Bike rental prediction

By Sujay Pawar

Completed on: 24th July 2019

Index

|  |  |
| --- | --- |
| **S.No** | **Description** |
| 1. | Problem Statement |
| 2. | Data analysis |
| 3. | Pre-processing |
| 4. | Feature selection |
| 5. | Feature scaling |
| 6. | Model development |

1. Problem statement

The objective of this project case is to prediction of bike rental count on daily based on the environmental and seasonal settings. We have provided with 15 columns and 731 observations which will help us to build a machine learning environment to predict future values if the same case scenario is faced. Data has been bifurcated in two years 2011 and 2012 and rest are the sub-components or supporting variables which helps us determine the outcome for given data set.

1.2 Data overview

Our motive revolves around preparing the best model to predict the future estimates of bike will be going on rent as per different weather conditions, temperature, seasons, holiday, working day etcetera. The mentioned factor does influence on the count of the bike to be rented as less number can be seen if we have rain or the wind speed is maximum and more to be rented if weather is pleasant and not too harsh.

We have total of 14 predictor columns which determine the outcome of count.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| instant | dteday | season | yr | mnth | holiday | weekday | workingday | weathersit |
| 1 | 01-01-11 | 1 | 0 | 1 | 0 | 6 | 0 | 2 |
| 2 | 02-01-11 | 1 | 0 | 1 | 0 | 0 | 0 | 2 |
| 3 | 03-01-11 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |
| 4 | 04-01-11 | 1 | 0 | 1 | 0 | 2 | 1 | 1 |
| 5 | 05-01-11 | 1 | 0 | 1 | 0 | 3 | 1 | 1 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| temp | atemp | hum | windspeed | casual | registered | **cnt** |
| 0.344167 | 0.363625 | 0.805833 | 0.160446 | 331 | 654 | **985** |
| 0.363478 | 0.353739 | 0.696087 | 0.248539 | 131 | 670 | **801** |
| 0.196364 | 0.189405 | 0.437273 | 0.248309 | 120 | 1229 | **1349** |
| 0.2 | 0.212122 | 0.590435 | 0.160296 | 108 | 1454 | **1562** |
| 0.226957 | 0.22927 | 0.436957 | 0.1869 | 82 | 1518 | **1600** |
| 0.204348 | 0.233209 | 0.518261 | 0.0895652 | 88 | 1518 | **1606** |
| 0.196522 | 0.208839 | 0.498696 | 0.168726 | 148 | 1362 | **1510** |
| 0.165 | 0.162254 | 0.535833 | 0.266804 | 68 | 891 | **959** |
| 0.138333 | 0.116175 | 0.434167 | 0.36195 | 54 | 768 | **822** |
| 0.150833 | 0.150888 | 0.482917 | 0.223267 | 41 | 1280 | **1321** |

2. Data Analysis

Understanding of the data helps us to make the most out of every variables to predict the outcome which in turns helps to manage a good accuracy out of our developed models. The more rigorous the detection of hidden patterns while analyzing the data the more efficient prediction done by the models. So analyzing the data plays an important role in overall effectiveness of our project and makes it’s a stepping stone to start our journey to get most out of the provided dataset.

3. Pre-processing

Before we starting predicting anything on the given datasets by applying our machine learning models we have to deep understand the data if all or only some of the data is useful which is also applicable on variables as not all provided variables always help to best predicting the outcome, as two different variables sometimes provide similar basis of information making the data redundant.

3.1 Missing values analysis

It can happen while collecting the data or feeding the data some fields are missed our or left blank intentionally as the observer do not have any answer or he has been provided with multiple choices to choose from. Such blank fields are missing values which are denoted as NA or nan in R and python respectively. This missing values will throw an error while computing any of our machine learning techniques so should be computed or deleted altogether.

If we need to fill the missing values with our own values for analysis we can use multiple computing solutions such as we can replace the na values with mean, median, mode or knn imputation. Also, we our particular column has more than 30% of the data to be missing, we can safely delete the column from our further analysis.

Fortunately, our dataset does not contain such missing values so we do not need to action anything further on this.

