Capstone Project 2
Retail Sales Prediction

Team Members

Mohammed Saif Khan

Kaustubh Kulkarni

Content

- Problem Statement
- Data Description
- Approach
- Datasets
- Exploratory Data Analysis
- Feature engineering
- Model Implementation
- Hyperparameter Tuning
- Feature Importance
- Conclusion

Problem statement

- Rossmann operates over 3000 drug stores in 7 European countries. Rossmann managers are tasked with predicting their sales for 6 weeks in advance.

  Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality.

  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 Description

We have two datasets given

1. **Rossmann Stores Data.csv** - historical data including Sales
2. **store.csv** - supplemental information about the stores

Data Fields:

- **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
- **StateHoliday** - 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
- **SchoolHoliday** - indicates if the (Store, Date) was affected by the closure of public schools

Data Description(Contd...)

Data Fields:

- **StoreType** - differentiates between 4 different store models: a, b, c, d
- **Assortment** - describes an assortment level: a = basic, b = extra, c = extended
- **CompetitionDistance** - distance in meters to the nearest competitor store
- **CompetitionOpenSince**[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 participating
- **Promo2Since**[Year/Week] - describes the year and calendar week when the store started participating in Promo2
- **PromoInterval** - 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

Approach

1) Data Exploration

2) Data Cleaning and Preparation

3) Exploratory Data Analysis

- Hypothesis Testing
- Independent variables vs dependent variables plot

4) Feature Engineering

5) Modelling

- Linear Regression Model
- Decision Tree
- Random Forest Model
- Hyper Parameter Tuning

6) Conclusion

Final Dataset

Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday	StoreType	Assortment	CompetitionDistance	Promo2
1	5	2015-07-31	5263	555	1	1	0	1	c	a	1270.0	0
2	5	2015-07-31	6064	625	1	1	0	1	a	a	570.0	1
3	5	2015-07-31	8314	821	1	1	0	1	a	a	14130.0	1
4	5	2015-07-31	13995	1498	1	1	0	1	c	c	620.0	0
5	5	2015-07-31	4822	559	1	1	0	1	a	a	29910.0	0

Exploratory Data Analysis

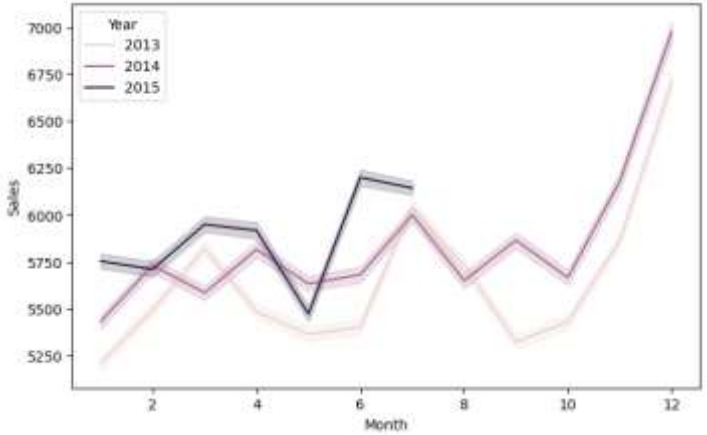
Hypothesis

After initial exploration of the data we came up with the following hypotheses.

- Stores should sell more over the years.
- Stores should sell less on weekends.
- Stores with closer competitors should sell less.
- Store type with a larger assortment of products should sell more.
- Number of Customers should have a positive correlation with Sales.
- Stores with promotion should have high Sales.

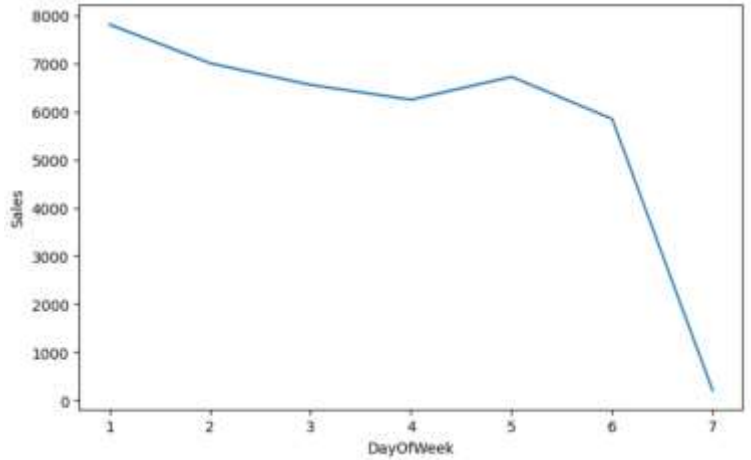
Hypothesis

H1: Stores should sell more over the years.



From the plot we can see the slight increase in Sales over the years. Hence, the Hypothesis is **True**.

H2: Sales should be lower on weekends.

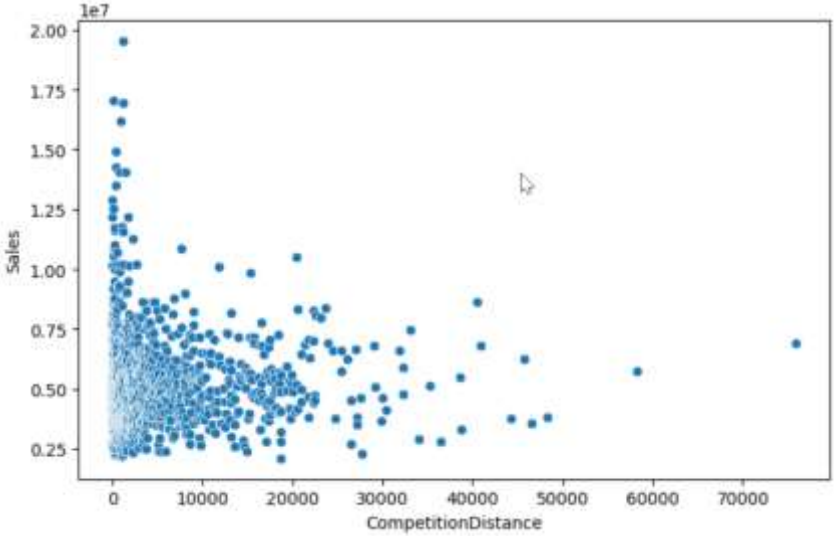


There is gradual decrease in the Sales over the week and the Sales are the lowest on weekend, as some stores remain close on weekends.

