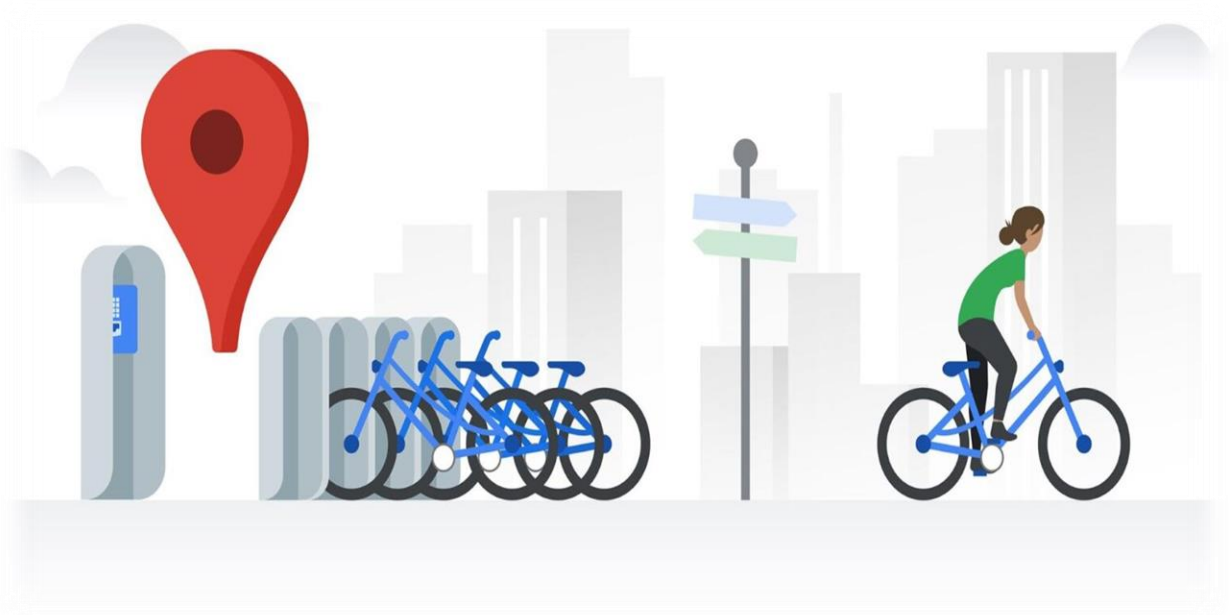
# CAPSTONE PROJECT

(SUPERVISED ML-REGRESSION)

## Seoul Bike Sharing Demand Prediction




**Team Members**

Mounika Dontula

Abhash Jain

# CONTENT

- Problem Statement
- Data Description
- Data Manipulation
- EDA-Analysis of Categorical Variables
- Analysis of Numerical Variables
- One Hot Encoding
- Regression Plot
- Machine Learning Algorithms(Regression)
- Comparison of different model
- Conclusion

# Problem Statement

Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

# Data Description

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

**Attribute Information:**

- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of the day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10m
- Dew point temperature - Celsius
- Solar radiation - MJ/m2
- Rainfall - mm

# Data Description….

•Snowfall - cm

•Seasons - Winter, Spring, Summer, Autumn

•Holiday - Holiday/No holiday

•Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

1. This dataset contains 8760 lines and 14 columns.
2. Numerical variables - temperature, humidity, wind, visibility, dew point temp, solar radiation, rainfall, snowfall.
3. Categorical variables - seasons, holiday and functioning day.
4. Rented bike column - which we need to predict for new observations.

# Data Manipulation

- **Handling null and duplicate values**
  - There are no null values and duplicate values in this data set.
- **Checking for Correlation**
  - With the help of heatmap correlation the correlation between Temperature and Dew point temperature are high.
- **Dropping some columns**
  - Due to high correlation between Temperature and Dew point temperature we are dropping the column of Dew Point Temperature.
- **Removing Outliers**
  - Some outliers are present in the data set so we are removing those outliers with the help of inter quartile range(IQR).
- **Handling null**
  - Initial there were no null values in this data set but after removing outliers data set consist null values those null values fillup by mean values of corresponding features.

