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. Here are some inference for their effect on dependant variable:

> Fall Season has more number of bookings.

> During the start of the month we can see that the bookings are drastically increasing and then there is a small number of decrease in the last 3 months.

> Clear weather attracted more bookings.

> Almost similar number of bookings during the weekdays.

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

Ans. drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. We can also use any of the columns to drop not only the first column.

For example, there are three dummy variables A,B, and C. Removing any one of the three columns will decide what data is about that particular column if both does not have data. So we don't need any third identification it will be clear with the help of two columns itself.

3. **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 with the target variable.

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

Ans. I have validated based on :

> Error terms should be normally distributed.

> Multicolinearity

> No visible pattern in Residual values and linear relationship should be visible

> No auto-correlation

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

Ans. The top three features are:

> temperature

> winter season

> September month

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   Ans. Linear regression is a statistical algorithm used to model the relationship between a dependent variable and one or more independent variables. It aims to find a linear equation that best fits the given data points and can be used to predict the value of the dependent variable based on the values of the independent variables.The linear regression algorithm assumes that there is a linear relationship between the independent variables (also called features or predictors) and the dependent variable (also called the target variable or response variable). The equation for a simple linear regression with one independent variable can be written as:$y=mx + c$. The goal of linear regression is to estimate the values of m and c that minimize the difference between the predicted values ($\hat{y}$) and the actual values (y) of the dependent variable for the given data points. This difference is often referred to as the error or residual. Linear regression is widely used in various fields for tasks

such as predicting sales, analyzing trends, and understanding the relationship between variables. Linear regression can be extended to multiple independent variables, resulting in multiple linear regression. The equation for multiple linear regression becomes:$y = m_1x_1 + m_2x_2 + ... + m_nx_n + c$.

2. **Explain the Anscombe's quartet in detail. (3 marks)**
   Ans. Anscombe's quartet is a collection of four datasets that were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and the limitations of relying solely on summary statistics. Despite having nearly identical summary statistics, the four datasets have distinct patterns and characteristics, highlighting the need for exploratory data analysis.Each dataset in Anscombe's quartet consists of 11 (x, y) points. Let's explore the characteristics of each dataset:

Dataset I:
- x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
- y: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68

Dataset I is a simple linear relationship between x and y. When plotted, the points form a relatively straight line with a positive slope. The data appears to fit well to a linear regression model.

Dataset II:
- x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
- y: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74

Dataset II also exhibits a linear relationship between x and y, but with a slight upward curvature. The linearity is less apparent compared to Dataset I.

Dataset III:
- x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
- y: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73

Dataset III contains an apparent non-linear relationship between x and y. It demonstrates a clear pattern of a quadratic curve, with a slight bend upward.

Dataset IV:
- x: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8
- y: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89

Dataset IV challenges the notion that a scatter plot always indicates a relationship. While the x values are mostly the same (8), the y values vary considerably.

However, when examining the summary statistics, they are very similar to Dataset I.
The purpose of Anscombe's quartet is to emphasize the importance of visual exploration of data and not relying solely on summary statistics. Despite sharing similar statistical properties, the four datasets have distinct patterns when plotted, illustrating that summary statistics alone may not capture the complexities and nuances of the data. Therefore, visualizing the data graphically is crucial for understanding its characteristics and making accurate interpretations.

3. **What is Pearson's R? (3 marks)**
   Pearson's correlation coefficient, commonly referred to as Pearson's R or simply the correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. Pearson's R is a value that ranges between -1 and 1. The sign of the coefficient indicates the direction of the relationship, while the magnitude represents the strength of the relationship.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**
   Ans. Scaling, in the context of data analysis and machine learning, refers to the process of transforming the values of variables to a specific range or distribution. It is performed to ensure that all variables have a comparable scale, which can be beneficial for various reasons during data analysis and model building. Scaling is performed because of Comparable Magnitudes, Convergence Speeds, Feature Importance and Interpretability. The difference between Normalised and Standardised scaling are as follows: While normalized scaling brings the values within a specific range(0,1), standardized scaling centers the data around the mean and adjusts the scale by the standard deviation(mean = 0, std = 1). Standardized scaling is often preferred in scenarios where the shape and spread of the distribution are important, and when algorithms or models assume a standard normal distribution.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   (3 marks)
   Ans. In some cases, the VIF value can be infinite or extremely large. This occurs when there is perfect multicollinearity among the predictor variables. Perfect multicollinearity means that one or more independent variables in the regression model can be perfectly

predicted by a linear combination of other independent variables.When perfect multicollinearity exists, the regression model becomes unidentifiable, resulting in mathematical problems during the estimation of the coefficients. Specifically, the problem arises when calculating the inverse of the matrix used to estimate the coefficients. In this situation, the determinant of the matrix becomes zero, making the matrix non-invertible and leading to infinite VIF values.

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

Ans. A Q-Q plot (quantile-quantile plot) is a graphical tool used to assess whether a given dataset follows a particular theoretical distribution, such as a normal distribution. It is a plot of the quantiles of the observed data against the quantiles expected from the theoretical distribution.

In a Q-Q plot, the x-axis represents the expected quantiles from the theoretical distribution, while the y-axis represents the observed quantiles from the dataset. If the dataset perfectly follows the theoretical distribution, the points in the Q-Q plot will fall on a straight line. Deviations from the straight line indicate deviations from the assumed distribution.

The use and importance of Q-Q plots in linear regression are as follows:

> Outlier Detection

> Model Comparison

>Assumption check

>Distribution Assessment