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:
   - Spring has the lowest bike rental ratio than other seasons.
   - In snow, people avoid traveling by bike, therefore, renting bikes is low.
   - Year 2019 has higher rentals than 2018.
   - There are no major differences in Weekend and working day bike rental count in any season and weather situation.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
   Ans: For example, if we have 3 unique variables in one categorical variable. So, without dropping the first variable, we convert it into a dummy variable, generating 3 different results as 00,01,10.

   Here if we drop the first variable it will generate 01 and 10 and we can still identify the third variable as 00. So, adding unnecessary columns also affects the model so it is important to use drop_first= True.

3. Looking at the pair plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   Ans: "atemp" 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: Will find the difference between y_actual and y_predicted and then draw a histogram of error terms and check if the residuals are following the normally distributed with a mean 0.

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: The main 3 features which are contributing are weathersit, season, windspeed.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   Ans: Linear regression is a widely used statistical algorithm used to model the relationship between a dependent variable and one or more independent variables. It aims to find the best-fit straight line that represents the linear relationship between the variables.

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

Data Preparation: Gather a dataset consisting of observations of the dependent variable (often denoted as 'y') and one or more independent variables (often denoted as 'x'). Ensure the data is cleaned, properly formatted, and any missing values are handled appropriately.

Model Representation: In linear regression, we assume a linear relationship between the independent variables and the dependent variable. For a simple linear regression with one independent variable, the equation is represented as: $y = mx + b$, where 'm' is the slope (coefficient) of the line, 'x' is the independent variable, and 'b' is the y-intercept.

Cost Function: The goal of linear regression is to minimize the difference between the predicted values and the actual values of the dependent variable. A common cost function used is the Mean Squared Error (MSE), which calculates the average squared difference between the predicted and actual values.

Gradient Descent: Gradient descent is an iterative optimization algorithm used to minimize the cost function. It starts with random values for the coefficients and iteratively adjusts them to reach the minimum cost. In each iteration, the algorithm calculates the gradient of the cost function with respect to the coefficients and updates the coefficients in the direction that minimizes the cost. This process continues until convergence is achieved.

Training the Model: To train the linear regression model, the dataset is fed into the algorithm. During training, the algorithm adjusts the coefficients using gradient descent to find the values that minimize the cost function.

Evaluation: Once the model is trained, it can be evaluated using various evaluation metrics such as the coefficient of determination (R-squared), Mean Absolute Error (MAE), or Root Mean Squared Error (RMSE) to assess its performance and accuracy in predicting the dependent variable.

Prediction: After the model is trained and evaluated, it can be used to make predictions on new, unseen data. By inputting values of the independent variables into the trained model, it calculates the predicted values of the dependent variable.

Linear regression is a versatile algorithm that can handle both simple linear relationships with one independent variable and multiple linear relationships with multiple independent variables.

It provides insights into the strength and direction of the relationships between variables and allows for predictions and inference based on the learned coefficients.

2. Explain Anscombe's quartet in detail. (3 marks)
   Ans: Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

   Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

3. What is Pearson's R? (3 marks)
   Ans: The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

   Although interpretations of the relationship strength (also known as effect size) vary between disciplines,

   The Pearson's correlation coefficient varies between -1 and +1 where:

   r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

   r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

   r = 0 means there is no linear association

   r > 0 <0. 5 means there is a weak association

   r >0. 5 <0. 8 means there is a moderate association

   r > 0.8 means there is a strong association

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

Ans: Scaling, in the context of machine learning and data preprocessing, refers to the transformation of features or variables to a specific range or distribution. It is performed to ensure that all variables are on a similar scale and have comparable importance during model training.

Scaling is necessary for several reasons:

- Avoiding Variable Bias
- Enhancing Model Performance
- Improving Convergence

Normalized scaling brings the values within a specified range, while standardized scaling standardizes the values to have a mean of 0 and a standard deviation of 1.

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

Ans: This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which leads to 1/(1-R2) infinity.

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

Ans: The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not.
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value.
For the reference purpose, a 45% line is also plotted, if the samples are from the same population, then the points are along this line.

Usage:
The Quantile-Quantile plot is used for the following purpose:

- Determine whether two samples are from the same population.
- Whether two samples have the same tail
- Whether two samples have the same distribution shape.
- Whether two samples have common location behavior.

How to Draw Q-Q plot

- Collect the data for plotting the quantile-quantile plot.
- Sort the data in ascending or descending order.
- Draw a normal distribution curve.
- Find the z-value (cut-off point) for each segment.
- Plot the dataset values against the normalizing cut-off points.

Advantages of Q-Q plot

- Since Q-Q plot is like a probability plot. So, while comparing two datasets the sample size need not to be equal.
- Since we need to normalize the dataset, so we don’t need to care about the dimensions of values.