Capstone Project

Retail Sales Prediction

Team Dataloft

Gaurav Yadav

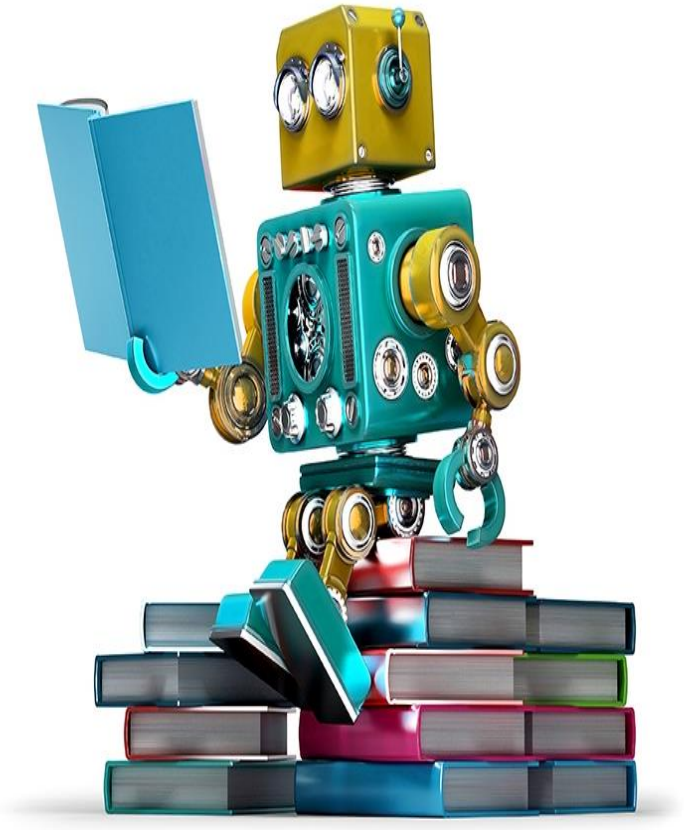
Mohan Vishe

Rahul Ray

Shambhuraj Desai

Contents

- Problem Statements
- Data Summary
- Data Description
- Data Pre-Processing
- Data Cleaning and Wrangling
- Exploratory Data Analysis
- Feature Engineering
- Linear Regression
- Lasso & Ridge Regression
- Decision Tree and Random Forest
- Feature Importance
- 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, 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.
- You are 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.

Data Summary

The data collected had two different data files:

1. Rossmann Stores Data: Historical data including Sales.
2. Store: Supplemental information about the stores.

{After merging both the datasets we have 1017209 number of records and 18 number of fields and our dataset period is from 1st Jan 2013 to 31st July 2015.}

Data Description

Most of the fields are self-explanatory. The following are descriptions of features.

- **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 we are predicting).
- **Customers** - the number of customers on a given day.
- **Open** - an indicator for whether the store was open: 0 = closed, 1 = open.

Continued..

- **State Holiday** - 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.
- **School Holiday** – indicates if the (Store, Date) was affected by the closure of public schools.
- **Store Type** – differentiates between 4 different store models: a, b, c, d.
- **Assortment** – describes an assortment level: a = basic, b = extra, c = extended.
- **Competition Distance** – distance in meters to the nearest competitor store.
- **Competition Open Since** [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.
- **Promo Interval** – 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, and November of any given year for that store.

