Retail Store Sales Analytics

Presentation Outline

- Problem Statement
- Overview of Dataset
- EDA and Key Insights
- Modelling & Evaluation of Models
- Conclusion and Future Works
- Jupyter Notebook
- Q&A

Business Problem

We are a big supermarket chain that runs about 1000 stores across the country. The regular sales forecast from area managers are often inaccurate and they uses sales dollars per square foot (store area) as the primarily metric to measure store performance.

The CEO has now tasked the Data Scientist team to provide some insights to the store sales performance metrics, what the chain can do to optimize their physical retail presence and if there is anyway to provide a estimate sales benchmark for the 1000 stores.

Data Science Questions

We attempt to answer these questions through this project:

1. Is there a correlation between store floor area vs store sales $?
2. Is there any other business insights we can provide to?
3. Can we use the available dataset to design a machine learning model to estimate store sales?

EDA & Business Insights

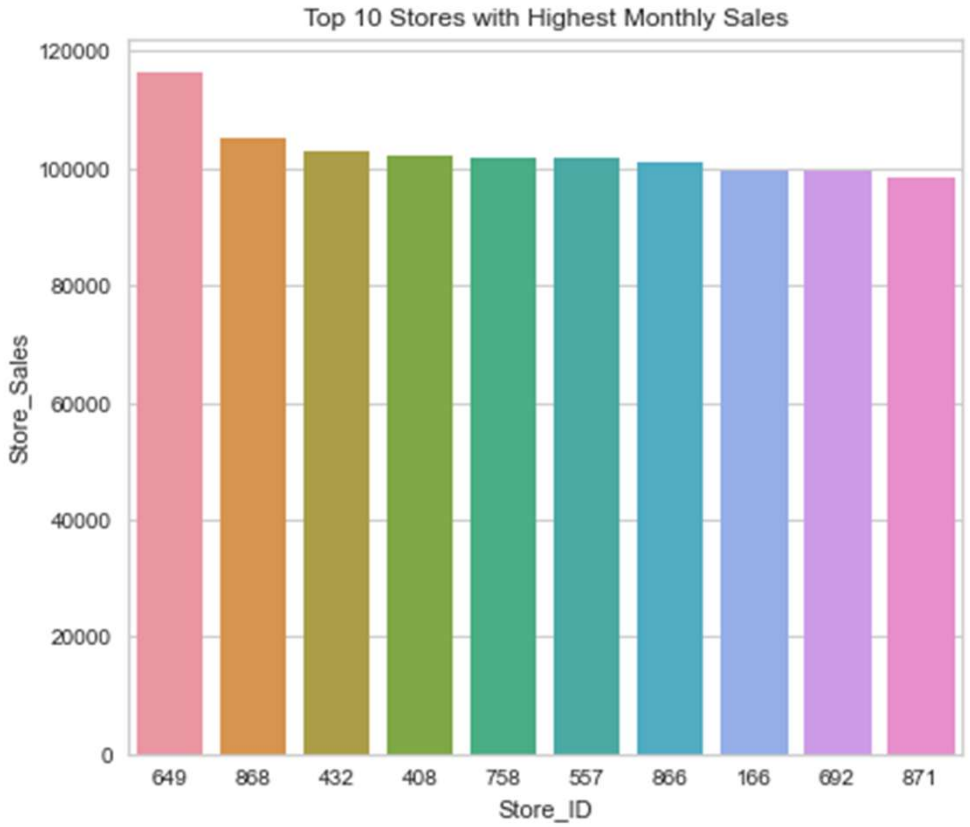
Overview of Dataset

Overview of Dataset:

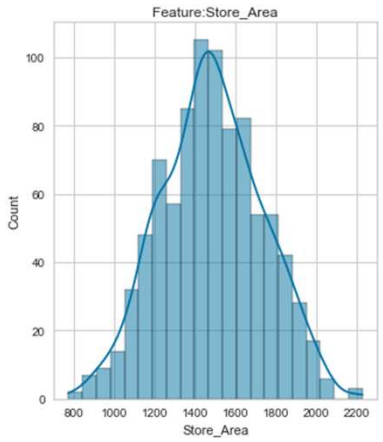
896 rows, 5 columns

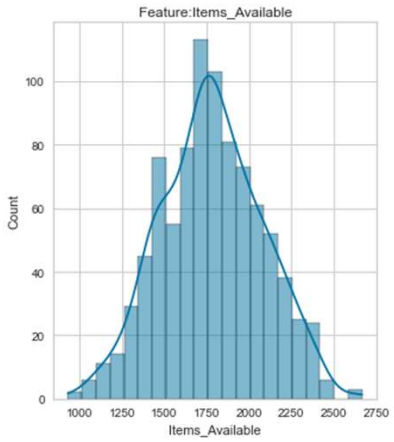
Column Name	Description
Store_ID	Numeric representation of the Store
Store_Area	Physical Area of the store in yard square
Items_Available	Number of different items available in the corresponding store
DailyCustomerCount	Number of average daily customers who visited to stores
Store_Sales	Sales in (US $) that stores made per month

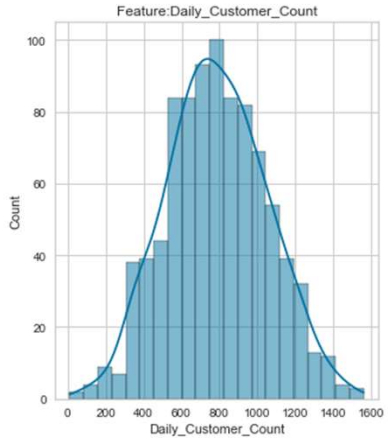
Top 10 Performing Stores - $100k

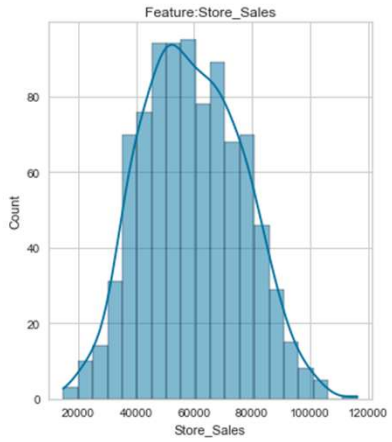


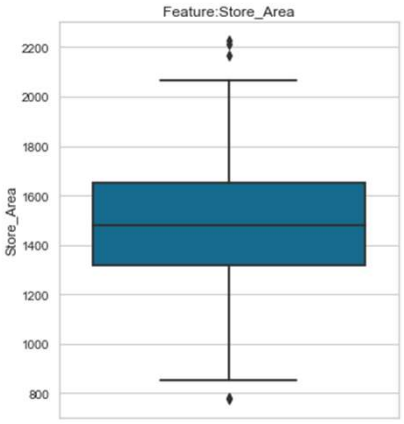
EDA – All Normally Distributed!

