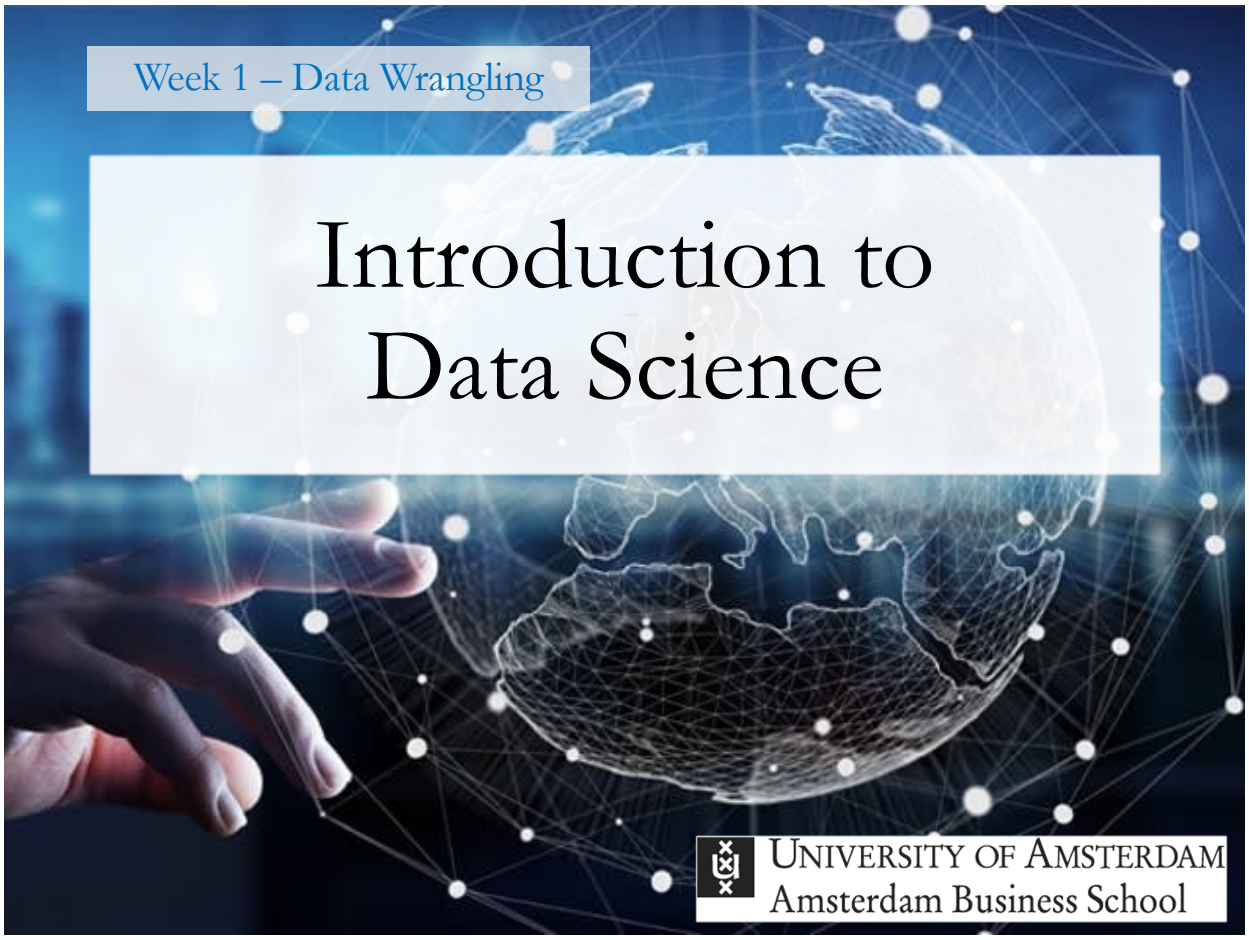


Contents

About me

Reza Mohammadi

Reza Mohammadi

- Assistant Professor in Business Analytics group at UvA
- Postdoc in Data Science at JADS
- PhD in Statistics at University of Groningen

Course coordinator:

- Data Wrangling - Minor Amsterdam Data Science & AI
- Data Analytics - Executive Master of Finance and Control
- Business Analytics - MBA

University homepage: https://www.uva.nl/profile/a.mohammadi

Email: A.Mohammadi@uva.nl

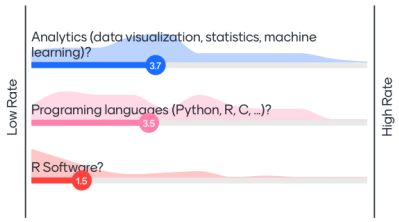
Office: M 4.11

About you

Mentimeter

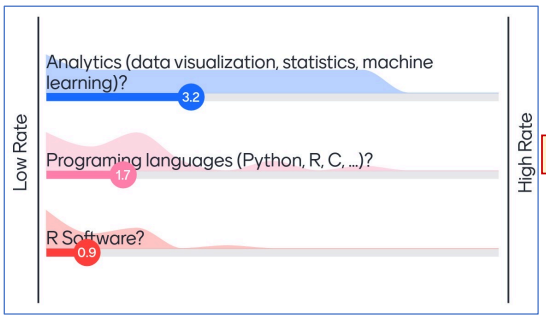
Go to **www.menti.com** and use the code **2238 6901**

How would you rate your knowledge level in the following elements?



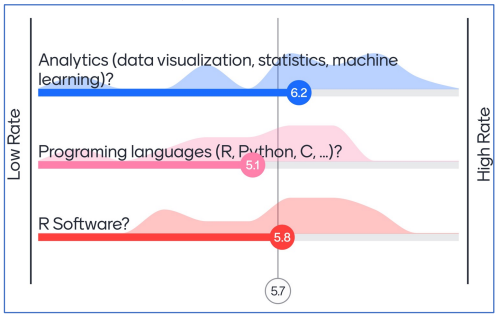
My job

How would you rate your knowledge level in the following elements?



First Lecture

Last lecture



My job is to take you form location A to location B

About you

- Your expectation about this course and "Data Science"
- Which parts (Analytics, Business, or Software) of Data Science do you think is important for your career?
- Do you have any exciting project(s) that you could use as a case study for this course?
- Did you watch the YouTube videos on Canvas about R & R-markdown? Are they useful to you?

Objectives

On completion of this course, you will be able to:

- explore and analyze different datasets and summarize the main characteristics of the data by using Exploratory Data Analysis approaches;
- derive and apply various hypotheses testing methods to address a research question and make informed decisions;
- apply the basic principles underlying statistical analysis methods, such as *regression analysis*;
- extract information from large datasets by using advanced Machine Learning methods such as *Decision Tree*, *Random Forest*, and *Regression Analysis*;
- apply ***R*** software as needed to reach the above objectives.

Course Organization

Week	Data	Subject	Material
Week 1	6 Sep	- Introduction to Data Science - Introduction to R and R-Markdown	Larose: chapters 1 & 2
Week 2	13 Sep	Exploratory Data Analysis	Larose: chapter 3
Week 3	20 Sep	Statistical Inference & Hypothesis Testing	Larose: chapters 4 & 5
Week 4	27 Sep	Classification & Model Evaluation	Larose: chapters 6 & 7
Week 5	4 Oct	Decision Trees	Larose: chapter 8
Week 6	11 Oct	Regression Analysis	Larose: chapter 5

Course Organization

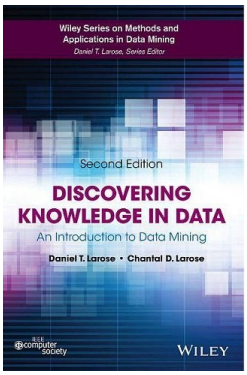
Weekly session:

- Lecture: to cover theory.
- Computer-lab: the theory will be practiced in applications by hand-on programming in R software.

