Consin Hu

Object-Oriented Programming

May 09, 2024

**Final Project Report**

**Introduction**

For this project, I chose to use the Fruit Prices of 2020 dataset obtained from Data.gov. The goal of this project was to analyze and visualize this dataset to derive insights into trends and patterns in fruit pricing in 2020. The dataset contains information on retail prices of various fruits across different measurements and different forms (dried, fresh, frozen). I will be discussing how transforming this raw data into a more structured format such as a chart helped me uncover valuable insights on fruit price trends that could be beneficial for consumers, retailers, and policymakers.

**ETL Process**

**Extract:**

I extracted the data from Data.gov as a CSV file and I chose to do so because I knew I wanted to open it on my laptop through Excel since it was a data heavy file.

**Transformations:**

The transformations I performed were data grouping, aggregation, and filtration which was accomplished through the pandas library.

1. Data grouping: I chose to group my data by form and retail price using the .groupby() function because I was interested in studying how different forms of fruits could affect prices.
2. Data aggregation: Once the data was grouped, I applied the aggregation function of mean( ) to calculate the summary statistic I wanted to study. In this case, I wanted to study the average price by form across different types of fruits of the fruit prices dataset.
3. Data filtering: Since $2.00 or below is a relatively normal price range for any form of fruits, I chose to filter my data to only include fruits priced on average $2.00 or above.

**Load:**

I chose to load my data first as another CSV file which I’ve included in this report in the Appendix. Then I chose to load my filtered data version of the original fruit dataset into pie chart format using Matplotlib. I chose this chart because it is a clean and clear data visualization allowing for quick understanding and immediate analysis. Percentages were used for defining pie chart sections for further ease of understanding.

**Findings**

After the first loading of the newly transformed data, the results came back that dried fruits were the most expensive form of fruits at an average of $6.21 with frozen fruits following at $3.50 and fresh at $2.08. Once the pie chart was assembled, the percentages revealed that 52.7% of the fruits above $2.00 were dried, 29.7% were frozen, and 17.7% were fresh. Although there wasn’t direct information about why the prices were distributed as such, an inference I made was that dried and frozen fruits were likely priced higher due to extra processing and packaging procedures.

**Challenges**

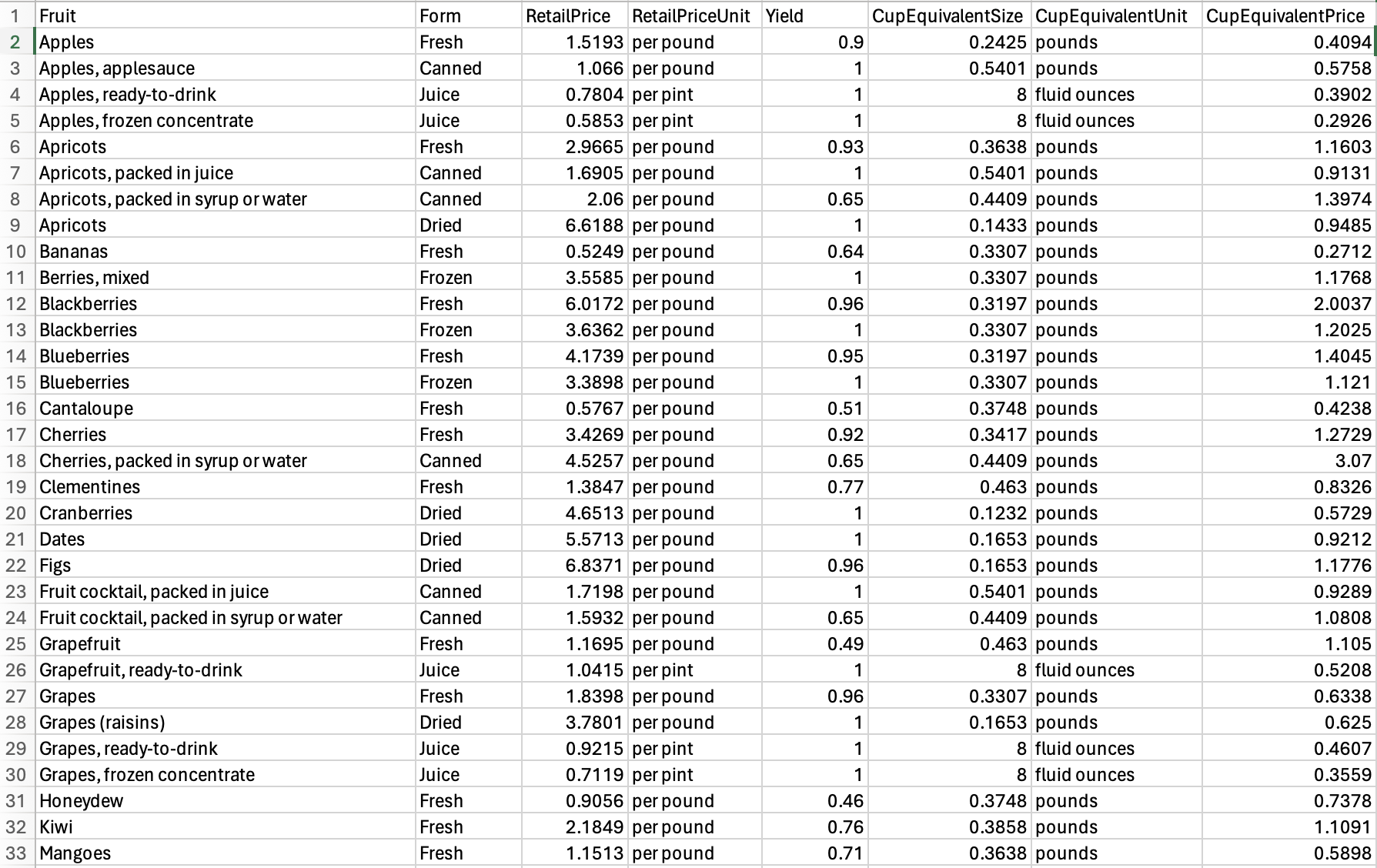
I was fortunate enough to not encounter too many challenges with this project. I think the main struggle was that this is the first project of its kind that I’ve completed so starting it was difficult and slow. I am relatively new to Python too so I often needed to look for aid online or through AI. Other than that, I think I mainly spent the majority of my time figuring out what perspective I wanted to take with the dataset and what I was interested in studying about it.

