Project 3

Background

In this project, it will be analysed a data set, containing information about different fruits and vegetables. There are 6 indicators described in Table 1:

category	**item**	**variety**	**date**	**price**	**unit**
shows whether the data entry is a fruit or a vegetable	shows which fruit or vegetable it is	shows the assortment of the item	shows the date on which the data is taken	shows the price of the item on the particular date	shows the measurement of the item - ”kg”, ”head”, ”twin” etc.

Table 1: Description of each header in the data set

Conducting the analysis requires loading the data file first, which is done by the MATLAB code shown in Figure 1.

```
fv=readtable('fruitandvegprices.csv'); %load all the data
```

Figure 1: MATLAB code for loading the data set

Task 1

1) Introduction to the problem

In the first task, a list of every unique element in the **item** category (List 1) and every unique **variety** for tomatoes, beans, lettuce, and carrots (List 2) should be generated.

2) Code and its implementation

List 1 can be created by the code in Figure 2. On line 1 we extract the data contained in the item header only. However, some repetitions may appear /for example, different varieties of tomatoes would still appear as 'tomatoes' in the item category/. Hence, the built-in MATLAB function **unique** is used, which ”returns the same data, but with no repetitions and in sorted order”. (The MathWorks, Inc., ”Unique values in array”)

```
1 all_items=fv.item; %list of all the items
2 dist_items=unique(all_items) %list of only the distinct items
```

Figure 2: MATLAB code for creating List 1

List 2 is made by the MATLAB code in Figure 3. Here, we first extract the data for the varieties of the tomatoes only /on line 2/ and then we remove the repetitions in the data /on line 3/. Similarly we do the same for the beans /lines 5 and 6/, for the lettuce /lines 8 and 9/ and carrots /lines 11 and 12/.

```
%Find distinct varieties for each of tomatoes, beans, lettuce and carrots
tom_var=fv.variety(strcmp(fv.item, 'tomatoes')); %all varieties of tomatoes
tom_unique_var=unique(tom_var) %unique varieties of tomatoes

beans_var=fv.variety(strcmp(fv.item, 'beans')); %all varieties of beans
beans_unique_var=unique(beans_var) %unique varieties of beans

let_var=fv.variety(strcmp(fv.item, 'lettuce')); %all varieties of lettuce
let_unique_var=unique(let_var) %unique varieties of lettuce

car_var=fv.variety(strcmp(fv.item, 'carrots')); %all varieties of carrots
car_unique_var=unique(car_var) %unique varieties of carrots
```

Figure 3: MATLAB code for creating List 2

3) Results

The output of Figure 2 to create List 1 is displayed in Figure 15 in the Appendix. The list is shown on lines 5 to 58 in Figure 15.

The output of Figure 3 is shown in Figure 14. It can be observed that for **tomatoes** there are 4 distinct varieties /lines 3 to 6/, for **beans** - 3 /lines 9 to 11/, for **lettuce** - 4 /lines 14 to 17/ and for **carrots** - just one /line 20/.

Task 2

1) Introduction to the problem

In the second task, the mean value for the **variety** of every kind of **tomato** should be determined.

2) Code and its implementation

The code used for this question is shown in Figure 4. Solving the problem requires extracting all the data for the tomatoes first. This includes the values for all categories and is done on line 1 in Figure 4. Therefore, it should be narrowed down to the varieties of the tomatoes and their prices only. Consequently, the 4 mean values for each variety should be found. The easiest way this could be done is by using Group Statistics since one of the input arguments could be chosen as 'mean' and it would be calculated directly by the software (The MathWorks, Inc., "Compute descriptive statistics of repeated measures data by group"). This calculation is done on line 2 in Figure 4.

```
1 tom_data=fv(strcmp(fv.item, 'tomatoes'),:); %all the data for the tomatoes
2 grpstats(tom_data,'variety','mean','DataVars',{'price'}) %find the mean price for each
      variety of the tomatoes
```

Figure 4: MATLAB code for finding the mean value for each variety of the tomatoes

3) Results

The output of the code in Figure 4 is displayed in Figure 5. It can be easily seen that the data set includes 206 **round**, 187 **vine**, 144 **cherry** and 71 **plum tomatoes** with mean prices 1.0996£, 1.4963£, 2.1984£ and 1.4224£ respectively.

```
ans =
  4 3  table
               variety     GroupCount   mean_price
              __________   __________   __________
    round     {'round' }      206         1.0996
    vine      {'vine' }       187         1.4963
    cherry    {'cherry'}      144         2.1984
    plum      {'plum' }        71         1.4224
```

