Superstore Project Report

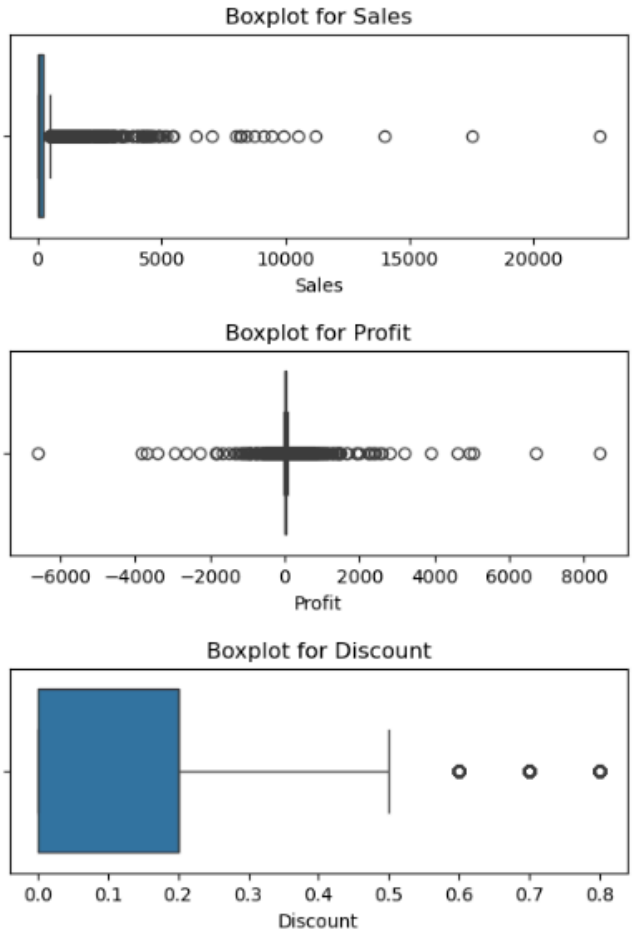
- **Problem Statement:**
  - With growing demands and an increase in competition in the market, our Superstore would like to understand which products, regions, categories and customer segments they should target or avoid to boost profitability for the store. What kind of prediction model with an accuracy of 80% or more, can we make that can help us increase profits in geographical areas and predict future sales trends in the next 30 days?

- **Data Wrangling:**
  - Data Cleaning
  - While cleaning the data our goal was to find which region and cities had the highest amount of sales. The dataset did not have any duplicates or missing values and we found that the dataset had almost 10,000 sales orders. After looking at the boxplot distribution of the sales, profit and discount columns, we concluded that the data was highly seasonal so no outliers were dropped. New York City and Los Angeles had the most product sales for the company compared to other cities with the West region having the highest volume of sales.
  -

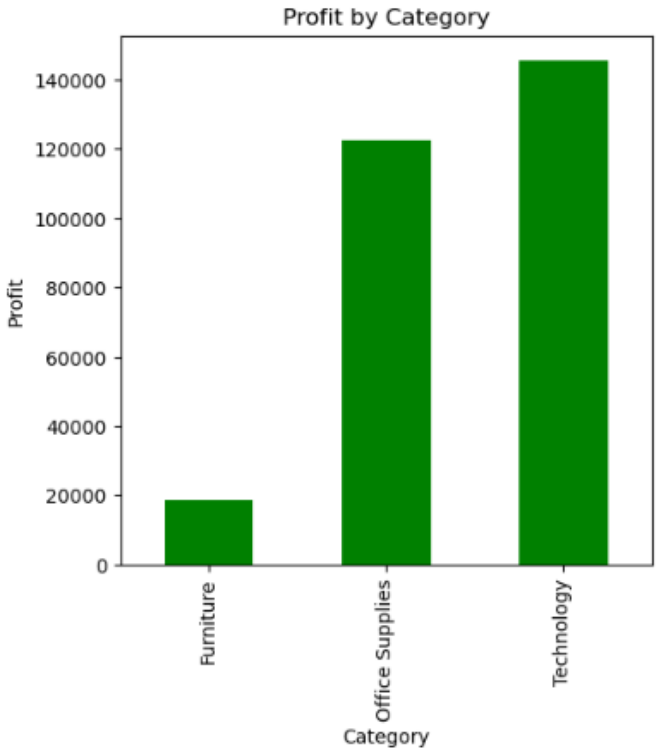
Region	
West	3203
East	2848
Central	2323
South	1620

