**Retail Sales Prediction**

**By**

**Akash raj**

**Data science trainees,**

**Alma Better, Bangalore**

**Abstract:**

Sales forecasting refers to the process of estimating demand for or sales of a particular product over a specific period of time. Businesses use sales forecasts to determine what revenue they will be generating in a particular timespan to empower themselves with powerful and strategic business plans. Important decisions such as budgets, hiring, incentives, goals, acquisitions and various other growth plans are affected by the revenue the company is going to make in the coming months and for these plans to be as effective as they are planned to be it is important for these forecasts to also be as good. The sales forecasts are also different from the sales-goals a company has. Sales-goals is what a company wants to happen to execute their future plans for the business. On the other hand sales forecasts are what is going to happen on the basis of past records, data, trends and various improvement measures taken. The work here predicts the sales for a drug store chain in the European market for a time period of six weeks and compares the results of machine learning algorithms.

**Keywords:**

Data Collection and Preprocessing

- Importing important libraries and modules  
- Data Cleaning  
- Missing Data Handling  
- Merging the Datasets

Exploratory Data Analysis

- Categorical Features  
- Numerical Features  
- EDA Conclusion

Feature Selection and Outlier Detection

- Feature Engineering  
- Outlier Detection and Treatment

Modeling

- Train Test Split

-Linear

-Polynomial

-Lasso

-Ridge

-Baseline Model

-Decision Tree

-Feature Importance

|  |
| --- |

**Problem Description**

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. 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 .

**Introduction:**

The interest for a product continues to change occasionally. No business can work on its monetary growth without assessing client interest and future demand of items precisely. Sales forecasting refers to the process of estimating demand for or sales of a particular product over a specific period of time. For a good sales forecast, it is extremely important to get a good dataset as well. Forecasts heavily depend on the past records, trends and patterns observed for sales of a particular store. The variations could be due to a number of reasons.

**Understanding the Data:**

First step involved is understanding the data and getting answers to some basic questions like; What is the data about? How many rows or observations are there in it? How many features are there in it? What are the data types? Are there any missing values? And anything that could be relevant and useful to our investigation. Let’s just understand the dataset first and the terms involved before proceeding further. Our dataset consists of two csv files, the first consists of historical data with 1017209 rows or observations and 9 columns with no null values. The second dataset was supplementary information about the stores with 1115 rows and 10 columns and a lot of missing values in a few columns. The data types were of integer, float and object in nature.

**Let’s define the features involved:**

### Data fields

### Most of the fields are self-explanatory. The following are descriptions for those that aren't.

#### 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

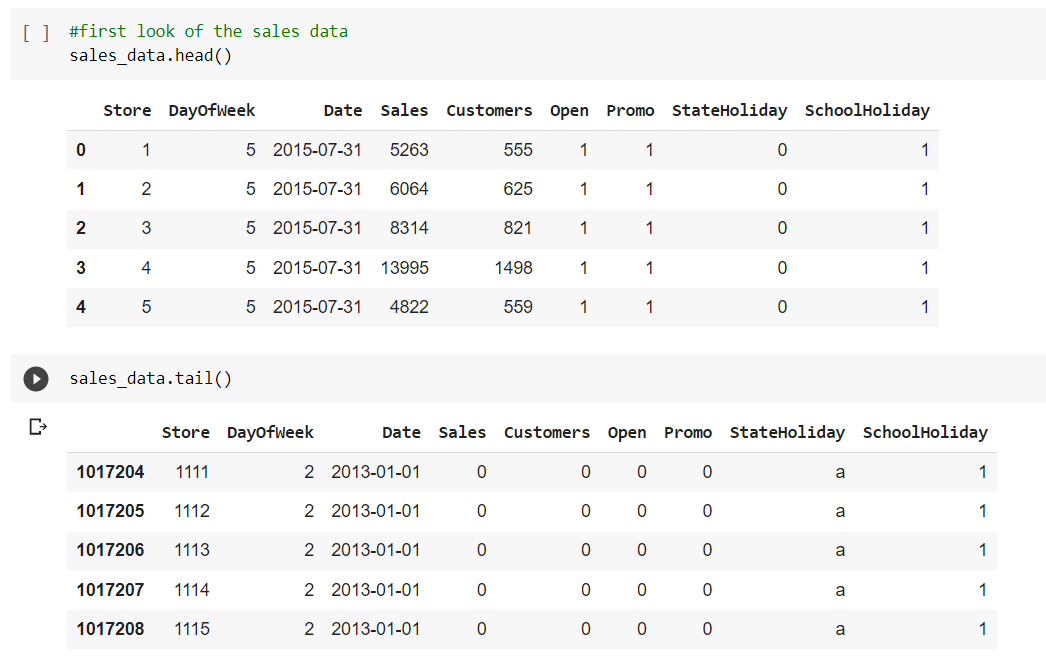
#### Promo - indicates whether a store is running a promo on that day

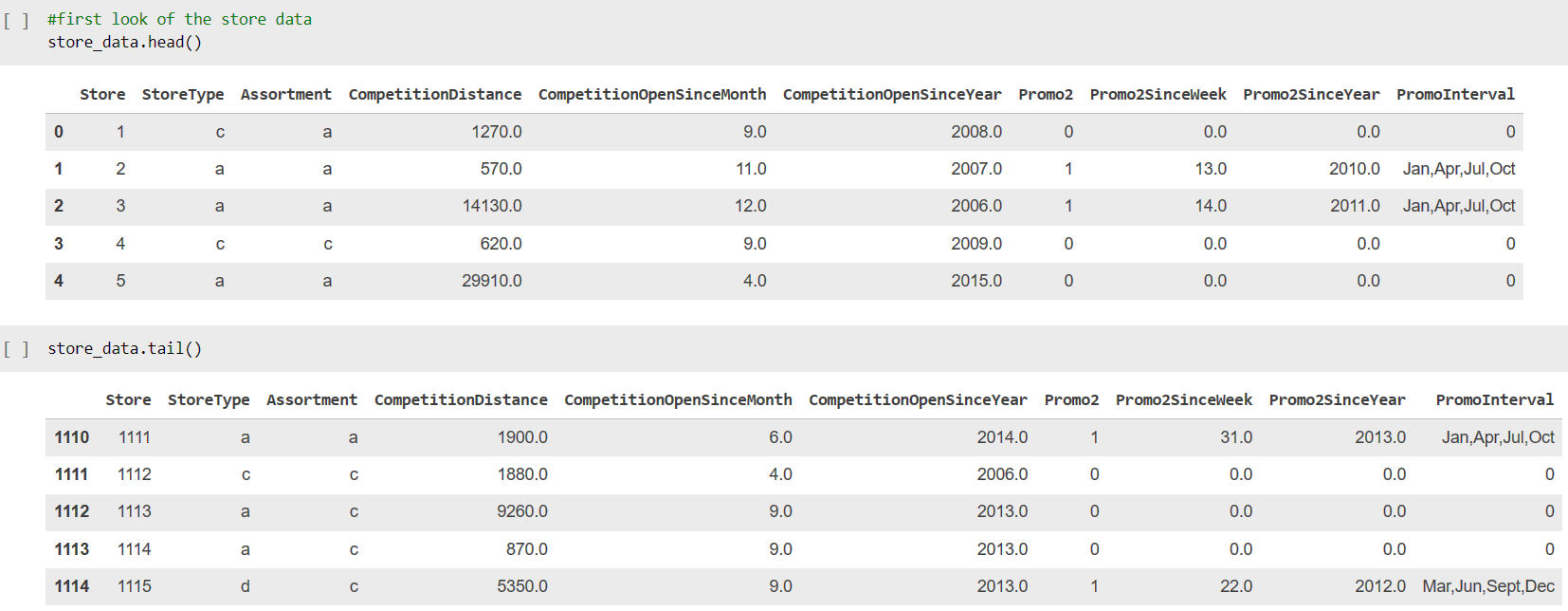
#### Promo2 - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating

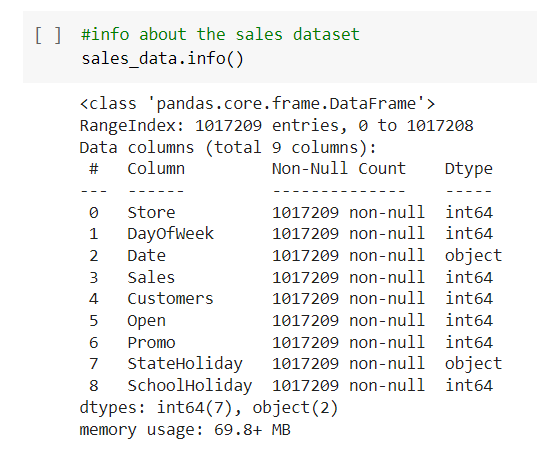
#### Promo2Since[Year/Week] - describes the year and calendar week when the store started participating in Promo2

#### 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

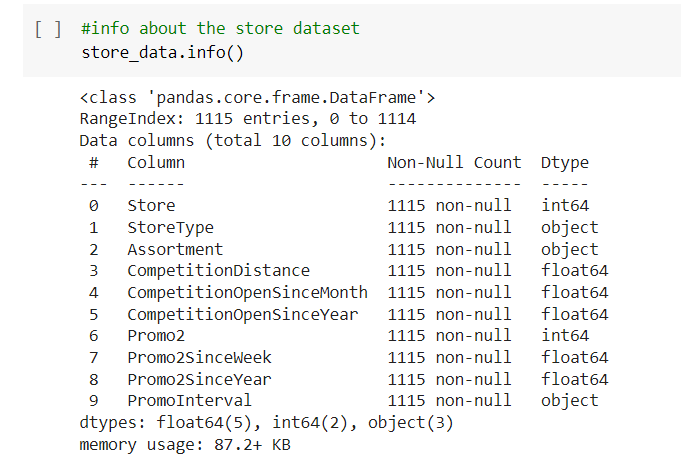
**Exploring Data:**

****

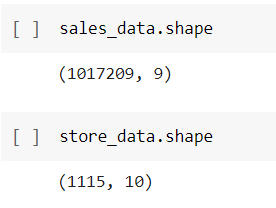
****

****

**There are 1017209 rows or observations and 9 columns in this dataset. There seems to be no null values in it. It has integer, datetime and object as data types.**

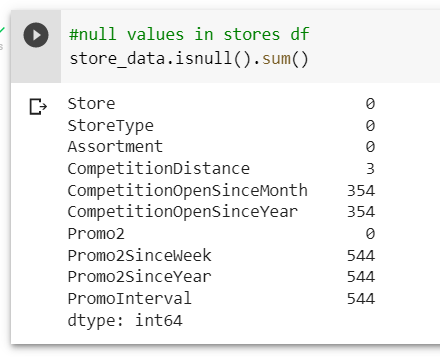
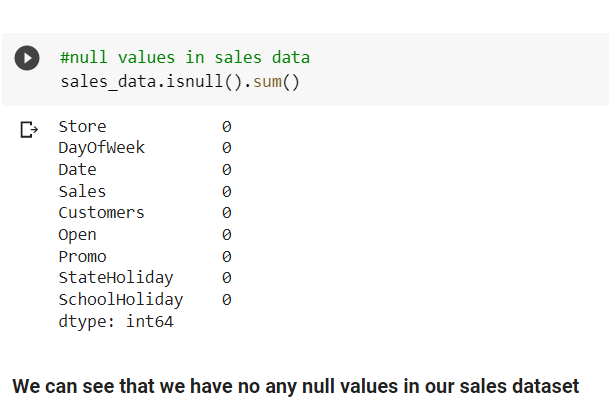
****

**There are 1115 rows and 10 columns. There are missing values in it and it is important to impute them with appropriate values in order to get good results later on.**

****

**Data Cleaning and Preprocessing:**

Handling missing values is an important skill in the data analysis process. If there are very few missing values compared to the size of the dataset, we may choose to drop rows that have missing values. Otherwise, it is better to replace them with appropriate values.



