Exploratory Data Analysis with Python

Exploratory data analysis (EDA) is an important step in the data analyzing process to understand the dataset better. By doing EDA, we can understand the main features of the data, the relationships between variables, and the variables that are relevant to our problem. EDA can also help us identify and handle missing or duplicate values, outliers, and errors in the data.

we're using Python to do data analysis because it has many libraries that can help us perform EDA, such as pandas, numpy, matplotlib, and seaborn. Pandas is a library for data manipulation and analysis. Numpy is a library for numerical computing. Matplotlib and seaborn are libraries for data visualization.

UNIT – I

Introduction to Data Science: Introduction to Data Science - Data Science Stages - Data Science Ecosystem - Tools used in Data Science - Data Science Workflow - Automated methods for Data Collection - Overview of Data - Sources of Data - Big Data - Data Categorization.

INTRODUCTION TO DATA SCIENCE:

Data Science is a multidisciplinary field that uses scientific methods, processes, algorithms, and systems to extract insights and knowledge from structured and unstructured data. It integrates techniques from statistics, computer science, and domain expertise.

DATA SCIENCE LIFE CYCLE:

Business Understanding: Ask relevant questions and define objectives for the problem that needs to be tackled.

Data Mining: Gather and scrape the data necessary for the projects.

Data Cleaning: Fix the inconsistencies within the data and handle the missing values.

Data Exploration: Form hypothesis about your defined problem by visually analyzing the data.

Feature Engineering: Select important features and construct more meaningful ones using the raw data that you have.

Predictive Modeling: Train Machine learning models, evaluate their performance, and use them to make predictions.

Data Visualization: Communicate the findings with key stakeholders using plots and interactive visualizations.

IMPORTANCE OF DATA SCIENCE:

- Data science helps brands to understand their customers in a much enhanced and empowered manner.
- It allows brands to communicate their story in such a engaging and powerful manner.
- Big data is a new field that is constantly growing and evolving.
- Its findings and results can be applied to almost any sector like travel, healthcare and education among others.
- Data science is accessible to almost all sectors.

USES:

Data Science is being used in almost all major industry. Here are some examples:

- Predicting customer preferences for personalized recommendations.
- Detecting fraud in financial transactions.
- Forecasting sales and market trends.
- Enhancing healthcare with predictive diagnostics and personalized treatments.
- Identifying risks and opportunities in investments.

STAGES OF DATA SCIENCE (OR) DATA SCIENCE WORKFLOW

The stages of data science are a structured process that helps to derive insights from data:

- Problem Identification
- Data Collection
- Data Cleaning and Preprocessing
- Exploratory Data Analysis (EDA)
- Modeling
- Evaluation
- Deployment and Communication

Problem Identification:

The first step in the data science project life cycle is to identify the problem that needs to be solved. This involves understanding the business requirements and the goals of the project. Once the problem has been identified, the data science team will plan the project by determining the data sources, the data collection process, and the analytical methods that will be used.

Data Collection:

The second step in the data science project life cycle is data collection. This involves collecting the data that will be used in the analysis. The data science team must ensure that the data is accurate, complete, and relevant to the problem being solved.

Data Cleaning and Preprocessing:

The third step in the data science project life cycle is data preparation. This involves cleaning and transforming the data to make it suitable for analysis. The data science team will remove any duplicates, missing values, or irrelevant data

from the dataset. They will also transform the data into a format that is suitable for analysis.

Exploratory Data Analysis (EDA):

The fourth step in the data science project life cycle is data analysis. This involves applying analytical methods to the data to extract insights and patterns. The data science team may use techniques such as regression analysis, clustering, or machine learning algorithms to analyze the data.

