Capstone Project - Regression

Bike Sharing Demand Prediction

-Abhinav Singh

FLOW OF THE PRESENTATION:

- OBJECTIVE
- DATASET DESCRIPTION
- EXPLORATORY DATA ANALYSIS (EDA)
- LINEAR MODELS
- DECISION TREE BASED MODELS
- CONCLUSION

OBJECTIVE:

In this age and time, many of the urban cities around the world harbor considerably large population. Owing to the large population,daily commuters need mobility options which are available close at hand, and that is why many urban cities around the world have introduced rental bikes system and one such city is ‘Seoul’, whose data we are going explore.

Providing the commuters with stable supply of rental bikes at each hour the day becomes a challenge. So, the objective of the project is to build a regression model that predicts the count of rental bikes required at each hour everyday.

DATASET DESCRIPTION:

Date : A particular day.
Rented Bike count - Count of bikes rented at each hour
Hour - Hour of the day (0-23)
Temperature- Temperature at particular hour in Celsius
Humidity - %age humidity at particular hour
Wind speed - Speed of wind at particular hour (m/s)
Visibility 10m - Visibility at particular hour.
Dew point temperature - Dew point temperature at particular hour in Celsius

Solar radiation - Solar radiation at particular hour (MJ/m2)
Rainfall - Rainfall received at particular hour (mm)
Snowfall - Snowfall received at particular hour (cm)
Seasons - Winter, Spring, Summer, Autumn
Holiday - Holiday/No holiday
Functional Day - Whether the service provider was functioning at particular day/hour (Yes/No).

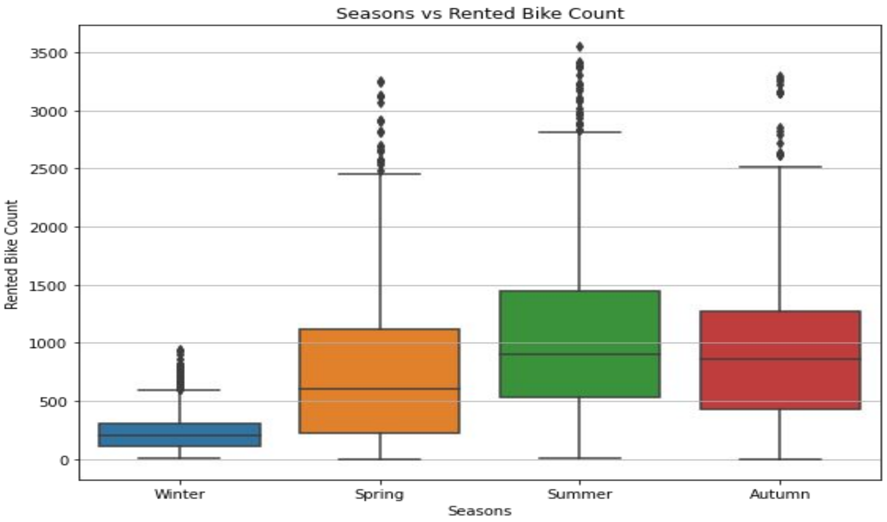
These are the 14 features present in our dataset and the total number of records present are 8,760.

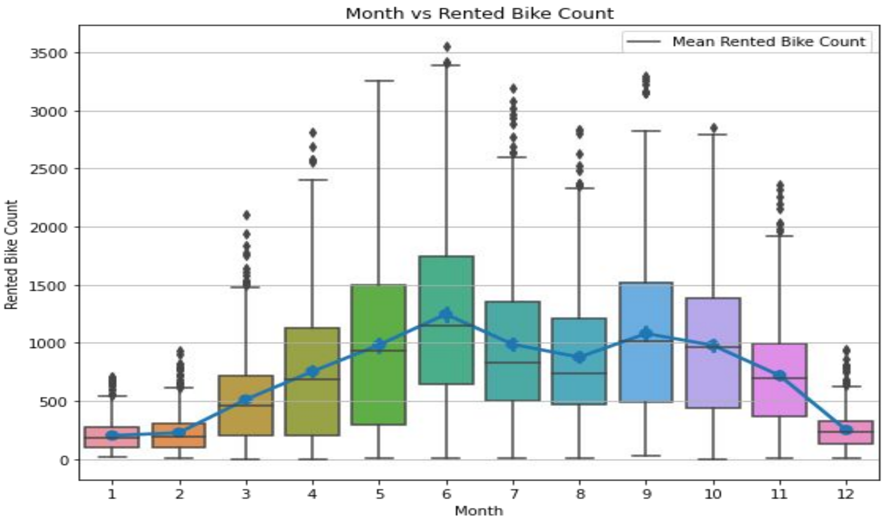
EDA:

The time duration of our dataset is from: December 2017 to November 2018.

Out of the 8,760 records, there are 295 records where the service provider was not functioning and the rented bike count is zero (0).

1) Season/Month wise bike demand:





Winter Season: December, January, February (12,1,2), Spring Season: March, April, May (3,4,5), Summer Season: June, July, August (6,7,8), Autumn Season: September, October, November (9,10,11)

