

Title: Day 8 - Bivariate Analysis and Correlation for Fake Bill Detection

Introduction:

This documentation delves into the tasks conducted on Day 8 of the project, which center around bivariate analysis and correlation in the context of the 'FAKEBILL_new.csv' dataset. The primary objective is to gain a deeper understanding of the dataset and to examine the relationships between independent variables and the dependent variable 'is_genuine,' which signifies whether a bill is authentic or counterfeit.

Task 1: Correlation Analysis

- The first task involves exploring the correlations between different attributes within the dataset.
- A correlation matrix is calculated to quantify the degree of linear association between these attributes.
- The correlation matrix is then displayed, providing a visual representation of the relationships among attributes. This allows for the identification of potential strong or weak correlations.

Task 2: Bivariate Analysis

- Bivariate analysis focuses on investigating the connections between each independent variable and the dependent variable 'is_genuine.'
- Four subtasks are executed to examine these relationships: boxplots, t-tests, crosstabs, and chi-square tests.

Task 2a: Boxplots

- Side-by-side boxplots are crafted to compare the distributions of each independent variable. These boxplots are grouped by the 'is_genuine' category, which distinguishes genuine and counterfeit bills.
- This graphical representation offers an immediate insight into the variations in attribute distributions between the two categories.

Task 2b: t-tests

- Independent two-sample t-tests are carried out for each independent variable concerning the 'is_genuine' variable.
- The t-statistic and p-value are computed for each test, shedding light on whether statistically significant differences exist in the means of these attributes between genuine and counterfeit bills.

Task 2c: Crosstabs

- Crosstabs are used to construct contingency tables that illustrate the relationships between the 'is_genuine' variable and each independent variable.
- These tables present the frequency of observations within each combination of variables, offering valuable insights into the interplay between attributes and their impact on bill authenticity.

Task 2d: Chi-Square Tests

- Chi-square tests are conducted to evaluate the association between the 'is_genuine' variable and each independent variable.
- The tests provide a chi-square statistic, p-value, degrees of freedom, and an expected frequencies table for each analysis. These statistics aid in comprehending the significance of the relationships.

Conclusion:

The tasks performed on Day 8 significantly contribute to a more profound understanding of the 'FAKEBILL_new.csv' dataset. Correlation analysis unveils the relationships between attributes, while bivariate analysis scrutinizes how independent variables influence the authenticity of bills. By employing a combination of visualization, statistical tests, and contingency tables, this analysis serves as a cornerstone for more advanced investigations and the development of predictive models tailored for the detection of counterfeit bills.