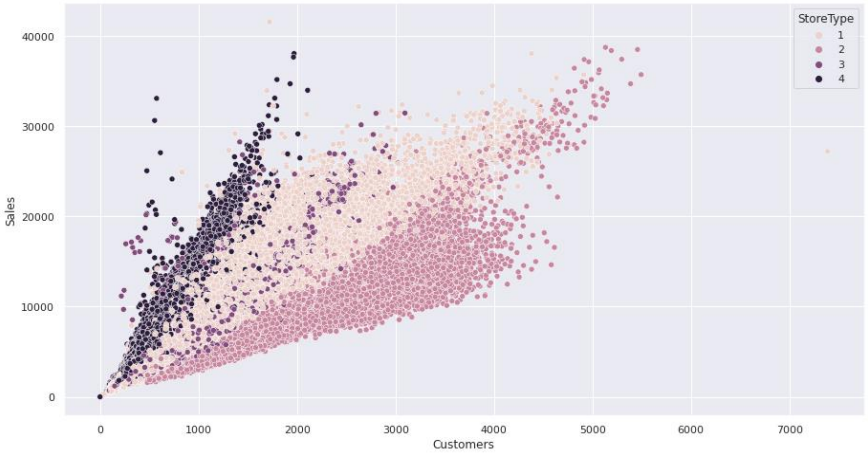
Data Preprocessing

Data Cleaning and Wrangling

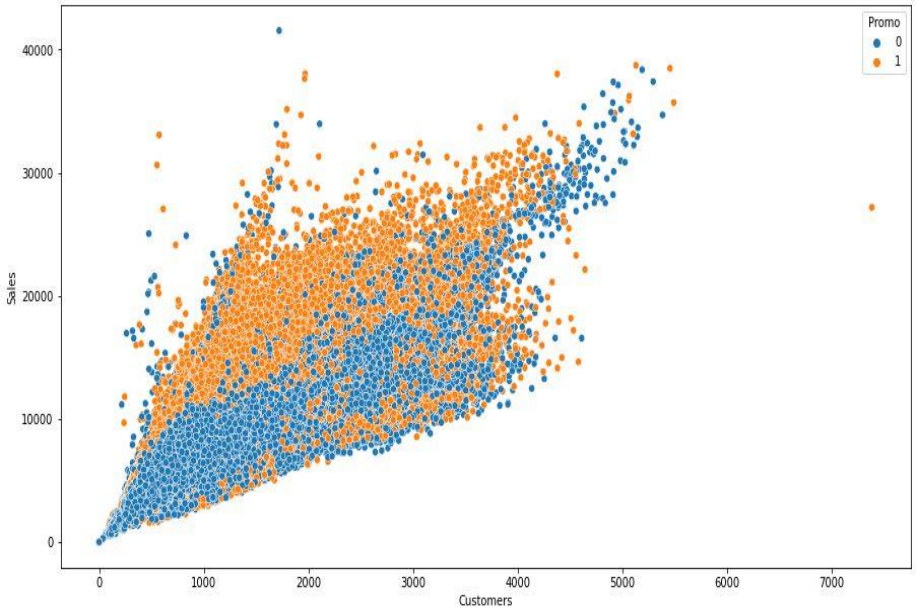
- **Missing Values:**
- Competition Distance has 3 missing values. Competition Open Since and Competition Open Since Year have 354 missing values. Promo2SinceWeek, Promo2SinceYear and PromoInterval have 544 missing values.
- We have replaced null values present in the Competition Distance column with median and the rest of 5 columns with Zero using fillna.
- **Merge both Datasets**: We have merged both the available dataset.
- **Changing Dtypes:** We have five object features namely Date, State Holiday, Assortment, Promo Interval. Using astype, we changed them to integer.
- **Encoding:** We Encoded State holiday using get_dummies.
- **Data Extraction:** We have extracted Date, Year, Month, Day, Week, Week of Year from Date column for further analysis and then dropped the Date column.

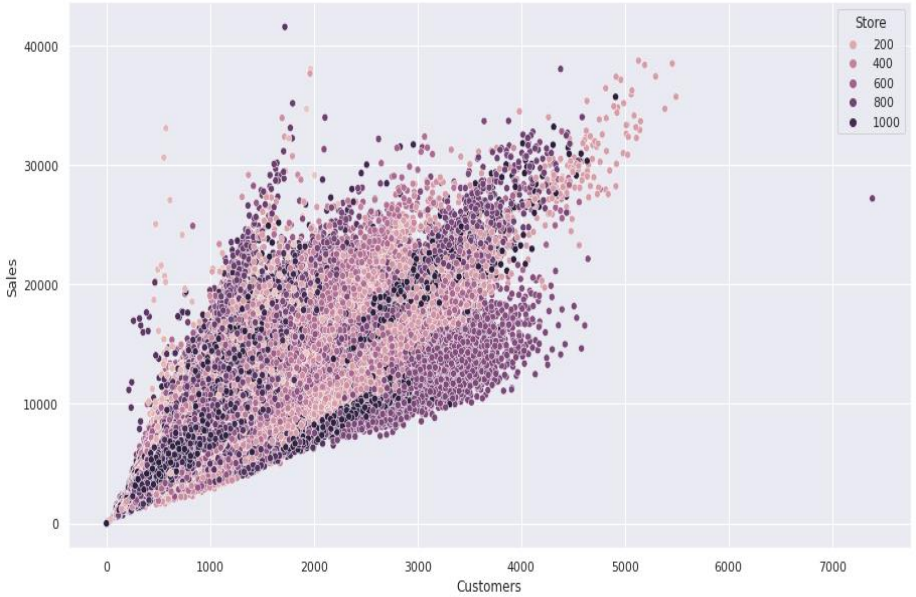
Customer vs Sales

- The scatter plot of customer and sales, here we can see relationship between the store types and assortment of sales.



