1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)
    1) As median of year. increased from 0 to 1. i.e. 2018 to 2019. From which we can say that demand of bike share increases in each year
    2) Demand almost same on Working day as well as non-working day however demand is decreased on holiday as it is indicated by median of holiday vs median on non-Holiday.
    3) There is highest demand on season on fall as well as low demand in rain fall.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
    Using "drop_first = True" is important because it helps in reducing the extra column while creating dummy variables to avoid multi-collinearity getting added to the model if all dummy variables are included.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
    'temp' has the highest correlation with the target variable 'cnt'. It is linearly related to the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
    After building the model on the training set, we checked for following assumptions.
    - Multicollinearity check with Variance Inflation Factor (VIF)
    - Checking R2 and adjusted R2 values for training and test set

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

    1)temp

    2)yr

    3)season

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
    Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables.
    It is one of the very basic forms of machine learning where we train a model to predict the behaviour of our data based on some variables.
    There are two types of linear regression algorithms:
    ▪ Simple Linear Regression: Single independent variable is used. o $Y = \beta 0 + \beta 1X$ is the line equation used for SLR.

▪ Multiple Linear Regression: Multiple independent variables are used. o Y= β0 + β1X1 +… + βpXp is the line equation for MLR.

Here, β0 = value of Y when X = 0 (Y intercept); β1, β2……, βp = Slope or the gradient.

2. Explain the Anscombe's quartet in detail. (3 marks)
3. 3. What is Pearson's R? (3 marks)

▪ The Pearson's R (also known as Pearson's correlation coefficients) measures the strength between the different variables and the relation with each other. It lies between -1

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

Scaling is a method to normalize the range of independent variables. It is performed to bring all the independent variables on a same scale. If scaling is not done, then it affects the coefficients.

The prediction and precision of prediction stays unaffected after scaling. There are two types of Scaling:

1. Min-Max scaling (aka Normalization):

   The Min-Max scaling normalizes the data within the range of 0 and 1. The Min-Max scaling helps to normalize the outliers as well.
2. Standardization:

   It converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1. 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)

If there is perfect correlation, then VIF is infinite whereas if all the independent variables are orthogonal to each other then VIF = 1. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1

VIF = 1/ 1–R^2
VIF=1/0
VIF=infinity

To solve this, 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)

The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another.

A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Whether the Distributions is Gaussian, Uniform, Exponential or even Pareto distribution