Regression Model to Predict Bike Sharing Demand

Manas Nayan Mukherjee

Cohart-Boston

# Abstract:- Rental Bike Sharing is the process of renting bicycles on an hourly, weekly, or membership-based basis. This phenomenon has seen its stock rise to considerable levels due to a global effort towards reducing the carbon footprint, leading to climate change, unprecedented natural disasters, ozone layer depletion, and other environmental anomalies.

**In our project, we choose to analyse a dataset pertaining to rental bike demand from the South Korean city of Seoul, comprising of climatic variables like temperature, humidity, rain, snowfall, dew point temperature, and others. An hourly rental bike count is the regress after thorough pre processing. To an extent, our linear model was able to explain the factors orchestrating the hourly demand for rental bikes.**

***Keywords: -*** *Data Mining, Linear Regression, Correlation Analysis, Bike Sharing Demand Prediction, Carbon Footprint.*

# INTRODUCTION

In the past decade, bike sharing has emerged as a sustainable, convenient, and generally affordable travel mode,and become an integral component of urban transportation systems in cities around the world. The trend is likely to continue under the ongoing COVID-19 pandemic, as biking is usually seen as a safer and healthier travel option.

Because of its positive effects on the environment, public health and trafﬁc congestion, bike sharing systems (BSS)should be promoted to play a bigger role in urban mobility.Currently, efﬁcient operations of BSS rely on the dynamic rebalancing of bikes between locations to better match the ever-changing demand patterns.

Therefore, the accurate short-term demand prediction at high spatial resolution is crucial because it is the basis to ensure the availability,efﬁciency, and user experience of bike sharing service. Recent years have seen growing interests in short-term demand prediction for intelligent BSS, with a particular focus on deep learning methods, because of their demonstrated effectiveness in extracting the complex and nonlinear knowl-edge hidden in large-scale mobility data.

In practice, BSS are often designed as feeders to public transport systems or support multi-modal transport connections. Extensive prior studies have explored the relationship between BSS usage with public transit, taxi and ride-hailing, and unrevealed signiﬁcant complementary or competitive relation-ships subject to trip purpose, trip length, availability, etc.

As a result, the demand for bike sharing will inevitably be inﬂuenced by other transportation modes, which should be considered in demand prediction. Incorporating demand information across modes can also help mitigate the data sparsity problem commonly seen in BSS, since bike sharing is rarely one of the primary travel modes in cities.

The speciﬁc contributions of this research are as follows:

1.We proposed a supervised learning model for bike sharing demand prediction based on inter-modal relationships taken from historical demand data.

2.We introduce a multi-relational graph neural network(MRGNN) to model spatial correlations between heterogeneous spatial units across different modes.

3.Extensive experiments are conducted based on real-world datasets from Seoul city, and the results demonstrate the superior performance of our proposed model compared to existing methods.

# PROPOSED MODEL

**The relationship between the dependent variable-"Rented Bike Count ‘and remaining columns (independent variable).**

For our Rental Cycle Dataset, the Pre-Processing was performed on Python library. The CSV file was loaded using read.csv () function. The missing data is checked using is info () function. Additionally, which () function was invoked to attain the index numbers of the missing values in the dataset. The output depicted that there were no missing values in our dataset.

Categorical variables- Seasons, Functioning Day, and Holiday- were converted coded into numerical depictions to fit our Linear Regression analysis. The transformed dataset is loaded as a fresh .csv file using the write.csv () function.

Other than Seasons, Holiday, and Functioning Day, Descriptive Statistics provide detailed information of numerical data in terms of Central Tendency, namely Mean, Median, and Mode.

Data dispersion is also explained via Standard Deviation. Also, the extreme values are represented by Maximum, Minimum, and Range.

For Categorical variables mentioned above, Central Tendency and dispersion become irrelevant. Hence, the description is done with the help of a matrix which shows percentage share of each category in a specific Category. Also, Cumulative percentage shows the validates the part of each category type.

Using Scatterplot, all the set of independent variables are plotted against the dependent variable, Rented Bike Count.

