# Bike Rental Predication

Aditya Tiwari

26th June, 2019

# Contents

# Chapter 1

# Introduction

### 1.1 Problem Statement

The objective of this Case is to Predication of bike rental count on daily based on the environmental and seasonal settings. Predicting demand of ride sharing systems has become a common problem in the modern world, with companies such as Uber and Ola emerging in the car-sharing service. Similarly, bike-sharing services have gained considerable traction in the past decade. Heavy street traffic in busy cities and a desire for an environmentally friendly form of transportation make biking an attractive alternative to traveling by car.

### 1.2 Data

Before exploring the data, we need to understand the relationship between variables. The dataset shows hourly rental data for two years (2011 and 2012). We are required to predict the total count of bikes rented during each day.

The features are:

#### Independent Variables

1. Instant (Record index)
2. Dteday (Date)
3. Season (1:spring, 2:summer, 3:fall, 4:winter)
4. Year (0: 2011, 1:2012)
5. Month (1 to 12)
6. Holiday : whether that day is holiday or not
7. Weekday : day of the week
8. Working-day : if day is neither weekend nor holiday , value is 1. Otherwise 0

9. Weather situation :
   —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
10. Normalized temperature in Celsius
11. Normalized feeling temperature in Celsius
12. Normalized humidity
13. Normalized wind speed

**Dependent Variables**

1. Count of casual users
2. Count of registered users
3. Count of total rental bikes including both casual and registered

# Chapter 2

# Methodology

## 2.1 Pre-processing

### 2.1.1 Data Exploration

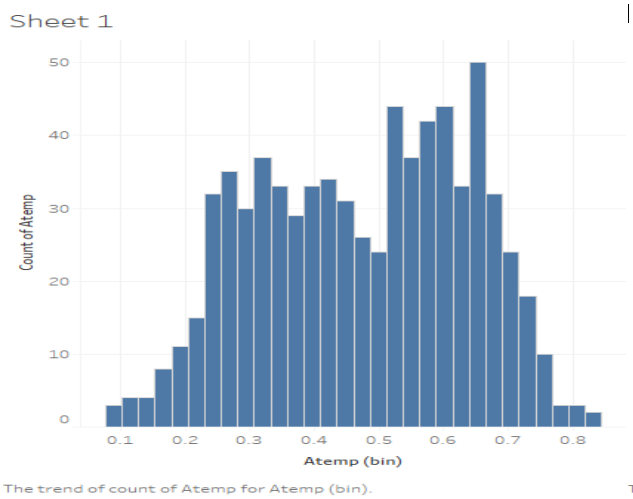
Any predictive modelling requires that we look at the data before we start modelling. However, in data mining terms looking at data refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis. To start this process, we will first try and look at the distributions of the variables

Firstly, we will make sure if the data type of each variable is appropriate.

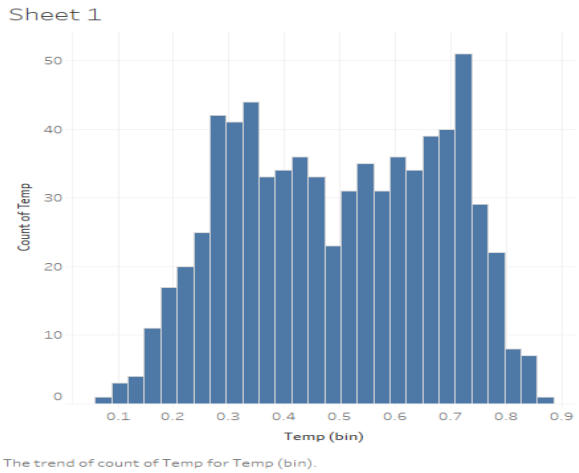
```
instant          int64
dteday          object
```

```
season          category
yr                 int64
mnth            category
holiday             bool
weekday         category
workingday          bool
weathersit      category
temp             float64
atemp            float64
hum              float64
windspeed        float64
casual             int64
registered         int64
cnt                int64
```

