My report for Project

Student ID no:1234567

June 16, 2021

1 Introduction

In this project we manipulated and analysed data about the price of bananas (from the given dataset **bananas.csv**) from many countries and over many years.

2 Question One

The first question asks us to output a list of distinct origins for the bananas. First we read the table into Matlab using the code

```
B = readtable('bananas-18jan21');
```

Figure 1: Read table into Matlab

and we will do this at the start of every subsequent question too. To find the distinct origins we can make use of Matlab's inbuilt **unique** function to find the unique elements of the Origin column (accessed by **B.Origin**). We then convert the output from the **cell** datatype to categorical to remove the curly brackets and quotation mark around the outputs.

The code outputs the 27 distinct origins of bananas from the datafile. Matlab presents the output in a vertical list and To conserve space in this report the output is presented here inline:

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

Figure 2: Output for question one

3 Question Two

The aim of this question is to find the three countries with the lowest average banana price over the last five years of data of that country and the three with the highest average, excluding all_bananas in both cases.

We begin by creating a list of the distinct origins, as we did in question one. Then to find the three countries with the lowest average we create (at this point empty) variables to store the three lowest averages (three_lowest_avgs) and the countries corresponding to those averages (three_lowest_countries).

Then we open a for loop, running from 1 to 27 (each number corresponding to one of the 27 origins). If the origin is all_bananas then we use continue to skip this iteration of the loop. We use Matlab's inbuilt find function to find the indices of the rows corresponding to the current origin.

```
indices = find(B.Origin == origin(k));
```

Figure 3: find row indices corresponding to current origin

Next, we want to find the row index of the data closest to five years from the latest data. We access the dates corresponding the current origin and save the latest date and create a duration array equal to five years. We then subtract the five years from the latest date. To find the date and row index of the data closest to five years from the latest data, find the minimum value of (abs(dates − earliest_required_date)). We do this using the code:

Next, we store all the banana prices for the current origin in an array x and the banana prices until five years from the latest data in an array z and we use Matlab's inbuilt mean function to find the average of the elements in z, using the lines:

Then, we use an if statement to check if three_lowest_avgs and three_lowest_countries have less than three elements. If they do then we add

```
dates = B.Date(indices);
latest_date = dates(1); five_years = years(5);
earliest_required_date = dates(1) - five_years;

%find the index of the data closest to 5 years ago from latest
[y I1] = min(abs(dates - earliest_required_date));
```

Figure 4: Find date closest to earliest required

```
x = B.Price(indices);
z = x(1:I11);
m = mean(z);
```

Figure 5: Find mean of prices over five years from latest data

the current mean and origin to three_lowest_avgs and three_lowest_countries, respectively. We do this with the code:

```
if length(three_lowest_avgs) < 3
     three_lowest_avgs = [three_lowest_avgs m];
     three_lowest_countries = [three_lowest_countries origin(i)];
```

Figure 6: The if statment

If three_lowest_avgs and three_lowest_countries have three elements, then, using an elseif statement we check if the current calculated mean is less than the maximum element of three_lowest_avgs. If it is we replace the maximum element with the current mean and the origin corresponding to the maximum element with the current origin. We do this with the code: Then, we use an if statement to check if three_lowest_avgs and
three_lowest_countries have less than three elements. If they do then we add the current mean and origin to three_lowest_avgs and three_lowest_countries, respectively. We do this with the code:

To find the three countries with the highest average, we use a method and code very similar to the above. The only difference is that in the elseif statement we compare the current mean to the minimum element of three_highest_avgs and then if the mean is greater we replace the minimum element and correpsonding origin with the current mean and origin. This difference can be seen in the code:

```
   elseif m < max(three_lowest_avgs)
        [M, I2] = max(three_lowest_avgs); %max value and index of it
        three_lowest_avgs(I2) = [];
        three_lowest_countries(I2) = [];
        three_lowest_avgs = [three_lowest_avgs m];
        three_lowest_countries = [three_lowest_countries origin(i)];
```

