Introduction to Data Science

Introduction to Data Science Workflow

Session 1

From Introduction to Analysis

- Define the Business Goal
- Collect and manage data
  - Read the data
  - Pre-Processing
  - Data Visualization
- Demo – Churn analysis (part 1)
- Build the model – Introduction
  - Machine Learning;
  - Supervised and Unsupervised Learning;
  - Introduction to models (Classification, Regression, Cluster and Association)

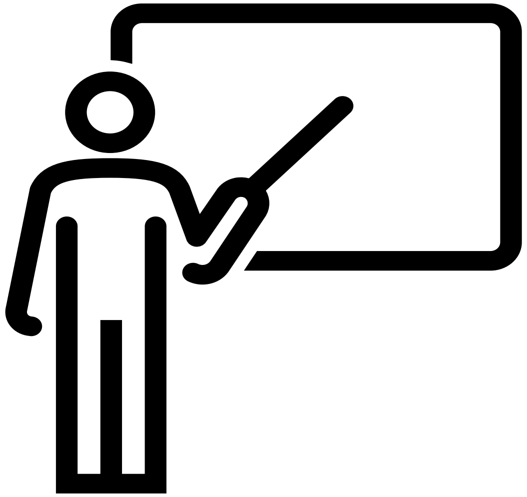
Session 2

Deep dive into Analysis

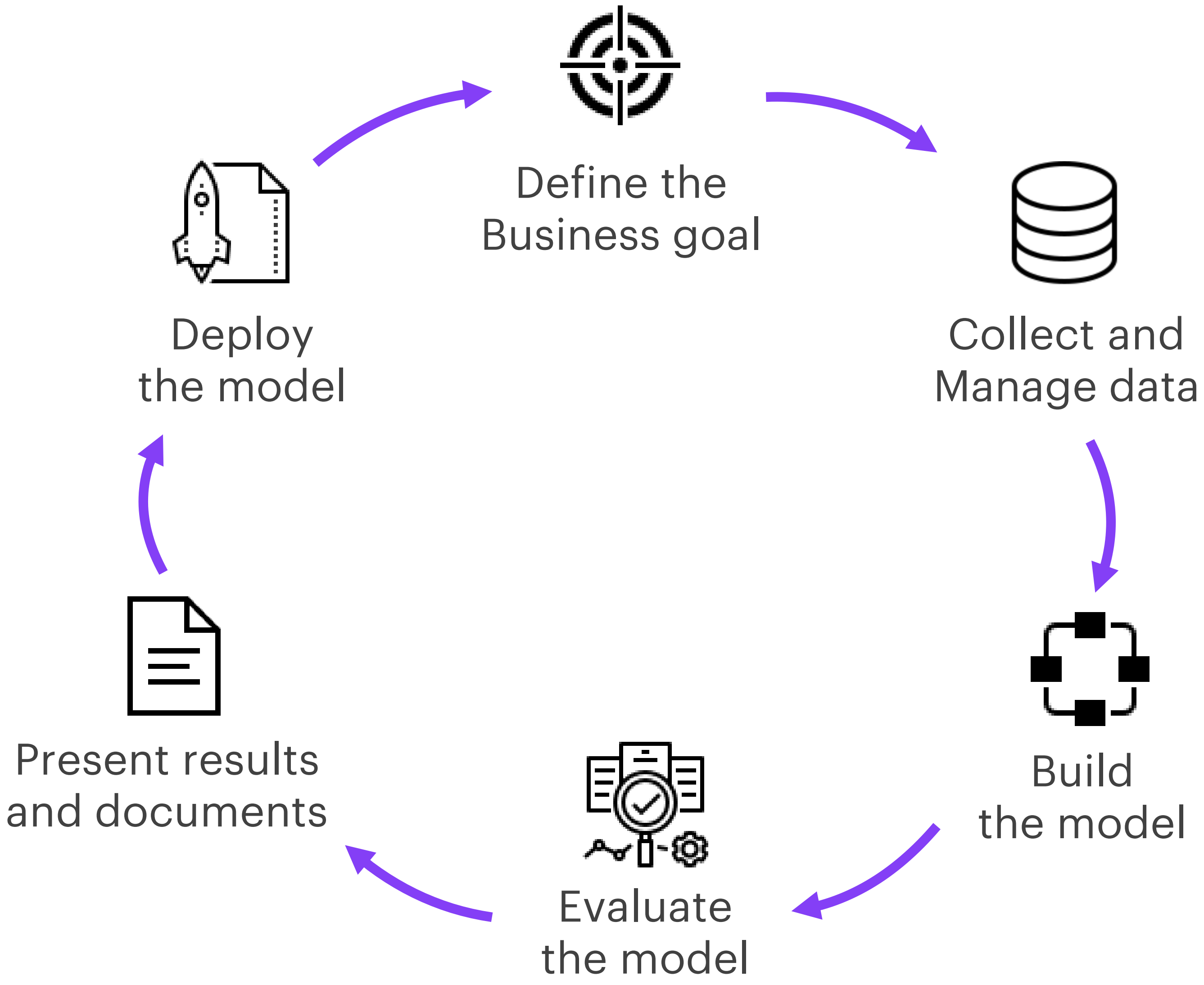
- Build the model – Deep dive
  - Training and test data
  - Linear Regression model
  - Classification models (Logistic Regression, Decision Tree and Random Forest)
  - Clustering models (k-Means, hierarchical clustering)
- Evaluate the model
- Present results and documents
- Demo – Churn analysis (part 2)
- Demo – Customer segmentation

