Capstone Project

Seoul Bike Sharing Demand Prediction

Sanjay Jaiswal

Content:

- Business Understanding
- Data Summary
- Feature Analysis
- Exploratory Data Analysis
- Implementing Algorithms
- Challenges
- Conclusions

Business Understanding:

- Bike rentals have became a popular service in recent years and it 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 people having no personal vehicles and also to avoid congested public transport which that's why they prefer rental bikes.

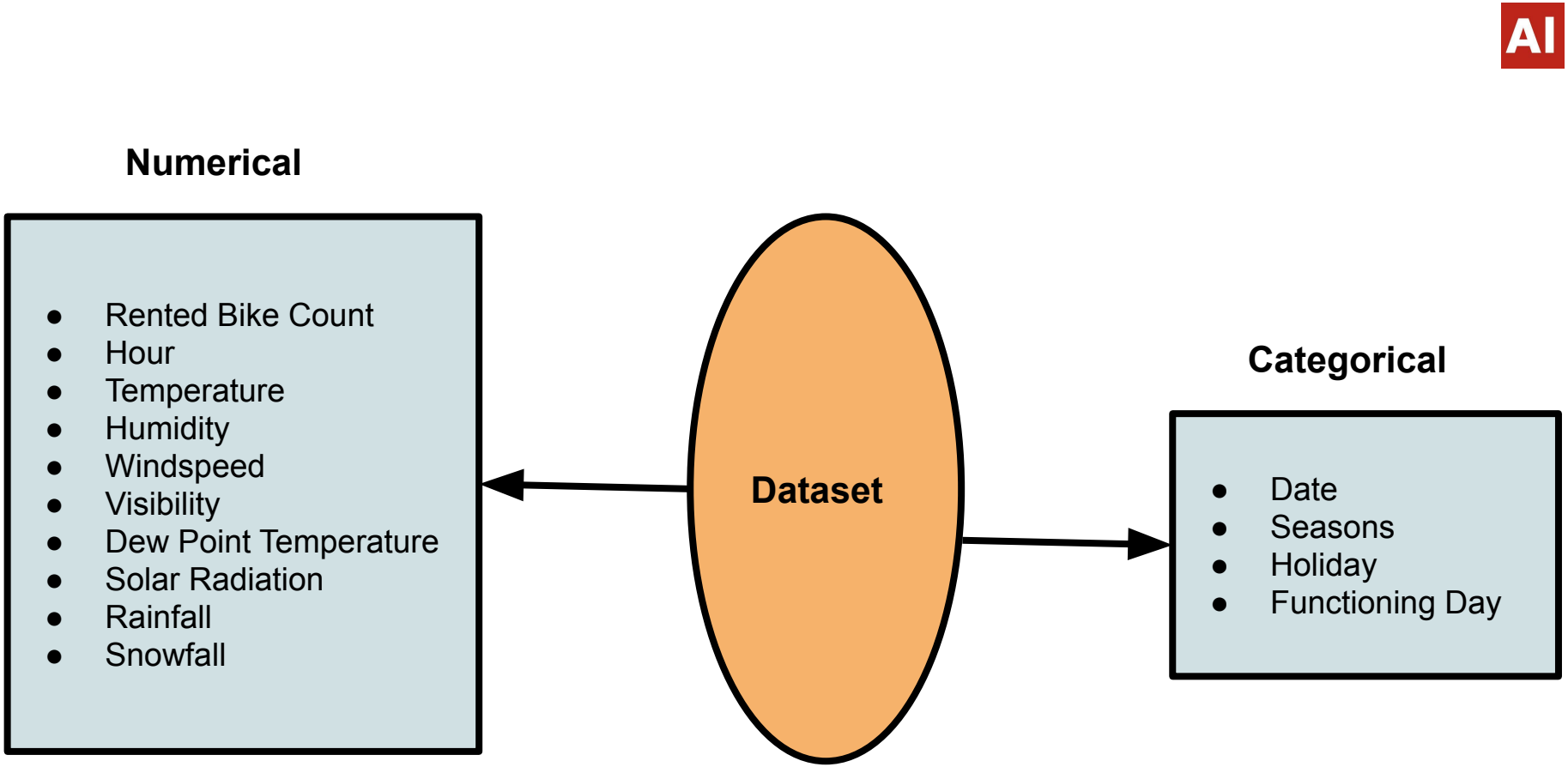
- Therefore, the business to strive and profit more, it has to be always ready and supply no. of bikes at different locations, to fulfil the demand.

- Our project goal is a preplanned set of bike count values that can be a handy solution to meet all demands.

Data Summary:

