CAPSTONE PROJECT-2
ON
RETAIL SALES PREDICTION

Rajat Mishra
(Cohort Warsaw)

PROBLEM STATEMENT

- Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Ross*mann 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.

- 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

WORK FLOW :

AI

So we will divide our work flow into following 3 steps.

- Data Collection and Understanding
- Data Cleaning and Manipulation
- Exploratory Data Analysis(EDA)
- Hypothesis Testing
- Feature engineering and Data preprocessing
- ML Model Implementation

EDA will be divided into following 3 analysis.

1) Univariate analysis: Univariate analysis is the simplest of the three analyses where the data you are analyzing is only one variable.

2) Bivariate analysis: Bivariate analysis is where you are comparing two variables to study their relationships.

3) Multivariate anlysis: Multivariate analysis is similar to Bivariate analysis but you are comparing more than two variables.

Hypothesis Testing is also main component in the project

Feature engineering and Data preprocessing includes such as handling null values,missing values as well as some new table creation,table manupulation etc

DATA COLLECTION AND UNDERSTANDING:

AI

Data Description:

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

DATA COLLECTION AND UNDERSTANDING:

AI

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

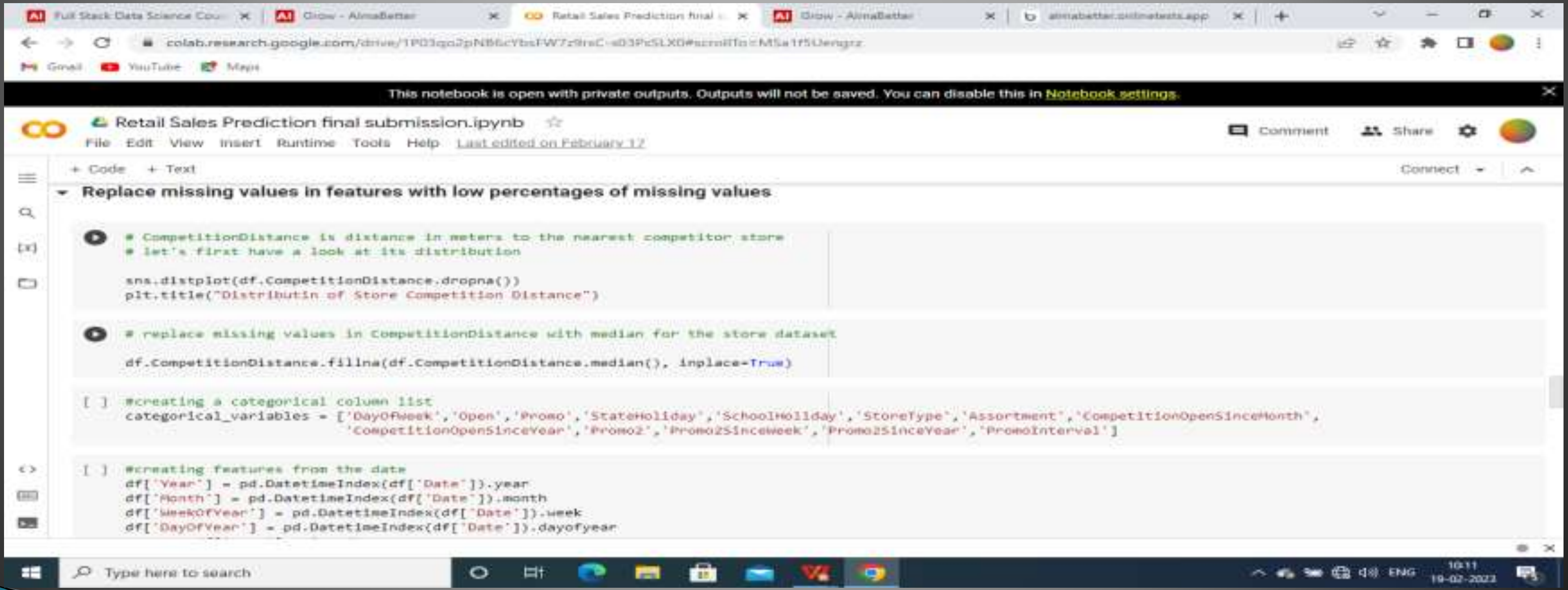
Promo2Since[Year/Week] – describes the year and calendar week when the store started participating in Promo2

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

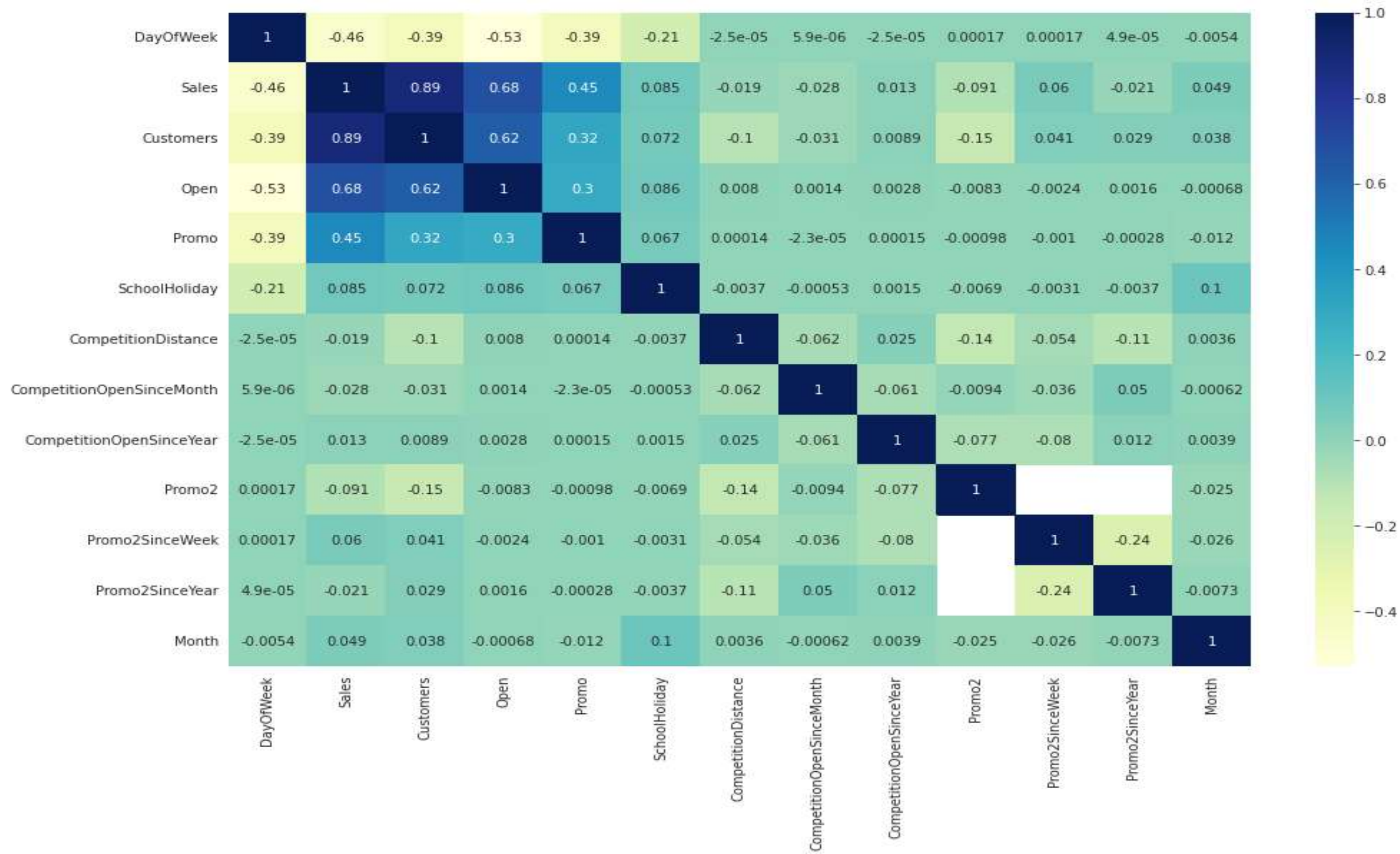
DATA MANUPULATION AND HANDLING:

