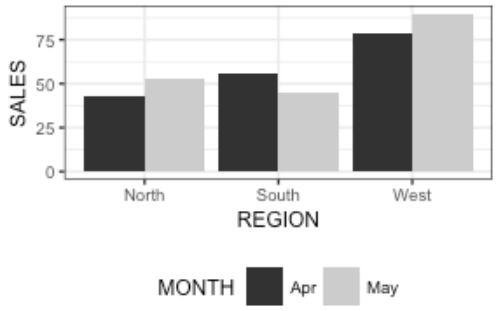
ANSWERS

1. The graph below was drawn from the data frame on the right using a layered grammar of graphics approach.



MONTH	REGION	SALES
Apr	North	43
Apr	West	79
Apr	South	56
May	North	53
May	West	90
May	South	45

Indicate the following:

a. all aesthetic mappings (use arrows) REGION → X, SALES → Y, FILL → MONTH

b. geom BAR

c. stat IDENTITY

d. position DODGE

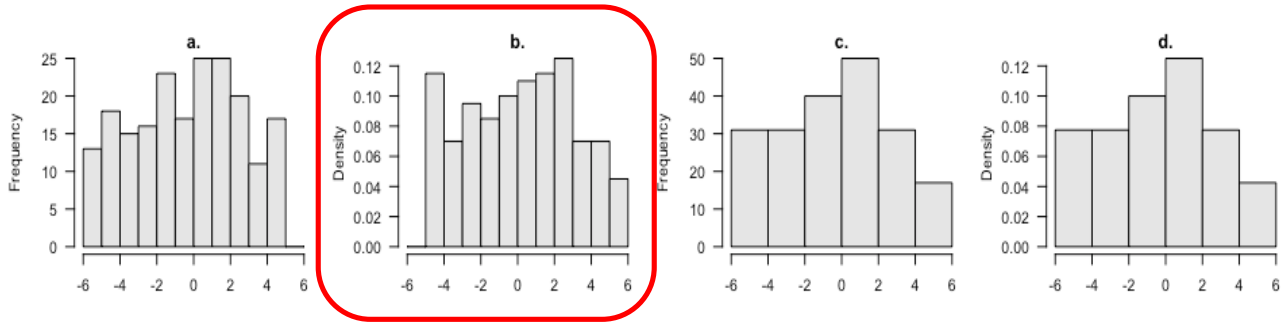
2. Would a diverging stacked bar chart be a good choice for plotting the following data on student loan default rates by school in North Carolina? Why or why not?

No, since the data values in the chart do not have a clear middle value and two extremes, they are not well suited to a diverging stacked bar chart display.

3-Year Cohort Default Rates [1]	FY2005	FY2006	FY2007	FY2008	FY2009	FY2010	FY2011
ASU	1.3	1.3	3.7	2.96	3	3.6	4.8
ECU	2.5	2.4	3.8	3.58	3.2	4.6	6.7
ECSU	17.7	19.9	19	20.24	16.8	22.1	22.9
FSU	16.5	13.2	20.8	21.29	17.4	17.2	13.2
NCA&T	14.5	14.3	16.3	12.93	13.5	16.8	18
NCCU	11.1	14.2	13.9	14.78	14.9	17	16.8
NCSU	2.2	1.8	3.8	2.79	3.2	3.6	4.1
UNCA	3.3	3.7	4	4.95	5.6	6.5	7.3
UNC-CH	0.4	0.5	1.6	1.16	0.8	1.6	2.3
UNCC	2.4	2	3.9	3.02	2.8	5.5	6.2
UNCG	3.1	2.9	4.3	3.69	3.9	6.7	7.5
UNCP	5.2	6.4	8.3	5.99	6.2	11.4	15.9
UNCW	3.3	5.1	4.9	4.8	4.5	5.4	6.7
UNCSA	4.7	4.2	5.6	3.78	6.4	7.9	6
WCU	4.3	5.7	6.5	6.37	6	9	8.1
WSSU	10.3	10.2	15.4	13.87	12.1	16.3	16

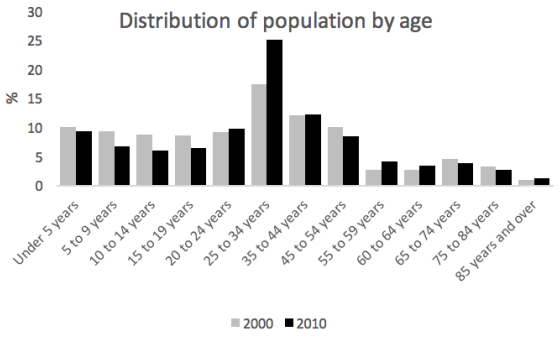
[1] Students who defaulted on loans within three years of leaving school, either through graduation or dropping out.

