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? (3 marks)
ans:
a) There are no missing / Null values either in columns or rows
b) All variable are multicollinear in nature and have high collinearity with target variable.
c) Combiation of VIF, correlation value and P-value will be the best option to understand the statstical importance and improve the models by dropping columns inorder to improve adjusted r^2.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
Ans: To avoid multicollinearity issues and to ensure the model's interpretability and reduces redundancy.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Ans : Registered Column shows highest correlation

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
Ans: a) Residual Analysis
b) Multicollinearity
c) Linearity of Residuals
d) Outlier treatment

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
Ans: a) Temparature b) Summer Season c) Winter Season.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
Linear regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables and aims to find the best-fitting straight line that predicts the dependent variable based on the independent variables. This is achieved by estimating the slope and intercept of the line that minimizes the sum of squared differences between observed and predicted values. The algorithm calculates these coefficients using techniques like ordinary least squares or gradient descent. Once determined, the model can make predictions by plugging in values of the independent variables into the linear equation.
y=mx+c

2. Explain the Anscombe's quartet in detail (3 marks)
Anscombe's quartet is a set of four datasets with nearly identical statistical properties, designed to illustrate the importance of visualizing data. Despite their similar summary statistics, the datasets exhibit vastly different relationships when plotted. This highlights the danger of relying solely on summary statistics, as they

may mask underlying patterns or anomalies in the data. Anscombe's quartet emphasizes the necessity of data visualization in understanding and interpreting datasets accurately, as visual inspection can reveal insights that summary statistics alone might overlook, thereby aiding in more informed decision-making and analysis.

3. What is Pearson’s R? (3 marks)
Pearson's correlation coefficient (denoted as Pearson's R) is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to +1, where +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. Pearson's R is calculated by dividing the covariance of the two variables by the product of their standard deviations.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Scaling is the process of transforming data to a common scale, often within a specific range. It's performed to ensure that variables with different units or magnitudes contribute equally to analyses, preventing biases. Normalized scaling rescales data to a range between 0 and 1, preserving the original distribution. Standardized scaling, on the other hand, transforms data to have a mean of 0 and a standard deviation of 1, facilitating comparisons between variables with different units. While normalized scaling maintains original ranges, standardized scaling centers data around a common point, enabling easier interpretation of the relative importance of variables in statistical analyses.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
Variance Inflation Factor (VIF) measures multicollinearity among predictor variables in regression analysis. VIF becomes infinite when one or more variables can be perfectly predicted by a linear combination of other variables, resulting in a perfect collinearity issue. This situation, known as multicollinearity, occurs when variables are highly correlated, making it impossible to estimate their unique effects accurately. Infinite VIF implies that the standard errors of regression coefficients are undefined due to perfect multicollinearity, making it challenging to interpret the impact of individual predictors accurately in the regression model.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
A Q-Q plot is a graphical tool used to assess whether a dataset follows a certain probability distribution, such as the normal distribution. It compares the quantiles of the dataset to those of a theoretical distribution. In linear regression, Q-Q plots help validate the assumption of normality of residuals, which is crucial for accurate inference and prediction. If the residuals deviate significantly from the diagonal line in the Q-Q plot, it suggests departures from normality, indicating potential issues with the model's assumptions. Correcting these deviations improves the reliability and validity of regression analyses.