Multiple Linear Regression Analysis on Superstore Sales Data

Executive Summary

Nicole Reiswig

College of Information Technology, Western Governors University

Dr. William Sewell

May 01, 2024

Multiple Linear Regression Analysis on Superstore Sales Data

Executive Summary

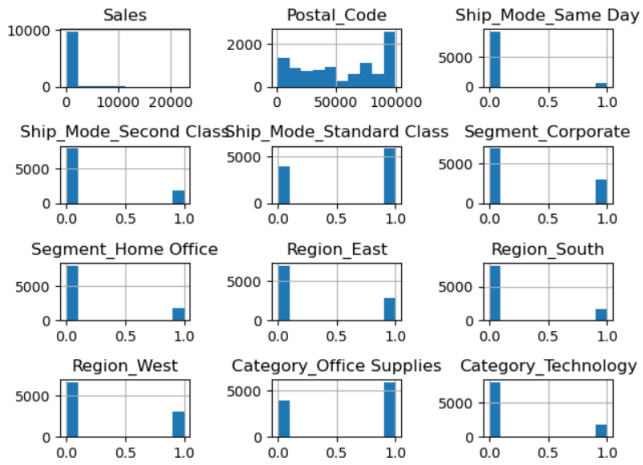
Superstores collect large amounts of sales data. The data collected in this application encompasses sales data over some time, product categories, geographic regions, and customer segments. Various attributes have the power to predict favorable business outcomes with insight on how to achieve the goal. This study aims to identify the attributes of superstore sales that are key factors in sales revenue and which attributes have statistical significance in predicting future sales revenue. The goal is to identify if a multiple linear regression model can be constructed utilizing this data set to provide these business insights. A summary of the hypothesis is: H_0 - A predictive regression model cannot be constructed from the superstore sales data, H_1 - A predictive regression model can be constructed from the superstore sales data.

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. The target variable was Sales. For this study, Row_ID, Order_ID, Customer_ID, and Product_ID were excluded as they would provide no statistical significance to this analysis.

Visualizations created on raw dataset:

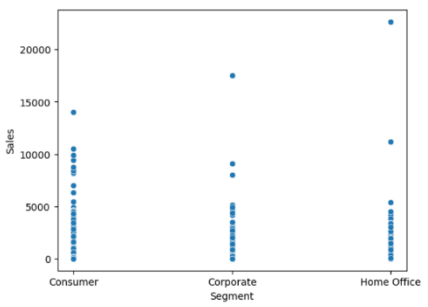
- Univariate- histogram and boxplot to view outliers and distribution

```
[51]: # Create histograms of variables
df[['Sales', 'Order_Date', 'Ship_Date', 'Postal_Code', 'Ship_Mode_Same Day', 'Ship_Mode_Second Class', 'Ship_Mode_Standard Class', 'Segment_Corporate',
plt.savefig('superstore_pyplot.jpg')
plt.tight_layout()
```

