Report - Analysis of UCL Dataset

I - Univariable Analysis - Teffahi Salah Eddine

The Project file is a Python notebook that performs Univariable Analysis on the 'goals.csv' dataset. The analysis focuses on the 'position' column.

Data Understanding

The 'position' column, which represents the position on the field where goals were scored from, is analyzed. The data is first inspected to understand its distribution and characteristics.

Descriptive Statistics

Descriptive statistics are calculated for the 'position' column. This includes measures of central tendency such as mean, median, and mode, measures of dispersion such as standard deviation, variance, range, interquartile range, and measures of shape such as skewness and kurtosis. These statistics provide a comprehensive summary of the 'position' column.

Frequency Distribution

The frequency distribution of the 'position' column is calculated and visualized using a histogram. This provides an understanding of how often each position is represented in the dataset.

Boxplot

A boxplot is created for the 'position' column to visualize its quartiles (including the median) and any outliers. This provides a summary of the distribution of the 'position' column.

Gini Index and Lorenz Curve

The Gini index is calculated and the Lorenz curve is plotted for the 'position' column. These are measures of inequality in the data, with the Gini index being a number between 0 and 1 (where 0 corresponds to perfect equality and 1 corresponds to perfect inequality) and the Lorenz curve being a graphical representation of the distribution of the data.

In conclusion, the univariable analysis performed in this notebook provides a comprehensive analysis of the 'position' column in the 'goals.csv' dataset, providing insights into its distribution, central tendency, dispersion, and inequality.

II - Bivariable Analysis - Boukader Imad Eddine

This part of the analysis performs Bivariable Analysis on the 'goals.csv' dataset. The analysis focuses on the relationship between the 'position' and 'goals_scored' columns.

Data Understanding

The 'position' and 'goals_scored' columns, which represent the position on the field where goals were scored from and the number of goals scored respectively, are analyzed. The data is first inspected to understand its distribution and characteristics.

Descriptive Statistics

Descriptive statistics are calculated for the 'position' and 'goals_scored' columns. This includes measures of central tendency such as mean, median, and mode, measures of dispersion such as standard deviation, variance, range, interquartile range, and measures of shape such as skewness and kurtosis. These statistics provide a comprehensive summary of the 'position' and 'goals_scored' columns.

Scatter Plot

A scatter plot is created to visualize the relationship between the 'position' and 'goals_scored' columns. This provides an understanding of how the number of goals scored varies with the position on the field.

Correlation

The correlation between the 'position' and 'goals_scored' columns is calculated. This provides a measure of how much the two variables move in relation to each other.

Covariance

The covariance between the 'position' and 'goals_scored' columns is calculated. This provides a measure of how much the two variables vary together.

Regression Analysis

A regression analysis is performed to model the relationship between the 'position' and 'goals_scored' columns. This includes calculating the regression coefficients, the coefficient of determination (R-squared), and performing a hypothesis test for the regression coefficients.

Residual Analysis

A residual analysis is performed to validate the assumptions of the regression model. This includes checking the normality of the residuals, the homoscedasticity of the residuals, and the independence of the residuals.

In conclusion, the bivariable analysis performed in this notebook provides a comprehensive analysis of the relationship between the 'position' and 'goals_scored' columns in the 'goals.csv' dataset, providing insights into their correlation, covariance, and the nature of their relationship.

III - MultiVariable Analysis Khaldi Abderrhmane

This part performs Principal Component Analysis (PCA) on the 'goals.csv' dataset. The analysis focuses on the following columns: 'right_foot', 'left_foot', 'inside_area', 'outside_areas', and 'penalties'.

Data Preprocessing

The data is first centered and scaled to ensure that all variables are on the same scale. This is crucial for PCA as it is a variance maximizing exercise. It involves subtracting the mean and dividing by the standard deviation for each value of each variable.

Covariance and Correlation Matrix

The covariance and correlation matrices are calculated to understand the linear relationships between different variables. The covariance matrix provides a measure of how much each of the variables is dependent on each other. The correlation matrix is another way to measure how different variables are dependent on each other.

Eigenvalue Decomposition

Eigenvalues and eigenvectors are calculated from the covariance matrix. These are used to identify the principal components of the data. The eigenvalues represent the distribution of the source data's variance amongst each of the principal components, whereas the eigenvectors represent the principal components themselves.

Selection of Principal Components

The principal components are selected based on the explained variance ratio, which is the proportion of the dataset's variance that lies along the axis of each principal component. The 'elbow method' is used to determine the optimal number of principal components, which is where the explained variance ratio starts to flatten out on a scree plot. In this case, the optimal number of principal components was found to be 3.

PCA Transformation

The data is then transformed into the first three principal components. This reduces the dimensionality of the data while preserving as much of the data's original variance as possible.

Visualization

The transformed data is visualized using a scatter plot, which shows the individuals' cloud. This provides a visual representation of the dataset along the principal components. The correlation circle is also plotted to show the correlation of the original variables with the principal components.

Quality and Contribution of Individuals

The quality of representation of each individual and their contribution to the principal components is calculated. This provides an understanding of how well each individual is represented by the principal components and their contribution to the principal components.

Correlation between Original Variables and Axes

The correlation between the original variables and the axes is calculated. This provides an understanding of how each original variable contributes to the principal components.

In conclusion, the PCA performed in this notebook provides a comprehensive analysis of the 'goals.csv' dataset, reducing its dimensionality and providing insights into the underlying structure of the data.