Capstone Project

Project Title

Seoul Bike Sharing Demand Prediction

BY:- Arkopravo Pradhan

CONTENTS OF THE PRESENTATION

- Problem Statement
- Data Summary
- Exploratory data analysis
- Data wrangling
- Machine Learning models
- Model Explanation
- Conclusion

Problem Statement

The contents in the data belongs to the city called Seoul.
Seoul, officially the Seoul Special City, is the capital and largest metropolis of South Korea. Seoul has a population of 9.7 million people, and forms the heart of the Seoul Capital Area with the surrounding Incheon metropolis and Gyeonggi province. Considered to be a global city, Seoul was the world's 4th largest metropolitan economy in 2014 after Tokyo, New York City and Los Angeles.

Bike rentals have became a popular service in recent years and it seems people are using it more often. With relatively cheaper rates and ease of pick up and drop at own convenience is what making this business excel. Mostly used by people having no personal vehicles and also to avoid congested public transport which follows its own time.

Therefore, the business to strive and profit more, it has to be always ready and supply no. of bikes at different locations, to fulfil the demand a pre planned set of bike count values can therefore, be a handy solution to meet all demands

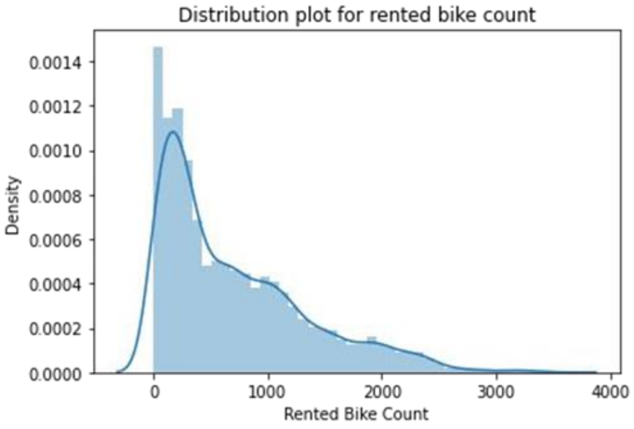
Data Summary

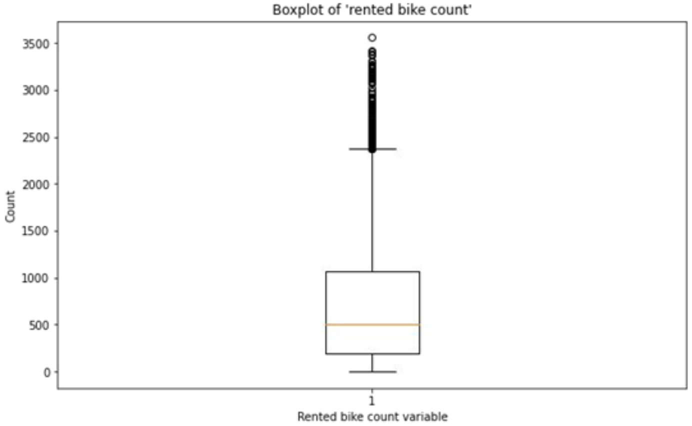
- We have a variable 'date' which tells us the day on which some bikes were rented
- We have a variable 'hour' which tells which hour of the day the bikes were rented.
- We have some numerical type variables such as temperature, humidity, wind, visibility, dew point, temp, solar radiation, rainfall, snowfall which tells the environment conditions at that particular hour of the day.
- We have some categorical variable such as seasons, is it a holiday and is it a functioning day or not.
- And finally we have 'rented bike count' variable which we need to predict for new observations given the other variables.

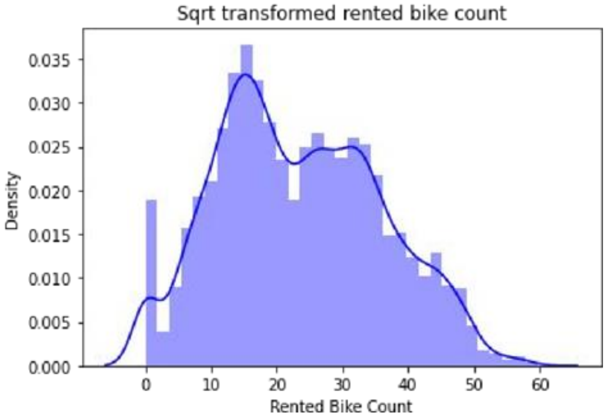
- Dataset dimension: – 8760 rows and 14 columns

EDA - Univariate analysis

- “Rental bike count”: – dependent variable
- Outliers are above 2500.
- Was moderately skewed(positive)
- Skewness: 1.153428 Skewness after transformation: 0.237362
- Kurtosis: 0.853387 Kurtosis after transformation: -0.657201

