Capstone Project- 2

Seoul Bike Sharing Demand Prediction

Team Members
Zunaid
ArunTeja Lonka
Upasana Kumari
Sukesh Shetty

Introduction & Business Understanding of Seoul bike services

Problem statement

Data Description and Pre-processing

Exploratory Data Analysis

Analysing numerical variables and Feature Processing

Model development using different algorithms

Model Performance

Challenges Faced

Results & Conclusions

Introduction to Seoul Public Bike

Seoul Public Bike

• Seoul Public Bikes are designed to be used by all including women, the elderly and the infirm. Made of light-weight and durable materials, the bicycles prioritize driving stability and user convenience.

Business Understanding

- Bike rentals have become a popular service in recants years and is seems people are using it more often. With relatively cheaper rates and ease of pick up and drop at own convenience is what making this business thrive.
- Mostly used by peoples having no personal vehicles and also to avoid congested public transport that's why they prefer rental bikes.
- Therefore the business to strive and profit more , it has to be always ready and supply number of bikes at different locations, to fulfill the demand.

Rental Stations

- Rental stations are installed by popular pedestrian areas, including subway entrances/exits, bus stops, residential complexes, public offices, schools, and banks.
- Rental stations are unmanned stands for the rental and return of bikes.
- Rental stations are installed in highly accessible areas near popular destinations.
- Users can rent and return bicycles at any rental station.

Problem statement ..

Problem Description

- 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.

- The main goal is to create a prediction model that can be used to anticipate the number of bike rentals each hour based on weather conditions. As a result, it would be easier to anticipate fast and accurately.

Data Description

- **The dataset contains weather information (Temperature, Humidity, Wind speed, Visibility, Dew point, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.**
- Here **'Rented Bike count** is our **dependent variable** and rest of all other variable are independent variable.

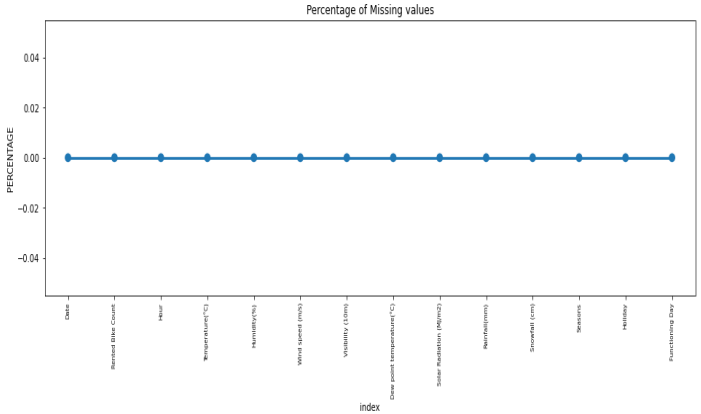
Attribute Information:

- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of the day
- Temperature-Temperature in Celsius
- Humidity - %
- Wind-speed - m/s
- Visibility - 10m
- Dew point temperature - Celsius
- Solar radiation - MJ/m2
- Rainfall - mm Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day – No Func(Non Functional Hours), Fun(Functional hours)

This Dataset contains 8760 rows and 14 columns.

Data Pre-processing

Insight from our Dataset after Data pre-processing



- There are no missing values present in our dataset.
- There are no duplicate values present in our dataset.
- There are no null values present in our dataset.
- And finally we have rented bike count which we have to predict for new observations.
- The dataset shows hourly rental data for one year(1 December 2017 to 30 November 2018) i.e. 365 days, so we consider as one single year.
- We convert the date column into 3 different column i.e. 'year', 'month', 'day'.
- We also change name of some columns for our convenience.

