This class is being conducted over Zoom. As the instructor, I will be recording this session. I have disabled the recording feature for others so that no one else will be able to record this session. I will be posting this session to the course's website.

If you have privacy concerns and do not wish to appear in the recording, you may turn video off (click "stop video") so that Zoom does not record you.

The chat box is always open for discussion and questions to the entire class. You may also send messages privately to the instructor or the TAs. Please note that Zoom saves all chat transcripts.

I create a live transcription of each session using Otter.ai. This means that Otter.ai will transcribe anything spoken over the Zoom audio. The transcript will be posted with the session video on the course website.

Introduction to Statistics

Stats 7

Mary Ryan

Aug. 4, 2020


Course website:

https://canvas.eee.uci.edu/courses/28451

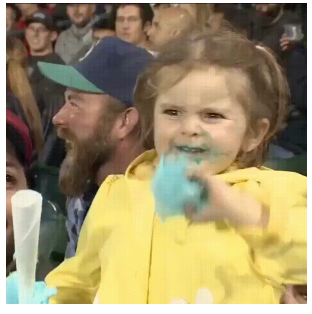

Slides can be found at:

https://maryryan.github.io/stats7-SS2-2020-slides/lec1/stats7-SS2-2020-lec1

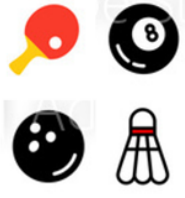
About the Teaching Team

- Instructor: Mary Ryan
  - 5th year Statistics PhD student
  - 📧: marymr@uci.edu
  - OH: Tuesdays & Thursdays, 4 - 5 p.m.

- TA: Kyle Conniff
  - 5th year Statistics PhD student
  - 📧: krconnif@uci.edu
  - OH: Wedednesdays, 11 a.m. - noon, and Sundays 7:30 - 8:30 p.m.



- TA: Jenifer Rim
  - 4th year Statistics PhD student
  - 📧: jsrim@uci.edu
  - OH: Fridays, 2 - 4 p.m.



Learning Objectives

Upon successful completion of the course, students should be able to:

- analyze and present data
- understand the basic design of and be able to evaluate experimental and observational studies
- have a basic understanding of fundamental probability principles and processes, and how they might appear in the real world
- identify and use appropriate statistical tools to answer and explore scientific questions
- be able to draw appropriate contextual conclusions based on results from statistical tools

Tentative Schedule

Week 1

- Descriptive statistics & data visualization
- Data sampling & bias

Week 2

- Regression
- Basic probability

Week 3

- Midterm
- Discrete probability distributions

Week 4

- Continuous probability distributions & sampling distributions
- One-sample inference

Week 5

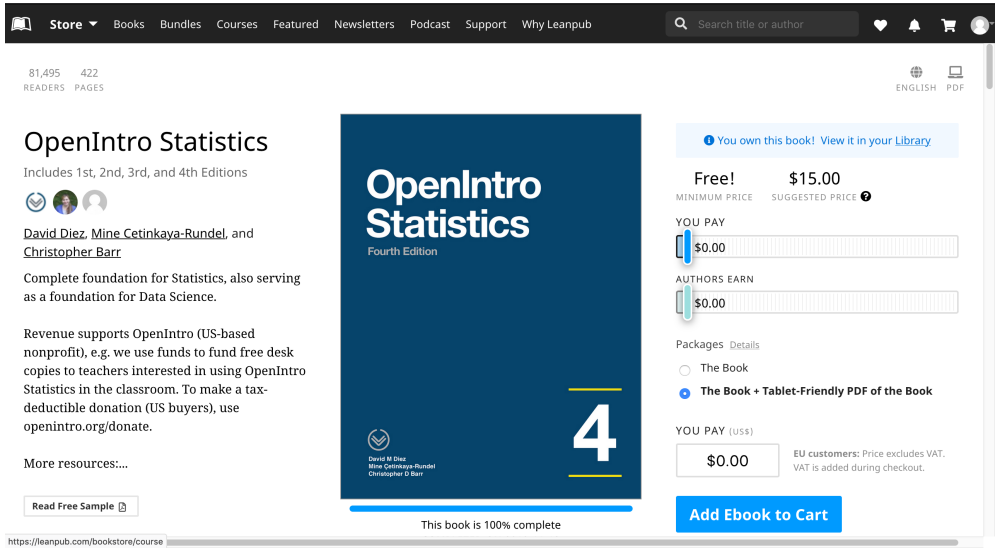
- Two-sample inference
- Inference for categorical data

Final Exam

Textbook

Recommended text: OpenIntro Statistics, 4th Edition

- PDF free to download



- Paperback copy $20 from Amazon

Assignments

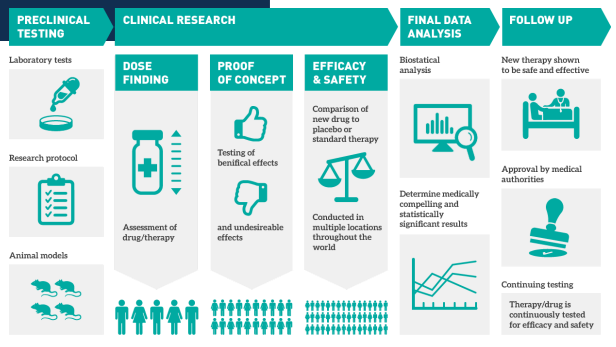
- Video Quizzes & Surveys: 5%
    - Short topic videos with embedded quiz questions to be completed before appropriate lecture, found on course website
- Homework: 25%
    - Weekly assignments, found on the course website
    - Open Tuesdays @ 1p
    - Due following Tuesday @ 12.59p (before lecture) via Canvas
    - For every day late, homework grade will suffer a 20% penalty
- Labs: 15%
    - Open Thursdays @ 4p, found on the course website
    - Due following Tuesday @ 12.59p (before lecture) via Canvas
- Midterm Exam: 25%
    - Date: Tuesday, Aug. 18 @ 1p
- Final Exam: 30%
    - Date: Tuesday, Sept. 8 @ 1p
    - Cumulative

If you have issues with either of the exam dates, let me know as soon as possible

What is Statistics?

What is Statistics?

United States® Census Bureau



BLS
BUREAU OF LABOR STATISTICS
U.S. DEPARTMENT OF LABOR

FiveThirtyEight

The Grammar of Science

Grammar takes a pile of words,

and helps turn it into a sentence that makes sense

In this class, we are going to learn about data, statistics,and how they impact our lives.

The Grammar of Science

Grammar takes a pile of words,

and helps turn it into a sentence that makes sense

In this class, we are going to learn about data, statistics,and how they impact our lives.

Statistics takes a pile of data,

and helps turn it into scientific conclusions that we can **interpret**

What Statistics Isn't

Just like all sentences that are grammatically correct aren't necessarily good at communicating their meaning...

Buffalo buffalo Buffalo buffalo

buffalo buffalo Buffalo buffalo

Statistics is not simply a machine where data goes in and truth pops out the other side

We need to use Statistics **conciously** and understand its **limitations**

Types of Data

- A variable is a measured characteristic of your data
    - A dataset is often made up of many different variables, of many different types

Types of Data

- A variable is a measured characteristic of your data
  - A dataset is often made up of many different variables, of many different types
- Categorical (qualitative) variables
  - Ordinal variable: categorical variable whose categories have a specific ordering
  - Nominal variable: categorical variable without any ordering
  - Binary variable: categorical variable with only 2 categories

Types of Data

- A variable is a measured characteristic of your data
  - A dataset is often made up of many different variables, of many different types
- Categorical (qualitative) variables
  - Ordinal variable: categorical variable whose categories have a specific ordering
  - Nominal variable: categorical variable without any ordering
  - Binary variable: categorical variable with only 2 categories
- Quantitative variables
  - Discrete variable: takes on numerical values in jumps
  - Continuous variable: takes on numerical values that can go out to infinite decimal points

Types of Data

1) Number of cats at the animal shelter

2) Species of pet

3) Weight

4) Mood

5) Diagnosis

6) Population of Orange County

7) Airplane ticket class

Types of Data

1) Number of cats at the animal shelter — Discrete

2) Species of pet

3) Weight

4) Mood

5) Diagnosis

6) Population of Orange County

7) Airplane ticket class

Types of Data

1) Number of cats at the animal shelter	Discrete
2) Species of pet	Nominal
3) Weight	
4) Mood	
5) Diagnosis	
6) Population of Orange County	
7) Airplane ticket class	

Types of Data

1) Number of cats at the animal shelter	Discrete
2) Species of pet	Nominal
3) Weight	Continuous
4) Mood	
5) Diagnosis	
6) Population of Orange County	
7) Airplane ticket class	

Types of Data

1) Number of cats at the animal shelter	Discrete
2) Species of pet	Nominal
3) Weight	Continuous
4) Mood	Nominal
5) Diagnosis	
6) Population of Orange County	
7) Airplane ticket class	

Types of Data

1) Number of cats at the animal shelter	Discrete
2) Species of pet	Nominal
3) Weight	Continuous
4) Mood	Nominal
5) Diagnosis	Nominal
6) Population of Orange County	
7) Airplane ticket class	

Types of Data

1) Number of cats at the animal shelter	Discrete
2) Species of pet	Nominal
3) Weight	Continuous
4) Mood	Nominal
5) Diagnosis	Nominal
6) Population of Orange County	Discrete
7) Airplane ticket class	

Types of Data

1) Number of cats at the animal shelter	Discrete
2) Species of pet	Nominal
3) Weight	Continuous
4) Mood	Nominal
5) Diagnosis	Nominal
6) Population of Orange County	Discrete
7) Airplane ticket class	Ordinal

Types of Data

1) Number of cats at the animal shelter	Discrete
2) Species of pet	Nominal
3) Weight	Continuous
4) Mood	Nominal
5) Diagnosis	Nominal
6) Population of Orange County	Discrete
7) Airplane ticket class	Ordinal

8) Money you spend to wash 1 load of laundry

Types of Data

1) Number of cats at the animal shelter	Discrete
2) Species of pet	Nominal
3) Weight	Continuous
4) Mood	Nominal
5) Diagnosis	Nominal
6) Population of Orange County	Discrete
7) Airplane ticket class	Ordinal
8) Money you spend to wash 1 load of laundry	Continuous

Types of Data

8) Money you spend to wash 1 load of laundry

- I spend $1.75 per wash cycle

Types of Data

8) Money you spend to wash 1 load of laundry

- I spend $1.75 per wash cycle
    - Looks discrete, right?