3. Three of the histograms below are right closed and one is right open. All are drawn with the same dataset. Which is the right open one? B



ANSWERS

4. Based on graph below, what is your best estimate of the ratio of 29 year olds to 22 year olds in this population in 2010?



% of 29 yo: approx. 25/10 = 2.5
% of 22 yo: approx. 10/5 = 2

ratio: 1.25

5. A dataset includes the following variables: SALESAMOUNT (num), STORE NAME (factor), TIME OF SALE (num), DAY OF WEEK (factor). A histogram of the SALESAMOUNT variable is bimodal. What would be a reasonable next graph in an effort to explain the cause of the bimodality?

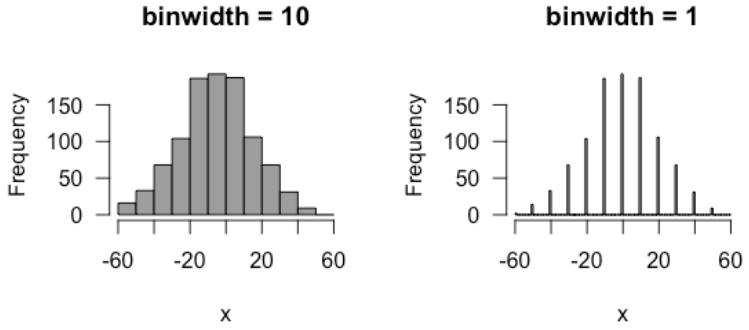
Anything that conditions SALESAMOUNT on another variable, such as:

- multiple histograms faceted on STORE NAME or DAY OF WEEK
- a scatterplot of SALESAMOUNT ~ TIME OF SALE.

6. Briefly describe a graphical technique to identify patterns of rounding in a dataset.

Gaps in a histogram with small bin widths will indicate rounding patterns.

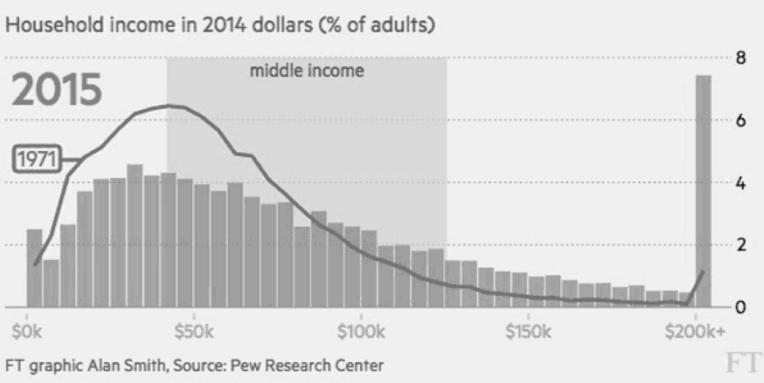
(Example:)



ANSWERS

7. Sketch a sample histogram that shows evidence of topcoding. Include axis and tick mark labels.

The histogram should have a large bin at the right end and a "+" category, as below:



8. Which of the following plot types shows information that is most similar in nature to that shown by a stem-and-leaf plot?

a. histogram

b. QQ plot

c. parallel coordinate plot

d. boxplot

9. Imagine you have data in the form shown below that indicates the number of coffee drinkers in a sample by quantity consumed per day (1-2 CUPS, 3-4 CUPS, 5+ CUPS) and by whether the drinkers view caffeine as harmful or beneficial (HARM, BENE). The graphics program that you plan to use requires that the data be tidy. What will the tidy form look like? Draw the tidy data frame, creating columns / column names as necessary.

