CAP 5768 – Intro to Data Science

Lecture 3: Statistics – Part 1

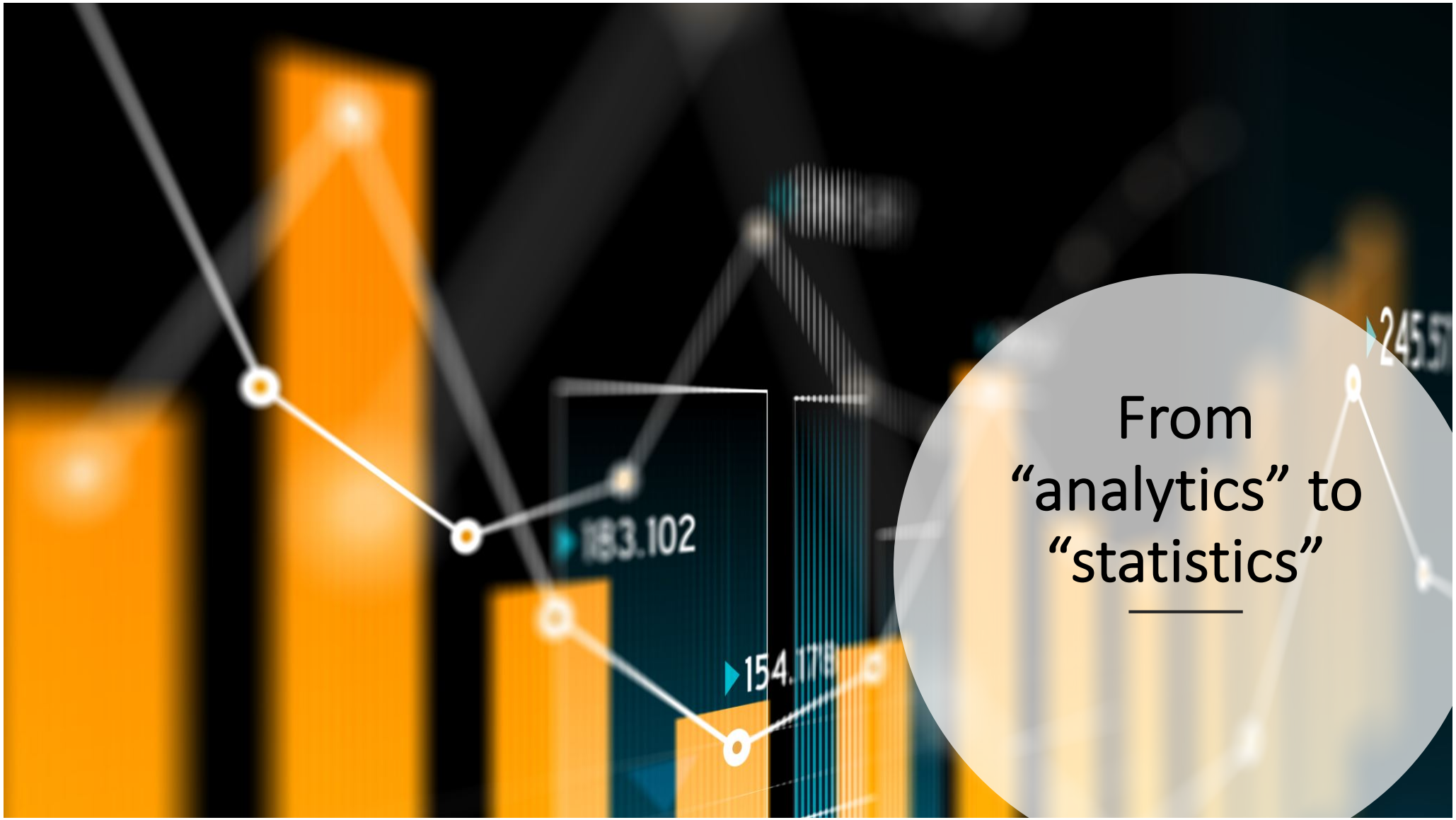
Oge Marques, PhD

Professor

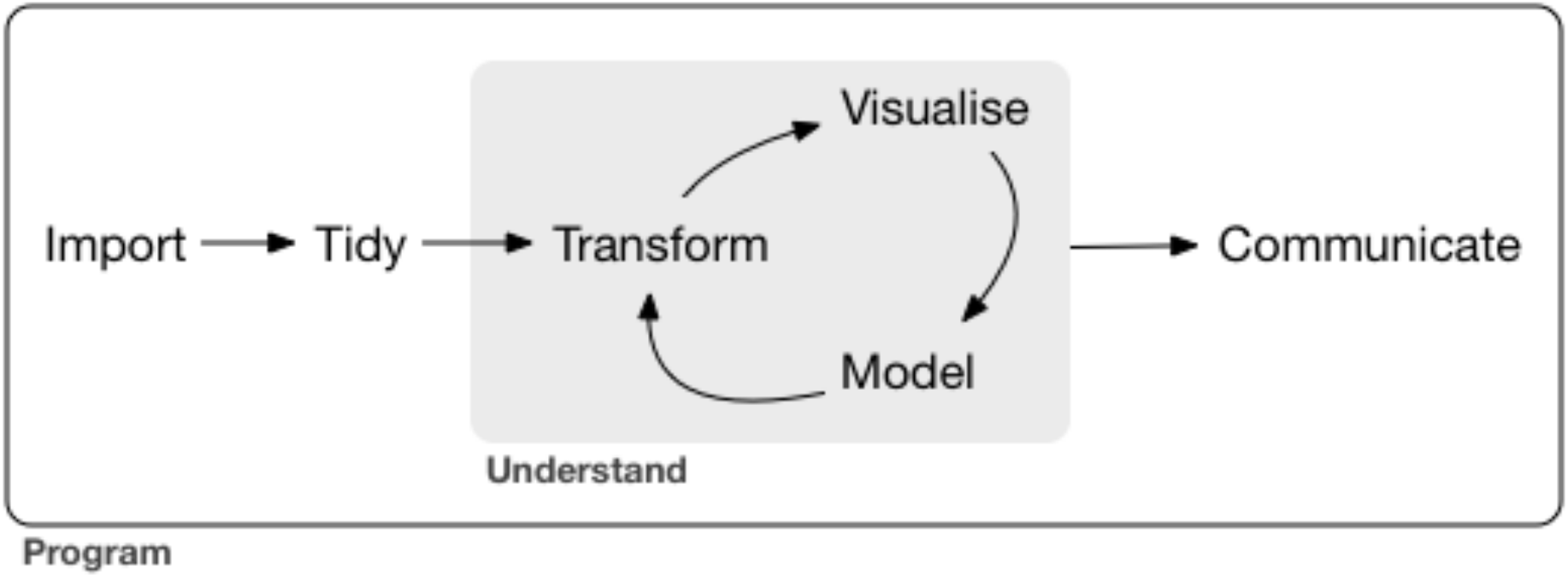
College of Engineering and Computer Science

College of Business

FAU FLORIDA ATLANTIC UNIVERSITY


From
“analytics” to
“statistics”

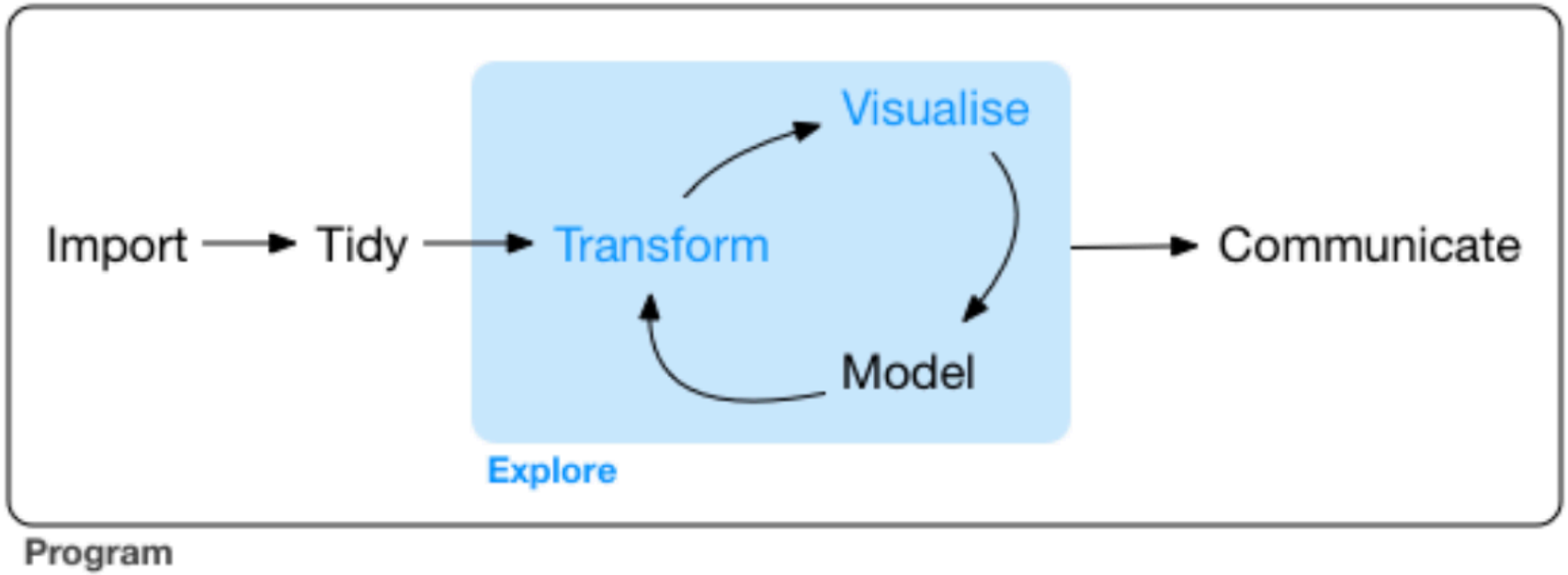
Data Science Workflow



Data Science Workflow

- **Import**: Take data stored in a file, database, or web API, and load it into a **data frame**.
- **Tidy:** Store it in a consistent form that matches the semantics of the dataset with the way it is stored. When your data is tidy, each column is a **variable**, and each row is an **observation**.
- **Transform:** Narrow in on observations of interest, create new variables that are functions of existing variables, and calculate a set of summary statistics.
  - Tidying + Transforming = Data Wrangling
- **Visualize**: Fundamental step for human analysis.
- **Model**: Build the ability to make predictions and inferences based on the data.
- **Communicate**: Share results, get feedback, promote reproducibility.
- **Programming**: The “glue” that keeps everything together.

