Unit II

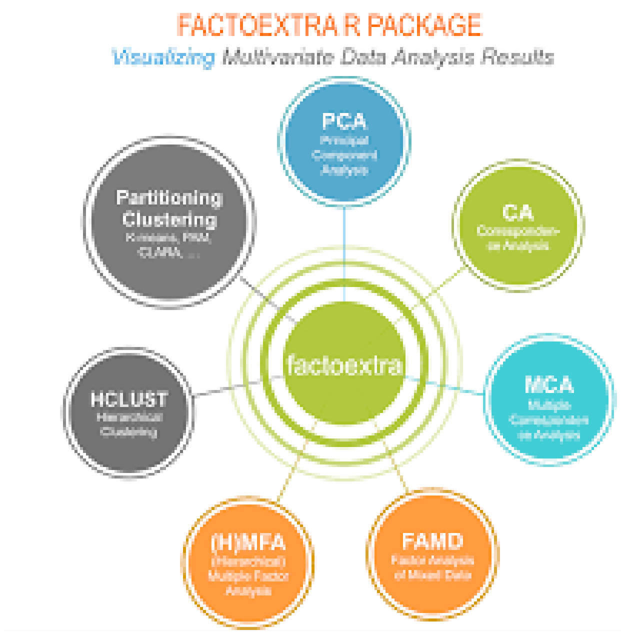
2. Multivariate frequencies

Multivariate frequencies in data analytics refer to the analysis of two or more variables simultaneously to understand their joint distribution and relationships. It involves examining the frequencies or counts of different combinations of values for the variables being studied. This analysis is particularly valuable when dealing with categorical or discrete variables.

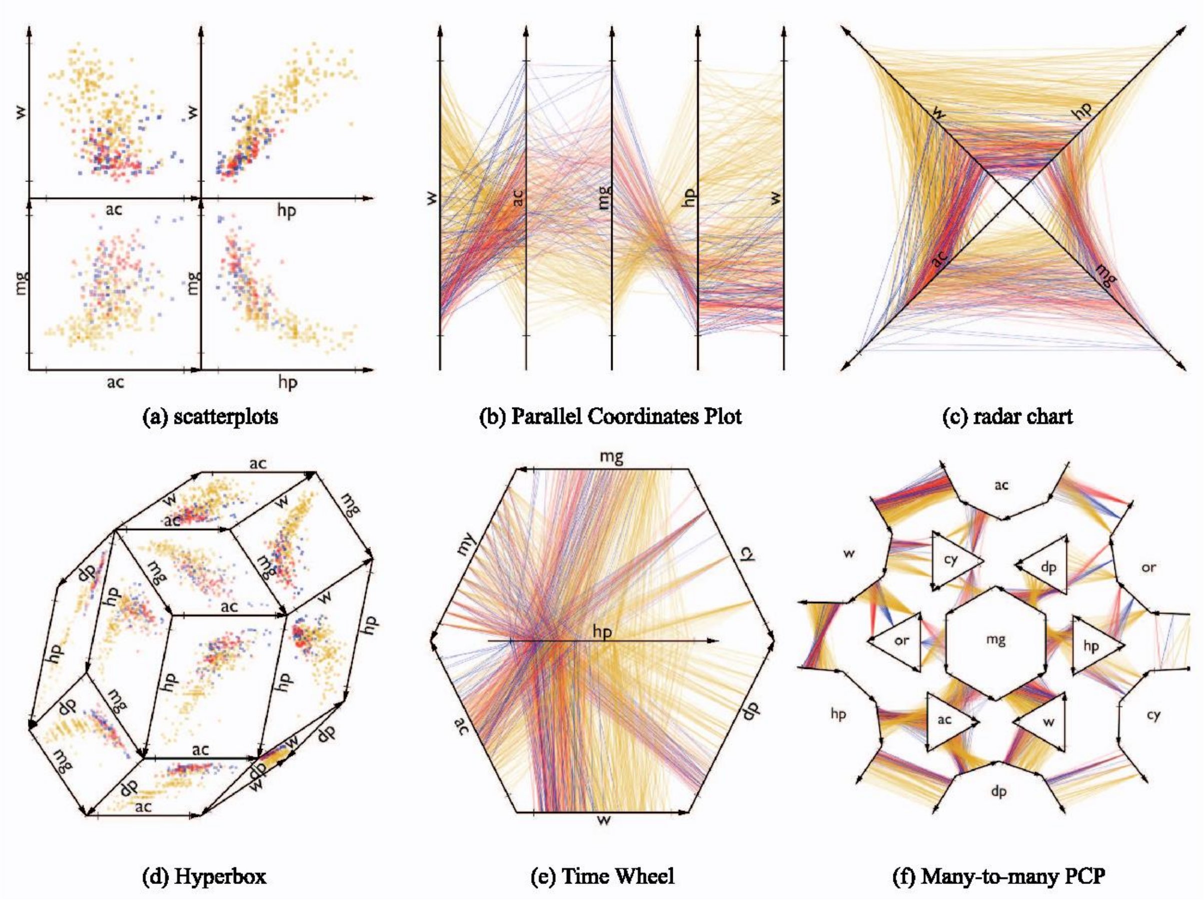
- Cross-Tabulation: Cross-tabulation, also known as a contingency table, is a common way to display multivariate frequencies. It shows the frequencies of each combination of values for two or more categorical variables. This allows analysts to observe patterns and associations between the variables.
- Understanding Associations: Multivariate frequencies help identify whether there are significant associations or dependencies between two or more categorical variables. For example, in a survey, we can cross-tabulate gender with a product preference to determine if there are gender-based preferences.
- Visualizing Multivariate Frequencies: The data from cross-tabulation can be visualized using stacked bar charts, heatmaps, or mosaic plots. Stacked bar charts show the frequencies of each category for one variable, segmented by the categories of the other variable. Heatmaps use color-coded cells to represent the frequencies for each combination of values, providing a visual summary of the relationships.
- Hypothesis Testing: Multivariate frequencies can be used for hypothesis testing to determine if there is a statistically significant relationship between the variables. Techniques like the chi-square test or Fisher's exact test can be applied to assess the independence or association between categorical variables.
- Real-World Applications: Multivariate frequencies are widely used in market research, social sciences, and customer segmentation. They help understand customer behavior, identify market trends, and reveal patterns in survey responses.
- Limitations: Multivariate frequencies are most suitable for analyzing categorical data. When dealing with continuous variables, other multivariate analysis techniques like scatter plots, correlation matrices, and regression models are more appropriate.

