Bike Sharing Demand Prediction

❖ INTRODUCTION:

Renting bike is a booming business due to various reasons. According to the data compiled by the team behind the Medina bike sharing map, there are currently over 3,000 bike share systems in cities around the globe Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis.

The benefits of bike sharing schemes include transport flexibility, reductions to vehicle emissions, health benefits, reduced congestion and fuel consumption, and financial savings for individuals.

1.Problem Statement:

Rented bikes are introduced in many urban cities for the enhancement of the mobility comfort. Our objective is to predict the count of bikes required at each hour throughout the day so that bikes are accessible to the public at right time and thus lessening the waiting time.

Data description

FEATURE	TYPE
Date: year-month-day	Date
Rented Bike Count	Int64
Hour	Int64
Temperature(°C)	Float64
Humidity (%)	Int64
Wind speed (m/s)	Float64
Visibility (10m)	Int64

Dew Point temperature (°C)	Float64
Solar Radiation (MJ/m2)	Float64
Rainfall (mm)	Float64
Snowfall(cm)	Float64
Seasons	Object
Holiday	Object
Functioning day	Object

- **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 that
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): At the beginning of the day and it ranges from -30.6°C to 27.2°C.

Solar Radiation (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

Functioning Day: If the day is a Functioning Day or not and it contains object data type yes and no.

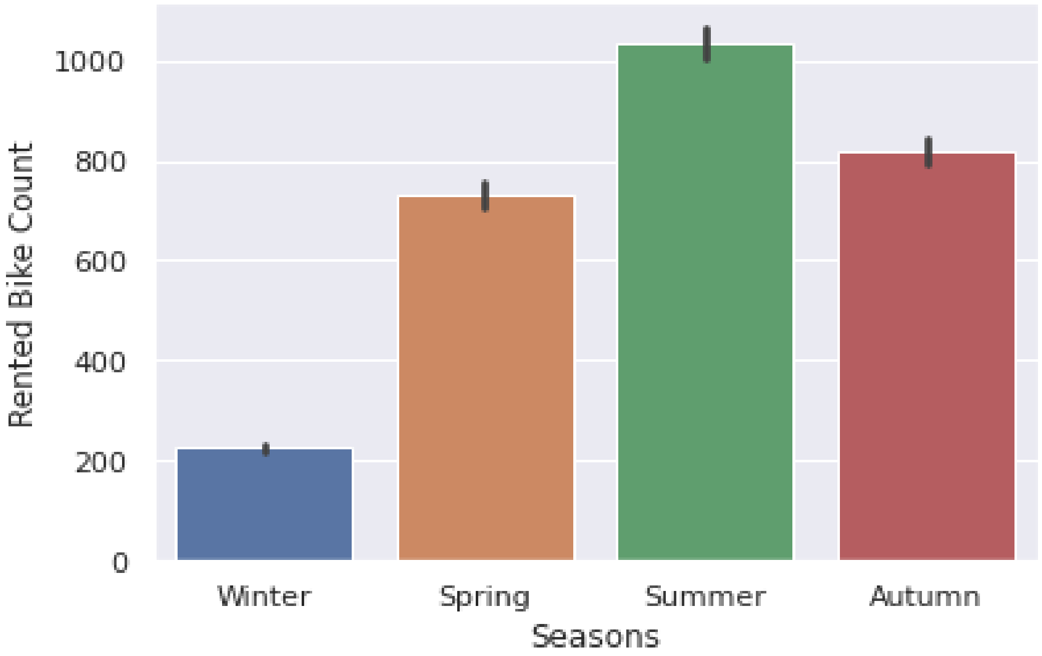
2.EDA (Exploratory data analysis)

The main purpose of EDA is to detect any errors, outliers as well as to understand different patterns in the data. It allows Analysts to understand the data better before making any assumptions. The outcomes of EDA help businesses to know their customers, expand their business and take decisions accordingly.

