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 analysis of categorical variables reveals that:

Year (yr): The demand for bikes is significantly higher in 2019 compared to 2018.

Season: Bike demand tends to be higher in summer and winter, whereas it is lower in spring.

Weather situation (weathersit): Adverse weather conditions (e.g., heavy rain, snow) have a significant negative impact on bike demand.

Holiday: Bike demand decreases on holidays compared to non-holidays.

Working Day: Bike demand slightly decreases on working days.

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: Using drop_first=True when creating dummy variables helps to avoid the so-called "dummy variable trap." This trap happens when there's multicollinearity among the dummy variables, which means they provide redundant information to the model. By dropping the first category, we ensure that the model doesn't include these redundant variables. This step allows us to interpret the regression coefficients more accurately and results in a more stable and reliable 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: From the pair-plot analysis, temperature (temp) has the highest correlation with the target variable (cnt). This indicates that as temperature increases, bike demand also tends to increase.

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:

To validate the assumptions of Linear Regression:

Linearity: Checked scatter plots of predictors against residuals to ensure linear relationships.

Homoscedasticity: Plotted residuals versus predicted values to check for constant variance.

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: Based on the final model, the top 3 features significantly contributing to bike demand are:

Year (yr)

Weather situation (weathersit)

Spring season (season_spring)

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Linear regression is a method used to understand and predict the relationship between one outcome (called the dependent variable) and one or more influencing factors (called independent variables). The goal is to find the best-fitting straight-line equation that explains how these factors influence the outcome
The coefficients are calculated using a process called least squares, which finds the values that minimize the gap (residuals) between the actual data and the predictions made by the equation. This way, the model is as accurate as possible in representing the observed data.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer:

set of four datasets that share almost identical statistical properties, such as the mean, variance, correlation, regression line, and R^2. However, when plotted, they look completely different. This demonstrates how relying only on numbers can give a misleading picture, and why visualizing data is essential.

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)

It's handy to use Pearson's R, or Pearson's correlation coefficient, to find the linear or straight-line relation of two variables with one another. That way you get a number that ranges between -1 and 1:

- If you get 1 then the relationship is absolutely perfect, positive that's one thing: if that goes up, the other will as well.

A value of -1 indicates a perfect negative relationship, in which one increases as the other decreases.

A value of 0 indicates no linear relationship whatever.

This coefficient is so useful in understanding both the
strength of the relationship and the direction that two variables are connected and whether they are close by or pretty much independent.

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 adjusts the range of features in the data to a standard range. It is done in order to ensure that each feature contributes equally to the model and to enhance the performance of algorithms sensitive to the scale of data.

Normalized scaling or Min-Max Scaling: Scales the data to a fixed range, typically in the range [0, 1].

Standardized scaling or Z-score normalization: Centers the data around the mean with a unit standard deviation.

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)

An infinite VIF value indicates perfect multicollinearity, meaning that one predictor variable is a perfect linear combination of one or more other predictors. This situation often arises due to exact duplication of variables or perfect correlation among predictors

---

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 is a graphical device to determine whether a given data set is distributed according to a known distribution, for example, normal distribution. It plots the quantiles of the data against the quantiles of a known theoretical distribution. In linear regression, a Q-Q plot of the residuals is applied to test the assumption of normality. If the residuals are approximately along the reference line, it means that the residuals are normally distributed and thus the assumption is satisfied.
