Predicting Superstore Sales

NICOLE REISWIG

Agenda

- Introduction & background
- Problem statement & hypothesis
- Data analytics process
- Data overview
- Data visualizations
- Summary of findings
- Limitations of techniques & tools
- Proposed action
- Expected benefit

Introduction & background

Master of science in data analytics

Master of business administration

Bachelor of science in healthcare management

Previously an Executive Director for long-term care facilities

Currently located in the Florida panhandle

From Washington state

PROBLEM STATEMENT & HYPOTHESIS

1. Can a multiple linear regression model be constructed based solely on the research data?

2. Null hypothesis H_0 : A predictive regression model cannot be constructed from the superstore sales data"

3. Alternate hypothesis H_1 : A predictive regression model can be constructed from the superstore sales data."

DATA ANALYTICS PROCESS

STAGE 01	STAGE 02	STAGE 03	STAGE 04	STAGE 05
Data acquisition	Data cleaning	Data wrangling	Exploratory data analysis	Predictive modeling

Cleaning:

Missing, null, duplicates

Wrangling:

One-hot encoding

Converting data type

EDA:
Visualizations
Qqplot, Shapiro-wilk
Correlation matrix
VIF

Predictive modeling:

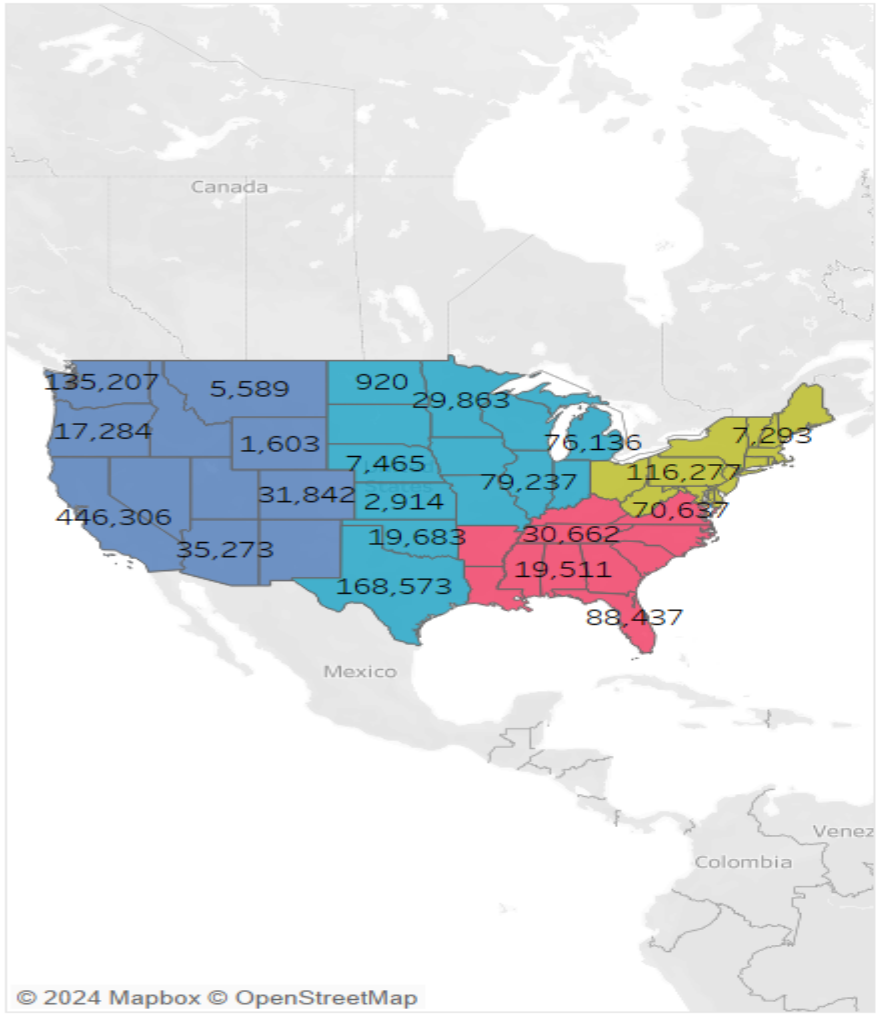
Multiple linear regression

Data overview

- Consisted of 9800 rows of superstore sales data
- 18 columns including sales data, product categories, geographic regions and customer segment
- Predictive modeling using multiple linear regression