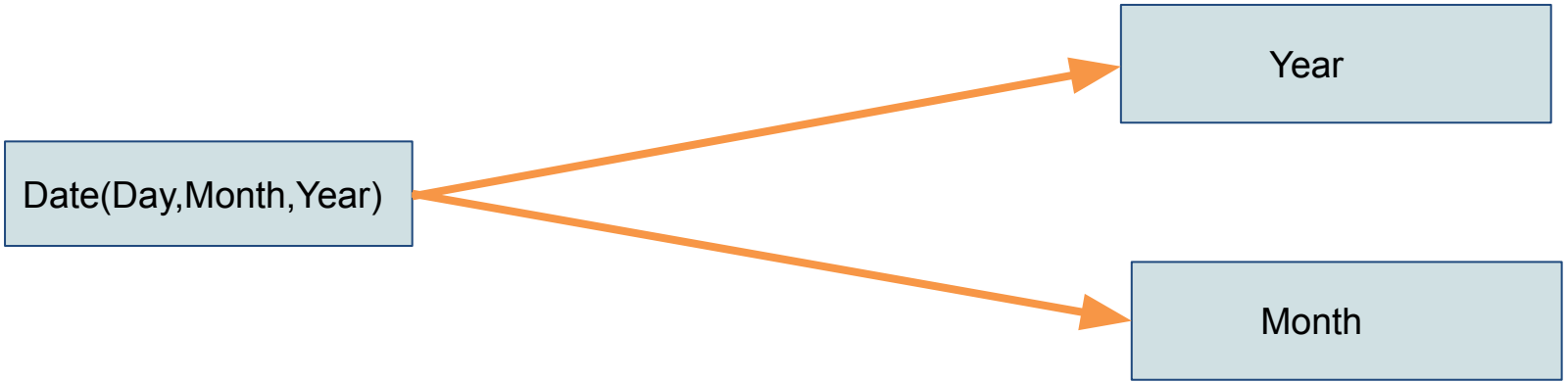
- This Dataset contains 8760 rows and 14 columns.
- Four categorical features Date, Seasons, Holiday, & Functioning Day.
- We have some numerical type variables such as temperature, humidity, wind, visibility, dew point temp, solar radiation, rainfall, snowfall which tells the environment conditions at that particular hour of the day.
- There is no missing values and duplicate value in the given dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8760 entries, 0 to 8759
Data columns (total 14 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Date                   8760 non-null   object
 1   Rented_Bike_Count      8760 non-null   int64
 2   Hour                   8760 non-null   int64
 3   Temperature            8760 non-null   float64
 4   Humidity               8760 non-null   int64
 5   Wind_speed             8760 non-null   float64
 6   Visibility             8760 non-null   int64