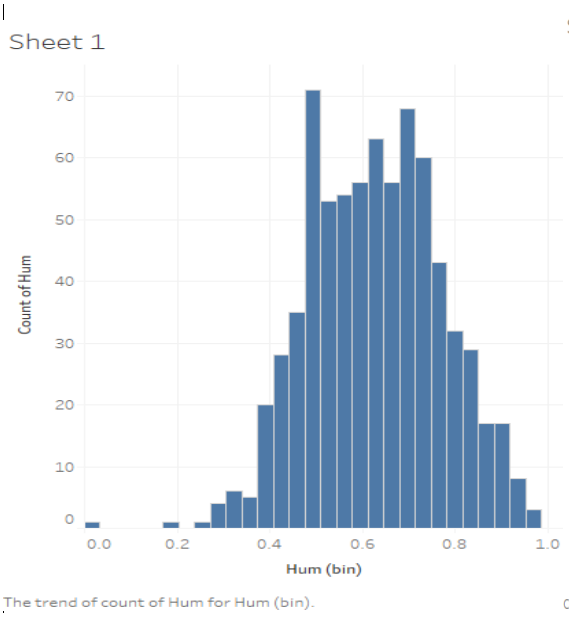
## 2.1.2 Data Distribution

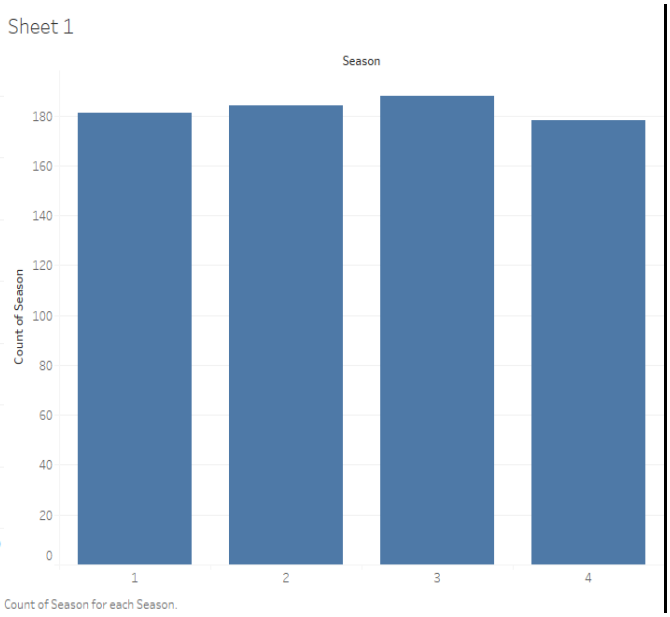



Histogram of data['atemp']



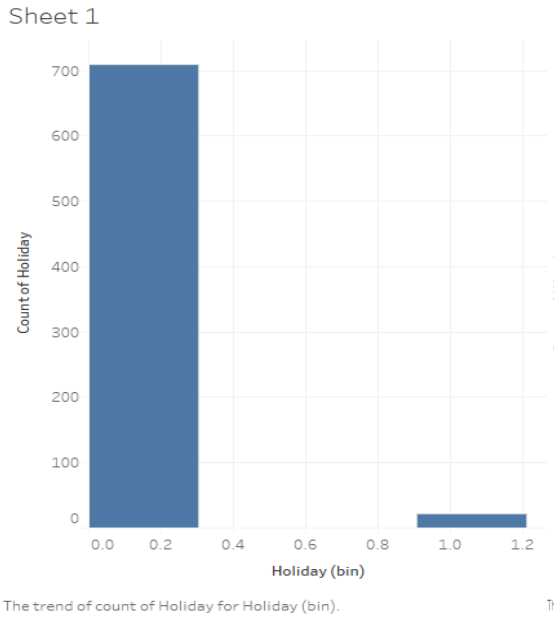

Histogram of data['temp']

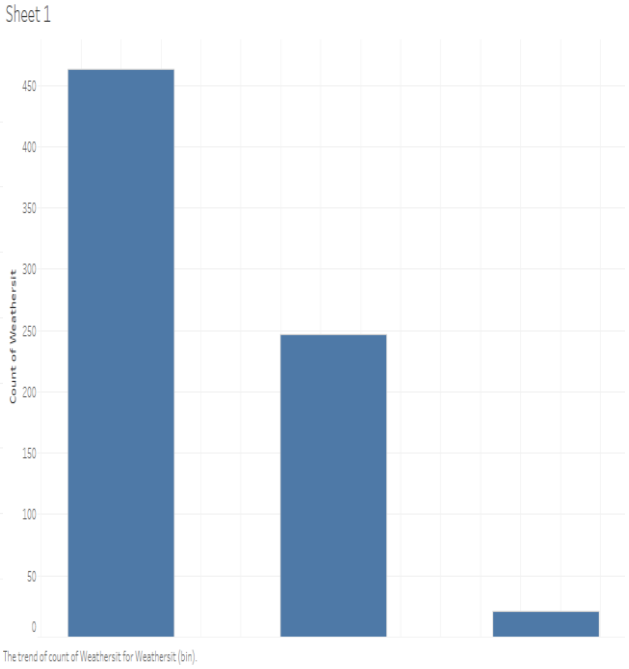



Histogram of data['Hum']

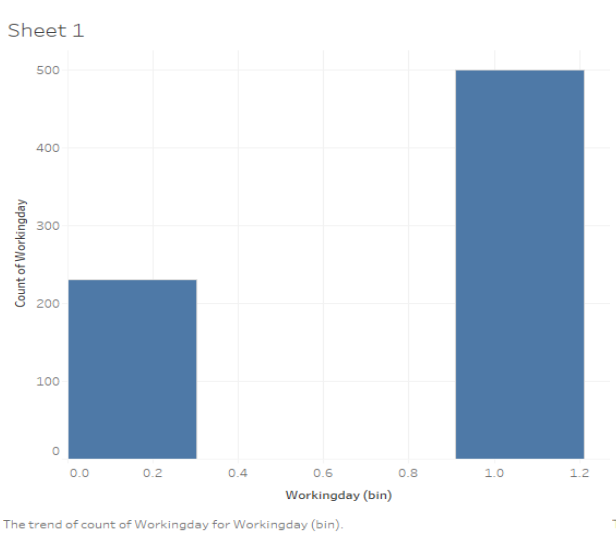
Histogram of data['season']



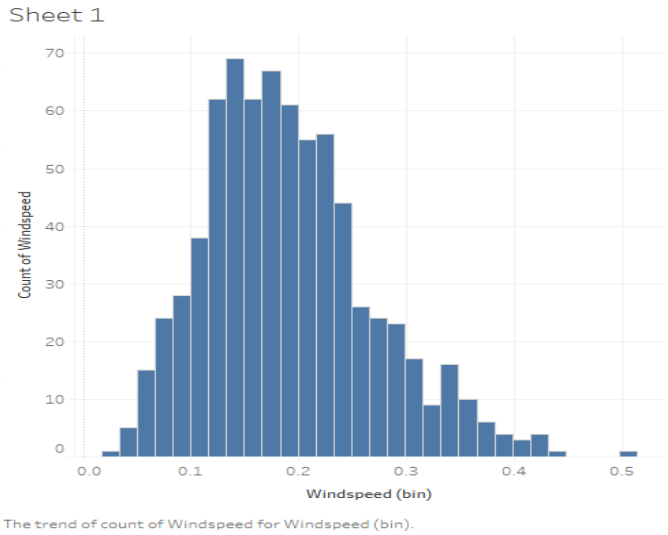




Histogram of data['Holiday']

