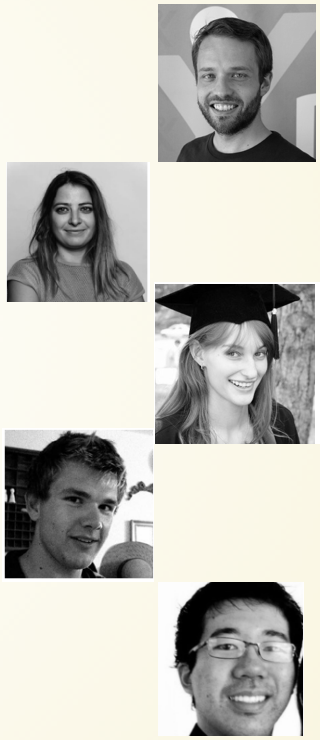
Explorative Data Analysis



- Looking at data
- Distributions
- **LUNCH**
- A case study
- What if too many dimensions?

Tereza Iofciu, Alisa Dammer, Jan Bilek, Caio Miyashiro, Philipp Kähler

Data Science, @mytaxi

Part 1

What to look for

Summarising data

Detailed view

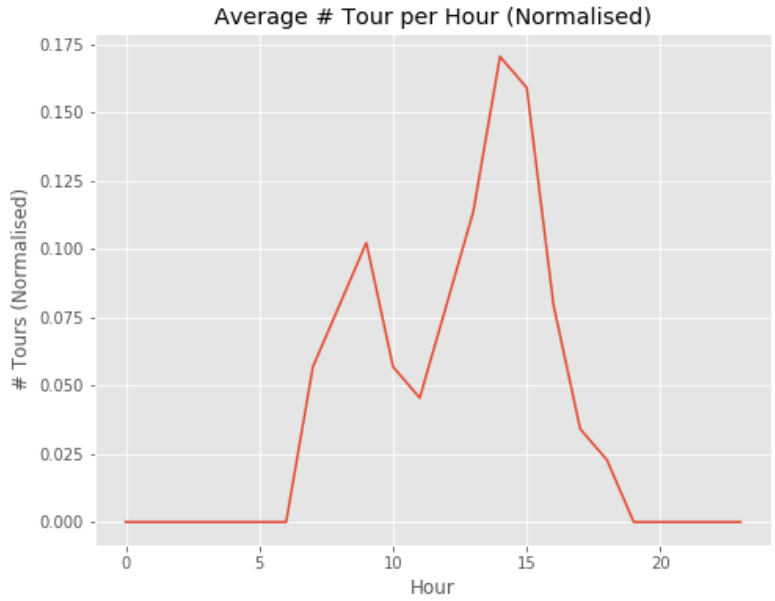
EXPLORATORY ANALYSIS = DETECTIVE WORK

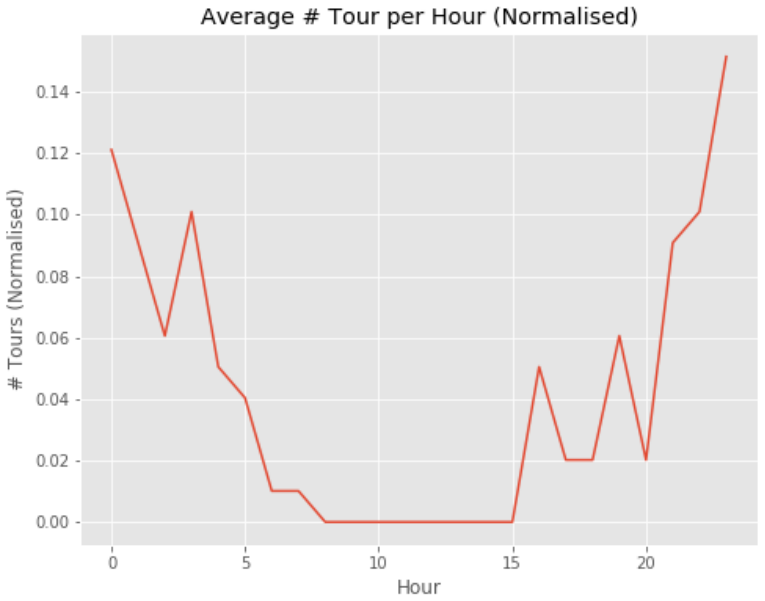
WHAT CAN WE LOOK FOR IN THE DATA

CLUES

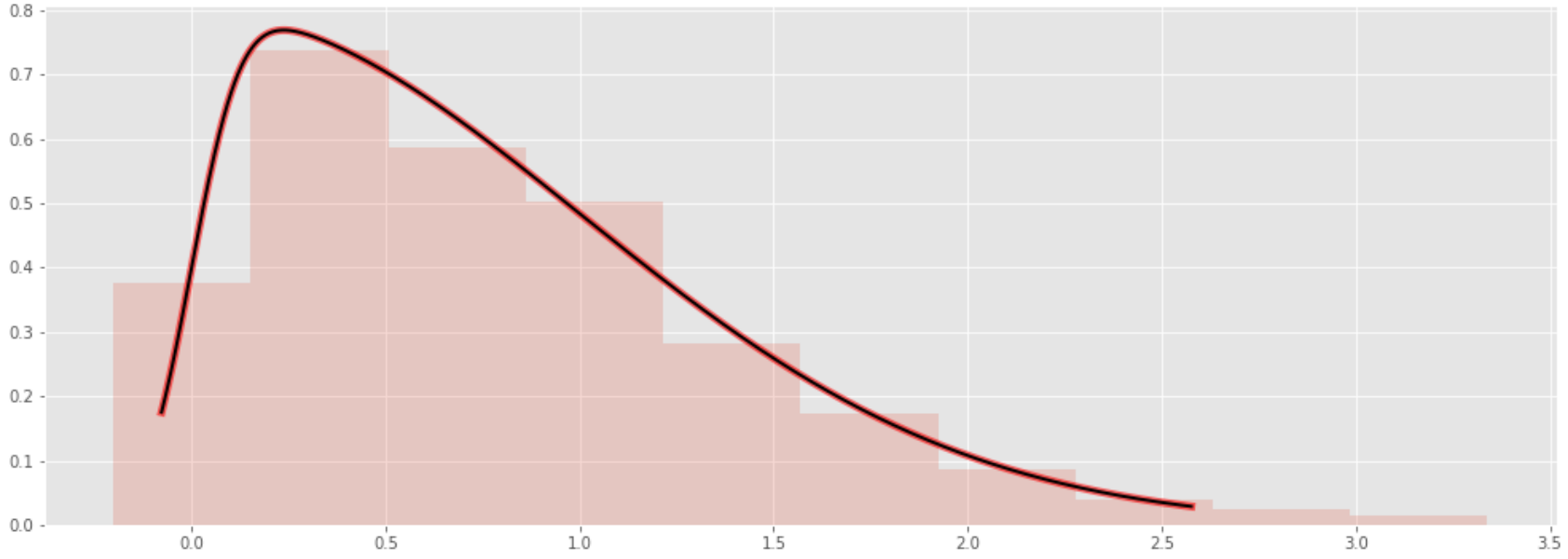
