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 an analysis on categorical columns using box plots and bar plots. The visualizations reveal several key observations:

- The fall season appears to have garnered more bookings, and across each season, the booking count experienced a significant increase from 2018 to 2019.
- The majority of bookings occurred during the months of May, June, July, August, September, and October. There is a noticeable upward trend from the beginning to the middle of the year, followed by a decline towards the year-end.
- Bookings were notably higher during clear weather conditions, which aligns with expectations.
- Thursday, Friday, Saturday, and Sunday saw a higher number of bookings compared to the earlier days of the week.
- The year 2019 saw a substantial increase in bookings compared to the previous year, indicating positive progress in terms of business.

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

Answer:

The parameter **`drop_first=True`** is utilized during the creation of dummy variables, specifically for categorical variables with multiple levels. Its purpose is to omit the first column generated during the dummy variable creation process.

For instance, consider a categorical column like "gender" with four variables: "Male," "Female," "Other," and "Unknown." If a person can only be classified as one of these three categories and "Unknown" encompasses any other cases, there's no need for a separate column for "Unknown." Therefore, by setting **`drop_first=True`**, the first column representing "Male" is dropped.

This parameter is situation-dependent. In cases where having one less dummy variable is appropriate, it helps reduce the number of columns. However, this isn't universally applicable. In scenarios where all possible categories need representation, such as a column for "Fav_genre" with values like "Rock," "Hip hop," "Pop," "Metal," and "Country," setting `drop_first=True` would not be suitable. This is because individuals may have more than one favorite genre, and dropping any of the columns would not accurately represent this diversity. Consequently, the default parameter is `drop_first=False` in such cases.

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

Answer:

`atemp` and `temp` have high 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 assumptions of the Linear Regression Model through the following five criteria:

1. Normality of Error Terms:
    - The error terms should exhibit a normal distribution.
2. Multicollinearity Check:
    - There should be no significant multicollinearity among the variables.
3. Linear Relationship Validation:
    - There should be a discernible linear relationship among the variables.
4. Homoscedasticity:
    - Residual values should not display any discernible pattern, indicating homoscedasticity.
5. Independence of Residuals:
    - There should be no evidence of autocorrelation in the residuals.

These assessments are crucial for ensuring the validity and reliability of the Linear Regression Model, as they collectively address the normality and distribution of errors, the relationship between variables, the absence of multicollinearity, the homogeneity of variance, and the 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:

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

- season_4
- season_3
- season_2

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 set of independent variables. The fundamental idea is that changes in the values of the independent variables (whether an increase or decrease) correspondingly lead to changes in the dependent variable. This relationship can be expressed mathematically through the equation Y = mX + c, where Y is the dependent variable, X is the independent variable, m is the slope representing the effect of X on Y, and c is the Y-intercept (constant).

The nature of the linear relationship can be classified as positive or negative:

- **Positive Linear Relationship:**
  - Occurs when both the independent and dependent variables increase together.
- **Negative Linear Relationship:**
  - Happens when an increase in the independent variable corresponds to a decrease in the dependent variable.

Linear regression comes in two types:

1. **Simple Linear Regression:**
   - Involves one independent variable.
2. **Multiple Linear Regression:**
   - Encompasses multiple independent variables.

There are several assumptions made by the Linear Regression model regarding the dataset:

1. **Multi-collinearity:**
   - Assumes minimal or no multi-collinearity, where independent variables do not exhibit strong dependencies among themselves.
2. **Auto-correlation:**
   - Assumes minimal or no auto-correlation, indicating that there is no significant dependency between residual errors.
3. **Relationship Between Variables:**
   - Assumes a linear relationship between response and feature variables.
4. **Normality of Error Terms:**
   - Assumes that error terms are normally distributed.
5. **Homoscedasticity:**
   - Requires that there be no discernible pattern in residual values, ensuring homoscedasticity.

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 (mean, variance, correlation, and linear regression) yet appear very different when graphed. This quartet was created by the statistician Francis Anscombe in 1973 to emphasize the importance of graphical data exploration in addition to numerical summaries. The datasets highlight the limitations of relying solely on summary statistics and the importance of visualizing data to understand its underlying patterns.

Here are the details of Anscombe's quartet:

1. Dataset I:
    - x-values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7
    - y-values: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82
    - Mean of x: 9.0, Mean of y: 7.5
    - Variance of x: 11.0, Variance of y: 4.12
    - Correlation between x and y: 0.816
    - Linear regression: y = 3.00 + 0.50 * x
2. Dataset II:
    - x-values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7
    - y-values: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26
    - Mean of x: 9.0, Mean of y: 7.5
    - Variance of x: 11.0, Variance of y: 4.12
    - Correlation between x and y: 0.816
    - Linear regression: y = 3.00 + 0.50 * x
3. Dataset III:
    - x-values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7
    - y-values: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42
    - Mean of x: 9.0, Mean of y: 7.5
    - Variance of x: 11.0, Variance of y: 4.12
    - Correlation between x and y: 0.816
    - Linear regression: y = 3.00 + 0.50 * x
4. Dataset IV:
    - x-values: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8
    - y-values: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91
    - Mean of x: 9.0, Mean of y: 7.5
    - Variance of x: 11.0, Variance of y: 4.12
    - Correlation between x and y: 0.816
    - Linear regression: y = 3.00 + 0.50 * x

Key observations from Anscombe's quartet:

- While the summary statistics are nearly identical, the datasets have different distributions and patterns when graphed.
- Emphasizes the importance of visualizing data to uncover patterns that may not be apparent in summary statistics alone.
- Highlights the limitations of relying solely on numerical measures without exploring the actual data distribution.
- Demonstrates that a single summary statistic may not capture the complexity of a dataset.

