Empirical Software Engineering Research

Data & Measures

Norman Peitek, Annabelle Bergum, Lina Lampel, Sven Apel

Learning Goals

- Understand different types and structures of data and various data sources
- Familiarize yourself with different operations and measures on data depending on their structure
- Obtain knowledge on how to visualize different types of data

Reviewing Homework

Yesterday's homework was to think about data/information that you collect from other people/services.

What is collected from you?

What kind of data is it and how is the data structured?

Research Data

“You cannot control what you cannot measure” (P. Gilb)

Independent of the type of research, they all require collecting research data that will be analyzed.

Research Data Sources

Research data can be acquired from different sources and depending on the research question, different sources need to be used.

- Observation: passive human observation, sensors, … to measure information
- Experimental: active intervention from the researcher to observe variables

- Simulated data: generated data by imitating real-world processes
- Derived data: use and combine existing data sources

Realistically, the source is driven by what *can* be accessed

Using the right source is critically important for the success of research

Primary vs. Secondary Data Collection

	Primary Data	Secondary Data
Example	Student collects (new) data by themselves, for example, through conducting an experiment	Student uses existing data from a prior study
Easiness	Difficult	Easier
Collection time	Time-consuming	Less time-consuming
Flexibility	High	Low

→ There are advantages and disadvantages for each type of data collection

Human versus Data Collection

	Human Focus	Data Focus
Example	Student interviews 10 programmers regarding their preferred programming language	Student analyzes commit history of an open-source project for time of day
Resources	Higher	Lower (typically, can be very high)
Reliability	Lower	Higher
Collection time	Time-consuming	Less time-consuming (typically)
Flexibility	Higher	Lower
Focus on…	Cognition	Behavior

→ There are advantages and disadvantages for each type of data collection

Different Types of Research Data

There are several ways to classify research data.

- Researchers must understand the kind of data they are working with because analysis techniques (including statistics) are typically only for specific data types.

- One main distinction is quantitative (= *numbers*) versus qualitative (= *words*) data.

Different Types of Research Data

Qualitative data is collected from interviews, surveys, code comments, reviews, …

- To some degree, qualitative data can be transformed into quantitative data (e.g., encoding comments into a sentiment score)
- Can be meaningful, but less precise as it requires subjective interpretation

Quantitative data typically is measured

- Precise, but possibly less meaningful

Many Measures for Constructs

A measure is mapping from a studied object to a scale. There can be multiple measures for the same construct.

Sometimes, it is impossible to find a reasonable measure of a construct

For example, code complexity:

- Complexity Metric LOC (quantitative data)
- Subjective rating in 1 out of 5 (quantitative data)
- Written review “complex because of confusing variable names and …” of a programmer (qualitative data)

Types of Qualitative Data

Qualitative (categorical) data represent data groups.

- Nominal
  - There is no (sensible) order between values
  - “yes”, “no”, “don’t know”
  - Gender, Colors, study degrees, …
- Ordinal
  - Order is important, but the specific values and the differences between them are not
  - “very bad”, “bad”, “neither”, “good”, “very good”
  - Ranking (race result)

Types of Quantitative Data

Quantitative data can be linear (100 is twice as much as 50, impossible for ranking)

- Discrete quantitative data (= integer)
  - Number of students in this course (15, 16, 17)
  - Money (1.59 Euro)
- Continuous quantitative data (= float)
  - Response time (Stop timer)
  - Distance

Where would age of a participant belong to?

Measurement Scales

	Categorized	Ranked	Evenly Spaced	Natural Zero	Examples
Nominal	✔	✘	✘	✘	Nationality, gender, car brands, …
Ordinal	✔	✔	✘	✘	Race results, programming ability (novice, intermediate, expert), Likert scale
Interval	✔	✔	✔	✘	Exam grades, Temperature in Celsius
Ratio	✔	✔	✔	✔	Height, Weight, Age, Temperature in Kelvin

Increase in richness

Selecting Measurement Scale

For some variables, the researcher can choose at what measurement scale data is collected. For example, income of a participant could be collected

- on a ratio level (exact income), or
- in brackets on an ordinal level (0€ - 9 999€, 10 000 – 19 999€, 20 000 € - 29999 Euro, …)

The ratio level provides more precise data than ordinal level.

- Why would researchers choose a less precise measurement scale?
- What is the danger of using ordinal level?

Examples of Data

Objective Data (of Artifacts)

- We can gather a vast amount of data from automated metrics
  - For example, the commit history of the Linux kernel (or any other open-source project)
- Many metrics are reliable and computationally cheap (e.g., LOC)
  - However, if applied to a huge data set, it can quickly compound
- Human elements can be more fuzzy
  - For example, code review texts: Are they friendly or overly critical? How would you measure it?
  - Often these measures are more difficult and computationally more expensive

Subjective Data

Subjective data is comparatively easy to collect (e.g., ask questions)

- Subjective data can be biased in many ways (e.g., self-censorship, …) from the researcher and participant
    - Depending on the research, the individuality of participants can be tricky
- But many questions can only be solved with a subjective view
    - For example, “Do you like pair programming?” is hard to answer with objective data
- Subjective data can complement objective data

Behavioral Data

Behavioral data can be mined (e.g., commit history, email lists, …) or actively collected (e.g., controlled experiment) and tends to provide more objective views on human behavior.

- Behavioral data can be useful to observe real-world scenarios in more detail
  - For example, detailed interaction history with an IDE or "How often do developers look at facebook?"
  - But finding participants for such experiments can be tricky (and there are ethical/privacy concerns)
- Behavioral data can show patterns, but often not explain them
  - *Why* did a programmer behave this way?

Physiological Data

There is a variety of physiological measures that provide insights into human emotional and cognitive states

- Heart rate, electrodermal activity: stress
- Eye movements: cognitive process, attention
- Sensors are becoming increasingly cheap
- Physiological measures often correlate with subjective data
    - For example, electrodermal activity records high stress levels that participants also reported on

Neuroimaging Data

Neuroimaging provides insights into brain activity.

- Most detailed data, but expensive to collect
- Will discuss in more detail in today's guest lecture

Descriptive Statistics for Quantitative Data

Practice Data Set

For the rest of the lecture, we will use a data set describing passengers of the Titanic.

Class	Sex	Age	Survived	Alive	Alone
3	Male	22	0	No	False
1	Female	38	1	Yes	False
3	Female	26	1	Yes	True
1	Female	35	1	Yes	False
3	Male	35	0	No	True
2	Male		0	No	True
1	Male	54	0	No	True
2	Female	2	0	No	False
…	…	…	…	…	…

The data set including the scripts will be uploaded under materials

For each column:
What is the measurement scale?
(nominal, ordinal, interval, ratio)

Do survived and alive describe the same data?

Survived, alive, and alone are all binary data with different encoding → unify

Descriptive Statistics

The human mind has limited capacity → Descriptive statistics summarize a (large) data set and are fundamentally important in research.

- For example, studies typically do not report details about every single participant, but provide a summary (e.g., average age)

Descriptive Statistics

(Frequency) Distribution

Measures of Central Tendency

Measures of Variability

Distribution

(Frequency) distribution summarizes the frequency of all values in a data set in absolute or relative terms.

Participant	Age	Gender
001	34	Woman
002	43	Woman
003	54	Man
004	75	Diverse
005	12	Woman
006	1	Man
James	23	Man
008	42	Woman
…	…	…

We recruited 42 participants (20 women, 20 men, 2 diverse)

Measures of Central Tendency

Measures of central tendency provide a single value representing the center of a (large) data set

There are three common options:

- Mean
- Median
- Mode

Measures of Central Tendency: Mean

(Arithmetic) Mean = (sum of all values) / (total number of values)

- Colloquially, the “average” typically refers the arithmetic mean
- Mean is ideal when the data is balanced, symmetrical, and has no outliers
- → Mean is problematic when the data is imbalanced, skewed, or has outliers

- Alternative version: geometric mean = *n*th root for the product of n numbers

Measures of Central Tendency: Median

Median = middle value of data set

What to do if n is even? For example: [1,2,3,4]?

Usage: Median is better than mean when

- There are few measurement values
- The data is skewed (non-normal distribution)
- The data contains (extreme) outliers

Measures of Central Tendency: Mode

Mode = value that occurs most often in data set

- Unlike mean and median, mode can be used for categorical variables
- For example
  - “What is your favorite color: Green, blue, or red?”

Measures of Central Tendency: Titanic Passengers

Let’s practice measures of central tendency with the Titanic data set.

