ECON 3818

Chapter 1

Kyle Butts

27 September 2021

Introduction

What to expect from the course

Objectives:

- Give you a background on statistical theory and their application
- Learn how to perform basic statistical analysis in the `R` programming language
- Prepare you to succeed in econometrics courses

Grading Summary

ASSIGNMENT	PERCENTAGE
Homework	10%
R Problem Sets	15%
R Project	15%
Midterm 1	20%
Midterm 2	20%
Final	20%

Expectations for Students

Homework & R Homework

- Weekly homeworks are assigned, the best way to learn mathematics is **practice, practice, practice**
- Absolutely NO late homework; will drop the lowest two homeworks and the lowest R homework.

Expectations for Students

Attendance

- Really important to attend lecture to get handle on new material, but no attendance will be taken.
- Clicker questions will count for extra credit

Expectations for Students

Recitation

- Recitation attendance is not mandatory, but if you are going to study for 1 hour/week, recitation is the best place to do it. You will walk through examples and really helpful for prepping for exams

Expectations for Students

R Project

- You will download real world data and perform basic statistical analysis and create data visualizations.

Expectations for Students

Midterms and Final Exam

- NO makeup midterms, weight of missed midterm will be added to final
- Must inform me of any accommodations **two weeks** before an exam

What is Statistics?

Statistics gives us a way of linking *economic theory* with the real world through data analysis

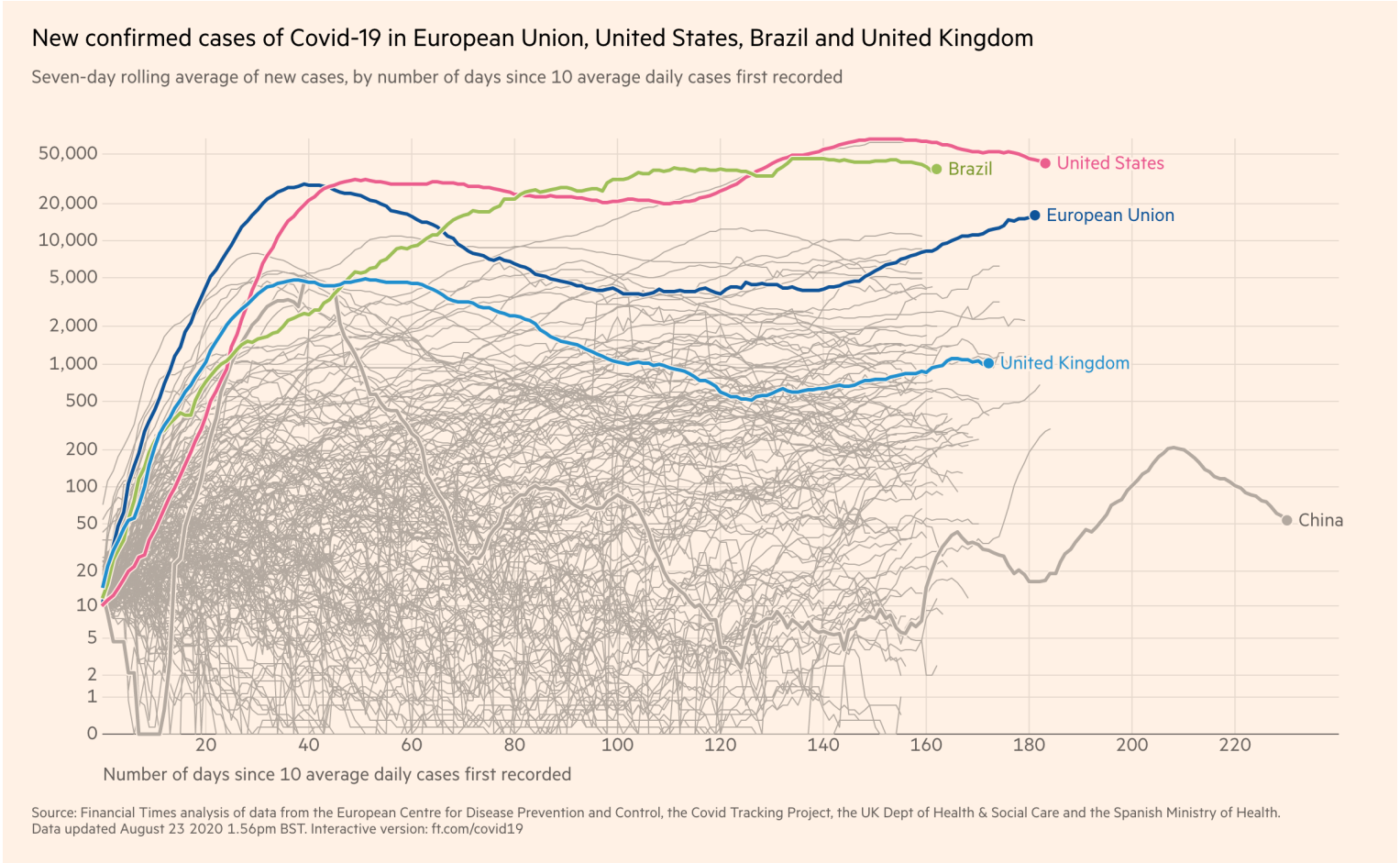
- How did the market react when interest rates went up?
- How did firms respond to a new government policy?

Statistics allows us to translate datasets into *usable* information

- Summary statistics help describe large groups of data
- Use statistics to make predictions
- Statistics helps us inform our decision making

