# Predicting Bike Sharing demand in upcoming Quarter/Year

**Praveen Kumar Anwla**

# Introduction:-

## 1.1 Problem Statement:-

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 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.

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. So Merchant can organize sufficient bikes on Kiosk.

## 1.2 Hypothesis Generation:-

Before exploring the data to understand the relationship between variables, we'd focus on hypothesis generation first.

Here are some of the hypothesis which I thought could influence the demand of bikes:

**Hourly trend:** There must be high demand during office timings. Early morning and late evening can have different trend (cyclist) and low demand during 10:00 pm to 4:00 am.

**Daily Trend:** Registered users demand more bike on weekdays as compared to weekend or holiday.

**Rain:** The demand of bikes will be lower on a rainy day as compared to a sunny day. Similarly, higher humidity will cause to lower the demand and vice versa.

**Temperature:** In India, temperature has negative correlation with bike demand. But, after looking at Washington's temperature graph, I presume it may have positive correlation.

**Pollution:** If the pollution level in a city starts soaring, people may start using Bike (it may be influenced by government / company policies or increased awareness).

**Time:** Total demand should have higher contribution of registered user as compared to casual because registered user base would increase over time.

**Traffic:** It can be positively correlated with Bike demand. Higher traffic may force people to use bike as compared to other road transport medium like car, taxi etc

## 1.3 Data:-

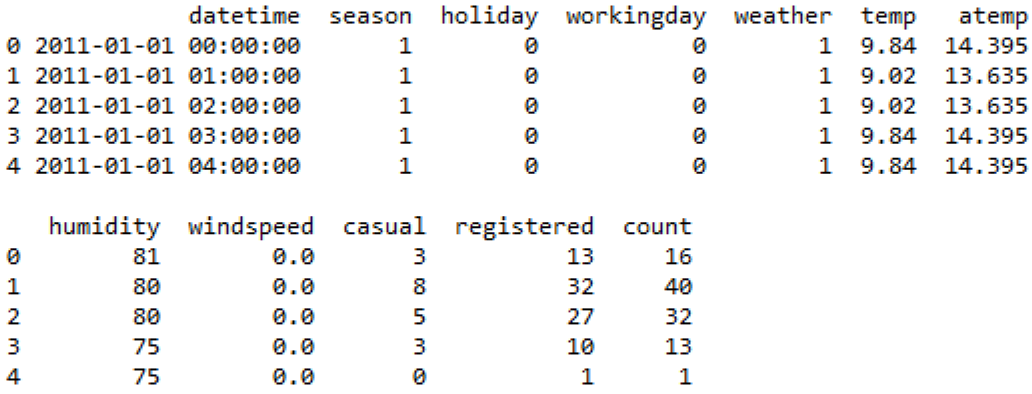
The dataset shows hourly rental data for two years (2011 and 2012). The training data set is for the first 19 days of each month. The test dataset is from 20th day to month's end. We are required to predict the total count of bikes rented during each hour covered by the test set.

In the training data set, they have separately given bike demand by registered, casual users and sum of both is given as count.

Training data set has 12 variables (see below) and Test has 9 (excluding registered, casual and count).

Our task is to build Regression models which will classify the predict the count of bikes depending on multiple variables / factors. Given below is a sample of the data set that we are using to predict the count of bikes

Bike Sharing Demand sample data:-

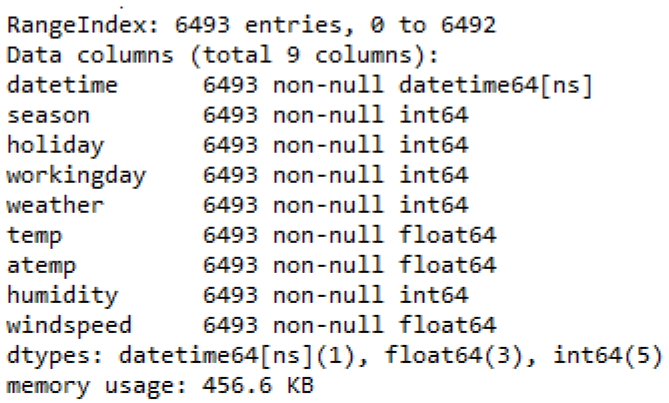


```
             datetime  season  holiday  workingday  weather  temp   atemp
0 2011-01-01 00:00:00       1        0           0        1  9.84  14.395
1 2011-01-01 01:00:00       1        0           0        1  9.02  13.635
2 2011-01-01 02:00:00       1        0           0        1  9.02  13.635
3 2011-01-01 03:00:00       1        0           0        1  9.84  14.395
4 2011-01-01 04:00:00       1        0           0        1  9.84  14.395

   humidity  windspeed  casual  registered  count
0        81        0.0       3          13     16
1        80        0.0       8          32     40
2        80        0.0       5          27     32
3        75        0.0       3          10     13
4        75        0.0       0           1      1
```

