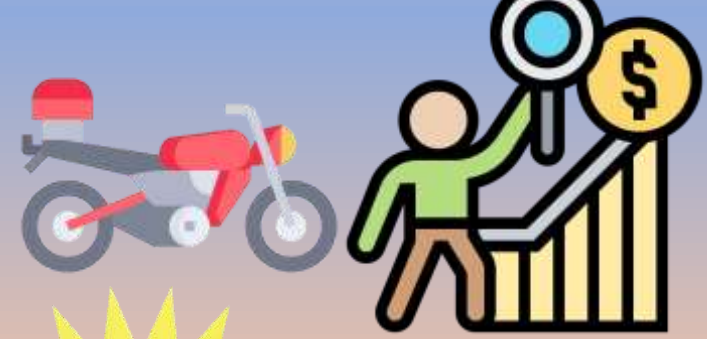
# Seoul Bike Sharing Demand Prediction Supervised Regression

Analyses of all features

# Weather related features




# Features

## 1. Numerical Features

- Date
- Temperature (C)
- Due Point Temperature (C)
- Wind speed (m/s)
- Visibility (10m)
- Solar Radiation (MJ/m2)
- Rainfall (mm)
- Snowfall (mm)
- Humidity (%)
- Hour

## 2. Categorical Features

- Seasons
- Functioning Day
- Holiday

8760 Rows

13 features + 1 Target variable

Rented Bike Count

# Checking Missing Values

## Missing Values occur due to:

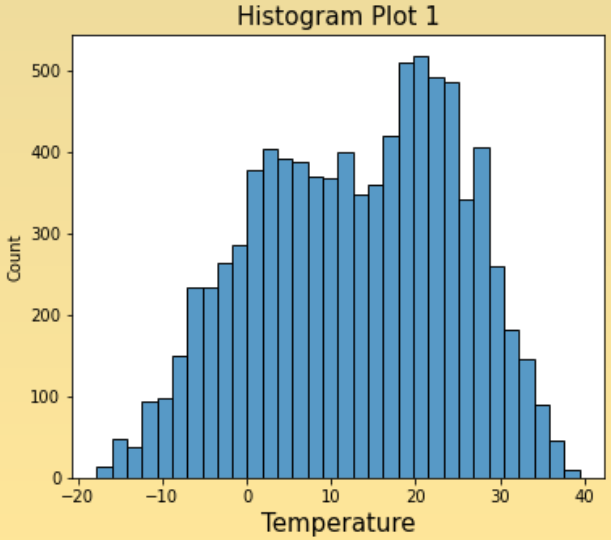
1. Human Error
2. Corrupt Data
3. Customer not willing to share the data

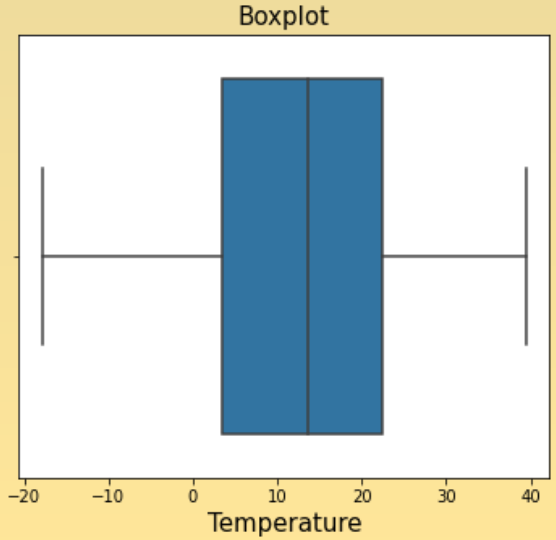
## Missing Values can be stored in the form of

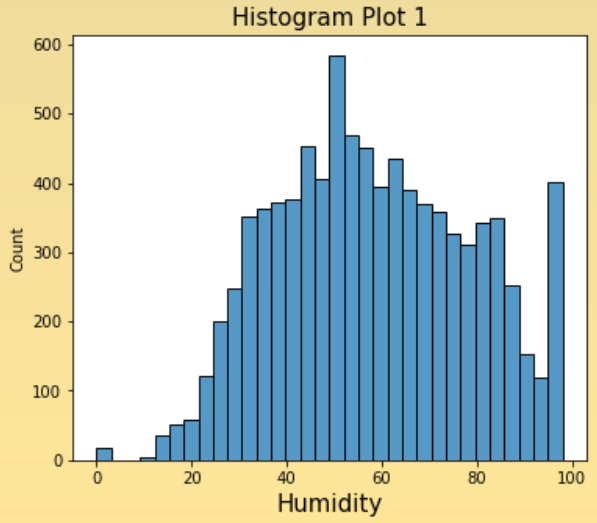

## Zero is present in most of all the features

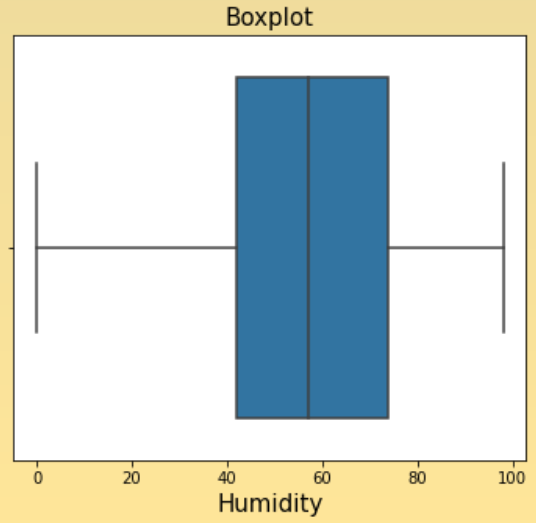
| | | |
|---|---|---|
| Snowfall (cm) | 8317 | 94.942922 |
| Rainfall(mm) | 8232 | 93.972603 |
| Solar Radiation (MJ/m2) | 4300 | 49.086758 |
| Hour | 365 | 4.166667 |
| Rented Bike Count | 295 | 3.367580 |
| Wind speed (m/s) | 74 | 0.844749 |
| Dew point temperature(°C) | 60 | 0.684932 |
| Temperature(°C) | 21 | 0.239726 |
| Humidity(%) | 17 | 0.194064 |

