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)

    1. Bike demand takes a dip in spring.

    2. Bike demand in the year 2019 is higher as compared to 2018

    3. Bile demand is high in the months from May to October

    4. Bike demand is high if weather is clear or with mist cloudy, while it is low when there is light rain or light snow.

    5. Bike demand does not change whether day is working day or not.

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

Ans. If we do not use drop _ first= True then dummy variables will be created it is important.

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

Ans. Atemp and temp both have same correlation with target variable

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

Ans. Linearity of relationship between response and predictor variables.

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. Temperature (temp) – coefficient value of 0.5636 indicated that a unit increase in temp variable increases the bike hire numbers by 0.5636 units.

Weather situation (weathersit)- coefficient value of -0.3070 indicated that with respect to weathersit1 a unit increase in weathersit3 variable decreases the bike hire numbers by 0.3070 units

Year (yr) coefficient value of 0.2308 indicated that a unit increase in year variable increases the bike numbers by 0.2308 units.

General Subjective Questions

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

   Linear regression is a supervised machine learning method that is used to fine linear equation that best describes the correlation of the explanatory variables with the equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of the best fit is an iterative process.

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

3. What is Pearson's R? (3 marks)
   In statistics, the pearson correlation coefficient is a correlation coefficient that measures linear correlation between two sets of data. It is ratio between the covariance of two variables and the product of their standard deviations. It is essentially a normalized measurement of the covariance, such that the result always has a value between -1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   It is a step of data pre-processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modeling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. It is important to note that scaling just affects the coefficients.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   (3 marks)
   An infinite VIF occurs when there is perfect multi-colinearity in the dataset. This means that the one predictor variable is an exact linear combination of other predictor variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   (3 marks)
   Q-Q plot compares the quantiles of the dataset to those of a theoretical distribution. In linear regression, it is used to assess the normality assumption of residuals.