# SHOPPER STOP FOOTWEAR

**Course**: ALY6020 Predictive Analytics

**Instructor**: Alakh Verma

**Case Study**: Retail Industry


**Group Members**:

**Pragati Koladiya** | **NUID**: 00102944

**Tanvi Bhagat** | **NUID**: 001083830

**Priyanka Kanukollu** | **NUID**: 001021111


# Agenda

- Introduction
- Exploratory Data Analysis (EDA)
- Pre-processing
- Models
- Comparisons of models
- Conclusion

# Aim

- Understanding the data set and analysis of the shopping pattern of the customers.
- Forecasting sales of the footwear store for year 2020 using Time Series.
- Implementing Naïve Bayes, Decision tree, and Random Forest to predict the ratings given by customer.

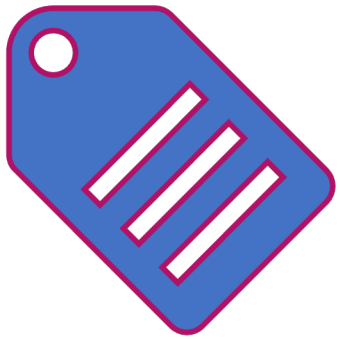
# Dataset Information

Data represents customer ratings based on the purchased footwear, and it shows the sales for every day within four years.

Dataset Includes ten features such as the date, Everyday sales, categories, brands, discount, size, etc.

The dataset contains sales information of the store from 1st Jan 2015 to 31st Dec 2019.

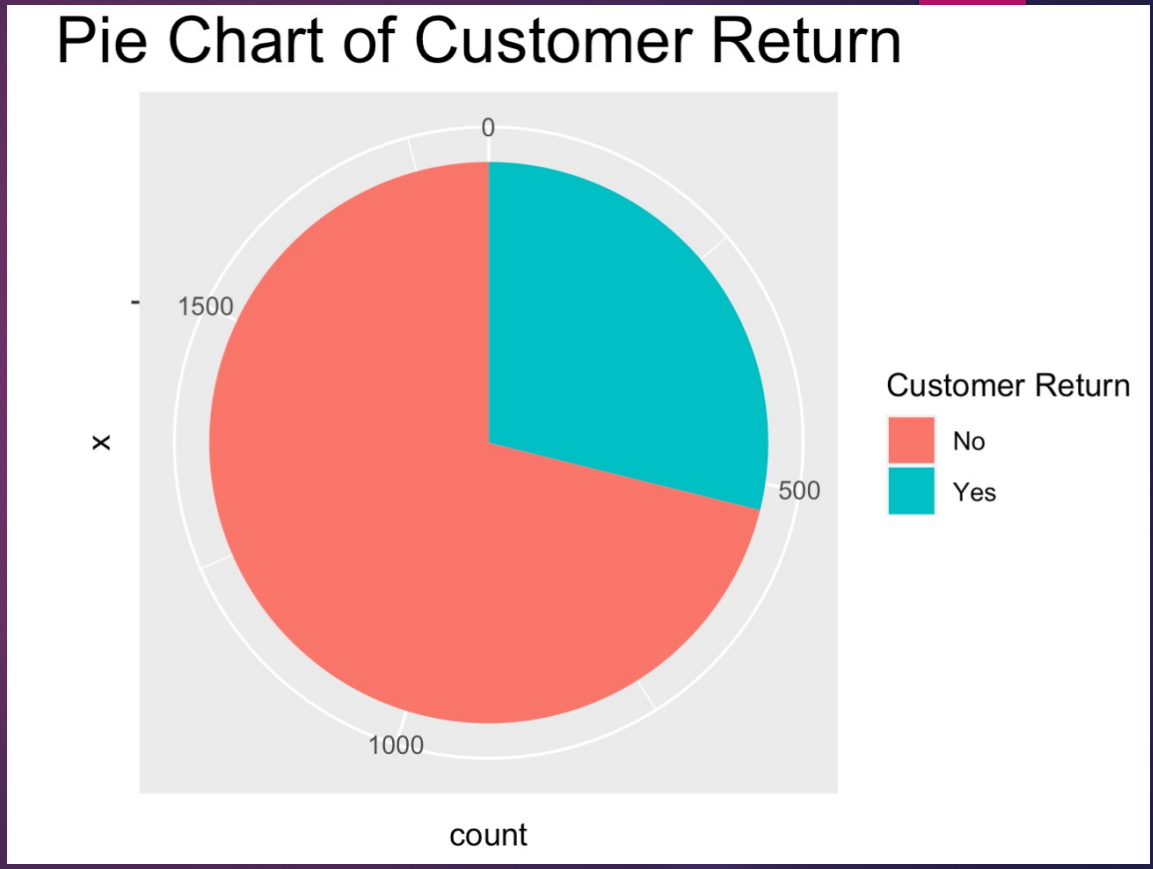
# Exploratory Data Analysis (EDA)



- Distribution of Product return?
- How many products are on discount?
- What is the products count w.r.t Ratings?
- Is there any effect of discount on sales?
- What is the footwear count by categories grouped by Customer Return?
- Brand-wise count of footwear w.r.t discount offered?
- Is there a link between product ratings and customer return?

