M.Tech Program

Advanced Industry Integrated Programs

Jointly offered by University and LTIMindTree

Data Engineering

Knowledge partner

LTIMindtree

Implementation partner

L&T EduTech

Course Objective

- Recognize data types and structures.
- Grasp big data fundamentals and analytics.
- Master data ingestion processes and tools.
- Understand exploratory data analysis techniques.
- Learn storage methods and data flow.

Modules

- Data Types & Formats
- Data Ingestion techniques
- Data Profiling & Visual Representation via various tools (Pandas)
- Storage and retrieval methods
- Data Lineage Analysis

Data Profiling & Visual Representation via various tools (Pandas)

Exploratory Data Analysis

Exploratory Data Analysis

Data Profiling

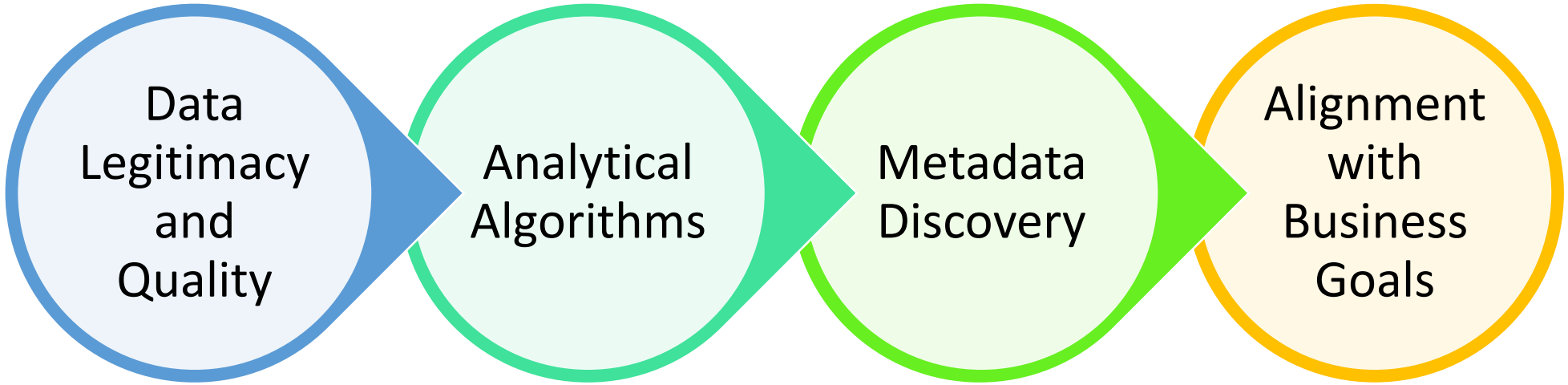
Exploratory Data Analysis

Data Profiling

- Data profiling involves examining and analyzing data to create useful summaries, providing a high-level overview that helps identify data quality issues, risks, and trends.
- This process offers critical insights that companies can use to their advantage.

Exploratory Data Analysis

Data Profiling – How it works



Exploratory Data Analysis

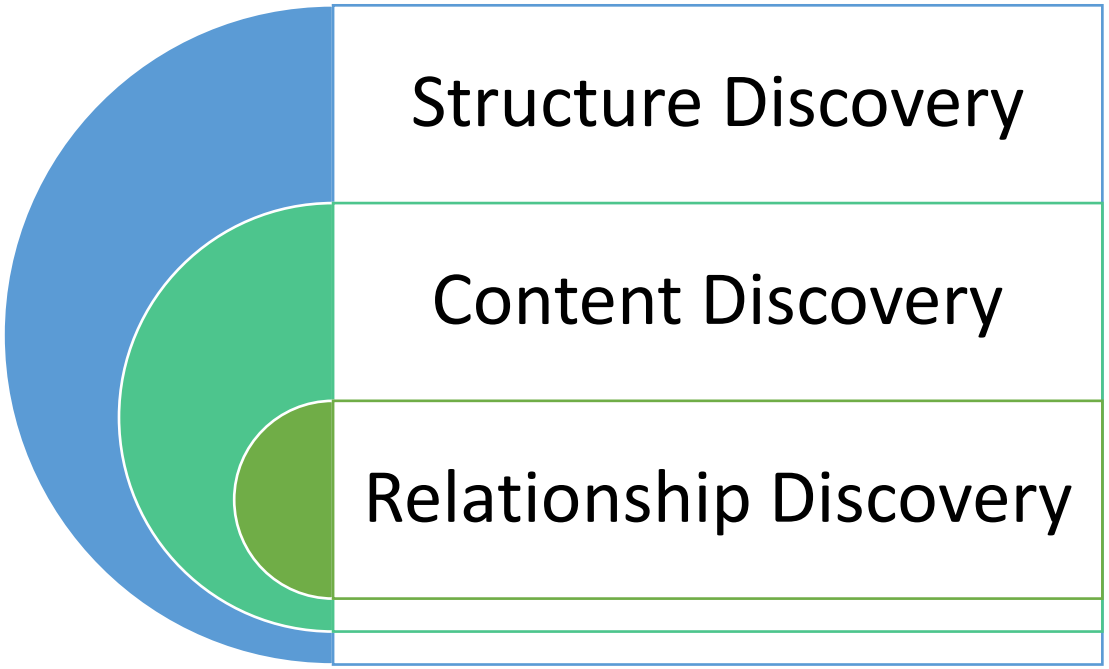
Benefits of Data Profiling

- Improved Data Quality and Credibility
- Predictive Decision Making
- Proactive Crisis Management
- Organized Data Sorting

Exploratory Data Analysis

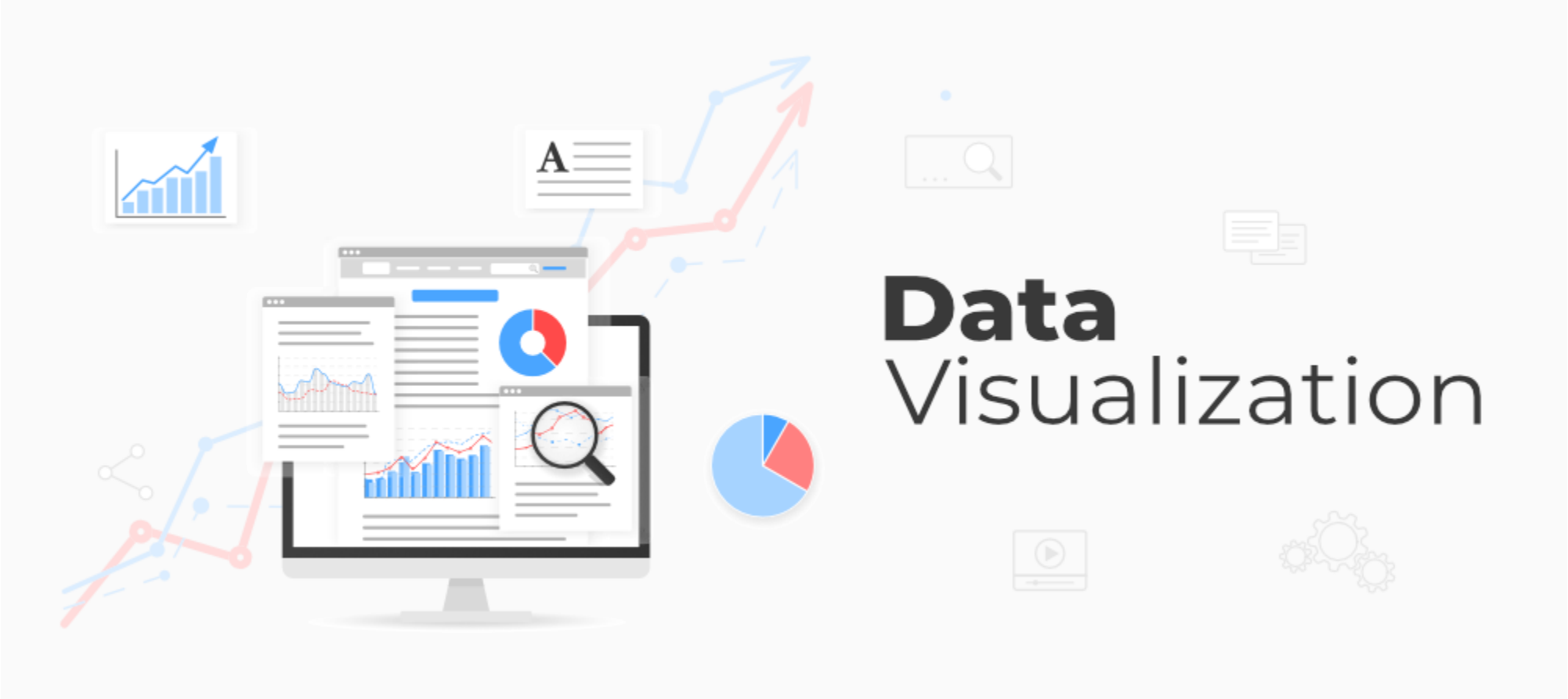
Types of Data Profiling

- Data profiling applications analyze databases by organizing and collecting information. These techniques can be categorized into three main types:



Exploratory Data Analysis

Data Visualization



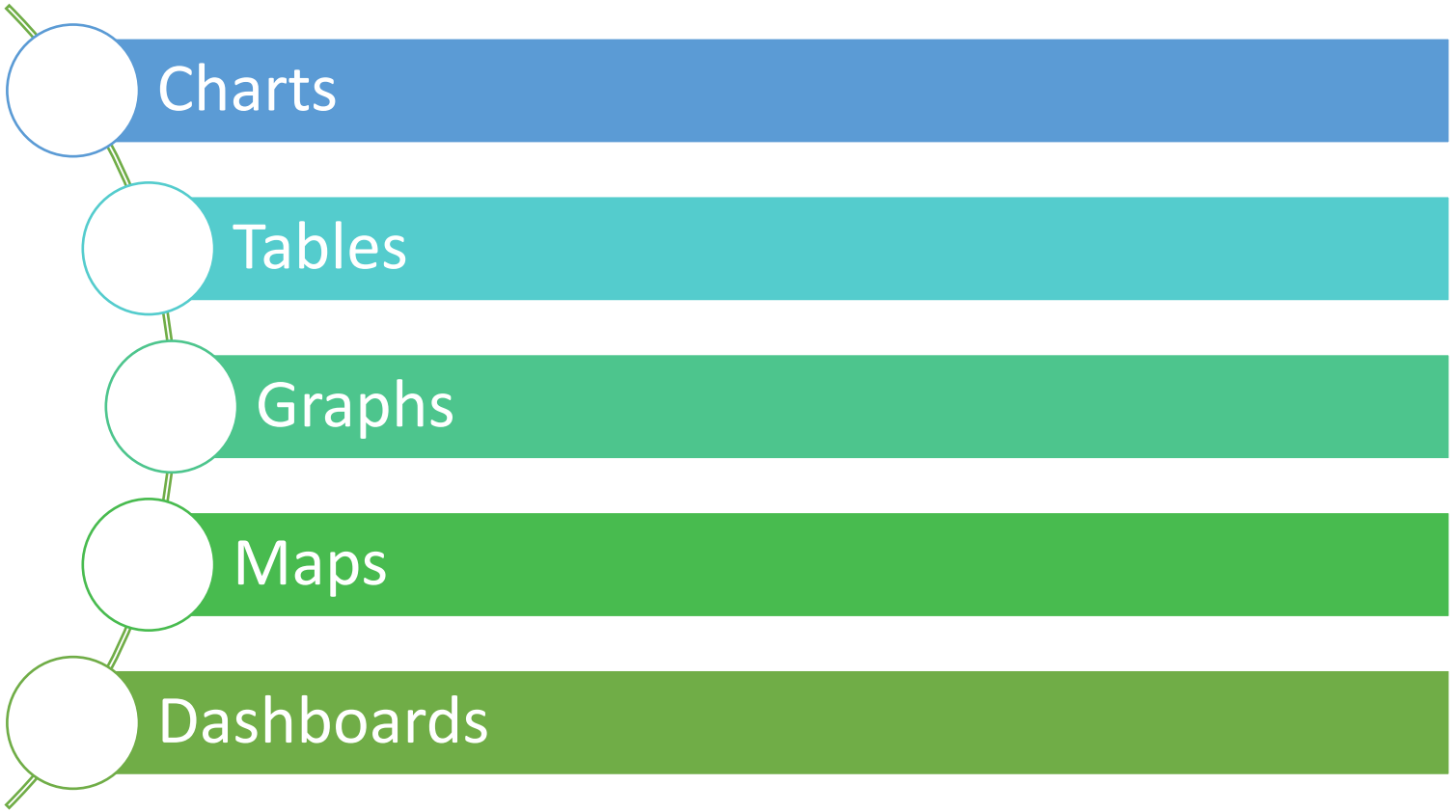
Exploratory Data Analysis

Data Visualization

- The graphical representation of information and data.
- Examples include charts, graphs, and maps.
- Purpose of Data Visualization:
  - Makes it easier to see and understand trends, patterns, and outliers in data.
  - Helps in analyzing large amounts of information.
  - Supports making data-driven decisions.

Exploratory Data Analysis

Common Types of Data Visualization



Exploratory Data Analysis

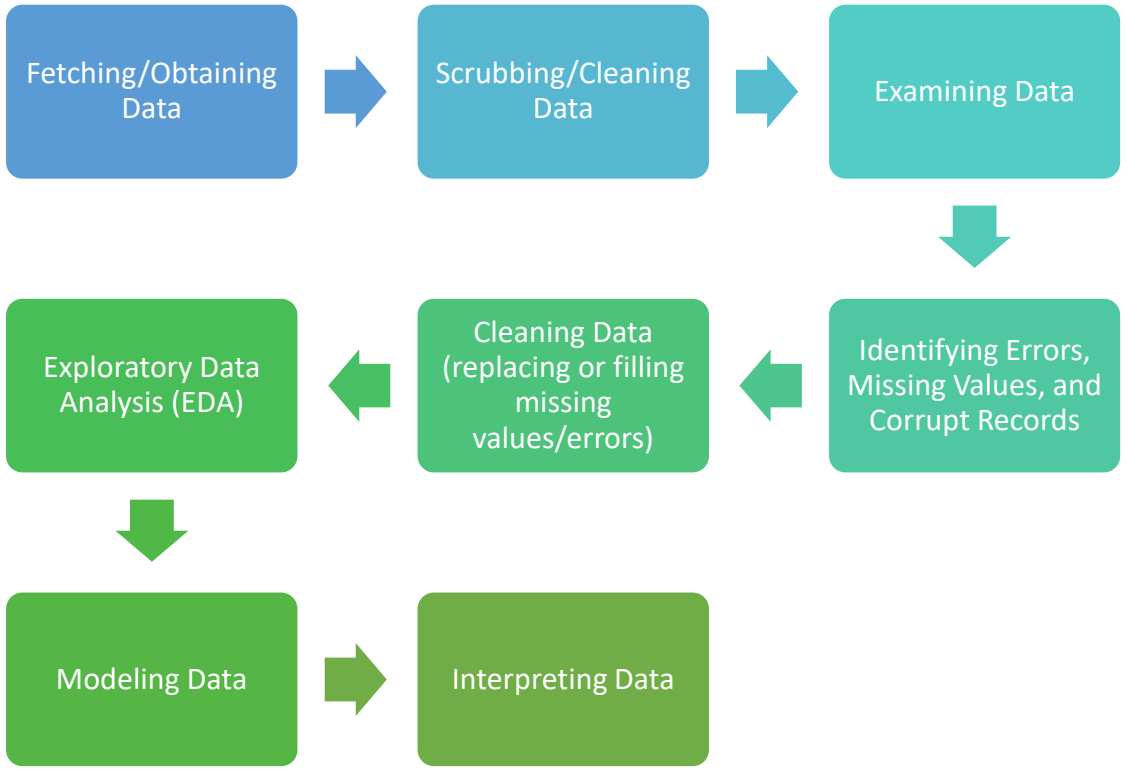
Data Visualization Important

- Data Visualization Discovers the Trends in Data
- Data Visualization Provides a Perspective on the Data
- Data Visualization Puts the Data into the Correct Context
- Data Visualization Saves Time
- Data Visualization Tells a Data Story

Exploratory Data Analysis

Data Science Pipeline

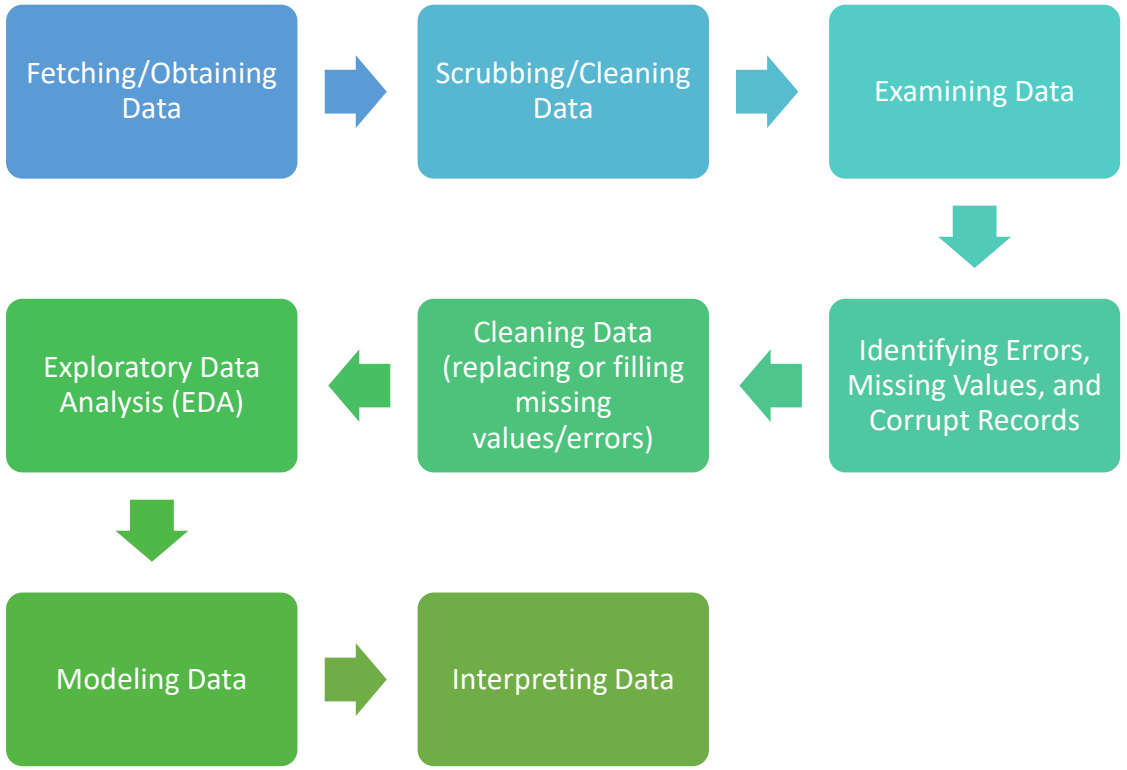
- Transforms raw data from various sources into an understandable format for analysis.



