Section 1: Exploratory Data Analysis

The structure of this chapter hopefully ensures a coherent flow to exploratory data analysis (EDA), starting with introduction to structured data and moving through univariate, bivariate, and multivariate analyses. It also addresses data quality issues such as missing data and pre-processing.

1.1 Elements of Structured Data Structured data refers to data organized into a specific format, making it easier to process, analyse, and manipulate. This often includes tables with rows and columns, where each row represents an observation, and each column represents a variable or feature.

1.2 Rectangular Data Rectangular data, also known as tabular data, is a common data format in which data is organized into rows and columns. Each row corresponds to an observation, while each column represents a variable or attribute.

1.3 Data Frames and Indexes A data frame is a two-dimensional data structure that stores data in a tabular format. It has labelled axes (rows and columns) and can contain heterogeneous data types. An index is a unique identifier for each row in the data frame, often a numeric or string value.

1.4 Non-Rectangular Data Structures Non-rectangular data structures include time-series, graphs, trees, and spatial data structures, which have their own unique organization and methods for analysis.

- **Time-series data:** Sequential data points collected over time, often with equally spaced intervals. Time-series data can be analysed using techniques such as moving averages, exponential smoothing, or time-series decomposition.
- **Graphs/networks:** Data structures composed of nodes (vertices) and edges (connections) that represent relationships between entities. Graphs can be analysed using graph theory algorithms, such as shortest path, community detection, or centrality measures.
- **Trees:** Hierarchical data structures with a root node and child nodes, representing relationships between entities in a hierarchical manner. Tree structures can be used for decision-making, classification, or organization of information.
- **Spatial data:** Data with a geographical or spatial component, often represented as points, lines, or polygons on a map. Spatial data can be analysed using techniques such as spatial autocorrelation, clustering, or spatial regression.

1.5 Estimates of Location Estimates of location provide a central value that represents a dataset. Common estimates of location include the mean, median, and robust estimates.

1.5.1 Mean The mean, or average, is the sum of all values in a dataset divided by the number of values. Formula: $\mu = \Sigma x/N$

1.5.2 Median The median is the middle value in a dataset when sorted.

1.5.3 Robust Estimates

Robust estimates, such as the trimmed mean and the weighted mean, are less sensitive to outliers and skewness in the data.

- **Trimmed Mean:** A variation of the mean where a certain percentage of the highest and lowest values are removed before calculating the average. This reduces the influence of outliers on the mean.

- **Weighted Mean:** The weighted mean takes into account the relative importance of each data point by assigning weights. It is calculated as the sum of the product of each data point and its corresponding weight divided by the sum of the weights. Formula: Weighted Mean = Σ(wx) / Σw

1.6 Example: Location Estimates of Population and Murder Rates

1.7 Estimates of Variability Estimates of variability measure the dispersion or spread of data points in a dataset.

1.7.1 Standard Deviation and Related Estimates The standard deviation is the square root of the variance, which is the average squared difference from the mean. Formula: $\sigma = \sqrt{(\Sigma(x-\mu)^2/N)}$

1.7.2 Estimates Based on Percentiles Percentiles divide the data into equal parts, such as quartiles (25th, 50th, and 75th percentiles) or deciles (10th, 20th, ..., 90th percentiles). The interquartile range (IQR) is the range between the first quartile (25%) and third quartile (75%).

- **Range:** The difference between the maximum and minimum values in a dataset.
- **Interquartile Range (IQR):** The range between the first quartile (Q1) and third quartile (Q3) in a dataset, representing the central 50% of the data.
- **Percentile Rank:** The percentage of data points below a given value in a dataset. For example, if a data point has a percentile rank of 70, it means that 70% of the data points are below that value.

1.8 Exploring the Data Distribution Various techniques can be used to visualize and explore the distribution of a dataset.

1.8.1 Percentiles and Boxplots - Boxplots display the five-number summary (minimum, first quartile, median, third quartile, maximum) and help visualize the spread and skewness of a dataset.

1.8.2 Frequency Tables and Histograms Frequency tables list the frequency of each value or interval in a dataset, while histograms graphically represent the frequency distribution using bars.

1.8.3 Density Plots and Estimates Density plots are smooth curves that represent the distribution of a dataset, often using kernel density estimation (KDE).

- Kernel Density Estimation (KDE): A non-parametric method used to estimate the probability density function of a random variable. It involves placing a kernel (a smooth, symmetric function) at each data point and summing the kernels to create a smooth curve that represents the distribution of the data. KDE can provide more detail about the shape of a distribution compared to histograms, and it allows for the estimation of continuous probability density functions from discrete data.

1.9 Exploring Binary and Categorical Data Binary and categorical data can be analysed using techniques such as mode, expected value, and probability.

1.9.1 Mode The mode is the most frequently occurring value in a dataset.

1.9.2 Expected Value The expected value is the weighted average of all possible values, where each value is multiplied by its corresponding probability.

1.9.3 Probability - Probability measures the likelihood of a particular outcome or event occurring within a given sample space.

- **Joint Probability:** The probability of two or more events occurring simultaneously, often represented as $P(A \cap B)$ for events A and B.
- **Conditional Probability:** The probability of an event occurring given that another event has already occurred, often represented as $P(A \mid B)$ for events A and B, meaning the probability of event A happening given that event B has occurred.

