Applied Data Analytics II – Course Project
Date completed: 04/10/2020
Research question: Bike Sharing Demand
Document name: Proposal

1.1. Project Background

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able to rent a bike from a one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world, such as Citi-bike and so on.

The data generated by these systems makes them attractive for researchers because the duration of travel, departure location, arrival location, and time elapsed is explicitly recorded. Bike sharing systems therefore function as a sensor network, which can be used for studying mobility in a city. In this competition, participants are asked to combine historical usage patterns with weather data in order to forecast bike rental demand in the Capital Bikeshare program in Washington, D.C.

1.2. Dataset Description

The dataset can be publicly obtained from Kaggle (Links: https://www.kaggle.com/c/bike-sharing-demand/overview/description), which contains bike hourly rental figures spanning two years (2011 and 2012), with variables such as season, holiday, working day, weather included in the dataset. The train dataset, which is composed of the first 19 days of each month in both 2011 and 2012, contains more than 10,800 rows of data, while the test dataset, consisted of the rest days from the twentieth day to the end day of the month, contains approximately 6,500 rows of data. The data fields are shown as follows.

Attribute	Descriptions
datatime	Hourly late + timestamp
season	1 = spring; 2 = summer; 3 = fall; 4 = winter
holiday	Whether the day is considered a holiday
workingday	Whether the day is neither a weekend nor holiday
weather	1: Clear, Few clouds, Partly cloudy, Partly cloudy 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
temp	Temperature in Celsius
atemp	“feel like” temperature in Celsius
humidity	Relative humidity
windspeed	Wind speed
casual	Number of non-registered user rentals initialed

registered	Number of registered user rentals initialed
count	Number of total rentals

Table 1 – Data fields explanation of the dataset

1.3. Research Purpose

I have known variables including data time, season, holiday, working day, weather, temperature, humidity, wind speed and so on, I will predict the total count of bikes rented during each hour covered by the test set. Based on the prediction, the company can determine the number of bikes they ought to put into the market in a single day in order to maximize the profit and minimize the cost.

1.4. Algorithm

The following algorithms will be adapted to predict the total count of bikes rented.

a. Linear Regression

Linear Regression fits a linear model with coefficients $w = (w_1, ..., w_p)$ to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.

b. Decision Tree Regressor

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

c. Random Forest Regressor

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

d. Gradient Boosting Regressor

Gradient Boosted Decision Trees (GBDT) is a generalization of boosting to arbitrary differentiable loss functions. GBDT is an accurate and effective off-the-shelf procedure that can be used for both regression and classification problems in a variety of areas including Web search ranking and ecology.

e. K Neighbours Regressor

Neighbors-based regression can be used in cases where the data labels are continuous rather than discrete variables. The label assigned to a query point is computed based on the mean of the labels of its nearest neighbors.

f. Bagging Regressor

A Bagging regressor is an ensemble meta-estimator that fits base regressors each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. Such a meta-estimator can typically be used as a way to reduce the variance of a black-box estimator (e.g., a decision tree), by introducing randomization into its construction procedure and then making an ensemble out of it.

1.5. Evaluation

I intend to Root Mean Squared Logarithmic Error (RMSLE) to evaluate the prediction result. Root Mean Square Logarithmic is the ratio (the log) between the actual values in the data and predicted values in the model. I select RMSLE instead of RMSE because in this case, under-prediction, which is likely to result in being lack of putting enough bikes into the market, is worse than an over-prediction. RMSLE can be calculated as the following formula.