	1-2 CUPS	3-4 CUPS	5+ CUPS
HARM	6	9	0
BENE	9	2	1

Correct as well to try to create a data frame of "cases" with 27 rows (each row = one person)

VIEW	CONSUMPTION	FREQ
HARM	1-2 CUPS	6
BENE	1-2 CUPS	9
HARM	3-4 CUPS	9
BENE	3-4 CUPS	2
HARM	5+ CUPS	0
BENE	5+ CUPS	1

ANSWERS

10. The categories for traveler ratings on TripAdvisor are "Excellent", "Very Good", "Average", "Poor", and "Terrible". What type of color scheme would be the best choice for displaying this measure, if it's important to easily distinguish between the positive and negative ratings? Either provide the type of color scheme or a specific example ("Green" for "Excellent", etc.)

a **diverging** palette

11. Name one technique for improving the chances that individuals with color vision deficiencies will be able to distinguish colors in the graphs you create.

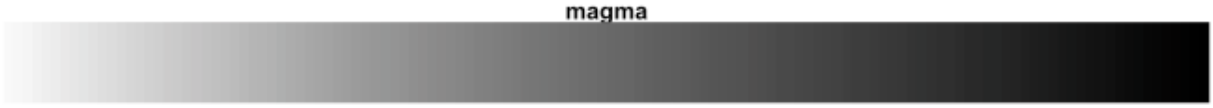
- use a color palette known to be color vision deficiency (CVD) friendly
- run your graphics through a CVD simulator such as VisCheck.com
- draw graphics in black & white
- use shapes as well as color to distinguish categories
- vary lightness

("Avoid red-green palettes" is not correct – we specifically discussed that the issue is much more complex than this.)

12. Is the grey palette shown below perceptually uniform? Why or why not?

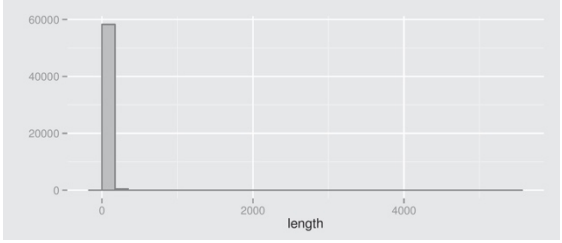
No. The transitions are not smooth. There are clear jumps and prominent bands.

(Compare to the following, for example.)



13. a) Describe the distribution of the variable displayed in the histogram to the right. (The axes were automatically chosen to fit the data, that is, not adjusted manually.)

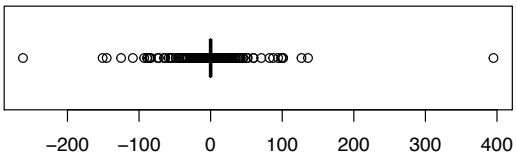
Most of the data is below 200 (approximately) with one or more high outliers.



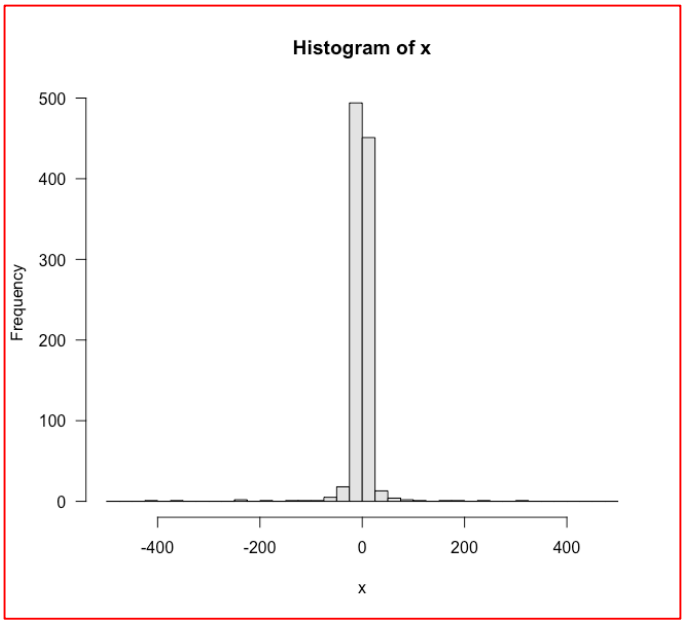
b) What would be a reasonable next step to get a better sense of the distribution of this variable?

