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: 1. Fall has the largest mean count of bikes as well as on an average higher value of distribution
2. Year 2009 saw higher sales than 2008.
3. Not significant difference between holiday and no holiday on sales.
4. Weather suitable for bike sales is 'clear, few clouds, partly cloudy

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

Ans: Because, for n levels in a variable n-1 levels of dummy variables are enough that is why we use 'drop_first=True'

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

Ans: 'temp'

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

Ans: I checked the Adj. R2 value, the p value of variables and VIF factors for variables. I removed variables one by one so all the variables have a p-value of less than 5% and VIF value less than 5.

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: 1. Sep, 2. Oct and 3. Yr.

General Subjective Questions

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

Ans: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

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

Ans: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points.

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

Ans: In Statistics, the Pearson's Correlation Coefficient is also referred to as **Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation**. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their

standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.
Pearson's Correlation Coefficient is named after Karl Pearson. He formulated the correlation coefficient from a related idea by Francis Galton in the 1880s.

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

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans: If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

6.What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans: In statistics, a Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. First, the set of intervals for the quantiles is chosen.