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)
   Ans: Below is the correlation of categorical variables on the dependent variable.
   Season (Summer, Fall and Winter) is showing positive coefficient of 0.11, 0.15 and 0.16 respectively.
   Light Snow is negatively correlated with negative coefficient of -0.19
   Mist + Cloudy weather is showing negative coefficient of -0.04

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   Ans: When converting categorical variable to dummy variable "n" categories can be identified by n-1 boolean.
   By using drop_first, 1st category can be explained by '0's for all other categories.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   Ans: "registered" variable has the highest correlation (0.945) with the target variable.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   Ans: The assumption of Linear Regression is validated through plotting Error term between Y_predicted and Y_actual and plotting distribution plot using bins to verify if it is normally distributed around 0.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**
   Ans: Top 3 features contributing significantly are listed below with their coefficients
   Casual (0.68),
   Weather with Light Snow (-0.19),
   WorkingDay (0.186),
   Winter (0.16)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   Ans: Linear Regression algorithm is to find the co-efficient of independent variables (x) with the dependent variable (y). The algorithm work with below assumptions;
   i) Data is uniformly distributed to avoid underfitting.
   ii) Dependent and independent variables are linearly correlated

   The algorithm works as below;

1. Remove Heteroscedasticity from the dataset. Remove any data which explains another data in the dataset.
2. Replace categorical variable with unordered list
3. Perform Train-Test split for model training and test
4. Standardize the train dataset using min-max scaling
5. Build Statistics model using recursive feature elimination
6. Remove features which has high P-Value and High VIF (Variance Inflation Factor) manually.
7. Repeat step 6 until we find the least number of feature defining target variable.
8. Evaluate the model on test-data set and check error term is normally distributed across 0.

2. Explain the Anscombe's quartet in detail. (3 marks)
   Ans: Anscombe's quarted are the list of variables which has same mean, Standard deviation and correlation coefficient but are qualitatively very different and defines relationship with target variable differently.

3. What is Pearson's R? (3 marks)
   Ans: Pearson's R is the correlation between 2 variables to measurement linear correlation and ranges between -1 and 1.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   Ans: Scaling is a standardizing technique to align all the features to common scale. Scaling is performed so the data can be feeded into machine algorithm for interpretation. Normalized Scaling is a technique to standardize data between -1 and 1 or 0 and 1. Standardized Scaling is a technique to scale data with mean as 0 and standard deviation of 1.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   Ans: VIF goes infinite when the denominator becomes 0, ie Ri-square term is 1. Which occurs only when variables are multi-collinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   Ans: Q-Q plot is a scatter plot between pair of variables. It explains the correlation of one variable over the other variable**.** If the variable are highly correlated then the pair of variables almost form a straight line.