Problem Statement

The objective of this Case is to Predication of bike rental count on daily based on the environmental and seasonal settings.

Bike rental data

	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	1	2011-01-01	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
1	2	2011-01-02	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
2	3	2011-01-03	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
3	4	2011-01-04	1	0	1	0	2	1	1	0.200000	0.212122	0.590435	0.160296	108	1454	1562
4	5	2011-01-05	1	0	1	0	3	1	1	0.226957	0.229270	0.436957	0.186900	82	1518	1600
5	6	2011-01-06	1	0	1	0	4	1	1	0.204348	0.233209	0.518261	0.089565	88	1518	1606
6	7	2011-01-07	1	0	1	0	5	1	2	0.196522	0.208839	0.498696	0.168726	148	1362	1510
7	8	2011-01-08	1	0	1	0	6	0	2	0.165000	0.162254	0.535833	0.266804	68	891	959
8	9	2011-01-09	1	0	1	0	0	0	1	0.138333	0.116175	0.434167	0.361950	54	768	822
9	10	2011-01-10	1	0	1	0	1	1	1	0.150833	0.150888	0.482917	0.223267	41	1280	1321

Method:

We want to extract the total number of people who rent a Bike daily based on Weather condition.

Exploratory Data Analysis - It includes following steps Looking into the data means visualizing the data through graphs and analyzing all variables.

- Visualization
- Missing value analysis
- Outlier analysis
- Correlation
- Feature scaling
- Dummy data create
- Feature sampling

Models:- Applied below models on preprocessed data

- Decision tree
- Random forest
- Linear regression

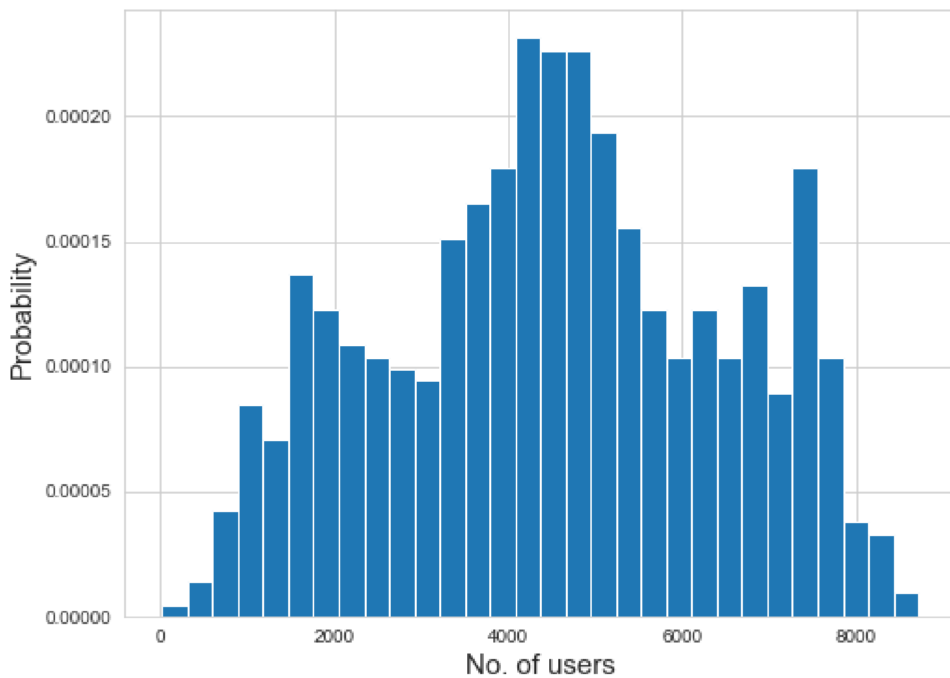
To see data information, datatype and number of observation data.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 731 entries, 0 to 730
Data columns (total 16 columns):
instant       731 non-null int64
dteday        731 non-null object
season        731 non-null int64
yr            731 non-null int64
mnth          731 non-null int64
holiday       731 non-null int64
weekday       731 non-null int64
workingday    731 non-null int64
weathersit    731 non-null int64
temp          731 non-null float64
atemp         731 non-null float64
hum           731 non-null float64
windspeed     731 non-null float64
casual        731 non-null int64
registered    731 non-null int64
cnt           731 non-null int64
dtypes: float64(4), int64(11), object(1)
memory usage: 88.6+ KB
```

There are variables of datatype float and int and date has object
Observations are 731 and variables are16
Check unique values