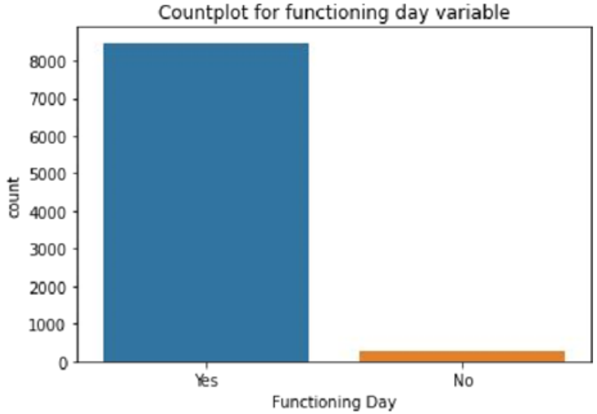


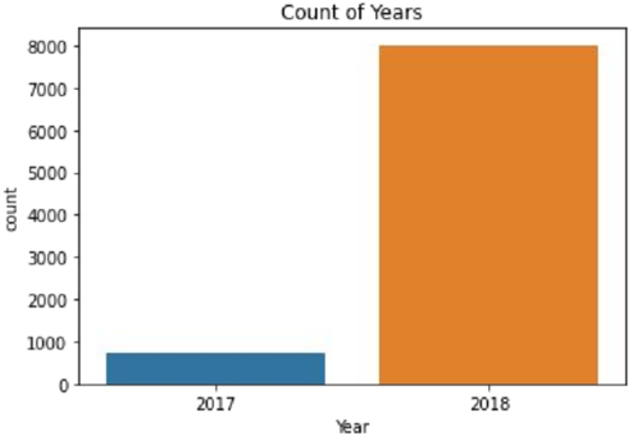


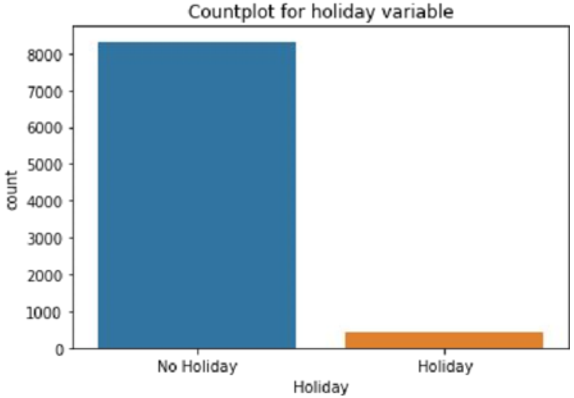


Univariate analysis - Categorical variables

- Functioning day and holiday had majority of one class around 97% and 95%.
- Majority of observations were from 2018.
- Decided to drop them.