TOOLS

APPEARANCES OF SEPARATION INTO GROUPS

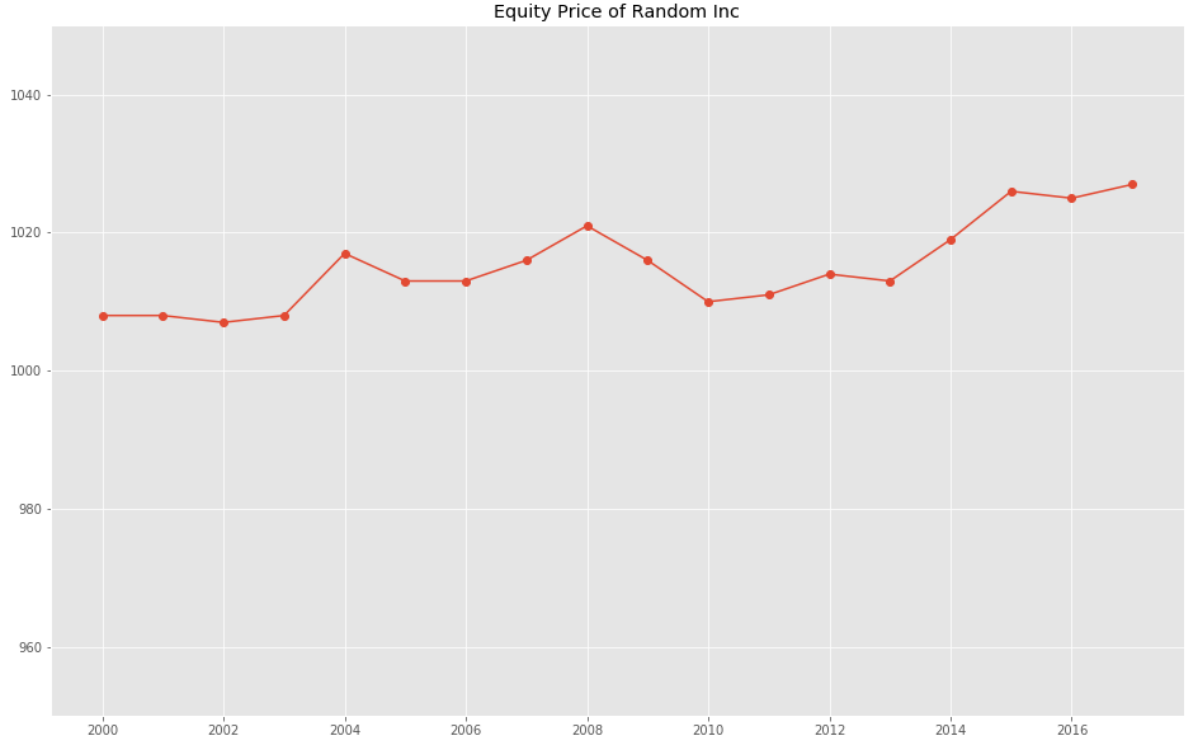




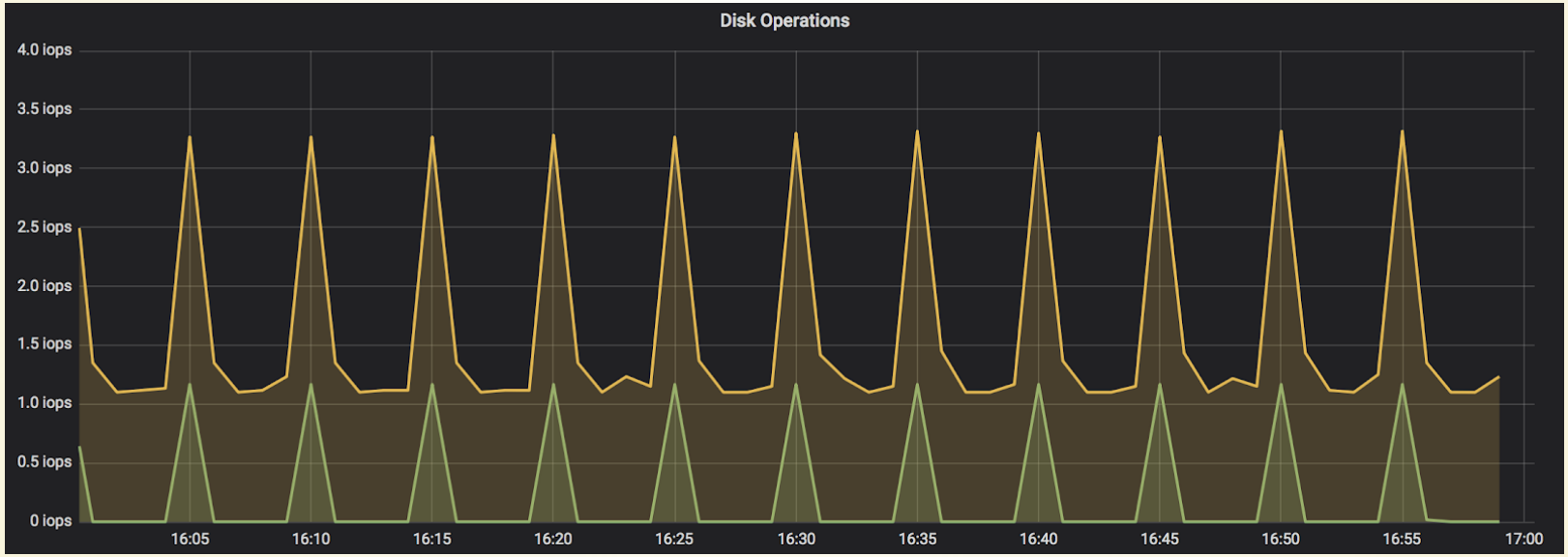
SKEWED DISTRIBUTIONS, GOING FURTHER IN ONE DIRECTION



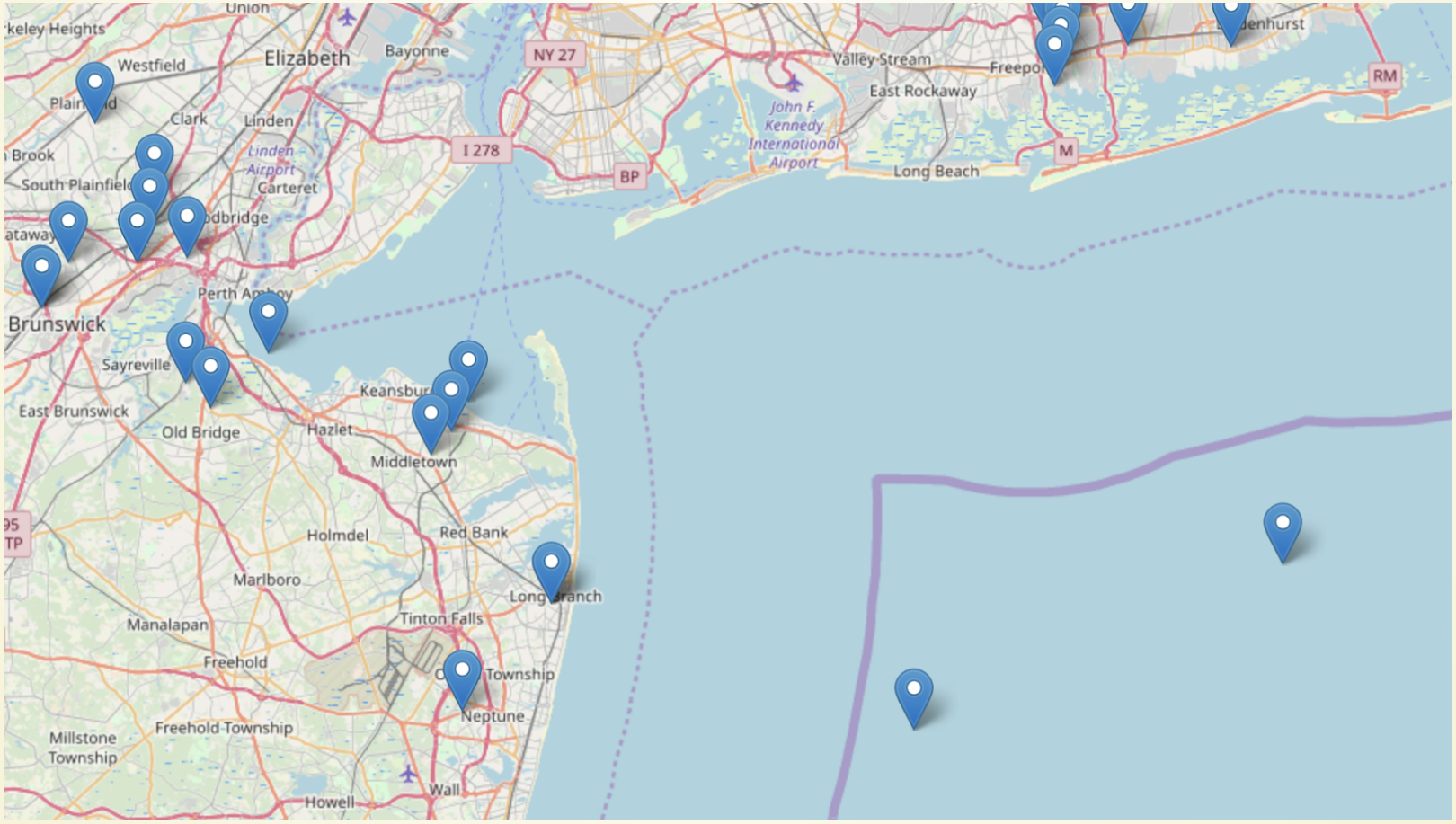
