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 final model we can see that the output variable 'cnt' is dependent on the following categorical variables 'yr', 'holiday', 'spring', 'winter' and the months such as dec, jan,july, sep, nov etc. So the effect of categorical variable on the dependent variable 'cnt' is very significant.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
   Ans: During dummy variables creation the first dummy variable can be inferred by the value of other variables. When the value of the rest of the variables are zero indicates the occurrence of the first variable.

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: Considering the P value and the VIF value. The model should not have any independent variable with a high p value and a VIF value more than 5 which indicates that the variable is insignificant and interdependent.

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: 'temp','yr','winter'

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   Ans: Linear regression is a popular algorithm used in statistical modeling and machine learning to establish a relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables and aims to find the best-fit line that minimizes the differences between the observed and predicted values. In this explanation, we'll focus on simple linear regression, which involves a single independent variable.
   Here's a step-by-step breakdown of the linear regression algorithm:
   1. Data Collection: Gather a dataset consisting of pairs of observations of the dependent variable (usually denoted as "y") and the independent variable (typically denoted as "x"). The dataset should have a sufficient number of samples to capture the relationship between the variables.
   2. Data Preprocessing: Perform any necessary preprocessing steps, such as handling missing values, handling outliers, scaling the data, or normalizing the variables. These steps ensure the data is suitable for analysis and prevents biases in the model.
   3. Model Representation: Linear regression assumes a linear relationship between the dependent variable and the independent variable. The relationship is represented as a linear equation:
      $y=b0+b1*x$

where:

- y represents the dependent variable.
- x represents the independent variable.
- b0 is the y-intercept (the value of y when x is 0).
- b1 is the slope of the line (the change in y corresponding to a unit change in x).

The goal of linear regression is to estimate the values of b0 and b1 that best fit the data.

1. Model Training: The training phase involves finding the optimal values of b0 and b1 that minimize the difference between the observed values of y and the predicted values from the linear equation. This is typically achieved using a technique called ordinary least squares (OLS) or other optimization algorithms. The OLS method minimizes the sum of squared errors (SSE) between the predicted and observed values.
2. Model Evaluation: Once the model is trained, it needs to be evaluated to assess its performance. Common evaluation metrics for linear regression include the coefficient of determination (R-squared), root mean squared error (RMSE), mean absolute error (MAE), and others. These metrics help measure how well the model fits the data and the accuracy of its predictions.
3. Prediction: After the model has been evaluated, it can be used for making predictions on new, unseen data. Given a new value of x, the model can estimate the corresponding value of y using the learned coefficients b0 and b1.
4. Interpretation: Linear regression also provides valuable insights into the relationship between the variables. The sign and magnitude of the coefficients b0 and b1 indicate the direction and strength of the relationship. A positive b1 suggests a positive relationship between x and y, whereas a negative b1 indicates an inverse relationship. The magnitude of b1 represents the rate of change in y for a unit change in x.

Linear regression is a simple yet powerful algorithm that is widely used for various applications including prediction, forecasting, and analysis.

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

Ans: Anscombe's quartet is a famous example in statistics that highlights the importance of visualizing data and not relying solely on summary statistics. It consists of four datasets that have nearly identical statistical properties, including means, variances, correlations, and regression lines, but are fundamentally different when graphically visualized.

The four datasets in Anscombe's quartet were created by the statistician Francis Anscombe in 1973.

The main purpose of Anscombe's quartet is to emphasize the importance of exploring and visualizing data before drawing conclusions based solely on summary statistics. It demonstrates that datasets with similar statistical properties can have vastly different patterns when plotted, leading to different interpretations and insights. Visualizing the data allows researchers to uncover nuances, identify outliers, and make more informed decisions about which statistical methods are appropriate for analysis.

Anscombe's quartet serves as a cautionary reminder that summary statistics alone may not provide a complete understanding of the data and that data visualization is a vital tool in data analysis.

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

Ans: Pearson's correlation coefficient, often denoted as Pearson's R or simply as R, is a

statistical measure that quantifies the strength and direction of the linear relationship between two variables. It was developed by Karl Pearson and is widely used to assess the degree of association between continuous variables.

Pearson's R ranges from -1 to 1, where:

- A value of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other decreases linearly.
- A value of 1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other also increases linearly.
- A value of 0 indicates no linear relationship between the variables.