Hence, the Hypothesis is **True**.

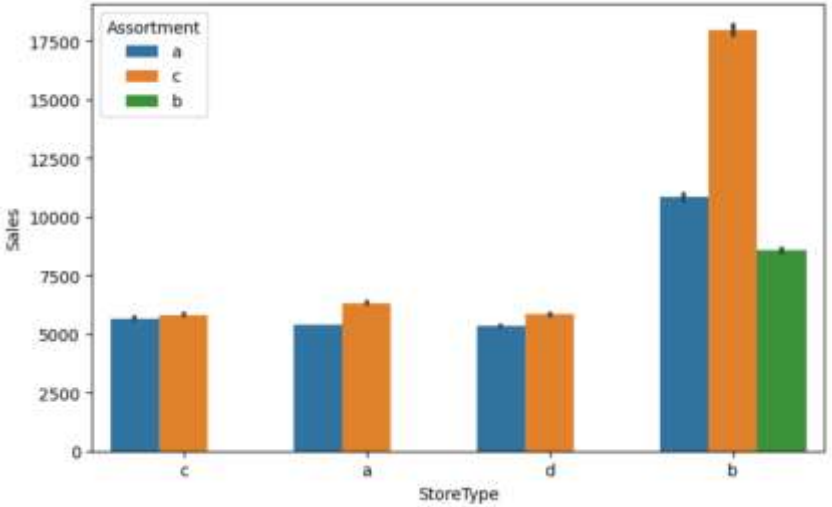
Hypothesis

H3: Stores with closer competitors should sell less.



From the plot we can deduce Stores with closer competition have the higher Sales. Hence ,the Hypothesis is **False**.

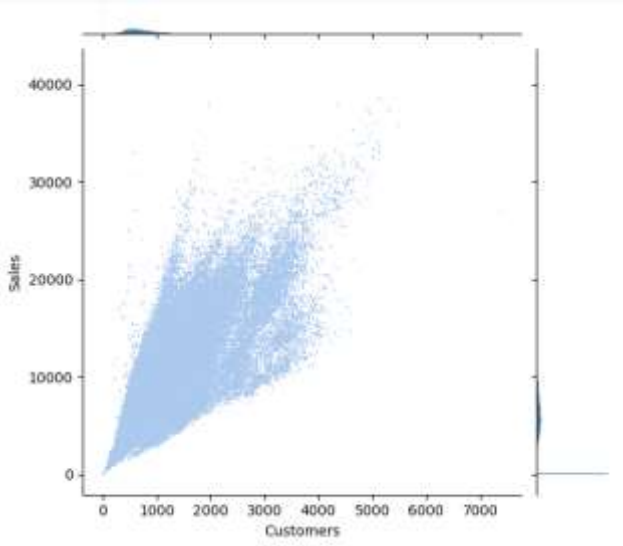
H4: Store type with a larger assortment of products should sell more.



Store type b has all the three types of assortments , thus the sales are highest. Hence , the Hypothesis is **True**.

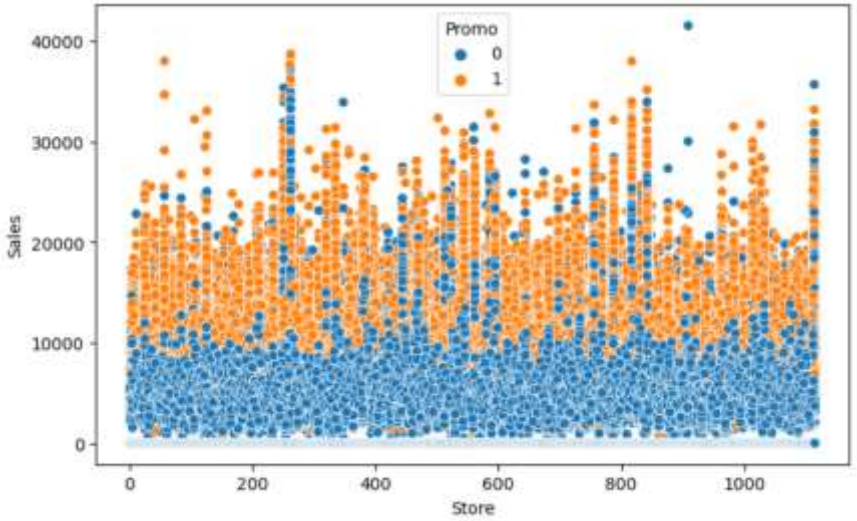
Hypothesis

H5: Number of Customers should have a positive correlation with Sales.



From the above plot we can see that Sales has linear relation with number of Customers.Hence , the Hypothesis is **True**.

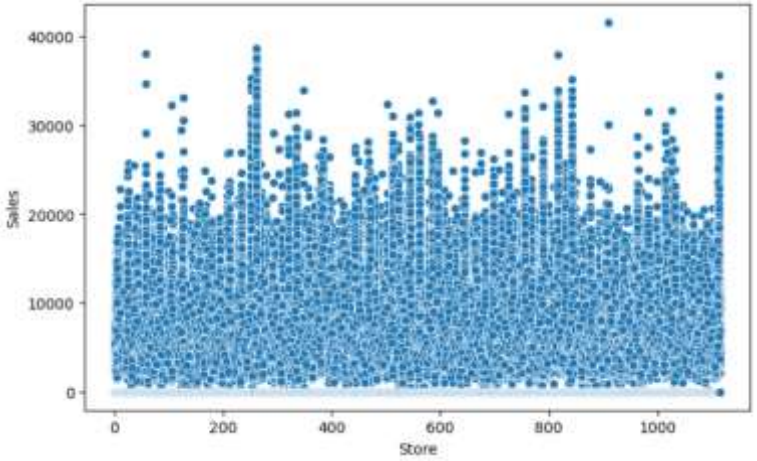
H6: Stores with promotion should have high Sales.