```
instant       731
dteday        731
season          4
yr              2
mnth           12
holiday         2
weekday         7
workingday      2
weathersit      3
temp          499
atemp         690
hum           595
windspeed     650
casual        606
registered    679
cnt           696
dtype: int64
```

Target variable is 'cnt'
Unique values of 'cnt' is 696



Checking for missing value

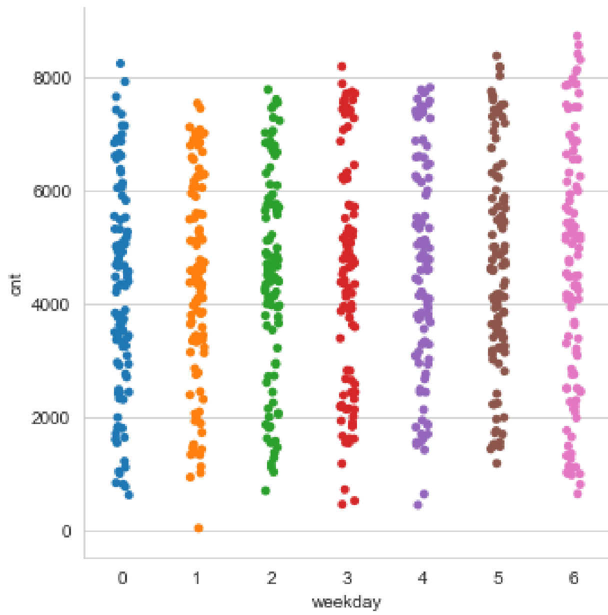
print(data.isnull().sum())

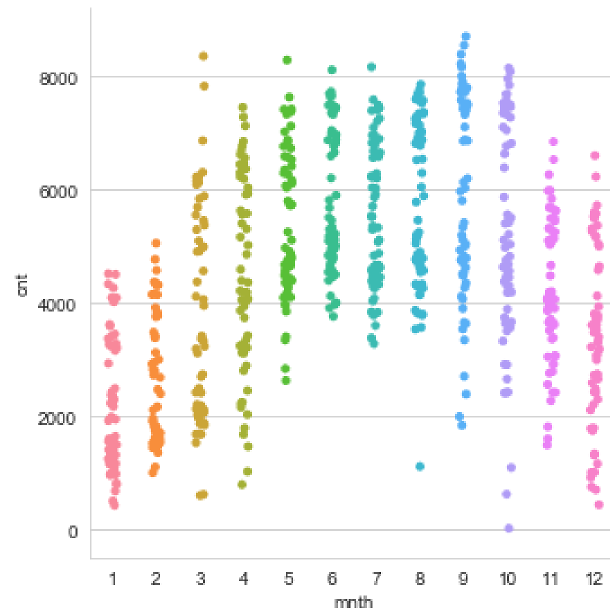
```
season         0
yr             0
mnth           0
holiday        0
weekday        0
workingday     0
weathersit     0
temp           0
atemp          0
