Understanding the Driving Factors of the Sales using Regression Models

LI LI

JUNE 2022

SPRINGBOARD

DATA SCIENCE CAREER TRACK

Problem & Context

- ❑ The Superstore Giant is seeking a better understanding of what works best for them and which products, regions, and categories they should target or avoid.
- ❑ The main interest in this project is to understand the key factors driving the sales.
- ❑ This project trained different types of regression models, both linear and non-linear, and models are evaluated and selected based on metrics, such as R2 and Rmse

Goals

- ❑ The primary stakeholders for this project would be owners, suppliers, investors, etc.
- ❑ Any sort of retail store which uses sales as a performance benchmark would find this project to be interesting and meaningful.

Key Results

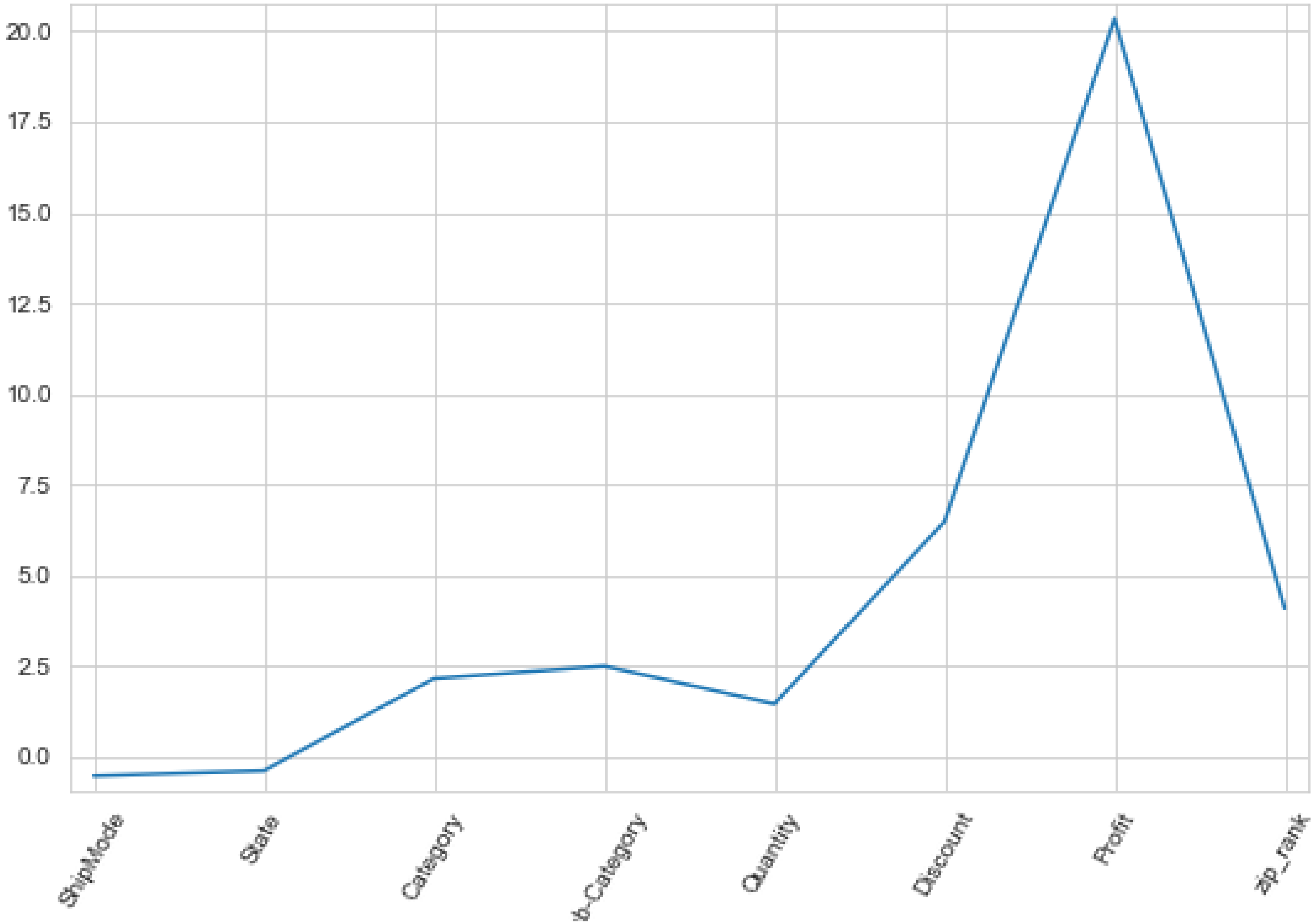
Important Features uncovered during modeling:

❑ Profit

❑ Quantity

❑ Postal code (city/state/region)

