Capstone Project-2

Seoul Bike Sharing Demand Prediction

Team Members:

Jasjot Kaur

Venkatesh A.

Parul Saini

Flow of Presentation

- Objective
- Data Summary
- Exploratory Data Analysis
- Data Preprocessing
- ML models
- Conclusion

Objective:

- For the enhancement of mobility comfort, rental bikes have been introduced in many urban cities.
- To facilitate the ease at the fullest, it is important to that the rental bikes are available and accessible to the public at the right time; also reducing the waiting time.
- Eventually, providing the city with a stable supply of rental bikes becomes a major concern.
- Thus, the prediction of bike count required at each hour is crucial for a stable supply of rental bikes.
- The objective of this project is **to build a ML model which is optimal and which is able to predict the bike count required based on available features.**

Workflow

- Null values
- Duplicated values
- Outliers
- Categorical data

- Answering hypothesized question

- Finding correlation
- Statistical modelling for feature selection

- Linear Regression
- Lasso Regression
- Ridge Regression
- Elastic-Net Regression
- Polynomial Regression
- Random Forest Regressor
- Gradient Boosting Regressor

- Low Bias
- Low Variance

Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables.

Data Summary

The dataset originally contains 14 features for 8760 records; the readings of these 14 parameters were recorded on different days and time of a day. The description of the columns are as follows:

- Date : The date of the day, type: str
- Rented Bike Count - Number of rented bikes per hour and it is also a dependent variable, type: int
- Hour - Hour of the day ranging from 0-23, type: int
- Temperature (°C)- Temperature in Celsius, type: float
- Humidity(%) - Humidity in the air in %, type: int
- Wind speed (m/s) - Speed of the wind in m/s, type: float
- Visibility (10m) - Visibility in m, type: int
- Dew point temperature(°C) - The temperature at which the water starts to condense out of the air, type: float
- Solar Radiation (MJ/m2) - Electromagnetic radiation emitted by the Sun, type: float
- Rainfall(mm) - Amount of rainfall in mm, type: float italicized text
- Snowfall(cm) - Amount of snowfall in cm, type: float
- Seasons - Season of the year, type: str
- Holiday - If the day is holiday or not, type: str
- Functioning Day - Whether the day is functional or not, type: str

Data Cleaning

- Null and duplicate values were not found in the dataset,
- There were no structural errors in the dataset,
- New columns were made using the existing columns and the irrelevant columns were removed,
- Datatypes of some columns were changed as per requirement

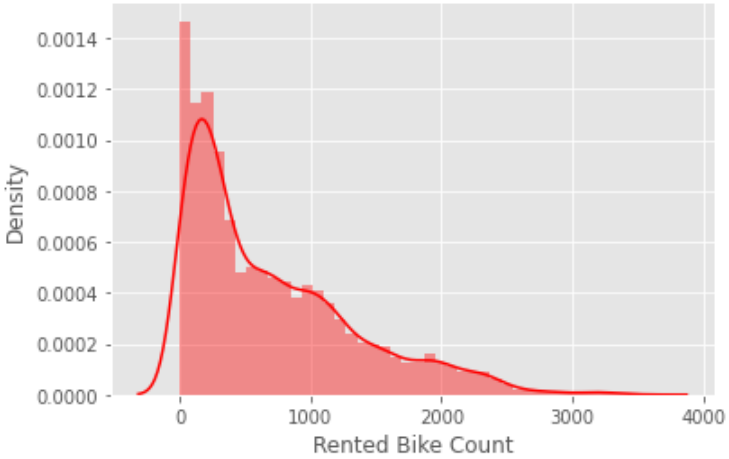
For example, the '*date*', '*month*' and '*year*' columns were formed from '*Date*' column, as shown below:

(dropping the '*Date'* column later)

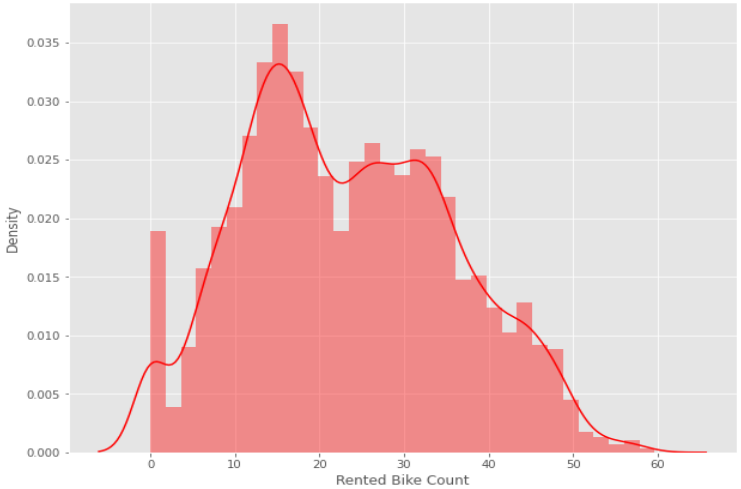
```
# Date columns to Date format conversion

df_b['Date']= pd.to_datetime(df_b['Date'])
```

```
df_b["date"] = pd.DatetimeIndex(df_b["Date"]).day
df_b['month'] = df_b['Date'].apply(lambda x : x.month)
df_b["year"] = pd.DatetimeIndex(df_b["Date"]).year
```