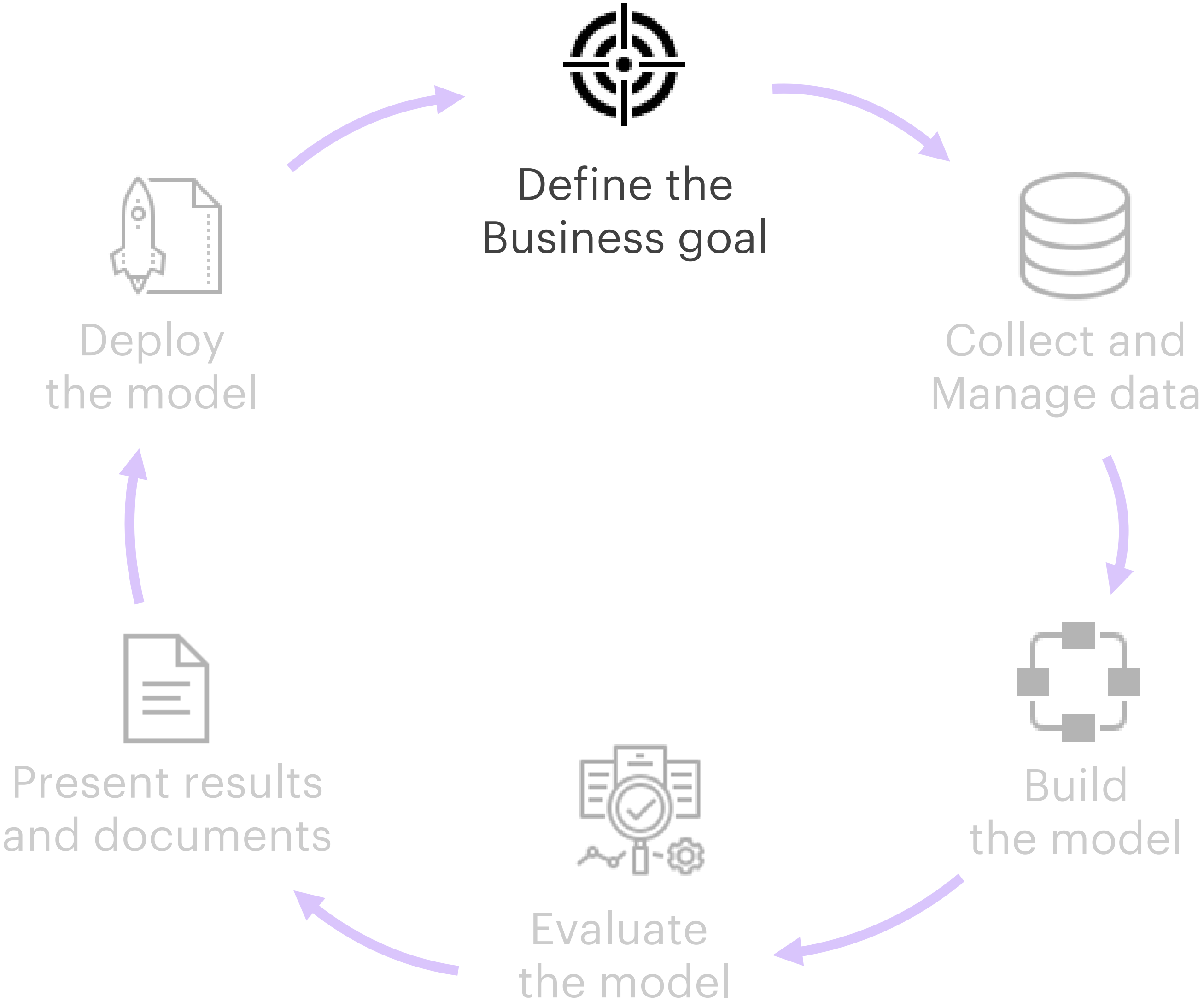
Session 1

The Data Science Workflow - From Introduction to Analysis



The Data Science Workflow





Session 1

Define the Business goal

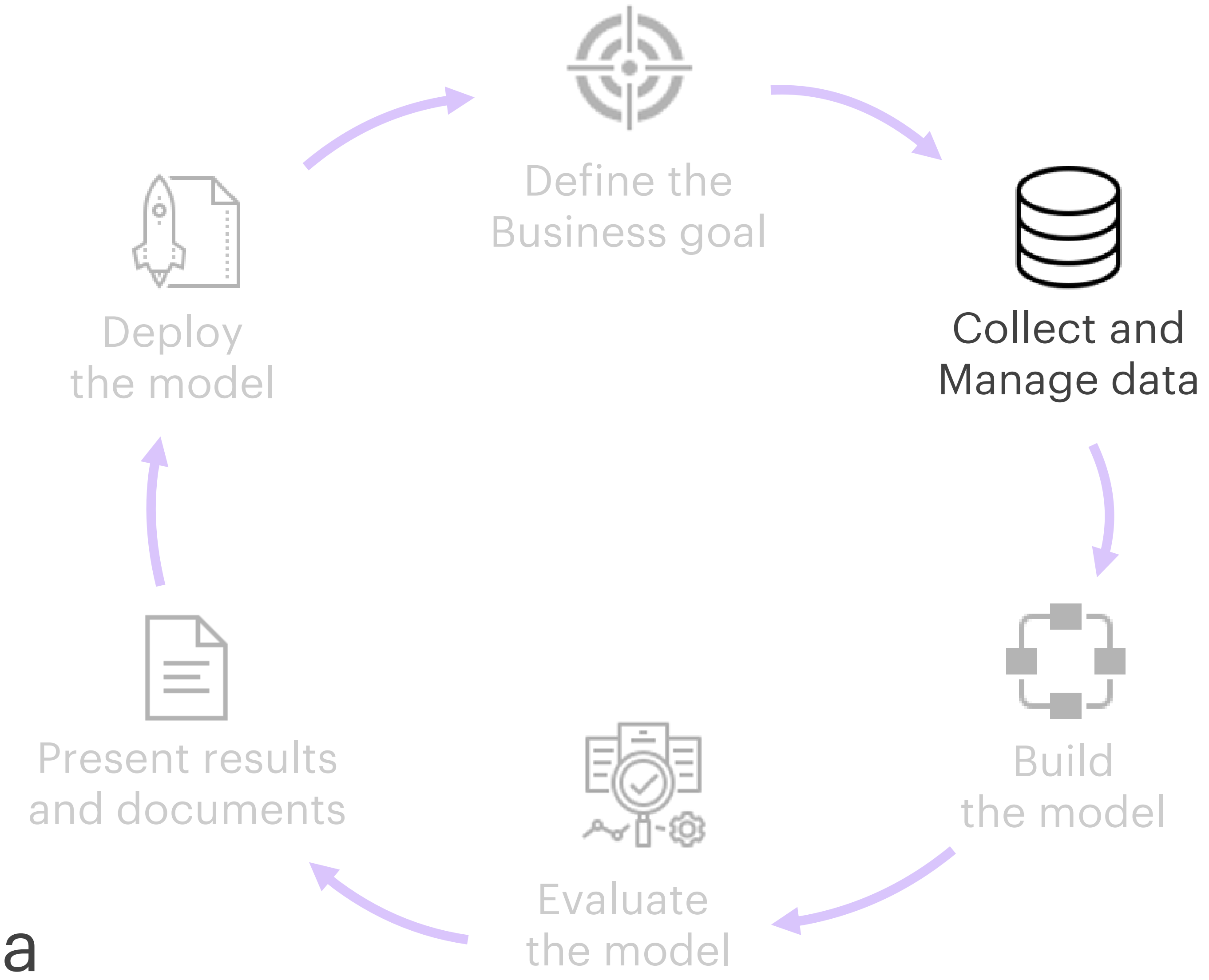
[Back to slide](#)

Business Goal

Understanding of how value and information flows in the business + Ability to identify business opportunities = Extraction of business-focused insights from data

Most common situations:

- Banking – risk classification
- Customer Segmentation
- Predict Customer Churn
- Time series – forecasting
- Language processing – entity relationship



Session 1

Collect and manage data

[Back to slide](#)

Collect and Manage data

Collect

- Read the data

Manage

- Pre-Processing
- Data Visualization

Collect and Manage data

Collect

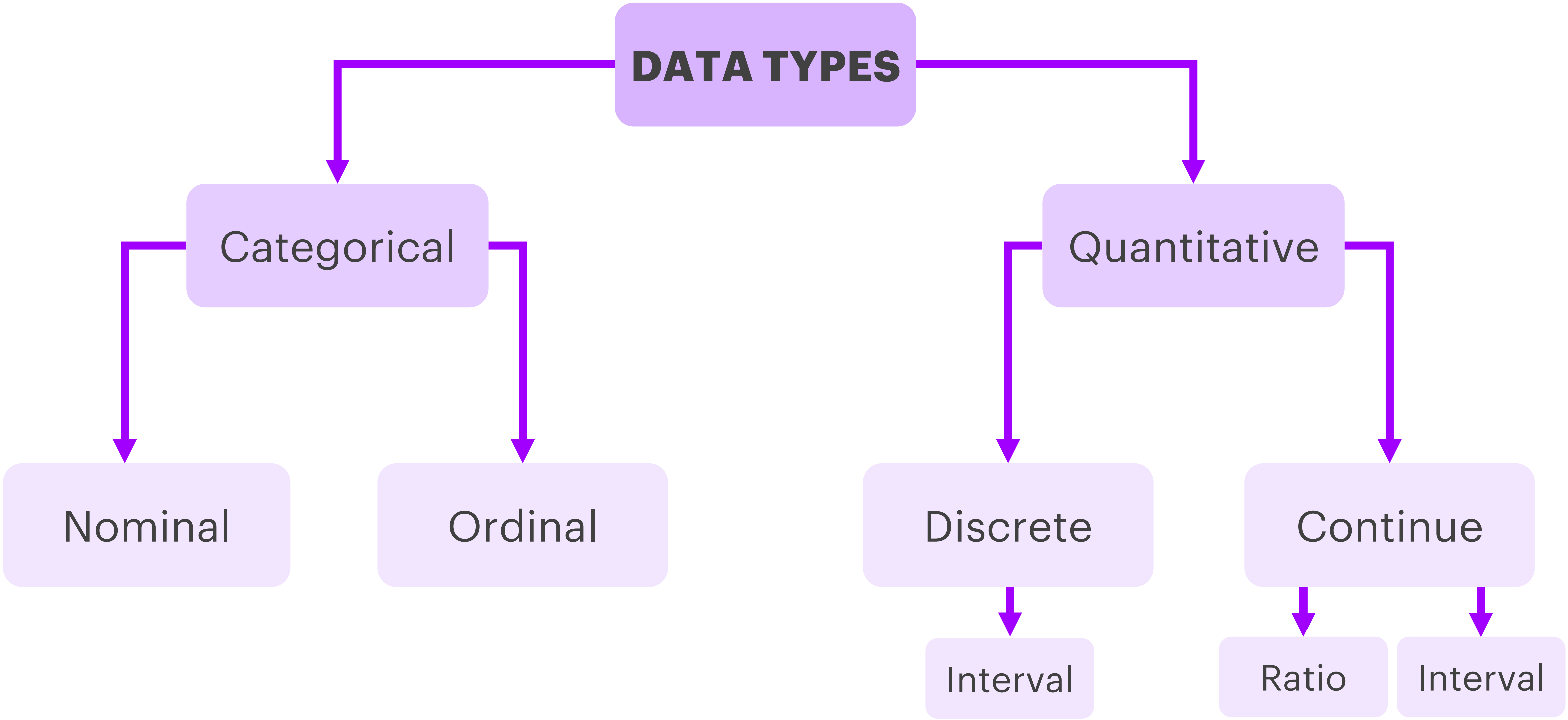
Manage

Read
the data

Pre-
Processing

Data
Visualization

Read the data



Categorical data types

Nominal

➢The values are a set of labels which don’t imply a quantity

➢**Examples:**
Gender, Nationality, Boolean, any identifier: e.g. patient (XXYXXX, XYXXXX), treatment (treatment 1, treatment 2)

➢**Operators: = and ≠**
when we compare the values we can only say if they are the same or different

Ordinal

➢The values are a set of labels which don’t imply a quantity, but imply a total/order relationship

➢**Examples:**
Bad, fair, good, excellent; the satisfaction

➢**Operators: = and ≠; <> and ≤≥**
when we compare the values we can also say if they are greater or less

Quantitative data types

Discrete

- A variable is defined discrete if it's possible to attribute an integer number to its values

- **Examples:** number of sons in a family, number of hats in a wordrobe, etc…

Continue

- A variable is defined continue if it's possible to attribute a real number to its values

- **Examples:**
Weight, Age..

