MED2056: Medical Data Analysis with Python

Assist. Prof. Huseyin TUNC

November 18, 2023

Table of Contents

What is Data Science?

- **Definition:** Data Science is an interdisciplinary field that involves extracting insights and knowledge from structured and unstructured data using scientific methods, processes, algorithms, and systems.
- **Components:**
  - **Data Collection:** Gathering raw data from various sources, including databases, APIs, sensors, and more.
  - **Data Cleaning and Preprocessing:** Ensuring data quality by handling missing values, outliers, and transforming data into a usable format.
  - **Exploratory Data Analysis (EDA):** Analyzing and visualizing data to discover patterns, trends, and relationships.
  - **Statistical Analysis and Machine Learning:** Applying statistical methods and machine learning algorithms for predictions, classifications, and clustering.
  - **Data Visualization:** Communicating findings through visual representations to aid in decision-making.

What is Data Science? (Cont'd)

- **Applications (Healthcare Focus):**
  - **Predictive Analytics in Patient Outcomes:**
    - Using historical patient data to predict potential health issues, hospital readmissions, and overall patient outcomes.
  - **Drug Discovery and Development:**
    - Analyzing molecular and genetic data to identify potential drug candidates and optimize drug development processes.
  - **Personalized Medicine:**
    - Tailoring medical treatments to individual characteristics, such as genetics, to enhance treatment efficacy and reduce side effects.
  - **Disease Surveillance and Epidemiology:**
    - Monitoring and predicting disease outbreaks, understanding disease patterns, and improving public health strategies.
  - **Healthcare Fraud Detection:**
    - Applying machine learning algorithms to detect fraudulent activities in healthcare insurance claims and billing.

What is Data Science? (Cont'd)

- **Applications (Healthcare Focus - Additional Examples):**
  - **Clinical Decision Support Systems:**
    - Developing intelligent systems that assist healthcare professionals in making clinical decisions by analyzing patient data, medical literature, and best practices.
  - **Image Analysis for Diagnostics:**
    - Utilizing machine learning algorithms to analyze medical images (e.g., X-rays, MRIs) for early detection and diagnosis of diseases such as cancer or neurological disorders.
  - **Remote Patient Monitoring:**
    - Implementing data-driven solutions for continuous monitoring of patients' health remotely, enabling timely interventions and reducing the need for frequent hospital visits.
  - **Genomic Data Analytics:**
    - Analyzing genomic data to identify genetic markers associated with diseases, enabling personalized treatment plans and advancements in precision medicine.
  - **Health Behavior Analysis:**
    - Applying data science techniques to understand and predict health-related behaviors, facilitating interventions for lifestyle modifications and preventive care.

Data Science vs. Traditional Statistics

- **Data Science vs. Traditional Statistics:**
  - **Scope and Objectives:**
    - **Traditional Statistics:** Primarily focused on analyzing and summarizing data, making inferences about populations based on sample data.
    - **Data Science:** Encompasses a broader range of activities, including data cleaning, feature engineering, machine learning, and the deployment of models.
  - **Data Handling:**
    - **Traditional Statistics:** Often assumes clean and well-structured datasets, with a focus on hypothesis testing and parameter estimation.
    - **Data Science:** Deals with messy, unstructured data, requiring substantial effort in data cleaning, preprocessing, and exploration.
  - **Exploration vs. Hypothesis Testing:**
    - **Traditional Statistics:** Emphasizes hypothesis testing to draw conclusions about population parameters.
    - **Data Science:** Prioritizes exploration and discovery, seeking patterns and insights in the data without a predefined hypothesis.

Data Science Lifecycle

- **Data Acquisition:**
  - Gathering data from various sources, such as databases, APIs, external datasets, or data generated within an organization.
- **Data Cleaning and Preprocessing:**
  - Cleaning and transforming raw data to ensure its quality, handle missing values, and prepare it for analysis.
- **Exploratory Data Analysis (EDA):**
  - Analyzing and visualizing data to understand patterns, trends, and relationships, guiding further analysis.
- **Modeling:**
  - Building and training machine learning models based on the insights gained from EDA, selecting appropriate algorithms for the task.
- **Evaluation:**
  - Assessing the performance of the models using metrics relevant to the specific problem, fine-tuning as necessary.
- **Deployment:**
  - Implementing the model in a real-world setting, integrating it into existing systems, and ensuring it works as intended.

