# CAPSTONE PROJECT

## RETAIL SALES PREDICTION


TEAM MEMBERS

Shivansh Yadav

Rishabh Kumar

Madhur Awasthi

(Data Science Trainees @ Almabetter)

# RETAIL SALES PREDICTION

- Introduction
- Data Pipeline
- Data Summary
- Data Exploration
- Merging of data
- Visualization and Analysis
- Feature Engineering
- Regression Modeling and Model selection
- Technologies used

# Introduction

- Retail sales contributes to the economy in large scale and provides the need of people. Many well known Famous brand Retail stores have spread all over countries, Being part of same chain some stores earn more and some earn lesser revenue.
- Rossmann operates over 3,000 drug stores in 7 European countries. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality.
- We need to train a model to predict Future sales.

# Data Pipeline

- **EDA** : In this part, we do some exploratory data analysis (EDA) on the data to see how is the data.
- **Data cleaning** : In this part we have removed or cleaned unnecessary data. Since there were some data with null or unwanted values.
- **Data merging**: In this part after removing unnecessary data from both .csv files we merged the data.
- **Data preparation :** In this part we converted categorical values into numerical to make data ready for modeling.
- **Data visualization** : In this part we visualize and analyze the data from which we get the trend and relation between features which is useful for prediction.
- **Data modeling :** In this part we trained and predicted data in different regression models for getting final model.
- **Model Selection** : In this part we selected model which gives best result.
- **Model Explainibilty** : In this part we see features with their weightage.

# Data Summary

- 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
- State Holiday - This column 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 - This column indicates if the (Store, Date) was affected by the closure of public schools.

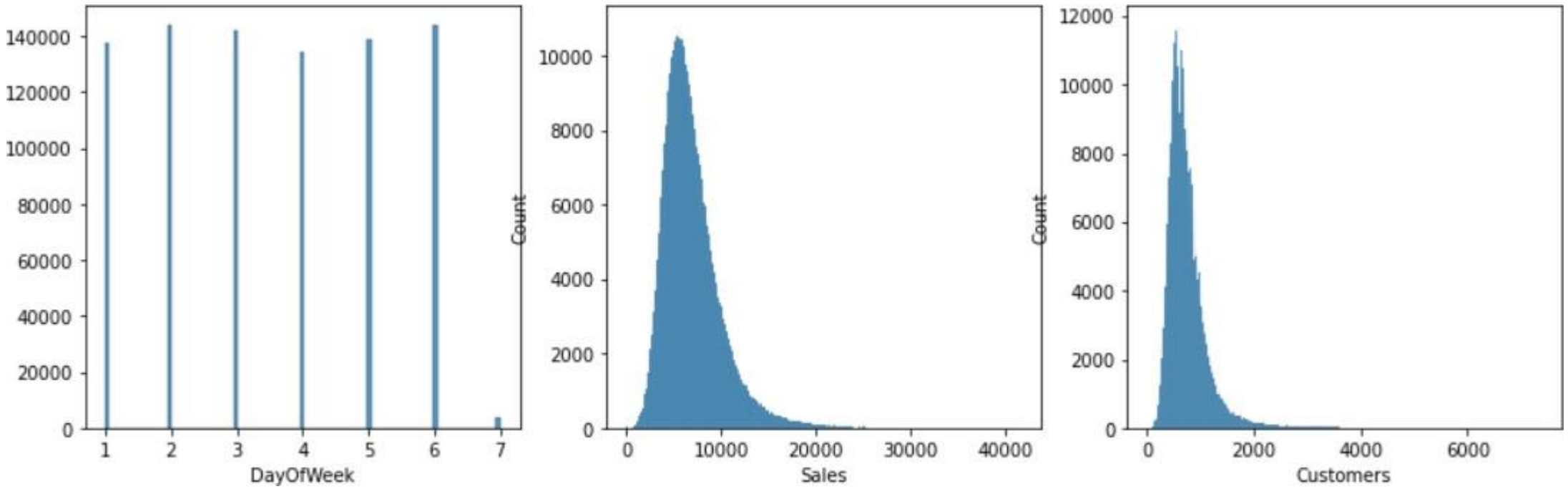
# Data Summary

- Store Type - This column differentiates between 4 different store models: a, b, c, d
- Assortment - This column describes an assortment level: a = basic, b = extra, c = extended.
- Competition Distance - This column distance in meters to the nearest competitor store
- Competition Open Since[Month/Year] - This column gives the approximate year and month of the time the nearest competitor was opened
- Promo - This column 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

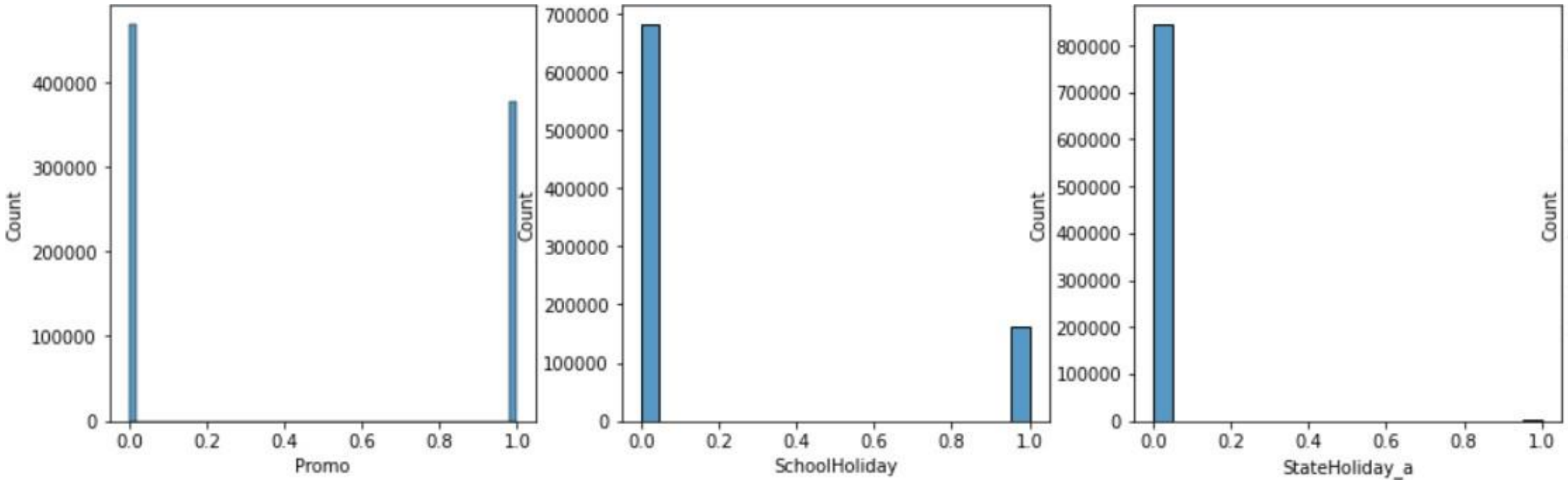
# Data Summary

- Promo2 Since[Year/Week] - This column describes the year and calendar week when the store started participating in Promo2
- Promo Interval - This column 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.

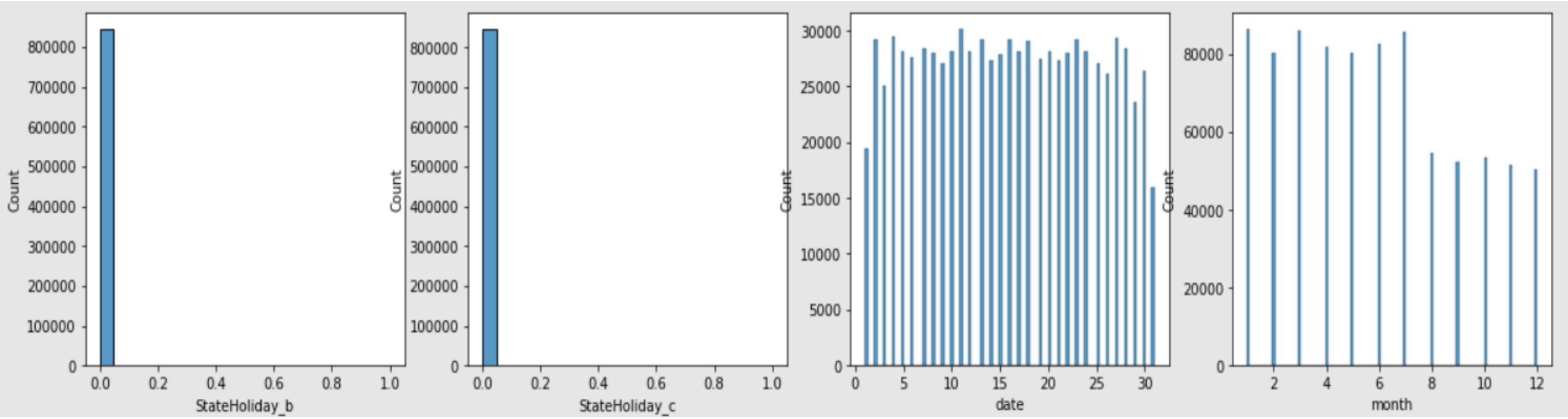
# Distribution of data for sales dataframe



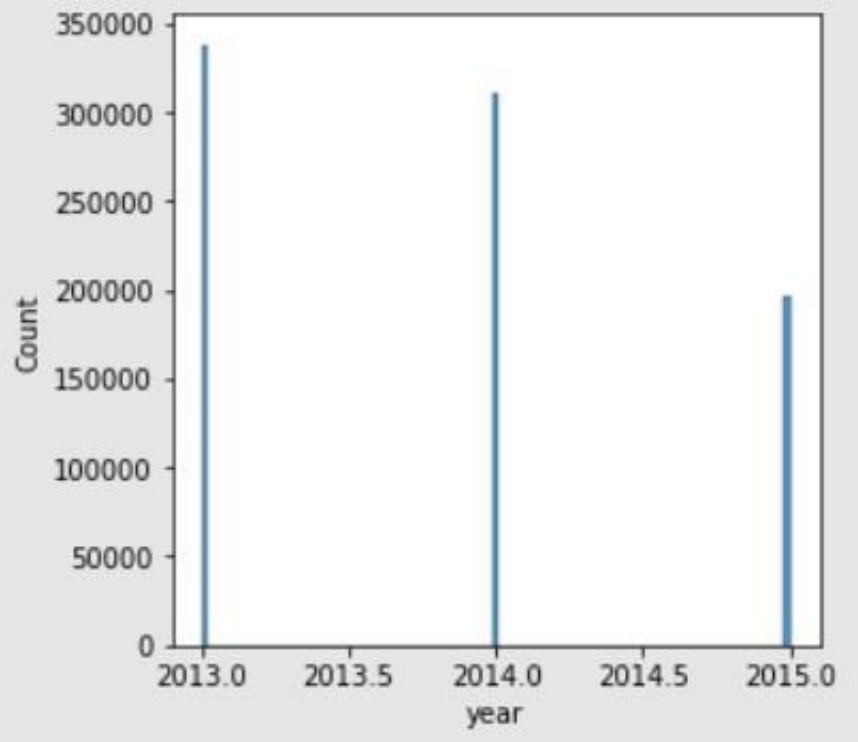
# Distribution of data for sales dataframe



