STA 141 - Exploratory Data Analysis and Visualization

Derek L. Sonderegger

September 16, 2019

Contents

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.

The source code for my notes, homework assignments, and other information is available on my on GitHub. In particular, there is a directory `data-raw` that contains all of the datasets that we'll use in these notes and homeworks.

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 as well as GitHub repositories for the book source and data used in the book.
  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.
  - Cognative 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.
  - Britian'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

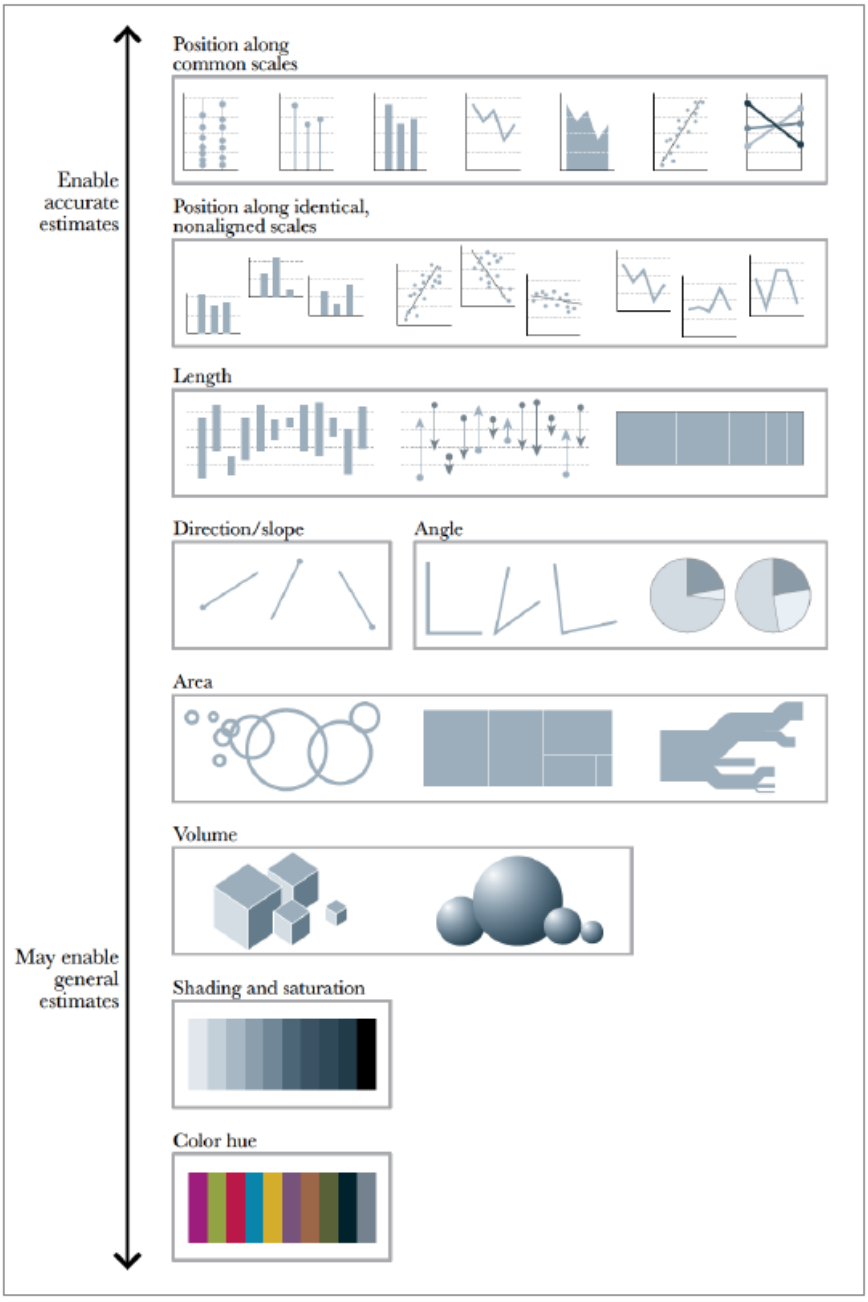
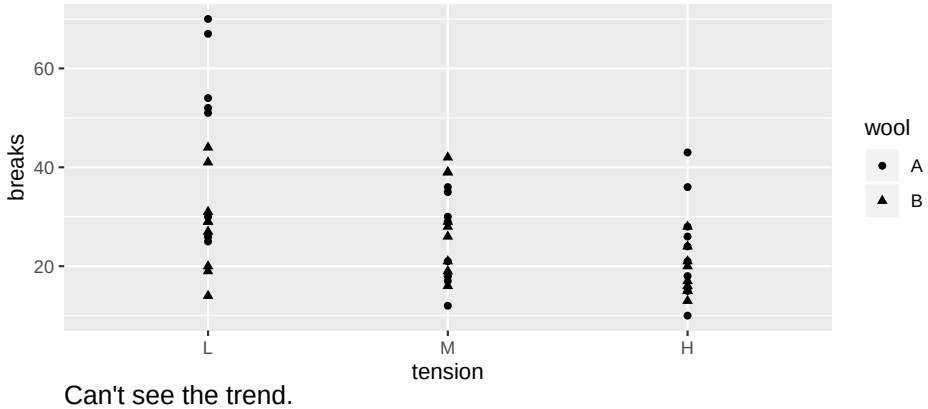


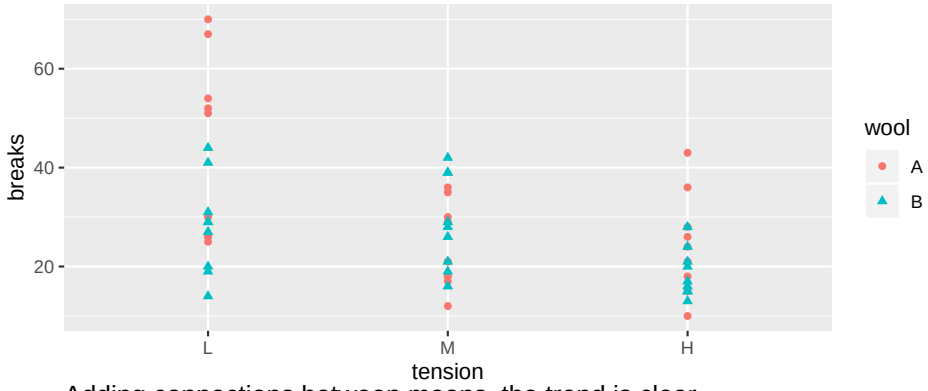
Figure 5.5 Scale of elementary perceptual tasks, inspired by William Cleveland and Robert McGill.