Types of Data

8) Money you spend to wash 1 load of laundry

- I spend $1.75 per wash cycle
  - Looks discrete, right?
  - Converted to Iraqi Dinar: 2080.75 ع.د

Types of Data

8) Money you spend to wash 1 load of laundry

- I spend $1.75 per wash cycle
  - Looks discrete, right?
  - **Converted to Iraqi Dinar: 2080.75 ع.د**
  - **Converted to Euro: € 1.5596**

Types of Data

8) Money you spend to wash 1 load of laundry

- I spend $1.75 per wash cycle
  - Looks discrete, right?
  - **Converted to Iraqi Dinar: 2080.75 ع.د**
  - **Converted to Euro: € 1.5596**
  - **Converted to Mexican Peso: $ 40.285735**

Types of Data

8) Money you spend to wash 1 load of laundry

- I spend $1.75 per wash cycle
  - Looks discrete, right?
  - Converted to Iraqi Dinar: 2080.75 ع.د
  - Converted to Euro: € 1.5596
  - Converted to Mexican Peso: $ 40.285735
- While we might think of currency in "discrete" US Dollar (or Dinar, or Euro, or Peso) units, any amount of money can be converted from one currency to another
  - Currency conversion doesn't respect the fact that you don't carry around $1/12^{th}$ cents
  - So we generally think of currency as a continuous variable

How Do We Describe Our Data?

For categorical data, we might want to:

- Know how many observations are in each caetgory
  - Counts
- Know how large a category is compared to all observations
  - Percentages
  - $\frac{\#\text{ observations in a category}}{\text{Total observations}} \times 100$

How Do We Describe Our Data?

For quantitative data, we might want to:

- Know what the smallest value of a variable is
  - **Minimum**
- Know what the largest value of a variable is
  - **Maximum**
- Know the distance between the maximum and the minimum
  - **Range**
- Know what the center of our data is...

Journey to the Center of the Data

- Mean
  - The "typical value" of the data
  - $\bar{x} = \frac{\text{sum of data points}}{\text{number of data points}} = \frac{\sum_{i=1}^{n} X_i}{n}$
- Median
  - The "exact middle" of the data
  - 50% of data points are below the median, 50% of data points are above the median
- Mode
  - The most common value in the data

Journey to the Center of the Data

- **Mean**
  - The "typical value" of the data
  - $\bar{x} = \frac{\text{sum of data points}}{\text{number of data points}} = \frac{\sum_{i=1}^{n} X_i}{n}$
- **Median**
  - The "exact middle" of the data
  - 50% of data points are below the median, 50% of data points are above the median
- **Mode**
  - The most common value in the data

Median

2, 5, 6, 8, 8, 8, 10, 13, 14, 16, 16, 19, 20, 21, 25

Median

2, 5, 6, 8, 8, 8, 10, 13, 14, 16, 16, 19, 20, 21, 25

Median

2, 5, 6, 8, 8, 8, 10, 13, 14, 16, 16, 19, 20, 21, 25

Median

2, 5, 6, 8, 8, 8, 10, 13, 14, 16, 16, 19, 20, 21, 25

Median

2, 5, 6, 8, 8, 8, 10, 13, 14, 16, 16, 19, 20, 21, 25

Median

2, 5, 6, 8, 8, 8, 10, 13, 14, 16, 16, 19, 20, 21, 25

2, 5, 6, 8, 8, 8, 10, 10, 13, 14, 16, 16, 19, 20, 21, 25

Median

2, 5, 6, 8, 8, 8, 10, 13, 14, 16, 16, 19, 20, 21, 25

2, 5, 6, 8, 8, 8, 10, 10, 13, 14, 16, 16, 19, 20, 21, 25

Median

2, 5, 6, 8, 8, 8, 10, 13, 14, 16, 16, 19, 20, 21, 25

2, 5, 6, 8, 8, 8, 10, 10, 13, 14, 16, 16, 19, 20, 21, 25

Median

2, 5, 6, 8, 8, 8, 10, **13,** 14, 16, 16, 19, 20, 21, 25

2, 5, 6, 8, 8, 8, 10, **10, 13,** 14, 16, 16, 19, 20, 21, 25

(10+13)/2 = 11.5

How Do We Describe Our Data?

- We might also want to know how much our observations vary
- Nice to compare how far away observations are from the center of the data (deviation)

$$x_i - \bar{x}$$

How Do We Describe Our Data?

- We might also want to know how much our observations vary
- Nice to compare how far away observations are from the center of the data (deviation)

$$x_i - \bar{x}$$

- But some observations are above the mean and some are below...
  - If we just add the deviations up some of them might cancel out
  - Square the deviations to remove the sign:

$$(x_i - \bar{x})^2$$

How Do We Describe Our Data?

- We might also want to know how much our observations vary
- Nice to compare how far away observations are from the center of the data (deviation)

$$x_i - \bar{x}$$

- But some observations are above the mean and some are below...
  - If we just add the deviations up some of them might cancel out
  - Square the deviations to remove the sign:

$$(x_i - \bar{x})^2$$

- But every observation has a different squared deviance. How do we get one metric?

How Do We Describe Our Data?

- We might also want to know how much our observations vary
- Nice to compare how far away observations are from the center of the data (deviation)

$$x_i - \bar{x}$$

- But some observations are above the mean and some are below...
  - If we just add the deviations up some of them might cancel out
  - Square the deviations to remove the sign:

$$(x_i - \bar{x})^2$$

- But every observation has a different squared deviance. How do we get one metric?
  - Take an average!
  - Variance (s^2) = $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$
  - Standard deviation (s) = $\sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$

Variance & Standard Deviation

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

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

- When variance/standard deviation is **large**, the observations vary from the mean a lot
- When variance/standard deviation is **small**, the observations generally stay close to the mean

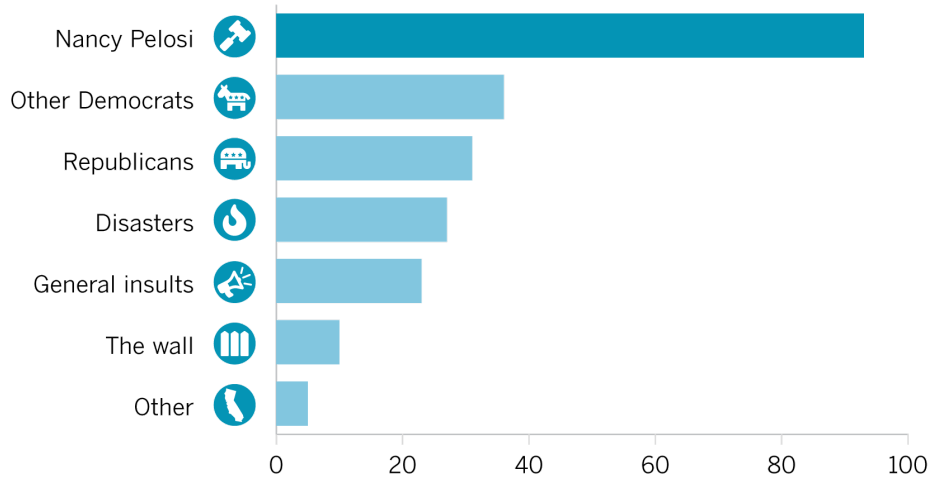
Data Visualization

- Summarizing data with metrics are nice, but people are much better at identifying differences/patterns graphically

Data Visualization

- Summarizing data with metrics are nice, but people are much better at identifying differences/patterns graphically

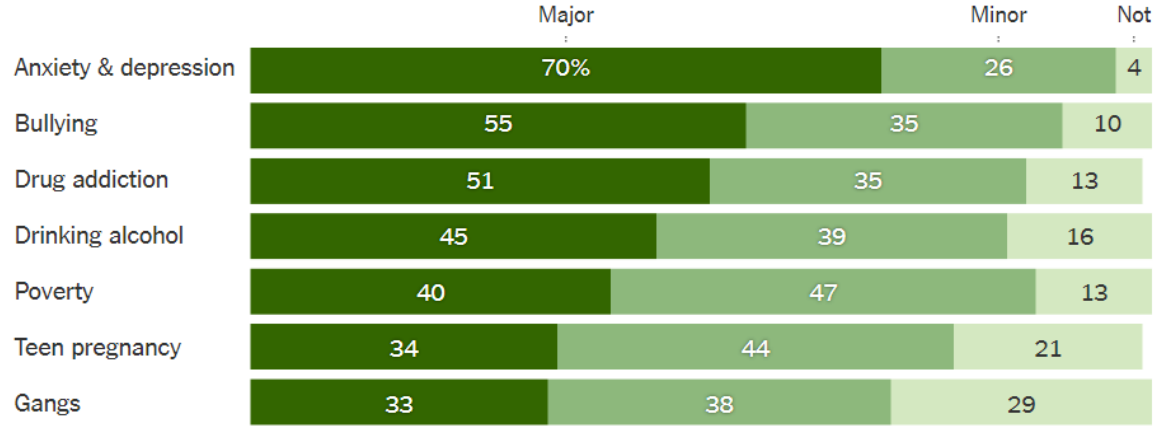
- Bar plot
  - Categories on one axis
  - Counts/percentages on the other axis
  - AKA: bar graph, bar chart, barplot