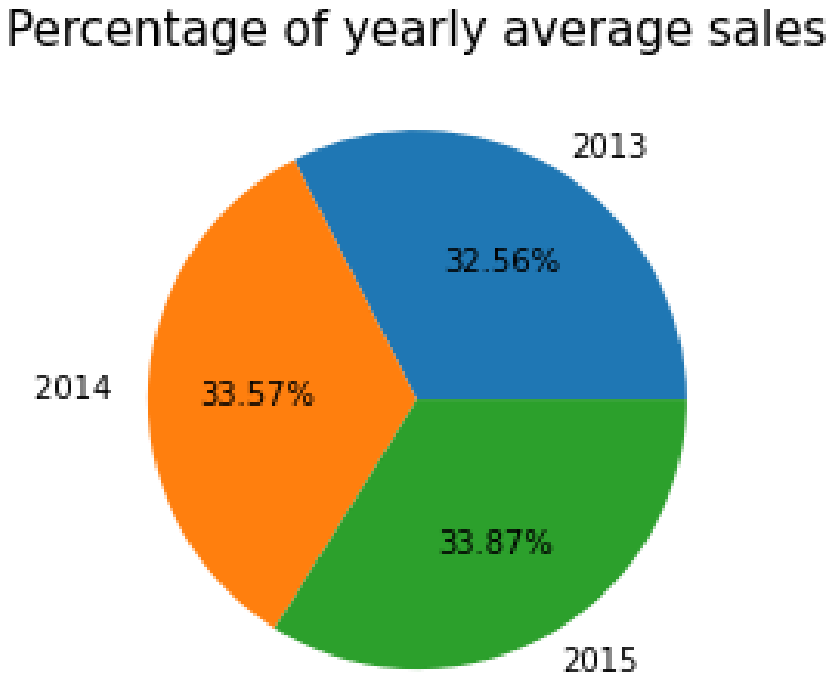
# Distribution of data for sales dataframe