From the above plot we can see that Stores with promotion usually have higher Sales.Hence , the Hypothesis is **True**.

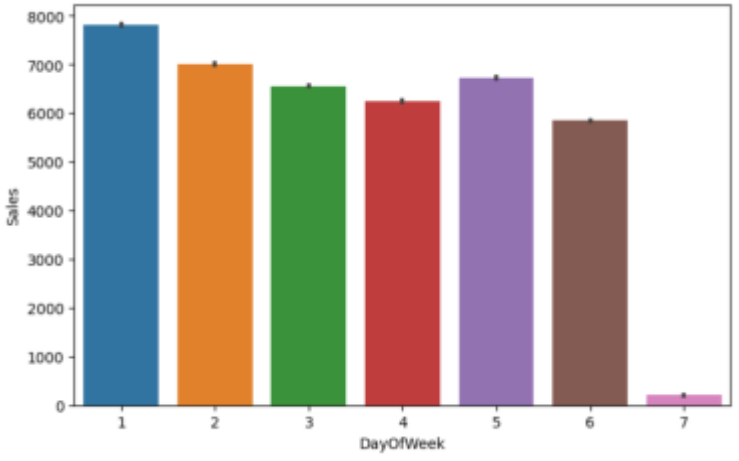
Independent vs Dependent variables

1. **Sales VS Stores**



Most of the times, Sales for the stores are below **20000**.

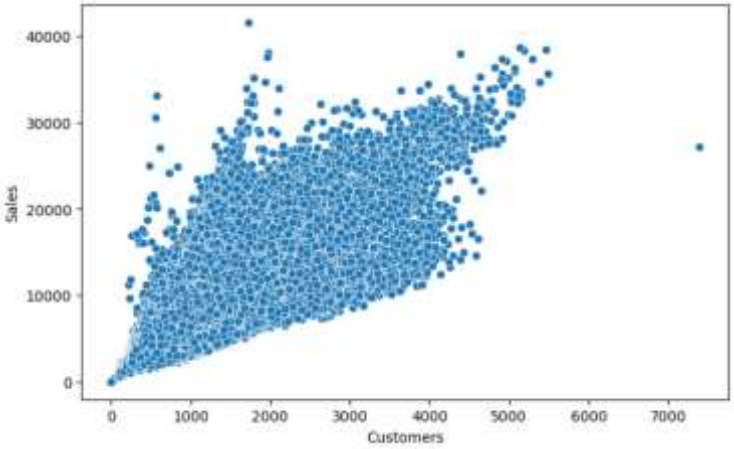
2. **Day of week vs Sales**



From the above plot we can say that Sales are highest on Mondays.

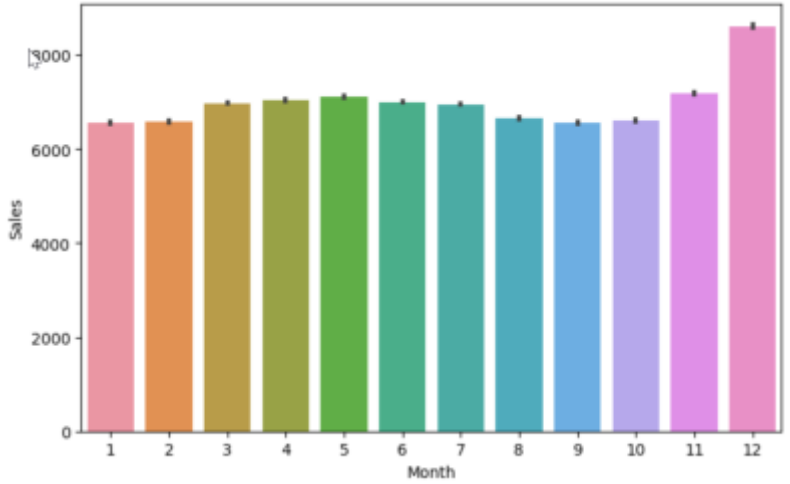
Independent vs Dependent variables AI

3. Customers VS Sales



There is a Linear relation between Sales and Customers.

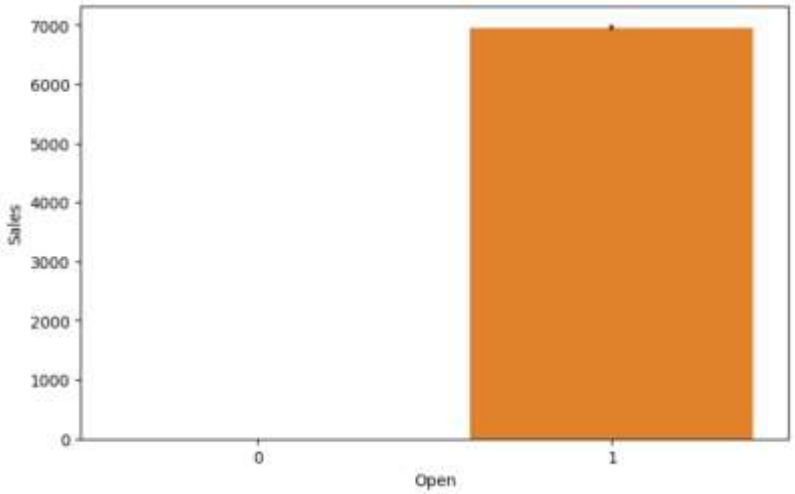
4. Sales per Month



Sales are highest in December maybe because of christmas month.

