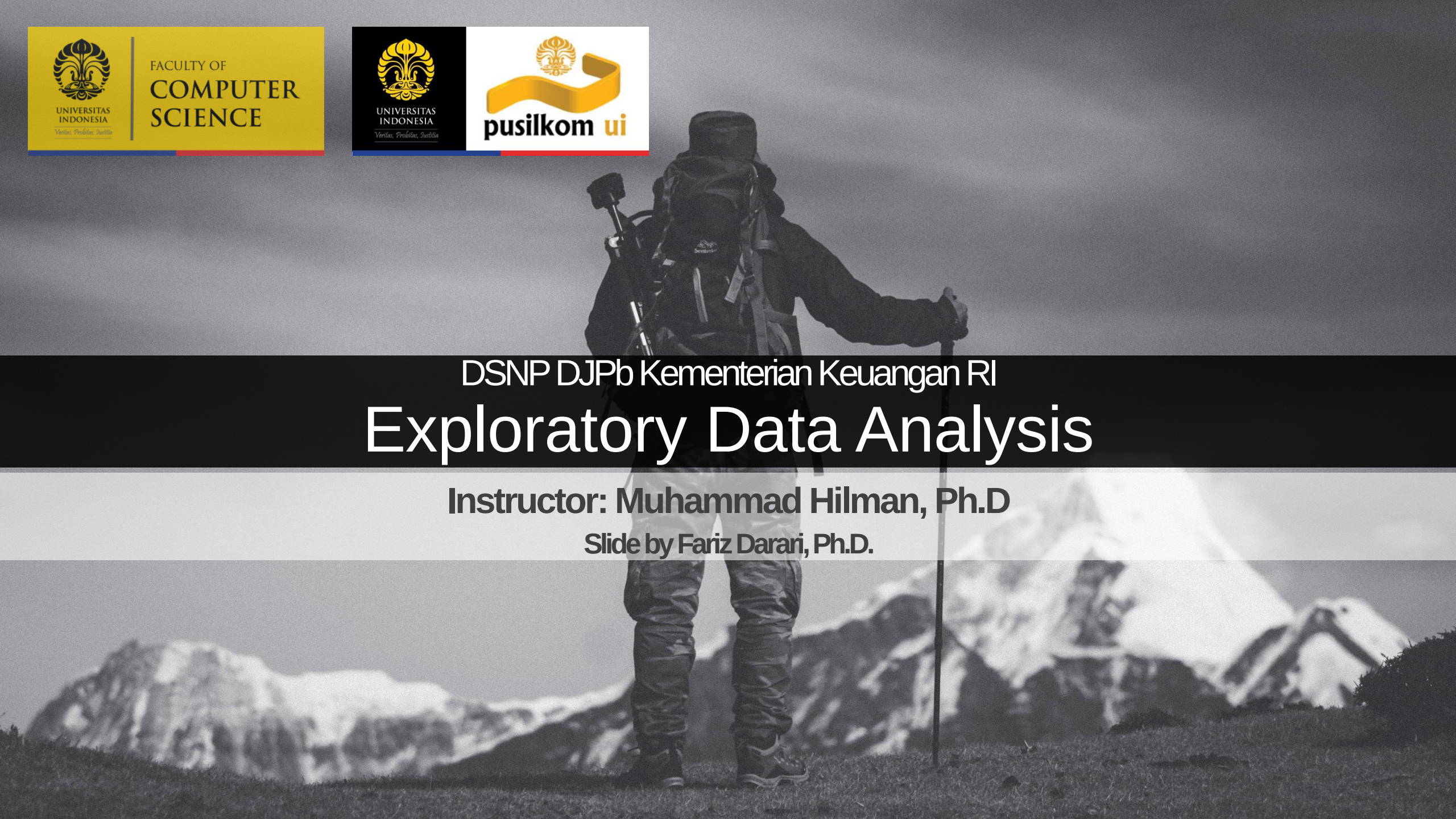


DSNP DJPb Kementerian Keuangan RI

Exploratory Data Analysis

Instructor: Muhammad Hilman, Ph.D

Slide by Fariz Darari, Ph.D.

Why love data science

- Data makes informed decisions – no more wild guessing!
- You can argue better backed up by data – no more losing an argument!
- It's exciting!
- It can be applied to ANY domain!
- **$$$$$**


Ask a Lot of Questions
Translate an ambiguous request into a concrete, well-defined problem
Identify business priorities & strategy decisions that will influence your work
01 Frame the Problem
02 Collect Raw Data
Identify All Available Datasets
Web, internal/ external databases, etc.
Extract Data Into Usable Format
.csv, .json, .xml, etc.
Identify Business Insights
Return back to the business problem
Visualize Your Findings
Keep it simple & priority-driven
Tell a Clear & Actionable Story
Effectively communicate to non-technical audiences
06 Commu-nicate Results
THE DATA SCIENCE PROCESS
03 Process the Data
Examine Data at a High-Level
Understand every column; identify errors, missing values & corrupt records
Clean the data
Throw away, replace, and/or filter corrupt /error prone / missing values
Create a Predictive Model
Use feature vectors from step #4
Evaluate & Refine Model
Perhaps return to step #2, 3, or 4
05 Perform In-Depth Analysis
04 Explore the Data
Play Around With the Data
Split, segment, & plot the data in different ways
Identify Patterns & Extract Features
Use statistics to identify & test significant variables

Exploration: Definition

- **Cambridge Dictionary**
  The activity of
  searching and finding out about something.

- **Oxford Dictionary**
  The action of exploring an **unfamiliar area.**

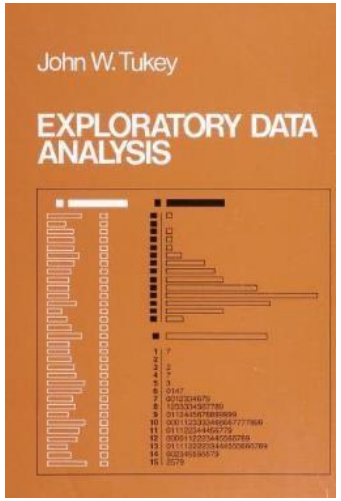
- **KBBI**
  Penjelajahan lapangan dengan tujuan memperoleh **pengetahuan lebih banyak (tentang suatu keadaan).**

EDA: Storytelling with data

Exploratory data analysis (EDA) is an approach:

- to **analyzing datasets**
- by **summarizing their main characteristics**
- often with **visual methods**.

The term EDA was coined by John W. Tukey in the book "Exploratory Data Analysis" in **1977.**



EDA: Take a peek at data

- EDA is a term for an initial analysis done with datasets.
- It's basically taking a peek at the data to understand more about what it represents and how to use it.
- It's often a precursor to more advanced data analytics techniques.

Any patterns you identify easily?

Months or Years	Hours of Sleep per Day	Number of Naps
0-3 months	14-17	3-5
4-12 months	12-16	2-3
1-2 years	11-14	1-2
3-5 years	10-13	0-1
6-12 years	9-12	0
13-18 years	8-10	0
18+ years	7+	0

Any patterns you identify easily?

Months or Years	Hours of Sleep per Day	Number of Naps
0-3 months	14-17	3-5
4-12 months	12-16	2-3
1-2 years	11-14	1-2
3-5 years	10-13	0-1
6-12 years	9-12	0
13-18 years	8-10	0
18+ years	7+	0

Any patterns you identify easily?

Status	Pengeluaran per Bulan
Miskin	354000
Rentan	354000-532000
Calon Kelas Menengah	532000-1200000
Kelas Menengah	1200000-6000000
Kelas Atas	>6000000

Status Sosial Ekonomi Penduduk Indonesia

04-02-2020

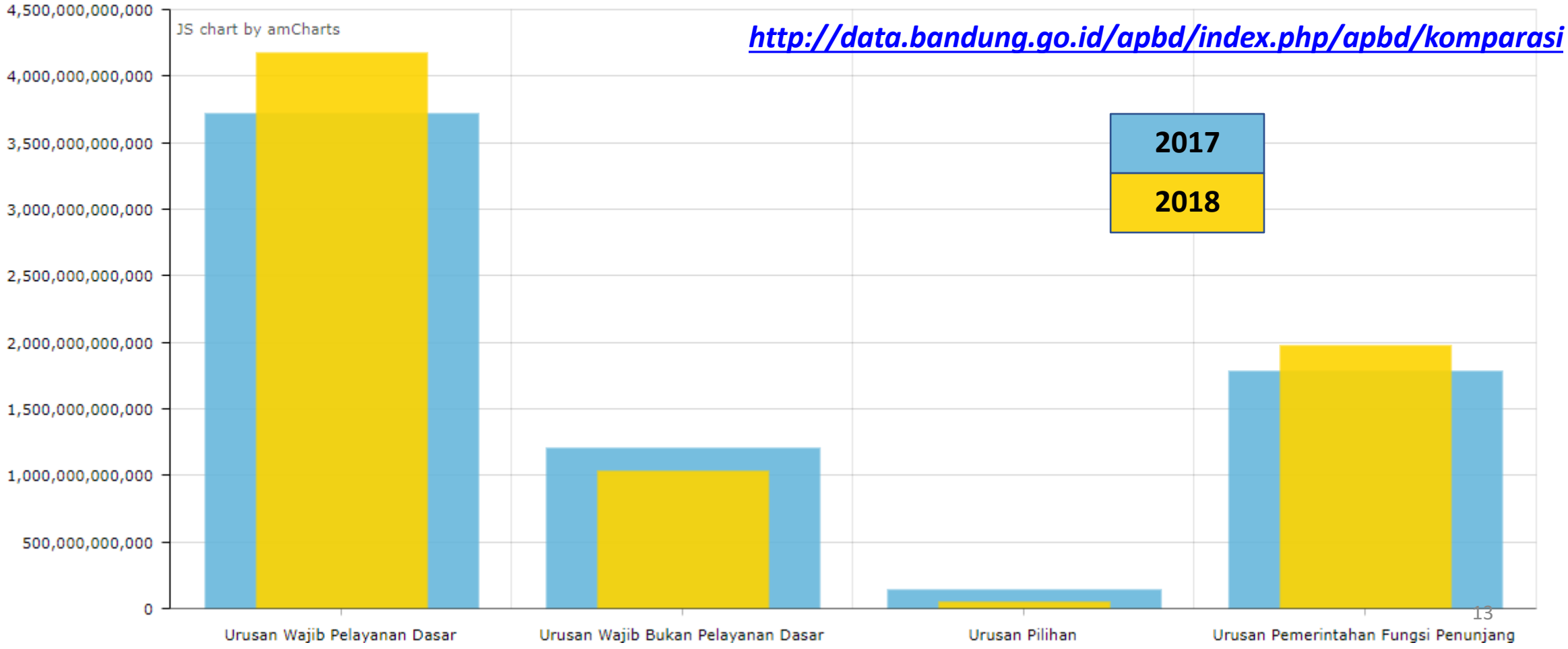
World Bank, BPS, Kemenkeu

EDA for Bandung's regional budget (APBD)

Tahun	Urusan	Nilai
2017	Wajib Pelayanan Dasar	3719118136664
2017	Wajib Bukan Pelayanan Dasar	1210033618392
2017	Pilihan	142725926491
2017	Pemerintahan Fungsi Penunjang	1784054610909
2018	Wajib Pelayanan Dasar	4178281927428
2018	Wajib Bukan Pelayanan Dasar	1033613747540
2018	Pilihan	50233286256
2018	Pemerintahan Fungsi Penunjang	1977684576079

EDA for Bandung's regional budget (APBD)

Komparasi Anggaran Urusan Pemerintahan tahun 2017 dan 2018



http://data.bandung.go.id/apbd/index.php/apbd/komparasi

Why EDA?

Raw Data

Why EDA?

Raw Data

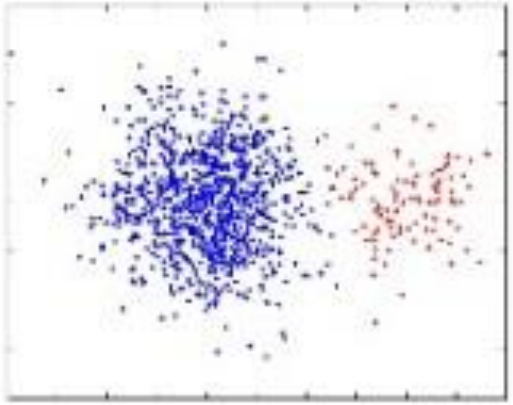
EDA

Why EDA?

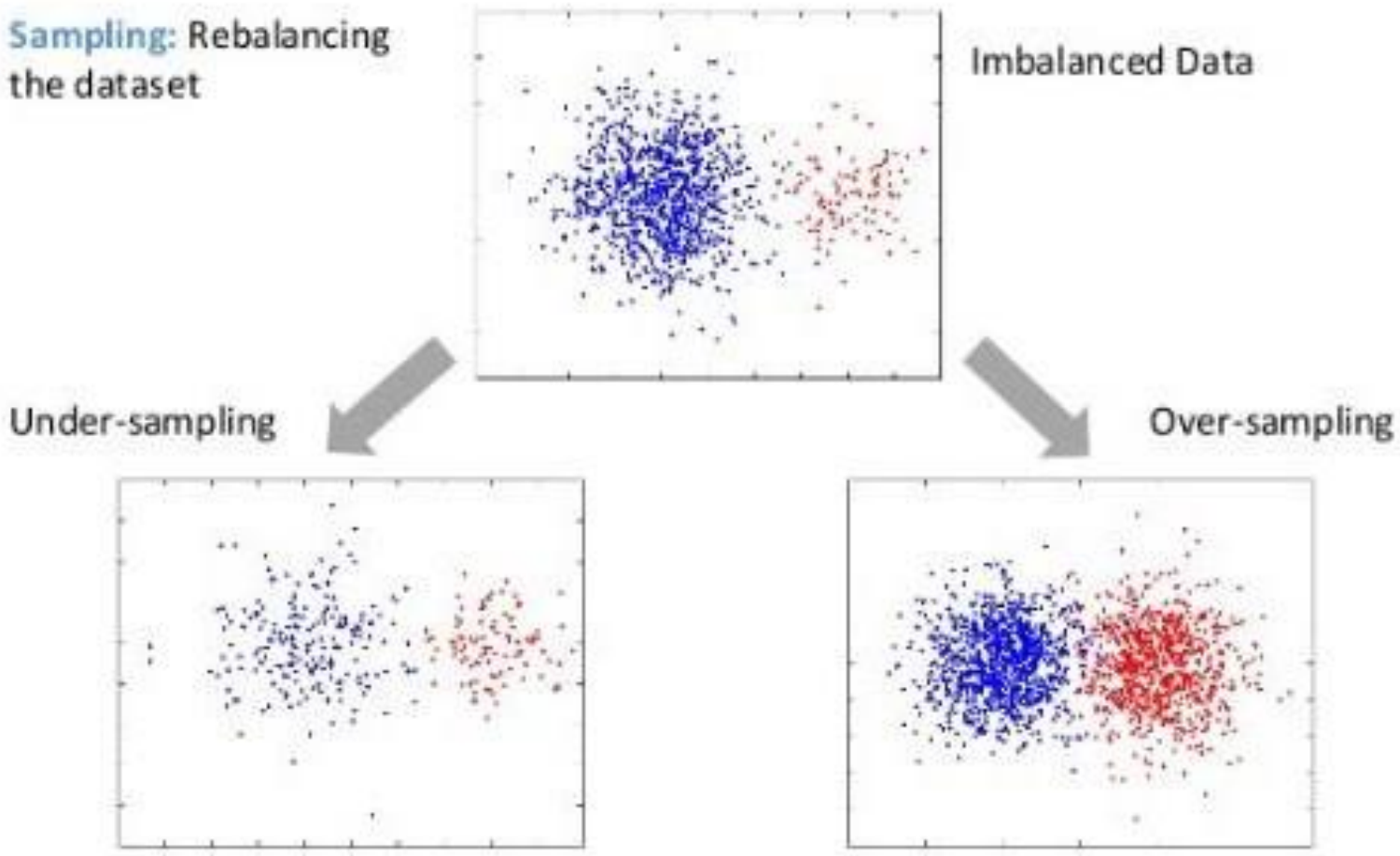
- We need to familiarize with a new dataset: How does it look like?
  - How many attributes, and of what kind?
  - Are there any missing values?
  - How are the values distributed?
  - Is our dataset imbalanced? (= if left untreated, our model can be biased)

- Hunting for something interesting: What catches your eyes?
  - Are there any outliers?
  - Are there any correlations between attributes?
  - How do the distributions compare between different samples?

For example, with EDA, we can detect imbalanced data



For example, with EDA, we can detect imbalanced data



EDA approaches

- **Descriptive statistics**
  - Central tendency
  - Measure of variation
  - Skewness & kurtosis
  - Correlations

- **Data visualizations**
  - Single attribute (univariate analysis):
    Barcharts, histogram, pie charts, donut charts
  - Multiple attributes (multivariate analysis):
    Scatter plots, bubble charts, linecharts, heatmaps

Quiz time: EDA on Bandung parking violations

1. Go to: http://data.bandung.go.id/dataset/pelanggaran-parkir-di-kota-bandung
2. Explore these two datasets:
    - Jumlah Pelanggaran Parkir - Tahun 2013
    - Jumlah Pelanggaran Parkir - Tahun 2014
3. Answer these questions:
    - How many columns and rows are there?
    - What are the column types?
    - Is there any missing data?

Descriptive statistics

Descriptive statistics: central tendency

(Arithmetic) Mean

Assuming we have *n* data values, labeled x_1 through x_n:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The arithmetic mean $\bar{x}$ is the sum of all data values (that is, from x_1 to x_n) divided by the number of values *n*.

Descriptive statistics: central tendency

Median

The **middle value** after all of the data values are put in an **ordered list**.

Median is robust: A few very high or low values have no effect on the median.

PS: If there are an even number of values, take the average of the two middle values.

Descriptive statistics: central tendency

Mode

Mode is the **most likely** or frequently occurring value.

Mode is suitable for categorical data, for example,
what is the mode of the music genres of a group of people?

Central tendency: Exercise 1

Consider the following field trip durations:
9, 10, 12, 12, 10, 11, 12, 13, 16, 18.

Compute the mean, median, and mode!

Central tendency: Exercise 1

Consider the following field trip durations:
9, 10, 12, 12, 10, 11, 12, 13, 16, 18.

Compute the mean, median, and mode!

$$\text{Mean} = \frac{9 + 10 + 12 + 12 + 10 + 11 + 12 + 13 + 16 + 18}{10} = 12.3$$

Consider the following ordered data values:

9, 10, 10, 11, 12 , 12, 12, 13, 16, 18

↑

$(n+1)/2 = (10+1)/2 =$ 5.5th value

Median = 12

Mode = 12

Central tendency: Exercise 2

Consider the following field trip durations:
9, 10, 12, 12, 10, 11, 12, 13, 16, 38.

Compute the mean, median, and mode!

Central tendency: Exercise 2

Consider the following field trip durations:
9, 10, 12, 12, 10, 11, 12, 13, 16, 38.

Compute the mean, median, and mode!

$$\text{Mean} = \frac{9 + 10 + 12 + 12 + 10 + 11 + 12 + 13 + 16 + 38}{10} = 14.3 \text{ days}$$

Ordered data values:

9, 10, 10, 11, 12 , 12, 12, 13, 16, 38

↑

Median = 12 days (not influenced by extreme values)

Mode = 12 days

Central tendency: Real-world exercise

Analyze the mean and median on the dataset
"Jumlah Pelanggaran Parkir - Tahun 2014" of Bandung city

No	Bulan	Tahun	Roda 4 (Mobil)	Roda 2 (Motor)	Total Kendaraan
1	Januari	2014	107	32	139
2	Februari	2014	146	1	147
3	Maret	2014	199	50	249
4	April	2014	222	42	264
5	Mei	2014	239	101	340
6	Juni	2014	150	48	198
7	Juli	2014	57	81	138
8	Agustus	2014	290	78	368
9	Septembe	2014	318	31	349
10	Oktober	2014	275	54	329
11	Novembe	2014	252	106	358
12	Desembe	2014	122	30	152

Central tendency: Real-world exercise

Analyze the mean and median on the dataset
"Jumlah Pelanggaran Parkir - Tahun 2014" of Bandung city

No	Bulan	Tahun	Roda 4 (Mobil)	Roda 2 (Motor)	Total Kendaraan
1	Januari	2014	107	32	139
2	Februari	2014	146	1	147
3	Maret	2014	199	50	249
4	April	2014	222	42	264
5	Mei	2014	239	101	340
6	Juni	2014	150	48	198
7	Juli	2014	57	81	138
8	Agustus	2014	290	78	368
9	Septembe	2014	318	31	349
10	Oktober	2014	275	54	329
11	Novembe	2014	252	106	358
12	Desembe	2014	122	30	152

MEAN	198,1	54,5	252,6
MEDIAN	210,5	49	256,5

Descriptive statistics: Measures of variation

Range

Range = max - min

The simplest measure of variation, often denoted by indicating the largest and smallest values separately.

Descriptive statistics: Measures of variation

Inter-Quartile Range (IQR)

Divides a dataset into quartiles:

- Q1 (lower quartile): 25th percentile
  - Median of lower half
- Q2 (median): 50th percentile
- Q3 (upper quartile): 75th percentile
  - Median of upper half

IQR = Q3-Q1

Inter-quartile range: Exercise 0

From the data (n = 7):
5, 7, 4, 4, 6, 2, 8

Q1 = ?
Q2 = ?
Q3 = ?

IQR = ? **Range = ?**

Inter-quartile range: Exercise 0

From the data (n = 7):
5, 7, 4, 4, 6, 2, 8 -> Sorted: **2, 4, 4, 5, 6, 7, 8**

Q1 = median of lower half = 4
Q2 = 5
Q3 = median of upper half = 7

IQR = Q3-Q1 = 3 **Range = 6**

Inter-quartile range: Exercise 1

From the data (n = 10):
25, 26, 29, 32, 35, 36, 38, 44, 49, 51

Q1 = ?
Q2 = ?
Q3 = ?

Inter-quartile range: Exercise 1

From the data (n = 10):
25, 26, 29, 32, 35, 36, 38, 44, 49, 51

Q1 = median of lower half = third smallest value = 29
Q2 = 35.5 (average of 5^{th} and 6^{th} observations)
Q3 = median of upper half = third largest value = 44

IQR = Q3-Q1 = 44-29 = 15 **Range = 51-25 = 26**

Inter-quartile range: Exercise 2

From the data (n = 10):
25, 26, 29, 32, 35, 36, 38, 44, 49, 510

Q1 = ?
Q2 = ?
Q3 = ?

Inter-quartile range: Exercise 2

From the data (n = 10):
25, 26, 29, 32, 35, 36, 38, 44, 49, 510

Q1 = median of lower half = third smallest value = 29
Q2 = 35.5 (average of 5^{th} and 6^{th} observations)
Q3 = median of upper half = third largest value = 44

IQR = Q3-Q1 = 44-29 = 15 **Range = 510-25 = 485**

Inter-quartile range: Exercise 3

From the data (n = 11):
2, 3, 4, 5, 6, 6, 6, 7, 7, 8, 9

Minimum = ?
Maximum = ?
Range = ?
Lower quartile (Q1) = ?
Median (Q2) = ?
Upper quartile (Q3) = ?
IQR = ?