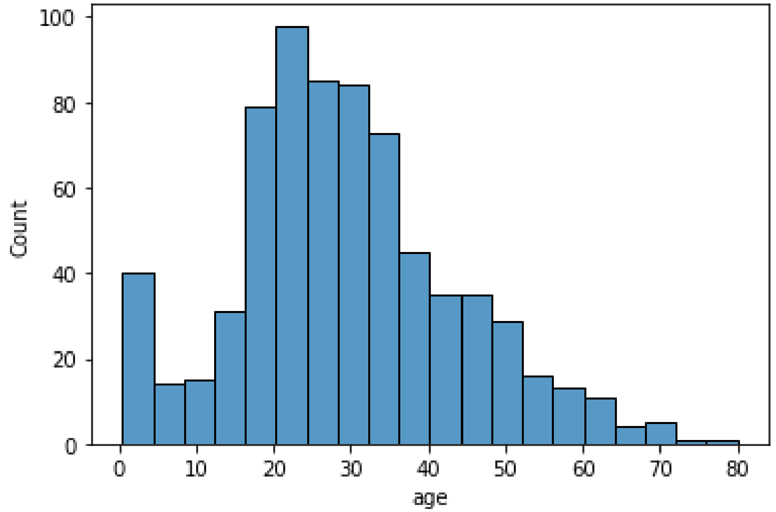
- For demos and example scripts, we use the Python library pandas for all the upcoming analysis and matplotlib/seaborn for the visualization.
- Many alternatives exist (R, Julia, SPSS, …)

Live Demo

Skewness in Data

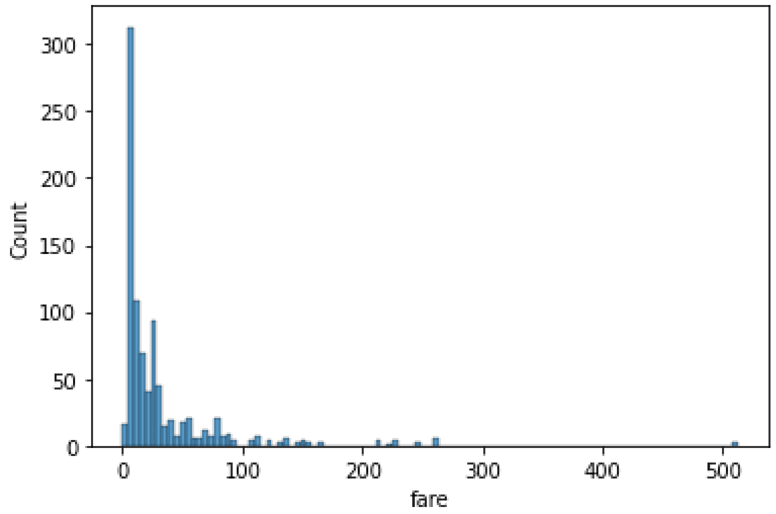
Age:

- 29.6 mean, 28 median → fairly close



Fare:

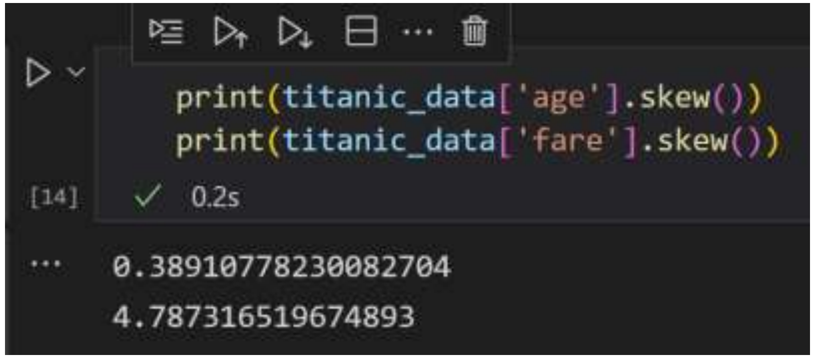
- 32 mean, 12 median → Positive skew



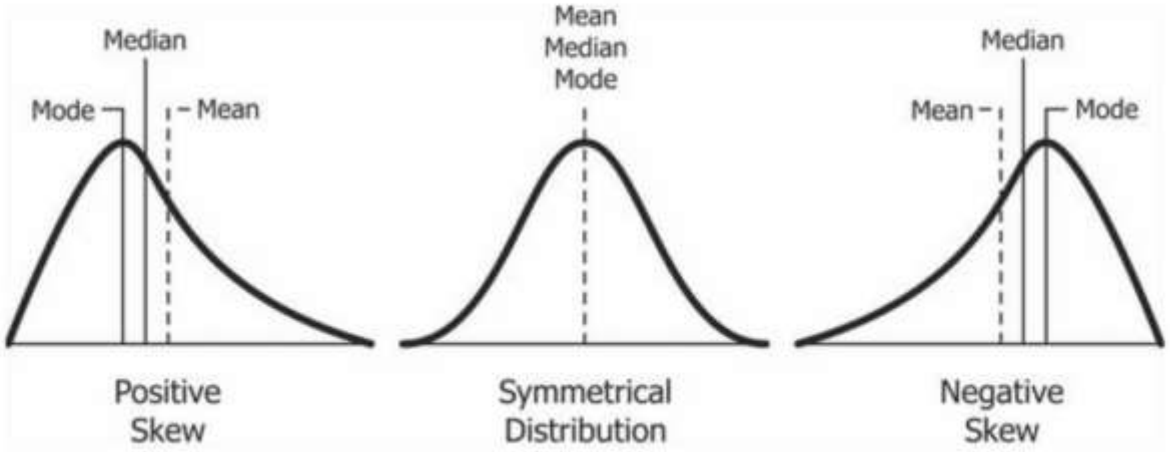
Skewness in Data

Skewness can bias the analysis

→ Impacts some statistical tools



```
print(titanic_data['age'].skew())
print(titanic_data['fare'].skew())
```

https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa

Shapiro-Wilk Test

- Many statistical tests assume normally distributed data
- Shapiro-Wilk is one test to check for a normal distribution
  - H0: data is normally distributed
  - → if $p < \alpha$, then the null hypothesis is rejected, and the data is *not* normally distributed
  - This is one test where researchers like to see large p-values

- `stats.shapiro(age_males["age"]))`
- → `ShapiroResult(statistic=0.989, pvalue=0.008)`
  - → non-normal distribution

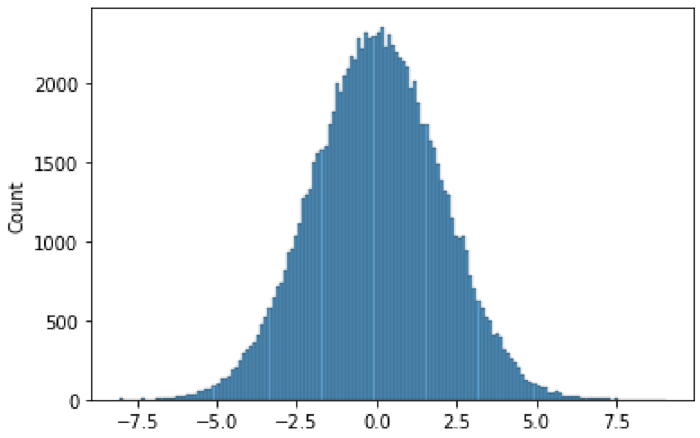
Measures of Variability

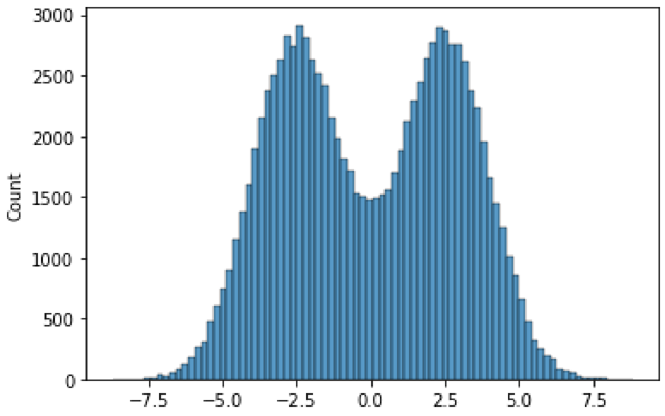
Measures of central tendency summarize the “center” of the data, but not how it spreads out. Measurements are often noisy, and we collect a distribution of similar measurements (rather than all the exact same).

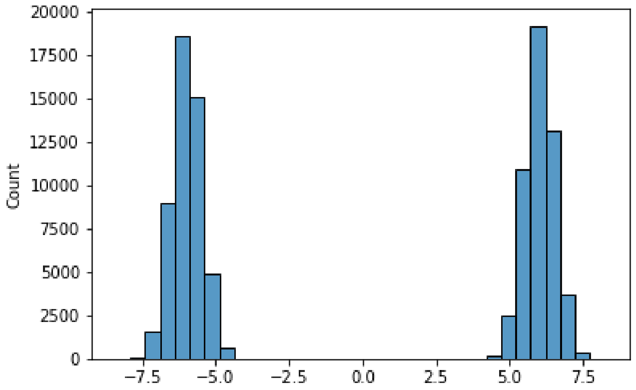
Typically, a measure of central tendency is therefore combined with a measure of variability (dispersion) to also show the noise/error in the data.

