**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](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html#sklearn.linear_model.LinearRegression) fits a linear model with coefficients w = (w1, ..., wp) 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](https://scikit-learn.org/stable/modules/tree.html#tree-classification) and [regression](https://scikit-learn.org/stable/modules/tree.html#tree-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**

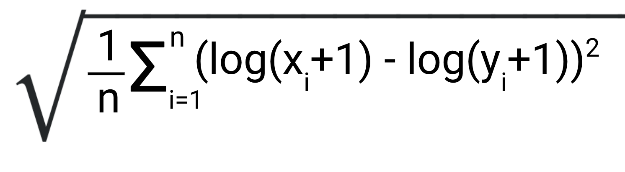
Neighbors-based regression can be used in cases where the data labels are continuous rather than discrete variables. The label assigned to a query point is computed based on the mean of the labels of its nearest neighbors.

**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.



**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.

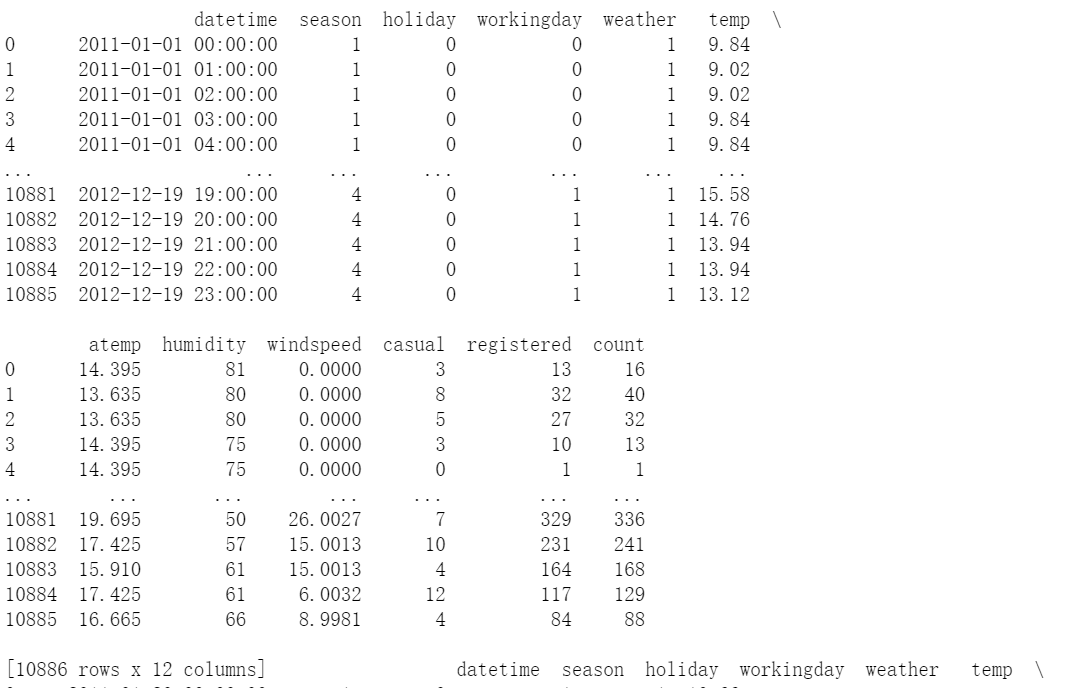


Figure 1

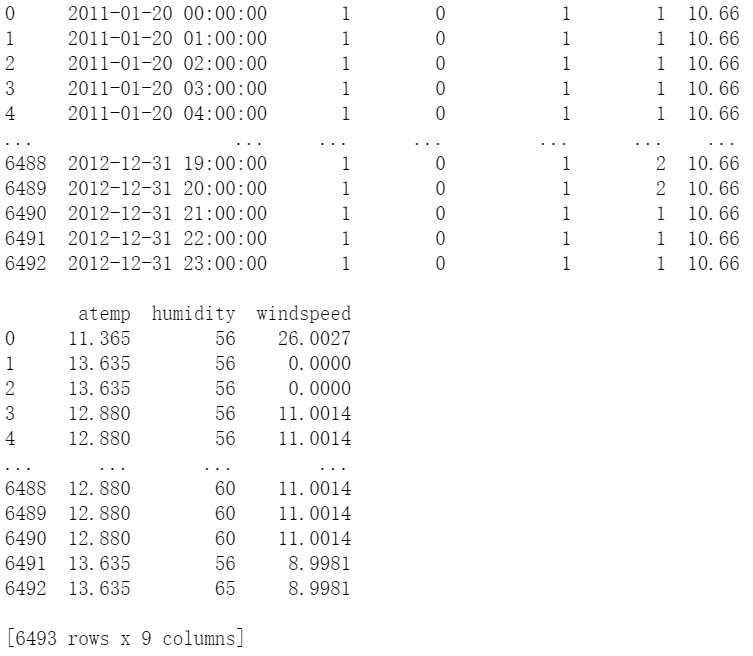


Figure 2

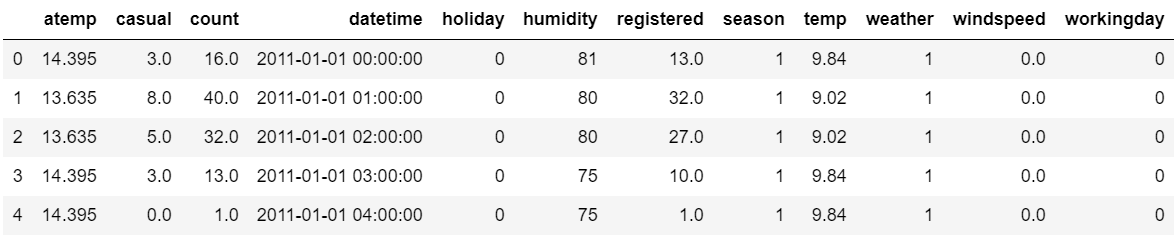


Figure 3

