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?**

   Done analysis on categorical variables using the barplot and boxplot .
   1. Seems like september has high bookings and most booking are from may to october.
   2. Fall has many bookings and summer comes next
   3. from thrusday to sunday we have little high bookings and there is no much difference in weekdays as such.
   4. Highest bookings are when the weather is clear and low when its snowing
   5. holidays have more bookings
   6. 2019 has more booking than 2018

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

   The drop_first=True parameter is important when creating dummy variables to address multicollinearity and improve interpretability. By dropping the first category, it avoids perfect multicollinearity and ensures a reference category for comparison. This simplifies the interpretation of the model and provides meaningful coefficients for each category relative to the dropped reference category.

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

   Temp variable has highest correlation with target variable

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

   I have validated using 5 assumptions:
   1. Normality of error terms
      Errors are normally distributed
   2. linear relationship validation
      linear relation should be visible in variables
   3. Multicollinearity
      There has to be insignificant multicolinearity
   4. Homoscedasticity
      No visible pattern

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. Fall
   3. September month

General Subjective Questions

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

   Linear regression is a widely used statistical algorithm for modeling the relationship between a dependent variable and one or more independent variables. It aims to find the best-fitting linear equation that represents the linear relationship between the variables. Here's a detailed explanation of the linear regression algorithm:

   1. **Data Preparation**:
      a. **Gather the dataset**: Collect the data containing the dependent variable(target variable) and independent variables (predictor variables) that you want to analyze and model.
      b. **Split the data**: Divide the dataset into a training set and a test set. The training set is used to train the linear regression model, while the test set is used to evaluate its performance.
   2. Model Representation:
      a. **Linear equation**: The linear regression algorithm represents the relationship between the dependent variable (y) and the independent variables (x) using a linear equation of the form: $y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$.
      b. β_o, β_1, β_2, ..., β_p are the coefficients (or weights) associated with each independent variable, representing the impact or effect of that variable on the dependent variable.
      c. p represents the number of independent variables in the model.
   3. **Model Training**:
      a. **Ordinary Least Squares (OLS)**: The most common method used to estimate the coefficients in linear regression is the Ordinary Least Squares algorithm. It aims to minimize the sum of squared differences between the actual values of the dependent variable and the predicted values by adjusting the coefficients.
      b. **Coefficient estimation**: The OLS algorithm calculates the coefficients by minimizing the residual sum of squares (RSS), which is the sum of the squared differences between the actual and predicted values.
      c. The algorithm uses matrix operations to efficiently estimate the coefficients. It calculates the coefficients as: $\beta = (X^T X)^{-1} X^T y$, where X is the matrix of independent variables, y is the vector of dependent variable values, and β is the vector of coefficients.
   4. **Model Evaluation:**
      a. Once the model is trained, it is evaluated using the test set to assess its performance and generalization ability.
      b. Evaluation metrics: Common evaluation metrics for linear regression include mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), R-squared (coefficient of determination), and adjusted R-squared.
      c. These metrics assess how well the linear regression model fits the data and how accurately it predicts the dependent variable.
   5. **Model Prediction**:After evaluating the model, it can be used to make predictions on new, unseen data. Given the values of the independent variables, the model calculates the predicted value of the dependent variable using the learned coefficients.

6. **Model Assumptions**:
    a. Linear regression relies on certain assumptions, such as linearity, independence of errors, constant variance of errors (homoscedasticity), and normality of errors.
    b. Violations of these assumptions can affect the accuracy and reliability of the linear regression model. Various diagnostic techniques are used to assess and address these assumptions.

Overall, the linear regression algorithm provides a straightforward and interpretable way to model and understand the relationship between variables. It is widely used in various fields, such as economics, finance, social sciences, and machine learning, for both prediction and inference tasks.

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

Anscombe's quartet refers to a collection of four datasets that have nearly identical statistical properties but exhibit distinct patterns when visualized. It highlights the importance of data visualization in understanding and interpreting statistical analyses. The quartet was introduced by the statistician Francis Anscombe in 1973. Let's explore each dataset in detail:

Dataset I:

This dataset consists of 11 data points with a linear relationship between the x and y variables.

