Capstone Project

Retail Sales Prediction

By

Baishanvee Mahato

Data Science Trainee- Almabetter, Bengaluru

Contents

- Problem statement
- First Look At The Data
- Exploratory Data Analysis
- Outliers Treatment
- Inferences from EDA
- Feature Engineering
- Building a Regression Model
- Conclusion

Problem Statement

Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality. With thousands of individual managers predicting sales based on their unique circumstances, the accuracy of results can be quite varied.

I have been provided with historical sales data for 1,115 Rossmann stores. The task is to forecast the "Sales" column for the test set. Note that some stores in the dataset were temporarily closed for refurbishment.

First Look At The Data

Details Of Dataset Provided:

Rossmann Stores Data.csv - historical data including Sales

store.csv - supplemental information about the stores

Data fields

- **Id** - an Id that represents a (Store, Date) duple within the test set
- **Store** - a unique Id for each store
- **Sales** - the turnover for any given day (this is what you are predicting)
- **Customers** - the number of customers on a given day
- **Open** - an indicator for whether the store was open: 0 = closed, 1 = open
- **StateHoliday** - indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None

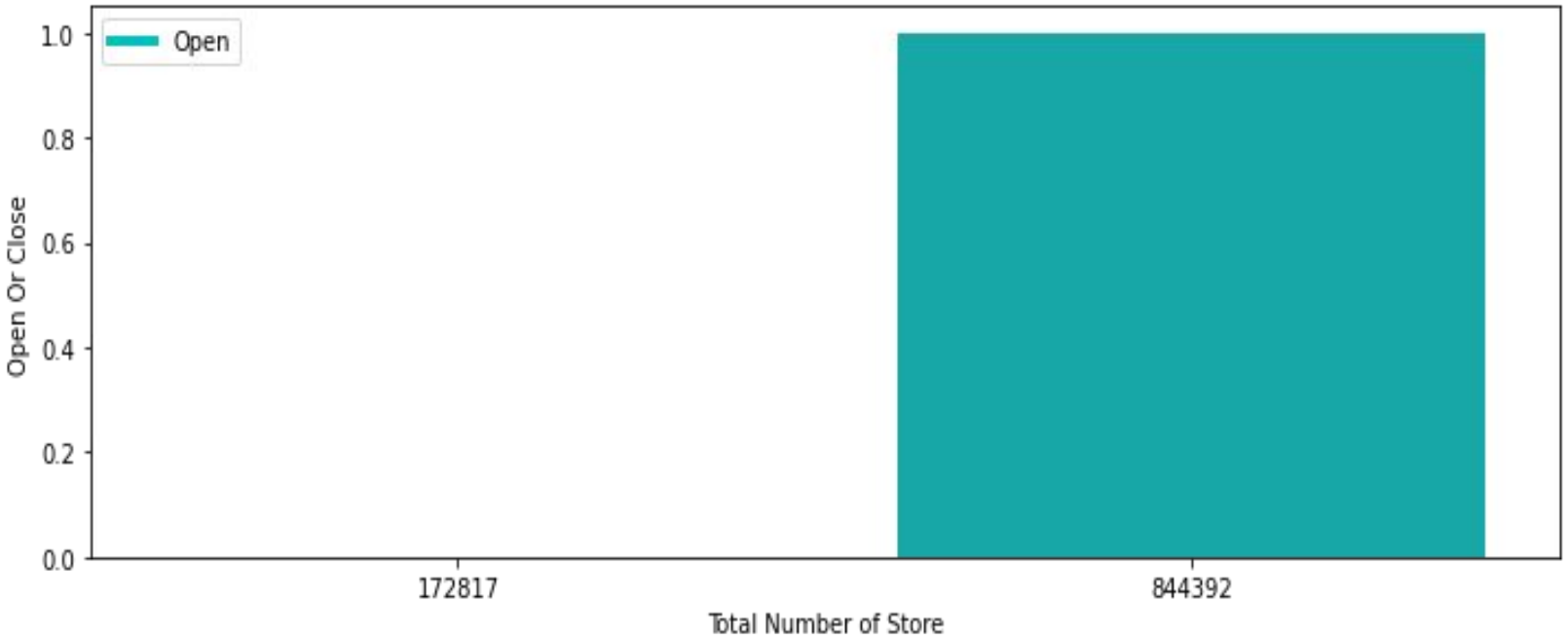
- **SchoolHoliday** - indicates if the (Store, Date) was affected by the closure of public schools
- **StoreType** - differentiates between 4 different store models: a, b, c, d
- **Assortment** - describes an assortment level: a = basic, b = extra, c = extended
- **CompetitionDistance** - distance in meters to the nearest competitor store
- **CompetitionOpenSince[Month/Year]** - gives the approximate year and month of the time the nearest competitor was opened
- **Promo** - indicates whether a store is running a promo on that day
- **Promo2** - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
- **Promo2Since[Year/Week]** - describes the year and calendar week when the store started participating in Promo2
- **PromoInterval** - describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew. E.g. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store