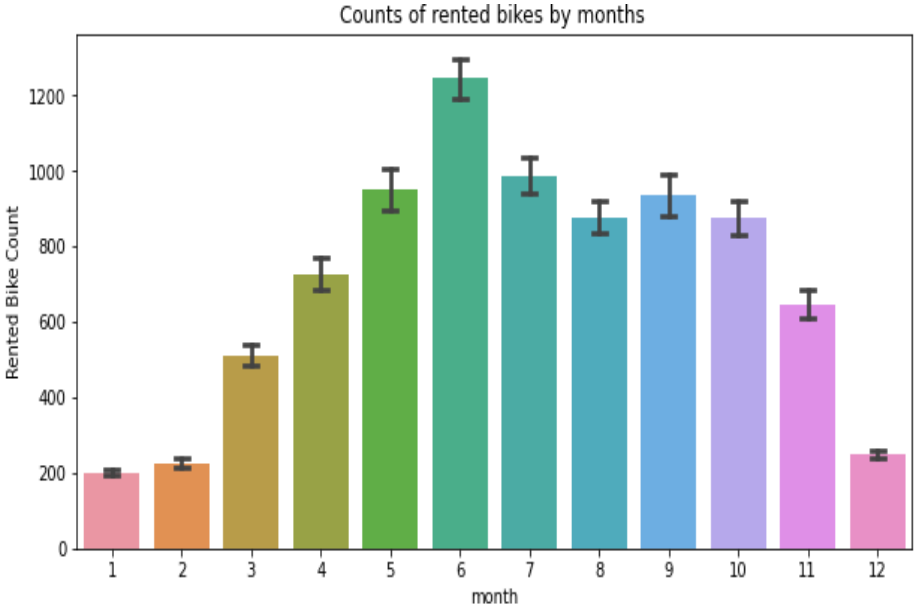
# EDA-Analysis of categorical variables

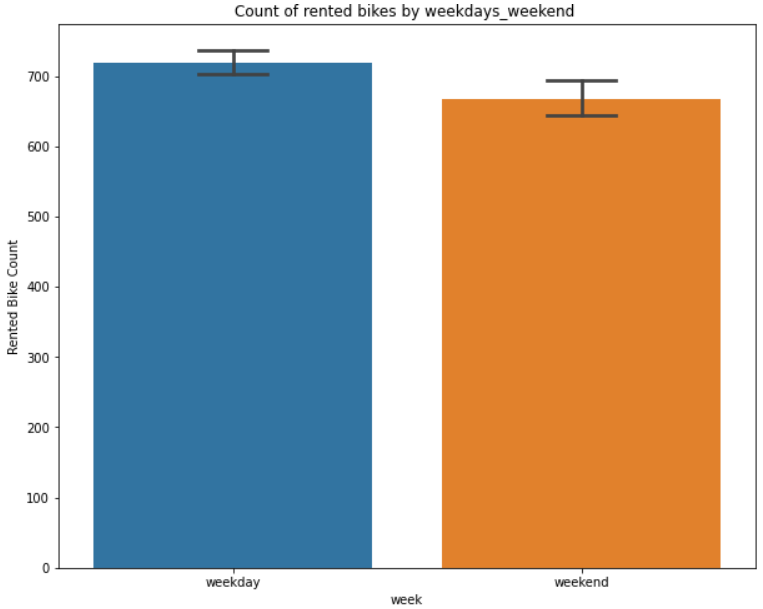
It is used to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.

**Rented Bike Count is a dependent or target variable.**

- Analysis of Rented Bike Count with respect to Month
- Analysis of Rented Bike Count with respect to Week
- Analysis of Rented Bike Count with respect to Hour
- Analysis of Rented Bike Count with respect to Functioning day
- Analysis of Rented Bike Count with respect to Holiday
- Analysis of Rented Bike Count with respect to Seasons

# EDA-Analysis of categorical variables....



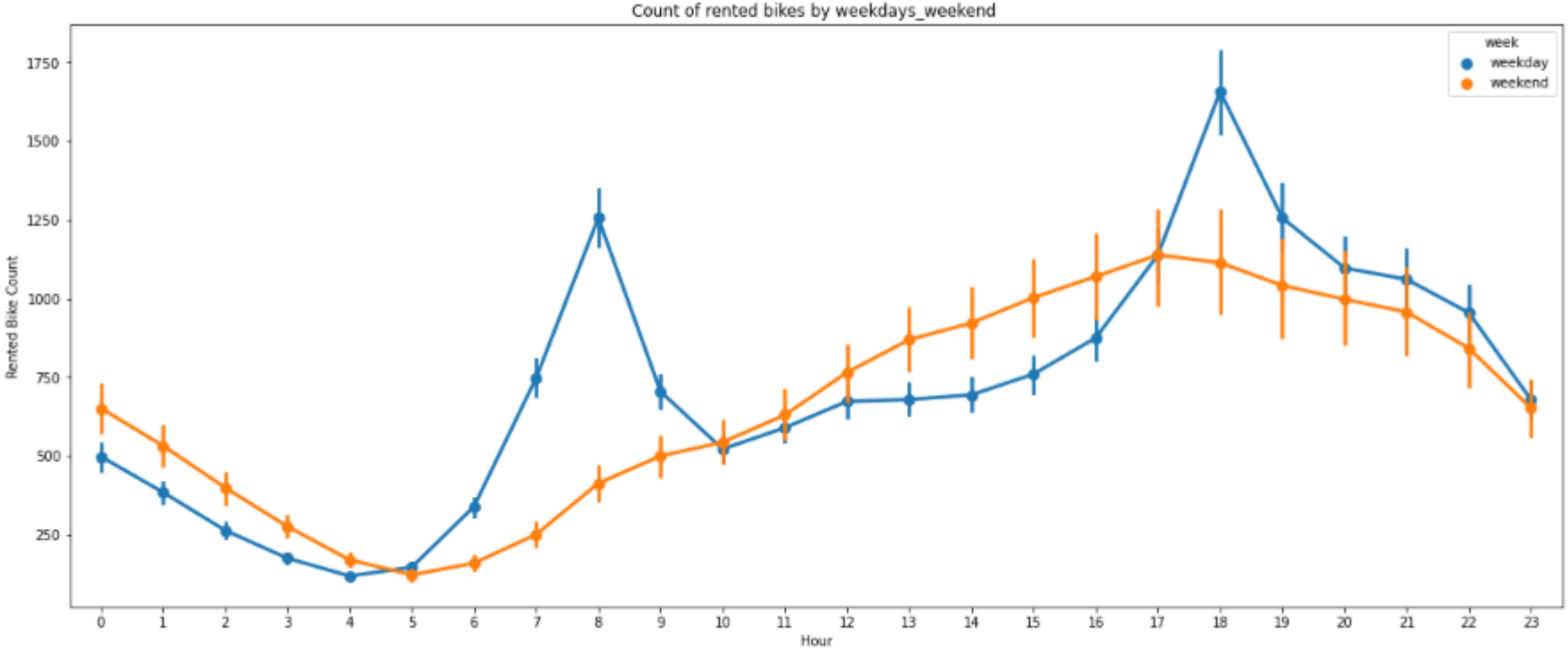

Rented Bike Count with respect to Month

With this Bar plot we can say that from the 5th month to 10th month the demand of the rented bike is high as compare to other months.

