**Trends in Fruit and Vegetable Affordability: A Cost Analysis Over Time**

*Harsh Patel*

**Data Analysis and Visualization 2**

*Farah Huassien*

# Introduction

Access to affordable and nutritious food is a critical aspect of public health and economic stability. Understanding the cost of fruits and vegetables is essential for assessing the affordability of a healthy diet, especially in the context of dietary guidelines and food security. The dataset "Fruits and Vegetable Prices" from the U.S. Department of Agriculture’s Economic Research Service (USDA, ERS) provides detailed estimates of average retail costs for over 150 fresh and processed fruits and vegetables in the United States. These estimates are based on retail scanner data collected from a range of stores, including grocery stores, supermarkets, and convenience stores, for the years 2013, 2016, 2020, and 2022.

The dataset provides insights into the costs associated with consuming fruits and vegetables per edible cup equivalent, considering factors such as inedible portions and cooking loss (U.S. Department of Agriculture, Economic Research Service, 2025). By analyzing this dataset, we can explore trends in food affordability, regional price variations, and the potential impact on dietary habits. The study of such data can inform policymakers, nutritionists, and economists in their efforts to promote healthier food choices and improve food accessibility.

This research aims to investigate whether specific fruit and vegetable categories have become affordable over time. By analyzing price trends across different years, we can identify patterns in cost fluctuations and determine which categories have experienced the most significant changes. This study will provide valuable insights into food affordability trends and their implications for consumer purchasing behavior.

# Initial Analysis:

Data Summarization/Visualization:

The dataset obtained from the USDA provides approximations of retail prices for various fruits and vegetables, per-pound costs. Due to the price differences between fruits and vegetables, we separated the data for fruits and vegetables into two separate data frames to help observe any differences between fruits and vegetables over time as well:

Datasets:

* Fruits: [Fruits\_Retail\_Cost.xlsx](https://sheridanc-my.sharepoint.com/:x:/g/personal/pate3837_shernet_sheridancollege_ca/EX1llCvy_79JqafKLmo-xM4BRYKn2fi3FO3ltjGCr-nKTA?e=X4YqHR)
* Vegetables: [Veg\_Retail\_Cost.xlsx](https://sheridanc-my.sharepoint.com/:x:/g/personal/pate3837_shernet_sheridancollege_ca/EQm9FRQA7I1AjsLjTXmo96QBOsnOeQ_i_2XxB4mWMGyVNQ?e=l142l7)

Descriptive Statistics:

Fruit:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Year:** | **2013** | **2016** | **2020** | **2022** |
| **Mean:** | 2.43 | 2.532 | 2.6362 | 3.018 |
| **IQR:** | 2.46 | 2.53 | 2.4072 | 2.82 |
| **Min:** | 0.33 | 0.320 | 0.3604 | 0.382 |
| **1st Quartile:** | 1.04 | 1.050 | 1.1513 | 1.337 |
| **Median:** | 1.65 | 1.790 | 1.897 | 2.179 |
| **3rd Quartile:** | 3.5 | 3.580 | 3.5585 | 4.157 |
| **Max:** | 8.5 | 10.160 | 10.5527 | 10.303 |

Vegetables:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Year:** | **2013** | **2016** | **2020** | **2022** |
| **Mean:** | 1.768 | 1.781 | 1.8922 | 2.113 |
| **IQR:** | 1.105 | 1.072 | 1.048 | 1.232 |
| **Min:** | 0.560 | 0.6 | 0.6682 | 0.797 |
| **1st Quartile:** | 1.097 | 1.080 | 1.1542 | 1.267 |
| **Median:** | 1.465 | 1.485 | 1.6864 | 1.854 |
| **3rd Quartile:** | 2.203 | 2.152 | 2.2027 | 2.5 |
| **Max:** | 5.860 | 6.070 | 6.7045 | 6.821 |

Data Types:

* Fruit/Vegetable item (example: Apple, Avocado, Lettuce)
* Year of price
* Form (Fresh, Canned, Frozen, or Dried)
* Cost (per pound or pint in the case of juices)

Analysis:

Based on the data shown from the descriptive statistics, it can be observed, in both fruits and vegetables, that there is a slight increase over the years in price. For fruits, the average price goes from $2.46 to $3.02, a 22.7% increase. Another important note is that the largest jump in the average cost occurs from 2020 to 2022, going from $2.64 to $3.02. Likewise, this larger increase in average cost is also reflected in vegetable prices, going from $1.89 to $2.11 in 2020 to 2022, respectively, again the largest increase during the COVID pandemic years. This will be further analyzed in the charts review and hypothesis testing.