The plot displays the data distribution of each dependent variable with respect to the hourly Rental Bike Count.

In addition to providing initial distribution of continuous data, the scatter plot also aids in identifying any noisy data or outliers which could be removed to gain an optimized linear model.

All the plots are made on Python Library using regplot() function.

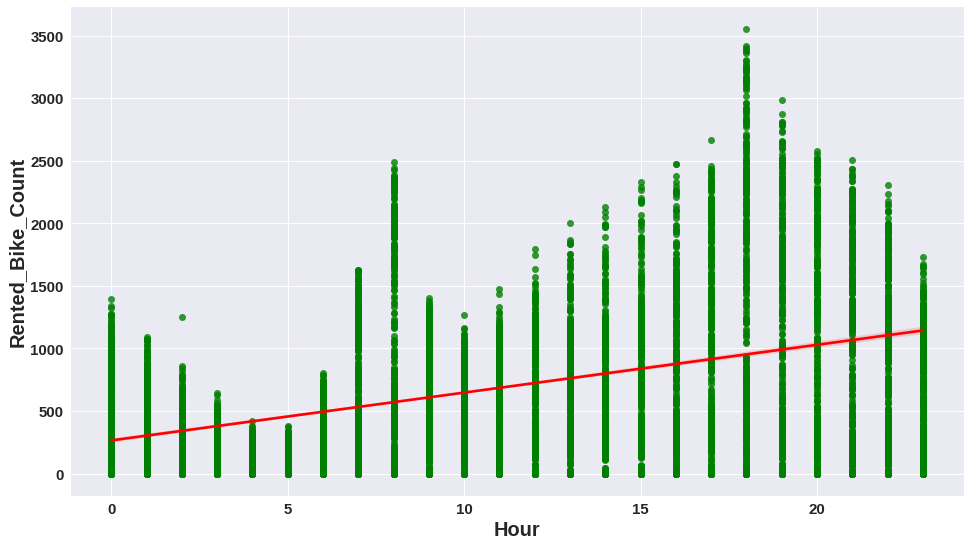


Figure 1. Scatter Plot between Rental Bike Count and Hour of the Day

From the above Scatter Chart, we can observe that data points are closely placed to each other, thereby forming dark linear patterns on the graph.

However, the only secluded point appears to be where the Rented Bike Count exceeds 3500. We have to closely monitor that point over its potential to be an outlier.

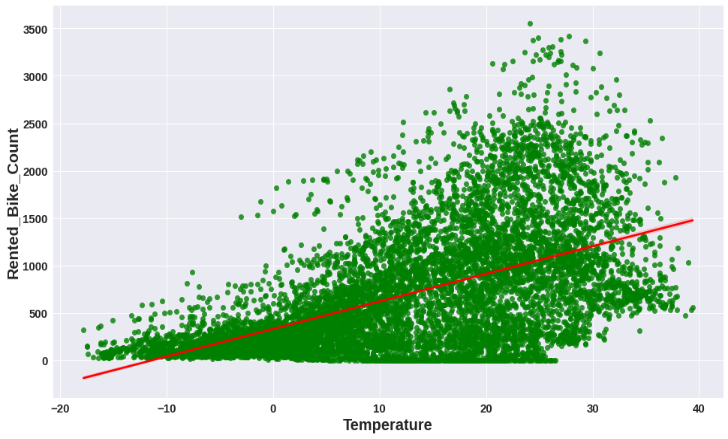


Figure 2. Scatter Plot between Rental Bike Count and Temperature

From the above distribution, Rental Bike Count is spread in form of a cloud which is dense around the region of -20 to 40C. The small tailing clusters towards the higher end of X axis shows that almost all the data points will affect our regression model.

