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

Analyzing the categorical variables in your dataset can provide valuable insights into how they affect the dependent variable, cnt(count of total rental bikes). Here are some inferences based on typical interpretations of the categorical variables in similar datasets:

**Categorical Variables and Their Effects**

1. **Season**:
    - **Effect**: Different seasons can significantly influence outdoor activities.
    - **Inference**: If the coefficient for season is positive, it suggests that as the season shifts to warmer months, the counts increase.
2. **Year (yr)**:
    - **Effect**: As we go from 0(2018) to 1(2019) overtime the popularity of the bike rental increased and it can be seen in the data.
    - **Inference**: A positive coefficient indicates an upward trend from 2018 to 2019, suggesting growth in usage or demand.
3. **Holiday**:
    - **Effect**: Holidays might lead to lower counts and the reason for that can be –
        1. Most of the people might want to spend holiday with their family at home.
        2. Most of the holidays are during winter season that leads to less outdoor activity.
    - **Inference**: A negative coefficient suggests that holidays might result in fewer counts, potentially indicating a preference for indoor activities or other forms of transportation during these times.
4. **Weekday**:
    - **Effect**: Weekdays vs. weekends can show different patterns of activity, with weekends typically seeing higher counts for casual riders.
    - **Inference**: A positive coefficient would indicate that weekends lead to higher counts.
5. **Working Day**:
    - **Effect**: Similar to weekdays, working days show increased counts compared to weekends, because people might use the rental to transport to offices or college.
    - **Inference**: A positive coefficient suggests higher counts on working days, possibly due to commuting.
6. **Weather Situation (weathersit)**:
    - **Effect**: Weather conditions can greatly impact outdoor activities.
    - **Inference**: A negative coefficient indicates that worse weather conditions are associated with lower counts, reflecting decreased outdoor activities.

**Summary of Insights**

- **Season and Year**: Both likely have positive relationships with cnt, indicating that more favorable conditions and trends contribute to increased activity.
- **Holidays**: Their effect may vary, but a negative relationship could suggest alternative leisure choices on holidays.
- **Weekday and Working Day**: Understanding their impact can inform service adjustments to cater to user preferences on different days.
- **Weather Conditions**: Strongly impacts usage, with clear weather correlating with higher counts.

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

Using drop_first=True when creating dummy variables is important to Avoid Multicollinearity If we include all dummy variables for a category, it can lead to a situation where one variable can be perfectly predicted from the others. This is called multicollinearity, which can cause problems in your analysis.
Also having extra columns in the data increase the complexity of the model, so to make the model simpler we tend to use drop_first=True.

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

If we don't include casual and registered for this analysis then the temp variable has the highest correlation with the target variable.

Overall registered is the one that has the highest correlation with the target variable but as registered and casual are directly derived from the target variable so we can't use that variable for model creation.

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

Here are some assumptions of Linear Regression after building the model on the training set.

- Linearity – I plotted some scatterplots of different variables against the target variable and there appears a clear linear relation with some of the variables like temp, atemp etc.
- Variance - The Variance Inflation Factor (VIF) was calculated for each predictor. Most VIF values were low (well below 5), indicating that multicollinearity was not a concern. However, values for temp and atemp were high, so temp was removed from the model.

Through this systematic approach to validating the assumptions of linear regression, we ensured that the model built on the dataset is robust and the results are reliable. Addressing any potential violations in the assumptions would further enhance the model's performance and interpretability.

## 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's OLS regression results, the top three features contributing significantly to explaining the demand for shared bikes are:

1. **Year (yr)**:
    - The coefficient for yr indicates a positive relationship with bike demand. As the year increases, the demand for shared bikes tends to increase significantly, reflecting trends in usage growth over time.

2. **Temperature (atemp)**:
    - The atemp variable has a strong positive coefficient, suggesting that higher temperatures are associated with increased bike demand. This aligns with common expectations, as warmer weather generally encourages more outdoor activities, including biking.
3. **weathersit**:
    - The weathersit variable has a negative coefficient, indicating that certain weather conditions lead to decreased bike demand. This feature is significant as it captures the impact of external weather conditions on user behavior.