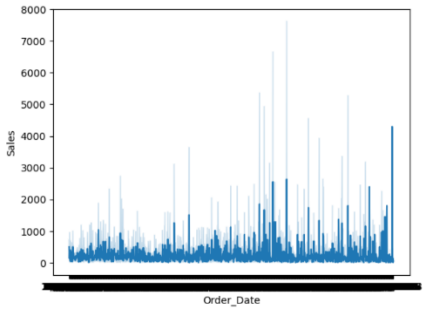


- Bivariate- scatterplot and lineplot to view relationships

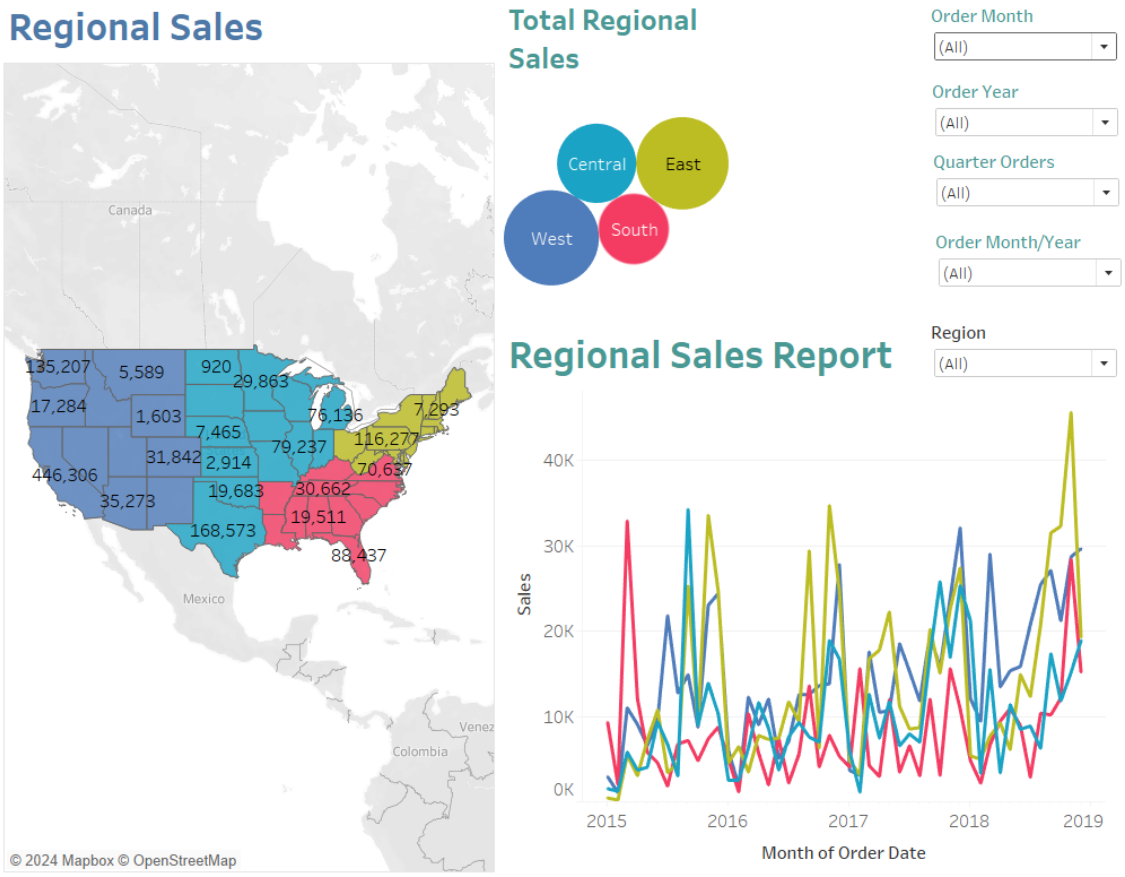
```
[32]: ##Add scatter plot for each variable before cleaning
sns.scatterplot(x="Segment",
                y="Sales",
                data=superstore)
plt.show()
```

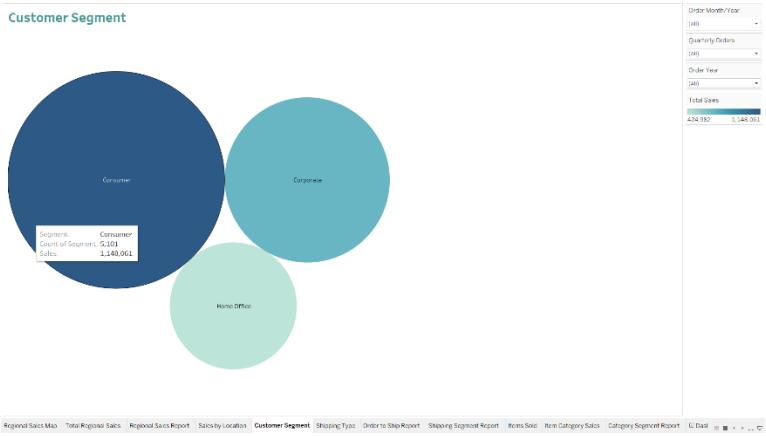



```
[76]: sns.lineplot(x="Order_Date",
                  y="Sales",
                  data=df)
plt.show()
```

