# 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:

- Bike Booking has shown an increase in the year 2019 compared to 2018. Indicates the increase in popularity.
- Bike booking seems to be comparatively high during season fall and followed by summer, Boombikes should consider this deploy more bikes during this period.
- Bike booking at its peak on Jun month during the year 2018 and Sep Month during the year 2019
- Decline of Bike booking is visible starting from Nov till Mar, Boombikes should deploy less Bikes during this period.
- Demand is high during non-holidays especially Wednesday to Friday.
- Demand is poor while the weather situation is Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds and no booking during Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

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

Ans:

Using drop_first=True during dummy variable creation s an important consideration when working with categorical variables.
It helps avoid multicollinearity, establishes a reference category, and enhances the interpretability of the model.

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

**Ans:**

Temp have highest correlation

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

**Ans**

Below Method helps to validate the assumptions after building the model.

- Independence of residuals: A random scatter of residuals indicates independence. If there is a pattern or correlation, it suggests the presence of omitted variables or violation of the independence assumption.

- Multicollinearity check-No Correlation between the independent Variable
VIF Should be less than five.

- Homoscedasticity of residuals: NO presence of a funnel shape or a distinct pattern.A uniform distribution of residuals, indicates homoscedasticity.

- Linearity of the relationship: Should not be any deviation from linear relationship

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:

   Below are the 3 Variable which is contributing significantly towards explaining the demand of the shared bikes

   - Temp
   - Winter
   - September

# General Subjective Questions

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

   Ans:

   Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature, is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression. The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).

   The algorithm aims to minimize the difference between the predicted values and the actual values of the target variable. It achieves this by calculating the optimal values for the coefficients and intercepts in the linear equation.

   Mathematically the relationship can be represented with the help of the following equation –

   $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$
   In this equation:

   y represents the target variable (dependent variable) that we want to predict.

   $x_1, x_2, ..., x_p$ represent the input features (independent variables) that we use to predict the target variable.

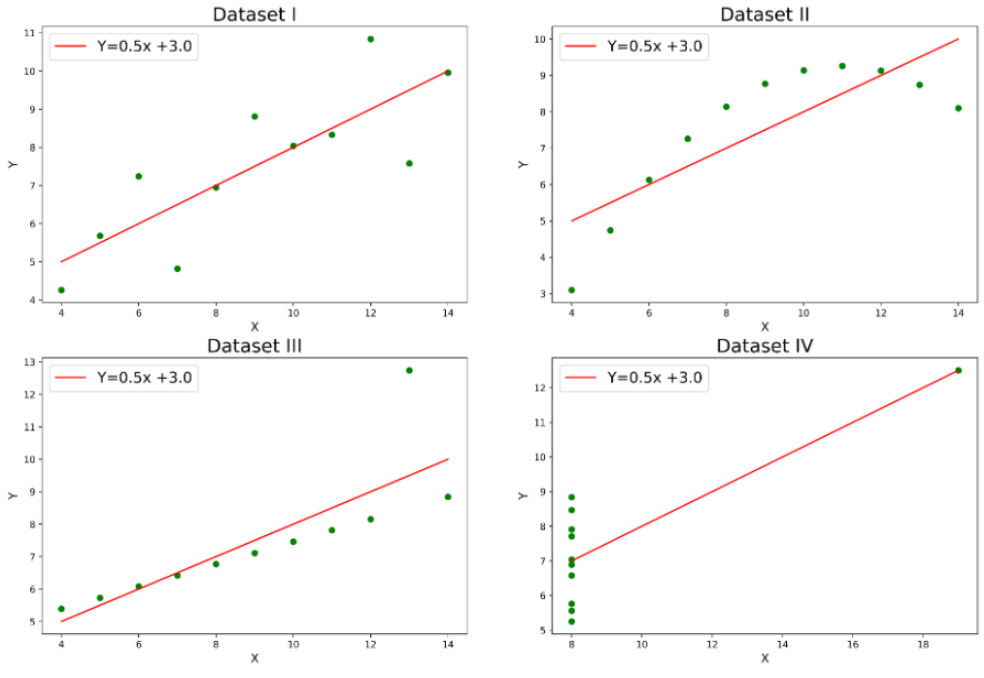
$\beta_o$, $\beta_1$, $\beta_2$, ..., $\beta_p$ are the coefficients or weights assigned to each input feature. $\beta_o$ is the intercept term, and $\beta_1$, $\beta_2$, ..., $\beta_p$ are the slopes that determine the influence of each feature on the target variable.

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

**Ans:**

Anscombe's quartet is a famous statistical example that consists of four datasets, each containing 11 data points. These datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and not relying solely on summary statistics.
The four datasets in Anscombe's quartet have nearly identical summary statistics, including means, variances, and correlation coefficients. However, when graphically visualized, the datasets exhibit significant differences, highlighting the limitations of relying solely on numerical summaries.