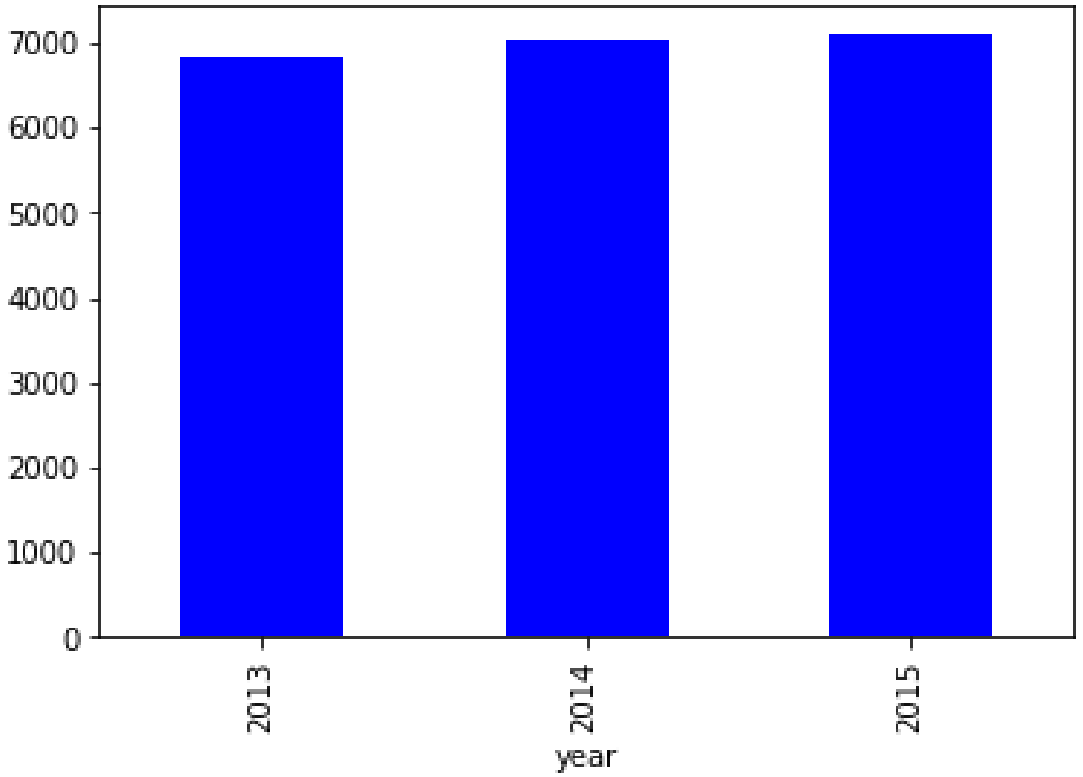


# Distribution of data for sales dataframe



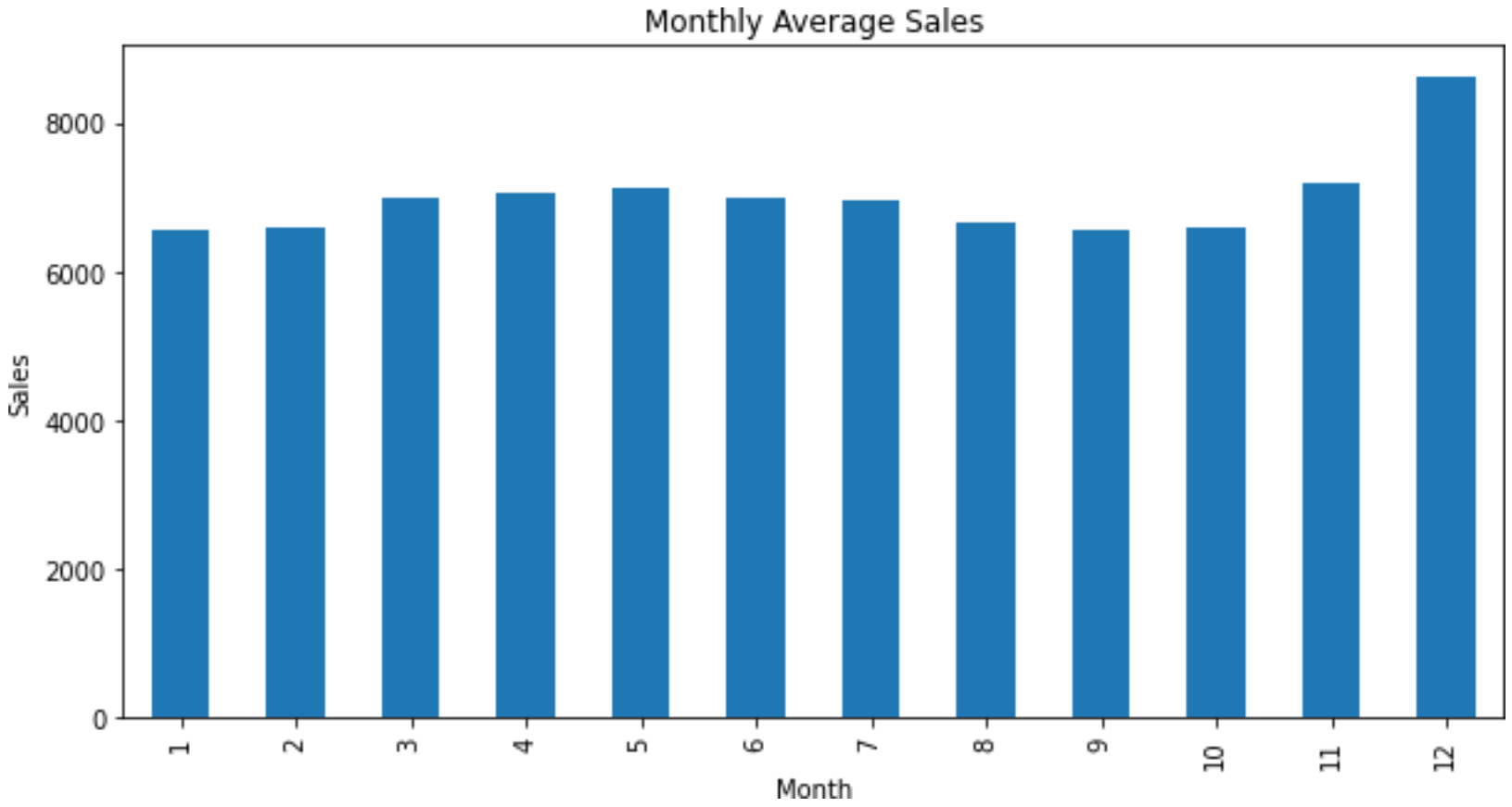
# Yearly average sales



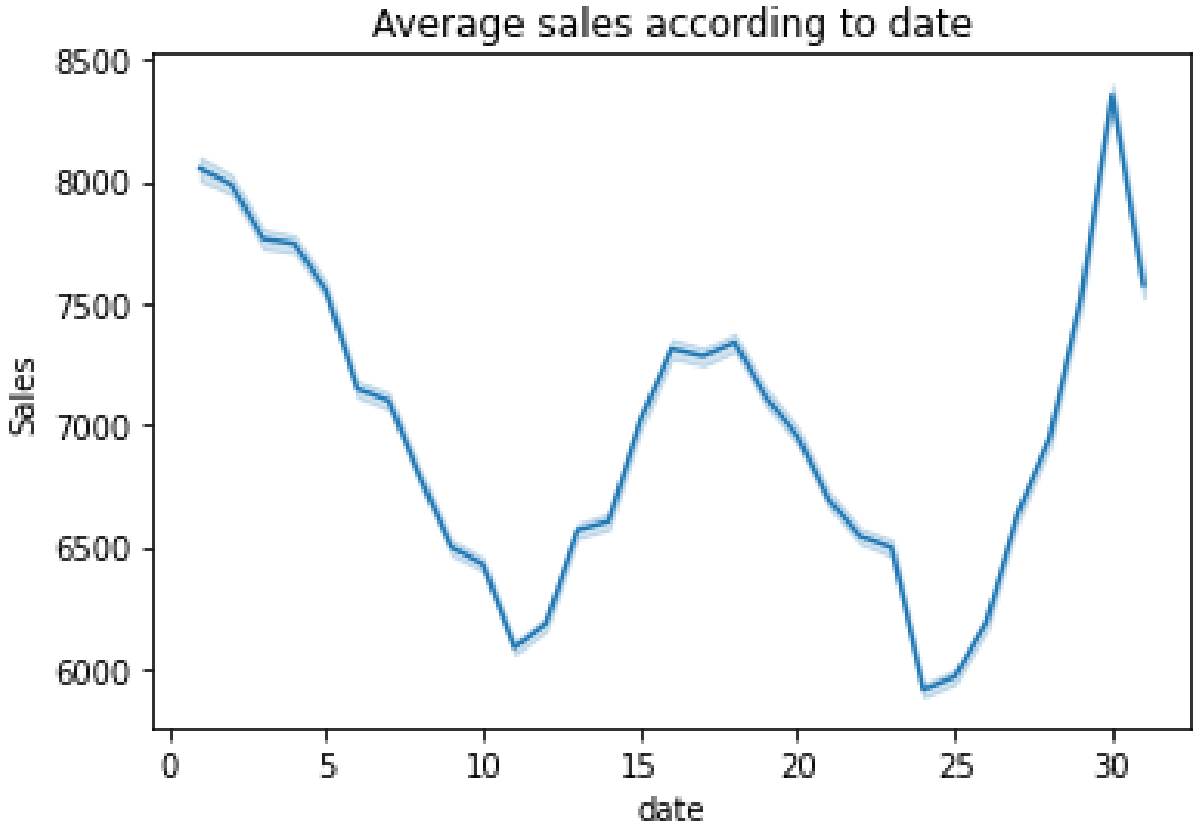

- As seen from this pie chart average sales (2015>2014>2013).But not varying much with year.

# Monthly average sales



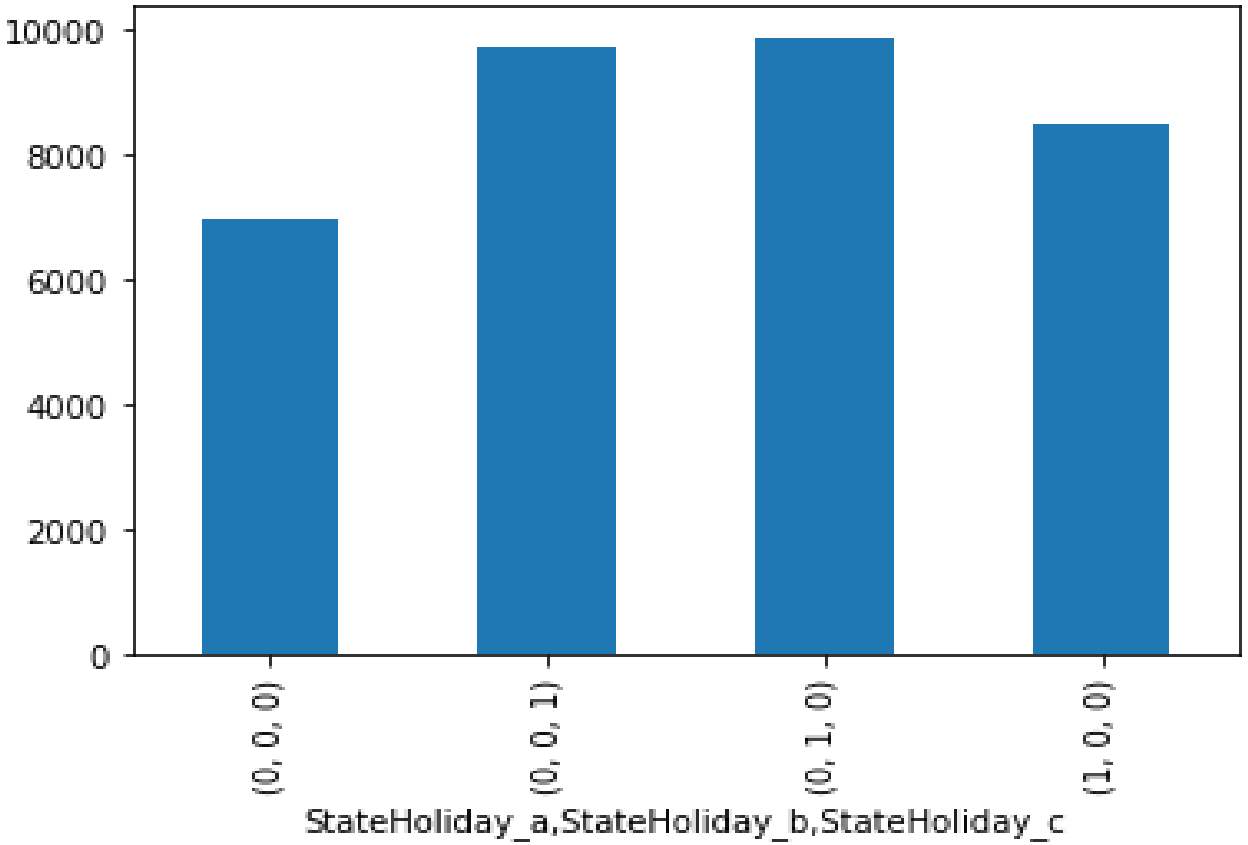

- Monthly average sales also not varying much.
  Only the **12th month** is showing exceptionally high sales.

# Average sales date wise



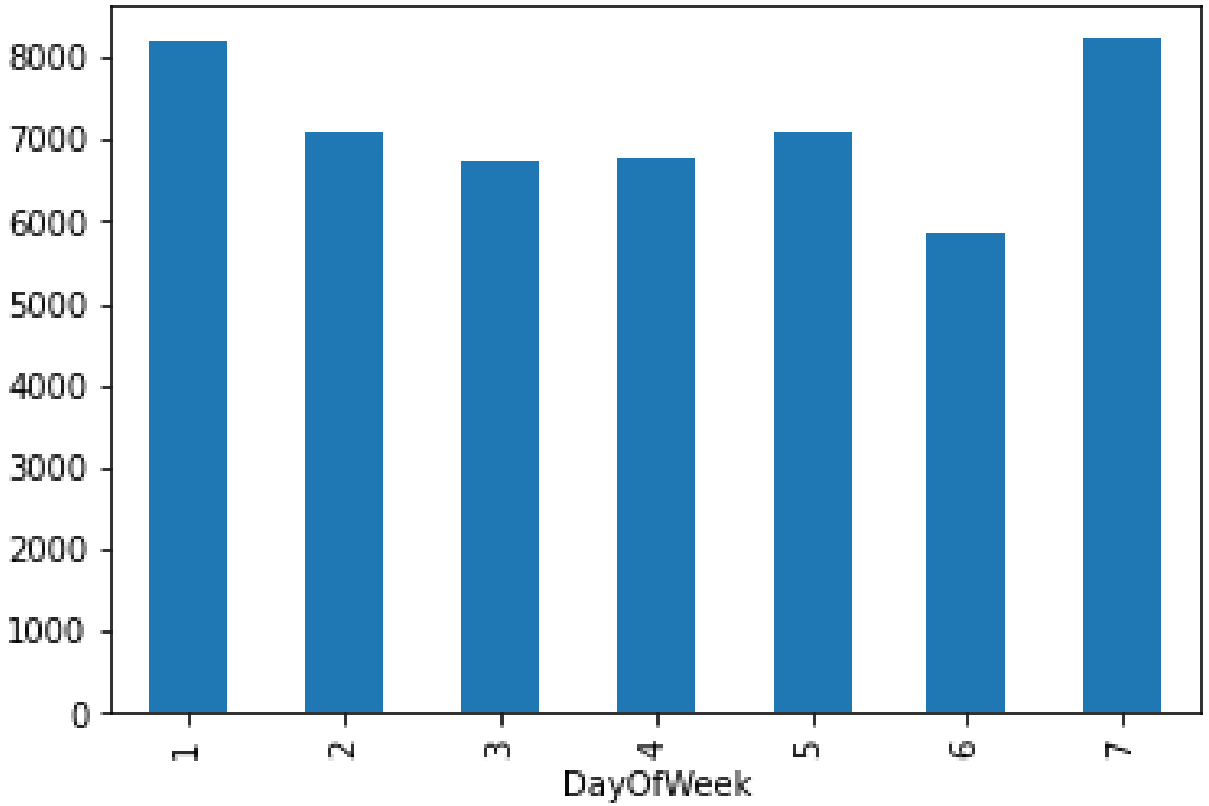

- As seen from above line chart sales is varying according to date.

# Average sales on state holidays



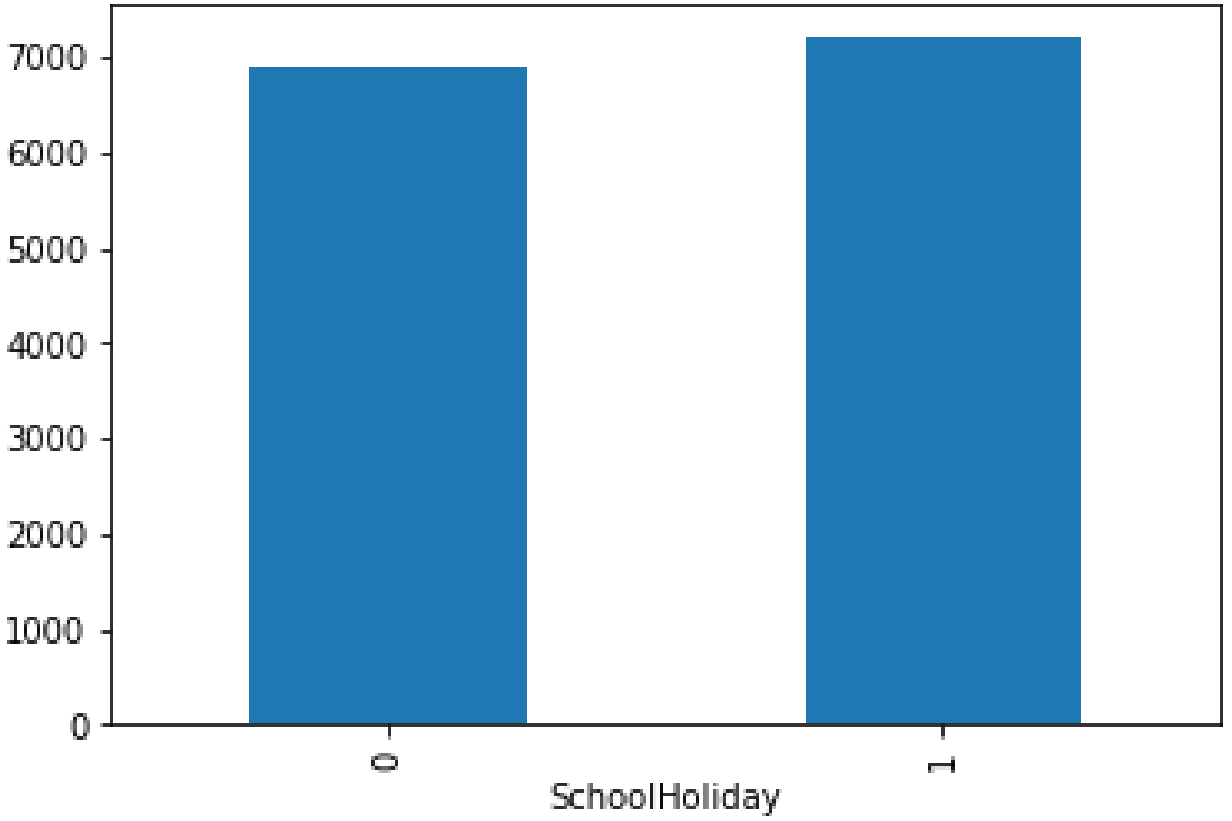

- We could see average **sales** on **state holidays are more** than non state holidays.
- Avg sales **(StateHoliday_b>StateHoliday_a>StateHoliday_c)**

