­

**Multiple Linear Regression Analysis on Superstore Sales Data**

Nicole Reiswig

College of Information Technology, Western Governors University

Dr. William Sewell

April 24, 2024

**Table of Contents**

Research Question 3

Data Collection 5

Data Extraction and Preparation 7

Analysis 24

Data Summary and Implications 32

References 34

**Research Question**

This analysis aims to answer the research question: "Can a multiple linear regression model be constructed based solely on the research data?" The research question will be answered utilizing multiple linear regression modeling. The goal of this analysis is to determine which variables in the given data set have a statistically significant impact on sales revenue. In determining which variables are most statistically significant in predicting sales revenue actionable insights can be made to increase sales revenue. The Null hypothesis is, "A predictive regression model cannot be constructed from the superstore sales data." The alternate hypothesis is, "A predictive regression model can be constructed from the superstore sales data."

Multiple linear regression is an appropriate technique to use for analyzing the research question because (Zach. 2020) multiple linear regression is the most commonly used technique in statistics and it is used to quantify the relationship between predictor variables and response variables. In this specific analysis, multiple linear analysis is used to quantify the relationship between the predictor variable Sales which is a continuous variable, and all the other variables in the superstore data set which are the independent variables. The multiple linear regression analysis provides insight into how strong a relationship is between the dependent and independent variables. Therefore the multiple linear regression analysis is a great technique for answering business questions with large data sets and finding the most significant factors affecting the business model.

Multiple linear regression has four assumptions according to (Zach, 2020) these assumptions are that there is a linear relationship between the x and y variables, residuals are independent, homoscedasticity, and normality are present. To meet these assumptions there are a couple of procedures that must take place. 1) A linear relationship between the x and y variables means there is a relationship in the form of a straight line. This is verified through bivariate graphs. 2) For the residuals to be independent, there must not be a correlation among the variables, they are independent and random. The residuals are verified through visualizations such as a correlation matrix and heatmap. 3) Homoscedasticity is when the error does not change across the values of the independent variables (Middleton, 2023), homoscedasticity is verified through VIF the variance inflation factor. 4) Lastly, normality is when the residuals are normally distributed and this is checked through univariate graphs commonly a histogram.

The contribution of this study to the field of data analytics and the MSDA program is to create a predictive model that can estimate future sales of the superstore so that the company can be adequately prepared to meet future business needs. This study will utilize a multiple linear regression model to analyze the significant predictor variables for future sales. Dong, Chen, Y., Gu, A., Chen, J., Li, L., Chen, Q., Li, S., & Xun, Q., (2020) found that multiple linear regression analysis was successful in identifying positively correlated relationships in sales data specifically with the time and review variables. In previous studies utilizing multiple linear regression for sales revenue, there has been a positive correlation between sales revenue and review ratings (Dong et al., 2020). Analyzing the superstore sales data through multiple linear regression to answer the research question is crucial because it can be determined which factors have the greatest influence on sales revenue. In knowing which factors have the greatest impact both positive and negative smart business decisions can be made to maximize sales revenue in the upcoming years.

# Data Collection

The data collected is owned by Bhanupratap Biswas, the author of the data is Bhanupratap Biswas a Kaggle expert. The dataset was collected from Kaggle.com and can be located at this web address:

<https://www.kaggle.com/datasets/bhanupratapbiswas/superstore-sales>.

This data set is updated on an annual basis. The superstore dataset has been licensed as ODC Public Domain Dedication and License which allows us to use this public data for this analysis (Open Knowledge, n.d.).

The data collection methodology was straightforward with no challenges once the dataset was selected. Finding the data set consisting of the minimum requirements for this study was a challenge and a disadvantage due to the time-consuming nature of the process. There are numerous amounts of data available to study. However, pages of data must be explored to find one that meets the criteria. The criteria for this analysis meant finding a dataset with a minimum of 7,000 rows of data and then also finding a dataset that could be utilized to answer the research question selected while also meeting the requirements of the chosen analysis technique. To overcome this challenge there were several topics of interest downloaded, cleaned, and explored before selecting this dataset to perform the analysis on. The advantage of this collection method was the ability to find a great dataset to work with for this analysis.

The superstore sales data consists of 9,800 rows, 18 columns, 4,922 unique Order IDs, 1,230 unique Order Dates, 1,320 unique Ship Dates, and 793 unique Customer IDs. The 18 columns in this data set are Row\_ID, Order\_ID, Order\_Date, Ship\_Date, Ship\_Mode, Customer\_ID, Customer\_Name, Segment, Country, City, State, Postal\_Code, Region, Product\_ID, Category, Sub\_Category, Product\_Name and Sales.

|  |  |  |
| --- | --- | --- |
| Field | Data Type | Variable Type |
| Row\_ID | Categorical | Independent |
| Order\_ID | Categorical | Independent |
| Order\_Date | Categorical | Independent |
| Ship\_Date | Categorical | Independent |
| Ship\_Mode | Categorical | Independent |
| Customer\_ID | Categorical | Independent |
| Customer\_Name | Categorical | Independent |
| Segment | Categorical | Independent |
| Country | Categorical | Independent |
| City | Categorical | Independent |
| State | Categorical | Independent |
| Postal\_Code | Numeric | Independent |
| Region | Categorical | Independent |
| Product\_ID | Categorical | Independent |
| Category | Categorical | Independent |
| Sub\_Category | Categorical | Independent |
| Product\_Name | Categorical | Independent |
| Sales | Numeric | Dependent |

The study's limitations are that the dataset does not include certain factors that could have been valuable to the study such as review ratings. This data is not included because it was not collected or required in data collection. According to Chattopadhyay (2016) multiple linear regression analysis increases customer satisfaction and loyalty, predicts the number of purchases made, discovers the relationship between customer wait times and the number of complaints, and is used to estimate customers' needs. The delimitations of the study are that Order\_ID, Customer\_ID, Customer\_Name, Country, and Product\_ID will be removed from the dataset. These fields will be removed because they do not have statistically significant value in this analysis. Multiple Regression (n.d.) recommends that when using multiple linear regression the analysis should be limited to five or fewer independent variables to prevent multicollinearity.