- Remove the outliers and redraw the histogram.
- Transform the length variable (such as log scale)

14. Draw a (rough) histogram of the dataset depicted in the boxplot below, which contains 1000 values.



Most of the data should be very close to zero, with a few outliers, as shown.



15. Which of the following is best for comparing medians of subgroups?

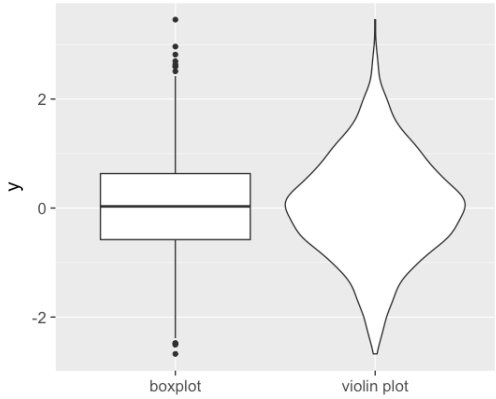
a. faceted histograms
b. faceted QQ plots
c. parallel coordinate plot
d. multiple boxplots

ANSWERS

16. Name one advantage of a violin plot over a boxplot. Either answer in general or provide a specific example.

Advantage of violin plot: better for viewing the shape of the distribution

Advantage of boxplot: better for comparing quartiles



17. What issue/problem does jittering of data points address?

overplotting (overlapping points)

18. Name one problem that can arise when drawing histograms of discrete data.

- equal bin widths might contain different numbers of discrete values
- since there's no standard on whether histogram bins should be right open or right closed, histograms with data values that fall on bin borders will be ambiguous

19. Describe one method for ordering the bars of a bar chart that in general is better than alphabetical ordering.

by length of bars (frequency)

20. a) Name a key difference in purpose between a bar chart and a histogram.

Bar charts show comparisons of values; histograms show shapes of distributions.

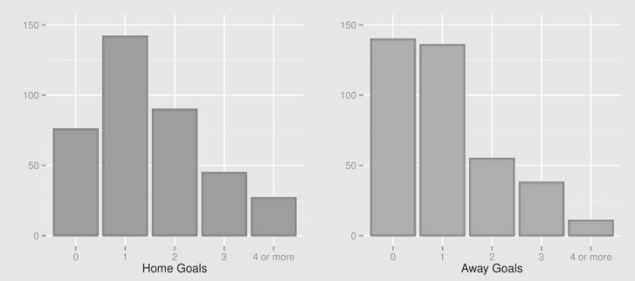
b) Name a key difference in appearance between a bar chart and a histogram.

Bar charts have spaces between bars; histograms don't, with the possible exception of histograms of discrete data.

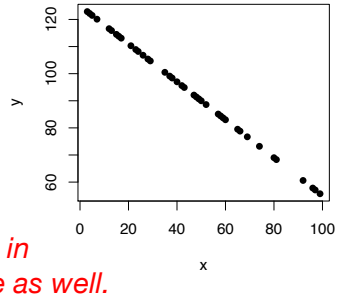
ANSWERS

21. In terms of good graphing practice, is it ever acceptable to include an "or more" bin in a histogram, such as "3000+"? Why or why not?

It's ok as long as there is not much data in the bin, such as here (the last category in each is "4 or more"):

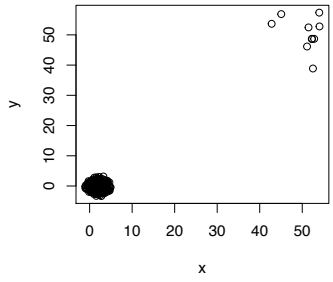


22. The scatterplot to the right is drawn from real world data. Describe the relationship between the variables.



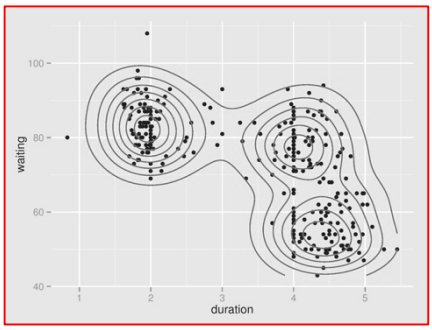
The best answer would be to point out that two measurements in real world data would never be perfectly correlated as shown, and therefore, one is very likely a direct transformation of the other (such as bodyfat and body density measurements, which came up in a homework problem.) However, negative correlation is acceptable as well.

