Capstone Project-2

Supervised ML (Regression) - Capstone Project

Bike Sharing Demand Prediction

Team Members:

Email id	Name
chetanjadhav2341@gmail.com	Chetan Jadhav
meghanars70@gmail.com	Meghana Rs
regorobin5@gmail.com	Robin Rego
poojaparsana158@gmail.com	Pooja Parsana

Project Details

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.

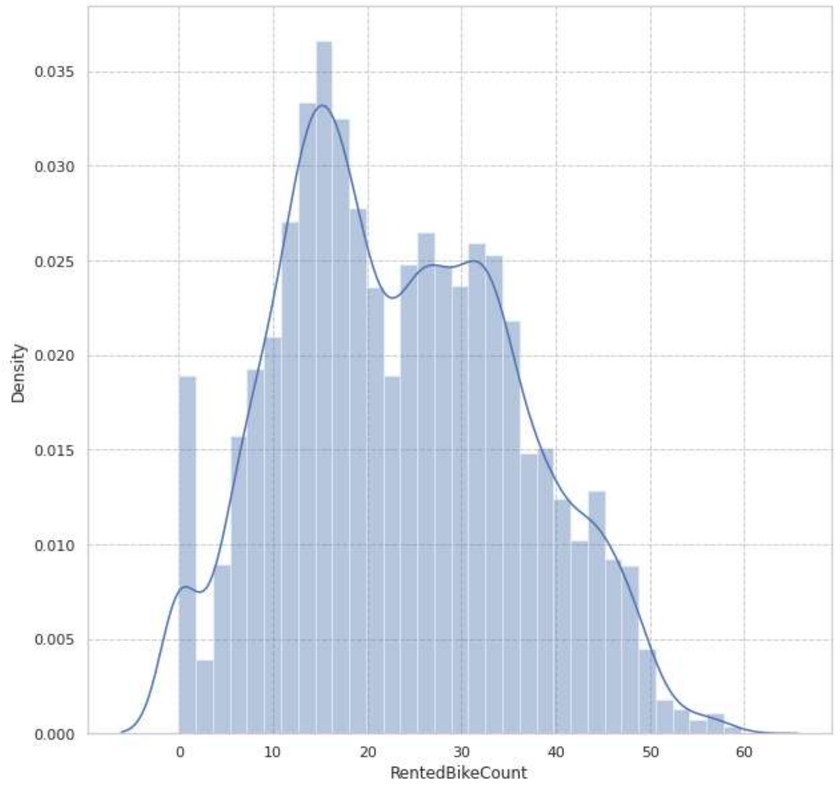
Steps performed

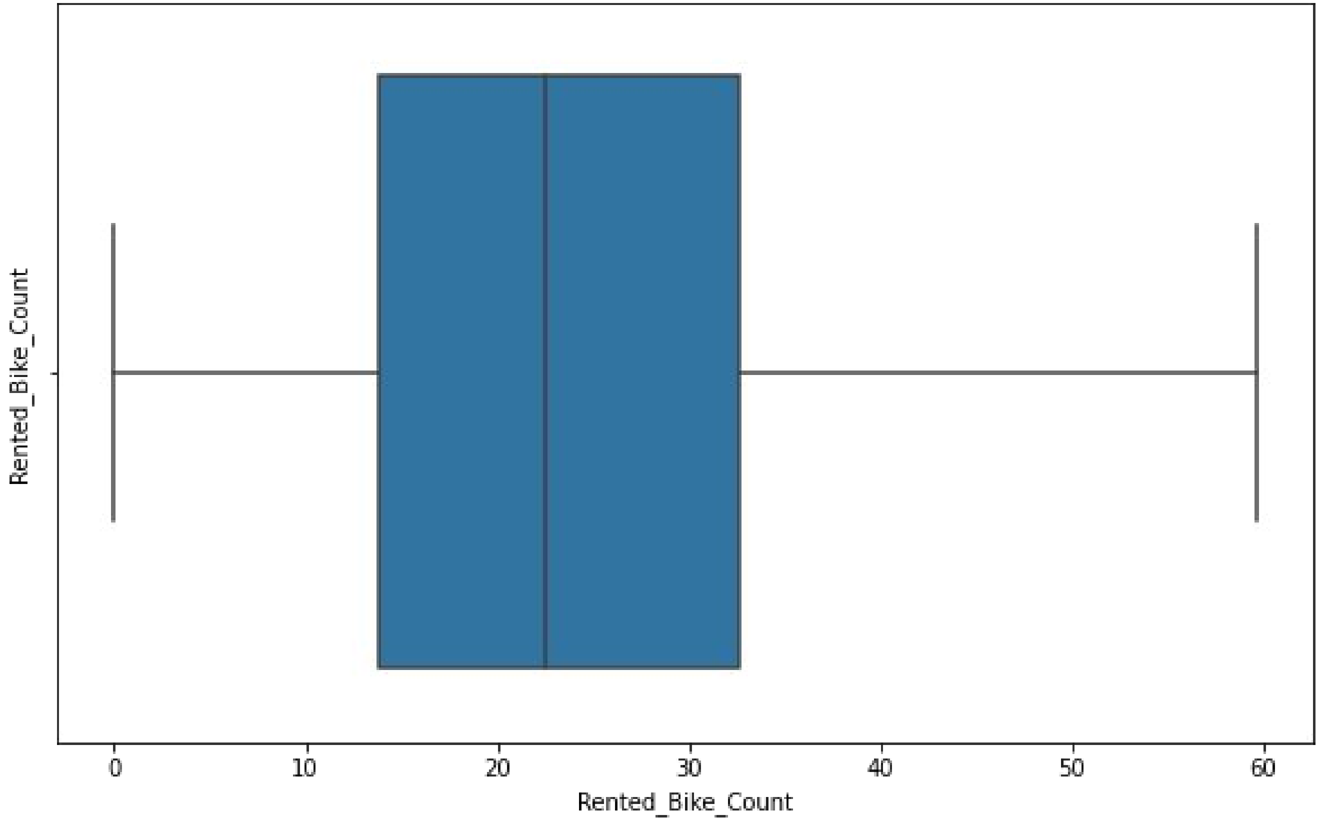
- Data cleaning
- Data visualizations
- Data preprocessing
- Model Implementation
- Evaluation metrics

Feature Summary

- Date: year-month-day
- Rented Bike Count -Count of bikes rented at each hour
- Hour: Hour of the day
- Temperature: (in Celsius)
- Humidity: (in %)
- Windspeed: m/s
- Visibility: 10m
- Dew Point Temperature (in celsius)
- Solar Radiation: MJ/m2
- Rainfall: mm
- Snowfall: cm
- Seasons : Winter, Spring, Summer,
- Autumn
- Holiday - No Holiday/ Holiday
- Functional Day - Yes/No

Analysis Of Rented Bike Column

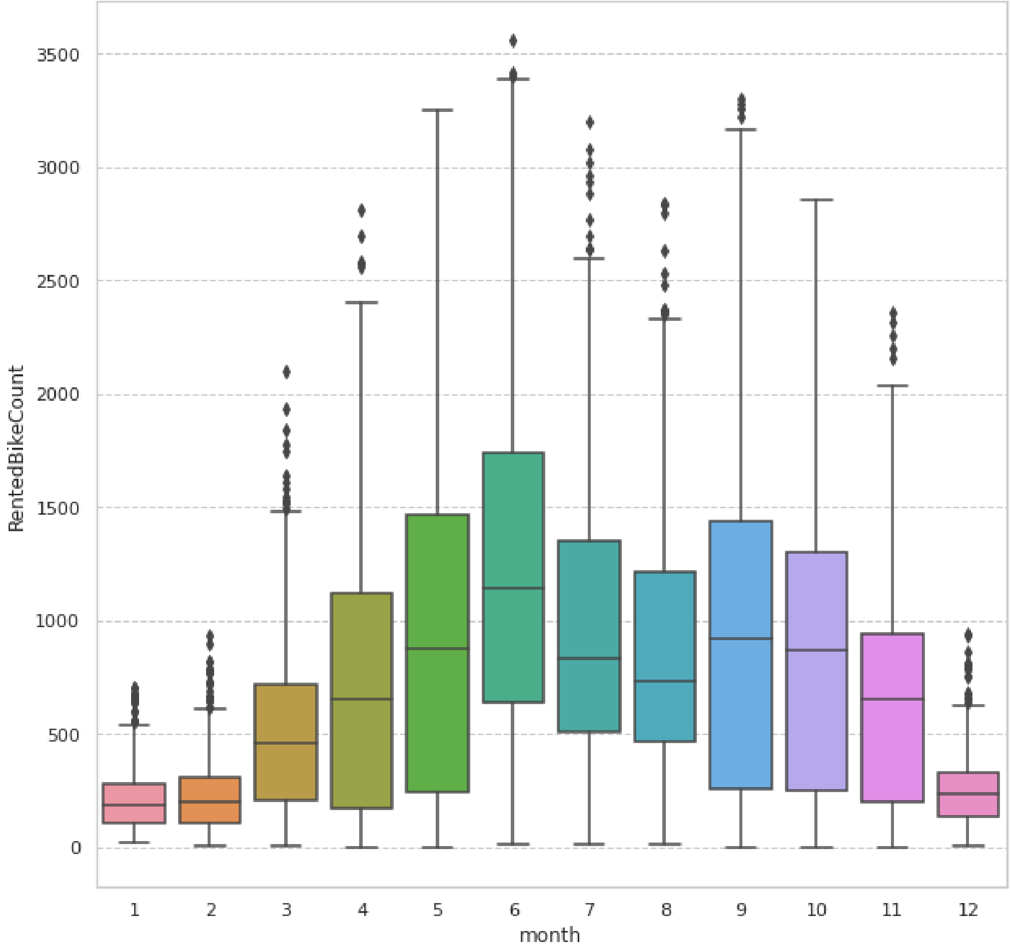




- **After applying square root transformation there is no outliers present in rented bike count column**

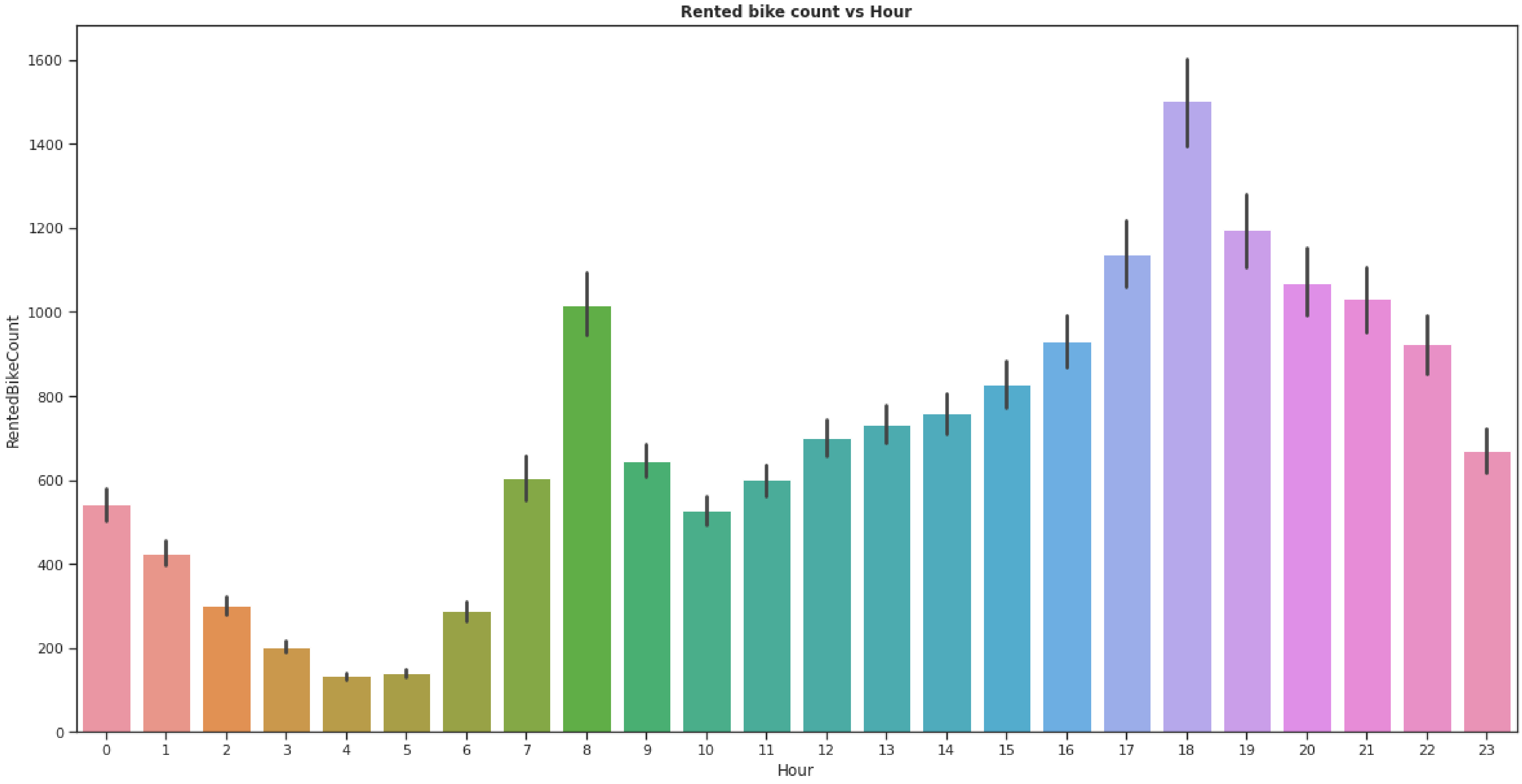
Exploratory Data Analysis

Month vs Rental Bike Count



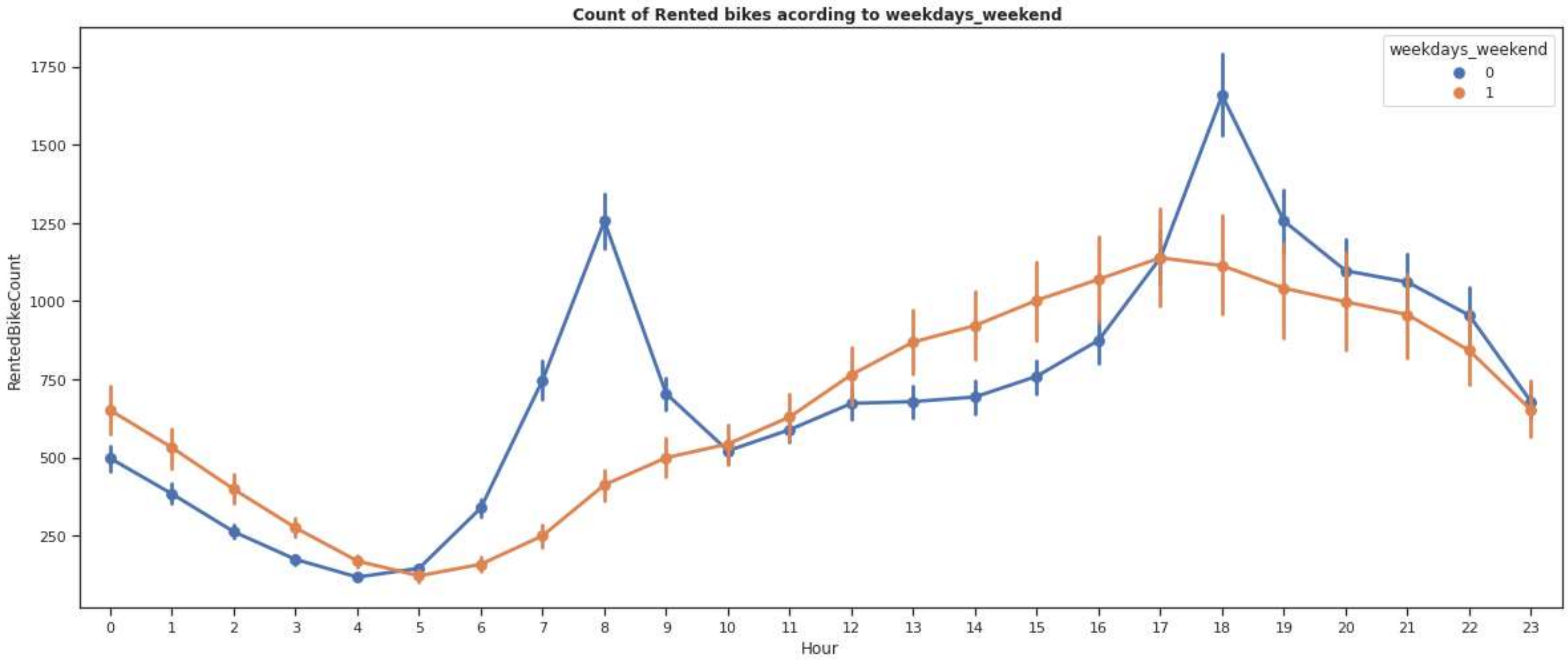
- **From above graph it is clear that the from the month 5 to 10 demand was increases compared to the other month. or this is the summer season.**

Hour vs Rental Bike Count



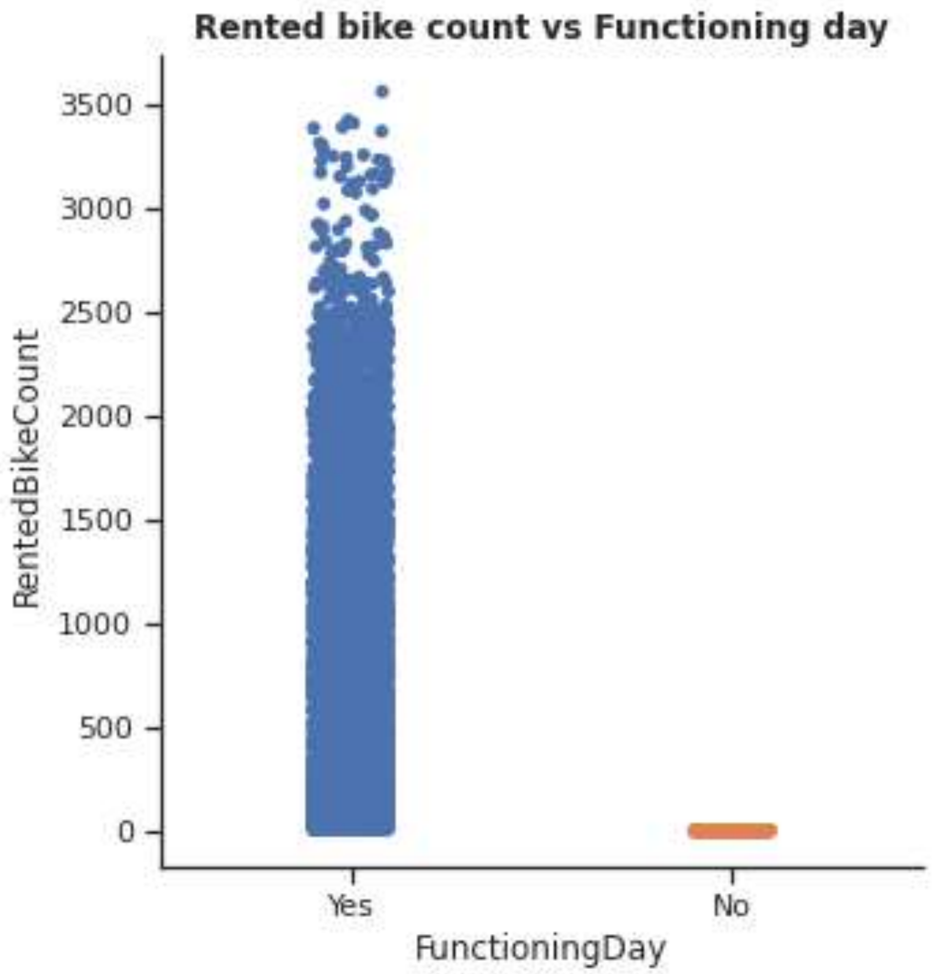
- **From above graph we can say that people use rented bikes during their working hour from7am to 9am and 5pm to 7pm.**

Weekdays and weekend vs Rental Bike Count



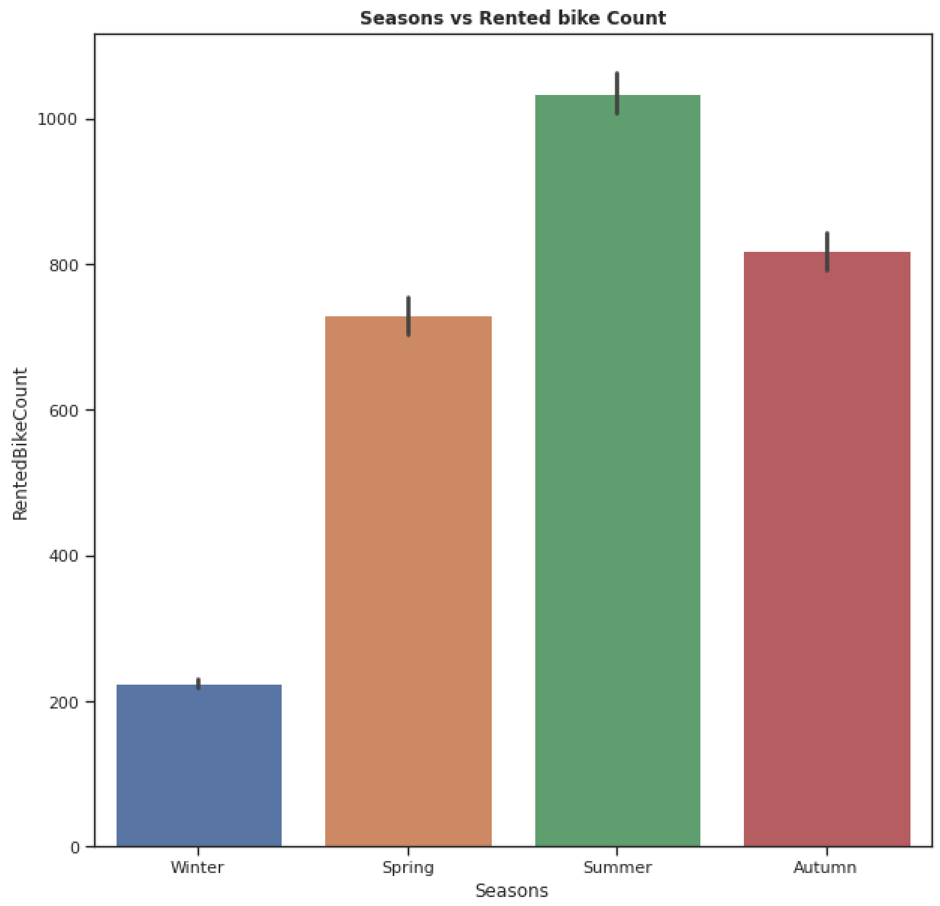
- From the above point plot we can say that in the weekdays which represent in blue colour show that the demand of the bike higher because of the office.
- The orange colour represent the weekend days, and it show that the demand of rented bikes are very low especially in the morning hour but when the evening start from 4 pm to 8 pm the demand slightly increases.

Functioning Day vs Rented bike count



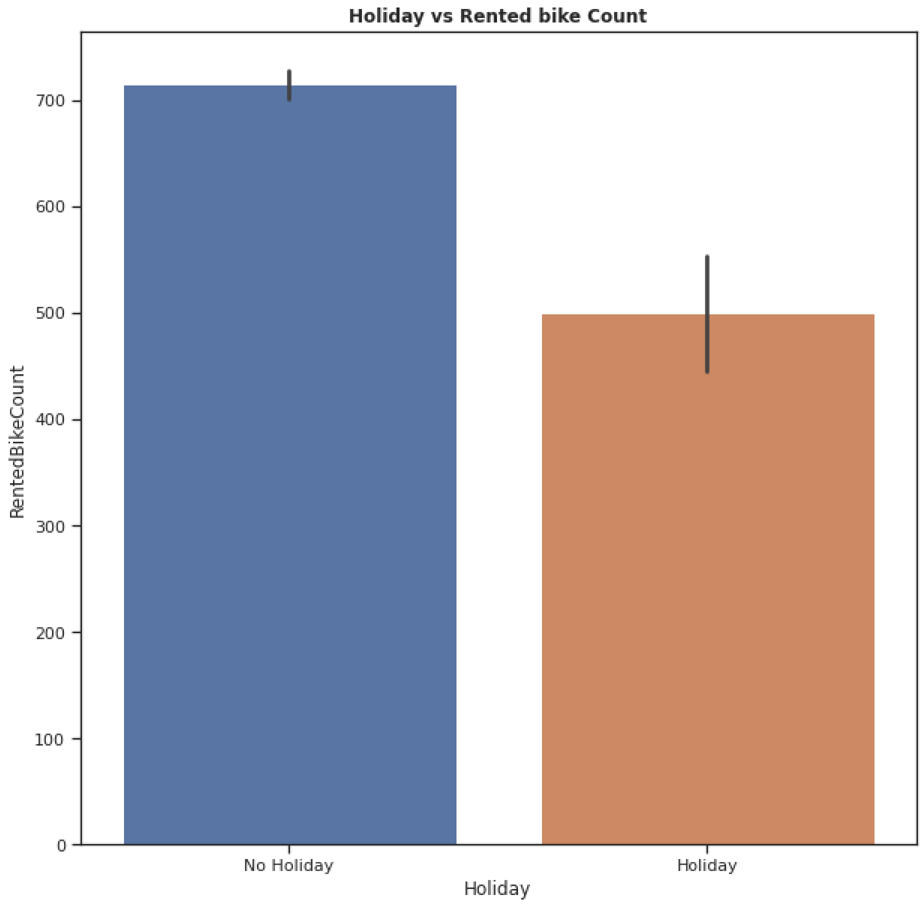
- **Above graph clearly shows that the, people don't use rented bike in no functioning day.**

Analysis Of Season Variable



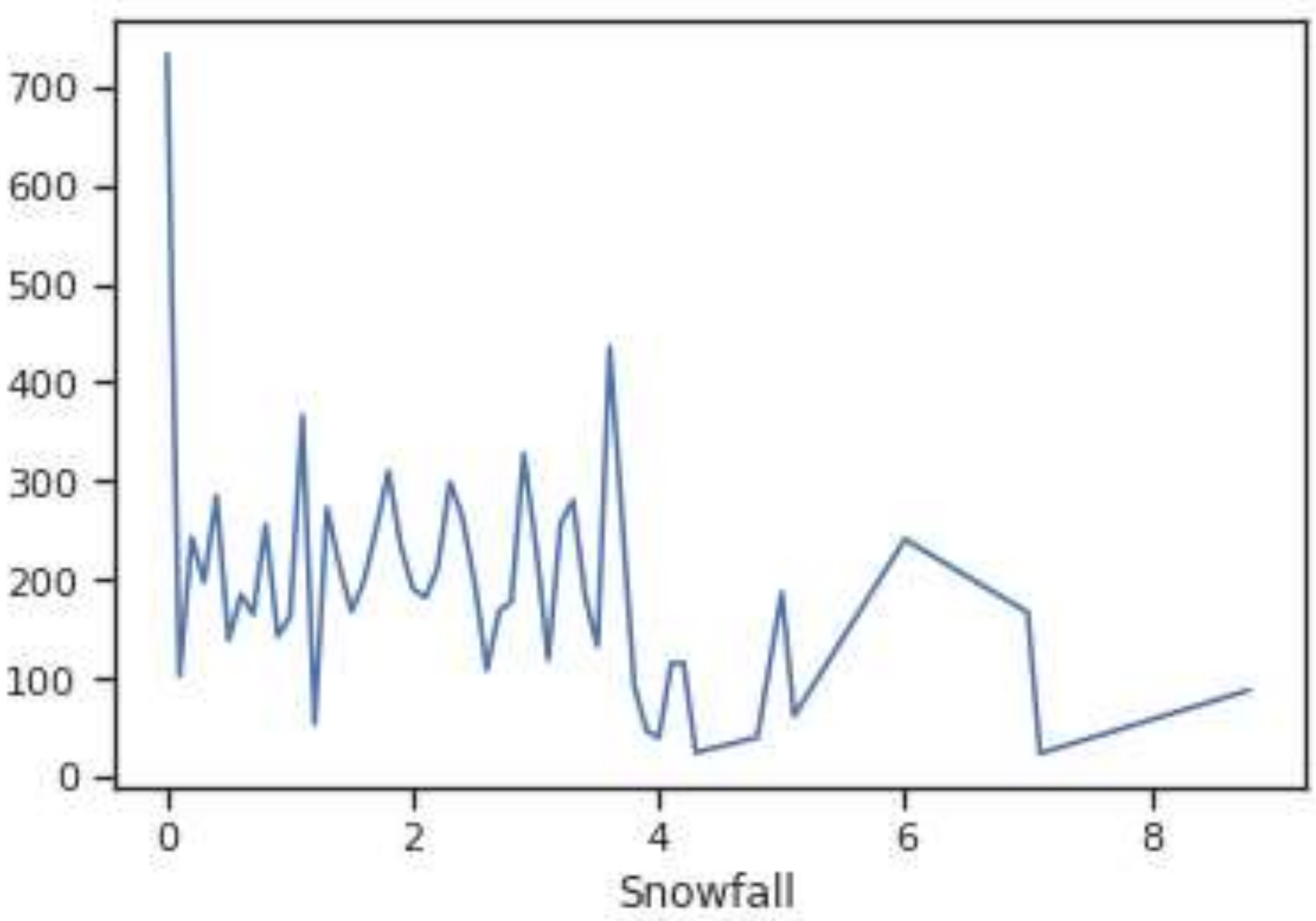
- In above graph we can see that the during summer season demand of rented bike was increase.
- And during the winter demand is very low.

Analysis of Holiday Variable



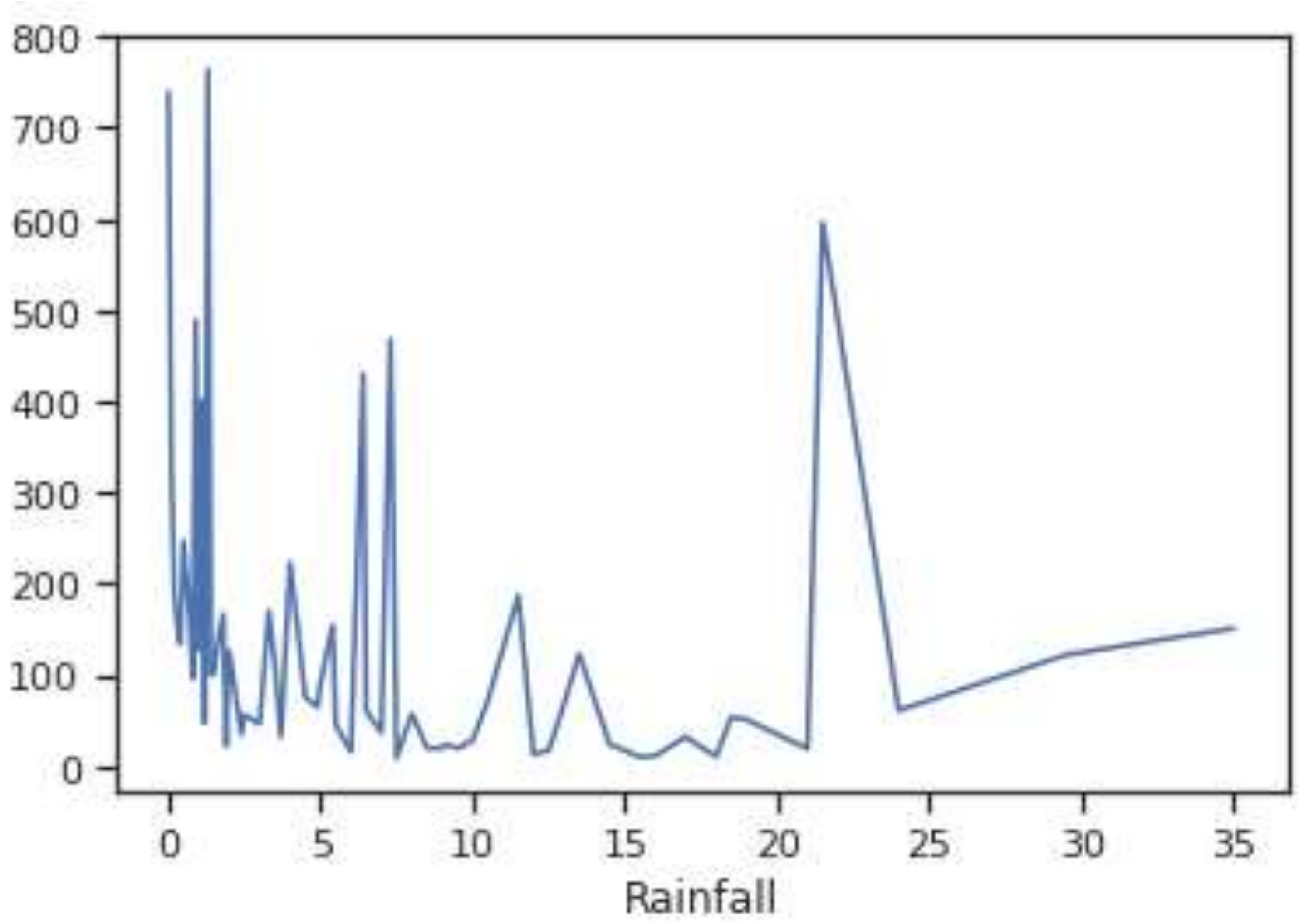
- **From above graph it is clear that the during a holiday demand of rented bike was decrease compared to the no holidays.**

Snowfall vs Rented Bike Count



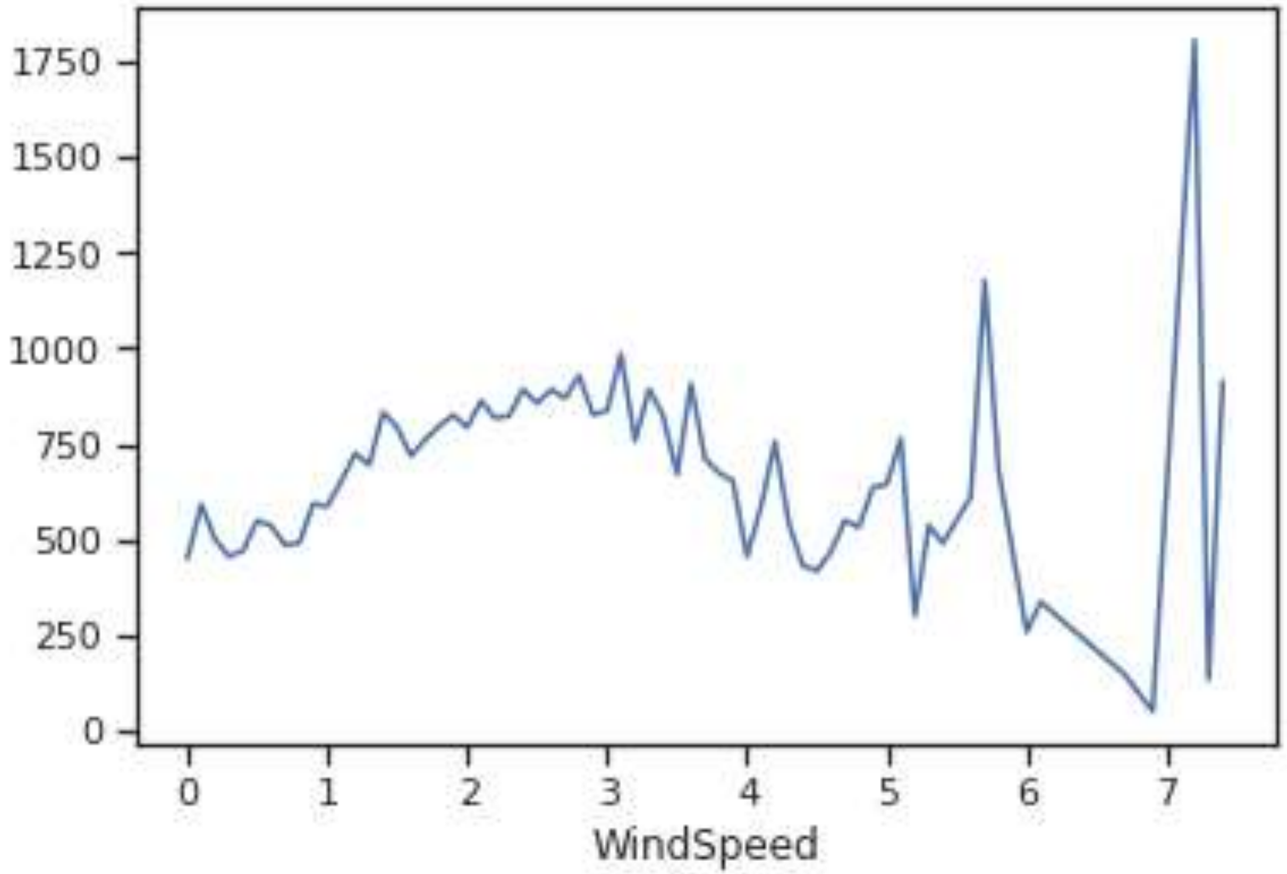
- **Graph shows increase in snowfall demand was decreases.**

Rainfall vs Rented Bike Count



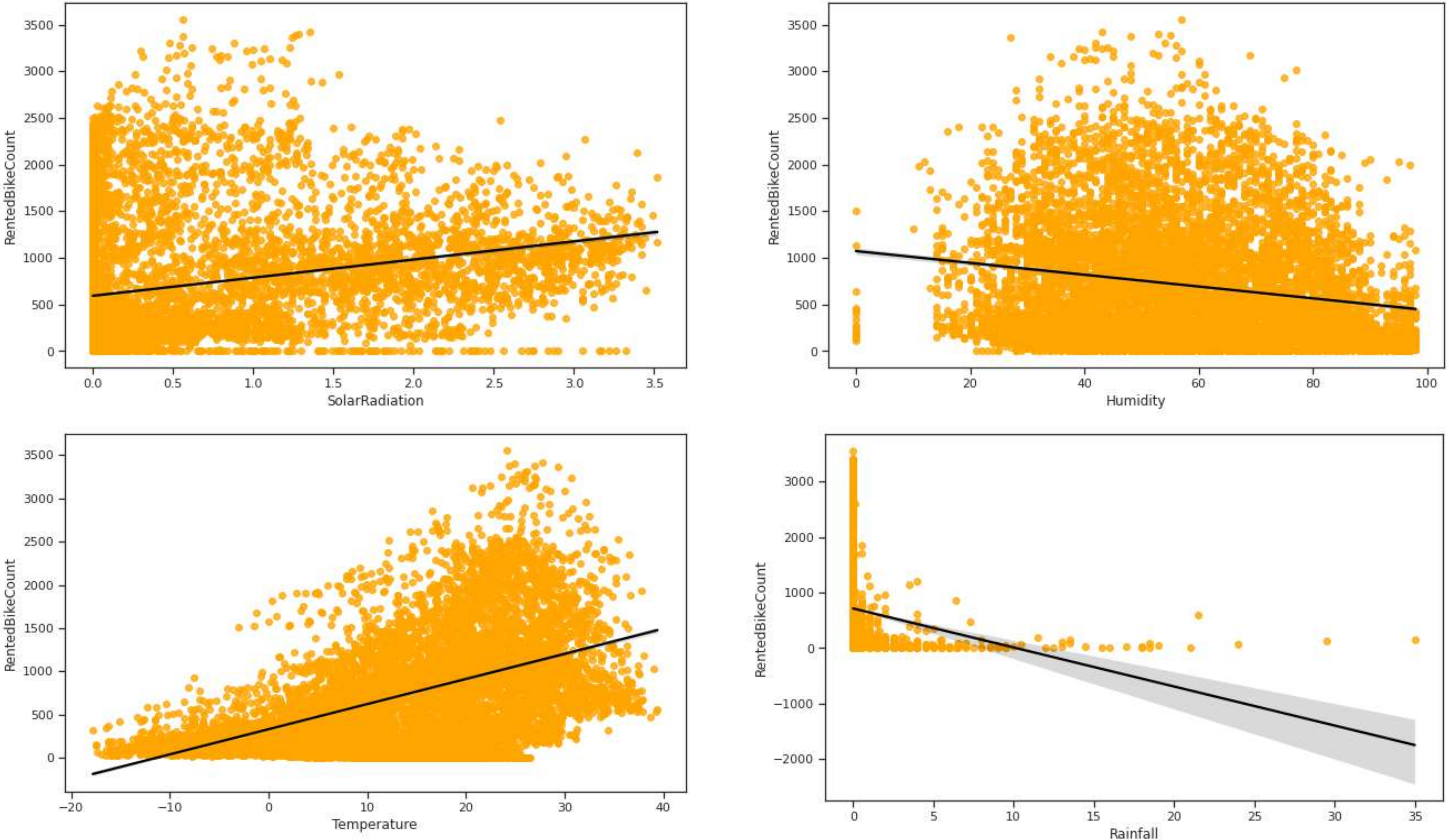
- **just like snowfall, when rainfall is increases demand of rented bike is decreases.**

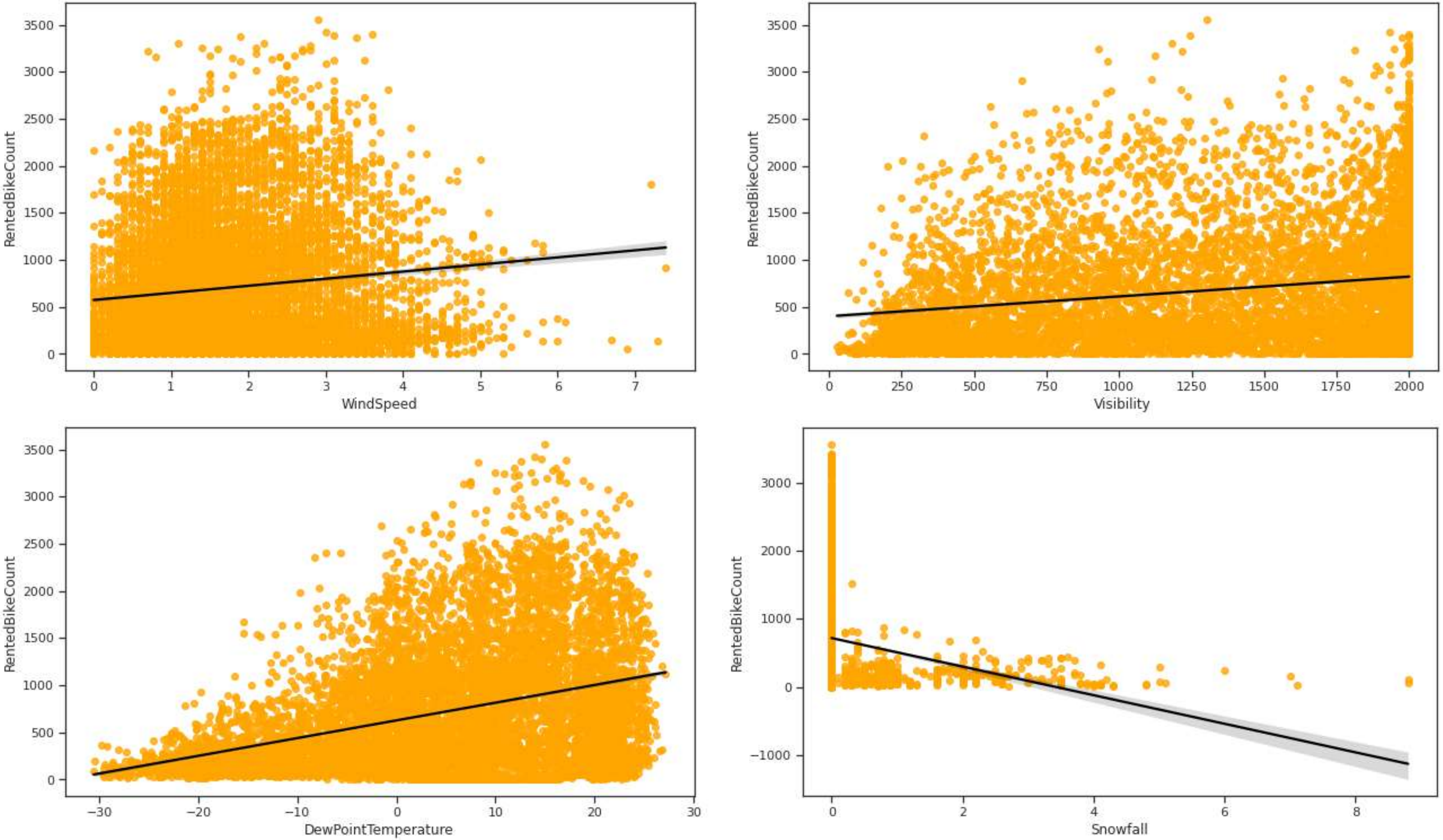
Windspeed vs Rented Bike Count



- 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.

Regression plots for numerical variable



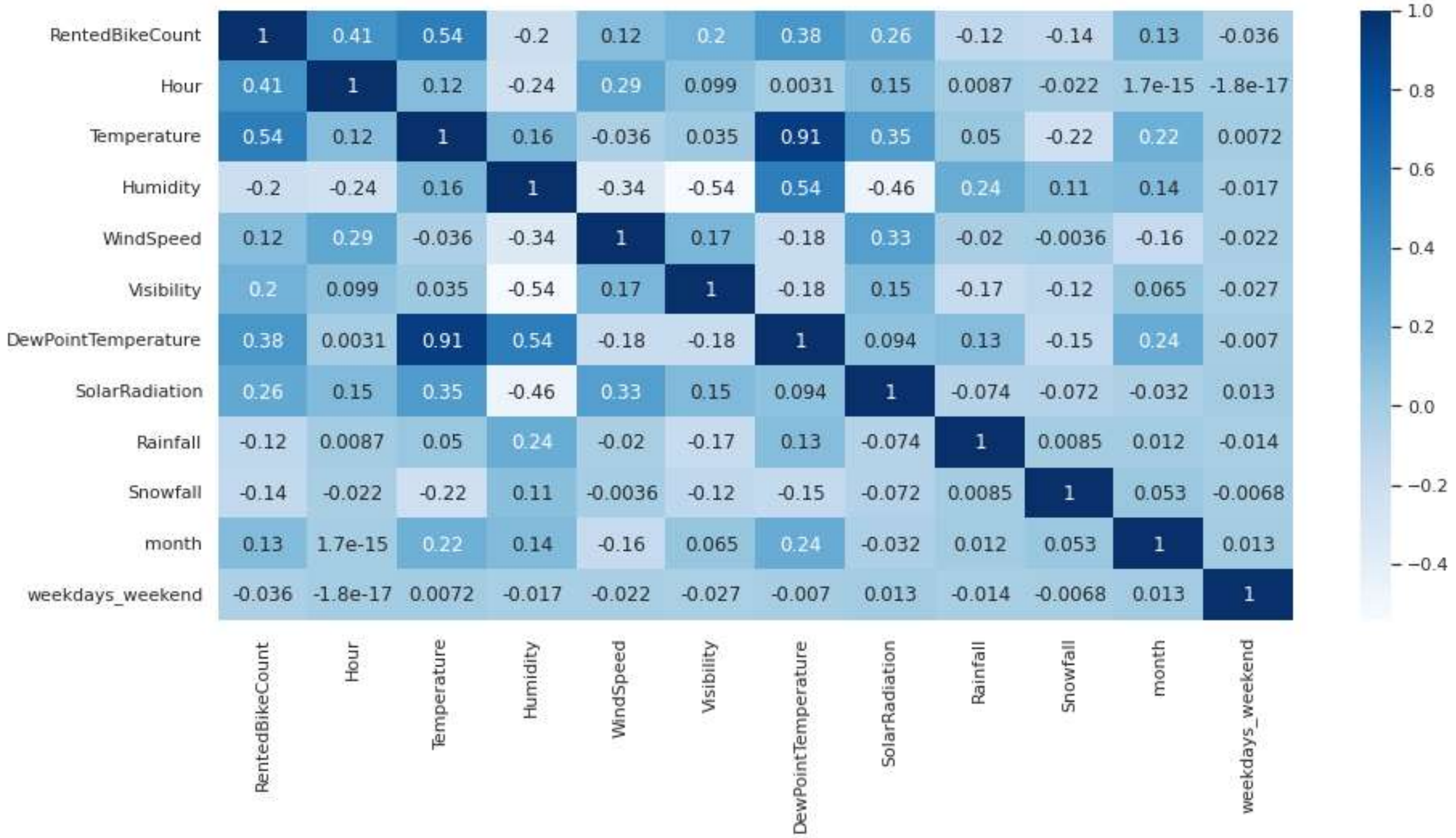


RentedBikeCount
WindSpeed
Visibility
DewPointTemperature
Snowfall

Analysis from regression plot of numerical variable

- **From the above regression plot of all numerical features we see that the columns 'Temperature', 'Wind_speed','Visibility', 'Dew_point_temperature', 'Solar_Radiation' are positively relation to the target variable.**
- **which means the rented bike count increases with increase of these features.**
- **'Rainfall', 'Snowfall', 'Humidity' these features are negatively related with the target variable which means the rented bike count decreases when these features increase.**

Heatmap



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

Model Building

- LINEAR REGRESSION
- LASSO REGRESSION
- RIDGE REGRESSION
- DECISION TREES REGRESSOR
- RANDOM FOREST REGRESSOR
- GRADIENT BOOSTED REGRESSOR
- GRADIENT BOOSTING REGRESSOR WITH GRIDSEARCHCV

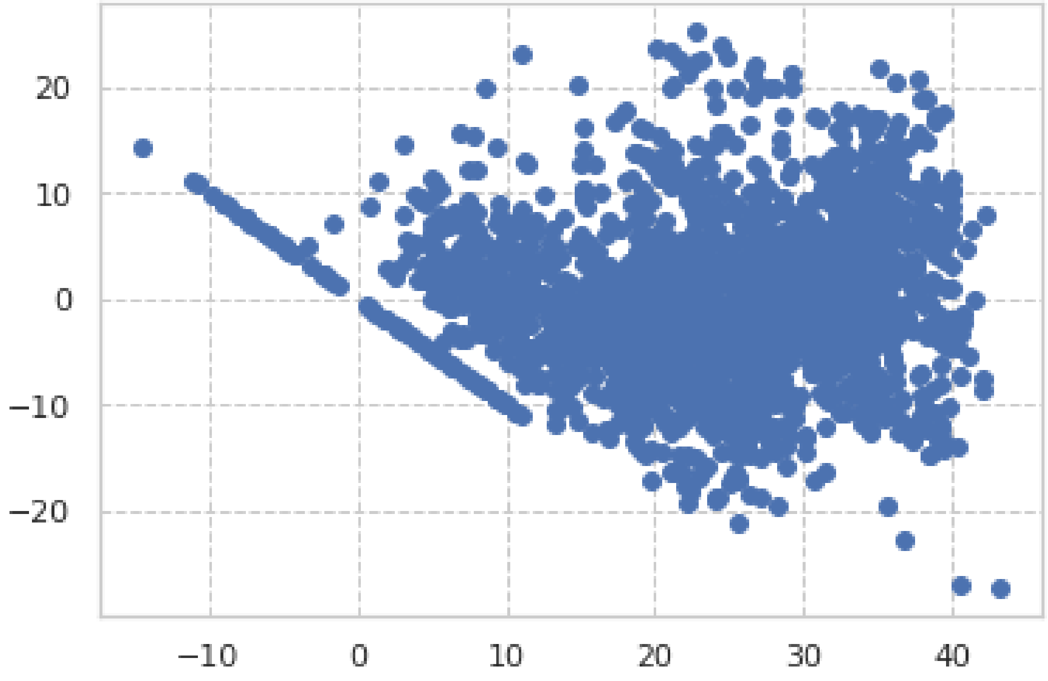
Linear Regression

Training Dataset Result

```
MSE : 53.39656510174565
RMSE : 7.307295334235893
MAE : 5.608650235369007
R2 : 0.6568484454l4216
Adjusted R2:0.65432178877667
```

Test Dataset Result

```
MSE : 51.969197542857664
RMSE : 7.208966468423727
MAE : 5.5449337368940235
R2 : 0.660738066202897
Adjusted R2 : 0.65824004920301
```



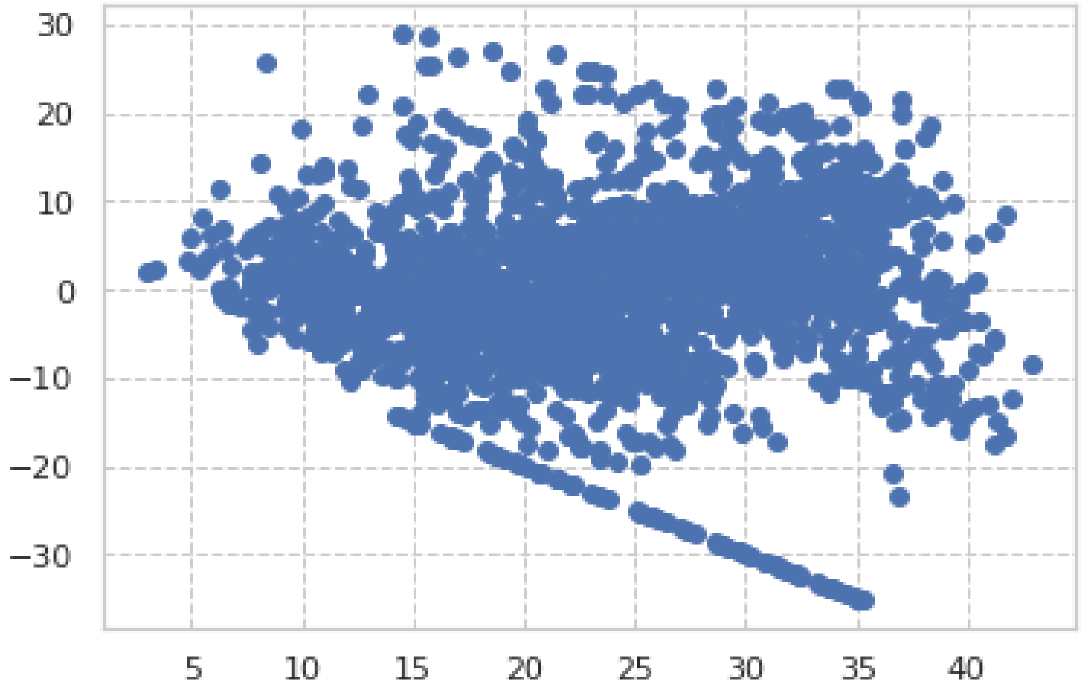
Lasso Regression

Training Dataset Result

MSE : 81.67121948733921
RMSE : 9.037213037620571
MAE : 6.714463134100375
R2 : 0.47514215795350834
Adjusted R2: 0.47127758111377

Test Dataset Result

MSE : 82.83241723530892
RMSE : 9.101231632878537
MAE : 6.695468179147453
R2 : 0.4592588036564451
Adjusted R2 : 0.455277276209828



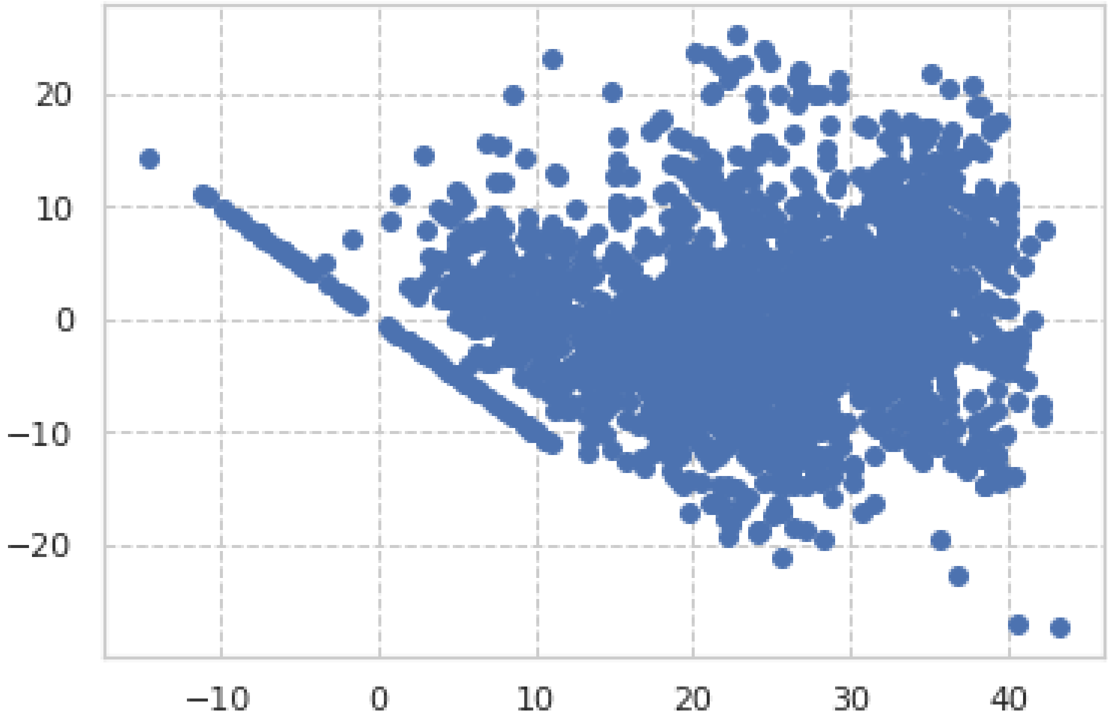
Ridge Regression

Training Dataset Result

```
MSE : 53.39657222149276
RMSE : 7.307295821402933
MAE : 5.608667022744986
R2 : 0.6568483996593574
Adjusted R2 : 0.654321742684921
```

Test Dataset Result

```
MSE : 51.97020619182089
RMSE : 7.209036426029549
MAE : 5.545003364375157
R2 : 0.6607314816063674
Adjusted R2 : 0.65823341612348
```



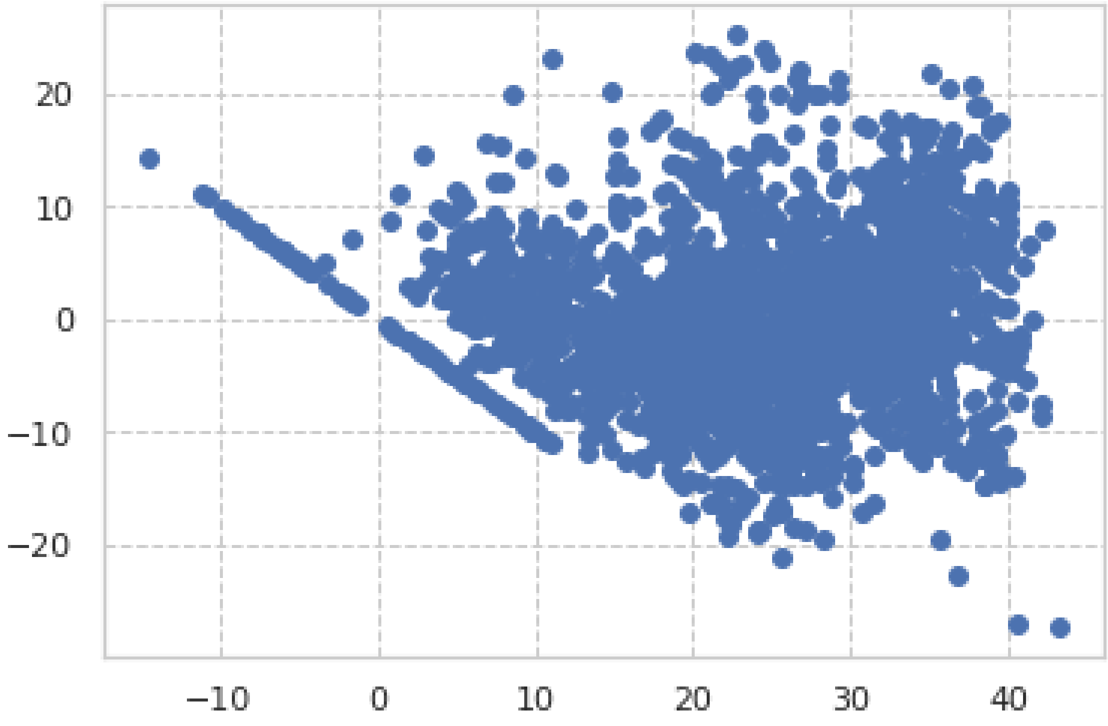
Decision Tree

Training Dataset Result

```
Model Score: 0.8478993907609244
MSE : 23.66782249625263
RMSE : 4.864958632532514
MAE : 3.501757760905517 7
R2 : 0.8478993907609244
Adjusted R2 : 0.8467794599059657
```

Test Dataset Result

```
MSE : 27.205370800338237
RMSE : 5.215876800724 71
MAE : 3.720949541591801 7
R2 : 0.8223996685771711
Adjusted R2 : 0.8210919809090784
```



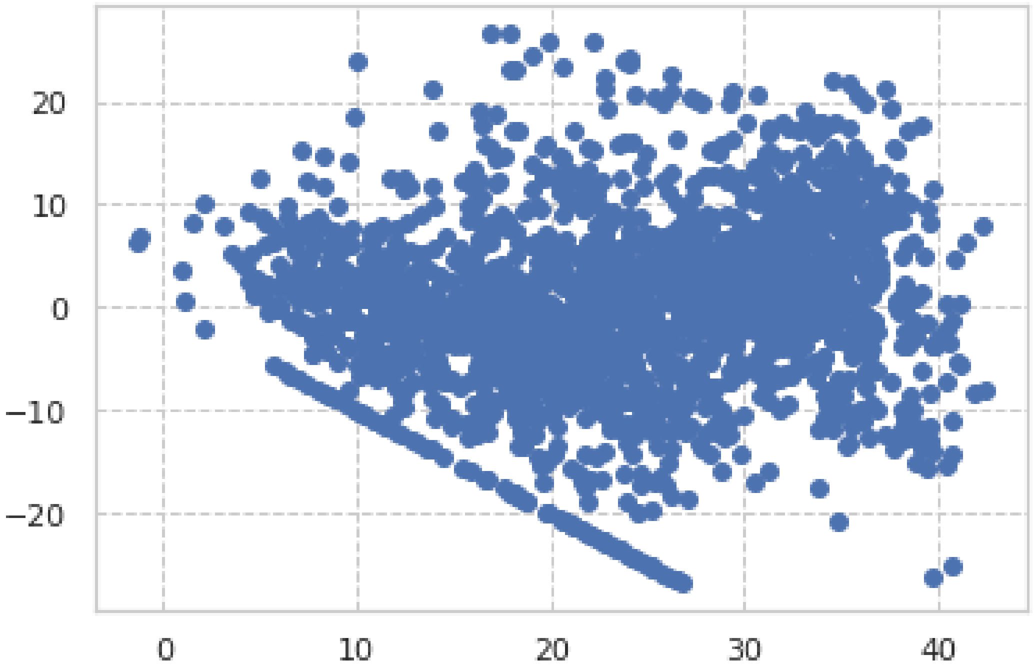
Elastic net Regression

Training Dataset Result

MSE : 64.43270985216326
RMSE : 8.026998807285526
MAE : 6.095440247072635
R2 : 0.585924965214 2559
Adjusted R2 :0.5828760924316

Test Dataset Result

MSE : 64.58120546973048
RMSE : 8.036243243564 15
MAE : 6.079004936296894 5
R2 : 0.578405176709 9109
Adjusted R2 : 0.57530093502898



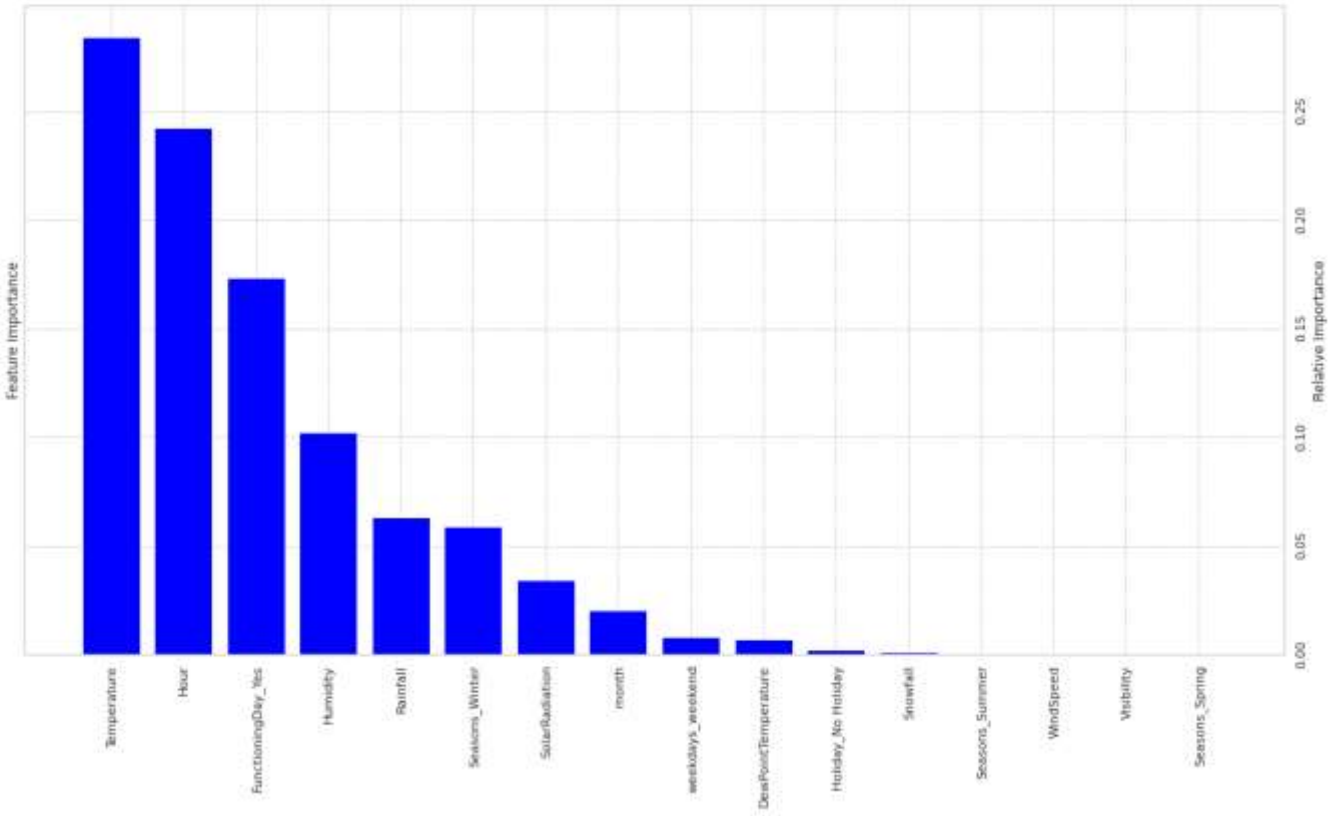
Gradient Boosting

Training Dataset Result

Model Score: 0.9001355144
MSE : 15.539549310552909
RMSE : 3.942023504566264
MAE : 2.821389156564207
R2 : 0.9001355144601924
Adjusted R2 : 0.8994002030

Test Dataset Result

MSE : 17.070899205089862
RMSE : 4.131694471411985
MAE : 2.945146990371217
R2 : 0.888558866017383
Adjusted R2 : 0.887738315228



Model Conclusion

level_0	level_1	Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Training set	0	Linear regression	5.609	53.397	7.307	0.657	0.65
Training set	1	Lasso regression	6.714	81.671	9.037	0.475	0.47
Training set	2	Ridge regression	5.609	53.397	7.307	0.657	0.65
Training set	3	Decision tree regression	3.502	23.668	4.865	0.848	0.85
Training set	4	Elastic net regression	6.095	64.433	8.027	0.586	0.58
Training set	5	Gradient boosting regression	2.821	15.54	3.942	0.9	0.9
Training set	6	Gradient Boosting grid search cv	1.452	4.687	2.165	0.97	0.97
Test set	0	Linear regression	5.545	51.969	7.209	0.661	0.66
Test set	1	Lasso regression	6.695	82.832	9.101	0.459	0.46
Test set	2	Ridge regression	5.545	51.97	7.209	0.661	0.66
Test set	3	Decision tree regression	3.721	27.205	5.216	0.822	0.82
Test set	4	Elastic net regression Test	6.079	64.581	8.036	0.578	0.58
Test set	5	Gradient boosting regression	2.945	17.071	4.132	0.889	0.89
Test set	6	Gradient Boosting grid search cv	1.872	9.003	3.0	0.941	0.94

Conclusion

- We analysis that Hour is the most important feature.
- Rented bike count is mostly related with the time of the day as it is peak at 10 am and 8pm.
- Also we observed that rented bike count is high during working days compared to nonworking days.
- 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 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.

- All metrics were evaluated for each model, MSE(Mean Squared Error), MAE (Mean Absolute Error), RMSE(Root Mean squared Error),R2 Score, Adjusted R2 Score.
- At the end, comparison of models stated that some models showed improvement or were able to handle the overfitting issues when hyperparameter tuning was performed.
- Adjusted R2 score was used to compare models as it is a special form of R2 score. Adjusted R2 indicates how well terms fit a curve or line, and also adjusts for the number of terms in a model.