Collect and Manage data

Collect | **Manage**

Read the data

Pre-Processing

Data Visualization

Pre-processing

Any type of processing performed to transform or encode raw data

- **Goal**: prepare it for another data processing procedure
- Pre-processing is a collection of techniques that allows to transform and prepare the data:
  - Handling missing values
  - Transforming Data
  - Centering and Scaling

Handling missing values

Why missing data

- Data were not collected
- The information is not applicable to that specific row

How to manage missing data

- Drop the value
- Imputing values (mean, median, mode)

Back to slide

Handling missing values

Dropping missing data

- **Pros**: we don't have to impute values so we analyze only real data
- **Cons**: we cannot drop the exclude the missing value only but the entire row… We lose information

Best option when there are only a few missing values

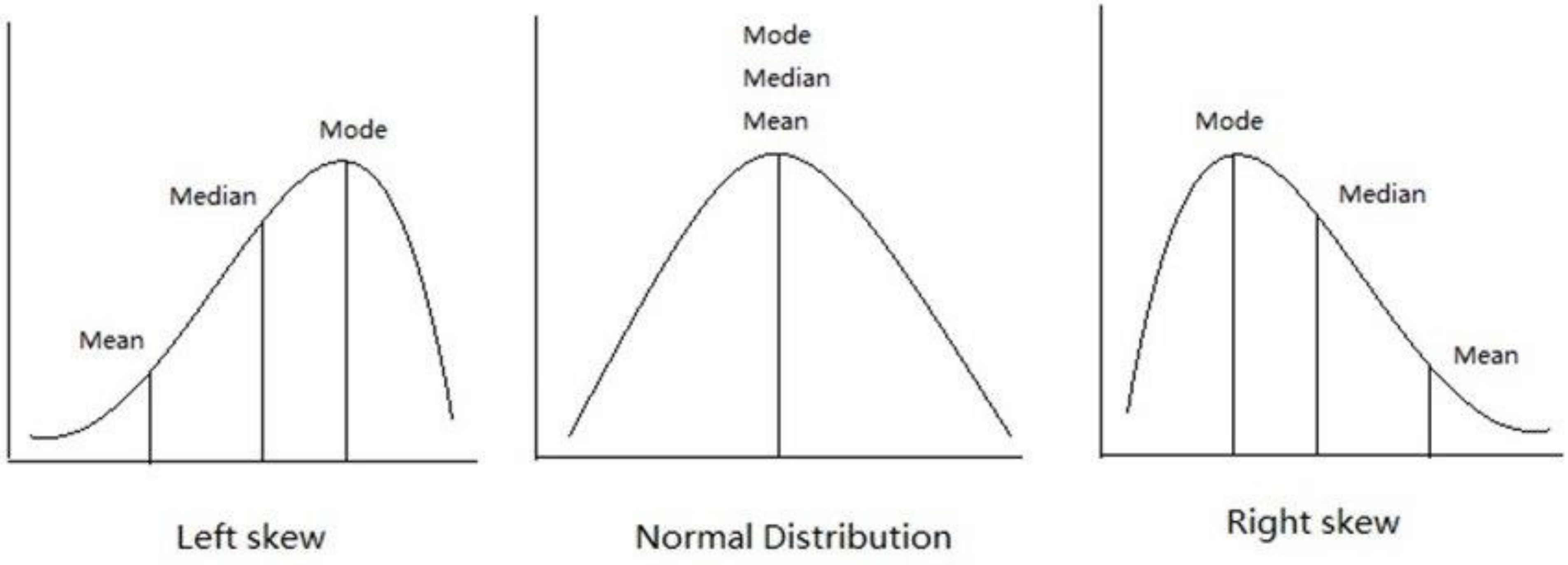
Handling missing values

Imputing values: it introduces some noise but allows not to lose info

- **mean:** quantitative variables, when the mean is representative of the sample
- **median:** quantitative variables, when there are outliers
- **mode:** categorical variables

Handling missing values

Imputing values



Transforming Data

The dummy variable

- For One hot Encoding is used a variable called dummy variable
- It is a binary variable that takes the value 0 or 1 to indicate the absence or presence of the categorical effect indicated by the feature

Back to slide

Transforming Data

Nominal to numeric (One hot Encoding)

ID	Color
1	Red
2	Green
3	Blue
4	Green

ID	Color_Red	Color_Green	Color_Blue
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

Transforming Data

➢ **Ordinal to numeric (Ordinal Encoding)**

ID	satisfaction_degreee
1	Poor
2	Fair
3	Good
4	Excelent

ID	satisfaction_degreee
1	1
2	2
3	3
4	4

Centering and Scaling

Normalization

Standardization

[Back to slide]

Centering and Scaling

- **Standardization**: allows to make variables comparable, because all features are centered around 0 and have variance 1.

$$Z_{stand} = \frac{x \; - mean(x)}{std(x)}$$

- **Normalization:** adjusts the values measured on different scales. The data points are shifted and rescaled so that they end up in a range of 0 to 1.

$$Z_{norm} = \frac{x \; - min(x)}{\max(x) - \min(x)}$$

Collect and Manage data

Collect | **Manage**

Read the data

Pre-Processing

Data Visualization

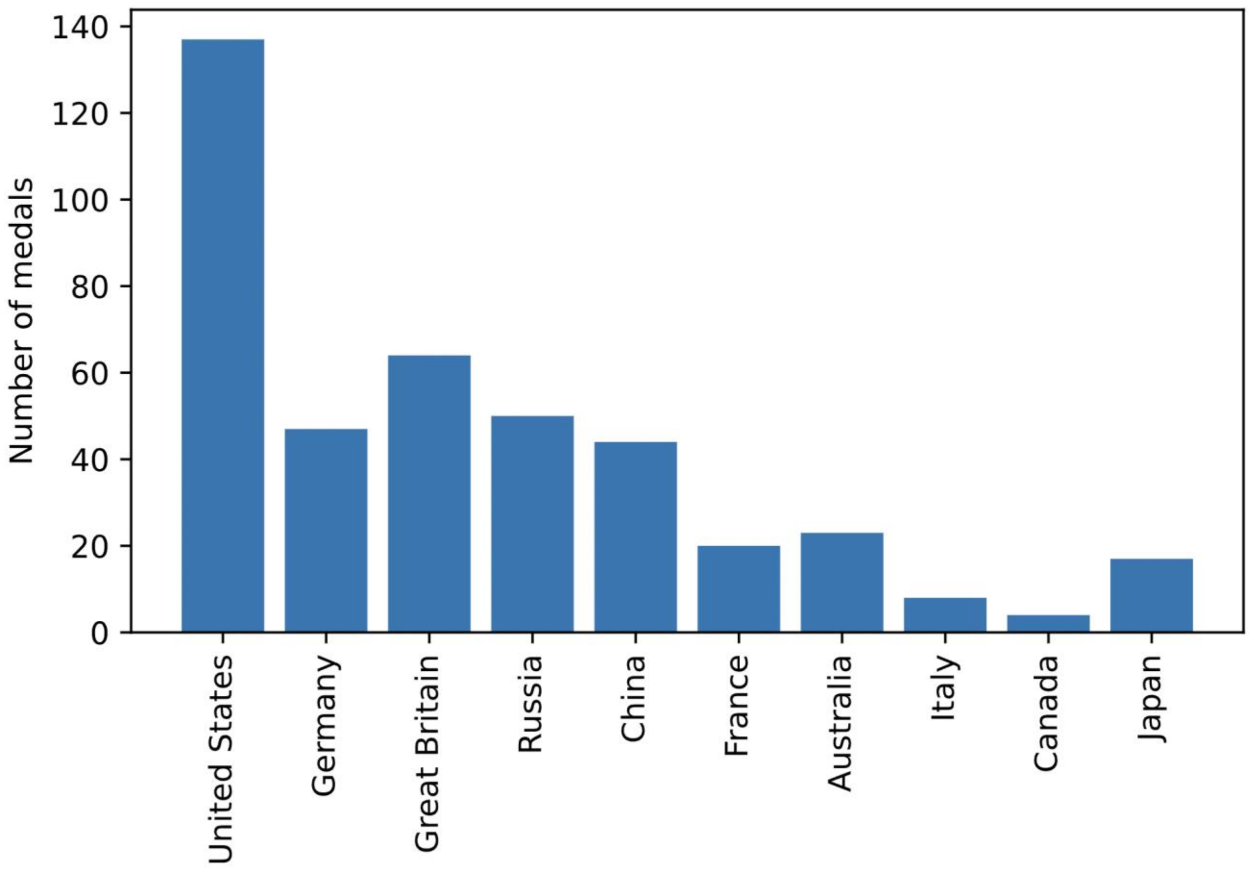
Data visualization

- Data Visualization is the graphical representation of information and data
- Easily way to share information to non-technical audiences without confusion
- Visualize patterns and relationship

Data visualization is part of EDA (Exploratory Data Analysis)

Matplotlib and Seaborn are the most commonly used libraries for data visualization.

