Capstone Project

Retail Sales Prediction

Content

1. **Wrangling the data**

- Dealing with Null Values
- Converting PromoOpen and competitionOpen in a simplified way
- Creating dummies for some columns
- Checking distribution of data in SALES column, using Skewness
- Checking distribution of different features, using Skewness
- Finalising the data by Scaling

2. **Implementing Regression techniques**

Problem Statement

Sales on a particular day for different stores

Data Summary

Data set name – Retail Sales Prediction

- We have two datasets

- **- Stores having different features – “Store”**
- **- Stores with sales on a particular day – “Rossmann Stores Data”**

Shape of combined Dataset- 1017209 rows, 23 columns

Columns - 'Store', 'DayOfWeek', 'Date', 'Sales', 'Customers', 'Open', 'Promo', 'StateHoliday', 'SchoolHoliday', 'Year', 'Month', 'Day', 'Week', 'WeekOfYear', 'StoreType', 'Assortment', 'CompetitionDistance', 'CompetitionOpenSinceMonth', 'CompetitionOpenSinceYear', 'Promo2', 'Promo2SinceWeek', 'Promo2SinceYear', 'PromoInterval'

Cleaning dataset

We can see only a few columns had null values.

- Column “CompetitionOpenSinceMonth” and “CompetitionOpenSinceYear” had null values – after exploring I got to know that I should replace the null by mode in this case.

- Column “CompetitionDistance” had null values – after exploring I got to know that I should replace the null by median in this case.

- Column “Promo2SinceWeek”, “Promo2SinceYear”, “PromoInterval” was having a lot of Null values, because those stores have not started any promotion, so they should be zero for our Dataset.

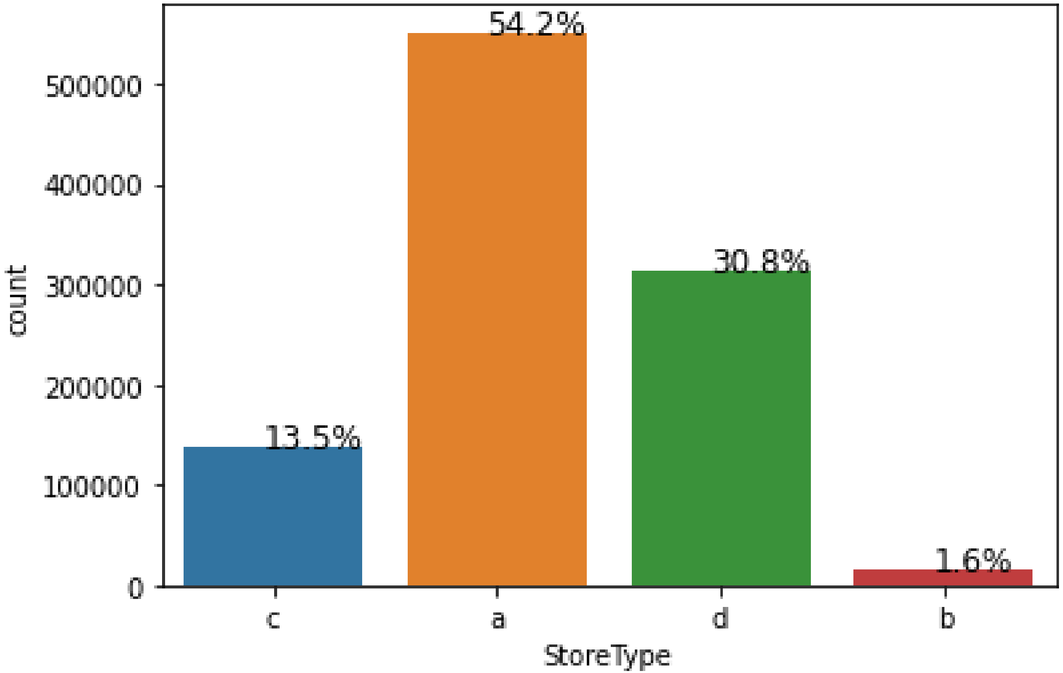
- I have split the “Date” into Month, Year, week of year, day, and Week

- Finally I have merged the two Datasets into one, named “df”

