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?

A.

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

A. removes the first column which is created for the actual column.

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

A. Temp

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

A.

- By Residual analysis, building a histogram plot between Error and Error terms, which we got from the predicted value Y-TRAINED sets
- Recursive feature elimination

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

General Subjective Questions

1. Explain the linear regression algorithm in detail.

A. A simple linear regression model attempts to explain the relationship between a dependent and an independent variable using a straight line.

Y = mx + c
y = how far up
x = how far along
m = Slope or Gradient (how steep the line is)
c = value of **y** when **x=0 or intercept point**

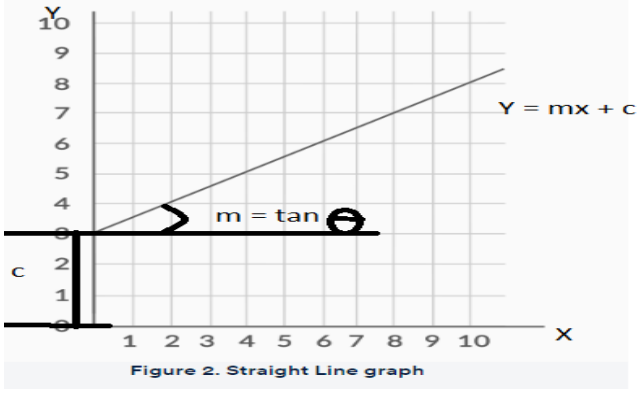

Figure 2. Straight Line graph

2. Explain the Anscombe's quartet in detail.

   Anscombe's Quartet can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on scatter plots.

3. What is Pearson's R?

A. The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   a) It is a step of data Pre-Processing which is applied to independent so it speeds up the calculation
   b) To keep data in equal level/range
   c) Normalized scaling = > x = (x – min(x))/(Max(x) – min(x))
      Standardized scaling => x= (x – mean(x))/sd(x)

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

A. This shows a perfect correlation between two independent variables.

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

A. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot.

   provides a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.