hum            0
windspeed     13
cnt            0
dtype: int64
```

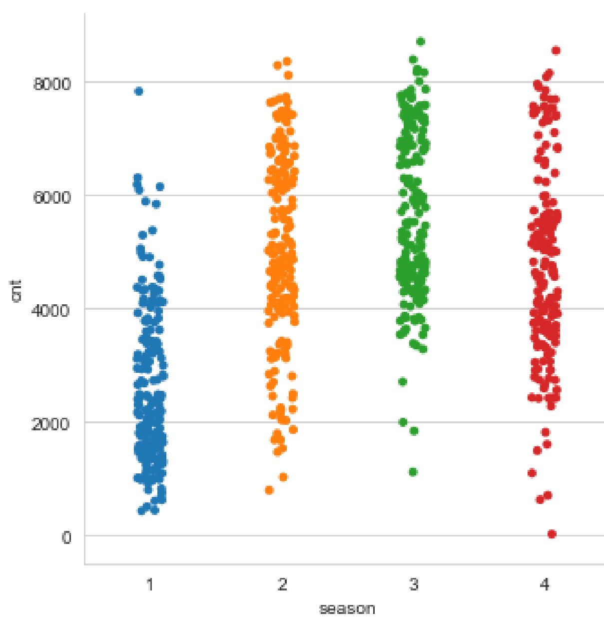
Windspeed variable has 13 missing values

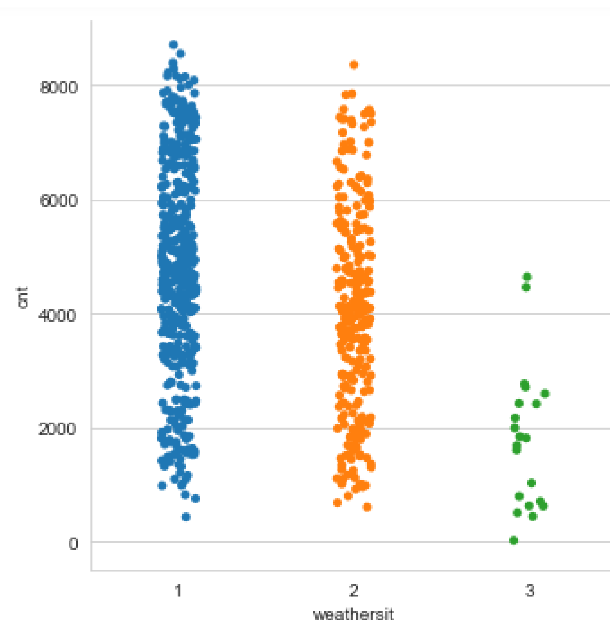
Data Understanding

Understanding the data set and to see how different features interact with each other and the target. First the amount of bike rental counts for each day of the week is analyzed.

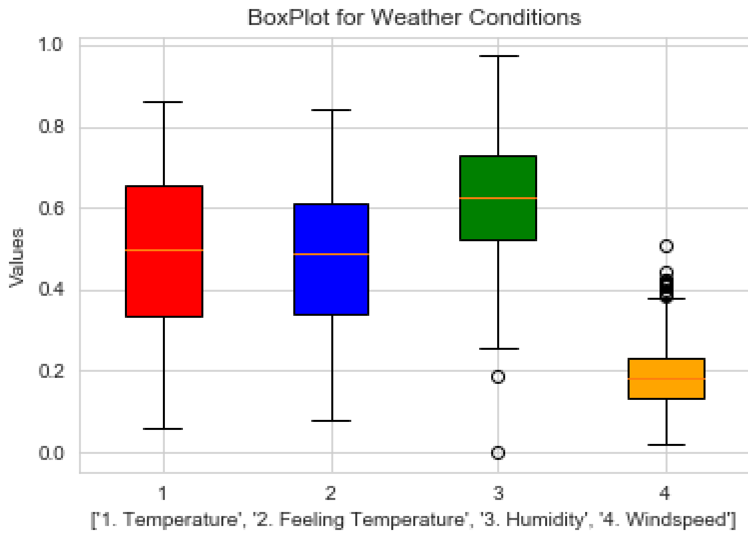








Outlier Analysis



Humidity and windspeed has outlier values
As windspeed has above 75th quartile so we have to deal with it

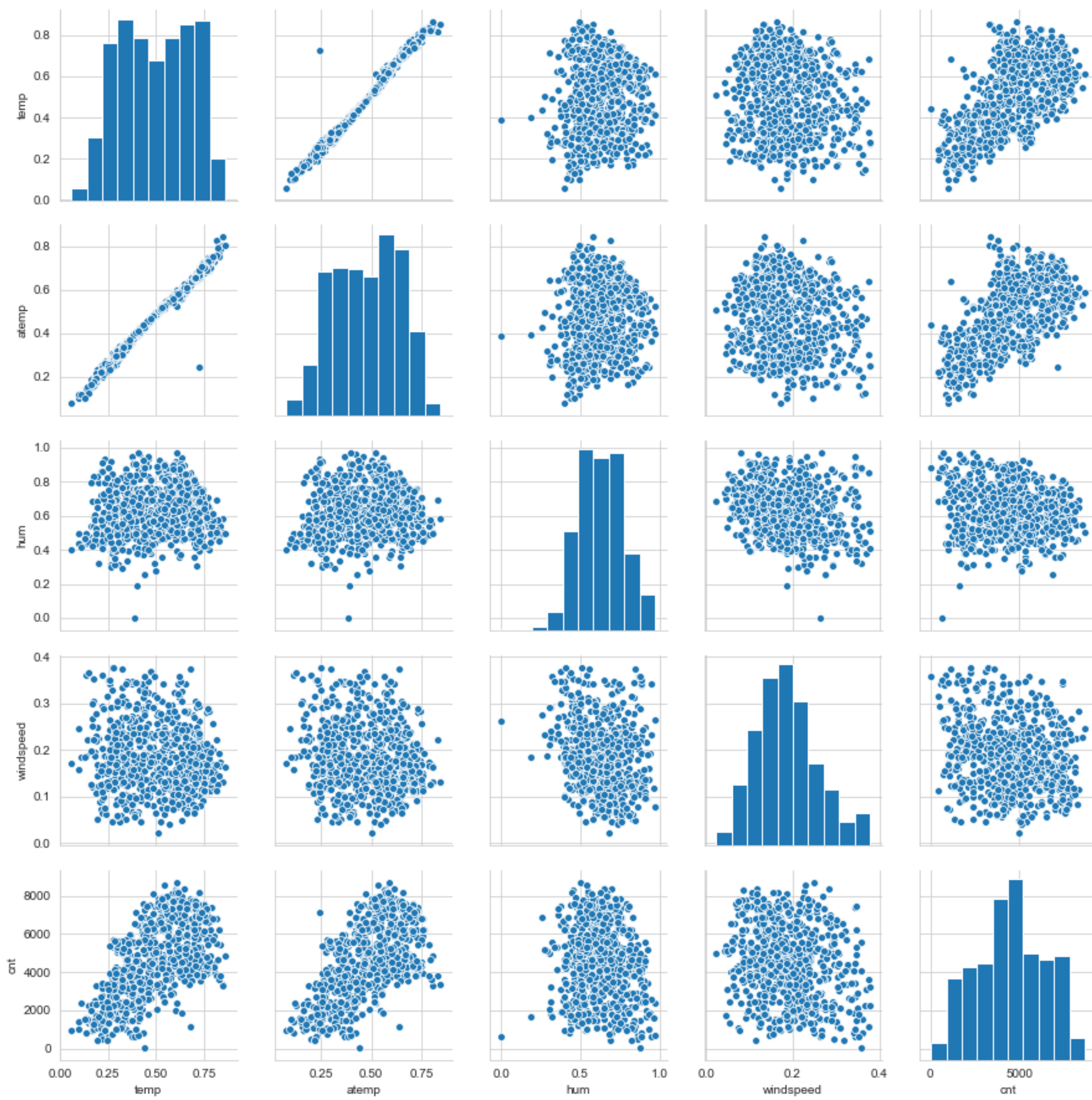
Using mean imputation method we imputed missing values in windspeed variable
