# 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**: From the analysis of the categorical variables in the dataset, it can be inferred that bike rental rates tend to be higher during the summer and fall seasons, with peak demand observed in the months of September and October. Additionally, bike rentals are more prominent on Saturdays, Wednesdays, and Thursdays, and the demand is higher in the year 2019 compared to 2018. Moreover, bike rentals are notably higher on holidays. These insights offer a deeper understanding of the factors influencing rental demand and can help in formulating targeted strategies for increased rentals.

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

**Ans**: The `drop_first=True` parameter is used to avoid redundancy when creating dummy variables. It removes the first category, thereby reducing the number of columns and preventing the dummy variable trap, which can lead to multicollinearity.

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

**Ans**:The `temp` variable shows the highest correlation with the target variable, indicating that temperature has the most significant impact on bike rentals.

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

**Ans**: The assumptions of linear regression were validated through multiple checks. First, we assessed the Variance Inflation Factor (VIF) to ensure there was no multicollinearity among the independent variables, with all variables having a VIF below 5. Second, we examined the error distribution of residuals, which was centered around 0 and followed a normal distribution, meeting the assumption of homoscedasticity and normality of errors. Finally, we confirmed the linear relationship between the dependent variable and the feature variables by visualizing scatter plots and the regression line, showing a linear trend in the data

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

**Ans**: The top three features that significantly contribute to the demand for shared bikes are temperature, year, and holiday variables. Temperature has the highest correlation with bike demand, as it directly influences people's willingness to rent bikes. The year variable indicates a growing trend in bike rentals, with 2019 showing higher demand than 2018. Additionally, holidays contribute positively to bike rentals, with more people renting bikes on such days, likely due to increased leisure and outdoor activities.

# General Subjective Questions

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

AnsLinear Regression is a machine learning algorithm used for supervised learning, primarily to predict a dependent variable (target) based on one or more independent variables (predictors). The core concept of linear regression is to establish a linear relationship between the dependent and independent variables, making it one of the most widely used techniques for predictive modeling.

There are two primary types of linear regression:

1. Simple Linear Regression: This is used when there is only one independent variable to predict the target variable. It attempts to fit a straight line (the regression line) that best describes the relationship between the single independent variable and the dependent variable. The regression line can be described by the equation:
   Y=β0+β1XY = \beta_0 + \beta_1 XY=β0+β1X
   where YYY is the predicted value of the dependent variable, XXX is the independent variable, β0\beta_0β0 is the intercept, and β1\beta_1β1 is the coefficient for XXX.
2. Multiple Linear Regression: This is used when there are multiple independent variables to predict the target variable. The relationship is still linear, but it involves more than one independent variable. The equation for multiple linear regression is:
   Y=β0+β1X1+β2X2+⋯+βnXnY = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_nY=β0+β1X1+β2X2+⋯+βnXn
   where X1,X2,…,XnX_1, X_2, \dots, X_nX1,X2,…,Xn are the independent variables and

β1,β2,…,βn\beta_1, \beta_2, \dots, \beta_nβ1,β2,…,βn are their respective coefficients.

In linear regression, the regression line represents the relationship between the dependent and independent variables. The direction of the relationship is determined by the sign of the coefficient:

- Positive Linear Relationship: When the dependent variable (Y) increases as the independent variable (X) increases. This is seen when the regression coefficient is positive.
- Negative Linear Relationship: When the dependent variable (Y) decreases as the independent variable (X) increases. This is seen when the regression coefficient is negative.

Linear regression is foundational in predicting numerical outcomes and is widely applied in various domains like economics, business, and social sciences.

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

**Ans**:Anscombe's quartet consists of four distinct datasets that each contain 11 data points. Despite having nearly identical simple descriptive statistics, such as mean, variance, and correlation, the datasets have very different distributions and appear vastly different when visualized graphically. This is a deliberate design by the statistician Francis Anscombe to highlight the importance of graphical data exploration in the analysis process.

The four datasets in the quartet show that relying solely on summary statistics, such as mean and correlation, can be misleading because these numbers can mask underlying patterns and trends in the data. By plotting the data, one can observe the presence of outliers, non-linearity, and other patterns that are not captured by summary statistics alone.

Anscombe's quartet emphasizes the importance of visually inspecting data before proceeding with analysis, as graphical representations provide deeper insights into the structure and characteristics of the data, ensuring more accurate and informed conclusions.

