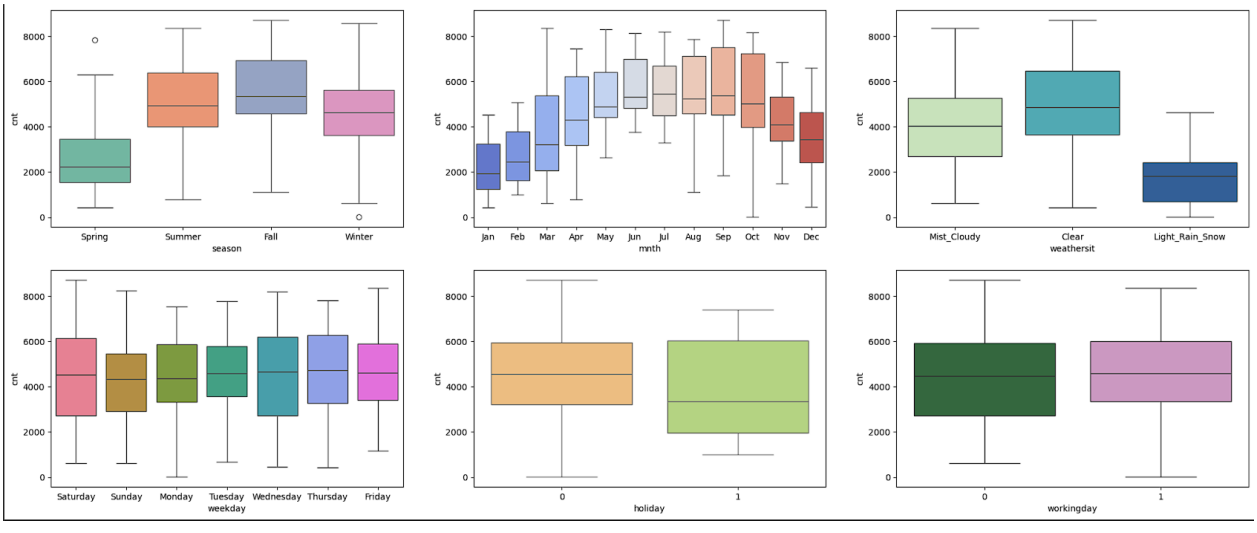
# Assignment-based Subjective Questions

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

Ans -



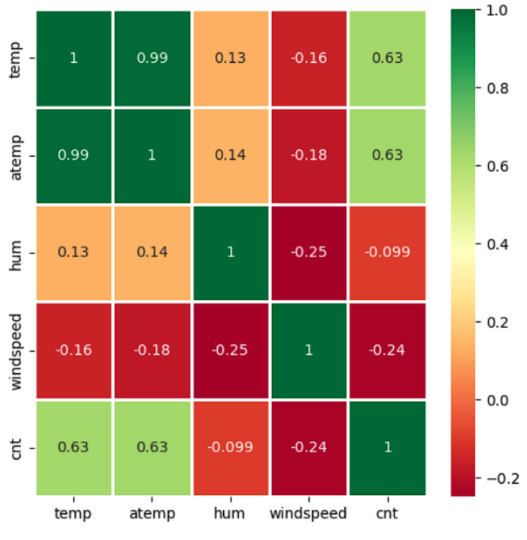

- **Season vs Bike Count**
  - Summer and Fall: These seasons see the highest median bike counts, indicating more bike rentals during these periods.
  - Spring: Has the lowest median and range in bike counts, suggesting fewer rentals in this season.
- **Month vs Bike Count**
  - June to September: These months show higher median bike counts, which aligns with the summer season.
  - January and December: Lower bike counts, consistent with the colder winter months.
- **Weather Situation vs Bike Count**
  - Clear Weather: Predictably, bike rentals are higher in clear weather.

- Mist/Cloudy: There is a slight decrease in the median bike count, indicating that overcast weather might reduce rentals somewhat.
- Light Rain/Snow: This condition significantly reduces the bike count, suggesting that adverse weather conditions strongly discourage bike rentals.

