EDA- AN OVERVIEW

-AJITHA

WHAT IS EDA?

EDA-Exploratory Data Analysis

Summarizes the data sets' main characteristics

To discover patterns

To spot anomalies

To test hypothesis

To check hypothesis

Determines if the chosen statistical techniques are appropriate

Developed by American mathematician John Tukey in the 1970s

EDA techniques continue to be a widely used method in the data discovery process today

BY WHOM?

WHY EDA?

Helps analysing before even making any assumptions

Identifies any obvious errors

Finds interesting relations among variables

Ensures the validity of the analysis produced by the data scientists

Helps stakeholders by confirming they are asking the right questions

Answers questions about standard deviations, categorical variables, and confidence intervals

TYPES OF EDA

- Univariate non-graphical
- Univariate graphical
- Multivariate nongraphical
- Multivariate graphical

UNIVARIATE NON-GRAPHICAL

This is simplest form of data analysis

The data being analyzed consists of just one variable

Doesn’t deal with causes or relationships

Describe the data and find patterns that exist within it.

UNIVARIATE GRAPHICAL METHODS

Stem-and-leaf plots: Display all data values along with the shape of the distribution.

Histograms: Bar plot representing the frequency or proportion of cases for a range of values.

Box plots: Graphically depict the five-number summary (minimum, first quartile, median, third quartile, maximum) of the data distribution.

MULTIVARIATE NONGRAPHICAL METHOD

Involve analyzing relationships between two or more variables through cross-tabulation or statistics.

Examples include contingency tables, correlation coefficients, or measures of association like chi-square tests.

MULTIVARIATE GRAPHICAL METHODS

Utilize graphics to display relationships between two or more sets of data.

Commonly used graphic: Grouped bar plot or bar chart, with each group representing one level of one variable and each bar within a group representing levels of another variable.

PANDAS:

Excels at data manipulation and cleaning.

Functions like fillna() for imputing missing values, drop() for removing rows/columns, and various indexing techniques for selecting specific data subsets.

Operations like merging datasets (concat(), join()) and data type conversion (astype()) become essential for data preparation.

