Prediction of Bike Rental

14-Sep-2019

Abhishek singh pushkar

Contents

Introduction

Problem Statement

The objective of this Case is to Predication of bike rental count on daily based on the environmental and seasonal settings. Using statistics we need to read the data and develop a machine-learning algorithm to predict the number of bikes hired based according to the given condition.

Data

The data given has 731 rows and 16 columns which includes

instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit
1	2011-01-01	1	0	1	0	6	0	2
2	2011-01-02	1	0	1	0	0	0	2
3	2011-01-03	1	0	1	0	1	1	1
4	2011-01-04	1	0	1	0	2	1	1
5	2011-01-05	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
0.200000	0.212122	0.590435	0.160296	108	1454	1562
0.226957	0.229270	0.436957	0.186900	82	1518	1600

Here our target is the cnt variable which is the count of no. of bikes hired. We can see from the above data that cnt variable is the sum of casual and registered variables. The casual represents casual no. of users who are random travellers and registered represents the no. of users who are registered by the company as bike users.

Methodology

Data preparation

We need to change the numerical value to the categorical value of the variable
season: Season
- 1: Spring
- 2: Summer
- 3: Fall
- 4: Winter

yr: Year
- 0: 2011
- 1:2012

holiday:
- 0: Working Day
- 1: Holiday

Workingday:
- 0: Holiday
- 1: Working Day

weathersit:
- 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
- 4: Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog

And changing the variables to proper name to increase the readability of the data

season	: Season
yr	: Year
mnth	: Month
holiday	: Holiday
weekday	: Weekday
workingday	: Working Day
weathersit	: Weather Condition
temp	:Temperature
atemp	: Feeling Temperature
hum	: Humidity
windspeed	: Wind Speed
casual	: Casual Users
registered	: Registered Users
cnt	: Count

Exploratory data analysis

Distribution of categorical variable with the target variable

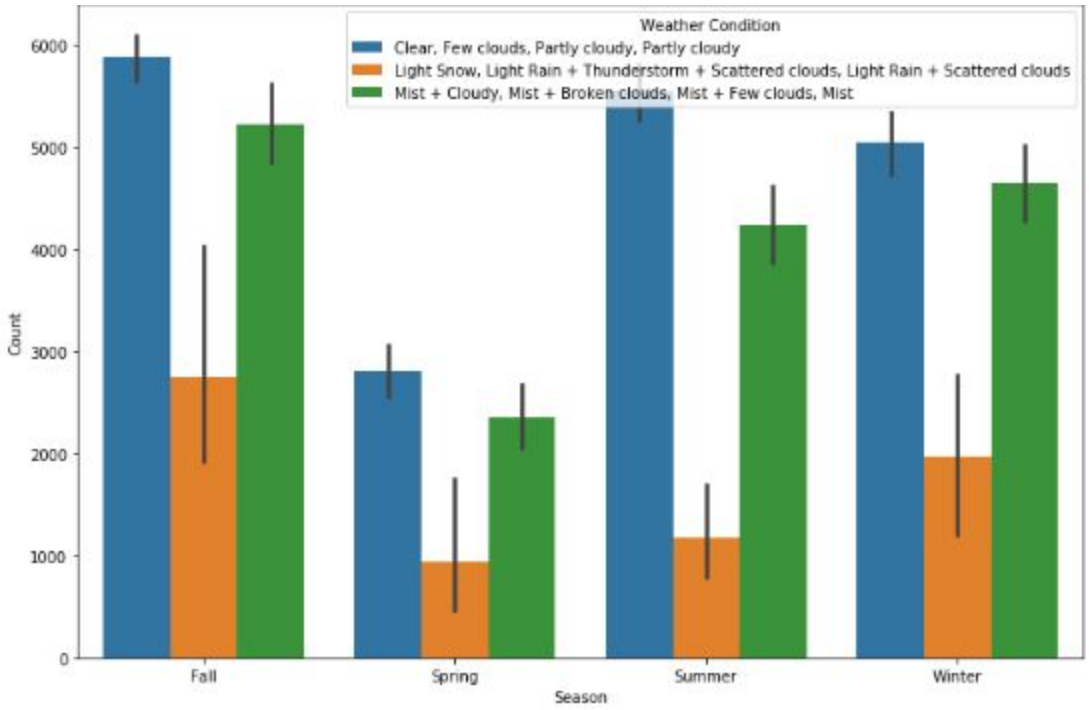


Fig. Bar graph showing the number of bikes (Count) hired season-wise based on weather condition.

