DA 620

Data Driven Decision and Business Intelligence

Capstone Project

- Deepthi Velakaturi

Table of Contents

Background.....3
Problem Scenario / Business Issue.....3
Objective / Goals of the Project.....4
Data Exploration / Data Visualization.....4
Data Manipulation.....4
Data Analysis and Visualization.....5
Methodology / Model Building.....11
Conclusion / Recommendations.....15
Bibliography / References.....16

Background

We have made use of the **Walmart Global Store** dataset which is available over the internet (***UCI machine learning repository***). This dataset has the Retail data for 4 years. It also has the sales data of 45 stores. It includes the details about the product and customer for sales analysis.

We will be making use of Python for our Analysis.

Problem Scenario / Business Issue

We will perform EDA on the dataset along with Sales performance analysis

1. What is the percentage of sales in the last year?
2. What is the sales rate for each month of the year?
3. How are the sales trending over time?
4. What are the top-selling categories?
5. What are the top-selling products (Top 10)?
6. Top Selling Sub categories?
7. Which ship mode is most frequently chosen by customers
8. Which states contribute the most to the sales revenue?
9. Where is the best place to put our new ad?
10. Who are our best customers (Top 5)?
11. Which region has the most orders?

We will also create a forecasting model to forecast / predict the sales and also a prediction model for sales prediction.

Objective / Goals of the Project

Goal of the project is to

- Provide a thorough analysis of the dataset
- Provide all the answers for the business problems and build a forecast / prediction model for sales analysis.

Data Exploration / Data Visualization

Multiple techniques are used to explore the dataset

Lists are

1. Shape of dataframe
2. Datatype info
3. Summary
4. List of Columns
5. Check for null values
6. Check the values for postal code which are ***NAN***

Data Manipulation

The following is done for Data Handling (Manipulation)

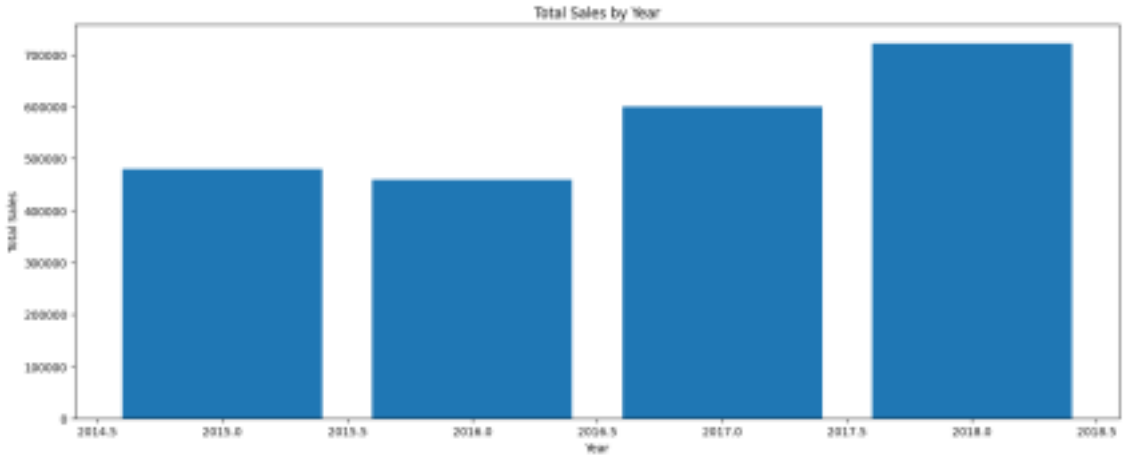
1. Removes duplicate rows if any from the Dataset
2. Change data type
3. Drop unnecessary columns
4. Sort values by Order Date
5. Fill ***NaN*** values