**Predictor Variables:-**

As we can see in the table below we have the following 9 variables/features, using which we have to correctly predict the count of the bikes:



```
RangeIndex: 6493 entries, 0 to 6492
Data columns (total 9 columns):
datetime      6493 non-null datetime64[ns]
season        6493 non-null int64
holiday       6493 non-null int64
workingday    6493 non-null int64
weather       6493 non-null int64
temp          6493 non-null float64
atemp         6493 non-null float64
humidity      6493 non-null int64
windspeed     6493 non-null float64
dtypes: datetime64[ns](1), float64(3), int64(5)
memory usage: 456.6 KB
```

# Chapter 2

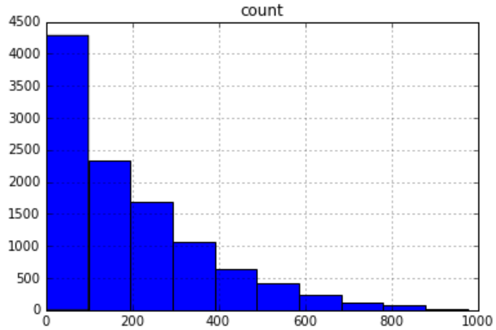
# Methodology

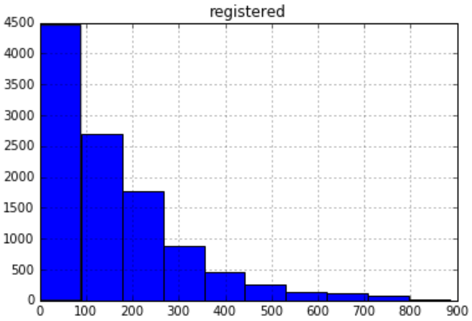
## 2.1 Data Pre Processing or Data Exploration:-

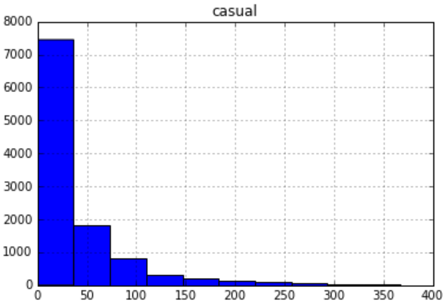
Any predictive modeling requires that we look at the data before we start modeling. However, in data mining terms looking at data refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis. To start this process we will first try and look at all the histograms of the variables. Most analysis like regression, require the data to be normally distributed. We can visualize that in a glance by looking at the distributions of the variable.

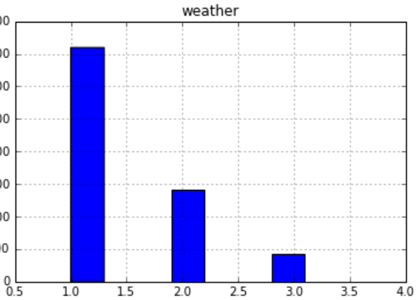
We see that dependent variables “**Count**”, **“registered**”, “**Casual**” are right skewed.

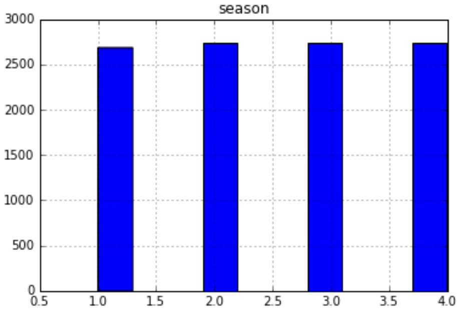
We should check for Missing values and Outliers in data, if there’re any.