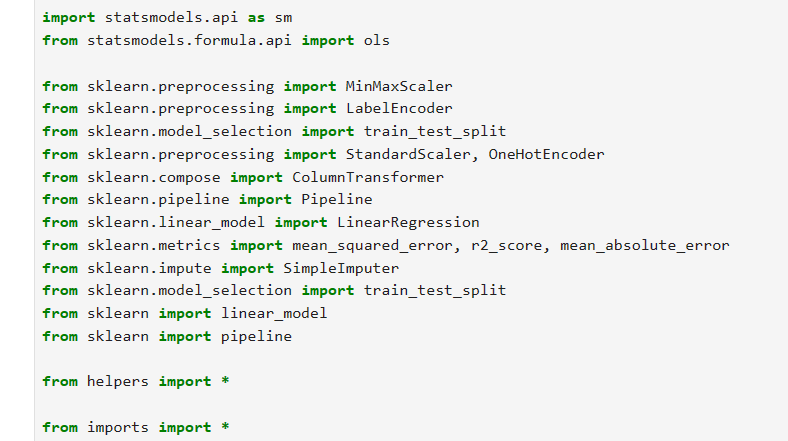
# Data Extraction and Preparation

The data analysis tools/techniques utilized are appropriate for this analysis. Python is an open-source programming language that in terms of usability is the better choice because the syntax it uses is similar to other languages and therefore is more versatile (Ozgur, Colliau, Rogers, Hughes, & Myer-Tyson, 2017). Python is a famous data analysis language used by data scientists and is the preferred language because of its highly interactive nature and its scientific ecosystem libraries (Siddiqui, Alkadri, & Khan, 2017). According to Siddiqui (2017), SAS is an expensive solution and is between a programming language and a scripting language. R is an open-source programming language that is more difficult to learn but is great with data manipulation, statistics, and graphics functionality built-in (Siddiqui, 2017). Ozgur (2017) notes that R is geared more towards statistical roles, resembles SAS, does not rely on a computer science background or coding, and is open-sourced, while Python is geared more towards the coding aspects of jobs. Utilizing Python provides a few advantages one being ease of implementation through libraries such as pandas and scikit learn. Pandas can handle data manipulation and scalability. Scikit Learn is a machine learning library that can perform multiple linear regression analyses. Multiple linear regression can consider multiple factors and produce accurate predictions. One disadvantage of the data extraction and preparation method is that R is geared more towards statistical roles and potentially could have produced the same results with fewer lines of code.

