Representing data

Experimental Research

Professor Velez (11/1/22)

Recap

- Learned fundamentals regarding experimental design
- Designed and analyzed a non-human subjects experiment
- Created an intervention that we will examine in the context of a class-wide experiment
- We are currently waiting for the IRB review

This week

- Revisit concepts we covered in the measurement week
- Create summaries and data visualizations
- **Crucial** early step in the research pipeline
  - Identity errors in coding
  - Highlight expected patterns
  - Uncover unexpected patterns

Summarizing data

- We've already been working with data summaries
- Means and standard deviations
  - Capture the central tendency of a variable
  - Capture the typical observed variation
- Statistics help simplify the fairly complex information we collect in the social sciences

Example

Find every ‘5’

```
91768679919603862559304865514 43
93536525027521413949126687660 13
50954446634735459936046364480 78
64256201711559067101210692187 01
05844482055613859025035730688 45
11721023918797942388926697645 99
69872279510159489260177591666 04
84362040365270292917077427170 51
32805064285531588631368683804 21
40551359067703297836717483928 74
```

Asking questions about our data

- How is the variable distributed?
  - What is the most likely value for this variable? **Mean**
  - What is the value at the midpoint? **Median**
  - What is the most common value? **Mode**
  - Are observations dispersed across the range of the variable? **Variance/Standard Deviation**
- How are variables related to each other?
  - When X is high (low), is Y also high (low)? **Covariance/correlation**

Different averages

- Mean
  - Sum of values divided by number of values
  - Sensitive to outliers
- Median
  - Midpoint of a variable (50th percentile)
  - Less sensitive to outliers
- Mode
  - Most common value
  - Can be helpful when working with nominal data

Polling Data					
Poll	Date	Sample	Republicans (R)	Democrats (D)	Spread
RCP Average	**10/12 - 10/28**	**--**	**48.0**	**45.1**	**Republicans +2.9**
CBS News Battleground Tracker	10/26 - 10/28	1899 LV	47	45	Republicans +2
Data for Progress (D)**	10/26 - 10/27	1217 LV	49	45	Republicans +4
Trafalgar Group (R)	10/25 - 10/27	1089 LV	48	42	Republicans +6
Rasmussen Reports	10/23 - 10/27	2500 LV	49	42	Republicans +7
InsiderAdvantage	10/25 - 10/25	750 LV	48	44	Republicans +4
Economist/YouGov	10/22 - 10/25	1114 LV	45	49	Democrats +4
Politico/Morning Consult	10/21 - 10/23	2005 RV	42	47	Democrats +5
USA Today/Suffolk	10/19 - 10/24	1000 LV	49	45	Republicans +4
Democracy Corps (D)	10/19 - 10/23	LV	50	48	Republicans +2
Emerson	10/18 - 10/19	1000 LV	46	41	Republicans +5
NBC News	10/14 - 10/18	LV	48	47	Republicans +1
Monmouth	10/13 - 10/17	756 RV	50	44	Republicans +6
CNBC	10/13 - 10/16	800 RV	48	46	Republicans +2
Harvard-Harris	10/12 - 10/13	LV	53	47	Republicans +6
All 2022 Generic Congressional Vote Polling Data					

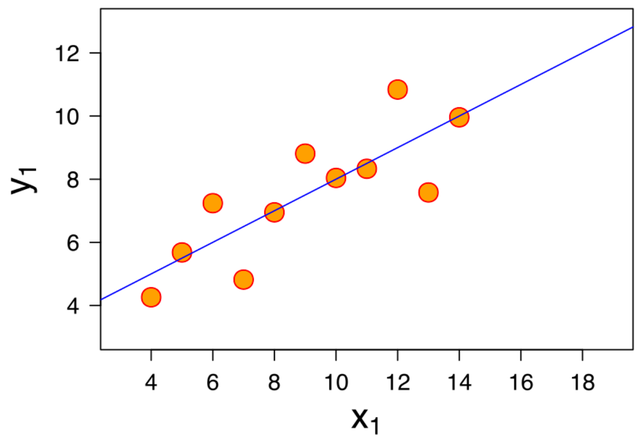
Sometimes, summary statistics obscure patterns that are hidden in the data

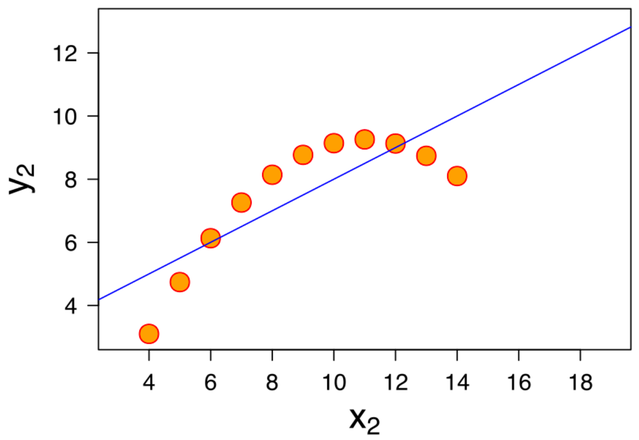
Example

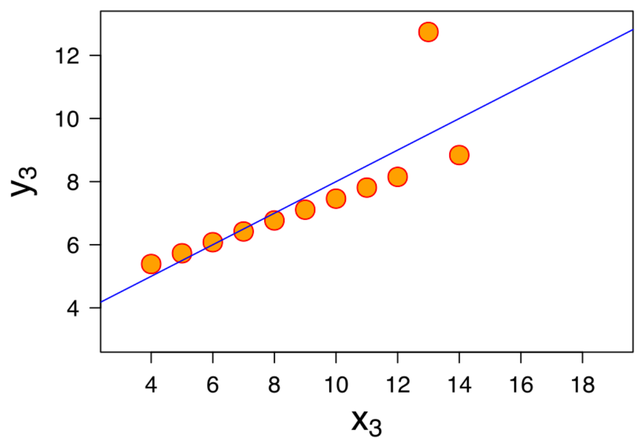
- Anscombe's Quartet
  - Four data sets
  - Near identical on several metrics

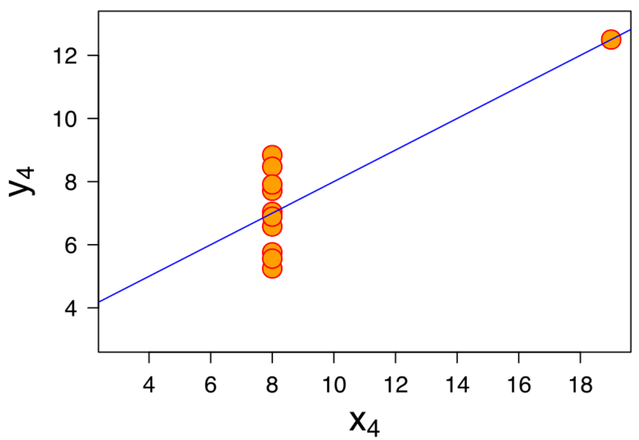
Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