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

Answer:

Pearson's correlation coefficient, denoted as r, serves as a quantitative measure of the strength of the linear relationship between two variables. This coefficient takes on values within the range of +1 to -1, providing insights into the nature and direction of their association.

- Positive Correlation ($r > 0$):
  - Indicates that as one variable increases, the other variable tends to increase as well.
  - The closer the value of r is to +1, the stronger the positive correlation.
- Negative Correlation ($r < 0$):
  - Suggests that as one variable increases, the other variable tends to decrease.
  - The closer the value of r is to -1, the stronger the negative correlation.
- Zero Correlation ($r = 0$):
  - Signifies no linear association between the two variables.
  - Indicates that changes in one variable do not predict changes in the other variable.

Pearson's correlation coefficient provides valuable insights into the direction and strength of the linear association between variables, aiding in the interpretation of their relationship within a dataset.

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

Answer:

Feature scaling is a preprocessing technique aimed at standardizing independent features in a dataset to a consistent range. This practice is crucial when dealing with variables that exhibit varying magnitudes, values, or units. If feature scaling is omitted, machine learning algorithms may disproportionately emphasize features with larger values, potentially leading to inaccurate model outcomes. In essence, feature scaling prevents the algorithm from assigning greater importance to variables with higher magnitudes.

Consider the following example to illustrate the necessity of feature scaling: Suppose an algorithm, without feature scaling, treats a value of 3000 meters as greater than 5 kilometers. In reality, this is an incorrect comparison due to the difference in units. Feature scaling addresses this issue by bringing all feature values to a common magnitude, ensuring that the algorithm interprets and weighs them appropriately.

In summary, feature scaling is employed during data preprocessing to:

1. Standardize features to a consistent range.
2. Mitigate the impact of varying magnitudes, values, or units.
3. Prevent the algorithm from assigning disproportionate importance to features with larger values.
4. Enhance the overall performance and accuracy of machine learning models by promoting fair consideration of all features.

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

Answer:

The Variance Inflation Factor (VIF) is a metric used to assess multicollinearity in a regression model. If there is a perfect correlation between independent variables, the VIF becomes infinite. A high VIF value indicates a strong correlation among variables, and a VIF of 4, for example, implies that the variance of the model coefficient is inflated by a factor of 4 due to multicollinearity.
In cases where the VIF is infinite, it signifies a perfect correlation between two independent variables. Perfect correlation leads to an R-squared (R2) value of 1, resulting in a calculation of 1 / (1 - R2) which equals infinity. To address this situation and resolve perfect multicollinearity, it is necessary to drop one of the variables from the dataset. Removing one of the correlated variables helps eliminate the issue of perfect multicollinearity, allowing for a more stable and interpretable regression model. This step is crucial for maintaining the integrity of the regression analysis and ensuring meaningful results.

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

Answer:

The quantile-quantile (q-q) plot is a graphical method used to assess whether two datasets originate from populations with a common distribution. This plot involves comparing the quantiles of one dataset against the quantiles of another, with a 45-degree reference line indicating where points should fall if the two datasets share the same distribution.

Key aspects of the q-q plot:

1. Quantiles Definition:
    - Quantiles represent the fraction or percentage of points below a given value. For instance, the 30% quantile is the point below which 30% of the data falls, and 70% falls above that value.
2. Reference Line:
    - A 45-degree reference line is plotted on the graph. If both datasets come from populations with the same distribution, the points on the q-q plot should approximately align with this reference line.
3. Interpretation:
    - Departure from the 45-degree reference line indicates a deviation from a common distribution. The greater the departure, the stronger the evidence that the two datasets have different distributions.

Importance of the q-q plot:

1. Distribution Assumption Justification:
    - It helps determine if the assumption of a common distribution for two datasets is justified. If the points align well with the reference line, it supports the assumption.
2. Pooling Estimators:
    - When datasets share a common distribution, location and scale estimators can be pooled to obtain estimates of the common location and scale. This is valuable for certain statistical analyses.
3. Understanding Differences:
    - If two samples differ, the q-q plot provides insight into the nature of these differences. It visually reveals how the datasets deviate from each other, offering more nuanced information compared to some analytical tests.

In summary, the q-q plot is a powerful tool for assessing the distributional similarity between two datasets. It aids in making informed decisions about the appropriateness of statistical assumptions and provides valuable insights into the nature of differences between datasets.