Modeling:

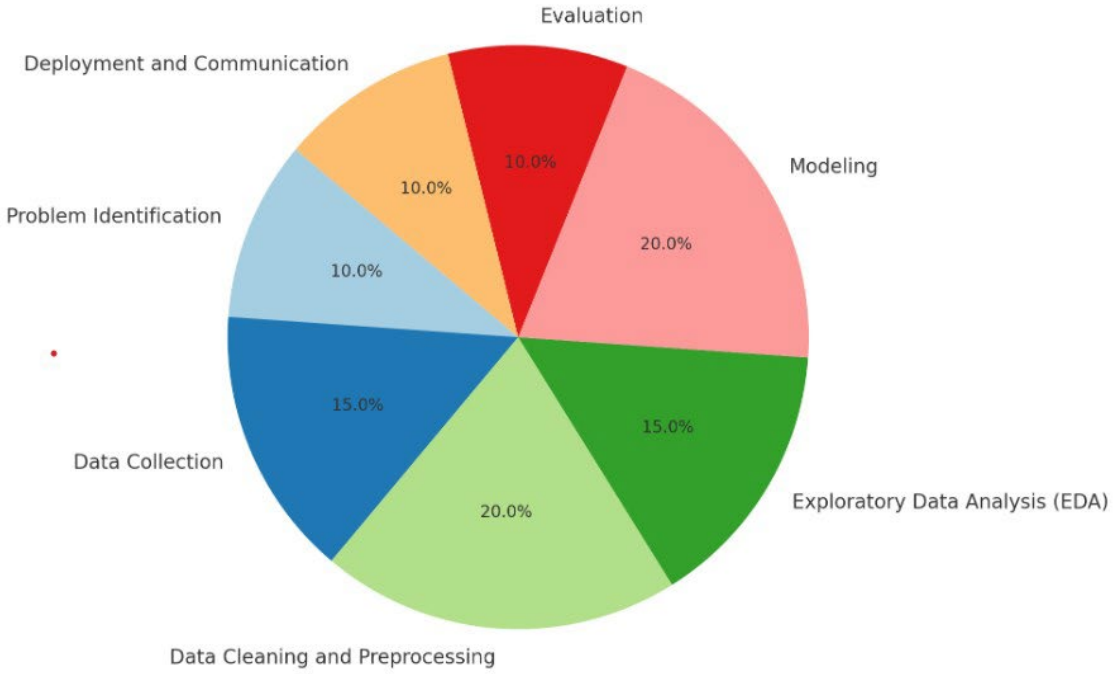
The fifth step in the data science project life cycle is model building. This involves building a predictive model that can be used to make predictions based on the data analysis. The data science team will use the insights and patterns from the data analysis to build a model that can predict future outcomes.

Evaluation:

The sixth step in the data science project life cycle is model evaluation. This involves evaluating the performance of the predictive model to ensure that it is accurate and reliable. The data science team will test the model using a validation dataset to determine its accuracy and performance.

Deployment and Communication:

The final step in the data science project life cycle is model deployment. This involves deploying the predictive model into production so that it can be used to make predictions in real-world scenarios. The deployment process involves integrating the model into the existing business processes and systems to ensure that it can be used effectively.



Stages of Data Science (or) Data science workflow

DATA SCIENCE ECO-SYSTEM

The Data Science Ecosystem refers to the collection of tools, technologies, frameworks, and processes used by data scientists to collect, analyze, visualize, and derive insights from data. It encompasses various elements that work together to enable data-driven decision-making.

- **Tools and technologies**

These include software applications, physical resources, and data storage tools like databases, data warehouses, and data lakes.

- **People**

These include data scientists, database administrators, and data analysts.

- **Data**

This includes data collected from various sources, such as customers, and data that is analyzed to inform business decisions.

- **Data science**

This includes the use of scientific methods, statistics, and algorithms to extract knowledge from data

A data science ecosystem can help businesses:

- Understand their customers
- Create better marketing, pricing, and operations strategies
- Make data-driven decisions
- Streamline data use
- Ensure consistent data quality and security
- Adapt to new needs

Key Components of the Ecosystem:

1. Programming Languages

Programming is the backbone of data science. Popular languages include:

Python:

- Libraries like Pandas, NumPy, and Matplotlib for data manipulation and visualization.
- Scikit-learn, TensorFlow for machine learning and AI.

R: Widely used for statistical computing and visualization.

SQL: Essential for querying and managing relational databases.

2.Data Storage & Management

Efficient storage is essential for managing large datasets effectively.

Relational Databases:

- MySQL, PostgreSQL, Oracle are used for managing structured data.
- They organize data in tables and maintain relationships between them.

NoSQL Databases:

- MongoDB, Cassandra are suitable for handling unstructured or semi-structured data.
- They help process large volumes of data quickly and efficiently.

Data Warehousing:

- Snowflake, Amazon Redshift are used for large-scale data storage and analysis.
- These are crucial for enterprise-level business data analytics.