When plotted, the data points form a clear upward linear trend, and a linear regression model would fit the data well.

The statistical properties, such as the means, variances, and correlation coefficient, closely resemble those of the other datasets in the quartet.

Dataset II:

Dataset II also contains 11 data points, but the relationship between x and y is nonlinear.

When plotted, the data appears to follow a curvilinear pattern, deviating from the linear relationship observed in Dataset I.

Despite the nonlinearity, the statistical properties remain similar to those of the other datasets.

Dataset III:

Dataset III comprises 11 data points, where all x values are the same except for one outlier.

When plotted, the data forms a linear pattern, similar to Dataset I. However, the outlier significantly affects the relationship between x and y.

The outlier demonstrates the importance of detecting and addressing outliers that can distort statistical analyses.

Dataset IV:

Dataset IV consists of 11 data points, where the relationship between x and y is strongly influenced by a single outlier.

When plotted, most of the data points form a linear relationship, but the outlier dramatically alters the regression line and correlation coefficient.

Dataset IV emphasizes the impact of influential points on statistical analyses, highlighting the need to identify and handle such observations.

The significance of Anscombe's quartet lies in its ability to illustrate that relying solely on summary statistics can be misleading. Despite having similar statistical properties, these datasets exhibit diverse patterns when visualized. It underscores the importance of data visualization to gain insights, detect patterns, outliers, and nonlinearity that summary statistics alone may not reveal.

Anscombe's quartet serves as a reminder to explore and visualize data thoroughly, enabling a deeper understanding of the relationships and characteristics within the dataset before drawing conclusions or making decisions based solely on statistical measures.

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

Pearson's correlation coefficient, often denoted as Pearson's R or simply as r, is a measure of the linear relationship between two continuous variables. It quantifies the strength and direction of the association between the variables.

Pearson's R ranges from -1 to 1. A value of 1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally. A value of -1 indicates a perfect negative linear relationship, where as one variable increases, the other variable decreases proportionally. A value of 0 suggests no linear relationship between the variables.

Pearson's R is calculated by dividing the covariance of the two variables by the product of their standard deviations. The formula for Pearson's correlation coefficient is:

$$r = (\Sigma[(X_i - \bar{X})(Y_i - \bar{Y})]) / [\sqrt{(\Sigma(X_i - \bar{X})^2)}\sqrt{(\Sigma(Y_i - \bar{Y})^2)}]$$

where:

X_i and Y_i represent individual data points of the variables X and Y, respectively.

$\bar{X}$ and $\bar{Y}$ denote the means of variables X and Y, respectively.

Σ represents the summation operator.

Pearson's R is commonly used in statistics, research, and data analysis to examine the relationship between variables and to determine the strength and direction of the association.

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

Scaling, in the context of data analysis and machine learning, refers to the process of transforming numerical variables to a consistent scale. It involves adjusting the values of the variables so that they are comparable and have similar ranges or distributions.

Scaling is performed for several reasons:

Comparison: Scaling allows for a fair and meaningful comparison between variables that may have different units or scales. When variables have significantly different ranges, their magnitudes can disproportionately influence certain analyses or algorithms.

Convergence: Many machine learning algorithms and optimization techniques rely on numerical optimization methods. Scaling can help these algorithms converge faster and more accurately by bringing variables to a similar scale. It prevents certain variables from dominating the optimization process due to their larger values.

Interpretation: Scaling can aid in the interpretability and understanding of the data. Variables with different scales can be challenging to interpret visually or when communicating the result to others. Scaling ensures that the variables are on a comparable scale, facilitating interpretation.

Normalized scaling and standardized scaling are two common scaling techniques:

Normalized Scaling (Min-Max Scaling): In normalized scaling, also known as min-max scaling, the values of the variable are transformed to a specific range, typically between 0 and 1. The formula for normalized scaling is:

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

Here, X represents the original value, X_min is the minimum value of the variable, and X_max is the maximum value of the variable. Normalized scaling preserves the relative relationships between values while constraining them within a specific range.

