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)
    - After pandemic, demand is twice next year.
    - Summer and fall seasons are the best time and has high demand.
    - During spring the demand goes down rapidly
    - Weekdays have high demand and holidays have negative impact on demand.
    - When the weather is snowy or misty, it has a negative impact on the demand
    - Lower windspeed leads to greater demand
    - From the month of May to October the demand keeps on increasing

2. **Why is it important to use drop_first=True during dummy variable creation?** (2 mark)
    - On using drop_first=True, during creation, the original column from which the dummy columns are derived will be dropped
    - Once dummy variables are create, the original column is of no use. So it must be removed

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?** (1 mark)
    - **Temp** column have 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)
    - R-square value is 0.83. which is a very good value representing a strong model
    - The probability of F-statistics is way less than 0. This assures that the model is valid and proves that the outcome is not by chance or luck
    - The p-value and VIF of all the features are within acceptable range
    - The error is normally distributed
    - The predicted value and target value is closer to the best fit line in a linear regression pattern

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?** (2 marks)
    - **Temperature, season, weather** are the top 3 features contributing significantly towards explaining the demand of the shared bikes

General Subjective Questions

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

- Linear regression is a statistical method that is used to predict a continuous dependent variable(target variable→x) based on one or more independent variables(predictor variables→(y)).
- This method is used in predictive analysis of machine learning model
- There are 2 types of regression
    - Simple Linear Regression – The target variable depends on a single independent variable

      $$Y = \beta_0 + \beta_1 X_1$$

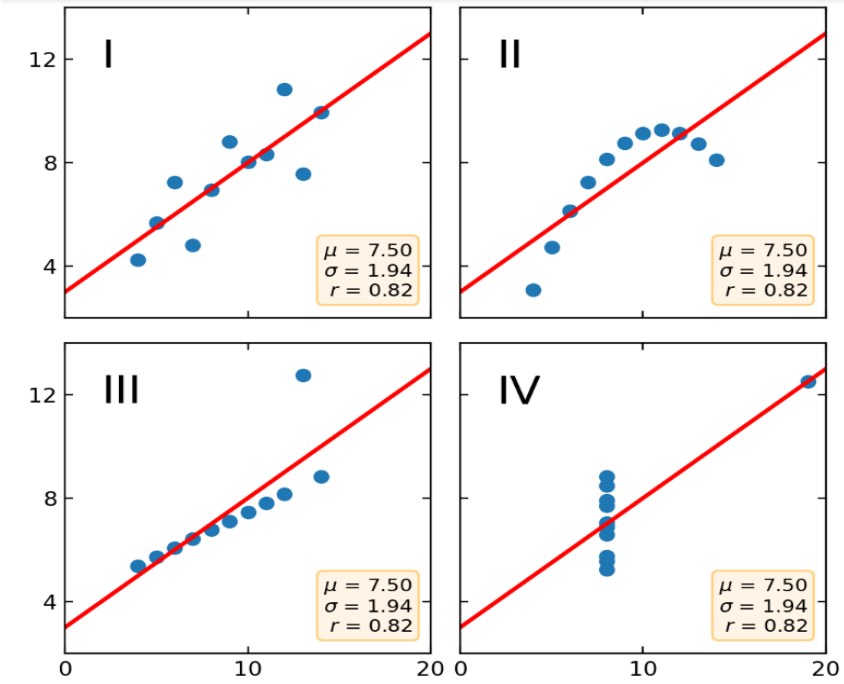
    - Multiple Linear Regression – The target variable depends on more than one independent variable

      $$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

- Rules defining Linear Regression
    - Little or no multicollinearity → The independent variables should not be much related to each other
    - No outliers should be present in data
    - Homoscedasticity → The error(difference between predicted variable and target variable), should be independent on each other

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

- Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.
- This explains the importance of data visualization before building a model only based on numbers.
- Anscombe's quartet suggests to check for anaomalies like outliers, data diversity



3. **What is Pearson's R?** (3 marks)
   - In statistics, the **Pearson correlation coefficient** (**PCC**) is a correlation coefficient that measures linear correlation between two sets of data.
   - It is a normalized **measurement** of the covariance, so the result always has a value between −1 and 1.
     - 1 indicates a strong positive relationship.
     - -1 indicates a strong negative relationship.
     - A result of zero indicates no relationship at all.
   - It describes the strength and direction of the linear relationship between two quantitative variables.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[\, n\sum x^2 - (\sum x)^2 \,][\, n\sum y^2 - (\sum y)^2 \,]}}$$

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?** (3 marks)
   - Feature scaling is a method used to normalize the range of independent variables or features of data
   - It refers to putting the feature values into the same range which would help in more appropriate model building
   - There are 2 types of Scaling
     - Normalisation or MinMax scaling → It consists of rescaling the range of features to scale the range between 0 and 1

$$x' = a + \frac{(x - \min(x))(b - a)}{\max(x) - \min(x)}$$

     - Standardisation → Feature standardization makes the values of each feature in the data have zero mean and unit variance

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?** (3 marks)
   - VIF is used to calculate the multicollinearity between variables.
   - So if the variables are extremely independent of each other then VIF will be 0.
   - But if there is perfect correlation between the independent variables then VIF will be infinity
   - For example, in the given assignment temp and atemp column had infinite VIF initially

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.** (3 marks)
   - A QQ plot is a scatterplot created by plotting two sets of quantiles against one another
   - If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x.