- Your task: Study the assigned literature every week; see course schedule. Do the weekly exercises.
- Attending to lectures and computer-labs are highly recommended but not mandatory.
- Submitting weekly exercises and main project are mandatory.
- I will not accept assignments by email, just online in Canvas.
- In principle the lectures and computer-labs will be fully on-location. Please do not count on Zoom/Team.
- Lectures will not be recorded.
- Slides will be available on Canvas on each *Friday* before lectures.

References

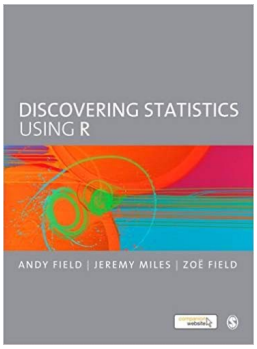
Book: Discovering Knowledge in Data (Larose).



- Bring the subject to life: by working in real data sets.
- Emphasis is on using theory in practice.
- Each chapter has a R-zone for R code.
- The book is intended as a reference source
- We cover chapters 1 till 8.
- You can access to the eBook for free by connect to the internet with the UvA or eduroam network.

Optional:

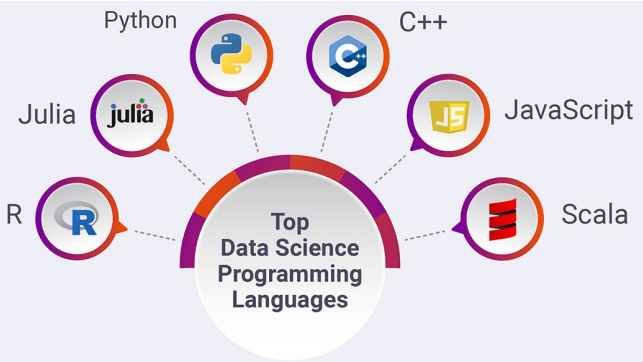
Book: Discovering Statistics Using R (Field et al.):



- It has a nice introduction to both basic statistics & R.
- It's self-study.

Why R Software?

- R is free and Open-source Tool.
- R is one of the leading tools for Data Science, Statistics, and Machine Learning.
- For data-driven businesses, lack of Data Scientists is a huge concern. Companies are using R as their core platform and are recruiting trained R programmers.
- R contains actual machine and statistical techniques; new techniques are made available in R very quickly.



Introduction to R

YouTube link: https://youtu.be/_V8eKsto3Ug

R Markdown

YouTube link: https://youtu.be/DNS7i2m4sB0

Grating

- **6 Weekly exercises (12%)**: would be done individually or in a group with a maximum of 3 individuals. The grade will account for 12% (2% for each) of the final grade.
- **Main project (28%):** would be done individually or in a group with a maximum of 3 individuals. The grade will account for 28% of the final grade.
- **Exam (60%):** is **multiple-choice** and is a closed book exam. Only a non-graphical calculator can be used. The result will account for 60% of the final grade.

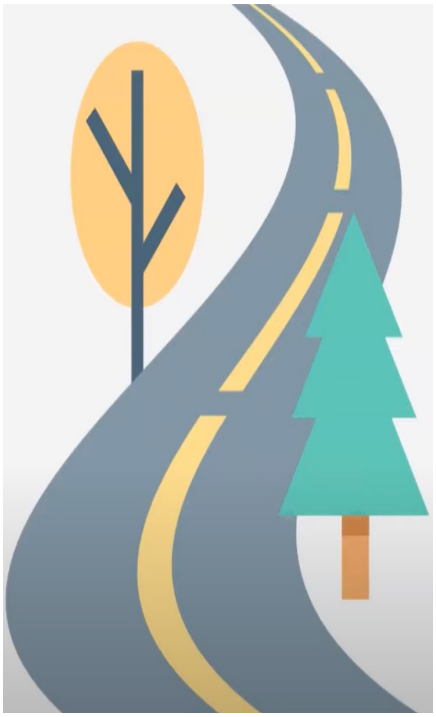
Study: Slides + Weekly Exercises + main reference book.

The main spirit behind the course structure (the weekly exercises and main project) is "**Learning by doing it!**"

Contents

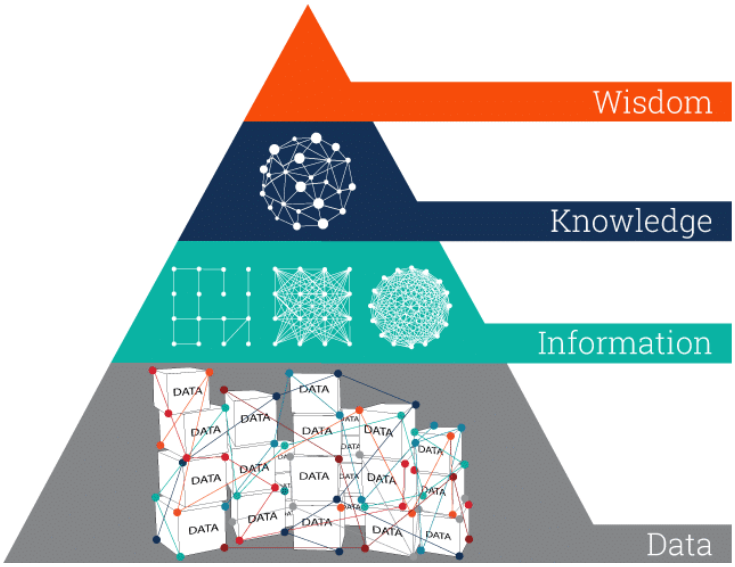
Introduction to Data Science

YouTube link: https://youtu.be/X3paOmcrTjQ



Wisdom Hierarchy

The more we enrich our data with meaning and context, the more knowledge and insights we get out of it so we can take better, informed and data-based decisions.

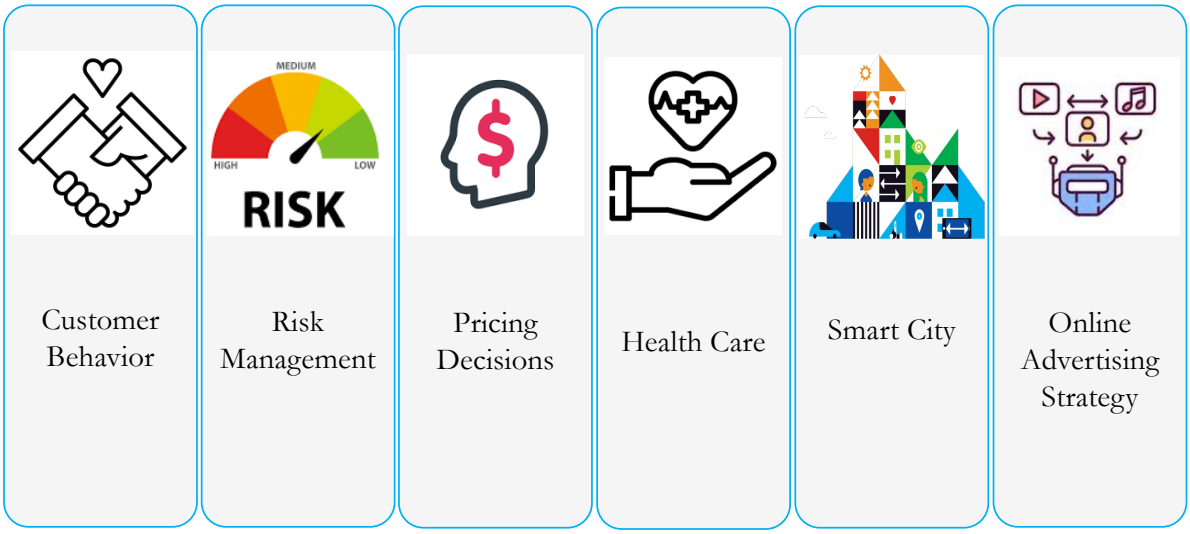


Each step up the pyramid answers questions about and adds value to the initial data.

