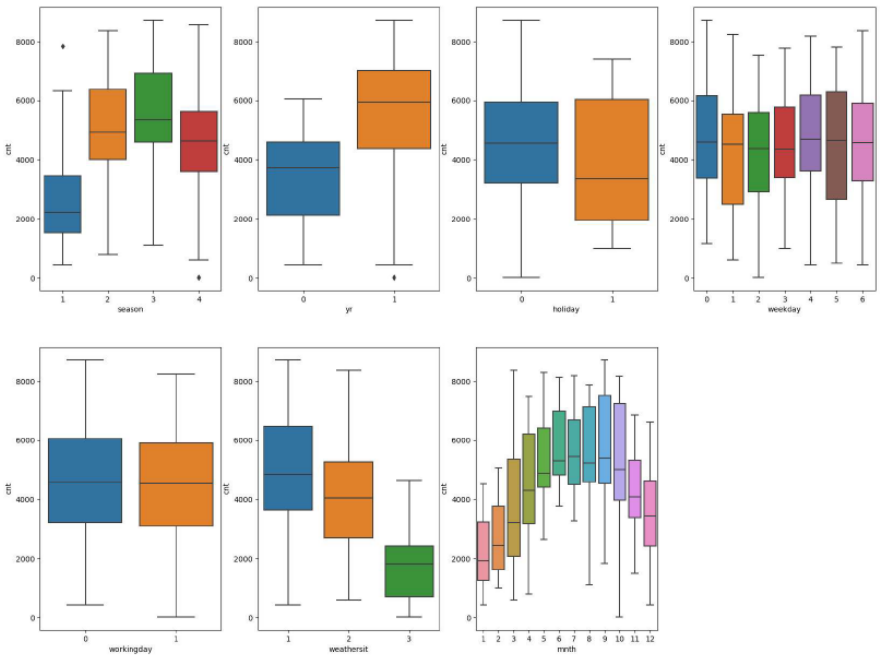
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer:

From the analysis of the categorical variables in the dataset, the following inferences about their effect on the dependent variable 'cnt' can be made:

- **Season**:
  - Fall and Summer are the most favorable seasons for bike rentals, showing the highest counts.
  - Spring has the least number of users.
  - The Winter season shows moderate usage.
  - Strategic advertising can target peak seasons to boost rentals further.



- **Year (yr)**:
  - There is a significant increase in bike rentals from 2018 to 2019, indicating growth in popularity or user base over time.
- **Month (mnth)**:
  - Bike rentals peak during the months of June, July, August, September, and October.
  - Rentals are comparatively lower in January and December, likely due to colder weather conditions.
- **Weekday**:
  - There is no significant pattern observed in the total count ('cnt') across different weekdays.
  - However, registered users tend to rent bikes more on weekdays, while casual users prefer weekends.
  - Thursday, Friday, Saturday, and Sunday see more bookings compared to the start of the week.
- **Working Day (workingday)**:
  - Registered users are more likely to rent bikes on working days, suggesting usage for daily commutes.
  - Casual users prefer non-working days.
  - Overall bike demand remains relatively constant regardless of whether it is a working day or not.
- **Holiday**:
  - Bike rentals decrease during holidays.

- Casual users are more likely to rent bikes on holidays compared to registered users.

- **Weather Situation (weathersit)**:
  - Clear or partly cloudy weather conditions attract the highest number of rentals.
  - Light mist or rain sees a decrease in bike usage.
  - There are no rentals recorded during heavy rain or snow, indicating adverse weather conditions deter bike usage.

