Capstone Project-2

Retail Sales Prediction

By

K Vidyasagar

Points to be discussed:

- Problem statement
- Data Summary
- Data Cleaning
- Univariate Analysis
- Bivariate Analysis
- Machine Learning Models Training and Testing
- Conclusions

Problem statement:

- Look at the given datasets and study the relationship between different features or trends in different features. Find the correlation between sales and other features and build an efficient machine learning model to predict future sales for given input variables.
- You are provided with historical sales data for 1,115 Rossmann stores. The task is to forecast the sales column for the test set.

Data Summary:

- Two data sets were given (Rossmann sales and store details data)
- **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
- **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
- **Store Type -** differentiates between 4 different store models: a, b, c, d

Data Summary:

- **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 participation.

Data Summary:

- **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, November of any given year for that store.
- **Date:** the day when transaction took place.
- **Day Of Week:** a particular day in a week.

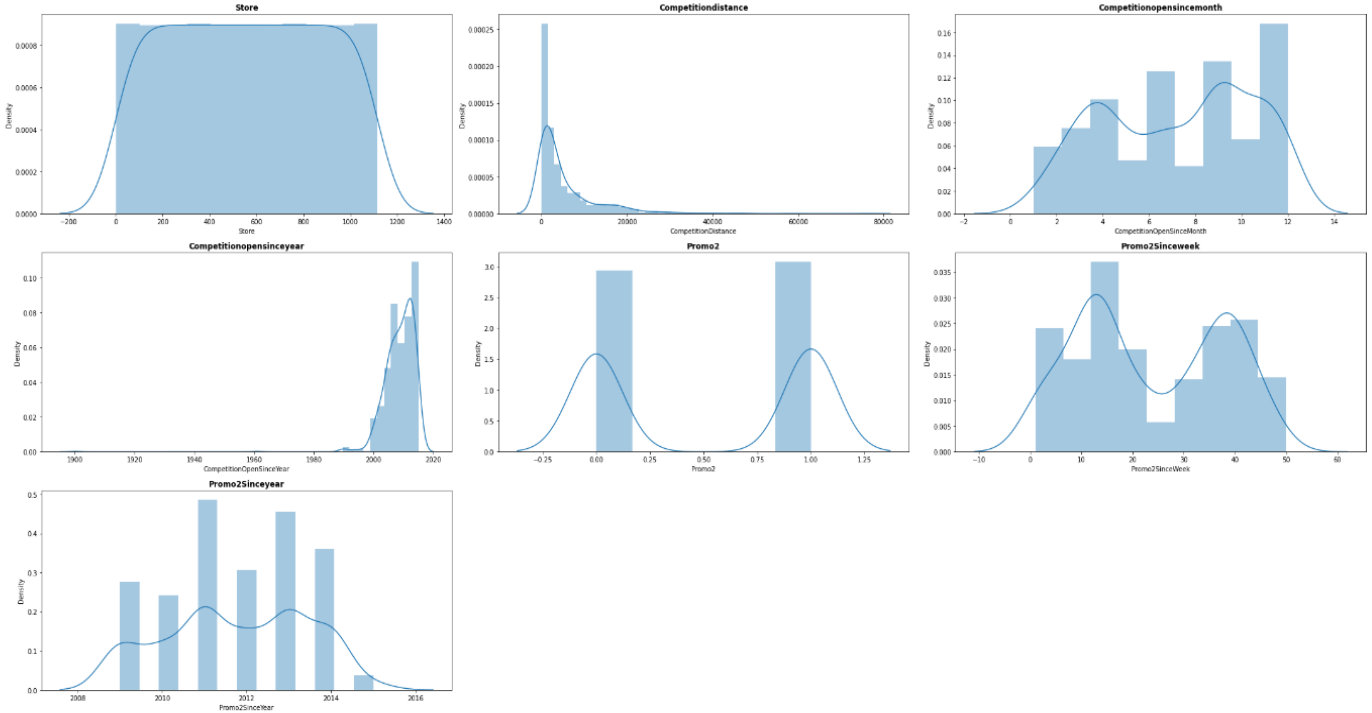
Data Cleaning:

```
# missing values?
store_df.isnull().sum()
```

```
Store                         0
StoreType                     0
Assortment                    0
CompetitionDistance           3
CompetitionOpenSinceMonth   354
CompetitionOpenSinceYear    354
Promo2                        0
Promo2SinceWeek             544
Promo2SinceYear             544
PromoInterval               544
dtype: int64
```

