knowledgehut
upGrad

Web Scraping and Exploratory Data Analysis

Pre-requisites

Hope you have gone through the self-learning content for this session on the PRISM portal.

By the End of this Session, You Will:

- Explore and understand data using Python and pandas' library.
- Extract data from websites using Python and the Beautiful Soup library.
- Visualize data in a clear and concise way.
- Apply the concepts of EDA and web scraping to real-world problems.

Key Takeaways From This Session

Automation and Efficiency: Web scraping allows for automated data extraction from websites, saving time and effort compared to manual data collection.

Ethical Considerations: It is important to use web scraping ethically and responsibly.

Data Understanding: EDA helps in understanding the dataset by examining its structure, content, and relationships between variables.

Descriptive Statistics: It includes measures of central tendency and variability and provides summary information to describe the dataset's key characteristics.

Pop Quiz

Q. Which of the following is not an ethical consideration when web scraping?

a. Respecting website terms of service

b. Avoiding unauthorized access to data

c. Overloading servers with excessive requests

d. Using a public API instead of web scraping

Pop Quiz

Q. Which of the following is not an ethical consideration when web scraping?

a. Respecting website terms of service

b. Avoiding unauthorized access to data

c. Overloading servers with excessive requests

d. Using a public API instead of web scraping

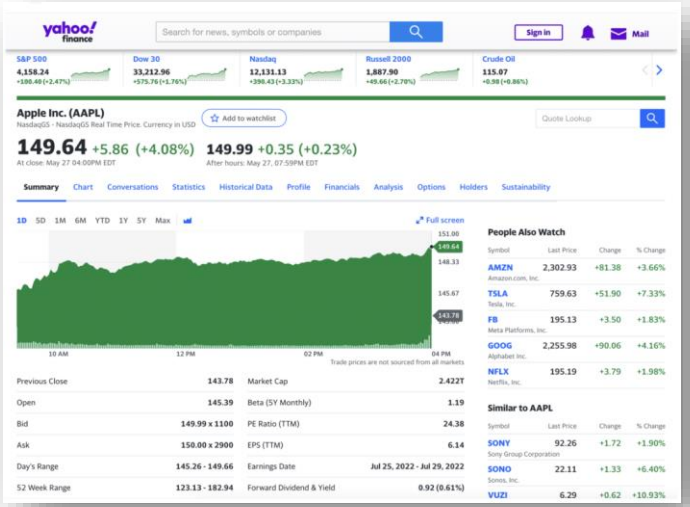
Hint:
Using a public API is not an ethical consideration but rather a technical consideration.

Introduction to Web Scraping

What is Web Scraping?

- Web scraping is like going to a library to find a book. The website is like the library, the HTML content is like the books, and the data elements that you need are like the specific books that you are looking for.
- Parsing libraries are like librarians. They help you to navigate the HTML DOM and extract the desired data elements.
- Web scraping can be used to automate the process of data collection and extract the specific data elements that you need. This saves you time and effort, and it allows you to focus on analyzing the data.

Example:
You are a data analyst who is interested in tracking the price of a particular stock. You could use web scraping to extract the stock price from a website like **Yahoo Finance**.



Concepts and Techniques of Web Scraping

Targeted Data Extraction:

Web scraping is the process of extracting specific data elements from web pages. This can be done using a variety of techniques, including HTML parsing, XPath, and CSS selectors.

HTML Parsing:

HTML parsing is the process of understanding the structure of HTML documents. This is essential for web scraping, as it allows you to navigate and extract data from HTML tags.

XPath and CSS Selectors:

XPath and CSS selectors are powerful techniques to locate and extract specific HTML elements based on their attributes or relative position within the document.

Handling Dynamic Content:

Websites with dynamic content may require additional techniques to scrape data that is loaded or modified dynamically. This can be done using APIs, JavaScript rendering, or browser automation tools.

Ethical Considerations:

Web scraping should be done responsibly and ethically. This means respecting website terms of service, robots.txt files, and ensuring that your scraping activities do not overload servers or infringe on user privacy.

Demo - Introduction to Web Scraping

Pop Quiz

Q. Which library is used for web scraping?

a. pandas

b. NumPy

c. Seaborn

d. Beautiful Soup

Pop Quiz

Q. Which library is used for web ccraping?

a. pandas

b. NumPy

c. Seaborn

d. Beautiful Soup

Introduction to Exploratory Data Analysis

What is an EDA?