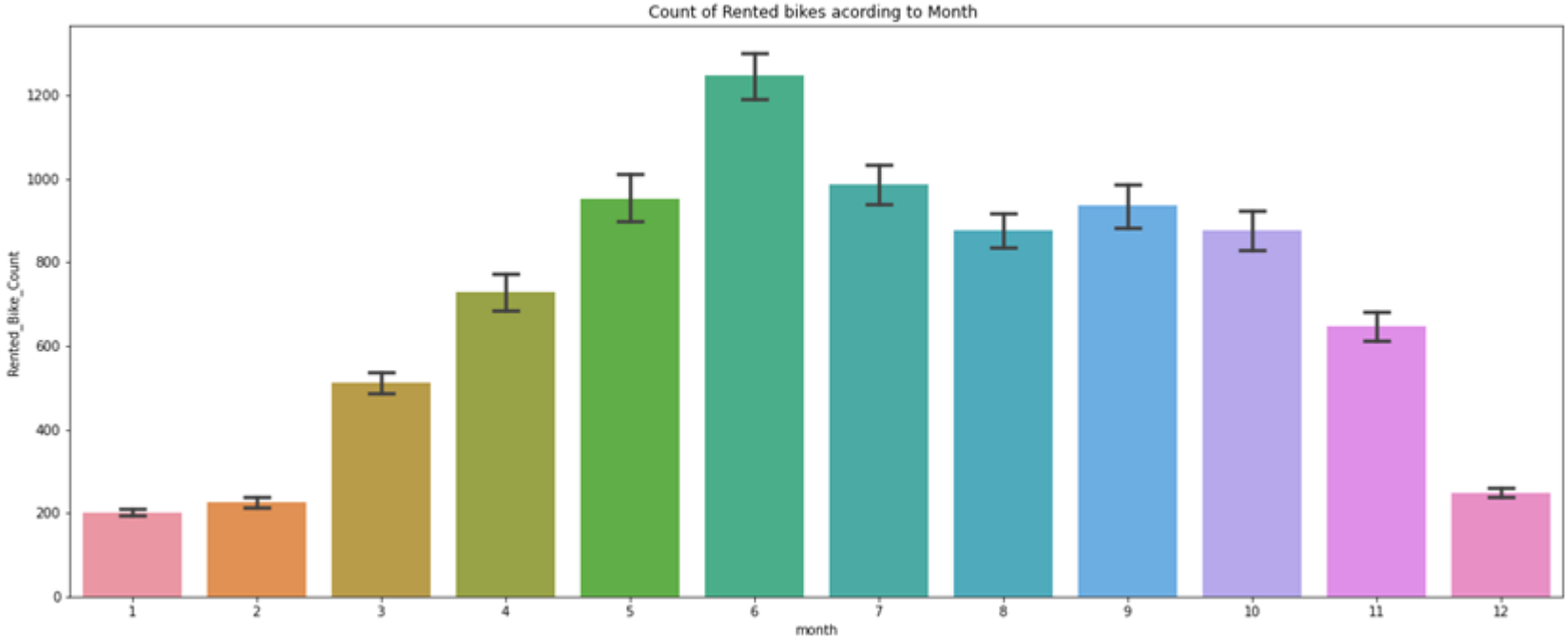
Exploratory Data Analysis

EDA is a process of visualizing and analyzing the data to extract the useful insight from it. In other words, EDA is the process of summarizing the data in order to gain the better understanding of the dataset.

We divided our EDA into several sub parts.

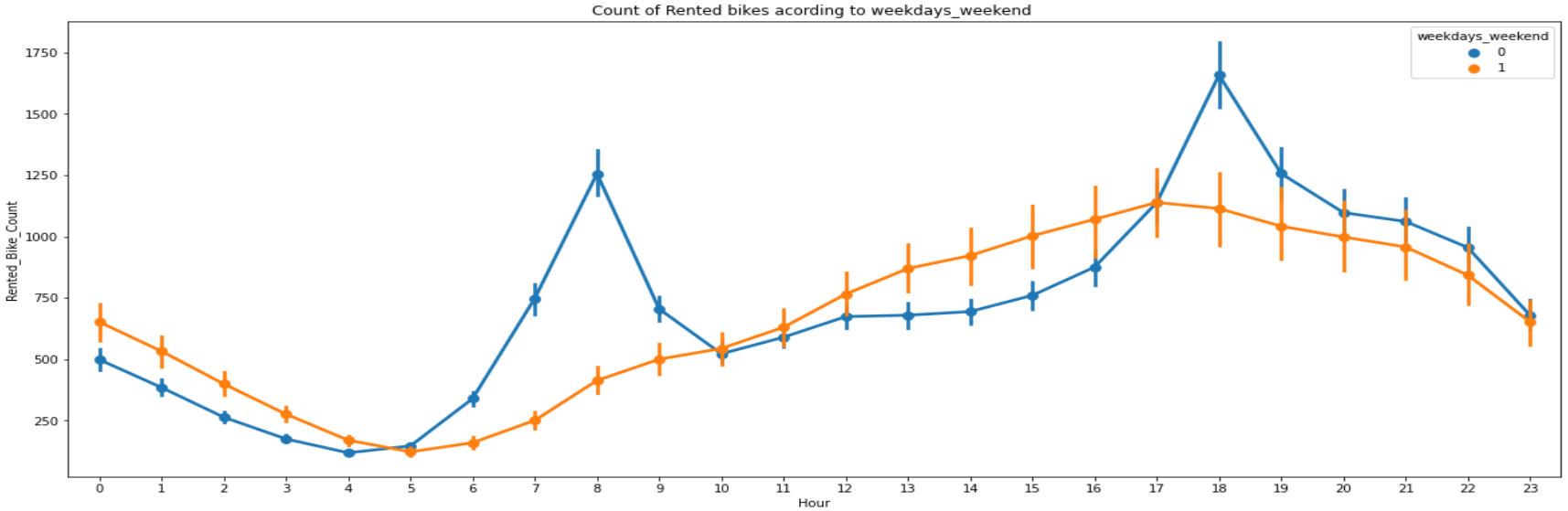
- Analysis of Categorical variable
- Analysis of Numerical Variable
- Distribution of Dependent variables

Analysis of Categorical variable (month)



From the above bar plot we can clearly say that from the month 5 to 10 the demand of the rented bike is high as compare to other months, these months are comes under the summer season.

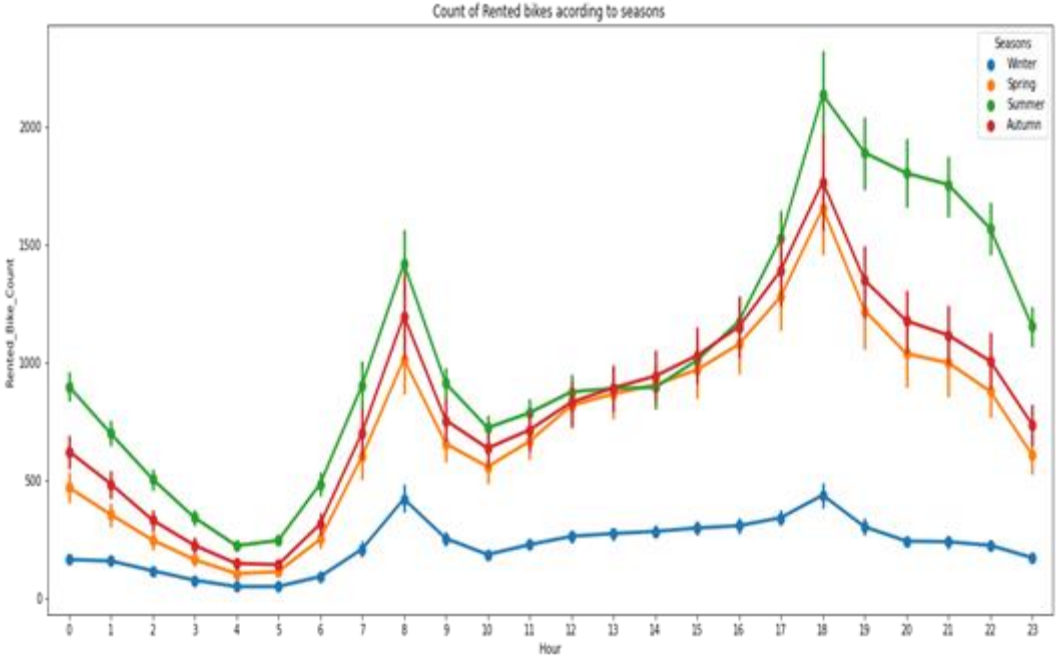
Analysis of Categorical variable (hour and weekdays)