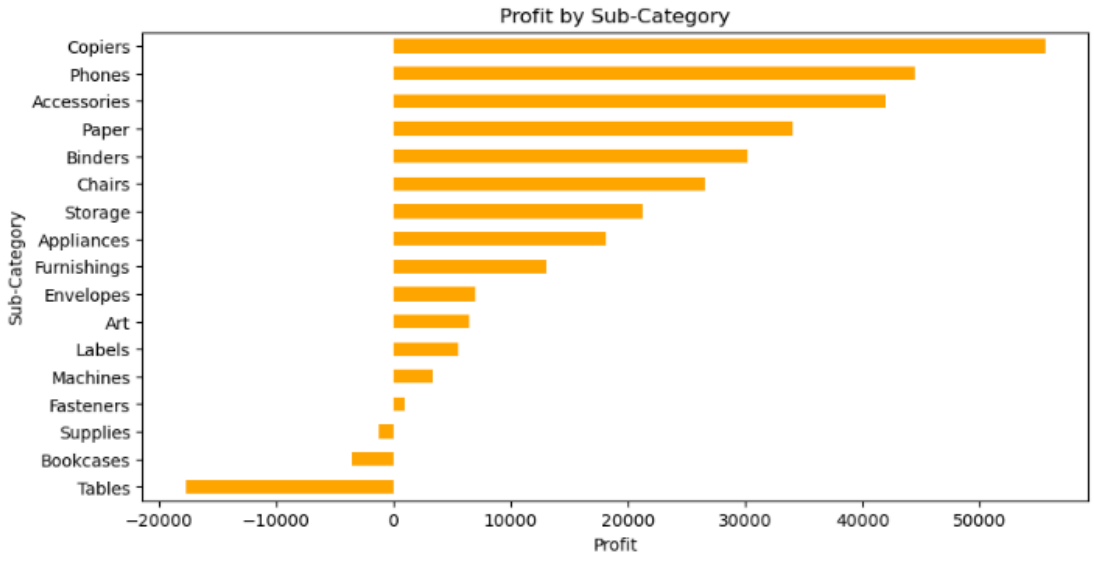
  -

City	
New York City	915
Los Angeles	747
Philadelphia	537
San Francisco	510
Seattle	428



- **Exploratory Data Analysis:**
  - [EDA Report]
  - While exploring the data we wanted to focus on “Sales, Profit and Discount” and what metrics affected these features. We found that the superstore made most of its profits selling copiers and the “Technology” category was the most profitable out of every other category. Our most profitable customers were “Home Office” and “Corporate” customers. Another interesting thing we found out about the data in regard to discounts is that heavily discounted items were negatively impacting our profits. Any discounted item with a 30% or more discount is likely to produce unprofitable results. “Tables” are the items that are hurting our profitability the most.





- **Model Preprocessing and Feature Engineering:**
  - Model Preprocessing Phase
  - We want to find a prediction model that will help us increase profits in geographical areas and predict future sales trends in the next 30 days so we need to find the important features in the dataset that are impacting our profitability. We encoded our categorical columns and dropped any columns that

would hinder our results in our prediction model. Our target for our model when training and splitting the data was the “Profit” column.