Inter-quartile range: Exercise 3

From the data (n = 11):

2, 3, 4, 5, 6, 6, 6, 7, 7, 8, 9

Minimum = 2

Maximum = 9

Range = 2 to 9 = 7

Lower quartile (Q1) = 4

Median (Q2) = 6

Upper quartile (Q3) = 7

IQR = 3

Inter-quartile range: Real-World Exercise

Analyze the inter-quartile information on the dataset "Jumlah Pelanggaran Parkir - Tahun 2014" of Bandung city

No	Bulan	Tahun	Roda 4 (Mobil)	Roda 2 (Motor)	Total Kendaraan
1	Januari	2014	107	32	139
2	Februari	2014	146	1	147
3	Maret	2014	199	50	249
4	April	2014	222	42	264
5	Mei	2014	239	101	340
6	Juni	2014	150	48	198
7	Juli	2014	57	81	138
8	Agustus	2014	290	78	368
9	Septembe	2014	318	31	349
10	Oktober	2014	275	54	329
11	Novembe	2014	252	106	358
12	Desembe	2014	122	30	152

Inter-quartile range: Real-World Exercise

Analyze the inter-quartile information on the dataset "Jumlah Pelanggaran Parkir - Tahun 2014" of Bandung city

No	Bulan	Tahun	Roda 4 (Mobil)	Roda 2 (Motor)	Total Kendaraan
1	Januari	2014	107	32	139
2	Februari	2014	146	1	147
3	Maret	2014	199	50	249
4	April	2014	222	42	264
5	Mei	2014	239	101	340
6	Juni	2014	150	48	198
7	Juli	2014	57	81	138
8	Agustus	2014	290	78	368
9	Septembe	2014	318	31	349
10	Oktober	2014	275	54	329
11	Novembe	2014	252	106	358
12	Desembe	2014	122	30	152

MEAN	198,1	54,5	252,6
MEDIAN	210,5	49	256,5
MODE	#N/A	#N/A	#N/A
MIN	57	1	138
MAX	318	106	368
RANGE	261	105	230
Q1	128	31,25	148,25
Q2	210,5	49	256,5
Q3	269,25	80,25	346,75
IQR	141,25	49	198,5

Outliers

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. Before abnormal observations can be singled out, it is necessary to characterize normal observations.

Outliers, according to IQR, are data points whose values are:

- less than Q1-1.5*IQR, or
- more than Q3+1.5*IQR

Outliers: Exercise

The dataset of N = 90 **ordered** observations as shown below is examined for outliers:

30, 171, 184, 201, 212, 250, 265, 270, 272, 289, 305, 306, 322, 322, 336, 346, 351, 370, 390, 404, 409, 411, 436, 437, 439, 441, 444, 448, 451, 453, 470, 480, 482, 487, 494, 495, 499, 503, 514, 521, 522, 527, 548, 550, 559, 560, 570, 572, 574, 578, 585, 592, 592, 607, 616, 618, 621, 629, 637, 638, 640, 656, 668, 707, 709, 719, 737, 739, 752, 758, 766, 792, 792, 794, 802, 818, 830, 832, 843, 858, 860, 869, 918, 925, 953, 991, 1000, 1005, 1068, 1441.

Q2 (Median) = 559.5

Q1 = 429.75

Q3 = 742.25

IQR = Q3-Q1 = 312.5

Are there any outliers?

Outliers: Exercise

The dataset of N = 90 **ordered** observations as shown below is examined for outliers:

30, 171, 184, 201, 212, 250, 265, 270, 272, 289, 305, 306, 322, 322, 336, 346, 351, 370, 390, 404, 409, 411, 436, 437, 439, 441, 444, 448, 451, 453, 470, 480, 482, 487, 494, 495, 499, 503, 514, 521, 522, 527, 548, 550, 559, 560, 570, 572, 574, 578, 585, 592, 592, 607, 616, 618, 621, 629, 637, 638, 640, 656, 668, 707, 709, 719, 737, 739, 752, 758, 766, 792, 792, 794, 802, 818, 830, 832, 843, 858, 860, 869, 918, 925, 953, 991, 1000, 1005, 1068, 1441.

Q2 (Median) = 559.5

Q1 = 429.75

Q3 = 742.25

IQR = Q3-Q1 = 312.5

Are there any outliers?

Outliers, according to IQR, are data points whose values are:

- less than Q1-1.5*IQR, or
- more than Q3+1.5*IQR

Outliers: Exercise

The dataset of N = 90 **ordered** observations as shown below is examined for outliers:

30, 171, 184, 201, 212, 250, 265, 270, 272, 289, 305, 306, 322, 322, 336, 346, 351, 370, 390, 404, 409, 411, 436, 437, 439, 441, 444, 448, 451, 453, 470, 480, 482, 487, 494, 495, 499, 503, 514, 521, 522, 527, 548, 550, 559, 560, 570, 572, 574, 578, 585, 592, 592, 607, 616, 618, 621, 629, 637, 638, 640, 656, 668, 707, 709, 719, 737, 739, 752, 758, 766, 792, 792, 794, 802, 818, 830, 832, 843, 858, 860, 869, 918, 925, 953, 991, 1000, 1005, 1068, **1441.**

Q2 (Median) = 559.5

Q1 = 429.75

Q3 = 742.25

IQR = Q3-Q1 = 312.5

Are there any outliers? Yes, see the one highlighted yellow above.

Outlier: Real-World Exercise

Is there any outlier on the dataset
"Jumlah Pelanggaran Parkir - Tahun 2014" of Bandung city?

No	Bulan	Tahun	Roda 4 (Mobil)	Roda 2 (Motor)	Total Kendaraan
1	Januari	2014	107	32	139
2	Februari	2014	146	1	147
3	Maret	2014	199	50	249
4	April	2014	222	42	264
5	Mei	2014	239	101	340
6	Juni	2014	150	48	198
7	Juli	2014	57	81	138
8	Agustus	2014	290	78	368
9	Septembe	2014	318	31	349
10	Oktober	2014	275	54	329
11	Novembe	2014	252	106	358
12	Desembe	2014	122	30	152

Outlier: Real-World Exercise

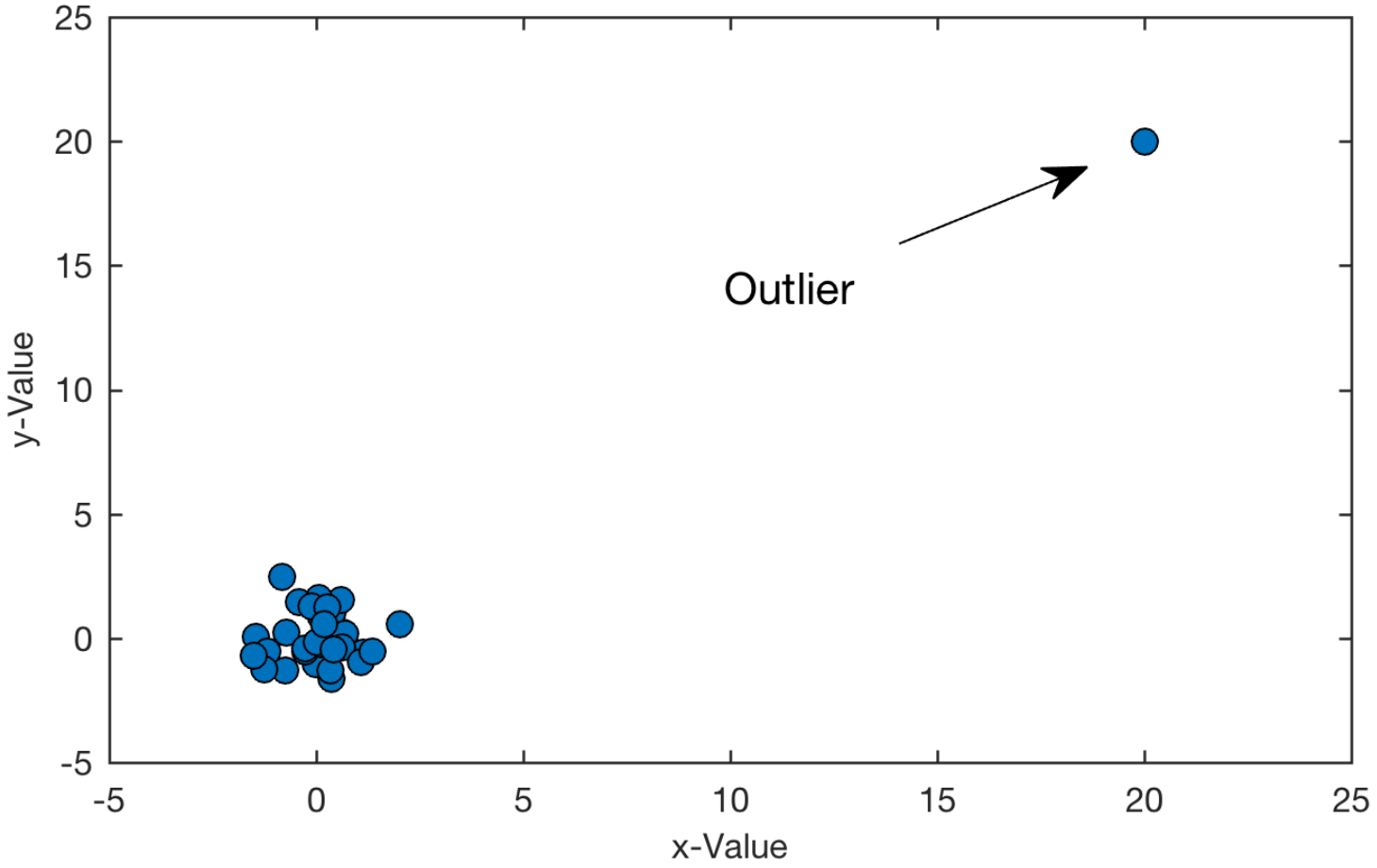
Is there any outlier on the dataset
"Jumlah Pelanggaran Parkir - Tahun 2014" of Bandung city?

No	Bulan	Tahun	Roda 4 (Mobil)	Roda 2 (Motor)	Total Kendaraan
1	Januari	2014	107	32	139
2	Februari	2014	146	1	147
3	Maret	2014	199	50	249
4	April	2014	222	42	264
5	Mei	2014	239	101	340
6	Juni	2014	150	48	198
7	Juli	2014	57	81	138
8	Agustus	2014	290	78	368
9	Septembe	2014	318	31	349
10	Oktober	2014	275	54	329
11	Novembe	2014	252	106	358
12	Desembe	2014	122	30	152

Answer = No

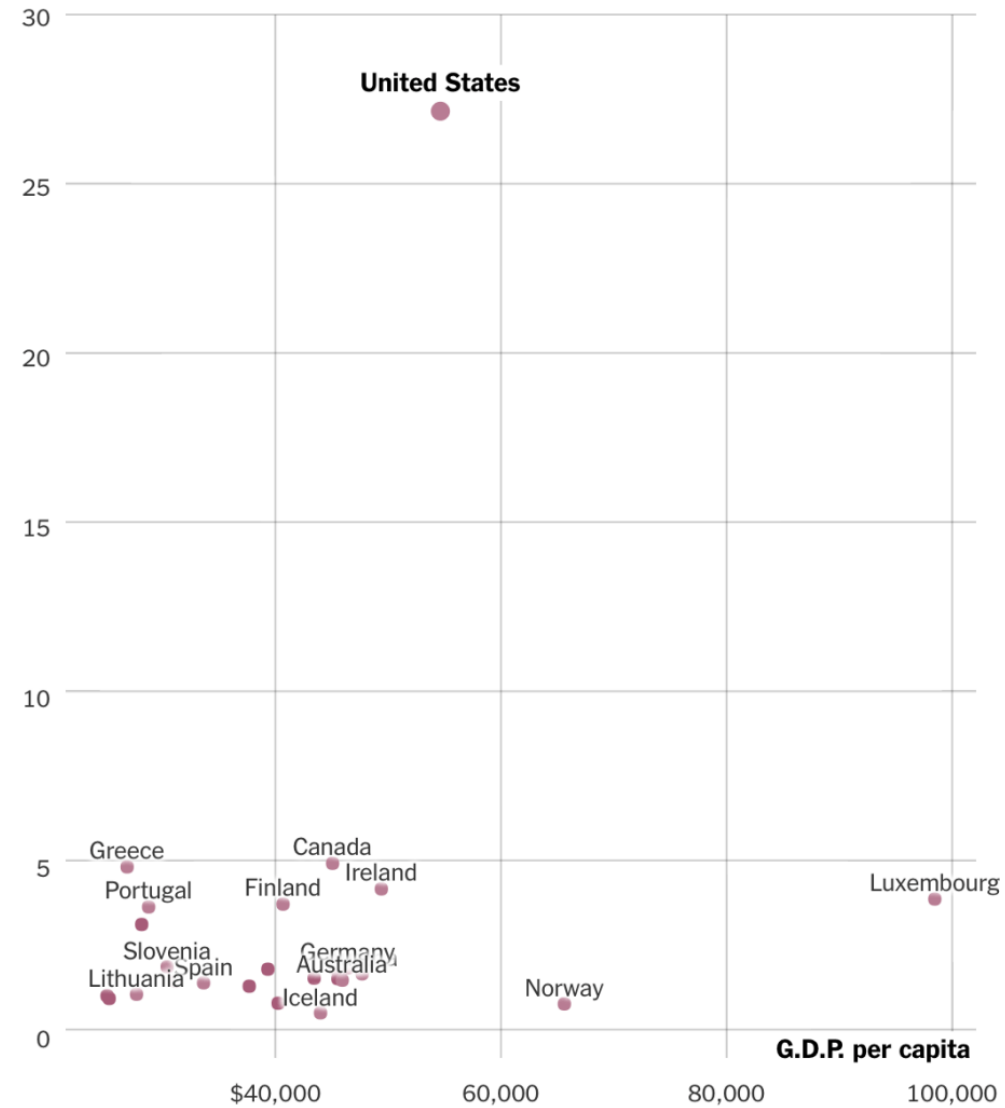
Lower outlier limit	-83,875	-42,25	-149,5
Upper outlier limit	481,125	153,75	644,5

Outliers: 2-Dimensional Data



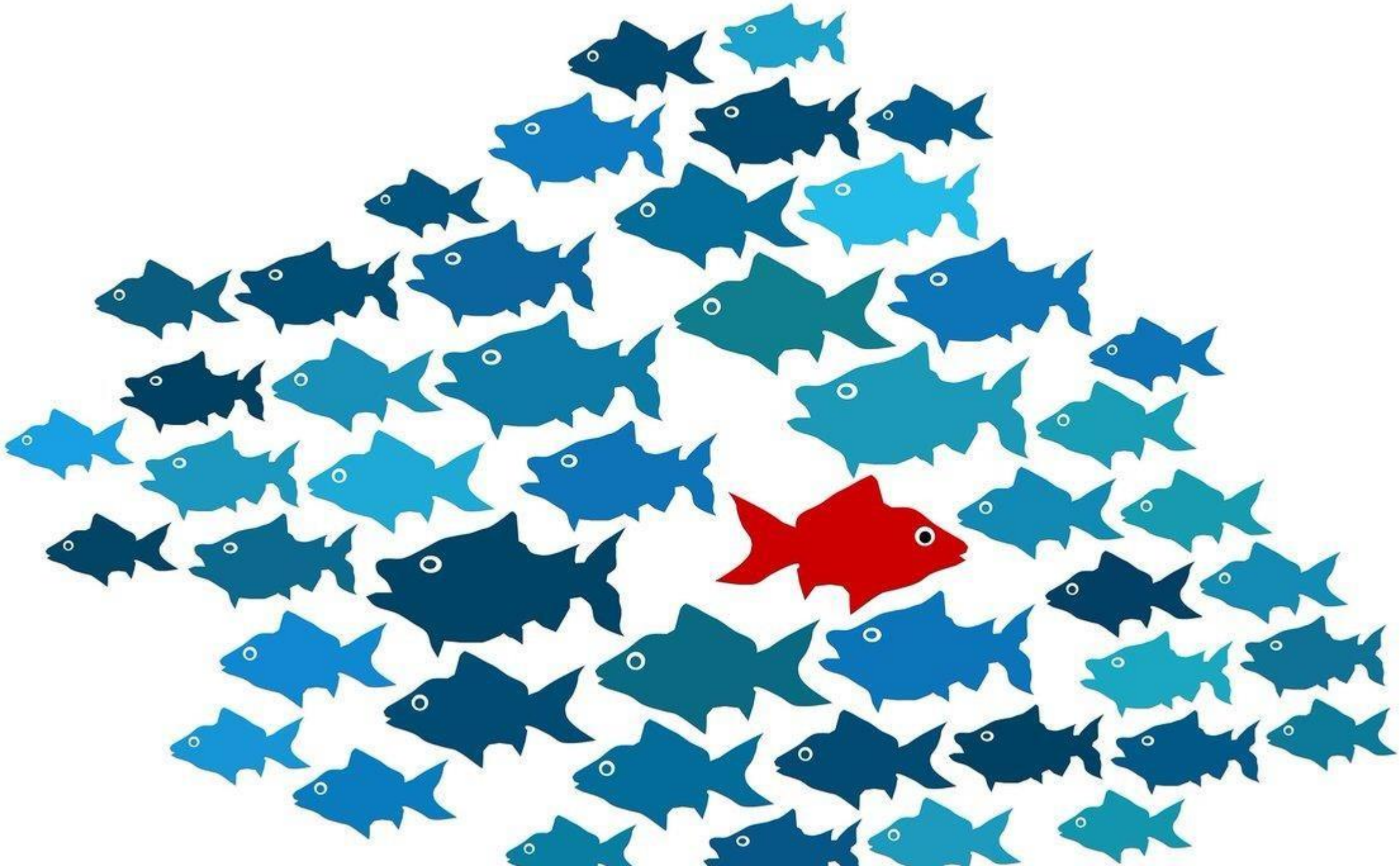
No Other Rich Western Country Comes Close

Gun homicides per day if each country had the same population as the U.S.



Shown are Western countries that have G.D.P. per capita over $25,000 and that make statistics on gun homicides available.

Sources: Small Arms Survey (2007–12 average); World Bank



Descriptive statistics: Measures of variation

Variance and Standard Deviation

Variance = Average of the squared deviation of the observations from the mean

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

Standard deviation s = Square root of the variance

Taking the root of the variance means the standard deviation is restored to the original unit of measure.

Intuition: If variance/std deviation is high, then your dataset has larger variability!

Variance and Standard Deviation: Exercise 1

From the data:

25, 26, 29, 32, 35, 36, 38, 44, 49, 51

Compute the variance and standard deviation?

Variance and Standard Deviation: Exercise 1

From the data:

25, 26, 29, 32, 35, 36, 38, 44, 49, 51

Compute the variance and standard deviation?

s^2 = 82.9

s = 9.1

Variance and Standard Deviation: Exercise 2

From the data:

25, 26, 27, 24, 25, 26, 28, 24, 25, 25

Compute the variance and standard deviation?

Variance and Standard Deviation: Exercise 2

From the data:

25, 26, 27, 24, 25, 26, 28, 24, 25, 25

Compute the variance and standard deviation?

