Capstone 1 Report: Seoul Bike Sharing Demand Prediction

Tzu-Ying Chen

# Problem

Seoul, South Korea is one of the world largest metropolitan area, with around 24 million people living in the city. The transportation within the city has become a public issue to be solved. Rental bikes are then introduced to Seoul for the eco-friendly solution for the commuters and the enhancement of mobility comfort. In order to gain revenue and achieve promised service level, the bike rental companies need to ensure the rental bikes are available and accessible while maximizing the usage and minimizing the idle time.

# Goal and Utility

The goal of this project is to analyze the bike rental activities at different time on different days and in different weather, and predict the appropriate bike count required at each hour for the stable supply of the rental bikes. The analysis and the prediction can ensure a stable supply of the rental bikes for the publics and help the bike sharing companies schedule regular maintenance for the bikes.

# Data

The data set for this project contains 8,760 observations including weather information, such as temperature, humidity, and windspeed, the number of bikes rented per hour, datetime information, such as holidays and date.

The data set is available on UCI Machine Learning Repository. The page contains information about the data set and its source and relevant scientific publications:

[https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand#](https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand)

# Approach

It is a supervised regression task. The count of the rental bikes will be our target variable, while the datetime, weather information, and holidays will be used as the predictive variables. Linear regression, ridge regression, or Random Forest regression are some of the machine learning models that could be used in this project. Data exploratory analysis will provide deeper understandings of the data and address the problem more clearly, and cross-validation could ensure the outcomes of the model is reliable.

# Deliverables

The deliverables for the project will include the source code, the data set, and a paper addressing the purpose, approach, findings and results of the project.

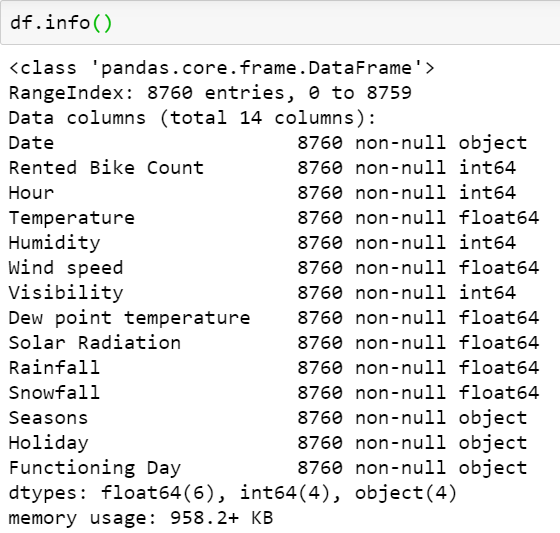
# Data Collection

The data for this capstone project contains public bikes rental information in Seoul Bike Sharing System and is acquired from UCI Machine Learning Repository. Link to the data source is below:

[https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand#](https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand)

The data set is in csv format, which could be directly read into python console using python pandas library.

# Data Cleansing and Data Wrangling

The only thing that I need to do when reading dataset was to set the correct encoding when parsing the data using pd.read\_csv. The data set comes fairly clean, with a total of 14 columns and 8,760 data points without any null values in the data set. There were not outliers, and do not need to drop and/or fill any missing values.