Data Exploration in Data Science

- **Importance of Data Exploration:**
  - Understanding the characteristics and structure of the data is crucial before applying advanced analytics or machine learning models.
- **Techniques in Data Exploration:**
  - **Descriptive Statistics:** Calculating measures such as mean, median, and standard deviation to summarize the main features of a dataset.
  - **Data Visualization:** Creating plots, charts, and graphs to visually represent patterns and relationships in the data.
  - **Correlation Analysis:** Examining the strength and direction of relationships between variables.
  - **Outlier Detection:** Identifying and handling anomalies in the data that may affect analysis.
  - **Feature Engineering:** Creating new features or transforming existing ones to improve the performance of machine learning models.
- **Tools for Data Exploration:**
  - **Python Libraries:** Pandas for data manipulation, Matplotlib and Seaborn for visualization, and NumPy for numerical operations.
  - **R Programming:** Utilizing R and its packages for statistical analysis and visualization.

Statistical Analysis in Data Science

- **Objective:** Utilize statistical methods to derive insights, patterns, and relationships from data.
- **Key Concepts:**
  - **Descriptive Statistics:** Summarize and describe the main features of a dataset.
  - **Visualization Techniques:** Represent data visually to aid in understanding and interpretation.
  - **Correlation:** Measure the strength and direction of relationships between variables.
  - **Regression:** Predict the value of one variable based on the values of others.
  - **Classification:** Categorize data into classes or groups based on certain features.

Descriptive Statistics

- **Mean ($\bar{x}$):** The average of a set of values.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- **Median:** The middle value in a sorted list of numbers.
- **Standard Deviation (σ):** A measure of the amount of variation or dispersion in a set of values.

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

- **First Quartile (Q1):** The median of the lower half of the dataset.
- **Third Quartile (Q3):** The median of the upper half of the dataset.
- **Interquartile Range (IQR):** The range between the first quartile (Q1) and the third quartile (Q3) in a dataset.

$$IQR = Q3 - Q1$$

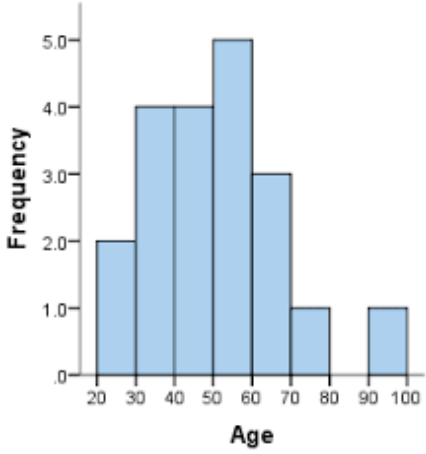
Descriptive Statistics (Cont'd)

Why they are useful (Cont'd):

- **Mean:** Sensitive to outliers, providing a representative value.
- **Median:** Robust to outliers, offering a central value.
- **Standard Deviation:** Quantifies the amount of variation in data points.
- **Interquartile Range (IQR):** Captures the spread of the middle 50% of the data, useful for detecting skewness.

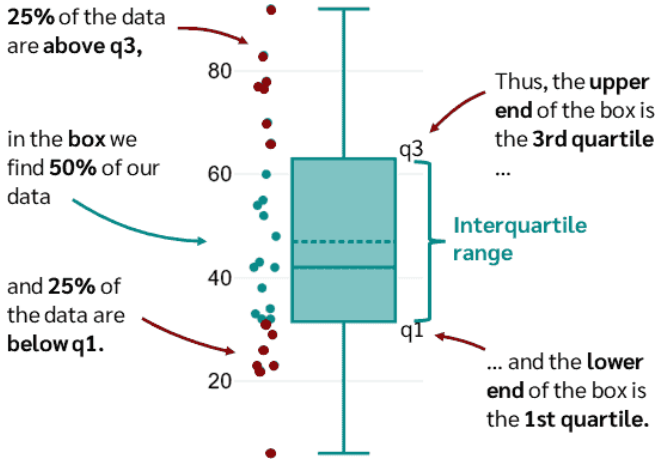
Visualization Techniques

- **Histograms:** Display the distribution of a continuous variable.



