Capstone Project

Bike Sharing Demand Prediction

By- Navneet Keshri

Points for Discussion

- Introduction
- Problem Statement
- Know Your Data
- Data Wrangling
- Data Visualization
- Feature Engineering
- ML Model Implementation
- Conclusion

Introduction

1. A growing number of metropolitan areas are now offering bike rentals as a means of enhancing mobility convenience. The public must have access to the rental bike at the appropriate time so that it reduces the amount of time people have to wait.
2. Providing the city with a consistent supply of rental bikes becomes a major concern at some point. The most important part is the expected hourly bikes count for the constant supply of rental bikes.
3. The membership, rental, and bike return procedures in a city are all automated by a network of locations in bike-sharing systems.
4. In this dataset, we predict the demand for the Bike Sharing Program in Seoul based on historical data.

Problem Statements

Predicting how many bikes will be needed at any given time and day

Reduce bikes wastes resources (both in terms of bike maintenance and the land required for parking)

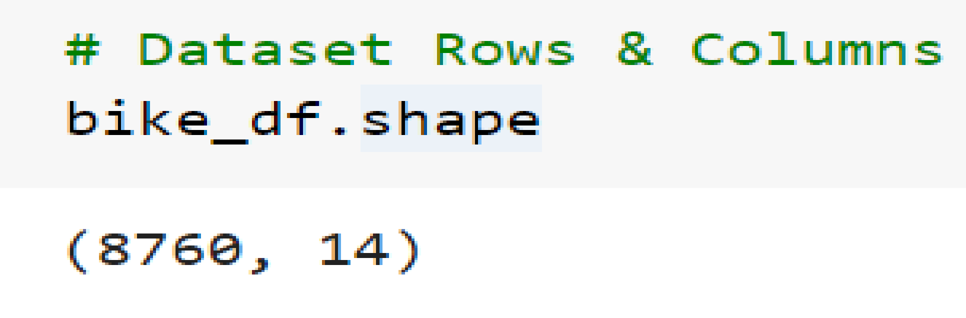
Reduce short-term loss due to the loss of immediate customers and long-term loss of future customer base.

Manage demand and supply equilibrium throughout the day.

Run the businesses effectively by estimating the demands

Know Your Data

The Seoul bike dataset has 14 columns and 8760 rows.

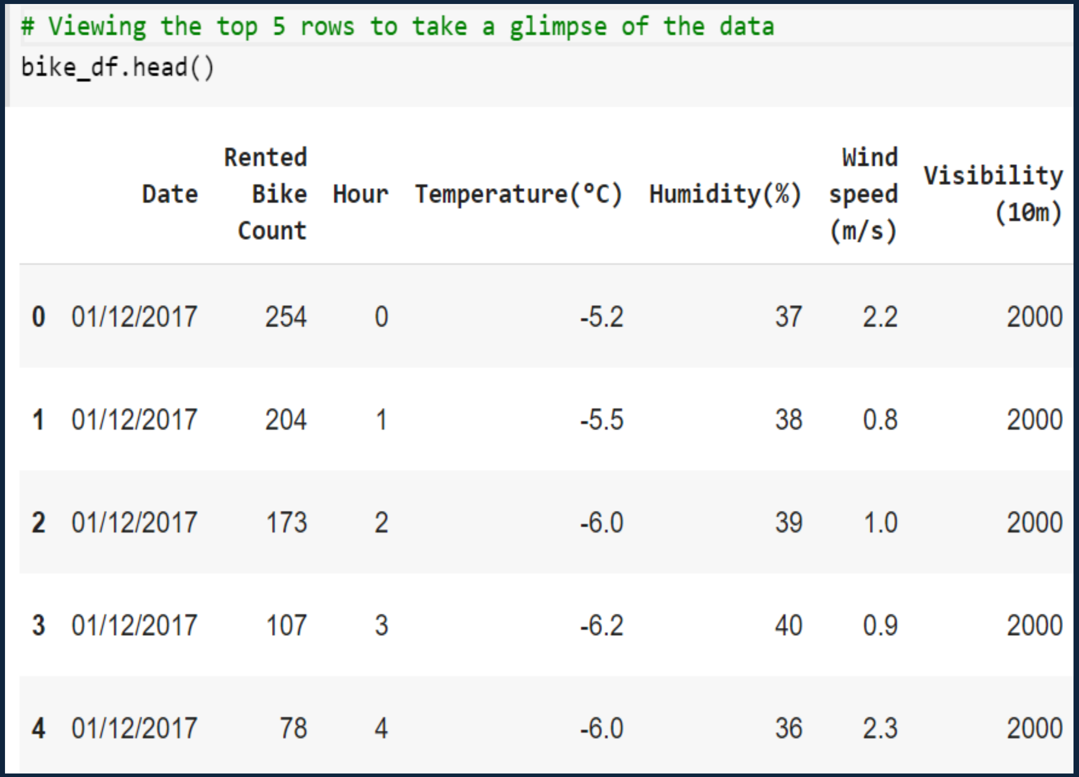


```
# Dataset Rows & Columns
bike_df.shape
```

```
(8760, 14)
```

Understanding Your Variable

1. Date: year-month-day of Bike sharing.
2. Rented Bike count: Count of bikes rented at each hour.
3. Hour: Hour of the day.
4. Temperature: Temperature in Celsius.
5. Humidity: Humidity in %
6. Windspeed: Windspeed in m/s.
7. Visibility: Visibility in 10m.

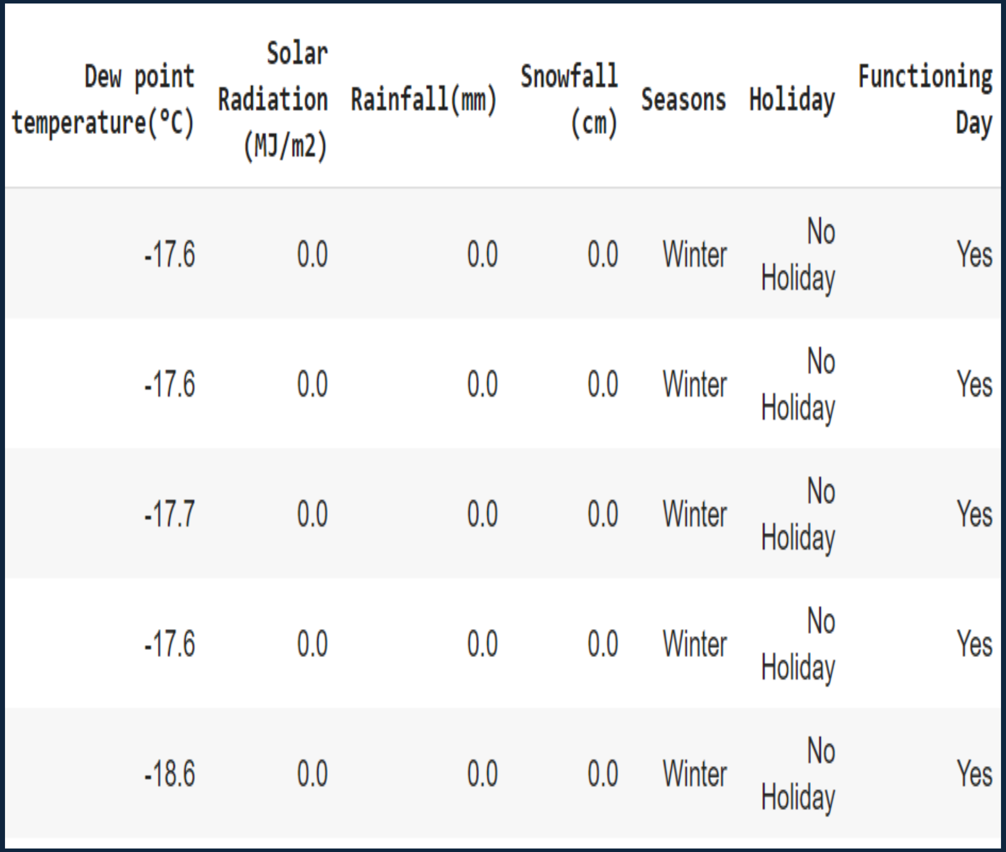


```
# Viewing the top 5 rows to take a glimpse of the data
bike_df.head()
```

	Date	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)
0	01/12/2017	254	0	-5.2	37	2.2	2000
1	01/12/2017	204	1	-5.5	38	0.8	2000
2	01/12/2017	173	2	-6.0	39	1.0	2000
3	01/12/2017	107	3	-6.2	40	0.9	2000
4	01/12/2017	78	4	-6.0	36	2.3	2000

