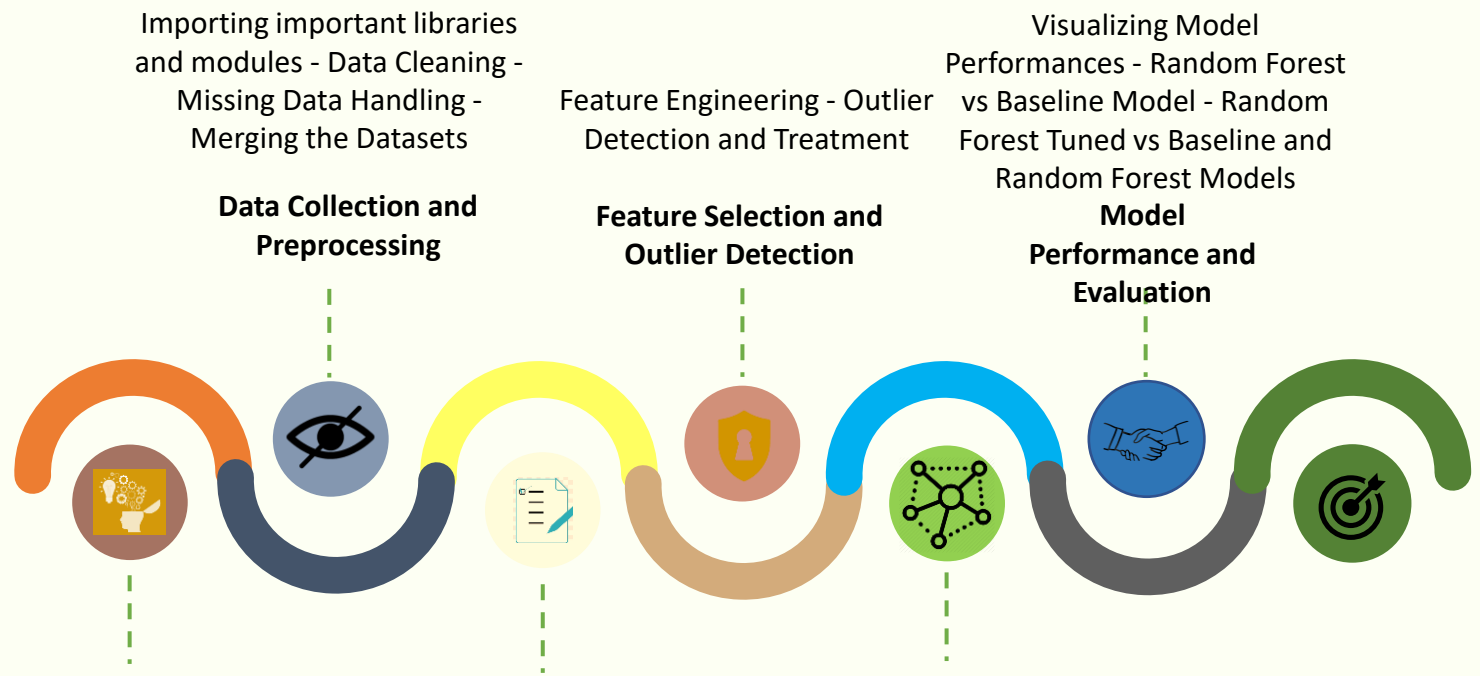
Capstone Project -2

Supervised Machine Learning – Regression

Jouher Lais Khan

Approach

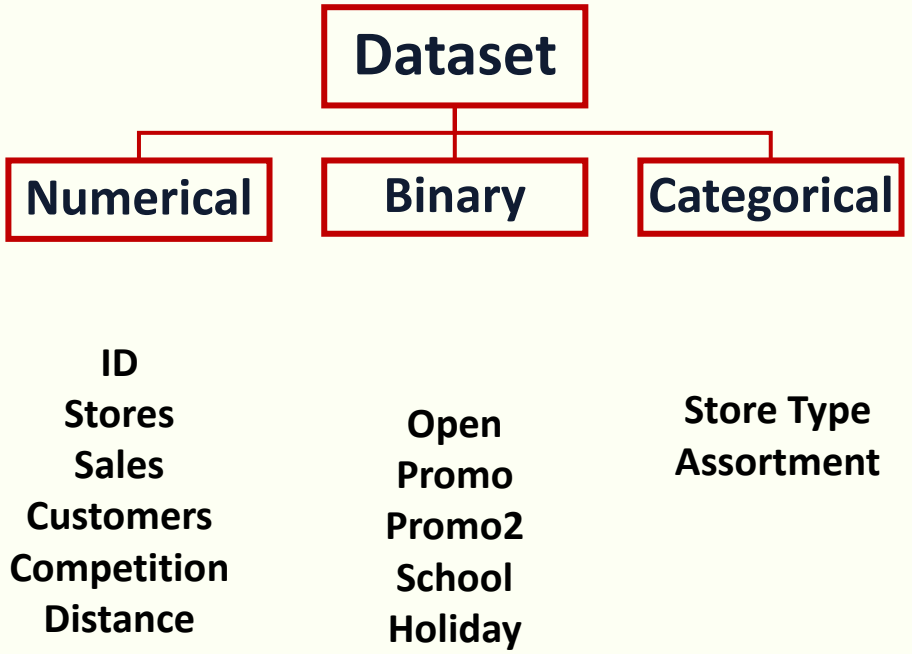

Importing important libraries and modules - Data Cleaning - Missing Data Handling - Merging the Datasets
Data Collection and Preprocessing
Feature Engineering - Outlier Detection and Treatment
Feature Selection and Outlier Detection
Visualizing Model Performances - Random Forest vs Baseline Model - Random Forest Tuned vs Baseline and Random Forest Models
Model Performance and Evaluation
Business Problem
Exploratory Data Analysis
Categorical Features - Continuous Features - EDA Conclusion
Modelling Train-Test Split
Linear Regression(OLS) - Lasso, Ridge, Elastic Net with Hyperparameter Tuning . Baseline Model - Decision Tree, Random Forest Model
Conclusion and Recommendations

Problem Description

1. 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.
2. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality.
3. With thousands of individual managers predicting sales based on their unique circumstances, the accuracy of results can be quite varied.
4. 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

