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)**
   - ⇨ The categorical variables like season, holiday, weekday, holiday and weather influence the dependent variable count.
     Below are the findings:
     - Bike are rented highest at fall.
     - Fall and summer witness almost similar amount of Bike rent.
     - Spring has least amount of Bike rented.
     - Bike are rented highest at fall.
     - Fall and summer witness almost similar amount of Bike rent.
     - Spring has least amount of Bike rented.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   - ⇨ Dropfirst=True is important as it reduces the removal of extra column during creation of a dummy variable.

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

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   - ⇨ Validation of R-squared : Higher the R-squared value the better.
   - ⇨ Probability(F-statistic): Next is to validate the probability of F-Statistic value. It has to be low value, which signifies the fitness of the model.
     If the Probability is more than 0.05 it signifies the model is not fit.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**
   - ⇨ Bike renting has increased over the year.
   - ⇨ Bike renting is highest at Sept and lowest at the starting of the month which increases as the month increase.
   - ⇨ Temperature has an direct effect on Bike renting. i.e. With increase in temperature the bike rent increases.

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   - ⇨ Linear Regression algorithm is one of mostly used Machine learning algorithm. It is based on supervised learning.
     It helps in predictive analysis based on the past data. The output variable predicted is a continuous variable.
     It helps in finding the relation between two or more continuous variable and predicting the possible future outcome.
     There are two types of Regression:
     a) Simple Linear regression: has only 1 independent variable.
     b) Multiple Linear regression: has more than 1 independent variable.

2. **Explain the Anscombe's quartet in detail. (3 marks)**
   - ⇨ Anscombe's quartet consists of four data set which seems almost identical statistical properties, yet differ from each other when plotted in a graph.
     It signifies the importance of Data visualization before choosing any Machine learning algorithm.

3. **What is Pearson's R? (3 marks)**
   - ⇨ Pearson's R also known as Pearson Corelation Coefficient is used to measure the linear corelation between two data set.
     It is used to measure how strong is the relation between two variables.
     The value of the Pearson Corelation Coefficient is always between -1 and 1.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**
   - ⇨ Scaling is the process to standardize or normalize the data to fit under a same range.
   - ⇨ It is a pre-processing process and is performed to handle large variance in the data set and to fit the data under the same range.
   - ⇨ Normalized scaling:
     It is a type of scaling, where maximum and minimum values are used for scaling.
     It is used when there is no or negligible number of outliers.
   - ⇨ Standardized Scaling:
     It is a type of scaling which uses Mean and Standard Deviation for scaling
     It is not affected by of outliers as Normalized scaling.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**
   - ⇨ Sometimes VIF value is infinite indicates that the variable is strongly co-related, i.e. R2=1
   - ⇨ VIF is calculated as :
     VIF=1/(1-R2)
     If the R2 value is 1, VIF would be infinite.

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

- ⇨ Q-Q plot also known as Quantile-Quantile plot is the graphical representation of two data set. It helps to determine if the two different data set come from population with same distribution.
- ⇨ Importance:
    - It helps to determine if the two different data set come from population with same distribution.
    - The sample size of two data set need not be same.