**Project 1:**

***Bike***

***Rental***

***Count***

**Submitted by- Tanya Shrivastava**

**Submitted on -03/10/2019**

**Contents**

**1 Introduction**

1.1 Problem Statement

1.2 Data

**2 Methodology**

2.1 Pre Processing

2.1.1 Exploratory Data Analysis

2.1.2 Missing Value Analysis

2.1.3 Outlier Analysis

2.1.4 Feature Selection

2.2 Modeling

2.2.1 Model Selection

2.2.2 Regression Trees

**3 Visualizations**

3.1 Visualization on seasonal condition

3.2 Visualization on result stored on weather conditions

1. **INTRODUCTION**

**1.1 Problem statement**

Choosing bike sharing system as a medium of transport will allow a more eco friendly way of transportation . A bike rental is a bicycle business that rents bikes for short periods of time. Bike rental shops rent by the day or week as well as by the hour, and these provide an excellent opportunity for people who don't have access to a vehicle:

Travellers

Tourists .

Specialized bike rental shops thus typically operate at beaches, parks, or other locations that tourists frequently visit . In this case, the fees are set to encourage renting the bikes for a few hours at a time, rarely more than a day. The objective of this Case is to predict the bike rental count based on the environmental and seasonal settings, So that required bikes would be arranged and managed by the shops according to environmental and seasonal conditions.

**1.2 Data**

Our task is to build regression models which will predict the count of bike rented depending on various environmental and seasonal conditions Given below is a sample of the data set that we are using to predict the count of bike rents:

Table 1.1: Sample Data (Columns: 1-8)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | instant | dteday | season | yr | mnth | holiday | weekday | workingday | | | |
|  |  |  |  |  |  |  |  |  |  |  |  |
| 1 | | 1/1/2011 | 1 | 0 | 1 | 0 | 6 | 0 |  |  |  |
| 2 | | 1/2/2011 | 1 | 0 | 1 | 0 | 0 | 0 |  |  |  |
| 3 | | 1/3/2011 | 1 | 0 | 1 | 0 | 1 | 1 |  |  |  |
| 4 | | 1/4/2011 | 1 | 0 | 1 | 0 | 2 | 1 |  |  |  |
| 5 | | 1/5/2011 | 1 | 0 | 1 | 0 | 3 | 1 |  |  |  |
| 6 | | 1/6/2011 | 1 | 0 | 1 | 0 | 4 | 1 |  |  |  |
|  |  |  | | | |  |  |  |  |  |  |
|  |  | Table 1.2: Sample Data (Columns: 7-16) | | | |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  | |  |  |
|  | weathersit | temp | atemp | hum | windspeed | casual | registered | cnt | | | |
|  |  |  |  |  |  |  |  |  |  |  |  |
| 2 | | 0.344167 | 0.363625 | 0.805833 | 0.160446 | 331 | 654 | 985 |  |  |  |
| 2 | | 0.363478 | 0.353739 | 0.696087 | 0.248539 | 131 | 670 | 801 |  |  |  |
| 1 | | 0.196364 | 0.189405 | 0.437273 | 0.248309 | 120 | 1229 | 1349 |  |  |  |
| 1 | | 0.2 | 0.212122 | 0.590435 | 0.160296 | 108 | 1454 | 1562 |  |  |  |
| 1 | | 0.226957 | 0.22927 | 0.436957 | 0.1869 | 82 | 1518 | 1600 |  |  |  |
| 1 | | 0.204348 | 0.233209 | 0.518261 | 0.0895652 | 88 | 1518 | 1606 |  |  |  |

Variables present in given dataset are instant, dteday, season, yr, mnth, holiday, weekday, workingday, weathersit, temp, atemp, hum, windspeed, casual, registered, cnt

The details of variable present in the dataset are as follows - instant: Record index

dteday: Date

season: Season (1:springer, 2:summer, 3:fall, 4:winter)

yr: Year (0: 2011, 1:2012)

mnth: Month (1 to 12)

hr: Hour (0 to 23)

holiday: weather day is holiday or not (extracted fromHoliday Schedule)

weekday: Day of the week

workingday: If day is neither weekend nor holiday is 1, otherwise is 0.