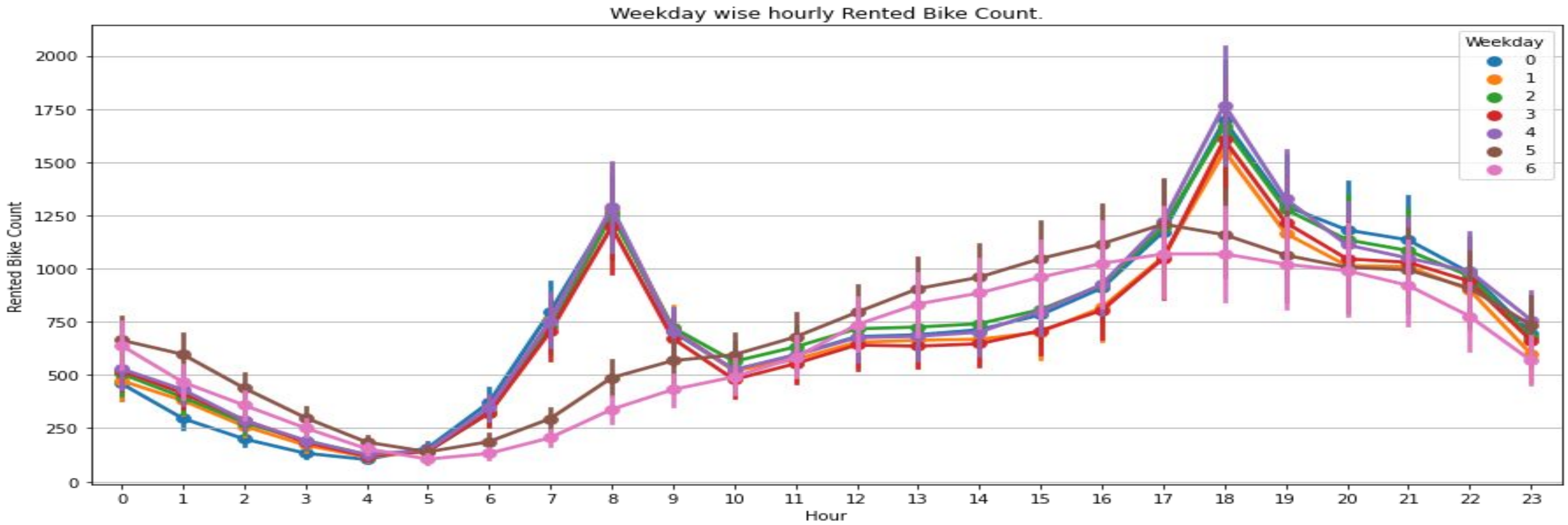
Seasons vs Rented Bike Count boxplot: The highest median bike demand is in Summer, followed by Autumn, Spring and Winter seasons.

Months vs Rented Bike Count boxplot:

- Similar to seasons, the demand in the winter months (12,1,2) is the lowest.
- Come Spring, the demand starts increasing nearly linearly, till June (6) in Summer season.
- After that the demand falls in the month of July & August (7 & 8), followed by a spike in the month of September (9) and then again falling in October & November.

Relation of bike demand with Seasons/Months is non monotonic.

2) Weekday wise hourly bike demand:

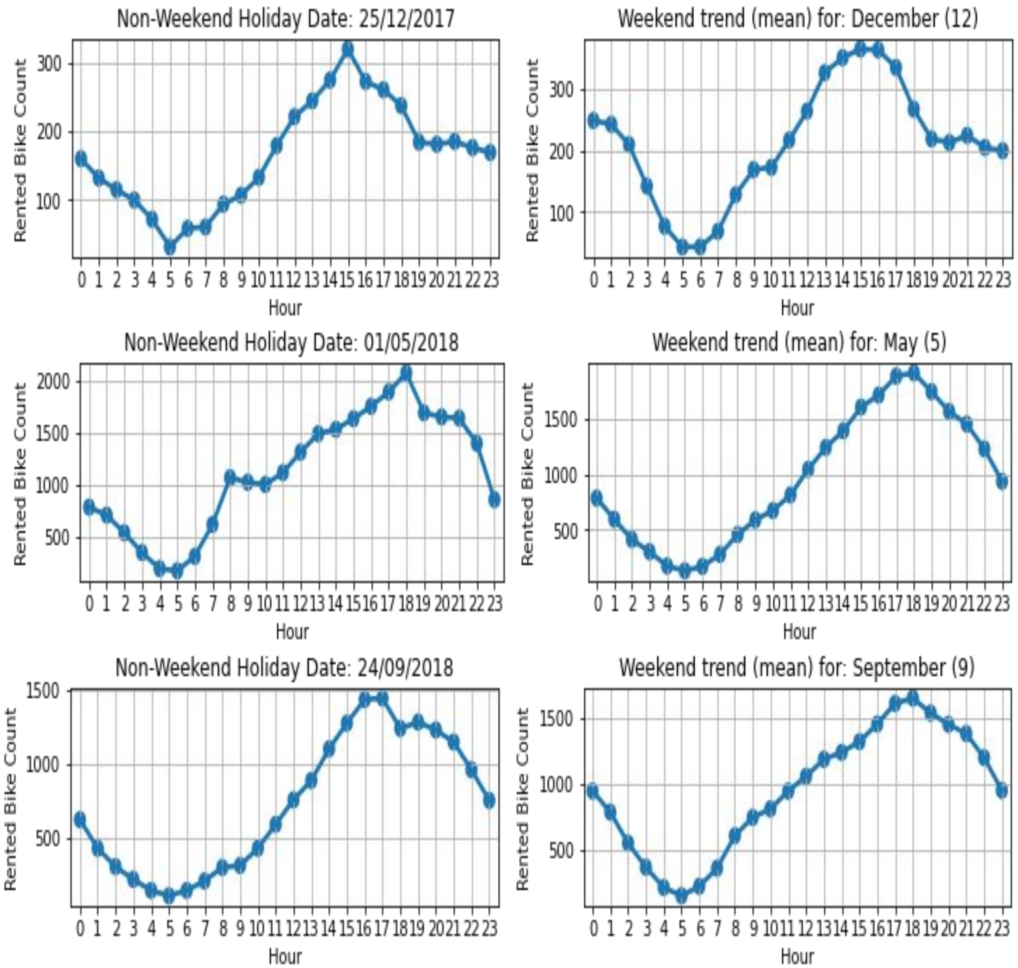


Over the year:

- The pattern of bike demands is similar for weekdays, ie Monday to Friday (0-4), with peak/rush around 8 am and then peak around 6pm, indicating these are office going and returning hours, respectively.
- The pattern of bike demand for weekends, ie Saturday & Sunday (5 & 6), is different than weekdays, with demand gradually increasing and reaching maximum around 5-6 pm, indicating that people have their weekends off.
- Post evening (after 6pm) and post midnight (0-5 am) hours are fairly similar in trend for all seven days of the week.

Overall, the relation of bike demand over the whole day is non monotonic.