The California-related topics the president tweets most frequently about (Sept. 17, 2019), via Priya Krishnakumar at The LA Times

Data Visualization

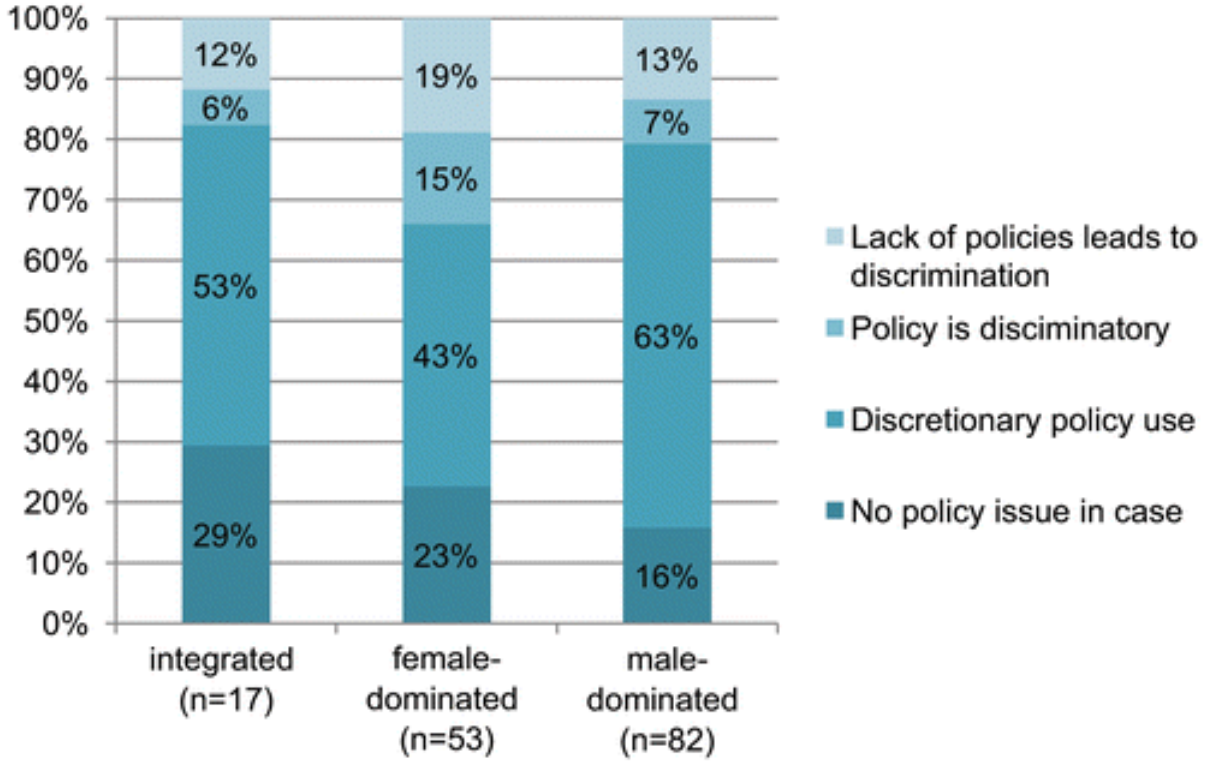
- Stacked bar plot
  - Categories on one axis
  - Percentage on other axis
  - Bar divided into percentages of how common 2nd categorical variable is within your axis category
  - Full bar should add up to 100%



Teenagers Say Depression and Anxiety Are Major Issues Among Their Peers (Feb. 29, 2019), via Karen Zraick at The New York Times

Data Visualization

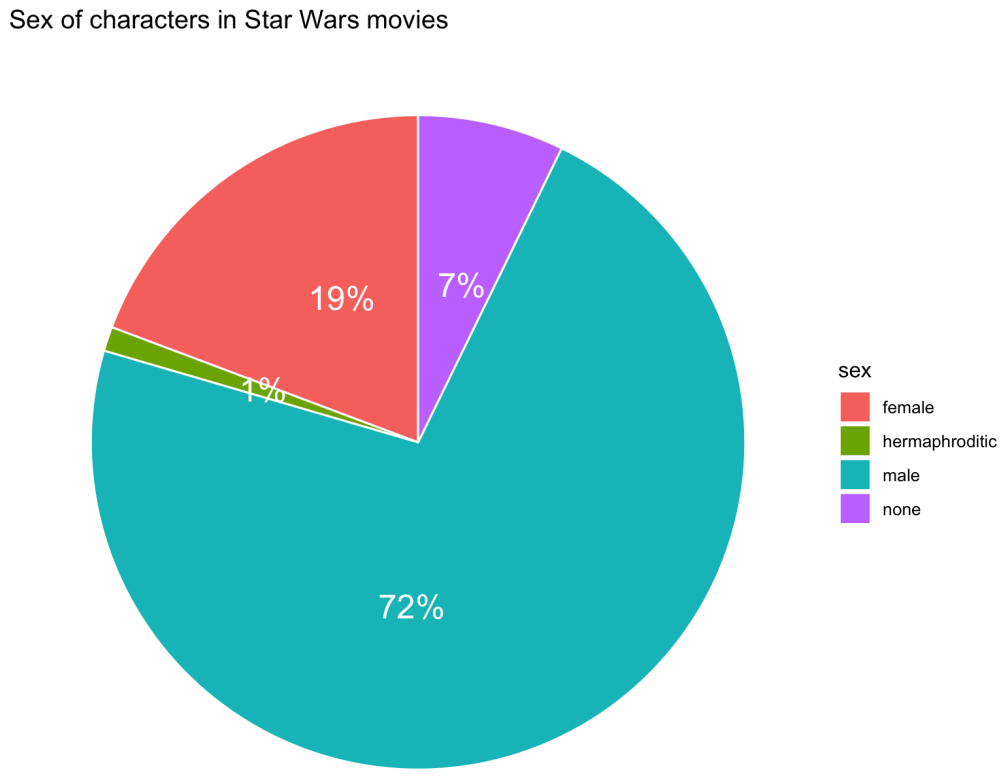
- Stacked bar chart
  - What's going on here?



Gender Discrimination at Work: Connecting Gender Stereotypes, Institutional Policies, and Gender Composition of Workplace (Dec. 5, 2011), via Donna Bobbitt-Zeher at Gender & Society

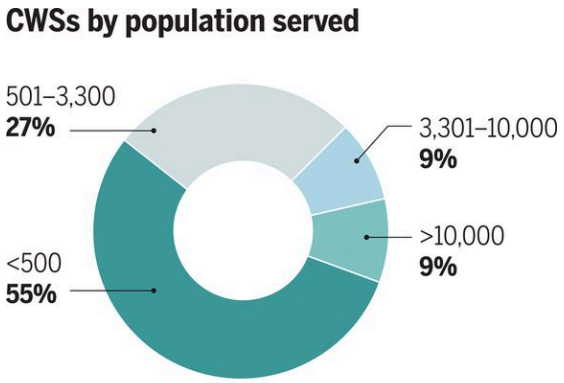
Data Visualization

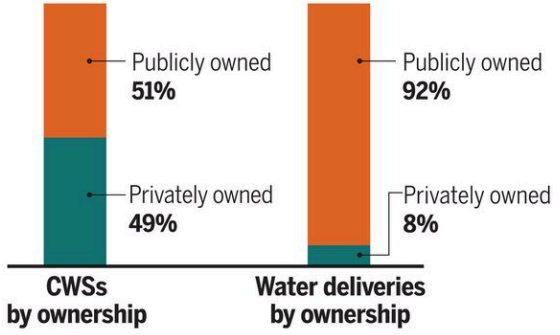
- Pie chart
  - Can also visualize percentages
  - No real axes
  - Total circle = 100%
  - Each wedge corresponds to percentage of total observations correspond to one category
  - People are much worse at judging volume by circle wedge than volume by box, though



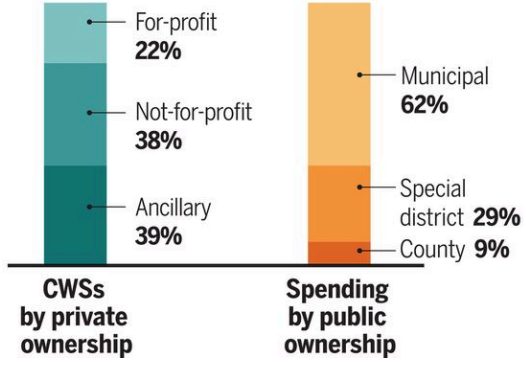
Data Visualization

- Pie chart & stacked bar plot
  - What's going on here?
  - CWS = community water system



