MATH36031 Project 3

10680317

Dec 2023

1 Introduction

In this project, we examine the `fruitandvegprices.csv` dataset, which meticulously records an array of fruits and vegetables, with their respective varieties, dates of record, prices, and units of measurement. **Figure** 1 shows the first 5 lines of the dataset, providing an introductory view of its composition and the type of data under analysis.

The project is divided into 5 tasks that ranging from basic data identification, such as listing unique items, to more complex analyses, including price variation studies and correlation assessments. Specifically, we explore the price dynamics of items like tomatoes and carrots, visualise data through box plots and time series analysis, and concluding with a correlation analysis between prices of carrots and tomatoes.

1	category	item	variety	date	price	unit
2	fruit	apples	bramleys_seedling	2023-10-13	1.14	kg
3	fruit	apples	coxs_orange_group	2023-10-13	1.21	kg
4	fruit	apples	egremont_russet	2023-10-13	1.44	kg
5	fruit	apples	gala	2023-10-13	1.21	kg

Figure 1: Problem description

2 Task1

2.1 Motivation

In task 1, our aim is to analyse the `fruitandvegprices.csv` dataset. Then generate a list of every unique entry under the `item` column, and identify all the distinct varieties associated with 4 items: Tomatoes, beans, lettuce and carrots.

2.2 Method

The full demonstration of the task 1 code can be found in **Appendix** (**Listing** 2). We examine the `fruitandvegprices.csv` dataset using `readtable` function and store it in variable `data`. For better precision, the numeric display format is set in `long`. In **Line** 3, we extract all unique entries from the `item` column of the `data` table. The `unique` function ensures each item is listed only once. This list of unique items is then displayed, followed by a count of these distinct items, providing a clear overview of the dataset's diversity.

Furthermore, we aim to identify and list the unique varieties for these specific items, a cell array

`specificItems` is defined for these items. From **Line** 10, a `for` loop iterates through each item in the `specificItems`. The code filters the dataset for each item, extracts its unique varieties using the `unique` function. Additionally, for each of these specific items, the code calculates and displays the number of their distinct varieties.

2.3 Output

The full demonstration of the task 1 output can be found in **Appendix** (**Listing** 2). The output of the code shows that there are 54 distinct items in total, and the number of distinct varieties for tomatoes, beans, lettuce, and carrots are 4, 3, 4 and 1 respectively. We want to make sure the output lists are correct. Therefore, we can read the csv file by using Excel, and using the "Sort & filter" section to double check our results. In this case it is correct.

3 Task2

3.1 Motivation

Task 2 focus on analysing the price dynamics of various tomato varieties, and calculate the average prices for each distinct variety of tomatoes. By calculating the average price of tomato varieties, we can identify pricing patterns of the market, also provides us with data-informed evidence to make informed decisions when adjusting production schedules, tailoring marketing campaigns, or setting competitive prices in the marketplace[1].

3.2 Method

The full demonstration for the task 1 code can be found in **Appendix** (**Listing** 3). We first filter the `data` to include rows of tomatoes from `item` column using `strcmp` function. This function is used for string comparison, and we store it in a new table called `tomatoesData`. Then we apply `grpstats` function to calculate the mean price for each tomato variety and display it. This is the function parameters in **Line** 6:

- `tomatoesData`: The tomato dataset.
- `variety`: Indicates that the data should be grouped by the different varieties of tomatoes.
- {`mean`}: Specifies that the mean of each group should be calculated.
- `DataVars`: Indicates which variable in `tomatoesData` should be used for the calculation, in this case: `price`.

3.3 Output

The full demonstration for the **Task** 2 output can be found in **Appendix** (**Listing** 4). The mean price for round, vine, cherry, plum are 1.100, 1.496, 2.198 and 1.422 respectively. Suggesting that round tomatoes may be among the more affordable options in the market, and cherry tomatoes has the highest mean price, which might due to the high consumer demand or lower supply levels.

4 Task3

4.1 Motivation

Building upon the insights gained in task 2, task 3 involves creating a box plot to illustrate the variations in tomato prices across different varieties. We continue our analysis using the same dataset `tomatoesData`

from **section** 3.