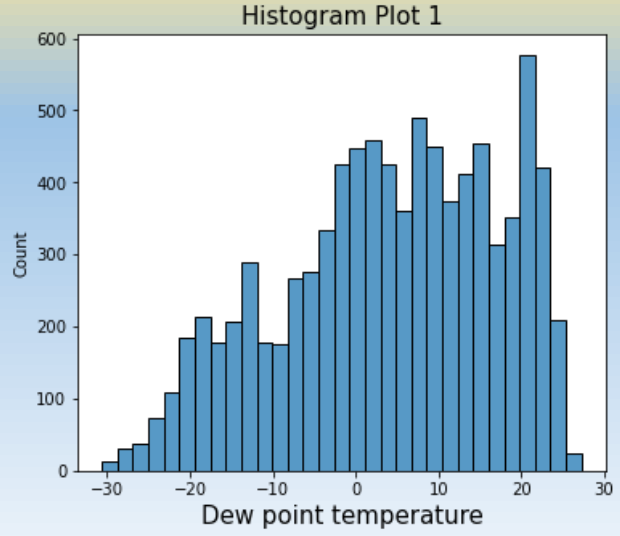
# Exploratory Data Analysis

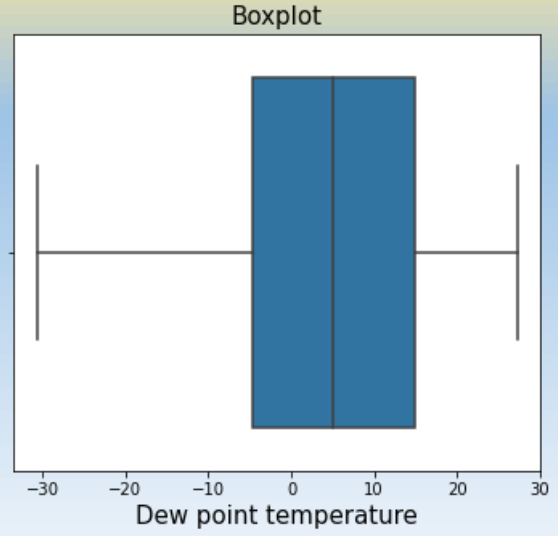


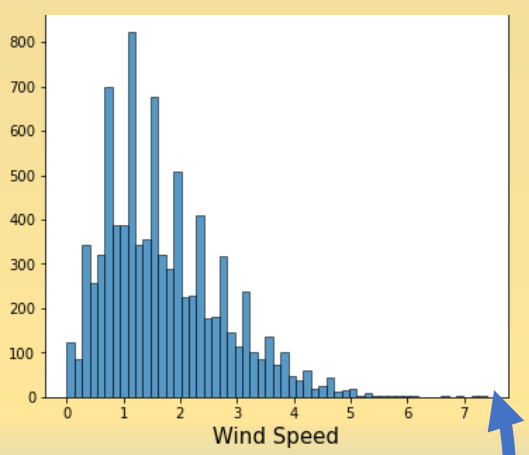

Temperature

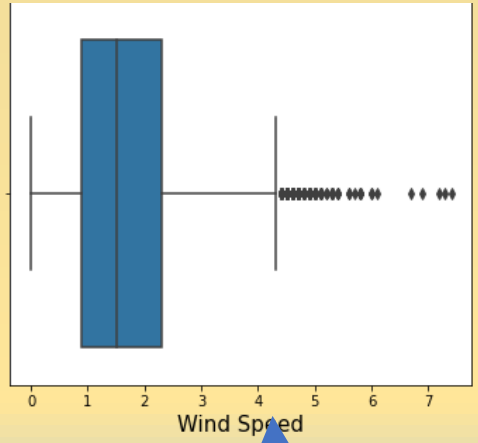
Due point Temperature

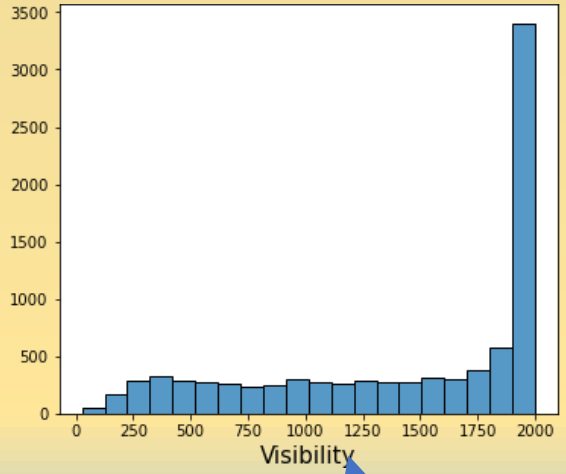
Humidity

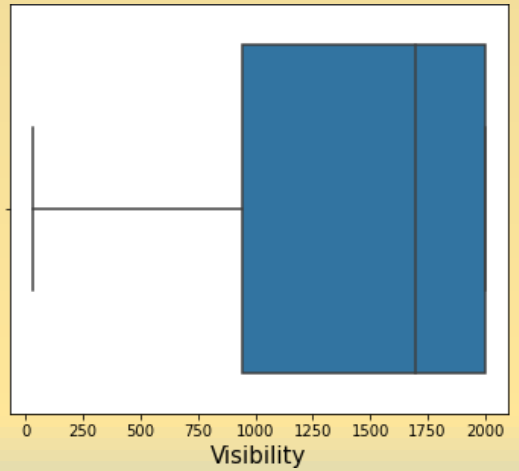
Only these three features have normal distribution

