# 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)**

   From your analysis of the categorical variables from the dataset:

   - the bike rental rates are increased in Summer and Falls.
   - It increased around 50% in 2019.
   - the bike rental rates are increases from May to Oct
   - the bike rental rates are higher on holidays (holidays=0)
   - the bike rental rates are higher weathersit then Mist+ Cloudy and decreases in Light Snow

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

   Using drop_first=True in dummy variable creation avoids multicollinearity and improves interpretability by removing one redundant dummy variable.

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

The temp and atemp variables have the highest correlation with the target variable.

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, several techniques were used:

1. Residual Analysis: The error distribution of residuals was examined. This involved analyzing the distribution of the differences between the actual values and the predicted values. By checking for normality and symmetry in the residual plot, we can assess if the assumption of normally distributed errors holds.
2. Linear Relationship: A scatter plot was created between one independent variable and the dependent variable. By visually inspecting the plot, we could observe if a straight line passing through the points could be seen. This helps us assess whether there is a linear relationship between the variables.

3. Homoscedasticity: The variance of error terms was examined. If the variance of error terms is constant across different levels of the independent variable, it indicates homoscedasticity. This assumption was checked to ensure that the variability of errors does not change systematically as we move along the range of the independent variable.
4. Absence of Multicollinearity: To check for multicollinearity, a heatmap and Variance Inflation Factor (VIF) were used. The heatmap visually represents the correlations between independent variables, helping to identify potential multicollinearity issues. The VIF provides a numerical measure of multicollinearity by quantifying how much the variance of the estimated regression coefficient is inflated due to multicollinearity.

By applying these techniques, we can assess whether the assumptions of Linear Regression are met and validate the model's performance on the training set.

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 size of the coefficients, are:

1. ‘Temperature (temp)’: Coefficient of 0.4396

2. ‘Year (yr)’: Coefficient of 0.2344

3. ‘Season_Winter’: Coefficient of 0.1322

## General Subjective Questions

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

Linear regression is a statistical algorithm used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables, meaning that the dependent variable can be expressed as a linear combination of the independent variables.

The algorithm aims to find the best-fitting line that represents the relationship between the variables. This line is called the regression line or the line of best fit. The regression line is determined by minimizing the sum of the squared differences between the actual values of the dependent variable and the predicted values from the line.

Here are the steps involved in the linear regression algorithm:

1. **Data Preparation**: Gather a dataset that includes both the dependent variable (also known as the target variable) and one or more independent variables (also known as predictor variables or features). Ensure the dataset is suitable for linear regression analysis, such as having numerical data and no missing values.
2. **Data Visualization**: Explore the data using scatter plots or other visualizations to understand the relationship between the dependent variable and each independent variable. This step helps identify any linear patterns or outliers in the data.
3. **Model Building**: Select the appropriate form of linear regression based on the problem at hand. There are different types of linear regression, such as simple linear regression (with one independent variable) and multiple linear regression (with multiple independent variables). Define the regression equation, which represents the relationship between the dependent variable and the independent variables.
4. **Model Training**: Use the dataset to estimate the coefficients (slope and intercept) of the regression equation. This is done by finding the values that minimize the sum of squared differences between the actual and predicted values.
5. **Model Evaluation**: Assess the performance of the model using various statistical measures such as R-squared (coefficient of determination), adjusted R-squared, mean squared error, and others. These measures indicate how well the model fits the data and can be used to compare different models.
6. **Model Prediction**: Once the model is built and evaluated, it can be used to make predictions on new, unseen data. The regression equation is applied to the new data to estimate the values of the dependent variable.

It's important to note that linear regression makes certain assumptions, such as linearity, independence of errors, constant variance of errors (homoscedasticity), and normal distribution of errors. Violations of these assumptions may affect the accuracy and reliability of the results. Therefore, it's necessary to check these assumptions before interpreting the results of a linear regression analysis.

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

Anscombe's quartet is a set of four datasets to demonstrate the importance of graphing data before analyzing it. Here are the key points:

1. Descriptive Similarity: Each dataset consists of eleven x,y points with nearly identical statistical properties: mean, variance, correlation, and regression line (y = 3.00 + 0.500x).

2. Graphic Diversity: When graphed, each dataset shows a distinctly different distribution or relationship. For example, one is linear, another is non-linear (curvilinear), the third has an outlier affecting the correlation, and the fourth has a distinct variance issue with one outlier affecting the regression line.

3. Importance of Visualization: This quartet stresses that despite similar statistical features, datasets can behave very differently and it's crucial to plot data visually to capture underlying patterns, outliers, or anomalies that statistics alone may miss. It serves as a cautionary tale against relying solely on numerical summaries for data analysis.

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

Pearson's R is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It ranges from -1 to +1, with a value of 1 indicating a perfect positive linear relationship, 0 indicating no linear relationship, and -1 indicating a perfect negative linear relationship.

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

Scaling is a data preprocessing technique used to transform variables into a specific range or distribution. It is performed to ensure that all variables are on a comparable scale, which can improve the performance of certain machine learning algorithms.

The main reason for scaling is to avoid the dominance of certain variables over others in the data analysis process. Variables with large values may have a disproportionate impact on the analysis and can overshadow the contribution of other variables. Scaling helps to address this issue by bringing all variables to a similar scale, allowing for a fair comparison and analysis.

There are two commonly used methods of scaling: normalized scaling and standardized scaling.

1. Normalized Scaling (also known as Min-Max scaling):
   - Normalized scaling transforms the values of variables to a range between 0 and 1.
   - It uses the minimum and maximum values of each variable to perform the transformation.
   - The formula for normalized scaling is:
     - scaled_value = (original_value - min_value) / (max_value - min_value)
   - Normalized scaling preserves the relative relationships between the values of the variables, but it may be sensitive to outliers.
2. Standardized Scaling (also known as Z-score scaling):
   - Standardized scaling transforms the values of variables to have a mean of 0 and a standard deviation of 1.
   - It uses the mean and standard deviation of each variable to perform the transformation.
   - The formula for standardized scaling is:
     - scaled_value = (original_value - mean) / standard_deviation
   - Standardized scaling centers the distribution of variables around zero and adjusts for differences in scale between variables.

- It preserves the relative relationships between the values of the variables and is less sensitive to outliers compared to normalized scaling.

In summary, scaling is performed to bring variables to a comparable scale and avoid dominance of certain variables in data analysis. Normalized scaling transforms values to a range between 0 and 1, while standardized scaling transforms values to have a mean of 0 and a standard deviation of 1.

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

The Variance Inflation Factor (VIF) is a measure used in regression analysis to assess multicollinearity, which is the presence of high correlation among predictor variables. The VIF equation is as follows:

VIF = 1 / (1 - R^2)

Where R^2 is the coefficient of determination of a particular predictor variable regressed on all other predictor variables.

When the VIF value is infinite, it indicates that there is perfect multicollinearity present in the regression model. This occurs when one or more predictor variables can be perfectly predicted by a linear combination of the other predictor variables. In other words, there is an exact linear relationship between two or more variables in the model.

Perfect multicollinearity can cause issues in regression analysis because it violates the assumptions of ordinary least squares (OLS) regression. It leads to unreliable coefficient estimates, large standard errors, and unstable model predictions. Therefore, it is important to detect and address multicollinearity in order to obtain accurate and meaningful results from regression analysis.

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

The quantile-quantile Q-Q plot is used to compare the quantiles of a sample distribution with a theoretical distribution. It helps determine if the dataset follows a specific distribution and if the errors in linear regression are normally distributed. It's important for assessing the assumptions of linear regression and identifying any deviations from normality.