weathersit: (extracted fromFreemeteo)

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: Normalized temperature in Celsius. The values are derived via (t-t\_min)/(t\_max-t\_min),

t\_min=-8, t\_max=+39 (only in hourly scale)

atemp: Normalized feeling temperature in Celsius. The values are derived via (t-t\_min)/(t\_maxt\_min),

t\_min=-16, t\_max=+50 (only in hourly scale)

hum: Normalized humidity. The values are divided to 100 (max)

windspeed: Normalized wind speed. The values are divided to 67 (max)

casual: count of casual users

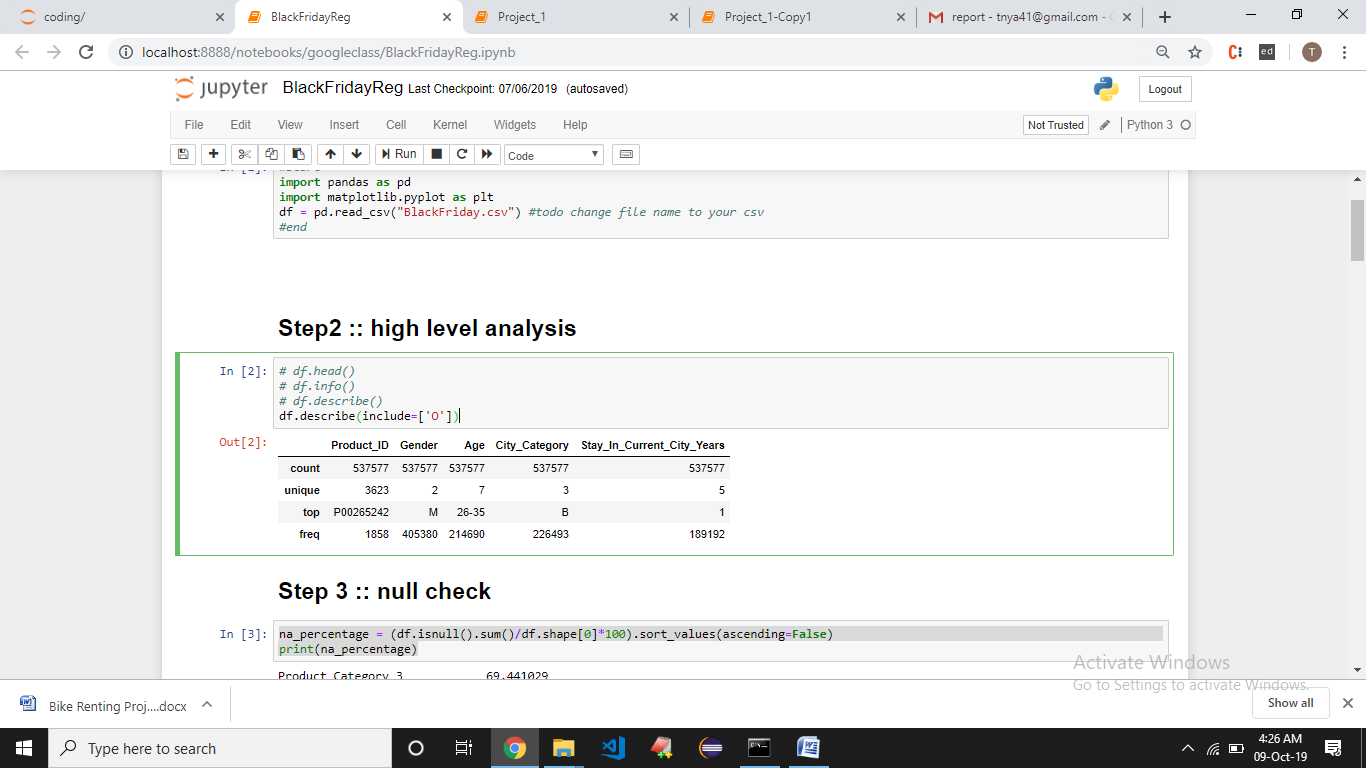
registered: count of registered users

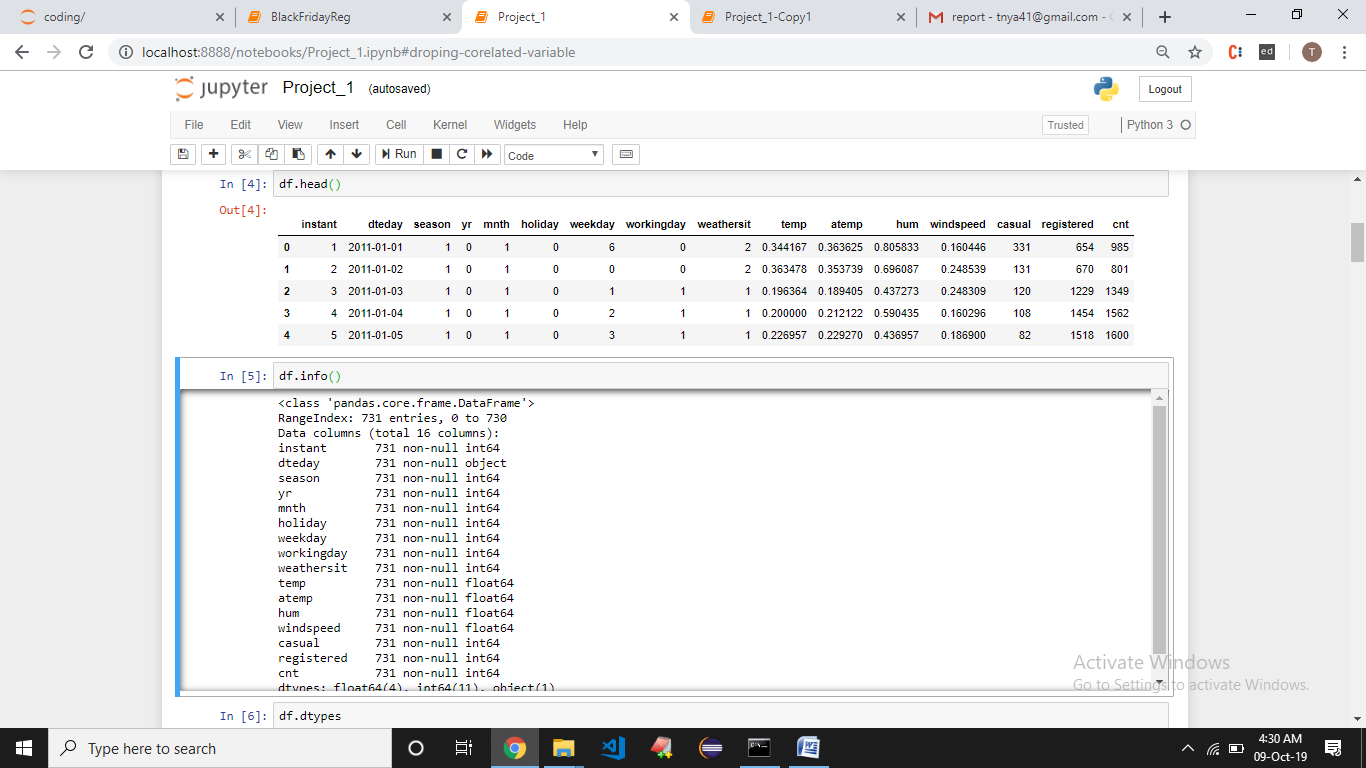
cnt: count of total rental bikes including both casual and registered

**2. Methodology**

**2.1 Pre Processing**

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.



****

**2.1.1 Exploratory Data Analysis**

In exploring the data we have

Converted season, mnth, workingday, weathersit into categorical variables Feature Engineering :Changed deday variables’s date value to day of date and converted to categorical variable having 31 levels as a month has 31

days.

Deleted instant variable as it is nothing but an index.

Omitted registered and casual variable as sum of registered and casual is the total count that is what we have to predict.

=

**2.1.2 Missing Value Analysis**

Missing value analysis is done to check is there any missing value present in given dataset. Missing values can be easily treated using various methods like mean, median method, knn method to impute missing value.