▾ Check Unique Values for each variable.

```
[ ] # Check Unique Values for each variable.
    for column in df.columns:
        unique_values = df[column].unique()
        print(f'{column}: {unique_values}')
```

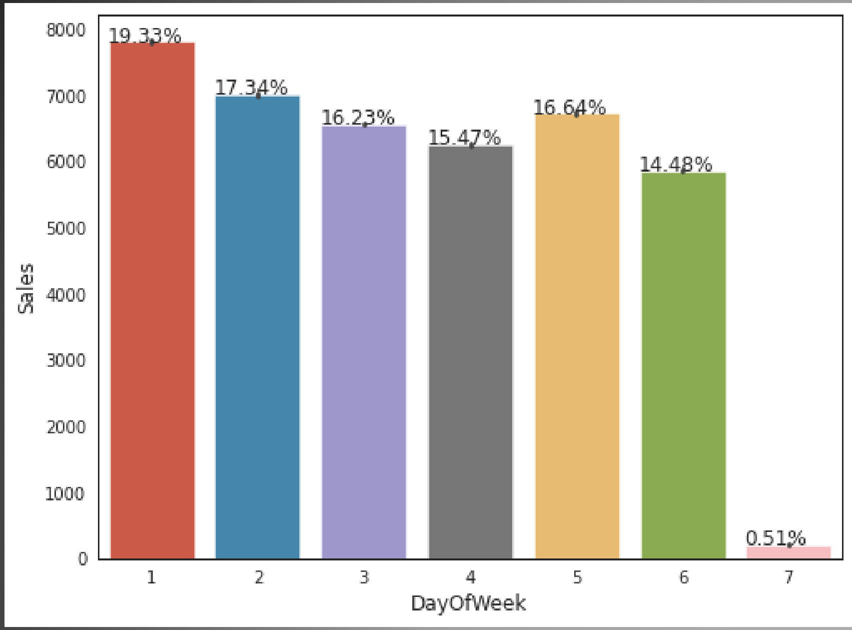


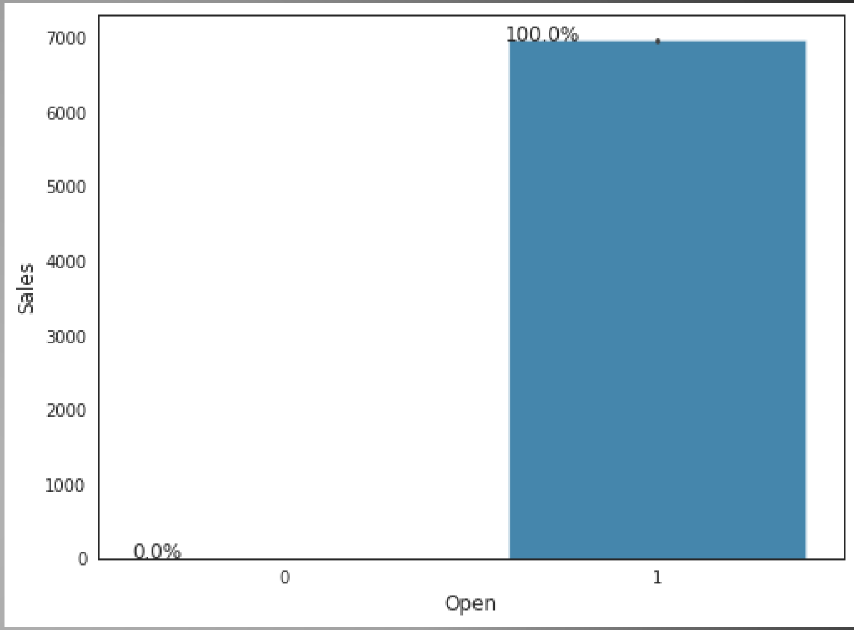
CORRELATION MATRIX

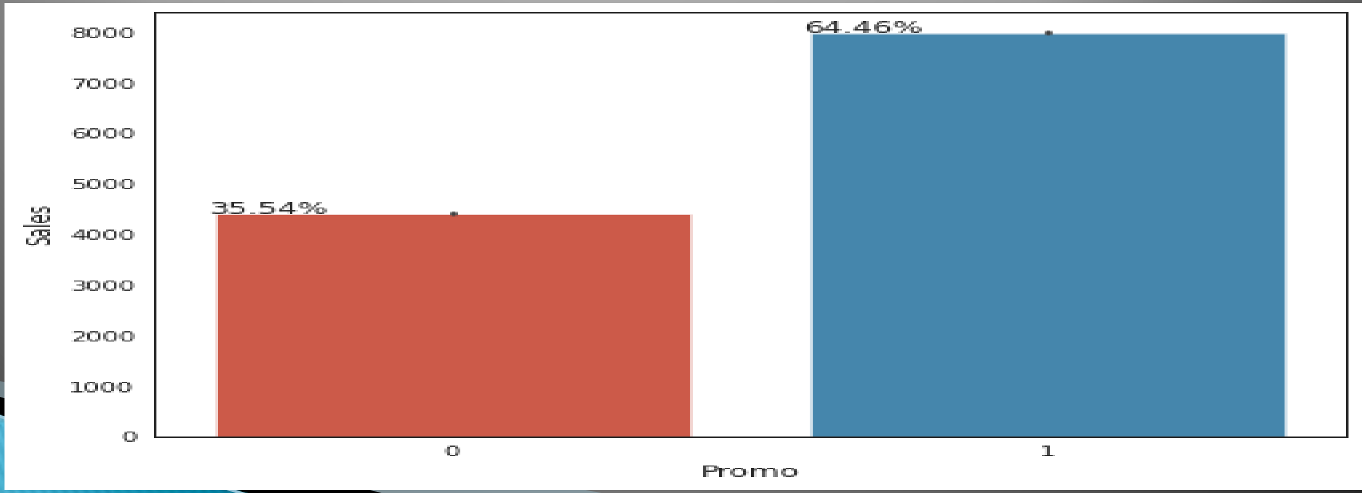


1.Day of the week has a negative correlation indicating low sales as the weekends, and promo, customers and open has positive correlation.
2.State Holiday has a negative correlation suggesting that stores are mostly closed on state holidays indicating low sales.
3.CompetitionDistance showing negative correlation suggests that as the distance increases sales reduce, which was also observed through the scatterplot earlier.
4.There's multicollinearity involved in the dataset as well. The features telling the same story like Promo2, Promo2 since week and year are showing multicollinearity.
5.The correlation matrix is agreeing with all the observations done earlier while exploring through barplots and scatterplots.

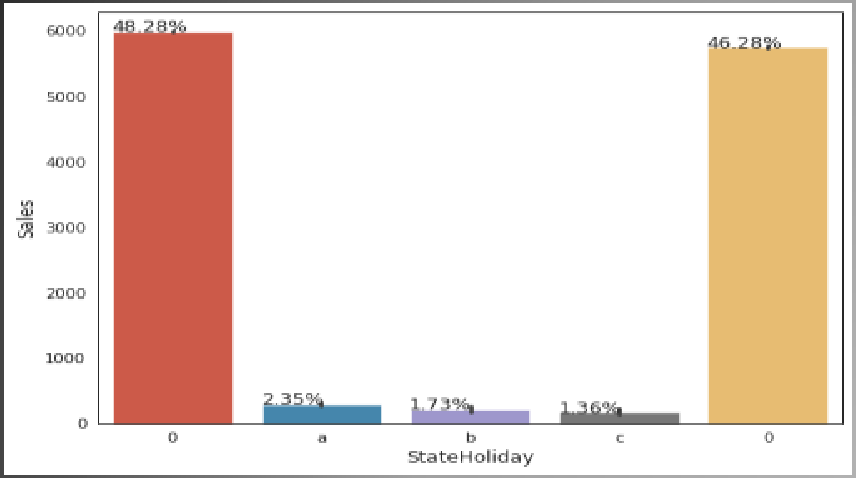
EXPLORATORY DATA ANALYSIS(EDA)