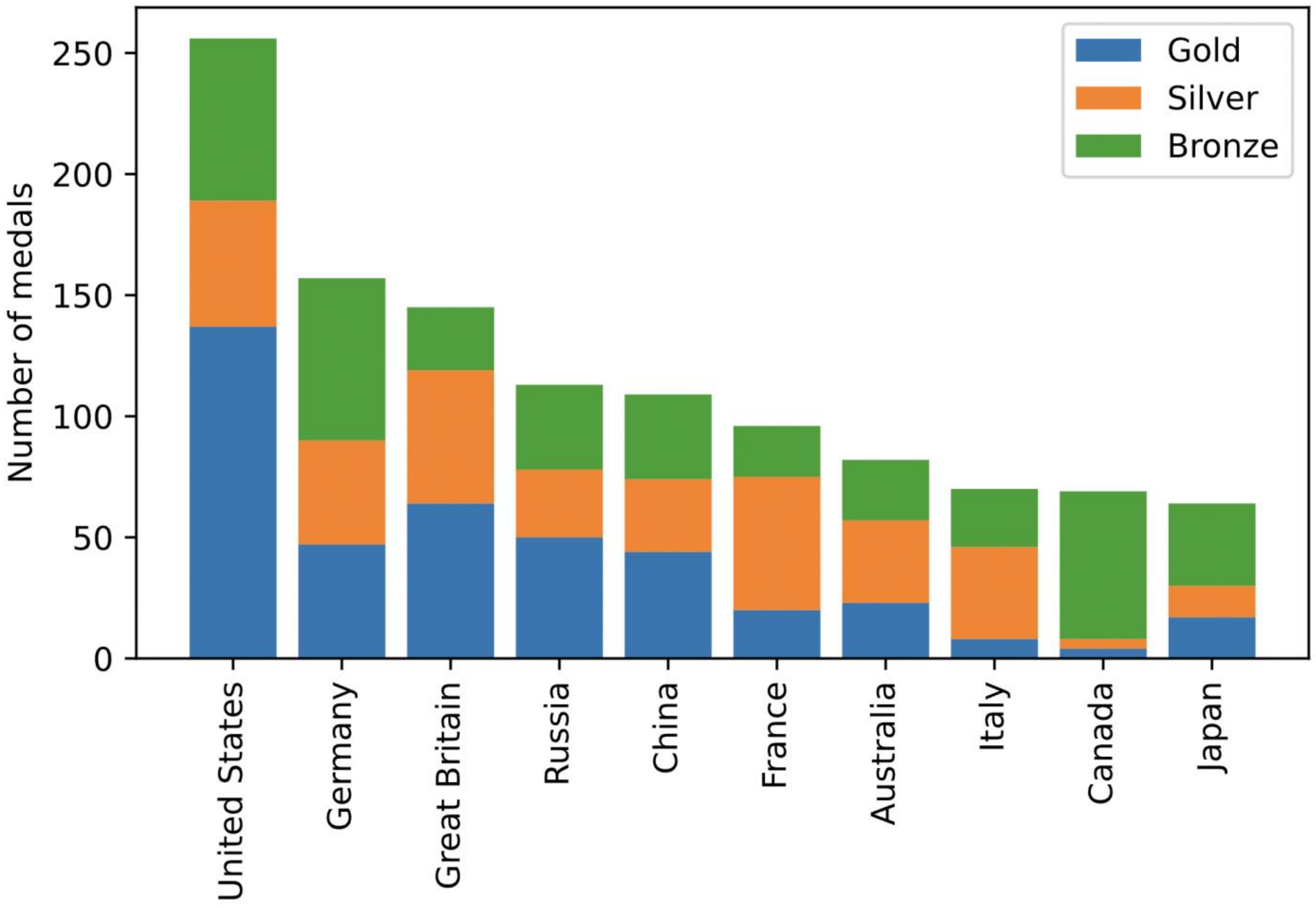
Data visualization



Bar charts (quantitative comparisons)

- Example: comparisons between the number of gold medals won by different countries in the 2016 Olympic games
- X axis: labels (name of the countries); Y axis: count of the values (number of gold medals won)

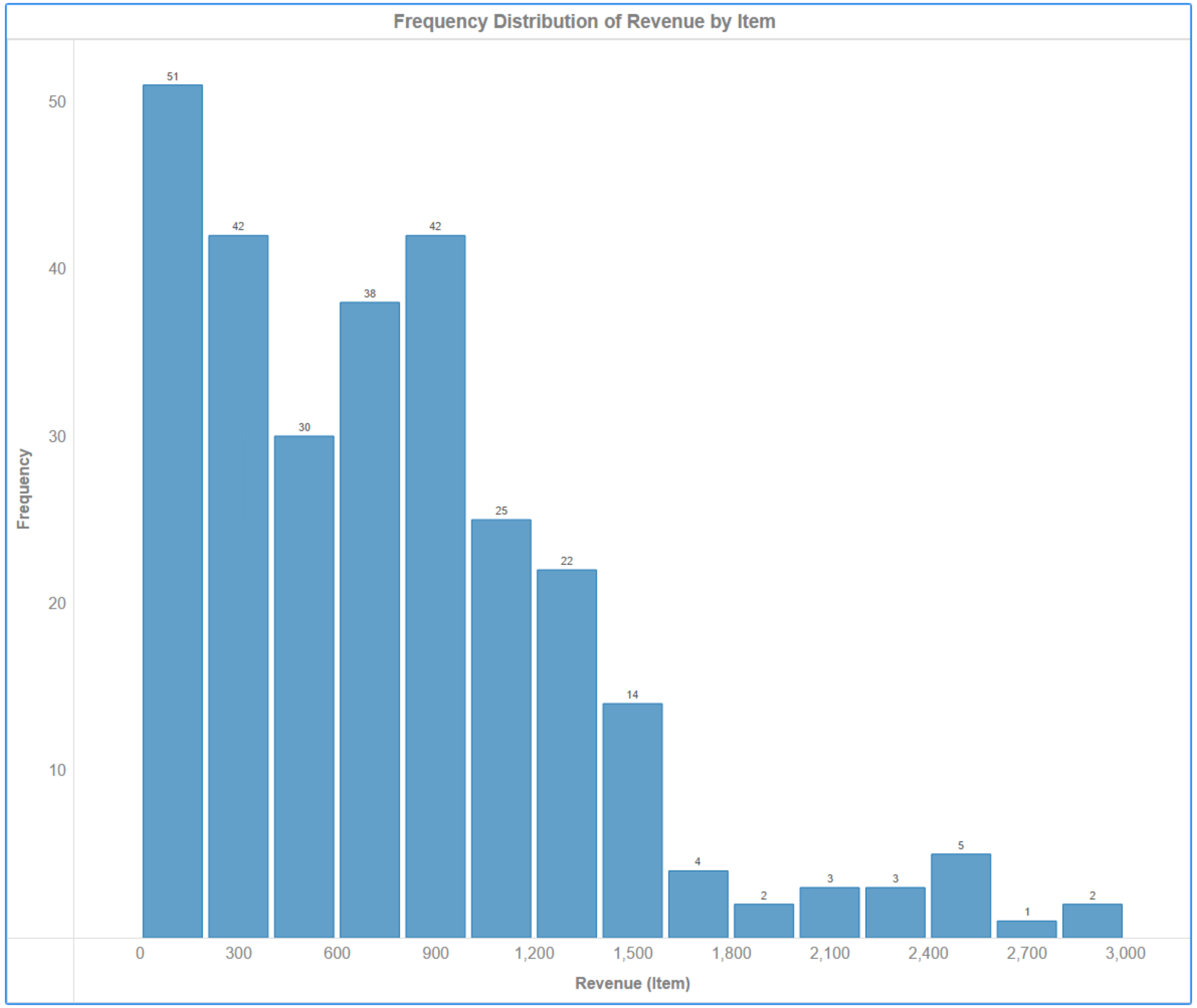
Data visualization



Bar charts (quantitative comparisons)