https://www.ontotext.com/knowledgehub/fundamentals/dikw-pyramid

Data Science Applications

No matter which industry you work in, there is room for analytics to add value:



Customer Behavior | Risk Management | Pricing Decisions | Health Care | Smart City | Online Advertising Strategy

What is Data Sciences?

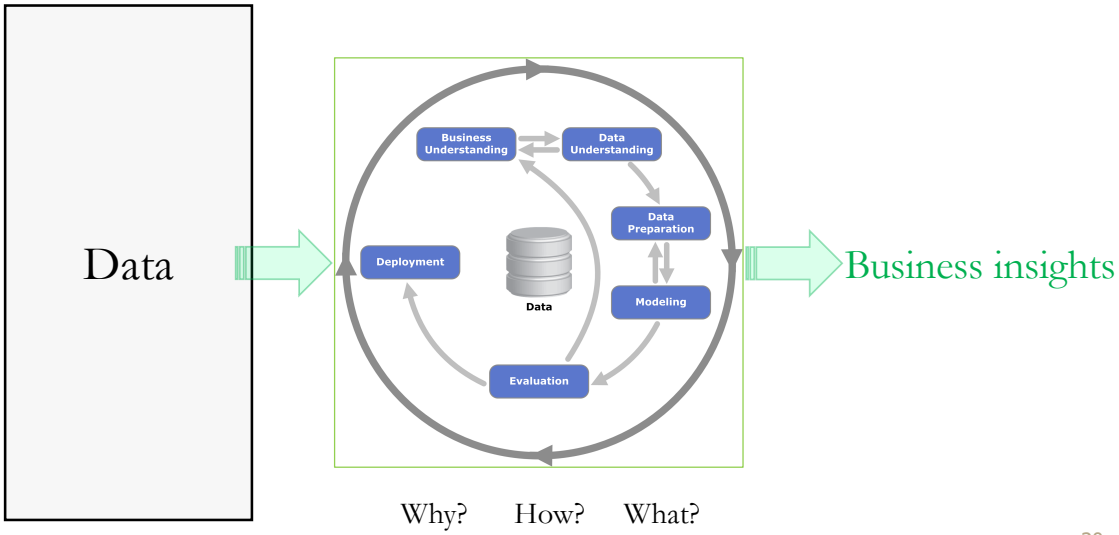
Data Science uses tools and techniques to turn data into meaningful business insights.

Goal: Use data to take better business decisions.

What is Data Science?

Data Science uses tools and techniques to turn data into meaningful business insights.

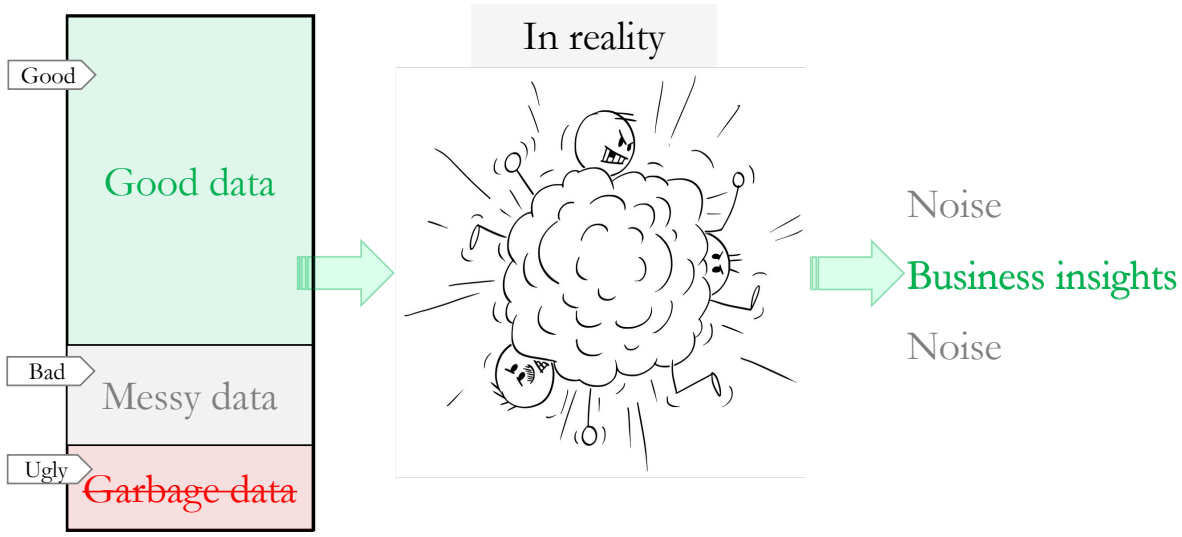
Goal: Use data to take better business decisions.



What is Data Science?

Data Science uses tools and techniques to turn data into meaningful business insights.

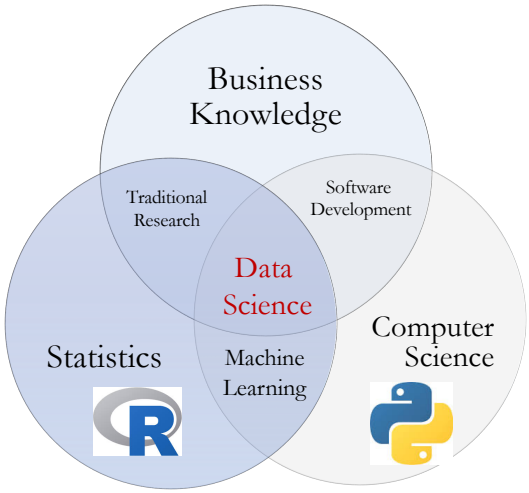
Goal: Use data to take better business decisions.



What is Data Science?

A lot of different terms are around analytics likes:

- Data Analytics
- Business Intelligence
- Data Science
- Decision Science
- Data Mining
- ….



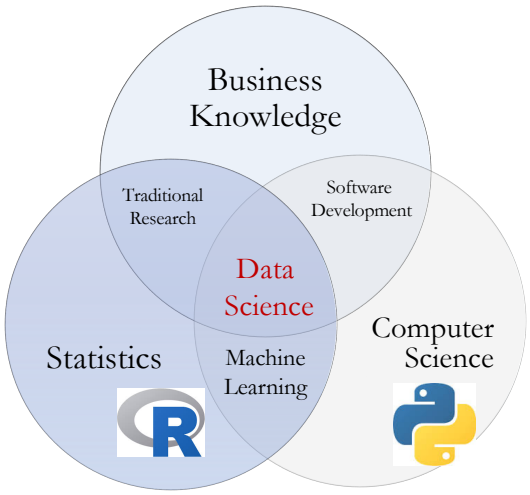
Although there are some differences, all of them support the same goal: "turning data into useful insight".

