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?**

   The following variables have significant effect on target variable “cnt”
   - “temp”
   - “hum”
   - “windspeed”
2. **Why is it important to use drop_first=True during dummy variable creation?**

   It is particularly important to avoid multicollinearity issues in regression analysis. Dropping one level of the categorical variable also helps in interpreting the intercept term in the regression model.
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

   - “registered”

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

   The assumptions of Linear Regression after building the model on the training set were validated through the following techniques
   - Residual analysis
   - Normality of residuals
   - Homoscedasticity
   - Independence of residuals
   - Outlier detection
   - Cross-validation

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

1- “temp”
2- “hum”
3- “windspeed”

General Subjective Questions

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

Linear regression relies on several assumptions, including linearity, independence of errors, constant variance of errors (homoscedasticity), and normality of errors. Linear regression is a simple yet powerful algorithm for modelling the relationship between dependent and independent variables. It is widely used in various fields, including economics, finance, healthcare, and social sciences, for prediction, inference, and understanding relationships between variables.

Broadly there will 5 steps.

1) Model formulation: In this step the linear regression model will be represented using intercept, coefficients and error term. The target variable (y) will be a function of independent variables (X1, X2 etc.), intercept and error term.
2) Model fitting: In this step the goal is to find the values of coefficients that minimize the difference between the actual and predicted values.
3) Least squares estimation: Minimizes the sum of the squared differences between the observed and predicted values.
4) Coefficient estimation: The coefficients are estimated using various optimization techniques such as ordinary least squares, gradient descent, or matrix methods. The intercept represents the value of the dependent variable when all independent variables are zero, and the other coefficients represent the change in the dependent variable for a one-unit change in the corresponding independent variable, holding other variables constant.
5) Model evaluation: Once the model is fitted, it is essential to evaluate its performance. Common metrics for evaluating linear

regression models include the coefficient of determination (R-squared), mean squared error (MSE), root mean squared error (RMSE), and others. These metrics help assess how well the model fits the data and how accurately it predicts the dependent variable.

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

The quartet consists of 4 data sets. Each data set contains 11 data pairs. When plotted they will have different appearances. However, all 4 data sets will exhibit similar summary statistics. From this it's apparent that the summary statistics alone will not the reveal the trends and inferences of the data set. It must be analysed in conjunction with the plots and figures.

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

Pearson's *r* is a correlation coefficient. It is a measure of linear relationship between 2 variables. It's range is -1 to +1. When it equals to "+1" it means a perfect positive linear relationship between variables. It means they are directly proportional. When it is equal to "-1", the variables are indirectly proportional to one another. When it's ZERO, no relationship between them.

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

Scaling means transforming the data into a similar scale. It is performed to improve the interpretation of machine learning models.

Normalized scaling rescales features to a fixed range between 0 and 1, while standardized scaling transforms features to have a mean of 0 and a standard deviation of 1.

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

VIF stands for variance inflation factor. It measures the multicollinearity in regression analysis. Higher the VIF, higher the quantity of variance of regression coefficient attributable to multicollinearity. SO if this VIF becomes infinite this indicates there is severe multicollinearity among the predictor variables.

This happens when there is perfect collinearity among the predictor variables.

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

Q-Q plot is a diagnostic tool in linear regression analysis, helping to assess the normality assumption of the residuals. Q-Q stands for Quantile-Quantile plot. It compares 2 or more probability distributions by plotting their quantiles against each other. Quantiles are division of probability distribution into equally probable intervals. For example, 25% percentile. It's a good fit if the data points are on the Q-Q diagonal line. If not on the 45 degree line, then it's not a good fit with probability distribution. These indicate skewness or heavy tails.