APPEARANCE OF UNEXPECTED POPULAR/UNPOPULAR VALUES



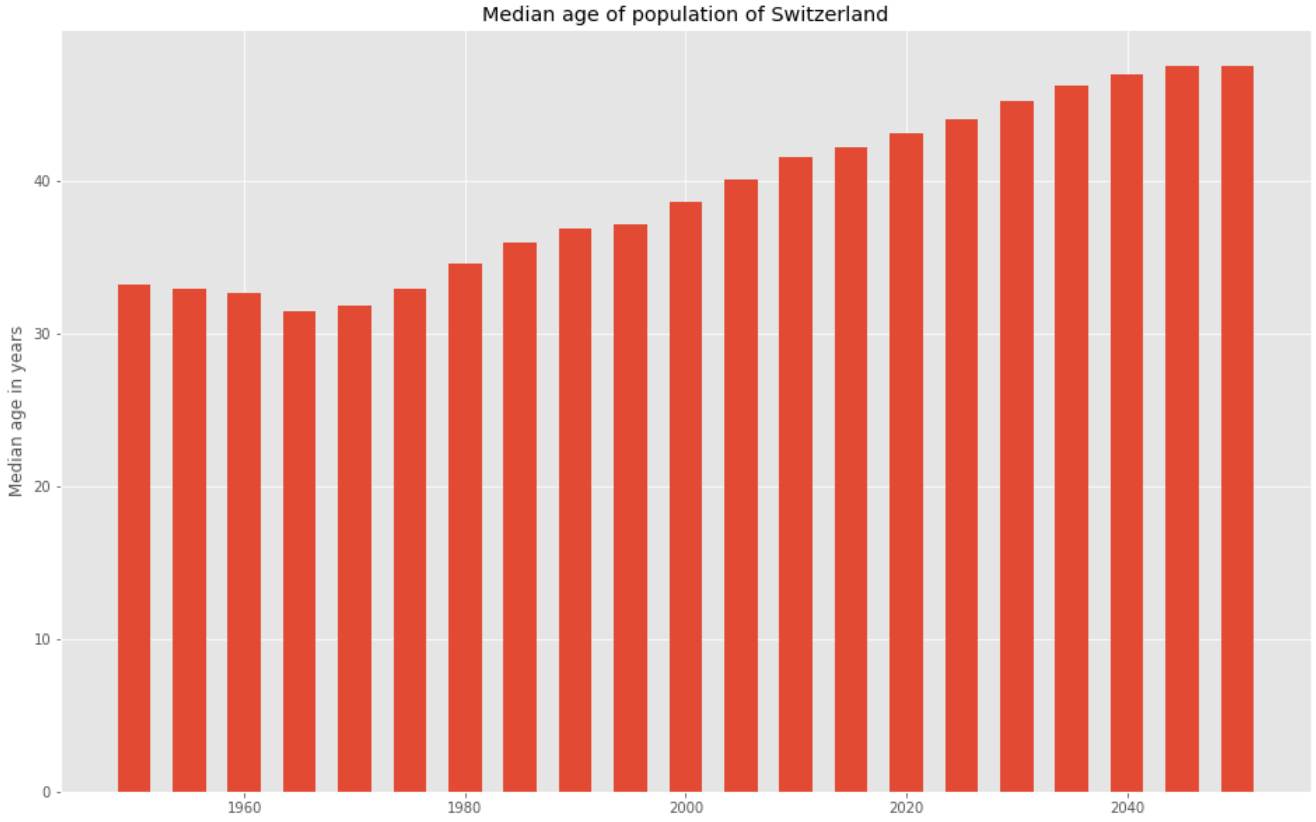
APPEARANCE OF UNEXPECTED POPULAR/UNPOPULAR VALUES



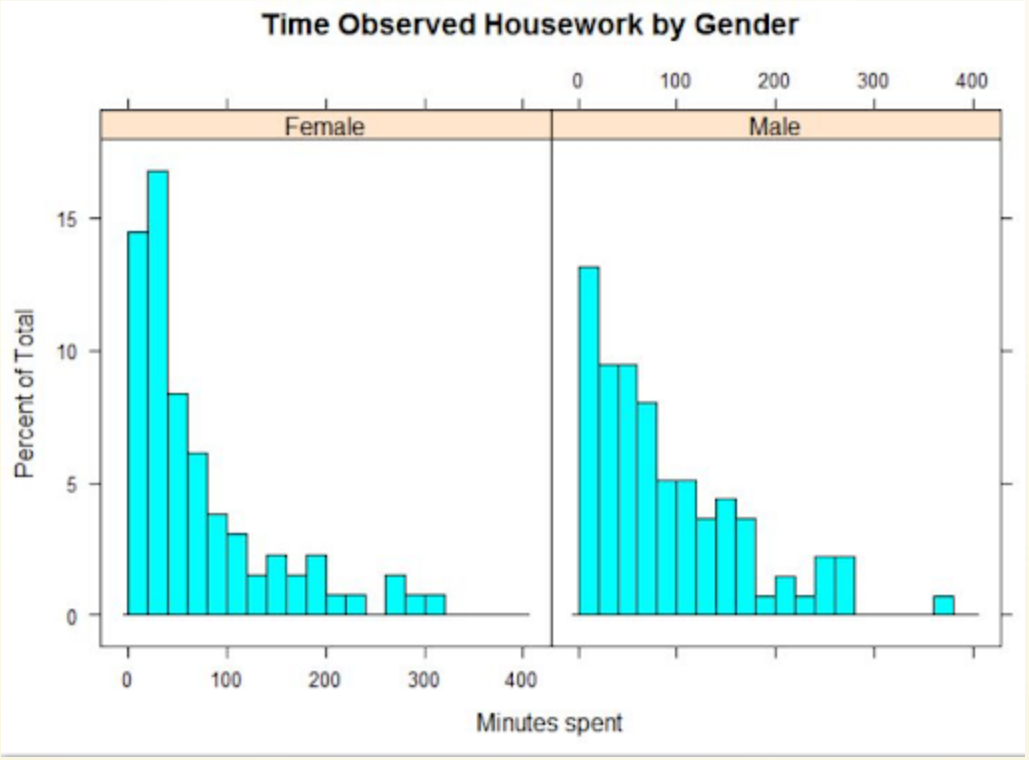
APPEARANCE OF UNEXPECTED POPULAR/UNPOPULAR VALUES



Where the values are “centered”