23. The graph to the right is a scatterplot of 1000 points. Briefly describe what you observe and what you would do next to get a better sense of the relationship between the variables.



Most of the points are overlapping. There appear to be two groups of data. I would divide the data between the ball on the left and the group of points on the right and draw new scatterplots.

24. What new information is gained by added density estimate contours to a scatterplot? Describe a hypothetical or real example of the added information.



Density estimate contours might show clusters in scatterplots.

ANSWERS

25. A regression line added to a scatterplot is, in essence, a visualization of:

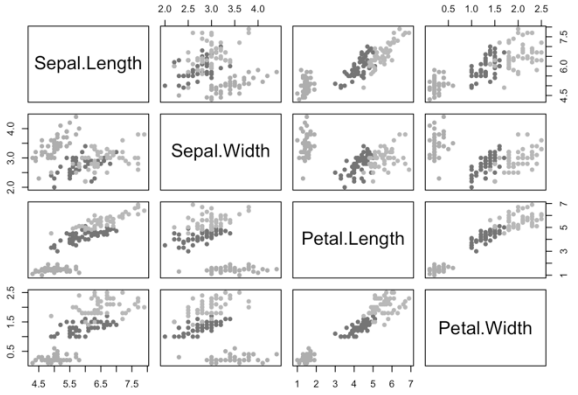
a. a split between clusters
b. a model
c. the interquartile range
d. the rejection region

26. Name one advantage of a parallel coordinates plot over a scatterplot matrix. Either answer in general or provide a specific example.

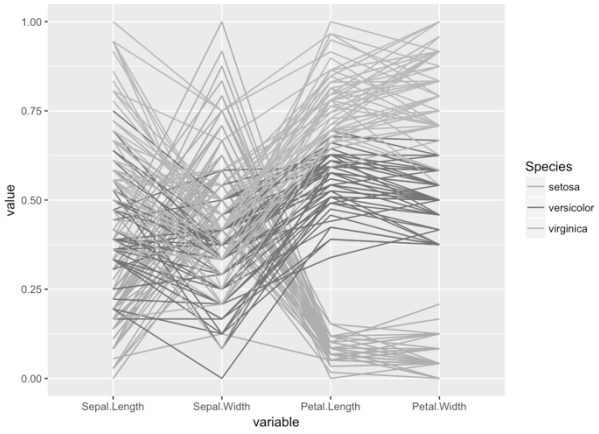
A scatterplot matrix only shows relationships between two variables at a time, whereas a parallel coordinates plot shows trends among multiple variables.

Example:

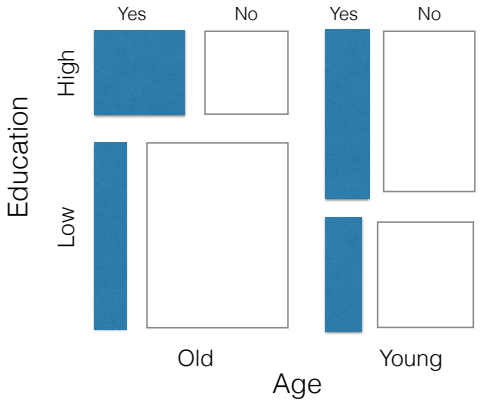
scatterplot matrix



parallel coordinates plot



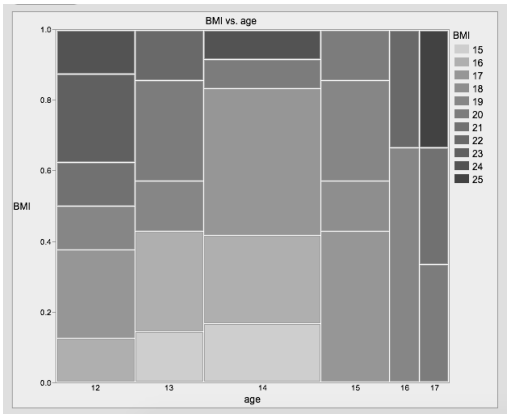
27. In the mosaic plot to the right, "Yes" and "No" refer to whether the individuals in the sample listen to (yes) or do not listen to (no) classical music. Do education and/or age appear to be correlated with classical music listening? Does there appear to be an interaction effect?