- Imagine you are a data scientist tasked with analyzing a new dataset. The dataset is large and complex, and you need to gain a quick understanding of its characteristics before you can start analyzing it.
- EDA is like taking a first look at the dataset. You are looking for patterns, trends, and outliers.
- You are also trying to understand the relationships between the different variables in the dataset.
- In this scenario, the dataset is like a new city. You have never been to this city before, and you need to get your bearings.
- You are looking for the major landmarks, the transportation system, and neighborhoods.
- EDA is a crucial step in the data analysis process. It helps you to understand the data and to identify the areas that need further investigation. It also helps you to generate hypotheses and to formulate research questions.

Steps Involved in EDA

1. **Initial Data Investigation:** This is like exploring the city for the first time. You are walking around, looking at the buildings, and trying to get a sense of the layout.

2. **Data Summary and Descriptive Statistics:** This is like learning about the city's population, economy, and history. You are also trying to understand the different neighborhoods and their connection.

3. **Data Visualization:** This is like creating a map of the city. You use different colors and symbols to represent the different neighborhoods and landmarks.

4. **Pattern Recognition and Relationships:** This is like looking for patterns in the city's traffic patterns, crime rates, or economic growth. You are also trying to understand how the different neighborhoods are connected.

5. **Hypothesis Generation:** This is like generating hypotheses about why the city is how it is. You ask questions like "Why is the crime rate so high in this neighborhood?" or "Why is the economy so strong in this neighborhood?"

Popular Python Libraries for EDA

pandas:

A powerful library for data manipulation and analysis. It provides data structures and functions for handling structured data.

NumPy:

A fundamental library for numerical computing in Python. It provides high-performance multidimensional array objects and functions for mathematical operations.

Matplotlib:

A popular data visualization library in Python. It offers a wide range of plotting functions to create various types of graphs and charts.

Seaborn:

A statistical data visualization library built on top of Matplotlib. It provides a high-level interface for creating aesthetically pleasing and informative statistical graphics.

Plotly:

An interactive visualization library that allows the creation of interactive, web-based visualizations. It offers a range of chart types, including scatter plots, bar charts, heat maps, and 3D plots.


Q&A

Poll Time

Q. Which of the following Python libraries is best suited for creating interactive, web-based visualizations?

a. pandas

b. NumPy

c. Matplotlib

d. Plotly

Poll Time

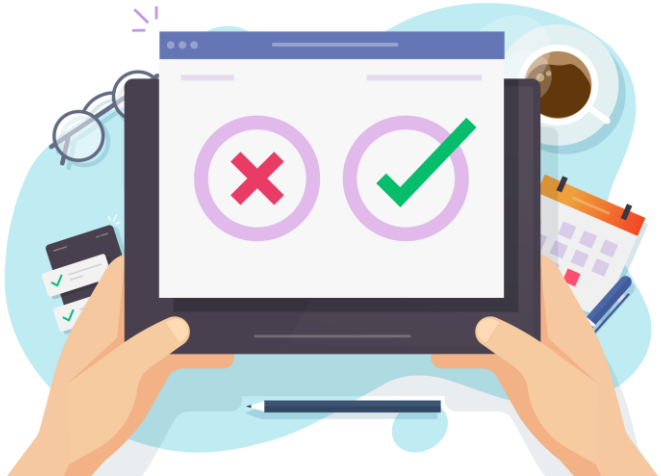
Q. Which of the following Python libraries is best suited for creating interactive, web-based visualizations?

