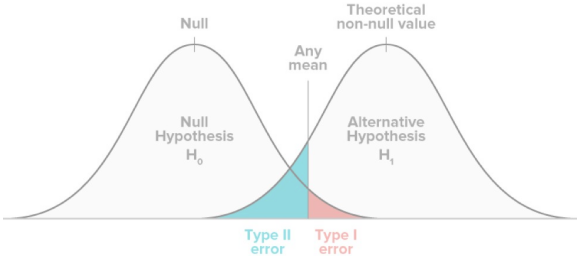
the shortcut × moz://a

DATA SCIENCE

TRAINING PROGRAM

# Hypothesis Testing

**Larissa Leite**
**Kodit.io**




kodit.io selko SILO.AI

# Outline

- What is hypothesis testing?
- Steps
  a. Make assumptions
  b. Conduct fact based tests
     - Types of tests
  c. Evaluate results
  d. Reach a conclusion

# What is hypothesis testing?

**First things first**

- Hypothesis ≋ assumption
  - If I…(do this to an *independent variable*)….then... (this will happen to the *dependent variable*)

- What types of assumptions?
  - **Distribution**
    - Normal
  - **Sampling**
    - Data sample is randomly selected, no bias
  - **Linearity**
    - Two variables have a linear relationship

# What is hypothesis testing?

- The process of verifying whether a hypothesis should be accepted or rejected by performing statistical tests
  - More specifically, whether an initial hypothesis can be rejected

- Test on the results of a survey or experiment to see if they are meaningful
  - What are the odds that the results happened by chance?

# What is hypothesis testing?

- In the industry, it is widely used to test the effect of a new feature, a new campaign
  - Basically, determining whether something has a significant impact on a given question that needs to be answered, or statement that needs to be validated:

*“We think customers order different quantities of products when offered a discount”*

- In data science, it is often used to test relationship between variables

# What is hypothesis testing?

**Steps**

1. Make assumptions
    a. Take an initial position
    b. Determine the alternate position
2. Conduct fact based tests
    a. Sampling
    b. Decide which test is appropriate
    c. Set acceptance criteria
3. Evaluate results
    a. Does the evaluation support the initial position? Are we confident that the result is not due to chance?
4. Reach one of the following conclusions:
    a. Reject the original position in favor of alternate position **or** fail to reject the initial position

# What is hypothesis testing?

**Steps**

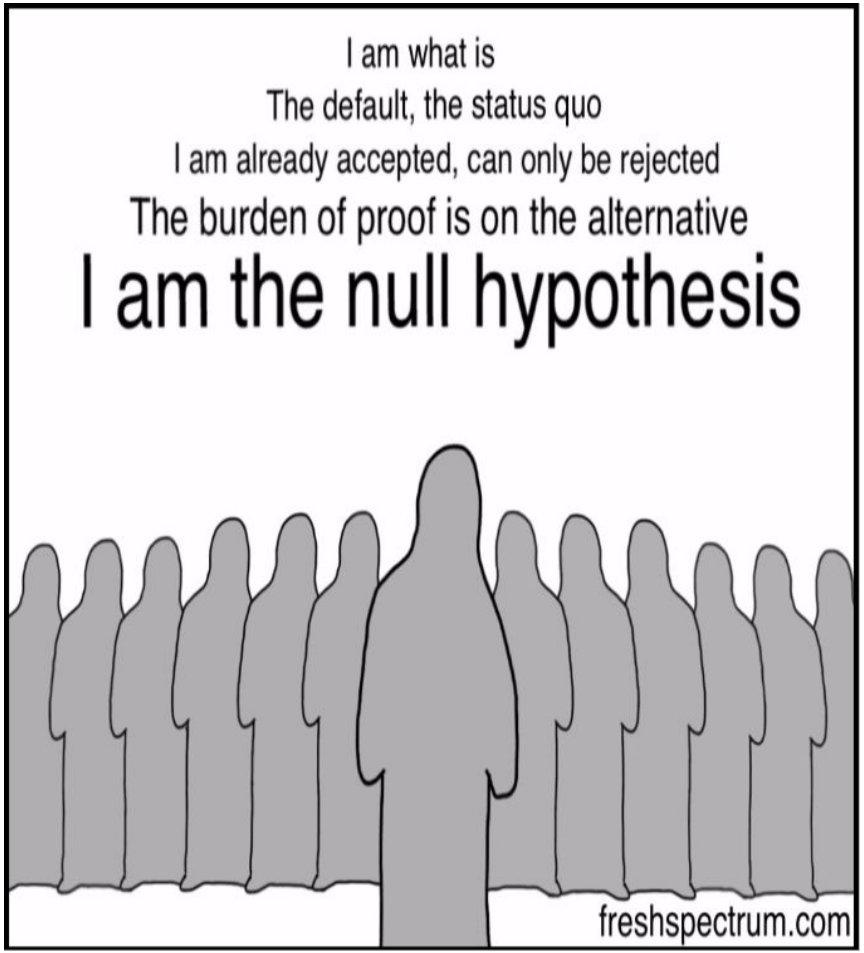
1. Make assumptions
   a. Take an initial position
   b. Determine the alternate position
2. Conduct fact based tests
   a. Sampling
   b. Decide which test is appropriate
   c. Set acceptance criteria
3. Evaluate results
   a. Does the evaluation support the initial position? Are we confident that the result is not due to chance?
4. Reach one of the following conclusions:
   a. Reject the original position in favor of alternate position **or** fail to reject the initial position

# H0: NULL hypothesis

- Initial position (default, what is already known) that needs to be proven wrong

# H1(a): Alternative hypothesis

- Contrary position to the NULL hypothesis


I am what is
The default, the status quo
I am already accepted, can only be rejected
The burden of proof is on the alternative
I am the null hypothesis
freshspectrum.com

# Example

*“We think customers order different quantities of products when offered a discount”*

H0: Discount does not have an effect on the number of products ordered by a customer.

H1: Discount has an effect on the number of products ordered by a customer.

# What is hypothesis testing?

**Steps**

1. Make assumptions
   a. Take an initial position
   b. Determine the alternate position
2. Conduct fact based tests
   a. Sampling
   b. Decide which test is appropriate
   c. Set acceptance criteria
3. Evaluate results
   a. Does the evaluation support the initial position? Are we confident that the result is not due to chance?
4. Reach one of the following conclusions:
   a. Reject the original position in favor of alternate position **or** fail to reject the initial position

# What is hypothesis testing?

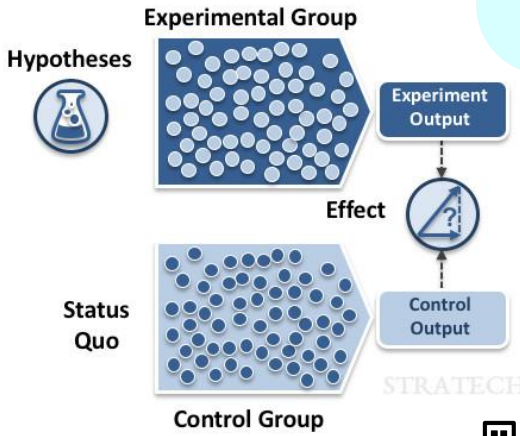
**Steps**

1. Make assumptions
   a. Take an initial position
   b. Determine the alternate position
2. Conduct fact based tests
   a. Sampling
   b. Decide which test is appropriate
   c. Set acceptance criteria
3. Evaluate results
   a. Does the evaluation support the initial position? Are we confident that the result is not due to chance?
4. Reach one of the following conclusions:
   a. Reject the original position in favor of alternate position **or** fail to reject the initial position

# Data collection
## Sampling

- Getting enough observations from a larger population that allows conclusions to be drawn
  - Often random
  - Representative
    - Stratified sampling

- Control *vs* experimental group

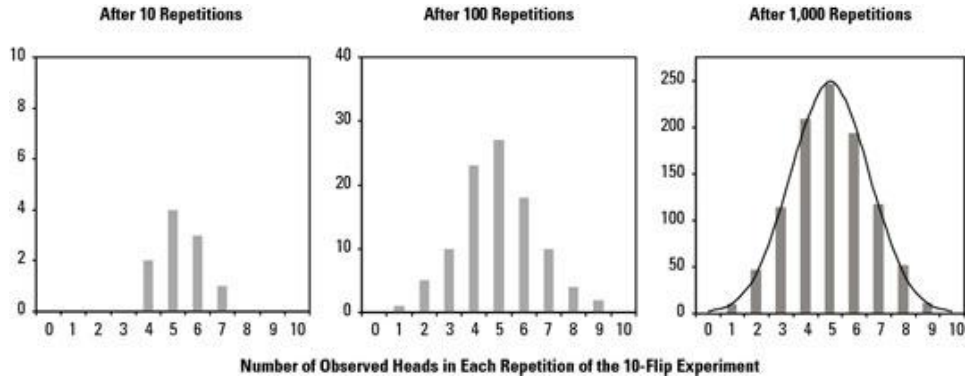


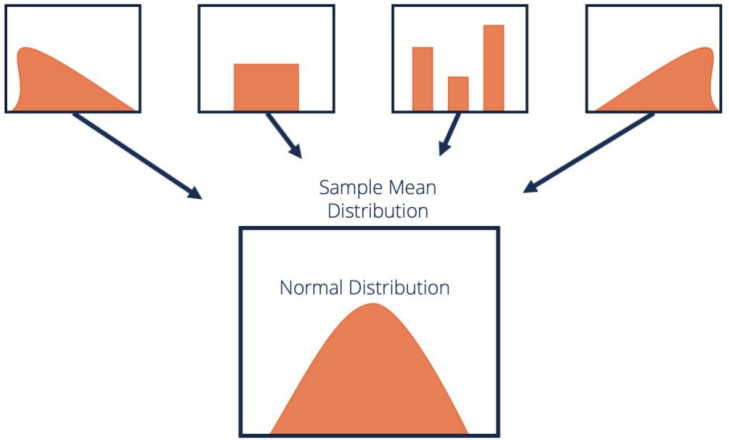
# Central Limit Theorem (CLT)

- The shape of the sampling distribution of means will be normal
  - The **mean** of this sampling distribution will be the population mean
  - The **variance** will be equal to the population variance divided by the sample size
  - The **standard** error will be square root of the variance

- The distribution of sample means is normal regardless of the distribution of the actual population, given a minimum sample size