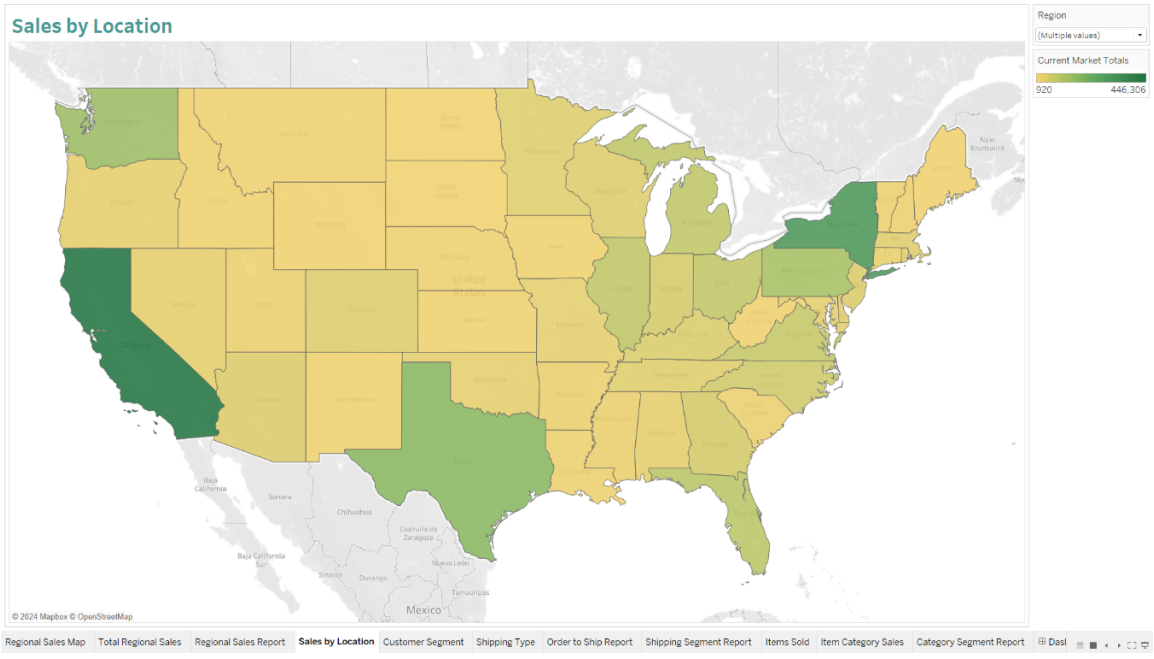


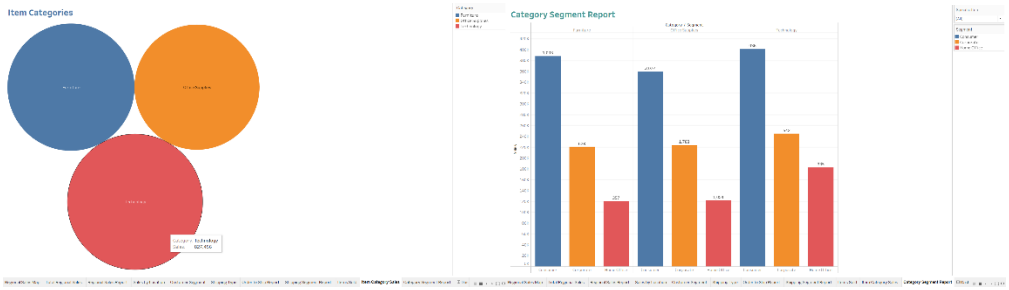
- Tableau visualizations- maps, reports and others