Figure 5: MATLAB output for the mean values of the varieties of the tomatoes

4) Code validation

In order to validate the code output in Figure 5, let us find the means without using the **grpstats** function as shown in Figure 6. We first extract the prices for all the tomatoes on line 1, then use **for** loop to find the prices for each variety separately at each iteration on lines 2 and 3 and finally, find the mean for each variety using these prices /lines 4 and 6/.

```
1 tom_prices = tom_data.price; %prices for all the tomatoes
2 for i=1:4
3     prices=tom_prices(strcmp(tom_var, tom_unique_var(i)));%finding all the prices for each
          variety of the tomato
4     mean_tom_pr(i)=mean(prices); %finding the mean of these prices for each variety
5 end
6 mean_tom_pr(1:i)
```

Figure 6: MATLAB code validation

Running the code in Figure 6, the output shown in Figure 7 is consistent [1] with the one in Figure 5.

```
ans =
    2.1984    1.4224    1.0996    1.4963
```

Figure 7: MATLAB code validation output

Task 3

1) Introduction to the problem

The third question requires generating a box plot of the differences of the prices of all the varieties of the **tomatoes**.

2) Code and its implementation

The code for this question is presented in Figure 8. Solving this question does not require much effort since we could simply use the built-in function for creating a box plot in MATLAB. The only thing that should be considered is that the first input argument should be a "numeric vector or a matrix", whereas the second one could also be a "categorical array". (The MathWorks, Inc., "Visualize summary statistics with box plot") Using the data for tomatoes only we defined on line 1 in Figure 4, we set the **price** category from this data as the first argument and the **variety** category as the second one /line 1 in Figure 8/.

[1] Note that the means are not outputted in the same order.

```
boxplot(tom_data.price, tom_data.variety); %generating a box plot
title('Variation of prices of tomato variaties'); %label the plot
xlabel('Variety of tomatoes'); %label the x-axis
ylabel('Price'); %label the y-axis
```

Figure 8: MATLAB code for generating a box plot of the prices of the different tomato varieties

3) Results and interpretation

The output we get after running the code in Figure 8 is displayed in Figure 9.

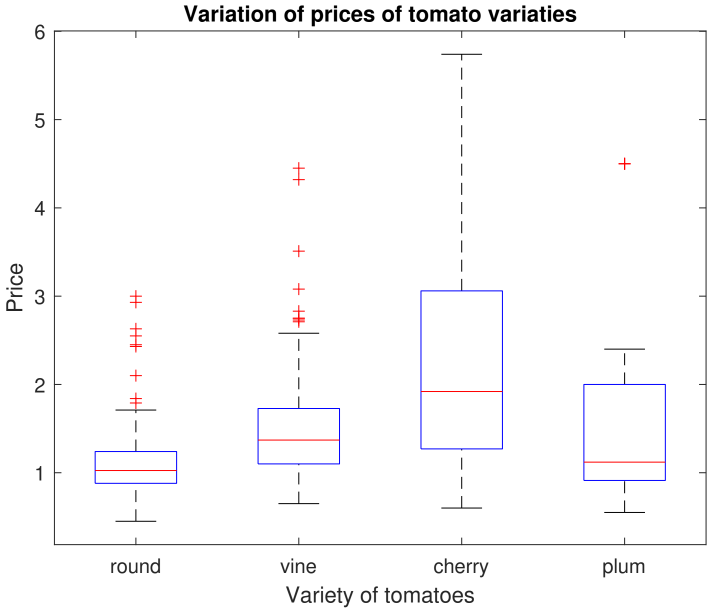

Figure 9: Variation of the price of each tomato variety

In Figure 9 we can observe that the distributions for the prices of round and plum tomatoes are fairly symmetric. Since "the longer the box and whiskers, the greater the variability of the distribution", cherry tomatoes have the greatest variability in their price, whereas round tomatoes have the smallest one. The highest median has again the price of the cherry tomatoes (with a quality value of approximately 2), followed by the one for vine, plum, and round tomatoes. Moreover, we can see that the price of the cherry tomatoes is right-skewed since the median is closer to the box's lower values and the upper whisker is longer, making it positively skewed. We can therefore conclude that the data has a tail that extends more towards the higher values. Considering the outliers, we can ignore the one for plum tomatoes. However, for round and vine tomatoes there are more outliers, reaching a price of approximately 1£ higher than their range (Frost, J., 2023, "Box Plot Explained with Examples"). In conclusion, the variability in the price of cherry tomatoes is the highest, followed by this of the vine, plum and round tomatoes.

Task 4

1) Introduction to the problem

In this problem, the data for round tomatoes will be analysed and more particularly, their time series and observable seasonal trends.

2) Time Series Analysis

There are 4 different types of components when analysing time series, presented in Table 2.