How widely the values are separated - domain knowledge



Data summaries

Statistics

Mean

Median

Mode

Range

Variance

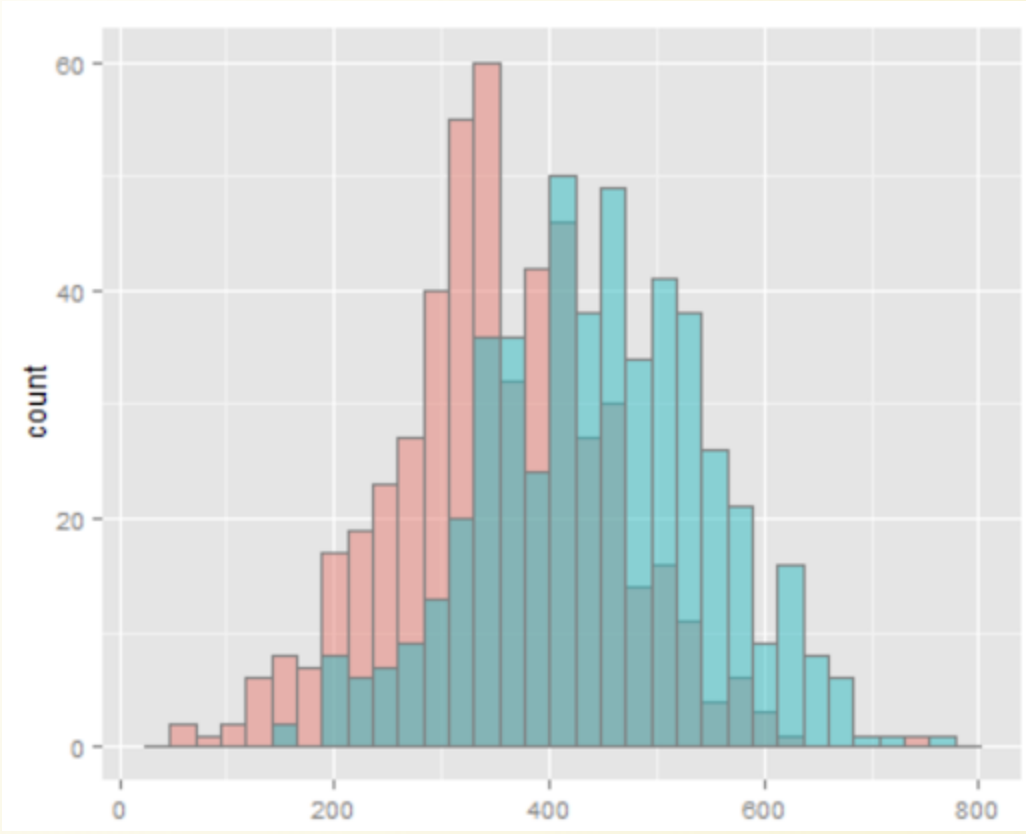
Standard deviation

Standard Error

Visualising and comparing data distributions

Histograms - plotting the frequency distribution of data

Most of the plots we've seen before

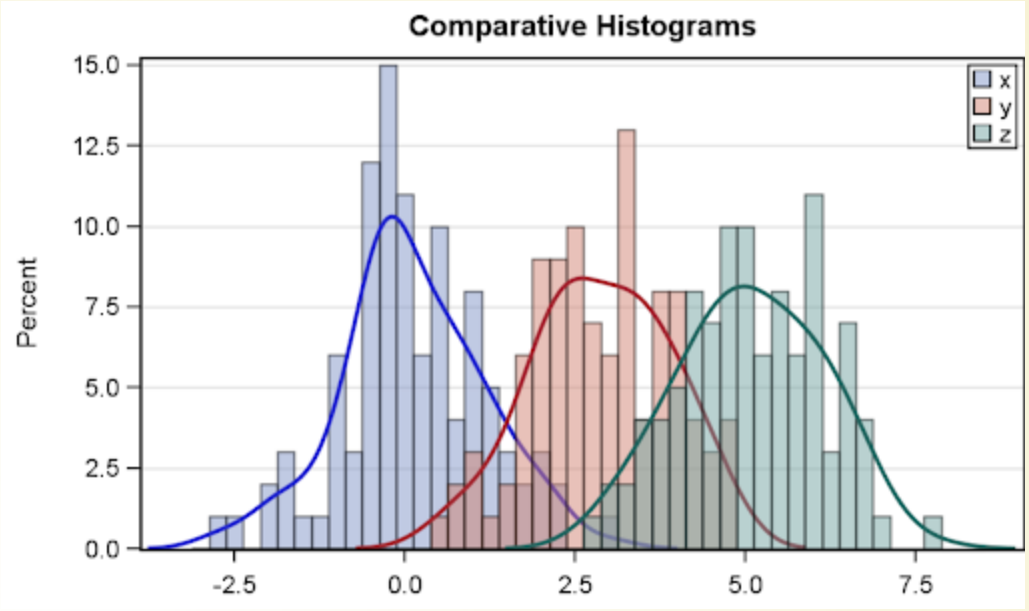


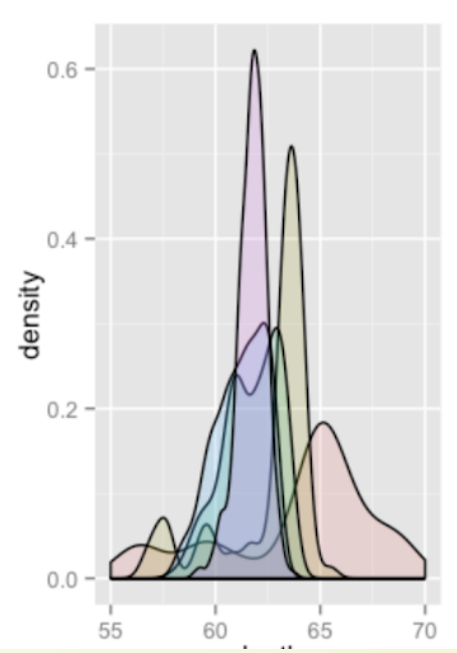
Visualising and comparing data distributions

