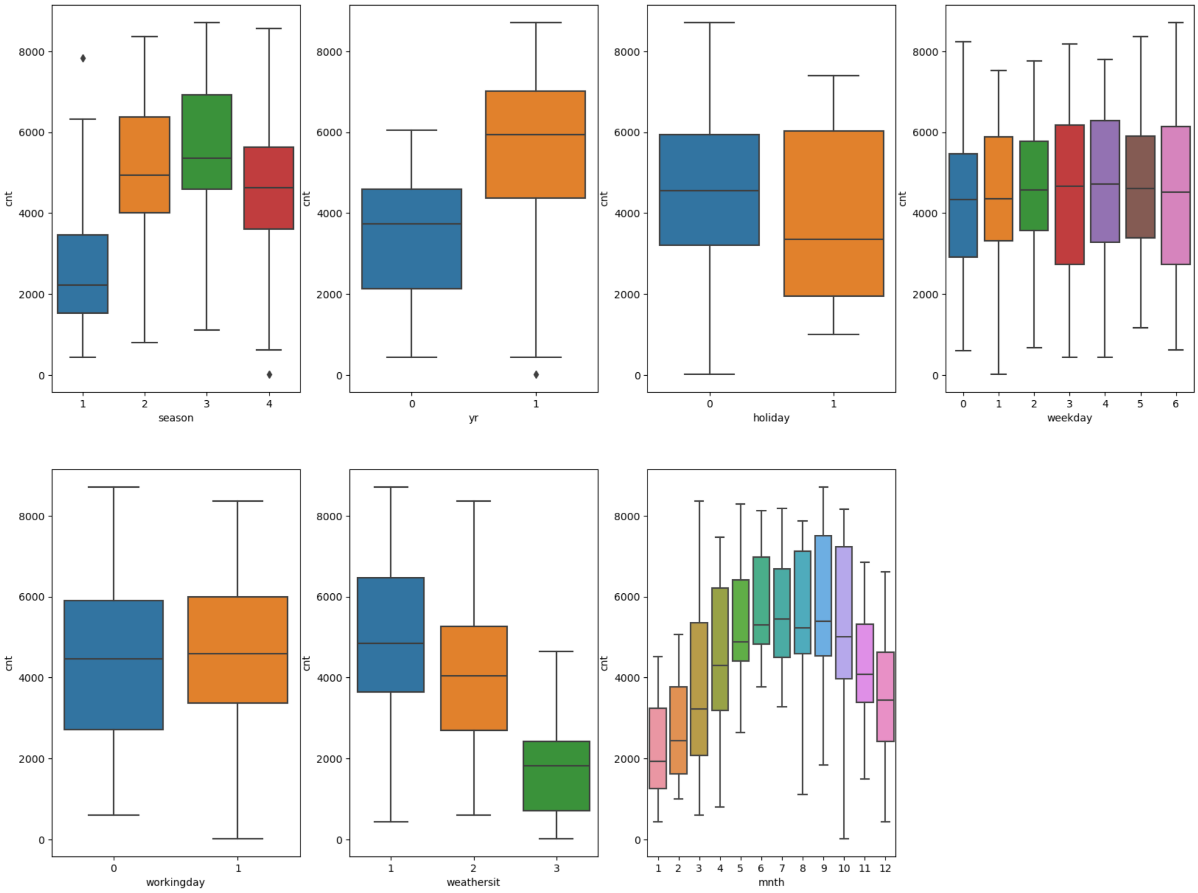
Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: Summary of the effects of categorical variables on the dependent variable, based on the boxplot visualizations:

- Season: Demand for bike rentals is highest during Fall (category 3) and lowest during Spring (category 1).
- Year (Yr): The year 2019 exhibited higher user counts compared to 2018, suggesting a potential trend of increasing demand over time.
- Holiday: Bike rentals tend to decrease during holidays, indicating a reduction in demand during these periods.
- Weekday: Demand for bike rentals remains relatively constant throughout the week, with no significant variation observed across weekdays.
- Workingday: The median count of users remains consistent throughout the week, with similar booking patterns observed on working and non-working days.
- Weather Situation (Weathersit): Demand for bike rentals is significantly lower during adverse weather conditions such as heavy rain or snow, with the highest counts observed during clear or partly cloudy weather.
- Month (Mnth): Rentals peak in September and decrease in December, aligning with weather patterns. Substantial snowfall in December may lead to reduced rentals, influencing user behavior.



Q2. Why is it important to use `drop_first=True` during dummy variable creation?