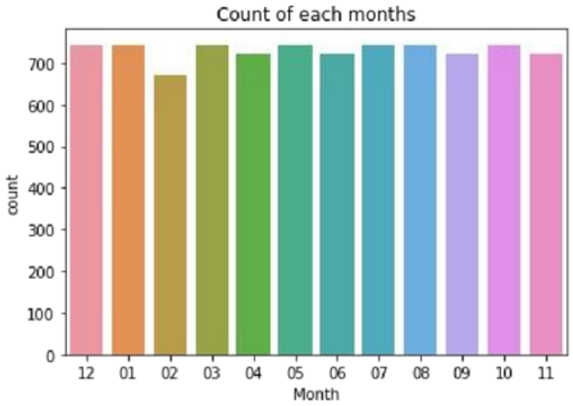


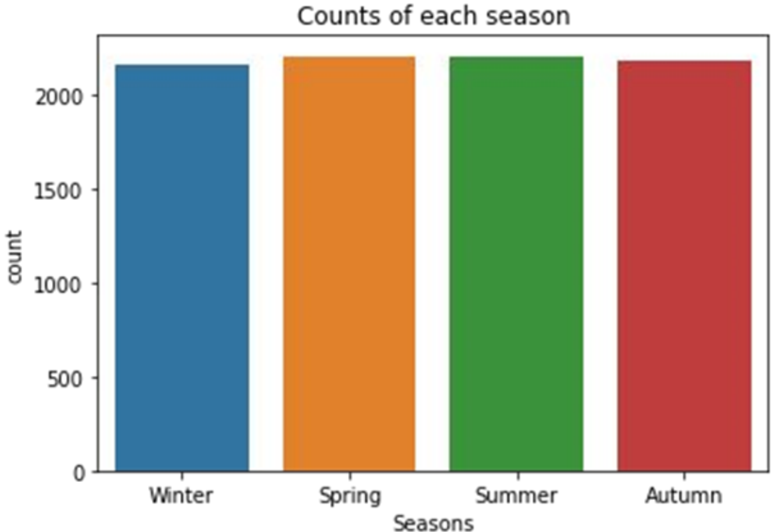


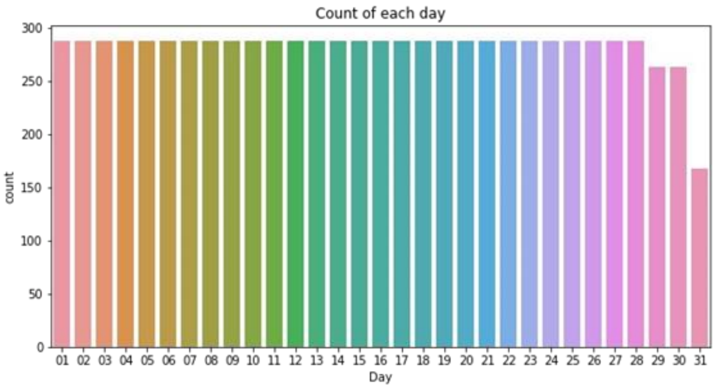


Univariate analysis- Categorical variables

- Seasons, day, month had almost equal no. of observations for each.