 7   Dew_point_temperature  8760 non-null   float64
 8   Solar_Radiation        8760 non-null   float64
 9   Rainfall               8760 non-null   float64
 10  Snowfall               8760 non-null   float64
 11  Seasons                8760 non-null   object
 12  Holiday                8760 non-null   object
 13  Functioning_Day        8760 non-null   object
dtypes: float64(6), int64(4), object(4)
memory usage: 958.2+ KB
```


Numerical
Rented Bike Count
Hour
Temperature
Humidity
Windspeed
Visibility
Dew Point Temperature
Solar Radiation
Rainfall
Snowfall
Dataset
Categorical
Date
Seasons
Holiday
Functioning Day

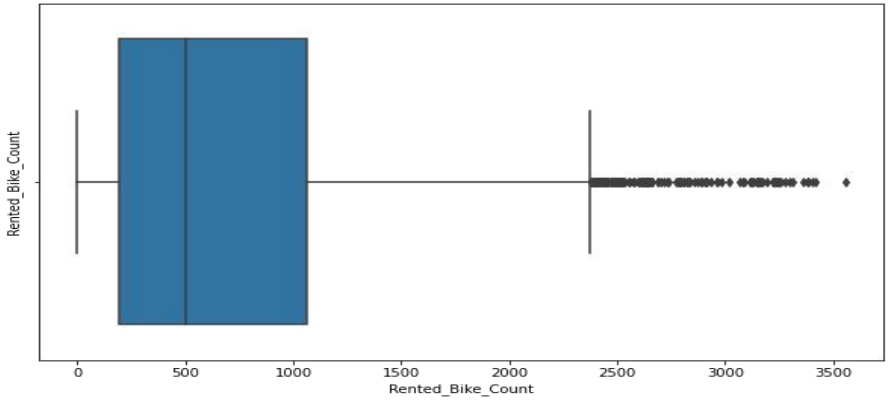
Feature Analysis:

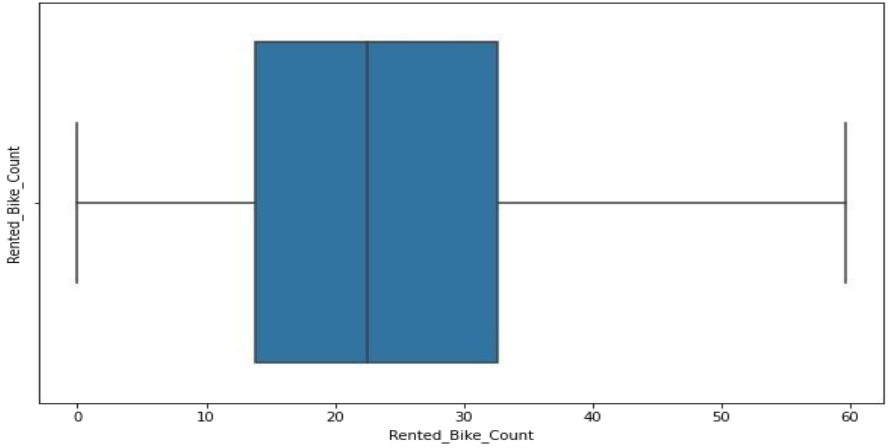
Feature engineering is the process of selecting, manipulating, and transforming initial variables into features that can be used in model training.



Analysis Of Rented Bike Count(Dependent Variable):

- The boxplot shows that Rented Bike Count has moderate right skewed.
