**SEOUL BIKE SHARING DEMAND PREDICTION**

**Sundar K**

**Data science trainee**

**Alma better**

|  |  |
| --- | --- |
|  | **CONTENTS** |
| **1** | **Abstract** |
| **2** | **Problem statement** |
| **3** | **Introduction** |
| **4** | **Methodology** |
| **5** | **Steps involved** |
| **6** | **Data analysis** |
| **7** | **Models** |
| **8** | **Result** |
| **9** | **Conclusion** |
| **10** | **References** |

**1.ABSTRACT**

The bike sharing system has brought wide convenience to residents in the city and serves as important tools to transport from one place to another place. For the bike sharing companies, they need to know the total users of bike, so they can release suitable number of bikes into the market. This paper uses visualization technology to visualize data and figure out the possible factors which can impact the total number of users. After completing the data analyzing, this paper figures out the season, weather, temperature, humanity and wind speed are the main factors which can have impacts on the total number of users. the second stages, this paper uses regression model, to predict the possible number of bike users. The input factors are season, weather, temperature, humanity and wind speed. By analyzing regression model results, the model has the best prediction, which can be used for real practice.

**2.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**.**

**3.INTRODUCTION**

Bicycle Sharing Systems are very green, healthy and cheap way to navigate from one place to another place. Now with the new methods of electronic sharing and registration, the whole process of bicycle sharing, from the rental to returning back has become much more automatic and convenient. Through the bicycle sharing system, the users can easily rent a bicycle from one place and return it in another place. Bicycle sharing companies such as Airbike, OBike and Mobike has become much more popular in the world in the past few years, due to the pro-health and environment-friendly mode of transport. The bike sharing systems can bring a lot of data, but the characteristics of the data have not been fully understood. Opposed to other data which is generated by transport devices such as subway and bus. The bike sharing system is a virtual sensor network since the bike can be used to detect the most important events in the city by monitor these data

**4.METHODOLOGY**

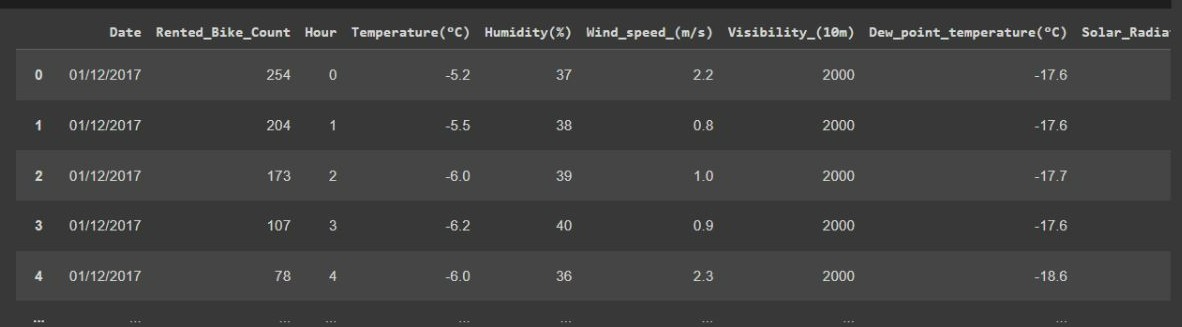
**4.1 DATA COLLECTION**

The dataset in this project is provided from alma better. which includes two years dataset 2017 to 2018.The dataset has 13 attributes including weather situation, date, weekday/public holiday, the count of number of bikes rented on that day and temperature

**4.2 DATA DESCRIPTION**

The bike sharing data is recorded on daily basis. The records are about 731 days. All the data information can be shown in Table 1.

Table 1.



In the bike sharing dataset, the dataset has following fields:

1. Instant -record index
2. Date : year-month-day
3. Rented Bike count - Count of bikes rented at each hour
4. Hour - Hour of the day
5. Temperature-Temperature in Celsius
6. Humidity - %
7. Windspeed - m/s
8. Visibility - 10m
9. Dew point temperature - Celsius
10. Solar radiation - MJ/m2
11. Rainfall - mm
12. Snowfall - cm
13. Seasons - Winter, Spring, Summer, Autumn
14. Holiday - Holiday/No holiday
15. Functional Day – NoFunc (Non Functional Hours), Fun(Functional hours)

**5.Steps involved:**

* **Exploratory Data Analysis**

After loading the dataset we performed this method by comparing our target variable that is Rented\_bike\_Count with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

* **Null values Treatment**

Our dataset contains a large number of null values which might tend to disturb our accuracy hence we replaced it with zero made them at the beginning of our project inorder to get a better result.

* **Encoding of categorical columns**

We used One Hot Encoding to produce binary integers of 0 and 1 to encode our categorical features because categorical features that are in string format cannot be understood by the machine and needs to be converted to numerical format.

**6 DATA ANALYSIS**

In order to figure out which factors have impact on the number of bikes. The dist plot, box plot, the bar plot and scatter plot will be drawn to show the influence on the number of bikes.