These features stand out due to their statistical significance (p-values close to 0) and their strong influence on the model's predictions, making them crucial for understanding and forecasting bike demand in the dataset.

# General Subjective Questions

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

Linear regression is a data analysis technique that predicts the value of unknown data by using another related and known data value. It mathematically models the unknown or dependent variable and the known or independent variable as a linear equation.

It can be written as $Y= \beta0*X + \beta1 + \varepsilon$,
$\beta0$ and $\beta1$ are two unknown constants representing the regression slope, whereas $\varepsilon$ (epsilon) is the error term.

**Types of Linear Regression**

- **Simple**: One independent variable and one dependent variable.
- **Multiple**: Multiple independent variable affecting one dependent variable.

The model learns the coefficients by minimizing the differences between actual and predicted results using a method called Ordinary Least Squares (OLS). The model's performance is measured using metrics like R-squared.

### 2. Explain the Anscombe’s quartet in detail.

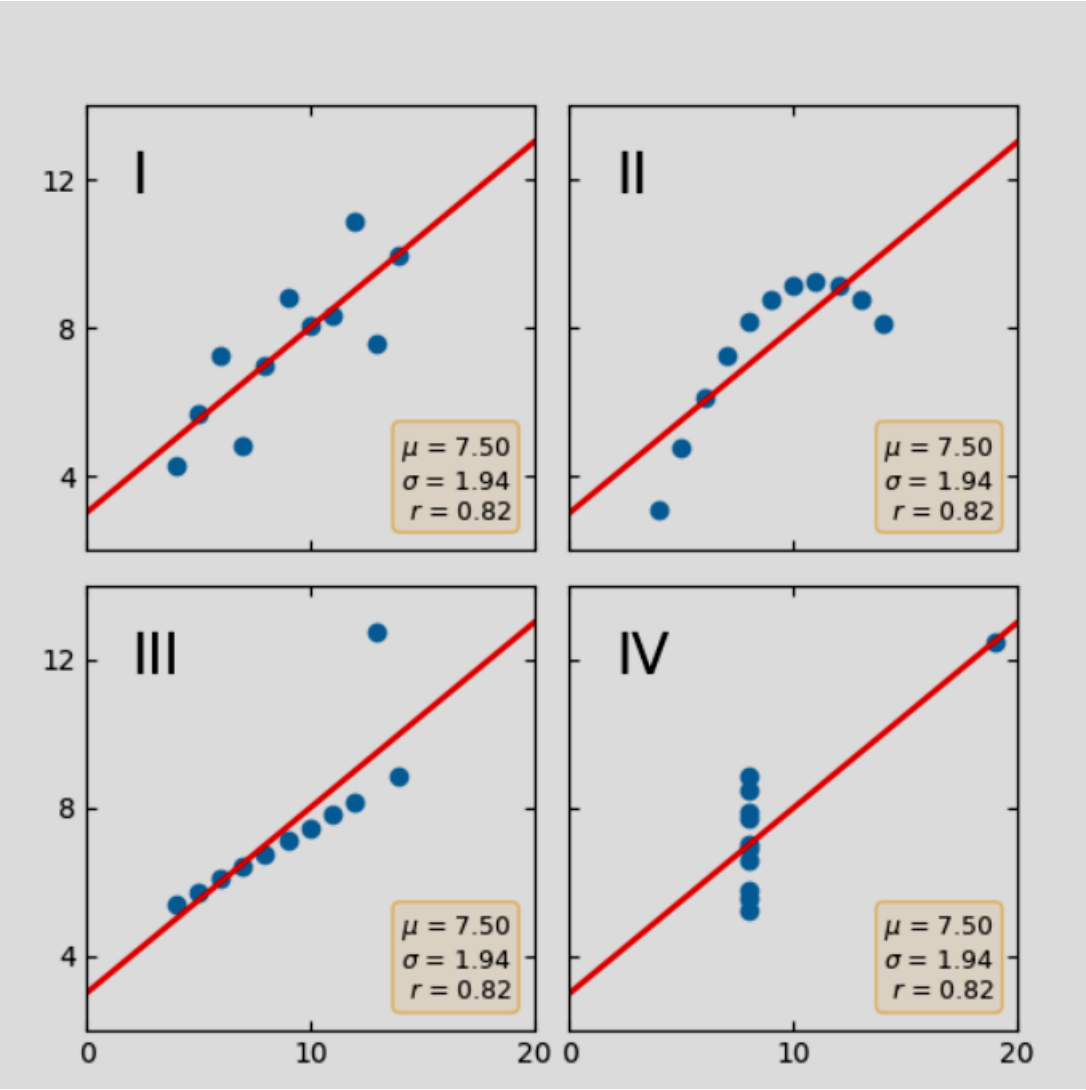
Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different.