Understanding Your Variable

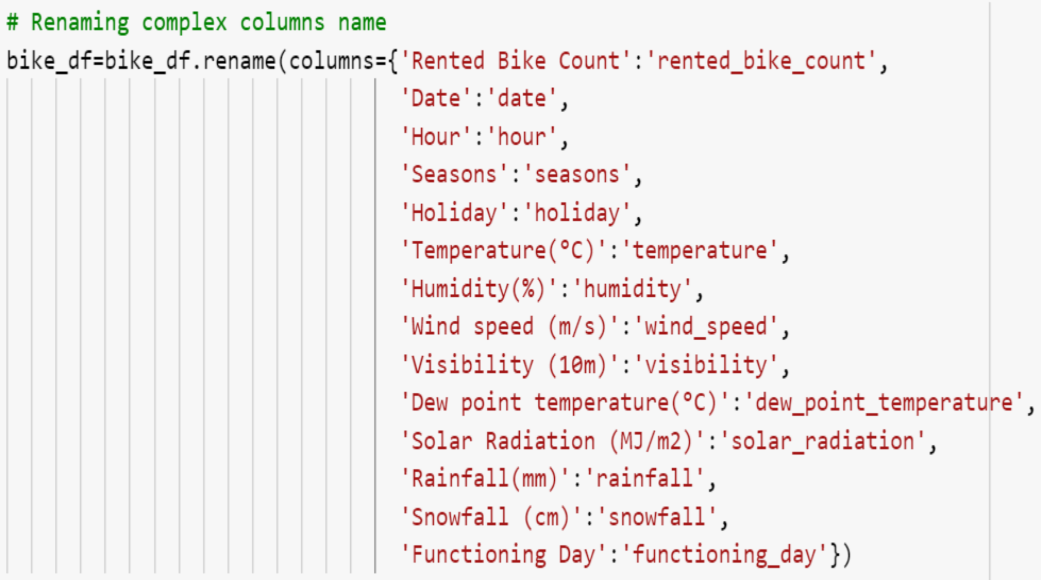
8. Dew point temperature: Dew point temperature in Celsius.
9. Solar radiation: Solar radiation in MJ/m.
10. Rainfall: Rainfall in mm.
11. Snowfall: Snowfall in cm.
12. Seasons: Winter, Spring, Summer, Autumn
13. Holiday: Holiday/No holiday.
14. Functional Day: No (Non-Functional Hours), Yes (Functional hours).



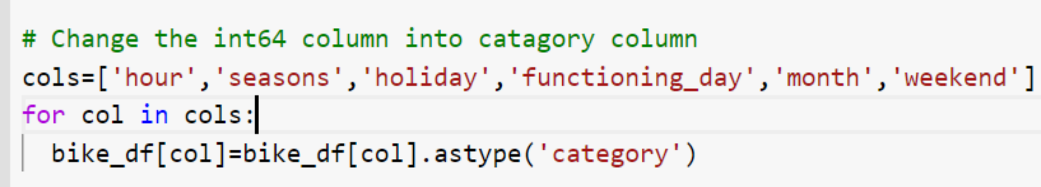
Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
-17.7	0.0	0.0	0.0	Winter	No Holiday	Yes
-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
-18.6	0.0	0.0	0.0	Winter	No Holiday	Yes

Data Wrangling

1. I did some basic data manipulations and renamed complex column names.
2. I checked for missing and duplicate values and fortunately there are no missing and duplicate values are there in our dataset. (Clean data)
3. I extracted and created 3 columns from the date which are year, month, and day.
4. From the day column, I created the weekend column (Saturday/Sunday =1, Other days = 0) to better understand the weekend demand of bike sharing count.
5. Although some columns appear to be of the integer type, the "Hour," "Month," and "Weekend" columns are actually of the category type. As a result, if we do not alter this data structure, we run the risk of being deceived by the values during subsequent analyses.



```
# Renaming complex columns name
bike_df=bike_df.rename(columns={'Rented Bike Count':'rented_bike_count',
                                'Date':'date',
                                'Hour':'hour',
                                'Seasons':'seasons',
                                'Holiday':'holiday',
                                'Temperature(°C)':'temperature',
                                'Humidity(%)':'humidity',
                                'Wind speed (m/s)':'wind_speed',
                                'Visibility (10m)':'visibility',
                                'Dew point temperature(°C)':'dew_point_temperature',
                                'Solar Radiation (MJ/m2)':'solar_radiation',
                                'Rainfall(mm)':'rainfall',
                                'Snowfall (cm)':'snowfall',
                                'Functioning Day':'functioning_day'})
```



```
# Change the int64 column into catagory column
cols=['hour','seasons','holiday','functioning_day','month','weekend']
for col in cols:
  bike_df[col]=bike_df[col].astype('category')
```

Data Visualization: Understand the relationship between variables

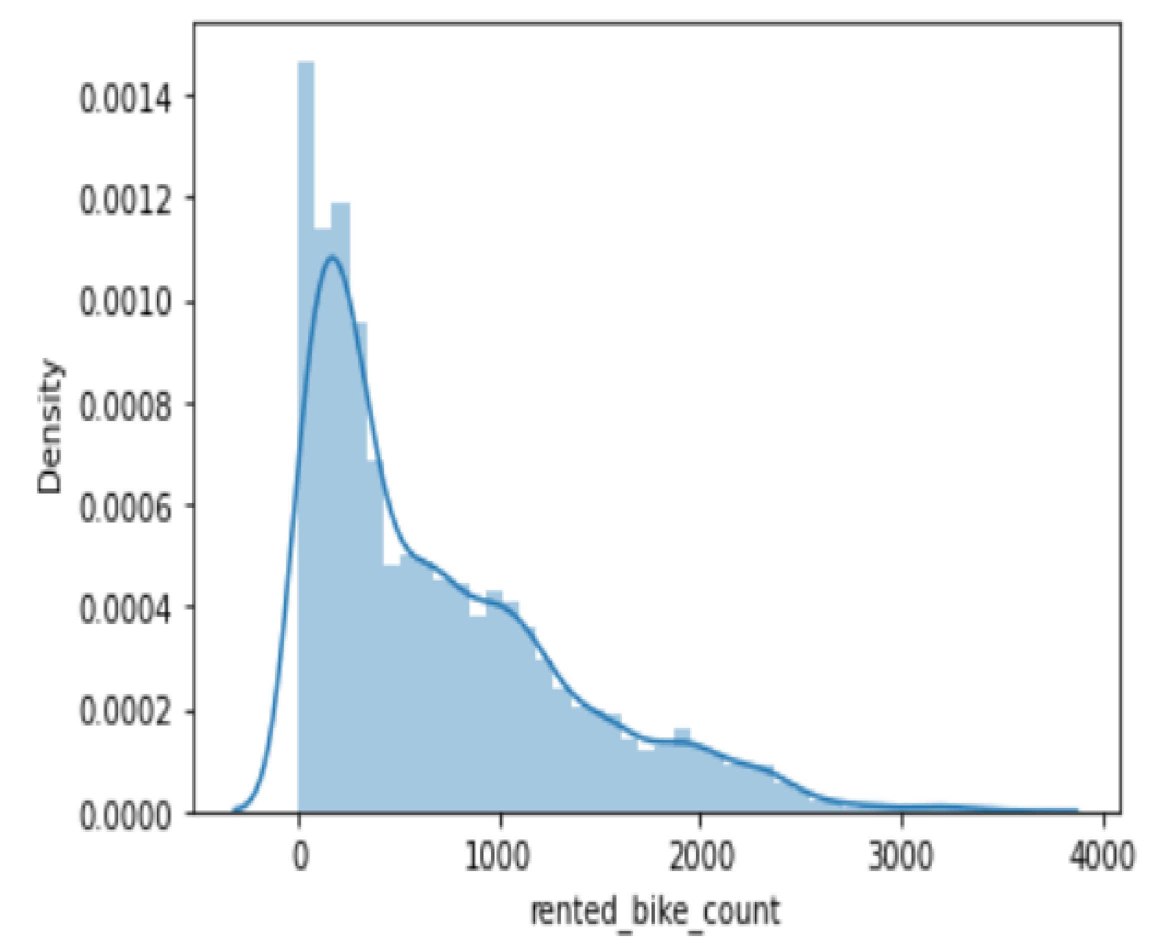


Chart 1 – Distribution Chart on Dependant Variable i.e., rented_bike_count (Univariate Analysis)

- **The distplot shows the data distribution of a variable in comparison to the density distribution**
- **rented_bike_count is positively skewed in the distribution**