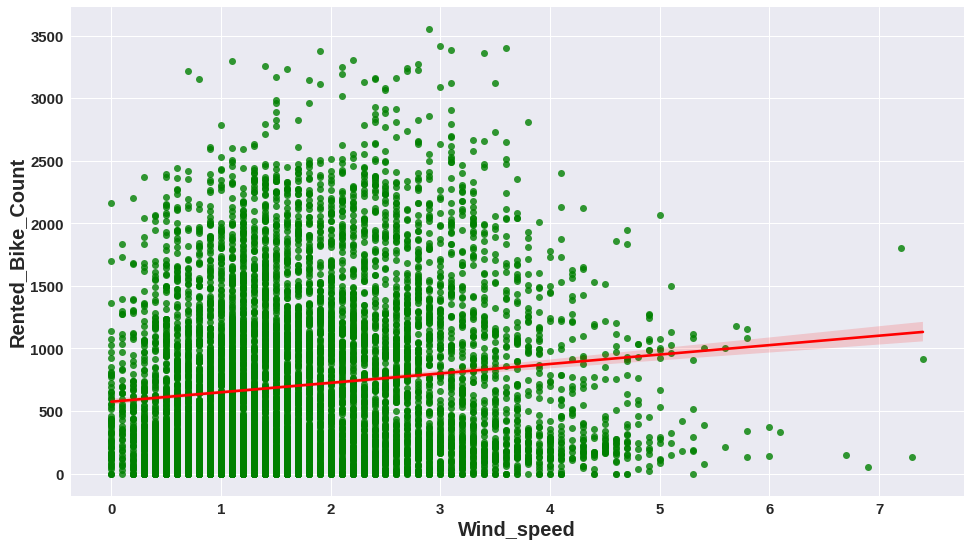


Figure 3. Scatter Plot between Rental Bike Count and Wind Speed

As identified above, in this distribution too, the data points form a prominent cloud around the Wind Speed lying between 0-5 m/s. However, the distribution starts to fade into secluded clusters till breeze of 6m/s.

Post that speed, the data seems to be isolated without any affect from the available data bunch. These points would be considered as potential outliers for our linear regression model.

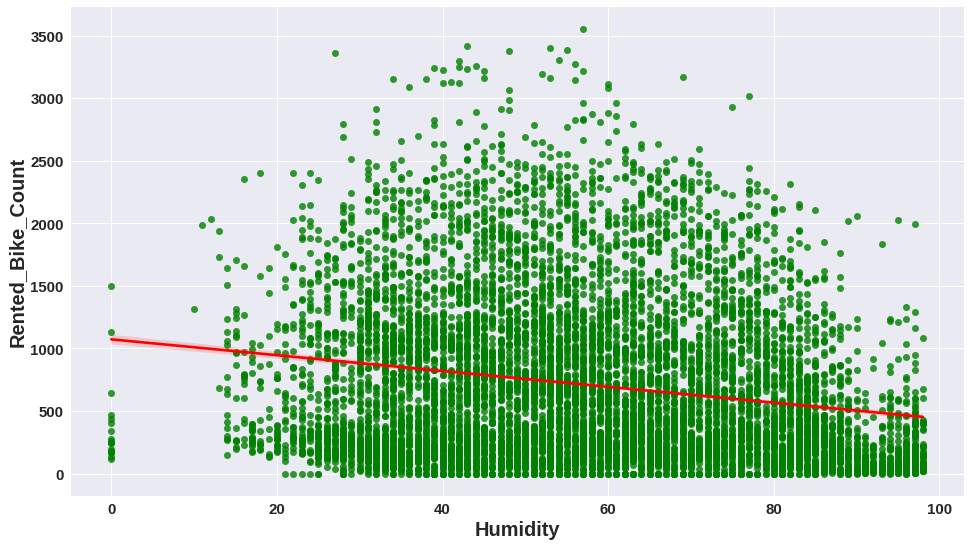


Figure 4. Scatter Plot between Rental Bike Count and Humidity

From the above Scatter Plot, it is evident that data points form a cloud for Humidity ranging between 20 to

100. Also, it understood that Humidity could not be equal to zero, realistically, meaning a possible data discrepancy.

Additionally, the single data point of Rental Bike Count above 3500 which would be taken into consideration during outlier analysis.