s^2 = 1.6

s = 1.26

Descriptive statistics: Skewness & kurtosis

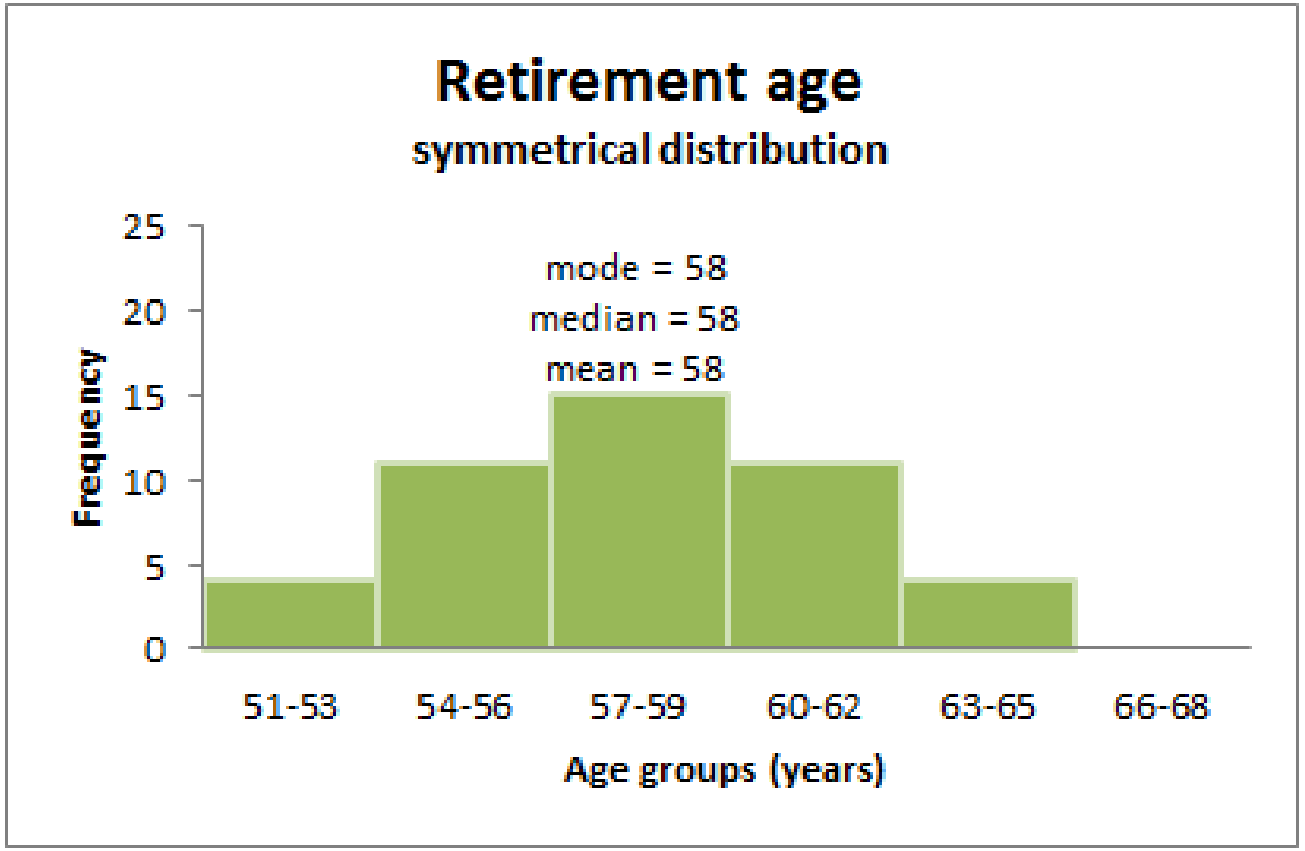
Skewness

A measure of asymmetry

Kurtosis

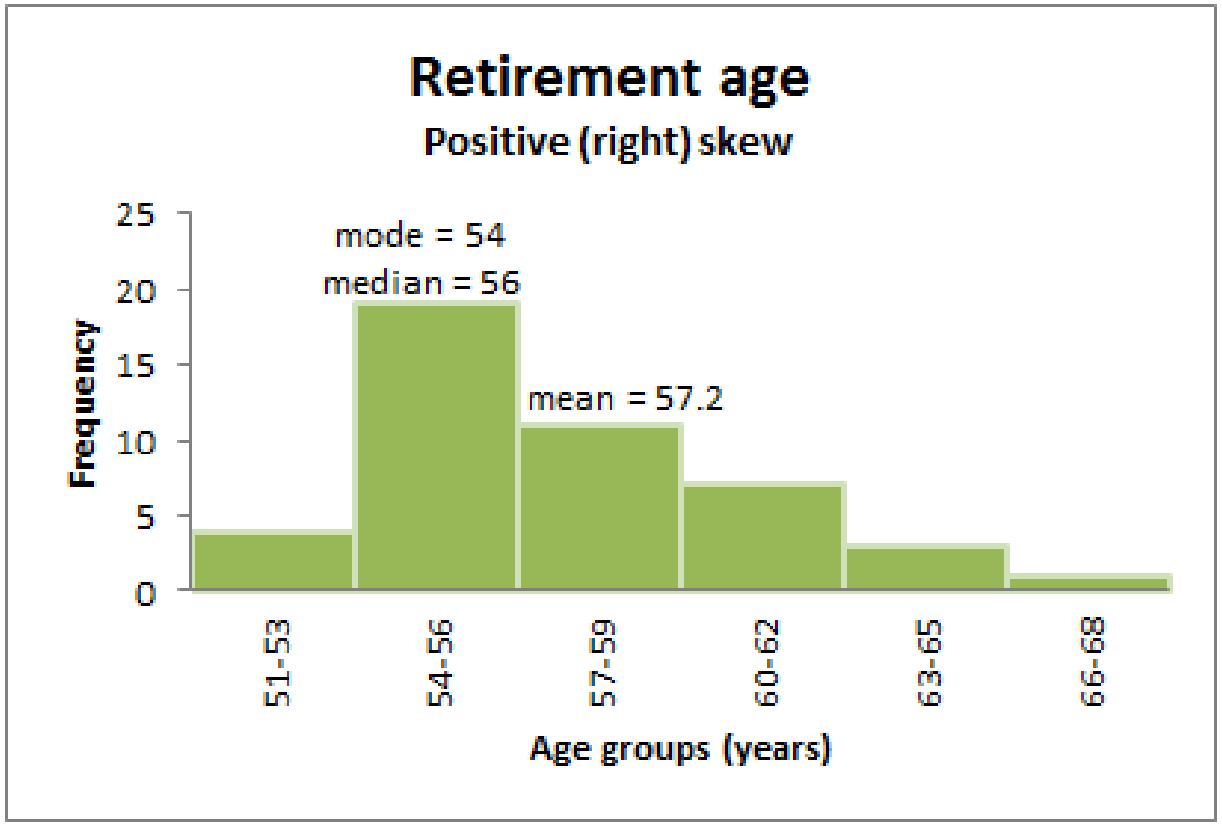
A measure of outliers

Descriptive statistics: Skewness



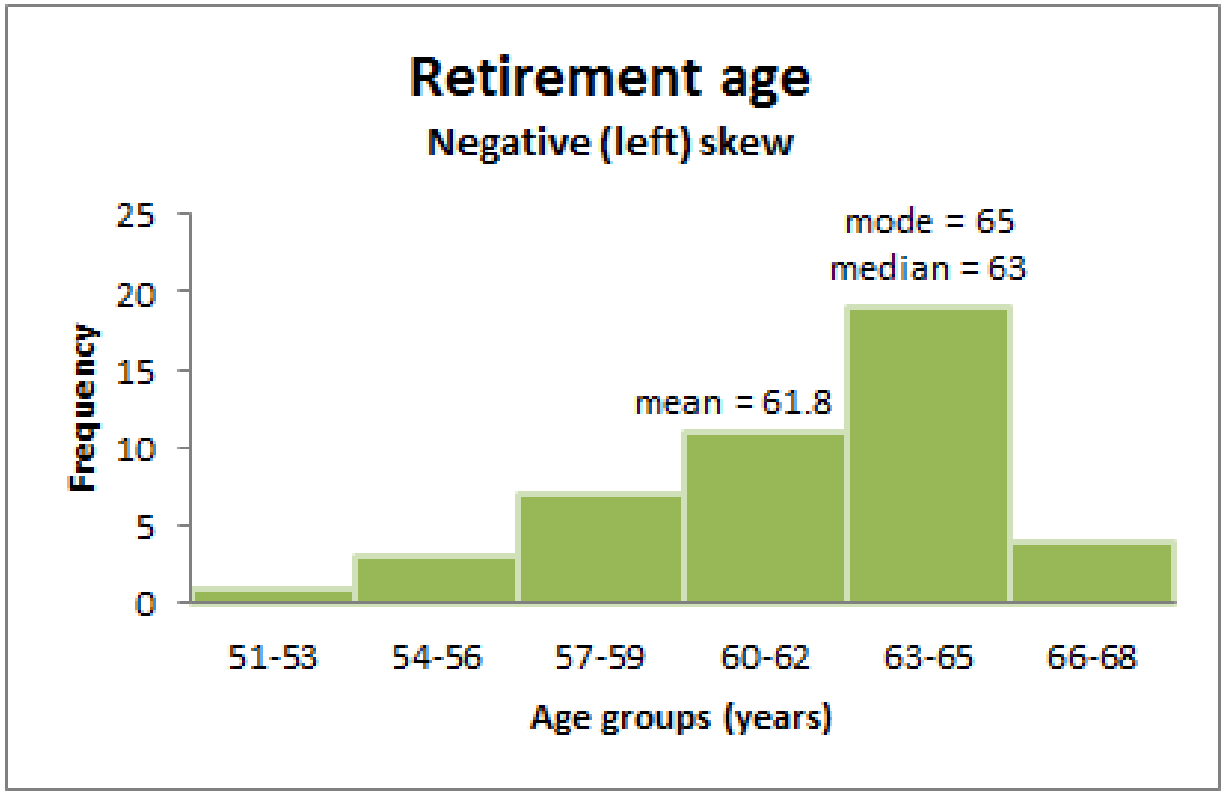
When a distribution is symmetrical, the mode, median and mean are all in the middle of the distribution.
The above graph shows a larger retirement age dataset with a distribution which is symmetrical.
The mode, median and mean all equal 58 years.

Descriptive statistics: Skewness



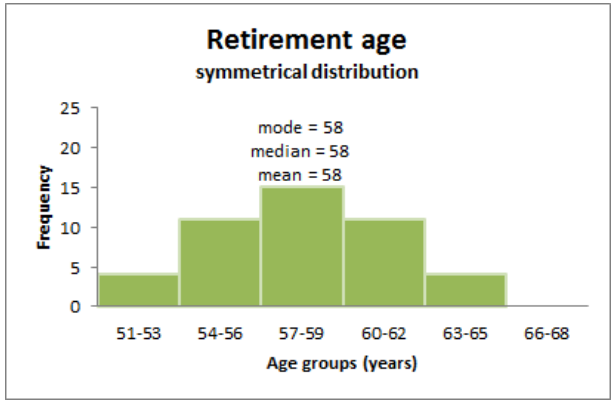
A distribution is said to be **positively or right skewed** when the tail on the right side of the distribution is longer than the left side. In a positively skewed distribution, it is common for the mean to be 'pulled' toward the right tail of the distribution. Although there are exceptions to this rule, generally, most of the values, including the median value, tend to be less than the mean value.

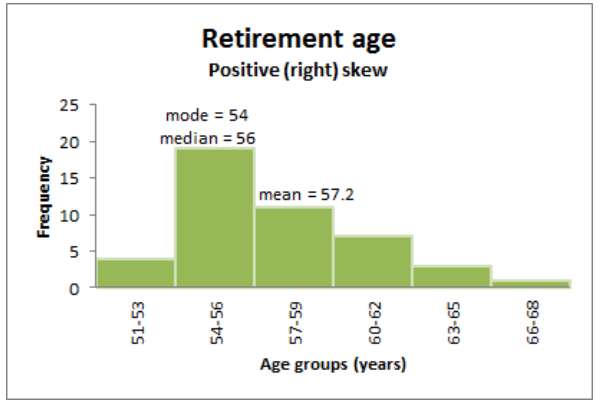
Descriptive statistics: Skewness

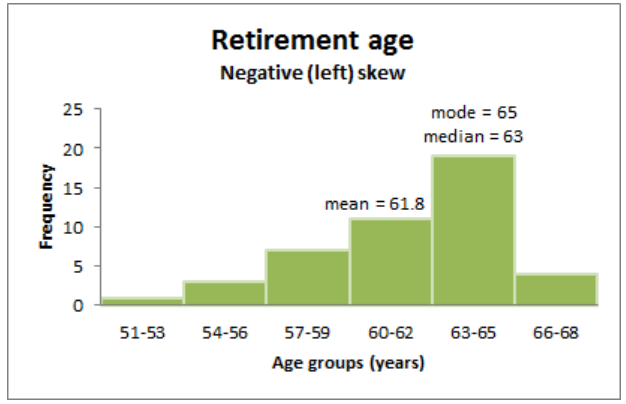


A distribution is said to be **negatively or left skewed** when the tail on the left side of the distribution is longer than the right side. In a negatively skewed distribution, it is common for the mean to be 'pulled' toward the left tail of the distribution. Although there are exceptions to this rule, generally, most of the values, including the median value, tend to be greater than the mean value.

Descriptive statistics: Skewness

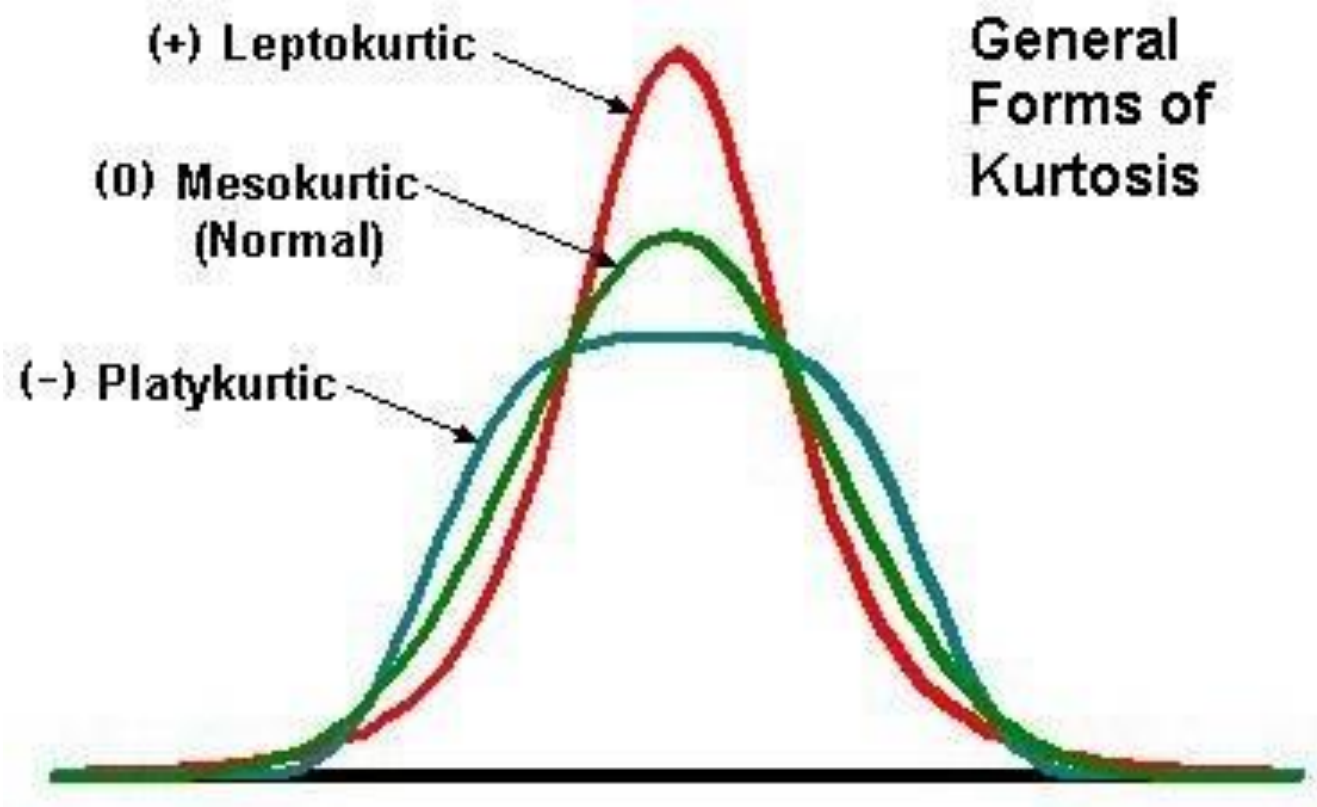






When a **distribution is skewed** the mode remains the most commonly occurring value, the median remains the middle value in the distribution, but **the mean is generally 'pulled' in the direction of the tails**. In a skewed distribution, the **median is often a preferred measure** of central tendency, as the mean is not usually in the middle of the distribution.

Descriptive statistics: Kurtosis

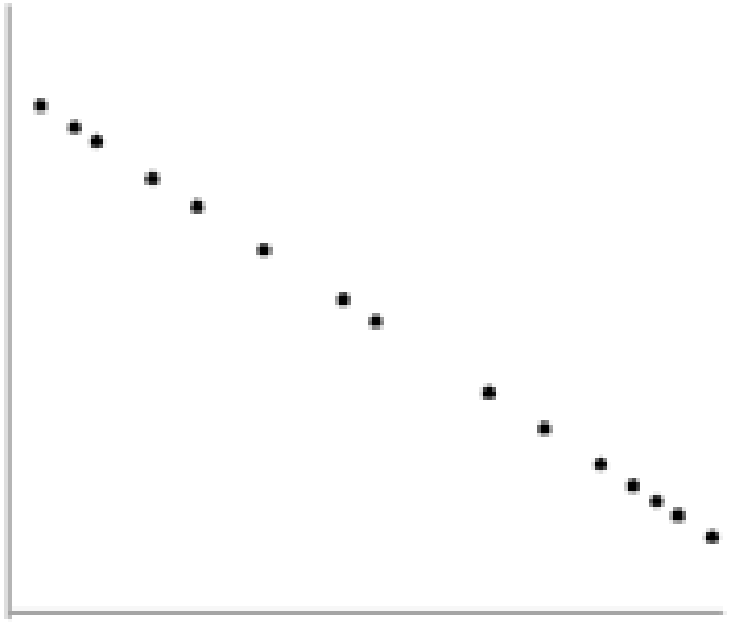


High kurtosis indicates the presence of outliers!

Correlation

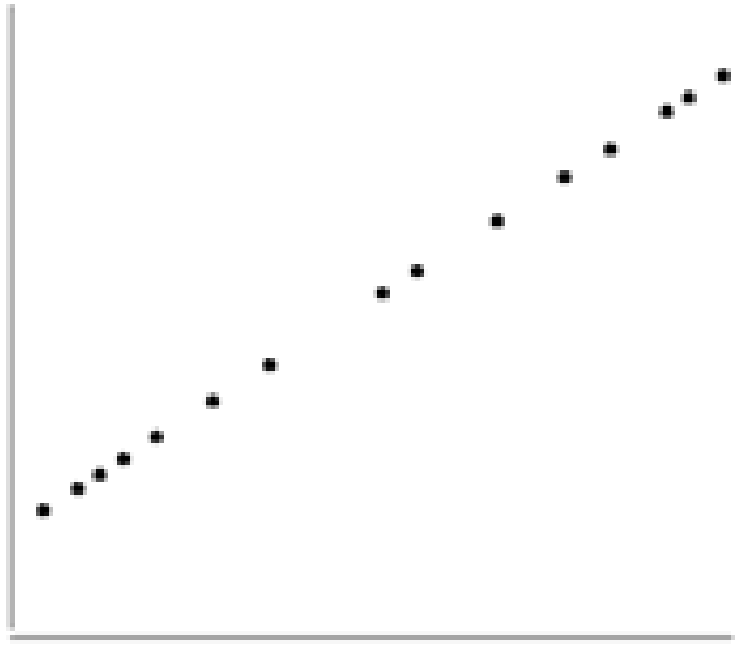
- It is a technique to investigate the relationship between two variables: that is, measures the strength of the association between the two variables
- Pearson's correlation coefficient (r) is a type of correlation coefficient
- Correlation coefficient returns a value between -1 and 1
  - -1 denotes strongest negative correlation
  - 0 denotes no correlation
  - 1 denotes strongest positive correlation

Correlation



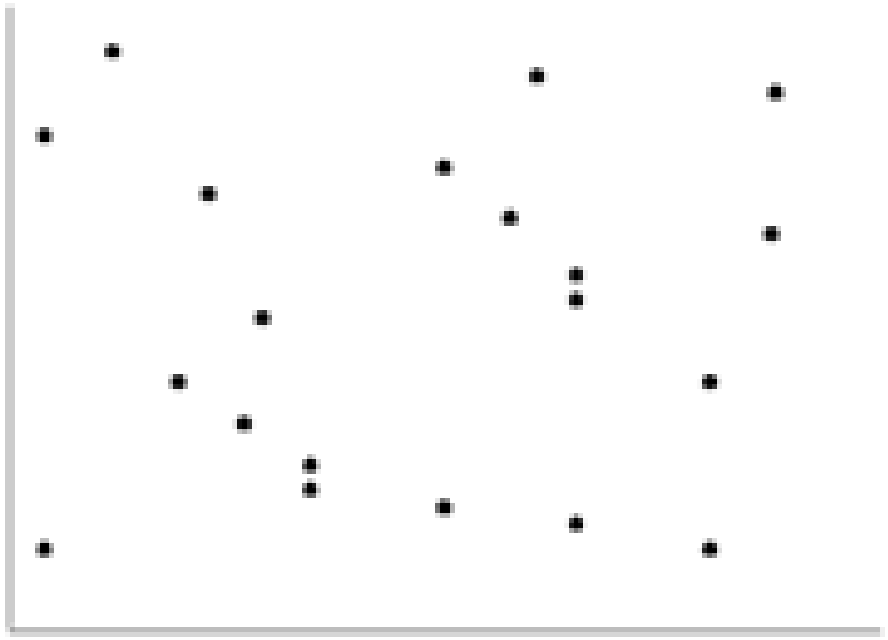
r = -1

Correlation



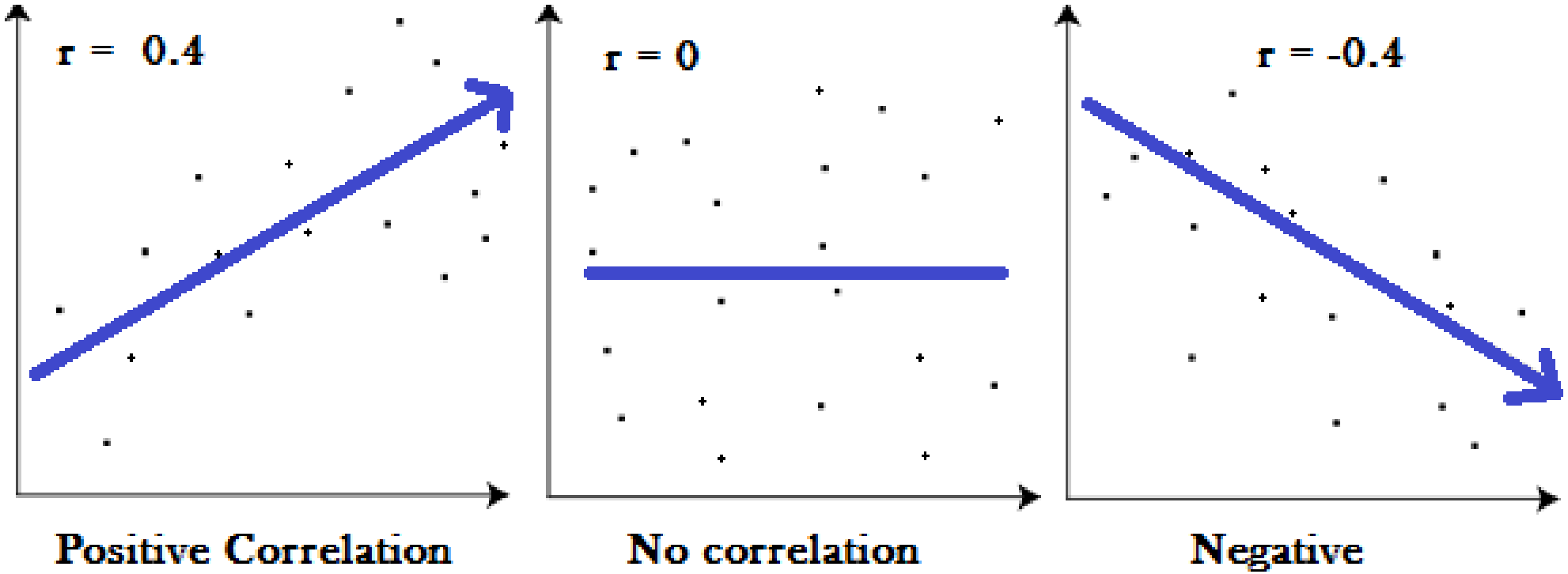
r = 1

Correlation



r = 0

Correlation



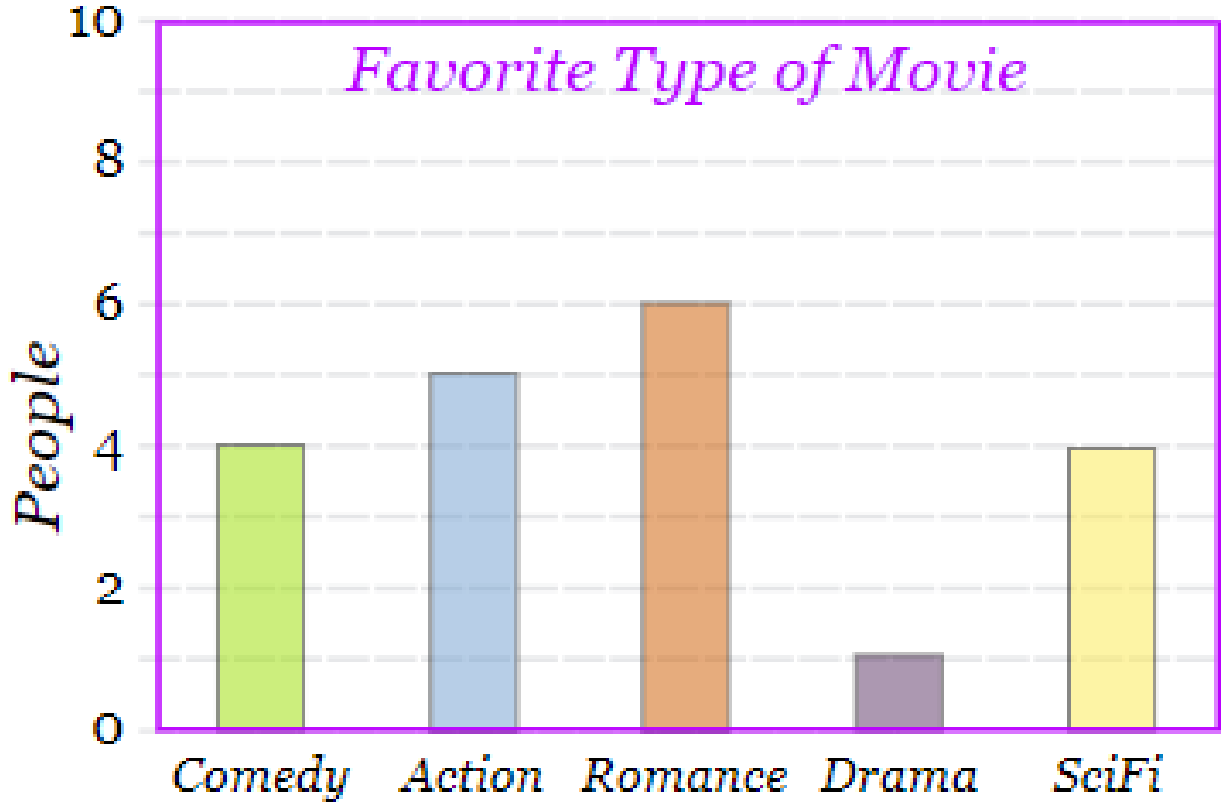
Data visualization

Why data visualization (data viz)?

- In technical documents or reports, one of the most important skills you need to have is the ability to make compelling data visualizations.
- The visualizations will typically carry the weight of your arguments.
- They need to be clear, attractive, and convincing.
- The difference between good and bad figures can be the difference between a highly influential or an obscure paper, a grant or contract won or lost, a job interview gone well or poorly.