	Trends	**Seasonality**	**Cycles**	**Noise**
Time	fixed	fixed	not fixed	not fixed
Nature	”gradual, upward/downward trend”	”swings between up or down, pattern repeatable”	”repeating up and down, no fixed period”	”high fluctuation, not repeatable”
Prediction capability	predictable	predictable	challenging	challenging

Table 2: Components of Time Series Data

(Pandian, S., 2023, ”Time Series Analysis and Forecasting — Data-Driven Insights”, ”Components of Time Series Analysis”)

3) Code and its implementation

For analysing the time series of round tomatoes, 2 categories will be needed - date and price. Therefore, we will first extract the data for these 2 indicators for the round tomatoes only as shown on line 1 in Figure 10. Here, we put as column values 4 : 5, because these are the columns of the desired categories. Since the output of this line will be shown as a table, we should convert the date to an array /line 2/ in order to plot it /line 3/.

```
1  round_tomatoes=tom_data(strcmp(tom_data.variety, 'round'),4:5); %date and price for round tomatoes
2  date=datetime(table2array(round_tomatoes(:,1))); %convert the data table into an array
3  plot(date,round_tomatoes.price) %plot the time series
```

Figure 10: MATLAB code for creating a plot of the time series of the round tomatoes

4) Results and interpretation

After running the code in Figure 10 we get the plot presented in Figure 11. Comparing the graph with the information given in Table 2, we can conclude that the component we have is seasonality.

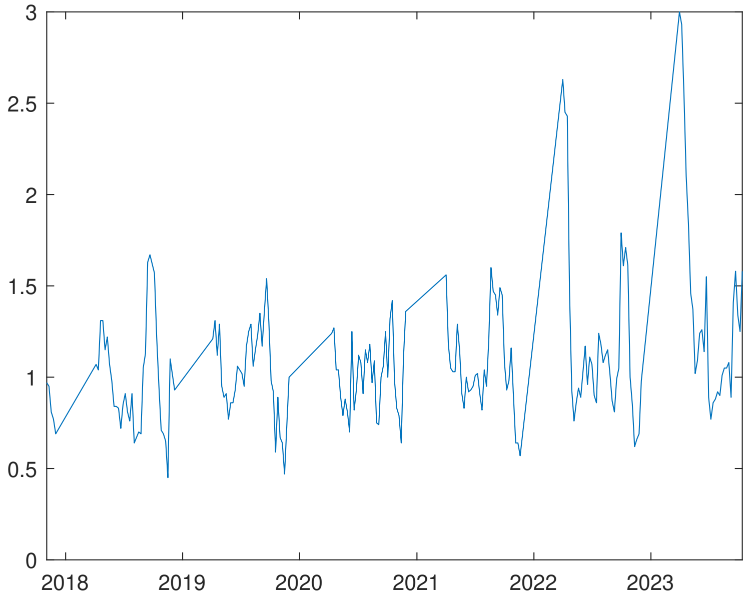

Figure 11: Time series for round tomatoes

From Figure 11 an observable trend is that in the first quarter of each year in the time frame from 2018 to 2023 there is an increase in the price of round tomatoes with a peak at the beginning of the second quarter. A decrease in the price follows after that with a trough reached during the third quarter of each year. Another rise follows with a peak at the end of the third quarter/the beginning of the fourth quarter. Similarly, as before, a drop comes after, followed by another growth of the price in the last month of the fourth quarter.

In conclusion, we can see that the price of the round tomatoes fluctuates significantly during the observed time period, and it ”swings between up or down”. However, it can be noted that there are 2 peaks each year - at the beginning of the spring and the autumn and 2 troughs - at the beginning of the summer and the winter. Therefore, it is ”pattern repeatable”, we can make predictions for the future /see Table 2/ and also conclude that weather could be one of the factors influencing the price.

Task 5

1) Introduction to the problem

In question 5, there should be calculated the correlation coefficients of the prices of **carrots** and **round tomatoes**.

2) Code and its implementation

The code needed for solving this question is shown in Figure 12. Before finding the correlation coefficients, we first need to extract the data for the prices of the carrots and the prices of the round tomatoes. This is done on lines 1 and 2 in Figure 12 respectively. Since the two data sets have different sizes we find the row indices /i.e. number of rows in the two vectors/ on lines 4 and 5 and the intersection of both on line 7. Hence, it is now possible to calculate the coefficients using the built-in function in MATLAB /line 9/.

```
1 carrot_prices=fv.price(strcmp(fv.item, 'carrots'),:); %prices of the carrots
2 tomatoes_round_prices=fv.price(strcmp(tom_data.variety, 'round'),:) %prices of the round
      tomatoes
3 % row indices
4 ind1 = find(isfinite(carrot_prices));
5 ind2 = find(isfinite(tomatoes_round_prices));
6 % intersection of the indices
7 ind = intersect(ind1, ind2);
8 % correlation coefficients of the data
9 corrcoef(carrot_prices(ind),tomatoes_round_prices(ind))
```