3. Big Data Technologies

These technologies efficiently handle the **Volume, Velocity, and Variety** of big data.

- **Hadoop:**
  - A framework used for distributed storage and processing of massive datasets.
  - Ideal for managing large-scale data efficiently.
- **Apache Spark:**
  - A powerful tool for in-memory processing.
  - Helps process data quickly for real-time analytics.

4. Data Visualization

Visualization helps make data insights accessible and actionable.

Business Intelligence Tools:

• **Tableau:** A popular tool for creating data visualizations, dashboards, and interactive graphs.

• **Power BI:** A Microsoft tool widely used for data analysis and visualization.

• **Looker:** Another powerful tool used for data analysis and business intelligence.

Python Libraries:

• **Matplotlib:** A key library for creating basic graphs and charts for data visualization.

• **Seaborn:** Built on top of Matplotlib, it helps create more attractive and informative graphs.

• **Plotly:** A powerful library for creating interactive visualizations.

5. Machine Learning and AI Tools

Frameworks and libraries to develop predictive and prescriptive models.

Scikit-learn: It's a Python library used for traditional machine learning tasks like classification, regression, clustering, etc. It provides simple and efficient tools for data analysis and modeling.

TensorFlow and PyTorch: These are powerful frameworks used for deep learning and building neural networks. TensorFlow, developed by Google, and PyTorch, developed by Facebook, are both widely used for tasks like image recognition, natural language processing, and more complex AI models.

H2O.ai: Automated machine learning (AutoML) platform.

6. Data Collection and Integration

Tools for gathering and merging data from diverse sources.

Web Scraping:

- **Beautiful Soup:** A Python library used for web scraping to collect data from websites. It is used to parse HTML and XML documents.
- **Scrapy:** A Python framework for web scraping that helps efficiently collect and process large amounts of data.

APIs:

- APIs are used to connect to services for data collection. For example, APIs are used to retrieve data from services like Google Maps or Twitter.

ETL Tools:

- **Talend:** An open-source tool used for data extraction, transformation, and loading (ETL).
- **Apache NiFi:** An open-source tool used for data processing and flow development.

Cloud Platforms

Cloud services offer scalability, flexibility, and advanced analytics tools.

1. **Amazon Web Services (AWS):**
    - **S3:** A cloud storage service used for data storage.
    - **SageMaker:** An AWS service used to create, train, and deploy machine learning and artificial intelligence models.
2. **Google Cloud Platform (GCP):**
    - **BigQuery:** A Google Cloud service used for analyzing large datasets.
3. **Microsoft Azure:**
    - **Azure ML:** A service on Azure used to manage machine learning workflows, train models, and deploy them.

8. Collaboration and Workflow Tools

Platforms and tools to enhance productivity and team collaboration.

Version Control: Git, GitHub, Bitbucket are used for code tracking & team collaboration.They help track changes in code and enable multiple developers to work together efficiently.

Notebooks: Jupyter Notebooks, Google Colab are used for code sharing & data visualization.Widely used in data science, machine learning, and research.

TOOLS USED IN DATA SCIENCE

Data science requires a variety of tools to perform tasks such as data collection, cleaning, analysis, visualization, modeling, and deployment. Here's a categorized overview of tools commonly used in data science:

1. Programming and Scripting Tools

- **Python:** Most popular language for data manipulation, analysis, and machine learning (e.g., libraries like Pandas, NumPy, Scikit-learn).
- **R:** Ideal for statistical analysis and data visualization.
- **SQL:** Essential for querying and managing structured data in databases.

2. Data Collection Tools

- **Beautiful Soup**: Python library for web scraping.
- **Scrapy**: Framework for extracting data from websites.
- **APIs**: Tools like Postman facilitate data retrieval from services (e.g., Twitter API).
- **ETL Tools**: Talend, Apache NiFi, and Informatica for extracting, transforming, and loading data.

3. Data Cleaning and Preprocessing Tools

- **OpenRefine**: For cleaning messy data.
- **Excel**: Commonly used for smaller datasets and initial cleaning.
- **Python Libraries**: Pandas and NumPy for handling missing values, normalization, etc.