# The dataset had a lot of nulls in the following columns:

# \*\* Out of 1115 entries there are missing values for the columns:\*\*

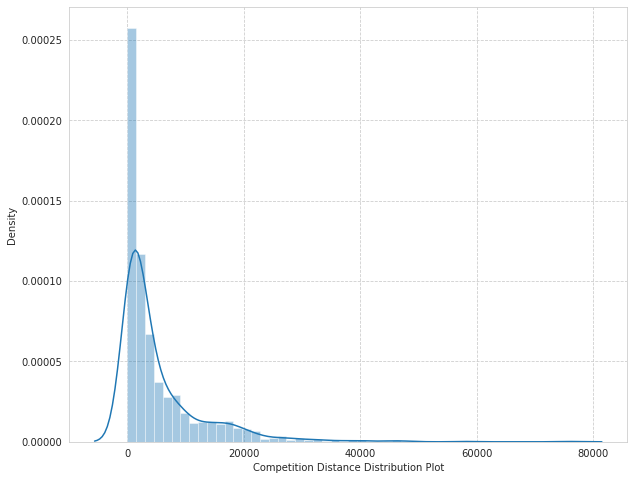
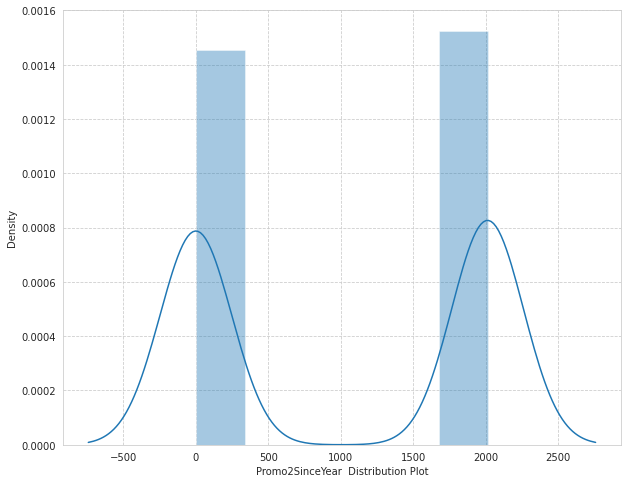
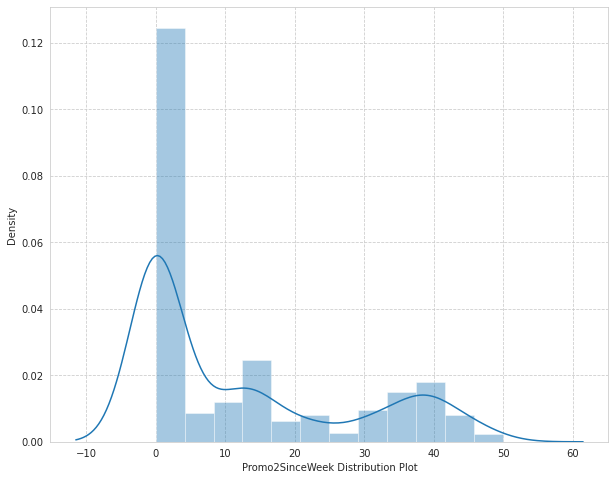
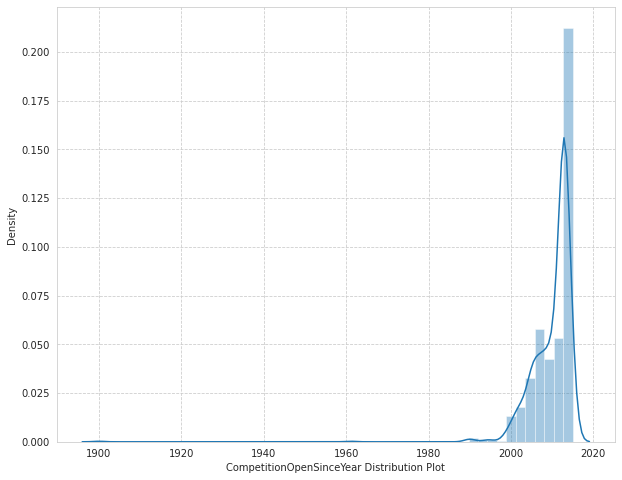
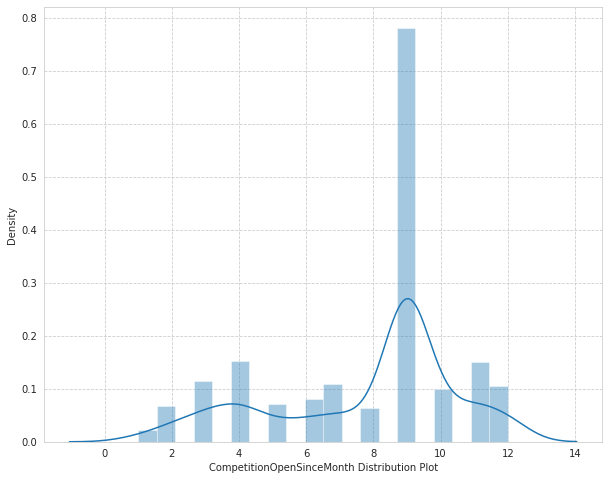
\*\* CompetitionDistance- distance in meters to the nearest competitor store, the distribution plot would give us an idea about the distances at which generally the stores are opened and we would impute the values accordingly

\*\* CompetitionOpenSinceMonth- gives the approximate month of the time the nearest competitor was opened, mode of the column would tell us the most occuring month

\*\* CompetitionOpenSinceYear- gives the approximate year of the time the nearest competitor was opened, mode of the column would tell us the most occuring month

\*\* Promo2SinceWeek, Promo2SinceYear and PromoInterval are NaN wherever Promo2 is 0 or False as can be seen in the first look of the dataset. They can be replaced with 0.

# Distribution plot of all null values :

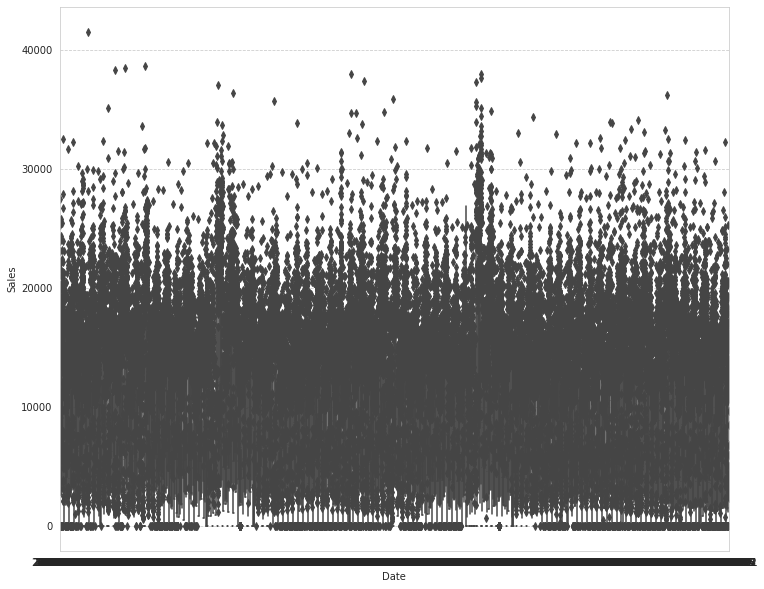


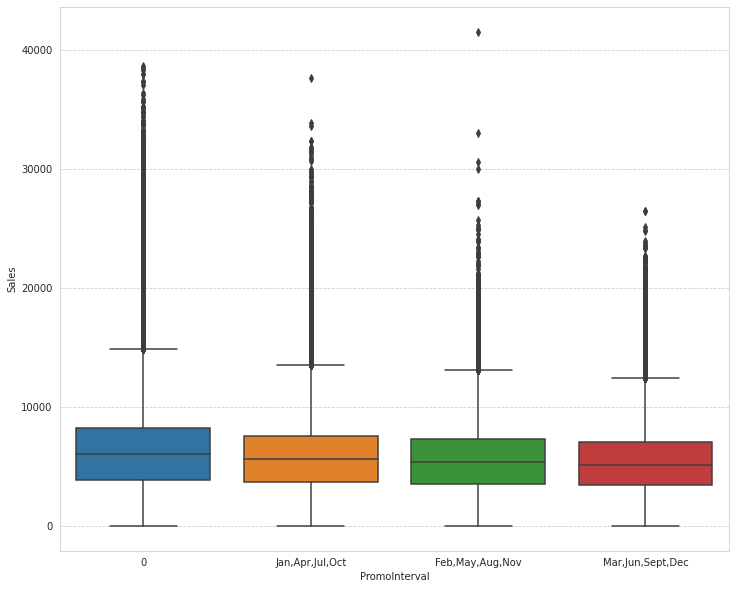
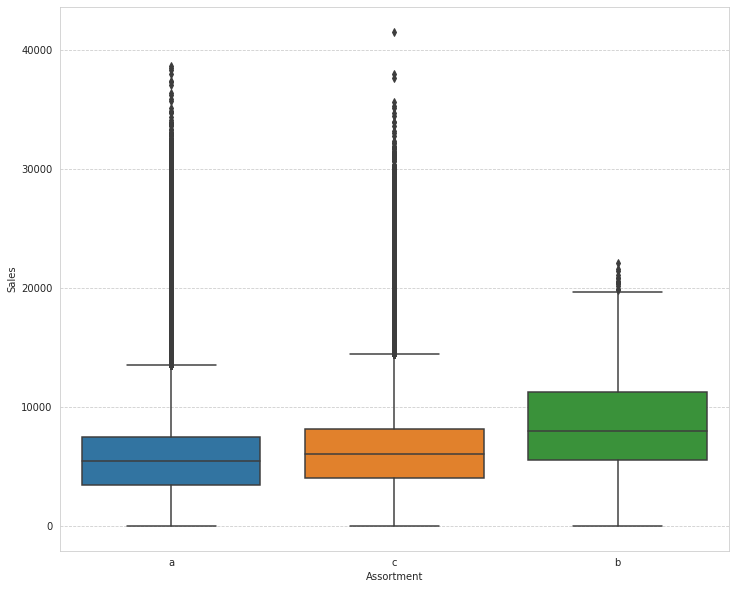
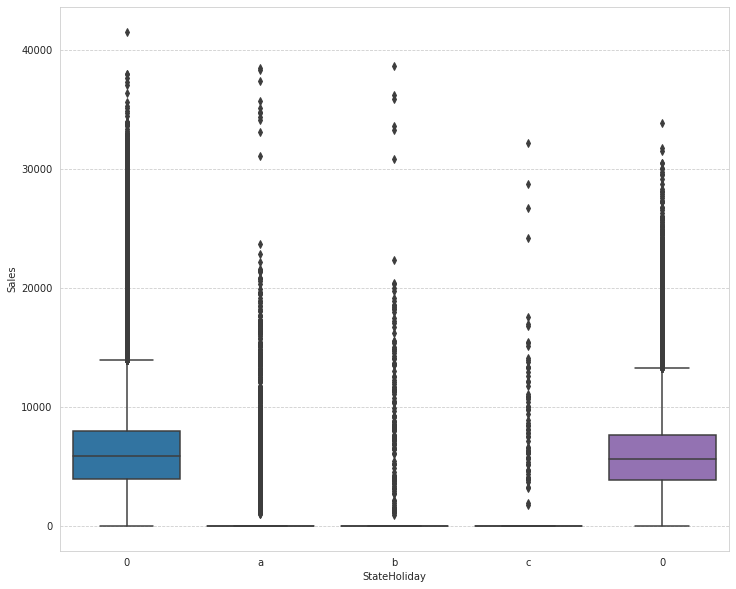
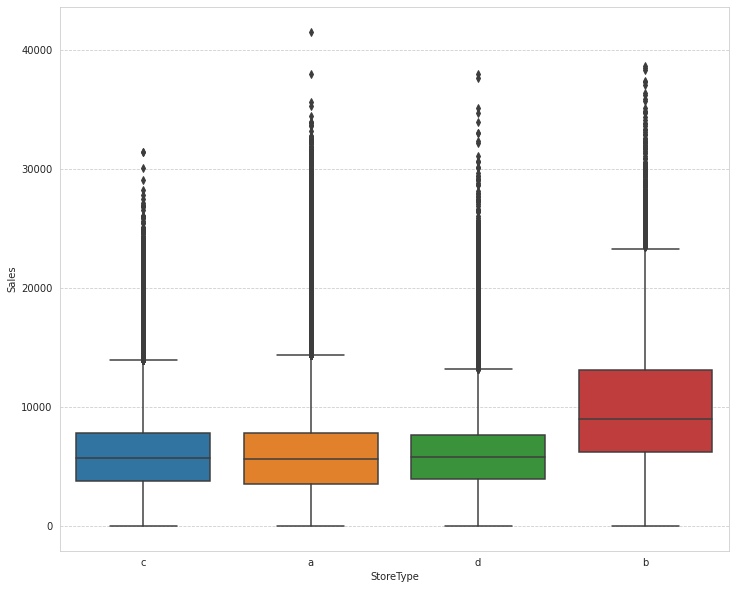
**Exploratory Data Analysis:**

Exploratory data analysis is a crucial part of data analysis. It involves exploring and analyzing the dataset given to find out patterns, trends and conclusions to make better decisions related to the data, often using statistical graphics and other data visualization tools to summarize the results. The visualization tools involved in the investigation are python libraries- matplotlib and seaborn. The goal here is to explore the relationships of different variables with ‘Sales’ to see what factors might be contributing to the high and low sales numbers. Approach