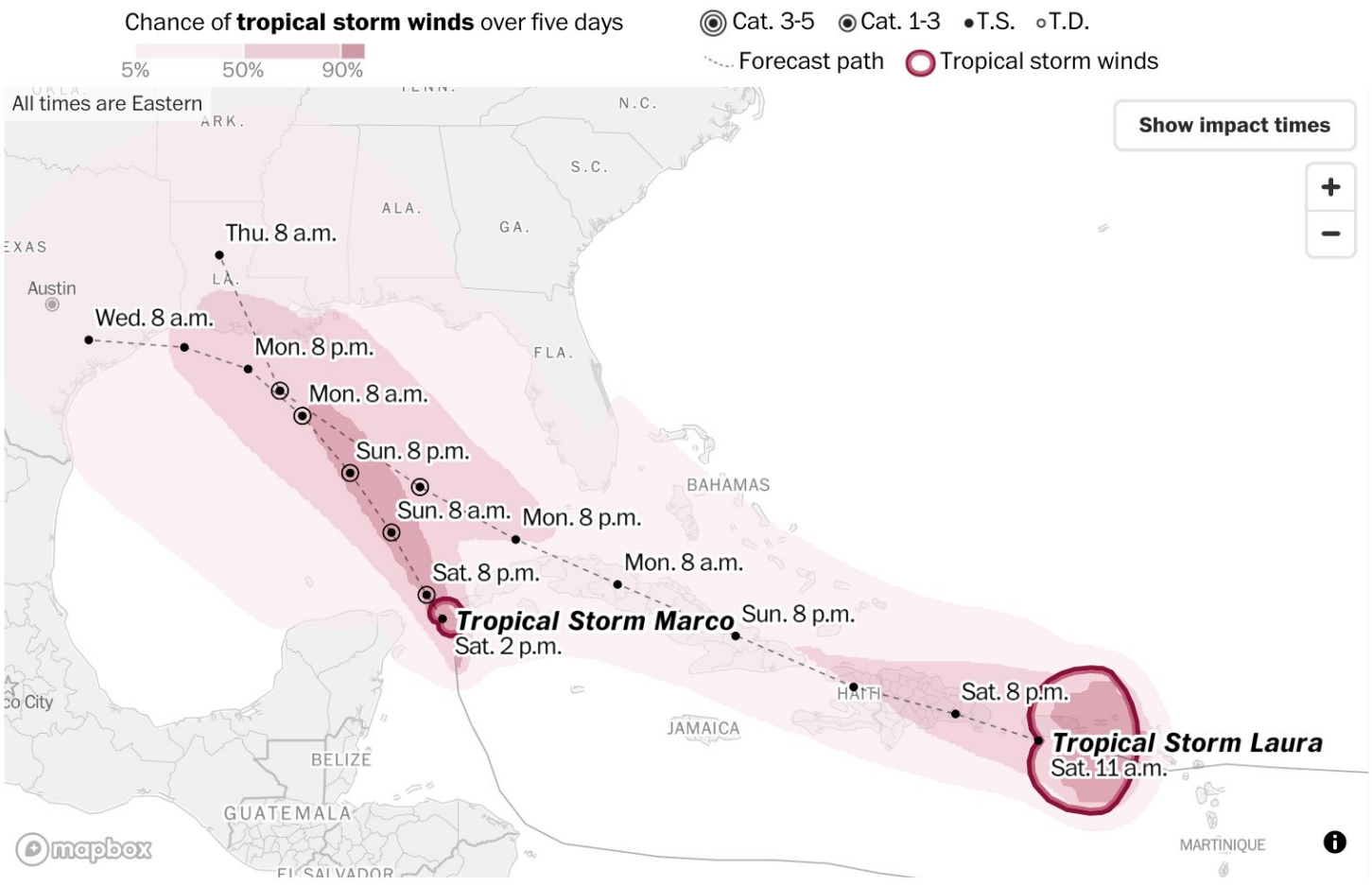
Real-world examples

Coronavirus Tracking



Real-world examples

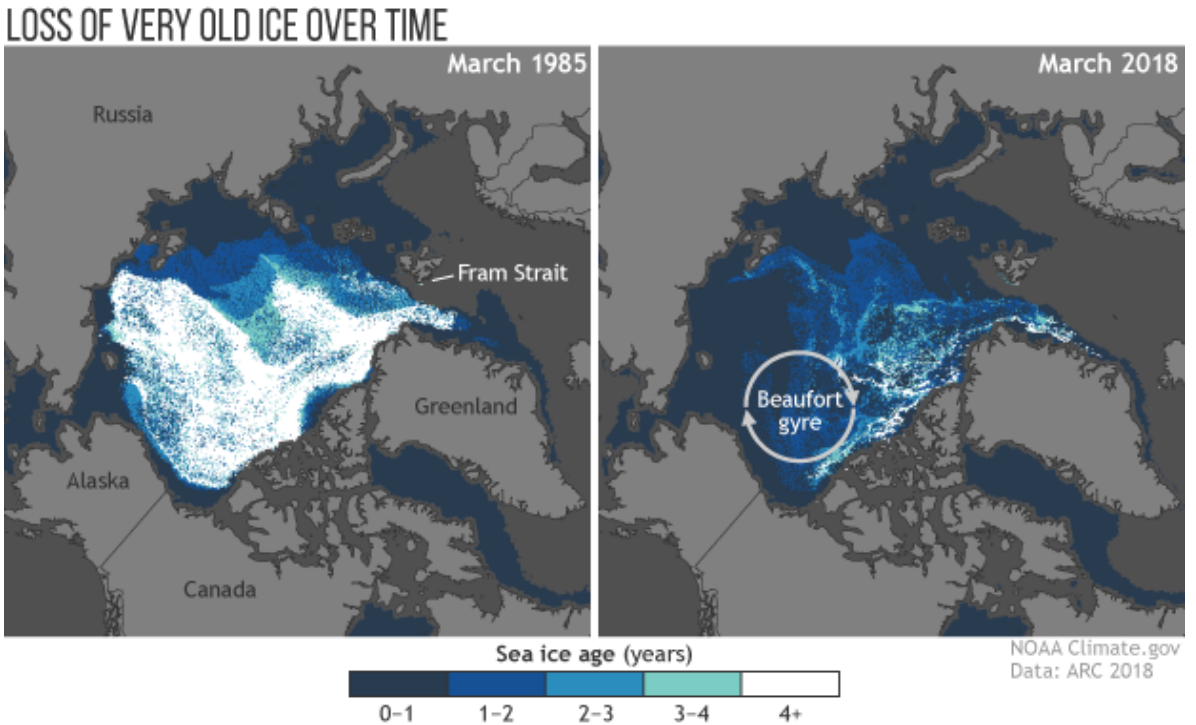
Weather Prediction



Source: National Weather Service. Note: Impact lines represent the earliest reasonable arrival time of tropical storm winds.

