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)**

1. Season vs cnt:

- **Spread**: The interquartile range (IQR) appears to be widest for summer and narrowest for winter, suggesting that there is more variability in counts during the summer and less in the winter.
- **Median**: The median value is highest in the summer, indicating that the count is typically higher in this season. The median is lowest in the spring.
- **Outliers**: There are a few outliers in the spring and winter, indicating occasional counts that are much higher than usual.

2. Month vs cnt:

- **Spread**: Some months, like June, July, and August, show a larger IQR, indicating more variability in counts. December, on the other hand, shows a smaller spread.
- **Median**: The medians are generally higher in the warmer months (May to October) and lower in the colder months (November to April), suggesting a possible seasonal trend in the counts.
- **Outliers**: There are outliers present in several months, with January, March, and December having noticeable lower outliers. These could represent days with significantly lower counts.

3. Week vs cnt:

- **Spread**: The spread of counts across days of the week seems fairly consistent, with no dramatic differences in the IQR.
- **Median**: There is a slight variation in median counts, but not significant. This suggests that the day of the week might not have a strong influence on the counts.
- **Outliers**: There are some outliers on Saturday, indicating occasional high counts that deviate from the typical range.

4. Weather vs cnt:

- **Spread**: Clear weather shows a wider spread compared to mist and light rain, suggesting more variability in counts during clear conditions.
- **Median**: The median count is highest for clear weather conditions, which might indicate that more events occur or more observations are recorded during clear weather. The median is notably lower for light rain.
- Outliers: There are no visible outliers for any of the weather conditions, which suggests that counts are relatively consistent within each weather category.

Overall Analysis:

- The counts are typically higher during the summer months, which could be due to more activities or events occurring during this season.
- The variability in counts is greater during the summer and clear weather conditions, which might suggest that factors such as events are influencing the counts more during these times.
- Lower counts in the winter and during light rain may indicate a seasonal effect or a weather-related reduction in activities.

- The day of the week does not seem to be a strong factor affecting counts, as evidenced by the relatively uniform spread and median values.

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

   Dropping the first level of a categorical variable makes it the reference category. This reference category serves as a baseline for interpreting the coefficients of the remaining dummy variables. The coefficients for the other levels of the categorical variable can be interpreted as how they differ from the reference category.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

   Temp and atemp has the highest correlation with target variable 'cnt'

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

   Using Residual Analysis, to plot the distribution of error terms which shows that the distribution is normal.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

   Temperature, Spring season and Humidity

General Subjective Questions

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

   Linear regression is a widely used statistical method in machine learning and statistics for modelling the relationship between a dependent variable (target) and one or more independent variables (features). It assumes that there is a linear relationship between the input variables and the target variable. Linear regression aims to find the best-fitting straight that minimizes the difference between the observed and predicted values of the target variable. Let's break down the linear regression algorithm in detail:

   1. Simple Linear Regression:

      - In simple linear regression, there is one independent variable (feature) and one dependent variable (target).

      y = b0 + b1 * x

      - `y` is the target variable.
      - `x` is the independent variable.
      - `b0` is the y-intercept (the value of `y` when `x` is 0).
      - `b1` is the slope of the line (how much `y` changes for a unit change in `x`).

   - The goal is to find the values of `b0` and `b1` that minimize the sum of squared differences between the observed values of `y` and the predicted values (least squares method).

   2. Multiple Linear Regression:

      - In multiple linear regression, there are multiple independent variables (features) and one dependent variable (target).

      y = b0 + (b1 * x1) + (b2 * x2) + ... + (bn * xn)

      - `y` is the target variable.
      - `x1`, `x2`, ..., `xn` are the independent variables.
      - `b0` is the y-intercept.
      - `b1`, `b2`, ..., `bn` are the coefficients for each independent variable.

   - The goal is to find the values of `b0`, `b1`, `b2`, ..., `bn` that minimize the sum of squared differences between the observed values of `y` and the predicted values.

