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)

Analysis of categorical variable have been done using box plot and bar plot. Below are the points we can infer from visualization:

- Fall season has more number of booking and number of booking have increased for every season from year 2018 to 2019
- Most booking have been done in month of June, July, aug, sep and oct. Increasing trend is observed in number of booking from start of the year to mid of the year till June, followed by steady line from June to sep, then decrease is observed afterwards.
- Number of booking of each weekday is almost same, not much difference
- Clear weather has more number of booking
- Number of Booking seems to be greater on holidays, which is obvious
- Number of booking is almost similar whether it is working day or not
- Number of booking in 2019 has drastically increased from that of in 2018

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 drop_first = True when creating dummy variables is important because it prevents the "dummy variable trap" by removing one of the dummy variables for each categorical feature, which reduces multicollinearity among the newly created dummy variables and ensures a stable regression model. Essentially, it avoids creating redundant information between dummy variables by selecting a reference level to compare against.

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)

'temp' variable has 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)

validation of Linear Regression after building the model on the training set has been done on below assumptions:

- Normality of the error
- Multicollinearity check
- Linear relationship validation
- Homoscedasticity

- Independence of residuals

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)
Top 3 features:

- Temp
- Winter
- 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)

<Your answer for Question 6 goes here>
Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It provides valuable insights for prediction and data analysis. This article will explore its types, assumptions, implementation, advantages, and evaluation metrics.
Linear regression is also a type of supervised machine-learning algorithm that learns from the labelled datasets and maps the data points with most optimized linear functions which can be used for prediction on new datasets. It computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation with observed data. It predicts the continuous output variables based on the independent input variable.
For example if we want to predict house price we consider various factor such as house age, distance from the main road, location, area and number of room, linear regression uses all these parameter to predict house price as it consider a linear relation between all these features and price of house.
The interpretability of linear regression is one of its greatest strengths. The model's equation offers clear coefficients that illustrate the influence of each independent variable on the dependent variable, enhancing our understanding of the underlying relationships. Its simplicity is a significant advantage; linear regression is transparent, easy to implement, and serves as a foundational concept for more advanced algorithms.
Our primary objective while using linear regression is to locate the best-fit line, which implies that the error between the predicted and actual values should be kept to a minimum. There will be the least error in the best-fit line.
The best Fit Line equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).

$$Y=mx+c$$

Where,

m = slope

x = independent variable

c = intercept

The assumptions of multiple linear regression are crucial for ensuring that the model provides reliable and valid results:

Linearity: The relationship between the dependent variable and each independent variable is linear.

Independence: The observations are independent of each other (no autocorrelation).

Homoscedasticity: The variance of the errors (residuals) is constant across all levels of the independent variables.

No Multicollinearity: The independent variables should not be highly correlated with each other.

Normality of Errors: The residuals (errors) should be approximately normally distributed.

No Autocorrelation: Residuals should not exhibit patterns or correlations over time (for time-series data).

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)

<Your answer for Question 7 goes here>

Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation) but different distributions and relationships between variables. It was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before drawing conclusions.

The quartet consists of four datasets with two variables (X and Y). Despite having nearly identical summary statistics—such as the same mean of X (mean(X) = 9), mean of Y (mean(Y) = 7.5), variance, and correlation (around 0.82)—the data distributions differ significantly when visualized.

Dataset 1: Displays a strong linear relationship between X and Y.

Dataset 2: Shows a parabolic relationship, with a curve rather than a straight line.

Dataset 3: Contains a linear relationship with one outlier that heavily influences the regression line.

Dataset 4: Has a vertical line where all X values are the same, and Y values vary, indicating no meaningful relationship.

Anscombe's Quartet emphasizes that summary statistics alone can be misleading. Visualizing data through plots (such as scatter plots) is essential to identify underlying patterns, trends, or anomalies in the data. This insight is crucial for proper data analysis and decision-making.