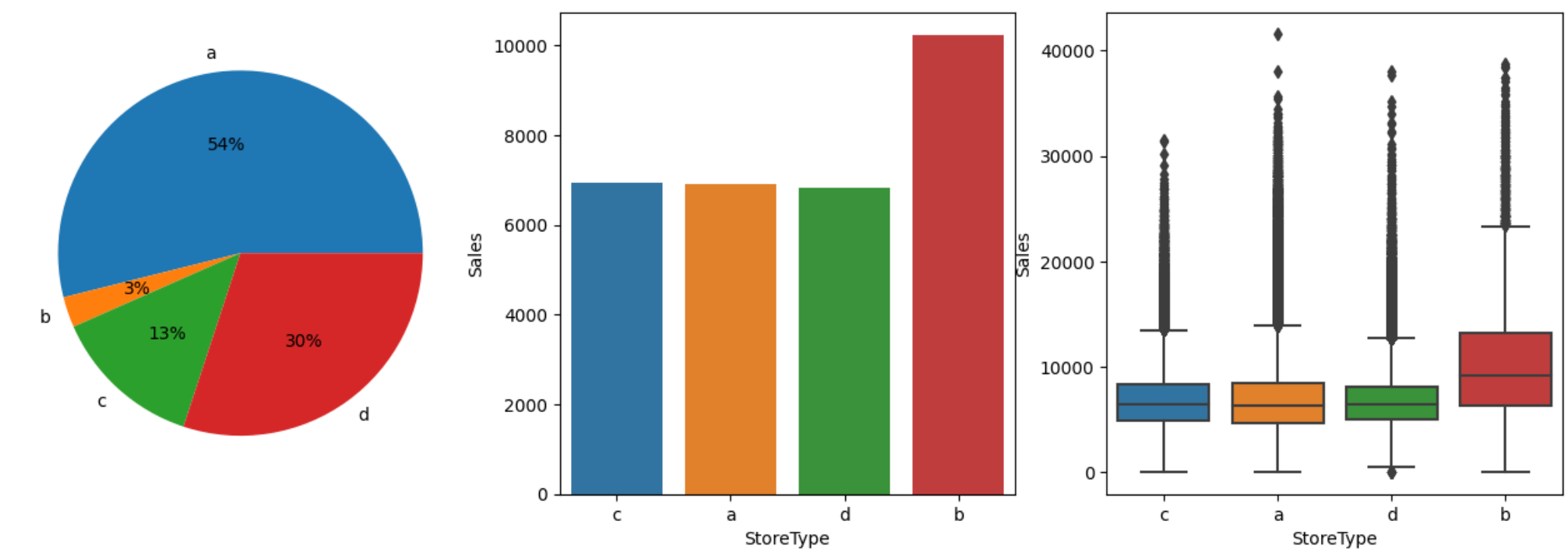
- **Id** - an Id that represents a (Store, Date) duple within the set
- **Store** - a unique Id for each store(Integer)
- **Sales** - the turnover for any given day (Dependent Variable)
- **Customers** - the number of customers on a given day
- **School Holiday** - indicates if the (Store, Date) was affected by the closure of public schools
- **Store Type** - differentiates between 4 different store models: a, b, c, d
- **Assortment** - describes an assortment level: a = basic, b = extra, c = extended..
- **Competition Distance** - distance in meters to the nearest competitor store
- **Promo** - indicates whether a store is running a promo on that day
- **Promo 2** - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating



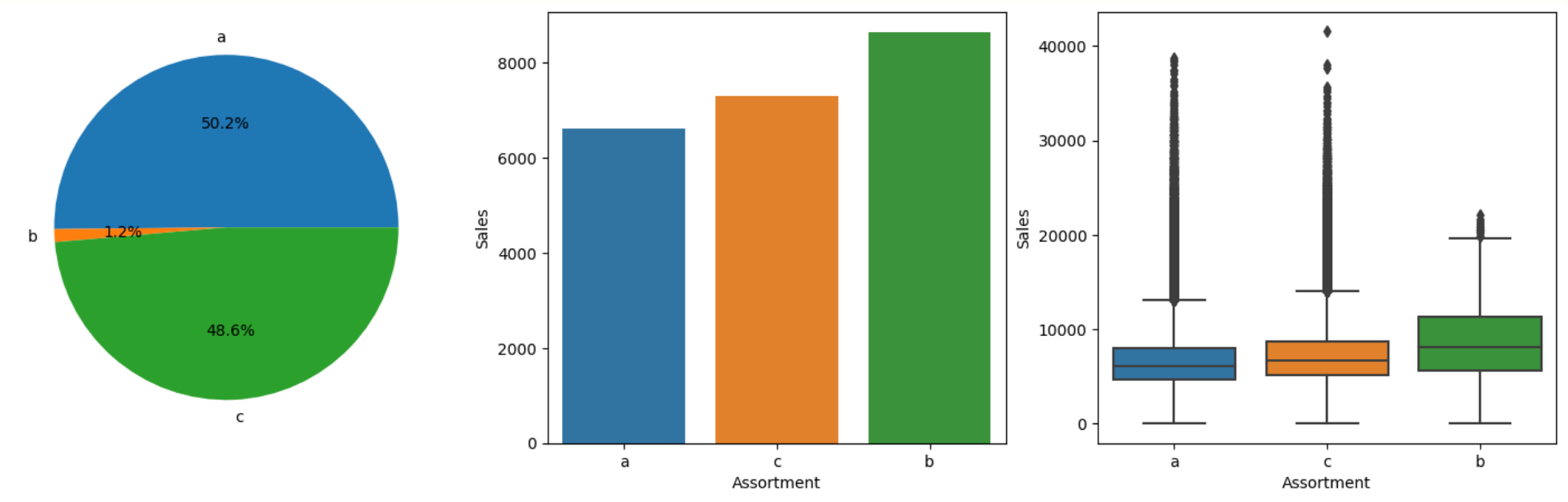
Data Processing

1. Since, when Store are closed there will be no sales, row where store are closed are dropped.
2. Competition Open Since, Promo 2 Since Week, Promo 2 Since Year, Promo Interval, Competition Open Since Month, Competition Open Since Year contains very number of null values so they are dropped.
3. State Holiday of category a, b, c contains only 0.1% of the dataset so rows where State Holiday are a, b, c are dropped.
4. Adding new columns such as week of year, Day, Month, Year.