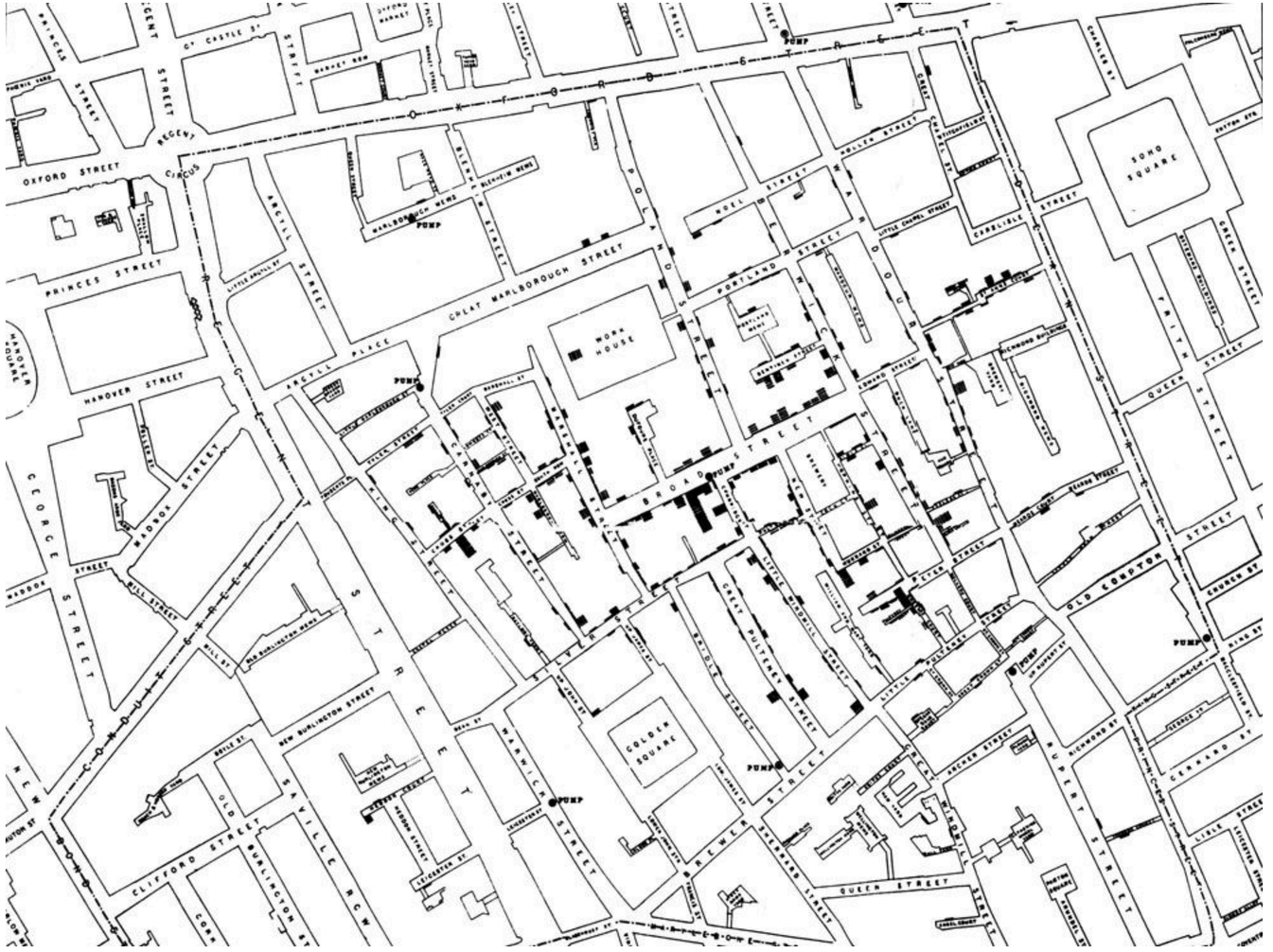
Example


12
10
8
6
4
y1
4 6 8 10 12 14 16 18
x1


12
10
8
6
4
y2
4 6 8 10 12 14 16 18
x2


12
10
8
6
4
y3
4 6 8 10 12 14 16 18
x3

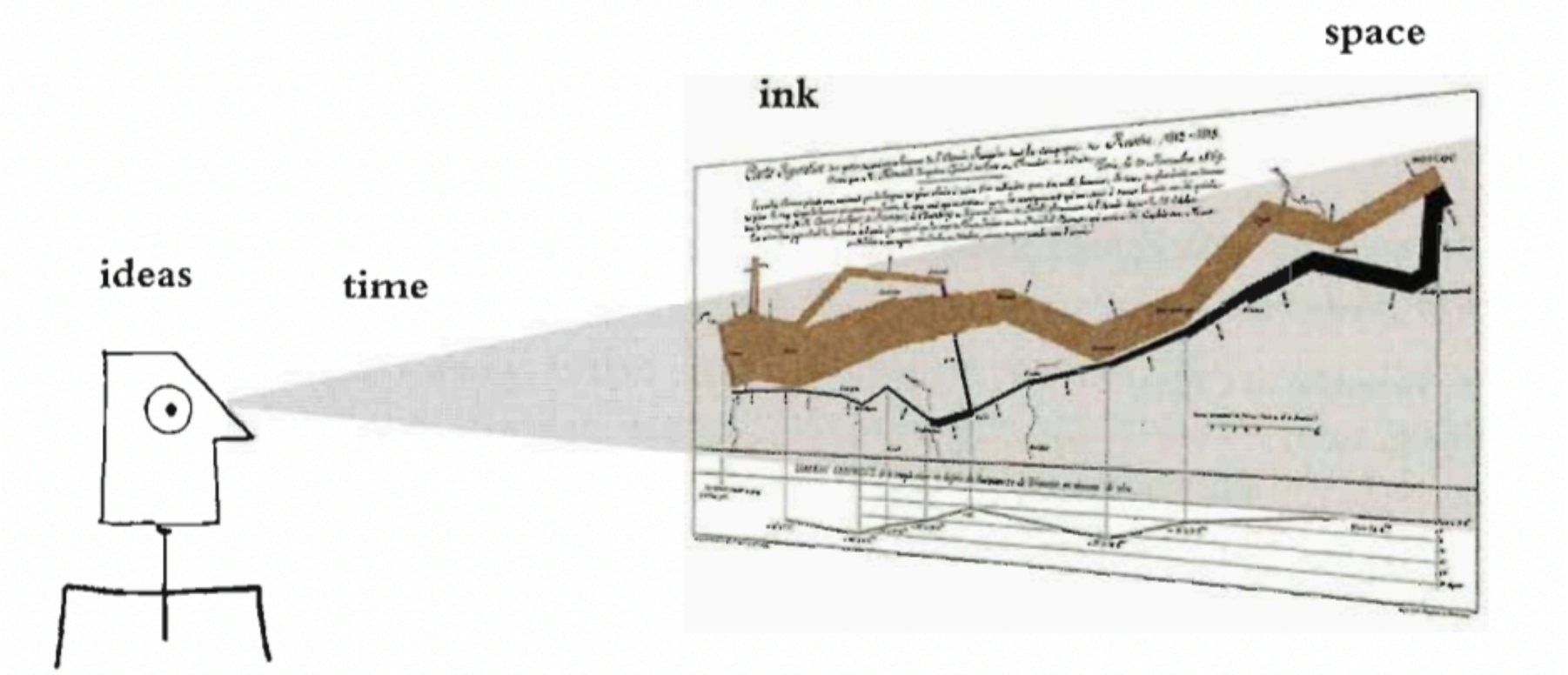

12
10
8
6
4
y4
4 6 8 10 12 14 16 18
x4


OXFORD STREET
REGENT STREET
PRINCES STREET
HANOVER STREET
GEORGE STREET
MADDOX STREET
ARGYLL PLACE
GREAT MARLBOROUGH STREET
POLAND STREET
BROAD STREET
WORK HOUSE
BERWICK STREET
GOLDEN SQUARE
WARWICK STREET
NEWEL STREET
QUEEN STREET
OLD COMPTON STREET
SOHO SQUARE
CARLISLE STREET
FRITH STREET
RUPERT STREET
CLIFFORD STREET
NEW BURLINGTON STREET
SAVILLE ROW
PUMP

Data visualization

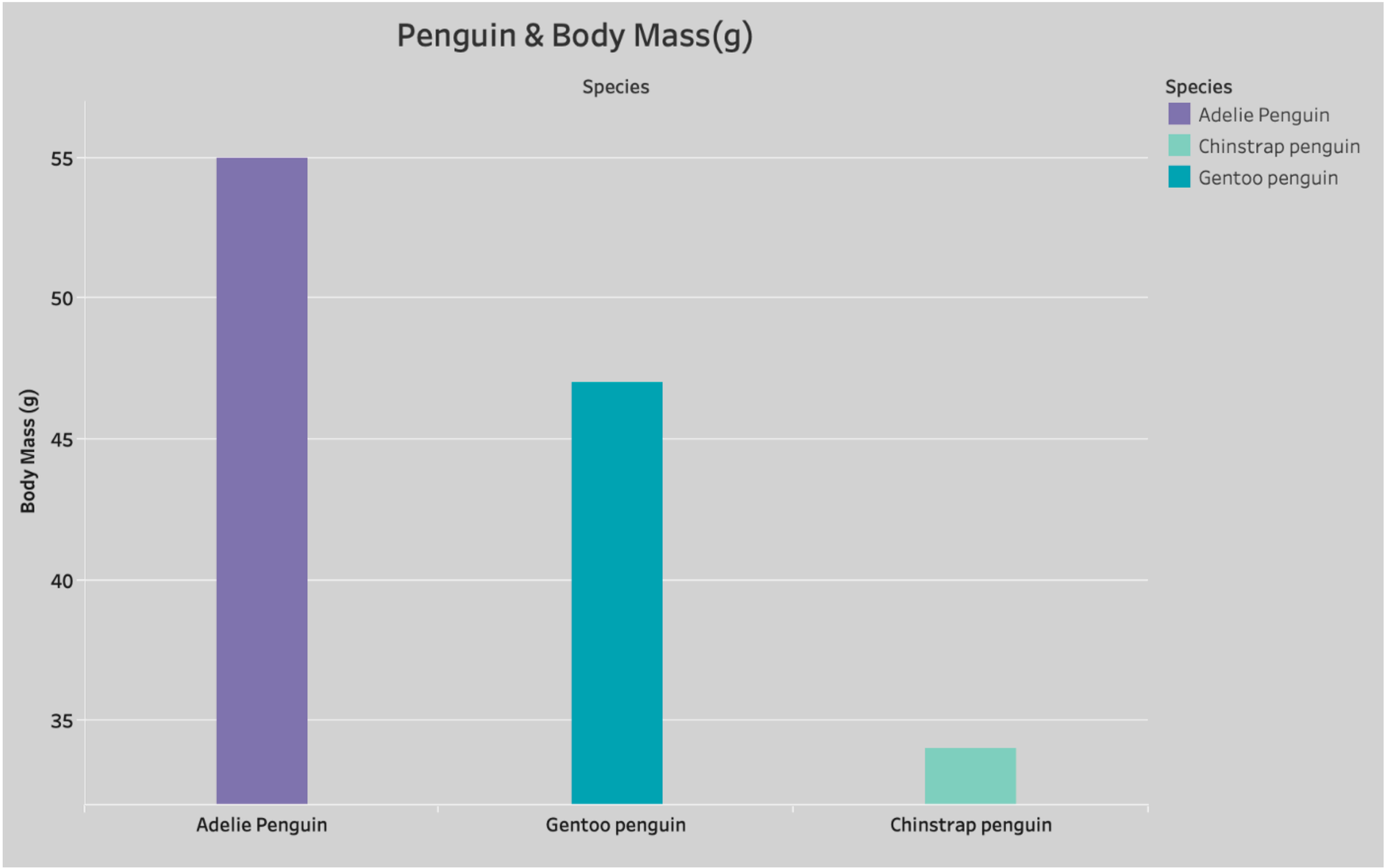
- What is data vis?
  - A visual representation of information
  - Graphics *reveal* data - Edward Tufte
- Principles of graphical excellence
  - Communicate complex ideas with clarity, precision, and efficiency
  - Viewer should get the greatest number of ideas in the shorter amount of time with the least ink in the least amount of space