Data Exploration



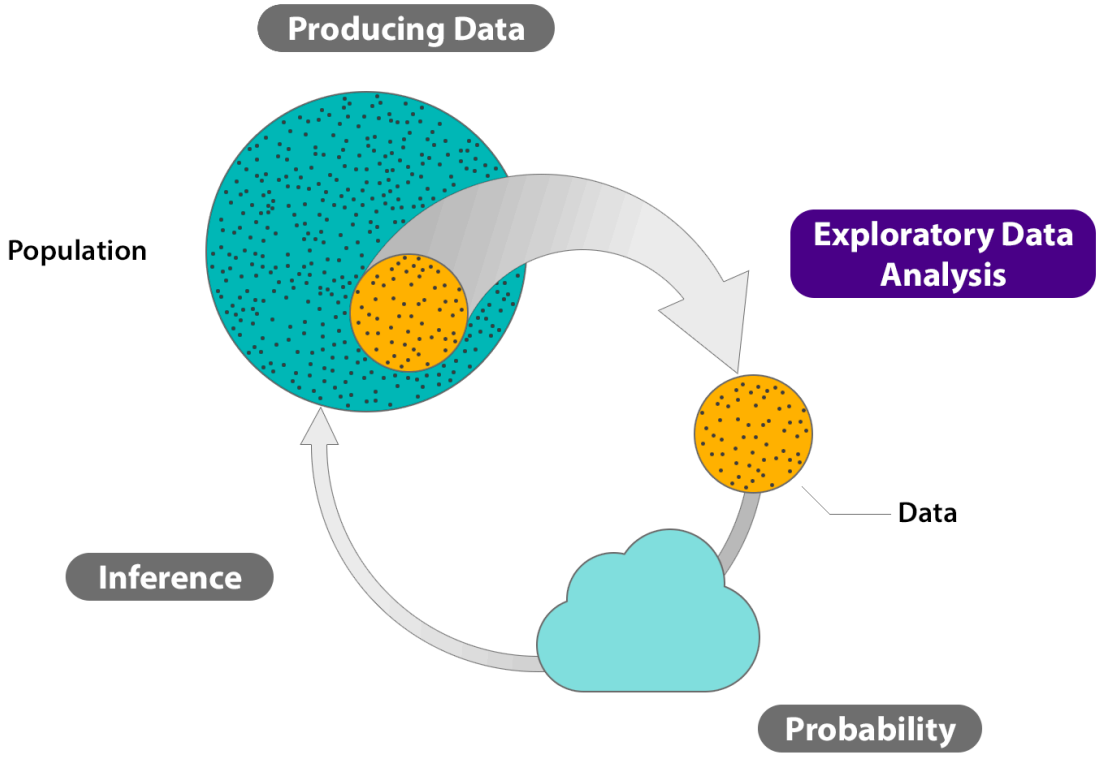
Statistical Thinking

Cassie Kozyrkov
Chief Decision Scientist, Google
@quaesita

Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.

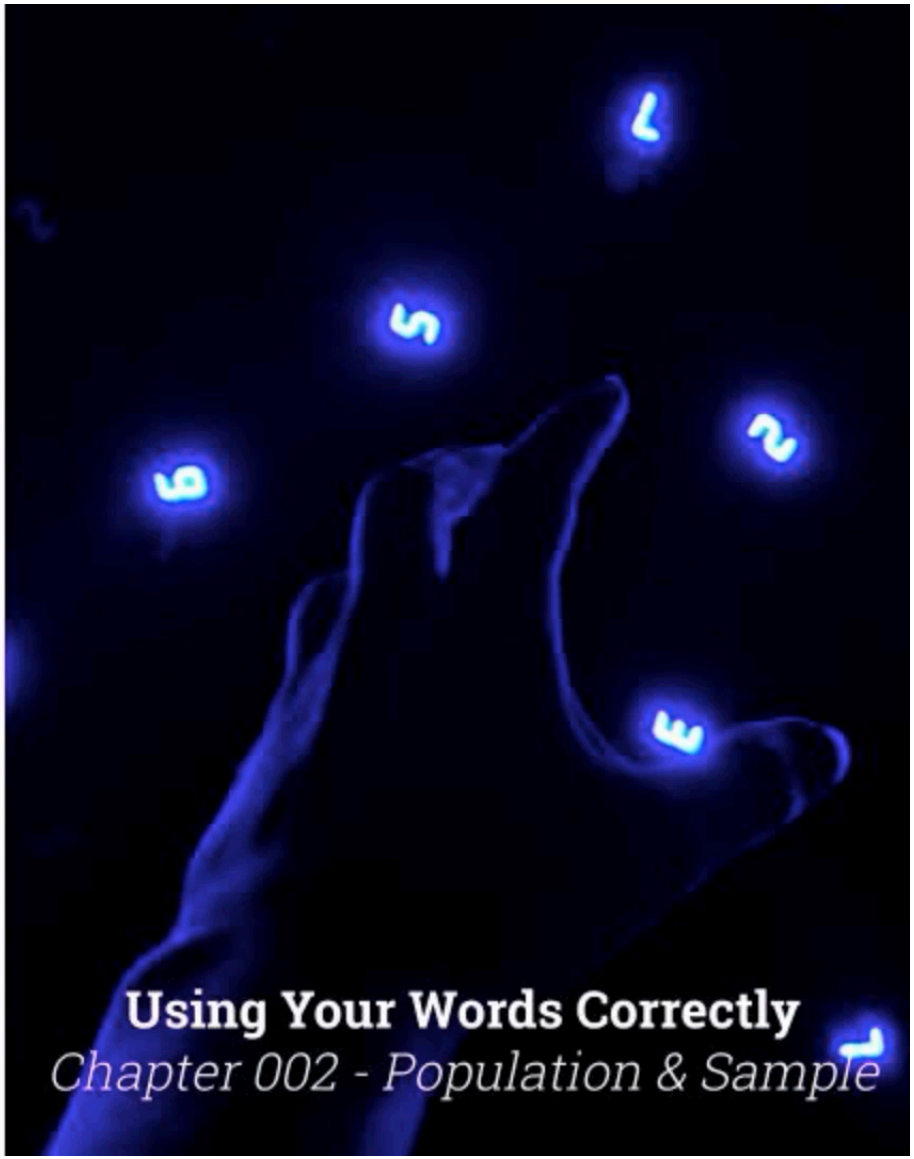
—John Tukey

Exploratory Data Analysis (EDA) in context



Statistical Thinking

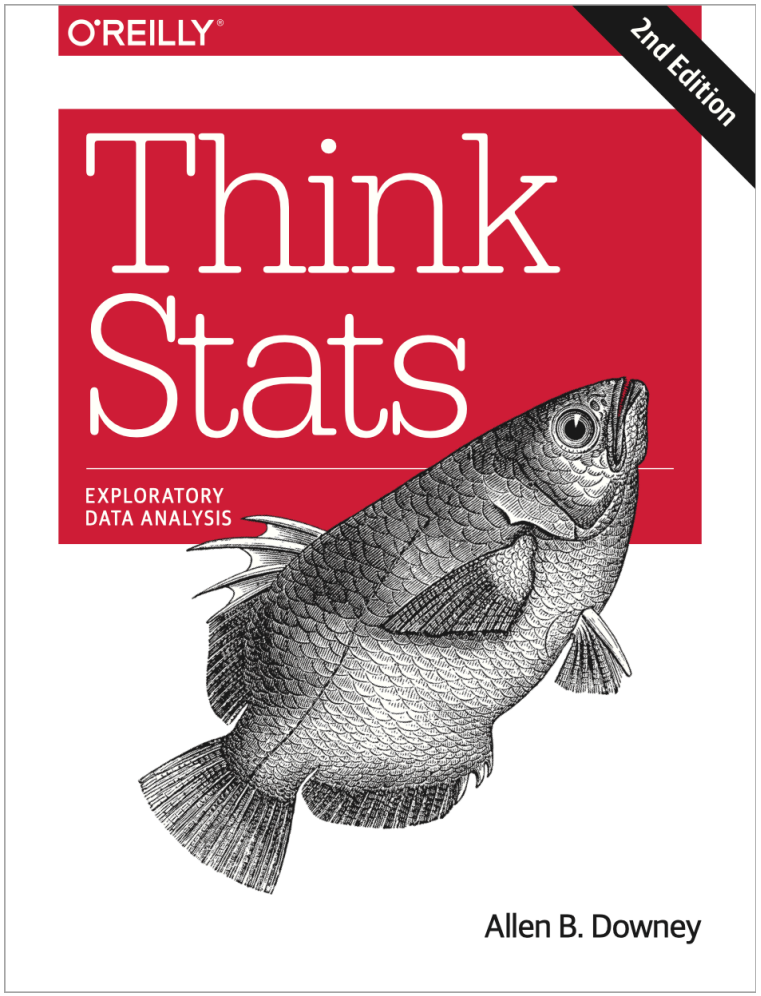
Cassie Kozyrkov
Chief Decision Scientist, Google
@quaesita