a. pandas

b. NumPy

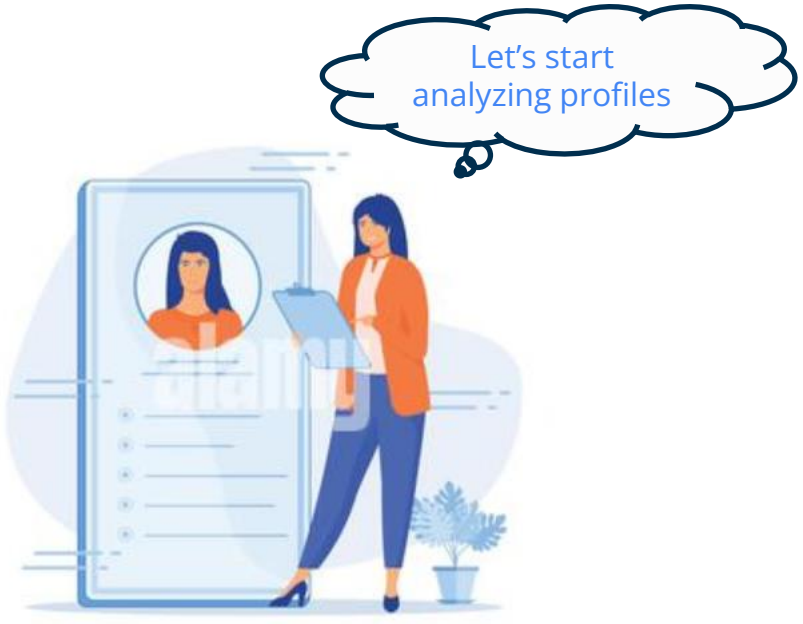
c. Matplotlib

d. Plotly



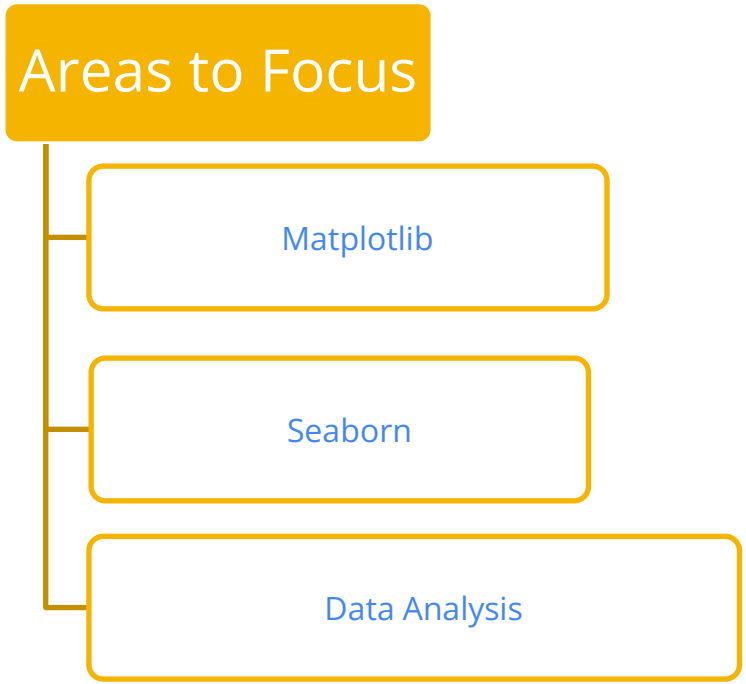
Case Study – Web Scraping and EDA

Problem Statement



Alice, a finance professional, wants to understand job profiles and mobile phone accessibility in Eastern Africa. She has data from 2016-2018 for Kenya, Uganda, Tanzania, and Rwanda. She will conduct EDA to answer basic data questions and gain insights.

Areas to Focus



Poll Time

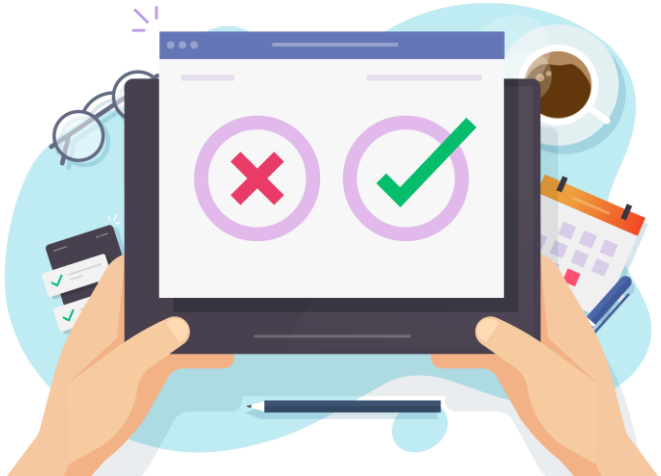
Q. Which of the following is not a basic data question that Alice wants to answer with EDA?

a. What are the most common job profiles in Eastern Africa?

b. What is the percentage of people who have a cell phone?

c. What is the percentage of people who have a bank account?

d. What is the percentage of people who have both a cell phone and a bank account?



