

Data Science

Chapter 1: Introduction to Data Visualization

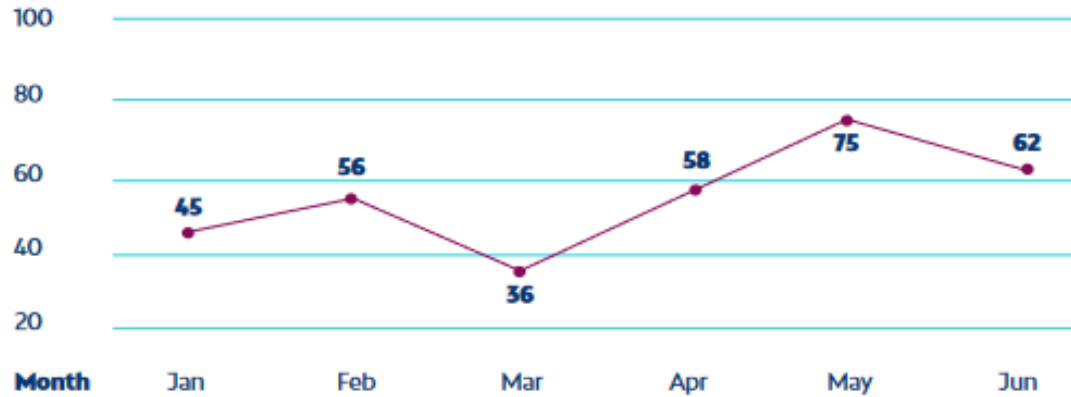
Chapter outline

This chapter covers

- Importance of Data Visualization
- Understanding/Discussing the flow of a data science process
- Understanding/Discussing the importance of data visualization during the process

Importance of Visualization

- we typically process images 60,000 times faster than a table or a text



Month	Jan	Feb	Mar	Apr	May	Jun
Sales	45	56	36	58	75	62

- This graphic below could clearly explain the country with the greatest demand for a certain product compared globally, in a concrete month.

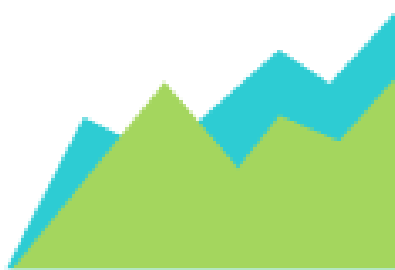


Importance of Visualization

- Data Visualization helps in interpreting relations between data
 - **Correlation:** Data with two or more variables that can demonstrate a positive or negative correlation with one another. For example: salaries based on level of education.



- **Series over time:** Here we can trace the changes in the values of a constant metric over the course of time. For example: monthly sales of a product over the course of two years



Types of Attributes

- Before we talk about visuals themselves, we must first understand the different kinds of data that can be visualized.
 - **Nominal – categorical** : Nominal values represent discrete units . It has no orders
 - Examples: eye color, sex (M, F).
 - **Ordinal – categorical** : Ordinal values represent discrete and ordered units.
 - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}, educational background (master, PhD)
 - **Interval – numerical** : Interval values hold no true zero and can represent values below zero..
 - Examples: temperatures in Celsius or Fahrenheit.
 - **Ratio – numeric** : Ratio values are the same as interval values, with the difference that they never fall below zero. More than this, we can say: $\text{time1} = 3 * \text{time2}$
 - Examples: temperature in Kelvin, length, time, counts

Motivation - example

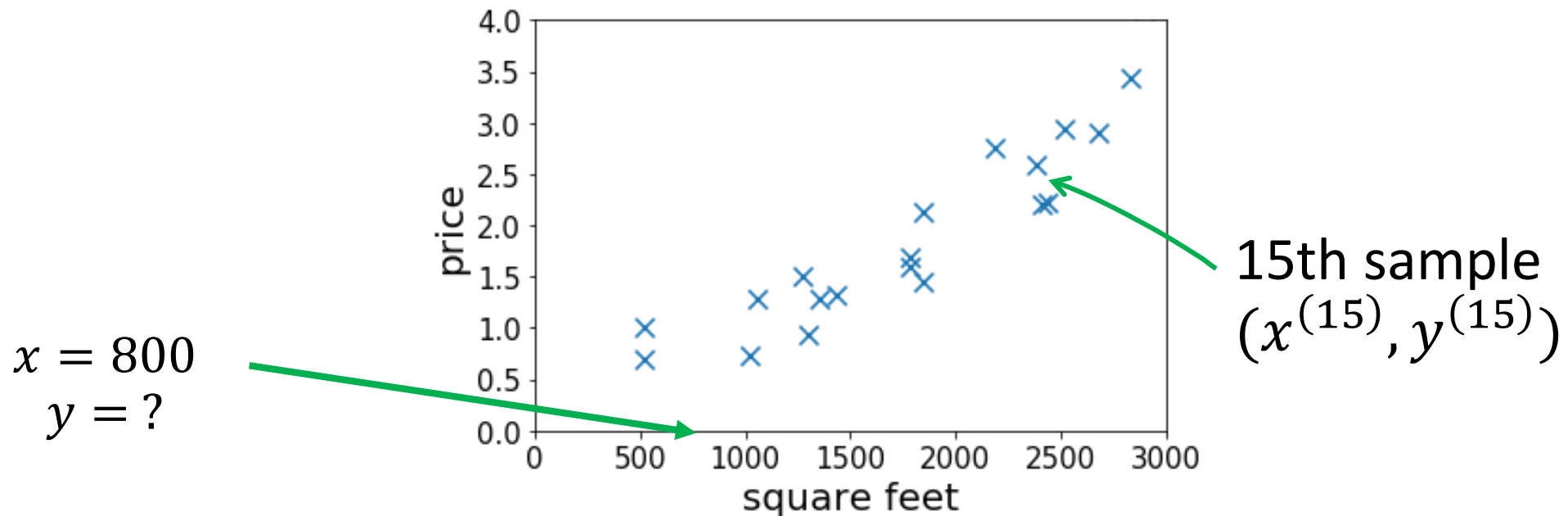
➤ Given: a dataset that contains n samples, where:

➤ x is the area in square feet

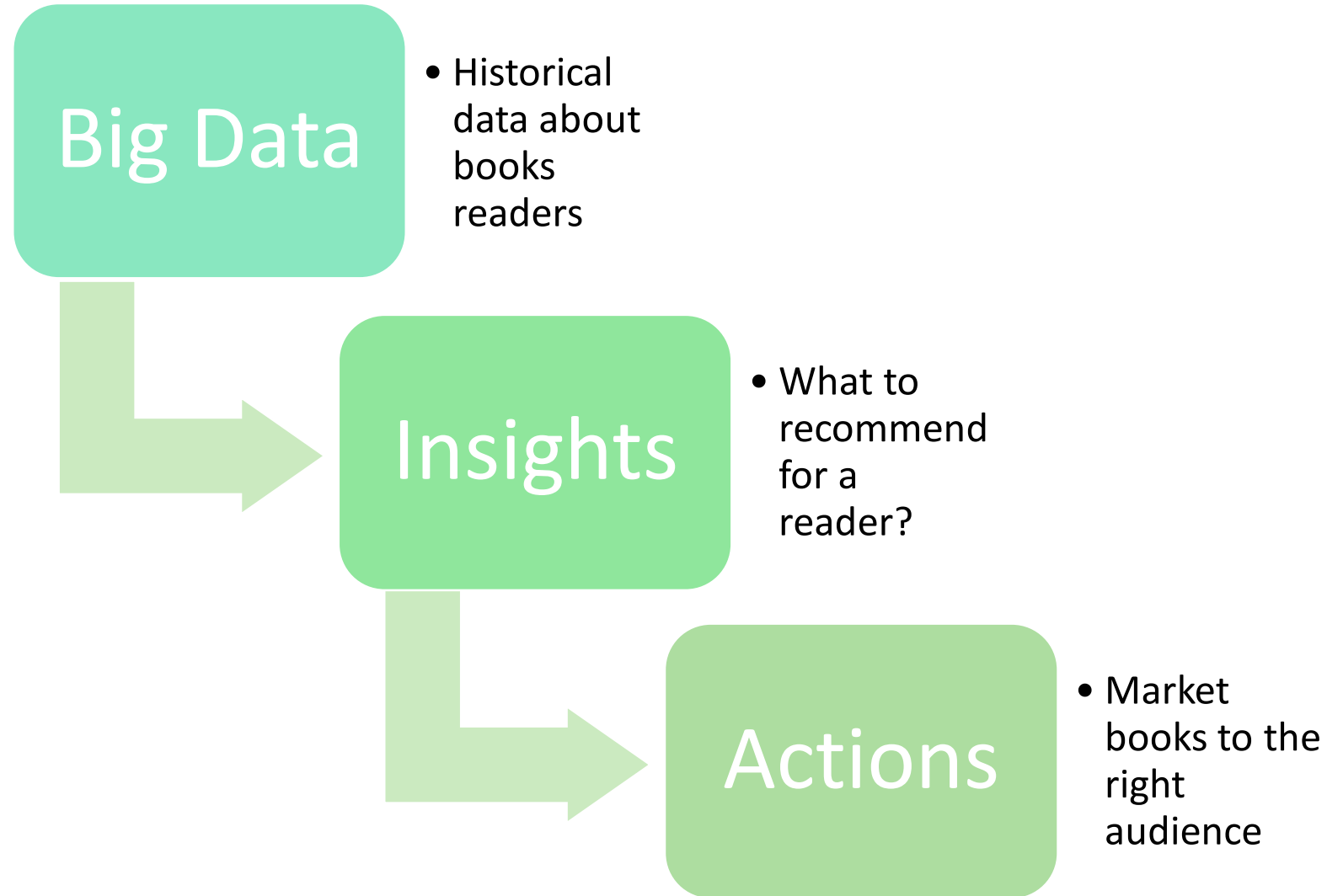
➤ y is the price

$$(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(n)}, y^{(n)})$$

➤ Task: if a residence has x square feet, predict its price!



Understanding the Data Science Benefits



Definitions

➤ Big data

Big data is a blanket term for any collection of data sets so large or complex that it becomes difficult to process them using traditional data management techniques such as, for example, the RDBMS (relational database management systems). The characteristics of big data are often referred to as the **three Vs**:

- **Volume** - How much data is there?
- **Variety** - How diverse are different types of data?
- **Velocity** - At what speed is new data generated

Often these characteristics are complemented with a **fourth V**:

- **Veracity** - How accurate is the data?

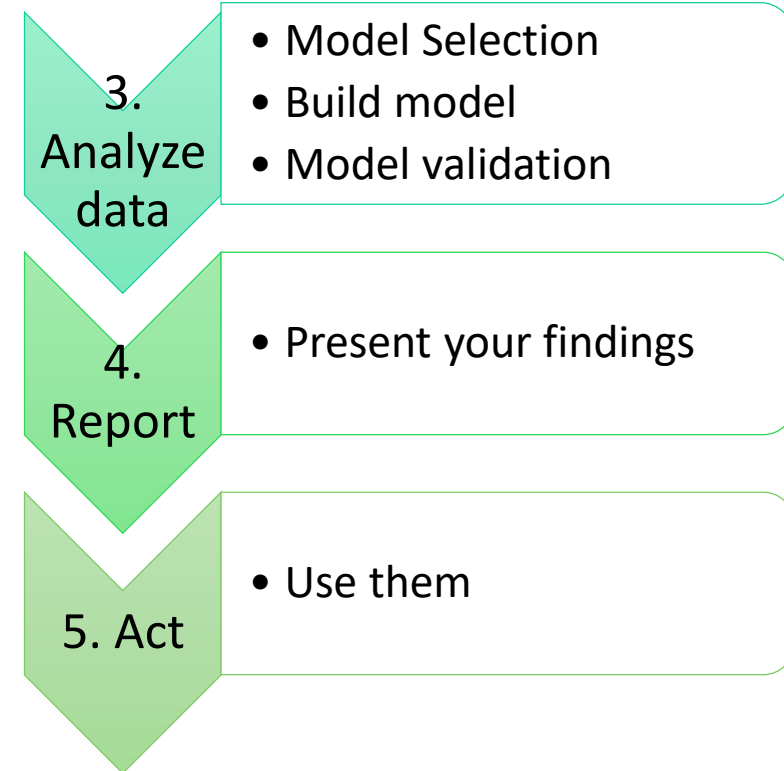
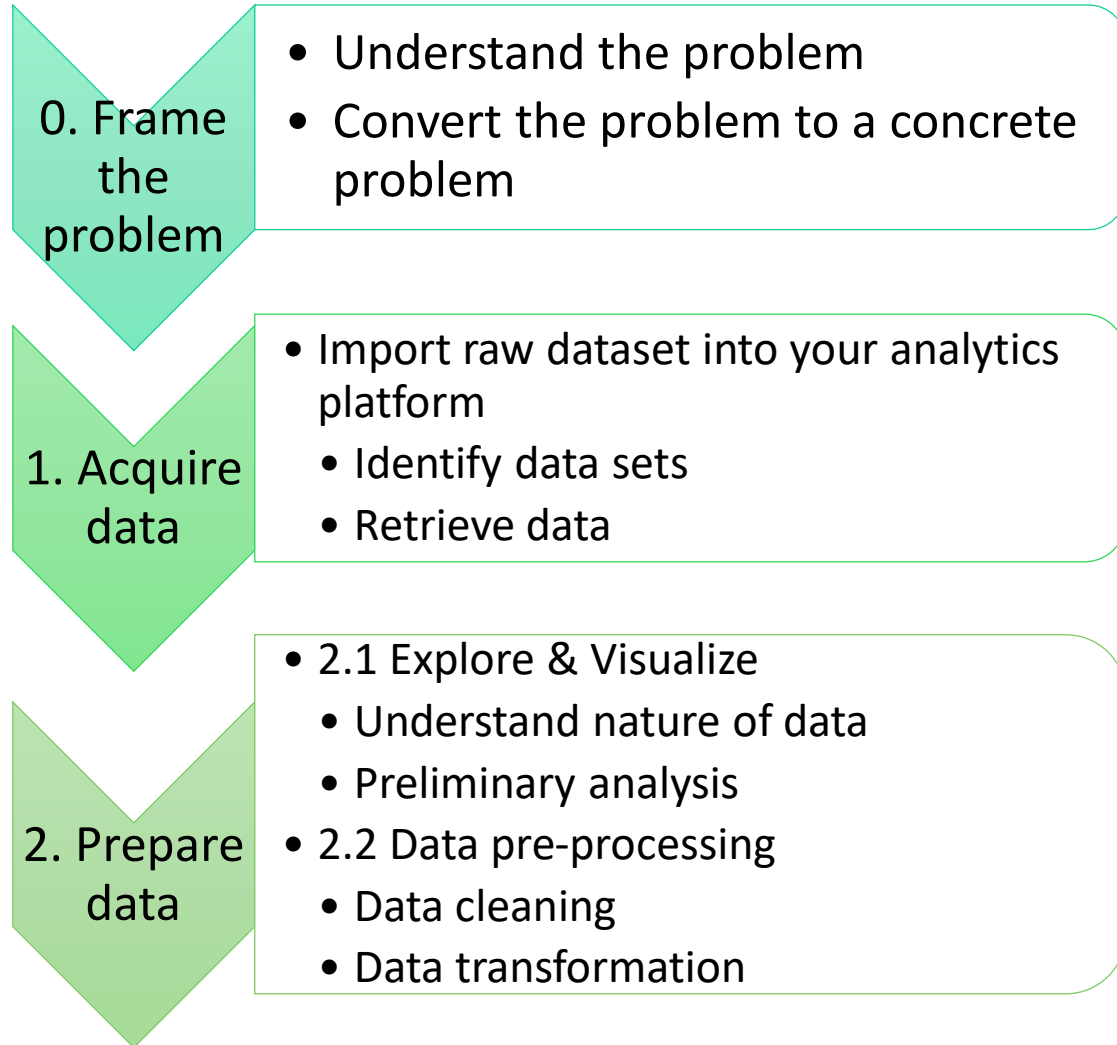
Definitions

➤ Data Science

Data science is an **inter-disciplinary** field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data.

You can think of the relationship between big data and data science as being like the relationship between crude oil and an oil refinery.