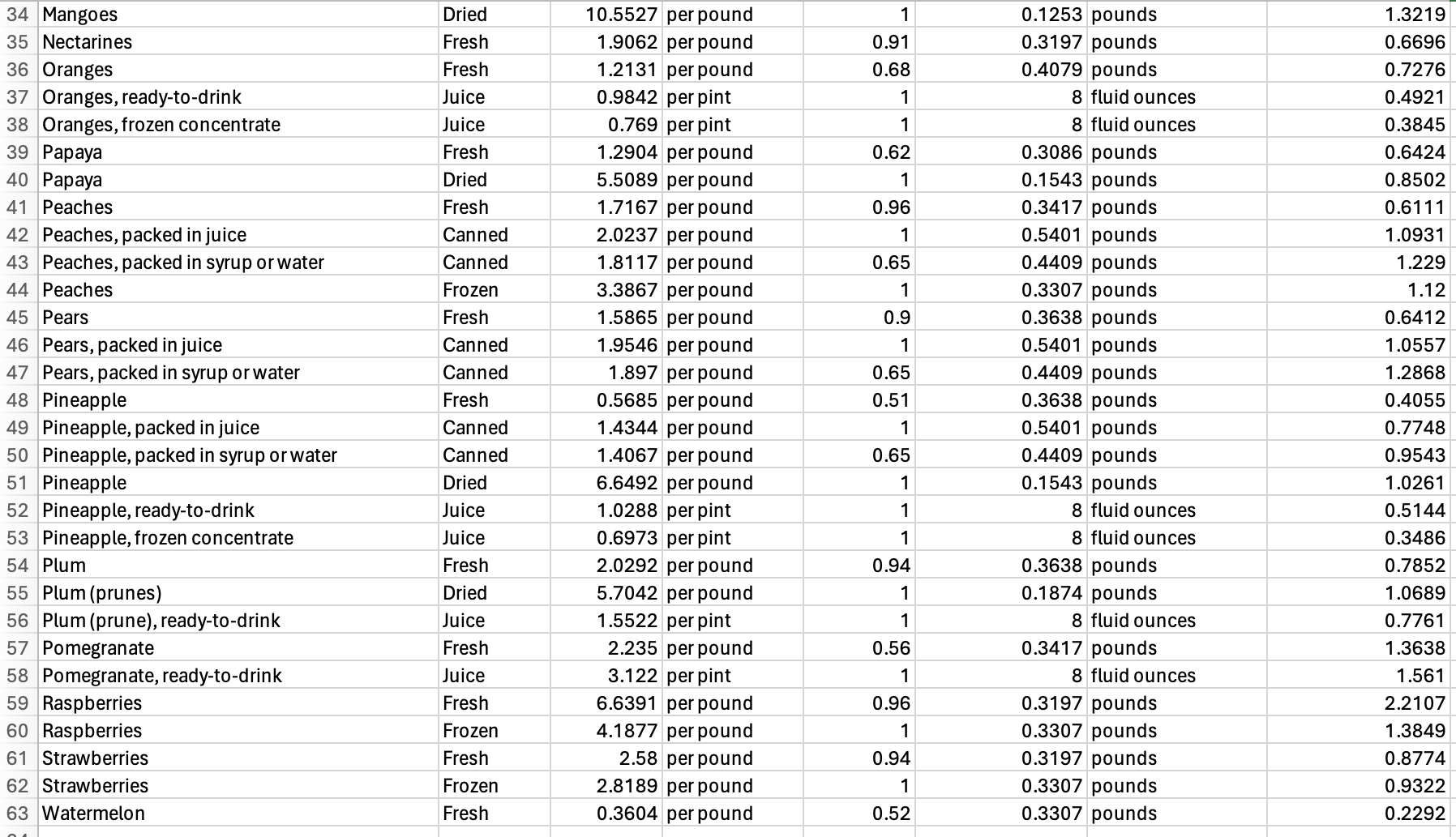
**Future Direction**

While the current analysis was able to provide me with information I was not aware of before, there are still several potential avenues for future exploration and refinement. If I were to continue working on this project, I could see myself performing more detailed and refined analysis techniques as well as including external information. For example, integrating socioeconomic data to analyze the relationship between fruit prices and demographic factors such as income, education, and food access could help me further uncover