Exploratory Data Analysis

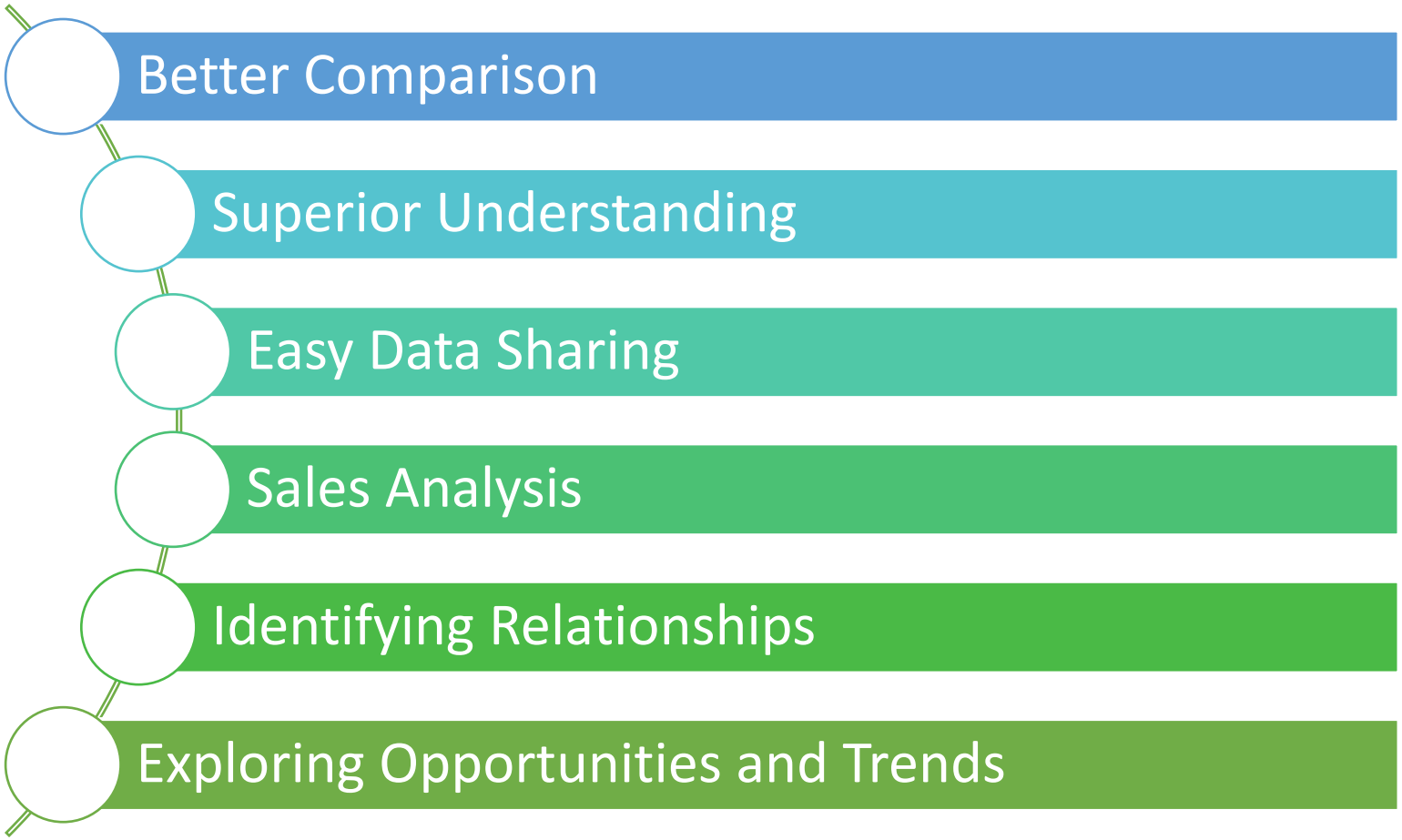
Data Science Pipeline

- Transforms raw data from various sources into an understandable format for analysis.



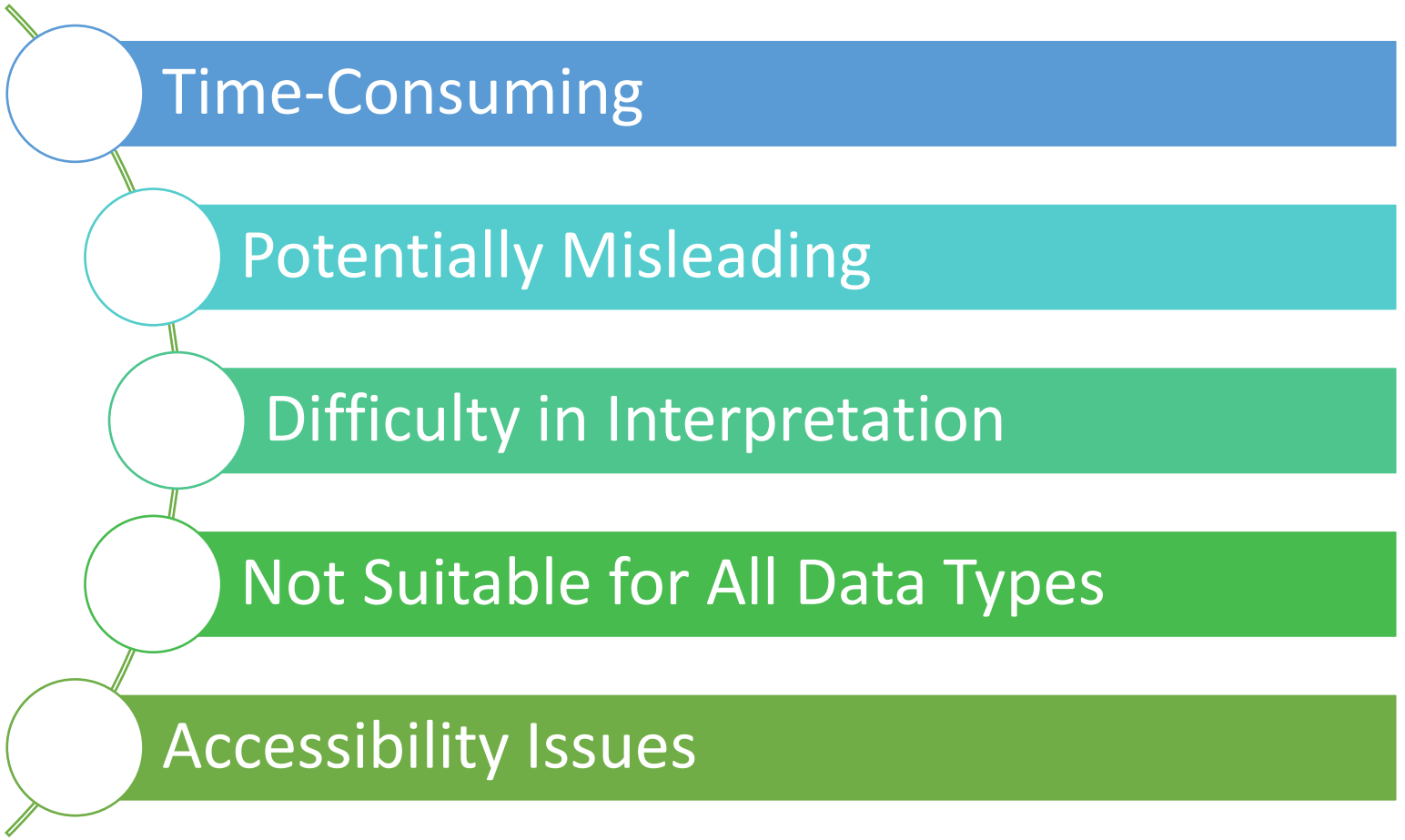
Exploratory Data Analysis

Advantages of Data Visualization



Exploratory Data Analysis

Disadvantages of Data Visualization



Exploratory Data Analysis

Data Visualization Tools

- Tableau
- Looker
- Zoho Analytics
- Sisense
- IBM Cognos Analytics
- Qlik Sense
- Domo
- Microsoft Power BI
- Klipfolio
- SAP Analytics Cloud

Exploratory Data Analysis

What is Exploratory Data Analysis

Purpose	• Analyzes datasets to understand their main characteristics.
Methods	• Summarizes data features. • Detects patterns. • Uncovers relationships.
Techniques	• Uses visual and statistical methods.
Benefits	• Provides insights • Helps formulate hypotheses for further analysis

Exploratory Data Analysis

Data Pre-processing and Feature Engineering?

- Data-Driven Processes
- Exploratory Data Analysis (EDA
- Data Pre-Processing
- Feature Engineering
- Importance of Feature Engineering
- Pre-Processing Focus

Exploratory Data Analysis

EDA Using Python

Step 1: Import Libraries

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import warnings

warnings.filterwarnings('ignore')
```

Exploratory Data Analysis

EDA Using Python

Step 2: Load the Dataset

```
data = pd.read_csv("used_cars.csv")
```

Exploratory Data Analysis

EDA Using Python

Step 3: Initial Data Inspection

- Get the shape (number of rows and columns) of the dataset.
- Display the first and last 5 rows.
- Use .info() to check data types and missing values.python

Exploratory Data Analysis

EDA Using Python

Step 3: Initial Data Inspection

```
print(data.shape)

print(data.head())

print(data.tail())