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)

<Your answer for Question 8 goes here>

Pearson's correlation coefficient, often denoted as r, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1, where:

- r = 1: Perfect positive linear relationship (as one variable increases, the other also increases proportionally).
- r = -1: Perfect negative linear relationship (as one variable increases, the other decreases proportionally).
- r = 0: No linear relationship (the variables do not change in a consistent pattern relative to each other).

The formula for Pearson's r is:

$$r = (\Sigma (Xi - \bar{X})(Yi - \bar{Y})) / (\sqrt{(\Sigma (Xi - \bar{X})^2 \Sigma (Yi - \bar{Y})^2)})$$

Where:
- Xi and Yi are the individual sample points from two variables X and Y,
- $\bar{X}$ and $\bar{Y}$ are the mean values of X and Y, respectively.

Pearson's r is sensitive to outliers, as extreme values can significantly distort the correlation. It assumes that the relationship between the variables is linear, and both variables should be normally distributed for the coefficient to be meaningful.

Pearson's r is widely used in fields such as economics, psychology, and biology to understand the degree of association between two variables, helping researchers to make informed decisions based on the strength and direction of the relationship.

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)

<Your answer for Question 9 goes here>

Scaling is the process of transforming features in a dataset to a common scale without distorting differences in the ranges of values. It ensures that no variable dominates others due to differences in magnitude, units, or variance.

Why Scaling is Performed:

- Improved Model Performance: Scaling ensures that algorithms, especially those that rely on distance metrics (e.g., k-NN, SVM), work effectively.
- Convergence: In gradient-based optimization methods (e.g., linear regression, neural networks), scaling speeds up convergence by ensuring that all features contribute equally.

Difference between Normalized and Standardized Scaling:

- Normalization (Min-Max Scaling): Rescales the data to a specific range, typically [0, 1]. It is sensitive to outliers.

$$X_{normalised} = (X - X_{min}) / (X_{max} - X_{min})$$

- Standardization (Z-score Scaling): Centers the data around zero by subtracting the mean and dividing by the standard deviation, resulting in a distribution with mean 0 and variance 1. It is less sensitive to outliers.

$$X_{std} = (X - \mu)/ \sigma$$

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)

<Your answer for Question 10 goes here>

The Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to multicollinearity with other predictor variables. A VIF value of infinity occurs when there is perfect multicollinearity between the variables. This happens when one predictor variable is a perfect linear function of another, meaning that the predictor can be predicted with zero error using the other predictors in the model.

Reasons for Infinite VIF:

Perfect Multicollinearity: If two or more variables are perfectly correlated (e.g. X1=2X2), one variable can be predicted exactly by the others, leading to infinite variance for the coefficient of the affected variable.

Redundant Variables: Including variables that represent the same underlying information or repeated measurements can cause perfect correlation, leading to an infinite VIF.

Linear Dependence: If the matrix of predictors has linearly dependent columns, the regression model cannot estimate unique coefficients, resulting in an infinite VIF.

In practice, if you encounter infinite VIFs, it indicates that you need to remove or combine the collinear variables to resolve the issue of perfect multicollinearity.

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)

<Your answer for Question 11 goes here>

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a particular theoretical distribution, commonly the normal distribution. It compares the quantiles of the observed data against the quantiles of a theoretical distribution. If the points on the plot lie approximately along a straight line, it suggests that the data follows the theoretical distribution. If the points deviate significantly from the line, the data may not follow the assumed distribution.

Use and Importance of a Q-Q Plot in Linear Regression:

1. Normality Assumption: One of the assumptions of linear regression is that the residuals (errors) should be normally distributed. A Q-Q plot helps to visually assess whether the residuals approximate a normal distribution. If the residuals deviate from the straight line, it suggests a violation of normality.

2. Detecting Outliers: The Q-Q plot can help identify outliers or extreme values in the residuals. Points that are far from the straight line indicate outliers that may impact the regression model's results.

3. Model Validation: By checking the normality of residuals, the Q-Q plot helps to validate the assumptions of linear regression. Violations of the normality assumption can affect hypothesis tests and confidence intervals.

In summary, the Q-Q plot is a diagnostic tool to check the residuals' normality and ensure that the linear regression assumptions are not violated, leading to more reliable model results.