Data Visualization: Understand the relationship between variables

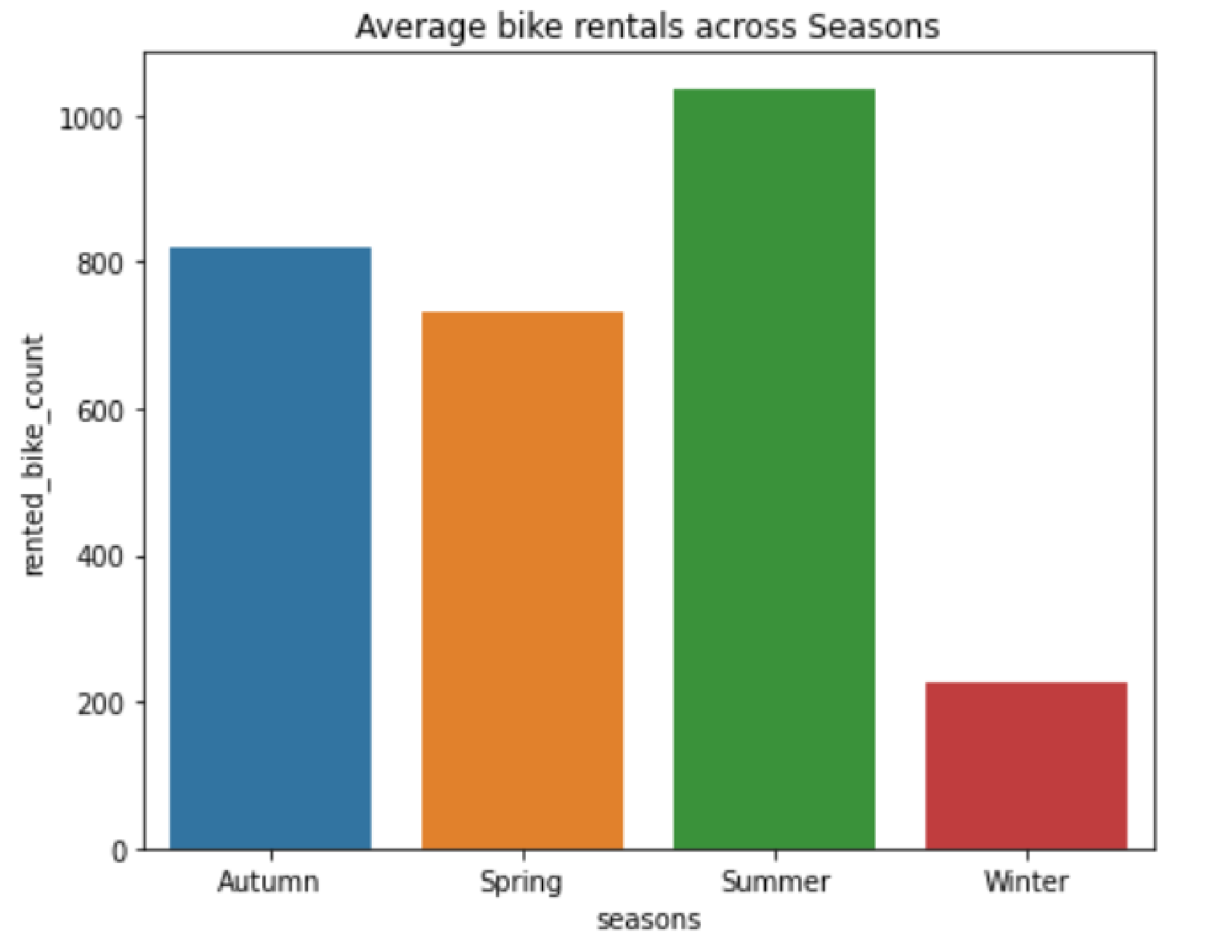


Chart 2 - Rented Bike Count Vs Seasons (Bivariate Analysis)

- The frequency counts of values at various levels of a categorical variable are depicted in bar charts.
- The most preferred season for the rented_bike_count is summer and the least preferred is winter which means that people prefer to rent bikes in warm seasons.

Data Visualization: Understand the relationship between variables

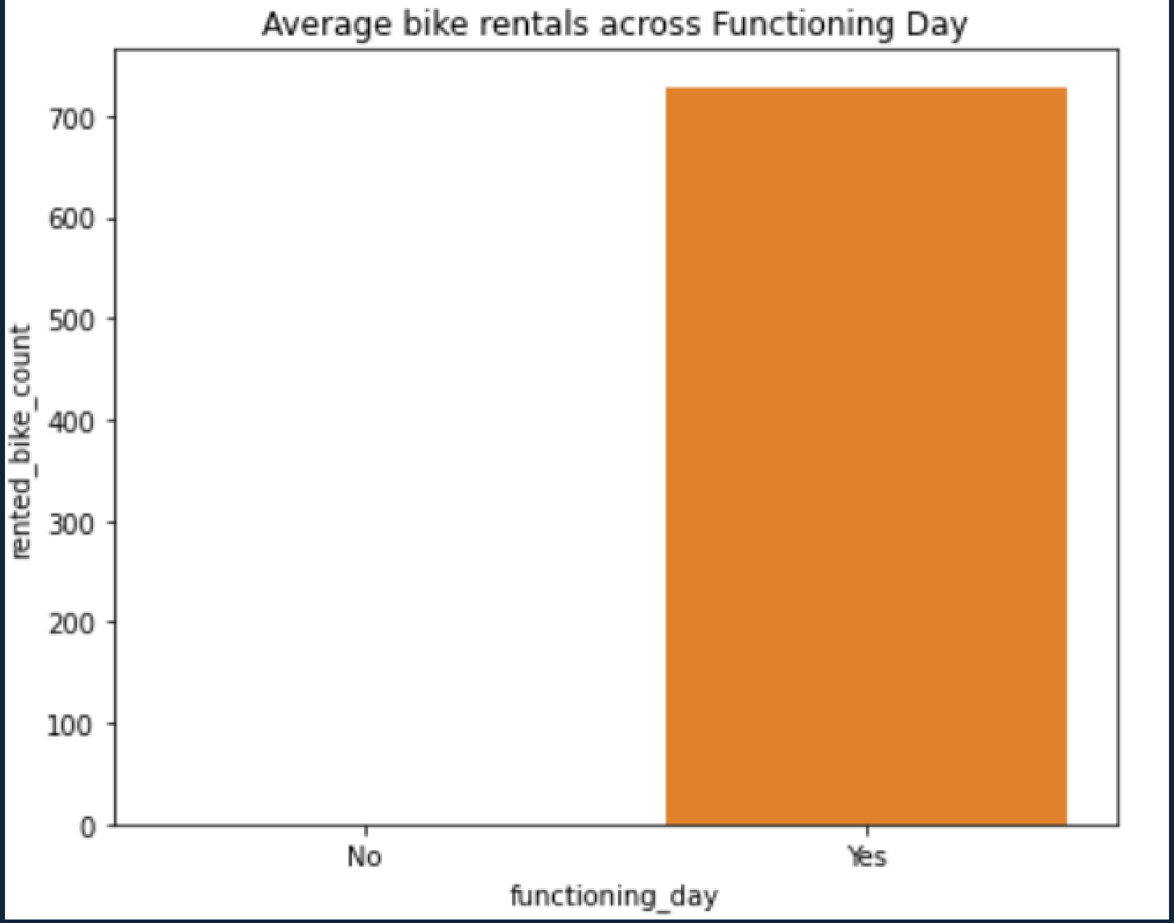


Chart 3 - Rented Bike Count Vs Functioning Day (Bivariate Analysis)

- **During business hours, there is a lot of demand for bike-sharing services, which could be because many customers use these bikes to get to work. When it is not a functioning_day, there is no demand for rented bikes**
- **functioning_day has a high demand for rented bikes and if the day is not a functioning day then there is no demand for bike sharing.**

