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?**

   Some of the inferences on the analysis and their affect on the dependant variable –

   a) The season of Fall has the highest median followed by summer as they have st weather conditions.
   b) The months of Fall – June to October have a higher median value
   c) The overall median for the weekdays and working days are the same.

2. **Why is it important to use drop_first=True during dummy variable creation?**

   It's important to use drop_first = true as it helps in reducing the extra column created during dummy variable creation.It helps to reduce the correlations created among dummy variables.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

   The numerical variable 'atemp' has the highest correlation with the target variable 'cnt' with a value of '0.65' followed by 'temp' with a value of '0.64'.

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

   We validate the assumptions of the Linear Regression by plotting a distplot of the residuals and analysing it to see if it is a normal distribution or not and if it has 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?**

   Top3 features contributing significantly towards explaining the demands of the shared bikes –

   a) Temp – A coefficient value of '0.5082' indicates that a unit increase in temp variable ,increases the bike hire numbers by '0.5082' units.
   b) Weathersit – A coefficient value of '-0.1572' indicates that a unit increase of this variable, decreases the bike hire numbers by '-0. 1572' units.
   c) Yr – A coefficient value of '0.2294' inidicates that a unit increase of this variable, increase the bike hire numbers by '0.2294' units.

General Subjective Questions

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

Linear regression is a widely used supervised machine learning algorithm for modeling the relationship between a dependent variable (target) and one or more independent variables (features or predictors). It assumes that the relationship between the variables is linear, meaning that changes in the target variable are proportional to changes in the independent variables. In this explanation, I'll describe simple linear regression, which involves a single independent variable. Multiple linear regression extends these concepts to multiple independent variables.

The linear regression model is represented by the equation:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

y: The dependent variable or target you want to predict.
x: The independent variable (feature) used for prediction.
β_0 (beta-zero): The intercept, representing the value of y when x is 0.
β_1 (beta-one): The slope or coefficient, representing how y changes when x changes.
ε (epsilon): The error term or residual, representing the difference between the predicted and actual values. It accounts for unexplained variation in y.

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

Anscombe's quartet is a set of four small datasets that have nearly identical simple descriptive statistics but exhibit vastly different distributions and appear quite distinct when graphically visualized.

Anscombe's quartet consists of four separate datasets, each containing 11 data points (pairs of x and y values). These datasets are named I, II, III, and IV.

Despite the different distributions and relationships in the four datasets, their basic summary statistics are nearly identical.

The true value of Anscombe's quartet lies in visualizing the data. When you plot these datasets, you can see how they are very different despite having similar summary statistics. Here's what makes each dataset unique:

Dataset I: Forms a clear linear relationship.
Dataset II: Forms a linear relationship but is influenced by an outlier.
Dataset III: Clearly demonstrates a nonlinear relationship.
Dataset IV: Contains an extreme outlier that skews the regression line.

Anscombe's quartet serves as a powerful reminder of the importance of data visualization in exploratory data analysis. It highlights that relying solely on summary statistics can lead to misleading conclusions about the data. Even when summary statistics appear consistent, the underlying patterns and relationships within the data may be vastly different.

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

Pearson's R, also known as the Pearson correlation coefficient or Pearson's correlation, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is named after its developer, Karl Pearson, and is widely used in statistics to assess the degree of association between two sets of data.

The Pearson correlation coefficient, denoted as "r," ranges from -1 to 1:

If r = 1, it indicates a perfect positive linear relationship, meaning that as one variable increases, the other also increases proportionally.

If r = -1, it indicates a perfect negative linear relationship, meaning that as one variable increases, the other decreases proportionally.

If r = 0, it indicates no linear relationship between the two variables.

$r=\sum(X_i-\bar{X})2\sum(Y_i-\bar{Y})2/ \text{ sqroot}(\sum(X_i-\bar{X})(Y_i-\bar{Y})$

where X_i and Yi are individual points from the datasets

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

Scaling is a data preprocessing technique used in machine learning and data analysis to transform numerical features or variables to a standard range or distribution. The primary goal of scaling is to ensure that all the features in a dataset have similar scales or magnitudes.

There are two common methods for scaling data: normalized scaling and standardized scaling, and they have different approaches:

Normalized Scaling (Min-Max Scaling): This method scales the features to a specific range, typically [0, 1]. The formula for min-max scaling is:

X_normalized = (X - X_min) / (X_max - X_min)

Standardized Scaling (Z-score Scaling): This method standardizes the features to have a mean of 0 and a standard deviation of 1. The formula for standardization is:

X_standardized = (X - X_mean) / X_std

In summary, both normalized scaling and standardized scaling are techniques used to bring features to a common scale. Normalized scaling constrains the data to a specific range, while standardized scaling centers the data around 0 with a standard deviation of 1. The choice between the two methods depends on the specific characteristics of the data and the requirements of the machine learning algorithm you are using.

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

The Variance Inflation Factor (VIF) is a measure used in regression analysis to assess multicollinearity among predictor variables. Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated with each other, making it difficult to discern the individual effects of each predictor on the target variable. VIF quantifies how much the variance of the estimated regression coefficients is increased due to multicollinearity.
A VIF value of infinity (or sometimes extremely large values) can occur when there is perfect multicollinearity in the model. Perfect multicollinearity means that one or more of the predictor variables can be exactly predicted by a linear combination of the other predictor variables. In other words, there is a perfect linear relationship among some of the predictors.

Here's why VIF can become infinite:

1) Mathematical Redundancy
2) Matrix Inversion Issues

To address the issue of infinite VIF values, you should examine your dataset and the variables included in your regression model. Here are some steps you can take:

Identify the Redundant Variables

Remove or Combine Variables

Reevaluate the Model

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

A Q-Q plot, short for Quantile-Quantile plot, is a graphical tool used in statistics to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It compares the quantiles of the observed data against the quantiles of the theoretical distribution being tested. The main use of a Q-Q plot is to visually check the assumption of normality in a dataset, but it can also be used to assess the fit of data to other probability distributions.

Importance in Linear Regression:

The Q-Q plot is important in linear regression for several reasons:

Assumption Checking: Linear regression models often assume that the residuals (the differences between observed and predicted values) are normally distributed. If this assumption is violated, it can affect the validity of regression results, such as p-values and confidence intervals. The Q-Q plot helps you assess whether this assumption holds.

Identifying Outliers: Outliers in the data can have a significant impact on regression results. A Q-Q plot can help identify outliers as data points that deviate significantly from the expected quantiles of the theoretical distribution.

Model Improvement: If the Q-Q plot indicates a departure from normality, it may suggest the need for data transformation or the use of robust regression techniques that can handle non-normally distributed data.

In summary, a Q-Q plot is a valuable diagnostic tool in linear regression that helps you assess the normality of residuals and detect outliers, ultimately leading to more accurate and reliable regression analyses.