Data Preprocessing



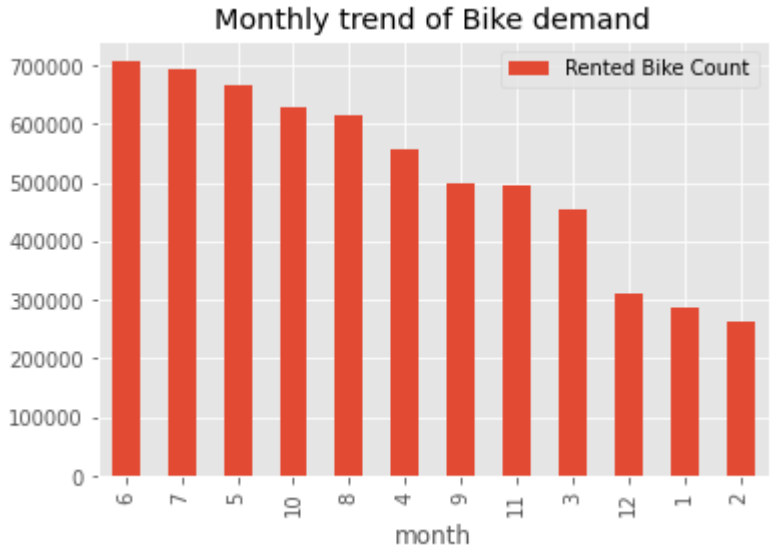
Right Skewed

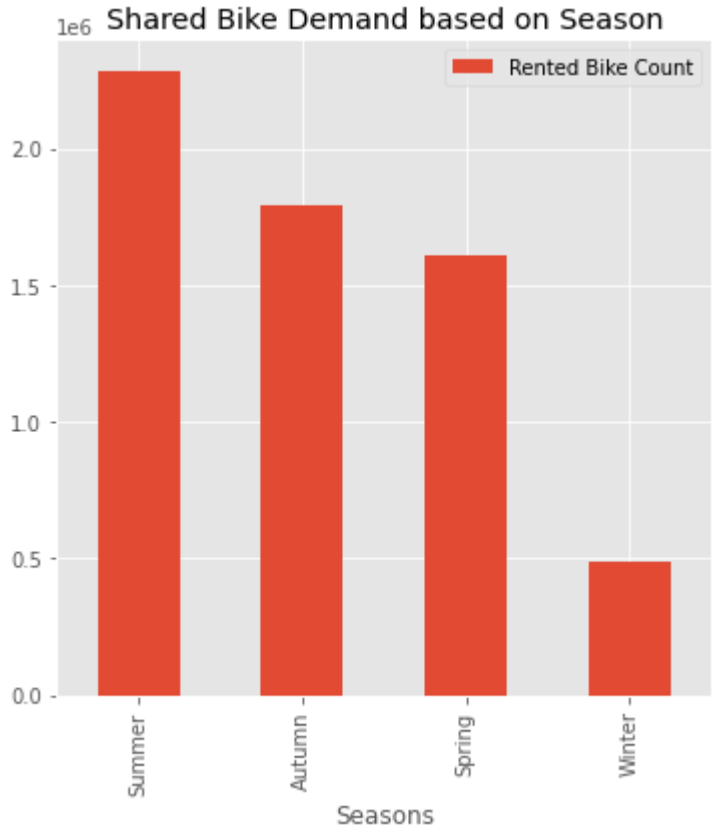


Normalised

- For the dependent variable: Rented Bike Count, ‘Square root’ transformation was applied to convert the positively skewed distribution to normal distribution.
- Normal distribution (also called as Gaussian distribution) says that the data should be normally distributed in nature.
- The shape of the distribution appears to be bell curve because the data frequency decreases as we move further away from the mean value to the either extremes the data.

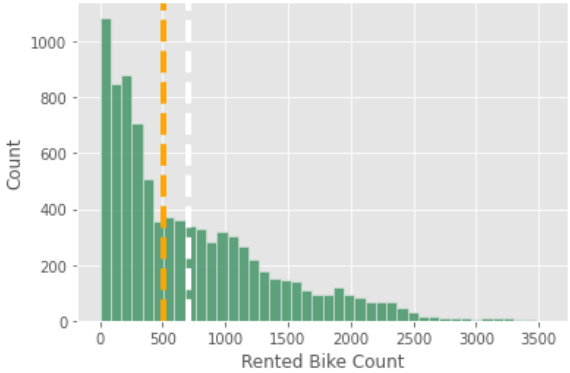
Exploratory Data Analysis

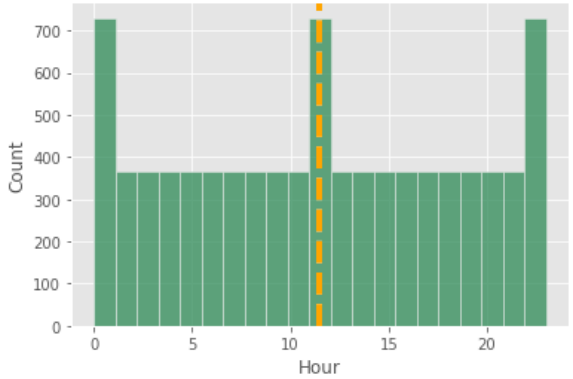


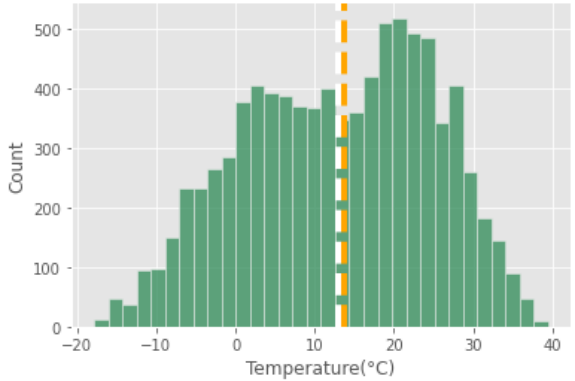


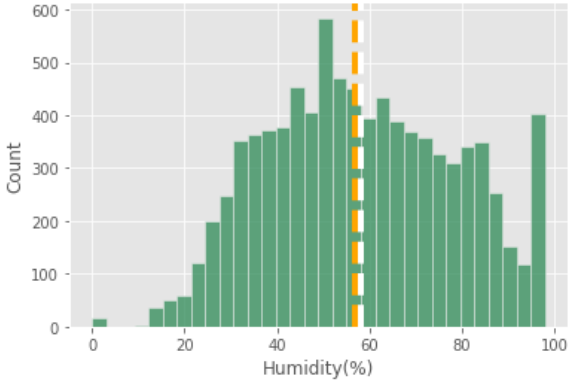
- It was found that most of the bikes were rented in Summer season during the months of May, June, July; while least bikes were rented in Winter season in the months of December, January and February.

