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)

Investigating the influence of categorical variables on the dependent variable requires a careful examination of the relationships between these factors and the outcome of interest. This approach enables us to detect underlying trends and subtle differences across various categories. Furthermore, employing visualization techniques helps uncover potential patterns, identify outliers, and reveal nuances in the data that might otherwise remain hidden.

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)

Using the parameter drop_first=True is crucial for avoiding the dummy variable trap in regression models. This trap occurs when the dummy variables for a categorical feature are perfectly collinear, which complicates coefficient estimation and interpretation. By omitting one category for each categorical variable, we ensure only the necessary set of dummy variables remains, thereby reducing multicollinearity and improving the model's overall stability.

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)

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

After training the linear regression model, we verified its assumptions in two primary ways. First, we conducted outlier detection by examining standardized residuals, which helped us identify and investigate data points that might disproportionately influence the model. Second, we performed a multicollinearity check by calculating the Variance Inflation Factor (VIF) for each predictor. High VIF values indicate problematic correlations among the independent variables, which can undermine the accuracy and interpretability of the model's coefficients.

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)

'temp', 'atemp', 'hum'

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 that models the linear relationship between one or more independent variables and a dependent variable. In simple linear regression, the model is:

$y=\beta 0+\beta 1x+\varepsilon$,

where β is the intercept, $\beta 1$ is the slope, and ε is the error term. It finds the best-fit line by minimizing the sum of squared residuals.

Multiple linear regression extends this to several predictors.

Key assumptions include linearity, independence, homoscedasticity, and normality of residuals. It's widely used for forecasting, trend analysis, and understanding how factors affect an outcome. By quantifying relationships, it aids decision-making in fields like finance, economics, and social sciences.

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 collection of four datasets with the same descriptive statistical characteristics, such as means, variance, R-squared, correlations, and linear regression lines, but distinct graph scatter plot representations. Eleven (x, y) points make up each dataset. The datasets were produced to illustrate the value of data visualisation and the fallibility of summary statistics on their own.

The significance of exploratory data analysis and the limitations of relying solely on summary statistics are demonstrated using Anscombe's quartet. Additionally, it highlights how important it is to use data visualisation in order to identify trends, outliers, and other important features that may not be readily apparent from summary statistics alone.

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)

Being a descriptive statistic, the Pearson correlation coefficient encapsulates a dataset's features. In particular, it explains the direction and magnitude of the linear relationship between two numerical variables.

It is a number between -1 and 1, where 0 denotes no association, +1 denotes a positive correlation, and -1 denotes a negative correlation. The direction of change in one variable is reflected in the other.

It can be expressed using the following formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

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 key data preprocessing step applied to the independent variables to standardize their values within a specific range, thereby improving computational efficiency. Datasets often include features with widely differing magnitudes, units, and ranges. Without scaling, algorithms may overemphasize magnitude rather than units, leading to inaccurate modeling. By scaling, all variables are brought to the same level of magnitude. Notably, scaling only affects the coefficients and does not influence statistical measures like t-statistics, F-statistics, p-values, or R-squared.

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 Variance Inflation Factor (VIF) is infinite, it implies that Ri2=1R_i^2 = 1Ri2=1, indicating a perfect linear relationship between the iii-th variable and the other independent variables. This is known as perfect multicollinearity, which occurs when one or more predictors can be exactly derived from the others. Consequently, the correlation matrix becomes singular, its inverse cannot be computed, and the VIF becomes infinite.

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 Quantile-Quantile (Q-Q) plot is a graphical method for comparing the quantiles of a sample distribution to those of a theoretical distribution, such as the normal, uniform, or exponential. By plotting these quantiles against each other, it becomes easier to see how closely the dataset matches the specified distribution. In linear regression, Q-Q plots are particularly useful for assessing the normality of residuals, offering a clear visual check on assumptions. Researchers and statisticians often use them in combination with other diagnostic techniques to ensure the robustness and validity of regression models.