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

Using drop_first=True during dummy variable creation is important for the following reasons:

- **Reduces Multi-collinearity**:
  - When we create dummy variables, each category level of a categorical variable is typically transformed into a separate column. For a categorical variable with nnn levels, nnn dummy variables are created.
  - Including all nnn dummy variables introduces multi-collinearity because one dummy variable can be predicted from the others. This high correlation can negatively impact the model's performance.
  - By setting drop_first=True, the first category level is dropped, resulting in n−1n-1n−1 dummy variables. This helps to avoid multi-collinearity by removing the redundant dummy variable.
- **Simplifies Interpretation**:
  - With n−1n-1n−1 dummy variables, the omitted category serves as the reference category. The model coefficients for the remaining dummy variables represent the change relative to this reference category.
  - This approach simplifies the interpretation of the model's coefficients.

Example:

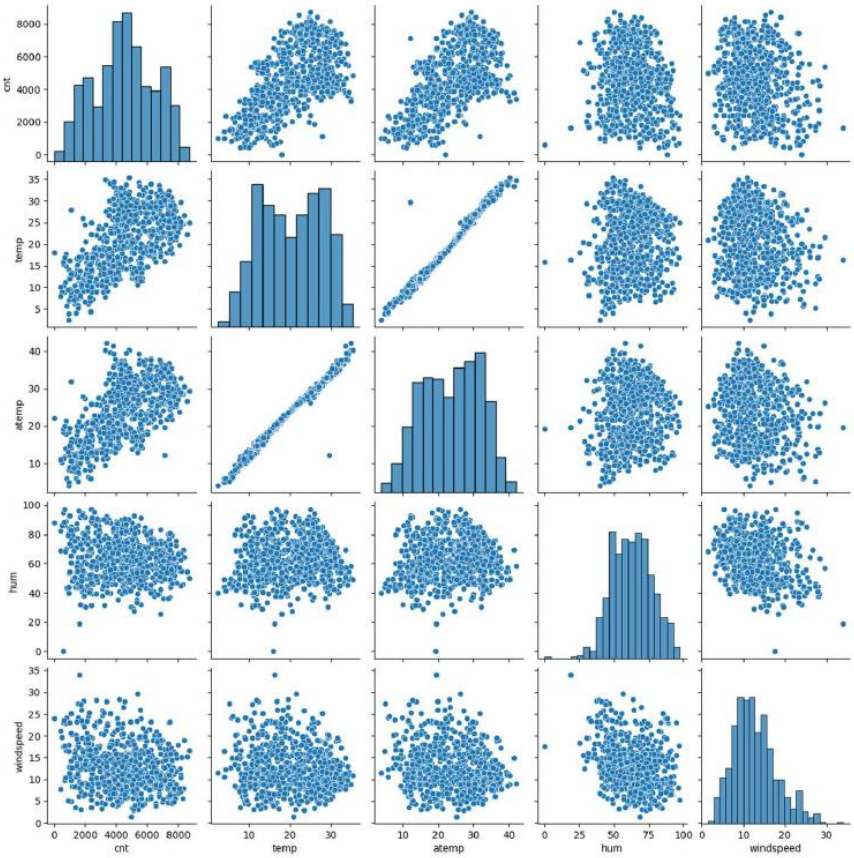
If a categorical variable has three levels: A, B, and C:

- Without drop_first=True, three dummy variables are created: A, B, and C.
- With drop_first=True, only two dummy variables are created: A and B. The presence of a zero in both indicates the third category, C.