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 data preprocessing, refers to the transformation of variables to a specific range or distribution. It is performed to bring different variables onto a comparable scale, eliminate bias, and enhance the performance of various data analysis and machine learning algorithms.

Scaling is necessary for several reasons:

1. Comparison: Variables may have different units or measurement scales, making direct comparisons difficult. Scaling enables meaningful comparisons by bringing variables to a common scale.
2. Avoidance of Biases: Some algorithms, such as distance-based methods (e.g., k-means clustering, k-nearest neighbors), are sensitive to the scale of variables. Variables with larger scales can dominate the algorithm's calculations and introduce biases. Scaling ensures that all variables contribute equally.
3. Gradient Descent Convergence: In optimization algorithms like gradient descent, scaling can help improve convergence rates by ensuring that variables have similar magnitudes. It helps prevent oscillations and speed up the learning process.

Two commonly used scaling techniques are normalized scaling and standardized scaling:

1. Normalized Scaling (Min-Max Scaling): Normalization rescales variables to a specific range, typically between 0 and 1. The formula for normalized scaling is:

$X' = (X - X\_min) / (X\_max - X\_min)$

where X' is the scaled value, X is the original value, X_min is the minimum value in the dataset, and X_max is the maximum value. This scaling technique retains the relative relationships between the data points and preserves the distribution shape.

2. Standardized Scaling (Z-score Scaling): Standardization transforms variables to have a mean of 0 and a standard deviation of 1. The formula for standardized scaling is:

$X' = (X - \mu) / \sigma$

where X' is the scaled value, X is the original value, μ is the mean of the variable, and σ is the standard deviation. Standardization centers the data around zero and scales it by the standard deviation. It maintains the shape of the distribution and is suitable when the data is expected to have outliers.

The main difference between normalized scaling and standardized scaling lies in the range and interpretation of the scaled values. Normalization scales variables to a specific range (e.g., 0 to 1), while standardization rescales variables with a mean of 0 and a standard deviation of 1. Normalized values are bounded, preserving the original data range, while standardized values have no fixed range and can have negative and positive values.

The choice between normalized scaling and standardized scaling depends on the specific requirements of the problem at hand. Normalization is useful when the range of variables is significant, and preserving the original range is important. Standardization is beneficial when variables have different means and variances, and ensuring zero mean and unit standard

deviation is desired.

Both scaling techniques play a crucial role in data preprocessing and enable more effective and reliable data analysis and modeling.

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

Ans: The VIF (Variance Inflation Factor) is a measure used to assess multicollinearity in regression analysis. Multicollinearity occurs when independent variables in a regression model are highly correlated with each other, which can cause issues in the interpretation and stability of the regression coefficients. The VIF quantifies the degree of multicollinearity by examining how much the variance of the estimated regression coefficient is inflated due to correlation with other variables.

In general, a VIF value greater than 1 indicates some level of multicollinearity, with higher values indicating stronger correlation. However, it is possible for the VIF to be infinite for a specific variable. This happens when one or more independent variables in the regression model are perfectly linearly dependent on a combination of the other variables. In other words, one or more variables can be expressed as an exact linear combination of the other variables in the model.

When this perfect multicollinearity occurs, it creates redundancy in the model, and it becomes impossible to estimate the regression coefficients accurately. In such cases, the VIF for the variable that is linearly dependent on others becomes infinite because the estimated variance of its regression coefficient is undefined due to the linear dependency. The infinite VIF indicates that the variable's contribution to the model is completely explained by other variables and provides no additional information.

Perfect multicollinearity can arise due to various reasons, such as including redundant variables, using derived variables that are linear combinations of other variables, or including variables that are too similar in their information content.

To address the issue of perfect multicollinearity, it is necessary to identify and remove the linearly dependent variables from the regression model. This can be done by examining the correlation matrix or performing techniques like backward elimination, ridge regression, or principal component analysis (PCA) to reduce the dimensionality of the data.

In summary, the VIF can be infinite when one or more variables in a regression model are perfectly linearly dependent on a combination of other variables. Detecting and resolving perfect multicollinearity is crucial to ensure the stability and interpretability of the regression model.

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

Ans: Q-Q plot, short for Quantile-Quantile plot, is a graphical tool used to assess the distributional similarity between a dataset and a theoretical distribution. It compares the quantiles of the dataset against the quantiles expected from a specified distribution, typically a normal distribution. Q-Q plots are commonly used in statistics and data analysis to evaluate the assumption of normality, which is often required by linear regression and other statistical methods.