

STA 141 - Exploratory Data Analysis and Visualization

Derek L. Sonderegger

September 05, 2019

Contents

Preface	5
1 Week 1	7
1.1 Introduction	7
1.2 EPTs and Gestalt	8
1.3 Practicum #1	13
2 Week 2	17
2.1 Amounts	17
2.2 Distributions	23
2.3 Proportions	29

Preface

These are the lecture notes for STA 141 - Exploratory Visual Data Analysis. This course is intended to teach students how to think critically about problems, examine data that can provide answers, and create graphs that are insightful, and ask follow-up questions to the visual analysis. Also, because disinformation can be disguised to seem credible, students will also be exposed to various data visualization tricks and statistical malarky that propagandists.

This course draws information from several websites and books.

- Alberto Cairo's book *The Truthful Art*. Alberto Cairo is the Knight Chair in Visual Journalism at the School of Communication of the University of Miami (UM), where he heads specializations in infographics and data visualization. You can find more about him at thefunctionalart.com. Also at his website, his blog highlights data visualizations in the news. Many of the graphs I will shown in this class are featured here.
- Claus O. Wilke has a book *Fundamentals of Data Visualization*. There is an online pre-print version of the book available [here](#). He also has a nice R package called *cowplot* that aids in making publication ready graphs using R's *ggplot2*.
- *Calling Bullshit* This is a 3-credit course taught at Univeristy of Washington. Their goal is to teach students to recognize bullshit provide another scientist a reason why a claim is bullshit.
- Amelia McNamara's SDS 136 course. Amelia is an Assistant Professor at University of St Thomas and I've been influenced by her presentations at national conferences. These are her notes from a data visulization course that she teaches. She is also pretty fun to follow on twitter.

Chapter 1

Week 1

1.1 Introduction

1.1.1 Initial thoughts.

- What is data?
- Why visualize it?
 - See relationships that raw data obscure.
 - Cognitive work to translate raw numbers into context between other data points are already done for you.
 - Good graphics translate particular numerical relationships into physical relationships which our brains are really good at processing.

1.1.2 Amazing Graphics

- UK Drug Poison These are the data from the UK related to drug overdose and misuse. Farther down the page gives a similar graph on suicide.
- Migration Patterns in Europe.

1.1.3 Bad Graphs

- While the New York Times is generally really quite good, this graph is quite misleading. Another view of the data is more fair, although the still depressing.
- I feel compelled to show a bad 3-d Excel graphic as well.
- Reuter's infamous gun deaths in Florida chart.

- An amusing case where a pie chart is ridiculous. This is the result of a survey that asks what pizza toppings are liked. In particular, a person can pick more than one topping and so the percentages don't sum to 100%.

1.1.4 Tableau or ???

- Tableau is a nice program that reads in data and can produce some very nice graphics and dashboards.
 - Licensing Questions?
 - What are dashboards? A series of related graphs, often with controls that allow you to explore the data.
 - Britain's Coal Use 2015-2019

1.2 EPTs and Gestalt

- Some Visual tasks are easier than others.

From Hadley Wickham's Stat 405 at Rice. (Slides 34 - 40) Effective Visualizations

1.2.1 Groupings / Gestalt

The way we organize our graphics can lead a viewer to create mental groups of marks.

Winona State's Data Visualization PowerPoint

Slides 30-50

- Enclosures
- Connections
- Proximity
- Similarity (color/shape)

Example: Warpbreaks While spinning wool into thread, if the tension on the wool isn't correctly set, the thread can break. Here we compare two different types of wool at three different tensions.

Cairo/EPT.bb

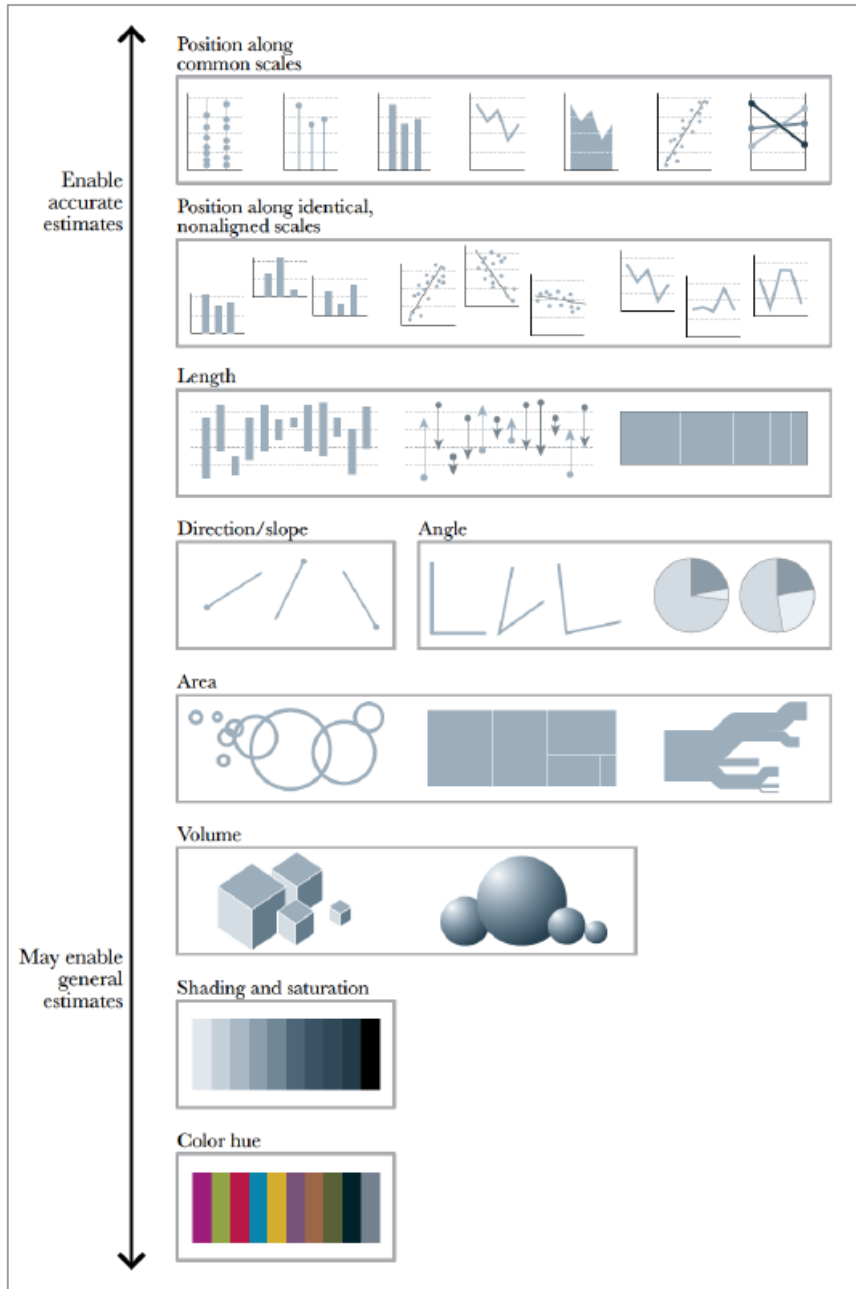
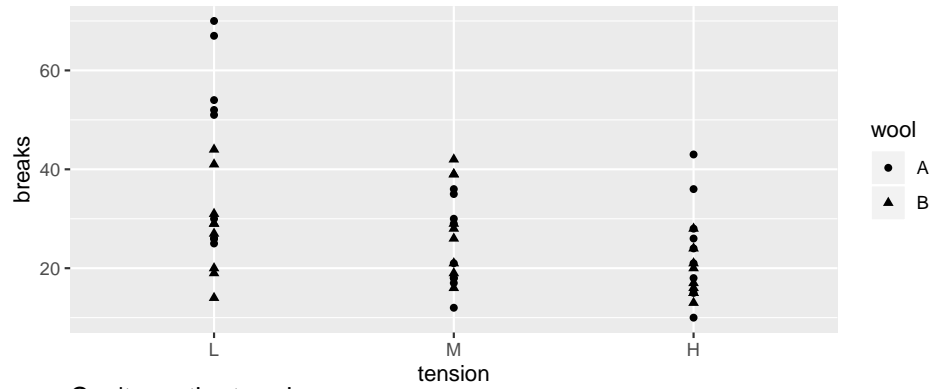


