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:

- The **fall season** has seen a significant increase in bookings, with a noticeable rise in bookings from **2018 to 2019** across all seasons.
- **Clear weather** contributed to a higher number of bookings, which is expected.
- **Thursdays, Fridays, Saturdays** recorded more bookings compared to the start of the week, indicating a preference for weekend activities.
- **Non-holidays** typically saw fewer bookings, which is understandable as people often prefer spending time at home with family during holidays.
- **2019** experienced a significant increase in bookings compared to the previous year, suggesting positive business growth.

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: When we use this one columns will be reduced while creating dummy variables. The preferred dummy variables should be (n-1) where n is number of cateegories

---

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: atemp

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: validated on linearity, multi collinearity etc;

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: year, temp and winter

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 a statistical method used to understand the relationship between a dependent variable and one or more independent variables. It works by fitting a straight line (or hyperplane in multiple dimensions) to the data that minimizes the difference between the observed and predicted values. The goal is to find the best-fitting line that can predict the dependent variable based on the independent variables. The model assumes that this relationship is linear and that changes in the independent variables will produce proportional changes in the dependent variable. Linear regression is commonly used for forecasting and estimating trends. It's simple, interpretable, and effective for modeling data with a clear linear pattern. However, it can be sensitive to outliers and requires assumptions such as no multicollinearity between the predictors.

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 is a set of four datasets that have nearly identical simple descriptive statistics but very different distributions and visual patterns. Created by statistician Francis Anscombe in 1973, the quartet highlights the importance of visualizing data before drawing conclusions. Each dataset in the quartet has the same mean, variance, correlation, and regression line, but when plotted, they reveal strikingly different relationships between variables. This demonstrates that summary statistics alone can be misleading and that visual inspection of data is crucial in statistical analysis. Anscombe's quartet is often used in teaching to emphasize the value of exploratory data analysis.

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's R, also known as the Pearson correlation coefficient, measures the linear relationship between two continuous variables. It ranges from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. A positive value suggests that as one variable increases, the other does as well, while a negative value indicates that as one increases, the other decreases. Pearson's R assumes that the

relationship between the variables is linear and that the data is normally distributed. It is widely used in statistics to assess the strength and direction of linear correlations.

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 features into a common scale to improve the performance of machine learning models. It ensures that features with different units or ranges don't disproportionately affect the model. Normalized scaling rescales data to a fixed range, typically [0, 1], while standardized scaling transforms data to have a mean of 0 and a standard deviation of 1. Both techniques help in models that are sensitive to the scale of the data, like regression and neural networks.

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)

A VIF (Variance Inflation Factor) can be infinite when there is perfect multicollinearity, meaning one predictor variable is a perfect linear function of others. This leads to an undefined or infinite VIF value, indicating that the variable cannot be independently estimated without redundancy. It typically occurs in cases of exact linear relationships among features.

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)

A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, such as a normal distribution. In linear regression, it helps assess the normality of residuals by plotting their quantiles against the quantiles of a normal distribution. If the points lie roughly along a straight line, it indicates that the residuals are normally distributed, which is an important assumption for valid regression results. This helps to validate the model's reliability and ensures accurate inferences.