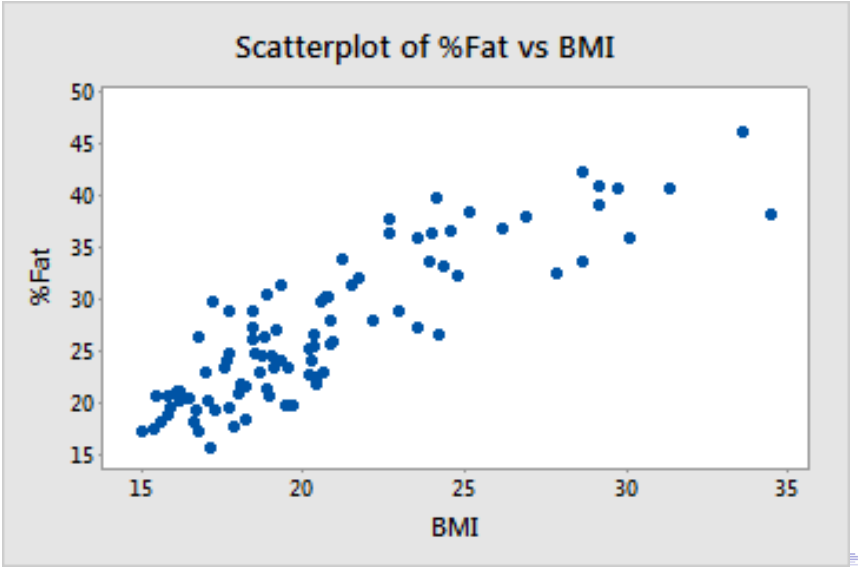
Visualization Techniques (Cont'd)

- **Box Plots:** Illustrate the distribution of a dataset and identify outliers.



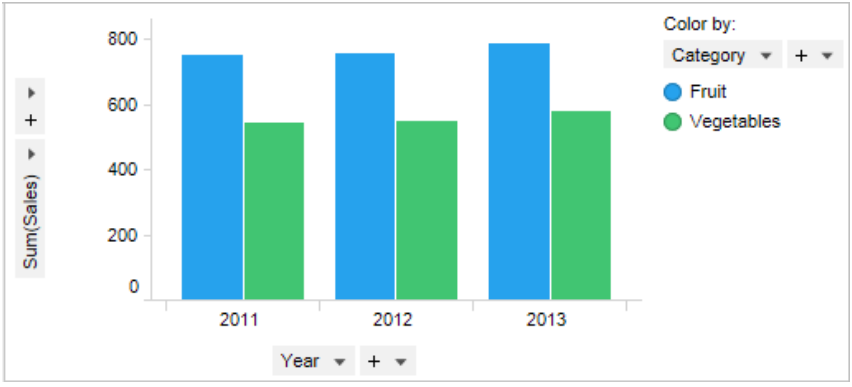
Visualization Techniques (Cont'd)

- **Scatter Plots:** Show the relationship between two continuous variables.



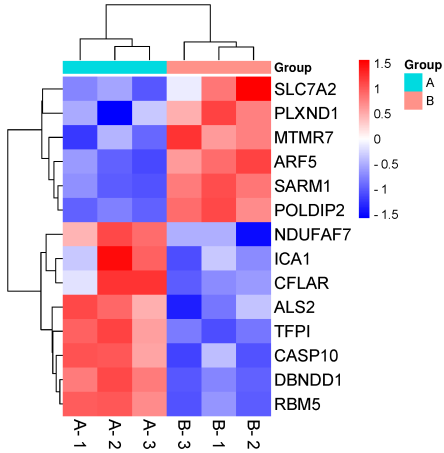
Visualization Techniques (Cont'd)

- **Bar Charts:** Present the distribution of a categorical variable.



Visualization Techniques (Cont'd)

- **Heatmaps:** Visualize the correlation matrix between variables or may give some other relational information.

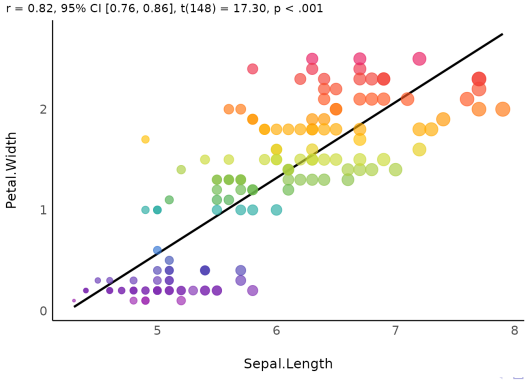


Correlation

- **Correlation Coefficient (r):** Quantifies the strength and direction of a linear relationship between two variables.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- **Example:** Correlation between sepal length and petal width.