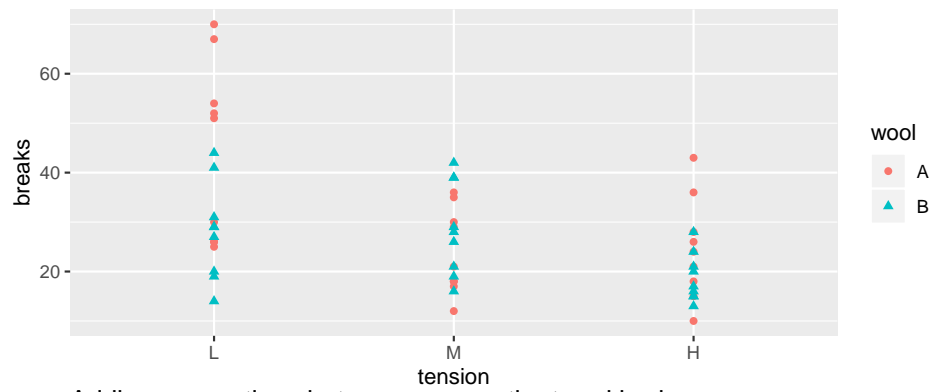
Figure 5.5 Scale of elementary perceptual tasks, inspired by William Cleveland and Robert McGill.

Figure 1.1: From Alberto Cairo's "The Truthful Art"

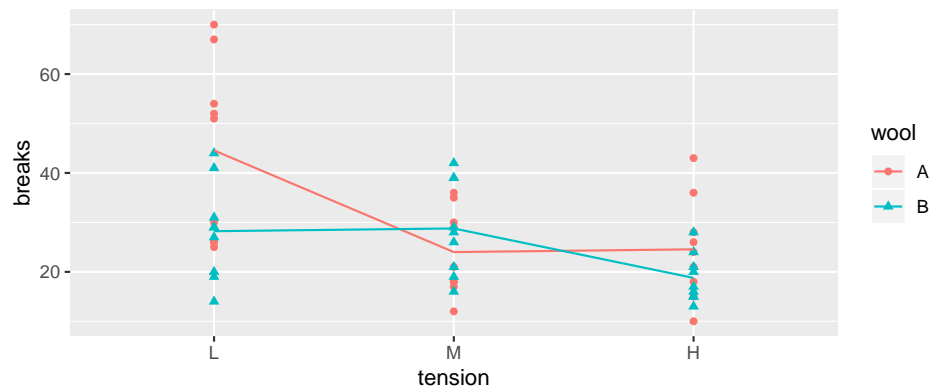
Can't easily distinguish wool types.

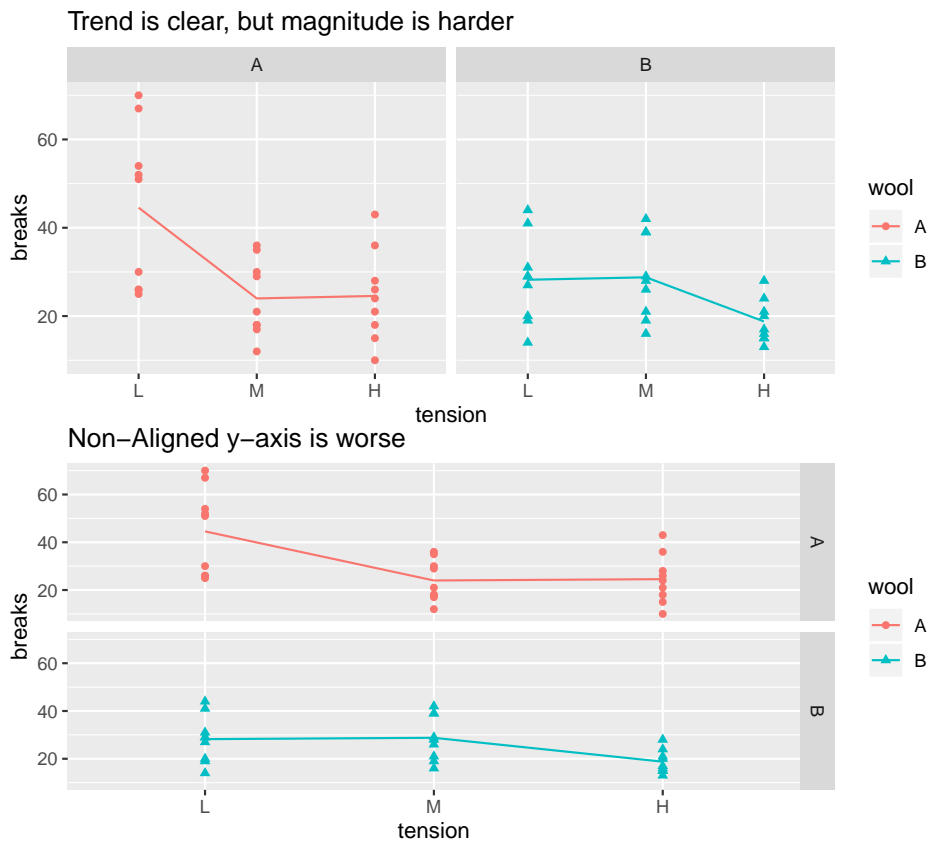


Can't see the trend.



Adding connections between means, the trend is clear





1.2.2 “Color” Scales

Defining Color really has three different attributes (From Wikipedia).

1.2.2.1 HSV Scale

- Hue: The attribute of a visual sensation according to which an area appears to be similar to one of the perceived colors: red, yellow, green, and blue, or to a combination of two of them.
 - Saturation: The “colorfulness of a stimulus relative to its own brightness”
 - Value: The “brightness relative to the brightness of a similarly illuminated white”
-
- Hue is appropriate for categorical variables.
 - Saturation and/or Value is appropriate for a quantitative variable scale.

