Class 07

DATA1220-55, Fall 2024

Sarah E. Grabinski

2024-09-13

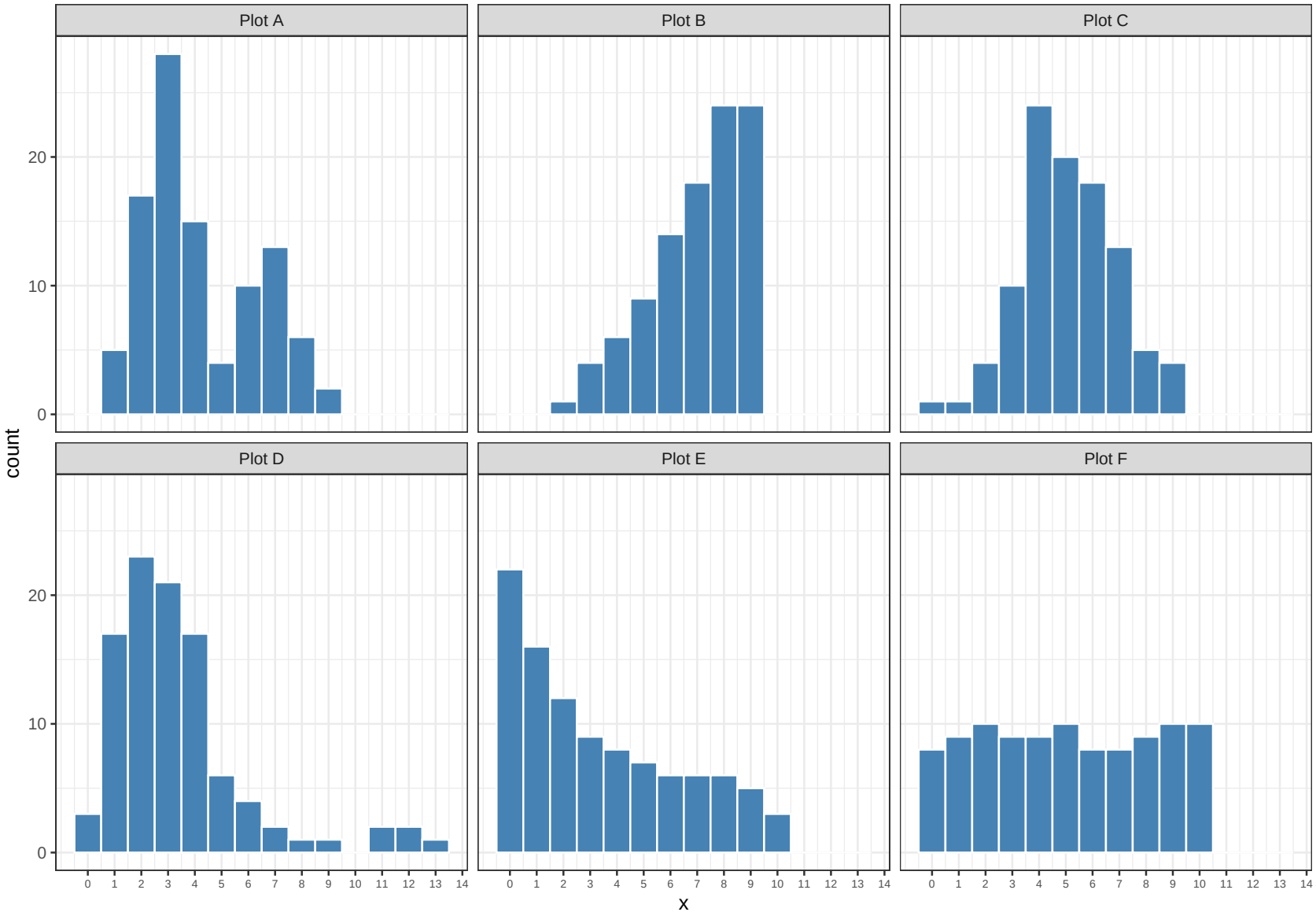
Last time...

- The ***normal distribution*** and when to use the ***mean + standard deviation***
- ***Robust statistics*** and when to use the ***median + interquartile range***
- The ***5-number summary*** and ***how to read a boxplot***
- Other distribution plots: ***dot plot***, ***histogram***, ***density plot***, ***violin plot***

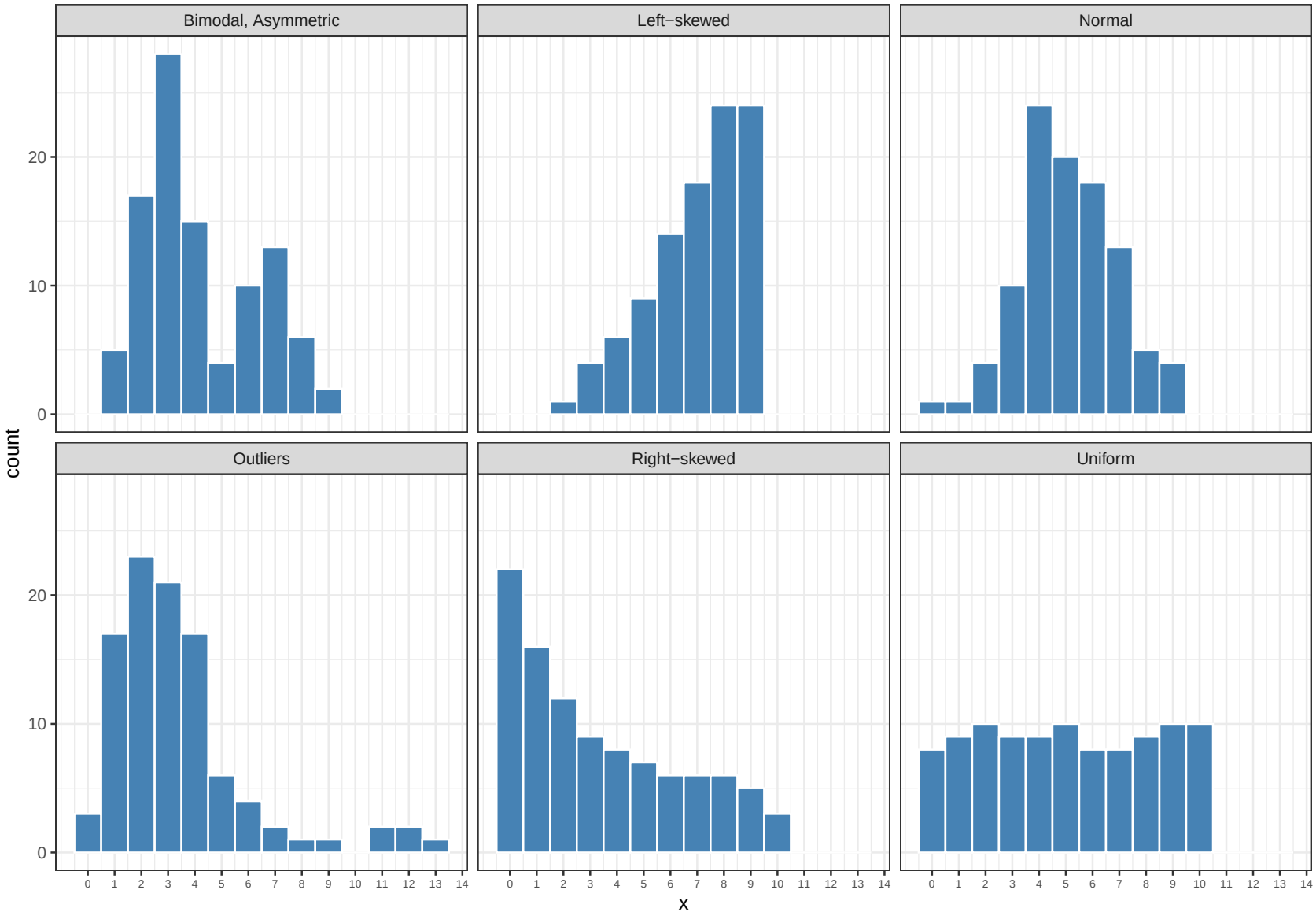
Distribution Checklist

- What is the **modality** of the distribution?
- How many "peaks" are there?
- Is the distribution **skewed** or **symmetric**?
- Is there a longer "tail" on the left or right side?
- Are there any **outliers**?
- How extreme are the most extreme values?
- What are the appropriate **summary statistics** for a distribution with this shape?
- Would the mean+standard deviation or the median+IQR more accurately describe this data?

Practice: Visualizing Distributions



Practice: Visualizing Distributions



Practice: Summary Statistics

Is the mean greater than, lesser than, or about equal to the mean?
What does that mean for the shape of the distribution?

	min	Q1	median	Q3	max	mean	sd	n	missing
	1	3	3.5	6	9	4.2	2.117746	100	0

Practice: Summary Statistics

The mean is greater than the median indicating this is an asymmetrical distribution that is skewed. Because the mean is greater than the median, this distribution is right-skewed towards numbers on the high end of the range.

	min	Q1	median	Q3	max	mean	sd	n	missing
	1	3	3.5	6	9	4.2	2.117746	100	0

Today

- Analyze contingency (e.g. 2x2) tables
- Summarizing categorical variables with proportions
- Visualizing data using categories
- Comparing data from 2+ variables

What is a contingency table?

A ***contingency table*** is a cross-tabulation of the ***frequency*** of observations across 2 categorical variables

Variable 2

Variable 1		Category 1	Category 2	Total
	Category 1	A	B	A + B
	Category 2	C	D	C + D
	Total	A + C	B + D	A + B + C + D

Figure 1: In this table, the values are the count of observations that belong both the corresponding row and column categories.

What is frequency?

When someone describes a category's ***frequency***, they may be referring to...

- ***Count***: the total number of times the category appears in the data
- ***Proportion***: the total number of times the category appears in the data divided by the number of observations

Calculating a proportion

Only a subset of observations belong to each level of a categorical variable. A proportion describes the count of the observations belonging to that category divided by the total number of observations (n).

$$\text{Proportion} = \frac{\text{count(category)}}{n}$$

Proportions can be any real number between a minimum of 0 ($\frac{0}{n}$) and maximum of 1 ($\frac{n}{n}$).