# Average sales by day of week



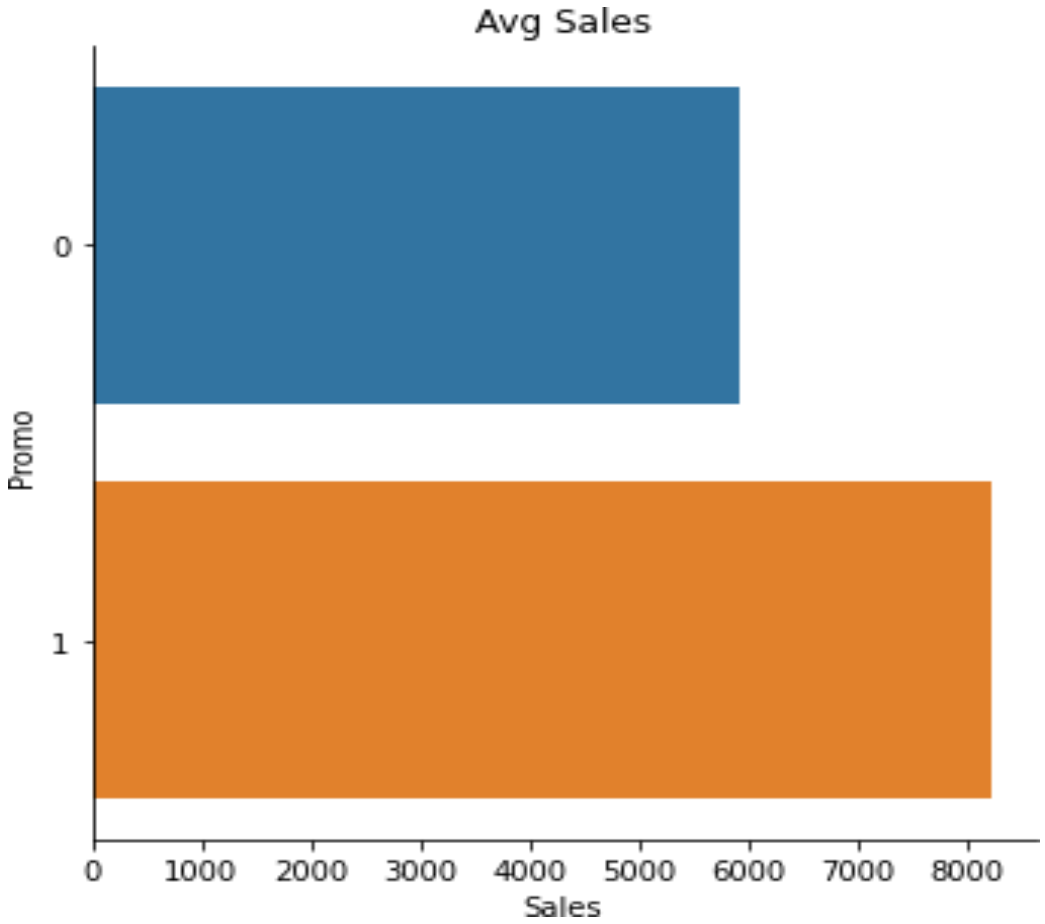

- We could see average sales on some day of week is more than others.
- Maximum average sales are on **(day7 > day1 > day2 > day5).**

# Average sales on school holiday vs Non school holiday



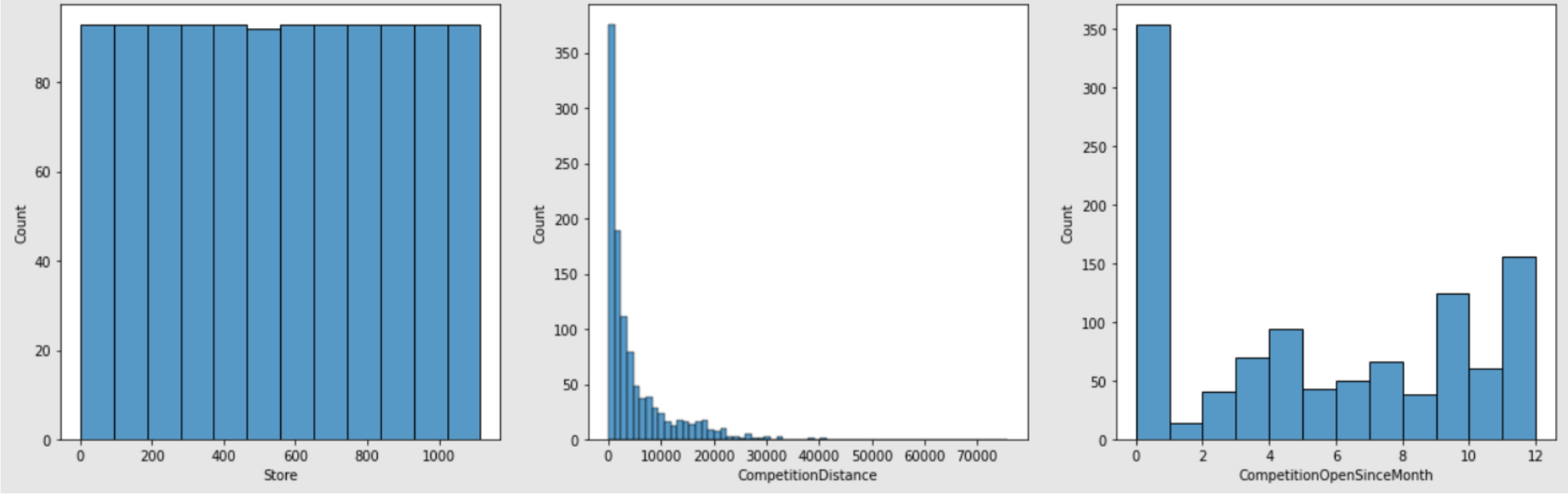

- From this bar plot, we could see on school holidays average sales is more than on normal day

# Average sales with or without promo

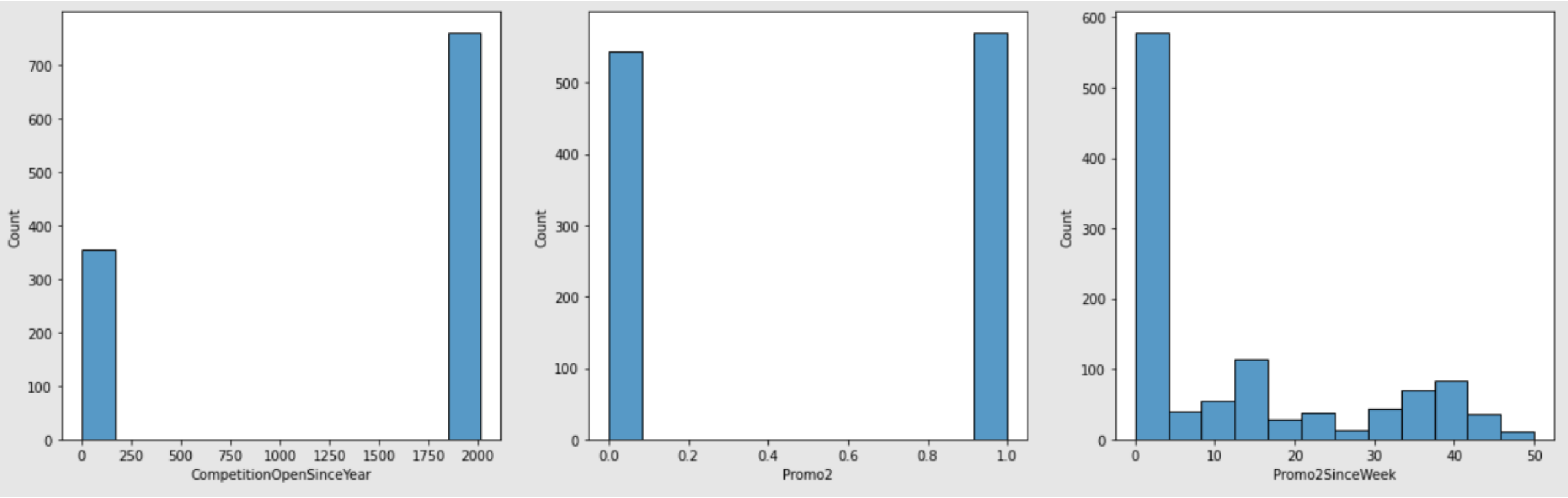



- From this plot, we could see that with promo average sales is more than without promo

