Data Science with Python

Data Science Overview

Learning Objectives

By the end of this lesson, you will be able to:

- Describe the process of data analytics and its steps
- List the skills and tools required for data analysis
- Understand the challenges of data analytics
- Explain exploratory data analysis technique
- Illustrate data visualization techniques
- Describe hypothesis testing

DATA AND
ARTIFICIAL INTELLIGENCE

Data Science

Why Data Science?

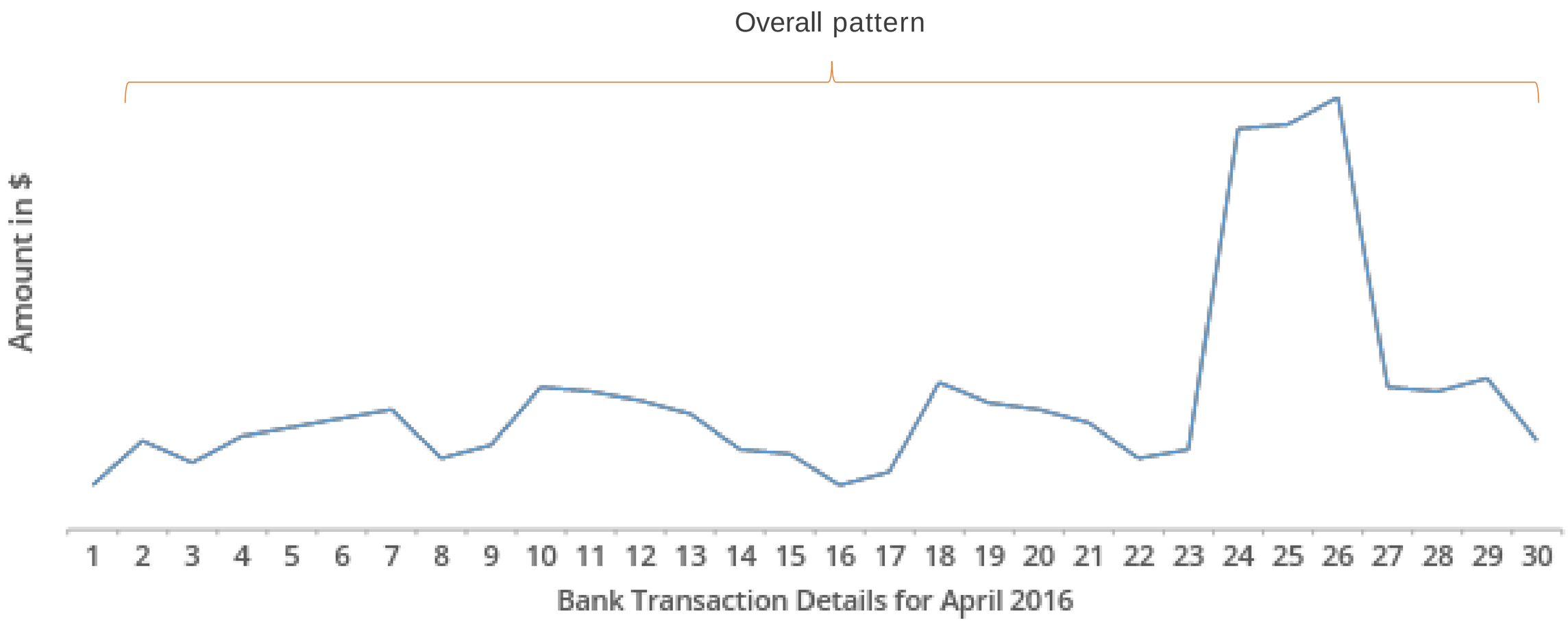
Data, by itself, is just an information source. Unless you understand it, you will not be able to use it effectively.

Date	Description	Deposit	Withdrawal	Balance
Apr 1	ATM Post Debit		100	$200,000
Apr 2	PayPal Transfer 231054	200		$202,000
Apr 3	Simplilearn course fee		150	$200,500
Apr 4	Starbucks Café		210	$198,400
Apr 5	Walmart TX		230	$196,100
Apr 6	eBay swiss watch 239		250	$193,600
Apr 7	Caterpillar black boots men		270	$190,900
Apr 8	Halo blue shirt 831		160	$189,300