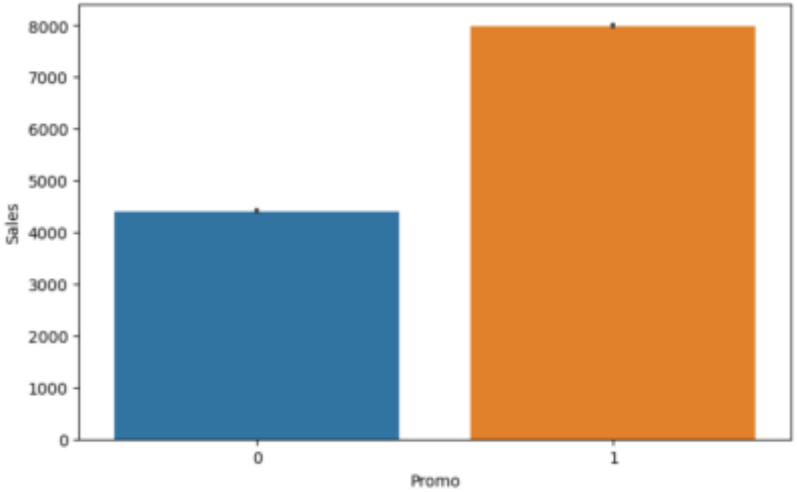
Independent vs Dependent variables

5. Sales Vs Open



Obvious that no sales when shops are closed.

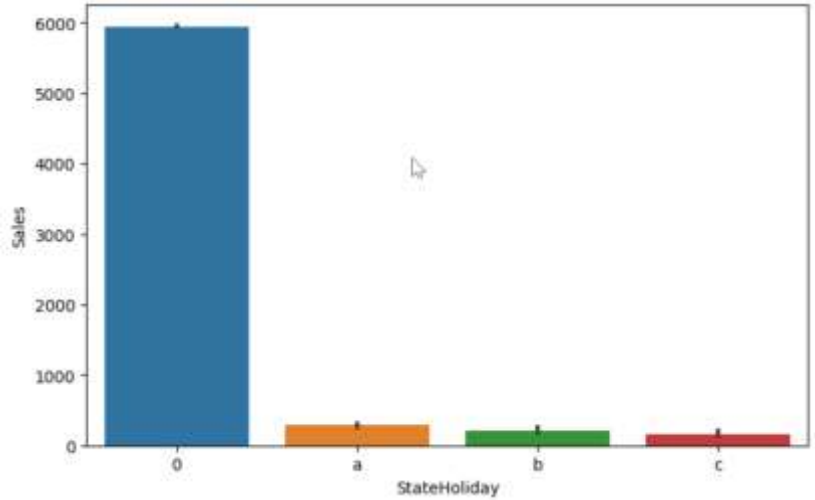
6. Promo VS Sales



The Stores with promotion has higher Sales.

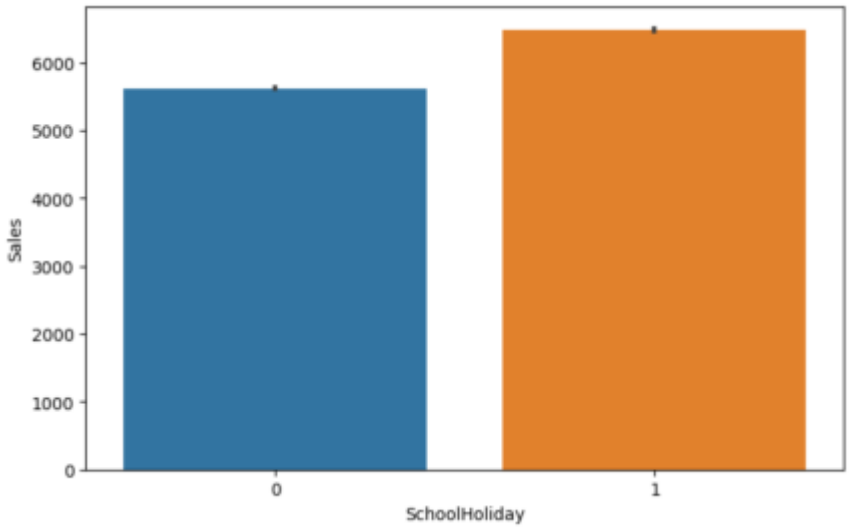
Independent vs Dependent variables

7. Sales VS StateHoliday



Lowest sales on state holidays as many shops remain closed.

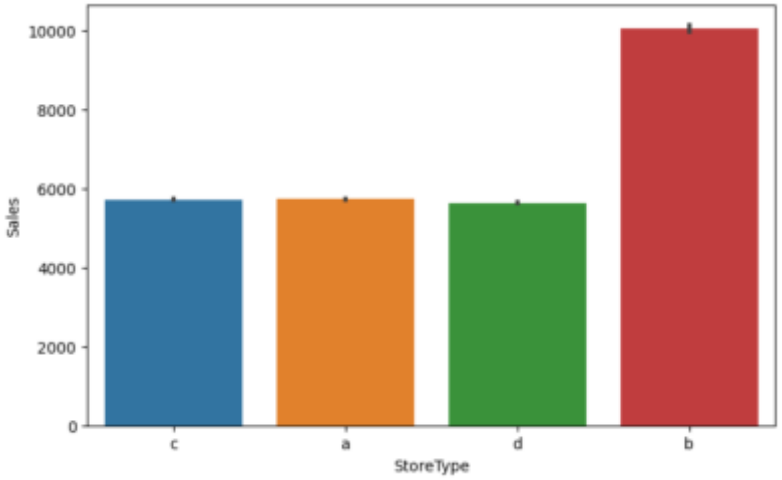
8. SchoolHoliday VS Sales



Though Stores were affected by school holidays they have higher sales in general.