# Exploratory Data Analysis (Continued)



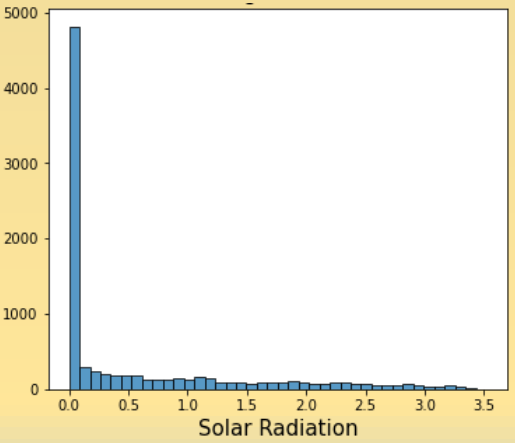



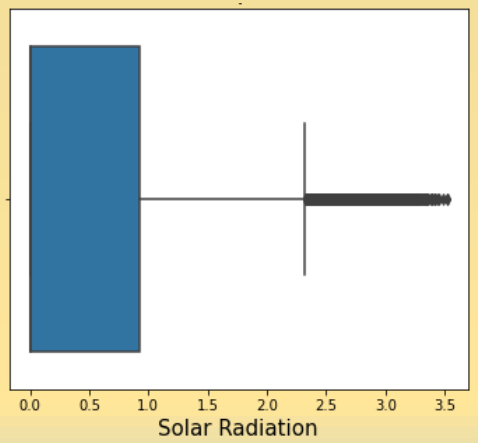

Outliers present

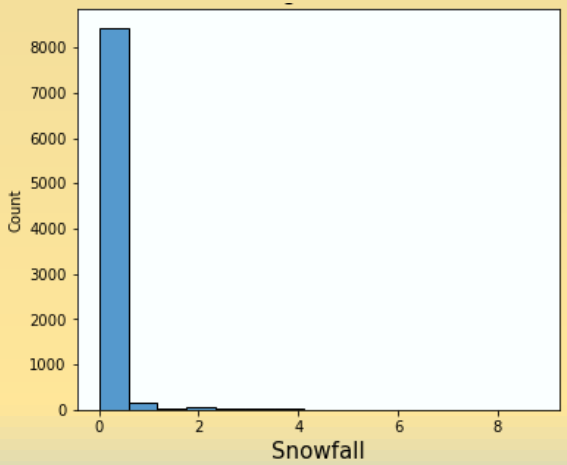
Not normally distributed

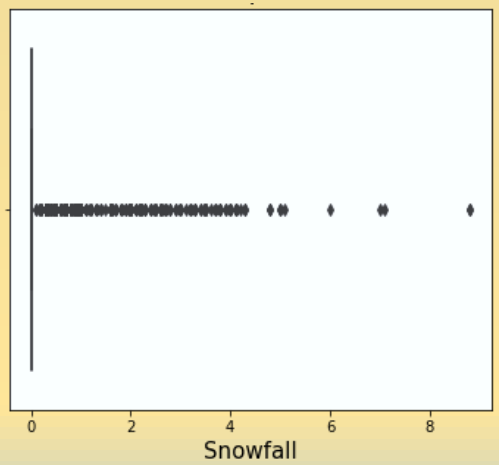
Wind speed and Visibility are not normally distributed and there are some outliers found in Wind speed

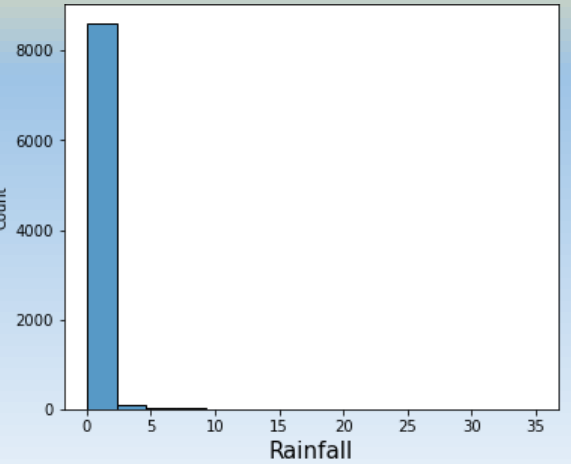
# Exploratory Data Analysis (Continued)

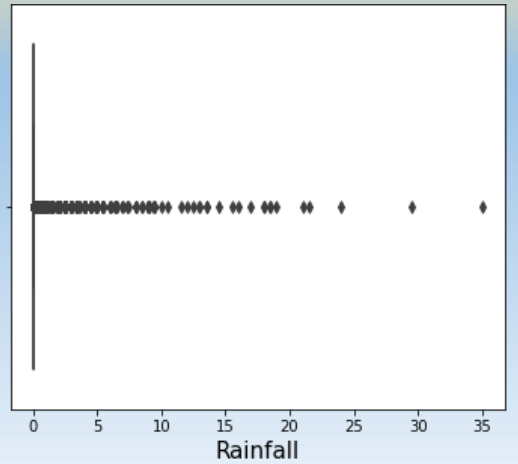


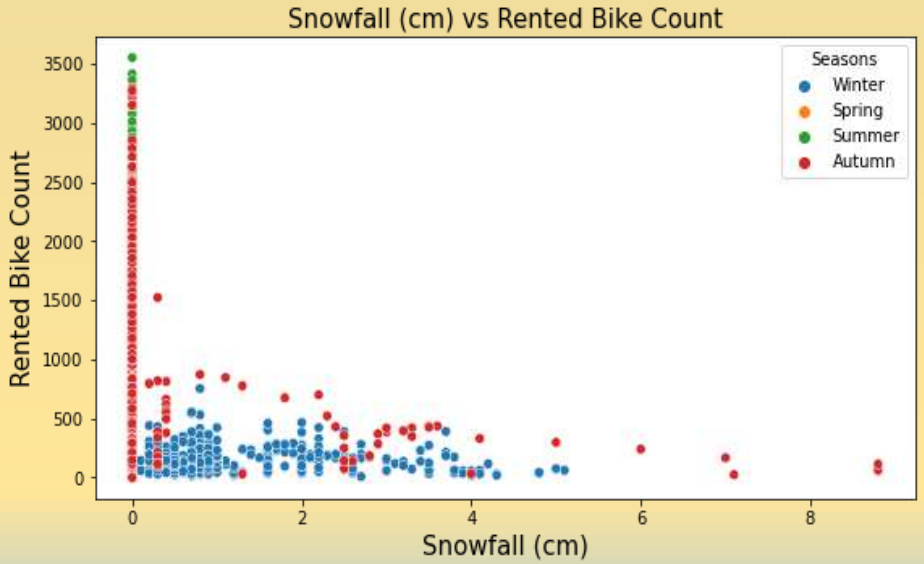



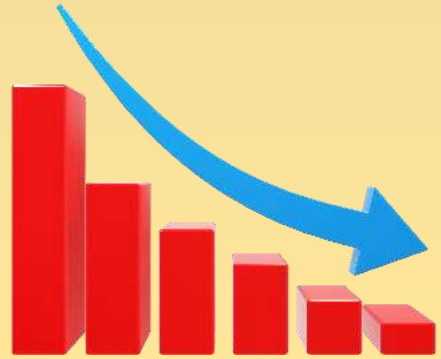

There is a lot outliers in the features i.e.

- Rain Fall
- Snow Fall
- Solar Radiation

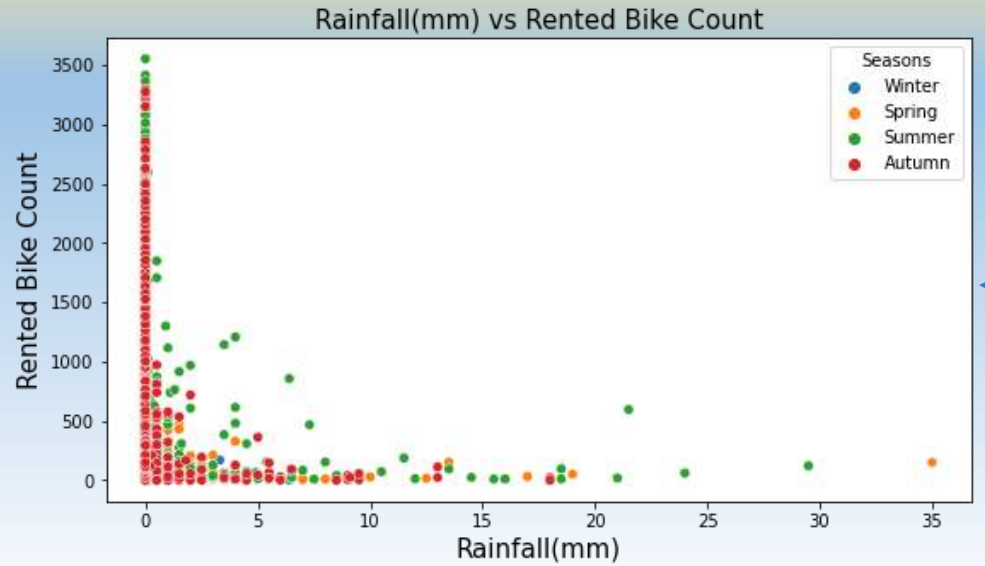
About 25% to 30% of data are outliers and they are natural according to the seasons