Data Visualization: Understand the relationship between variables

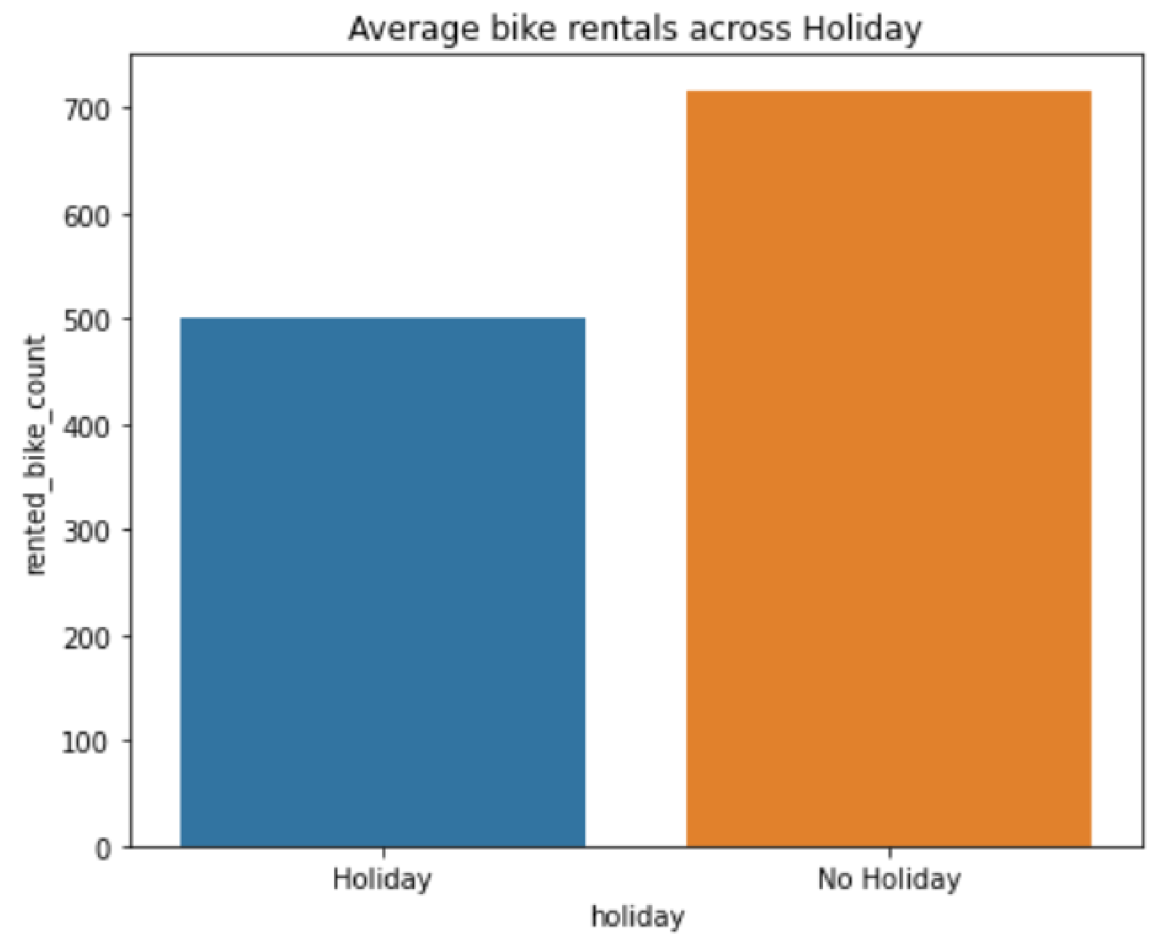


Chart 4 - Rented Bike Count Vs Holiday (Bivariate Analysis)

- When there is no holiday, demand for bike sharing is higher than when there is a holiday,
- This is a clear indication that Bike Sharing is preferred for business-related purposes and people book bikes to get to work more frequently than going for a trip.

Data Visualization: Understand the relationship between variables

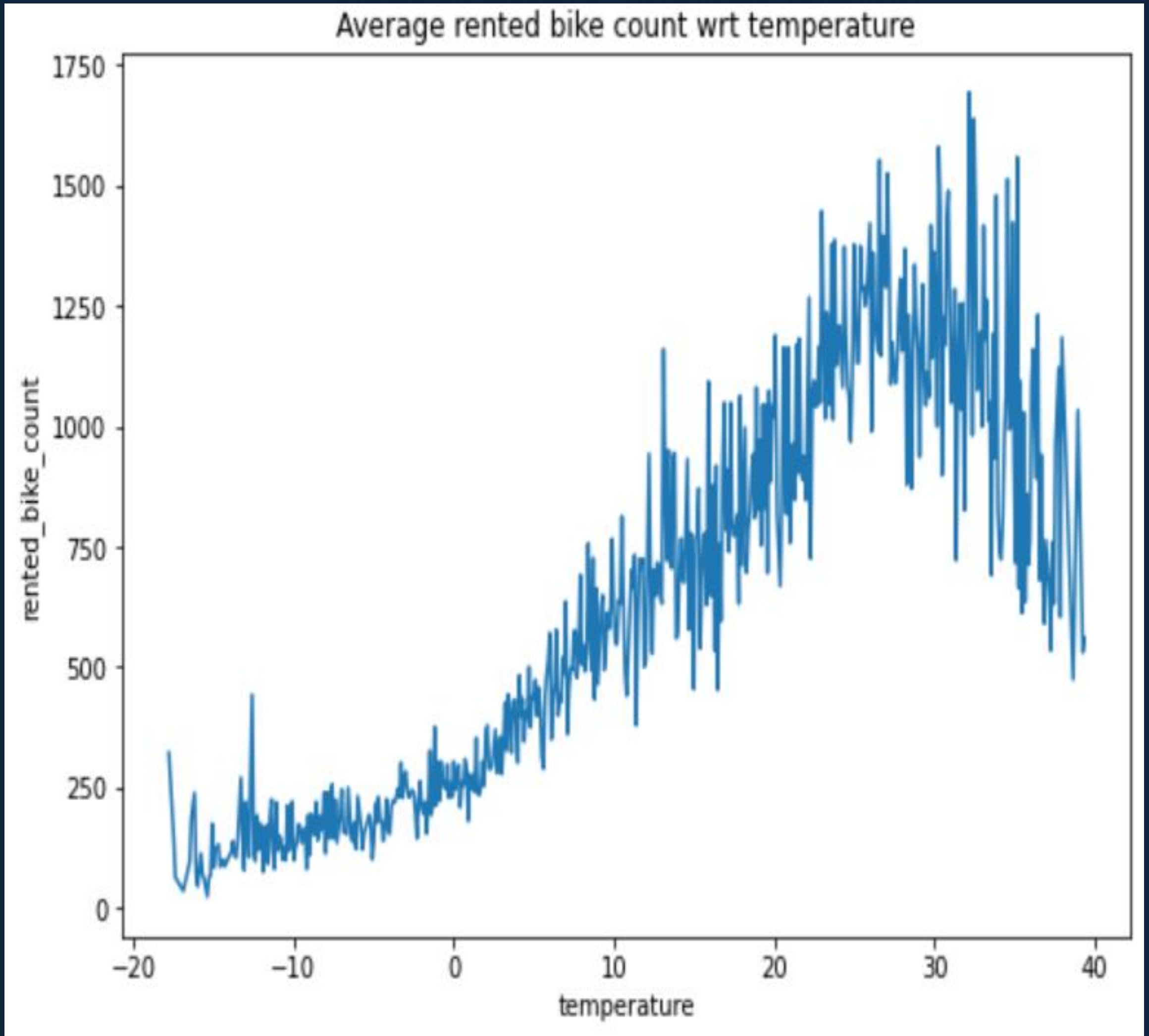


Chart 5 - Rented Bike Count Vs Temperature (Bivariate Analysis)

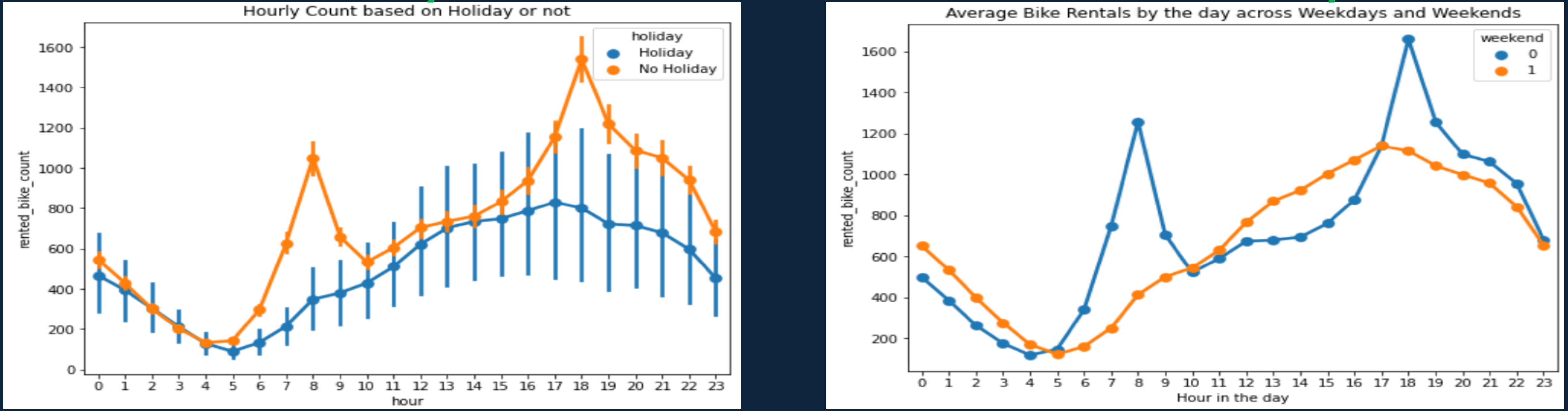
- Since temperature is not a categorical variable where a bar chart can be used, but it is a continuous numerical variable hence A line chart is used to show how temperature has changed over time.

- We can see from the line plot that the average number of bikes rented with temperature increases steadily, with a slight decrease at the highest temperature.

- This confirms our analysis of the season’s column, which revealed that people prefer renting bikes in warm temperatures.