Highlights from

“Think Stats 2/e”

by Allen B. Downey



Exploratory data analysis

- […] “The process I use when I start working with a dataset”:

  - **Importing and cleaning** the data
  - **Single variable explorations**: start by examining one variable at a time, finding out what the variables mean, looking at distributions of the values, and choosing appropriate summary statistics.
  - **Pair-wise explorations**: to identify possible relationships between variables, look at tables and scatter plots, and compute correlations and linear fits.
  - **Multivariate analysis**: if there are apparent relationships between variables, use multiple regression to add control variables and investigate more complex relationships.
  - **Estimation and hypothesis testing**: When reporting statistical results, it is important to answer three questions: How big is the effect? How much variability should we expect if we run the same measurement again? Is it possible that the apparent effect is due to chance?
  - **Visualization**: During exploration, visualization is an important tool for finding possible relationships and effects. Then if an apparent effect holds up to scrutiny, visualization is an effective way to communicate results.

Chapter 1: Exploratory Data Analysis

Anecdotal evidence vs. statistical approach

National Survey of Family Growth (NSFG)

Data frames and variables

Data cleaning

Validation

Interpretation

Chapter 1 Highlights

- The thesis of this book: “[…] data combined with practical methods can answer questions and guide decisions under uncertainty.”
- Reasons why anecdotal evidence usually fails
- The tools of statistical approach to data analysis
  - Data collection
  - Descriptive statistics
  - Exploratory Data Analysis (EDA)
  - Estimation
  - Hypothesis Testing

Chapter 1 Highlights

- NSFG Cycle 6 study
  - Goals
  - Population and sample
  - Python code for importing data
  - Python code for EDA using Pandas DataFrames
  - Relevant variables
  - Data cleaning
  - Validation (against the NSFG codebook)
  - Interpretation