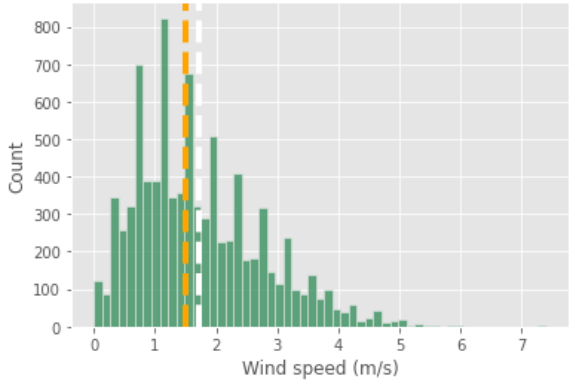
Visualizing Numeric Features:

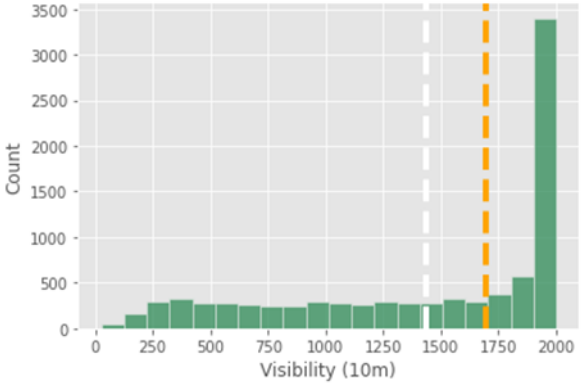




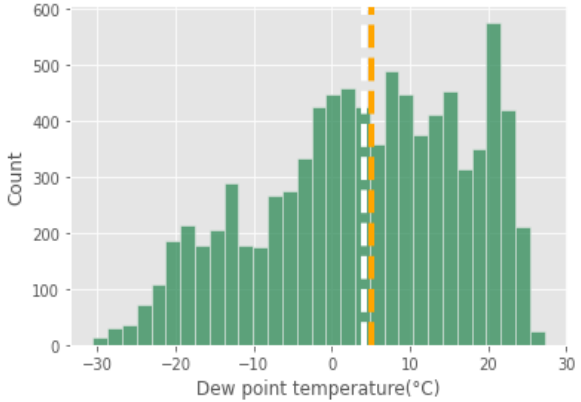


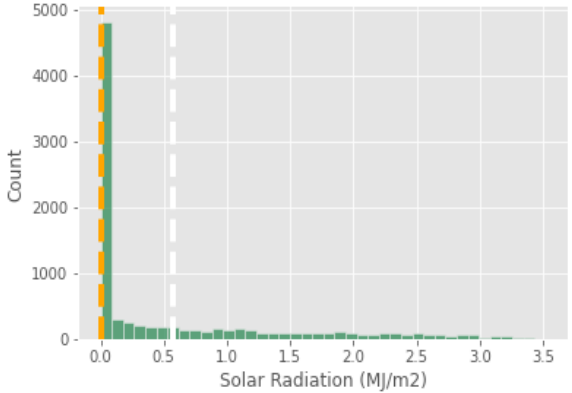


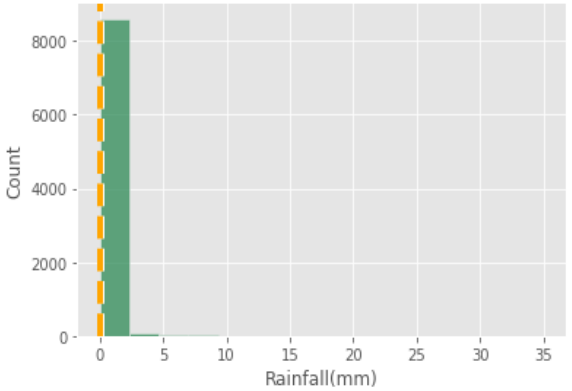


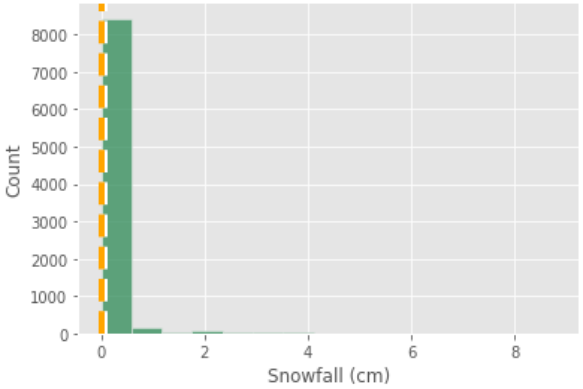


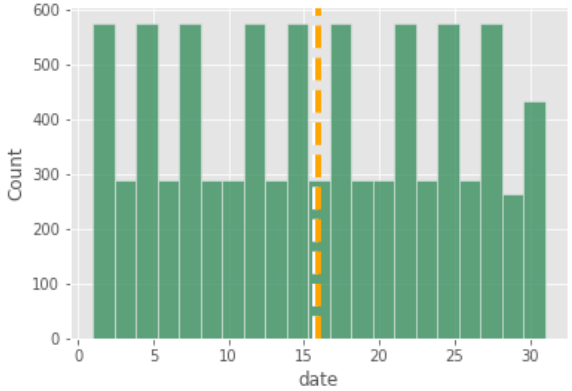
Visualizing Numeric Features:

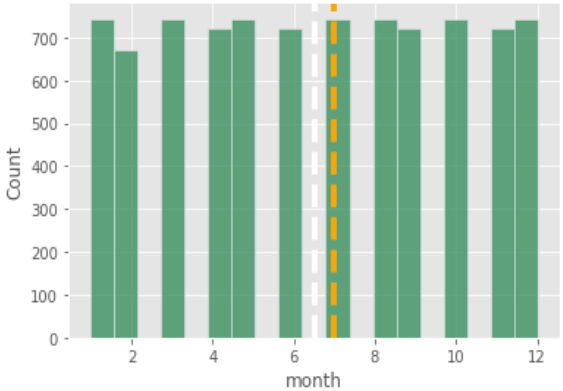




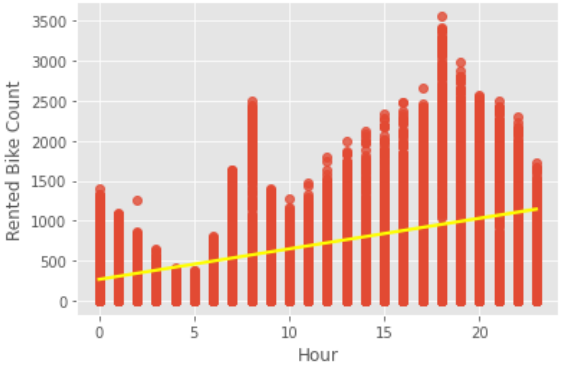


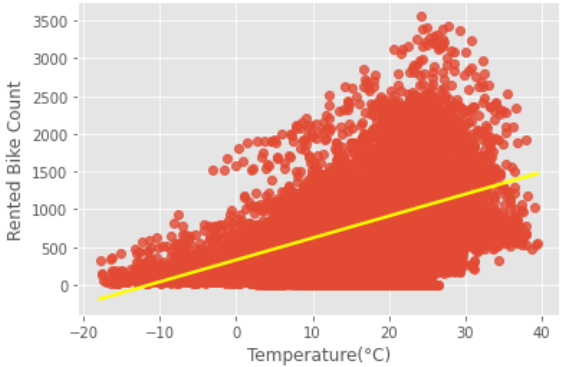


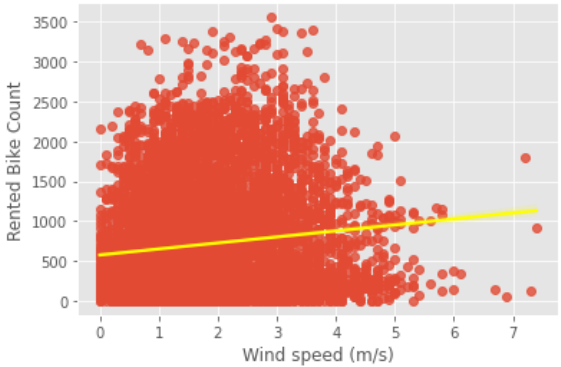


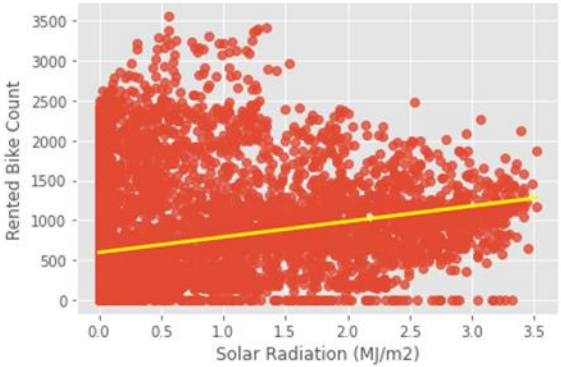


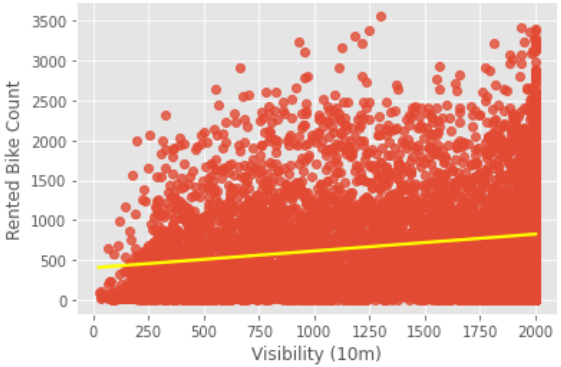
Relationship between Dependent and Independent (Numerical) variables:

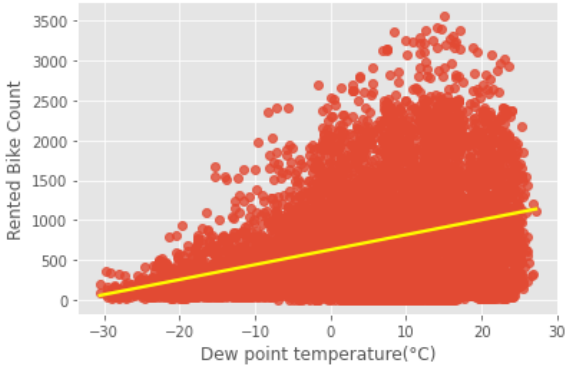




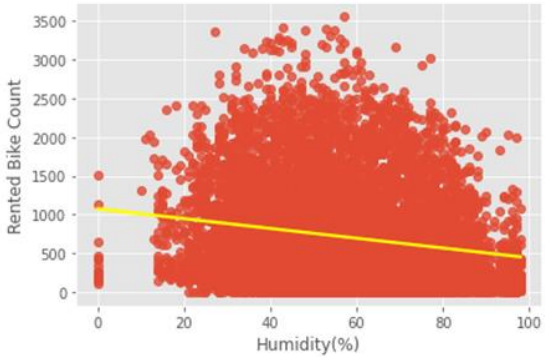


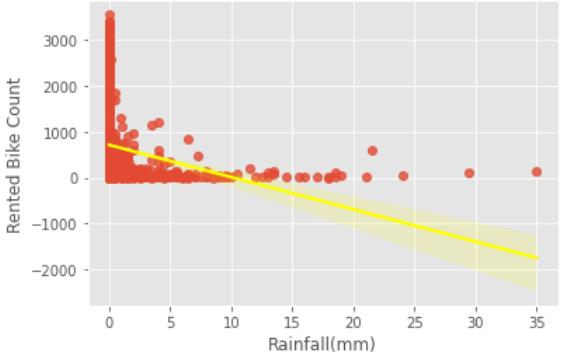


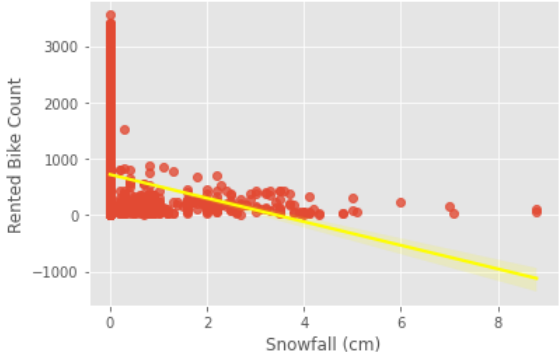


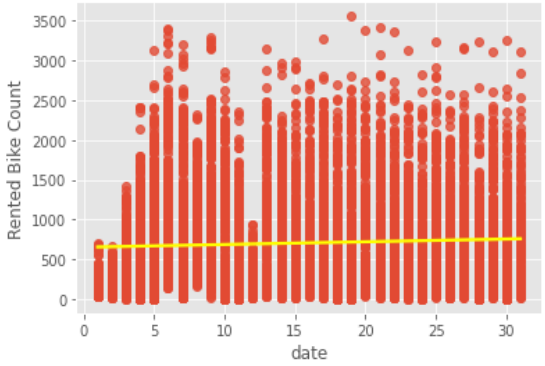


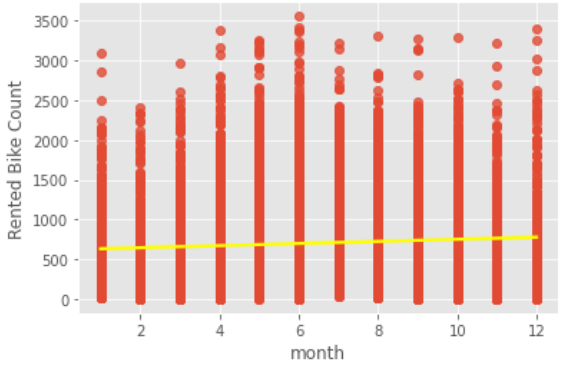
Relationship between Dependent and Independent (Numerical) variables:

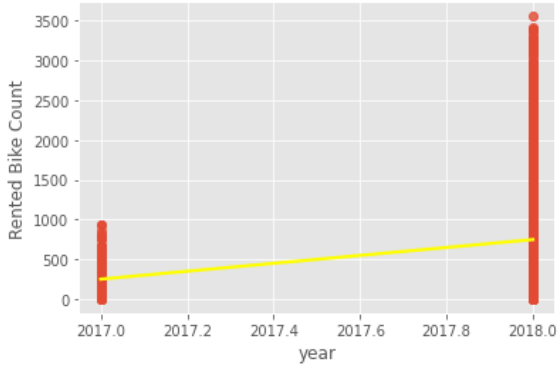






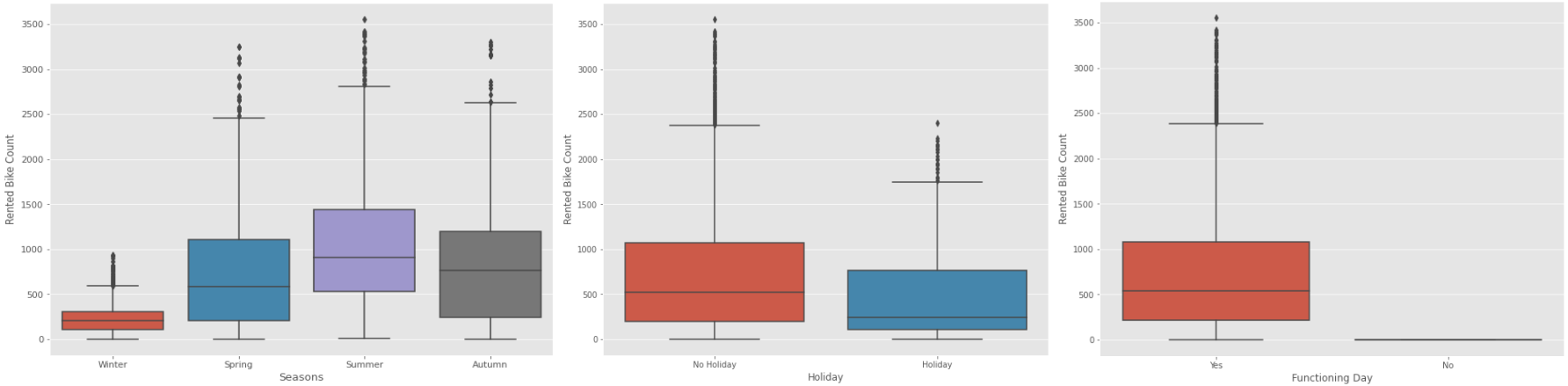






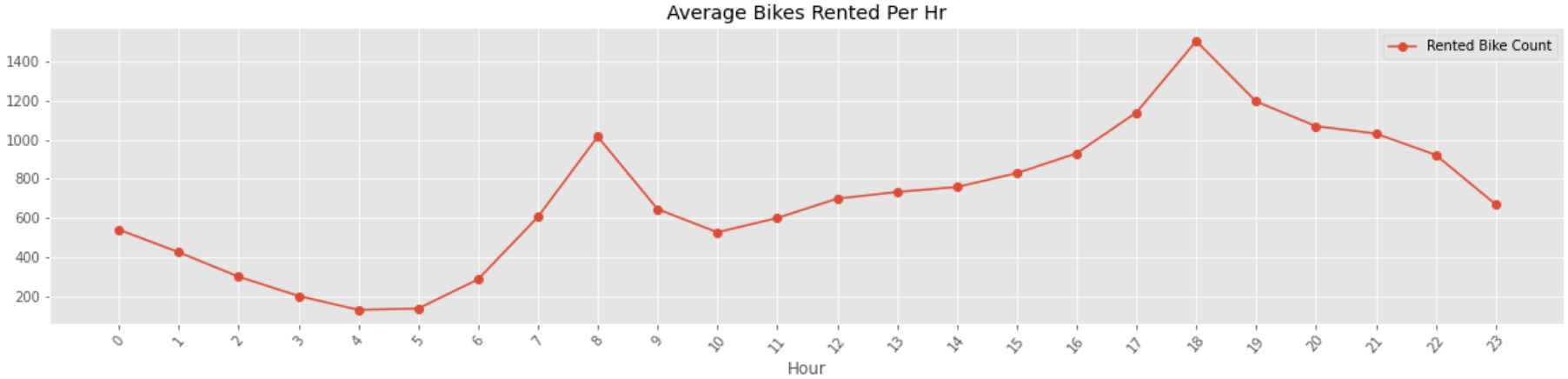
EDA

Relationship between Dependent and Independent (Categorical) variables:



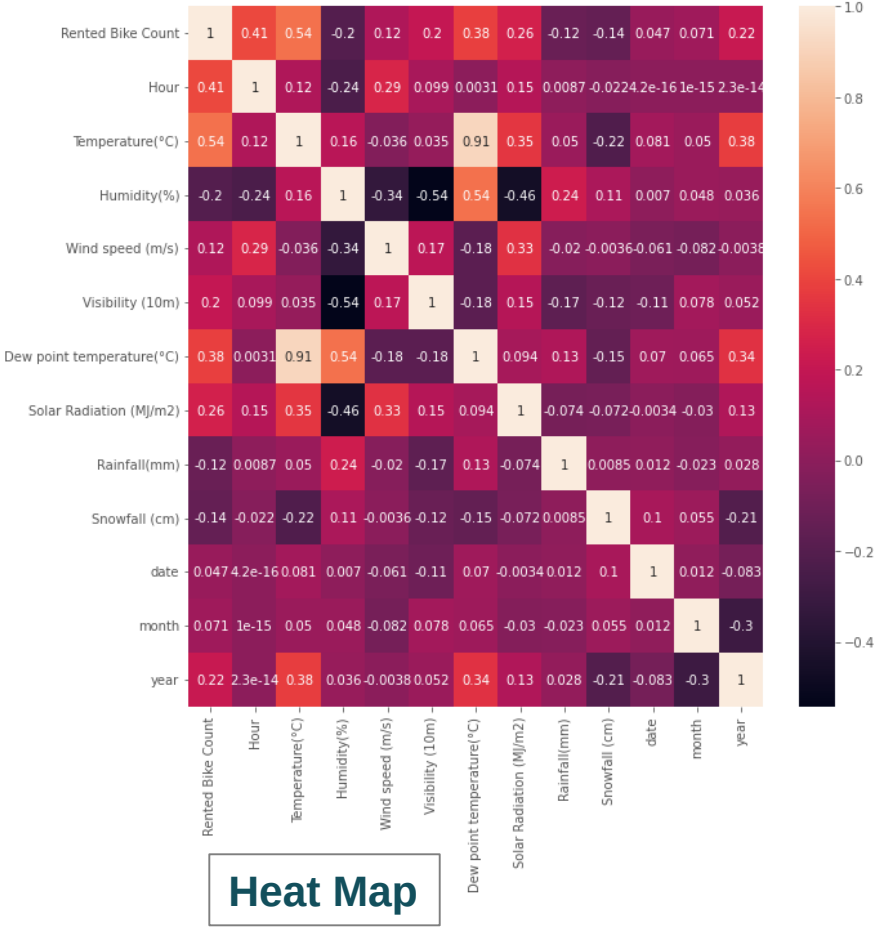
- We can clearly see that there is less demand of rented bike during winter season.
- While bikes were rented both on holidays and non-holiday (working) days, the rental bike demands was seen more during working days.
- The functioning day recorded to have most bikes rented; this may indicate that either the data have been collected on the functioning day or there was no service of bike rentals during non-functioning days.