# Central Limit Theorem (CLT)




*Source: corporatefinanceinstitute.com*

# What is hypothesis testing?

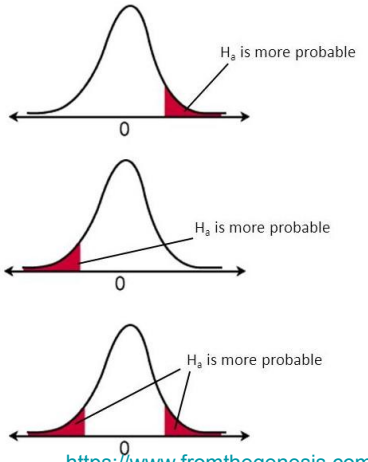
**Steps**

1. Make assumptions
   a. Take an initial position
   b. Determine the alternate position
2. Conduct fact based tests
   a. Sampling
   b. Decide which test is appropriate
   c. Set acceptance criteria
3. Evaluate results
   a. Does the evaluation support the initial position? Are we confident that the result is not due to chance?
4. Reach one of the following conclusions:
   a. Reject the original position in favor of alternate position **or** fail to reject the initial position

# How to test our hypothesis?

**Decide which test is appropriate**

- A test statistic is selected in order to quantify, within observed data, the distinction between the null and the alternative hypothesis




**Right-tail test**

$H_a: \mu > \text{value}$

**Left-tail test**

$H_a: \mu < \text{value}$

**Two-tail test**

$H_a: \mu \neq \text{value}$

https://www.fromthegenesis.com/why-hypothesis-testing/

# How to test our hypothesis?

**Decide which test is appropriate**

- Highly depends on the type of the problem and on the data

- **Check the test's assumptions before using them!**

# How to test our hypothesis?

**Z-test**

- Compares the sample to the population
- Test proportion
  - Outlier removal
- Assumptions
  - Data points should be independent from each other
  - Random sampling from a population
  - Sample is assumed to be normally distributed (N > 30)
  - Population mean and standard deviation are known
- Z-score is measured in standard deviations (standardization)

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

$\overline{X}$ = sample mean
$\mu$ = population mean
$\sigma$ = population standard deviation
n = sample size

# How to test our hypothesis?

## T-test

- When comparing two samples: is there a statistically significant difference?
- Assumptions
  - Population standard deviation is unknown, samples are from a normally distributed population
  - Variance between the samples must be equal, otherwise use Welch's t-test
- T-score
  - Ratio of the difference between two groups: the larger the t-score, the larger the difference
- Different types
  - Independent samples t-test which compares mean for two different groups
  - Paired sample t-test which compares means from the same group at different times
- Hypothesis:
  - H0: the means of the samples are equal
  - H1: the means of the samples are different

# How to test our hypothesis?

## ANOVA

- **Analysis of variance**, used to compare multiple samples with a single test
  - Compares the difference between the three or more samples/groups of a single independent variable
- Assumptions
  - The samples are independent
  - Each sample is from a normally distributed population
  - Variance between the samples must be equal
  - If the assumptions are not met, Kruskal-Wallis test could be an alternative

- Hypothesis:
  - H0: all pairs of samples are same, i.e. all sample means are equal
  - H1: at least one pair of samples is significantly different

# How to test our hypothesis?

## Chi-square

- Used to compare categorical variables

- Goodness of fit test
  - Determines if a sample matches the population

- A chi-square test for two independent variables
- Hypothesis:
  - H0: Variable A and Variable B are independent
  - H1: Variable A and Variable B are not independent

# How to test our hypothesis?

**Chi-square**

## Example

- In an election survey, voters might be classified by gender (male or female) and voting preference (Democrat, Republican, or Independent)

- We could use a chi-square test for independence to determine whether gender is related to voting preference
  - Correlation between categorical variables

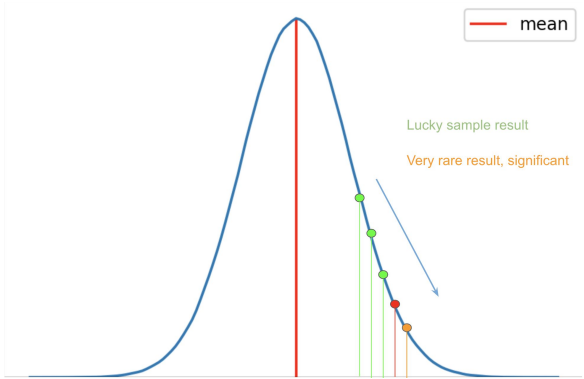
# What is hypothesis testing?

**Steps**

1. Make assumptions
    a. Take an initial position
    b. Determine the alternate position
2. Conduct fact based tests
    a. Sampling
    b. Decide which test is appropriate
    c. Set acceptance criteria
3. Evaluate results
    a. Does the evaluation support the initial position? Are we confident that the result is not due to chance?
4. Reach one of the following conclusions:
    a. Reject the original position in favor of alternate position **or** fail to reject the initial position

# Significance level

- Output from the test statistics
- Select a significance level (α), a probability threshold below which the null hypothesis will be rejected. Common values are 5% (standard) and 1%



# What is hypothesis testing?

**Steps**

1. Make assumptions
   a. Take an initial position
   b. Determine the alternate position
2. Conduct fact based tests
   a. Sampling
   b. Decide which test is appropriate
   c. Set acceptance criteria
3. Evaluate results
   a. Does the evaluation support the initial position? Are we confident that the result is not due to chance?
4. Reach one of the following conclusions:
   a. Reject the original position in favor of alternate position **or** fail to reject the initial position

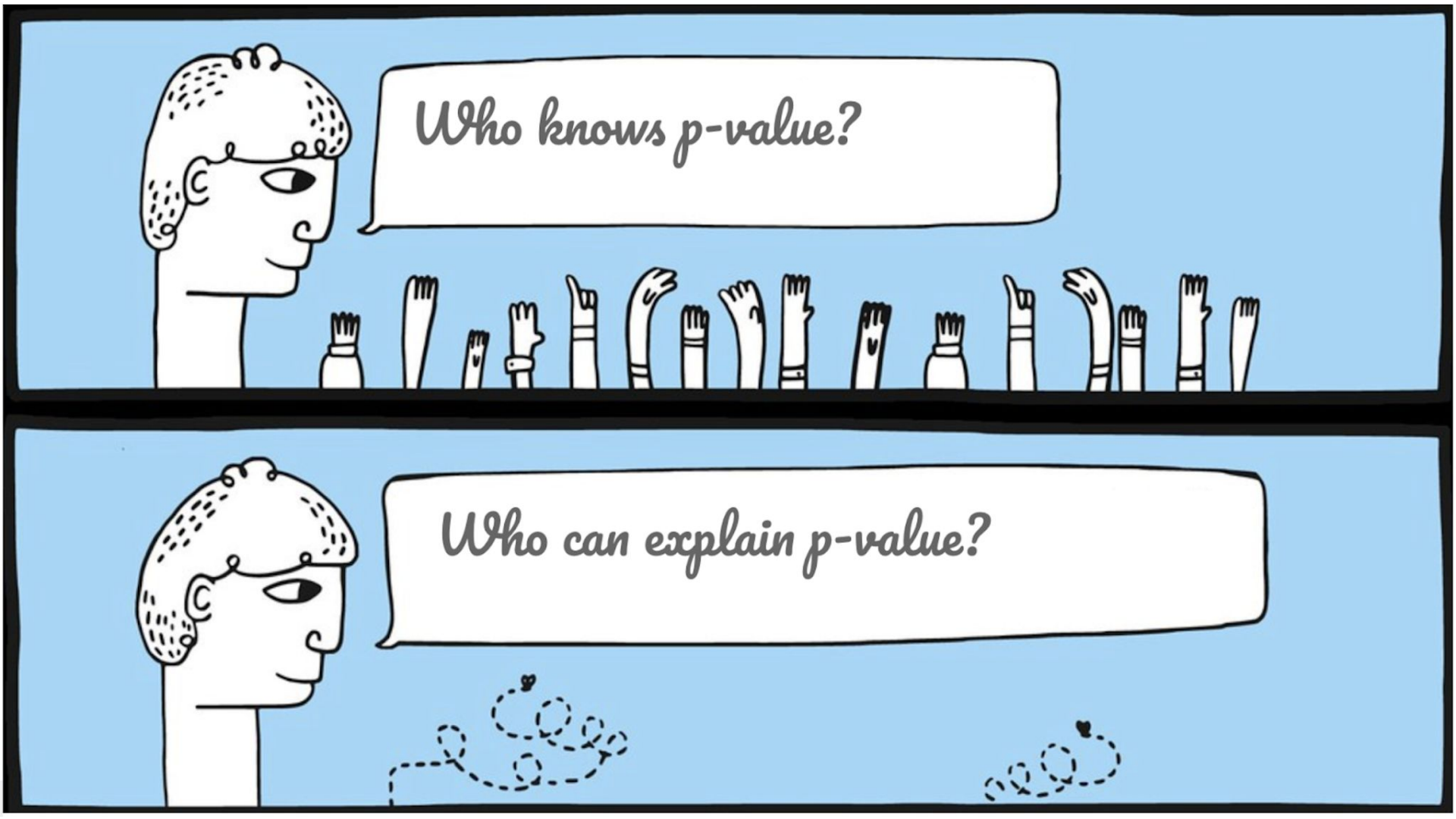
# How to test our hypothesis?

## Tests results

- From the scores, we can calculate the **p-values**
  - Z-table, t-table

- Libraries often provide the score and the corresponding p-value as part of the output of an statistical test

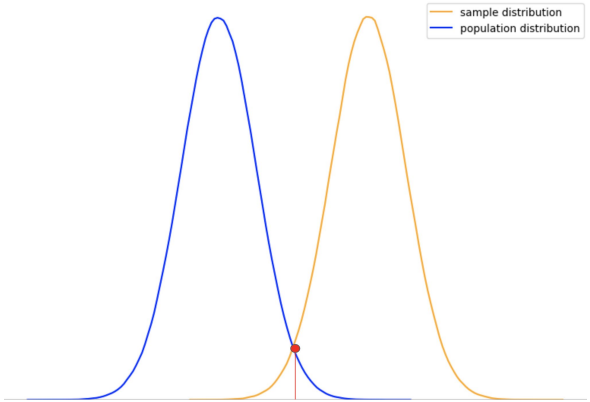
```
from scipy.stats import ...
```


Who knows p-value?
Who can explain p-value?

# How to test our hypothesis?

## p-values

- Probability value (p-value) is the probability that, when the null hypothesis is true, the statistical-test result would be equal to (or more extreme than) the actual observed results
- p-value does not hold any value by itself
  - A large p-value implies that sample scores are more aligned or similar to the population score: our data is highly consistent with our null hypothesis



# How to test our hypothesis?

## p-values

- How do you interpret a p-value?
- How much importance should we place in the p-value?
- How will you explain the significance of p-value to a non-data science person (a stakeholder for example)?