Chapter 2: Distributions

Histograms

Outliers

Summary statistics

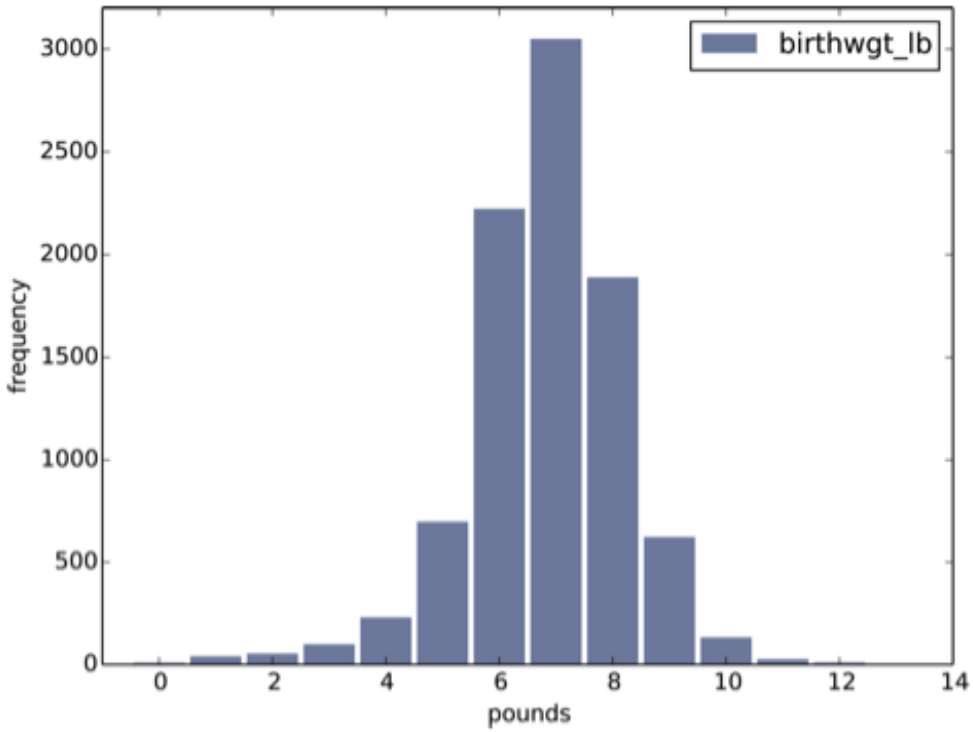
Variance

Cohen’s effect size

Reporting results

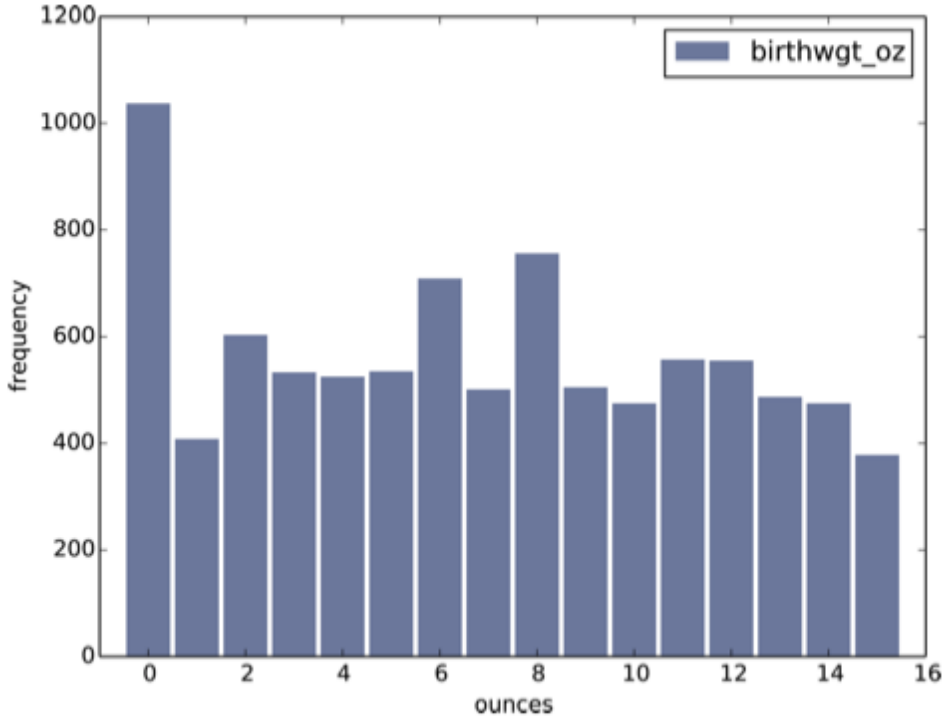
Chapter 2 Highlights

Histograms



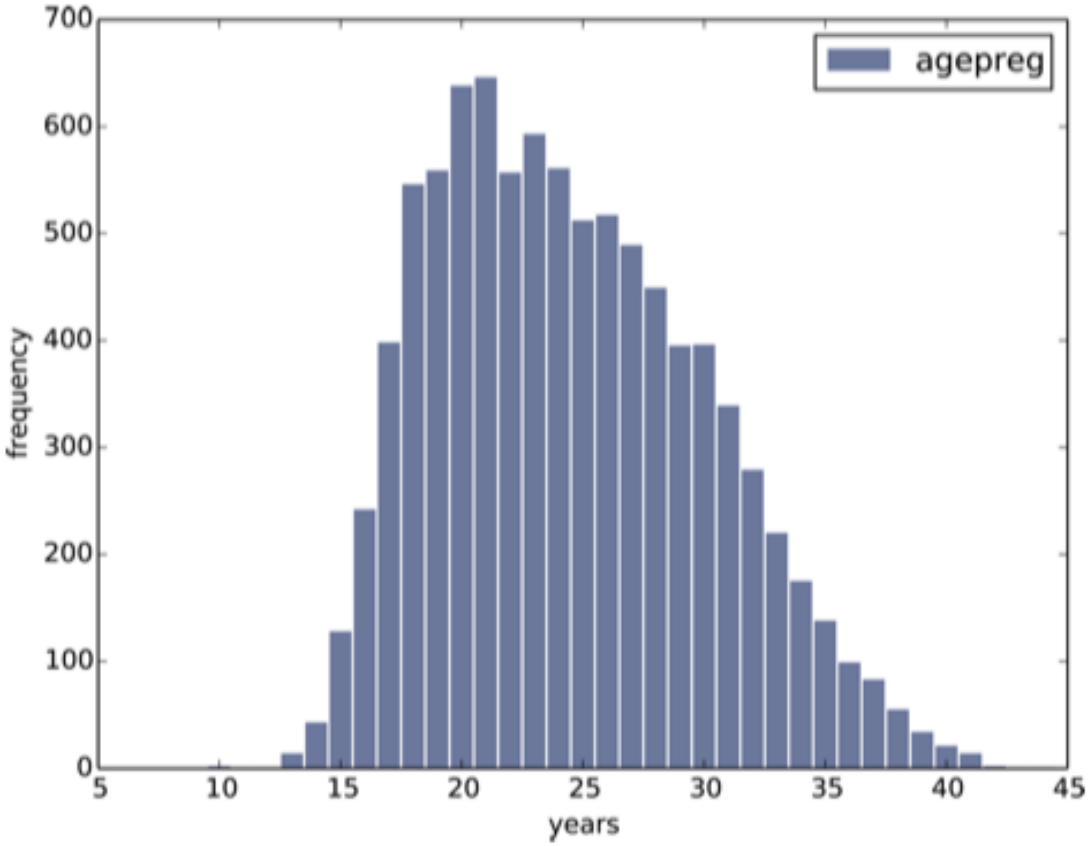
Chapter 2 Highlights

Histograms



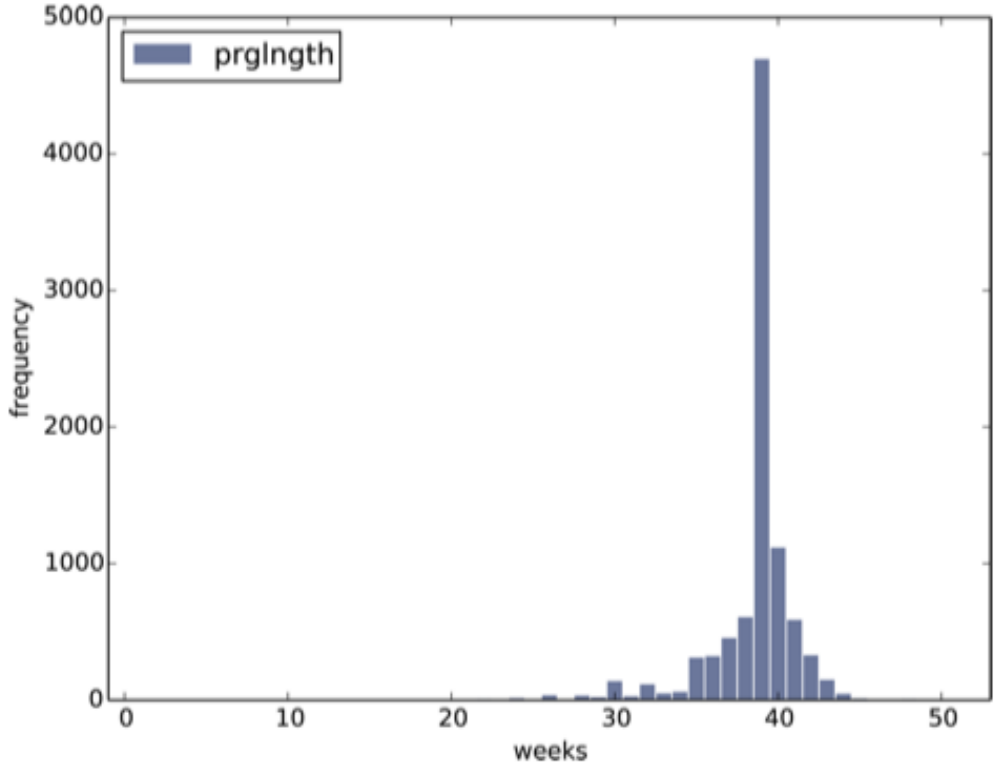
Chapter 2 Highlights

Histograms



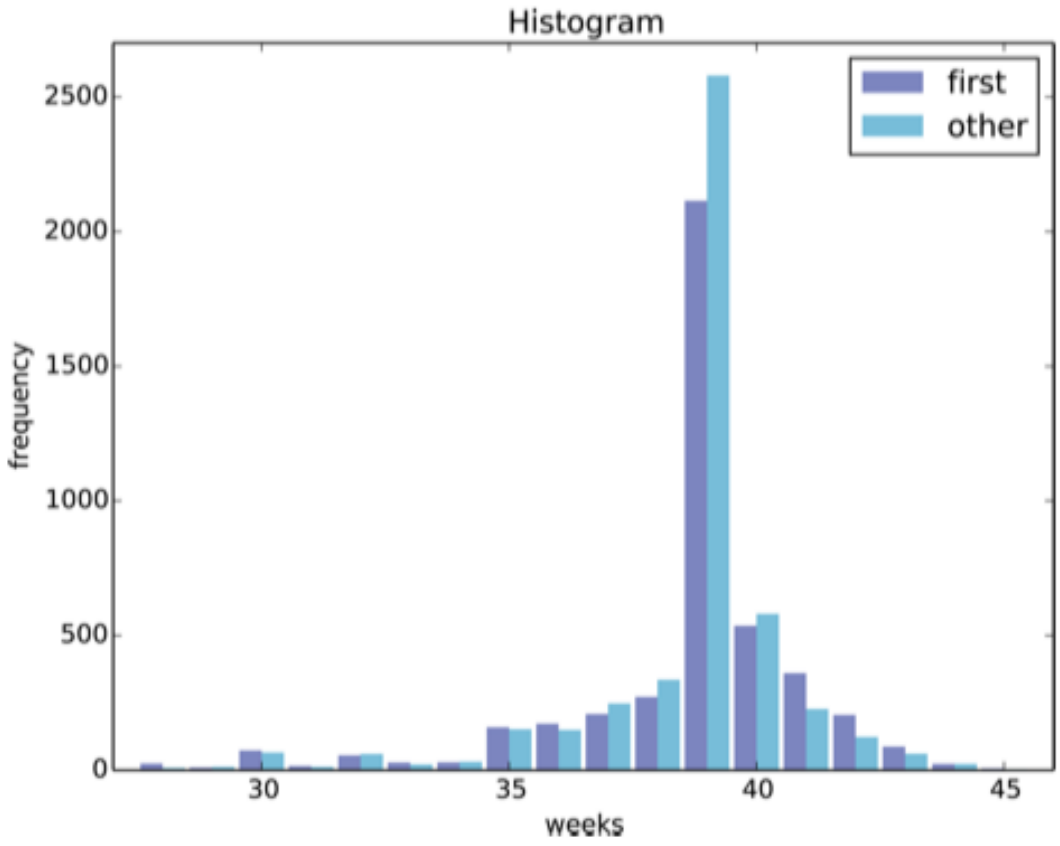
Chapter 2 Highlights

Histograms



Chapter 2 Highlights

Histograms



Chapter 2 Highlights

- **Summary statistics**
  - central tendency
    - Do the values tend to cluster around a particular point?
  - modes
    - Is there more than one cluster?
  - spread
    - How much variability is there in the values?
  - tails
    - How quickly do the probabilities drop off as we move away from the modes?
  - outliers
    - Are there extreme values far from the modes?

Chapter 3: Probability Mass Functions (PMFs)

Normalization

The class size paradox

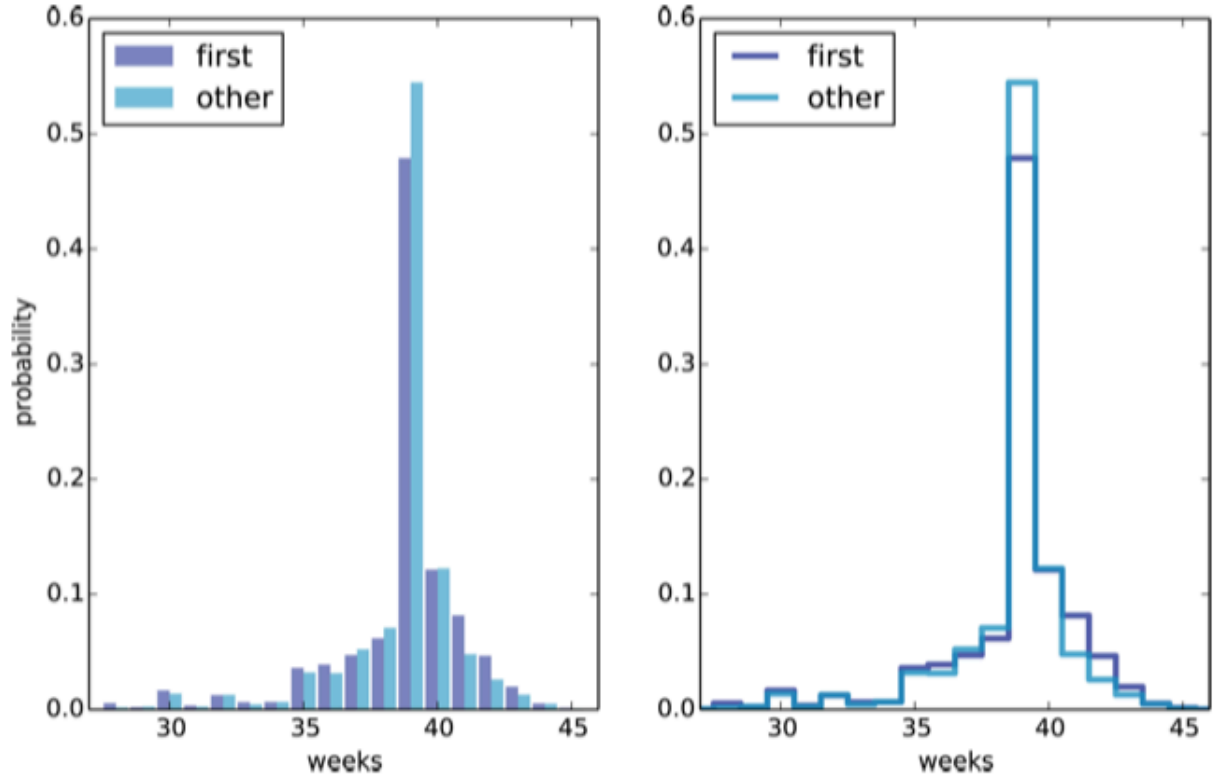


Figure 3-1. PMF of pregnancy lengths for first babies and others, using bar graphs and step functions