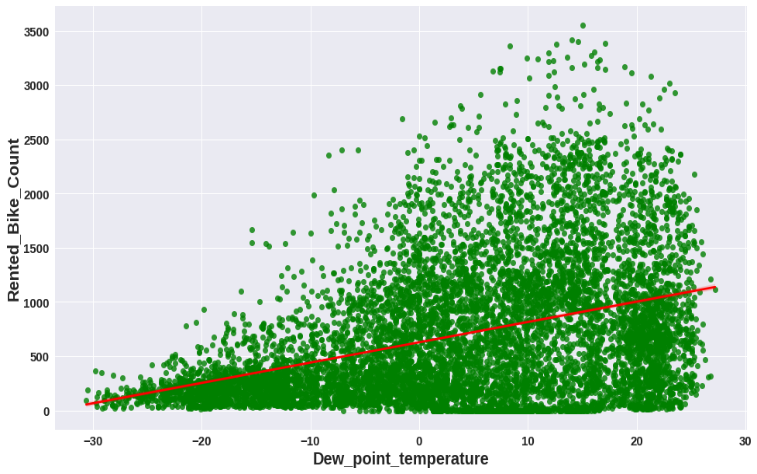


Figure 5. Scatter Plot between Rental Bike Count and Dew Point Temperature

From the above Scatter Plot, the formed data point cloud depicts that Dew Point Temperatures did not make any significant impact until Rental Bike Count reached 500.

After that, there is lightening of the data shade which suggests that Dew Point Temperature made visible effect as number of Rental Bike Counts increased till 3500, where a single data point, again, appears to be secluded.

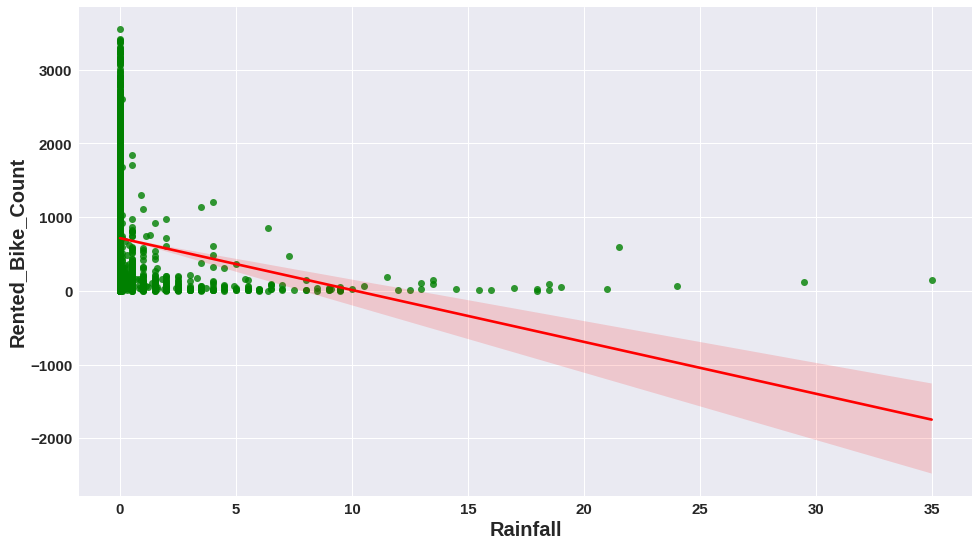


Figure 6. Scatter Plot between Rental Bike Count and Rainfall

As Cycle is an open way to commute between places, Rainfall, almost, will have an inverse relation with the Rental count. The Scatter Plot above also suggests that a significant number of counts lies along the dates when Rainfall was equal to 0 mm.

The scattered clusters above 10 mm to 35 mm show a minimal rise of Rental Bikes from zero. The single data point at 35mm remains an exception and suggests orders for recreational purposes or any other relevant cause. Hence, points above 20 should be considered during outlier analysis.