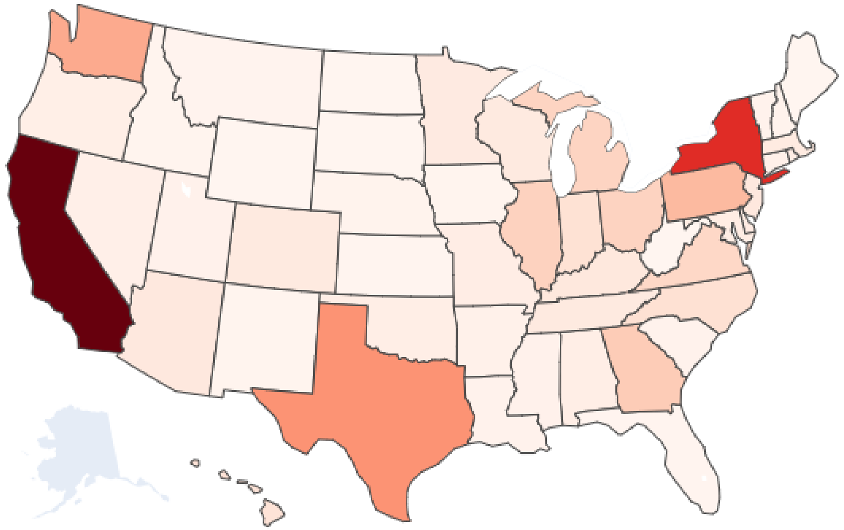
❑ Discount

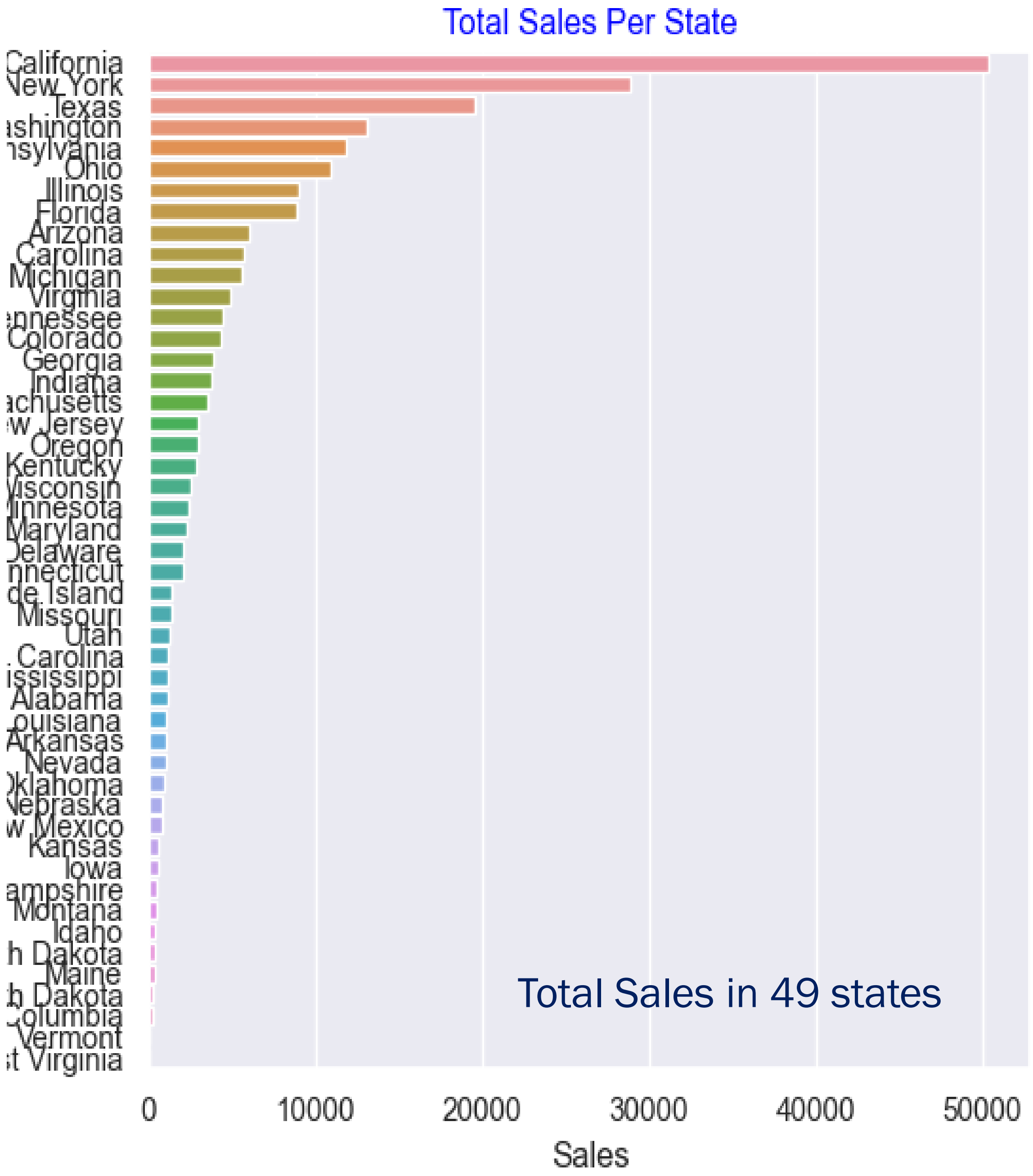


Preliminary Data Analysis

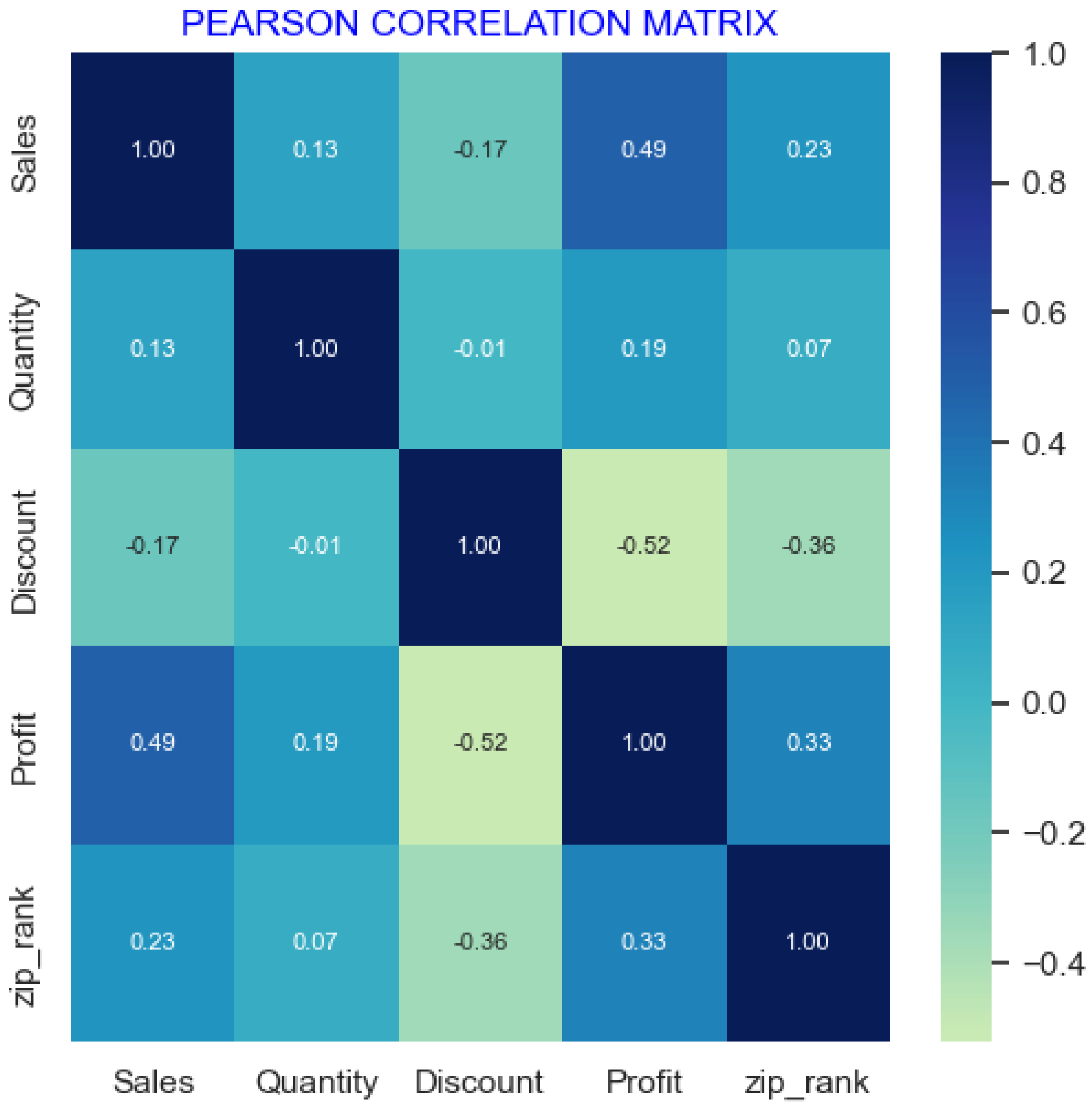
- ❑ The data set is taken from Kaggle (https://www.kaggle.com/datasets/vivek468/superstore-dataset-final)
- ❑ It contains a superstore sales and profits information between 2014 -2018