# How to test our hypothesis?

## p-values in Data Science

- p-value is an important metric in the process of feature selection
  - Find out the best subset of the independent variables to build the model
  - Throwing in redundant and non-contributing variables adds complexity to the model
    - They can reduce the model performance in terms of accuracy and runtime

# How to test our hypothesis?

**p-values in Data Science**

## Example

- Consider a dataset that contains the following information about different startups, for which we wanna **predict the profit**:
  - `State California`
  - `State Florida`
  - `R&D spend`
  - `Administration`
  - `Marketing spend`
- **H0**: The independent variable has no significant effect over the target variable
- **H1**: The independent variable has a significant effect on the target variable

# How to test our hypothesis?

## p-values in Data Science

### Example

- All the variables, except `R&D Spend` have a p-value over 0.05

- Can they be removed from the dataset?

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | y | R-squared: | 0.951 |
| Model: | OLS | Adj. R-squared: | 0.945 |
| Method: | Least Squares | F-statistic: | 169.9 |
| Date: | Tue, 03 Sep 2019 | Prob (F-statistic): | 1.34e-27 |
| Time: | 09:22:23 | Log-Likelihood: | -525.38 |
| No. Observations: | 50 | AIC: | 1063. |
| Df Residuals: | 44 | BIC: | 1074. |
| Df Model: | 5 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 5.013e+04 | 6884.820 | 7.281 | 0.000 | 3.62e+04 | 6.4e+04 |
| data[0] | 198.7888 | 3371.007 | 0.059 | 0.953 | -6595.030 | 6992.607 |
| data[1] | -41.8870 | 3256.039 | -0.013 | 0.990 | -6604.003 | 6520.229 |
| data[2] | 0.8060 | 0.046 | 17.369 | 0.000 | 0.712 | 0.900 |
| data[3] | -0.0270 | 0.052 | -0.517 | 0.608 | -0.132 | 0.078 |
| data[4] | 0.0270 | 0.017 | 1.574 | 0.123 | -0.008 | 0.062 |

# What is hypothesis testing?

**Steps**

1. Make assumptions
    a. Take an initial position
    b. Determine the alternate position
2. Conduct fact based tests
    a. Sampling
    b. Decide which test is appropriate
    c. Set acceptance criteria
3. Evaluate results
    a. Does the evaluation support the initial position? Are we confident that the result is not due to chance?
4. Reach one of the following conclusions:
    a. Reject the original position in favor of alternate position **or** fail to reject the initial position

# Outcomes

1. Reject the NULL hypothesis
    - If there are statistically significant evidences that suggest that the alternate hypothesis is valid, then the NULL hypothesis is rejected

2. Fail to reject the NULL hypothesis

# Type I and Type II errors

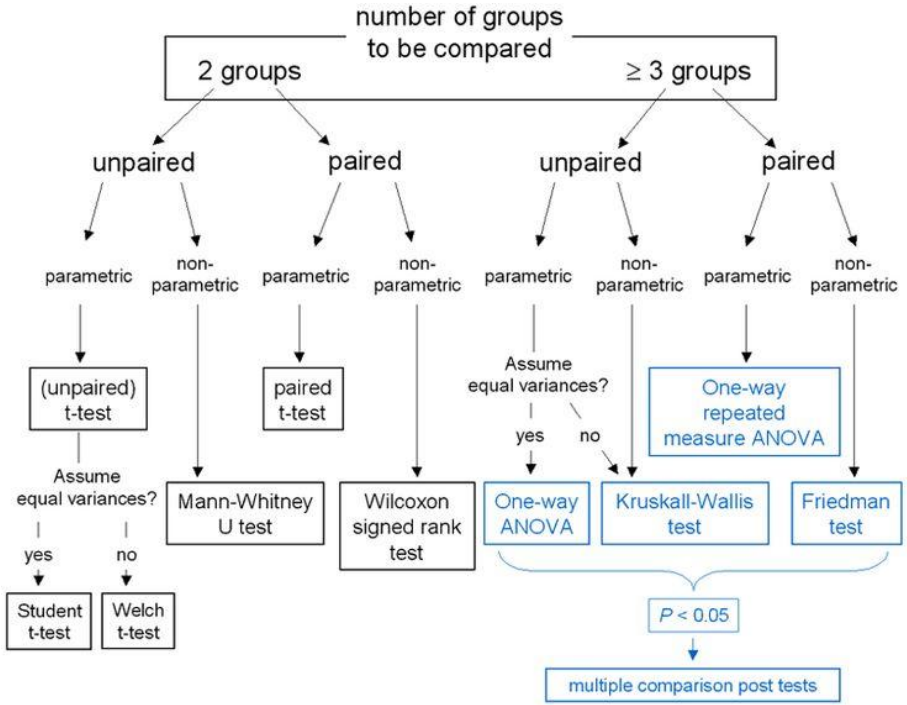
- Type I error (false positive)
  - Rejection of a true null hypothesis
  - Usually leads to the conclusion that a supposed effect or relationship exists when in fact it does not

- Type II error (false negative)
  - The failure to reject a false null hypothesis

| Table of error types | | Null hypothesis ($H_0$) is | |
|---|---|---|---|
| | | **True** | **False** |
| **Decision about null hypothesis ($H_0$)** | **Don't reject** | Correct inference (true negative) (probability = 1 - $\alpha$) | Type II error (false negative) (probability = $\beta$) |
| | **Reject** | Type I error (false positive) (probability = $\alpha$) | Correct inference (true positive) (probability = 1 - $\beta$) |

# Summary

Continuous variables



# References


[1] Test statistic

[2] Hypothesis testing

[3] Your guide to master hypothesis testing in statistics

[4] T-test

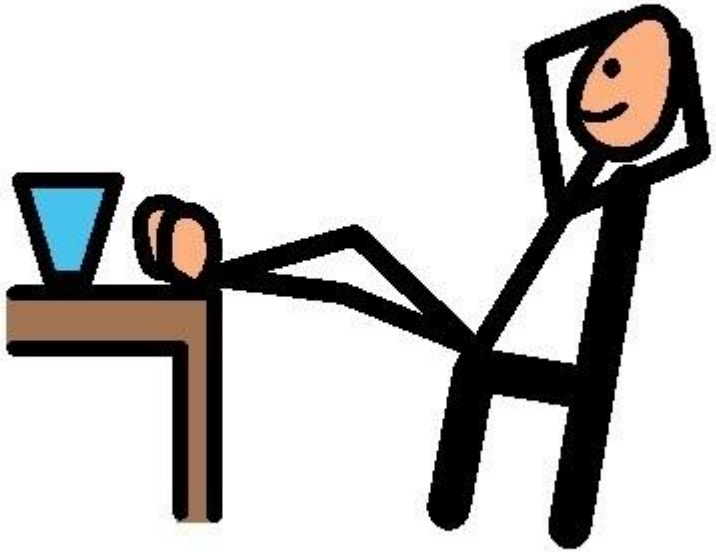
[5] Python for Data Analysis: Hypothesis testing and t-test

[6] Chi-square test for Machine Learning

[7] Statistical hypothesis testing in Python

[8] Type I and Type II errors

# break



the shortcut × moz://a

DATA SCIENCE

TRAINING PROGRAM

# (More) Data Visualization

**Larissa Leite**
**Kodit.io**

kodit.io selko SILO.AI

# Visualization

- Know **what** you want to communicate and **who** you want to communicate to

- Should be concise
  - All the information
  - As simply as possible
- Should be perceivable
  - Easy to interpret

# Visualization

## Libraries

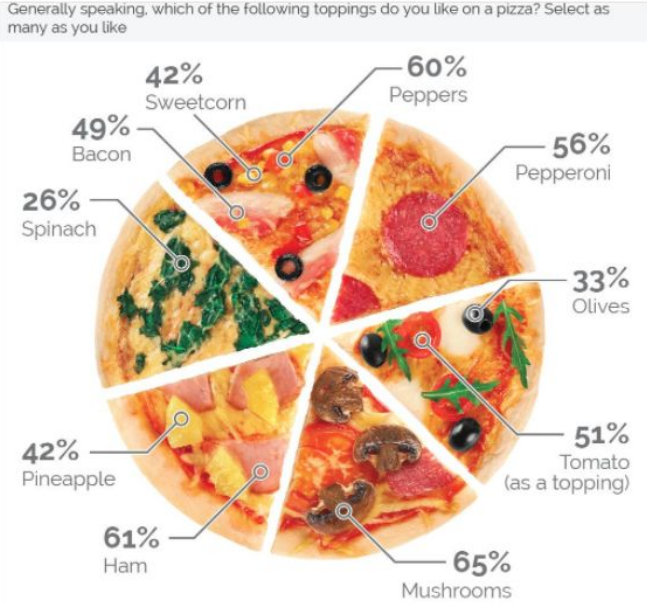
```
import matplotlib.pyplot as plt
```

- Sometimes external libraries provide more options
- Seaborn
  - More visualizations than matplotlib; visually better defaults and color schemes
- Bokeh
  - Browser based; interactive visualizations
- Plot.ly
  - Online service with Python API
- D3.js
- Highcharts, Sigma JS (graphs)
- Tableau

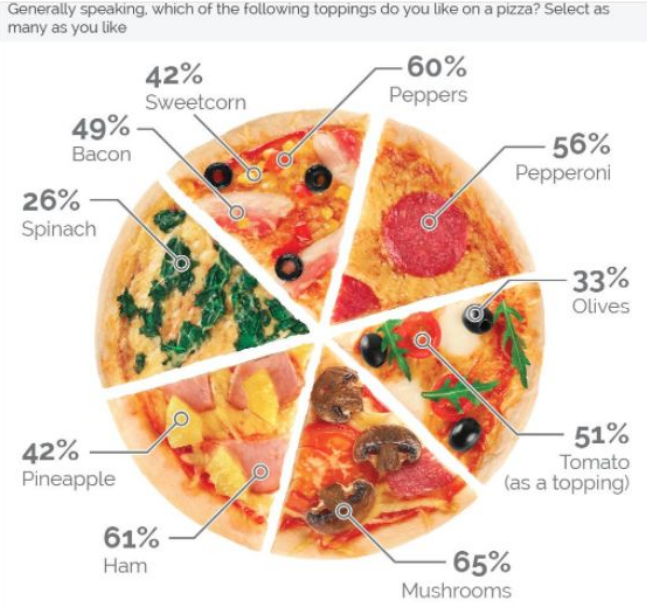
# Visualization

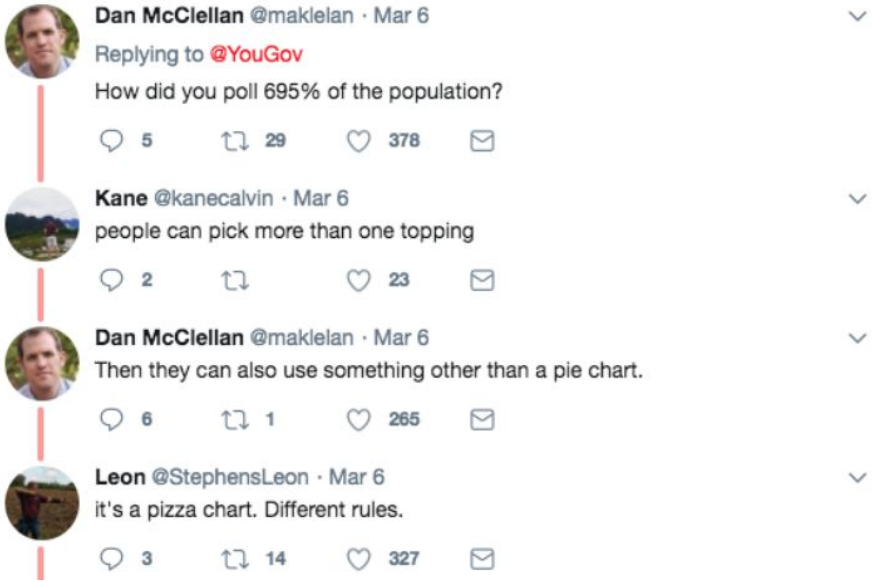
- ☑ Univariate distribution plots (histograms, densities)
- ☑ Scatter plots
- ☑ Summary statistic plots (bar charts)