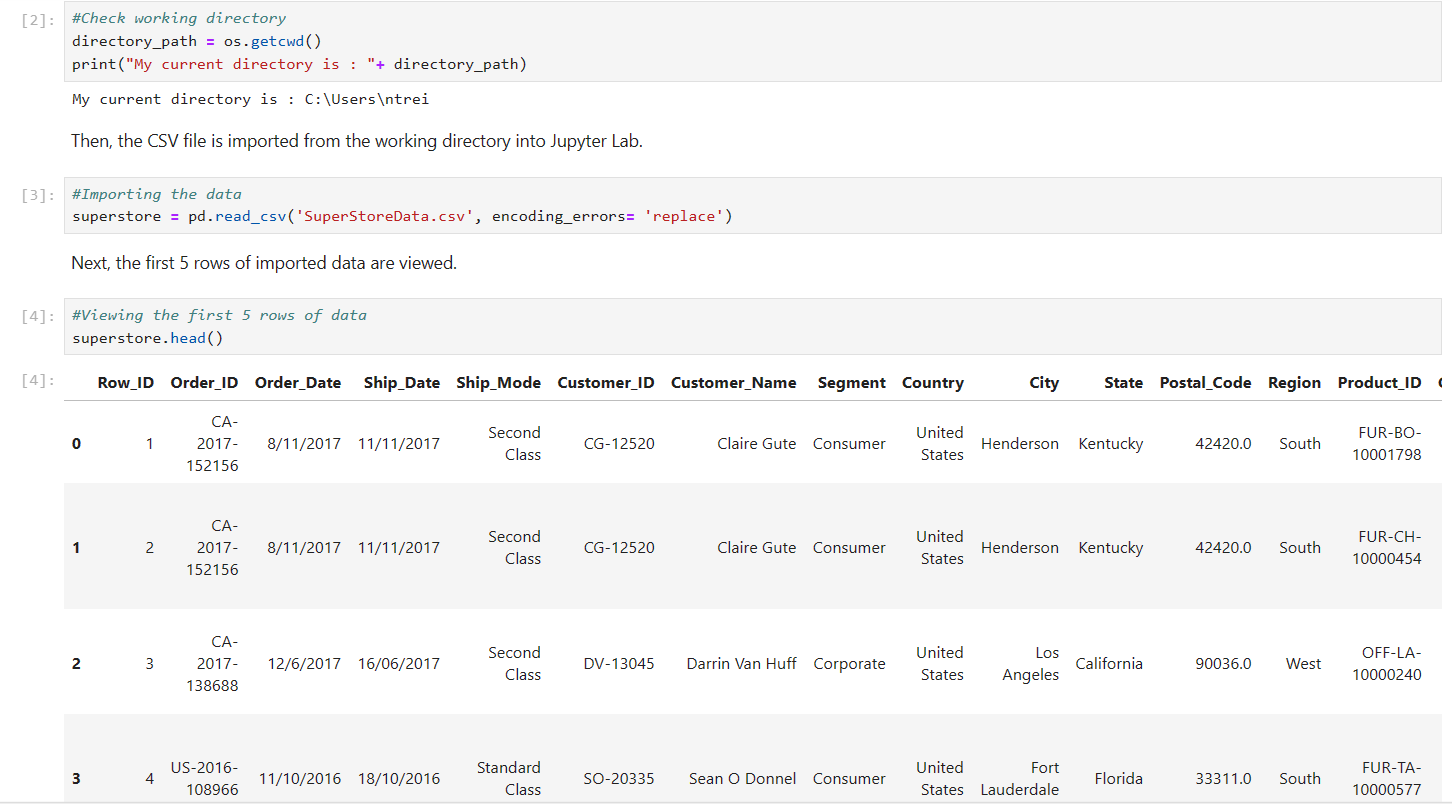
The data was extracted from the publicly available CSV file on Kaggle.com which shows the superstore data set with data only in the United States. The data was imported into Jupyter Lab utilizing Python and Anaconda. The first step once in the lab environment was to import the proper libraries and packages to begin the analysis. Pandas, numpy, matplotlib, seaborn, os, patsy,statsmodels, scipy, missingno, warnings, datetime, tensorflow, sklearn, imports, and helpers were imported. These libraries provide these functions for this analysis, matplotlib provides visualizations, numpy provides arrays, Keras provides deep neural networks, TensorFlow is a machine learning library, scipy is used for mathematic and scientific computations, pandas is a software package, os is an operating system, seaborn is used for statistical graphs, patsy is used for describing statistical models, statsmodels provides classes and functions for statistical models, missingno visualize missing data, warnings provide the ability to ignore warnings, datetime provides classes for manipulating dates and times, and sklearn is a machine learning library.

A screenshot of a computer program

Description automatically generated



The working directory was checked, the data was imported and then the data was viewed.



The data was checked for missing, null, or duplicate values. The data quality is high as it has no missing, null, or duplicate values. Overall data sparsity is 0%. The data contains both qualitative and quantitative values as required for multiple linear regression analysis.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

The unique values for each column were calculated. The shape of encoding each unique value for each column in the data set was viewed. The cardinality is too great to use them all without proliferation.

A screenshot of a computer

Description automatically generated

Before manipulating the data it’s important to first have visualizations to examine outliers, distribution shape and spread, relationships between variables, and correlation (Walker, 2020).

A screenshot of a computer

Description automatically generatedA screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generated

Univariate and bivariate visualizations were created to view the distribution and spread of the data and to check for patterns. All categorical variables were converted using encoding to binary variables so that they could be utilized in the multiple linear regression model.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

The original dataset was uploaded to Tableau Desktop and the data was explored. Reports were created looking into regional sales. The image below shows which states are part of which region. Each state's sales are shown on the map. There is an interactive feature that allows for the month, year, month/year combination, or quarterly sales to be viewed individually. The report shows the total regional sales with the West having the highest profits and the South with the least. The report also shows a line graph displaying this same information but over time. There is a report that visualizes which states have the most and least sales. These reports were also created to visualize shipping trends, segment trends, and product trends. A screenshot of a graph

Description automatically generated

A map of the united states

Description automatically generatedA screenshot of a computer screen

Description automatically generatedA screenshot of a computer screen

Description automatically generatedA screenshot of a graph

Description automatically generatedA graph with different colored bars

Description automatically generatedA graph of different colored bars

Description automatically generated with medium confidenceA screenshot of a computer screen