- It also shows that we have detect outliers in Rented Bike Count column.
- Since the assumption of linear regression is that 'the distribution of dependent variable should be normal. so, we should perform Square root operation to make it normal.
- After applying Square root to the skewed Rented Bike Count, here we get almost normal distribution.
- After applying Square root to the Rented Bike Count column, we find that there is no outliers present

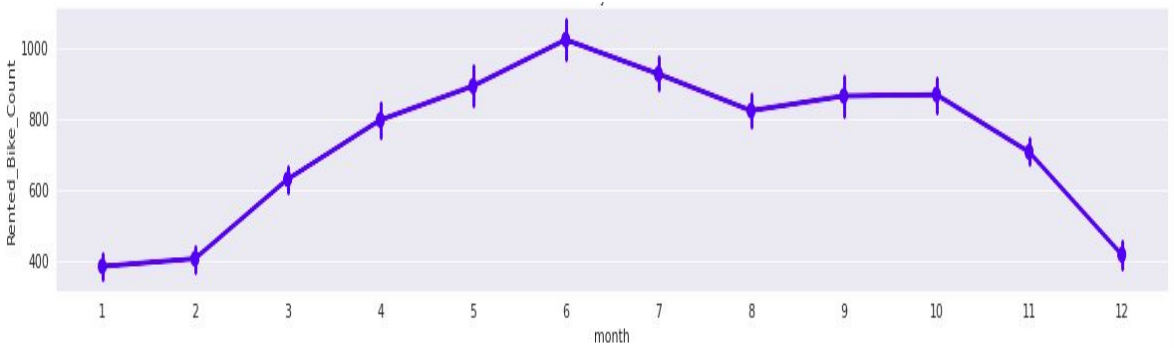


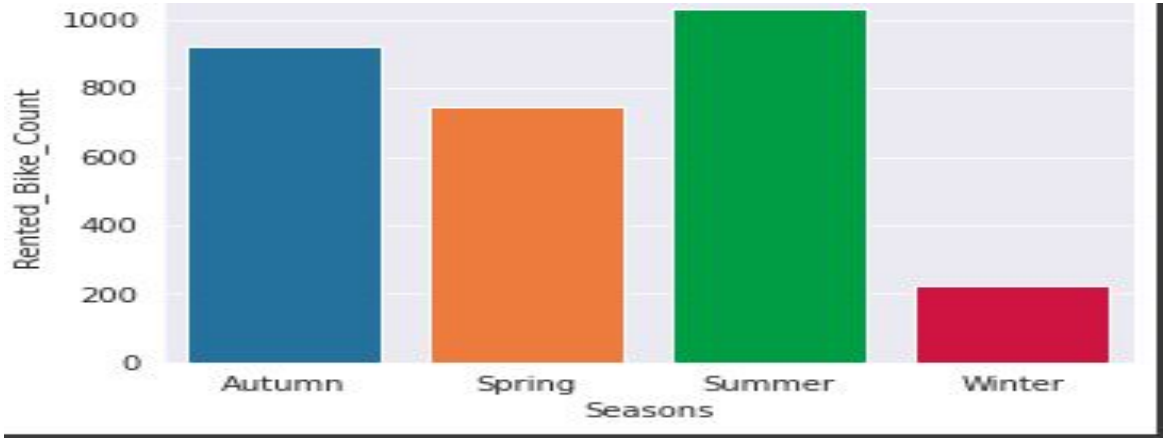


Analysis of Season Variable:

- From the point 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 inside the summer season.

- And from the barchart we can say that average rented bike counts are more in summer season as compared to other season.

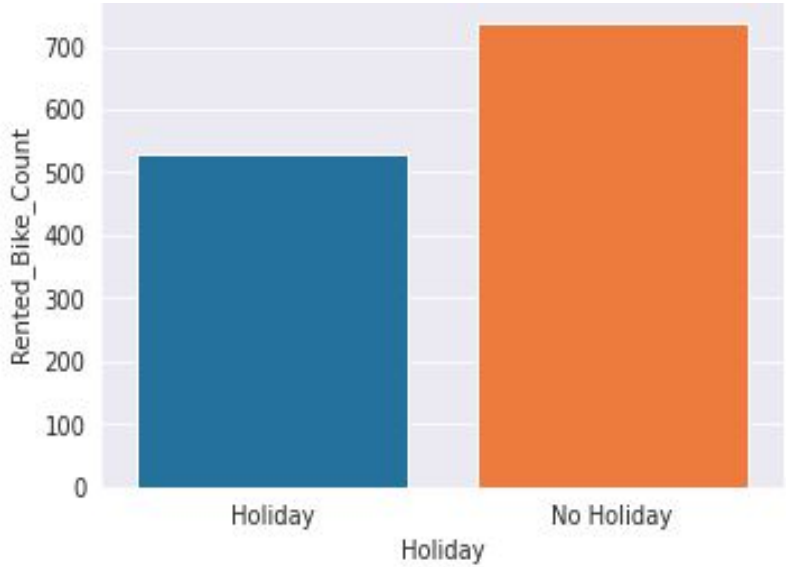




Analysis of Holiday Variable:

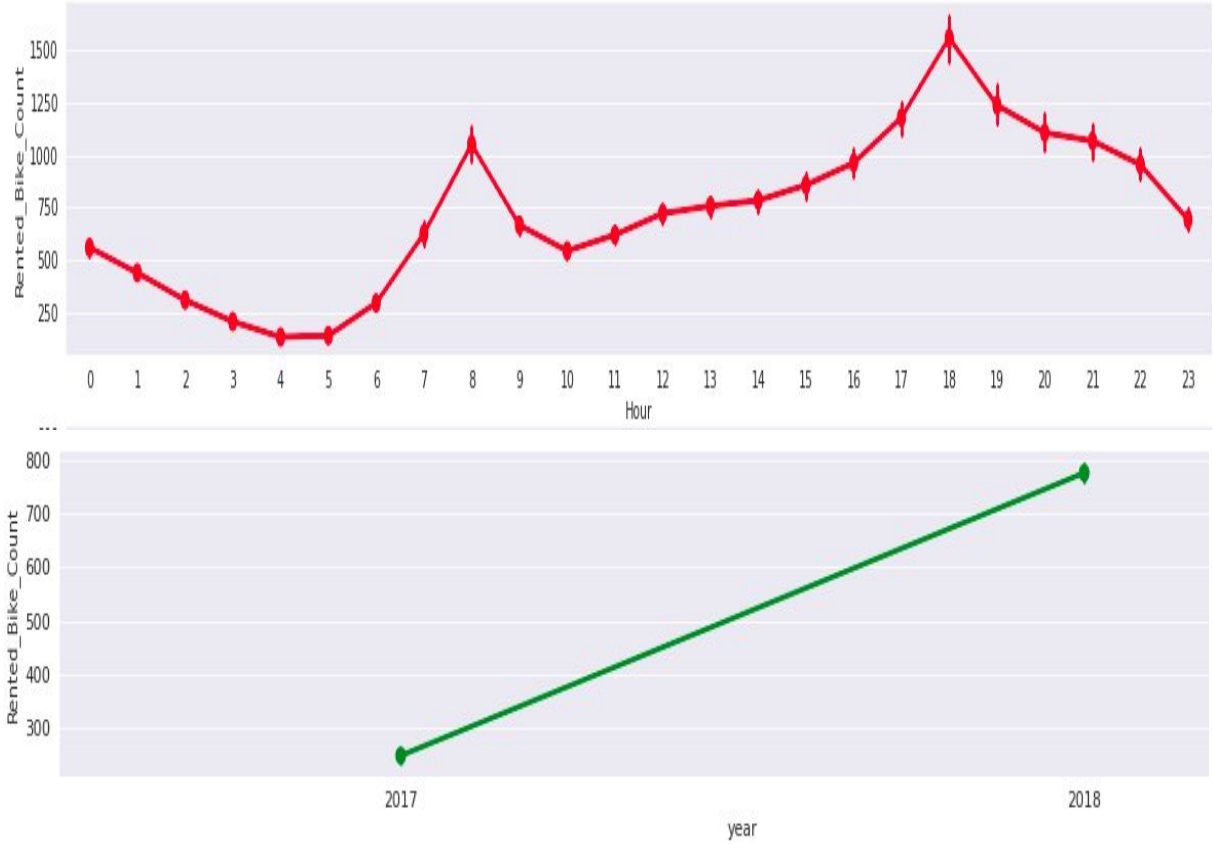