Overall patterns not clearly visible

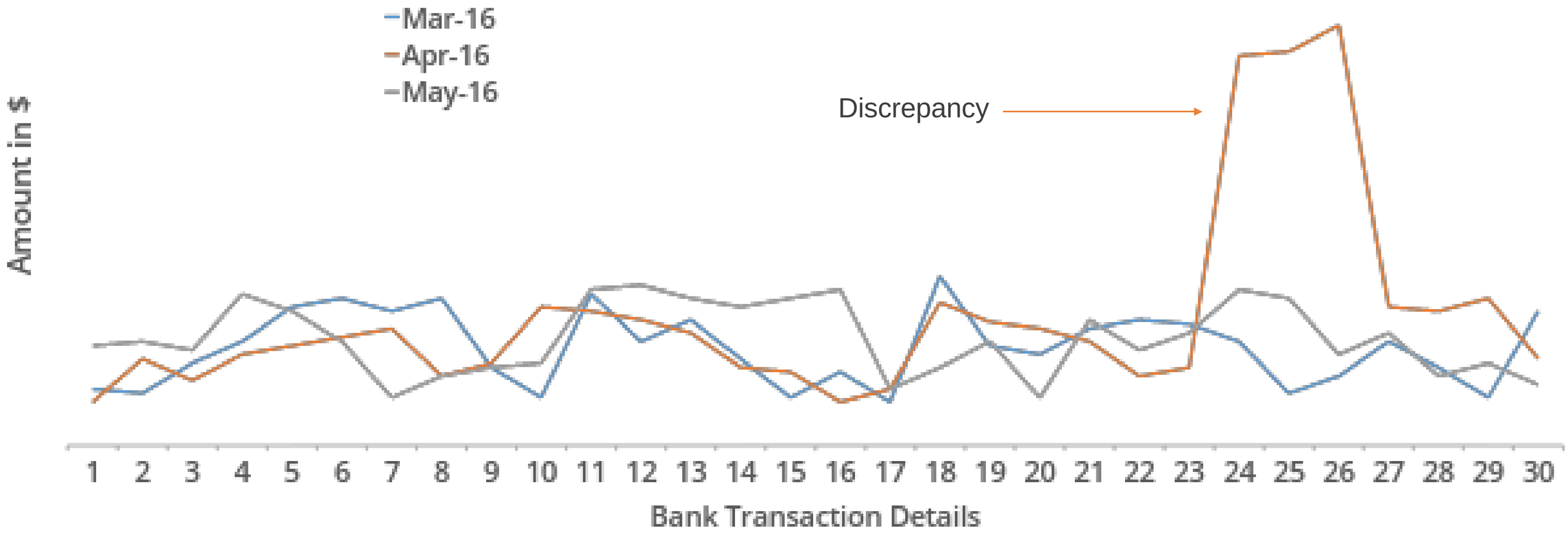
Why Data Science?

When the transaction details are presented as a line chart, the deposit and withdrawal patterns become apparent.



Why Data Science?

When the transaction details are presented as a line chart, the deposit and withdrawal patterns become apparent. It helps view and analyze general trends and discrepancies.