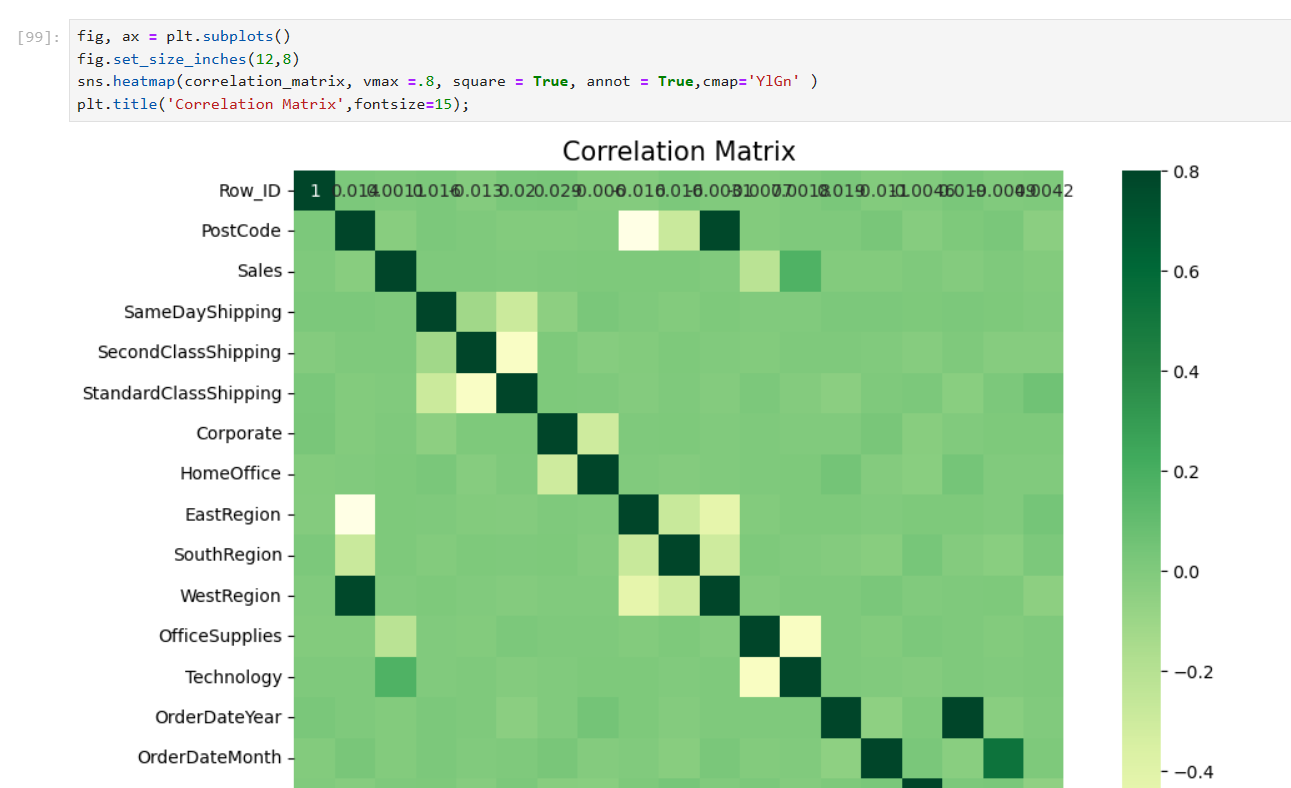
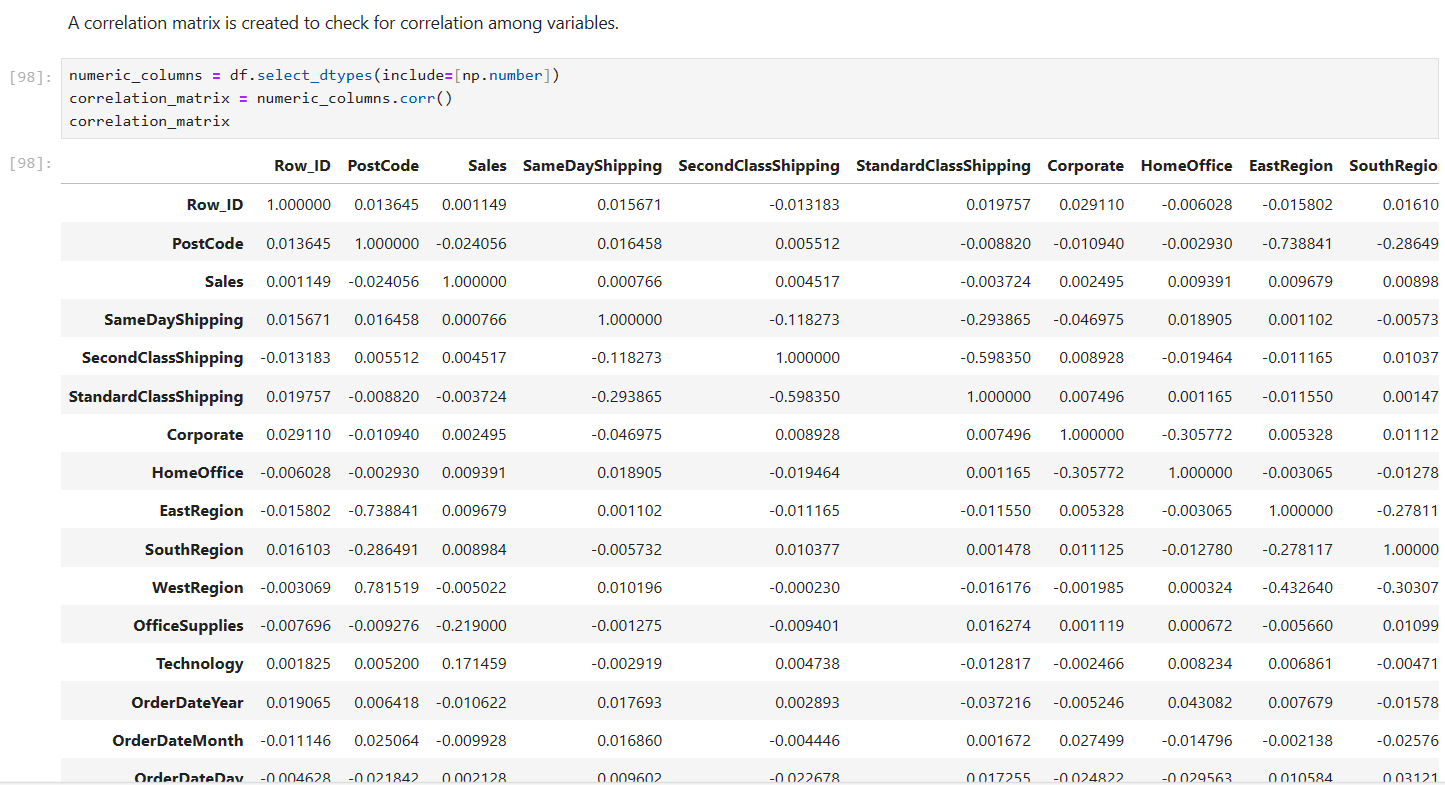
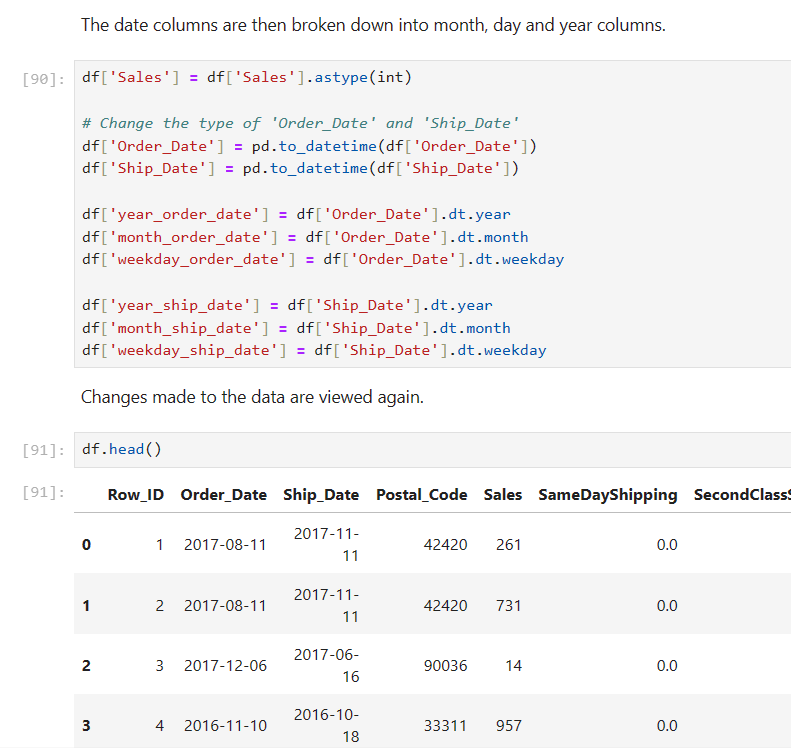
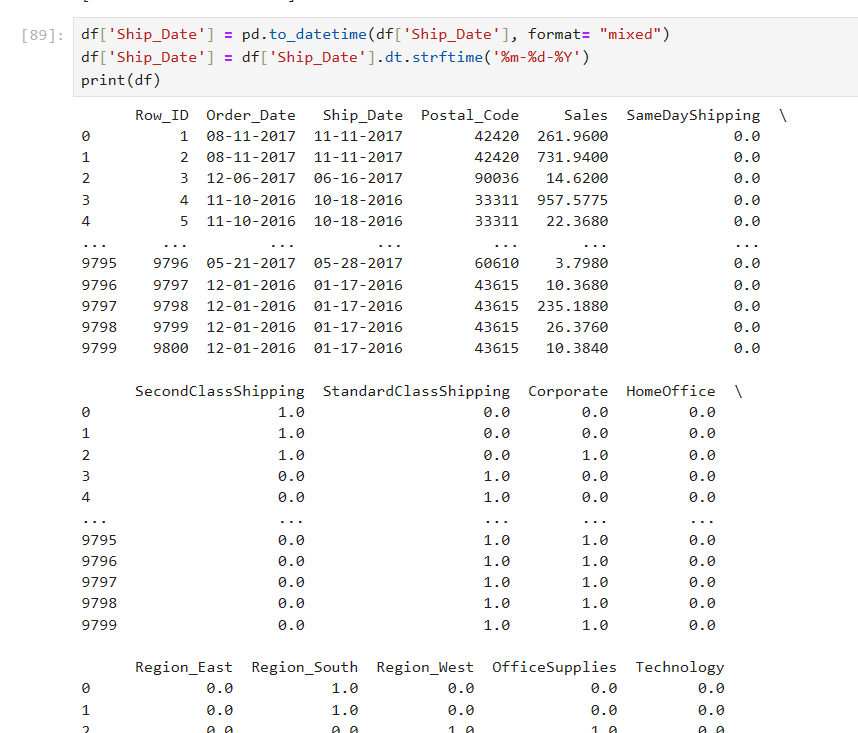