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

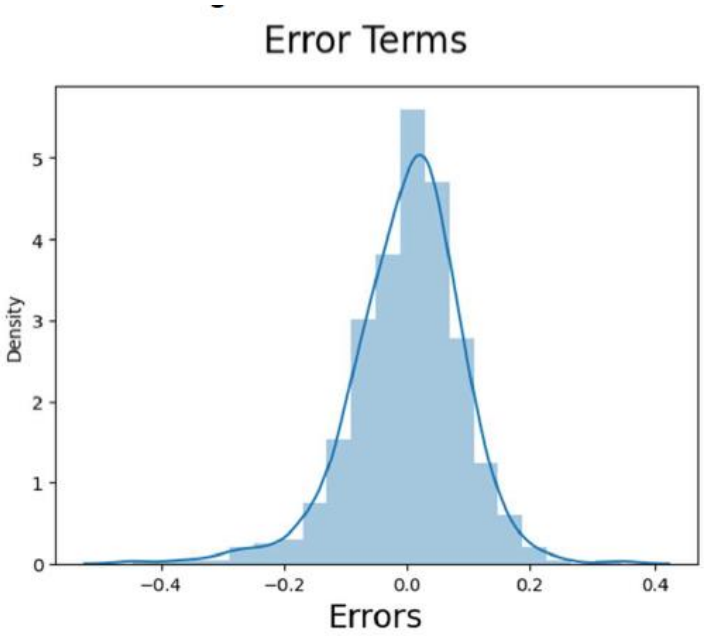
Among the numerical variables, the variable **'temp'** (temperature) has the highest correlation with the target variable **'cnt'** (count of bike rentals).

- **Reasoning**:
    - The 'temp' variable shows the strongest relationship with the number of bike rentals.
    - While 'atemp' (feels like temperature) also shows a high correlation, it is derived from 'temp', humidity, and windspeed, so 'temp' is the primary variable to consider.
    - Other variables have lower correlations with the target variable compared to 'temp'.



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

To validate the assumptions of Linear Regression after building the model on the training set, I conducted several checks:



- **Linearity**:
    - Ensured a linear relationship between the independent and dependent variables by visualizing the numeric variables using a pairplot.
    - Verified that the actual vs. predicted plot showed points distributed symmetrically around the diagonal line.
- **Independence of Errors**:
    - Checked that the error terms were independent of each other by ensuring there was no specific pattern observed in the residuals versus predicted values plot.
- **Normality of Error Terms**:
    - Validated that the residuals followed a normal distribution and were centered around zero (mean = 0) using a histogram and distribution plot of the residuals.
- **Homoscedasticity**:
    - Ensured that the residuals had constant variance by plotting the residuals and verifying that there was no visible pattern (i.e., the residuals were spread evenly).
- **Multicollinearity**:
    - Checked for insignificant multicollinearity among the independent variables by calculating the Variance Inflation Factor (VIF) to quantify how strongly the feature variables were associated with each other.

These steps helped ensure that the Linear Regression model met the necessary assumptions for reliable predictions.

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

The top 3 features contributing significantly towards explaining the demand for shared bikes based on the final model are:

1. **Temperature (temp)**: This feature has the highest positive impact on bike demand, indicating that higher temperatures lead to increased bike rentals.
2. **Year (yr)**: The demand for shared bikes has grown year over year, showing a significant positive effect on bike rentals in 2019 compared to 2018.
3. **Weather Situation (weathersit)**: Adverse weather conditions such as light snow and rain negatively impact bike demand, showing a significant decrease in rentals during such conditions.

General Subjective Questions

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

Linear Regression Algorithm Explanation

Linear Regression is a fundamental statistical and machine learning technique used to model and analyze the relationship between a dependent variable and one or more independent variables. The goal is to find the best-fitting linear relationship that can predict the dependent variable based on the input features.

Types of Linear Regression:

1. **Simple Linear Regression**:
    - Involves a single independent variable.
    - The relationship is modeled using the equation: Y=β0+β1XY = \beta_0 + \beta_1XY=β0+β1X
        - YYY is the dependent variable.
        - XXX is the independent variable.
        - β0\beta_0β0 (intercept) is the value of YYY when X=0X = 0X=0.
        - β1\beta_1β1 (slope) represents the change in YYY for a one-unit change in XXX.
2. **Multiple Linear Regression**:
    - Involves multiple independent variables.
    - The relationship is modeled using the equation: Y=β0+β1X1+β2X2+⋯+βpXpY = \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_pX_pY=β0+β1X1+β2X2+⋯+βpXp
        - X1,X2,…,XpX_1, X_2, \ldots, X_pX1,X2,…,Xp are the independent variables.
        - β0\beta_0β0 is the intercept, and β1,β2,…,βp\beta_1, \beta_2, \ldots, \beta_pβ1,β2,…,βp are the coefficients of the respective variables.