Measures of Variability

- Let's look at three randomly generated samples:



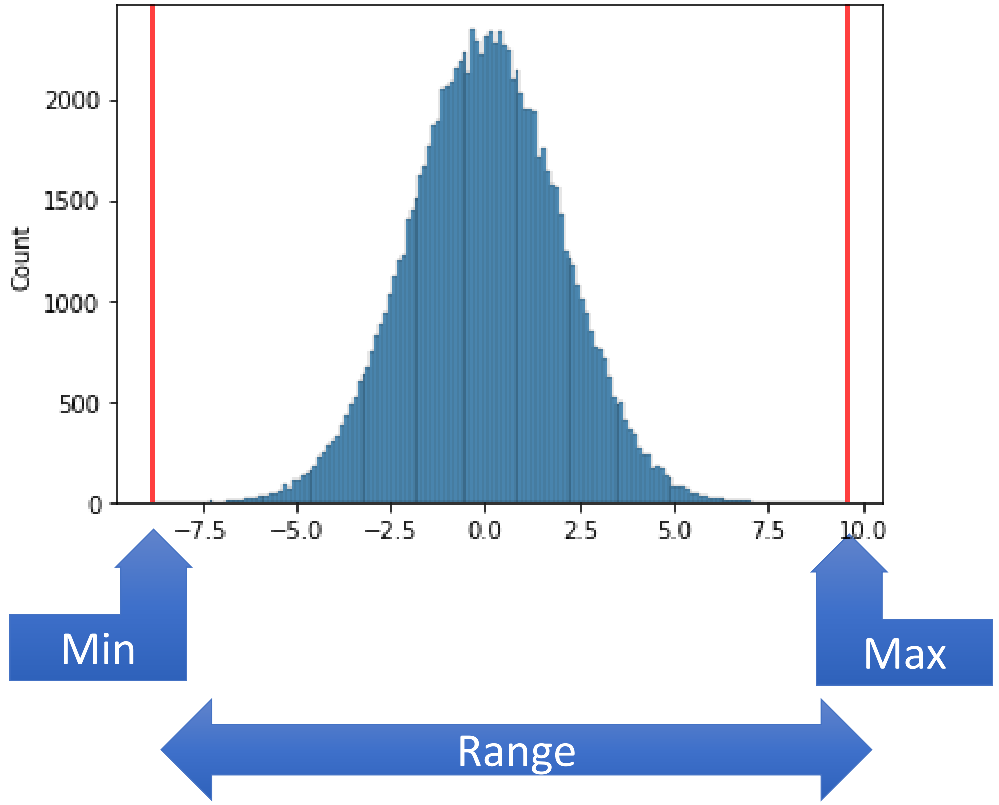




The mean is in all three cases "0".
But clearly, the *variability* is much different,

Measures of Variability: Min, Max, Range

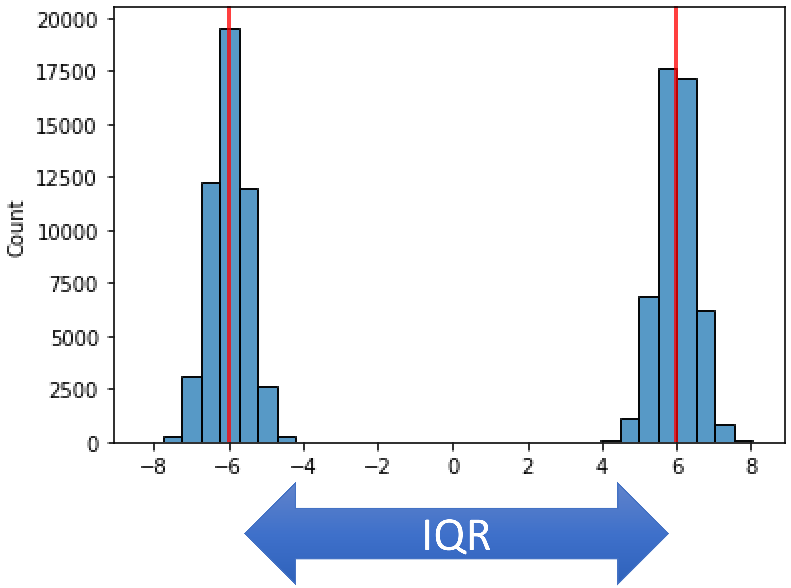
The most basic option to describe variability is describing the range (minimum to maximum value).

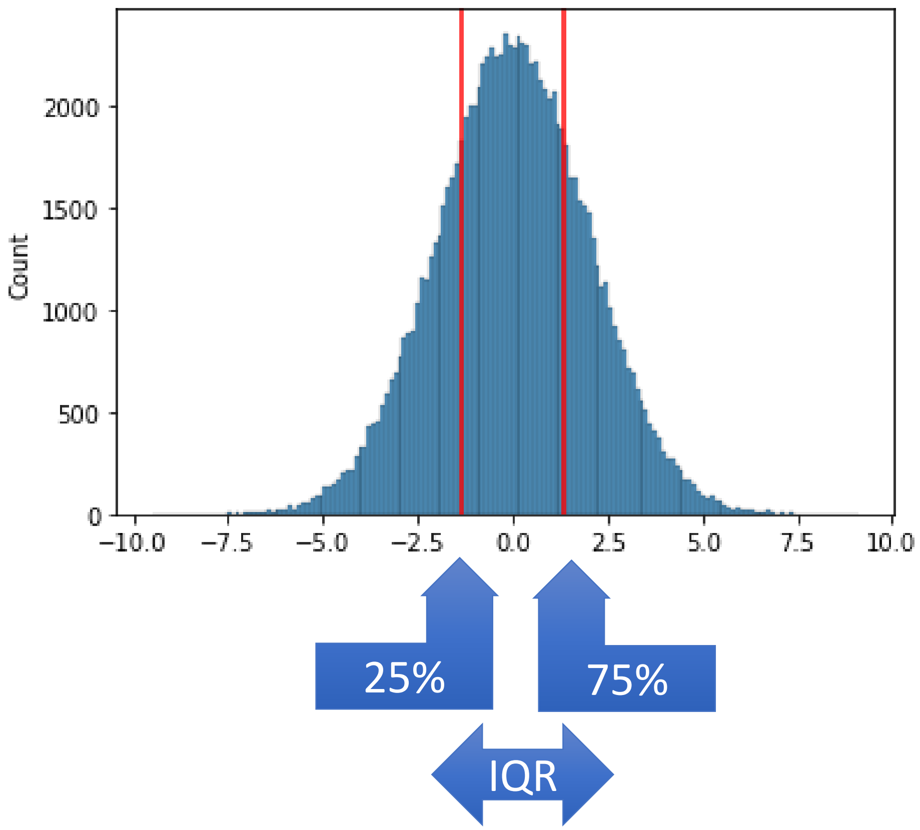


Measures of Variability: Interquartile Range

IQR: Apply the concept of the median again to each half of the data, distance between each half's median

IQR: 75 percentile – 25 percentile





Measures of Variability: Variance and SD

Variance is the average of squared deviations from the mean.

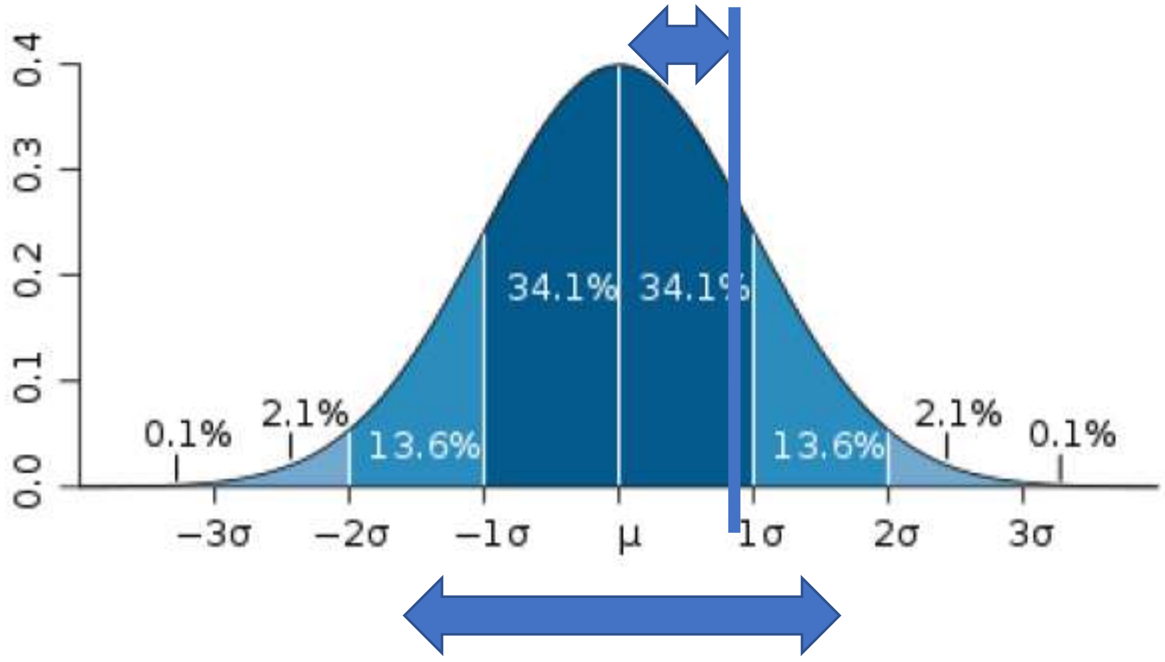
Standard deviation (SD) is the square root of variance.