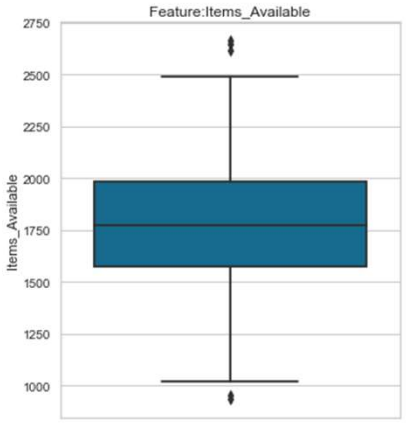


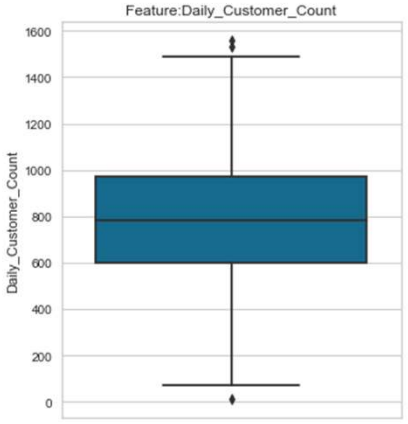


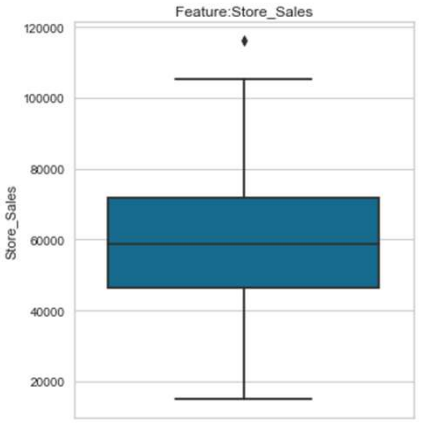




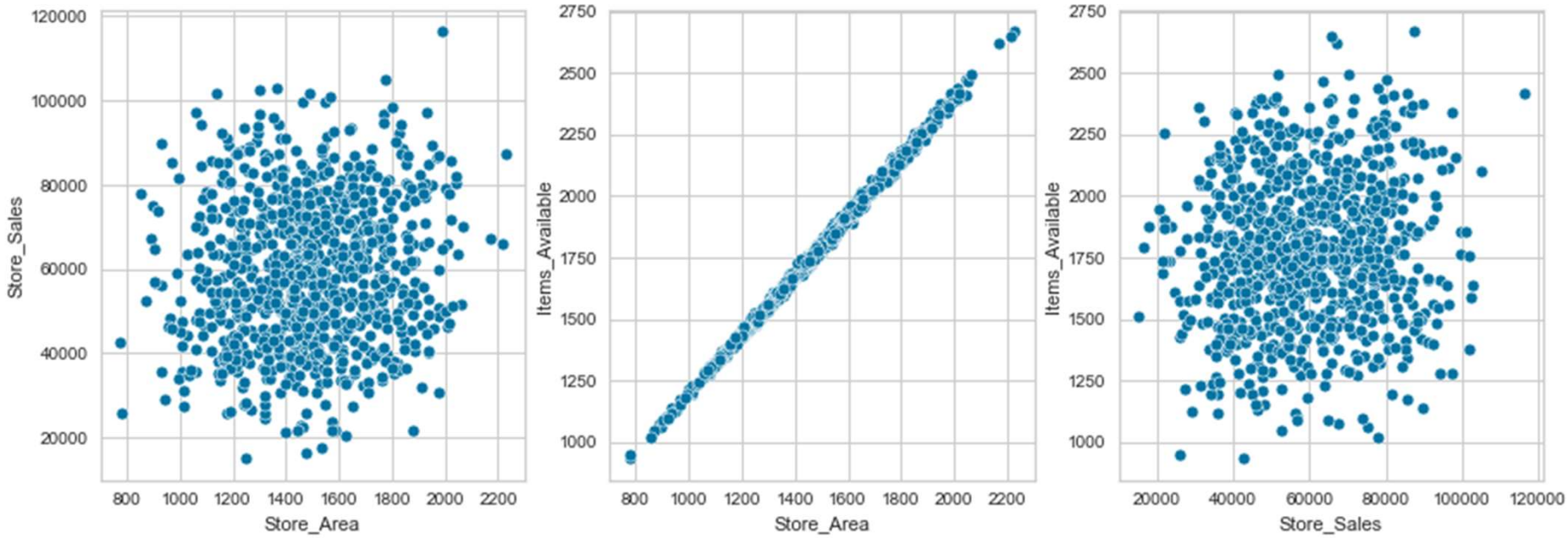






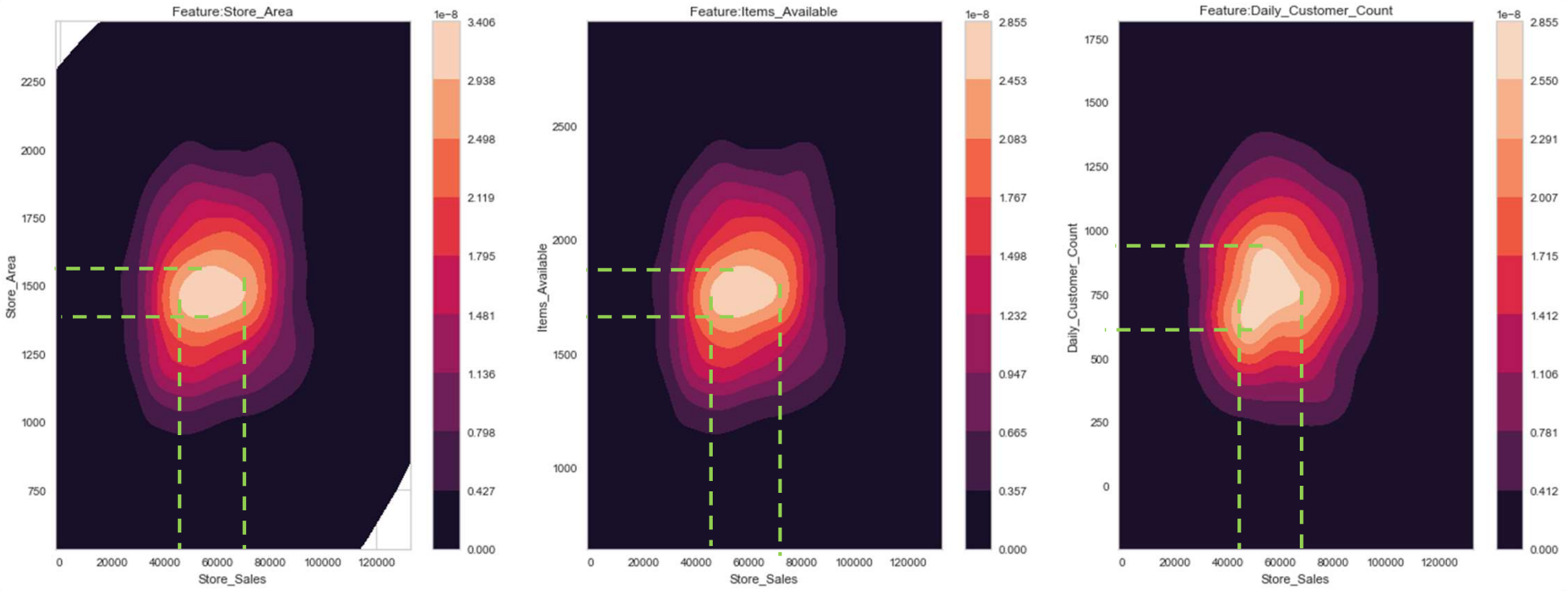


Store Area correlated to Store Sales?



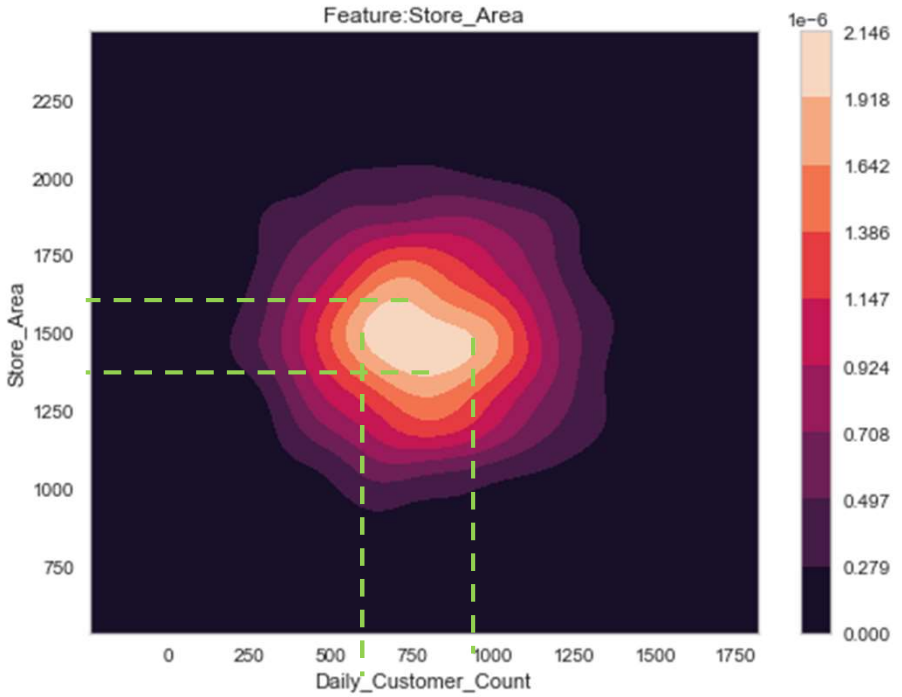
- Store_Area and Items_Available are strongly positively correlated. (This means that larger stores have