Histograms - plotting the frequency distribution of data

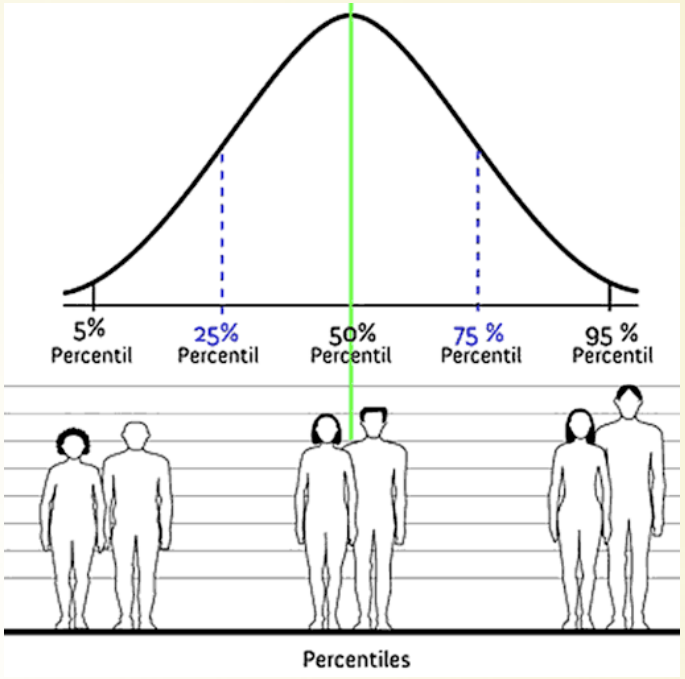
Most of the plots we've seen before

Work best for comparing maximum 3-4 groups





Data summaries



5 number summary:

extremes, median, quartiles

Ranges:

- between quartiles
- between extremes

median

1st quartile 3rd quartile

Minimum Maximum

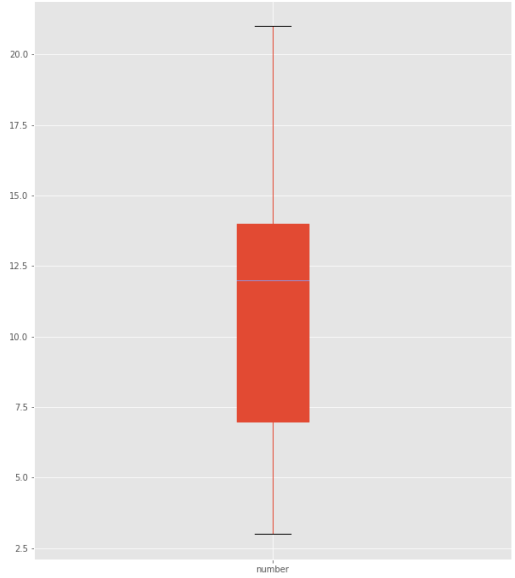
Data summaries

$$numbers = [3, 5, 7, 8, 12, 13, 14, 18, 21]$$

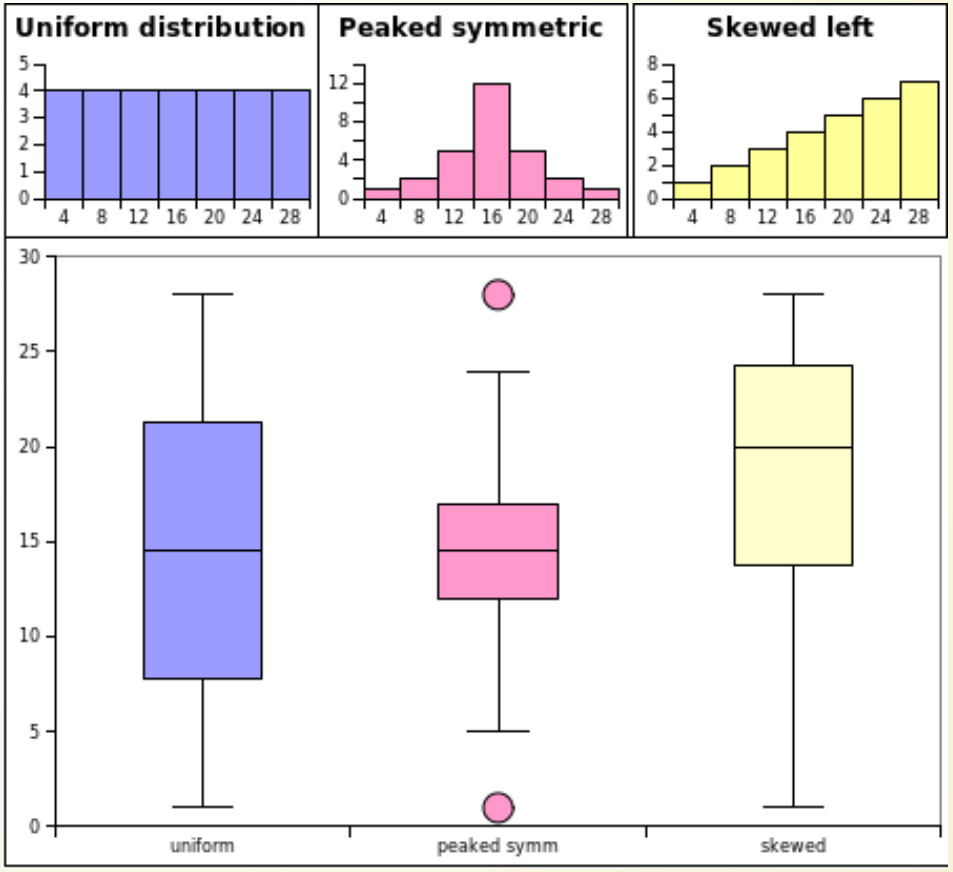
extremes: 3, 21

median: 12

quartiles: (5 + 7)/2, (14+18)/2



What does it mean?