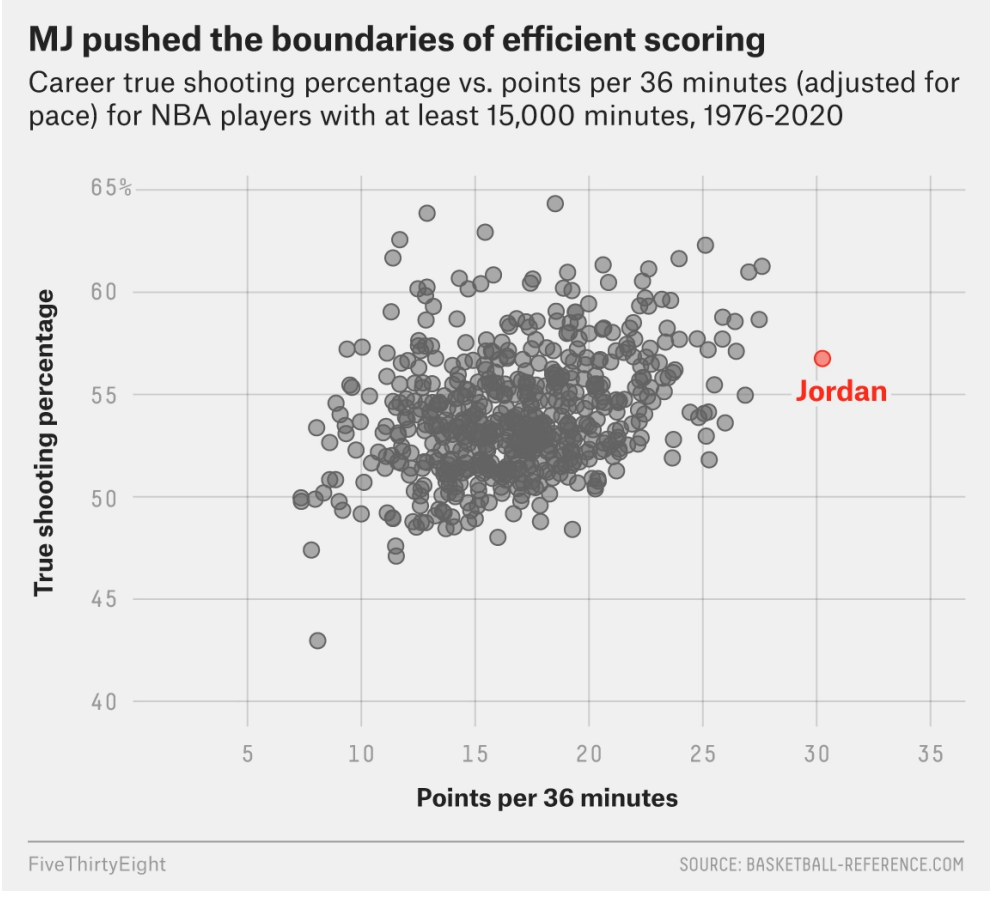


The effects of drinking water service fragmentation on drought-related water security (April 17, 2020), via Megan Mullin at Science



Data Visualization

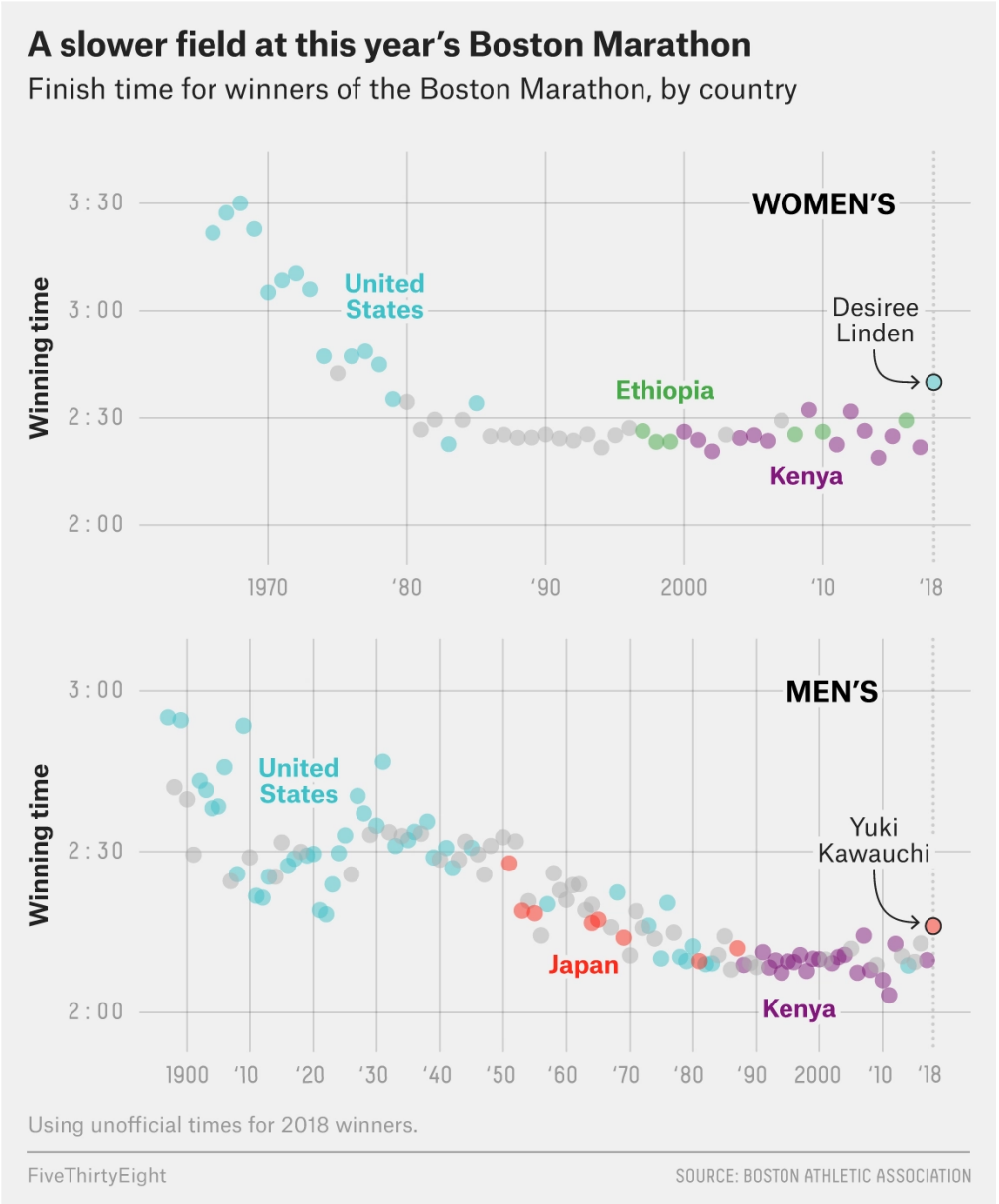
- Scatter plot
  - Quantitative variable on one axis
  - Another quantitative variable on other axis
  - Observations shown as points



Why Michael Jordan Was The Best (April 17, 2020), via Neil Paine at FiveThirtyEight

Data Visualization

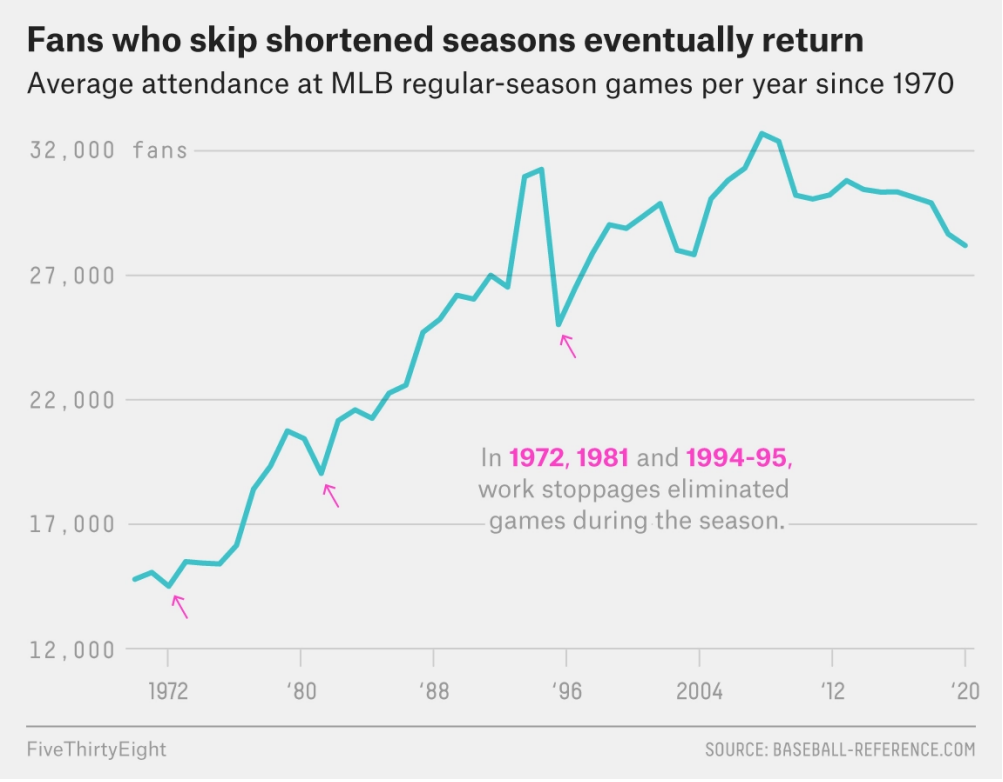
- Scatter plot
  - What's going on here?