# Visualization

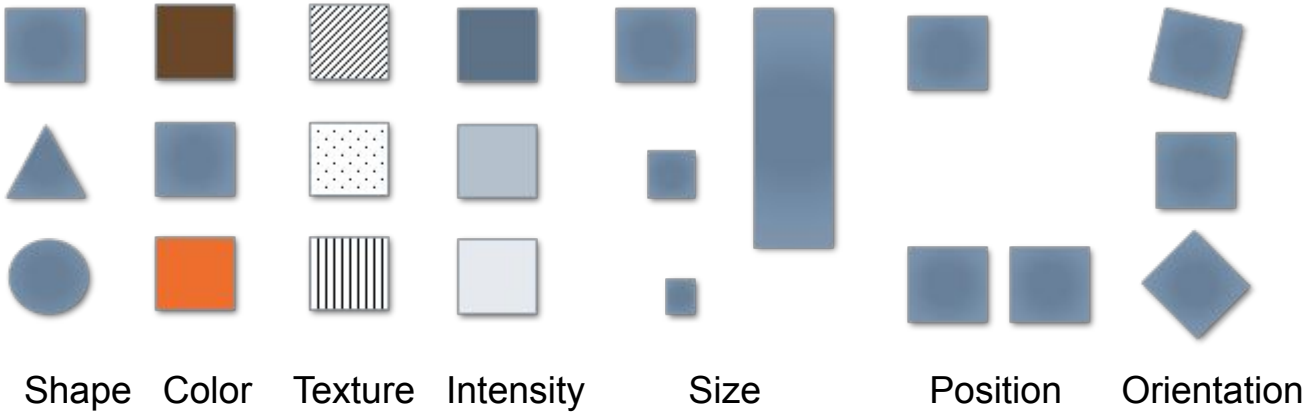


# Visualization

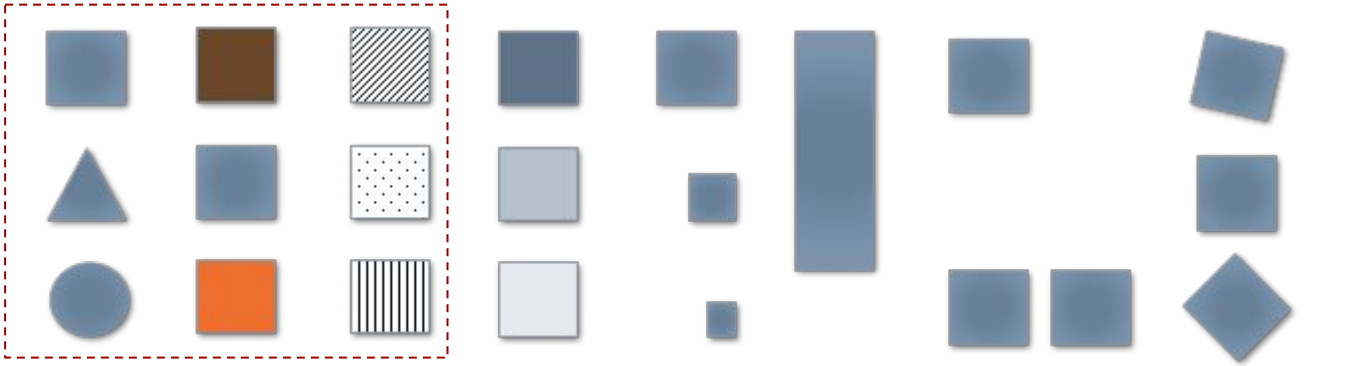


# Visual variables



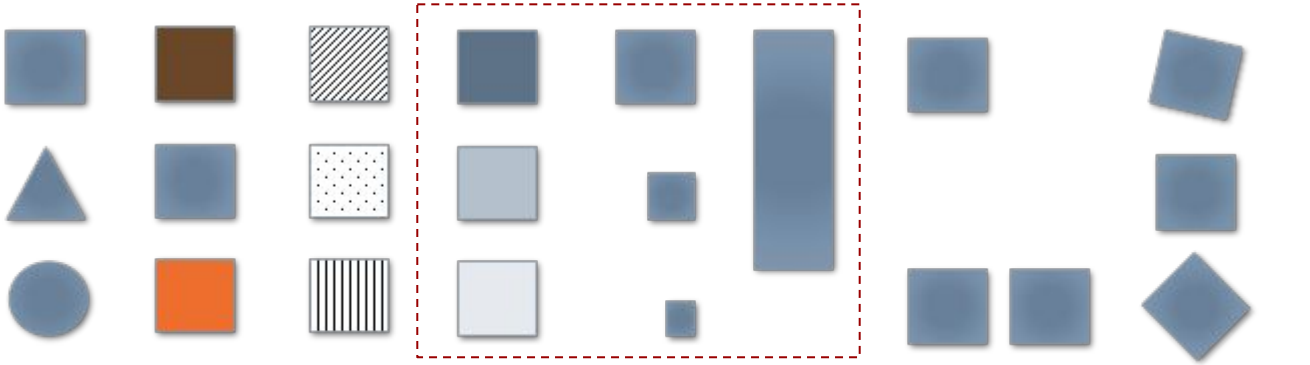
# Visual variables



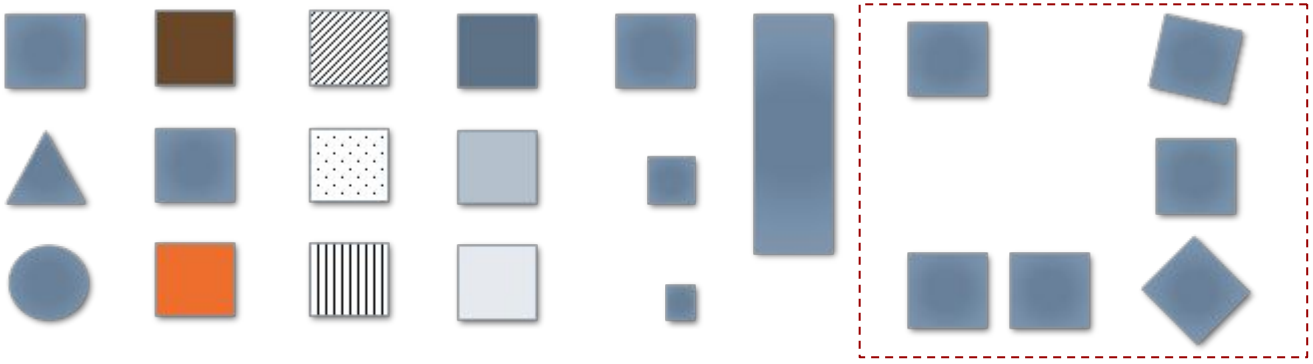
Shape Color Texture Intensity Size Position Orientation

Distinguish different categories

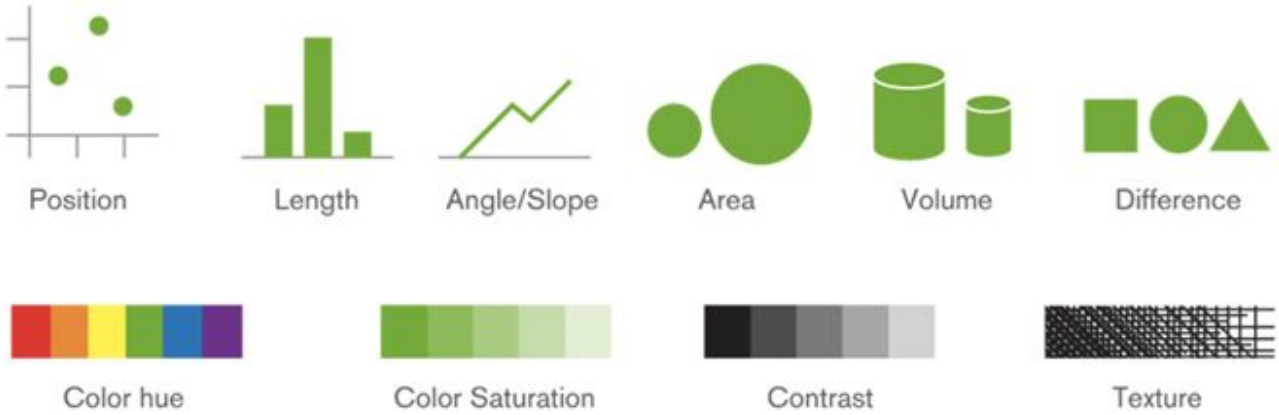
# Visual variables



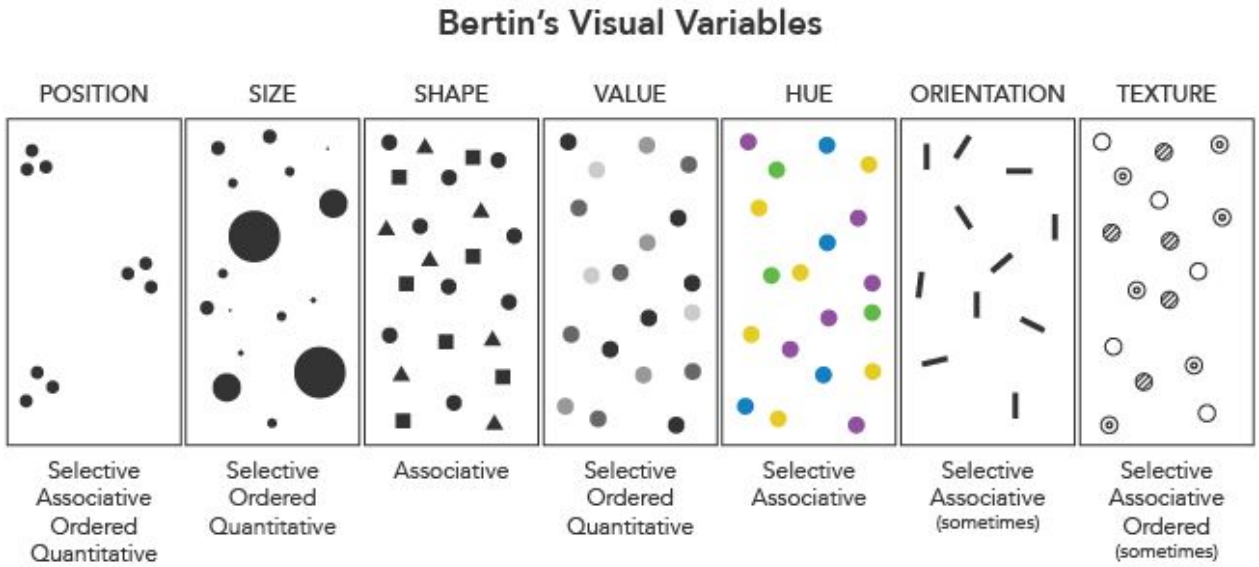
# Visual variables



# Visualization



# Visualization



# Visualization

## Color

- Given any 2 colors, make it intuitively obvious which represents “higher” and which represents “lower”