Rented Bike Count with respect to Week

With this bar plot we can say that in the week days which represent in blue color show that the demand of the bike higher because of the office working day.

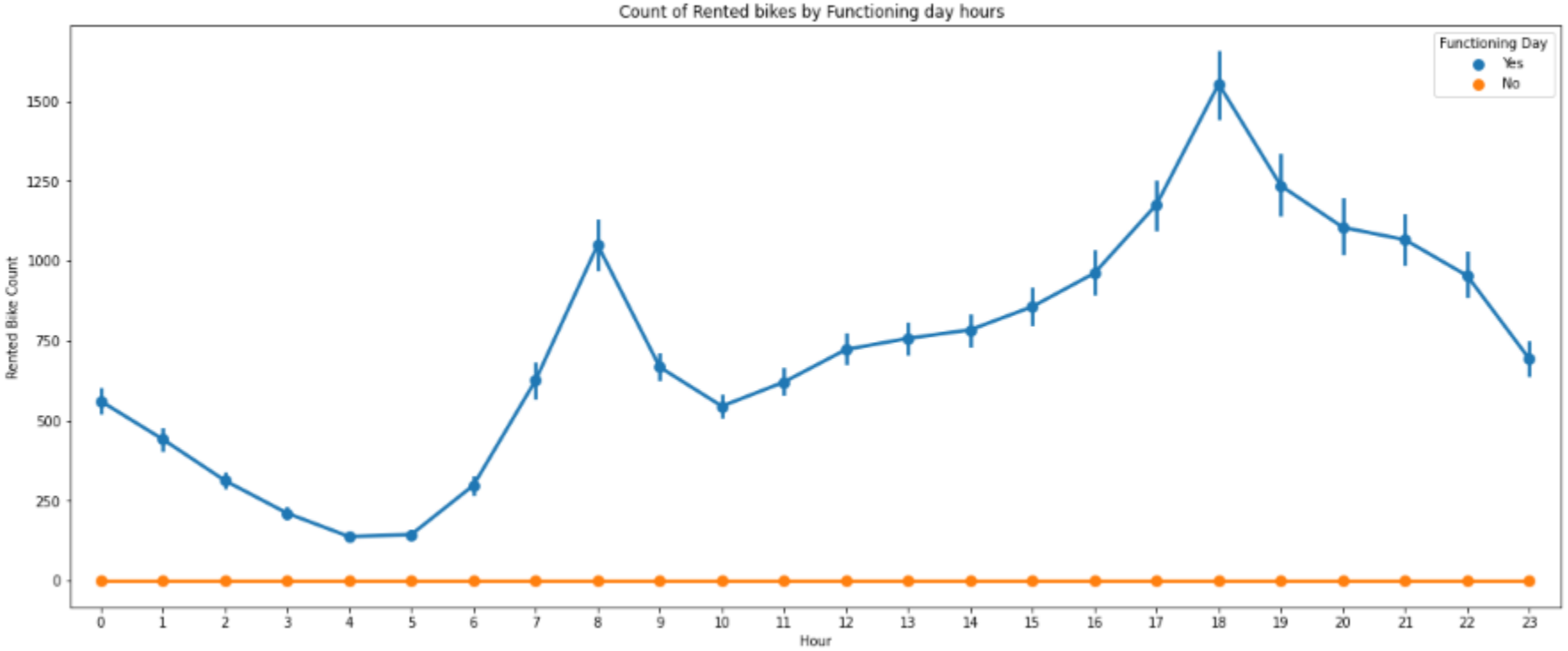
# EDA-Analysis of categorical variables....




Rented Bike Count with respect to Hour

From this point plot we can see that Peak Time are 7 am to 9 am and 5 pm to 7 pm and show that demand of rented bikes are very low in the morning hour but in the evening from 4 pm to 8 pm the demand increases.

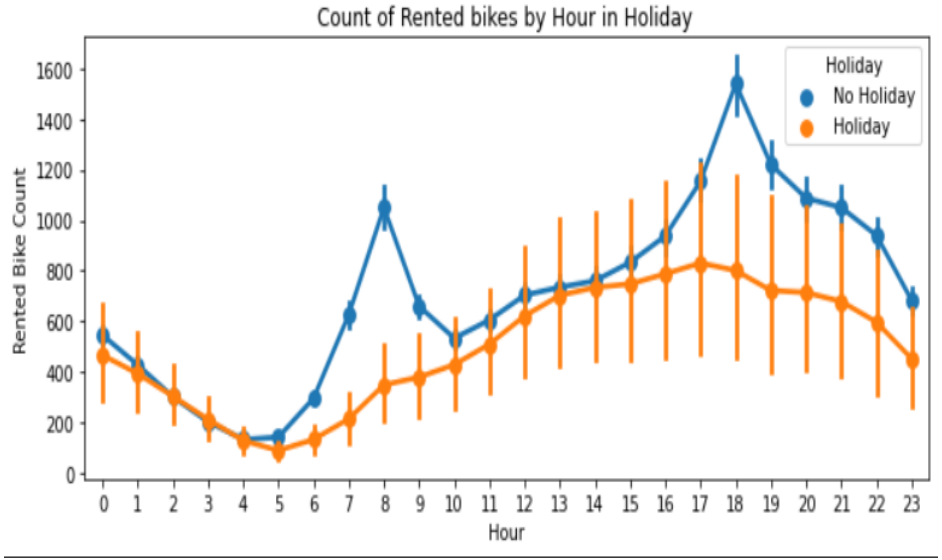
# EDA-Analysis of categorical variables....