- Advantage of SD over variance: same unit as original data (not squared)

Accuracy & Precision

Accuracy:
Deviation of observed mean from true mean

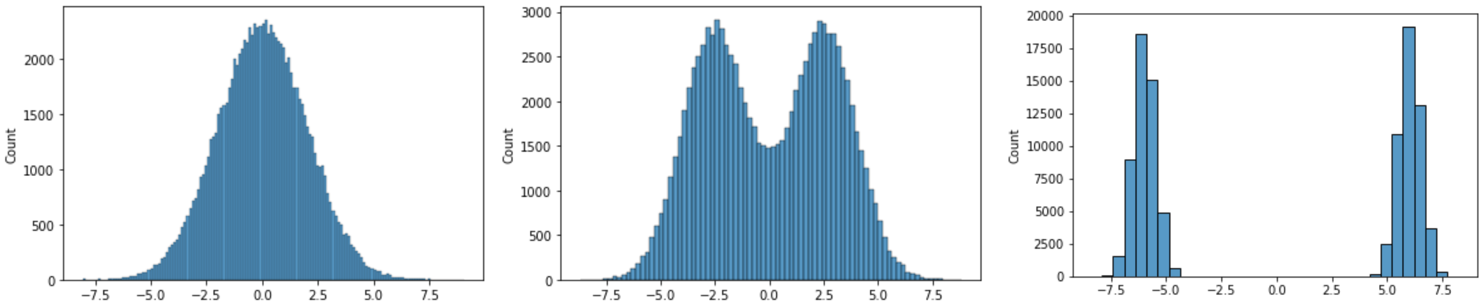
Important when measuring response time



Precision:
Dispersion around mean

Cause of measurement errors is unclear

Measures of Variability: Example



	Left Plot	Center Plot	Right Plot
Mean	0.0	0.0	0.0
Median	0.0	0.0	0.0
Min/Max	-8 / 8	-8 / 8	-8 / 8
IQR	2.7	5.0	12
Variance	4.0	9.0	36
Standard deviation	2.0	3.0	6

Central tendency and even min/max ranges are identical

Measures of variability can summarize the different data distribution

Measures of Variability: Titanic Passengers

Live Demo

	Age	Fare
Mean	29.6	32.2
Median	28	14.4
Min/Max	0.42 / 80	0 / 512
IQR	17.9	23
Variance	211	2469
Standard deviation	14	49

IQR responds less strongly to outliers than variance/standard deviation

Descriptive Statistics & Measurement Scales

	Mathematical Operations				Measures of Central Tendency			Measures of Variability			
	Equality	Comparison	Addition & Subtraction	Multiplication & Division	Mode	Median	Mean	Range	Interquartile Range	Standard Deviation	Variance
Nominal	✔	✘	✘	✘	✔	✘	✘	✘	✘	✘	✘
Ordinal	✔	✔	✘	✘	✔	✔	✘	✔	✔	✘	✘
Interval	✔	✔	✔	✘	✔	✔	✔	✔	✔	✔	✔
Ratio	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔

Be Cautious with Automation

```
titanic_data.describe(include='all')
```

[12] ✓ 0.6s

	survived	pclass	sex	age	fare	embarked	class	embark_town	alive	alone
count	891.000000	891.000000	891	714.000000	891.000000	889	891	889	891	891
unique	NaN	NaN	2	NaN	NaN	3	3	3	2	2
top	NaN	NaN	male	NaN	NaN	S	Third	Southampton	no	True
freq	NaN	NaN	577	NaN	NaN	644	491	644	549	537
mean	0.383838	2.308642	NaN	29.699118	32.204208	NaN	NaN	NaN	NaN	NaN
std	0.486592	0.836071	NaN	14.526497	49.693429	NaN	NaN	NaN	NaN	NaN
min	0.000000	1.000000	NaN	0.420000	0.000000	NaN	NaN	NaN	NaN	NaN
25%	0.000000	2.000000	NaN	20.125000	7.910400	NaN	NaN	NaN	NaN	NaN
50%	0.000000	3.000000	NaN	28.000000	14.454200	NaN	NaN	NaN	NaN	NaN
75%	1.000000	3.000000	NaN	38.000000	31.000000	NaN	NaN	NaN	NaN	NaN
max	1.000000	3.000000	NaN	80.000000	512.329200	NaN	NaN	NaN	NaN	NaN

Before Anything Else: Understand Your Data!

1. Get an overview and familiarize yourself with the data
2. Apply the right measures for the right types of data
3. Use multiple measures, at least for informative purposes
   - For example, don't just blindly use the mean if you have outliers
4. Remember to deal with missing values and possible outliers
   - We will discuss this in more detail in tomorrow's data analysis lecture

One Perspective Is Insufficient

When you understand and/or report on data, use **multiple** perspectives

- Most common: n, mean + SD

- Typically, a table with all information is too much
  - Sometimes there is a full report in appendix/replication package

Showing all raw data easily overwhelms humans, but there is a close alternative

→ Use visualization!

~~Homework~~ Breakwork

Imagine we have four data sets with x/y coordinates. The discussed measures for all four data sets are the same:

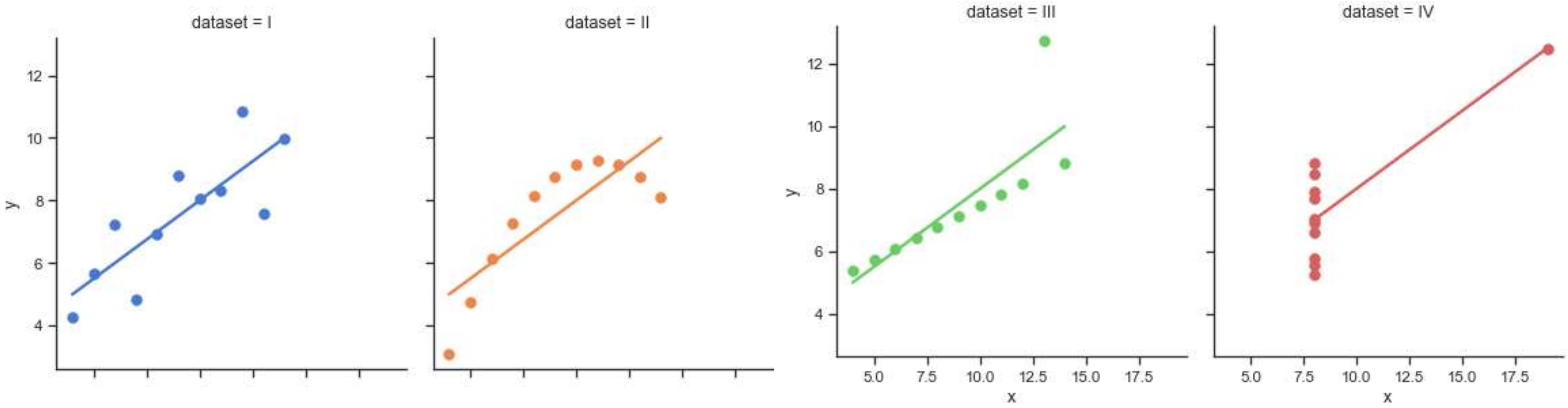
- Mean + SD of x: 9.00 + 11.00
- Mean + SD of y: 7.50 + 4.13
- Correlation identical

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
…	…	…	…	…	…	…	…

→ Please discuss over the break whether the four data sets are identical! For example, if you conduct a study with four groups: are they sufficiently similar?

Limitations of Descriptive Statistics

- So far, we have used descriptive statistics to create an overview of a data set
  - But those can be misleading



- Use visualization as complementary tool to descriptive statistics!

Basic Data Visualization

Data Visualization

In the last lecture, we discussed several important measures to understand and summarize data:

- Distribution
- Measures of central tendency
- Measures of variability

Let‘s turn dry numbers into understandable visualizations!