Impact of Promo of Sales and Customers

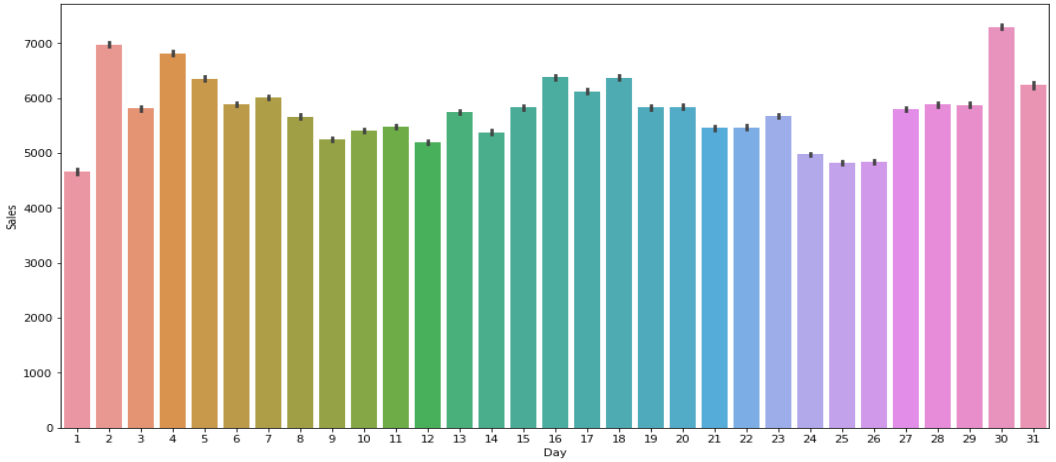


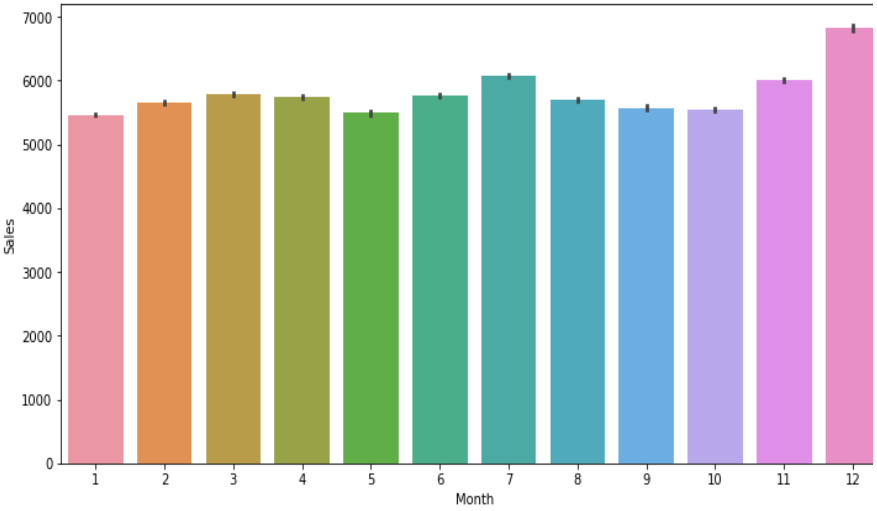


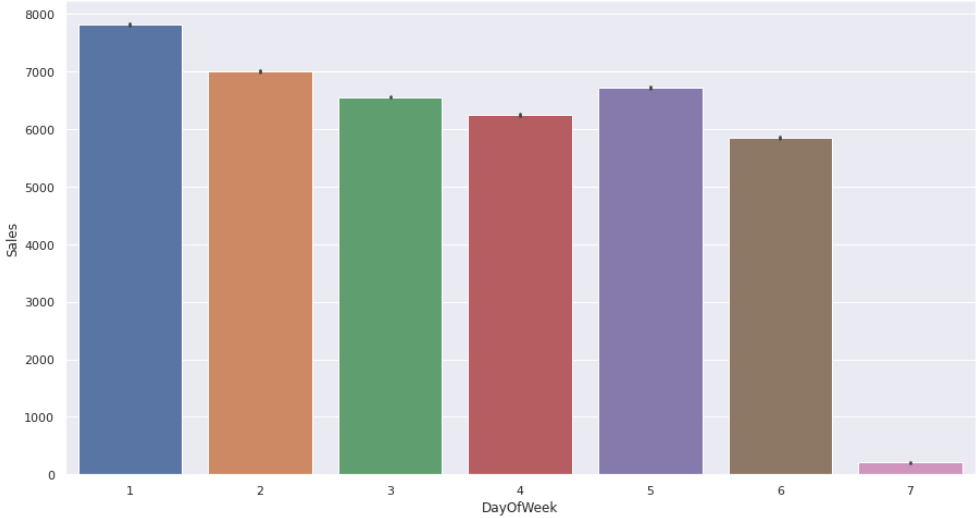
There is a linear relationship between sales and customers and whenever promo was there, sales and customers are higher compare to it was not which means promo has a good impact on sales.

Sales Analysis

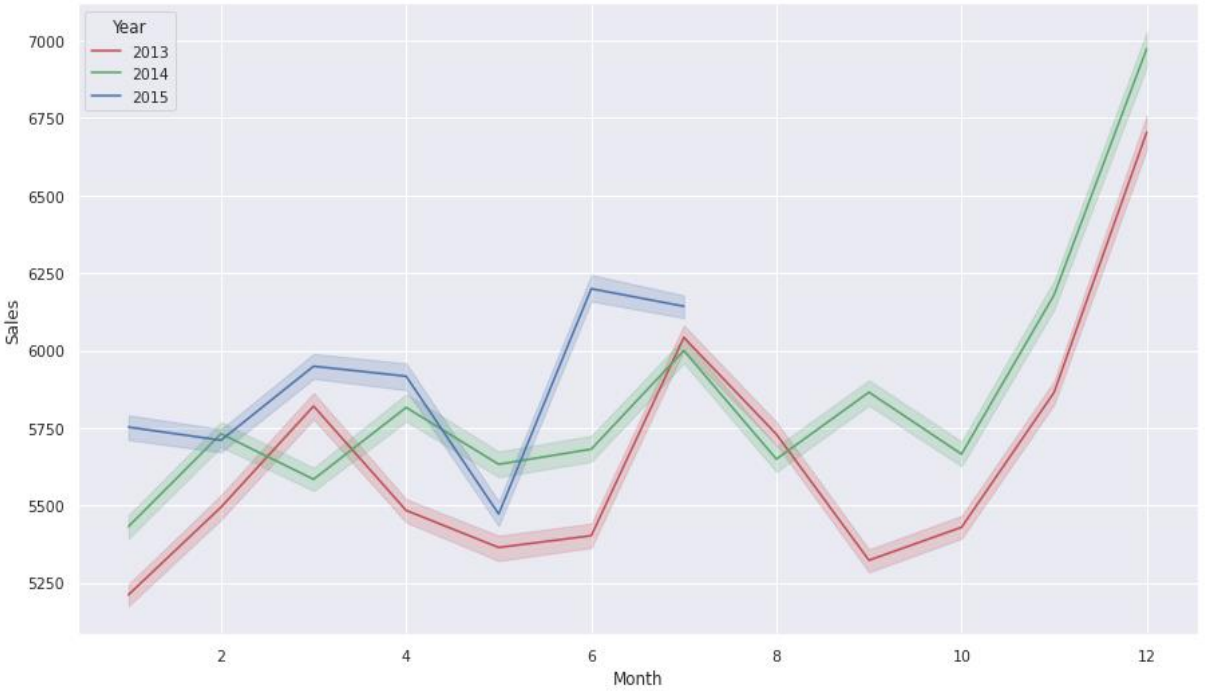
- The sales in December month is the highest among other months.
- From these plot of Day of Week, Sales starts declining from Tuesday to Friday and almost zero on Sunday.