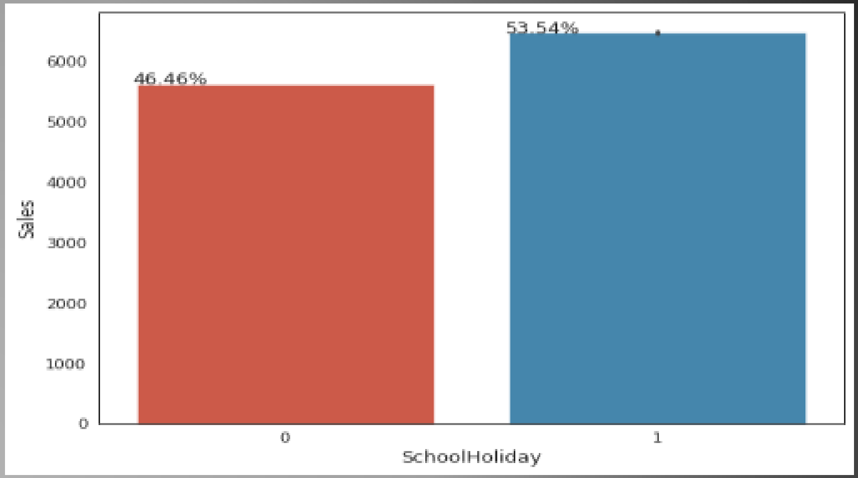


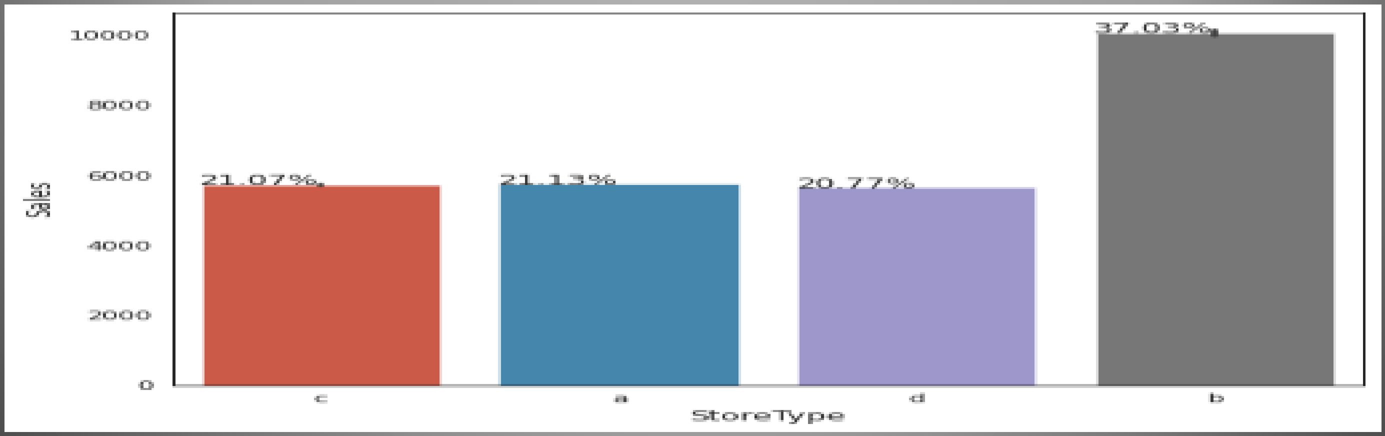




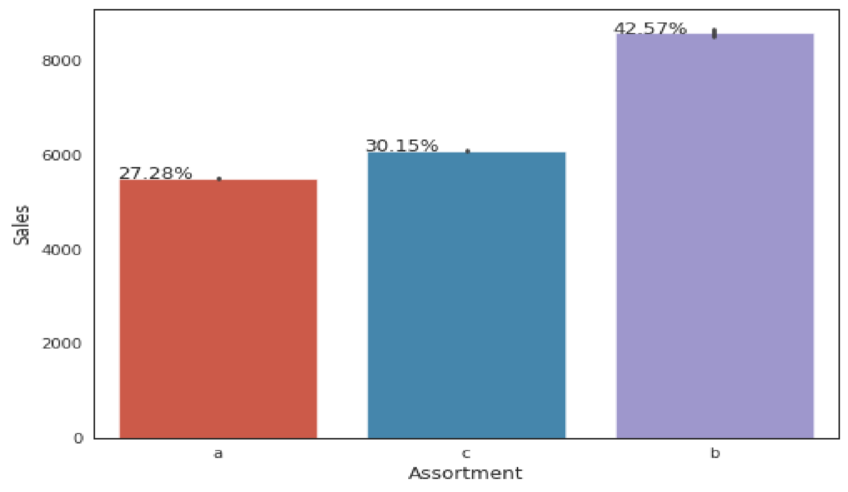
EXPLORATORY DATA ANALYSIS(EDA)