- Other columns are already cleaned with no null values

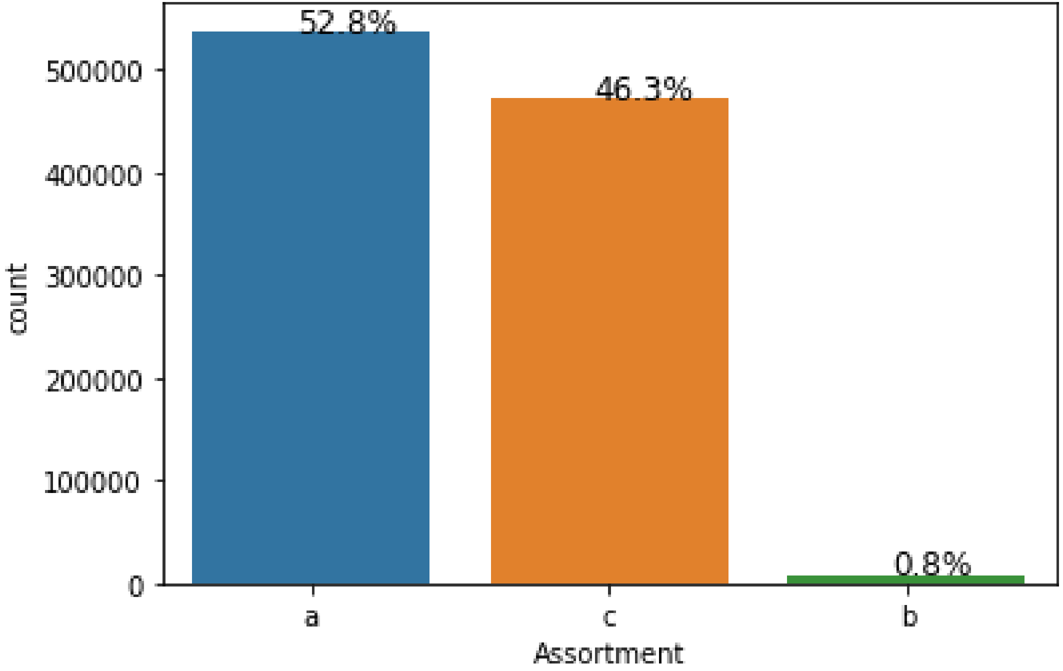
Stores

There are 4 different type of stores among which 54% stores are of type – a which is maximum, and the least is type - b