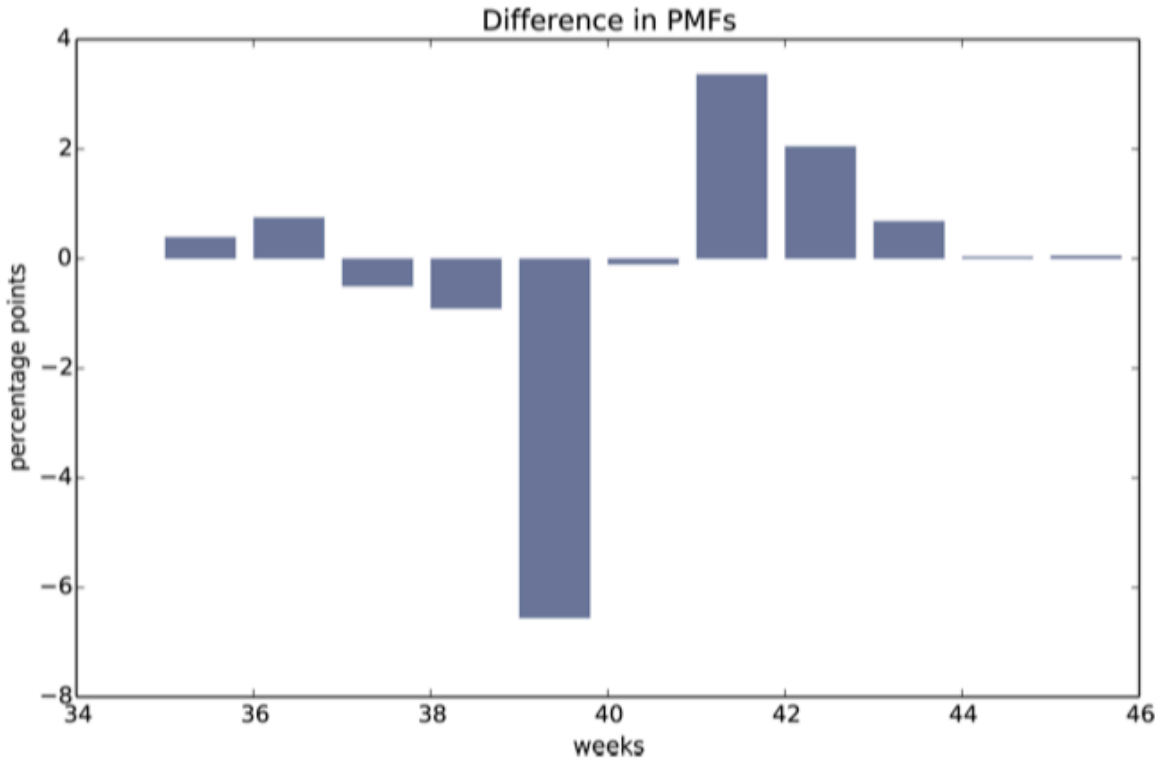


Figure 3-2. Difference, in percentage points, by week

The class size paradox

- Easy to understand
  - See http://www.greenteapress.com/thinkbayes/html/thinkbayes009.html for more

- But somewhat confusing implementation / explanation
  - See https://stats.stackexchange.com/questions/303411/biased-distribution-described-in-think-stats

Chapter 4: Cumulative Distribution Functions (CDFs)

Percentiles and percentile ranks

CDFs: representations

Comparing CDFs

Percentile-based statistics

Interquartile range (IQR)

Percentiles and percentile ranks

```
def PercentileRank(scores, your_score):
    count = 0
    for score in scores:
        if score <= your_score:
            count += 1

    percentile_rank = 100.0 * count / len(scores)
    return percentile_rank
```

Percentiles and percentile ranks

```
def Percentile(scores, percentile_rank):
    scores.sort()
    for score in scores:
        if PercentileRank(scores, score) >= percentile_rank:
            return score
```

Percentiles and percentile ranks

PercentileRank takes a value and computes its percentile rank in a set of values

Percentile takes a percentile rank and computes the corresponding value.

Cumulative Distribution Function (CDF)

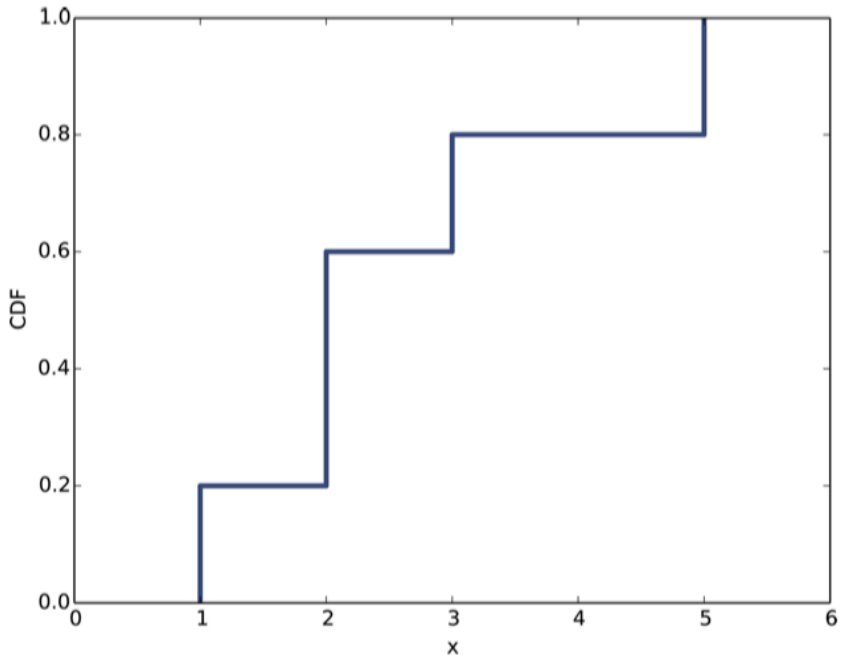
- The function that maps a value to its percentile rank

```
def EvalCdf(t, x):
    count = 0.0
    for value in t:
        if value <= x:
            count += 1

    prob = count / len(t)
    return prob
```

$$CDF(0) = 0$$
$$CDF(1) = 0.2$$
$$CDF(2) = 0.6$$
$$CDF(3) = 0.8$$
$$CDF(4) = 0.8$$
$$CDF(5) = 1$$

[1, 2, 2, 3, 5]