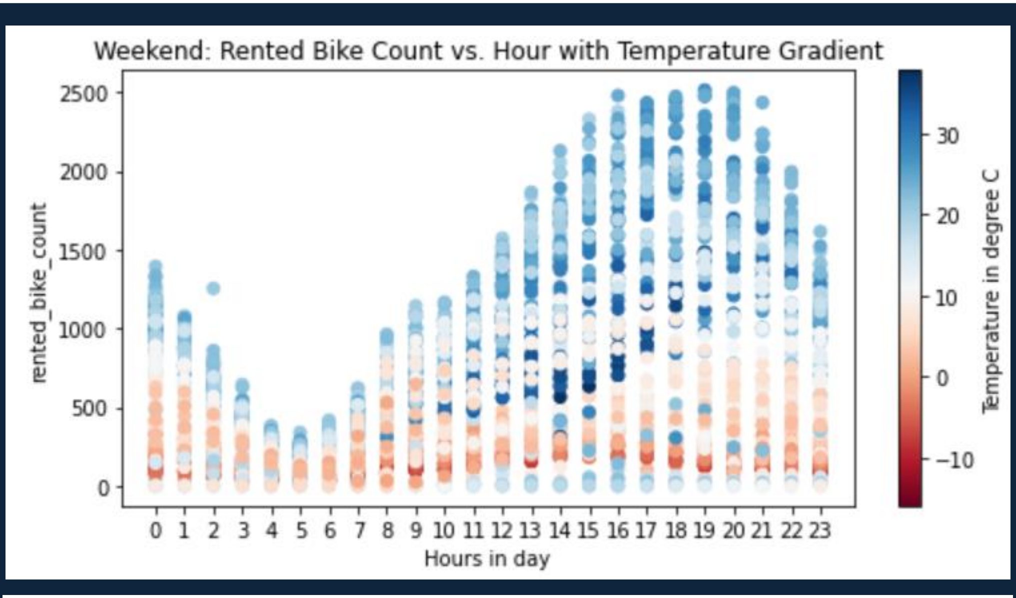
Data Visualization: Understand the relationship between variables

Chart 6 - Hourly Rented Bike Counts Distribution (Multivariate Analysis)



- Point charts are useful for clearly displaying quantitative data.
- Higher reservations can be seen at around 8 am and 5 pm (office hours) and close to 0 reservations very early in the morning.
- Working Day: The first pattern is where there is a peak in the rentals at around 8 am and another at around 6 pm. These correspond to working local bikers who typically go to work on a working day.
- Non-Working Day: Second pattern where there are more or less uniform rentals across the day with a peak at around noon time. These correspond to probably tourists who typically are casual users who rent/drop off bikes uniformly during the day and tour the city.

Data Visualization: Understand the relationship between variables



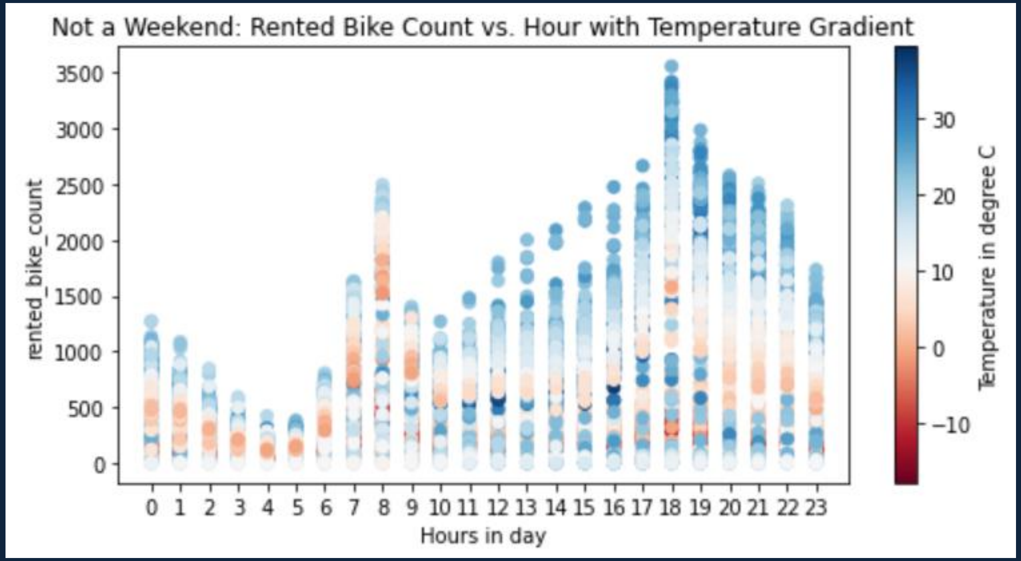


Chart 7 - Weekend: Rented bike count Vs Hour with Temperature (Multivariate Analysis)

- The scatter plot is most frequently used to illustrate the nature of the relationship between two or more variables.
- As can be seen from the preceding, a greater number of people generally prefer biking in temperatures between moderate and high
- Depending on the temperature and time of day, this analysis can be used daily for business purposes.

Data Visualization: Understand the relationship between variables

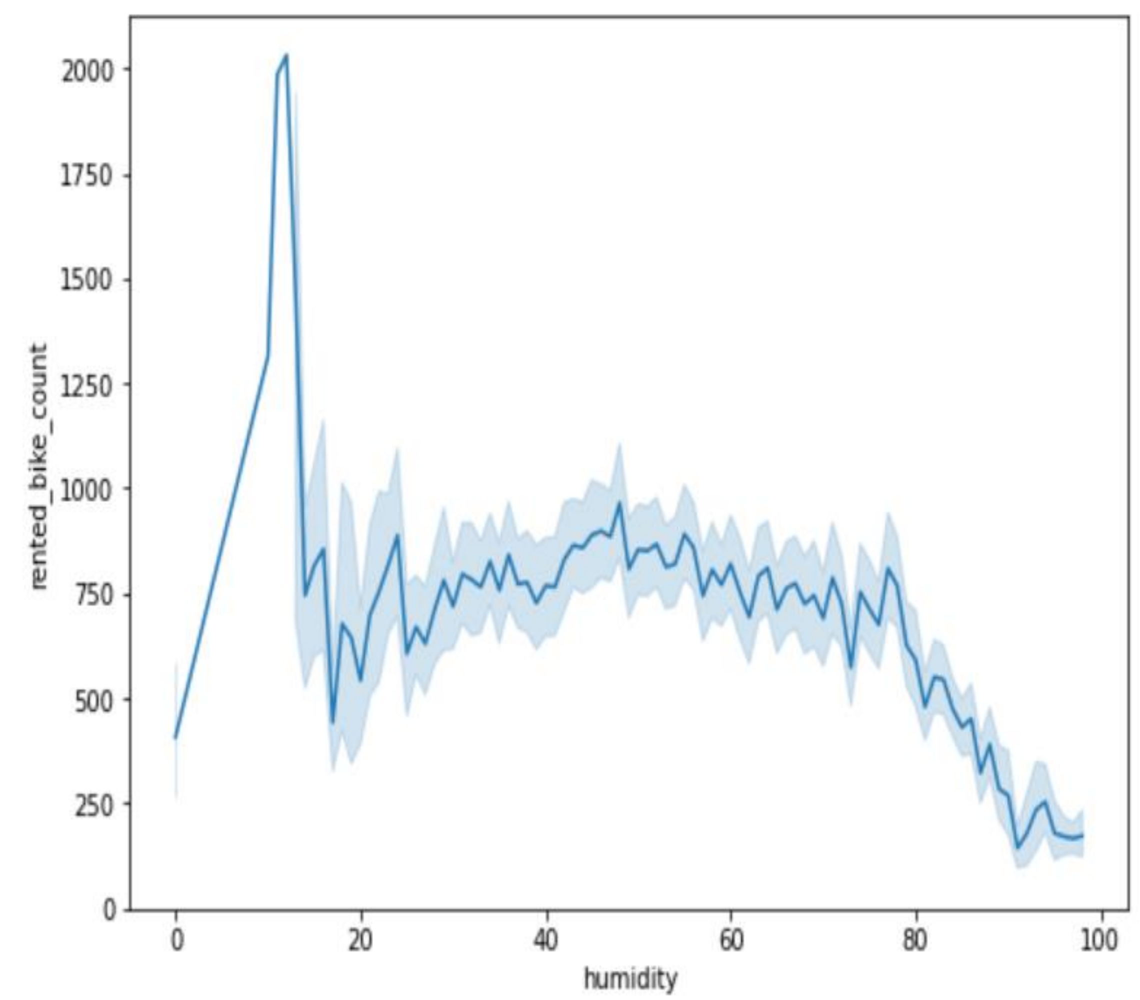


Chart 8 - Rented Bike Count Vs Humidity (Bivariate Analysis)

- We can see from the line plot above that the average number of bikes rented with humidity goes up and down sharply. For the number of rented bikes in demand, the most preferred humid environment is 0-20.