Figure 7: The elseif statment for lowest averages

```
   elseif m > min(three_highest_avgs)
        [M1, I21] = min(three_highest_avgs); %min value and index of it
        three_highest_avgs(I21) = [];
        three_highest_countries(I21) = [];
        three_highest_avgs = [three_highest_avgs m];
        three_highest_countries = [three_highest_countries origin(k)];
```

Figure 8: The elseif statment for highest averages

We obtain the outputs for three countries with highest and three with lowest averages:

```
   three_lowest_countries =

  1x3 categorical array

     surinam      venezuela      windward_isles


three_highest_countries =

  1x3 categorical array

     costa_rica      dollar_bananas      panama
```

Figure 9: Output for question two

4 Question Three

The aim of this question is to plot the banana prices for **Costa Rica**, **Ecuador** and **Windward Isles** across the years, as well as **all_bananas** on the same plot.

First, we convert **B.Origin** to cetegorical format, so we can use it to find the row indices we need. We do this with the code:

```
B.Origin = categorical(B.Origin);
costa_rica = find(B.Origin == 'costa_rica'); %row indices for costa rica
windward_isles = find(B.Origin == 'windward_isles'); %" " windward
isles
ecuador = find(B.Origin == 'ecuador'); %" " ecuador
all_bananas = find(B.Origin == 'all_bananas'); %" " all_bananas
```

Figure 10: find row indices

Then, we make a plot with the date and price on the x- and y-axis, respectively. We do this for all four given origins on the same graph, using hold on. We give the plot a title ("Variation of price for selected countries"), a legend and we give each line on the graph a different colour, so its easier to differentiate the lines. This is done using the commands:

```
plot(B.Date(costa_rica), B.Price(costa_rica), 'k*-')
hold on
plot(B.Date(windward_isles), B.Price(windward_isles), 'r*-')
plot(B.Date(ecuador), B.Price(ecuador), 'b*-')
plot(B.Date(all_bananas), B.Price(all_bananas), 'g*-')
lg = legend('Costa Rica', 'Windward Isles', 'Ecuador', 'All Bananas');
title('Variation of price for selected countries');
xlabel('Date');
ylabel("Price in   's");
hold off
```

Figure 11: code for plot

The resulting plot:

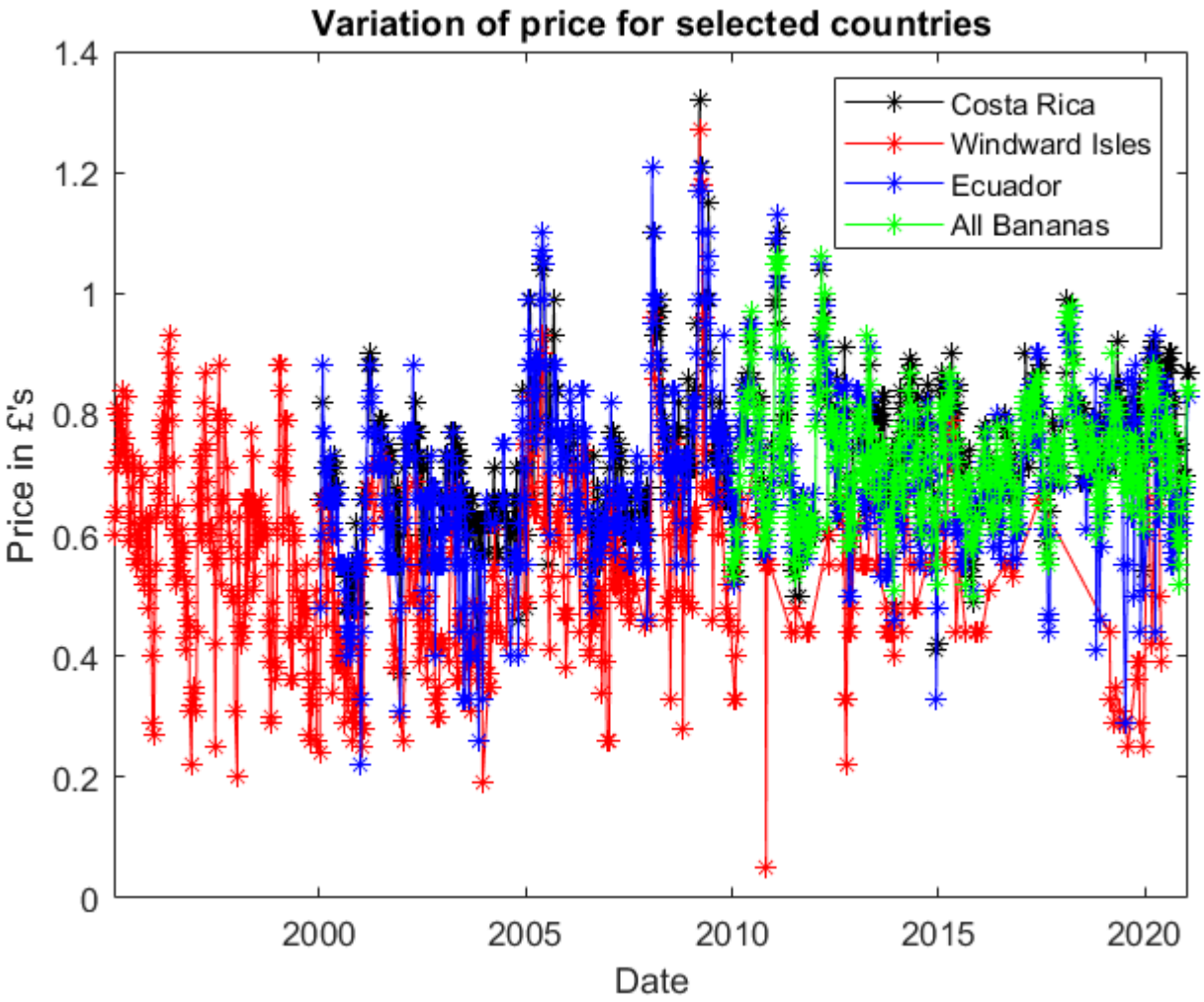

Figure 12: Plot for question three

As can be seen, the figure is very busy and dense. The Fast Fourier Transform (FFT) could have been used to reduce the data needed whilst retaining the important information but I couldn't figure out how to use it to a satisfactory standard.

5 Question Four

This question is very similar to question three. The main difference here is that we only want to consider the price between the dates between $2016-01-01$ and $2020-12-31$. To do this, we convert the table to a timetable and use the *timerange* function with appropriate arguments to reduce the timetable to the requested dates. This is done using the code:

```
B = table2timetable(B);
B = B(timerange('2016-01-01', '2020-12-31'), :);
```

Figure 13: table2timetable and reduction to required dates

The rest of the code is almost identical to question three's, with minor changes (e.g. different $x-label$ and axis spacing). The resulting plot can be seen in figure 14. Again, like with question three, we could have used the FFT to make the plot look nicer and simpler.

As for season trends, Ecuador's banana price is lower in the second half of the year than it is in the first, in all of the years on the graph and All Bananas follows pretty much the same trend. Costa Rica's price is erratic with no clear trends. There's not enough data points for Windward Isle to support any sort of conclusion there.

