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)

   Answer: Below is the inference from the analysis:

   a. in summer sale increase by 647.41 times
   b. in winter it increases by 1093.89 times
   c. when it snows rental decreases by around 2000 times
   d. when its mist rental decreases by around 450times

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

   Answer: in order to remove duplicate columns, which would further lead to collinearity

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

   Answer: temp has the maximum correlation

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

   Answer: we can validate by doing residual analysis of training data and then curving histogram to check normality of the curve.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer: 1. Temp outside

2. weather condition like snow

3. wind speed

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Answer: Linear regression is a basic form of machine learning where we train a model to predict the behaviour of the data based on some variables. Here y = mx+c, m is coefficient and c is intercept

   We divide the data into training and test dataset and use different methods like RFE etc for modelling

2. Explain the Anscombe's quartet in detail.

   Answer: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points

3. What is Pearson's R?

   Answer: Pearson R is a measure of the strength of the linear relationship between two variables on a sample which can range from -1 to 1

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

   Answer: Scaling is a technique to standardize the independent features present in the data in a fixed range. To bring

independent variable to same level which can help in managing units.

- Min-Max Normalization: This technique re-scales a feature or observation value with distribution value between 0 and 1.
- Standardization: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

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

   Answer: When there is perfect collinearity and R is 1

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

   Answer: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data which 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.