valuable insights on fruit price trends which in turn could be even more beneficial for consumers, retailers, and policymakers. I could accomplish more refined analysis with more data transformation techniques like data cleaning, data standardization and normalization, and data and time processing.

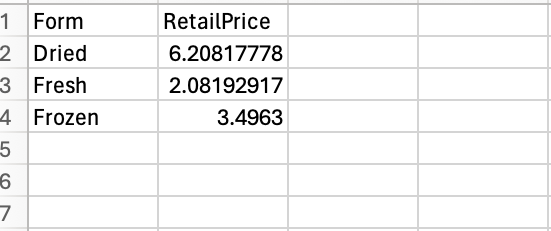
**Appendix**

Dataset in Excel before transformation:

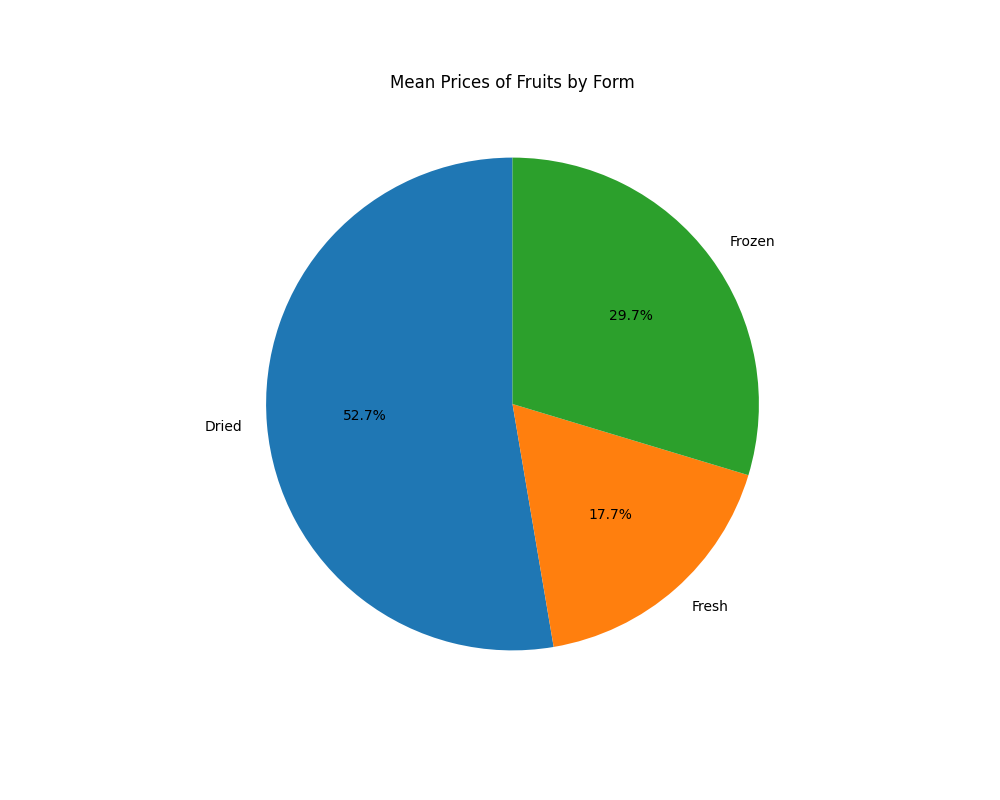




Dataset in Excel after transformations:



Pie Chart:



**References:**

Link to dataset: https://catalog.data.gov/dataset/fruit-and-vegetable-prices