Assignment-Based Subjective Questions

1. Q: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer: Through a thorough examination of categorical variables, one can draw insights into their influence on the dependent variable. Observing patterns and variations among different categories helps discern how each group contributes to the overall variation in the dependent variable.

2. Q: Why is it important to use drop_first=True during dummy variable creation? (2 marks)

Answer:It is crucial to incorporate `drop_first=True` in dummy variable creation to prevent multicollinearity issues. By excluding one dummy column, it mitigates the dummy variable trap, ensuring that the model remains interpretable and stable.

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

Answer:By carefully examining the pair-plot among numerical variables, the goal is to identify the variable exhibiting the strongest linear correlation with the target variable.

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

Answer: The validation process involves a comprehensive check of assumptions such as linearity, independence, homoscedasticity, and normality of residuals. Diagnostic plots, thorough residual analysis, and statistical tests are employed to ensure that the residuals are normally distributed and exhibit constant variance across predicted values.

5.Q: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer: In determining the top 3 features, emphasis is placed on coefficients or importance scores in the final model. The evaluation considers both statistical significance and practical relevance to identify features significantly contributing to the demand for shared bikes.

General Subjective Questions:

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

Answer:The idea of linear regression model is to come up with a best fit line for the data points provided. The main parts of linear regression involve 1. a set of independent variabes 2. a dependent variable Y whose value needs to be determined while predicting with the help of new dataset The linear regression technique uses Ordinal least Squares method, however gradient descent can also be used as a cost function The linear regression works well when the problem is linear in nature .if a curve can be drawn then it becomes polynomial regression different types of equation such quadratic and cubic and much more can be applied The main assumptions of linear Regression are as follows as per definition Linearity: The relationship between independent and dependent variable is linear. Homoscedasticity: The variance of residual is constant for any value of the independent variable. Independence: Each Observations are independent of each other. Normality: the residuals are normally distributed. With mean 0

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

Answer: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven x and y points. It demonstrates both the importance of graphing data before analyzing it and the effect of outliers and other observations

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

Answer:Pearsons coefficient of correlation – is a measure which tries to explain relationship between 2 variables. It ranges between -1 to +1 0 -> no relationship between two variable Between 0 to -1 negative correlation increase in a variable lead to decrease in other variable Between 0 to +1 positive correlation increase in one variable also leads to increase in another variable

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

Answer:Scaling is technique to standardize the data often the ml algorithms consider large values to be very large and smaller value as very small without considering the unit leading to misinterpretation Normalized scaling uses min and max values to

represent the range of values between 0 and 1 Standardized scaling uses z score which is based on mean and standard diviation Normalized scaling is used when distribution is not known Standardized scaling is used data follows gausian or normalized distribution

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

Answer: Infinite VIF values occur due to perfect multicollinearity, where one variable is a perfect linear combination of others. This situation arises when it becomes impossible to distinguish the individual impact of correlated variables.VIF is used to solve multicollinearity issue when R =1 strong correlation VIF tends to infinity

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

Answer: A Q-Q (Quantile-Quantile) plot compares the quantiles of observed residuals with those of a theoretical distribution. It serves to assess the normality of residuals in linear regression. Points close to the diagonal line indicate normality, while deviations provide insights into potential non-normality, crucial for understanding model assumptions.