Another note is that these prices are being generalized and might now all have increased the same amount from 2013 to 2022. For example, Watermelon is very cheap at $0.33 and only increased to $0.38, a 15% increase. Whereas dried Papaya went from $4.57 to $7.48, a 63% increase. So, the increase in price will not be evenly distributed among all fruits and vegetables. Also, the different forms of fruits/vegetables have their prices listed separately, as in many cases, dried items tend to be the most expensive.

Data Graphs:

Each graph lists all prices of fruits and vegetables for each year. Upon initial observations, some items did not appear in the different years' datasets, and were removed accordingly as their price difference in years cannot be observed properly:

Fruits:

A graph of multiple colored lines

AI-generated content may be incorrect.A graph with different colored lines

AI-generated content may be incorrect.A graph of a graph

AI-generated content may be incorrect.A graph with different colored lines

AI-generated content may be incorrect.

Vegetables:

A graph with multiple colored lines

AI-generated content may be incorrect.A graph with multiple colored lines

AI-generated content may be incorrect.A graph with different colored lines

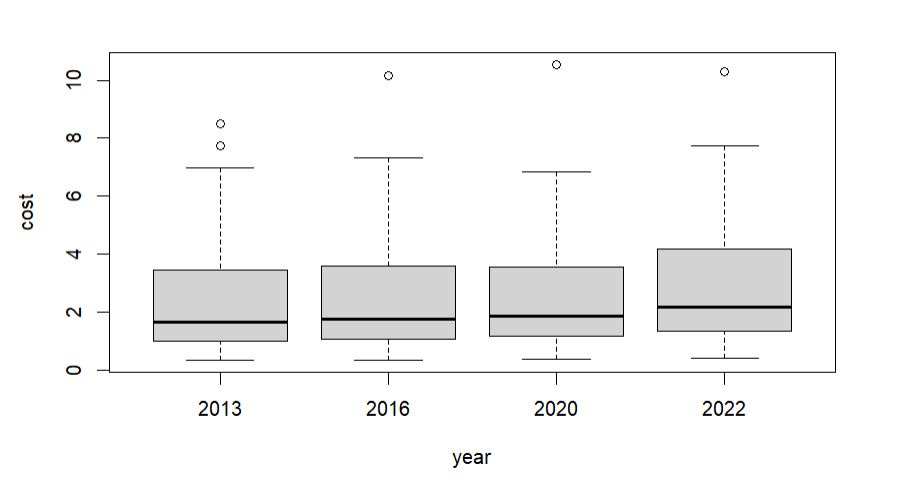
AI-generated content may be incorrect.A graph with multiple colored lines

AI-generated content may be incorrect.

Boxplot visualization:

Fruits:

When a boxplot representing the prices of fruits each year is shown, apart from the 2020 to 2022 gap, the increases in costs appear much smaller, with dried mangoes appearing as a strong outlier every year. However, the cost jump from 2020 to 2022 is very apparent here, shifting the first quartile, median, and third quartile up a noticeable amount.



Vegetables:

When a boxplot representing the prices of vegetables each year is shown, a similar jump can be observed from 2020 to 2022.

# 

# The vegetable data also contained many more outliers, which may affect the mean price to appear higher, however, there is only a small difference in the median and mean for vegetables throughout the years. Fruits had a larger difference in median and mean values, with a lower minimum and higher maximum value than vegetables. This shows that fruits tend to have a larger price disparity than vegetables. When generalizing the cost to an average for each year, it may not accurately reflect the cost changes for individually more expensive/cheaper fruits. The data also has gaps between years; having data for foods from the missing years would help give a more accurate picture of the changes in cost. Also, different parts of the world would experience different changes in prices; as such, the observations made on this dataset might not apply the same to people in other parts of the world.

# Hypothesis Testing

Before conducting further analysis, we wanted to check if there existed any price trend for ten years. So, we plotted a line graph to understand the price trend over a timespan of ten years.