In R function(x){sum(is.na(x))} is the function used to check the sum of missing values.

In python bike\_train.isnull().sum() is used to detect any missing value



There is no missing value found in given dataset.

**2.1.3 Outlier Analysis**

Outlier analysis is done to handle all inconsistent observations present in given dataset. As outlier analysis can only be done on continuous variable.

Figure 2.1 and 2.2 are visualization of numeric variable present in our dataset to detect outliers using boxplot. Outliers will be detected with red color

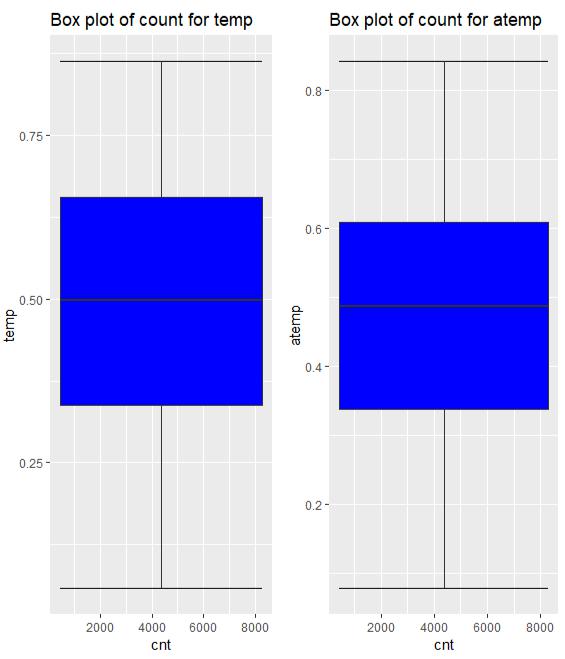


Figure 2.1 Boxplot graph of temp and atemp variables

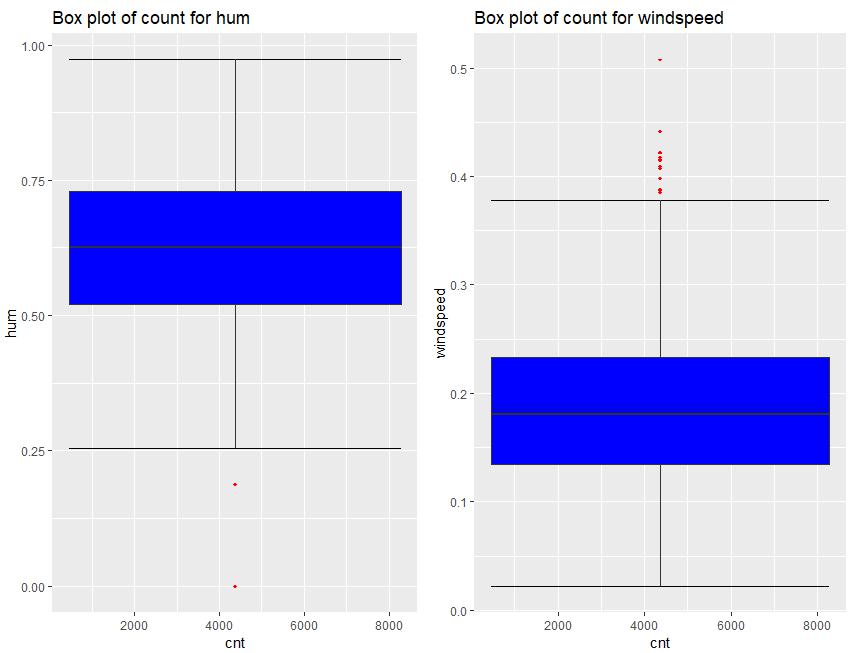


Figure 2.2 Boxplot graph of hum and windspeed variables

According to above visualizations there is no outlier found in temp and atemp variable but there are few outliers found in windspeed and hum variable.

As windspeed variable defines the windspeed on a particular day and hum defines the humidity of that day so we can neglect these outliers because both these variable define environmental condition. Due to drastic change in weather like strome, heavy rain condition.