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(x_i+1) - \log(y_i+1))^2}$$

2. Initial Exploration

I have initially explored the dataset by performing the following steps.

a. Importing the libraries and dataset, combining the training dataset and testing dataset and observed the data types.

```
                  datetime  season  holiday  workingday  weather   temp  \
0      2011-01-01 00:00:00       1        0           0        1   9.84
1      2011-01-01 01:00:00       1        0           0        1   9.02
2      2011-01-01 02:00:00       1        0           0        1   9.02
3      2011-01-01 03:00:00       1        0           0        1   9.84
4      2011-01-01 04:00:00       1        0           0        1   9.84
...                    ...     ...      ...         ...      ...    ...
10881  2012-12-19 19:00:00       4        0           1        1  15.58
10882  2012-12-19 20:00:00       4        0           1        1  14.76
10883  2012-12-19 21:00:00       4        0           1        1  13.94
10884  2012-12-19 22:00:00       4        0           1        1  13.94
10885  2012-12-19 23:00:00       4        0           1        1  13.12

        atemp  humidity  windspeed  casual  registered  count
0      14.395        81     0.0000       3          13     16
1      13.635        80     0.0000       8          32     40
2      13.635        80     0.0000       5          27     32
3      14.395        75     0.0000       3          10     13
4      14.395        75     0.0000       0           1      1
...       ...       ...        ...     ...         ...    ...
10881  19.695        50    26.0027       7         329    336
10882  17.425        57    15.0013      10         231    241
10883  15.910        61    15.0013       4         164    168
10884  17.425        61     6.0032      12         117    129
10885  16.665        66     8.9981       4          84     88

[10886 rows x 12 columns]                  datetime  season  holiday  workingday  weather   temp  \
```

Figure 1

```
0     2011-01-20 00:00:00       1        0           1        1  10.66
1     2011-01-20 01:00:00       1        0           1        1  10.66
2     2011-01-20 02:00:00       1        0           1        1  10.66
3     2011-01-20 03:00:00       1        0           1        1  10.66
4     2011-01-20 04:00:00       1        0           1        1  10.66
...                   ...     ...      ...         ...      ...    ...
6488  2012-12-31 19:00:00       1        0           1        2  10.66
6489  2012-12-31 20:00:00       1        0           1        2  10.66
6490  2012-12-31 21:00:00       1        0           1        1  10.66
6491  2012-12-31 22:00:00       1        0           1        1  10.66
6492  2012-12-31 23:00:00       1        0           1        1  10.66

       atemp  humidity  windspeed
0     11.365        56    26.0027
1     13.635        56     0.0000
2     13.635        56     0.0000
3     12.880        56    11.0014
4     12.880        56    11.0014
...      ...       ...        ...
6488  12.880        60    11.0014
6489  12.880        60    11.0014
6490  12.880        60    11.0014
6491  13.635        56     8.9981
6492  13.635        65     8.9981

[6493 rows x 9 columns]
```

Figure 2

	atemp	casual	count	datetime	holiday	humidity	registered	season	temp	weather	windspeed	workingday
0	14.395	3.0	16.0	2011-01-01 00:00:00	0	81	13.0	1	9.84	1	0.0	0
1	13.635	8.0	40.0	2011-01-01 01:00:00	0	80	32.0	1	9.02	1	0.0	0
2	13.635	5.0	32.0	2011-01-01 02:00:00	0	80	27.0	1	9.02	1	0.0	0
3	14.395	3.0	13.0	2011-01-01 03:00:00	0	75	10.0	1	9.84	1	0.0	0
4	14.395	0.0	1.0	2011-01-01 04:00:00	0	75	1.0	1	9.84	1	0.0	0