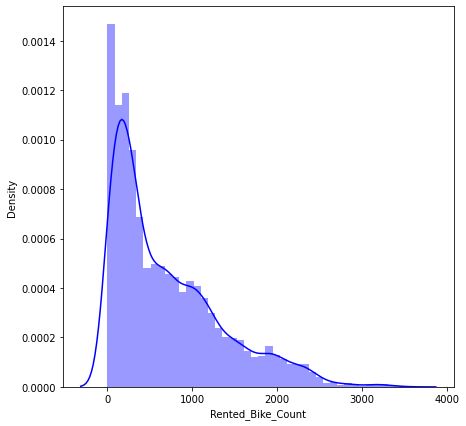


Figure 1.a the distplot

In the project Figure 1.b shows that there is a relationship between density and number of users, the density has 0.0014 which as the highest number

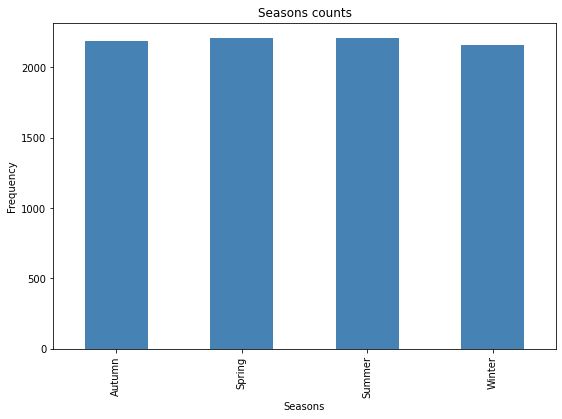
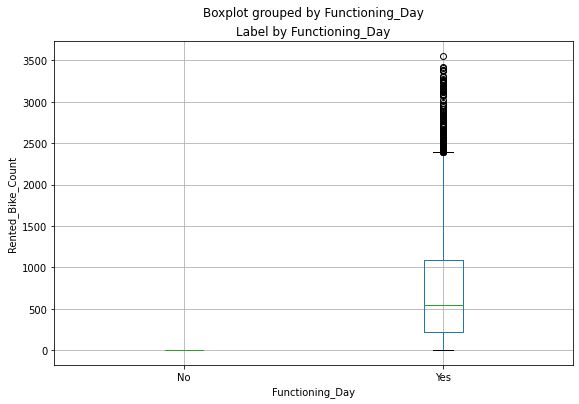


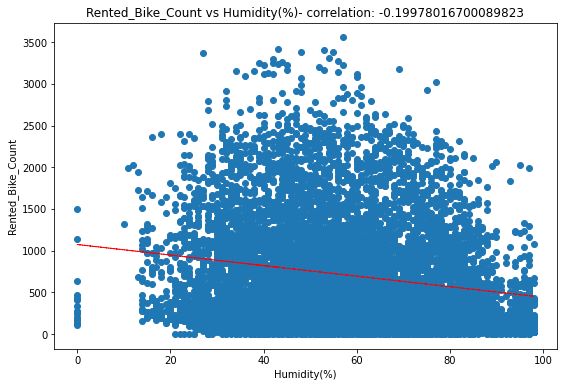
Figure 1.b the bar plot

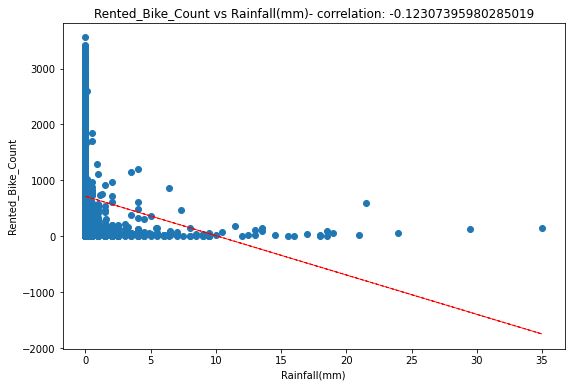


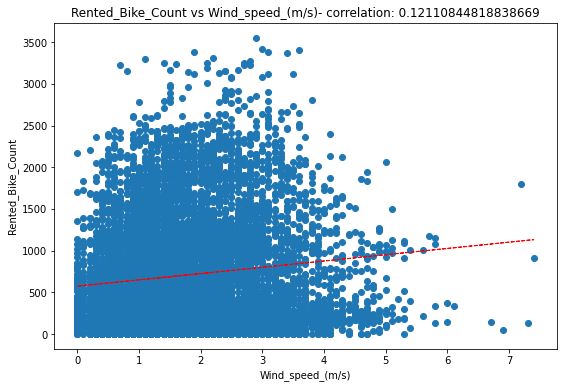
1.b shows that there is a relationship between season and frequency, the frequency has 2000 which as the highest number

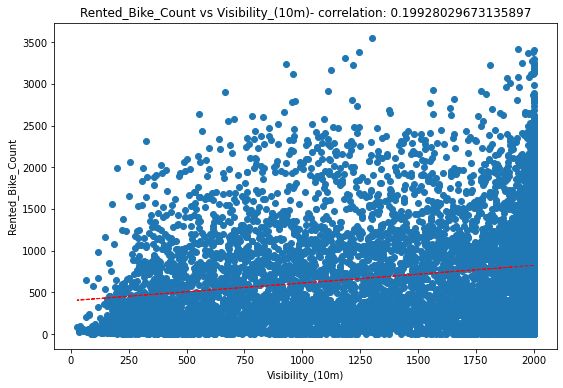
Figure 1.c

In the project Figure 1.c shows that there is a relationship between functioning day and number of users.









