

---

# Statistical and Correlation Analysis of Avocado Sales Volume and Prices in the US Market, Insights for Uncovering Key Factors of Statistical Data Discovery

---

**Cheima Ali Bensaad**

*Dr. University of Hertfordshire Computer science- Data science department, UK*

*ORCID ID:0000-0003-1760-2940*

*Email: [Cheima.alibensaad@gmail.com](mailto:Cheima.alibensaad@gmail.com)*

---

**DOI:** <https://doi.org/10.19275/RSEP186>

---

**Article Type:** Original/Research Paper

---

## **Article History**

Received: 30 April 2024    Revised: 5 June 2024    Accepted: 14 June 2024    Available Online: 30 June 2024

---

**Keywords:** Avocado sales and prices, Correlation analysis, statistical distribution, Non-parametric test.

**JEL classification:** Q11, C82, C15

---

**Citation:** Cheima, A. B. (2024). Statistical and Correlation Analysis of Avocado Sales Volume and Prices in the US Market, Insights for Uncovering Key Factors of Statistical Data Discovery, *Review of Socio-Economic Perspectives*, 9(1), 199-205.

---

**Copyright © The Author(s) 2024** This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

---

## **Abstract**

Avocado consumption has witnessed rapid growth because of the change in lifestyle, leading to significant fluctuations in sales volumes and prices. The research delves into a statistical analysis of avocado sales in the US market from 2015 to 2016, with a specific focus on exploring the parameters of the statistical correlation between avocado sales volume and their prices. The primary research consists of comprehensive data analysis and visualisation to understand the statistical distribution, the trend and the relationship between avocado sales volumes and prices across various US markets using R language. The correlation analysis with non-parametric statistical Spearman test reveals a significant negative correlation between the two variables and rejects the null hypothesis based on no correlation between those variables. However, it is emphasised that correlation does not imply causality, which secondly interests us in uncovering the key factor behind the findings.

The research has not only identified the major factors and principles of the need for efficient statistical methods and techniques to enhance knowledge discovery in terms of data collecting, handling, and preprocessing the data but also highlighted unusual sale-price patterns and distribution that require more investigation in terms of quality, consistency, and interpretability. This underscores the necessity of conducting more research to find optimal prices better and understand the factors influencing avocado sales volumes and market tendency.

---

## 1. Introduction

This study is highly relevant in today's economic context and, thus, understanding aspects of the market dynamics of avocado sales concerning their pricing and significant implications. Avocados have become a core ingredient in many US households, with their popularity driven by their perceived health benefits and versatility in various culinary applications (Johnson,2016). This surge in demand has led to price volatility and supply chain disruptions (Kim ,2017) and, therefore, the need for a deeper understanding of the factors influencing avocado sales (Smith,2020). This study aims to provide a comprehensive statistical and correlation analysis of avocado sales volumes and prices, incorporating various fields. Through visualization and statistical analysis, including descriptive, correlation, and non-parametric methods, we aim to discover and extract key knowledge factors. By combining these methods with economic and field knowledge interpretation, we achieve more rounded and accurate results. Focusing on the sales aspect, the study follows the sequential phases of the data analytics lifecycle.

This research offers insights into the market dynamics of avocado sales and pricing, with significant implications. Utilising historical data on avocado prices and sales volumes across multiple U.S. markets, this study seeks to answer key research questions about the correlation between price and sales volume as *"Is there a correlation between the avocado sales volume and avocado prices in multiple US regional markets from 2015 to 2016?"*. To guide this study, we have formulated the following hypotheses: the null hypothesis (H0), *"There is no correlation between the avocado sales volume and avocado prices in multiple US regional markets from 2015 to 2016."* while alternative hypothesis (H1) : *"There is a correlation between the avocado sales volume and avocado prices in multiple US regional markets from 2015 to 2016"*.

