## 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?**

Categorical variables can have a significant effect on the dependent variable by influencing its outcomes based on the different groups or categories they represent. By analyzing the distribution of the dependent variable across these categories, you can identify patterns and relationships. For example, in a dataset predicting diabetes, if a larger percentage of people with a "Non-Vegetarian" diet are diabetic compared to "Vegan" individuals, this suggests that diet type affects the outcome. Additionally, statistical tests like the Chi-Square test can confirm whether there is a strong association between a categorical variable and the target. Visualization techniques, such as bar plots or stacked charts, also help in spotting these trends. Moreover, using methods like dummy variables in regression or feature importance in models can quantify the exact influence of each category on the dependent variable.

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

Using drop_first=True during dummy variable creation is important to avoid **multicollinearity**, which occurs when one variable can be predicted from others, making the model unstable. By dropping one category, we prevent redundancy in the dataset while still retaining all the necessary information to distinguish between categories. This helps improve the performance and interpretability of the model.

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

The numerical variable that has the highest correlation with the target variable can be identified by observing the pair-plot, which shows scatter plots and correlation coefficients. The variable with the most linear-looking relationship or the highest correlation coefficient in the pair-plot will have the strongest correlation with the target.

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

To validate the assumptions of linear regression after building the model on the training set, the following steps are taken:
1. Linearity: Check if the relationship between the independent variables and the target variable is linear. This is done by plotting the residuals vs. predicted values. If the residuals are randomly scattered around zero, the linearity assumption holds.
2. Homoscedasticity: Ensure constant variance of residuals by analyzing the same residuals vs. predicted values plot. If the spread of residuals is consistent across the range of predictions, the assumption of homoscedasticity is satisfied.
3. Normality of Residuals: Assess the normality of residuals using a Q-Q plot or histogram. If the residuals follow a straight line in the Q-Q plot or the histogram resembles a normal distribution, the assumption is valid.

These steps confirm that the key assumptions of linear regression—linearity, homoscedasticity, and normality of residuals—are satisfied.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

Based on the final model, the top three features that significantly contribute to explaining the demand for shared bikes are likely to include:
1. Weather Conditions: Variables such as temperature and humidity can greatly influence bike usage, as pleasant weather typically leads to higher demand for shared bikes.
2. Day of the Week: This feature often reflects patterns in user behavior, with demand typically increasing on weekends compared to weekdays, as people tend to engage in recreational activities.
3. Time of Day: The time (e.g., peak hours vs. off-peak hours) plays a critical role in determining bike demand, with higher usage during commuting hours in the morning and evening.
These features provide valuable insights into the factors that drive bike-sharing demand and help in making accurate predictions.

## General Subjective Questions

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

Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (features) by fitting a linear equation to observed data.

**Key Components:**

1. Equation of the Line: The relationship is expressed as:
   $$
   Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon
   $$
   Where:
   - $Y$ is the dependent variable.
   - $X_1, X_2, ..., X_n$ are the independent variables.
   - $\beta_0$ is the intercept (the expected value of $Y$ when all $X$ are 0).
   - $\beta_1, \beta_2, ..., \beta_n$ are the coefficients (indicating the change in $Y$ for a unit change in $X$).
   - $\epsilon$ is the error term (captures the variability in $Y$ not explained by $X$).

2. Objective: The goal of linear regression is to find the best-fitting line by minimizing the difference between the predicted values ($\hat{Y}$) and the actual values ($Y$). This is typically achieved using the Ordinary Least Squares (OLS) method, which minimizes the sum of the squared residuals (the differences between observed and predicted values).

3. Assumptions: Linear regression relies on several assumptions:
   - Linearity: The relationship between the independent and dependent variables is linear.
   - Independence: Observations are independent of each other.

- Homoscedasticity: The residuals (errors) have constant variance at every level of $X$.
- Normality: The residuals of the model should be approximately normally distributed.

4. Evaluation: After fitting the model, its performance is evaluated using metrics such as:
- R-squared: Indicates the proportion of variance in the dependent variable explained by the independent variables.
- Mean Absolute Error (MAE): Average of the absolute differences between predicted and actual values.
- Root Mean Squared Error (RMSE): Measures the square root of the average of squared differences, giving higher weight to larger errors.

