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)
   - Month and Week of the days variable does not have any effect on the dependent variable.
   - On the other hand season and weather plays major role on the dependent variable.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
   - Dummy variables are created to represent the categorical variable in 0 & 1
   - e.g. If there are 3 possible grades for the assignment. Pass, Fail, Distinction

#	Pass	Fail	Distinction
1	1	0	0
2	0	1	0
3	0	0	0

   - Even after removing the first column, it is possible to represent the grades using dummy variables.
   - Dropping the first column reduces the extra column created during dummy variable creation and reduces the 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 variables have highest corelation with target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   - Errors are normally distributed
   - Error terms are independent of each other
   - Error terms have constant variance
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   - Humidity
   - Temperature
   - Weekend

General Subjective Questions

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

- Linear Regression is a machine learning model based on supervised learning.
- It tries to predict the dependant variable based on one or more independent variables.
- This regression technique tries to find out the relationship between dependent and independent variable.
- Hypothesis function for linear regression: y = B0 + B1 * x

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

- Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set.
- It illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.
- Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

3. What is Pearson's R?(3 marks)

- Pearson's r is a numerical summary of the strength of the linear association between the variables.
- it assigns a value between − 1 and 1, where 0 is no correlation, 1 is total positive correlation, and − 1 is total negative correlation.
- This is interpreted as follows: a correlation value of 0.7 between two variables would indicate that a significant and positive relationship exists between the two. A positive correlation signifies that if variable A goes up, then B will also go up, whereas if the value of the correlation is negative, then if A increases, B decreases.

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

- **In machine learning,** feature scaling **refers to putting the feature values into the same range.** Scaling is extremely important for the algorithms considering the distances between observations like k-nearest neighbors. On the other hand, rule-based algorithms like decision trees are not affected by feature scaling.
- **In** normalization**, we map the minimum feature value to 0 and the maximum to 1. Hence, the feature values are mapped into the [0, 1] range:**

  z = x – x(min) / x(max) – x(min)
- **In** standardization**, we don't enforce the data into a definite range. Instead, we transform to have a mean of 0 and a standard deviation of 1:**

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

- If there is a perfect corelation between 2 independent variables then VIF becomes infinite.
- In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.
- To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

5. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

- *Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.*
- *This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.*