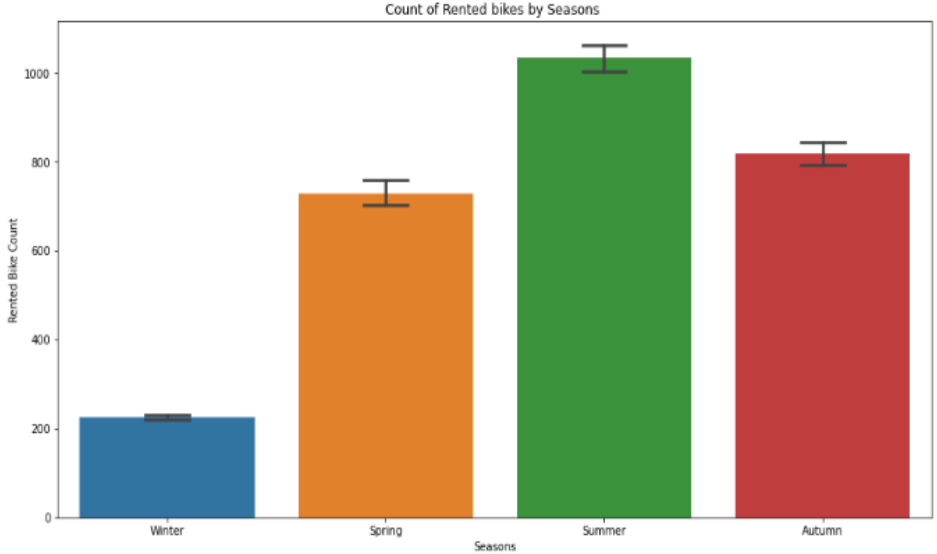



Rented Bike Count with respect to Functioning day

The point plot shows the use of rented bike in functioning day or not, and it also shows that, Peoples do not use rented bikes in no functioning day.

# EDA-Analysis of categorical variables....




## Analysis of Rented Bike Count with respect to Holiday

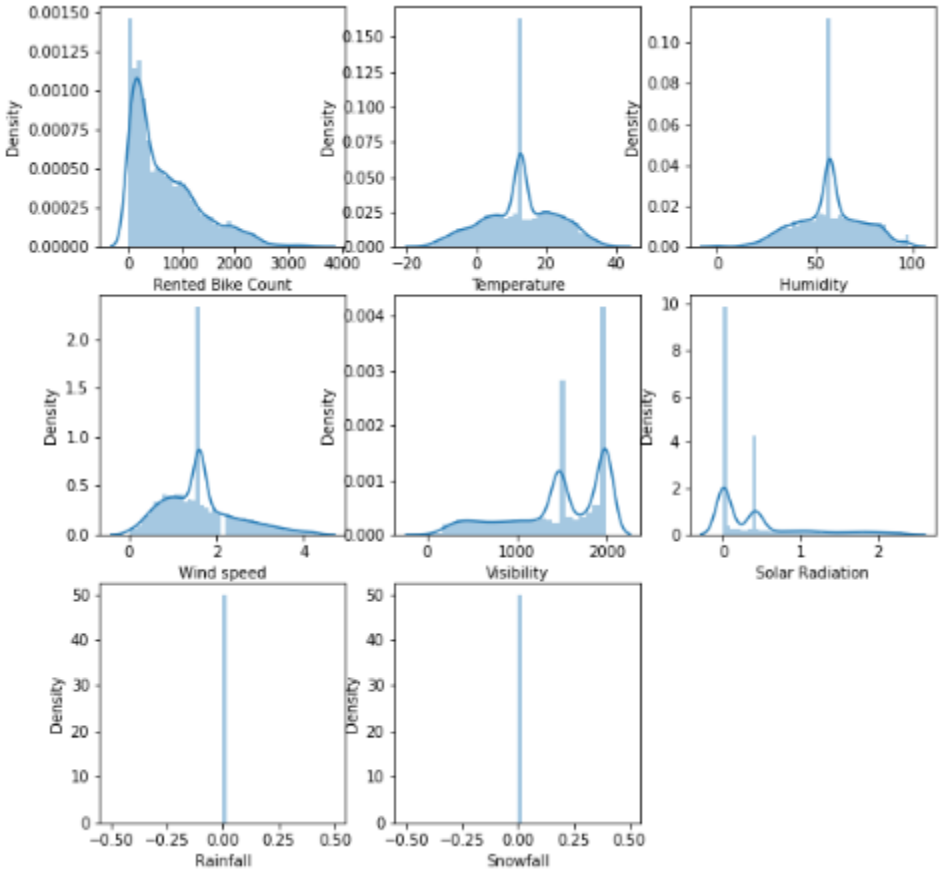
The point plot shows the use of rented bike in a holiday, and it also shows that in holiday people uses the rented bike from 2pm-8pm.