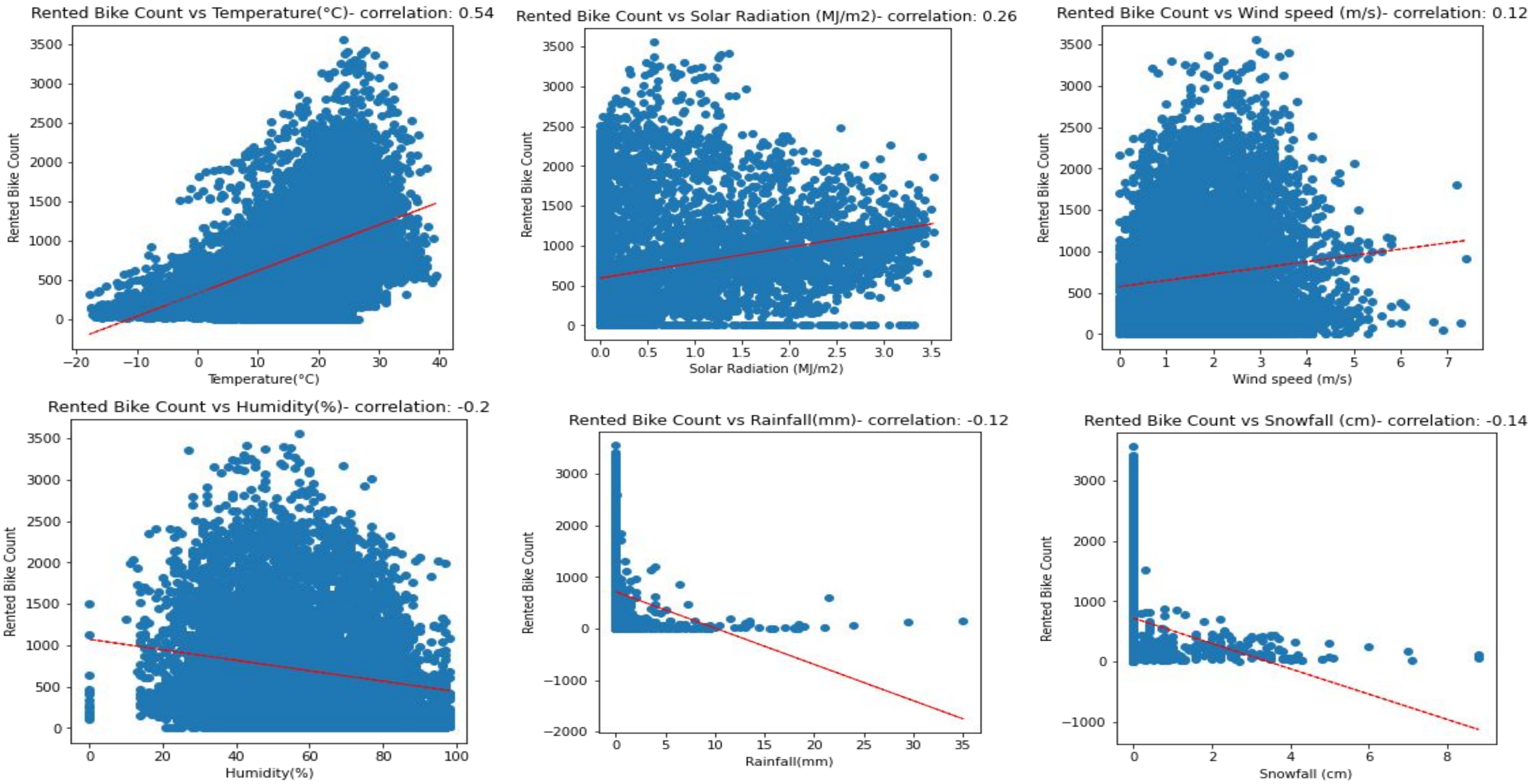
3) Bike demand on non-weekend Holidays:



- On left, we are looking at the demand of bikes on some of the non weekend holidays across months compared to weekend trend in that particular month

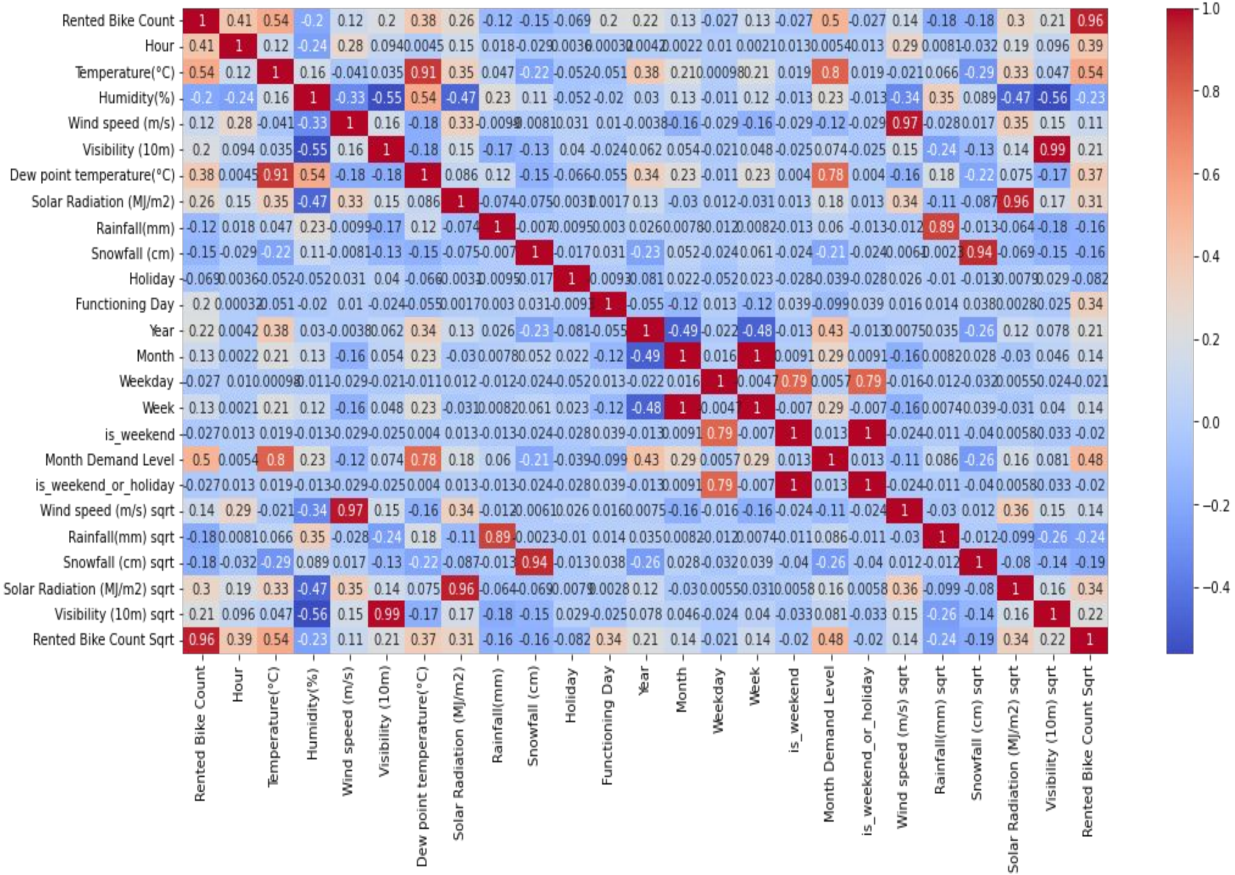
- Overall, it was observed that the trend of rented bike demand was similar to that of on weekends.