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

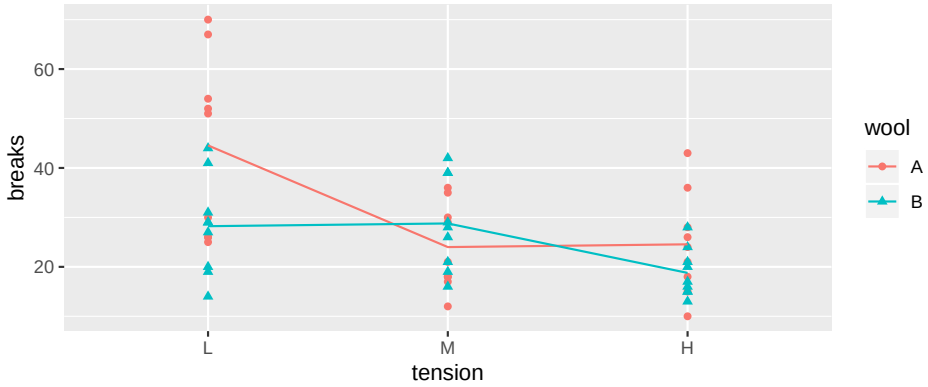
Can't easily distiguish wool types.

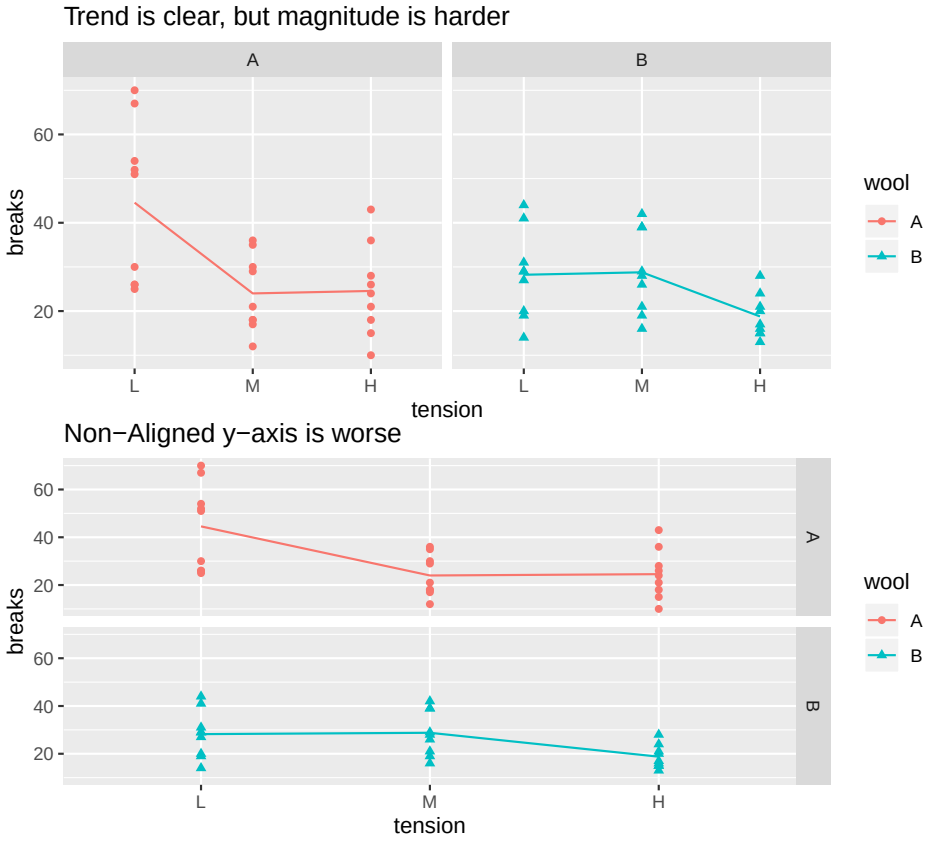


Can't see the trend.



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 approriate for a quantitative variable scale.

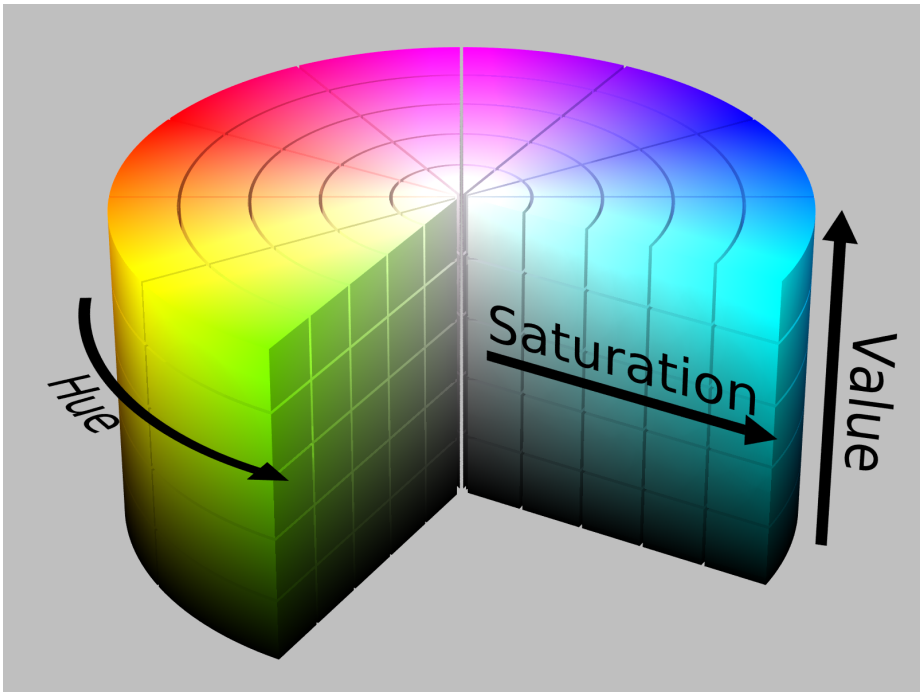


Figure 1.2: HSV Cylinder from Wikipedia

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.

