**Capstone Project**

**Retail Sales Prediction**

**Shajad**

**Contents:**

* **Abstract**
* **Introduction**
* **Problem Statement**
* **Steps Involved**
* **Data Visualization**
* **Methods and algorithm that we used**
* **Conclusion**
* **References**

**Abstract:** 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.

**Data Description**

Rossmann Stores Data.csv - historical data including Sales store.csv - supplemental information about the stores DatafieldsMost of the fields are self-explanatory.

* Id - an Id that represents a (Store, Date) duple within the set
* Store - a unique Id for each store
* Sales - the turnover for any given day (Dependent Variable)
* 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. An assortment strategy in retailing involves the number and type of products that stores display for purchase by consumers.
* 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

**Introduction:**

we are provided with historical sales data for 1,115 Rossmann stores. The task is to forecast the "Sales" column for the test set. In this data some stores were temporarily closed for refurbishment.

**Problem Statement:**

This data set contains information of stores that information includes such as Store, Sales, Customer, Open, State holiday, School holiday, Store types. All personally identifying information has been removed from the data.

Explore and analyze the data to discover important factors that govern the .

We will tackle the problem statement in the following steps:

Step 1: Data Overview.

Step 2:Data Preparation and Cleaning.

Step 3: Visualising the Univariate and Bivariate features.

Step 4: Correlation Analysis.

Step 5: Concluding Analysis.

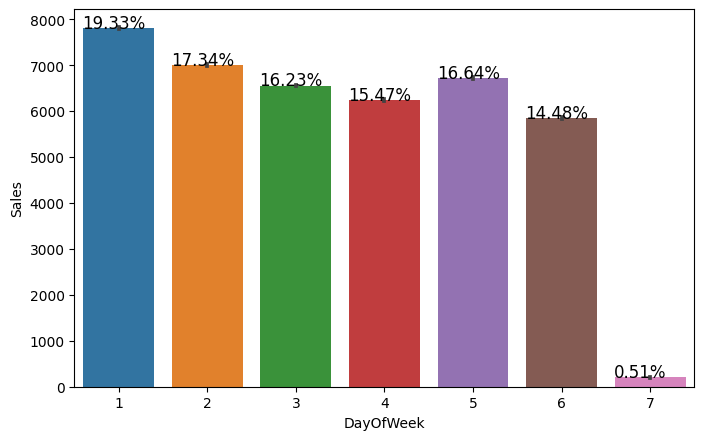
**Steps Involved:**

* **Importing Packages:** Importing the various libraries that will help us analyze our dataset properly with visual graphs.
* **Data Overview :** We load the data and go over the basic features, shape and datatypes of various variables.
* **Data Preparation and Cleaning:** We use various features of python to create combined features of date and time and other variables which can be simplified. We also drop columns and rows which have a lot of null values. Finally we take care of outliers.
* **Univariate and Bivariate Analysis:** We use seaborn and matplotlib to plot graphs starting with one variable graphs and then plotting two or more variable graphs to understand the variables and it spread and range with its frequency
* **Concluding Analysis:** Here after plotting and analyzing all the data we finally make predictions and remarks about our analyzes.