data['windspeed'] = data['windspeed'].fillna(data['windspeed'].mean())

print(data.isnull().sum())

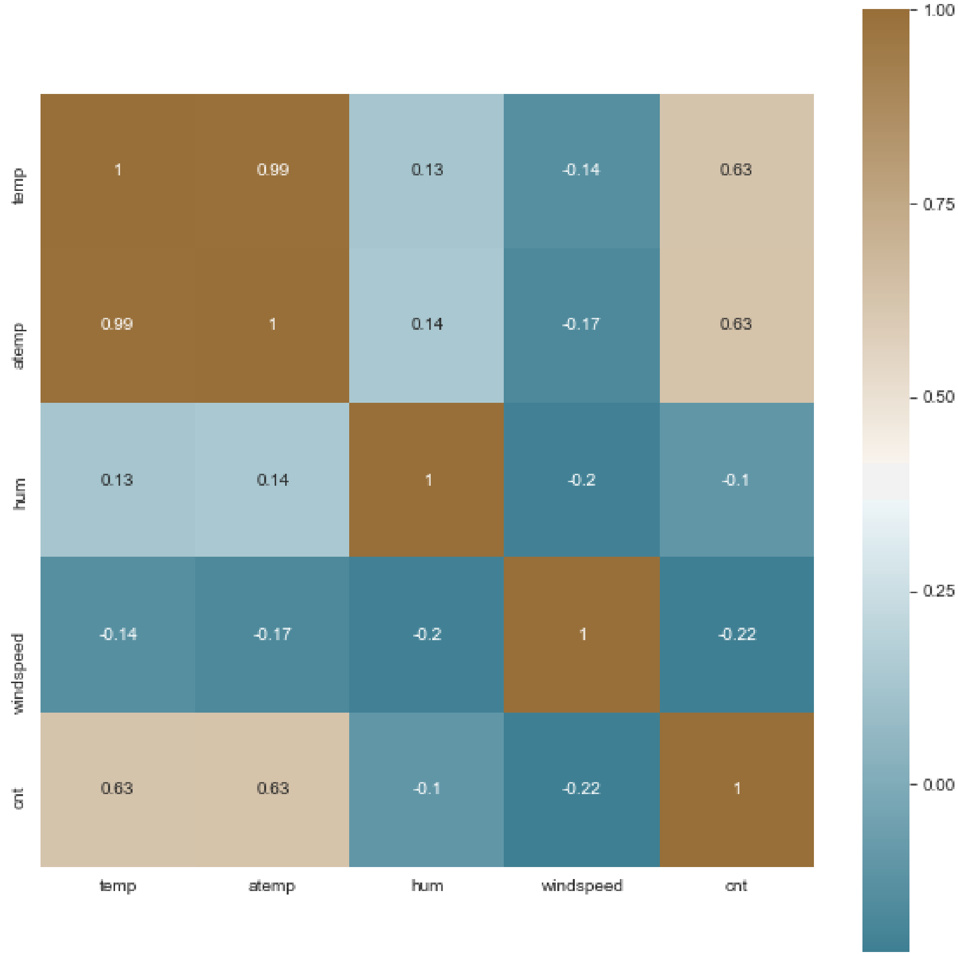
```
season        0
yr            0
mnth          0
holiday       0
weekday       0
workingday    0
weathersit    0
temp          0
atemp         0
hum           0
windspeed     0
cnt           0
dtype: int64
```

Now there are no missing values

Pairplots



Correlation Plot



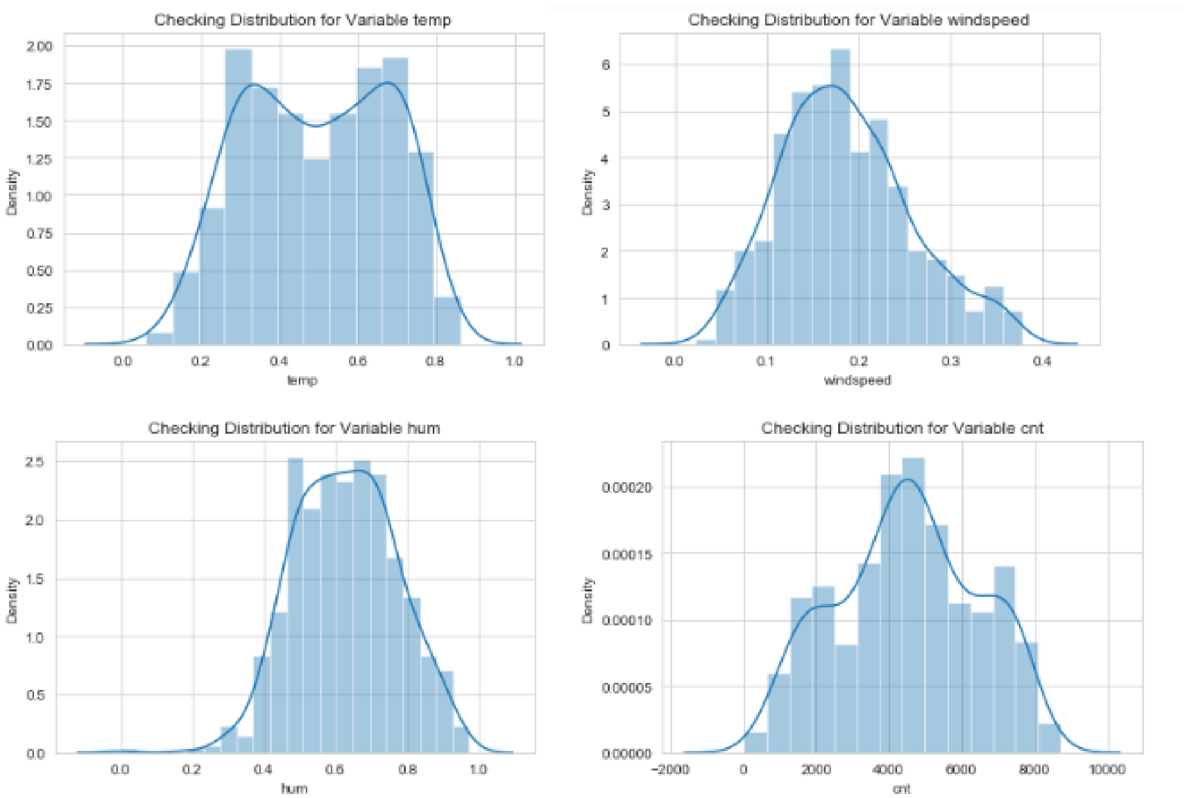
temp
atemp
hum
windspeed
cnt
1
0.99
0.13
-0.14
0.63
0.99
1
0.14
-0.17
0.63
0.13
0.14
1
-0.2
-0.1
-0.14
-0.17
-0.2
1
-0.22
0.63
0.63
-0.1
-0.22
1
1.00
0.75
0.50
0.25
0.00

ANOVA test for P values