4) Bike demand vs Numerical Variables:



LINEAR MODELS:

Feature Selection- Numerical Variables:



Highly correlated features-

- Temperature & Dew Pt. Temperature: **0.91**
- Humidity & Visibility: **-0.55**
- Humidity & Solar Radiation: **-0.47**

	variables	VIF
0	Temperature(°C)	3.125607
1	Humidity(%)	2.633308
2	Rainfall(mm) sqrt	1.118176
3	Snowfall (cm) sqrt	1.184295
4	Solar Radiation (MJ/m2) sqrt	1.749015

Numerical features after VIF analysis.

Statsmodel OLS regression:

```
==============================================================================
Dep. Variable:      Rented Bike Count   R-squared:                       0.760
Model:                            OLS   Adj. R-squared:                  0.759
Method:                 Least Squares   F-statistic:                     966.6
Date:                Wed, 25 May 2022   Prob (F-statistic):               0.00
Time:                        16:46:04   Log-Likelihood:                -19772.
No. Observations:                6132   AIC:                         3.959e+04
Df Residuals:                    6111   BIC:                         3.973e+04
Df Model:                          20
Covariance Type:            nonrobust
=================================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                        -1.9448      0.351     -5.541      0.000      -2.633      -1.257
Temperature(°C)               2.3011      0.175     13.128      0.000       1.958       2.645
Rainfall(mm) sqrt            -2.8240      0.086    -33.022      0.000      -2.992      -2.656
Solar Radiation (MJ/m2) sqrt  2.4612      0.198     12.454      0.000       2.074       2.849
Humidity(%)                  -1.8707      0.114    -16.468      0.000      -2.093      -1.648
Snowfall (cm) sqrt           -0.0337      0.085     -0.398      0.691      -0.200       0.132
is_weekend                   -1.5120      0.174     -8.699      0.000      -1.853      -1.171
Holiday                      -3.3738      0.374     -9.018      0.000      -4.107      -2.640
Functioning Day              28.3309      0.461     61.501      0.000      27.428      29.234
Month Demand Level_1         -8.7246      0.287    -30.391      0.000      -9.287      -8.162
Month Demand Level_2         -4.3092      0.265    -16.240      0.000      -4.829      -3.789
Month Demand Level_3         -0.0105      0.192     -0.055      0.956      -0.388       0.367
Month Demand Level_4          1.8884      0.189      9.987      0.000       1.518       2.259
Month Demand Level_5          3.4603      0.266     12.986      0.000       2.938       3.983
Month Demand Level_6          5.7508      0.284     20.246      0.000       5.194       6.308
hour_window_0-2              -1.7715      0.253     -6.994      0.000      -2.268      -1.275
hour_window_12-14            -3.8612      0.320    -12.054      0.000      -4.489      -3.233
hour_window_15-17             0.5904      0.262      2.249      0.025       0.076       1.105
hour_window_18-20             8.9999      0.227     39.623      0.000       8.555       9.445
hour_window_21-23             5.3299      0.260     20.520      0.000       4.821       5.839
hour_window_3-5              -8.5598      0.254    -33.743      0.000      -9.057      -8.063
hour_window_6-8               0.9156      0.229      3.993      0.000       0.466       1.365
hour_window_9-11             -3.5881      0.277    -12.975      0.000      -4.130      -3.046
==============================================================================
Omnibus:                      121.087   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              245.833
Skew:                           0.078   Prob(JB):                     4.15e-54
Kurtosis:                       3.968   Cond. No.                     2.57e+16
==============================================================================
```