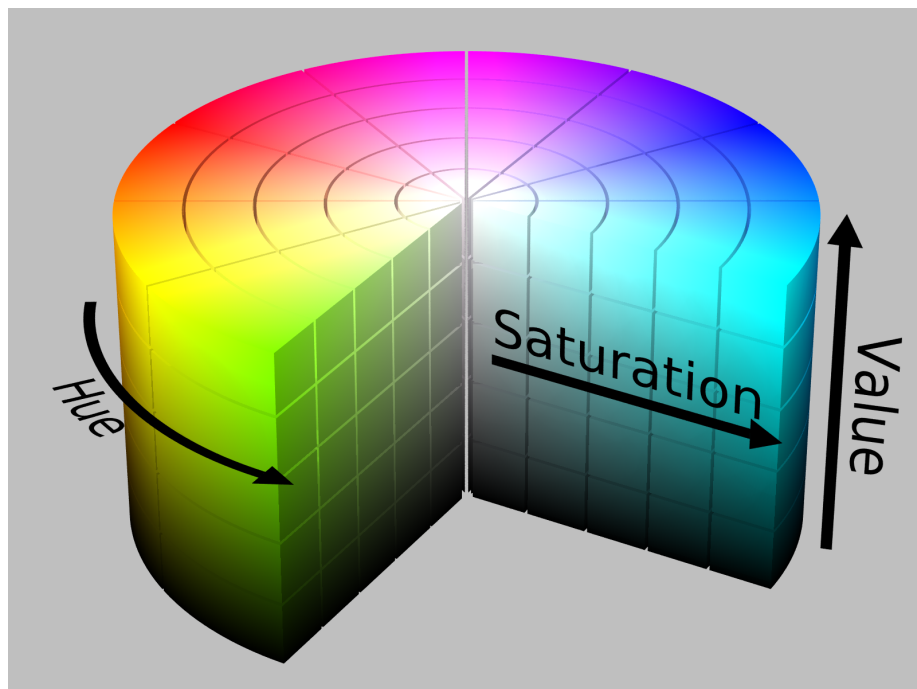


Figure 1.2: HSV Cylinder from Wikipedia

Neither R nor Tableau make it particularly easy to map these aspects, so we won't get too deep into it.

1.3 Practicum #1

1.3.1 How to Store Data

Data is commonly stored in spreadsheets.

- Columns are variables of interest
- Rows are observations.

Example: A dataset we'll call **iris** which has 150 observations of three species of iris. Each observation measured the length and width of both the petals and sepals.

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

Example: A dataset we'll call **GradeBook** that has records of how well a student performed on exams. I'll refer to this storage as the *wide* orientation.

StudentID	Exam 1	Exam 2	Final Exam
1	87	87	81
2	91	88	85
3	88	79	92
4	91	97	94
5	100	83	90
6	85	79	81

Or I could have stored the information in the following manner, which I'll refer to as the *long* orientation.

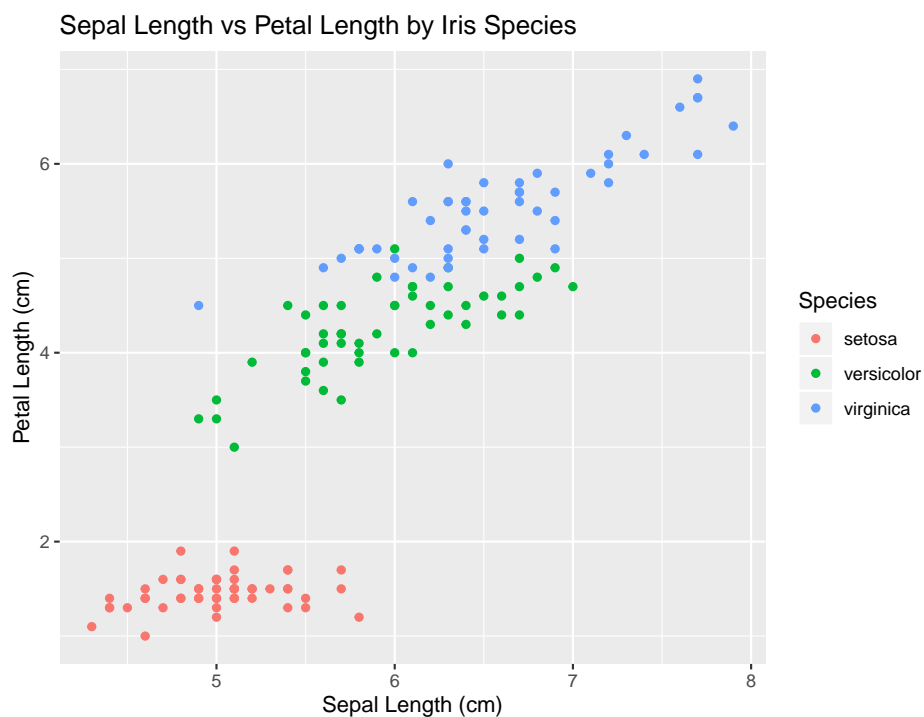
StudentID	Assesment	Score
1	Exam 1	91

StudentID	Assesment	Score
1	Exam 2	90
1	Final Exam	87
2	Exam 1	73
2	Exam 2	76
2	Final Exam	53

1.3.2 Tableau

1.3.2.1 Task 1: Dragging variables onto destination

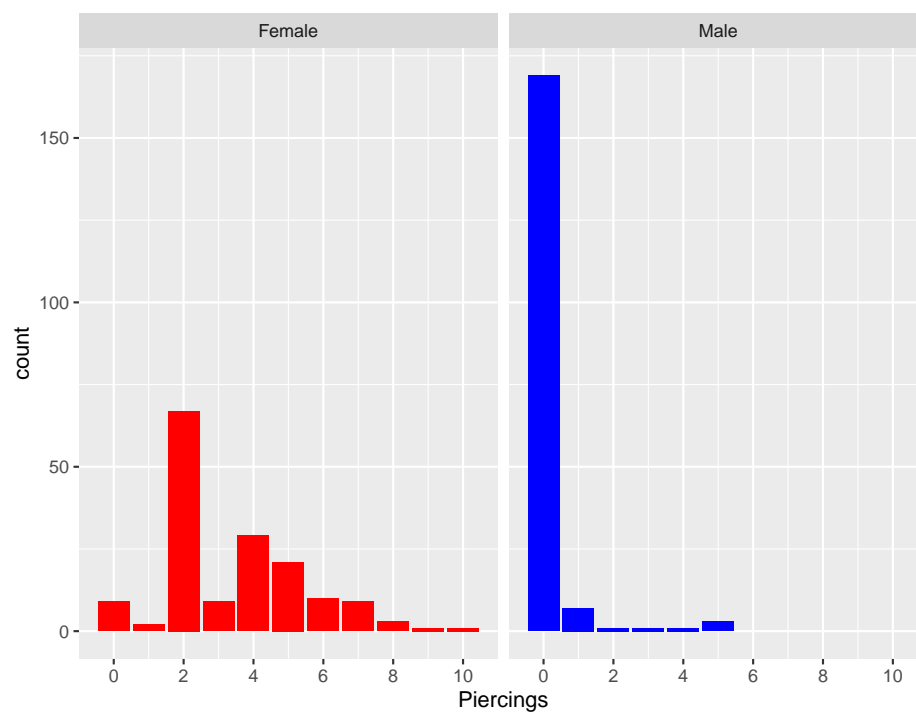
First we'll make a scatterplot with some colors.