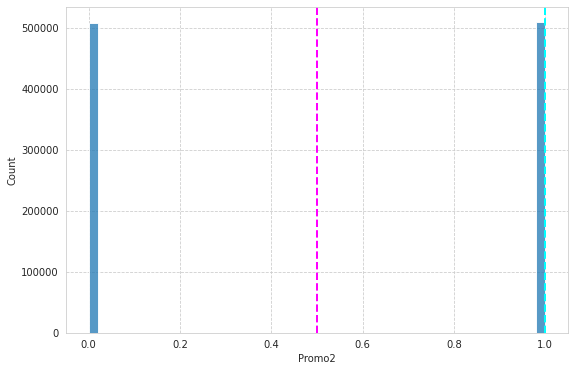
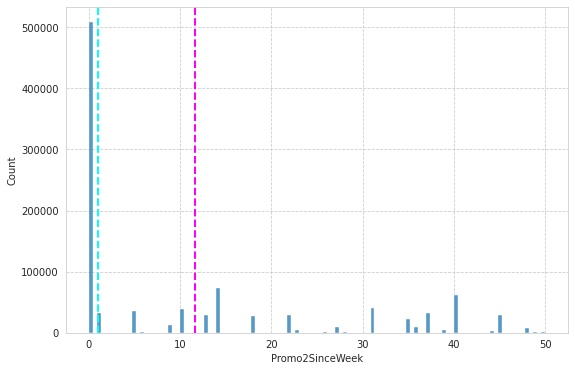
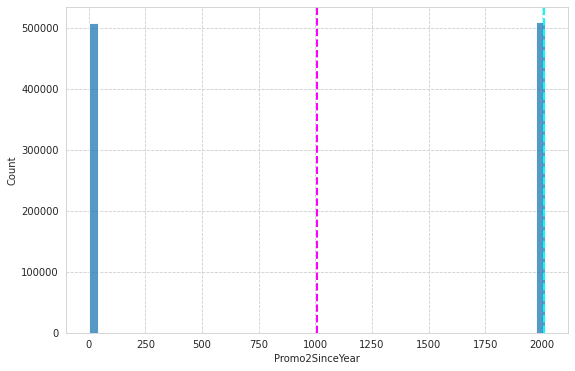
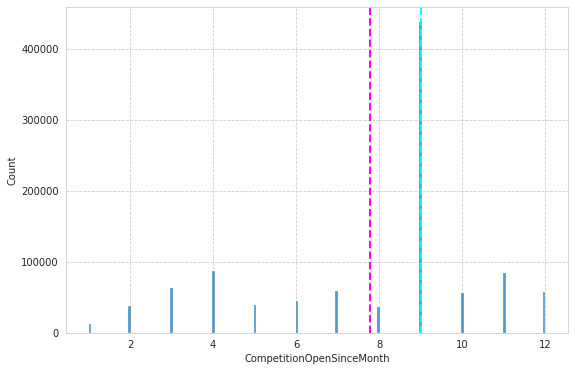
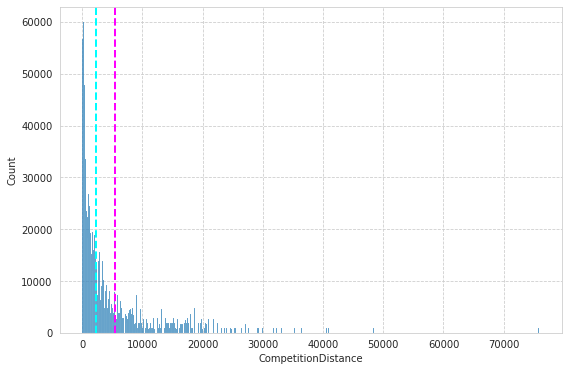
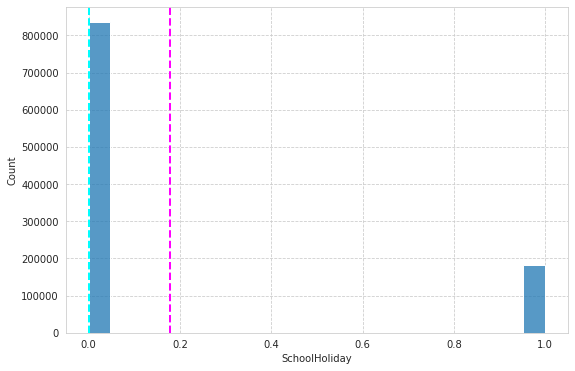
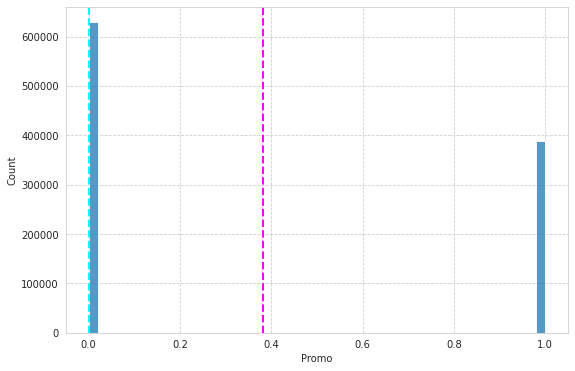
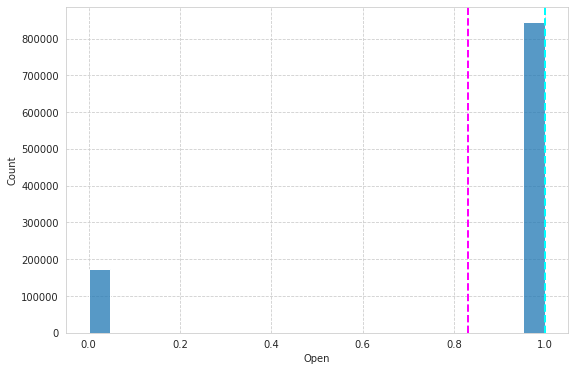
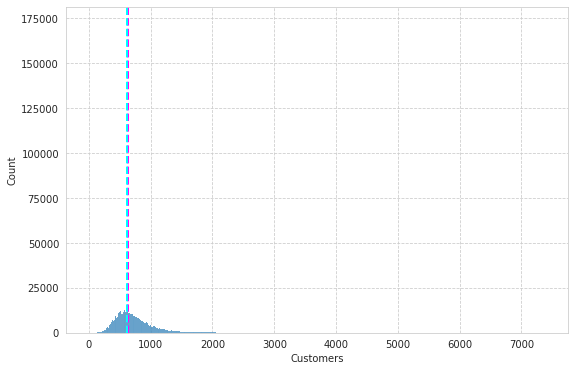
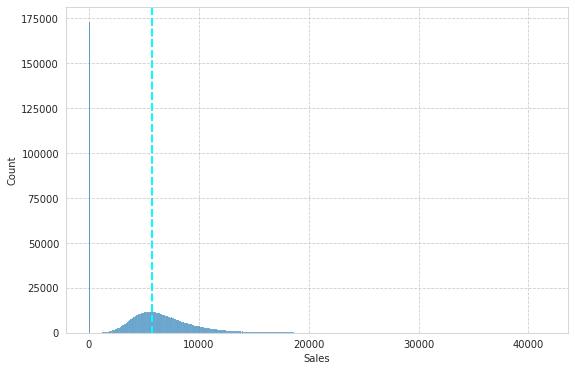
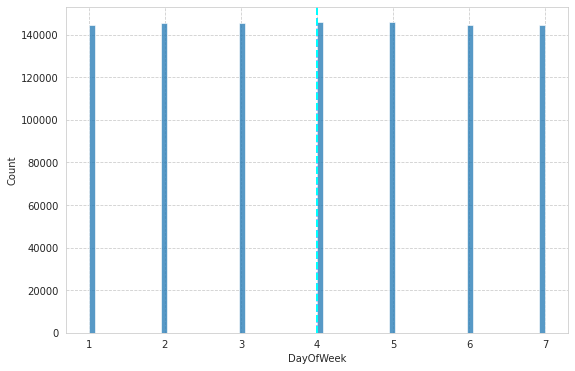
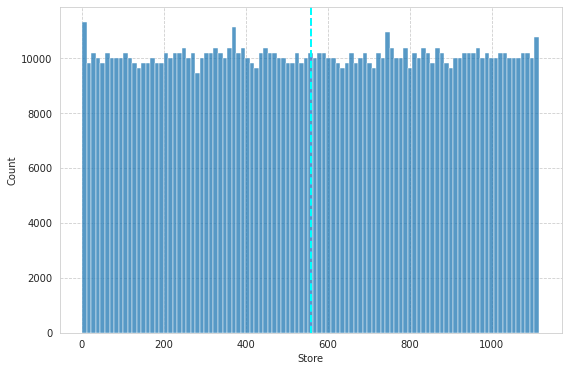
**Categorical Insights:**