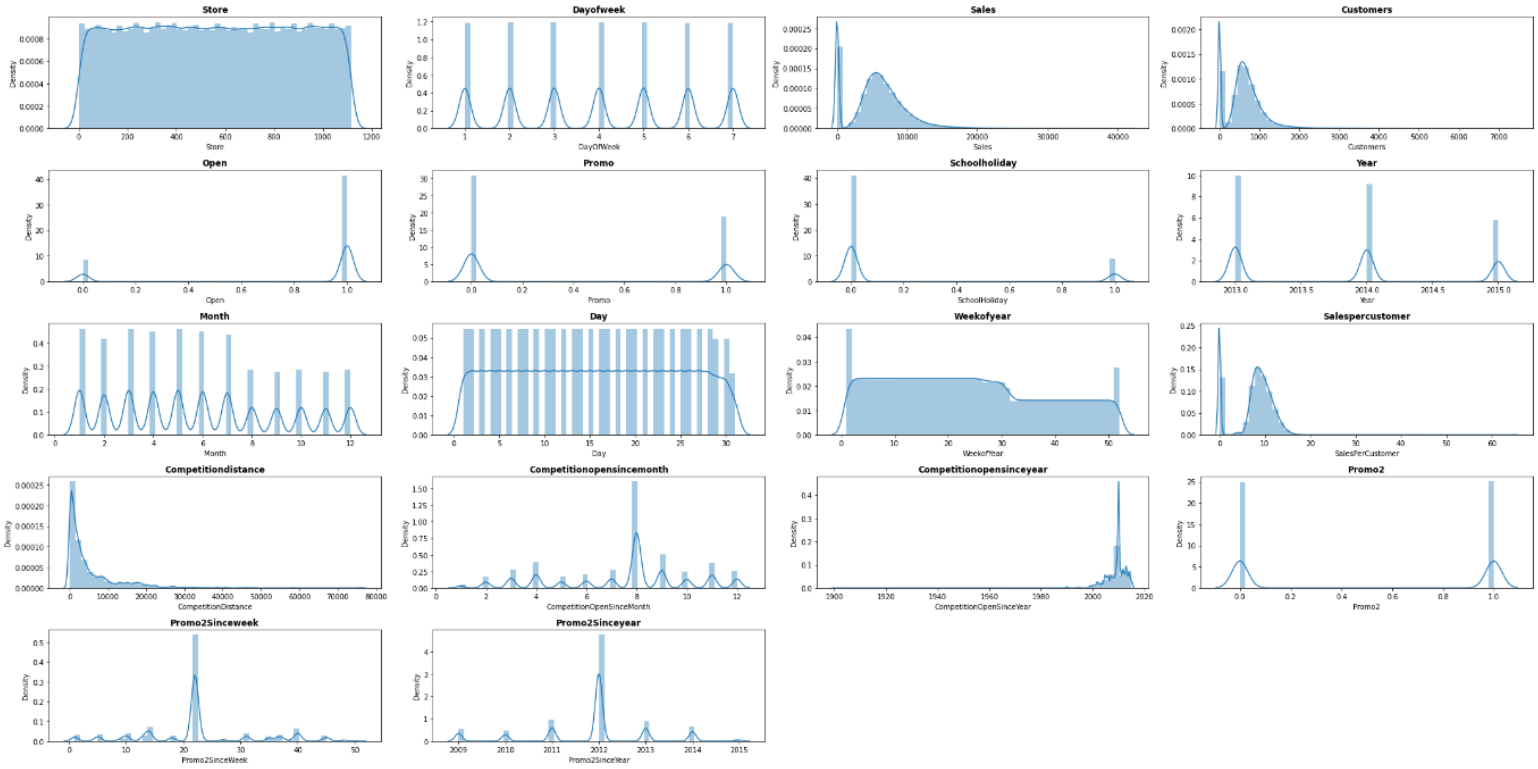
```
# let us replace null values in store_df
store_df['CompetitionOpenSinceMonth'].fillna(store_df['CompetitionOpenSinceMonth'].median(), inplace = True)
store_df['CompetitionOpenSinceYear'].fillna(store_df['CompetitionOpenSinceYear'].median(), inplace = True)
store_df['Promo2SinceWeek'].fillna(store_df['Promo2SinceWeek'].median(), inplace = True)
store_df['Promo2SinceYear'].fillna(store_df['Promo2SinceYear'].median(), inplace = True)
store_df['PromoInterval'].fillna('Not Known', inplace = True)
```

Univariate Analysis:



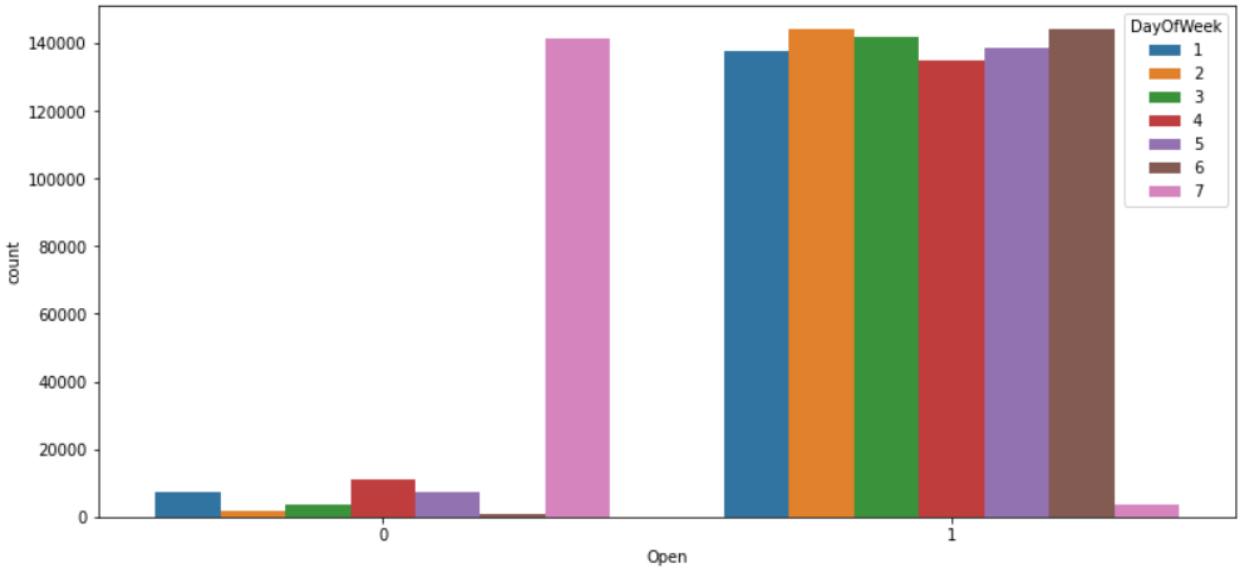
Distribution of numerical features in store data set