The iris data I used for this graph is available here: data: <https://raw.githubusercontent.com/dereksonderegger/141/master/data-raw/iris.csv>

1.3.2.2 Task 2: Modifying how a variable is displayed

https://raw.githubusercontent.com/dereksonderegger/141/master/data-raw/Lock5_GPAGender.csv



1.3.2.3 Task 3: Reorder categorical variable levels

<https://raw.githubusercontent.com/dereksonderegger/141/master/data-raw/warpbreaks.csv>



Chapter 2

Week 2

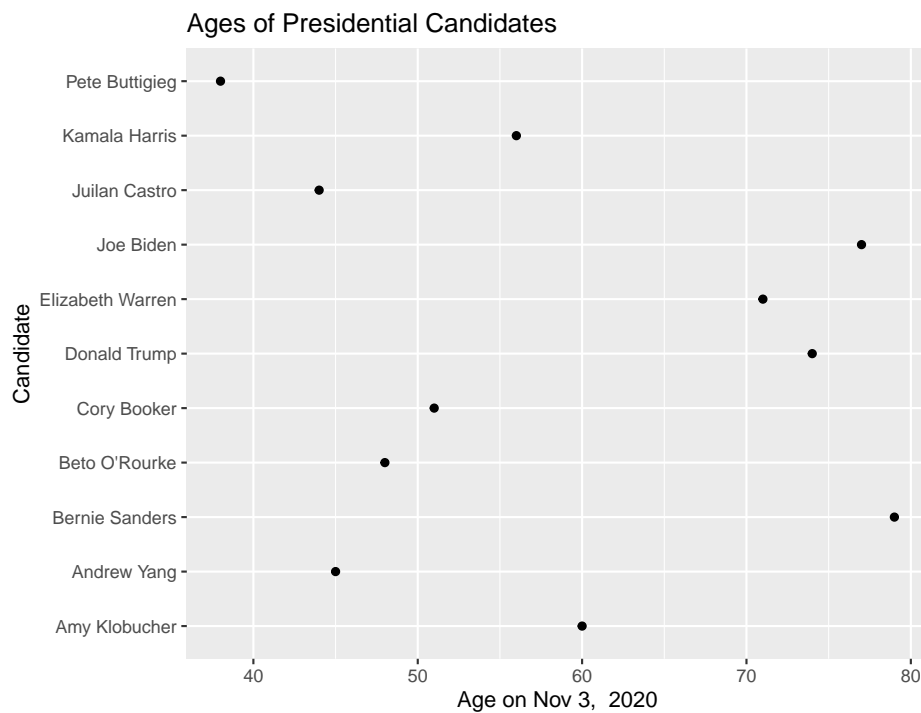
Assignments

1. Read Chapter 6 and 7 from Claus Wilke's Fundamentals of Data Visualization book.
2. Watch Hans Roslings TED talk. Comment on two aspects of the video that stand out to you regarding how he presented his data and how he talked through the visualization with the audience.
3. Graph the NOAA CO₂ data over time. The National Oceanic and Atmospheric Administration has monthly CO₂ atmospheric levels data available to the public at NOAA's website. However, this is a little obnoxious to get into Tableau, so I've done some data wrangling for you already. The `Date2` column has the date information encoded using a continuous decimal scale. CO₂ is measured in parts per million.
 - a) Plot CO₂ over time while showing the monthly trend. Should we use area or lines? Why?
 - b) Create a graph that shows a single mark per year (so average over all observations in a year). Do you like this better? Explain your reasoning.
 - c) Create a follow-up graphic that shows the monthly trend. Explain what you want to demonstrate and why you chose to display the information as you did.

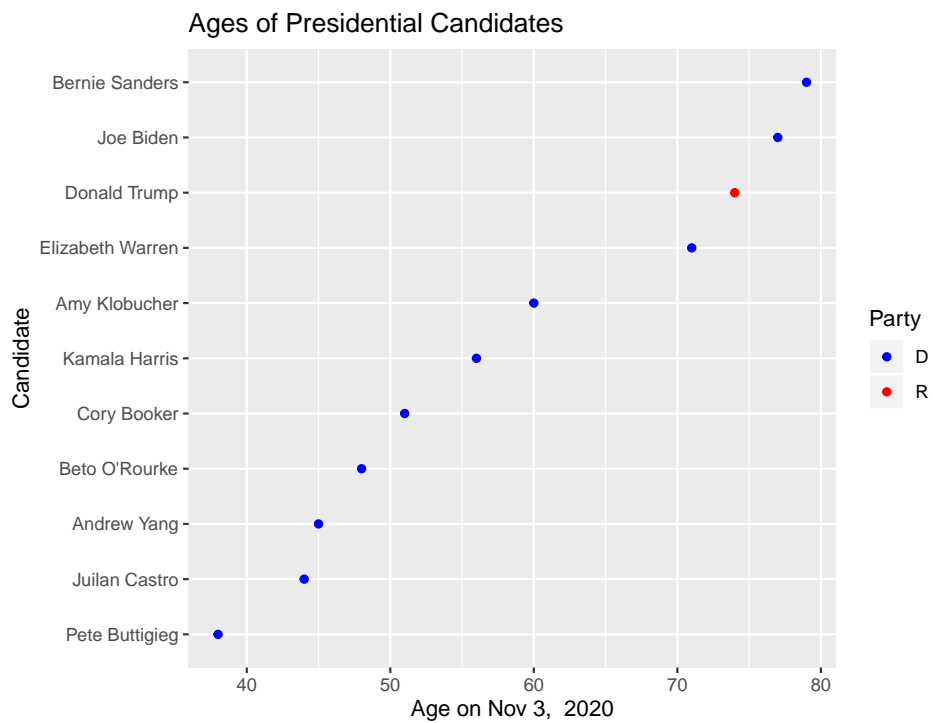
2.1 Amounts

The 2020 presidential candidate field has a wide range of ages. The New York Times has a nice article showing the candidate ages. I grabbed a few of the

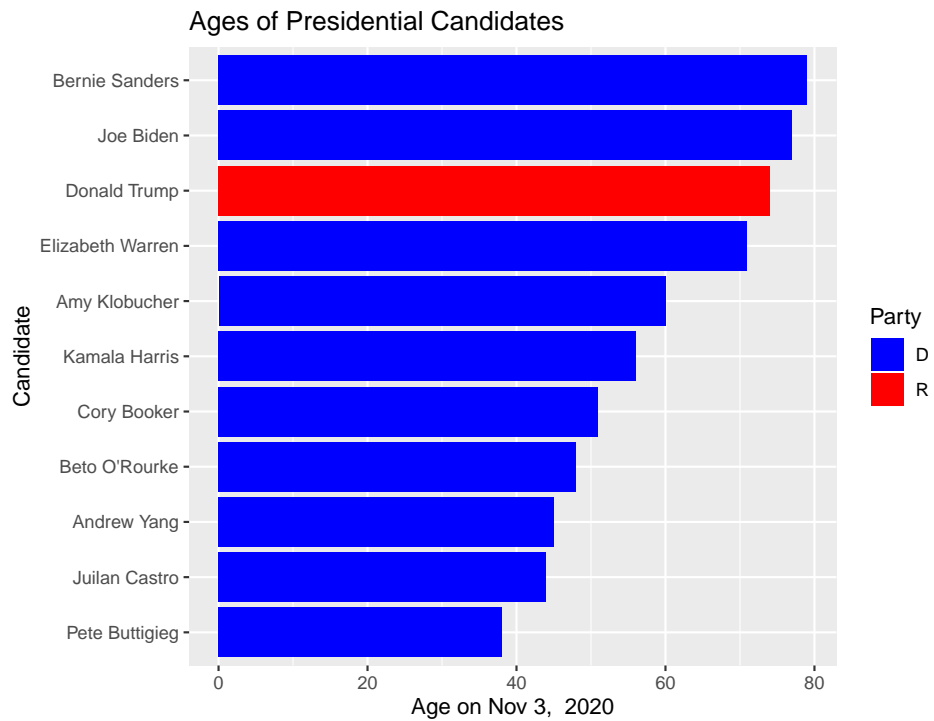
most prominent candidates and pulled their birthdays from Wikipedia and then calculated their age on election day.