**ploting Box plot to visualize and trying to get information from plot**

****

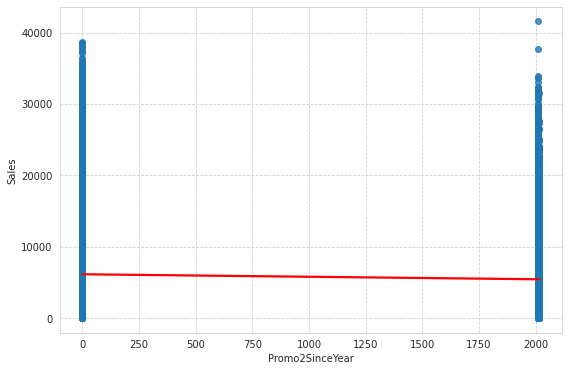
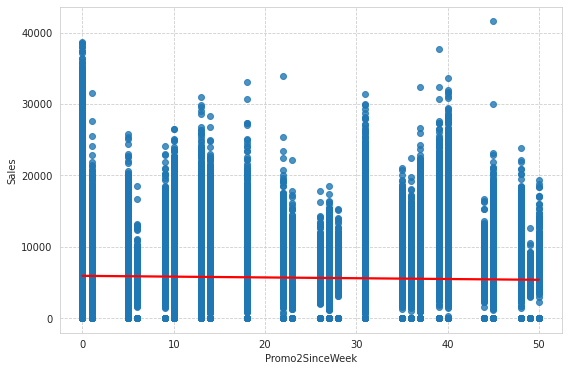
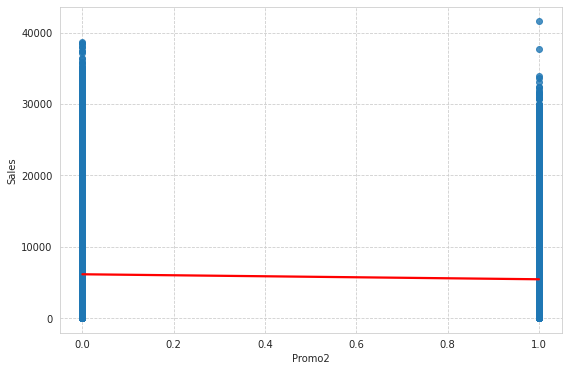
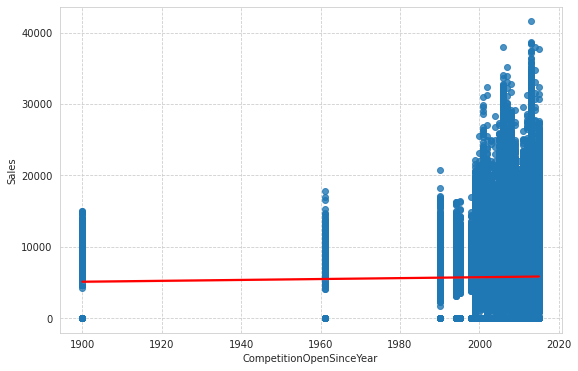
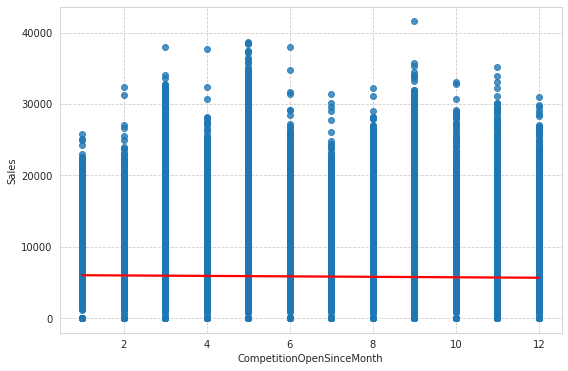
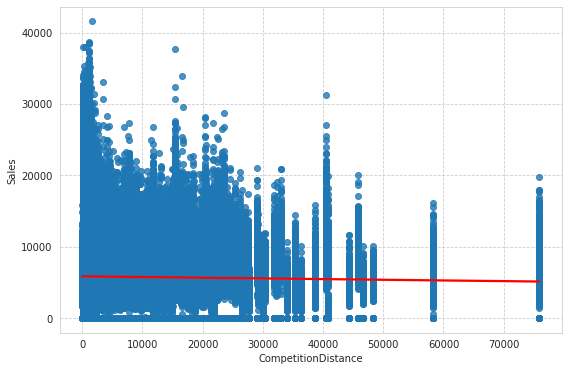
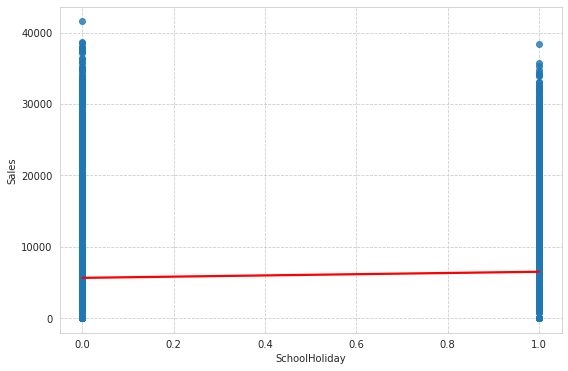
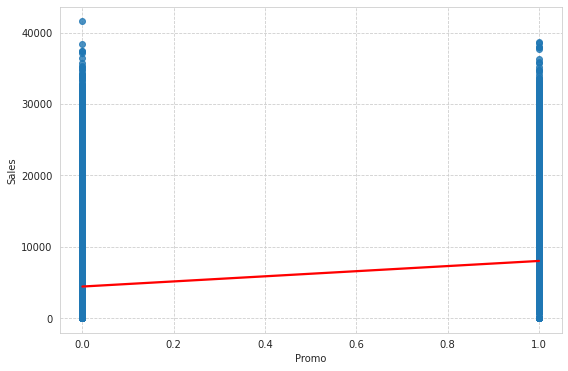
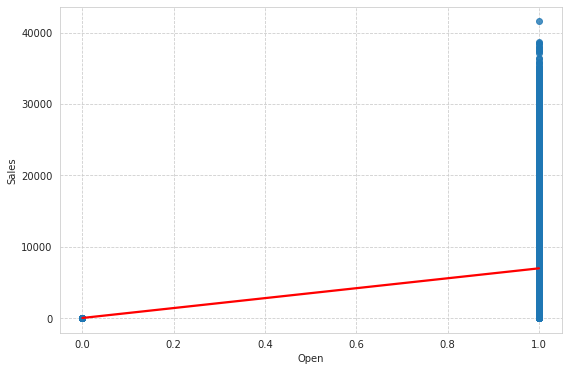
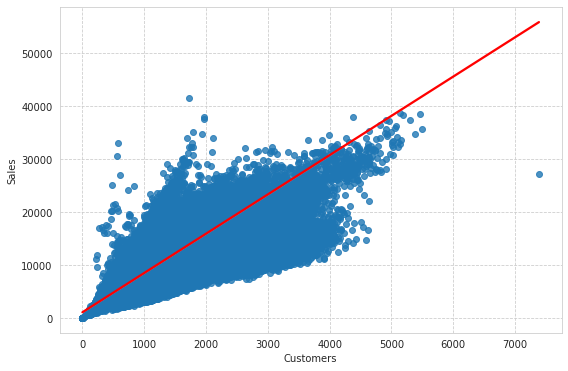
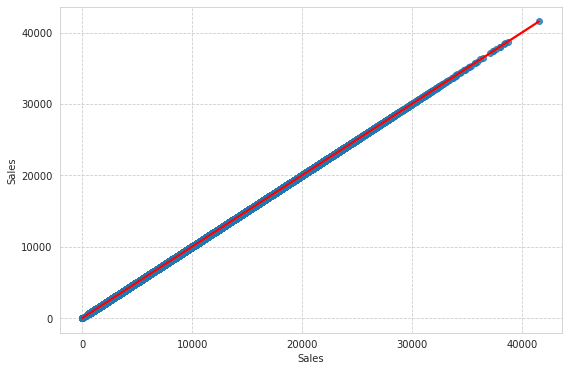
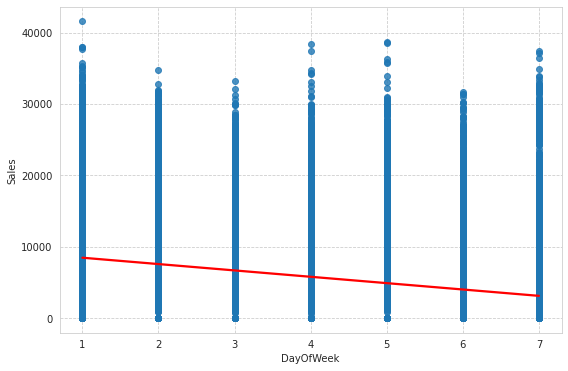
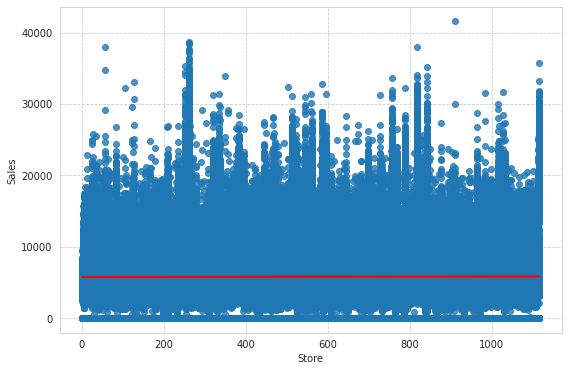
****

**#plotting histogram of dataset of numerical variables:**

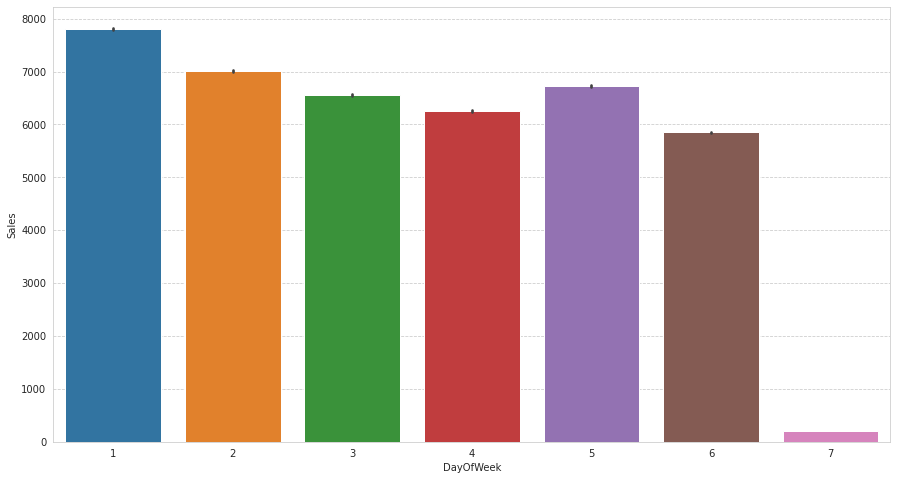
****

**# ploting Regression plot of each columns of dataset**

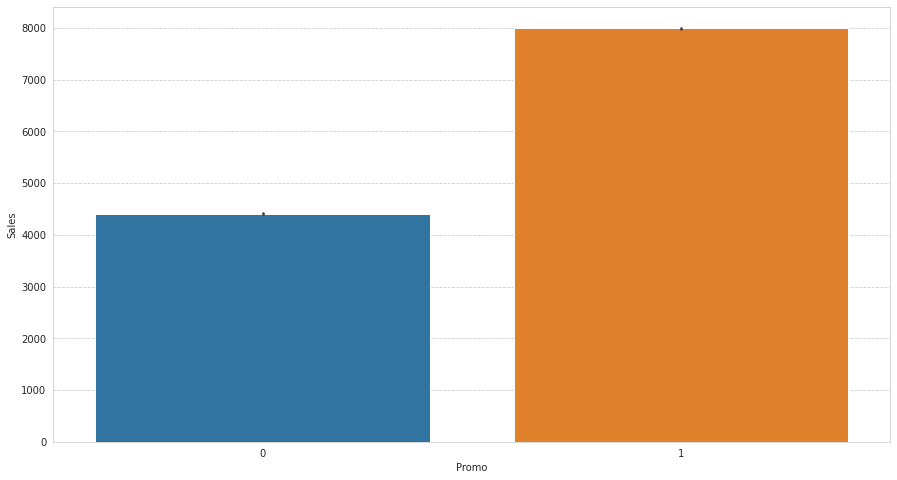
**v/s sales count columns**

****

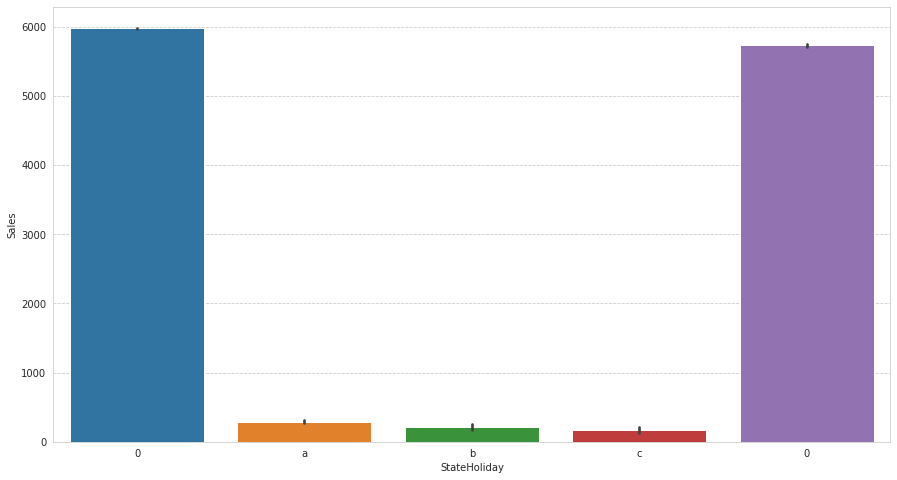
**barplots visualization of each columns of dataset v/s Sales count columns:**

****

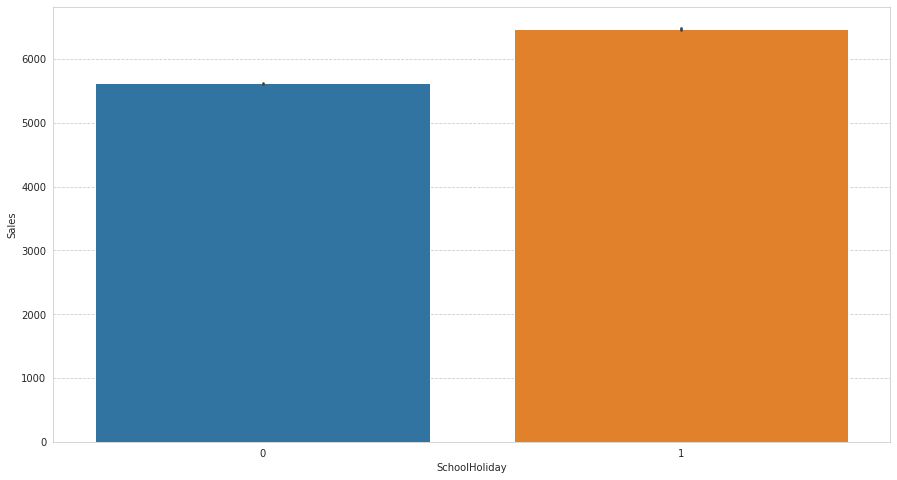
**Here it can be deduced that there were more sales on Monday, probably because shops generally remain closed on Sundays which had the lowest sales in a week. This validates the hypothesis about this feature.**

****

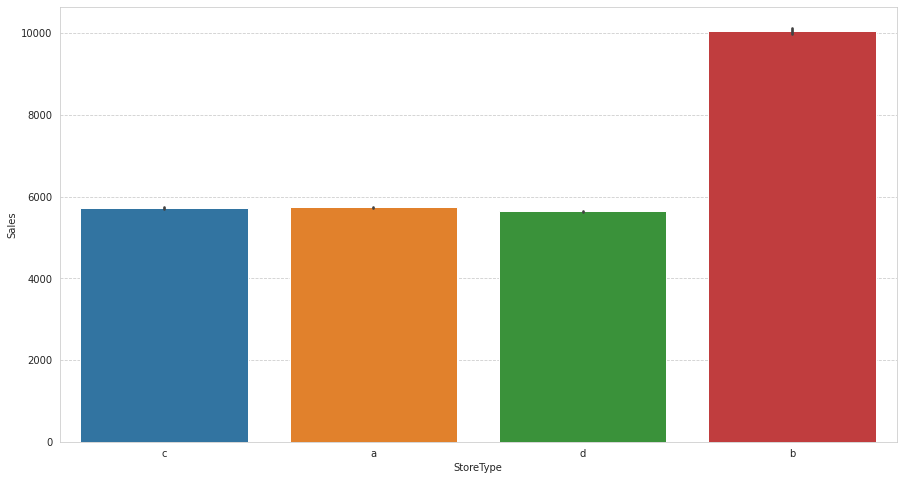
**Promotion has a positive effect on Sales indicating high sales for stores with Promo=1**