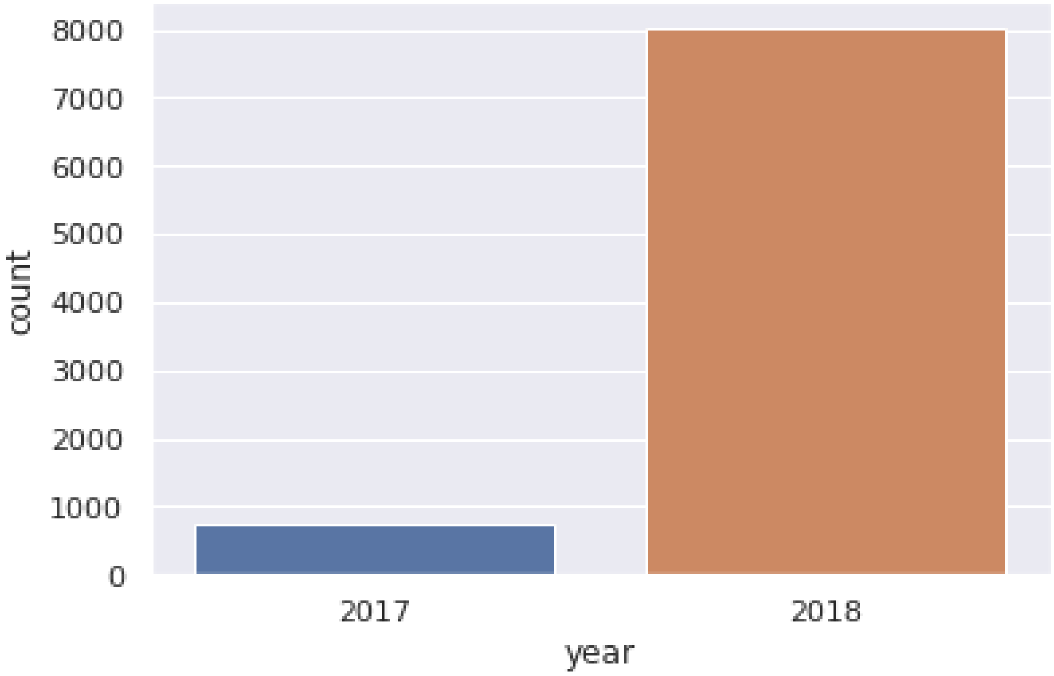
For effective EDA we extracted the year, day of the week and month from the Date column