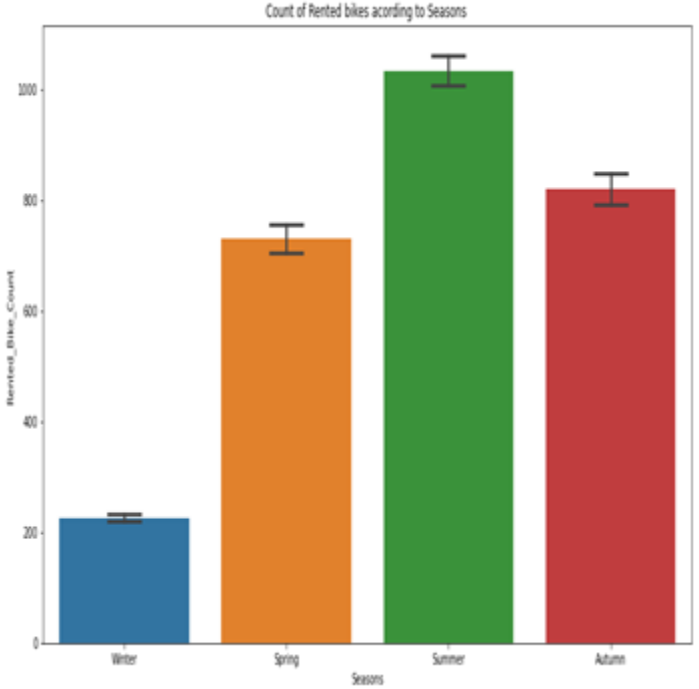


In the above point plot the blue line represent the weak days and orange line represent the weak ends.
Peak hour in weak days are from 7am to 9am and 5pm to 7pm.

On weak ends the demand of rented bikes are very low specially in the morning hours but when the evening start from 2pm to 8pm the demand of bikes increases slightly.

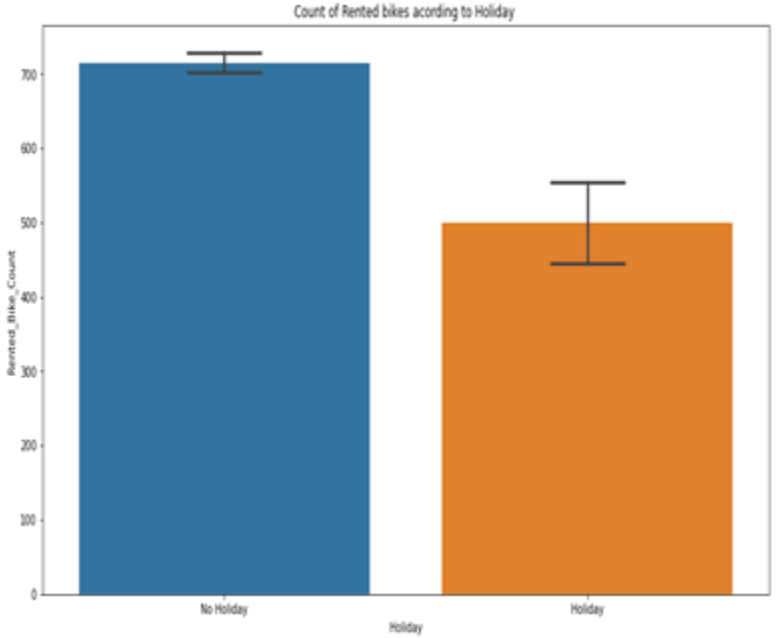
Analysis of Categorical variable (seasons)

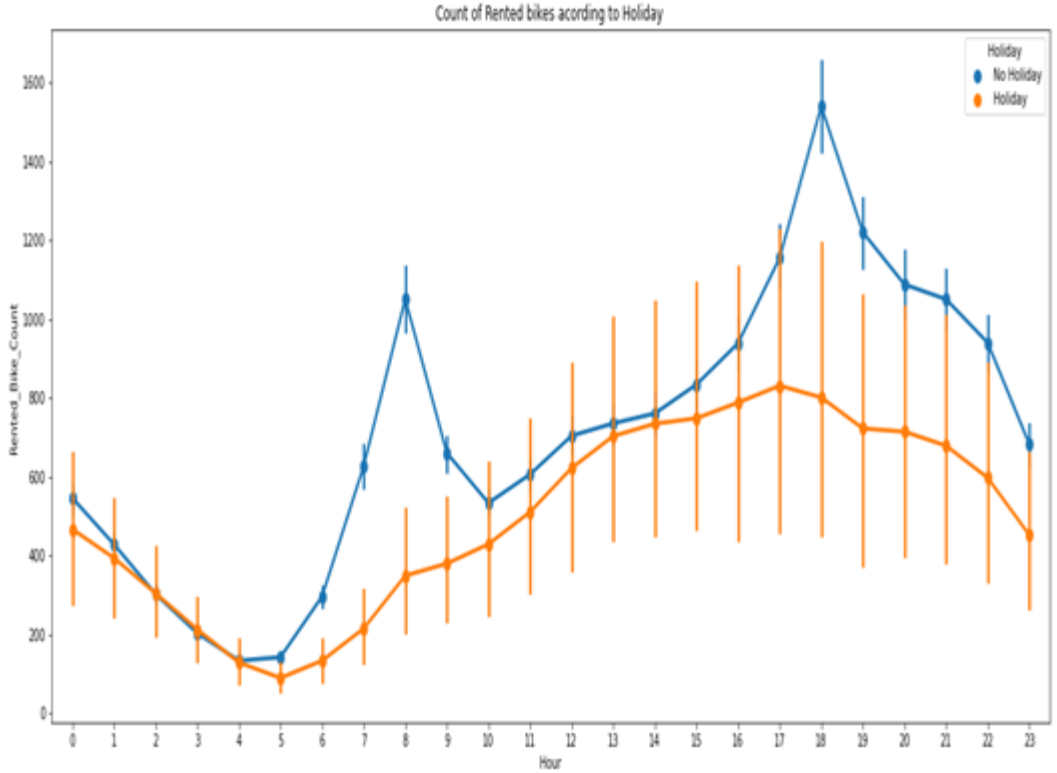




- *In the above bar plot and point plot which shows the use of rented bike in in four different seasons, and it clearly shows that,*
- *In summer season the use of rented bike is high and peak time is 7am-9am and 5pm-7pm.*
- *In winter season the use of rented bike is very low because of snowfall.*