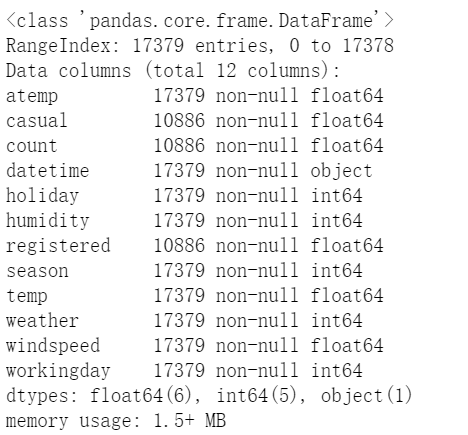


Figure 4

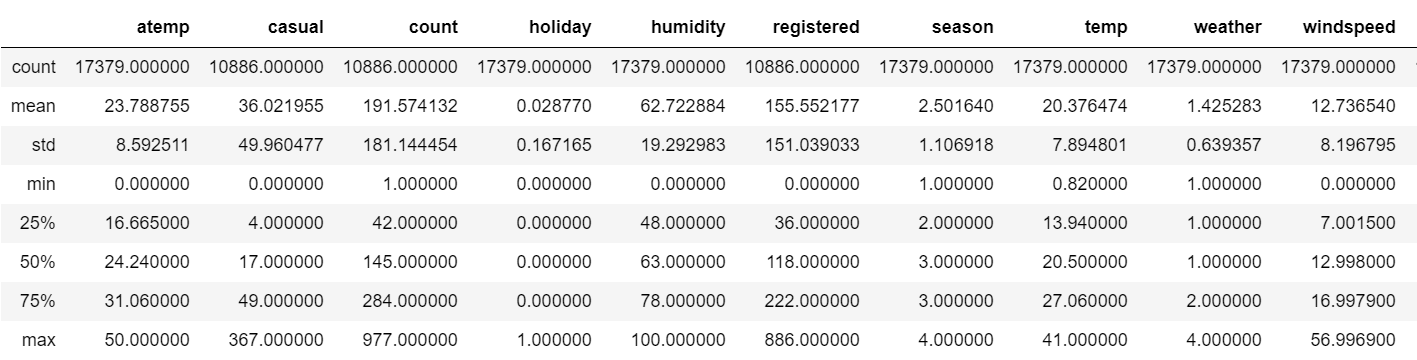


Figure 5

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

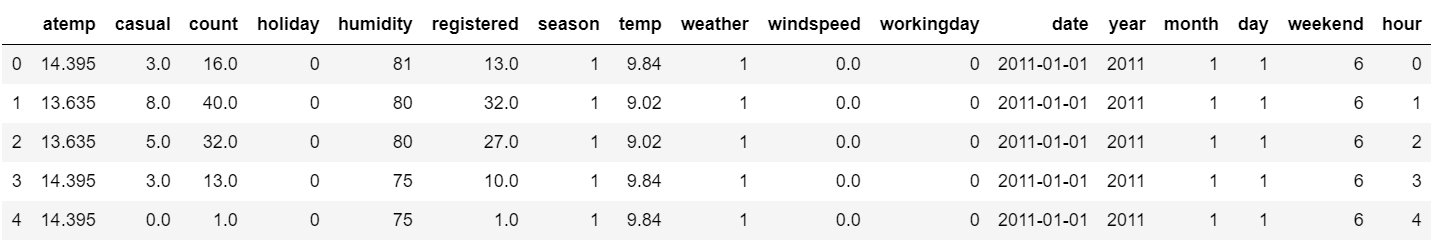


Figure 6

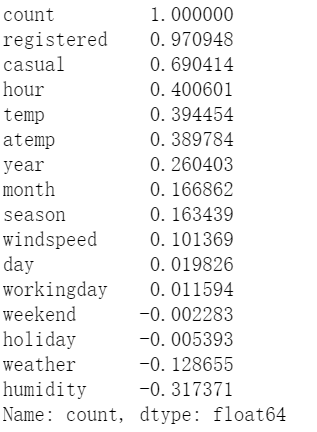


Figure 7

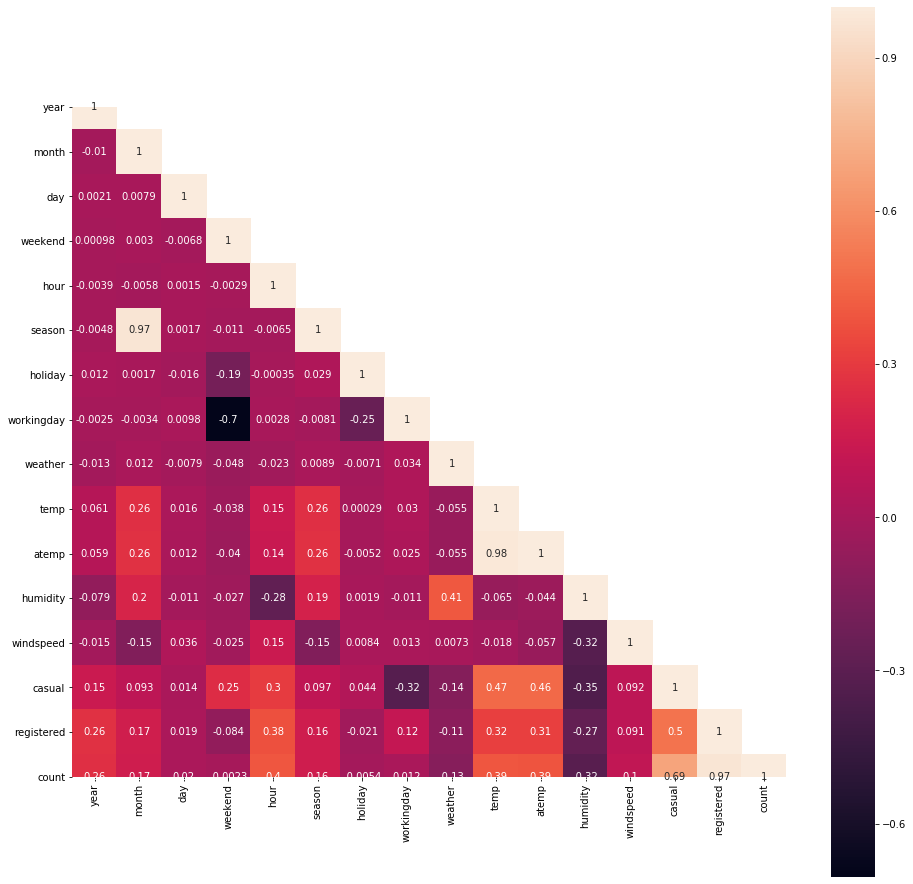
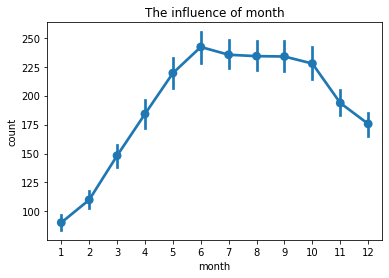
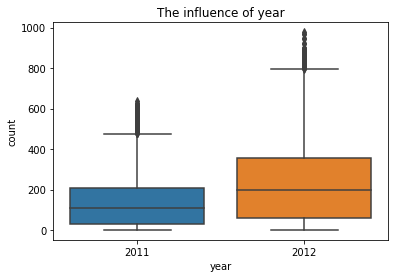
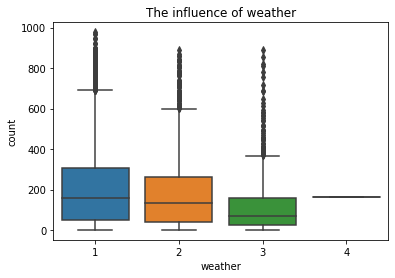
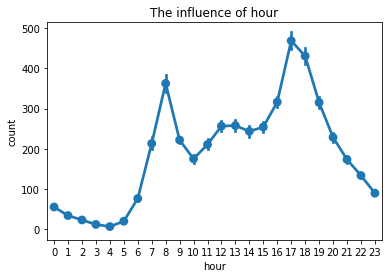
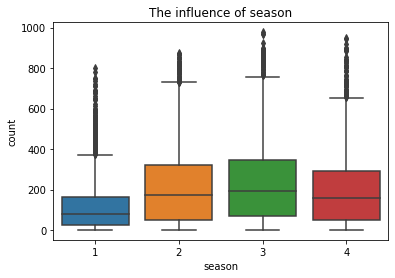