Gray scale

Intensity interpolation

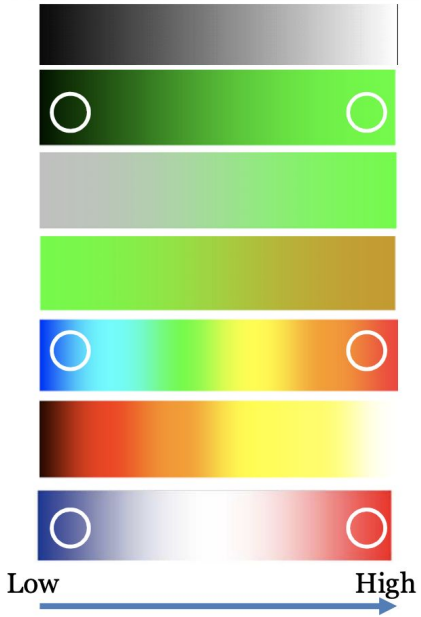
Saturation interpolation

Two-color interpolation

Rainbow scale

Heated object interpolation

Blue-White-Red



# Visualization

## Color

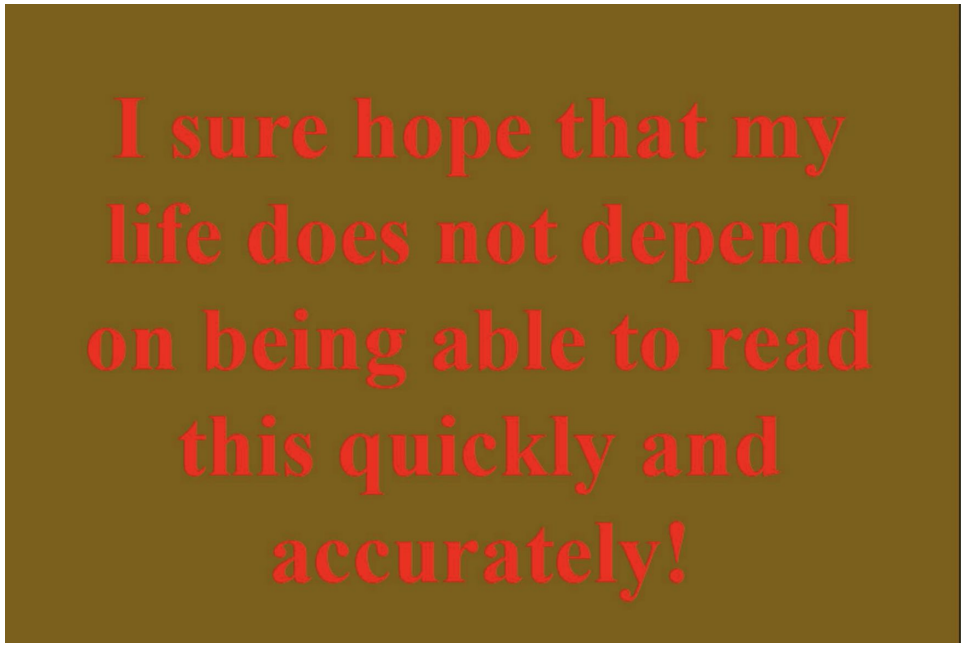
- Do not attempt to fight pre-established color meanings

| Red | Green | Blue |
| --- | --- | --- |
| Stop | On | Cool |
| Off | Plants/nature | Safe |
| Dangerous | Moving | Nitrogen |
| Hot | Money | |
| High stress | | |
| Money loss | | |

# Visualization

## Color

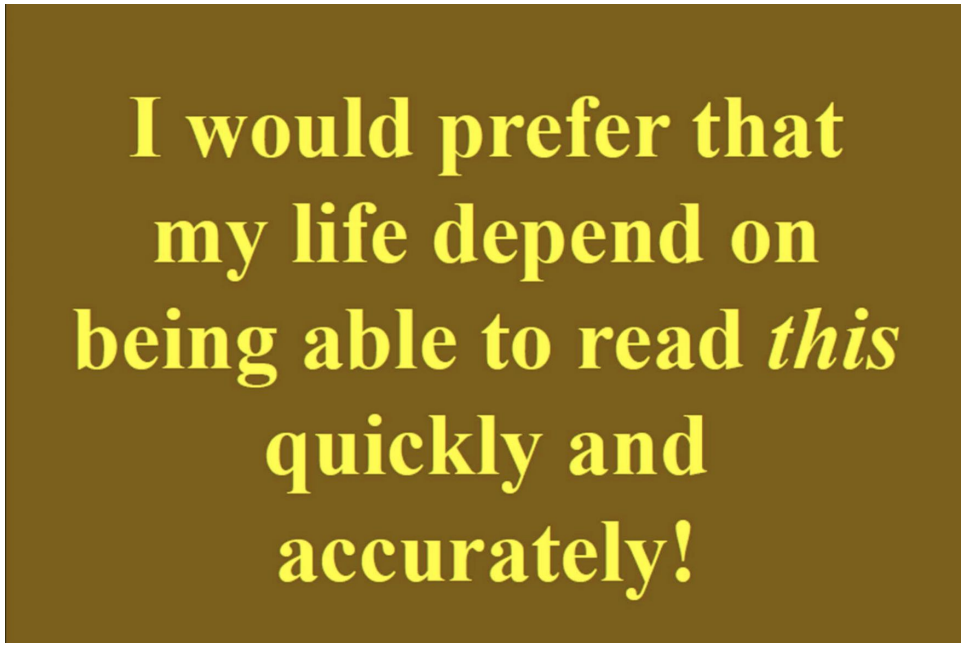
- Attention to contrast!!!



# Visualization

## Color

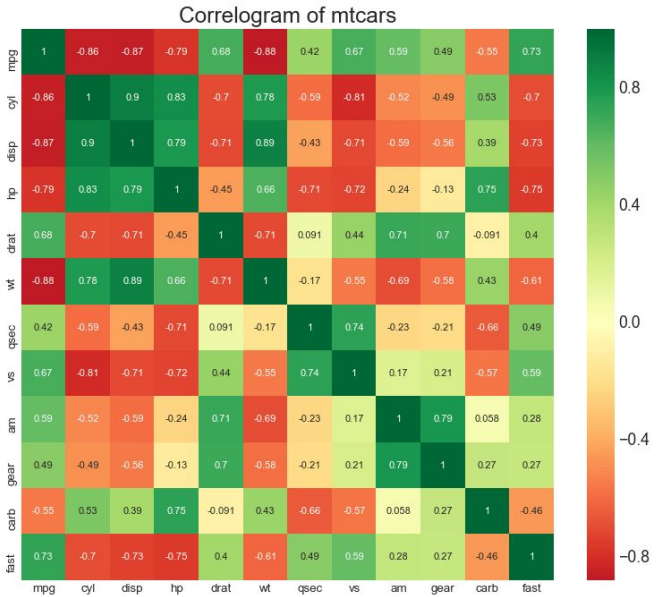
- Attention to contrast!!!