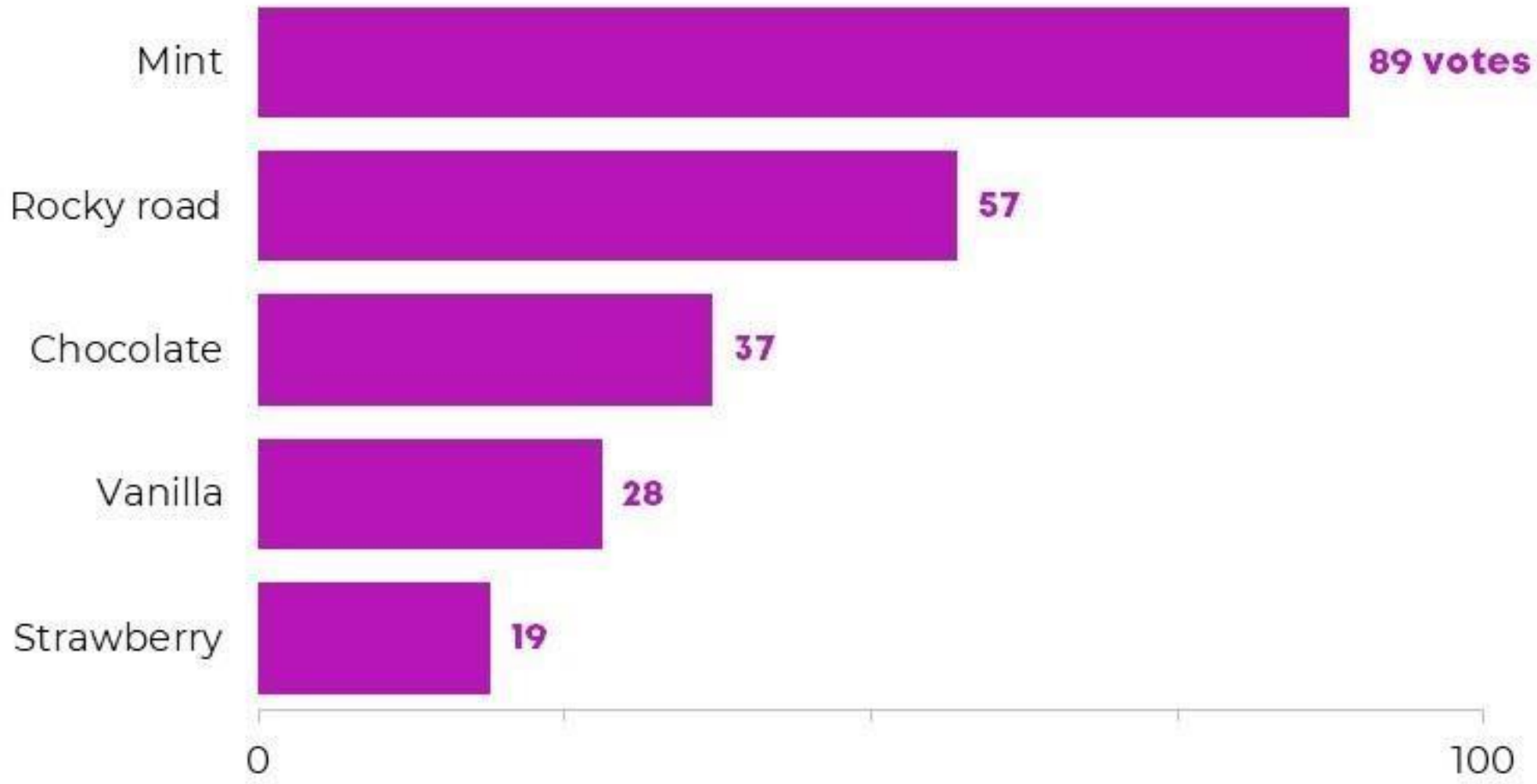
Data visualization: Barcharts

Table: Favorite Type of Movie				
Comedy	Action	Romance	Drama	SciFi
4	5	6	1	4



A barplot (or barchart) is one of the most common types of plot. It shows the relationship between a categorical variable and a numerical variable.

Data visualization: Barcharts



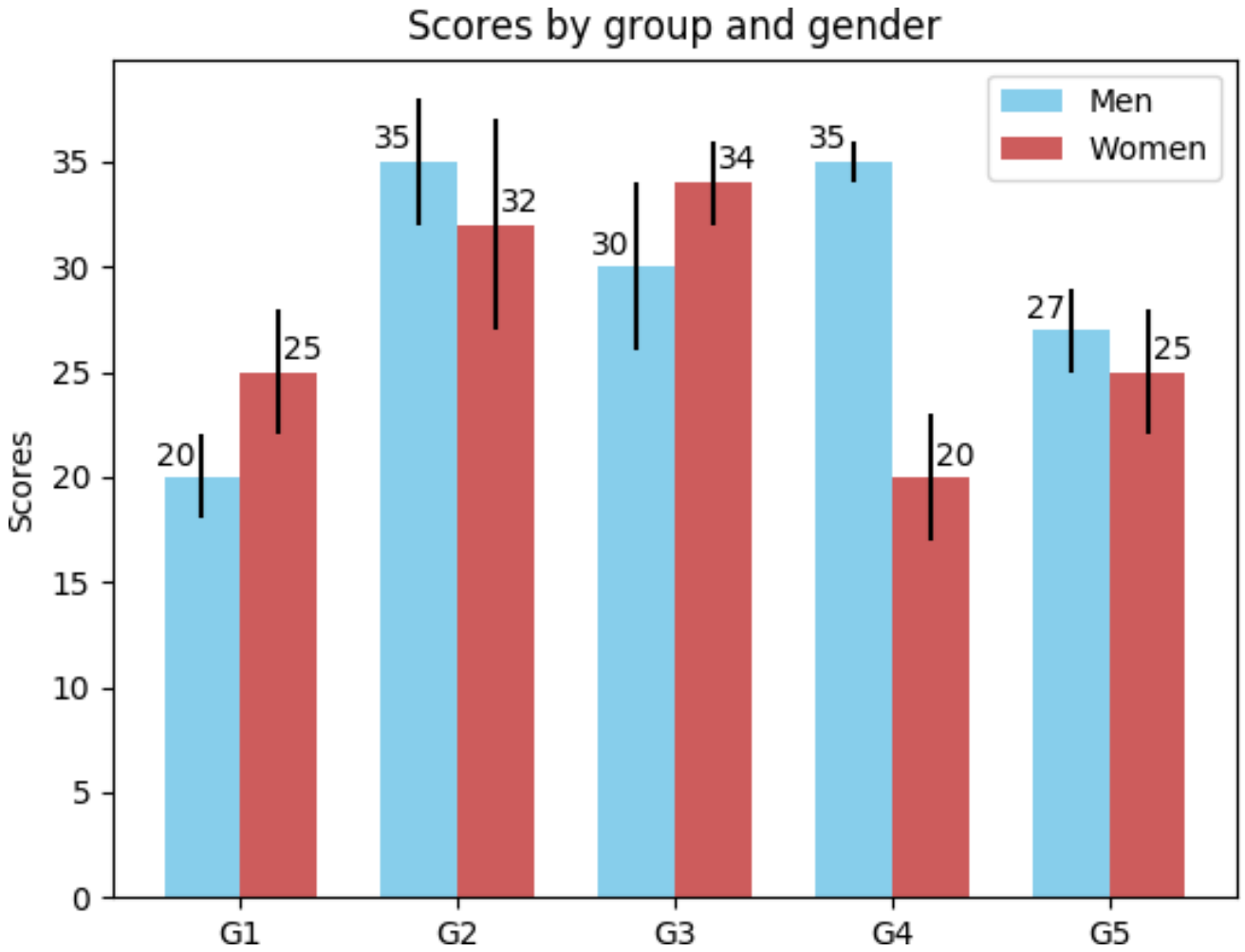
Barcharts with horizontal bars, can you guess why this format is sometimes necessary?

The World's Deadliest Earthquakes

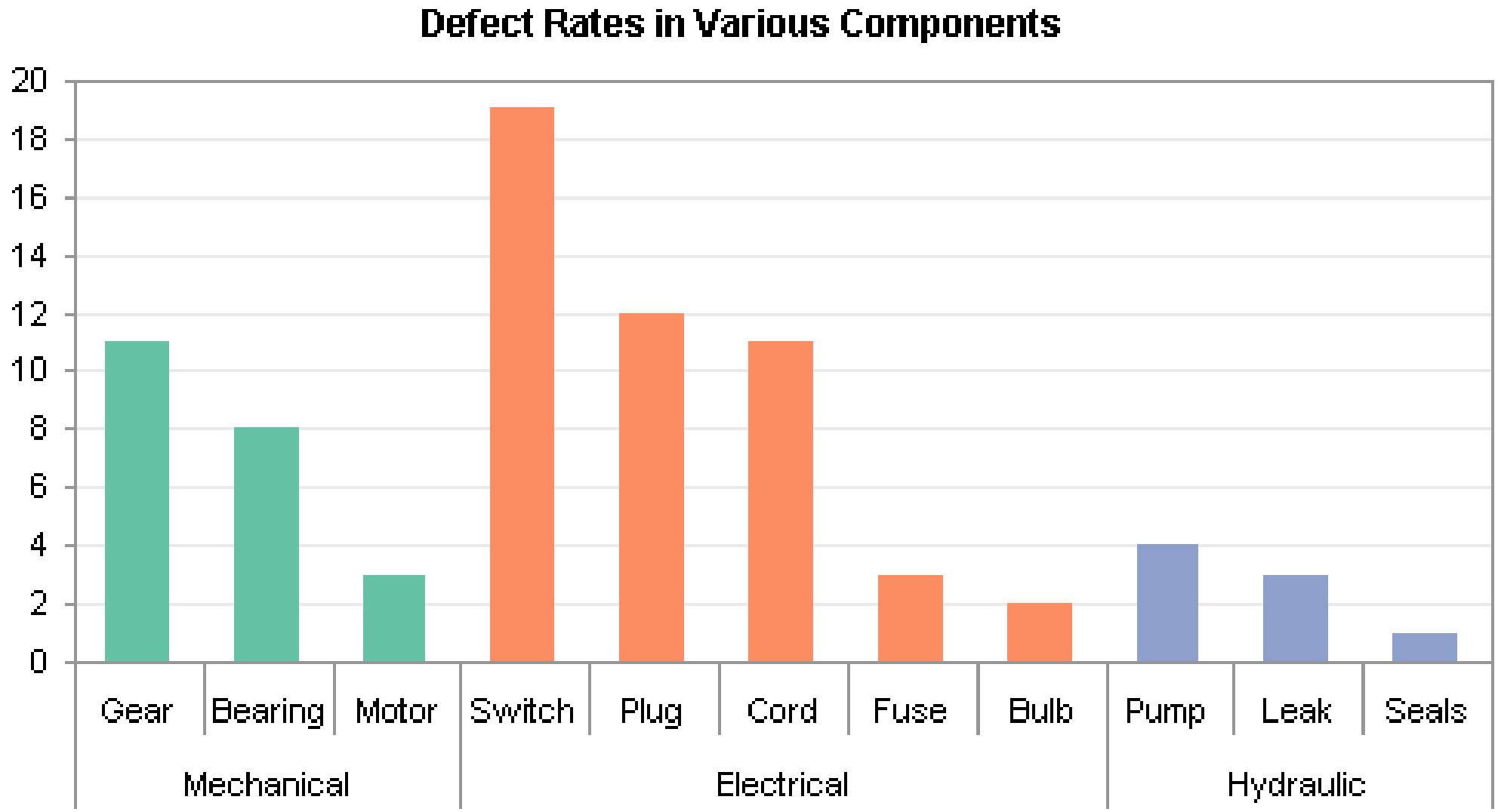
Earthquakes since 1900 which caused the most deaths*

* Figures are rounded. Some figures are estimates and could vary.
As of January 8, 2020
Sources: US Geological Survey, Wikipedia

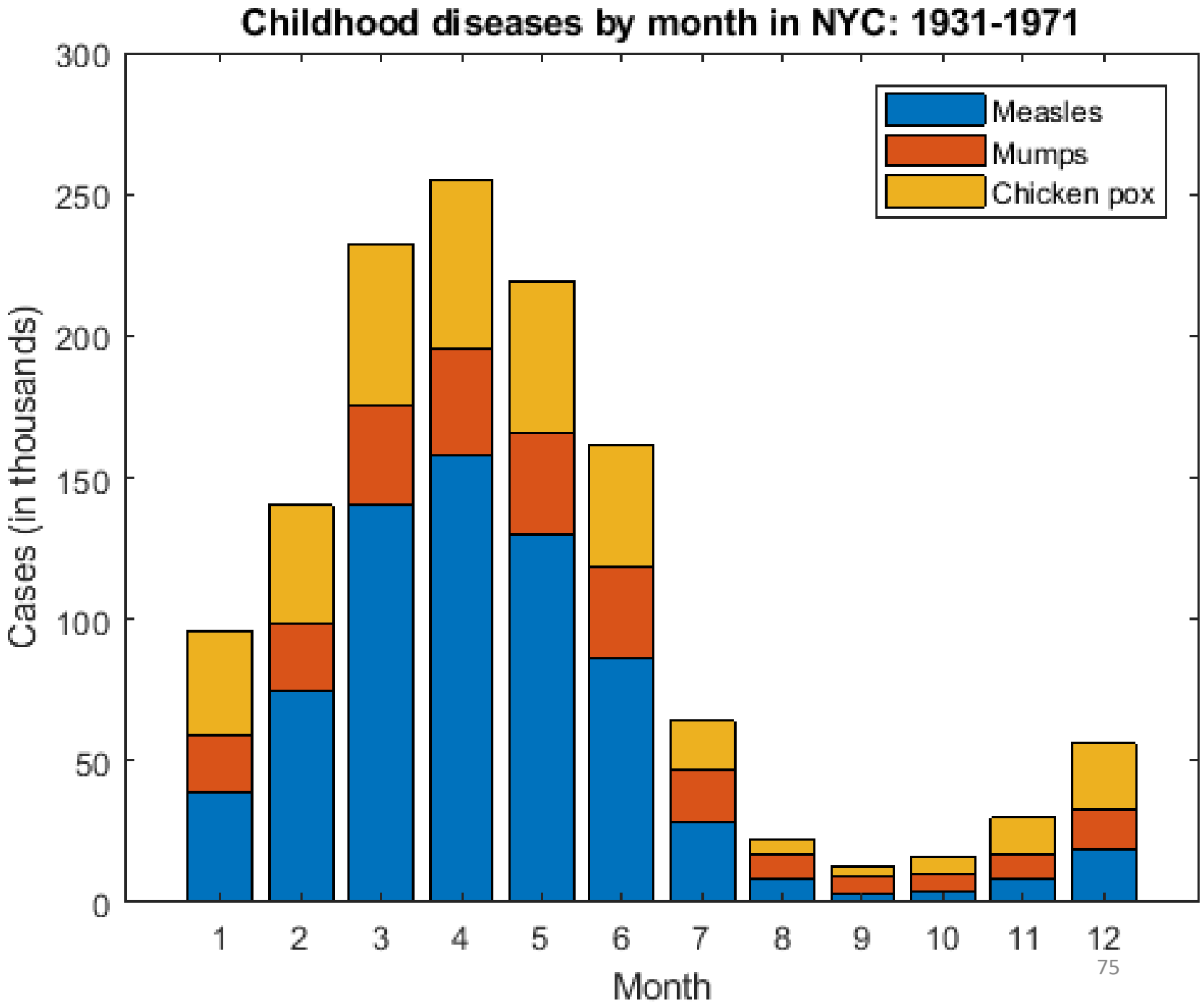
Data visualization: Barcharts - Contrasting



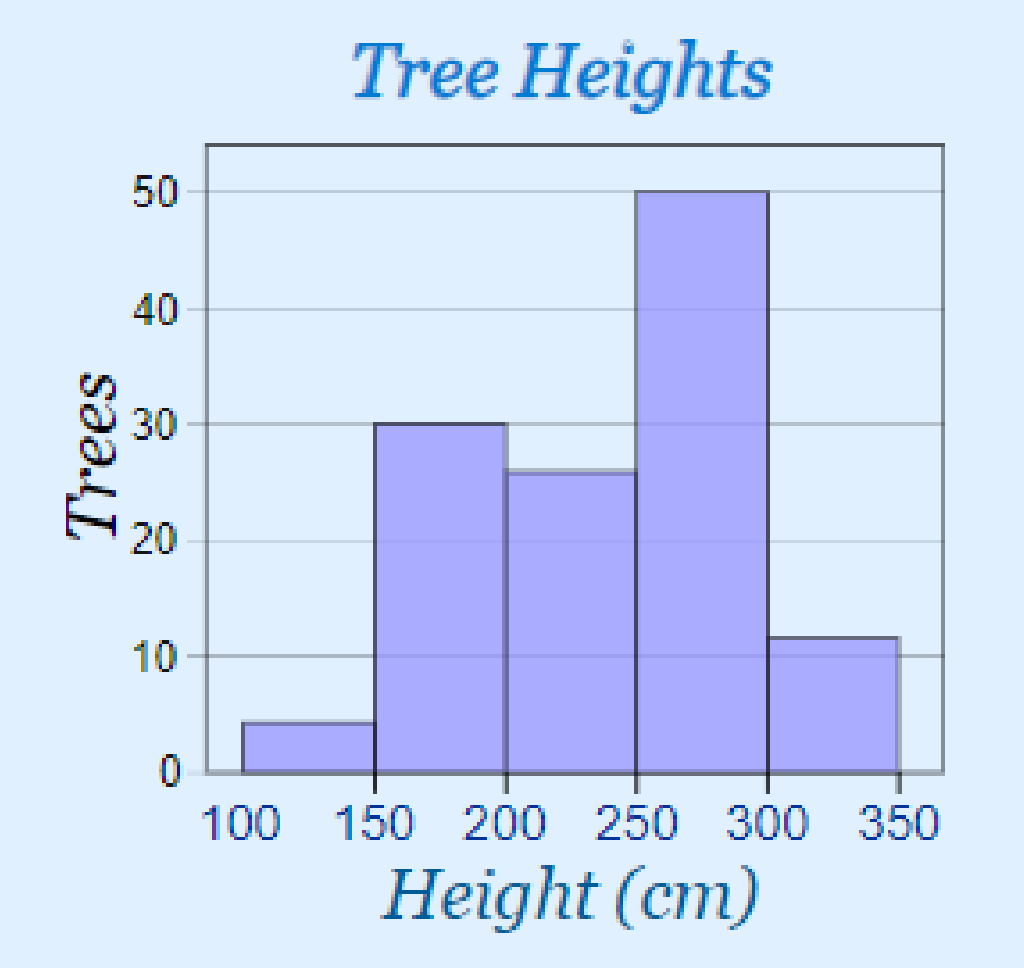
Data visualization: Barcharts - Contrasting



Stacked barcharts



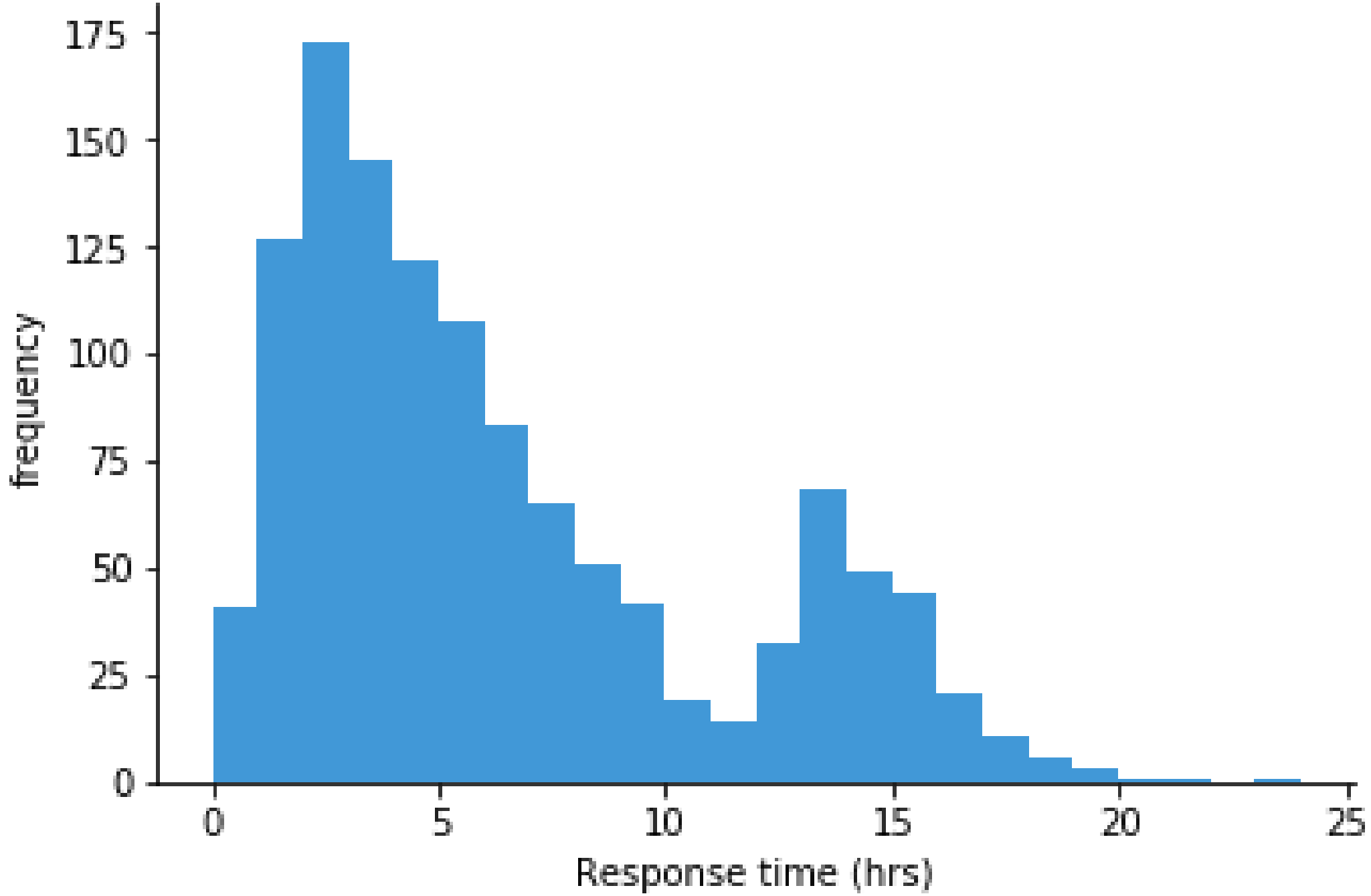
Data visualization: Histograms

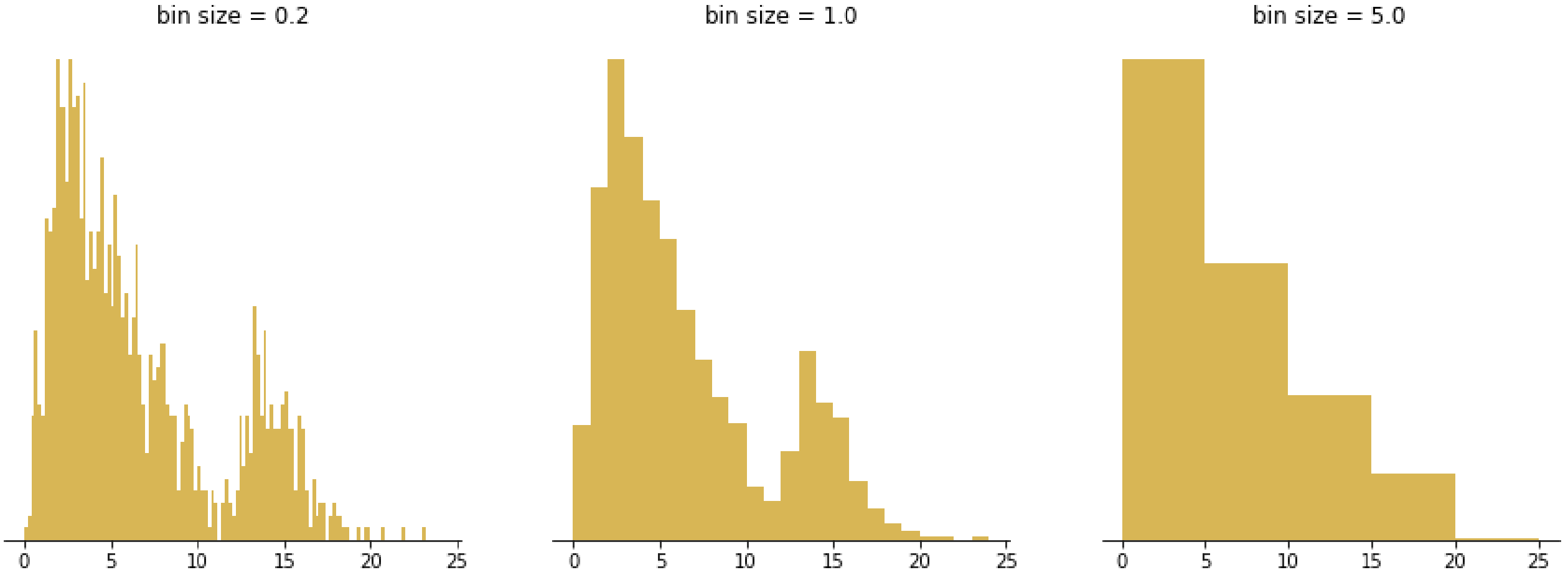


An histogram is an accurate graphical representation of the **distribution of numerical data**.