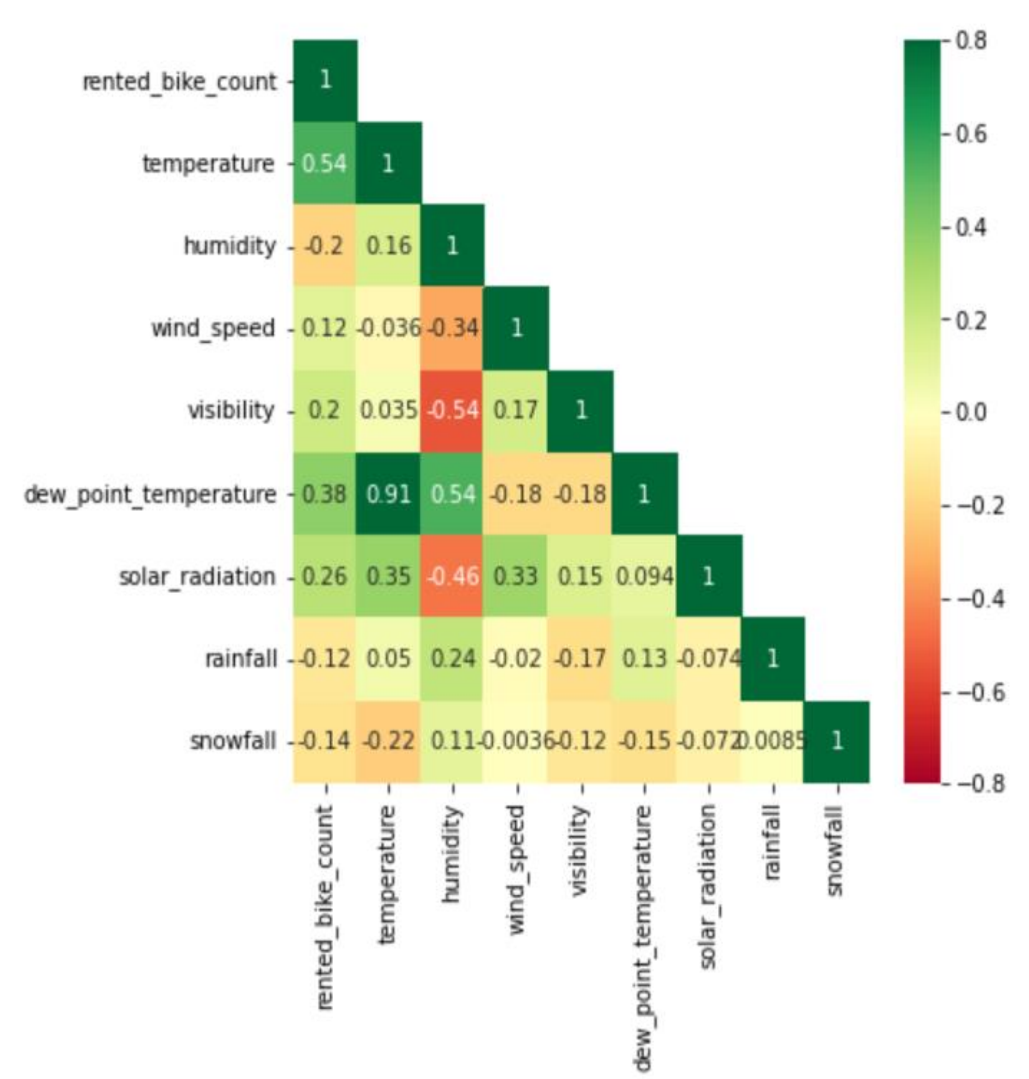
- When the humidity in the air is between 0 and 20, the distributor can increase the number of bikes that can be rented to manage demand.

- There are some other variables but those are not so important for analysis purposes.

Outlier Analysis

1. **Z score > 4 Pruning:** Z-score indicates how much a given value differs from the standard deviation, if it is greater than 4 we are considering those rows as outliers.
2. **Heatmap Plot:** it makes it simpler to visualize the relationship between variables and the dependent variable when conducting an analysis

- **Very Highly Correlated (0.7 - 0.9):** temperature and dew_point_temperature are very highly correlated as expected.
- **Moderately Correlated (0.5 - 0.7):** We see a moderate correlation between humidity and dew_point_temperature and temperature and rented_bike_count. This is probably only true for the range of temperatures provided.
- **Negative Correlation (less than 0):** We see a negative correlation between visibility and humidity solar radiation action and humidity. The more the humidity, the fewer people prefer to bike.
- **Low Correlation(near zero):** rented_bike_count has a weak dependence on windspeed, snowfall, and rainfall.



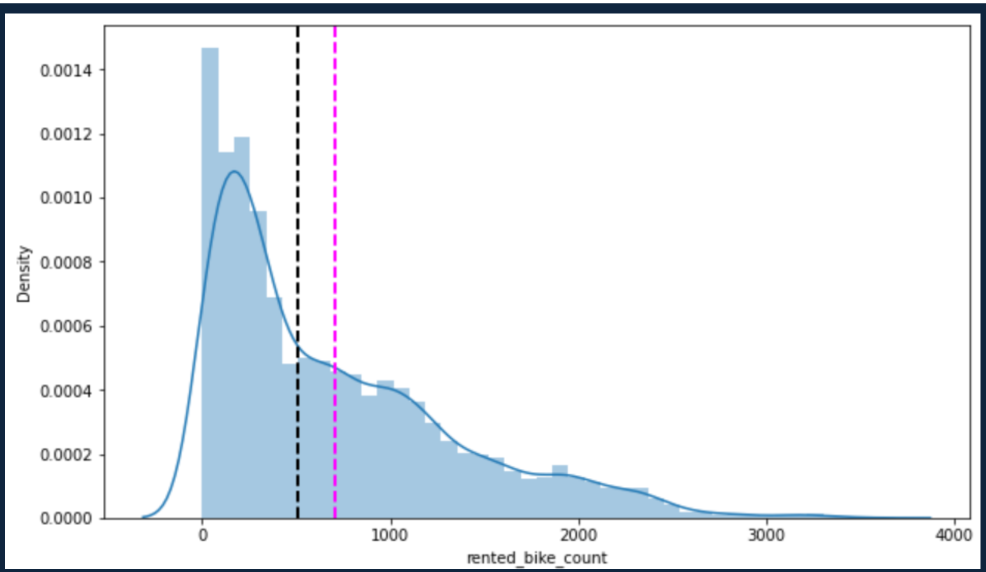
Feature Engineering and Data Pre-Processing

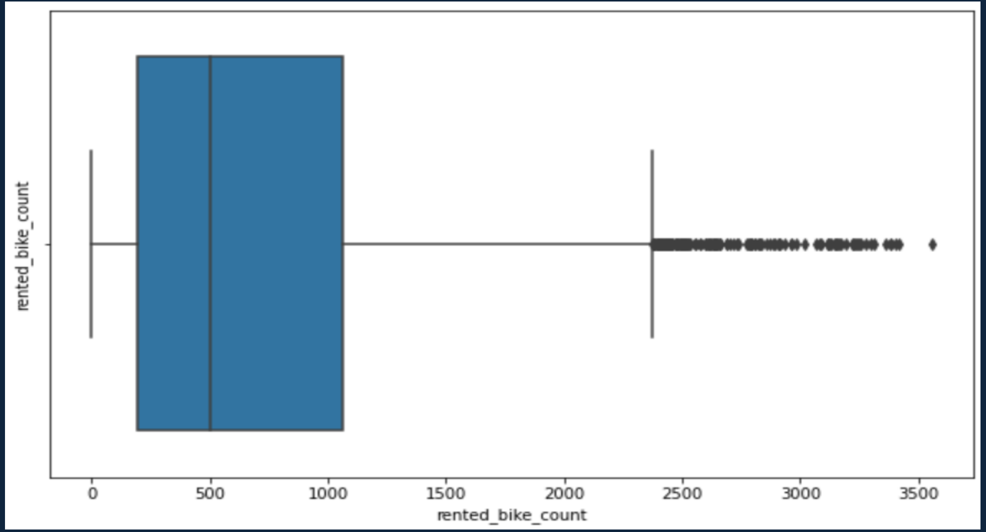
1. **Drop Column:**

- We are dropping some columns which are directly linked to any other column like season can be directly linked to month, holiday, day, date, and functioning_day with the weekend column.
- We are dropping highly correlated columns like temperature and dew_point_temperature. wind speed, rainfall, and snowfall are very poorly correlated with the rented_bike_count. Hence drop this column

2. **Categorical encoding**: Encoding categorical data is a process of converting categorical data into integer format so that the data with converted categorical values can be provided to the models to give and improve the predictions. We did this with month, hour, and other columns.

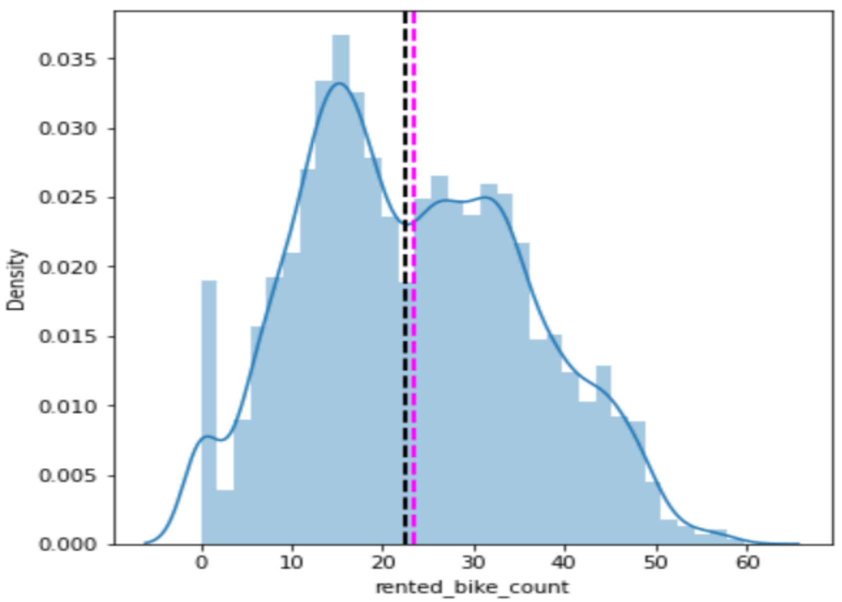
Feature Engineering and Data Pre-Processing