Data visualization

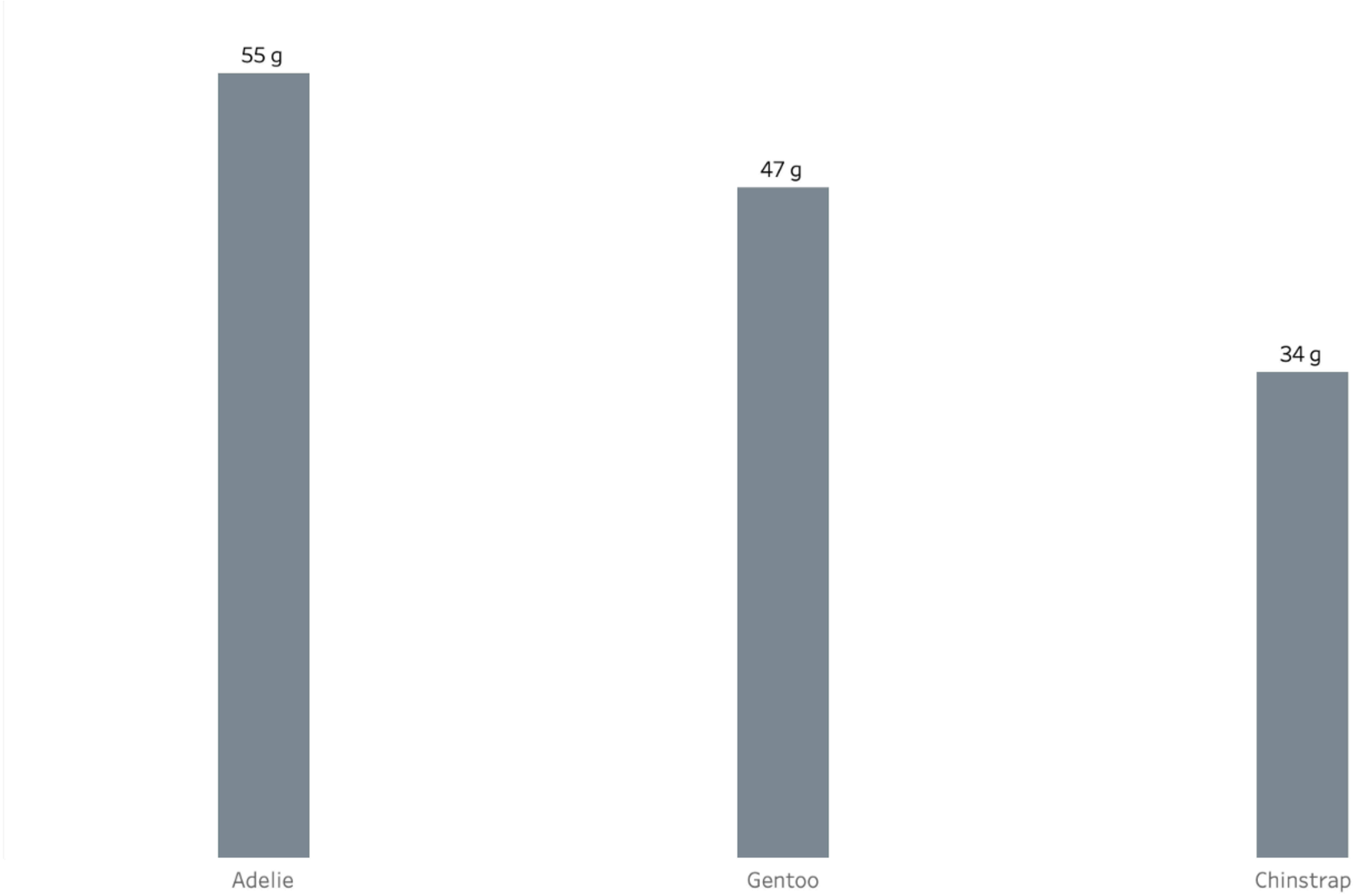


Data-ink Ratio

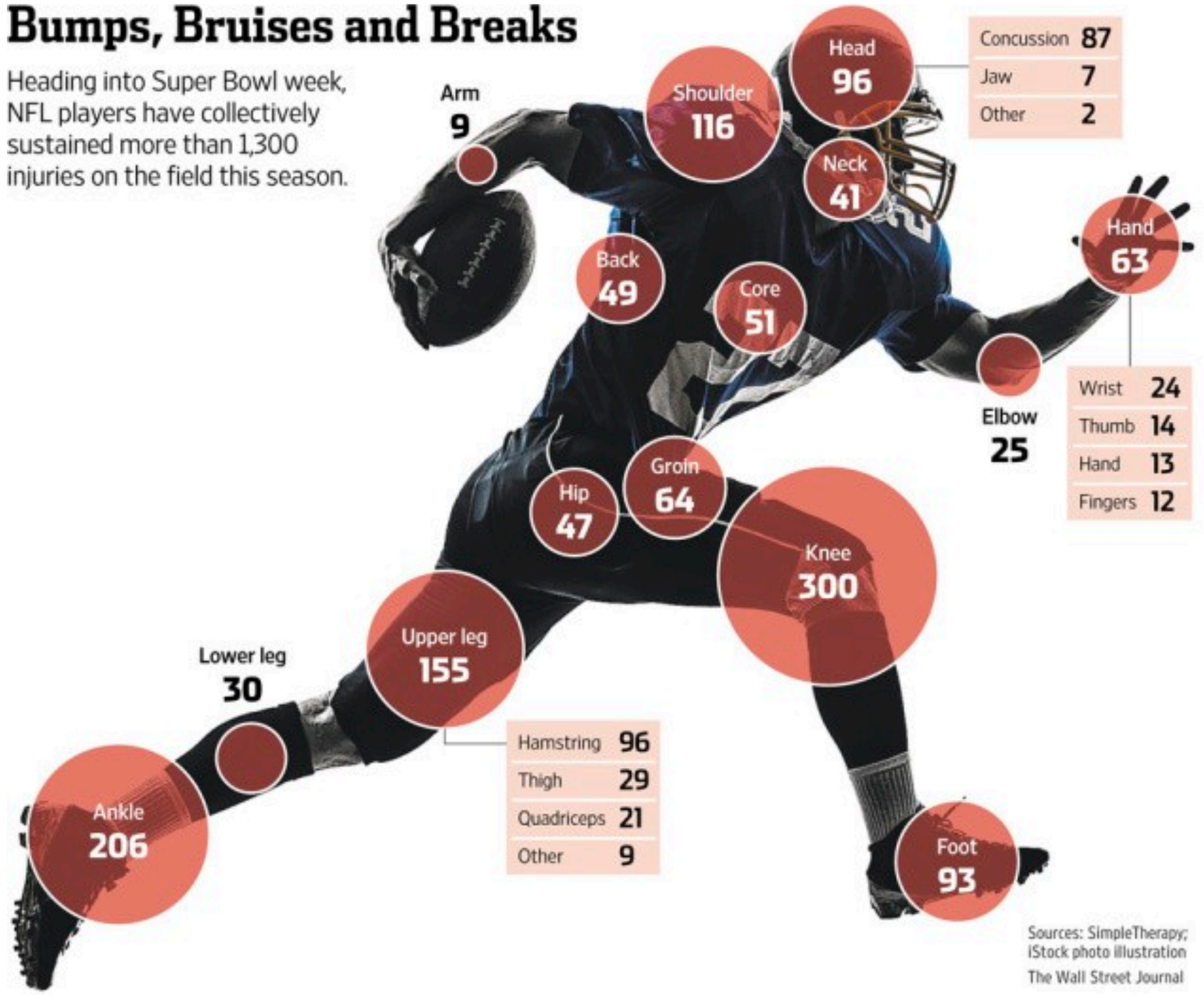
- Tufte argues that good data vis should maximize of the data-ink ratio
  - Data-ink: portion of ink that we can attribute to the data itself
  - Total-ink: comprised of data-ink and non-data ink.
  - Given today's technology, we can replace ink with pixels
- This is a principle of parsimony. Only visualize what is essential.


Penguin & Body Mass(g)
Species
Species
Adelie Penguin
Chinstrap penguin
Gentoo penguin
Body Mass (g)
55
50
45
40
35
Adelie Penguin
Gentoo penguin
Chinstrap penguin

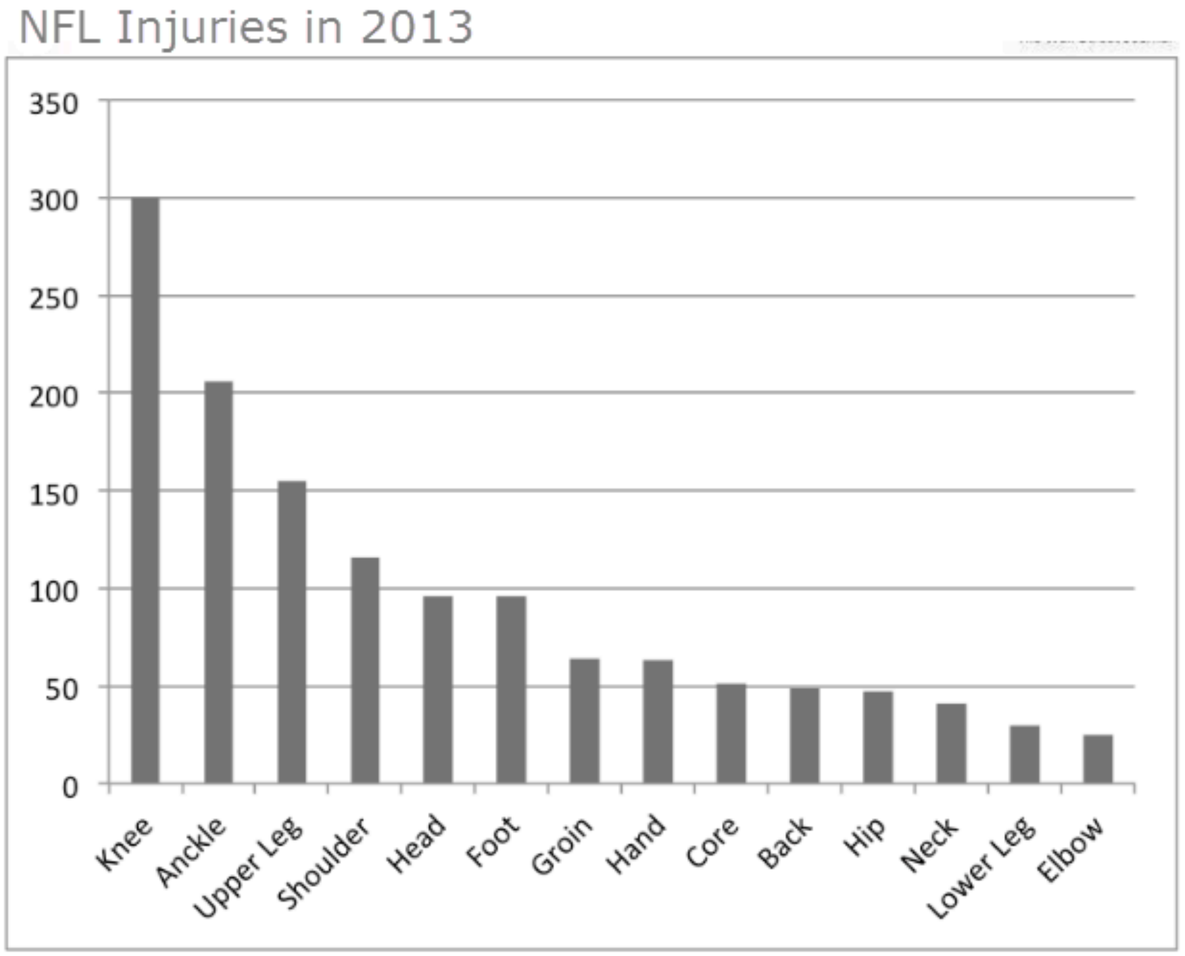
Penguin Species & Body Mass



Minimize “chart junk”



Minimize “chart junk”