- **Weekday vs Bike Count**
  - Weekdays (Monday to Friday): The median bike count is fairly consistent across the workweek, with a slight dip on Tuesday.
  - Weekends (Saturday and Sunday): Slightly higher bike counts on Saturdays and Sundays compared to weekdays, indicating more leisure bike usage.
- **Holiday vs Bike Count**
  - Non-Holidays: These days have a higher median bike count, likely due to regular commuting.
  - Holidays: The median count is slightly lower, suggesting reduced bike rentals when people aren't commuting to work.
- **Working Day vs Bike Count**
  - Working Days: The bike count is fairly consistent, likely due to regular commutes.
  - Non-Working Days: Slightly lower median, indicating fewer rentals on days off, which might be due to a reduction in commuting-related bike usage.

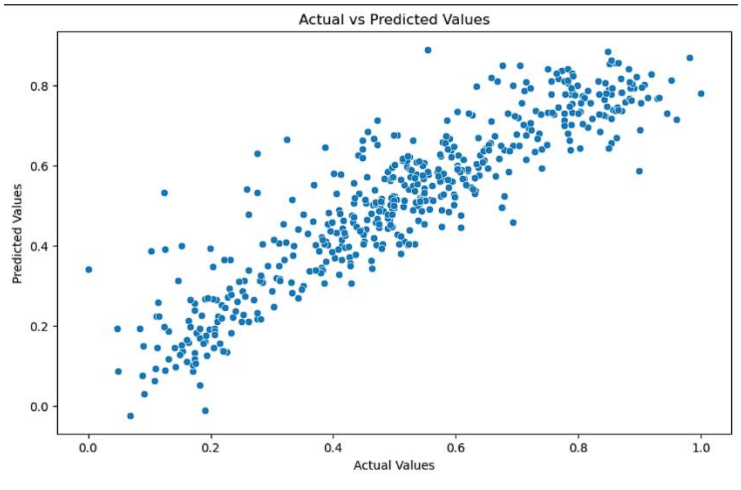
2) Why is it important to use **drop_first=True** during dummy variable creation?
   - If you don't drop the first column then your dummy variables will be correlated (redundant). This may affect some models adversely and the effect is stronger when the cardinality is smaller.
   - For example, iterative models may have trouble converging and lists of variable importances may be distorted.
   - Another reason is, if we have all dummy variables it leads to Multicollinearity between the dummy variables. To keep this under control, we lose one column

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



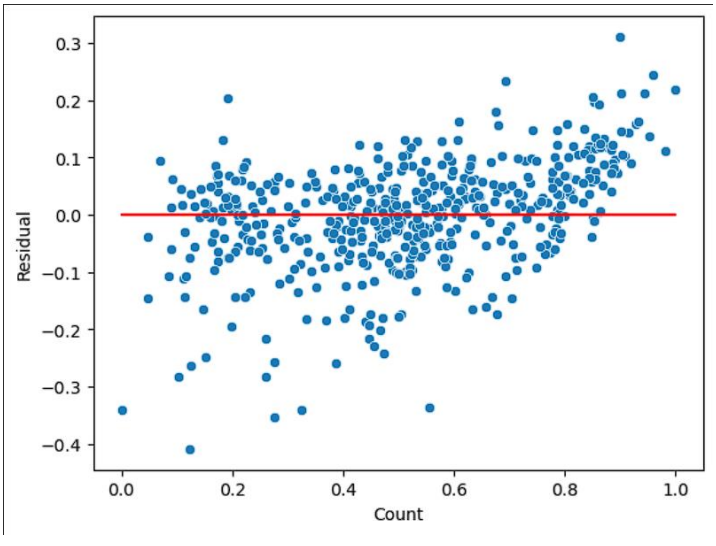
- Temperature (temp and atemp)
    - High Correlation Between temp and atemp: The scatter plot shows a strong positive correlation between temp (actual temperature) and atemp (feels-like temperature). This indicates that these two variables move closely together, as expected.
    - Relationship with cnt (count): Both temp and atemp show a positive correlation with cnt (bike rentals). As temperature increases, the number of bike rentals also tends to increase, suggesting that more favorable weather encourages bike usage.
- Humidity (hum)
    - No Strong Correlation with cnt: The scatter plot between hum and cnt does not show a clear linear relationship, indicating that humidity does not have a strong direct impact on the number of bike rentals.
    - Distribution of Humidity: The histogram for hum shows that the data is relatively uniformly distributed, with a slight concentration in the 60-80 range.
- Windspeed
    - No Strong Correlation with cnt: Similar to humidity, windspeed does not show a strong correlation with bike rentals. The scatter plot suggests that windspeed might not significantly impact bike usage.
    - Windspeed Distribution: The histogram shows that most windspeed values are concentrated at lower speeds, with very few occurrences at higher wind speeds.
- cnt (Count of Bike Rentals)
    - Overall Distribution: The histogram for cnt shows that bike rentals are fairly normally distributed, with a slight skew towards the lower end. This suggests that while most days see moderate bike rentals, there are fewer days with extremely high or low rentals.
    - Relationship with Weather Conditions: The pair plots show that weather conditions like temperature have a more significant impact on bike rentals compared to other factors like humidity and windspeed.