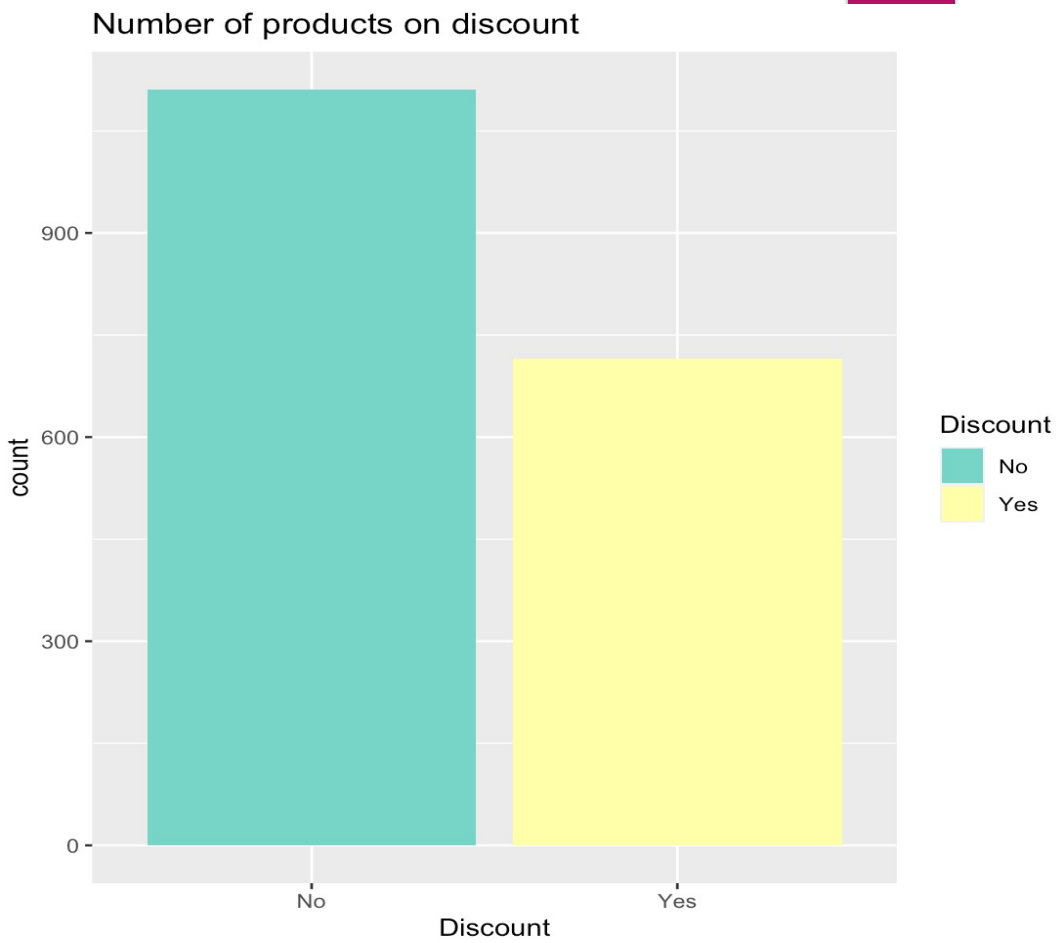
## Distribution of Product return?

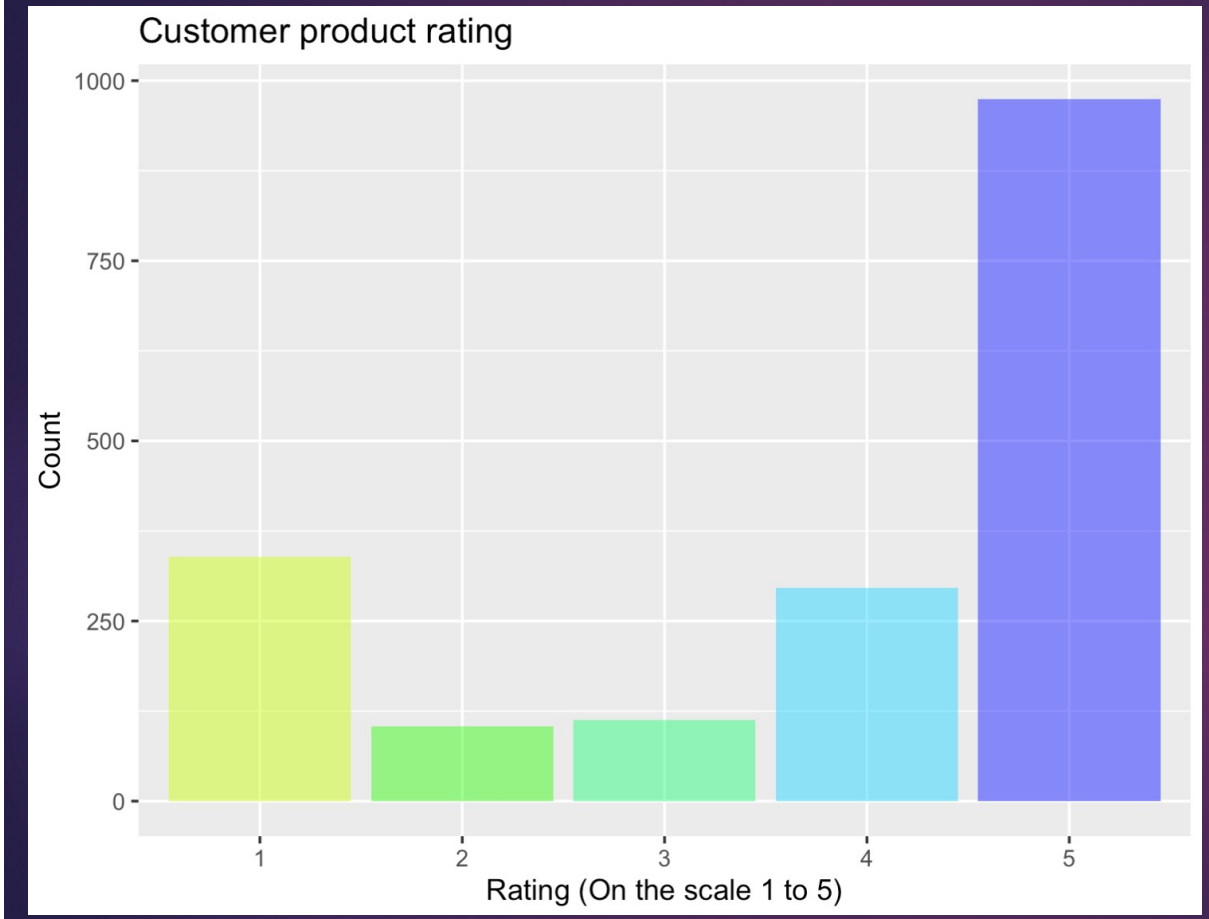
- The shop has approximately 30% product return rate.
- This is one of the important aspect that any retail store should focus on.