Figure 3

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17379 entries, 0 to 17378
Data columns (total 12 columns):
atemp         17379 non-null float64
casual        10886 non-null float64
count         10886 non-null float64
datetime      17379 non-null object
holiday       17379 non-null int64
humidity      17379 non-null int64
registered    10886 non-null float64
season        17379 non-null int64
temp          17379 non-null float64
weather       17379 non-null int64
windspeed     17379 non-null float64
workingday    17379 non-null int64
dtypes: float64(6), int64(5), object(1)
memory usage: 1.5+ MB
```

Figure 4

	atemp	casual	count	holiday	humidity	registered	season	temp	weather	windspeed
count	17379.000000	10886.000000	10886.000000	17379.000000	17379.000000	10886.000000	17379.000000	17379.000000	17379.000000	17379.000000
mean	23.788755	36.021955	191.574132	0.028770	62.722884	155.552177	2.501640	20.376474	1.425283	12.736540
std	8.592511	49.960477	181.144454	0.167165	19.292983	151.039033	1.106918	7.894801	0.639357	8.196795
min	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	1.000000	0.820000	1.000000	0.000000
25%	16.665000	4.000000	42.000000	0.000000	48.000000	36.000000	2.000000	13.940000	1.000000	7.001500
50%	24.240000	17.000000	145.000000	0.000000	63.000000	118.000000	3.000000	20.500000	1.000000	12.998000
75%	31.060000	49.000000	284.000000	0.000000	78.000000	222.000000	3.000000	27.060000	2.000000	16.997900
max	50.000000	367.000000	977.000000	1.000000	100.000000	886.000000	4.000000	41.000000	4.000000	56.996900

Figure 5

b. Defining datetime as date, year, month, day, weekend, hour; Calculating the correlation between each attribute and count.

	atemp	casual	count	holiday	humidity	registered	season	temp	weather	windspeed	workingday	date	year	month	day	weekend	hour
0	14.395	3.0	16.0	0	81	13.0	1	9.84	1	0.0	0	2011-01-01	2011	1	1	6	0
1	13.635	8.0	40.0	0	80	32.0	1	9.02	1	0.0	0	2011-01-01	2011	1	1	6	1
2	13.635	5.0	32.0	0	80	27.0	1	9.02	1	0.0	0	2011-01-01	2011	1	1	6	2
3	14.395	3.0	13.0	0	75	10.0	1	9.84	1	0.0	0	2011-01-01	2011	1	1	6	3
4	14.395	0.0	1.0	0	75	1.0	1	9.84	1	0.0	0	2011-01-01	2011	1	1	6	4

Figure 6

