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: <Your answer for Question 1 goes below this line> (Do not edit)

1. Seasons – Season 1 sees lowest number of bookings, 3 and 4 see high bookings and are similar
2. Holiday – much lesser bookings on holidays
3. Yr - Have seen good year-on-year growth
4. Weekdays – not much pattern across weekdays
5. Months – not much pattern across months
6. Weathersit – in weather situation 3, getting almost no bookings

---

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: <Your answer for Question 2 goes below this line> (Do not edit)

1. Since it is not adding any new value in the data, other binary variables suffice the definition
2. To reduce the number of variables to be learnt by the model

---

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: <Your answer for Question 3 goes below this line> (Do not edit)

1. Temp and atemp both show same level of correlation – 0.63 highest metric

---

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: <Your answer for Question 4 goes below this line> (Do not edit)

1. Assumptions are built by looking at the data, meaning feature level correlation with the target. Feature level coefficients after training the model are compared with one-one correlation to conclude model learnt correct parameters
2. EG –
    a. Atemp has 0.63 correlation – and trained model has coeff of 0.344 – positive and high coeff validates the hypothesis
    b. Year-on-year change – in EDA shows increase in bookings in second year – validated by model coeff of 1.08
    c. Humidity and wind speed – has negative one-one correlation and have negative coeff of -0.1 and -0.08 in the model

---

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: <Your answer for Question 5 goes below this line> (Do not edit)

1. Weathersit_3 : snow/rainy weather has highest negative correlation of -1.19, meaning less bookings in hard weather
2. Year : because of year on year increase, it has high positive correlation of 1.08
3. Season_4 : winter season has high number of bookings

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)

In a linear regression model, we assume a linear relationship between the dependent and independent variables. In single linear regression, we try to fit a straight line between dependent and independent variable and in multi linear regression, this is a multi dimension plane.
This line or hyperplane is fit in a way to reduce the delta (error) between the line/plane and individual points of the target variable.

- Simple Linear Regression: This involves a single independent variable X used to predict a single dependent variable Y. The relationship is represented as a straight line.
- Multiple Linear Regression: This involves multiple independent variables \(X_1, X_2, \dots, X_n \) used to predict the dependent variable Y. Here, the prediction is a weighted sum of the inputs.

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 designed by statistician Francis Anscombe to demonstrate the importance of data visualization in statistical analysis. Despite having nearly identical basic statistical properties, each dataset in the quartet exhibits a very different pattern when visualized. This highlights the limitations of relying solely on summary statistics to understand data and the importance of graphing data to reveal patterns, outliers, and relationships.

- Mean of X and Y: The mean of both X and Y is almost the same across all four datasets.
- Variance of X and Y: The variance of both X and Y is nearly identical across all datasets.
- Correlation between X and Y: The correlation coefficient between X and Y is approximately the same for each dataset.
- Linear Regression Equation: The linear regression line (slope and intercept) for predicting Y from X is very similar in all datasets.

Even after all these similarities, data in 4 dataset is different, which one can see only by visualizing each

---

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 strength and direction of the linear relationship between two continuous variables. Developed by Karl Pearson, this statistic ranges from -1 to 1, where:

- 1 indicates a perfect positive linear relationship (as one variable increases, the other also increases).
- -1 indicates a perfect negative linear relationship (as one variable increases, the other decreases).
- 0 indicates no linear relationship.

Pearson's R is calculated by dividing the covariance of the two variables by the product of their standard deviations, resulting in a dimensionless measure of correlation.

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)

1. Scaling is done to bring all features on the same scale, to ensure that while model being trained it does not over value a certain feature only because of its scale and not by its correlation with target variable
2. It overall ensures correct parameter coefficients while training
3. In normalized scaling – scaled values are almost normally distributed between 0 and 1 with 50% of data lying in 1 std from median
4. In standard scaler – we subtract mean of values from each value and divide it by min subtracted from max. Thus, output has values between -1 and 1
5. Standard scaler handles outliers better by reducing the scales around tail

---

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)

If the VIF for a variable is infinite, it usually indicates perfect multicollinearity—meaning that the variable is a perfect linear combination of one or more other variables.

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 plot, or quantile-quantile plot, is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. In a Q-Q plot, quantiles of the observed data are plotted against the quantiles of a specified theoretical distribution.

In Quantiles from the data are calculated and plotted on the y-axis, while the quantiles from the theoretical distribution are plotted on the x-axis.

If the data distribution matches the theoretical distribution, the points in the Q-Q plot should lie approximately along a straight line at a 45° angle.

In linear regression, several assumptions must be satisfied to obtain reliable model estimates. One of the key assumptions is that the residuals (errors) follow a normal distribution. This is important for accurate hypothesis testing, confidence intervals, and prediction intervals. A Q-Q plot helps verify this assumption by assessing the normality of residuals.