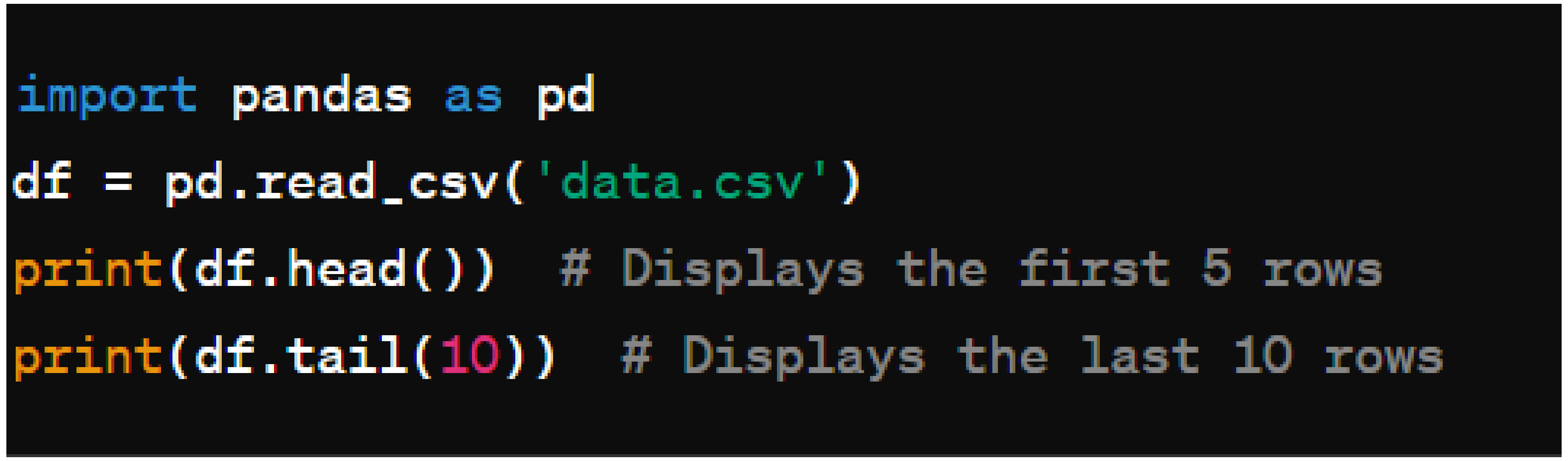
DATA WRANGLING WITH PYTHON LIBRARIES

NUMPY

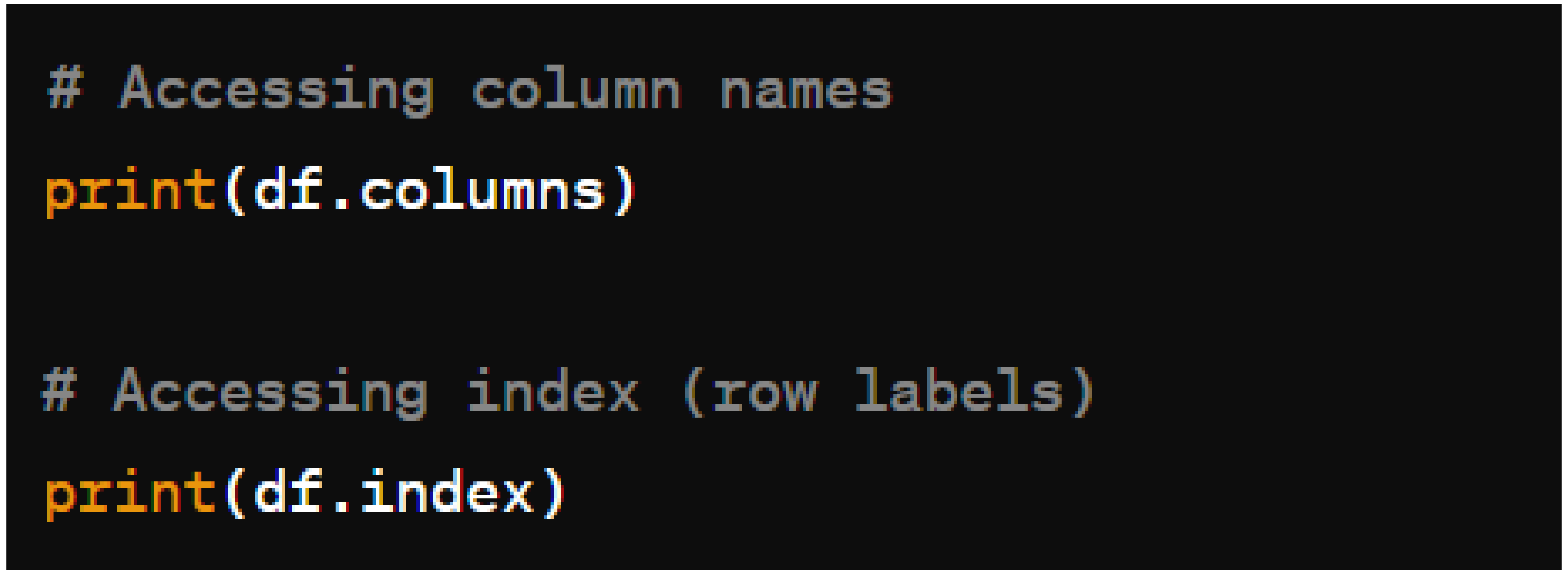
For numerical computations, NumPy provides efficient arrays and vectorized operations.

Can calculate summary statistics like mean, standard deviation, percentiles using functions like mean(), std(), percentile().

Code Snippets

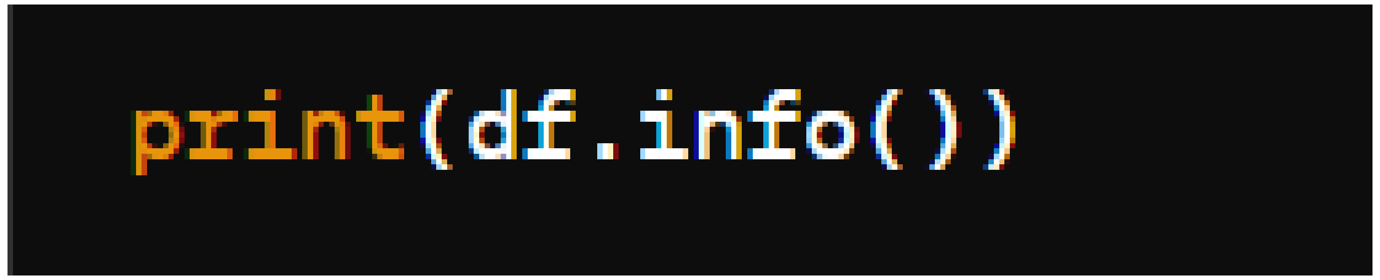


```python
import pandas as pd
df = pd.read_csv('data.csv')
print(df.head())  # Displays the first 5 rows
print(df.tail(10))  # Displays the last 10 rows
```