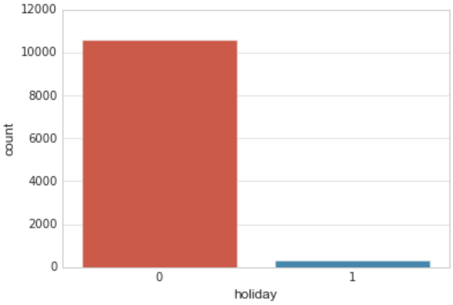

count
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
200
400
600
800
1000

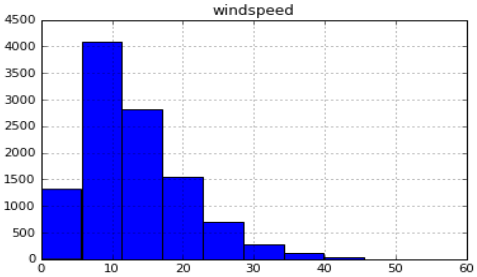

registered
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
100
200
300
400
500
600
700
800
900

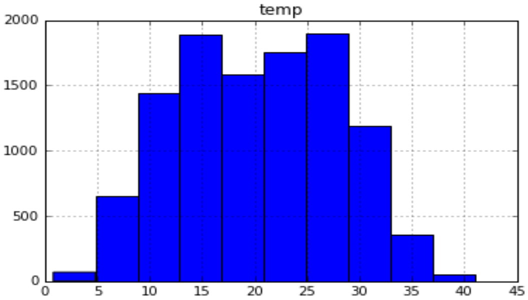

casual
8000
7000
6000
5000
4000
3000
2000
1000
0
0
50
100
150
200
250
300
350
400

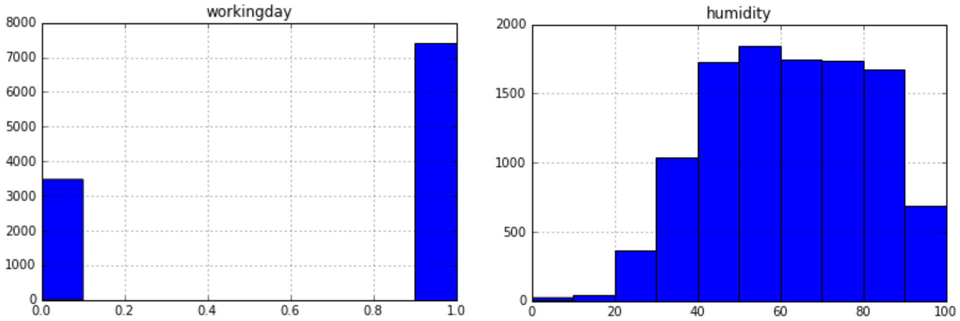

weather
8000
7000
6000
5000
4000
3000
2000
1000
0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0


season
3000
2500
2000
1500
1000
500
0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0


12000
10000
8000
6000
4000
2000
0
count
0
1
holiday


windspeed
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
10
20
30
40
50
60