Histogram of data['weathersit']




Histogram of data['workingday']

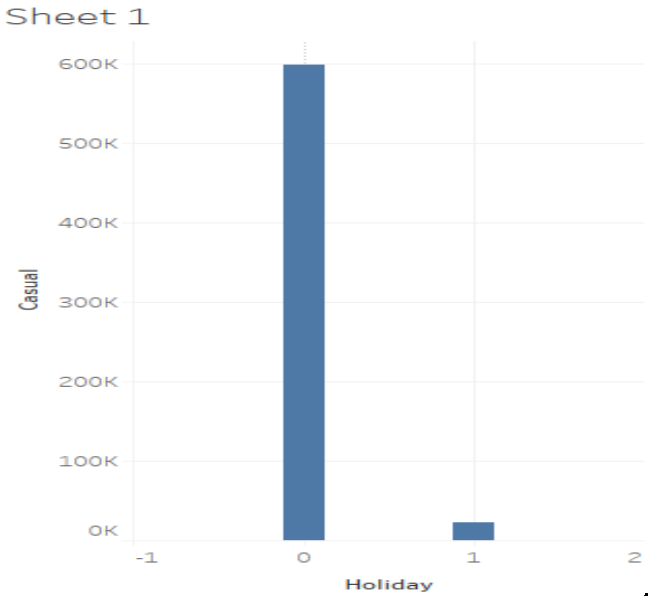



Histogram of data['windspeed']

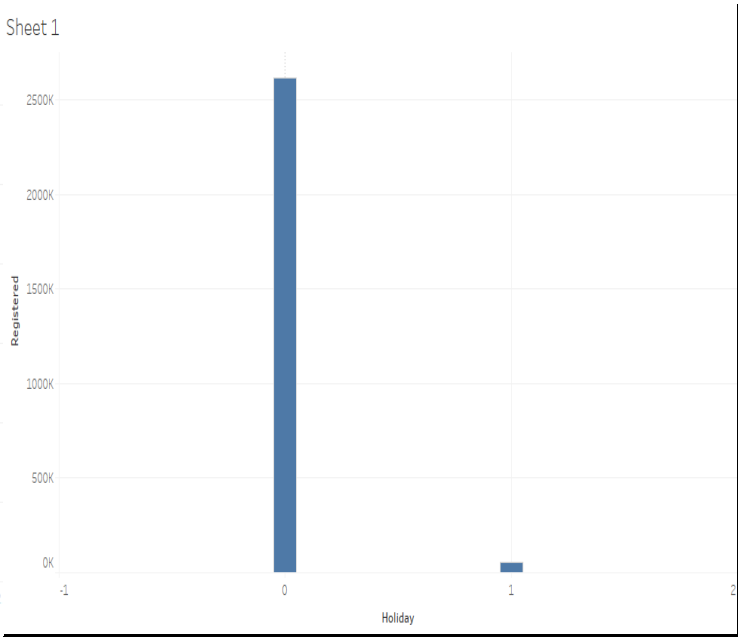
Few inferences can be drawn by looking at these histograms:

- Season has four categories of almost equal distribution
- Weather 1 has higher contribution i.e. mostly clear weather, and no contribution of weather 4.
- Variables temp, atemp, humidity and windspeed looks naturally distributed.

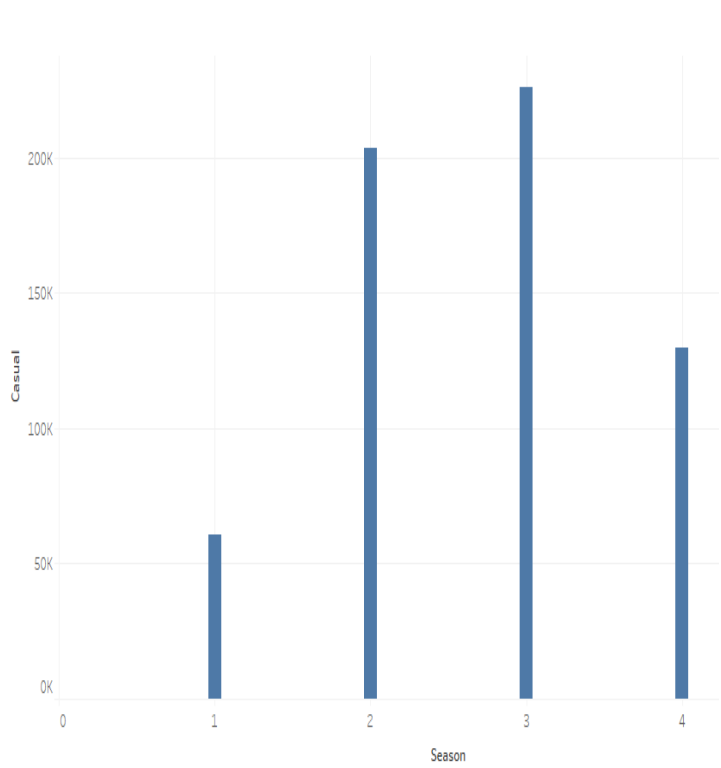
We will further look at the behaviour of the independent variables on the dependent variables, casual and registered separately.