data.info()
```

Exploratory Data Analysis

EDA Using Python

Step 4: Check for Duplicates

Use .nunique() to find unique values in each column.

data.nunique()

Exploratory Data Analysis

EDA Using Python

Step 5: Handle Missing Values

- Identify missing values using data.isnull().sum().
- Calculate the percentage of missing values.
- data.isnull().sum()
- (data.isnull().sum() / len(data)) * 100

Exploratory Data Analysis

EDA Using Python

Step 6: Data Reduction

- Drop irrelevant columns (e.g., ID columns).
- data = data.drop(['S.No.'], axis=1)

Exploratory Data Analysis

EDA Using Python

Step 7: Feature Engineering

- Create new features like car age from the year.
- from datetime import date
- data['Car_Age'] = date.today().year - data['Year']
- Split names to get brand and model.
- data['Brand'] = data.Name.str.split().str.get(0)
- data['Model'] = data.Name.str.split().str.get(1) + data.Name.str.split().str.get(2)

Exploratory Data Analysis

EDA Using Python

Step 8: Data Cleaning

- Correct inconsistent brand names.
- data["Brand"].replace({"ISUZU": "Isuzu", "Mini": "Mini Cooper", "Land": "Land Rover"}, inplace=True)

Exploratory Data Analysis

EDA Using Python

Step 9: Statistical Summary

- Use data.describe().T to get statistics summary for numerical columns.Use data.describe(include='all').T for all columns.
- data.describe().T
- data.describe(include='all').T

Exploratory Data Analysis

EDA Using Python

Step 10: Separate Numerical and Categorical Variables

- cat_cols = data.select_dtypes(include=['object']).columns
- num_cols = data.select_dtypes(include=np.number).columns.tolist()
- print("Categorical Variables:", cat_cols)
- print("Numerical Variables:", num_cols)

Exploratory Data Analysis

EDA Using Python

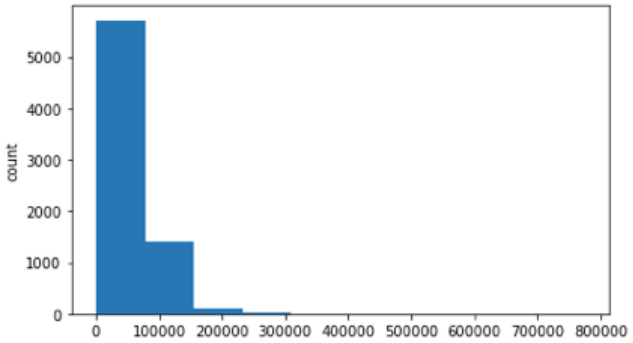
Step 11: Univariate Analysis

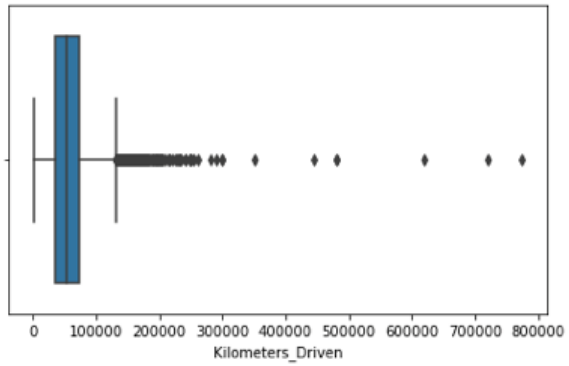
- Visualize numerical variables using histograms and box plots.
- for col in num_cols:
- plt.figure(figsize=(15, 4))
- plt.subplot(1, 2, 1)
- data[col].hist(grid=False)
- plt.subplot(1, 2, 2)
- sns.boxplot(x=data[col])
- plt.show()

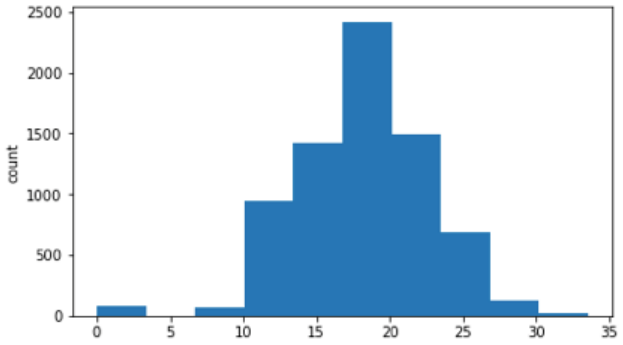
Exploratory Data Analysis

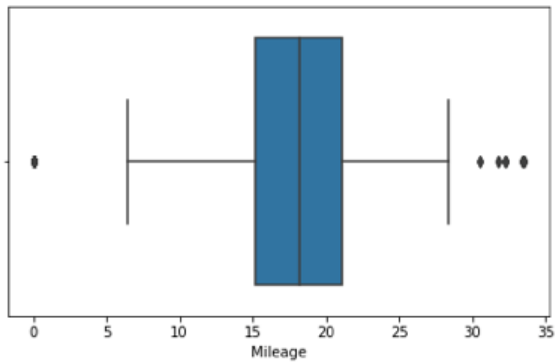
EDA Using Python

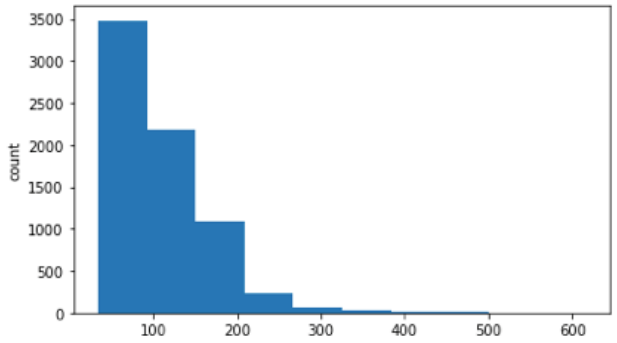
Step 11: Univariate Analysis

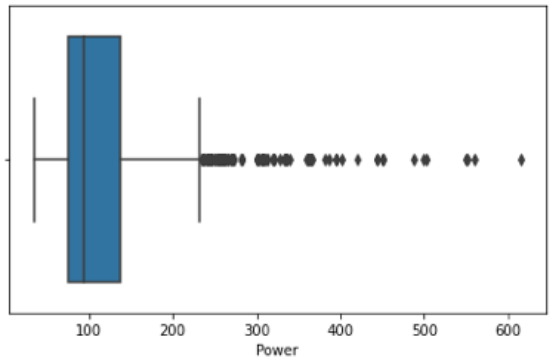












Exploratory Data Analysis

EDA Using Python

Step 11: Visualize categorical variables using count plots.

- fig, axes = plt.subplots(3, 2, figsize=(18, 18))
- sns.countplot(ax=axes[0, 0], x='Fuel_Type', data=data)
- sns.countplot(ax=axes[0, 1], x='Transmission', data=data)
- sns.countplot(ax=axes[1, 0], x='Owner_Type', data=data)
- sns.countplot(ax=axes[1, 1], x='Location', data=data)
- sns.countplot(ax=axes[2, 0], x='Brand', data=data.head(20))
- sns.countplot(ax=axes[2, 1], x='Model', data=data.head(20))
- plt.show()

Exploratory Data Analysis

EDA Using Python

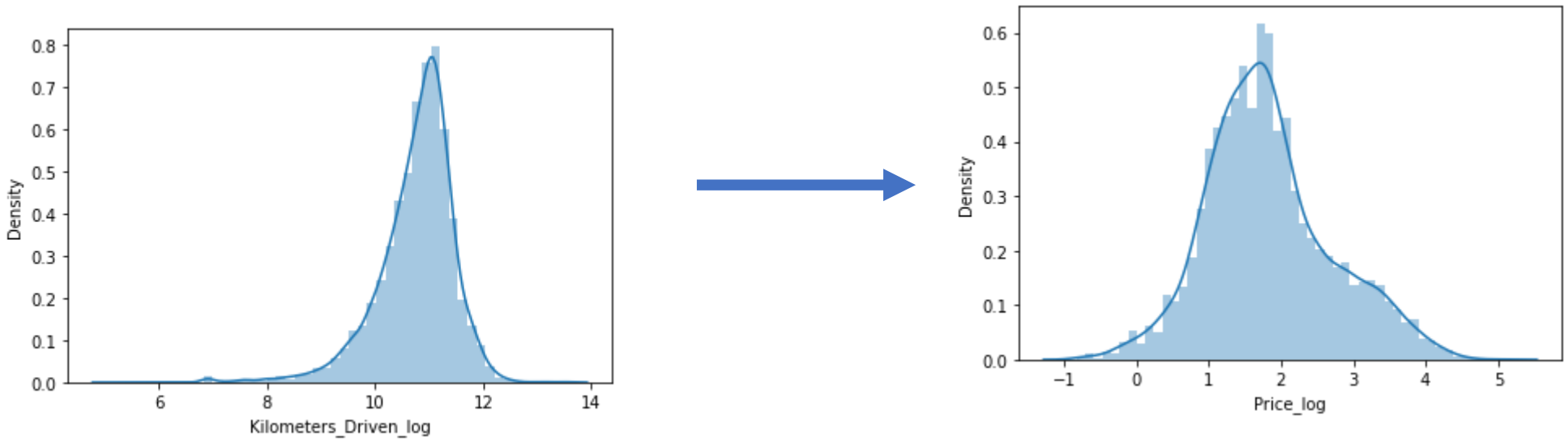
Step 12: Data Transformation

- Apply log transformation to skewed data.
- def log_transform(data, cols):
- for col in cols:
- data[col + '_log'] = np.log(data[col] + 1)
- log_transform(data, ['Kilometers_Driven', 'Price'])
- sns.distplot(data["Kilometers_Driven_log"], axlabel="Kilometers_Driven_log")
- plt.show()

Exploratory Data Analysis

EDA Using Python

Step 12: Data Transformation



Exploratory Data Analysis

EDA Using Python

Step 13: Bivariate Analysis

- Use pair plots and bar plots to examine relationships between variables.
- sns.pairplot(data=data.drop(['Kilometers_Driven', 'Price'], axis=1))
- plt.show()

Exploratory Data Analysis

EDA Using Python

Step 14: Multivariate Analysis

- Use a heat map to visualize correlations.
- plt.figure(figsize=(12, 7))sns.heatmap(data.drop(['Kilometers_Driven', 'Price'], axis=1).corr(), annot=True, vmin=-1, vmax=1)plt.show()

Exploratory Data Analysis

EDA Using Python

Step 15: Impute Missing Values

- Impute missing values based on domain knowledge and patterns in the data.
- data['Mileage'].fillna(value=np.mean(data['Mileage']), inplace=True)
- data['Seats'] = data.groupby(['Model', 'Brand'])['Seats'].apply(lambda x: x.fillna(x.median()))
- data['Engine'] = data.groupby(['Brand', 'Model'])['Engine'].apply(lambda x: x.fillna(x.median()))
- data['Power'] = data.groupby(['Brand', 'Model'])['Power'].apply(lambda x: x.fillna(x.median()))

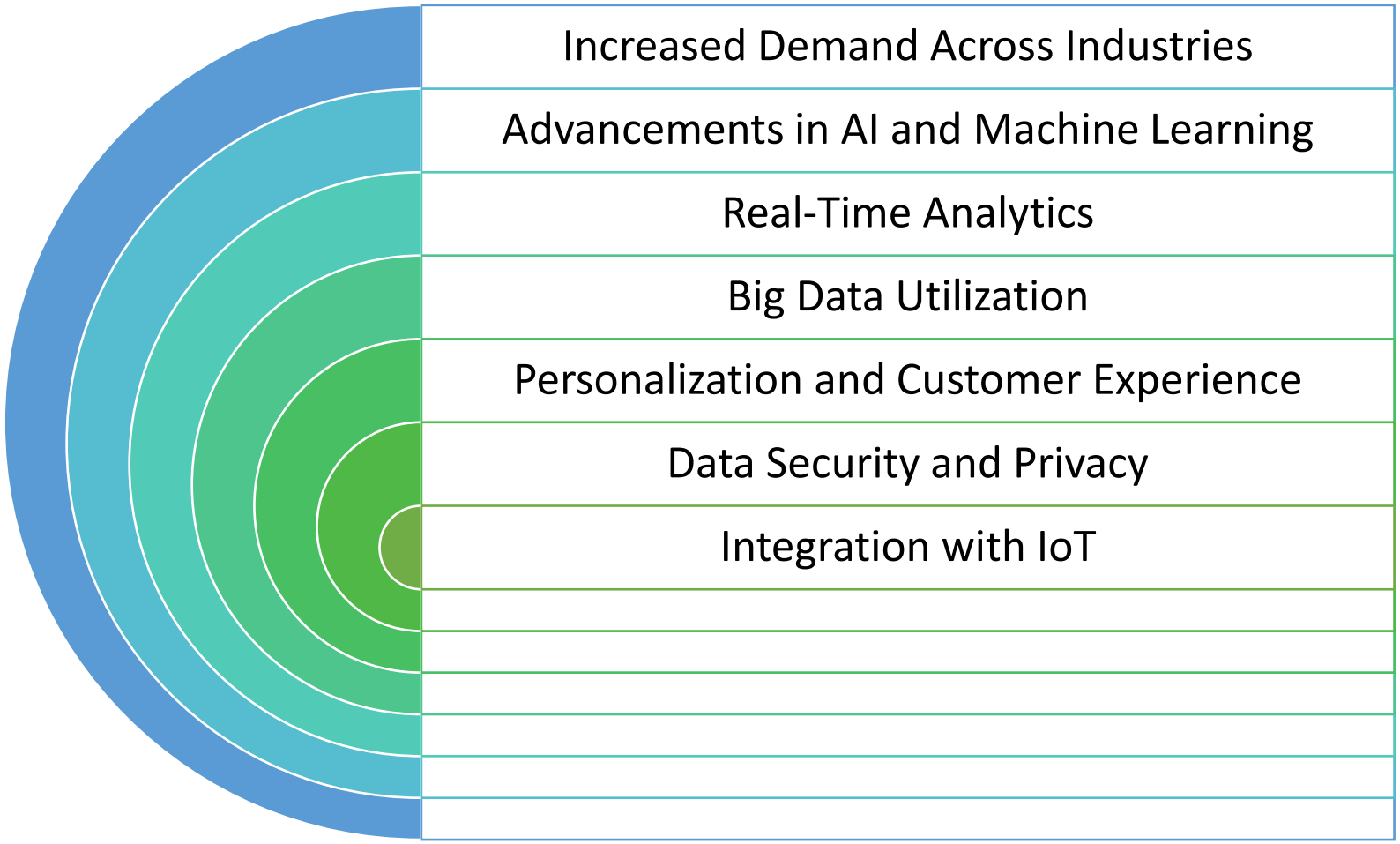
Exploratory Data Analysis

Market Analysis with Exploratory Data Analysis

- Analyze Purchase Patterns and Preferences
- Understanding Customer Journeys and Touchpoints
- Identifying Market Segments Based on Behavior
- Analyzing Conversion Rates and Customer Engagement
- Optimizing Marketing Strategies Based on EDA Insights

Exploratory Data Analysis

Future Scope of Data Analytics



Exploratory Data Analysis

Data Analysis with Python

Specify Data Requirements	• Determine what data is needed
Prepare or Collect Data	• Gather the data required for analysis.
Clean and Process	• Handle missing data, remove duplicates, and ensure data quality
Analyze	• Use statistical and computational techniques to explore data.
Share	• Present findings using visualizations and reports
Act or Report	• Make decisions or recommendations based on the analysis.

Exploratory Data Analysis

Data Analysis with Python

Creating NumPy Arrays