This study follows the CRISP-DM Cross-Industry Standard Process for Data Mining methodology (Paul Moggridge. 2024) and data visualisation cycles , starting from understanding the data and problem, followed by data preparation and preprocessing, through to statistical analysis, and finally testing and validating the findings. The paper is structured into sections that provide a summary of the data description and visualization, followed by data analysis, and includes an in-depth interpretation of the findings along with the limitations and suggestions for future research. All steps were performed using R code in RStudio version 4.3.1. The R programming language is widely used for statistical analysis and data visualisation in agricultural economics (Wiley, 2021).

## 2. Literature Review

Before delving into the methodology and findings of this study, it is essential to first review the existing literature on avocado market dynamics, including previous research on sales volume, pricing strategies, and consumer preferences.

Building upon Feldman's insights and Dominy, N. (2016), they have conducted a qualitative and descriptive analysis of historical avocado marketing strategies on U.S. consumer demand. Investigate the impact of Strategic marketing, including health benefit promotions which found an increase of avocado consumption after an adaptive marketing strategy effectively boosted consumer demand. As highlighted by Fuerte and al(2019), they combined quantitative surveys with choice experiments. Utilised statistical techniques such as logistic regression to analyse survey data and estimate willingness to pay. Spanish consumers value locally produced and organic avocados. Younger consumers are willing to pay more for premium attributes. In line with the findings of Matthews and Murphy (2013) , they analyse the price elasticity of demand for avocados in the U.S and investigate the price elasticity of demand for avocados and how consumers respond to price fluctuations. Has found Demand is price elastic, with significant responsiveness to price changes, especially during off-peak seasons. It has been suggested that Producers and retailers should consider pricing strategies carefully to avoid large drops in sales. Targeted promotions can help stabilise demand. In term of technical concept, Stephen Few (2009). study emphasis on simplicity, clarity, and integrity in visualisations is a valuable reminder that often the most basic tools are the most effective. He focuses on a variety of visualisation techniques and advocates for the use of graphs and charts such as scatterplot and histograms, arguing that these tools are often sufficient for most data analysis tasks.

The research gap and difference between the current study and previous work are that our study employs correlation analysis, such as Pearson or Spearman correlation depending on data distribution, to uncover relationships between sales volume and prices. In contrast, previous studies took different approaches: Feldman (2016) focused on qualitative analysis of marketing strategies; Fuerte et al (2019) examined consumer preferences and willingness to pay; and Matthews & Murphy (2013) analysed price elasticity. This study provides a quantitative analysis of the correlation between avocado sales volume and prices, offering a new perspective on market dynamics.

### 3. Data Understanding and Preprocessing

This section consists of the first stage of data analytics cycle including data description, understanding and preprocessing. The dataset is based on historical data on avocado prices and sales volume in multiple US markets and contains the average prices of avocados in dollars (\$) and their corresponding total volume sales (quantity) over a year aiming to gain insights into the relationship between quantity sold and prices. Using the **dataset avoado** which was imported directly from the Kaggle platform and originally, (Kiggins, 2018).

Initial data(row data) includes 18250 rows (observations) which reflects the weekly information covering a 4-year period starting in 2015 and ending in 2018. Furthermore, the data involves 12 features (columns) including, prices, volume, region, type of avocado, and quantity of bags sold by type and size. After the data preprocessing step, the processed dataset we are looking into becomes 53 weeks and covers 2 features as final process of data dimensionality by feature selection .