temp
2000
1500
1000
500
0
0
5
10
15
20
25
30
35
40
45

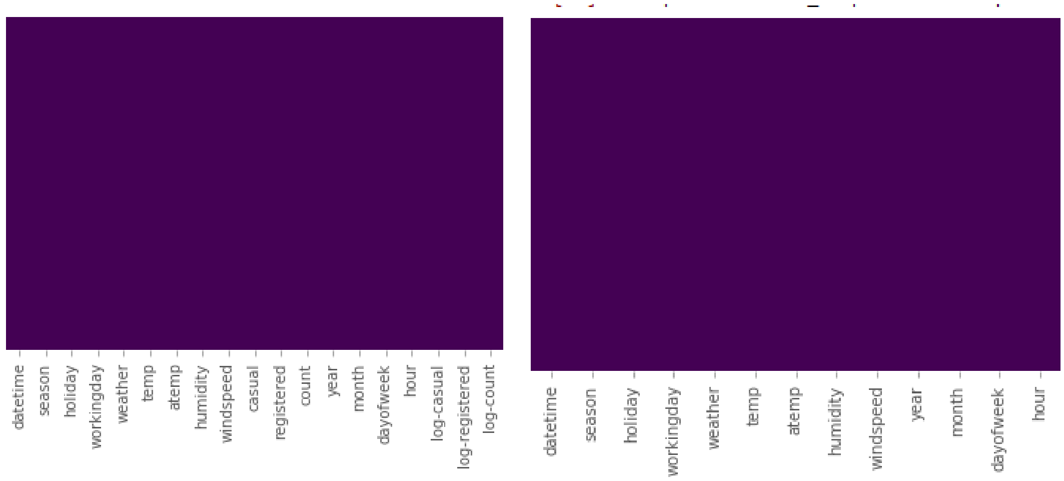


Few inferences can be drawn by looking at these histograms:

- Season has four categories of almost equal distribution
- Weather 1 has higher contribution i.e. mostly clear weather.
- As expected, mostly working days and variable holiday is also showing a similar inference. You can use the code above to look at the distribution in detail. Here you can generate a variable for weekday using holiday and working day. Incase, if both have zero values, then it must be a working day.
- Variables temp, atemp, humidity and windspeed  looks naturally distributed

### 2.2 Check for Missing Values:-

We see that neither train (Plot in left) nor test data set (Plot in right) has missing values.



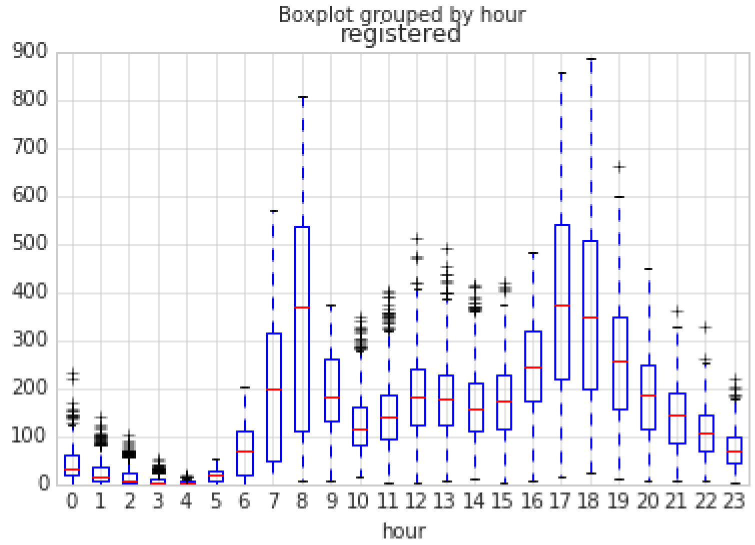
## 2.3 Hypothesis Testing:-

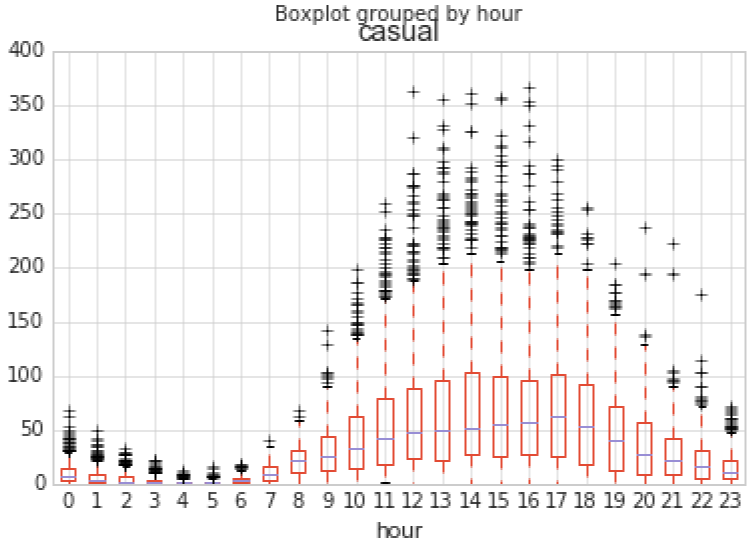
Let's test the hypothesis which we had generated earlier.  Here I have added some additional hypothesis from the dataset. Let's test them one by one:

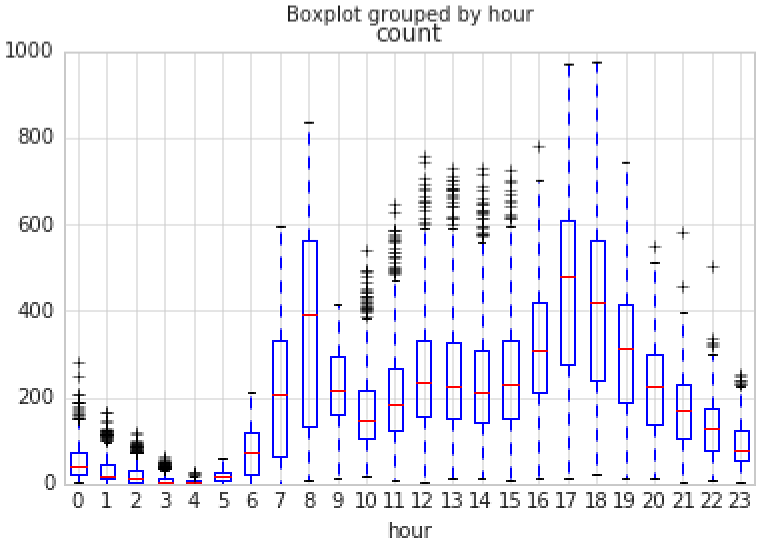
**Hourly trend**: We don't have the variable 'hour' with us right now. But we can extract it using the datetime column.

In below figure, as we can see the trend of bike demand over hours. As we can see bike demands are high between 8 A.M to 10 P.M.

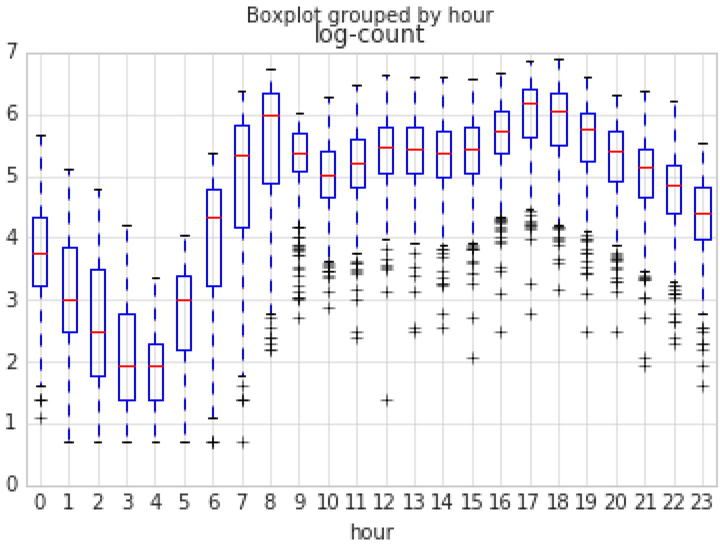
Below, as we can see that registered users have similar trend as count. Whereas, casual users have different trend. Thus, we can say that ‘hour’ is significant variable and our hypothesis is ‘true’.