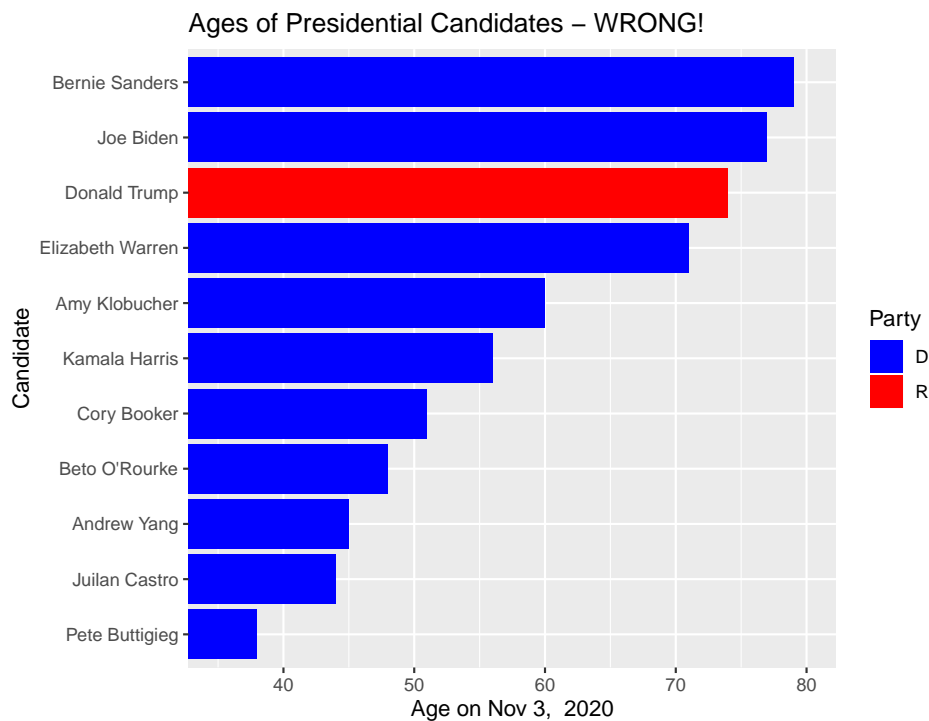
The order of the candidates is useless. Here we have ordered them alphabetically when we should try to think about an ordering that improves clarity. Lets switch to sorting the candidates by age.



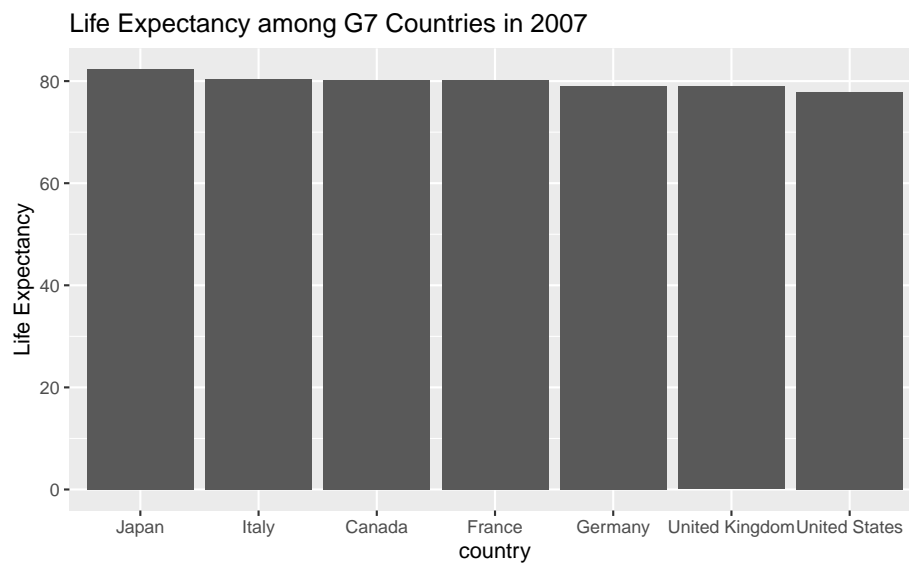
This isn't too bad, but it fails to visually impress the differences. A bar chart should visually impress the ages based on the length of the bar so that we can't have to keep looking at the Age axis.



What would be *dishonest* is if we were to chop off the bars at 35 or 40 to make the age difference between Buttigieg and Warren, Trump, Biden and Sanders seem huge.

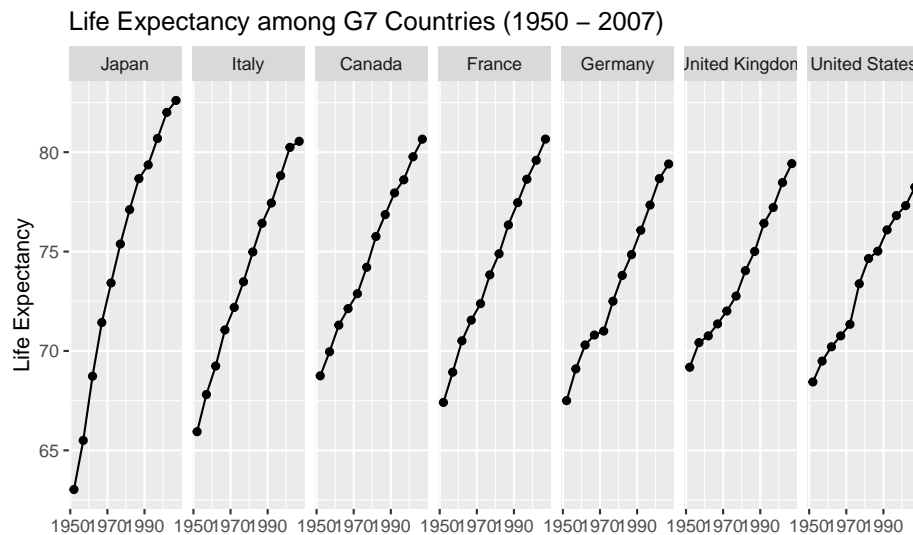


Often we need to graph some value and want to know how it varies among *two* different categories. In these cases, we have to employ some sort of grouping strategy.



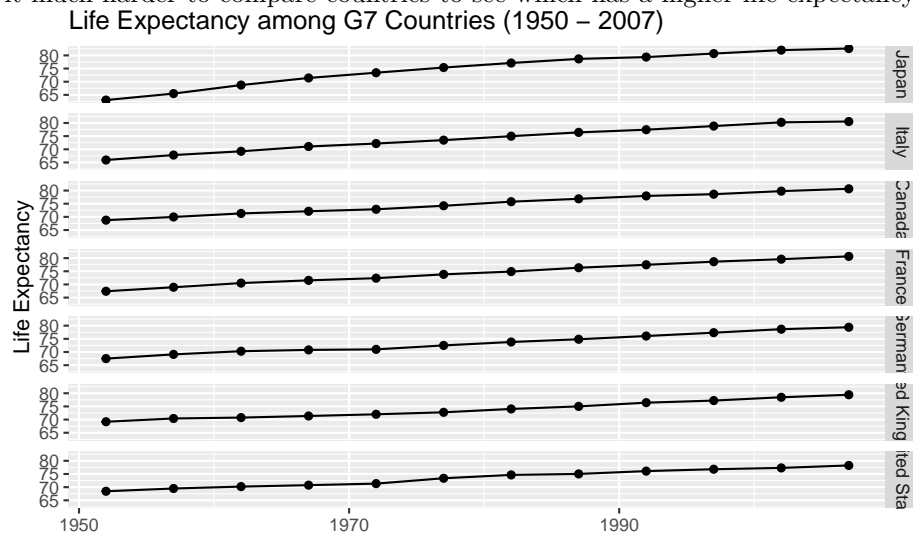
The bar chart here is obscuring the differences in life expectancies because the

numbers are so close. In this case, I think points make more sense. Also I want to see how life expectancy has changed since World War II.