- Creating a stack bar chart where we add silver and bronze medals, differentiating the medals by color

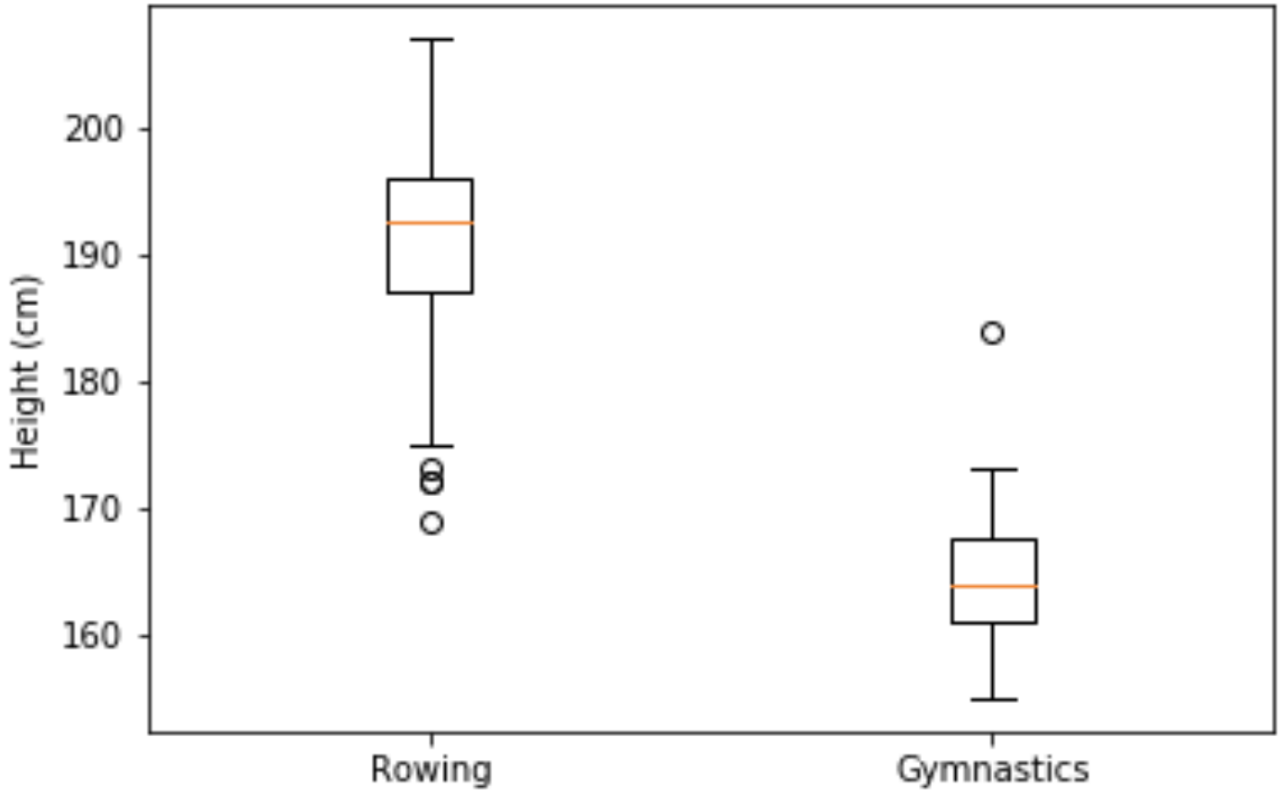
Data visualization



Histogram

- Dataset of the household income.
- Quantity of family by income.
- X axis: Revenue in $
  Y axis: number of family.

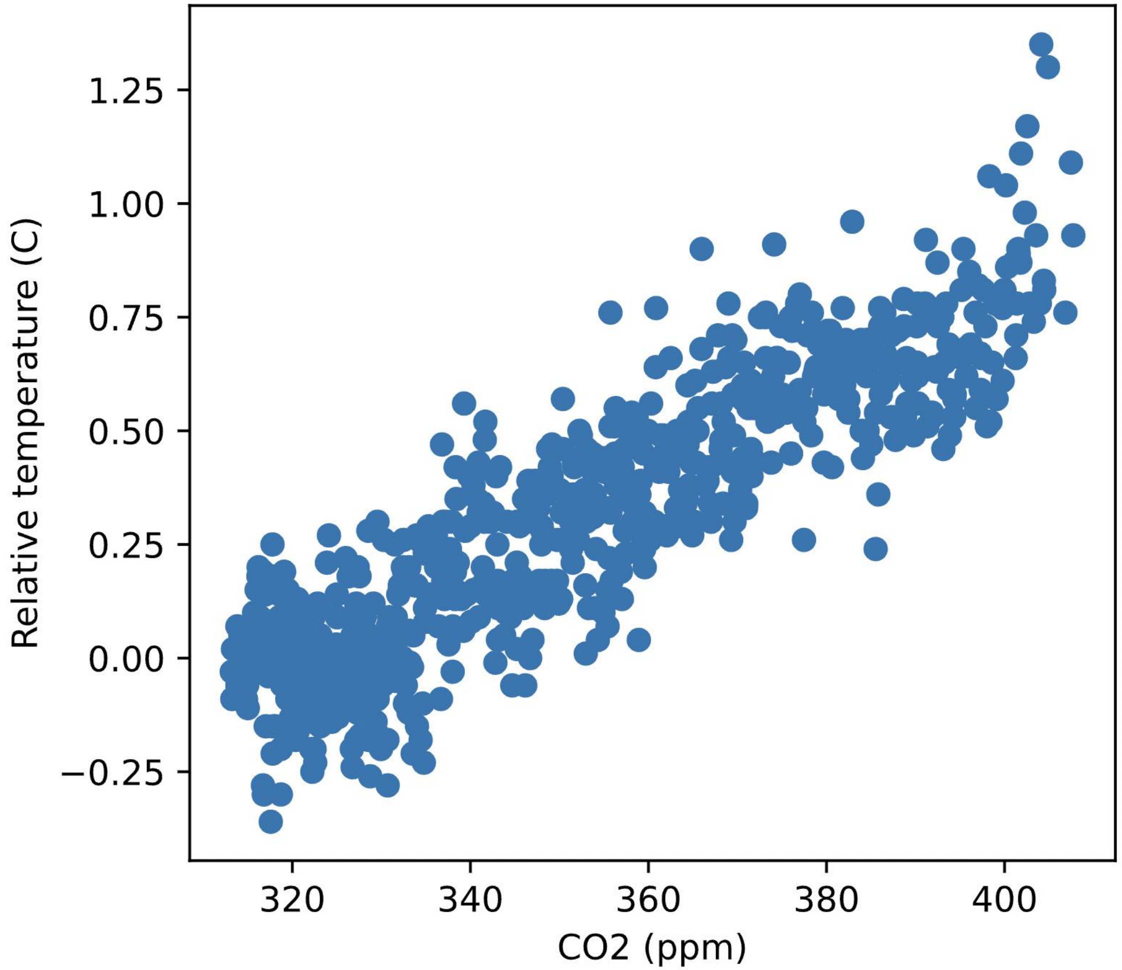
Data visualization



Boxplot

- The redline indicates the **median** height, the edges of the box portion at the center indicate the inter-quartile range of the data, between the 25th and the 75th **percentiles**.
- The **whiskers** at the end of the bar indicate the size of the inter-quartile range beyond the 25th and the 75th percentiles
- Points that appear outside the whiskers are **outliers**

