# Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer abouttheir effect on the dependent variable? (3 marks)**

   -- During EDA analysis on 'weathersit' column, I observed that the during '4: Heavy Rain + I ce Pallets + Thunderstorm + Mist, Snow + Fog', users we not at all riding using the app.

   -- Users mostly used on weather conditions like '1: Clear, Few clouds, Partly cloudy, Partly cloudy' followed by '2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist'

   -- Users are less likely to use during weather conditions like '3: Light Snow, Light Rain + Th understorm + Scattered clouds, Light Rain + Scattered clouds'

   -- Fall season has most ridership by users.

   -- During working day or non-working days the usage i.e., numbers of users using remaine d the same.

   -- The mean of user ridership is higher in 2019 than in 2018.

   -- After doing a segmented Analysis on hue being the year, we can see that the ridership i. e., usage count is significantly more in 2019 compared to 2018 in all parameters.

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

   - For 'k' Categorical levels, we only require (k-1) dummies. So for this we use drop_first=True while creating dummies for categorical columns. This helps in simplifying our model.

   - For example, in the data, the column "season" has 4 values
     : season (1:spring, 2:summer, 3:fall, 4:winter).

     So we only require (k-1) i.e., (4-1) = 3 dummies.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   - Based on pair plot, 'temp' and 'atemp' has equal amount of correlation with target variable 'cnt'.

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

   There are few ways for validating:
   - Calculate the VIF (Variance Inflation Factor) for the independent variables to check for Multicollinearity.
   - Residual: Error terms should ne normally distributed and mean must be around zero.
   - There should not be any visible patterns in residual 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)**

- Based on the observations from my final model, the top three features contributing are:
  'temp'
  'mnth_9' i.e., September month and
  Season ('Fall')

# General Subjective Questions

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

Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent(y) and independent variable(x).

Linear Regression is of two types: Simple and Multiple.

Simple Linear Regression is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable Whereas, In Multiple Linear Regression there are more than one independent variables for the model to find the relationship.

Simple Linear Relationship: $Y = MX + C$

Where, C = Constant / Intercept

M = Slope

Multiple Linear Relationship: $Y = b0 + b1x1 + b2x2 + b3bx3 + ….. + bnxn$

Linear Relation ship can be +ve or -ve as well.

Assumptions:

a. Linearity: It states that the dependent variable Y should be linearly related to independent variables.

b. Homoscedasticity: The variance of the error terms should be constant i.e., the spread of residuals should be constant for all values of X. This assumption can be checked by plotting a residual plot.

c. No Multicollinearity: The variables should be independent of each other i.e., no correlation should be there between the independent variables.

d. The error terms should be normally distributed

e. No Autocorrelation: The error terms (yact – ypred) should be independent of each other.

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

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of

using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

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

The Pearson correlation coefficient (r ) is the most common way of measuring a linear correlation.
Pearson's r is used to illustrate the relationship between two continuous variables, such as years of education completed and income.
The correlation between two variables using Pearson's r will always be between –1 and +1.
A correlation coefficient of 0 means that there is no relationship, either positive or negative, between these two variables.
A correlation coefficient of +1 means that there is a perfect positive correlation, or relationship, between these two variables. In the case of +1, as one variable increases, the second variable increases in exactly the same level or proportion. Likewise, as one variable decreases, the second variable would decrease in exactly the same level or proportion.
A correlation coefficient of –1 means that there is a perfect negative correlation, or relationship, between two variables. In this case, as one variable increases, the second variable decreases in exactly the same level or proportion. Also, as one variable decreases, the other would increase in exactly the same level or proportion.

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

Scaling is a data preprocessing technique used in statistics and machine learning to transform variables or features within a dataset to a consistent range or scale, making it easier for machine learning algorithms to work effectively and avoid bias due to varying units or scales.

Scaling is done to make sure that all the different values or features in your data are in a similar range, so that mathematical operations and comparisons between them are fair and accurate.

Normalized scaling (Min-Max scaling) transforms data to a specific range, typically between 0 and 1, preserving the relative relationships between data points.

X_normalized = (X - X_min) / (X_max - X_min)

Here, X is the original data point, X_min is the minimum value in the dataset, and X_max is the maximum value.

Normalization is useful when you want to preserve the relationships between data points but need to constrain them to a specific range.

Standardized scaling (Z-score scaling) standardizes data to have a mean of 0 and a standard deviation of 1, centering the data and ensuring consistent spread, often used for algorithms assuming a normal distribution.

X_standardized = (X - X_mean) / X_std

Here, X is the original data point, X_mean is the mean of the dataset, and X_std is the standard deviation.
Standardization is useful when you want to center the data around zero and ensure that it has a consistent spread.

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

The VIF (Variance Inflation Factor) can become infinite when two or more variables in a regression model are so strongly related that one of them can be perfectly predicted using the others. It's like having two pieces of information that are exactly the same, making it impossible for the model to tell them apart. This creates problems, and we need to find and fix this by mostly dropping 1 variable, to get reliable results.

VIF = 1 / (1- R^2)

When R-Squared (R^2) = 1 then VIF = Infinite.

R^2 = 1 shows a perfect relation between 2 independent variables.

**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 determine if two datasets come from population with a common distribution.
A Q-Q plot is a graph used in linear regression to check if the data's distribution is close to a normal (bell-shaped) curve.

Used in Assumption Checking:

- One of the key assumptions in linear regression analysis is that the residuals (the differences between observed and predicted values) follow a normal distribution. If this assumption is violated, it can affect the validity of statistical tests and confidence intervals associated with the regression model. A Q-Q plot is a visual tool used to check if the residuals are approximately normally distributed.

If the points on the plot follow a straight line, it's good; if they curve, it suggests the data might not be normal, which can affect the reliability of the regression analysis.

Important in:

- Linear regression assumes that the residuals are normally distributed with a mean of zero. If the Q-Q plot shows deviations from a straight line, it suggests that the normality assumption may be violated. Detecting such violations is crucial for assessing the validity of regression results.
- Q-Q plots can help identify outliers in the data.