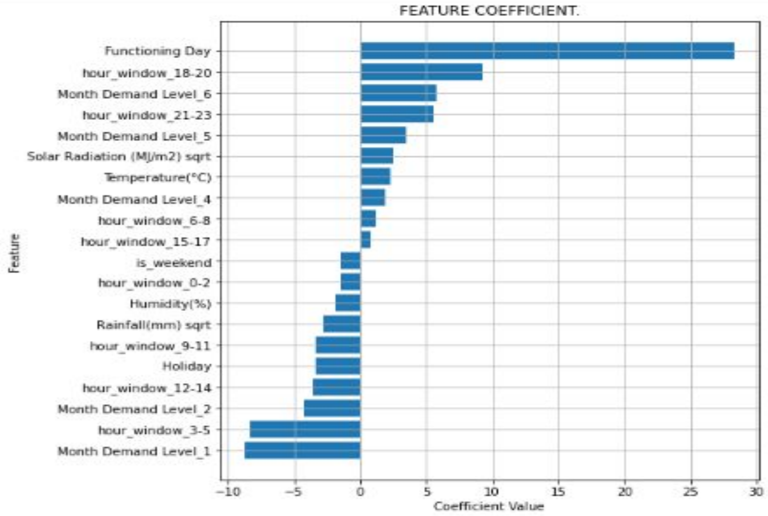
```
==============================================================================
Dep. Variable:      Rented Bike Count   R-squared:                       0.760
Model:                            OLS   Adj. R-squared:                  0.759
Method:                 Least Squares   F-statistic:                     1018.
Date:                Wed, 25 May 2022   Prob (F-statistic):               0.00
Time:                        16:53:01   Log-Likelihood:                -19772.
No. Observations:                6132   AIC:                         3.958e+04
Df Residuals:                    6112   BIC:                         3.972e+04
Df Model:                          19
Covariance Type:            nonrobust
=================================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                        -1.9497      0.423     -4.605      0.000      -2.780      -1.120
Temperature(°C)               2.3092      0.174     13.263      0.000       1.968       2.650
Rainfall(mm) sqrt            -2.8216      0.085    -33.077      0.000      -2.989      -2.654
Solar Radiation (MJ/m2) sqrt  2.4569      0.197     12.452      0.000       2.070       2.844
Humidity(%)                  -1.8805      0.111    -16.961      0.000      -2.098      -1.663
is_weekend                   -1.5095      0.174     -8.690      0.000      -1.850      -1.169
Holiday                      -3.3689      0.374     -9.010      0.000      -4.102      -2.636
Functioning Day              28.3272      0.461     61.509      0.000      27.424      29.230
Month Demand Level_1         -8.7296      0.309    -28.292      0.000      -9.334      -8.125
Month Demand Level_2         -4.2893      0.333    -12.878      0.000      -4.942      -3.636
Month Demand Level_4          1.9004      0.285      6.667      0.000       1.342       2.459
Month Demand Level_5          3.4705      0.370      9.376      0.000       2.745       4.196
Month Demand Level_6          5.7628      0.378     15.260      0.000       5.022       6.503
hour_window_0-2              -1.7706      0.256     -6.929      0.000      -2.272      -1.270
hour_window_12-14            -3.8636      0.321    -12.050      0.000      -4.492      -3.235
hour_window_15-17             0.5877      0.263      2.236      0.025       0.072       1.103
hour_window_18-20             8.9958      0.229     39.305      0.000       8.547       9.444
hour_window_21-23             5.3299      0.262     20.355      0.000       4.817       5.843
hour_window_3-5              -8.5582      0.256    -33.429      0.000      -9.060      -8.056
hour_window_6-8               0.9186      0.232      3.960      0.000       0.464       1.373
hour_window_9-11             -3.5892      0.278    -12.911      0.000      -4.134      -3.044
==============================================================================
Omnibus:                      120.667   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              244.560
Skew:                           0.078   Prob(JB):                     7.84e-54
Kurtosis:                       3.966   Cond. No.                     3.52e+15
==============================================================================
```

- On the left, we are seeing regression with numerical features post VIF analysis, and we see that 'Snowfall' is statistically insignificant with '**p-value**' of **0.691** (>0.05).
  In the categorical features, 'Month Demand Level_3' is statistically insignificant with '**p-value**' of **0.956** (>0.05).
- On the right, we are seeing regression excluding the statistically insignificant features where the remaining features are statistically significant.

LINEAR REGRESSION:

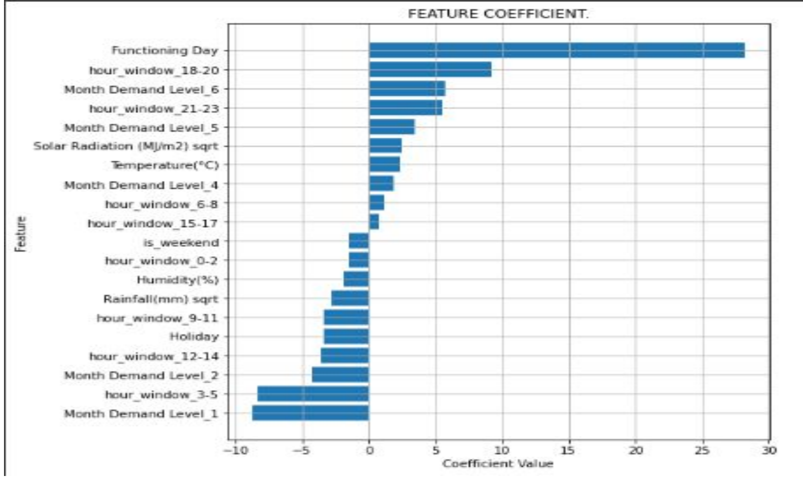
```
**** RESULTS for LINEAR REGRESSION:****

TRAINING SET:
---------------
R2 train (Linear Regression): 0.7221
Adj. R2 train (Linear Regression): 0.7212
MSE train (Linear Regression): 115628.30971727562
RMSE train (Linear Regression): 340.0416293886318
----------------------------------------------------------------------------

TESTING SET:
---------------
R2 test(Linear Regression): 0.7009
Adj. R2 test (Linear Regression): 0.6986
MSE test (Linear Regression): 124288.06762325743
RMSE test (Linear Regression): 352.54512849173994
----------------------------------------------------------------------------
```

```
Model Intecept: -2.1934135736234097
```