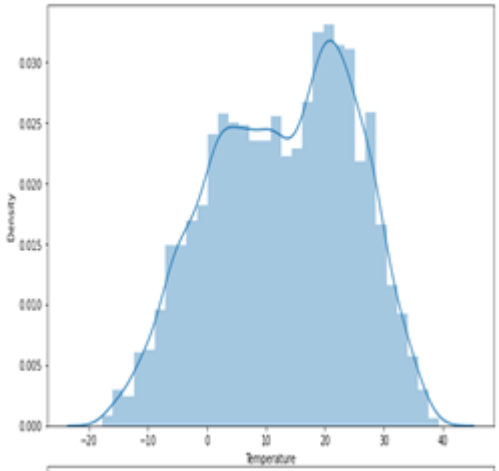
Analysis of Categorical variable (holiday)

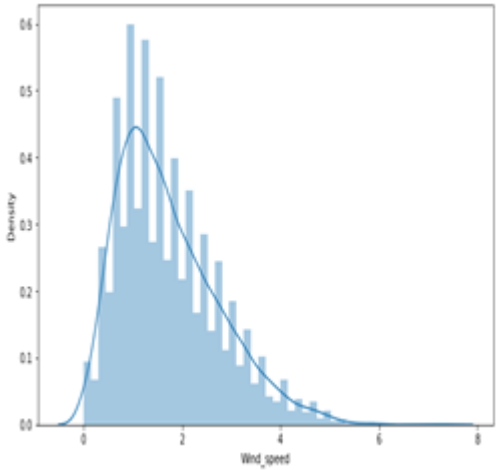


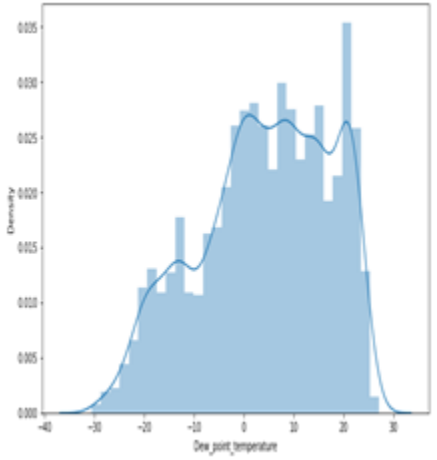


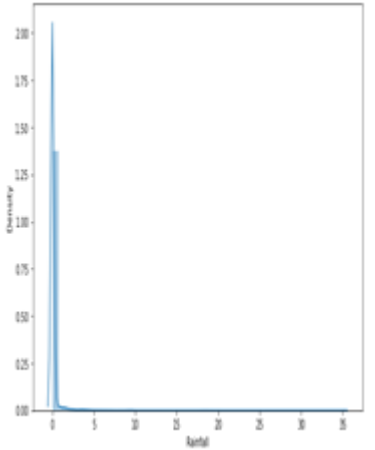
- *In the above bar plot and point plot shows that the rented bike count on Holidays and No Holidays,*
- *It is clearly shows that there is less rides occur on an Holiday.*
- *Peak hour on Holidays are from 2pm to 8pm.*

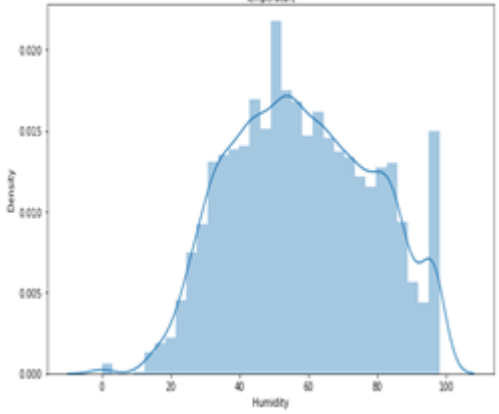
Distribution of our Independent Variable

