EDA -
Pharmaceutical Sales Prediction across multiple stores

10 Academy assignment

DATA EXPLORATORY FINDINGS

Niyomukiza Thamar

Motivation/ why doing this

- Data exploration is the initial step in data analysis where the user use data visualization and statistical techniques for describing dataset characterizations to better understand the nature of the data.
- For this case, I plan to perform the following:
    - **Exploration of customer purchasing behavior**
    - Univariate analysis: for continuous variables, build box plots or histograms for each variable independently; for categorical variables, build bar charts to show the frequencies
    - Bi-variable analysis - determine the interaction between variables by building visualization tools
    - Multi-variable analysis
    - Detect and treat missing values
    - Detect and treat outliers
    - Perform Descriptive statistics: e.g., Mean

Understanding the Data

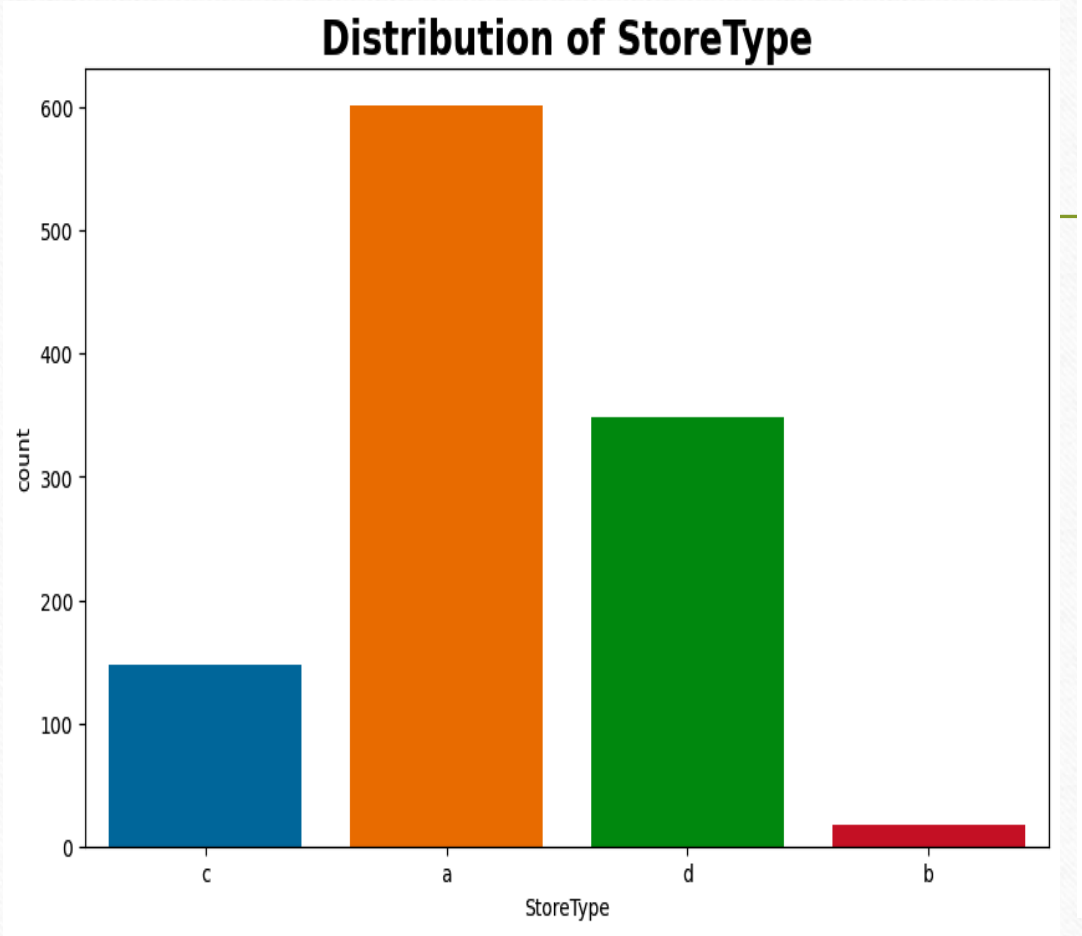
```
round(df.describe().T,2)
```

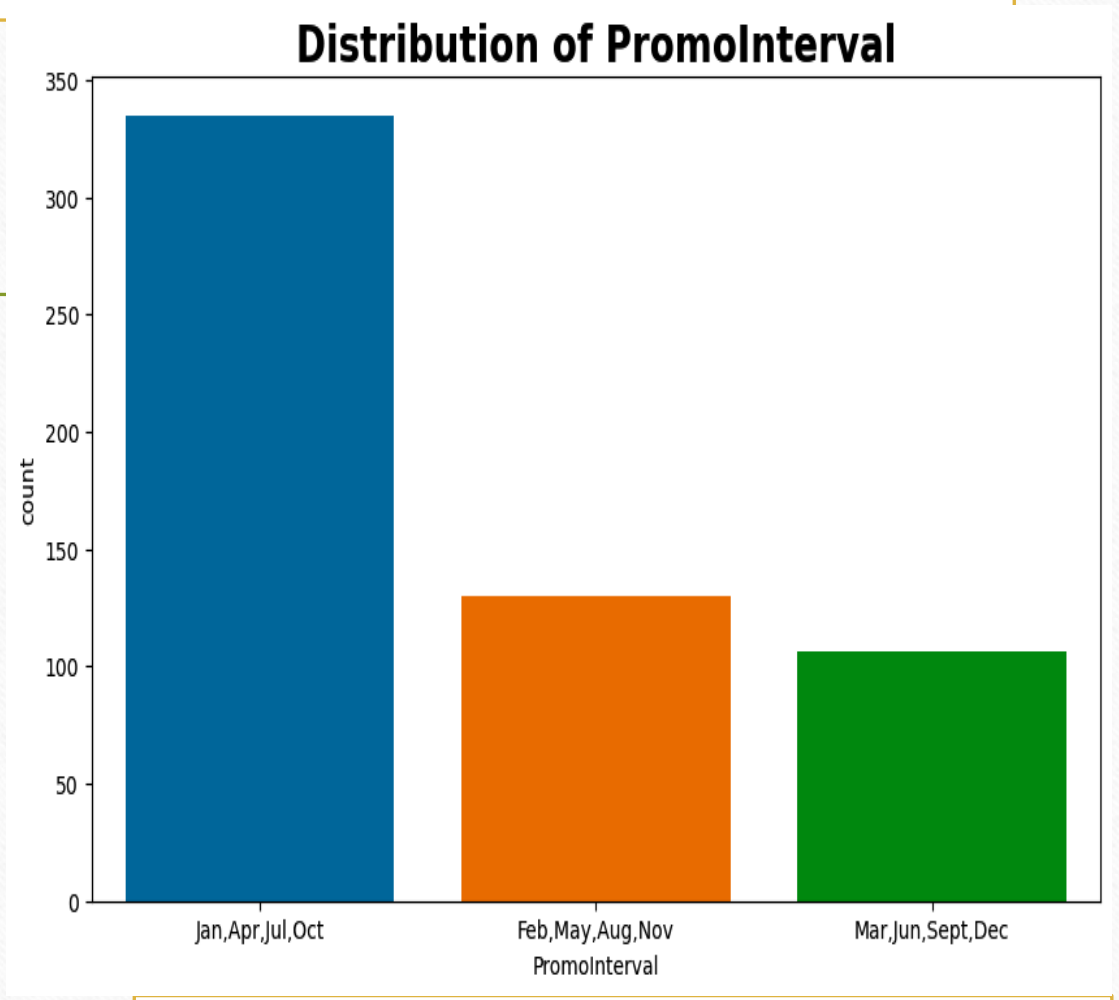
✓ 0.5s Python

	count	mean	std	min	25%	50%	75%	max
Store	1017209.0	558.43	321.91	1.0	280.0	558.0	838.0	1115.0
DayOfWeek	1017209.0	4.00	2.00	1.0	2.0	4.0	6.0	7.0
Sales	1017209.0	5773.82	3849.93	0.0	3727.0	5744.0	7856.0	41551.0
Customers	1017209.0	633.15	464.41	0.0	405.0	609.0	837.0	7388.0
Open	1017209.0	0.83	0.38	0.0	1.0	1.0	1.0	1.0
Promo	1017209.0	0.38	0.49	0.0	0.0	0.0	1.0	1.0
SchoolHoliday	1017209.0	0.18	0.38	0.0	0.0	0.0	0.0	1.0
CompetitionDistance	1014567.0	5430.09	7715.32	20.0	710.0	2330.0	6890.0	75860.0
CompetitionOpenSinceMonth	693861.0	7.22	3.21	1.0	4.0	8.0	10.0	12.0
CompetitionOpenSinceYear	693861.0	2008.69	5.99	1900.0	2006.0	2010.0	2013.0	2015.0
Promo2	1017209.0	0.50	0.50	0.0	0.0	1.0	1.0	1.0
Promo2SinceWeek	509178.0	23.27	14.10	1.0	13.0	22.0	37.0	50.0
Promo2SinceYear	509178.0	2011.75	1.66	2009.0	2011.0	2012.0	2013.0	2015.0

There are in total 1115 stores
- sales feature having 3849.93 volatility and
- feature customers having 464.41 volatility
- with a mean of 57773.82 and 633.15 respectively

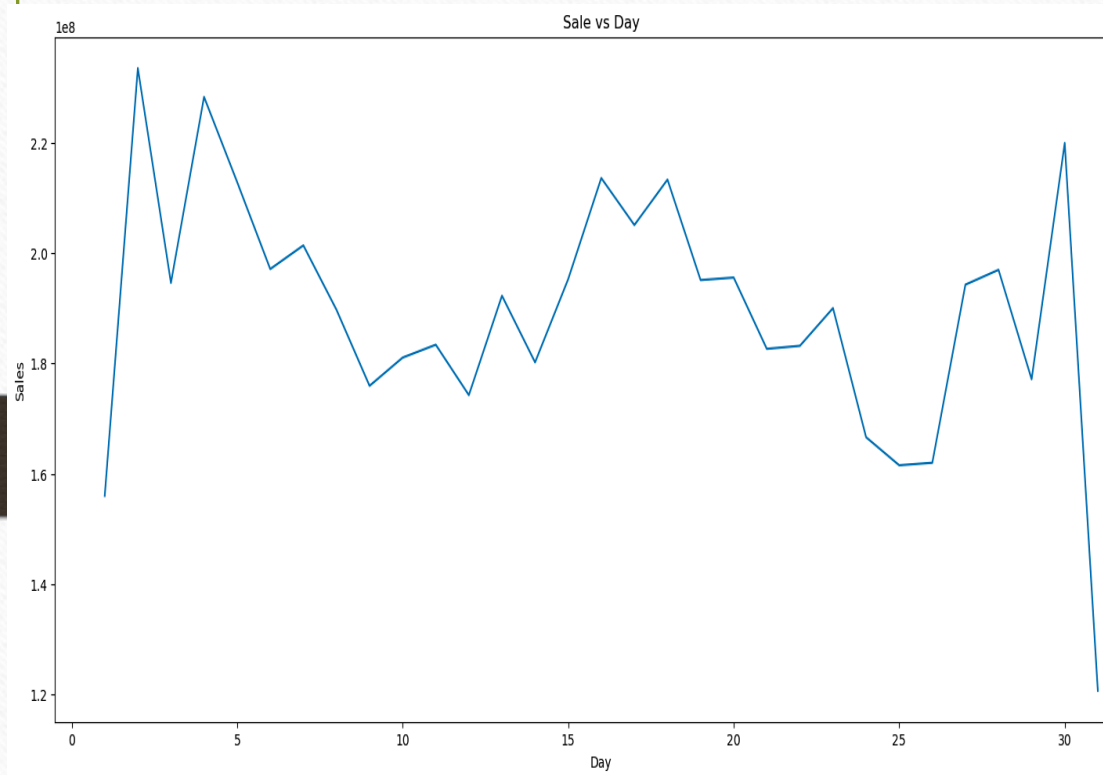
The promotions distribution

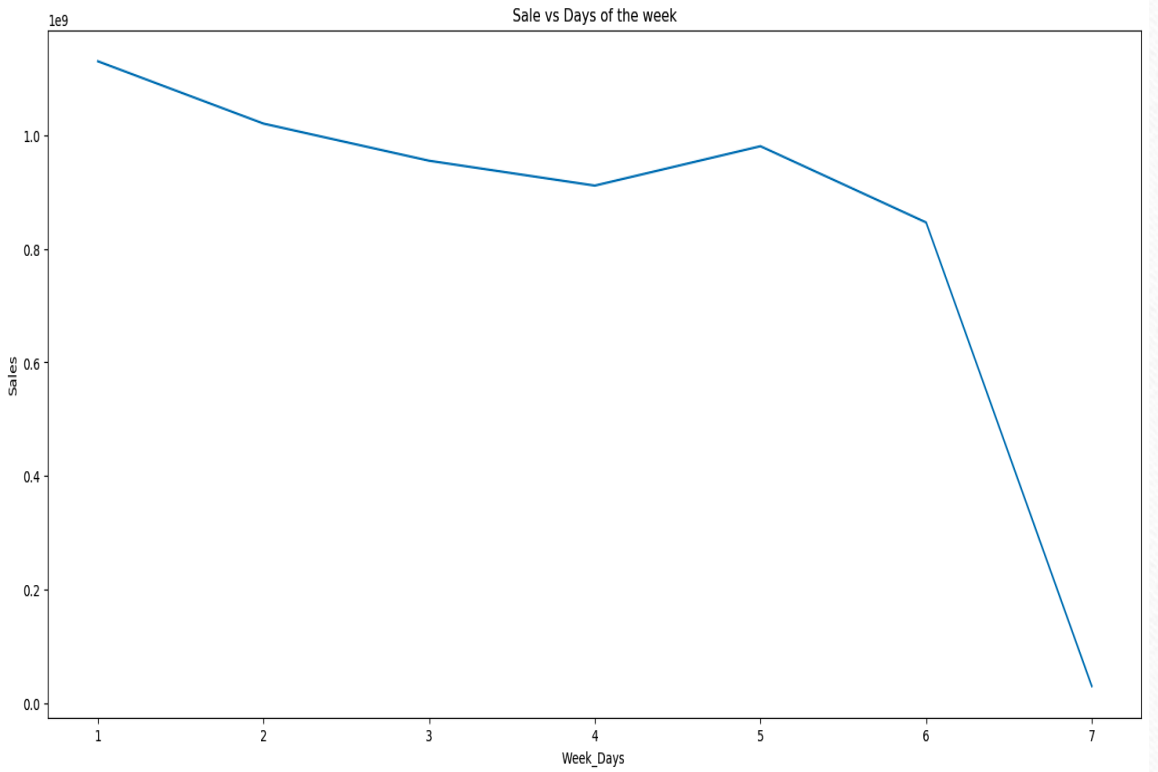




Store a wins the count and Promotioninteval of Jan, Apr,jul,oct also is higher than other periods

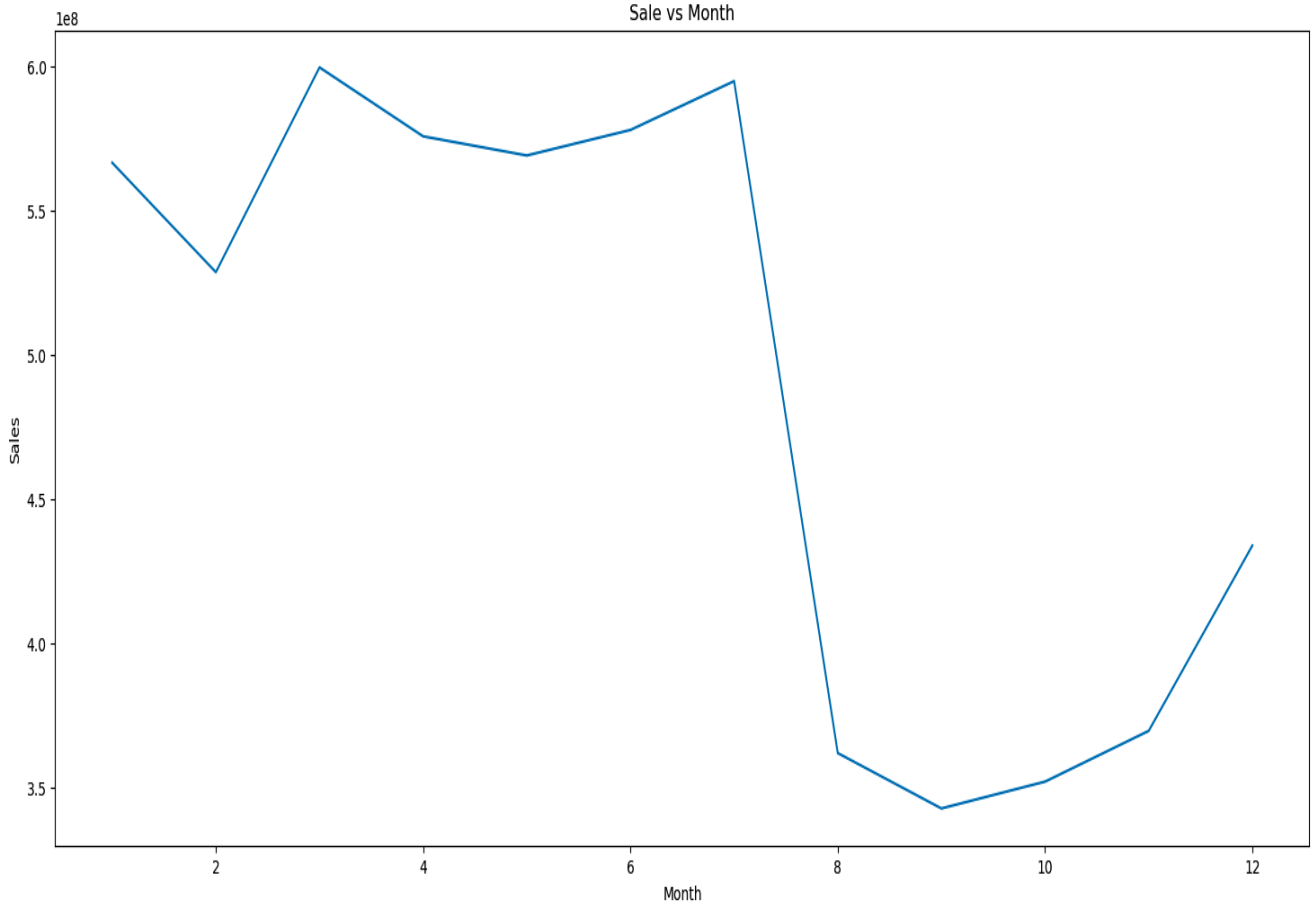
Understand sales behaviors

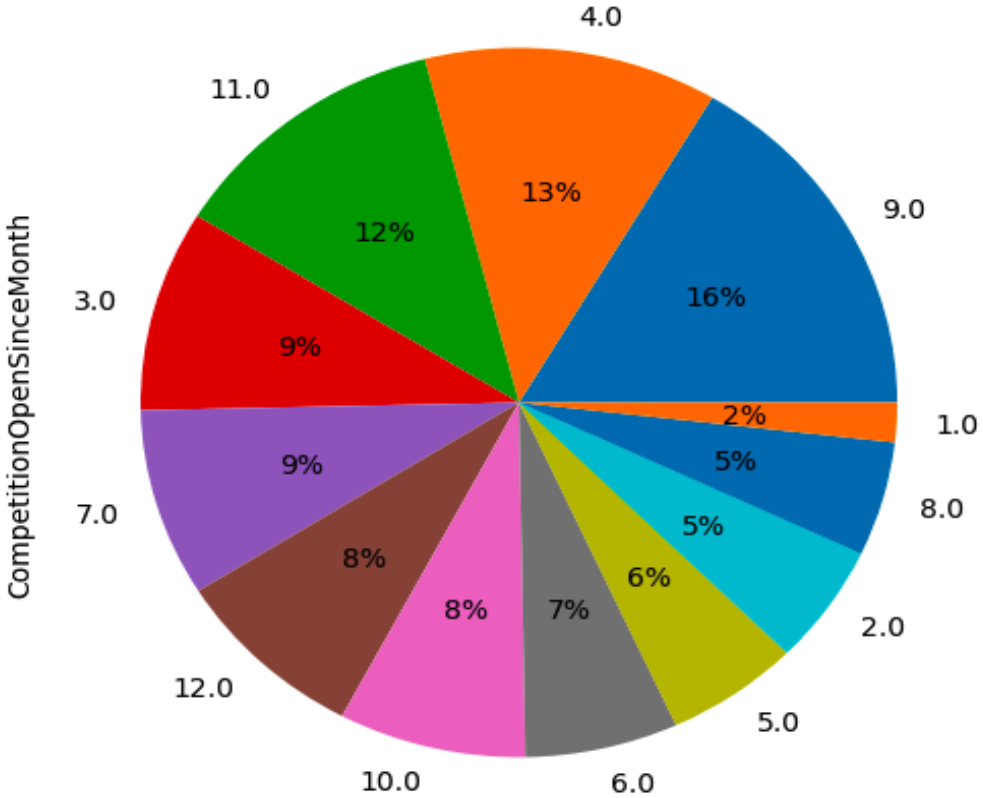




Most Sales are done in the beginning of the month with end of the month being the lowest.
Sales are more in the beginning of the week than the end

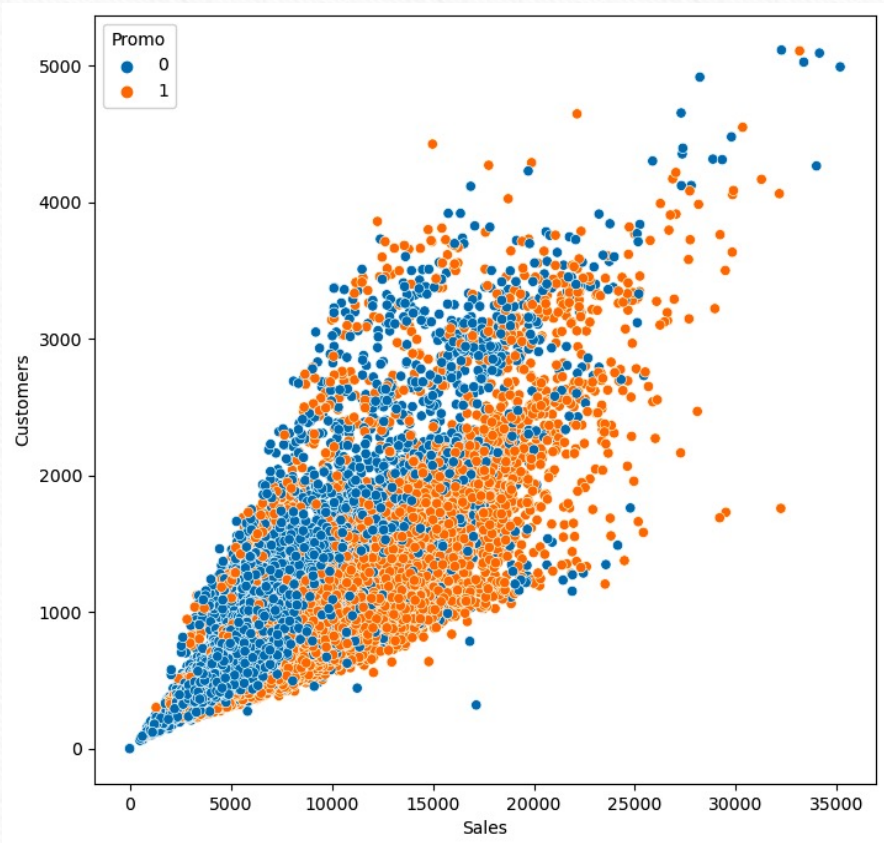
Sales behavior Cont'd

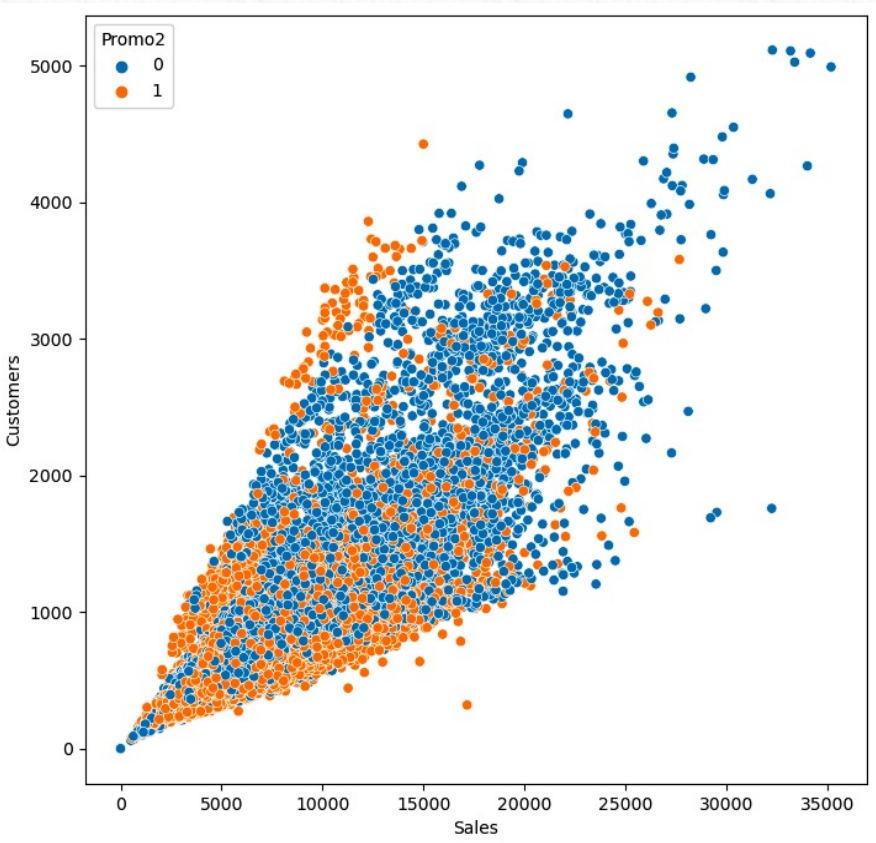




Sales relatively are lower by the end of year. Could it be because of the competitors increase influence during the last quarter? 16% in September and 12% in December ?

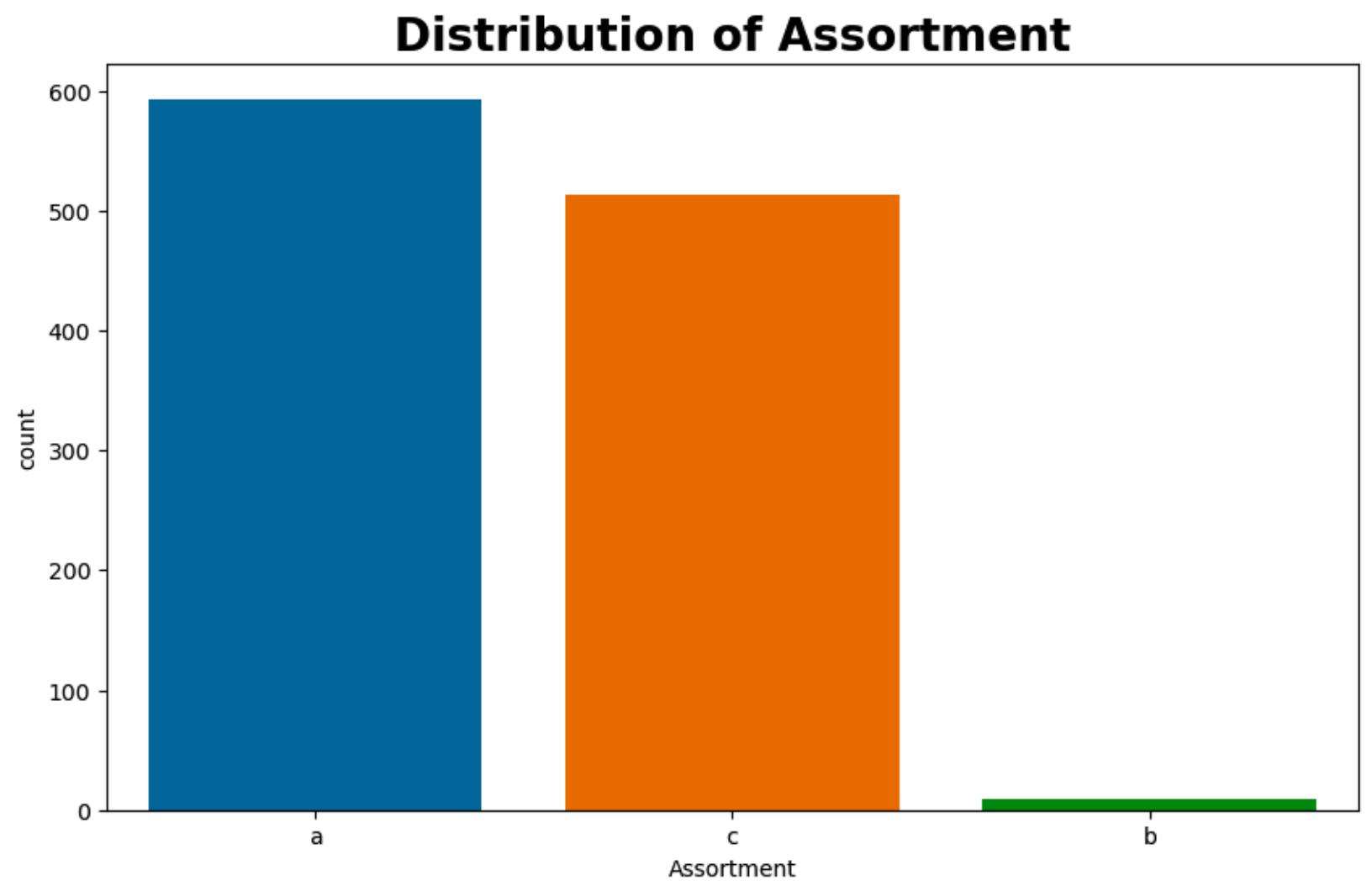
Promo types influence on sales and customers





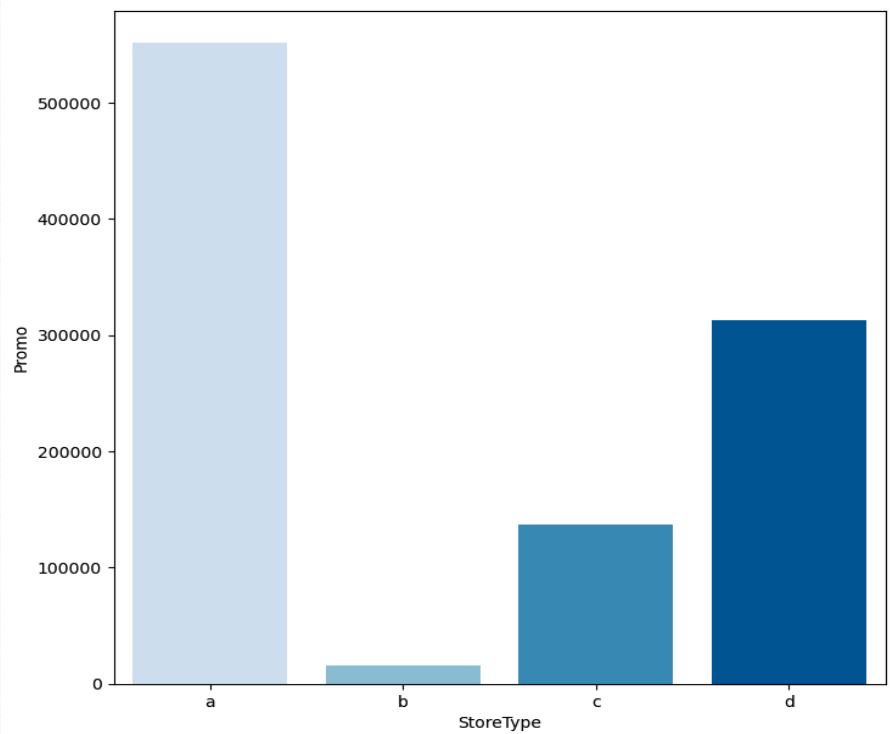
Seems like Promo1 was more successful for the stores! Lets check the Sales for each promo on average.

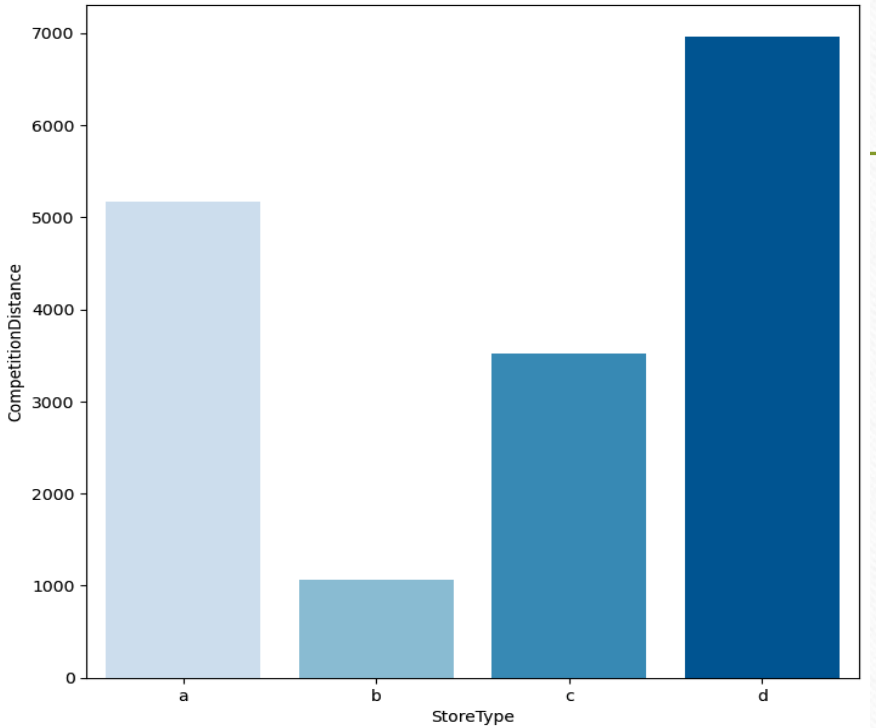
Assortment types



Looking at the graph Assortment a has higher counts. But assortment a and c should be used for selecting the product category to avail to the customers

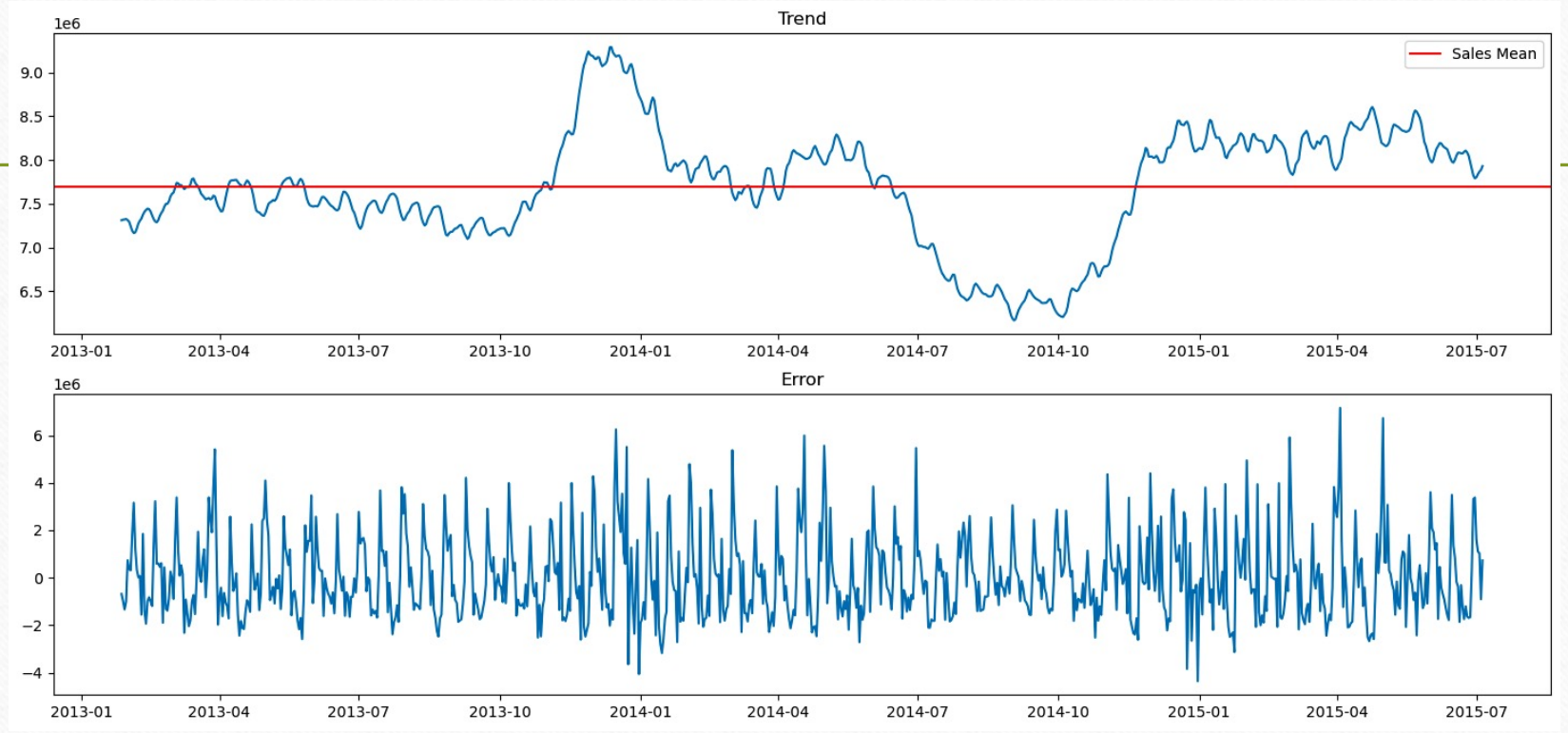
Variable transformations





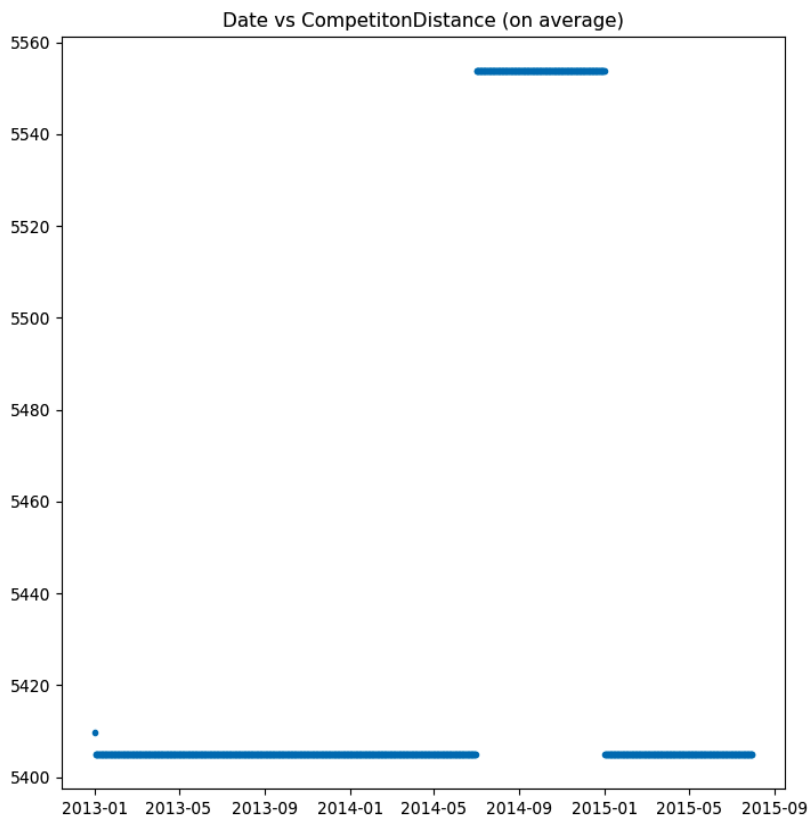
- Store A did the most Promo'1's despite being on average top second in comparison to other stores with regard to Competition Distance (distance in meters to the nearest competitor store). Hence, I think it is fair to say promos are a big deal. Other factors could be seasonality, trend etc. Let's see about trend!

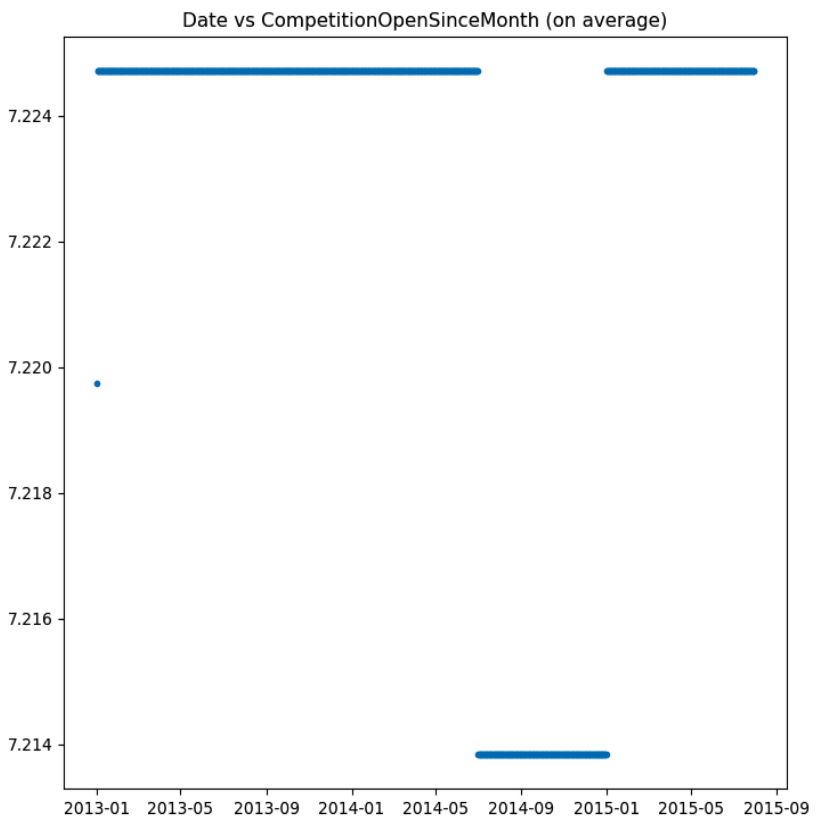
Trends sales by average over the years



- 2015 has been a good year as the trend line is above the average line by the end of 2014. Beginning of 2014 is a huge peak, I wonder what derived that?

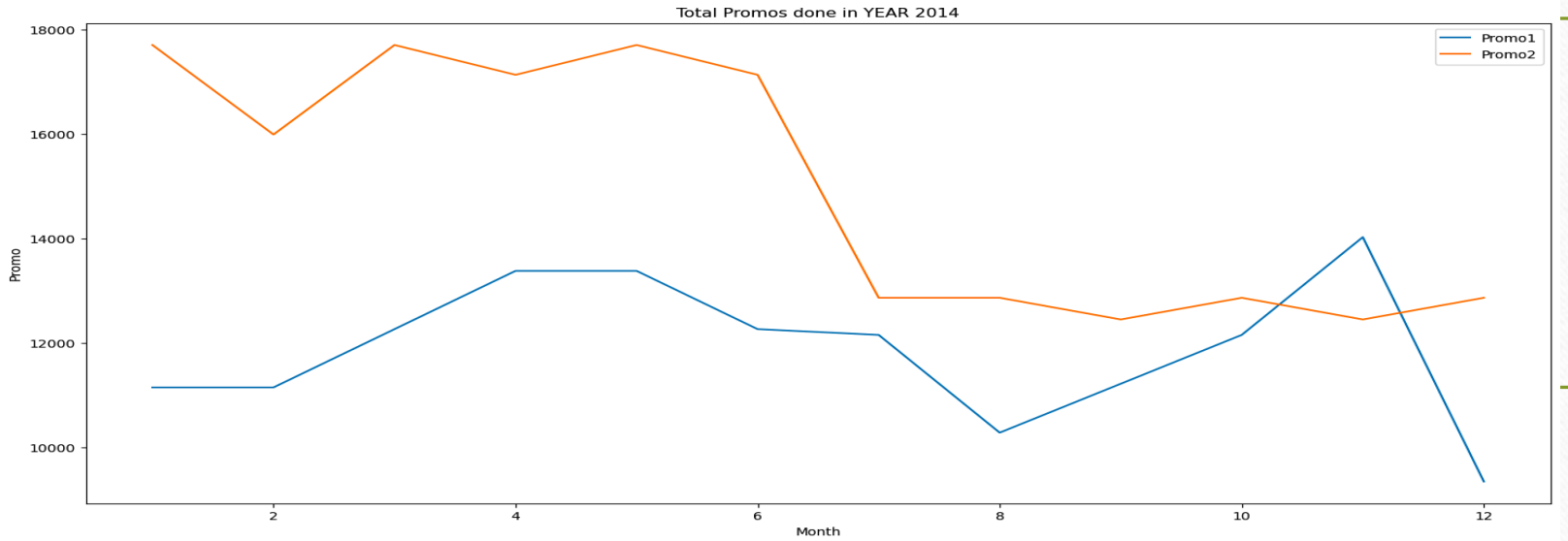
Distance to the next competitor effect

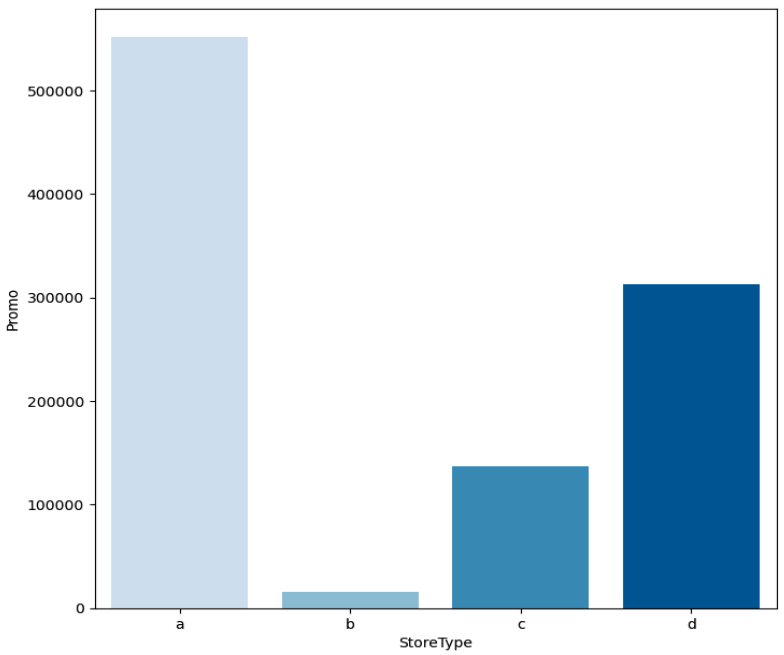


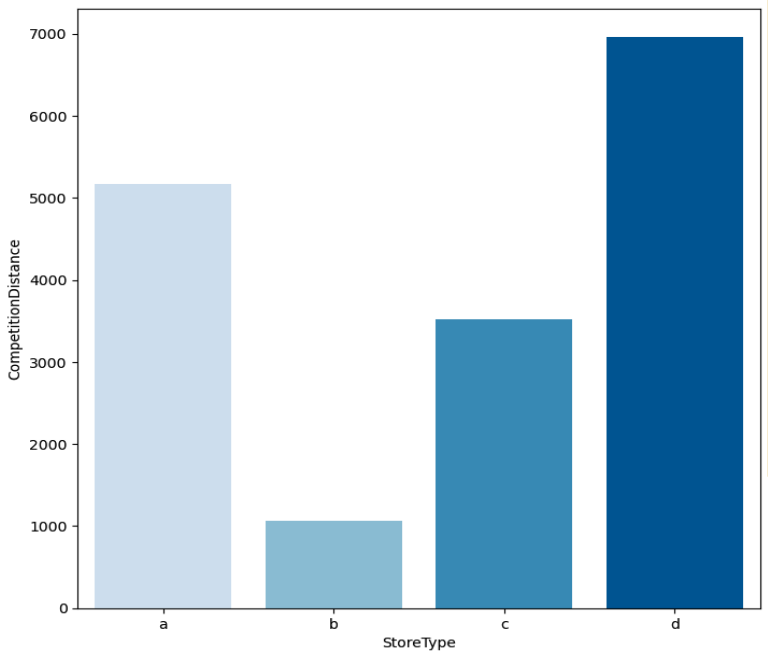


Seems like there was a new competitor near the end of 2014 and since the distance also relatively increased it could be maybe change of location, but these are just assumptions. It could be useful for the model to interpret such behavior in the future for the stores.

Lets see if the stores had done less promos when the trend was going down.



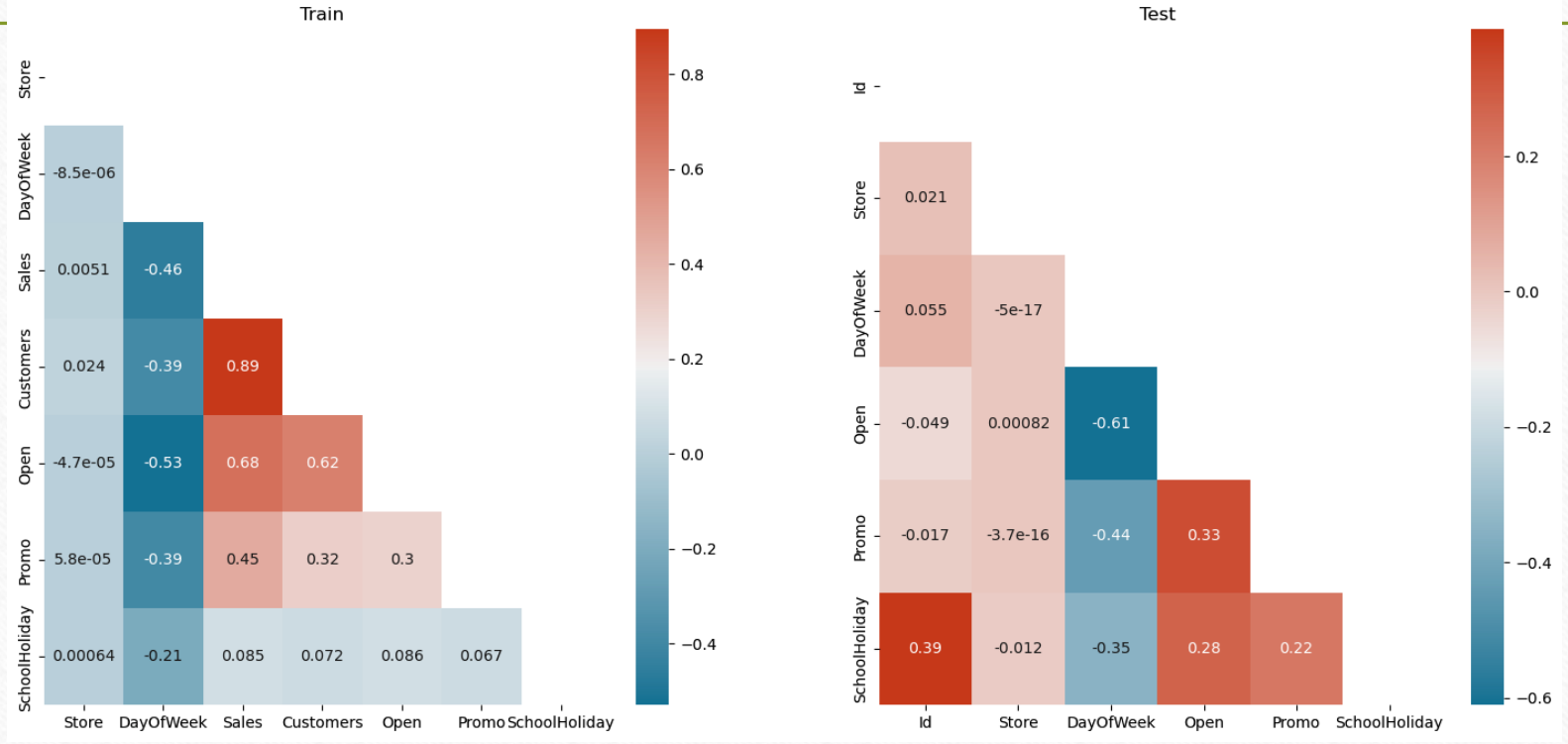




Looking at the second plot, it seems that in year 2014 there was no enough promation justifying the reason of sales being below the average

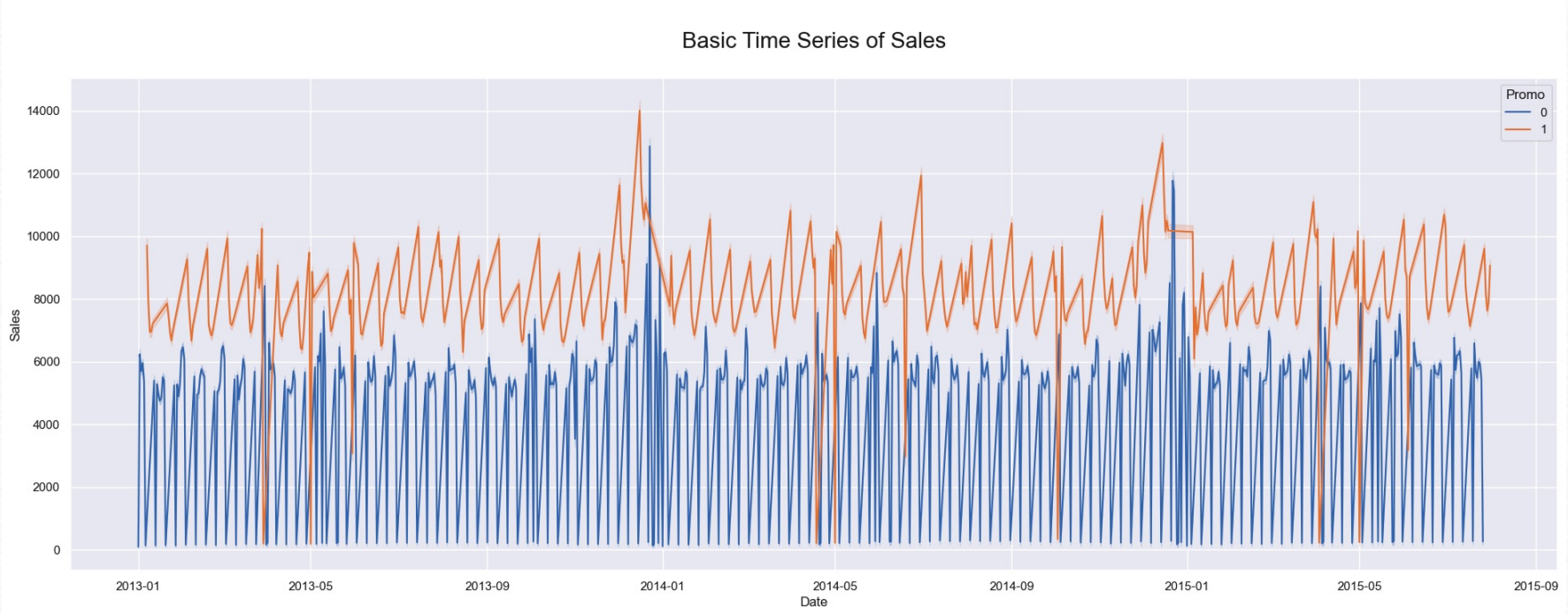
Correlation

- Sales are highly correlated with feature Customers and feature Open
- and moderately correlated with Promo

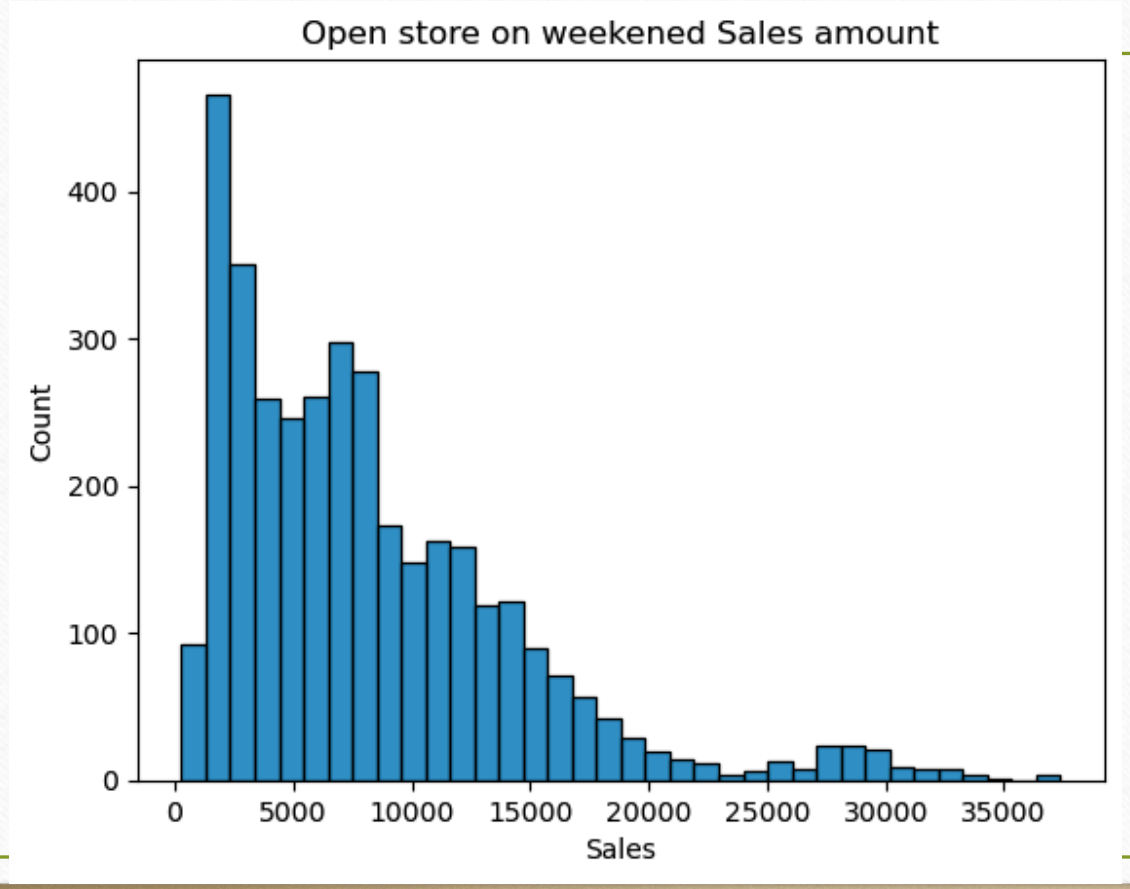


Promotion type improve sales

prom 1 is winning over the all years



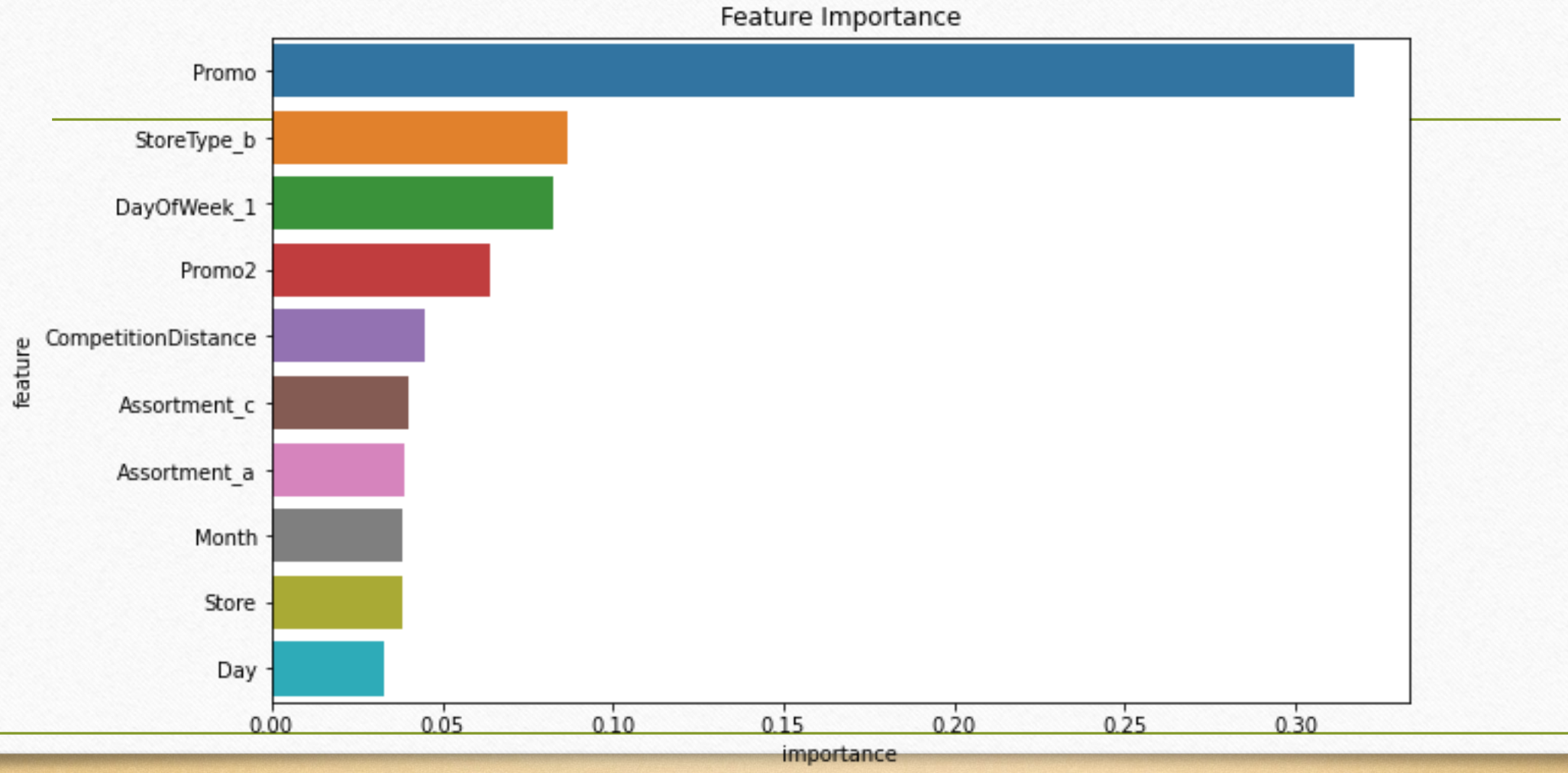
Investigate week sales



Sales are not higher, but also not negligable

ML Using xboost

- Feature importance



Conclusion

- We have sen that customers and sales increased if promotions are offered.
- We also have seen that assortment types have to be used before availing a certain product in a certain area
- Competitors also have an impact to our business, offering promotions will help improve sales in this case
- We need to continue investigate more our dataset then perfom User Engagement analysis.
- Predictive analysis using machine learning

Limitations

- Due to time limitation I couldn't proceed to Deep learning
- I was not able to also include Dashboards


Thank You!
alt