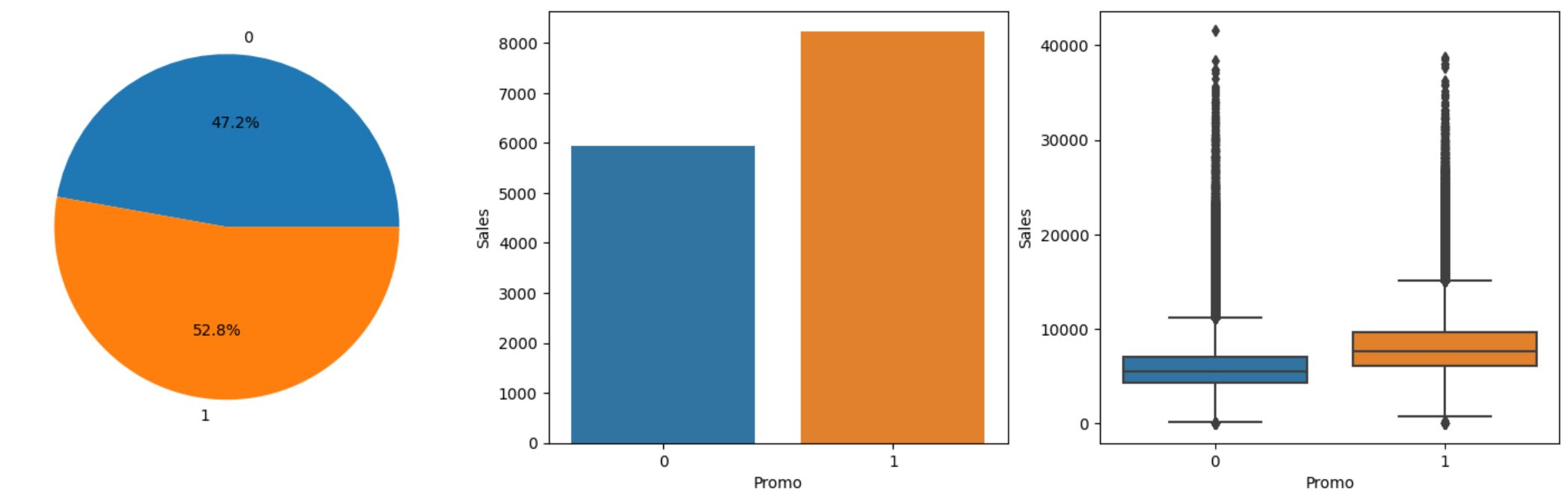
Exploratory Data Analysis
Store Type



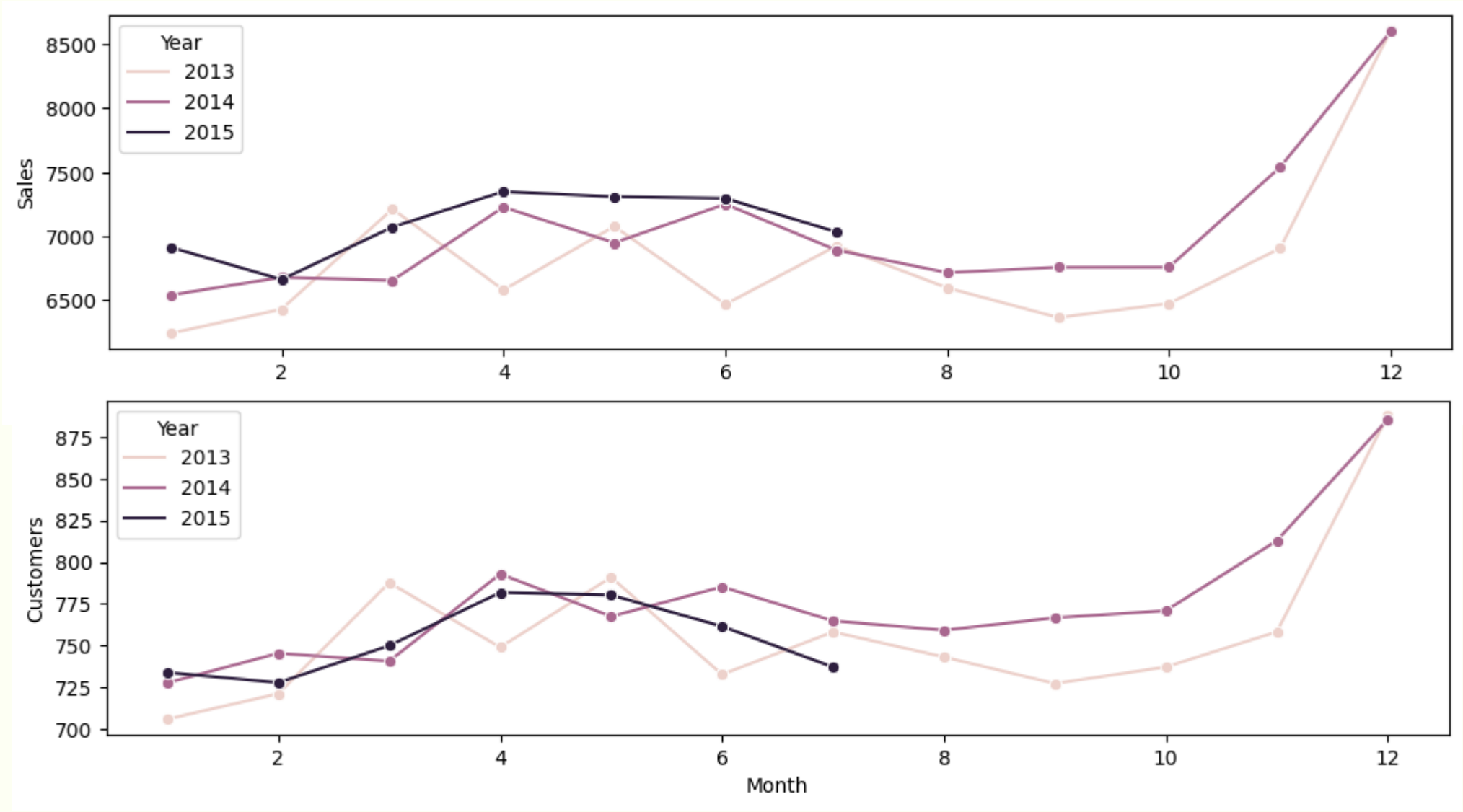
Assortment


a
50.2%
b
1.2%
48.6%
c
Sales
8000
6000
4000
2000
0
a
c
b
Assortment
Sales
40000
30000
20000
10000
0
a
c
b
Assortment