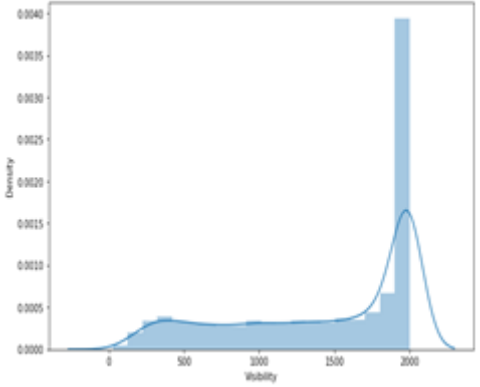


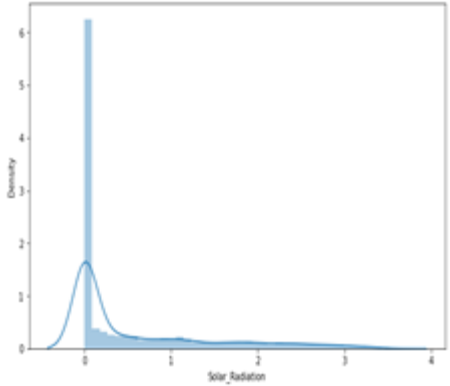


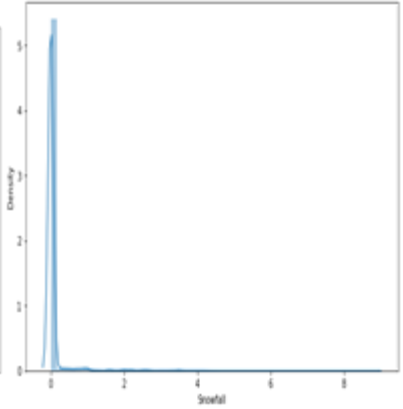




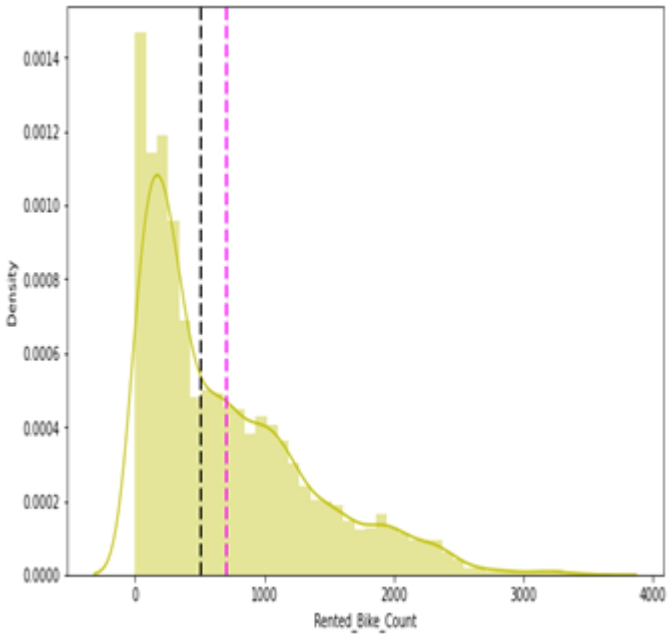




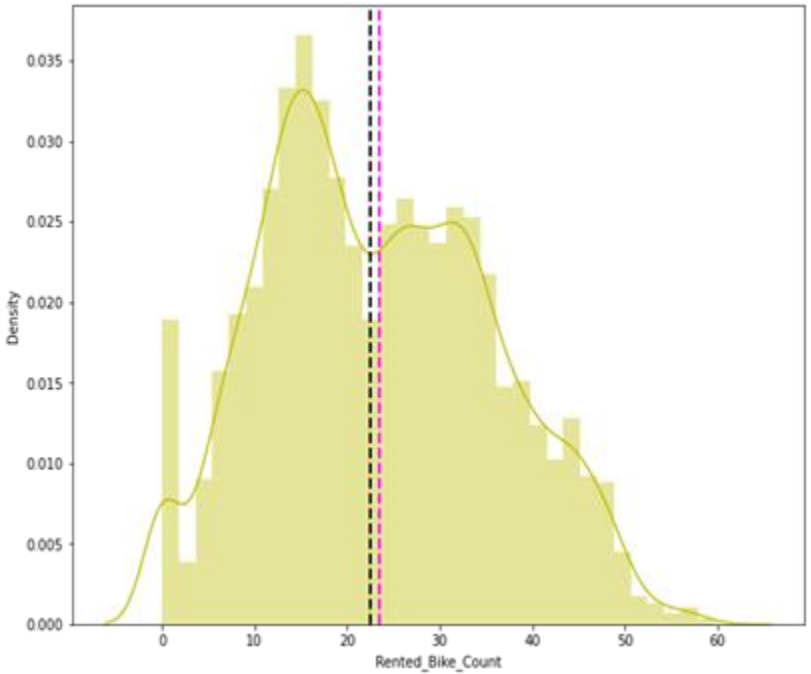




Analysing numerical variables and Feature Processing



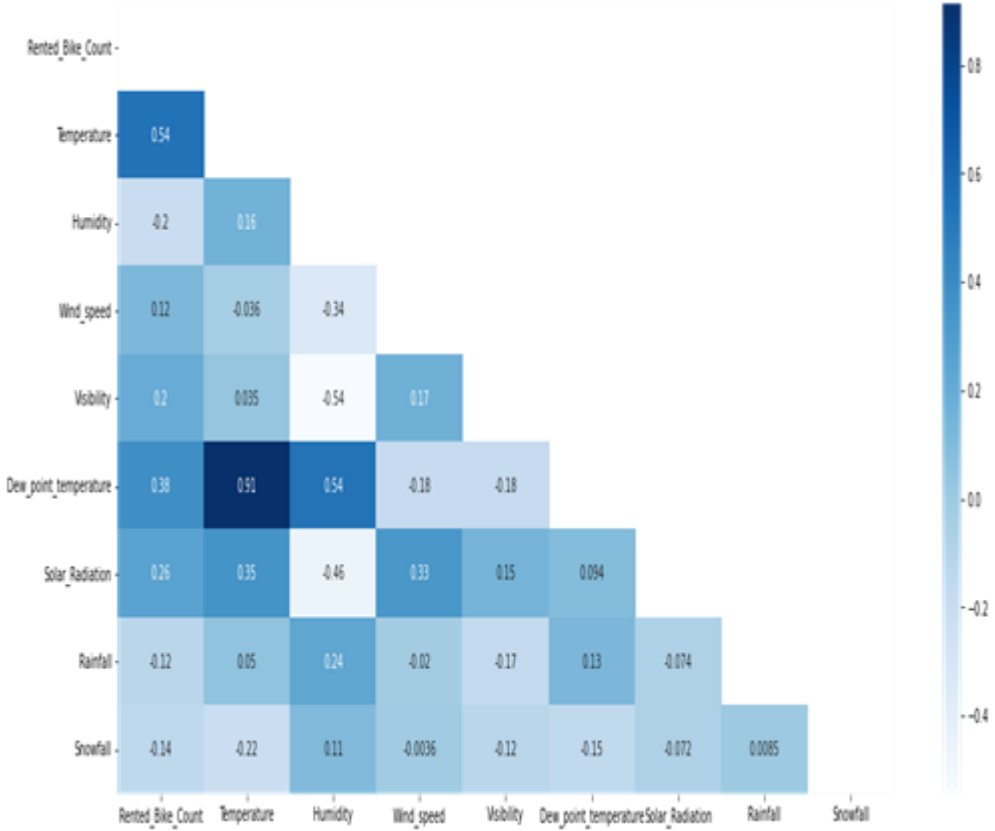
distribution of dependent variable



• *The above graph shows that Rented Bike Count has moderate right skewness. Since the assumption of linear regression is that 'the distribution of dependent variable has to be normal', so we should perform some operation to make it normal.*

• *Since we have generic rule of applying Square root for the skewed variable in order to make it normal .After applying Square root to the skewed Rented Bike Count, here we get almost normal distribution.*

correletion betweeen variables using Correlation heatmap



We can observe on the heatmap that on the target variable line the most positively correlated variables to the rent are :

- the temperature
- the dew point temperature
- the solar radiation

And most negatively correlated variables are:

- Humidity
- Rainfall

From the above correlation heatmap, We see that there is a positive correlation between columns 'Temperature' and 'Dew point temperature' i.e. 0.91 so even if we drop this column then it don't affects the outcome of our analysis. And they have the same variations.. so we can drop the column 'Dew point temperature(°C)'.

