**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) 2. Why is it important to use **drop\_first=True**during dummy variable creation? (2 mark) 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation

with the target variable? (1 mark) 4. How did you validate the assumptions of Linear Regression after building the model on the

training set? (3 marks) 5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes? (2 marks)

**General Subjective Questions**

1. Explain the linear regression algorithm in detail. (4 marks) 2. Explain the Anscombe’s quartet in detail. (3 marks) 3. What is Pearson’s R? (3 marks) 4. What is scaling? Why is scaling performed? What is the difference between normalized scaling

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

(3 marks)

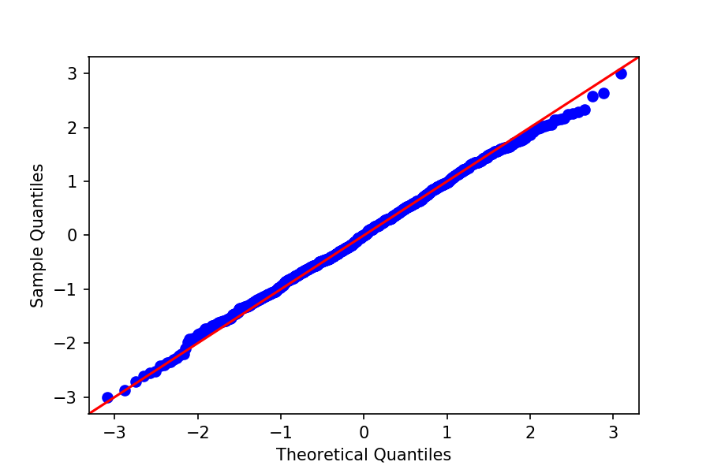
**6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

Q-Q plot, also known as Quantile-Quantile plot, is a graphical plot created by plotting two sets of quantiles against one.

It is used for determining if two datasets come from populations with a common distribution.

The q-q plot is used to answer the following questions:

* Do two data sets come from populations with a common distribution?
* Do two data sets have common location and scale?
* Do two data sets have similar distributional shapes?
* Do two data sets have similar tail behaviour?



For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members’ age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line (45-degree line). The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

The advantages of the q-q plot are:

* The sample sizes do not need to be equal.
* Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.