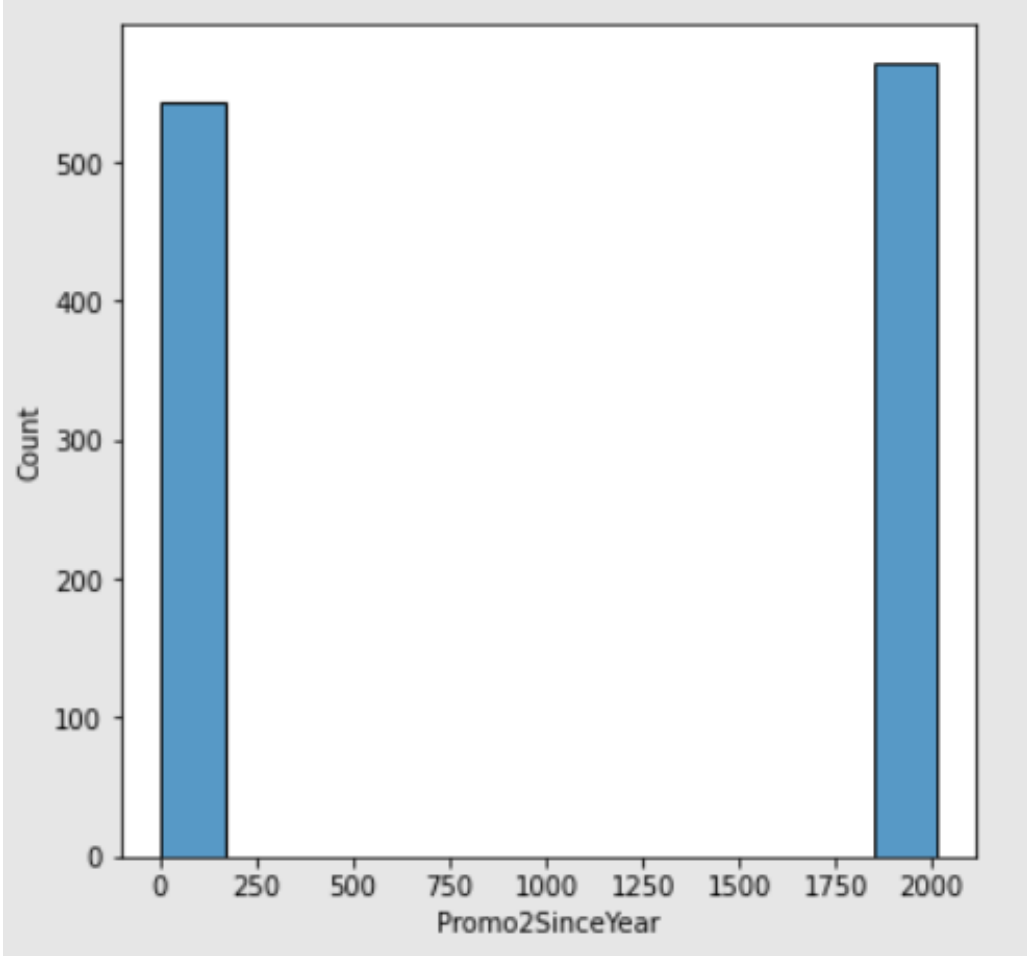
# Distribution of data for store dataframe



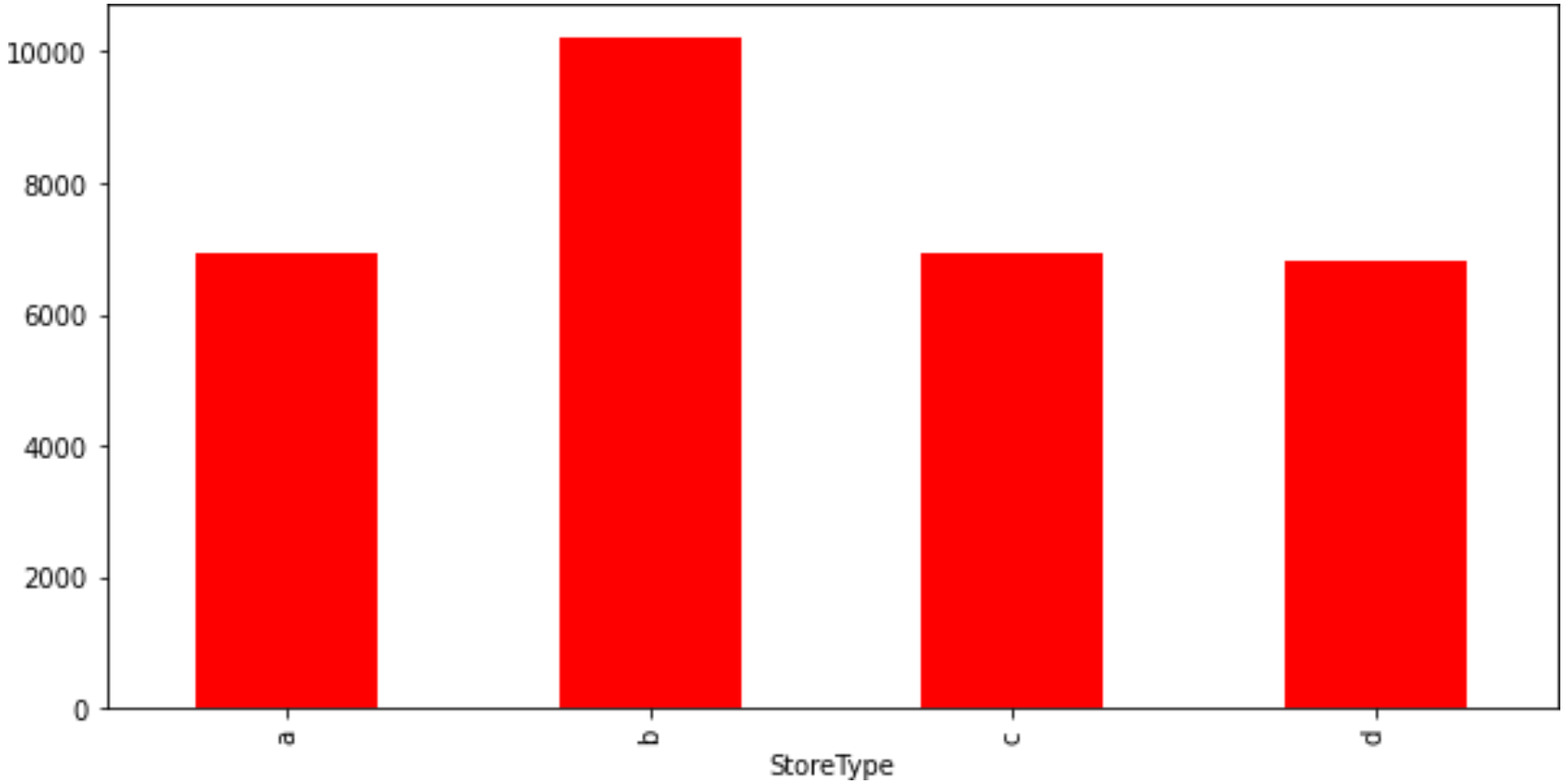
# Distribution of data for store dataframe



