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) Solution:**

- Bookings appear to have increased over the season of fall. And from 2018 to 2019, the number of bike reservations has significantly increased.
- The months of June, July, August, September, and October saw the highest number of bike reservations. It began to rise at the beginning of the year and continued until the middle of the year, at which point it gradually began to fall as the year came to a conclusion.
- Bookings increased when the weather was clear, which is charismatic.
- Bookings are higher on Thursday, Friday, and Sunday than on other days of the week.
- Bookings tend to be less frequent during holidays, which makes sense given that people may enjoy spending time with their loved ones during these times.
- The number of reservations appeared to be roughly equal on working and non-working days.
- 2019 brought in more bookings

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

Solution:

As we are aware, we may represent a column with 'n' categories using 'n-1' dummies. pd.get_dummies will generate 'n' dummies using the 'drop_first=True' setting. The remaining dummy variables become independent and allow for obvious comparisons between categories by ensuring that one dummy variable is eliminated as a reference category.

It's crucial to prevent multicollinearity problems, enhance model interpretability, and lessen model complexity while creating dummy variables.

For instance, using pd.get_dummies without drop_first=True would produce three dummy columns if the column "students" has three categories: "Class A," "Class B," and "Class C." The first dummy column ('Class A') is dropped when drop_first=True is used, allowing us to express these categories with only two dummies rather than three. This strategy streamlines the model.

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

(1 mark)

Solution:

'temperature' is the variable which has the highest correlation in the target variable.

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

I have assessed the validity of the Linear Regression Model based on the following five assumptions:

1. **Normality of error terms:**

- The assumption is that the residuals (error terms) should follow a normal distribution. - The residuals are guaranteed to have a symmetrical distribution with a constant variance due to normality.

2. **Multicollinearity check:**

- Multicollinearity refers to the presence of high correlations between predictor variables in the regression model.

- Significant multicollinearity can result in unstable parameter estimations and make it difficult to comprehend the impacts of individual variables, therefore it must be found and addressed. - In order to find and remove highly correlated predictors, I calculated the correlation matrix and looked at the variance inflation factor (VIF) values.

3. **Validation of linear relationship:**

- This assumption assumes a linear relationship between the predictor variables and the response variable.

- I investigated the relationship using scatter plots and by evaluating the linearity of residuals in relation to expected values.

- A linear relationship is indicated by a clearly linear pattern in scatter plots or by a haphazard distribution of residuals around the line.

4. **Homoscedasticity:**

- Homoscedasticity is the presumption that the residual variance is constant for all levels of the predictor variables.

5. **Independence of residuals:**

- This assumption assumes that the residuals are independent of each other, indicating no correlation or autocorrelation.

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

Solution:

Three key characteristics primarily affect the demand for shared bikes: Winter season, temperature, and September

General Subjective Questions

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

Simply said, linear regression is a supervised machine learning model that identifies the linear relationship between the dependent (y) and independent (x) variables by determining the best fit linear line between them.

There are two varieties of linear regression: simple and multiple.

When only one independent variable is present, simple linear regression must be used to determine its linear connection to the dependent variable.

In contrast, several independent variables are used in multiple linear regression to identify relationships.

Finding the best-fit linear line and the ideal intercept and coefficient values so that the error is minimised is the major goal of a linear regression model.

Error is the discrepancy between actual and predicted values, and the objective is to minimise this discrepancy.

Equation of Simple Linear Regression, where b_o is the intercept, b_1 is coefficient or slope, x is the independent variable and y is the dependent variable.

$$y = b_o + b_1 x$$

Equation of Multiple Linear Regression, where bo is the intercept, $b_1, b_2, b_3, b_4 \ldots, b_n$ are coefficients or slopes of the independent variables $x_1, x_2, x_3, x_4 \ldots, x_n$ and y is the dependent variable.

Y = b0 + b1x1+ b2x2 + .. + bnxn

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

Four datasets that share the same statistical characteristics but display different graphical patterns make up Anscombe's quartet. The statistician Francis Anscombe developed it in 1973 to draw attention to the value of data visualisation and to warn against relying exclusively on summary statistics.

The four datasets in Anscombe's quartet consist of eleven pairs of x and y values. Despite having the same mean, variance, correlation, and linear regression parameters, the datasets have different distributions and relationships between the variables.

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

The Pearson correlation coefficient, often known as Pearson's R, is a statistical metric that expresses the linear relationship between two continuous variables. It accepts values between -1 and 1, and is represented by the symbol "r".

One may calculate Pearson's R using the following formula:

r = (((Xi - X_mean) * (Yi - Y_mean))) / (sqrt((Xi - X_mean))*sqrt((Yi - Y_mean)*sqrt((Xi - X_mean)))

Where: - Xi and Yi are the two variables' individual values.

The two variables' means are X_mean and Y_mean.

- The summation symbol is indicated by.

Principal features of Pearson's R

1. Range: R has a value between -1 and 1. A value of 1 denotes a linear relationship between the variables that is positive, a value of 0 denotes a linear relationship that is negative, and so on.

2. Relationship Strength: The size of r represents how strong the linear relationship is. A value that is nearer to -1 or 1 denotes a stronger linear correlation, whereas values that are nearer to 0 denote a weaker connection.

3. Significance: A hypothesis test can be used to establish the statistical significance of Pearson's R. It aids in determining if the correlation was accidental or statistically significant. This determination is based on the p-value of the correlation coefficient. A p-value less than the selected level of significance (for example, 0.05) denotes a significant correlation.

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

In machine learning, the preprocessing step known as scaling entails converting a dataset's features to a uniform scale. It makes sure that all feature ranges are comparable, which is crucial for several machine learning algorithms and data analysis methods. Scaling is done to solve problems with the size and units of the features as well as to stop larger features from

influencing learning more than smaller features.

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

Solution:

The occurrence of a Variance Inflation Factor (VIF) with an infinite value is referred to as "perfect multicollinearity." It occurs when one or more independent variables in a regression model have an exact linear connection. Accurate computation of the VIF values for the impacted variables is not possible due to perfect multicollinearity because it results in unstable parameter estimates.

Several such situations that can lead to perfect multicollinearity and infinite VIF values include the following:

1. Duplicate or Redundant Variables: Redundancy occurs when the information provided by two or more variables in a dataset is identical or perfectly correlated. For instance, complete multicollinearity would be introduced if "height in inches" and "height in centimetres" were both independent variables.

2. Problems with Data Transformation: Improper data transformations can also produce perfect multicollinearity. For instance, you can achieve perfect multicollinearity if you divide a continuous variable into categorical bins and add each bin as an independent variable.

It is significant to highlight that whereas high VIF values (but not infinite ones) imply considerable multicollinearity between variables, infinite VIF values denote the presence of complete multicollinearity. In these situations, it is wise to evaluate how multicollinearity affects the model's performance and think about resolving it by deleting or changing the associated variables.

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

Solution:

Quantile-quantile plots, often known as Q-Q plots, are a graphical tool for evaluating how closely a dataset resembles a theoretical distribution like the normal distribution. In order to visually compare and identify deviations from the predicted distribution, it compares the quantiles of the actual data against the quantiles of the theoretical distribution.

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

1. Examining the assumption of normality
2. Finding outliers and skewness
3. Model Evaluation and Assumption Checking
4. Distribution Comparison

In conclusion, the Q-Q plot is a useful tool in linear regression for testing the validity of the normality assumption, identifying skewness and outliers, assessing the suitability of the model, and comparing distributions. It aids in comprehending the distributional properties of the residuals and sheds light on possible problems that could compromise the accuracy and dependability of the linear regression study.