Exploratory Data Analysis

Regional Sales



Total Regional Sales

Order Month
(All)

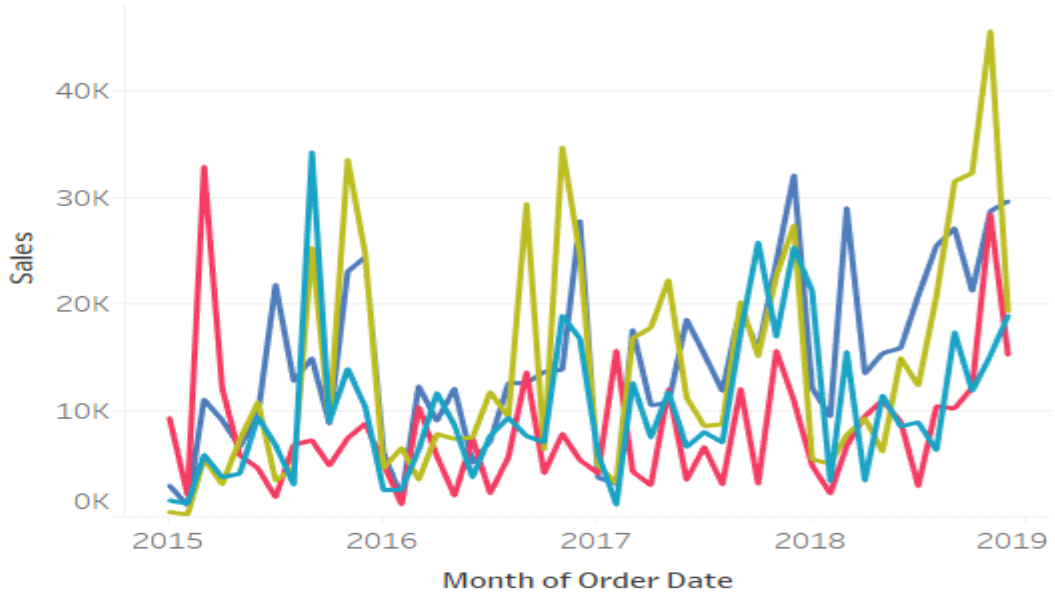
Order Year
(All)

Quarter Orders
(All)

Order Month/Year
(All)

Region
(All)

Regional Sales Report



Correlation

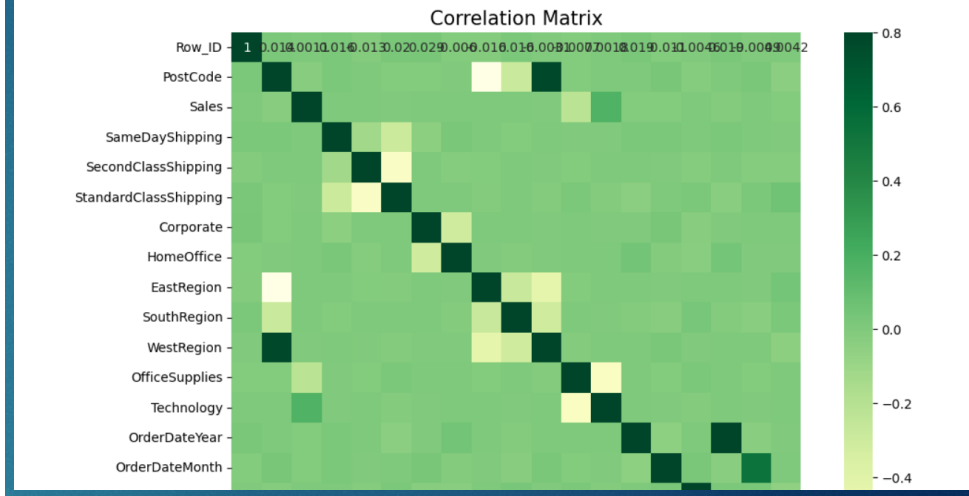
A correlation matrix is created to check for correlation among variables.