- From the barchart, we observe that large number of bikes are being rented when there is a working day/No Holiday and more often in summer season.

- Even in general also, bikes are being rented more in the working day itself regardless of the seasons.



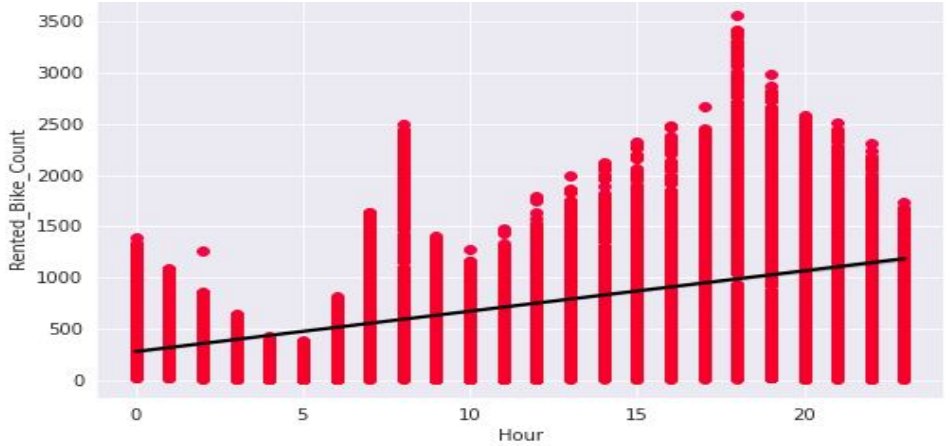
Analysis of Hour and Year Variable:

- In the given plot which shows the no. of rented bike varies according to the hours.
- There must be high demand during the office timings around 8 A.M. and 6 P.M., also for early morning and late evening we are having a relatively different trends
- Compare to 2017 there was a increase in count of bike rent in 2018.

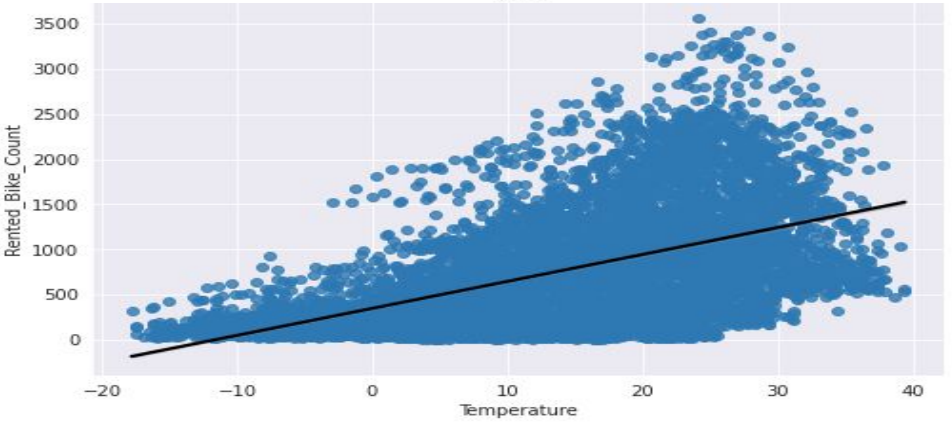


Analysis of Dependent Variable Vs Hour & Temp:

- As we can see the regplot we can say that there must be high demand during the office timings around 8 A.M. and 6 P.M. also for early morning and late evening we are having a relatively different trends.



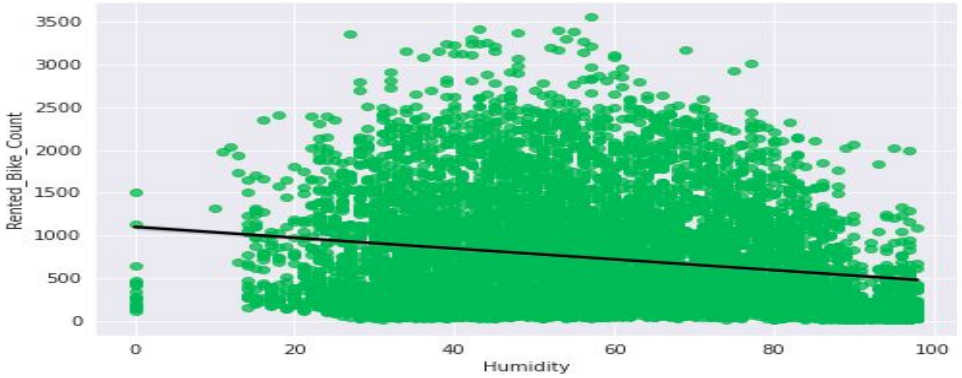