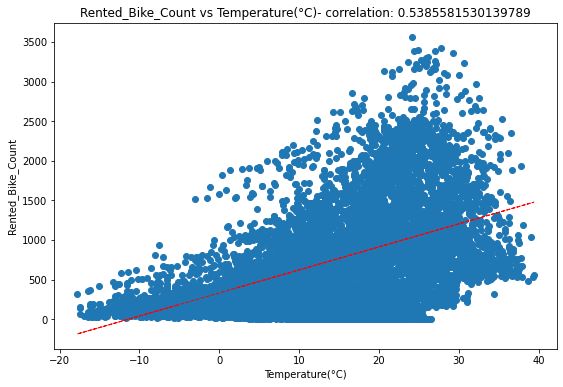


Figure 1.d corelation between different factors and users

**7. MODELS**

**5.1 LINEAR REGRESSION MODEL**

In order to predict the number of bikes in the future, this paper will build a multiple linear regression model. The multiple linear regression model is to model the correlation between two or more variable and a response variable by fitting a linear equation to observed data.

A multiple linear regression model with k predictor variable x1,x2,…,xp and a response Y, can be written as

𝑦 = 𝛽0 + 𝛽1𝑥1 + 𝛽2𝑥2 + ⋯ +𝛽𝑝𝑥𝑝 + 𝜖 (1)

The 𝜖 is the error term which can be used to estimate the accuracy of the linear regression model.

When giving a data set{𝑦𝑖 , 𝑥𝑖1, … 𝑥𝑖𝑝} 𝑖=1 𝑛 of the n statistical units, a linear regression model assume that the relationship between the dependent variable y and the given x is linear. The 𝜖 is the error variable. Thus the model can be written as

𝑦𝑖 = 𝛽0 + 𝛽1𝑥𝑖1 + 𝛽2𝑥𝑖2 + ⋯ +𝛽𝑝𝑥𝑖𝑝 + 𝜖𝑖 𝑖 = 1, … , 𝑛 (2)

The formula (2) can be denoted as matrix notation as

𝑦 = 𝑋𝛽 + 𝜖 (3)

where 𝑦 = { 𝑦1 𝑦2 . . . 𝑦𝑛} , 𝑋 = { 𝑥1 𝑇 𝑥2 𝑇 . . . 𝑥𝑛 𝑇} = { 1 𝑥11 . . . 𝑥1𝑝 1 𝑥21 . . . 𝑥2𝑝 . . . . . . . . . . . . . . . . . . 1 𝑥𝑛1 . . . 𝑥𝑛𝑝} , 𝛽 = { 𝛽0 𝛽1 . . . 𝛽𝑝} , 𝜖 = { 𝜖0 𝜖1 . . . 𝜖𝑝} 3.2

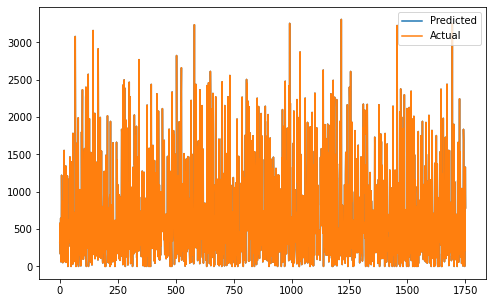
**5.2 VARIABLE SELECTION IN LINEAR REGRESSION MODEL**

A variable selection method should to select the best variable for a special purpose such as prediction. The best should find the balance between the goodness of fit and the number of variables. Earlier selection standard uses the residual sums of squares which have overfit problems (Miller, 2002). Akaike found a Akaike Information Criterion method which can be used for model or variable selection via Kullback-Leibler divergence (Gutierrez & Heming, 2018). The AIC can be given as

𝐴𝐼𝐶𝑃 = 𝑛𝑙𝑛(𝜎𝑚𝑙𝑒 2 ) + 2𝑝

Where 𝜎𝑚𝑙𝑒 2 is the maximum likelihood estimate of 𝜎 2. By using this selection standard, the model with the smallest AIC can be deemed as the best

**8 Result**



**9 Conclusion**

That's it! We reached the end of our exercise.

Starting with loading the data so far we have done EDA , null values treatment, encoding of categorical columns, feature selection and then model building. It is quite evident from the results that Linear regression is the best model that can be used for the Bike Sharing Demand Prediction since the performance metrics (MSE,RMSE) shows lower and (r2,adjusted\_r2) show a higher value

**10 Reference**

1. MachineLearningMastery
2. GeeksforGeeks
3. Analytics Vidhya