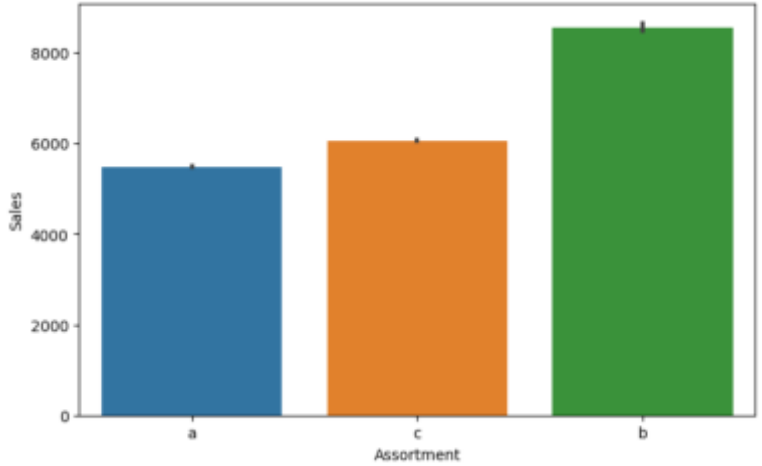
Independent vs Dependent variables

9. StoreType VS Sales



Store type b has the highest Sales.

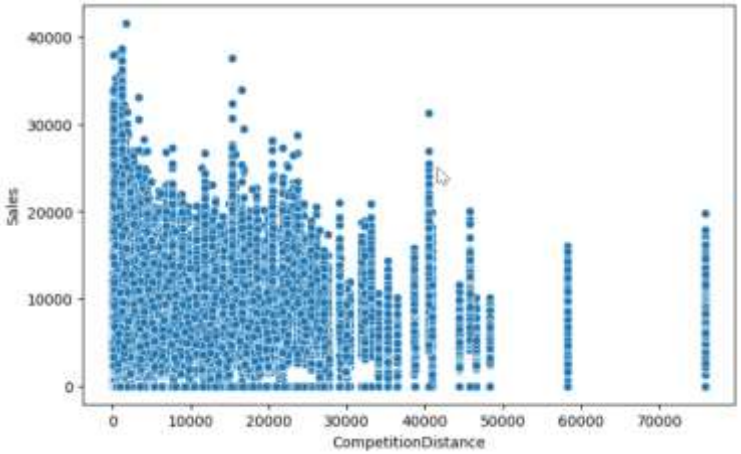
10. Assortment VS Sales



The Stores with **Assortment type b** have comparatively high sales.

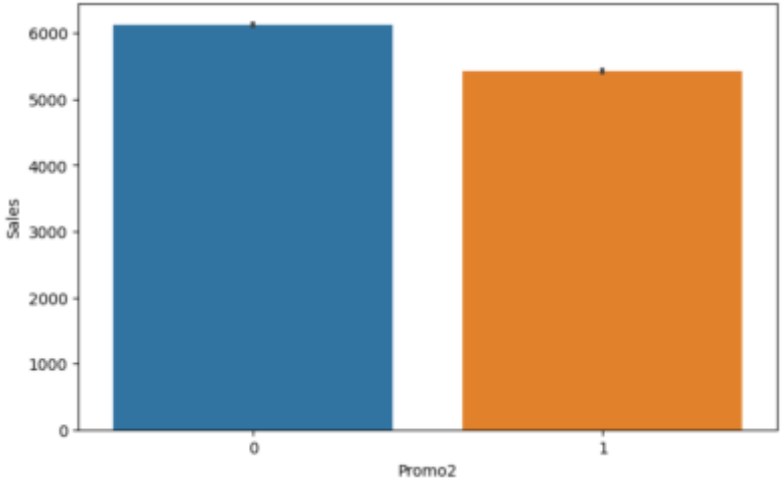
Independent vs Dependent variables

11. CompetitionDistance VS Sales



Most of the competitive stores are located close to each other and they have high sales as well.

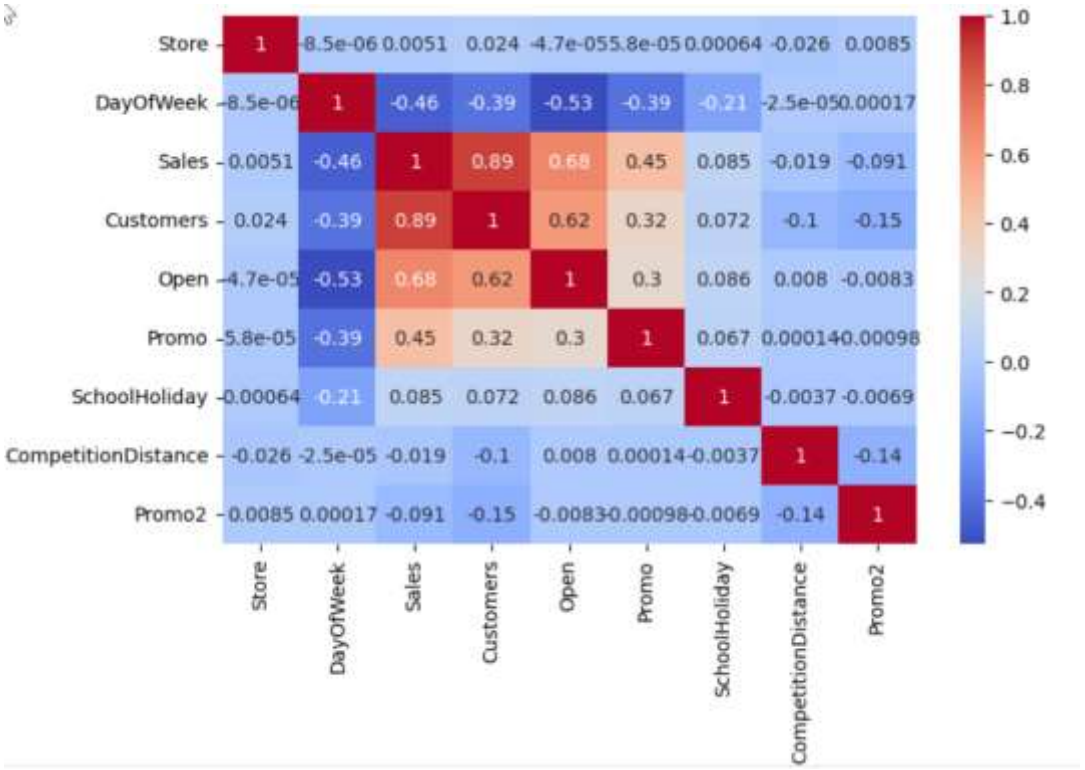
12. Sales VS Promo2



Store sales don't seem to benefit that much by long term promotion.