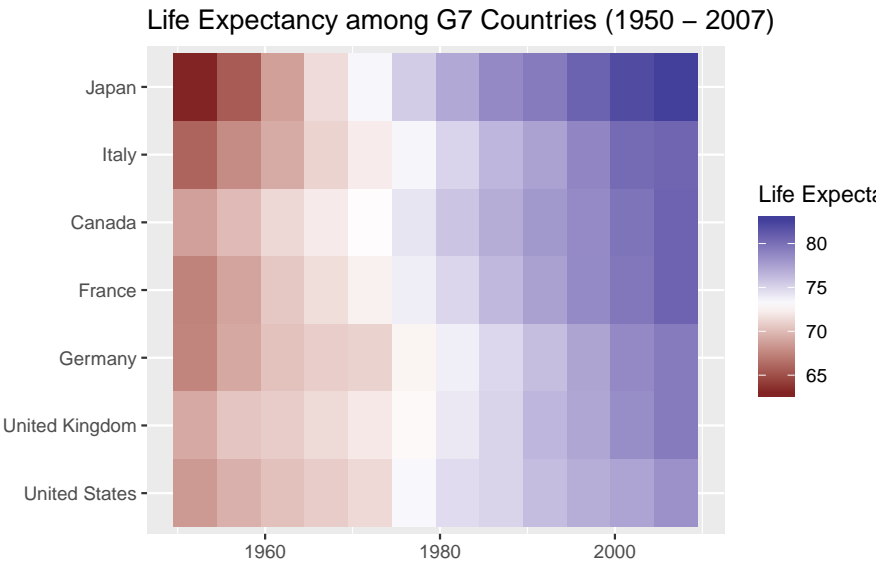


In the above graph, I am grouping countries both by enclosure and with a physical path connection. The reader tends to see the line as a whole object and compare the line max/min and slope among the seven countries.

We might consider changing the faceting to stack the countries, but this makes it much harder to compare countries to see which has a higher life expectancy.



A heat map makes it easier to see which country has the highest life expectancy,



but we lose precision in the actual values.

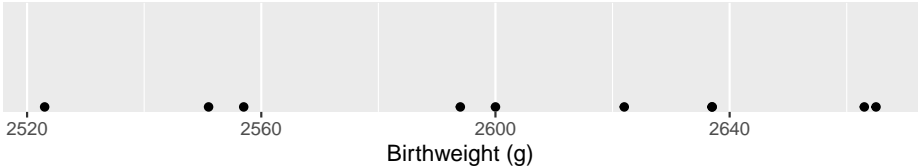
2.2 Distributions

Given a single variable, I often want to know what values are common and what values are rare. To visualize this, we will primarily compare marks along a common axis (the most accurate EPT!)

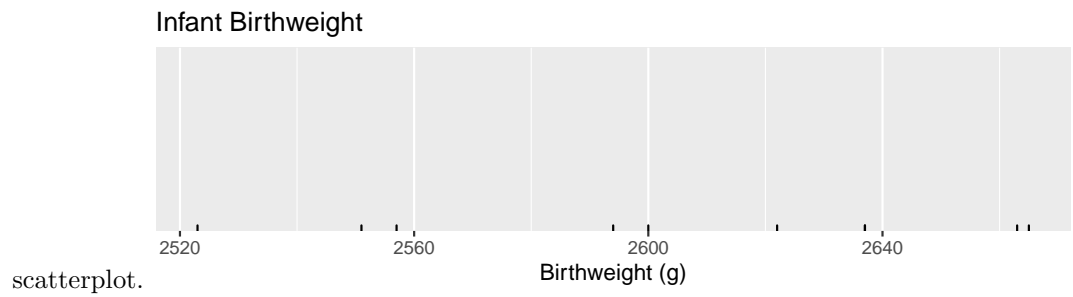
2.2.1 Small samples

If we only have a few observations, then we can just graph them along an axis.

Infant Birthweight

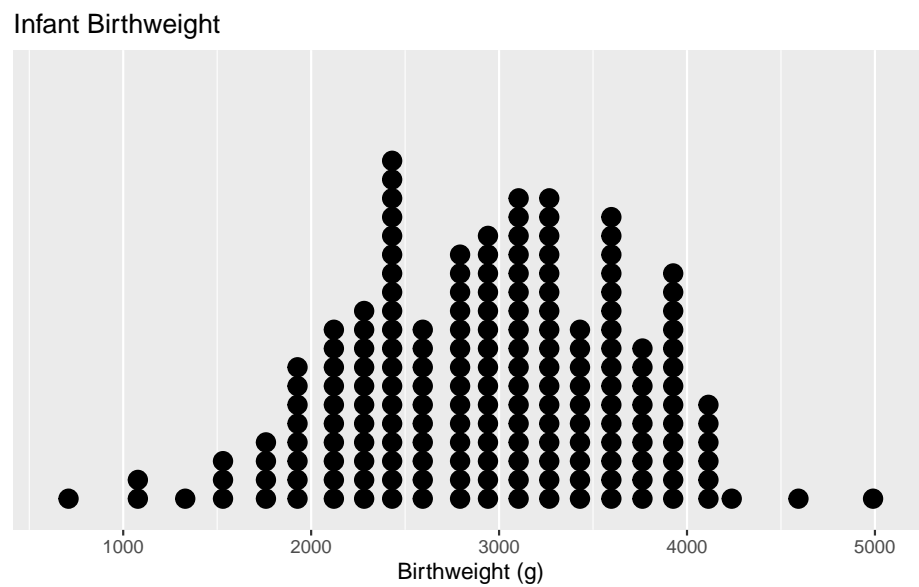


Another trick that works with more data, is to not use dots but rather lines. This is called a rugplot. This is often used in conjunction with another graph such as a