- **Below are few highlights from the analysis-**

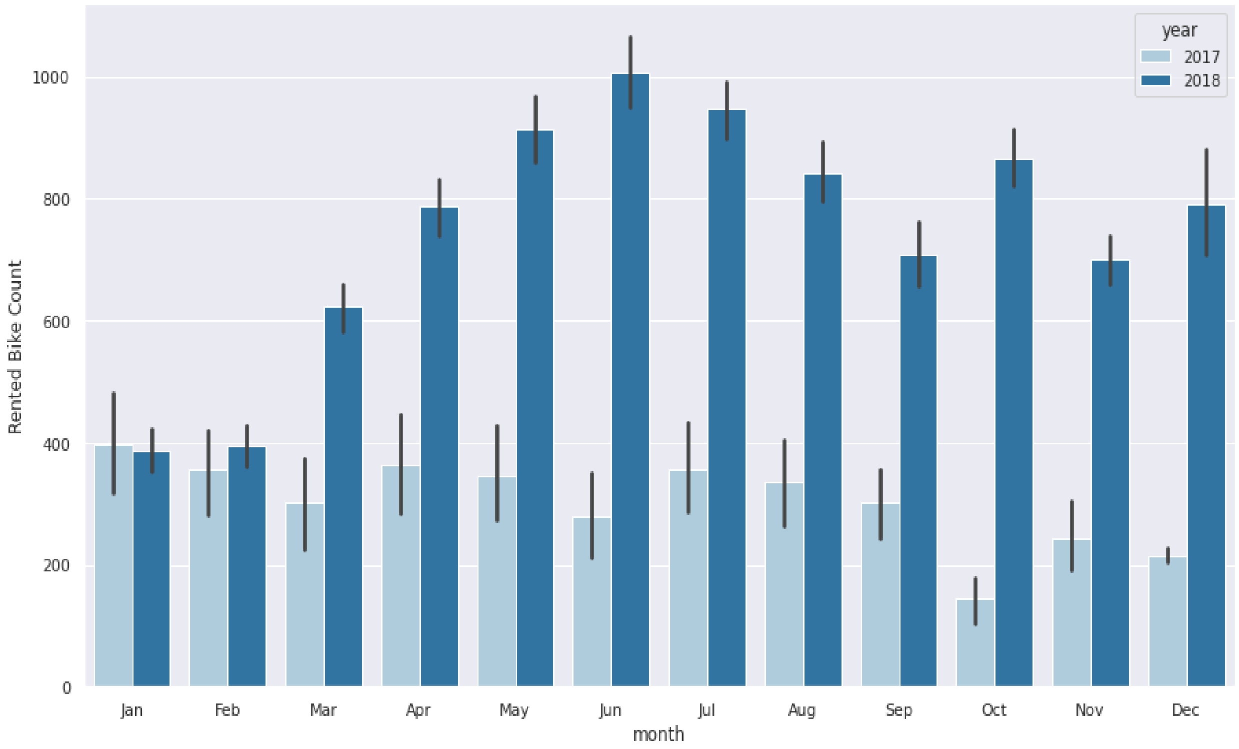
a) Season wise demand



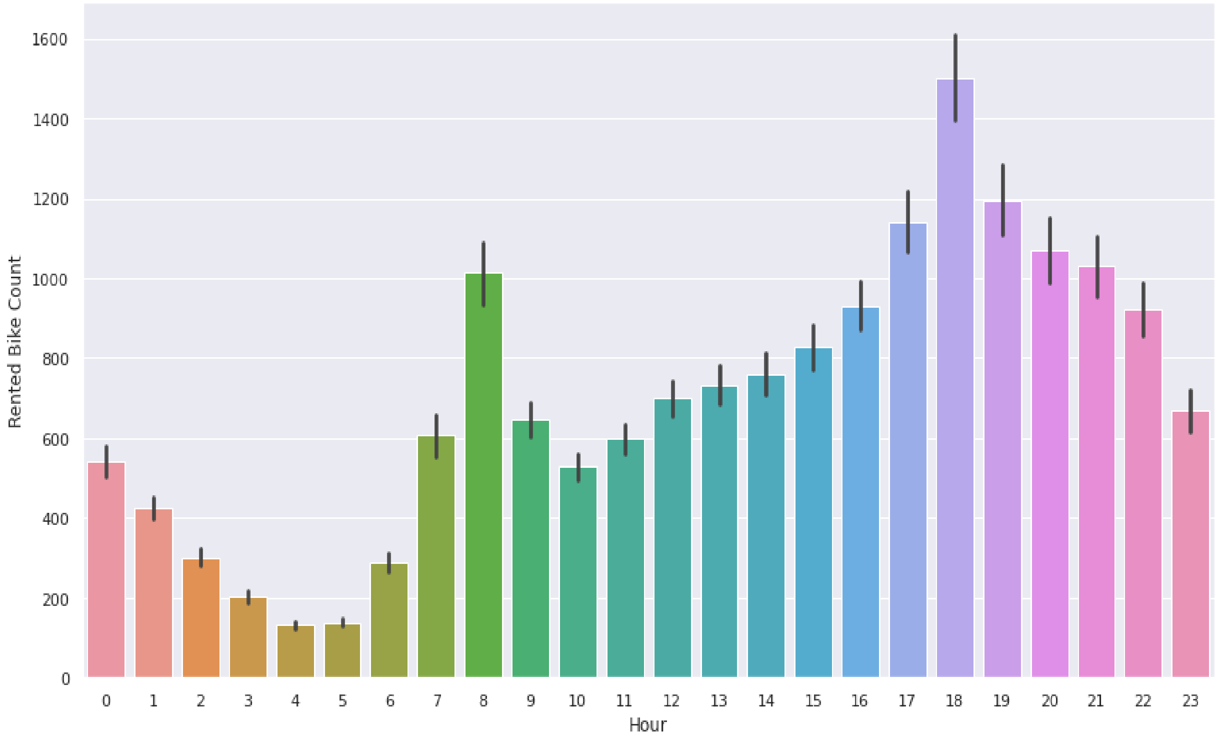
b) Year wise demand



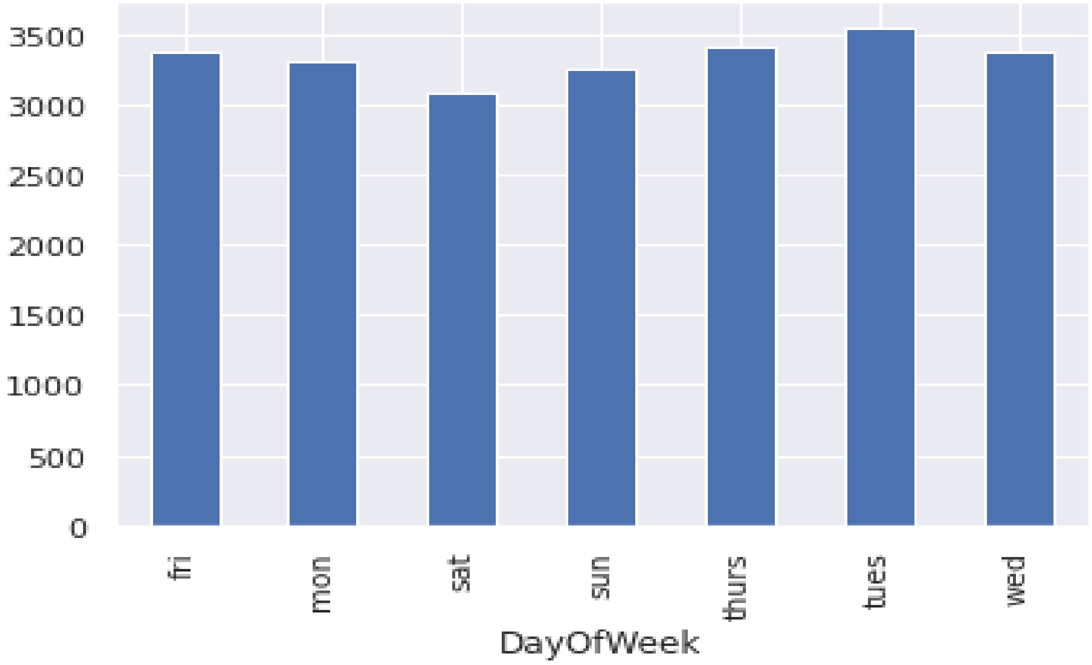
c) Month wise demand



d) Hour wise demand



e) Day wise demand



3.Data Pre-processing:

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.
We performed the data pre-processing in following manner:

1. We looked for the null or duplicate values and found none in the given dataset
2. there were outliers in many features of our dataset so treated the outliers by removing them and filled the NaN values with mean of the respective column
3.We converted the categorical features into numeric by creating dummy variables.

We created dummy variables for season, day of the week, year, month, holiday and functioning day columns

5. We checked the correlation of features with our target variable and removed the highly correlated variables. The most highly correlated variables were Dew point temperature with a positive correlation of 0.91. We dropped snowfall and rainfall as well due to presence of NaN values after getting rid of outliers.

4.Training the Model:

a) Assigning the dependent and independent variables
b) Splitting the model into train and test sets.
c) Transforming data using minmaxscaler.
d) Fitting linear regression on train set.
e) Getting the predicted dependent variable values from the model.

5.Evaluating metrics of our models:

A. Getting MSE, RMSE, R2-SCORE, ADJUSTED-R2 SCORE for different models used.

- MSE - the mean squared error or mean squared deviation of an estimator measures the average of the squares of the errors.
- RMSE - Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are
- R2-SCORE - R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.
- ADJUSTED-R2 SCORE - Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected.

6.Models used:

- **Linear regression:**

Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

In linear regression, the relationships are modelled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis

- **Lasso regression model**

  Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e., models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination. The acronym “LASSO” stands for Least Absolute Shrinkage and Selection Operator. Lasso solutions are quadratic programming problems, which are best solved with software (like MATLAB). The goal of the algorithm is to minimize:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

- **Decision tree regression model**