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

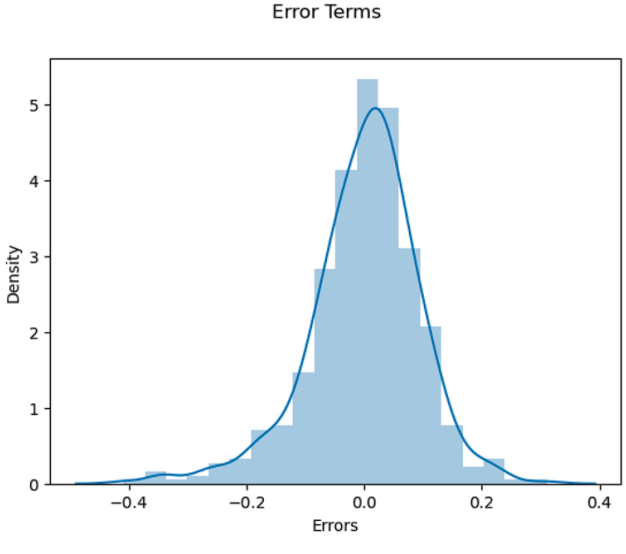


- The linearity is validated by looking at the points distributed symmetrically around the diagonal line of the actual vs predicted plot as shown in the above figure.

Homoscedasticity




- Homoscedasticity refers to the condition where the variance of the errors (residuals) is constant across all levels of the independent variables.
- It is a key assumption in linear regression that ensures that the model's residuals are evenly distributed around zero, without showing patterns or trends.
- There is no visible pattern in residual values shown in the above graph, thus homoscedasticity is well preserved




**Normality of Residuals:**

- The histogram of the residuals should ideally resemble a normal distribution (bell-shaped curve). This indicates that the errors are normally distributed, which is a key assumption in linear regression.

**Centering Around Zero:**

- The residuals should be centered around zero, meaning the model does not systematically overestimate or underestimate the target variable. This suggests that the model's predictions are unbiased.

**Spread of Residuals:**

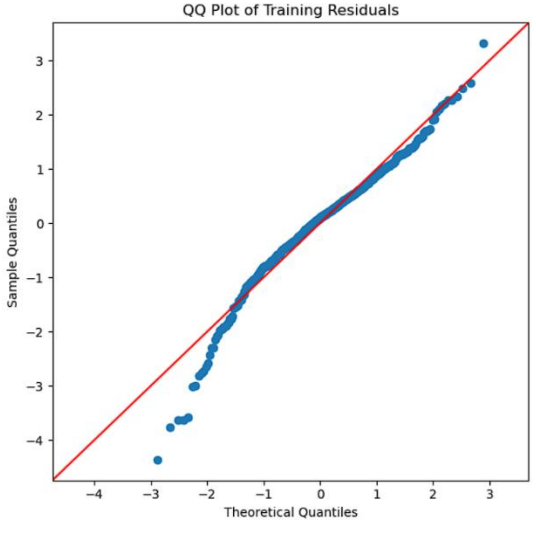
- The spread of the residuals (width of the histogram) provides information about the variance of the errors. A smaller spread indicates that the model's predictions are more precise.