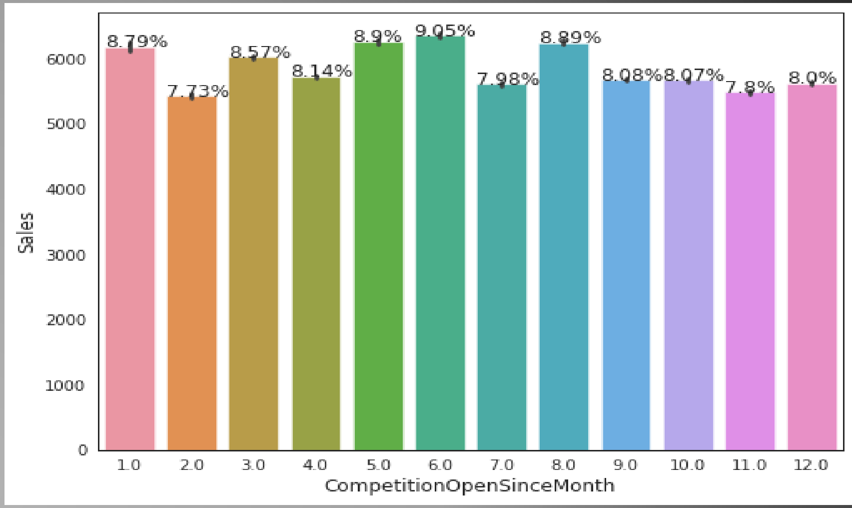


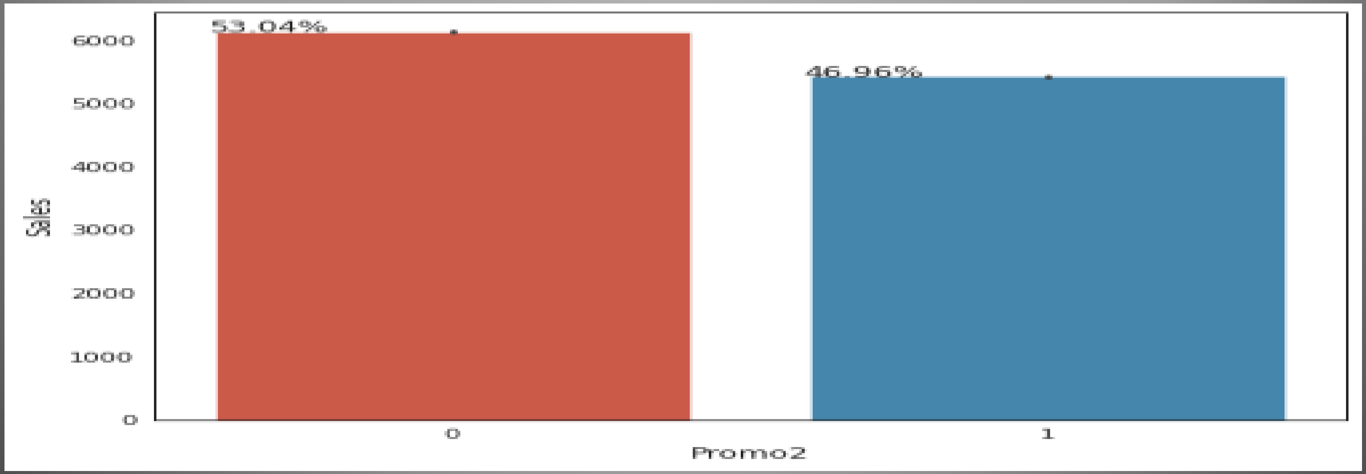




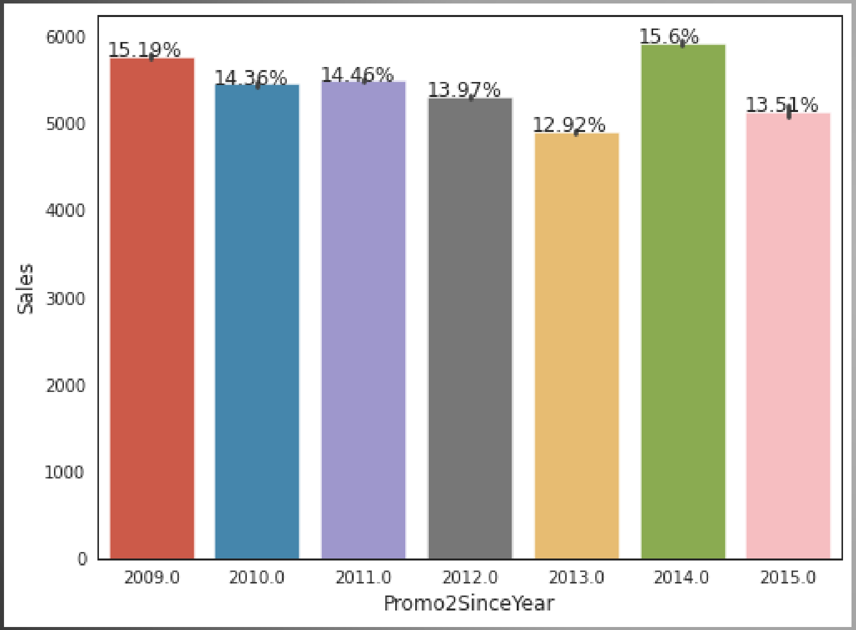
EXPLORATORY DATA ANALYSIS(EDA)

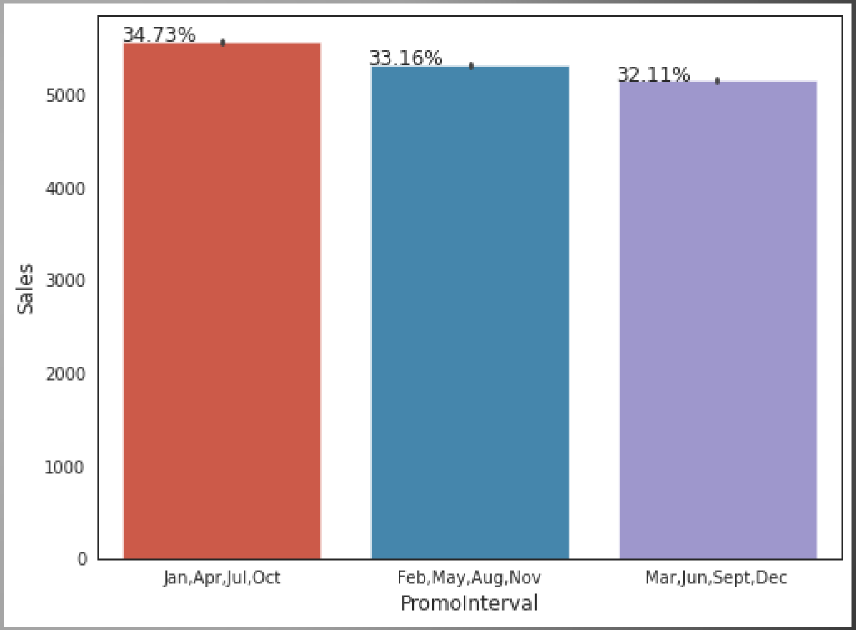






EXPLORATORY DATA ANALYSIS(EDA)





EXPLORATORY DATA ANALYSIS(EDA)

AI

Observations –

There were more sales on Monday, probably because shops generally remain closed on Sundays.

