# 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?

- **Season:** The box plots indicate that the count of total rentals varies significantly across different seasons, with higher median rentals in certain seasons (likely spring and summer).
- **Year (yr):** There's a noticeable difference in the number of rentals between the two years, with one year showing higher median rentals than the other.
- **Month (mnth)**: Similar to seasons, the number of rentals varies across months, reflecting seasonal trends.
- **Holiday:** The box plots show that non-holiday days tend to have higher counts of rentals compared to holiday days, which is intuitive as more people might use bikes for commuting on regular days.
- **Weekday:** The distribution of rentals across weekdays does not show significant variation, suggesting that the day of the week may not strongly influence rental counts.
- **Workingday**: Working days seem to have a slightly higher median rental count compared to non-working days, which aligns with the expectation that bikes are used more for commuting on working days.
- **Weather Situation (weathersit):** The box plots indicate that the count of rentals varies with different weather conditions, with clear days having higher rental counts compared to days with adverse weather conditions.

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

Using drop_first=True when creating dummy variables is crucial for the following reasons:

- **Preventing multicollinearity:** This can distort regression analysis by creating a perfect linear relationship among independent variables.
- **Model Interpretability:** It also enhances model interpretability by setting a reference category, making the effects of other categories easier to understand in comparison.

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

temp

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

- **Linearity:** The bivariate analysis showing a linear pattern between few predictors and the target validates the linearity assumption.
- **Normality of Residuals:** The nearly normal distribution of residuals, as seen in the histogram and confirmed by the Q-Q plot, supports the normality assumption.
- **Homoscedasticity:** The absence of a clear pattern or funnel shape in the scatter plot of residuals against predicted values indicates constant variance (homoscedasticity) of the residuals.
- **Independence of Residuals:** The random scatter of residuals around the horizontal axis in the scatter plot suggests that residuals are independent and not correlated with each other or with the independent variables.
- **No Multicollinearity:** I assessed multicollinearity among the chosen independent variables using Variance Inflation Factor (VIF) values, considering a VIF above 5 as indicative of significant multicollinearity.

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

- Temperature (temp)
- Year (yr)
- weathersit_Light Precipitation (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)

# General Subjective Questions

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

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The core idea is to find the linear relationship that best predicts the dependent variable from the independent variables.

### The linear equation

The linear equation for a simple linear regression (with one independent variable) is represented as:
$\mathbf{y = \beta_0 + \beta_1 x + \varepsilon}$
where
y is the dependent variable,
x is the independent variable,
$\beta_0$ is the y-intercept,
$\beta_1$ is the slope of the line (representing the change in y for a one-unit change in x),
ε is the error term (the difference between the observed and predicted values).

For multiple linear regression (with more than one independent variable), the equation expands to:
$\mathbf{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta n x n + \varepsilon}$

### Estimating Coefficients

The coefficients ($\beta_1$, $\beta_2$, etc.) are estimated during the training process. The most common method for this is the Ordinary Least Squares (OLS) approach, which minimizes the sum of the squared differences between the observed and predicted values (the sum of squared residuals).

### Steps in the algorithm

1. Model Specification: Define the dependent and independent variables.
2. Coefficient Estimation: Use OLS or another estimation technique to calculate the coefficients that minimize the sum of squared residuals.
3. Model Fitting: Apply the estimated coefficients to the linear equation to make predictions.

4. Assumption Validation: Check the key assumptions of linear regression (linearity, normality of residuals, homoscedasticity, independence, and no multicollinearity) to ensure the model is valid.
5. Interpretation: Analyze the coefficients to understand the impact of each independent variable on the dependent variable.
6. Prediction: Use the fitted model to predict the dependent variable from new data.

**Model Evaluation**

Model performance can be evaluated using metrics like R-squared and mean squared error (MSE) for the residuals.

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

Anscombe's quartet consists of four distinct datasets, each with eleven points, crafted by Francis Anscombe in 1973 to highlight the importance of data visualization alongside summary statistics. Despite having nearly identical means, variances, correlations, and regression lines, their plots reveal markedly different distributions:

1. Dataset 1 displays a classic linear relationship, aligning well with linear regression assumptions.
2. Dataset 2 shows a clear non-linear pattern, demonstrating that a linear model might not always be suitable.
3. Dataset 3 appears linear but is significantly influenced by an outlier, affecting the regression line.
4. Dataset 4 features a vertical cluster of points with one distant outlier, heavily skewing the regression line.

From these datasets, several key lessons emerge:

- Graphical analysis is crucial as it can uncover underlying patterns and anomalies that summary statistics might not reveal.
- Outliers can significantly alter the results of our analysis, potentially leading to misleading conclusions.
- Identical statistical measures can represent vastly different datasets, highlighting the limitations of relying solely on these summaries.
- It's essential to ensure that the data meets the assumptions of our analytical models, such as linearity and homoscedasticity, to avoid invalid conclusions.

## 3. What is Pearson's R?

Pearson's R, also known as Pearson's correlation coefficient, is a statistical measure that quantifies the linear correlation between two variables, ranging from -1 to +1. A value of +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 implies no linear correlation.

*The formula for Pearson's R is:*
$r = \Sigma(xi - \bar{x})(yi - \bar{y}) / sqrt[\Sigma(xi - \bar{x})^2\Sigma(yi - \bar{y})^2]$
*where xi and yi are the individual sample points indexed with i, $\bar{x}$ and $\bar{y}$ are the mean values of the two sets of data.*

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

Scaling is a technique used in data preprocessing to adjust the range of variable values.

Scaling adjusts the range of data values to ensure equal contribution to machine learning models, enhancing performance and convergence speed, especially for distance-based and gradient descent algorithms.

- **Normalized Scaling:** rescales features to a 0-1 range, ideal for bounding variables but sensitive to outliers.
- **Standardized Scaling:** centers features around mean 0 with standard deviation 1, reducing outlier impact and fitting algorithms assuming normal distribution.

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

The Variance Inflation Factor (VIF) measures how much the variance of an estimated regression coefficient increases if the predictors are correlated. A VIF value becomes infinite when there is perfect multicollinearity in the data, meaning at least one independent variable is an exact linear combination of others.

For instance, if we have two variables X1 and X2 in our model, and X2 = aX1 + b , the VIF for X1 and X2 could be infinite.

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

A Quantile-Quantile plot, or Q-Q plot, is a type of scatter plot that compares two sets of quantiles against one another. The first set of quantiles comes from the data whose distribution we're testing, and the second set represents the quantiles from a theoretical distribution we want to compare it to.

In linear regression, a Q-Q plot is vital for checking the normal distribution of residuals, key for valid statistical tests and confidence intervals. It helps identify outliers that might skew the model and checks for equal variance (homoscedasticity) in residuals. Thus, a Q-Q plot is essential for confirming model assumptions, guiding necessary data transformations, and ensuring the model's reliability.