****

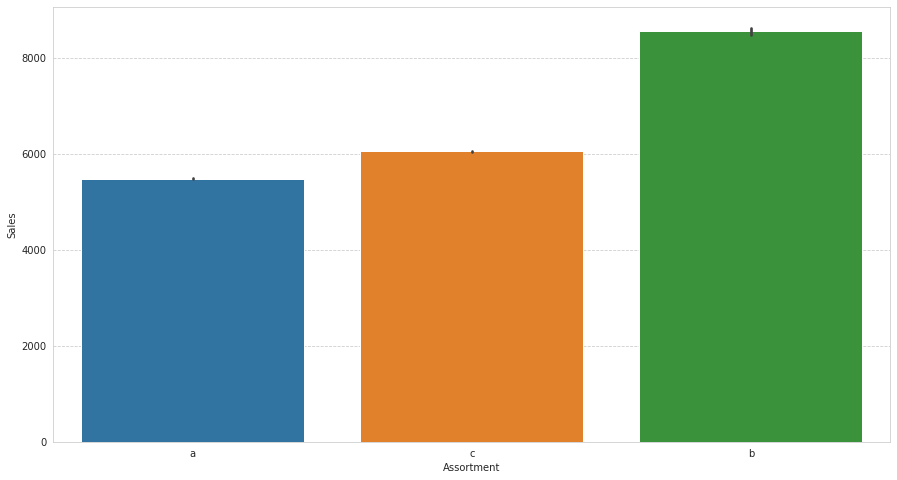
**Sales were low whenever there was a State Holiday indicating only a few stores were open on these days.** **Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None. Lowest of Sales were seen on state holidays especially on Christmas.**

****

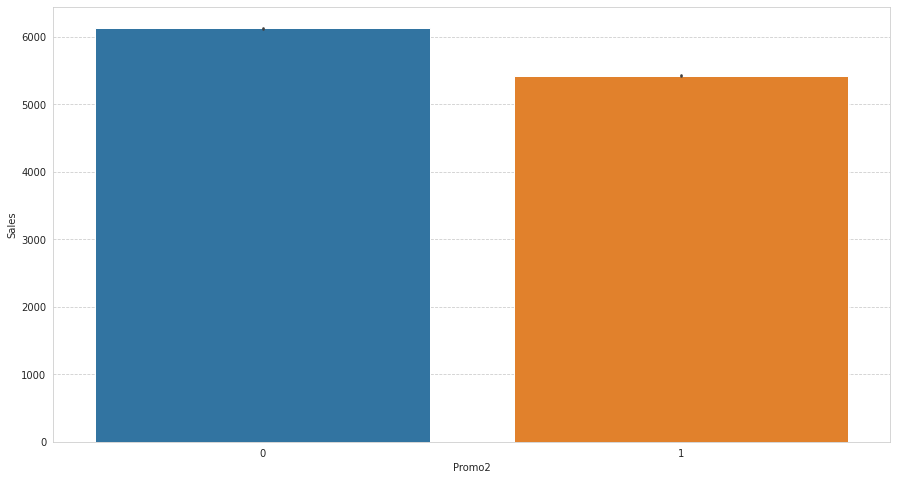
**Sales were high on an average on School Holidays indicating School Holidays weren’t compulsory by the law and comparatively more sales were recorded on holidays.**

****

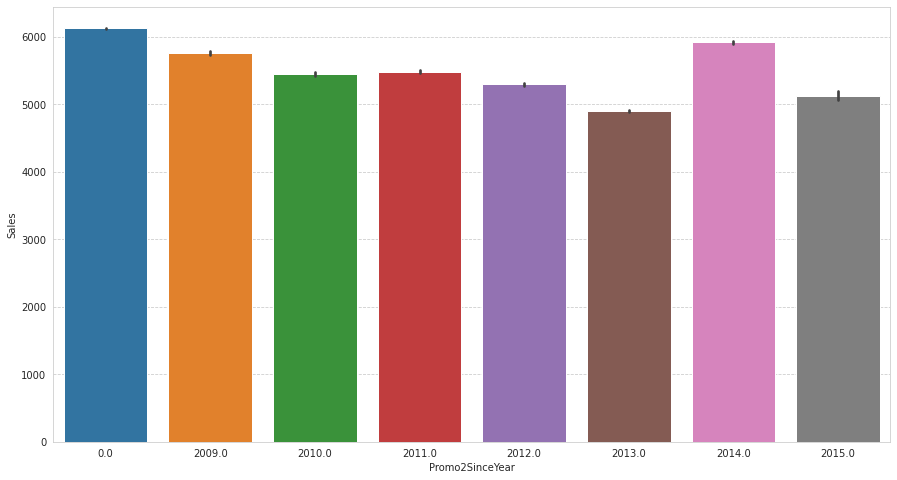
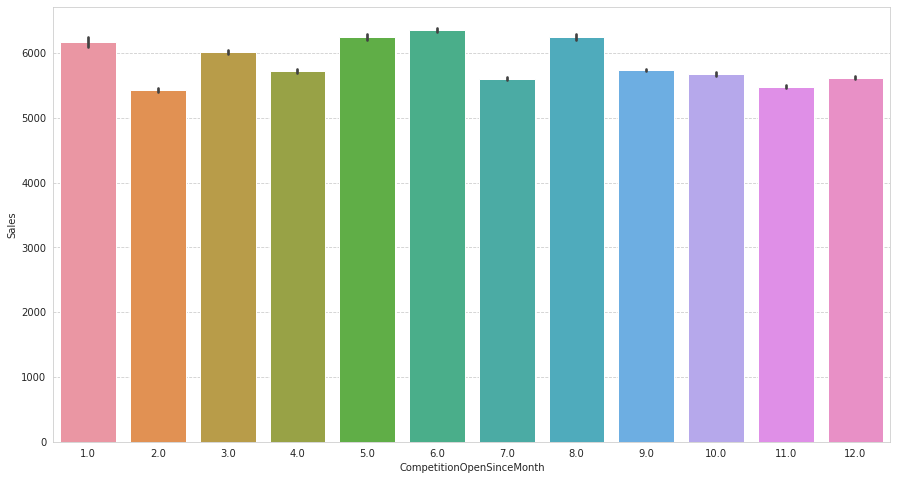
**A bar plot represents an estimate of central tendency for a numeric variable with the height of each rectangle. The store type b has the highest sales on an average.**

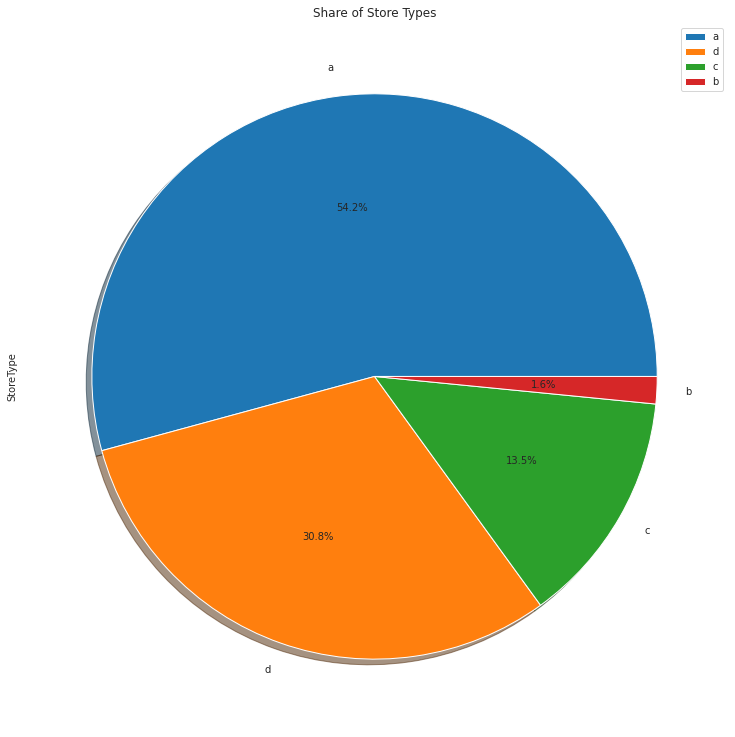
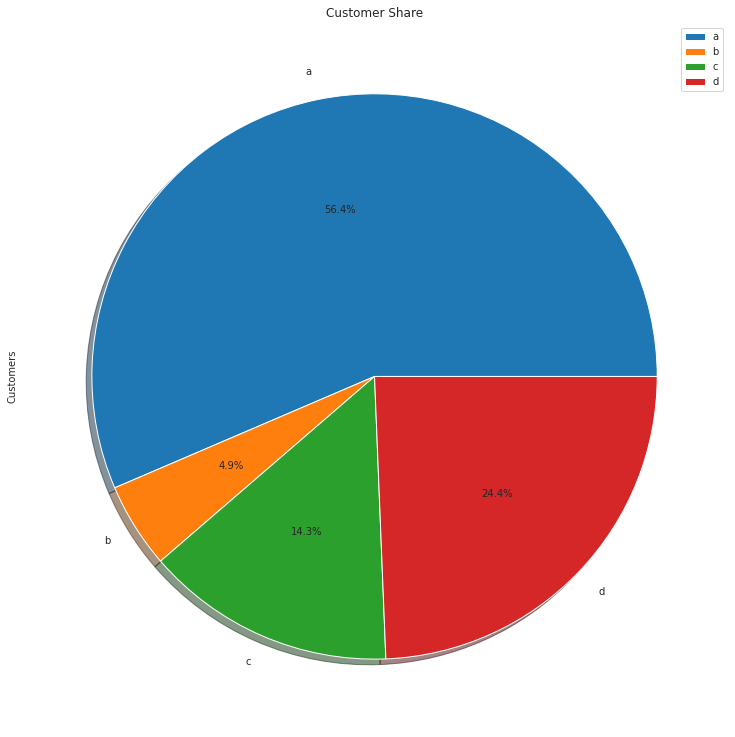
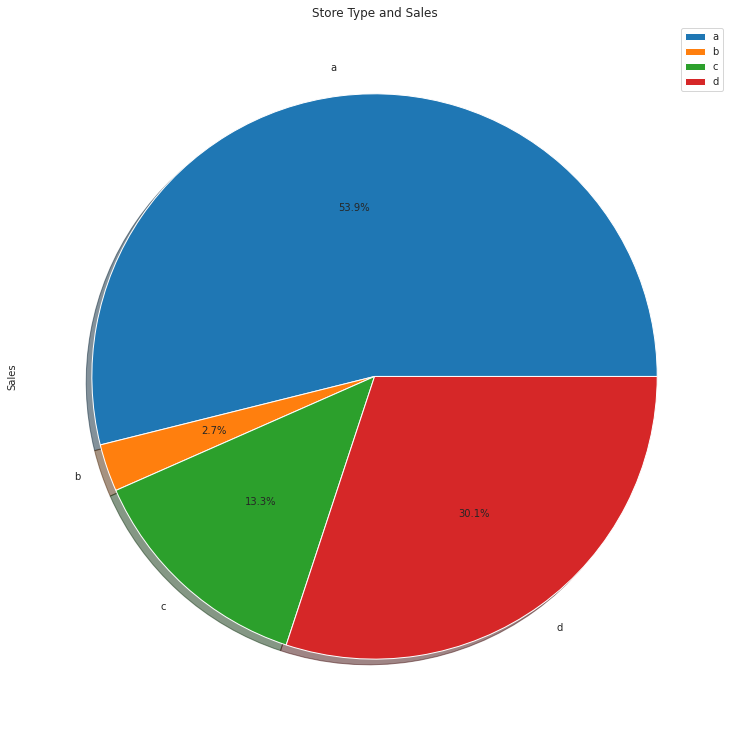
****