4. Data Analysis Tools

- **Jupyter Notebooks**: Interactive environment for data analysis and visualization.
- **Google Colab**: Cloud-based notebook for Python with GPU support.
- **RStudio**: Integrated development environment (IDE) for R programming.

5. Data Visualization Tools

- **Tableau**: For creating interactive dashboards and visualizations.
- **Power BI**: Business intelligence tool for data visualization.
- **Python Libraries**: Matplotlib, Seaborn, and Plotly for creating custom plots.
- **ggplot2**: Visualization library in R

6. Machine Learning and AI Tools

- **Scikit-learn**: For implementing traditional machine learning algorithms.
- **TensorFlow**: Deep learning framework for neural networks.
- **PyTorch**: Alternative to TensorFlow for deep learning tasks.
- **H2O.ai**: Platform for scalable machine learning and AutoML.

7. Big Data Tools

- **Apache Hadoop**: Framework for distributed storage and processing of big data.
- **Apache Spark**: In-memory big data processing tool.
- **Hive**: SQL-like interface for querying large datasets in Hadoop.

8. Cloud Platforms

- **Amazon Web Services (AWS)**: Services like S3 for storage, SageMaker for ML.
- **Google Cloud Platform (GCP)**: BigQuery for data analysis, AutoML for model building.
- **Microsoft Azure**: Azure ML for machine learning workflows.

9. Data Storage and Management Tools

- **Relational Databases**: MySQL, PostgreSQL, SQLite for structured data.
- **NoSQL Databases**: MongoDB, Cassandra for unstructured or semi-structured data.
- **Data Warehouses**: Snowflake, Amazon Redshift, Google BigQuery.

10. Workflow and Collaboration Tools

- **Git**: For version control and collaboration.
- **GitHub**: Repository hosting for collaborative coding.
- **Slack/Trello**: For communication and project management.

11. Automation Tools

- **Airflow**: Workflow automation for scheduling tasks.
- **Knime**: For building and deploying data workflows visually.
- **RapidMiner**: Drag-and-drop platform for machine learning and data prep.

12. Statistical Tools

- **SPSS**: Widely used for statistical analysis.
- **SAS**: Advanced analytics and statistical modelling.

- **MATLAB**: Used in numerical computing and advanced statistical modeling.

These tools are chosen based on the specific requirements of a project, the size of the data, and the complexity of the tasks involved.

AUTOMATED METHODS FOR DATA COLLECTION:

Automated data collection refers to the process of gathering data efficiently without manual intervention using specialized tools and techniques. These methods ensure scalability, accuracy, and time efficiency.

1. Web Scraping

One common method is web scraping, which extracts data from websites using tools like Beautiful Soup, Scrapy, or Selenium, often applied to tasks such as monitoring e-commerce prices or gathering reviews. (or) Extracting data from websites using scripts or tools.

- **Tools:** Beautiful Soup, Scrapy, Selenium.
- **Applications:**
    - Collecting product prices from e-commerce sites.
    - Gathering reviews or comments from social media or forums.

2. APIs (Application Programming Interfaces)

Another popular approach is using APIs (Application Programming Interfaces) to retrieve structured data directly from web services, such as weather updates from OpenWeather API or user data from social media platforms like Twitter. (or) Retrieving structured data directly from web services or applications.

- **Tools**: Postman, Python requests library.
- **Applications**:

- Fetching weather data from Open Weather API.
- Retrieving user data from social media APIs (e.g., Twitter API).

3. IoT Devices and Sensors

IoT devices and sensors are also significant contributors, collecting real-time data such as temperature from smart thermostats or health metrics from wearable devices. (or) Collecting real-time data from Internet of Things (IoT) devices or sensors.

- **Examples**:
  - Smart thermostats collecting temperature data.
  - Wearable devices recording health metrics.

4. Log Files and System Monitoring Tools

Log files and system monitoring tools like Splunk and ELK Stack enable the analysis of server or application logs for performance and behavioral insights. (or) Analyzing server or application logs for insights.

- **Tools**: Splunk, ELK Stack (Elasticsearch, Logstash, Kibana).
- **Applications**:
  - Monitoring user behavior on websites.
  - Detecting anomalies in system performance.

5. Automated Surveys and Forms

For user feedback and market research, automated surveys and forms via platforms like Google Forms or Typeform are widely used. (or) Using online platforms to distribute and collect survey responses.