A graph with numbers and a line

AI-generated content may be incorrect.

We can see here that the prices for both fruits and vegetables followed a linear trend until 2020. But after 2020 it increased at a steep rate from 2.5 to 2.9 approximately.

But it is not sufficient to say that price increased relatively from 2020 with conducting a proper analysis. So, we decided to conduct a hypothesis test to determine if this steep increase was statistically significant.

## ANOVA Test

We utilized ANOVA test to determine if there was statistically significant evidence that the price significantly increased or not.

The ANOVA test was conducted with response variable being Avg\_Cost and independent variable being Year.

Results:

* *Fruits:*

*A black text on a white background

AI-generated content may be incorrect.*

* *Vegetables:*

*A screenshot of a computer

AI-generated content may be incorrect.*

**Insights**

**1. P-values:**

* Fruits: **p = 0.103**
* Vegetables: **p = 0.107**

Both p-values are **greater than 0.05**, so:

**Fail to reject the null hypothesis** — there's no statistically significant difference in mean prices across years at the 5% significance level.

**2. F-values:**

* F = 8.216 (Fruits), F = 7.855 (Vegetables)
* These are moderately high, indicating some variance is explained by year, but not enough to be statistically significant.

**Conclusion:**

Although there appears to be some change in average cost over the years, the evidence is not strong enough (at the 5% level) to say these changes are statistically significant for either fruits or vegetables.

# Further Clarity

To better clarify the results, we applied a linear regression model to assess whether fruit and vegetable prices have significantly changed over time. To facilitate this analysis, the dataset was transformed from a wide format (where each year had a separate column) to a long format, allowing us to model **Year** as a continuous variable and account for differences across individual fruits and vegetables. This approach enabled us to capture both the overall time-based trend and the unique pricing patterns of each item through interaction terms.

Transformed Datasets:

* Fruits: [Fruits\_Retail\_Cost\_reshaped (1).xlsx](https://sheridanc-my.sharepoint.com/:x:/g/personal/pate3837_shernet_sheridancollege_ca/Efv73Kpmu51InASsf-mFfs8Bwp4pCllj7pfxaVD1J3mqRg?e=n9PRkZ)
* Vegetables: [Veg\_Retail\_Cost\_reshaped (1).xlsx](https://sheridanc-my.sharepoint.com/:x:/g/personal/pate3837_shernet_sheridancollege_ca/EWQH3_RFKUpLiTXzqchdNCEBJHvdYg1e7sn5XxSiMto5LQ?e=nkbsA2)

**Results**

* Fruits:

A black text on a white background

AI-generated content may be incorrect.

* Vegetables:

A close-up of numbers

AI-generated content may be incorrect.

**Interpretation of Results**

1. **Fruits:**

1. Multiple R-squared: 0.6959

* About 69.6% of the variation in fruit retail costs is explained by the model.
* This is a strong R-squared, suggesting that the model captures most of the pricing patterns.

2. Adjusted R-squared: 0.4722

* 47.2% of the variation is still explained after penalizing for model complexity — moderate to good fit.
* The drop from 69.6% to 47.2% suggests that some variables may not be adding much predictive value

1. p-value = 2.948e-10
   * highly significant (< 0.05), meaning the model overall is statistically significant.
2. **Vegetables:**
3. Multiple R-squared: 0.7876

* The model explains 78.8% of the variance in vegetable retail costs — very strong explanatory power.

1. Adjusted R-squared: 0.6753

* This is a high adjusted R², suggesting that even with many predictors, the model remains robust.

1. p-value < 2.2e-16
   * highly significant (< 0.05), meaning the model overall is statistically significant.

The linear regression models developed for both fruits and vegetables provide strong statistical evidence that retail prices have changed over time, with patterns varying by item type. The fruit model achieved a multiple R-squared of 69.6% and an adjusted R-squared of 47.2%, indicating a moderate to good fit despite the model's complexity.

The lower adjusted R-squared in the fruit model may be partly attributed to a smaller sample size compared to the vegetable model, which can reduce the model's ability to generalize. In contrast, the vegetable model performed even better, with a multiple R-squared of 78.8% and an adjusted R-squared of 67.5%, suggesting strong explanatory power. In both cases, the F-statistics were highly significant (p < 0.001), confirming that the combination of year, item type, and their interaction significantly influences retail prices.