EDA

Exploratory Data Analysis

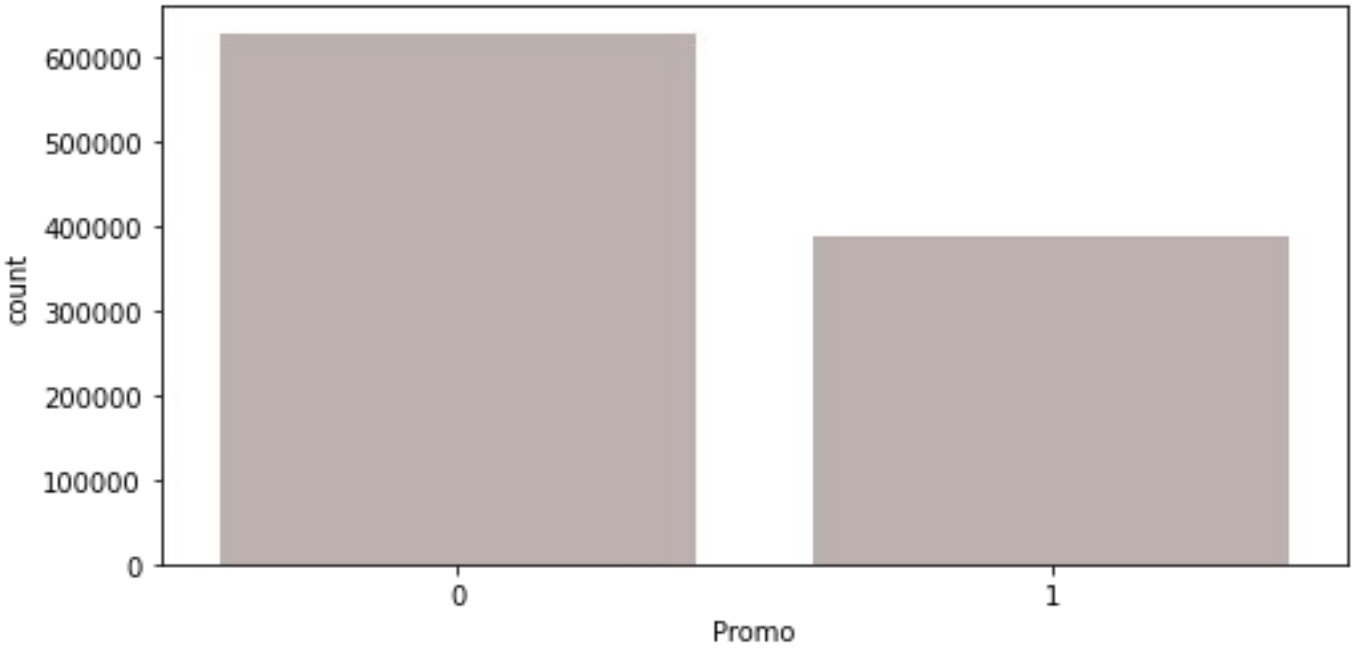
The Given dataset is from 1 January 2013 to 31 July 2015 , i.e. approx. 2.5 yrs.

How many open or closed stores are in the data?



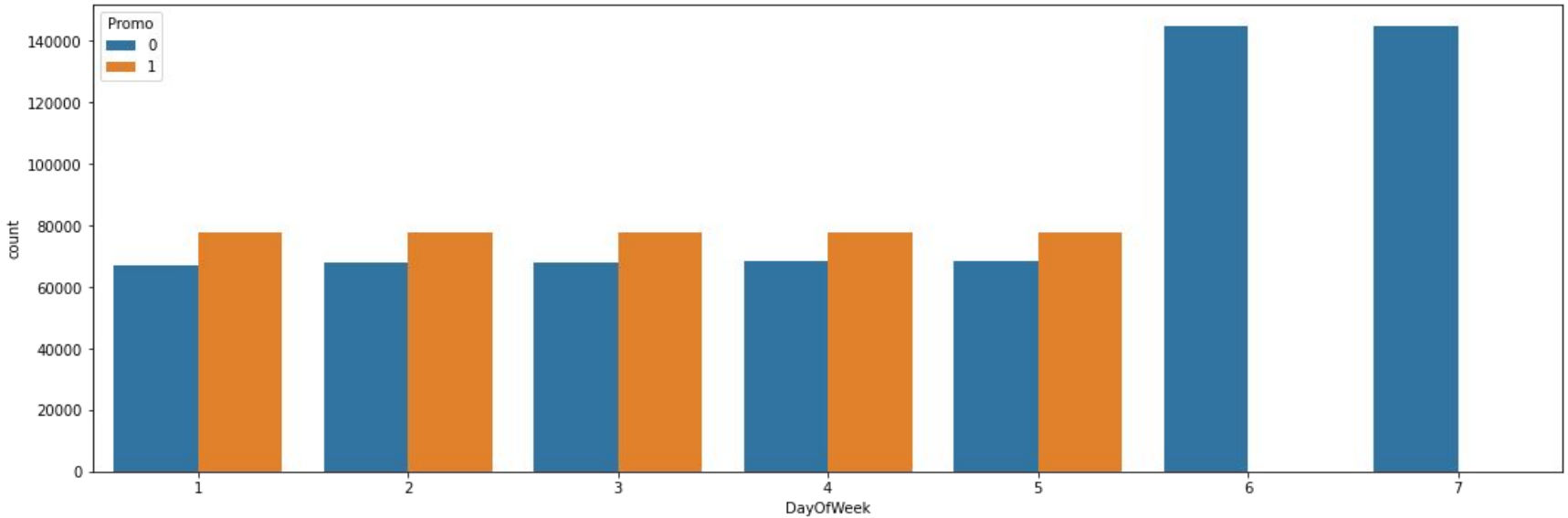
There are 172817 closed stores and there are 844392 open stores.

How many Promo is in the data?