- **Tools**: Google Forms, Typeform, SurveyMonkey.

- **Applications**:
    - Collecting customer feedback.
    - Conducting market research.

6. Data Streaming Services

In real-time analytics, data streaming services such as Apache Kafka and Amazon Kinesis capture live data streams, including stock market trends and social media feeds. (or) Capturing real-time data from streaming platforms.

- **Tools**: Apache Kafka, Amazon Kinesis, Google Pub/Sub.
- **Applications**:
    - Tracking stock market data.
    - Processing social media live feeds.

7. ETL Tools (Extract, Transform, Load)

ETL (Extract, Transform, Load) tools like Talend and Apache NiFi automate data integration and migration, consolidating data from various sources for centralized analytics. (or) Automating data integration from multiple sources into a centralized system.

- **Tools**: Talend, Apache NiFi, Informatica.
- **Applications**:
    - Migrating data between systems.
    - Consolidating business data for analytics.

8. Cloud-Based Data Pipelines

Additionally, cloud-based data pipelines on platforms like AWS, Google Dataflow, and Azure Data Factory streamline workflows for large-scale data

processing and machine learning applications. (or) Automating data workflows on cloud platforms.

- **Tools**: AWS Data Pipeline, Google Dataflow, Azure Data Factory.
- **Applications**:
    - Processing large-scale data in real time.
    - Enabling machine learning pipelines.

9. Email and Document Parsing

Lastly, email and document parsing tools extract structured information from unstructured sources such as emails or PDF reports, automating tasks like invoice processing or data extraction from reports. (or) Extracting structured information from emails or documents.

- **Tools**: PyPDF2 for PDFs, email parsers like Parsio.
- **Applications**:
    - Automating invoice processing.
    - Extracting data from reports.

Benefits of Automated Data Collection

- **Efficiency**: Saves time and resources.
- **Scalability**: Handles large-scale data effortlessly.
- **Consistency**: Ensures uniform data collection processes.
- **Real-Time Capability**: Enables continuous data flow for immediate insights.

Overview of Data

Data refers to raw facts, figures, or information collected for analysis. It forms the foundation of insights, patterns, and decisions in areas like business, science, and technology. By processing and interpreting data, organizations or individuals can make informed decisions.

Types of Data:

1. **Qualitative Data:**
    - Descriptive data that is non-numeric.
    - It can capture characteristics or qualities, like colors, textures, labels, or opinions.
    - **Example:** "The sky is blue," or "The customer is satisfied."
2. **Quantitative Data:**
    - Numerical data that can be measured or counted.
    - It allows for statistical analysis and can be used for more objective comparisons or conclusions.
    - **Example:** "The temperature is 75°F," or "The sales revenue is $10,000."

Data Characteristics:

1. **Accuracy:**
    - This refers to how close the data is to the actual, real-world value. Accurate data ensures reliable and truthful conclusions can be drawn from it.
    - **Example:** If you're recording temperatures, accurate data means it reflects the actual temperature measured.

2. **Completeness:**
    - Complete data includes all necessary and relevant information needed to draw meaningful insights.
    - **Example:** If you're tracking sales, missing data points like a sale amount or a date would
    - make the dataset incomplete**.**
3. **Consistency:**
    - Consistent data means that the information is the same across different databases or systems, without contradictions.
    - **Example:** If a customer's name is spelled differently in different parts of the system, that data is inconsistent.
4. **Timeliness:**
    - Timely data is up-to-date and relevant for the current time period or decision-making process.
    - **Example:** If you're tracking stock prices, timely data reflects the most recent market conditions, not outdated figures.

Sources of Data

Data can come from numerous sources, both structured and unstructured, and they can be categorized based on their origin.

1. Primary Data:

- **Definition:** This is data that is directly collected for a specific purpose by the researcher or organization.
- **Examples:**
    - Surveys: Asking people questions to collect information.
    - Interviews: One-on-one or group conversations to gather insights.
    - Experiments: Controlled tests to gather data on specific variables.

- Observations: Directly watching and recording behaviors or events.

2. Secondary Data:

- **Definition:** This data has already been collected by someone else and is repurposed for new research or analysis.
- **Examples**:
  - Government Reports: Data published by government agencies on various topics.
  - Academic Papers: Studies and research conducted by scholars.
  - Market Research Reports: Data provided by firms that analyze consumer behavior, trends, and markets.