## Analysis of Rented Bike Count with respect to Seasons

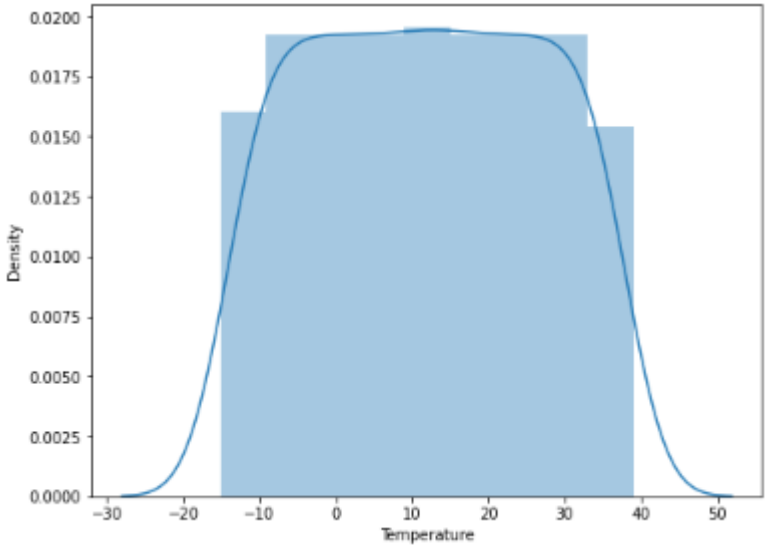
The bar plot shows the use of rented bike in in four different seasons, and it also shows that in summer season the use of rented bike is high and in winter season the use of rented bike is very low because of cold and snowfall.

# Analysis of Numerical Variables

From this distplot we can observe the density distribution among the columns like Rented Bike Count, Temperature, Humidity, Wind Speed, Visibility, Solar Radiation, rainfall and Snowfall.

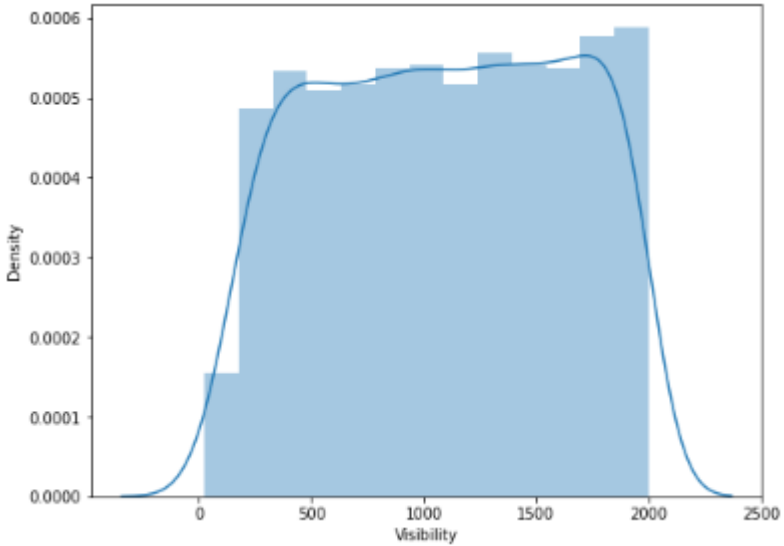


# Analysis of Numerical Variables




**Count of bikes rented in different temperature**

The plot shows that people tend to rent bikes when the temperature is between -5 to 25 degrees.




**Count of bikes rented in different visibility ranges**

The plot shows that people tend to rent bikes when the visibility is between 300 to 1700.

# One Hot Encoding

The main aim of One Hot encoding is to produce binary integers of 0 and 1 to encode our categorical features.

| | Date | Rented Bike Count | Hour | Temperature | Humidity | Wind speed | Visibility | Solar Radiation | Rainfall | Snowfall | ... | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | weekday | weekend |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 01/12/2017 | 254 | 0 | -5.2 | 37.0 | 2.2 | 2000.0 | 0.0 | 0.0 | 0.0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 1 | 01/12/2017 | 204 | 1 | -5.5 | 38.0 | 0.8 | 2000.0 | 0.0 | 0.0 | 0.0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 2 | 01/12/2017 | 173 | 2 | -6.0 | 39.0 | 1.0 | 2000.0 | 0.0 | 0.0 | 0.0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 3 | 01/12/2017 | 107 | 3 | -6.2 | 40.0 | 0.9 | 2000.0 | 0.0 | 0.0 | 0.0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| 4 | 01/12/2017 | 78 | 4 | -6.0 | 36.0 | 2.3 | 2000.0 | 0.0 | 0.0 | 0.0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |

5 rows × 37 columns

We can observe that all the values of each column is converted to numerical values(binary values).

