and plotted instead (Figure 4(b)). The rationale for this is that, for example, inflation is a long-term process and over twenty years or more, can be affected approximately equally within a year, i.e., each month is affected equally by the inflation in that year.
The graph shows that the price of bananas remains at a high level from February to June and at a low level during the rest of the year. It is worth noting that Colombia is a country close proximity to the Equator, its climate is generally tropical and isothermal (without any real change of seasons). Temperatures vary a little throughout the year. The most significant climatic factor is the amount of annual precipitation, which varies seasonally with distinct rainy and dry seasons. Since Magdalena is an economically important region of banana production in Colombia, then focus on the rainfall in Magdalena. There are distinct rainy (September–October to April–May) and dry (April–May to September–October) seasons at Magdalena [4]. The length of the dry season is approximately the same as the period of high banana prices, with a time difference of around two months, which may indicate a time lag in market feedback. This is reasonable, as drought conditions can negatively impact banana yields [5].
In conclusion, this analysis has revealed that the price of bananas in Colombia tends to rise in the months leading up to the dry season, from February to June. This seasonal pattern appears to be relatively stable from year to year, indicating that it may be influenced by climatic factors such as the amount of rainfall in the region.



(a) power spectrum against period (b) monthly mean price

Figure 4: time series analysis

2.5 Question 5

The goal of this question is to calculate the correlation coefficients between the fluctuations in the prices of bananas in Colombia and Costa Rica. As in the previous analysis, the first step is to remove any outliers from the data. Since the two data sets are not sampled at the same time, the time series for the two countries is taken as a union and used as the basis for resampling. This minimizes the effect of time and reduces inaccuracies due to interpolation. The resulting data is shown in Figure 5.
A correlation coefficient is a measure of the strength and direction of the relationship between two variables. The value of the coefficient can range from -1 to 1, with -1 indicating a perfect

negative relationship, 0 indicating no relationship, and 1 indicating a perfect positive relationship. A correlation coefficient of 0.72923 is considered to be moderate, as it is not a perfect relationship but still indicates a noticeable connection between the price of these two origins.

```
The correlation coefficient between Colombia and Costa rica is 0.72923.
```