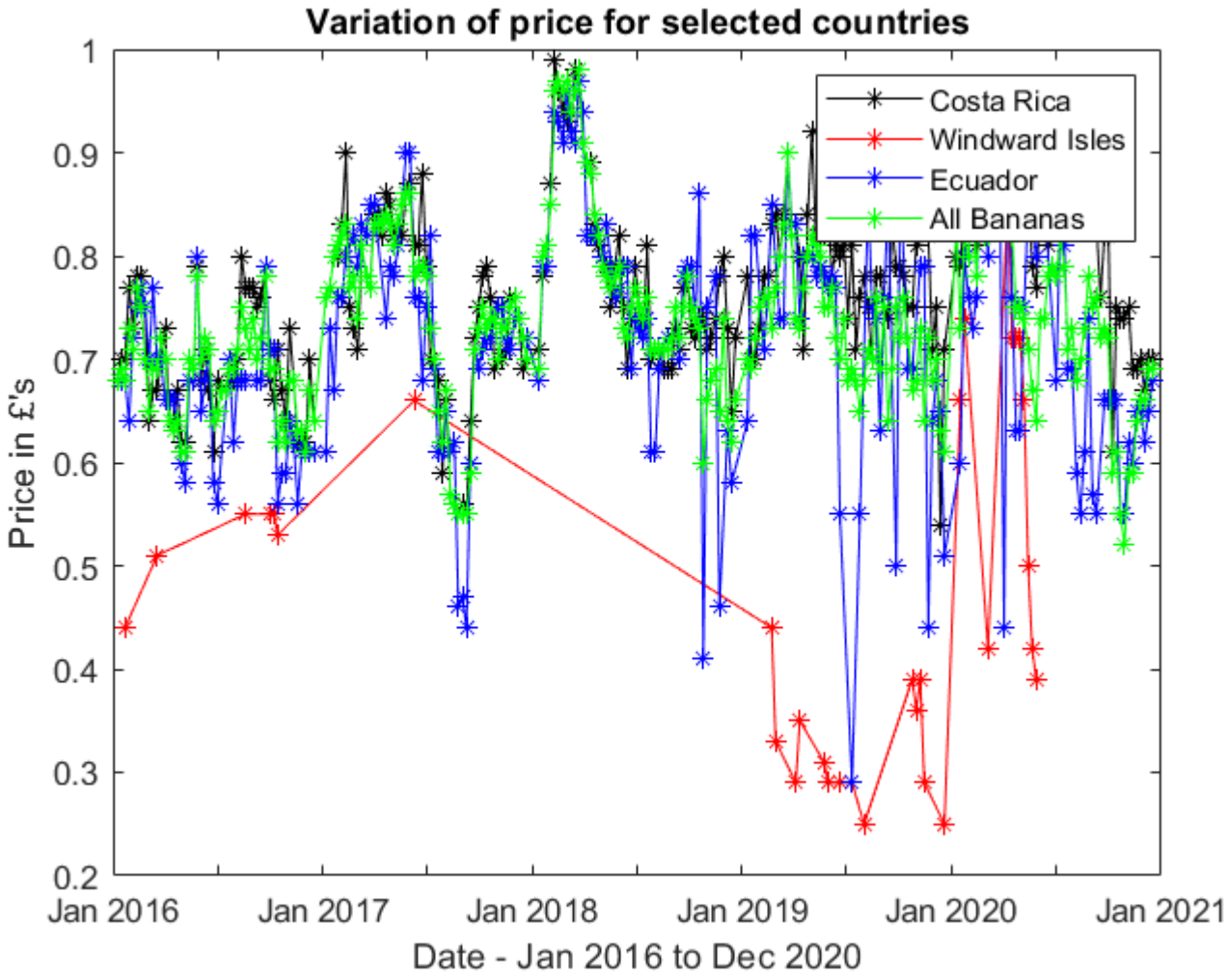

Figure 14: Plot for question four

6 Appendix

```
B = readtable('bananas-18jan21'); %read data into Matlab

origin = unique(B.Origin); %create list of distinct of distinct origins
origin = categorical(origin) %convert from cell aray to categorial.
```