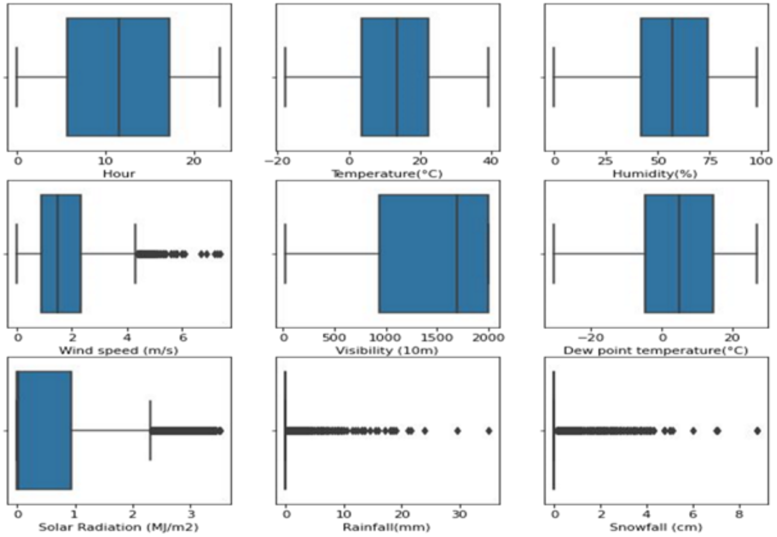


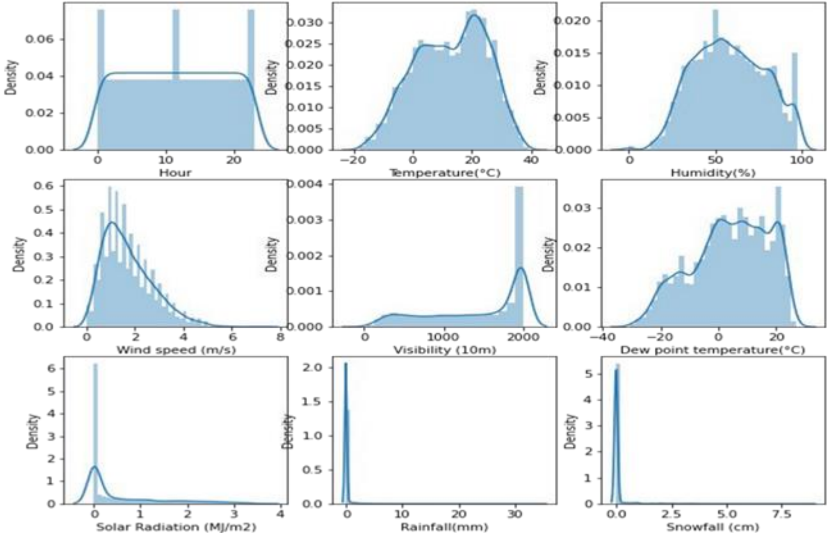




Univariate analysis - Numerical variables

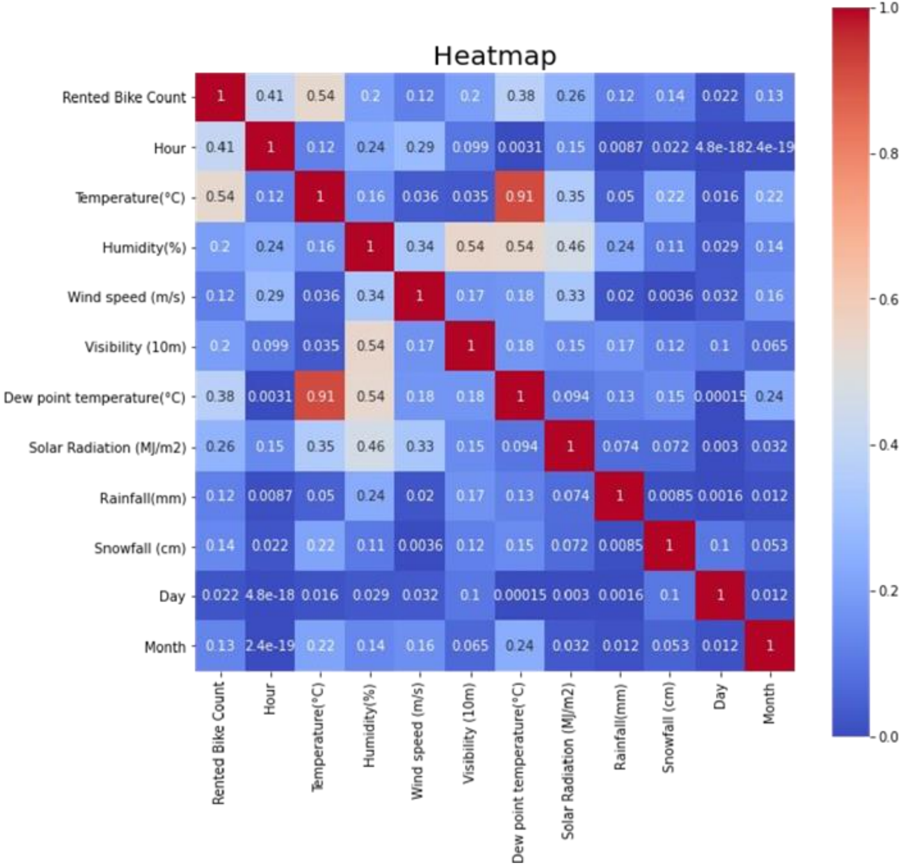
- ❖ Variables such as snowfall, rainfall had mostly zero values.
- ❖ Decided to drop them.





Multivariate analysis - Correlation

- Dew point temperature and Temperature were highly correlated.
- Linear regression assumes that independent variables must show some linear relationship with dependent variable.
- No such relationship seen here.
- Linear regression might not perform well.