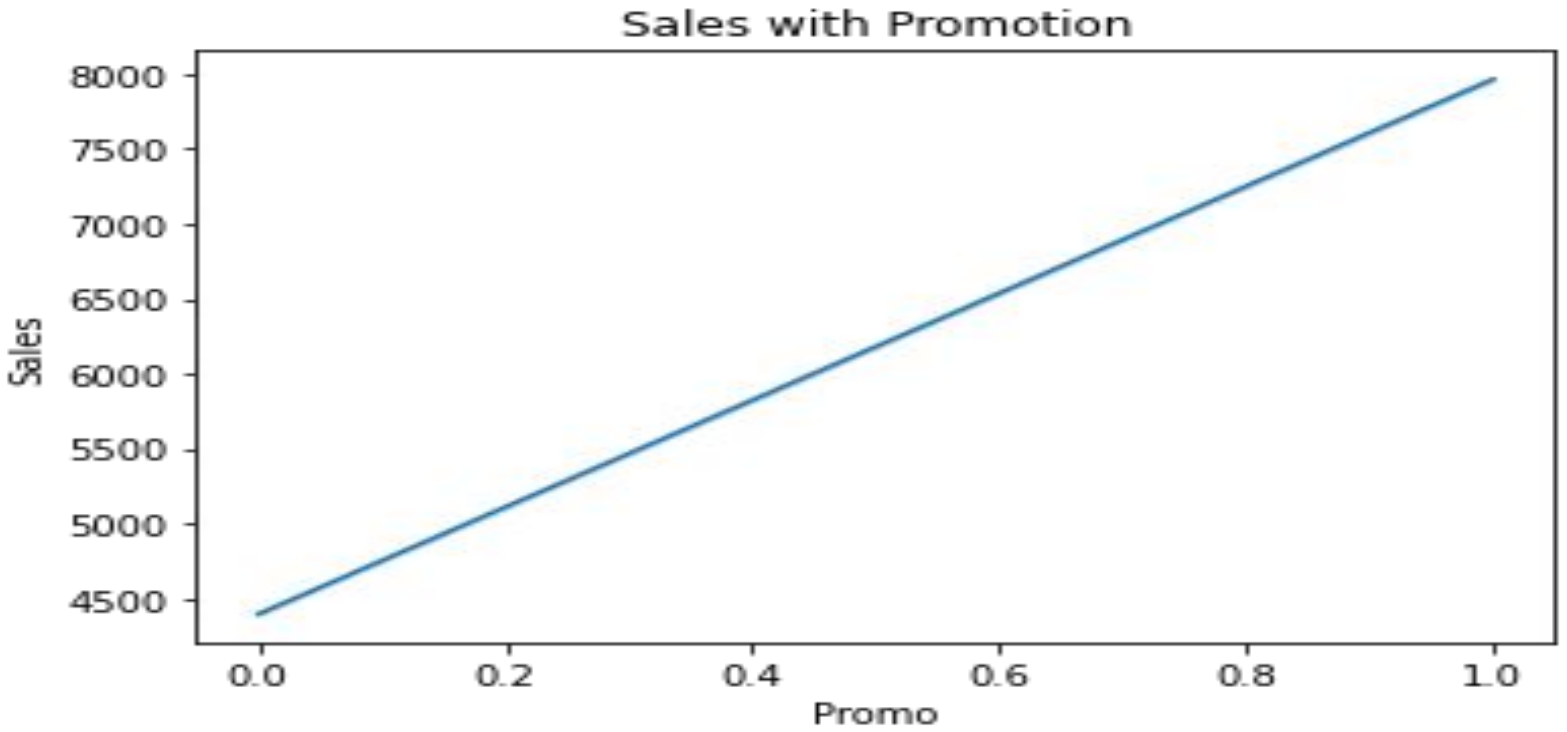
There are 629129 no promotions, and there are 388080 promotions in the data.

Which day has the Promotion?



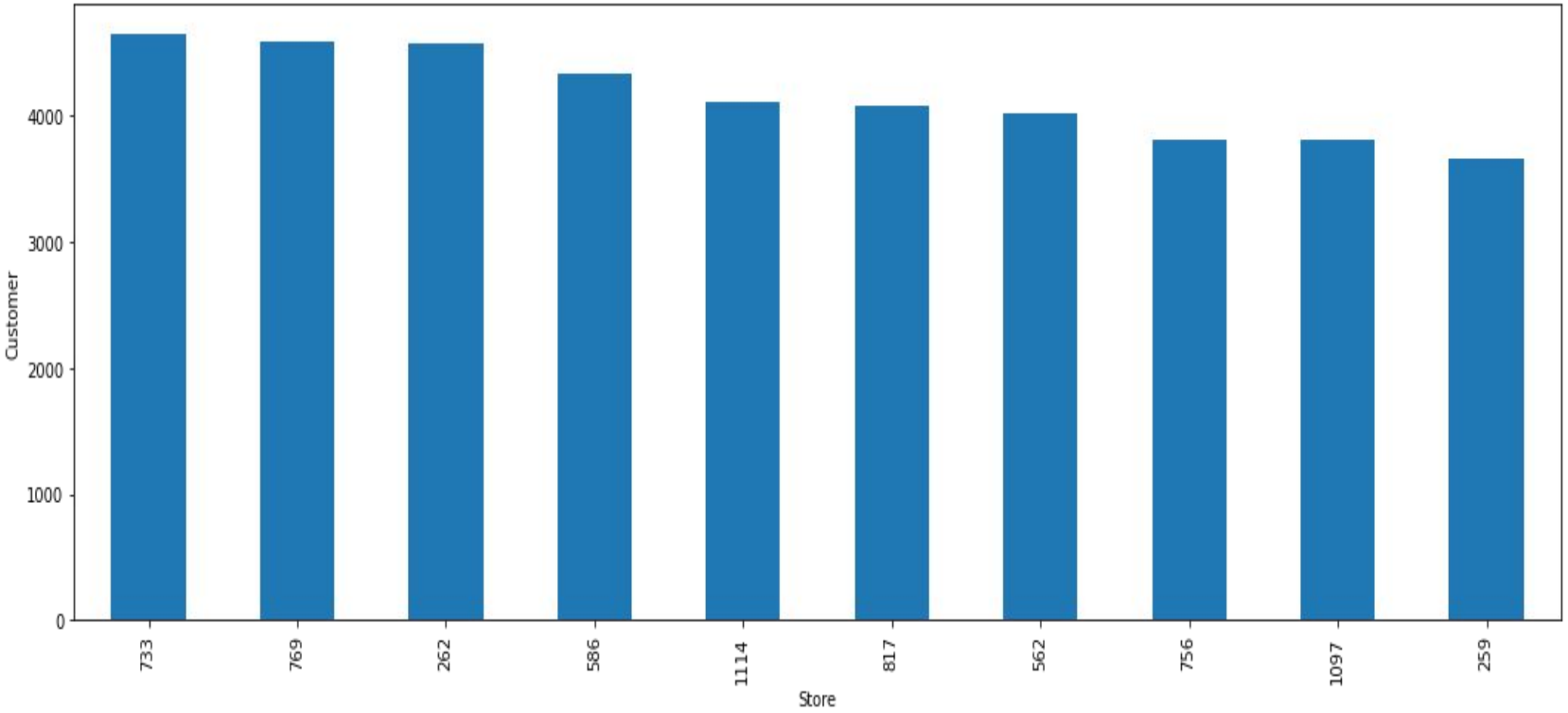
All of the weekday has a promotion. The weekend has no promotion.