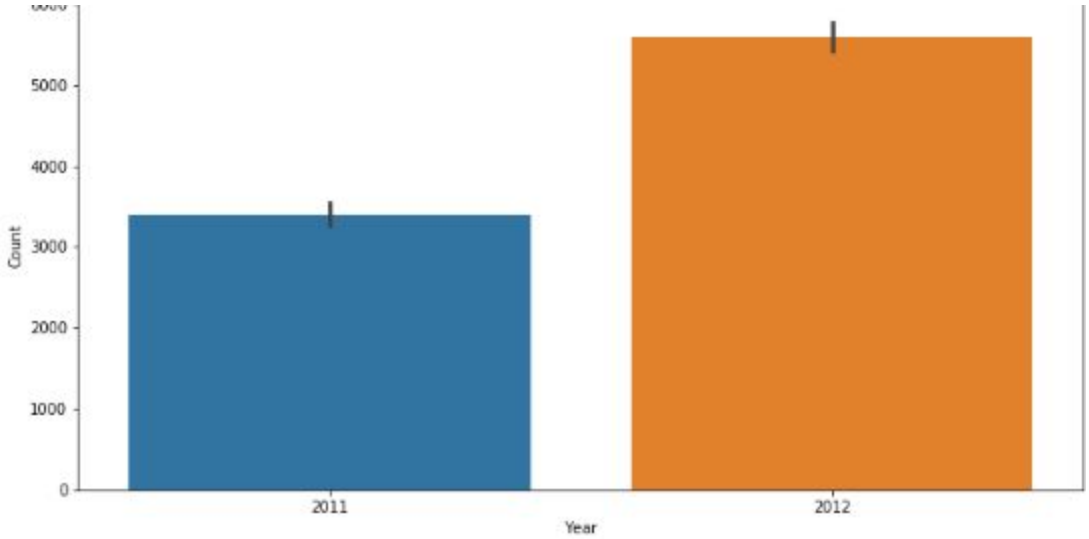


Fig. Bar graph showing the no. of bike hired yearly.

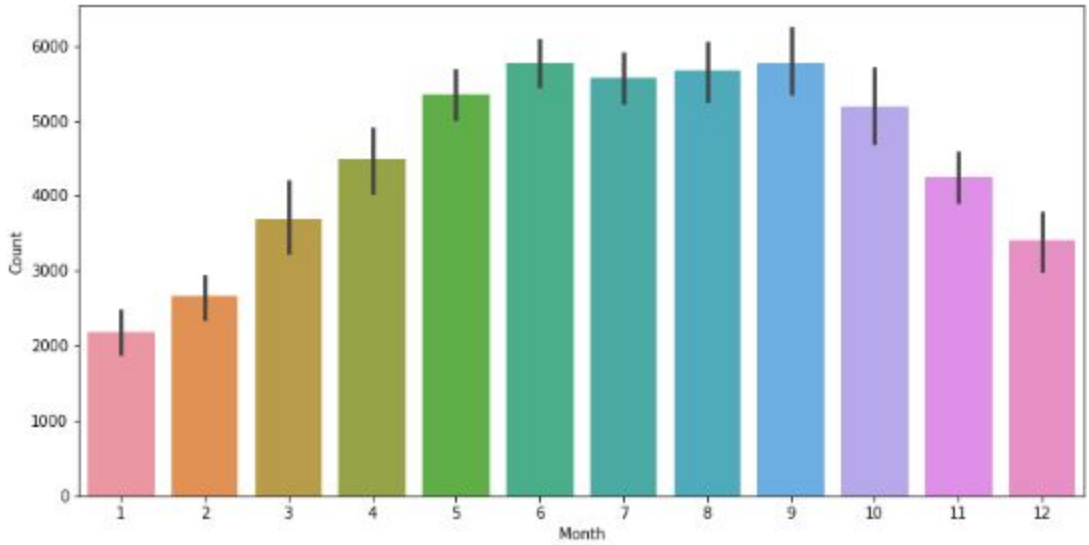


Fig. Bar graph showing no. of bikes hired month wise.