Data Analysis and Visualization

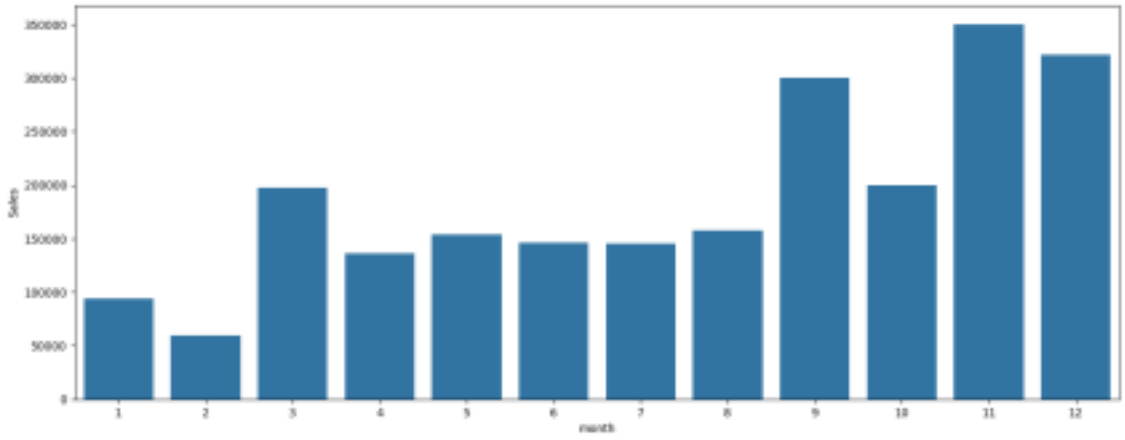
All the questions from the project problems are analyzed in a Python notebook sequentially.

We will perform EDA on the dataset along with Sales performance analysis

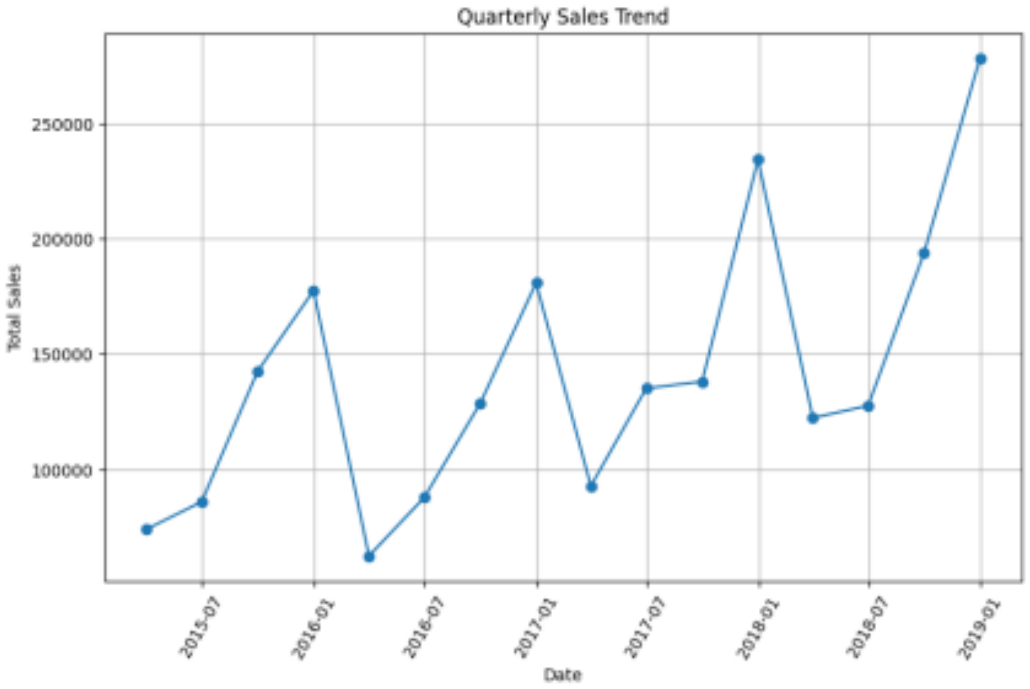
1. What is the percentage of sales in the last year?