Data Science

Data science combines data analysis, statistics, machine learning, and related methodology in order to manage and understand the data deluge associated with the emergence of information technology.

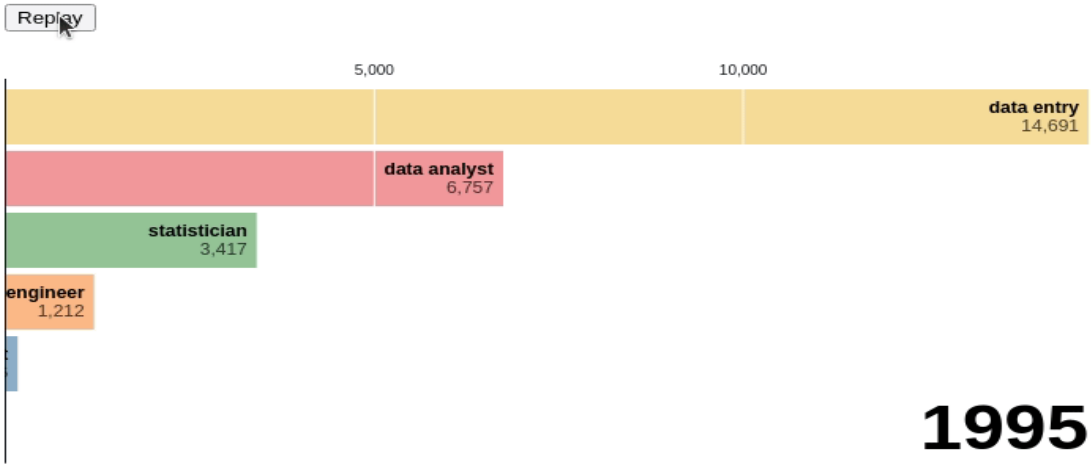
Data scientists are tasked with presenting digital information in a way that depicts its practical value in data-driven decision-making.

however, they don't typically endeavor to solve specific questions in the way that business analysts do when seeking out business analytics insights.

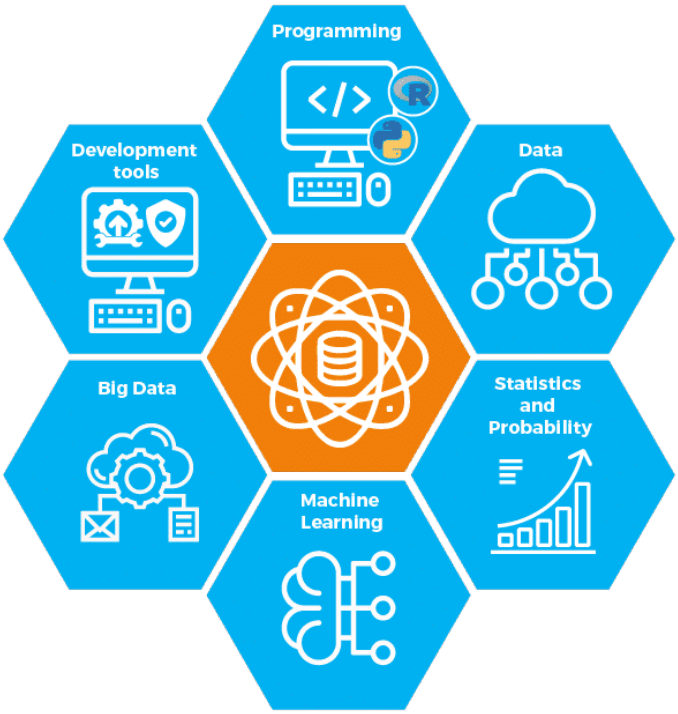


Data Science

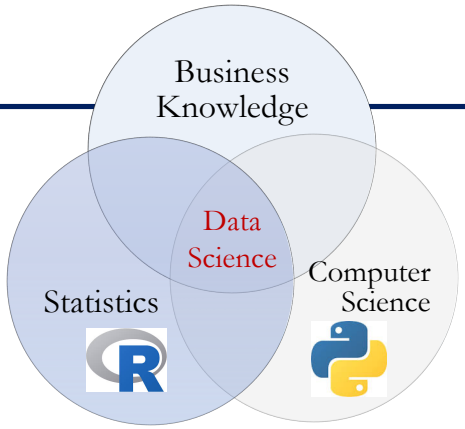
Rankings of different data-related positions:

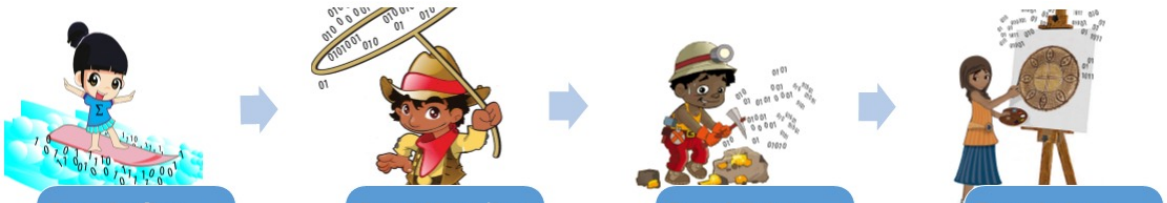


Data Science Components



Data Wrangling





Data Surfing
- Ponder Questions
- Explore Potential Data

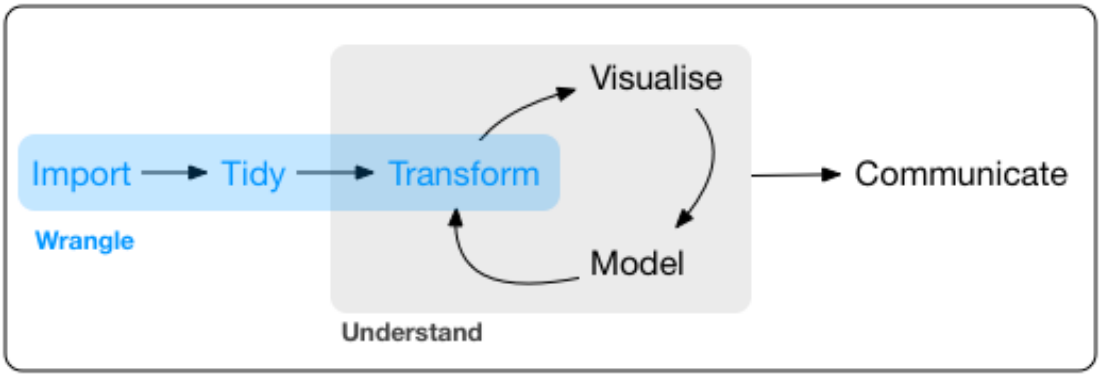
Data Wrangling
- Gather Data
- Clean the Data
- Connect Data
- Transform Data

Data Mining
- Choose Algorithm
- Use Algorithm
- Test Algorithm
- Refine Algorithm

Data Artistry
- Use Discovery
- Share Discovery
- Present Discovery

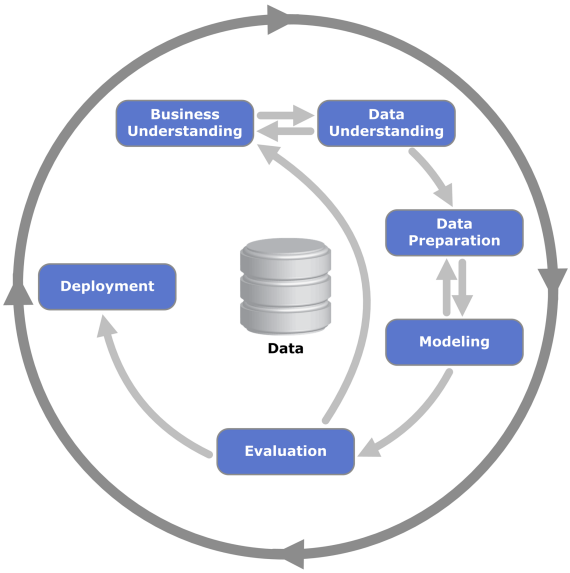
Data wrangling is …

- The process of transforming "raw" data into data that can be analyzed to generate valid actionable insights.
- Getting your data into a form that is natural to work with.
- Tidying and transforming data.