[98]:
```
numeric_columns = df.select_dtypes(include=[np.number])
correlation_matrix = numeric_columns.corr()
correlation_matrix
```

[98]:

	Row_ID	PostCode	Sales	SameDayShipping	SecondClassShipping	StandardClassShipping	Corporate	HomeOffice	EastRegion	SouthRegio
Row_ID	1.000000	0.013645	0.001149	0.015671	-0.013183	0.019757	0.029110	-0.006028	-0.015802	0.01610
PostCode	0.013645	1.000000	-0.024056	0.016458	0.005512	-0.008820	-0.010940	-0.002930	-0.738841	-0.28649
Sales	0.001149	-0.024056	1.000000	0.000766	0.004517	-0.003724	0.002495	0.009391	0.009679	0.00898
SameDayShipping	0.015671	0.016458	0.000766	1.000000	-0.118273	-0.293865	-0.046975	0.018905	0.001102	-0.00573
SecondClassShipping	-0.013183	0.005512	0.004517	-0.118273	1.000000	-0.598350	0.008928	-0.019464	-0.011165	0.01037
StandardClassShipping	0.019757	-0.008820	-0.003724	-0.293865	-0.598350	1.000000	0.007496	0.001165	-0.011550	0.00147
Corporate	0.029110	-0.010940	0.002495	-0.046975	0.008928	0.007496	1.000000	-0.305772	0.005328	0.01112
HomeOffice	-0.006028	-0.002930	0.009391	0.018905	-0.019464	0.001165	-0.305772	1.000000	-0.003065	-0.01278
EastRegion	-0.015802	-0.738841	0.009679	0.001102	-0.011165	-0.011550	0.005328	-0.003065	1.000000	-0.27811
SouthRegion	0.016103	-0.286491	0.008984	-0.005732	0.010377	0.001478	0.011125	-0.012780	-0.278117	1.00000
WestRegion	-0.003069	0.781519	-0.005022	0.010196	-0.000230	-0.016176	-0.001985	0.000324	-0.432640	-0.30307
OfficeSupplies	-0.007696	-0.009276	-0.219000	-0.001275	-0.009401	0.016274	0.001119	0.000672	-0.005660	0.01099
Technology	0.001825	0.005200	0.171459	-0.002919	0.004738	-0.012817	-0.002466	0.008234	0.006861	-0.00471
OrderDateYear	0.019065	0.006418	-0.010622	0.017693	0.002893	-0.037216	-0.005246	0.043082	0.007679	-0.01578
OrderDateMonth	-0.011146	0.025064	-0.009928	0.016860	-0.004446	0.001672	0.027499	-0.014796	-0.002138	-0.02576
OrderDateDay	-0.004628	-0.021842	0.002128	0.009602	-0.022678	0.017255	-0.024822	-0.029563	0.010584	0.03121

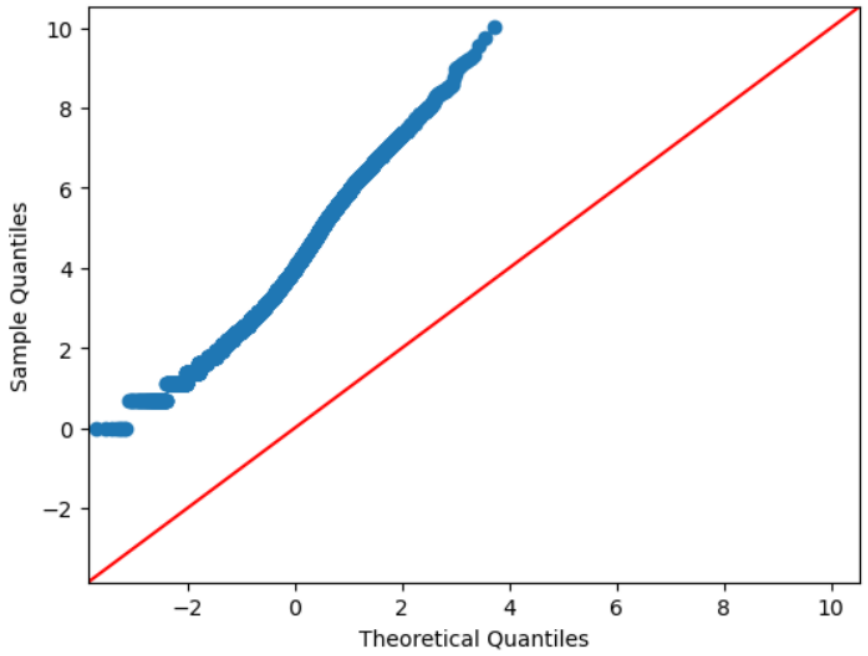
[99]:
```
fig, ax = plt.subplots()
fig.set_size_inches(12,8)
sns.heatmap(correlation_matrix, vmax =.8, square = True, annot = True,cmap='YlGn' )
plt.title('Correlation Matrix',fontsize=15);
```