By following these steps, linear regression provides a simple yet powerful approach for predicting outcomes based on linear relationships in the data.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets that have identical statistical properties but vastly different distributions and graphical representations. Created by the statistician Francis Anscombe in 1973, the quartet illustrates the importance of data visualization in understanding the characteristics of data, as relying solely on summary statistics can be misleading.

### Key Features of Anscombe's Quartet:

1. Identical Summary Statistics:
Each of the four datasets has the same:
- Mean of $x$ values: 9
- Mean of $y$ values: 7.5
- Variance of $x$: 11
- Variance of $y$: 4.125
- Correlation coefficient ($r$): Approximately 0.816

Despite these similarities, the datasets differ significantly in their distributions and relationships between $x$ and $y$.

2. Distinct Graphical Representations:
- Dataset I: A linear relationship with some scatter around the line.
- Dataset II: A strong linear relationship, but with a significant outlier that dramatically affects the regression line.
- Dataset III: A non-linear relationship, where the points form a curve rather than a straight line, demonstrating that a linear model would be inappropriate.
- Dataset IV: Similar to Dataset II but with a vertical line of points (a clear outlier), showing how a single point can heavily influence regression analysis.

3. Illustration of Data Visualization Importance:
Anscombe's quartet emphasizes that summary statistics (like mean, variance, and correlation) can be misleading when interpreting data. Visualizing the data through scatter plots can reveal patterns, trends, and anomalies that are not apparent from numerical summaries alone. This is a

crucial lesson in statistical analysis and data science, underscoring the necessity of combining statistical measures with visual exploration for accurate data interpretation.

In summary, Anscombe's quartet serves as a powerful reminder of the importance of data visualization and the potential pitfalls of relying solely on summary statistics to understand data relationships.

3. What is Pearson's R?

Pearson's $r$, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. Here's a detailed explanation:

## Key Features of Pearson's $r$:

1. Range and Interpretation:
   - The Pearson correlation coefficient ranges from $-1$ to $1$:
     - $r = 1$: Perfect positive correlation, meaning that as one variable increases, the other variable also increases proportionally.
     - $r = -1$: Perfect negative correlation, indicating that as one variable increases, the other decreases proportionally.
     - $r = 0$: No correlation, suggesting that there is no linear relationship between the two variables.
   - Values between $0$ and $1$ indicate a positive correlation, while values between $-1$ and $0$ indicate a negative correlation. The closer the absolute value of $r$ is to $1$, the stronger the linear relationship.

2. Formula:
   The Pearson correlation coefficient is calculated using the formula:
   $$
   r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}
   $$
   Where:
   - $n$ = number of data points
   - $x$ and $y$ = individual data points of the two variables
   - The numerator represents the covariance of the two variables, while the denominator standardizes this value by the standard deviations of both variables.

3. Assumptions:
   - Linearity: Pearson's $r$ assumes a linear relationship between the two variables, making it less suitable for non-linear associations.
   - Normality: The variables should ideally be normally distributed, especially in smaller samples, to ensure the reliability of the correlation coefficient.
   - Homoscedasticity: The variability of one variable should be consistent across the range of the other variable.

In summary, Pearson's $r$ is a vital tool in statistics for assessing the strength and direction of linear relationships between two continuous variables, but it is essential to consider its assumptions and limitations for accurate interpretation.

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

Scaling refers to the process of transforming the features of a dataset into a specific range or distribution to ensure that they contribute equally to the analysis. This is particularly important in machine learning algorithms, as features on different scales can lead to biased models and affect performance.

### Reasons for Scaling:

1. Improved Model Performance: Many machine learning algorithms, particularly those based on distance metrics (like k-nearest neighbors and clustering), are sensitive to the scale of the data. Scaling helps improve the performance of these algorithms.

2. Convergence in Optimization: Algorithms that use gradient descent, such as linear regression and neural networks, can converge faster when features are on similar scales. Scaling can help avoid issues like the vanishing or exploding gradients.

3. Equal Weighting: In algorithms that assign weights based on feature importance, scaling ensures that no single feature disproportionately influences the model due to its magnitude.