LET'S GET DOWN TO BUSINESS

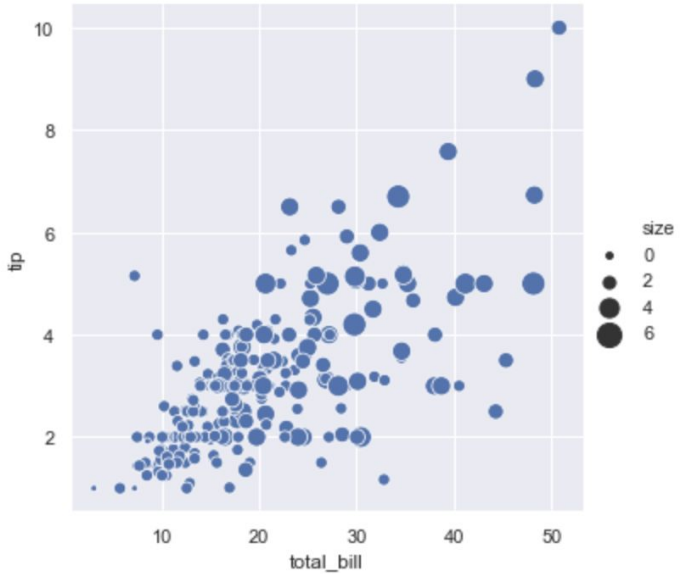
# Visualization
## Correlogram

Plot correlations between (all) the variables, a specific kind of heat map



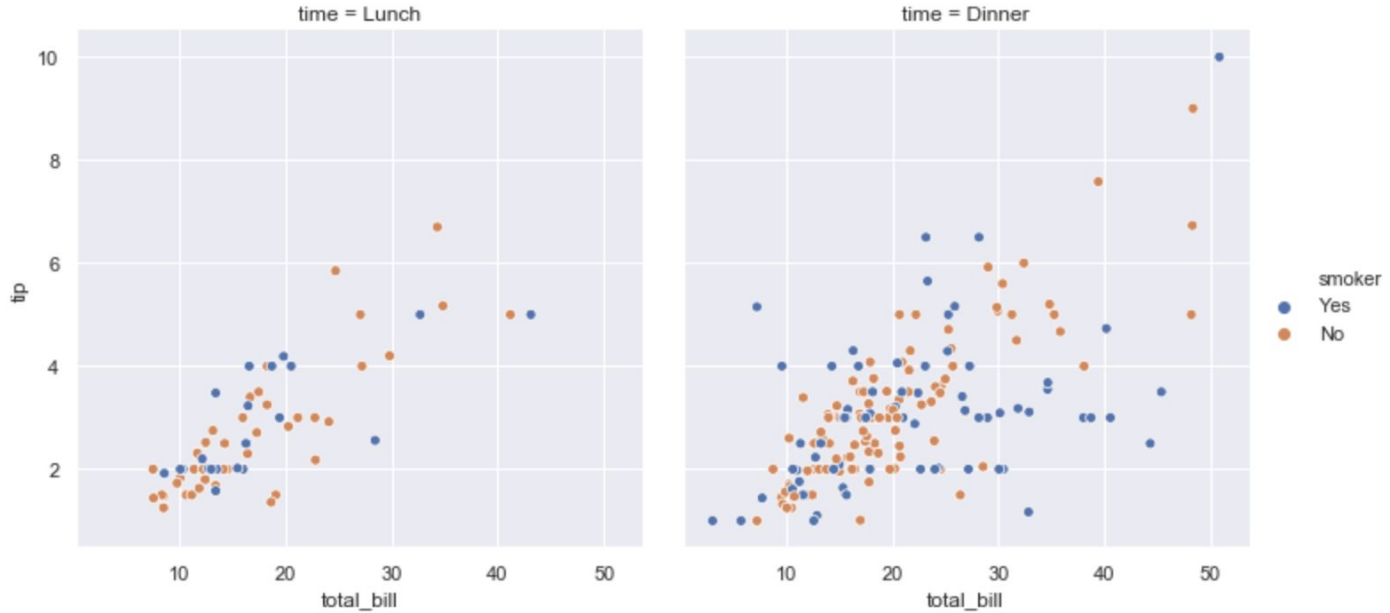
# Visualization

Relationship between two or more numerical variables



# Visualization

Relationship between categorical and numerical variables

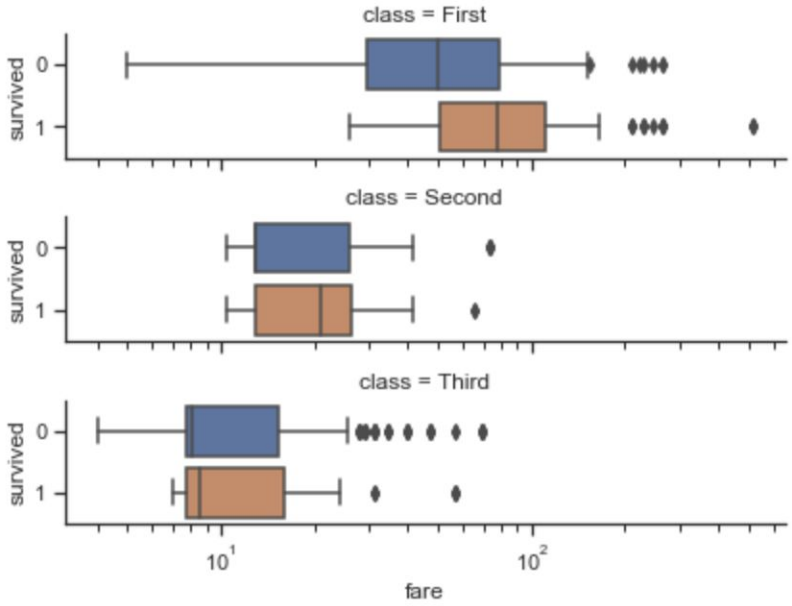


# Visualization
## Facet plot

Multivariate relationships

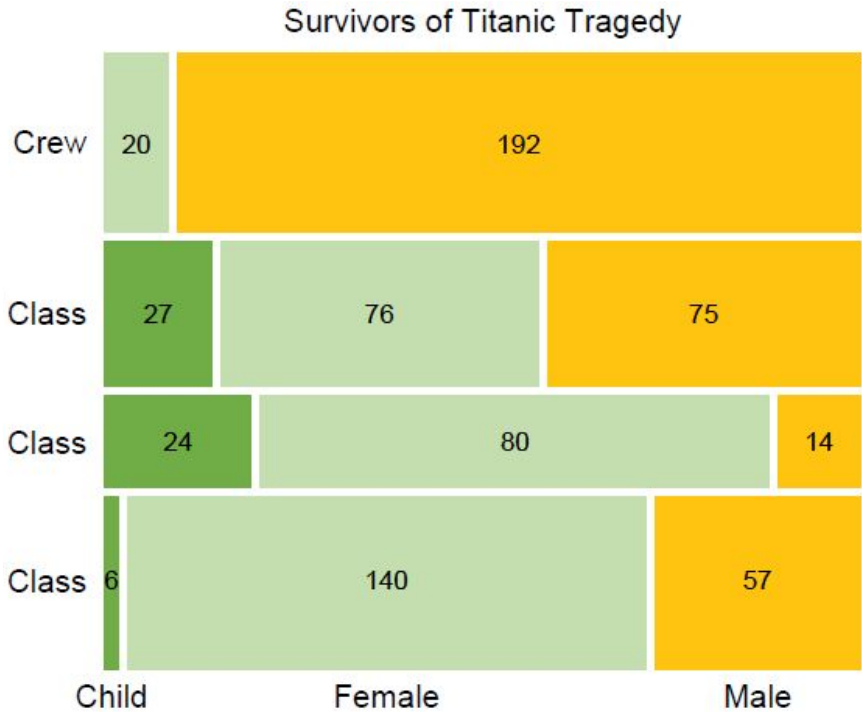
```
g = sns.catplot(x="fare", y="survived", row="class",
                kind="box", orient="h",
                height=1.5, aspect=4,
                data=titanic.query("fare > 0"))


g.set(xscale="log")
```



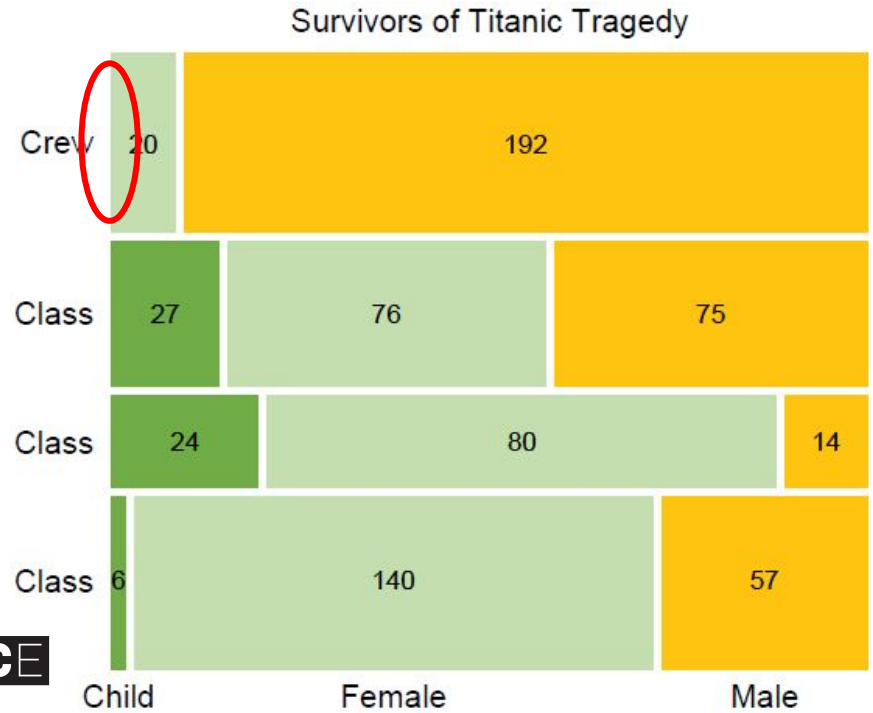
# Visualization
## Mosaic Plot

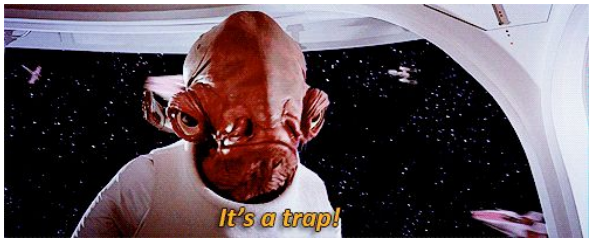
Relationship between two or more categorical variables



# Visualization
## Mosaic Plot

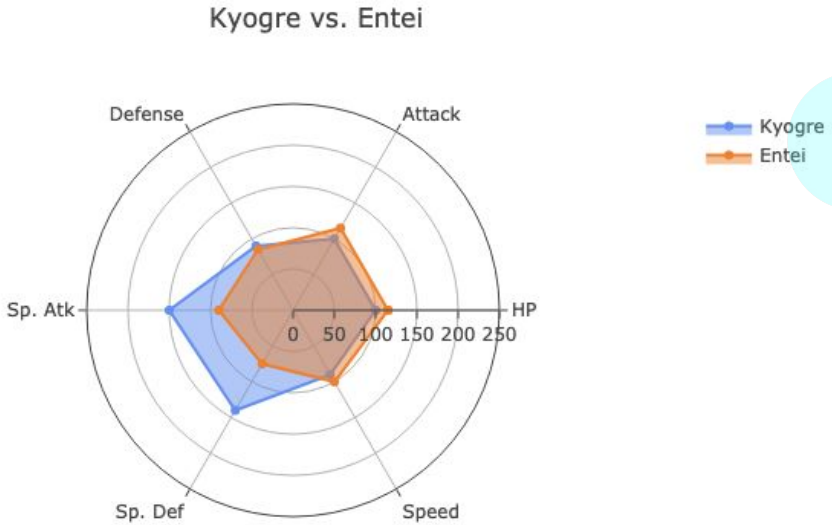
Relationship between two or more categorical variables



# Visualization
## Radar chart

Spider web that indicates variables and their values used to compare two (or more) entities; thresholds are often used

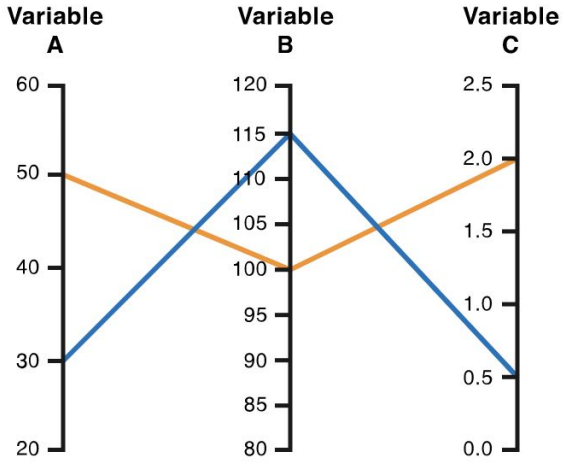


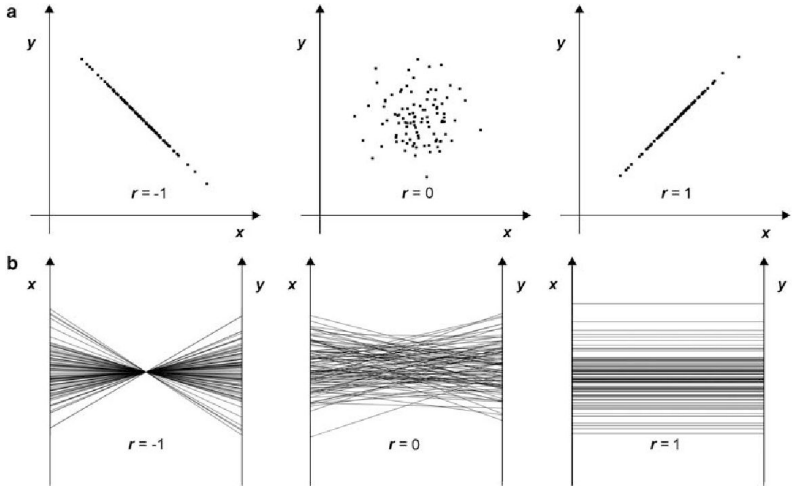
# Visualization

## Parallel coordinates

Used to compare many numerical variables and their relationships

| Data | | | |
|---|---|---|---|
| | Variable A | Variable B | Variable C |
| Item 1 | 50 | 100 | 2.0 |
| Item 2 | 30 | 115 | 0.5 |