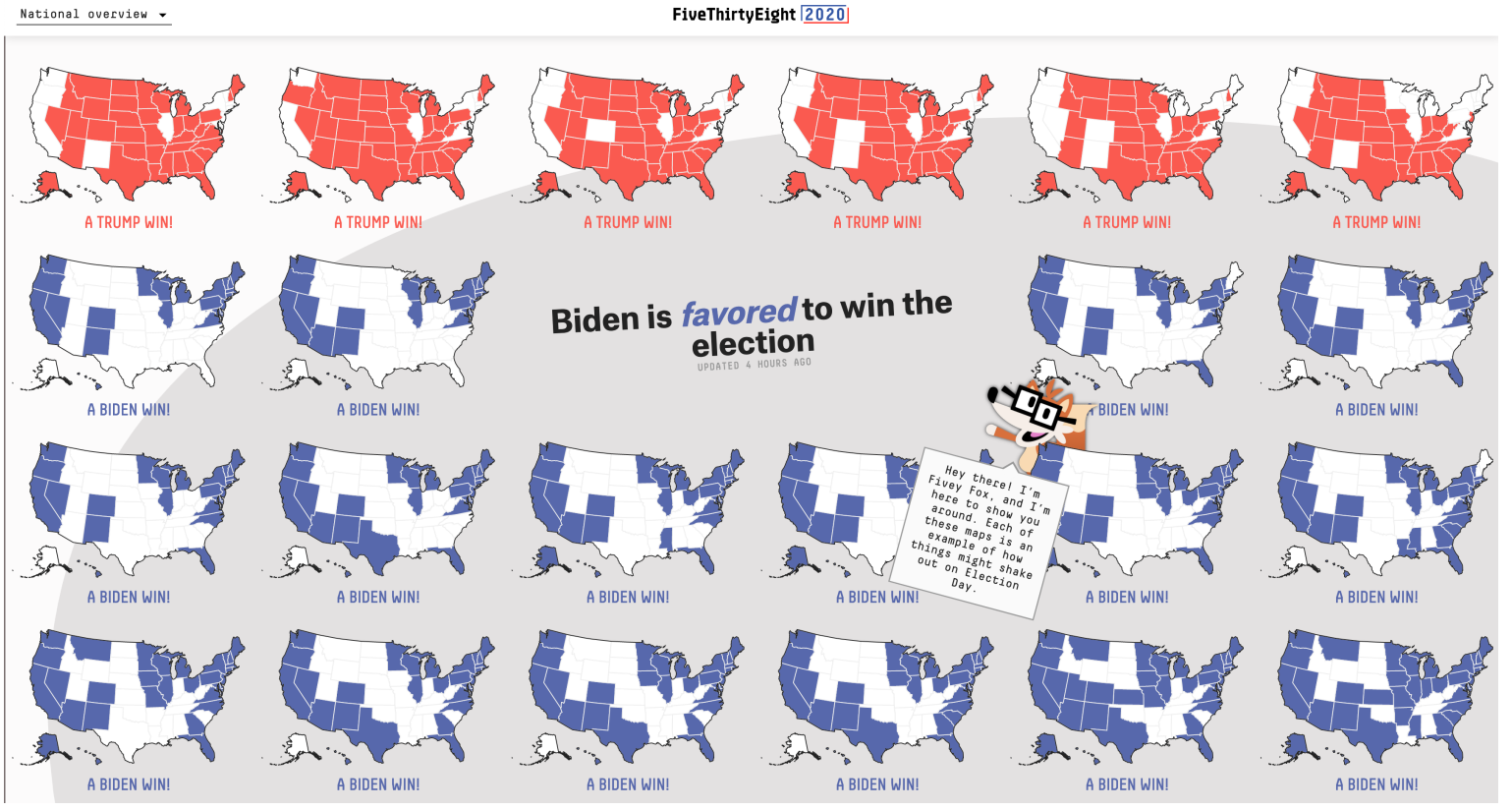
Real-world examples

Climate Change



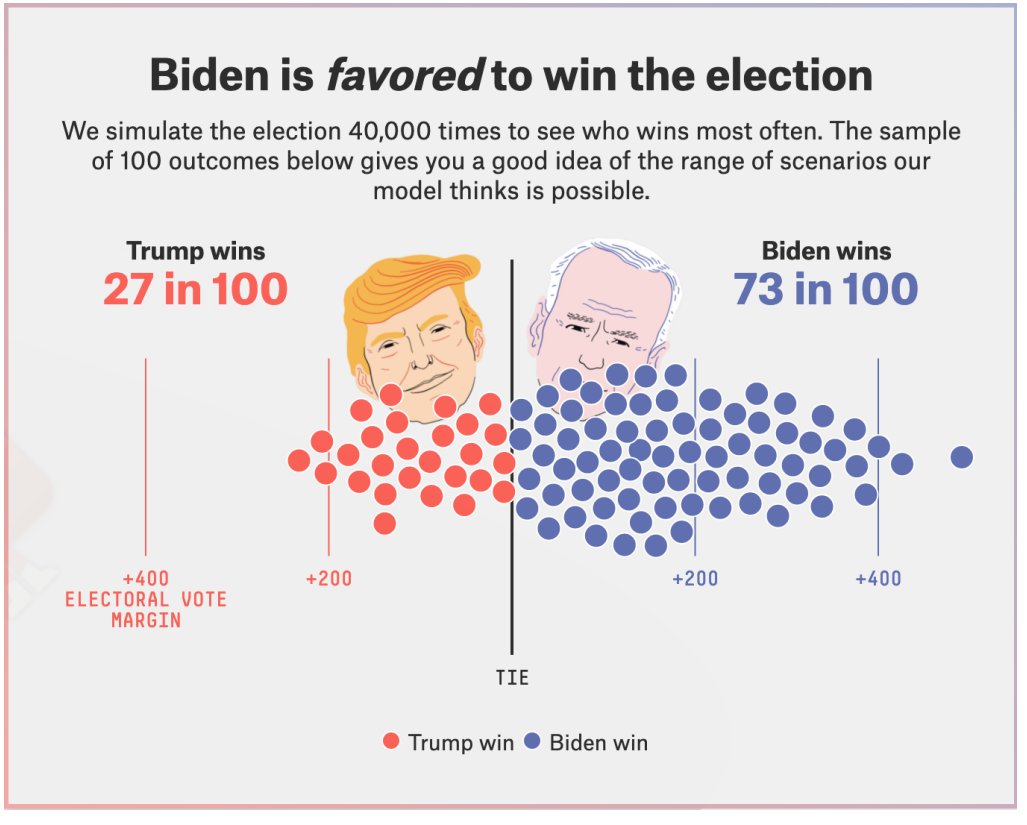
Real-world examples

Election Predictions



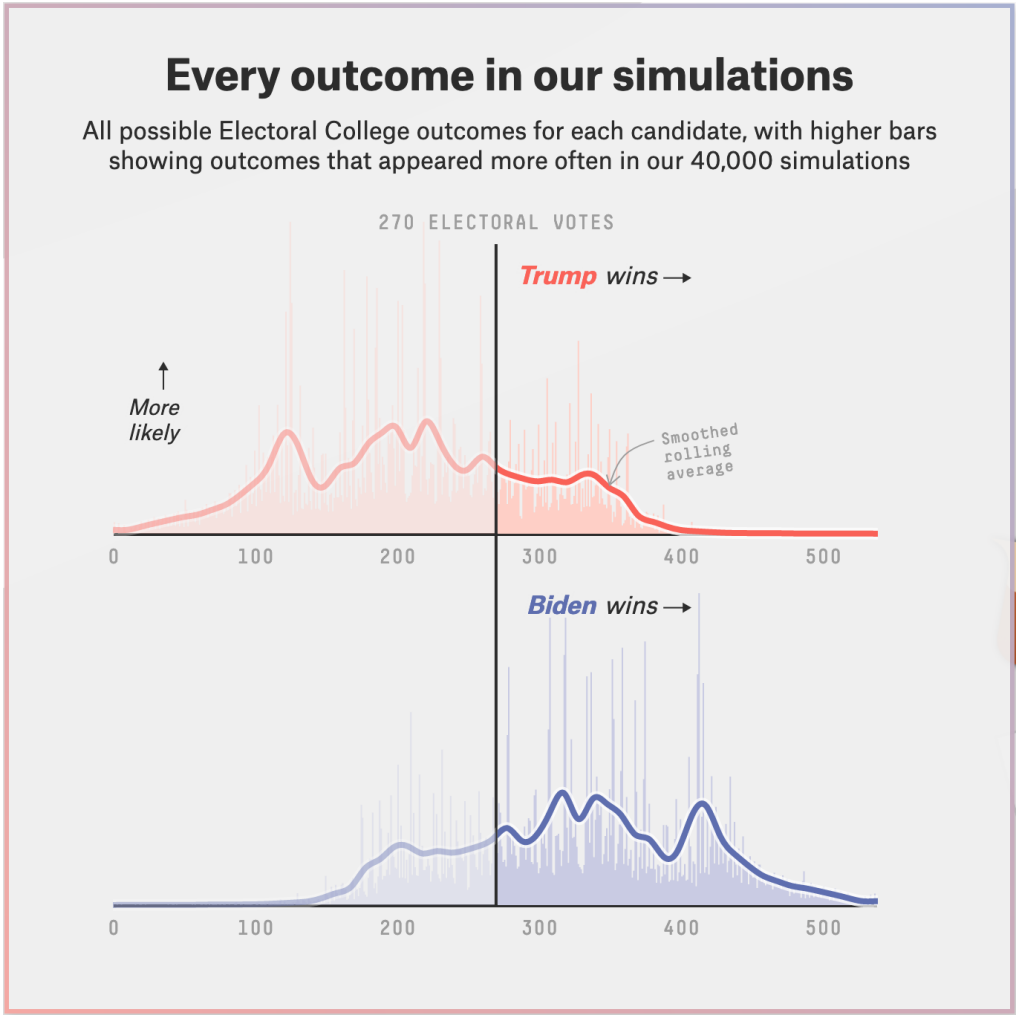
Real-world examples

Election Predictions



Real-world examples

Election Predictions



Other uses of Statistics

Statistics is used in a variety of different ways/fields

- Financial markets
- Science/medical research
- Purchasing insurance --- how risky are you to insure?
- Sports -- who do you draft?

Chapter 1: Picturing Distributions with Graphs

Statistics

Statistics: the science of data. Deals with the collection, organization, analysis, interpretation and presentation of data

- Use statistics to identify patterns and trends in the data in order to inform decision-making

Observation: an individual unit of analysis in the dataset

- ***Examples****: person, state, country, etc.*

Variable: characteristic of an observation

- ***Examples****: age, population, GDP, etc.*

NBA Salaries

PLAYER NAME	POSITION	TEAM	SALARY	CONTRACT LENGTH
Stephen Curry	Point Guard	Golden State Warriors	40200000	5 years
Russell Westbrook	Point Guard	Houston Rockets	38500000	5 years
Chris Paul	Point Guard	Oklahoma City Thunder	38500000	4 years
Lebron James	Small Forward	Los Angeles Lakers	37400000	4 years
James Harden	Shooting Guard	Houston Rockets	38200000	4 years
Kevin Durant	Small Forward	Brooklyn Nets	37200000	4 years

Type of Variables

Categorical variable: takes on a unique value for each possible category or trait

- ***Examples****: race, political party, dog breed, etc.*

Quantitative variable: measured on a numeric scale

- ex[Examples: income, unemployment rate, weight, etc.]
- Variables may be either **discrete** (countable) or **continuous** (uncountable)

NBA Salaries

PLAYER NAME	POSITION	TEAM	SALARY	CONTRACT LENGTH
Stephen Curry	Point Guard	Golden State Warriors	40200000	5 years
Russell Westbrook	Point Guard	Houston Rockets	38500000	5 years
Chris Paul	Point Guard	Oklahoma City Thunder	38500000	4 years
Lebron James	Small Forward	Los Angeles Lakers	37400000	4 years