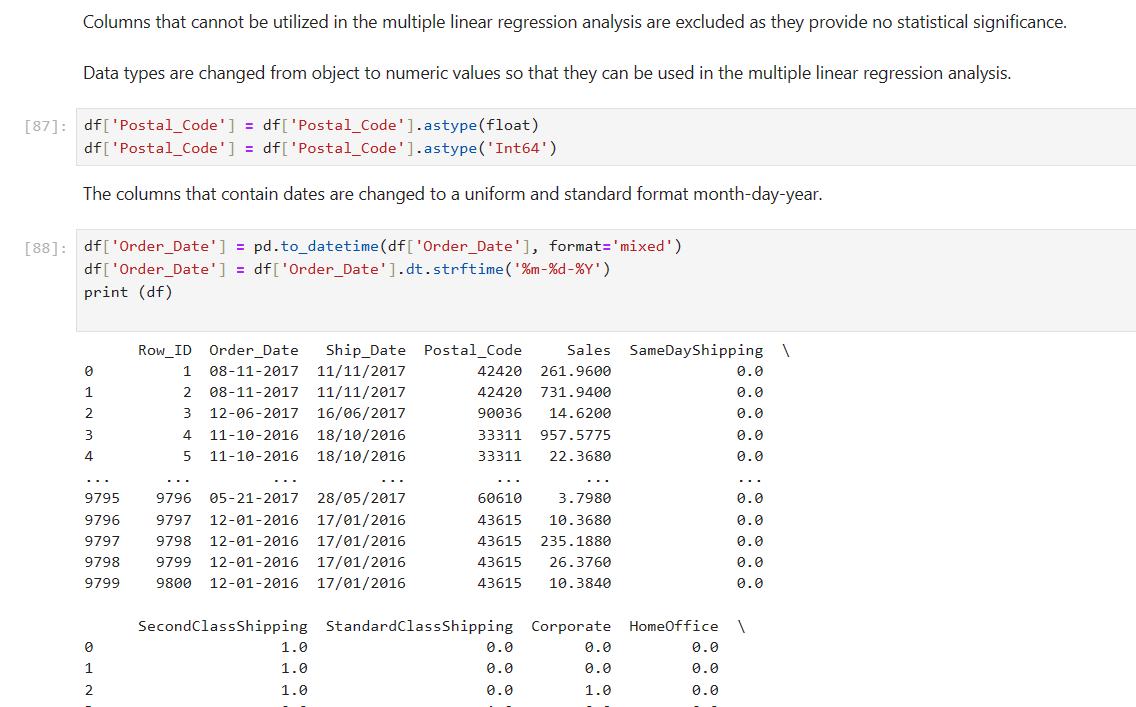
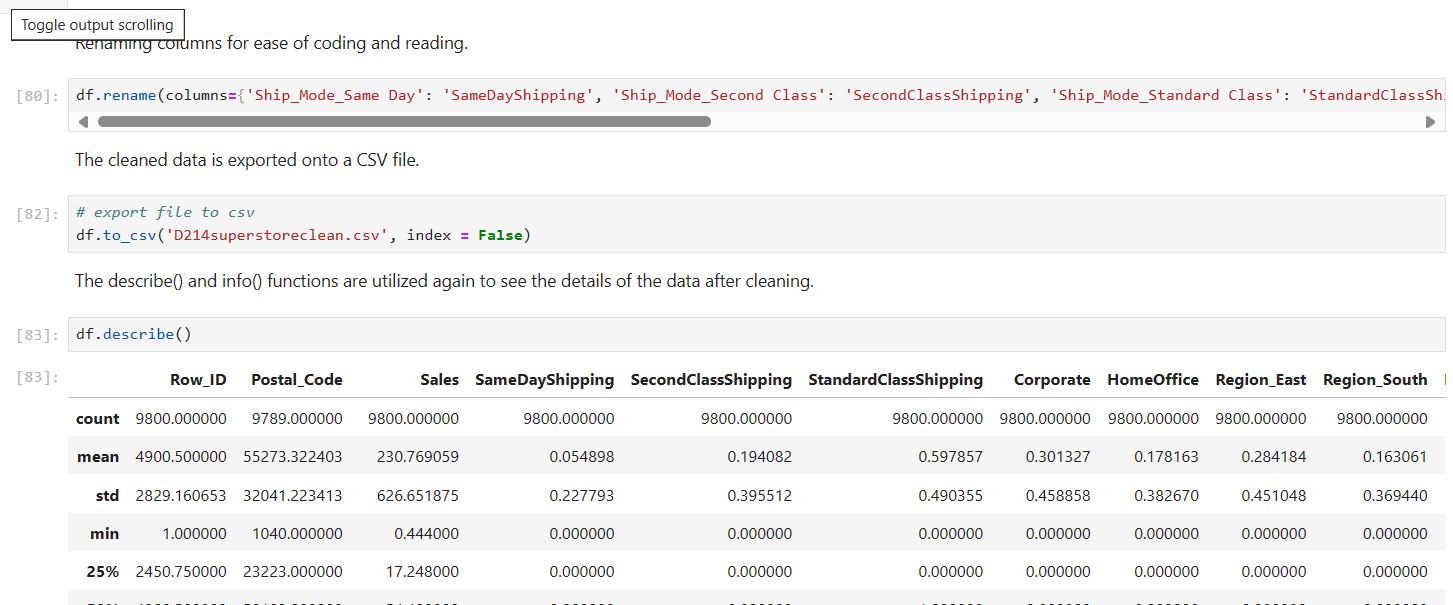
Description automatically generatedA graph of different colored columns