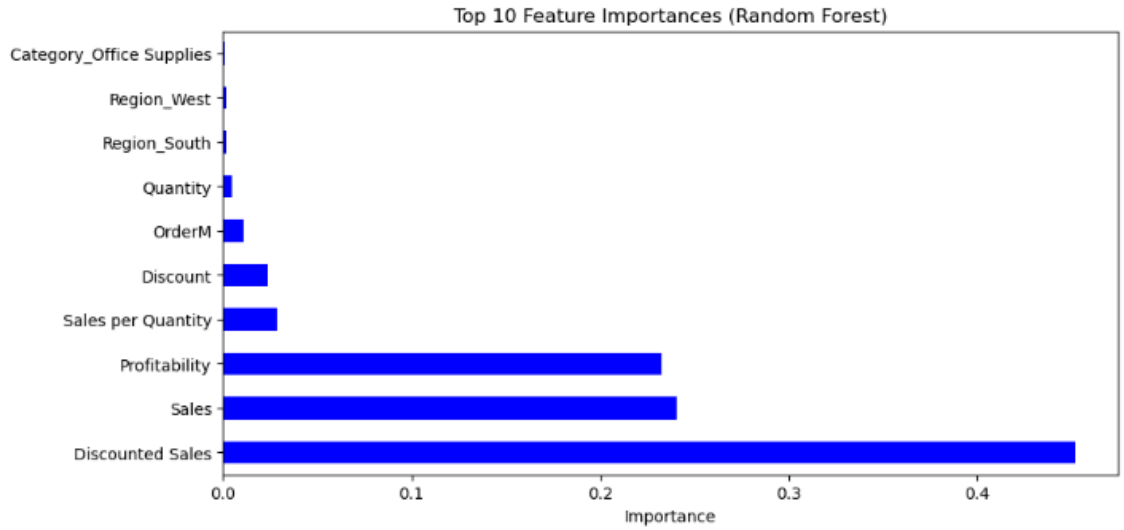
```
Features of Interest: Index(['Sales', 'Quantity', 'Discount', 'OrderY', 'OrderM', 'Discounted Sales',
       'Sales per Quantity', 'Profitability', 'Ship Mode_Same Day',
       'Ship Mode_Second Class', 'Ship Mode_Standard Class',
       'Segment_Corporate', 'Segment_Home Office', 'Region_East',
       'Region_South', 'Region_West', 'Category_Office Supplies',
       'Category_Technology'],
      dtype='object')
```

- **Modeling:**
  - Machine Learning Models
  - For our prediction model testing we decided to test Random Forest Regression and XG Boost since the sales data is large and has a lot of features . In the Random Forest model we used a Grid Search to perform hyperparameter tuning and find the best parameters for our model. We found that the best tree depth for our model was a max depth of 10. After training the data and fitting it to the model, the test data performance accuracy came out to be 94%. The Random Forest model helped us conclude that “Discounted Sales” was the most important feature impacting our profitability. The XG Boost model produced a result of 75% accuracy performance.

```
#Predict on Test Data
best_rf = grid_search.best_estimator_
y_pred_tuned = best_rf.predict(X_test)

print("Test RMSE :", np.sqrt(mean_squared_error(y_test, y_pred_tuned)))
print("Test R2 Score :", r2_score(y_test, y_pred_tuned))
```

```
Test RMSE : 51.52603920018055
Test R2 Score : 0.9409406147965207
```



```
#XG Boost Model
model_xgb = XGBRegressor(random_state=123)
model_xgb.fit(X_train, y_train)
y_pred_xgb = model_xgb.predict(X_test)
print("XGBoost Test RMSE:", np.sqrt(mean_squared_error(y_test, y_pred_xgb)))
print("XGBoost Test R2 Score:", r2_score(y_test, y_pred_xgb))
```