James Harden	Shooting Guard	Houston Rockets	38200000	4 years
Kevin Durant	Small Forward	Brooklyn Nets	37200000	4 years

Clicker Question

Given the following dataset, which of these statements is correct?

Statwide Electricity Stats

STATE	YEAR	ELECTRICITY SALES	GOVERNMENT	RENEWABLE CAPACITY (MWH)
AK	2000	$5M	D	0
AL	2000	$77M	D	493
AR	2000	$36M	R	369
AZ	2000	$64M	R	1
CA	2000	$220M	D	3053
CO	2000	$47M	R	29
CT	2000	$34M	R	262

a. Electricity sales, renewable capacity and state are all quantitative variables

b. Government, state are both categorical variables

c. All variables are categorical

d. All variables are quantitative

Dummy Variables

Often time in datasets **dummy variables**, or *indicator variables*, are used to describe categorical variables.

- ***Example**: the "Government" variable as 0 for D and 1 for R.*

Sometimes Dummy/indicator variables put observations into categories, even though they are numerical in value

- ***Example**: Years of schooling into "HS Degree" dummy (years $ >12 $)*

Distribution of a Variable

Distribution of a variable: tells us *what values* it takes and *how often* it takes these values

- lists all possible outcomes of variable and their associated frequencies

Statwide Electricity Stats

STATE	YEAR	ELECTRICITY SALES	GOVERNMENT	RENEWABLE CAPACITY (MWH)
AK	2000	$5M	D	0
AL	2000	$77M	D	493
AR	2000	$36M	R	369
AZ	2000	$64M	R	1
CA	2000	$220M	D	3053
CO	2000	$47M	R	29
CT	2000	$34M	R	262

What is the distribution of Government?

Distribution of a Variable

Statwide Electricity Stats

STATE	YEAR	ELECTRICITY SALES	GOVERNMENT	RENEWABLE CAPACITY (MWH)
AK	2000	$5M	D	0
AL	2000	$77M	D	493
AR	2000	$36M	R	369
AZ	2000	$64M	R	1
CA	2000	$220M	D	3053
CO	2000	$47M	R	29
CT	2000	$34M	R	262

Distribution of Government: D - 3/7 and R - 4/7

Visualizing Categorical Variable

Distribution of categorical variable lists the categories and gives **the count/percent** of individuals who fall into each category.

- Often visualize distributions of categorical variables using **pie charts** or **bar graphs**.