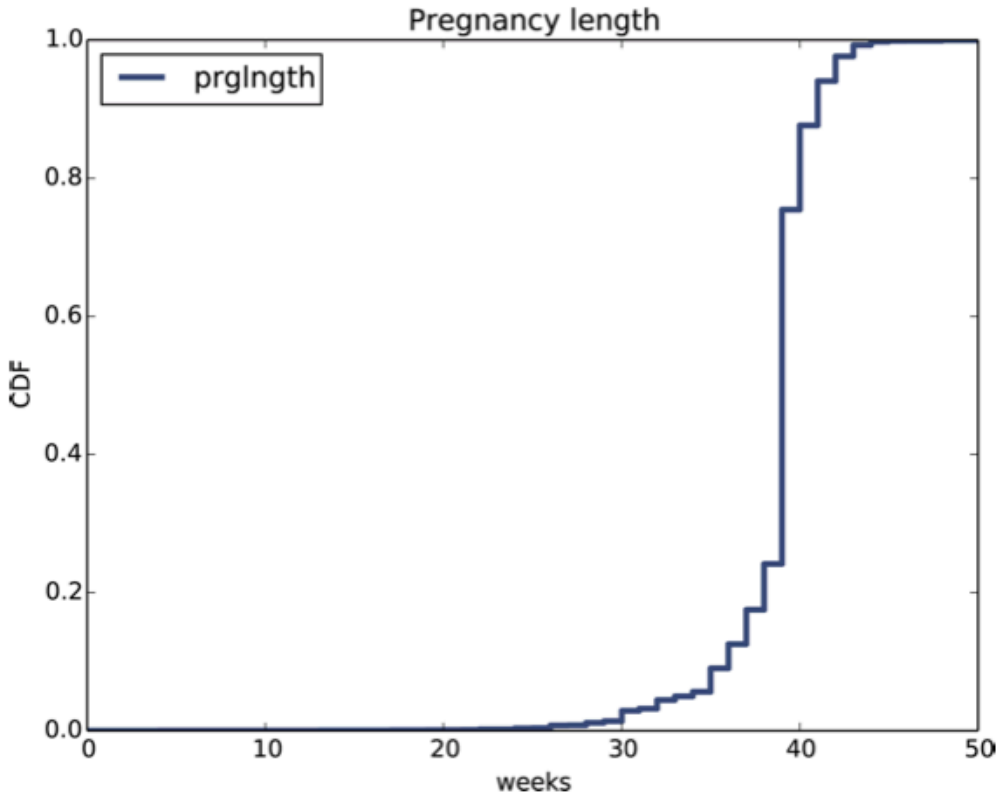


Example: CDF of pregnancy length

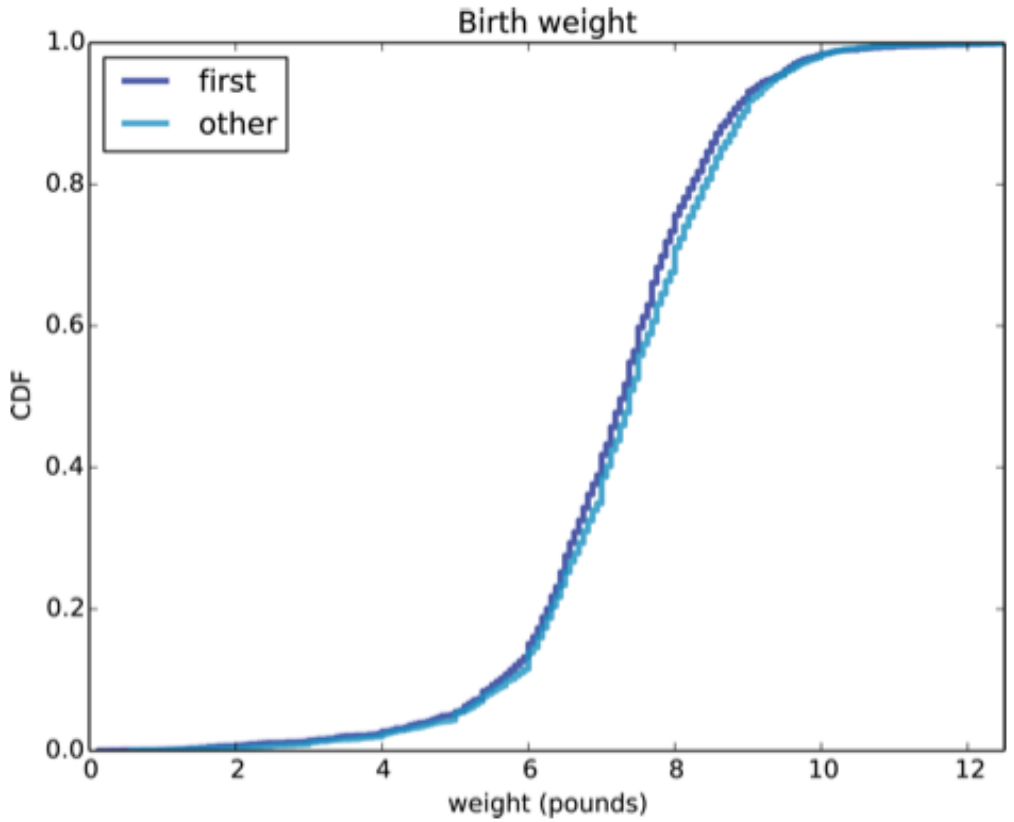
~10% of pregnancies are shorter than 36 weeks.

~90% are shorter than 41 weeks.

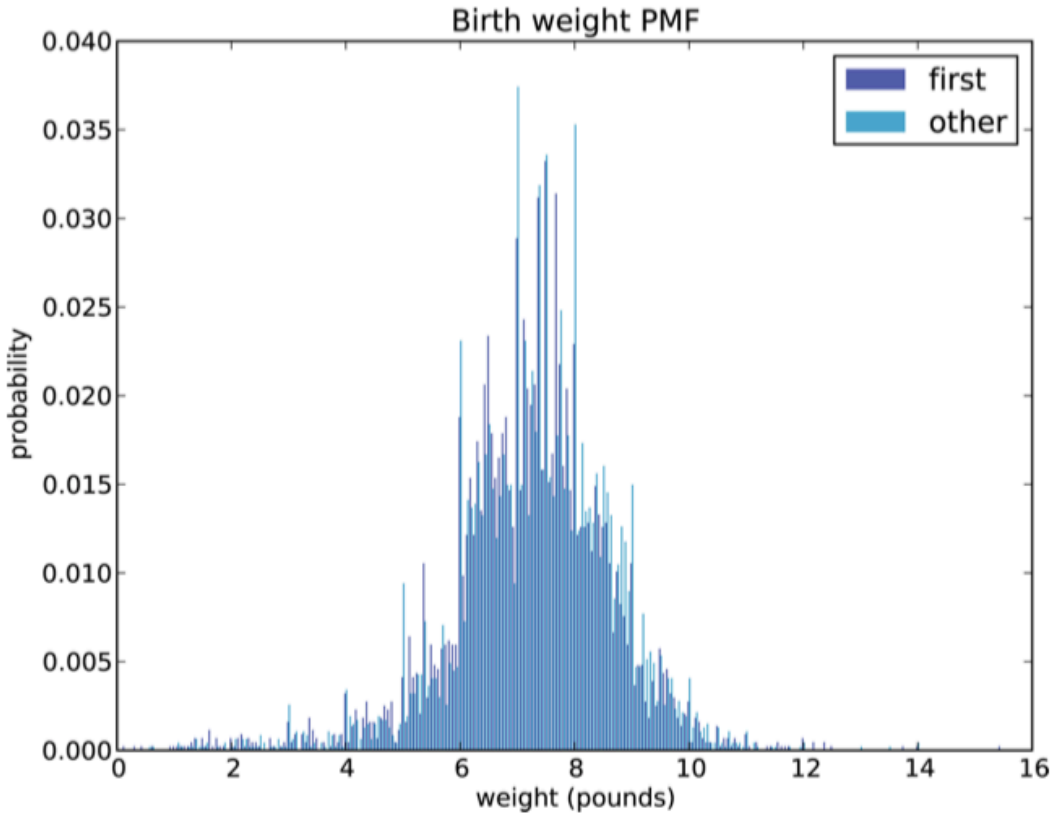
Common values appear as steep or vertical sections of the CDF.