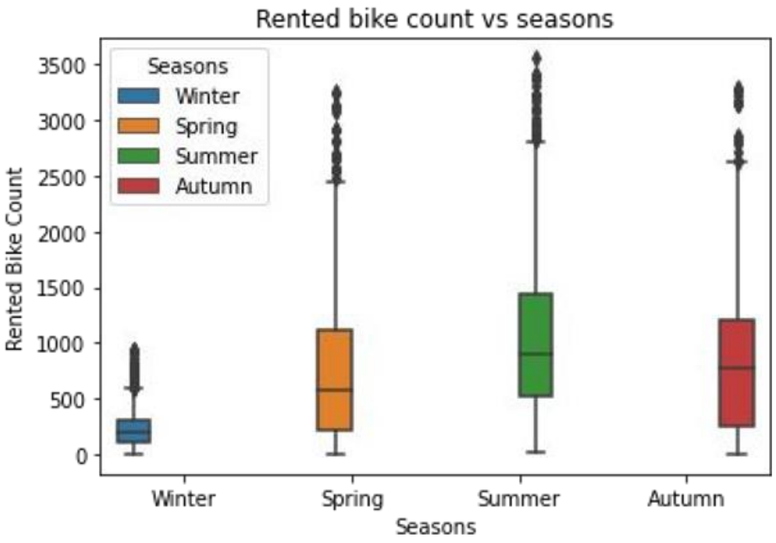


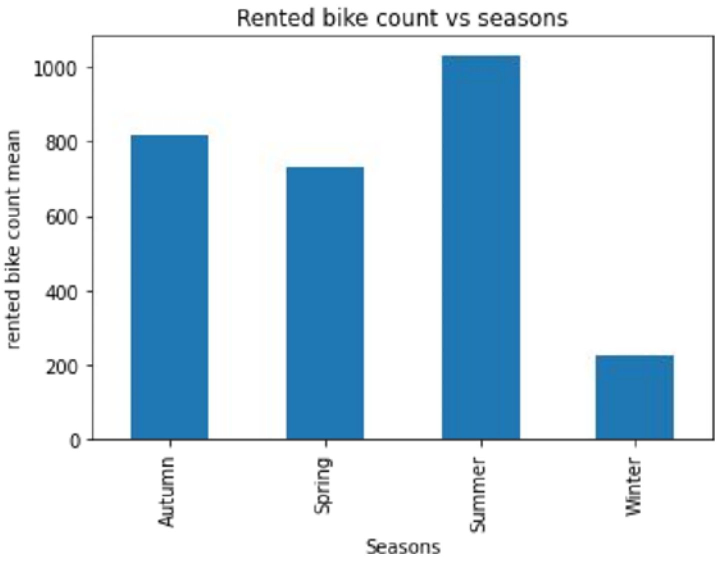
Heatmap

	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Day	Month
Rented Bike Count	1	0.41	0.54	0.2	0.12	0.2	0.38	0.26	0.12	0.14	0.022	0.13
Hour	0.41	1	0.12	0.24	0.29	0.099	0.0031	0.15	0.0087	0.022	4.8e-18	2.4e-19
Temperature(°C)	0.54	0.12	1	0.16	0.036	0.035	0.91	0.35	0.05	0.22	0.016	0.22
Humidity(%)	0.2	0.24	0.16	1	0.34	0.54	0.54	0.46	0.24	0.11	0.029	0.14
Wind speed (m/s)	0.12	0.29	0.036	0.34	1	0.17	0.18	0.33	0.02	0.0036	0.032	0.16
Visibility (10m)	0.2	0.099	0.035	0.54	0.17	1	0.18	0.15	0.17	0.12	0.1	0.065
Dew point temperature(°C)	0.38	0.0031	0.91	0.54	0.18	0.18	1	0.094	0.13	0.15	0.00015	0.24
Solar Radiation (MJ/m2)	0.26	0.15	0.35	0.46	0.33	0.15	0.094	1	0.074	0.072	0.003	0.032
Rainfall(mm)	0.12	0.0087	0.05	0.24	0.02	0.17	0.13	0.074	1	0.0085	0.0016	0.012
Snowfall (cm)	0.14	0.022	0.22	0.11	0.0036	0.12	0.15	0.072	0.0085	1	0.1	0.053
Day	0.022	4.8e-18	0.016	0.029	0.032	0.1	0.00015	0.003	0.0016	0.1	1	0.012
Month	0.13	2.4e-19	0.22	0.14	0.16	0.065	0.24	0.032	0.012	0.053	0.012	1

