Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

   Answer
   I have conducted a comprehensive analysis on categorical columns, employing both box plots and bar plots. The ensuing visualizations yield several discernible patterns:

   - The fall season emerges as a particularly attractive period for bookings, accompanied by a noteworthy surge in booking counts across each season from 2018 to 2019.
   - Predominantly, bookings concentrate in the months of May, June, July, August, September, and October. A discernible upward trajectory is observed from the initial months to the mid-year, followed by a gradual descent towards the year's conclusion.
   - Unambiguously, clear weather conditions are associated with heightened booking activity.
   - Thursdays, Fridays, Saturdays, and Sundays exhibit a markedly higher volume of bookings compared to the early days of the week.
   - During non-holiday periods, there is a discernible reduction in booking frequency, a plausible observation considering individuals' inclination to spend holidays at home with their families.
   - Booking frequencies manifest a near parity between working days and non-working days.
   - Notably, the year 2019 attracts a significantly greater number of bookings compared to the preceding year, indicative of positive business advancement.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**

   Answer

- The utilization of drop_first = True holds significance as it aids in mitigating the generation of an additional column during the creation of dummy variables, consequently diminishing correlations among these variables.
    - The syntax for this parameter is as follows:
- drop_first: bool, default False, indicating whether to obtain k-1 dummies out of k categorical levels by excluding the first level.
- To illustrate, consider a scenario where a categorical column possesses three distinct values, and the objective is to create dummy variables for that column. When one variable is not A or B, it implicitly implies C. Therefore, there is no necessity for the third variable to discern the presence of C. Utilizing drop_first = True facilitates this streamlined representation.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

   Answer

- The variable 'temp' exhibits the strongest correlation with the target variable.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 Marks)**

   Answer

   I have assessed the validity of the Linear Regression Model based on the following five assumptions:

   - **Normality of Error Terms**

   The assumption entails that error terms should conform to a normal distribution.

   - **Multicollinearity Check**

   It is essential to ensure that there is no significant multicollinearity among variables.

   - **Linear Relationship Validation**

   The presence of a visible linear relationship among variables is a crucial aspect of validation.

   - **Homoscedasticity**

   The absence of discernible patterns in residual values is essential to meet the assumption of homoscedasticity.

   - **Independence of Residuals**

   Ensuring no auto-correlation is crucial to fulfill the assumption of independence of residuals.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 Marks)**

Answer

The three primary features that play a significant role in elucidating the demand for shared bikes are as follows:

- Temperature (temp)
- Winter
- September (sep)

General Subjective Questions

1. **Explain the linear regression algorithm in detail? (4 marks)**

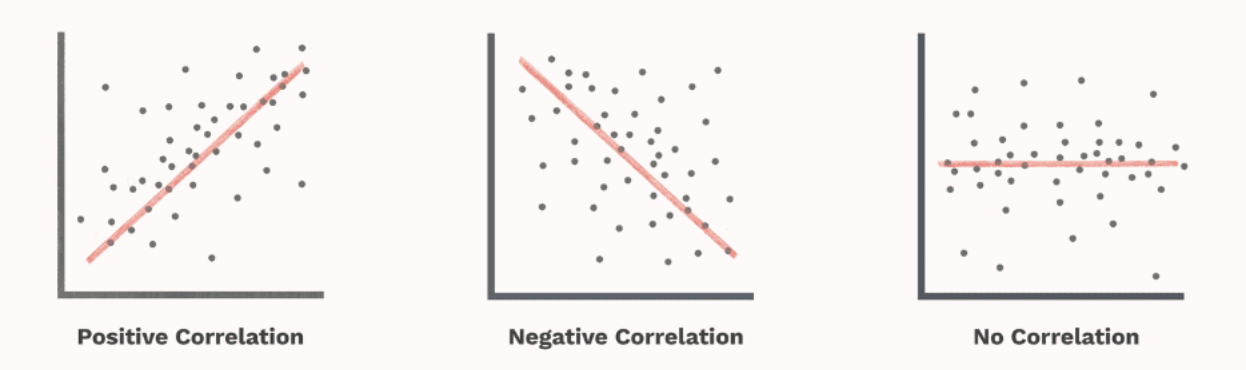
Answer

Linear regression is a statistical model that examines the linear association between a dependent variable and a given set of independent variables. In the context of this model, a linear relationship implies that as the values of one or more independent variables undergo changes, the corresponding values of the dependent variable will similarly increase or decrease.

Mathematically, this relationship is expressed by the equation Y = mX + c, where:

- Y is the dependent variable under prediction.
- X is the independent variable utilized for making predictions.
- m represents the slope of the regression line, depicting the impact of X on Y.
- c is a constant known as the Y-intercept, signifying the value of Y when X equals zero.