Univariate Analysis:



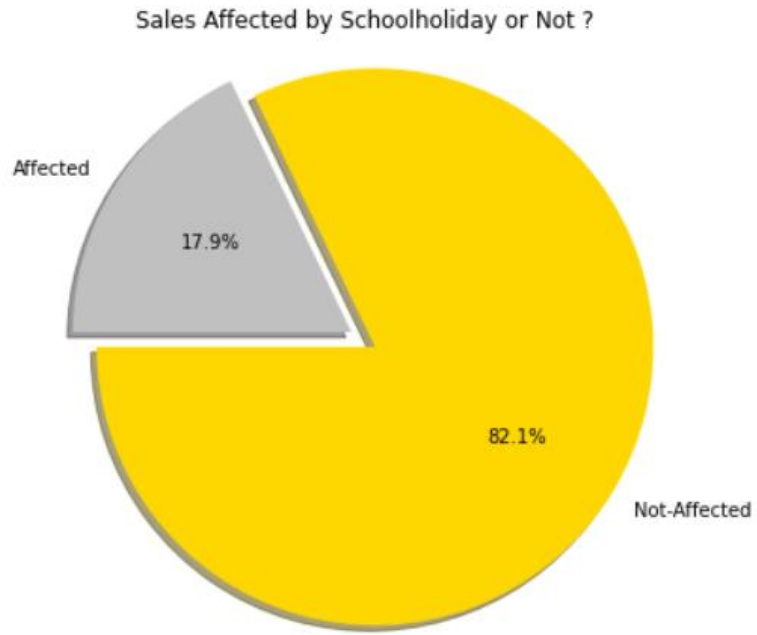
Distribution of numerical features in our final data set

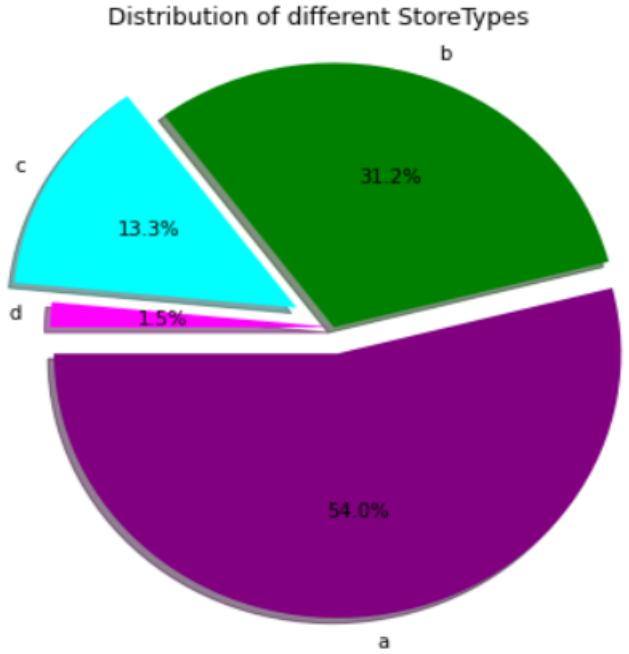
Univariate Analysis:



Count of shops opened or closed on days of week

Univariate Analysis:

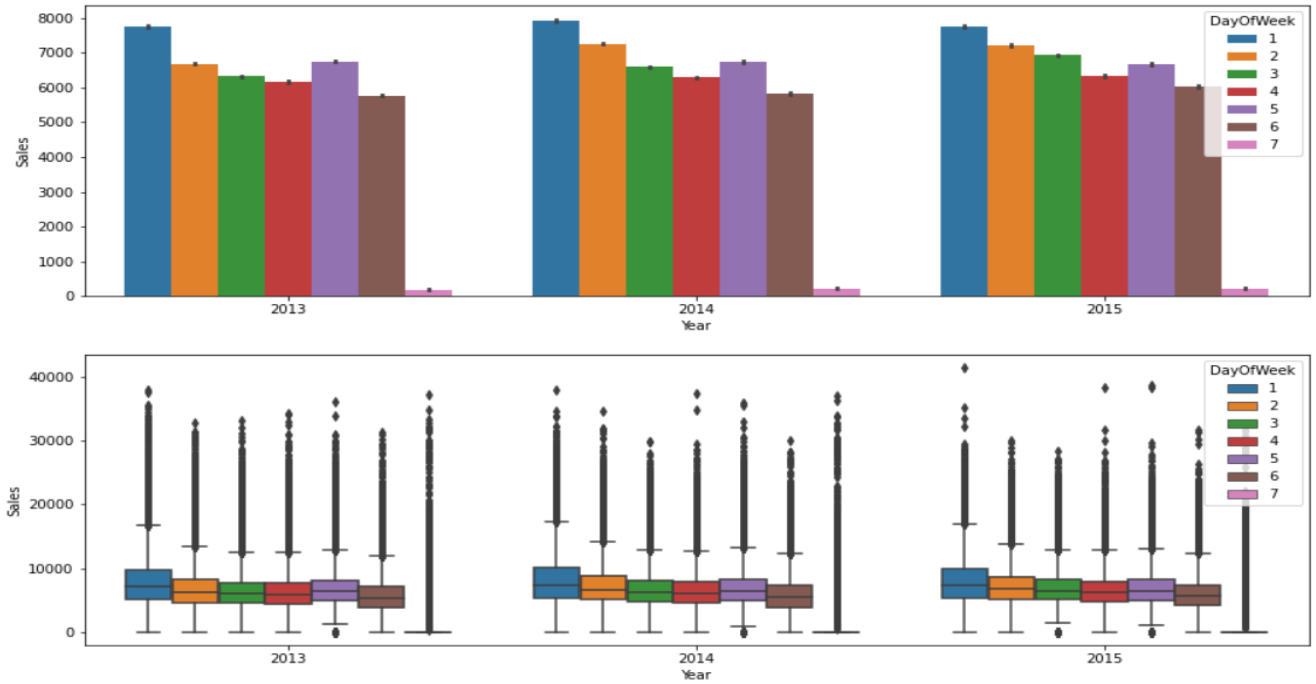




Count of shops opened or closed on days of week

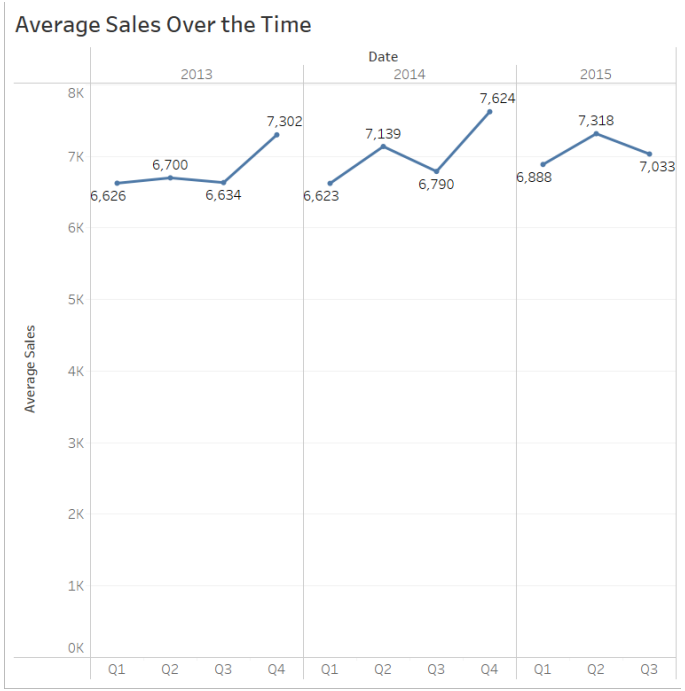
Distribution of store types

Bivariate Analysis:



Sales in a week every year

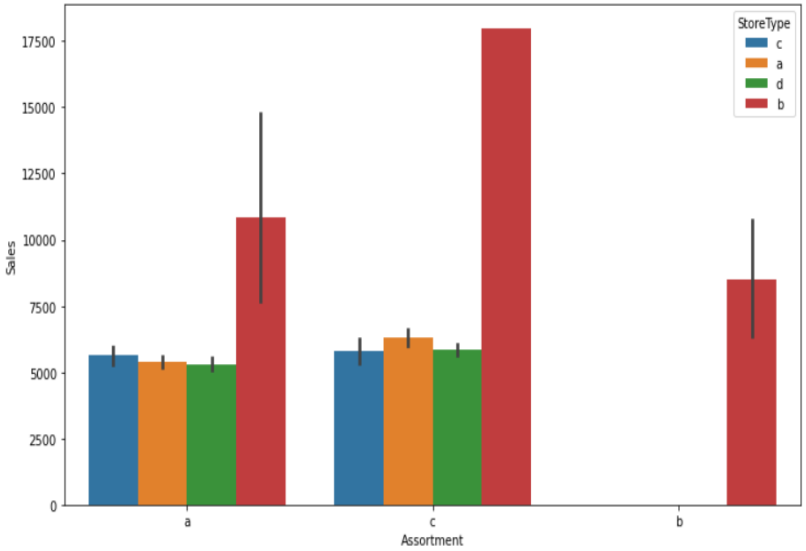
Bivariate Analysis:



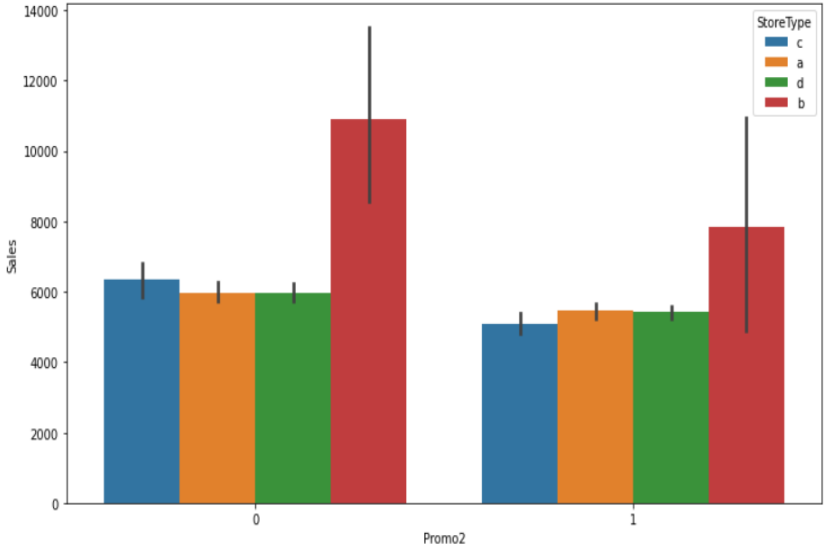
Average Sales in a quarter

Sales Vs Customers

Bivariate Analysis:

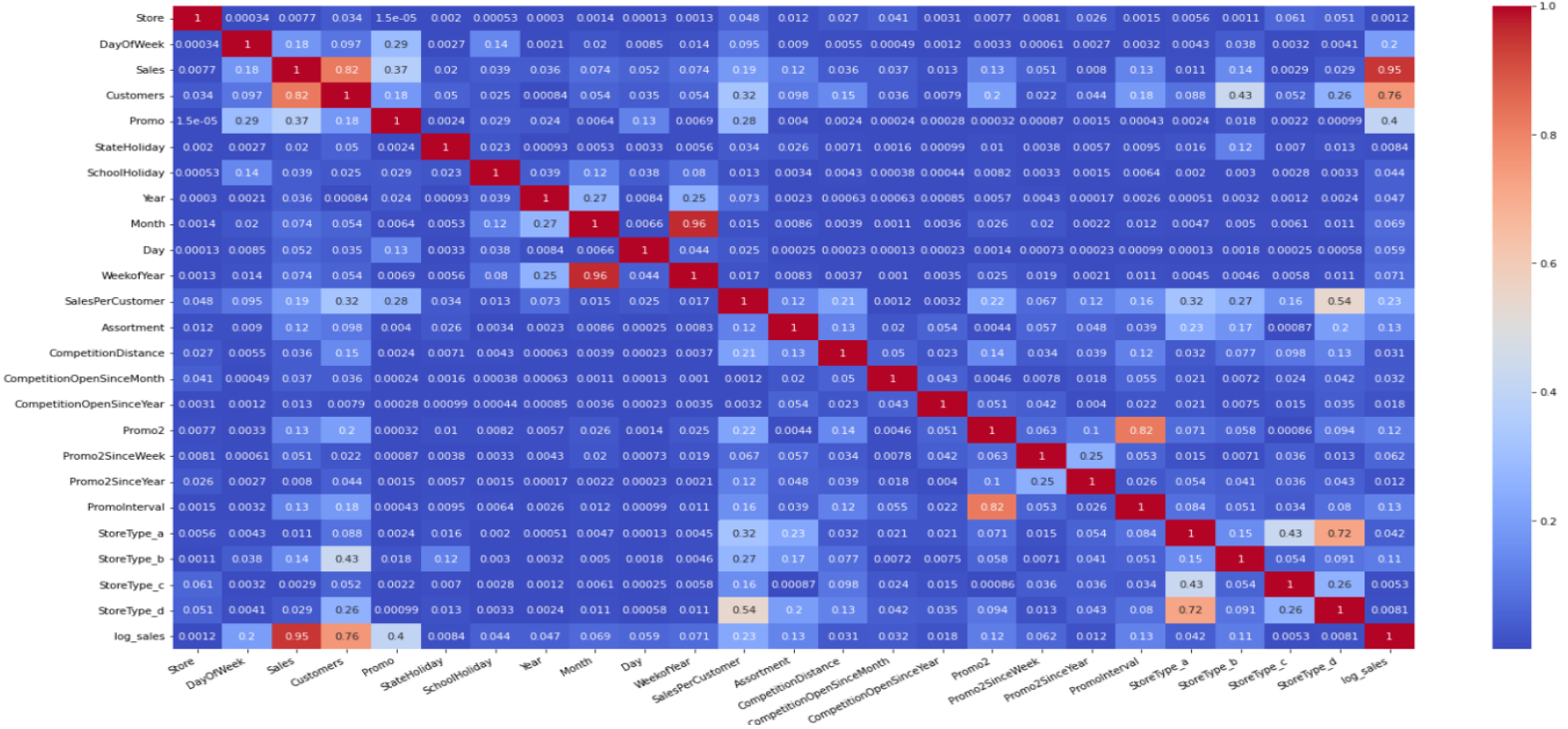


Average Sales for store type and assortment



Average Sales for store type and promo2

Bivariate Analysis:



Correlation Matrix

Bivariate Analysis:

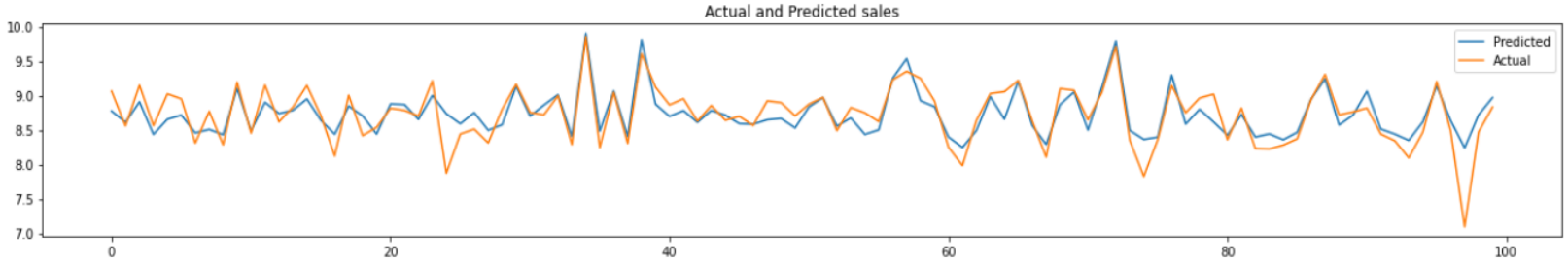
	variables	VIF
0	StoreType_a	5.809873e+06
1	StoreType_d	3.289187e+06
2	StoreType_c	1.435935e+06
3	StoreType_b	1.978117e+05
4	Month	1.405801e+01
5	WeekofYear	1.385696e+01
6	Promo2	3.370507e+00
7	PromoInterval	3.226180e+00
8	SalesPerCustomer	2.004016e+00
9	Customers	1.540769e+00
10	Promo	1.345064e+00
11	Assortment	1.140743e+00
12	CompetitionDistance	1.134116e+00
13	DayOfWeek	1.120923e+00
14	Promo2SinceYear	1.108198e+00
15	Year	1.093354e+00
16	Promo2SinceWeek	1.086224e+00
17	Day	1.060074e+00
18	SchoolHoliday	1.055205e+00
19	CompetitionOpenSinceMonth	1.019778e+00
20	StateHoliday	1.016211e+00
21	CompetitionOpenSinceYear	1.012580e+00
22	Store	1.010046e+00