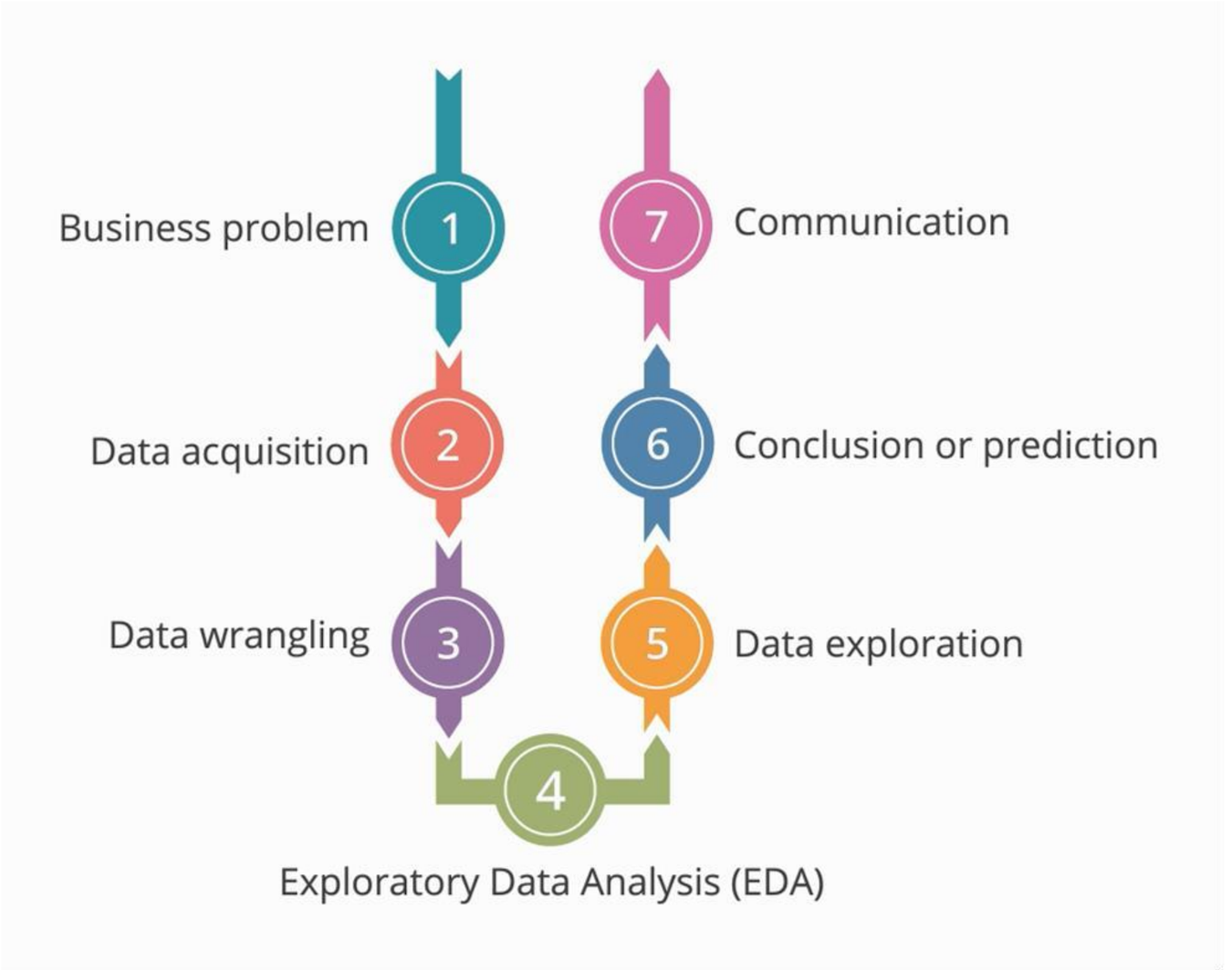
Introduction to Data Science

Data analytics is a combination of processes to extract information from datasets.

DATA AND
ARTIFICIAL INTELLIGENCE

Data Science Process

Data Science Process



Data Science: Skills and Tools

Question or Business Problem

- Ability to ask appropriate questions and know the business
- Domain knowledge
- Passion for data
- Analytical approach

Data Acquisition

- Beautiful Soup for web scraping
- CSV or other file knowledge
- NumPy
- Pandas
- Database

Data Wrangling

- CSV or other file knowledge
- NumPy
- Pandas
- Database
- SciPy

Data Exploration

- NumPy
- SciPy
- Pandas
- Matplotlib

Conclusion or Predictions

- Scikit-Learn, the main machine learning library
- CSV or other file knowledge
- NumPy
- Pandas
- Database
- SciPy

Communication or Data Visualization

- Pandas
- Database
- Matplotlib
- PPT
- CSV or other file knowledge

Business Problem

The process of analytics begins with questions or business problems of the stakeholders.

Business problems trigger the need to analyze data and find answers.

Data Acquisition

Collect data from various sources for analysis to answer the question raised in step 1.

Data Scientist expertise:

- File handling
- Web scraping

Twitter, Facebook, LinkedIn, and other social media and information sites provide streaming APIs.

Server logs can be extracted from enterprise system servers to analyze and optimize application performance.

Data Wrangling and Exploration

Data wrangling is the most important phase of the data analytics process.

Data cleansing

Data manipulation

Data discovery

Data pattern

Data Wrangling

Data Exploration

Data Wrangling: Challenges

This phase includes data cleansing, data manipulation, data aggregation, data split, and reshaping of data.

Causes of challenges in the data wrangling phase:

- Unexpected data format
- Erroneous data
- Voluminous data to be manipulated
- Classifying data into linear or clustered
- Determining relationship between observation, feature, and response

Data wrangling is the most challenging phase and takes upto 70% of the Data Scientist’s time.

Data Exploration: Model Selection

Model selection is based on the overall data analysis process to draw conclusions and make accurate predictions.

Model selection

- Is based on the overall data analysis process
- Should be accurate to avoid iterations
- Depends on pattern identification and algorithms
- Depends on hypothesis building and testing
- Leads to building mathematical statistical functions

Exploratory Data Analysis (EDA)

Approach

EDA approach studies the data to recommend suitable models that best fit the data.

Focus

The focus is on data and its structure, outliers, and models suggested by the data.

Assumptions

EDA techniques make minimal or no assumptions. They present all the underlying data without any data loss.