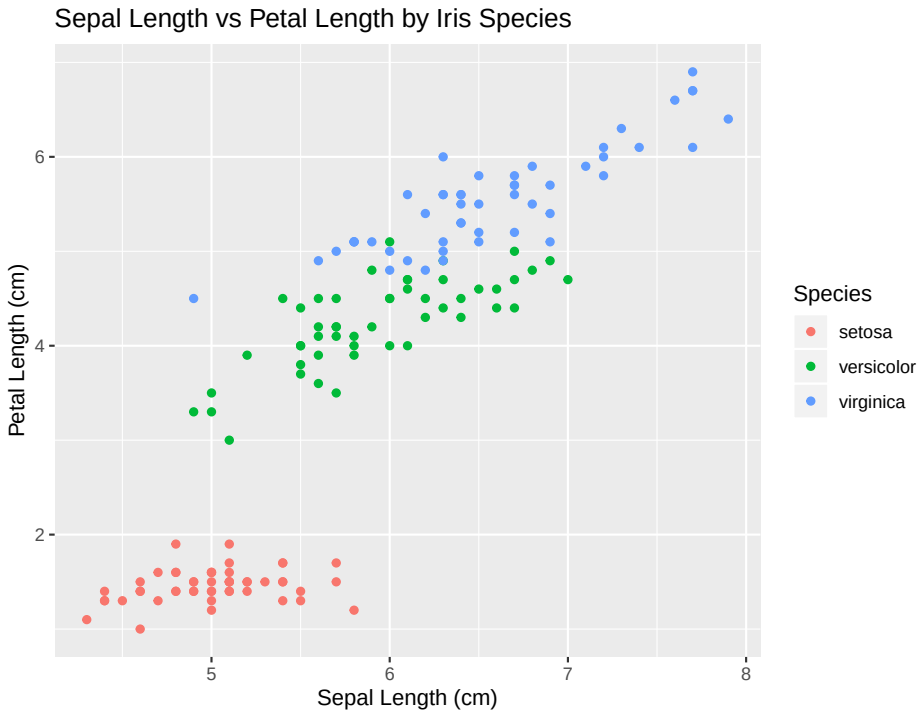
StudentID	Assesment	Score
1	Exam 1	91

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

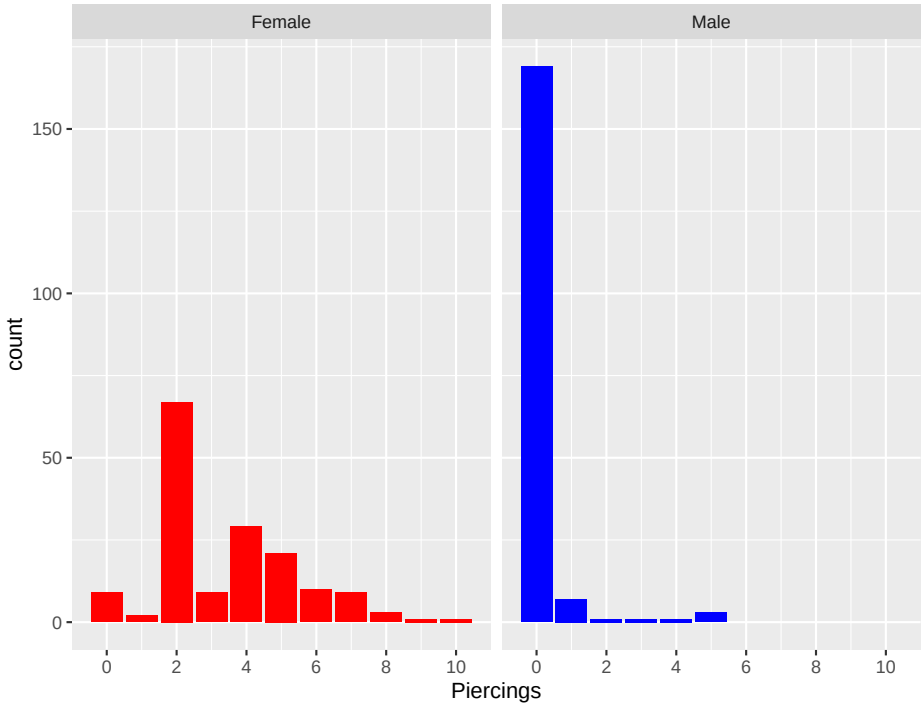
First we'll make a scatterplot with some colors.