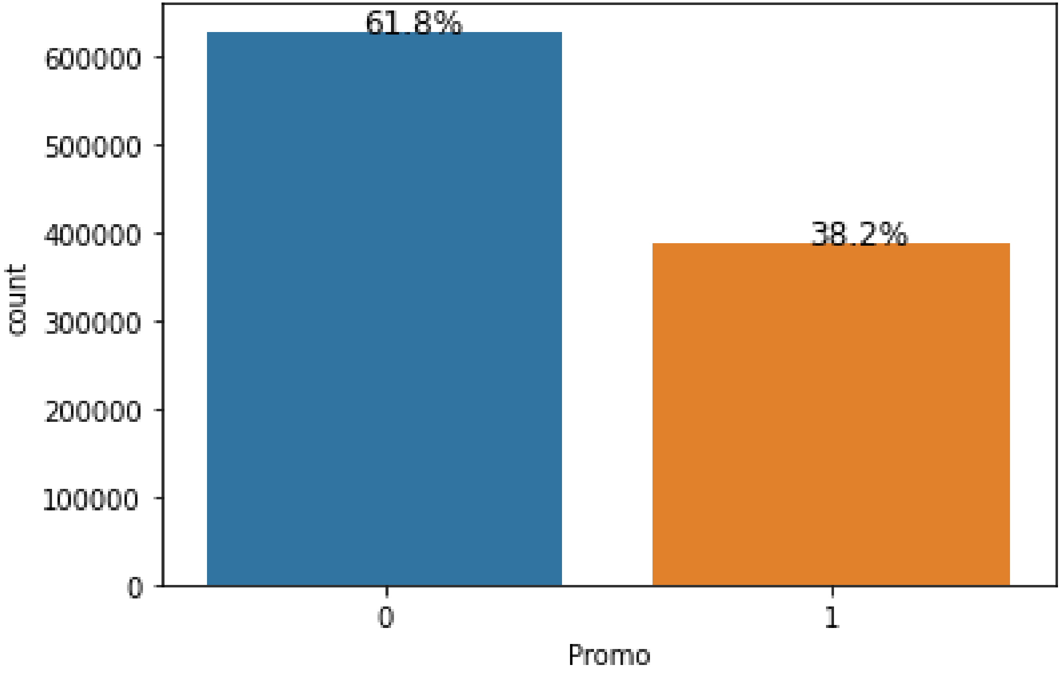
Stores with assortment level

There are 3 different type of assortment level among which 52% stores are of assortment type – a which is maximum, and the least is type - b



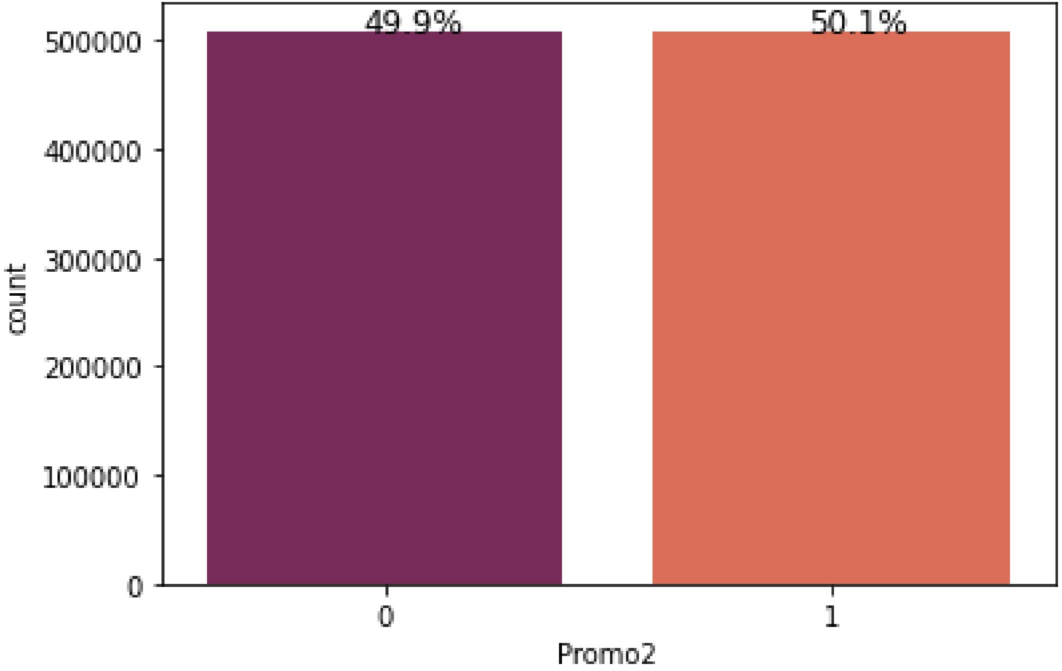
Stores

There are 1115 different stores among which 38% stores are running promo and 62% are not



Stores

From these 38% i.e. 424 stores, 50% are having promo in continuation.
212 stores have started the 2nd onwards round