```
import numpy as np

b = np.empty(2, dtype=int)
print("Matrix b : \n", b)

a = np.empty([2, 2], dtype=int)
print("\nMatrix a : \n", a)

c = np.empty([3, 3])
print("\nMatrix c : \n", c)
```

Exploratory Data Analysis

Data Analysis with Python

Using numpy.zeros()

```
import numpy as np

b = np.zeros(2, dtype=int)
print("Matrix b : \n", b)

a = np.zeros([2, 2], dtype=int)
print("\nMatrix a : \n", a)

c = np.zeros([3, 3])
print("\nMatrix c : \n", c)
```

Exploratory Data Analysis

Data Analysis with Python

NumPy Array Indexing and Slicing

```
import numpy as np

a = np.arange(10, 1, -2)
print("\nA sequential array with a negative step: \n", a)

newarr = a[np.array([3, 1, 2])]
print("\nElements at these indices are:\n", newarr)
```

```
import numpy as np

a = np.arange(20)
print("\nArray is:\n ", a)

print("\n a[-8:17:1] = ", a[-8:17:1])
print("\n a[10:] = ", a[10:])
```

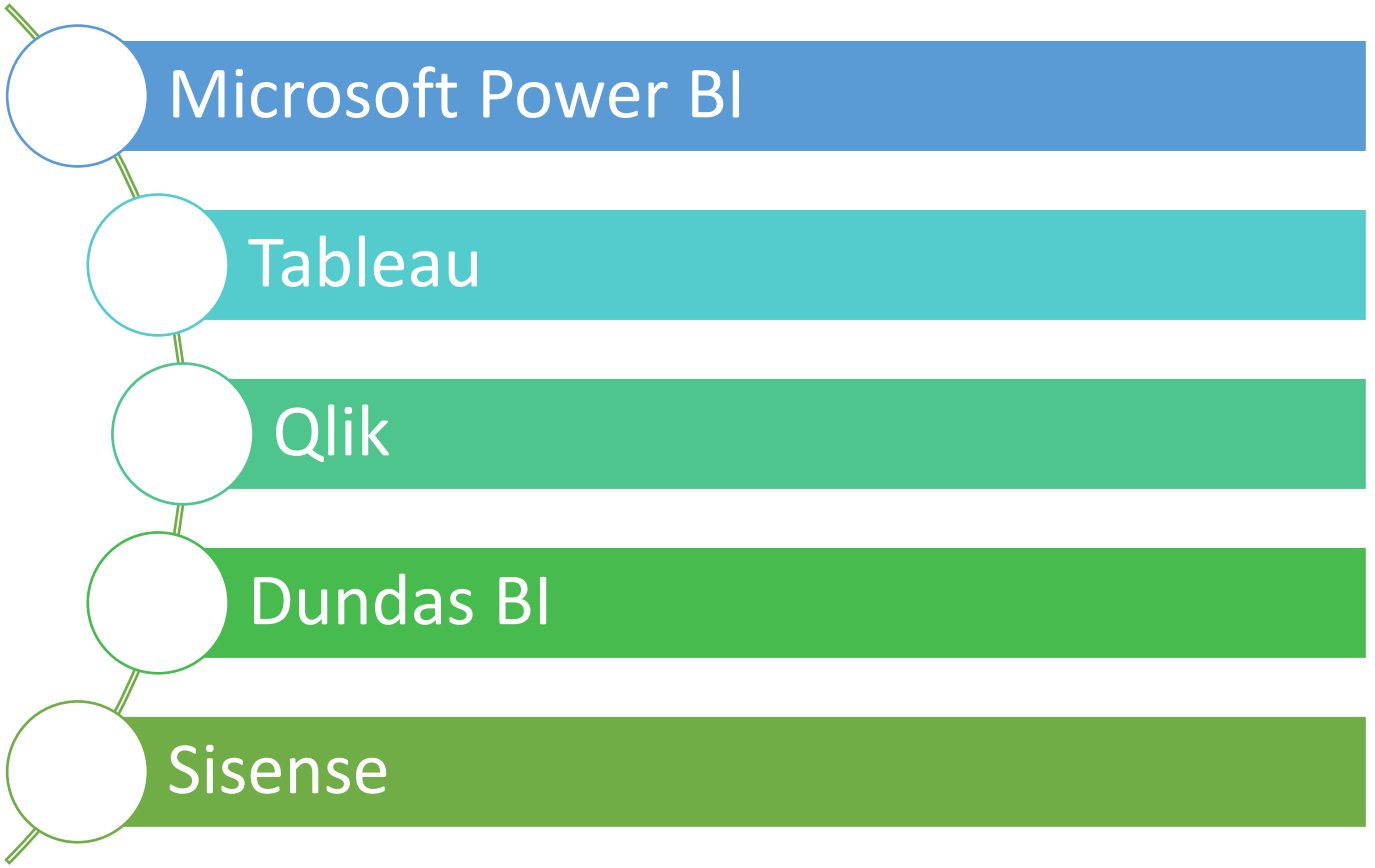
Exploratory Data Analysis

Analyzing Data Using Pandas

- Pandas is a Python library used for handling relational or labeled data.
- It offers various data structures to manipulate data and time series.
- Pandas is built on top of the NumPy library.
- The library is typically imported using the alias pd: import pandas as pd.
- Using pd as an alias helps to write less code.
- However, using an alias is not mandatory.
- Pandas provides two main data structures for manipulating data:
- Series: a one-dimensional labeled array.
- DataFrame: a two-dimensional labeled data structure (like a table).

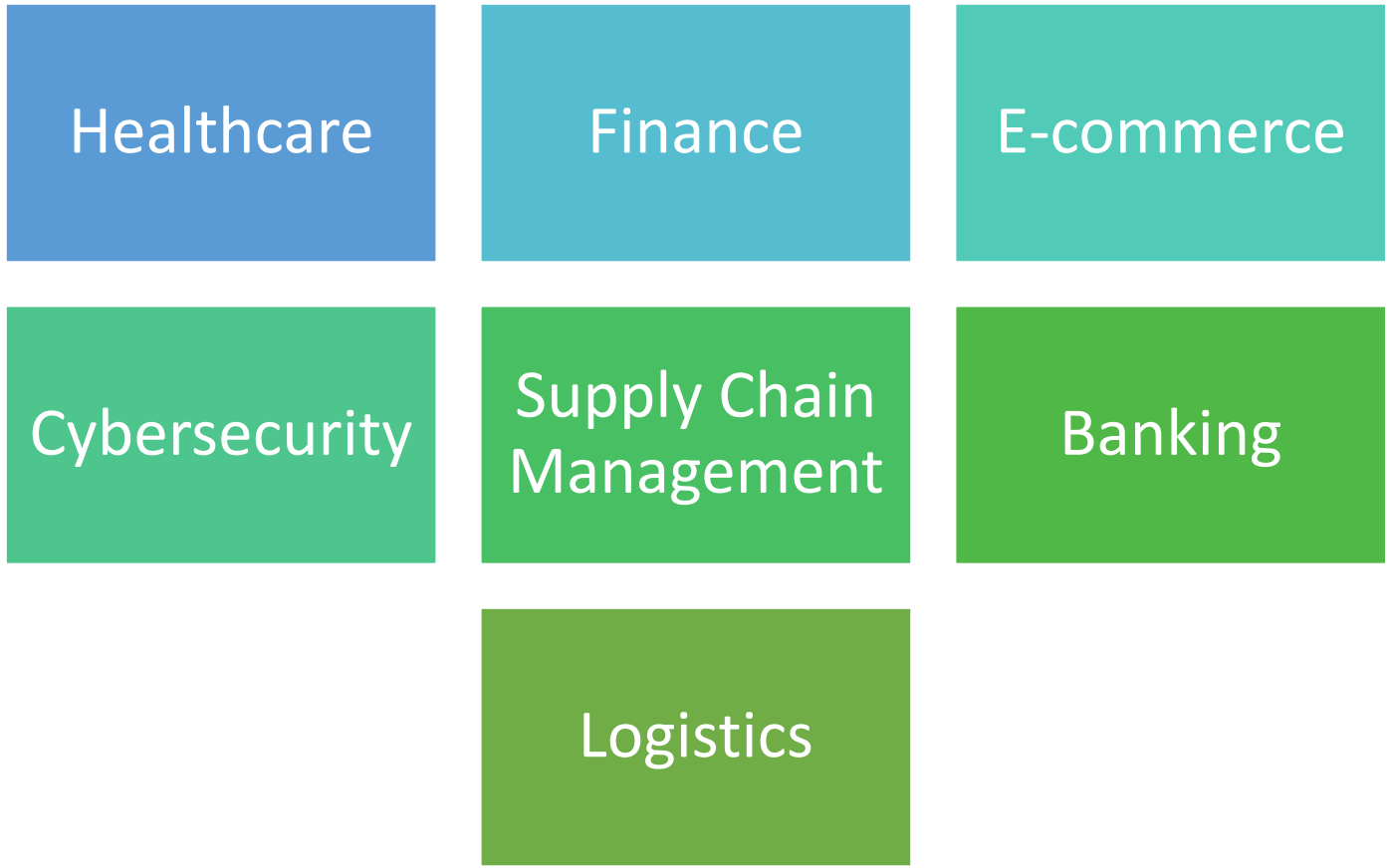
Exploratory Data Analysis

Top Business Intelligence Tools



Exploratory Data Analysis

Application of Data Analytics



Exploratory Data Analysis

Retrieving in Data Analytics

Data Sources

- Databases: SQL, NoSQL databases.
- APIs: Public or private APIs.
- Web Scraping: Extracting data from websites.
- Files: CSV, Excel, JSON, XML.

Data Extraction Tools

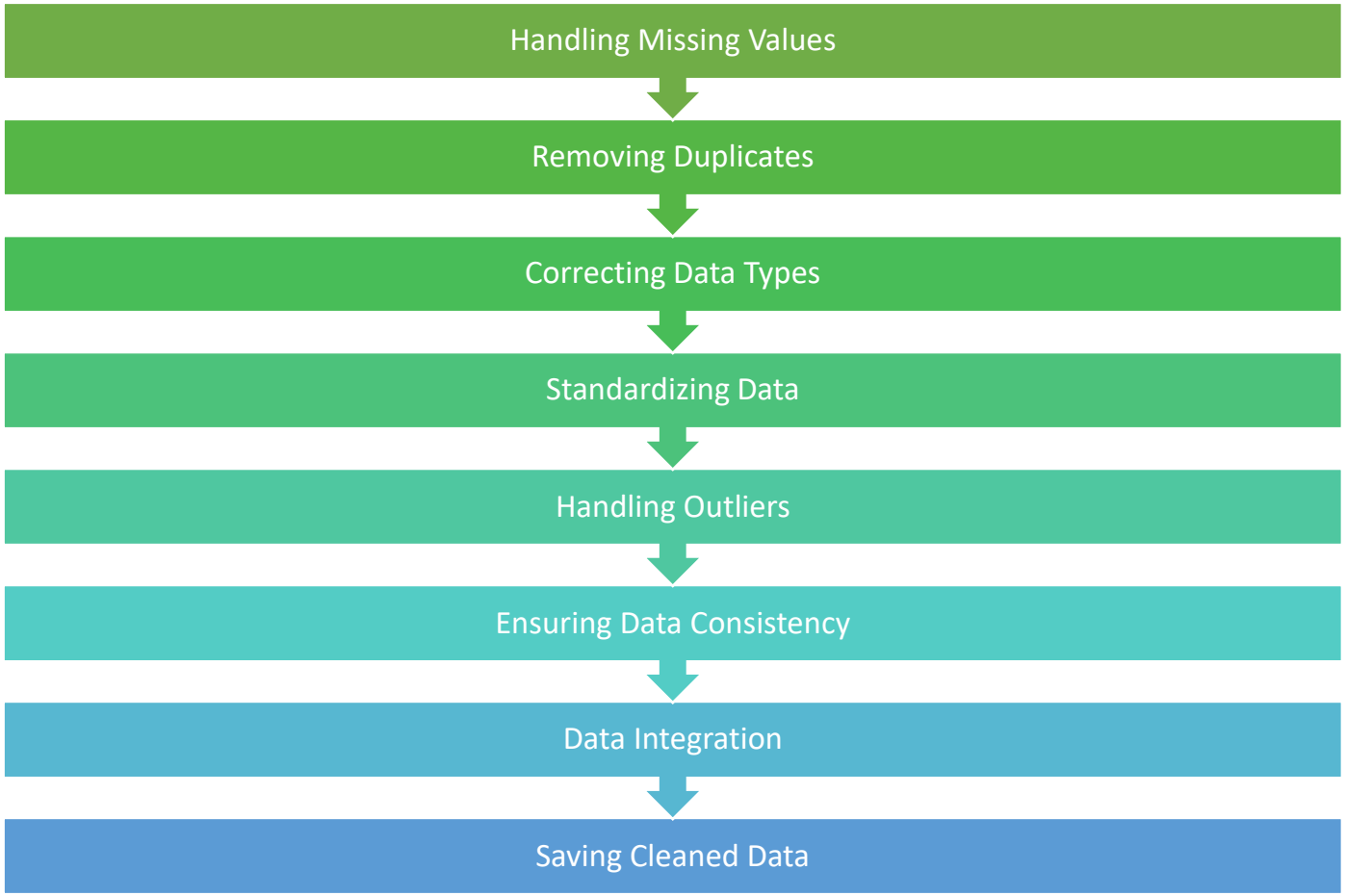
- SQL Queries: Used to fetch data from relational databases.
- API Clients: Tools like Postman or Python libraries (requests, BeautifulSoup) for API data.
- Web Scraping Libraries: BeautifulSoup, Scrapy, Selenium for scraping websites.
- Data Import Functions: Pandas functions like read_csv(), read_excel(), read_json() for importing data from files.

Automating Data Retrieval

- Scripts: Python scripts for automated data fetching.
- Scheduling Tools: Cron jobs, Airflow for scheduling and automating data retrieval processes.

Exploratory Data Analysis

Cleaning Data in Data Analytics



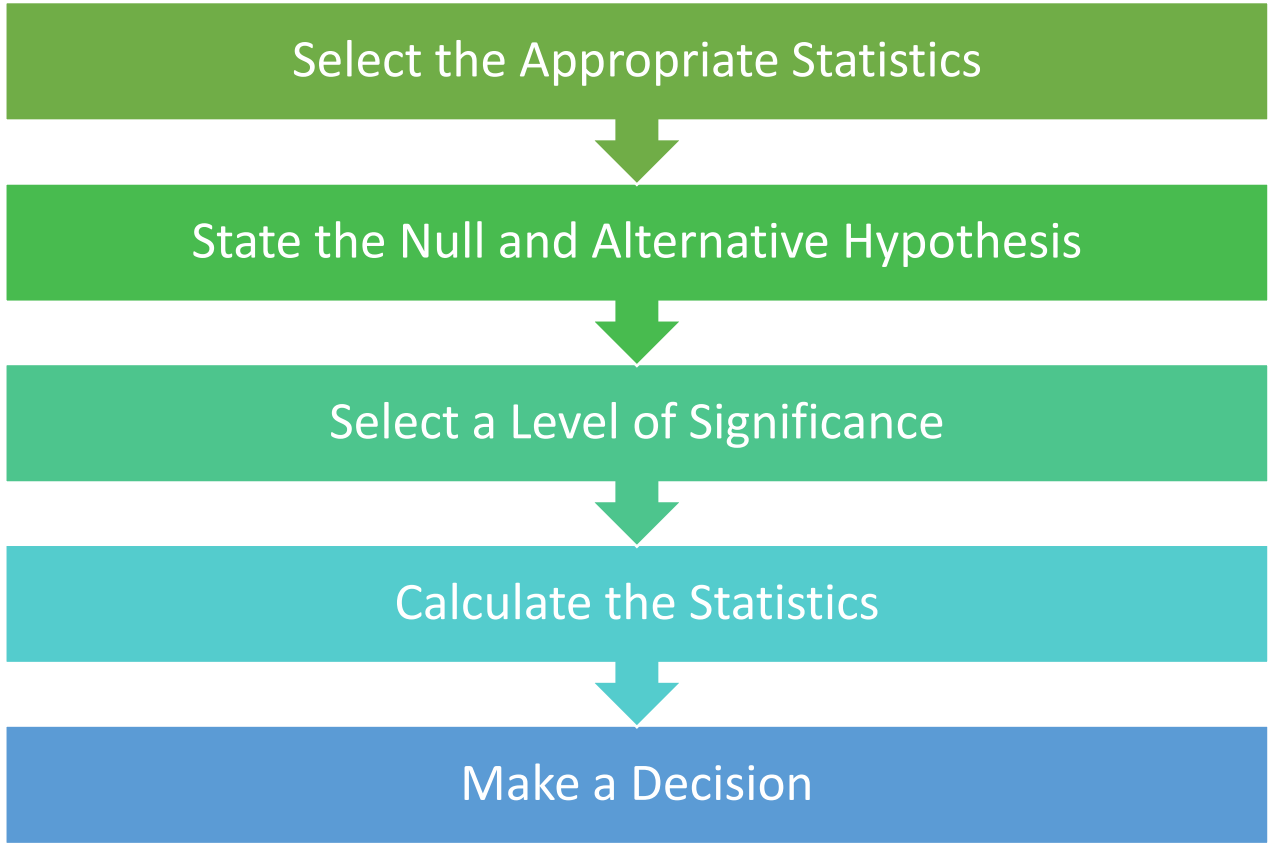
Exploratory Data Analysis

Hypothesis Testing in Inferential Statistics

- Hypothesis testing is an inferential statistical method.
- It uses sample data to make decisions or solve assumptions about a population parameter.
- A population parameter is a characteristic that describes a population.
- Hypothesis testing helps determine if the sample data supports a specific assumption about the population.

Exploratory Data Analysis

Steps taken to conduct a Hypothesis Testing



Exploratory Data Analysis

Hypothesis Testing in Inferential Statistics

- Examples of its application include:
  - Estimating the average salary of domestic workers.
  - Determining if COVID-19 has less impact on people who take vitamin C.
  - Checking if a manufacturer consistently produces 50 liters of milk.
- It helps us answer questions about the population based on the information we gather from the sample.

Descriptive Statistics

Descriptive Statistics

Define Descriptive Statistics

- Descriptive statistics summarizes or describes the characteristics of a data set.
- Descriptive statistics consists of three basic categories of measures: measures of central tendency, measures of variability (or spread), and frequency distribution.
- Measures of central tendency describe the center of the data set (mean, median, mode).
- Measures of variability describe the dispersion of the data set (variance, standard deviation).
- Measures of frequency distribution describe the occurrence of data within the data set (count).

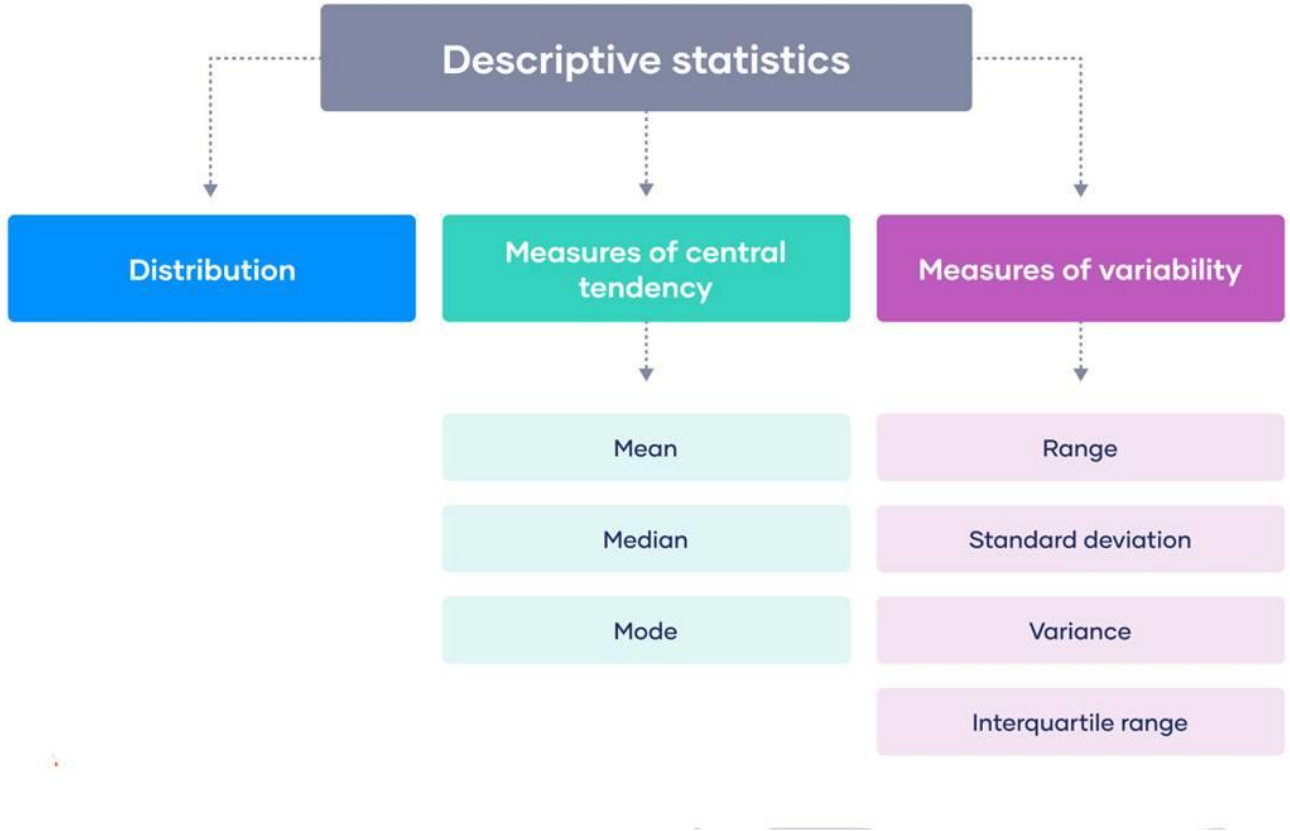
Descriptive Statistics

Define Descriptive Statistics

- Descriptive statistics summarizes or describes the characteristics of a data set.
- Descriptive statistics consists of three basic categories of measures: measures of central tendency, measures of variability (or spread), and frequency distribution.
- Measures of central tendency describe the center of the data set (mean, median, mode).
- Measures of variability describe the dispersion of the data set (variance, standard deviation).
- Measures of frequency distribution describe the occurrence of data within the data set (count).

Descriptive Statistics

Types of Descriptive Statistics



Descriptive Statistics

Distribution

- Distribution (or frequency distribution) refers to the quantity of times a data point occurs. Alternatively, it is the measurement of a data point failing to occur. Consider a data set: male, male, female, female, female, other. The distribution of this data can be classified as:
- The number of males in the data set is 2.
- The number of females in the data set is 3.
- The number of individuals identifying as other is 1.
- The number of non-males is 4.

Descriptive Statistics

Central Tendency

- Measures of central tendency focus on the average or middle values of data sets.
- Measures of variability focus on the dispersion of data.
- Both measures use graphs, tables, and discussions to help people understand the analyzed data.