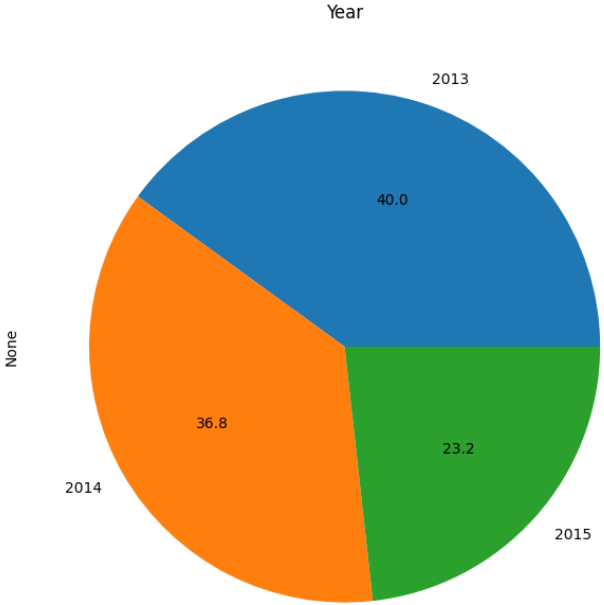






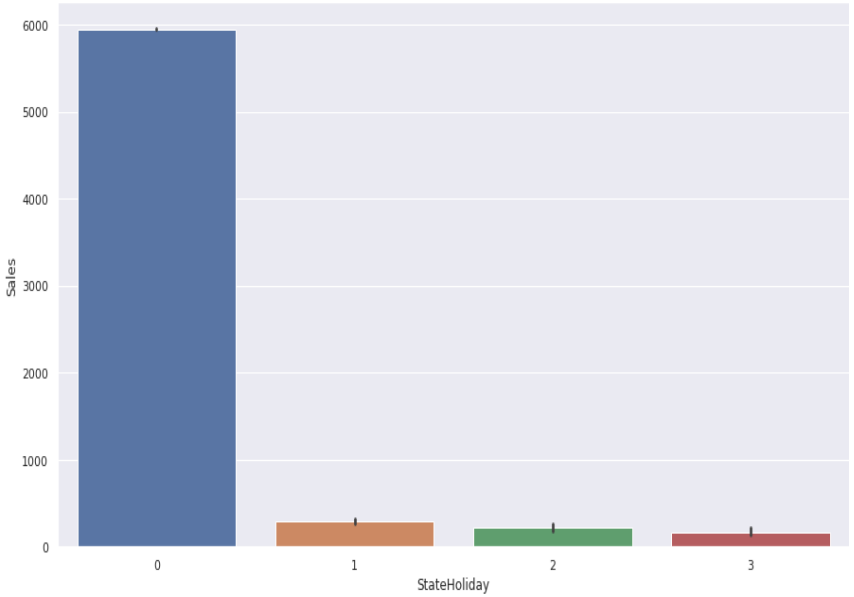
Yearly Sales Analysis

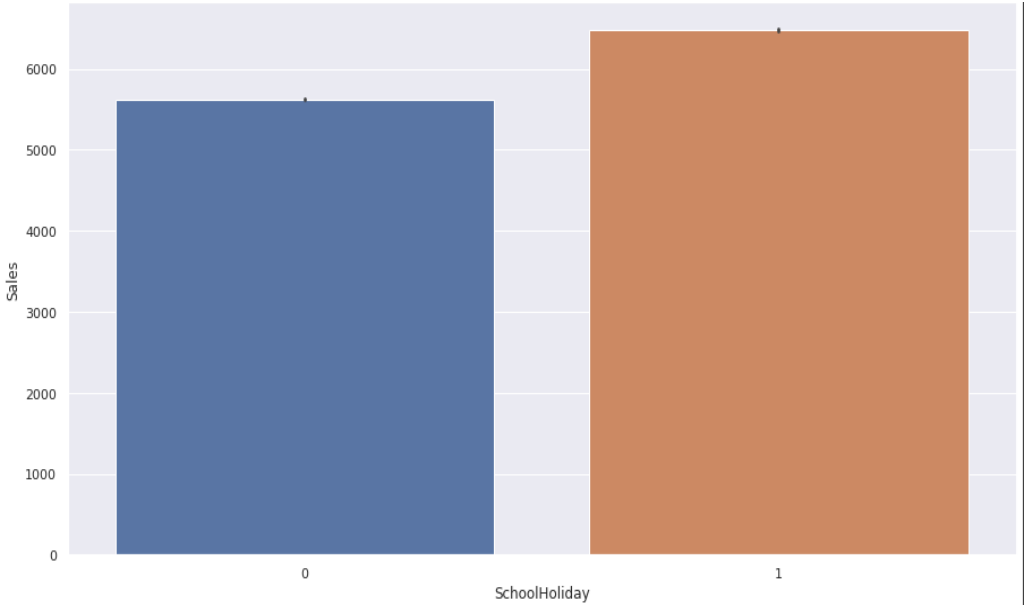




- From this Line plot sales at different months are different, corresponding to 3 years.
- The sales in the month of December is the highest sales among others.