It could be seen that the Promo leads to more sales.

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. Lowest of Sales were seen on state holidays especially on Christmas.

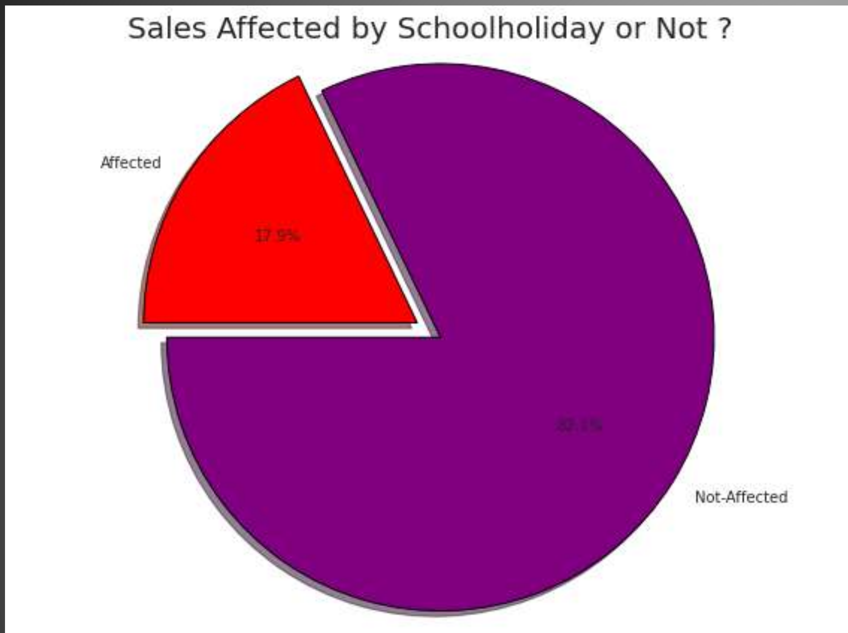
More stores were open on School Holidays than on State Holidays and hence had more sales than State Holidays.

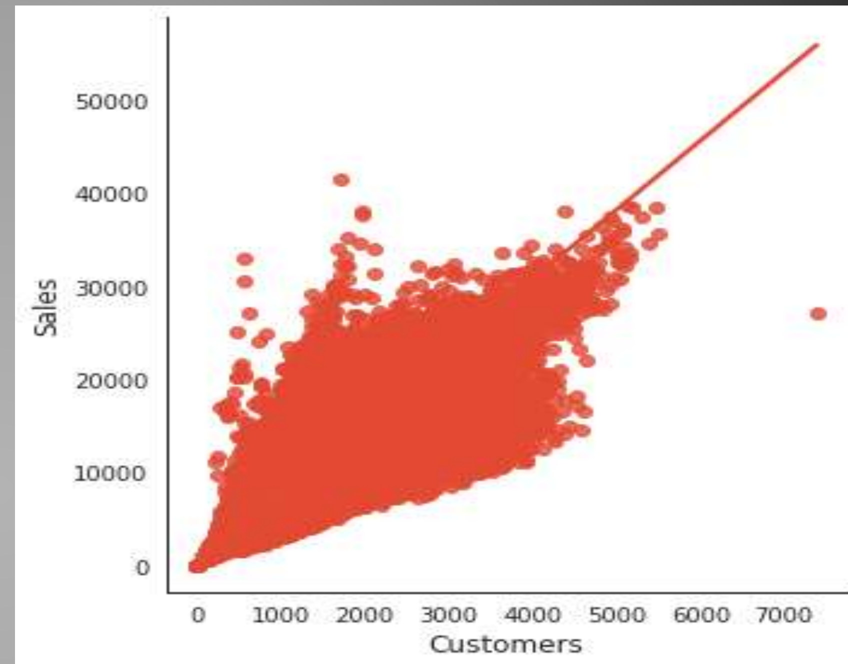
On an average Store type B had the highest sales.