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

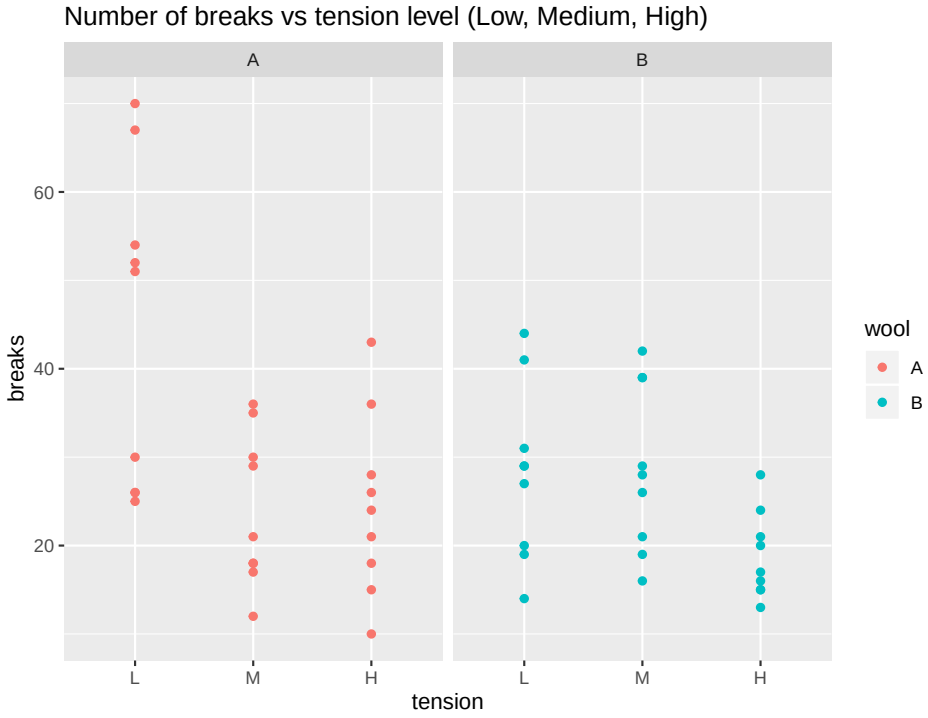
1.3.2.2 Task 2: Modifying how a variable is displayed

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


Number of breaks vs tension level (Low, Medium, High)
A
B
breaks
60
40
20
L
M
H
tension
wool
A
B

Chapter 2

Week 2

Assignments

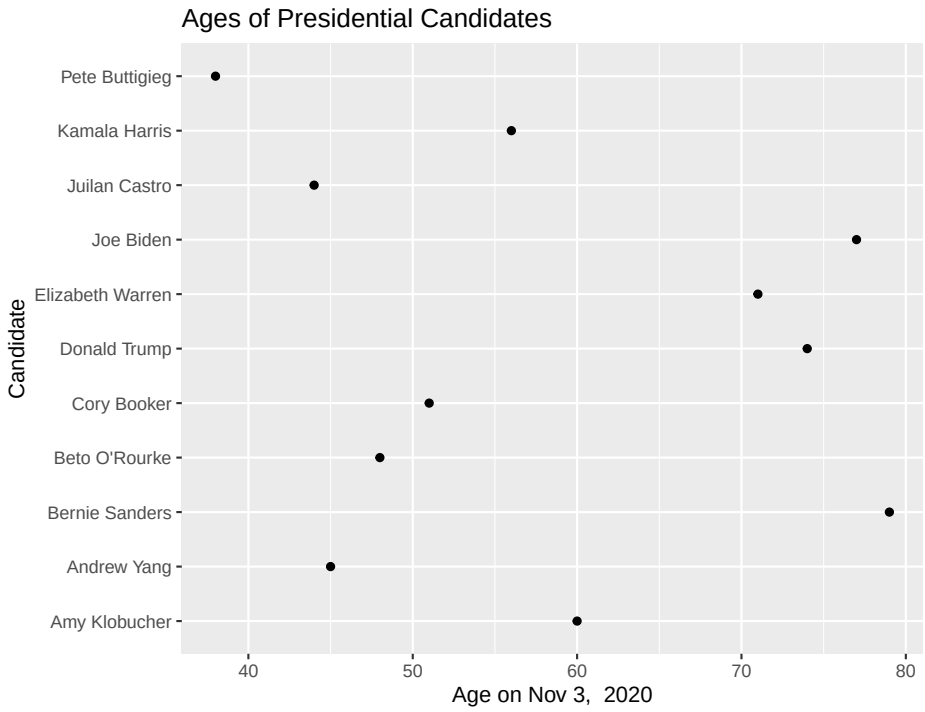
1. Read Chapter 6 and 7 from Claus Wilke's Fundamentals of Data Visualiztion book.
2. Watch Hans Roslings TED talk. Comment on two aspects of the videa that stand out to you regarding how he presented his data and how he talked through the visualization with the audience.
3. Graph the NOAA CO_2 data over time. The Nationa Oceanic and Atmospheric Adminstration has monthly CO_2 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. CO_2 is measured in parts per million.

    a) Plot CO_2 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.