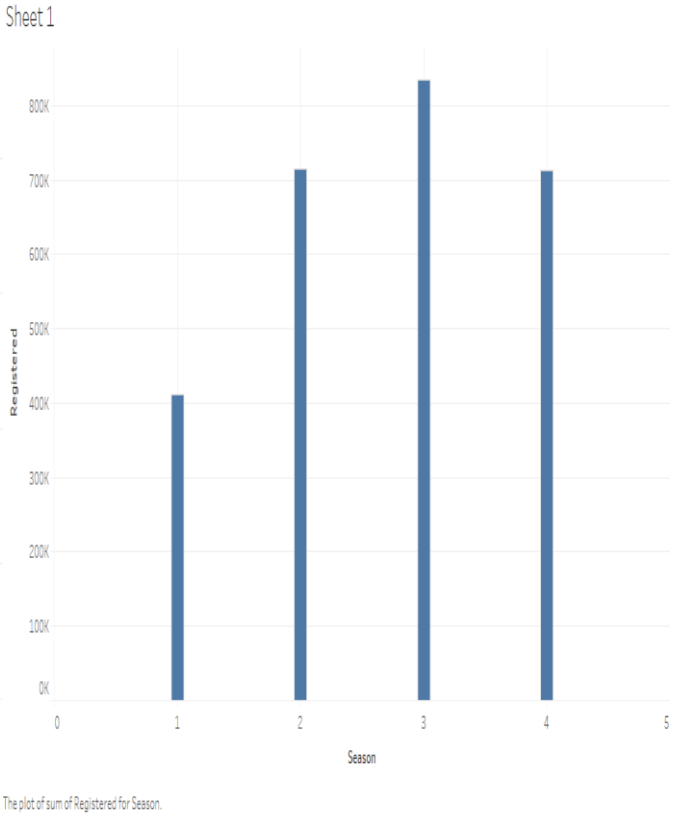

**casual vs holiday**



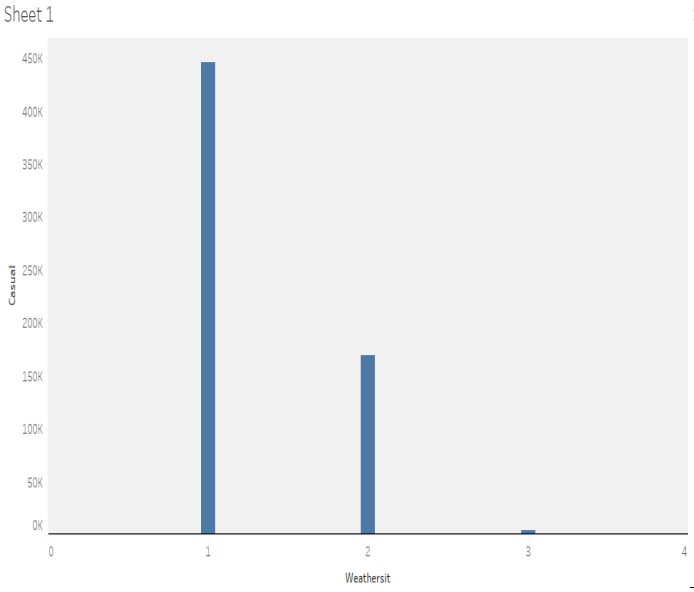

**registered vs holiday**



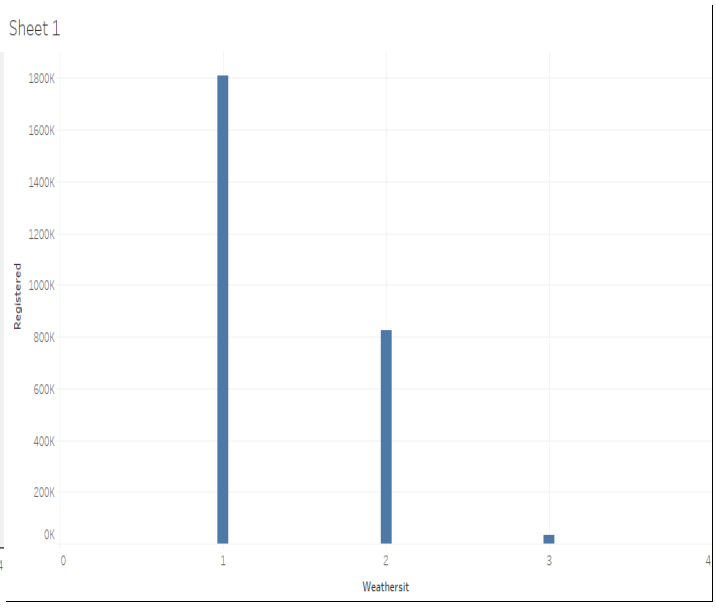

**casual vs season**



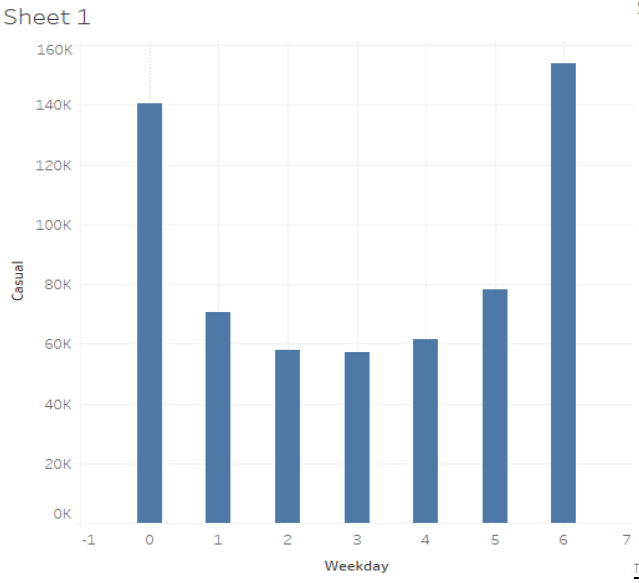

**registered vs season**



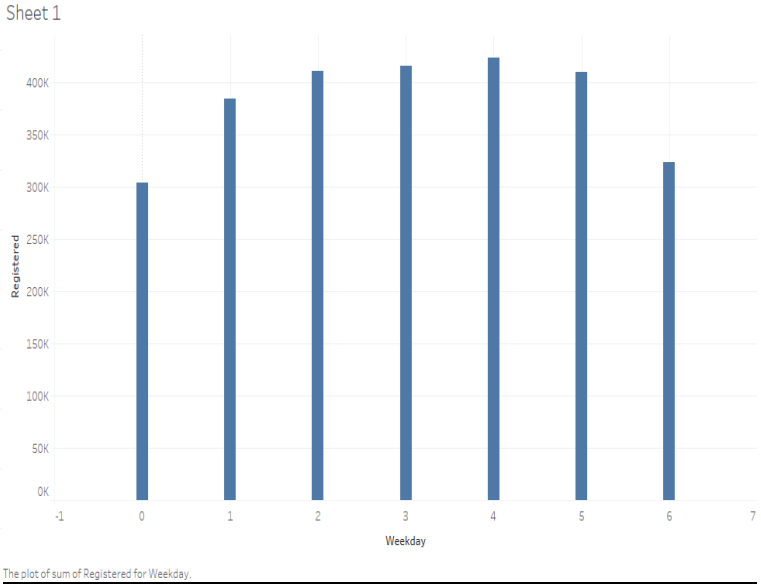

**casual vs weathersit**



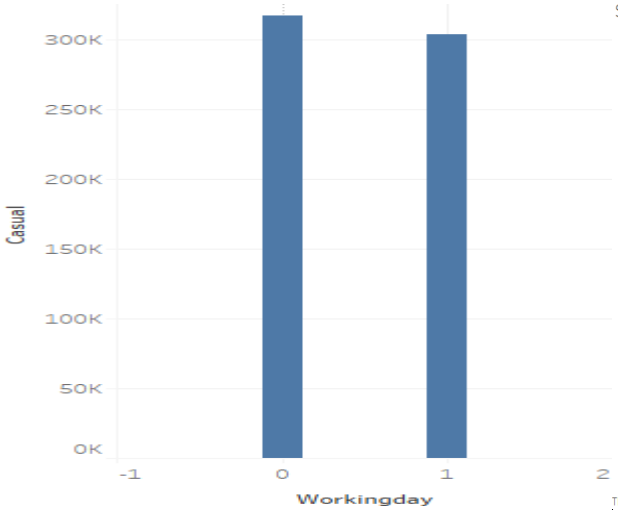

**registered vs weathersit**



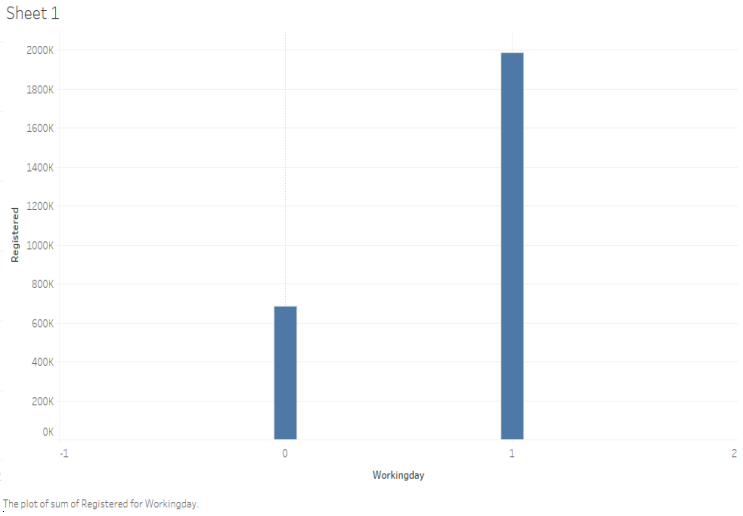