Calculating Proportions by row

Variable 2

Variable 1		Category 1	Category 2	Total
	Category 1	A / (A + C)	B / (B + D)	1
	Category 2	C / (A + C)	D / (B + D)	1
	Total	(A + C) / (A + B + C + D)	(B + D) / (A + B + C + D)	1

Figure 2: The row totals are all 1, which is the maximum value of a proportion. This indicates that the denominator for the proportions is the row total for each cell.

Calculating Proportions by Column

Variable 2

Variable 1		Category 1	Category 2	Total
	Category 1	A / (A + B)	B / (A + B)	(A + B) / (A + B + C + D)
	Category 2	C / (C + D)	D / (C + D)	(C + D) / (A + B + C + D)
	Total	1	1	1

Figure 3: The column totals are all 1, which is the maximum value of a proportion. This indicates that the denominator for the proportions is the column total for each cell.

Example: Field Goals

In the NFL, kickers can play a pivotal role in determining the outcome of a game. Much time and money has been devoted to determining the factors which lead to a successful kick. Past performance is one factor which could affect whether a kick is good or not. Are kickers more or less likely to make the extra point kick if they missed their last kick?

Contingency Table: Counts

last_kick	Didn't Make Current Kick	Made Current Kic
Didn't Make Last Kick	47	2
Made Last Kick	15	11
Total	62	13

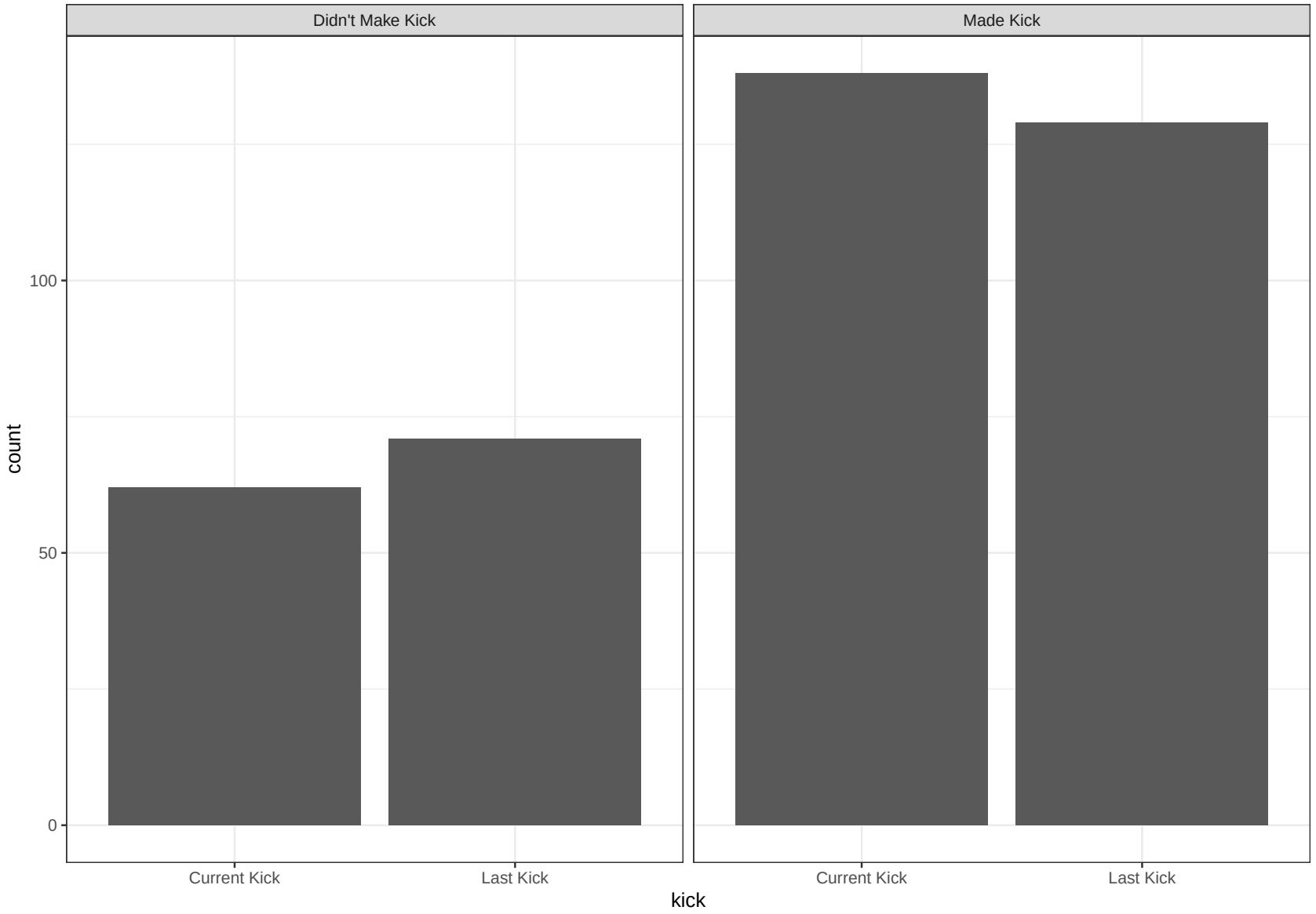
Contingency Table: Proportions by Row

last_kick	Didn't Make Current Kick	Made Current Kic
Didn't Make Last Kick	0.6619718	0.338028
Made Last Kick	0.1162791	0.883720
Total	0.3100000	0.690000

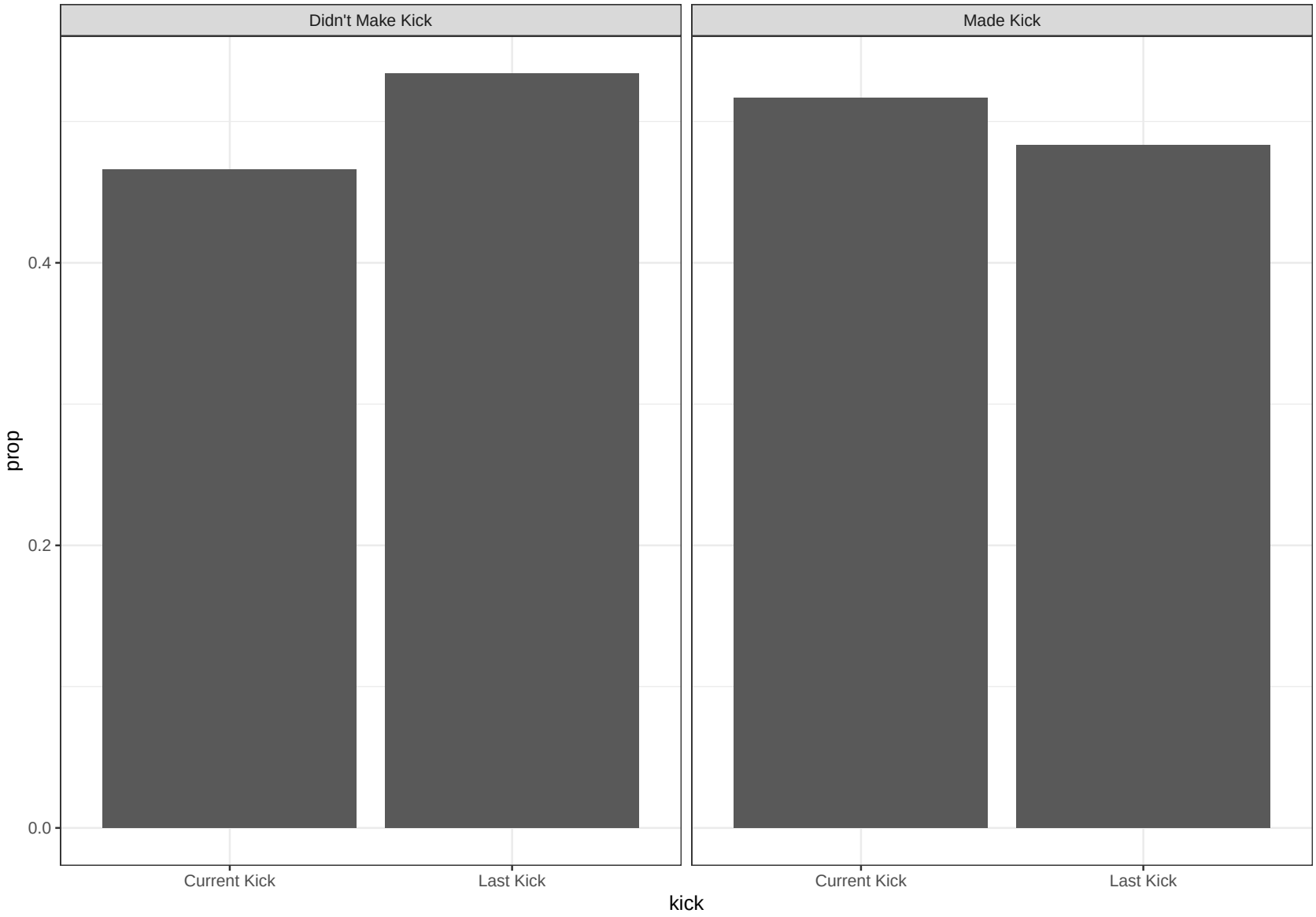
Contingency Table: Proportions by Column

last_kick	Didn't Make Current Kick	Made Current Kic
Didn't Make Last Kick	0.7580645	0.17391
Made Last Kick	0.2419355	0.82608
Total	1.0000000	1.00000