- Sales were low in 2014 from July to September due to refurbishment.

Holidays Analysis

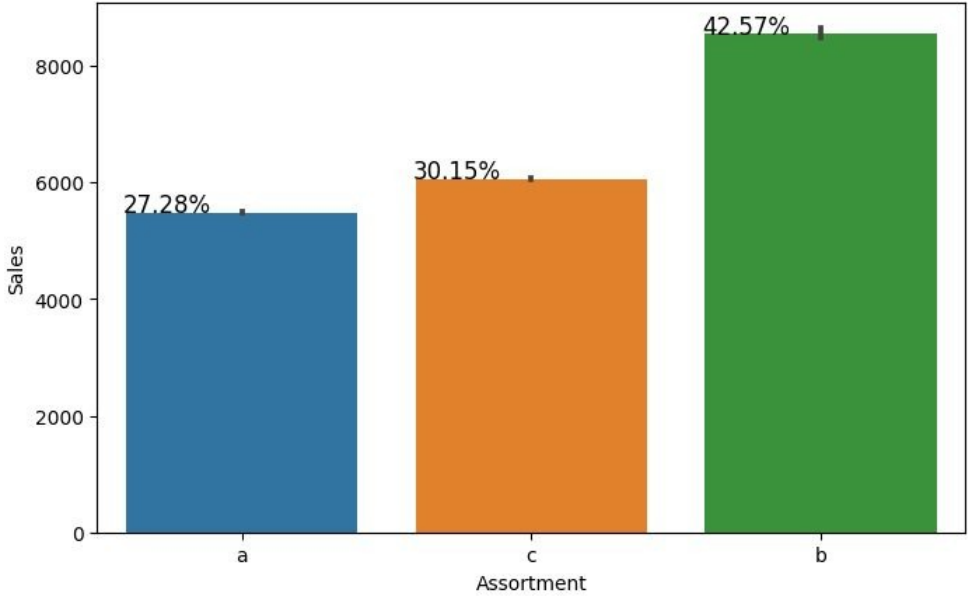


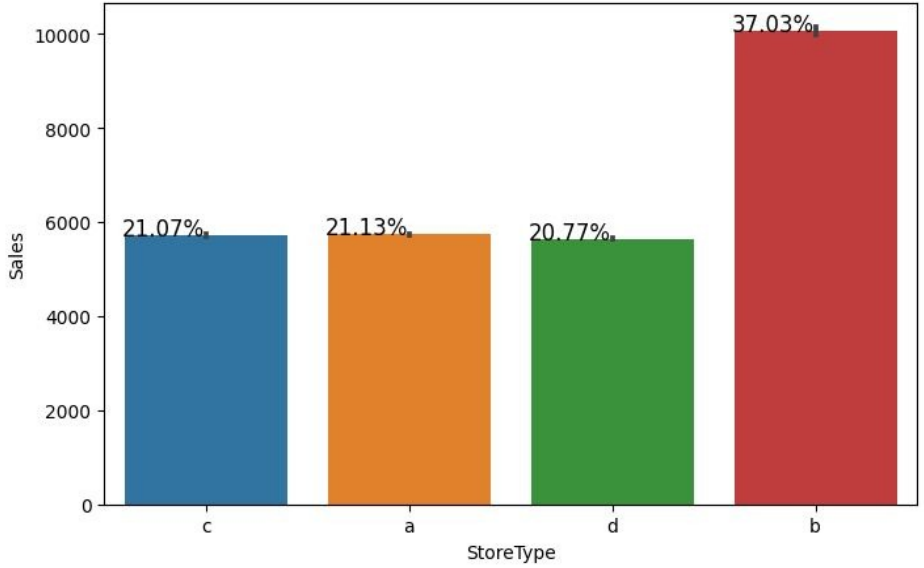


- The Sales at **state Holiday** are low as the stores are closed at state holidays.
- The Sales at **school Holiday** are higher as the stores are open at school holiday.