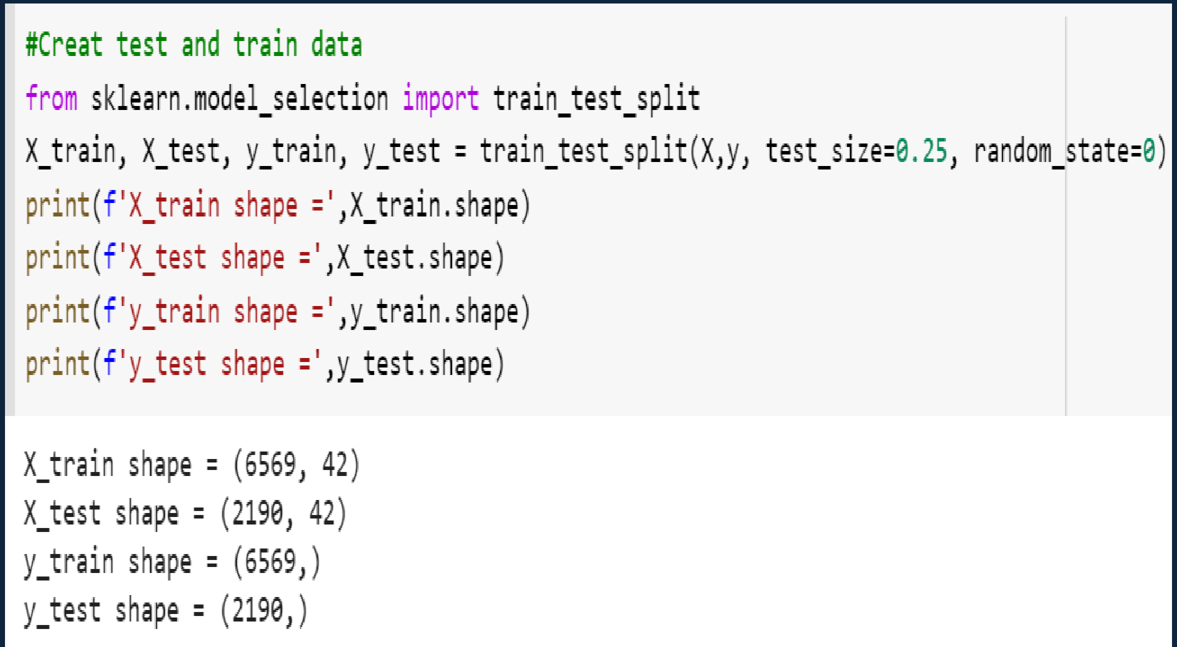
3. **Normalize our dependent variable: Right skewness is moderate in the Rented Bike Count, as shown in the graph above. Since "the distribution of the dependent variable has to be normal" is the assumption of linear regression, we should carry out square root normalization to make it normal.**

(Almost normal)



Feature Engineering and Data Pre-Processing

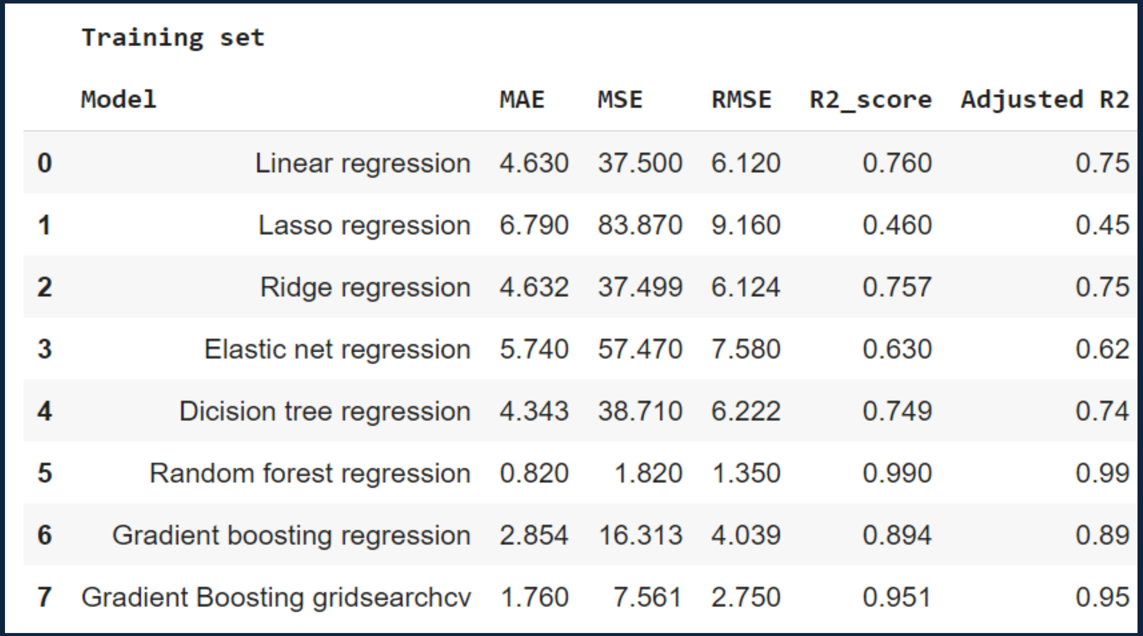
4. **Data Splitting:** Our dataset is small so we are using a 75:25 ratio, which means 75% of our data will be used for training to build our model and we will validate the model result with our 25% data.
5. **Evaluation metrics used:** We are estimating the mean squared error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R-squared (R2), and Adjusted R-squared for our training dataset as well as our test dataset to compare the results of our models.
   Because R square estimates the relationship between the movements of a dependent variable and those of an independent variable, R square is the best evaluation method for predicting the rented_bike_count.



```
#Creat test and train data
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.25, random_state=0)
print(f'X_train shape =',X_train.shape)
print(f'X_test shape =',X_test.shape)
print(f'y_train shape =',y_train.shape)
print(f'y_test shape =',y_test.shape)
```

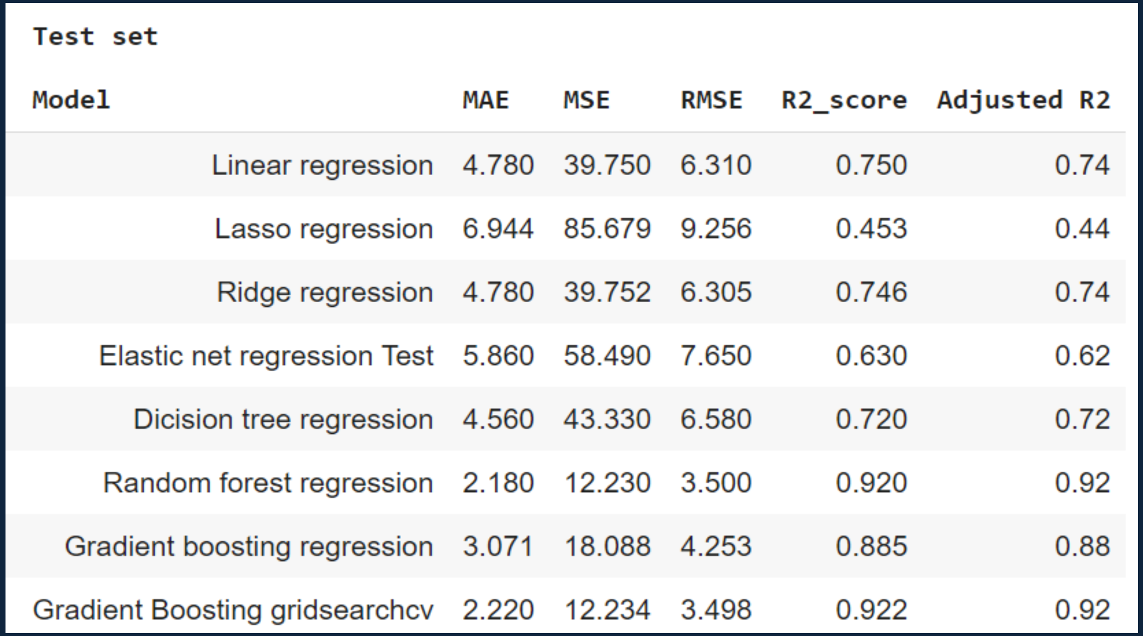
```
X_train shape = (6569, 42)
X_test shape = (2190, 42)
y_train shape = (6569,)
y_test shape = (2190,)
```

ML Model Implementation



Training set

	Model	MAE	MSE	RMSE	R2_score	Adjusted R2
0	Linear regression	4.630	37.500	6.120	0.760	0.75
1	Lasso regression	6.790	83.870	9.160	0.460	0.45
2	Ridge regression	4.632	37.499	6.124	0.757	0.75
3	Elastic net regression	5.740	57.470	7.580	0.630	0.62
4	Dicision tree regression	4.343	38.710	6.222	0.749	0.74
5	Random forest regression	0.820	1.820	1.350	0.990	0.99
6	Gradient boosting regression	2.854	16.313	4.039	0.894	0.89
7	Gradient Boosting gridsearchcv	1.760	7.561	2.750	0.951	0.95