Average bikes rented per hour:



- There was high rise of rented Bikes from 8:00 a.m to 9:00 p.m; indicating that people preferred rented bike during rush hours.
- We can clearly see that rental bike demand peaked at 8 a.m in the morning and 6:00 p.m. in the evening. So we can say that that during office opening and closing time there is much high demand of bike rentals.

Correlation Visualisation of Features:



Heat Map

From the heat map and the calculated VIF it was observed that there is high correlation between temperature, humidity and dew point temperature.

VIF of features after dropping 'Rented Bike Count', 'year' and 'Dew point temperature(°C)'

	variables	VIF
0	Hour	3.998419
1	Temperature(°C)	3.236167
2	Humidity(%)	6.757926
3	Wind speed (m/s)	4.621365
4	Visibility (10m)	5.455330
5	Solar Radiation (MJ/m2)	2.280208
6	Rainfall(mm)	1.081555
7	Snowfall (cm)	1.136671
8	date	3.849545
9	month	4.603431

Dummy Variables using One-Hot Encoding:

```
#creating Dummy variable for categorical columns; One-Hot Encoding

dummy_categories= pd.get_dummies(categorical_features)
```

```
dummy_categories
```

	Seasons_Autumn	Seasons_Spring	Seasons_Summer	Seasons_Winter	Holiday_Holiday	Holiday_No Holiday	Functioning Day_No	Functioning Day_Yes
0	0	0	0	1	0	1	0	1
1	0	0	0	1	0	1	0	1
2	0	0	0	1	0	1	0	1

- Creating dummy variables for categorical variables: ‘Seasons’, ‘Holiday’ and ‘Functioning day’ using one hot encoding
- A one hot encoding allows the representation of categorical data to be more expressive.
- Many machine learning algorithms cannot work with categorical data directly.
- The categories must be converted into numbers.
- This is required for both input and output variables that are categorical.

Splitting the data into Train and Test

```
#splitting

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=0)
```

Scaling the data

```
# Transforming data

scaler = MinMaxScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

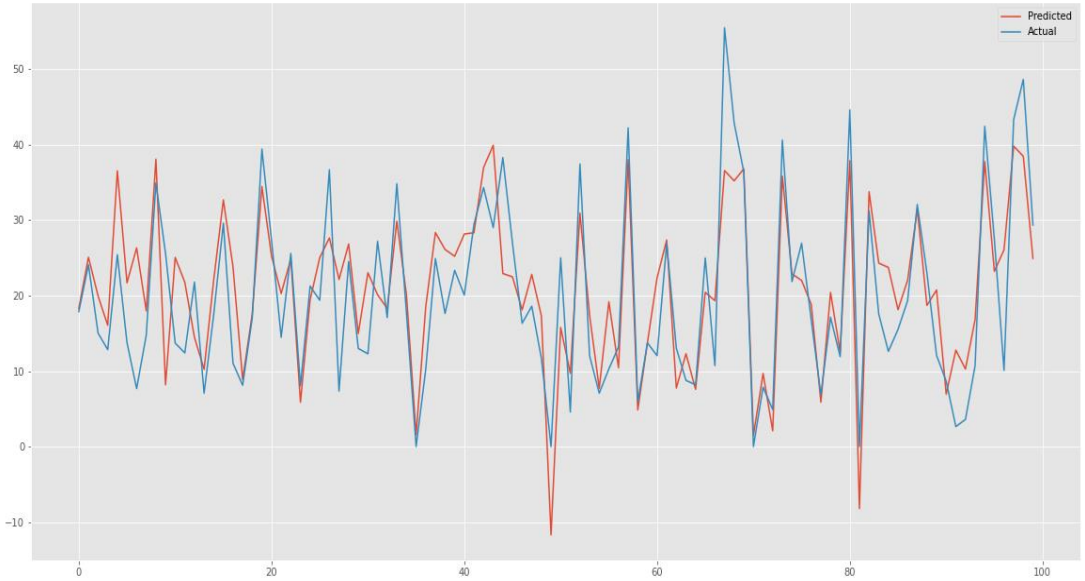
Models used

- Linear Regression
- Lasso Regression
- Ridge Regression
- Elastic-Net Regression
- Polynomial Regression
- Random Forest Regressor
- Gradient Boosting Regressor

To improve models' performance Hyperparameter tuning was done using GridSearchCV

1. Linear Regression

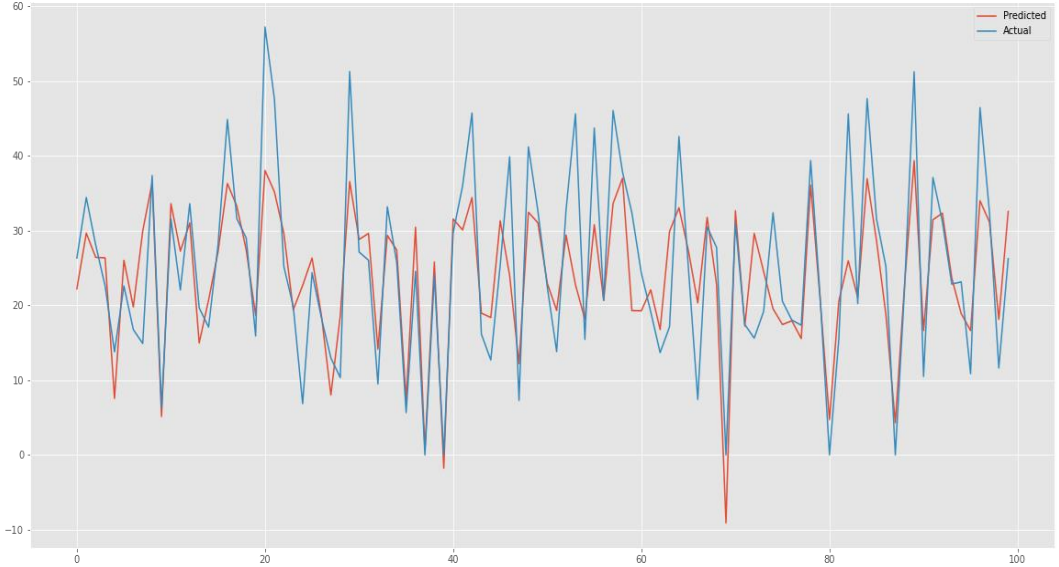
Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent and independent variable, thus predicting the value of a variable based on the value of another variable.