# Data Story

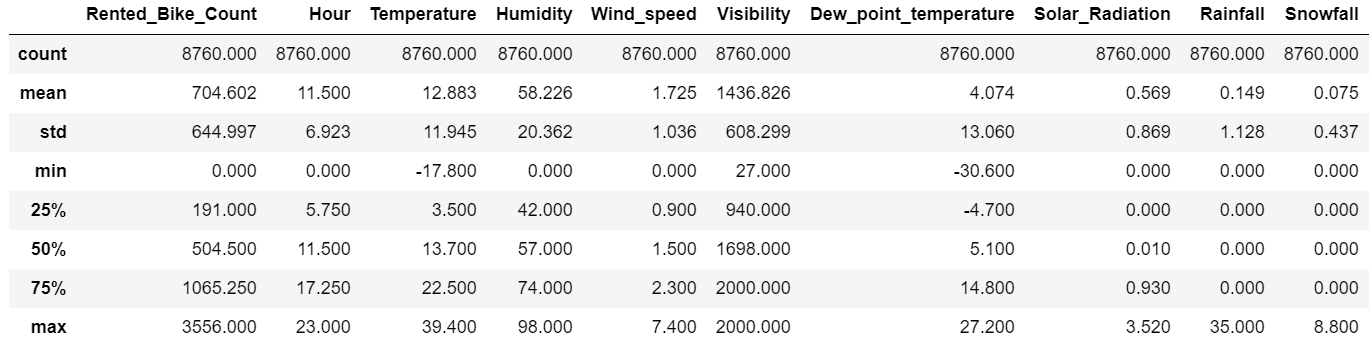
## Data Type

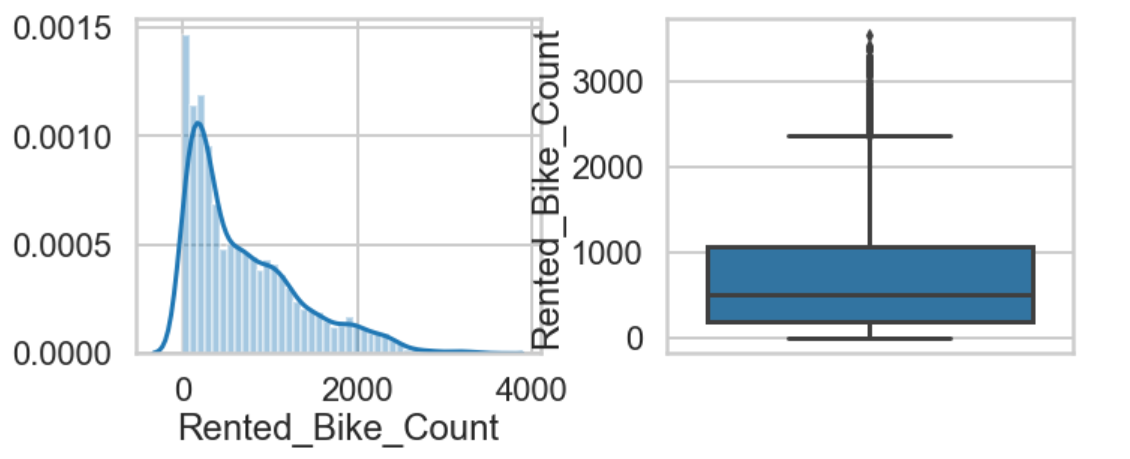
The goal of the project is to predict rental bike count for each hour, hence the target variable for this project is rented\_bike\_count. The predictor variables include temperature, humidity, wind speed, dew point temperature, solar radiation, rainfall, snowfall, holiday, functioning day, seasons, date, and hour. The following chart show the overview of the data set in terms of the data type and category.

## Exploratory Data Analysis

### Univariate Analysis

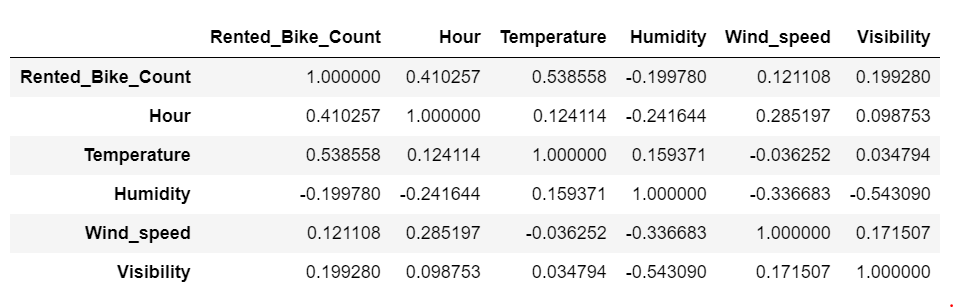
When conducting univariate analysis, variables are analyzed one at a time. For the continuous variable, we can observe its distribution and identify the outliers. In the project, there are 10 continuous variables in the data set.