Poll Time

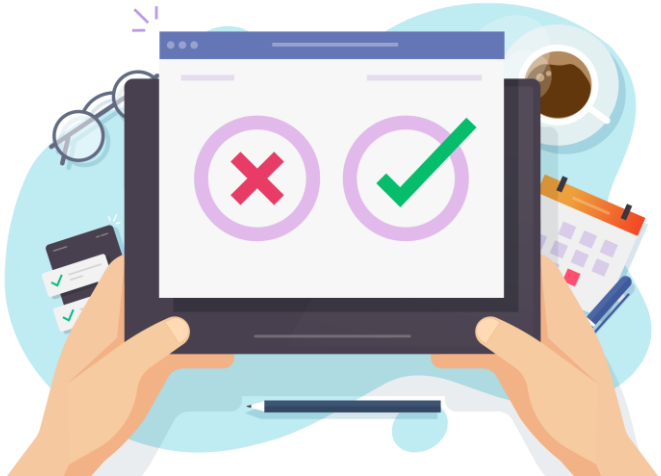
Q. Which of the following is not a basic data question that Alice wants to answer with EDA?

a. What are the most common job profiles in Eastern Africa?

b. What is the percentage of people who have a cell phone?

c. What is the percentage of people who have a bank account?

d. What is the percentage of people who have both a cell phone and a bank account?




Q&A

Summary

- Exploratory Data Analysis (EDA) is a process of analyzing and visualizing data to gain insights and understand its characteristics, patterns, and relationships.
- Web Scraping is the automated extraction of data from websites. It involves parsing HTML content, accessing web pages, and extracting targeted data elements using libraries like Beautiful Soup or Scrapy.
- EDA can help analysts identify trends, outliers, correlations, and potential data issues. It is a crucial step to guide further analysis, model-building, and decision-making processes based on a comprehensive understanding of the dataset.
- Web Scraping can empower analysts to gather data from multiple sources efficiently, automate data collection processes, and integrate web data into their analyses. However, it is important to consider ethical considerations, respect website policies, and ensure responsible scraping practices.

Activity 1

Prerequisites:

- Basic knowledge of Python
- Familiarity with the Beautiful Soup library
- A web browser

Problem Statement:

You are a data analyst who is interested in scraping data from the website **www.worldometers.info/population/:https://www.worldometers.info/population/**

You want to extract the following data for each country:

1. Country name
2. Population
3. Population growth rate

Activity 1

Steps to Solve:

1. Install the Beautiful Soup library.
2. Open a web browser and navigate to the website **www.worldometers.info/population/: https://www.worldometers.info/population/.**
3. Use the Beautiful Soup library to parse the HTML content of the website.
4. Extract the data for each country as specified in the problem statement.
5. Save the data to a CSV file.

Activity 2

Prerequisites:

- Basic knowledge of Python
- Familiarity with pandas library
- A dataset of your choice

Problem Statement:

You are a data analyst who is interested in conducting Exploratory Data Analysis (EDA) on a dataset of your choice. You want to answer the following questions:

1. What are the main features of the dataset?
2. What are the distributions of the different features?
3. Are there any outliers in the dataset?
4. Are there any correlations between the different features?

Activity 2

Steps to Solve:

1. Import pandas library.
2. Load the dataset into a pandas DataFrame.
3. Use the describe() method to get a summary of the dataset.
4. Use the hist() method to plot the distributions of the different features.
5. Use the boxplot() method to identify outliers in the dataset.
6. Use the corr() method to calculate the correlations between the different features.

Next Session:
EDA – Case Study

THANK YOU

Please complete your assessments and review the self-learning content for this session on the **PRISM** portal.

knowledgehut
upGrad

EDA – Case Study

Pre-requisites

Hope you have gone through the self-learning content for this session on the PRISM portal.

By the End of this Session, You Will:

- Learn how to explore and visualize data.
- Identify and present interesting insights from data.
- Communicate your findings in a clear and concise way.
- Apply the concepts of EDA to a real-world problem.

Poll Time

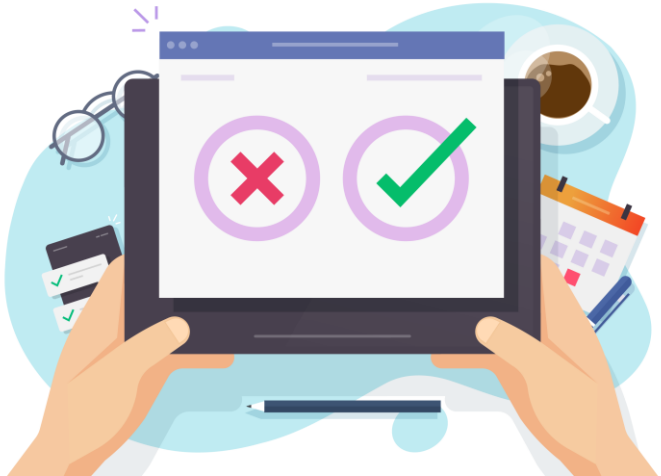
Q. Which of the following is not a common visualization used in EDA?