```
MSE : 173790.18680622437
RMSE : 416.8815021156784
R2 : 0.5833466162751417
Adjusted R2 :
0.5800855170470927
```

2. Lasso Regression

Least Absolute Shrinkage and Selection Operator



L1 Regression is an extension of linear regression that adds a regularization penalty to the loss function during training. It performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.

```
MSE : 173932.50237390096
RMSE : 417.0521578578643
R2 : 0.5830054217352261
Adjusted R2 :  0.5797416520158425
```

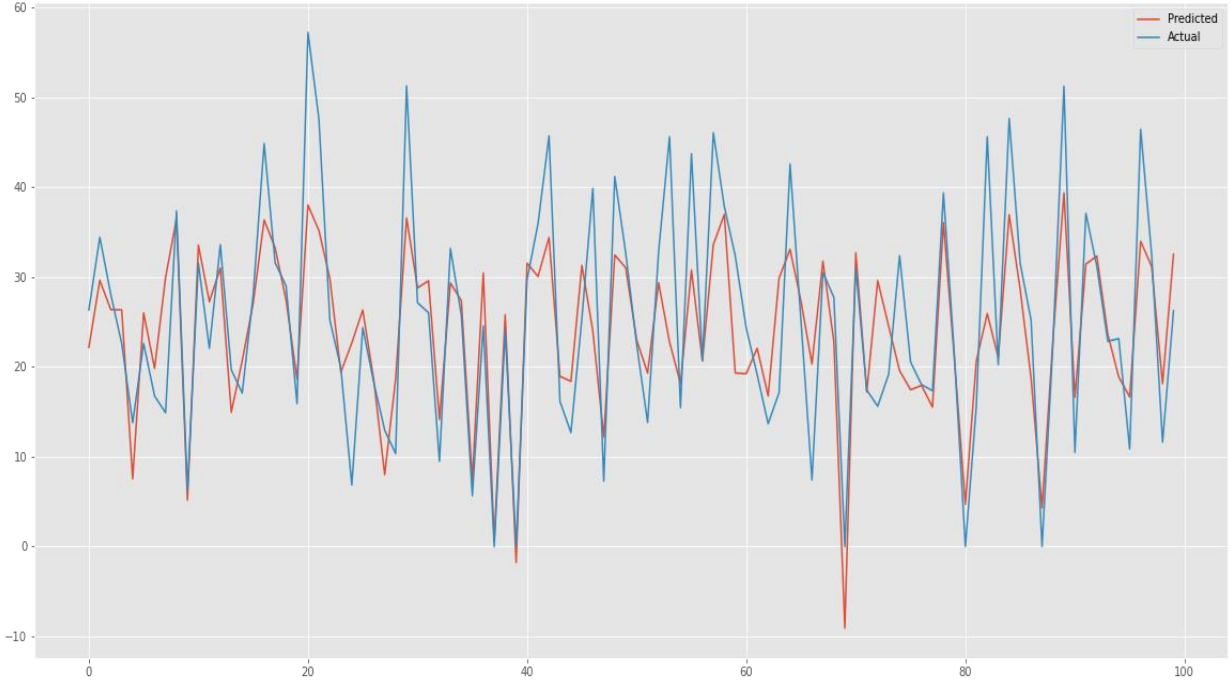
```
The best fit alpha value is found out to be : {'alpha': 0.001}
Using  {'alpha': 0.001}  the negative mean squared error is: - 53.68784100286467
Train Score :-53.36052478633948
Test score :-53.239858545492176
MSE : 53.239858545492176
RMSE : 7.296564845562066
r2_score : 0.6629672043035522
Adjusted R2 : 0.6603292864735155
```

After GridsearchCV

3. Ridge Regression

L2 Regularization

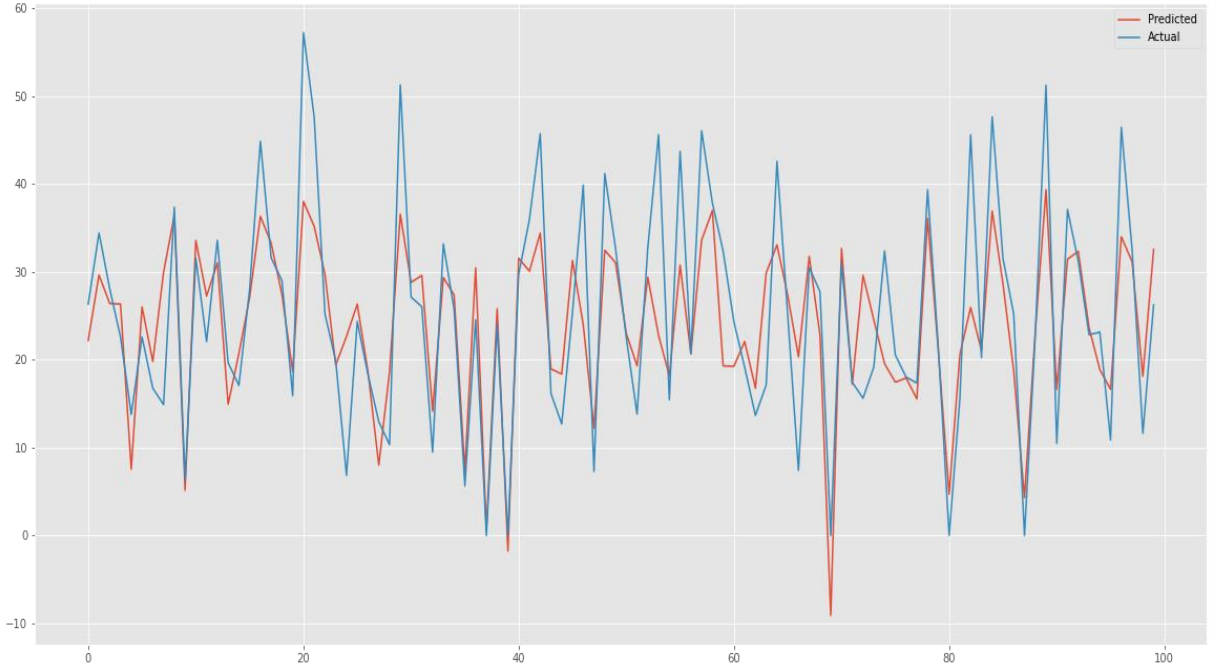
Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where the independent variables are highly correlated. It adds “squared magnitude” of coefficient as penalty term to the loss function during training.