3. Multivariate data visualization

Multivariate data visualization in data analytics focuses on representing and exploring relationships between multiple variables in a single visual representation. It allows data analysts and data scientists to gain insights into complex datasets by visualizing how multiple variables interact with each other. Multivariate data visualization techniques help in identifying patterns, correlations, clusters, and outliers, and facilitate better understanding and communication of the data's structure.


FACTOEXTRA R PACKAGE
Visualizing Multivariate Data Analysis Results
PCA
Principal Component Analysis
CA
Correspondence Analysis
MCA
Multiple Correspondence Analysis
FAMD
Factor Analysis of Mixed Data
(H)MFA
(Hierarchical) Multiple Factor Analysis
HCLUST
Hierarchical Clustering
Partitioning Clustering
K-means, PAM, CLARA, ...
factoextra

- Scatter Plot Matrix: A scatter plot matrix displays scatter plots for all possible pairs of continuous variables in the dataset. Each plot represents the relationship between two variables, allowing analysts to quickly identify patterns and correlations between multiple pairs of variables.
- Parallel Coordinates Plot: This type of plot is useful for visualizing high-dimensional data. It uses parallel axes to represent each variable, and data points are connected by lines. Parallel coordinates plots help identify trends and patterns in the data, especially in situations where there are many variables involved.
- Heatmap: A heatmap is a graphical representation of a matrix where each cell's color represents the magnitude of the value. It is commonly used to display correlation matrices, showing the strength and direction of correlations between multiple variables.
- 3D Scatter Plot: When dealing with three continuous variables, a 3D scatter plot can be used to visualize the relationships among them in a three-dimensional space. However, interpreting 3D plots can be challenging, especially with more than three variables.
- Radar Chart: A radar chart (also known as a spider chart or star plot) displays multivariate data on a two-dimensional chart with multiple axes emanating from the center. Each axis represents a different variable, and data points are plotted relative to the axes. Radar charts are useful for comparing multiple variables across different categories.
- Glyph Plot: Glyph plots use symbols (glyphs) to represent multiple variables simultaneously. Different properties of the symbols, such as size, shape, and color, can represent various dimensions of the data.
- Chernoff Faces: Chernoff faces are a creative and unusual way to visualize multivariate data. Each data point is represented by a human face, and different facial features (e.g., eyes, mouth, nose) correspond to different variables' values.



(a) scatterplots

(b) Parallel Coordinates Plot

(c) radar chart

(d) Hyperbox

(e) Time Wheel

(f) Many-to-many PCP

4. Multivariate statistics

Multivariate statistics in data analytics involves analyzing and modeling relationships among multiple variables simultaneously. Unlike univariate statistics that focus on a single variable, multivariate statistics consider the joint behavior of two or more variables to gain deeper insights and make more informed decisions. Multivariate techniques are valuable when dealing with complex datasets with interdependent variables.

- Multivariate Regression Analysis: Extends the concept of simple linear regression to include multiple independent variables to predict a dependent variable. It helps identify the relationships between the dependent variable and multiple predictors, allowing for more sophisticated modeling.

- Principal Component Analysis (PCA): A dimensionality reduction technique used to transform high-dimensional data into a lower-dimensional space while retaining as much of the original variance as possible. PCA helps identify the most significant patterns or principal components in the data, making it easier to visualize and analyze complex datasets.

- Factor Analysis: A method used to explore the underlying structure of observed variables by identifying latent factors that explain the common variance among them. Factor analysis helps reduce the number of variables and provides insights into the underlying constructs in the data.
- Canonical Correlation Analysis (CCA): Measures the relationship between two sets of variables to identify patterns of association between them. CCA is useful when dealing with multiple sets of variables that are potentially related to each other.
- Multivariate Analysis of Variance (MANOVA): Extends the concept of analysis of variance (ANOVA) to multiple dependent variables to determine whether there are significant differences among groups based on different independent variables.
- Cluster Analysis: Clustering methods group similar data points together based on their characteristics. It helps identify natural clusters or segments within the data, aiding in data segmentation and pattern recognition.
- Discriminant Analysis: Discriminant analysis is used to classify data into predefined groups based on their features. It is commonly employed in classification problems when the classes are known.
- Multivariate Time Series Analysis: Deals with analyzing and forecasting multiple time series variables simultaneously. Techniques like Vector Autoregression (VAR) and Vector Error Correction Model (VECM) are used in this context.