# How many products are on discount?

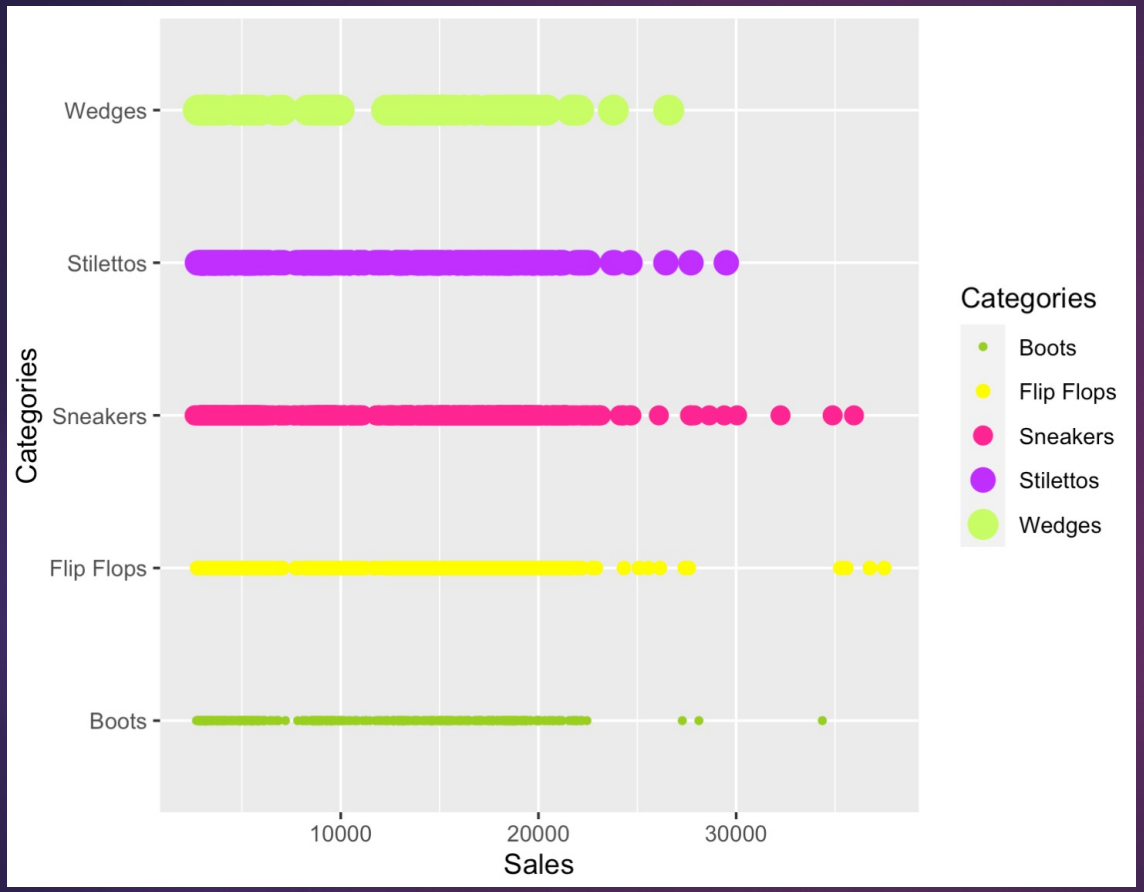
- Out of the 1826 footwears, the store offers discount on approximately 750 of them.
- The remaining 1076 pairs of footwear have no discounts and can be termed as NEW ARRIVALS.






## What is the products count w.r.t ratings?

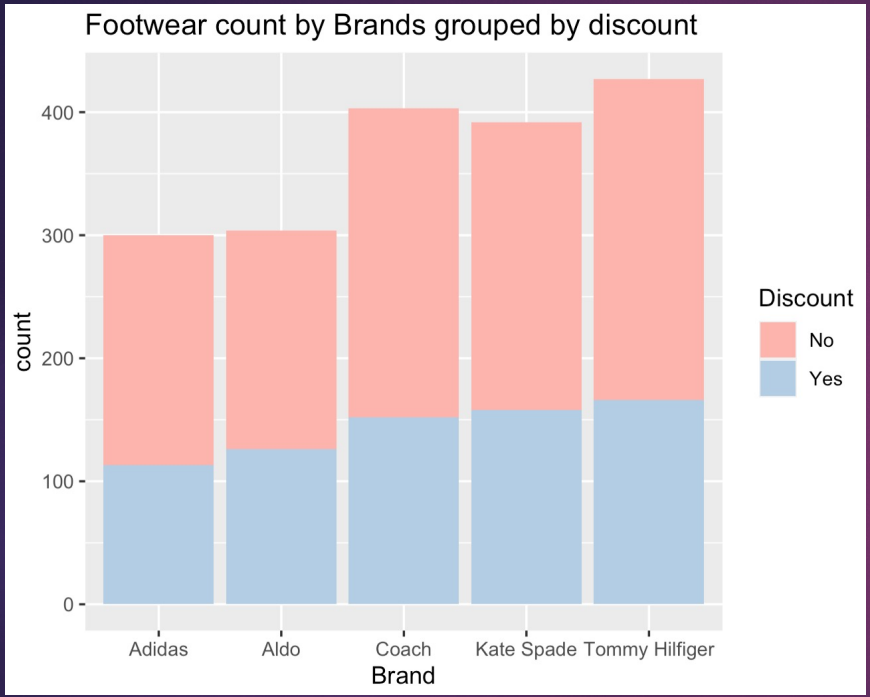
- The store has received rating 5 from maximum of the customers(around 950).
- Cannot neglect the fact that rating 1 has also around 300 count.
- The store management should work on this.