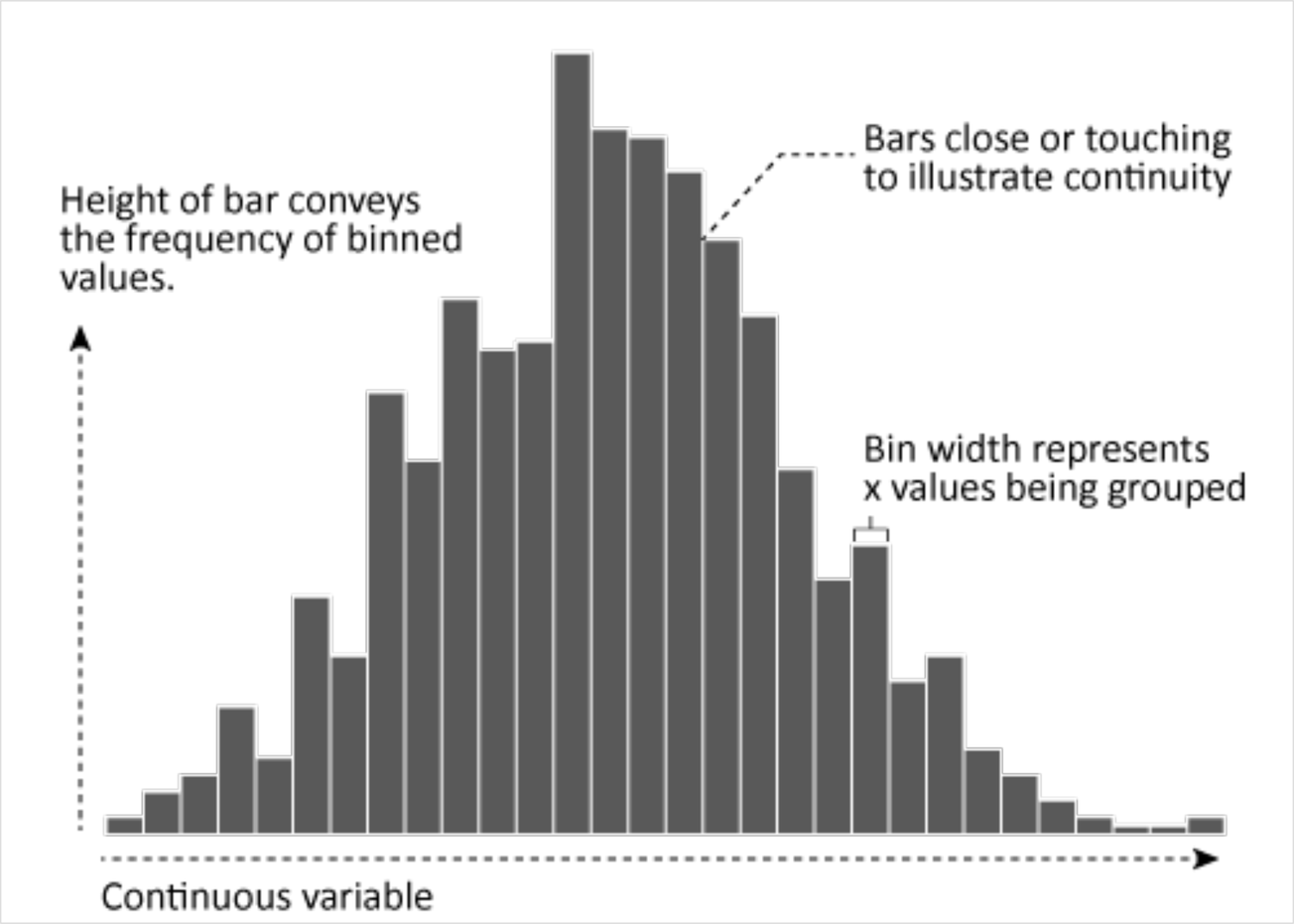
Different Visualizations

- Now that we have some criteria to evaluate visualizations, what do different visualizations look like? What do they accomplish?

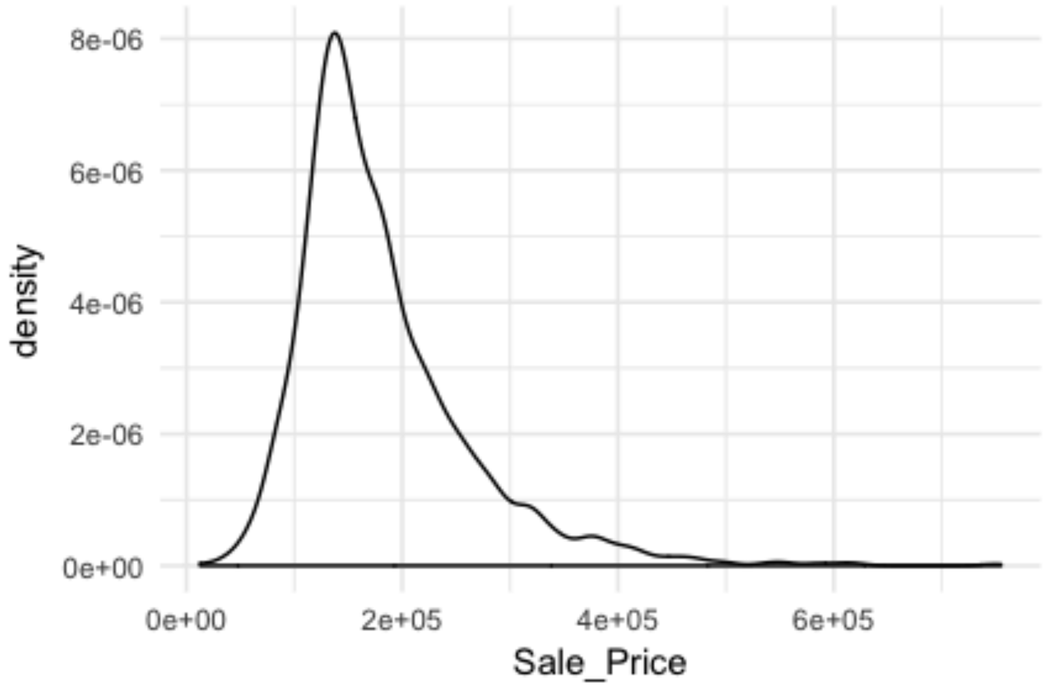
Univariate Plots

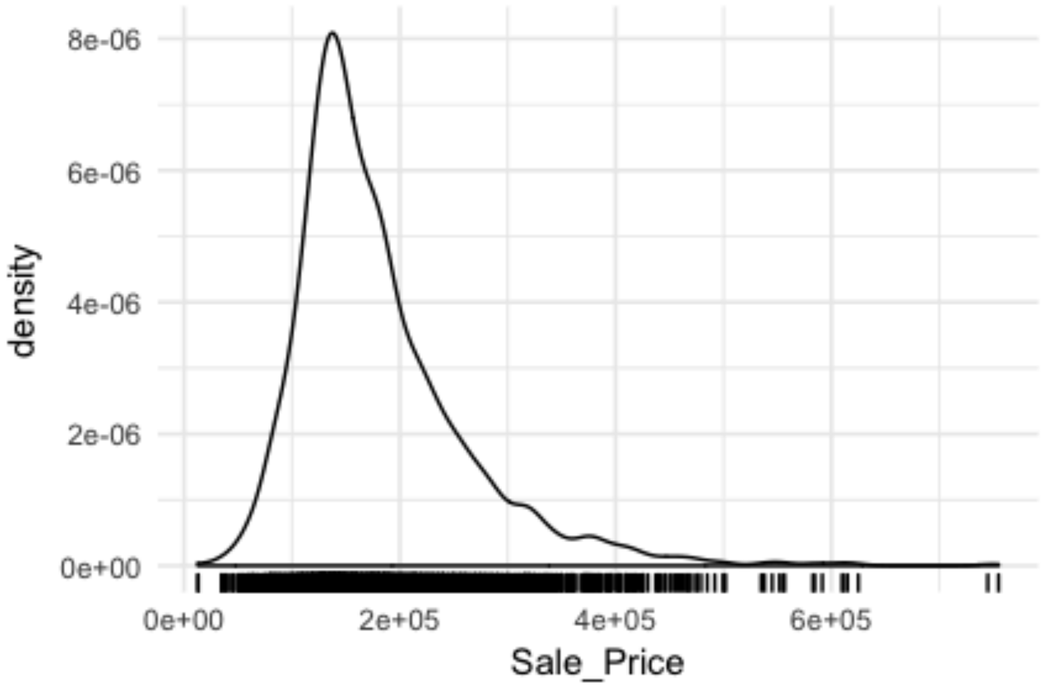
- The following plots help us summarize a single variable
- How is it distributed?
- Which values are most common?