# Exploratory Data Analysis (Continued)



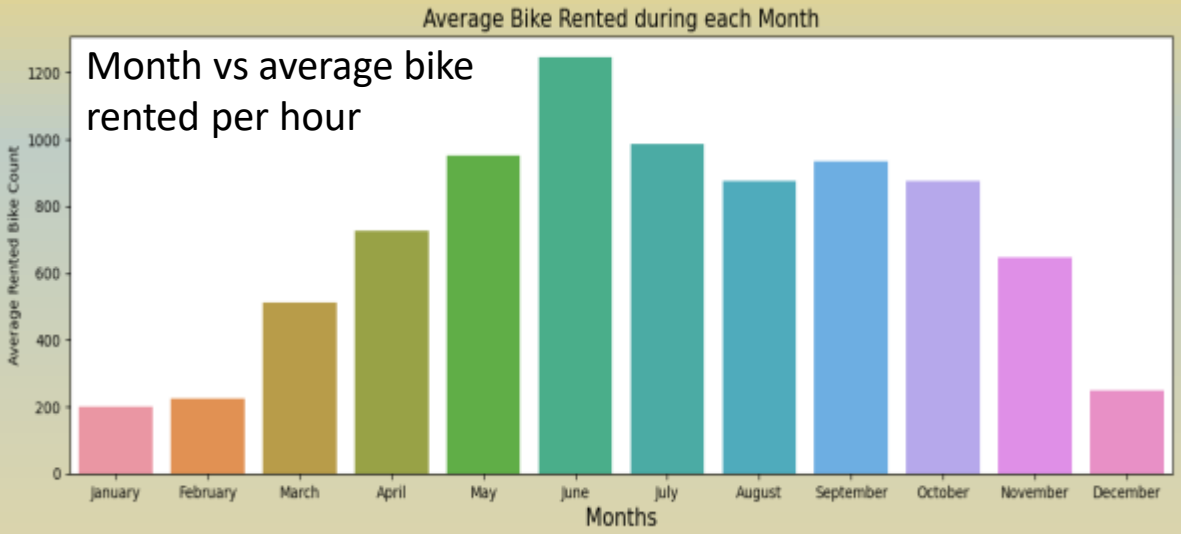



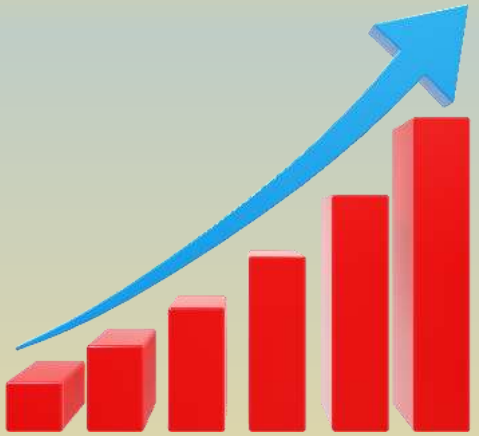
Winter & Monsoon



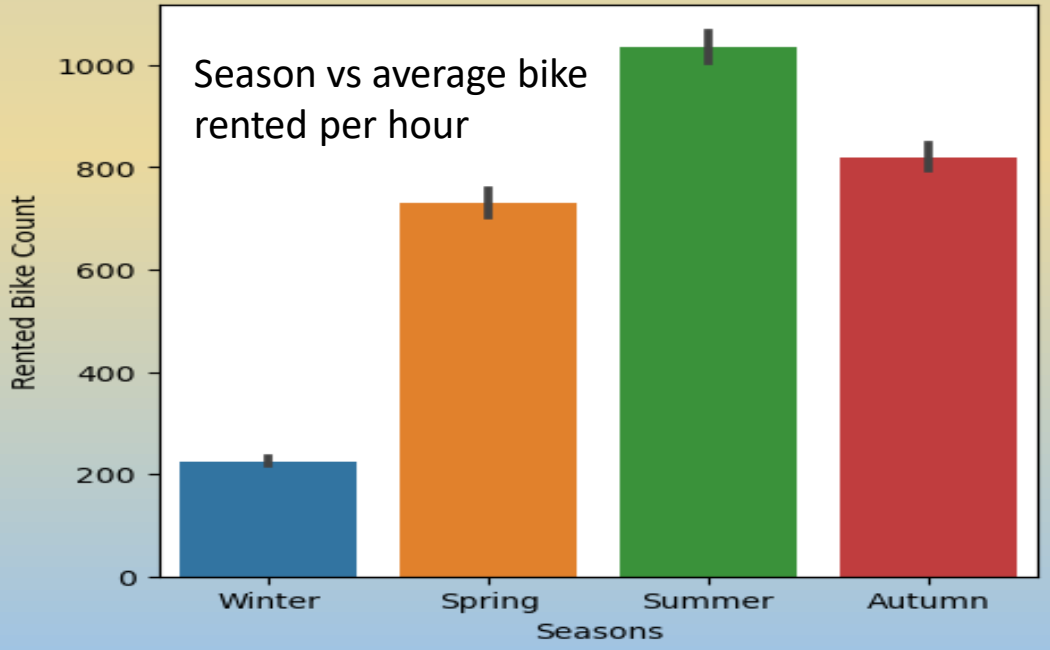

Negative Correlation

# Exploratory Data Analysis (Continued)



Autumn & Summer and spring



# Conclusion on visual analysis of data

1. Out of all 10 numerical features Date and hour are not random.

2. Only 3 features are normal distributed and they have no outliers i.e. Temperature, Due point Temperature, Humidity

3. Left all other features like Solar radiation, Rainfall, Snowfall, Visibility and Windspeed are not normally distributed and have lot of outliers

**Solar Radiation**

Most of the values recorded during summer or autumn will be treated as outliers Because maximum of values recorded will be zero or close to zero

