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)**
   - Based on dataset usage of bikes in 2019 are more than 2018
   - Spring season has the lowest demand
   - In weekdays bike usage is high
2. **Why is it important to use drop_first=True during dummy variable creation? (2 marks)**
   - Helps to reduce the correlations created 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)**
   - Registered has the highest correlation
4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   - There should be no correlation between the residuals
   - The error terms must be normally distributed
5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**
   - Negative correlation in (Light rain Light snow Thunderstorm)
   - Positive correlation in (month 3 and 10)

==========

s**General Subjective Questions**

1. **Explain the linear regression algorithm in detail. (4 marks)**
   - we train a model to predict the behavior of your data based on some variables. Regression models is a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables.
   - To calculate best-fit line linear regression Y = Mx + C where m is intercept and c is coefficients.
   - If there is a single input variable, then it is called simple linear regression.
   - If there is more than one input variable, it is called multiple linear regression.
2. **Explain the Anscombe's quartet in detail. (3 marks)**
   - Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed.
   - The basic thing to analyze about these datasets is that they all share the same descriptive statistics but different graphical representation.
3. **What is Pearson's R? (3 marks)**
   - Represents the relationship between two variables that are measured on the same interval or ratio scale and represented in the same way as a correlation coefficient that is used in linear regression, ranging from -1 to +1
   - Positive correlations indicate that both variables move in the same direction. Conversely, a value of -1 represents a perfect negative relationship.

- Negative correlations indicate that as one variable increases, the other decreases; they are inversely related. A zero indicates no correlation.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**
   - A step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.
   - If scaling is not performed, then algorithm considers one and this will lead to incorrect modelling.
   - It brings all of the data in the range of 0 and 1, sklearn.preprocessing.MinMaxScaler helps to implement normalization

     x = x-min(x) / max(x) – min(x)
   - Standardization replaces the values by their Z scores and sklearn.preprocessing.scale helps to implement standardization

     x = x-mean(x) / sd(x)
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.
   - If the VIF is infinite, the that shows a perfect correlation between two independent variables
   - To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**
   - Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile
   - To find out if two sets of data come from the same distribution the points will fall on that reference line.
   - This helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.