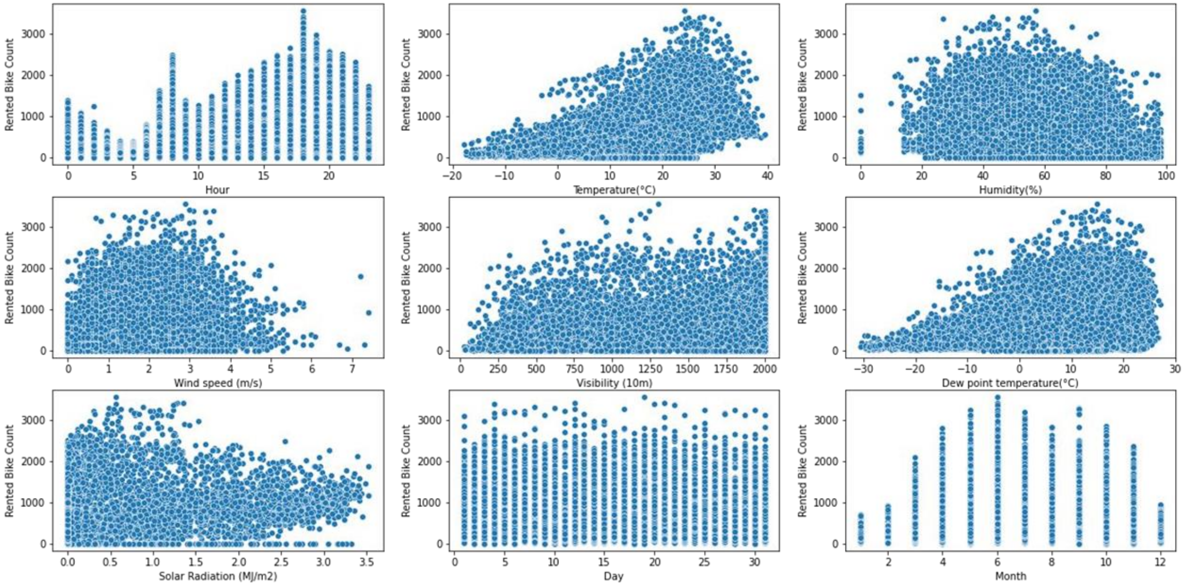
Multivariate analysis

- Demand for bikes was high during summer compared to winter



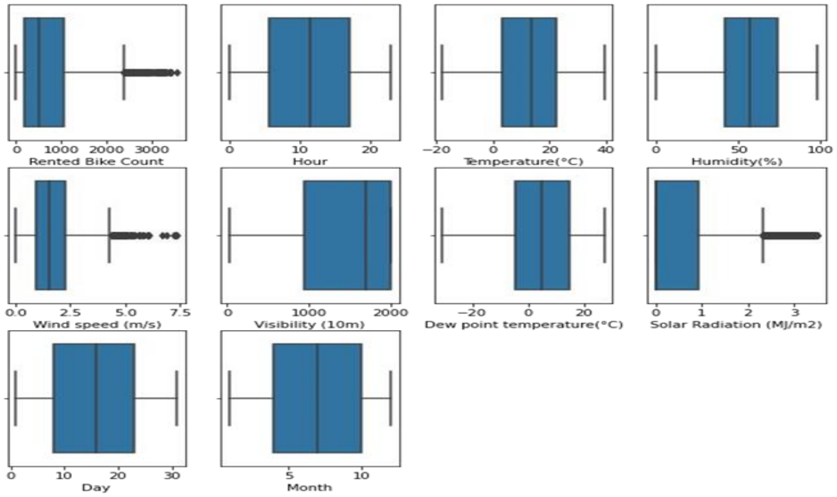


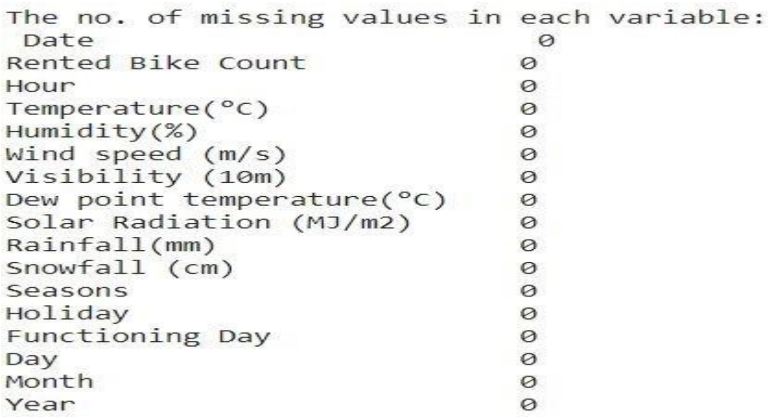
Multivariate analysis



Data Wrangling - missing values and outliers

- No missing values
- Tackled outliers in rented bike count by applying transformation
- Windspeed and solar radiation had outliers, but they were not that far from maximum values.



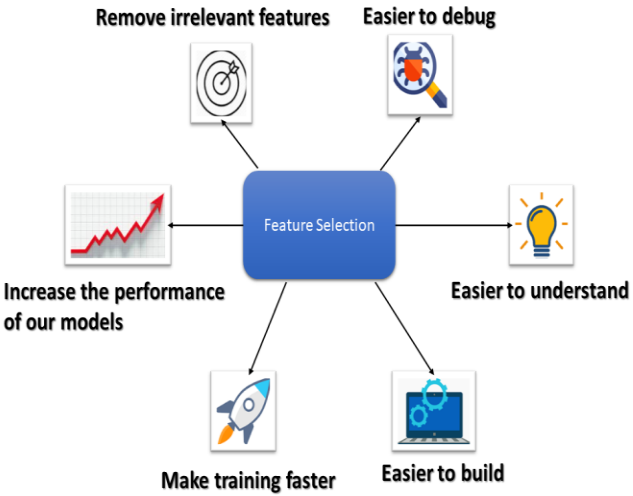


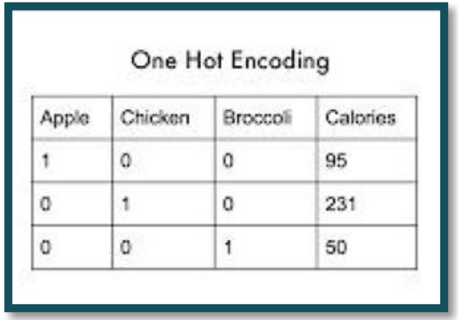
```
The no. of missing values in each variable:
 Date                          0
Rented Bike Count             0
Hour                          0
Temperature(°C)               0
Humidity(%)                   0
Wind speed (m/s)              0
Visibility (10m)              0
Dew point temperature(°C)     0
Solar Radiation (MJ/m2)       0
Rainfall(mm)                  0
Snowfall (cm)                 0
Seasons                       0
Holiday                       0
Functioning Day               0
Day                           0
Month                         0
Year                          0
```

Data wrangling - feature selection

Removed columns :