Store Type and Assortment Analysis

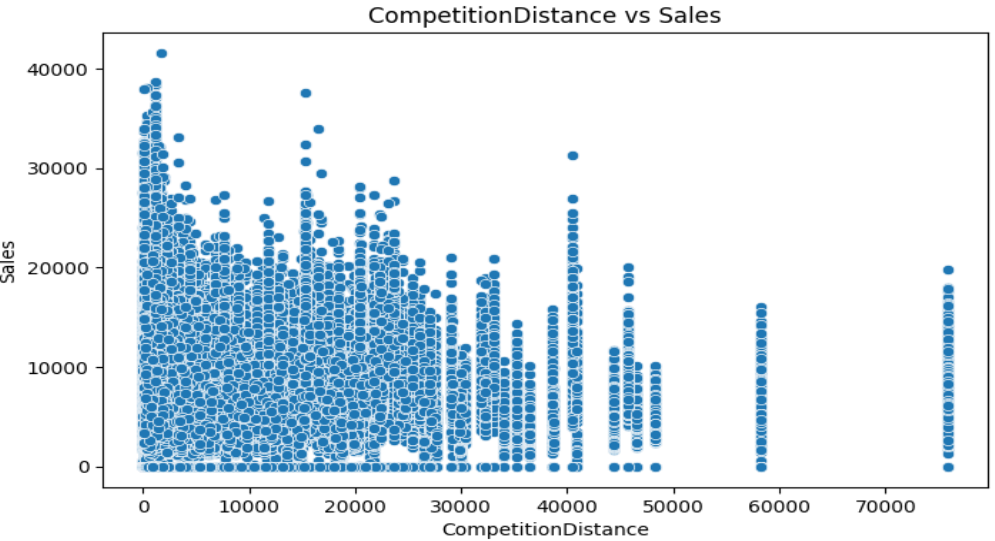
- The “b” store type had the highest number of sales.
- Type of Store plays an important role in the opening pattern of stores. All Type ‘b’ stores never closed except for refurbishment or other reasons.
- Assortment “b” had the highest number of sales.

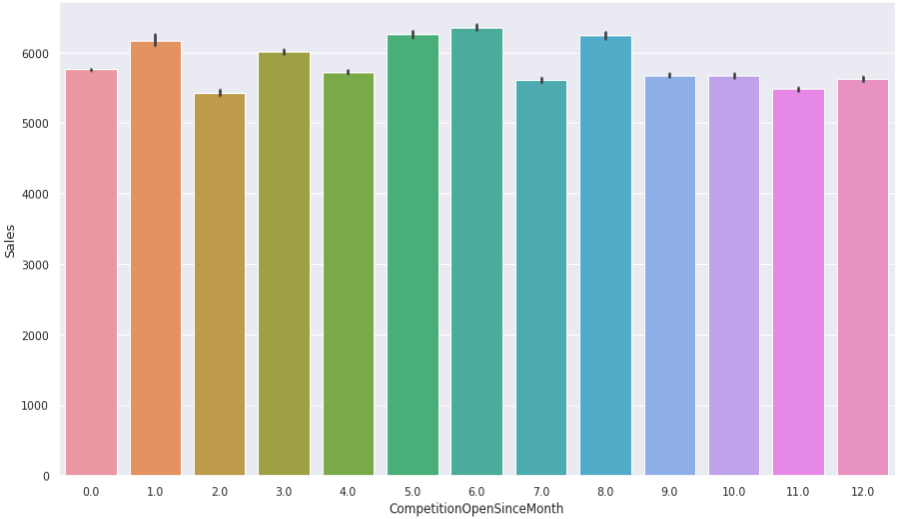




Competition Analysis

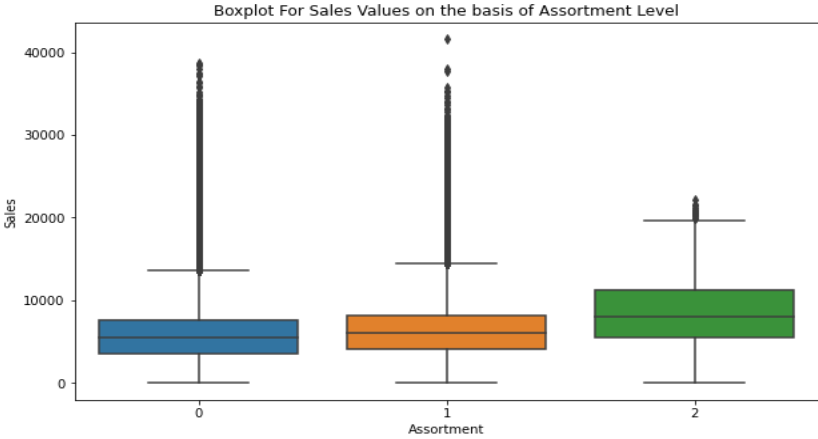
- Mostly stores were not that far from competitors and the stores were densely located near each other and surprisingly sales were higher when competition was nearer.