It takes as input one numerical variable.

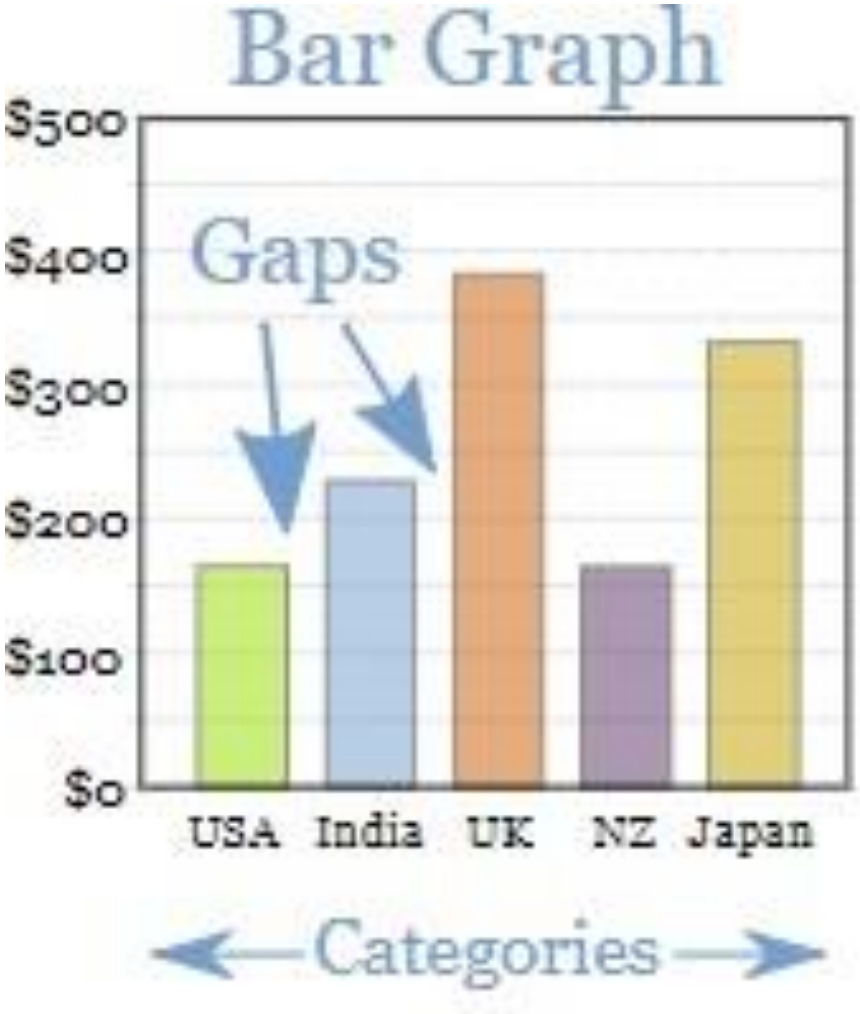
The variable is **cut into several bins**, and **the number of observations per bin** is represented by **the height of the bar**.


175
150
125
100
75
50
25
0
frequency
0
5
10
15
20
25
Response time (hrs)


bin size = 0.2
bin size = 1.0
bin size = 5.0
0
5
10
15
20
25

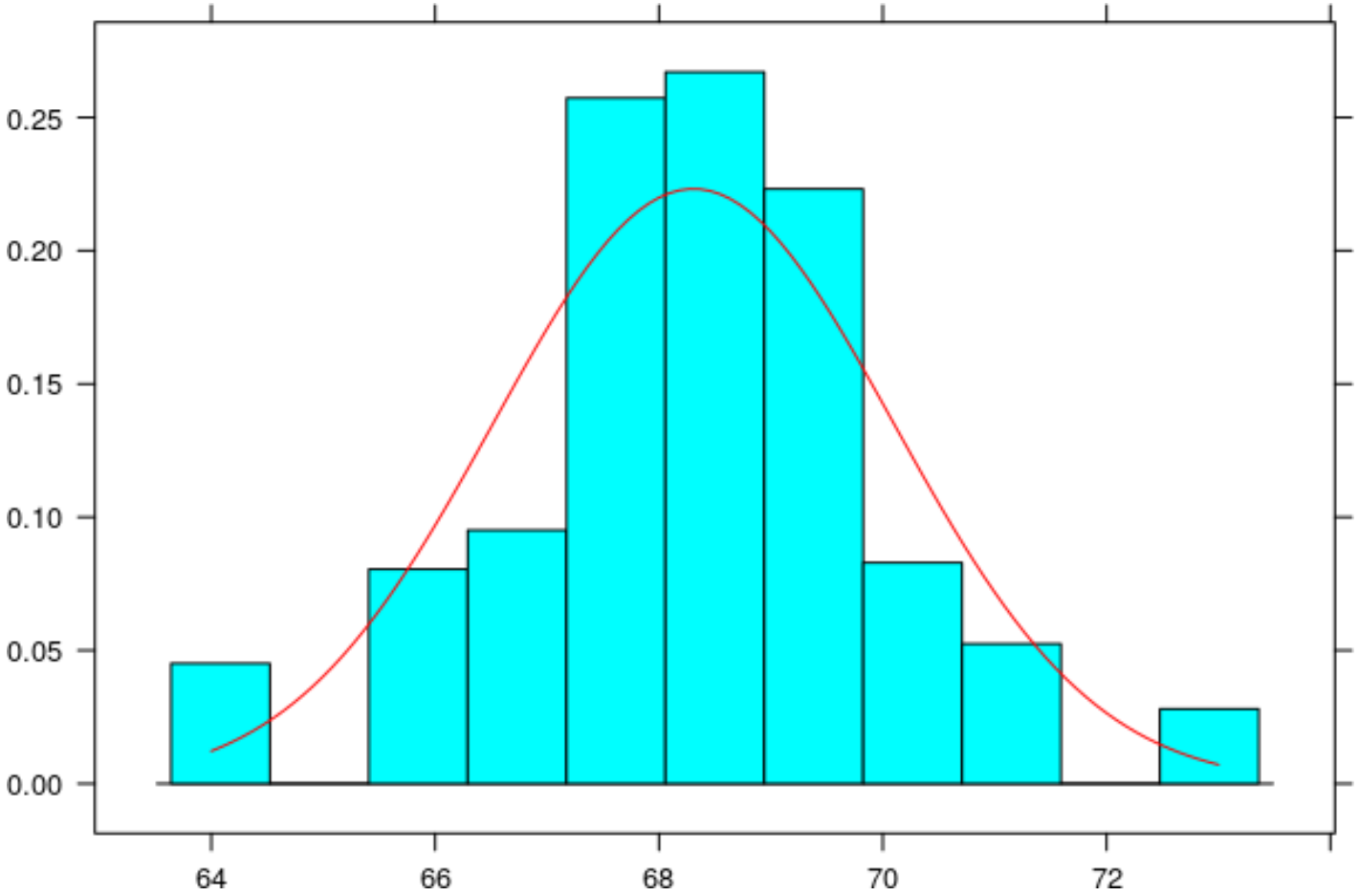
Barcharts vs. Histograms

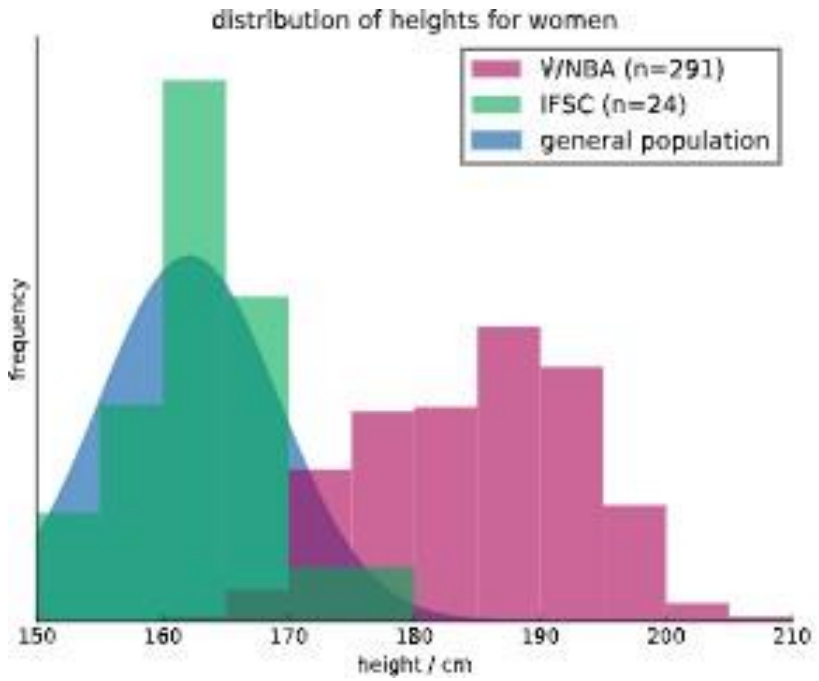
While barcharts are for categorical data, histograms are for continuous data (time, height, etc)

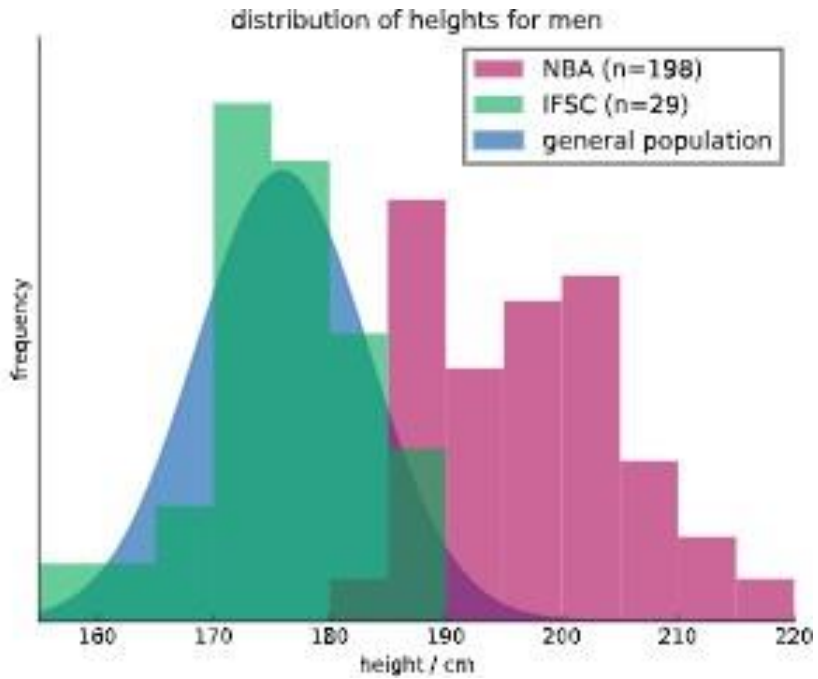


Data visualization: Density plots

Also known as: Smoothed histograms

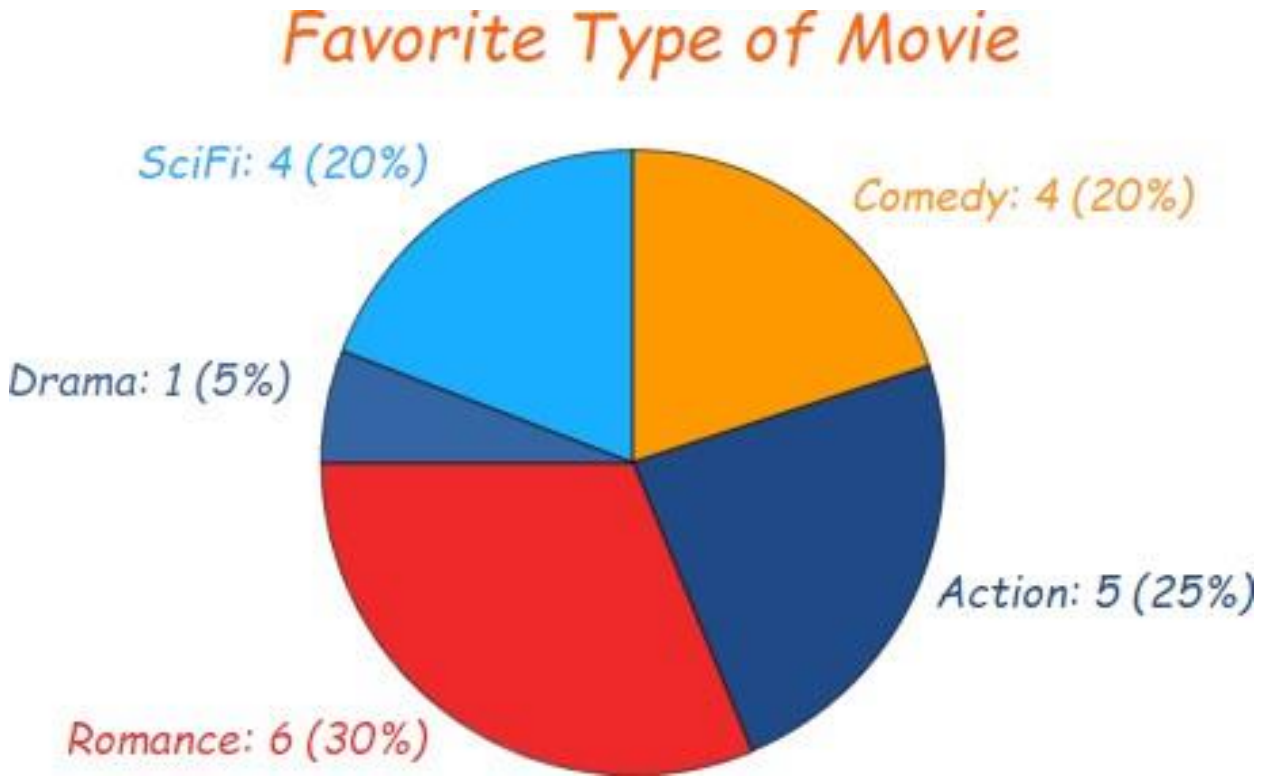



distribution of heights for women
W/NBA (n=291)
IFSC (n=24)
general population
frequency
150
160
170
180
190
200
210
height / cm

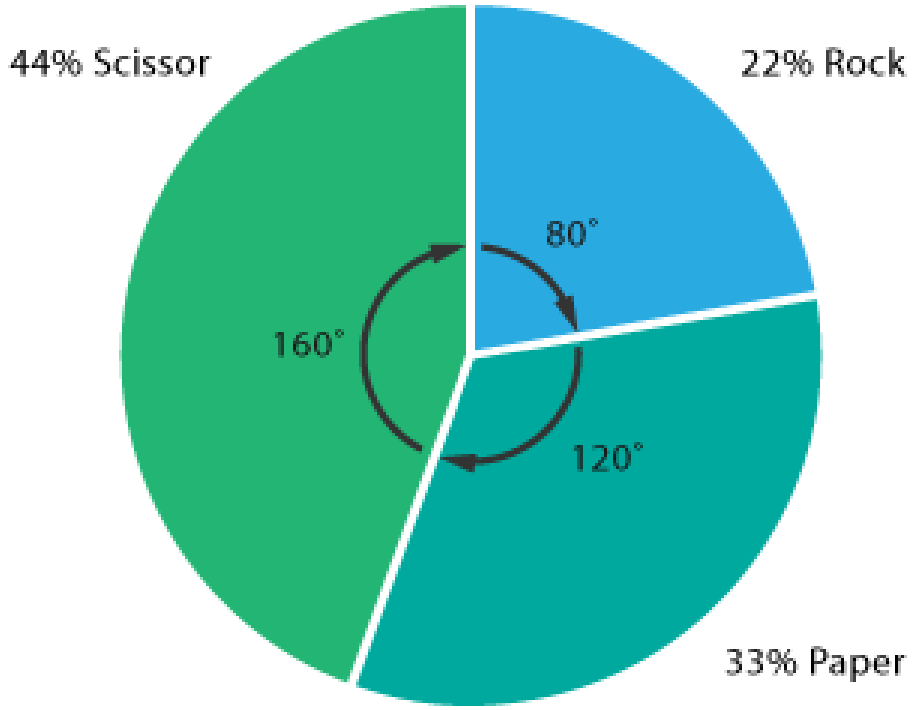

distribution of heights for men
NBA (n=198)
IFSC (n=29)
general population
frequency
160
170
180
190
200
210
220
height / cm

Data visualization: Piecharts

Table: Favorite Type of Movie				
Comedy	**Action**	**Romance**	**Drama**	**SciFi**
4	5	6	1	4

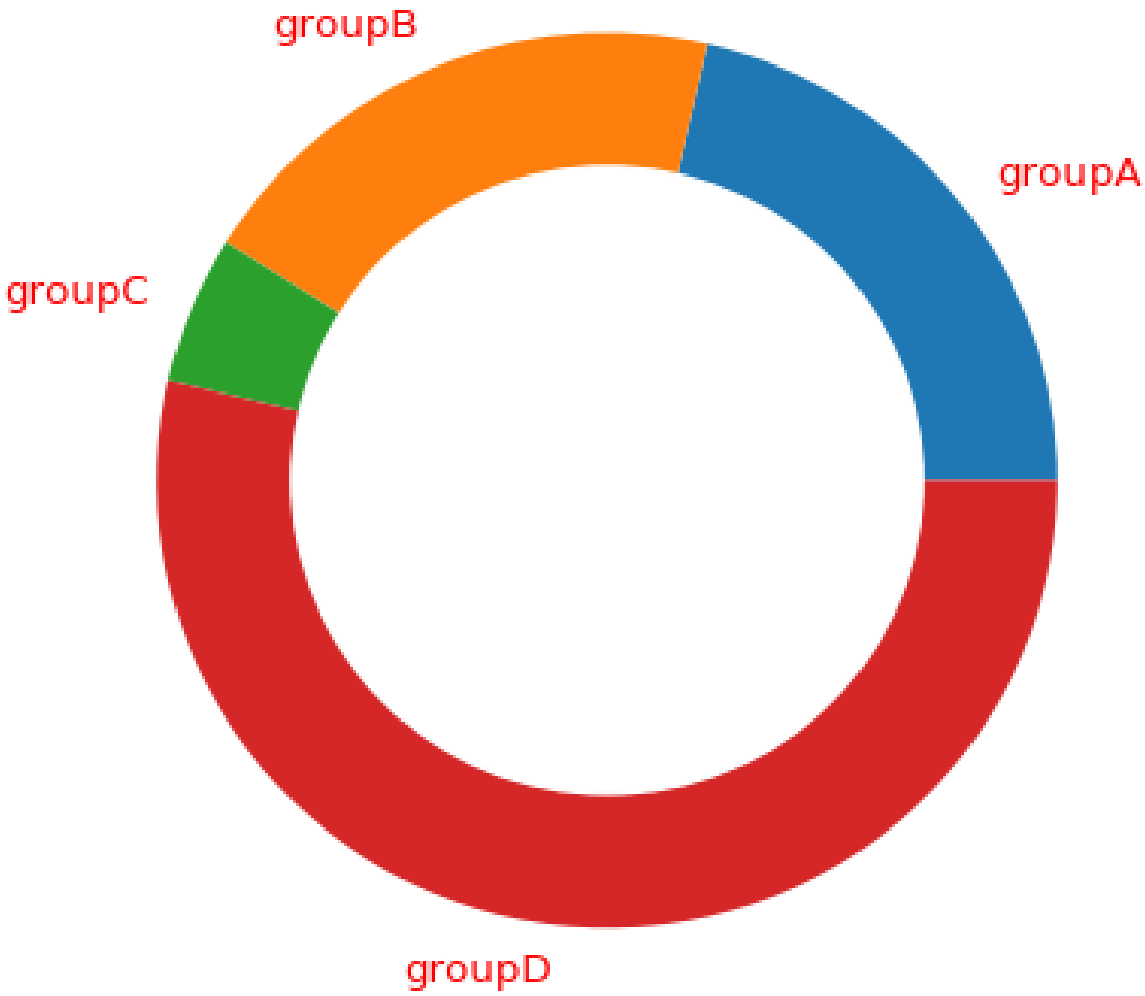


Data visualization: Piecharts

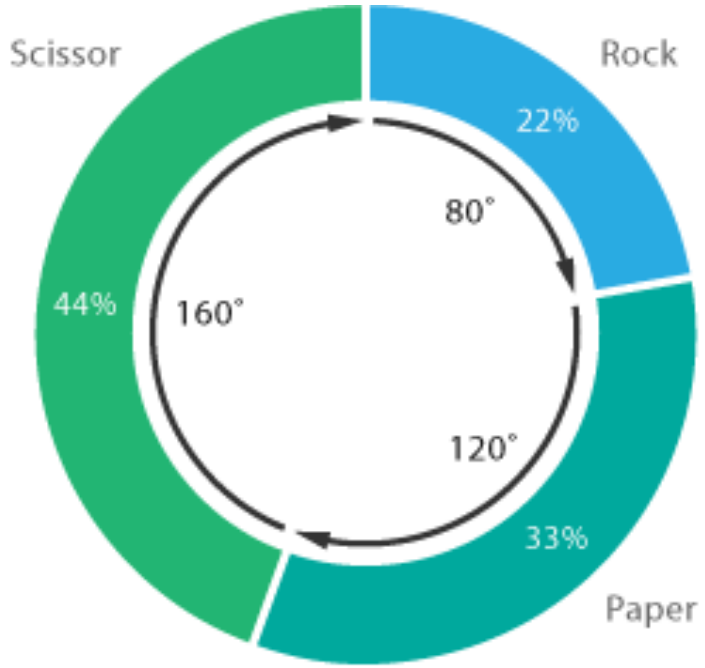


Data			
Rock	Paper	Scissor	TOTAL
2	3	4	9
To calculate percentages			
2/9=22%	3/9=33%	4/9=44%	100%
Degrees for each "pie slice"			
(2/9) x 360 = 80°	(3/9) x 360 = 120°	(4/9) x 360 = 160°	360°

Data visualization: Donut plots (similar to piecharts)