**2.1.4 Feature Selection**

Feature selection analysis is done to Select subsets of relevant features (variables, predictors) to be in model construction.

As our target variable is continuous so we can only go for correlation check. As chi-square test is only for categorical variable.

Figure 2.4 show a correlation plot for all numeric variable present in dataset

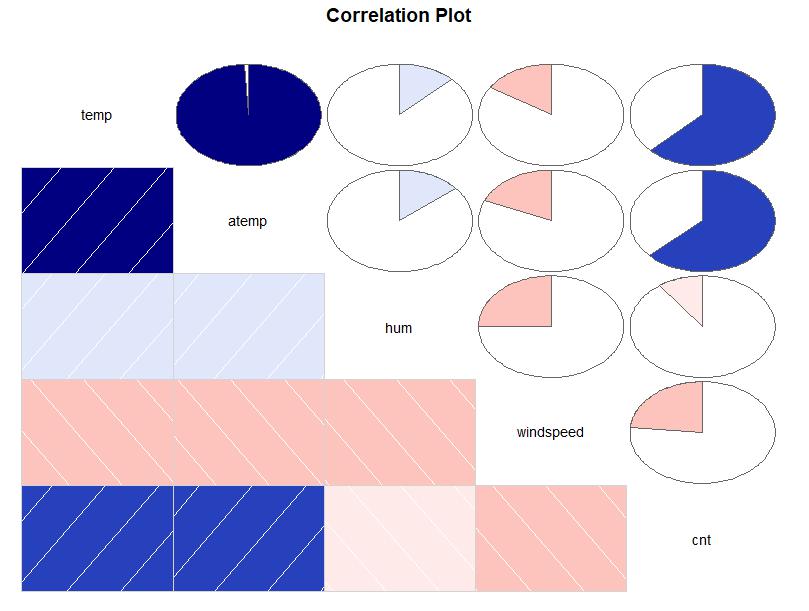


Figure 2.4 correlation plot

In above visualization we can see that only 2 variables are highly correlated with each other. Dark blue color represent highly correlated and light color represent very less correlated so we have a choice to remove either temp or atemp because these variables contains nearly equal information. So I have removed atemp variable from dataset.

**2.1.4 Feature Scaling**

Feature scaling includes two functions normalization and standardization. It is done reduce unwanted variation either within or between variables and to bring all of the variables into proportion with one another.

In given dataset all numeric values are already present in normalized form.

**2.2 Modeling**

**2.2.1 Model Selection**

In this case we have to predict the count of bike renting according to environmental and

seasonal condition. So the target variable here is a continuous variable. For Continuous

we can use various Regression models. Model having less error rate and more

accuracy will be our final model.

Models built are

Random Forest (with 200 trees)

**2.2.2 Random Forest**

In Random forest we have divided the dataset into train and test part using random sampling. For this model we have divided the dataset into train and test part using random sampling Where train contains 80% data of data set and test contains 20% data and contains 12 variable where 12th variable is the target variable.