# Distribution of data for store dataframe

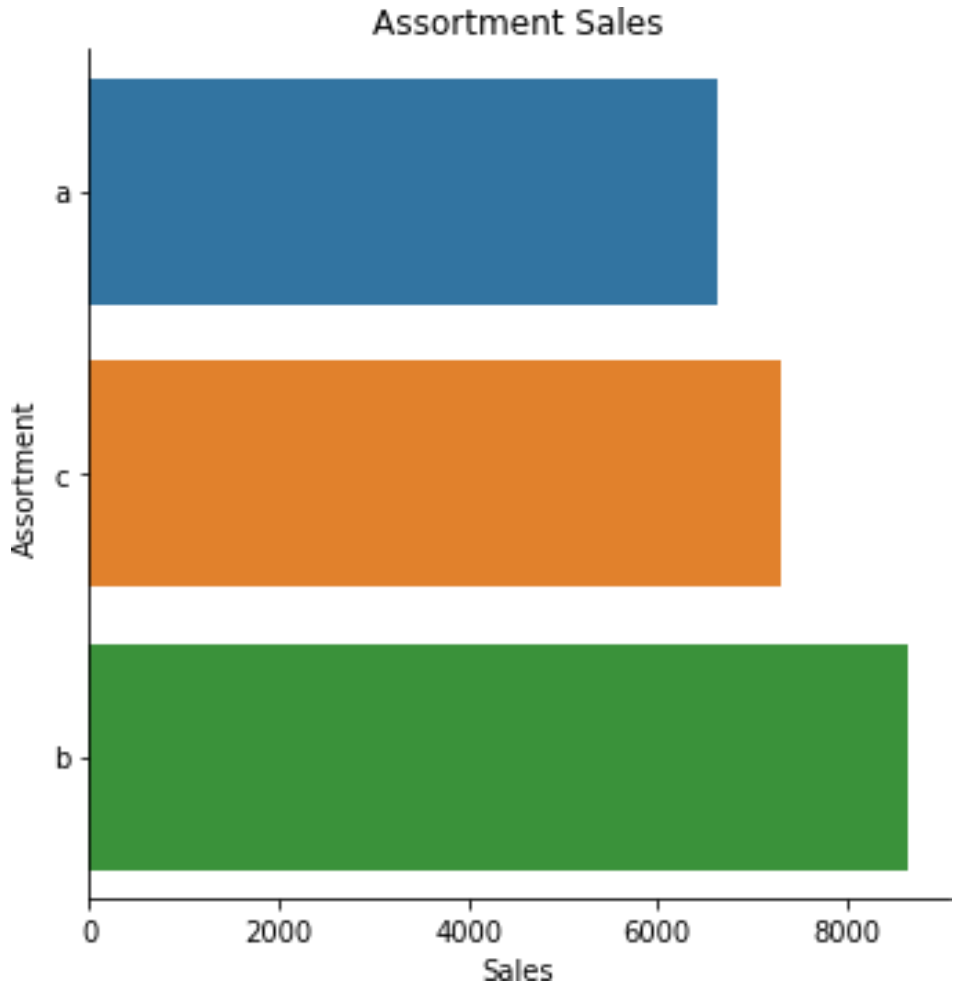


# Average sales according to storetype



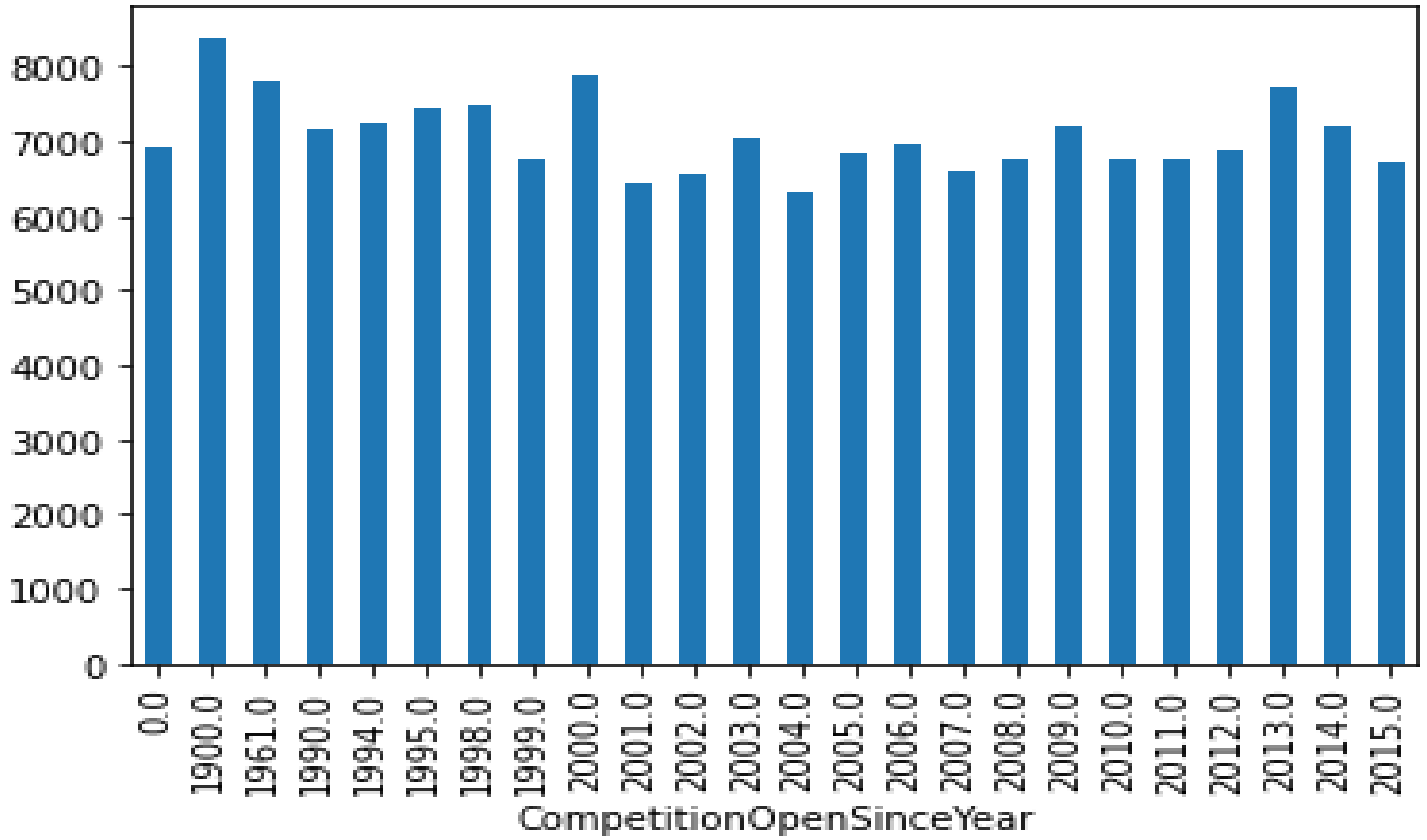

- As seen from above plot sales of store **type B** are much **higher.**

# Average sales assortment wise



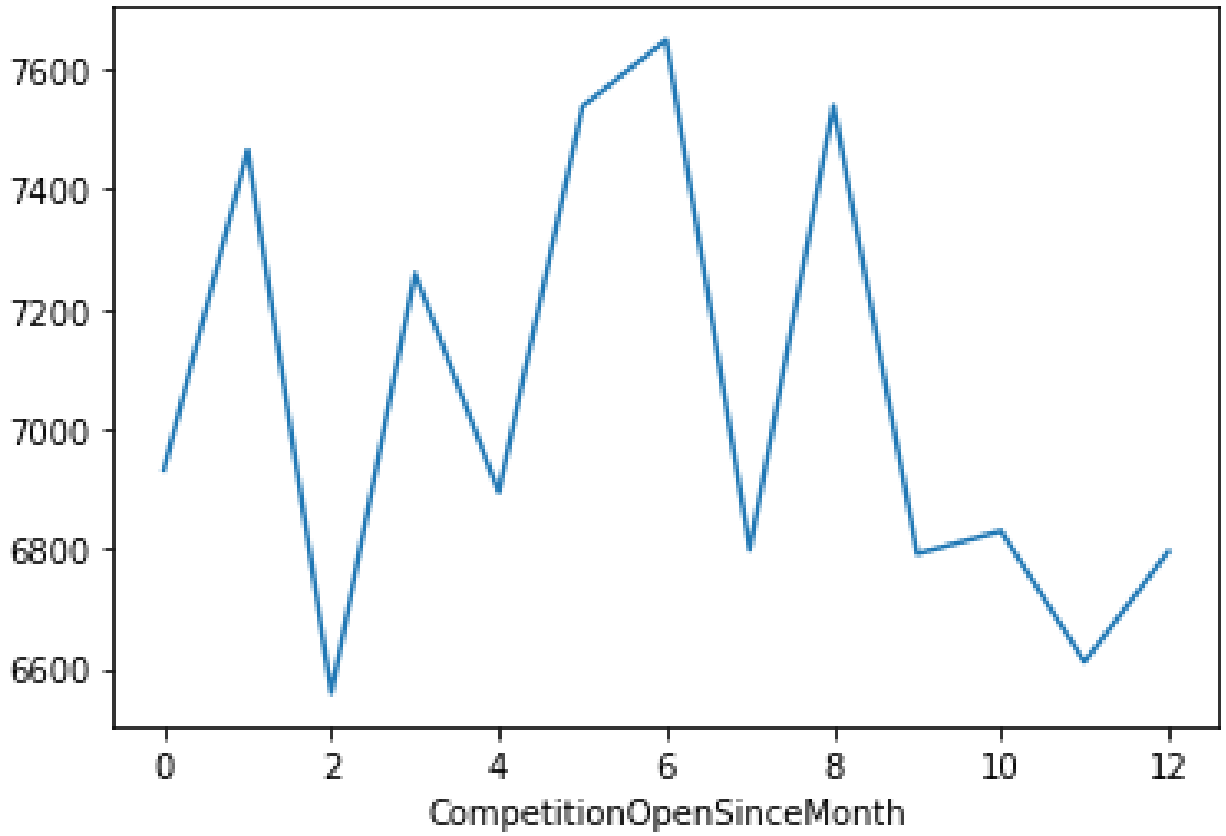

- As seen assortment sales are varying and max sales is with **assortment b.**

# Average sales competition open since year wise

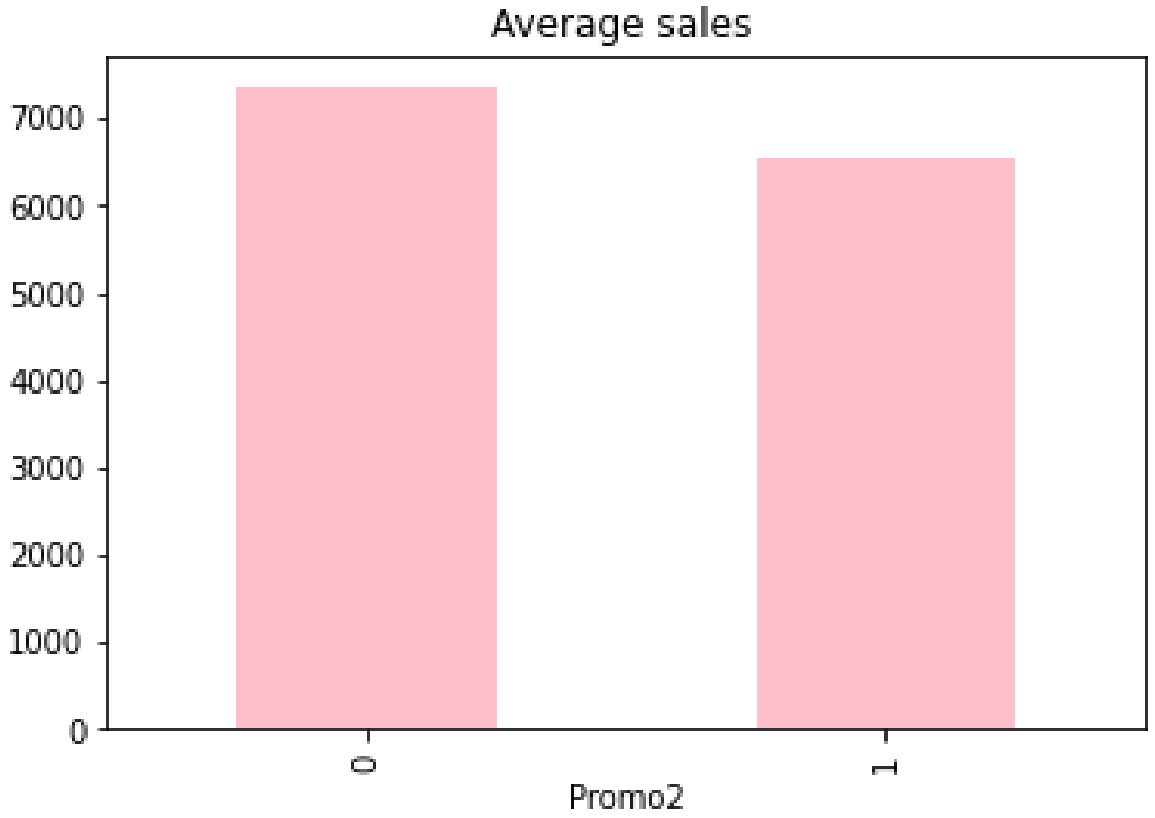



- As seen from above plot average sales of this columns is probably same , so we are removing this column.