Figure 5: the correlation coefficient between Colombia and Costa Rica

3 Appendix - MATLAB code

```
format short
a=readtable('bananas.csv');a=sortrows(a,2);% full data
origin=unique(a.Origin); % answer for Q1
units=unique(a.Units); % answer for Q1
disp(['There are ' num2str(height(origin)) ' different origins, ' ...
    'with unit ' units{:} ', they are:'])
disp(cell2table(origin,'VariableNames',{'Origin'}))% answer for Q1
b=[];x=[];Q2=cell(height(origin),2);% initial variables
q3={'colombia','costa_rica','dominican_republic','honduras', ...
    'jamaica','windward_isles','mexico'};% need box plots
for c=1:height(origin)
    name=repmat(origin(c),height(a),1); % vector for find one by one
    inx=find(strcmp(a.Origin,name));b=a{inx,3};% the price for each country
    Q2{c,1}=cell2mat(origin(c));Q2{c,2}=mean(b);% storage the value in a cell
    if ismember(origin(c),q3) % table for Q3 to draw box plots
        x=[x;a(inx,3) a(inx,1)];
    end

end
Q2=cell2table(Q2,'VariableNames',{'Origin','Mean'});% add variable names
disp('The mean price for each origin is shown below:')
disp(Q2)% answer for Q2
figure
boxplot(table2array(x(:,1)),table2array(x(:,2)))% plot for Q3
title('Box plot for comparasion');
ylabel('price £/kg');

% Q4 Time Series
q4={'colombia'};q5={'costa_rica'};% same step as previous to extract data
name=repmat(q4,height(a),1);name2=repmat(q5,height(a),1);
inx=find(strcmp(a.Origin,name));COL=[];COL=[COL;a(inx,3) a(inx,2)];
inx2=find(strcmp(a.Origin,name2));COS=[];COS=[COS;a(inx2,3) a(inx2,2)];
IQR_COL=prctile(COL.Price,75)-prctile(COL.Price,25);% prepare for outliers
IQR_COS=prctile(COS.Price,75)-prctile(COS.Price,25);
% values greater than the upper bound or
```

```
% less than the lower bound are outliers
inx_o_L=find(COL.Price>prctile(COL.Price,75)+1.5*IQR_COL ...
    |COL.Price<prctile(COL.Price,25)-1.5*IQR_COL);% index for outliers
COL(inx_o_L,:)=[];% remove the outliers
inx_o_S=find(COS.Price>prctile(COS.Price,75)+1.5*IQR_COS ...
    |COS.Price<prctile(COS.Price,25)-1.5*IQR_COS);
COS(inx_o_S,:)=[];% remove the outliers
xq = (COL.Date(1):days(7):COL.Date(end))';% resampling time series
V = interp1(COL.Date, COL.Price, xq, 'spline');% resampling
y = fft(V);
y(1) = [];% remove the sum of the data
n = length(y);
power = abs(y(1:ceil(n/2))).^2; % power of first half of transform data
maxfreq = 1/2;                  % maximum frequency
freq = (1:n/2)/(n/2)*maxfreq; % equally spaced frequency grid
period=1./freq;               % convenient vision
index=find(power==max(power));
mainPeriodStr=num2str(period(index)); % find the strongest frequency
figure
plot(period,power) % if the x-axis is freq, then is PSD
xlim([0 70]); % zoom in on max power
xlabel('period (Weeks/Cycle)')
ylabel('Power')
hold on
plot(period(index),power(index),'r.', 'MarkerSize',25)
title('power against period figure');
text(2,1200,'Period is the reciprocal of the frequency.','FontSize',15);
text(period(index)+2,power(index),['Period = ',mainPeriodStr]);
legend({'trend','the highest power point'},'Location','northwest');
mon=[];
for i=1:12
    inx=find(xq.Month==i);
    mon=[mon mean(V(inx))];
end
figure
plot(1:12,mon,'LineWidth',1.3)% trend over 20 years
hold on
years=unique(year(COL.Date));
for i=1:length(years)
    y=[];
    inx=find(year(COL.Date)==years(i));
    y=[y;COL(inx,:)];
    mons=[];
    for j=1:12
        inx=find(y.Date.Month==j);
        mons=[mons mean(y{inx,1})];
```

```
    end
    plot(1:12,mons,'LineWidth',0.8,'Color',[0.8 0.8 0.8])% trend in one year
    hold on
end
ylim([0.65 0.8])
xlabel('Month')
ylabel('Price')
legend('mean price over 20 years','each year monthly price')
title('Monthly mean price')% answer for Q4 (2 pics)


% Q5
xq=(unique([COL.Date;COS.Date]))';
A=interp1(COL.Date,COL.Price,xq,'spline');
B=interp1(COS.Date,COS.Price,xq,'spline');
R=corrcoef(normalize(A),normalize(B));
disp(['The correlation coefficient between Colombia and Costa rica is ' ...
    num2str(R(2)) '.'])% answer for Q5
```

References

[1] K. Potter, H. Hagen, A. Kerren, and P. Dannenmann, "Methods for presenting statistical information: The box plot." in *VLUDS*, 2006, pp. 97–106.

[2] H. J. Nussbaumer, "The fast fourier transform," in *Fast Fourier Transform and Convolution Algorithms*. Springer, 1981, pp. 80–111.

[3] (2022, December). [Online]. Available: https://uk.mathworks.com/help/matlab/math/using-fft.html?searchHighlight=fft%20sunspot&s_tid=srchtitle_fft%2520sunspot_1

[4] S. Zubelzu, N. Panigrahi, A. J. Thompson, and J. W. Knox, "Modelling water fluxes to improve banana irrigation scheduling and management in magdalena, colombia," *Irrigation Science*, pp. 1–11, 2022.

[5] O. R. Salau, M. Momoh, O. A. Olaleye, and R. S. Owoeye, "Effects of changes in temperature, rainfall and relative humidity on banana production in ondo state, nigeria," *World Scientific News*, no. 44, pp. 143–154, 2016.