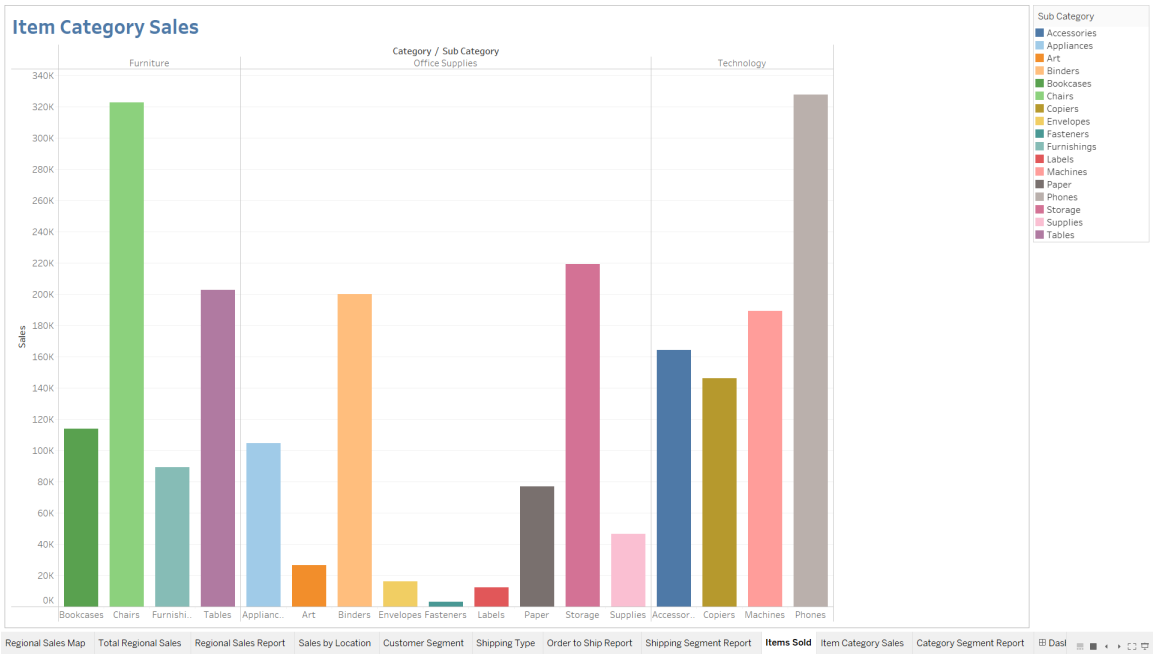



Customer Segment
Consumer
Corporate
Home Office
Segment Consumer
Count of Segment... 5,101
Sales 1,140,061
Order Month / Year
Quarter by Orders
Order Year
Total Sales
474,982 1,148,061
Regional Sales Map
Total Regional Sales
Regional Sales Report
Sales by Location
Customer Segment
Shipping Type
Order to Ship Report
Shipping Segment Report
Items Sold
Item Category Sales
Category Segment Report
Dasl


Shipping Mode
Standard Class
Second Class
First Class
Same Day
SUM(Sales)
125,210 1,340,931
Regional Sales Map
Total Regional Sales
Regional Sales Report
Sales by Location
Customer Segment
Shipping Type
Order to Ship Report
Shipping Segment Report
Items Sold
Item Category Sales
Category Segment Report
Dasl


Sales by Location
Region
(Multiple values)
Current Market Totals
920 446,306
© 2024 Mapbox © OpenStreetMap
Regional Sales Map
Total Regional Sales
Regional Sales Report
Sales by Location
Customer Segment
Shipping Type
Order to Ship Report
Shipping Segment Report
Items Sold
Item Category Sales
Category Segment Report
Dasl





The variable preparation steps included:

- Verifying there were no missing, null, or duplicate values
- Excluding the categorical variables with more than 4 levels as the cardinality would be too great and did not want to proliferate
- One-hot encoding of all other categorical variables to binary variables 0/1 utilizing the K-1 method
- Renaming columns for ease of analysis and readability
- Changing the data type from object or datetime to numeric variables

At the end of the variable preparation process sixteen major independent variables were identified and retained for future analysis: SameDayShipping, SecondClassShipping, StandardClassShipping, Corporate, HomeOffice, EastRegion, WestRegion, SouthRegion,

OfficeSupplies, Technology, OrderDateYear, OrderDateMonth, OrderDateDay, ShipDateYear, ShipDateMonth, and ShipDateDay. The data was then split into training and test sets.

Multiple Linear Regression was used to identify the statistically significant independent variables and interaction effects, build the model, and score new observations. 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 Sales ~ SameDayShipping + SecondClassShipping + Corporate + HomeOffice + EastRegion + SouthRegion + WestRegion + OfficeSupplies + Technology + OrderDateDay + ShipDateDay. 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, Sales ~ SameDayShipping + OfficeSupplies + Technology. The r squared remained at .051 with an adjusted r squared of .050. There was no improvement in this model regarding 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.

```
[169]: LMRF = ols(formula="Sales ~ SameDayShipping + OfficeSupplies + Technology", data=df).fit()

[170]: # Get summary statistics
       print(LMRF.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.051
Model:                            OLS   Adj. R-squared:                  0.050
Method:                 Least Squares   F-statistic:                     174.6
Date:                Fri, 26 Apr 2024   Prob (F-statistic):          2.65e-110
Time:                        13:33:01   Log-Likelihood:                -76765.
No. Observations:                9800   AIC:                         1.535e+05
Df Residuals:                    9796   BIC:                         1.536e+05
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept        349.9707     13.483     25.957      0.000     323.541     376.400
SameDayShipping    1.9985     27.079      0.074      0.941     -51.082      55.079
OfficeSupplies  -231.2553     15.573    -14.850      0.000    -261.782    -200.729
Technology       105.5980     19.623      5.381      0.000      67.132     144.064
==============================================================================
Omnibus:                    17958.041   Durbin-Watson:                   1.983
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         42603794.132
Skew:                          13.391   Prob(JB):                         0.00
Kurtosis:                     324.898   Cond. No.                         5.32
==============================================================================
```

```
[197]: # Make predictions on the test data
       y2_pred = model_pipeline.predict(X2_test)

[199]: # Evaluate the model using MSE, R-squared, MAE
       rmse2 = mean_squared_error(y2_test, y2_pred)
       mae2 = mean_absolute_error(y2_test, y2_pred)
       r22 = r2_score(y2_test, y2_pred)

[200]: # Store the results
       rmse_results[model_name] = rmse2
       mae_results[model_name] = mae2
       r2score_results[model_name] = r22

[201]: print("Model :", model_name)
       print(f"RMSE : {rmse2:.2f}")
       print(f"MAE : {mae2:.2f}")
       print(f"r2 : {r22:.2f}")
       print("--------")
```

```
Model : Linear Regression
RMSE : 1.63
MAE : 0.44
r2 : 0.04
--------
```

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. Previous studies show that review data can help increase customer satisfaction which increases sales revenue. The delimitations of the study are that Order_ID, Customer_ID, Customer_Name, Country, and Product_ID will be removed from the dataset. The variable country was removed because the data was only collected in the United States and the other columns removed provided no statistically significant value to this analysis.

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. Through exploratory data analysis, it was also discovered that the overall largest segment was the consumer, the preferred shipping method was standard class, and the East and West regions were the top performers.

Future studies of this dataset should include a market basket analysis or another type of classification model. Another model that could be beneficial is a time series analysis.

The benefits of the study are 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.

References:

Anseur, A. (2024). *Superstore Sales EDA + ML.* Kaggle.com. Retrieved April 16, 2024, from Superstore Sales | EDA + ML (kaggle.com)