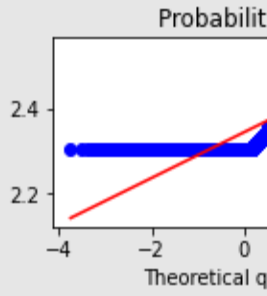
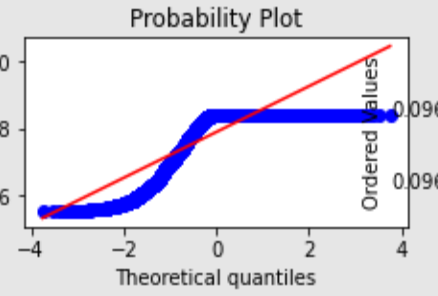
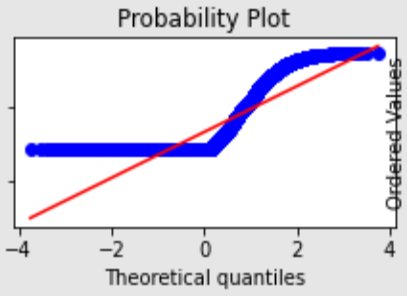
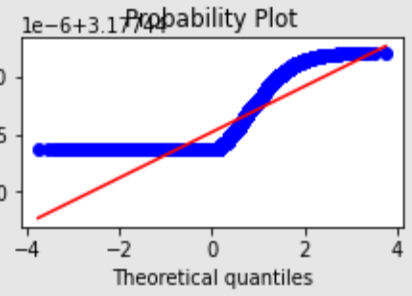
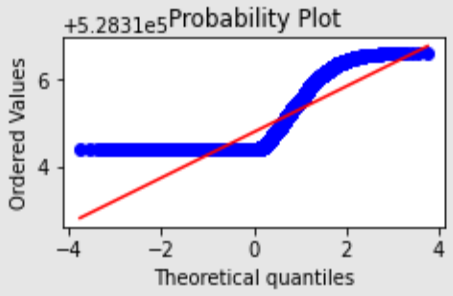
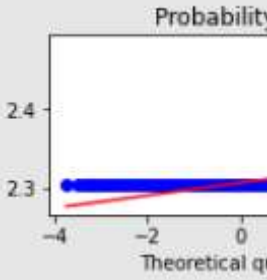
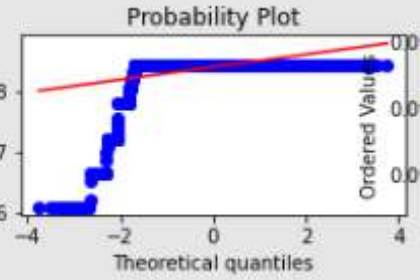
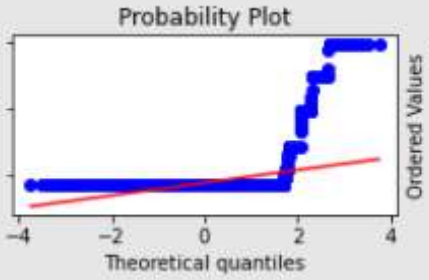
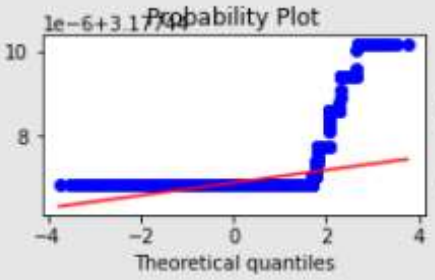
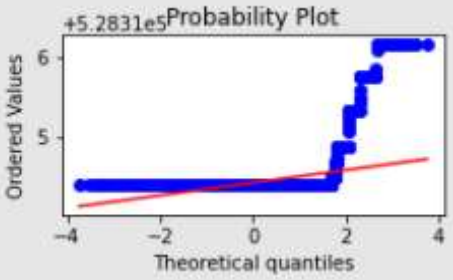
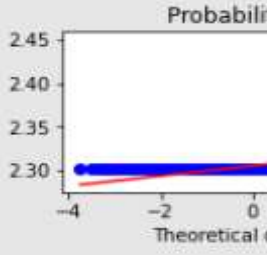
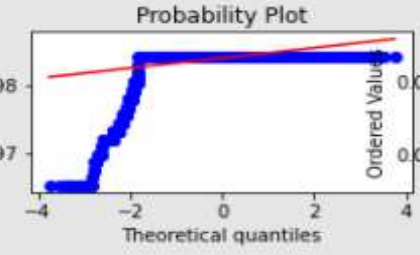
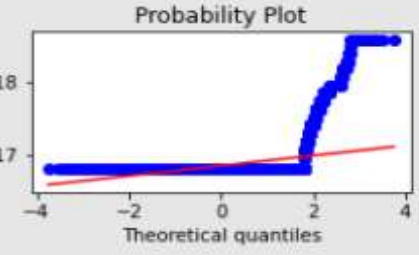
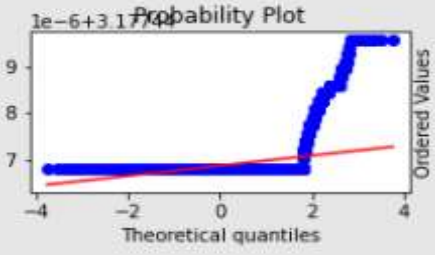
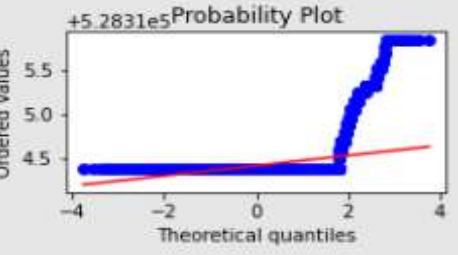
**Rainfall** **Snowfall**

Most of the values recorded during winter or monsoon will be treated as outliers Because maximum of values recorded will be zero or close to zero (mm)

If we try to remove all the outliers we might end up losing around 50% to 70%

# Attempt to make distribution normal by transformation

Was not able to transform data to achieve normality using transformations like:

| | Log Transformation | Inverse Log Transformation | Reciprocal Transformation | Square root Transformation | Exponential Transformation |
|---|---|---|---|---|---|
| 1. SOLAR RADIATION |  |  |  |  |  |
| 2. RAIN FALL |  |  |  |  |  |
| 3. SNOW FALL |  |  |  |  |  |

# Conclusion on visual analysis of data (Continued)

Due to Presence of lot of outliers in features like Rain Fall, Snowfall, Solar Radiation, Wind speed, visibility

Rain Fall Snow Fall solar radiation, Wind speed , Visibility are not normally distributed

**Linear Models does not perform well** !!!!!

Linear Regression

**Sensitive to Outliers**

Hence we can try models Like:

1. **Decision Tree**
2. **Random Forest**
3. **AdaBoost**
4. **Gradient Boost**
5. **XGBoost**

Insensitive to Outliers

# Attempt for Linear regression model (performance check)

## Handled Outliers in Data for best performance of Linear Regression

## Procedure for Handling Outliers

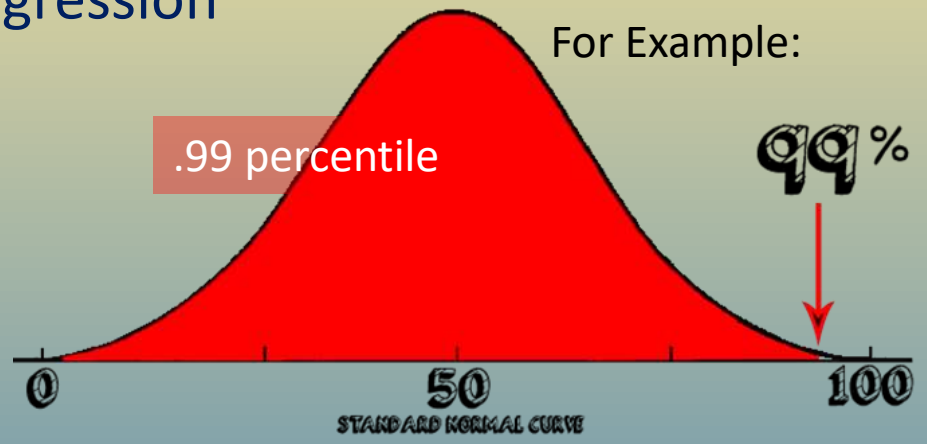
Removing values greater than .99 percentile in Wind speed