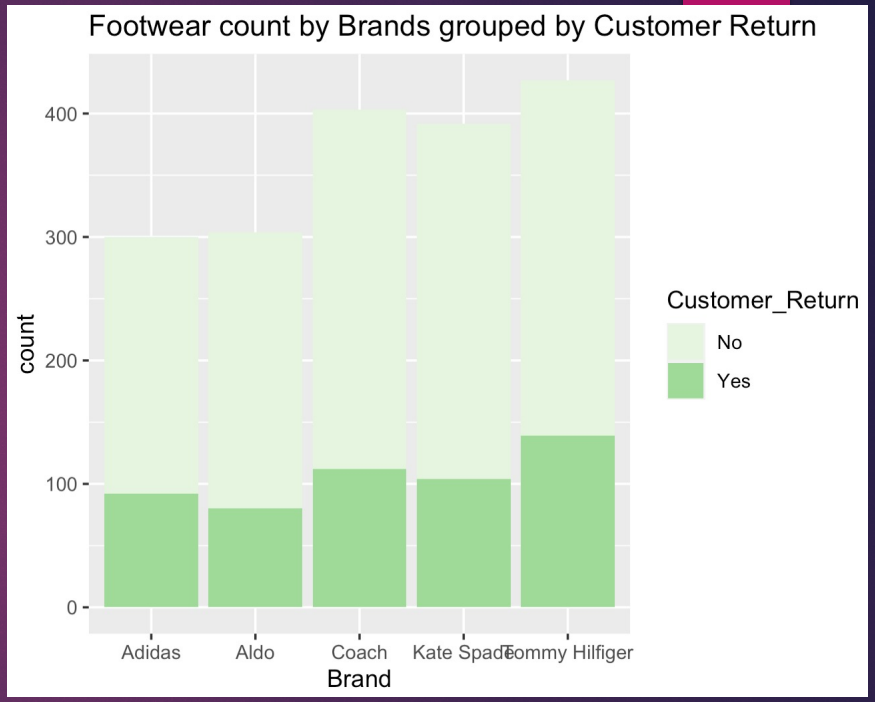



# Distribution of sale w.r.t footwear categories?

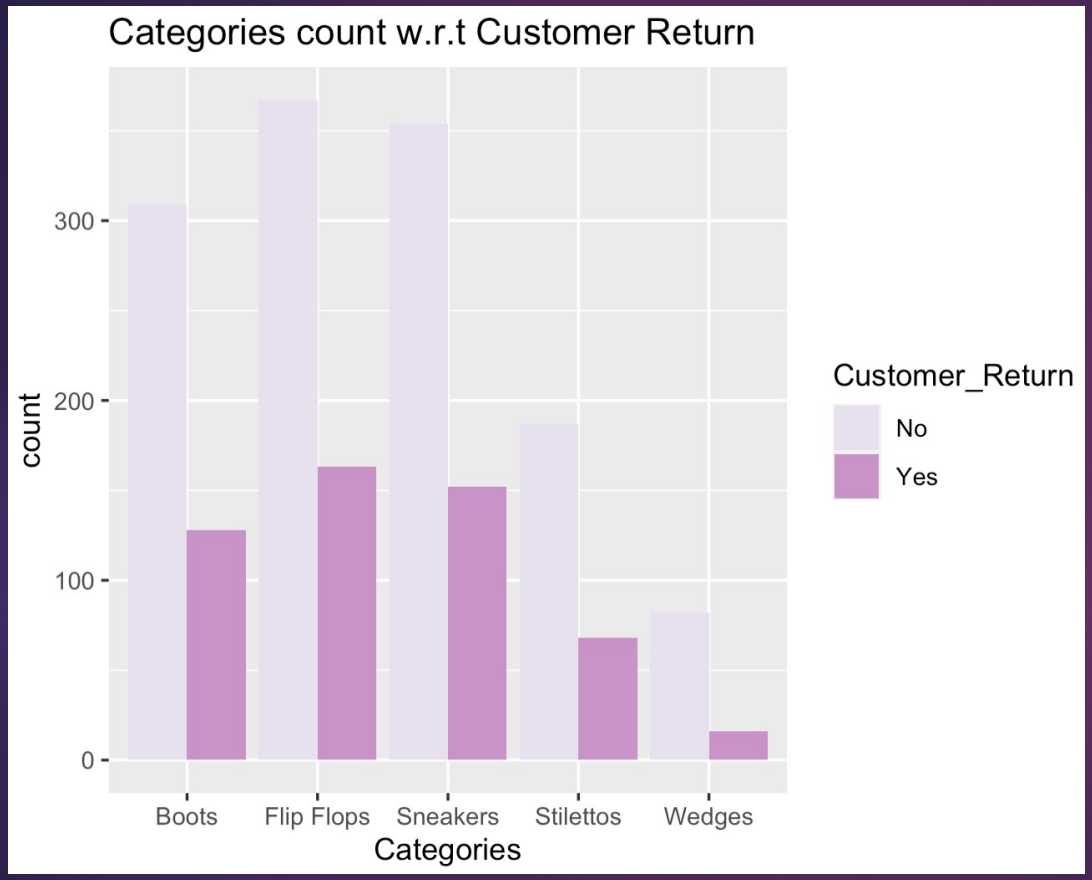
- The store has 5 footwear categories.
- Sneakers seems to be doing a good business as it is the category with the highest sales followed by Stilettos and Flip flops.
- Boots category seem to contribute a little less to the sales compared to the other categories.