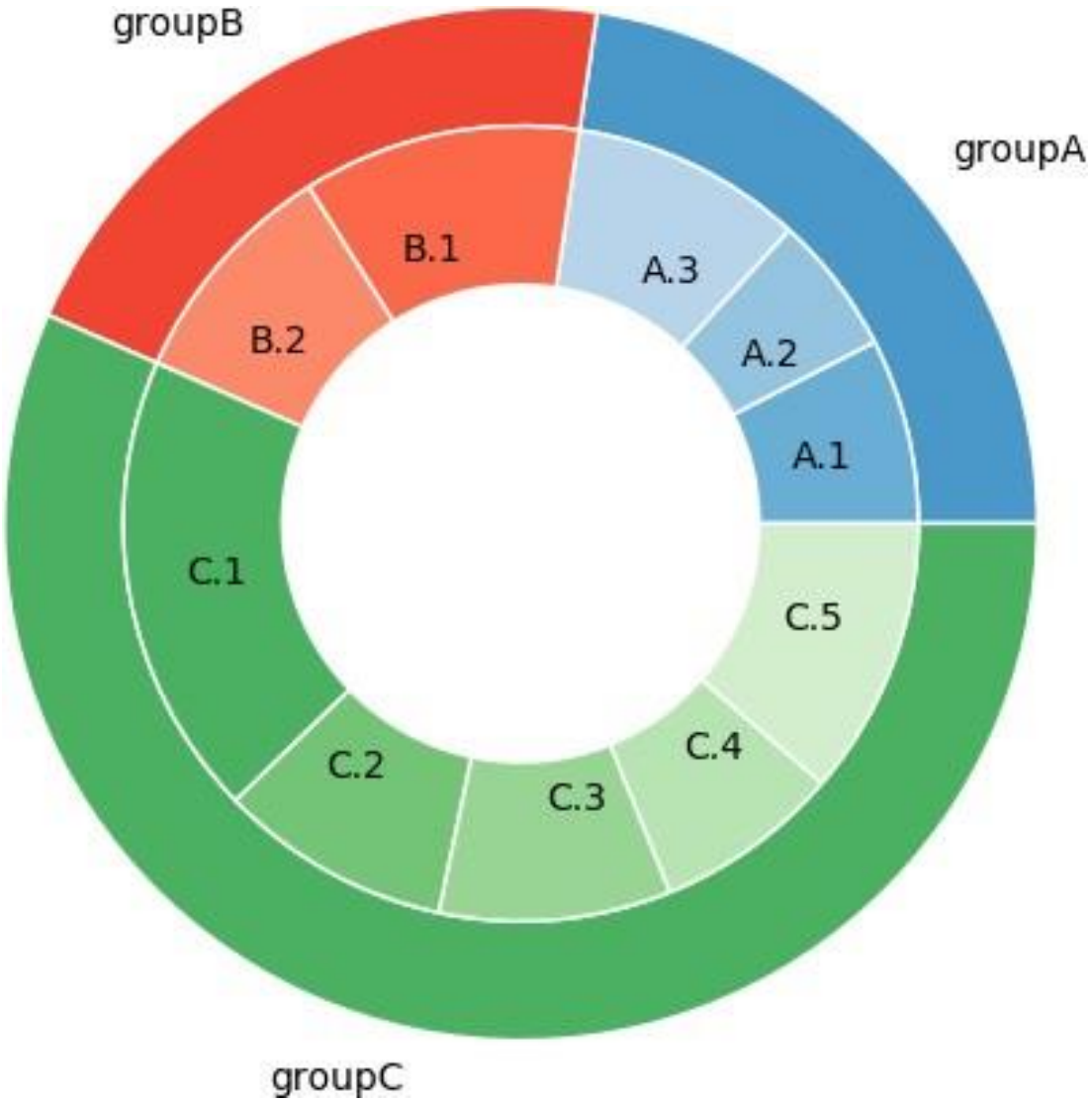


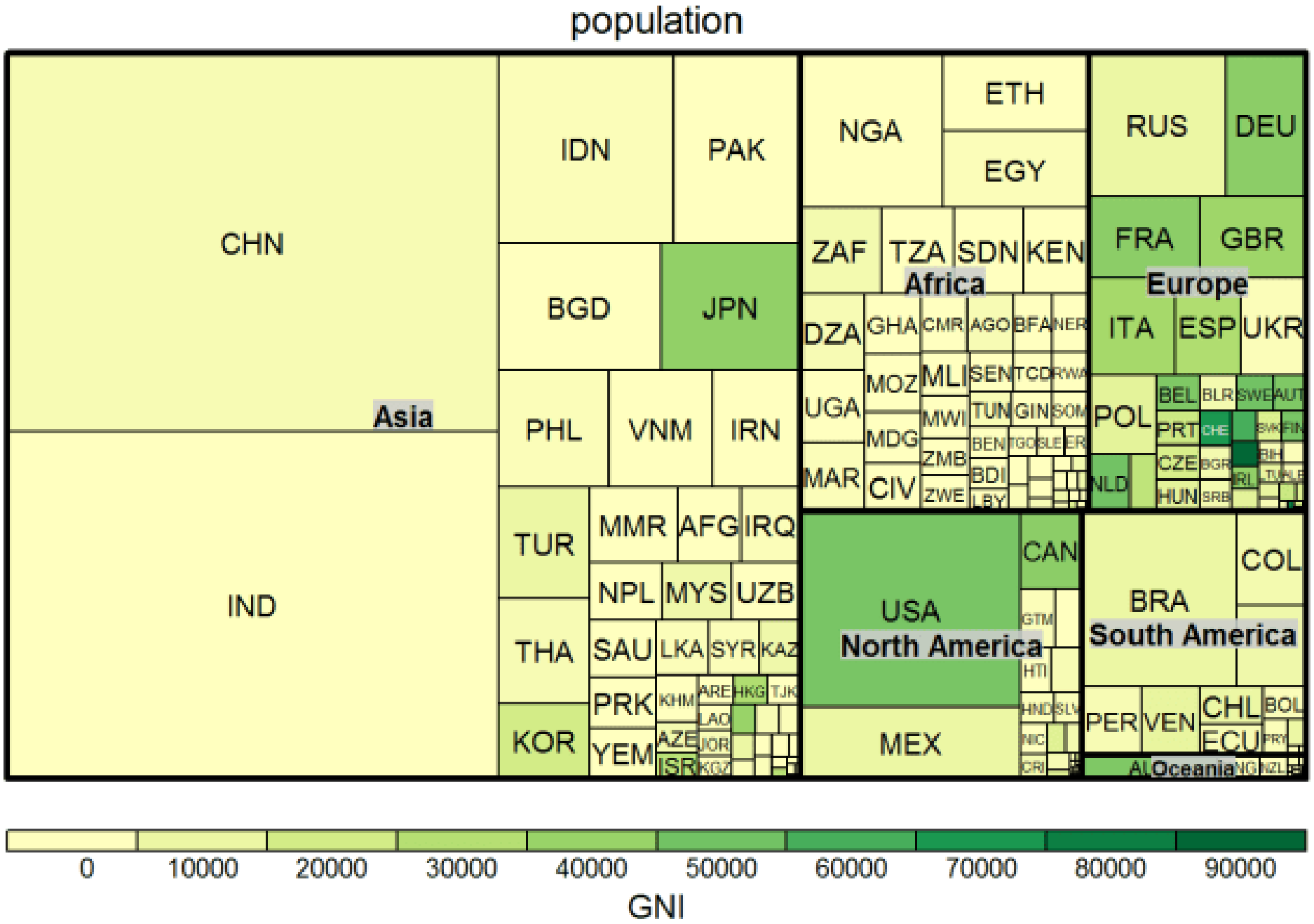
Data visualization: Donut plots (similar to piecharts)



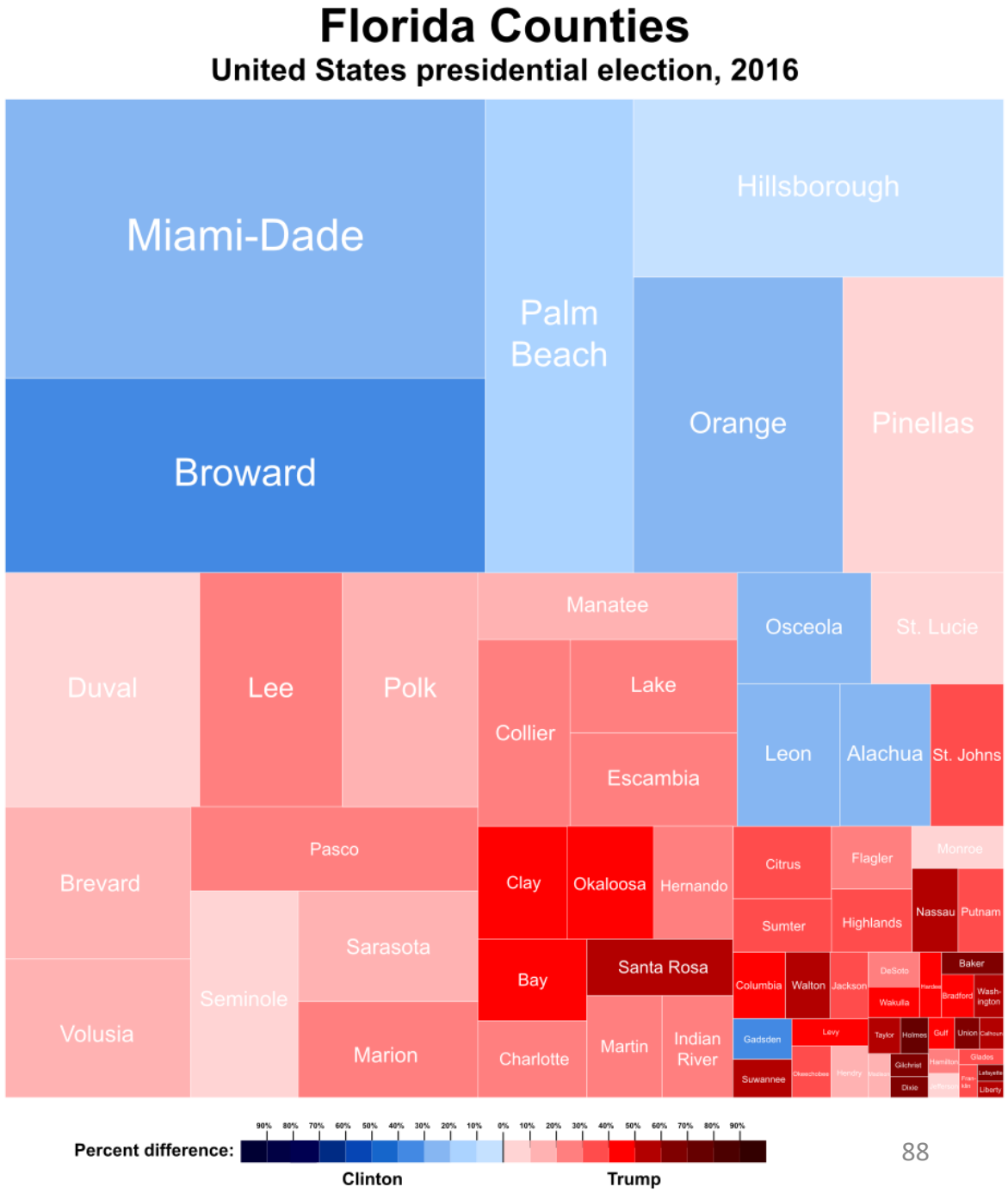
Data			
Rock	Paper	Scissor	TOTAL
2	3	4	9
To calculate percentages			
2/9 = **22%**	3/9 = **33%**	4/9 = **44%**	100%
Degrees for each "donut slice"			
(2/9) x 360 = **80°**	(3/9) x 360 = **120°**	(4/9) x 360 = **160°**	360°

Data visualization: Donut with subplots

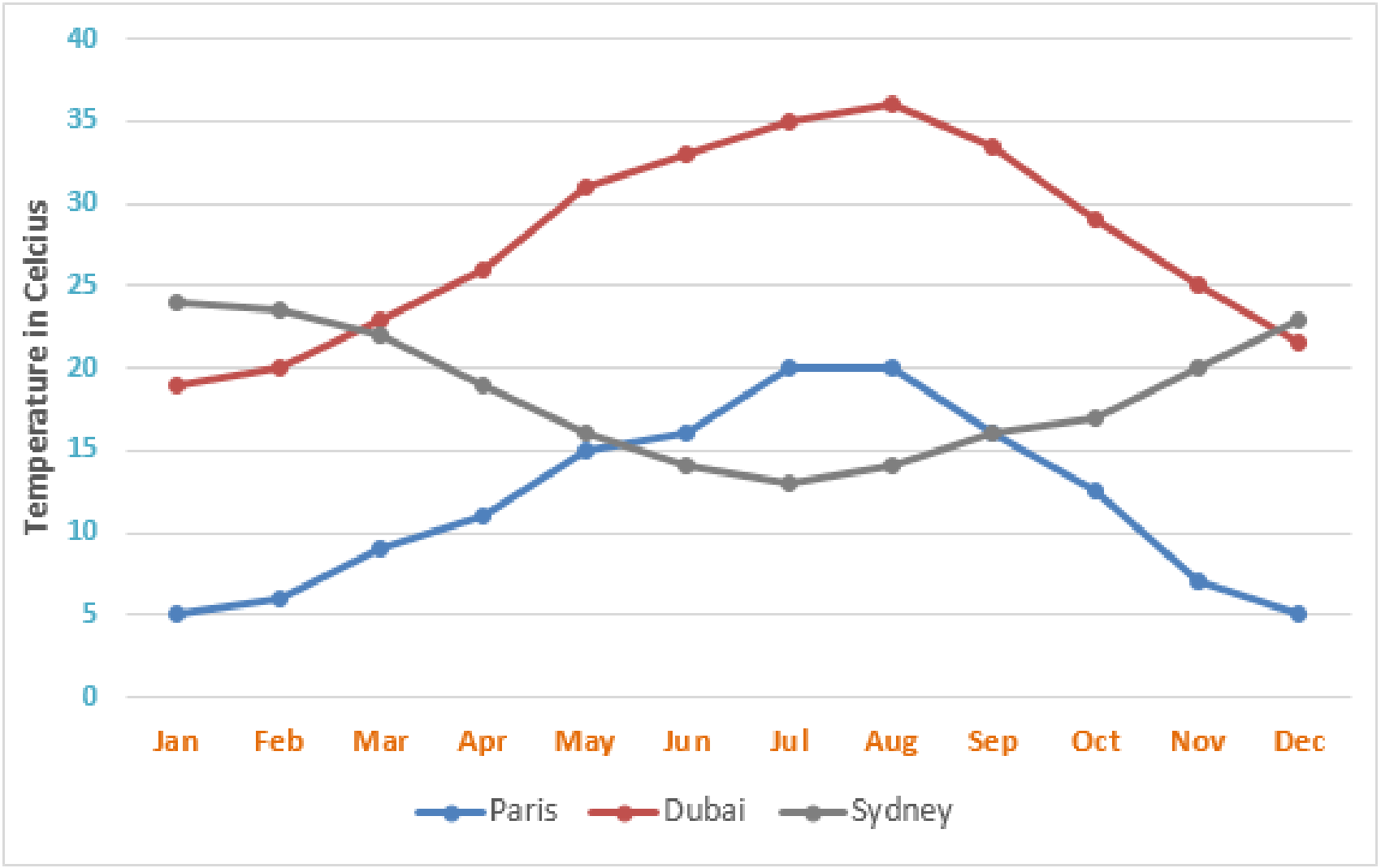



population
CHN
Asia
IND
IDN
PAK
BGD
JPN
PHL
VNM
IRN
TUR
MMR
AFG
IRQ
NPL
MYS
UZB
THA
SAU
LKA
SYR
KAZ
PRK
KHM
ARE
HKG
TJK
LAO
KOR
YEM
AZE
JOR
ISR
KGZ
NGA
ETH
EGY
ZAF
TZA
SDN
KEN
Africa
DZA
GHA
CMR
AGO
BFA
NER
UGA
MOZ
MLI
SEN
TCD
RWA
MWI
TUN
GIN
SOM
MDG
BEN
TGO
SLE
ER
MAR
CIV
ZMB
ZWE
BDI
LBY
RUS
DEU
FRA
GBR
Europe
ITA
ESP
UKR
POL
BEL
BLR
SWE
AUT
PRT
CHE
SVK
FIN
CZE
BGR
BIH
IRL
NLD
HUN
SRB
USA
North America
CAN
GTM
HTI
HND
SLV
NIC
CRI
MEX
BRA
South America
COL
PER
VEN
CHL
ECU
BOL
PRY
Oceania
AUS
NZL
0
10000
20000
30000
40000
50000
60000
70000
80000
90000
GNI