Figure 12: MATLAB code for finding the correlation coefficients of the prices of carrots and round tomatoes

3) Results and interpretation

Running the code in Figure 12 gives the output in Figure 13.

```
1 ans =
2     1.0000   -0.0430
3    -0.0430    1.0000
```

Figure 13: MATLAB code output for finding the correlation coefficients of the carrot and round tomatoes price

On lines 2 and 3 in Figure 13 we notice that the correlation coefficient we wanted to find is -0.0430. Since the value is really close to zero, we conclude that the relationship between the prices of the carrots and those of the round tomatoes is negative and very weak. Therefore, there is almost no correlation between them. The diagonal of the matrix in Figure 13 contains 1's as it calculates the relationship between the prices of the same vegetables with themselves. (Frost, J., 2023, ”Interpreting Correlation Coefficients”)

References

The MathWorks Inc. (2023b) - ”Unique values in array - MATLAB unique”, ”Description”. [Online], accessed 1 December 2023, < https://uk.mathworks.com/help/matlab/ref/double.unique.html >.

The MathWorks Inc. (2023b) - ”Compute descriptive statistics of repeated measures data by group”, ”Input Arguments”. [Online], accessed 1 December 2023, < https://uk.mathworks.com/help/stats/repeatedmeasuresmodel.grpstats.html >.

The MathWorks Inc. (2023b) - ”Visualize summary statistics with box plot”, ”Input Arguments”. [Online], accessed 1 December 2023, < https://uk.mathworks.com/help/stats/boxplot.html >.

Frost, J. (2023) - ”Box Plot Explained with Examples”. [Online], accessed 3 December 2023, < https://statisticsbyjim.com/graphs/box-plot/ >.

Pandian, S. (2023) - ”Time Series Analysis and Forecasting — Data-Driven Insights”, ”Components of Time Series Analysis”. [Online], accessed 3 December 2023, < https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-to-time-series-analysis/ >.

Frost, J. (2023) - ”Interpreting Correlation Coefficients”. [Online], accessed 3 December 2023, < https://statisticsbyjim.com/basics/correlations/ >.

Appendix

Code listings

[All the codes have been tested in MATLAB.]

```
tom_unique_var =
  4 1  cell array
    {'cherry'}
    {'plum'  }
    {'round' }
    {'vine'  }
beans_unique_var =
  3 1  cell array
    {'broad'                 }
    {'dwarf_french_or_kidney'}
    {'runner_climbing'       }
let_unique_var =
  4 1  cell array
    {'butterhead_indoor' }
    {'cos'               }
    {'crisp_iceberg_type'}
    {'little_gem'        }
car_unique_var =
  1 1  cell array
    {'topped_washed'}
```

Figure 14: MATLAB code output for List 2

```
dist_items =

  54 1  cell array

    {'alstromeria'       }
    {'apples'            }
    {'asparagus'         }
    {'beans'             }
    {'beetroot'          }
    {'blackberries'      }
    {'blueberries'       }
    {'brussels_sprouts'  }
    {'cabbage'           }
    {'calabrese'         }
    {'capsicum'          }
    {'carrots'           }
    {'cauliflower'       }
    {'celeriac'          }
    {'celery'            }
    {'cherries'          }
    {'chinese_leaf'      }
    {'chrysanthemum'     }
    {'coriander'         }
    {'courgettes'        }
    {'cucumbers'         }
    {'curly_kale'        }
    {'currants'          }
    {'cyclamen'          }
    {'geranium'          }
    {'gladioli'          }
    {'gooseberries'      }
    {'leeks'             }
    {'lettuce'           }
    {'lillies'           }
    {'mixed_babyleaf_salad'}
    {'narcissus'         }
    {'onion'             }
    {'pak_choi'          }
    {'parsnips'          }
    {'pears'             }
    {'peas'              }
    {'peony'             }
    {'plums'             }
    {'poinsettia'        }
    {'raspberries'       }
    {'rhubarb'           }
    {'rocket'            }
    {'spinach_leaf'      }
    {'spring_greens'     }
    {'stocks'            }
    {'strawberries'      }
    {'swede'             }
    {'sweet_williams'    }
    {'sweetcorn'         }
    {'tomatoes'          }
    {'tulips'            }
    {'turnip'            }
    {'watercress'        }
```

Figure 15: MATLAB code output for List 1