a. Bar chart

b. Line chart

c. Screen Plot

d. Heat map



Poll Time

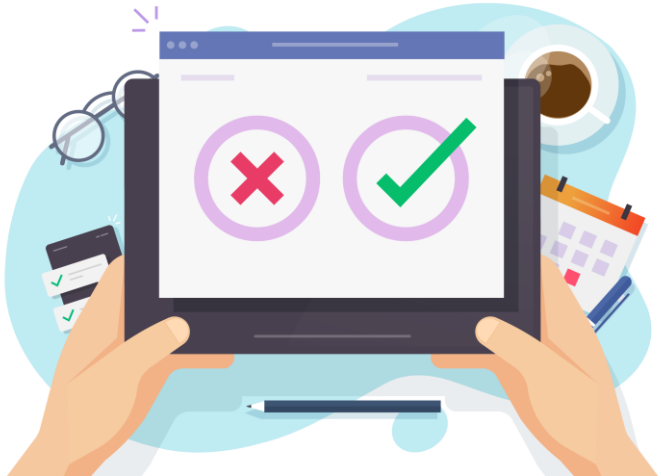
Q. Which of the following is not a common visualization used in EDA?

a. Bar chart

b. Line chart

c. Screen Plot

d. Heat map



Case Study – Problem Statement

Recap

- Fundamentals of Web Scraping
- Scraping data using BeautifulSoup.
- Identifying and dealing with missing values.
- Basics of Exploratory Data Analysis.

What Are You Going to Build?

- Exploratory Data Analysis (EDA) techniques to gain insights from data.
- Formulating relevant data queries and conducting data analysis.
- Data cleaning and handling to ensure data accuracy.
- Creating informative data visualizations for meaningful representations.
- Interpreting data insights to make informed decisions.

Poll Time

Q. What type of plot is used to analyze gender-wise customer distribution?

a. Bar Plot

b. Scatter Plot

c. Line Plot

d. Histogram

Poll Time

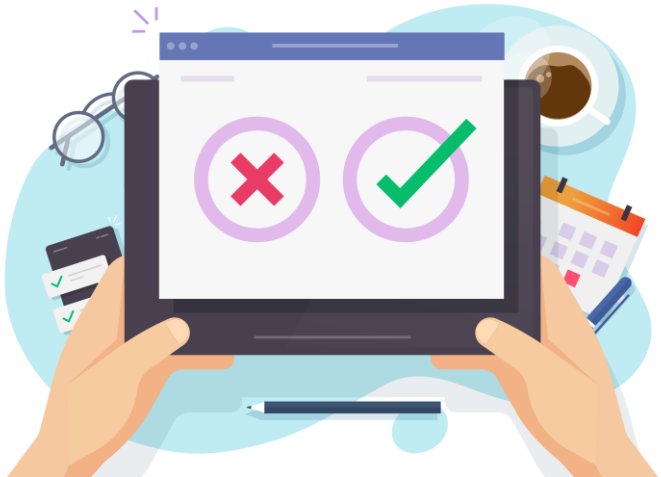
Q. What type of plot is used to analyze gender-wise customer distribution?

a. Bar Plot

b. Scatter Plot

c. Line Plot

d. Histogram



Hands-on: Case Study Questions

Poll Time

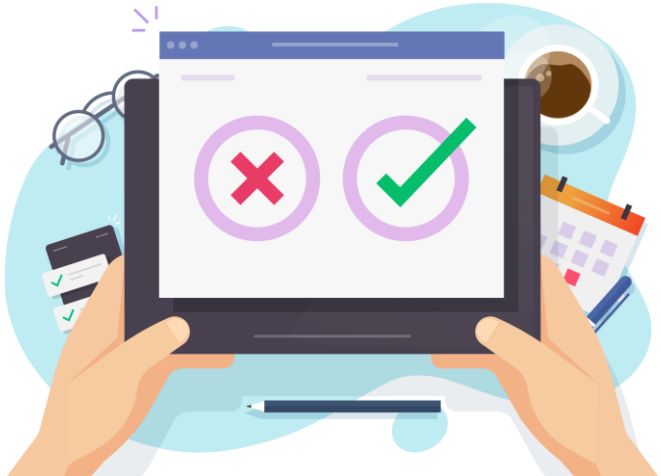
Q. How can the maximum and minimum job categories of customers be found?