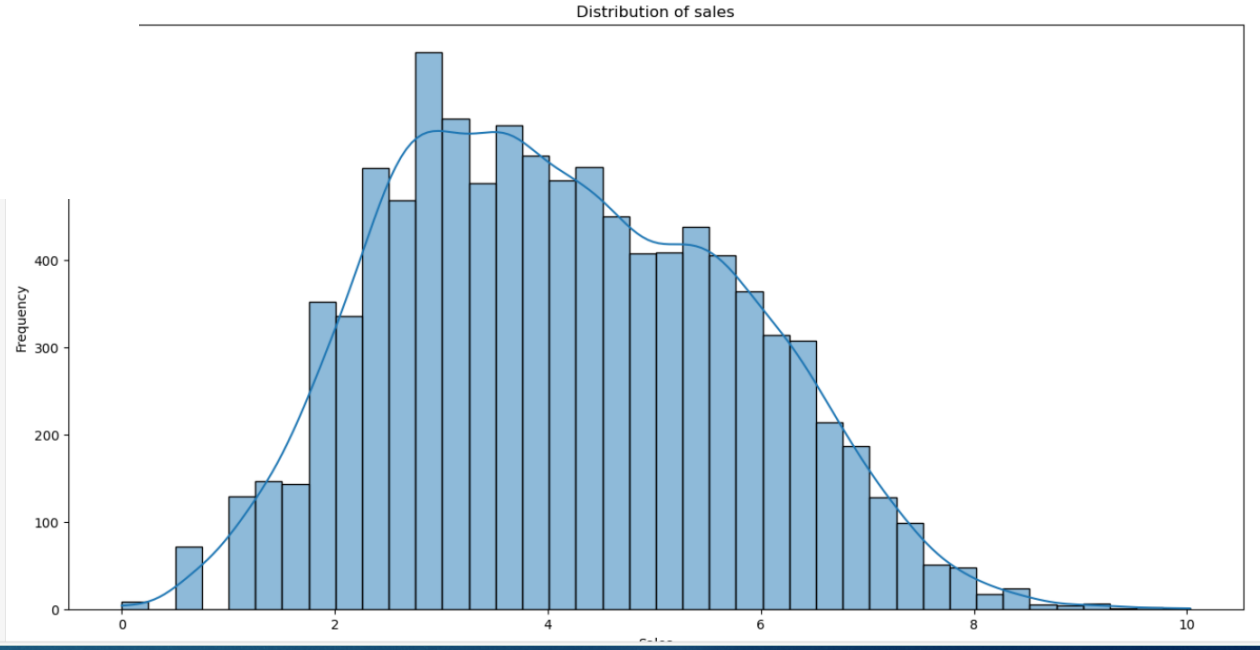


Visualizing Data

A QQ plot was created, although not required for multiple linear regression analysis it is a good way to visualize the data.

[101]:
```
# Create the Q-Q plot with a 45-degree line
fig = sm.qqplot(data=np.log1p(df['Sales']), line='45')
plt.show()
```





Multiple Linear Regression

[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
--------
```

RMSE stands for root mean squared error is 1.63. The closer to zero the rmse score is the more accurate the prediction is

MAE stands for mean absolute error is .44. The closer the mae is to zero the more accurate the models' prediction.

R-squared is 0.04. The closer to 1 the r squared value is the better fit the regression is.

Limitations and Benefits

Benefit	Limitation
The multiple linear regression analysis uses historic sales data to find patterns between the independent and dependent variables to predict future sales.	The lack of review data from consumers
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.	A delimitation would be the automatic removal of several columns prior to analysis due to the lack of statistical significance

Recommendations:

>Perform market basket analysis or another classification model

>Perform a time series analysis

>Collect review data from the superstore customers

Thank You!

STUDENT ID 008756001