more items available & smaller stores have small amount of items.)
- There seem to be almost 0 correlation of Sales with any features available in the data set.

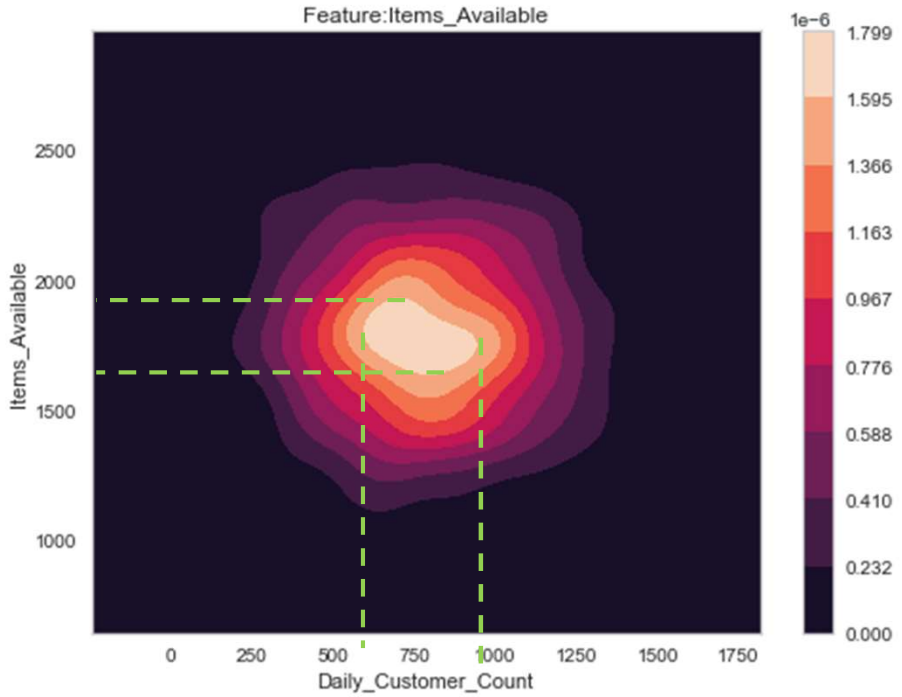
Store Area vs Other Features



- For Store_Sales from $40,000 to 70,000, the Store Area can range from 1350 to 1600 and items available ranges from 1600 to 1900 and daily customer counts between 650 to 950.