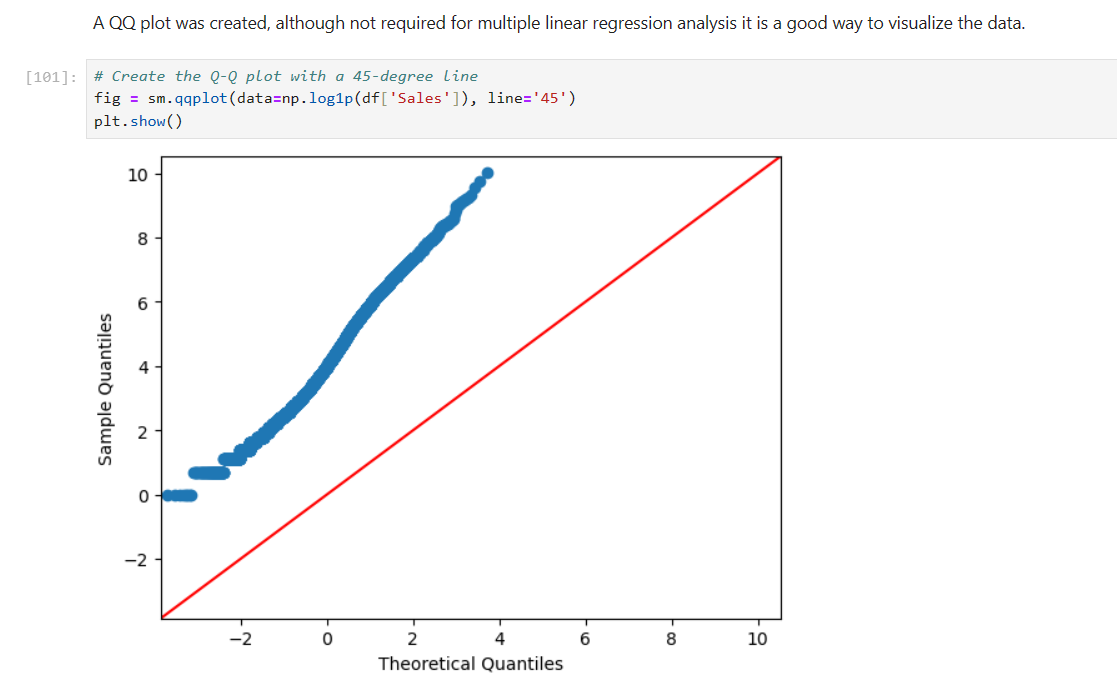
Description automatically generated with medium confidenceA screen shot of a graph