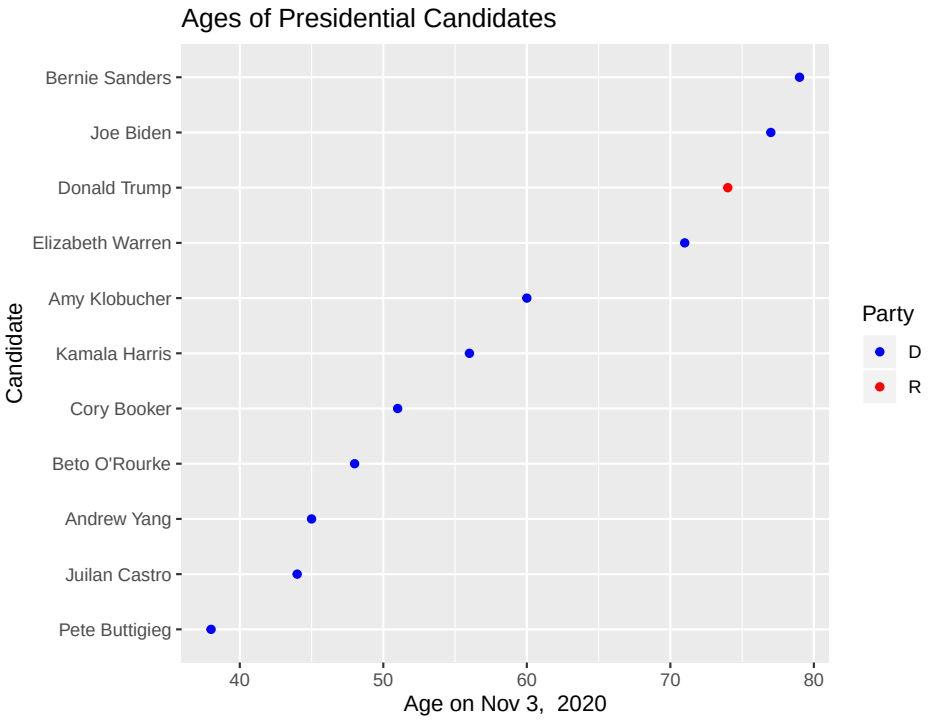
2.1 Amounts

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

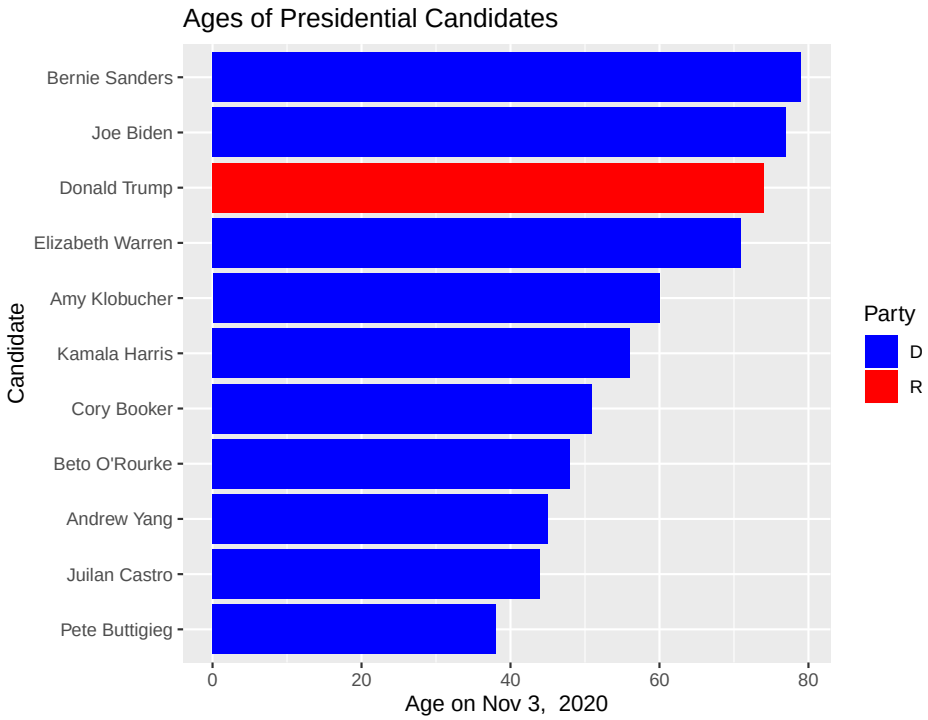
most prominent candidates and pulled their birthdays from Wikipedia and then calculated their age on election day.



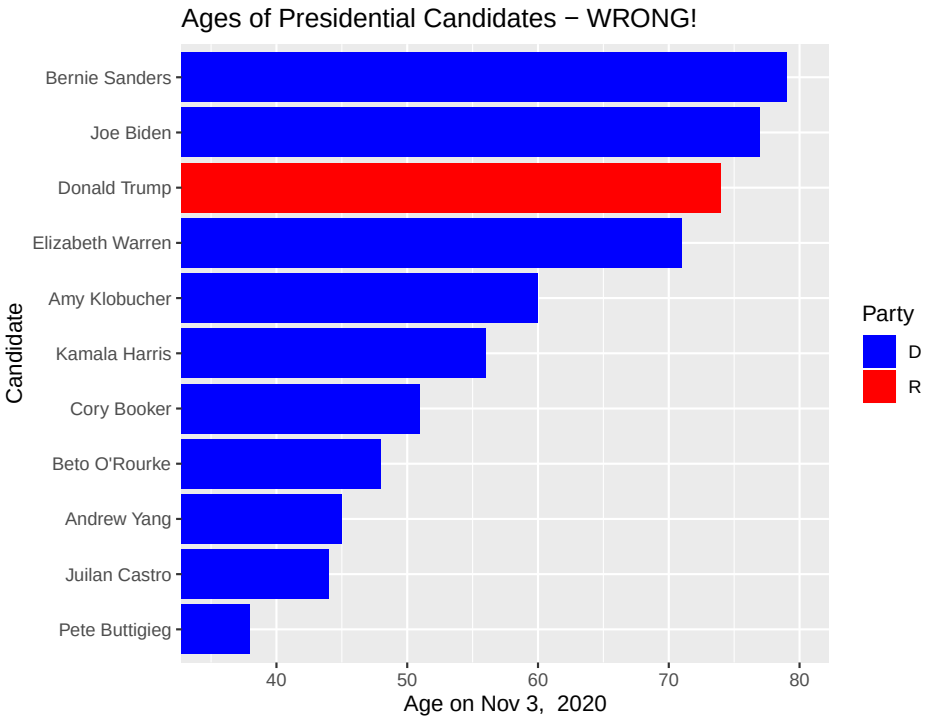
The order of the candidates is useless. Here we have ordered them alphabetically when we should try to think about an ordering that improves clarity. Lets switch to sorting the candidates by age.