# Average sales competition open since month wise

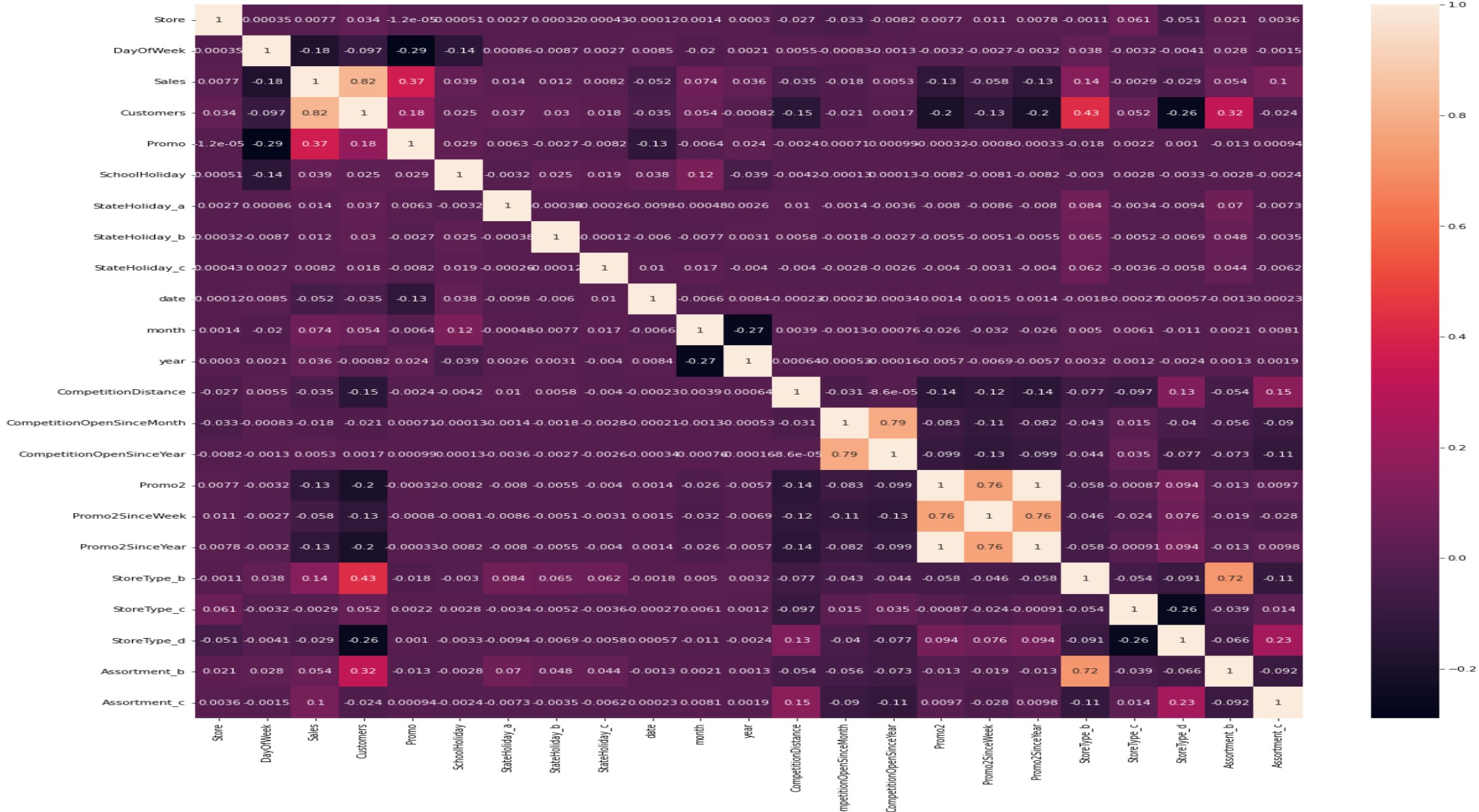


# Average sales with or without Promo2




- As seen from above plot Promo2 affecting sales negatively.

# CORRELATION



# Feature Selection

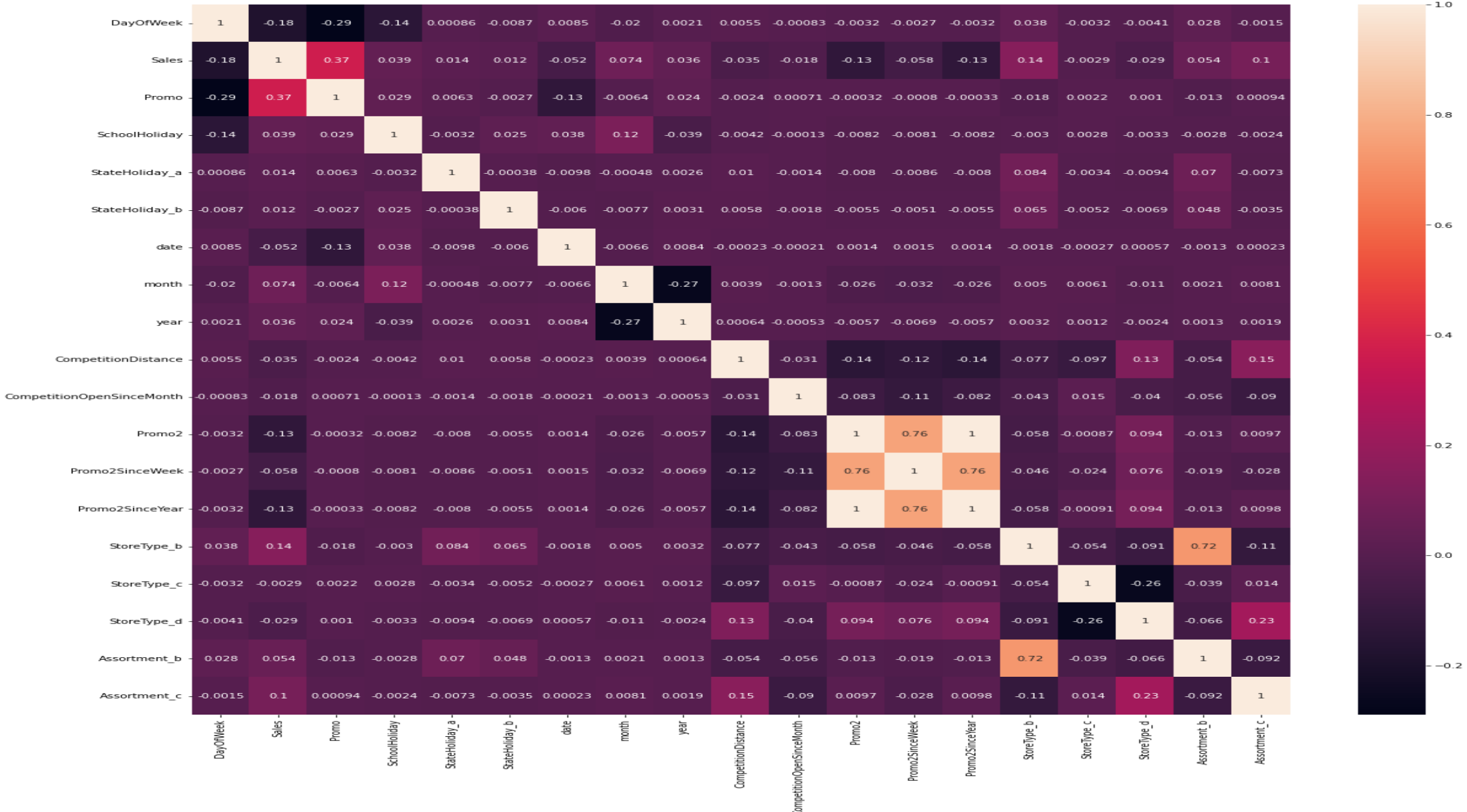
- The customers column is highly correlated but we remove this column because we cannot predict customers in future.
- There are some columns that seems to be least correlated with sales i.e. Compitionopensinceyear ,Stateholiday_C and Store(store_id) so we also drop these columns.
- Promo2 and Promo2SinceYear showing multicollinearity so we are removing any one column from these two.

# Feature Selection

Features we are selecting for our model are:-

I. DayOfWeek
II. Promo
III. SchoolHoliday
IV. StateHoliday
V. Date
VI. Month
VII. Year
VIII. CompetetionDistance
IX. CompetitionOpenSinceMonth
X. Promo2
XI. Promo2SinceWeek
XII. StoreType
XIII. Assortment