**Tabl. 1:** Overview of the Avocado dataset table head highlighting target variables:

Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge	Bag type	year	region
29/11/2015	1.94	831.69	0	94.73	0	736.96	736.96	0	0	organic	2015	Albany
15/03/2015	1.63	1777.09	0	1209.68	0	567.41	366.67	200.74	0	organic	2015	Boise
01/02/2015	1.43	1780.76	0	999.31	0	781.45	283.33	498.12	0	organic	2015	Boise
11/01/2015	1.44	2378.68	0	1923.4	0	455.28	170	285.28	0	organic	2015	Boise
27/12/2015	1.47	5043.15	0	166.37	0	4876.78	2751.87	2124.91	0	organic	2015	BuffaloRochester
13/12/2015	1.45	4317.94	0	130.62	0	4187.32	1502.72	2684.6	0	organic	2015	BuffaloRochester
22/03/2015	1.93	530.96	0	147.63	0	383.33	383.33	0	0	organic	2015	BuffaloRochester
11/01/2015	1.59	2078.49	0	143.51	0	1934.98	1934.98	0	0	organic	2015	BuffaloRochester
04/01/2015	1.73	379.82	0	59.82	0	320	320	0	0	organic	2015	BuffaloRochester
17/05/2015	1.85	1338.84	0	969.48	0	369.36	356.67	12.69	0	organic	2015	GrandRapids
19/04/2015	1.59	1377.1	0	963.77	0	413.33	413.33	0	0	organic	2015	GrandRapids
12/04/2015	1.86	823.5	0	606.83	0	216.67	216.67	0	0	organic	2015	GrandRapids
26/07/2015	1.03	2995.57	0	1295.04	0	1700.53	40	1660.53	0	organic	2015	Louisville
16/07/2015	1.18	3688.04	0	1436.55	0	1233.46	523.23	1160.13	0	organic	2015	Louisville

**Source:** Kiggins, 2018, [www.kaggle.com/datasets/neuromusic/avocado-prices](http://www.kaggle.com/datasets/neuromusic/avocado-prices)

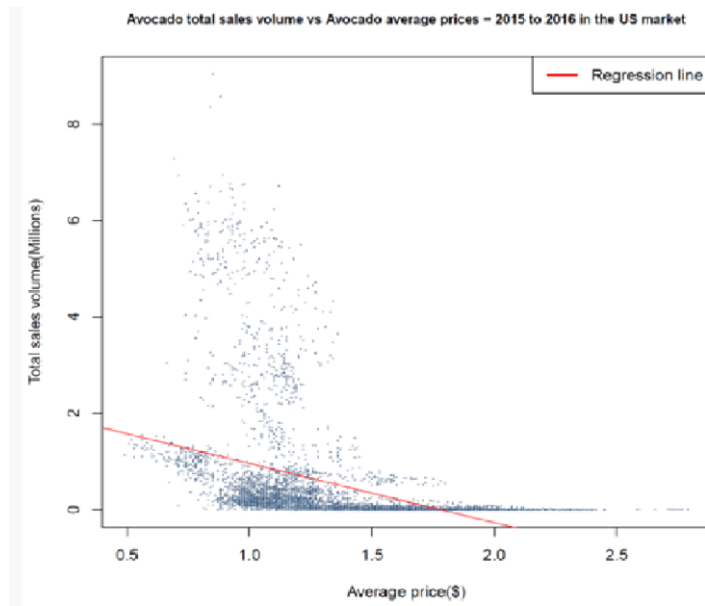
The columns in yellow highlighting target variables. Summarising and describing the main features of a dataset consists on calculating the average total price and mode both equal to \$1.58 with the mode is and average total sales volume is 3251.646 .

To clean and preprocess the data, we undertook several steps. First, we ensured that the selected variables contained no missing values. We then addressed outliers and noise by identifying high and low sales volume values. Initial data representation revealed numerous outliers, necessitating data normalization and standardisation to ensure scalability and address these anomalies. Additionally, we counted repeated values to determine if the dataset was imbalanced. Throughout this process, we utilised the R programming language to execute all data cleaning and preprocessing tasks efficiently. These steps ensured that the dataset was clean, balanced, and prepared for further analysis.

### 4. Data Visualization

The visualisation step for this data cycle is conducted to explore and represent visually the dataset features; the histogram and scatter plot are the most suitable statistical tools/techniques (Roxy Peck.2021) to display and identify the frequency, trends, and patterns and visualise the relationship.