- **Bayes' Theorem:** A method for updating probabilities based on new evidence, often used in statistical inference and machine learning. It is represented as P(A | B) = P(B | A) * P(A) / P(B), where P(A | B) is the posterior probability, P(B | A) is the likelihood, P(A) is the prior probability, and P(B) is the evidence.

1.10 Correlation - Correlation measures the strength and direction of a linear relationship between two variables.

1.10.1 Scatterplots - Scatterplots are graphical representations of the relationship between two quantitative variables, with each data point represented as a dot in the Cartesian plane.

- **Scatterplot best-fit line:** A line that represents the underlying relationship between two variables in a scatterplot. This can be a simple linear regression line or a more complex curve fitted using techniques such as polynomial regression or locally weighted regression (LOESS).
- **Scatterplot matrix (SPLOM):** A square grid of scatterplots showing all possible pairwise combinations of variables in a dataset. Each scatterplot displays the relationship between two different variables, allowing for the simultaneous exploration of multiple relationships.

1.11 Exploring Two or More Variables Several techniques can be used to explore relationships between multiple variables.

1.11.1 Hexagonal Binning and Contours (plotting numeric versus numeric data) Hexagonal binning and contour plots can be used to visualize the relationship between two numeric variables, especially when there is a large amount of data or overlapping points.

- **Hexagonal Binning:** A technique for aggregating data points in two-dimensional space by dividing the space into hexagonal bins. The hexagons can be coloured or shaded to represent the density or count of data points within each bin. This technique is particularly useful for visualizing large datasets or scatterplots with overlapping points, as it reduces overplotting and enables the identification of patterns and trends.
- **Contour Plots:** A graphical representation of the relationship between two numeric variables, where contour lines are used to connect points with equal values. In the context of Exploratory Data Analysis, contour plots can be used to visualize the density of data points or to estimate the probability density function of a bivariate distribution. Contour plots can help identify.

1.11.2 Two Categorical Variables A cross-tabulation or heatmap can be used to visualize the relationship between two categorical variables.

- **Cross-tabulation:** A technique for summarizing the relationship between two categorical variables by creating a contingency table. The table shows the frequency of observations for each combination of categories, allowing for the identification of patterns and relationships between the variables.
- **Chi-squared test:** A statistical test used to determine if there is a significant association between two categorical variables in a cross-tabulation. The test compares the observed frequencies in the contingency table to the expected frequencies under the assumption of independence between the variables.

1.11.3 Categorical and Numeric Data Boxplots, violin plots, and bar charts can be used to visualize the relationship between categorical and numeric data.

1.11.4 Visualizing Multiple Variables Parallel coordinates plots, scatterplot matrices, and multivariate visualization techniques can help explore relationships between multiple variables simultaneously.

- **Parallel Coordinates Plot:** A visualization technique for multivariate data, where each variable is represented on a separate vertical axis, and individual data points are connected by lines across the axes. This allows for the identification of patterns, trends, and potential clusters in the data.
- **Scatterplot Matrix:** A matrix of scatterplots displaying all pairs of variables, providing a comprehensive view of the relationships between multiple variables. This can help identify correlations, trends, and outliers in the data.

1.12 Summary Exploratory Data Analysis (EDA) is an essential first step in data analysis, providing a comprehensive understanding of the data's structure, characteristics, and relationships. By using various techniques such as location and variability estimates, data visualization, and correlation analysis, we can identify patterns, anomalies, and insights that guide further analysis and model building.

1.13 Handling Missing Data Missing data is a common issue in real-world datasets and can affect the quality of insights derived from Exploratory Data Analysis. Several methods can be employed to handle missing data:

- **Deletion:** Remove observations with missing values, either by listwise deletion (removing the entire row) or pairwise deletion (removing only the specific data point). This method can be appropriate when the amount of missing data is minimal and missing values are missing at random (MAR).
- **Imputation:** Fill in the missing values using various techniques such as mean imputation, median imputation, mode imputation, or more advanced methods like k-Nearest Neighbours (k-NN) imputation or regression imputation. Imputation helps retain the structure of the data but may introduce bias or reduce variance.
- **Model-based approaches:** Use statistical models like Expectation-Maximization (EM) or Multiple Imputation by Chained Equations (MICE) to estimate missing values based on the observed data. These methods can provide more accurate estimates but may be computationally intensive.

1.14 Data Pre-processing Before performing Exploratory Data Analysis, it is essential to pre-process the data to ensure its quality and suitability for analysis. Some common data pre-processing steps include:

- **Data cleaning: Identify** and correct errors, inconsistencies, and inaccuracies in the data, such as duplicate records, typos, or incorrect data types.
- **Data transformation:** Apply functions or operations to the data to improve its interpretability or suitability for analysis. Common transformations include normalization, standardization, log transformations, and power transformations.
- **Feature engineering:** Create new variables or features from existing ones to improve the quality of insights or model performance. Techniques for feature engineering include binning, aggregation, interaction terms, and domain-specific transformations.
- **Encoding categorical variables:** Convert categorical variables into a suitable format for analysis or modelling, such as one-hot encoding, label encoding, or target encoding.