Objective:

The objective of linear regression is to determine the best-fit line that minimizes the error between the predicted values and the actual values. This is achieved by finding the optimal values for the coefficients (β\betaβ).

Cost Function:

- **Mean Squared Error (MSE)** is commonly used as the cost function: MSE=1n∑i=1n(yi−y^i)2\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2MSE=n1i=1∑n(yi−y^i)2 where nnn is the number of observations, yiy_iyi is the actual value, and y^i\hat{y}_iy^i is the predicted value.
- The goal is to minimize the MSE to find the best-fit line.

Methods to Minimize Cost Function:

1. **Gradient Descent**:
    - An iterative optimization algorithm that adjusts the coefficients to minimize the cost function. The weights are updated using: β=β−α∂J(β)∂β\beta = \beta - \alpha \frac{\partial J(\beta)}{\partial \beta}β=β−α∂β∂J(β) where α\alphaα is the learning rate, and ∂J(β)∂β\frac{\partial J(\beta)}{\partial \beta}∂β∂J(β) is the gradient of the cost function.
2. **Ordinary Least Squares (OLS)**:
    - A method to find the coefficients by minimizing the Residual Sum of Squares (RSS): RSS=∑i=1n(yi−y^i)2\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2RSS=i=1∑n(yi−y^i)2

Assumptions:

- **Linearity**: The relationship between the dependent and independent variables should be linear.
- **Independence**: Residuals (errors) should be independent.
- **Homoscedasticity**: Residuals should have constant variance.
- **Normality**: Residuals should be normally distributed.

Limitations:

- Assumes a linear relationship between input and output variables, which may not always hold.
- Sensitive to outliers and multicollinearity among independent variables.

linear regression provides a straightforward and effective way to understand relationships between variables and predict outcomes. It involves fitting a line to data points to minimize prediction error, with a variety of methods available for optimizing the fit.

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

Anscombe’s Quartet

Explanation:

Anscombe's Quartet, introduced by statistician Francis Anscombe in 1973, consists of four datasets, each containing eleven (x, y) pairs. Despite having nearly identical descriptive statistics, these datasets display drastically different patterns when graphed. The quartet was designed to illustrate that statistical summaries alone can be misleading and that visualizing data is crucial for accurate analysis.

Key Points:

1. **Identical Descriptive Statistics:**
    - All four datasets share the same summary statistics:
        - Mean of x: 9
        - Mean of y: 7.50
        - Variance of x: 11
        - Variance of y: 4.13
        - Correlation coefficient between x and y: 0.816
2. **Different Visual Representations:**
    - **Dataset I:** Shows a clear linear relationship between x and y. The data points form a well-fitting linear model, which aligns with the regression line.
    - **Dataset II:** Although the summary statistics are similar, the data points are not normally distributed, and the relationship between x and y is not linear. This dataset highlights the importance of checking data distribution.
    - **Dataset III:** Features a linear distribution similar to Dataset I, but the presence of a single outlier significantly impacts the regression line. This outlier skews the results, emphasizing how influential points can affect analysis.
    - **Dataset IV:** Contains a high-leverage point that skews the correlation coefficient. Even though most data points suggest no strong relationship between x and y, this single point creates a misleadingly high correlation coefficient.

Importance of Visualization:

- **Visualizing Data:** Anscombe's Quartet underscores the importance of visualizing data before drawing conclusions based on statistical measures alone. The visual plots reveal patterns and anomalies that summary statistics might obscure.
- **Impact of Outliers:** The quartet demonstrates how outliers or influential points can distort statistical results, leading to incorrect interpretations. Properly addressing outliers and understanding their influence is crucial for accurate data analysis.