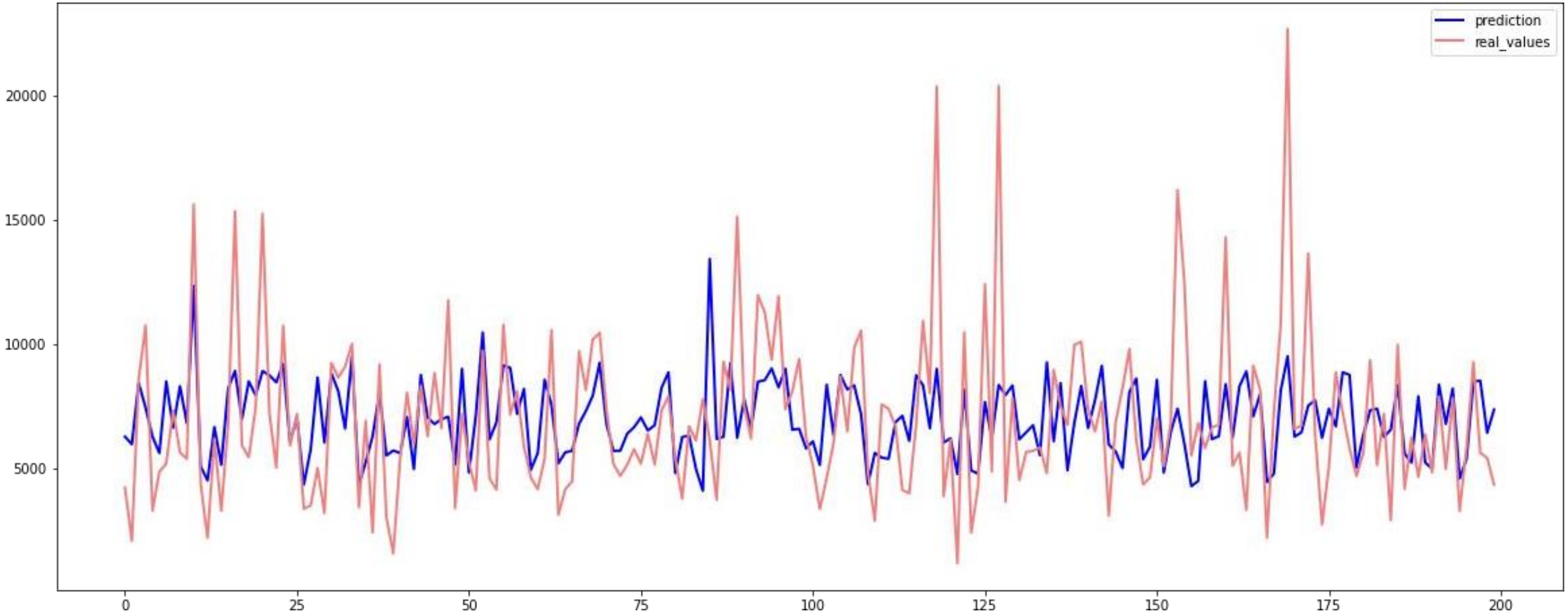
# Correlation after feature engineering



# We used some models and find which model is best fitting with data. Models we used are given below.

- Linear Regression
- Decision Tree
- XGBoost
- Random Forest

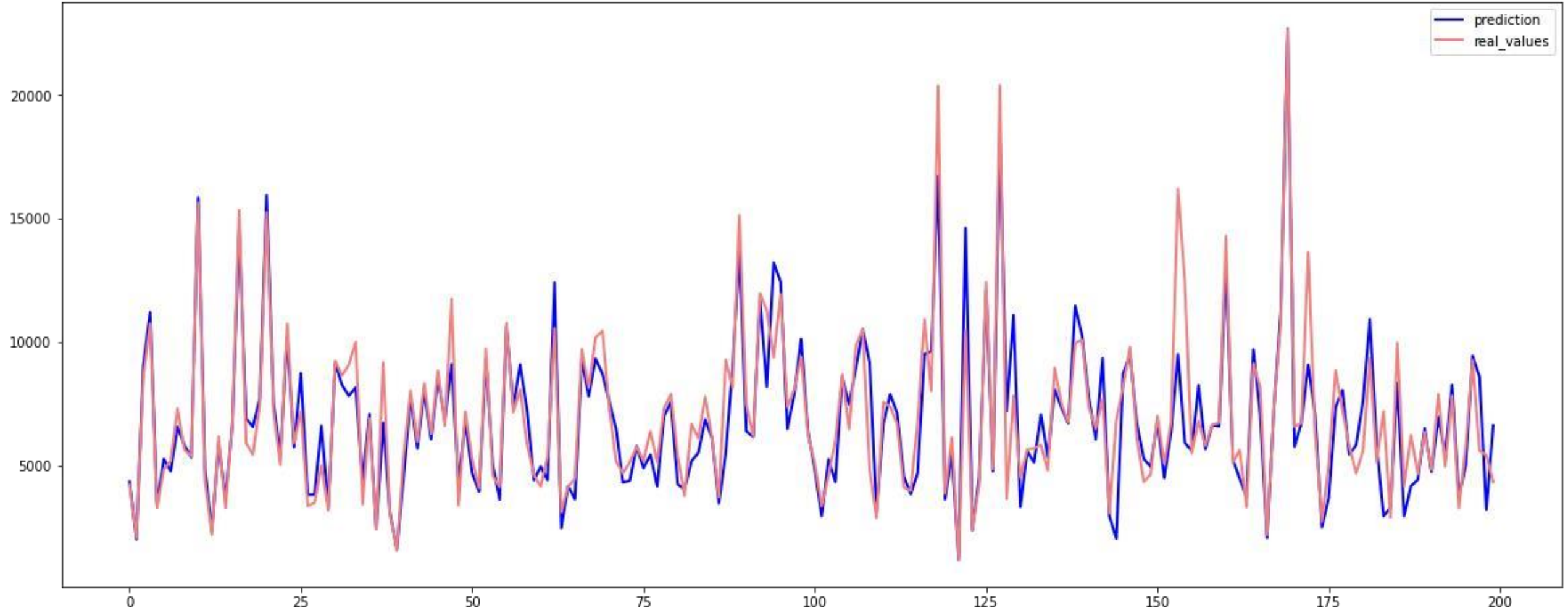
# Linear regression




R2 score : 0.2109690140757124 5

Mean Squared Error : 2750.8171177596555

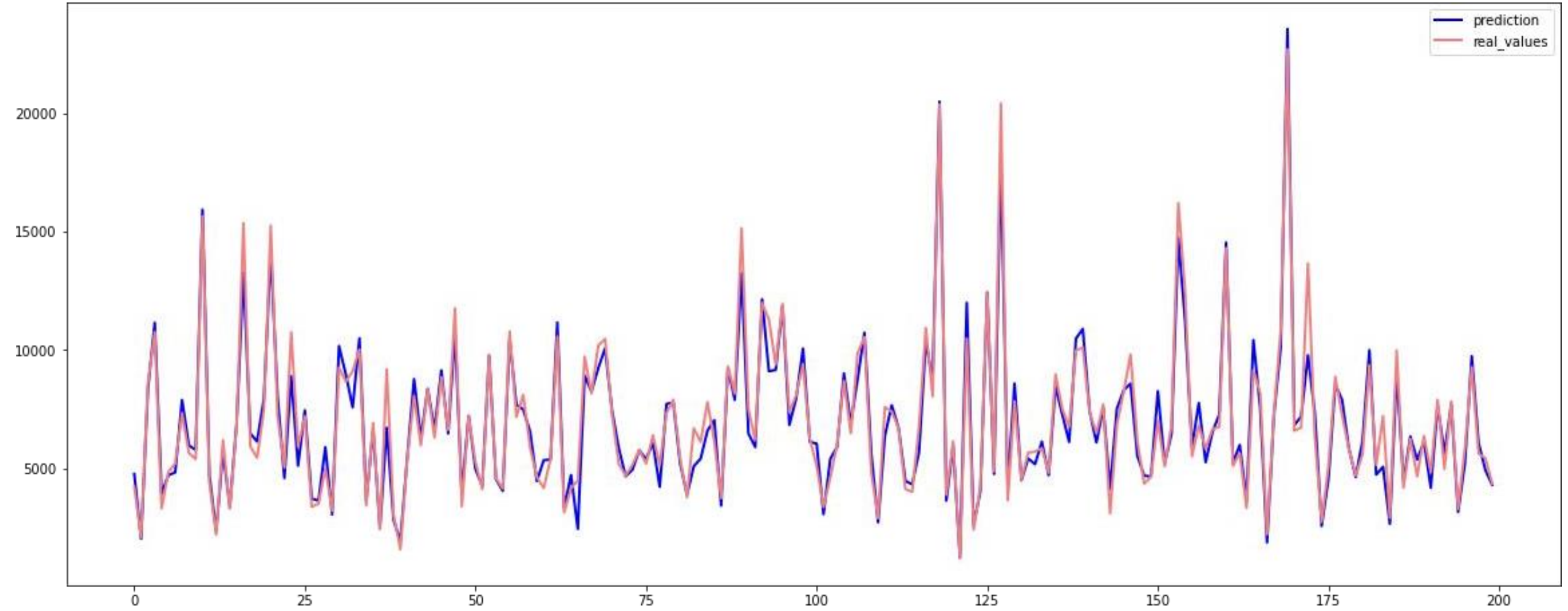
# Decision Tree




Accuracy score : 0.9970735015558014
R2 score : 0.8017666074277554

Mean Absolute Error : 856.039877213427
Root Mean Squared Error : 1378.805781065069

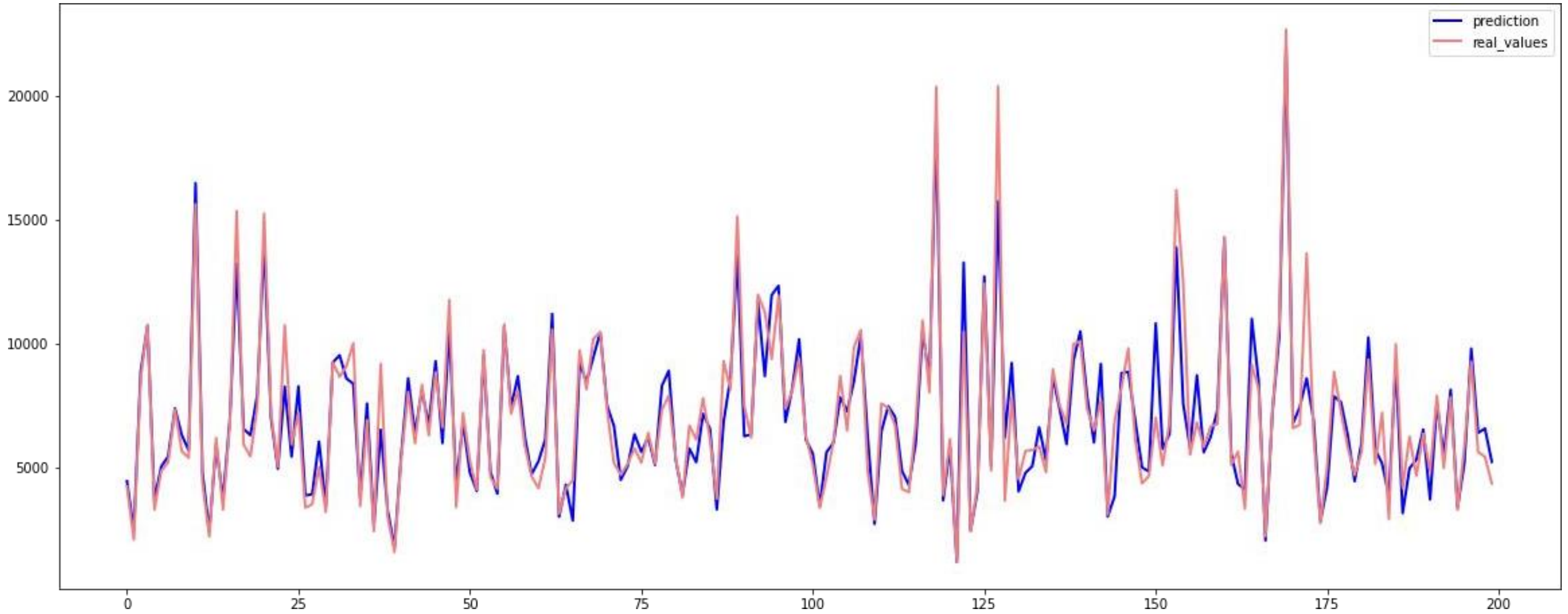
# XGBoost




Accuracy score : 0.942437460428324
R2 score : 0.9424374605624231

Mean Absolute Error : 496.7095188366338
Root Mean Squared Error : 742.9928969806351

# Random Forest




Accuracy score : 0.8945691889339075
R2 score : 0.8662517846224456

Mean Absolute Error : 704.3839797356239
Root Mean Squared Error : 1132.55339638993

# Model Explainability of XGboost

- We took XGboost as our final modal because it was giving least error and best r2 score, below is the weightage of features in our model.

| Weight | Feature |
|---|---|
| 0.1896 | f12 |
| 0.1576 | f1 |
| 0.1106 | f8 |
| 0.0768 | f9 |
| 0.0757 | f10 |
| 0.0652 | f16 |
| 0.0568 | f13 |
| 0.0541 | f11 |
| 0.0457 | f14 |
| 0.0443 | f15 |
| 0.0351 | f0 |
| 0.0205 | f6 |
| 0.0197 | f5 |
| 0.0165 | f4 |
| 0.0141 | f3 |
| 0.0095 | f7 |
| 0.0083 | f2 |

| | index | 0 |
|---|---|---|
| 0 | 0 | DayOfWeek |
| 1 | 1 | Promo |
| 2 | 2 | SchoolHoliday |
| 3 | 3 | StateHoliday_a |
| 4 | 4 | StateHoliday_b |
| 5 | 5 | date |
| 6 | 6 | month |
| 7 | 7 | year |
| 8 | 8 | CompetitionDistance |
| 9 | 9 | CompetitionOpenSinceMonth |
| 10 | 10 | Promo2 |
| 11 | 11 | Promo2SinceWeek |
| 12 | 12 | StoreType_b |
| 13 | 13 | StoreType_c |
| 14 | 14 | StoreType_d |
| 15 | 15 | Assortment_b |
| 16 | 16 | Assortment_c |

# Technology Used

- Python(Programming Language)
- Libraries used:1) Pandas
  2)Numpy
  3)Matplotlib
  4)Seaborn
  5)Sklearn

# THANK YOU