Description automatically generated

# Analysis

Pandas has many tools that were used to clean the data including calculating summary statistics, changing series values conditionally, evaluating and cleaning strings, working with dates, and missing value imputation (Walker, 2020). For ease of reading and interpreting columns were renamed. All data types were converted to numeric so that the multiple linear analysis could be performed. The k-1 method was utilized when one-hot encoding. Columns that served no purpose in the analysis such as Customer\_ID, Product\_ID, etc. were excluded from the data frame. Datetime data was reformatted into month, year, and day columns for workability. A correlation matrix was calculated with a heat mat to visualize the correlation and satisfy the assumptions on multiple linear regression analysis. The distribution of the sales data was visualized with a histogram. A Q-Q plot and Shapiro-Wilk were calculated to determine normality and to view the test statistic. Although it was not required for multiple regression analysis it will provide visualization of the data. ANOVA was performed on all variables, although not required it provides insight into the data. The variance inflation factor was calculated to satisfy the assumptions of multiple linear regression analysis. A VIF of 1 means there is no correlation between the predictor and remaining predictor variables, a VIF greater than 5 warrants an investigation, and a VIF of 10 or greater is a sign of multicollinearity requiring correction. Based on the VIF scores it is advisable that in the reduced model variables that should be removed from the multiple regression model are OrderDateYear, OrderDateMonth, ShipDateYear, and ShipDateMonth. StandardClassShipping is very close to the benchmark at 4.94 and may also be excluded. The data was split into training and test sets and the initial model was calculated using the kitchen sink approach, where all variables are included. The initial model was then reduced based on the multiple linear regression model assumptions and tests performed previously such as VIF. The model was reduced until all assumptions were met and only statistically significant variables remained in the final model.

A graph with blue lines

Description automatically generated

The data analysis technique that was employed was Multiple Linear Regression Analysis. Multiple linear regression analysis is an appropriate technique because the data consists of one continuous dependent variable and several independent variables (Mishra, Pandey, Singh, Keshri, 2019). The analysis was performed using Python in the Jupyter Lab environment. The goals and expectations of the study were to predict future sales revenue utilizing the provided superstore sales dataset. The multiple linear regression analysis uses historic sales data to find patterns between the independent and dependent variables to predict future sales. Predicting future sales revenue and the variables that have the most statistically significant impact on the sales revenue provides actionable insights for smart business decisions.

