ASSIGNMENT BASED SUBJECTIVE QUESTIONS

1) **. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Ans :From the boxplot, it can be observed from the categorical variable like season that the demand of bike is in summer and fall season and the month which are June, July and August are also in high demand for bikes.

2) **Why is it important to use drop_first=True during dummy variable creation?**

Ans : There is a need to do drop_first = True during dummy variablecreation because to prevent the multi-colinearity between the target and independent variables and it is pretty cyclic p[rocess.

3) **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

Ans : Recalling the pair plot of numerical variables, it can be observed that **Temperature** variable is highly correlated with the target variable.

4) **How did you validate the assumptions of Linear Regression after building the model on the training set ?**

Ans: We can validate the assumptions of linear regression after building the model on the training set by (1)Linearity between the independent and dependent variable (2) Independence of errors (3) Homoscadaticity assumption (4) Normality of Residual errors (5) Multi-colinearity (6) Outliers and Inferential points.

5) **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes ?**

Ans: Based on P-value method, The top 3 P-value variables are the most significant which are temp, season_suimmer and mnth_Nov. This all stats are only based on summary of P-value.

GENERAL SUBJECTIVE QUESTIONS

6) **Explain the linear regression algorithm in detail.**

Ans: Linear regression is a supervised machine learning method that is used by the Train Using AutoML tool and finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.

The following is an equation of a resulting linear regression :

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots$$

Where, y = Dependent variable, b0 = intercept point and x = Independent variable.

7) **Explain the Anscombe's quartet in detail.**

Ans: Anscombe 's quartet is used to illustrate the importatnce of exploratory data analysis and drawbacks of only depending on the summary statastics. It also emphasizes the importance of using data visualization to spot trends, outliers and other crucial details that might not be obvious from summary statastics only.

8) **What is Pearson's R?**

Ans: Pearsons 's R is also known as Pearson 's correlation coefficient. Pearson 's R is a measure of strength and direction between two variables and it also indicates the correlation between independent and dependent variables.

9) **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Ans: The two most discussed scaling methods are Normalization and Standardization. Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

10) **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

Ans: If there is a perfect correlation, then the VIF is infinity because VIF = 1/(1 – R Squared). It indicated that a perfect correlation between the dependent and independent variable or two variables.

11) **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

Ans: Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential. The main aim of Q-Q plot is to check the normality or wether a distribution is normal or not.