Comparing CDFs



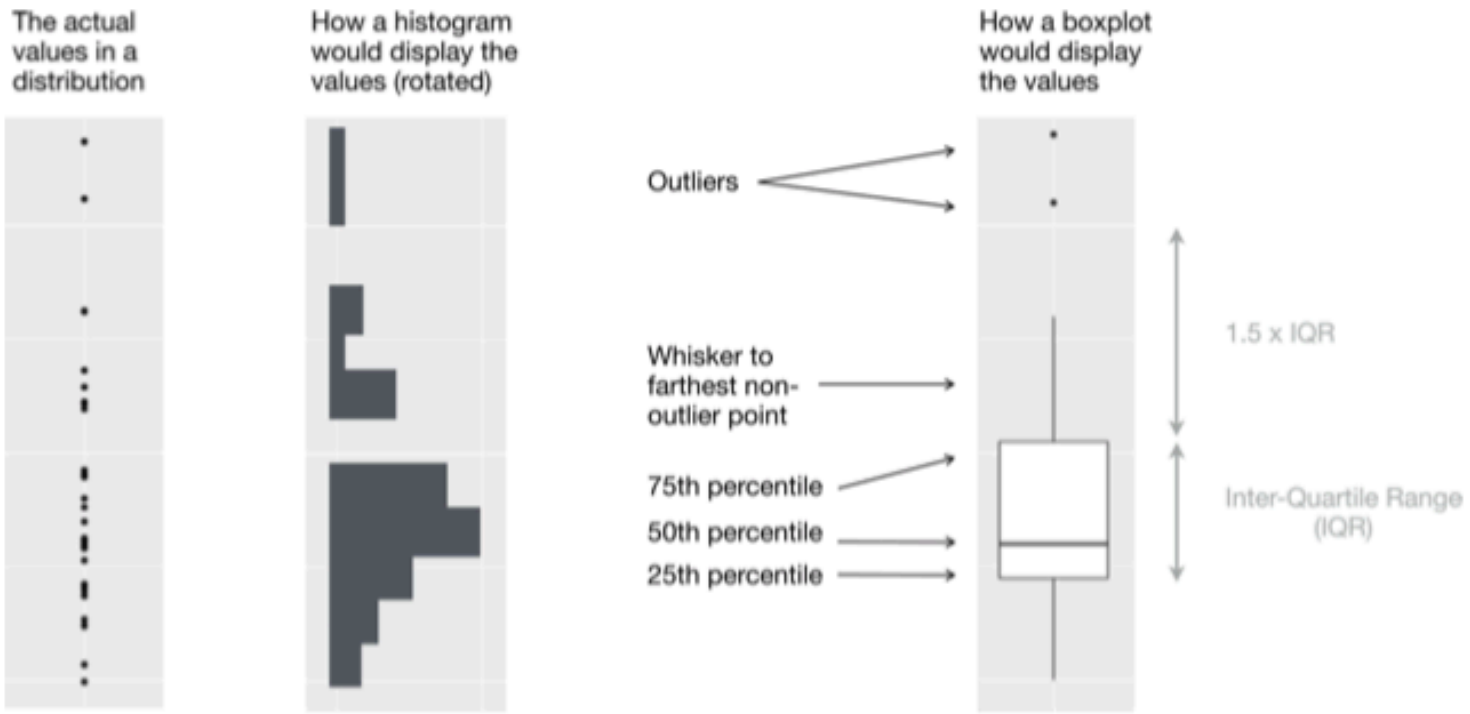
Comparing histograms



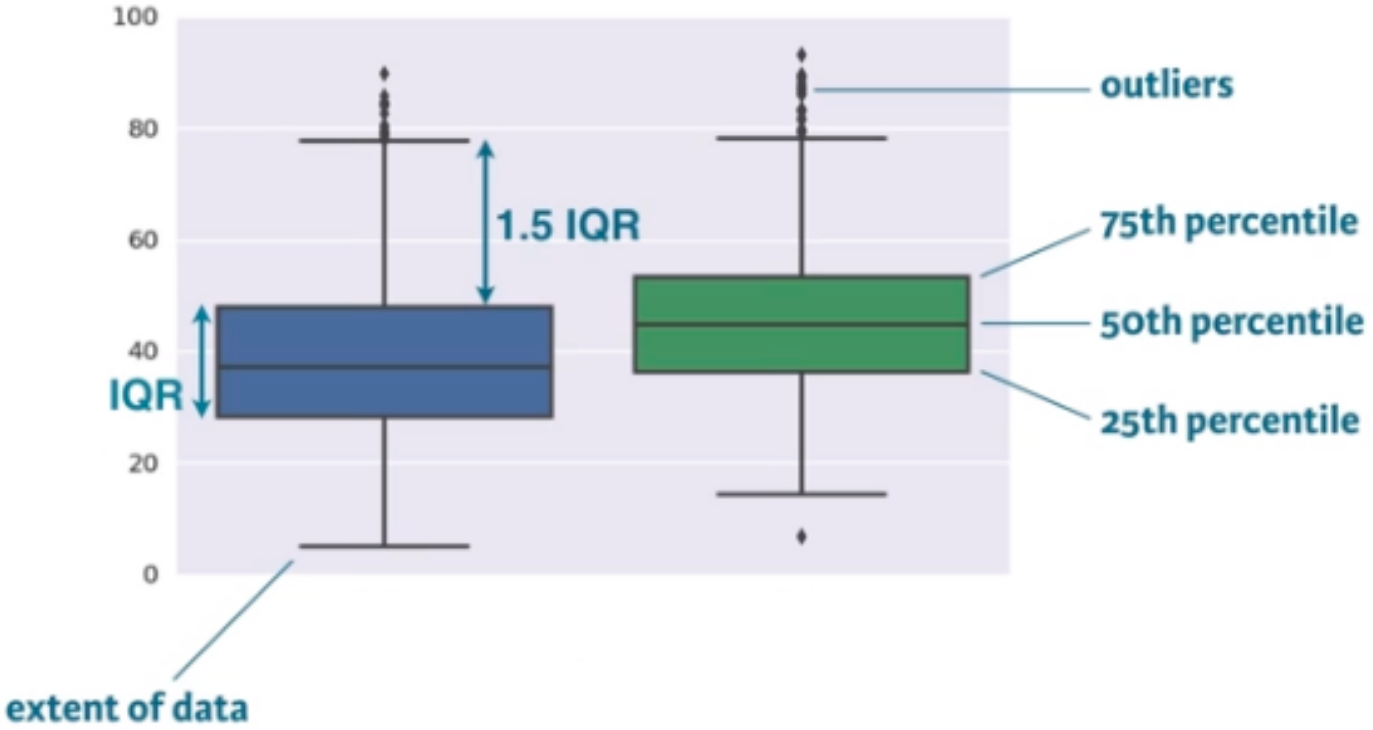
Percentile-based statistics

- Median = 50th percentile
  - A measure of central tendency of the distribution
- Interquartile Range (IQR) = the difference between the 75th and 25th percentiles
  - A measure of the spread of the distribution

Box(-and-whisker) plots



Box(-and-whisker) plots

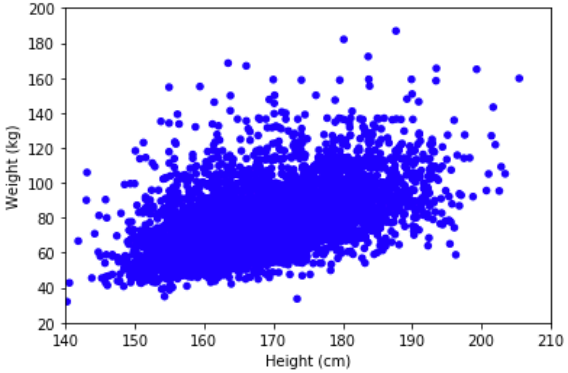


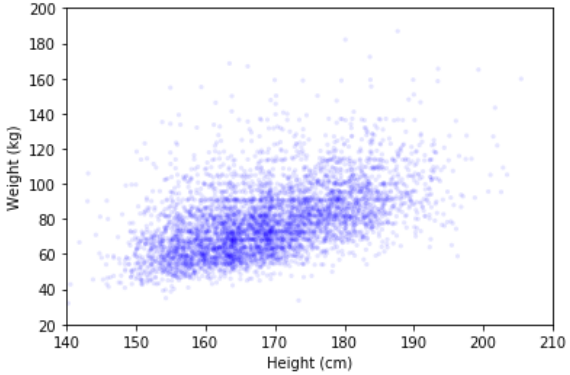
Chapter 7: Relationships between variables

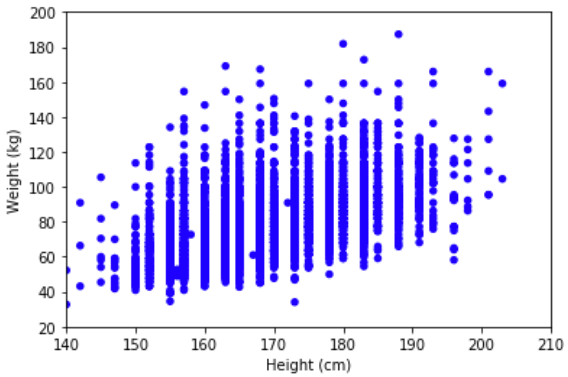
- Scatter plots
- Correlation
- Covariance
- Pearson’s correlation
- Nonlinear relationships
- Spearman’s rank correlation
- Correlation and causation

Dataset

- The National Center for Chronic Disease Prevention and Health Promotion conducts an annual survey as part of the Behavioral Risk Factor Surveillance System (BRFSS).
- In 2008, they interviewed 414,509 respondents and asked about their demographics, health, and health risks.
- Among the data they collected are the weights in kilograms of 398,484 respondents.
- The repository for the book contains **CDBRFS08.ASC.gz**, a fixed-width ASCII file that contains data from the BRFSS, and **brfss.py**, which reads the file and analyzes the data.



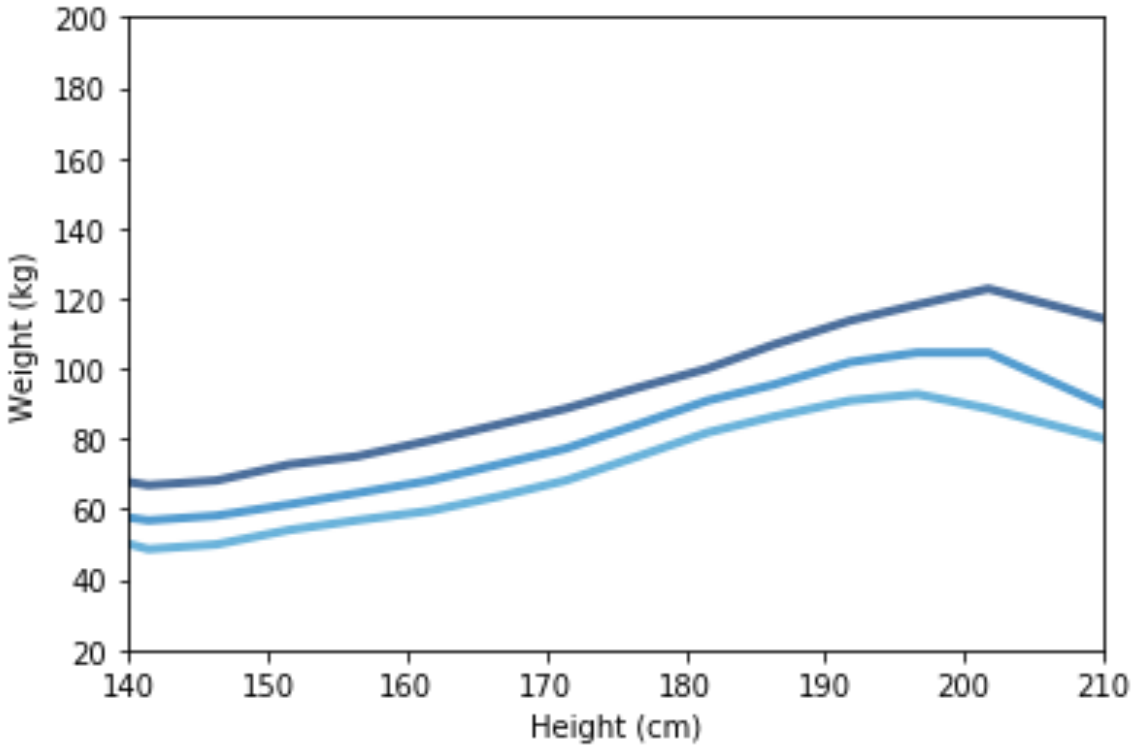




Scatter plots: weight vs. height

Plotting percentiles

- Sometimes a better way to get a sense of the relationship between variables is to divide the dataset into groups using one variable, and then plot percentiles of the other variable.



Correlation and covariance

Correlation is a statistic intended to quantify the strength of the relationship between two variables.

Covariance is a measure of the tendency of two variables to vary together.