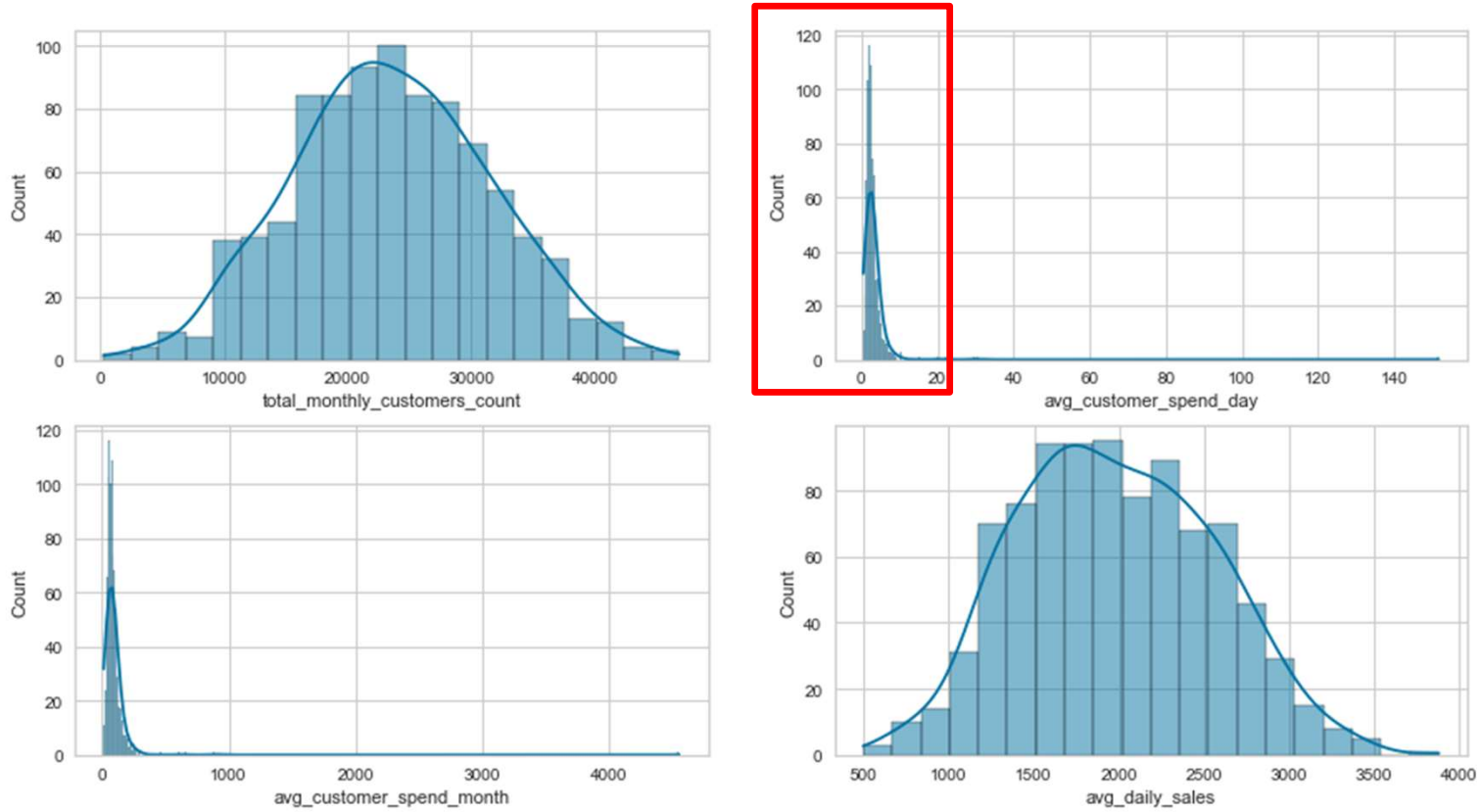
Customer Counts vs Other Features





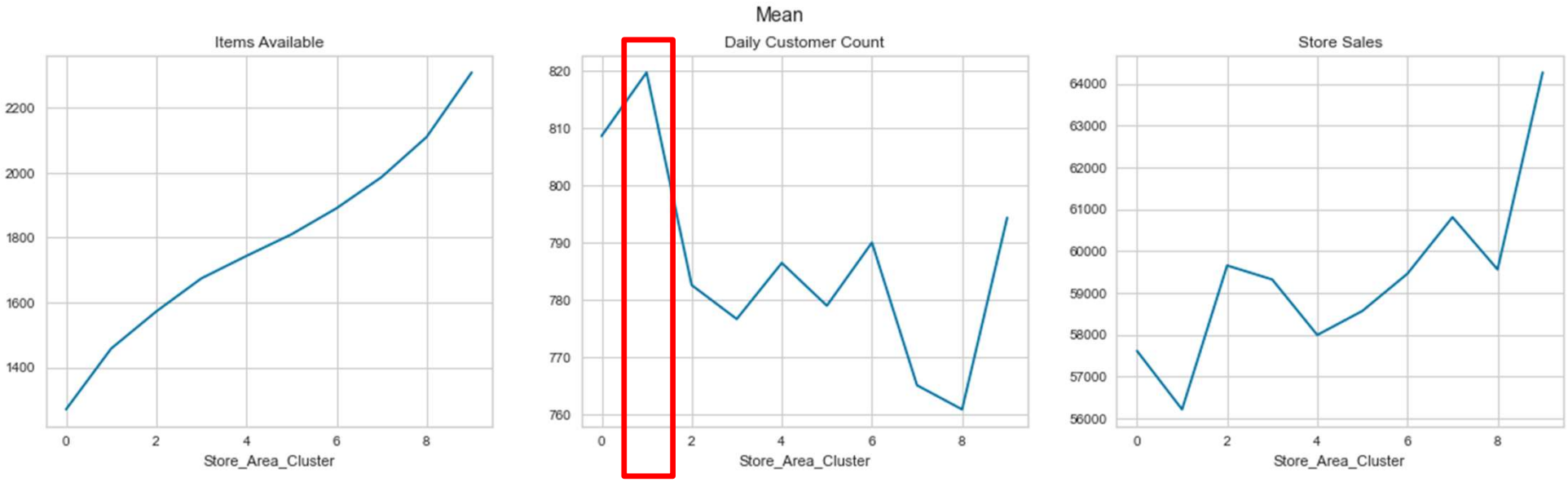
For Customer_Count ranges from 550 to 1000, the store area is usually 1400 to 1600 sq yard and Items_Available ranges from 1600 to 1900

Most of the customers spend less than $3 in a store per day (< $100 per month) Which store?



