# Assignment-based Subjective Questions

***Q1.*** *From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)*
**Ans1.** Seasons, years and months have positive effect on the dependent variable count.

***Q2.*** *Why is it important to use* ***drop_first=True*** *during dummy variable creation? (2 mark)*
**Ans2.**It is important to use drop_first=True, as it drop the first dummy variable created. And this first variable can be inferred from rest of the dummy variables created.

***Q3.*** *Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)*
**Ans3.** temp(temperature in Celsius) and atemp(feeling temperature in Celsius) have the highest correlation with target variable cnt(count of total rental bikes)

***Q4.*** *How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)*
**Ans4.** Validating the assumptions of linear regression is an essential step to ensure the model's reliability and generalizability. Here are several common methods to validate the assumptions of linear regression after building the model on the training set:

**a.)Residual Analysis**:

- Plotting the residuals (the differences between observed and predicted values) against the predicted values or the independent variables. The residuals should be randomly distributed around zero without any discernible pattern.
- Checking for heteroscedasticity, which refers to the presence of unequal variance in the residuals across different levels of the independent variables.

**b.)Normality of Residuals**:

- Checking whether the residuals follow a normal distribution. This can be assessed through Q-Q plots, histograms of residuals, or formal statistical tests such as the Shapiro-Wilk test.

**c.)Linearity**:

- Verifying that the relationship between the independent variables and the dependent variable is linear. This can be examined through scatter plots of the independent variables against the dependent variable.

***Q5.*** *Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)*
**Ans5.** temp(temperature in Celsius), Year, and RC(Weathersit value 3 Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds which I named as RC) are the top 3 features contributing significantly towards explaining the demand of the shared bikes based on the coefficient of model for each feature.

# General Subjective Questions

***Q1.****Explain the linear regression algorithm in detail. (4 marks)*
**Ans1.** Linear regression is a statistical method used for modeling the relationship between a dependent variable (target) and one or more independent variables (features). It assumes a linear relationship between the independent variables and the dependent variable. Here's a detailed explanation of the linear regression algorithm:

1. **Assumptions**:
    - **Linearity**: The relationship between the independent variables and the dependent variable is linear.
    - **Independence**: The observations are independent of each other.
    - **Homoscedasticity**: The variance of the residuals (the differences between observed and predicted values) is constant across all levels of the independent variables.
    - **Normality of Residuals**: The residuals are normally distributed.
    - **No Multicollinearity**: The independent variables are not highly correlated with each other.
2. **Model Representation**:
    - The linear regression model can be represented as: $Y=\beta0+\beta1X1+\beta2X2+...+\beta nXn+\epsilon Y=\beta0+\beta1X1+\beta2X2+...+\beta nXn+\epsilon$
        - Y is the dependent variable.
        - X1,X2,...,XnX1,X2,...,Xn are the independent variables.

        - $\beta 0, \beta 1, \ldots, \beta n \beta 0, \beta 1, \ldots, \beta n$ are the coefficients (parameters) representing the effect of each independent variable on the dependent variable.
        - $\epsilon\epsilon$ is the error term representing the difference between the observed and predicted values.
3. **Objective**:
    - The objective of linear regression is to estimate the coefficients $\beta 0, \beta 1, \ldots, \beta n$ that minimize the sum of squared differences between the observed and predicted values (i.e., minimize the residual sum of squares).
4. **Parameter Estimation**:
    - The coefficients are estimated using methods such as Ordinary Least Squares (OLS) or gradient descent.
    - OLS estimates the coefficients by minimizing the sum of squared residuals analytically.
    - Gradient descent iteratively updates the coefficients to minimize the cost function (sum of squared residuals).
5. **Model Evaluation**:
    - The model's performance is evaluated using metrics such as R-squared, adjusted R-squared, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and others.
    - These metrics assess how well the model fits the data and whether it generalizes to unseen data.
6. **Interpretation**:
    - Once the model is trained, the coefficients can be interpreted to understand the relationship between the independent variables and the dependent variable.
    - A positive coefficient indicates a positive relationship, while a negative coefficient indicates a negative relationship.
    - The magnitude of the coefficient reflects the strength of the relationship, with larger coefficients indicating a stronger effect.
7. **Predictions**:
    - After model evaluation, the trained model can be used to make predictions on new data by plugging in the values of the independent variables into the regression equation.

Linear regression is a simple, yet powerful method widely used for predictive modeling, inference, and understanding the relationships between variables in various fields such as economics, finance, social sciences, and machine learning.

***Q2.*** *Explain the Anscombe's quartet in detail. (3 marks)*

**Ans2.** Anscombe's quartet is a set of four datasets that have nearly identical statistical properties but appear very different when graphed. It was created by the statistician Francis Anscombe in 1973 to illustrate the importance of graphical data exploration and the limitations of relying solely on summary statistics.