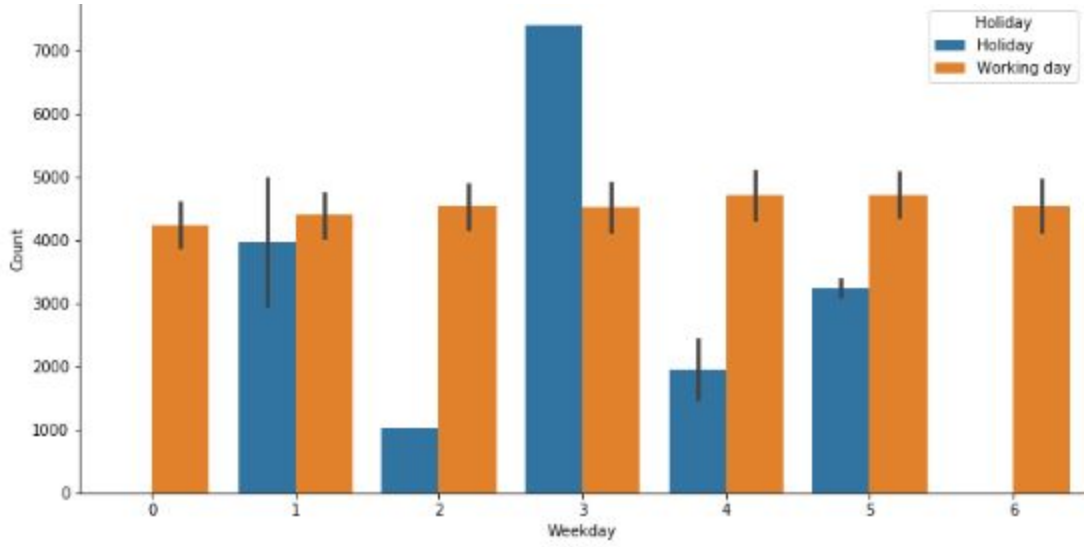
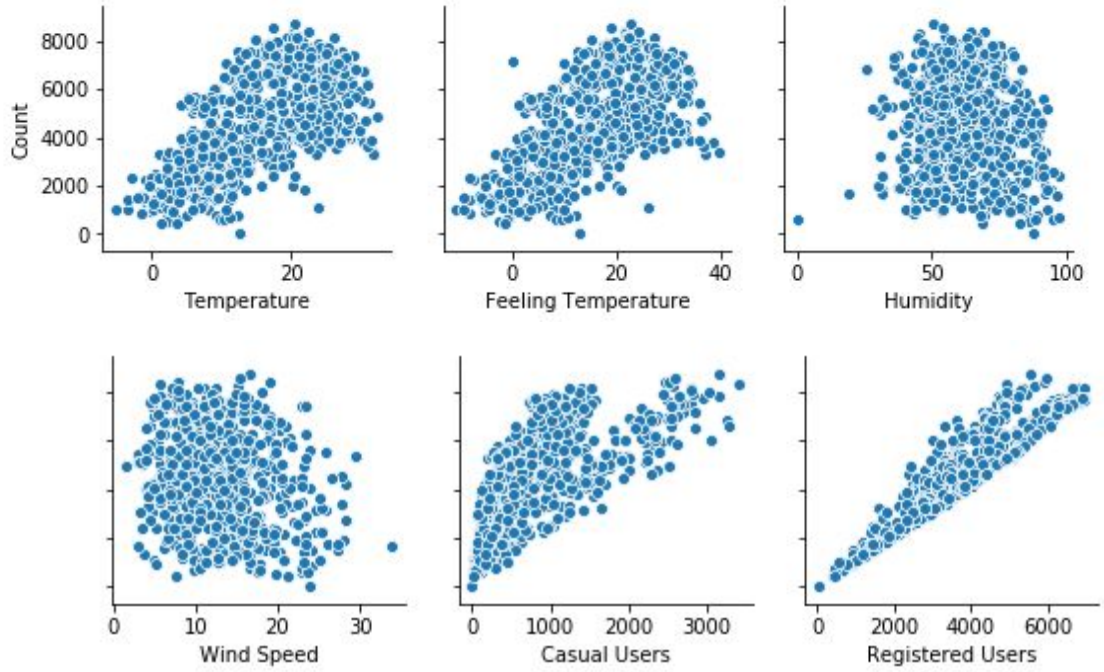


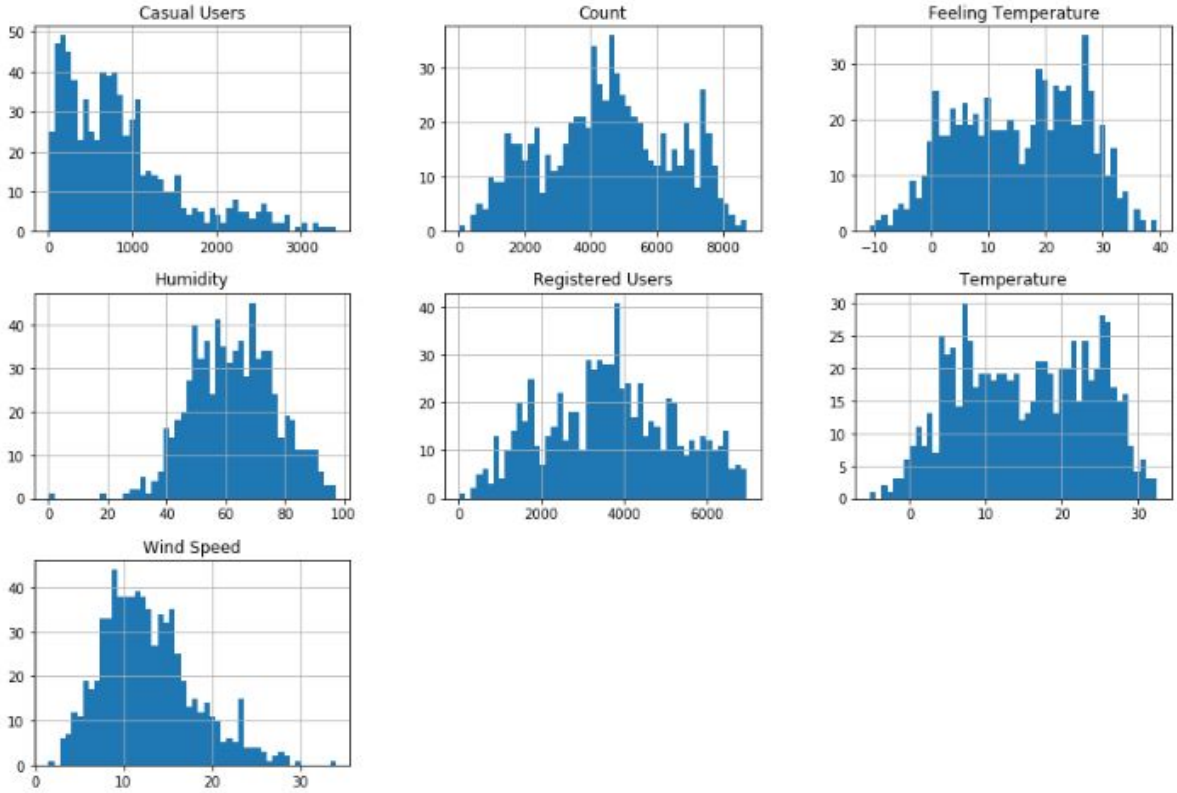
Fig. Bar graph showing no. of bikes hired weekly with separation based on weather the day was a holiday or working day.