This Year's Boston Marathon Was Slooooooowww (April 16, 2018), via Gus Wezerek at FiveThirtyEight

Data Visualization

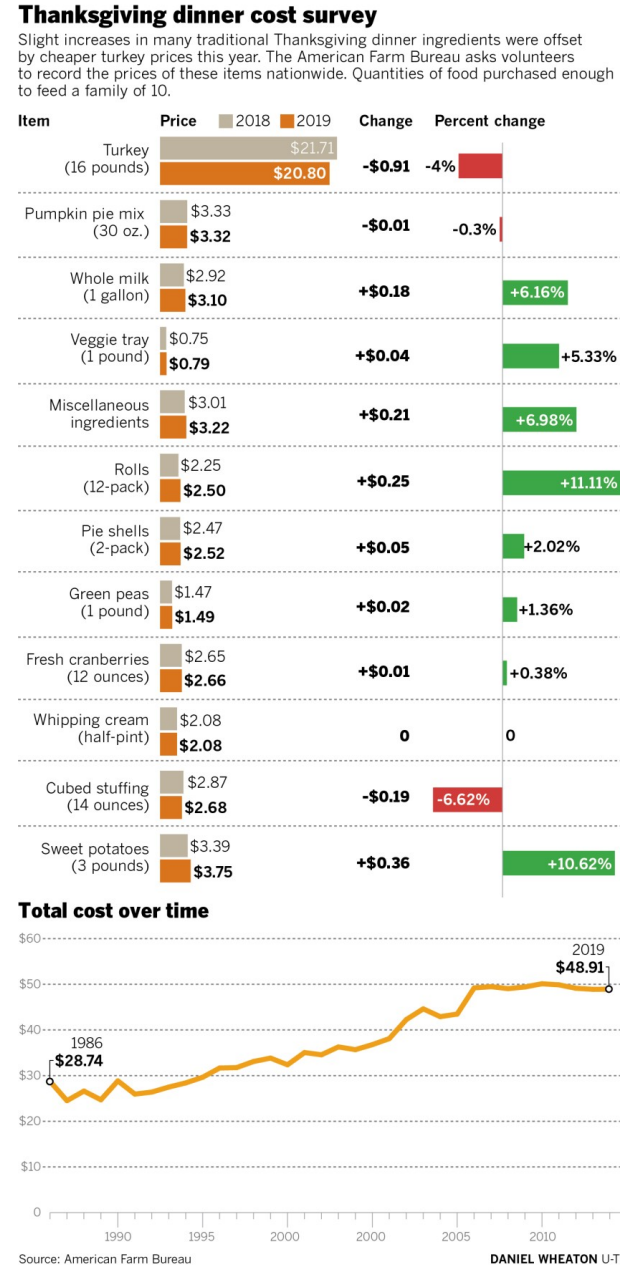
- Line plot
  - Time on one axis (usually x-axis)
  - Quantitative variable on other axis
  - Line connecting each observation (usually the peaks in the line)
  - Good for observing trends over time



Do Baseball's Labor Fights Drive Fans Away? (June 12, 2020), via Travis Sawchik at FiveThirtyEight

Data Visualization

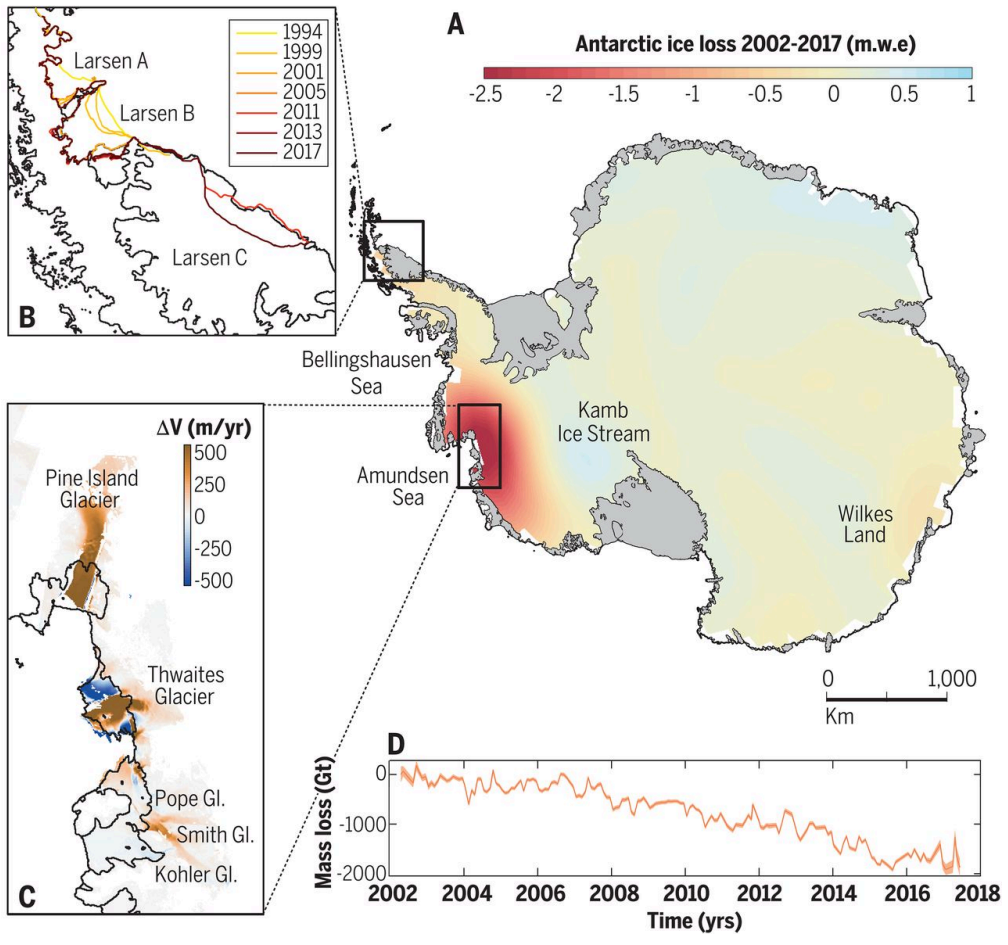
- Line plots
  - What's going on here?



Thanksgiving dinner costs remain stable (Nov. 27, 2019), via Daniel Wheaton at The San Diego Union-Tribune

Data Visualization

- Line plots
  - What's going on here (part D)?

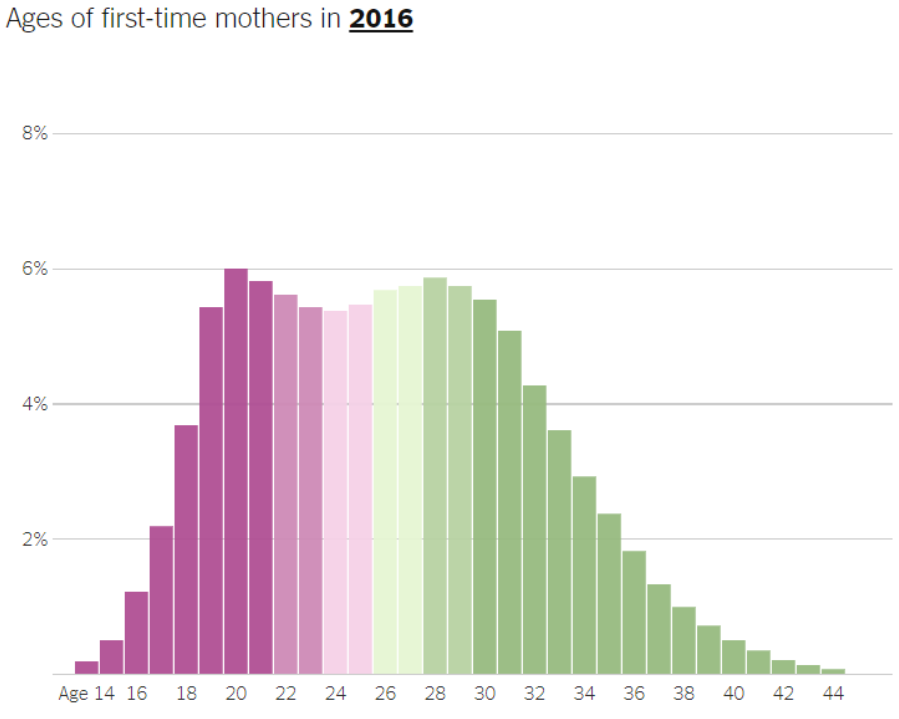


History, mass loss, structure, and dynamic behavior of the Antarctic Ice Sheet (March 20, 2020),

via Robin E. Bell & Helene Seroussi at Science

Data Visualization

- Histogram
  - Like a bar plot, but instead of counts of observations in categories we have counts of observations in intervals (bins) of a quantitative variable
  - You can change the interval width (and the # of bars) for histograms; you can't combine or divide categories for bar plots
  - Good for figuring out the data's shape



The age of first-time mothers in 2016 (Aug. 4, 2018), via Quoctrung Bui at The New York Times

Data Visualization

- Histogram
  - What's going on here?

"All the News That's Fit to Print"

The New York Times

Late Edition

Today, clouds and sunshine, stray showers, windy, chilly, high 49. Tonight, mostly clear, brisk, cold, low 39. Tomorrow, clouds and sunshine, high 62. Weather map is on Page C8.