```
                 sum_sq     df           F        PR(>F)
season    4.517974e+08    1.0  143.967653  2.133997e-30
Residual  2.287738e+09  729.0         NaN           NaN
                 sum_sq     df           F        PR(>F)
yr        8.798289e+08    1.0  344.890586  2.483540e-63
Residual  1.859706e+09  729.0         NaN           NaN
                 sum_sq     df          F        PR(>F)
mnth      2.147445e+08    1.0  62.004625  1.243112e-14
Residual  2.524791e+09  729.0        NaN           NaN
                 sum_sq     df         F    PR(>F)
holiday   1.279749e+07    1.0  3.421441  0.064759
Residual  2.726738e+09  729.0       NaN       NaN
                 sum_sq     df         F    PR(>F)
weekday   1.246109e+07    1.0  3.331091  0.068391
Residual  2.727074e+09  729.0       NaN       NaN
                   sum_sq     df         F    PR(>F)
workingday  1.024604e+07    1.0  2.736742  0.098495
Residual    2.729289e+09  729.0       NaN       NaN
                   sum_sq     df          F        PR(>F)
weathersit  2.422888e+08    1.0  70.729298  2.150976e-16
Residual    2.497247e+09  729.0        NaN           NaN
```

'temp' and 'atemp' are correlated so one of them should be removed

Feature Scaling



Data before scaling

	season	yr	mnth	weathersit	temp	hum	windspeed	cnt
0	1	0	1	2	0.344167	0.805833	0.160446	985
1	1	0	1	2	0.363478	0.696087	0.248539	801
2	1	0	1	1	0.196364	0.437273	0.248309	1349
3	1	0	1	1	0.200000	0.590435	0.160296	1562
4	1	0	1	1	0.226957	0.436957	0.186900	1600

Data after scaling

	season	yr	mnth	weathersit	temp	hum	windspeed	cnt
0	1	0	1	2	-0.826097	1.249316	-0.364668	985
1	1	0	1	2	-0.720601	0.478785	0.873479	801
2	1	0	1	1	-1.633538	-1.338358	0.870246	1349
3	1	0	1	1	-1.613675	-0.263001	-0.366777	1562
4	1	0	1	1	-1.466410	-1.340576	0.007143	1600

Applying machine learning algorithms

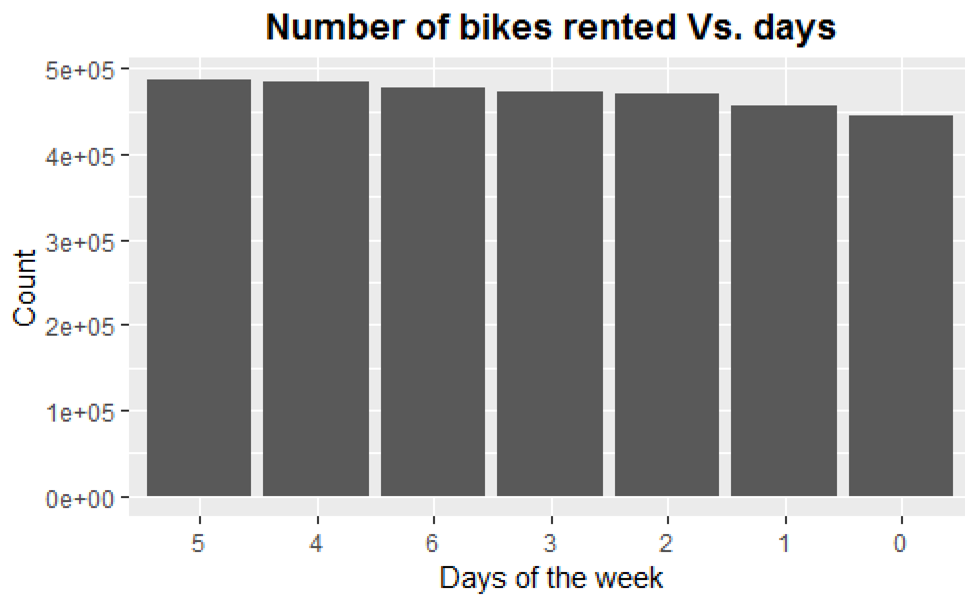
```
Decision Tree Model :
RMSE = 997.3873927346699
RSquared test = 0.7073525764693427
MAPE = 25.707144204754727