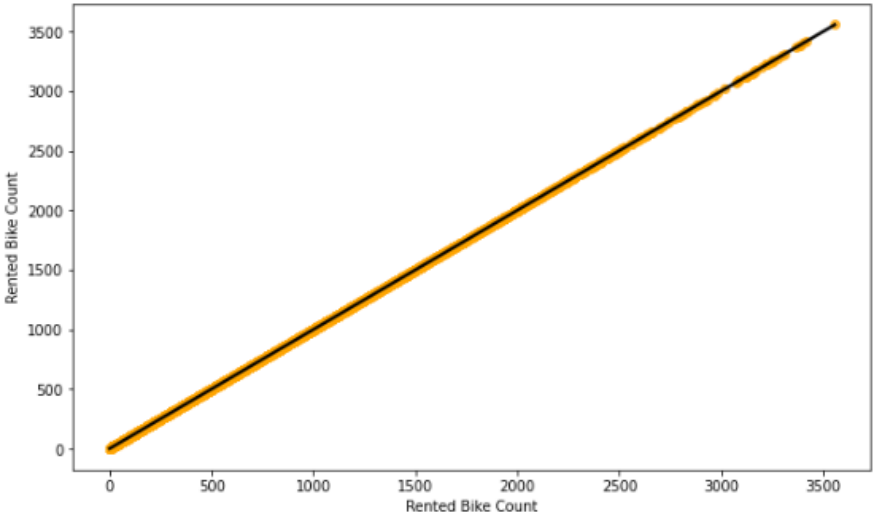
# Check for Multicollinearity

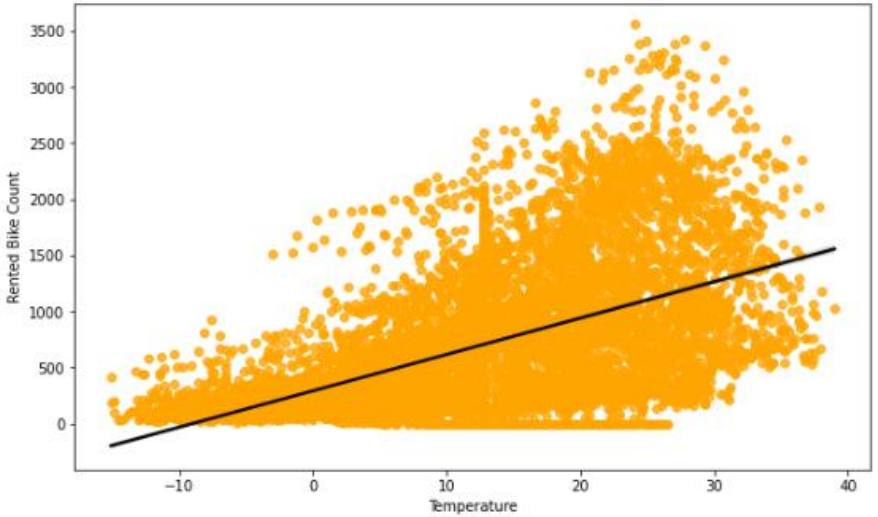
- A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model.
- We can see the variance inflation factor for various columns features.

| | feature | VIF |
|---|---|---|
| 0 | Rented Bike Count | 2.22 |
| 1 | Hour | 1.41 |
| 2 | Temperature | 3.97 |
| 3 | Humidity | 2.33 |
| 4 | Wind speed | 1.29 |
| 5 | Visibility | 1.75 |
| 6 | Solar Radiation | 1.62 |
| 7 | Rainfall | NaN |
| 8 | Snowfall | NaN |
| 9 | Autumn | inf |
| 10 | Spring | inf |
| 11 | Summer | inf |
| 12 | Winter | inf |
| 13 | No Holiday | 1.05 |
| 14 | No | inf |
| 15 | Yes | inf |
| 16 | 1 | inf |
| 17 | 2 | inf |
| 18 | 3 | inf |
| 19 | 4 | inf |
| 20 | 5 | inf |
| 21 | 6 | inf |
| 22 | 7 | inf |
| 23 | 8 | inf |
| 24 | 9 | inf |
| 25 | 10 | inf |
| 26 | 11 | inf |
| 27 | 12 | inf |
| 28 | weekday | inf |
| 29 | weekend | inf |

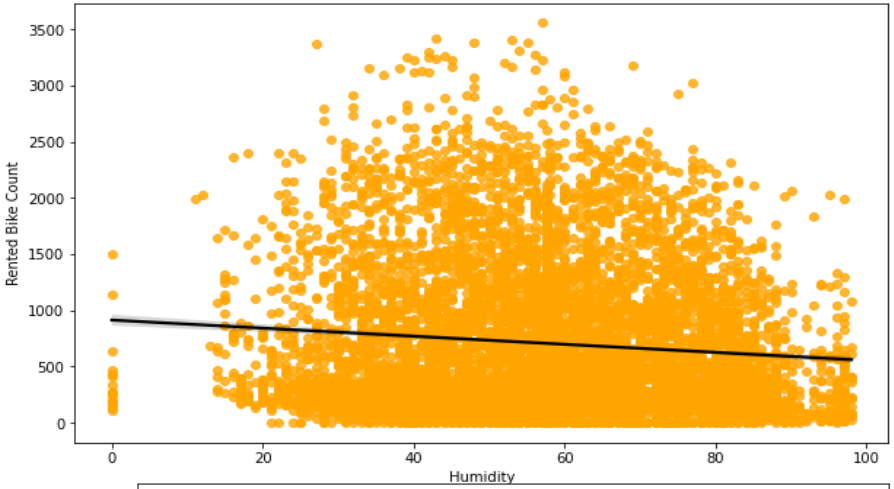
# Regression Plots

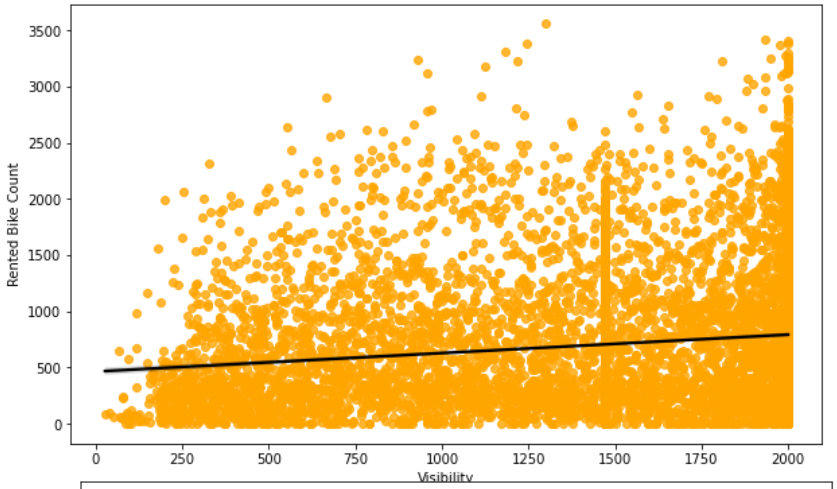
Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variable.