It is observed from these graphs that most of the bikes were hired in the fall season followed by fall. Users were active when the sky was clear or with little rain. In 2012 there are more numbers of users, that is because of the increase in the number of registered users.

Distribution of Continuous variable with target variable



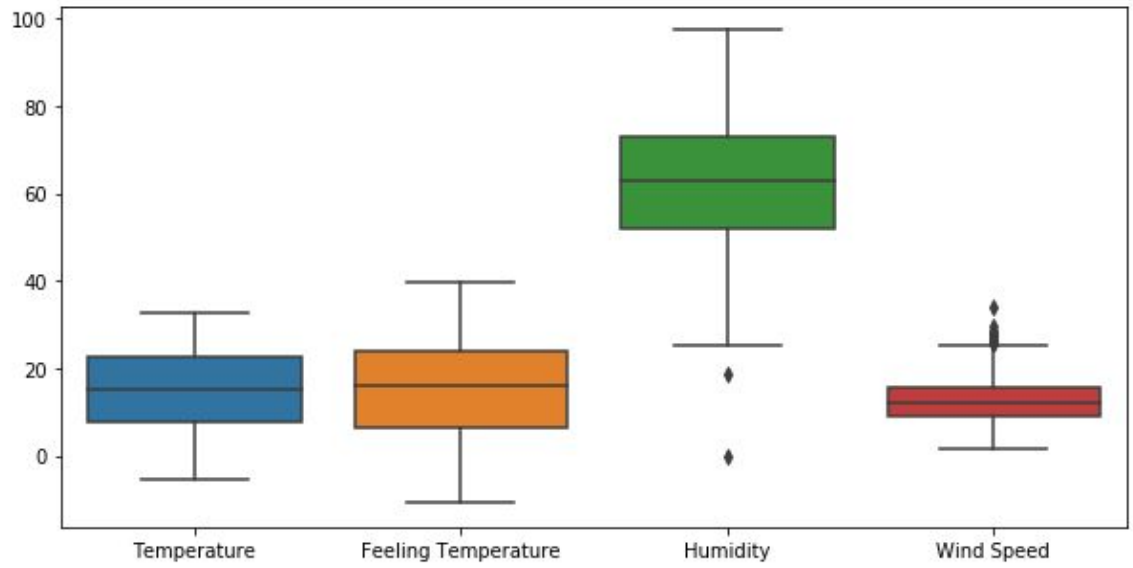
Distribution of continuous variables

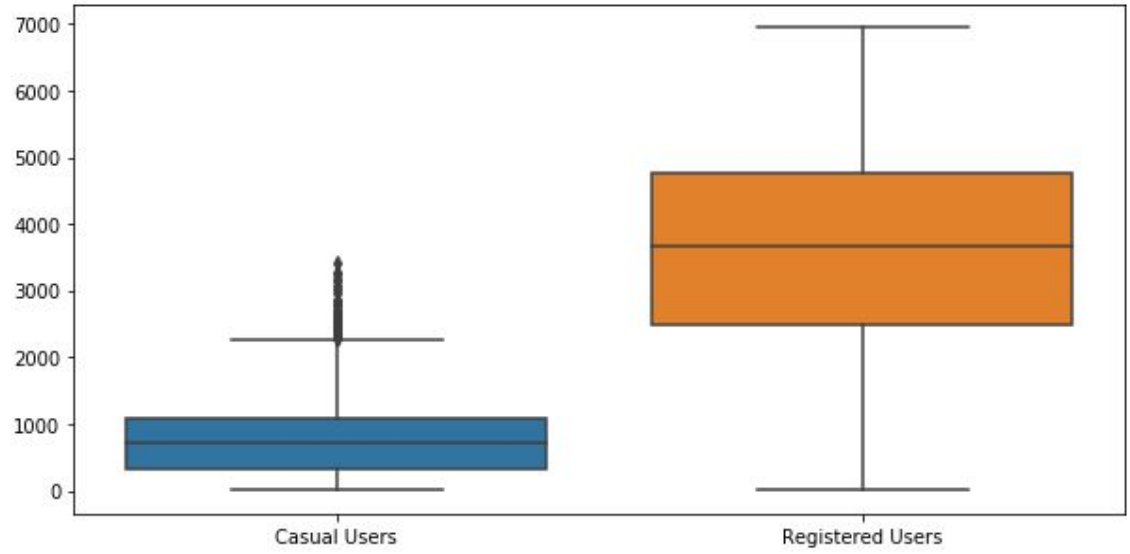


Missing Value Analysis

The data do not contain any missing values.

