# 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)

Ans: - The season has a significant impact on bike rentals, with fall (season 3) having the highest demand.

- Bike rentals have been on the rise in the year 2019 compared to 2018.

- Bike demand is high in the months from May to October, with March, May, October, and September being peak rental months.

- Clear or misty cloudy weather conditions are associated with higher bike rentals, while light rain or light snow leads to lower demand.

- Bike demand remains relatively consistent throughout the weekdays.

- Bike demand doesn't significantly change based on whether it's a working day or not.

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

Ans: Creating k-1 dummy variables from a categorical variable with k categories is crucial to avoid redundancy and reduce multicollinearity in regression analysis. It simplifies the interpretation of coefficients and the intercept term while efficiently encoding categorical information. For instance, in a case with three categories (e.g., Furnished, Semi-furnished, Unfurnished), having two dummy variables (Furnished and Semi-furnished) is sufficient, as the absence of both implies the third category (Unfurnished). This practice optimizes data representation and enhances the reliability of regression results.

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

Ans: atemp and temp both have same correlation with target variable of 0.63 which is the highest among all numerical variables.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: Linearity: Assumes a linear relationship between predictor and response variables.
Normality: Expects the errors to follow a normal distribution.
Constant Variance: Requires consistent spread of errors (homoscedasticity).
Low Multicollinearity: Expects low correlation between predictor variables (measured by low VIF values).

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans:

- Holiday
- yr
- weathersit_Moderate

# General Subjective Questions

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

Ans: Linear regression is a statistical and machine learning algorithm used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). It assumes a linear relationship and aims to find the best-fitting line or hyperplane that minimizes the difference between predicted and actual values. This is done by optimizing coefficients to minimize a cost function (e.g., MSE)

Simple Linear Regression (SLR) models the relationship between a single dependent variable and a single independent variable by finding the best-fitting straight line that minimizes the difference between predicted and actual values.
In SLR, the linear equation is represented as:

y=mx+c
Where:
y is the target variable.
x is the predictor variable.
m is the slope of the line (the coefficient).
c is the y-intercept (the constant).

Multiple Linear Regression (MLR) extends SLR to consider multiple independent variables, allowing for more complex modeling by incorporating several predictors that simultaneously influence the dependent variable.
In MLR, the linear equation is represented as:

y= c+m1x1+m2x2+m3x3……mnxn
n defines number of independent variables.
y is the target variable.
x is the predictor variable.
m is the slope of the line (the coefficient).
c is the y-intercept (the constant).

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

Ans: Anscombe's quartet is a set of four small datasets with nearly identical simple descriptive statistics (mean, variance, correlation, etc.), but they exhibit different relationships when plotted. It highlights the importance of visualizing data. Each dataset in the quartet emphasizes the need for exploratory data analysis and not relying solely on summary statistics.
Anscombe's quartet consists of four datasets, each containing 11 data points. These datasets were created by the statistician Francis Anscombe to illustrate the concept that datasets with similar statistical properties can lead to vastly different insights when graphed.
Dataset I:
This dataset forms a simple linear relationship when plotted. It demonstrates that relying solely on summary statistics like correlation can be misleading.

Dataset II:
Dataset II also exhibits a linear relationship but with an outlier that significantly affects the regression line. It emphasizes the importance of identifying and handling outliers in data analysis.

Dataset III:

This dataset forms a non-linear relationship when plotted. It highlights that linear regression may not be appropriate for all datasets and the importance of considering alternative regression models.

Dataset IV:
Dataset IV consists of two distinct clusters of data points with different linear relationships. It illustrates that datasets can have multiple subgroups or patterns, which may require separate analysis.

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

Ans: Pearson's R referred as Pearson's correlation coefficient measures the strength and direction of the linear relationship of two continuous variables. It ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), where 0 equals no linear correlation. It helps assess the degree to which two variables change together. If R is positive, it signifies a positive linear correlation, meaning that as one variable increases, the other tends to increase as well. Similarly, if R is negative, it indicates a negative linear correlation, implying that as one variable increases, the other tends to decrease.
Pearson's correlation is frequently used in various fields, including statistics, economics, and social sciences, to assess how two variables are related. For example, it can be used to examine the correlation between a person's age and their income. Pearson's correlation assumes that the relationship between the variables is linear. If the relationship is non-linear, Pearson's R may not accurately capture the association.

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

Ans: Scaling is a fundamental data preprocessing technique used in statistics and machine learning. It involves transforming data to a standardized range, ensuring that variables with different units, magnitudes, or scales can be compared on an equal footing.
The primary purpose of scaling is to eliminate any potential bias that may arise from the inherent differences in the measurement units of various variables. By bringing all variables to a common scale, scaling helps prevent certain variables from dominating the analysis or influencing algorithms disproportionately due to their larger values. Scaling is performed to enhance the effectiveness of data analysis and modeling, as it aids in improving algorithm convergence, reducing sensitivity to feature magnitudes, and facilitating the interpretation of results.

**Normalized Scaling**: Normalized Scaling also known as min-max scaling, is one of the common scaling techniques. It rescales data to a range typically between 0 and 1. The formula used involves subtracting the minimum value of the variable from each data point and then dividing by the range (the difference between the maximum and minimum values). Normalization preserves the relative differences in the data, ensuring that the relationship between values remains the same.

**Standardized Scaling:** Standardized Scaling or z-score scaling, is another widely used scaling method. It scales data to have a mean of 0 and a standard deviation of 1. To achieve this, it subtracts the mean of the variable from each data point and divides by the standard deviation. Standardization centers the data, making it have a similar mean and scale, which is especially important when working with algorithms that are sensitive to variable magnitudes or when comparing variables with different units.

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

Ans: The value of VIF (Variance Inflation Factor) can become infinite when there is a perfect or nearperfect multicollinearity issue in the regression model. Perfect multicollinearity occurs when one or more predictor variables can be exactly predicted by a linear combination of other predictor variables within the model. When perfect multicollinearity is present, the VIF for at least one of the correlated variables becomes infinite because the formula for calculating VIF involves division by zero. Here's why this happens:

VIF=1/1-R^2

When perfect multicollinearity exists, R^2 becomes equal to 1 because the correlated variable can be perfectly predicted by other variables. When R^2 is 1, the denominator in the VIF formula becomes 0 resulting in division by zero and yielding an infinite VIF value.

Essentially, infinite VIF signifies that the model cannot handle the perfect multicollinearity issue, making it mathematically unsolvable and highlighting the need to address multicollinearity by either eliminating one of the correlated variables or employing techniques designed to mitigate this problem.

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

Ans: A Q-Q (Quantile-Quantile) plot is a graphical tool to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It compares the quantiles of the dataset to those of the theoretical distribution. Here are some important points to note:

Validity of Statistical Inference: Ensuring that the residuals follow a normal distribution is crucial for the validity of statistical inference in linear regression. Many statistical tests and confidence intervals rely on the assumption of normality

Model Evaluation: The Q-Q plot provides a visual representation of how closely the residuals approximate a normal distribution. If the points on the plot closely align with a straight line (the diagonal line), it suggests that the residuals exhibit a normal distribution.

Identifying Departures from Normality: On the Q-Q plot, departures from the diagonal line can indicate deviations from normality. For example, if the points deviate upward or downward at the tails, it may suggest heavy-tailed or skewed residuals.