- From the given regplot we can see that people like to ride bikes when it is pretty hot temp around 25°C.

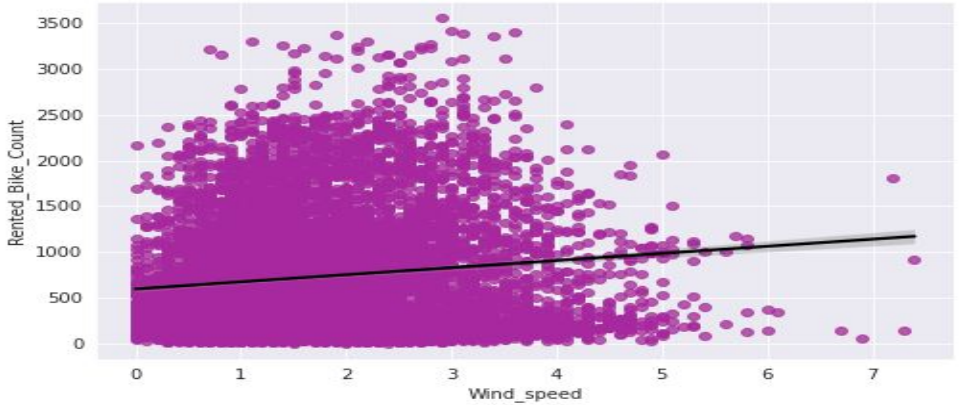


Analysis of Dependent Variable Vs Humidity & Wind Speed Temp:

- As we can see the regplot we can say that Humidity acts as a deterrent(a thing that discourages or is intended to discourage someone from doing something) to a bike ride. The bike count decreases when the humidity increases

- In wind speed plot that the demand of rented bike is uniformly distribute despite of wind speed but when the speed of wind was 7 m/s then the demand of bike also increase that clearly means peoples love to ride bikes when its little windy

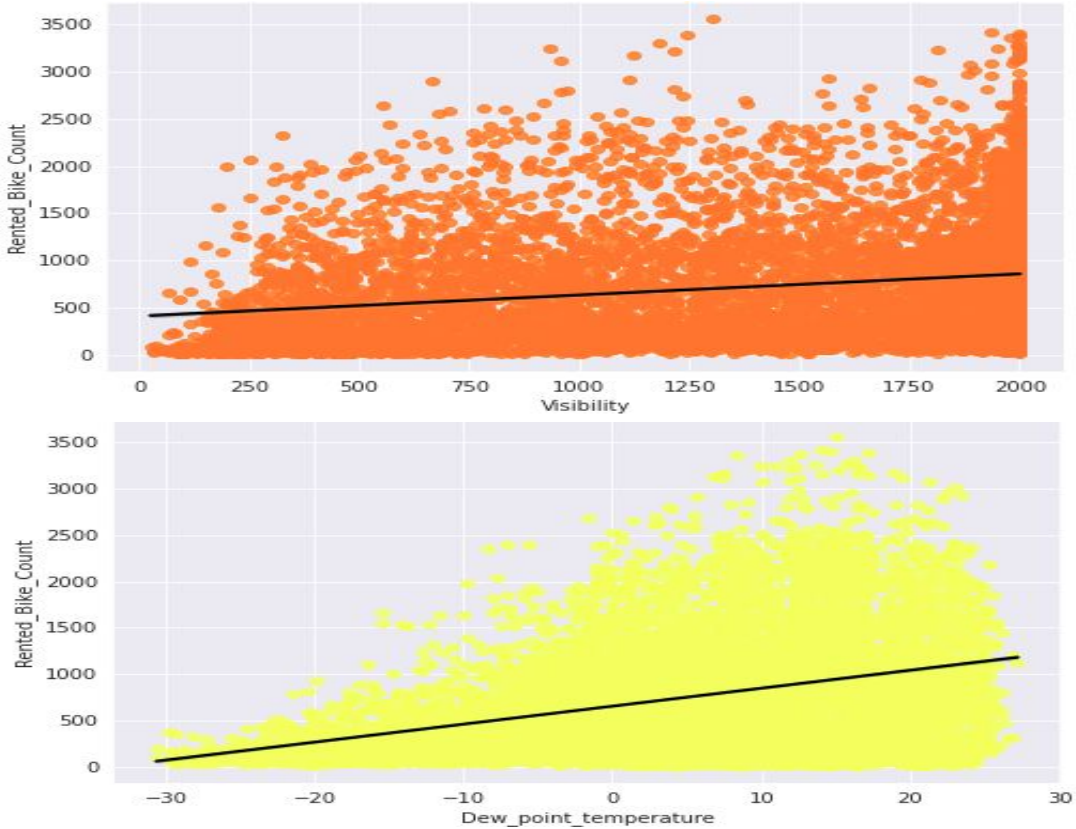




Analysis of Dependent Variable Vs Visibility & Dew Point Temp:

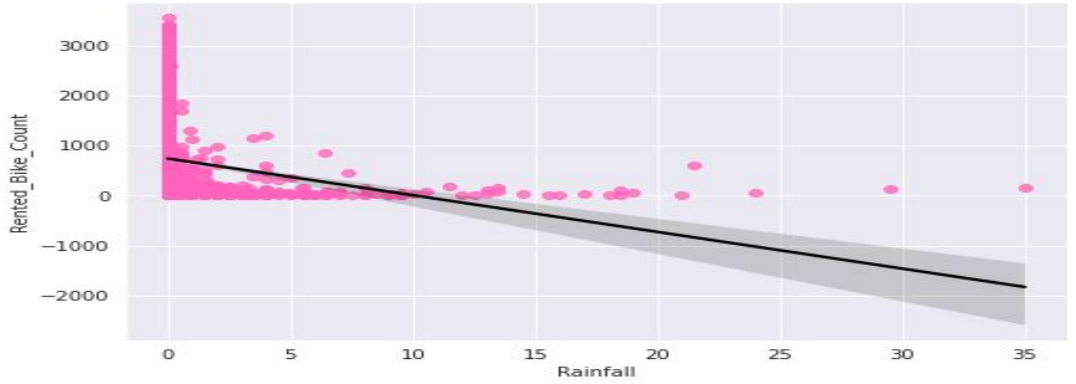
- As we can see the regplot we can say that If there is low visibility, people won't prefer to ride the bike. So,as the visibility increases , the number of bike count also increases.