**Detection of Outliers:**

- If the histogram shows extreme values far from zero, this could indicate the presence of outliers or high-leverage points, which may need further investigation.

**Symmetry:**

- The histogram should be roughly symmetric. Asymmetry could suggest that the error terms are skewed, indicating potential issues with the model's assumptions.




**Assessing Normality:**

- The Q-Q plot is used to assess whether the residuals follow a normal distribution. In a Q-Q plot, the residuals are plotted against a theoretical normal distribution. If the residuals are normally distributed, the points should fall approximately along the reference line (usually a 45-degree line).

**Model Diagnostics:**

- The Q-Q plot helps in diagnosing the model fit. If the residuals are not normally distributed, it could indicate that the model is not appropriate for the data, and alternative models or transformations of the data might be needed.

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

- **Temperature (temp):** With a coefficient of 0.564438, temperature is the most influential factor in the model. A higher temperature strongly correlates with an increase in bike rentals, suggesting that warmer weather encourages more people to rent bikes.
- **Light Rain or Snow (Light_rainsnow):** This feature has the most significant negative impact on bike rentals, with a coefficient of -0.307082. Adverse weather conditions like light rain or snow considerably reduce the number of bike rentals, making it a crucial factor to consider when predicting demand.
- **Year (yr):** The year variable, with a coefficient of 0.230252, indicates that bike rentals are expected to increase over time. This reflects a potential growth trend in the popularity of bike-sharing, making it one of the top contributors to demand.

# General Subjective Questions

1) Explain the linear regression algorithm in detail

**1. Introduction to Linear Regression:** Linear regression is a fundamental algorithm in machine learning used to model the relationship between a dependent variable (target) and one or more independent variables (features). The goal is to find the best-fitting straight line (in simple linear regression) or hyperplane (in multiple linear regression) that predicts the target variable based on the input features.

**2. The Mathematical Formula:** The basic form of a linear regression equation is:
Y=β0+β1X1+β2X2+⋯+βnXn+ϵY = \beta_0 + \beta_1X_1 + \beta_2X_2 + \dots + \beta_nX_n + \epsilonY=β0+β1X1+β2X2+⋯+βnXn+ϵ

- **YYY**: The predicted value (dependent variable).
- **X1,X2,...,XnX_1, X_2, \dots, X_nX1,X2,...,Xn**: The independent variables (features).
- **β0\beta_0β0**: The intercept, representing the predicted value when all features are zero.
- **β1,β2,...,βn\beta_1, \beta_2, \dots, \beta_nβ1,β2,...,βn**: The coefficients that represent the slope of the line, showing how much YYY changes with a one-unit change in each XXX.
- **ϵ\epsilonϵ**: The error term, representing the difference between the actual and predicted values.

**3. How the Algorithm Works:**

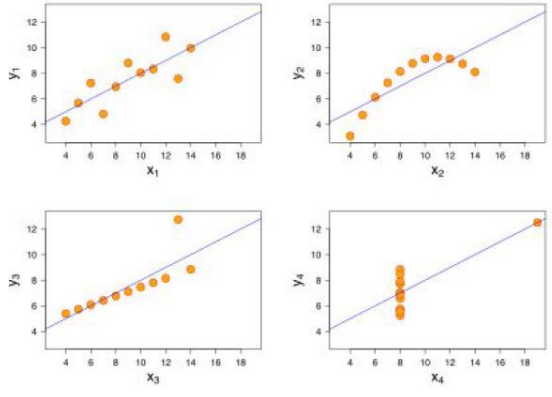
- **Training the Model**: During training, the algorithm finds the best values for the coefficients (β\betaβs) by minimizing the difference between the actual and predicted values. This is done using a method called **Ordinary Least Squares (OLS)**, which minimizes the sum of the squared differences between the actual and predicted values (called residuals).
- **Prediction**: Once the model is trained, you can use it to predict the target variable for new data by plugging the feature values into the equation.