a. By calculating the mean of job categories

b. By using a line chart

c. By identifying the most frequent and least frequent job categories

d. By using a Scatter Plot



Poll Time

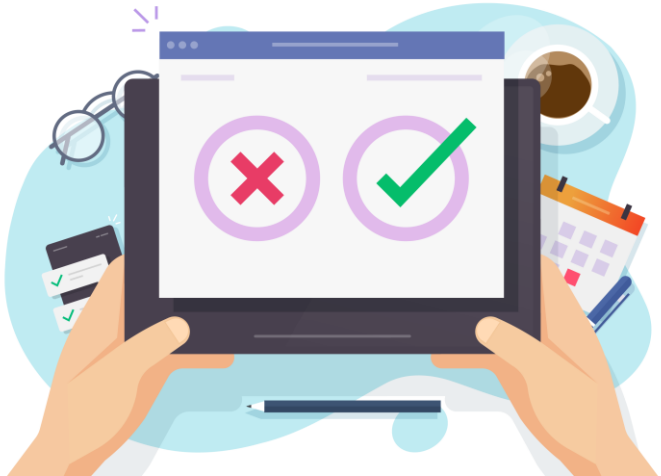
Q. How can the maximum and minimum job categories of customers be found?

a. By calculating the mean of job categories

b. By using a line chart

c. By identifying the most frequent and least frequent job categories

d. By using a Scatter Plot




Q&A

How Do Things Work in the Real World?

How Would You Go About This Twist?

Poll Time

Q. You are a mobile phone manufacturer. You want to know if there is enough demand for 5G phones. Which of the following would help you gather more information?

a. Conduct a survey of mobile phone users

b. Look at the sales figures for 5G phones

c. Analyze social media data

d. All of the listed

Poll Time

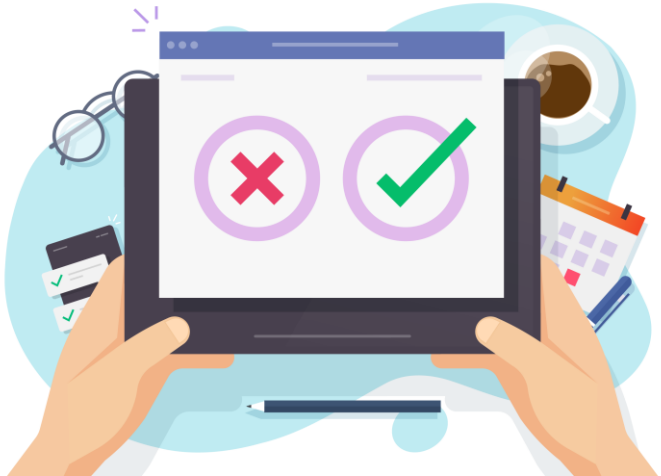
Q. You are a mobile phone manufacturer. You want to know if there is enough demand for 5G phones. Which of the following would help you gather more information?

a. Conduct a survey of mobile phone users

b. Look at the sales figures for 5G phones

c. Analyze social media data

d. All of the listed



Summary

- EDA is a process of exploring and understanding data. It is a systematic way of looking at data to identify patterns, trends, and relationships.
- It can identify patterns, trends, and relationships in data.
- EDA can be used to inform decision-making. There are a number of different tools and techniques that can be used for EDA.
- EDA is an iterative process. It is a repeated process as you learn more about the data. You may identify new patterns, trends, or relationships as you explore the data.

Activity 1

Prerequisite:

Basic knowledge of Python
Familiarity with the pandas library
Access to a dataset of patient medical records

Problem Statement:

You are a healthcare data scientist, and you are trying to identify patients who are at risk for developing diabetes. You have access to a dataset of patient medical records, which includes information such as age, gender, height, weight, blood pressure, and blood sugar levels.

Data:
https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Activity 1

Steps to Solve:

1. Load the dataset of patient medical records into a pandas DataFrame.

2. Explore the distribution of the data by plotting histograms and boxplots.

3. Look at the relationships between different features by plotting scatter plots and correlation matrices.

4. Identify any interesting insights from the data and summarize your findings in a report.


Q&A

Session Feedback

Next Session:
Data Visualization Using Python

THANK YOU

Please complete your assessments and review the self-learning content for this session on the **PRISM** portal.

knowledgehut
upGrad