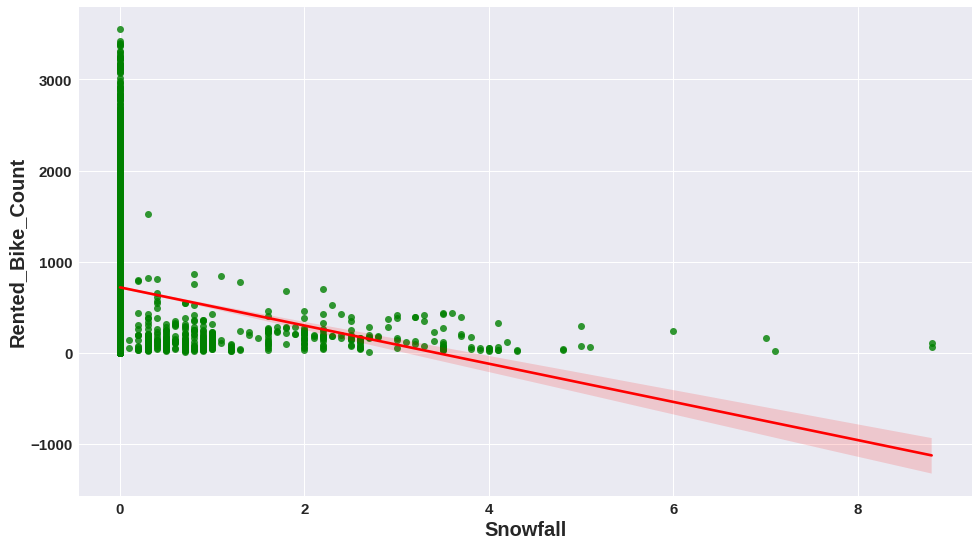


Figure 7. Scatter Plot between Rental Bike Count and Snowfall

As observed in the case of Rainfall, 0 cm Snowfall dominated the Rental Count distribution and clusters lying till

4 cm. Similar to Rainfall, a few data points above 6 cm suggest Rental cycles for recreation or any other relevant cause.

Above scatterplots present a clear picture about how outliers affect the regression model for our dataset. On monitoring data points on multiple visualizations, R was used to trim our dataset initially containing 8700 entries, ending up on 8567 observations.

1. *Correlation Analysis*

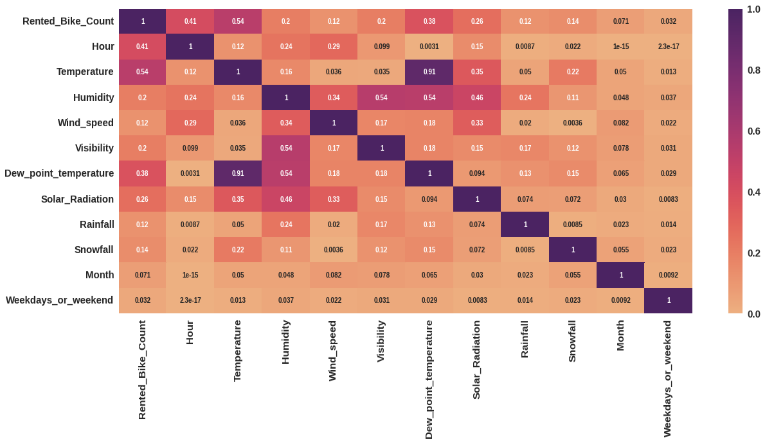


Figure 8. Correlational Analysis of the dataset variables from the above Correlation graph, we can observe that

Temperature and Dew Point Temperature are highly correlated, thereby one of the variables would have to be removed from our Regression model, depending on the significance of each variable.

# RESULTS AND DISCUSSIONS

***Linear Regression:***

Before the outlier treatment we obtained a Regression Model on the dataset containing 8760 observations. The parameters saw a slight improvement after Outlier Treatment and Correlation analysis.

To refine our model, we cleaned a few outliers to obtain an efficient Regression line. Rental Orders above 2500 were removed from our dataset, owing to the scattered distribution leading to noisy data.