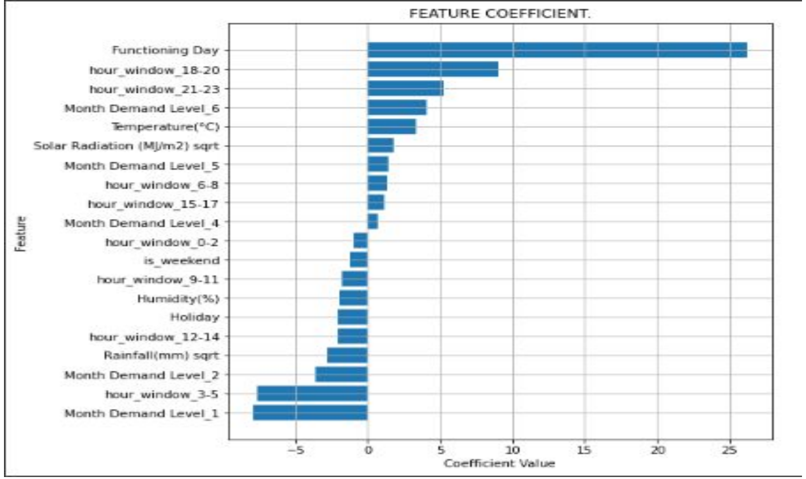
RIDGE REGRESSION:

```
**** Results for RIDGE REGRESSION from Grid Search: ****

 The best estimator across ALL searched params:
 Ridge(alpha=0.75)

 The best score across ALL searched params:
 -37.371657606982794

 The best parameters across ALL searched params:
 {'alpha': 0.75}
----------------------------------------------------------------------------

TRAINING SET:
---------------
R2 train (Ridge Regression): 0.7218
Adj. R2 train (Ridge Regression): 0.7209
MSE train (Ridge Regression): 115753.6502783094
RMSE train (Ridge Regression): 340.22588125877405
----------------------------------------------------------------------------

TESTING SET:
---------------
R2 test(Ridge Regression): 0.7007
Adj. R2 test (Ridge Regression): 0.6984
MSE test (Ridge Regression): 124395.49733987435
RMSE test (Ridge Regression): 352.69745865241833
----------------------------------------------------------------------------

Model Intecept: -2.0747526541485506
```



LASSO REGRESSION:

```
**** Results for LASSO REGRESSION from Grid Search: ****

 The best estimator across ALL searched params:
 Lasso(alpha=0.05)

 The best score across ALL searched params:
 -38.07928733919868

 The best parameters across ALL searched params:
 {'alpha': 0.05}
----------------------------------------------------------------------------

TRAINING SET:
---------------
R2 train (Lasso Regression): 0.7013
Adj. R2 train (Lasso Regression): 0.7003
MSE train (Lasso Regression): 124290.83750325158
RMSE train (Lasso Regression): 352.54905687471575
----------------------------------------------------------------------------

TESTING SET:
---------------
R2 test(Lasso Regression): 0.6812
Adj. R2 test (Lasso Regression): 0.6788
MSE test (Lasso Regression): 132488.4446757722
RMSE test (Lasso Regression): 363.9896216594261
----------------------------------------------------------------------------

Model Intecept: -0.3675980027846464
```

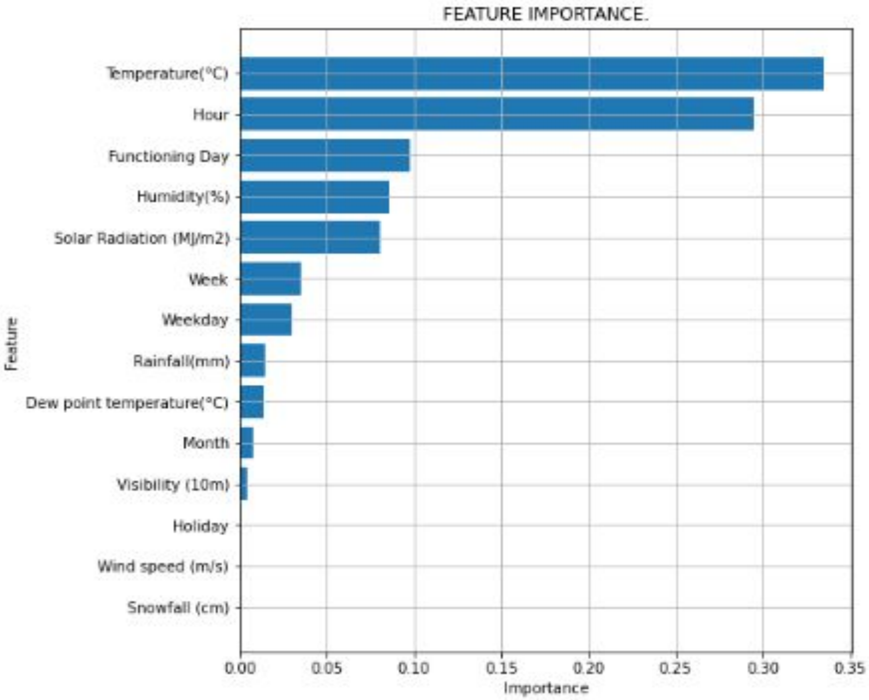


DECISION TREE BASED MODELS:

Since Decision Trees are non parametric models, original dataset with some extracted ‘time’ features was used, ie features like ‘Hour’ & ‘Month’ were not converted into categorical features like ‘Hour Window’ & ‘Month Demand level’.

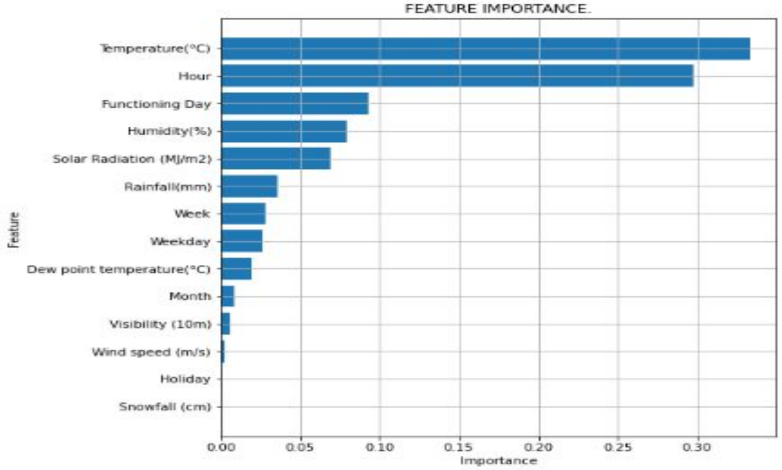
DECISION TREE REGRESSION:

```
**** Results for DECISION TREE REGRESSION from Grid Search: ****

 The best estimator across ALL searched params:
 DecisionTreeRegressor(max_depth=12, min_samples_leaf=8, min_samples_split=25,
                       random_state=0)

 The best score across ALL searched params:
 0.854215166884479

 The best parameters across ALL searched params:
 {'max_depth': 12, 'min_samples_leaf': 8, 'min_samples_split': 25}
----------------------------------------------------------------------------

TRAINING SET:
---------------
R2 train (Decision Tree Regression): 0.8746
Adj. R2 train (Decision Tree Regression): 0.8743
MSE train (Decision Tree Regression): 34115.0326931135
RMSE train (Decision Tree Regression): 184.70255193990553
----------------------------------------------------------------------------

TESTING SET:
---------------
R2 test(Decision Tree Regression): 0.8056
Adj. R2 test (Decision Tree Regression): 0.8046
MSE test (Decision Tree Regression): 56127.71693183033
RMSE test (Decision Tree Regression): 236.91288891031303
----------------------------------------------------------------------------
```