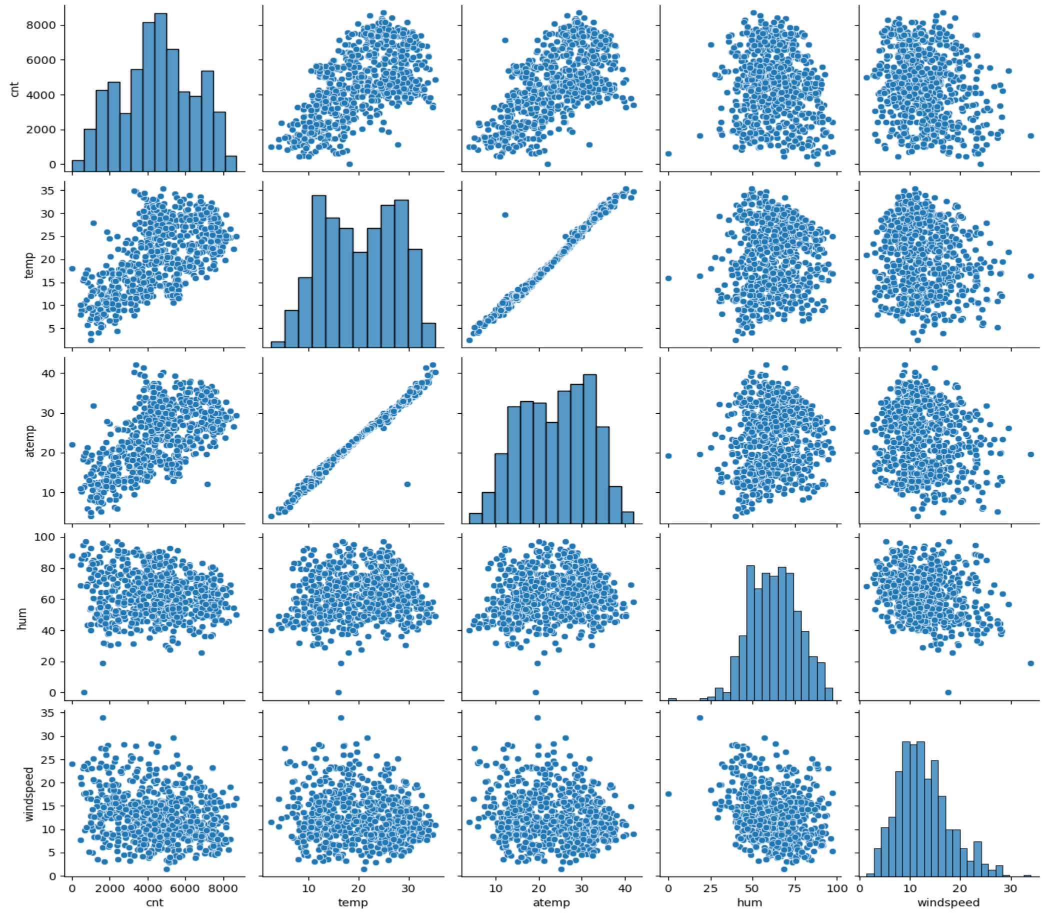
Ans: When generating dummy variables for categorical features, it's crucial to employ **drop_first=True** to maintain model efficiency and accuracy. This parameter ensures that we create n-1 dummy columns for a categorical variable with n levels, effectively avoiding multicollinearity issues and reducing redundancy in the model.

Let's illustrate with an example: consider a categorical variable representing different types of housing conditions, such as furnished, semi-furnished, and unfurnished. By encoding this variable into dummy variables, we aim to capture the presence or absence of each housing condition. However, including dummy variables for all three conditions introduces redundancy. For instance, if a property is neither furnished nor semi-furnished, it's automatically unfurnished. Therefore, including a separate dummy variable for unfurnished would duplicate information already captured by the absence of the furnished and semi-furnished categories.

By dropping the first dummy variable (e.g., unfurnished), we maintain model efficiency while ensuring that each category is uniquely represented. This approach improves the model's interpretability and performance by minimizing correlations among dummy variables and avoiding multicollinearity, ultimately leading to more reliable predictions.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: ‘atemp’ and ‘temp’ had the highest correlation with the target variable.



Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: I have validated the assumptions of the Linear Regression Model based on the following five criteria:

1. **Normality of Error Terms**: I have ensured that the error terms are normally distributed, which is crucial for the accuracy of the model's predictions.
2. **Multicollinearity Check**: I have verified that there is insignificant multicollinearity among the independent variables, ensuring that each predictor contributes unique information to the model.
3. **Homoscedasticity**: I have confirmed that there is no discernible pattern in the residual values, indicating consistent variability across the range of predicted values.
4. **Independence of Residuals**: I have checked for autocorrelation in the residuals to ensure that each observation's error term is independent of those from other observations.
5. **Linear Relationship Validation**: I have validated the linearity among variables, ensuring that the relationship between the independent and dependent variables is adequately captured by a linear model.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

- temp (with coefficient value of - 0.472823)
- season_winter (with coefficient value of - 0.079699)
- yr (with coefficient value of - 0.234361)

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans: Linear regression is a fundamental tool in machine learning used for supervised learning. It helps us understand the relationship between a continuous dependent variable (y) and one or more independent variables (x). The core idea lies in finding a best-fitting straight line through a set of data points.

Formulas Unveiling the Line:

Here's a breakdown of linear regression with key formulas:

1. The Linear Regression Equation:

This equation represents the best-fitting line:

$\hat{y} = \theta_0 + \theta_1 x$

where:

- $\hat{y}$ (pronounced "y hat") is the predicted value of y for a given value of x.
- θ_0 (theta naught) is the intercept, the y-axis value where the line crosses.
- θ_1 (theta one) is the slope, representing the change in y for a unit change in x.

2. Finding the Optimal θ_0 and θ_1:

We aim to minimize the difference between the predicted values ($\hat{y}$) and the actual values (y) of the dependent variable. A common method uses the Mean Squared Error (MSE):

$MSE = (1/n) * \sum(y - \hat{y})^2$

where:

- n is the number of data points.
- Σ (sigma) represents the summation over all data points.

Finding the optimal θ_0 and θ_1 that minimize MSE involves calculus. However, efficient algorithms like gradient descent or the normal equation can solve for these values.

3. Normal Equation for Linear Regression:

This formula provides a closed-form solution for θ_0 and θ_1:

- $\theta_1 = (\Sigma xy - (\Sigma x)(\Sigma y)) / (\Sigma x^2 - (\Sigma x)^2)$
- $\theta_0 = (\Sigma y) - (\theta_1)(\Sigma x)$

where:

- Σxy is the sum of the product of x and y values across all data points.
- Σx and Σy are the sums of x and y values, respectively.
- Σx^2 is the sum of squared x values across all data points.

Simple Linear Regression (SLR): A Single Guiding Light

SLR, a subset of linear regression, focuses on modeling the relationship between a single independent variable (x) and a continuous dependent variable (y). It's the foundation for understanding MLR and is represented by the same formulas mentioned above.

Multiple Linear Regression (MLR): Illuminating Relationships with Many Lights

MLR is an extension of linear regression that tackles the scenario with multiple independent variables (x1, x2, ..., xn) influencing the continuous dependent variable (y). Here, the model attempts to find a hyperplane (a flat, multidimensional plane) that best fits the data points in a higher dimensional space.

MLR Equation:

The equation for MLR reflects this multi-variable influence:

$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$

where:

- The same notations for $\hat{y}$, θ_o, and θ_n apply as in SLR.
- θ_1 to θ_n are the coefficients (slopes) for each respective input variable (x_1 to x xn).

Finding optimal coefficients in MLR involves solving a system of linear equations or using specialized algorithms for higher dimensions.

Q2. Explain the Anscombe’s quartet in detail.

Ans: Anscombe's quartet is a famous set of four datasets that were created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data before drawing conclusions from statistical models. Despite having very similar descriptive statistics, the datasets exhibit vastly different properties when plotted, highlighting the limitations of relying solely on summary statistics.

The quartet consists of four datasets, each containing 11 (x, y) pairs of data points. Here's a brief overview of each dataset:

1. Dataset I: This dataset forms a simple linear relationship between x and y, with a slope of approximately 0.5 and an intercept of around 3. There are no obvious outliers, and the relationship appears to be well-behaved.
2. Dataset II: This dataset also follows a linear relationship, but with a different slope and intercept compared to Dataset I. However, there is one outlier point that significantly deviates from the linear trend, influencing the regression line if fitted.
3. Dataset III: Unlike the previous datasets, Dataset III forms a non-linear relationship, resembling a quadratic curve. The outlier from Dataset II is retained but is not as influential due to the curve's shape.
4. Dataset IV: Dataset IV is characterized by a perfect linear relationship except for one outlier point, which significantly deviates from the linear trend. Removing this outlier would dramatically alter the regression line.