Does Promotion affect the Sales Price?



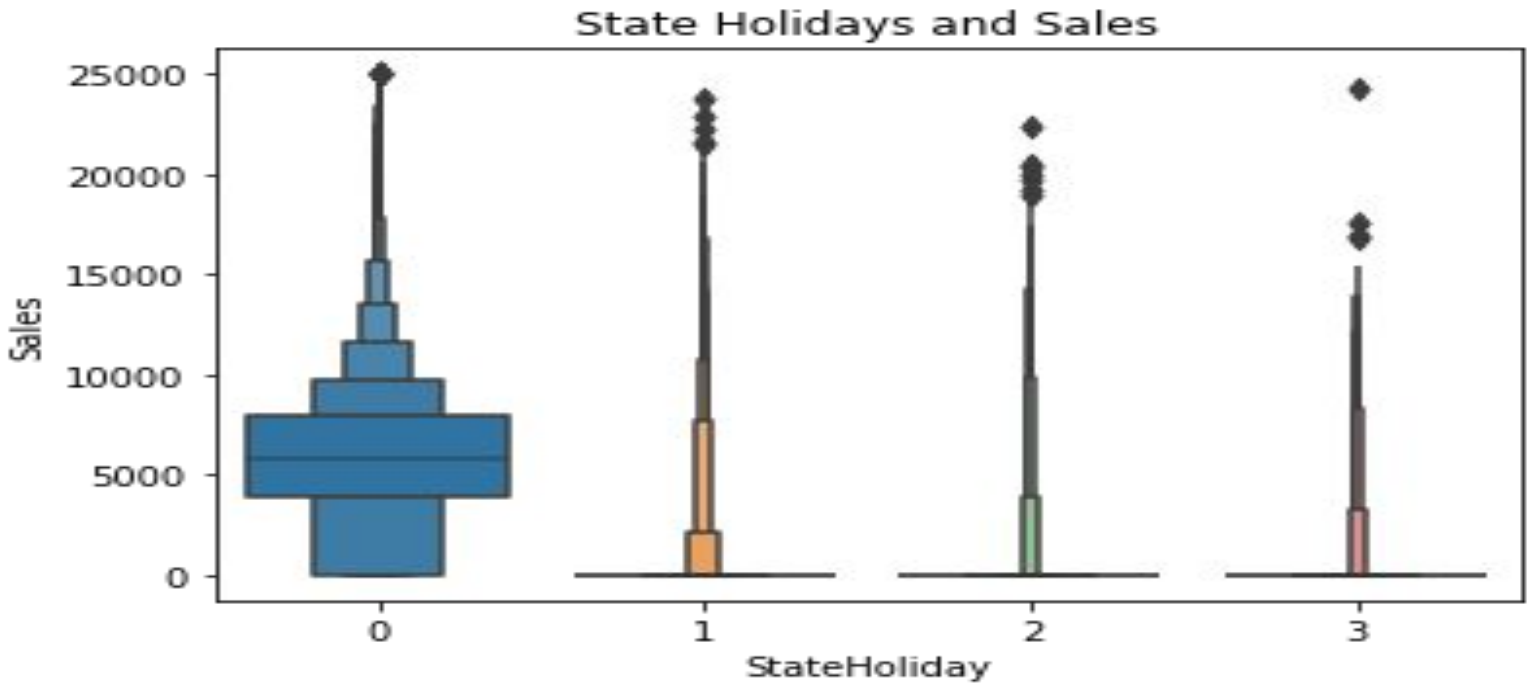
The sales price has the highest when there is a promotion.

Which store has the more Customers?