VOL. CLXIX ... No. 58,688 © 2020 The New York Times Company NEW YORK, SATURDAY, MAY 9, 2020 $3.00

U.S. UNEMPLOYMENT IS WORST SINCE DEPRESSION

MONTHLY CHANGE IN JOBS SINCE THE END OF WORLD WAR II

NATIONAL UNEMPLOYMENT RATE

APRIL UNEMPLOYMENT BY RACE BY GENDER BY AGE GROUP

14.7%

14.2 16.7 18.9 13.5 16.2 11.5 12.3 13.6 14.5 25.7

White Black Hispanic Men Women 35 to 44 45 to 54 55+ 25 to 34 20 to 24

Georgia Killing Puts Spotlight on a Police Force's Troubled History

Long Path to Arrests of Ex-Officer and Son in Black Man's Death

April's Rate of 14.7% Touches All Parts of Economy

By NELSON D. SCHWARTZ and BEN CASSELMAN

NEWS ANALYSIS

In Flynn Case, Russia Inquiry Is Barr's Target

By MARK MAZZETTI

WHITE HOUSE MEMO

If West Wing Still Isn't Safe, Is Any Office?

By PETER BAKER and MICHAEL CROWLEY

A Mexico City crematory. Regional virus deaths are officially low.

As Official Toll Ignores Reality, Mexico's Hospitals Are Overrun

By AZAM AHMED

A Walk a Day to Get By

Moscow's Paused Celebration

G.O.P. Convention Qualms

Politics and the Justice Dept.

Banker Shakes Up Suriname

A Good Walk, Unspoiled

Anyone Up to Shoot Hoops?

New York's Return to the Arts

A Shot at Stardom, Revised

Timothy Egan

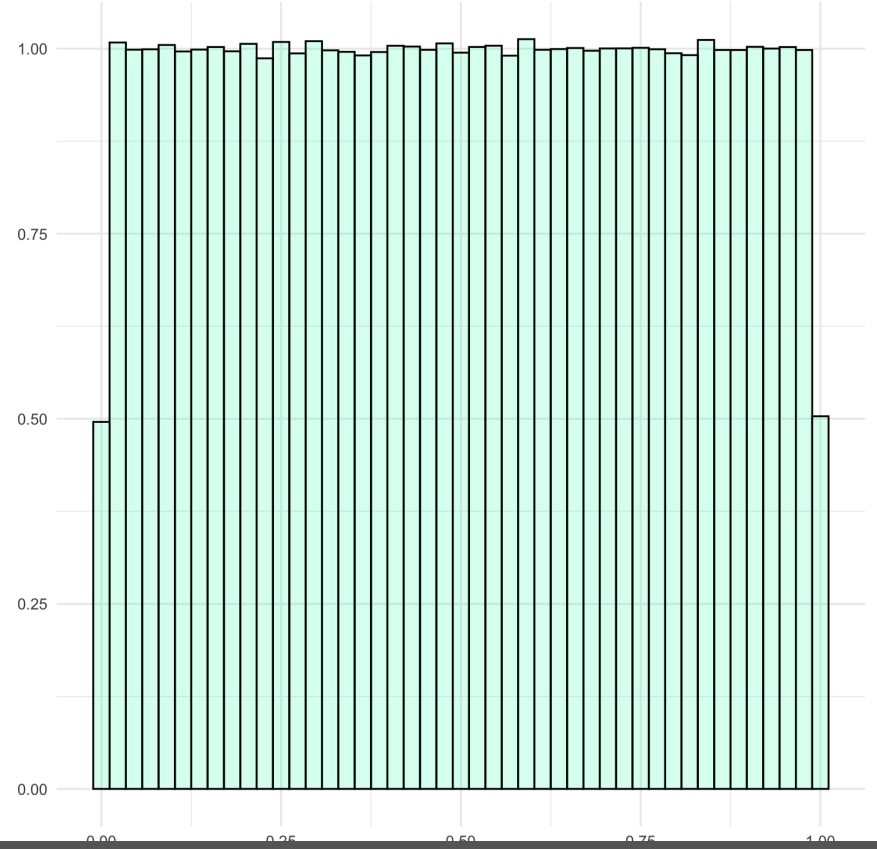
20,500,000 jobs lost in April

Monthly change in jobs since WWII (May 9, 2020), via Guilbert Gates, Ella Koeze, and Bill Marsh at The New York Times