EDA Techniques

Quantitative:
Provides numeric outputs for the inputted data
Graphical:
Uses statistical functions for graphical output

EDA: Quantitative Technique

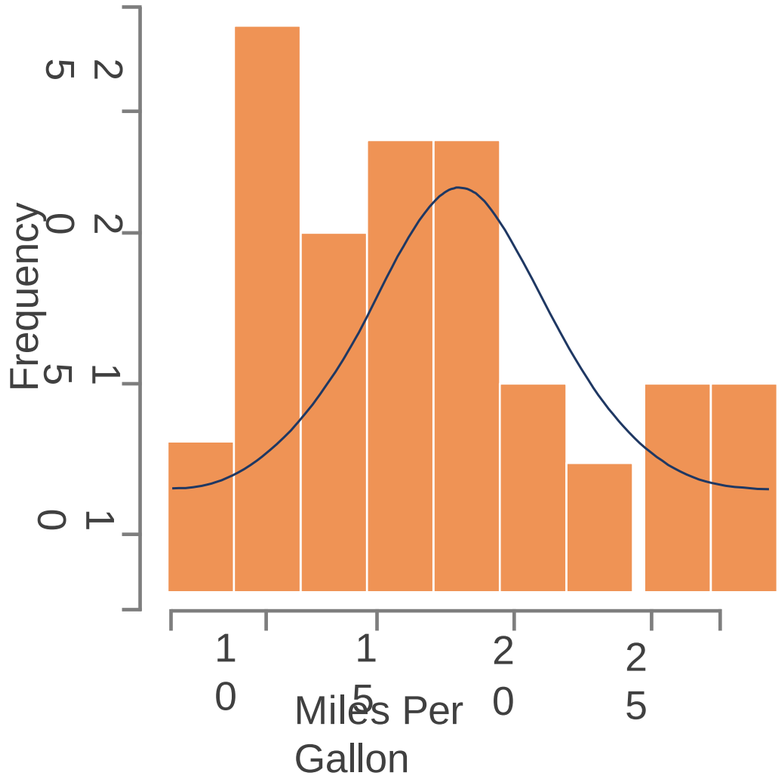
EDA quantitative technique has two goals: measurement of central tendency and spread of data.

	Measurement of Central Tendency
Mean	Mean is the point which indicates how centralized the data points are. • Suitable for symmetric distributions
Median	Median is the exact middle value. • Suitable for skewed distributions and for catching outliers in the dataset
Mode	Mode is the most common value in the data (frequency).

	Measurement of Spread
Variance	Variance is approximately the mean of the squares of the deviations.
Standard deviation	Standard deviation is the square root of the variance.
Interquartile range	Interquartile range is the distance between the 75^{th} and 25^{th} percentile. It's essentially the middle 50% of the data.

EDA: Graphical Technique

Histograms and scatter plots are two popular graphical techniques to depict data.

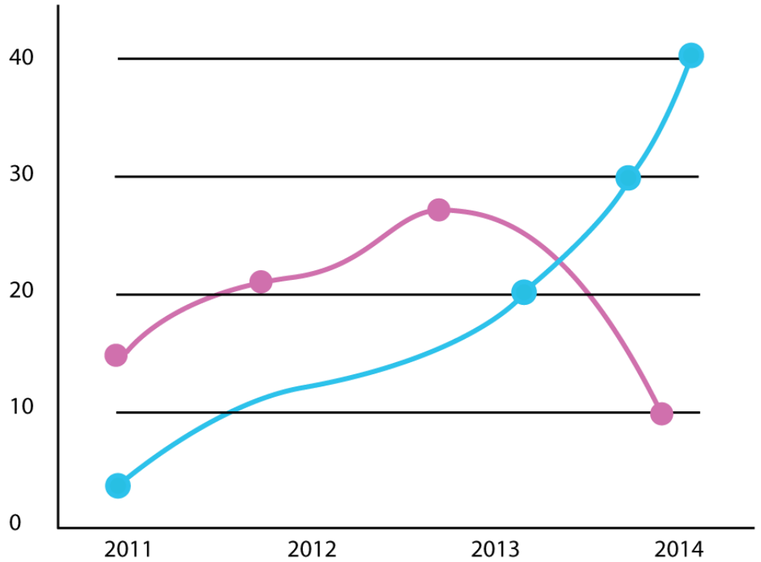


Histogram graphically summarizes the distribution of a univariate dataset.

It shows:

- the center or location of data (mean, median, or mode)
- the spread of data
- the skewness of data
- the presence of outliers
- the presence of multiple modes in the data

EDA: Graphical Technique



Scatter plot represents relationships between two variables.
It can answer these questions visually:

- Are variables X and Y related?
- Are variables X and Y linearly related?
- Are variables X and Y nonlinearly related?
- Does change in variation of Y depend on X?
- Are there outliers?