Figure 15: Code for question one

```
B = readtable('bananas-18jan21'); %read data into Matlab
origin = unique(B.Origin); %create list of distinct of distinct origins
origin = categorical(origin); %convert from cell aray to categorial.
three_lowest_avgs = [];
three_lowest_countries = [];
for i = 1:27
    if origin(i) == 'all_bananas' %we don't want to include all_bananas
        continue
    end
    indices = find(B.Origin == origin(i)); %the indices where the data is
                                           %for origin(i)
    dates = B.Date(indices);
    latest_date = dates(1); five_years = years(5);
    earliest_required_date = dates(1) - five_years;
    %find the index of the data closest to 5 years ago from latest
    [y I1] = min(abs(dates - earliest_required_date));
    x = B.Price(indices);
    z = x(1:I1);
    m = mean(z);
    if length(three_lowest_avgs) < 3
        three_lowest_avgs = [three_lowest_avgs m];
        three_lowest_countries = [three_lowest_countries origin(i)];
    elseif m < max(three_lowest_avgs)
        [M, I2] = max(three_lowest_avgs); %max value and index of it
        three_lowest_avgs(I2) = [];
        three_lowest_countries(I2) = [];
        three_lowest_avgs = [three_lowest_avgs m];
        three_lowest_countries = [three_lowest_countries origin(i)];
    end
end
three_lowest_countries
three_highest_avgs = [];
three_highest_countries = [];
for k = 1:27
    if origin(k) == 'all_bananas' %we don't want to include all_bananas
        continue
    end
    indices = find(B.Origin == origin(k)); %the indices where the data is
                                           %for origin(k)
    dates = B.Date(indices);
    latest_date = dates(1); five_years = years(5);
    earliest_required_date = dates(1) - five_years;
    %find the index of the data closest to 5 years ago from latest
    [y1 I11] = min(abs(dates - earliest_required_date));
    x = B.Price(indices);
    z = x(1:I11);
    m = mean(z);
    if length(three_highest_avgs) < 3
        three_highest_avgs = [three_highest_avgs m];
        three_highest_countries = [three_highest_countries origin(k)];
    elseif m > min(three_highest_avgs)
        [M1, I21] = min(three_highest_avgs); %min value and index of it
        three_highest_avgs(I21) = [];
        three_highest_countries(I21) = [];
        three_highest_avgs = [three_highest_avgs m];
        three_highest_countries = [three_highest_countries origin(k)];
    end
end

three_highest_countries
```


Figure 16: Code for question two (with blank lines removed for brevity

```
B = readtable('bananas-18jan21'); %read data into Matlab

B.Origin = categorical(B.Origin);
costa_rica = find(B.Origin == 'costa_rica'); %row indices for costa rica
windward_isles = find(B.Origin == 'windward_isles'); %" " " windward
isles
ecuador = find(B.Origin == 'ecuador'); %" " " ecuador
all_bananas = find(B.Origin == 'all_bananas'); %" " " all_bananas

plot(B.Date(costa_rica), B.Price(costa_rica), 'k*-')
hold on
plot(B.Date(windward_isles), B.Price(windward_isles), 'r*-')
plot(B.Date(ecuador), B.Price(ecuador), 'b*-')
plot(B.Date(all_bananas), B.Price(all_bananas), 'g*-')
lg = legend('Costa Rica', 'Windward Isles', 'Ecuador', 'All Bananas');
title('Variation of price for selected countries');
xlabel('Date');
ylabel("Price in   's");
hold off
```

Figure 17: Code for question three

```
B = readtable('bananas-18jan21'); %read data into Matlab
B = table2timetable(B);
B = B(timerange('2016-01-01', '2020-12-31'), :);

B.Origin = categorical(B.Origin);
costa_rica = find(B.Origin == 'costa_rica');
windward_isles = find(B.Origin == 'windward_isles');
ecuador = find(B.Origin == 'ecuador');
all_bananas = find(B.Origin == 'all_bananas');

plot(B.Date(costa_rica), B.Price(costa_rica), 'k*-')
hold on
plot(B.Date(windward_isles), B.Price(windward_isles), 'r*-')
plot(B.Date(ecuador), B.Price(ecuador), 'b*-')
plot(B.Date(all_bananas), B.Price(all_bananas), 'g*-')
lg = legend('Costa Rica', 'Windward Isles', 'Ecuador', 'All Bananas');
title('Variation of price for selected countries');
xlabel('Date - Jan 2016 to Dec 2020');
ylabel("Price in   's");
hold off
```

Figure 18: Code for question three