- Measures of central tendency describe the center position of a data set.
- This involves analyzing the frequency of each data point in the distribution.
- The center is described using the mean, median, or mode.
- These measures highlight the most common patterns in the data set.

Descriptive Statistics

Measures of Variability

- Measures of variability (or spread) analyze how dispersed the data is.
- While measures of central tendency give the average, they don't show data distribution.
- For example, the average might be 65, but data points can range from 1 to 100.
- Measures of variability describe the shape and spread of the data set.
- Examples include range, quartiles, absolute deviation, and variance.
- Consider the data set: 5, 19, 24, 62, 91, 100.
- The range is 95, calculated by subtracting the lowest number (5) from the highest (100).

Descriptive Statistics

Population

- Population refers to the complete set of individuals or items of interest in a study.
- It can include people, animals, plants, objects, or other groupings.
- For example, if a researcher studies the dietary patterns of all adults in a country, the population is all adults in that country.

Descriptive Statistics

Ways of Collecting Data From a Population

- Collecting data from a whole population can be difficult, especially if it's large or spread out. Here are some methods researchers use:
- **Census:**
  - Collect data from every individual or item in the population.
  - Provides accurate and comprehensive information.
  - Time-consuming, costly, and impractical for large populations.

Descriptive Statistics

Ways of Collecting Data From a Population

- **Administrative Data:**
  - Use existing records or databases from government agencies, organizations, or institutions.
  - Examples include census data, tax records, healthcare records, and educational records.
- **Surveys:**
  - Use questionnaires or interviews with a representative sample or the entire population if feasible.
  - Gather information directly from individuals about their opinions, behaviors, and characteristics.

Descriptive Statistics

Ways of Collecting Data From a Population

- **Direct Observation:**
  - Observe and record information about individuals or items first hand.
  - Common in anthropology, ecology, and sociology to observe natural behaviors in their environments.

- **Remote Sensing:**
  - Use technologies like satellites, drones, or sensors to collect data about environmental characteristics or phenomena.
  - Useful for studying large geographic areas or inaccessible locations.

Descriptive Statistics

Ways of Collecting Data From a Population

- **Social Media and Web Data:**
  - Analyze data from social media platforms, websites, or online communities.
  - Understand behaviors, preferences, and interactions of digital populations.

- **Physical Measurements:**
  - Take physical measurements or samples from individuals or items.
  - Common in biology, medicine, and engineering to collect objective data on physical characteristics or properties.

Descriptive Statistics

What Is a Sample?

- A sample is a subset of individuals, items, or observations from a larger population.
- It represents the characteristics of the larger group.
- A sample is a smaller, manageable portion of the population.
- It is studied to make inferences about the entire population.

Descriptive Statistics

Reasons for Sampling

- **Practicality:**
  - Studying an entire population is often impractical due to time, cost, and logistics constraints.
  - Sampling allows for meaningful insights from a smaller, manageable subset.
- **Efficiency:**
  - Focuses resources on a subset of the population, saving time and resources.
  - Provides valuable information without needing data from every individual or item.
- **Generalizability:**
  - Proper sampling allows valid inferences about the entire population.
  - A representative sample lets researchers generalize findings with confidence.

Descriptive Statistics

Reasons for Sampling

- **Accuracy:**
  - Sampling methods aim to reduce bias and increase precision.
  - Techniques like randomization help ensure the sample accurately reflects the population, reducing skewed or erroneous results.
- **Ethical Considerations:**
  - Studying an entire population may be unethical or impractical, especially with sensitive topics or vulnerable groups.
  - Sampling minimizes potential harm and adheres to ethical guidelines while still providing valuable research.

Descriptive Statistics

What is a Variable?

- **Definition:** A variable is a characteristic, number, or quantity that can change or vary. In research and statistics, variables are used to represent data that can take on different values or attributes.

- **Key Points:**
- **Changeable:** Variables can vary among different individuals or over time.
- **Data Representation:** They are used to collect, analyze, and interpret data.
- **Types:** Variables can be classified into different types based on their nature and the kind of data they represent.

Descriptive Statistics

Types of Variables

- **Qualitative (Categorical) Variables :**
- Nominal: Categories without a specific order (e.g., gender, eye color).
- Ordinal: Categories with a specific order (e.g., education level, satisfaction rating).
- **Quantitative (Numerical) Variables:**
- **Discrete:** Countable values, often whole numbers (e.g., number of children, number of **cars).**
- **Continuous:** Any value within a range, including decimals (e.g., height, weight, temperature).