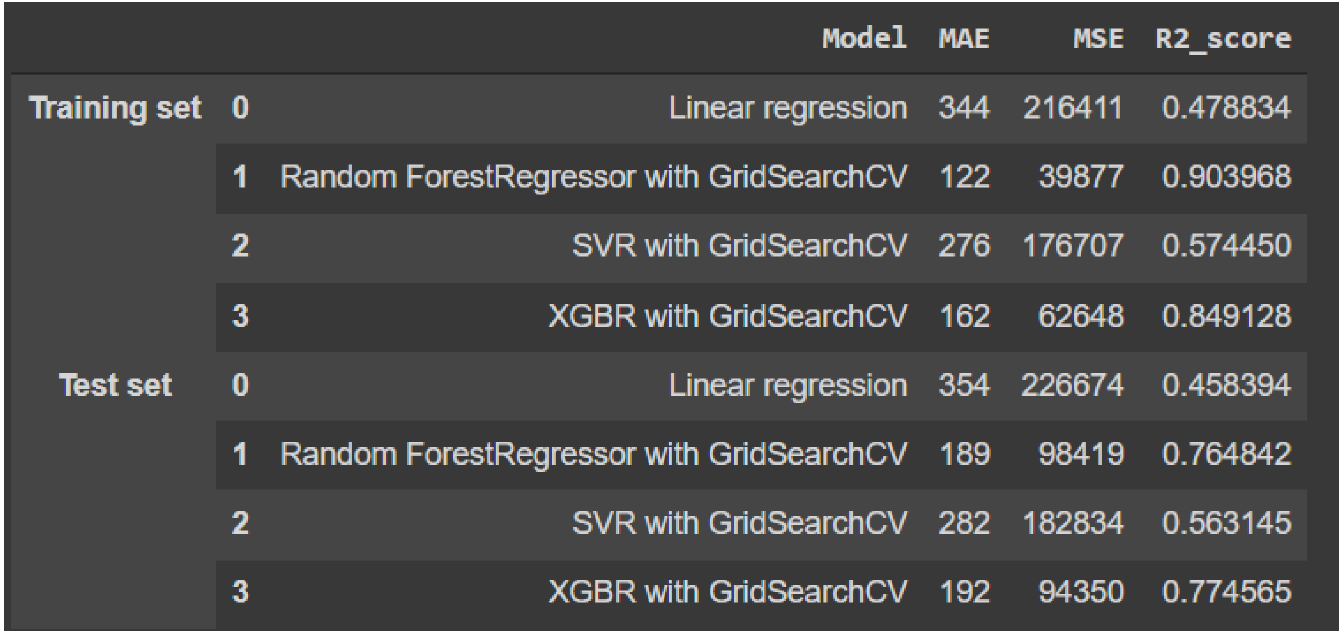
- Functioning day, Holiday, Year - Had majority of one class
- Snowfall, Rainfall - Had mostly 0
- Dew point temperature - highly correlated with temperature Encoding:
- One hot encoded Season




One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

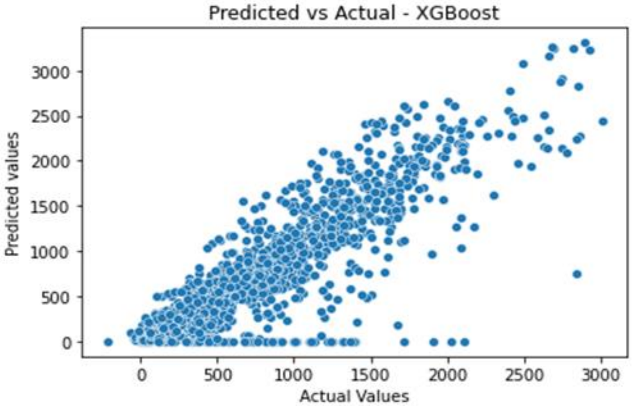
Machine learning Models

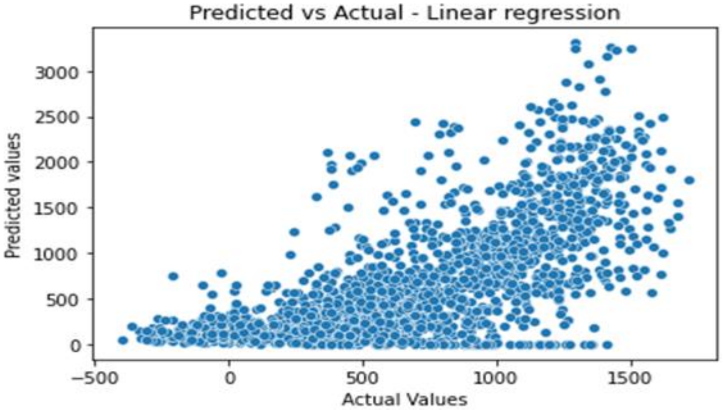


		Model	MAE	MSE	R2_score
Training set	0	Linear regression	344	216411	0.478834
	1	Random ForestRegressor with GridSearchCV	122	39877	0.903968
	2	SVR with GridSearchCV	276	176707	0.574450
	3	XGBR with GridSearchCV	162	62648	0.849128
Test set	0	Linear regression	354	226674	0.458394
	1	Random ForestRegressor with GridSearchCV	189	98419	0.764842
	2	SVR with GridSearchCV	282	182834	0.563145
	3	XGBR with GridSearchCV	192	94350	0.774565

Machine Learning Models

- 1. Linear Regression : Underfit Model
- 2. SVR with GridsearchCV : Below Average Model
- 3. Random ForestRegressor with GridsearchCV : Overfit Model
- 4. XGBR with GridsearchCV : Best Model(According to observation)





