# **Project\_Title:-Seoul Bike Sharing Demand Prediction**

# **Lalit Ahirrao**

# **AniketGajmal**

# **Rushikesh Pawar**

# **Prasad Ghegade**

# **Samarth Gangurde**

**ABSTRACT:**

These bike sharing systems have become widely popular across different cities in the world.The elements of affordability and accessibility have helped in the promotion and expansion of the concept of a bike rental system. And it definitely comes across as a win-win for just about anyone who is willing to skip the expensive cab services for a bike.Discussions around the positive effects of bike sharing often focus on environmental sustainability and the personal health benefits experienced by users. When new bike share schemes are introduced, there’s often a reference to the “greening” of the city; ideally, more bikes can equal fewer cars, which leads to a reduction in greenhouse gasses and an improvement in local air quality

**PROBLEM STATEMENT:**

Bike share systems have been growing in popularity across the nations. The concept is simple. There are racks of bikes set up around the city, and people can rent a bike for a short period of time, even if only to get from point A to point B.Covid-19 has had a significant impact on shared mobility and more particularly

on the use of shared bikes.As people reassess ground transportation options

in the face of the COVID-19 pandemic, many are choosing isolated modes such as sharing bikes over public transportation. This analysis focuses on factors that will predict the overall demand of bikes, maximize The availability of bikes to the customer, minimize the time of waiting to get a bike on rent.

**DATA DESCRIPTION:**

The data description phase starts with an initial data collection and proceeds with activities in order to get familiar with the data. Identifying data quality problems, discovering first insights into the data and detecting interesting subsets to form hypotheses from hidden information are activities of this step. Data which is collected from a rented bike provider company from Seoul to get analyzed, involves usage details of customers from. The data was taken from a rented bike Provider Company. It has 8760 rows and 14 columns. Most columns related to hourly bike count for rent. Other column was indicative of weather conditions affecting bike count per hour.

**ATTRIBUTES :**

* Date : year-month-day
* Rented Bike count - Count of bikes rented at each hour
* Hour - Hour of he day
* Temperature-Temperature in Celsius
* Humidity - %
* Wind Speed - 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)

**FEATURE BREAKDOWN:**

**Date**: The date of the day, during 365 days from 01/12/2017 to 30/11/2018, formatting in DD/MM/YYYY, we need to convert into date-time format.

**Rented Bike Count**: Number of rented bikes per hour which our dependent variable and we need to predict tha*t*

**Hour:** The hour of the day, starting from 0-23 it's in a digital time format

**Temperature (°C):** *Temperature of the weather in Celsius and it varies from -17***°**C to 39.4**°**C*.*

**Humidity (%)**: Availability of Humidity *in the air during the booking and ranges from 0 to 98%.*

**Wind speed (m/s):** Speed of the wind while booking and ranges from 0 to 7.4m/s.

**Visibility (10m):** Visibility to the eyes during driving in “m” and ranges from 27m to 2000m.

**Dew point temperature: (°C)**:*Temperature At the beginning of the day*and it ranges from -30.6**°**C to 27.2**°**C.

**SolarRadiation(MJ/m2)**

Sun contribution or solar radiation during ride booking which varies from 0 to 3.5 MJ/m2.

**Rainfall (mm):** The amount of rainfall during bike booking which ranges from 0 to 35mm.

**Snowfall (cm):** Amount of snowing in cm during the booking in cm and ranges from 0 to 8.8 cm.

**Seasons:** Seasons of the year and total there are 4 distinct seasons I.e. summer, autumn, spring and winter.

**Holiday:** If the day is holiday period or not and there are 2 types of data that is holiday and no holiday

**DATA SOURCING:**

Data Sourcing is the process of finding and loading the data into our system. Broadly there are two ways in which we can find data.

1. Private Data

2. Public Data

Data collected from several sources must be stored in the correct format and transferred to the right information technology personnel within a company. As mentioned previously, data can be collected from several objects on several events using different types of sensors and storage tools.

## **PREPROCESSING OF DATASET:**

Preprocessing of Data can refer to manipulation or dropping of data before it is used in order to ensure or enhance performance.as name suggest we prefer the preprocessing of data before training the model.Data preprocessing includes cleaning, Instance selection, normalization, transformation, feature extraction and selection, etc. The product of data preprocessing is the final training set

**DATA CLEANING:**

After completing the Data Sourcing, the next step in the process of EDA is Data Cleaning. It is very important to get rid of the irregularities and clean the data after sourcing it into our system.

Irregularities are of different types of data.

* · Missing Values
* · Incorrect Format
* · Incorrect Headers
* · Anomalies/Outliers

**DATA TRANSFORMATION:**

Data transformation is the process of normalizing and aggregating the data to

further improve the efficiency and accuracy of data mining.