DATA AND
ARTIFICIAL INTELLIGENCE

Hypothesis

Conclusion or Prediction

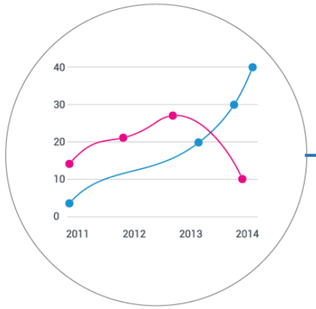
This step involves reaching a conclusion and making predictions based on the data analysis.

This phase:

- Involves heavy use of mathematical and statistical functions
- Requires model selection, training, and testing to help in forecasting
- Is called “machine learning” as data analysis is fully or semi-automated with minimal or no human intervention

Meaning of Hypothesis

Hypothesis is used to establish the relationship between dependent and independent variables.



Data Exploration Stage

Hypothesis building begins in the data exploration stage but becomes more mature in the conclusion or prediction phase.

Conclusion and Prediction

Key Considerations of Hypothesis Building
Testable explanations of a problem or observation
Used in quantitative and qualitative analyses to provide research solutions
Involves two variables, one dependent on another
Independent variable manipulated by the researcher
Dependent variable changes when the independent variable changes

Hypothesis Building Using Feature Engineering

Domain knowledge leads to hypothesis building using feature engineering.

Feature engineering involves domain expertise to:

- Make sense of data
- Construct new features from raw data automatically
- Construct new features from raw data manually

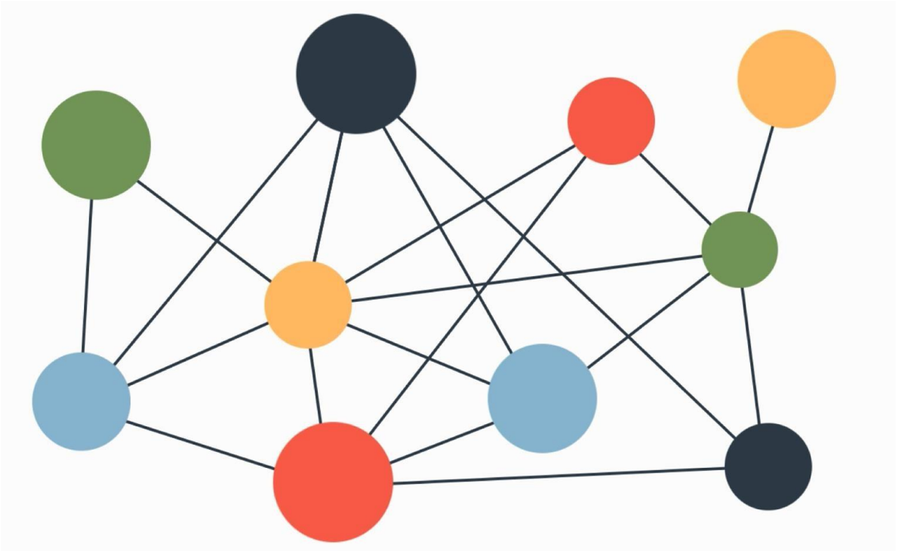
Hypothesis Building Using a Model

There are three phases to hypothesis building which are model building, model evaluation, and model deployment.

Phase 1: Model Building
- Identify best input variables
- Evaluate the model's capacity to forecast with these variables

Phase 2: Model Evaluation
- Train and test the model for accuracy
- Optimize model accuracy, performance, and comparison with other models

Phase 3: Model Deployment
- Use the model for prediction
- Use the model to compare actual outcome with expectations