- **Binary (Dichotomous) Variables:**
- **Definition**: Variables with only two possible values (e.g., yes/no, true/false).

Descriptive Statistics

Importance of Variables:

- **Measurement and Analysis:** Variables are essential for measuring and analyzing different aspects of research subjects.
- **Data Collection:** They help in organizing and categorizing data for better understanding and interpretation.
- **Hypothesis Testing:** Variables are crucial in testing hypotheses and drawing conclusions in research studies.

Descriptive Statistics

Excel for Data Analysis

- Data analysis helps make better judgments.
- Microsoft Excel is widely used for data analysis.
- Pivot tables are Excel's most popular analytic tool.
- Excel provides a user-friendly platform for organizing and interpreting data.
- Mastering Excel for data analysis enhances your ability to derive insights and make strategic decisions.
- Excel allows you to examine and interpret data from various sources.

Descriptive Statistics

Excel for Data Analysis

Key Excel tools for data analysis include:

- Conditional Formatting
- Ranges
- Tables
- Text functions
- Date functions
- Time functions
- Financial functions
- Subtotals
- Quick Analysis
- Formula Auditing
- Inquire Tool
- What-if Analysis
- Solvers
- Data Model
- PowerPivot
- PowerView
- PowerMap

Descriptive Statistics

Types of Data Analysis With Microsoft Excel

- **Descriptive Analysis**
  - **Summary Statistics:** Calculate the mean, median, mode, standard deviation, and range using functions like AVERAGE, MEDIAN, MODE, STDEV, and MAX-MIN.
  - **Data Visualization:** Create charts and graphs (e.g., bar charts, histograms, pie charts) to visualize data distributions and trends.
  - **PivotTables:** Summarize large datasets by creating PivotTables to calculate totals, averages, and other summary statistics quickly.

Descriptive Statistics

Types of Data Analysis With Microsoft Excel

- **Exploratory Data Analysis**
- **EDA involves analyzing data sets to find patterns, relationships, and anomalies.**
  - **Sorting and Filtering:** Sort data and apply filters to explore subsets of data.
  - **Conditional Formatting:** Highlight patterns and outliers using conditional formatting rules.
  - **Scatter Plots:** Create scatter plots to explore relationships between two numerical variables.
  - **Box Plots:** Use box plots to visualize the distribution and identify outliers.

Descriptive Statistics

Types of Data Analysis With Microsoft Excel

- **Inferential Analysis**
- **Inferential analysis involves making inferences and predictions about a population based on a sample of data.**
  - **Hypothesis Testing:** Use the Analysis ToolPak to perform t-tests, ANOVA, chi-square tests, and other statistical tests.
  - **Confidence Intervals:** Calculate confidence intervals to estimate population parameters.

Descriptive Statistics

Types of Data Analysis With Microsoft Excel

- **Predictive Analysis**
- **Predictive analysis uses historical data to predict future outcomes.**
  - **Regression Analysis:** Use the Analysis Tool Pak to perform linear and multiple regression analysis to predict the value of a dependent variable based on one or more independent variables.
  - **Trend Analysis:** Use trendlines in charts to identify trends and make forecasts.

Descriptive Statistics

Types of Data Analysis With Microsoft Excel

- **Prescriptive Analysis**
- **Prescriptive analysis provides recommendations for actions based on data.**
  - **What-If Analysis:** Use tools like Scenario Manager, Goal Seek, and Data Tables to explore different scenarios and their outcomes.
  - **Solver:** Optimize decision-making by finding the best solution to a problem with constraints using Solver.

Descriptive Statistics

Types of Data Analysis With Microsoft Excel

- **Diagnostic Analysis**
- **Diagnostic analysis aims to determine why something happened by identifying causes and correlations.**
  - **Correlation Analysis:** Calculate correlation coefficients to measure the strength and direction of relationships between variables.
  - **Drill-Down Analysis:** Use PivotTables to drill down into data for more detailed analysis.

Missing Data - What can be expected, what are you losing

Missing Data - What can be expected, what are you loosing

Why is it important to handle missing data?

- Real-world data often has missing values.
- Missing data can be due to loss, corruption, or specific reasons.
- Missing data reduces the predictive power of your model.
- Applying algorithms with missing data introduces bias in parameter estimation.
- Handling missing data is essential to ensure confidence in your results.

Missing Data - What can be expected, what are you loosing

Types of Missing Values

- Missing data can occur for various reasons and are categorized into three main types: Missing Completely at Random (MCAR), Missing At Random (MAR), and Not Missing at Random (NMAR).

Missing Data - What can be expected, what are you loosing

Types of Missing Values

- **Missing Completely at Random (MCAR):**
- Missing data has no pattern and is independent of other variables.
- Example: Data lost due to carelessness during collection.
- Analysis is not biased if data is MCAR, but this situation is rare and should not be assumed without strong evidence.

Missing Data - What can be expected, what are you loosing

Types of Missing Values

- **Missing At Random (MAR):**
- Missing data occurs within specific subsets and can be predicted by other variables.