Label Encoding:

A dataset may contain various type of values, sometimes it consists of categorical values. So, in-order to use those categorical value for programming efficiently we create dummy variables.

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.

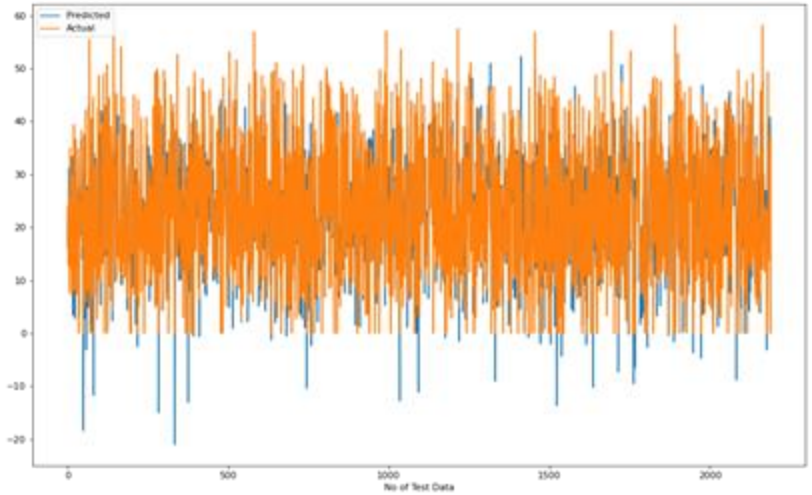
```
##Example function
def one_hot_encoding(data, column):
    data = pd.concat([data, pd.get_dum
mies(data[column], prefix=column, drop
_first=True)], axis=1)
    data = data.drop([column], axis=1)
    return data
```

Model development using different algorithms

We implemented 7 machine learning algorithms

- Linear regression
- Lasso regression
- Ridge regression
- Elastic net regression
- Decision tree regression
- Random forest regression
- Gradient boosting regression
- Gradient Boosting GridsearchCV

LINEAR REGRESSION



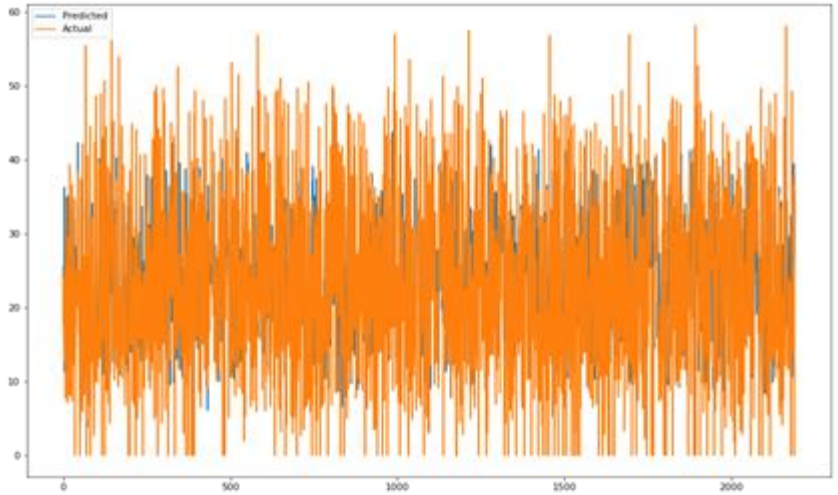
Train set results:

MSE : 35.07751288189293
RMSE : 5.9226271942350825
MAE : 4.474024092996787
R2 : 0.7722101548255267
Adjusted R2 :
0.7672119649454145

Test set results:

MSE : 33.27533089591926
RMSE : 5.76847734639907
MAE : 4.410178475318181
R2 : 0.7893518482962683
Adjusted R2 :
0.7847297833429184

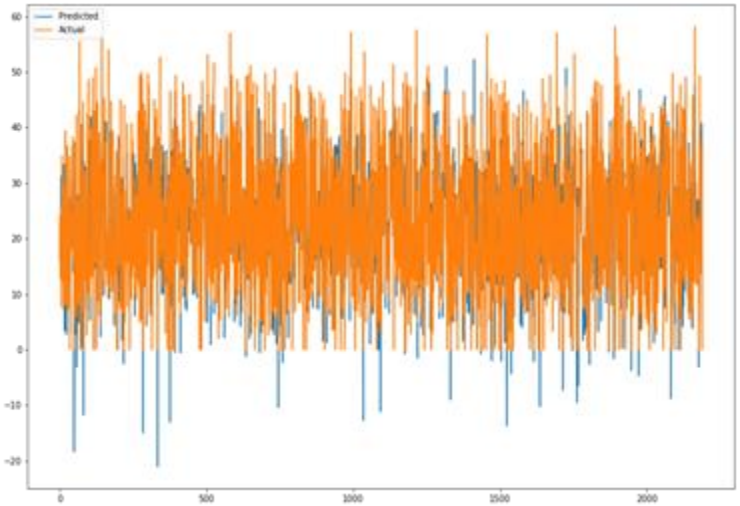
LASSO REGRESSION



MSE : 91.59423336097032
RMSE :9.570487623991283
MAE : 7.255041571454952
R2 :0.40519624904934015
Adjusted R2 :
0.3921449996120475