Data visualization

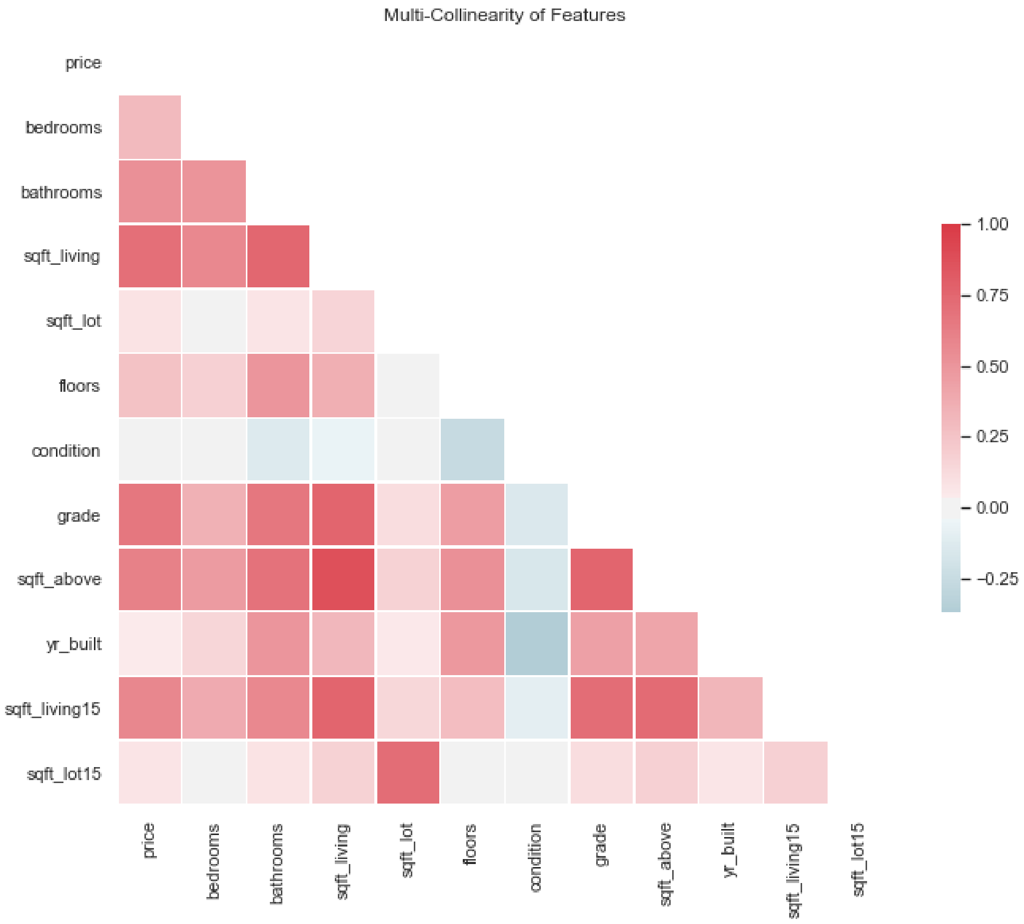


Scatter plot

- Used for bivariate comparisons (compare the values of different variables across observations)

- e.g. Climate change dataset: relation between the increase of temperature and the increase of carbon concentration

Data visualization

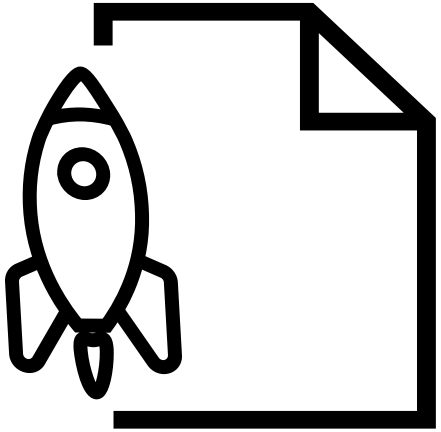


Correlation Matrix

- Displays the correlation coefficients for different variables.
- Allows to summarize a large dataset and to identify and visualize patterns in the given data.
- e.g. sqft_living and bedrooms are **positively correlated** because bigger homes typically have more bedrooms. yr_built and condition are **negatively correlated** because an older house starts to deteriorate as time passes.

Session 1

DEMO – Churn Analysis



[Back to slide](#)

Define the Business goal

Analyze the rate with which customers quit the product, site, or service.

Why customer churn analysis?

- One of the biggest concerns of any company
- One of the most common data science business goal

Define the Business goal

Main challenges of the customer churn analysis

- What is the likelihood of an active customer leaving an organization?
- What are key indicators of a customer churn?
- What retention strategies can be implemented to diminish prospective customer churn?

Setup notebook

[*LINK to Python Script*]

python™

NumPy

High-level mathematical functions to operate on large matrices and multidimensional arrays