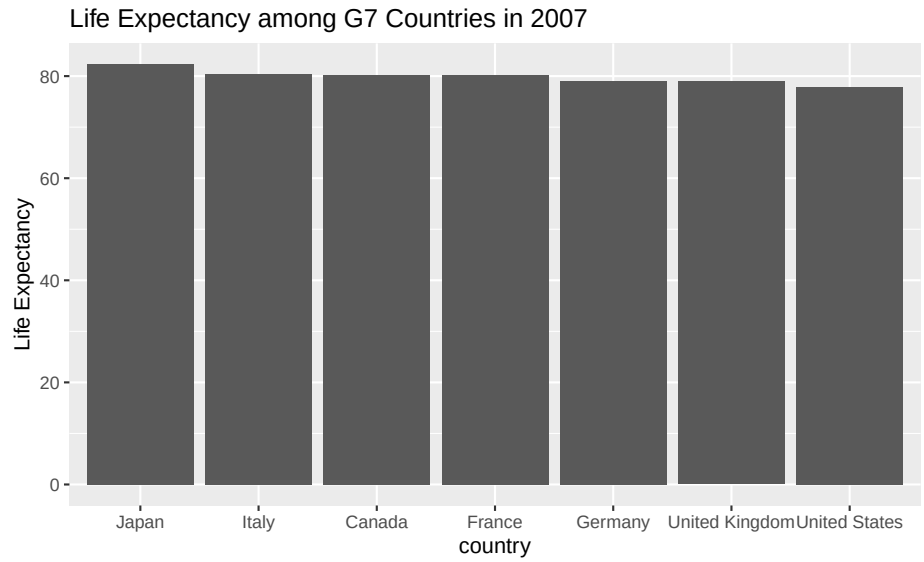
This isn't too bad, but it fails to visually impress the differences. A bar chart should visually impress the ages based on the length of the bar so that we con't have to keep looking at the Age axis.



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.

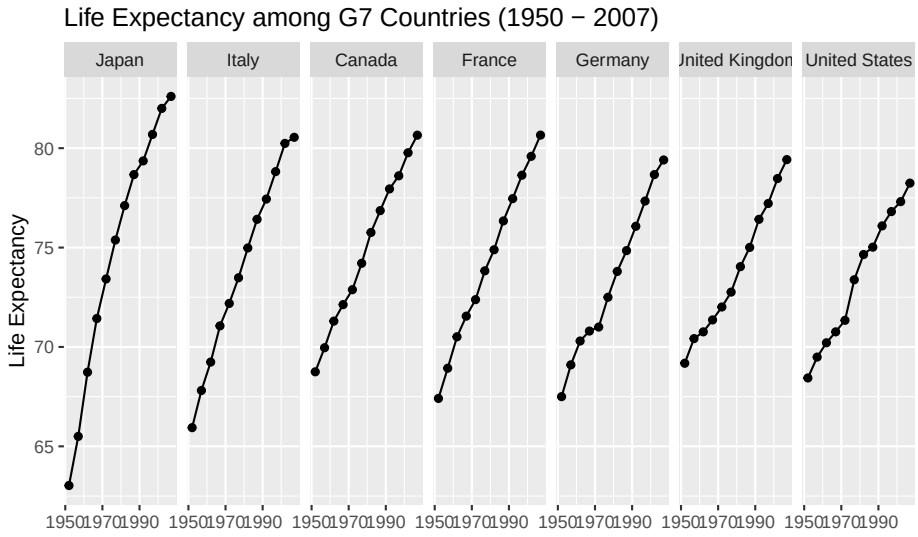


Often we need to graph some value and want to know how it varies among *two* different categories. In these cases, we have to employ some sort of grouping strategy.



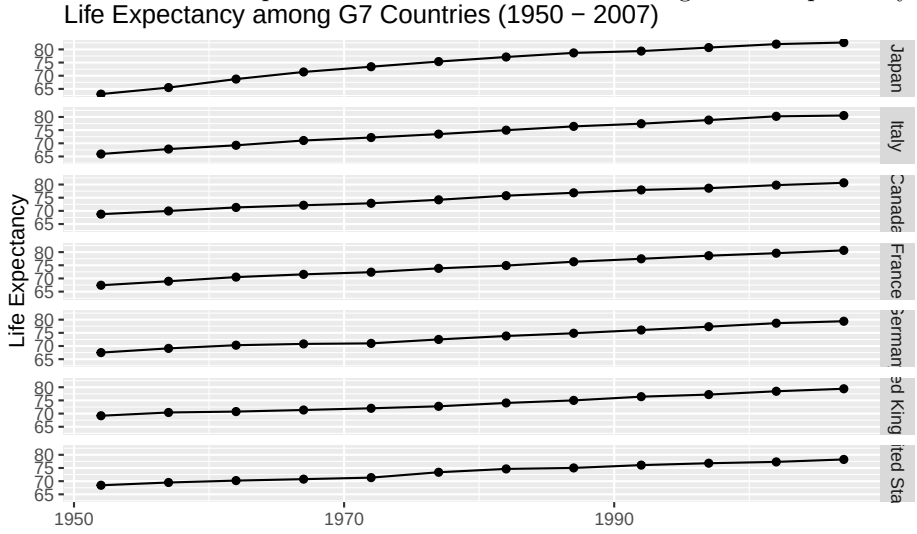
The bar chart here is obscuring the differences in life expectancies because the

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.

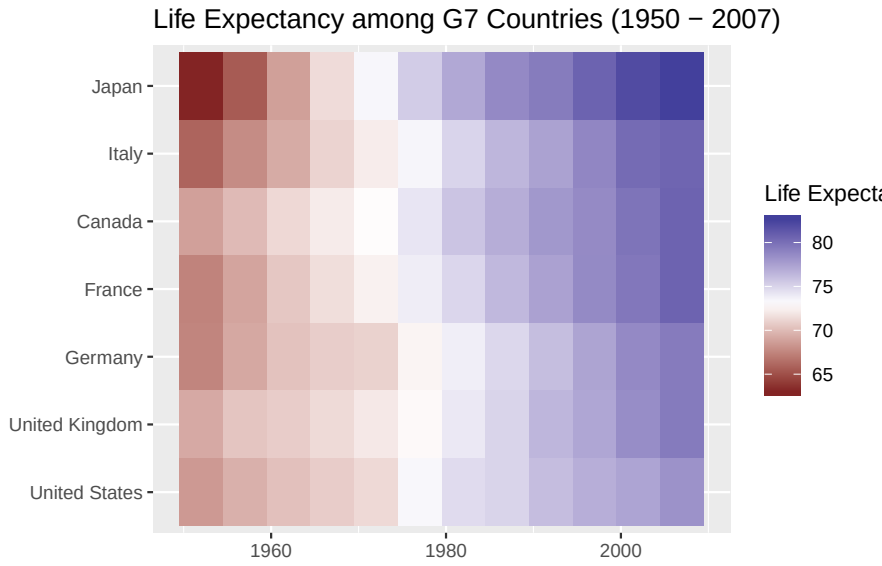


In the above graph, I am grouping countries both by enclosure and with a physical path connection. The reader tends to see the line as a whole object and compare the line max/min and slope among the seven countries.