### Types of Scaling:

1. Normalized Scaling (Min-Max Scaling):
   - Definition: Normalization rescales the data to a fixed range, typically $[0, 1]$. The formula for normalization is:
     $$
     X' = \frac{X - \text{min}(X)}{\text{max}(X) - \text{min}(X)}
     $$
   - Use Case: This technique is useful when the distribution of the data is not Gaussian and when it’s essential to keep all values within a specific range.

2. Standardized Scaling (Z-score Normalization):
   - Definition: Standardization rescales the data to have a mean of $0$ and a standard deviation of $1$. The formula for standardization is:
     $$
     X' = \frac{X - \mu}{\sigma}
     $$
     where $\mu$ is the mean of the feature values and $\sigma$ is the standard deviation.
   - Use Case: Standardization is preferred when the data follows a Gaussian distribution, as it retains information about the distribution and is less affected by outliers compared to normalization.

### Key Differences:

- Range: Normalized scaling results in values between $0$ and $1$, while standardized scaling produces values with a mean of $0$ and a standard deviation of $1$.
- Sensitivity to Outliers: Normalization can be heavily influenced by outliers, as it scales based on the minimum and maximum values. In contrast, standardization is less sensitive to outliers due to its reliance on the mean and standard deviation.

In summary, scaling is essential for preparing data for machine learning models, with normalization and standardization being two common techniques used to ensure that features are on comparable scales.

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

Variance Inflation Factor (VIF) is a measure used to quantify the extent of multicollinearity in regression analysis. It helps assess how much the variance of an estimated regression coefficient increases due to collinearity with other predictors. When calculating VIF, an infinite value indicates a severe case of multicollinearity. Here's why this can happen:

### Reasons for Infinite VIF:

1. Perfect Multicollinearity:
   - Definition: Perfect multicollinearity occurs when one predictor variable is an exact linear combination of one or more other predictor variables. This means that at least one of the predictors can be expressed as a linear function of others.
   - Implication: When perfect multicollinearity exists, the matrix used in regression calculations becomes singular (non-invertible). As a result, the VIF for the perfectly collinear variable becomes infinite, indicating that its contribution to the regression model cannot be determined reliably.

2. Redundant Features:
   - If two or more variables in the model provide redundant information (i.e., they contain the same information), it leads to perfect multicollinearity. For instance, if both variables represent the same concept or are derived from each other, the variance associated with one of them will inflate due to their interdependence.

3. Issues with Data Collection:
   - Sometimes, data collection methods might introduce perfect collinearity. For example, if measurements are taken from a predefined sample that inherently contains duplicated or highly correlated features, it could lead to infinite VIF values.

### Consequences:

- When VIF values are infinite, it implies that the regression coefficients are not estimable, which means that the model cannot be fitted accurately, making it necessary to remove or combine collinear predictors to resolve the issue.

In summary, infinite VIF occurs primarily due to perfect multicollinearity among predictor variables, leading to redundancy in information and causing issues in the regression analysis. This emphasizes the importance of identifying and addressing multicollinearity before fitting regression models.

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

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, most commonly the normal distribution. In a Q-Q plot, the quantiles of the observed data are plotted against the quantiles of the expected distribution. If the points on the plot fall approximately along a straight line, it suggests that the data follows the specified distribution.

### Use and Importance of Q-Q Plots in Linear Regression:

1. Normality of Residuals:
   - In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) should be normally distributed. A Q-Q plot helps assess this assumption visually. If the residuals deviate significantly from the straight line in the plot, it indicates that they may not be normally distributed, which could affect the validity of statistical inferences made from the model.

2. Identifying Outliers:
   - Q-Q plots can help in identifying outliers in the data. Points that lie far away from the reference line may represent outliers or extreme values that could influence the regression model's performance. Identifying and potentially addressing these outliers is crucial for building a robust linear regression model.

3. Model Diagnostics:
   - By evaluating the normality of residuals through a Q-Q plot, analysts can conduct diagnostics to determine whether the assumptions of linear regression are met. If the normality assumption is violated, transformations of the dependent variable (e.g., logarithmic or square root transformations) may be necessary, or a different modeling approach (e.g., non-linear regression) may be considered.

### Summary:

In summary, a Q-Q plot is a vital tool in linear regression analysis for checking the normality of residuals, identifying outliers, and performing model diagnostics. By ensuring that the assumptions of linear regression are satisfied, analysts can make more reliable predictions and statistical inferences.