| I | | II | | III | | IV | |
|---|---|---|---|---|---|---|---|
| x | y | x | y | x | y | x | y |
| 10.0 | 8.04 | 10.0 | 9.14 | 10.0 | 7.46 | 8.0 | 6.58 |
| 8.0 | 6.95 | 8.0 | 8.14 | 8.0 | 6.77 | 8.0 | 5.76 |
| 13.0 | 7.58 | 13.0 | 8.74 | 13.0 | 12.74 | 8.0 | 7.71 |
| 9.0 | 8.81 | 9.0 | 8.77 | 9.0 | 7.11 | 8.0 | 8.84 |
| 11.0 | 8.33 | 11.0 | 9.26 | 11.0 | 7.81 | 8.0 | 8.47 |
| 14.0 | 9.96 | 14.0 | 8.10 | 14.0 | 8.84 | 8.0 | 7.04 |
| 6.0 | 7.24 | 6.0 | 6.13 | 6.0 | 6.08 | 8.0 | 5.25 |
| 4.0 | 4.26 | 4.0 | 3.10 | 4.0 | 5.39 | 19.0 | 12.50 |
| 12.0 | 10.84 | 12.0 | 9.13 | 12.0 | 8.15 | 8.0 | 5.56 |
| 7.0 | 4.82 | 7.0 | 7.26 | 7.0 | 6.42 | 8.0 | 7.91 |
| 5.0 | 5.68 | 5.0 | 4.74 | 5.0 | 5.73 | 8.0 | 6.89 |




*Anscombe's quartet Plot*

## Explanation of this output:

- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated to be far away from that line.
- Finally, the fourth one(bottom right) shows an example of when one high-leverage point is enough to produce a high correlation coefficient.

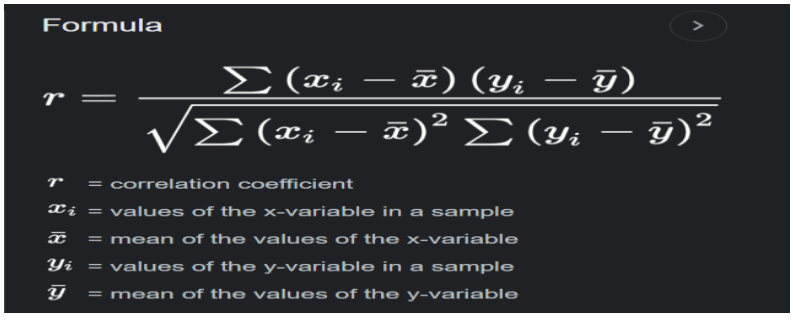
**3. What is Pearson's R? (3 marks)**

**Ans**

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It was developed by Karl Pearson, a British mathematician and statistician.

The Pearson correlation coefficient, denoted as "r," ranges between -1 and 1, where:

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




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

Ans:

Scaling, in the context of data preprocessing, refers to the process of transforming numerical variables to a specific scale or range. It is performed to ensure that all variables are on a similar scale and have comparable magnitudes, which can benefit various machine learning algorithms and statistical techniques. Scaling helps in avoiding biased or distorted results caused by variables with different scales or units.

The primary reasons for performing scaling are as follows:

- **Normalization of Variables:**
- **Enhancement of Algorithm Performance:**

- **Facilitation of Interpretation**

**Normalized Scaling**: Normalization scales the variables to a range between 0 and 1. It is achieved by subtracting the minimum value of the variable from each data point and then dividing it by the range (maximum value minus the minimum value). This ensures that the variable's minimum value becomes 0, and the maximum value becomes 1. Normalization can be expressed using the formula:
x_normalized = (x - min(x)) / (max(x) - min(x))

**Standardized Scaling:** Standardization, also known as z-score scaling, transforms the variables to have zero mean and unit variance. It involves subtracting the mean of the variable from each data point and then dividing it by the standard deviation. This centers the variable distribution around zero and adjusts its scale to have a standard deviation of 1. Standardization can be expressed using the formula:
x_standardized = (x - mean(x)) / std(x)

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

Ans:

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. It quantifies the extent to which the variance of the estimated regression coefficient is inflated due to the presence of correlation among the predictor variables.

In some cases, the VIF value can become infinite. This happens when one or more of the predictor variables in the regression model are perfectly correlated or linearly dependent on a combination of other predictor variables. This scenario is known as perfect multicollinearity.

Perfect multicollinearity can arise due to various reasons, including:

- Data errors or mistakes in data collection.
- Including derived variables that are linear combinations of other variables
- Adding dummy variables for categorical variables without excluding one reference category.

To handle perfect multicollinearity, one or more of the correlated variables should be removed from the model.
This can be done by either:

- Identifying and removing redundant variables manually based on domain knowledge and understanding of the data.
- Using variable selection techniques (e.g., stepwise regression, LASSO, ridge regression) that automatically select a subset of variables based on their relevance and contribution to the model.

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

**Ans:** Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other.
It compares the quantiles of a dataset against the quantiles of a theoretical distribution, often the standard normal distribution. The main purpose of a Q-Q plot is to visually inspect whether the data deviates from a specific distributional assumption.

Here's how the Q-Q plot is constructed and interpreted:

- First, the residuals are calculated by subtracting the predicted values from the observed values in the linear regression model.
- The residuals are then sorted in ascending order.
- Next, the quantiles of the standard normal distribution are calculated at the same probabilities as the ordered residuals. These quantiles serve as the expected values if the residuals are normally distributed.
- Finally, a scatter plot is created, where the x-axis represents the expected quantiles from the standard normal distribution, and the y-axis represents the observed quantiles from the residuals. Each point on the plot corresponds to a pair of quantiles.

- If the residuals are normally distributed, the points on the Q-Q plot should fall roughly along a straight line. Any deviation from linearity suggests a departure from normality. Specifically, the following scenarios may be observed:

- If the points on the plot deviate from a straight line in a systematic way (e.g., curved or S-shaped pattern), it indicates a departure from normality and suggests that the residuals do not follow a normal distribution.
- If the points on the plot deviate from a straight line only in the tails, it suggests heavy-tailed or light-tailed residuals.
- If the points on the plot deviate from a straight line near the center, it indicates a skewness in the distribution of residuals.