Histogram

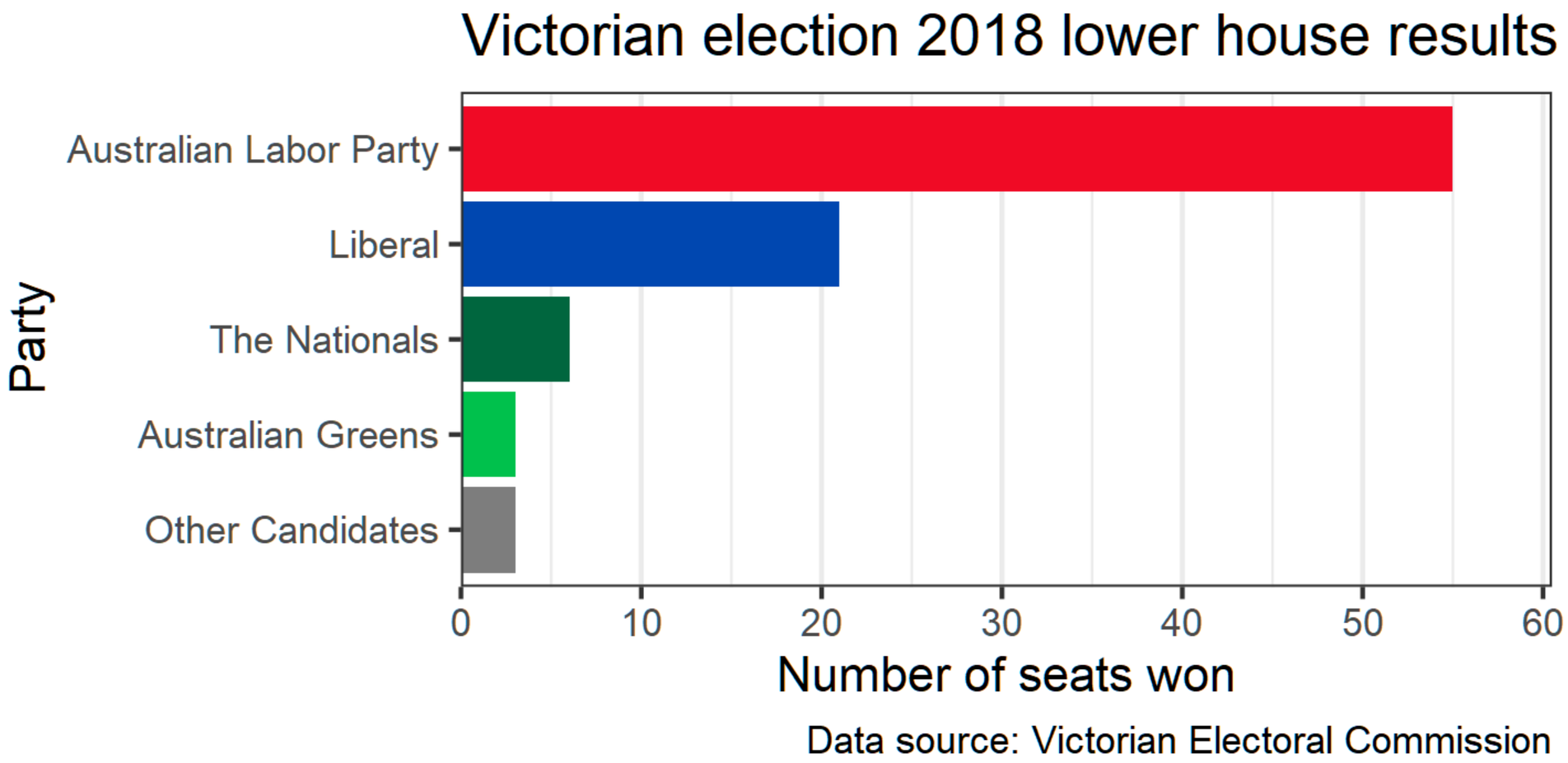

Height of bar conveys the frequency of binned values.
Bars close or touching to illustrate continuity
Bin width represents x values being grouped
Continuous variable

Density

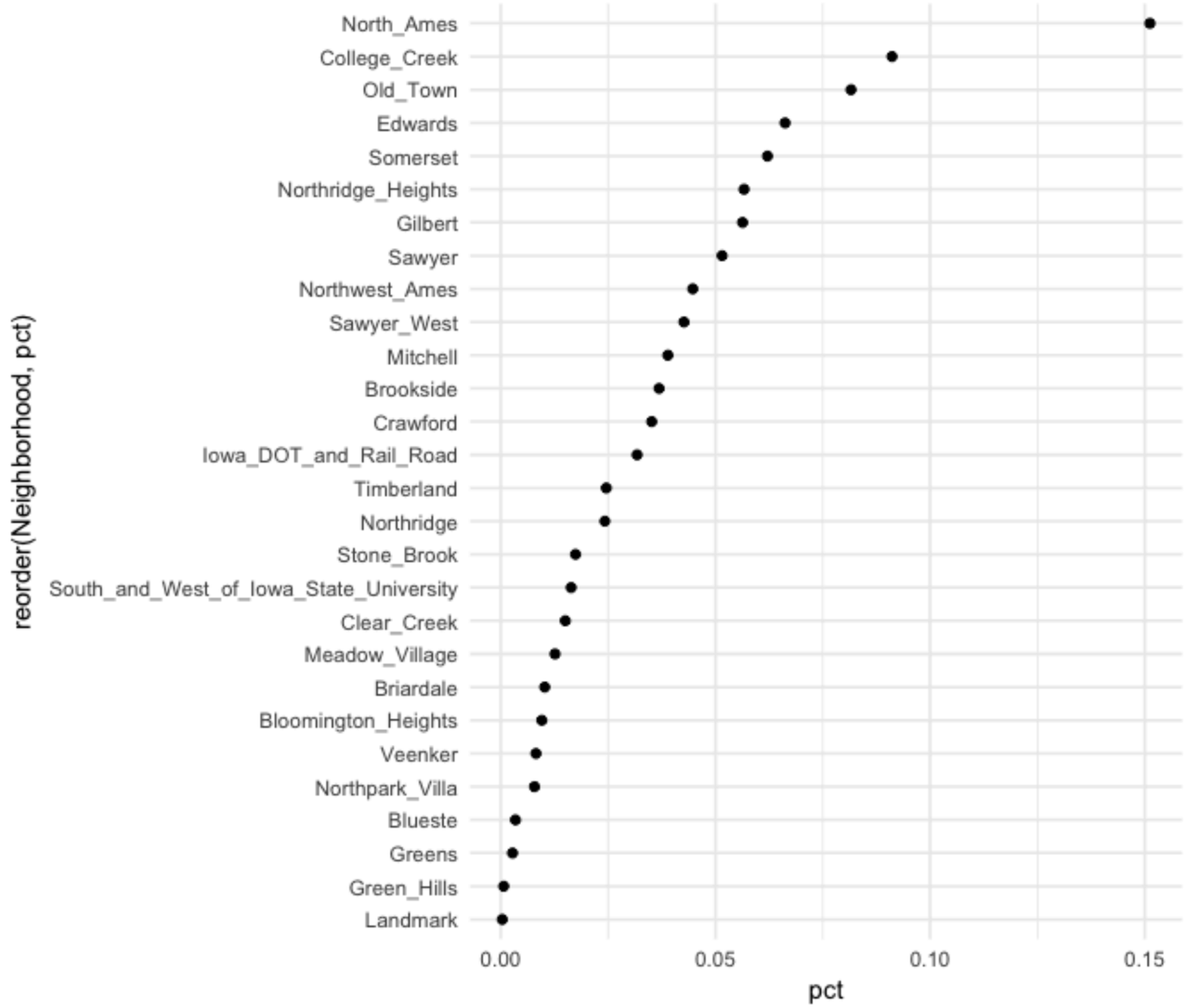

8e-06
6e-06
4e-06
2e-06
0e+00
density
0e+00
2e+05
4e+05
6e+05
Sale_Price


8e-06
6e-06
4e-06
2e-06
0e+00
density
0e+00
2e+05
4e+05
6e+05
Sale_Price

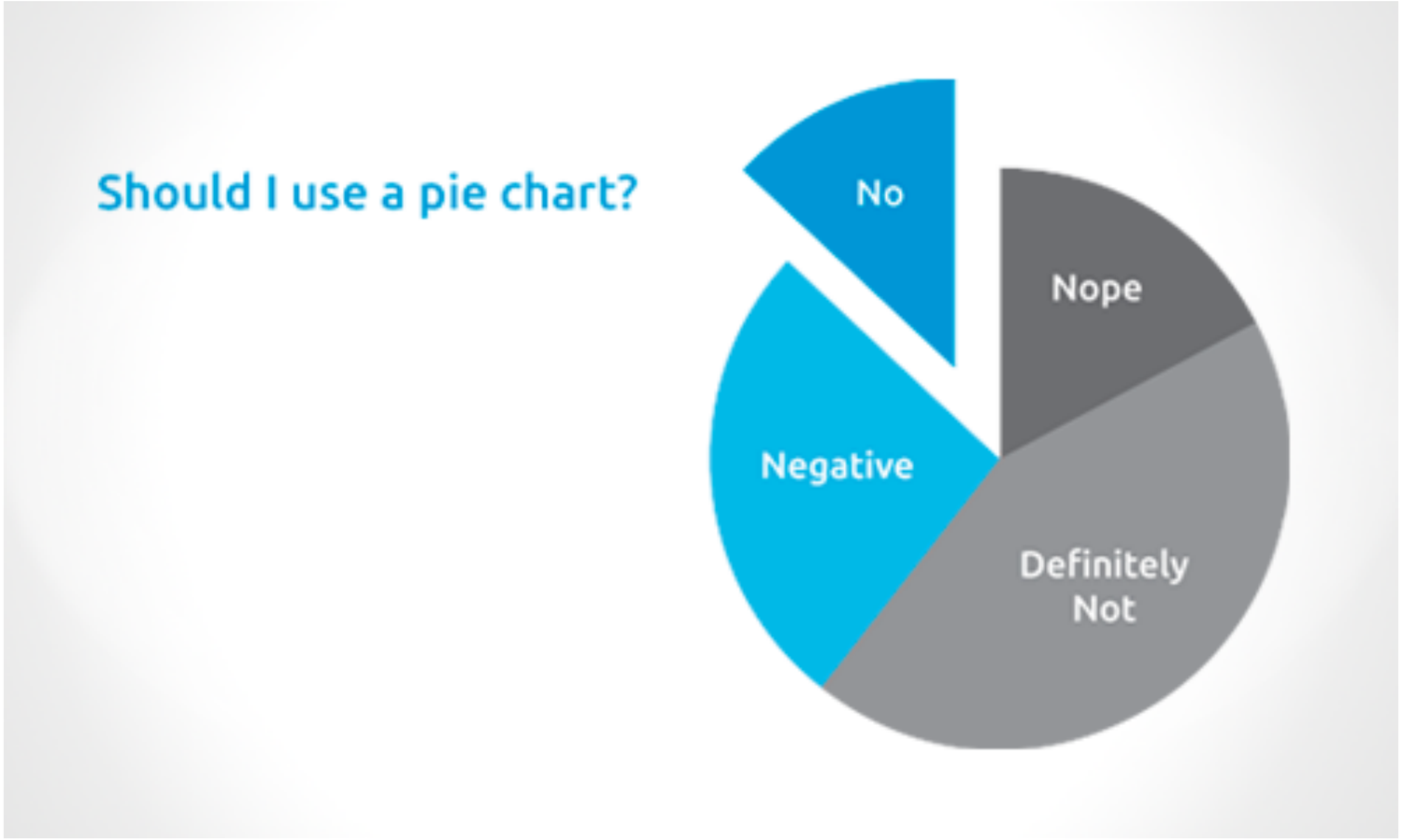
Bar Plots


Victorian election 2018 lower house results
Australian Labor Party
Liberal
The Nationals
Australian Greens
Other Candidates
Party
0
10
20
30
40
50
60
Number of seats won
Data source: Victorian Electoral Commission