Correlation: Pearson and Spearman

- Challenges:
  1. the variables we want to compare are often not expressed in the same units
  2. even if they are in the same units, they come from different distributions.
- Solutions:
  1. Transform each value to a standard score, which is the number of standard deviations from the mean.
     This transform leads to the "**Pearson product-moment correlation coefficient**."
  2. Transform each value to its rank, which is its index in the sorted list of values.
     This transform leads to the "**Spearman rank correlation coefficient**."

Pearson's correlation

- Pearson's correlation is always between -1 and +1 (including both).
  - If ρ is positive, we say that the correlation is positive, which means that when one variable is high, the other tends to be high.
  - If ρ is negative, the correlation is negative, so when one variable is high, the other is low.

- Only works for linear relationships (see next slide)

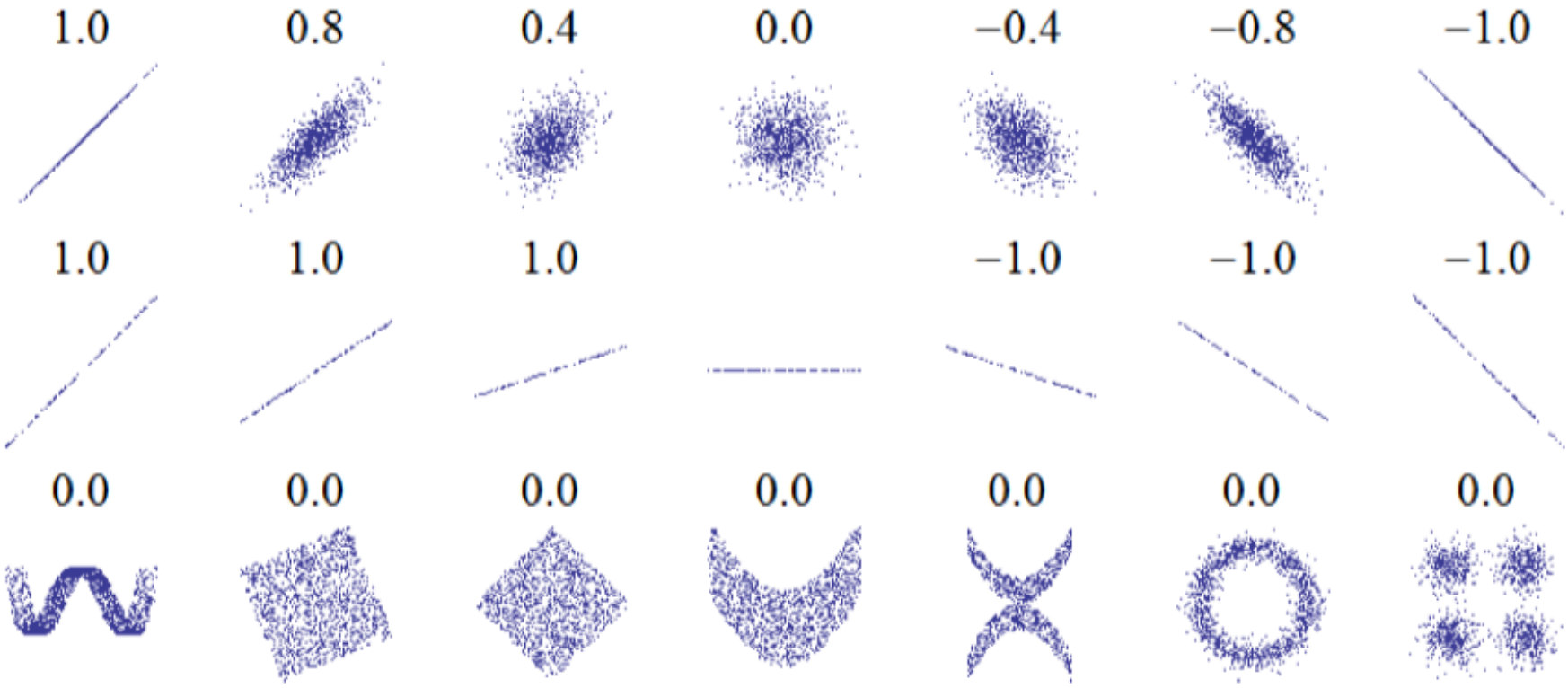


Figure 7-4. Examples of datasets with a range of correlations.

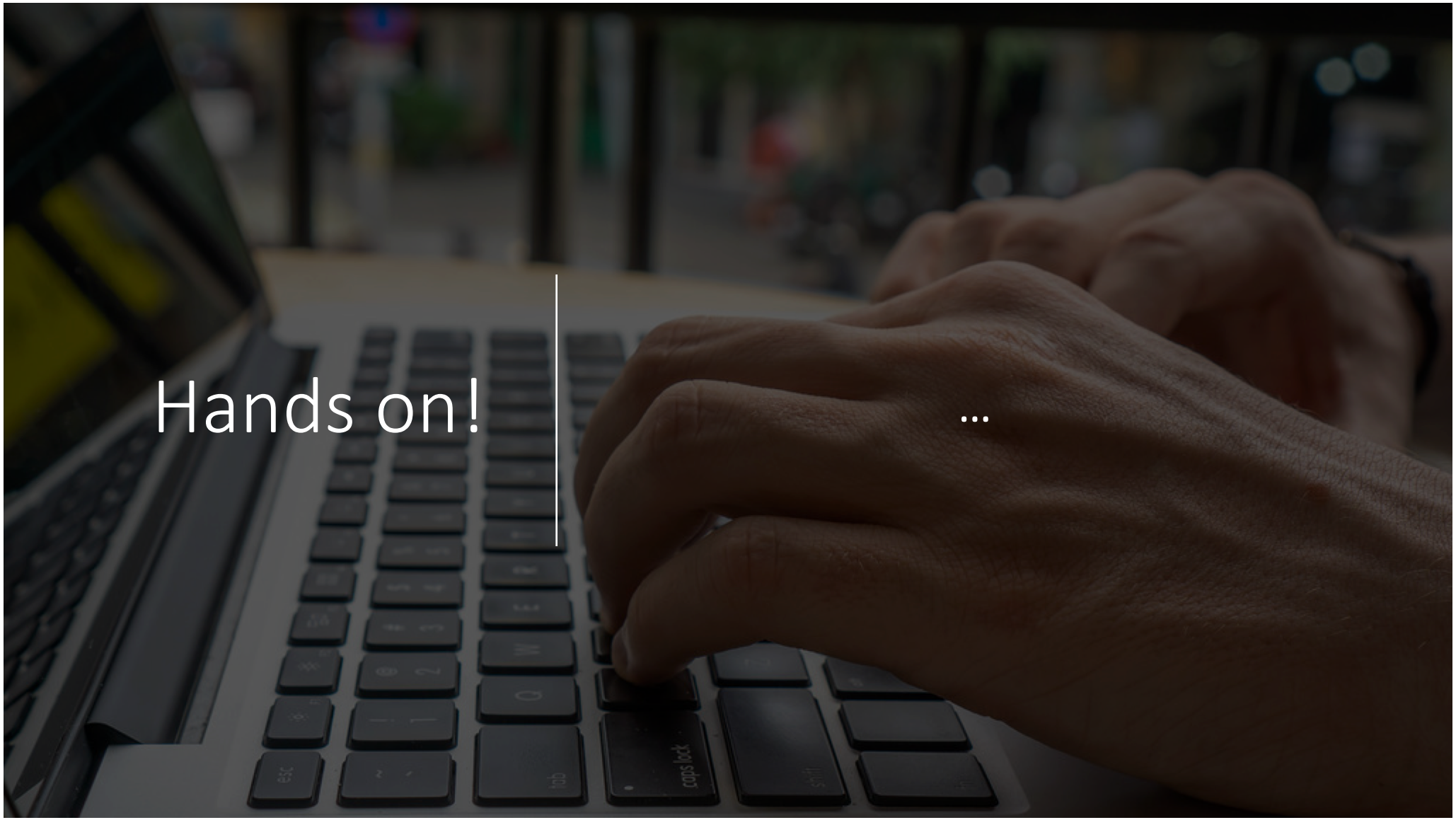
Spearman’s rank correlation

- Spearman’s rank correlation is an alternative to Pearson’s correlation that mitigates the effect of outliers and skewed distributions.

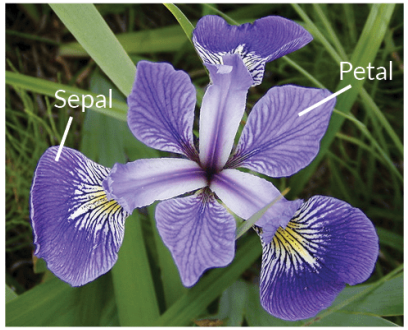
Correlation and causation

- If variables A and B are correlated, there are three possible explanations:
  - A causes B,
  - B causes A, or
  - some other set of factors causes both A and B.
- These explanations are called “causal relationships”.

- Randomized controlled trials
  - Subjects are assigned randomly to two (or more) groups: a treatment group that receives some kind of intervention, like a new medicine, and a control group that receives no intervention, or another treatment whose effects are known.


Hands on!
…

Assignment 3: Statistics – Part 1



Iris Versicolor

Iris Setosa

Iris Virginica

Goals:

- To transition from data analytics to basic statistical analysis.
- To practice the computation and displaying of summary statistics, percentiles, PMFs and (E)CDFs.
- To expand upon the prior experience of manipulating, summarizing, and visualizing small datasets.
- To display and interpret bee swarm plots and box-and-whisker plots.
- To visualize and compute pairwise correlations among variables in the dataset.

Concluding remarks