Anscombe's Quartet serves as a powerful reminder that statistical properties do not always tell the full story. Visualization provides critical insights into the structure and relationships within the data, making it an essential step in data analysis.

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

Pearson's R (Pearson's Correlation Coefficient)

Definition: Pearson's R, also known as Pearson's correlation coefficient, is a measure of the strength and direction of the linear relationship between two variables. It quantifies how well the relationship between the variables can be described by a straight line.

Formula: r=n(∑xy)−(∑x)(∑y)[n∑x2−(∑x)2][n∑y2−(∑y)2]r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}r=[n∑x2−(∑x)2][n∑y2−(∑y)2]n(∑xy)−(∑x)(∑y)

Where:

- nnn = Number of paired scores
- ∑xy\sum xy∑xy = Sum of the product of paired scores

- ∑x\sum x∑x = Sum of x scores
- ∑y\sum y∑y = Sum of y scores
- ∑x2\sum x^2∑x2 = Sum of squared x scores
- ∑y2\sum y^2∑y2 = Sum of squared y scores

Interpretation:

- **Range:** Pearson's R values range from -1 to 1.
    - **+1:** Perfect positive linear relationship (as one variable increases, the other increases proportionally).
    - **-1:** Perfect negative linear relationship (as one variable increases, the other decreases proportionally).
    - **0:** No linear relationship (the variables do not show any consistent linear trend).

Characteristics:

- **Direction:** A positive value indicates a positive linear relationship, while a negative value indicates a negative linear relationship.
- **Strength:** The magnitude of the coefficient (regardless of sign) indicates the strength of the linear relationship. Values closer to ±1 indicate a stronger linear relationship, while values closer to 0 indicate a weaker linear relationship.
- **Applicability:** Pearson's R is best used for linear relationships and may not accurately represent the relationship if it is nonlinear.

Pearson's R is a critical measure in statistics and data analysis for understanding the degree and direction of a linear relationship between two variables. It provides insights into how changes in one variable relate to changes in another, helping to identify patterns and correlations in data.

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

Scaling is a data pre-processing technique used to normalize or standardize the range of independent variables or features. It adjusts the features so that they fall within a specific range or have specific statistical properties, ensuring that all features contribute equally to the analysis and preventing any single feature from dominating due to its scale.

Why is Scaling Performed?

- **Uniformity:** Scaling ensures that features with different magnitudes or units are transformed to a common scale, which is crucial for algorithms that rely on distance measurements (e.g., K-Nearest Neighbors, clustering) or gradient-based optimization (e.g., linear regression).
- **Algorithm Performance:** It helps algorithms converge faster and more reliably by avoiding issues where features with larger ranges disproportionately affect the results.
- **Coefficient Comparability:** In regression models, scaling helps in comparing the significance of features by putting them on the same scale.

Types of Scaling:

1. **Normalization (Min-Max Scaling):**
    - **Purpose:** Maps the data to a specific range, typically [0, 1] or [-1, 1].
    - **Formula:** Normalized x=x−min(x)max(x)−min(x)\text{Normalized } x = \frac{x - \text{min}(x)}{\text{max}(x) - \text{min}(x)}Normalized x=max(x)−min(x)x−min(x)

- **Use Case:** Useful when features are on different scales and when the data does not follow a Gaussian distribution.
- **Sensitivity to Outliers:** Affected by outliers since the min and max values can be significantly skewed by them.

2. **Standardization (Z-Score Normalization):**
   - **Purpose:** Transforms the data so that it has a mean of 0 and a standard deviation of 1.
   - **Formula:** Standardized x=x−mean(x)sd(x)\text{Standardized } x = \frac{x - \text{mean}(x)}{\text{sd}(x)}Standardized x=sd(x)x−mean(x)
   - **Use Case:** Ideal when data follows a Gaussian distribution, but also works well for non-Gaussian distributions. It is less affected by outliers as it does not bound the data within a specific range.
   - **Sensitivity to Outliers:** Less affected by outliers compared to normalization.