**casual vs weekday**



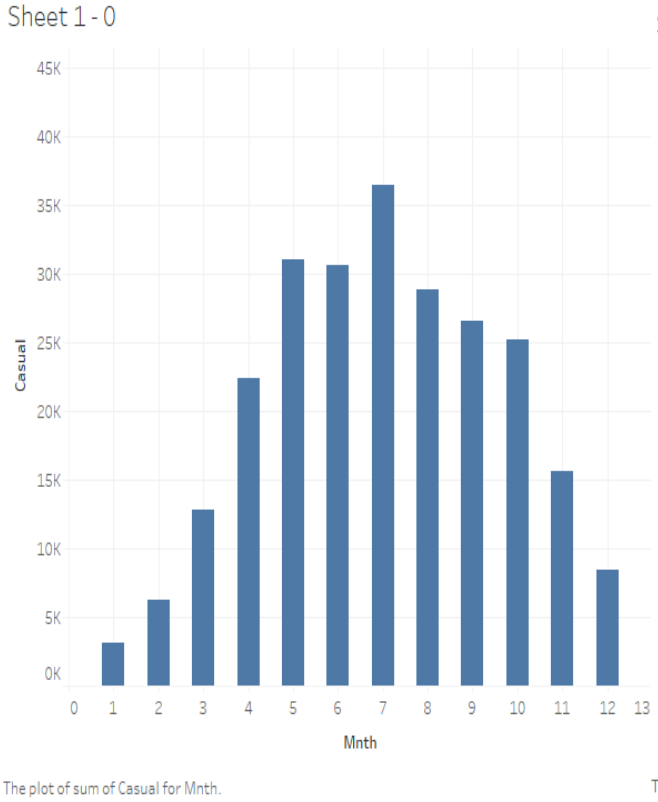

**registered vs weekday**



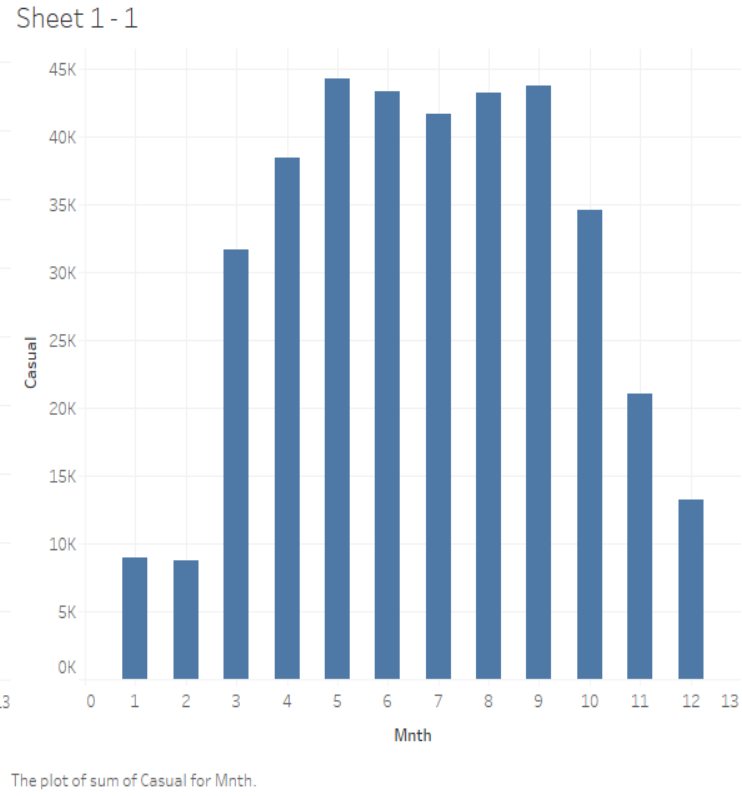

**casual vs workingday**



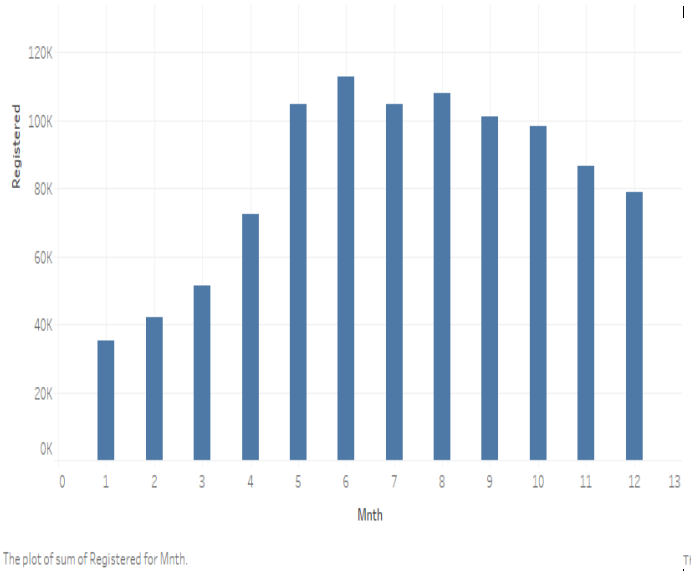

**registered vs workingday**



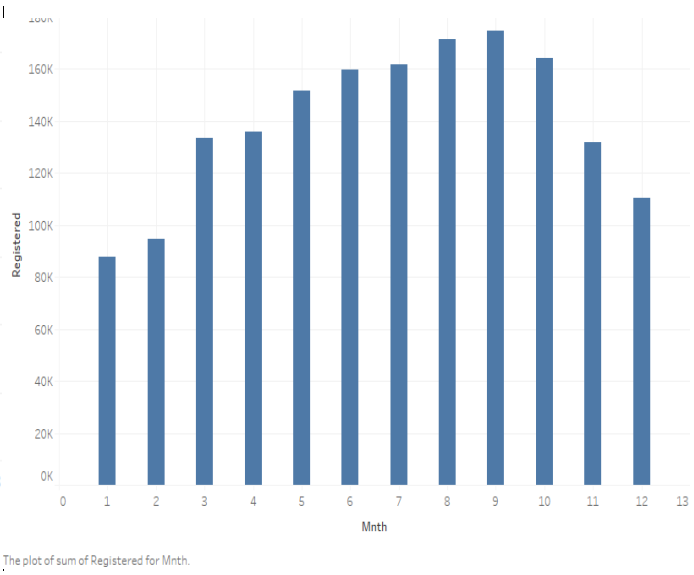

**casual vs mnth (year:2011)**




**casual vs mnth(year:2012)**




**registered vs mnth(year:2011)**



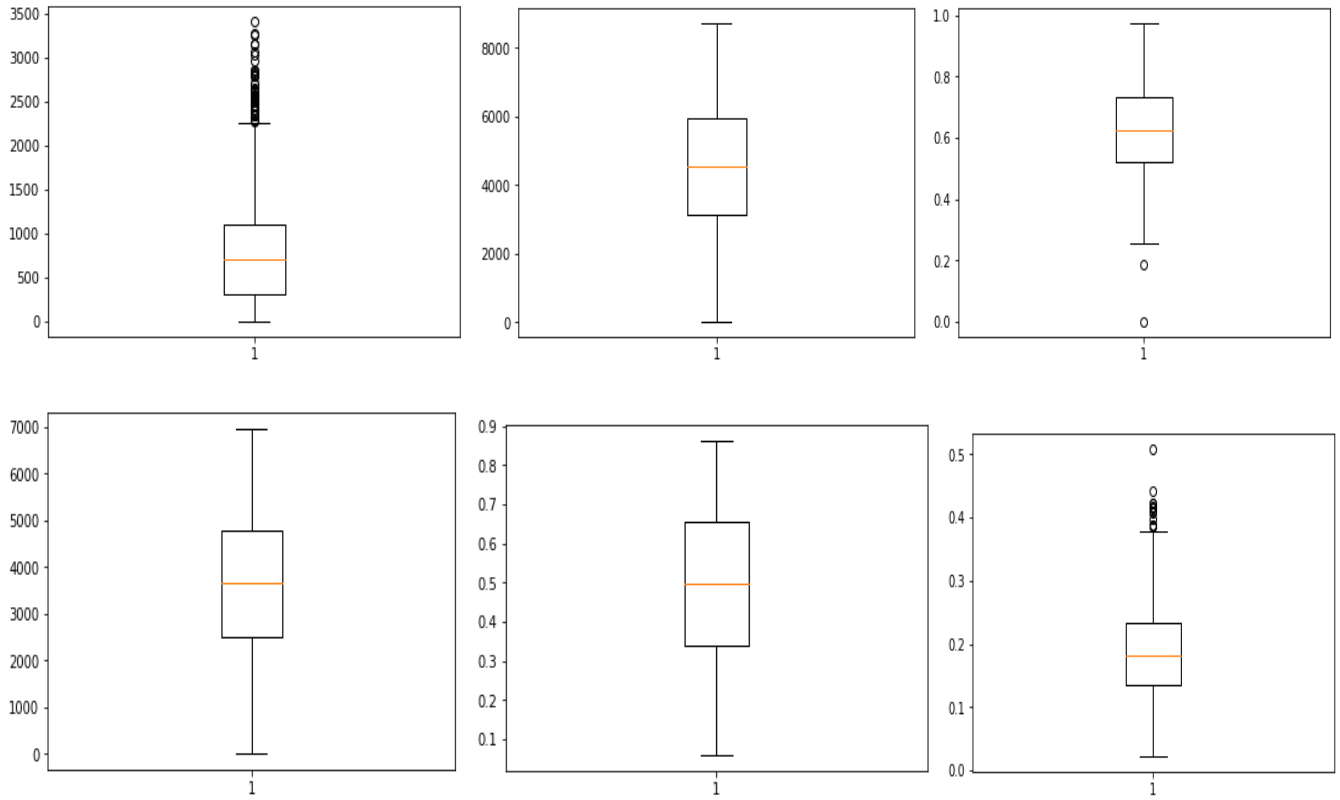

**registered vs mnth(year:2012)**

From the above figures, we can easily imply that both casual and registered users are behaving identical with each variable except 'weekday'&'workingday' variable.

Registered users are seen to use bike more likely on weekdays while casual users most likely uses bikes on weekends.

Also it being seen that demand of bikes has been risen up with the time.

### 2.1.3 Outlier Analysis




There is not too much outliers present in the data, the figure with maximum number of outliers are the count of casual users, which is a dependent variable So, they are not treated as outliers. 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).

The variable 'hum' is having 2 outliers which may interest us to know more about data or maybe not. It will be decided after the feature engineering we will do.

The last figure you see is the windspeed with quite amount of outliers.we will look further into them on feature engineering.

While, other features are clearly free of outliers.