**4. Assumptions of Linear Regression:**

- **Linearity**: The relationship between the dependent and independent variables should be linear.
- **Independence**: Observations should be independent of each other.
- **Homoscedasticity**: The variance of the error terms should be constant across all levels of the independent variables.
- **Normality**: The error terms should be normally distributed.

**5. Evaluation**: The performance of a linear regression model is typically evaluated using metrics like **Mean Squared Error (MSE)**, **R-squared**, and **Adjusted R-squared**. These metrics tell us how well the model explains the variation in the target variable and how accurate the predictions are. In summary, linear regression is a simple yet powerful tool for predicting a continuous outcome based on one or more features. It's widely used in many applications because it's easy to understand, interpret, and implement.

2) Explain the Anscombe's quartet in detail.




**1. Introduction to Anscombe's Quartet:** Anscombe's Quartet is a set of four datasets that have nearly identical statistical properties (mean, variance, correlation, etc.), yet they have very different distributions and appear very different when graphed. The quartet was created by statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before analyzing it.
**2. The Four Datasets:** Each of the four datasets in Anscombe's Quartet consists of 11 (x, y) pairs. Despite having similar statistical summaries, they illustrate different types of data relationships and outliers that cannot be captured through summary statistics alone.
Here's a summary of the key similarities between the datasets:

- **Mean of X**: All datasets have a mean of 9 for the x-values.
- **Mean of Y**: All datasets have a mean of 7.5 for the y-values.
- **Variance**: The variance of the x-values and y-values is the same across all datasets.
- **Correlation**: The correlation coefficient between x and y is about 0.82 in all datasets.
- **Linear Regression Line**: Each dataset has a linear regression line with an almost identical equation, y=3+0.5xy = 3 + 0.5xy=3+0.5x.

**3. The Importance of Visualization:** Although the four datasets have the same statistical properties, their scatter plots look very different:

- **Dataset 1**: Shows a typical linear relationship between x and y. This is what you'd expect to see if you were fitting a simple linear regression model.
- **Dataset 2**: Has a clear nonlinear relationship. The y-values increase as x increases, but not in a straight line. A linear model would not be appropriate here.
- **Dataset 3**: Contains an outlier that strongly influences the regression line. Without this outlier, the data would show no correlation. This dataset emphasizes how outliers can affect statistical measures like correlation and regression coefficients.
- **Dataset 4**: All the data points are vertically aligned except for one outlier. This creates a situation where the correlation appears strong, but the relationship between x and y is not actually linear.

3) What is Pearson's R?

**1. Definition:** Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two continuous variables. It quantifies how strongly the variables are related and the direction of that relationship. The value of Pearson's R ranges from -1 to 1.

**2. Interpretation of Values:**

- **+1**: A Pearson's R value of +1 indicates a perfect positive linear relationship between the two variables. As one variable increases, the other also increases proportionally.
- **0**: A value of 0 indicates no linear relationship between the variables. Changes in one variable do not predict changes in the other.
- **-1**: A value of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other decreases proportionally.

**3. Formula:** Pearson's R is calculated using the following formula:

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

Where:

- $X_i$ and $Y_i$ are the individual data points for the two variables.
- $\overline{X}$ and $\overline{Y}$ are the mean values of the two variables.
- The numerator measures the covariance between the variables, while the denominator normalizes this by the product of the standard deviations of the two variables.

**4. Applications:** Pearson's R is widely used in statistics, data analysis, and machine learning to assess the strength and direction of the linear relationship between two variables. It helps in understanding how one variable can be predicted from another.