Data Science process



In reality you won't progress in a linear way.
Don't be slave to the process.

Step 0: Frame the problem

“A well-defined data science problem, is a problem half-solved”

- The first thing you have to do before you solve a problem is to define exactly what it is. You need to ask questions:
 - Who is your client? - **librarian**
 - What exactly is the client asking you to solve? - **recommend books to readers**
 - How can you translate their ambiguous request into a concrete, well-defined problem?
- **data about Books? Readers? Sales? Stock... ?**
- It's important that at the end of this stage, you have all of the information and context you need to solve this problem.

Step 1: Acquire data

Data comes from many places with many ways to access it



Step 1: Acquire data

Source

Traditional Databases



Access

SQL & query browsers



Step 1: Acquire data

Source

Text files (text, excel, ...)



Access

Scripting languages



Step 1: Acquire data

Source

Remote data



Access

Web services

SOAP

REST

Web
Socket

Step 1: Acquire data

Source

noSQL storage



Access

Web Services + API

Step 2: Prepare data

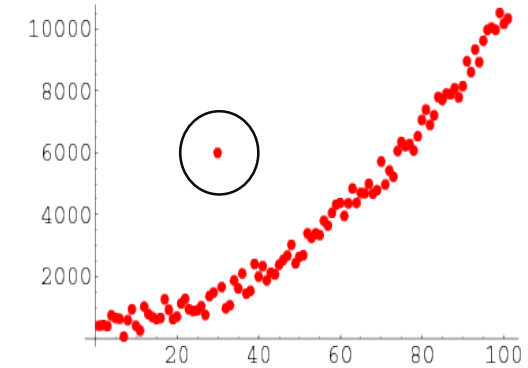
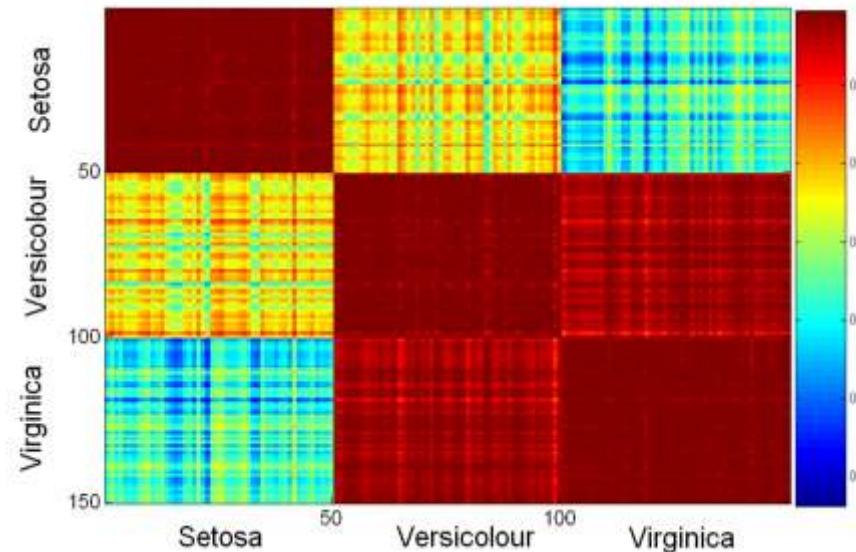
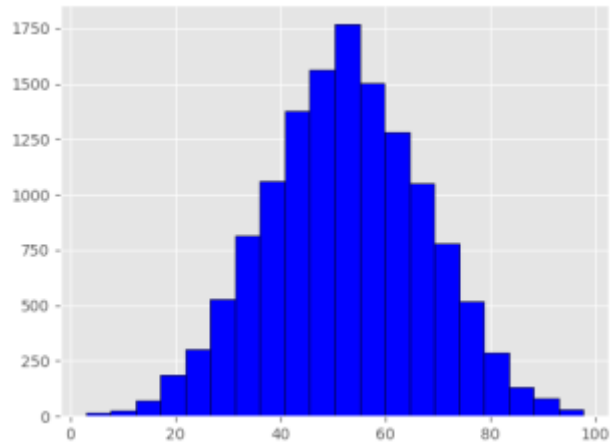
“ Always remember: Garbage in = Garbage out ! “

- 2.1 Explore & Visualize
 - Understand nature of data
 - Preliminary analysis
- 2.2 Data pre-processing
 - Data cleaning
 - Data transformation

Step 2: Prepare data

➤ Step 2.1: Explore & Visualize

- Understand nature of data
- Preliminary analysis



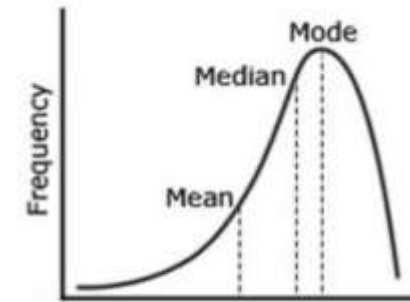
Step 2: Prepare data

Descriptive statistics

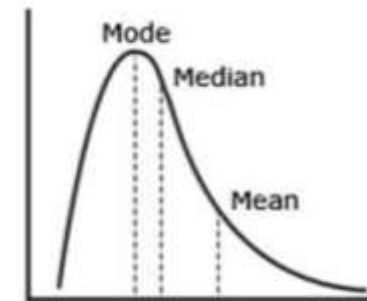
Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire or a sample of a population.

Descriptive statistics are broken down into 2 categories:

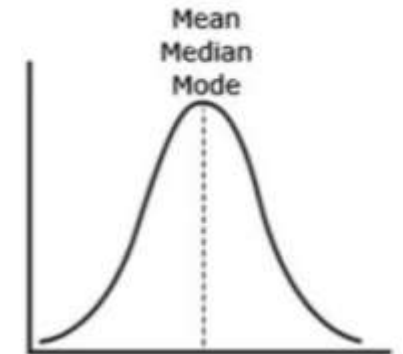
- Measures of central tendency: the mean, median and mode
- Measures of variability (spread): standard deviation, variance, minimum and maximum variables, and kurtosis and skewness.



Negatively Skewed



Positively Skewed

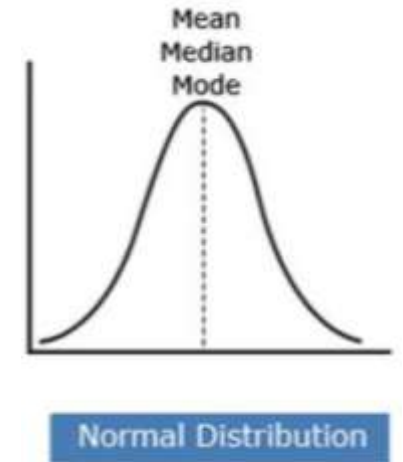
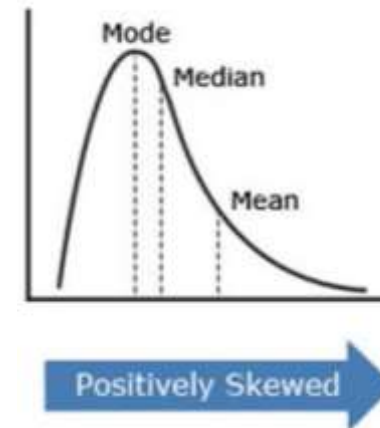
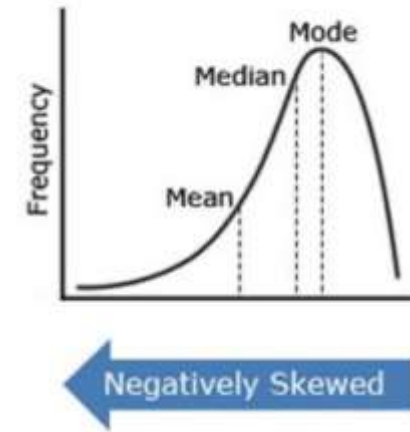