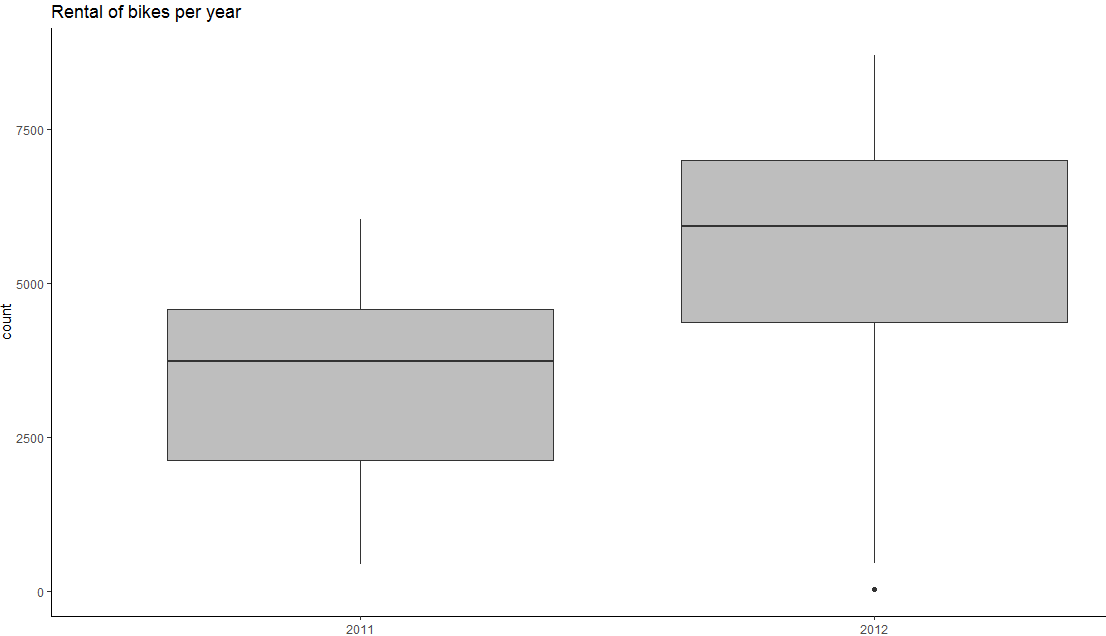
3.2 Outlier analysis

In simple words, an outlier is a field which is very different from rest of the data present in the column. To explain an outlier further we can take an example of stars in the sky. What is the probability of a shooting star amongst all the normal stars, can be one in a million, which makes the shooting star an outlier from rest of the stars or in our case rest of the column data. To give another example, in a survey people were given 4 choices with fourth been as ‘other’ and once other is selected the surveyor can input his own answer. So among 1000 of surveyors only 2 or 3 selected the other as option which makes their answers as an outliers with respect to rest of the answers. However, it not safe to treat every outliers as un-useful data as they can even be valid sometimes while in other cases an outlier can be completely unrelated to rest of the data.