Figure 8

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



 Figure 9

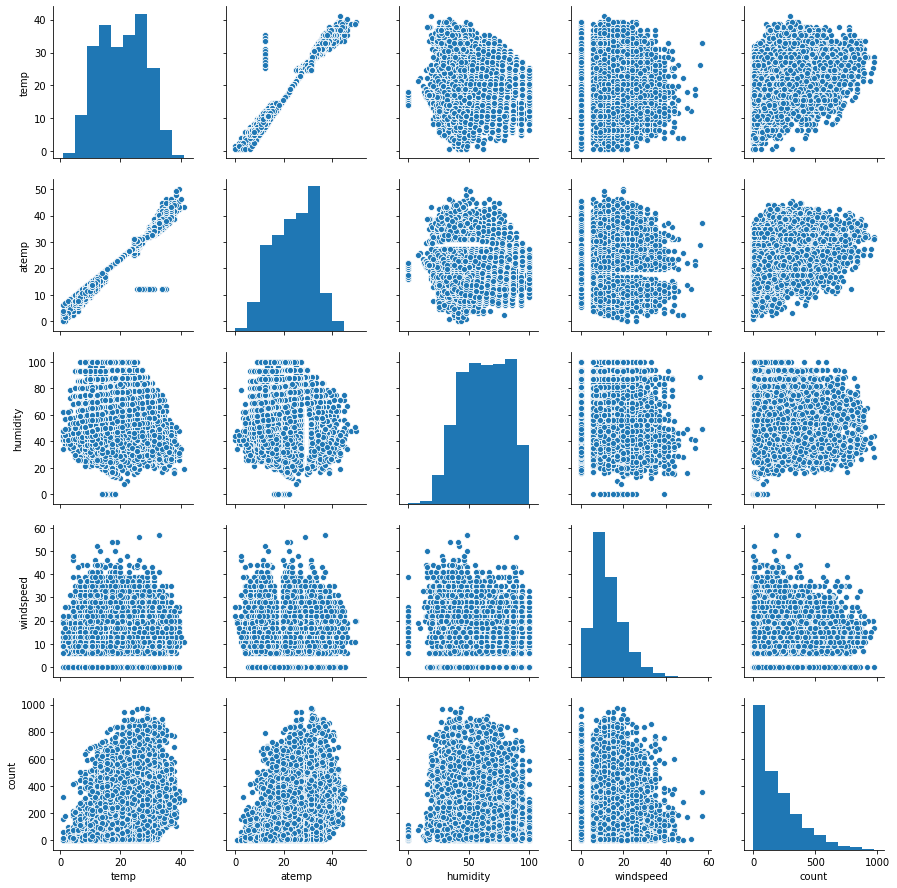


Figure 10