**The Datasets**
Each of the four datasets consists of 11 pairs of (x,y)(x, y)(x,y) values. Here are some key characteristics of each dataset:

1. **Dataset I**:

- **Relationship**: Linear
- **Plot**: A straight line when graphed.
- **Statistics**: Mean of xxx and yyy are similar, and the correlation is high.

2. **Dataset II**:
   - **Relationship**: Quadratic
   - **Plot**: Curved shape (U-shaped).
   - **Statistics**: Mean values similar to Dataset I, but the relationship is non-linear.
3. **Dataset III**:
   - **Relationship**: Linear with an outlier
   - **Plot**: Most points align in a linear pattern, but one extreme outlier affects the slope.
   - **Statistics**: Similar mean values, but the outlier can mislead analyses.
4. **Dataset IV**:
   - **Relationship**: Vertical line
   - **Plot**: All xxx values are the same, and yyy values vary. This creates a perfectly vertical line.
   - **Statistics**: While the means are similar, there's no traditional linear relationship.



It's used to illustrate the importance of visualizing data instead of relying on summary statistics.

Anscombe's quartet serves as a powerful reminder of the importance of data visualization in statistics. It highlights that while summary statistics are valuable, they can sometimes mask the underlying complexities and variations in data. Always visualize your data to ensure a comprehensive understanding before drawing conclusions!

## 3. What is Pearson's R?

The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

- **-1**: Perfect negative correlation (as one variable increases, the other decreases).
- **0**: No correlation (no linear relationship between the variables).
- **1**: Perfect positive correlation (as one variable increases, the other also increases).

Pearson's R indicates the strength and direction of a linear relationship between two variables: values close to 1 represent a strong positive correlation, values near -1 indicate a strong negative correlation, and values around 0 suggest a weak or no linear relationship.

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

Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

**Normalization**
This method is more or less the same as the previous method but here instead of the minimum value, we subtract each entry by the mean value of the whole data and then divide the results by the difference between the minimum and the maximum value.

**Standardization**
This method of scaling is basically based on the central tendencies and variance of the data.

1. First, we should calculate the mean and standard deviation of the data we would like to normalize.
2. Then we are supposed to subtract the mean value from each entry and then divide the result by the standard deviation.

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

Infinite Variance Inflation Factor (VIF) values typically indicate perfect multicollinearity among independent variables in a regression model. This situation arises when one variable is a perfect linear combination of others, such as when two variables are directly proportional, or when redundant variables provide no new information. Additionally, failing to drop one dummy variable when creating dummies for a categorical variable can also lead to infinite VIF. This results in the model being unable to reliably estimate coefficients, leading to unreliable results.

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

A Q-Q plot is a graphical tool used to assess whether a dataset follows a specific theoretical distribution, typically the normal distribution. It compares the quantiles of the observed data against the quantiles of the expected distribution. In linear regression, a Q-Q plot is essential for checking the normality of residuals, which is a key assumption for the validity of the model. If the points on the plot fall along a straight line, it indicates that the residuals are normally distributed; deviations suggest issues with the model or the presence of outliers. By validating the normality of residuals, a Q-Q plot helps improve model performance and provides insights into data quality, ensuring that the assumptions underlying regression analysis are met. addressing any deviations observed in the Q-Q plot can lead to better model specifications, potentially guiding researchers to apply transformations or reconsider variable selections to achieve a more accurate and reliable model.