→ We will be working with the titanic data set again

Data Visualization

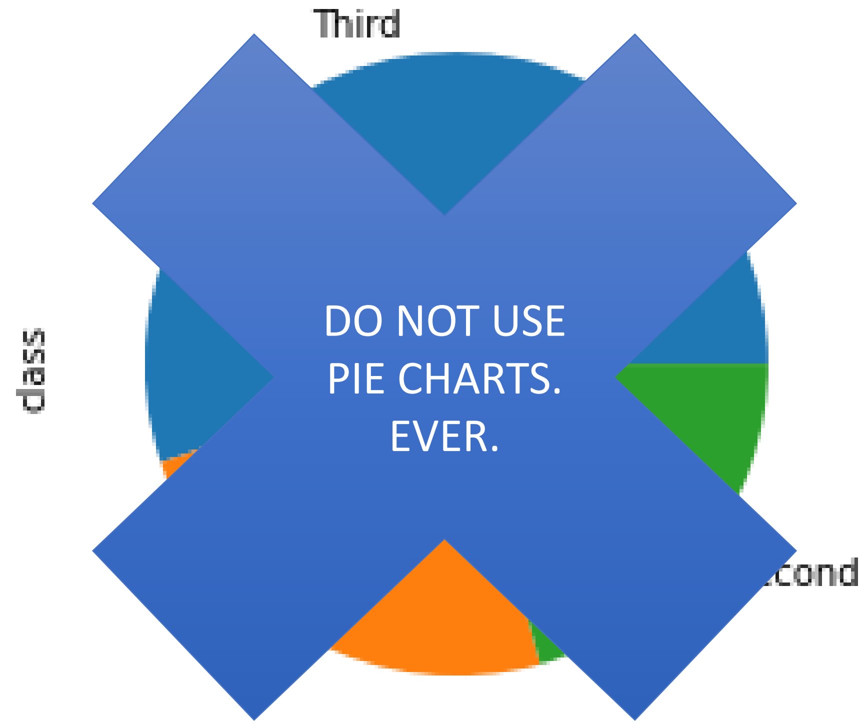
Before we start, let's reflect on what kind of data visualizations you have used. What kind of plots have you created?

Which tools did you use?

Should you use Microsoft Excel?

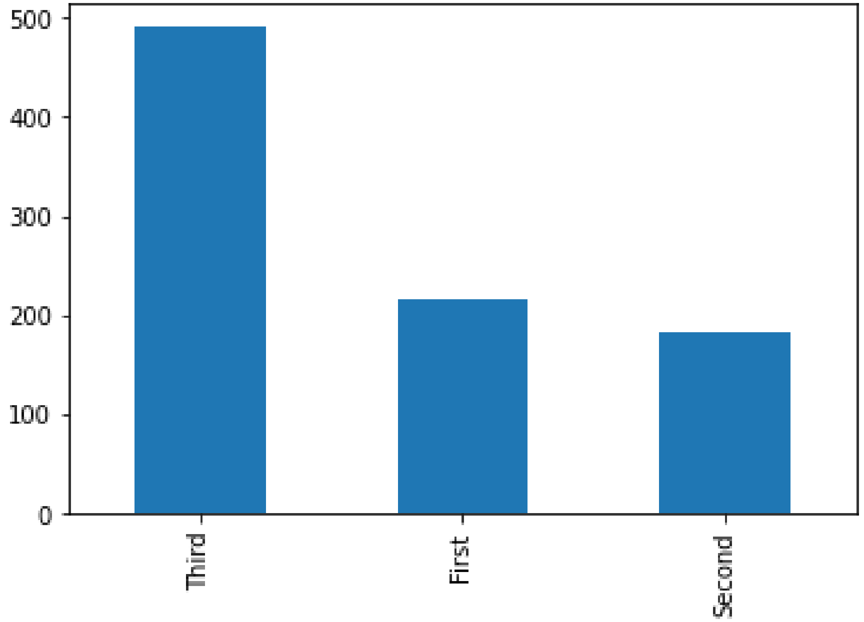
Frequency Distribution: Pie Charts

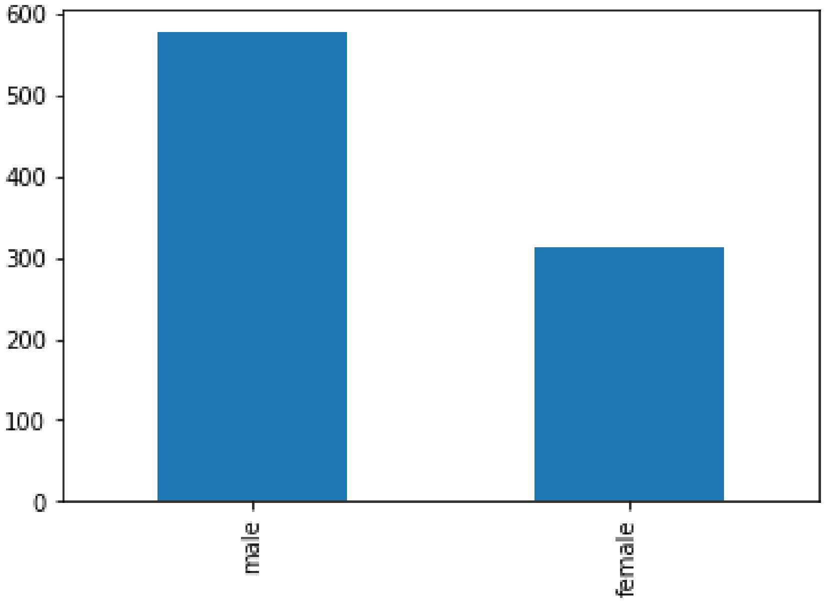
Distribution can be shown with pie charts

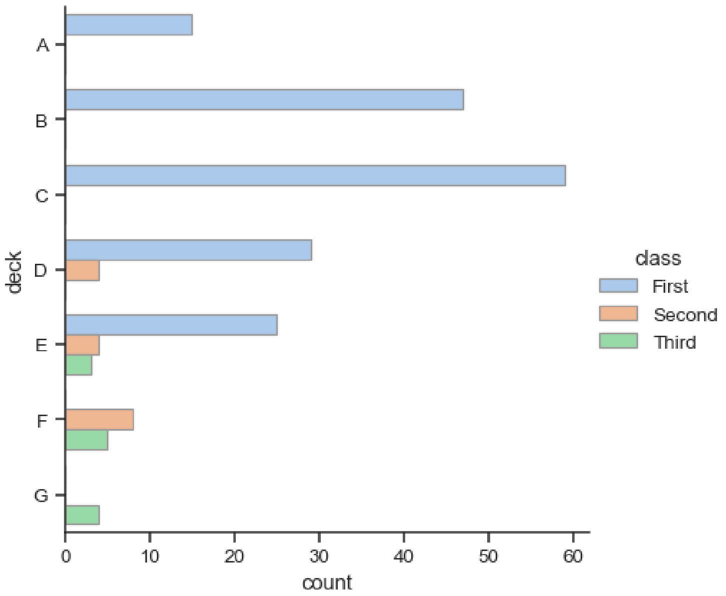


Frequency Distribution: Bar Plots

Bar plots represent the proportion of data in corresponding bar heights.