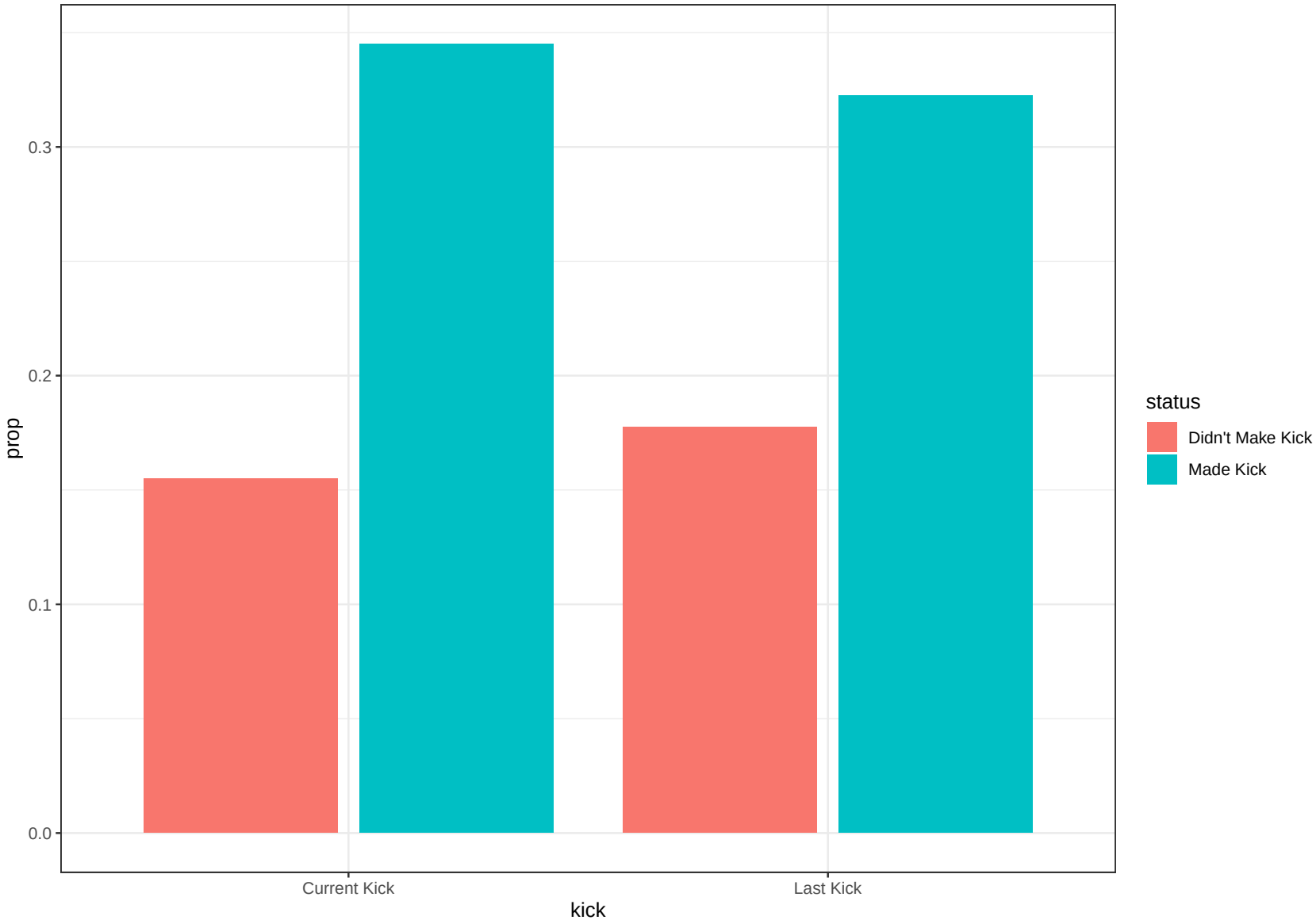
Visualizing Frequencies: Bar Plot (Counts)



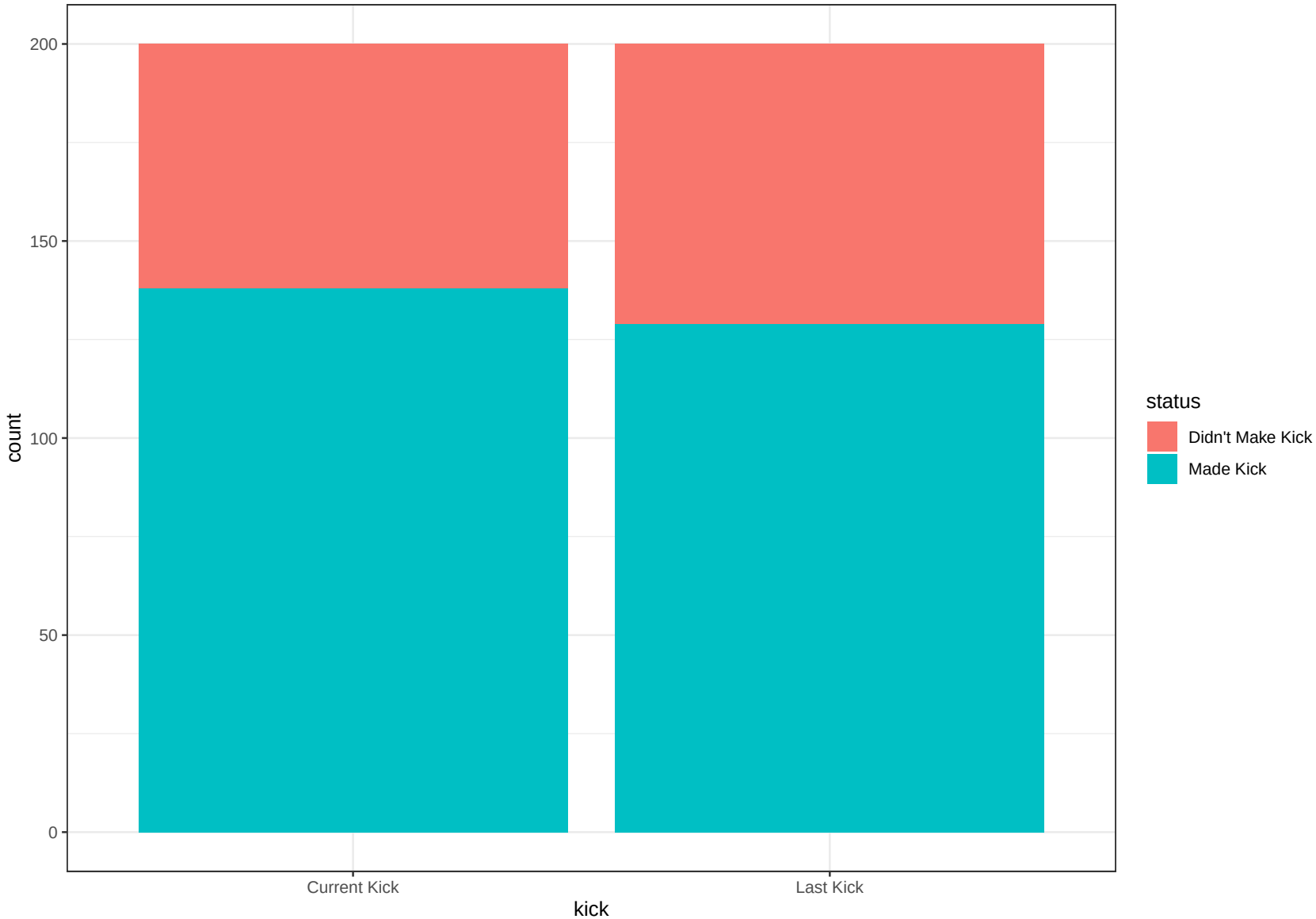
Visualizing Frequencies: Bar Plot (Proportions)



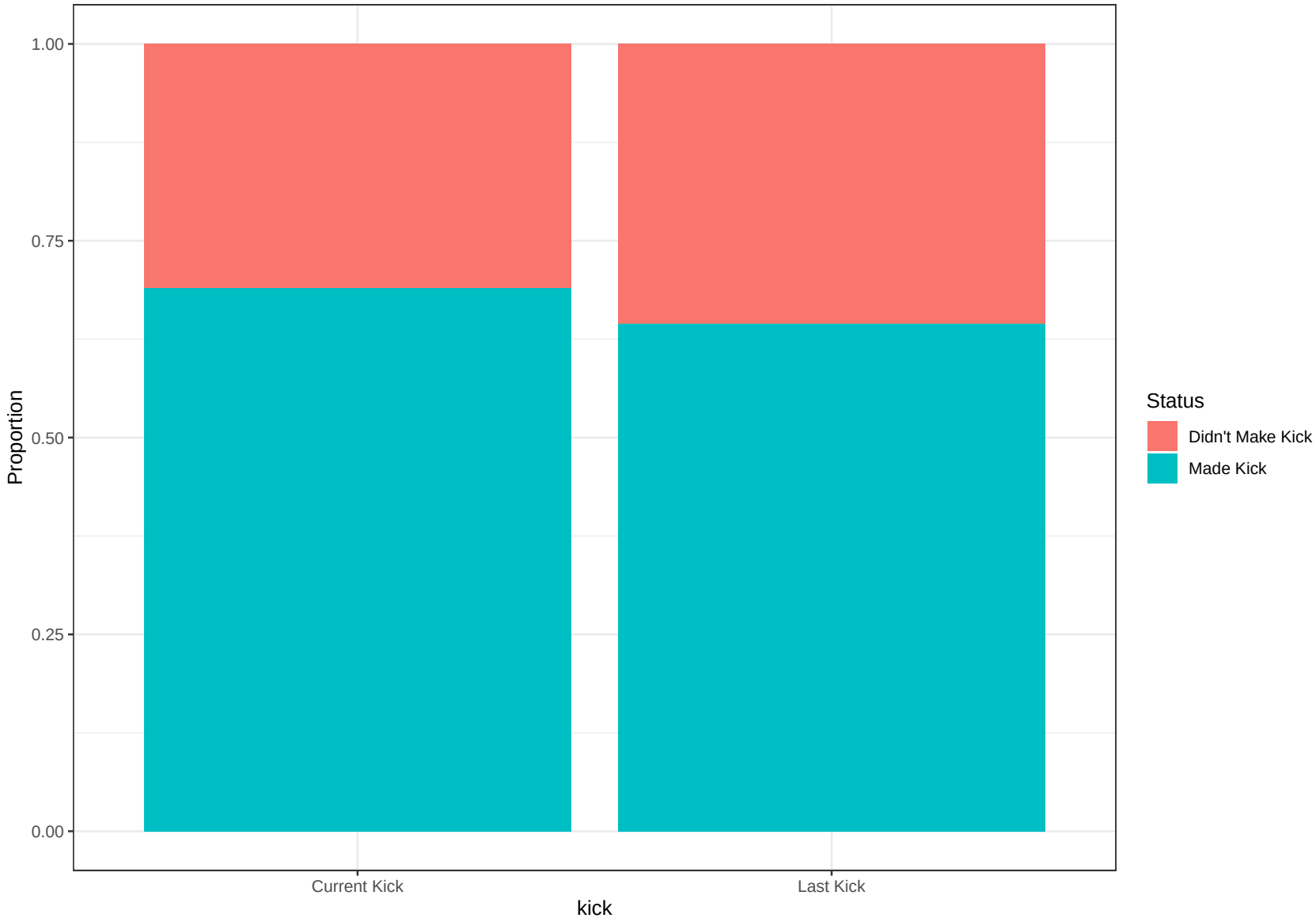
Comparing 2 (Categorical) Variables: Side-By-Side Bar Plot