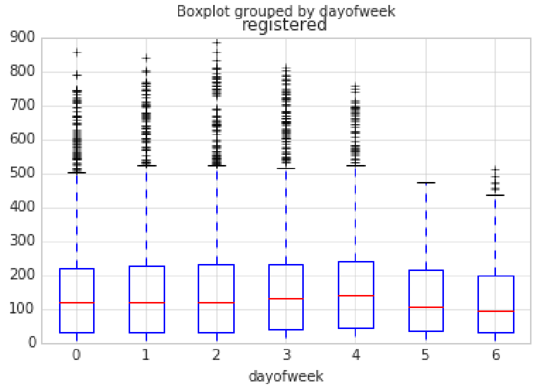


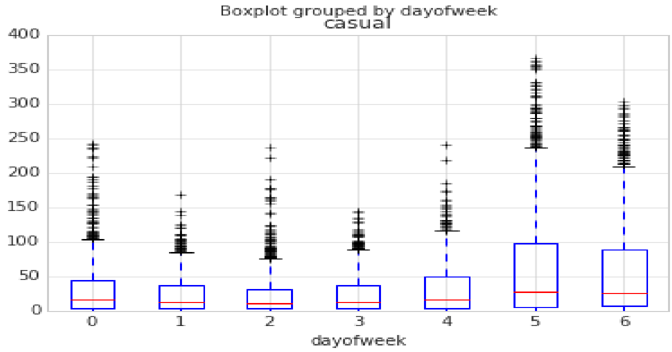
We’ve noticed that there are a lot of outliers while plotting the count of registered and casual users. These values are not generated due to error, so we consider them as natural outliers. They might be a result of groups of people taking up cycling (who are not registered). To treat such outliers, we will use logarithm transformation. Let’s look at the similar plot after log transformation.

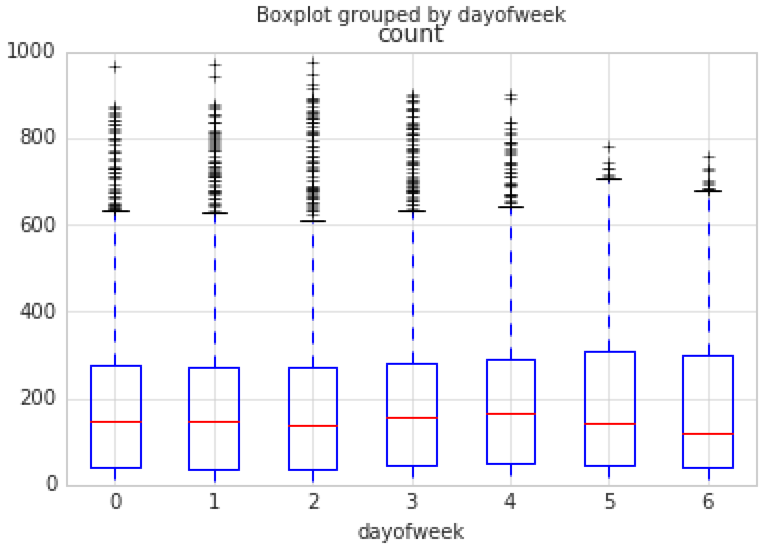


**Daily Trend:** Like Hour, we will generate a variable for day from datetime variable and after that we’ll plot it.

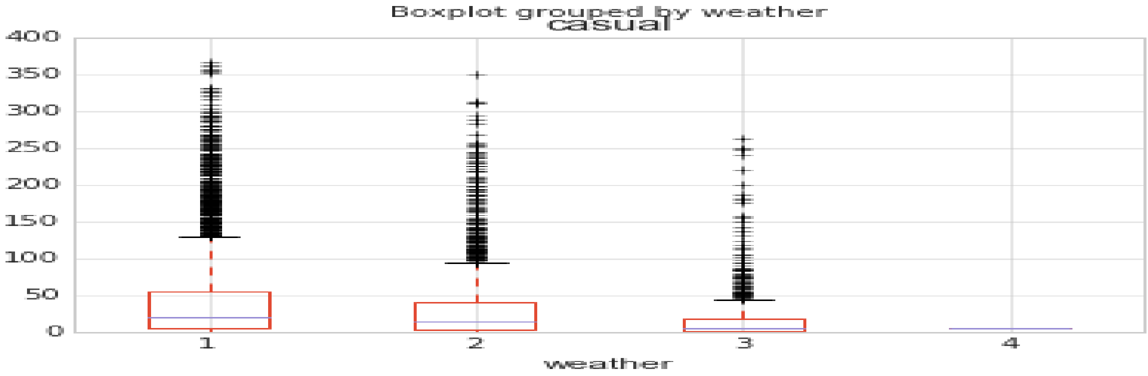
While looking at the plot, I can say that the demand of causal users increases over weekend.