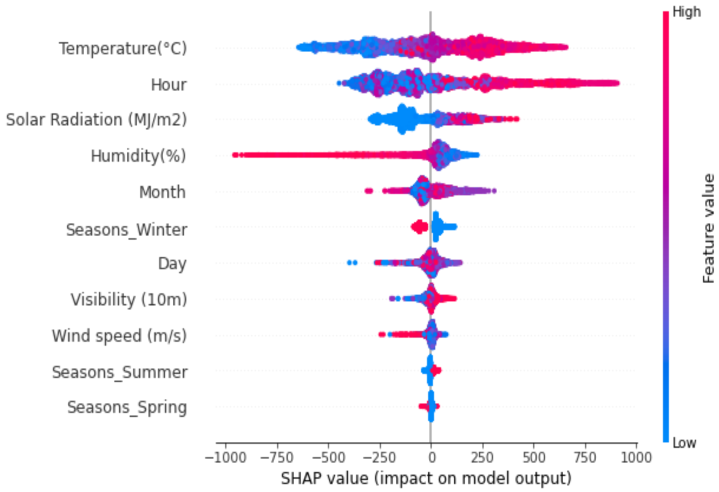
Best Model

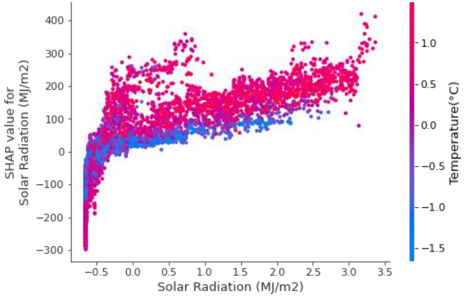
- This the best-case scenario.
- Best Parameters:
- 'gamma': 0, 'learning_rate': 0.27, 'max_depth': 4, 'min_child_weight': 1, 'n_estimators': 48, 'subsample': 0.6
- Training set: The evaluation metric values for training set - XGBR with GridSearchCV:
- The MAE of training set = 161.74367356417727
- The MSE of training set = 62648.394378423785
- The R2_score of training set = 0.8491283627560362

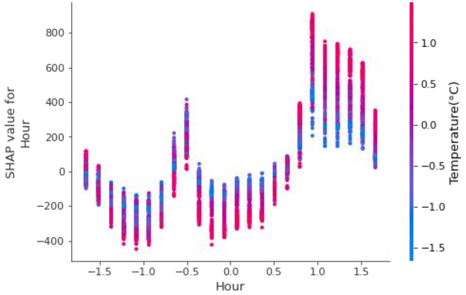
- Test Set: The evaluation metric values for test set - XGBR:
- The MAE of test set = 192.1705759935183
- The MSE of test set = 94349.74074271113
- The R2_score of test set = 0.7745647586402594

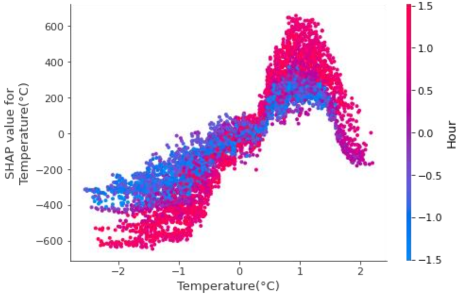
Model Explanation

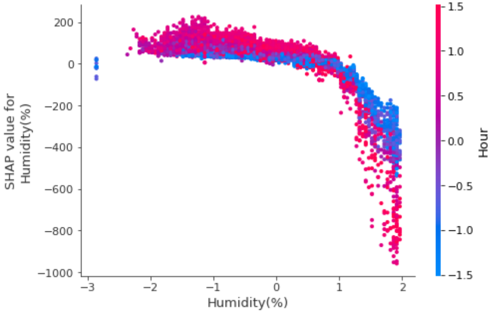
- The most important features were Temperature, Hour, Humidity, Month.











Conclusion

- We saw underfitting scenario in Linear regression and overfitting in Random Forest but the best performance was given by the XGBoost model.
- One of the challenges faced was to tune the hyperparameters, and find the best values which gave a better model.
- We also implemented shap techniques to understand the working of our XGBoost model and found out:
  1. Temperature was the second most important feature. Demand for bikes was higher when temperature was high.
  2. Hour of the day had the most impact on predicting values. Demand was high during evening and night hours.
  3. Demand was high for lower values of windspeed and solar radiation.
  4. Demand was less in winters as compared to other seasons.
- As this data is time dependent, the values for variables like temperature, windspeed, solar radiation etc., will not always be consistent. Therefore, there will be scenarios where the model might not perform well. As Machine learning is an exponentially evolving field, we will have to be prepared for all contingencies and also keep checking our model from time to time.

Thank You!