Standardized Scaling (Z-Score Scaling): Standardized scaling, also called z-score scaling or standardization, transforms the values of the variable to have a mean of 0 and a standard deviation of 1. The formula for standardized scaling is:

X_standardized = (X - X_mean) / X_std

Here, X represents the original value, X_mean is the mean of the variable, and X_std is the standard deviation of the variable. Standardized scaling ensures that the variable has zero mean and unit variance, making it suitable for certain statistical analyses and algorithms that assume normally distributed data.

The choice between normalized scaling and standardized scaling depends on the specific requirements of the analysis or the algorithm being used. Normalized scaling is useful when preserving the original range and relative relationships is important, while standardized scaling is beneficial when data normalization and achieving a standard distribution are desired.

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

The occurrence of an infinite value of VIF (Variance Inflation Factor) typically happens due to perfect multicollinearity in the data. Perfect multicollinearity refers to a situation where one or more independent variables in a regression model can be perfectly predicted from a linear combination of other variables.

Here are three reasons why an infinite VIF value may arise:

Redundant variables: When one variable is a perfect linear combination of other variables in the model, it creates redundancy. In such cases, the VIF calculation involves division by zero, leading to an infinite VIF value.

Singular design matrix: In some cases, the design matrix used in regression analysis becomes singular, which means it is not invertible. This happens when the independent variables are perfectly correlated, resulting in a matrix with linearly dependent columns. When calculating the inverse of such a matrix to compute the VIF, it leads to an infinite value.

Perfectly predicted variable: If one independent variable can be precisely predicted from a linear combination of the other variables, it introduces perfect multicollinearity. In this scenario, when calculating the VIF for the variable that can be predicted, it results in division by zero, resulting in an infinite VIF value.

It's important to note that an infinite VIF value is an indication of a severe problem in the regression model. Perfect multicollinearity hampers the estimation and interpretation of coefficients and undermines the reliability of the model. To address this issue, one needs to identify and resolve the multicollinearity problem, such as by removing redundant variables or transforming variables to reduce their correlation.

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

A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess whether a dataset follows a particular theoretical distribution. It compares the quantiles of the dataset against the quantiles of the expected distribution, typically a normal distribution.

The Q-Q plot is constructed by plotting the observed quantiles of the dataset on the y-axis against the corresponding quantiles of the theoretical distribution on the x-axis. If the data points on the plot closely follow a straight line, it indicates that the dataset is well approximated by the theoretical distribution. Deviations from the straight line suggest departures from the expected distribution.

In the context of linear regression, Q-Q plots are particularly useful for evaluating the assumption of normality of residuals. Residuals are the differences between the observed values and the predicted values from the regression model. The Q-Q plot of residuals allows us to assess whether the residuals are normally distributed.

The use and importance of a Q-Q plot in linear regression can be summarized as follows:

Assessing normality: The Q-Q plot helps us visually assess whether the residuals follow a normal distribution. If the points on the Q-Q plot deviate significantly from the straight line, it suggests that the residuals may not be normally distributed. Departures from normality can affect the validity of statistical tests and the reliability of regression estimates.

Detecting outliers and skewness: In addition to evaluating normality, the Q-Q plot can also reveal the presence of outliers or skewness in the residuals. Outliers are data points that significantly deviate from the expected pattern, while skewness indicates a departure from symmetry. These departures can influence the assumptions of linear regression and may warrant further investigation or data transformation.

Model diagnostics: Q-Q plots are an important diagnostic tool in linear regression. By examining the Q-Q plot of residuals, we can identify potential issues with the model assumptions and determine if any corrective actions are necessary. If the residuals do not exhibit a normal distribution, it may be necessary to explore alternative modeling techniques or consider transformations to improve the model's performance.

Overall, the Q-Q plot provides a visual assessment of how well the residuals conform to the assumption of normality in linear regression. It helps to validate the model and identify potential problems that need to be addressed for accurate and reliable analysis.