Contents

6 Steps of Data Science

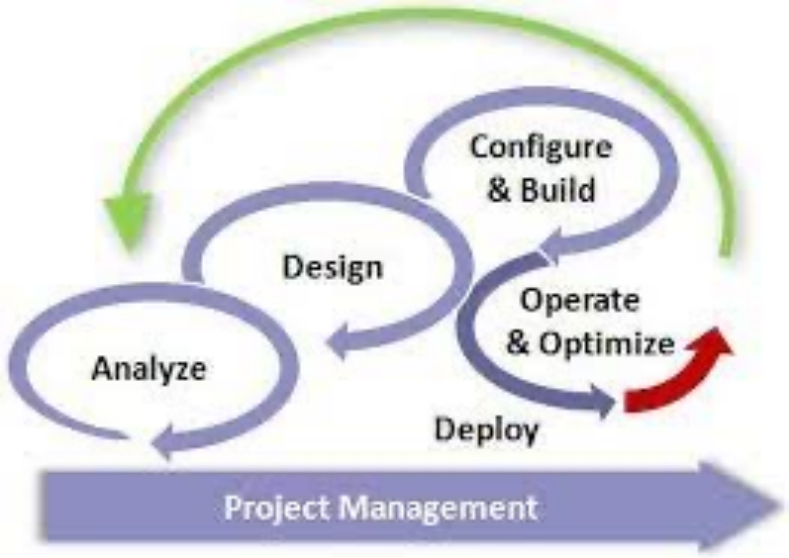


6-Step of Data Science process is a general problem-solving strategy of business unit research unit, as adaptive life cycle.

Also, known as Cross-industry standard process for data mining (CRISP-DM)

6 Steps of Data Science

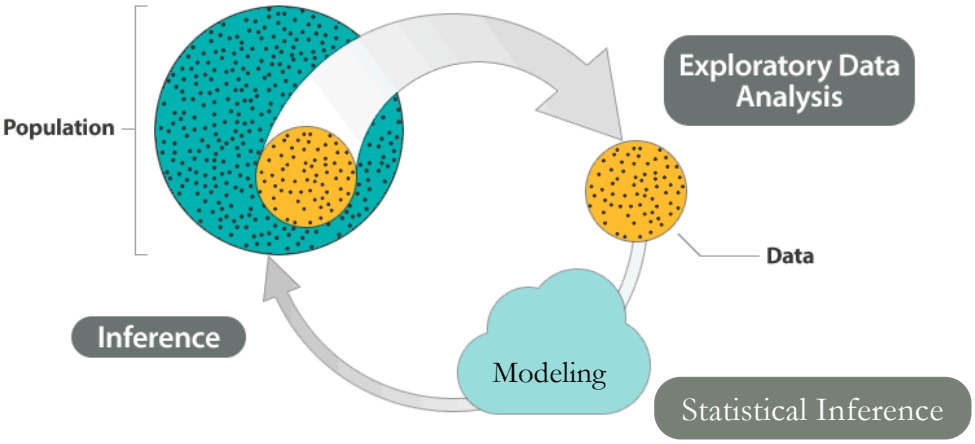
Analytics Solutions Unified Method



https://arxiv.org/pdf/1907.04461.pdf

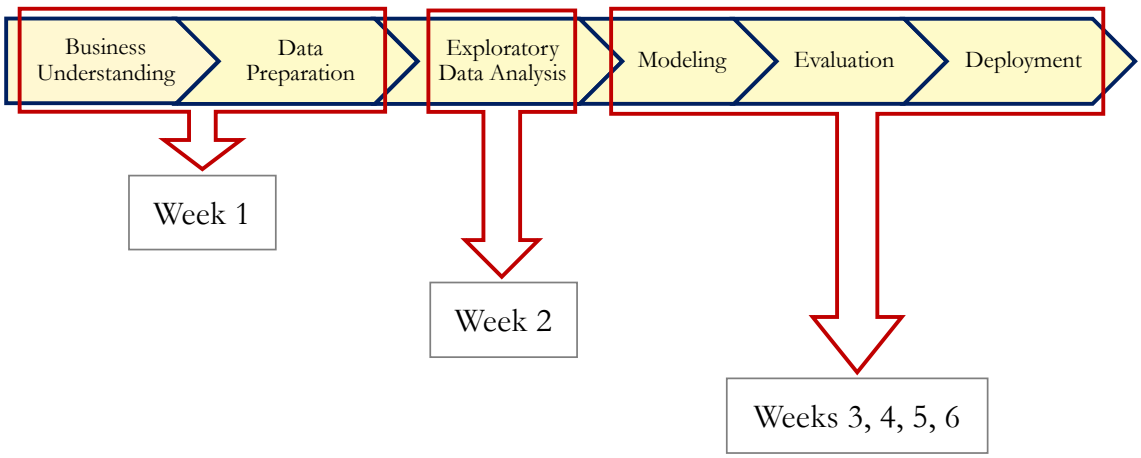
6-Step of Data Science Process

6-Step of Data Science process is a general problem-solving strategy of business/research unit, which is adaptive life cycle.



6 Steps of Data Science

6-Step of Data Science process is a general problem-solving strategy of business unit, which is adaptive life cycle.



6-Step Data Science Process

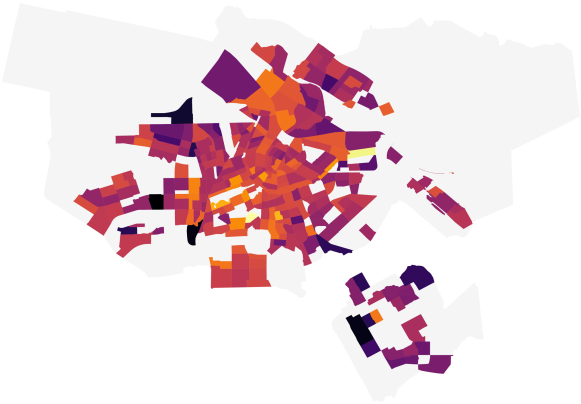
1) Business/Research Understanding Phase
   - Define project requirements and objectives.
   - Translate objectives into a data analytics problem definition.
   - Prepare a preliminary strategy to meet objectives.
2) Data Preparation Phase
   - Clean and prepare data so it is ready for modeling tools.
   - Perform transformation of certain variables, if needed.
3) Exploratory data analysis
   - Analyzing data to summarize their main characteristics, often with. visual methods.
   - Seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

6-Step Data Science Process

4) Modeling Phase
   - Select and apply one or more modeling techniques.
   - Calibrate model settings to optimize results.
5) Evaluation Phase
   - Evaluate one or more models for effectiveness.
   - Determine whether defined objectives are achieved.
6) Deployment Phase
   - Make use of the models created.
   - Simple deployment example: generate a report.
   - In businesses, the customer often carries out the deployment based on your model.