It is noteworthy that the nature of the linear relationship can be either positive or negative:



Positive Linear Relationship

Characterized by an increase in both the independent and dependent variables simultaneously. This concept is visually illustrated in the accompanying graph.

Negative Linear Relationship

Termed as such when an increase in the independent variable corresponds to a decrease in the dependent variable. This concept is elucidated through the graphical representation.

Assumptions

The Linear Regression model relies on several assumptions regarding the dataset:

Multi-collinearity

- The model assumes minimal or no multi-collinearity within the data. Multi-collinearity arises when there is dependence among the independent variables or features.

Auto-correlation

- Another assumption of the Linear Regression model is the presence of minimal or no auto-correlation in the data. Auto-correlation occurs when there is a dependency between residual errors.

Relationship Between Variables

- The model assumes a linear relationship between response and feature variables.

Normality of Error Terms

- The assumption is that error terms should follow a normal distribution.

Homoscedasticity

- There should be no discernible pattern in the residual values.

2. Explain the Anscombe's quartet in detail? (3 marks)

Answer

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics but differ significantly when graphed. It was created by the statistician Francis Anscombe in 1973 to emphasize the importance of graphing data before analyzing it and to demonstrate the limitations of relying solely on summary statistics.

Here are the key features of Anscombe's quartet:

Dataset Characteristics

Anscombe's quartet consists of four sets of (x, y) data points, labeled I, II, III, and IV.

Each dataset has 11 points.

Descriptive Statistics

Despite the differences in the actual data points, all four datasets share identical or very similar simple descriptive statistics:

- Mean of x: 9.0

- Variance of x: 11.0
- Mean of y: 7.50
- Variance of y: 4.12
- Correlation between x and y: 0.816 (approximately)

Graphical Representation

When plotted, each dataset forms a distinct pattern, illustrating that the relationship between x and y can vary significantly even when summary statistics are similar.

- Dataset I: Linear relationship.
- Dataset II: Non-linear, but still well described by a linear regression.
- Dataset III: Outlier strongly influences the correlation coefficient.
- Dataset IV: No correlation, but a strong linear relationship is observed when one outlier is removed.

Implications

The quartet highlights the importance of visualizing data and not relying solely on summary statistics.

It serves as a cautionary example against drawing conclusions based solely on numerical measures, such as means and correlations.

Educational Value

Anscombe's quartet is often used in statistics education to teach the principles of exploratory data analysis and the impact of outliers on statistical properties.

In summary, Anscombe's quartet demonstrates the necessity of complementing statistical analysis with graphical exploration to gain a more comprehensive understanding of the relationships within a dataset.

3. **What is Pearson's R? (3 marks)**

Answer

Pearson's correlation coefficient, often denoted as (r) or Pearson's (r), is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It was introduced by Karl Pearson and is widely used in statistics to assess the degree of linear association between two variables.

Here are key points about Pearson's correlation coefficient:

1. Range
   - Pearson's (r) ranges from -1 to 1.
   - (r = 1): Perfect positive linear correlation.
   - (r = -1): Perfect negative linear correlation.

- (r = 0): No linear correlation.

2. Formula
- The formula for Pearson's correlation coefficient between variables (X) and (Y) with sample size (n) is given by:
  [r = frac{sum{(X_i - bar{X})(Y_i - bar{Y})}}{sqrt{sum{(X_i - bar{X})^2} sum{(Y_i - bar{Y})^2}}}]
- Where (bar{X}) and (bar{Y}) are the means of (X) and (Y) respectively.

3. Interpretation
- The sign of (r) indicates the direction of the relationship:
  - Positive (r) implies a positive linear relationship.
  - Negative (r) implies a negative linear relationship.
- The magnitude of (r) indicates the strength of the relationship. The closer (r) is to 1 (or -1), the stronger the linear correlation.

4. Assumptions
- Pearson's correlation assumes a linear relationship between variables.
- It is sensitive to outliers.

5. Use Cases
- Pearson's (r) is commonly used in various fields, including psychology, biology, finance, and social sciences, to measure the strength and direction of linear relationships between variables.

In summary, Pearson's correlation coefficient is a statistical tool used to quantify the linear relationship between two continuous variables, providing insights into the strength and direction of their association.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

Answer

Definition
Feature Scaling stands as a pivotal preprocessing technique designed to standardize independent features within a dataset, harmonizing them within a defined range. This meticulous approach addresses the challenges arising from disparate magnitudes, values, or units present in the data.