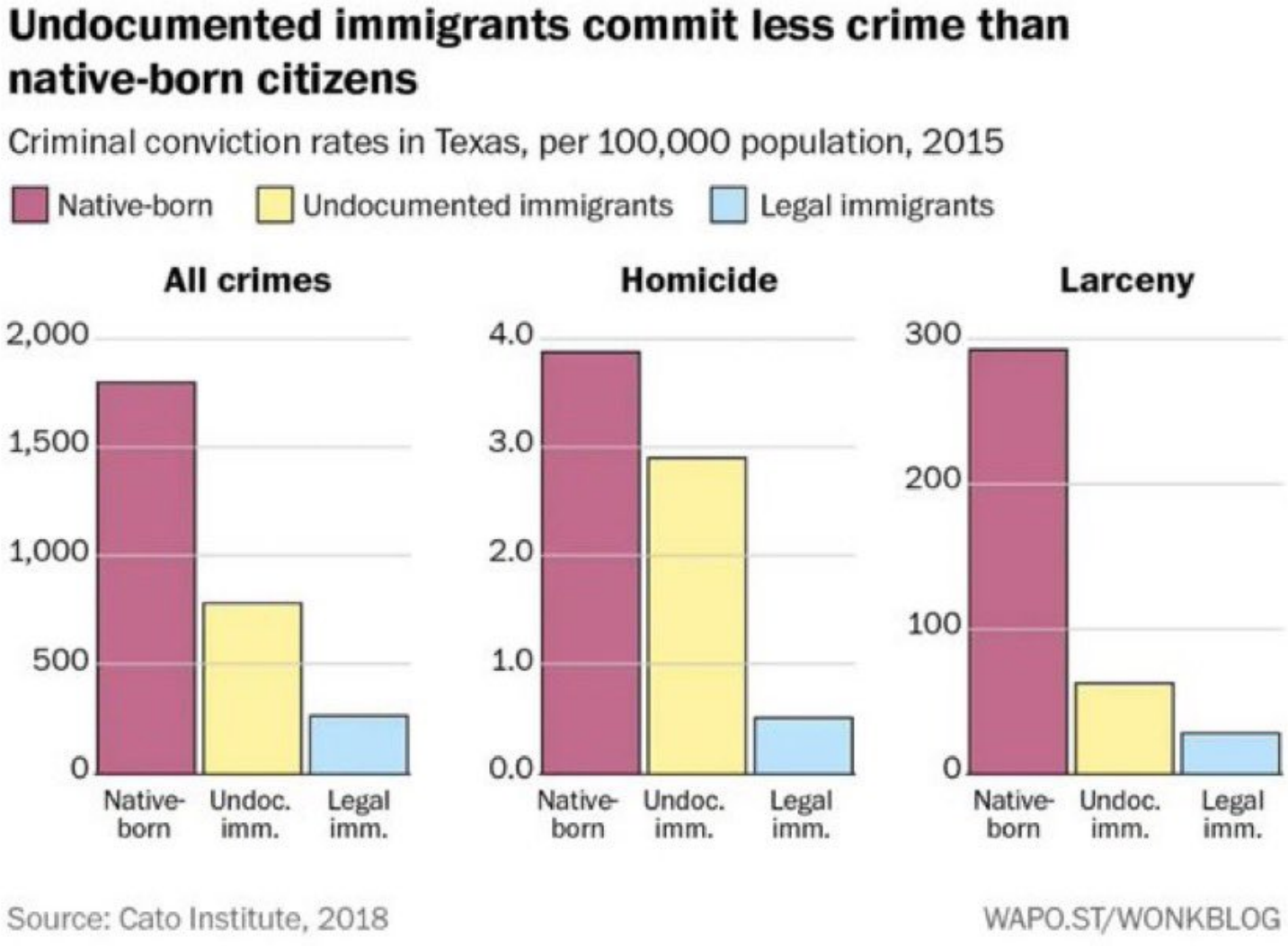
Comparing 2 (Categorical) Variables: Stacked Bar Plot (counts)



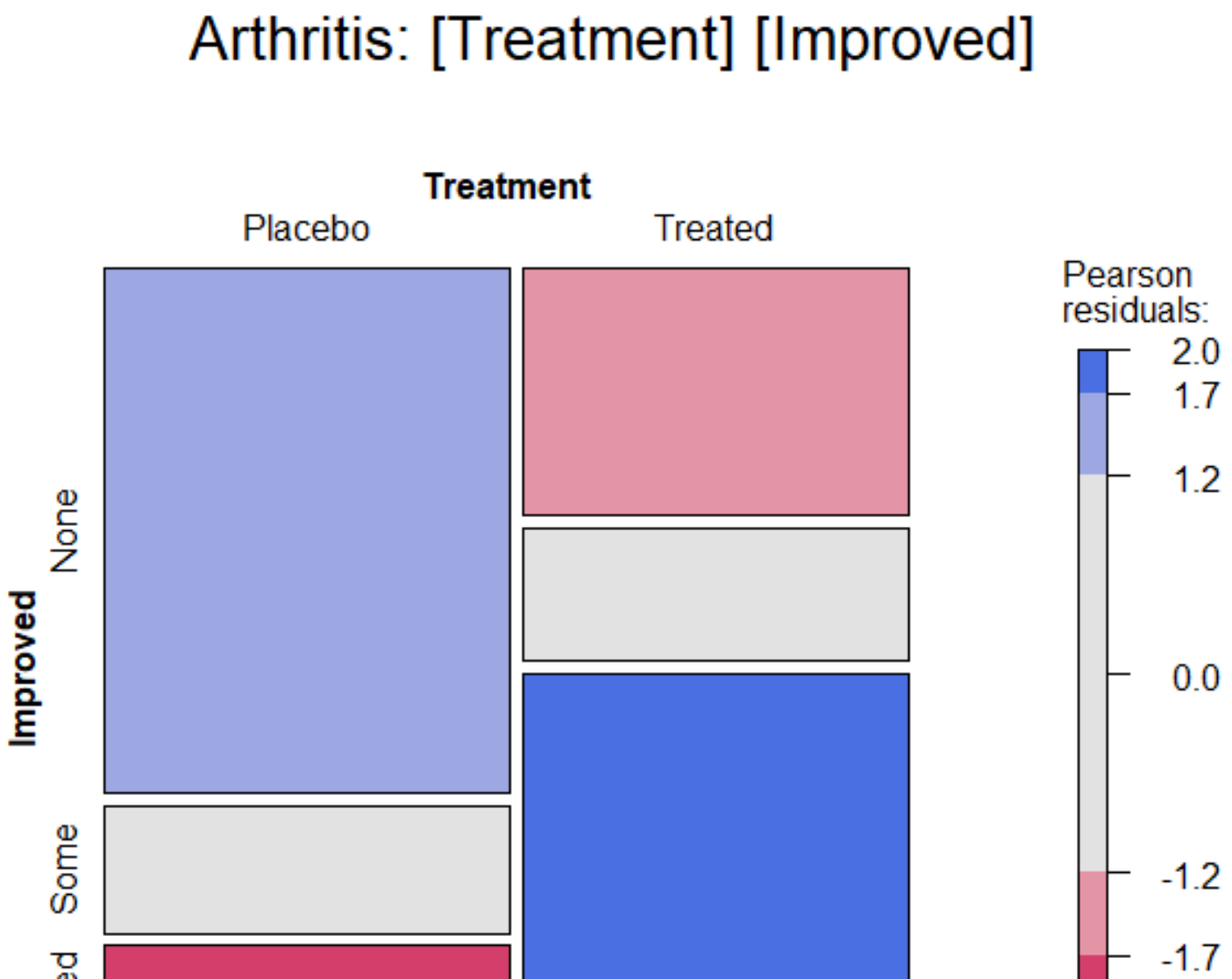
Comparing 2 (Categorical) Variables: Stacked Bar Plot (proportions)



Bar plots in the news this week...



Comparing 2 (Categorical) Variables: Mosaic Plot (2 Variables)



Comparing a categorical variable and a numeric variable

- Histograms + density curves
- Boxplot + violin plot
- Rain cloud plot

Which type of plot(s) are shown in these figures?