**DATA DEDUPLICATION:**

It is very likely that your dataset contains duplicate rows. Removing them is essential to enhance the quality of the dataset.

**MISSING VALUES:**

There is a representation of each service and product for each customer. Missing values may occur because not all customers have the same subscription. Some of them may have a number of services and others may have something different. In addition, there are some columns related to system configurations and these columns may have null values but in our orange telecom data set there are no null values present

If there are missing values in the Dataset before doing any statistical analysis, we need to handle those missing values.

There are mainly three types of missing values.

1. MCAR (Missing completely at random): These values do not depend on any other features.

2. MAR (Missing at random): These values may be dependent on some other features.

MNAR (Missing not at random): These missing values have some reason for why they are missing.

**DROPPING MISSING VALUES:**

One of the ways to handle missing values is to simply remove them from our dataset. We have know that we can use the isnull() and notnull() functions from the pandas library to determine null values

**HANDLING OUTLIERS:**

Outlier treatment is another important step in data pre-processing (one may prefer doing missing value treatment once outliers are treated, especially if using mean imputation as outlier can skew the data). Outlier as such is an observation that lies on an abnormal distance from other values or any observation far away from the mass of data or the overall pattern. Outliers can be mild and extreme with the extreme being away from the source by a great deal. Also, an outlier can be looked for in each variable (Univariate Outlier) or can be looked for in relation to other variables (Bivariate Outlier).

**EXPLORATORY DATA ANALYSIS:**

EDA means trying to understand the given data much better, so that we can make some sense out of it. Using univariate frequency analysis was conducted to describe key characteristics of each feature including, minimum and maximum value, average, standard deviation and others. It was also used to produce a value distribution and identify missing values, and outliers.

EDA is a process of examining the available dataset to discover patterns, spot anomalies, test hypotheses, and check assumptions using statistical measures. In this chapter, we are going to discuss the steps involved in performing top notch exploratory data analysis

In statistics, A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing tasked in Python uses data visualization to draw meaningful patterns and insights

**UNIVARIATE ANALYSIS:**

If we analyze data over a single variable/column from a dataset, it is known as Univariate Analysis. Univariate analysis looks at one feature at a time. When we analyze a feature independently, we are usually mostly interested in the distribution of its values and ignore other features in the dataset

Univariate analysis is the simplest form of analyzing data. It means that our data has only one type of variable and that we perform analysis over it. The main purpose of univariate analysis is to take data, summarize that data, and find patterns among the values. It doesn't deal with causes or relationships between the values. Several techniques that describe the patterns found in univariate data include central tendency (that is the mean, mode, and median) and dispersion (that is, the range, variance, maximum and minimum quartiles (including the interquartile range), and standard deviation).

**BIVARIATE ANALYSIS:**

If we analyze data by taking two variables/columns into consideration from a dataset, it is known as Bivariate Analysis.

**Numeric-Numeric Analysis:**

Analyzing the two numeric variables from a dataset is known as numeric-numeric analysis. We can analyze it in three different ways.

· Scatter Plot

· Pair Plot

· Correlation Matrix

·**Numeric - Categorical Analysis:**

Analyzing the one numeric variable and one categorical variable from a dataset is known as numeric-categorical analysis. We analyze those mainly using mean, median, and box plots.

**MULTIVARIATE ANALYSIS:**

Multivariate analysis is the analysis of three or more variables. This allows us to look at correlations (that is, how one variable changes with respect to another) and attempt to make predictions for future behavior more accurately than with bivariate analysis.

One common way of plotting multivariate data is to make a matrix scatter plot, known as a pair plot. A matrix plot or pair plot shows each pair of variables plotted against each other. The pair plot allows us to see both the distribution of single variables and the relationships between two variables

## **CREATING DUMMY VARIABLES:**

A dummy variable (aka, an indicator variable) is a numeric variable that represents categorical data, such as Hour, month, season, etc.

Technically, dummy variables are quantitative variables. Their range of values is small; they can take on only two quantitative values. As a practical matter, regression results are easiest to interpret when dummy variables are limited to two specific values, 1 or 0. Typically, 1 represents the presence of a qualitative attribute, and 0 represents the absence.

The number of dummy variables required to represent a particular categorical variable depends on the number of values that the categorical variable can assume. To represent a categorical variable that can assume k different values, a researcher would need to define k - 1 dummy variables.

For example, suppose we are interested in Holiday, a categorical variable that might assume two values - its Holiday or its not Holiday. We could represent Holiday with two dummy variables:

X1 = 1, if it's Holiday otherwise,

X1 = 0, if it's not Holiday .