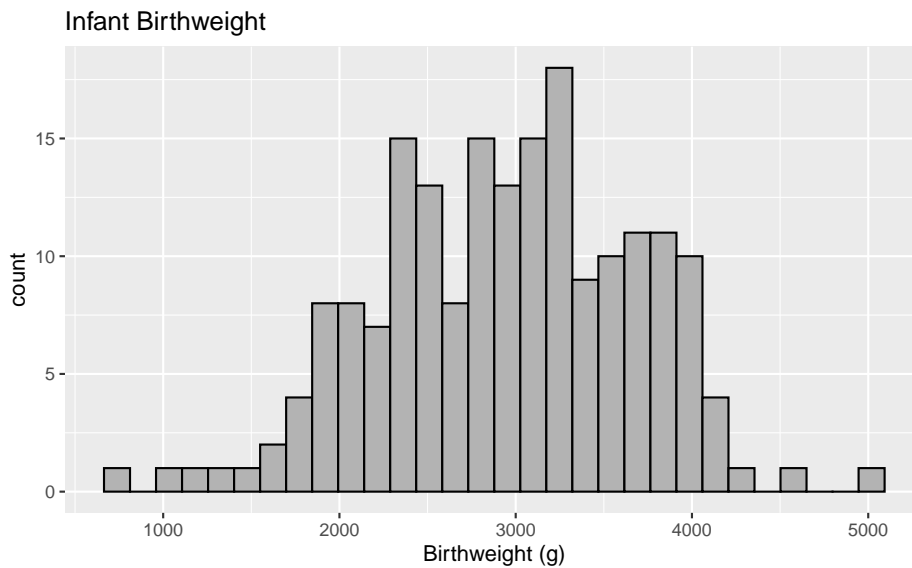


2.2.2 Histograms

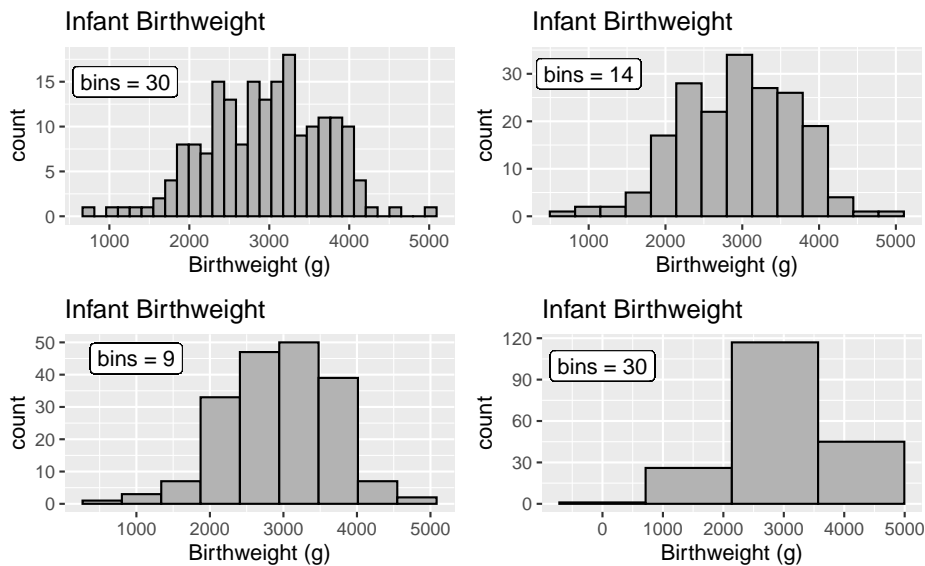
When we have a moderate size of data, graphing dots exactly on an axis doesn't work and results in overplotting and it is difficult to see where the data cluster. Instead we'll stack the dots in columns along the axis and call this a dotplot.



Each dot represents an observation, but the x-values have been rounded into group values. So we have lost some precision. Another common version of this is a histogram, where the y-axis represents how many observations fall into each bin.

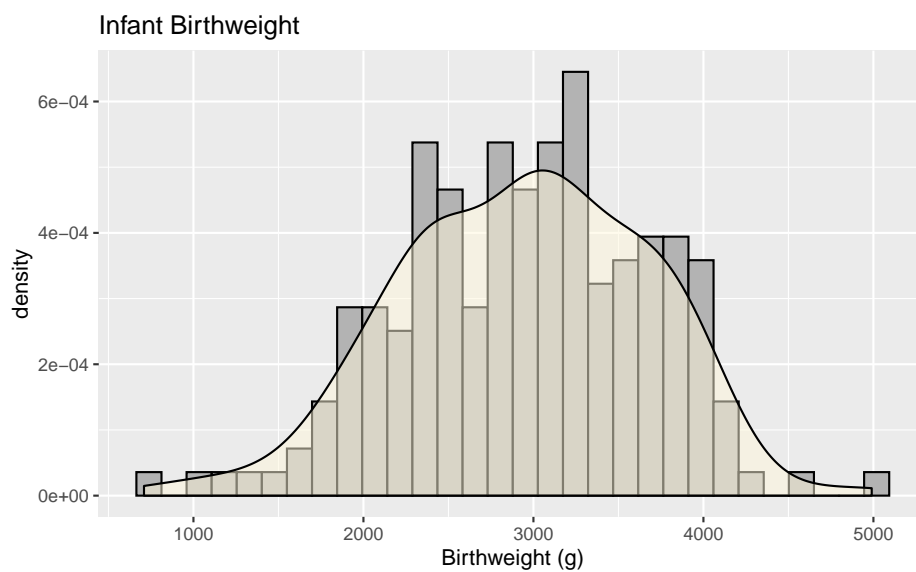


The choice of how many bins to include can make a dramatic difference in a graph. In particular, I don't believe that there is any biological reason to think the dip near 2700 grams is real. I believe that is actually just an artifact of the data I have. Instead we should consider changing the number of bins.



2.2.3 Density plots

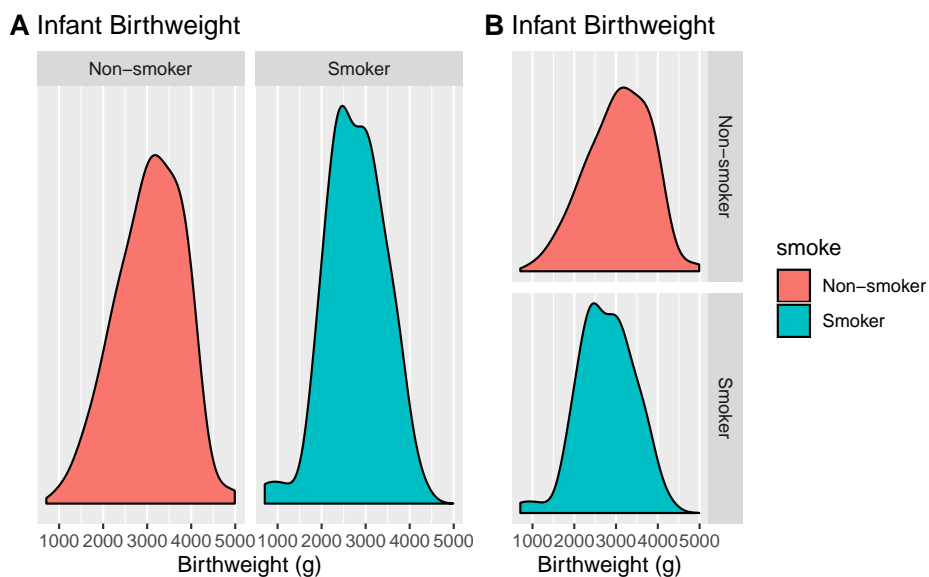
Histograms suffer from being too angular or pointy. Another solution is to call a kernel density smoother that mathematically smooths over the heights of the his-



togram bars.

2.2.4 Faceting

One of my favorite ways to display multiple distributions is to group each distribution into it's own plot in a process often referred to as faceting.