Normal Distribution

Step 2: Prepare data

Extra notes

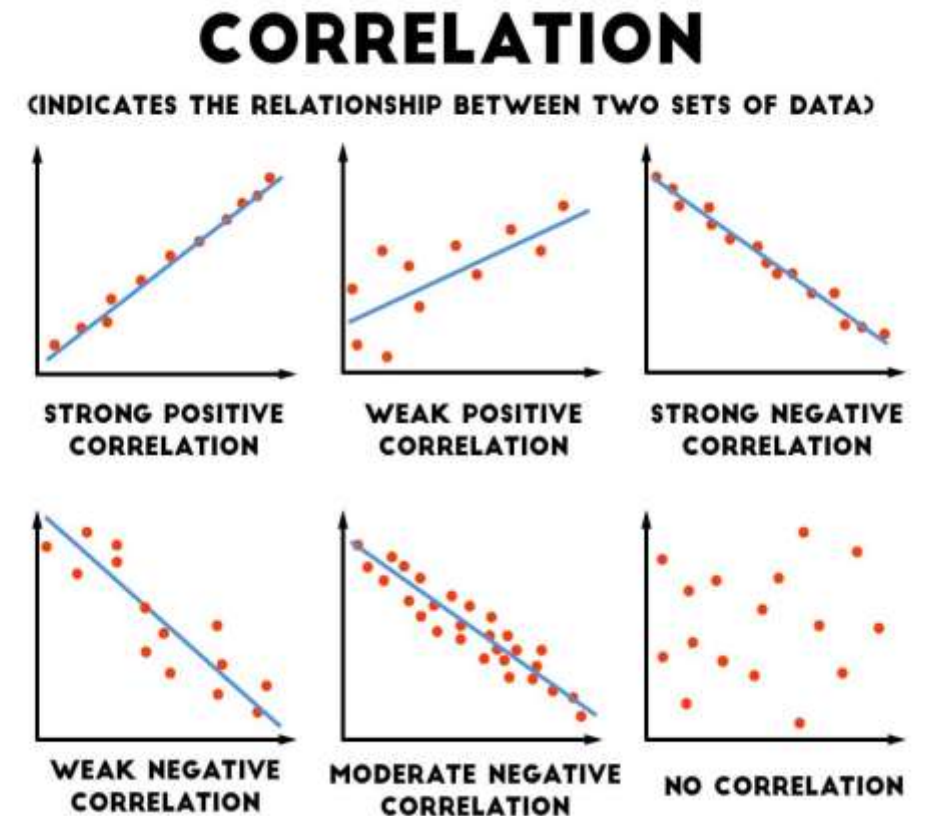
- Gaussian distribution (Normal) is the most important probability distribution in statistics because it fits many natural phenomena like age, height, test-scores, IQ scores, sum of the rolls of two dices and so on.
- Datasets with Gaussian distributions makes applicable to a variety of methods that fall under parametric statistics.
- Conclusions and summaries derived from such analysis are intuitive and easy to explain to audiences with basic knowledge of statistics



Step 2: Prepare data

Correlation

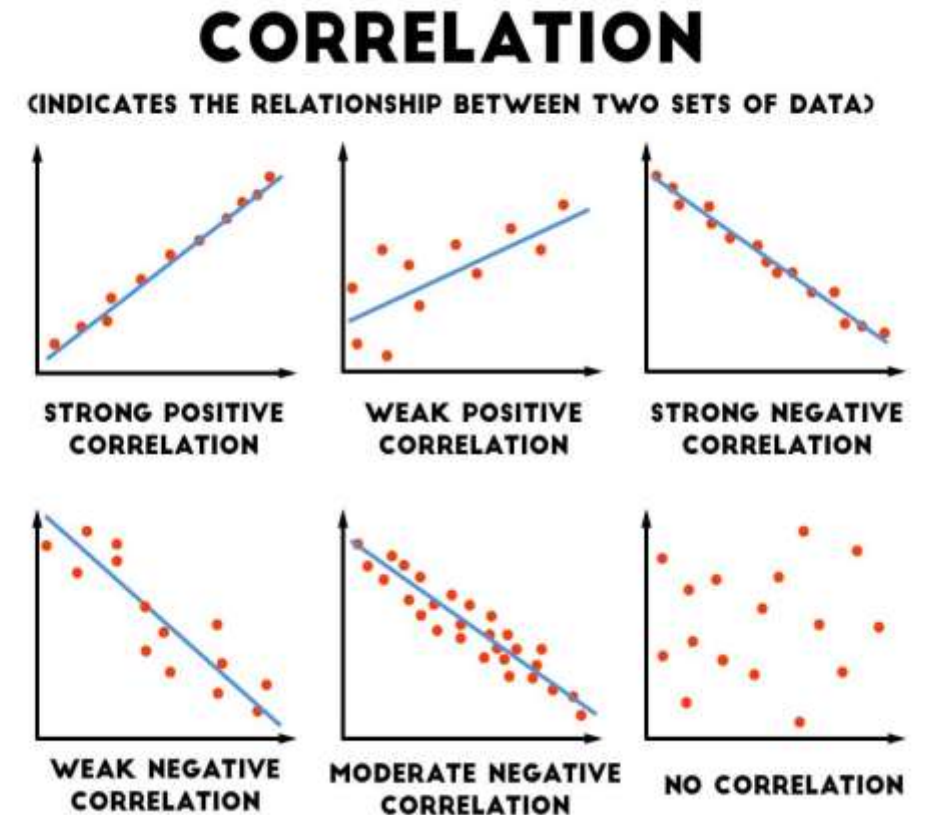
- Correlation explains how one or more variables are related to each other. These variables can be input data features which have been used to forecast our target variable.
- Two features (variables) can be positively correlated with each other. It means that when the value of one variable increases then the value of the other variable(s) also increases.



Step 2: Prepare data

Correlation

- Sometimes we are trying to forecast a variable y e.g. stock price, and we spend a huge amount of time gathering data for the features x_1 (company sales) and x_2 (company revenue) that can help us forecast the variable y . However the two features might be strongly positively correlated with each other.
- Thus it is suffice to only gather data for one of the features and feed it into our data science model. Not only it can save us effort in gathering and cleaning the data, it can also speed up the time it takes to train the model.