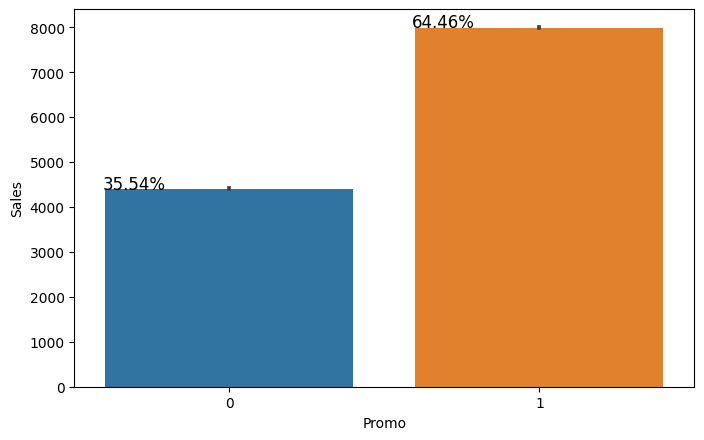
**Data Visualization:**

Lets understand various relation among target and other variables

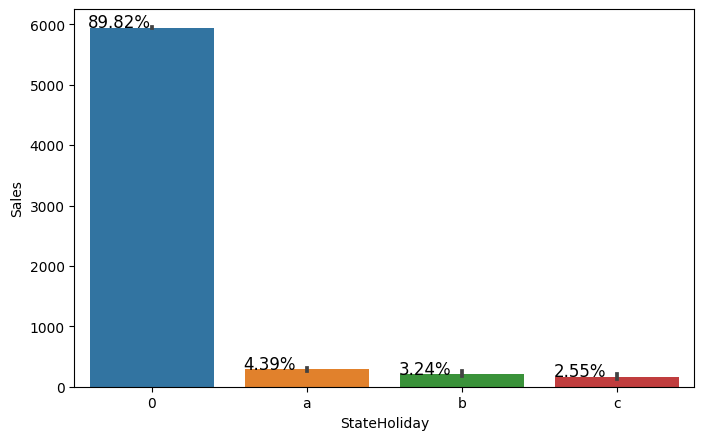
7th day of week the graph show some sales it means that there are some stores are open on sunday



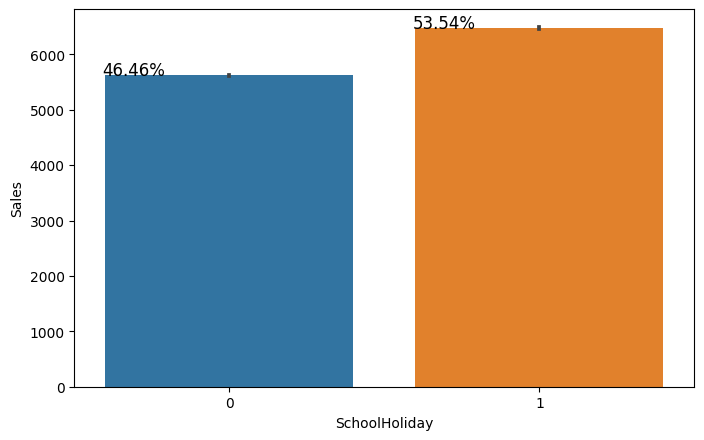
Promo leads to increase in sales of the stores



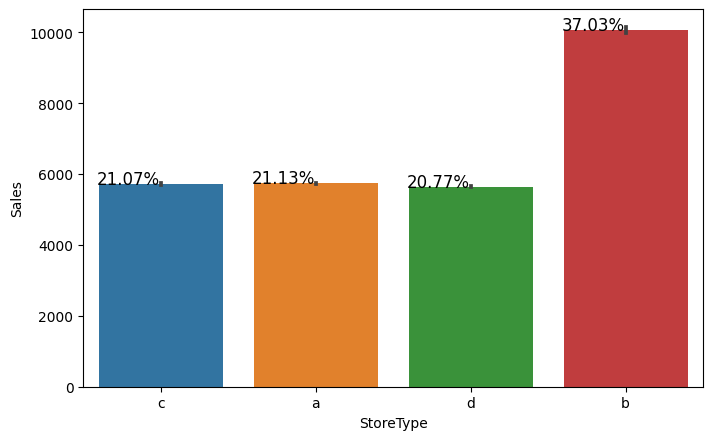
In this graph we show that during the state holiday the sales is at its minimum or near to zero



During the School holiday the sales remains the same



B type of store have dubble sale then other store



**Conclusion:**

* The positive relation between the promo and the sales
* in 7th days of the week the sale in likely minimum because of most of the stores are close to Sunday
* the sale in 1st day of the week in sale is likely maximum sales of the week because the store are closed on Sunday
* sale and open have the direct relationship between the open and the sale
* promo is leads the sale and make the sale approximately double
* on state holiday sales become approximately 0 but during the school holiday the sales remains the same
* B type store have more sales then other and Assortment B have the more sales
* In every month sale is always more then 5000

**References:**

* GeeksforGeeks
* Stackoverflow
* Almabetter
* Youtube
* Github