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. the total columns which are relevant for the prediction of target variable are ['const', 'yr', 'holiday', 'workingday', 'temp', 'hum', 'windspeed',

'season\_2', 'season\_4', 'weekday\_6', 'weathersit\_2', 'weathersit\_3',

'mnth\_3', 'mnth\_8', 'mnth\_9', 'mnth\_10'] and the R-Square is 0.815 on test data

1. **Why is it important to use drop\_first=True during dummy variable creation?** (2 mark)

Ans. by using drop\_first= True, it will drop the one column which has been create while dummification, this will reduce the redundant column

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

Ans. Temp variable has the highest correlation.

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

Ans. by checking the p value and Vif

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

Ans. yr, mnth\_3, Holiday

General Subjective Questions

1. **Explain the linear regression algorithm in detail**. (4 marks)

Ans. **Linear regression** is also a type of machine-learning algorithm more specifically a **supervised machine-learning algorithm** that learns from the labelled datasets and maps the data points to the most optimized linear functions. which can be used for prediction on new datasets.

First of we should know what supervised machine learning algorithms is. It is a type of machine learning where the algorithm learns from labelled data.  Labeled data means the dataset whose respective target value is already known. Supervised learning has two types:

* **Classification**: It predicts the class of the dataset based on the independent input variable. Class is the categorical or discrete values. like the image of an animal is a cat or dog?
* **Regression**: It predicts the continuous output variables based on the independent input variable. like the prediction of house prices based on different parameters like house age, distance from the main road, location, area, etc.

1. **Explain the Anscombe’s quartet in detail.** (3 marks)

Ans. Anscombe's quartet is a set of four datasets that have similar summary statistics, but appear very different when graphed. This demonstrates the importance of plotting data before analyzing it and building a model.

1. **What is Pearson’s R?** (3 marks)

Ans. The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction.

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

Ans. Feature scaling, also known as data normalization, is a technique used in machine learning to standardize the range of variables in a dataset. The goal is to make all variables have similar value ranges, which helps to eliminate biases caused by differences in measurement units, ranges, and distributions.

Scaling is a preprocessing step in machine learning that helps improve the accuracy and performance of models by normalizing data and making features equally important

Normalized scaling

Also known as Min-Max scaling, this technique scales numerical data to a range of 0 to 1. It's useful when the data distribution is unknown or not normally distributed. Normalization is good for maintaining the original shape of the data

Standardized scaling

This technique transforms data into a standard normal distribution, with a mean of 0 and a standard deviation of 1. It's useful when the data follows a Gaussian distribution. Standardization is good for optimizing algorithms like gradient descent, and for clustering models and distance-based classifiers.

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

Ans. A Variance Inflation Factor (VIF) of infinity indicates that there is perfect correlation between variables, and that one variable can be expressed as a linear combination of the other variables

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

Ans. he quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.