MSE : 96.7750714044618
RMSE 9.837432155011886
MAE :7.455895061963607
R2 :0.3873692800799008
Adjusted R2 :
0.37392686932535146

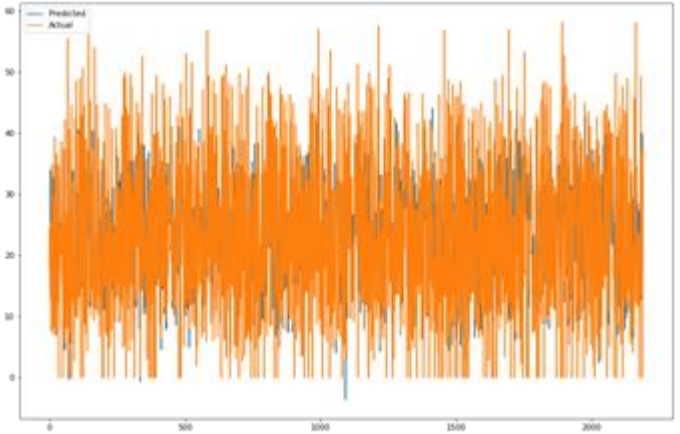
RIDGE REGRESSION



Train set results:
MSE : 35.07752456136463
RMSE : 5.922628180239296
MAE : 4.474125776125378
R2 : 0.7722100789802107
Adjusted R2 : 0.7672118879

Test set results:
MSE : 33.27678426818438
RMSE : 5.768603320404722
MAE : 4.410414932539515
R2 : 0.7893426477812578
Adjusted R2 : 0.784720389

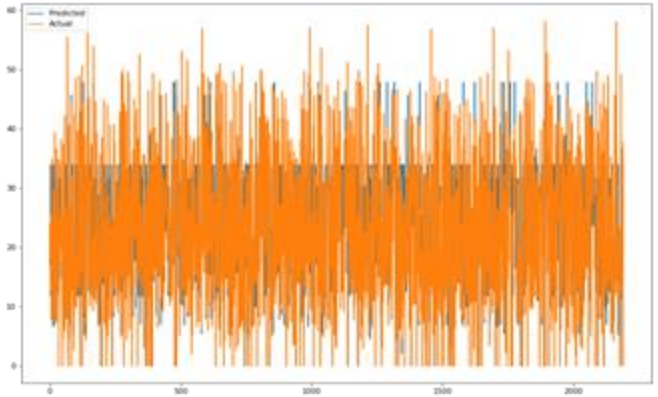
ELASTIC NET REGRESSION



Train set results:
MSE : 57.5742035398887
RMSE : 7.587766703048315
MAE : 5.792276538970546
R2 : 0.6261189054494012
Adjusted R2 : 0.6179151652

Test set results:
MSE : 59.45120536350042
RMSE : 7.710460774538
MAE : 5.873612334800099
R2 : 0.62364652169
Adjusted R2 : 0.6153885321

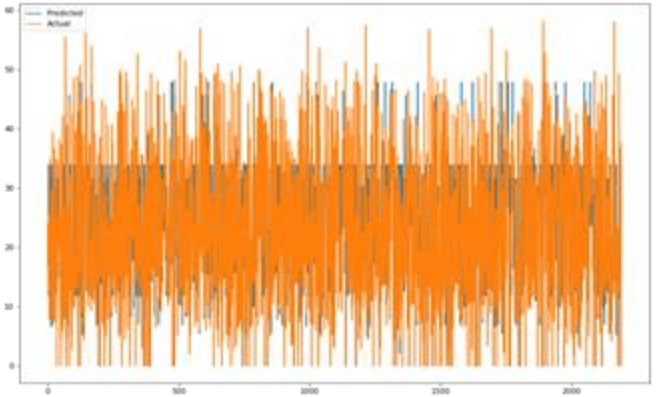
DECISION TREE



Train set results:
MSE : 43.47944139910931
RMSE : 6.593894251435135
MAE : 4.833796381744261
R2 : 0.7176488749951184
Adjusted R2 : 0.7114534955015472

Test set results:
MSE : 50.424498696751435
RMSE : 7.101020961576682
MAE : 5.196762423651914
R2 : 0.6807897272522531
Adjusted R2 : 0.6737855802778627