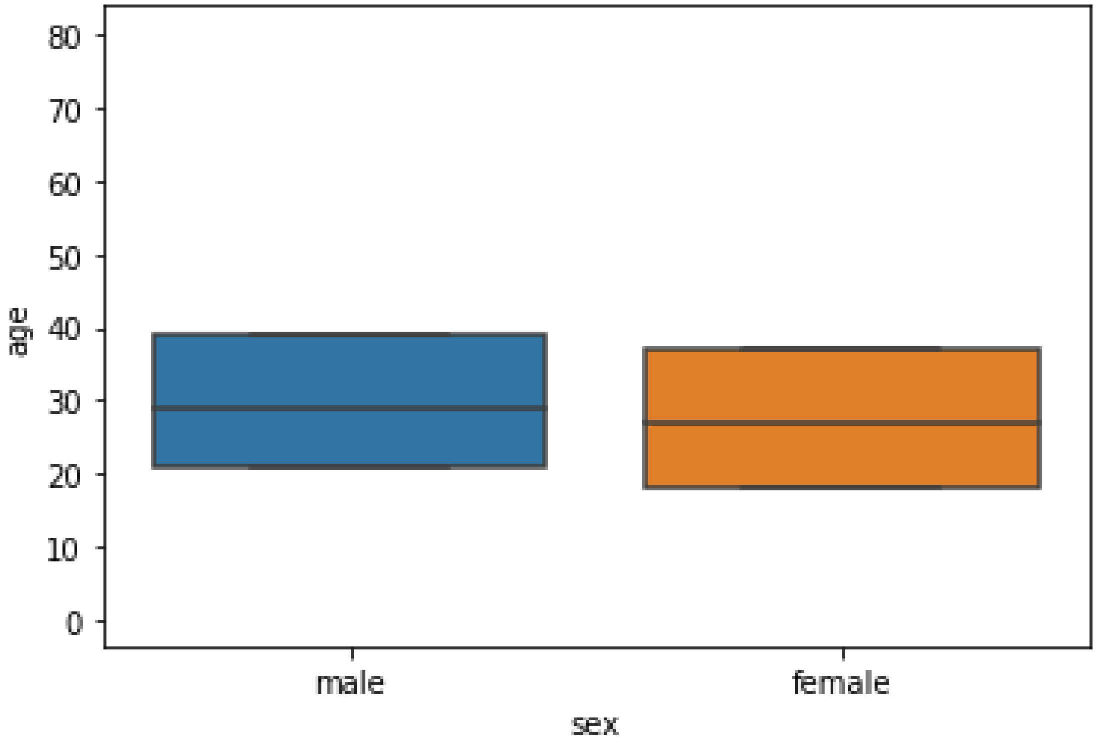






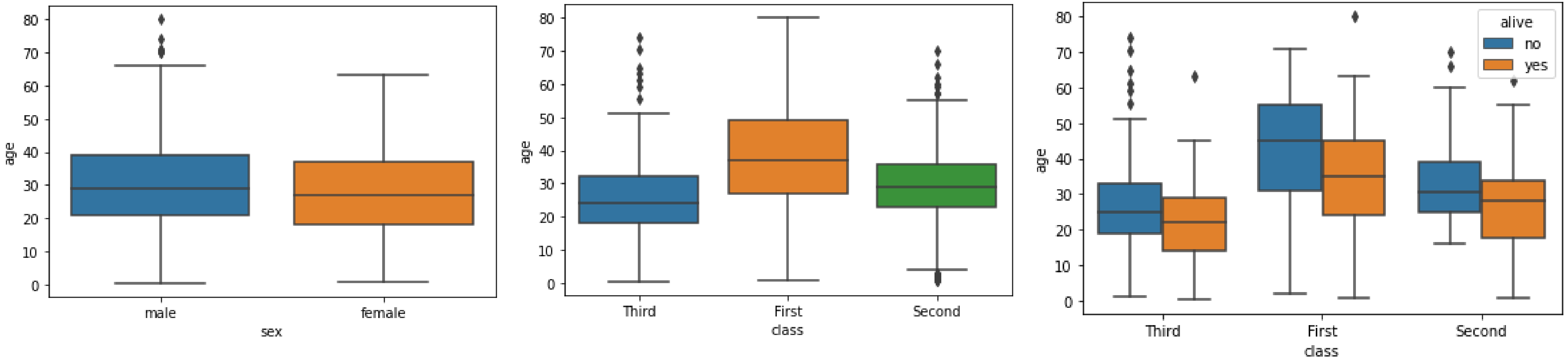
Distribution: Box Plots

Box plots visualize the median and IQR boundaries of the data.



Distribution: Box-and-Whisker Plots

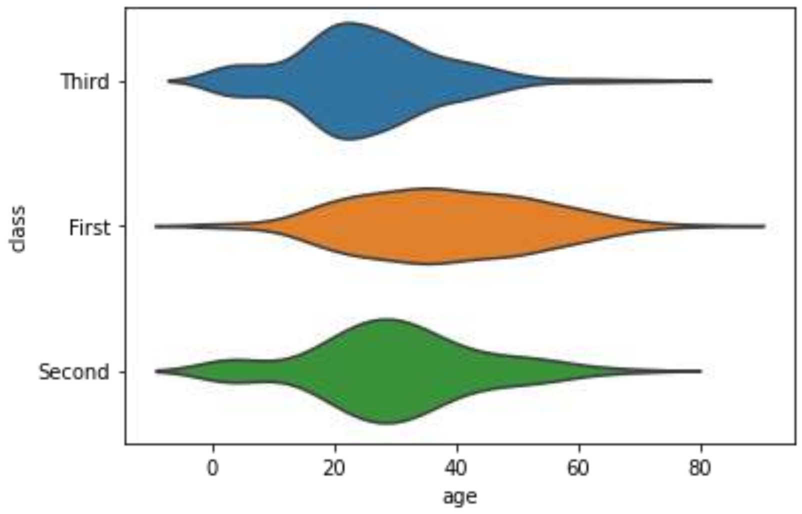
A box-and-whisker plot shows even more information. The whiskers can represent different things, but typically 1.5 IQR. Raw data outside can also be plotted

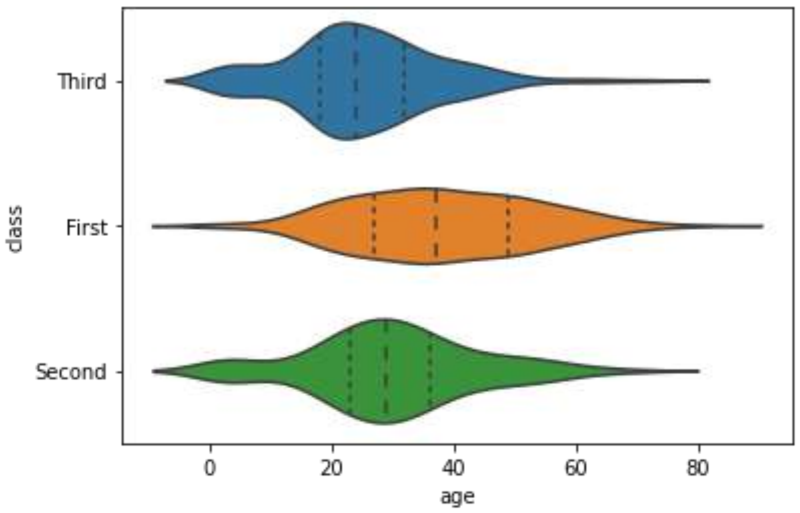


Distribution: Violin Plots

Violin plots are an extension of box plots that also show the density.

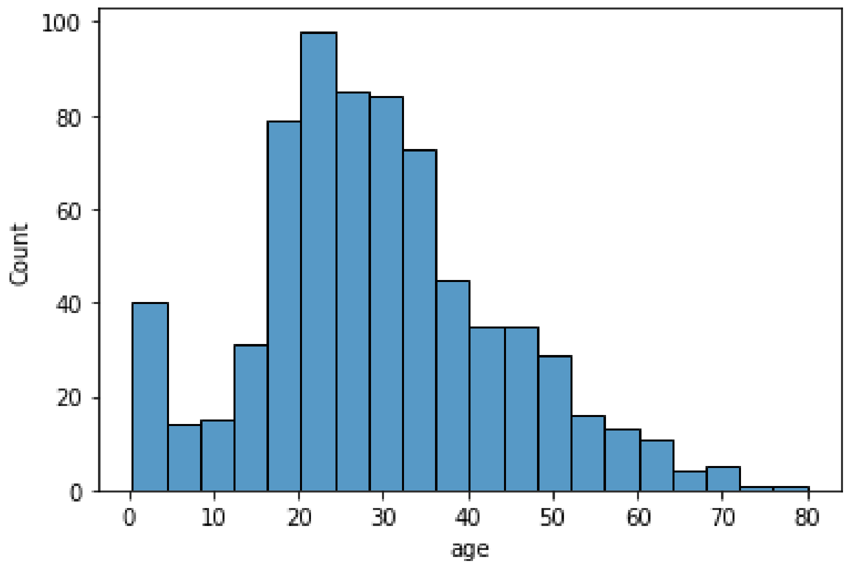
Violin plots can also be combined with descriptive statistics (e.g., median, IQR)

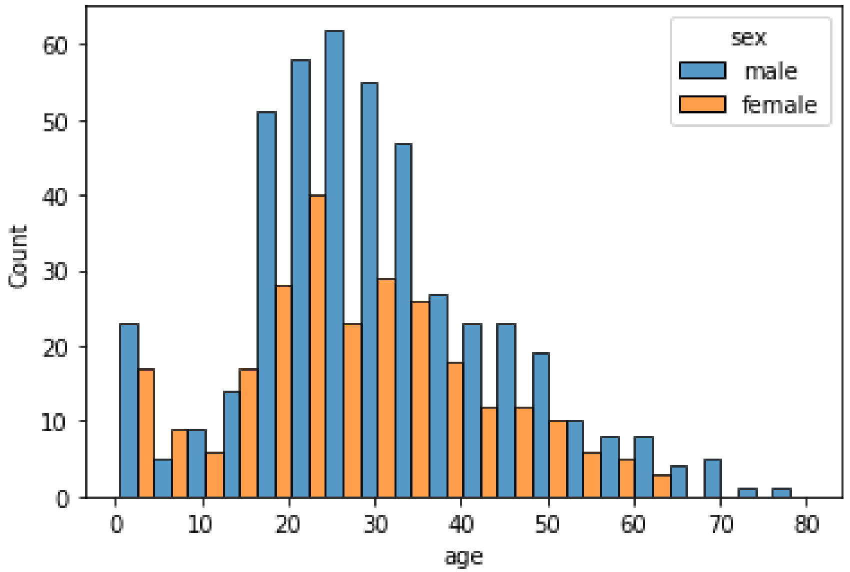




Distribution: Histograms

Histograms contain more information than box plots and are often preferred.