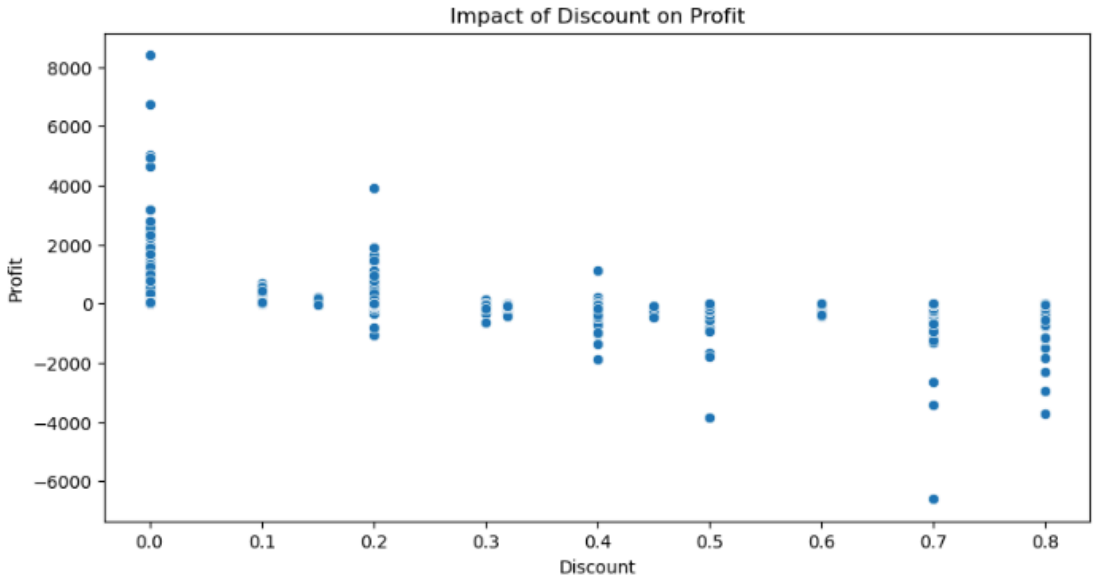
```
XGBoost Test RMSE: 105.32085432561489
XGBoost Test R2 Score: 0.7532460297563277
```

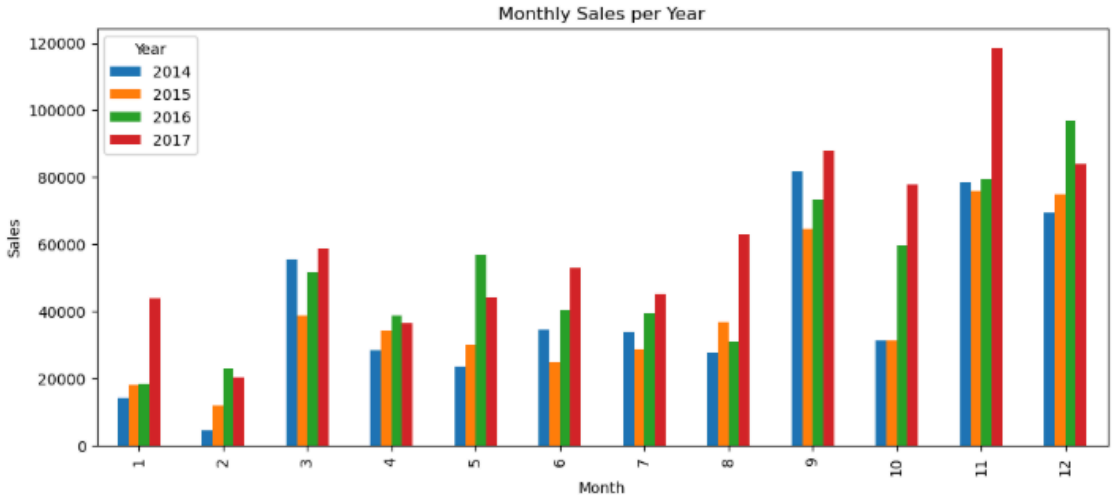
- **Winning Model and Scenario Modeling:**
  - Our goal was to produce a predictive model of 80% accuracy so our winning model is the Random Forest Model with a prediction accuracy of 94%. The XG Boost Model did not produce the results we were hoping for. We ran a Cross Validation score and our Random Forest Model explains 70% of the variance in our target “Profit” on the test data.

```
scores = cross_val_score(best_rf, X, y, cv=5, scoring='r2')
print("Cross-Validated R2 Mean:", scores.mean())
```

```
Cross-Validated R2 Mean: 0.7057826682021355
```