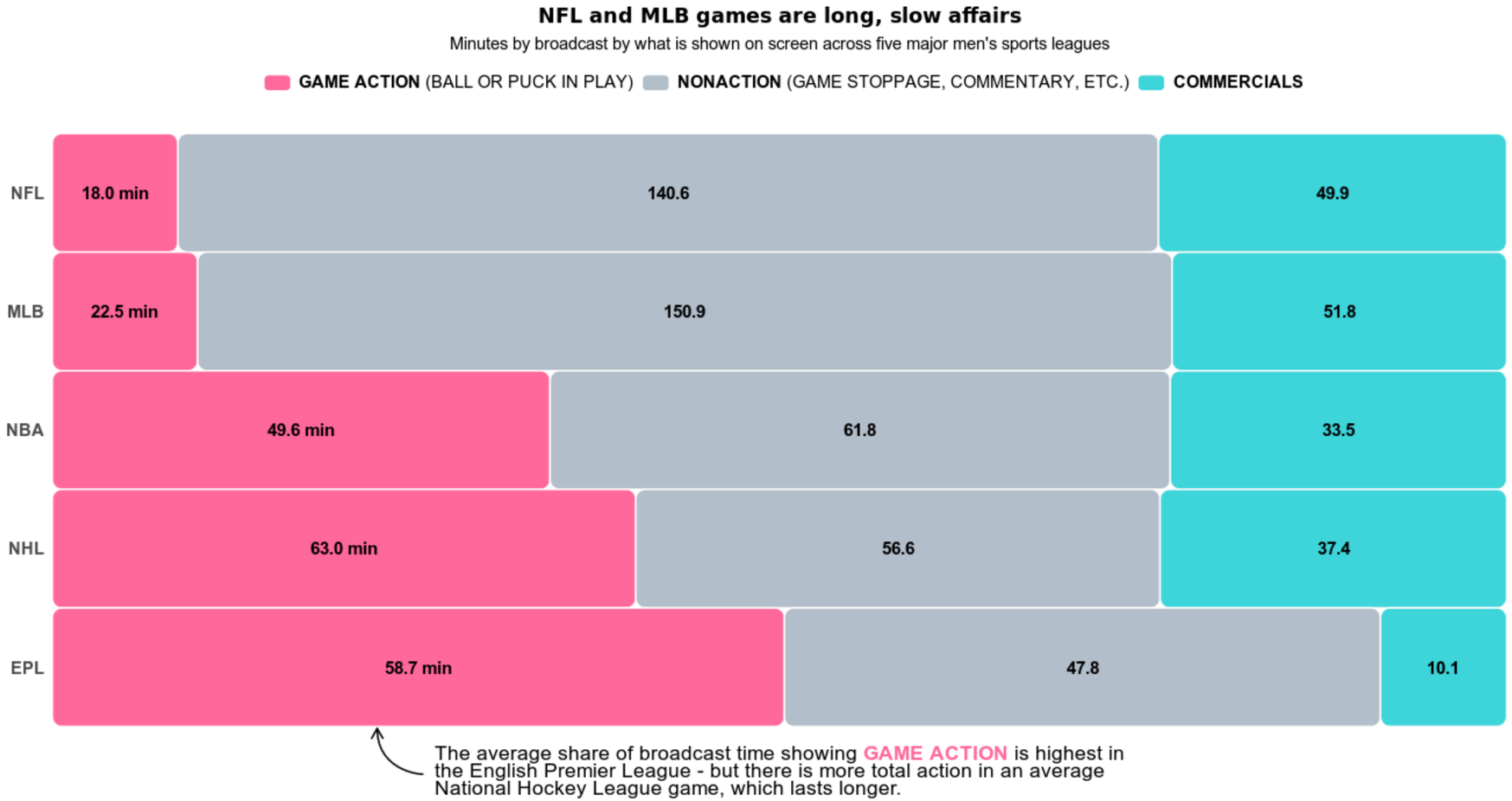
Dot Plot


North_Ames
College_Creek
Old_Town
Edwards
Somerset
Northridge_Heights
Gilbert
Sawyer
Northwest_Ames
Sawyer_West
Mitchell
Brookside
Crawford
Iowa_DOT_and_Rail_Road
Timberland
Northridge
Stone_Brook
South_and_West_of_Iowa_State_University
Clear_Creek
Meadow_Village
Briardale
Bloomington_Heights
Veenker
Northpark_Villa
Blueste
Greens
Green_Hills
Landmark
reorder(Neighborhood, pct)
0.00
0.05
0.10
0.15
pct

Pie Chart



Stacked Bar Chart



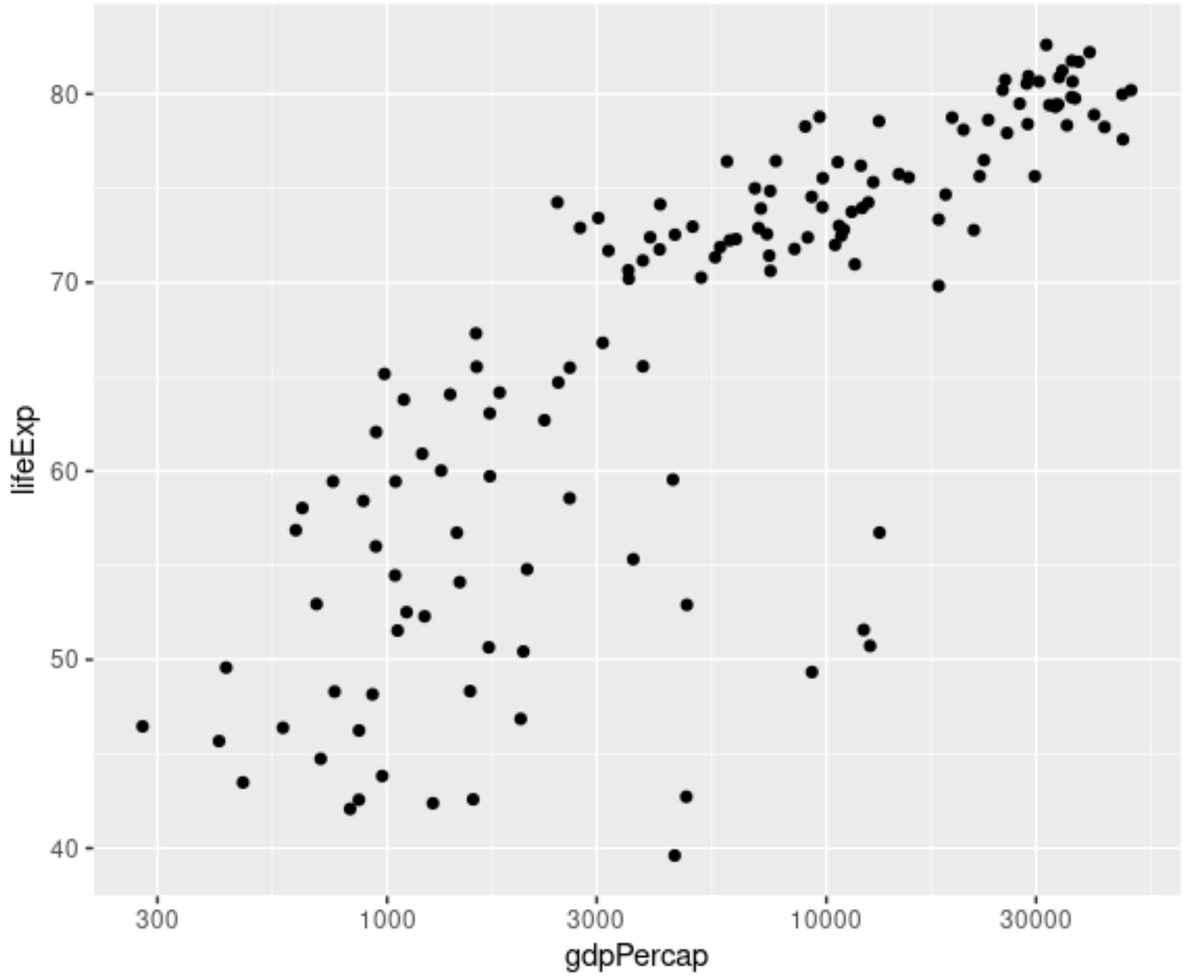
Games that we included: 10 NFL regular-season games between Nov. 7 amd Nov. 18, 2019. 17 MLB postseason games, including all the games in the 2019 ACLS, NLCS, and World Series; 10 NBA regular-season games between Nov. 6 and Nov. 15, 2019; 10 NHL regular-season games between Nov. 5 and Nov. 19, 2019, including three overtime games; and seven English Premier League games between Nov. 9 and Nov. 23, 2019. NBA game action includes free throws, so the action time exceeds the game time.