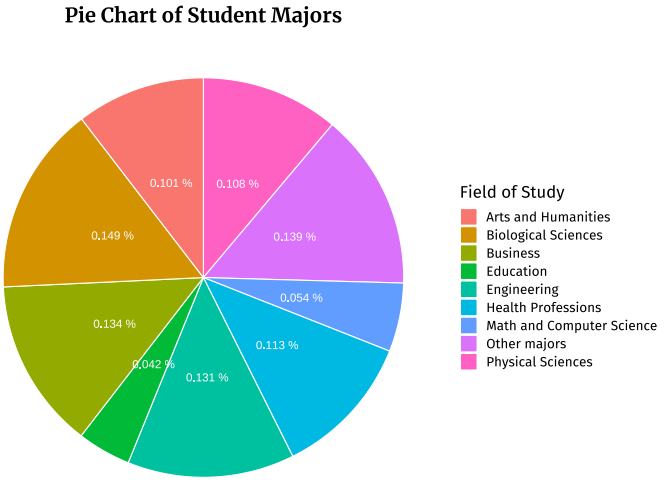
Examples

Distribution of CU Boulder Students

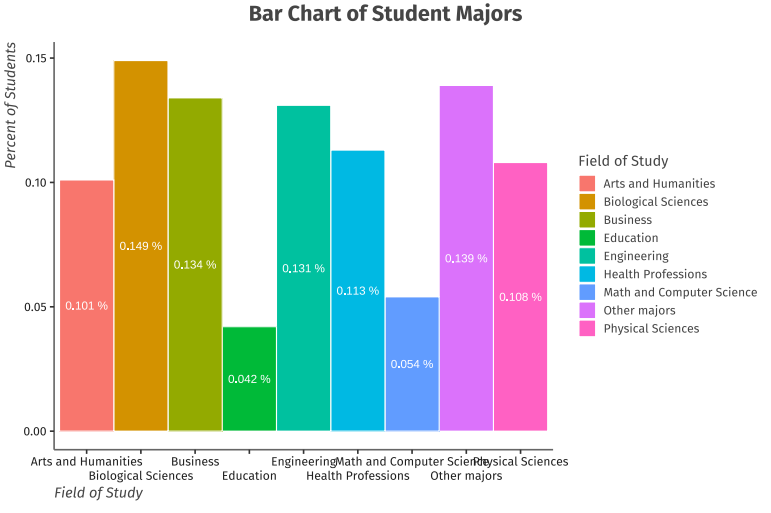
FIELD OF STUDY	PERCENT OF STUDENTS
Arts and Humanities	10.1%
Biological Sciences	14.9%
Business	13.4%
Education	4.2%
Engineering	13.1%
Health Professions	11.3%
Math and Computer Science	5.4%
Physical Sciences	10.8%
Other majors	13.9%

Examples

Pie Chart



Bar Chart



Visualizing Continuous Variable

Distribution of a variable: tells us *what values* it takes and *how often* it takes these values

- Often visualize distributions of continuous variables using **histograms**, **stemplots**, or **time plots** if variable is measured over time

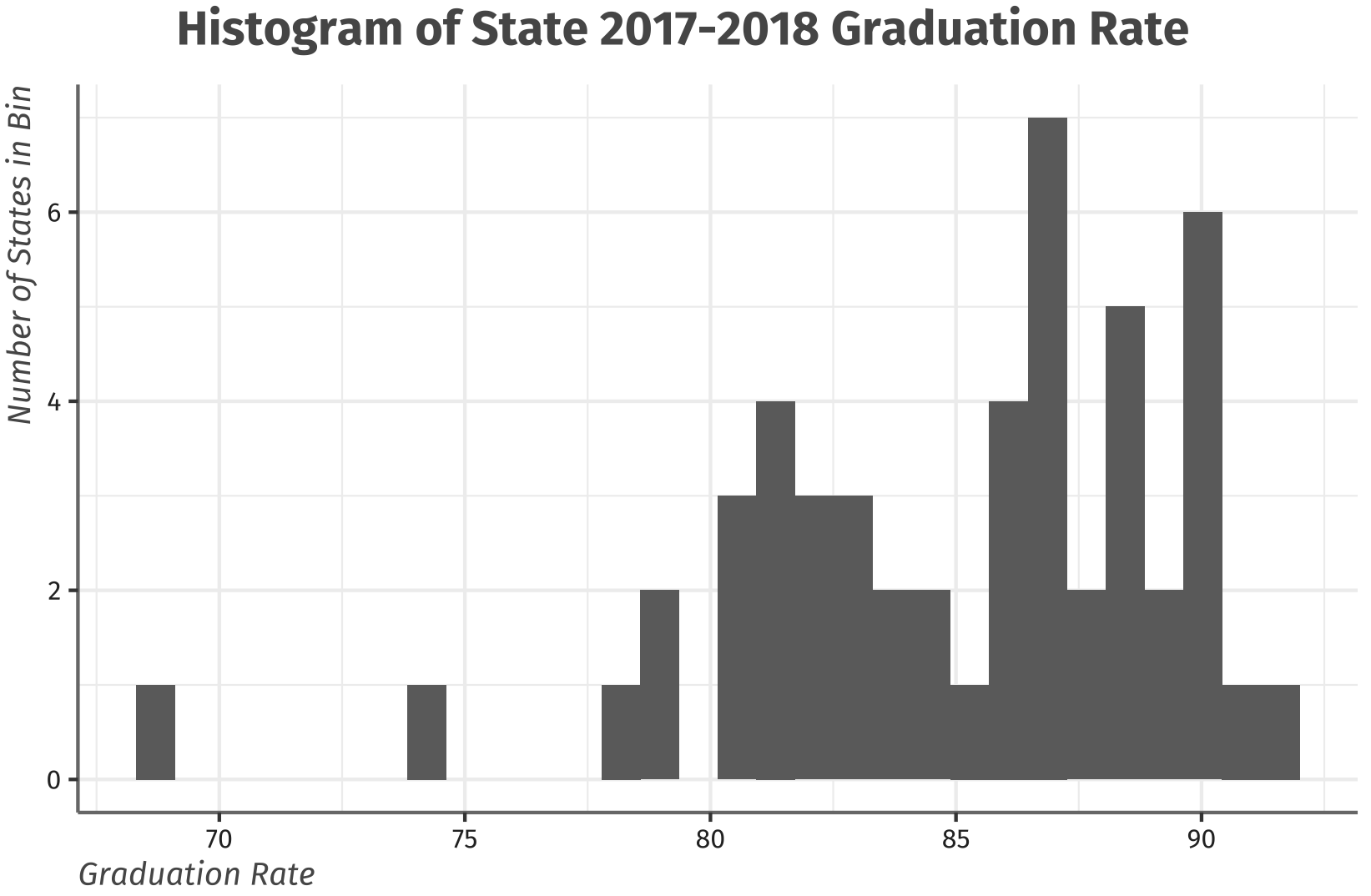
Histogram

- A **histogram** shows the distribution of a continuous variable by using bars whose height represents number of individuals who take on a value within a particular interval (bin)
    - Appropriate for variables that take on many different values or have large number of observations
- To make a histogram:
    - Divide the possible values into intervals (bins) of equal widths
    - Count how many observations fall into each interval (bin)
    - For each interval, draw a bar whose height is equivalent to the number (or percent) of observations in each interval