Correlation

- Sales and Customers shows strong positive correlation.
- Sales and Open shows strong positive correlation.
- Sales and Promo shows moderate positive correlation.
- Sales and DayOfWeek shows moderate negative correlation.
- DayOfWeek shows negative correlation with most of the columns.



Heatmap

Feature Engineering

Store	DayOfWeek	Sales	Customers	Promo	StateHoliday	SchoolHoliday	Promo2	Competition	WeekNumber
1	5	5263	555	1	0	1	0	1	31
2	5	6064	625	1	0	1	1	1	31
3	5	8314	821	1	0	1	1	0	31
4	5	13995	1498	1	0	1	0	1	31
5	5	4822	559	1	0	1	0	0	31

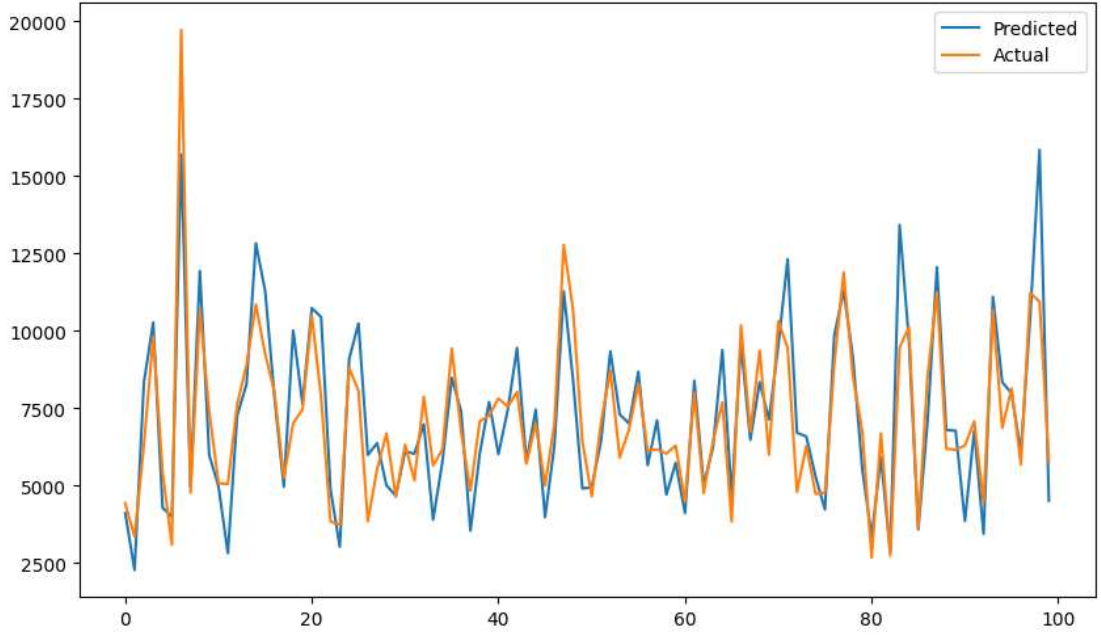
StoreType_a	StoreType_b	StoreType_c	StoreType_d	Assortment_a	Assortment_b	Assortment_c
0	0	1	0	1	0	0
1	0	0	0	1	0	0
1	0	0	0	1	0	0
0	0	1	0	0	0	1
1	0	0	0	1	0	0

```
Int64Index: 844338 entries, 0 to 1017190
Data columns (total 17 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   Store           844338 non-null  int64
 1   DayOfWeek       844338 non-null  int64
 2   Customers       844338 non-null  int64
 3   Promo           844338 non-null  int64
 4   StateHoliday    844338 non-null  int64
 5   SchoolHoliday   844338 non-null  int64
 6   Promo2          844338 non-null  int64
 7   Competition     844338 non-null  int64
 8   WeekNumber      844338 non-null  int64
 9   StoreType_a     844338 non-null  uint8
 10  StoreType_b     844338 non-null  uint8
 11  StoreType_c     844338 non-null  uint8
 12  StoreType_d     844338 non-null  uint8
 13  Assortment_a    844338 non-null  uint8
 14  Assortment_b    844338 non-null  uint8
 15  Assortment_c    844338 non-null  uint8
 16  Sales           844338 non-null  int64
dtypes: int64(10), uint8(7)
```