```
MSE : 173945.47720075367
RMSE : 417.0677129684743
R2 : 0.5829743152290949
Adjusted R2 :  0.5797103020425823
```

4. Elastic-Net Regression

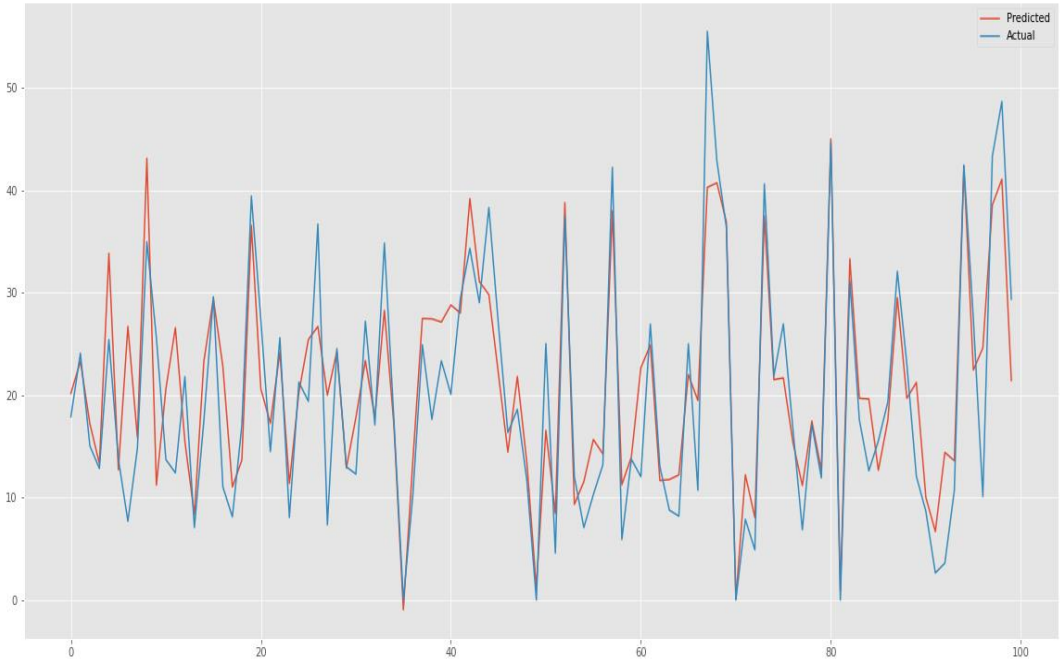
The elastic net is a regularized regression method that linearly combines the L_1 and L_2 penalties of the lasso and ridge methods.




```
MSE : 173880.81653682108
RMSE : 416.9901875785821
R2 : 0.5831293359751826
Adjusted R2 :  0.5798665361186348
```

5. Polynomial Regression

Polynomial Regression is a form of Linear regression known as a special case of Multiple linear regression which estimates the relationship as an nth degree polynomial. Polynomial Regression is sensitive to outliers so the presence of one or two outliers can also badly affect the performance.




```
MSE : 114889.36574372787
RMSE : 338.95333859357083
R2 : 0.7245584237474759
Adjusted R2 :  0.7224025734729396
```

6. Gradient Boosting Regressor

Gradient Boosting Machine (GBM) builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function by building an additive mode, using multiple decision trees of fixed size as weak learners or weak predictive models.

- Gradient Boosting algorithm is used to generate an ensemble model by combining the weak learners or weak predictive models.
- Used to train models for both regression and classification problem.

Boosting fit a sequence of weak learners − models that are only slightly better than random guessing, such as small decision trees − to weighted versions of the data. More weight is given to examples that were misclassified by earlier rounds.

```
Train Score :0.8878359518046239
Test score :0.8733636335154036
MSE : 20.004291346256295
RMSE : 4.472615716362887
r2_score : 0.8733636335154036
Adjusted R2 : 0.8723724649011134
```

7. Random Forest Regressor

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

```
Train Score :0.9867518223003453
Test score :0.9046748579297142
MSE : 15.058169841198175
RMSE : 3.8804857738688048
r2_score : 0.9046748579297142
Adjusted R2 : 0.9039287587514477
```

After GridsearchCV

```
Train Score :0.8603376640570762
Test score :0.8393758458516731
MSE : 25.373219921150795
RMSE : 5.037183729143776
r2_score : 0.8393758458516731
Adjusted R2 : 0.8381186586414883
```

Conclusion

- Bikes were found to be rented more on working days (i.e. non-holidays) and mostly during rush hours of 8:00 am (in morning) till 9:00 pm (in night).
- The demand of rented bikes was more during day time than nights. Further, high surge was observed in bike rental demand in the morning at 8:00 a.m. and 6:00 pm in the evening. This indicates that mostly people prefer to transit to their work place using bikes at 8:00 am (morning) and return back at 6:00 pm (evening).
- When the rainfall was less, people have booked more bikes except some few cases.
- The Temperature, Hour & Humidity are the most important features that positively drive the total rented bikes count.

Conclusion

- After performing the various linear regression models, it was found that ensemble models provided far more better results than linear regression model and its subsidiaries.
- Random forest regressor proved to be the best model with good train and test scores of 0.98 and 0.96 respectively, and model performance metrics including MSE : 14.85, RMSE: 3.85, and higher similar r2 and adjusted r2 scores of 0.90.
- In conclusion, the demand prediction for the given Seoul bike sharing dataset can be safely done using Random forest regressor.

THANK YOU