Outlier detection





Here we can see the outliers in Humidity, Wind Speed and Casual Users.

Outlier Removal

Outliers can be removed Inter Quarentile Range. IQR is calculated with min and max value of a variable. Any value outside the min and max is considered an outlier.

Feature Selection

Here we need to select the feature which is valuable for the prediction task. Any variable with high collinearity should be discarded. From the above image, we can see that Temperature and Feeling Temperature are highly correlated so dropping Feeling Temperature. Also, Registered users is highly correlated and casual users is of no use so dropping these two variables.

Sampling

Split the dataset into 80 per cent training data and 20 per cent test data.

Modeling

The target variable is continuous therefore the models should be regression type.

Tested with three algorithms:

1. Linear Regression
2. Decision Tree Regressor
3. Random Forest Regressor

Evaluation

For score evaluation below are calculated

1. Mean Absolute Error
2. Mean Squared Error
3. Mean Absolute Percentage Error
4. R squared value

And the value of these scores are:

1. Mean Absolute Error
    a. LR : 24.75
    b. DTR : 24.42
    c. RFR : 21.72

2. Mean Squared Error
    a. LR : 818.01
    b. DTR : 867.84
    c. RFR : 667.70

3. Mean Absolute Percentage Error
    a. LR : 22.05
    b. DTR : 23.09
    c. RFR : 20.00

4. R Squared value
    a. LR : 0.76
    b. DTR : 0.73
    c. RFR : 0.82

Conclusion

Calculation of the different scores tells us how often an algorithm can predict future cases. These are the definition of how these scores evaluate on the data provided.

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

MSE is a quadratic scoring rule that also measures the average magnitude of the error.

MAPE is a measure of prediction accuracy It usually expresses accuracy as a percentage and is defined by the formula.

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

So considering these different scores of evaluation and comparing them among different algorithms it concludes that Random forest is the best fit for the dataset.

Appendix A

Distribution of categorical variable with the target variable

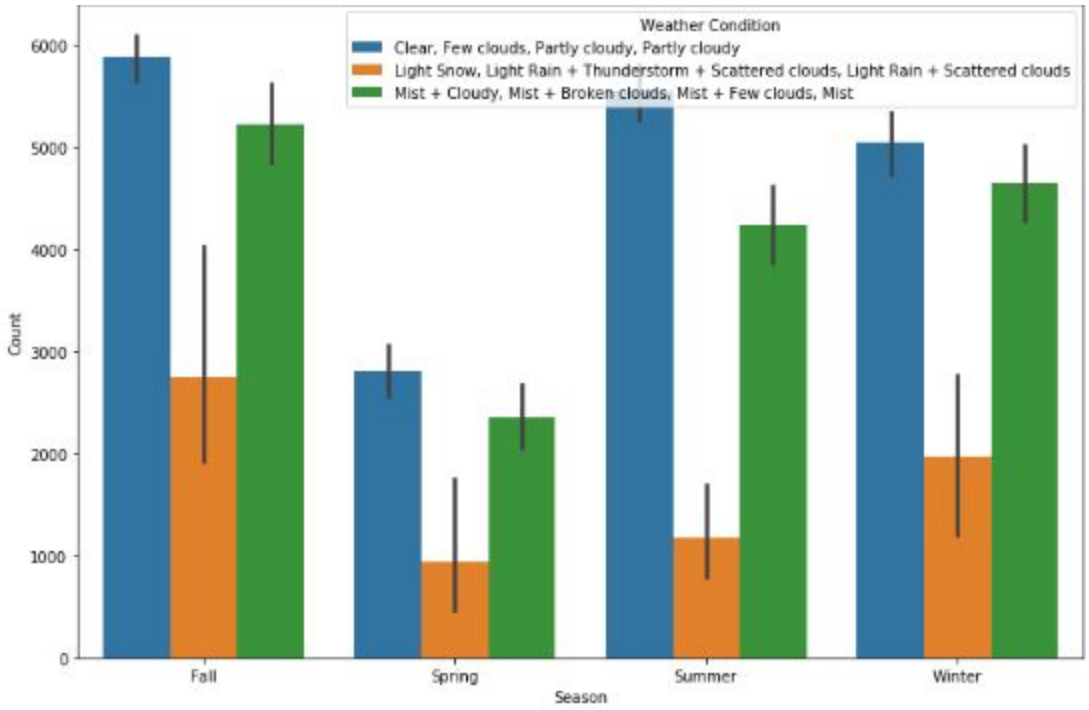


Fig. Bar graph showing the number of bikes (Count) hired season-wise based on weather condition.

