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)**
   - There was a gradual increase in the usage of bike rentals from the year 2018 and 2019
   - The weekdays column was not significant for the model building
   - Holidays field was not that effective when compared to the workingdays fields

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   - The usage of drop_first=True will create (n-1) Indicator variable during the dummy variable creation using pd.get_dummies(), here n stands for the number of unique values in a categorical variable

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

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   - The assumptions of the Linear Regression model was validated by
     i. The error terms was normally distributed
     ii. Pattern of error
     iii. Multicollinearity
5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**
   - The top features of contribution were atemp, year, month

General Subjective Questions

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

   - Linear regression is a basic machine learning algorithm that makes use of the equation of a straight line with one independent variable in the case of a Simple Linear regression or multiple independent variables in the case of a Multiple Linear Regression.
   - The equation is given by y = β0 + β1.X1 +…..+βn.Xn+ error
   - There are certain assumptions that are made for the Linear regression model to stay valid with one independent variable
     - o There should be a linear relationship between X and y

  - Error term should be normally distributed
  - Each error term should be independent i.e., no visible pattern
  - Error term should have constant variance
- For Multiple Linear Regression in addition to the assumptions for SLR the below assumptions should also be considered
  - Model should not overfit or underfit
  - Multicollinearity i.e., the association between the Predictor variables should be taken into consideration

2. **Explain the Anscombe's quartet in detail. (3 marks)**
   - Anscombe's quartet is a group of four dataset which are identical in descriptive statistics, whereas they have certain peculiar features in the dataset that fools the regression model when built
   - Example let us assume a simple linear model with the variation in the β0 and β1 coefficients, the model built using them will have a best fit line that will drastically vary on the scatter plot

3. **What is Pearson's R? (3 marks)**
   - Pearson's coefficient is a measure of the association between two continuous independent variables, and it is based on the method of covariance i.e., gives information about the magnitude and the direction of association.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**
   - Scaling is the concept of reducing the value between a range of value in order to have a small value of coefficients. There are many ways to do scaling e.g. MinMaxScaler (normalized), StandardScaler, etc
   - The major difference between Minmax and standard scaler is as follows

MinMaxScaler (normalized)	StandardScaler
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**
   - VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity.
   - When the VIF goes infinity it means there is a perfect correlation and the two comparing features are just the same.

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 plot of the quantiles of the first data set against the quantiles of the second data set. statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.
   - The importance of q-q plot is
     i. It can be used with sample sizes.
     ii. Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.