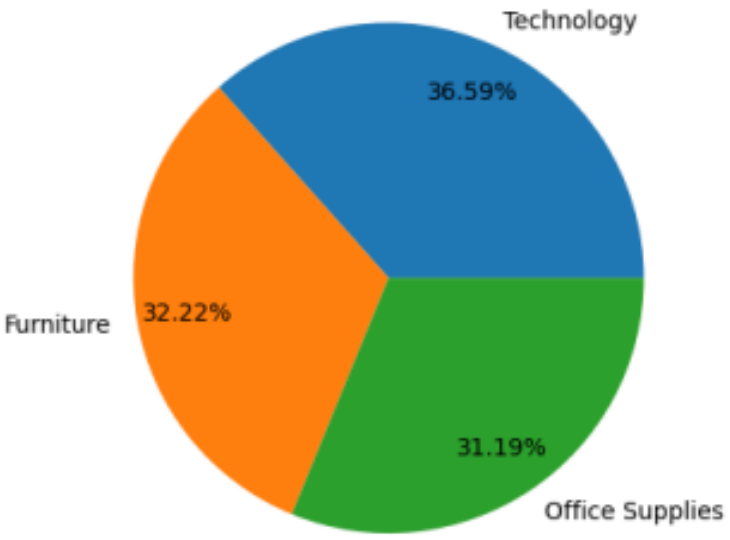
2. What is the sales rate for each month of the year?



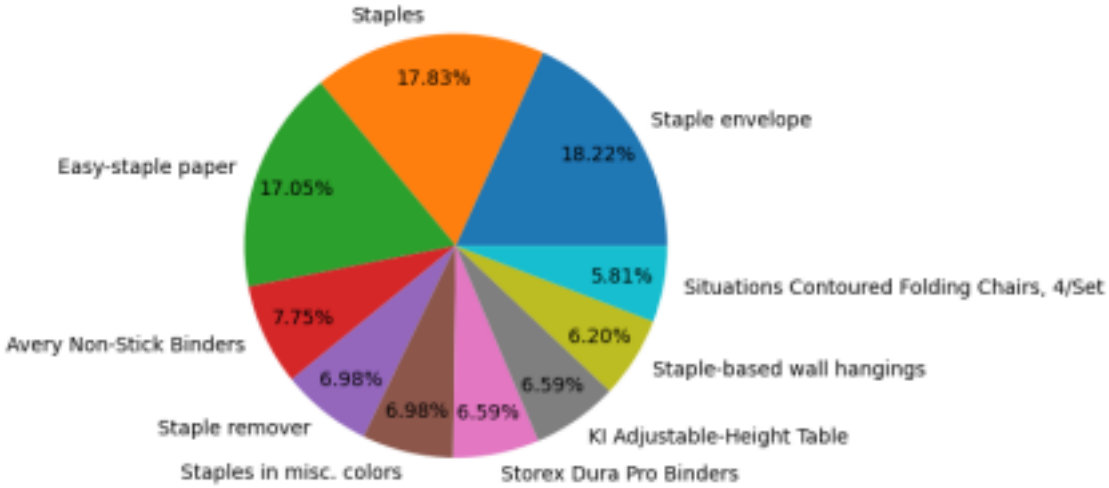
3. How are the sales trending over time?



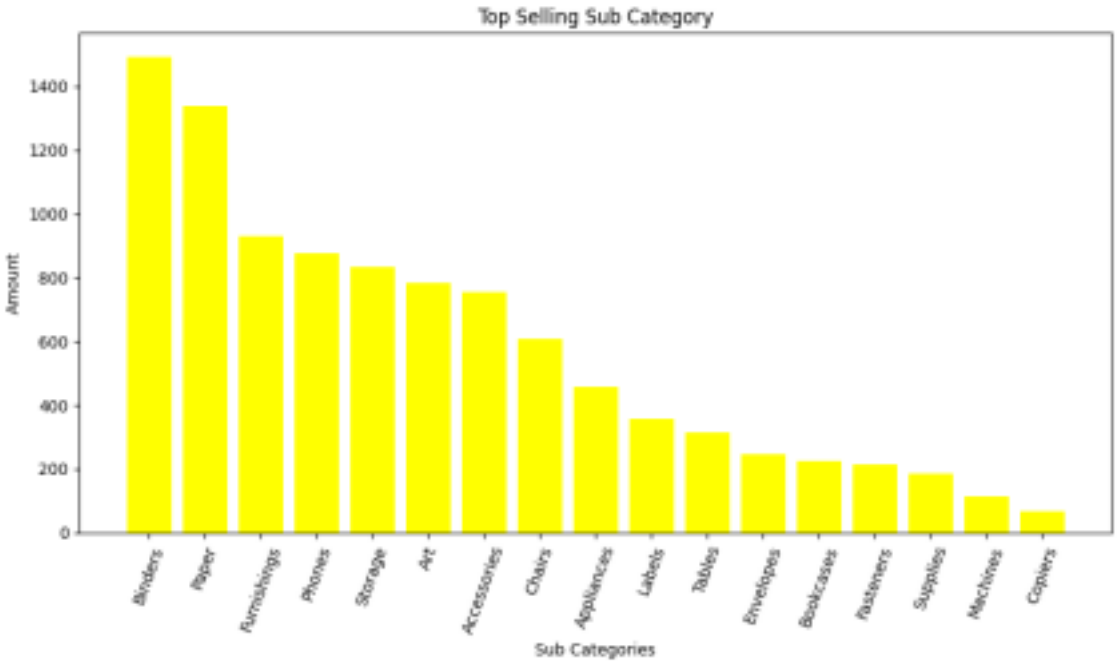
4. What are the top-selling categories?