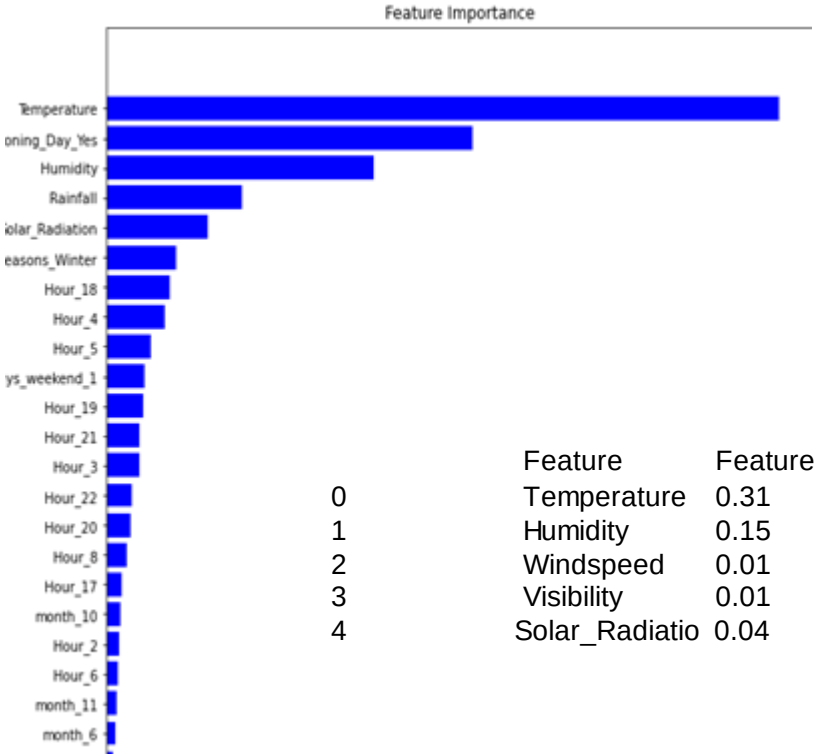
RANDOM FOREST



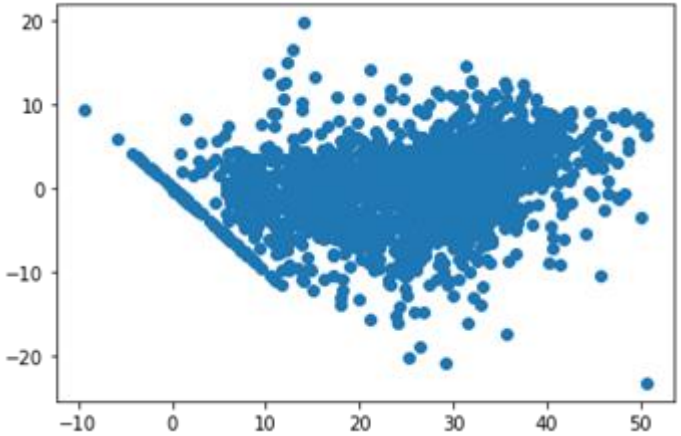
Train set results:
MSE : 1.5776988364569622
RMSE : 1.2560648217576043
MAE : 0.7983299095968202
R2 : 0.9897545822333945
Adjusted R2 : 0.9895297761479461

Test set results:
MSE : 12.685885000312213
RMSE : 3.561725003465626
MAE : 2.2077215853951526
R2 : 0.9196925122577947
Adjusted R2 : 0.9179303965136847

GRADIENT BOOSTING



	Feature	Feature Importance
0	Temperature	0.31
1	Humidity	0.15
2	Windspeed	0.01
3	Visibility	0.01
4	Solar_Radiatio	0.04



Train set result:

MSE : 18.64801713184794
RMSE : 4.3183349953249275
MAE : 3.269003569273124
R2 : 0.8789016499095264
Adjusted R2 : 0.8762444965695393

Test set results:

MSE : 21.28944184250869
RMSE : 4.6140483138463875
MAE : 3.492858786559991
R2 : 0.8652280396863458
Adjusted R2 : 0.8622708584843188

Conclusions:

- Hour of the day holds the most important feature.
- Bike rental count is mostly correlated with the time of the day as it is peak at 10am morning and 8pm at evening
- We observed that bike rental count is high during working days than non-working days.
- We see that people generally prefer to rent bike at moderate to high temperatures, and when little windy
- It is observed that highest number bike rentals counts in Autumn & Summer seasons and the lowest in winter season.
- We observed that the highest number of bike rentals on a clear day and the lowest on a snowy or rainy day.
- We observed that with increasing humidity, the number of bike rental counts decreases.

Results & Conclusions:

- ✔ During the time of our analysis, we initially did EDA on all the features of our dataset.
- ✔ We first analysed our dependent variable, 'Rented Bike Count' and also transformed it.
- ✔ Next we analysed categorical variable and dropped the variable who had majority of one class, we also analysed numerical variable, found out the correlation, distribution and their relationship with the dependent variable.
- ✔ We also removed some numerical features who had mostly 0 values and hot encoded the categorical variables.
    - Next we implemented 7 machine learning algorithms Linear Regression, lasso, ridge, elastic net, decision tree, Random Forest and XGBoost.
    - We did hyperparameter tuning to improve our model performance.
    - The results of our evaluation are:

Training set Results:

		Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Training set	**0**	Linear regression	4.474	35.078	5.923	0.772	0.77
	1	Lasso regression	7.255	91.594	9.570	0.405	0.39
	2	Ridge regression	4.474	35.078	5.923	0.772	0.77
	3	Elastic net regression	5.792	57.574	7.588	0.626	0.62
	4	Dicision tree regression	5.716	56.456	7.514	0.633	0.63
	5	Random forest regression	0.805	1.592	1.262	0.990	0.99
	6	Gradient boosting regression	3.269	18.648	4.318	0.879	0.88
	7	Gradient Boosting gridsearchcv	1.849	7.455	2.730	0.952	0.95

Test set Results

		Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Test set	**0**	Linear regression	4.410	33.275	5.768	0.789	0.78
	1	Lasso regression	7.456	96.775	9.837	0.387	0.37
	2	Ridge regression	4.410	33.277	5.769	0.789	0.78
	3	Elastic net regression Test	5.874	59.451	7.710	0.624	0.62
	4	Decision tree regression	5.872	60.654	7.788	0.616	0.61
	5	Random forest regression	2.211	12.761	3.572	0.919	0.92
	6	Gradient boosting regression	3.493	21.289	4.614	0.865	0.86
	7	Gradient Boosting gridsearchcv	2.401	12.393	3.520	0.922	0.92

Results & Conclusions:

- ✔ No overfitting is seen.
- ✔ Random forest Regressor and Gradient Boosting gridsearchcv gives the highest R2 score of 99% and 95% respectively for Train Set and 92% for Test set.
- ✔ Feature Importance value for Random Forest and Gradient Boost are different.
- ✔ We can deploy this model.
- ✔ However, this is not the ultimate end. As this data is time dependent, the values for variables like temperature, windspeed, solar radiation etc., will not always be consistent. Therefore, there will be scenarios where the model might not perform well.
- ✔ As Machine learning is an exponentially evolving field, we will have to be prepared for all contingencies and also keep checking our model from time to time.
- ✔ Therefore, having a quality knowledge and keeping pace with the ever evolving ML field would surely help one to stay a step ahead in future.

Challenges Faced

- Understanding Problem Statement
- Reading the dataset and Understanding the meaning of some columns
- Designing multiple visualisations to summarize the information in the dataset and successfully communicate the results and trends to clints/readers
- Machine learning models usually required high computation power

Refences and links

- ✔ Alma better
- ✔ analytics Vidhya
- ✔ Kaggle
- ✔ Quora
- ✔ Stack over flow