# Brands Vs discount & customer return ?



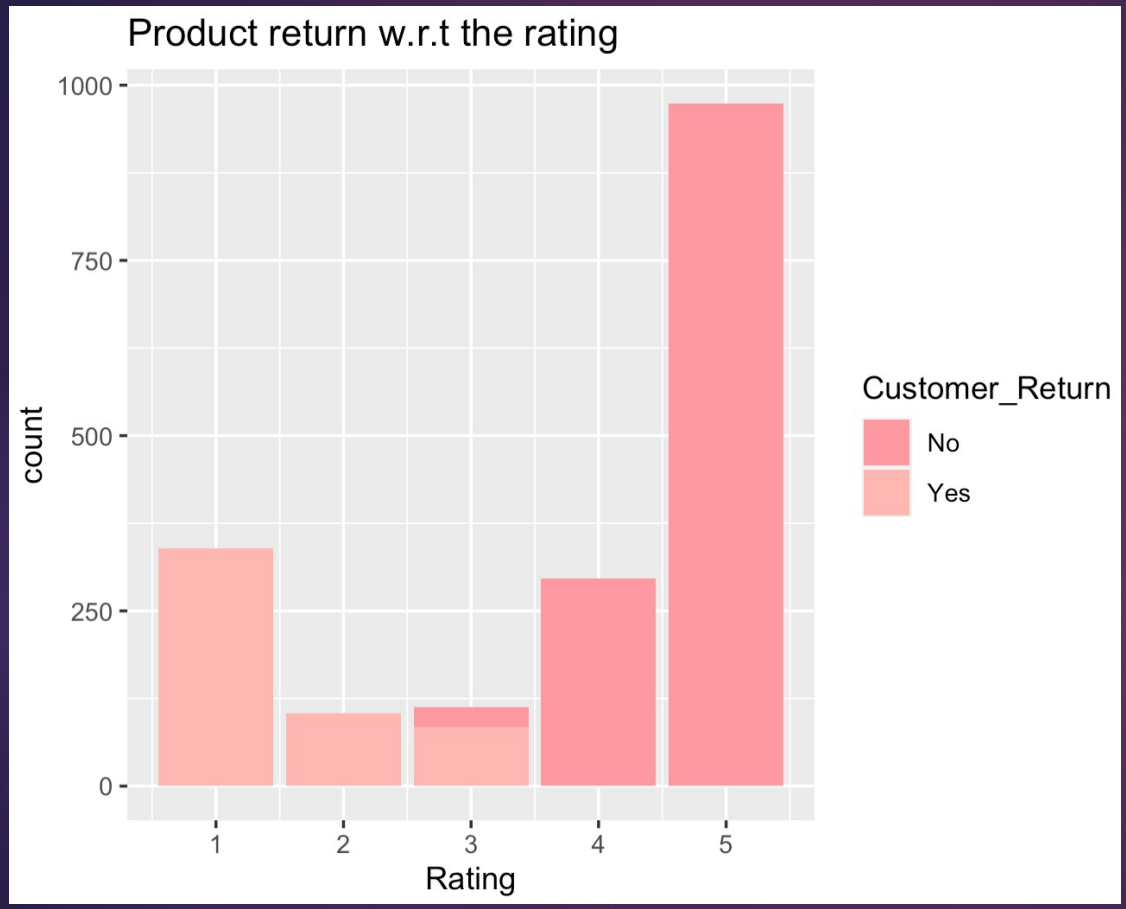




- Highest number of discounts is offered by Tommy Hilfiger followed by Kate Spade.
- But Tommy Hilfiger also has the highest number of customer returns.




## Categories vs customer return?

- Flip flops have the highest number of returns followed by sneakers.
- Very few customers return wedges and stilettos.




# Distribution of rating according to customer return.

- Majority of the Customers who gave 3 as the rating returned the footwear.

# PREPROCESSING

**Checking Null and Missing Values:**
The dataset has no null and missing values.

**Dropping columns:**
To leverage data for better efficiency and accuracy, we dropped serial numbers and Calendar dates for model implementation(Naïve Bayes, Decision Tree, and Random Forest).

**Label Encoder:**
To transform non-numeric values in data such as Categories, brand, color, discount to numeric.

**Feature Selection:**
To identify the correlation between the columns.

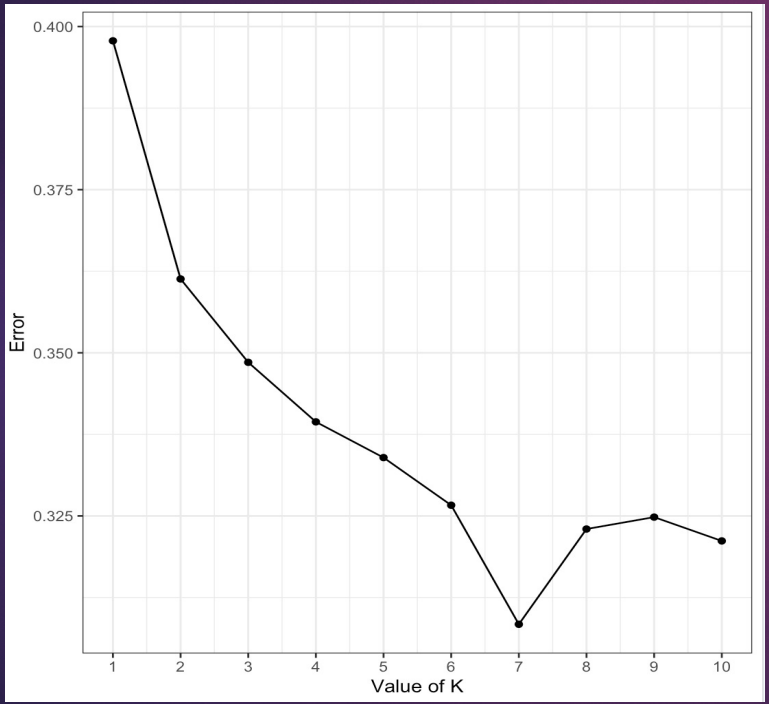
**Categorizing:**
The rating column has values from 1 to 5. So, we categorized it to 0 and 1 where1,2,3 = 0 and 4,5 =1.

# Models Implemented

- **KNN:** It captures information of all training cases and classifies new cases based on a similarity.
- **Naive Bayes Model:** Performs well in cases of categorical target variable compared to numerical variable.
- **Decision Tree:** The goal is to create a training model that can use to predict the target variable by simple decision rules inferred from prior data.
- **Random Forest Classifier:** Provides higher accuracy through cross validation. If there are more trees, it won't allow over-fitting trees in the model.
- **Time-series Forecasting:** To examine how the changes are associated with the sales variable over time. And the primary focus is to uncover patterns in data to predict future sales of the store.

# K Nearest Neighbor(KNN)

Goal: To predict customer rating of the different footwears. Target variable is Rating_categorical.




```
#KNN implementation
predicted.Rating <- knn(train[1:10],
                        test[1:10],
                        train$Rating_categorical,
                        k=1)
```

- The error rate is the lowest when the value of k is 7 and it is the highest when k is 1.
- That means, the model gives approximately 90% Accuracy for k = 7.