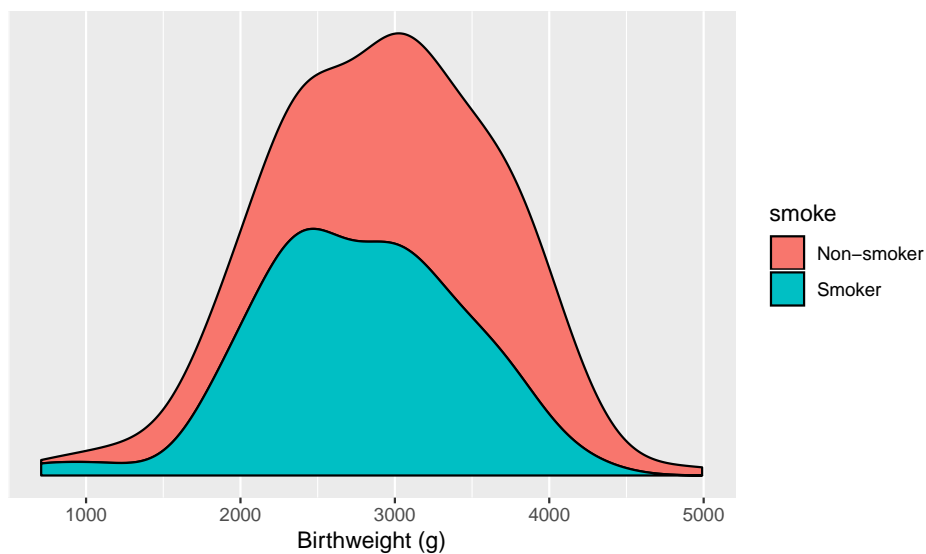


By choosing to put the two graphs on top of the other, it becomes clear that the smoker's tend to give birth to smaller infants. This fact isn't clear in the side-by-side graphs.

2.2.5 Stacking

Stacking the distribution involves laying each distribution on top of each other, so that the zero of the top curve follows the curve on the bottom. You can visualize the B chart having the Non-smoker density graph just melt onto the smoker density.

Infant Birthweight

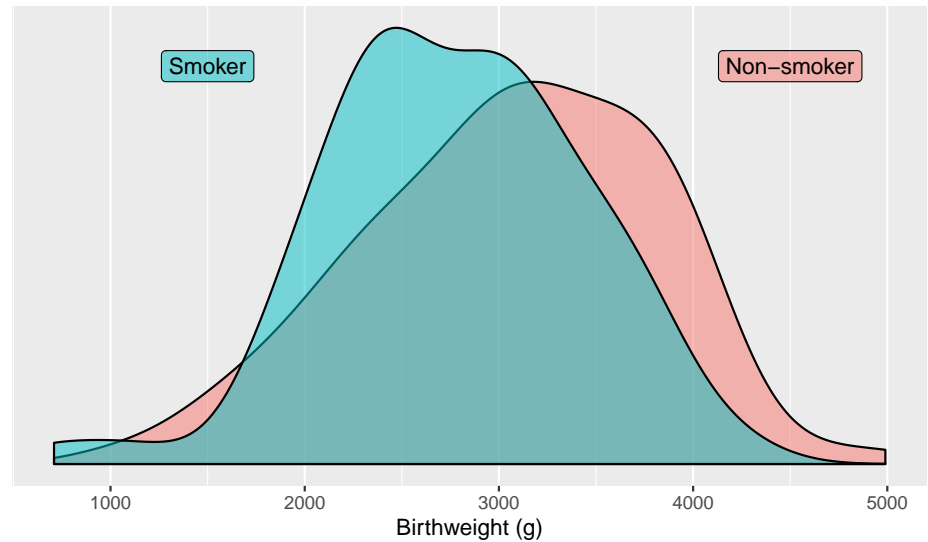


I really don't like this graph because it is very hard to see where the peak of the non-smoker curve is. This stacking trick works well enough when we have proportions but isn't good here.

2.2.6 Overlapping curves

Another option is to graph the densities, but allow them to overlap each other

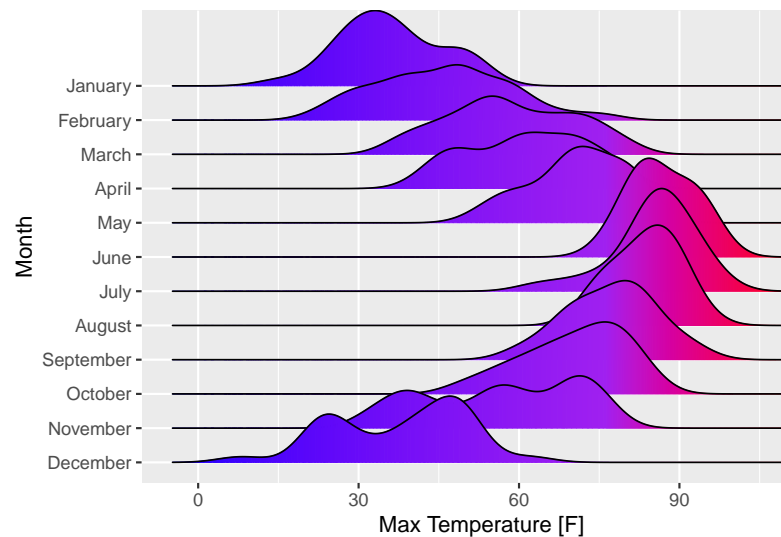
Infant Birthweight



and be a bit see-thru.

For seeing shifts in the center of the distribution, overlapping curves is quite powerful.

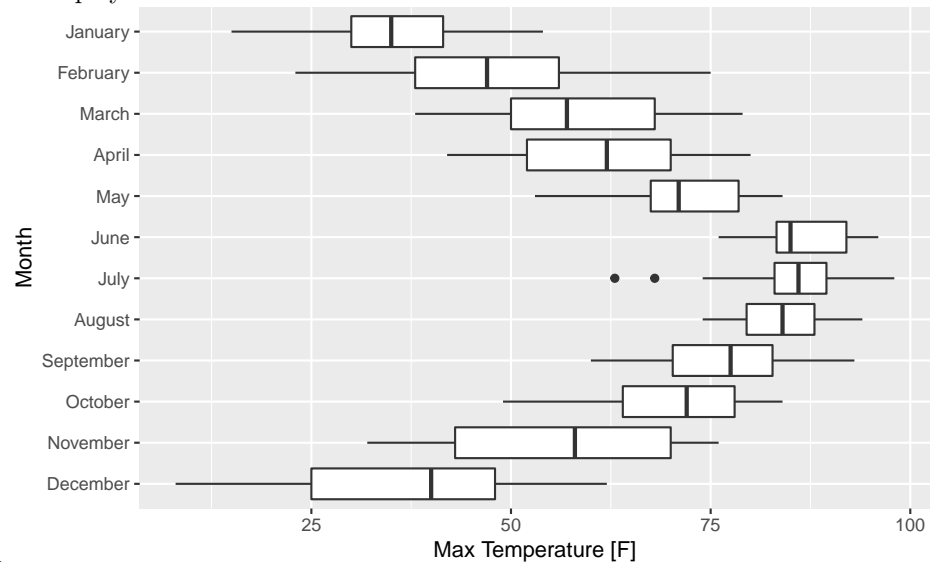
For another nice example, we can look at the density of the daily maximum temperatures in Lincoln NE, 2016



perature in Lincoln Nebraska in 2016.

2.2.7 Boxplots

Boxplots are a traditional way to display a distribution and the box contains the



middle 50% of the data points.

Notice that in density plots, there were two peaks in December with the lower peak corresponding to a cold snap. However that detail is lost in the boxplots.

2.3 Proportions

A good pie chart from reddit/r/dataisbeautiful member u/foiltape.

Blood Type Distribution in the United States



Figure 2.1: A Good Pie Chart