Model Implementation

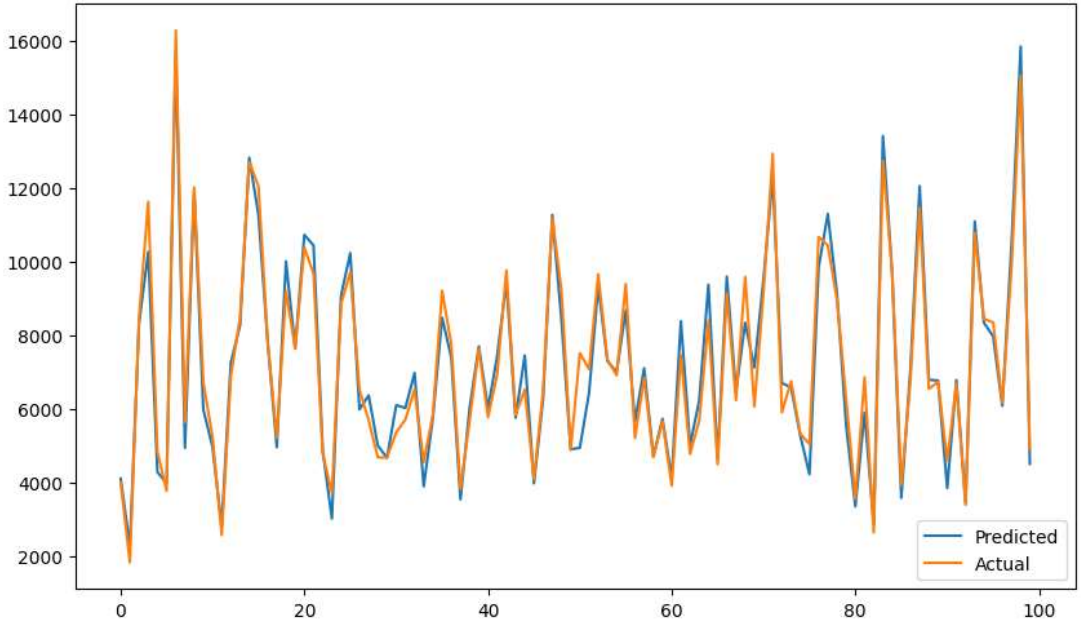
Linear Regression

Model Name	MAE Train	MAE Test	MAPE Train	MAPE Test	RMSE Train	RMSE Test	R2_Score Train	R2_Score Test	Adj_r2 Train	Adj_r2 Test
Linear Regression	944.8021	941.918021	0.14411	0.144346	1299.767071	1292.875092	0.825033	0.824893	0.82503	0.82489



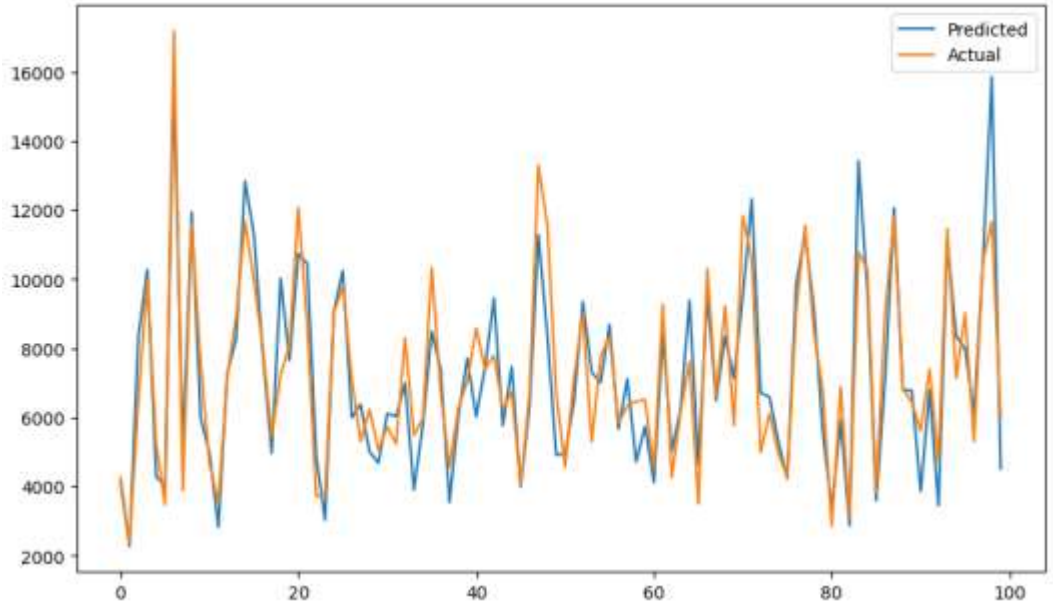
Decision Tree

Model Name	MAE Train	MAE Test	MAPE Train	MAPE Test	RMSE Train	RMSE Test	R2_Score Train	R2_Score Test	Adj_r2 Train	Adj_r2 Test
Decision Tree	0.915721	491.972325	0.000162	0.072283	16.850997	753.976317	0.999971	0.940447	0.999971	0.940446



Random Forest

Model Name	MAE Train	MAE Test	MAPE Train	MAPE Test	RMSE Train	RMSE Test	R2_Score Train	R2_Score Test	Adj_r2 Train	Adj_r2 Test
Random Forest Regression	853.533346	856.337341	0.12849	0.129219	1154.801786	1157.777003	0.861885	0.859577	0.861883	0.859574

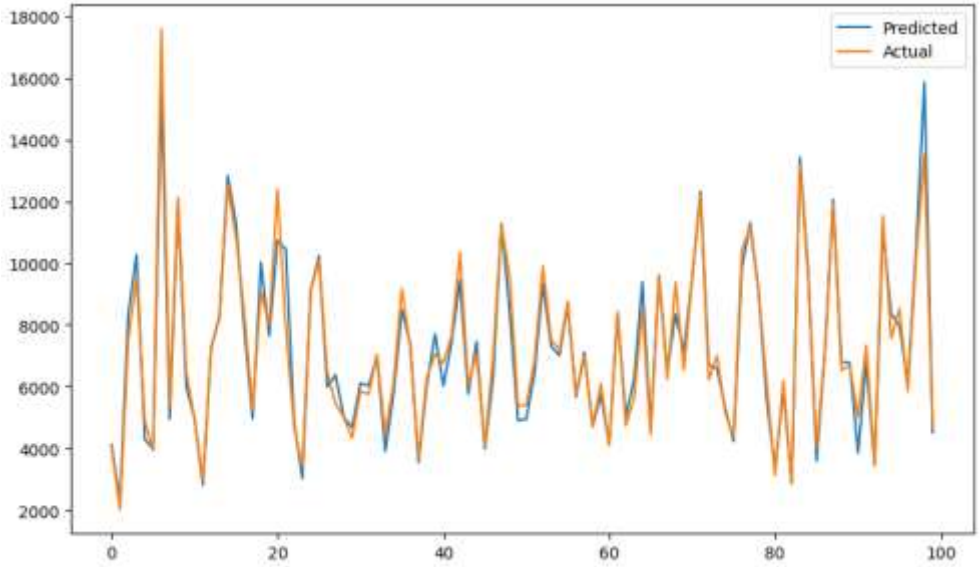


Random Forest after tuning

Model Name	MAE Train	MAE Test	MAPE Train	MAPE Test	RMSE Train	RMSE Test	R2_Score Train	R2_Score Test	Adj_r2 Train	Adj_r2 Test
Random Forest Regression hyperparameter tuned	305.661222	379.293586	0.044648	0.055516	457.793898	573.587229	0.978295	0.965534	0.978294	0.965534