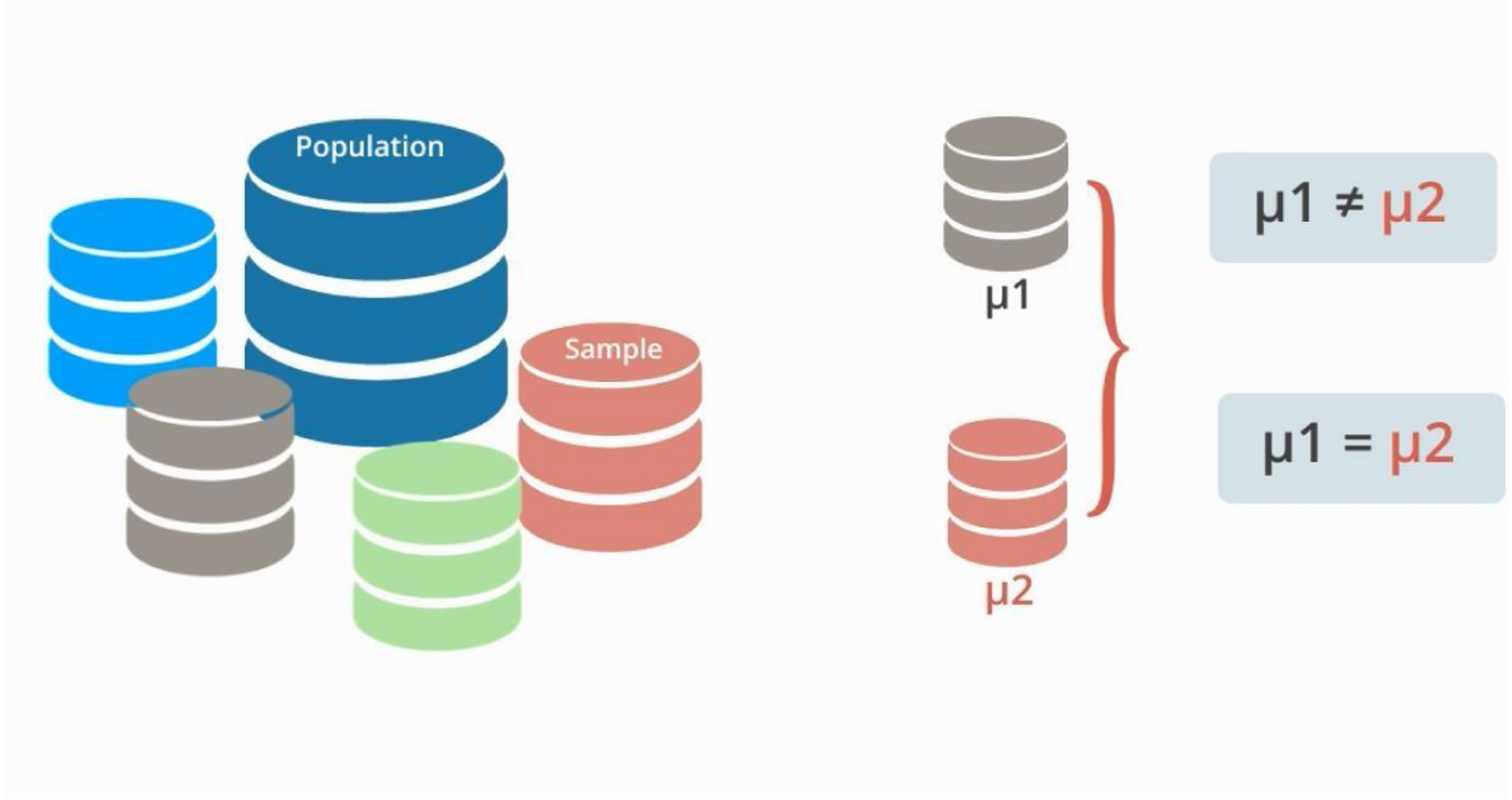


Hypothesis Testing

Draw two samples from the population and calculate the difference between their means.

Calculating the difference between the two means is called hypothesis testing.