Q3. What is Pearson’s R?

Ans: Pearson's R, denoted by "r", is a statistical measure that quantifies the linear relationship between two continuous variables. It reveals both the strength and direction of this relationship, ranging from -1 to +1.

The Formula:

Pearson's R is calculated using the following formula:

$r = (\Sigma(xy) - (\Sigma x)(\Sigma y)) / (\sqrt{(\Sigma x^2 - (\Sigma x)^2)} * \sqrt{(\Sigma y^2 - (\Sigma y)^2)})$

- Σ (sigma) represents summation over all data points.
- x and y are the individual data points for the two variables.
- xy represents the product of corresponding x and y values for each data point.
- Σx is the sum of all x values.
- Σy is the sum of all y values.
- Σx^2 is the sum of squared x values (x multiplied by itself for each data point and then summed).

- Σy^2 is the sum of squared y values (y multiplied by itself for each data point and then summed).

What the Formula Does:

1. Numerator: It calculates the covariance between x and y. Covariance measures how much the variables tend to vary together. A positive covariance suggests they move in the same direction, while a negative covariance indicates they move in opposite directions.
2. Denominator: It calculates the product of the standard deviations of x and y. The standard deviation measures the spread of data points around the mean.

Result:

- Values closer to +1: Indicate a strong positive correlation. As one variable increases, the other tends to increase as well.
- Values closer to -1: Indicate a strong negative correlation. As one variable increases, the other tends to decrease.
- A value of 0: Indicates no linear correlation between the variables.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Feature scaling is a preprocessing technique used to standardize or normalize the range of independent variables or features in a dataset. It involves transforming the values of features into a similar scale to ensure that no single feature dominates the others in terms of magnitude. Scaling is essential in many machine learning algorithms, particularly those that involve distance-based calculations or optimization algorithms, to improve their performance and convergence.

There are two common methods of feature scaling: normalization and standardization.

1. Normalization:
    - Normalization is suitable when the distribution of the data does not follow a Gaussian (normal) distribution.
    - In normalization, each feature is scaled independently to have values between 0 and 1, or within a specified range.
    - It is useful for algorithms that do not make assumptions about the distribution of the data, such as K-Nearest Neighbors (KNN) and neural networks.
    - Normalization maintains the relative relationships between values in each feature but does not eliminate outliers.
2. Standardization:
    - Standardization is appropriate when the data follows a Gaussian distribution, although it can be applied to any distribution.
    - In standardization, each feature is scaled to have a mean of 0 and a standard deviation of 1.

- Unlike normalization, standardization does not bound the range of values, making it robust to outliers.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: The Variance Inflation Factor (VIF) is a measure used to quantify how much the variance of a coefficient estimate in a regression model is inflated by collinearity among the independent variables. It provides a quantitative assessment of the extent to which the feature variables are correlated with each other, which is crucial for assessing the reliability of the linear model.

The formula for calculating VIF is ***VIF = 1 / (1 - R^2)*** R^2 is the R-square value of the independent variable being assessed for collinearity with other independent variables. If an independent variable can be perfectly explained by other independent variables, its R-square value will be 1, leading to VIF approaching infinity.

Interpreting the VIF values:

- VIF = 1: Indicates that the variable is not correlated with other variables.
- VIF between 1 and 5: Suggests moderate correlation between the variable and other variables.
- VIF greater than 5: Indicates high correlation between the variable and other variables, indicating potential multicollinearity issues.

For example, a VIF of 1.9 means that the variance of a particular coefficient is 90% larger than expected if there was no multicollinearity present. This inflation in variance indicates the degree to which the coefficient's precision is compromised due to collinearity.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans: A Q-Q plot, short for quantile-quantile plot, is a graphical method used to compare the shapes of distributions by plotting the quantiles of one dataset against the quantiles of another dataset. It provides a visual assessment of how well the distributions of two datasets match each other.

In a Q-Q plot, if both sets of quantiles come from the same distribution, the points on the plot will form a line that is approximately straight. This line indicates that the distributions have similar characteristics in terms of location, scale, shape, and tail behavior.

Key questions that a Q-Q plot can help answer include:

- Whether two datasets come from populations with a common distribution.
- Whether two datasets have similar location (central tendency) and scale (spread).
- Whether two datasets exhibit similar shapes in their distributions.
- Whether two datasets have similar tail behavior, indicating the presence of outliers or extreme values.