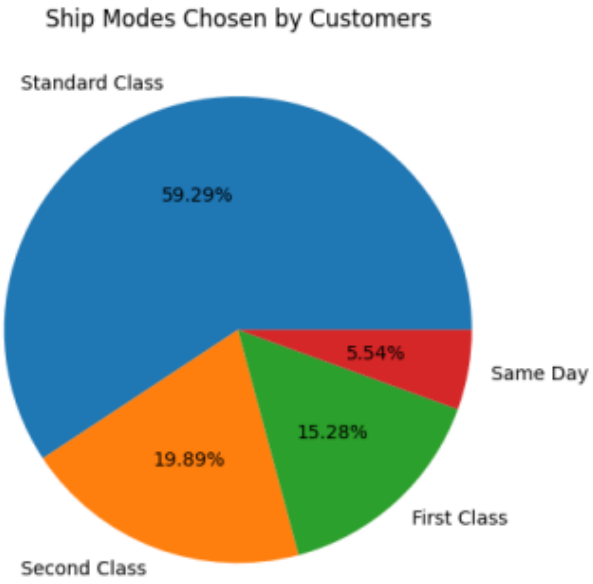
5. What are the top-selling products (Top 10)?



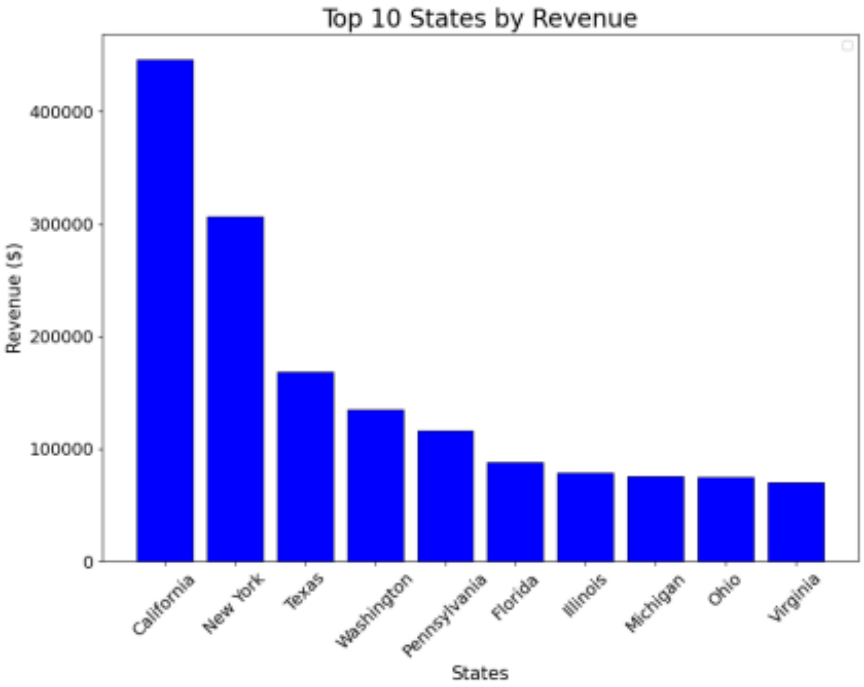
6. Top Selling Sub categories?