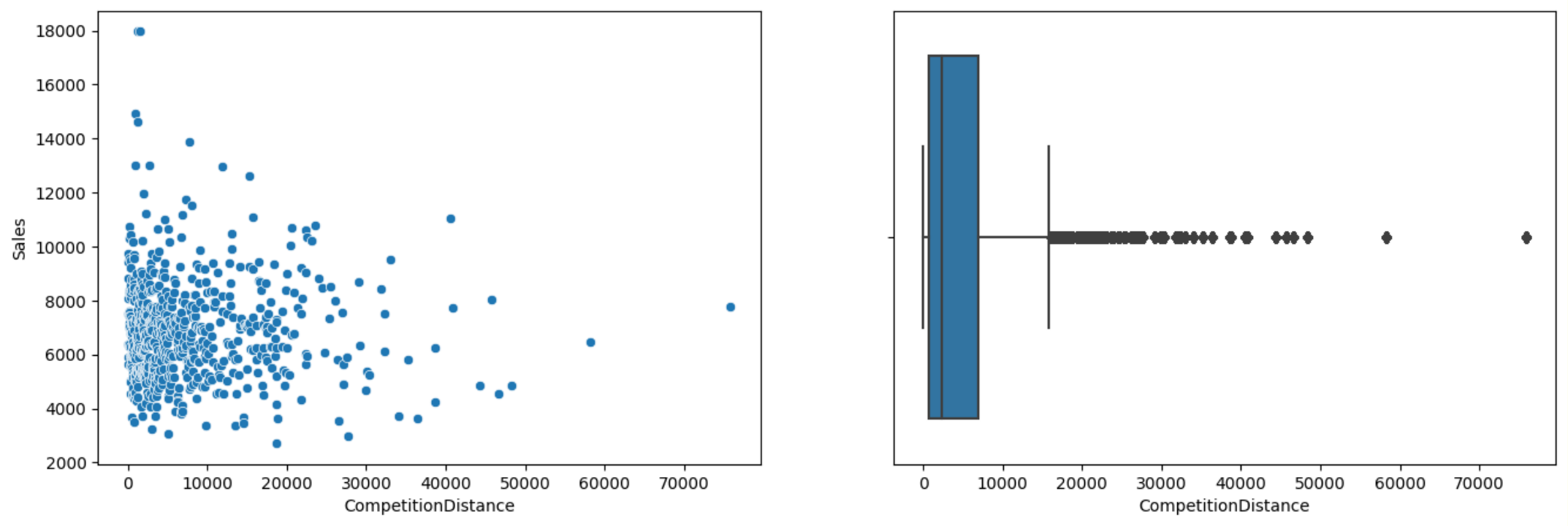
Promo


0
47.2%
52.8%
1
Sales
Promo
0
1000
2000
3000
4000
5000
6000
7000
8000
0
1
Sales
Promo
0
10000
20000
30000
40000
0
1

Month vs Sales, Customers

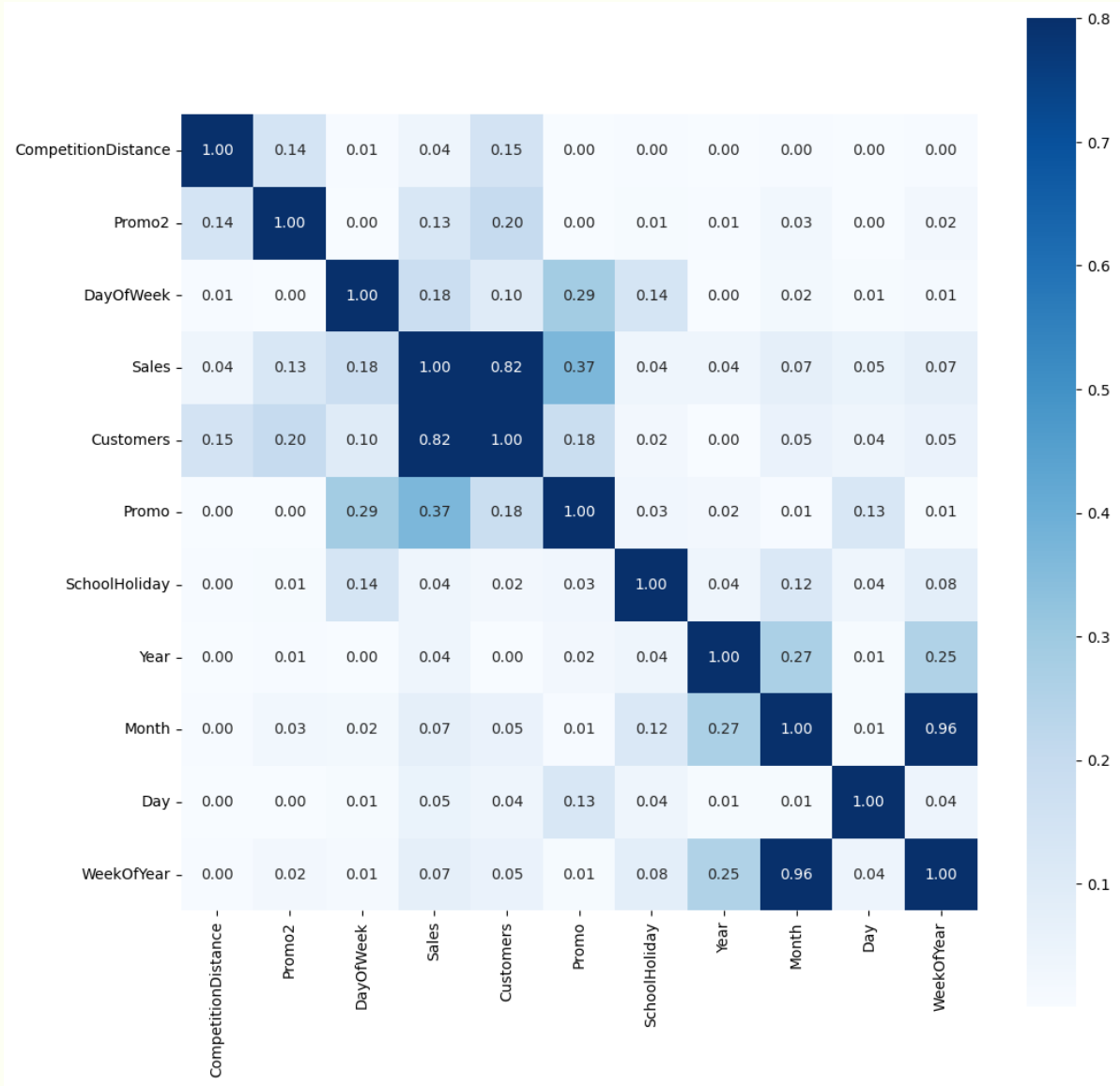


Competition Distance vs Sales



Correlation Map

Month has 0.96 correlation with week of year, this is because weeks of year is like subset of months of year.
Sales has highest correlation with the Customers



