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:

1. 2019 saw a notable increase in bookings compared to 2018, showing strong business growth and an expanding customer base.
2. Fall saw the highest number of bookings, and across all seasons, rentals increased significantly from 2018 to 2019.
3. The peak months for bookings were May, June, July, August, September, and October, with demand rising steadily from the start of the year, peaking mid-year, and then gradually declining toward the year's end.
4. Clear weather naturally attracted more bookings, as people preferred riding in good conditions.
5. Thursdays, Fridays, Saturdays, and Sundays had higher booking numbers compared to the start of the week, indicating that weekends and pre-weekend days are more popular for bike rentals.
6. Fewer bookings occurred on non-holidays, which makes sense on holidays, people likely have more free time for leisure activities, including bike rides.

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: It helps to reduce the extra columns created during the dummy variables and hence avoid redundancy.

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 with the target variable

---

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: Mainly I have checked the VIF and residuals to validate the assumptions of linear regression.

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 significantly towards the demand of the shared bikes were: temp, winter, and Sep.

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 supervised learning algorithm used to predict a target (dependent variable) based on one or more input features (independent variables) by establishing a linear relationship between them. It comes in two types: Simple Linear Regression, which involves a single independent variable, and Multiple Linear Regression, which uses multiple independent variables to predict the target. The relationship between these variables is represented by a regression line, where a positive linear relationship means the target variable increases as the independent variable increases, while a negative linear relationship means the target decreases as the independent variable increases. This technique is widely applied in fields such as finance, sales forecasting, and trend analysis due to its simplicity and effectiveness.

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 the 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's Correlation Coefficient is a statistical measure that quantifies the strength and direction of a linear relationship between two variables. It ranges from -1 to +1, where +1 indicates a perfect positive correlation (both variables increase together), 0 means no correlation (changes in one variable do not affect the other), and -1 represents a perfect negative correlation (one variable increases as the other decreases). This coefficient is widely used in fields like economics, healthcare, and social sciences to analyze relationships between variables, such as height and weight, stock prices, or temperature and energy consumption. However, it is important to remember that correlation does not imply causation, meaning that even if two variables are strongly correlated, one does not necessarily cause changes in the other.

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 a technique performed in pre-processing during building a machine learning model to standardize the independent feature variables in the dataset in a fixed range. The dataset could have several features which are highly ranging between high magnitudes and units. If there is no scaling performed on this data, it leads to incorrect modelling as there will be some mismatches in the units of all the features involved in the model. The difference between normalization and standardization is that while normalization brings all the data points in a range between 0 and 1, standardization replaces the values with their Z scores.

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 when there is a perfect correlation between the two independent variables. The R-squared value is 1 in this case. This leads to VIF infinity as VIF equals 1/(1-R2). This concept suggests that there is a problem of multi-collinearity and one of these variables need to be dropped 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.