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)**
   - Yr-Count of bike rented is significantly increased and popular in the 2019 when compared to the year 2018.
   - Weathersit- Count of bikes rented is high during the “Clear weather”.
   - Season- count of bikes rented is high in “Fall” and secondly high in “Summer”.
   - Holiday- Bikes are rented high on non-holidays compared to holidays.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   - For a variable of n levels, dummy variable can be represented by n-1. With the removing of first column also, data can be represented.
   - drop_first=True is used to eliminate the redundant columns.
   - Also, to avoid multicollinearity among the dummy variables.

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

“temp” has the highest correlation with the target variable “cnt” of about 0.5928.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   - The distribution of calculated residual values (error terms) is plotted.
   - And it follows a Normal distribution and centered at mean 0
   - Also, a linear relationship is seen between actual and predicted 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)**
   - Temp

- Year
- Humidity

General Subjective Questions

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

A linear regression algorithm explains the relationship between the independent variables and dependent variable using a straight line. It works on the numerical variables, hence categorical variables need to be converted to numerical using the term called “dummy variables”. And following steps are carried,

- The given dataset is divided into train and test set. And, then the train data is divided into independent and dependent(target) dataset.
- A linear model is fitted in the training dataset. To find the coefficient of the best fit line, minimizing of the cost function is done. Example of cost function is residual sum of squares.
- In case of multiple independent variables, the predicted variable is a “hyperplane” instead of line.
- Model is built and fit with the significant P and r2 values.
- Residual Analysis is done and assumptions are verified by the distribution plot.
- The predicted variable is then compared with test dataset and assumptions are made finally.

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

- Anscombe’s quartet comprises of four datasets which has nearly similar descriptive statistical values but have quite different distribution when visualized graphically.
- The statistics consists of (mean, sample variance of x and y, correlation coefficient, linear regression line and r-square vale).

- Anscombe’s quartet shows multiple datasets which has similar statistical properties but still can be vastly different from one another when it is graphed.

3. What is Pearson’s R? (3 marks)

- Pearson’s R measures the strength of association or correlation between two variables.
- It is the covariance of two variables divided by the product of their standard deviation.
- It has a value from (+1 to -1).
- A value of 1 means a positive correlation. That is, if one variable increases, the other variable will also increase.
- A value of 0 means no correlation among the variable.
- A value of -1 means a negative correlation. That is, if one variable increases, the other variable will decrease.

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

- Scaling of variable is done to keep the variable in certain range.
- It is a preprocessing step in linear regression analysis. And the reason for scaling the variables is for the faster computation of gradient descent. The time taken for gradient descent algorithm is low (when the data range is between 0-1) and high (when the data range is between 0-1000)
- **Normalized scaling-** Called MinMax scaling, the value is bound between 0 to 1. The outliers are scaled too. Good for non-gaussian distribution.
- **standardized scaling-** Values are centered at mean with a unit standard deviation. Values are not bounded, Good for gaussian distribution and do not affect outliers.

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

- VIF value becomes infinite if R-squared value is 1.

- It means there is a perfect correlation between the features.
- The formula for calculating VIF = 1/1-R2.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**
    - A Q-Q plot is a scatter plot of two sets of quantiles against each other.
    - It is mainly used to check if the two sets of data came from the same distribution.
    - It is the visual check on the data.
    - If the data is from the same source, it will appear as a Straight line.