Multi Correlation(VIF)

Month, Year, and Week Of Year has very VIF , so we will remove month and year.
After removing month, year we can see all the VIF is lower than 5.

	variables	VIF
0	Competition Distance	1.572305
1	Promo2	2.151655
2	Day Of Week	5.769767
3	Customers	5.189010
4	Promo	2.074781
5	School Holiday	1.307653
6	Year	22.659440
7	Month	57.066157
8	Day	4.584708
9	Week Of Year	51.265398

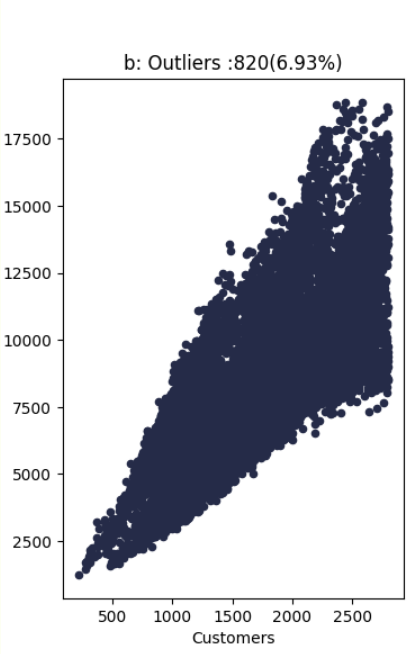
	Variables	VIF
0	Competition Distance	1.458452
1	Promo2	1.879426
2	DayOfWeek	3.891415
3	Customers	3.939596
4	Promo	1.874417
5	School Holiday	1.253272
6	Day	3.623503
7	Week Of Year	3.382762

EDA Conclusion

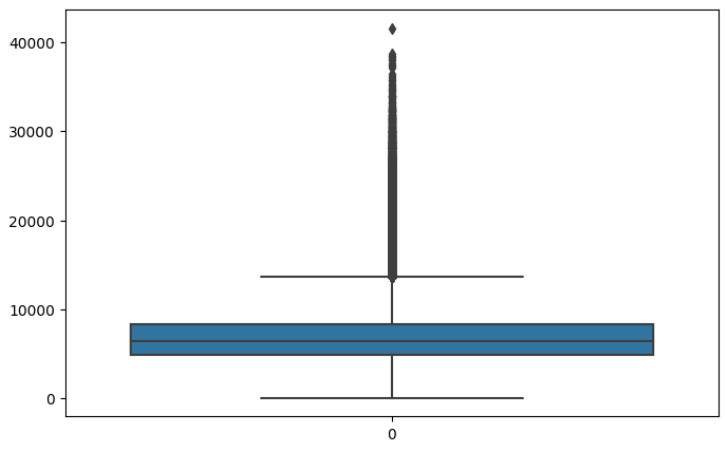
1. Mondays have most sales since most of the Sundays are closed.
2. Promotions seem to have a significant effect on sales but not for the number of customers.
3. Despite school holidays comprising only 19% of the total data points, the average sales during school holidays surpass those during no holidays.
4. Promo1 does not have significant role on sales
5. It is advisable to spend more on promos to get higher returns.
6. Store type b has higher sales and customers per store than other store types . More Store type b must be opened.
7. Assortment b is available only at store type b and it has more sales and customers thanany other assortment.
8. More assortment b must be stocked to meet the demands of customers.
9. Weekly sales and customers peak at the mid-December. It may be guessed that people buy drugs in advance just before the shops close for the holiday season.
10. In cases where there is less competition distance, it appears that sales values tend to be higher. This might be attributed to the possibility that in areas with higher demand, multiple stores are situated.
11. Sales are highest during December , this is because of Christmas and in this month harshest winter start in Europe so more people become sick.

Outliers Removal

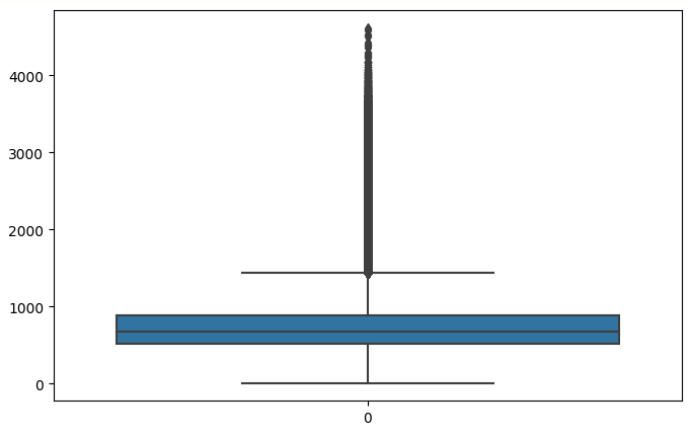
1. Numerical Feature like Sales, Competition Distance, Customers contains high number of outliers. We have removed extreme outliers up to 1% from them.
2. In Categorical Feature only Store Type of category b have outliers more than 5%.