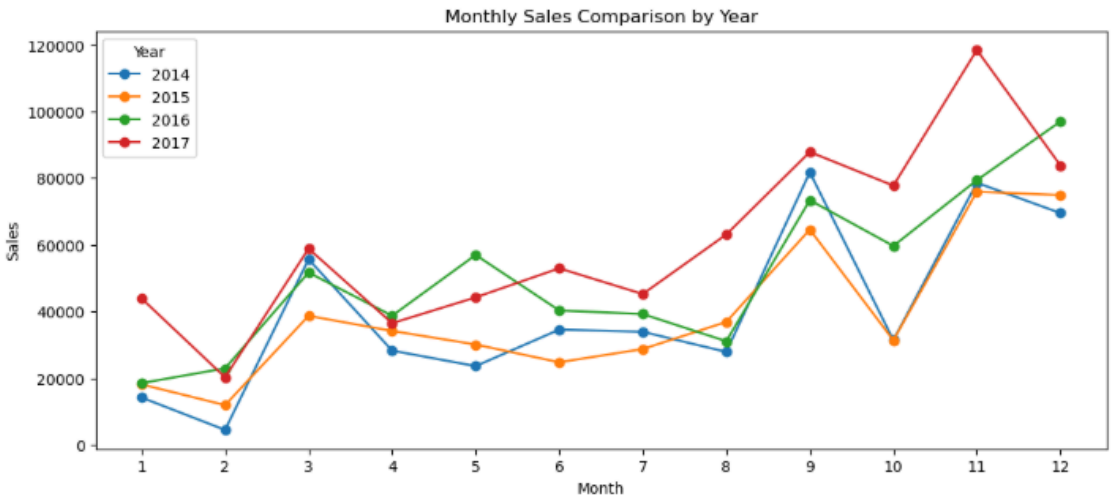
- **Recommendations:**
  - Since “Discounted Sales” was the most important feature impacting our profitability, seeing what items are heavily discounted and decreasing those items to a 20% or lower discount will potentially increase profitability for the superstore.

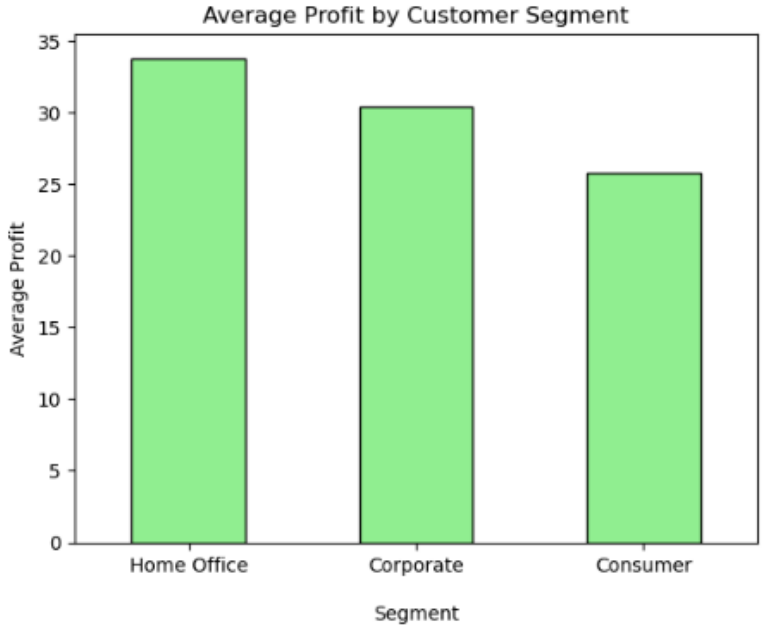


- The monthly sales trend shows sales spikes in December of each year, reaching record highs in December 2017. This pattern strongly suggests that end-of-year holiday promotions and campaigns are the primary drivers of sales growth.
- Additionally, there are noticeable dips in sales during the spring and late summer months, indicating where possible customer demand has decreased. To capitalize on these trends, the company should invest more in marketing of the holiday season to maximize profits.
- For the spring and summer months, launching targeted promotions or new product releases could help smooth out sales fluctuations and maintain steady revenue throughout the year.
- 

Conclusion:

- It is important to note that the Superstore is already operating profitably with sales increasing every year. The Random Forest Model was our winning predictive model scoring a 94% performance accuracy. After exploring the data the trend seems to be that more sales equals more profits without making any changes and decreasing our heavily discounted items can boost profitability. Our most profitable customers are in the “Home Office” and "Corporate" segments so offering those customers a discount of 20% or lower can potentially produce more sales.

- 

- 

- **Future Scope of Work:**
  - If the business executives find the Random Forest Model to be useful in their decision making to increase profitability, we could automate the model's use and create an excel plugin for efficient usability.