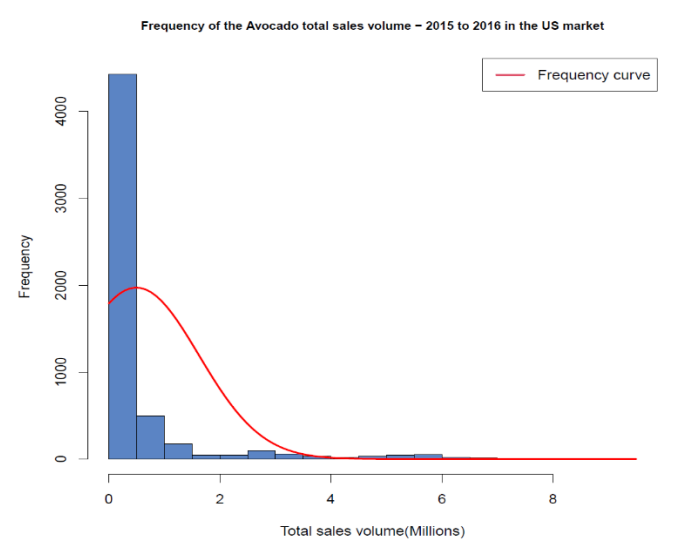
Firstly, we aim to analyse the data visually without relying on complex modelling. Simple models, such as linear regression, are often the most effective for understanding trends and data correlations. By fitting a line to the preprocessed data points, we can inspect the trend, as illustrated in the following scatter plot of total sales volumes against various prices. This approach helps identify how price changes correlate with sales volumes and aids in modelling the relationship between the dependent variable and the independent variables.



**Fig. 1** Scatter plot of the avocado total volume sale and average prices using RStudio 4.3.1

**Source:** Authors calculation and plotting using avocado dataset from Kiggins, 2018 and Kaggel database.

We assume using a function generated by a true underlying line (trendline) for:  $y = \beta_0 + \beta_1x$  (Stephen, 2009), represented as: Total sales volume =  $\beta_0 + \beta_1$  average price. This line indicates visually the overall trend. The **scatterplot** shows a decreasing line. We can notice that most points (p, vol) are gathering around the 0.5 million of total sales volume for the price of \$1 which may indicate the best price for more quantities sold. Furthermore, this graph reveals some outliers despite the data cleaning process. These outliers are represented by values significantly deviating from the expected range, with some data points showing high volumes between 4 and 6 million units. However, these points still cluster around the price range of \$1 to \$1.5.



**Fig. 2** Histogram of the frequency and data distribution of avocado total volume sale using RStudio 4.3.1

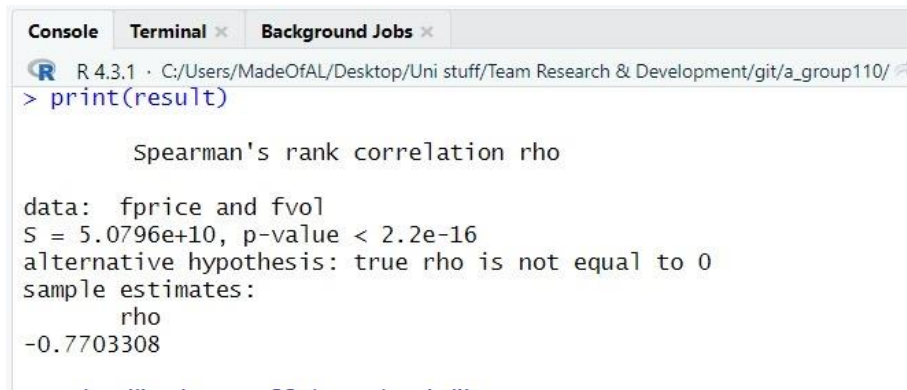
**Source:** Authors calculation and plotting using avocado dataset from Kiggins, 2018 and Kaggel database.

The **Histogram** depicts the frequency distribution of the target variable (total avocado sale volume). This histogram shows that this data distribution is skewed toward the right, rather than following the usual “bell curve” of a normal distribution as would be expected for such market analysis (Kulick and Wright 2008)[5]. The normal red curve overlay does not follow the shape of the underlying data, so the data is not normally distributed and in fact, does not have a symmetric parametric. The skewness and shape reveal more occurrences of lower volume of sales as data is concentrated on a higher level on the left side. For instance, this could be due to a situation where observation clusters at lower volume, but there are a few instances of extremely high volume.

We have tried to Examine how different regions in the U.S. market may exhibit varying degrees of sensitivity to price changes and correlation by contributing to further plots , however the results are mostly same as the general result .