Cluster 1 Alert:

Highest Daily Customer Count but Lowest Sales !



Most Daily Customers but Lowest Sales due to Lowest Average Purchase per Customer!

Cluster	Store Area Range	Average Number of Items Available	Average Daily Customer Count	Average Sales	avg_purchase
0	(774.999, 1165.0]	1,281	791	57,323	86
1	(1165.0, 1252.0]	1,458	**818**	**56,687**	76
2	(1252.0, 1359.0]	1,572	783	59,658	**147**
3	(1359.0, 1429.0]	1,674	777	59,324	86
4	(1429.0, 1477.0]	1,743	785	58,465	91
5	(1477.0, 1539.0]	1,809	780	59,036	86
6	(1539.0, 1614.5]	1,890	790	59,462	87
7	(1614.5, 1703.0]	1,987	765	60,813	101
8	(1703.0, 1828.0]	2,110	761	59,563	88
9	(1828.0, 2229.0]	2,308	788	63,727	96

Key Insights

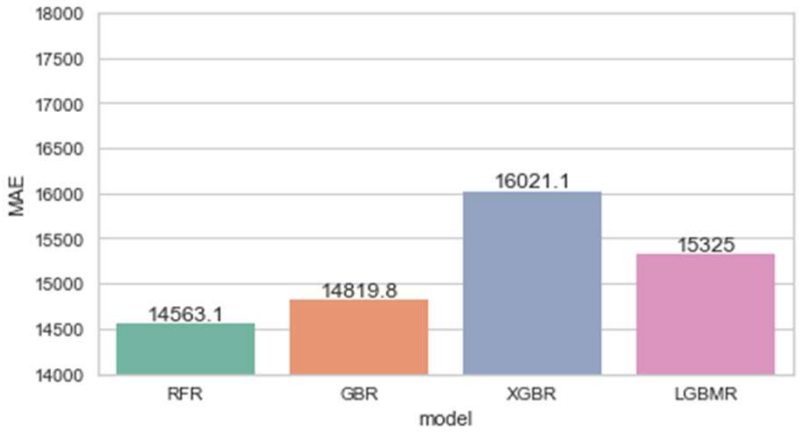
1. There seem to be 0 linear correlation of Store Sales with any features available in the data set.
2. Store_Area and Items_Available are strongly positively correlated. (This means that larger stores have more items available & smaller stores have small amount of items.)
3. Most of the customers spend less than 10 dollars in a store per day.
4. Optimal store area appears to be between 1252 and 1359 (Cluster2) as it has the highest average purchase even though there is not as many daily customers.
5. We do not recommend the highest store area range even though it has more customer and higher historical sales; physical store usually have very high overheads cost ratio, without this information, we would err on the side of caution.

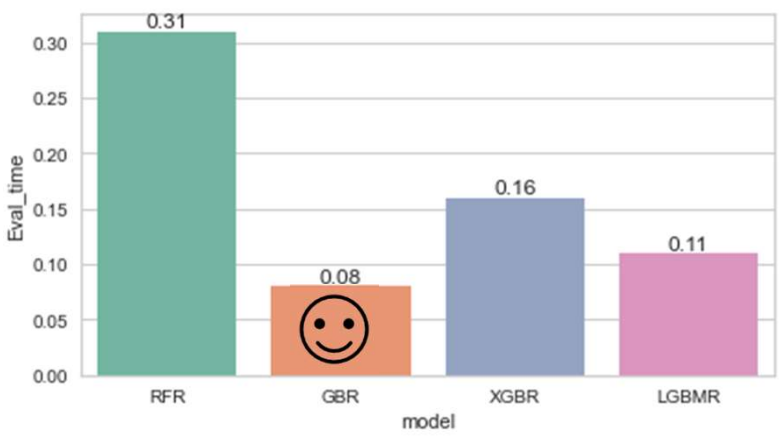
Store Sales Forecast

Modelling for Store Sales Forecast

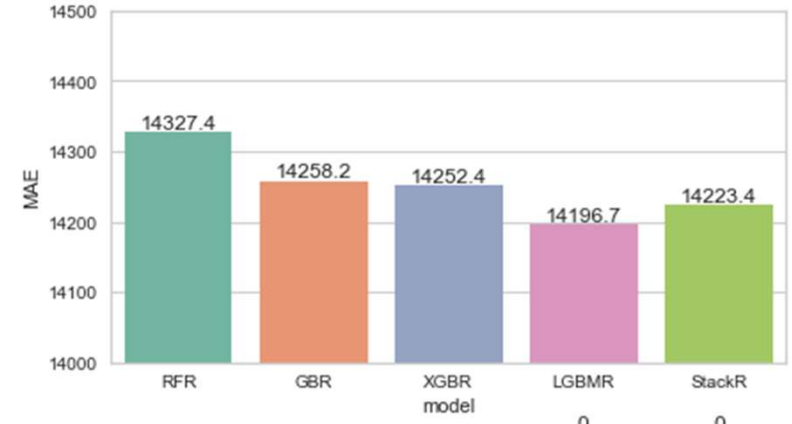
1. Regression Models
    - Random Forest Regressor
    - Gradient Boosting Regressor
    - XG Boosting Regressor
    - Light GBM Regressor
    - Stacking Regressor
2. RandomSearchCV
3. Evaluation of Models
4. Conclusion

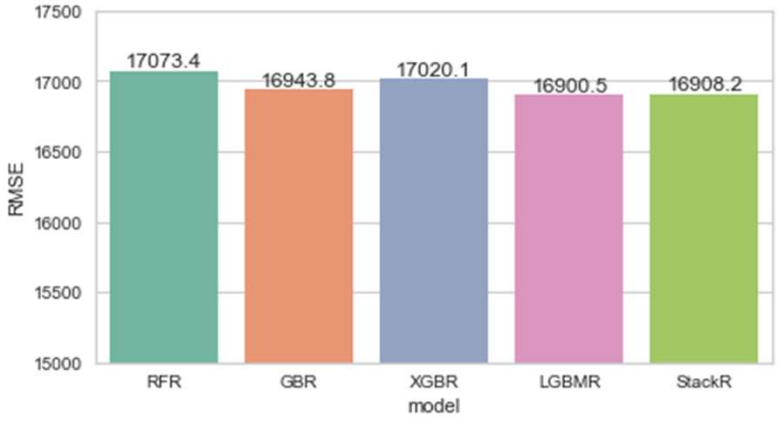
Base Model – RMSE 17k, negative R2?