Random Forest Model :
RMSE = 569.3767118168532
RSquared test = 0.904628995618636
MAPE= 13.426577692653508

Linear Regression Model
RMSE= 736.2047259447532
RSquared test= 0.8405538055300172
MAPE= 17.217590042129967
```

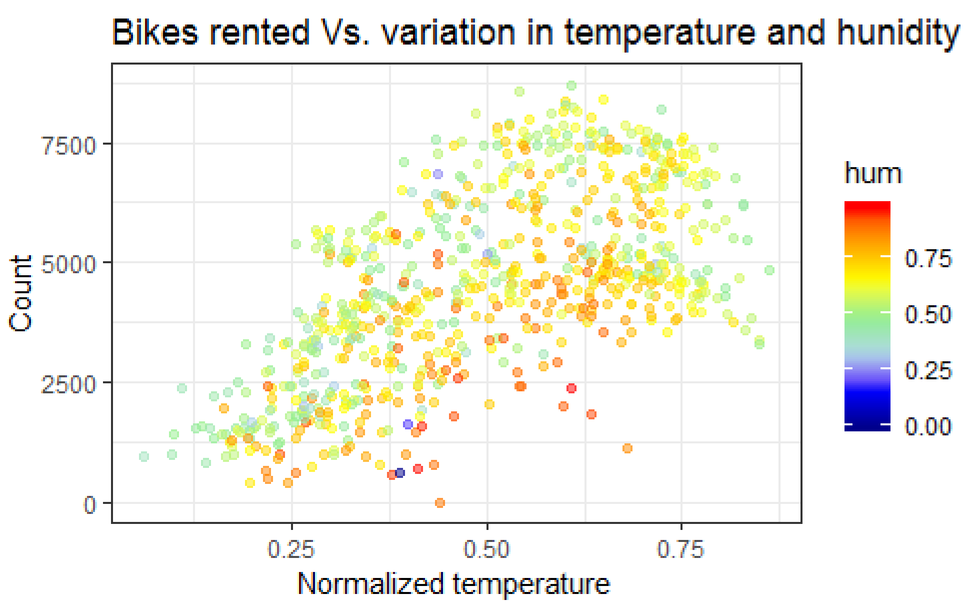
Conclusion

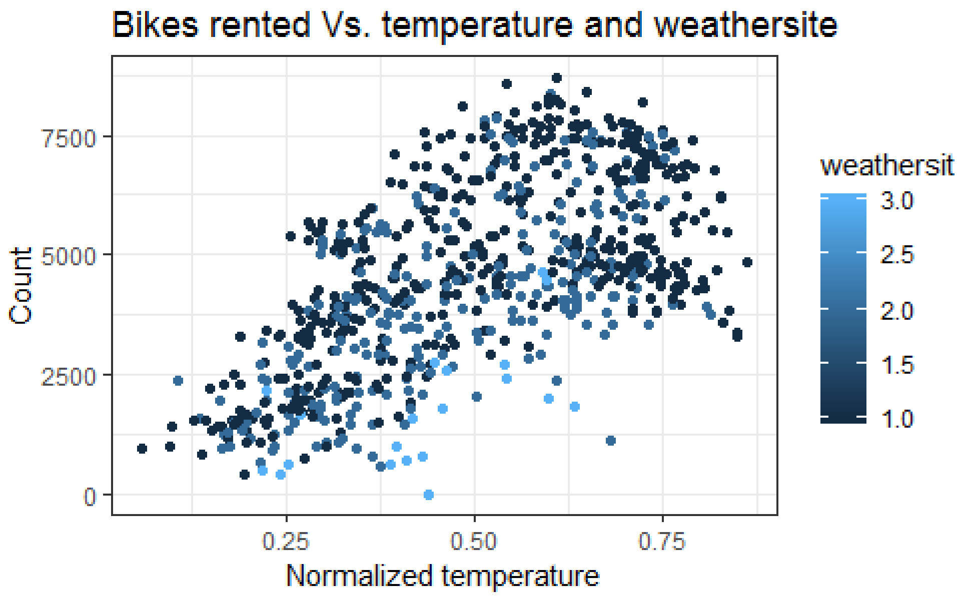
Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction **errors**). Residuals are a measure of how far from the regression line data points are, RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.
Whereas **R**-**squared** is a relative measure of fit, **RMSE** is an absolute measure of fit. As the square root of a variance, **RMSE** can be interpreted as the standard deviation of the unexplained variance and has the useful property of being in the same units as the response variable.
The mean absolute percentage error (MAPE), also known as mean absolute percentage deviation (MAPD), is a measure of prediction accuracy of a forecasting method in statistics, for example in trend estimation, also used as a Loss function for regression problems in Machine Learning.
Lower values of **RMSE and MAPE** and higher value of **R-Squared Value** indicate better fit.

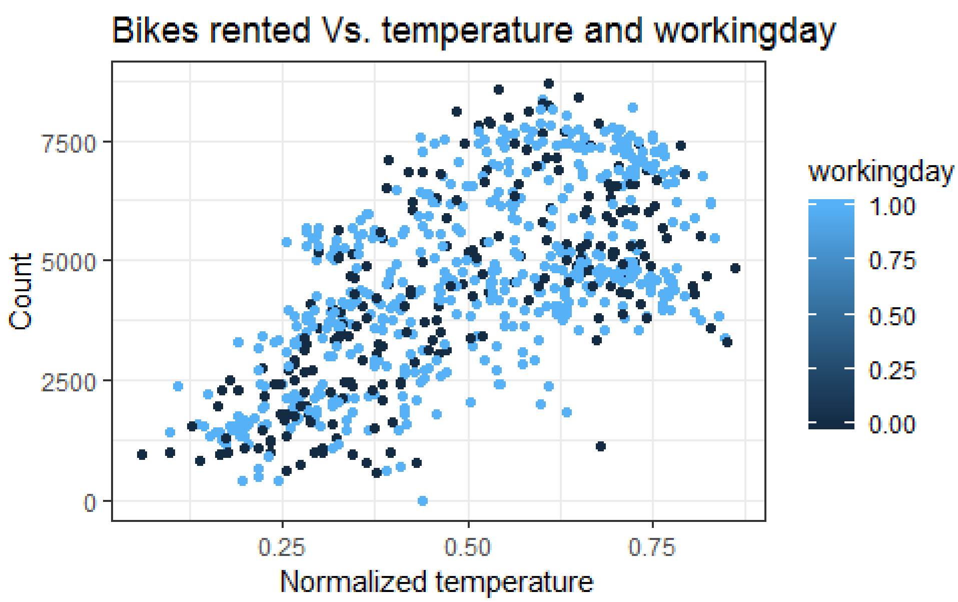
Choosing Random Forest as a method

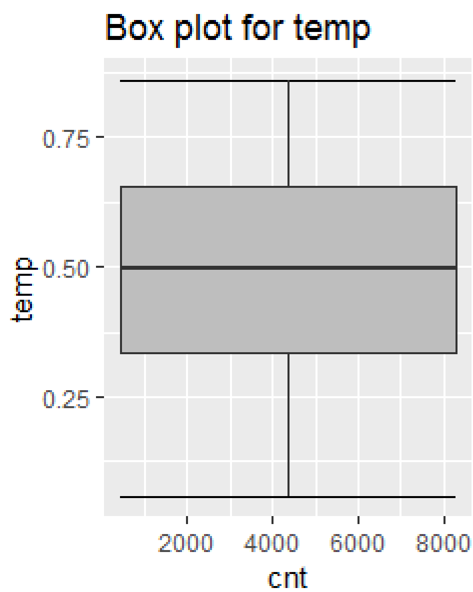
Extra figures of R code

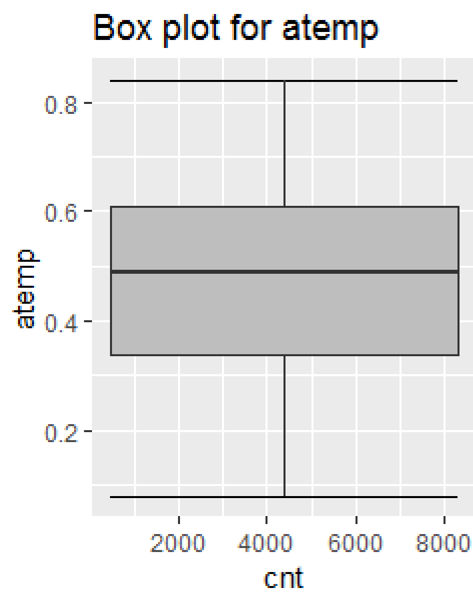

Number of bikes rented Vs. days
5e+05
4e+05
3e+05
2e+05
1e+05
0e+00
Count
5
4
6
3
2
1
0
Days of the week

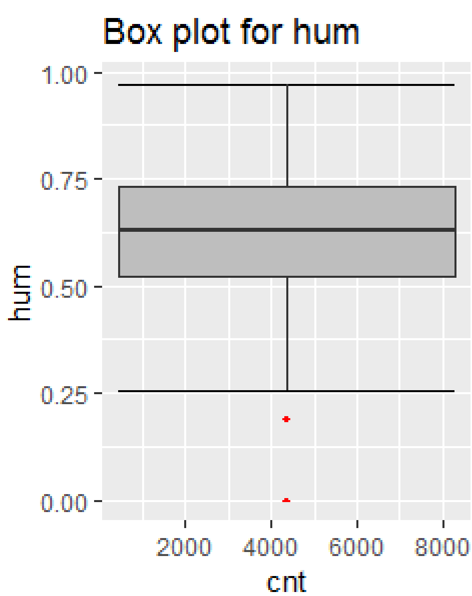

Bikes rented Vs. variation in temperature and hunidity
7500
5000
2500
0
Count
0.25
0.50
0.75
Normalized temperature
hum
0.75
0.50
0.25
0.00

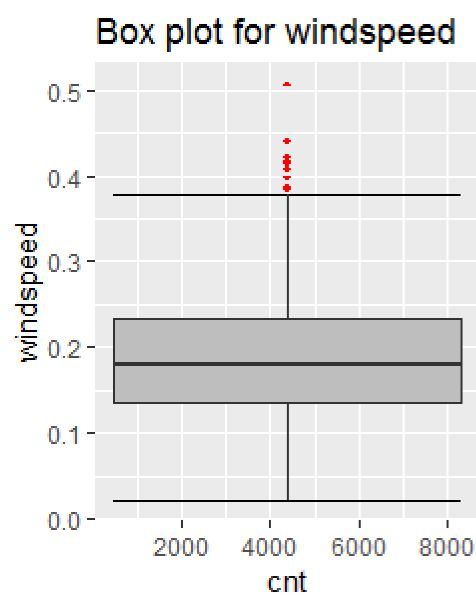

Bikes rented Vs. temperature and weathersite
Count
7500
5000
2500
0
0.25
0.50
0.75
Normalized temperature
weathersit
3.0
2.5
2.0
1.5
1.0


Bikes rented Vs. temperature and workingday
Count
7500
5000
2500
0
0.25
0.50
0.75
Normalized temperature
workingday
1.00
0.75
0.50
0.25
0.00

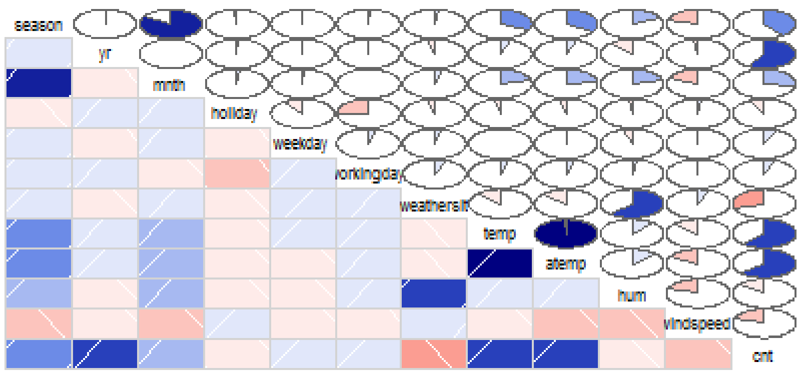

Box plot for temp
temp
0.75
0.50
0.25
2000
4000
6000
8000
cnt