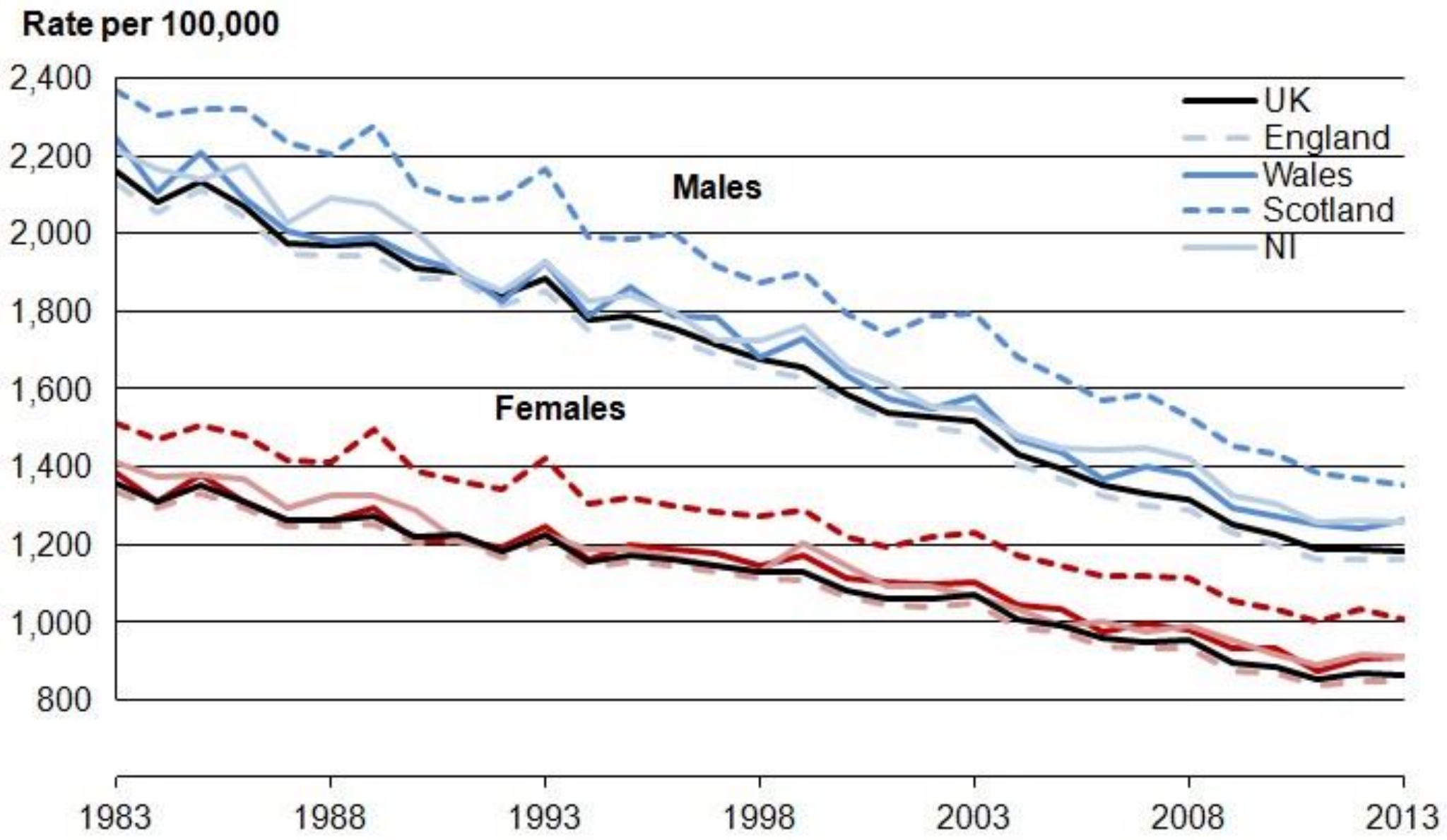
Data visualization: Treemaps



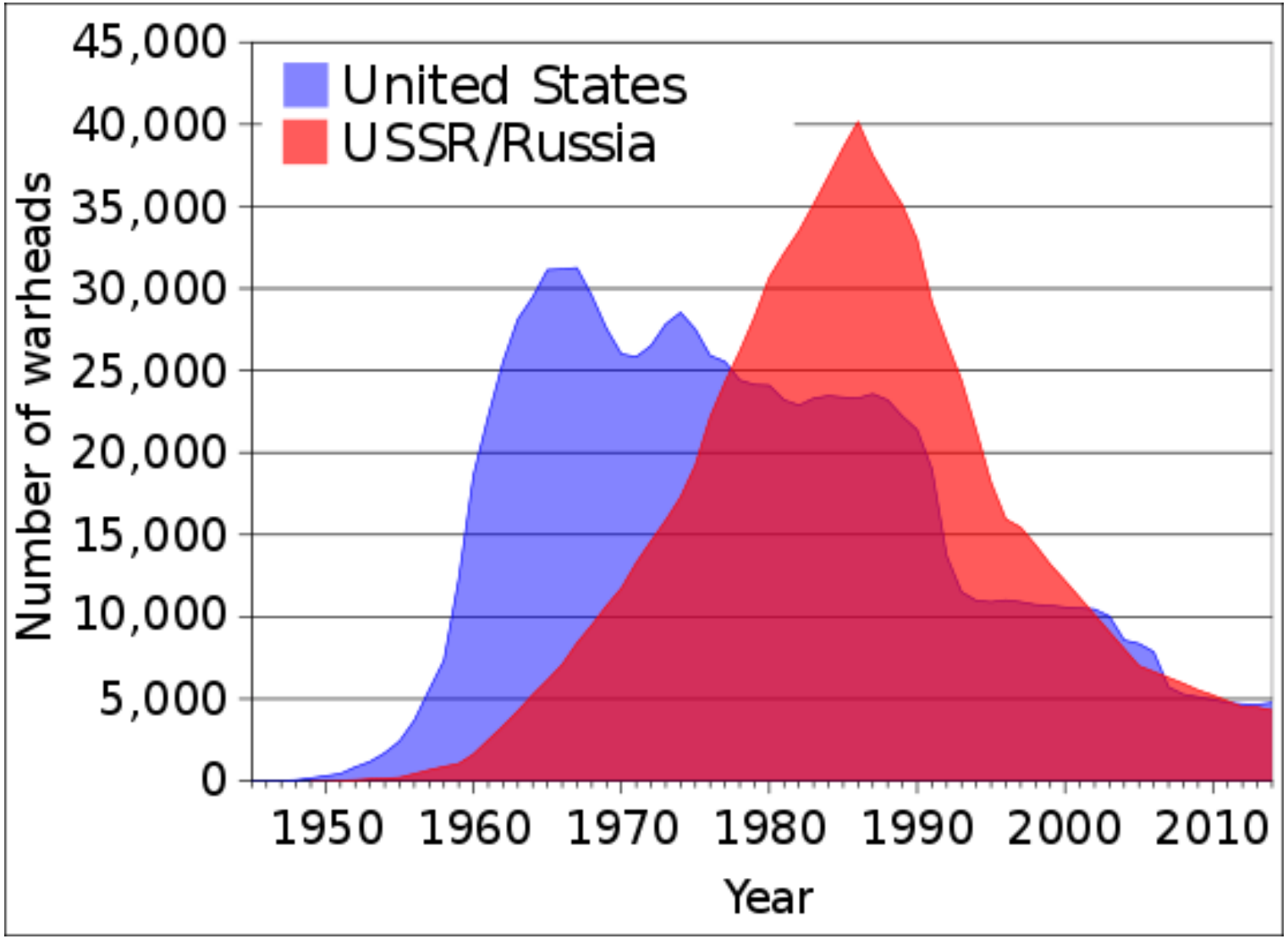
Data visualization: Linecharts



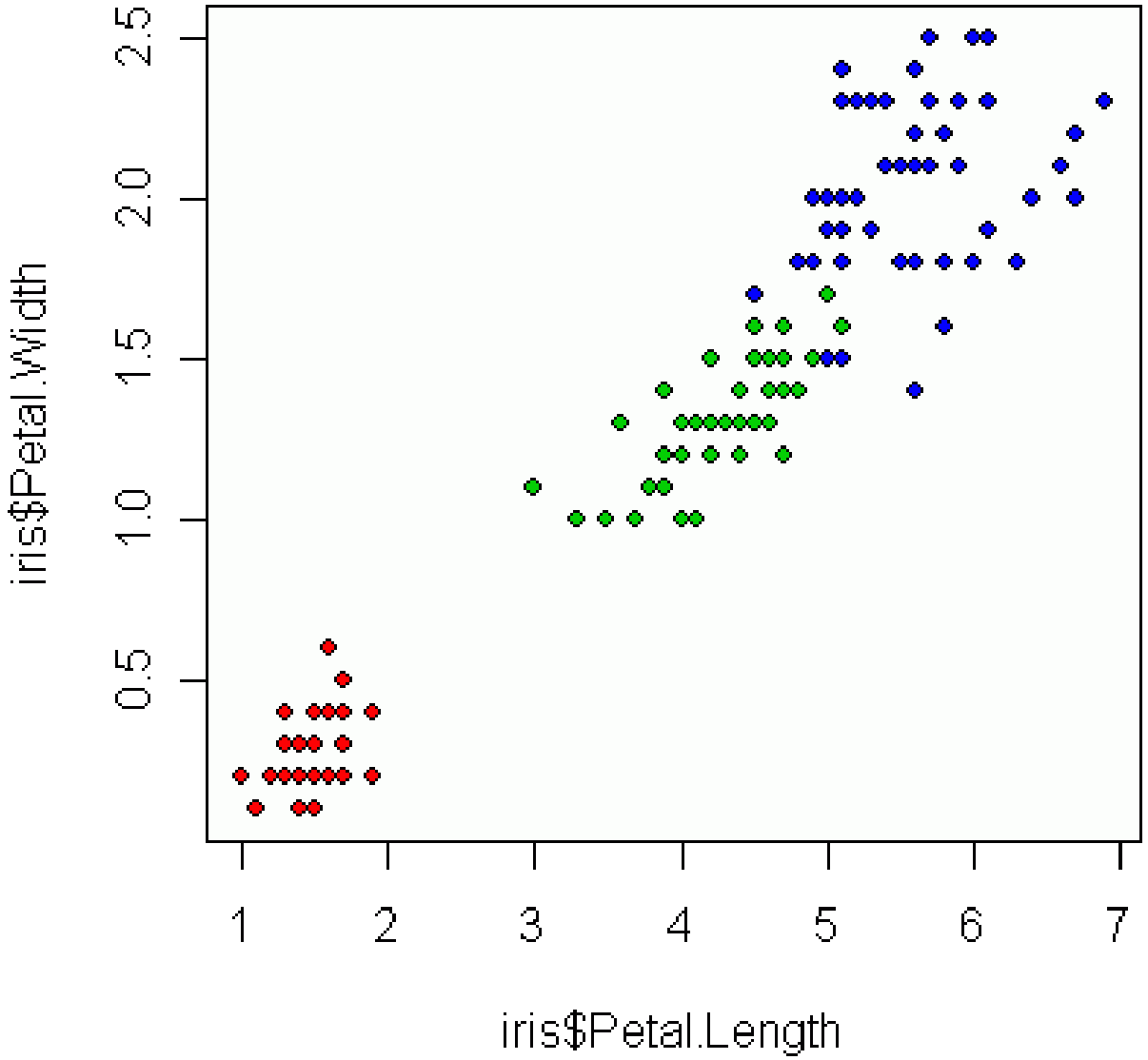
Age-standardized mortality rates



Data visualization: Area graphs



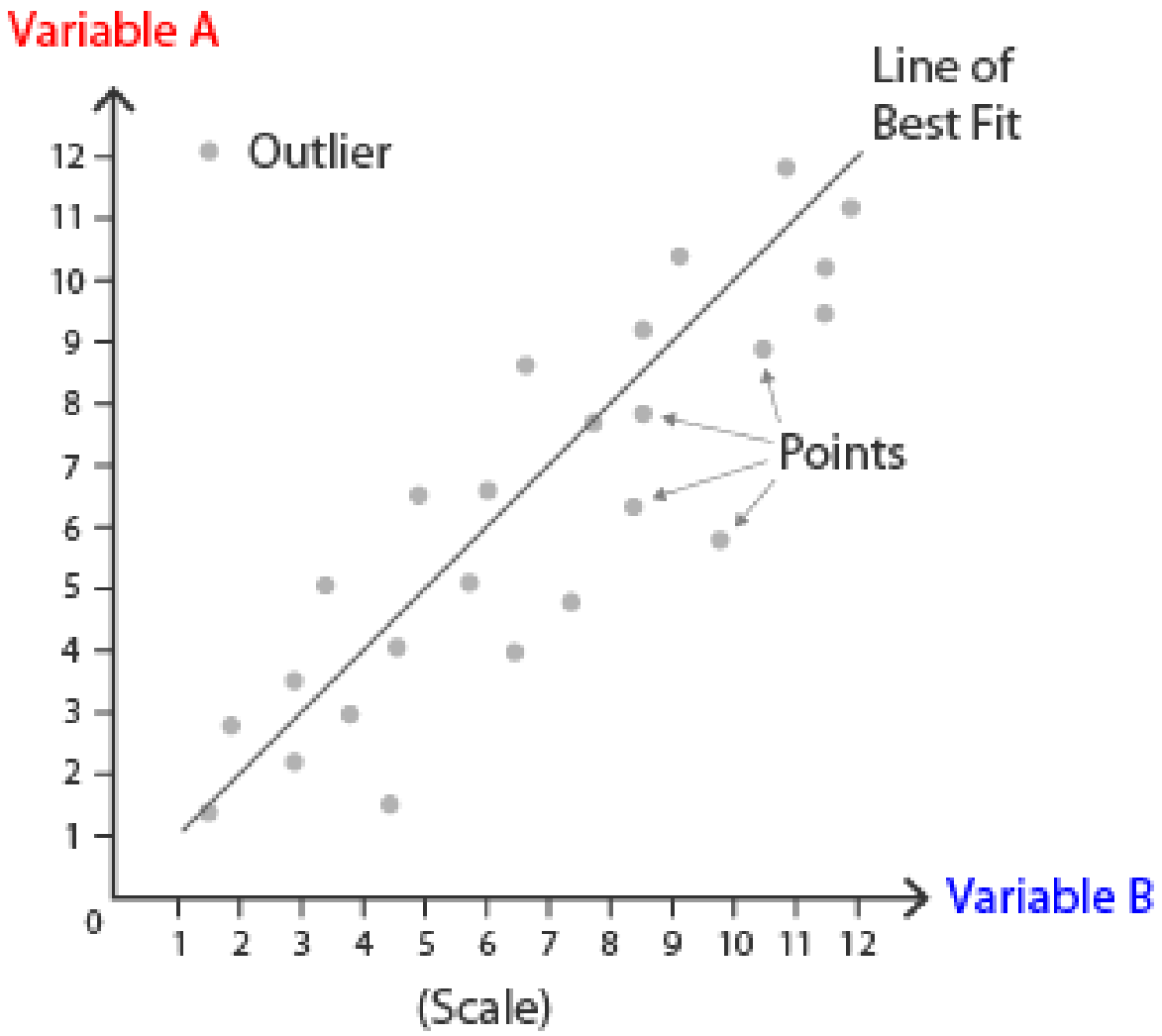
Data visualization: Scatterplots

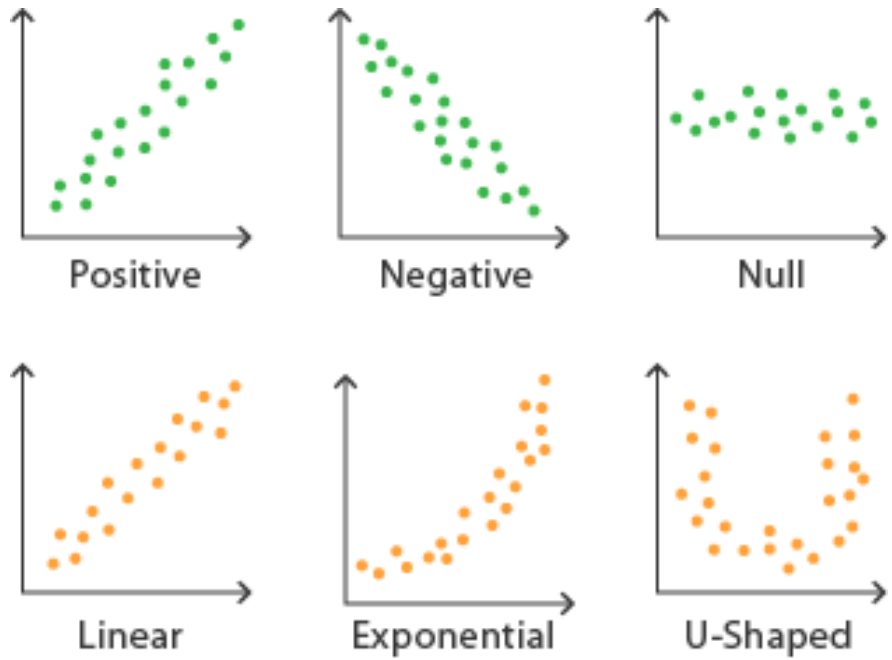


Scatter plot - is used to display a set of bi-variate data (two variables), usually displayed before computing a linear correlation coefficient or fitting a regression/classification line.

The scatter shows the strength of the relationship between the two numerical variables.

Data visualization: Scatterplots

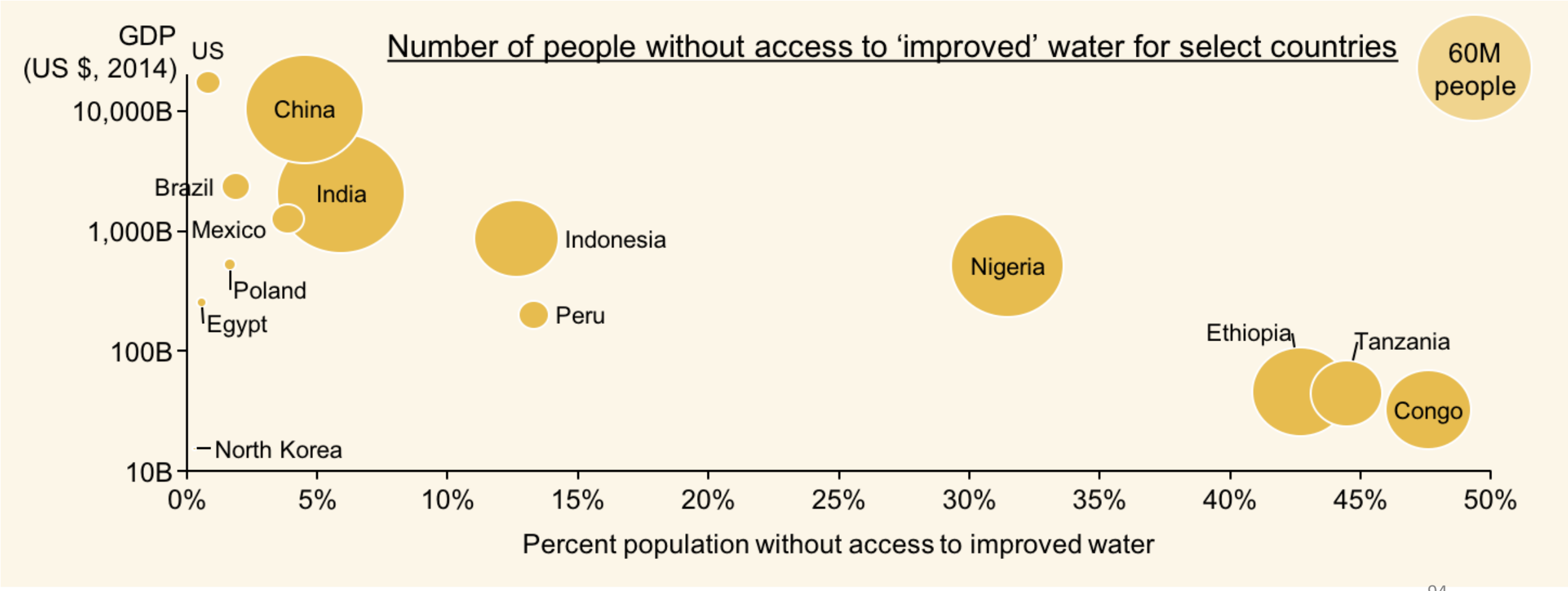




Correlation Strength:

Data visualization: Bubble plots

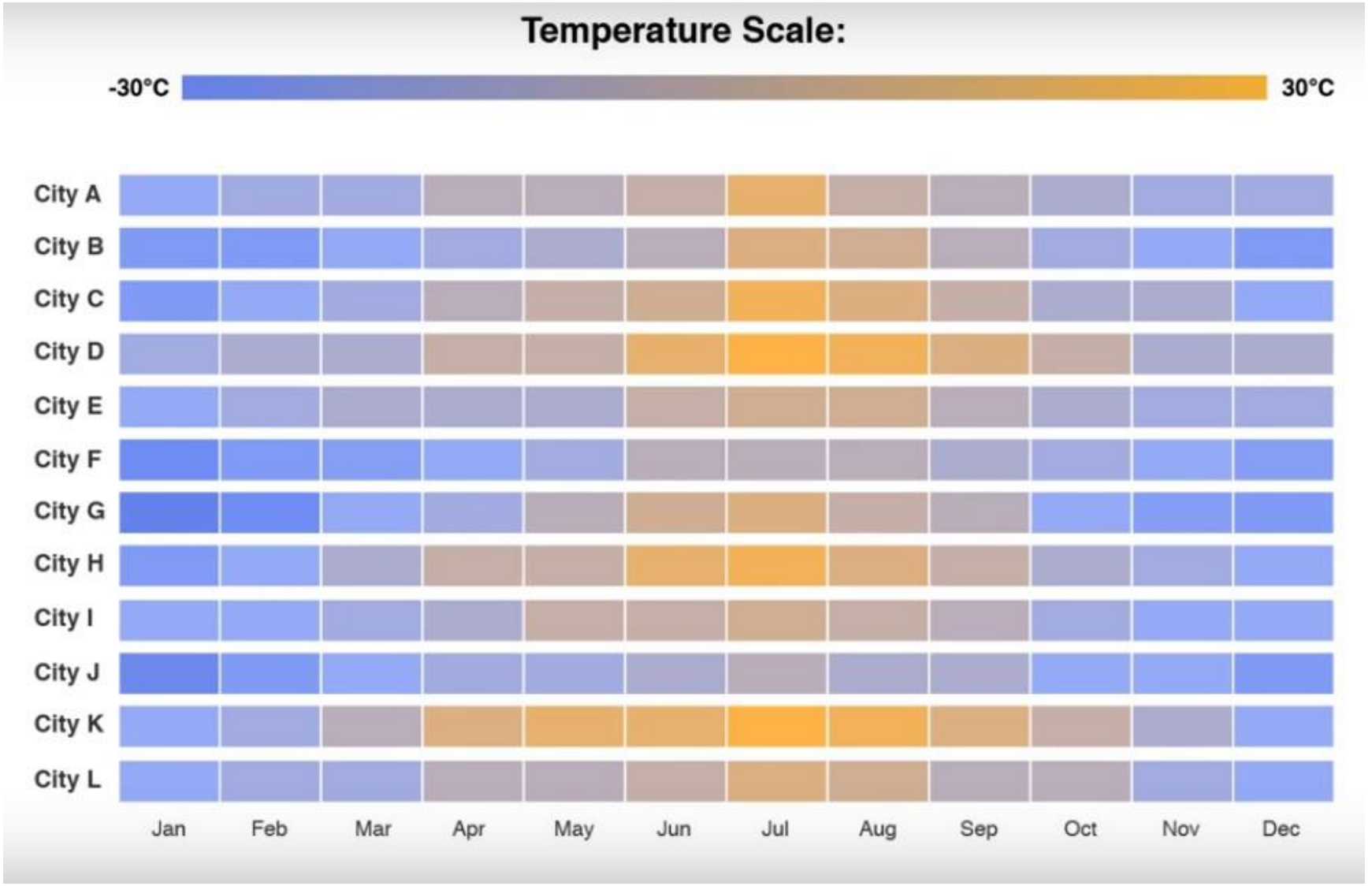
A bubble plot is a scatterplot where a third dimension is added: the value of an additional variable is represented through the size of the dots. You need 3 numerical variables as input: one is represented by the X axis, one by the Y axis, and one by the size.

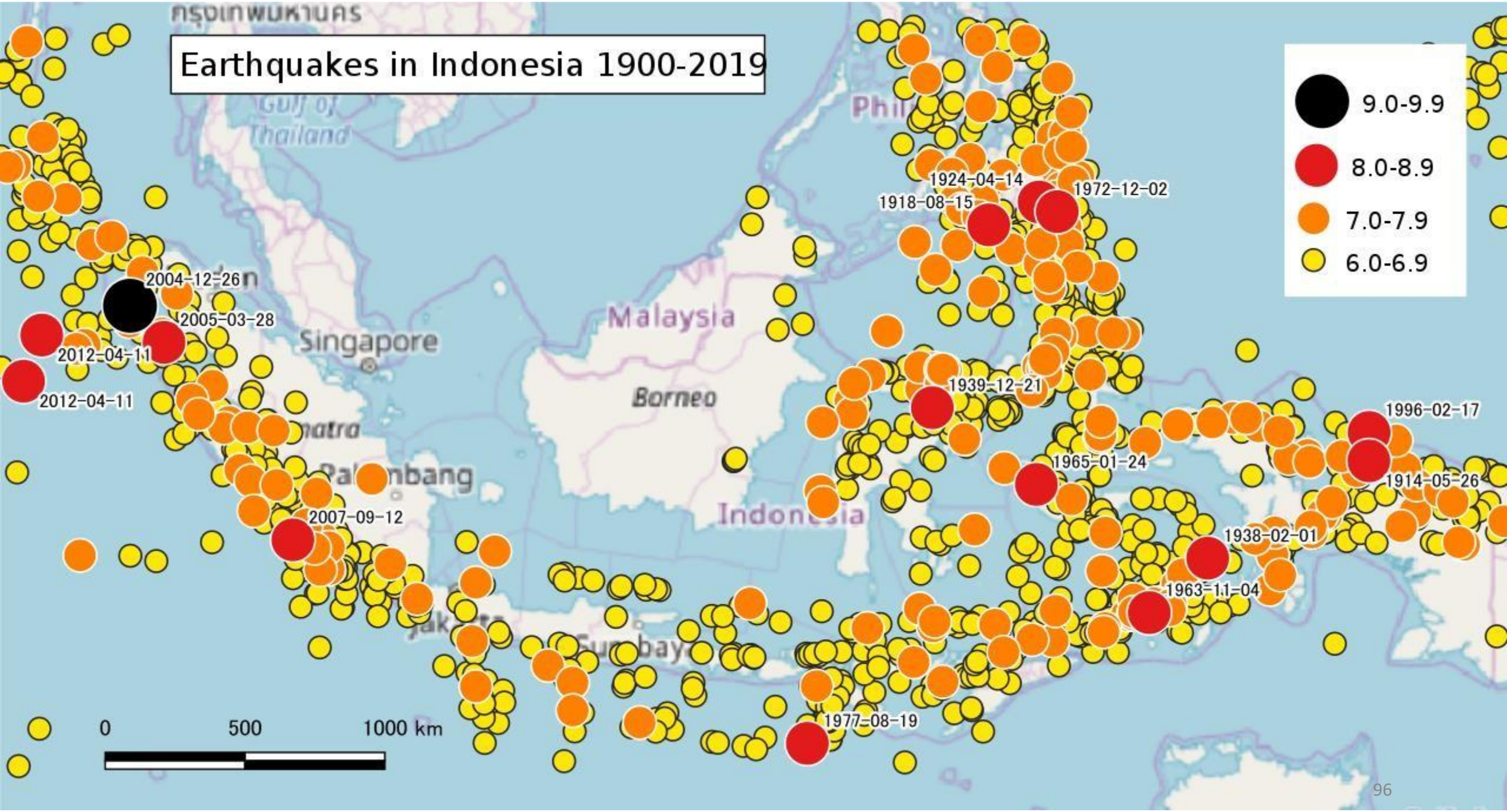


Source: National Geographic (2015 data)

Note: 'Improved' water is water from pipes or wells that are protected from contamination

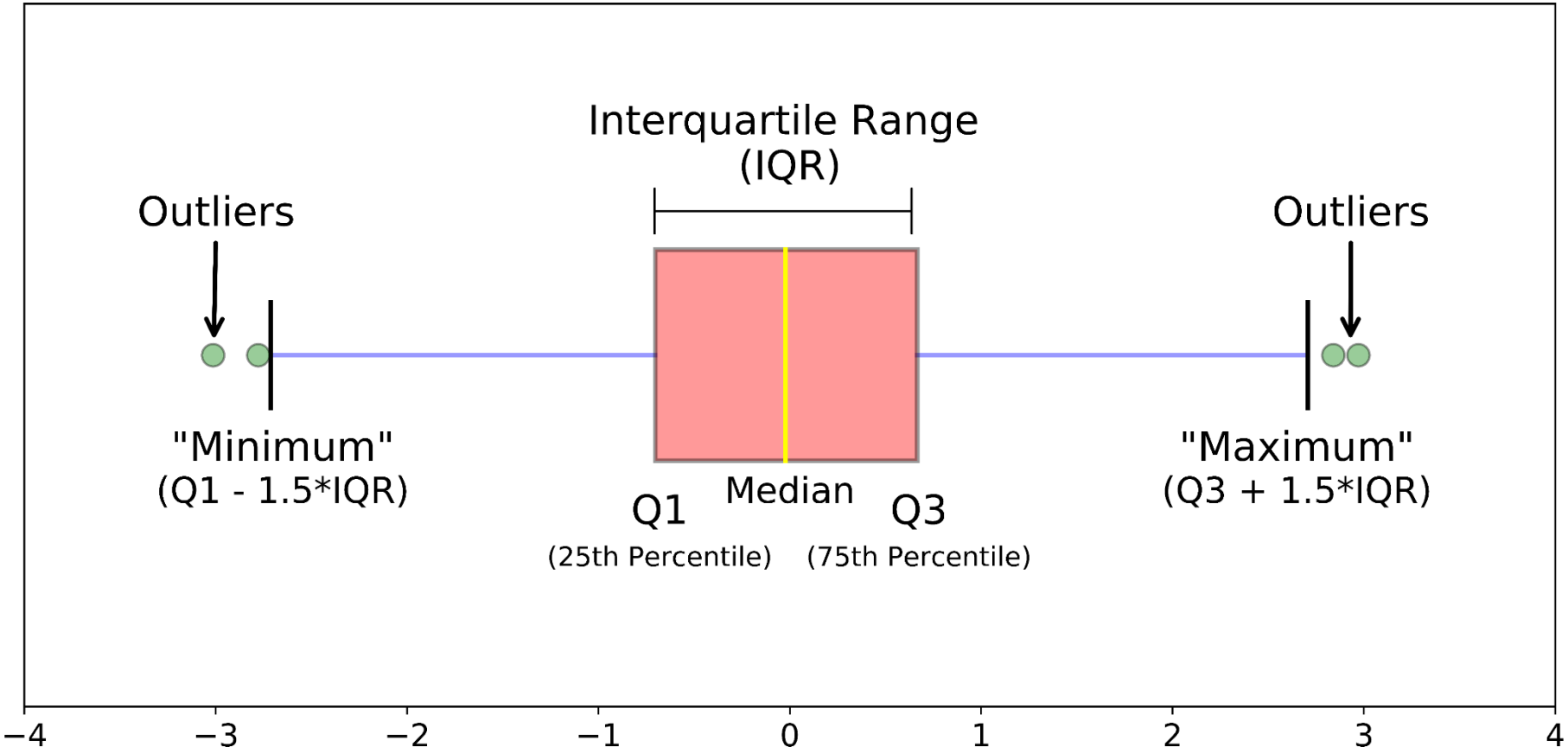
Data visualization: Heatmap



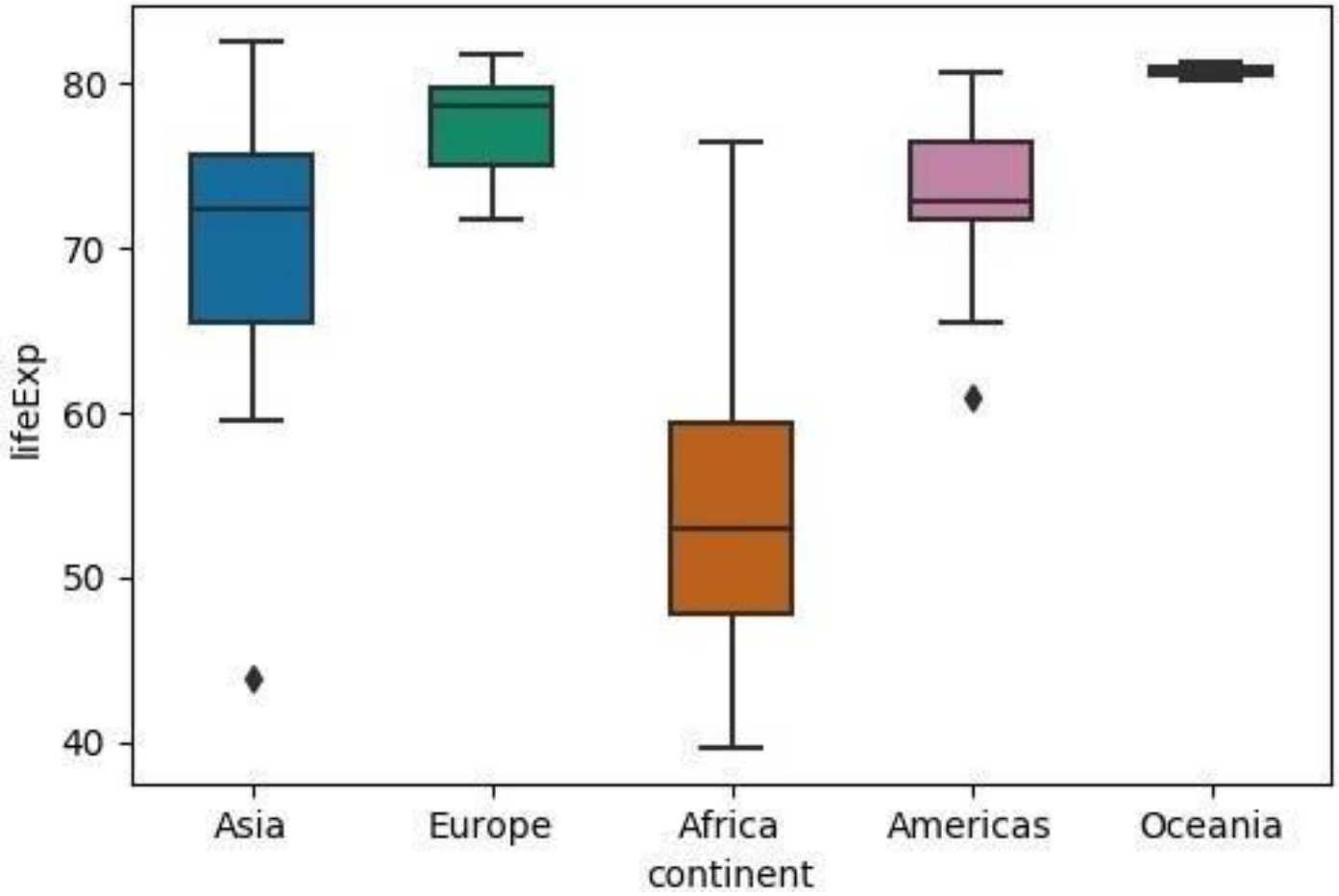

Earthquakes in Indonesia 1900-2019
9.0-9.9
8.0-8.9
7.0-7.9
6.0-6.9
2004-12-26
2005-03-28
2012-04-11
2012-04-11
2007-09-12
1924-04-14
1918-08-15
1972-12-02
1939-12-21
1965-01-24
1996-02-17
1914-05-26
1938-02-01
1963-11-04
1977-08-19
Singapore
Malaysia
Borneo
Indonesia
Thailand
0
500
1000 km

Data visualization: Box Plots

A box plot is a graphical display of
selected summary measures for numeric variables.