Highest average sales were seen with Assortment levels-b which is 'extra'.

With Promo2, slightly more sales were seen without it which indicates there are many stores not participating in promo.

EXPLORATORY DATA ANALYSIS(EDA)



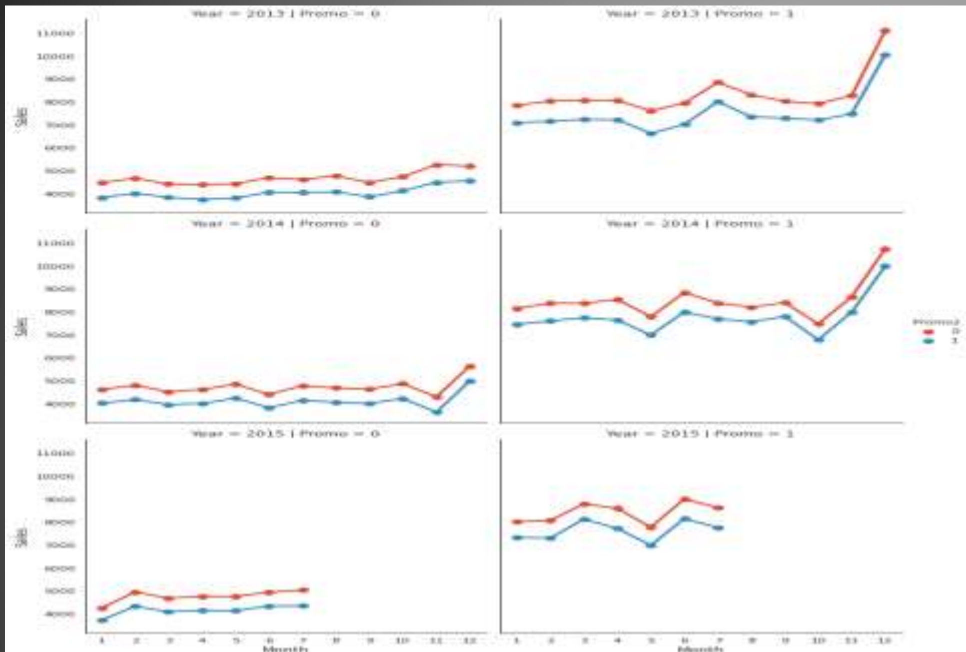


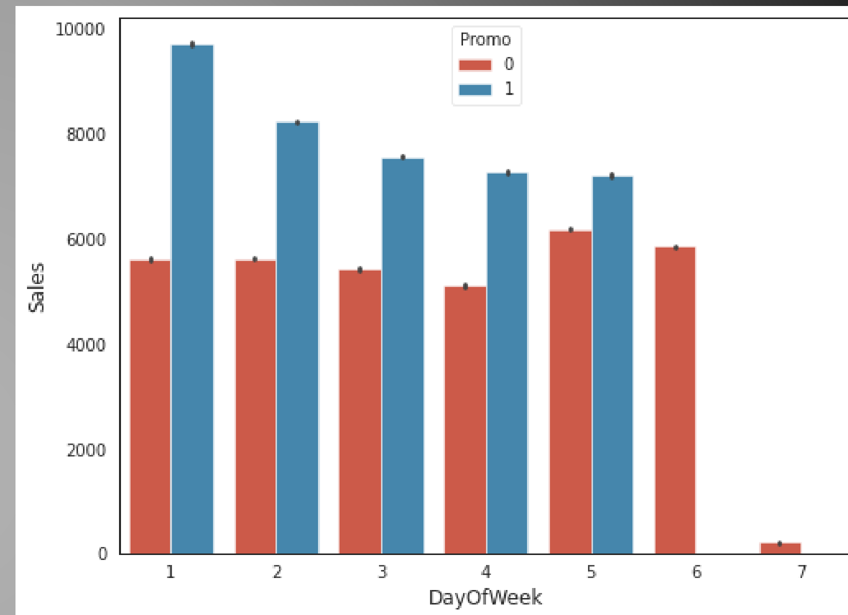
observations

1.82.1% sales are not affected and only 17.9% sales is affected because of schoo holiday

2.As we can see their is linear relationship between customers and sales as customers increasing sales also increasing

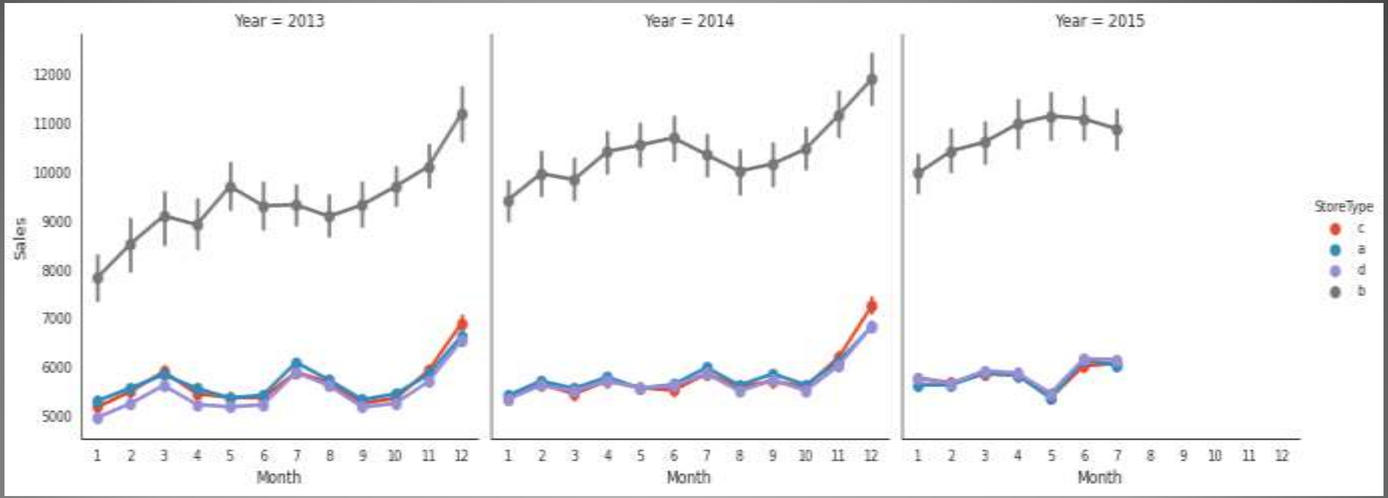
EXPLORATORY DATA ANALYSIS(EDA)