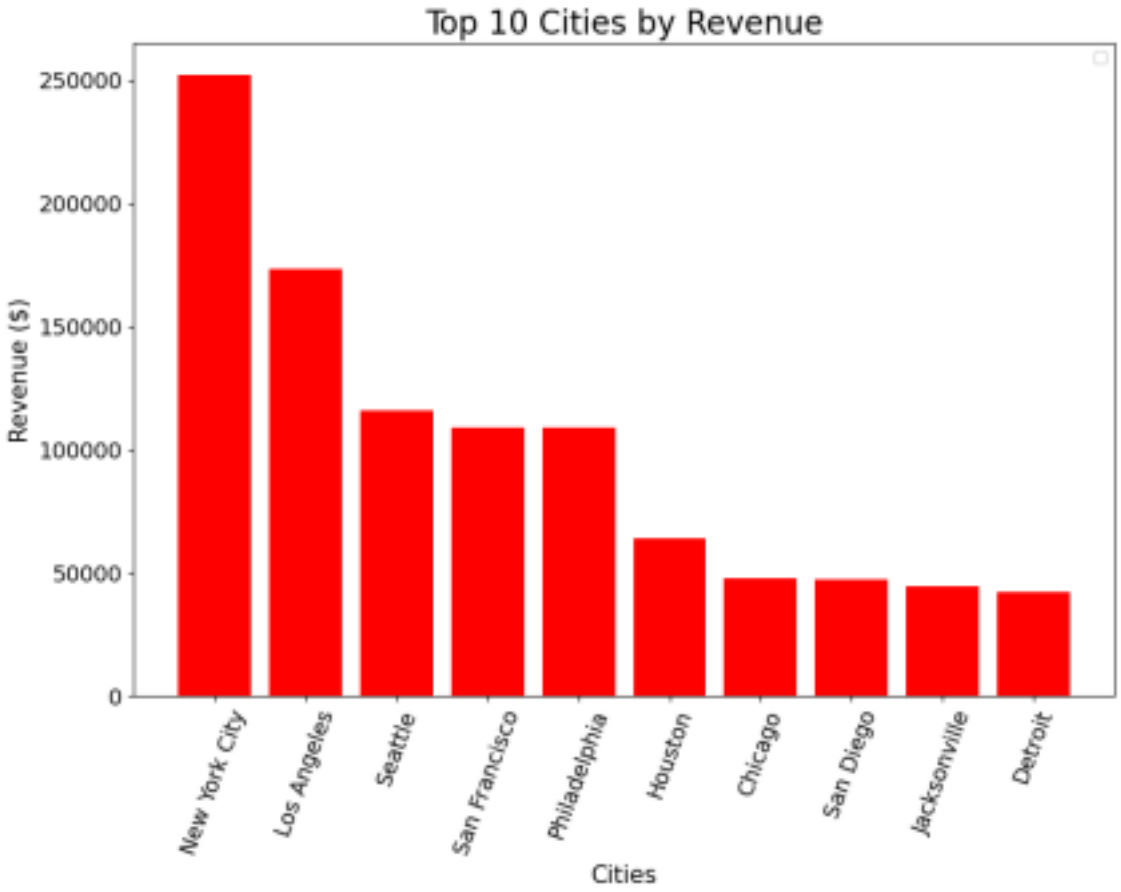
7. Which ship mode is most frequently chosen by customers?



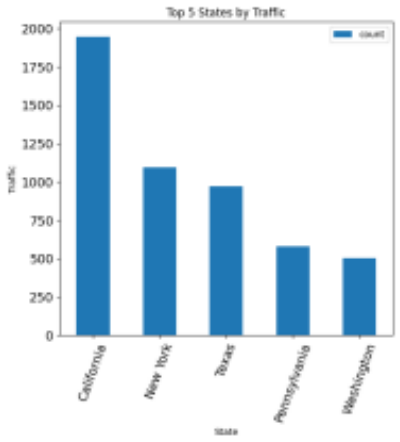
8. Which states contribute the most to the sales revenue?