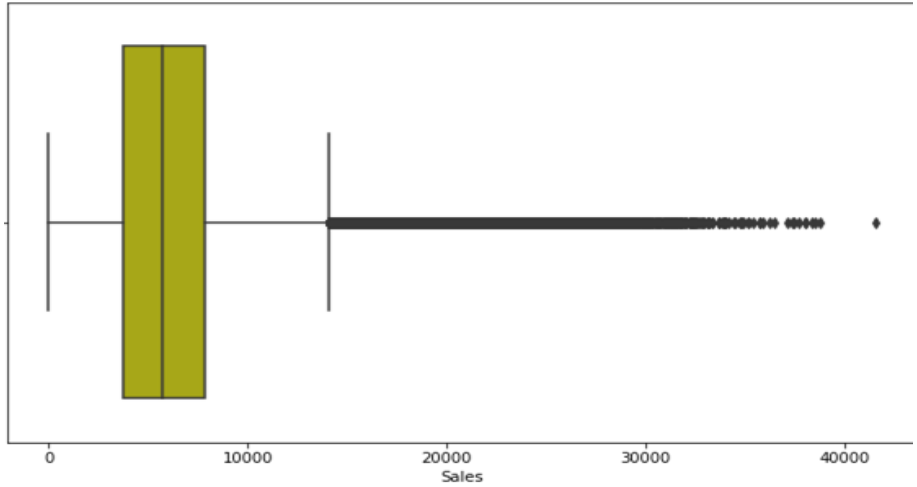


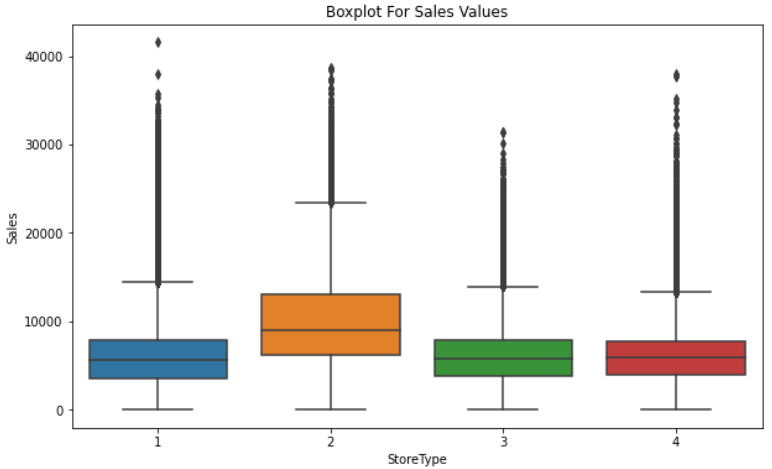


Outliers

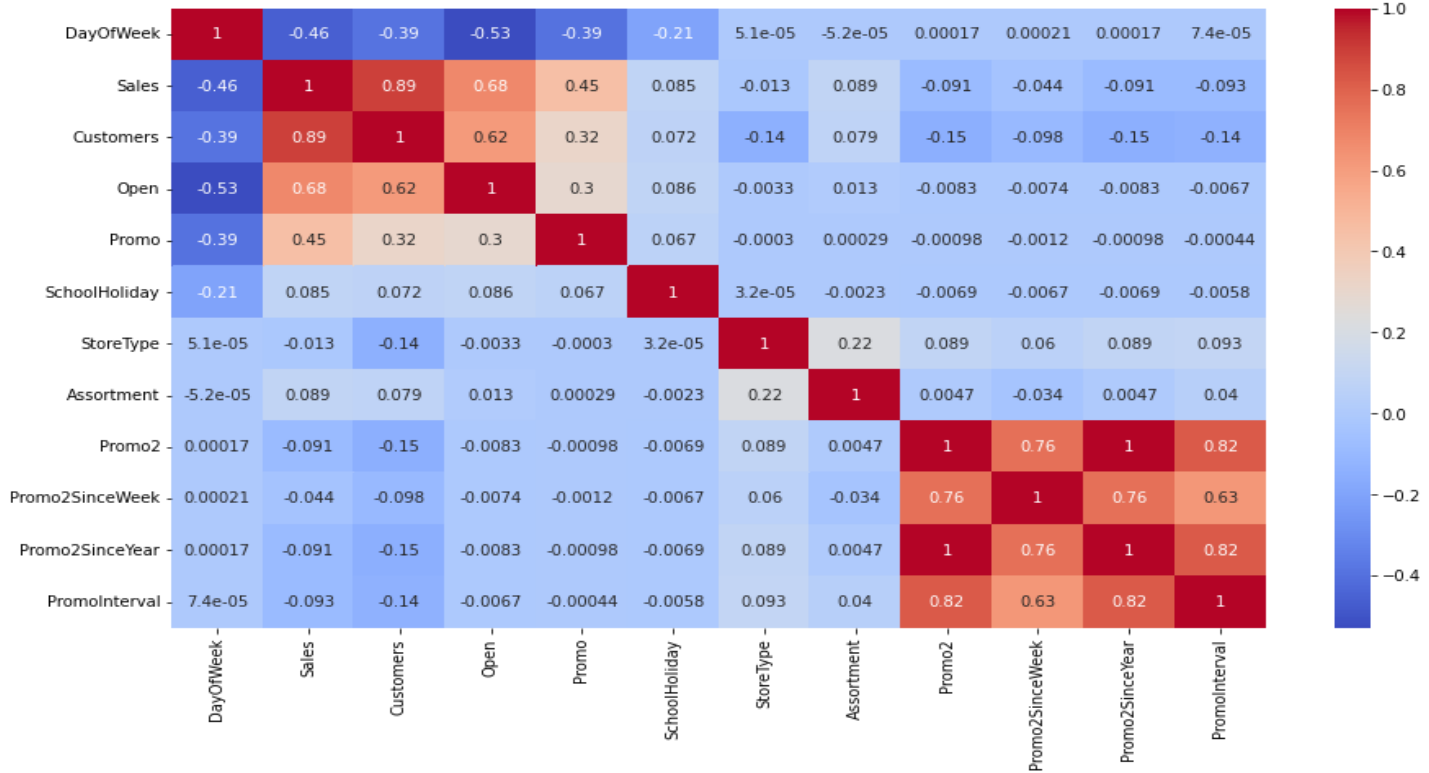
- We used box plot to identify the outliers.