Box plot for atemp
atemp
0.8
0.6
0.4
0.2
2000
4000
6000
8000
cnt


Box plot for hum
hum
1.00
0.75
0.50
0.25
0.00
2000
4000
6000
8000
cnt


Box plot for windspeed
windspeed
0.5
0.4
0.3
0.2
0.1
0.0
2000
4000
6000
8000
cnt

CORRELATION PLOT



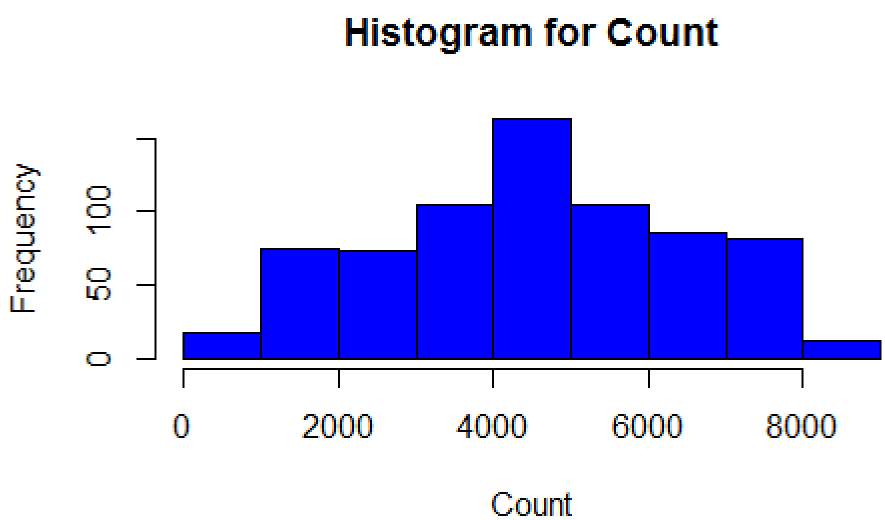
Temp and atemp are highly correlated as we see dark blue color

ANOVA test

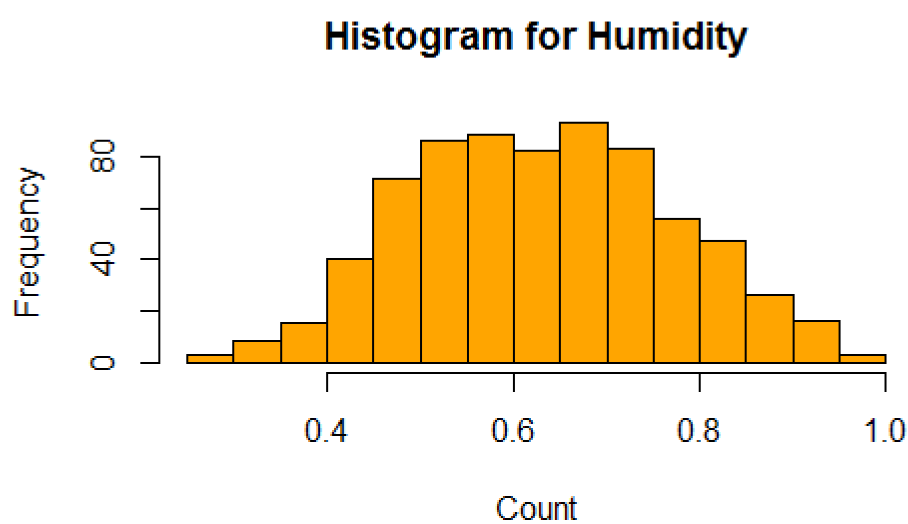
```
> summary(aov(formula = cnt~season,data = df))
             Df    Sum Sq   Mean Sq F value Pr(>F)    
season        1 4.268e+08 426760312   135.6 <2e-16 ***
Residuals   715 2.250e+09   3146948                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(aov(formula = cnt~yr,data = df))
             Df    Sum Sq   Mean Sq F value Pr(>F)    
yr            1 8.813e+08 881327066     351 <2e-16 ***
Residuals   715 1.796e+09   2511190                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(aov(formula = cnt~mnth,data = df))
             Df    Sum Sq   Mean Sq F value   Pr(>F)    
mnth          1 2.035e+08 203533335   58.84 5.61e-14 ***
Residuals   715 2.473e+09   3459153                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(aov(formula = cnt~holiday,data = df))
             Df    Sum Sq  Mean Sq F value Pr(>F)  
holiday       1 1.377e+07 13770983   3.697 0.0549 .
Residuals   715 2.663e+09  3724555                 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(aov(formula = cnt~weekday,data = df))
             Df    Sum Sq  Mean Sq F value Pr(>F)  
weekday       1 1.381e+07 13809167   3.708 0.0546 .
Residuals   715 2.663e+09  3724502                 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> summary(aov(formula = cnt~workingday,data = df))
             Df    Sum Sq Mean Sq F value Pr(>F)
workingday    1 8.494e+06 8494340   2.276  0.132
Residuals   715 2.668e+09 3731935
> summary(aov(formula = cnt~weathersit,data = df))
             Df    Sum Sq   Mean Sq F value Pr(>F)
weathersit    1 2.432e+08 243197751   71.45 <2e-16 ***
Residuals   715 2.434e+09   3403678
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