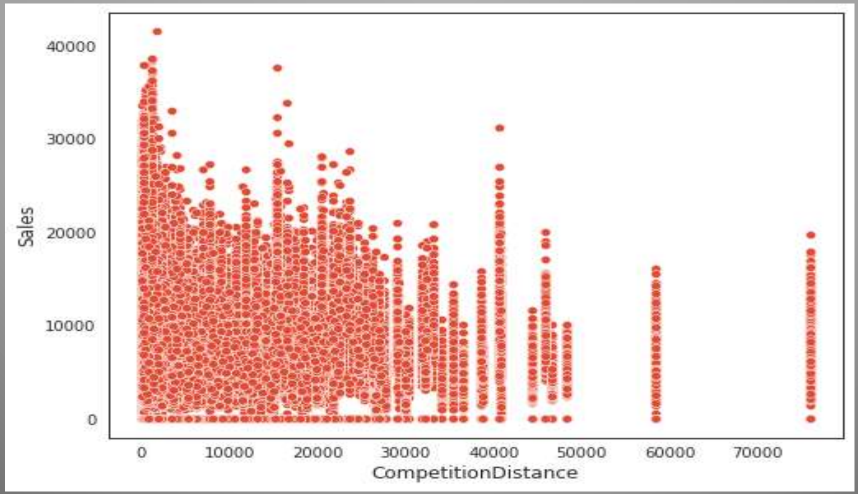




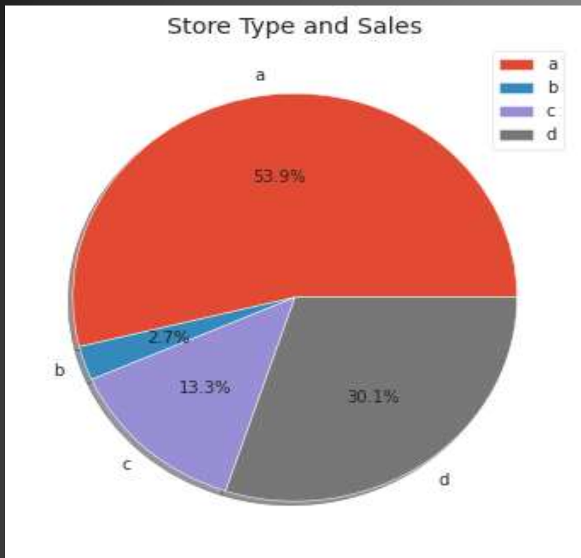
1.Here we can see that if their is no promo the sales is very less and if promo running their the sales is high.
2.Their is large diffrence on monday and it is decreasing day by day and on sunday their is no sales so it shwing less.

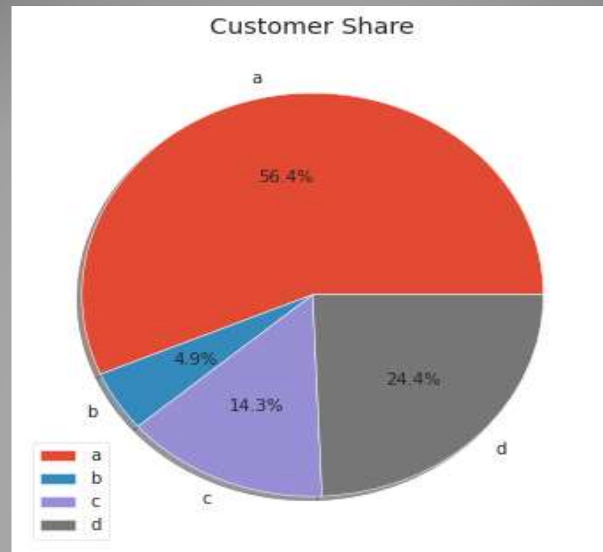
EXPLORATORY DATA ANALYSIS(EDA)

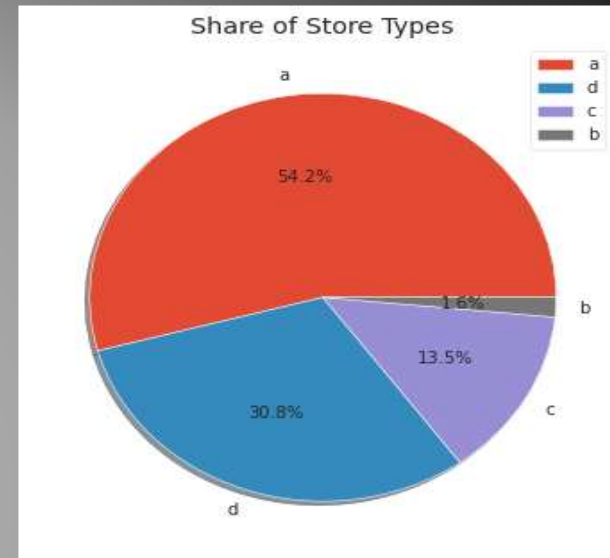




EXPLORATORY DATA ANALYSIS(EDA)







Observations –

1.In 2013 and 2014 their is some increasing in the sales but in 2015 their is some decreasing in trend of sales over the months

2.From the above scatter plot it can be observed that mostly the competitor stores weren't that far from each other and the stores densely located near each other saw more sales

3.A bar plot represents an estimate of central tendency for a numeric variable with the height of each rectangle. Earlier it was seen that the store type b had the highest sales on an average because the default estimation function to the barplot is mean.

4.But upon further exploration it can be clearly observed that the highest sales belonged to the store type a due to the high number of type a stores in our dataset. Store type a and c had a similar kind of sales and customer share.