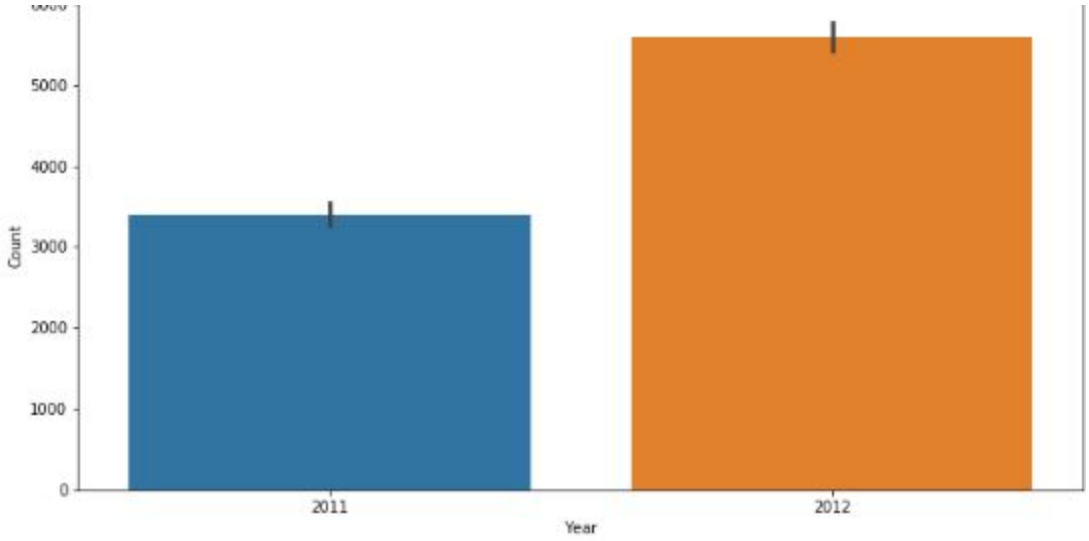


Fig. Bar graph showing the no. of bike hired yearly.

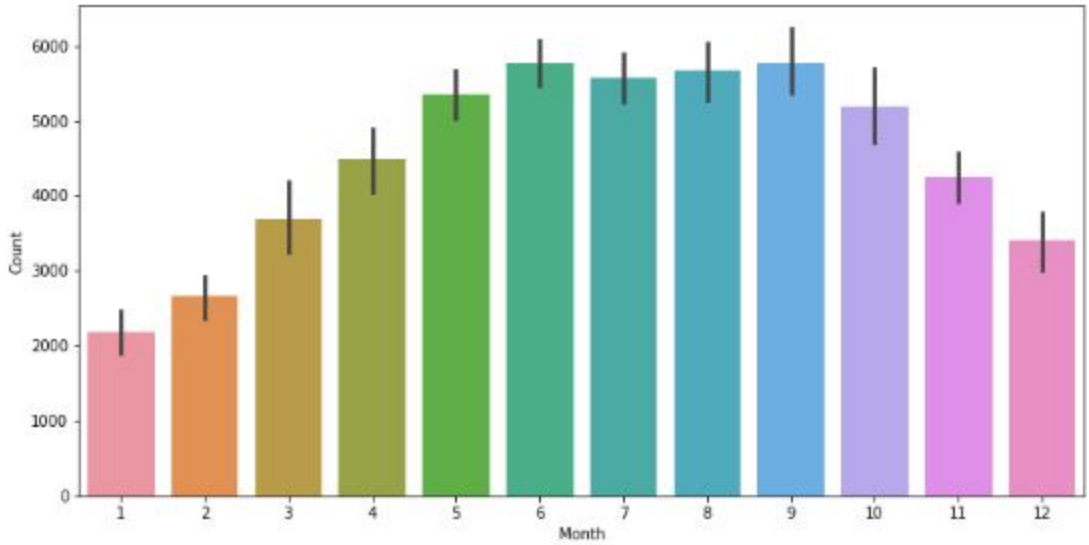


Fig. Bar graph showing no. of bikes hired month wise.