Correlation between features



- `Customers, Sales, Open, Promo` are high and positively correlated with each other.
- Where `DayOFWeek` has negative correlation with this features.
- ``'Promo2', 'Promo2SinceWeek', 'Promo2SinceYear'`` has some correlation with each other.

Feature Engineering

Before proceeding to modelling we did some feature engineering to simplify and speed up the data transformation and to enhance model accuracy.

Feature Selection

- **Dependent Variable:** Sales
- **Independent Variables: 'Store', 'Day Of Week', 'Sales', 'Customers', 'Open', 'Promo', 'State Holiday', 'School Holiday', 'Store Type', 'Assortment','Promo2', 'Month' & 'Promo Interval',**

- **Scaling numerical feature:** We used MinMaxScalar to transform numerical feature.
- **Encoding:** We Encoded State holiday using get_dummies.

Comparison of Models

Algorithm	Train Score	Test Score
Linear Regression	0.867129	0.867895
Lasso Regression	0.866983	0.867701
Ridge Regression	0.867128	0.867891
Decision Tree	0.999257	0.972821
Decision Tree with Hyper Parameter Tuning	0.986933	0.979324
Random Forest	0.997199	0.983467
Random Forest with Hyper Parameter Tuning	0.997180	0.983123

R Squared Score is used for the Comparison of models:

- The R Squared score of all Linear Regression Algorithm is 0.86 even with Regularization.
- The R Squared score of the Decision Tree Regressor model we got 0.97 on the test set which is also good.
- The Random Forest regressor model performed 0.98 and is the most optimal model.
- We deployed Random Forest, for Sales forecasting for the next 6 weeks.

Feature Importance

The importance features are Customers, Store Type, Competition Distance & Store.

Feature	Feature Importance
Customers	0.86
StoreType	0.03
CompetitionDistance	0.03
Store	0.02
Promo	0.02
Assortment	0.02
DayOfWeek	0.01
CompetitionOpenSinceMonth	0.01
CompetitionOpenSinceYear	0.01
Promo2SinceWeek	0.01
StateHoliday	0.00
SchoolHoliday	0.00
Promo2	0.00
Promo2SinceYear	0.00
PromoInterval	0.00

Best Estimators

Best Parameters for Decision Tree Regressor
{min_samples_leaf=8,min_samples_split=5}

Best Parameters for Random Forest Regressor
{n_estimators=80,min_samples_split=2, min_samples_leaf=1}

Conclusion From EDA

- From plot sales and competition Open Since Month shows sales go increasing from November and highest in month December.
- From plot Sales and day of week, Sales highest on Monday and start declining from Tuesday to Saturday and on Sunday Sales are almost near to Zero.
- Plot between Promotion and Sales shows that promotion helps in increasing Sales.
- Type of Store plays an important role in opening pattern of stores.
- All Type ‘b’ stores never closed except for refurbishment or other reason.
- All Type ‘b’ stores have comparatively higher sales and it mostly constant with peaks appears on weekends.
- We can observe that most of the stores remain closed during State Holidays.
- The number of stores opened during School Holidays was more than that were opened during State Holidays.
- The sales in the month of December are the highest sales among others.
- The Promotion increases the sales so we should focus on that factor.
- As the customers are positively correlated with sales so we have to increase the frequency of customers by offers.
- The sales for store type B is higher than any other stores.

Conclusion

In our analysis, we initially did EDA on all the features of our dataset. We first analyzed our dependent variable, 'Sales' and also transformed it. Next, we analyzed the categorical variable and replaced null values, we also analyzed the numerical variable, found out the correlation, distribution and their relationship with the dependent variable using corr () Function. We also removed some numerical features that had mostly 0 values and hot encoded the categorical variables.

Next, we implemented six machine learning algorithms Linear Regression, lasso, ridge, decision tree, Random Forest. We did hyperparameter tuning to improve our model performance.

- The Sales are highest on Monday and start declining from Tuesday to Saturday and on Sunday Sales are almost near to Zero.
- Those Stores who take participate in Promotion got their Sales increased.
- Type of Store plays an important role in the opening pattern of stores. All Type ‘b’ stores never closed except for refurbishment or other reason.
- We can observe that most of the stores remain closed during State holidays. But it is interesting to note that the number of stores opened during School Holidays was more than that were opened during State Holidays.
- The R Squared score of all Linear Regression Algorithm is 0.86 even with Regularization.
- The R Squared score of the Decision Tree Regressor model we got 0.97 on the test set which is also good.
- The Random Forest regressor model performed 0.98 which is very well among others.
- The Random forest regressor model is the most optimal model and can be deployed.