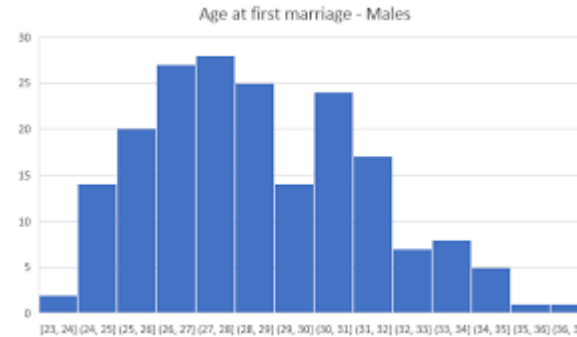
Step 2: Prepare data

Data visualization

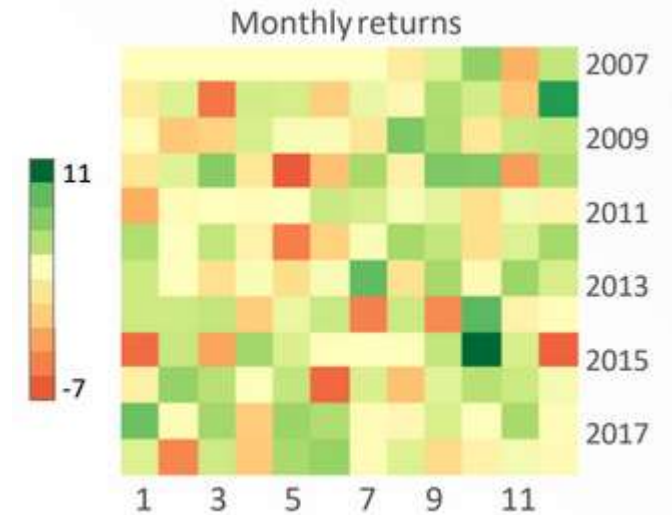
Line graphs



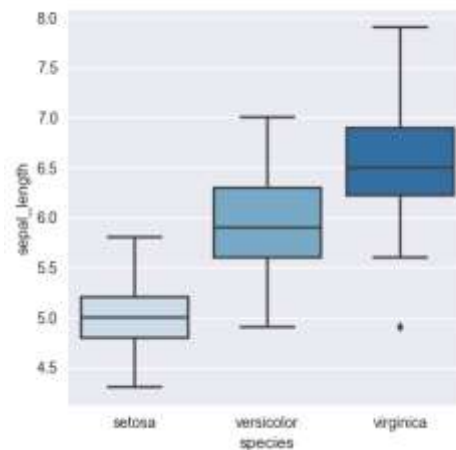
Histogram



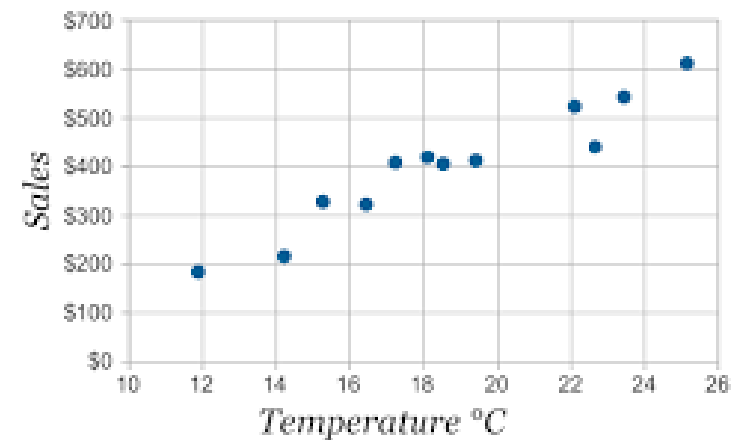
Heat Maps



Boxplots



Scatter plots



Step 2: Prepare data

➤ **Step 2.2: Data pre-processing**

- Data cleaning
- Data transformation

➤ Real world data is messy!

- High dimensionality
- Inconsistent values (example: 2 different addresses for same client ID)
- Duplicate records
- Missing values
- Invalid dataOutliers

Step 2: Prepare data

Addressing data quality issues

- For missing and invalid values:
 - Remove data
 - Replace the missing value with the mean or median or mode
- For duplicate records:
 - Merge
 - Choose the one with more data
 - Remove a record
- For outliers:
 - Remove it (most of the times)

Most of times, you need business knowledge to pre-process the data

Step 2: Prepare data

Features selection

- Select features
- Add feature(s) : example, add the feature country (deducted from the address feature)
- Combine features : example instead of length and width features → area feature

Step 2: Prepare data

Data scaling

- Feature scaling is a method used to standardize the range of features of data.
- Consider a data set containing two features(attributes), age(x1), and income(x2). Where age ranges from 0–100, while income ranges from 0–20,000 and higher. So, these two features are in very different ranges. When we do further analysis, the attributed income will intrinsically influence the result more due to its larger value.
- Sample s1 (20 , 1000) Sample s2(20, 1100) Sample s3(30, 1000)
- $d(s1,s2) \gg \gg \gg d(s1, s3)$ but in reality they are not

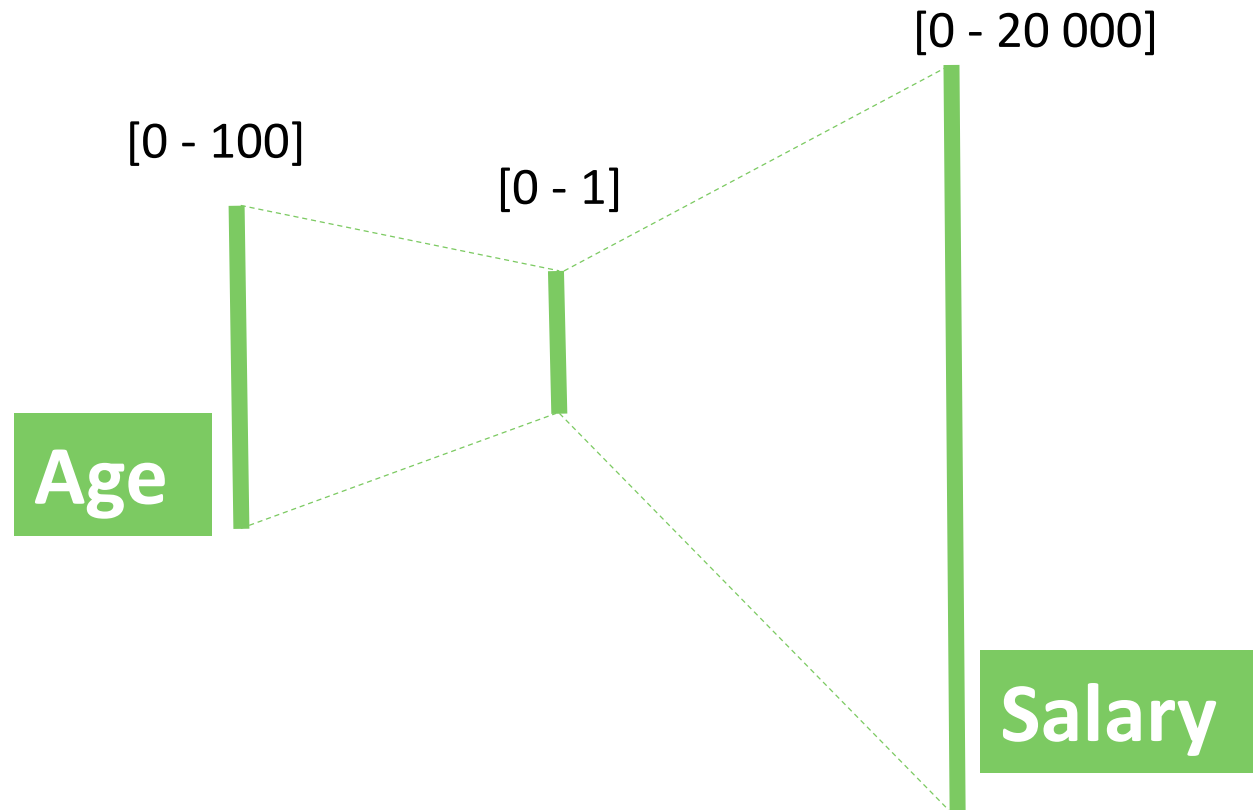
Step 2: Prepare data

Data scaling

In scaling (also called min-max scaling), you transform the data such that the features are within a specific range e.g. [0, 1].

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

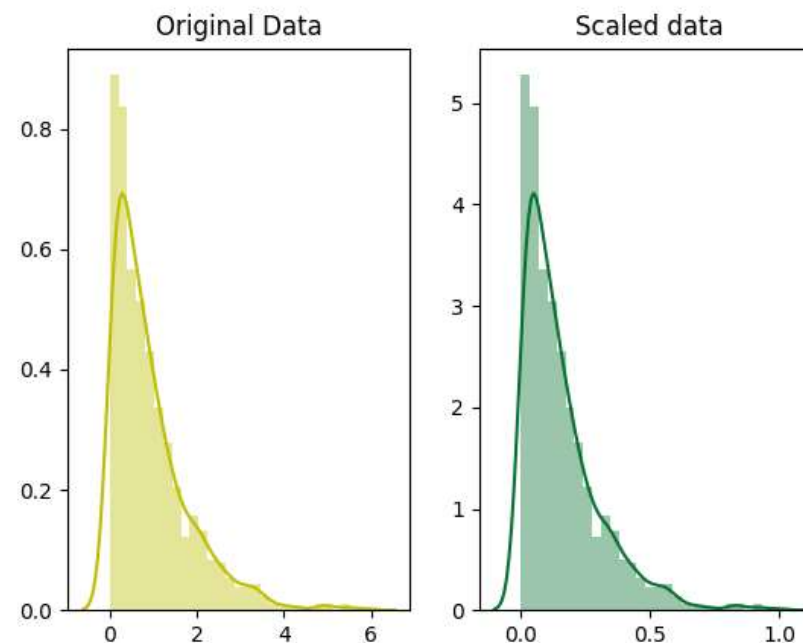
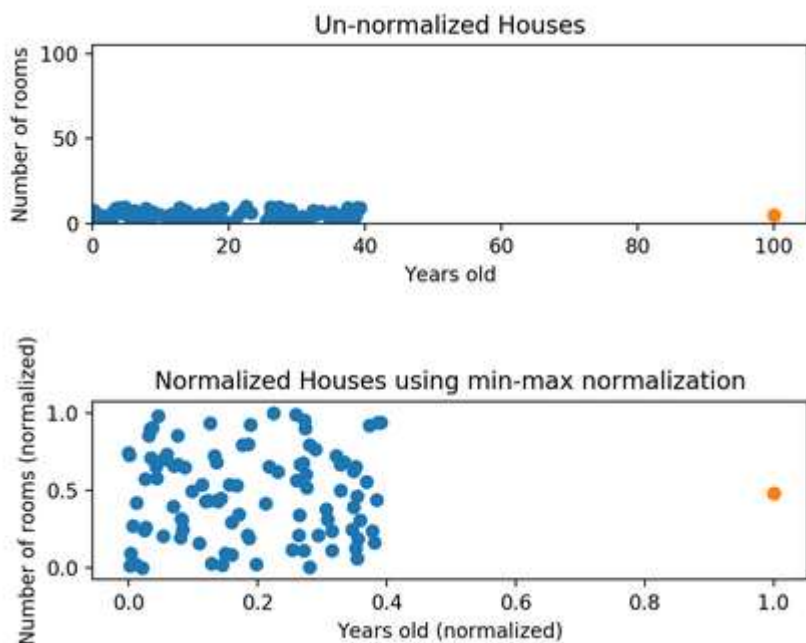
where x' is the scaled value.