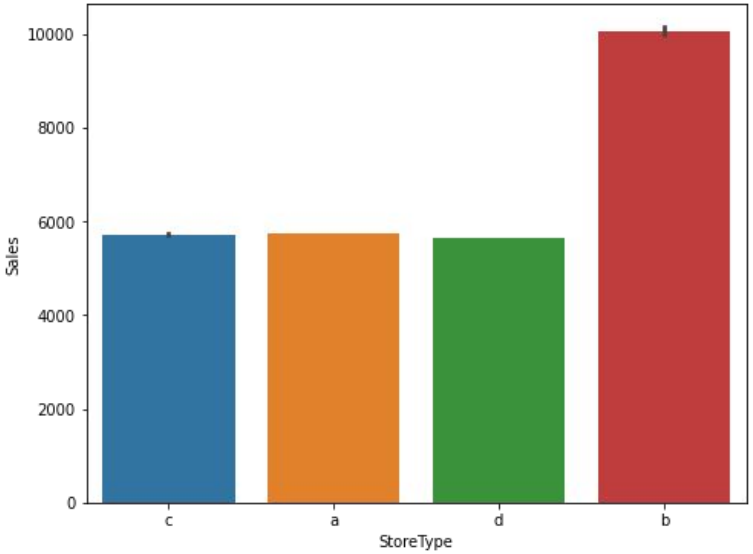
In the graph, the store which is 733 number has the maximum number of customers.

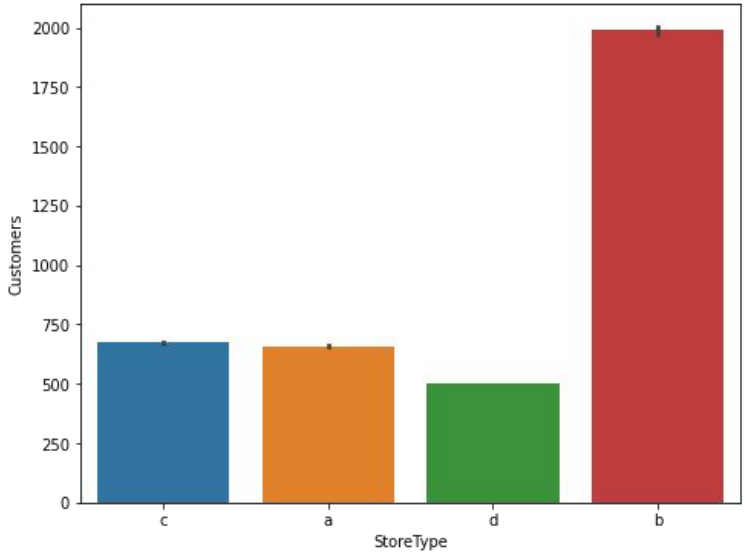
State Holidays and Sales



There are no sales during the public holiday period. Therefore, important holidays affect sales.

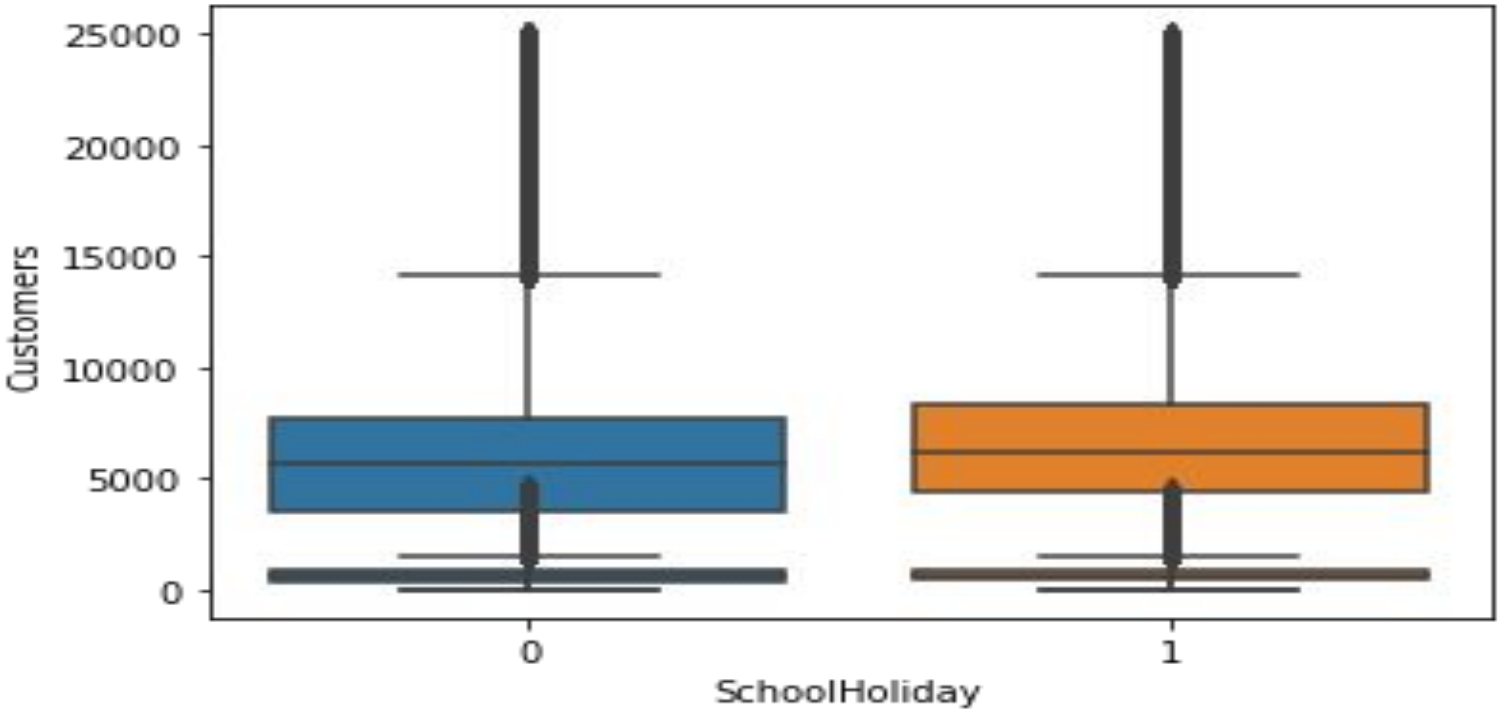
Sales Vs State Holidays





In the left figure, the b store has the highest sales. a, c, d stores have the same sales prices approximately. When we look at the right figure, b has the highest number of customers with a clear difference. In this figure, only the d store has the lowest number of customers but it has the same sales prices compared to a and c stores.