3. Internal Data:

- **Definition:** This is data generated within an organization, often related to its operations and activities.
- **Examples:**
  - Company Sales Data: Information on the company's sales performance.
  - Employee Records: Data about the organization's staff, such as performance, attendance, and salaries.
  - Financial Transactions: Data related to the company's income, expenses, and profits.

4. External Data:

- **Definition:** This is data that comes from outside an organization, typically from third-party sources.
- **Examples:**
  - Social Media Data: Information gathered from platforms like Facebook, Twitter, Instagram, etc.
  - Public Datasets: Data made available by government bodies, NGOs, or research organizations**.**

- Data from External Partners: Data shared by other companies or entities that the organization collaborates with.

5. Big Data Sources:

- **Definition:** Big Data refers to large volumes of data generated at high speeds, often in real-time, that require advanced processing techniques.
- **Examples:**
  - Social Media Platforms: Data generated from user interactions, posts, likes, comments, etc.
  - IoT Devices: Data collected from Internet of Things devices like smart home devices, sensors, and wearables.
  - E-commerce Platforms: Data from online shopping activities, including customer preferences, purchase history, and browsing behavior.

Big Data

Big data refers to extremely large datasets that are complex, varied, and grow at an exponential rate. Traditional data processing tools cannot efficiently manage, store, or analyze these datasets. The concept of big data is commonly defined by the 5 V's:

1. **Volume:** Refers to the massive size of data generated from various sources, such as social media, IoT devices, and e-commerce platforms. For example, Facebook generates terabytes of data daily.
2. **Velocity:** Indicates the speed at which data is generated and processed. Real-time data, like stock market feeds or live social media updates, highlights the need for fast processing.
3. **Variety:** Describes the different types of data, including structured (databases), semi-structured (JSON, XML), and unstructured (images, videos, social media posts).

4. **Veracity:** Reflects the uncertainty and reliability of data. Data must be cleansed and validated to ensure accuracy for analysis.
5. **Value:** Highlights the importance of extracting meaningful insights from raw data for business or societal benefits.

Applications of Big Data

- Healthcare: Analyzing patient data for improved diagnosis and personalized treatments.
- E-commerce: Optimizing customer experience through personalized recommendations.
- Finance: Fraud detection and risk analysis in real-time.
- Transportation: Enhancing logistics and traffic management with IoT data.

Challenges in Big Data

1. Data storage and management require scalable solutions like Hadoop or cloud platforms.
2. Data security and privacy concerns arise due to the sensitive nature of data.
3. Integration of varied datasets from multiple sources is complex.

Big Data - Data Categorization

Big data refers to the massive volume of structured, semi-structured, and unstructured data generated daily from various sources. Data categorization is the process of organizing this data into meaningful groups based on its characteristics, which is crucial for analysis and decision-making.

Types of Big Data Categorization:

1. **Based on Data Format:**
    - Structured Data: Organized data stored in rows and columns (e.g., relational databases). Example: Customer data, financial transactions.
    - Semi-Structured Data: Data that does not conform fully to a structured format but still contains tags or markers. Example: JSON, XML, logs.
    - Unstructured Data: Data without a predefined format or organization. Example: Images, videos, social media posts.
2. **Based on Data Source:**
    - Human-Generated Data: Data created by human activities. Example: Social media content, emails, or documents.
    - Machine-Generated Data: Data generated automatically by machines or sensors. Example: IoT data, server logs, GPS data.
3. **Based on Purpose:**
    - Transactional Data: Data that records business transactions. Example: Online shopping orders, payments.
    - Analytical Data: Data used for insights and decision-making. Example: Historical sales data, forecasting data.

Importance of Data Categorization:

1. Enables efficient storage and retrieval.
2. Facilitates better data analysis and insights.
3. Enhances data management and governance.
4. Improves decision-making by identifying patterns and trends.

Examples in Real Life:

- In healthcare, patient records (structured data) and medical imaging (unstructured data) are categorized for diagnosis and treatment.

- Social media companies categorize user data (likes, shares, posts) to improve user experience and target advertisements.

In conclusion, data categorization in Big Data is a critical process that helps organizations organize vast amounts of data, making it usable and insightful for various application**s.**