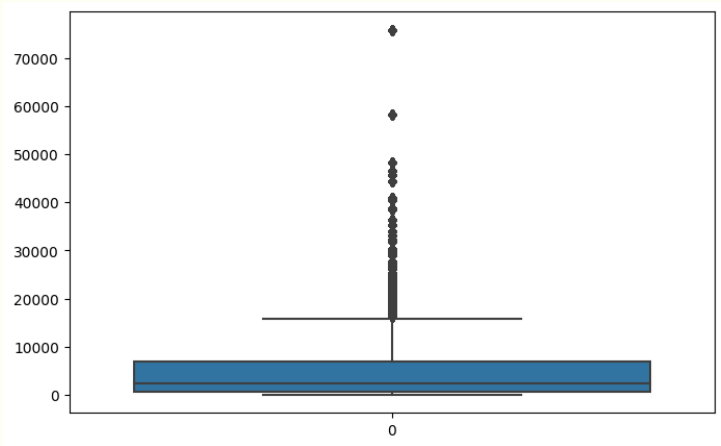
Store Type (b)



Sales



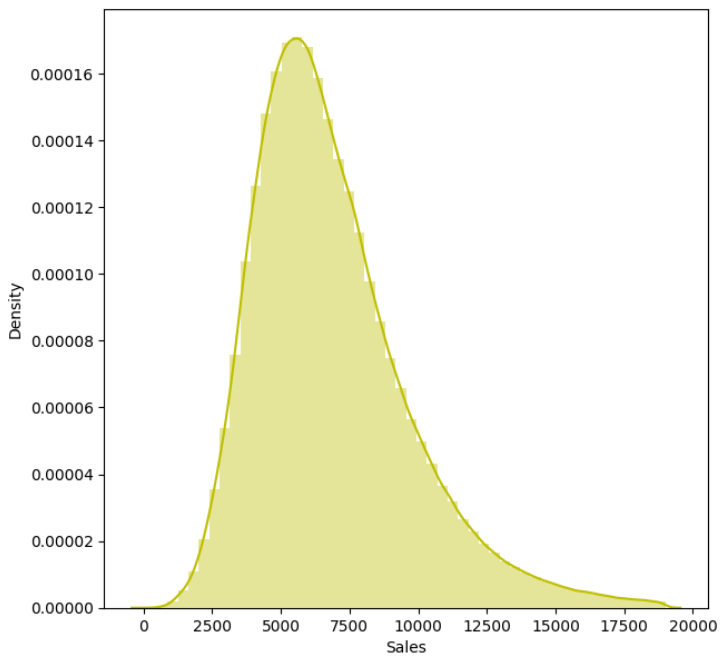
Customers

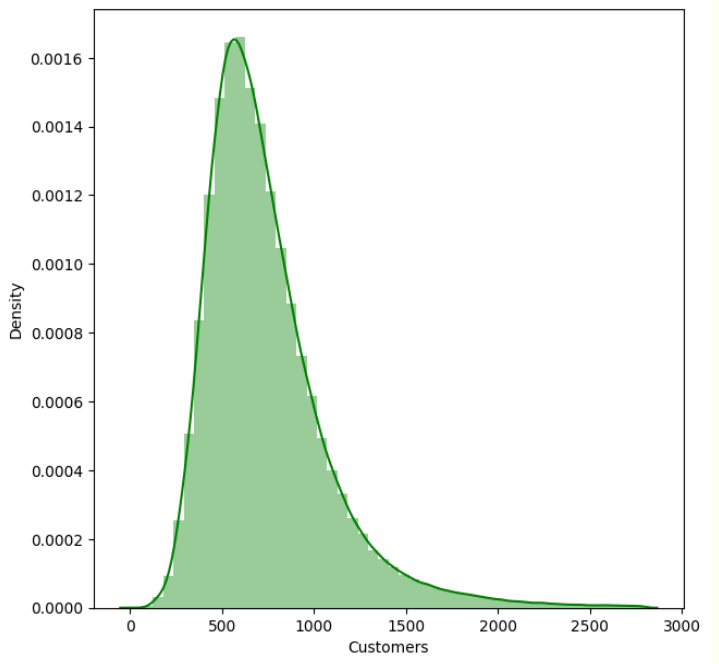


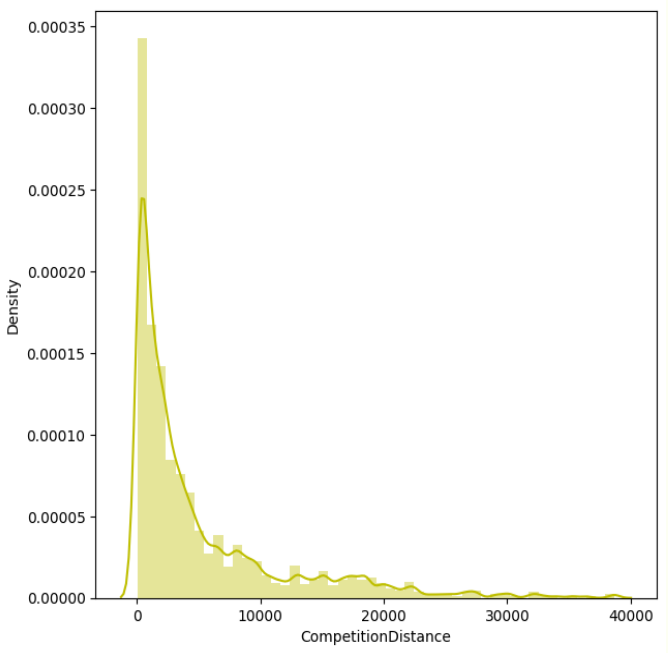
Competition Distance

Feature Transformation

Sales, Customers and Competition Distance are positively skewed, so we will apply log transformation to all of this features as they do not contains 0 values.