4.2 Method

The full demonstration for task 3 code can be found in **Appendix** (**Listing** 5). In **Line** 3, the `boxplot` function is used to create the box plot. The first argument, `tomatoesData.price` specifies the data to be plotted, which in this case is the price of tomatoes. The second argument, `tomatoesData.variety` is used to group the data by tomato variety. This means the box plot displays a separate box for each variety of tomato, showing the distribution of prices within each variety.

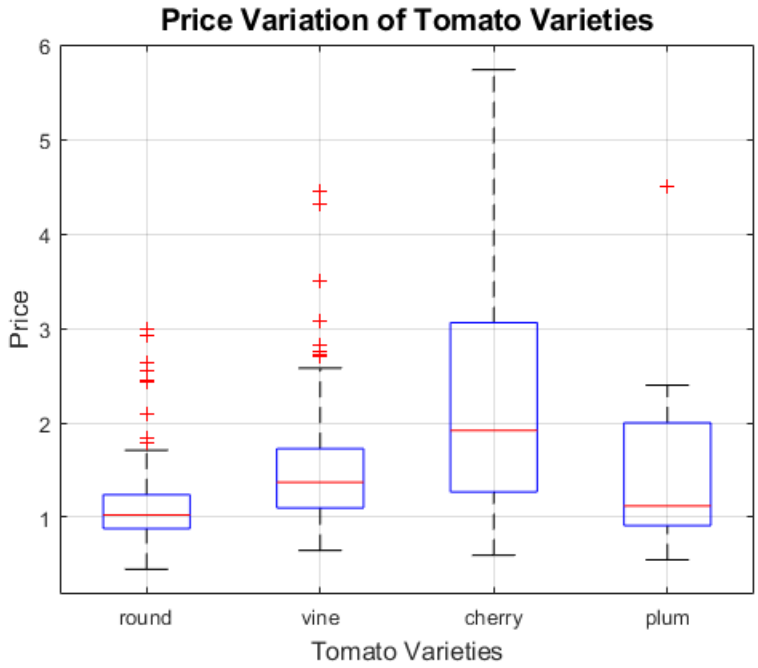

Figure 2: The Box plot for price variation of tomato varieties

4.3 Output

Figure 2 is the boxplot of task 3. It displays the price variation across four different varieties of tomatoes:

- **Round Tomatoes:** The median price is the lowest at 1.025, and the narrow interquartile range (IQR) from 0.88 to 1.24 suggests less variability. There are 9 outliers, indicating some instances of higher prices.
- **Vine Tomatoes:** This variety has a higher median price of 1.37 with an expanded IQR, indicating greater price variability. Outliers above Upper IQR (1.7275) are noted, suggesting sporadic higher prices.
- **Cherry Tomatoes:** It has the highest median price of 1.92 and the largest IQR, suggesting a wide range of prices. Also has the greatest maximum price (5.74) and it is the only one that does not have outliers.
- **Plum Tomatoes:** It has a median price of 1.12, positioned between the round and vine varieties, with a moderate IQR reflecting reasonable price variability. 3 outliers are observed, indicating relatively stable pricing.

Overall, the box plot illustrates that cherry tomatoes typically have the highest prices and the greatest variability, while round tomatoes are generally the cheapest with less price fluctuation. Vine and plum tomatoes are moderately priced, but there are a few instances where they are outside the usual price range.

The reasons for the existence of outliers could be due to decrease in supply because of unfavorable growing conditions, or seasonal holidays increased demand for specific varieties, which both cases can push prices up.

5 Task4

5.1 Motivation

Task 4 focus into price trends of round tomatoes, aiming to graphically represent the trend over time using a time series plot, potentially revealing insights into seasonal effects, supply and demand changes, and other economic factors influencing tomato prices.

5.2 Method

The full demonstration for task 4 code can be found in **Appendix** (**Listing** 6). We continue to use the `tomatoesData` from **section** 3. In **Line** 3, this line filters the `tomatoesData` to include round variety only, and stores in a new variable called `roundTomatoesData`. Then **Line** 5 converts the `data` column in `roundTomatoesData` from string to MATLAB's date-time format (year-month-day). The `plot` function in **Line** 7 then maps this data: `roundTomatoesData.date` (x-axis) represents dates and `roundTomatoesData.price` (y-axis) represents prices. A clear title and axis labels are applied, and `datetick` function is used to format the x-axis labels as dates. Lastly, the `axis tight` added to remove the extra margin around the data.