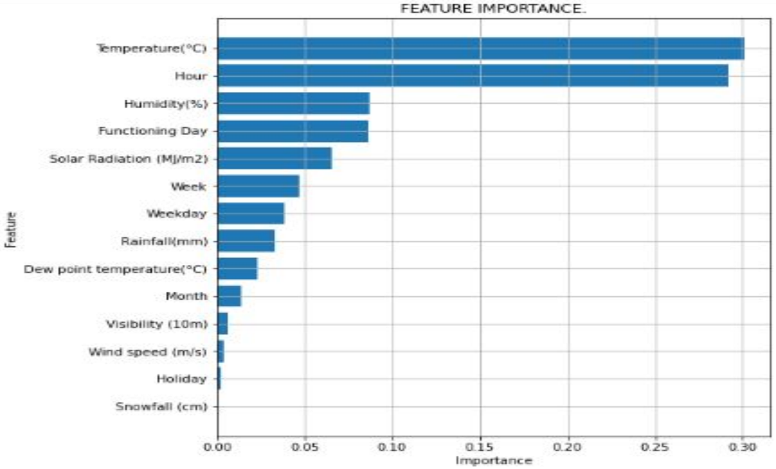
RANDOM FOREST REGRESSION:

```
**** Results for RANDOM FOREST REGRESSOR from Grid Search: ****

 The best estimator across ALL searched params:
 RandomForestRegressor(max_depth=10, max_features=12, min_samples_leaf=5,
                       min_samples_split=10, n_estimators=103)

 The best score across ALL searched params:
 0.8954406197897271

 The best parameters across ALL searched params:
 {'max_depth': 10, 'max_features': 12, 'min_samples_leaf': 5, 'min_samples_split': 10, 'n_estimators': 103}
--------------------------------------------------------------------------------

TRAINING SET:
---------------
R2 train (Random Forest Regressor): 0.863
Adj. R2 train (Random Forest Regressor): 0.8627
MSE train (Random Forest Regressor): 29708.526598040557
RMSE train (Random Forest Regressor): 172.36161579087312
--------------------------------------------------------------------------------

TESTING SET:
---------------
R2 test(Random Forest Regressor): 0.8286
Adj. R2 test (Random Forest Regressor): 0.8277
MSE test (Random Forest Regressor): 42370.729997496506
RMSE test (Random Forest Regressor): 205.84151670034038
```



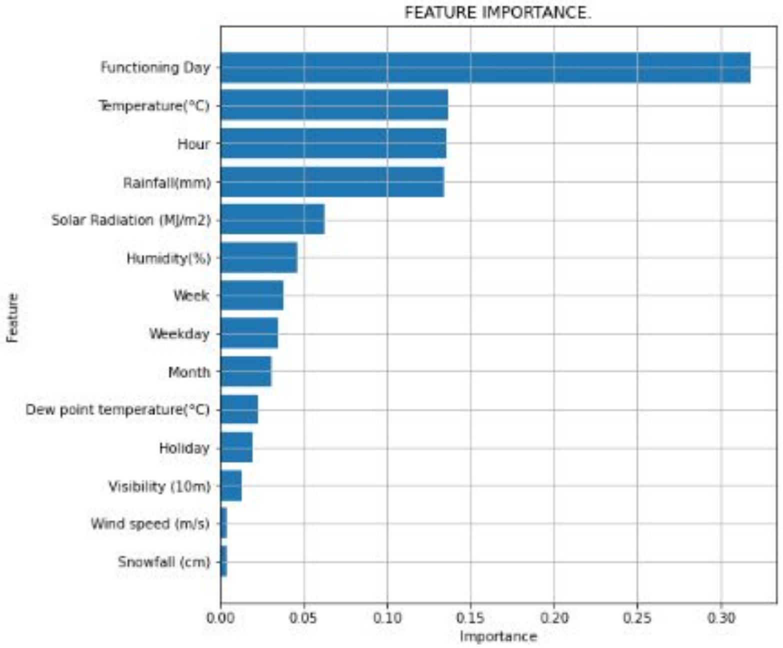
GRADIENT BOOSTING REGRESSION:

```
**** Results for GRADIENT BOOSTING REGRESSOR from Grid Search: ****

 The best estimator across ALL searched params:
 GradientBoostingRegressor(max_depth=7, max_features=10, min_samples_leaf=8,
                           min_samples_split=15, n_estimators=95,
                           random_state=0)

 The best score across ALL searched params:
 0.9373294420179166

 The best parameters across ALL searched params:
 {'max_depth': 7, 'max_features': 10, 'min_samples_leaf': 8, 'min_samples_split': 15, 'n_estimators': 95}
--------------------------------------------------------------------------------

TRAINING SET:
---------------
R2 train (Gradient Boosting Regressor): 0.961
Adj. R2 train (Gradient Boosting Regressor): 0.9609
MSE train (Gradient Boosting Regressor): 9832.11513580283
RMSE train (Gradient Boosting Regressor): 99.15702262473813
--------------------------------------------------------------------------------

TESTING SET:
---------------
R2 test(Gradient Boosting Regressor): 0.9089
Adj. R2 test (Gradient Boosting Regressor): 0.9084
MSE test (Gradient Boosting Regressor): 23254.456838905982
RMSE test (Gradient Boosting Regressor): 152.49412066996544
--------------------------------------------------------------------------------
```