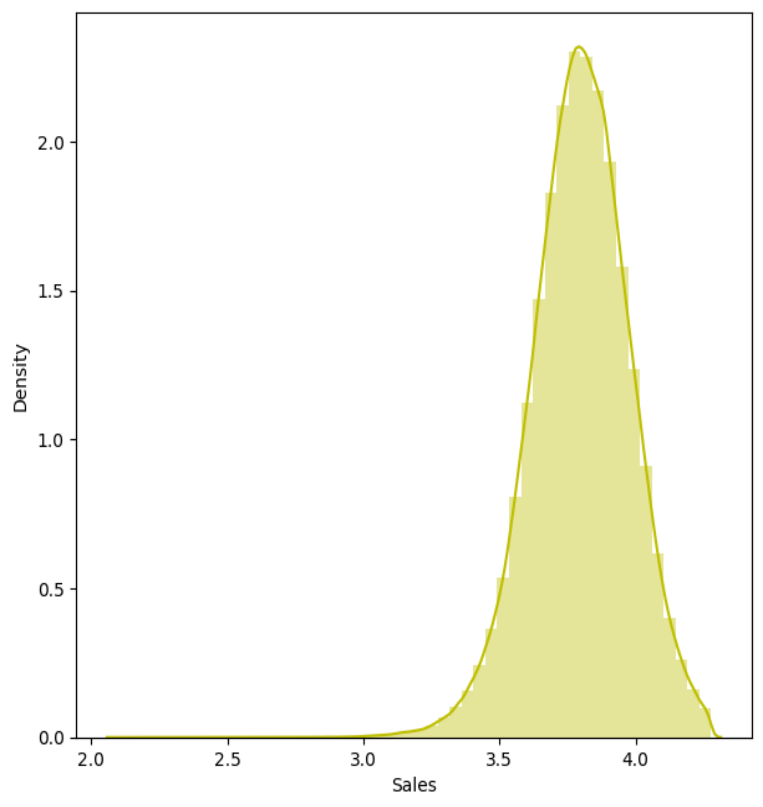


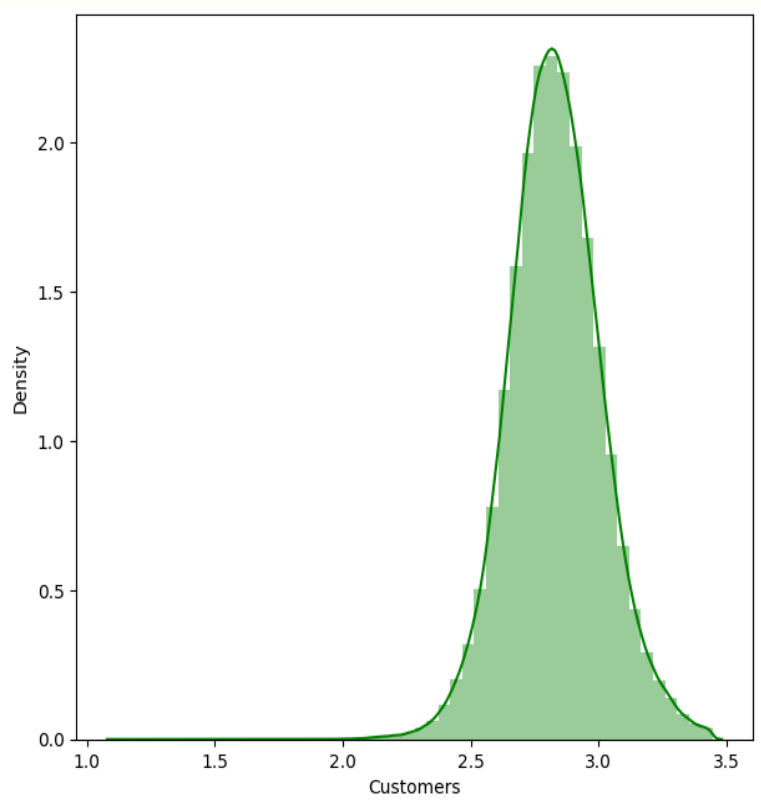


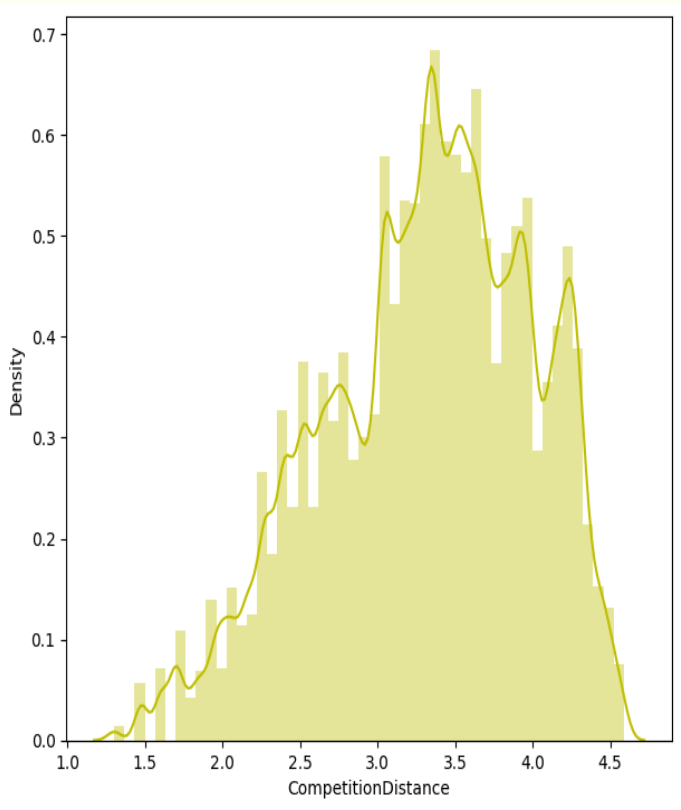


Feature's after log transformation.

Customers, Sales transformation are nearly normal and skewness are mostly removed.
Competition Distance skewness are removed but transformation is distorted normal distribution.







Feature Encoding

1. Creating copy of Data Frame for Modelling.
2. Creating Dummy Variables of Categorical columns.
3. Creating list of final features which will be used in modelling.
4. Creating Sales as dependent variables and features as independent variable.
5. Train-Test Split