5.3 Output

The time-series graph can be found in **Figure** 3. In this graph, the time-series plot of round tomato prices shows a clear seasonal trend, with price peaks typically occurring in the spring, such as the annual peak in April. Conversely, autumn sees a significant price decline, which coincides with an expected post-harvest increase in supply. Prices tend to be more stable in the winter months, while the summer months show a mixed pattern, which may be due to changes in consumer demand or growing conditions.

This cyclical behaviour suggests that the price of round tomatoes is influenced by seasonal factors that may be related to the agricultural cycle of planting, growing and harvesting, as well as seasonal changes in demand. The pricing trends for round tomatoes, as evident in the time-series graph, demonstrate clear seasonal influences:

- **Spring Peaks:** Prices increase in spring due to likely supply shortages after winter and heightened demand for fresh produce, coupled with increased initial production costs.
- **Autumn Declines:** The autumn period sees a reduction in prices, aligning with the abundance of supply from the harvest season and a potential shift in consumer preferences towards other seasonal produce.
- **Winter Stability:** Prices tend to stabilize during winter, suggesting a balance between supply and demand, possibly facilitated by controlled cultivation methods like greenhouse farming.
- **Summer Variability:** Summer prices exhibit fluctuations, affected by changes in consumer demand and the variability in growing conditions.

This pattern indicates that factors such as supply-demand dynamics, consumer behavior, and production aspects play significant roles in influencing the pricing of round tomatoes throughout the year[2].

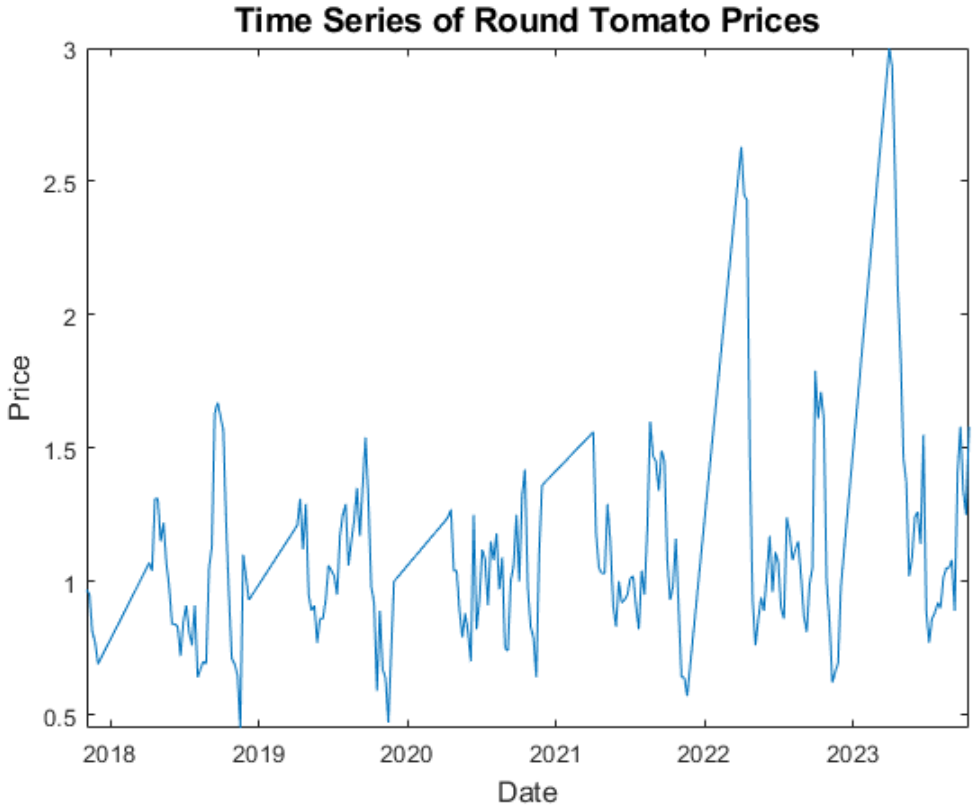

Figure 3: The time series of round tomato prices

6 Task5

6.1 Motivation

In Task 5, our analysis focused on understanding the relationship that exists potentially between the prices of two different produce items, carrots and round tomatoes.