**Assortment type b has the highest sales on an average.**

****

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

****

****

**Observation:**

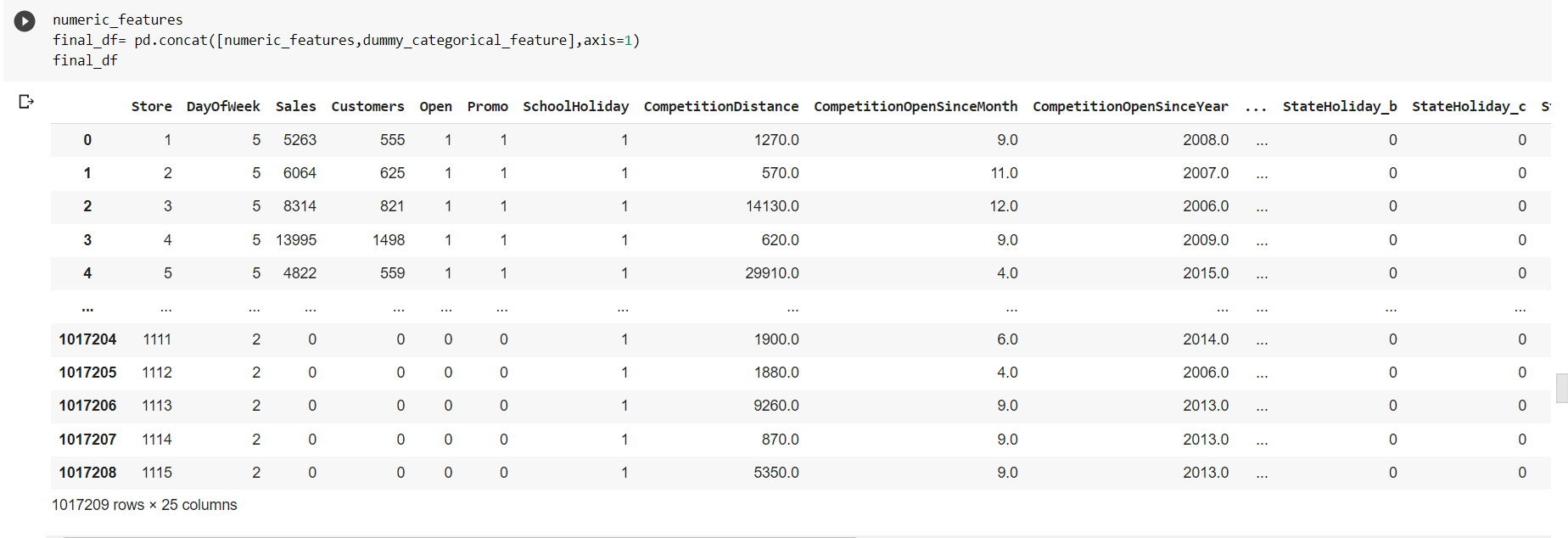
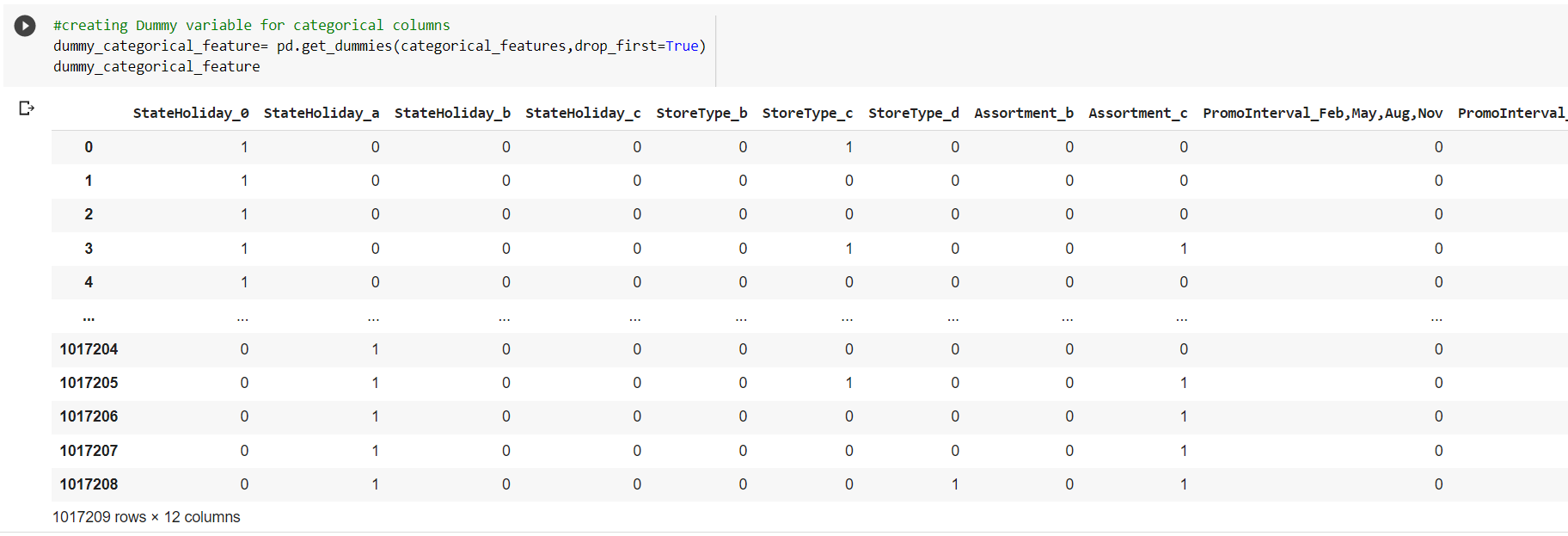
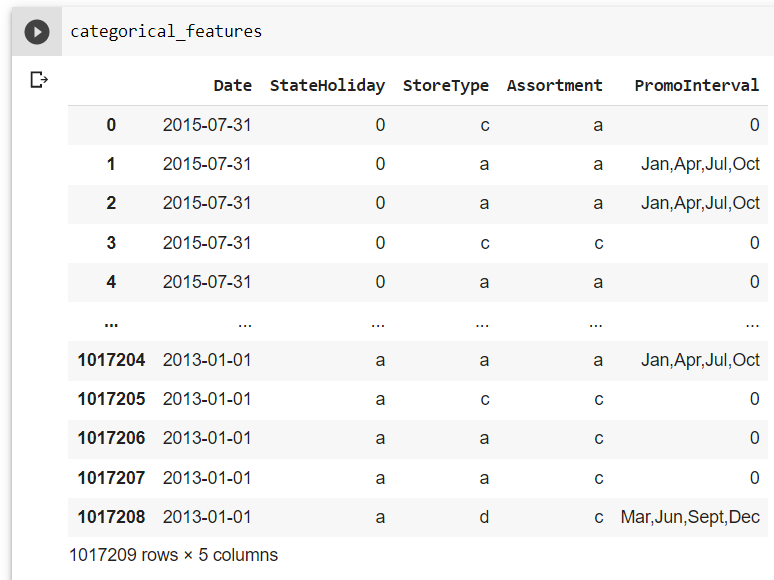
* A bar plot represents an estimate of central tendency for a numeric variable with the height of each rectangle. Earlier it was seen that the store type b had the highest sales on an average because the default estimation function to the barplot is mean.
* But upon further exploration it can be clearly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

**Feature engineering :**

Feature engineering refers to **manipulation — addition, deletion, combination, mutation — of your data set to improve machine learning model training, leading to better performance and greater accuracy**. Effective feature engineering is based on sound knowledge of the business problem and the available data sources.

**One hot encoding:**

one-hot encoding is **a frequently used method to deal with categorical data**. Because many machine learning models need their input variables to be numeric, categorical variables need to be transformed in the pre-processing part. Categorical data can be either nominal or ordinal.

****

# Ways to Detect and Remove the Outliers:

Outliers are those **data points that are significantly different from the rest of the dataset**. They are often abnormal observations that skew the data distribution, and arise due to inconsistent data entry, or erroneous observations.

*a****box plot****is a method for graphically depicting groups of numerical data through their quartiles. Box plots may also have****lines extending vertically****from the boxes (*whiskers*)****indicating******variability****outside the upper and lower quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram.****Outliers****may be****plotted****as****individual****points.*

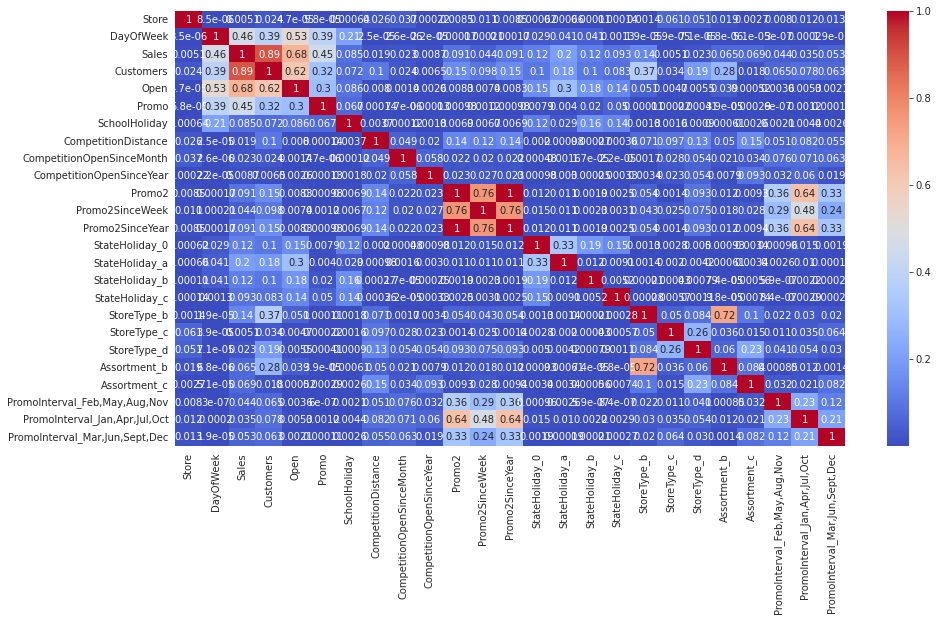
Above definition suggests, that if there is an outlier it will plotted as point in boxplot but other population will be grouped together and display as boxes. Let’s try and see it ourselves.

# 

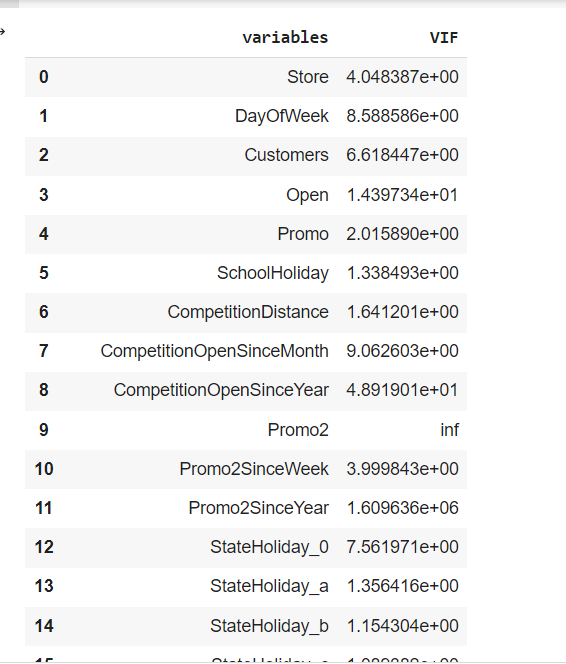
# 

# Correlation:

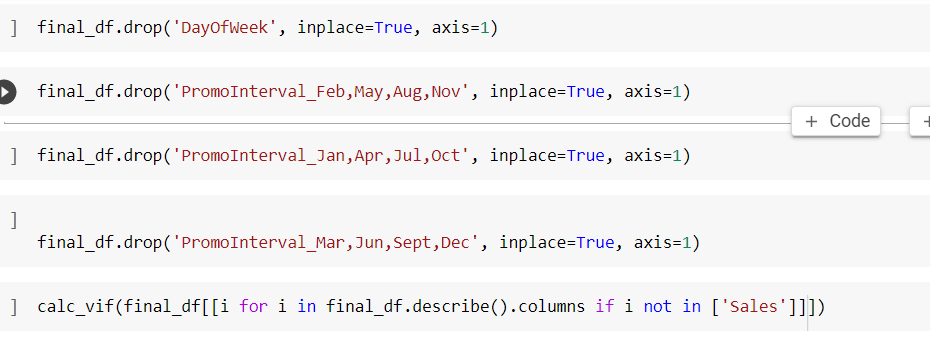
# Correlation is a statistical term used to measure the degree in which two variables move in relation to each other. A perfect positive correlation means that the correlation coefficient is exactly 1. This implies that as one variable moves, either up or down, the other moves in the same direction. A perfect negative correlation means that two variables move in opposite directions, while a zero correlation implies no linear relationship at all.