```
count          1.000000
registered     0.970948
casual         0.690414
hour           0.400601
temp           0.394454
atemp          0.389784
year           0.260403
month          0.166862
season         0.163439
windspeed      0.101369
day            0.019826
workingday     0.011594
weekend       -0.002283
holiday       -0.005393
weather       -0.128655
humidity      -0.317371
Name: count, dtype: float64
```

Figure 7

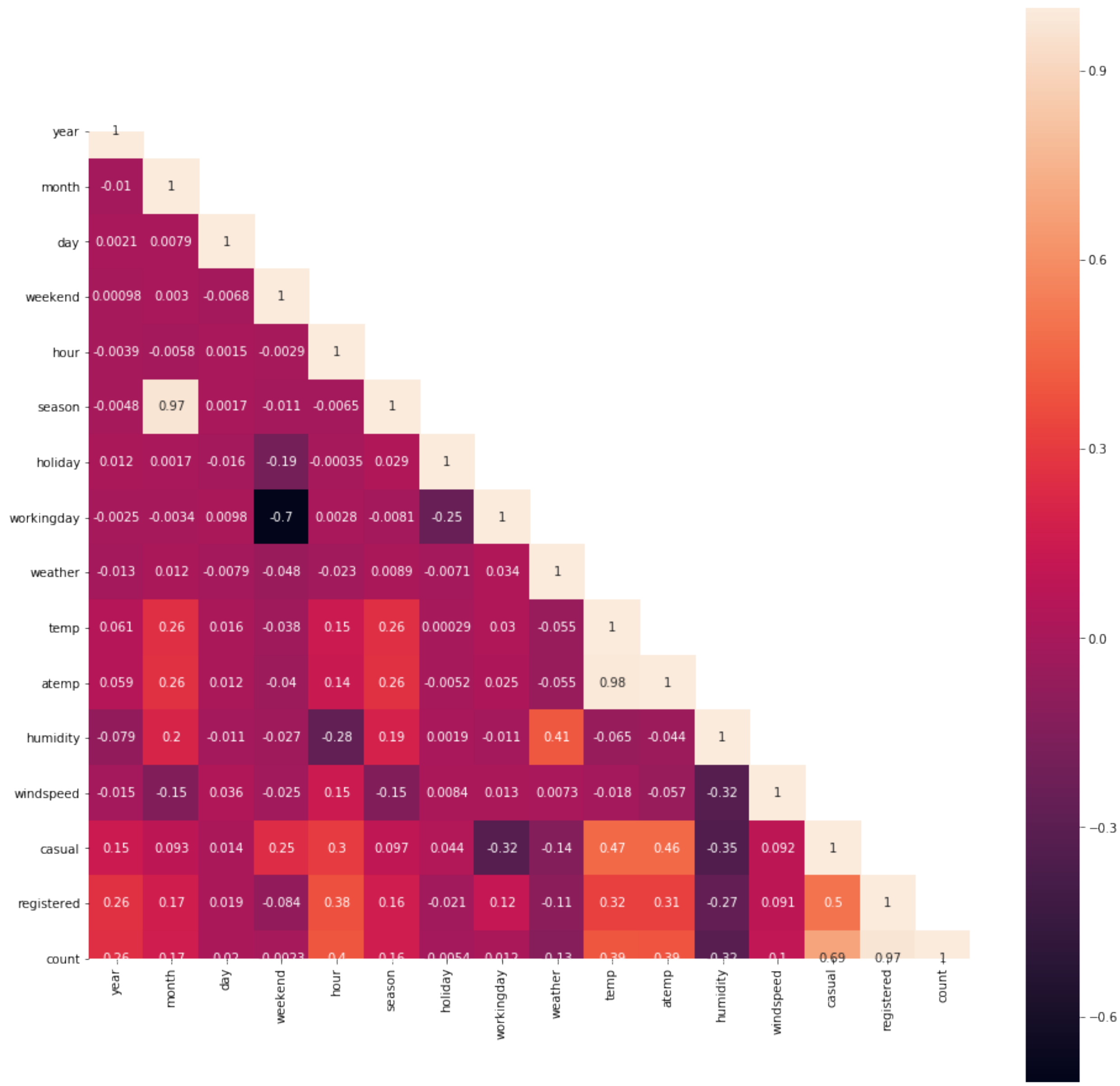

Figure 8

c. Visualizing each attribute to further analyze its impact on count.

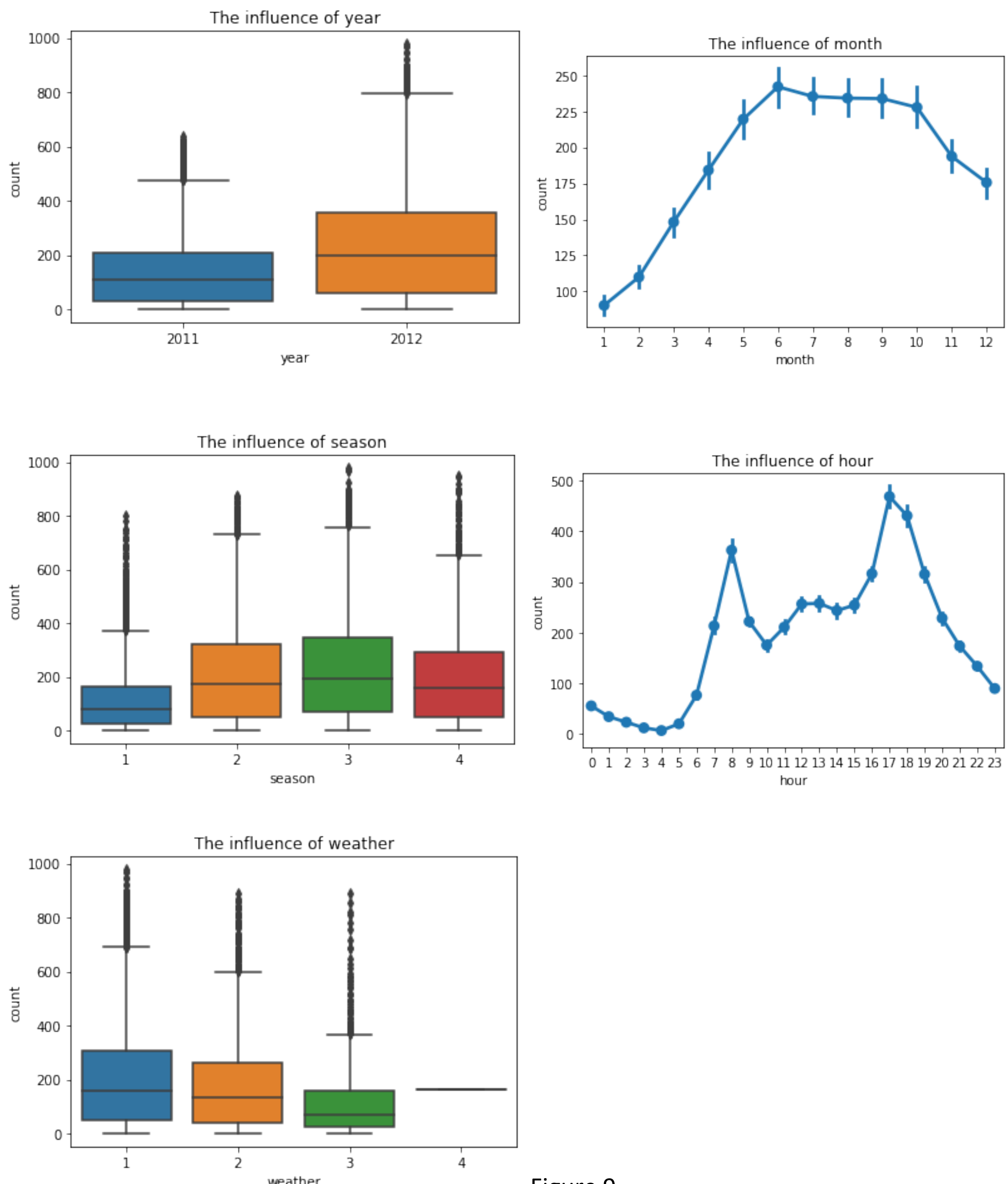

Figure 9

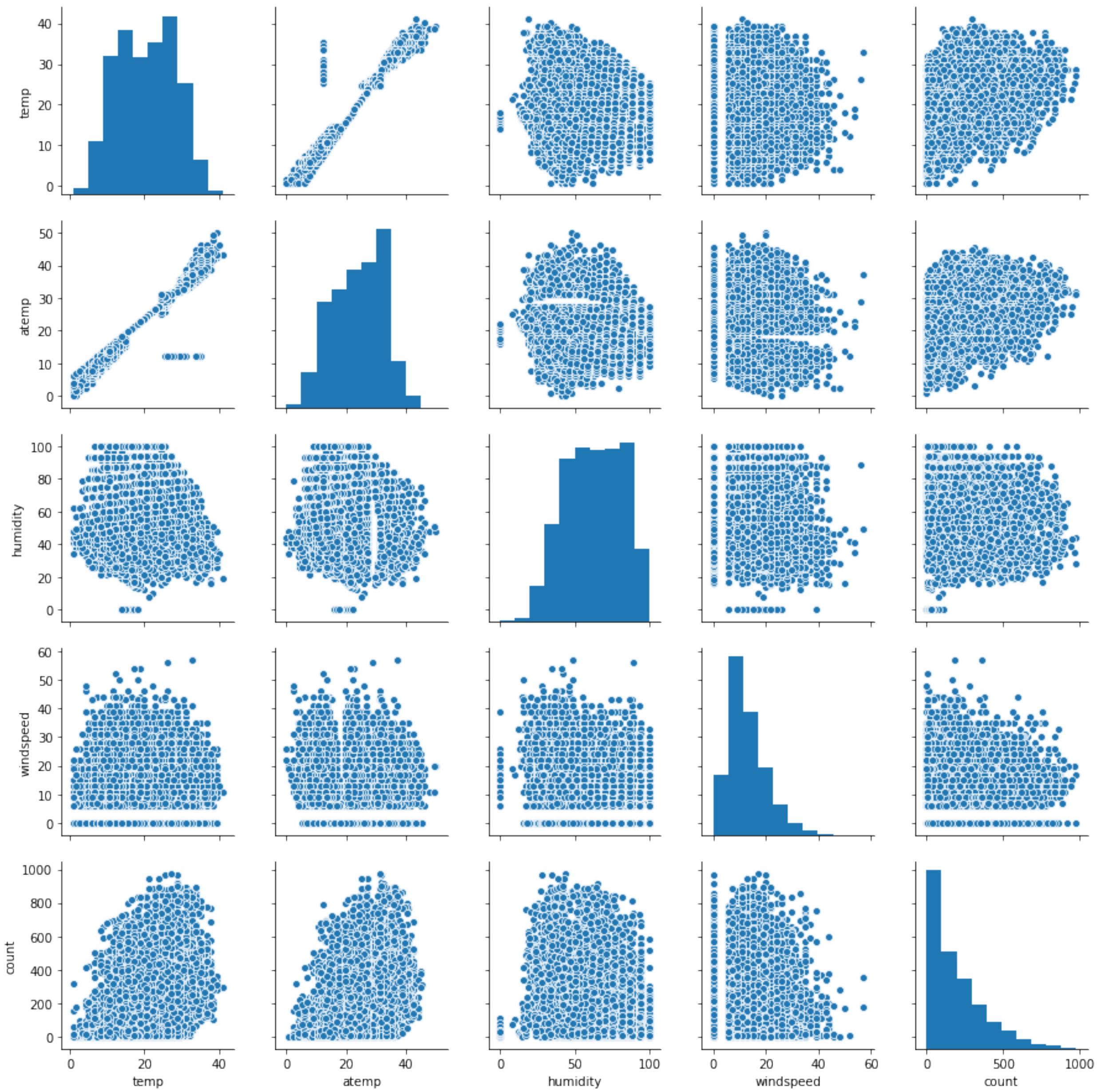

Figure 10

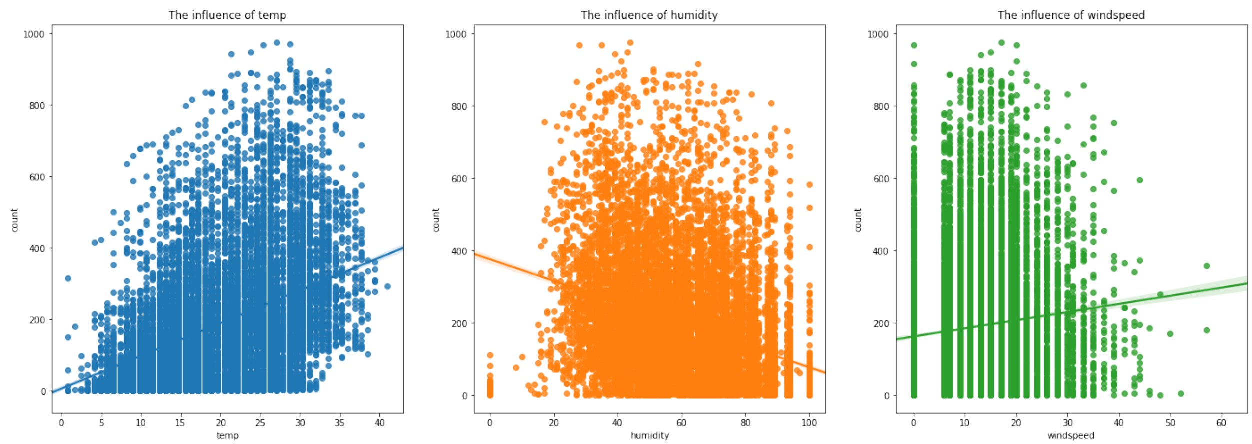

Figure 11

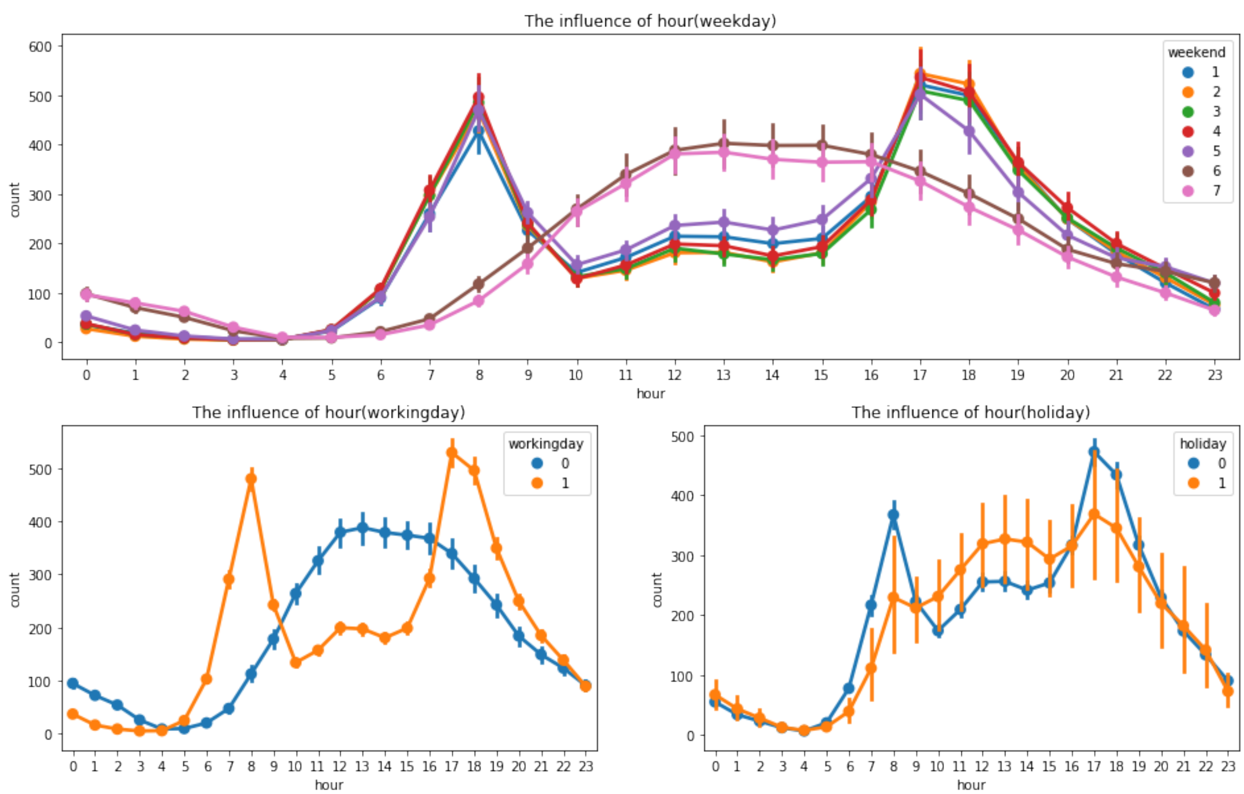

Figure 12

3. Further exploration

a. Data preparation

I have initially observed the data structure, and will check whether there exists missing data, and deal with outliers if needed in the formal project.

b. Analyzing the data and selecting appropriate attributes

I have initially calculated the correlation and explored the relationship between each predictors and count, I will further analyze the possible relationship by visualizing the data and thereby select the significant attributes.

c. Selecting and training the model; predicting the test dataset and evaluating the model.

Based on my purpose, I intend to use Linear Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, K Neighbours Regressor, Bagging Regressor to predict the count, and plan to use RMSLE to evaluate the result to determine the most suitable model.