Removing values greater than .95 percentile in Solar Radiation

Removing values greater than .95 percentile in Snowfall

Removing values greater than .95 percentile in Rainfall

For Example:




```
Total 1193 rows is been removed due to outliers
```

8760 Rows

13% Data Loss

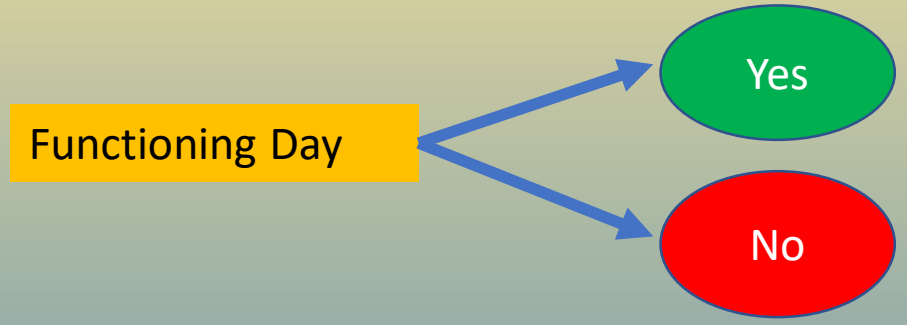
## Handling Multi-Collinearity between features

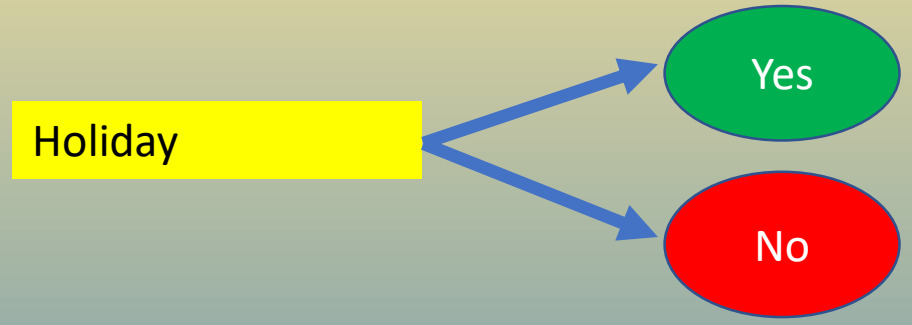
| VIF | Features |
|---|---|
| 7.529879 | Humidity(%) |
| 5.771939 | Visibility (10m) |
| 4.981123 | Month |
| 4.781143 | Wind speed (m/s) |
| 3.935537 | Hour |

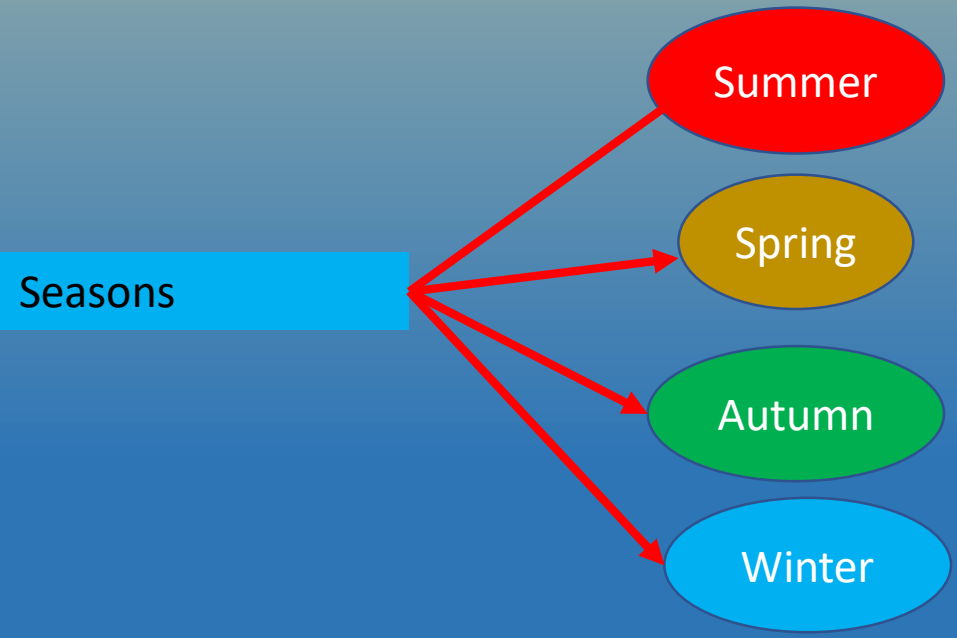
| VIF | Features |
|---|---|
| 1.081249 | Rainfall(mm) |
| 1.123575 | Snowfall (cm) |
| 1.697150 | Dew point temperature(C) |
| 1.969341 | Solar Radiation (MJ/m2) |
| 3.839832 | Day |

Using Variation Inflation Factor method

# Encoding Categorical Features




One hot Encoding for Season Feature

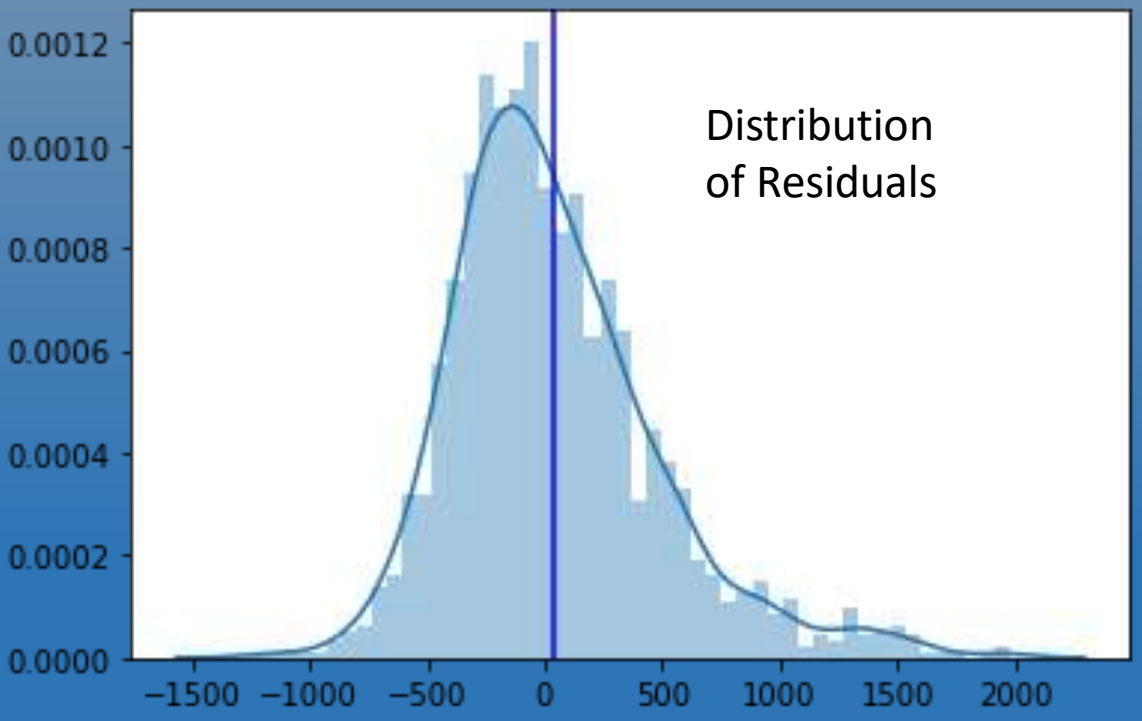
# Performance of Linear Regression

```
LinearRegression
MAE: 331.8812713565375
MSE: 198775.981568335
RMSE: 445.84300103100753
R2_score_train: 55.58%
R2_test: 56.92%
--------------------
```