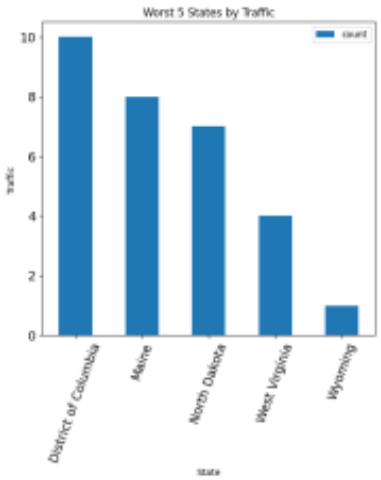


9. Which states contribute the most to the sales revenue?

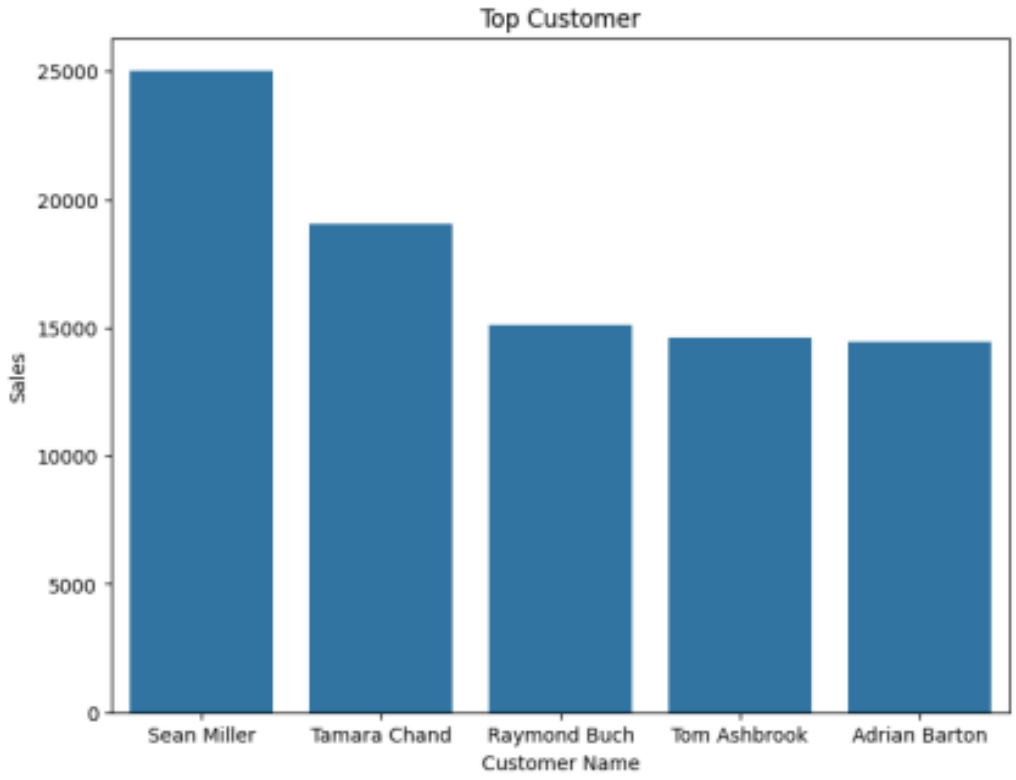


10. Where is the best place to put our new ad?

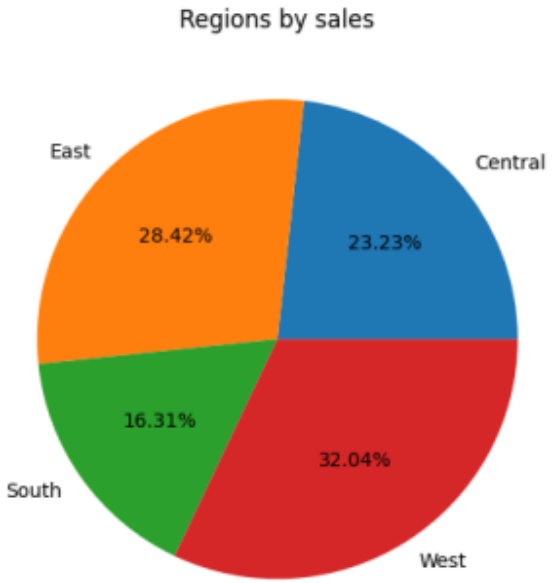




11. Who are our best customers (Top 5)?



12. Which region has the most orders?



Methodology / Model Building

Here we have build 2 forecasting models and 2 regression models

Forecasting models

Modeling with Prophet

Prophet is an open-source forecasting tool developed by Facebook that is designed for forecasting time series data with strong seasonal patterns and multiple seasonality. It's particularly useful for working with time series data that includes holidays and other special events. Prophet was created to be user-friendly and to handle many of the challenges that arise when making time series forecasts.

