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 analysis of the categorical variables from the dataset it could be inferred the bike rental rates are likely to be higher in summer and the fall season and are more prominent in the months of September and October, more so in the days of Sat, Wed and Thurs and in the year of 2019.

Additionally we could discern that bike rental are higher on holidays.

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

Ans: In the context of creating dummy variables for categorical features, the parameter **drop_first=True** is used to avoid multicollinearity in regression models. By setting **drop_first=True**, we are instructing to drop the first level of each categorical variable, leaving only **n-1** columns. This eliminates the perfect multicollinearity issue because the dropped category becomes the reference category, and the information about it is captured implicitly in the intercept term of the regression model.

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

Ans: The temp variable has the highest correlation with the target variable.

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

Ans: Validated the assumptions of linear regression by checking the VIF, error distribution of residuals as well as homoscedacity and linear relationship between the dependent variable and a feature variable.

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

Ans: The top 3 features contributing significantly towards the demand of the shared bikes are the temperature, year and the season variables.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear Regression is an ML algorithm used for supervised learning. This method is used to model the relationship between a dependent variable and one or more independent variables. The goal of linear regression is to find the best-fitting linear relationship that minimizes the sum of squared

differences between the observed and predicted values of the dependent variable. The equation for a simple linear regression, with one independent variable, takes the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{n-1} x_{n-1} + \beta_n x_n$$

There are two types of linear regression- simple linear regression and multiple linear regression.

a. Simple linear regression is used when a single independent variable is used to predict the value of the target variable.
b. Multiple Linear Regression is when multiple independent variables are used to predict the numerical value of the target variable.

A linear line showing the relationship between the dependent and independent variables is called a regression line. A positive linear relationship is when the dependent variable on the Y-axis along with the independent variable in the X-axis. However, if dependent variables value decreases with increase in independent variable value increase in X-axis, it is a negative linear relationship.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet consists of four data sets that have nearly identical simple descriptive statistics but have very different distributions and appear very different when presented graphically. Each dataset consists of eleven points. The primary purpose of Anscombe's quartet is to illustrate the importance of looking at a set of data graphically before beginning the analysis process as the statistics merely does not give the an accurate representation of two datasets being compared.

3. What is Pearson's R?

Ans: Pearson's Correlation Coefficient is used to establish a linear relationship between two quantities. It gives an indication of the measure of strength between two variables and the value of the coefficient can be between -1 and +1.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Scaling is the process of transforming numerical variables to a standardized range. It is performed to ensure that variables with different scales do not disproportionately influence machine learning models. Normalized scaling (min-max scaling) transforms data to a specific range, typically [0, 1]. Standardized scaling (z-score normalization) scales data to have a mean of 0 and a standard deviation of 1. While normalized scaling maintains the original distribution and is suitable for algorithms sensitive to variable ranges, standardized scaling is robust against outliers and suitable for algorithms that assume normally distributed data. Both methods enhance model performance and interpretability.

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

Ans: The value of VIF is infinite when there is a perfect correlation between the two independent variables. The Rsquared value is 1 in this case. This leads to VIF infinity as VIF equals to 1/(1-R2). This concept suggests that is there is a problem of multi-collinearity and one of these variables need to be dropped in order to define a working model for regression.

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

Ans: The quantile-quantile (Q-Q) plot are used to plot quantiles of a sample distribution with a theoretical distribution to determine if any dataset concerned follows any distribution such as normal, uniform or exponential distribution. It helps us determine if two datasets follow the same kind of distribution. It also helps to find out if the errors in dataset are normal in nature or not.