6.2 Method

The full demonstration for task 5 code can be found in **Appendix** (**Listing** 7). In Task 5, we aim to calculate the correlation coefficient between the prices of carrots and round tomatoes. Initially, it filters the main dataset to separate out data for carrots and round tomatoes. Then converts the `Date` column of these two subsets to date-time format and sorts them by date. The code next employs the `fillmissing` function to handle any missing values in the `price` columns of both datasets, using linear interpolation to estimate and fill these gaps. The intersect function is then used to identify and extract the dates common to both datasets, ensuring a consistent comparison basis. Corresponding prices for these common dates are selected from each subset. Finally, the correlation coefficient between the carrot and tomato prices for these dates is calculated and displayed. This coefficient quantifies the strength and direction of the linear relationship between the prices of these two items.

6.3 Output

The full demonstration for task 5 code can be found in **Appendix** (**Listing** 8). The results of task 5 show that the correlation coefficient between carrot and round tomato prices is 0.1986, suggesting a weak positive linear relationship. This suggests that although there is some correlation between the two as the price of carrots has increased slightly in line with the price of tomatoes, the relationship is not strong or significant. This correlation implies that the prices of the two commodities are influenced more by individual factors than by each other[3]. The positive sign of the coefficient indicates that, generally, as the price of carrots increase, the price of round tomatoes also tends to increase. This correlation coefficient could be due to a variety of factors such as different supply chains, varying demand drivers, or disparate market conditions affecting each product.

7 Appendix

```
%% Task 1
format long;
distinctItems = unique(data.item);  % List of all distinct entries under the 'item' header
disp('Distinct Items:');                        % Display distinct items
disp(distinctItems');
totalDistinctItems = length(distinctItems);
fprintf('\nTotal Number of Distinct Items: %d\n', totalDistinctItems);
% Display total number of distinct items
specificItems = {'tomatoes', 'beans', 'lettuce', 'carrots'};
for i = 1:length(specificItems)                 % List distinct varieties for specific items
    item = specificItems{i};
    varieties = unique(data.variety(strcmp(data.item, item)));
    numVarieties = length(varieties);             % Count of distinct varieties
    varietiesStr = strjoin(varieties', ', ');     % Join varieties as a single string
    fprintf('\nDistinct varieties for %s:', item);
    fprintf('%s\n', varietiesStr);
    fprintf('Total number of distinct varieties: %d\n', numVarieties);
end
```

Listing 1: Code for Task1.

```
Distinct Items:
  Columns 1 through 10

    {'alstromeria'}  {'apples'}  {'asparagus'}  {'beans'}  {'beetroot'}  {'blackberries'}  {'
        blueberries'}  {'brussels_sprouts'}  {'cabbage'}  {'calabrese'}

  Columns 11 through 20

    {'capsicum'}  {'carrots'}  {'cauliflower'}  {'celeriac'}  {'celery'}  {'cherries'}  {'
        chinese_leaf'}  {'chrysanthemum'}  {'coriander'}  {'courgettes'}

  Columns 21 through 30

    {'cucumbers'}  {'curly_kale'}  {'currants'}  {'cyclamen'}  {'geranium'}  {'gladioli'}  {'
        gooseberries'}  {'leeks'}  {'lettuce'}  {'lillies'}

  Columns 31 through 41

    {'mixed_babyleaf_salad'} {'narcissus'}  {'onion'}  {'pak_choi'}  {'parsnips'}  {'pears'}
         {'peas'}  {'peony'}  {'plums'}  {'poinsettia'}  {'raspberries'}
```

```
  Columns 42 through 51

    {'rhubarb'}  {'rocket'}  {'spinach_leaf'}  {'spring_greens'}  {'stocks'}  {'strawberries'}
          {'swede'}  {'sweet_williams'}  {'sweetcorn'}  {'tomatoes'}

  Columns 52 through 54

    {'tulips'}  {'turnip'}  {'watercress'}

Total Number of Distinct Items: 54
Distinct varieties for tomatoes:cherry, plum, round, vine
Total number of distinct varieties: 4
Distinct varieties for beans:broad, dwarf_french_or_kidney, runner_climbing
Total number of distinct varieties: 3
Distinct varieties for lettuce:butterhead_indoor, cos, crisp_iceberg_type, little_gem
Total number of distinct varieties: 4
Distinct varieties for carrots:topped_washed
Total number of distinct varieties: 1
```