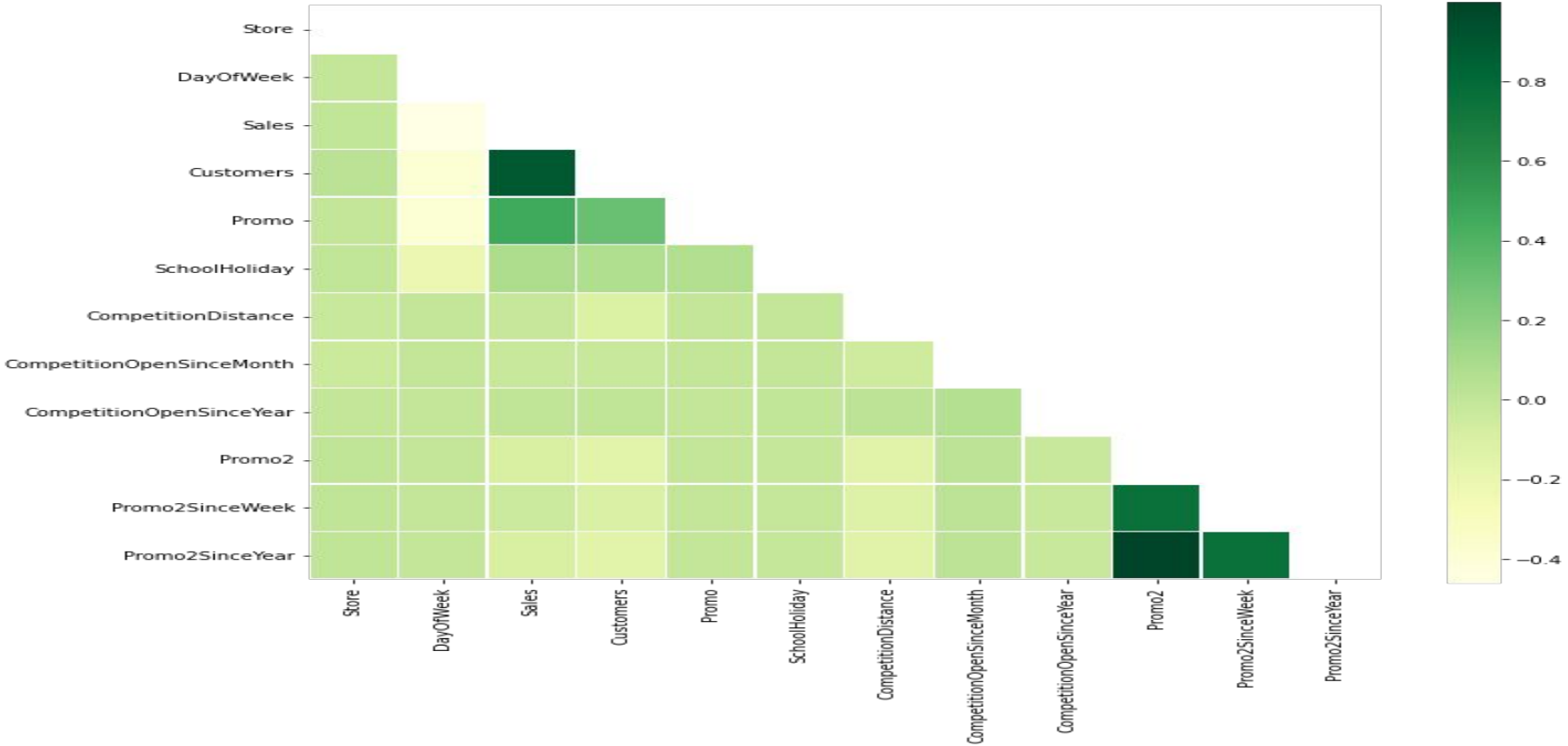
Sales Vs School Holidays



When we look at the figures, O shows the no school holidays and 1 shows the school holidays. we can clearly see that the school holiday increases the sales price and customers proportionally.

Correlation Analysis

Correlation analysis shows the relationship between 2 features. This analysis helps us to understand which features are positive or negative relationships.



In the above figure, the heat map shows the relationship between two features. Each two features values must be between -1 and 1. When we explain the map:

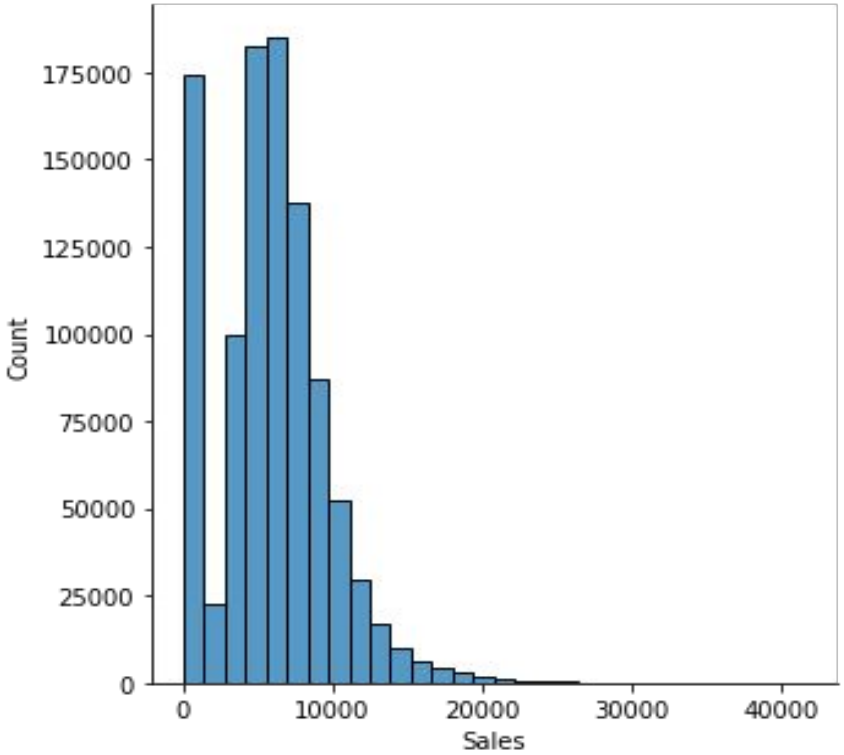
-1 means that there is a negative relationship between 2 features. For example, Customers and PromoInterval have a negative relationship. The colour of the value is light blue.

0 means that there is no relationship between 2 features.

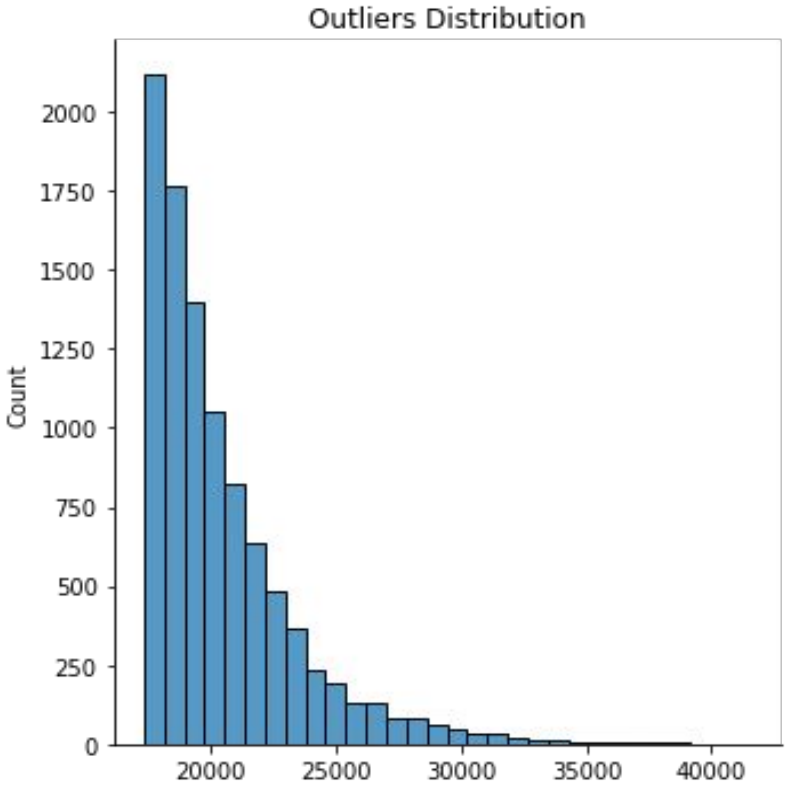
1 means that there is a positive relationship between 2 features. For example, Promo2 and Promo2SinceYear have a strong positive relationship.

Outliers Treatment

An outlier is **something separate or different from the crowd**. Outliers can be a result of a mistake during data collection or it can be just an indication of variance in your data.



From the above graph we can observe that sales >25k are very less, so it might be an outlier.



Total percentage of 0 in dataset: 16.995%
Total percentage of sales >25k in dataset: 0.075%

As we can see that we have very less percentage of sales data points that are >25k so we can drop them.

Inferences from EDA

- All of the weekday has a promotion and the weekend has no promotion.
- The sales price has the highest when there is a promotion.
- All Type ‘b’ stores have comparatively higher sales and it mostly constant with peaks appears on weekends.
- Majority of Stores remains closed on state holidays.
- School holiday increases the sales price and customers proportionally.

Feature Engineering

I have done feature engineering by:

- **Converting Categorical Variable to Numeric:**

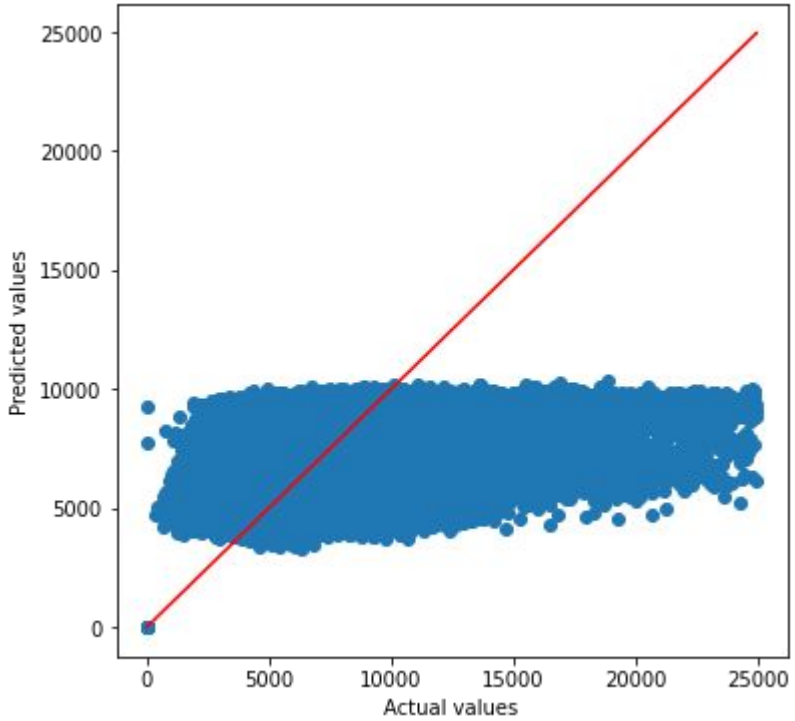
1. StateHoliday column has values 0 & "0", So, we need to change values with 0 to "0".
2. PromoInterval column has values 0 & "0", So, we need to change values with 0 to "0".
3. Encoded all categorical varibale to numeric values.