Similarly, Rainfall, Snowfall, Solar Radiation, and Wind Speed entries exceeding 10mm, 4cm, 3.5 MJ/m2, and 5m/s respectively were removed from our dataset, too. Our final dataset comprises of 8567 observations.

* The Mean Absolute Error (MAE) is 5.83.
* The Mean Squared Error (MSE) is 58.624.
* The Root Mean Squared Error (RMSE) is 7.656.
* The R2 Score is 0.618.

On testing data R2 score is 0.6183 which is almost close to training data R2 score. Hence, we can say that our model performance is good, and overfitting is not observed. We need to improve our model performance.

***Lasso Regression*:**

* The Mean Absolute Error (MAE) is 5.86.
* The Mean Squared Error (MSE) is 60.464.
* The Root Mean Squared Error (RMSE) is 7.77.
* The R2 Score is 0.61129.

On testing data R2 score is 0. 61129 which is almost close to training data R2 score. Hence, we can say that our model performance is good, and overfitting is not observed. We need to improve our model performance.

***Ridge Regression:***

* The best alpha value is {'alpha': 0.1}
* The best negative mean squared error value is 60.73

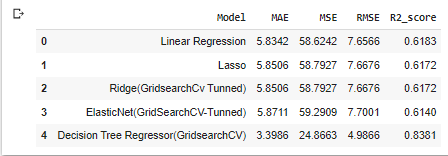
Best params are applied on the above ridge model. (Score on ridge is after hyperparameter tuning.)

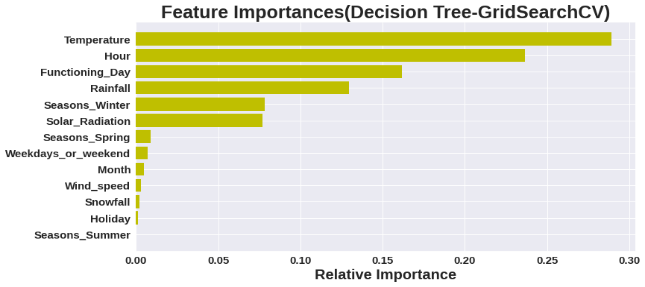
***ElasticNet Regression:***

* The Mean Absolute Error (MAE) is 5.871.
* The Mean Squared Error (MSE) is 59.29.
* The Root Mean Squared Error (RMSE) is 7.70.
* The R2 Score is 0.6139.
* The best alpha value is {'alpha': 0.0001, 'l1\_ratio': 0.5}
* The best negative mean squared error value is -60.7320

Its range is 0 < = l1\_ratio < = 1. If l1\_ratio = 1, the penalty would be L1 penalty. If l1\_ratio = 0, the penalty would be an L2 penalty. \*If the value of l1 ratio is between 0 and 1, the penalty would be the combination of L1 and L2.

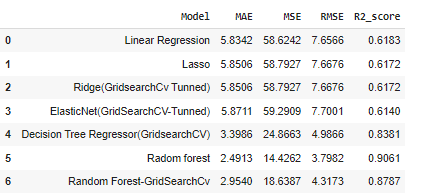
***Decision Tree Regression:***

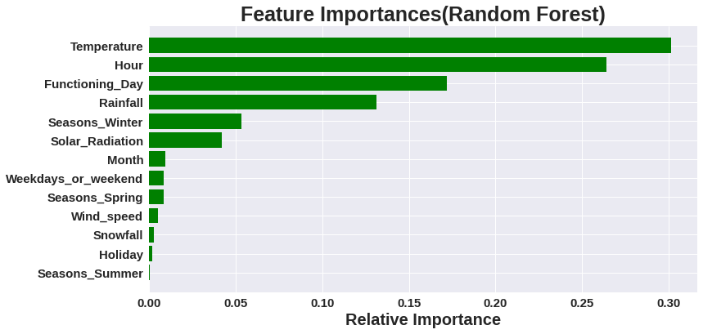




We have applied this best parameter to above Decision tree regressor model.

