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)

- Season: Fall has the highest demand for rental bikes.
- The demand for bikes increases till June. Demand Peaks in september then it is decreasing.
- Not much difference between weekdays and workingdays demands as mostly everyone has working days on weekdays.
- Clear weathersit has highest demand
- Demand is lowest on holiday.
- Demand is growing every year.

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

We don't need high correlations created among dummy variables and drop_first=True to avoid creating extra columns to achieve this.

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

'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)

I have validated using the following 4 assumptions

- Distribution of errors
- Multicollinearity
- Patterns in residual values
- Linear relationship validation

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

temp , workingday and season_summer

General Subjective Questions

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

One type of regression analysis where a dependent variable and an independent variable have a linear relationship is called linear regression. The fundamental objective of linear regression is to take into account the provided data points and create the trend line that best fits the data.

Let's imagine we have a dataset with details on the connection between X and Y. Numerous observations are made and documented regarding X and Y. Our training data will be this. Our objective is to create a model that can forecast the Y value given the X value. A regression line that will produce the least amount of error is obtained using the training data. Then, new data is applied using this linear equation. In other words, if we input X, our model should be able to predict Y with the least amount of error.The following equation is a representation of the linear regression model.

y=b0+b1X+e

Let's assume that there is a relationship between a student's study time and grades; regression analysis can explain this relationship. Regression analysis will provide us with a relationship that can be represented as a graph and used to predict the outcome of your data.

The goal here is to create a best fit trend line either by using mean squared error(MSE) to calculate the error model.

After the model is built we can determine the accuracy of the model by comparing the values and the predicted data.

R^2 and adjusted R^2 score can be used to determine the created model's performance.

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

Francis Anscombe, a statistician, created the Anscombe's quartet in 1973 to demonstrate the value of charting data before analyzing it and creating your model. The statistical observations in these four data sets are essentially identical, giving the same information (concerning variance and mean) for each x and y point in each data set. These data sets, however, seem substantially differently from one another when plotted.

The quartet by Anscombe emphasizes the value of data visualization before applying various algorithms to create models. In order to identify the numerous anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.), it is suggested that the data characteristics be plotted. Additionally, because it cannot handle any other type of data collection, linear regression may only be regarded as a fit for data that have linear connections.

The significance of visualization in data analysis is emphasized by this quartet. A thorough understanding of the dataset's structure can be obtained by looking at the data.

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

A numerical evaluation of the strength of the linear connection between the variables is provided by Pearson's r. The correlation coefficient will be positive if the variables tend to rise and fall together. The correlation coefficient will be negative if the variables have a tendency to rise and fall in opposition, with low values of one variable correlated with high values of the other.

R, the Pearson correlation coefficient, has a range of possible values between +1 and -1. There is no link between the two variables, as indicated by a value of 0. Positive associations have values greater than 0, meaning that if one variable's value rises, so does the value of the other. A result that is less than 0 denotes a negative connection, meaning that when one variable's value rises, the value of the other variable falls.

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

Feature scaling is a method for uniformly distributing the independent features in the data over a predetermined range. It is done as part of the pre-processing of the data to deal with extremely variable magnitudes, values, or units.

In the absence of feature scaling, a machine learning algorithm would often prioritize larger values over smaller ones, regardless of the unit of measurement.

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

VIF = infinity if there is perfect correlation. A high VIF score denotes a strong connection between the variables.

The presence of multicollinearity causes the variance of the model coefficient to be exaggerated by a factor of 6 if the VIF is 6.

VIF displays a complete correlation between two independent variables when its value is infinite. If the correlation is perfect, we have R-squared (R2) = 1, which results in 1/ (1-R2) infinite. To fix this, we must remove one of the variables from the dataset that is the source of this ideal multicollinearity.

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

A graphical method for assessing if two data sets originate from populations with a common distribution is the quantile-quantile (q-q) plot.

This can be used to find if two data sets have come from different or similar population distributions. We can use this to get more insights into the nature of difference between two data sets.