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,



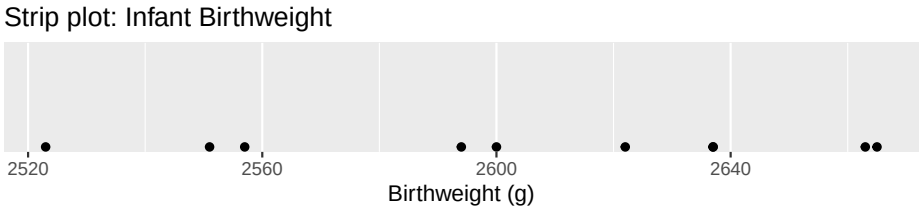
but we lose precision in the actual values.

2.2 Distributions

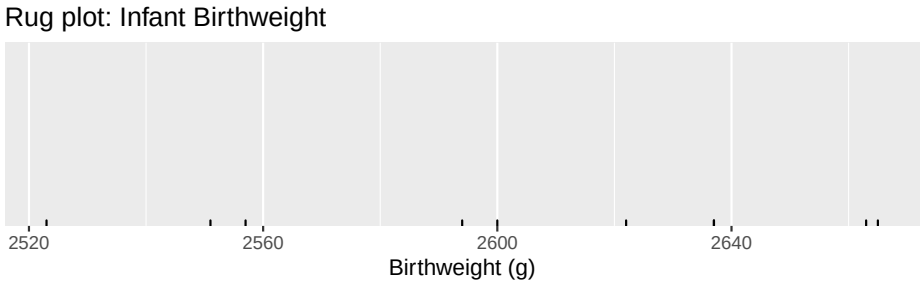
Given a single variable, I often want to know what values are common and what values are rare. To visualize this, we will primarily compare marks along a common axis (the most accurate EPT!)

2.2.1 Small samples

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



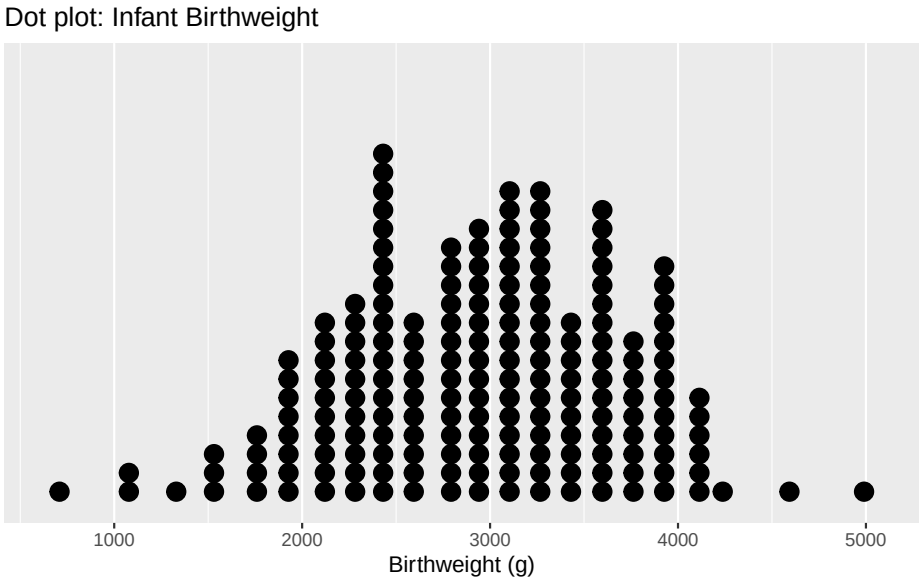
Another trick that works with more data, is to not use dots but rather lines. This is called a rugplot. This is often used in conjunction with another graph such as a



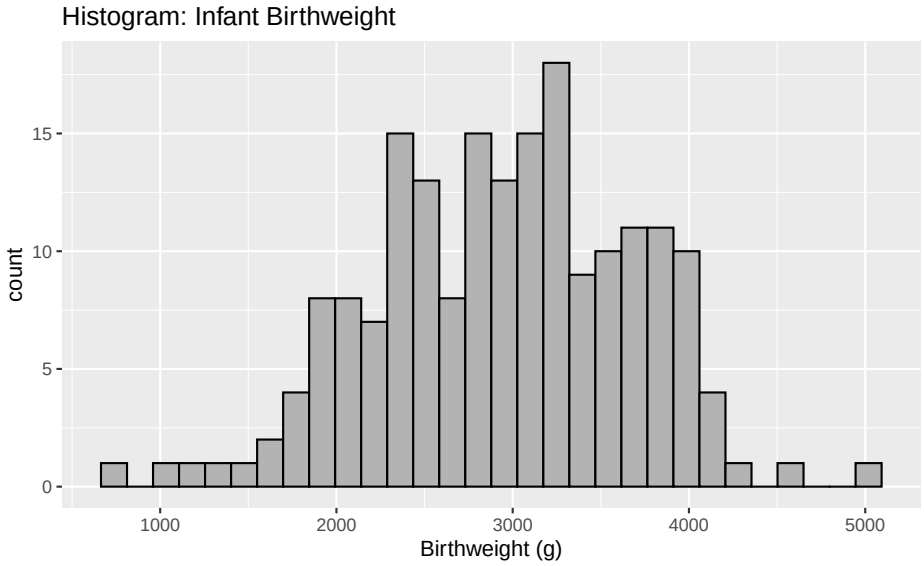
scatterplot.