The four datasets in Anscombe's quartet share the following statistical properties:

1. Each dataset consists of 11 (x, y) pairs.
2. Each dataset has the same mean and variance for both the x and y variables.
3. Each dataset has the same correlation coefficient between the x and y variables.
4. Each dataset yields the same linear regression line when fit to a linear model.

Despite these similarities, the datasets have very different patterns when plotted. Here's a brief description of each dataset:

1. **Dataset I**:
    - It forms a clear linear relationship between x and y.
    - All data points lie close to the regression line, and there are no outliers.
2. **Dataset II**:
    - It also exhibits a linear relationship, but with one outlier that significantly affects the regression line.
    - Removing the outlier would result in a nearly perfect linear relationship.
3. **Dataset III**:
    - It forms a non-linear relationship with a distinct pattern resembling a quadratic curve.
    - Despite having the same mean, variance, and correlation as the other datasets, it highlights the importance of considering non-linear relationships.
4. **Dataset IV**:
    - It consists of several clusters of data points, with each cluster having its own linear relationship.
    - The presence of distinct groups demonstrates the importance of identifying and analyzing subgroup relationships within a dataset.

The significance of Anscombe's quartet lies in its demonstration of the limitations of summary statistics. While summary statistics such as mean, variance, and

correlation provide valuable insights into the overall characteristics of a dataset, they may overlook important patterns or outliers that become apparent through visualization. Anscombe's quartet underscores the importance of graphical exploration in data analysis, encouraging researchers to plot their data before drawing conclusions or fitting models.

By highlighting the dangers of blindly relying on summary statistics, Anscombe's quartet has become a classic example in statistics and data science education, emphasizing the need for both numerical and visual analysis techniques in data exploration and interpretation.

***Q3.*** *What is Pearson's R? (3 marks) :*

**Ans3.** Pearson's correlation coefficient, often denoted by r, is a statistic that measures the strength and direction of the linear relationship between two continuous variables. It was developed by the British statistician Karl Pearson in the late 19th century and is one of the most widely used measures of correlation.

Pearson's r can range from -1 to 1:

- $r=1r=1$ indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable increases proportionally.
- $r=-1r=-1$ indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.
- $r=0r=0$ indicates no linear relationship between the variables.

The formula for Pearson's correlation coefficient between two variables X and Y with n data points is:

$$r=\sum i=1n(Xi-X^-)(Yi-Y^-)\sum i=1n(Xi-X^-)2\sum i=1n(Yi-Y^-)2r=\sum i=1n(Xi-X^-)2$$

$$\sum i=1n(Yi-Y^-)2$$

$$\sum i=1n(Xi-X^-)(Yi-Y^-)$$

Where:

- Xi and Yi are the individual data points.
- $X^-$ and $Y^-$ are the means of the X and Y variables, respectively.

Pearson's correlation coefficient is particularly useful for assessing the strength and direction of linear relationships between variables, but it assumes that the

relationship is linear and that the variables are approximately normally distributed. It is sensitive to outliers and can be affected by extreme values.

Pearson's r is widely used in various fields, including psychology, economics, biology, and social sciences, to examine relationships between variables, identify patterns, and make predictions. However, it's essential to interpret r within the context of the specific dataset and research question and to consider other factors such as causality, non-linearity, and the presence of outliers.

***Q4.*** *What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)*

**Ans4.** Scaling is the process of transforming data to fit within a specific scale, often to make it more manageable or comparable. It involves adjusting the range of values of variables while preserving their relationships with each other. Scaling is commonly performed in data preprocessing before applying certain machine learning algorithms or statistical analyses.

**Reasons for Scaling:**

1. **Normalization of Data**: Scaling ensures that all variables are on a similar scale, typically between 0 and 1 or -1 and 1. This normalization prevents variables with larger scales from dominating those with smaller scales in certain algorithms, such as distance-based methods like k-nearest neighbors or clustering algorithms.
2. **Improvement of Convergence**: Many optimization algorithms, such as gradient descent, converge more quickly when variables are on similar scales. Scaling helps prevent these algorithms from getting stuck in local minima or taking longer to converge due to differences in variable scales.
3. **Enhancement of Interpretability**: Scaling makes it easier to interpret the coefficients or feature importance of different variables in a model. Without scaling, the coefficients may not be directly comparable, making it challenging to assess the relative importance of each variable.
4. **Reduction of Numerical Instability**: Some algorithms, such as principal component analysis (PCA), are sensitive to the scale of variables. Scaling prevents numerical instability and ensures that the results of these algorithms are not heavily influenced by differences in variable scales.