These results indicates that pricing trends differ across fruits and vegetables and have evolved over time.

To check if the model fits well will the data before finalizing the insights we conducted normality assumption test for the model.

## Normality assumption for model:

Fruits:

A graph of a number of objects

AI-generated content may be incorrect.

Vegetables:

A graph of a number of objects

AI-generated content may be incorrect.

**Interpretation of Residual-Leverage Plot**

**1. Fruits**

1. **Standardized Residuals:**
   1. Most residuals lie between -2 and +2, indicating the model generally fits well.
   2. However, a few points like 24, 21, and 276 are above +2 or +4, marking them as potential outliers.
2. **Cook’s Distance:**
   1. Cook’s distance line is at 0.5.
   2. Points outside this curved area are potentially influential data points (especially point 276 could be influential).
3. **Trend Line:**
   1. The red line is relatively flat, suggesting no clear pattern in the residuals with respect to leverage — which is a good sign.
   2. A curving trend might indicate non-linearity, which we don't see here.
4. **Vegetables**
5. **Standardized Residuals:**
6. The majority of data points lie near the center with standardized residuals between -2 and +2.
7. This suggests that for most fruits, the model’s predictions are fairly accurate.
8. **Cook’s Distance:**
   1. Cook’s distance line is at 0.5.
   2. Points outside this curved area are potentially influential data points (especially point 128 could be influential).
9. **Trend Line:**
   1. The red line is relatively flat, suggesting no clear pattern in the residuals with respect to leverage — which is a good sign.

The residuals vs leverage plot for the fruit model shows that while most observations fit the model well, a few have both high leverage and high residuals.

# Final Insights

Although the initial ANOVA results did not show a statistically significant difference in mean prices across the selected years (p > 0.05), this was based on a simplified model using only four yearly averages — limiting statistical power due to the small sample size (df = 2).

In contrast, the linear regression models, built on reshaped and expanded data across all fruits and vegetables over time, revealed significant relationships between year, item type, and their interactions. These models used a much larger number of observations, increasing statistical sensitivity and allowing us to detect nuanced trends.

Thus, while the ANOVA suggested no strong average-level differences between years alone, the linear regression indicates that price trends do exist, especially within specific fruits or vegetables, and those trends vary over time. This shows that aggregated ANOVA results may mask item-specific variations that a regression model is better suited to uncover.

# Future Improvements

Following feedback from the professor, we conducted a more in-depth assessment of our linear regression models by testing the normality assumption of residuals — a key requirement for reliable inference in linear models. Using the Shapiro-Wilk test, we found that the residuals for both the fruit and vegetable models yielded extremely low p-values (< 0.05). This led us to reject the null hypothesis of normality, indicating that the residuals are not normally distributed.

**Shapiro Test:**

A screenshot of a computer code

AI-generated content may be incorrect.

To visually confirm this, we generated Q-Q (quantile-quantile) plots for both models:

**QQ Plots:**

Fruits:

A graph of a normal q-q plot

AI-generated content may be incorrect.

Vegetables:

A graph with a line and numbers

AI-generated content may be incorrect.

**Fruit Model**

* The Q-Q plot exhibited strong deviations from the diagonal line, particularly at both tails of the distribution.
* This suggests a heavy-tailed distribution, confirming the results of the Shapiro-Wilk test.
* The presence of multiple points diverging significantly (below -1 and above +2) highlights a lack of normality and potential outliers.

**Vegetable Model**

* The Q-Q plot also showed deviation from normality, though less pronounced than in the fruit model.
* Some mild skewness and heavy tails were observed, suggesting the residuals are closer to normal, but still not ideal.

For future model improvements, we suggest:

1) Transforming the retail cost data using a logarithm to address skewness and improve normality.

2) Employing robust regression to minimize the impact of outliers and enhance estimate reliability when data isn't normally distributed.

# References

U.S. Department of Agriculture, Economic Research Service. (2025). *Fruits and Vegetables Prices. Retrieved from* [*https://www.ers.usda.gov/data-products/fruit-and-vegetable-prices*](https://www.ers.usda.gov/data-products/fruit-and-vegetable-prices)