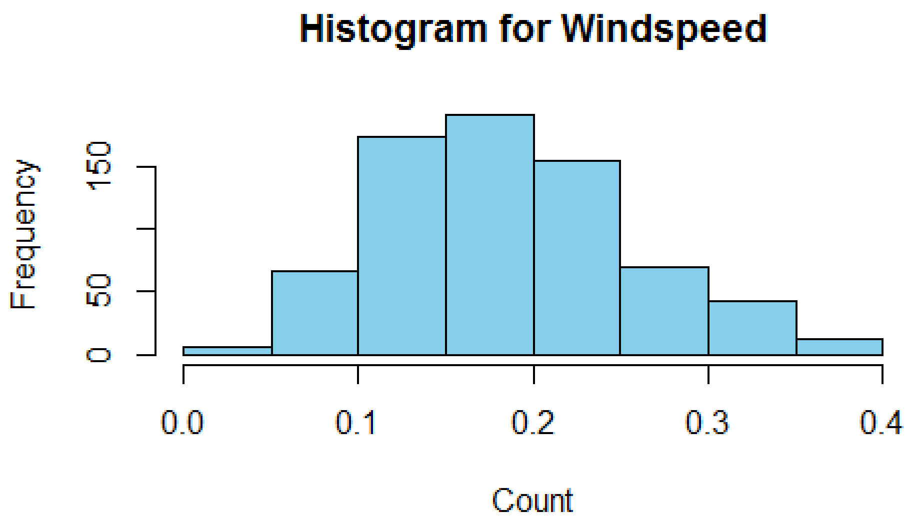
hist(df$cnt, col="blue", xlab="Count", main="Histogram for Count")



hist(df$hum, col="orange", xlab="Count", main="Histogram for Humidity")



```
hist(df$windspeed, col="sky blue", xlab="Count", main="Histogram for Windspeed")
```




```
hist(df$temp, col="red", xlab="Count", main="Histogram for Temperature")
```