seaborn

matplotlib

Creating static, animated, and interactive data visualizations

pandas

The most common Python library for Data manipulation and analysis

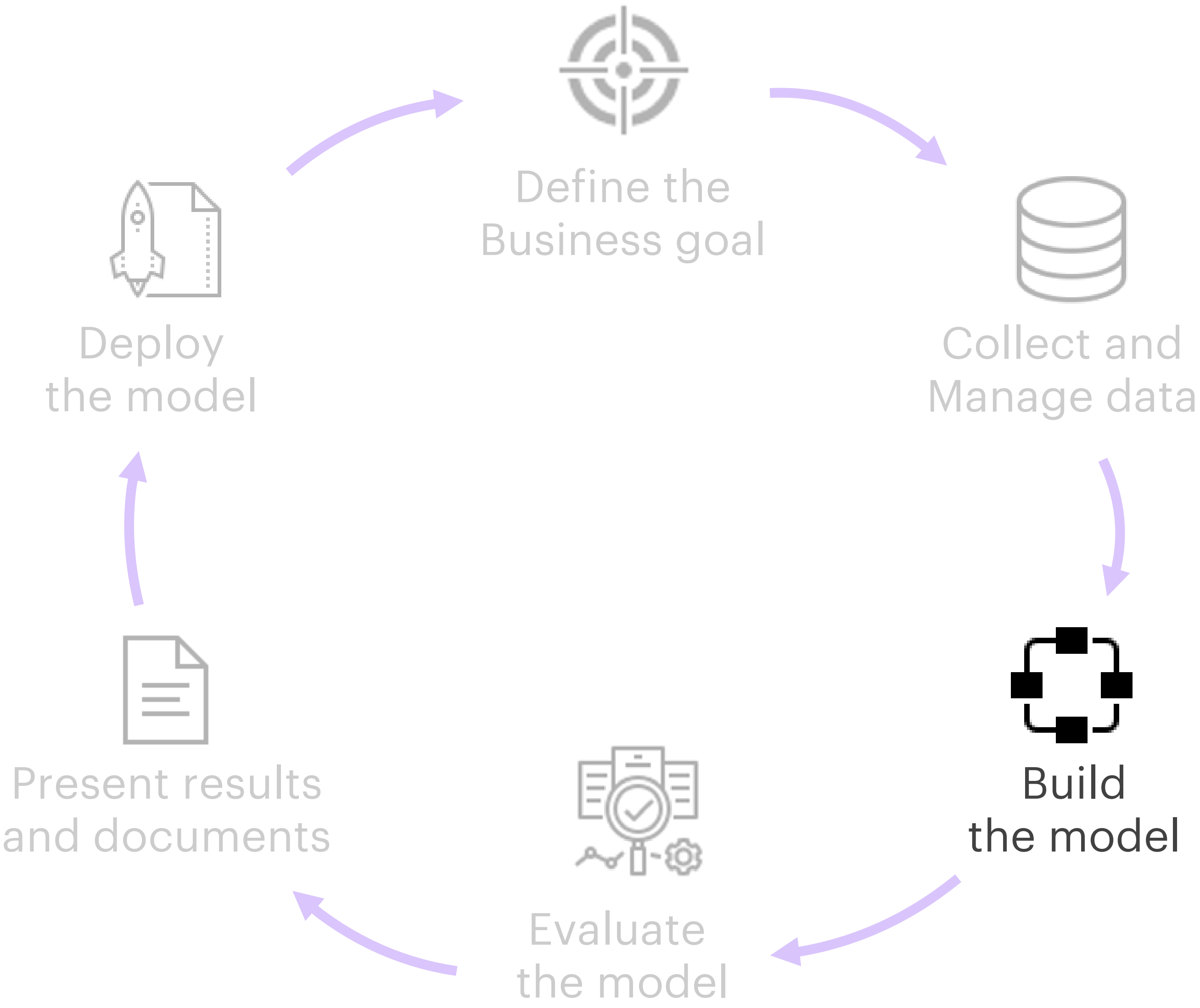
scikit learn

Simple and efficient tools for predictive data analysis

Requirements: no missing values & numeric value

Pre-processing data first

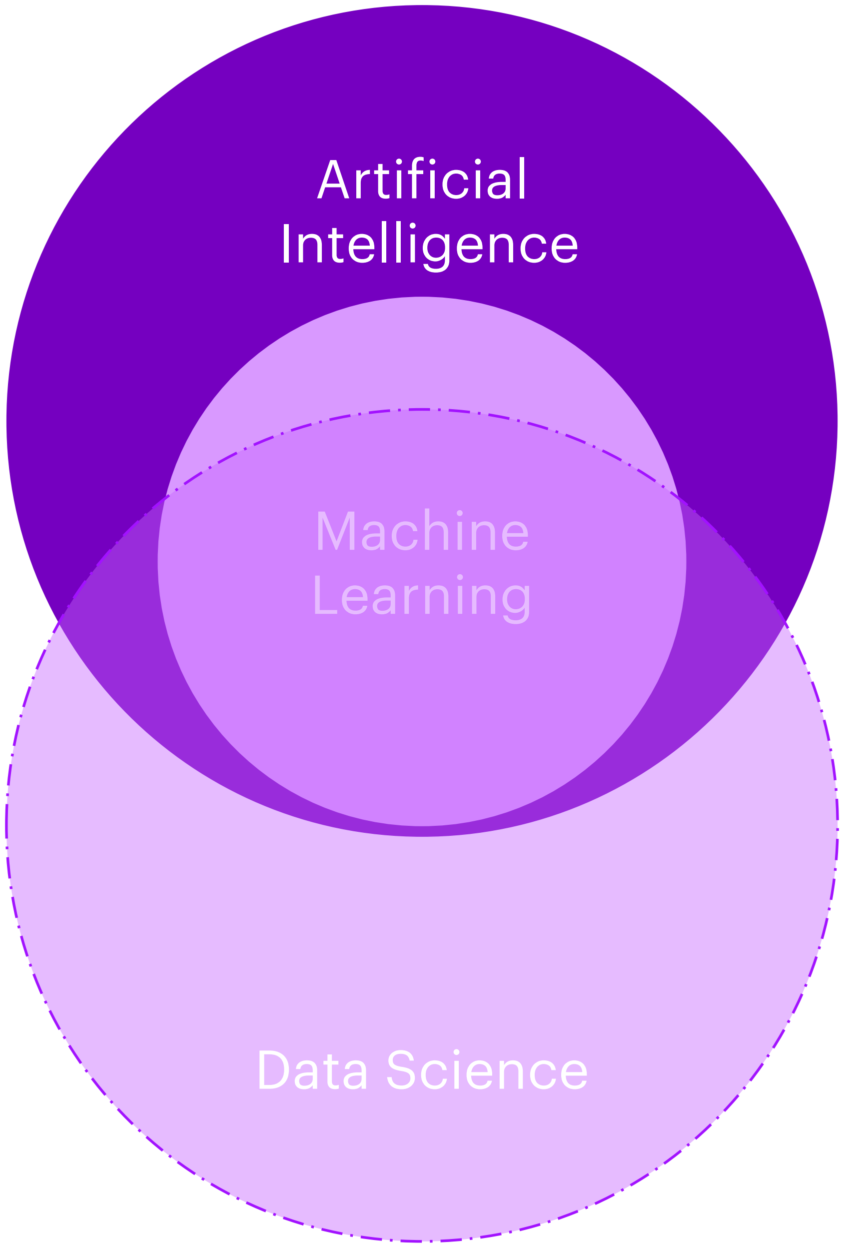
Note: a Library is a set of predefined functions or data structures designed to be connected to a software program through an appropriate connection



Session 1

Build the model – Introduction

Back to slide



Machine Learning

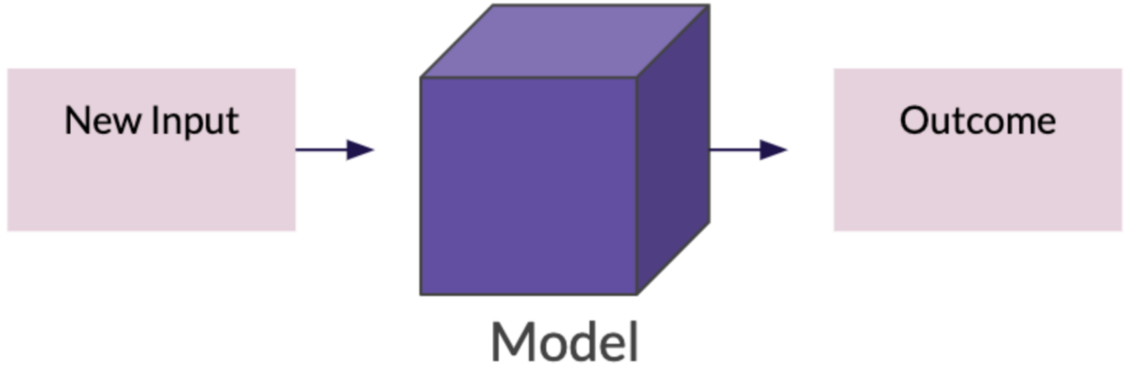
- ML is an important tool for data science work, since it helps to making discoveries and creating insights from data
- In ML computers makes inferences, predictions and find patterns from data without being explicitly programmed
- Machine learning learn patterns from existing data and applies them to the new data

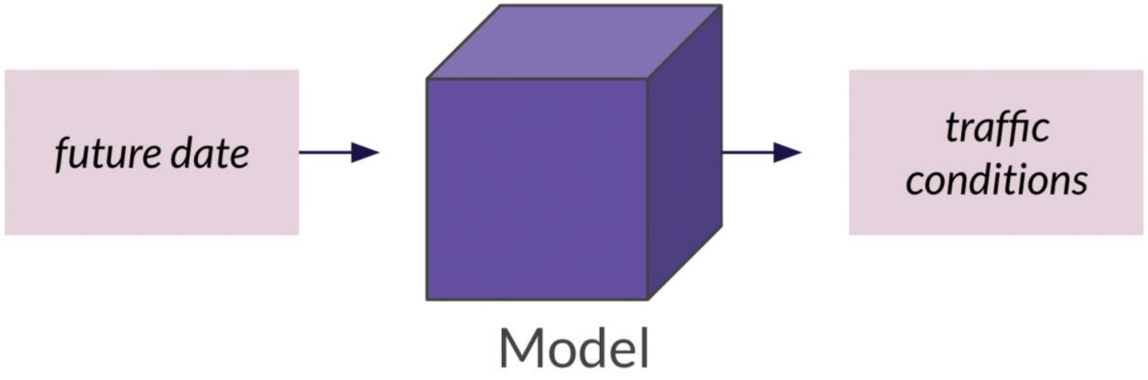
Machine Learning models

- **Machine Learning model**: statistical representation of a real world process (e.g. changes in traffic every hour), that is modeled using data

- We enter new input in a model to get an output

- e.g. make a model based on historical traffic data to predict how heavy the traffic will be in the future





Build the model - Introduction

Machine Learning

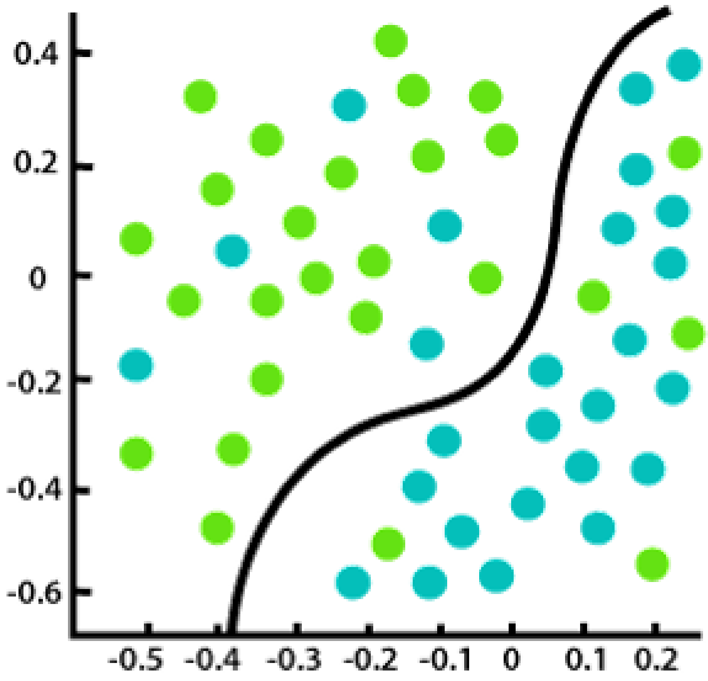
- Supervised learning
- Unsupervised learning