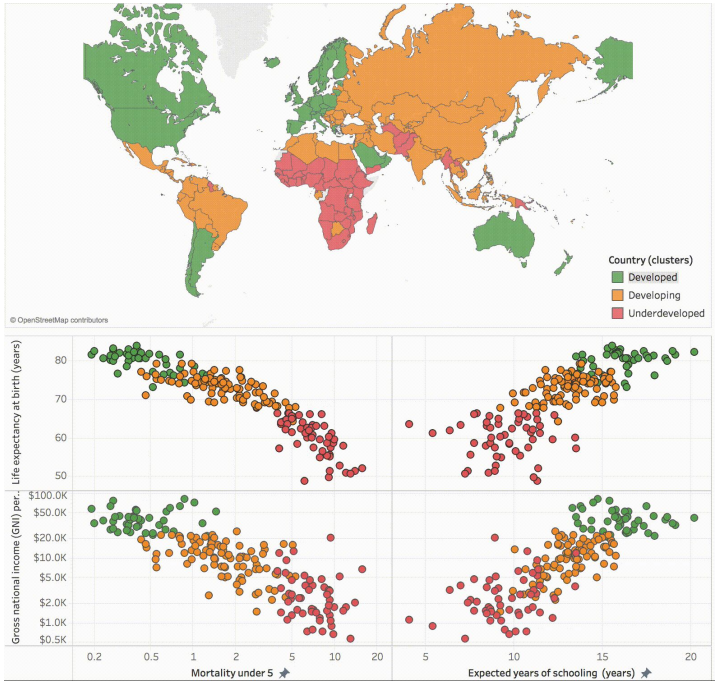

Negative correlation

No correlation

Positive correlation

# Visualization
## Clustering

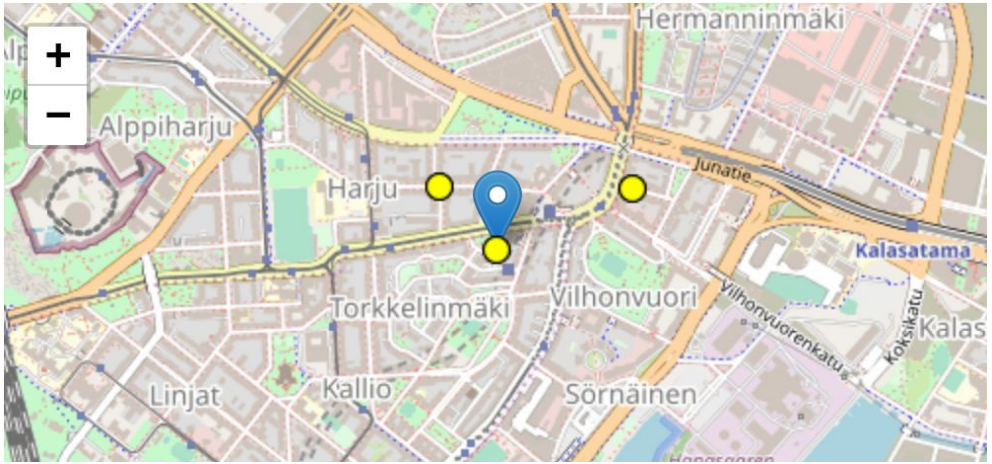
If it gets too cluttered…. grouping can help!



# Visualization
## Maps

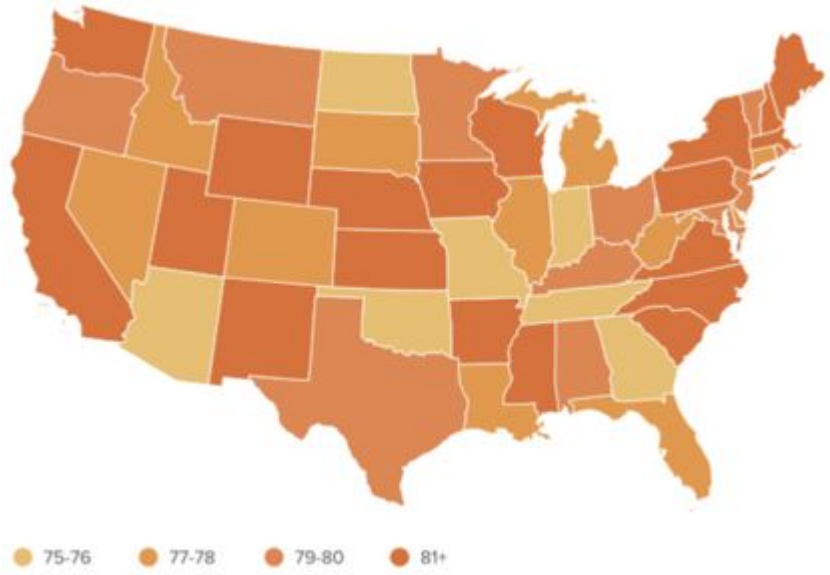
Leaflet.js, Mapbox

Zoomed-in map with points of interest

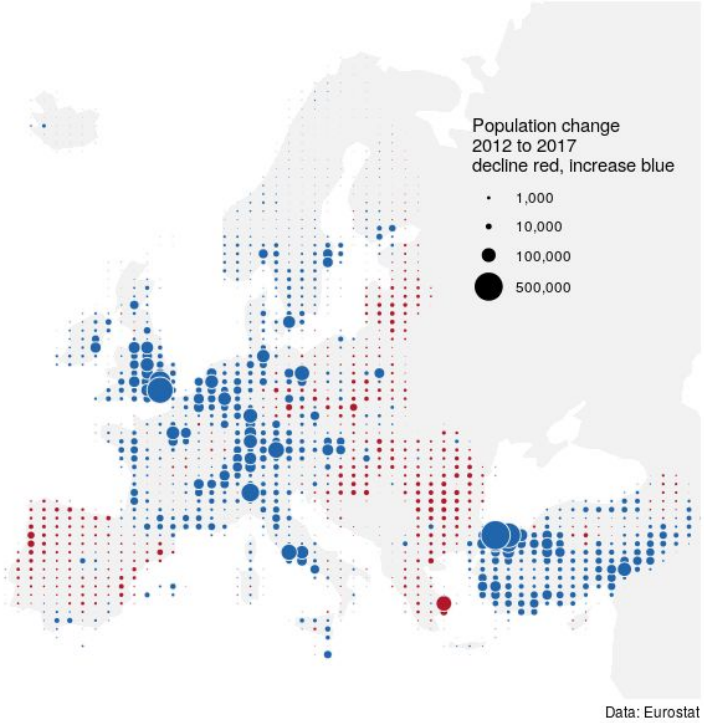


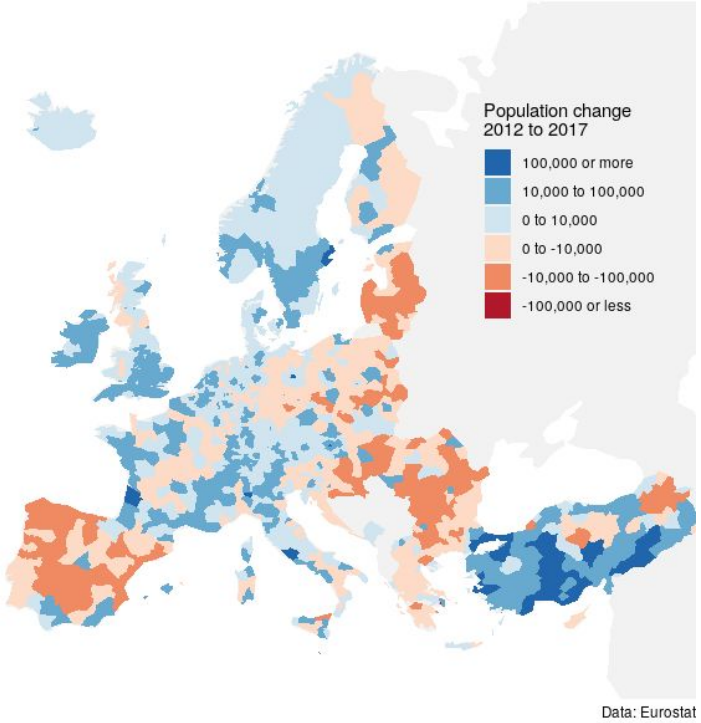

Number of sales by state

What about the color selection? Is it easy to tell them apart?