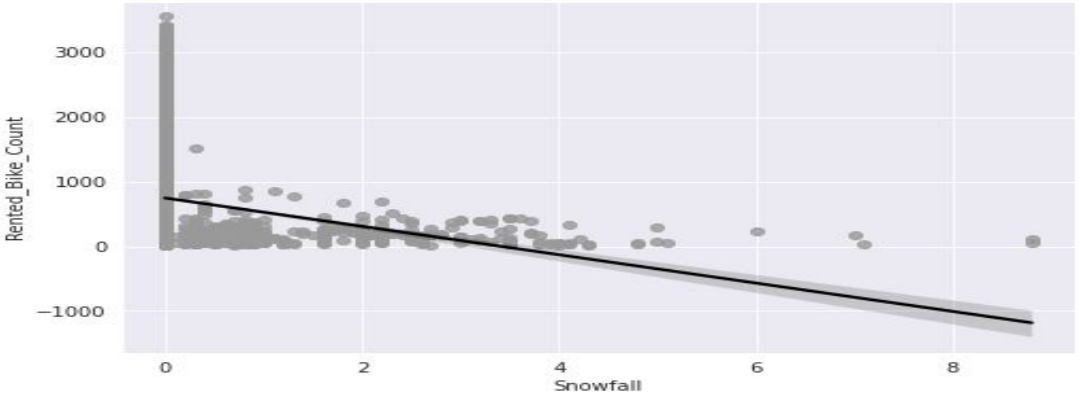
- From the given regplot we can see that people like to ride bikes when dew point temp increases.



Analysis of Dependent Variable Vs Rainfall & Snowfall:

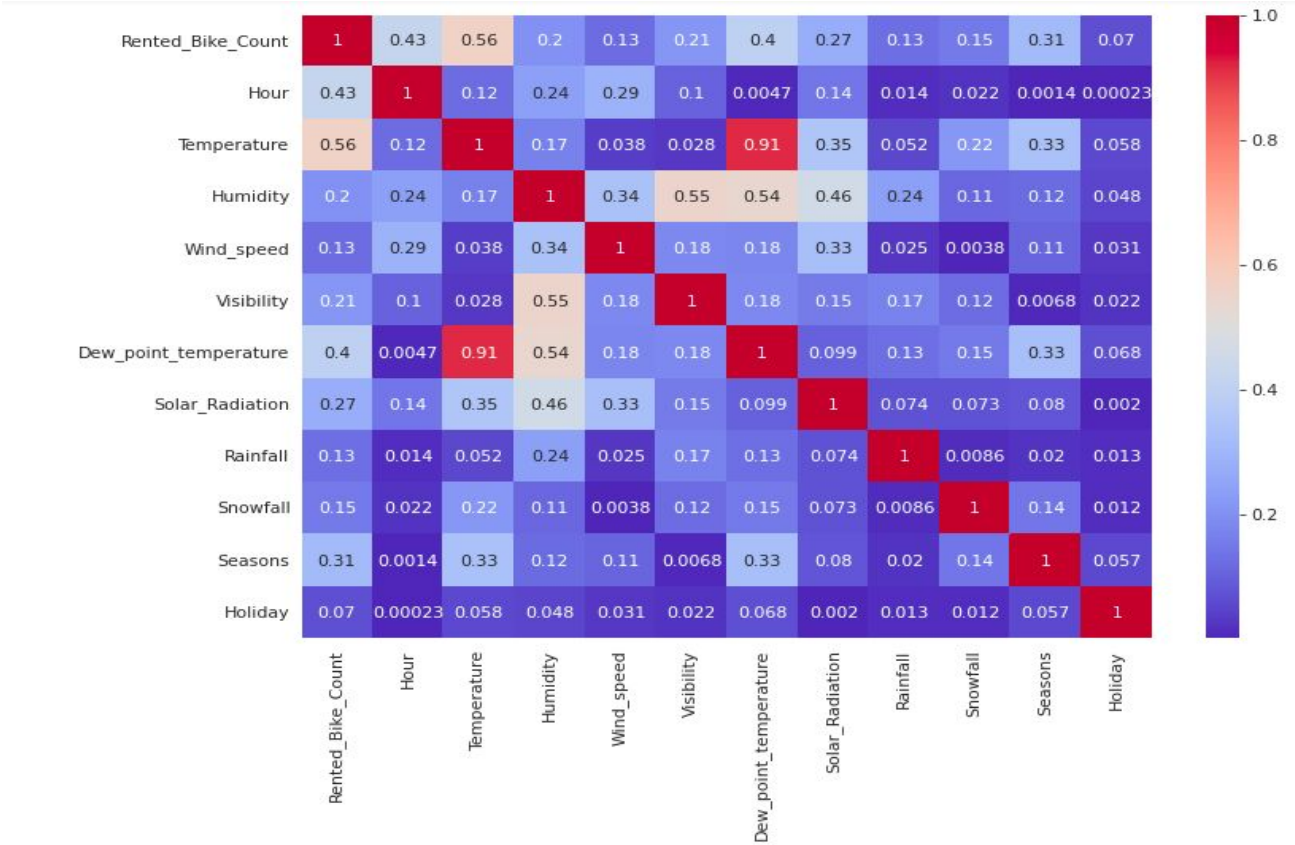
- If there is rainfall/Snowfall, people don't prefer to travel out. And, hence the bike count decreases.
- The amount of rented bike is very low When we have more than 4 cm of snow, the bike rents is much lower.





Correlation Matrix:

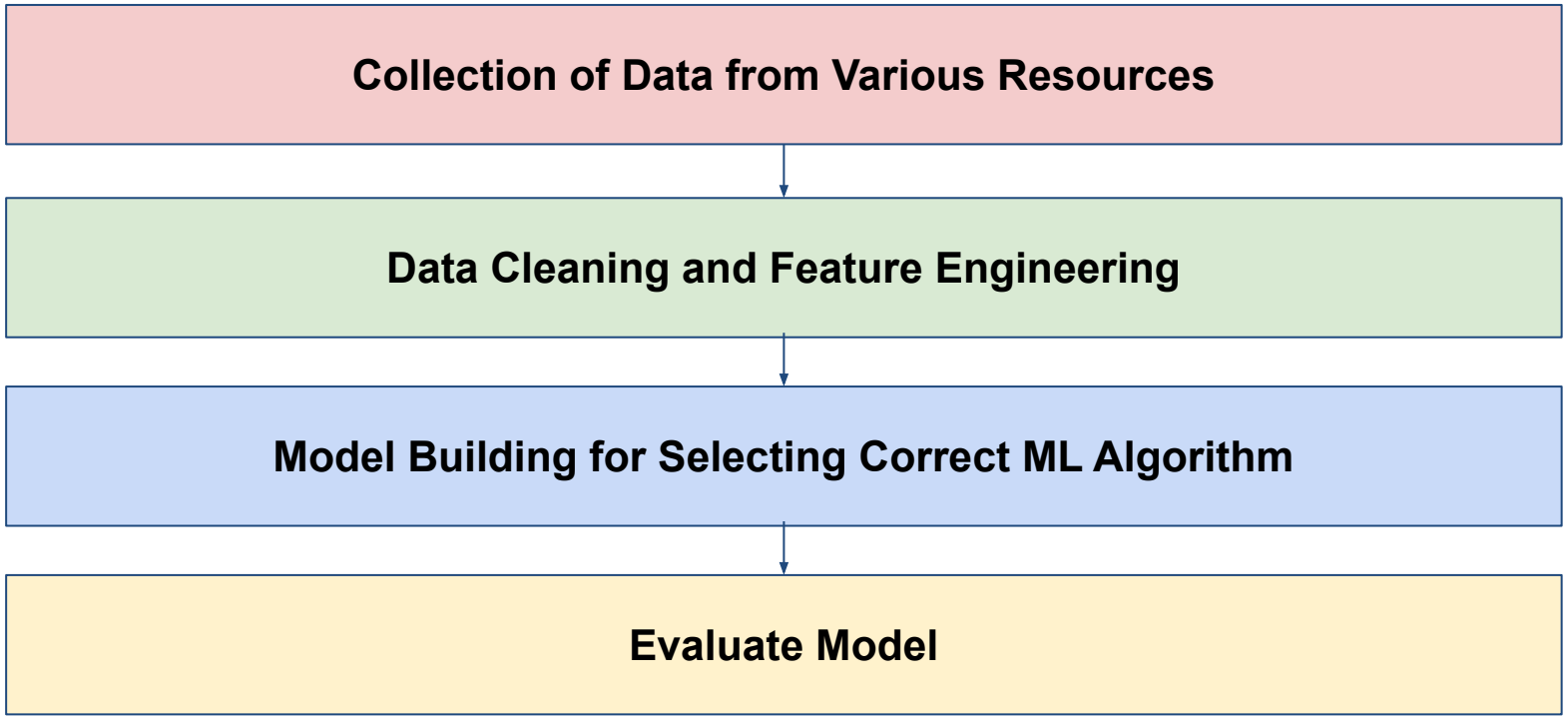
- **Variables like Dew Point Temperature, and Temperature are highly correlated.**



Model Building:

- Linear Regression
- Lasso Regression
- Ridge Regression
- Decision Tree Regression
- Random Forest Regression
- Gradient Boosting Regression
- Gradient Boosting Regression with GridSearchCV
- XGB Regression

Machine Learning Process Flow:



Hyperparameter tuning:

- Lasso Regression