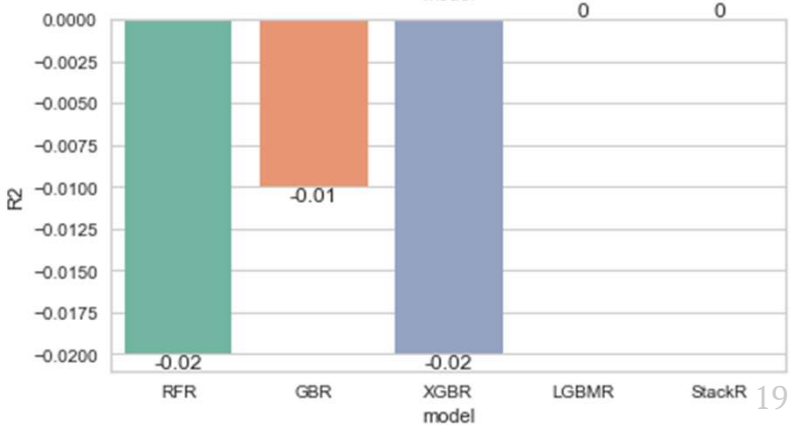


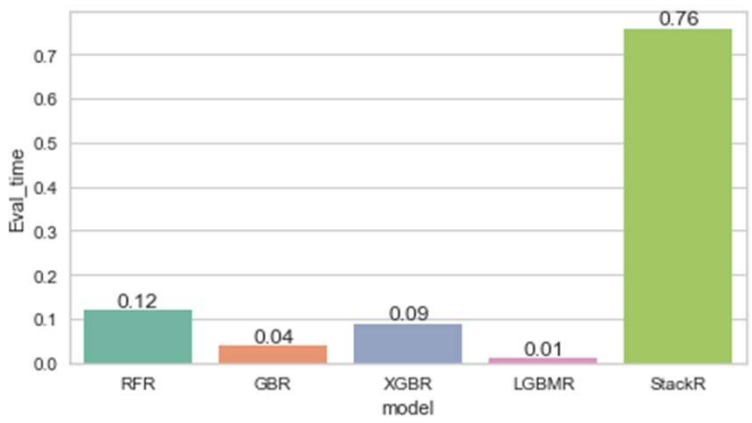


Tuned Model – Still Suboptimal ?

