Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <From my analysis we can infer that the bike rentals are higher in the summer and in fall months when compared to the winter months. In the Sat, Wed and Thursday have the greater number of bikes rented when compared to the other days in the week. The year 2019 has the greater number of bikes rented when compared to year of 2018, also when the weather is clear the bike rentals are high.> (Do not edit)

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <drop first=True will help in reducing the extra column created during the process of dummy variable creation, this helps in avoiding dummy variable trap which occurs in multicollinearity > (Do not edit)

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <The temp variable has the highest correlation > (Do not edit)

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <I have validated it using the VIF , removed the variables which has the highest VIF and p-score , checked error distribution of residual also checked the linear relationship of dependent variable > (Do not edit)

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <The top 3 features contributing are temperature, year and weathersit (mist+cloudy)> (Do not edit)

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)

Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Linear Regression is an ML algorithm used for the supervised learning. It is used in predicting the target variables based on the provided dependent variables, the linear relationship between dependent variable and other given independent variables will be seen. There are two types of linear regression 1. Simple linear regression 2. Multiple linear regression. Simple linear regression is used when we have one dependent variable whereas multiple linear regression is used if we have more than one dependent variable and if the dependent variable establishes the strong correlation to the target variable. There are two types of linear relationship one is positive and the other one is the negative relationship; both the linear relationship can be used in the ML algorithm >

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

< Anscombe's quartet consists of four data sets that have nearly identical simple descriptive statistics but have very different distributions and appear very different when presented graphically. Each dataset consists of eleven points. The primary purpose of Anscombe's quartet is to illustrate the importance of looking at a set of data graphically before beginning the analysis process as the statistics merely does not give an accurate representation of two datasets being compared.
>

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

< Pearson correlation coefficient is a statistical measure that quantifies linear relationship between two continuous variables, the usage of this is highly found on statistics , machine learning and in data analysis >

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Scaling is the process of transforming the values of a dataset to specific range , It is used in adjusting the feature values so that they are comparable . The scaling is performed for the following reasons

1.To improve model performance
2.To avoid feature dominance

3.To improve Interpretability

4.For distance based algorithms

The Major difference between normalized scaling and the standardization scaling is in the normalized scaling the scaled data ranges from 0 to 1 where as in the standardization centers data to mean 0 and scales to unit variance. In the normalization data's are bounded in the fixed range , in the normalization data will have mean of 0 and std of 1 , in the normalization no assumption against the data distribution is made , in the standardization assumes data is approximately normal. >

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

< The value of VIF is infinite if there is a perfect correlation between the two independent variables. The R-squared value will be 1 in this case. This leads to VIF infinity as VIF equals to 1/(1-R2). This concept suggests that is there is a problem of multi-collinearity and one of these variables need to be dropped in order to define a working model for regression.>

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

< The quantile-quantile (Q-Q) plot are used to plot quantiles of a sample distribution with a theoretical distribution to determine if any dataset concerned follows any distribution such as normal, uniform or exponential distribution. It helps us determine if two datasets follow the same kind of distribution. It also helps to find out if the errors in dataset are normal in nature or not.>