Step 2: Prepare data

Extra notes

- Scaling is important in the algorithms such as support vector machines (SVM) and k-nearest neighbors (KNN) where distance between the data points is important.
- Scaling does not change the distribution of your data !



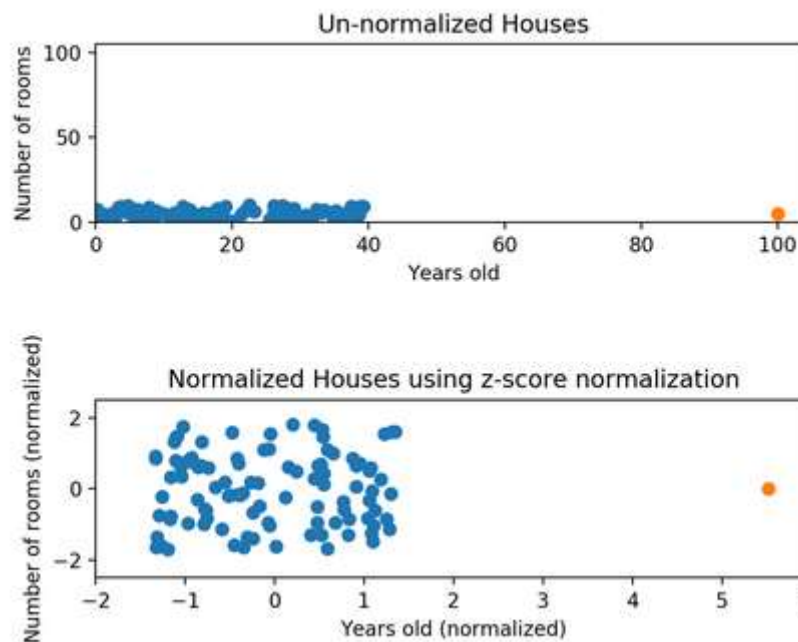
Step 2: Prepare data

Standardization

- In Standardization (also called z-score normalization), you transform your data such that the resulting distribution has a mean of 0 and a standard deviation of 1.
- It's widely used in SVM, logistics regression and neural networks.

$$x' = \frac{x - x_{mean}}{\sigma}$$

where x' is the standardized value.



Step 2: Prepare data

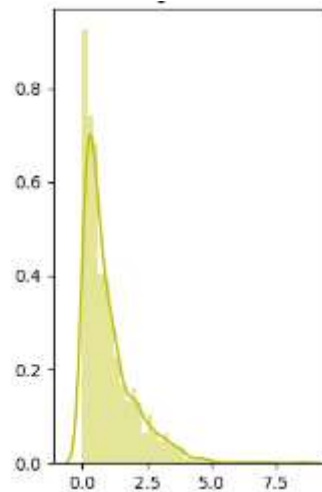
Normalization

- In normalization you're mostly changing the shape of the distribution of your data.
- You need to normalize our data if you're going use a machine learning or statistics technique that assumes that data is normally distributed e.g. linear regression, linear discriminant analysis (LDA) and Gaussian Naive Bayes.

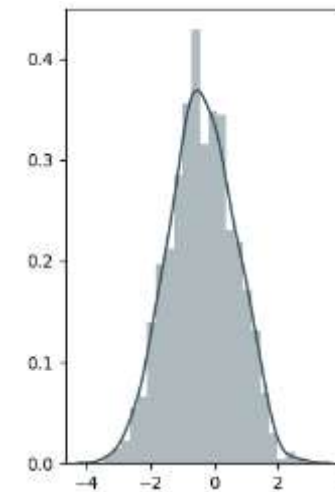
$$x' = \frac{x - x_{mean}}{x_{max} - x_{min}}$$

where x' is the normalized value.

Original data



Normalized data

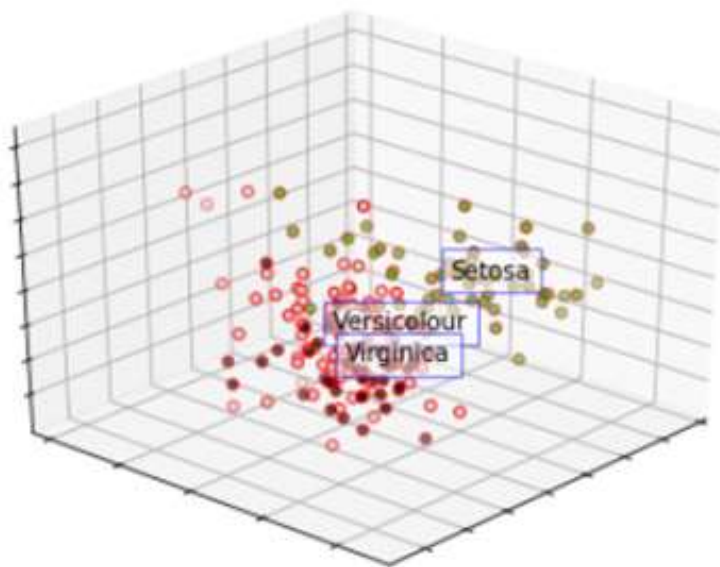


Step 2: Prepare data

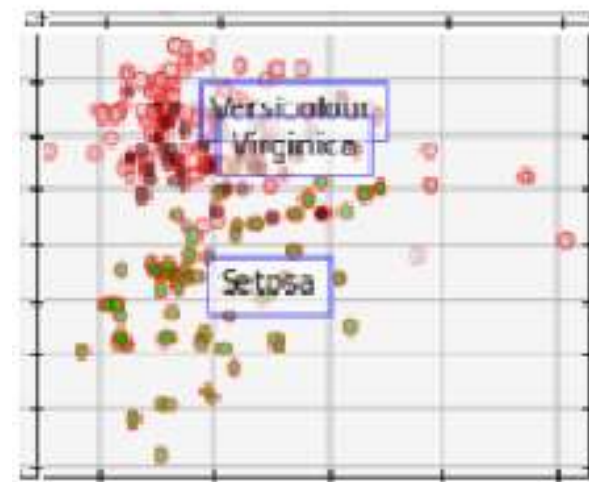
Dimensionality reduction

- is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension.

Original data – 3D



Data – 2D



Step 3: Analyze

“ Always remember: No perfect model to solve all problems kinds!”