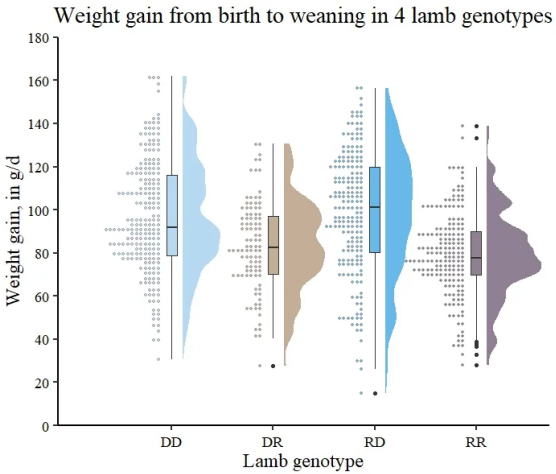


Figure 4: Figure A

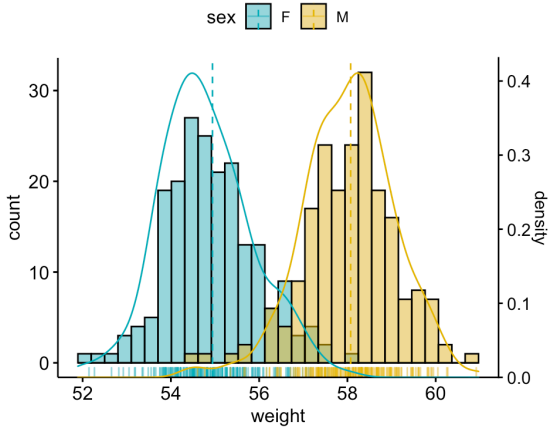


Figure 5: Figure B

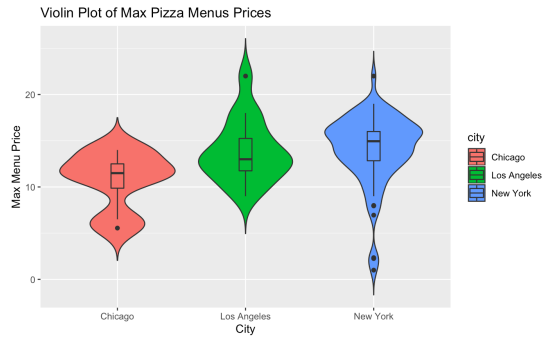
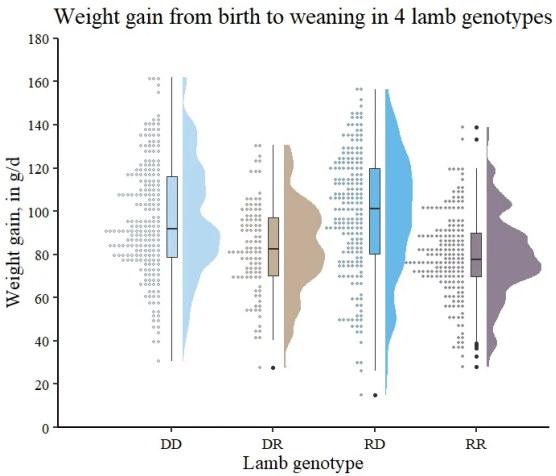
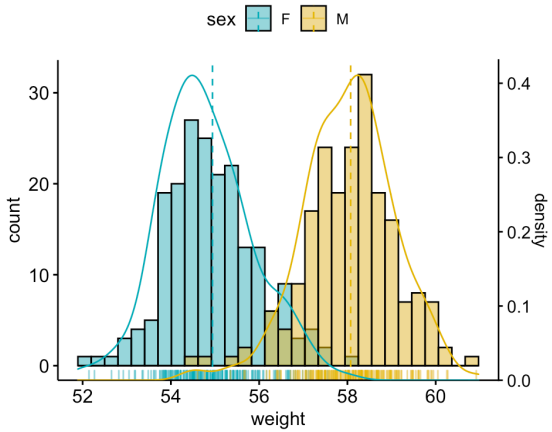


Figure 6: Figure C

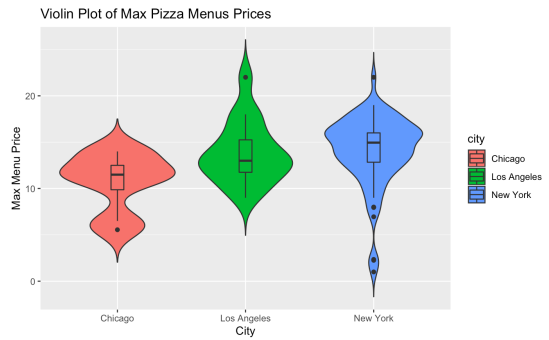
Which type of plot(s) are shown in these figures?



Answer: Rain Cloud Plot (Density plot + Boxplot + Dot plot)

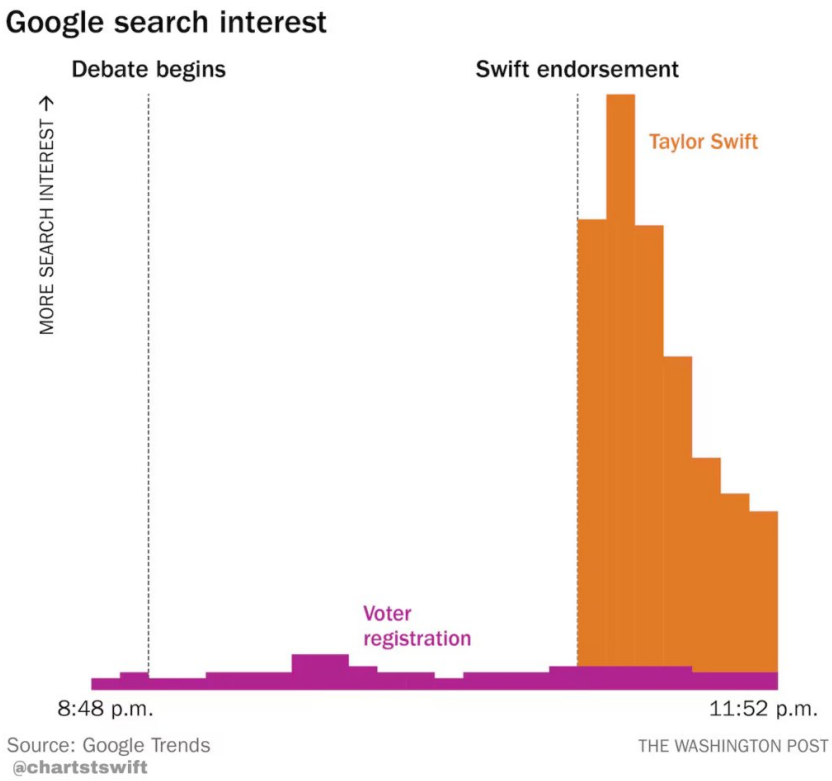


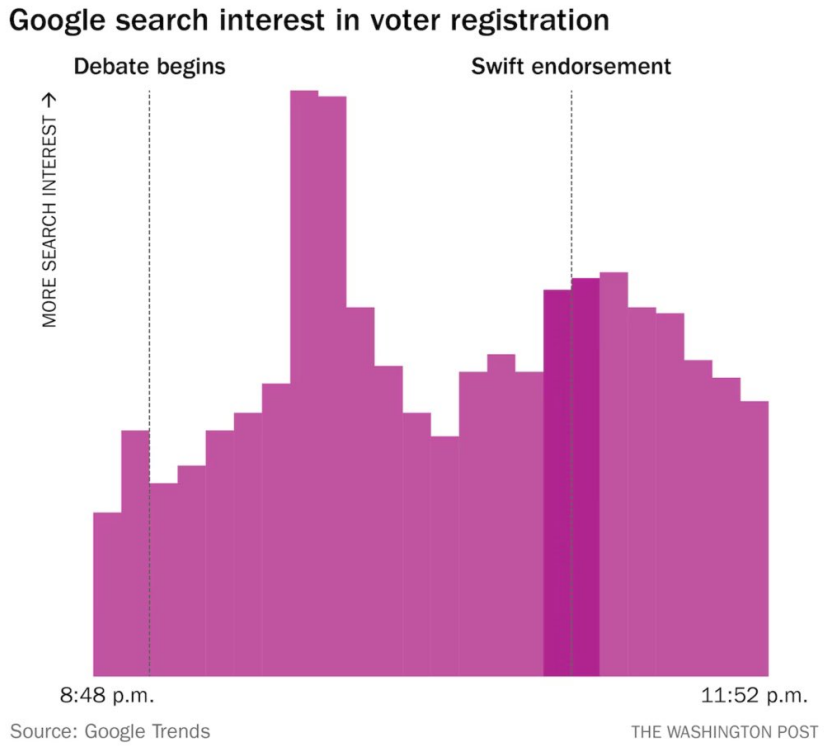
Answer: Histogram + Density Curve (+ Rug Plot, at bottom)



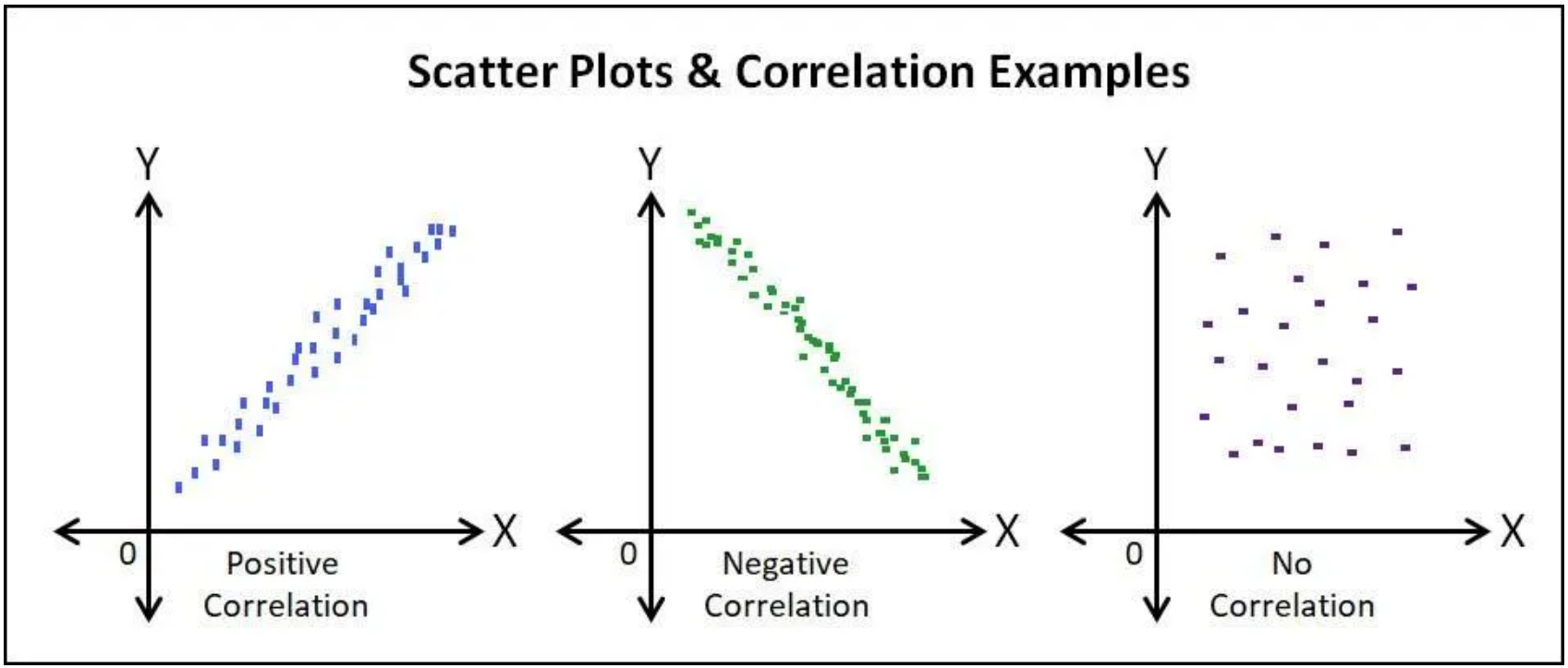
Answer: Boxplot + Violin plot

Distributions in the news this week...





