Approach Document :

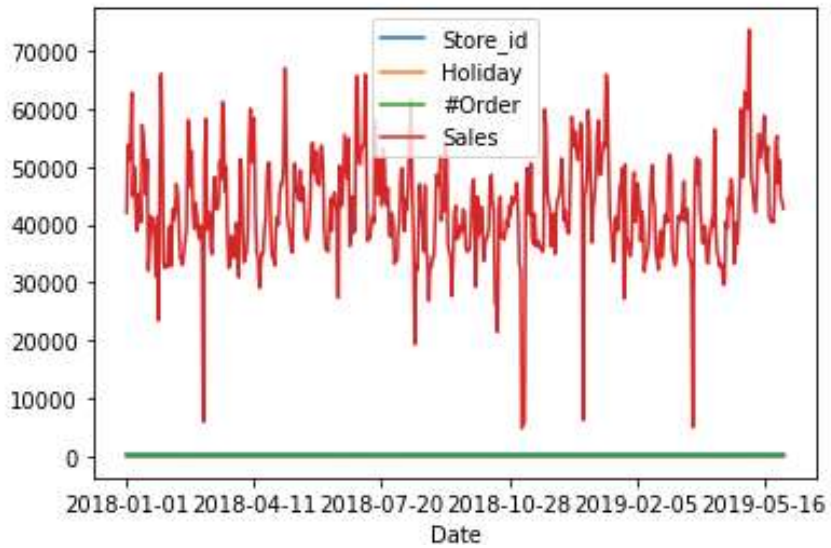
Shashikant Singh (shashikant.singh501@gmail.com)

EDA :

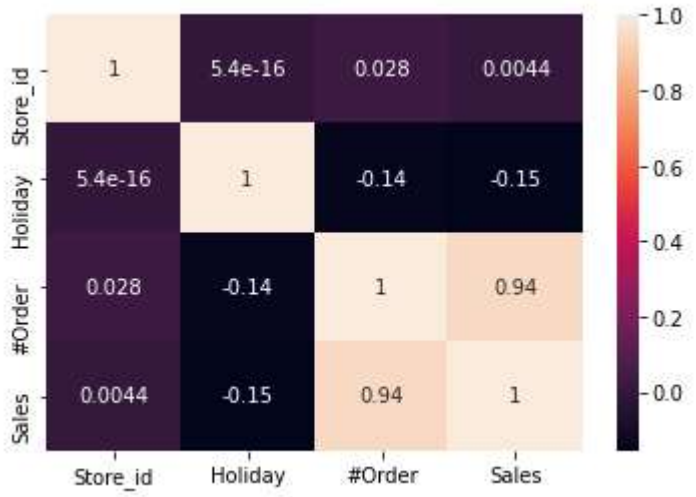
- From the EDA it was found out that there are outliers in Sales and #Order

- There are no null values in the dataset
- There are 188340 data points in the dataset with 10 features initially
- Sales data is following some seasonality.

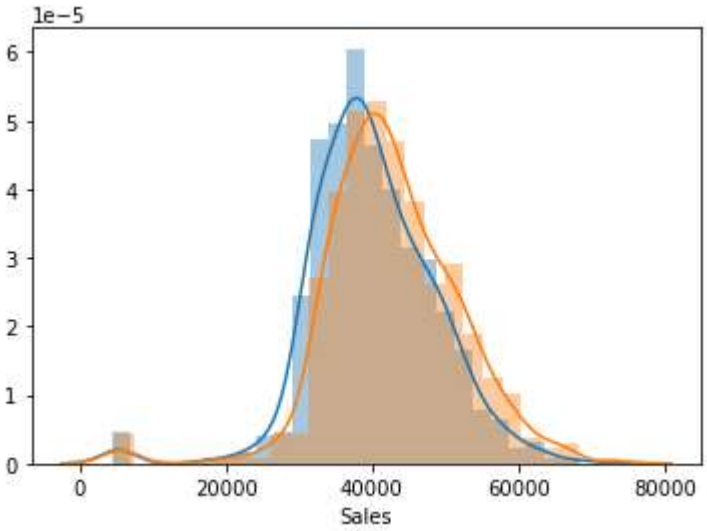
```
<matplotlib.axes._subplots.AxesSubplot at 0x1e81856ef10>
```

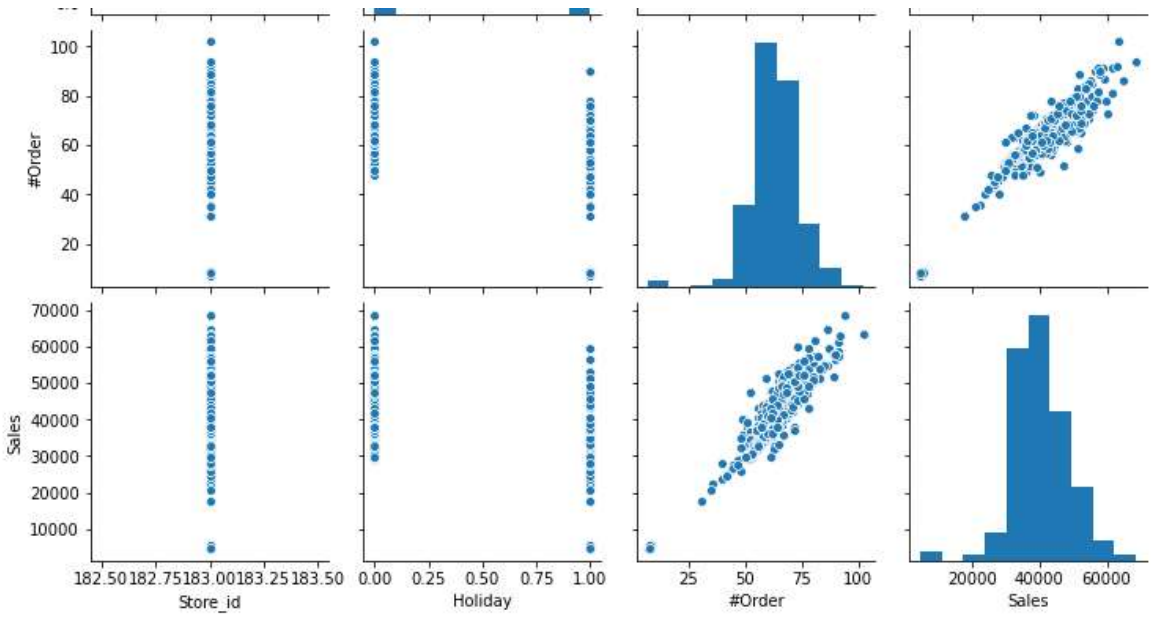


- There is high correlation between number of orders and Sales



- Sales follows bell curve.





Approach :

- Since #Order has high correlation with Sales and it was not available in the final test dataset, I have dropped it to make a workable model
- Since sales data is normally distributed and other paraments are independent of other, I can go ahead with model.
- To add more features in model , I divided the date in year, month, day, day of week, is month start, is month end, quarter.
- Created dummies for the training data set.
- Divided the training dataset into 80:20.
- First Decision Tree model is used which gave the accuracy of around 52%.
- Next I used Random Forest model with 100 estimators.
- Random forest model gave an average accuracy of 64% on train and 53% on Test dataset.
- To further improve the accuracy of model number of times paraments of random forest were changed but on average it's accuracy remained same.
- Next I moved to use XGBoost Regressor,
- Using XGBoost Regressor gave an average accuracy of 70.37% on train and 53.28% on test dataset.
- Tried improving the accuracy of model using different features and changing paraments.
- But the average accuracy of model remained at 70.37% on train and 53.28% on test dataset.
- I go ahead with this model to predict the sales for test final dataset.

Thank You.