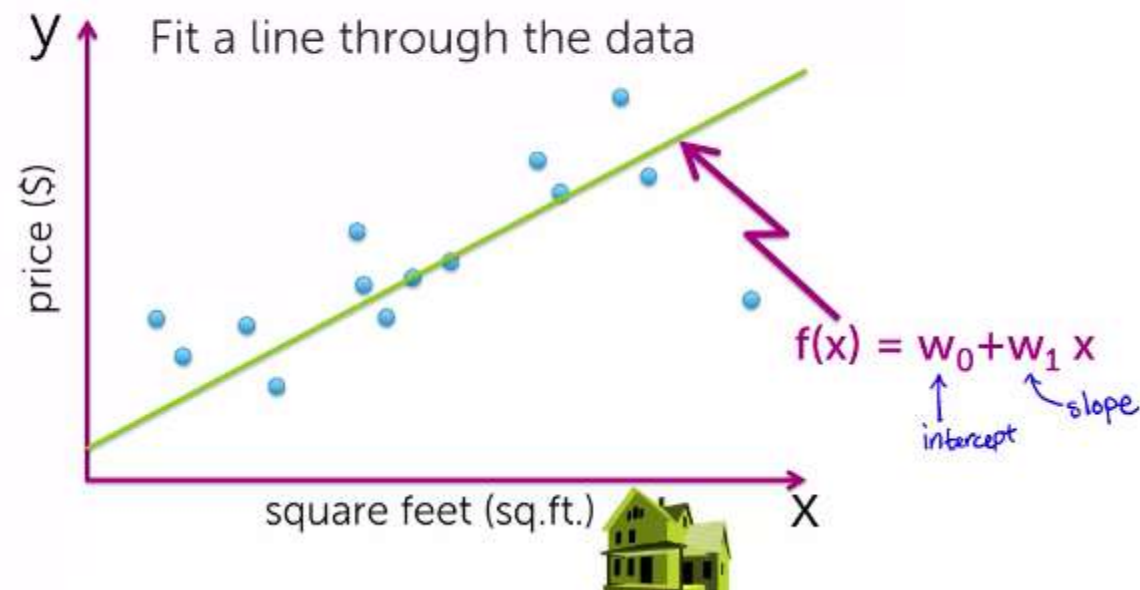
- Data analysis involves building a **model** from your data, which is called input data. The input data is used by the analysis technique to build a model.
- 4 categories of models:
 - Regression
 - Classification
 - Clustering
 - Association

Step 3: Analyze

Regression

- You have data about houses:
 - Size of the house in square feet (X1)
 - Number of rooms in the house (X2)
 - City where the house is located (X3)
 - Age of the house (X4)
 - Price
- Predict the price of a house

Use a **linear** regression model

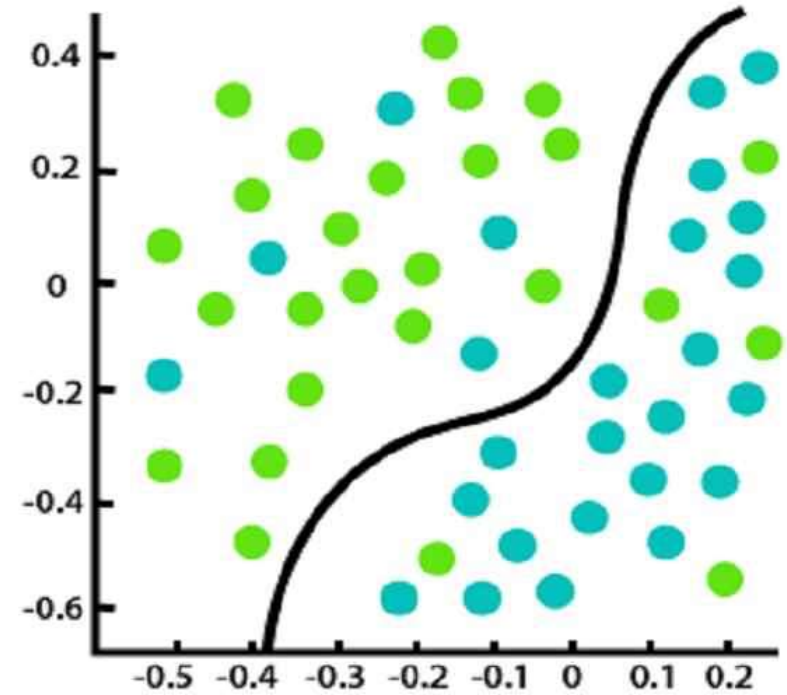


The outcome is continuous
(numerical)

Step 3: Analyze

Classification

- Predicting tumor cells as benign or malignant
- Classifying credit card transactions as legitimate or fraudulent
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil
- Categorizing news stories as finance, weather, entertainment, sports



The outcome is category

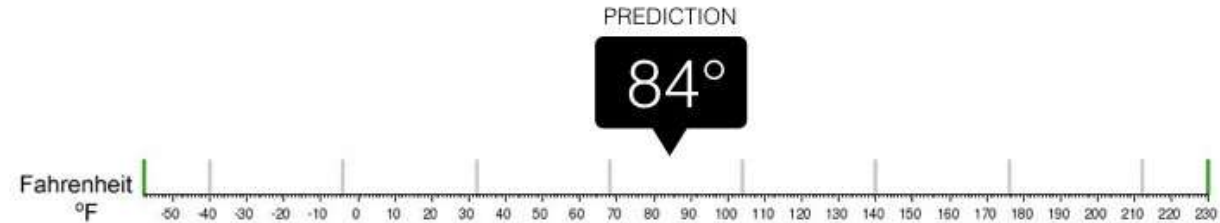
Step 3: Analyze

Classification vs Regression



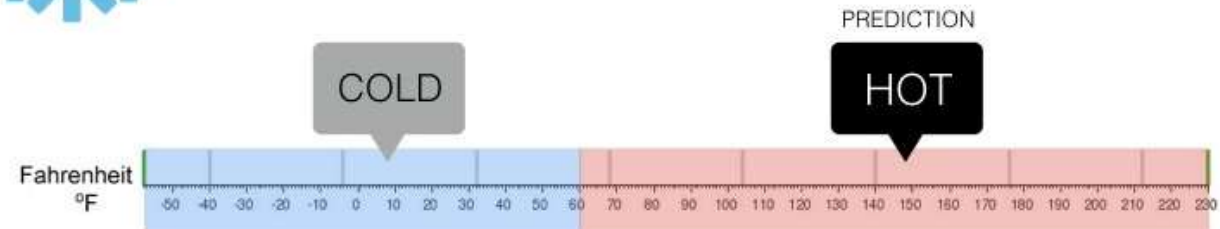
Regression

What is the temperature going to be tomorrow?



