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 : We can infer that the Boom bike's rental rates can benefit based on :-

* season_spring
* month_september
* temp
* weathersit_light snow
* year_2019

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

Ans : We use drop_first=True because this helps in reducing the extra column when we were creating dummy variables.

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

Ans: Highest correlation noticed is with temp (ie; temperature)

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: By checking the VIF, error distribution of residuals and lr (linear regression)between feature & dependent variables

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: Top 3 are :-

* Year
* Season
* Temperature

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans: Linear regression is a powerful statistical method for modelling the relationship between variables, which is done by fitting a straight line to the data, it allows for prediction and interpretation of the relationship between the dependent and independent variables. It is essential to assess the assumptions and evaluate the performance of the model to ensure its validity and reliability. Linear regression algorithm can be explained in steps :-

* Assumption of Linearity
* Objective
* Ordinary Least Squares (OLS)
* Estimation of Coefficients
* Interpretation of Coefficients
* Model Evaluation & Performance
* Assumption
* Conclusion

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans: Anscombe's quartet is a set of four datasets which were designed to have nearly identical summary statistics, such as mean, variance, correlation, and regression coefficients, yet they exhibit vastly different patterns when plotted graphically. The purpose of Anscombe's quartet is to illustrate the importance of visualising data and the limitations of relying solely on summary statistics in statistical analysis.

3. What is Pearson's R? (3 marks)

Ans: It is a statistical measure that quantifies the strength and direction of the linear relationship between two variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardised scaling? (3 marks)

Ans: It is a preprocessing step in data analysis and machine learning where the values of features (variables) are transformed to fit within a specific range or distribution. The goal of scaling is to standardise the range of features in the dataset to ensure that no single feature dominates the analysis simply because of its larger scale.

Scaling is performed for reasons:

* Equal Weightage
* Convergence
* Regularisation

The main difference between normalised scaling and standardised scaling lies in the range of values they produce and their sensitivity to outliers.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans: When the correlation between predictor variables is perfect (i.e., they are perfectly collinear), it leads to multicollinearity issues, and the VIF calculation may result in infinite values. This happens because the formula for VIF involves dividing by zero or extremely small values, which leads to infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans: It is a graphical tool used to compare the distribution of a sample to a known probability distribution, typically the normal distribution. It is a graphical method to assess whether a given set of data follows a certain theoretical distribution. Q-Q stands for quantile-quantile plot.