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?**

   - The demand for the bike is increased in the Year 2019 compared to Year 2018.
   - The demand for the bike is more during Summer, Fall and Winter compare to Spring Season

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

   - Dummy variables are useful because it allows us to use a single regression equation to represent multiple groups
   - drop_first=True is important to use, as it helps in dropping the additional column created during dummy variable creation.

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

   - From the pair plot that was defined in the python file, the temp variable has highest positive correlation with target variable cnt

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

   - To validate the assumptions of Linear Regression, build a scatter plot between y_test and y_pred after building the training set. It will show the Linear Regression model that is built is suitable for the dataset.

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

   - The top 3 features contributing significantly towards the demands of the share bikes are

     - Year 2019 (Positive Correlation) – More bikes are rented during this year
     - Temp (Positive Correlation) – Temp plays an important role in bike rentals
     - Light Snow (Negative Correlation) – During Snow time bike rentals are not that good

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   - Linear Regression model is a Machine Learning algorithm used for supervised learning.
   - Linear regression is used to predict a target variable (Y) based on the given independent variables(X). In this target variable is dependent on the independent variables.
   - A linear relationship between a dependent variable and the other given independent variables is found using this regression technique

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
- Even though they look similar but there is some uniqueness in the dataset
- They have very different distributions and appear differently when plotted on scatter plots

3. **What is Pearson's R?**

- The Pearson correlation coefficient (PCC) — also known as Pearson's R, is a measure of linear correlation between two sets of data.
- It is the ratio between the covariance of two variables and the product of their standard deviations;
- It is essentially a normalized measurement of the covariance; their result always has a value between −1 and 1.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

- *Scaling is step to make the independent variables normalize and bring to a particular range*
- It also helps in speeding up the calculations in an algorithm
- ***Normalization*** *rescales* the values into a range of [0,1] whereas ***Standardization*** *rescales* data with mean of 0 and a standard deviation of 1
- ***Normalization*** *scaling is useful when we don't know about the distribution whereas* ***Standardization*** *scaling is useful when the feature distribution is normal*

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

- When the correlation is perfect between two variables then VIF will be equal to infinity.
- During perfect correlation, the R2 will be 1, which lead to 1/(1-R2) infinity.
- An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

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

- Quantile-Quantile (Q-Q) plot, is a graphical tool which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.
- Also, it helps to determine if two data sets come from populations with a common distribution