6-Step Data Science Process

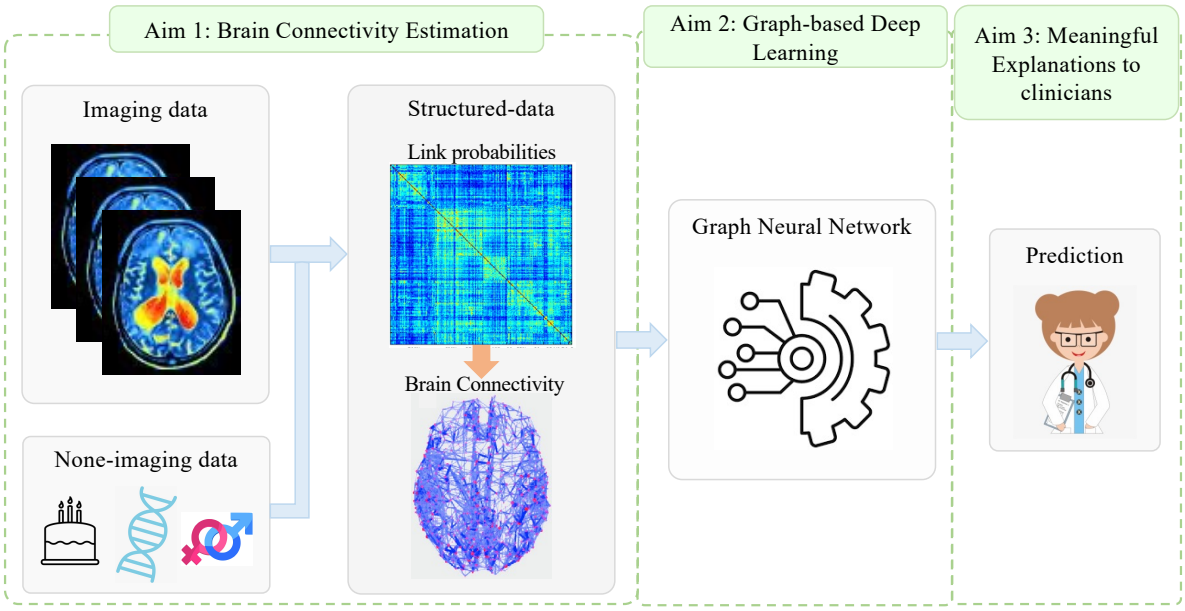
Example: Rental Price Prediction in Amsterdam using machine learning techniques.



The colors represent how expensive are to rent the houses in these areas in Amsterdam.

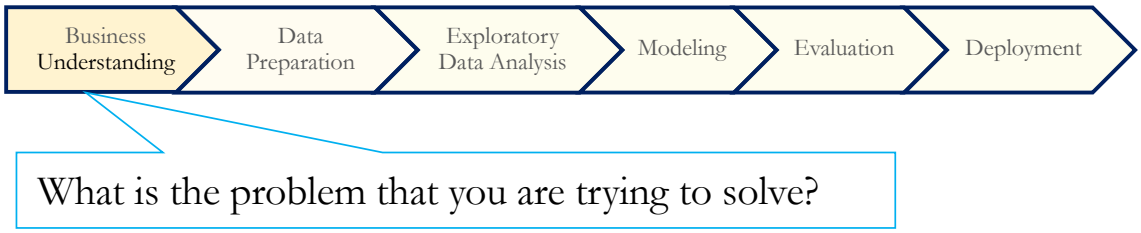
6-Step Data Science Process

Example: Machine learning techniques for Early Diagnosis of Alzheimer's



Contents

Business Understanding



In Business understanding phase we basically:

1. Understands the business process;
2. Define project requirements and objectives;
3. Translate objectives into a data analytics problem definition;
4. Prepare a preliminary strategy to meet objectives.

Business Understanding

Example: Tech Company

Intl.Mins	Intl.Calls	Intl.Charge	CustServ.Calls	Churn
10.0	3	2.70	1	False
13.7	3	3.70	1	False
12.2	5	3.29	0	False
6.6	7	1.78	2	False
10.1	3	2.73	3	False
6.3	6	1.70	0	False
7.5	7	2.03	3	False
7.1	6	1.92	0	False
8.7	4	2.35	1	False
11.2	5	3.02	0	False
12.7	6	3.43	4	True
9.1	5	2.46	0	False
11.2	2	3.02	1	False

Churn Data

Business insights:

Predict behavior to retain customers.

This dataset is availed in the R package "liver":

Business Understanding

Example: Bank marketing campaigns

campaign	pdays	previous	poutcome	deposit
1	330	1	failure	no
4	-1	0	unknown	no
1	-1	0	unknown	no
2	176	3	failure	no
1	330	2	other	no
2	-1	0	unknown	no
2	-1	0	unknown	no
1	147	2	failure	no
1	-1	0	unknown	no
2	-1	0	unknown	no
2	-1	0	unknown	no

Bank Data

Business insights:

Predict if client will subscribe a term deposit.

Business Understanding

Example: Credit Risk

nr_loans	age	marital_status	income	risk
3	34	other	28060.70	bad loss
2	37	other	28009.34	bad loss
2	29	other	27614.60	bad loss
2	33	other	27287.18	bad loss
2	39	other	26954.06	bad loss
2	28	other	26271.86	bad loss
3	28	other	40445.00	bad loss
2	25	other	23888.30	bad loss
2	41	other	35249.25	bad loss
2	26	other	23518.88	bad loss
2	30	other	23575.22	bad loss
2	23	other	23447.54	bad loss
2	47	other	23058.02	bad loss
2	23	other	22039.40	bad loss

Risk Data

Business insights:

Predict weather customer is Credit Risky or Credit Worthy

Contents

Data Preparation

- Clean and prepare data so it is ready for modeling tools.
- Perform transformation of certain variables, if needed.

Raw data are often unprocessed, incomplete, noisy and may contain:

- Obsolete/redundant fields
- Missing values
- Outliers
- Data in a form not suitable for data analysis
- Values not consistent with policy or common sense

Data Preprocess

For data analytics purposes, database values must undergo data cleaning and data transformation.

Minimize GIGO (Garbage In → Garbage Out).

- IF GIGO is minimized → THEN Garbage results Out from model is minimized.

Effort for data preparation ranges around 10%-60% of data analysis process – depending on the dataset.

Data?

features / fields / attributes/ variables

ID	age	work class	education	Marital status	occupation	gender	income
101	25	Private	7	Married	Sales	F	<=50K
102	38	Private	9	Separated	Other-service	?	<=50K
103	34	Gov	12	Never Married	Prof-specialty	?	>50K
104	44	?	10	Married	Adm-clerical	M	<=50K
105	48	Self-emp	6	?	Sales	F	>50K

- Here we have 5 unit or observation which in this dataset are customers.
- For each unit (customer/observation) we have measured 8 features (variables, attributes).

What type of variables we have?

➢Numerical:

- Continuous (entities get a distinct score), e.g. temperature, body length.
- Discrete (counts), e.g.: number of defects.