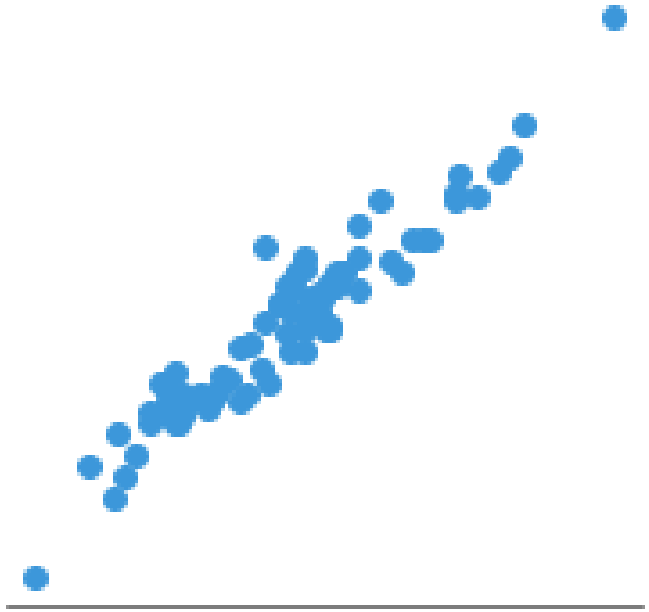
Comparing 2 Numerical Variables: Scatter plot



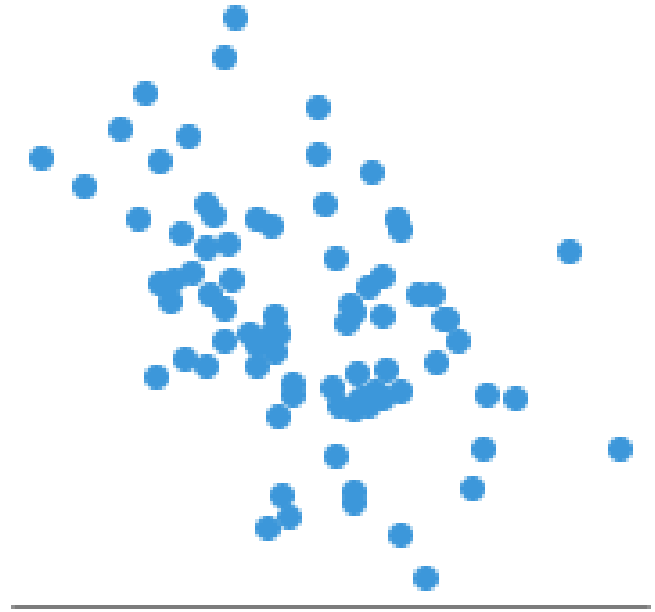
More ways to describe the relationship between 2 numerical variables

- Is the relationship between the 2 variables ***linear*** or ***nonlinear ?***
- Is the relationship ***strong*** or ***weak***?

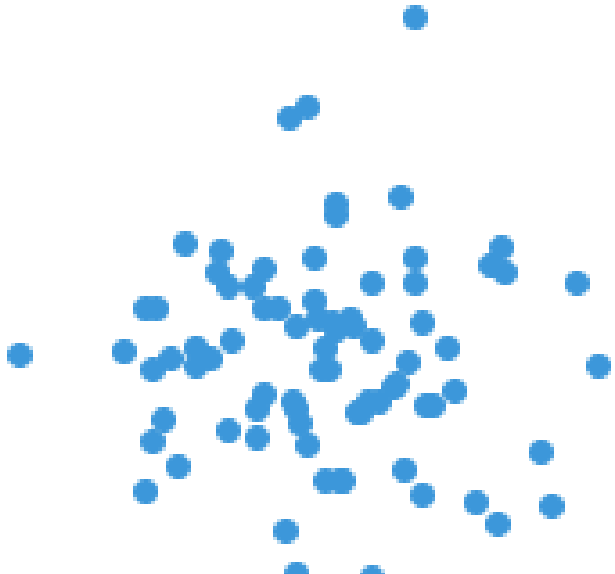
Example: describing numerical variable relationships