Regression

- **Simple Linear Regression:** Models the relationship between a dependent variable y and an independent variable x.

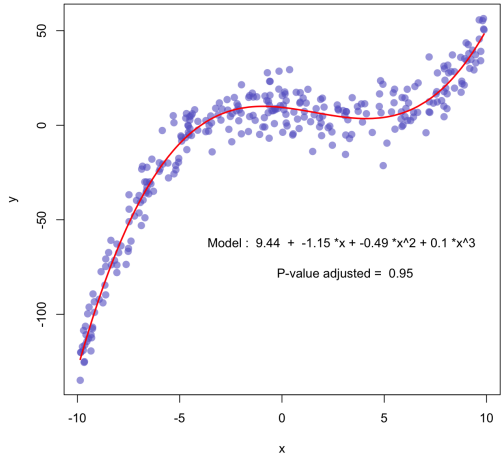
$$y = \beta_0 + \beta_1 x + \varepsilon$$

- **Multiple Linear Regression:** Extends simple linear regression to multiple independent variables.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

Regression

Example: Predicting house prices based on features like square footage and number of bedrooms.



Introduction to Machine Learning

- **What is Machine Learning (ML)?**
  - ML is a field of artificial intelligence that focuses on developing algorithms and models that enable computers to learn patterns and make decisions without explicit programming.
- **Types of Machine Learning:**
  - **Supervised Learning:** The model is trained on a labeled dataset, where the input data and corresponding output are provided.
  - **Unsupervised Learning:** The model is given unlabeled data and must find patterns or relationships without explicit guidance.
  - **Reinforcement Learning:** The model learns through interaction with an environment, receiving feedback in the form of rewards or penalties.
- **Common Machine Learning Tasks:**
  - **Classification:** Assigning labels to input data.
  - **Regression:** Predicting a continuous output.
  - **Clustering:** Grouping similar data points.
  - **Dimensionality Reduction:** Reducing the number of features in a dataset.

Popular Data Science Tools

- **Python:** Widely used programming language for data science and machine learning. Popular libraries include NumPy, Pandas, Matplotlib, and Scikit-Learn.
- **R:** Statistical programming language with extensive packages for data manipulation, visualization, and statistical modeling.
- **Jupyter Notebooks:** Interactive notebooks that allow combining code, visualizations, and text, facilitating data exploration and analysis.
- **SQL:** Structured Query Language for managing and querying relational databases, essential for extracting and manipulating structured data.
- **Tableau:** Data visualization tool that simplifies complex data into interactive and shareable visualizations.
- **Excel:** Widely used spreadsheet software with built-in data analysis tools, suitable for basic data manipulation and visualization.

Popular Data Science Tools

- **Git:** Version control system for tracking changes in code, collaborating with others, and maintaining project history.
- **Hadoop:** Distributed storage and processing framework, particularly useful for handling large-scale datasets.
- **TensorFlow and PyTorch:** Libraries for building and training deep learning models.
- **Apache Spark:** Fast and general-purpose cluster-computing framework for big data processing.

Ethics and Privacy in Data Science

- **Data Collection:**
  - Ensure transparency about what data is being collected and for what purpose.
  - Obtain informed consent from individuals before collecting their data.
  - Minimize the collection of sensitive or personally identifiable information.
- **Data Storage and Security:**
  - Implement robust security measures to protect stored data from unauthorized access.
  - Adhere to industry standards and regulations regarding data encryption and storage.
- **Data Use and Sharing:**
  - Clearly define the intended use of collected data and avoid using it for purposes beyond the scope of consent.
  - Anonymize or aggregate data whenever possible to protect individual privacy.
  - Exercise caution when sharing data, ensuring compliance with privacy laws and regulations.

Ethics and Privacy in Data Science

- **Bias and Fairness:**
  - Be aware of and mitigate biases in data that can lead to unfair or discriminatory outcomes.
  - Regularly audit and assess models for bias, and adjust algorithms to promote fairness.
- **Accountability and Transparency:**
  - Establish clear accountability for decisions made by algorithms and models.
  - Communicate transparently with stakeholders about the methods and processes used in data science projects.

Thank you!