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: From the box plots and data visualization done between the dependent variable and categorical variables, we have clear indication with the season Fall and Summer has high demand comparatively. Likewise, Year 2019 has more demand than 2018 which says the demand in coming years may increase. Holiday and weekday rental usage has no clear indication, but holiday rental usage spread is more than the workingday. Again, weathersit has direct effect on target variable.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
   Ans: It is to avoid dummy variable trap by dropping the first category and instead of creating dummies for each category it will create one less (n-1) dummies. This will reduce the multicollinearity in regression models.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   Ans: registered 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? (3 marks)
   Ans: After building the model, the predictions are run on training data first to see how well the train data is learned by the model and how well it can predict. Once predicted, graph is plotted between the y_pred and y_actual training data and the error terms are normally distributed mean centered at 0. Also, validated VIF to check the multicollinearity of the predictors with each other. Linear plot between y_test and y_pred.

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: Temperature, Year, Spring (Season) are the top 3 features mainly explaining the demand for shared bikes. We can know this by extracting the highest coefficients from the stats model, we can understand what are the top features effecting the target variable.

General Subjective Questions

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

Ans: Linear regression is a fundamental algorithm in machine learning and statistics used to model the relationship between a dependent variable and one or more independent variables. The goal is to find the best-fitting straight line (or hyperplane in higher dimensions) that describes the relationship between the variables.

Here's a detailed explanation of the linear regression algorithm:

1. Model Representation: The linear regression model is represented by the equation: [y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon] where (y) is the dependent variable, (x_1, x_2, \ldots, x_n) are the independent variables, (\beta_0) is the

intercept, (\beta_1, \beta_2, \ldots, \beta_n) are the coefficients, and (\epsilon) is the error term.

2. Assumptions: Linear regression makes several key assumptions:
   - Linearity: The relationship between the dependent and independent variables is linear.
   - Independence: The residuals (errors) are independent.
   - Homoscedasticity: The residuals have constant variance.
   - Normality: The residuals are normally distributed.
3. Estimation of Coefficients: The coefficients (\beta_0, \beta_1, \ldots, \beta_n) are estimated using the method of least squares, which minimizes the sum of the squared differences between the observed and predicted values of the dependent variable. Mathematically, this is represented as: [\min_{\beta} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2] where (y_i) is the observed value, (\hat{y}_i) is the predicted value, and (m) is the number of observations.
4. Fitting the Model: Once the coefficients are estimated, the model can be used to make predictions. The fitted model is represented by the equation: [\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n]
5. Model Evaluation: The performance of the linear regression model is evaluated using metrics such as R-squared, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics help assess how well the model fits the data and how accurately it predicts the dependent variable.
6. Validation: To ensure the model's reliability, it is important to validate the assumptions of linear regression. This can be done by checking the residual plots, performing statistical tests, and using cross-validation techniques.

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

   Ans: Anscombe's quartet consists of four datasets that have the same mean, variance, and correlation, but look very different when plotted. Each dataset consists of eleven (x, y) points. The quartet was created to illustrate the importance of data visualization and how relying solely on summary statistics can be misleading.

   Here are the key points about each dataset in Anscombe's quartet:
   1. Dataset I: Fits the linear regression model well, showing a simple linear relationship.
   2. Dataset II: Shows a non-linear relationship, making the linear regression model inappropriate.
   3. Dataset III: Contains an outlier that affects the regression line, demonstrating the influence of outliers.
   4. Dataset IV: Has a high-leverage point that creates a high correlation coefficient, even though the other data points do not indicate any relationship

3. What is Pearson’s R? (3 marks)

   Ans: Its Pearson correlation coefficient, measure between 2 variables. Its between -1 and 1 that measures the strength and direction of relationship between 2 variables. Value > 0 is positive association and negative association

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

and standardized scaling? (3 marks)

Ans: Scaling is a technique done to bring the whole dataset to one scale.

Normalized scaling: This scales the data to a fixed range usually between 0 and 1, Its also known as MinMax scaling

Standardized scaling: This technique scales that data to have mean of 0 and standard deviation 1. Also known as Z-score normalization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Ans: When there is perfect collinearity between the independent variables. This happens when R-squared value in the VIF that it can be separately calculated to indicate the perfect collinearity.
   (3 marks)
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   (3 marks)

   Ans: A Q-Q plot, or quantile-quantile plot, is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, such as the normal distribution. It plots the quantiles of the dataset against the quantiles of the theoretical distribution. If the data follows the theoretical distribution, the points on the Q-Q plot will lie approximately along a straight line.

   In the context of linear regression, a Q-Q plot is used to check the normality assumption of the residuals. The normality of residuals is one of the key assumptions in linear regression, and it ensures that the error terms are normally distributed. By plotting the residuals on a Q-Q plot, you can visually assess whether they follow a normal distribution. If the points on the Q-Q plot lie along a straight line, it indicates that the residuals are normally distributed, and the normality assumption is met. If the points deviate significantly from the line, it suggests that the residuals are not normally distributed, and the normality assumption may be violated