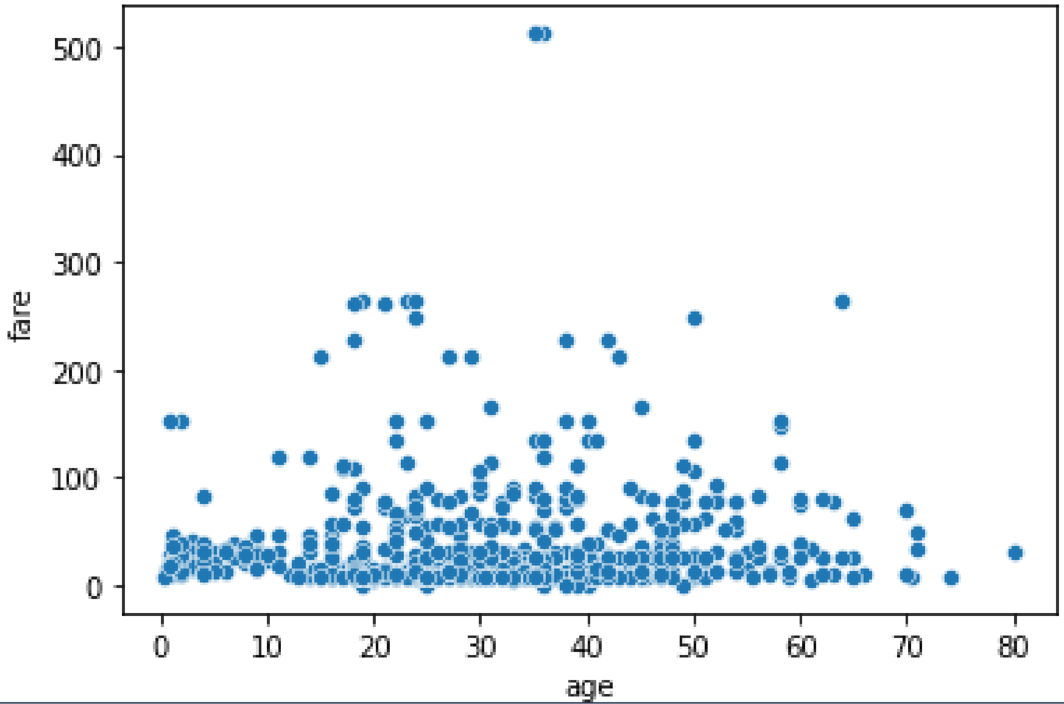


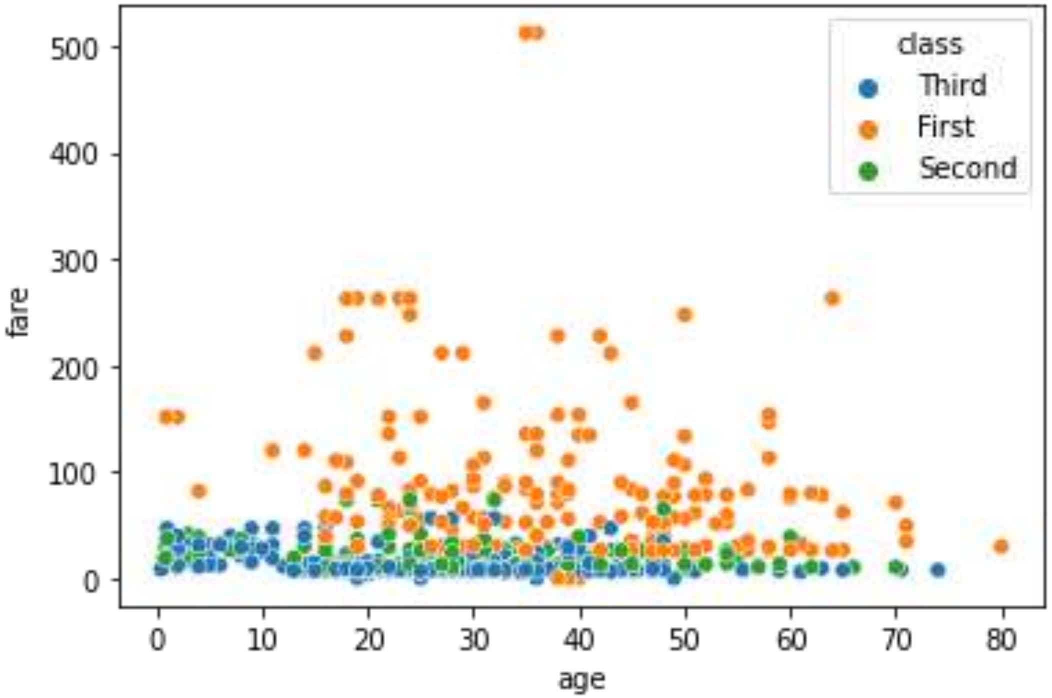


Relating 2 Variables: Scatterplots

A scatterplot shows the relationship between two variables.

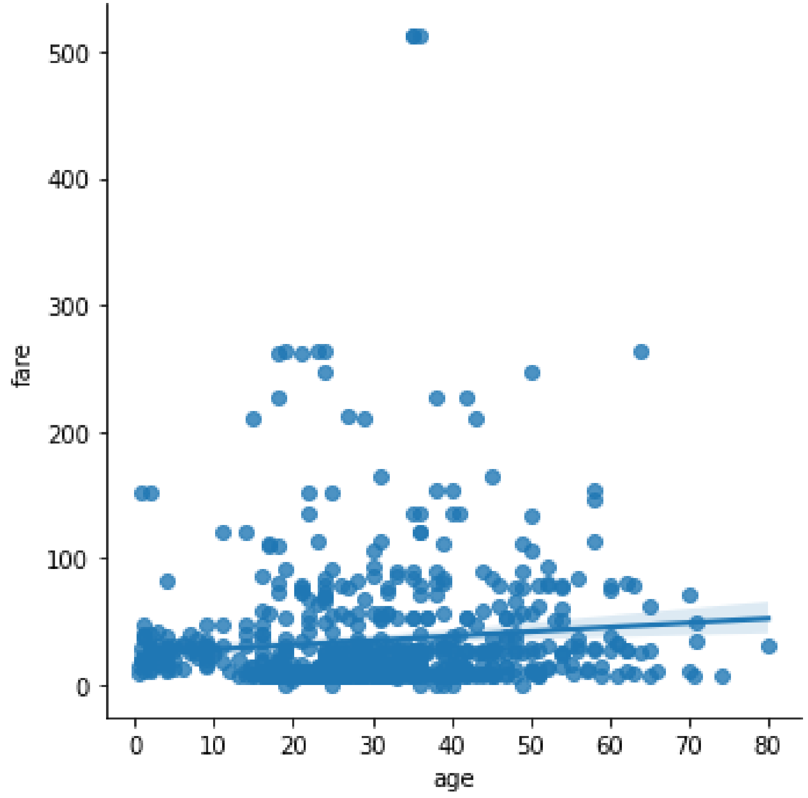
Additional semantic dimensions can easily be added.