Exploratory Data Analysis (EDA)





Total Sales in 49 states



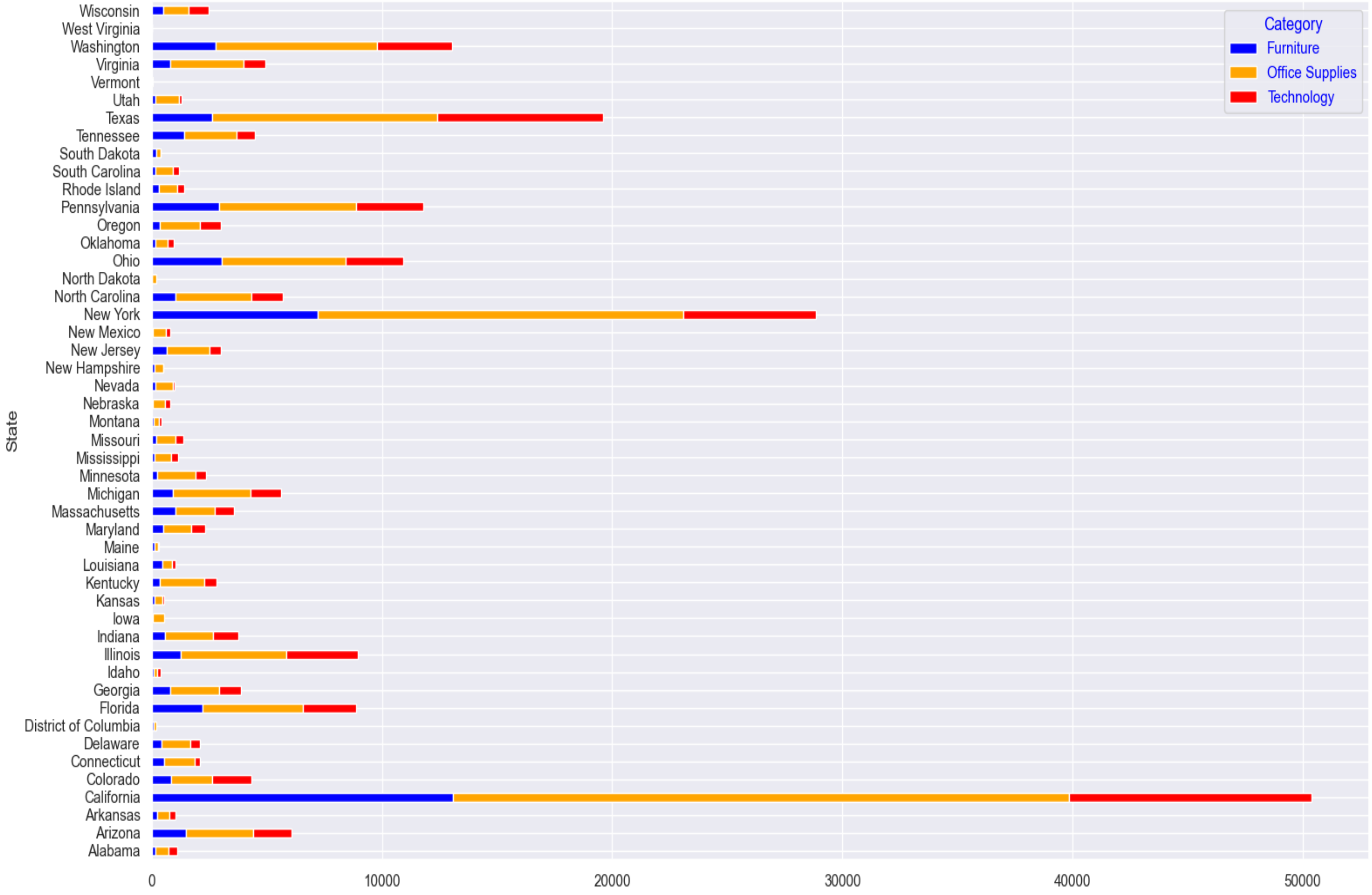
Correlation Coefficients

- Profit: 0.49
- Quantity: 0.13
- Discounts: -0.17
- Zip_rank: 0.23

States of Largest Sales

The most important states are:

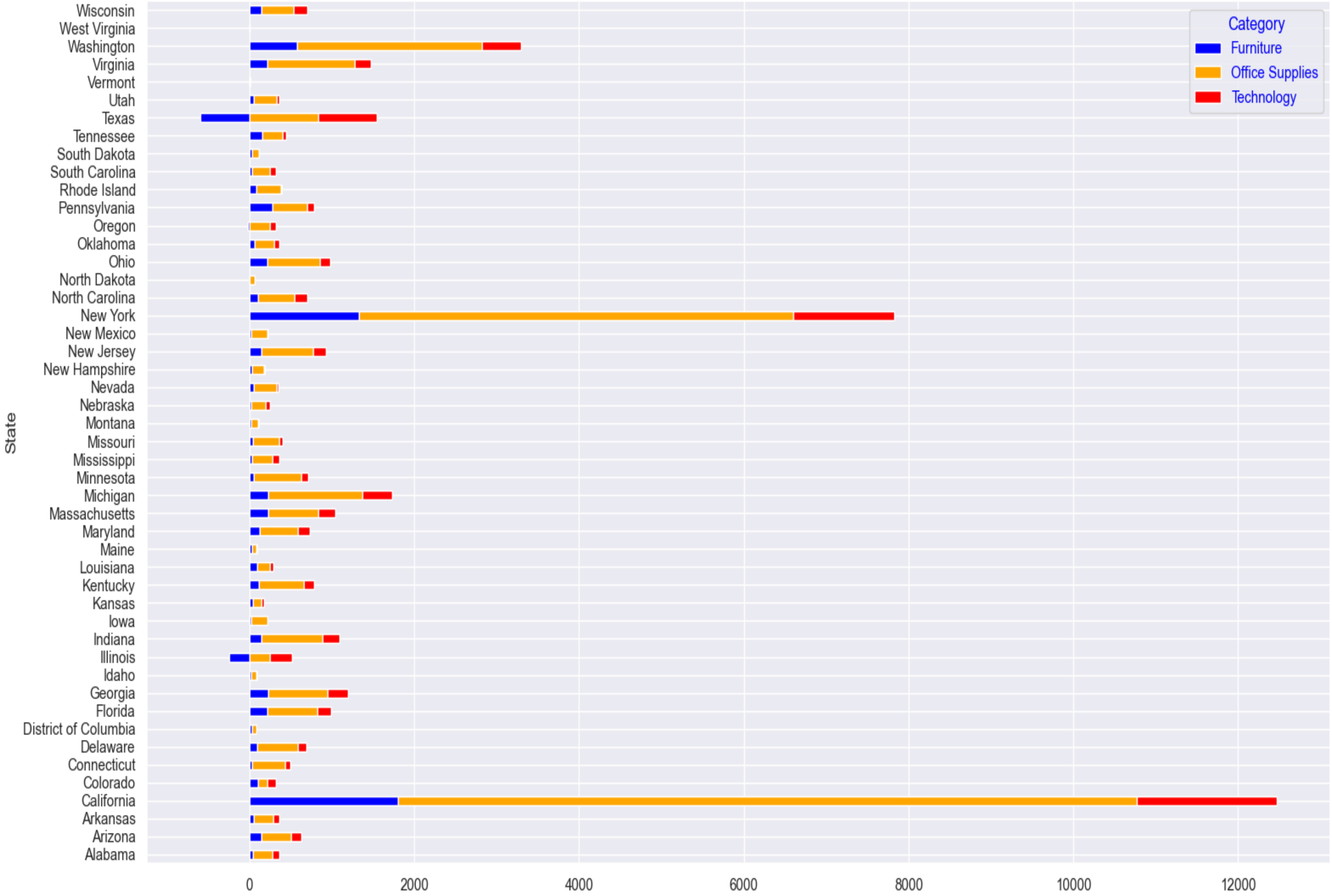
- California in the West Region
- New York in the East region