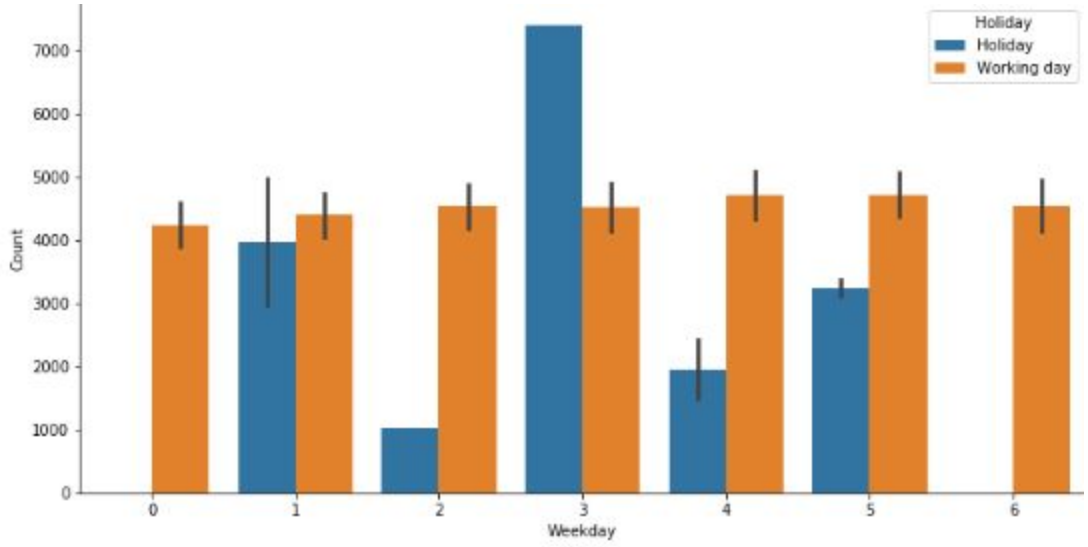
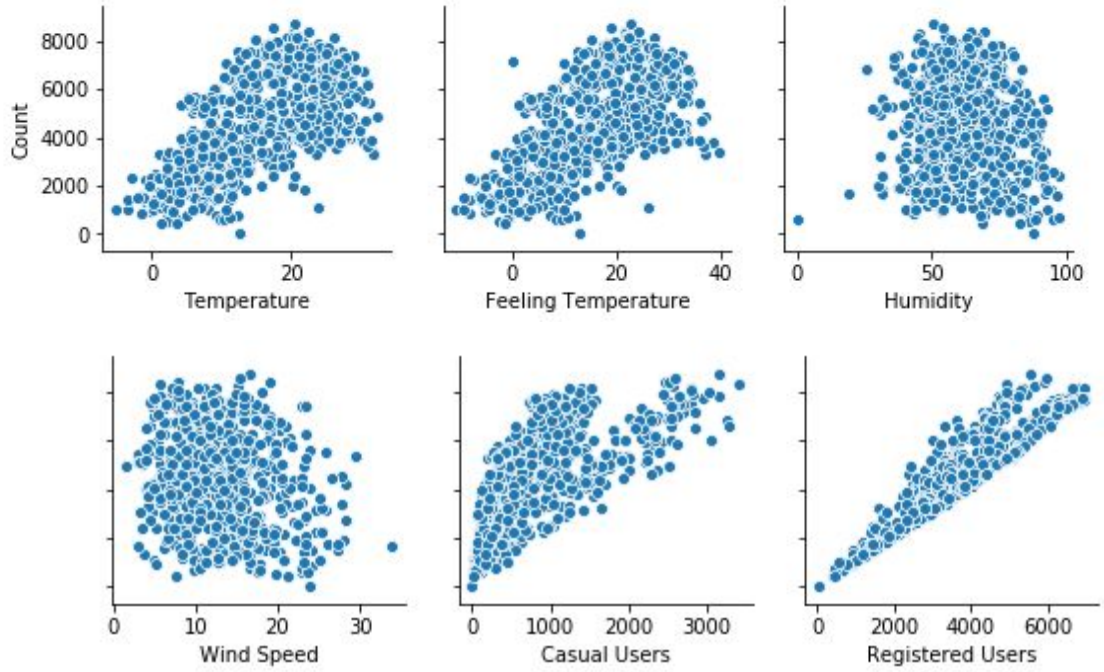
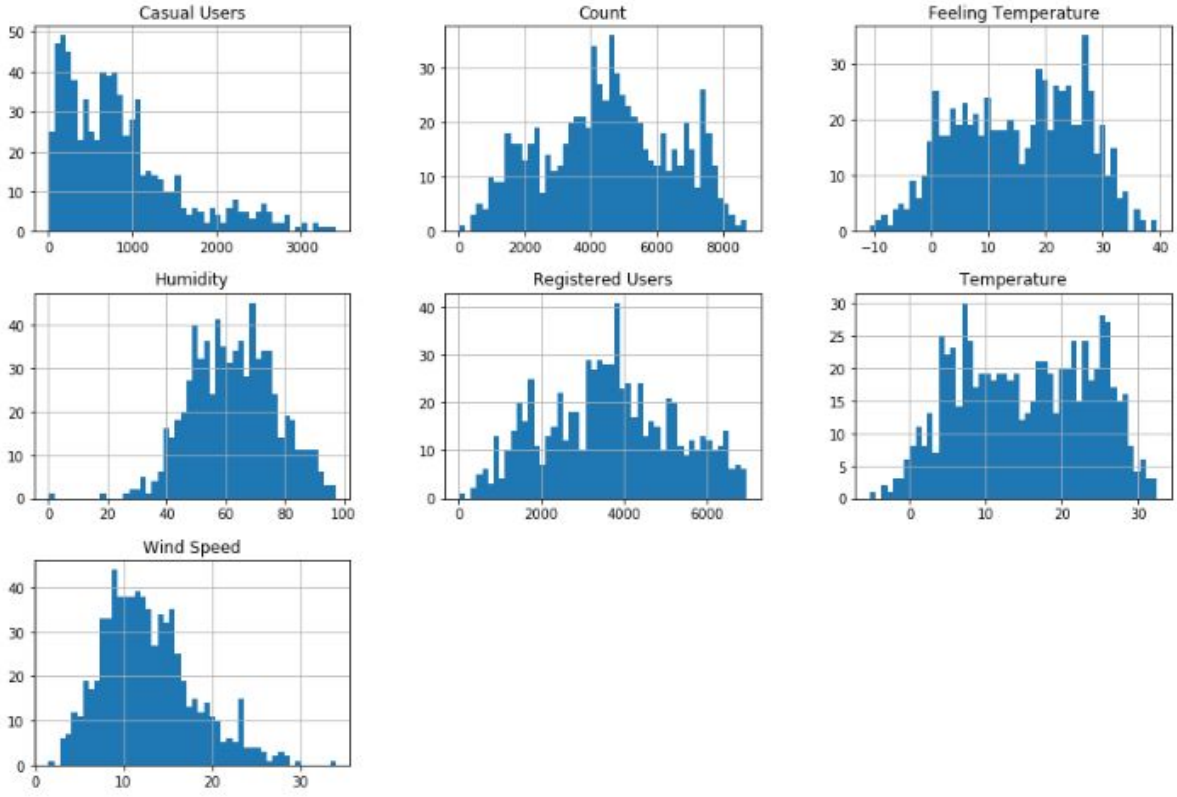