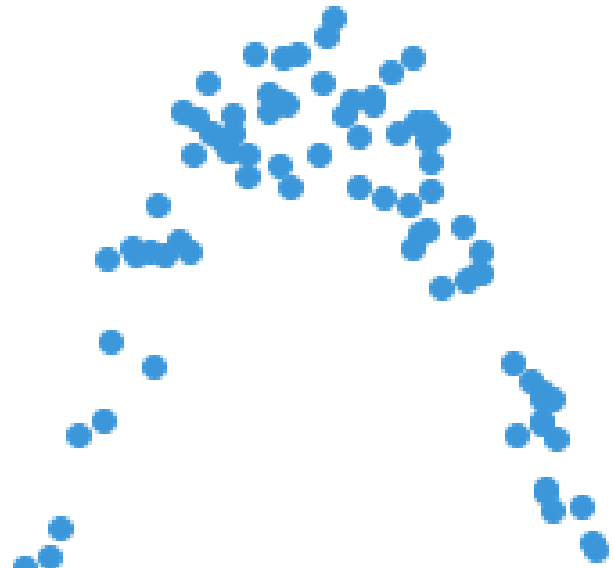


strong, positive, linear

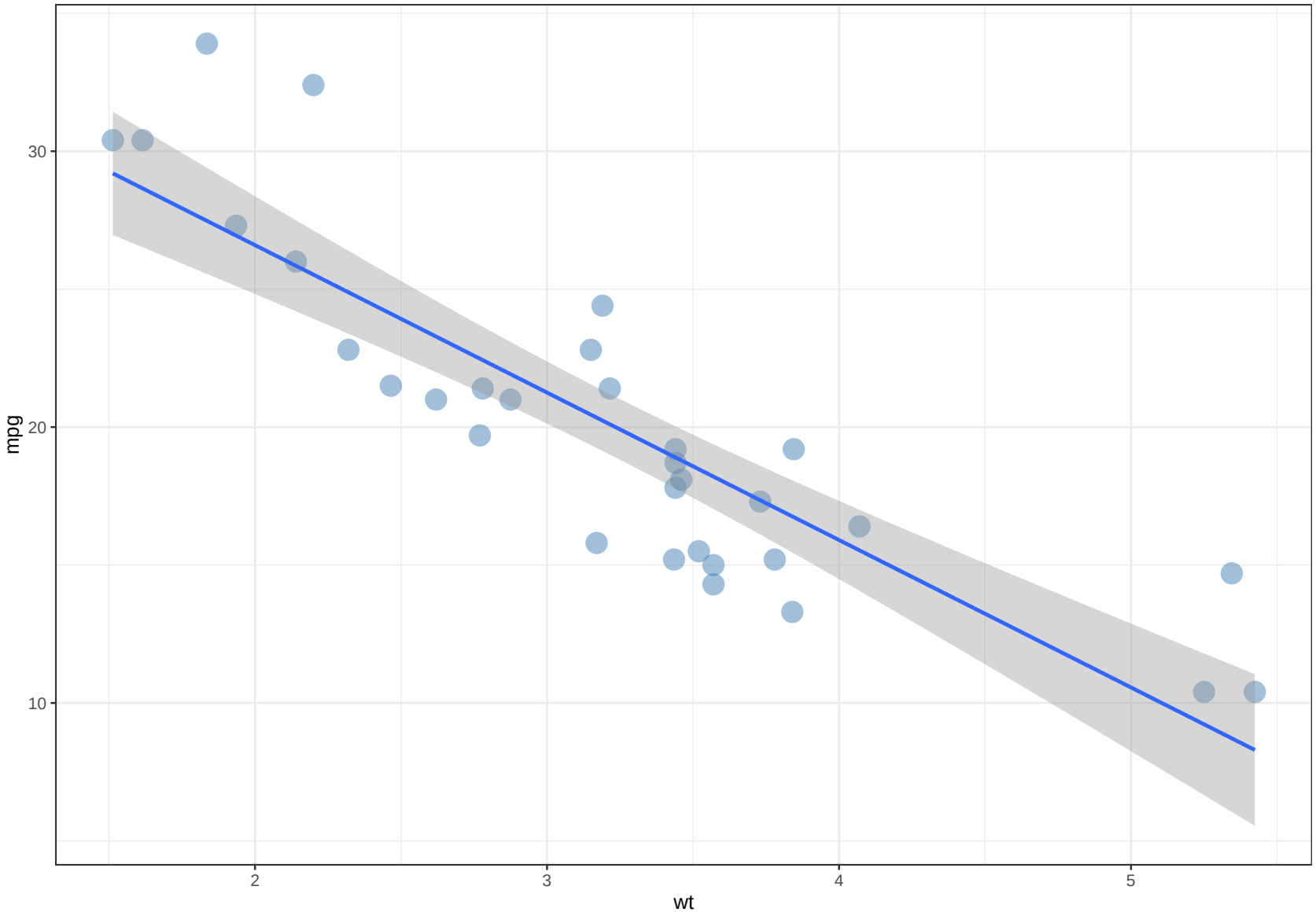


moderate, negative, linear

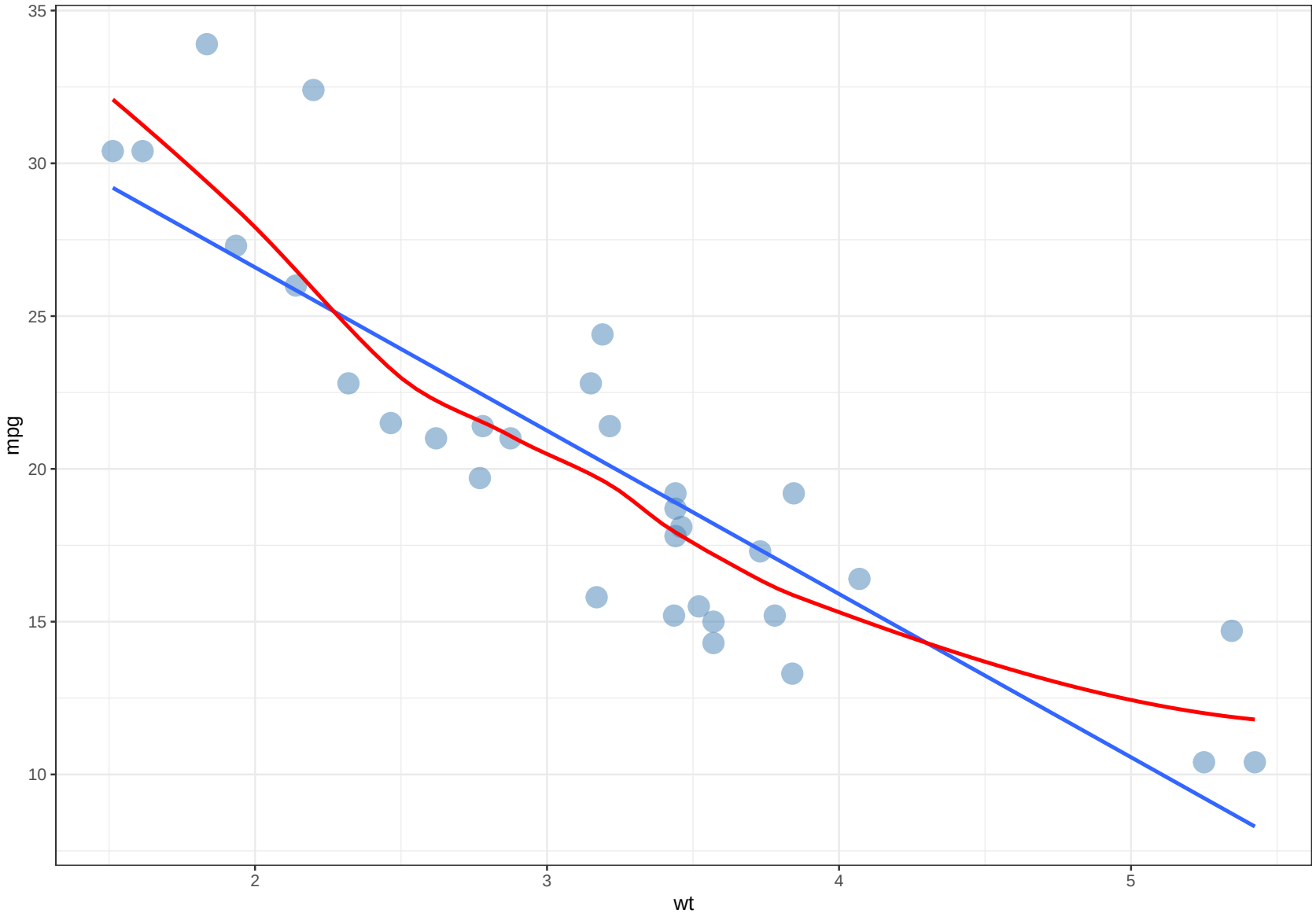




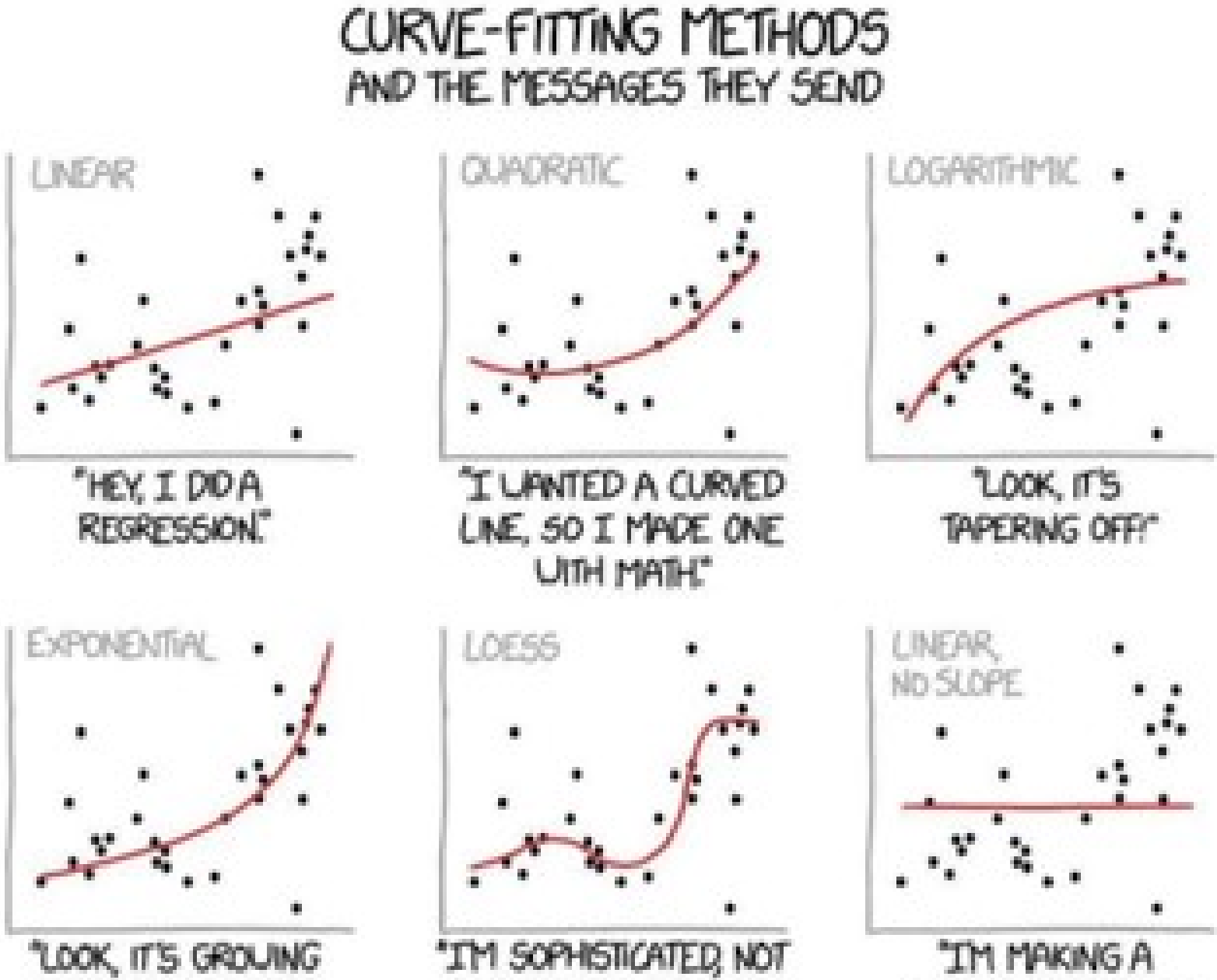
Visualizing the relationship between 2 numerical variables: add a regression line!



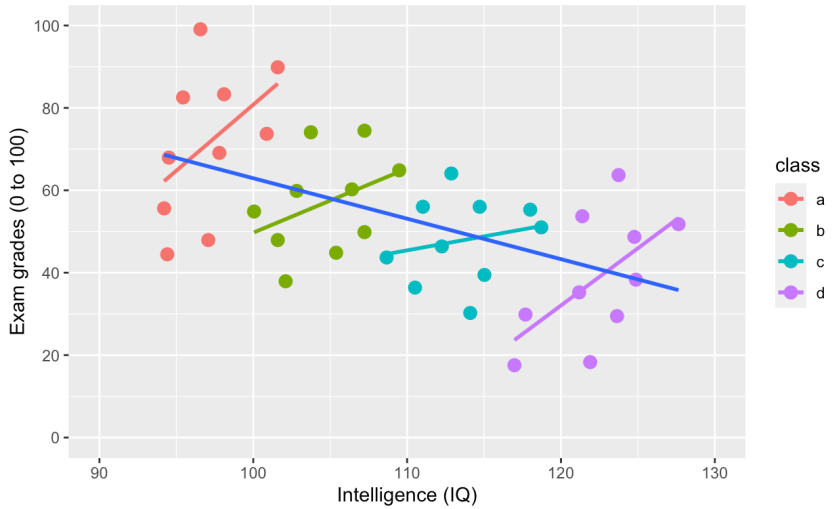
Does this line better "fit" the data?