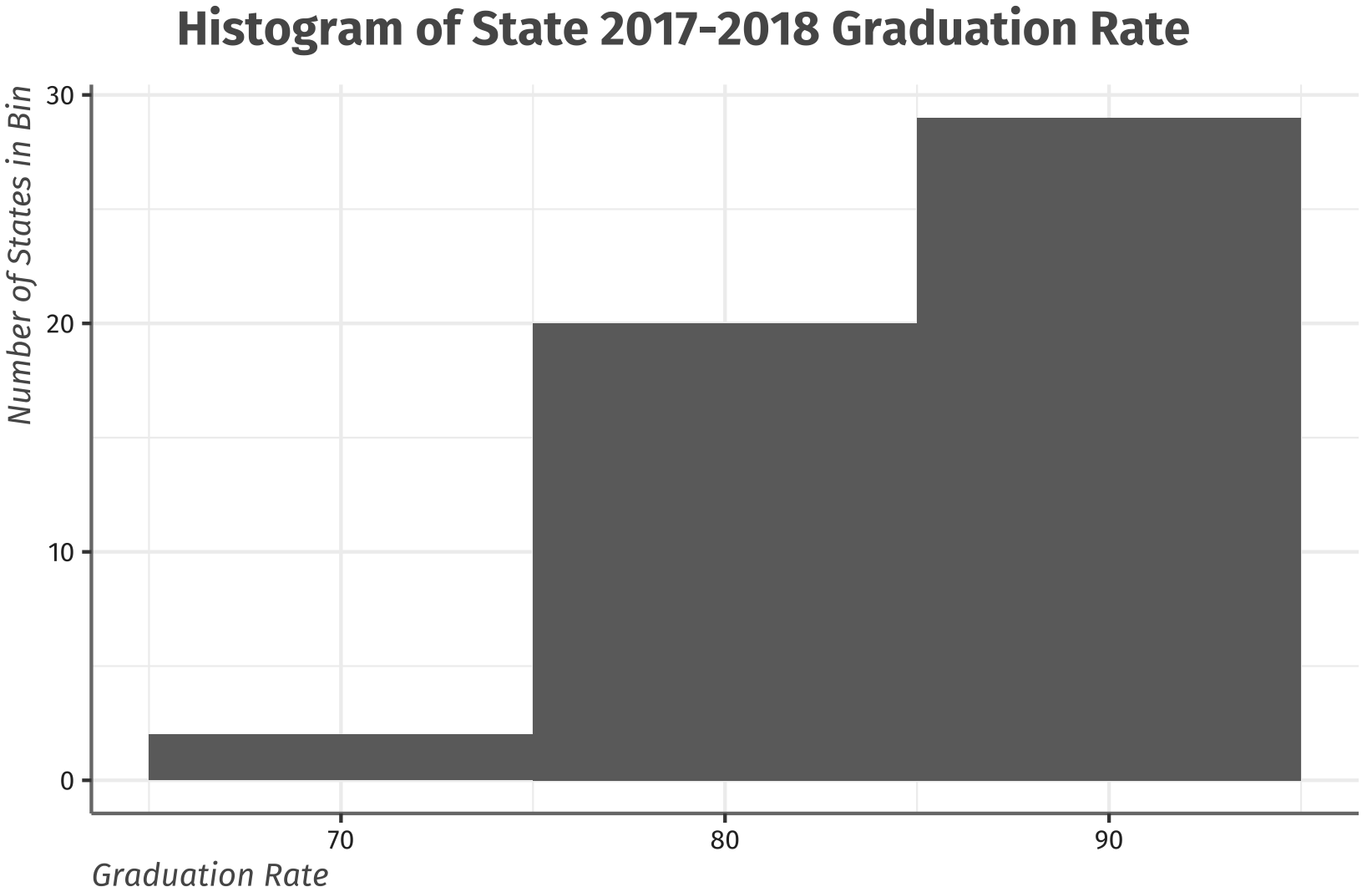
State-level Graduation Rates

	STATE	GRADUATION RATE
1	Alabama	90.0%
2	Alaska	78.5%
3	Arizona	78.7%
4	Arkansas	89.2%
5	California	83.0%
6..50		
51	Wyoming	81.7%