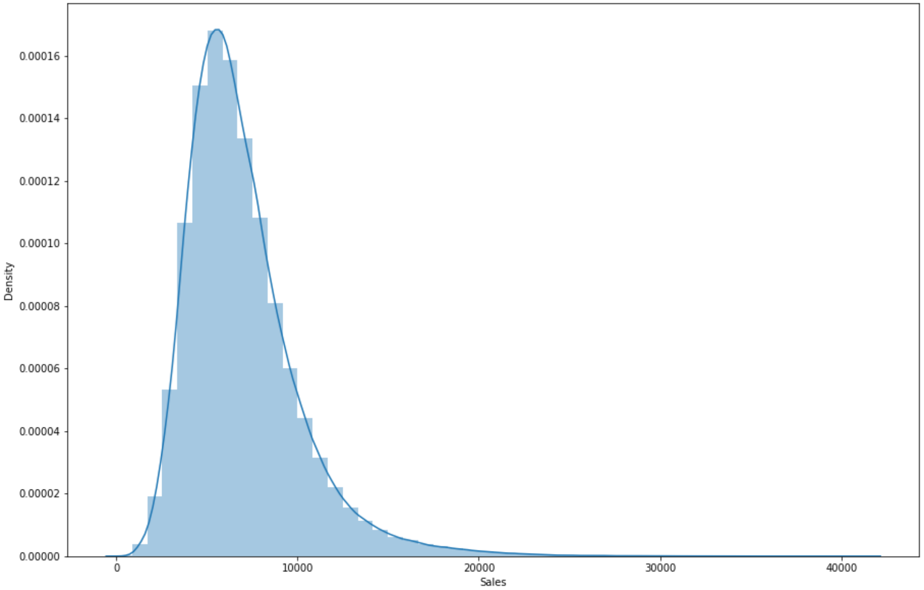
Average Sales by Store

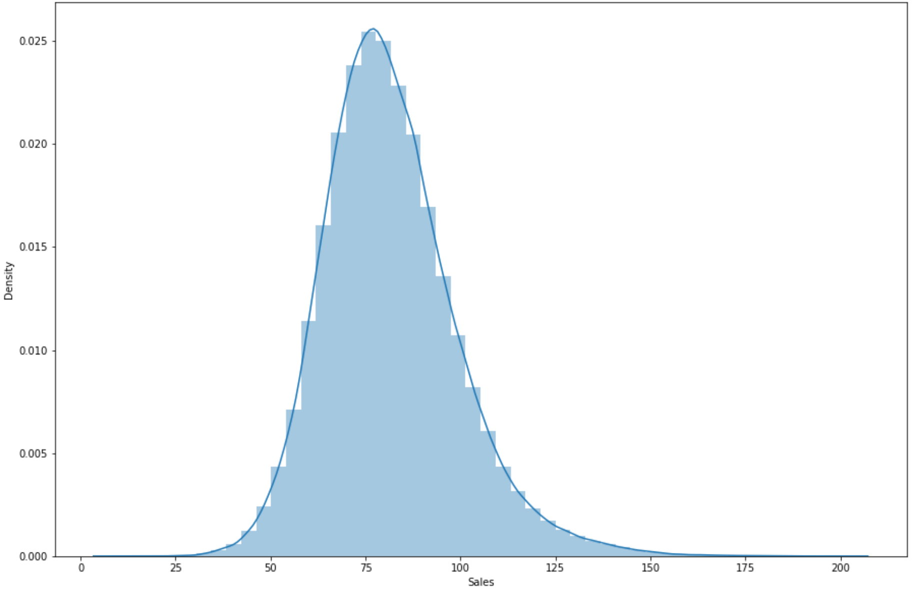
Average Sales by store Type –a : 5738
Average Sales by store Type –b : 10058
Average Sales by store Type –c : 5723
Average Sales by store Type –d : 5641

So we can say maximum sales by store type “b”, but also the number of store with type “b” is minimum so we should consider type “a”

Sales distribution

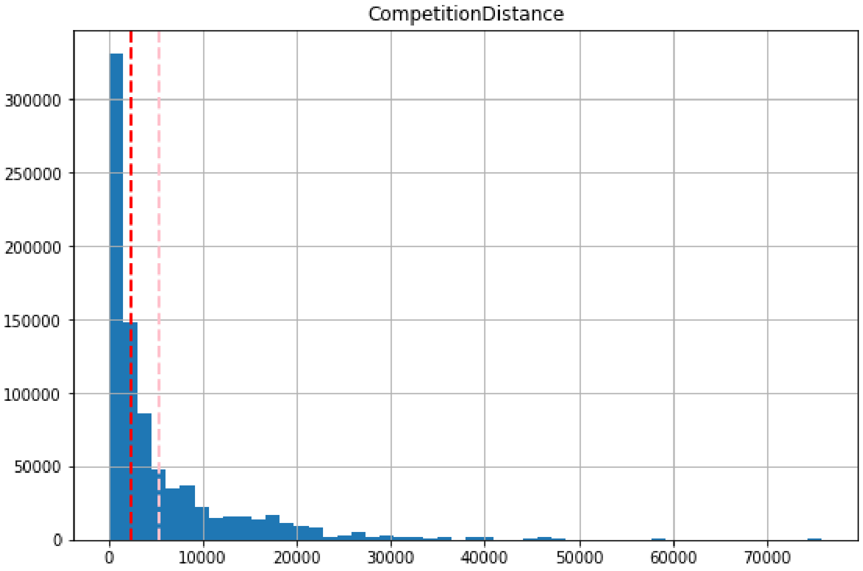
Right skewed changes to approximately normal distribution using Sqrt

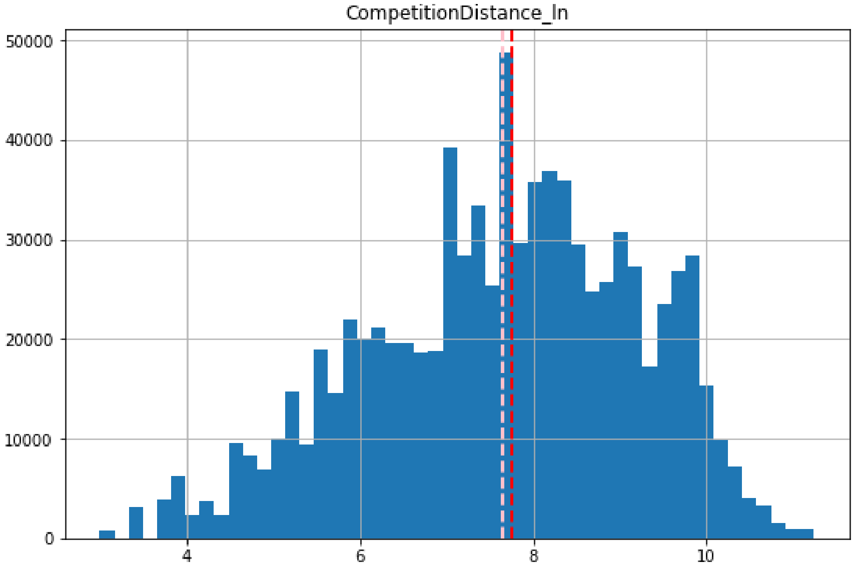




Distribution in independent features

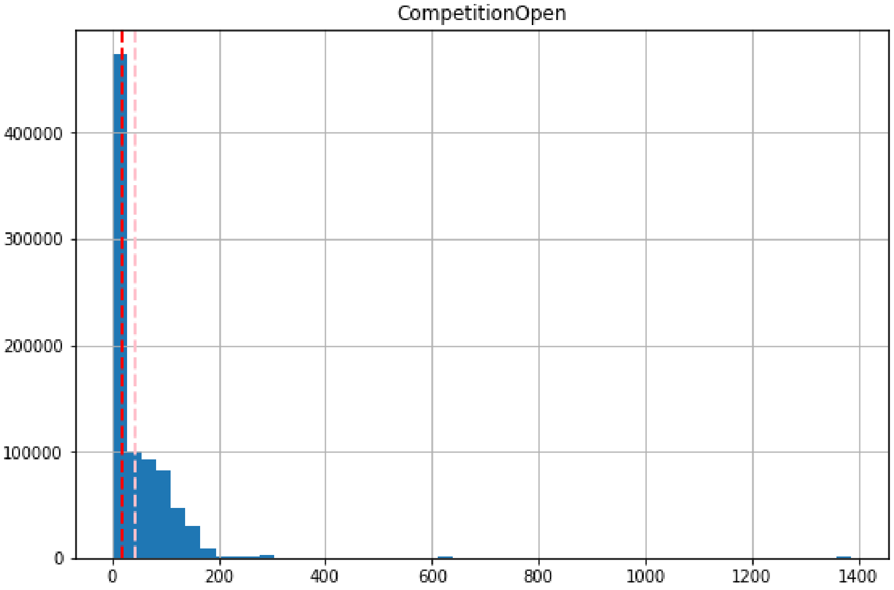
Right skewed changes to approximately normal distribution using Sqrt

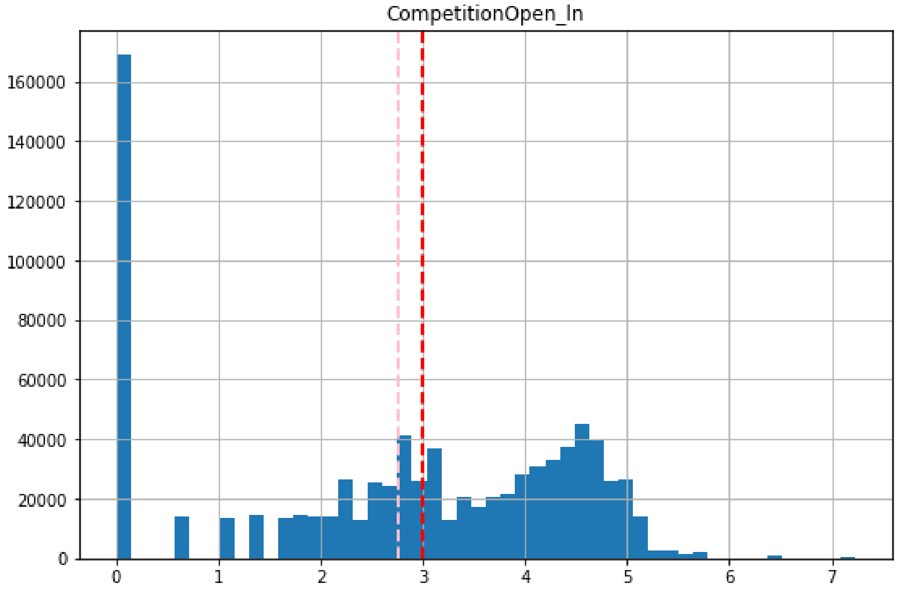




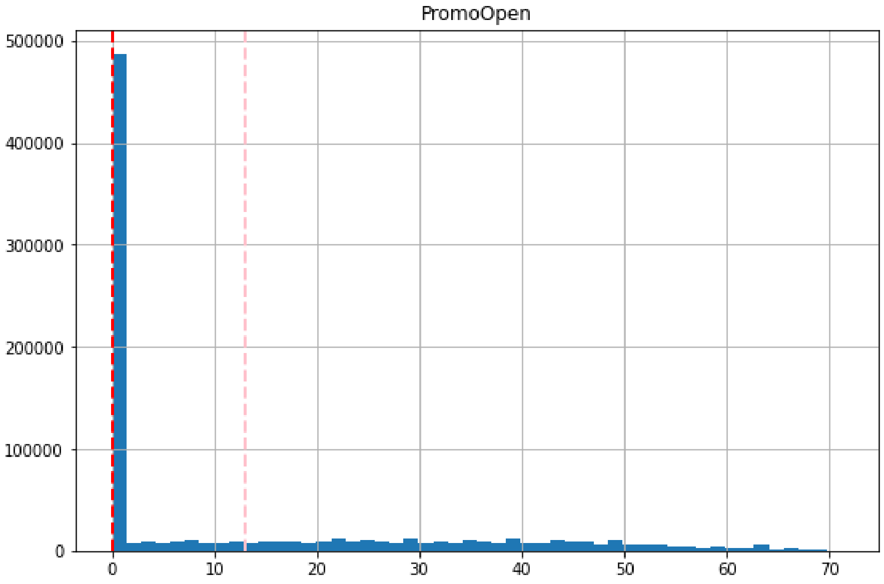
Distribution in independent features

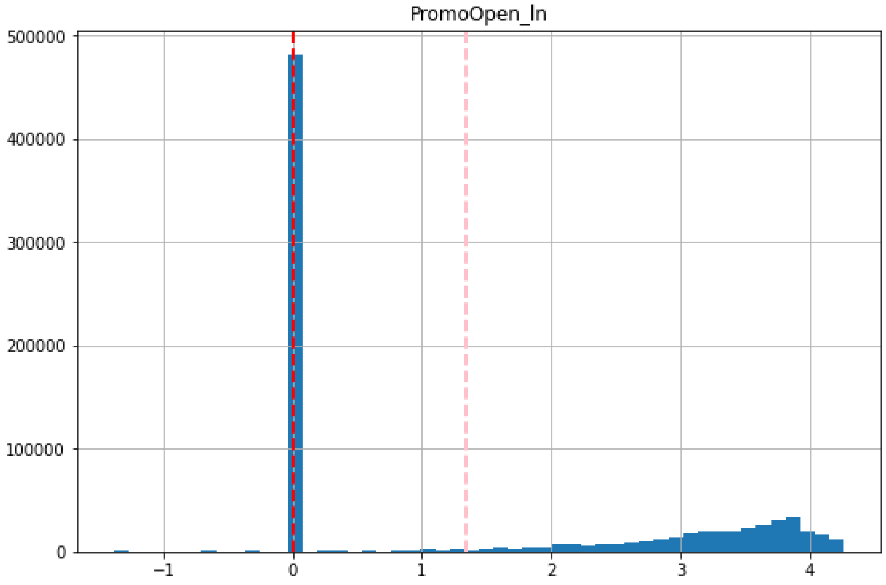
Right skewed but after transformation changes to a bit left