  Linear model trees combine linear models and decision trees to create a hybrid model that produces better predictions and leads to better insights than either model alone. A linear model tree is simply a decision tree with linear models at its nodes. This can be seen as a piecewise linear model with knots learned via a decision tree algorithm. LMTs can be used for regression problems (e.g., with linear regression models instead of population means) or classification problems (e.g., with logistic regression instead of population modes).

- **Random forest regression model**

  Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned.[1][2] Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

7.Models Performance on Test set:

Model	**R2 score**
Linear regression	0.587017
Lasso regression	0.587702
Decision tree regressor	0.783501
Decision tree GridsearchCV	0.801389
Random forest regressor	0.836237

8.Challenges faced:

- Pre-processing the data was one of the challenges we faced which includes removing highly correlated variables from the data so as to not hinder the performance of our regression model.
- Exploring all the columns and calculating VIF for multicollinearity was challenging because it might decrease the model’s performance.
- Selecting the appropriate models to maximize the accuracy of our predictions was one of the challenges faced.

9.Libraries used:

Numpy: NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, Fourier transform, and matrices

Pandas: Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

Matplotlib: Matplotlib is a cross-platform, data visualization and graphical plotting library for Python

Seaborn: It is used for data visualization and exploratory data analysis. Seaborn works easily with data frames and the Pandas library. The graphs created can also be customized easily.

Scikit-learn is a library in Python that provides many unsupervised and supervised learning algorithms. Scikit-learn is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modelling including classification, regression, clustering and dimensionality reduction

10.Conclusion:

We performed the EDA and understood the features which affect most to bike sharing demand. Further we converted our data into a machine-readable format so can predict the count of rented bikes required at a particular hour throughout the day. We cleaned our data and fit the model.

We concluded the following points:

- Maximum numbers of bike are rented during the year 2018.
- Demand of the rented bike is maximum during functioning day and no holidays.
- Summer seasons is the season of highest rented bikes whereas winter has least number of bikes rented.
- Peak time for renting bikes is between 7 to 9 am in the morning and 5 to 7 pm in the evening. This could be due to the fact that maximum bikes are rented by office going people.
- The model that performed best on the given dataset is Random Forest with an r2 score of 0.83.
- Linear regression performed the worst out of all the models with an r2 score of just 0.58.
- Decision tree is also performed really good with an r2 score of 0.80