****

**Calculate VIF and remove drop columns VIF greater than 5**

****

**Drop columns**

****

**Modeling:**

**Factors affecting in choosing the model: Determining which algorithm to use depends on many factors like the problem statement and the kind of output you want, type and size of the data, the available computational time, number of features, and observations in the data, to name a few.**

**Train-Test Split:**

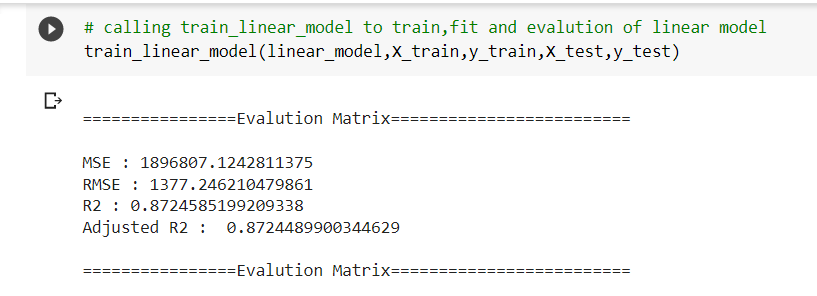
**In machine learning, train/test split splits the data randomly, as there’s no dependence from one observation to the other. That’s not the case with time series data. Here, it’s important to use values at the rear of the dataset for testing and everything else for training. The latest six weeks were kept as a testing set and the rest of the historical data was used in the training set.**

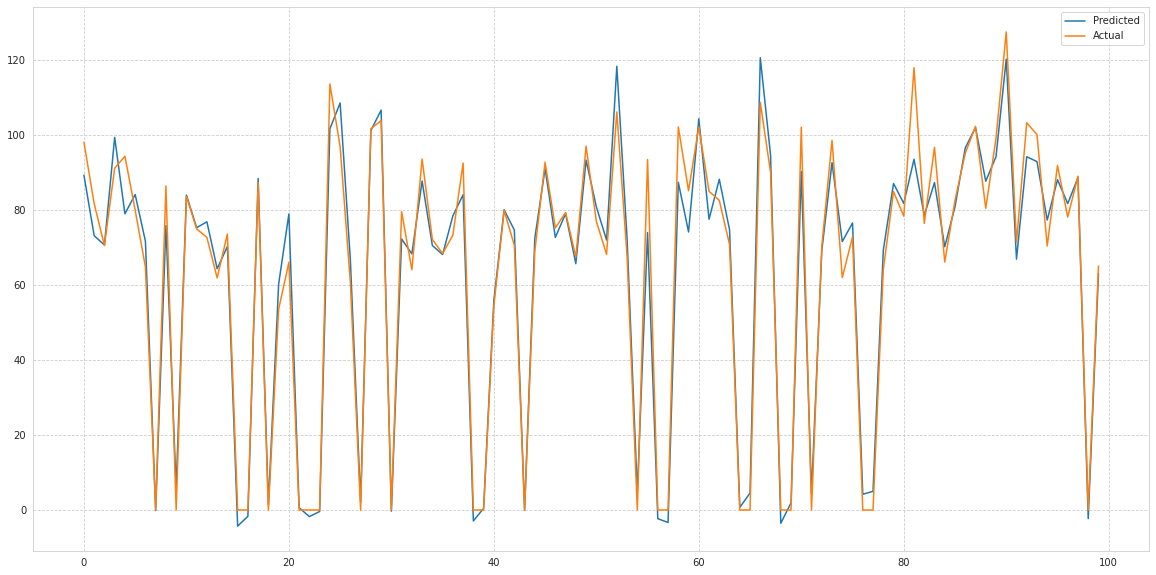
# Linear Regression :

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales**

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables.



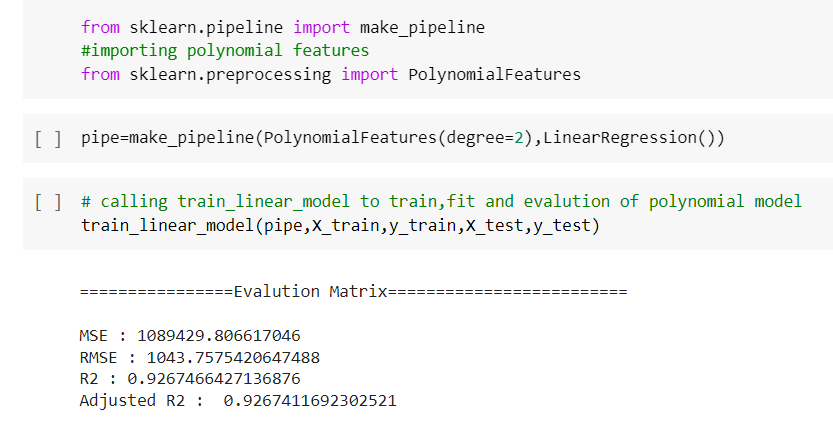


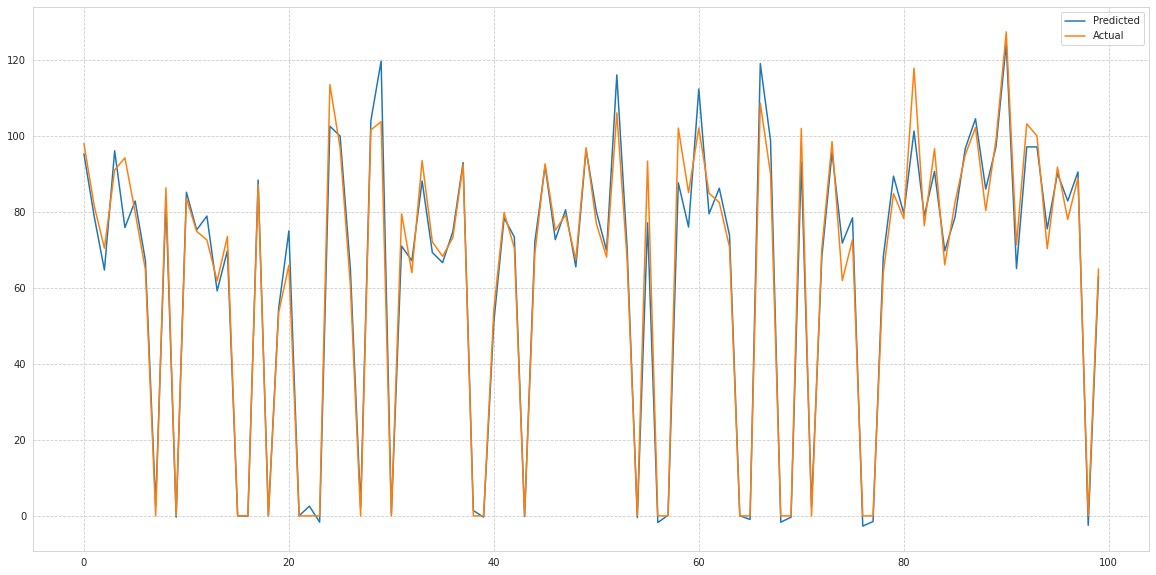
Polynomial Regression

* Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial. The Polynomial Regression equation is given below:

**y= b0+b1x1+ b2x12+ b2x13+...... bnx1n**

* It is also called the special case of Multiple Linear Regression in ML. Because we add some polynomial terms to the Multiple Linear regression equation to convert it into Polynomial Regression.
* It is a linear model with some modification in order to increase the accuracy.
* The dataset used in Polynomial regression for training is of non-linear nature.
* It makes use of a linear regression model to fit the complicated and non-linear functions and datasets.
* **Hence, *"In Polynomial regression, the original features are converted into Polynomial features of required degree (2,3,..,n) and then modeled using a linear model."***





## Regularization

Regularization is one of the most important concepts of machine learning. It is a technique to prevent the model from overfitting by adding extra information to it.

Sometimes the [machine learning](https://www.javatpoint.com/machine-learning) model performs well with the training data but does not perform well with the test data. It means the model is not able to predict the output when deals with unseen data by introducing noise in the output, and hence the model is called overfitted. This problem can be deal with the help of a regularization technique.

This technique can be used in such a way that it will allow to maintain all variables or features in the model by reducing the magnitude of the variables. Hence, it maintains accuracy as well as a generalization of the model.

It mainly regularizes or reduces the coefficient of features toward zero. In simple words, "In regularization technique, we reduce the magnitude of the features by keeping the same number of features."

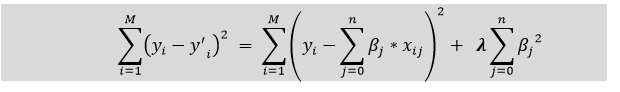
## Techniques of Regularization

There are mainly two types of regularization techniques, which are given below:

* **Ridge Regression**
* **Lasso Regression**

### Ridge Regression

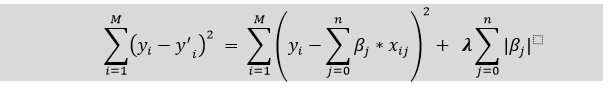
* Ridge regression is one of the types of linear regression in which a small amount of bias is introduced so that we can get better long-term predictions.
* Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called as **L2 regularization**.
* In this technique, the cost function is altered by adding the penalty term to it. The amount of bias added to the model is called **Ridge Regression penalty**. We can calculate it by multiplying with the lambda to the squared weight of each individual feature.
* The equation for the cost function in ridge regression will be:



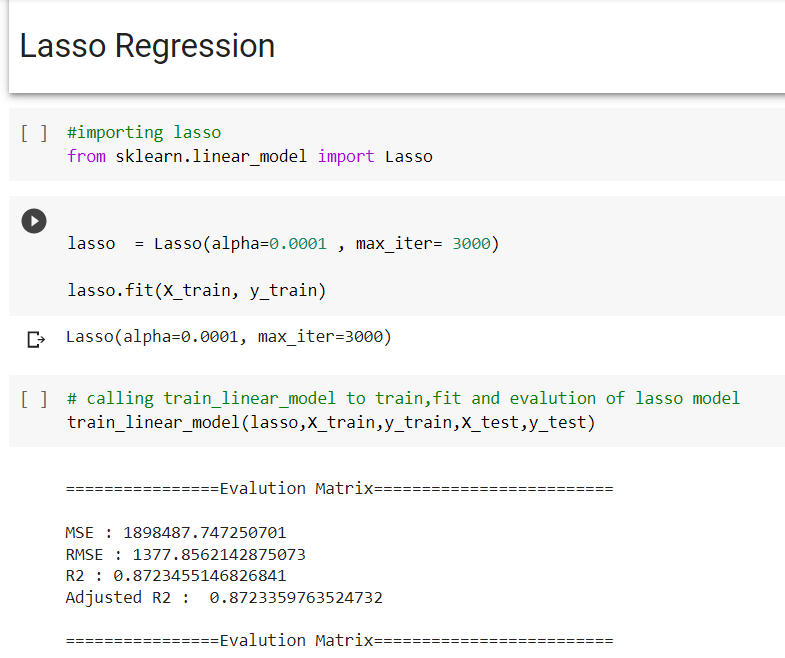
* In the above equation, the penalty term regularizes the coefficients of the model, and hence ridge regression reduces the amplitudes of the coefficients that decreases the complexity of the model.

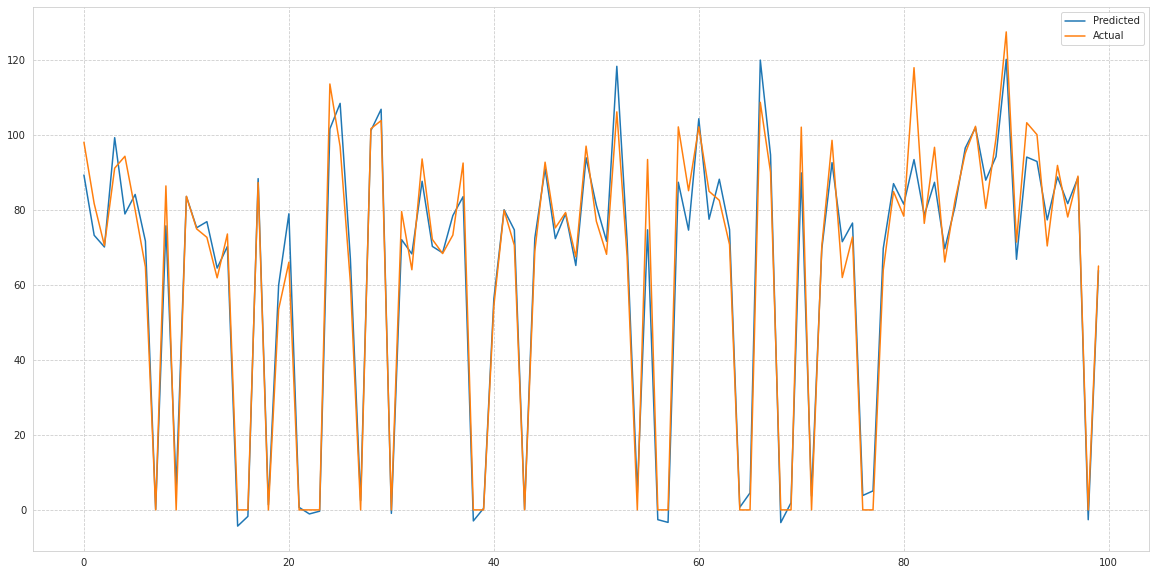
### Lasso Regression:

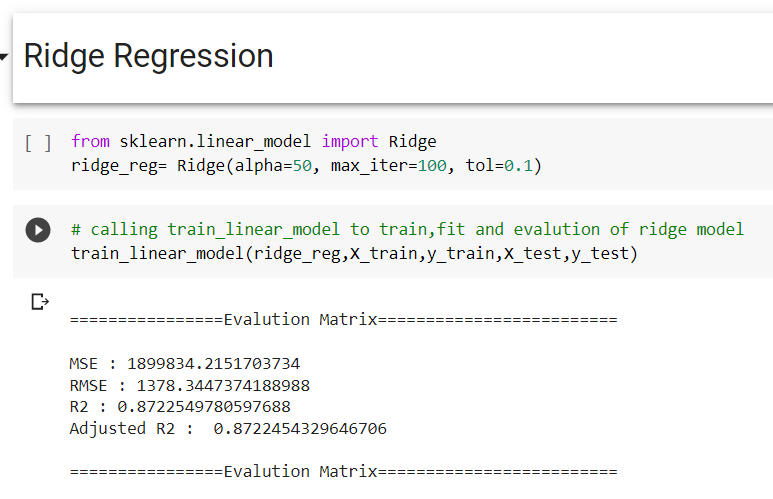
* Lasso regression is another regularization technique to reduce the complexity of the model. It stands for **Least Absolute and Selection Operator.**
* It is similar to the Ridge Regression except that the penalty term contains only the absolute weights instead of a square of weights.
* Since it takes absolute values, hence, it can shrink the slope to 0, whereas Ridge Regression can only shrink it near to 0.
* It is also called as **L1 regularization.** The equation for the cost function of Lasso regression will be:

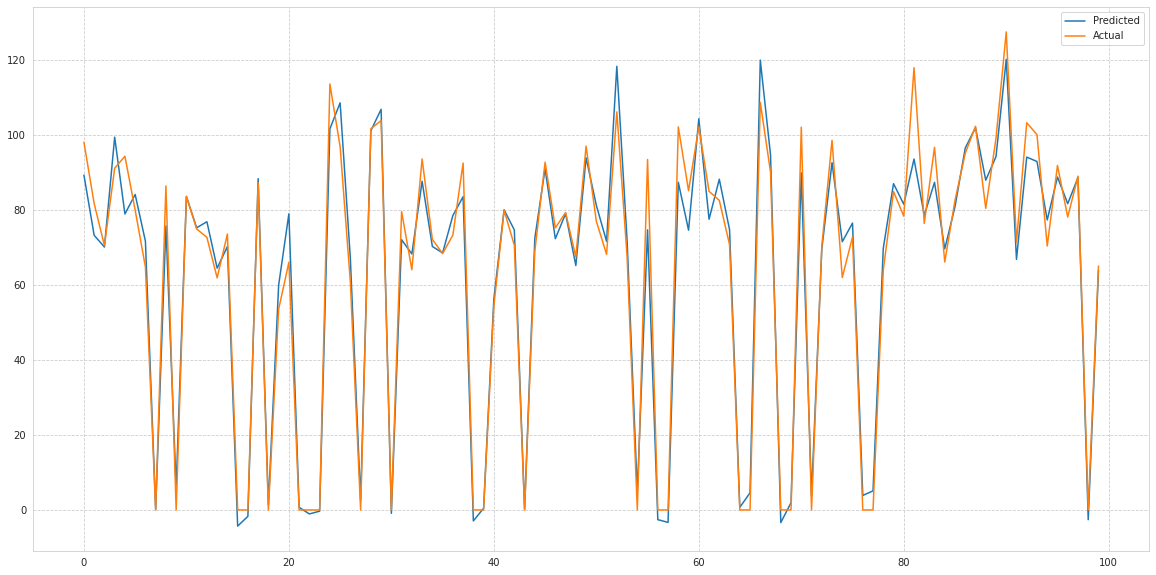


* Some of the features in this technique are completely neglected for model evaluation.
* Hence, the Lasso regression can help us to reduce the overfitting in the model as well as the feature selection.



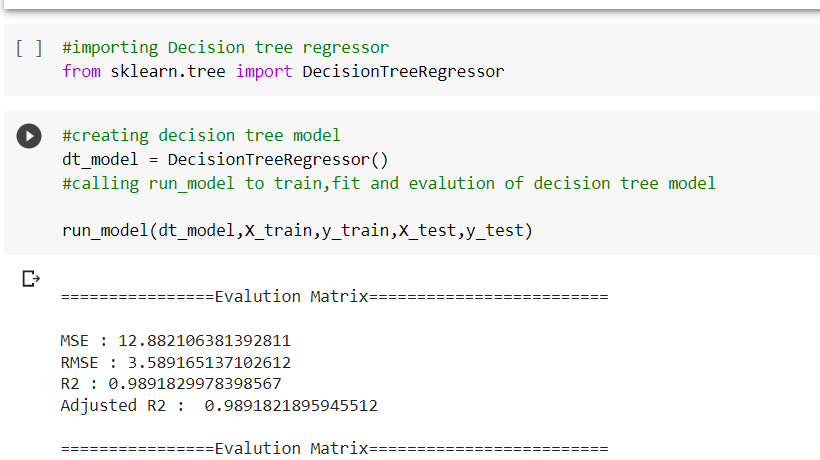


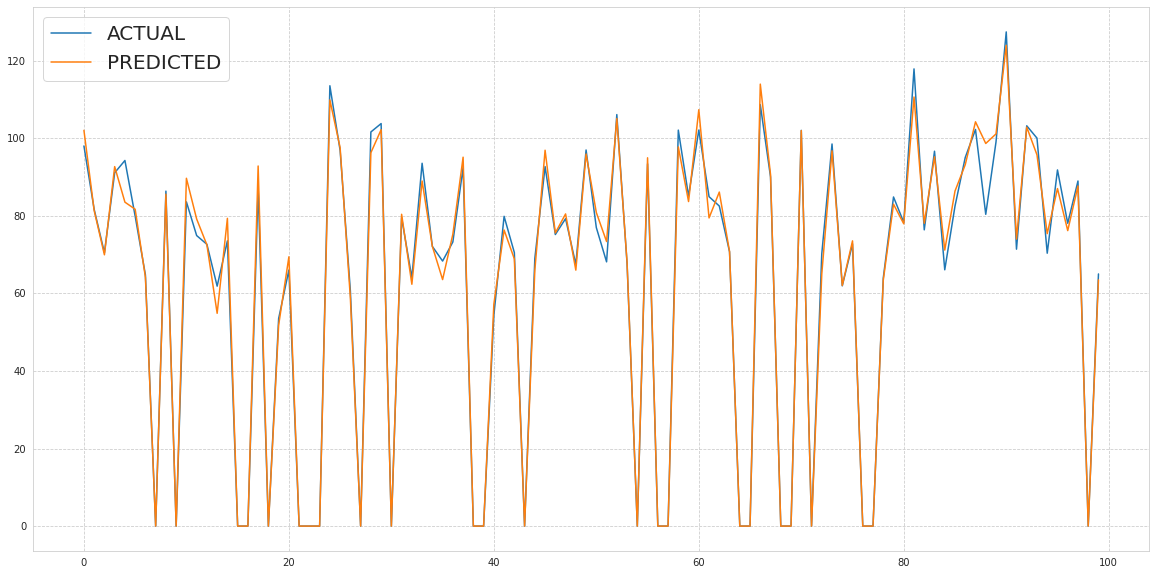


****

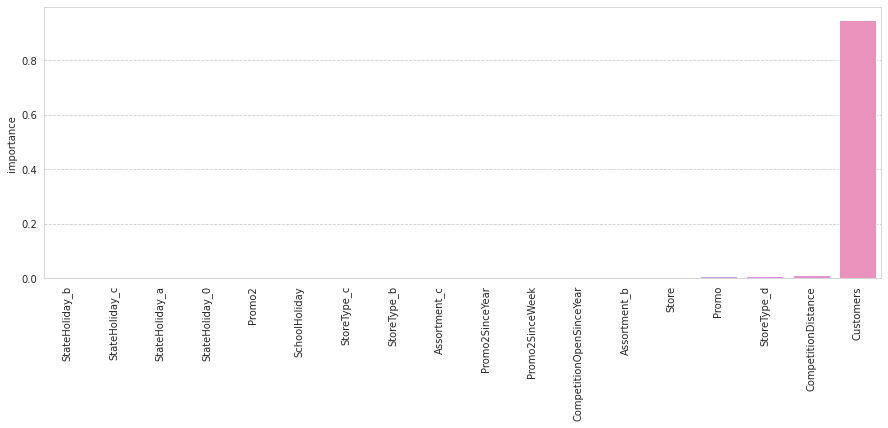
# Decision Tree

Decision Tree is a supervised learning method used in data mining for classification and regression methods. It is a tree that helps us in decision-making purposes. The decision tree creates classification or regression models as a tree structure. It separates a data set into smaller subsets, and at the same time, the decision tree is steadily developed. The final tree is a tree with the decision nodes and leaf nodes. A decision node has at least two branches. The leaf nodes show a classification or decision. We can't accomplish more split on leaf nodes-The uppermost decision node in a tree that relates to the best predictor called the root node. Decision trees can deal with both categorical and numerical data.



****

**features importance of decision tree**

****

**Evaluation Metrics:**

# ● ****Mean Absolute Error****

Mean Absolute Error is the average of the difference between the Original Values and the Predicted Values. It gives us the measure of how far the predictions were from the actual output. However, they don’t gives us any idea of the direction of the error i.e. whether we are under predicting the data or over predicting the data. Mathematically, it is represented as :

https://miro.medium.com/max/379/1*qak4Dadzs7pO0hnz4q8O8Q.gif

# ● ****Mean Squared Error****

Mean Squared Error(MSE) is quite similar to Mean Absolute Error, the only difference being that MSE takes the average of the **square**of the difference between the original values and the predicted values. The advantage of MSE being that it is easier to compute the gradient, whereas Mean Absolute Error requires complicated linear programming tools to compute the gradient. As, we take square of the error, the effect of larger errors become more pronounced then smaller error, hence the model can now focus more on the larger errors.

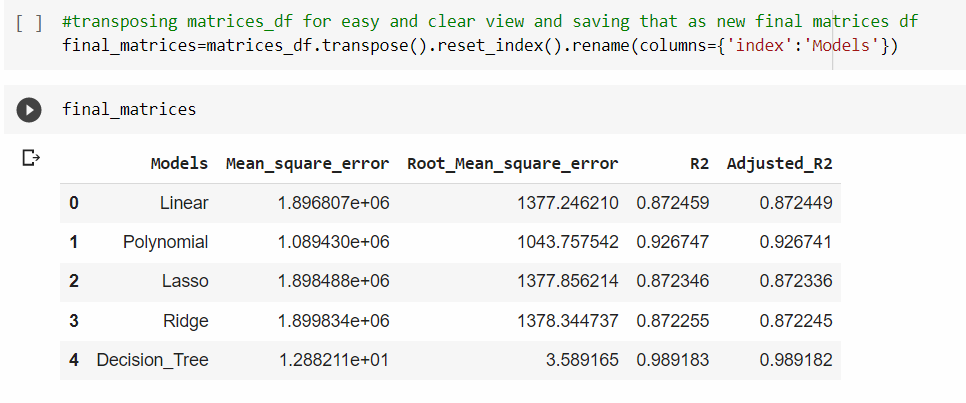
https://miro.medium.com/max/390/1*okvAVQNY6s5cMHxrqUzM5A.gif

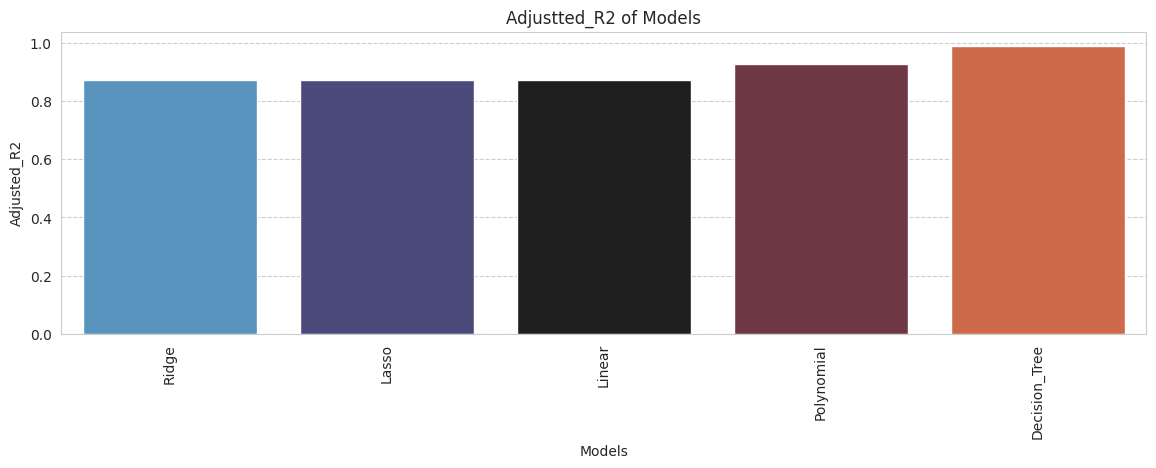
## ● Root Mean Squared Error or RMSE

**RMSE is the standard deviation of the errors which occur when a prediction is made on a dataset. This is the same as MSE (Mean Squared Error) but the root of the value is considered while determining the accuracy of the model.**

**● R Squared (R^2)- R2 score is a metric that tells the performance of your model, not the loss in an absolute sense that how well did your model perform. Hence, R2 squared is also known as Coefficient of Determination or sometimes also known as Goodness of fit. It’s value ranges from 0 to 1. It can be negative if the model is performing worse than the base.**

**● Adjusted R Squared- The disadvantage of the R2 score is while adding new features in data the R2 score starts increasing or remains constant but it never decreases because It assumes that while adding more data variance of data increases. Adjusted R^2 is adjusted for this disadvantage and shows the real value.**

****

****

# Final Conclusion:

**It is quite evident from the results that polynomial and Desission Tree is the best model that can be used for the Rossmann Sales Prediction since the performance metrics (mse,rmse) shows lower and (r2,adjusted\_r2) show a higher value for the polynomial and Desission Tree**