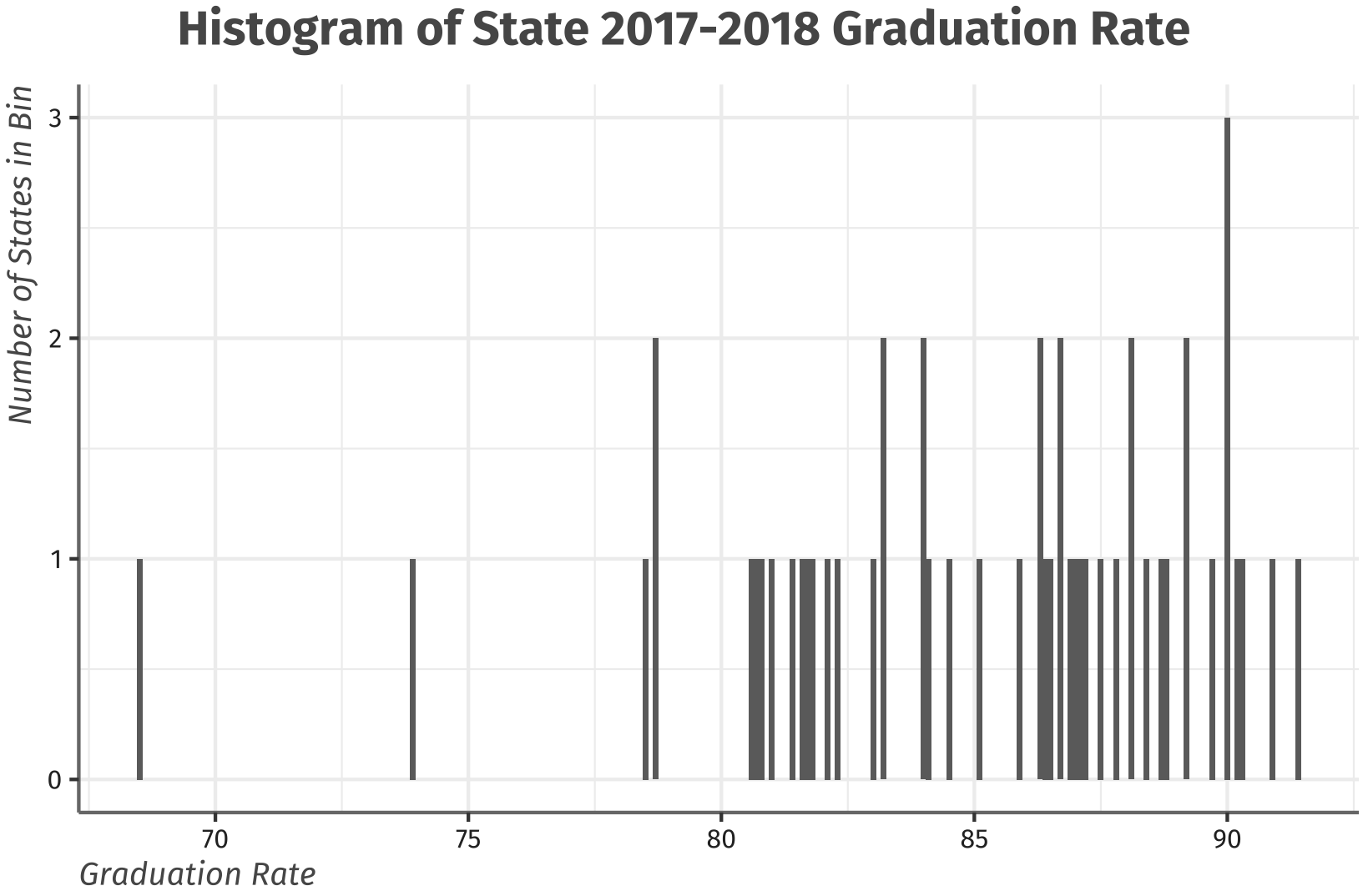
Graduation Rates



Less Informative



Too Many Bins



Interpreting Histogram

How to interpret histograms:

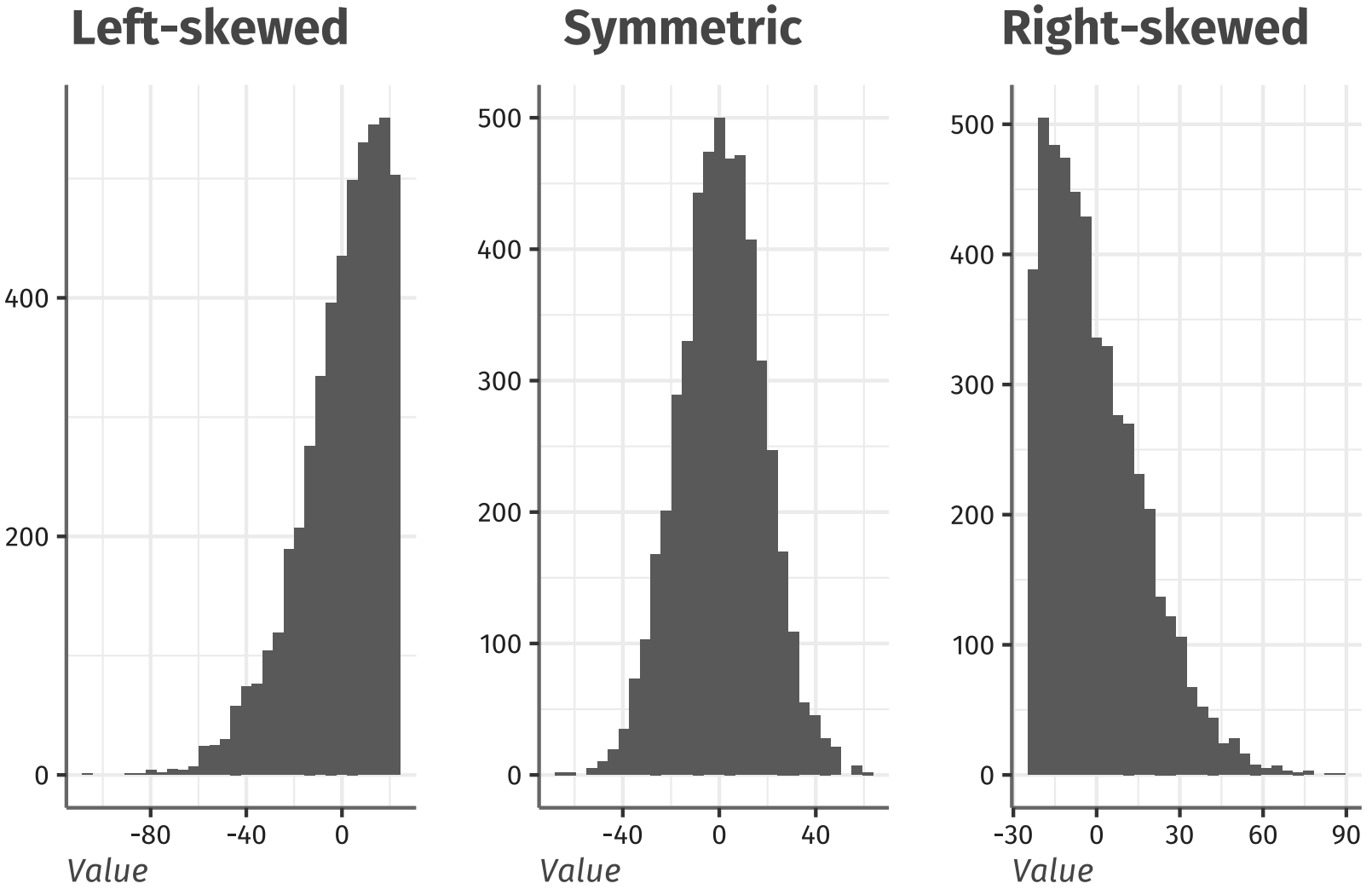
- Look for overall pattern and striking deviations from that pattern
    - An important kind of deviation is an **outlier**, an observation that falls *outside the overall pattern*
- Describe the pattern by its **shape**, **center**, and **variability** (or spread)