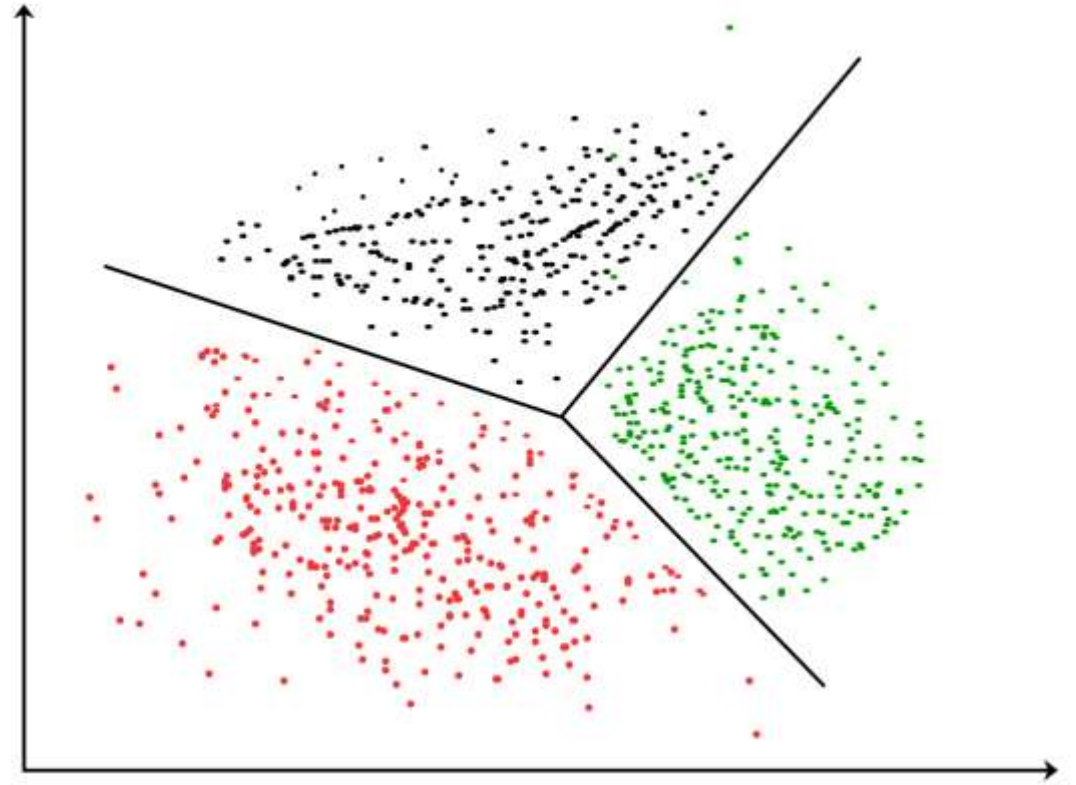
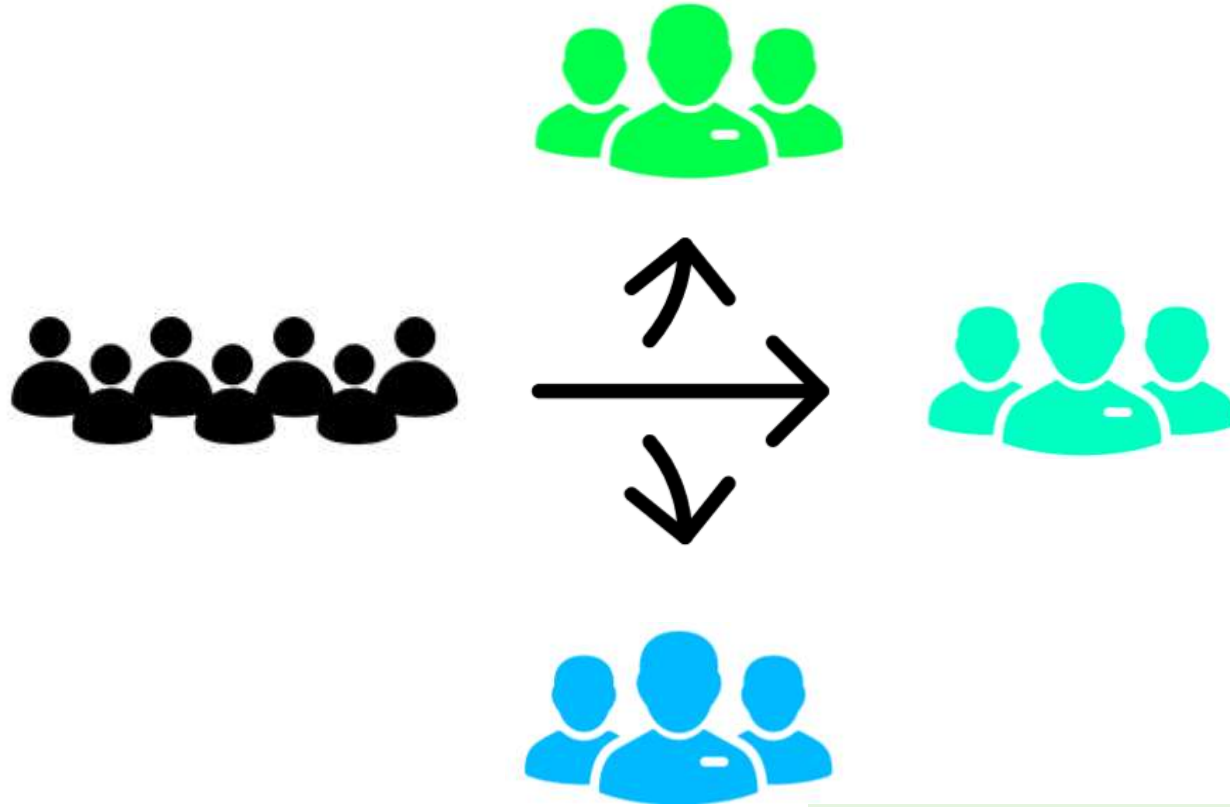
Classification

Will it be Cold or Hot tomorrow?



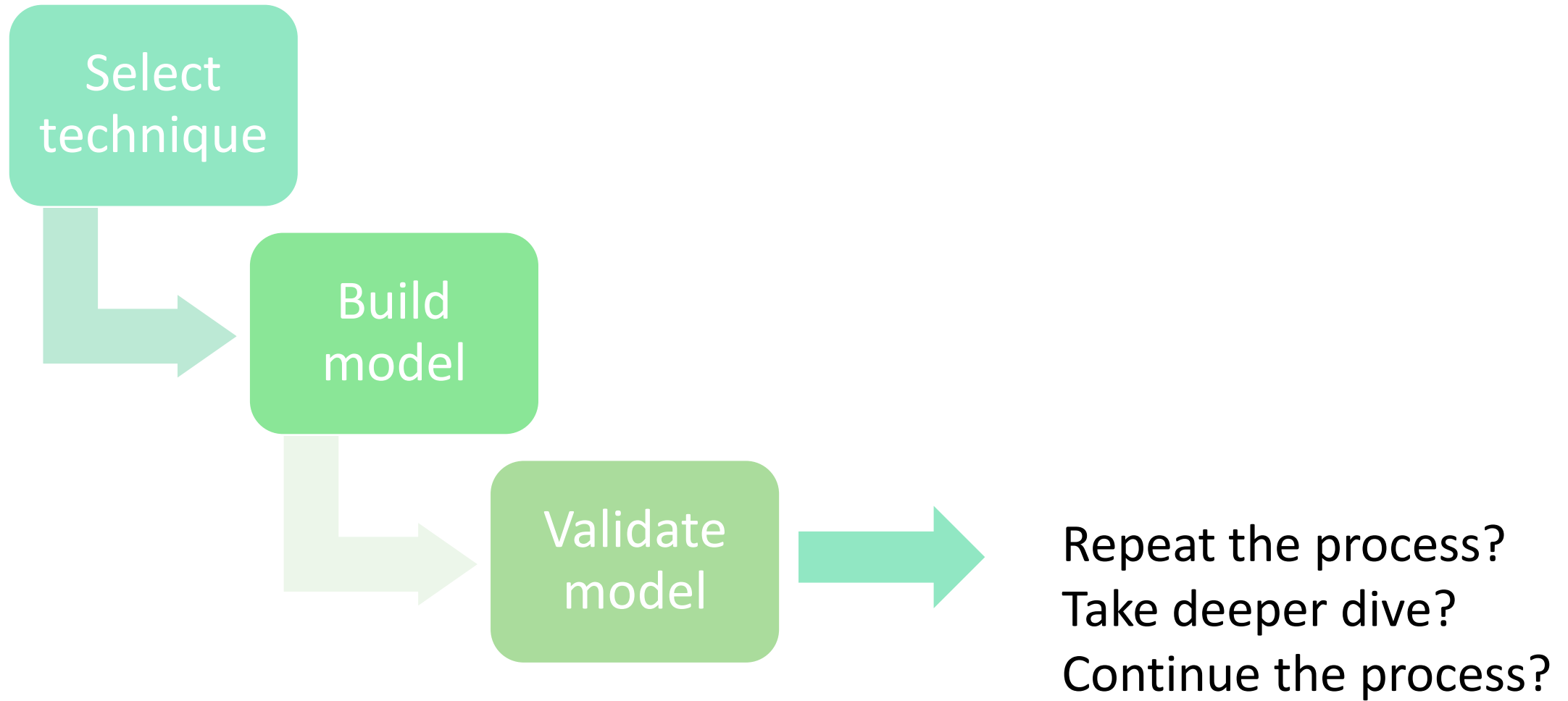
Step 3: Analyze

Clustering



Organize data into groups (clusters)

Step 3: Analyze



Step 4: Report

“Communication is one of the most underrated skills a data scientist can have. “

- Translate your findings into business concerns
- Use visualizations whenever applicable
- Focus on the result

Step 4: Report

Why visuals?

- A picture is worth a thousand words
- When creating a visual, always remember:
 - Less is more effective
 - Less is more attractive
 - Less is more impactful

Check this link: <https://www.darkhorseanalytics.com/blog/data-looks-better-naked>

Step 5: Act

- Ultimately, you've been called upon to create a solution throughout the data science process. After a proper communication, decision will be made:
 - Act on your proposals
 - Do not act on your proposals, then revisit the process!