### 5. Data Statistical Analysis:

This section represents the statistical analysis step for the DM-KDD data cycle using inferential statistics technique such hypothesis testing, determining the statistical method the most appropriate and in line with our previously expressed research question is mainly related to the aim to grasp the relationships we want to explore along with the type of data distribution that we got from the visualisation step. The dataset is not normally distributed the non-parametric test commonly and widely used is **Spearman's Rho** test (John Noll, 2023), the following outcome of the results by using **R** are shown in this figure:



```

Console Terminal x Background Jobs x
R 4.3.1 · C:/Users/MadeOfAL/Desktop/Uni stuff/Team Research & Development/git/a_group110/
> print(result)

Spearman's rank correlation rho

data:  fprice and fvol
S = 5.0796e+10, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
-0.7703308
. . . . .

```

**Fig. 3 :** Hypothesis test and correlation coefficient using R code of R studio 4.3.1.

**Source:** Author statistical calculation and testing, codes from Dr Noll and all (2023) , Handbook Team research and development materials, University of Hertfordshire .

The P value reported by the R is **p-value < 0.05**, so we **reject the null hypothesis (H0)**. The result of the hypothesis test proves statistically that there is a correlation between both variables represented by **the alternative hypothesis (H1)**. In addition to this analysis, the outcome of calculating the correlation coefficient helps to assess the strength and direction of the relationships of the features: The correlation coefficient **r** (sales, Total volume, Vol) is related to X the independent variable prices noted as (Average price) , (Dalgaard,2008). **r Corr= -0.77**, which indicates **a significant strong negative correlation between the avocado sales volume and avocado prices**. While the avocado price increases with one unit the sales decrease by 0.77 unit.

### 6. Result discussion and conclusion:

Sales and prices in this data are significantly negatively correlated which means while prices have an upward trend increasing the sales volume will be the inverse decreasing which in turn proves and demonstrates the sensitivity of the market to price fluctuations.

This result complies with the mechanisms of the market supply and purchase/demand which gives the sense that one variable is generally impacting on the other and has a sense of causality. In addition, our results can prove the purchase/sale elasticity of the product (Brown,2019). However, for our avocado dataset, this is not systematically valid for the causality concept as not Just because there is a significant correlation between the two variables, it does necessarily imply a causal relationship between them.

Avocado prices can be influenced by various external factors such as weather conditions, demand, supply, and regional market dynamics. For instance, when delving into row (initial) data we can observe challenges related to the fact that some weeks have the same quantity of avocados sold at varying prices within the same year. Conversely, there are instances where identical prices result in different quantities sold. Furthermore, a cheap unit price of \$1 or a high price of \$1.5 generates different quantities that have been sold/purchased less and more, respectively. This inconsistency indicates a lack of adherence to a price optimization pattern, suggesting potential influences such as seasonality as this data is related to the time, or fluctuations in supply and demand within the market (Garcia,2018).Those kind of challenges in understanding data the question of the quality, consistency, and interpretability' s principles.

In addition, the outliers that are figuring in this data can be also explained by those factors. In this research we have proceed to data cleaning and preprocessing of those outliers which has been delt with dimentiality reduction, feature selection and data scaling techniques.

Furthermore, the data exploration and mining for knowledge extraction have revealed a hidden key factor that emphasises in turn the importance of deploying the principles of KDD (Paul Moggridge, 2024) and the key statistical techniques to implement . It seems there are biased variables (information) that can affect the overall result related to the data itself, for example: the type /size of bags collected and sold does not give details on what was taken into consideration for the data collection phase, the fact that each avocado price is an average price and sale is represented as the total average of the different bags can also smoothen or bias the statistical analysis. This may raise the challenge of the methods employed in s data collection step, such as how retailers count their items. Recognising the importance of well-structured data driven by selecting the most relevant data collection techniques and methods can enhance the data/business understanding, processing, and analysis steps of the data mining cycle.