Hands on

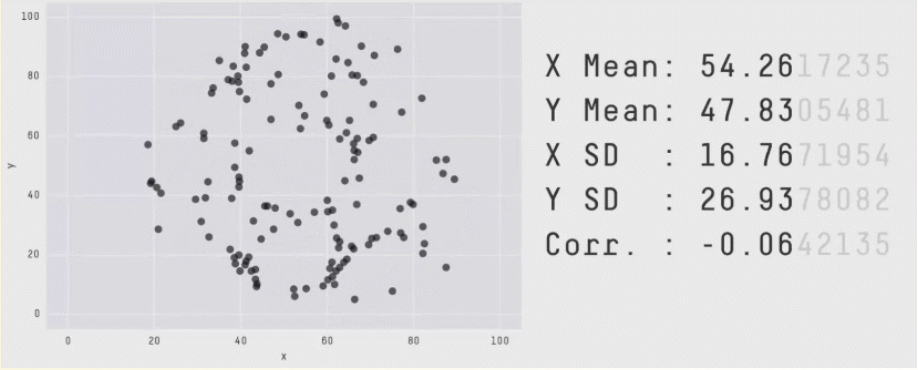
pen

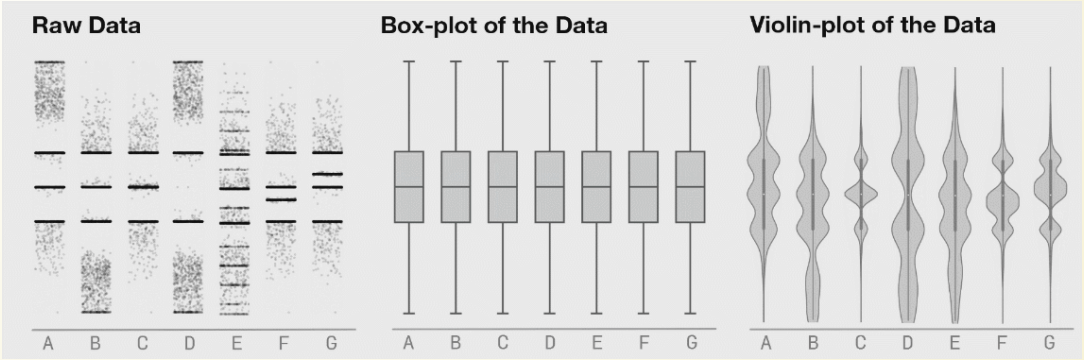
paper

dataset

Repo: github.com/terezaif/workshops_data_exploration

Summary or Details

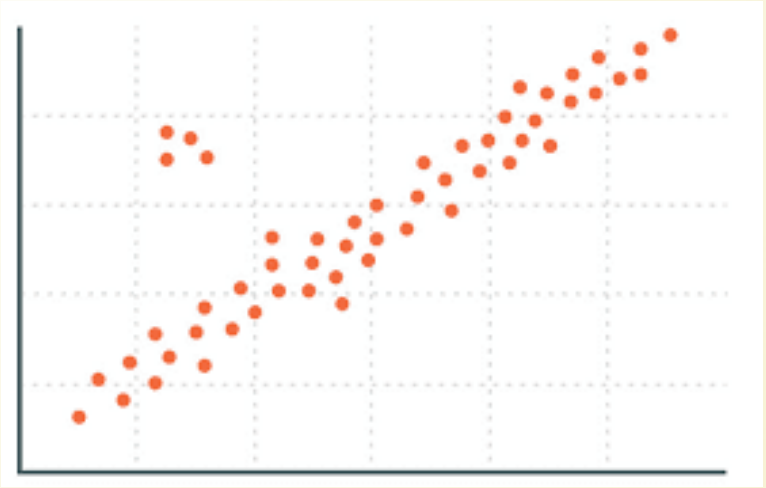




@JustinMatejka

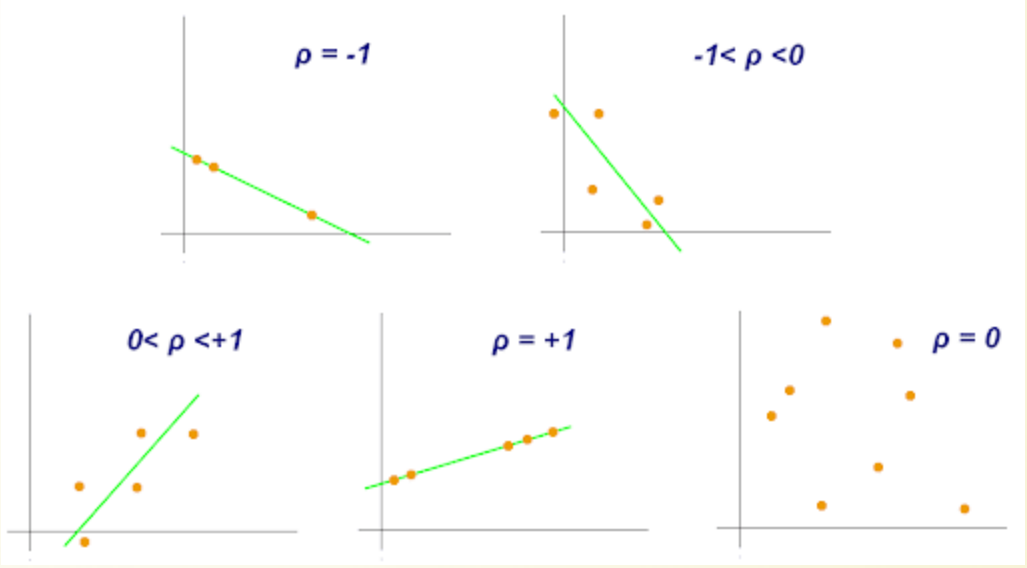
Scatter Plots

- Used to visualise relationship between two numeric variables
- Also called as correlation plot
- Can encode multiple dimensions by color, size
- It visually answers the question:
  - “How are these variables related?”
  - “When variable X grows, what happens to variable Y? With which intensity?”