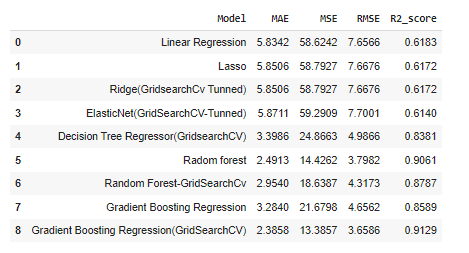
***Random Forest Regression:***

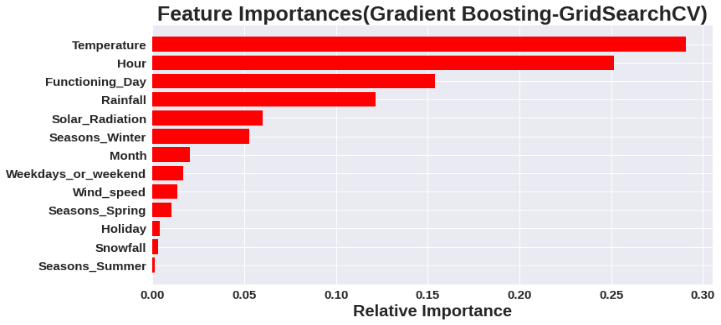
******



By looking at the graph we can say that temperature and hours plays very important role on bike rentals.

***Gradient Boosting Regression:***

****



1. **CHALLENGES**

The core challenges of the bike sharing prediction problem are the modeling of the complex spatial and temporal dependencies.

Firstly, the complex spatial dependencies, the bike sharing usage data distribution is highly imbalanced with varying demands and supplies in different locations of a city. For example, those famous spots are usually accompanied with a much higher demand for various transportation modes including shared bikes than other locations.

Secondly, for the complex temporal dependencies, the bike sharing usage demand may burst in morning evening peak hours because of commuters. But such patterns would not appear in weekends or holidays. Other external factors also make the bike sharing usage patterns complex, e.g., social events or weather.

Thirdly, the complex spatial and temporal dependencies are vulnerable to the changes of bike share systems or nearby environments, for example, the addition of a new bike station or a new shopping mall.

Dealing with such big dataset is quite difficult sometimes , finding missing values made things some more complicated, defining a function which is used to annotate the histogram percent according to their respective count taken a big notch of this obstacle part. Coming to the visualization part , more or less makes our challenges addresses to code in such a way to visualize the graphs as per rows and columns.

# CONCLUSION

**In this project, we trained a model to predict the number of bike rentals at any hour of the year given the weather conditions. The data set was obtained from the Seoul city. which contained the historical bike usage pattern with weather data.**

Firstly, we have done exploratory data analysis on the data set. We looked for missing data values (none were found) and outliers and appropriately modify them. We also perform correlation analysis to extract out the important and relevant feature set and later perform feature engineering to modify few existing columns and drop out irrelevant ones. We also found that:

1. working or non-working day: we see rental patterns same be it peak office hours 7am to 10am) between peak hours 4 pm to 7 pm.
2. Hour of the day: Bike rental count is mostly correlated with the time of the day. As indicated above, the count reaches a high point during peak hours on a working day and is mostly uniform during the day on a non-working day.
3. Temperature: People generally prefer to bike at moderate to high temperatures.
4. Season: We see the highest number bike rentals in summer (March to June) seasons and the lowest in winter (january to feb) seasons.

Secondly, as this is the regression problem so tried to predict the continuous value. We have used 6 regression models which are linear, lasso, ridge, decision tree, random forest and gradient boost and further we found that:

1. the relation between the dependent and independent variables should be almost linear,
2. mean of residuals should be zero or close to 0 as much as possible. It is done to check whether our line is actually the line of 'best fit',
3. there should be homoscedasticity or equal variance in regression model. This assumption means that the variance around the regression line is the same for all values of the predictor variable (X),
4. there should not be multicollinearity in regression model. Multicollinearity generally occurs when there are high correlations between two or more independent variables,
5. before and after applying these model we checked our regression assumptions by distribution of residuals, scatter plot of actual and predicted values, removing multi-colinearity among independent variables.