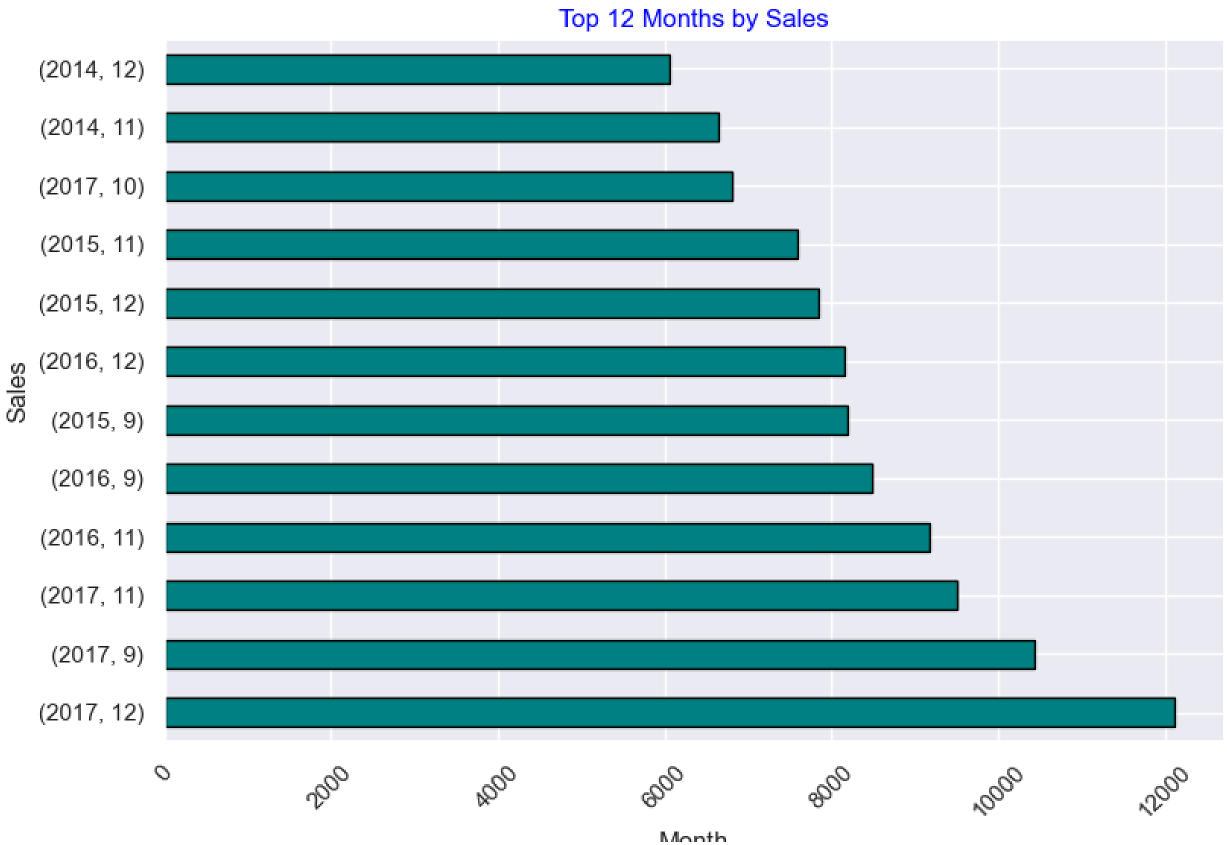


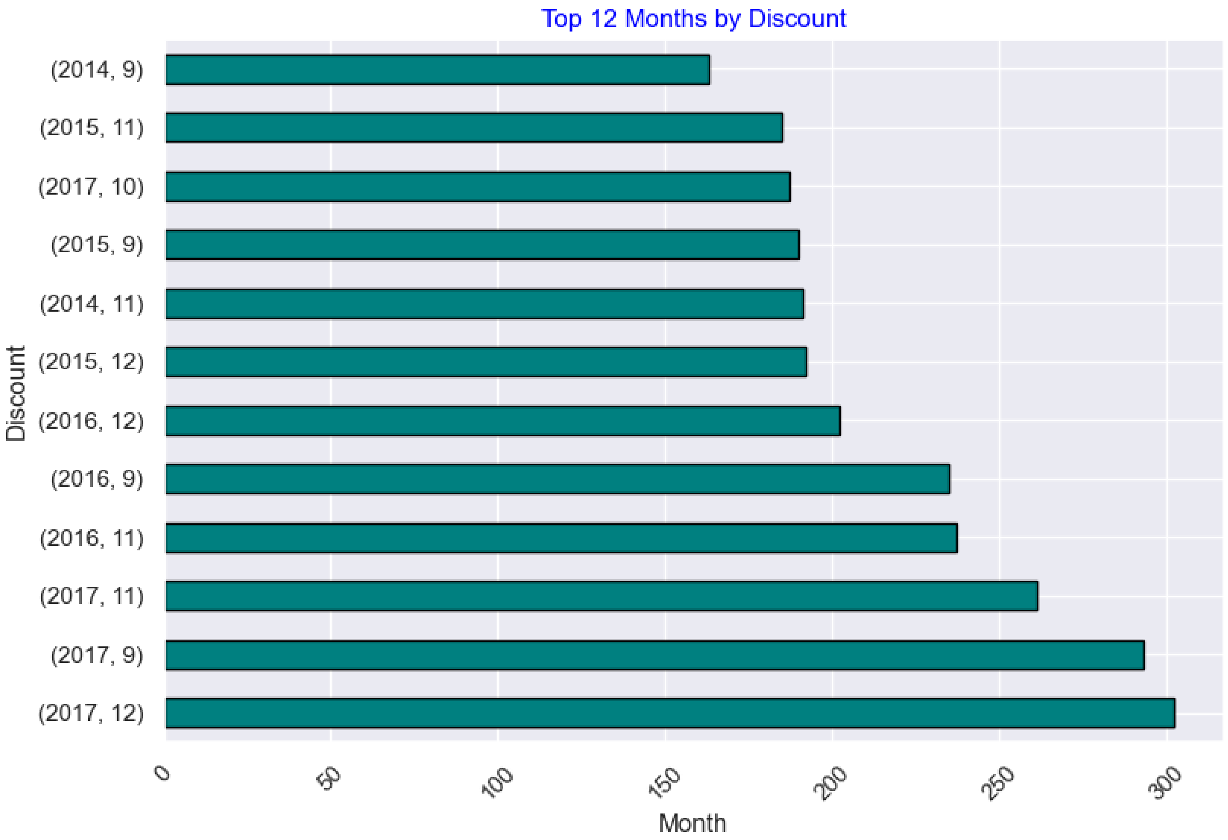
Most Profitable States

The most profitable states are:

- California in the West Region
- New York in the East region

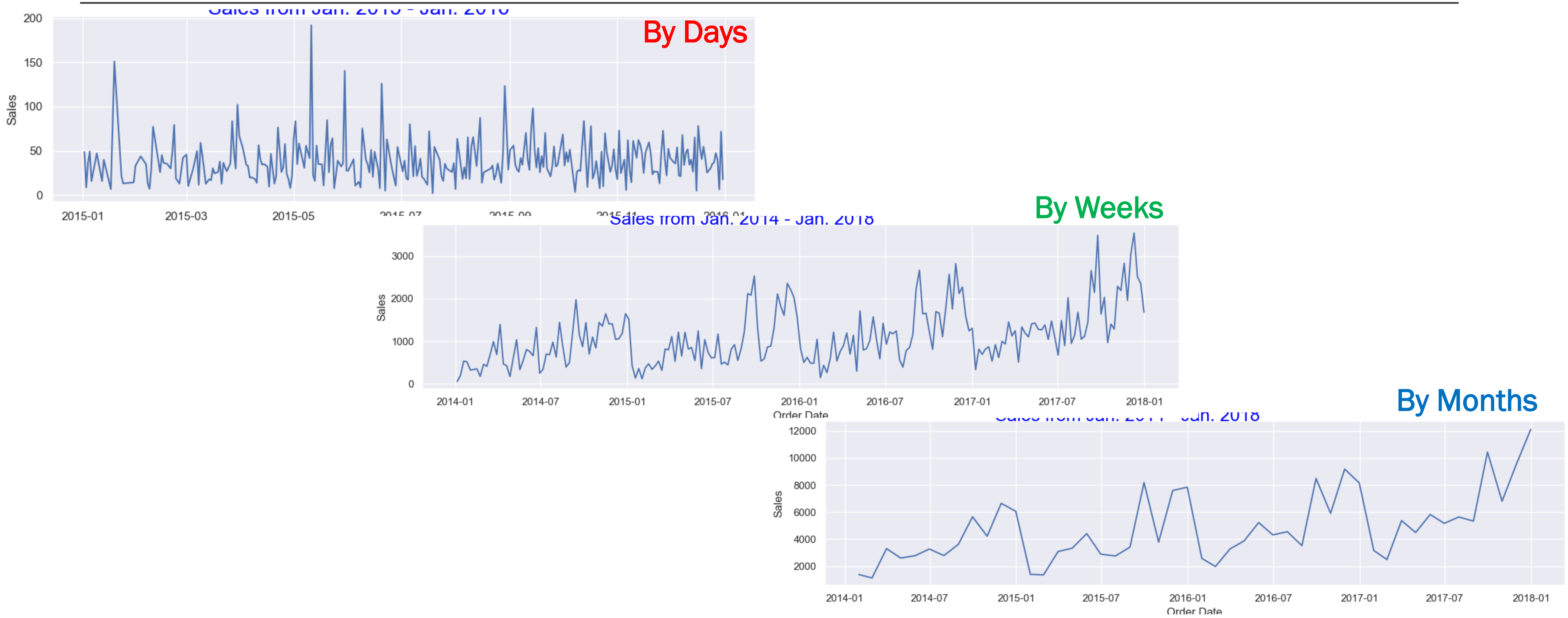


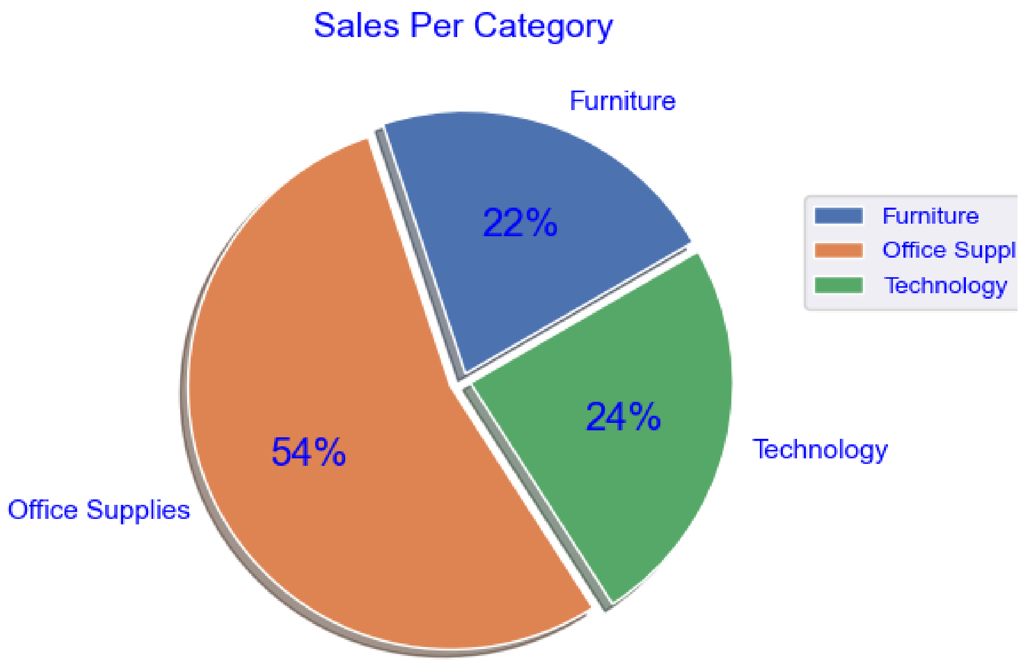


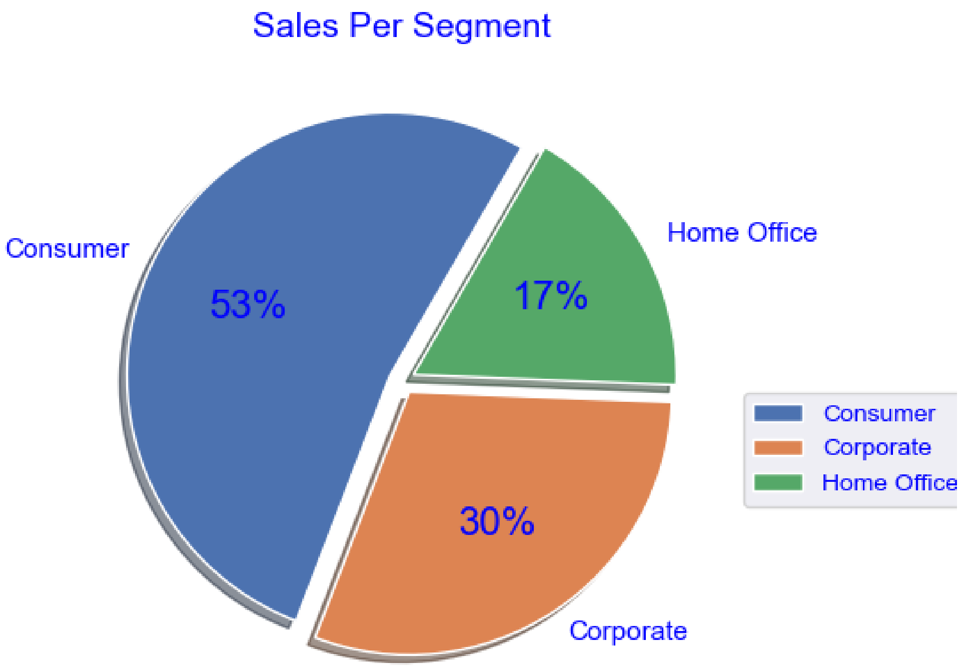


Big Sales in September, November and December

Sales by Days, by Weeks, By Months from Jan. 2014 – Jan. 2018

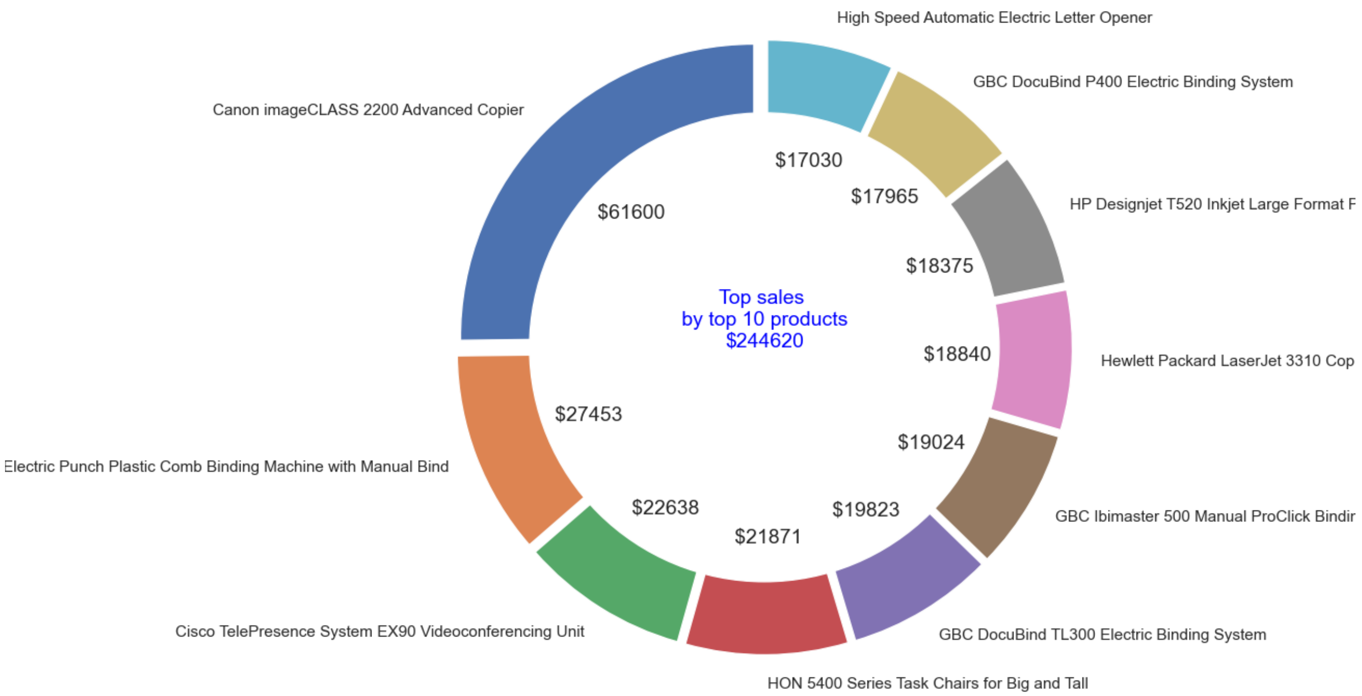


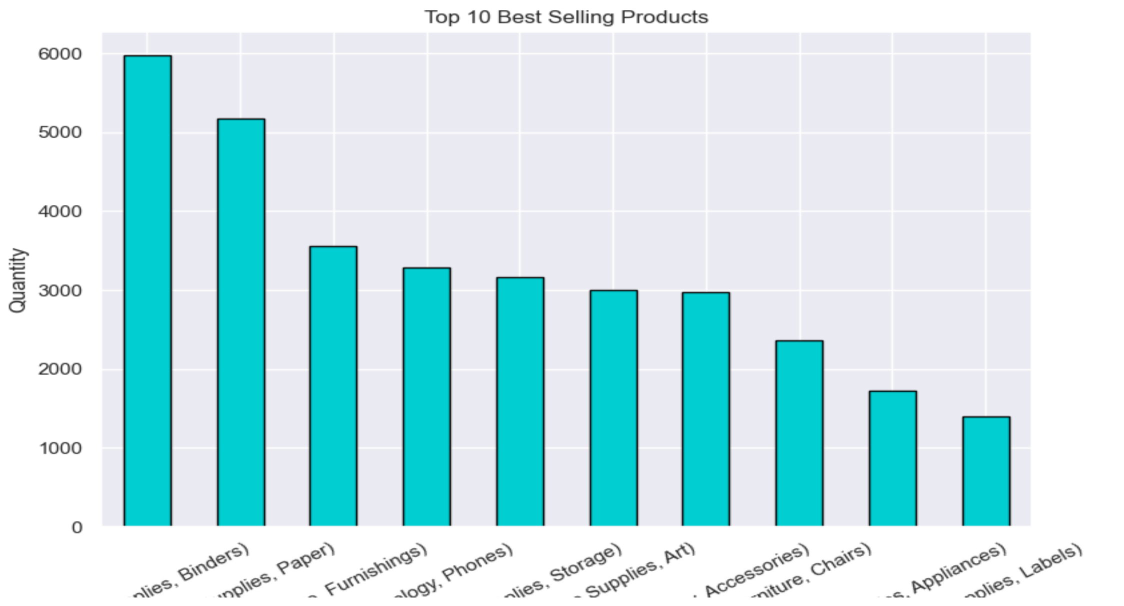


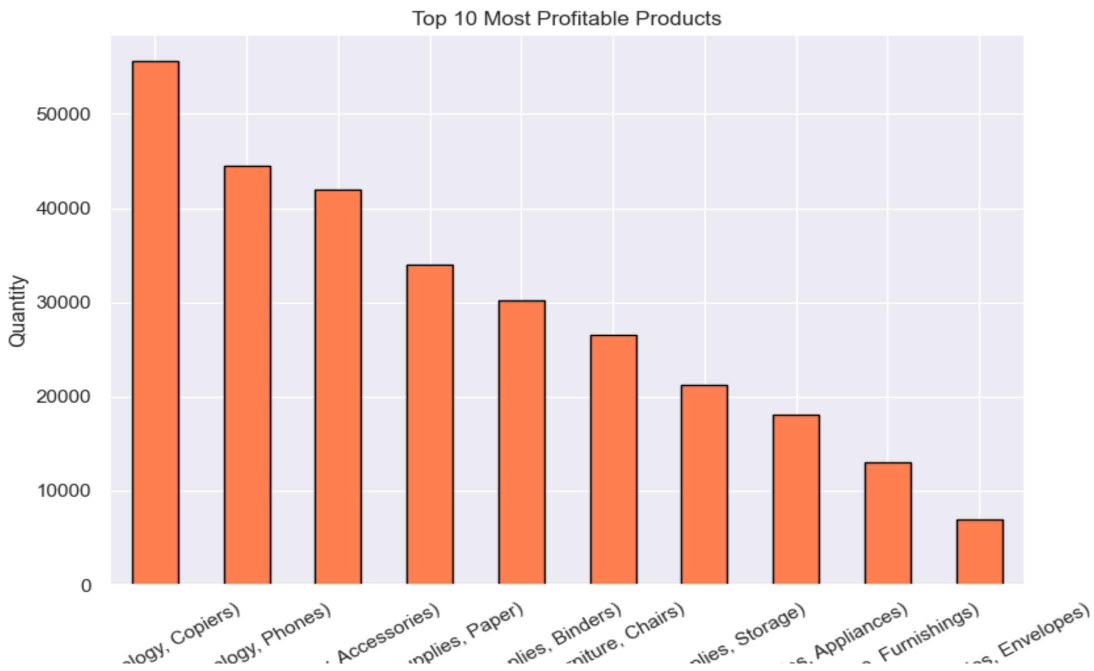


Category and Segment Most Contributed to Sales

Best Selling and Most Profitable Products



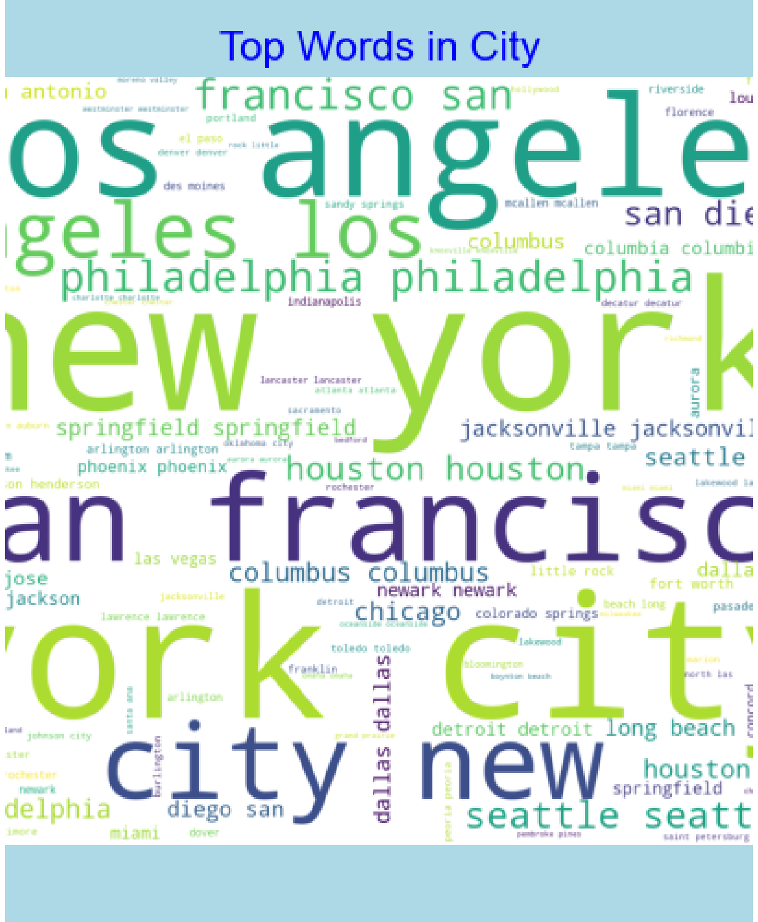


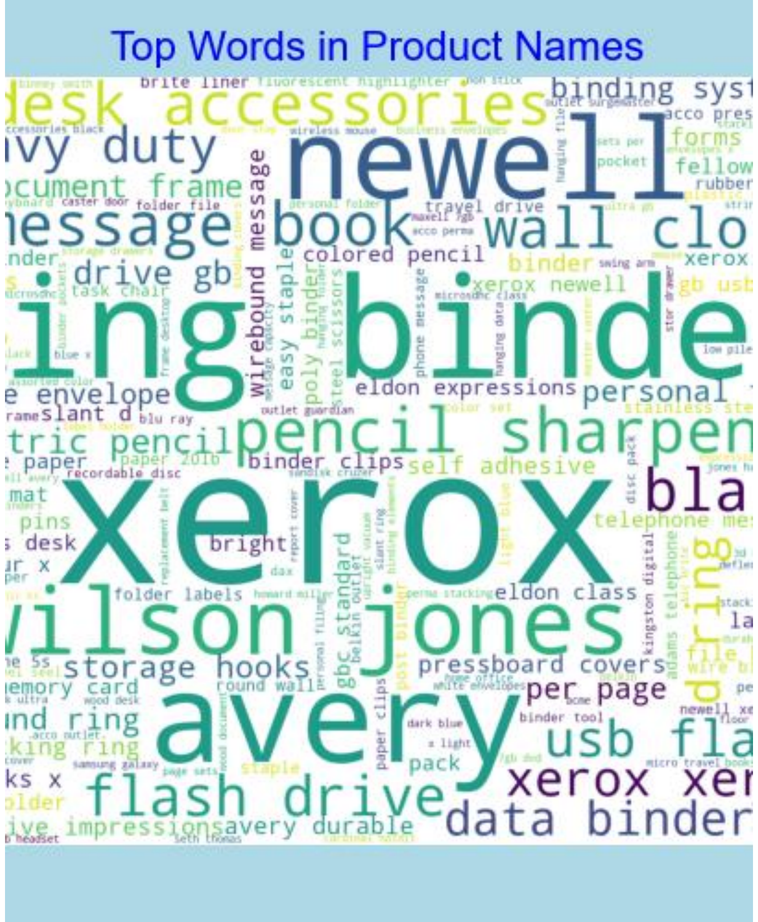