Shapes of Distributions

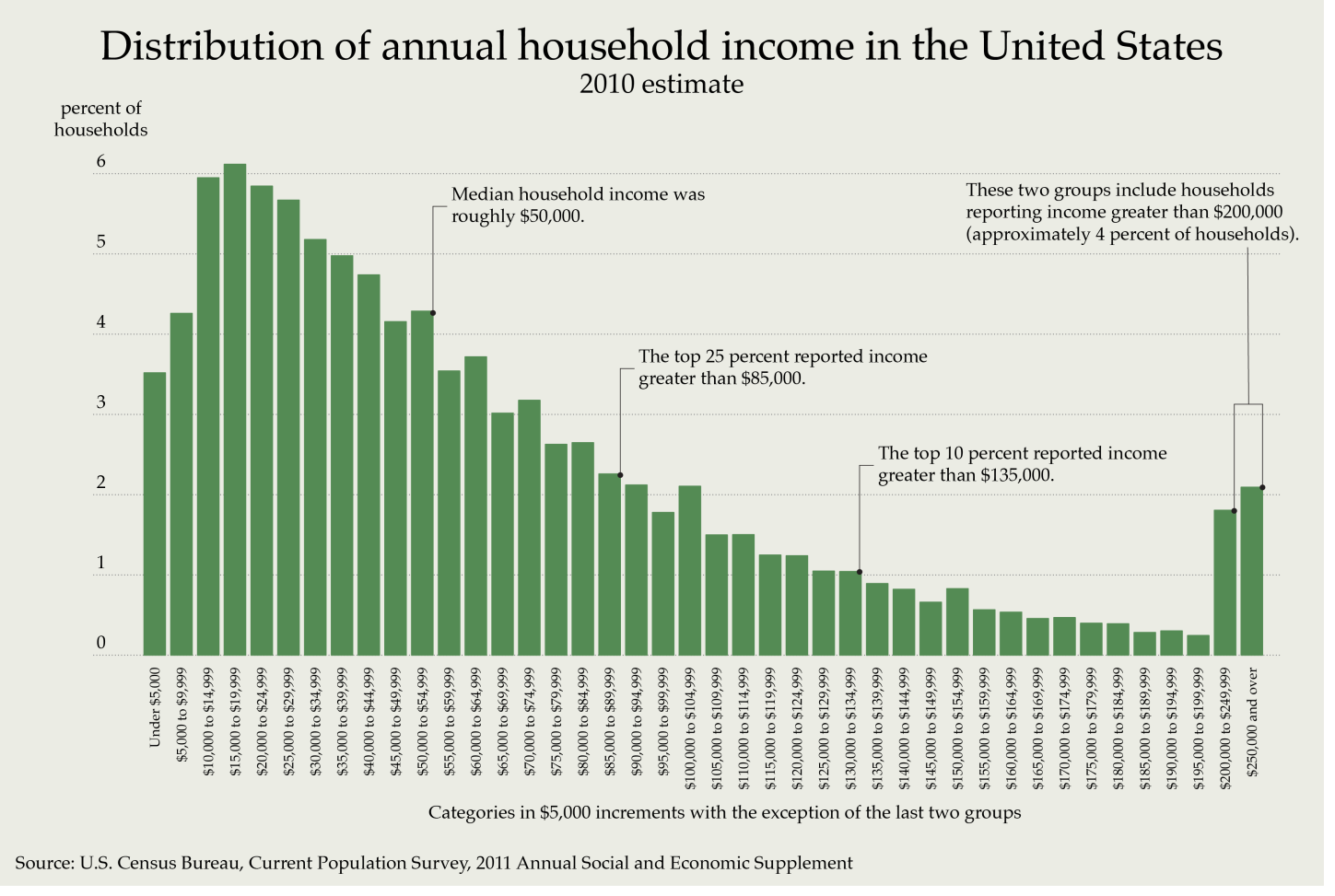
We describe the shape of the distribution as

- symmetric: the right and left sides of the graph are approximately mirror images of each other
- right-skewed: the right side of the graph (containing the half of the observations with larger values) is much longer than the left side
- left-skewed: the left side of the graph is much longer than the right side

Skewness Examples



Clicker Question



a. symmetric b. left-skewed c. right-skewed

Clicker Question

For which of the following variables would you need to use a histogram instead of a bar graph?

a. month of birth

b. distance from nearest metropolitan area

c. employment status

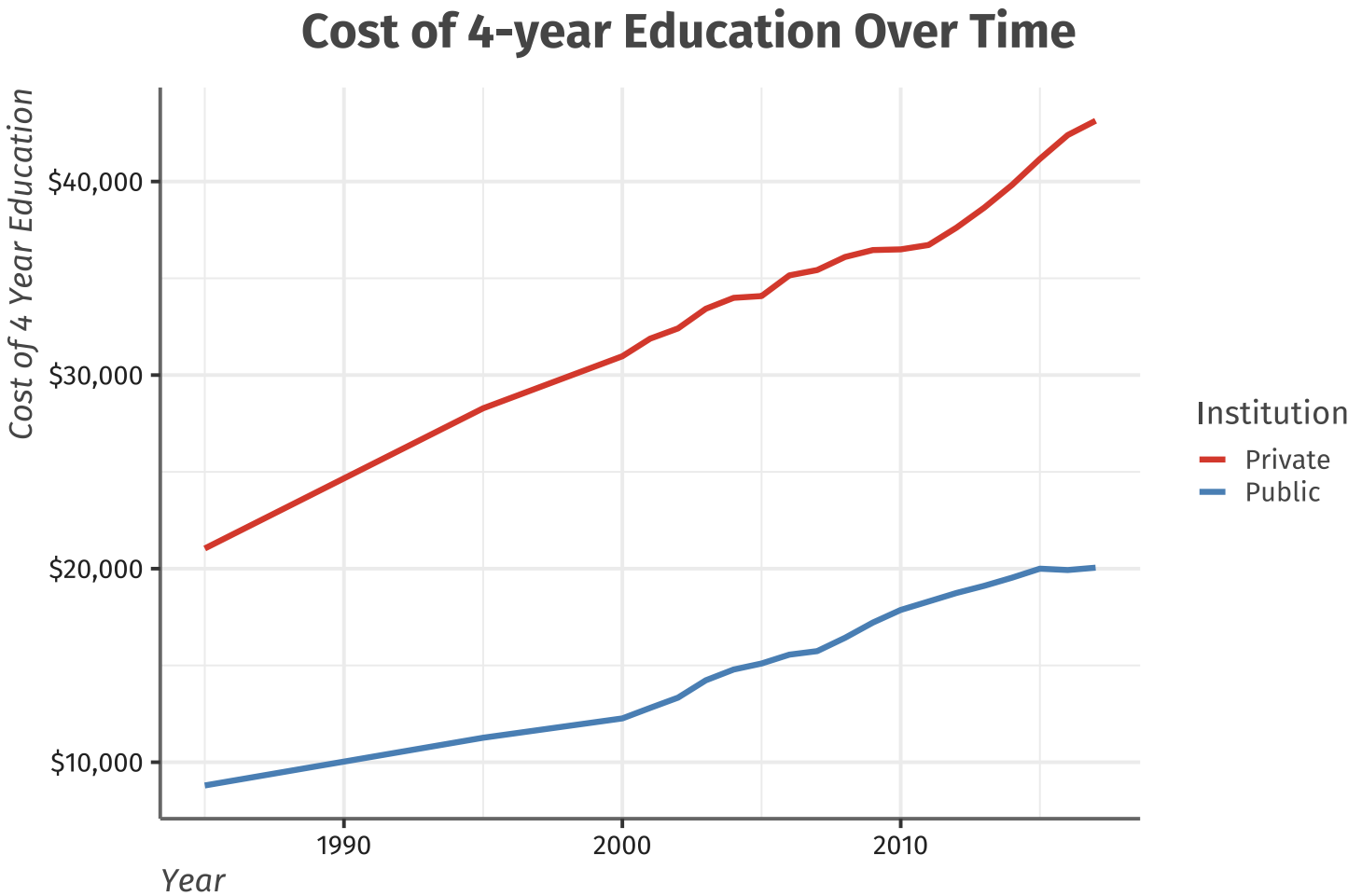
d. none of the above

Time Plots

Time Series is a connected line plotting the value of the variable over time

- Shows behavior over time which emphasizes *trends*
- Time is always on the **horizontal** axis, variable being measured on **vertical** axis
- Shows *trends* and *deviations from trends*
    - Also want to look for seasonal variation

Time Series Plots



Deviation from Trends