# Naive Bayes Model

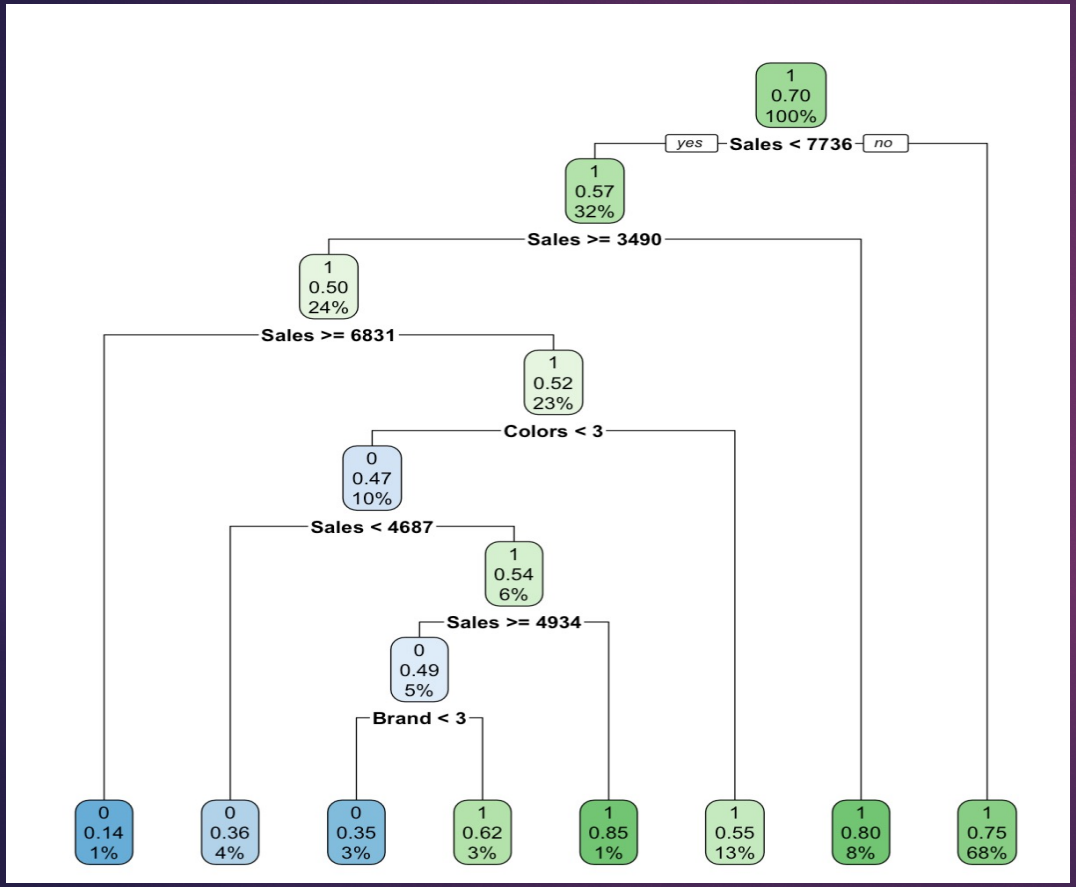
The aim is to predict the customer rating of each product (i.e. if the rating is good or bad).

```
#implementing naivebayes model
NaiveBayes = naiveBayes(Rating_categorical ~ . ,data = training_set)
NaiveBayes

# Predicting the Test set results
prediction <-predict(NaiveBayes, test_set)

#Model evaluation using Confusion Matrix
ConfusionMatrix = table(test_set[,9], prediction)
ConfusionMatrix
```

```
> Accuracy
[1] 0.6958425
```

The accuracy of approximately 69.5%.

# Decision Tree



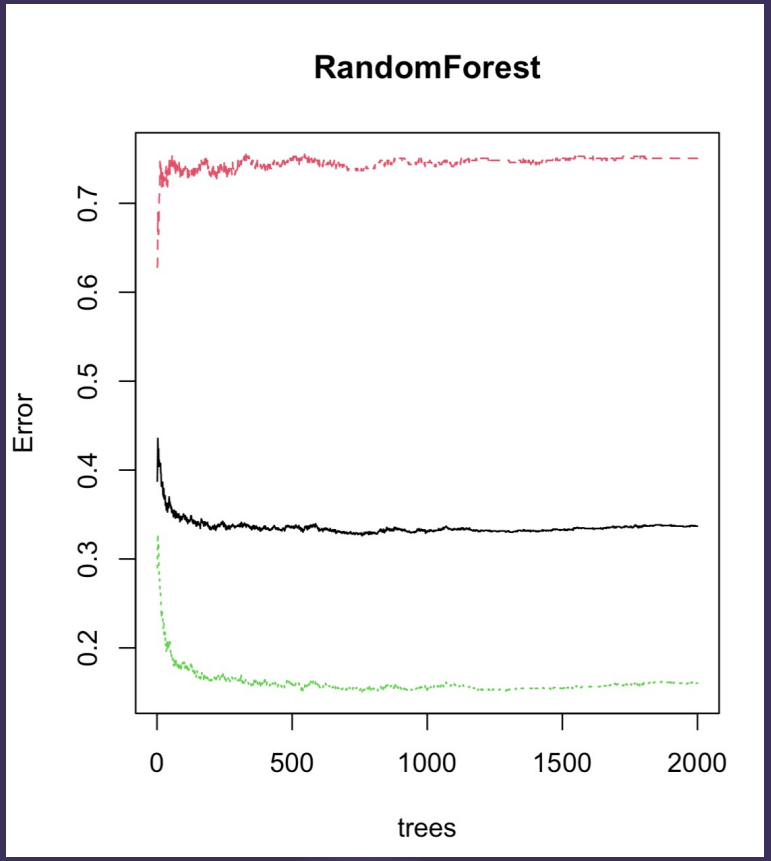

- A decision tree is constructed by recursive partitioning — starting from the root node , each node can be split into left and right child nodes.
- As we can see most of the decisions are made using sales features.
- The model indicates 93% chances that customers are going to give reviews as good whereas more than 7% shows the rating will be bad.
- The accuracy of decision tree is around 70%.

# Random Forest

- Random forest combines hundreds or thousands of decision trees, trains each one on a slightly different set of the observations, splitting nodes in each tree considering a limited number of the features.

- Upon plotting the model, we can see that the **Out-of-Bag error** rate is consistent after about 150 trees and thus we fix the number of trees to 150 instead of plotting 2000 trees.

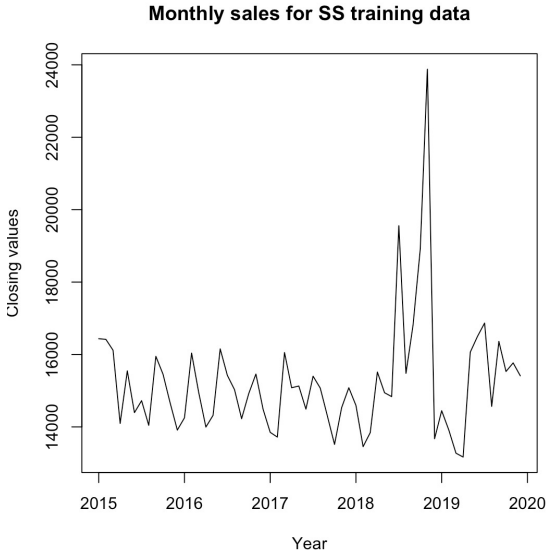
- The Random Forest model gives an accuracy of 82%.