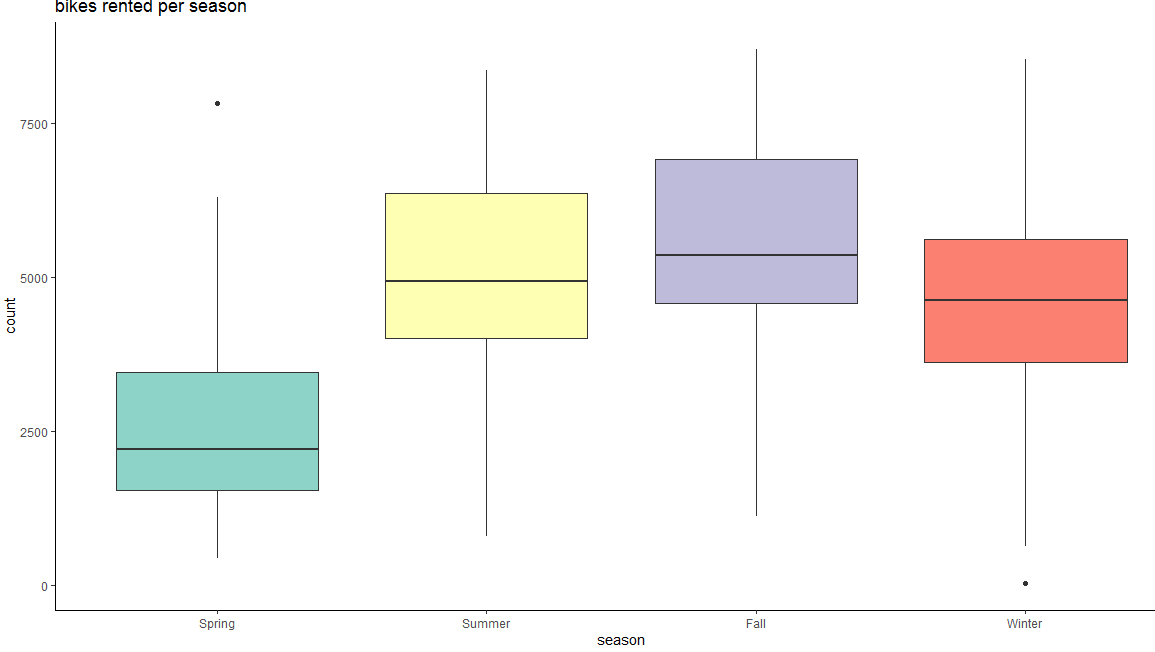
In the dataset provided, we came across 2 and 13 such cases from wind speed and humidity which has been labeled as outlier from our analysis.

Under EDA, we also visualize the data and plot the graphs which provides us with the relation between various columns and count of the bike was rented in previous instances. This step provides a proper insights which we need to have when proceeding further for building various models and understanding relations between the variables.