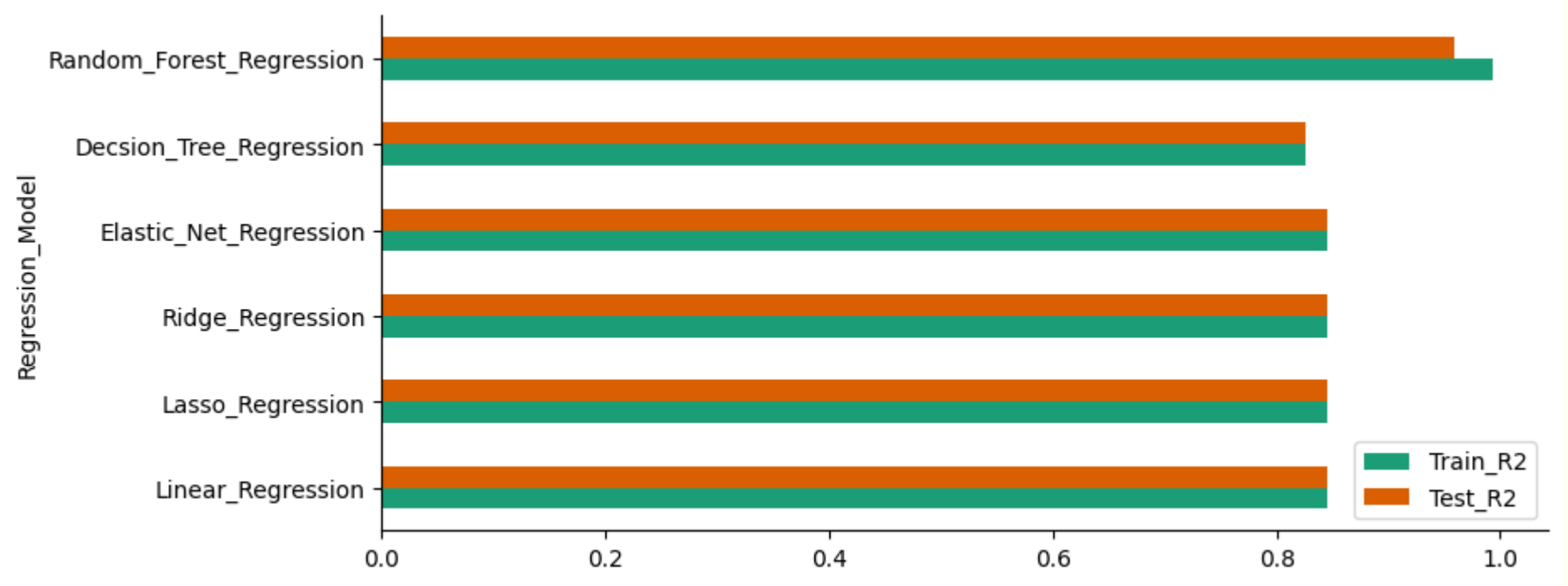
Models Implemented

1. Linear Regression (OLS)
2. Lasso Regression with Hyperparameter Tuning
3. Ridge Regression with Hyperparameter Tuning
4. Elastic Net Regression with Hyperparameter Tuning
5. Decision Tree Regression with Hyperparameter Tuning
6. Random Forest Regression

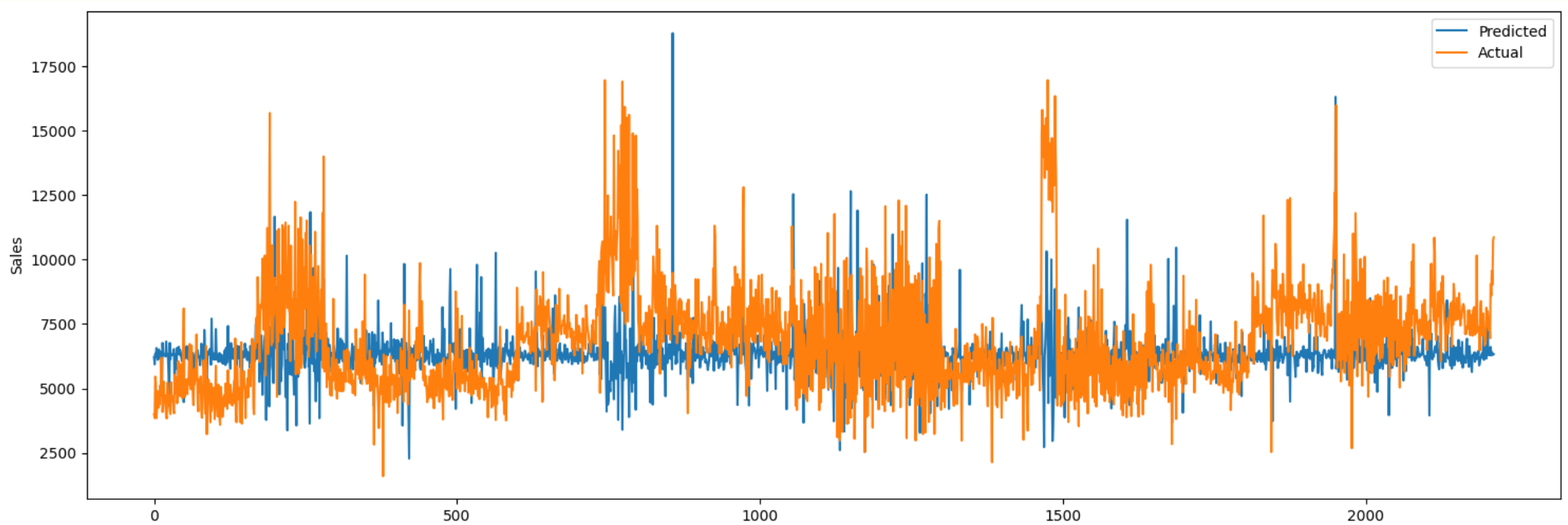
Model Performance

	Regression Model	Train MSE	Train RMSE	Train R2	Train Adjusted R2	Test MSE	Test RMSE	Test R2	Test Adjusted R2
0	Linear Regression	1.214859e+06	1102.206527	0.844978	0.844975	1.217627e+06	1103.461202	0.844943	0.844929
1	Lasso Regression	1.214485e+06	1102.036953	0.845026	0.845012	1.217348e+06	1103.334986	0.844979	0.844965
2	Ridge Regression	1.214486e+06	1102.037084	0.845026	0.845022	1.217348e+06	1103.335147	0.844979	0.844965
3	Elastic Net Regression	1.214485e+06	1102.036949	0.845026	0.845022	1.217348e+06	1103.334982	0.844979	0.844965
4	Decision Tree Regression	1.362364e+06	1167.203328	0.826156	0.826152	1.367349e+06	1169.337164	0.825877	0.825861
5	Random Forest Regression	5.596015e+04	236.558981	0.992859	0.992859	3.200686e+05	565.746016	0.959241	0.959238

Model Performance



Predicted vs Actual Sales



Conclusions of Modelling

1. The linear regression model is least accurate as it has very high coefficient of Assortment categories and Store type categories and it neglected features like customers , promotions which has positive correlation with sales , so we will use hyperparameter tuning to impose penalties on coefficients.
2. Decision Tree Model density distribution plot of sales varies highly with real data of sales.
3. Random Forest Regression has 99% accuracy for train data but 96% for test data, so this type of model cant be trusted , as the difference between train -test is very high
4. The most accurate models are Ridge , Lasso and Elastic-Net Regression , there train-test performances are almost similar and coefficient's are also similar.
5. The week of year line plot shows that Predicted Sales follows Actual Sales , with variation of mostly 700 dollars , except for last 2 week of the year

Thank You