Modeling with SARIMA

Seasonal Autoregressive Integrated Moving Average (SARIMA) is a time series forecasting model that extends the capabilities of the Autoregressive Integrated Moving Average (ARIMA) model to account for both seasonality and non-seasonal components in time series data. SARIMA is a powerful method for modeling and forecasting time series data that exhibit both trend and seasonal patterns.

Regression models

Ridge Regression

Ridge Regression is a variation of linear regression that addresses some of the issues of linear regression. Linear regression can be prone to overfitting when the number of independent variables is large, this is because the coefficients of the independent variables can become very large leading to a complex model that fits the noise of the data. Ridge Regression solves this issue by adding a term to the linear regression equation called L2 regularization term, also known as Ridge Penalty,

LASSO Regression

Similar to Ridge regression, LASSO (Least Absolute Shrinkage And Selection Operator) is another variation of linear regression that addresses some of the issues of linear regression.

It is used to solve the problem of overfitting when the number of independent variables is large. Lasso Regression adds a term to the linear regression equation called L1 regularization term, also known as Lasso Penalty, which is the sum of the absolute values of the coefficients multiplied by a regularization parameter lambda.

Model Selection

1. R^2 Score

R^2 score, also known as the coefficient of determination, gives the measure of how good a model fits to a given dataset. It indicates how close the predicted values are to the actual values.The R^2 value ranges from $-\infty$ to 1.

A model with negative R^2 value indicates that the best fit line is performing worse than the average fit line.

2. Mean Absolute Error (MAE)

Mean Absolute Error is the average of the sum of absolute difference between the actual values and the predicted values. Mean Absolute Error is not sensitive to outliers. MAE should be used when you are solving a regression problem and don't want outliers to play a big role in the prediction.

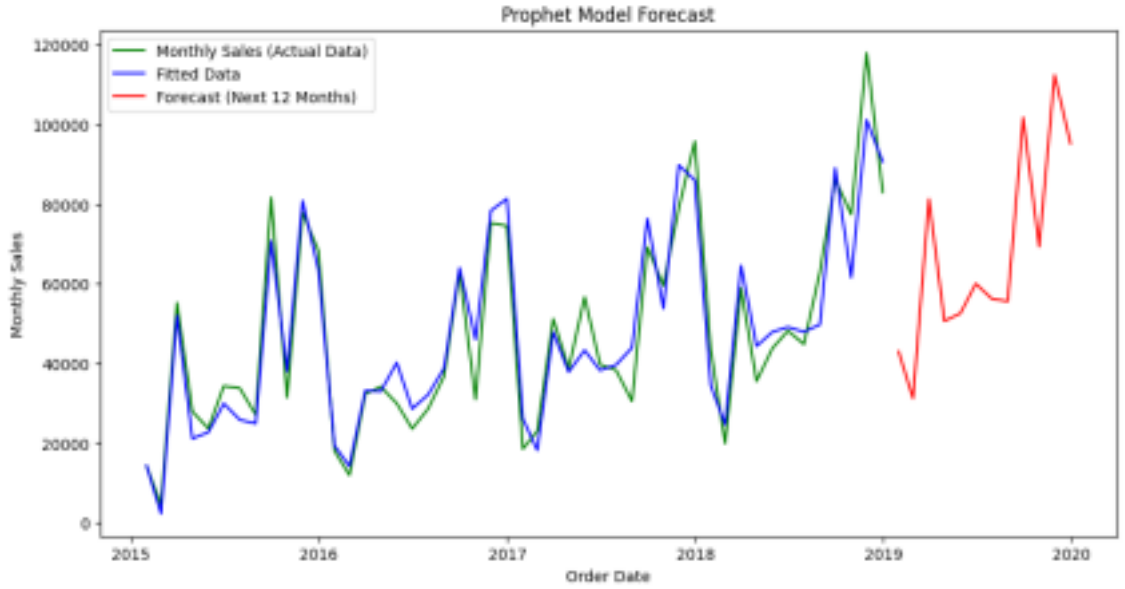
3. Explained Variance Score

This computes the explained variance regression score, if $\hat{y}$ is the estimated target output and y the corresponding (correct) target output and Var is variance (square of the standard deviation) Here, 1 is the best evaluation score possible for a model, and <0 are considered to be not properly trained models.