Data visualization: Box Plots



Outliers: Visualized

The data set of N = 90 ordered observations as shown below is examined for outliers:

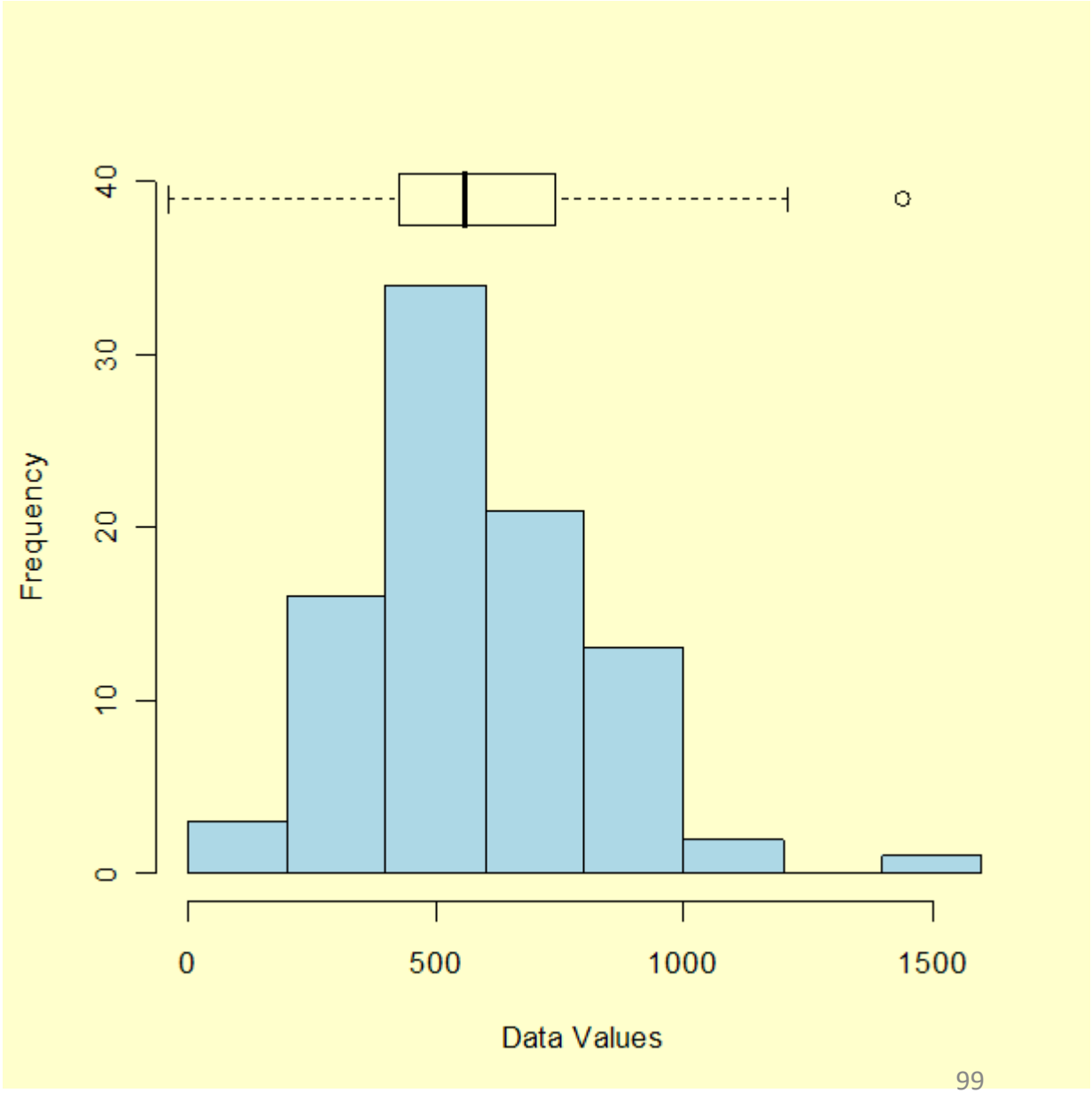
30, 171, 184, 201, 212, 250, 265, 270, 272, 289, 305, 306, 322, 322, 336, 346, 351, 370, 390, 404, 409, 411, 436, 437, 439, 441, 444, 448, 451, 453, 470, 480, 482, 487, 494, 495, 499, 503, 514, 521, 522, 527, 548, 550, 559, 560, 570, 572, 574, 578, 585, 592, 592, 607, 616, 618, 621, 629, 637, 638, 640, 656, 668, 707, 709, 719, 737, 739, 752, 758, 766, 792, 792, 794, 802, 818, 830, 832, 843, 858, 860, 869, 918, 925, 953, 991, 1000, 1005, 1068, 1441.

Q1 = 429.75

Q2 (Median) = 559.5

Q3 = 742.25

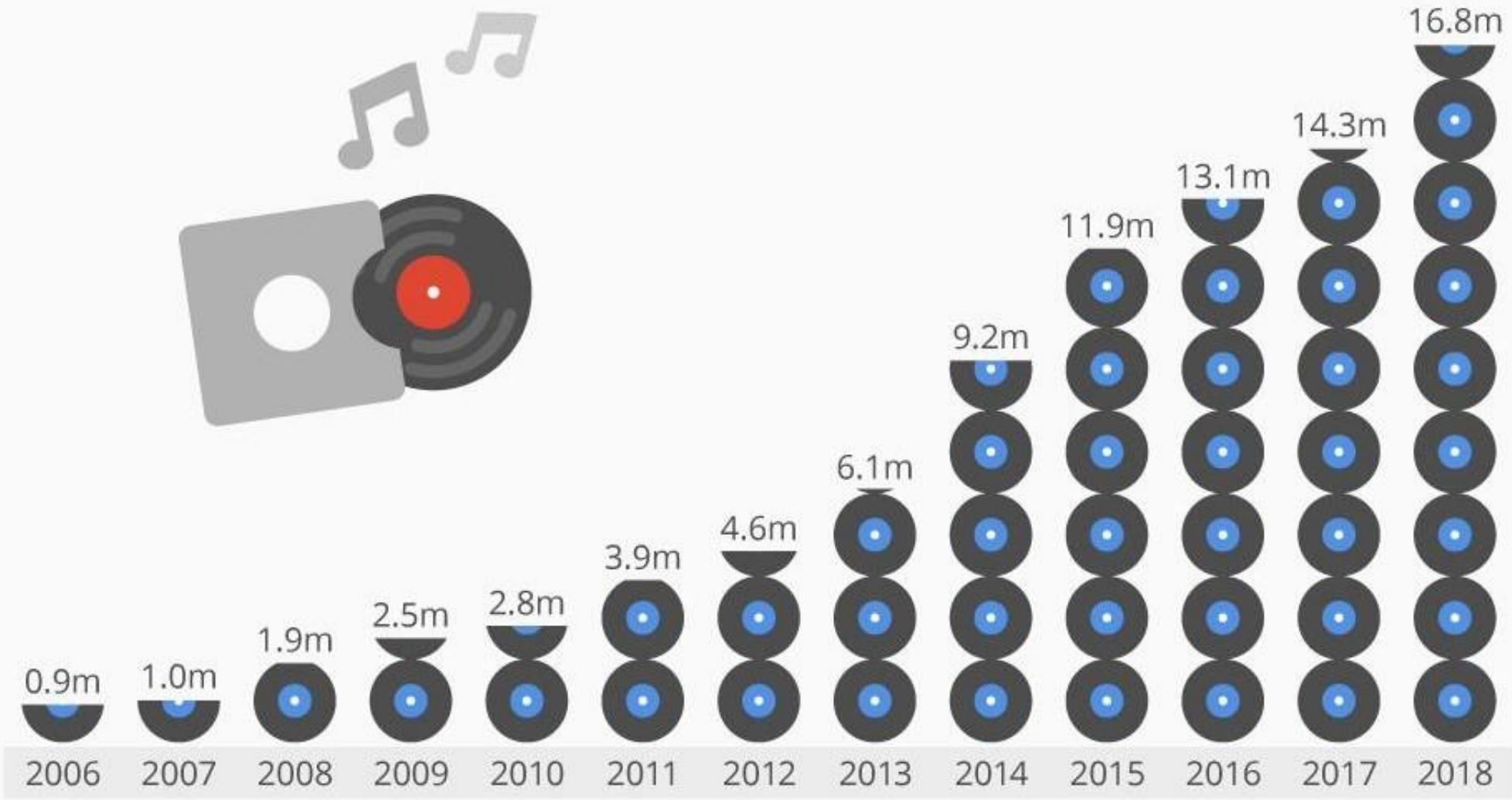
IQR = Q3-Q1 = 312.5



Bonus: Some interesting visualizations

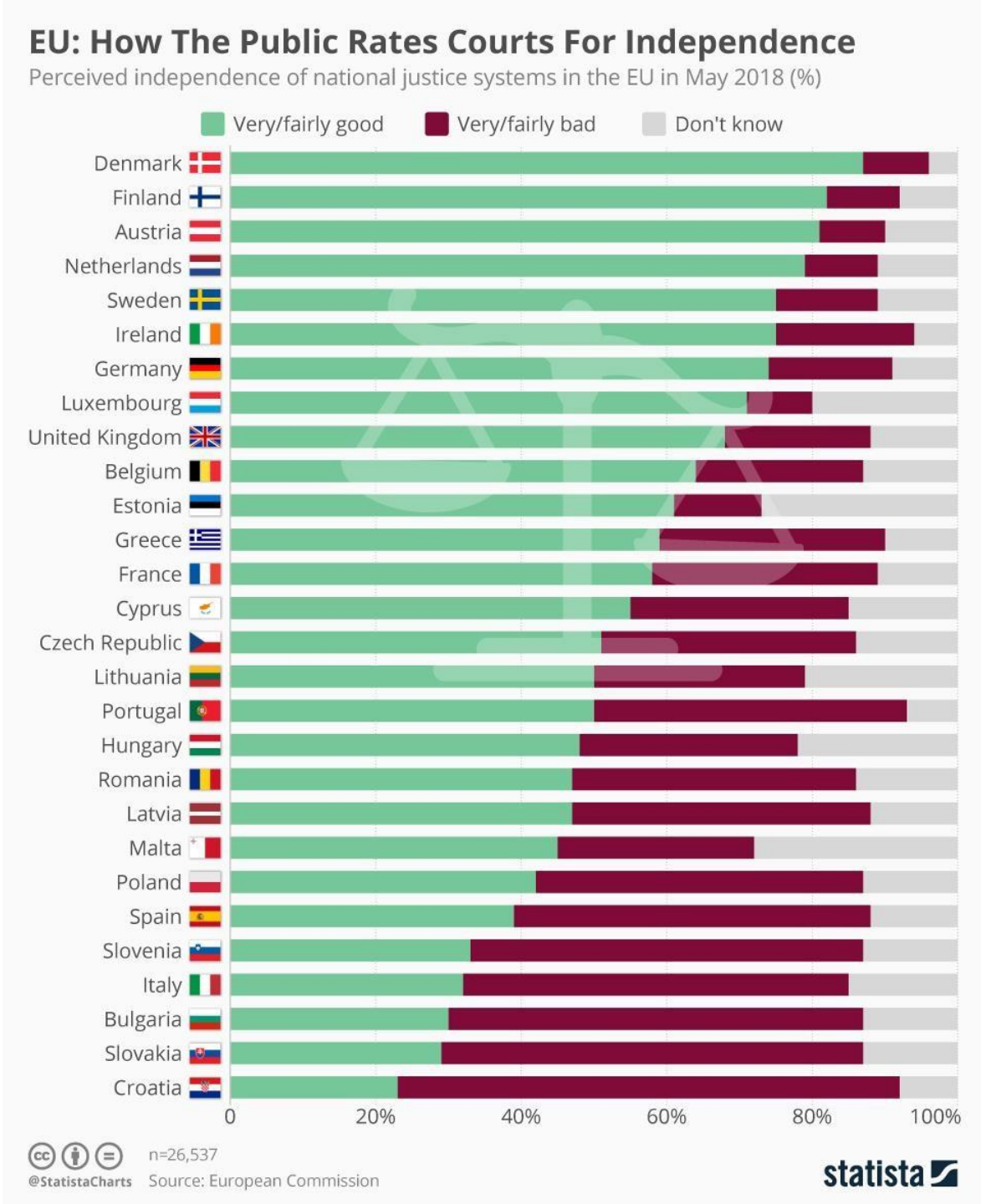
The Surprising Comeback of Vinyl Records

Vinyl LP unit sales in the United States from 2006 to 2018



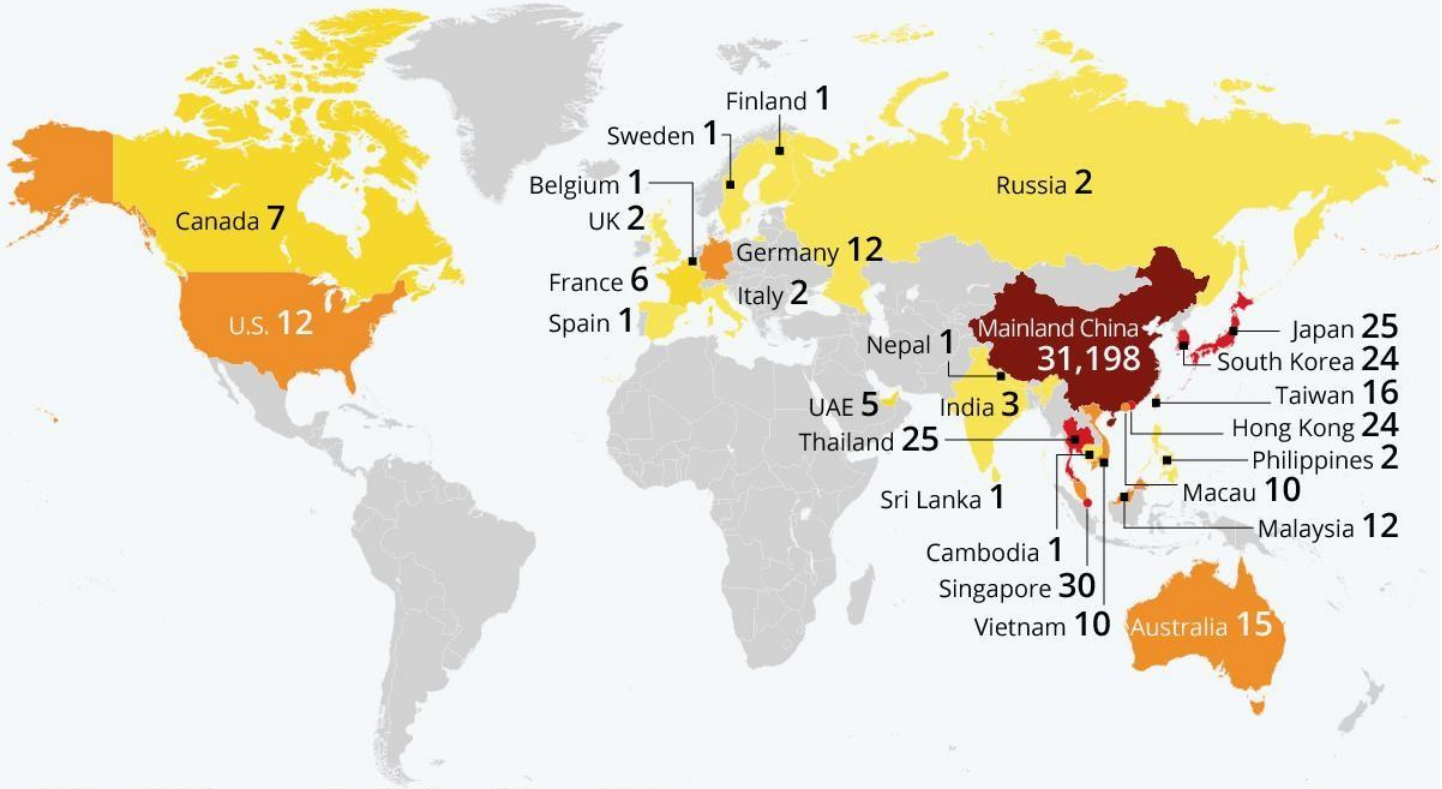
 Source: Nielsen

statista


EU: How The Public Rates Courts For Independence
Perceived independence of national justice systems in the EU in May 2018 (%)
Very/fairly good
Very/fairly bad
Don't know
Denmark
Finland
Austria
Netherlands
Sweden
Ireland
Germany
Luxembourg
United Kingdom
Belgium
Estonia
Greece
France
Cyprus
Czech Republic
Lithuania
Portugal
Hungary
Romania
Latvia
Malta
Poland
Spain
Slovenia
Italy
Bulgaria
Slovakia
Croatia
0
20%
40%
60%
80%
100%
n=26,537
@StatistaCharts
Source: European Commission
statista

Where The Coronavirus Has Been Confirmed

Locations by number of confirmed Wuhan coronavirus cases*



* As of February 7, 2020 at 10am CET
Source: Johns Hopkins University

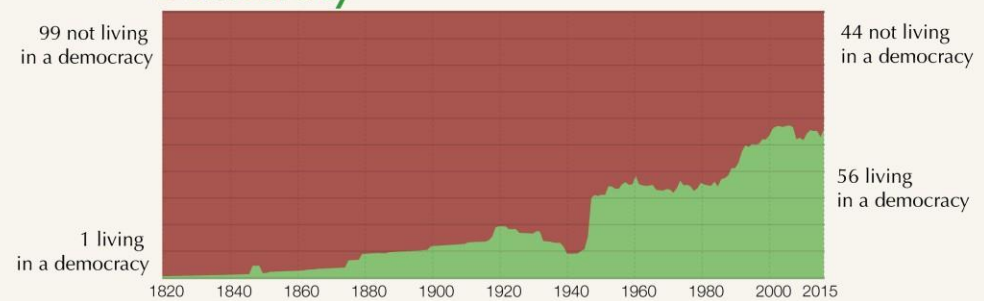
statista

The World as 100 People
over the last two centuries

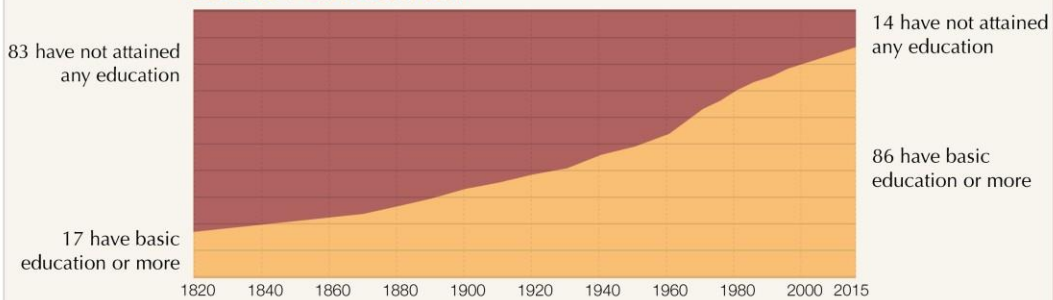
Our World in Data

Extreme Poverty

Democracy

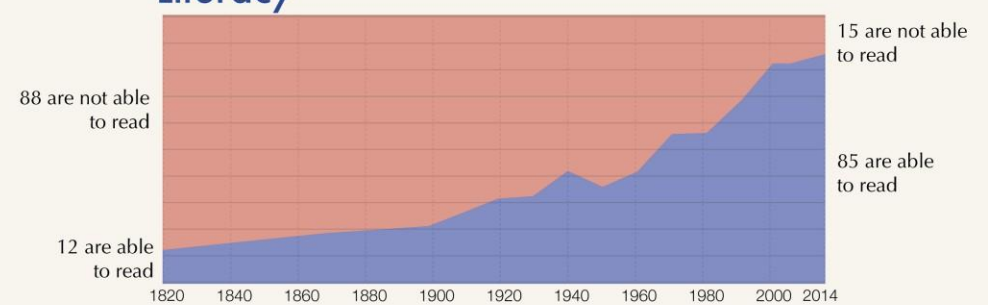


Basic Education



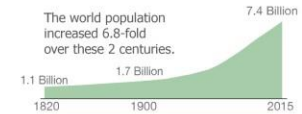
Vaccination against diphtheria, pertussis (whooping cough), and tetanus

Literacy



Child Mortality

Data sources:
Extreme Poverty: Bourguignon & Morrison (2002) up to 1970 – World Bank 1981 and later (2015 is a projection).
Vaccination: WHO (Global data are available for 1980 to 2015 – the DPT3 vaccination was licenced in 1949)
Education: OECD for the period 1820 to 1960. IIASA for the time thereafter.
Literacy: OECD for the period 1820 to 1990. UNESCO for 2004 and later.
Democracy: Politiy IV index (own calcluation of global population share)
Colonialism: Wimmer and Min (own calcluation of global population share)
Continent: HYDE database
Child mortality: up to 1960 own caluclations based on Gapminder; World Bank thereafter



All these visualizations are from OurWorldInData.org an online publication that presents the empirical evidence on how the world is changing.


THANK YOU