Images Generated by Wordcloud

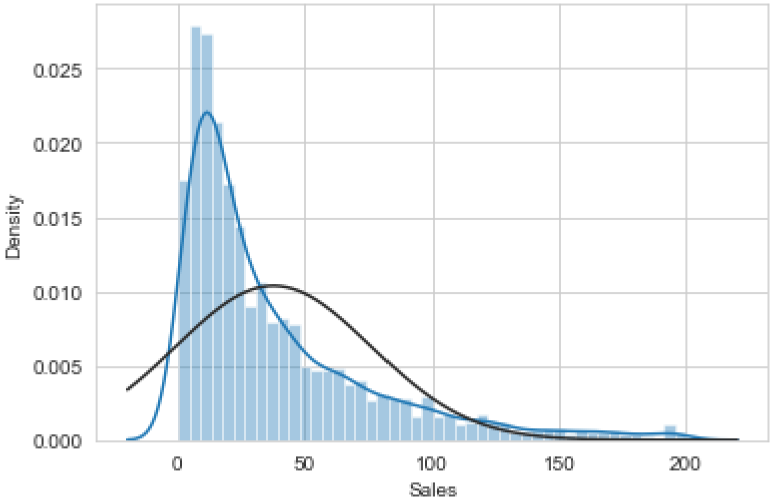
- **"xerox" and "ring binder"** are the two most important words that were repeatedly mentioned in the product names.

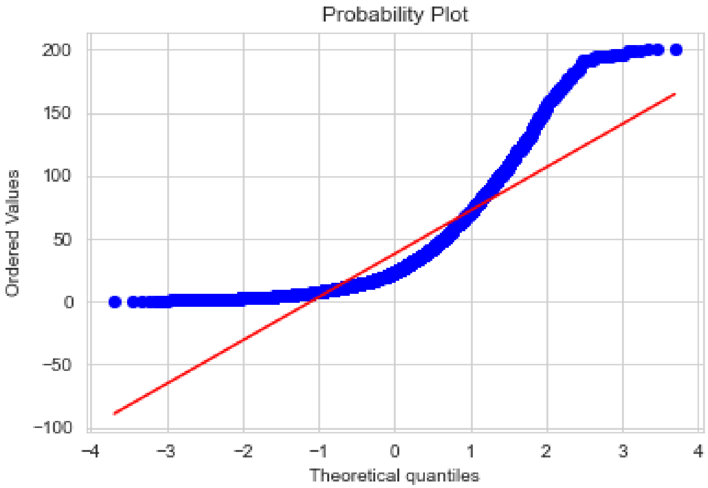
- **'New York'** in the East region, **'Los Angles' and ' San Francisco'** in the West region repeatedly appeared in the orders which make sense as those are the most densely populated cities.





Target Variable – ‘Sales’



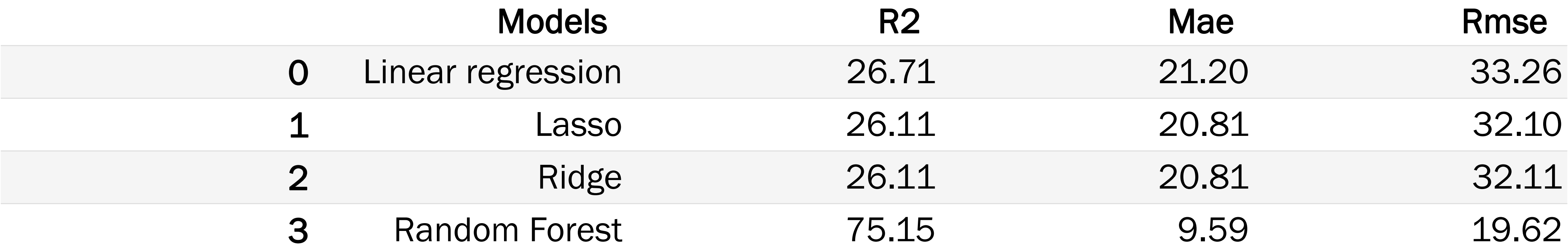


Regression Modeling

The machine learning models chosen were the following:

- ❑ Linear regression model
- ❑ Ridge regression
- ❑ Lasso Regression
- ❑ Random Forest regressor

Model Evaluation and Selection



	Models	R2	Mae	Rmse
0	Linear regression	26.71	21.20	33.26
1	Lasso	26.11	20.81	32.10
2	Ridge	26.11	20.81	32.11
3	Random Forest	75.15	9.59	19.62

- The results show that the best performances in terms of R squared (R^2) and Root Mean Square Error (RMSE) correspond to random forest Regressor.

- According to the result of R^2 and Root Mean Squared of these 5 models, the relationship between the features and the target variable is not clearly linear and shows the presence of non-linearity in the data..

Summary

- Since the data was collected from 2014 - 2018, this model might not fully reflect all the price changes recently in the market. Moreover, features that were used in developing this algorithm, might not be enough to sufficiently describe the sales. The sales range is mainly less than ~$100, which is a bit small in my opinion.

- And lastly, the market of densely populated urban areas is definitely different from that of the rural area. This means more data needs to be collected for different regions across all 49 states in the US in order to efficiently predict the sales for any particular area.