➢Categorical (entities are divided into distinct categories):

- Binary variable (two outcomes), e.g. dead or alive.
- Ordinal variable, e.g. bad, intermediate, good.
- Nominal variable, e.g. whether someone is an omnivore, vegetarian or vegan.

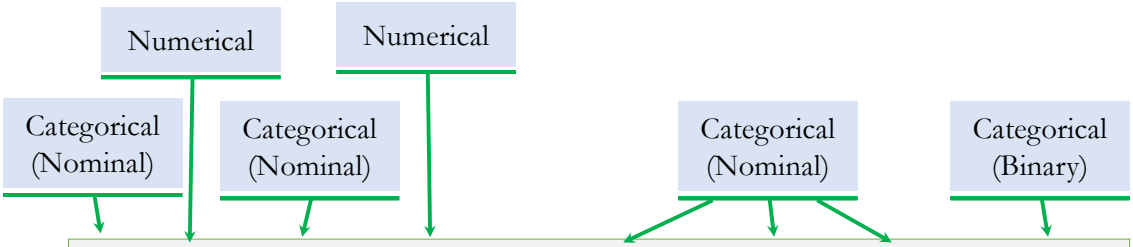
What type of variables we have?

What type of variables are they?

ID	age	work class	education	Marital status	occupation	gender	income
101	25	Private	7	Married	Sales	F	<=50K
102	38	Private	9	Separated	Other-service	?	<=50K
103	34	Gov	12	Never Married	Prof-specialty	?	>50K
104	44	?	10	Married	Adm-clerical	M	<=50K
105	48	Self-emp	6	?	Sales	F	>50K

What type of variables we have?

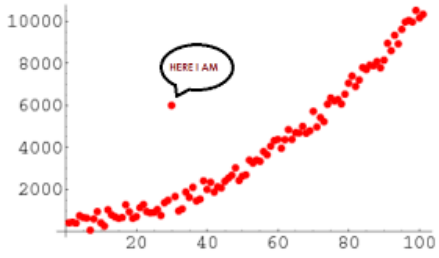
What type of variables are they?



ID	age	work class	education	Marital status	occupation	gender	income
101	25	Private	7	Married	Sales	F	<=50K
102	38	Private	9	Separated	Other-service	?	<=50K
103	34	Gov	12	Never Married	Prof-specialty	?	>50K
104	44	?	10	Married	Adm-clerical	M	<=50K
105	48	Self-emp	6	?	Sales	F	>50K

Outliers

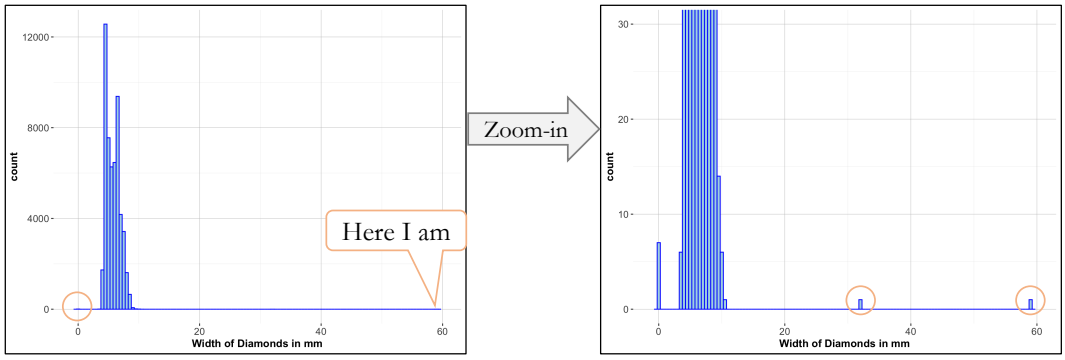
- Outliers or unusual values are observations that are unusual and go against the trend of the remaining data.
- Outliers may represent errors in data entry.
- Outlines may suggest important new science.
- Even if an outlier is a valid data point, certain statistical methods are very sensitive to outliers and may produce unstable results.
- Often, it is easiest to identify outliers by graphing the data.



Identify Outliers

Histogram: Outlier visualize in the histogram of <u>numeric</u> feature values; they may be the values on the tails.

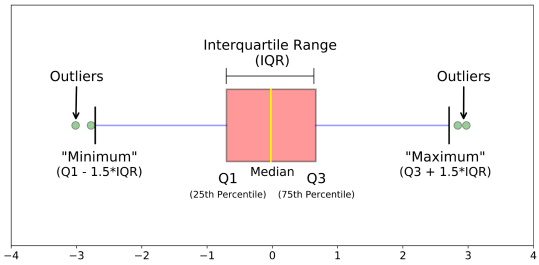
Example: Histogram shows the width of diamonds.



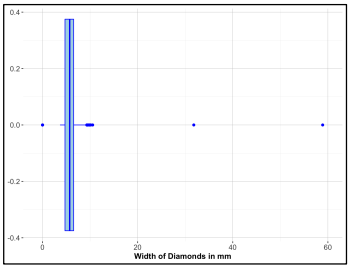
- Some of diamonds have width of zero and some width of 60 mm!
- Should we doubt the validity of this value?

Identify Outliers

Boxplot: We can observe the outliers using Boxplot. Boxplot represents the distribution of the feature.



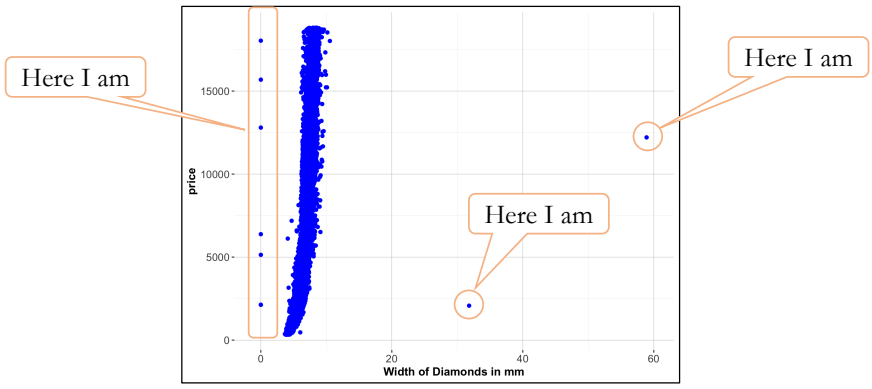
Example: Boxplot shows the width of diamonds.



Identify Outliers

Two-dimensional scatter plots help determine outliers in more than one variable.

Example: Scatter plot of diamond prices vas widths.



- One diamonds with width of 60 mm price of around 13K?

Handling outliers

- Drop the entire row with the outliers. (NOT recommended)

```
diamonds2 = diamonds %>%
            filter( between( y, 3, 20 ) )
```

- Replace outlines with missing values. (Recommended)

```
diamonds2 = diamonds %>%
            mutate( y = ifelse( y < 3 | y > 20, NA, y ) )
```

In this way we treat the outliers as missing values. Note, in R, we show missing values with NA (Not Available).