Data Shapes

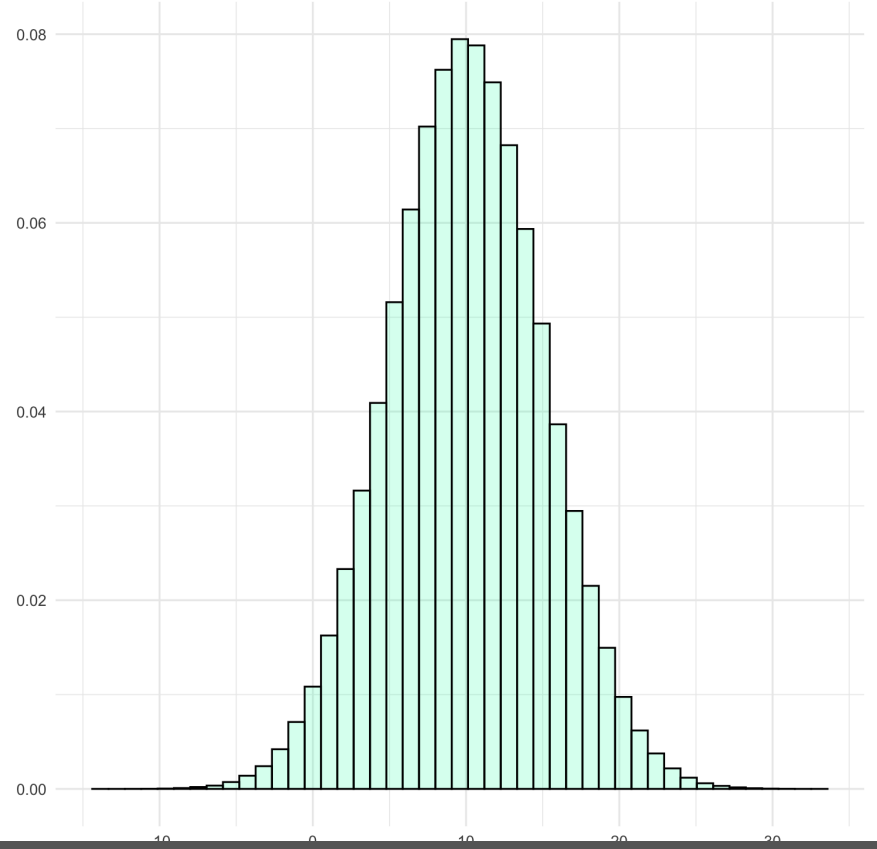
Uniform

- No defined peaks



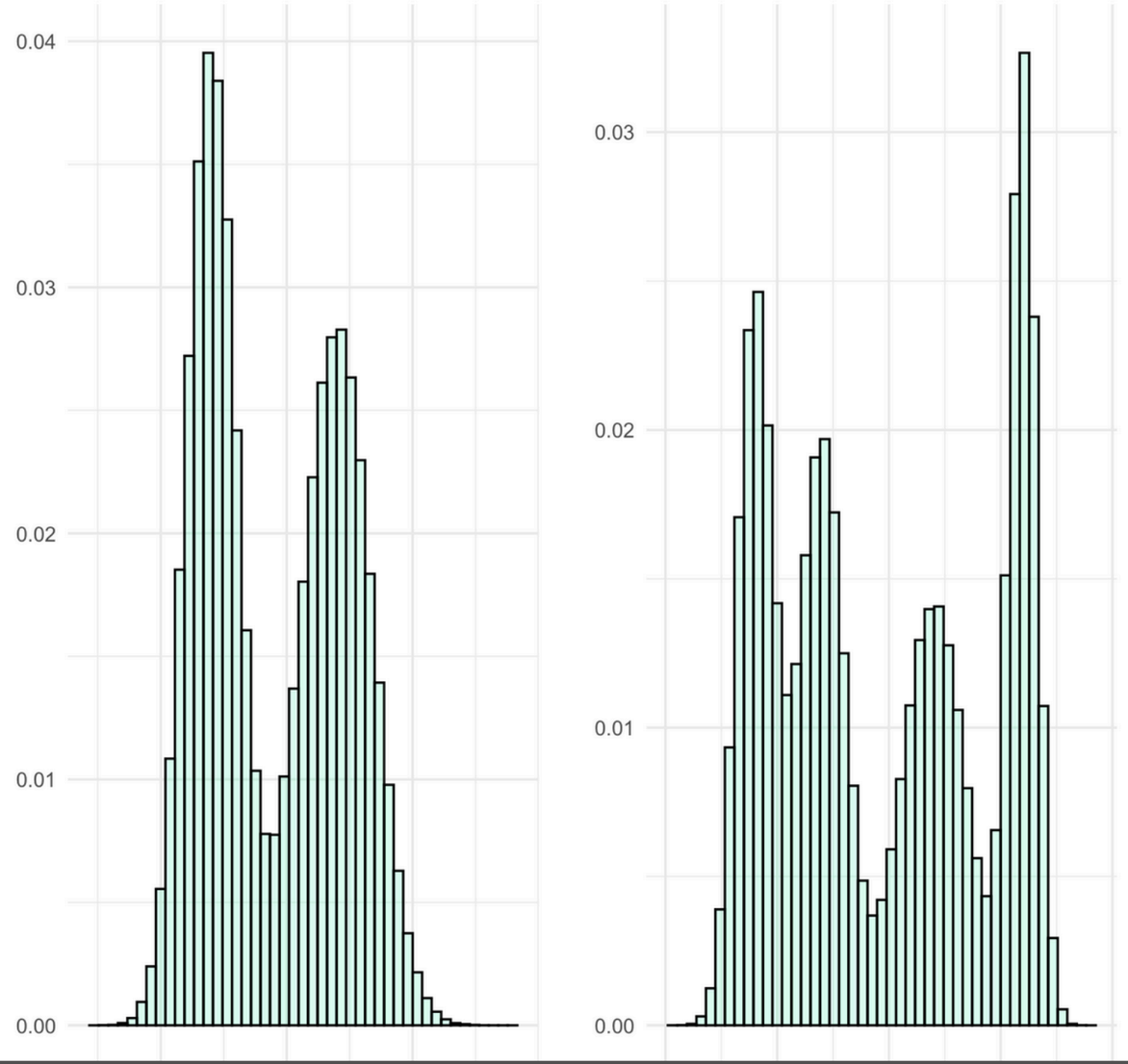
Unimodal

- One defined peak



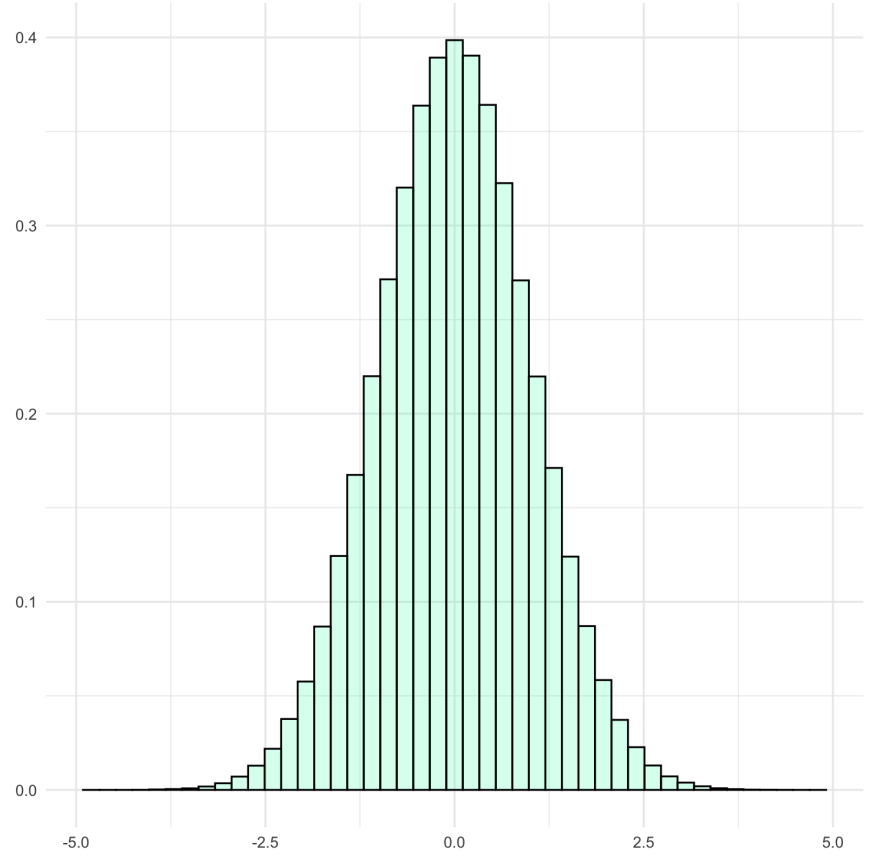
Data Shapes

Multimodal: Multiple defined peaks

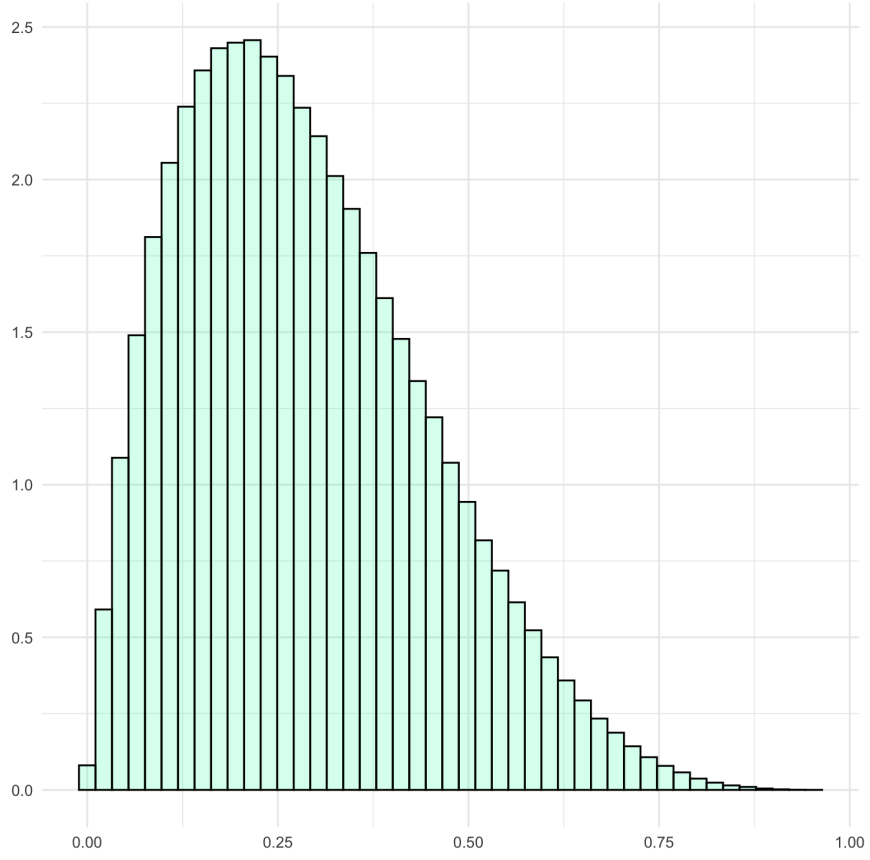


Data Shapes

Symmetrical

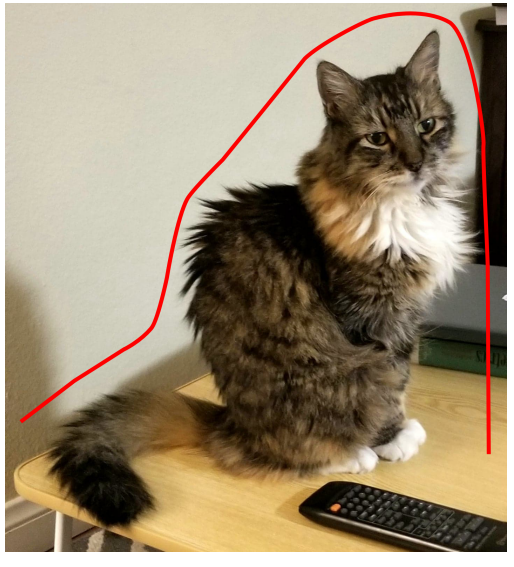


Skewed

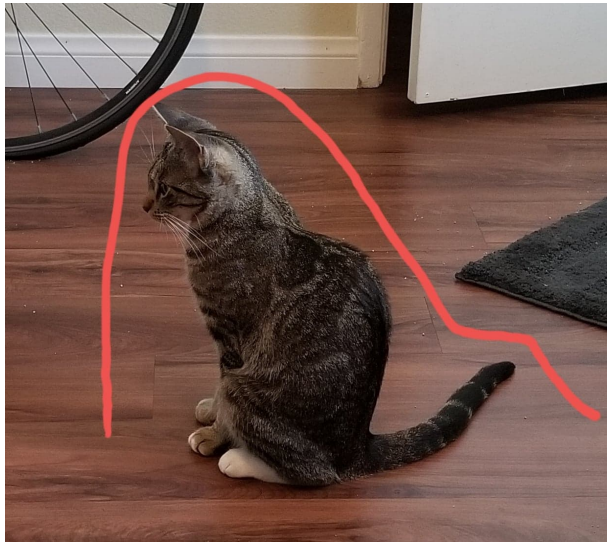


Skew

- Skew is what happens when the mean and median aren't exactly the same
- Right (positive) skew
  - Mean to the right (greater than) median
- Left (negative) skew
  - Mean to the left (less than) median
- When in doubt, look for the "tail"



Left skew



Right skew

Five Number Summary

- Getting a 10,000-foot view of the data's shape
- The five number summary contains:
  - Minimum
  - **1st quartile** (25^{th} percentile)
    - 25% of data are below the 1st quartile
  - Median (50^{th} percentile)
  - **3rd quartile** (75^{th} percentile)
    - 75% of data are below the 3rd quartile
  - Maximum

