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)

1. Temperature affects the Business positively,
2. Raining, Humidity, Windspeed and Cloudy affects the Business negatively.
3. The Demand of Bikes is more in the Winter and Summer season, mostly user don't like to travel using Bikes in Rainy Day or Rainy Season.
4. Saturday rentals are more than other usual Days
5. High demand in 2019 than 2018

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

This reduces additional column created during dummy variable creation as the column can anyways be represented with n-1 dimensions

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

Tmp has the highest correlation

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

1. Error terms were normally distributed
2. No or low multicollinearity between variables
3. Homoscedasticity - No patterns in residual values
4. Independence of residuals

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. Winter
3. Sep

General Subjective Questions

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

Linear regression is a statistical method that uses a linear model to predict the value of a dependent variable based on the values of one or more independent variables. The linear model is a mathematical equation that describes the relationship between the dependent variable and the independent variables. The goal of linear regression is to find the best possible linear model that fits the data.

The linear regression algorithm works by minimizing the sum of squared errors (SSE). The SSE is a measure of how far the data points are from the linear model. The linear regression

algorithm finds the line that minimizes the SSE.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet is a set of four data sets that have the same mean, standard deviation, correlation coefficient, and regression line. However, the data sets look very different, which shows that these statistics alone are not enough to understand the relationship between two variables.

The four data sets are:

1. $y = x + 1$
2. $y = 3 + 0.5x + \epsilon$, where ϵ is normally distributed with mean 0 and standard deviation 0.1
3. $y = 4 + 0.5x + \epsilon$, where ϵ is uniformly distributed between -0.1 and 0.1
4. $y = 5 + 0.5x + \epsilon$, where ϵ is a mixture of two normal distributions, one with mean 0 and standard deviation 0.1 and the other with mean 0 and standard deviation 0.2

The first data set is a perfect linear relationship. The second data set is a linear relationship with some noise. The third data set is a linear relationship with a lot of noise. The fourth data set is not a linear relationship at all.

Anscombe's quartet shows that it is important to look at the data itself, not just the statistics, when trying to understand the relationship between two variables.

3. What is Pearson's R? (3 marks)

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

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

Scaling is the process of transforming data so that it has a common scale. This is often done in machine learning to make the data more compatible with different algorithms. Scaling can also be used to normalize data, which is the process of making the data have a mean of 0 and a standard deviation of 1.

There are two main types of scaling: normalized scaling and standardized scaling. Normalized scaling is the process of dividing each data point by its standard deviation. Standardized scaling is the process of subtracting the mean from each data point and then dividing it by the standard deviation.

Scaling is performed for several reasons:

- To make the data more compatible with different algorithms.
- To normalize the data, which can help to improve the accuracy of machine learning models.
- To reduce the effect of outliers on the data.

The difference between normalized scaling and standardized scaling is that normalized scaling divides each data point by its standard deviation, while standardized scaling subtracts the mean from each data point and then divides it by the standard deviation.

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

The value of VIF (variance inflation factor) can be infinite when there is perfect multicollinearity between two or more independent variables. This means that the variables are perfectly correlated with each other, and so there is no way to distinguish their individual effects on the dependent variable.

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

A Q-Q plot is a graphical comparison of two probability distributions. It is a type of quantile-quantile plot, which means that it plots the quantiles of one distribution against the quantiles of the other distribution. A Q-Q plot is used to assess the shape of the two distributions and to see if they are similar. In linear regression, a Q-Q plot can be used to assess the normality of the residuals. If the residuals are normally distributed, the Q-Q plot will be a straight line. If the residuals are not normally distributed, the Q-Q plot will deviate from a straight line.