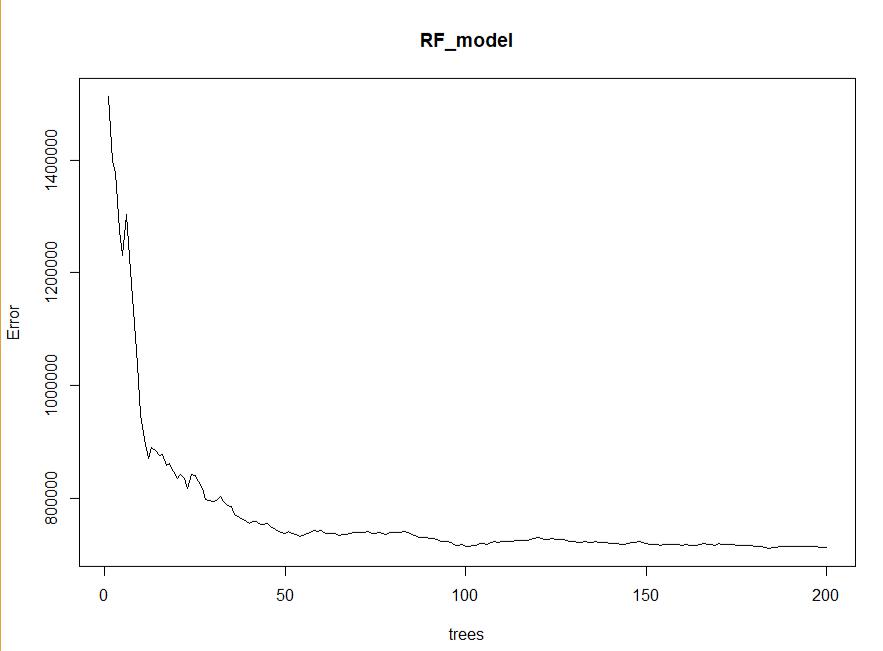


Figure2.2.2

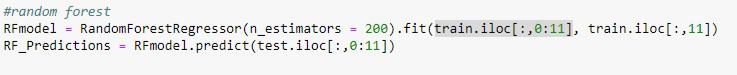
Above Figure2.2.2 represents the curve of error rate as the number of trees increases.

After 200 trees the error rate reaches to be constant.

In this model we are using 200 trees to predict the target variable.

Creating Model

In Python

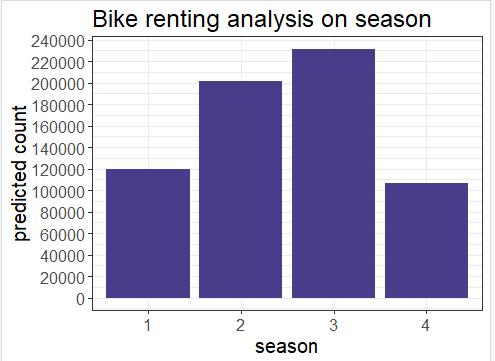


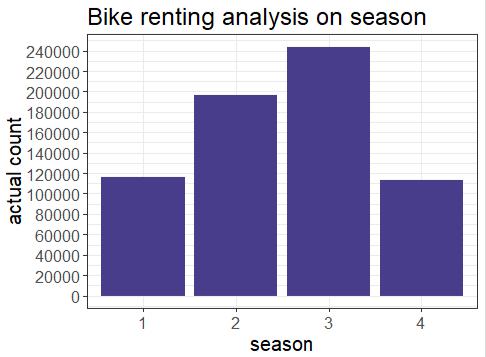
**In R**



**3.Visualizations**

**3.1 Visualization on result stored on seasonal settings**

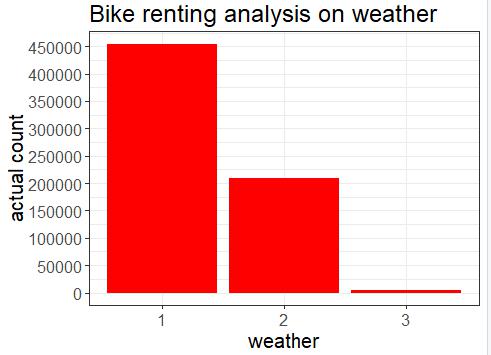




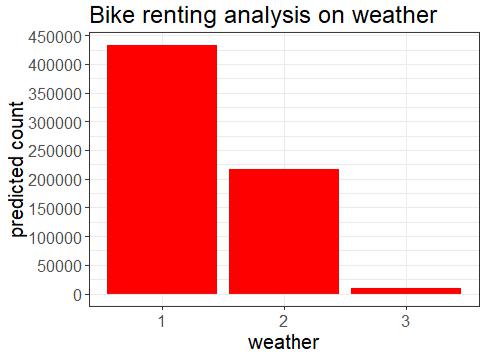
season: Season (1:springer, 2:summer, 3:fall, 4:winter)

Above two bar graph represents the comparison of predicted count value and actual count value based on seasonal condition.

**4.2 Visualization on result stored on weather conditions**



18



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

Above bar graph shows predicted count and actual count based on weather conditions

According to Seasonal and weather condition bar graph we can clearly notice that fall season that is autumn and where weather conditions are clear, few or partly cloudy on these conditions bike rent count is quite high than any other condition.

19