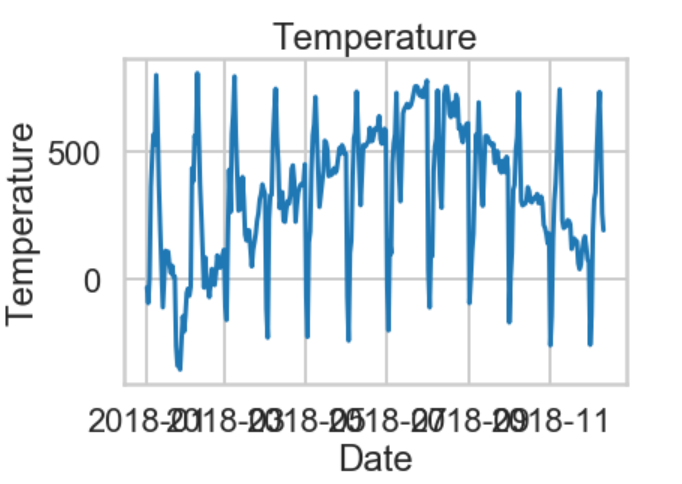
In the above data overview, we can summarize that there are average of 705 bikes are rented on an hourly basis, but the data has a fairly high standard deviation, indicating that there is a high variation in bike rentals. In addition, the distribution of bike rental amount shows left skewed, and the box chart providing the idea of there are some extreme cases.

### Bivariate Analysis

When conducting bivariate analysis, we are examining the relationship of two variables. When looking at the correlation between two continuous variables, there is no strong relationship between any of the two.

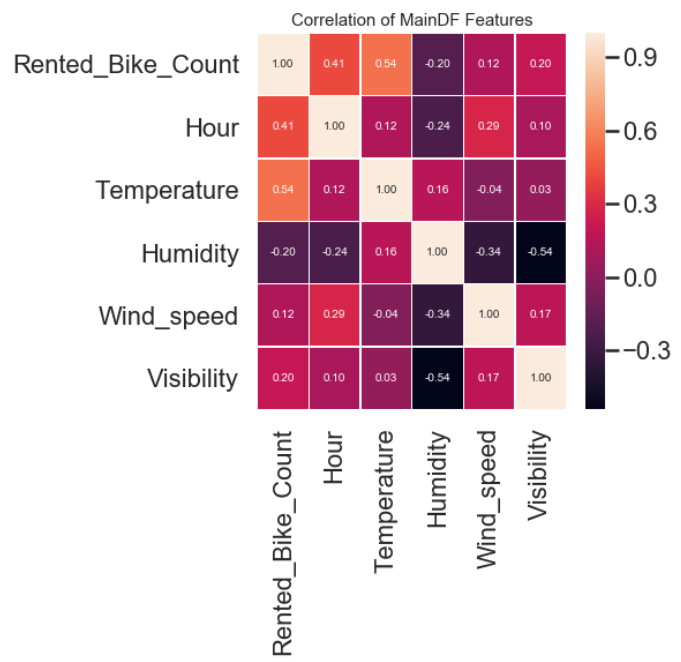


However, when plotting the time series data on temperature, we can see a trend. One noticeable thing is, I have to exclude 2017 data due as the data set doesn’t include full year of data. In addition, the chart below is aggregated into date level instead of the hourly level.



# Inferential Statistics Analysis

In the data set, it is showing a relatively strong correlation between rented bike count, the target variable, and hour and temperature, the two predictive variables. Hence, the two features will be used when training the model.

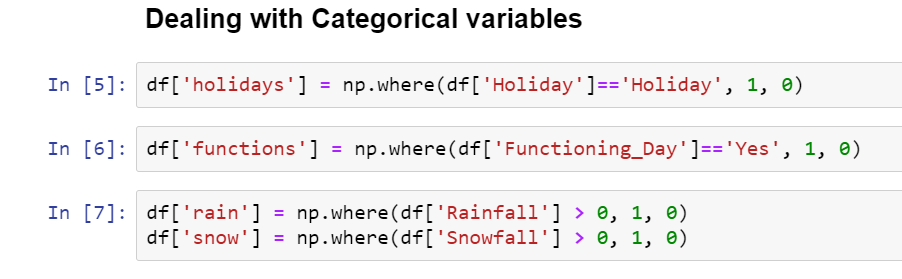


I am also interested in seeing whether the rental activities on holiday is similar to the non-holiday days. Hence, the Null hypothesis is the average of rental bike count on the holiday and the rental bike count on the non-holiday is the same. The alternative hypothesis then is the average of the bike counts are different. The t-value returns -7.59, indicating we reject the null hypothesis—the rental activities are quite different between holiday season and non-holiday season.