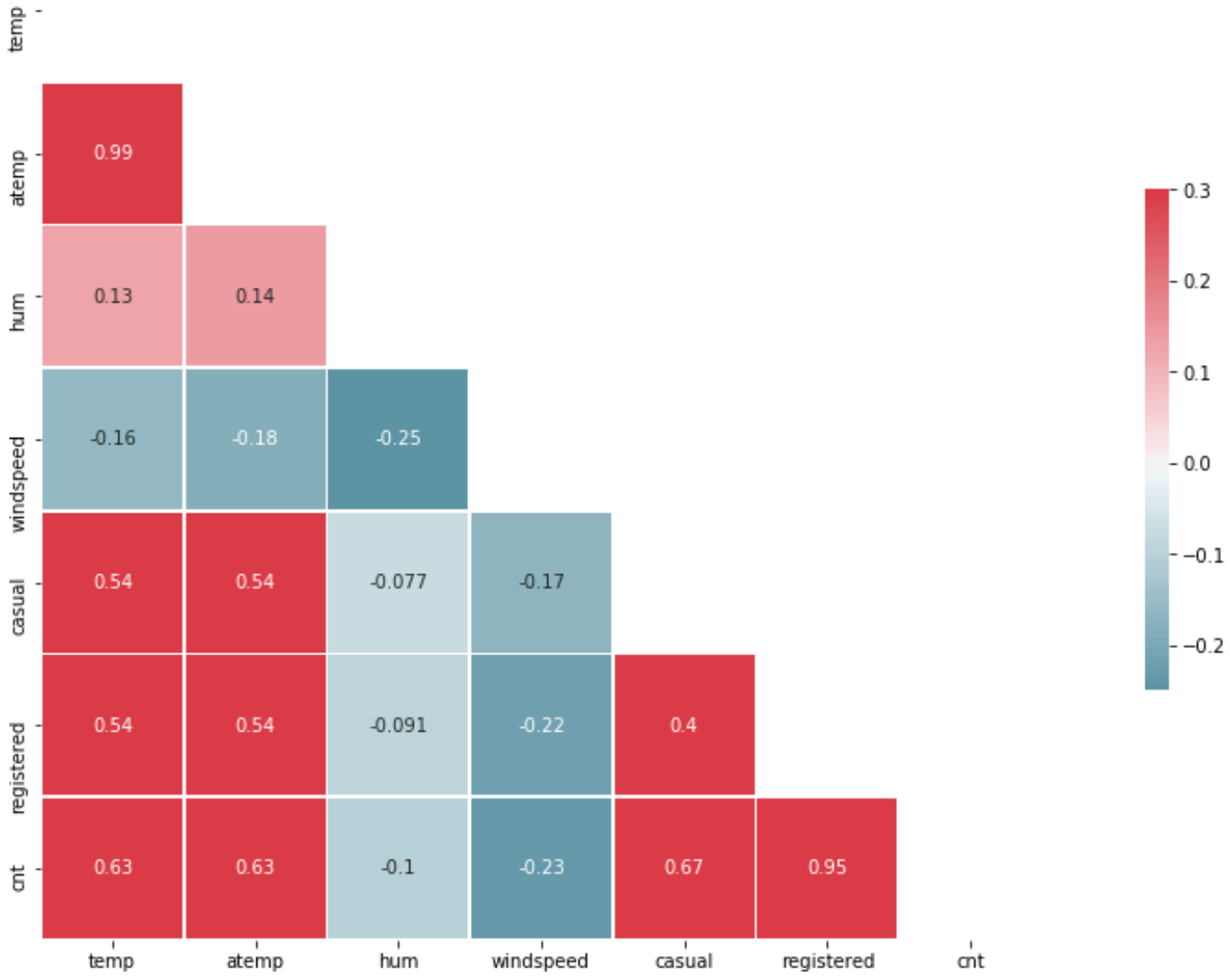
### 2.1.4 Feature Selection

We have got total of 16 variables out of which 3 are dependent('casual','registered','cnt') and rest are independent.

For the model be efficient, we need the independent variables to less correlated with each other, whereas independent variables to be highly correlated with the dependent variable(s).

We will find correlation between following continuous variables:

['temp','atemp','hum','windspeed','casual','registered','cnt']




Here, ‘atemp’ and ‘temp’ are having correlation of 0.99 which is totally not acceptable

‘hum’ is having very less correlation with the dependent variables

So, ‘atemp’ and ‘hum’ will provide more complexity to the model and may overfit the model. So, we will drop ‘hum’ and ‘atemp’.

Further, ‘dteday’ also not required for the model as those information already carried out by ‘yr’ ,‘mnth’ and ‘weekday’

‘instant’ will also be dropped.

Now, we will be left with 9 independent variables.

## 2.2 Modeling

### 2.2.1 Model Selection

As Casual and Registered users show almost identical behaviours except few changes. We will use same model for both.

The dependent variable can fall in either of the four categories:

1. Nominal

2. Ordinal

3. Interval

4. Ratio

In our case number of users, is Ratio the only predictive analysis that we perform is Regression.

We will go from simplest to complex.

The dependent variable in our model is a continuous variable i.e., Count of bike rentals. Hence the models that we choose are Linear Regression, Decision Tree and Random Forest. The error metric chosen for the problem statement is Mean Absolute Percentage Error (MAPE).

### 2.2.2 Multiple Linear Regression

Multiple linear regression is the most common form of linear regression analysis. As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables. The independent variables can be continuous or categorical.

Splitting Data into 70%-30%(train-test), Multiple Linear Regression model fits very well in the data with an **R^2 value of 0.819 and MAPE of 15.89%.** This model is fitting the data pretty well and considered quite impressive.

### 2.2.3 Decision Tree Regressor:

Decision Tree Regressor is a Regression algorithm which forms tree structure to decide the outcome/prediction of the Input.

Applying Decision Tree Regressor on bike rental data, it also performed very well**. MAPE value of 16.01% and R^2 value of 0.98**. This model is more efficient in explaining the bike rental count.

### 2.2.4 Random Forest:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

We have used Random Forest Regressor here and it has been the best performer over all as compared to Linear Regression and Decision Tree with **MAPE value of 12.56% and R^2 value of 0.97**. This has been the most accurate model of all.

# Chapter 4: Conclusion

Now that we have a few models for predicting the target variable, we need to decide which one to choose. There are several criteria that exist for evaluating and comparing models. We can compare the models using any of the following criteria:

1. Predictive Performance
2. Interpretability
3. Computational Efficiency

In our case of Bike count prediction Data, Interpretability and Computation Efficiency, do not hold much significance. Therefore, we will use Predictive performance as the criteria to compare and evaluate models. Predictive performance can be measured by comparing Predictions of the models with real values of the target variables, and calculating some average error measure.

### 4.1 Mean Absolute Percentage Error (MAPE)

MAPE is one of the error measures used to calculate the predictive performance of the model. We will apply this measure to our models that we have generated in the previous section.

**Linear Regression Model: MAPE = 15.89%**

**Decision Tree: MAE = 16.01%**

**Random Forest: MAE = 12.56%**

based on the above error metrics, Random Forest is the better model for our analysis. Hence Random Forest is chosen as the model for prediction of bike rental count.