Best Parameters

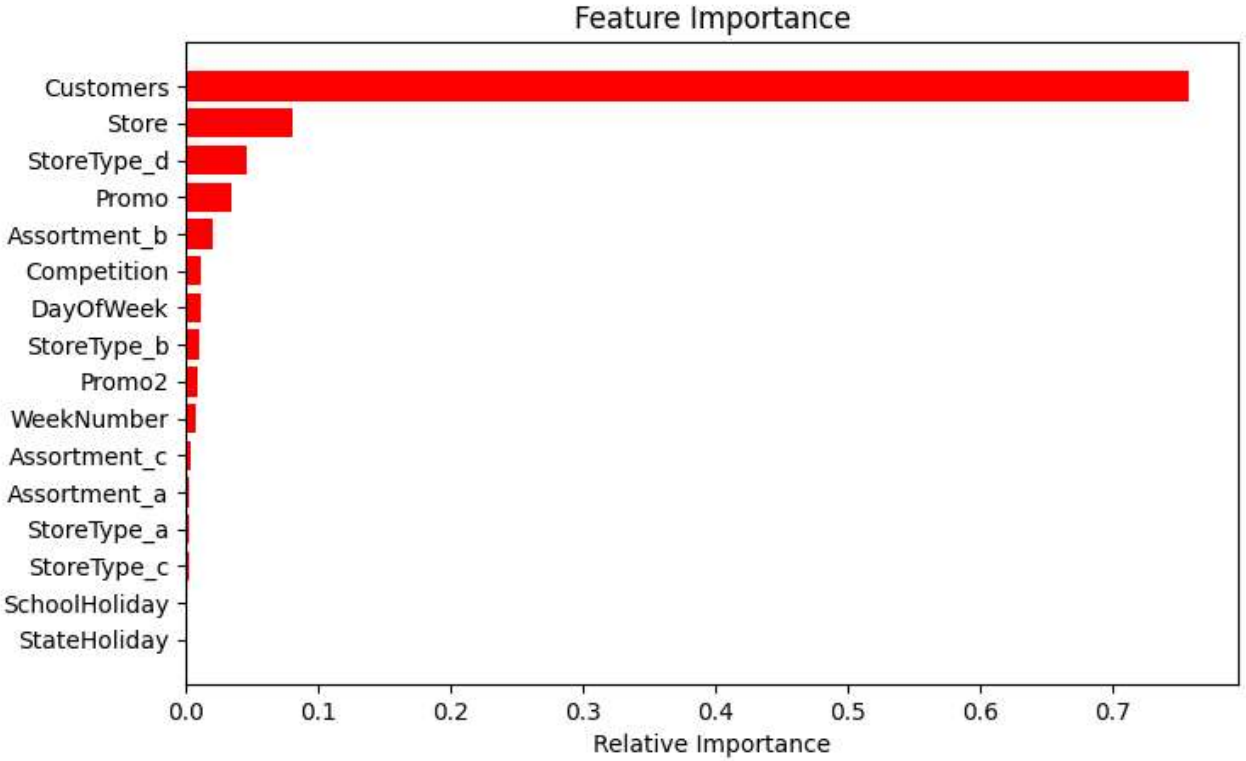
Max depth	50
Min Sample leaf	1
Min sample split	20
n estimators	150



Model Metrics Dataset

Model Name	MAE Train	MAE Test	MAPE Train	MAPE Test	RMSE Train	RMSE Test	R2_Score Train	R2_Score Test	Adj_r2 Train	Adj_r2 Test
Linear Regression	944.802100	941.918021	0.144110	0.144346	1299.767071	1292.875092	0.825033	0.824893	0.825030	0.824890
Decision Tree	0.915721	491.972325	0.000162	0.072283	16.850997	753.976317	0.999971	0.940447	0.999971	0.940446
Random Forest Regression	853.533346	856.337341	0.128490	0.129219	1154.801786	1157.777003	0.861885	0.859577	0.861883	0.859574
Random Forest Regression hyperparameter tuned	305.661222	379.293586	0.044648	0.055516	457.793898	573.587229	0.978295	0.965534	0.978294	0.965534

Feature Importance



Features like customers, store, promo, and store type b have more importance.

Conclusion

Important points from Exploratory Data Analysis :

- Sales column shows positive correlation with Customers, Open and Promo columns.
- Sales and DayOfWeek shows negative correlation.
- Sales for the Stores increases over the years.
- Sales decrease over weekends.
- Store with closer competition have more sales.
- Stores with all types of assortments tend to have higher sales.
- Sales increases with an increase in customers.
- We can an increase in sales for the stores that did promotion.

Conclusion

Prediction Summary For Models :

- Linear Regression Model shows an accuracy of 82.5% on train set and 82.4% on test set.
- Decision Tree Regression Model shows an accuracy of 99.9% on train set and 94.0% on test set.
- Random Forest Regression Model shows an accuracy of 86.1% on train set and 85.9% on test set.
- Hyperparameter Tunned Random Forest Regression Model shows an accuracy of 97.7% on train set and 96.5% on test set.
- The best model is the Hyperparameter Tunned Random Forest Regression Model.
- Decision Tree Regression Model has the highest accuracy but since it is overfitting therefore we don't consider this as the best model.

Conclusion

Important Features and Some Suggestions :

- The top five important features are Customers, Store, StoreType_d, Promo and Assortment_b.
- We can say that if the number of customers increase then there will be an increase in sales so stores should adapt methods for attracting customers.
- It is obvious that with more new stores there will be more sales.
- Store Owner should open stores that are from the storetype d.
- Now we know that if a store does promotion then the sales will increase so stores should increase their promotions.
- Stores should have assortments of assortment b as this would increase their sales.

THANK YOU!