There appears to be an interaction effect: people in the **old** Age group with **high** Education level listen to classical music at a higher rate than people in groups formed by all the other combinations of Age and Education levels, which all appear to listen to classical music at approximately the same rate.

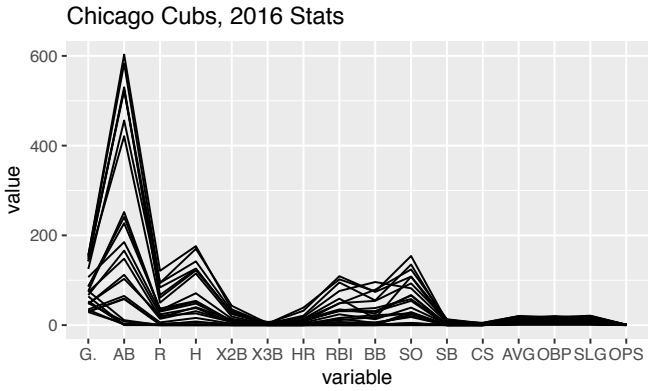
ANSWERS

28. BMI and age are continuous variables. What type of plot would be a better choice for exploring the relationship between these two variables than the one on the right? Provide a reasonable hypothesis for why a mosaic plot was chosen in this case.



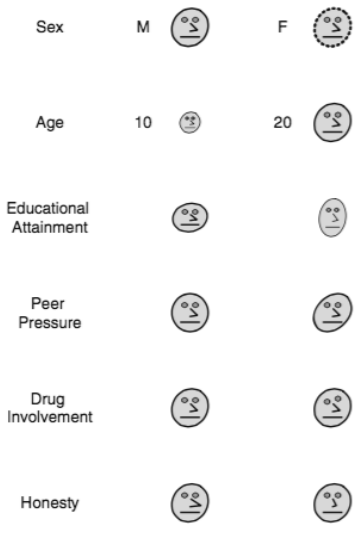
A scatterplot would be better. It is likely that the data was rounded to the nearest integer so it wouldn't have been possible to create a meaningful scatterplot. Therefore, a mosaic plot, generally used for categorical data, was employed instead.

29. What single adjustment (data transformation or parameter change) would most likely improve one's ability to detect patterns in the plot to the right?



Change the scale: either normalize the data or transform to a scale of 0 to 1.

30. Critique the graph on the right.



It's impossible to read. There's no legend explaining what any of the facial features mean, and the meaning is not intuitive at all.

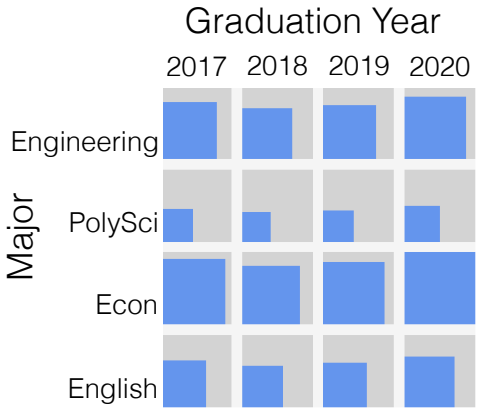
ANSWERS

31. A slope graph is a simple version of a(n):

a. mosaic plot
b. alluvial diagram
c. scatterplot matrix
d. parallel coordinates plot

32. The darker boxes in the fluctuation diagram to the right represent the number of students in a particular college by graduation year and major. Does graduation year appear to be correlated with major at this college?

No. While the class sizes appear to differ slightly, within each class, the same proportion of students are majoring in each of the majors listed.



33. Name one advantage of a Cleveland dot plot over a bar chart. Either answer in general or provide a specific example.

- A dot plot can show many more categories.
- A dot plot doesn't need to start at zero whereas a bar chart does.
- A dot is less cluttered.