**Poor Performance by Linear Regression Model**
**Not more than 70%**

Low Performance on Train Data set

1. RIDGE X
2. LASSO X
3. ELASTIC NET X




Mean of residuals is 41

Median of residuals is -35

# Hyper tuning models

- Decision Tree
- Random Forest
- Ada Boost
- XGboost
- Gradient Boost

**Decision Tree: 80 %**

a. max depth (Best 12)
b. max-leaf nodes (Best: None)
c. min samples leaf (Best:10)
d. splitter (Best)

**Random Forest: 85 %**

a. n estimators (Best 80)
b. max-leaf nodes (Best: None)
c. min samples leaf (Best:9)
d. max depth (None)

**XG Boost: 88 %**

a. lambda (Best 8)
b. max depth (Best 8)
c. gamma (Best 2.0)
d. learning rate (Best 1)
e. eta (Best 0.2)
f. alpha (Best 1.0)

**Ada Boost: 67 %**

a. n estimators (Best 80)
b. loss (Best square)
c. Learning Rate (Best 0.1)

**Gradient Boost 84 %**

a. n estimators (Best 80)
b. min samples leaf (Best 8)
c. max-leaf nodes (Best None)
d. learning rate (Best 1)
e. max features (Best 7)

# Model Performance

MAE — Mean Absolute Error

MSE — Mean Squared Error

RMSE — Root Mean squared Error

R2_test — R2 score

| | MAE | MSE | RMSE | R2_test |
|---|---|---|---|---|
| XGBRegressor | 127.603121 | 48342.519273 | 219.869323 | 88.4% |
| RandomForestRegressor | 149.119942 | 59442.133188 | 243.807574 | 85.73% |
| GradientBoostingRegressor | 176.008939 | 64348.651005 | 253.670359 | 84.56% |
| DecisionTreeRegressor | 170.133235 | 80004.770261 | 282.851145 | 80.8% |
| AdaBoostRegressor | 274.256987 | 136382.867764 | 369.300511 | 67.27% |

**Adaboost is performing poor**

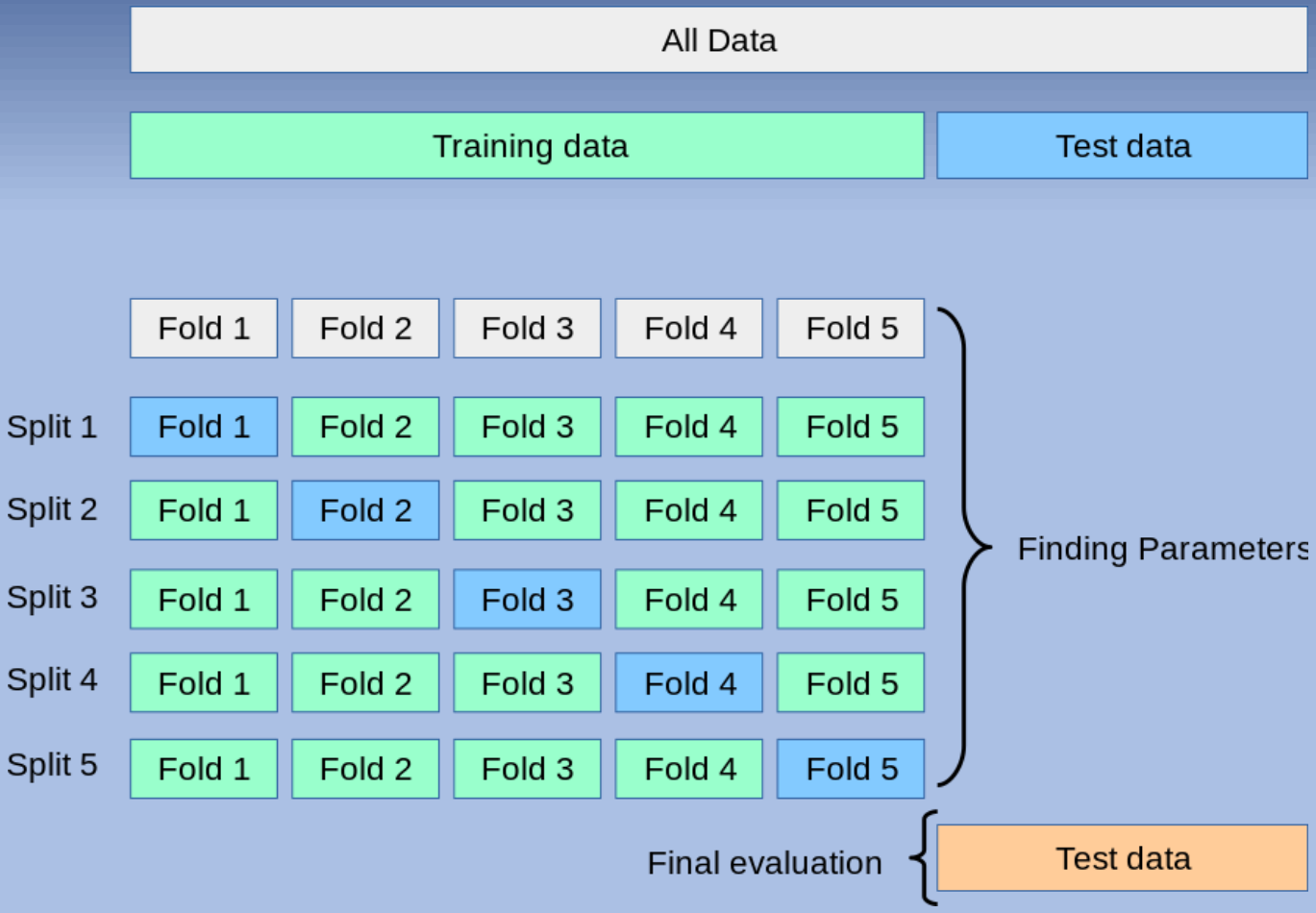
**These are the 4 models that are performing well on the data giving more than 70 % accuracy**