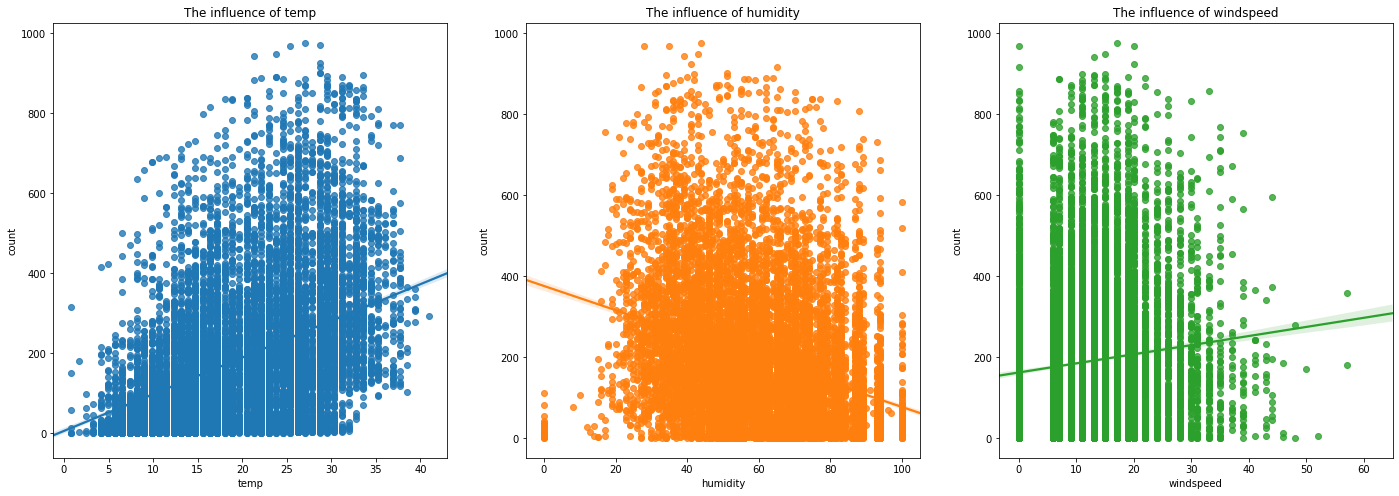


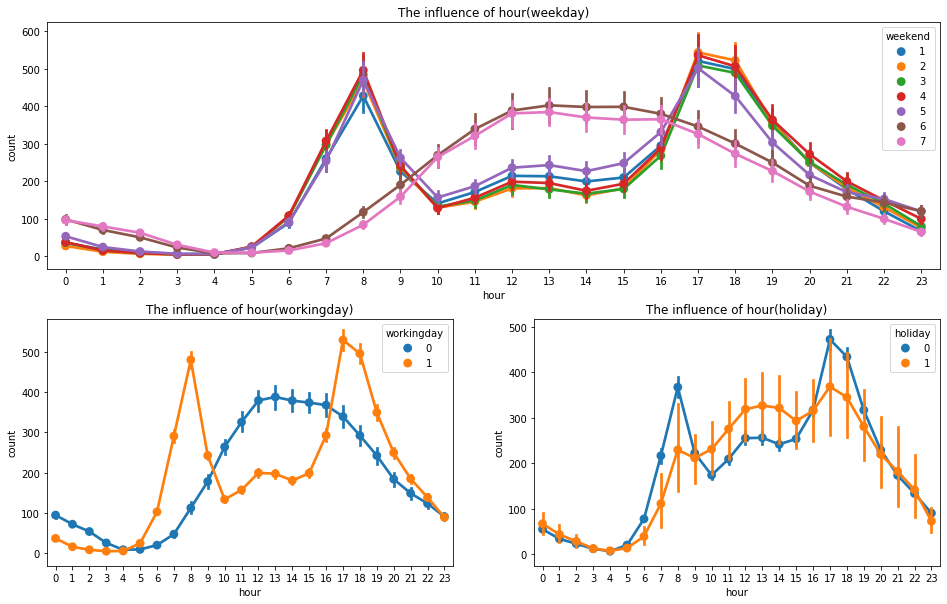
Figure 11

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.