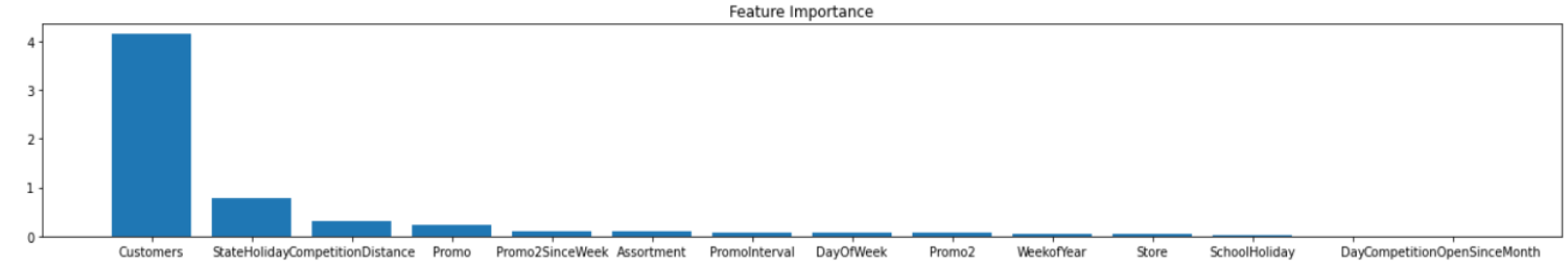
	variables	VIF
0	CompetitionOpenSinceMonth	6.821399
1	Promo2	6.186038
2	Promo2SinceWeek	5.403561
3	PromoInterval	5.280952
4	DayOfWeek	4.727793
5	Customers	4.669523
6	Day	3.975561
7	Store	3.699444
8	WeekofYear	3.532193
9	Promo	1.956727
10	Assortment	1.936679
11	CompetitionDistance	1.536476
12	SchoolHoliday	1.262902
13	StateHoliday	1.004358

$$VIF = \frac{1}{1 - R^2}$$

VIF Values

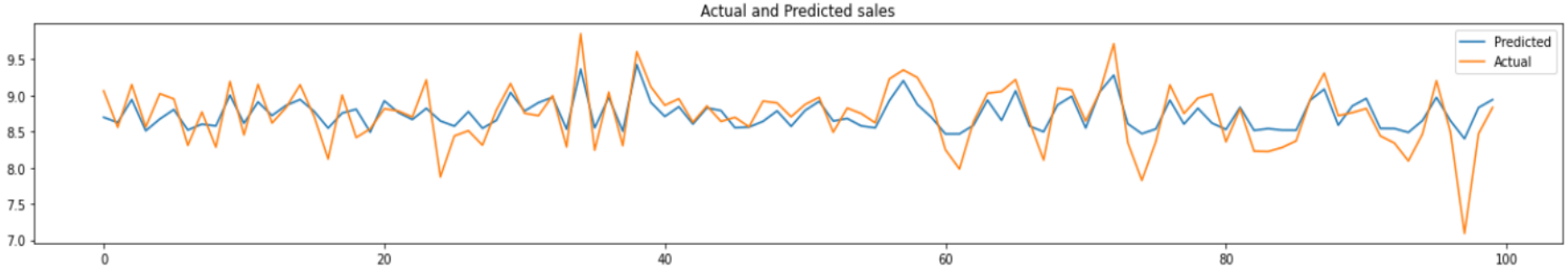
Machine Learning Models Training and Testing:

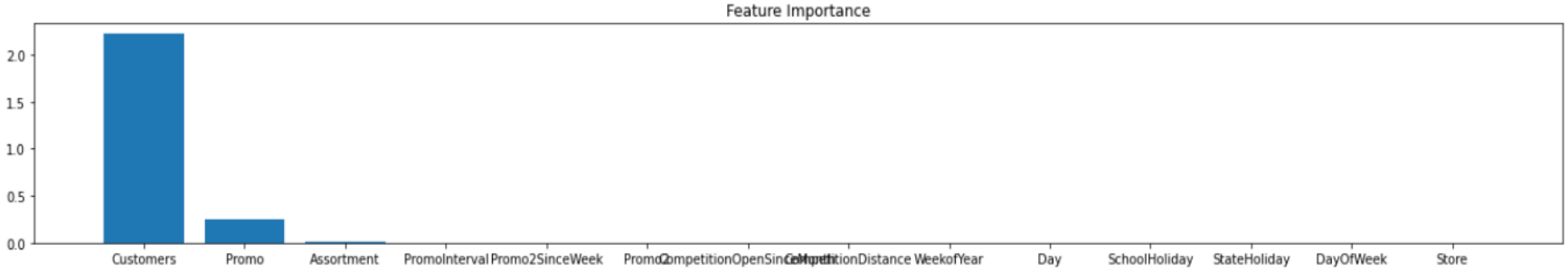





```
LinearRegression()
```

Machine Learning Models Training and Testing:

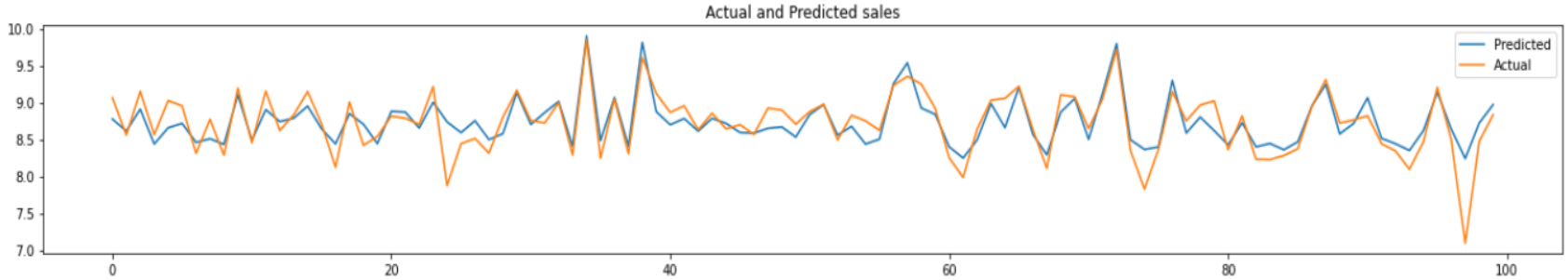


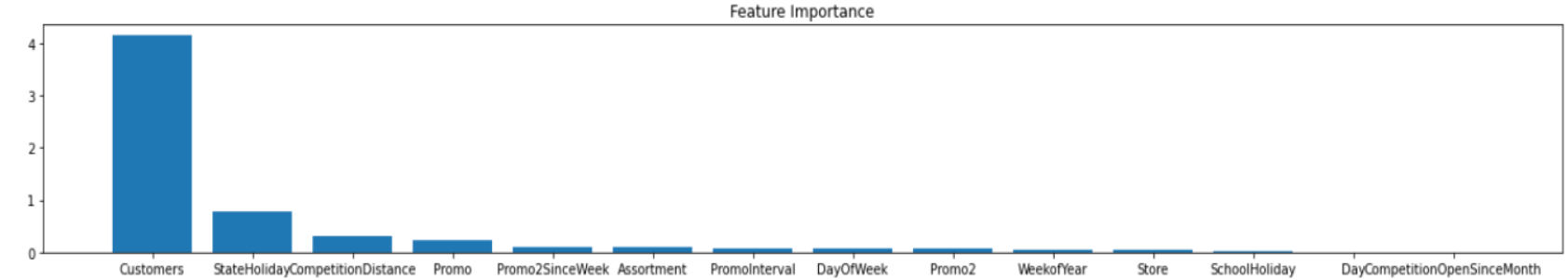



```
Lasso(alpha=0.01, max_iter=3000)
```

Machine Learning Models Training and Testing:

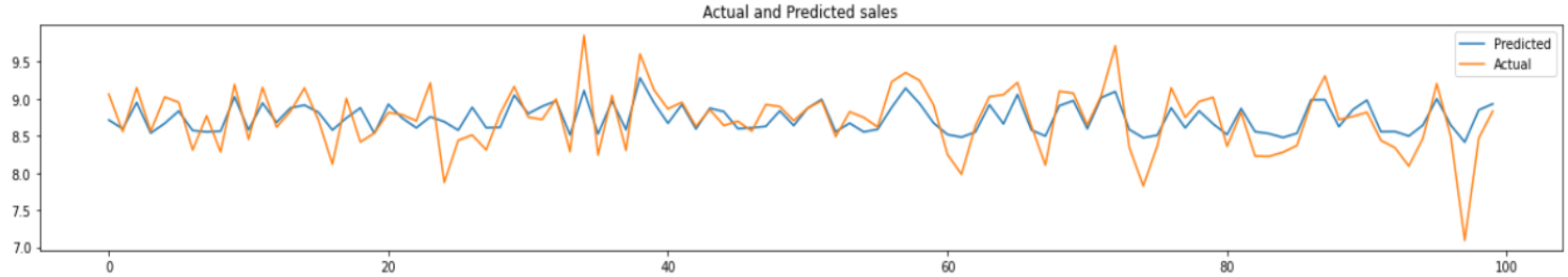
-

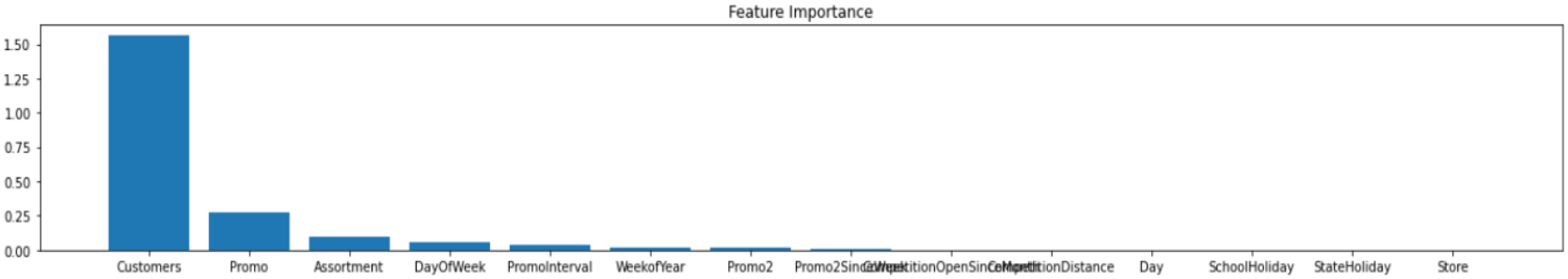





```
Ridge(alpha=0.01)
```

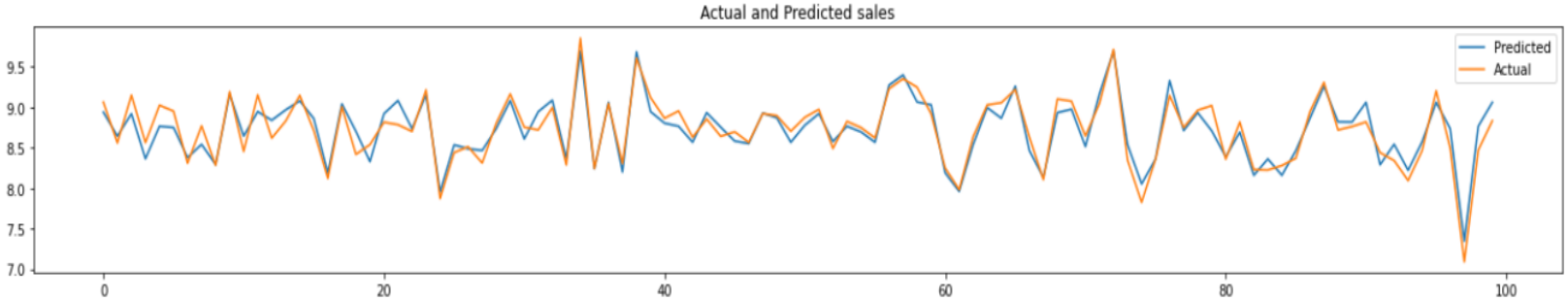
Machine Learning Models Training and Testing:

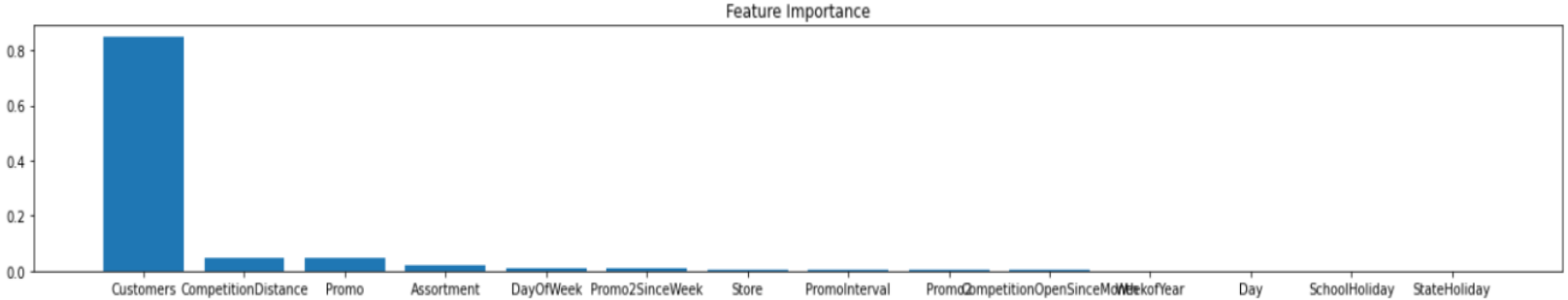





```
ElasticNet(alpha=0.01)
```

Machine Learning Models Training and Testing:






```
DecisionTreeRegressor(max_depth=10, min_samples_leaf=40, min_samples_split=50,
                      random_state=1)
```

Conclusions:

1. From correlation matrix, we can say that 'Customers' feature is highly correlated to Sales (dependent Variable).
2. The 'Month' feature was removed instead of week of year because these both features were correlated and 'Month' is less correlated with 'Sales' compared to later one.
3. In linear regression, Customers is the most influencing feature and State Holiday is at the second place.
4. In Decision Tree Regressor, Customers is the most influencing feature and Competition Distance is at the second place.
5. We find that among these five algorithms, Decision Tree Regression giving good results.

Conclusions:

6. RMSE Comparisons (For Test dataset):
   - A. Linear Regression: 0.2456
   - B. Decision Tree Regressor: 0.157
   - C. Lasso Regressor: 0.286
   - D. Ridge Regressor: 0.245
   - E. Elastic Net Regressor: 0.306

7. R2 Score of test dataset:
   - A. Linear Regression: 0.666
   - B. Decision Tree Regressor: 0.863
   - C. Lasso Regressor: 0.544
   - D. Ridge Regressor: 0.666
   - E. Elastic Net Regressor: 0.480

8. Adjusted R2 of test dataset:
   - A. Linear Regression: 0.666
   - B. Decision Tree Regressor: 0.863
   - C. Lasso Regressor: 0.545
   - D. Ridge Regressor: 0.666
   - E. Elastic Net Regressor: 0.480

Thank You!