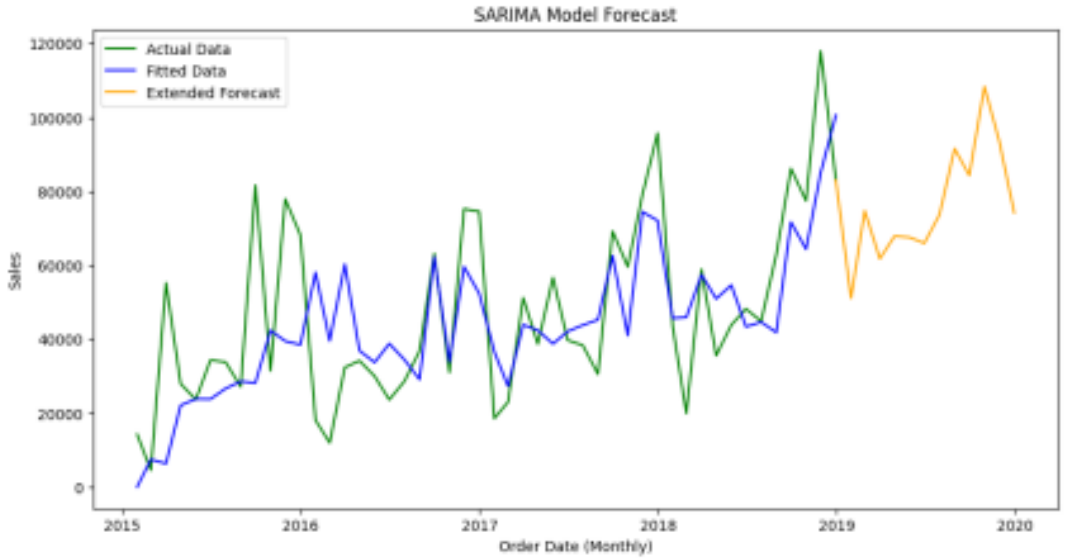
Forecast model

For this we have calculated an R2 score to check the accuracy of the model.

For Prophet: R-squared score for fitted data is 0.91



For SARIMA: R-squared score for fitted data is 0.40



So Prophet is much better than SARIMA

For regression model

For this we have calculated an R2 score, MAE and Explained Variance Score to check the accuracy of the model.

For Ridge Regression
R2: 0.393833167723699 03
MAE: 1039.8681226918943
Explained Variance Score: 0.3938570442532675

For LASSO Regression
R2: 0.36467217559443177
MAE: 1067.22721055767
Explained Variance Score:0.3647242651433018

So Ridge is slightly better than LASSO

Conclusion / Recommendations

We can use the Prophet model for forecasting the sales data, since its accuracy is much better.

Based on the data analysis we have the following findings on Sales performance analysis and Business issues.

1. The sales for the current year is **20.3%** more in comparison of last year
2. **September, November and December** have the highest Sales.
3. Sales are at their peak during **January** of almost every year. Almost every year, after January, sales decrease but start to increase again from April to January.
4. Top-selling category is '**Technology'**
5. Top-selling product is **'Canon image CLASS 2200 Advanced Copier'**
6. Top-selling Sub Categories is **'Binders'**
7. **'Standard Class'** ship mode is most frequently used
8. **California** contribute the most to the sales revenue
9. We can open a new branch in California, as it is the largest state in terms of the number of orders, so it would be better to build a branch there to handle the big number of orders faster.
10. **Sean Miller** is our best customers
11. **West** region has the most orders

Bibliography / References

For Dataset

UC Irvine Machine Learning Repository https://archive.ics.uci.edu/

Technical References

https://www.geeksforgeeks.org/
https://www.analyticsvidhya.com/
https://pieriantraining.com/
https://medium.com/