3. Training the Model:

- Training involves finding the values of the coefficients (`b0`, `b1`, `b2`, ..., `bn`) that best fit the data.
- Typically, this is done using optimization techniques like gradient descent or analytical methods like the normal equation.
- The optimization process aims to minimize the cost function, which measures the error between the predicted values and the actual target values.

4. Making Predictions:

- Once the model is trained, you can use it to make predictions on new, unseen data.
- Plug the values of the independent variables into the regression equation to predict the target variable.

5. Evaluation:

- To assess the performance of the linear regression model, various evaluation metrics can be used, such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (R2), etc.
- These metrics help measure the goodness of fit and the predictive accuracy of the model.

6. Assumptions of Linear Regression:

- Linear relationship: Assumes that the relationship between independent and dependent variables is linear.
- Independence: Assumes that the residuals (the differences between observed and predicted values) are independent of each other.
- Homoscedasticity: Assumes that the variance of the residuals is constant across all levels of the independent variables.
- No multicollinearity: Assumes that independent variables are not highly correlated with each other.

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

Anscombe's quartet is a set of four small datasets that have nearly identical simple descriptive statistics, but they exhibit vastly different characteristics when visualized or subjected to more in-depth statistical analysis. It was created to illustrate the importance of visualizing data and the limitations of relying solely on summary statistics. The quartet is often used to emphasize the concept that data should not be summarized solely by numerical measures, as visual inspection can reveal hidden patterns and relationships.

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

Pearson's correlation coefficient, "r," is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. Pearson's R ranges from -1 to 1 and has the following interpretation:

- A positive value of Pearson's R (closer to 1) indicates a strong positive linear relationship. This means that as one variable increases, the other tends to increase as well.

- A negative value of Pearson's R (closer to -1) indicates a strong negative linear relationship. This means that as one variable increases, the other tends to decrease.

- A Pearson's R value of 0 suggests no linear relationship between the two variables. In other words, they are not linearly correlated.

The formula for calculating Pearson's correlation coefficient for two variables, X and Y, with n data points, is as follows:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Pearson's R measures the strength and direction of the linear relationship, but it does not capture nonlinear relationships or dependencies between variables. Additionally, it assumes that the variables are normally distributed and that there are no outliers that could unduly influence the results. When these assumptions are violated, other correlation measures or data transformations may be more appropriate.

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 preprocessing refers to the process of transforming numerical features (variables) in a dataset to a common scale or range.

1. Why Scaling is Performed:

- Magnitude Consistency: Scaling ensures that all variables have similar magnitudes or scales. When features have vastly different ranges, some variables can dominate the others during certain computations or modeling processes, leading to biased results.

2. Normalized Scaling:

- Range: In normalized scaling, each variable is scaled to a specific range, typically between 0 and 1. This is achieved by transforming each data point using the formula:

$$X_normalized = \frac{X - X_min}{X_max - X_min}$$

- Outcome: Normalized scaling ensures that all variables have values between 0 and 1, preserving the relative proportions of data while bringing them into a consistent range.

3. Standardized Scaling:

- Mean and Standard Deviation: In standardized scaling, each variable is transformed to have a mean (average) of 0 and a standard deviation of 1. This is done by subtracting the mean and dividing by the standard deviation:

$$X_standardized = \frac{X - \mu}{\sigma}$$

- Outcome: Standardized scaling centers the data around 0 and scales it to have a spread of

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

The VIF can become infinite (or extremely large) when perfect multicollinearity exists in the dataset. Perfect multicollinearity occurs when one or more independent variables in the regression model can be exactly predicted from a linear combination of other independent variables.

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 is a valuable tool for assessing the goodness of fit between observed data and a theoretical distribution, particularly the normal distribution in the context of linear regression. It helps ensure that key assumptions of the regression model are met, aids in the detection of skewness and outliers, and provides insights into the data's distribution, which can guide data preprocessing and model selection decisions.