```
Lasso(alpha=0.01)
```

- Ridge Regression

```
Ridge(alpha=0.1)
```

- Decision Tree Regression

```
DecisionTreeRegressor(criterion='mse', max_depth=8, max_features=9,
                      max_leaf_nodes=100)
```

- Random Forest Regression

```
RandomForestRegressor(max_depth=6, n_estimators=200)
```

- XGB Regression

```
XGBRegressor(learning_rate=0.01, n_estimators=1000)
```

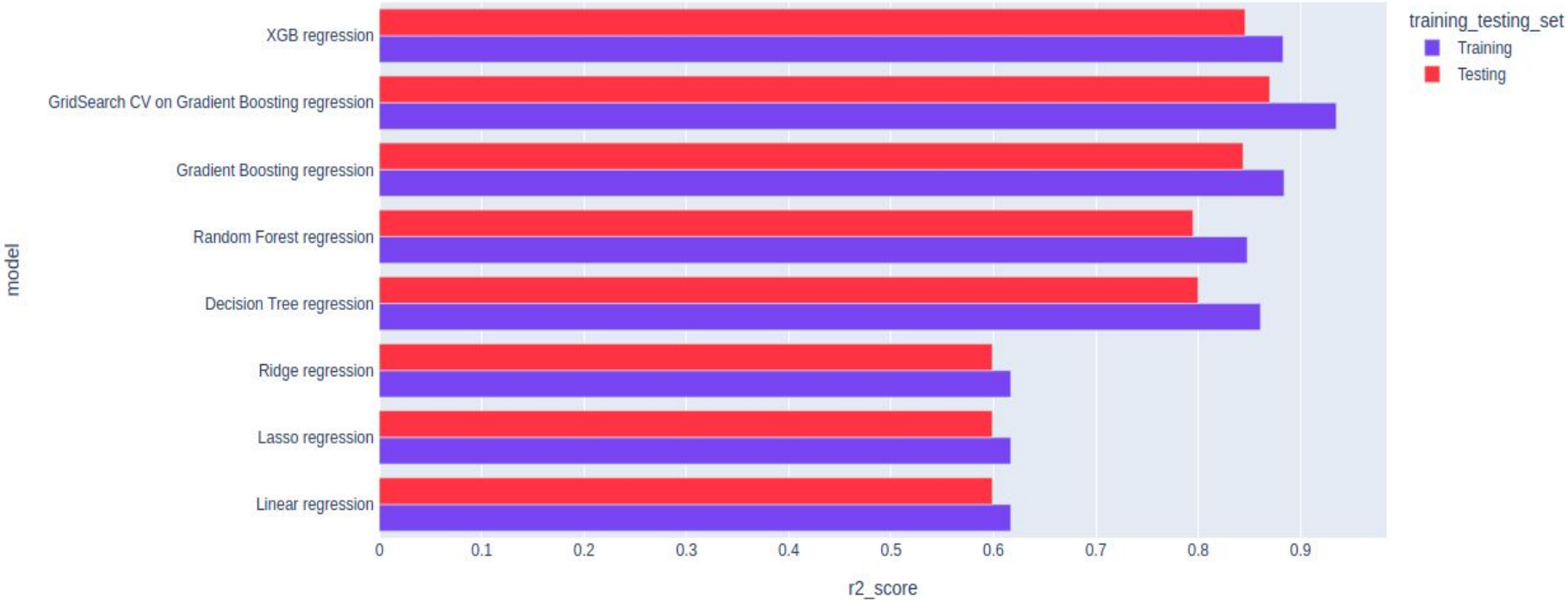
- Gradient Boosting Regression with GridSearchCV

```
{'max_depth': 8,
 'min_samples_leaf': 40,
 'min_samples_split': 50,
 'n_estimators': 100}
```

Evaluating models:

		Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Training set	0	Linear regression	5.634000	54.138000	7.358000	0.617000	0.610000
	1	Lasso regression	5.633000	54.139000	7.358000	0.617000	0.610000
	2	Ridge regression	5.634000	54.138000	7.358000	0.617000	0.610000
	3	Decision Tree regression	3.204000	19.616000	4.429000	0.861000	0.860000
	4	Random Forest regression	3.347000	21.514000	4.638000	0.848000	0.850000
	5	Gradient Boosting regression	2.902000	16.316000	4.039000	0.884000	0.880000
	6	GridSearch CV on Gradient Boosting regression	2.127000	9.207000	3.034000	0.935000	0.930000
	7	XGB regression	2.918000	16.460000	4.057000	0.883000	0.880000
Testing set	0	Linear regression	5.665000	55.285000	7.435000	0.599000	0.600000
	1	Lasso regression	5.666000	55.278000	7.435000	0.599000	0.600000
	2	Ridge regression	5.665000	55.285000	7.435000	0.599000	0.600000
	3	Decision Tree regression	3.682000	27.591000	5.253000	0.800000	0.800000
	4	Random Forest regression	3.774000	28.248000	5.315000	0.795000	0.790000
	5	Gradient Boosting regression	3.281000	21.463000	4.633000	0.844000	0.840000
	6	GridSearch CV on Gradient Boosting regression	2.886000	17.903000	4.231000	0.870000	0.870000
	7	XGB regression	3.270000	21.310000	4.616000	0.846000	0.840000

Comparing different ML Models:



- Linear regression,Lasso regression,Ridge regression have similar R2 score.
- **XGB regression**, **Gradient Boosting regression**, **Decision Tree regression** and **GridSearchCV on Gradient Boosting regression** has >=80% R2_score on testing set.
- **XGB regression** and **Gradient Boosting regression** are performing very well.

Challenges:

- Large Dataset to handle.
- Needs to plot lot of Graphs to analyse.
- Feature engineering
- Feature selection
- Optimising the model
- Carefully tuned Hyperparameters as it affects the R2 score.

Conclusion:

- Bike rental count is mostly correlated with the time of the day as it is peak at 8 am morning and 6 pm at evening.
- We observed that bike rental count is high during working days than non-working day.
- We see that people generally prefer to bike at moderate to high temperatures, and when little windy.
- It is observed that highest number bike rentals counts in Autumn & Summer seasons & the lowest in winter and spring season.
- We observed that the highest number of bike rentals on a clear day and the lowest on a snowy or rainy day and also increasing humidity, the number of bike rental counts decreases.
- Linear regression,Lasso regression,Ridge regression have similar R2 score.
- XGB regression, Gradient Boosting regression, Decision Tree regression and GridSearchCV on Gradient Boosting regression has >=80% R2_score on testing set.
- XGB regression and Gradient Boosting regression are performing very well.


thank you