- **Looking at the scenerio where the Stores are open and yet there is no sales on that day:**

  There are total 30 days where there is no record of sales even without any holidays, So removed these data points.

Building a Regression Model

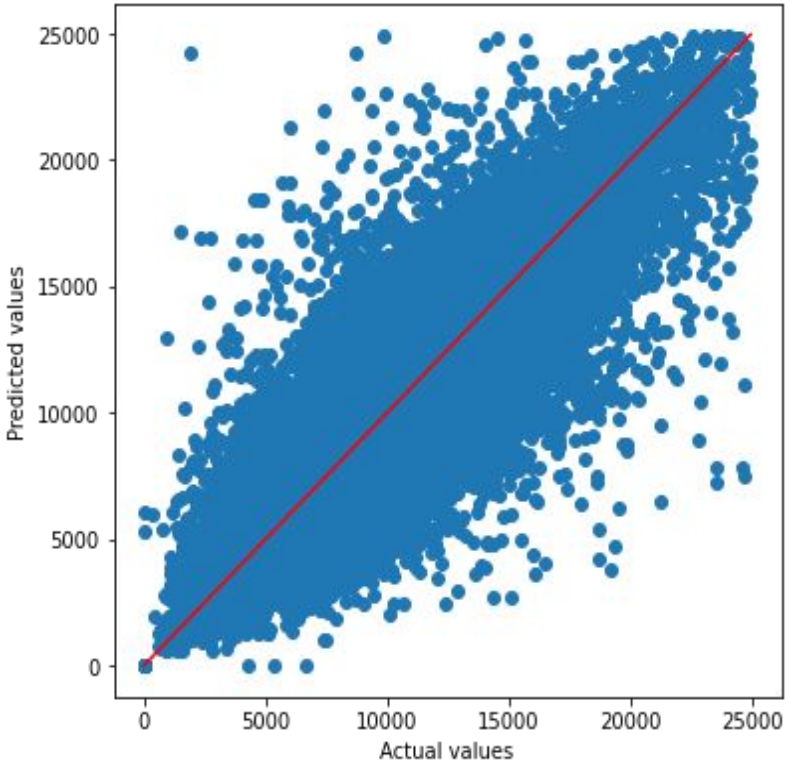
Linear Regression Algorithm:



r2_score: 0.7773037453225706
Mean absolute error: 995.40
Root mean squared error: 1923.5239720480379

From the plot we can see that Linear regression model is performing badly as its not making any predictions more than 10000 even for 25000 sales.

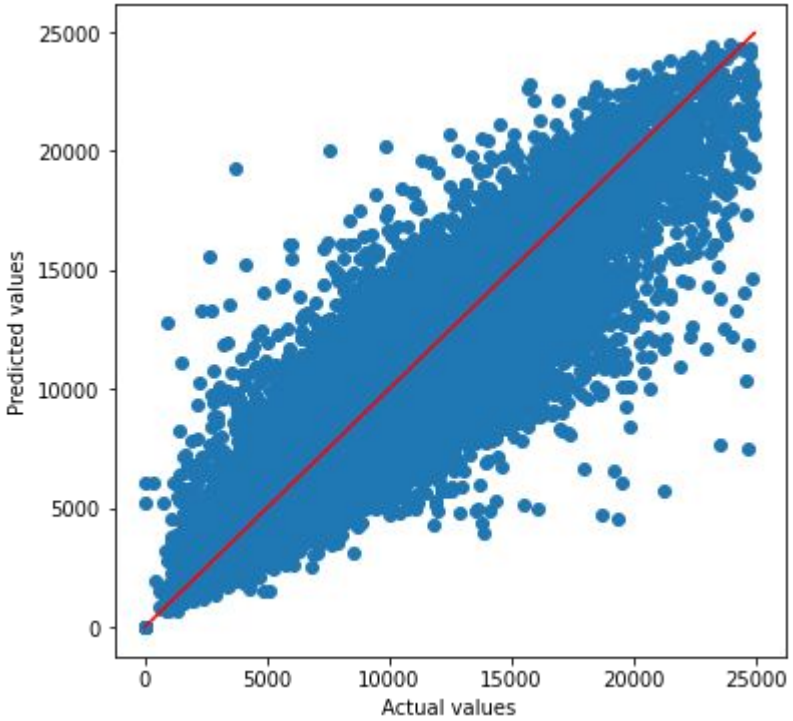
Decision Tree Regressor:



r2_score: 0.9510573626281227
Mean absolute error: 423.04
Root mean squared error: 901.7472587950255

The decision tree regressor performing well as compared to Linear Regression.

Random Forest Regressor:



r2_score: 0.9654193085804352
Mean absolute error: 357.44
Root mean squared error: 757.9804123580951

Random Forest regressor had the lowest error as compared to other models, which means it is better at predicting sales than other models.

CONCLUSION

Random Forest Regressor is the best for this Dataset Problem.