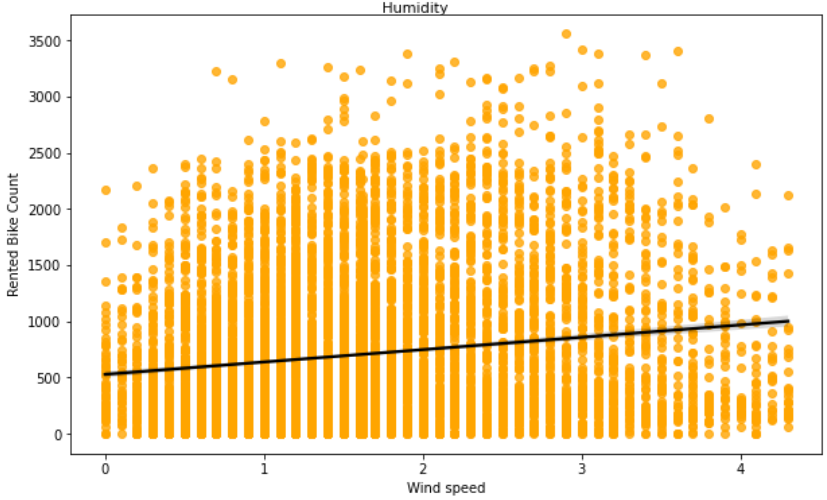


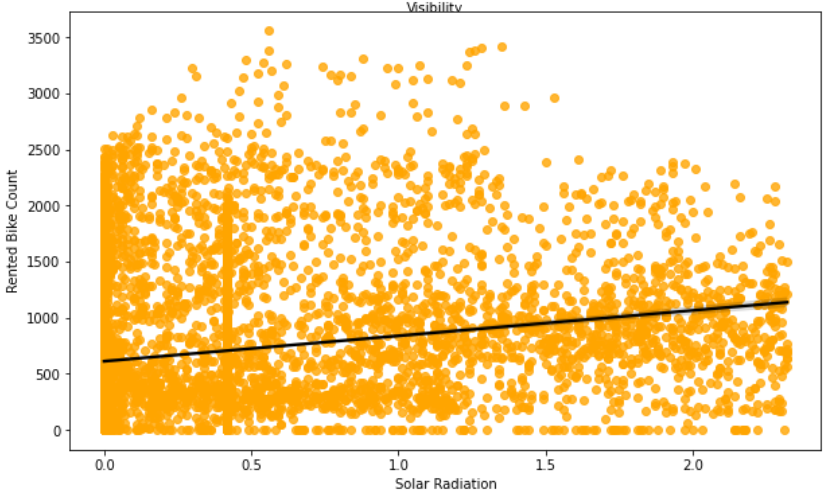



# Regression Plots….



# Machine Learning Algorithms(Regression)

These are the algorithms we have taken to convert the data set in to a model.

- **Linear Regression Model(L1 and L2 Regularisation)**
- **linear regression with elastic net**
- **Decision tree with grid search cv**
- **Random Forest Regressor with GridSearchCV**
- **Gradient Boosting Regressor with GridSearchCV**

# Linear Regression Model(L1 and L2 Regularisation)

## L1 Regularization :

A linear regression model that implements L1 norm for regularisation is called **lasso regression .** It adds the “absolute value of magnitude” of the coefficient as a penalty term to the loss function.

## L2 Regularisation :

A linear regression model that implements (squared) L2 norm for regularisation is called **ridge regression**. It adds the “squared magnitude” of the coefficient as the penalty term to the loss function.

```
MSE : 61.132589005575994
RMSE : 7.818733209771004
MAE : 5.911752830176522
R2 : 0.6030111076694047
Adjusted R2 : 0.5980533370436295
```

L1 Regularisation(Lasso Regression)
MSE,RMSE,MAE,R2,Adjusted R2 are not at its best in Lasso Regression.

```
MSE : 60.83384712272395
RMSE : 7.799605574817482
R2 : 0.6148937632653055
Adjusted R2 :  0.6100843884309686
```

L2 Regularisation(Ridge Regression)
The MSE,RMSE,R2,Adjusted R2 are also not at its best in Ridge Regression.

# Linear regression with elastic net

- Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models.
- The technique combines both the lasso and ridge regression methods by learning from their shortcomings to improve the regularization of statistical models.

```
MSE : 61.369571181250464
RMSE : 7.8338733191985215
MAE : 5.93382066577593
R2 : 0.6014721692250583
```

We can observe that Linear regression with Elastic Net is also not at its best compared to lasso and ridge regression.

# Decision tree with grid search cv

- Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output.
- Decision trees are prone to overfitting. To overcome this issue, we need to carefully adjust the hyperparameters of decision trees.we use Grid search cv for selecting best hyperparameters of decision trees.

```
The r2 score of decision tree is 0.7619882672759375
the r2 score of decision tree with hyper perameteres tunning is 0.8013257980814106
```

The R2 score for decision tree is low when compared to decision tree with hyperparameters.