Hypothesis Testing



Alternative Hypothesis

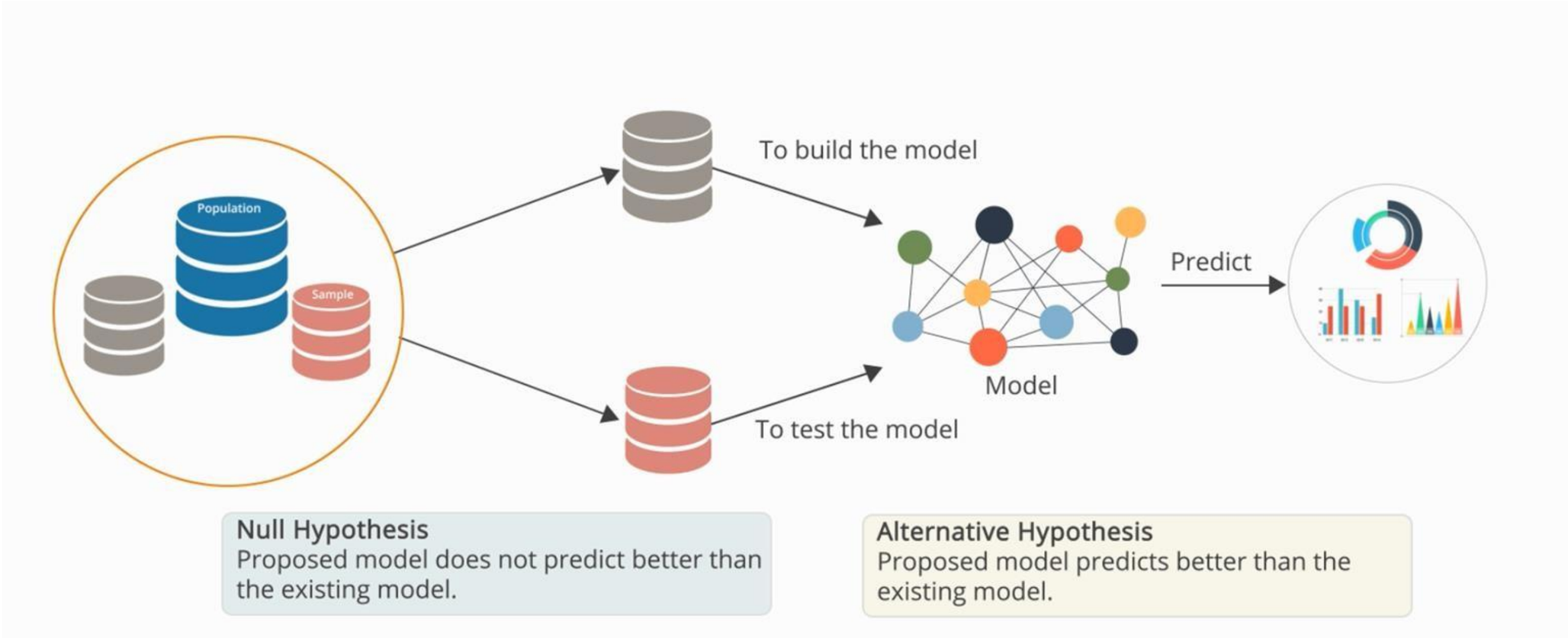
- Proposed model outcome is accurate and matches the data
- There is a difference between the means of S1 and S2

Null Hypothesis

- Opposite to the alternative hypothesis
- There is no difference between the means of S1 and S2

Hypothesis Testing Process

Choosing the training and test dataset and evaluating them with the null and alternative hypothesis.



Usually the training dataset is between 60% to 80% of the big dataset and the test dataset is between 20% to 40% of the big dataset.

DATA AND
ARTIFICIAL INTELLIGENCE

Data Visualization

Communication

The last step of data analysis is communication, where the analyzed data is formally presented to stakeholders.

Forms of data analysis presentations:

- Visual graphs
- Plotting maps
- Reports
- White paper reports
- PowerPoint presentations

Data Visualization

Data visualization techniques are used for effective communication of data.




Benefits of data visualization:

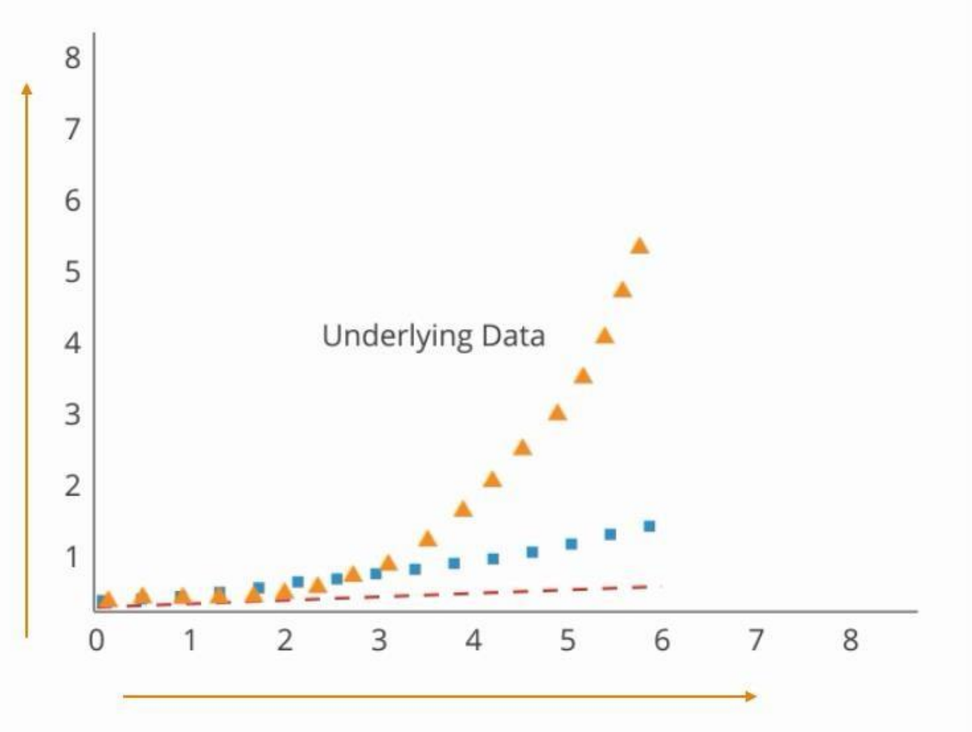
- Simplifies quantitative information through visuals
- Shows the relationship between data points and variables
- Identifies patterns
- Establishes trends