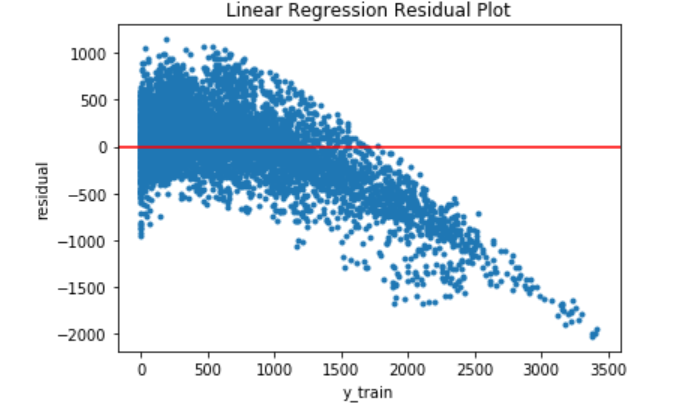
In addition to the holiday impact, I am also interested in the season impact to the rental bike activities, especially on summer and winter. The t-value returns 10.26, indicating we accept the null hypothesis—there is no significant difference between the rental activity in summer and in winter.

# In Depth Analysis

The data set consists of the count of hourly rental bike and its corresponding weather and datetime information. Hence, the first model that was used to give it a shot is linear regression model. When using linear regression model, it is necessary to transform the categorical variables into a set of true/false (binary) dummy variables. In addition to the “Holiday” and “functioning\_day” features, “Rainfall” and “snowfall” features are also transformed to binary variables for modeling.

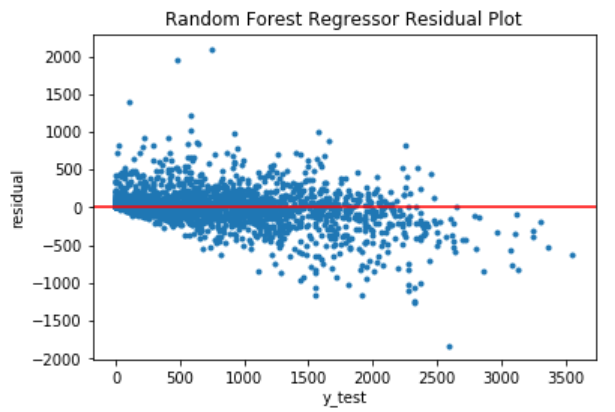


Linear Regression

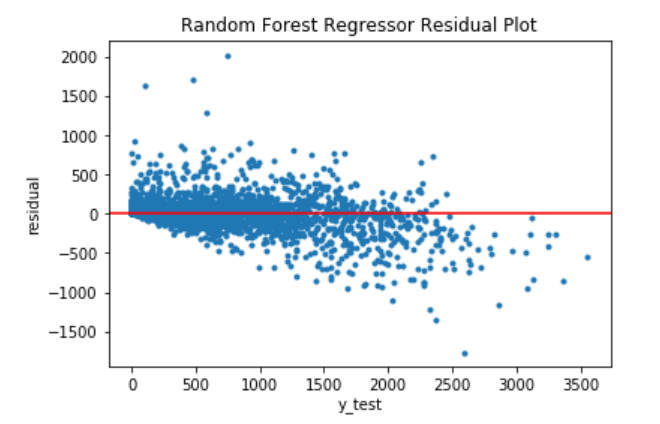
Rented bike count is the used as the target variable, while all other columns become the features to the model. The training result for linear regression is not that optimal. The r-squared is showing 0.549, with some of the features, such as Visibility and the log of visibility, not showing significance in the model. Then, I took out those features with less t-values and re-train the model. The R-squared did not improve. The graph below is the residual plot on the training dataset. The mean absolute error is 321.93. In conclusion, linear regression model does not perform well for this problem.

Random Forest Model

The random forest regressor model, on the other hand, performs very well on predicting the rented bike count. With the default hyperparameters, the model is already scoring very high, with 0.97 on the training dataset, and 0.867 on the testing dataset. The mean absolute error is 139, comparing to the one from linear regression, the random forest is doing so much better. The graph below is the residual plot of the random forest model. The error is range from -2000 to 2000 in this case.



In order to make the model perform better, the hyperparameters tuning is necessary. GridSearch cross validation is used in the tuning. The new parameters are improving the mean absolute error from 139 to 134.



# Summary and Next Step

Random forest regressor model is a good option for the bike rental prediction problem. The model can better help the rental bike company manage their rental bike supply and demand and achieve greater profits. The company can also work on maintenance of the bike by avoiding high demand hours.

In addition, there are some future improvements that can be done in the project, such as better hyperparameters tuning to get better prediction result, trying different machine learning model, such as SVM and gradient boost model.