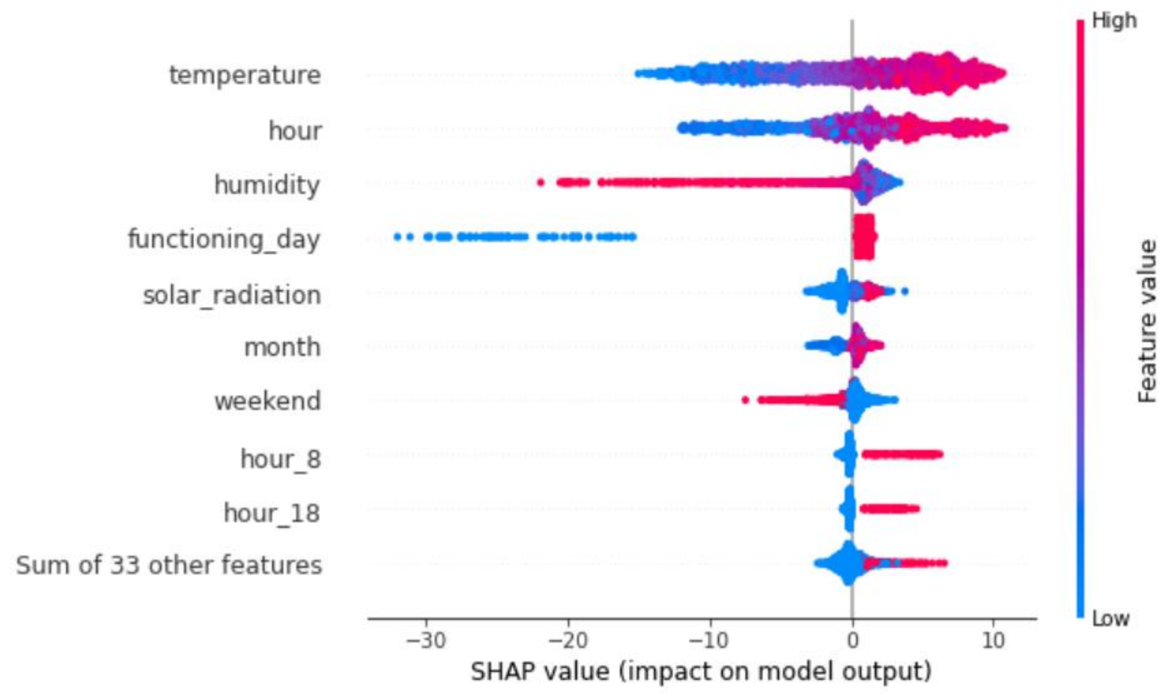
Test set

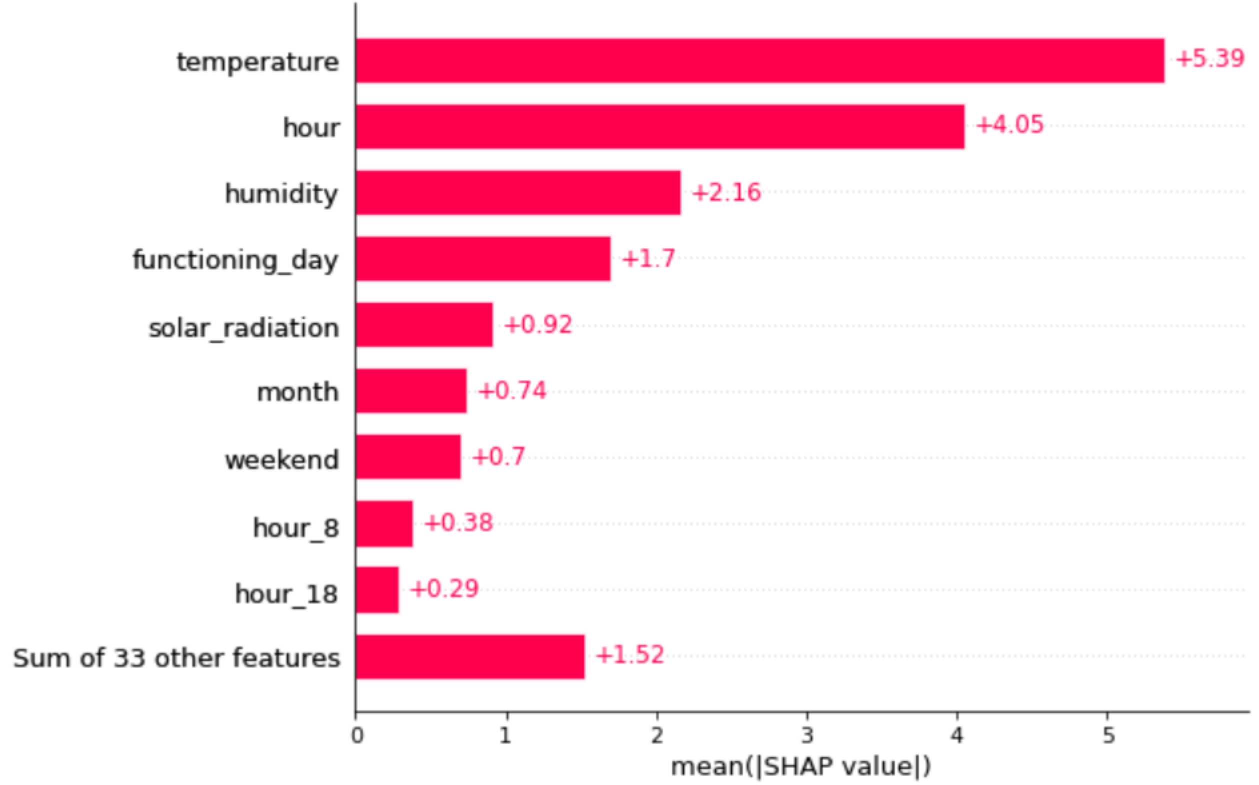
Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Linear regression	4.780	39.750	6.310	0.750	0.74
Lasso regression	6.944	85.679	9.256	0.453	0.44
Ridge regression	4.780	39.752	6.305	0.746	0.74
Elastic net regression Test	5.860	58.490	7.650	0.630	0.62
Dicision tree regression	4.560	43.330	6.580	0.720	0.72
Random forest regression	2.180	12.230	3.500	0.920	0.92
Gradient boosting regression	3.071	18.088	4.253	0.885	0.88
Gradient Boosting gridsearchcv	2.220	12.234	3.498	0.922	0.92

1. We have implemented 8 regression models to predict bike sharing demand and here is the result of all these models' training and test datasets. We also did hyperparameter tuning using Grid Search CV.
2. I chose the Random Forest model because, I need a better prediction for the number of rented bikes, and time is not a constraint.
3. Random Forest has the best R2 scores for the Test Set (0.92) and the Training Set (0.99), the lowest MSE, RMSE, and MAE in both training and test datasets.
4. Apart from these points Random Forest has the closest R square score for training and test datasets which shows it has the lowest overfitting.

ML Model Implementation

Shap Explainer: **Using the Shap explainer we can conclude that temperature is the most important feature followed by hour, humidity, functioning day and other variables**





Conclusion

Major Findings

1. We see 2 rental patterns across the day in bike rentals count - first for a Working Day where the rental count is high at peak office hours (8 am and 5 pm) and the second for a Non-working day where the rental count is more or less uniform across the day with a peak at around noon.
2. Temperature: People generally prefer to bike at moderate to high temperatures. We see the highest rental counts between 32 to 36 degrees Celcius
3. Season: We see the highest number of bike rentals in the Spring (July to September) and Summer (April to June) Seasons and the lowest in the Winter (January to March) season.
4. As one would expect, we see the highest number of bike rentals on a clear day and the lowest on a snowy or rainy day
5. Humidity: With increasing humidity, we see a decrease in the bike rental count.

Conclusion

Here are some solutions to manage Bike Sharing Demand

- The majority of rentals are for daily commutes to workplaces and colleges. Therefore open additional stations near these landmarks to reach their primary customers.
- While planning for extra bikes to stations the peak rental hours must be considered, i.e. 7–9 am and 5–6 pm.
- Start a new renting program for premium customers to increase business.
- Utilize the ML model to cater to demand efficiently.
- Be ready for 2 types of patterns in demand which are for working days and non-working days.
- Maintenance activities for bikes should be done at night due to the low usage of bikes during the night time. Removing some bikes from the streets at night time will not cause trouble for the customers.
- May start giving discounts to bookings if they book the bike in advance.
- Be proactive with communication. Ask for feedback often.
- Periodically throw Offers to retain customers.
- Define a roadmap for new customers.
- Stay competitive.

However, this is not the ultimate end. As Machine learning is an exponentially evolving field, having quality knowledge and keeping pace with the ever-evolving ML field would surely help one to stay a step ahead in the future.

> “Life may not be about your bike, but it sure can help you get through it.”
>
> Hallman

Thank you

Navneet Keshri

navneet2409jnv@gmail.com

linkedin.com/in/navneet-keshri-28650918b