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

**1. What is Scaling?** Scaling is the process of transforming the features of your data so that they fall within a specific range or have a specific distribution. This is especially important in machine learning models where features with different units or scales can disproportionately influence the model's performance.

**2. Why is Scaling Performed?**

- **Improving Model Performance:** Many machine learning algorithms, especially those based on distance (like K-nearest neighbors, SVMs) and gradient descent (like linear regression, neural networks), perform better when the input features are on a similar scale.
- **Convergence Speed:** For optimization algorithms, scaling helps in achieving faster convergence during training by ensuring that features contribute more uniformly to the cost function.

- **Consistency:** It ensures that no single feature dominates others simply because of its larger scale, leading to more reliable and interpretable models.

**3. Difference Between Normalized Scaling and Standardized Scaling:**

- **Normalized Scaling:**
  - **Definition:** Normalization (Min-Max Scaling) transforms the data to a fixed range, usually [0, 1]. It scales the data based on the minimum and maximum values of each feature.
  - **Formula:**

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

  - **Use Case:** Normalization is useful when you need to ensure that all features have the same scale, particularly in algorithms that don't assume normally distributed data (e.g., K-means clustering, neural networks).
- **Standardized Scaling:**
  - **Definition:** Standardization (Z-score normalization) transforms the data so that it has a mean of 0 and a standard deviation of 1. It adjusts the data based on the mean and standard deviation of the feature.
  - **Formula:**

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

  - **Use Case:** Standardization is typically used when you need the data to have a standard normal distribution (mean = 0, standard deviation = 1). This is important for algorithms like Principal Component Analysis (PCA), linear regression, and SVMs.

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

**1. Understanding VIF:** Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in a set of multiple regression variables. It quantifies how much the variance of a regression coefficient is inflated due to the correlation between the independent variables.

**2. Why VIF Can Be Infinite:** A VIF value becomes infinite when there is perfect multicollinearity between one independent variable and a combination of the other independent variables. This means that one variable is an exact linear combination of the others, leading to the denominator in the VIF calculation becoming zero.

**3. How It Happens:**

- **Perfect Collinearity:** When an independent variable can be perfectly predicted from one or more of the other independent variables, the correlation matrix becomes singular, making

the determinant of the matrix zero. This causes the VIF calculation to produce an infinite value.

- **Example:** If you have two variables where one is a perfect multiple of the other (e.g., X2 = 2*X1), the VIF for both variables will be infinite because they provide redundant information to the model.

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

1. **Assessing Normality:**
    - The Q-Q plot is used to assess whether the residuals follow a normal distribution. In a Q-Q plot, the residuals are plotted against a theoretical normal distribution. If the residuals are normally distributed, the points should fall approximately along the reference line (usually a 45-degree line).
2. **Identifying Deviations from Normality:**
    - **Systematic Deviations:** If the points deviate systematically from the line (e.g., a curve), it suggests that the residuals are not normally distributed. This could indicate issues like skewness or kurtosis in the residuals.
    - **Heavy Tails:** If the points at the ends of the plot (the tails) deviate significantly from the line, it suggests that the residuals have heavier tails than a normal distribution, meaning there are more extreme values than expected.
3. **Detecting Outliers:**
    - Points that fall far from the reference line, especially towards the ends of the plot, could indicate the presence of outliers in the data. These outliers can influence the model's predictions and might require further investigation.
4. **Model Diagnostics:**
    - The Q-Q plot helps in diagnosing the model fit. If the residuals are not normally distributed, it could indicate that the model is not appropriate for the data, and alternative models or transformations of the data might be needed.
5. **Implications for Inference:**
    - Normality of residuals is crucial for the validity of statistical tests and confidence intervals in linear regression. Deviations from normality, as indicated by the Q-Q plot, might affect the reliability of these inferences.
6. **Handling Non-Normal Residuals:**
    - If the Q-Q plot shows that the residuals are not normally distributed, consider using techniques like data transformation (e.g., log transformation), adding polynomial terms, or switching to a more robust regression method.