Correlations - the Pearson coefficient

Summarise the linear relationship between 2 variables in just 1 number.

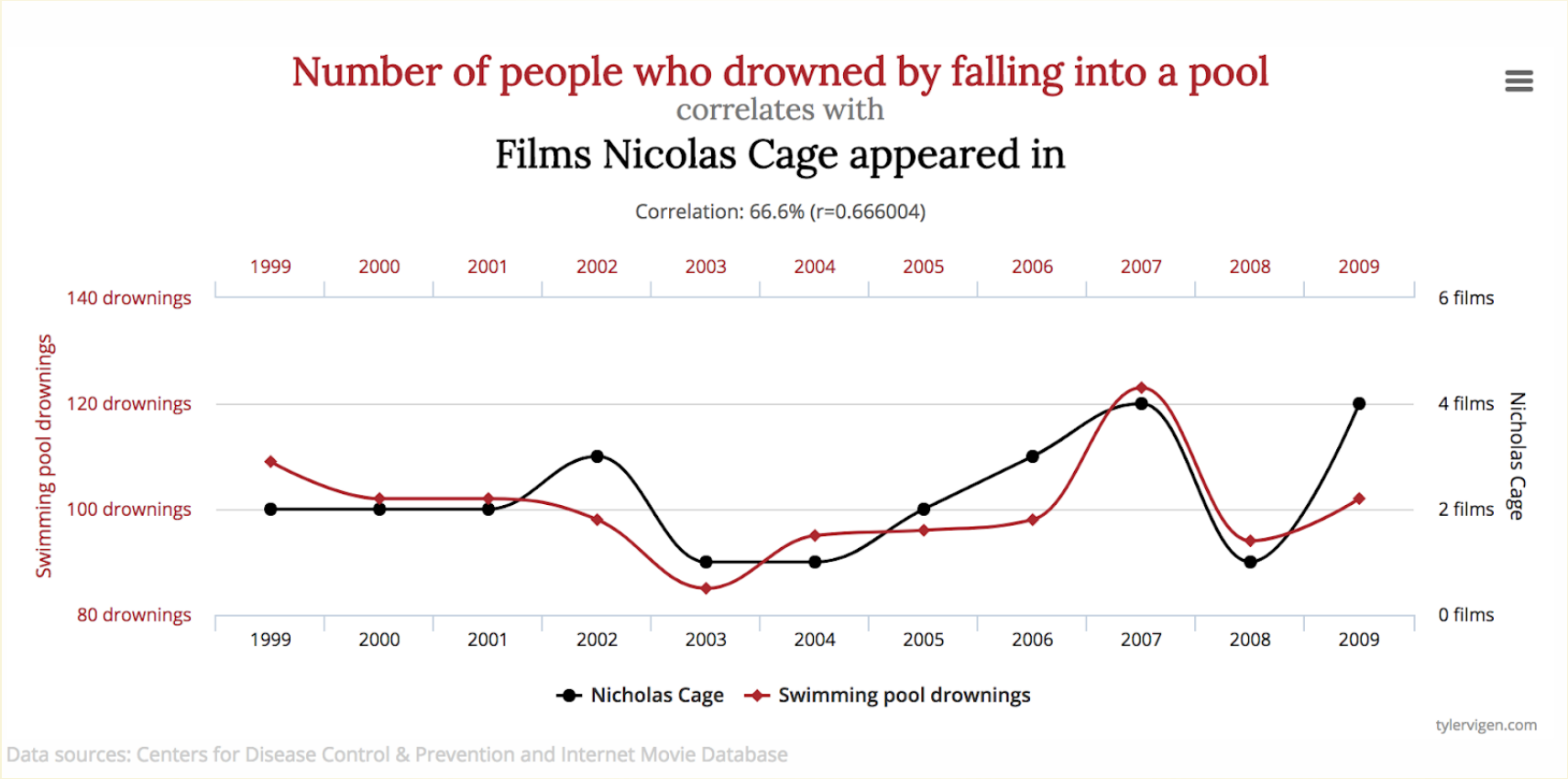


Answer questions from scatter plot in the following way:

“How strong is the relationship between two variables, *e.g.*, Temperature and Ice Cream Sold?”

Correlation != Causation

http://www.tylervigen.com/spurious-correlations



Hands on - to go

Plot the scatter plots:

Dataset 1: https://www.kaggle.com/mustafaali96/weight-height - File weight_height_15.csv

Dataset 2: Wine Dataset https://www.kaggle.com/zynicide/wine-reviews - File points_price_15.csv

Dataset 3: Iris Dataset https://archive.ics.uci.edu/ml/datasets/iris - File sepal_length_petal_length_15.csv

Dataset 4: Boston Housing Dataset https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html - File lstat_crim_15

Dataset 5: Motor Trend Car Road Tests https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html - File hp_mpg_15.csv

Next up Distributions

Things to read:

Awesome blog: http://benalexkeen.com/blog/

Exploratory Data Analysis by John W. Tukey

Dinosaur Plots :)
https://www.autodeskresearch.com/publications/samestats