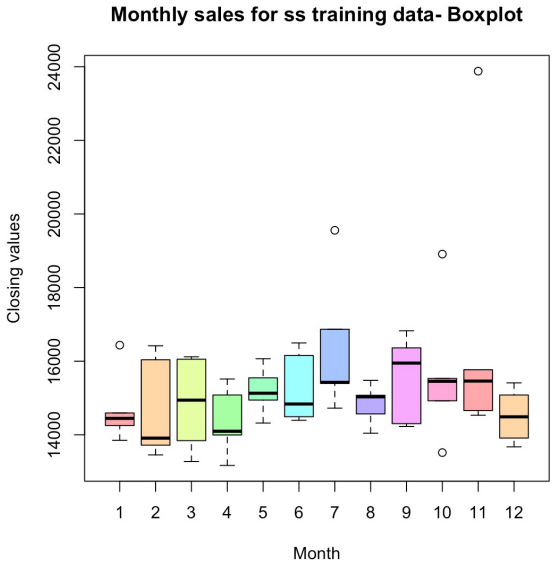


# Comparison of model accuracy

| ML MODELS IMPLEMENTED | ACCURACY |
|---|---|
| K-nearest Neighbors | 60% (for k=1)<br>90% (for optimum k = 7) |
| Naïve Bayes | 69.58% |
| Decision Trees | 70% |
| Random Forest | 82% |

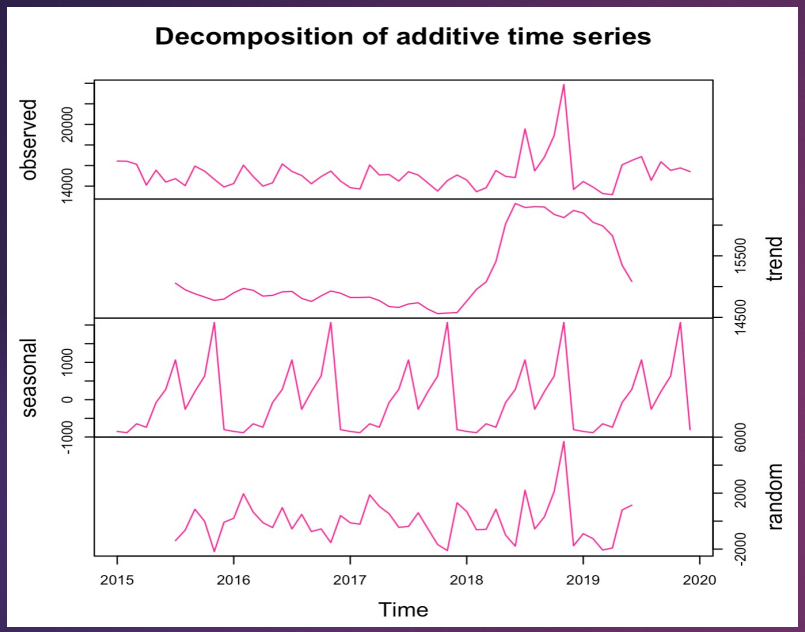
# Time Series Forecasting

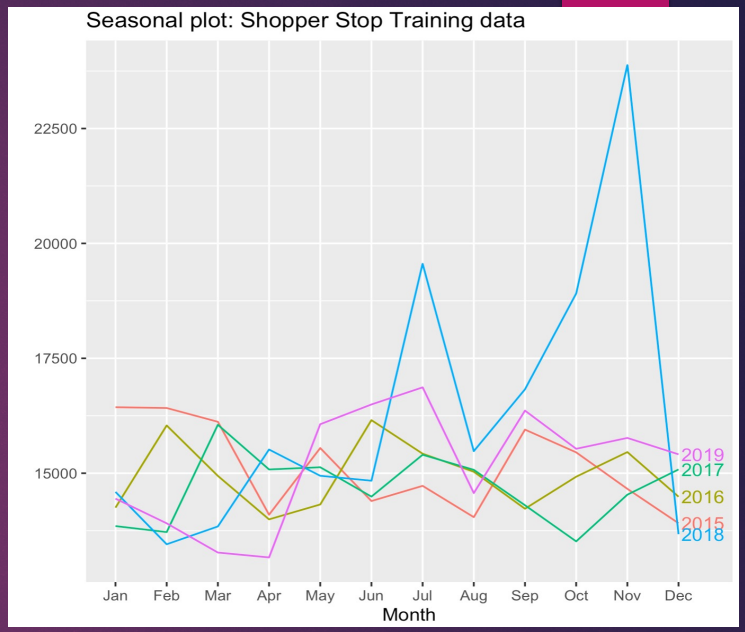



- Time series provide predictions about future trends based on historical data.
- Monthly Analysis:
  - From the graph, the sales are mostly flat across years except for mid 2018 to end 2018, where there is an unusual spike in sales figures.
- Seasonal Trends:
  - The boxplot shows the distribution of sales for each month over the five years period. For each bar, boundaries highlight the quantile regions.
  - Average sales figures are more for September. And over the years , February and March, seem to have a widespread in the range of sale figures when compared to other months.

# Decomposition




- Data decomposition provides insight into data and can help comprehend the data better. From the decomposed data we observe, trend, seasonal and random components.
- The Plot is the overlay of sales data per month over five years.