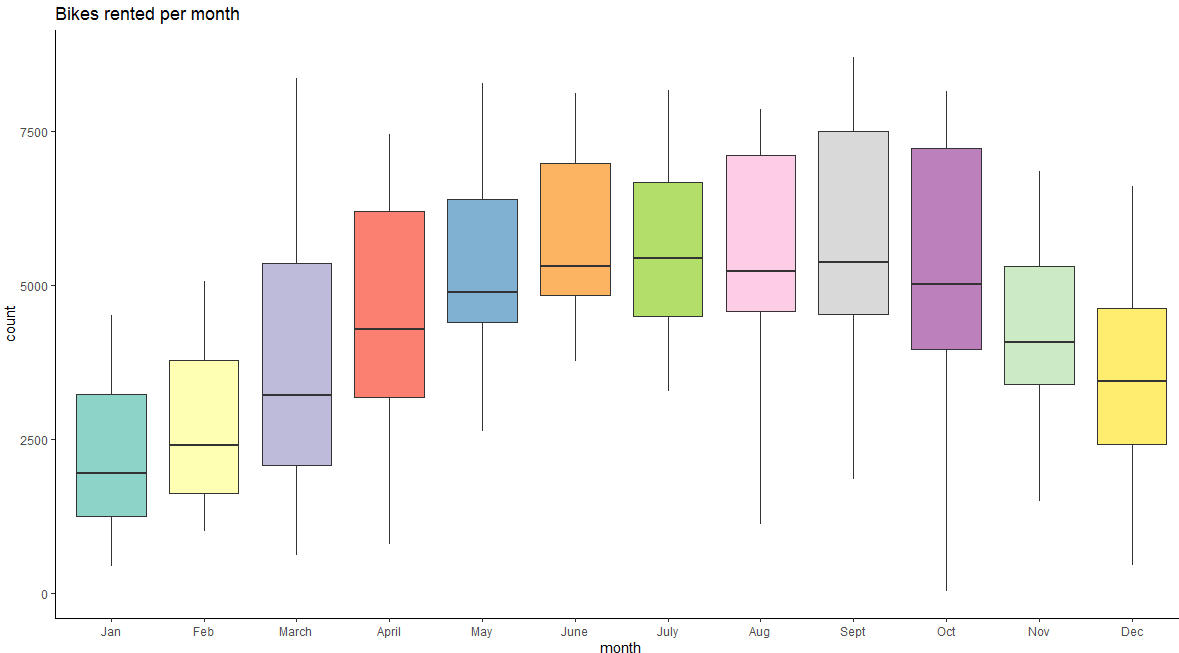
1. Bikes rented in year 2011 and 2012

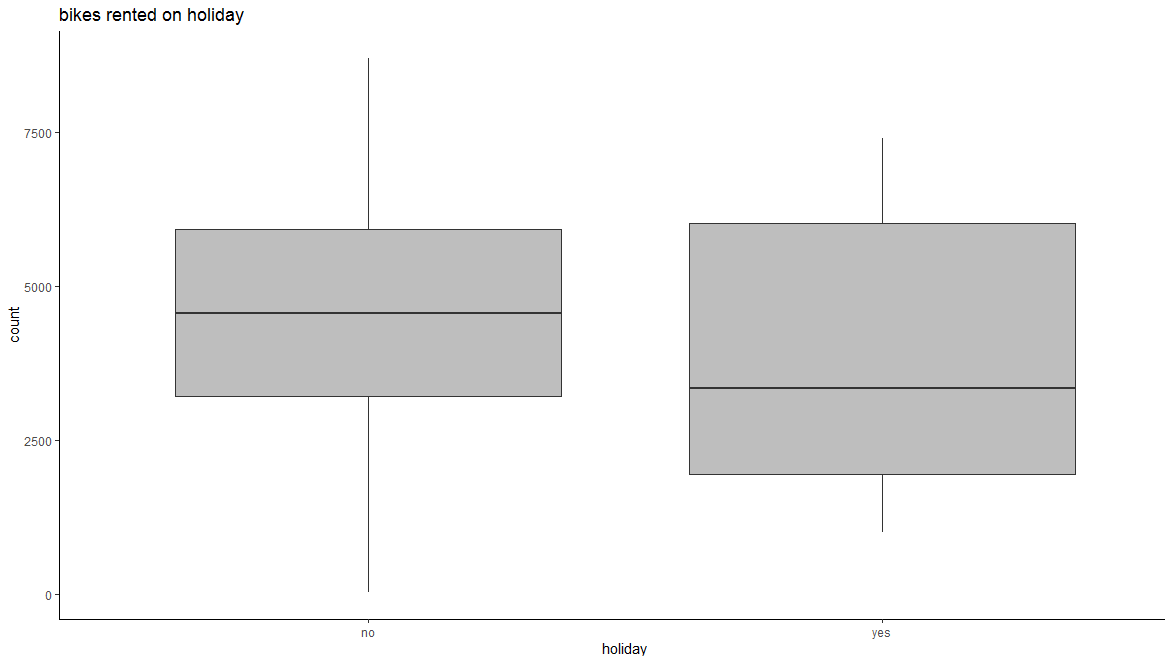


2. Bikes rented as per season

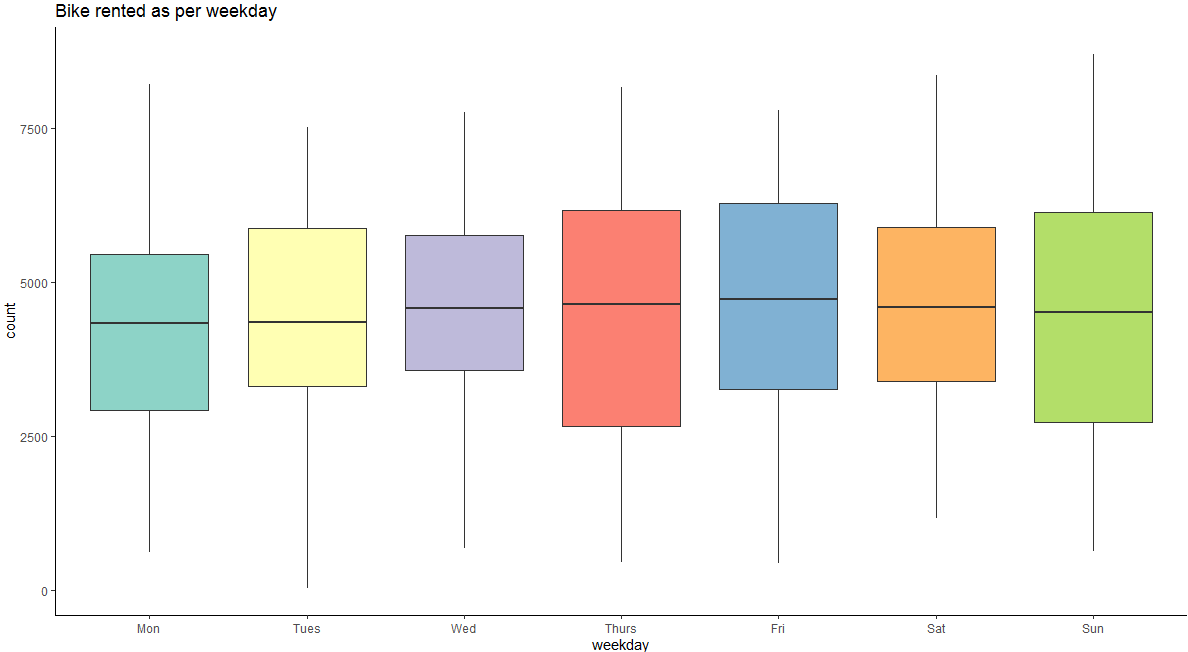


3. Count of bikes rented per month

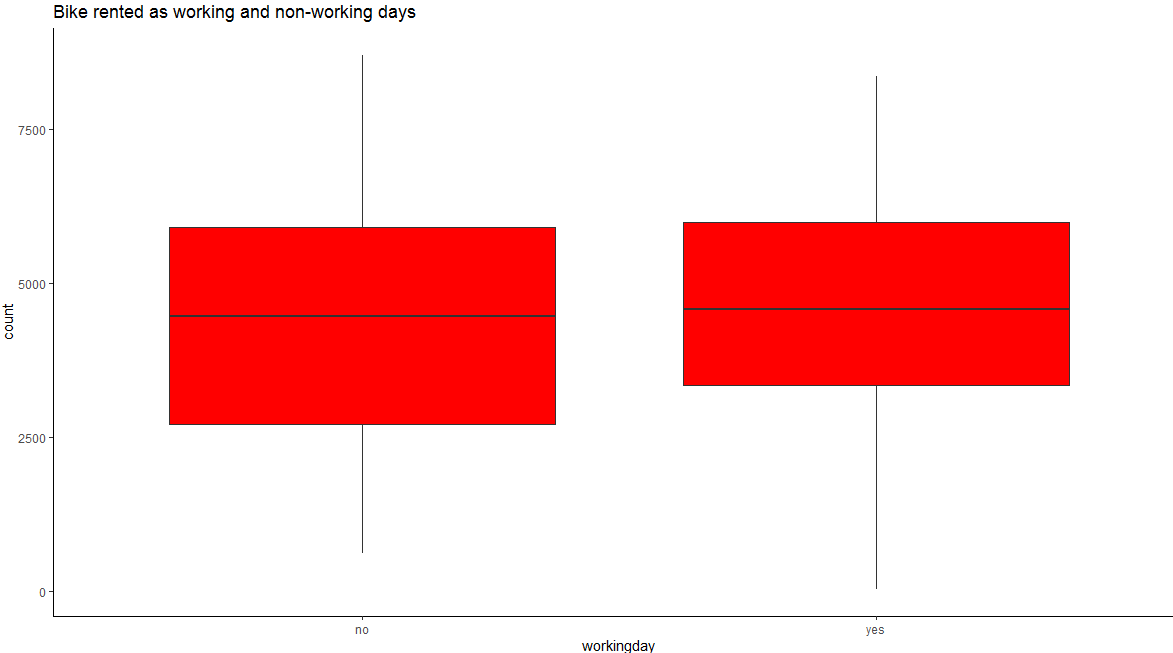


4. Count of bikes were rented on holiday

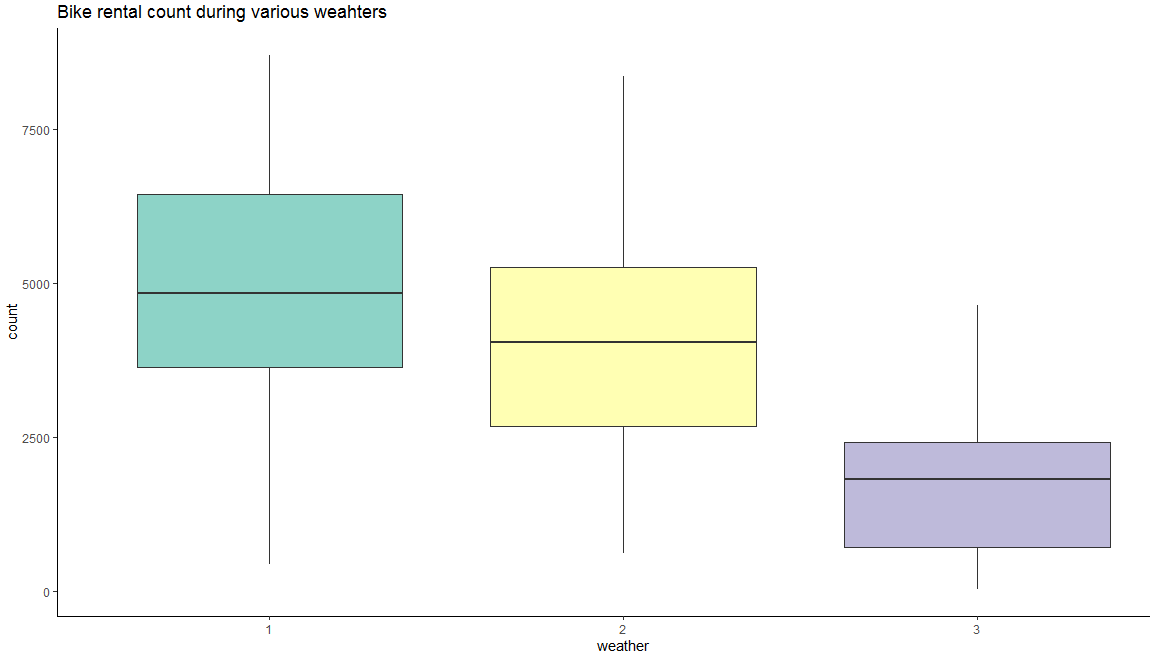
5. Count of bikes rented every day from the week



6. Count of bikes rented on working vs non-working days



7. Count of bikes rented during various weathers



4. Feature selection

This plays an important part when we develop our models as while providing the dataset it will be a very less chance of checking for redundancies, duplicate values, missing values and any unforeseen entries. Not every variable plays a key role to detect the outcome. We can remove unwanted variables, in short variables whose correlation is 1 or close to 1 are considered to be carrying same information so we can only subset only one from both.