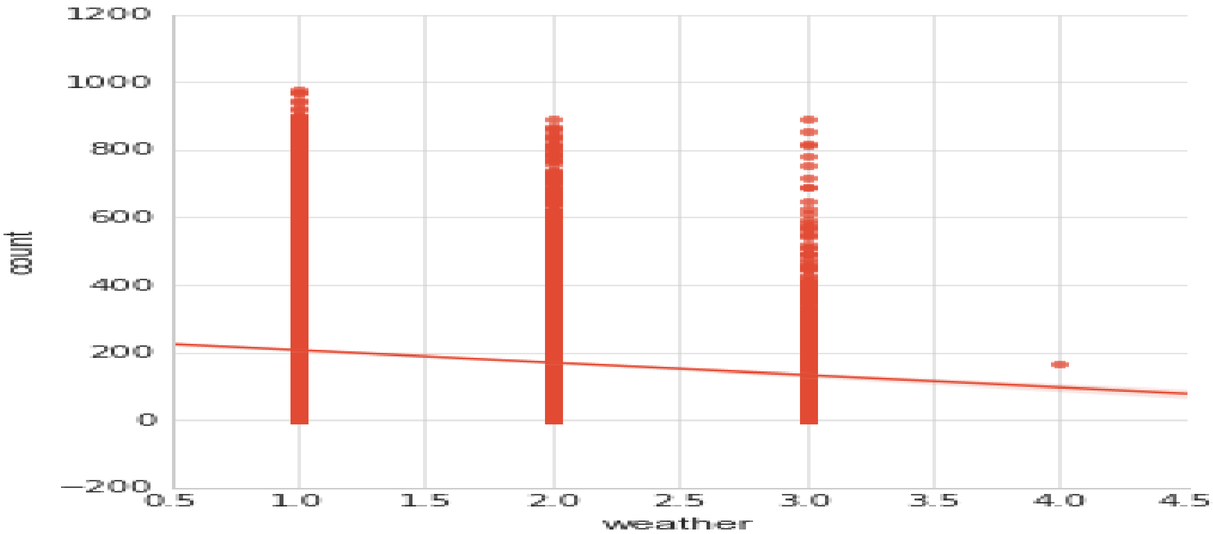






**Rain:** We don't have the 'rain' variable with us but have 'weather' which is sufficient to test our hypothesis. As per variable description, weather 3 represents light rain and weather 4 represents heavy rain. Take a look at the plot:

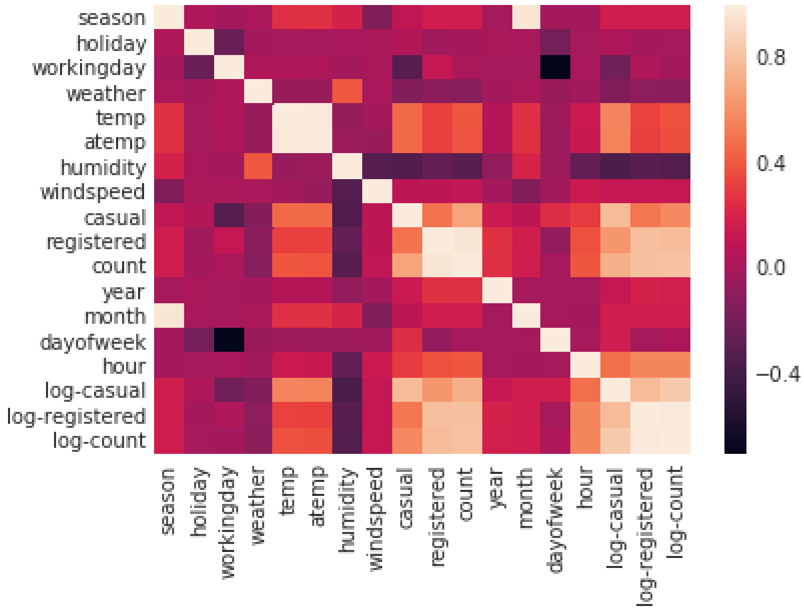


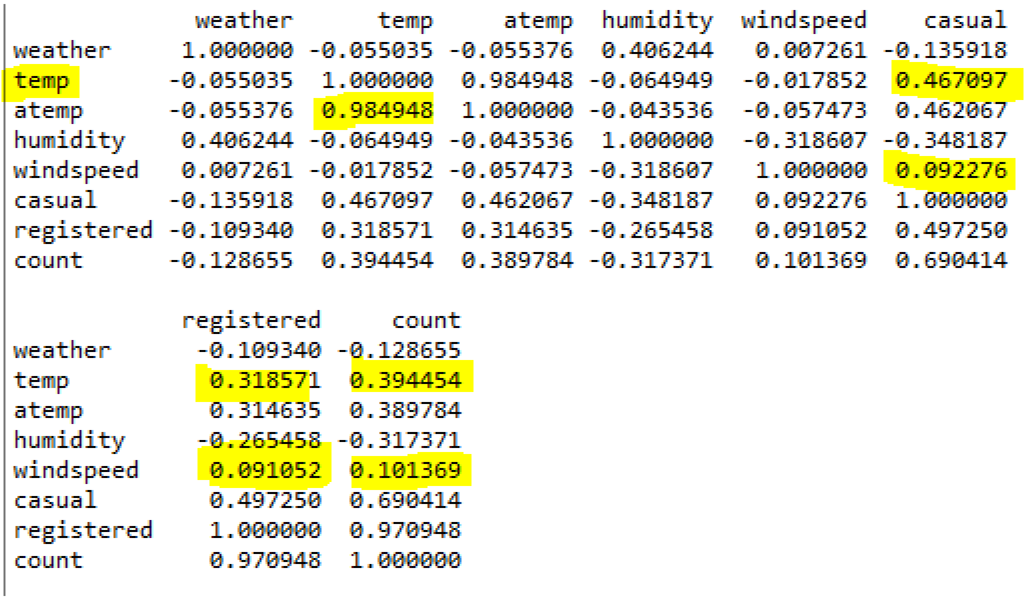


**Temperature, Windspeed and Humidity**: These are continuous variables so we can look at the correlation factor to validate hypothesis.

Here are a few inferences you can draw by looking at the above histograms:

- Variable temp is positively correlated with dependent variables (casual is more compare to registered)
- Variable atemp is highly correlated with temp.
- Windspeed has lower correlation as compared to temp and humidity





```
             weather      temp     atemp  humidity  windspeed    casual
weather     1.000000 -0.055035 -0.055376  0.406244   0.007261 -0.135918
temp       -0.055035  1.000000  0.984948 -0.064949  -0.017852  0.467097
atemp      -0.055376  0.984948  1.000000 -0.043536  -0.057473  0.462067
humidity    0.406244 -0.064949 -0.043536  1.000000  -0.318607 -0.348187
windspeed   0.007261 -0.017852 -0.057473 -0.318607   1.000000  0.092276
casual     -0.135918  0.467097  0.462067 -0.348187   0.092276  1.000000
registered -0.109340  0.318571  0.314635 -0.265458   0.091052  0.497250
count      -0.128655  0.394454  0.389784 -0.317371   0.101369  0.690414

            registered     count
weather      -0.109340 -0.128655
temp          0.318571  0.394454
atemp         0.314635  0.389784
humidity     -0.265458 -0.317371
windspeed     0.091052  0.101369
casual        0.497250  0.690414
registered    1.000000  0.970948
count         0.970948  1.000000
```

## 3. Model Building

As we know that dependent variables have natural outliers so we will predict log of dependent variables.

### 1) Random Forest:-

I've used Random Forest algorithm to predict the demands of bike. mean_absolute_percentage_error (MAPE) is used as evaluation metric for this model. I've got **2.31 % MAPE**

### 2)GBM(Gradient Bossting Method):-

I've used GBM (Gradient Boosting Method)  algorithm to predict the demands of bike. mean_absolute_percentage_error (MAPE) is used as evaluation metric for this model. I've got 0.025%  MAPE.

MAPE for GBM is far lesser than Random Forest MAPE.

We'll go with **GBM** for this problem based on result of chosen evaluation metric.