# Data Summary and Implications

The analysis implies that a multiple linear regression model can be constructed with the superstore sales data. An initial multiple linear regression model was constructed using the kitchen sink approach and putting all variables into the model. The root mean square error or rmse score for that model was 640497.38, the mean absolute error or mae score was 273.77, and the r squared score was 0.04. Another multiple linear regression model was calculated this one without the intercept. The r squared for this model was .164, and the adjusted r squared was .163. The variance inflation factor or VIF has been run and variables with a VIF of greater than 10 will be considered for removal in the reduced model due to a high VIF meaning there is multicollinearity. Variables with P values of greater than .05 will be considered for removal in the reduced model. The first reduced model consisted of LMR = ols(formula="Sales ~ SameDayShipping + SecondClassShipping + Corporate + HomeOffice + EastRegion + SouthRegion + WestRegion + OfficeSupplies + Technology + OrderDateDay + ShipDateDay", data=df).fit(). The r squared of this model was .051 and the adjusted r squared was .050. The model could be improved further as the research suggests that the best multiple linear regression model would have five or fewer variables. The next reduced model included LMRF = ols(formula="Sales ~ SameDayShipping + OfficeSupplies + Technology", data=df).fit(). The r squared remained at .051 with an adjusted r squared of .050. There was no improvement in this model in regards to the statistic r squared and adjusted r squared. The reduced model was then calculated with no intercept. The r squared then changed to .106 with an adjusted r squared of .106. The data frame was then standardized and the final multiple linear regression model was calculated. The rmse of this model was 1.63, mae was .44 and the r squared was .04. The closer to zero the rmse score is the more accurate the prediction is. The closer the mae is to zero the more accurate the models' prediction. The closer to 1 the r squared value is the better fit the regression is. This indicates that the variable office supplies and technology have the greatest impact on the sales revenue. Based on these findings it is recommended that the sale of technology and office supplies be prioritized in future sale efforts as they produce the greatest return of the variables in this study. Future studies of this dataset should include a market basket analysis or another type of classification model. A limitation of the study is the lack of review data for products and superstore locations. To improve the dataset reviews of the products and/or store locations could be beneficial in future studies on increasing sales revenue.

**References**

Anseur, A. (2024). *Superstore Sales EDA + ML.* Kaggle.com. Retrieved April 16, 2024, from [Superstore Sales | EDA + ML (kaggle.com)](https://www.kaggle.com/code/adelanseur/superstore-sales-eda-ml)

Chattopadhyay, R. (2016). *Effective Businesss Solutions With Big Data Analytics: Key for Business Growth.* Globsyn Management Journal, 10(1/2), 87–96. Retrieved April 16, 2024, from [https://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=132005022&  
site=eds-live&scope=site&authtype=shib&custid=ns017578](https://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=132005022&site=eds-live&scope=site&authtype=shib&custid=ns017578)

Darlington, R. & Hayes, A. (2017). *Regression Analysis and Linear Models : Concepts, Applications, and Implementation.* The Guilford Press. Retrieved April 16, 2024, from [https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1340168&  
site=eds-live&scope=site&authtype=shib&custid=ns017578](https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1340168&site=eds-live&scope=site&authtype=shib&custid=ns017578)

Dong, J., Chen, Y., Gu, A., Chen, J., Li, L., Chen, Q., Li, S., & Xun, Q. (2020). *Potential Trend for Online Shopping Data Based on the Linear Regression and Sentiment Analysis.* Mathematical Problems in  
Engineering, 1–11. Retrieved April 16, 2024, from <https://doi.org/10.1155/2020/4591260>

Mishra, P., Pandey, C. M., Singh, U., Keshri, A., & Sabaretnam, M. (2019). *Selection of appropriate statistical methods for data analysis.* Annals of Cardiac Anaesthesia, 22(3), 297–301. Retrieved April 16, 2024, from  
<https://doi.org/10.4103/aca.ACA_248_18>

Multiple Regression. (n.d.). *Multiple Regression.* Csulb.edu. Retrieved April 16, 2024, from [Multiple Regression (csulb.edu)](https://home.csulb.edu/~msaintg/ppa696/696regmx.htm)

Open Knowledge. (n.d.). *Open Data Commons Public Domain.* Opendatacommons.org. [Open Data Commons Public Domain Dedication and License (PDDL) v1.0 — Open Data Commons: legal tools for open data](https://opendatacommons.org/licenses/pddl/1-0/)

Ozgur, C., Colliau, T., Rogers, G., Hughes, Z., & Myer-Tyson, E. “Bennie.” (2017). *MatLab vs. Python vs. R.* Journal of Data Science, 15(3), 355–371.Retrieved April 16, 2024, from [https://search.ebscohost.com/login.aspx?direct=true&db=asn&AN=125220011&  
site=eds-live&scope=site&authtype=shib&custid=ns017578](https://search.ebscohost.com/login.aspx?direct=true&db=asn&AN=125220011&site=eds-live&scope=site&authtype=shib&custid=ns017578)

Siddiqui, T., Alkadri, M., & Khan, N. A. (2017). *Review of Programming Languages and Tools for Big Data Analytics.* International Journal of Advanced Research in Computer Science, 8(5), 1112–1118. Retrieved April 16, 2024, from [https://search.ebscohost.com/login.aspx?direct=true&db=asf&AN=124636531&  
site=eds-live&scope=site&authtype=shib&custid=ns017578](https://search.ebscohost.com/login.aspx?direct=true&db=asf&AN=124636531&site=eds-live&scope=site&authtype=shib&custid=ns017578)

Walker, M. (2020). *Python Data Cleaning Cookbook : Modern Techniques and Python Tools to Detect and Remove Dirty Data and Extract Key Insights.* Packt Publishing. Retrieved April 16, 2024, from [https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=2706043&  
site=eds-live&scope=site&authtype=shib&custid=ns017578](https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=2706043&site=eds-live&scope=site&authtype=shib&custid=ns017578)

Zach. (2020). *The four assumptions of linear regression.*

[The Four Assumptions of Linear Regression - Statology](https://www.statology.org/linear-regression-assumptions/)

Zach. (2020). *Four examples of using multiple linear regression.*

[4 Examples of Using Linear Regression in Real Life - Statology](https://www.statology.org/linear-regression-real-life-examples/)

Zach. (2020). *Multiple linear regression by hand.*

[Multiple Linear Regression by Hand (Step-by-Step) - Statology](https://www.statology.org/multiple-linear-regression-by-hand/)