Handling Missing Data

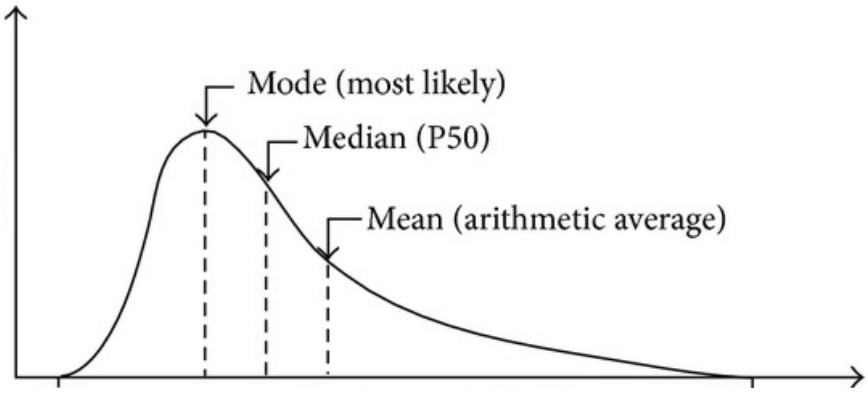
Missing values pose problems to data analysis methods.

- Delete Records Containing Missing Values?
  - Dangerous, as pattern of missing values may be systematic.
  - Valuable information in other fields lost.
  - As much as 80% of the records lost if 5% of data values are missing from a data set of 30 variables.
- Imputation: is a method to fill in the missing values with estimated ones:
  - Imputation with mean/median/mode.
  - Imputation with random values.
  - Imputation using prediction models.

Handling Missing Data

Imputation with Mean/Mode/Median is one of the most frequently used methods. It consists of replacing the missing data for a given attribute by the mean or median (quantitative attribute) or mode (qualitative attribute) of all known values of that variable.

- Imputation with Mode is used for the categorical attributes.
- Imputation with Mean and Median is used for the numerical attributes.



Handling Missing Data

Imputation with random values: Replace missing values (NAs) with values randomly taken from underlying distribution.

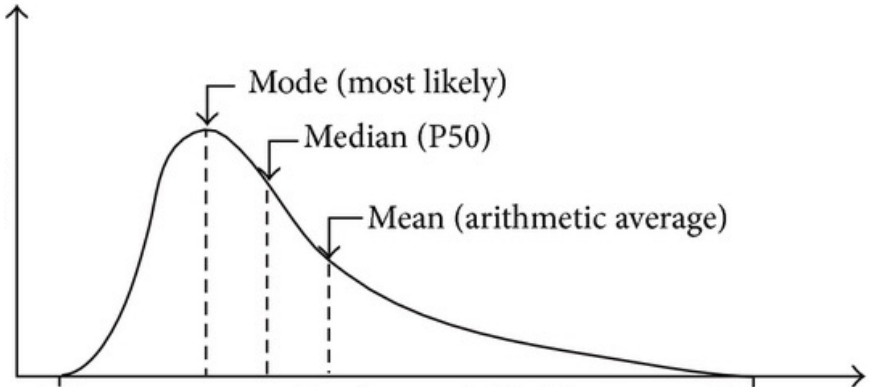
- Benefit: Measures of location and spread remain closer to original.

See the exercises for week 1 on Canvas for practical examples.

Measures of Center

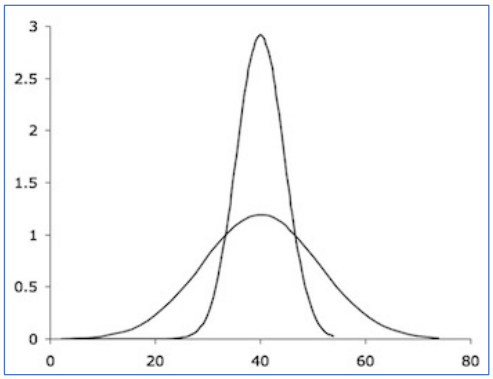
Estimate where the center of a particular variable lies. Most common measures of center are:

- Mean: Average of the valid values for a variable.
- Median: value in the middle, when data are sorted in order.
- Mode: value occurring with the greatest frequency.



Measures of Spread

Measures of location not enough to summarize a variable. For example, in the plot bellow, measures of center do not provide a complete picture.



Measures of spread complete the picture by describing how spread out the data values of each distribution are.

Typical measures of spread:

- Range (maximum – minimum).
- Standard Deviation – the "typical" distance between a field value and the mean:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n-1}}$$

Data Transformation

Variables tend to have ranges different from each other. Fer example, in "diamonds" dataset:

- carat range: [0.2, 5]
- Price of diamonds range: [326, 18823]

Some machine learning algorithms (i.e. the k-nearest neighbor algorithm) are adversely affected by differences in variable ranges.

Variables with greater ranges tend to have larger influence on the data model's results. Thus, numeric field values should be normalized.

Two of the prevalent methods will be reviewed

- Min-Max Normalization
- Z-score Standardization

Min-Max Normalization

Determines how much greater the field value is than the minimum value. Scales this difference by the field's range:

$$X^* = \frac{X - \min(X)}{\text{range}(X)} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

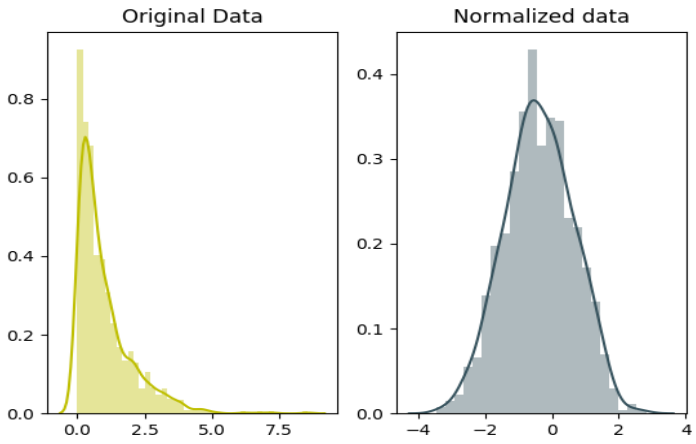
where X* refer to the normalized form of the variable X.

NOTE: Min-Max normalization will always range between 0 and 1.

Z-score Standardization

Widely used in statistical analysis. Takes difference between the field value and the field mean value. Scales this difference by the field's standard deviation.

$$X^* = \frac{X - \text{mean}(X)}{\text{SD}(X)}$$



Flag Variables

Some numerical methods require predictors to be numeric. For example, Regression requires recoding categorical variable into one or more flag variables.

A flag variable (as a dummy or indicator variable) is a categorical variable with only two values: 0 or 1.

Example: Categorical variable gender can be converted as:

If gender = female, then gender_flag = 1;

If gender = male. , then gender_flag = 0.

NOTE: If a categorical variable is nominal and has more than two categories ($\mathbf{k > 2}$), then we need to define $\mathbf{k - 1}$ dummy variables. For example, gender in three categories: 'female', 'male', and 'LGBT'.

Convert categorical variables to numeric

Why not transforming the categorical variable region into a single numerical variable? For example:

Region	Region_num
North	1
East	2
South	3
West	4

This is a common and hazardous error. The algorithm now assumes:

- The four regions are ordered
- West > South > East > North
- West is three times closer to South compared to north, etc.

This practice should be avoided, except with categorical variables that are clearly ordered, such as with a variable survey_response with values always, usually, sometimes, never.

Practical lab

- Exercises of Week 1 on Canvas > Week 1
- Online exercise on DataCamp:

  https://learn.datacamp.com/courses/reporting-with-rmarkdown

- Other useful links:

  https://rmarkdown.rstudio.com/lesson-1.html

  http://r4ds.had.co.nz/r-markdown.html