Significance
The meticulous application of Feature Scaling is indispensable, preventing machine learning algorithms from inadvertently assigning disproportionate weight to larger values. This practice ensures a nuanced interpretation of smaller values, independent of their actual unit.

Exemplary Scenario

In scenarios where algorithms overlook feature scaling, a value of 3000 meters may inaccurately be deemed greater than 5 kilometers. Feature scaling rectifies such discrepancies by aligning all values to uniform magnitudes.

Methods of Feature Scaling

Normalized Scaling	Standardized Scaling
Scaling Technique: Utilizes minimum and maximum values of features for scaling.	*Scaling Technique:* Applies mean and standard deviation for scaling.
Use Case: Deployed when features manifest varying scales.	*Use Case:* Implemented to achieve a zero mean and unit standard deviation.
Scale Range: Values scaled between [0, 1] or [-1, 1].	*Scale Range:* Not confined to a specific range.
Outlier Sensitivity: Highly responsive to outliers.	*Outlier Sensitivity:* Demonstrates significantly lower sensitivity to outliers.
Scikit-Learn Tool: Employs the MinMaxScaler transformer for normalization.	*Scikit-Learn Tool:* Incorporates the StandardScaler transformer for standardization.

In summary, Feature Scaling emerges as a fundamental preprocessing step, ensuring uniformity in feature magnitudes. The dichotomy between Normalized Scaling and Standardized Scaling reveals nuanced considerations, each method serving distinctive purposes. Scikit-Learn offers dedicated transformers, namely MinMaxScaler and StandardScaler, for the precise implementation of these scaling techniques.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

Answer

The occurrence of infinite values in the Variance Inflation Factor (VIF) is typically associated with perfect multicollinearity among the independent variables in a regression model. Perfect multicollinearity happens when one or more independent variables in the model can be exactly predicted by a linear combination of the others.

The formula for VIF is:

[text{VIF}(^{beta}_j) = frac{1}{1 - R_j^2}]

where (R_j^2) is the (R^2) value obtained by regressing the (j)-th independent variable against all the other independent variables. If (R_j^2) is equal to 1, it implies that the (j)-th variable can be perfectly predicted by the other variables, leading to division by zero in the VIF formula.

In practical terms, this situation often arises when there is a linear relationship among some or all of the independent variables, making the design matrix singular and causing issues in estimating the

coefficients. When the VIF becomes infinite, it indicates a severe problem with multicollinearity in the regression model.

To address this issue, it is essential to inspect the data for redundant or highly correlated independent variables and consider appropriate measures, such as removing one of the correlated variables or employing regularization techniques, to mitigate multicollinearity and stabilize the regression model.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression? (3 marks)**

Answer

A Q-Q (Quantile-Quantile) plot is a graphical tool used in statistics to assess whether a dataset follows a particular theoretical distribution. It compares the quantiles of the observed data against the quantiles of a chosen probability distribution. In the context of linear regression, Q-Q plots are particularly valuable for checking the normality of residuals.

Key Components of a Q-Q Plot:
- X-axis: Theoretical quantiles from a standard normal distribution.
- Y-axis: Observed quantiles from the dataset.

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

1. Normality Assessment:
   - Q-Q plots are employed to visually inspect whether the residuals of a linear regression model conform to a normal distribution. Normality of residuals is a crucial assumption for valid inference and hypothesis testing in linear regression.

2. Pattern Recognition:
   - In an ideal scenario, if the residuals follow a normal distribution, the points on the Q-Q plot should fall along a straight line. Deviations from this straight line indicate departures from normality.

3. Identification of Outliers:
   - Outliers can significantly impact the normality of residuals. A Q-Q plot allows for the identification of extreme values or outliers in the tails of the distribution, influencing the overall normality assessment.

4. Model Assumption Validation:
   - Checking the normality of residuals is essential to validate one of the key assumptions of linear regression. Violations of this assumption might affect the accuracy and reliability of statistical inferences drawn from the model.

5. Statistical Inference
   - When conducting hypothesis tests or constructing confidence intervals, assuming normality allows for the application of parametric statistical tests. Q-Q plots help ensure that these assumptions are met, enhancing the validity of statistical inferences.

Interpretation of Q-Q Plot in Linear Regression:

- If the points in the Q-Q plot closely follow a straight line, it suggests that the residuals are approximately normally distributed.
- Deviations or curvature in the plot may indicate non-normality, prompting further investigation and potentially the need for remedial actions.

In summary, Q-Q plots play a vital role in assessing the normality of residuals in linear regression. They provide a visual diagnostic tool for validating assumptions, identifying outliers, and ensuring the reliability of statistical inferences derived from the regression model.