- Example: Older people (above 45) are less likely to answer questions about time spent on Netflix in a survey.
- The 'Age' variable predicts the presence/absence of data, but not the missing values themselves.
- MAR is more common than MCAR.

Missing Data - What can be expected, what are you loosing

Types of Missing Values

- **Not Missing at Random (NMAR):**
- Missing data is related to the unobserved value itself, leading to potential bias.Example: In a survey on social media addiction, those who are addicted may intentionally not respond.Common methods like dropping rows/columns or imputation won't work.Requires deep domain knowledge to address properly.

Summary

Data Profiling & Visual Representation via various tools (Pandas)

Data Profiling and Visualization	Understanding data profiling as an initial step in data analysis. Visualization techniques to present insights from data effectively.
Data Analysis (EDA) with Pandas	Using Pandas library in Python for initial data analysis. Steps include loading data, summarizing main characteristics, and visualizing data.
Exploratory Data Analysis (EDA)	Data cleaning to handle missing values and outliers. Univariate and bivariate analysis to understand distributions and relationships.

Data Profiling & Visual Representation via various tools (Pandas)

Market Analysis with Exploratory Data Analysis	Applying EDA techniques to understand market trends and consumer behavior.
Data Analysis (EDA) with Pandas	Overview of data analytics as a field leveraging data for insights.
Top Business Intelligence Tools	Their capabilities in data visualization, dashboards, and interactive reporting.

Data Profiling & Visual Representation via various tools (Pandas)

Application of Data Analytics	Practical applications across industries such as finance, healthcare, and marketing.
Retrieving and Cleaning Data	Techniques for data extraction from various sources like databases and APIs.
Exploratory Data Analysis and Feature Engineering	Importance of feature engineering in enhancing predictive models.

Data Profiling & Visual Representation via various tools (Pandas)

Inferential Statistics and Hypothesis Testing	Using statistical methods to make inferences about populations based on samples.
Descriptive Statistics	Includes measures like mean, median, mode, and variability measures (standard deviation, variance).
Types of Descriptive Statistics	Different types include measures of central tendency (mean, median, mode) and measures of dispersion (range, variance, standard deviation).

Data Profiling & Visual Representation via various tools (Pandas)

Statistical Methods for Describing Data Characteristics	Techniques such as frequency distributions, percentiles, and correlation coefficients.
Real-World Applications of Descriptive Statistics using Excel	Using Excel for basic statistical analysis like mean, median, and standard deviation.
Types of Missing Data and Handling Techniques	Types include missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).

Knowledge Check

Data Profiling & Visual Representation via various tools (Pandas)

Q1 : What is the primary purpose of data profiling in data analysis?

A. To summarize data using statistical methods

B. To clean and preprocess data

C. To understand the structure and quality of data

D. To build predictive models

Data Profiling & Visual Representation via various tools (Pandas)

Q2 : Which library is commonly used for performing EDA in Python?

A. Matplotlib

B. Pandas

C. Scikit-learn

D. TensorFlow

Data Profiling & Visual Representation via various tools (Pandas)

Q3 : Which of the following is NOT a step typically involved in EDA?

A. Data cleaning

B. Hypothesis testing

C. Data visualization

D. Summary statistics

Data Profiling & Visual Representation via various tools (Pandas)

Q4 : In market analysis using EDA, what type of analysis helps identify trends and patterns in consumer behavior?

A. Inferential statistics

B. Descriptive statistics

C. Time series analysis

D. Predictive modeling

Data Profiling & Visual Representation via various tools (Pandas)

Q5 : What is a key future trend in data analytics?

A. Decreasing reliance on machine learning

B. Increased use of static reports

C. Integration of artificial intelligence

D. Reduction in data collection

Data Profiling & Visual Representation via various tools (Pandas)

Q6 : Which Python library is widely used for machine learning tasks in data analytics?

A. Pandas

B. NumPy

C. Matplotlib

D. Scikit-learn

Data Profiling & Visual Representation via various tools (Pandas)

Q7 : What is a primary feature of business intelligence (BI) tools like Tableau and Power BI?

A. Statistical hypothesis testing

B. Data extraction from APIs

C. Interactive data visualization

D. Natural language processing

Data Profiling & Visual Representation via various tools (Pandas)

Q8 : Which industry commonly uses predictive modeling and segmentation techniques based on customer data?

A. Agriculture

B. Retail

C. Construction

D. Hospitality

Data Profiling & Visual Representation via various tools (Pandas)

Q9 : What is the purpose of data cleaning in the data analysis process?

A. To reduce the size of the dataset

B. To enhance data quality and consistency

C. To perform data visualization

D. To build predictive models

Data Profiling & Visual Representation via various tools (Pandas)

Q10 : Which measure of central tendency is affected most by outliers?

A. Mean

B. Median

C. Mode

D. Range

?

Thank You !!!