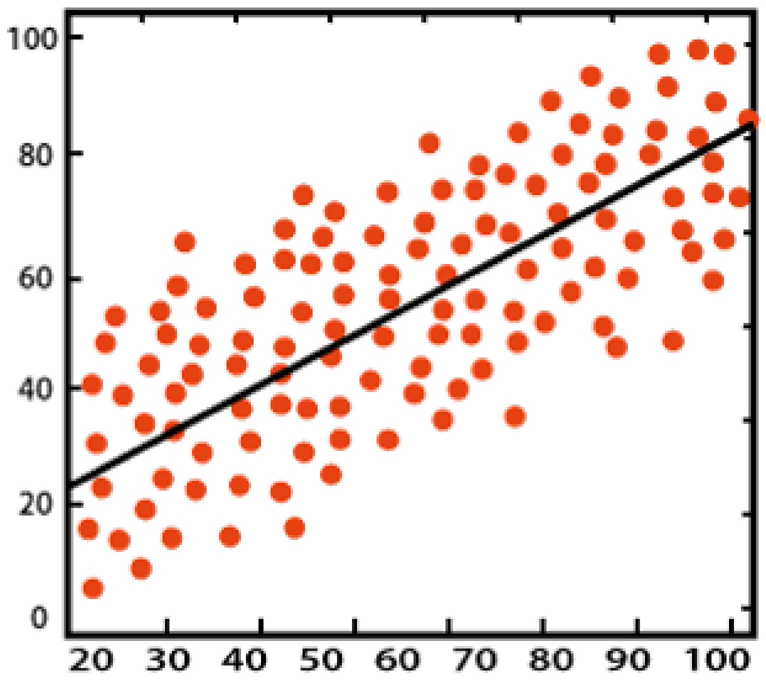
Supervised Learning models

- **Supervised learning:** A type of machine learning where the values to be predicted are already defined in a known variable
- Supervised learning models uses features to **predict the value of a target variable**
- e.g. Predicting the basketball player position basing on their points, assists, steals per game

	Features					Target variable
	points_per_game	assists_per_game	rebounds_per_game	steals_per_game	blocks_per_game	position
0	26.9	6.6	4.5	1.1	0.4	Point Guard
1	13	1.7	4	0.4	1.3	Center
2	17.6	2.3	7.9	1.00	0.8	Power Forward
3	22.6	4.5	4.4	1.2	0.4	Shooting Guard

Supervised Learning models





Classification **Regression**

Classification

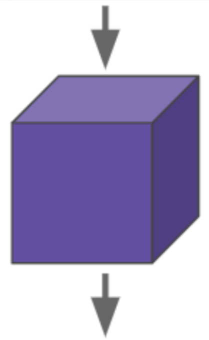
- It means to assign a category to an observation basing on many characteristics
- e.g. Predicting if the patient has heart disease basing on different features

Age	Sex	Cholesterol	Cigarettes per day	Family history of heart disease	Chest pain type	Blood sugar	Heart disease
55	M	221	5	True	typical angina	118	True
50	F	196	0	False	non-anginal pain	98	False
53	F	215	0	True	asymptomatic	110	True

Classification

- We input features to train the model
- We give to the model a new input (e.g. a new patient)
- The model output is a prediction based on the features.

Age	Sex	Cholesterol	Cigarettes per day	Family history of heart disease	Chest pain type	Blood sugar	Heart disease
65	F	208	2	False	typical angina	105	???



Heart disease
False

Regression

Reading ID	Humidity rate	Temperature in ˚C
0	0.89	7.388889
1	0.86	7.227778
2	0.89	9.377778
3	0.83	5.944444
...	...	...

- Linear Regression models allow to finds out a relationship between an x independent variable (input) and y dependent variable (output).
- e.g. Predicting temperature basing on humidity

Unsupervised Learning models

- **Unsupervised learning:** A class of machine learning techniques to discover patterns in data.
- It allows the model to work on its own to discover patterns and information that was previously undetected.

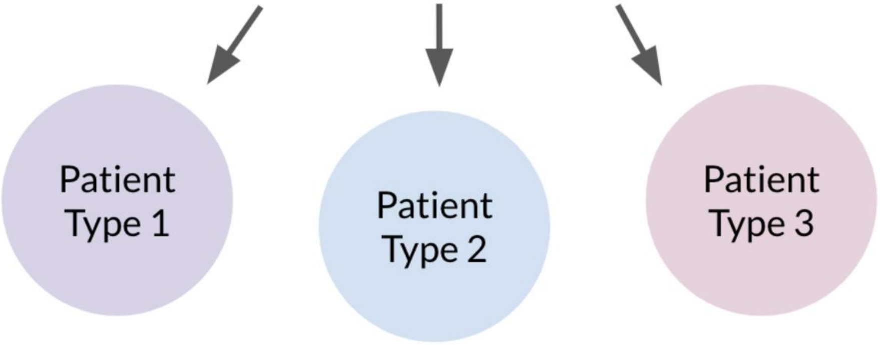
Clustering

- We input our dataset into a clustering model to **get categories of patients with feature similarity**

- e.g. starting from a group of patients with heart disease and create a cluster of patients based on: cholesterol, blood sugar level and age range

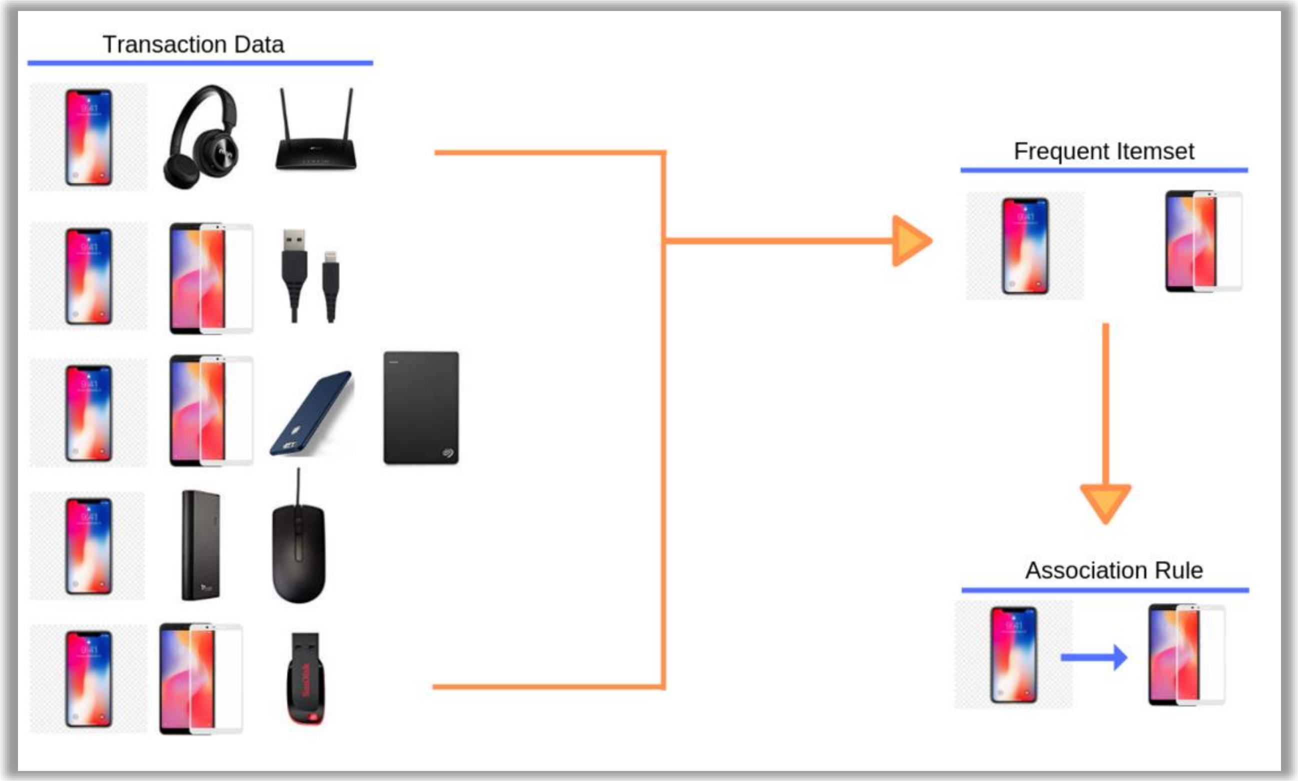
- We don't know these categories and their number before running the model

Age	Sex	Cholesterol	Cigarettes per day	Family history of heart disease	Chest pain type	Blood sugar	Heart disease
55	M	221	5	True	typical angina	118	True
53	F	199	0	True	non-anginal pain	98	True
53	F	215	0	True	asymptomatic	110	True
62	M	245	3	False	typical angina	126	True
...	...	...	...	...	...	...	...



Association

- It means **finding relationship between** events that happens together
- Used for example for market basket analysis: finding objects that are bought together



References

See the links on the Github file

Session 2

The Data Science Workflow -
Deep Dive into Analysis

Introduction to Data Science Workflow

Session 2

Deep dive into Analysis

- Build the model – Deep dive
  - Training and test data
  - Linear Regression model
  - Classification models (Logistic Regression, Decision Tree and Random Forest)
  - Clustering models (k-Means, hierarchical clustering)
- Evaluate the model
- Present results and documents
- Demo – Churn analysis (part 2)
- Demo – Customer segmentation

Q & A

Thank You