In our case too we have two variables which has more or less same information, which were temp and atemp. So for our model to be optimal we can choose either of the variable.



5. Feature scaling

While executing our machine learning models it is a good practice to bring all the numerical predictor variable to a certain range which matches to the range of other numerical variables. This would help the model to develop a better working model and all the variable can be imputed based on equal scale.

As, we already had the numerical variables standardized from the raw file this step was not needed in our project. Casual and Registered were in normal numeric form but as addition of both the variable gives us our dependent variable we will not be using in our model development.

Also, to be noted that majority of the bikes were rented by Registered users than casual.

6. Model development

All the above steps helps to get a model which can be then be applied on future cases to predict an outcome. Hence, all the execution shall be performed at utmost interest of the model. Data processing plays a vital role in determining the output result of the machine learning algorithms and helps us to predict not exact but approximate values on future data once fed to selected model.

Our dependent variable was an continuous variable so we have to apply our regression techniques to predict the outcome.

Models used to develop our cases are as per below

1. Decision tree (Python and R)

2. Random forest (Python and R)

3. Ridge regression (Python and R)

4. Lasso regression (Python and R)

5. Elastic net regression (Python)

Note:

1. The above project has been made solely by me and no individual help was taken on any of the step

2. Lot of information was accessed from google websites while developing the models and understanding other aspects of project.