Listing 2: Output for Task1.

```
%% Task 2
tomatoesData = data(strcmp(data.item, 'tomatoes'), :);% Filter data for tomatoes
% Use grpstats to calculate mean price for each variety
meanPrices = grpstats(tomatoesData, 'variety', {'mean'}, 'DataVars', 'price');
disp(meanPrices); % Display the mean prices for each variety of tomatoes
```

Listing 3: Code for Task2.

```
              variety     GroupCount      mean_price
            ----------   ----------   ----------------

    round    {'round' }      206        1.09961165048544
    vine     {'vine' }       187        1.49625668449198
    cherry   {'cherry'}      144        2.19840277777778
    plum     {'plum' }        71        1.42239436619718
```

Listing 4: Output for Task2.

```
%% Task 3
% Create a box plot for the price of different tomato varieties
boxplot(tomatoesData.price, tomatoesData.variety)
title('Price Variation of Tomato Varieties','FontSize',14)
xlabel('Tomato Varieties','FontSize',12)
ylabel('Price','FontSize',12)
grid on
```

Listing 5: Code for Task3 Box plot.

```
%% Task 4
% Filter data for tomatoes round variety
roundTomatoesData = tomatoesData(strcmp(tomatoesData.variety, 'round'), :);
% Convert the date column to datetime format
roundTomatoesData.date = datetime(roundTomatoesData.date, 'InputFormat', 'yyyy-MM-dd');
% Plot the time series directly using datetime objects
```

```
plot(roundTomatoesData.date, roundTomatoesData.price)
title('Time Series of Round Tomato Prices')
xlabel('Date')
ylabel('Price')
% Improve the readability of the x-axis dates
datetick('x', 'yyyy-mm-dd')
axis tight % Removes the extra margin around the data
```

Listing 6: The time series code of round tomato prices.

```
%% Task 5
carrotData = data(strcmp(data.item, 'carrots'), :); % Find carrots data
roundTomatoData = data(strcmp(data.item, 'tomatoes') & strcmp(data.variety, 'round'), :);
carrotData.date = datetime(carrotData.date);      % Convert dates to datetime and sort by date
roundTomatoData.date = datetime(roundTomatoData.date);
carrotData = sortrows(carrotData, 'date');
roundTomatoData = sortrows(roundTomatoData, 'date');
% Fill missing values in the price data
carrotData.price = fillmissing(carrotData.price, 'linear');
roundTomatoData.price = fillmissing(roundTomatoData.price, 'linear');
% Find common dates in both datasets
[commonDates, ia, ib] = intersect(carrotData.date, roundTomatoData.date);
% Select prices for common dates
carrotPrices = carrotData.price(ia);
tomatoPrices = roundTomatoData.price(ib);
% Calculate the correlation coefficient for the common date prices
corrCoeff = corrcoef(carrotPrices, tomatoPrices);
% Display the correlation coefficient
disp('Correlation Coefficient between Carrot Prices and Round Tomato Prices:');
disp(corrCoeff(1,2));
```

Listing 7: Code for Task5.

```
Correlation Coefficient between Carrot Prices and Round Tomato Prices:
   0.198614939415480
```

Listing 8: Output for Task5.

References

[1] Reference listSamoggia, A., Grillini, G. and Del Prete, M. (2021). Price Fairness of Processed Tomato Agro-Food Chain: The Italian Consumers' Perception Perspective. Foods, 10(5), p.984. doi:https://doi.org/10.3390/foods10050984. Available at: https://www.mdpi.com/2304-8158/10/5/984

[2] Quora. (n.d.). What factors contribute to price seasonality in tomatoes? [online] Available at: https://www.quora.com/What-factors-contribute-to-price-seasonality-in-tomatoes

[3] Nickolas, S. (2021). What does it mean if the correlation coefficient is positive, negative, or zero? [online] Investopedia Available at: https://www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp