Capstone Project

Seoul Bike Sharing Demand Prediction

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.

Key Steps:

- Defining the problem statement
- EDA and data visualization
- Data preprocessing
- Feature selection
- Preparing Dataset for model
- Applying model
- Model validation and selection

Why is predictive analytics useful for bike sharing company ?

- Bike Availablebality every time and everywhere and avoiding over-capacities.
- Bike position : how and when.
- Reduction of bottlenecks caused by regular bike maintenance.
- High availability of bikes increase the customer satisfaction.

Dataset:

Rows: 8760
Columns: 14

	Date	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Function
0	01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	
1	01/12/2017	204	1	-5.5	38	0.8	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	
2	01/12/2017	173	2	-6.0	39	1.0	2000	-17.7	0.0	0.0	0.0	Winter	No Holiday	
3	01/12/2017	107	3	-6.2	40	0.9	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	
4	01/12/2017	78	4	-6.0	36	2.3	2000	-18.6	0.0	0.0	0.0	Winter	No Holiday	

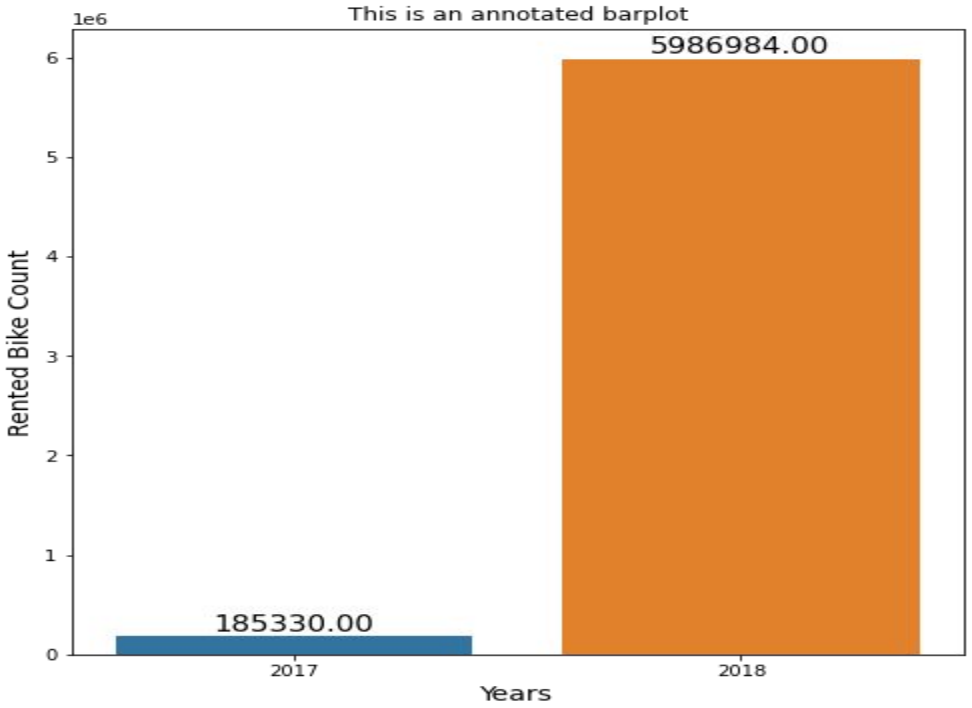
Variable Names:

- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of he day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - 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 - NoFunc(Non Functional Hours), Fun(Functional hours)

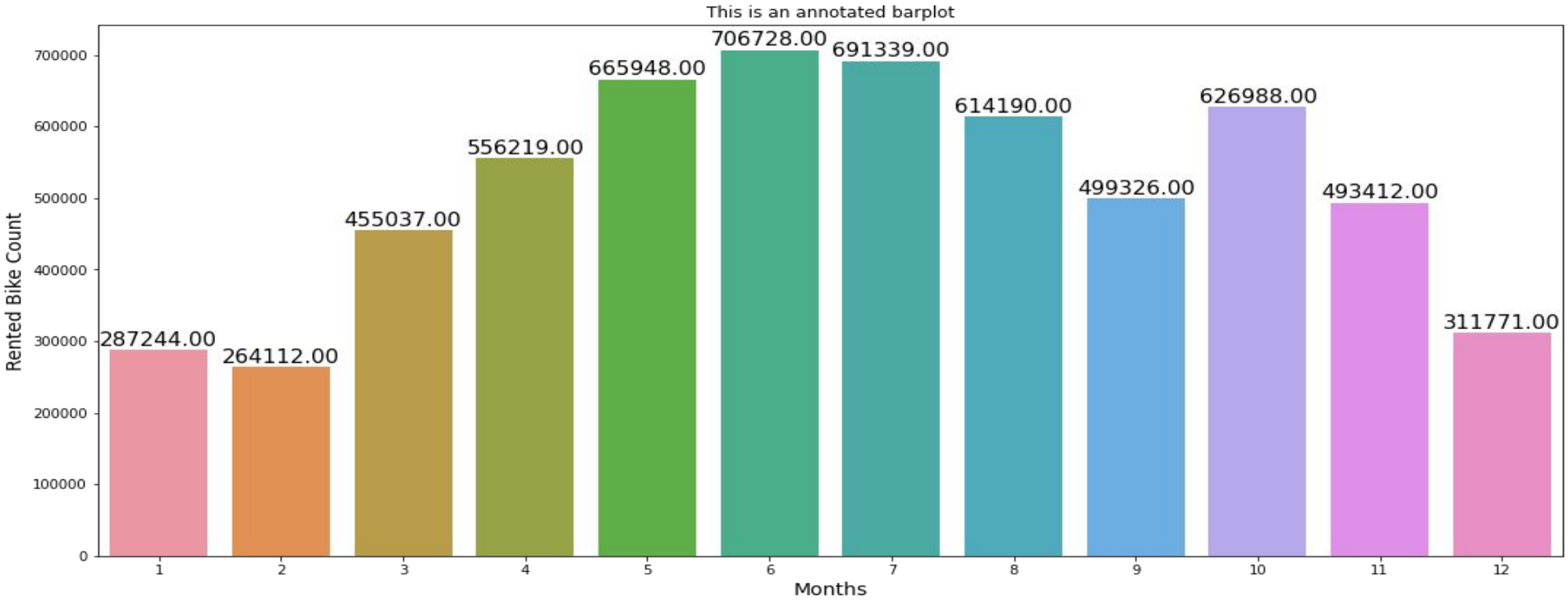
Exploratory Data Analysis:

EDA is used for analyzing what the data can tell us before the modeling or by applying any set of instructions/code.

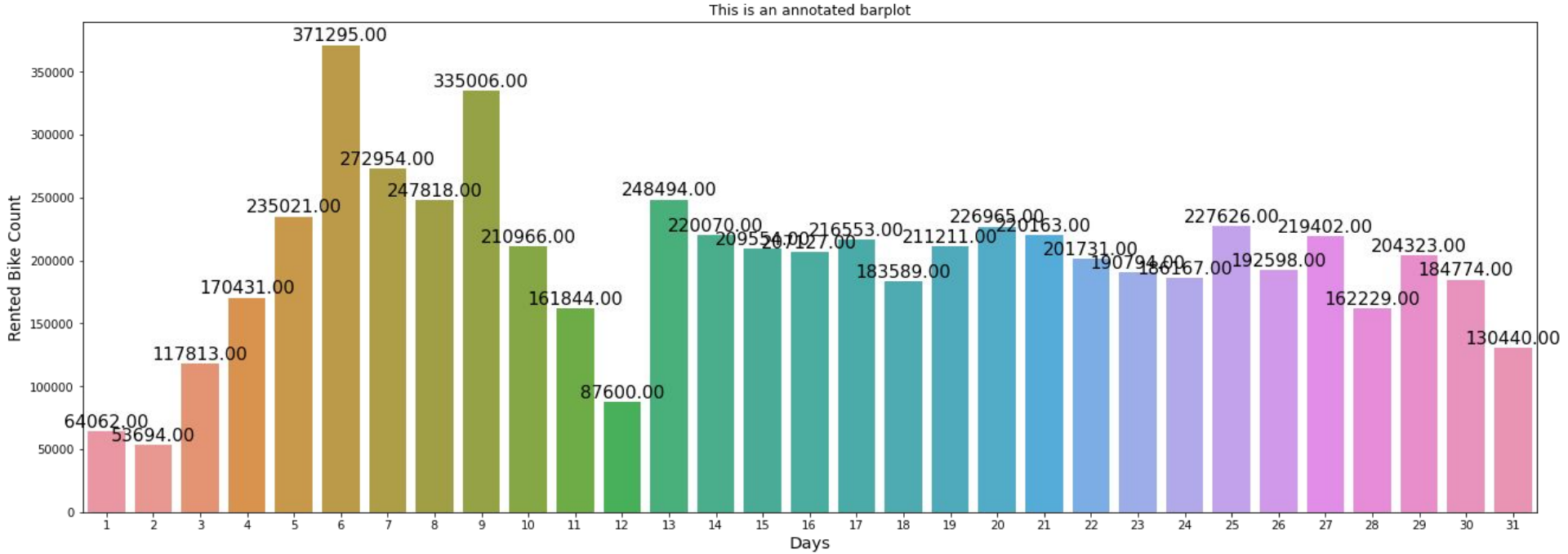
Rented Bike Count Per Year:



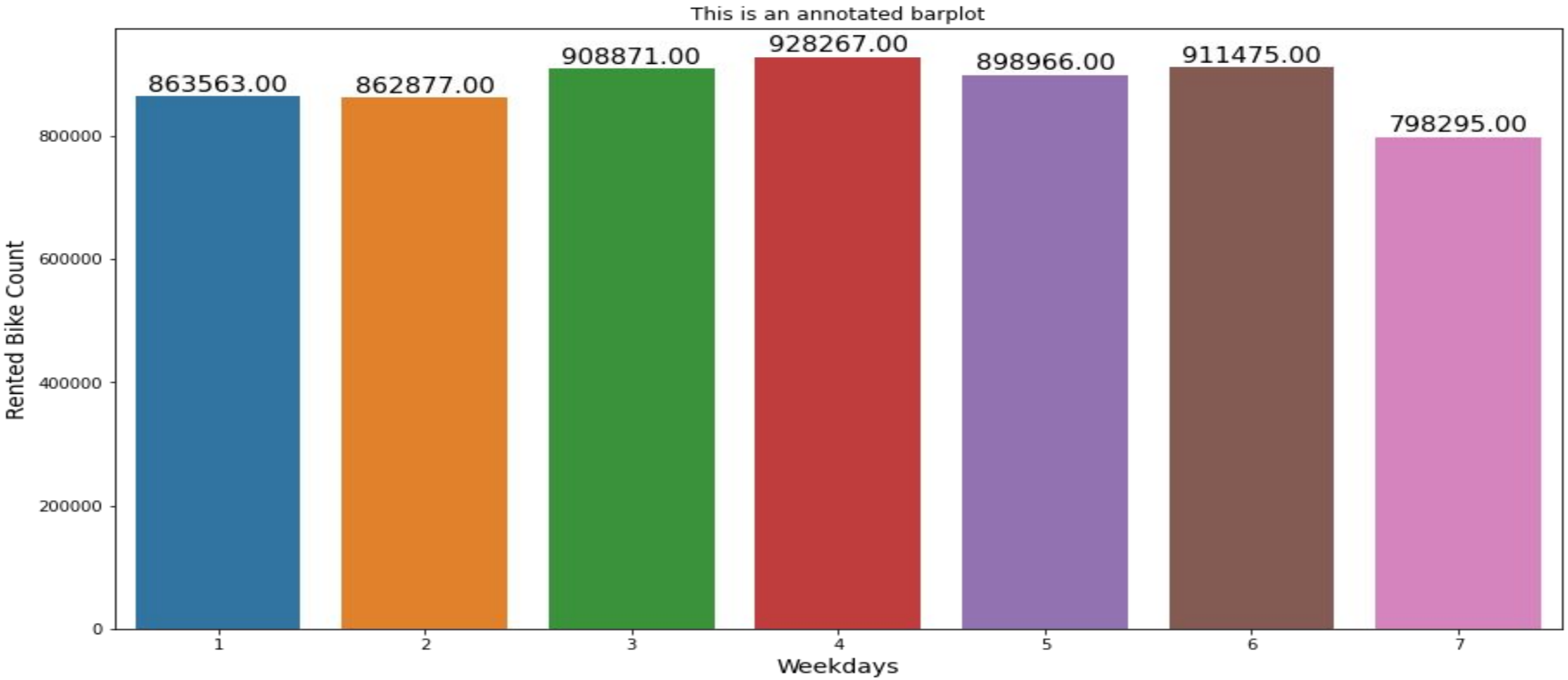
Rented Bike Count Per Month:



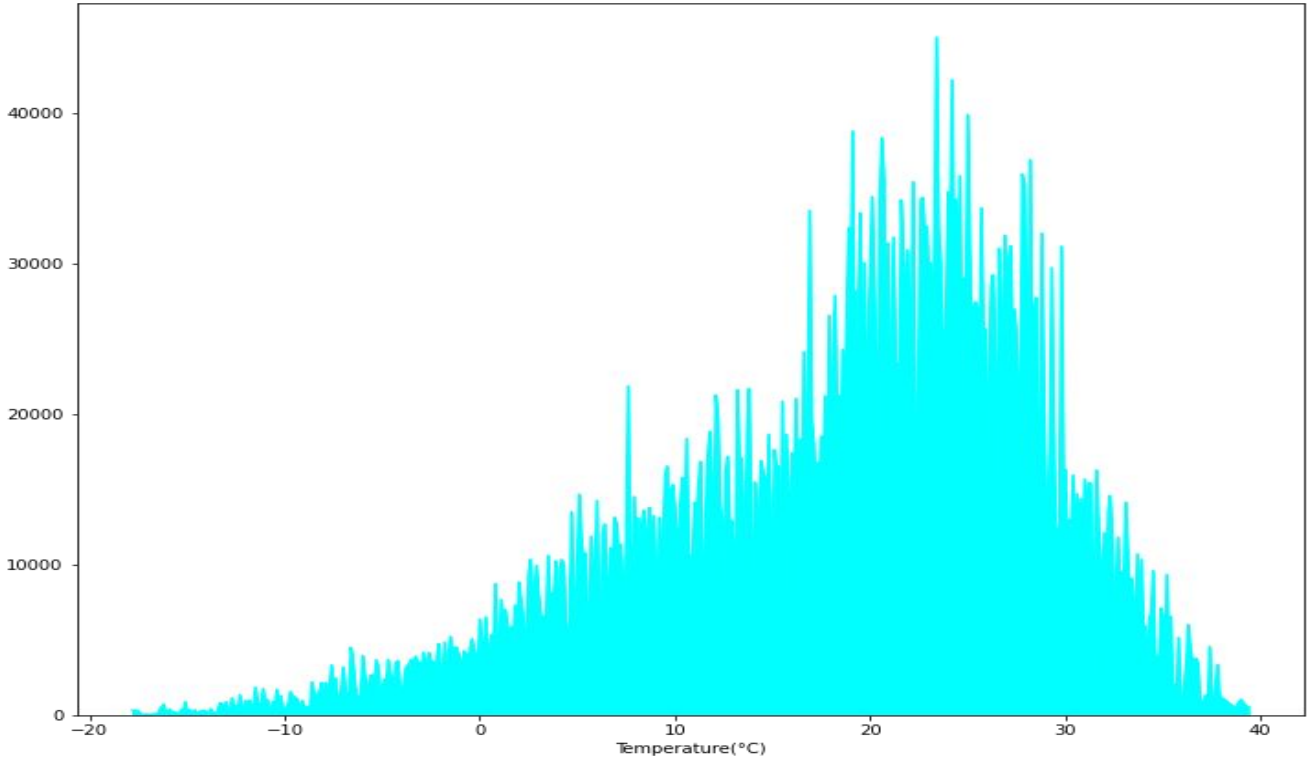
Rented Bike Count Per Day:



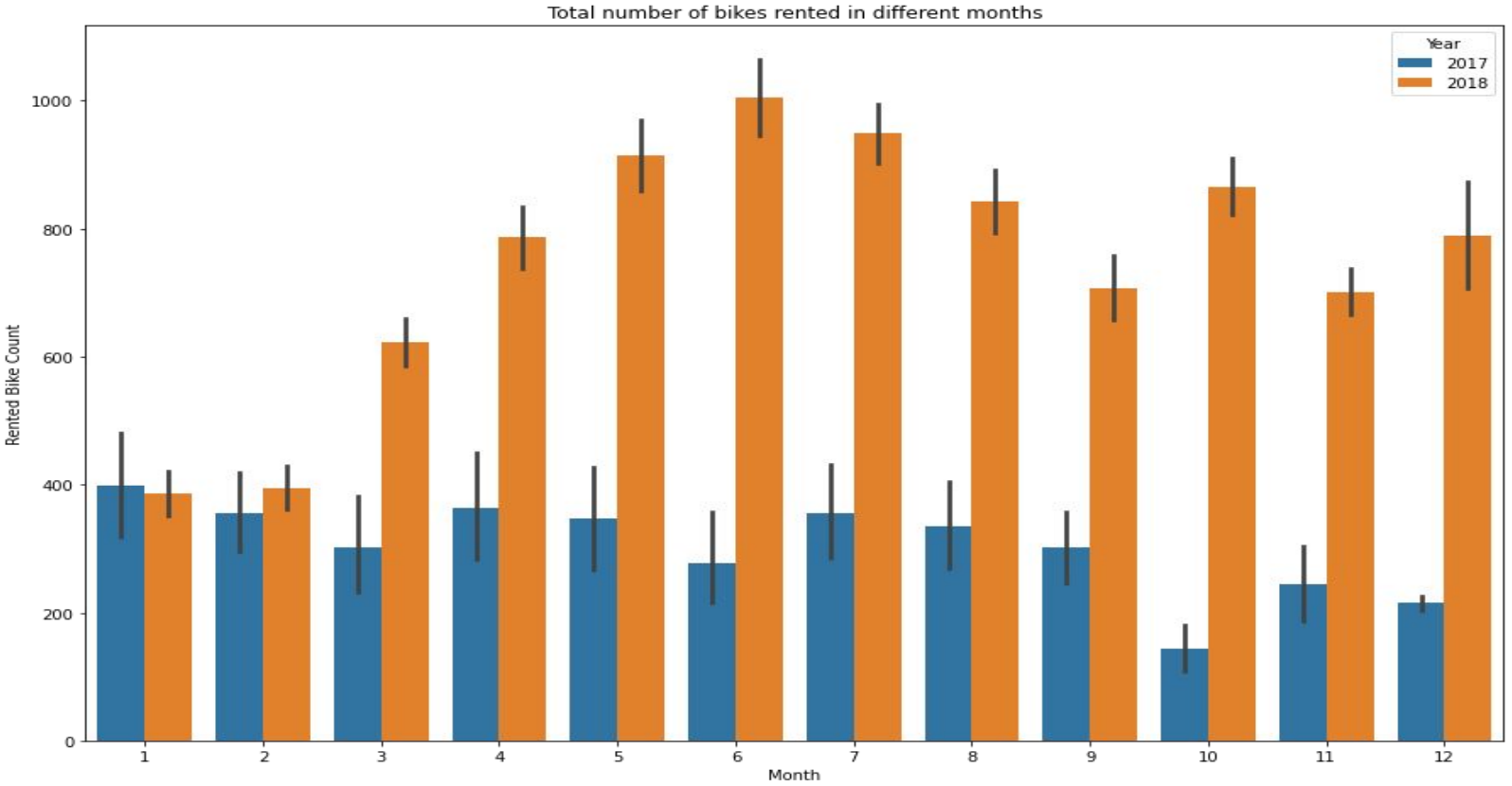
Rented Bike Count In Weekdays:



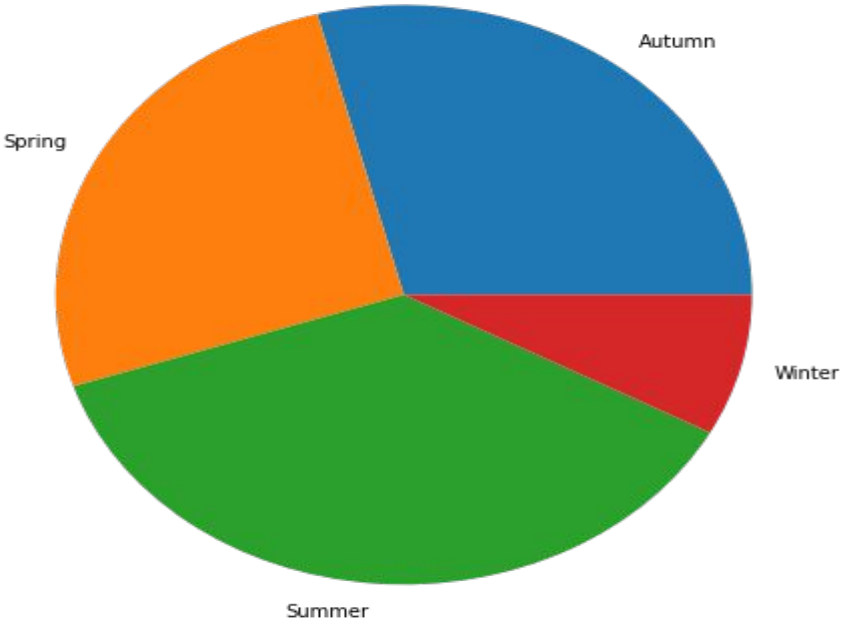
Rented Bike Count With Respect To Temperature:

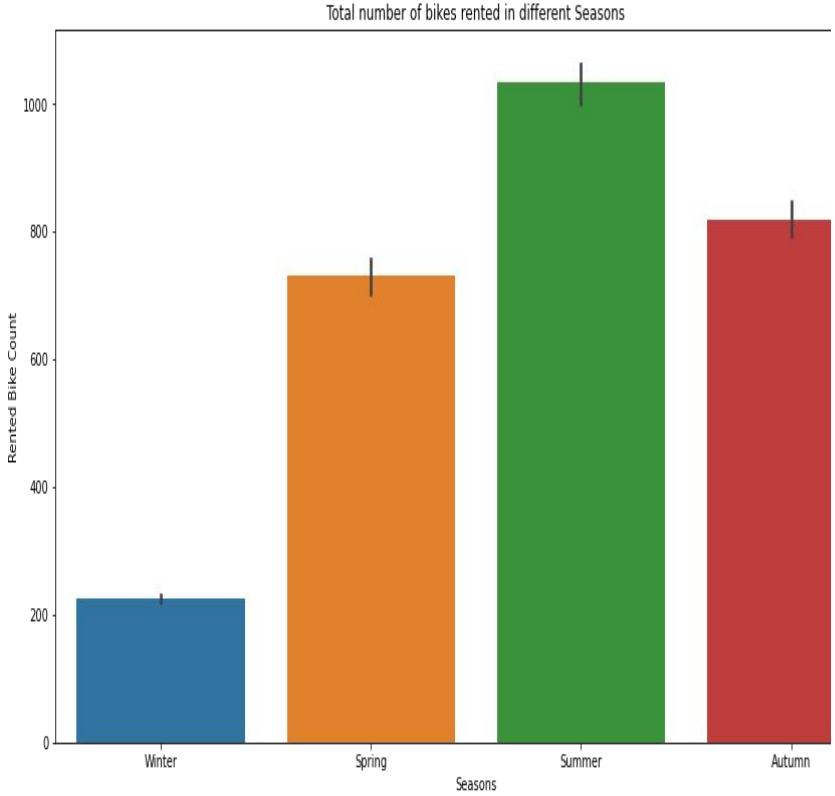


Rented Bike Count In Month:

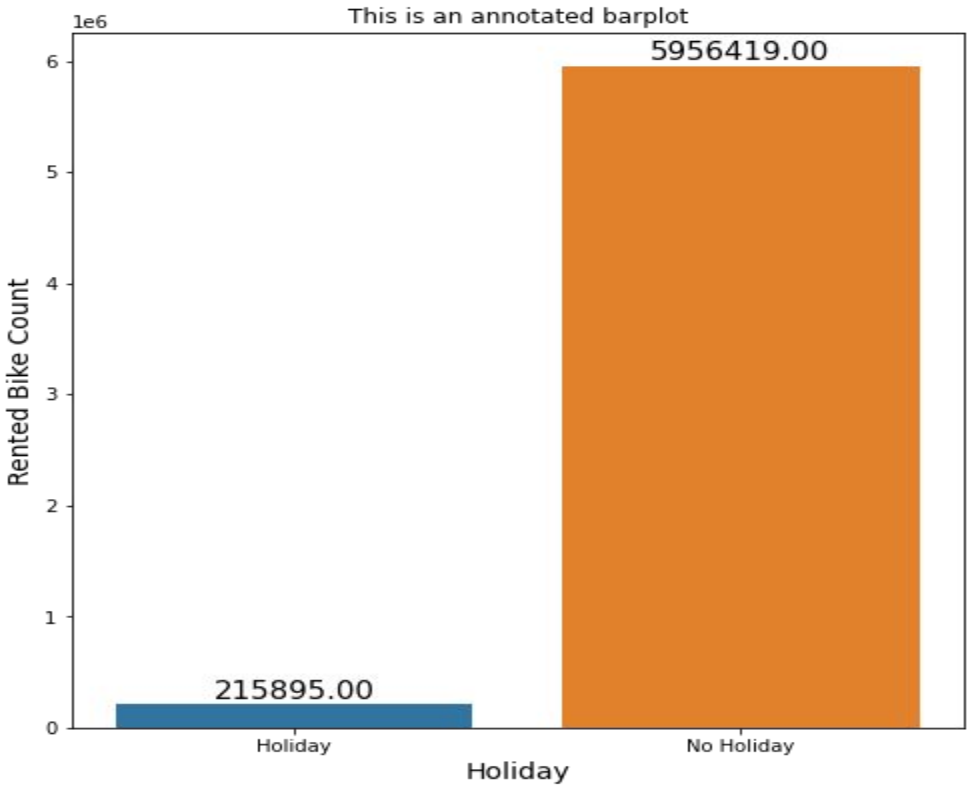


Rented Bike Count In Different Seasons:

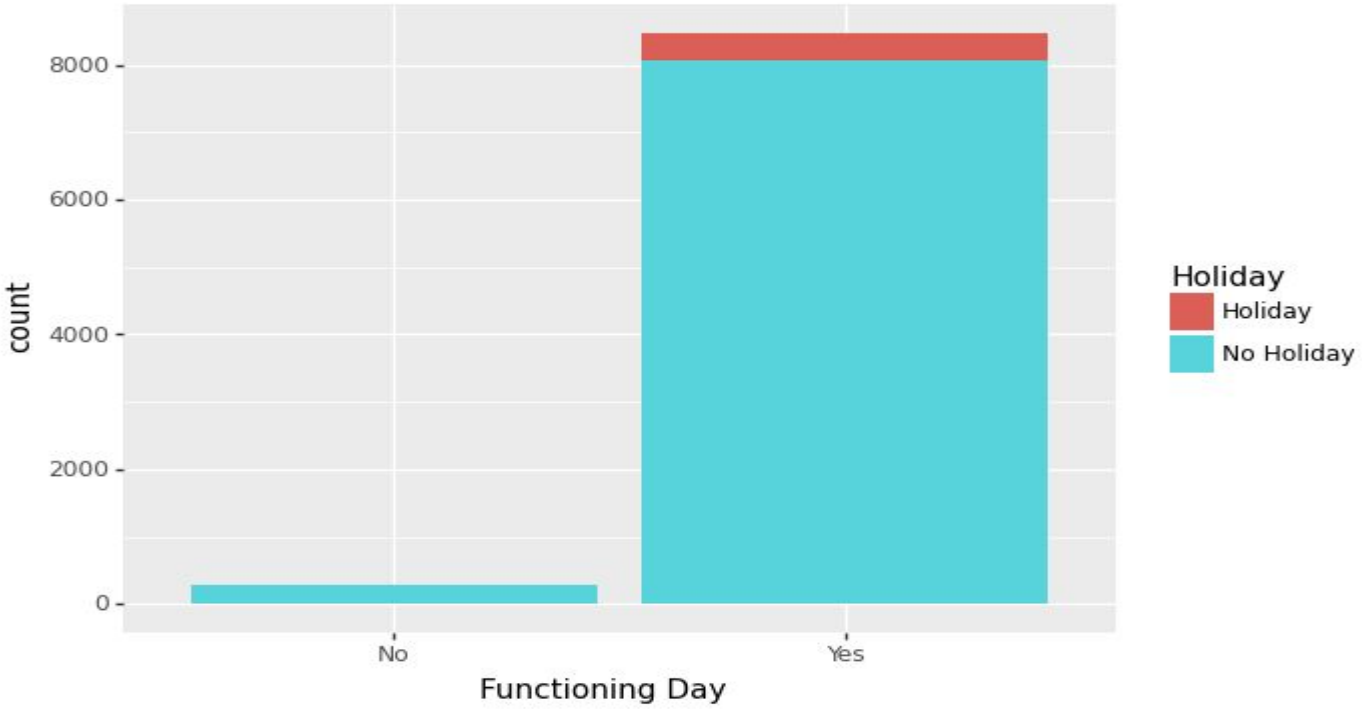




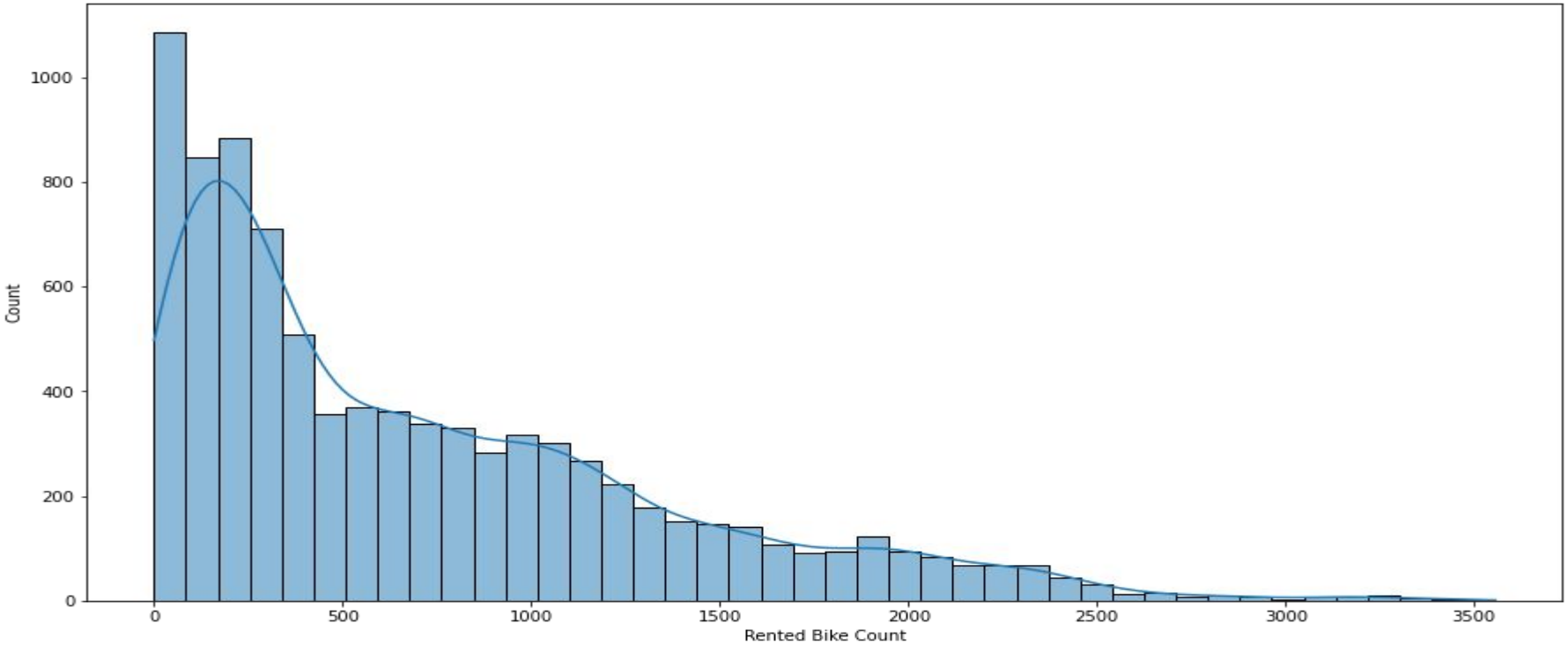
Rented Bike Count On Holiday and On No Holiday:



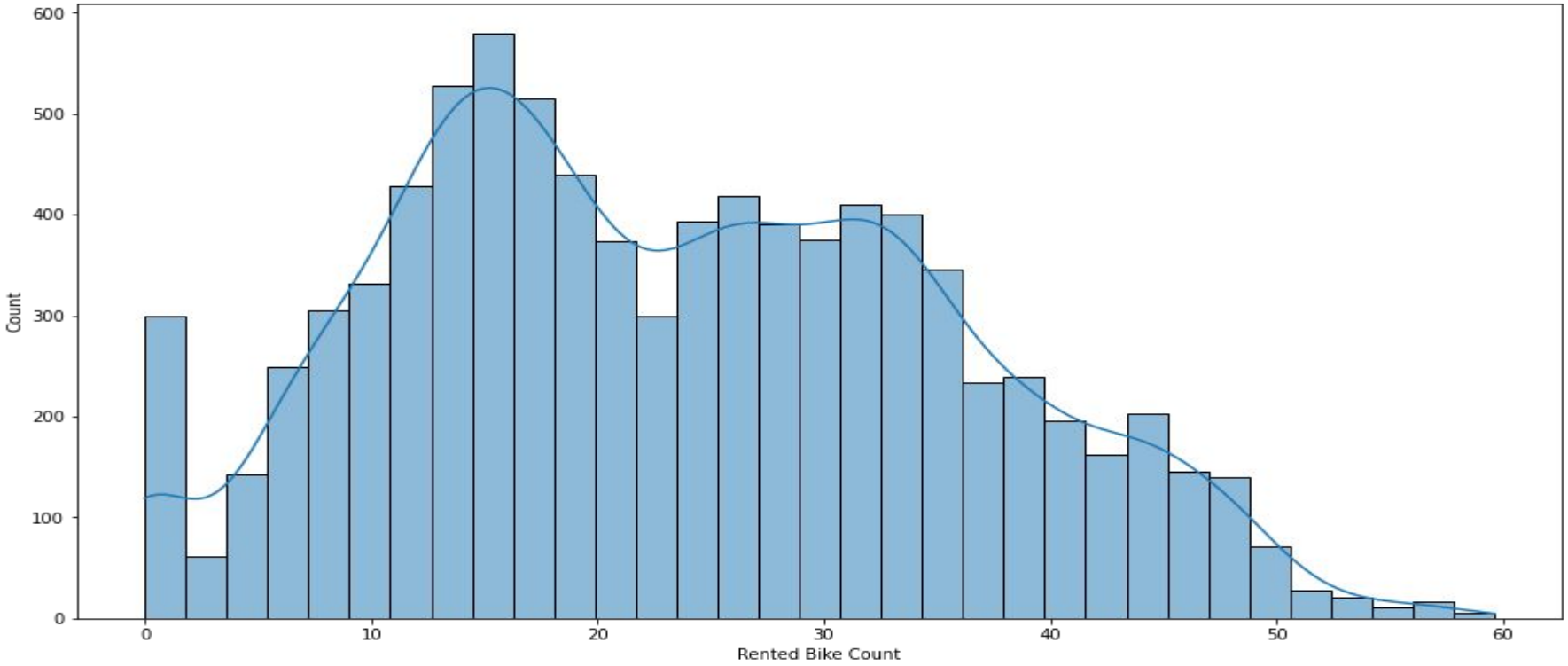
Rented Bike Count On Functioning Day:



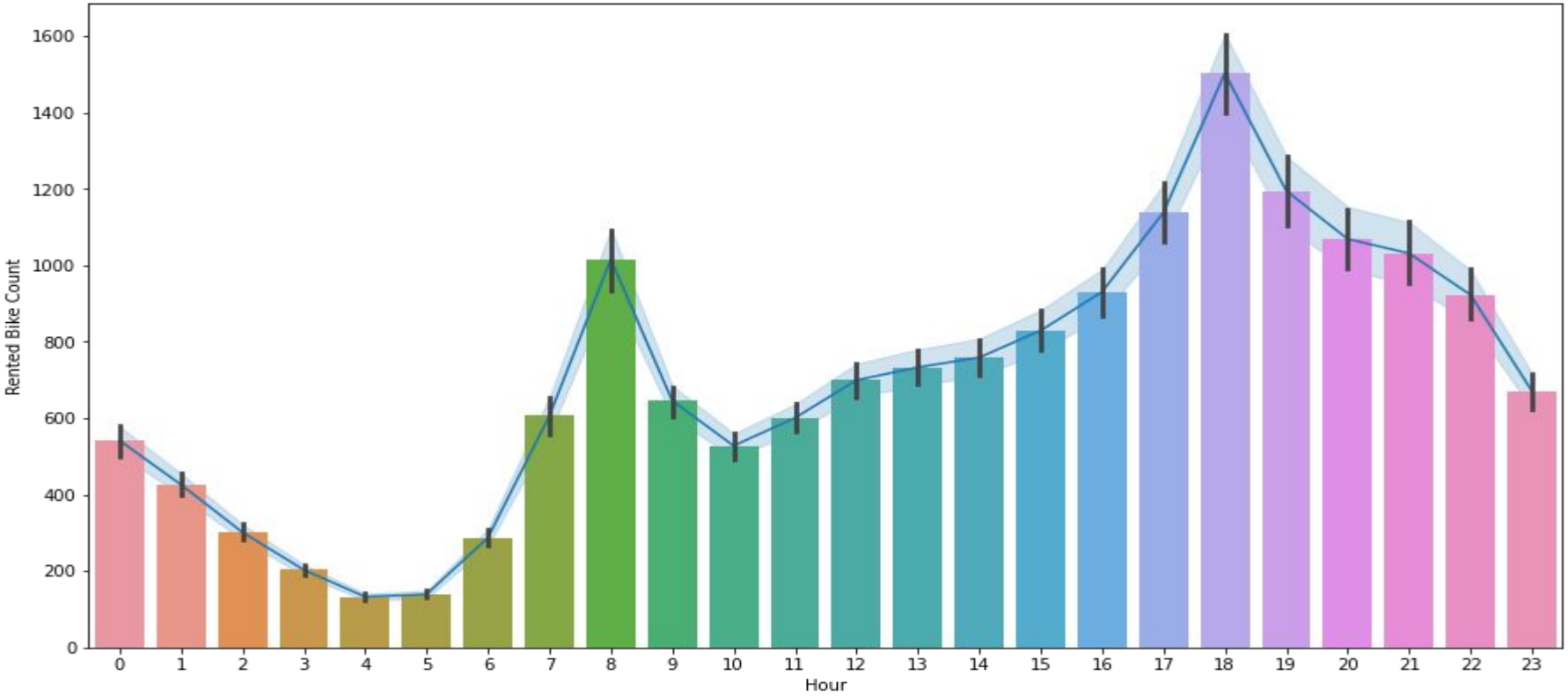
Positively Skewed Data:



After Square Root Transformation:



Rented Bike Count Per Hour:



Data Preprocessing:

Data After Converting to Numerical :

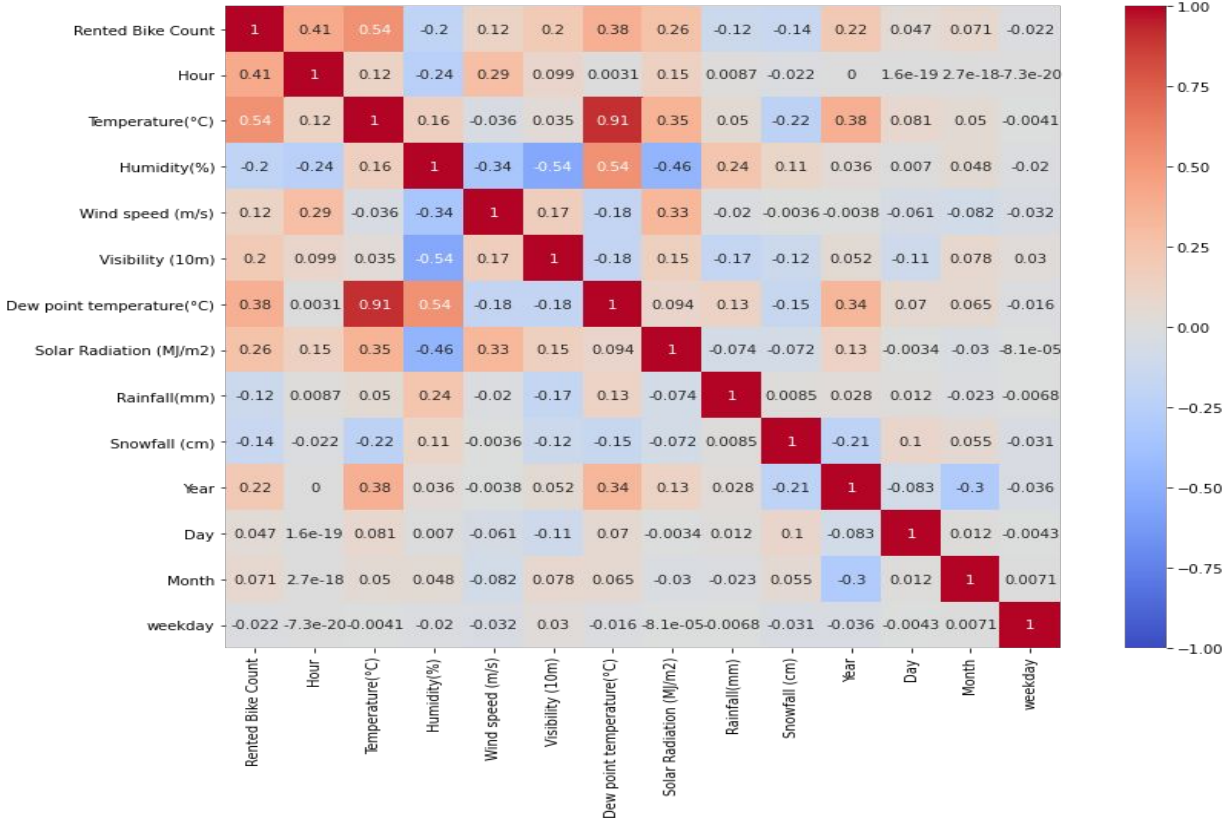
Rows: 8760

Columns: 30

Which columns we have converted ?

a) Seasons
b) Holiday
c) Functioning day
d) Date
e) Month

Correlation:



Models Used:

1) Linear Regression
2) Ridge Regression
3) Lasso Regression
4) Decision Tree Regressor
5) Random Forest
6) Gradient Boost
7) XGBoost

Model Validation and Selection:

Observation 1:

As Linear Regression is not giving us great result, we tried doing regularization we've used Ridge and Lasso Regression but still we did not reached to that extent.

Observation 2:

Then we tried using tree based models, we've used Decision tree regressor and Random Forest still we did not got good scores.

Model Validation and Selection:

Observation 3:

Gradient Boosting and XG Boost came for rescue this models have performed equally good in terms of R-squared and Root Mean Squared Error as this are ensemble models.

Model Validation and Selection:

The test accuracy of Gradient Boosting and XG Boost is almost the same. We can choose any model for the Bike Sharing Demand Prediction. But, XG Boost has some added advantages over Gradient Boosting like regularization.
If we choose Gradient Boosting for our model then the best hyperparameters are:

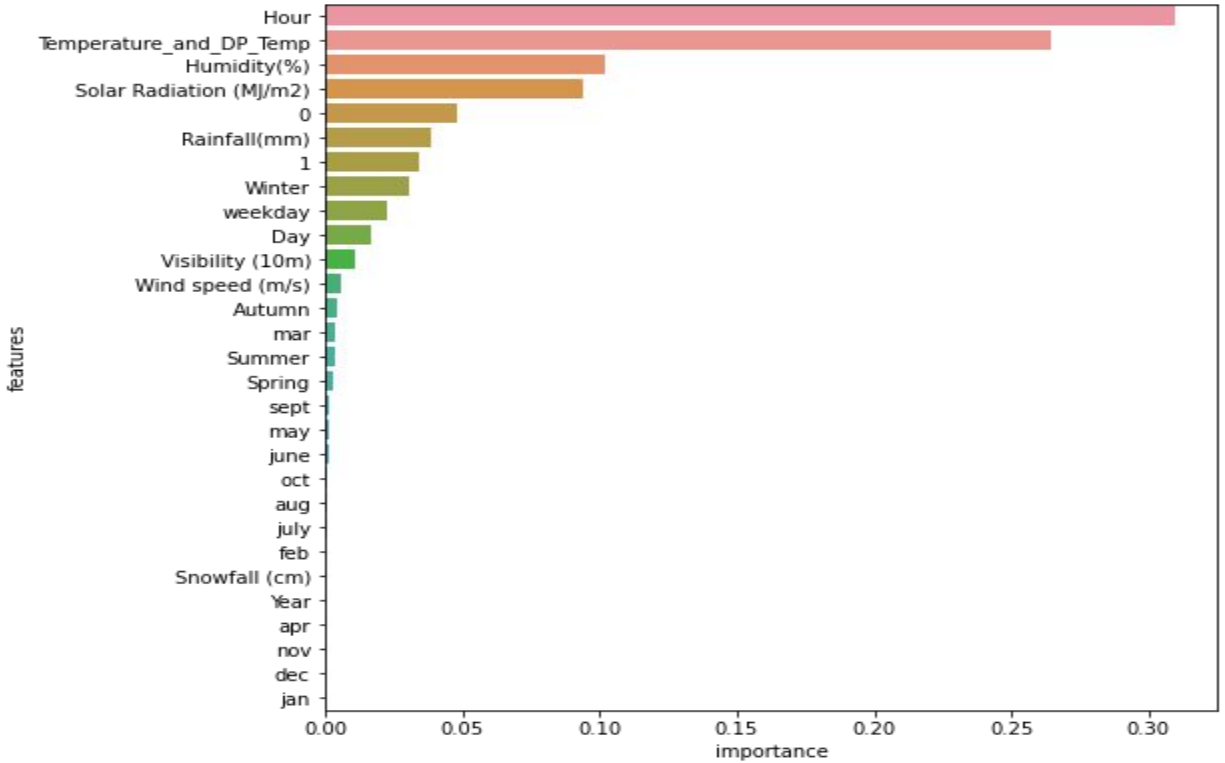
```
'max_depth': 10,
'min_samples_leaf': 40,
'min_samples_split': 50,
'n_estimators': 200
```

Model Validation and Selection:

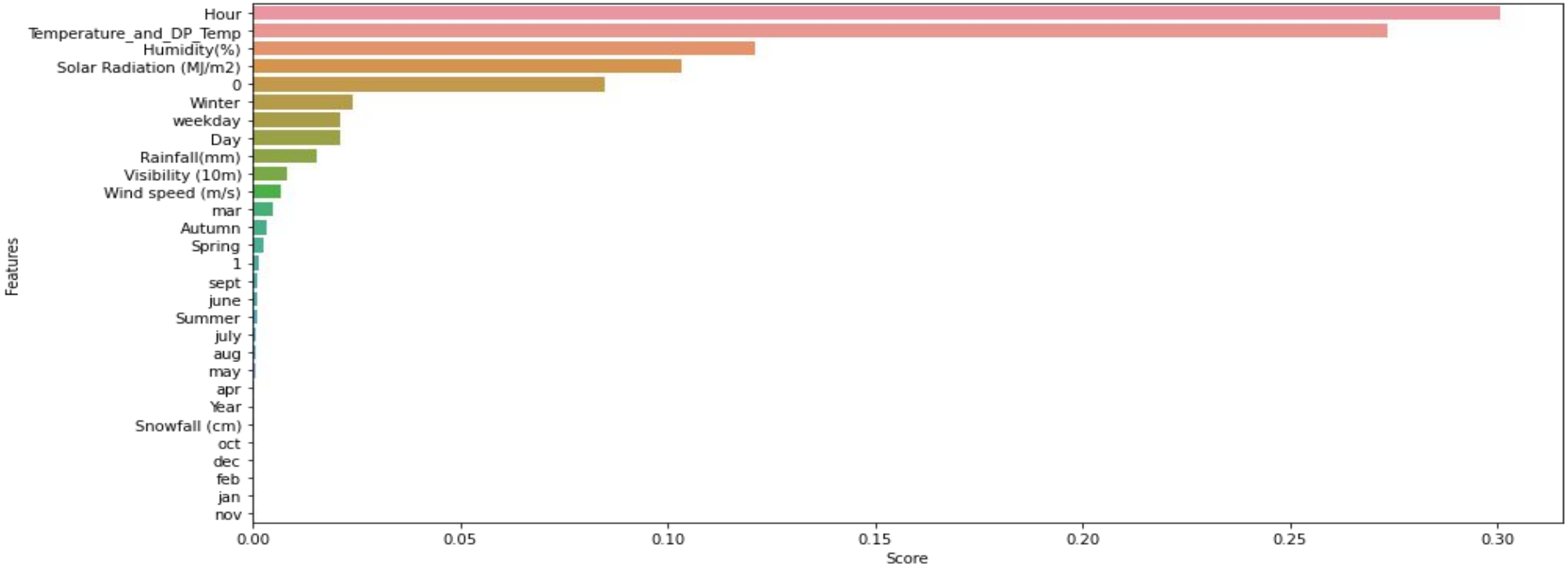
If we choose the XGBoost instead of Gradient Boosting, then best hyperparameters for the algorithm are:

```
'colsample_bytree': 0.7
'learning_rate': 0.05
'max_depth': 7
'min_child_weight': 4
'n_estimators': 500
'nthread': 4 'objective': 'reg:linear'
'silent': 1
'subsample': 0.7
```

Feature Importance for Gradient Boost:



Feature Importance for XGBoost:



Evaluation Metrics:

	Model_Name	MSE	R-squared
0	Linear Regression	185566.230050	0.556616
1	Ridge Regresion	186847.345537	0.556616
2	Lasso Regression	185567.406018	0.556616
3	Decision Tree	64046.284952	0.826155
4	Random Forest	142709.494512	0.659016
5	Gradient Boosting	31071.860688	0.925762
6	XGBoost	30076.949464	0.928135

Challenges:

~ Execution Takes Time

~ As there was no Linearity between independent and Dependent variable our linear regression model did not performed well.

~ There was Multicollinearity available between variables.

Conclusion:

- As it was stated in the problem statement, the business just started out in 2017. So the number of bikes rented in 2017 were too small.
- We can see in year 2018 the rented bike count was 5986984 which is greater than 2017.
- We can see in 6th month or in june the rented bike count is 706728 which is highest and in 2nd month or in feb the count was lowest which is 264112.
- We can see the rented bike count is highest on 6th day of the month which is 371295 and lowest on 2nd day of the month which is 53694.
- We can see on 4th day of week the rented bike count is 928267 which is highest.
- There's is a whooping increase in number of bike rents in year 2018. In the last month the demand decreases in 2018 but increases in it seen to be increasing in the end of 2017. It is like this because, in 2017 the demand is taking off and we can see the pattern as it is still inceasing in the beginning months of 2018. There is a decline in the end of the year. This could be repercussions of winter season as well.
- With pie and bar plot we can say in summer the rented bike count was high as compared to other seasons and lowest in winter season.This is because when temperature decreases amount of snowfall increases due to which people avoid getting out that is the reason in summer rented bike count increases.
- An ironic insight, all the holidays are falling on the functioning Days.

Conclusion:

- We can say on no holiday the rented bike count is much more high than on holiday.
- With the graph we can say on 18th hour of the day there is a huge spike in the count of rented bike which is approx. 1600
- People prefer to take bike ride more often when the temperature is near about 25 degrees Celcius. we can easliy conclude that the people gave more preference to bike riding in summers as compared to other seasons.
- The rise in demand started from the end of 2017 that too in the winter season of the year. The observer may find it weird because demand decreased in the end of 2018. Actually for this situation it can be said that, as the business grew to april 2018 it had increased exponentially as compared to 2017. So, we can say that in winter 2017 demand increased but it wasn't still upto the mark of it's full potential. With simple heuristics for future as well if everything else in independent variables remains constant we can say that, the demand will decrease in december but with the proportionate to the overall demand of that year.
- The number of business hours of the day and the demand for rented bikes were most correlated. It's common sense too.
- Highest number of bike rented at the 18th hour of day.
- After trying combinations of features with linear regression the model underfitted. It seemed obvious because data is spread too much. It didn't seem practical to fit a line.
- With this pair plot we can seee there is no relation between independent variable and dependent variables, so our linear regression model will not work well on this data.

Conclusion:

- With ridge the train score for alpha=0.01 came to be 0.56 and the test score for alpha=0.01 came to be 0.55. train score for alpha=100 came to be 0.56 and the test score for alpha=100 came to be 0.55. for both alpha 0.01 and 100 the train and test value came to be 0.56 and 0.55 respectively.
- With lasso training score came to be 0.56 and test score came to be 0.55. number of features used is 26. training score for alpha=0.01 came to be 0.56 and test score for alpha =0.01 came to be 0.55. number of features used: for alpha =0.01 is 28. training score for alpha=0.0001 came out to be 0.56 and test score for alpha =0.0001 came to be 0.55. number of features used: for alpha =0.0001 is 28.
- With Decision tree we reached at the model r squared value of 0.84. We only fitted with minimum number of leaf hyperparameter. With default parameters it overfitted and reached r-squared at 1 with train dataset but 0.83 with test.
- With random forest our r2 score came out to be 0.67 on training set and 0.64 on test set.
- Gredient boost came for the rescue to help us get best accuracy to approximate numbers of rented bikes demand. By increasing the number of trees we could overfit it to 1 r squared accuracy. But it was plausible results of training r-square at 0.97 and test r-square value at 92 also with adjusted r-square with 0.92.
- The Feature_importance was almost the same in both the tree based models. Gredient boost fine-tunes with error of the prior trees this is why it gets better accuracies.
- HOUR and TEMPERATURE_AND_DP_TEMPERATURE column are the main columns helping in prediction.
- With XGBoost Regressor we can see the r2 score on training data came to be 0.99 and on test data it is 0.92 which is quit good.


THANK
YOU!