FivethirtyEight SOURCE: UNIVERSITY OF TEXAS AT AUSTIN SPORTS ANALYTICS COURSE

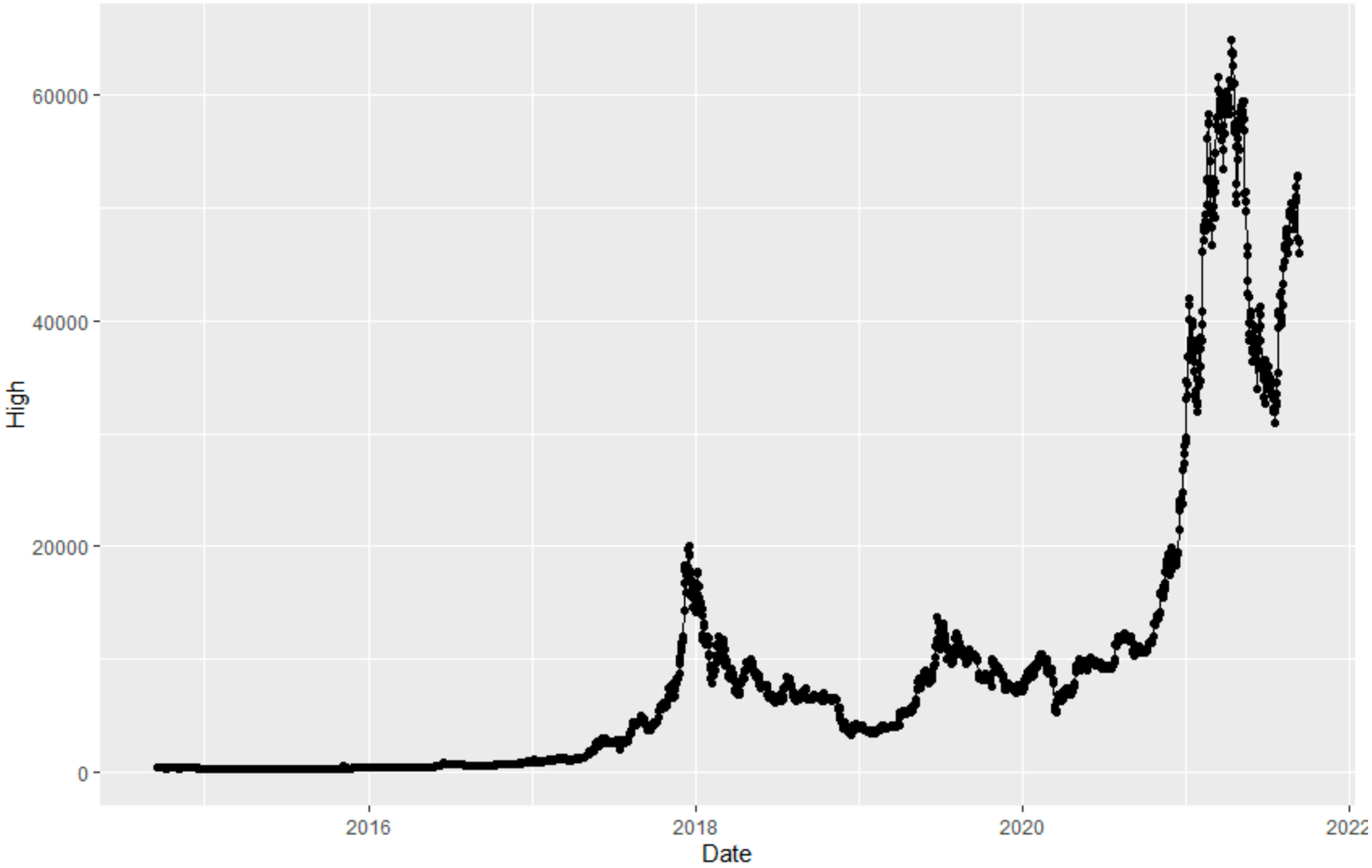
Multivariate plots

- Plots that summarize the relationship between two or more variables

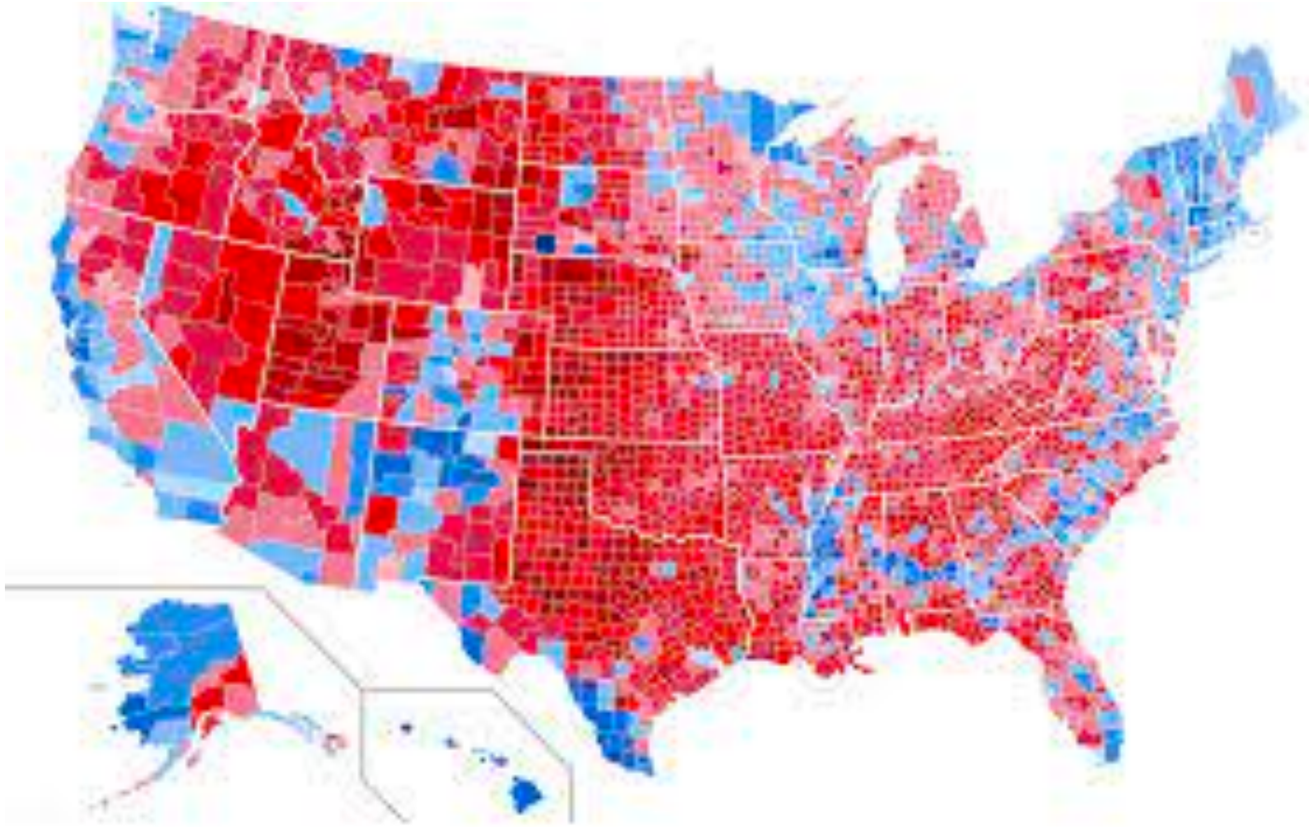
Scatterplot



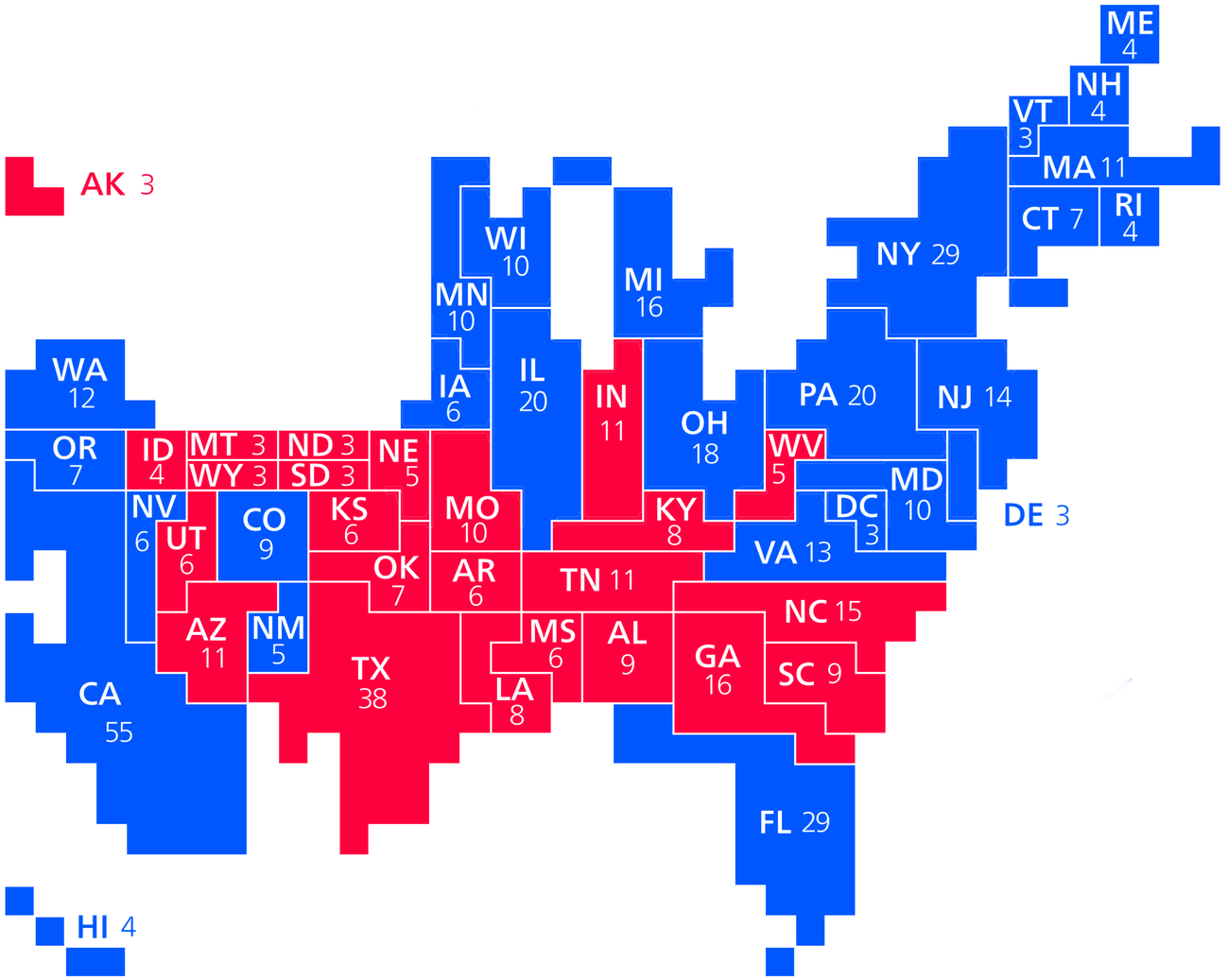
Time Series



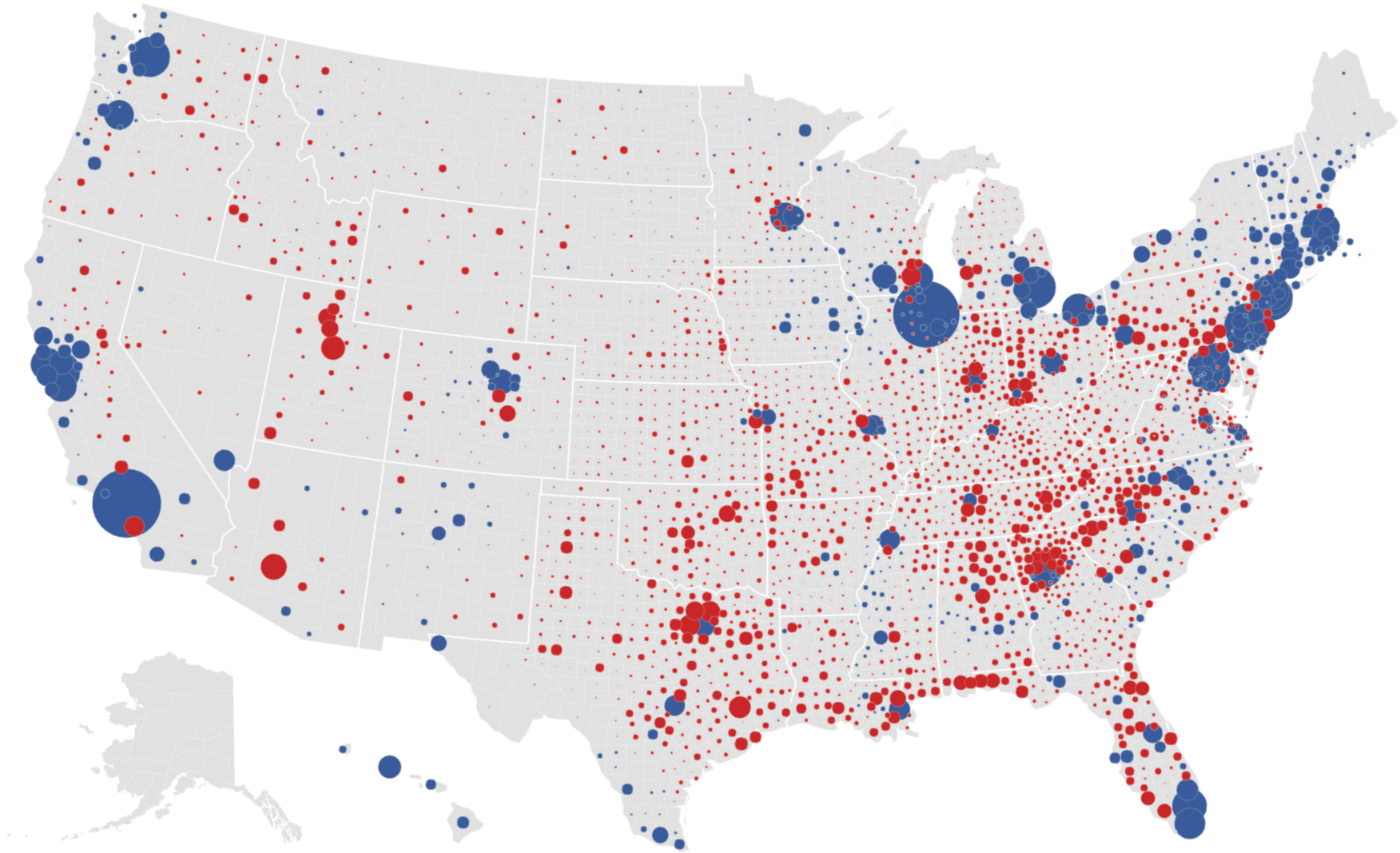
Choropleth Map



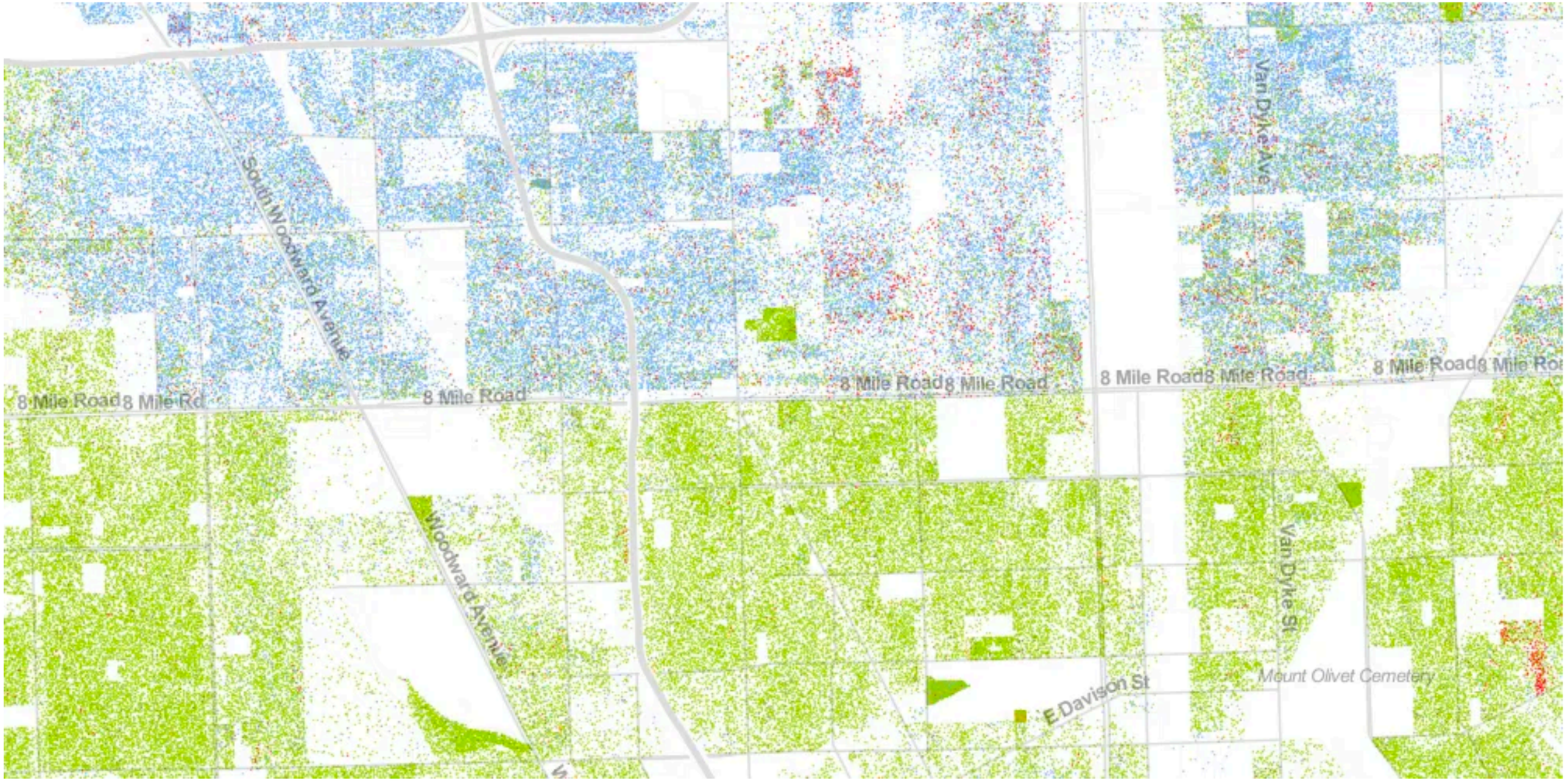
Cartogram


AK 3
ME 4
NH 4
VT 3
MA 11
CT 7
RI 4
NY 29
WI 10
MN 10
MI 16
WA 12
IA 6
IL 20
IN 11
OH 18
PA 20
NJ 14
OR 7
ID 4
MT 3
ND 3