**Difference between Normalized Scaling and Standardized Scaling:**

1. **Normalized Scaling**:

- Normalization scales the values of variables to a fixed range, typically between 0 and 1. It is achieved by subtracting the minimum value and dividing by the range (maximum value minus minimum value) of each variable.
- Normalization is useful when the distribution of the variables is not necessarily Gaussian and the algorithms being used do not assume Gaussian distribution.

2. **Standardized Scaling**:
   - Standardization scales the values of variables to have a mean of 0 and a standard deviation of 1. It is achieved by subtracting the mean and dividing by the standard deviation of each variable.
   - Standardization is more appropriate when the variables have different units or distributions and when the algorithms being used assume that the variables are normally distributed.

In summary, while both normalized scaling and standardized scaling adjust the scale of variables, they differ in the range of values they produce and their suitability for different types of data and algorithms. Normalization is suitable for algorithms that do not assume Gaussian distribution, while standardization is more appropriate for algorithms that assume Gaussian distribution or when the variables have different units.

***Q5.*** *You might have observed that sometimes the value of VIF is infinite. Why does this happen?*
(3 marks)
**Ans5.** The Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression analysis. Multicollinearity occurs when independent variables in a regression model are highly correlated with each other, which can lead to inflated standard errors and unreliable coefficient estimates. The VIF quantifies the severity of multicollinearity by calculating how much the variance of a regression coefficient is inflated due to multicollinearity.

The formula for VIF for a particular independent variable in a regression model is:

VIFi=1 / (1−Ri^2)

Where:

- Ri^2 is the coefficient of determination (R-squared) of the regression model with the ith independent variable as the dependent variable and all other independent variables as predictors.

When the value of Ri2 is close to 1, indicating a strong linear relationship between the ith independent variable and the other independent variables in the model, the denominator of the VIF formula becomes close to 0. As a result, the VIF becomes very large, approaching infinity.

The occurrence of an infinite VIF typically indicates perfect multicollinearity, where one or more independent variables in the regression model can be perfectly predicted by a linear combination of the other independent variables. In other words, one or more independent variables are redundant because they are linearly dependent on other variables in the model.

Perfect multicollinearity often arises due to data coding or measurement errors, inappropriate model specification, or including variables that are transformations or combinations of other variables already in the model. It can lead to numerical instability in estimation procedures and unreliable results.

To address the issue of infinite VIF, it's essential to examine the variables in the regression model and identify the source of multicollinearity. This may involve removing redundant variables, transforming variables, or redefining the model specification to mitigate multicollinearity and ensure the stability and reliability of the regression estimates.

***Q6.*** *What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)*

**Ans6.** A Q-Q (quantile-quantile) plot is a graphical method used to assess whether a dataset follows a particular probability distribution, such as the normal distribution. It compares the quantiles of the dataset's empirical distribution (observed data) against the quantiles of a theoretical distribution (expected data), typically the normal distribution. If the observed data closely follows the expected distribution, the points on the Q-Q plot will fall approximately along a straight line.

Here's how a Q-Q plot works:

1. **Sorting the Data**: The dataset's values are sorted in ascending order.
2. **Calculating Quantiles**: Quantiles are calculated for both the observed data and the theoretical distribution. These quantiles represent points that divide the data into equal-sized intervals. For example, the median represents the 50th percentile, dividing the data into two equal halves.
3. **Plotting Points**: The quantiles of the observed data are plotted on the x-axis, while the quantiles of the theoretical distribution are plotted on the y-axis.

4. **Visual Assessment**: If the data closely follows the theoretical distribution, the points on the Q-Q plot will form a straight line. Deviations from a straight line indicate departures from the expected distribution.

**Importance of Q-Q Plot in Linear Regression:**

1. **Assumption Checking**: Q-Q plots are commonly used in linear regression to assess the normality of residuals. In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) are normally distributed. By examining the Q-Q plot of residuals, you can determine whether this assumption holds true. If the residuals deviate significantly from a straight line, it suggests that the normality assumption may be violated.
2. **Diagnostic Tool**: Q-Q plots serve as a diagnostic tool to identify departures from normality and other distributional assumptions. They provide a visual indication of the distributional characteristics of the residuals, allowing you to identify potential issues such as skewness, heavy tails, or outliers.
3. **Model Improvement**: Identifying deviations from normality through Q-Q plots can guide model improvement strategies. For example, if the residuals exhibit non-normality, you may consider transforming the response variable or including additional predictor variables to better capture the data's distributional characteristics.
4. **Model Interpretation**: Ensuring that the residuals follow a normal distribution is important for the validity of statistical inference in linear regression. Q-Q plots help validate the assumptions underlying hypothesis tests, confidence intervals, and other inferential procedures based on the regression model.

Overall, Q-Q plots play a crucial role in assessing the validity of linear regression models by providing insights into the distributional characteristics of residuals and facilitating informed decisions regarding model adequacy and improvement.