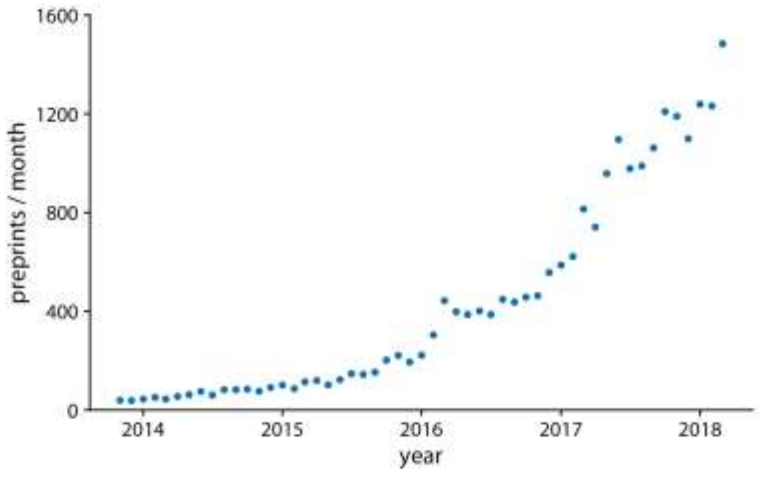
Relating 2 Variables: Regression Lines

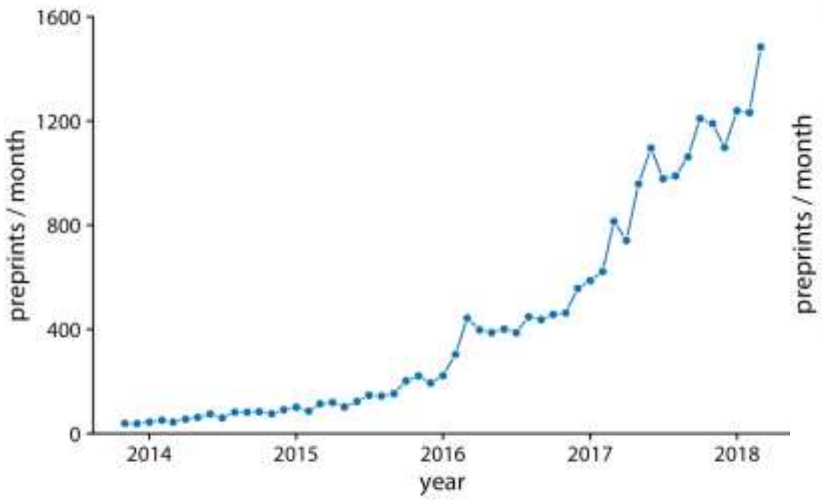
Scatterplots can be enhanced with regression lines to show a linear tendency in the relationship.

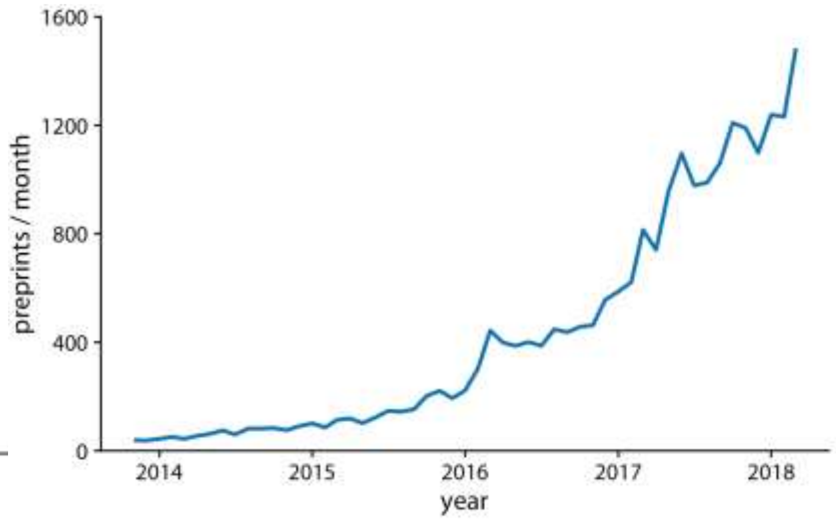


Time Series: Line Plots

Relating a variable over time





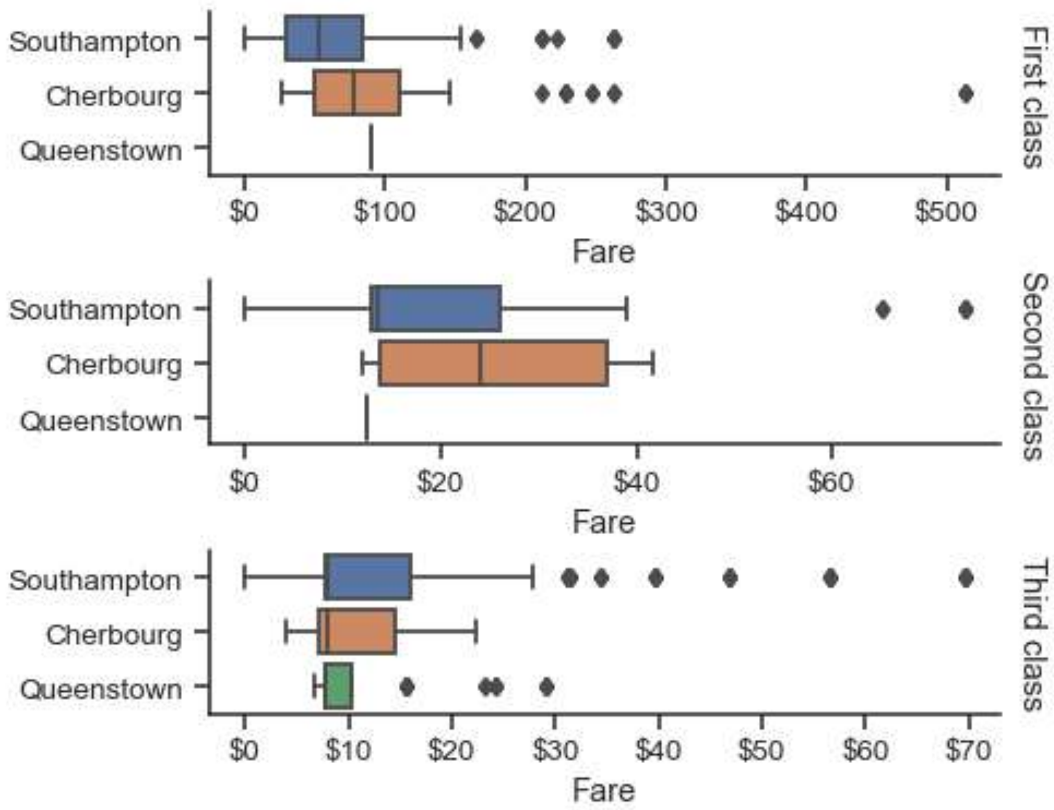


Extreme Variety of Vizualisations

So far, we looked at several options to visualize quantitative data

- There are endless options when it comes to visualization data.
- There is also lots you can do wrong

→ Visualization is something that can make a report/thesis stand out (positively or negatively)



General Tips for Visualizations

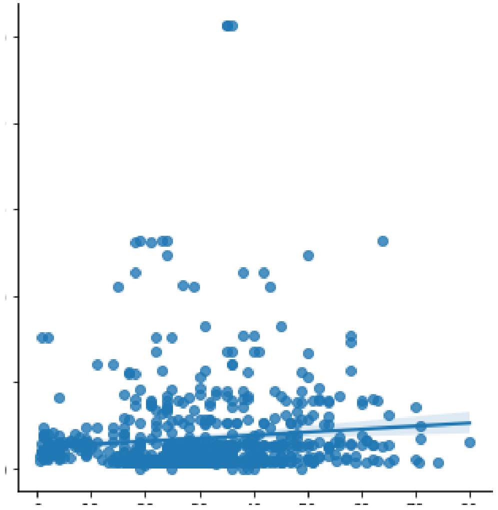
If you are unsure what visualization is suitable, review the literature (of papers that are close to your research).

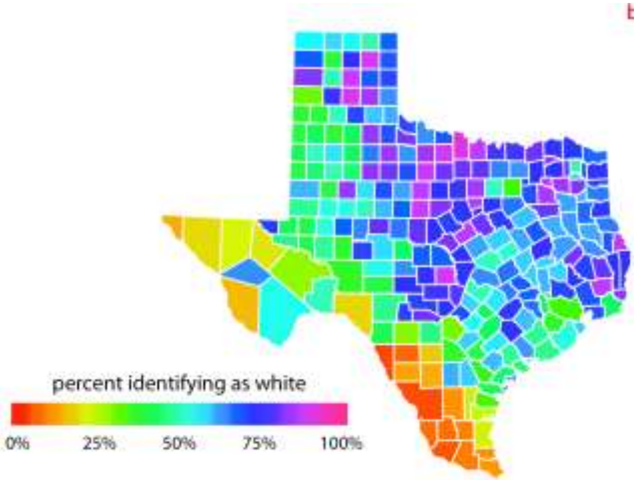
→ Focus on their selection of visualizations. Are they effective at presenting information?

→ But, many papers lack in quality visualizations. If there are no good visualizations, do not feel that you must use the same style

- (If they use pie charts, you do not need to use them too)

General Ground Rules for Visualizations





Labels must be

- On each axis
- Understandable
- Readable (font size!)

Reasonable color scheme

No misleading visualization

All of these examples are bad. Will revisit this topic on Friday

Further Reading

Check the sample code in the materials (apologies for bad code)

https://clauswilke.com/dataviz/ ($60 book on visualizations available for free online)

Python: seaborn https://seaborn.pydata.org/index.html

R: ggplot2 https://www.rdocumentation.org/packages/ggplot2/versions/3.4.0

Julia: Plots https://docs.juliaplots.org/latest/