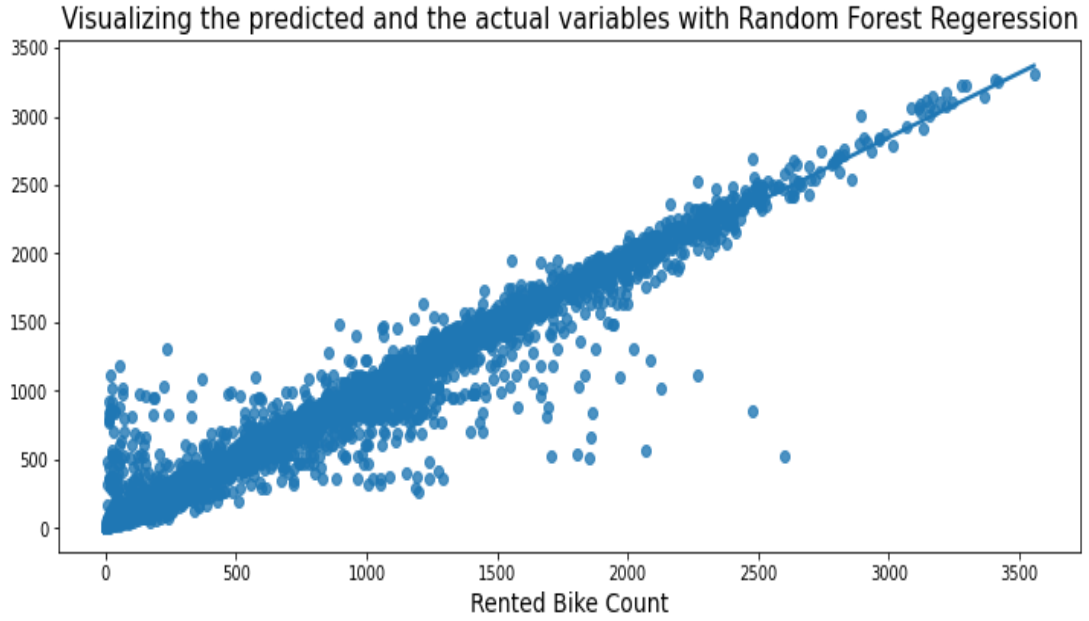
# Random Forest Regressor

- Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees
- One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression.

```
MSE= 7.548469131956553
RMSE= 2.747447748721812
R2_Score_train= 0.9509809996888273
MSE= 19.900384576287564
RMSE= 4.460984709264039
R2_Score_test= 0.8740214111679199
```

- We can observe that R2 score for training data is greater than the testing data.

The regplot shows the actual and predicted values with Random Forest Regression which is at its best.



# Gradient Boosting Regressor

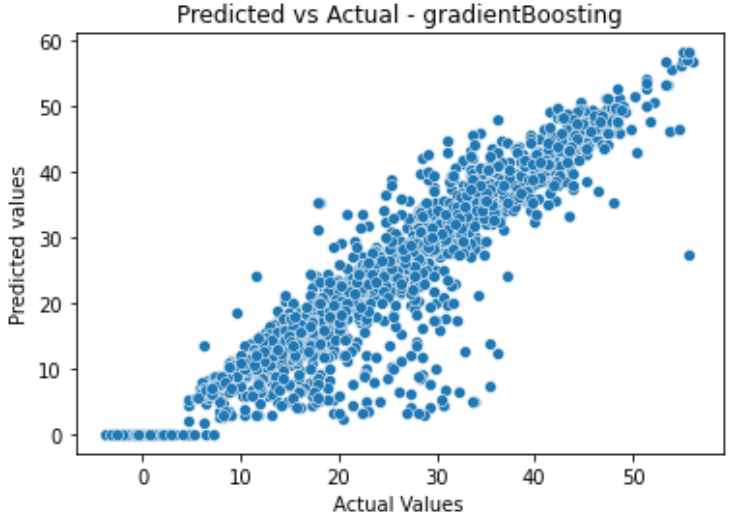
- Gradient boosting Regression calculates the difference between the current prediction and the known correct target value. This difference is called residual.
- This model creates a forest of 1000 trees with maximum depth of 3 and least square loss. The hyperparameters used for training the models are the following: n_estimators: Number of trees used for boosting. max_depth: Maximum depth of the tree.

```
r2_score(y_train, y_pred_train_g_g)
```

```
0.9106108001677612
```

```
gradient=r2_score(y_test, y_pred_g_g)
gradient
```

```
0.8694957936148591
```




- We can observe the R2 score for training data is higher than the testing data.

The scatter plot shows the Predicted Vs Actual values with Gradient Boosting Regressor.

# Comparison of different model

- We are comparing the R2 score for the models we used In Regression.
- R2 score is **(total variance explained by model) / total variance**.” So if it is 100%, the two variables are perfectly correlated, i.e., with no variance at all.

| | Lasso_model | Ridge_model | Random forest | Gradient Boosting | DecisionTree | elasticnet |
|---|---|---|---|---|---|---|
| r2_value | 0.603011 | 0.614894 | 0.874021 | 0.869496 | 0.801326 | 0.601472 |

# Conclusion

- Hour of the day holds most importance among all the features for prediction of dataset.
- It is observed that highest number bike rentals counts in Autumn/fall Summer Seasons and the lowest in Spring season.
- We observed that the highest number of bike rentals on a clear day and the lowest on a snowy or rainy day.
- As we can see the top 5 important features of our dataset are: Season_winter, Temperature, Hour, Season_autumn and Humidity.
- Peoples do not use rented bikes in no functioning day.
- People tend to rent bikes when the temperature is between -5 to 25 degrees.
- People tend to rent bikes when the visibility is between 300 to 1700.
- Linear Regression, Lasso and Ridge are not at its best.
- The above experiments we can conclude that gradient boosting and random forest regressor with using hyperparameters we got the best results.


Thank you