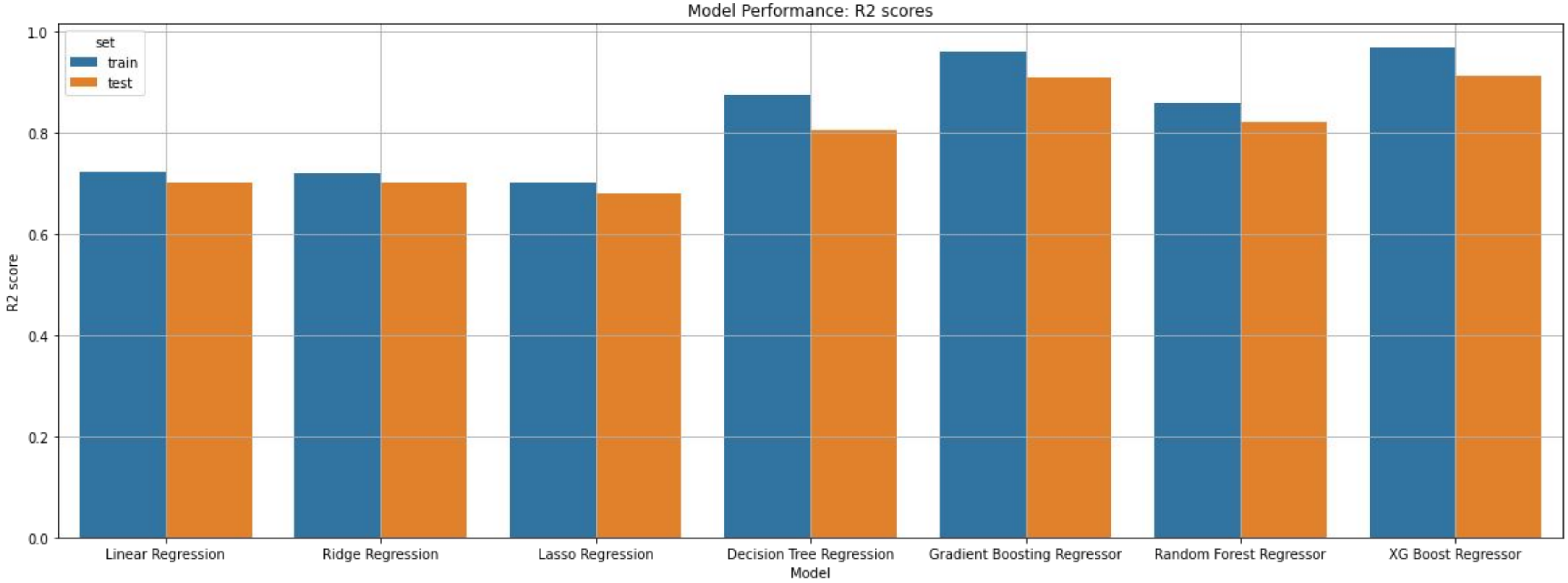


EXTREME GRADIENT BOOSTING REGRESSION:

```
**** Results for XG BOOST REGRESSOR from Grid Search: ****

 The best estimator across ALL searched params:
 XGBRegressor(colsample_bytree=0.85, max_depth=7, n_estimators=95,
             objective='reg:squarederror', reg_lambda=2.75)

 The best score across ALL searched params:
 0.9371475542019205

 The best parameters across ALL searched params:
 {'colsample_bytree': 0.85, 'max_depth': 7, 'n_estimators': 95, 'reg_lambda': 2.75}
----------------------------------------------------------------------------

TRAINING SET:
---------------
R2 train (XG Boost Regressor): 0.9678
Adj. R2 train (XG Boost Regressor): 0.9677
MSE train (XG Boost Regressor): 8672.925070480685
RMSE train (XG Boost Regressor): 93.12854057957037
----------------------------------------------------------------------------

TESTING SET:
---------------
R2 test(XG Boost Regressor): 0.9138
Adj. R2 test (XG Boost Regressor): 0.9133
MSE test (XG Boost Regressor): 23082.283588471983
RMSE test (XG Boost Regressor): 151.92854764155413
```



MODEL PERFORMANCE COMPARISON:



- Linear models (Linear, Ridge & Lasso) perform similarly, with R2 train of around 0.7 and R2 test slightly lower than that.
- Decision Tree Regression (DTR) and Random Forest Regression (RFG) perform similar and better than Linear models,but R2 test of RFG is better than DTR.
- Overall, Gradient Boosting (GBR) and Extreme Gradient Boosting (XGBR) regressor perform similar and the best, but R2 test of the XGBR (0.913) is better than GBR (0.908) and XGBR takes much lesser time to train.

CONCLUSIONS:

- The bike demand across months is non-monotonic, with lowest demand in winters, increasing till June in summers and then increases and decreases till November in autumn season.
- The pattern of bike demand is different for weekdays and weekends.
- On weekdays, there is a huge surge/ spike during morning 8 am and evening 6 pm indicating those being office going and returning hours.
- On weekends, there is no peak in the morning as weekdays, indicating that people have their weekends off at work and the maximum demand is in the evening, but lesser than normal weekdays.
- The relation of hour of the day and rented bike count is non-monotonic.
- The demand of bikes on non-weekend holidays follow similar pattern to that on weekends.
- Temperature has the highest positive correlation of 0.54 with bike demand. Wind speed and Solar radiation are also positively correlated.
- Humidity, Rainfall and Snowfall are negatively correlated with bike demand.
- All linear models performed almost similar, with R2 score of around 0.7 for both test and train test.
- All tree based models perform better than Linear models, with Extreme Gradient Boosting Regressor giving the best results: R2 train- 0.967, R2 test- 0.913.

THANK YOU!