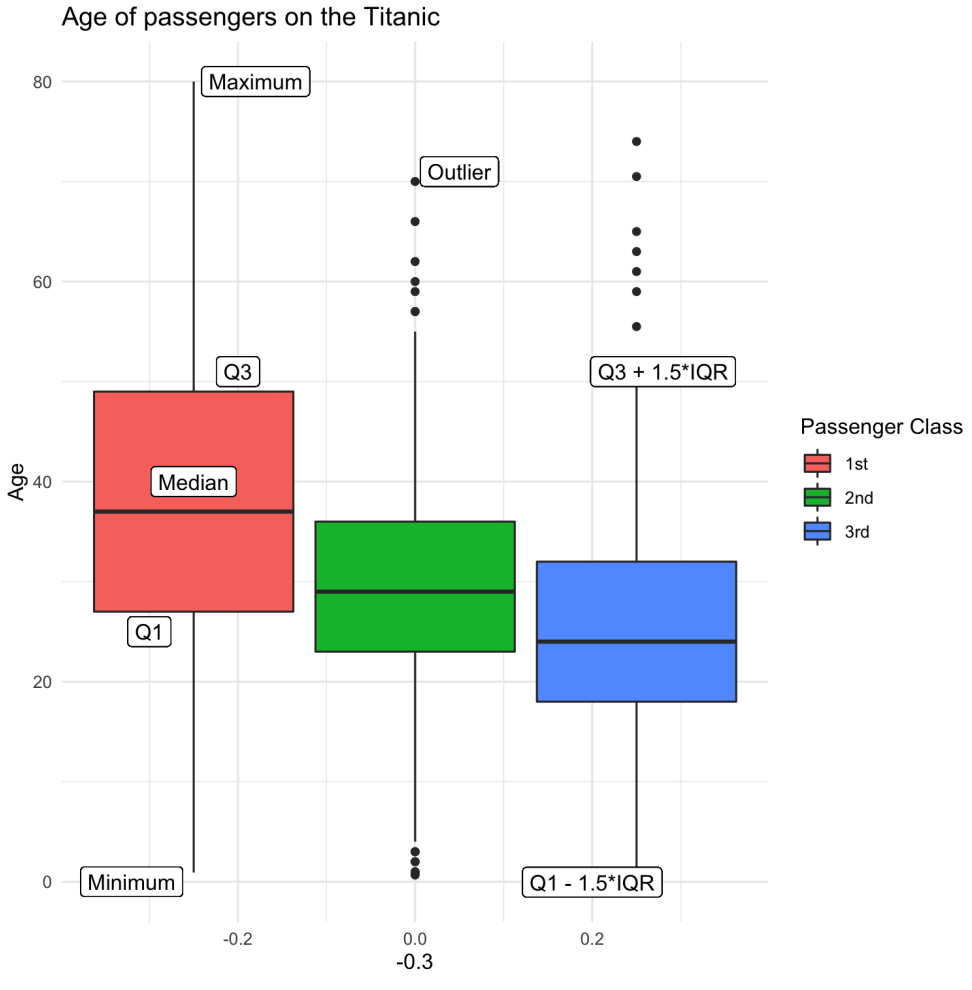
Data Visualization

- **Boxplot**
  - Bottom line of box at 1st quartile
  - Top line of box at 3rd quartile
  - Line in middle of box at median
  - Box "whiskers" extend from minimum to maximum, or to "outlier fences"

$$\text{Upper fence:} Q3 + 1.5 \times (Q3 - Q1)$$

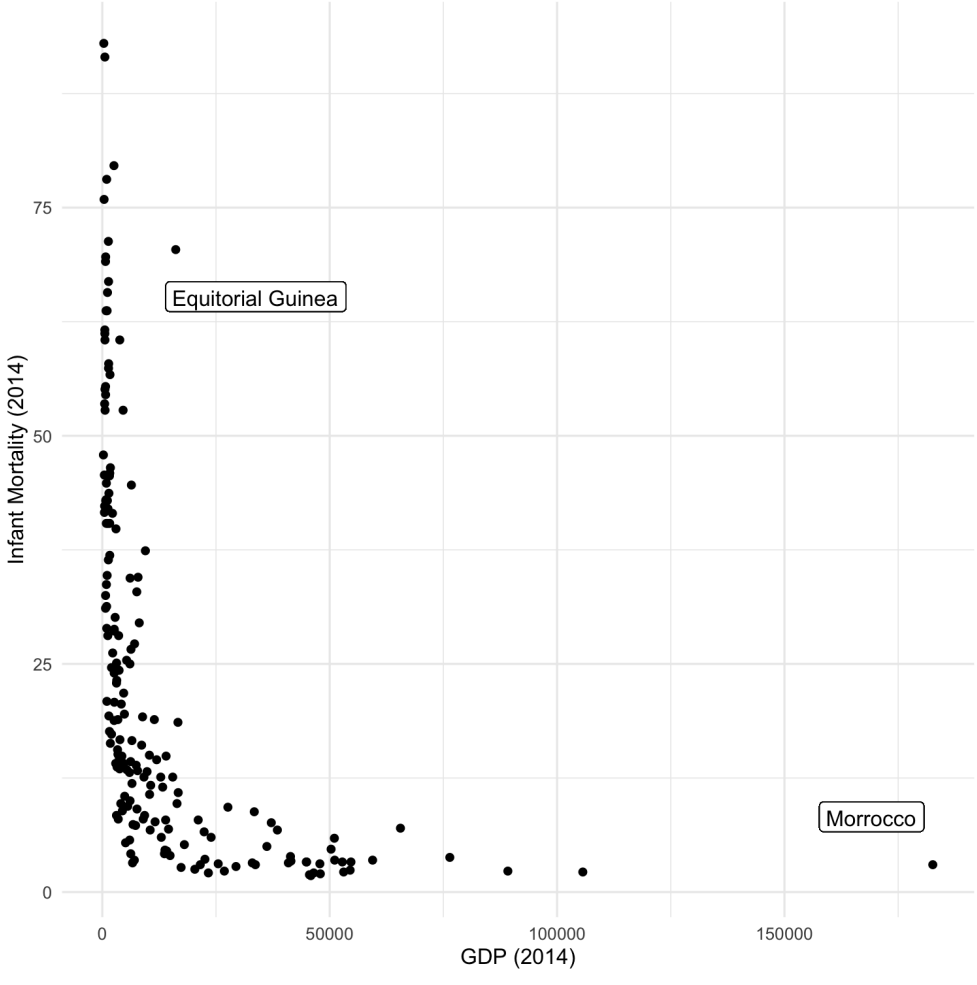
$$\text{Lower fence:} Q1 - 1.5 \times (Q3 - Q1)$$

  - Dots are outliers
    - Observations above $Q3 + 1.5 \times IQR$ or below $Q1 - 1.5 \times IQR$



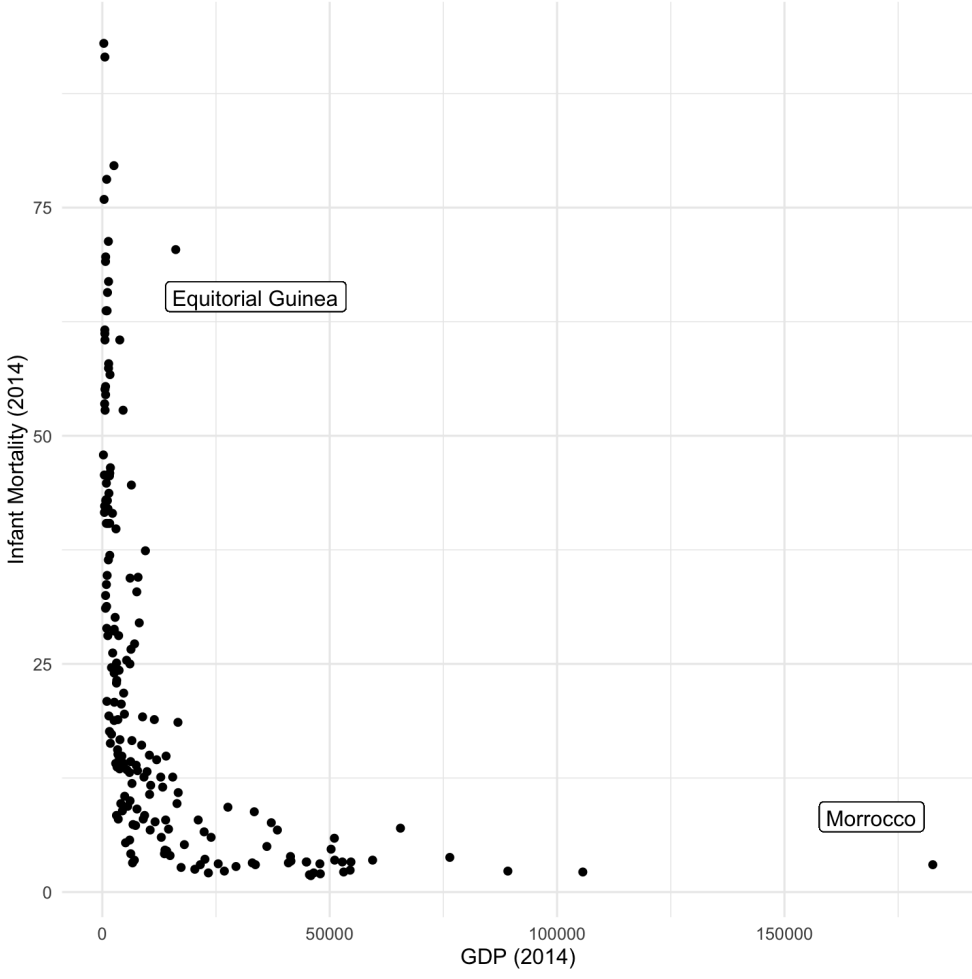
Outliers

- Outliers are data points that are strangely far away from the rest of the data
- Could signal that the values of those observations are wrong
    - Data entry error, recorded in inches instead of centimeters, etc.
- Could just be strange observations
- Outliers will create skew, pull your mean away from the median



Outliers

- Outliers are data points that are strangely far away from the rest of the data
- Could signal that the values of those observations are wrong
    - Data entry error, recorded in inches instead of centimeters, etc.
- Could just be strange observations
- Outliers will create skew, pull your mean away from the median
- **What do we do with them?**



Data Visualization

- Histogram vs. boxplot
  - Histograms give a more detailed view of data shape than boxplots
  - Because its less detailed, boxplots can be better for comparing the data shapes of multiple groups