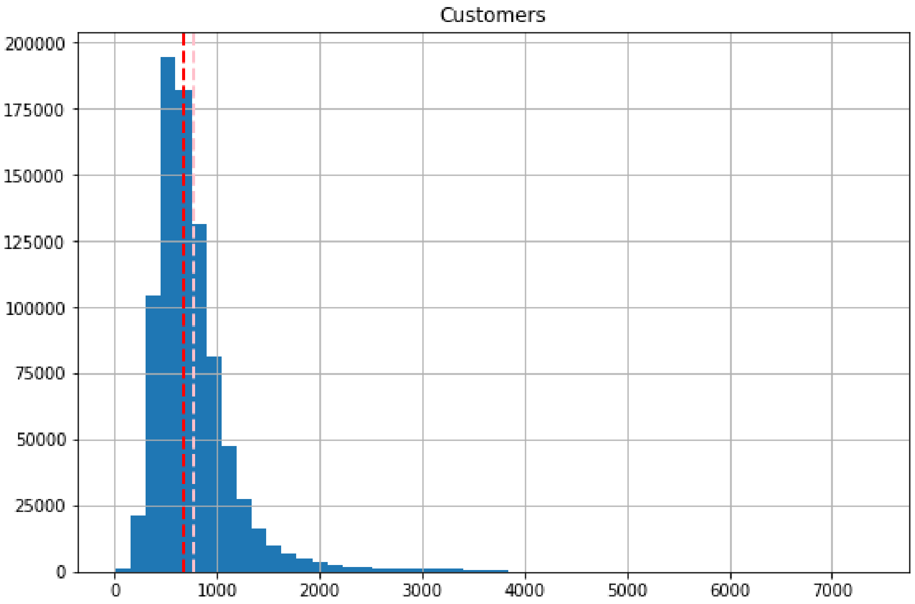


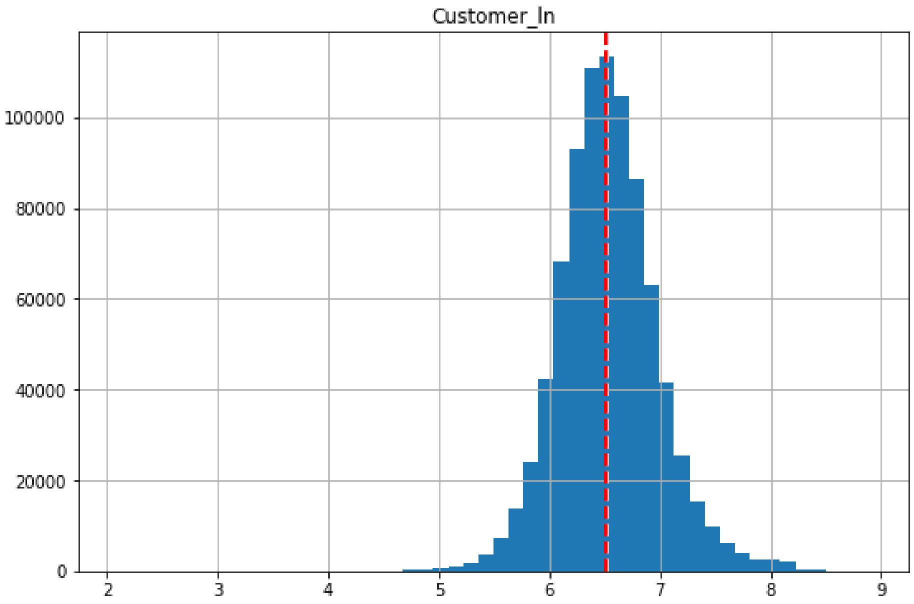
Distribution in independent features





Distribution in independent features



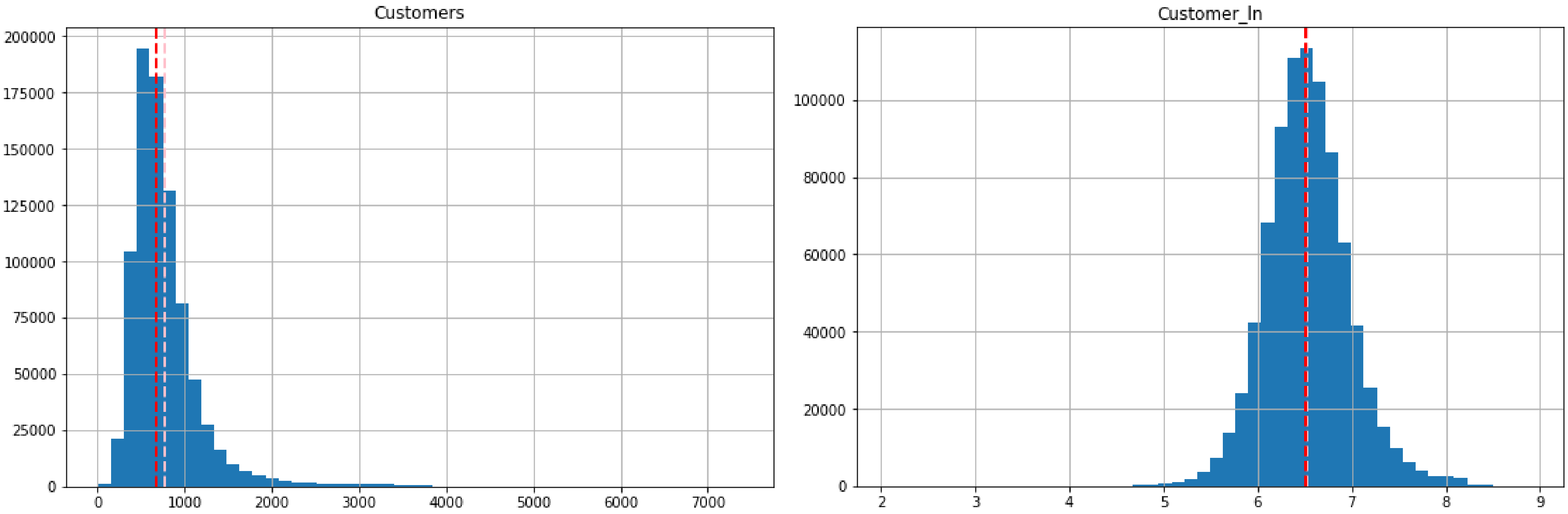


Regressions

Accuracies by different algorithms:-

- **Linear Regression – 82%**
- **Lasso – 73%**
- **Decision Tree – 94.5%**
- **Cross Validation with Decision Tree – 94.2%**

Distribution in independent features



Conclusions

- There are 4 different type of stores among which 54% stores are of type – a which is maximum, and the least is type – b

- There are 1115 different stores among which 38% stores are running promo and 62% are not

- we can say maximum sales by store type “b”, but also the number of store with type “b” is minimum so we should consider type “a”

- From 212 stores, store number - 158, 277, 370, 612, 637, 808, 960 had run the promo for maximum number of months i.e. 71 months

- Store number 815 has a competition from year 1900, so more than 100 years

- We can make these PromoOpen negative values to zeros, because they have not started the promos at that time

- Decision tree is the best for this dataset problem