## 7. Conclusion

Overall, this work has provided a visualisation, technical, statistical, and economic interpretation of the results, along with potential challenges and further research suggestions. The research has offered valuable insights into avocado sales, price fluctuations, and the impact of dataset analysis. The visualisation reveals hidden, unexplained, and unexpected factors influencing the relationship between price and sales volume, emphasising the importance of effective visualisations in uncovering trends and insights that might be missed through numerical analysis alone. Although the analysis reveals significant correlation, it does not automatically imply causality. Accurate data is foundational to reliable analysis, and statistical methods must be robust against data imperfections, ensuring that results are clear, interpretable, and actionable.

Additionally, this research underscores the necessity of integrating field knowledge to understand differences in findings regarding price and volume. Previous research, such as Mathews (2013), suggests that demand is price elastic, with significant responsiveness to price changes, especially during off-peak seasons. However, our analysis reveals instances where the volume demanded does not correspond to the same price elasticity. This discrepancy highlights the importance of considering major factors and principles, including data quality, consistency, complexity, and volume. Efficient statistical methods are crucial to processing and analysing information effectively, ensuring data is handled quickly and accurately without sacrificing precision Stephen Few (2009).

In conclusion, as a summary of the conducted analysis and visualisation steps, future research can further explore correlations by including additional data features and introducing predictive models and optimisation techniques. Time series analysis can be employed to identify seasonal patterns, and correlation analysis can be compared across single region markets. By understanding how price impacts demand, this project aids strategic planning for both individuals and retailers, assisting in decisions about the best quantities to sell or purchase. Determining the optimal price for maximising sales is a complex issue that cannot be answered solely based on the dataset. However, by integrating the key factors discussed and employing appropriate statistical methods, we can enhance knowledge discovery and make informed, data-driven decisions about the correlation between avocado sales volume and prices.

## References

- Brown, M. E., & Green, S. R. (2008). *Exploring Seasonal Patterns in Avocado Prices: A Time Series Approach*. *IEEE Transactions on Food Science and Technology*, 55(3), 789-800.
- Dalgaard, Peter. (2020). *Introductory Statistics with R*. 2nd ed. Statistics and Computing. Springer.
- Feldman, D., & Dominy, N. (2016). The Evolution of Avocado Marketing in the United States. *Journal of Food Distribution Research*, 47(2), 1-8.
- Fuerte, P., Alcon, F., & Garcia Cohegrus, P. (2019). Consumer Preferences and Willingness to Pay for Avocados in Spain. *Spanish Journal of Agricultural Research*, 17(1), e0105.
- Garcia, L. H., & Martinez, E. D. (2018). *Consumer Preferences and Willingness to Pay for Avocado Attributes: A Market Research Study*. *IEEE Transactions on Consumer Behavior*, 42(2), 567-578.
- Johnson, P. A., & White, C. R. (2016). *Avocado Price Forecasting Using Regression Analysis: A Case Study of US Markets*. *IEEE Transactions on Business Analytics*, 21(1), 234-247.
- Kiggins, J. *Avocado Prices: Historical data on avocado prices and sales volume in multiple US markets*. Available at: <https://www.kaggle.com/datasets/neuromusic/avocado-prices/> 2018 , (Accessed: 29 December 2023).
- Kim, Y., & Lee, S. (2017). *Optimising Avocado Prices for Sustainable Supply Chain Management*. *IEEE Journal of Sustainable Agriculture*, 33(4), 789-802.

- Matthews, W. A., & Murphy, E. (2013). Price Elasticity of Demand for Avocados: An Analysis of Consumer Response to Price Changes in the U.S. Market. *Agricultural and Resource Economics Review*, 42(1), 50-64.
- Noll John, Sarah Beecham. (2023). Team research and development materials, University of Hertfordshire, UK.
- Paul Moggridge. (2024). Data mining module materials, University of Hertfordshire, UK.
- Roxy Peck, Chris Olsen, I Jay L. Devore.(2021). *Introduction to Statistics and Data Analysis*. 4th Edition. Brooks/Cole.
- Smith, J., & Johnson, A. Analyzing Avocado Price Trends in the US Market.(2020). *IEEE Transactions on Agriculture*, 67(5), 1234-1245.
- Stephen Few.(2009). *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press.
- Wiley R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org>.