Fig. Bar graph showing no. of bikes hired weekly with separation based on weather the day was a holiday or working day.

Distribution of Continuous variable with target variable



Distribution of continuous variables



Outlier detection

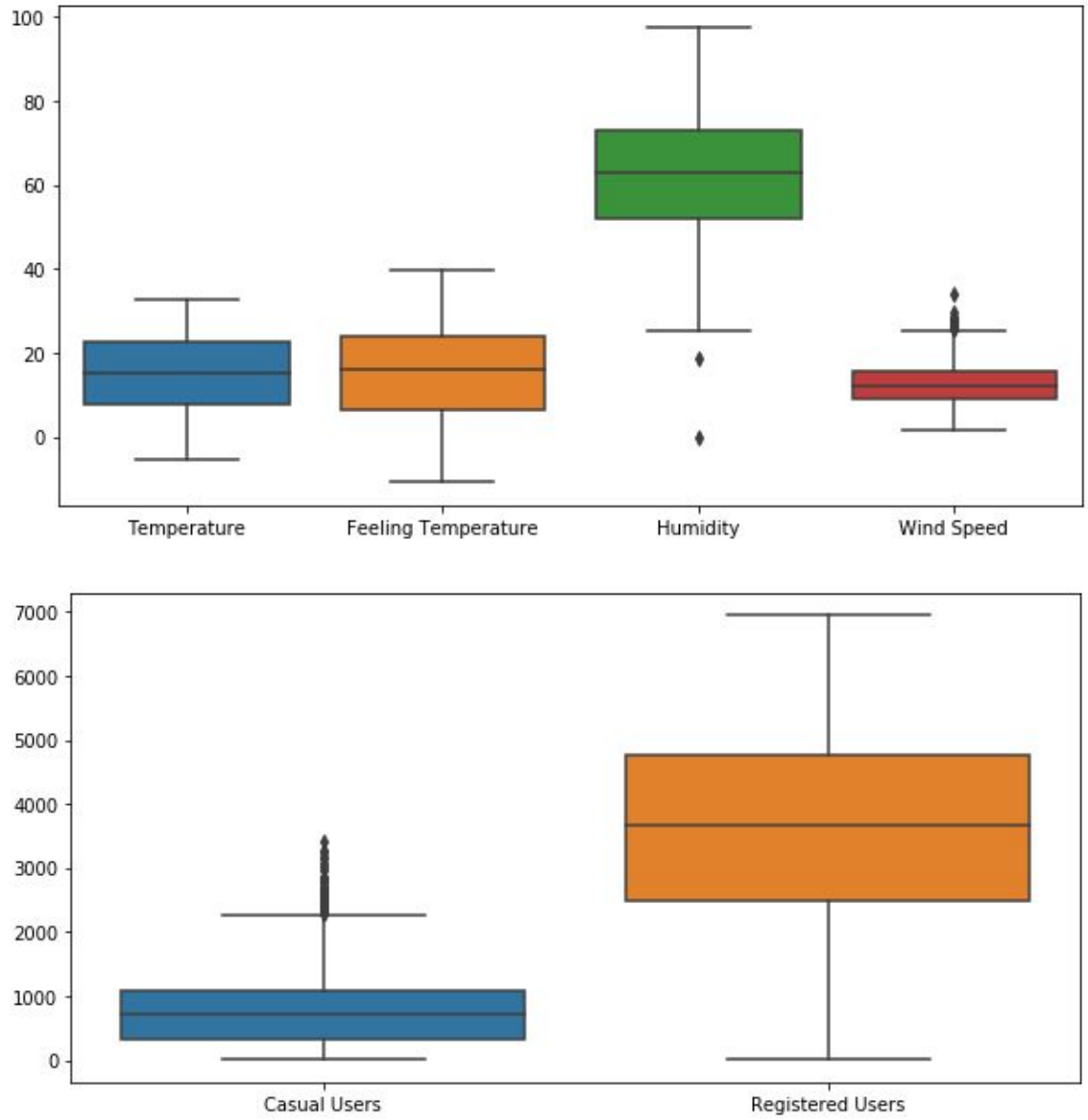


Fig. Boxplots for outlier detection

Feature Selection

Fig. Correlation matrix