Key Differences:

- **Range:** Normalization scales data to a fixed range (usually [0, 1]), whereas standardization does not have a fixed range.
- **Application:** Normalization is useful for non-Gaussian distributions and features on different scales, while standardization is suitable for Gaussian distributions and less sensitive to outliers.
- **Impact of Outliers:** Normalization is more affected by outliers, while standardization is less sensitive.

Scaling is essential in data preprocessing to ensure features contribute equally to the analysis and to improve algorithm performance. Normalization and standardization are two common methods, each with distinct applications and characteristics depending on the data distribution and the specific needs of the analysis.

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

The Variance Inflation Factor (VIF) measures how much the variance of an estimated regression coefficient increases due to multicollinearity. It helps in detecting how much a particular independent variable is correlated with other independent variables in the model.

Formula:

VIF=11−R2\text{VIF} = \frac{1}{1 - R^2}VIF=1−R21

where R2R^2R2 is the R-squared value obtained by regressing the independent variable of interest on all other independent variables.

Why Does VIF Become Infinite?

- **Perfect Multicollinearity:** When there is perfect multicollinearity between two or more independent variables, it means that one variable is a perfect linear combination of others. This results in R2=1R^2 = 1R2=1 when the variable is regressed against the others.
- **VIF Calculation:** Given the formula VIF=11−R2\text{VIF} = \frac{1}{1 - R^2}VIF=1−R21, if R2=1R^2 = 1R2=1, the denominator becomes zero, making the VIF value approach infinity. This indicates a perfect correlation, meaning the variable's variance is infinitely inflated due to its perfect correlation with other variables.

Key Points:

- **Implication of Infinite VIF:** An infinite VIF value signifies perfect multicollinearity, where the variable in question can be perfectly predicted from other independent variables, making it redundant.
- **Solution:** To resolve this issue, you should remove one of the perfectly correlated variables to eliminate the multicollinearity.

- VIF becomes infinite when there is perfect multicollinearity among the independent variables, resulting in R2=1R^2 = 1R2=1.
- Perfect correlation causes the VIF formula to yield an infinite value, indicating severe multicollinearity.
- Address this issue by identifying and removing redundant variables to improve the model's stability and interpretability.

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 (Quantile-Quantile) plot is a graphical tool used to compare the quantiles of two probability distributions. It plots the quantiles of one dataset against the quantiles of another dataset or a theoretical distribution.

Uses:

1. **Comparison of Distributions:**
    - To check if two datasets come from the same distribution, you compare their quantiles. If the points fall approximately along a straight line, the datasets are likely from the same distribution.
2. **Normality Check:**
    - In linear regression, it's used to assess whether the residuals (errors) follow a normal distribution, which is an assumption for many statistical tests.

Importance in Linear Regression:

1. **Distribution Consistency:**
    - Ensures that training and testing datasets come from populations with similar distributions. This is crucial for the generalizability of the model.
2. **Validation of Assumptions:**
    - Helps verify if the residuals of the regression model are normally distributed, which affects the validity of statistical tests and confidence intervals.

Interpreting Q-Q Plots:

- **Straight Line:** Points falling along a 45-degree reference line suggest that the datasets have the same distribution.
- **Deviations from Line:**
    - **Above Line:** Data points higher than expected.
    - **Below Line:** Data points lower than expected.
    - **S-Shaped Curve:** Indicates different distribution shapes, such as skewness or heavy tails.

Advantages:

- **Detection of Distributional Aspects:** Identifies shifts in location, scale, symmetry, and presence of outliers.
- **Versatility:** Useful for comparing sample data to theoretical distributions or between different datasets.

A Q-Q plot is essential for validating distribution assumptions in linear regression and ensuring the compatibility of training and testing datasets. It helps in understanding whether the datasets have similar distributions and in diagnosing potential issues with model assumptions.