# MODEL IMPLEMENTATION (ARIMA)

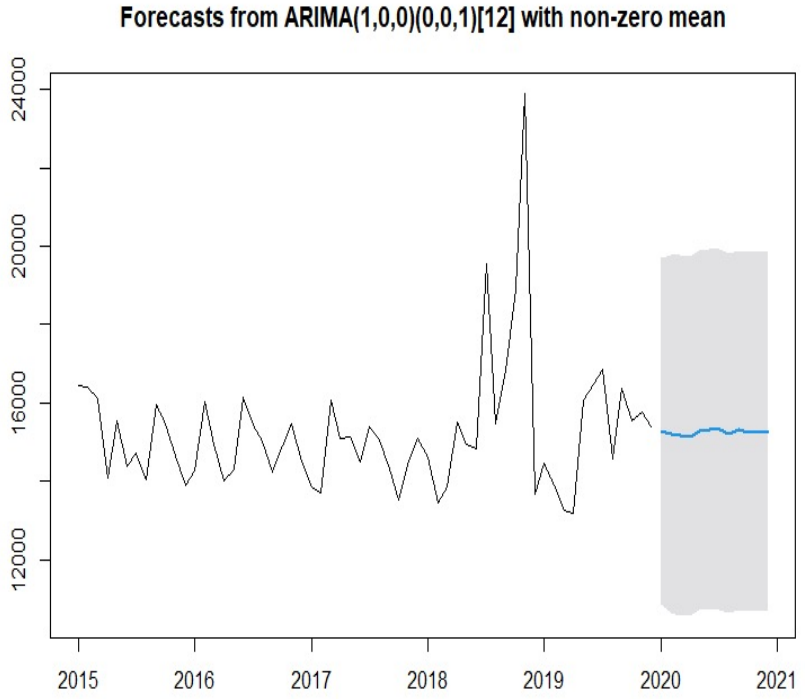
- To satisfy the time series assumptions, we need to remove the non-stationary from the data and check the Seasonality.

```
#Considering the order(1,0,0) as the best model
fitARIMA <- arima(SSdata_train, order=c(1,0,0),
                  seasonal = list(order = c(0,0,1), period = 12),method="ML")
fitARIMA
```

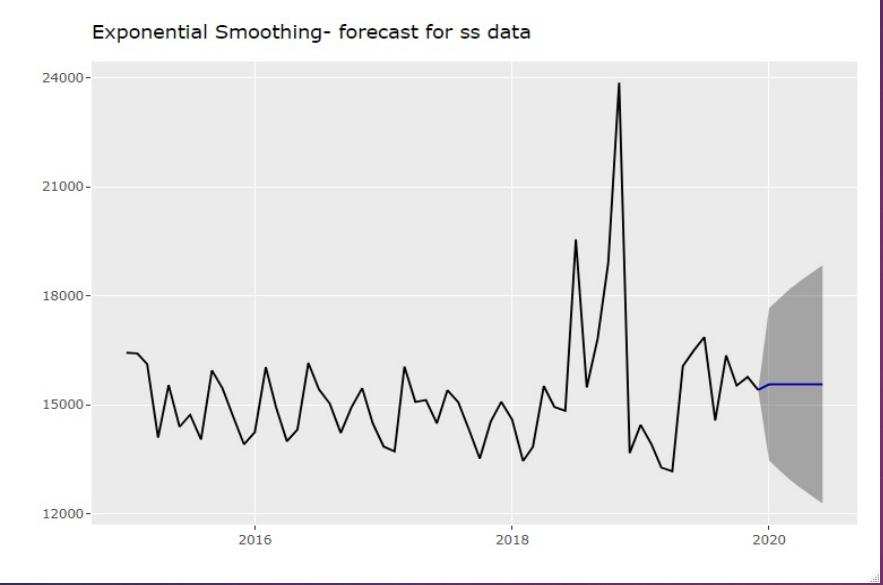
The model gives:
Log Likelihood = -527 and AIC = 1062

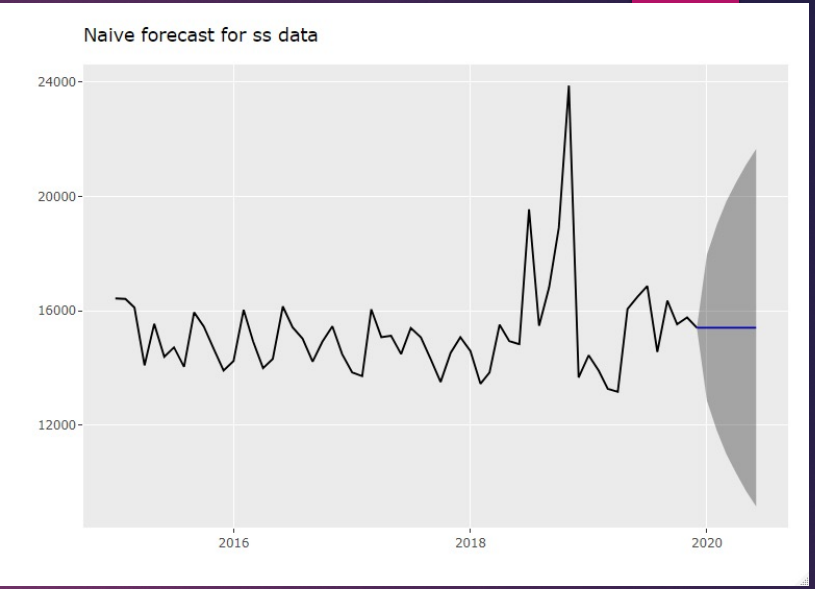
RMSE= 1577.423

```
> #Predicting the next 5 months sales
> predict(fitARIMA,n.ahead = 5)
$pred
          Jan      Feb      Mar      Apr      May
2020 15276.59 15205.15 15159.60 15146.41 15293.30
```



# Comparison with other Models:




RMSE = 1746.319

RMSE = 1991.616

In conclusion, on comparing RMSE, AIC, BIC values, Arima model performs better than other models.

How can the store benefit from implementing the above machine learning models?

- PRICE OPTIMIZATION
- DEMAND PREDICTION
- LOGISTICS SUPPORT
- PERSONALIZED OFFERS
- TREND ANALYSIS
- CHURN RATE PREDICTION
- BOOST SALES
- BETTER INVENTORY MANAGEMENT

# Conclusion

- In all performance metrics, Random Forest model has the highest accuracy in predicting the customer rating for the various footwears.
- Time series can successfully predict the future sales of the footwear store.
- The store witnessed a hike in sales between July and November,2018.
- The store has maximum high ratings for majority of their products, but they need to work on the Customer return aspect.
- The footwears that are on discount are not making much difference to the store's sales. So the store management needs to change their discount pattern.

# Links & References


1. https://www.rdocumentation.org/packages/e1071/versions/1.7-3/topics/naiveBayes
2. https://towardsdatascience.com/random-forest-in-r-f66adf80ec9
3. https://medium.com/@Synced/how-random-forest-algorithm-works-in-machine-learning-3c0fe15b6674
4. https://datascienceplus.com/time-series-analysis-using-arima-model-in-r
5. http://rstudio-pubs-static.s3.amazonaws.com/318411_18399592759841f2a151e445adb851c7.html

# Thank You

BE HOME, STAY SAFE.