**3. What is Pearson's R?**

**Ans**: Pearson's Correlation Coefficient (denoted as r) is a statistical measure used to determine

the strength and direction of the linear relationship between two continuous variables. It quantifies how well the variables are related, with the coefficient value ranging from -1 to +1.

- **r = +1** indicates a perfect positive linear relationship, where both variables increase together in a straight-line pattern.
- **r = -1** indicates a perfect negative linear relationship, where one variable increases while the other decreases in a perfectly linear fashion.
- **r = 0** indicates no linear relationship between the variables, meaning the variables are uncorrelated.

Values closer to +1 or -1 indicate a stronger linear relationship, while values near 0 suggest a weaker or no linear relationship. Pearson's correlation assumes that the relationship between the two variables is linear and both variables are approximately normally distributed.

In practice, Pearson's Correlation Coefficient helps to assess the degree of association between two variables, which can be important in regression analysis and other statistical methods.

**4. What is scaling? Why is scaling performed? What is the difference between**

**normalized scaling and standardized scaling?**

**Ans**: Scaling is a crucial pre-processing technique in machine learning that standardizes the independent feature variables in a dataset to a common range or distribution. It ensures that all features contribute equally to the model, preventing some features from dominating others due to differences in their scales.

In a dataset, feature variables can have varying ranges and units, which may cause issues in the model. Without scaling, features with larger magnitudes could disproportionately affect the model's performance, leading to incorrect results. Scaling ensures that all variables are on a similar scale, helping the algorithm perform better, especially in models like linear regression, SVM, and KNN.

**Normalization vs. Standardization:**

- **Normalization** involves transforming data to a specific range, typically [0, 1], using methods like min-max scaling. It is often used when the distribution of the data is not Gaussian and needs to be bounded within a range.
- **Standardization**, on the other hand, transforms the data to have a mean of 0 and a standard deviation of 1, resulting in Z-scores. This technique is useful when the data follows a Gaussian distribution, and the variance of each feature matters for the model's

performance.

In summary, scaling prepares the data for the machine learning model by eliminating biases introduced by different feature ranges or units, ensuring that all features contribute equally to the model.

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

**Ans**:The **Variance Inflation Factor (VIF)** is a metric used to detect the extent of multi-collinearity in a regression model. It quantifies how much the variance of a regression coefficient is inflated due to the correlation with other independent variables. A high VIF value suggests that a particular variable is highly correlated with one or more other variables in the model, leading to multi-collinearity.

When there is a perfect correlation between two independent variables, the **R-squared ($R^2$)** value becomes 1, meaning the independent variables can be perfectly predicted from each other. In this case, the **VIF** value becomes infinite because the formula for VIF is given by:

VIF=11−R2VIF = \frac{1}{1 - R^2}VIF=1−R21

When R2=1R^2 = 1R2=1, this leads to a division by zero, resulting in an infinite VIF. This indicates a serious problem of multi-collinearity, where the independent variables are not providing unique information to the model.

To resolve this issue, it is recommended to drop one of the perfectly correlated variables from the model. Removing one of the correlated variables reduces redundancy, helping to stabilize the regression model and improve its interpretability. This is crucial for building a well-defined, reliable regression model.

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

**Ans**: A **Quantile-Quantile (Q-Q) plot** is a graphical tool used to compare the distribution of a sample data set with a theoretical distribution, such as normal, uniform, or exponential distribution. The plot displays the quantiles of the sample data on the x-axis and the quantiles of the theoretical distribution on the y-axis. If the data points in the Q-Q plot fall approximately along a straight line, this indicates that the sample data follows the theoretical distribution.

Q-Q plots are primarily used for:

1. **Checking Distribution**: They help in determining whether the data follows a specific distribution, such as normality. For example, in a normal Q-Q plot, if the data points align along a straight line, the data is likely to follow a normal distribution.
2. **Identifying Deviations from Distribution**: Deviations from the straight line in the Q-Q plot suggest that the data does not follow the assumed distribution (e.g., if the points curve away, it might suggest skewness or heavy tails).
3. **Error Distribution**: Q-Q plots are useful for examining if the residuals (errors) from a model are normally distributed. This assumption is critical in many statistical methods, as violations of normality can affect the validity of inferences drawn from the model.

In summary, Q-Q plots provide valuable insight into the underlying distribution of data and whether certain assumptions, such as normality of errors, are met in statistical models.