1.XGBRegressor
2.RandomForestRegressor
3.GradientBoostingRegressor
4.DecisionTreeRegressor

# Feature Importance

| | Decision Tree | Random Forest | AdaBoost | Gradient Boost | XGBoost |
|---|---|---|---|---|---|
| 0 | Temperature(C) | Temperature(C) | Hour | Hour | Seasons_Winter |
| 1 | Hour | Hour | Temperature(C) | Seasons_Winter | Functioning Day |
| 2 | Humidity(%) | Functioning Day | Solar Radiation (MJ/m2) | Temperature(C) | Rainfall(mm) |
| 3 | Functioning Day | Solar Radiation (MJ/m2) | Functioning Day | Rainfall(mm) | Hour |
| 4 | Solar Radiation (MJ/m2) | Humidity(%) | Humidity(%) | Functioning Day | Temperature(C) |
| 5 | Dew point temperature(C) | Rainfall(mm) | Rainfall(mm) | Dew point temperature(C) | Solar Radiation (MJ/m2) |
| 6 | Seasons_Winter | Dew point temperature(C) | Seasons_Winter | Visibility (10m) | Humidity(%) |
| 7 | Month | Seasons_Winter | Dew point temperature(C) | Solar Radiation (MJ/m2) | Month |
| 8 | Rainfall(mm) | Month | Month | Month | Holiday |
| 9 | Day | Day | Wind speed (m/s) | Humidity(%) | Dew point temperature(C) |
| 10 | Wind speed (m/s) | Visibility (10m) | Seasons_Summer | Day | Seasons_Summer |

# Cross validating model's score range on dataset




K Fold :50

1. Result score of 50 shuffled split data is been calculated with best parameters of tuned model obtained from hypertuning
2. Mean and Standard Deviation of the scores are calculated
3. From the mean and standard deviation of scores we can calculate the confidence interval
4. Rather than giving the point estimate Its better to give clarity about the model performance range

# Estimating performance range of models

For 95% Confidence

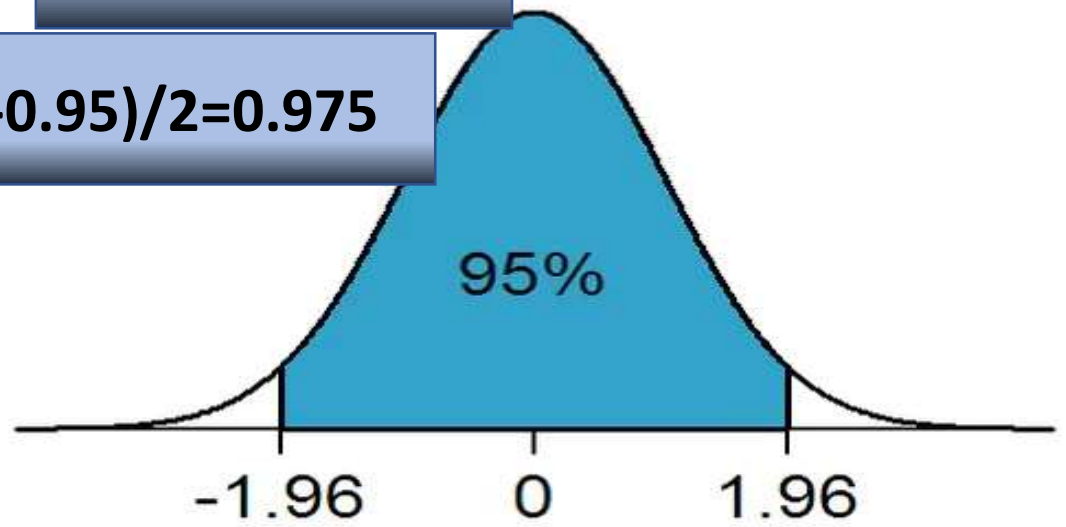
Formula for confidence Intervals | Mean(score) - 1.96 x Stdev(score) | to | Mean(score) + 1.96 x Stdev(score)

**Why 1.96?? !!**

| z | 0 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 |
|---|---|---|---|---|---|---|---|
| +0 | .50000 | .50399 | .50798 | .51197 | .51595 | .51994 | .52 92 |
| +0.1 | .53983 | .54380 | .54776 | .55172 | .55567 | .55966 | .56 60 |
| +0.2 | .57926 | .58317 | .58706 | .59095 | .59483 | .59871 | .60 57 |
| +0.3 | .61791 | .62172 | .62552 | .62930 | .63307 | .63683 | .64 58 |
| +0.4 | .65542 | .65910 | .66276 | .66640 | .67003 | .67364 | .67 24 |
| +0.5 | .69146 | .69497 | .69847 | .70194 | .70540 | .70884 | .71 26 |
| +0.6 | .72575 | .72907 | .73237 | .73565 | .73891 | .74215 | .74 37 |
| +0.7 | .75804 | .76115 | .76424 | .76730 | .77035 | .77337 | .77 37 |
| +0.8 | .78814 | .79103 | .79389 | .79673 | .79955 | .80234 | .80 11 |
| +0.9 | .81594 | .81859 | .82121 | .82381 | .82639 | .82894 | .83 47 |
| +1 | .84134 | .84375 | .84614 | .84849 | .85083 | .85314 | .85 43 |
| +1.1 | .86433 | .86650 | .86864 | .87076 | .87286 | .87493 | .87 98 |
| +1.2 | .88493 | .88686 | .88877 | .89065 | .89251 | .89435 | .89 17 |
| +1.3 | .90320 | .90490 | .90658 | .90824 | .90988 | .91149 | .91 08 |
| +1.4 | .91924 | .92073 | .92220 | .92364 | .92507 | .92647 | .92 85 |
| +1.5 | .93319 | .93448 | .93574 | .93699 | .93822 | .93943 | .94 62 |
| +1.6 | .94520 | .94630 | .94738 | .94845 | .94950 | .95053 | .95 54 |
| +1.7 | .95543 | .95637 | .95728 | .95818 | .95907 | .95994 | .96 80 |
| +1.8 | .96407 | .96485 | .96562 | .96638 | .96712 | .96784 | [illegible] |
| +1.9 | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | .97500 |

**Adding both i.e. 1.9+.6=1.96**

**Area=(1-0.95)/2=0.975**



# Model Performance Range

| | Mean Accuracy | Std Dev of Accuracy | Best Accuracy | C.I. of 95% |
|---|---|---|---|---|
| Decision Tree | 0.819849 | 0.052824 | 0.904558 | 71.63% to 92.34% |
| XGBoost | 0.898732 | 0.033192 | 0.952706 | 83.37% to 96.38% |
| Gradient Boost | 0.846267 | 0.031706 | 0.905249 | 78.41% to 90.84% |
| Random Forest | 0.873148 | 0.032985 | 0.924269 | 80.85% to 93.78% |

Conclusion: We can conclude that the all these models gives performance between the specified range in 95 % of the cases