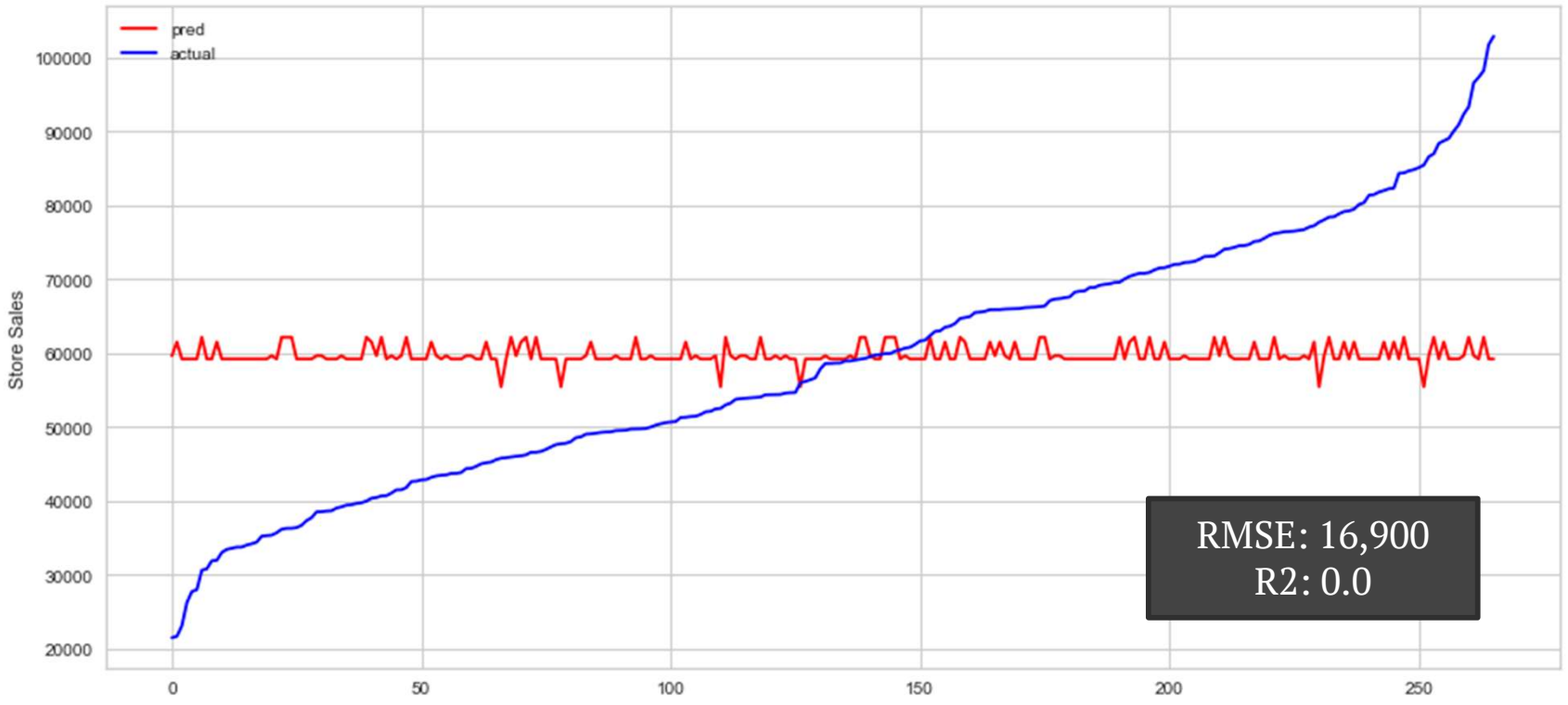






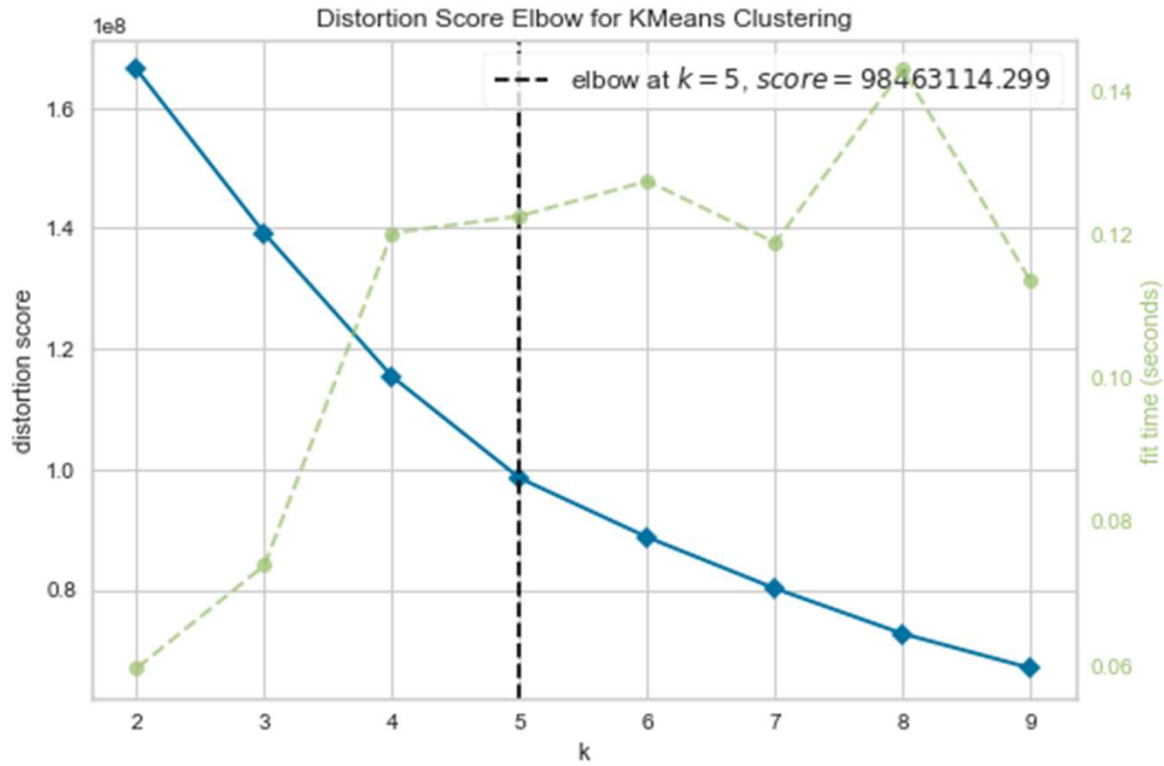


One of the Models- LightGBM

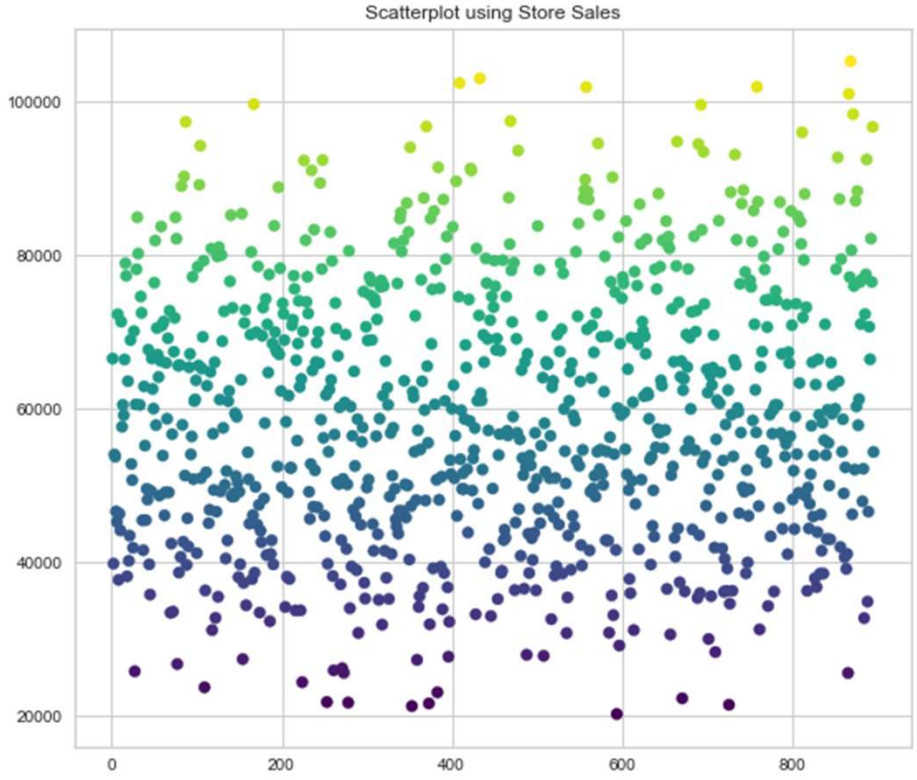


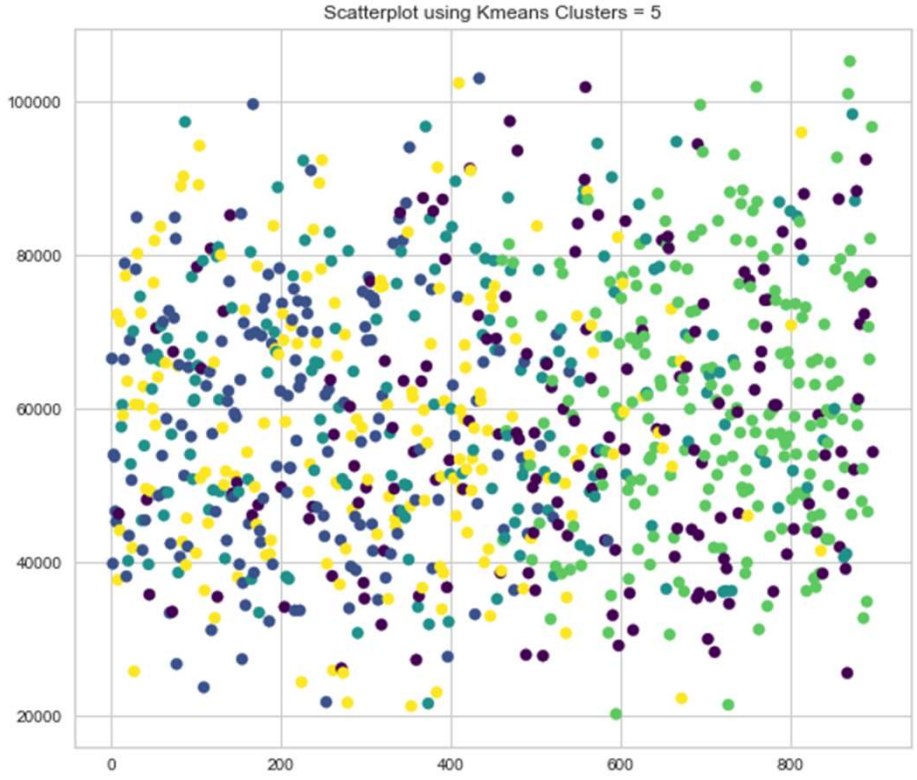
Clustering via KMeans

Optimal Clusters = 5

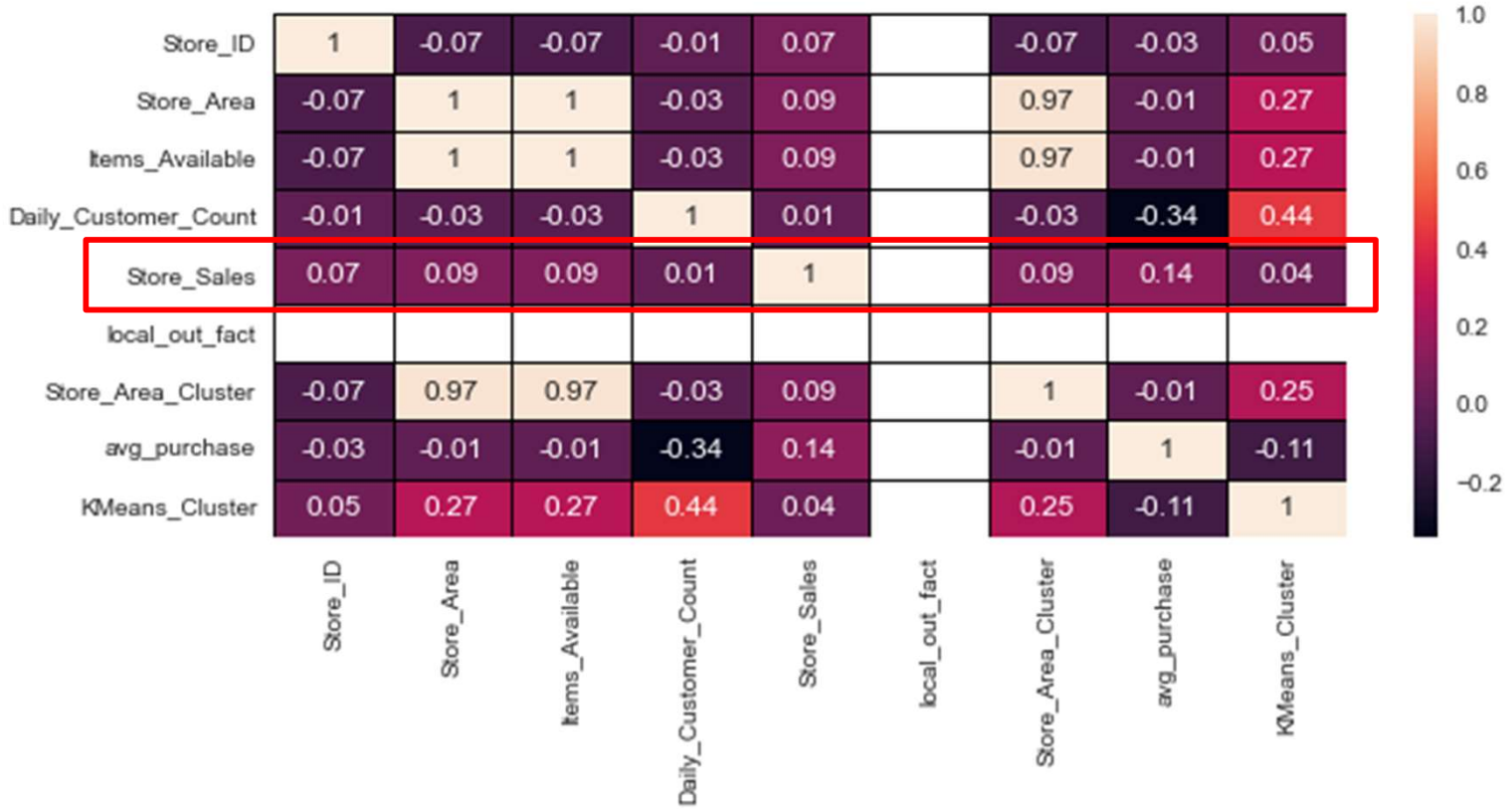


Clustering – no distinct clusters resembling Store Sales Clusters





Correlation Map for All Columns



Answers to Data Science Questions

1. Is there a correlation between store floor area vs store sales $?
   Answer: No linear correlations
2. Is there any other business insights we can provide to?
   Answer: Refer to Slide 15
3. Can we use the available dataset to design a machine learning model to estimate store sales?
   Answer: The current dataset might be adequate for basic descriptive & high level diagnostic analytics[1] but not insufficient for predictive analytics (to design a ML model for store sales forecasting.)

[1] https://online.hbs.edu/blog/post/types-of-data-analysis

Conclusion & Future Works

1. The RMSE need further working, and might require more features to be added (such as area code, name of area manager), considering the dataset only has 5 columns which has very little correlations and limited feature engineering.
2. Though there is a new column on Kmeans cluster, it has minimal linearity vs Store Sales as observed in the scatterplots.

Jupyter Notebook