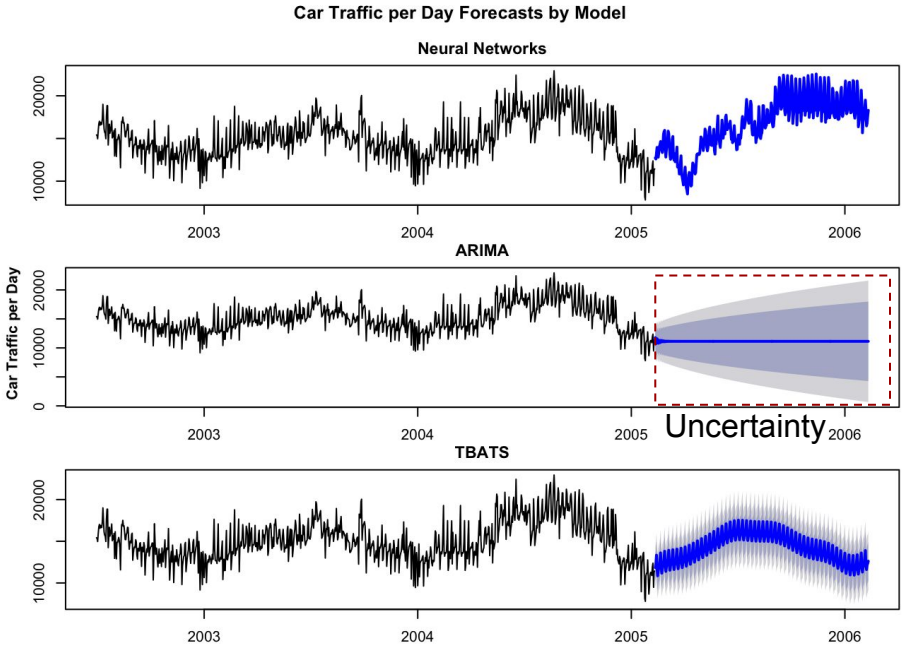
# Visualization
## Maps

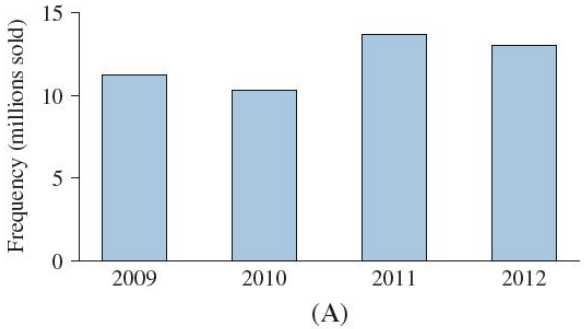


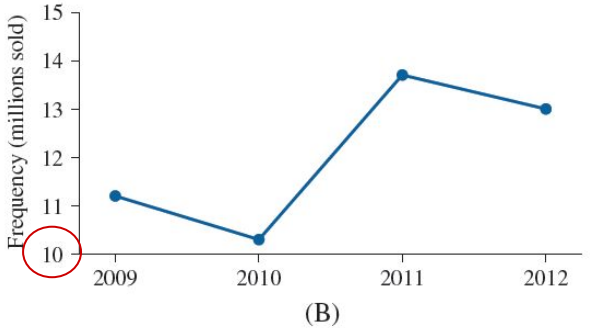



# Visualization

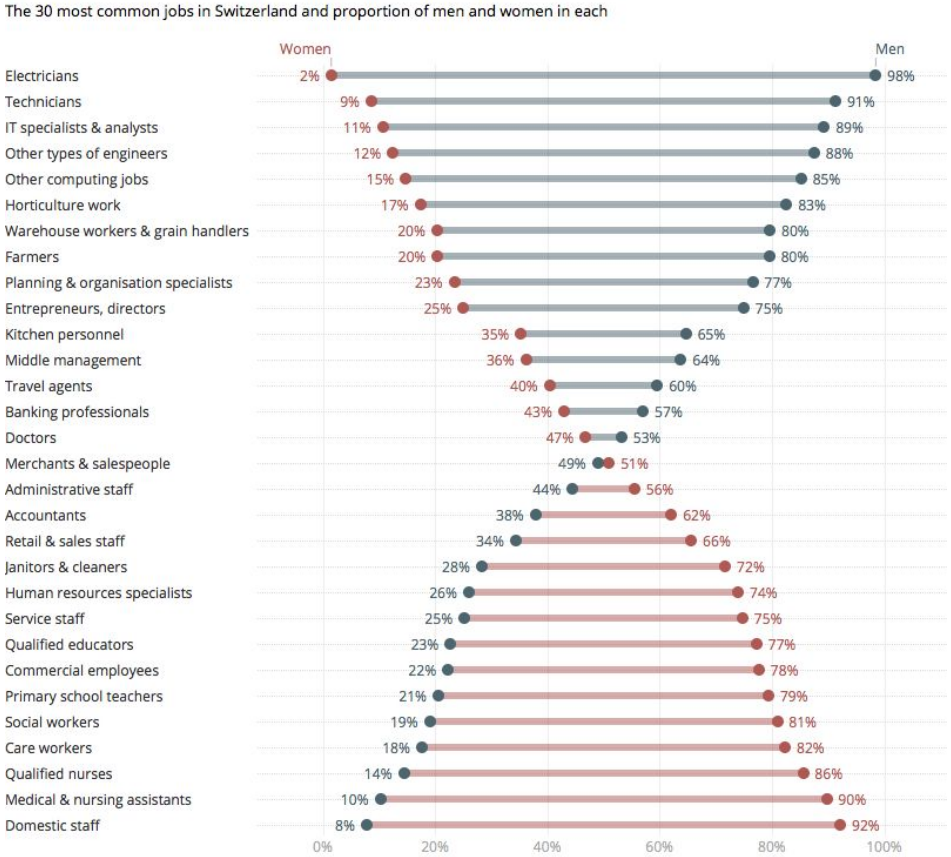
## Time Series



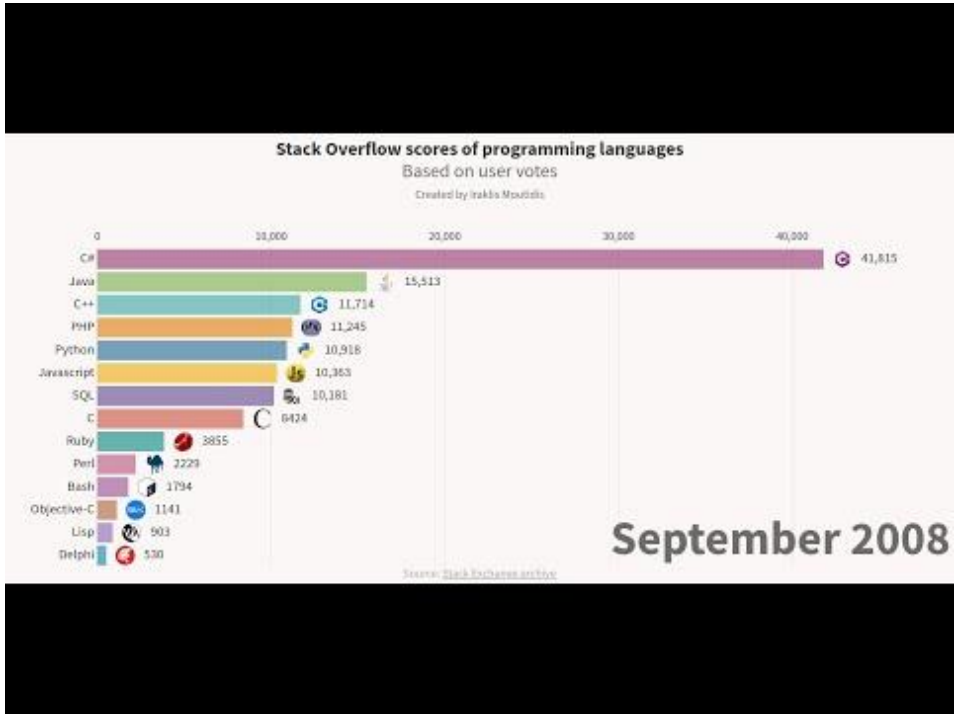



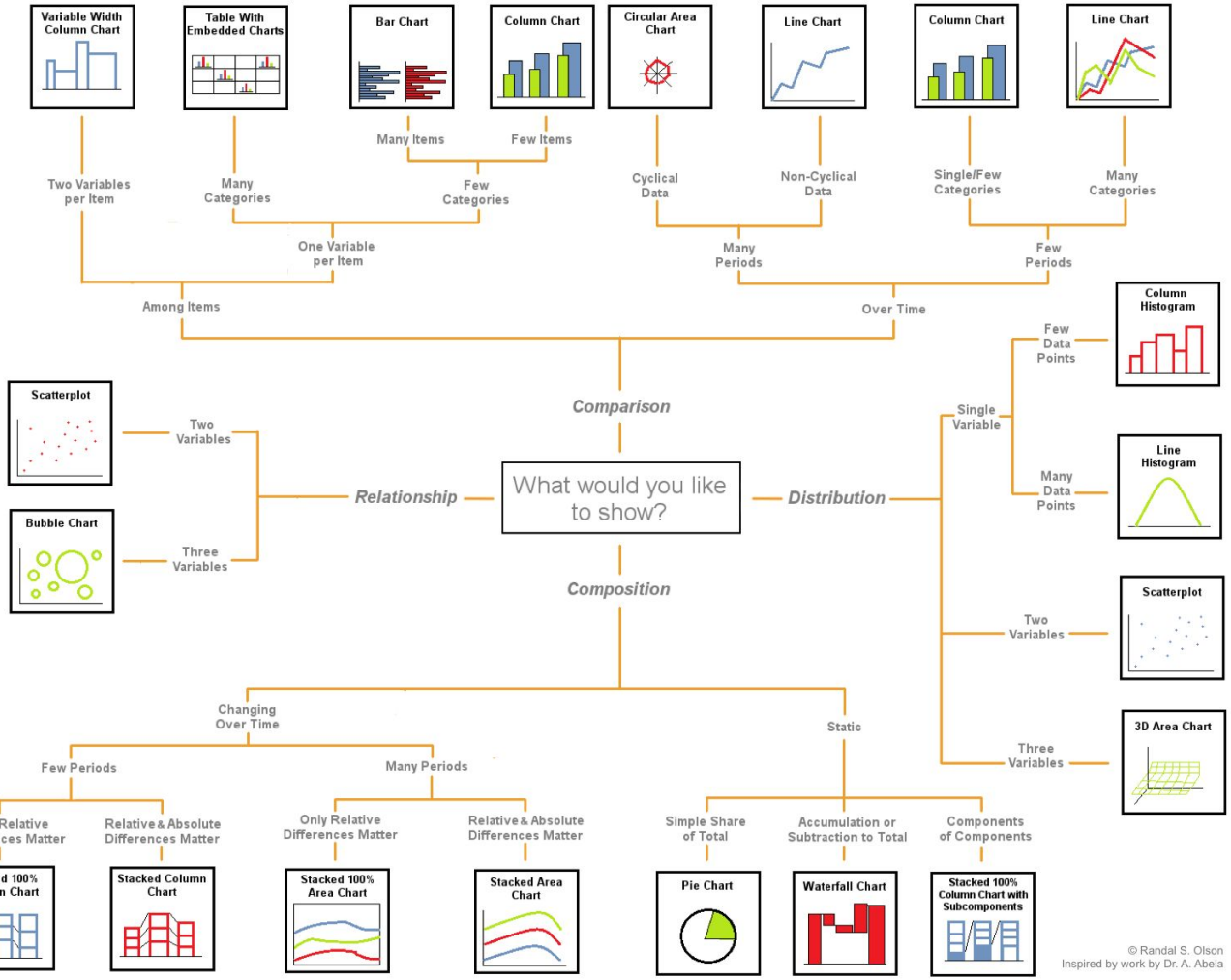
# Visualization



# Visualization




# The chart selector — some basic chart suggestions




© Randal S. Olson
Inspired by work by Dr. A. Abela

# References

[1] The Data Visualization Catalogue

[2] From Data to Viz

[3] Information Visualization

[4] Introduction to Information Visualization

[5] Tamara Munzner: Data visualization talks

[6] 20 visualization tools

[7] Seaborn tutorial