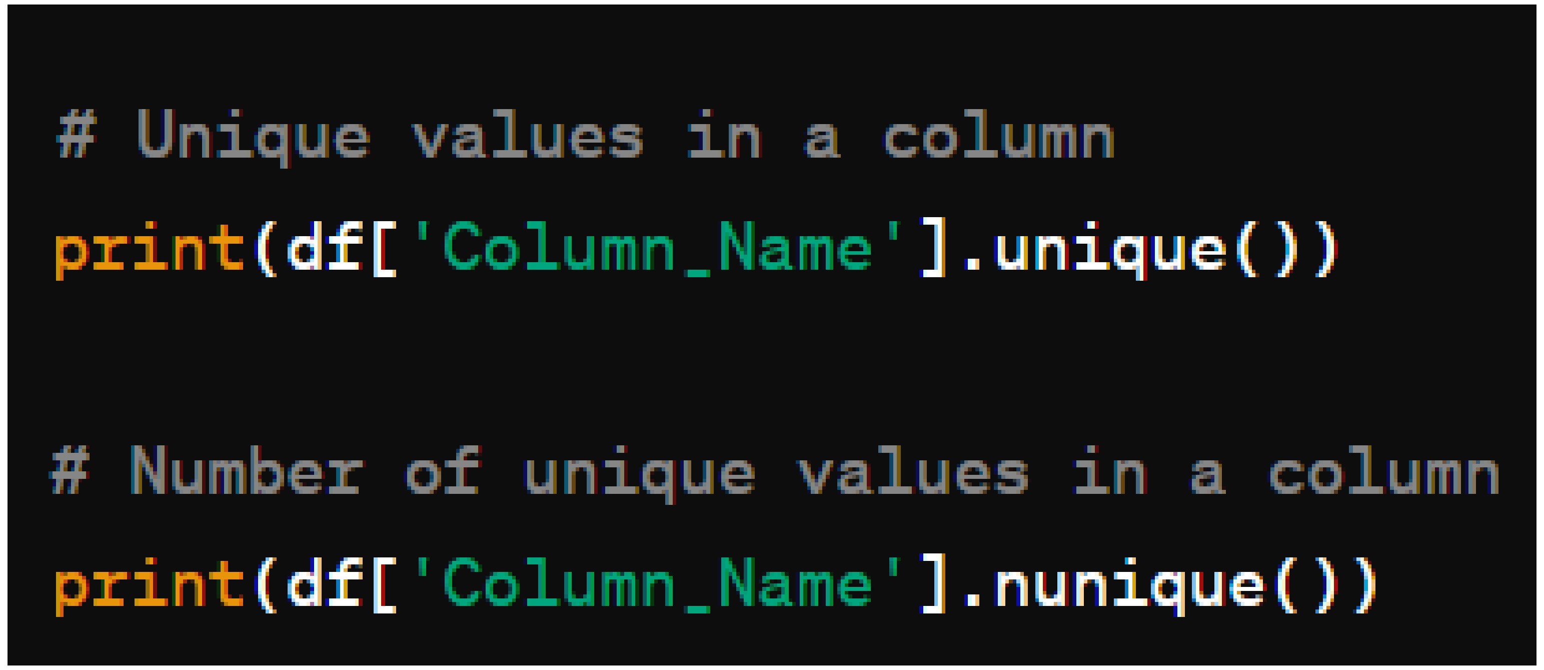


```python
# Accessing column names
print(df.columns)

# Accessing index (row labels)
print(df.index)
```



```python
print(df.info())
```



```python
# Unique values in a column
print(df['Column_Name'].unique())

# Number of unique values in a column
print(df['Column_Name'].nunique())
```

VISUALIZATION POWERHOUSES: MATPLOTLIB AND SEABORN

MATPLOTLIB:

This fundamental library offers extensive customization for creating various plots. Functions like hist() generate histograms, boxplot() creates boxplots, and scatter() allows for visualizing relationships between variables.

You can control plot elements like colors, labels, and legends for clear communication of insights.

SEABORN

Built on top of Matplotlib, Seaborn provides high-level functions for creating aesthetically pleasing and informative visualizations.

It offers specialized functions for specific plots like violin plots, jointplots (combining scatter and histograms) and heatmaps (visualizing correlations across multiple variables).

DATA ACQUISITION

Pandas is used to import data from various file formats such as CSV, Excel, SQL databases, JSON, etc.

read_csv(),

read_excel(),

read_sql()

DATA INSPECTION AND EXPLORATION

Once data is loaded, Pandas allows for quick inspection and exploration of the dataset.

Functions like head(), tail(), info(), describe(), and shape are used to get an overview of the dataset, its structure, and basic statistics.

DATA CLEANING AND PREPROCESSING

Pandas offers powerful tools for data cleaning and preprocessing, which are essential steps in EDA.

It provides methods for handling missing values (dropna(), fillna()), removing duplicates (drop_duplicates()), and transforming data (apply(), map()).

INDEXING AND SELECTION

Pandas allows for easy indexing and selection of data, enabling users to extract subsets of data for analysis.

Techniques like label-based indexing (loc[]), position-based indexing (iloc[]), and boolean indexing are commonly used.

DESCRIPTIVE STATISTICS

Pandas provides functions to compute descriptive statistics for numerical data, such as mean, median, standard deviation, etc.

Methods like mean(), median(), std(), min(), max(), and quantile() are used to calculate these statistics.

THANK YOU!