NE 5
WY 3
SD 3
WV 5
MD 10
DE 3
NV 6
UT 6
CO 9
KS 6
MO 10
KY 8
DC 3
VA 13
OK 7
AR 6
TN 11
NC 15
AZ 11
NM 5
TX 38
MS 6
AL 9
GA 16
SC 9
LA 8
CA 55
FL 29
HI 4

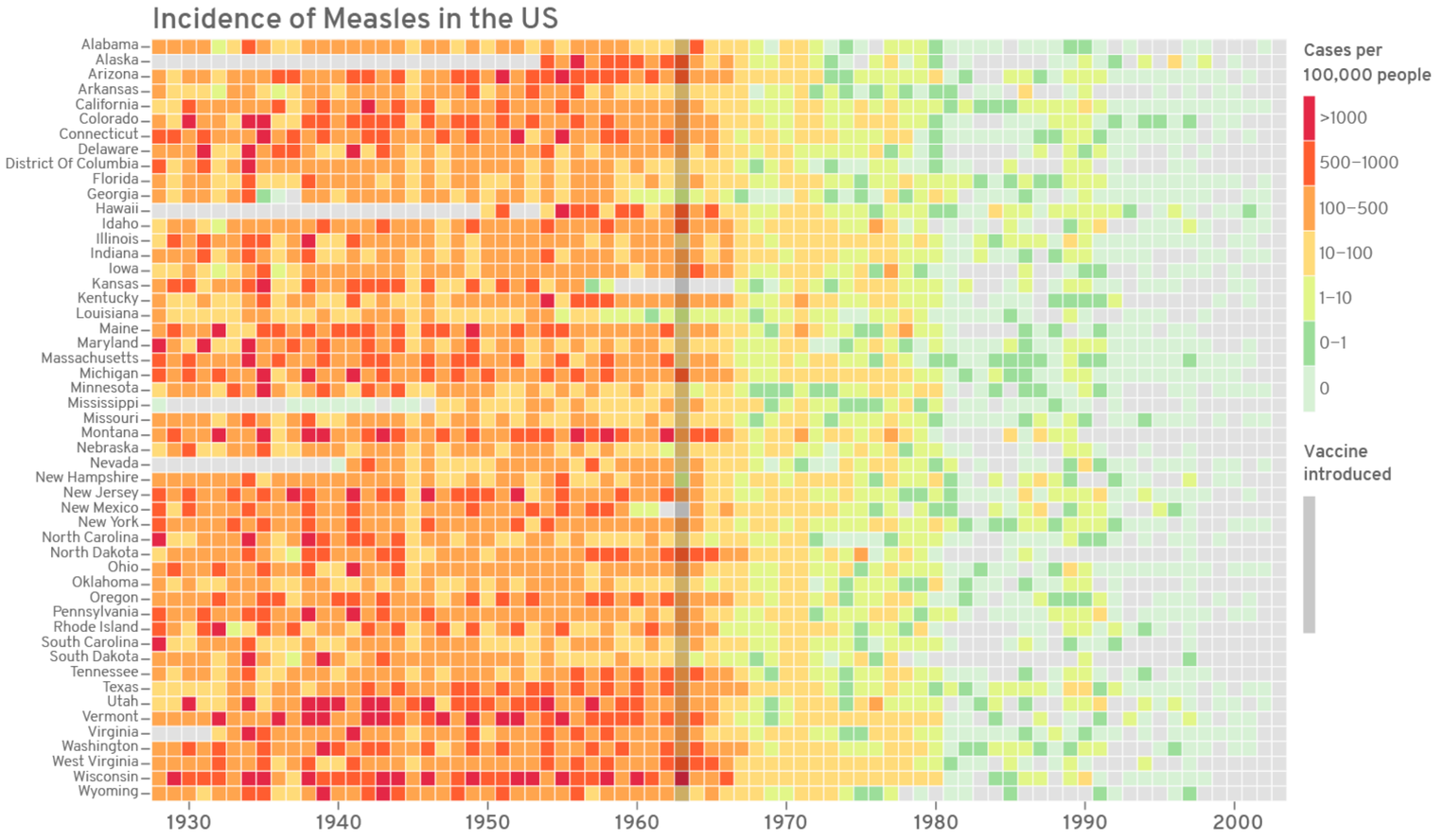
Proportional Symbol Maps



Dot Distribution Map



Heatmap


Incidence of Measles in the US
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District Of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming
1930
1940
1950
1960
1970
1980
1990
2000
Cases per 100,000 people
>1000
500–1000
100–500
10–100
1–10
0–1
0
Vaccine introduced

Group Assignment