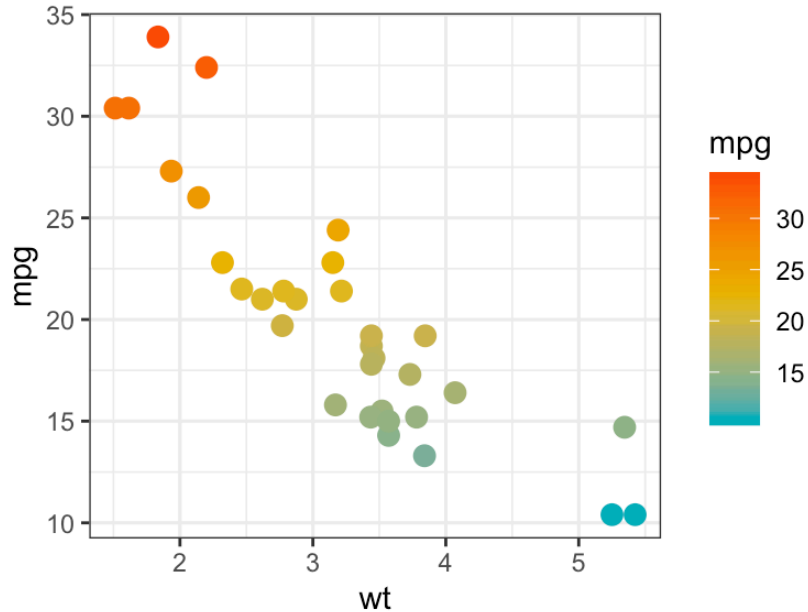


Linear vs Nonlinear Relationships

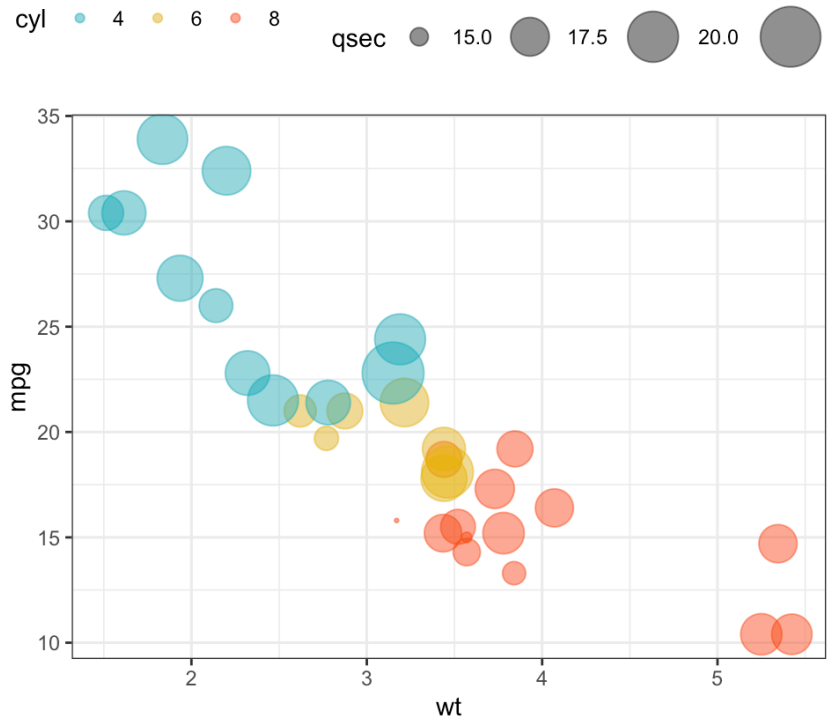


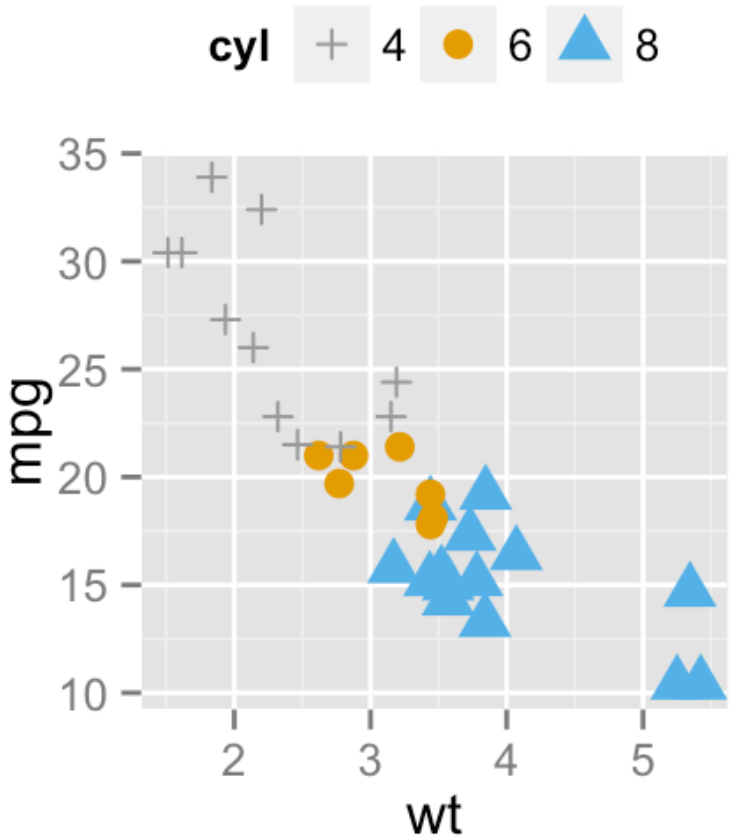
Can I compare 3 or more variables in a single plot?



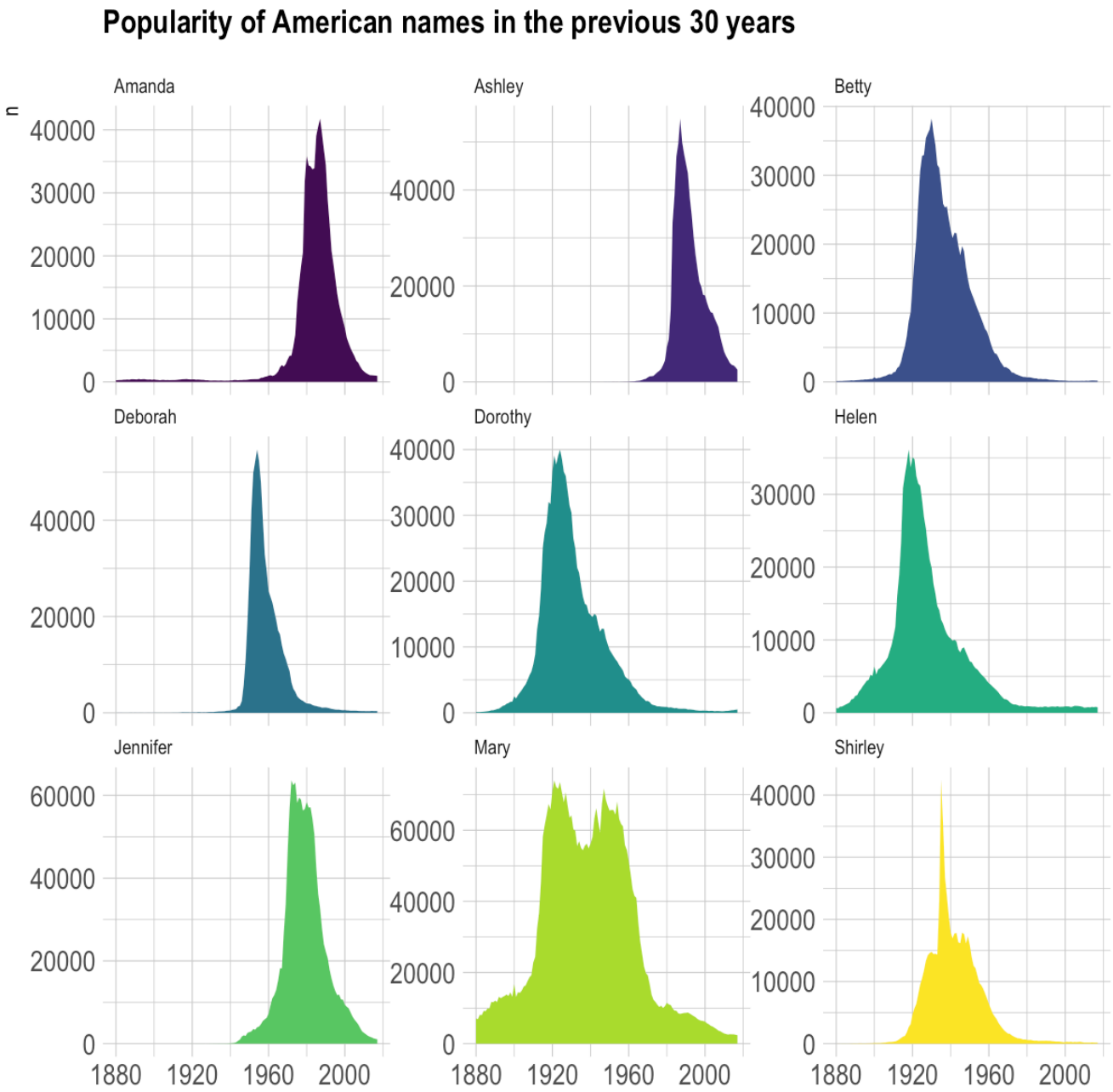


Changing shapes and sizes of points





Split into separate plots with faceting



Homework 2

- Describing numerical distributions: modality, skew, outliers
- Describing numerical distributions: appropriate summary statistics
- Matching numerical distributions to their summary statistics, reading a boxplot
- Calculating proportions from a contingency table

Homework 2

- Will be posted Friday 9/13/2024 after class
- Will be due Friday 9/20/2024 by 6:00pm
- I will post another instructional video for this homework, in addition to the one I did last time. If you're having trouble getting started, try watching both.
- Late policy: "This homework is due by 6:00pm on Monday, 9/9/24. No credit will be lost for assignments received by 7:00pm to account for issues with uploading. 10% of the points will be deducted from assignments received by 9:00am on Tuesday, 9/10/24. Assignments turned in after this point are only eligible for 50% credit, so it benefits you to turn in whatever you have completed by the due date."

How can I get help with homework?

- ***Read the textbook.*** Many of you are asking for additional examples. Luckily, there are tons we didn't go over in the textbook.
- ***Look at the homework early***. I can see in Canvas that many students didn't download the documents until 1-2 days before it was due. That's not a lot of time to get help.
- ***Ask a question on our Campuswire class feed.*** I'm only one person, and I may not be able to give you a prompt answer. However, the 28 other people in the class might be able to.
- ***Come to office hours.*** I will be available after class Monday 9/23/2024 and Wednesday 9/25/2024 from 2:30pm - 4:00pm. If you cannot make it, reach out to me to try and schedule an appointment.

Next Week: Chapter 3, Probability

- What is probability?
- Disjoint vs not disjoint sets
- Probability distributions
- Complements and independence