**ALGORITHMS WE USED:**

* LINEAR REGRESSION
* LASSO REGRESSION
* DECISION TREE
* RANDOM FOREST
* GRADIENT BOOSTING

## **LINEAR REGRESSION:**

Let’s know what linear regression is. It is very important and used for easy analysis of the dependency of two variables. One variable will be considered to be an explanatory variable, while others will be considered to be a dependent variable. Linear regression is a linear method for modeling the relationship between the independent variables and dependent variables. The linearity of the learned relationship makes the interpretation very easy. Linear regression models have long been used by people as statisticians, computer scientists, etc. who tackle quantitative problems.

The Formula of Linear Regression Let’s know what a linear regression equation is. The formula for linear regression equation is given by:

## y = a + bx

a and b can be computed by the following formulas:

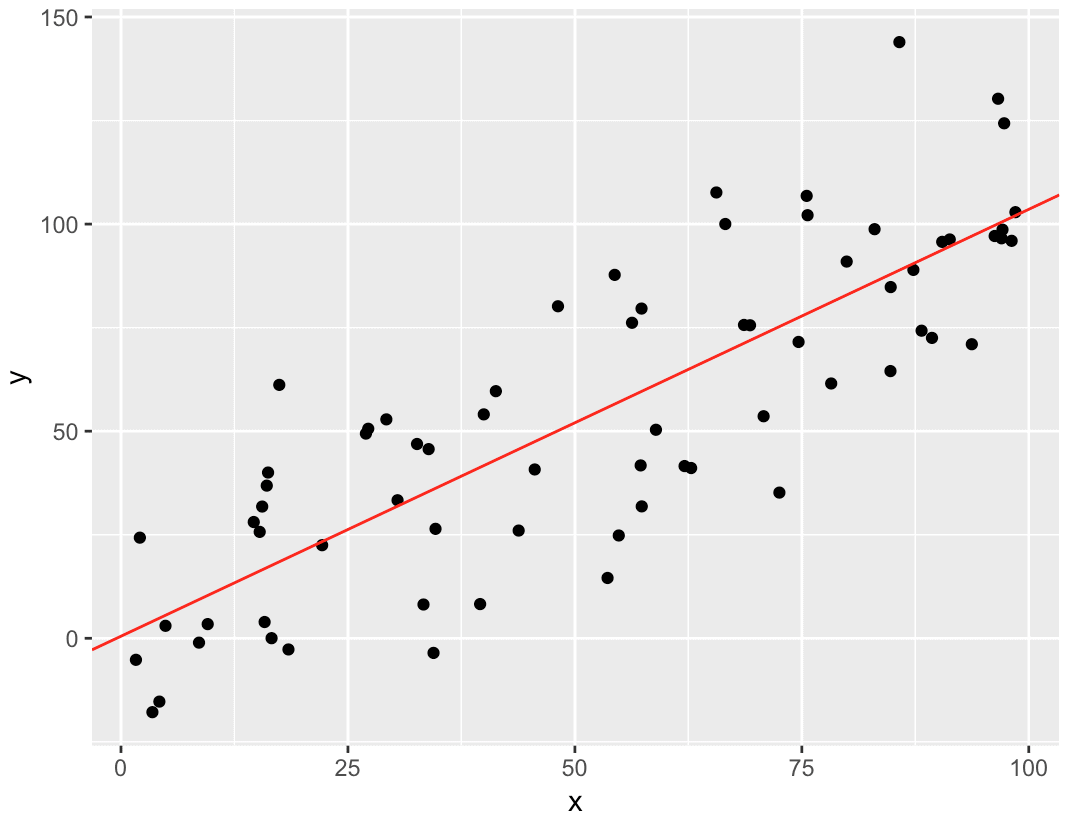
## b= n∑xy−(∑x)(∑y) / n∑x2−(∑x)2

## a= ∑y−b(∑x) / n

Where

x and y are the variables for which we will make the regression line.

* b = Slope of the line.
* a = Y-intercept of the line.
* X = Values of the first data set.
* Y = Values of the second data set.

****

In the above diagram the black dots we see are the distribution of 'Y' w.r.t 'X' there is no straight line that runs through all the data points So, the objective here is to fit the best fit of straight line that will try to minimize the error between the expected and actual value

**REFERENCES:**

* [EDA: Exploratory Data Analysis | Introduction to Exploratory Data Analysis (analyticsvidhya.com)](https://www.analyticsvidhya.com/blog/2021/02/introduction-to-exploratory-data-analysis-eda/)
* [End to end Case Study: Bike-sharing demand prediction | by Pawan Reddy Ulindala | Towards Data Science (medium.com)](https://medium.com/towards-data-science/end-to-end-case-study-bike-sharing-demand-dataset-53201926c8db)