Examples of data visualization:

- Presenting information about new and existing customers on the website and their behavior when they access the website
- Representing web traffic pattern for the website, for example, more activity on the website in the morning than in the evening

Plotting

Plotting is a data visualization technique used to represent underlying data through graphics.



Features of plotting:

- Plotting is like telling a story about data using different colors, shapes, and sizes.
- Plotting shows the relationship between variables.
- Example:
  - Change in value of Y results in change in value of X
  - X is independent of y

Data Types for Plotting

Numerical Data

There are two types of numerical data:
Discrete Data: Distinct or counted values
Example: Number of employees in a company or number of students in a class

Continuous Data: Values within a range that can be measured
Example: Height can be measured in feet or inches and weight can be measured in pounds or kilograms

Categorical Data

There are two types of categorical data:
Cluster or group: Grouped values
Example: Students can be divided into different groups based on height: tall, medium, and short

Ordinal: Grouped values as per ranks
Example: A ranking system; a three-point scale with ranks like agree, strongly agree, and disagree

Time Series

Data is measured in time blocks such as date, month, year, and time (hours, minutes, and seconds)

Types of Plot

Different data types can be visualized using various plotting techniques.

DATA AND
ARTIFICIAL INTELLIGENCE

Knowledge Check

Knowledge Check 1

What is the goal of data acquisition?

Select all that apply.

a. Collect data from various data sources

b. Answer business questions through graphics

c. Collect web server logs

d. Scrape the web through web APIs

What is the goal of data acquisition?

Select all that apply.

a. Collect data from various data sources

b. Answer business questions through graphics

c. Collect web server logs

d. Scrape the web through web APIs

The correct answer is **a, c, d**

Data acquisition is a process to collect data from various data sources, such as RDBMS and NoSQL databases, collect web server logs, and also scrape the web through web APIs.

Knowledge Check 2

What is exploratory data analysis technique?

Select all that apply.

a. Analysis of data using quantitative techniques

b. Conducted only on a small subset of data

c. Analysis of data using graphical techniques

d. Suggests models that best fit the data

What is exploratory data analysis technique?

Select all that apply.

a. Analysis of data using quantitative techniques

b. Conducted only on a small subset of data

c. Analysis of data using graphical techniques

d. Suggests models that best fit the data

The correct answer is **a, c, d**

Most EDA techniques are graphical in nature with a few quantitative techniques and also, suggest models that best fit the data. They use almost the entire data with minimum and no assumptions.

Knowledge Check

3

Which plotting technique is used for continuous data?

Select all that apply.

a. Regression plot

b. Line chart

c. Histogram

d. Heat map

Which plotting technique is used for continuous data?

Select all that apply.

a. Regression plot

b. Line chart

c. Histogram

d. Heat map

The correct answer is **b, c**

Line charts and histograms are used to plot continuous data.

Knowledge Check 4

Which Python library is the main machine learning library?

a. Pandas

b. Matplotlib

c. Scikit-learn

d. NumPy

Which Python library is the main machine learning library?

a. Pandas

b. Matplotlib

c. Scikit-learn

d. NumPy

The correct answer is **c**

Scikit-learn is the main machine learning library in Python.

Knowledge Check 5

Which of the following includes data transformation, merging, aggregation, group by operation, and reshaping?

a. Data acquisition

b. Data visualization

c. Data wrangling

d. Machine learning

Which of the following includes data transformation, merging, aggregation, group by operation, and reshaping?

a. Data acquisition

b. Data visualization

c. Data wrangling

d. Machine learning

The correct answer is **c**

Data wrangling includes data transformation, merging, aggregation, group by operation, and reshaping.

Key Takeaways

You are now able to:

- Describe the process of data analytics and its steps
- List the skills and tools required for data analysis
- Understand the challenges of data analytics
- Explain exploratory data analysis technique
- Illustrate data visualization techniques
- Describe hypothesis testing

DATA AND
ARTIFICIAL INTELLIGENCE

Thank You