5.Interesting insight to note is that store type b with highest average sales and per store revenue generation looks healthy and a reason for that would be all three kinds of assortment strategies involved which was seen earlie

FEATURE ENGINEERING

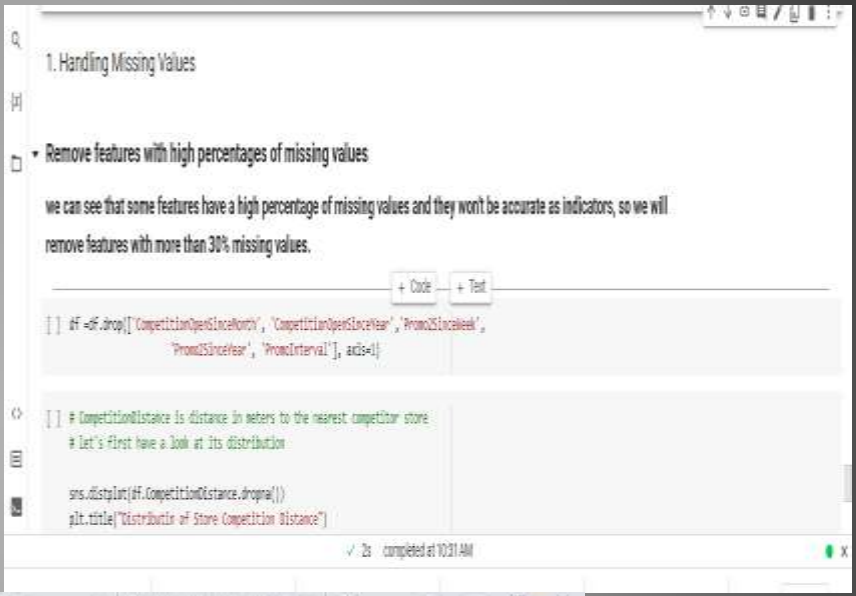
Replace missing values in features with low percentages of missing values

```
# CompetitionDistance is distance in meters to the nearest competitor store
# let's first have a look at its distribution

sns.distplot(df.CompetitionDistance.dropna())
plt.title("Distributin of Store Competition Distance")
```

```
# replace missing values in CompetitionDistance with median for the store dataset

df.CompetitionDistance.fillna(df.CompetitionDistance.median(), inplace=True)
```

```
#creating a categorical column list
categorical_variables = ['DayOfWeek','Open','Promo','StateHoliday','SchoolHoliday','StoreType','Assortment','CompetitionOpenSinceMonth',
                         'CompetitionOpenSinceYear','Promo2','Promo2SinceWeek','Promo2SinceYear','PromoInterval']
```

```
#creating features from the date
df['Year'] = pd.DatetimeIndex(df['Date']).year
df['Month'] = pd.DatetimeIndex(df['Date']).month
df['WeekOfYear'] = pd.DatetimeIndex(df['Date']).week
df['DayOfYear'] = pd.DatetimeIndex(df['Date']).dayofyear
```



1. Handling Missing Values

Remove features with high percentages of missing values

we can see that some features have a high percentage of missing values and they won't be accurate as indicators, so we will remove features with more than 30% missing values.

```
df =df.drop(['CompetitionOpenSinceMonth', 'CompetitionOpenSinceYear','Promo2SinceWeek',
             'Promo2SinceYear', 'PromoInterval'], axis=1)
```

```
# CompetitionDistance is distance in meters to the nearest competitor store
# let's first have a look at its distribution

sns.distplot(df.CompetitionDistance.dropna())
plt.title("Distributin of Store Competition Distance")
```

Retail Sales Prediction final submission.ipynb

This notebook is open with private outputs. Outputs will not be saved. You can disable this in Notebook settings.

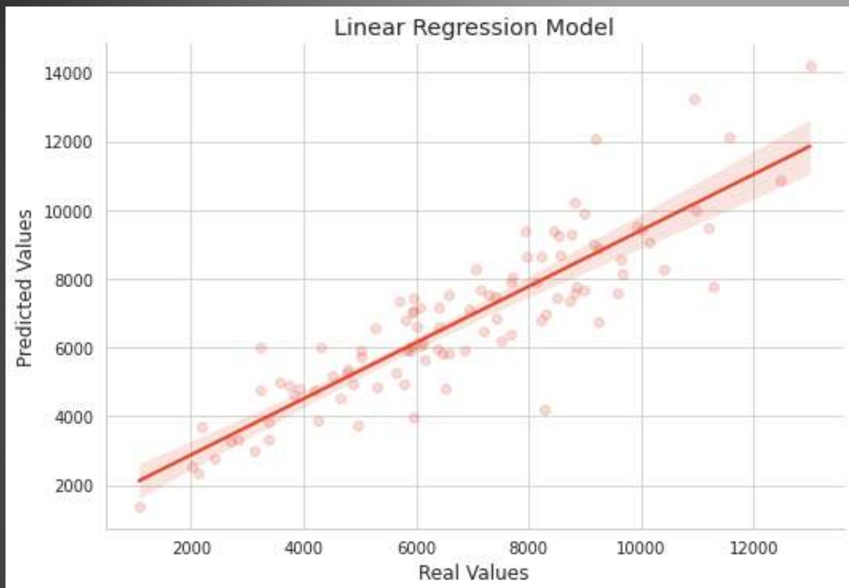
The distribution is right skewed, so we'll replace missing values with the median.

```
# replace missing values in CompetitionDistance with median for the store dataset
df.CompetitionDistance.fillna(df.CompetitionDistance.median(), inplace=True)
```

Outliers Handling

```
#removing outliers
def remove_outlier(df_in, col_name):
    q1 = df_in[col_name].quantile(0.25)
    q3 = df_in[col_name].quantile(0.75)
    iqr = q3-q1 #Interquartile range
    fence_low = q1-1.5*iqr
    fence_high = q3+1.5*iqr
    df_out = df_in.loc[(df_in[col_name] > fence_low) & (df_in[col_name] < fence_high)]
```

ML Model Implementation



Conclusion

The MSE and R2 score are commonly used evaluation metrics for regression models. In this case, the Linear Regression and Lasso Regression models have very similar performance, with the Lasso Regression model having a slightly lower MSE and a slightly higher R2 score.

The mean squared error (MSE) measures the average squared difference between the predicted and actual values, where a lower MSE indicates better performance. The R-squared (R2) score measures the proportion of the variance in the dependent variable that is predictable from the independent variables, where a higher R2 score indicates better performance.

THANK YOU...!!