2.2.2 Histograms

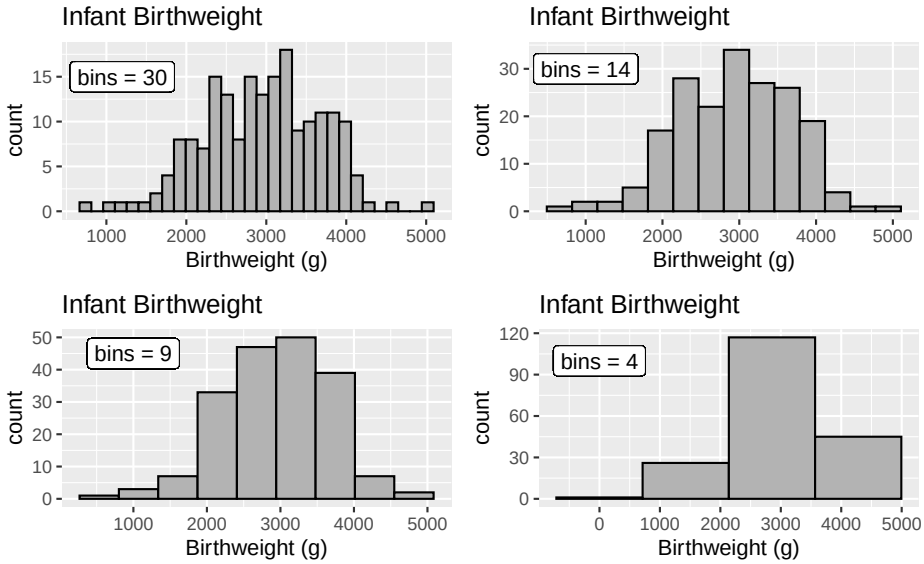
When we have a moderate size of data, graphing dots exactly on an axis doesn't work and results in overplotting and it is difficult to see where the data cluster. Instead we'll stack the dots in columns along the axis and call this a dotplot.



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.

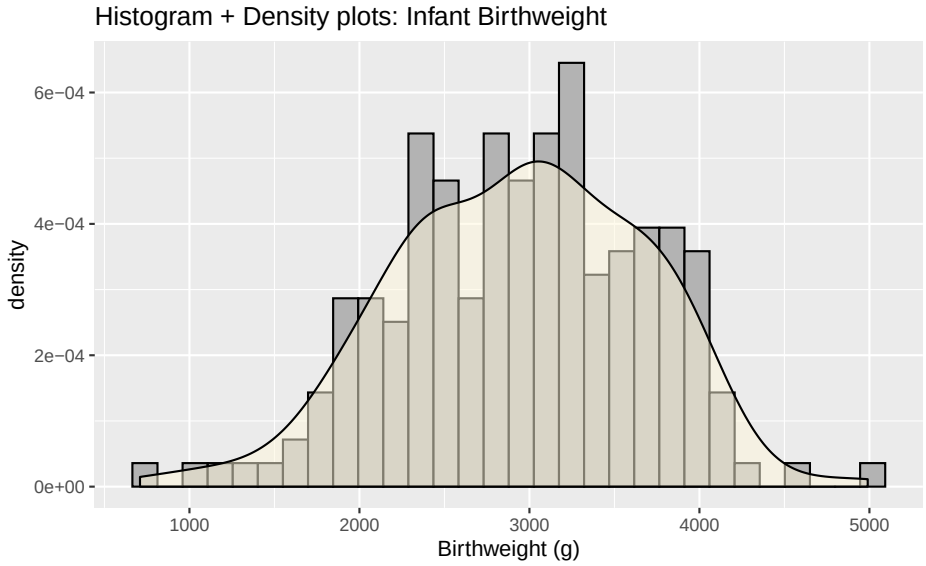


The choice of how many bins to include can make a dramatic difference in a graph. In particular, I don't believe that there is any biological reason to think the dip near 2700 grams is real. I believe that is is actually just an artifact of the data I have. Instead we should consider changing the number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



2.2.3 Density plots

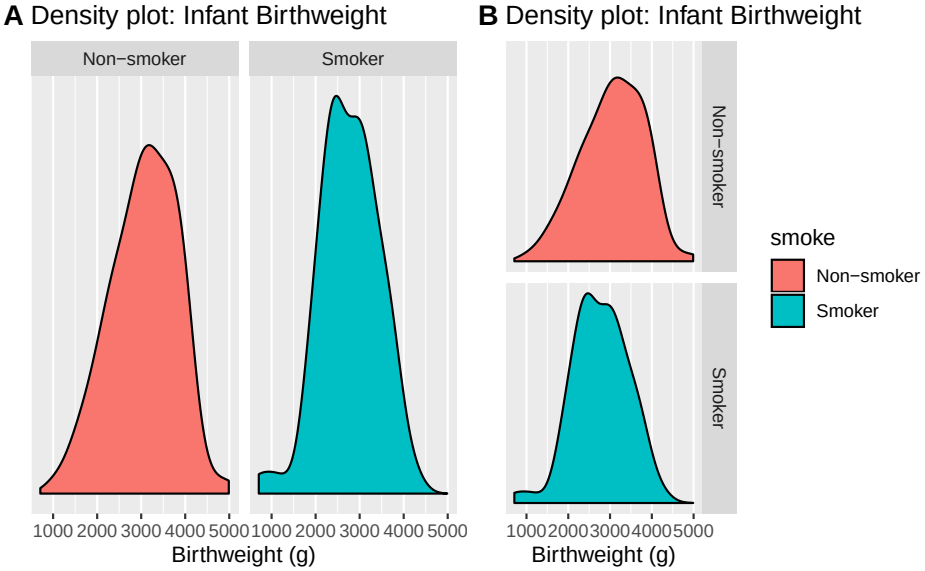
Histograms suffer from being to angular or pointy. Another solution is call a kernel density smoother that mathematically smooths over the heights of the his-



togram bars.

2.2.4 Faceting

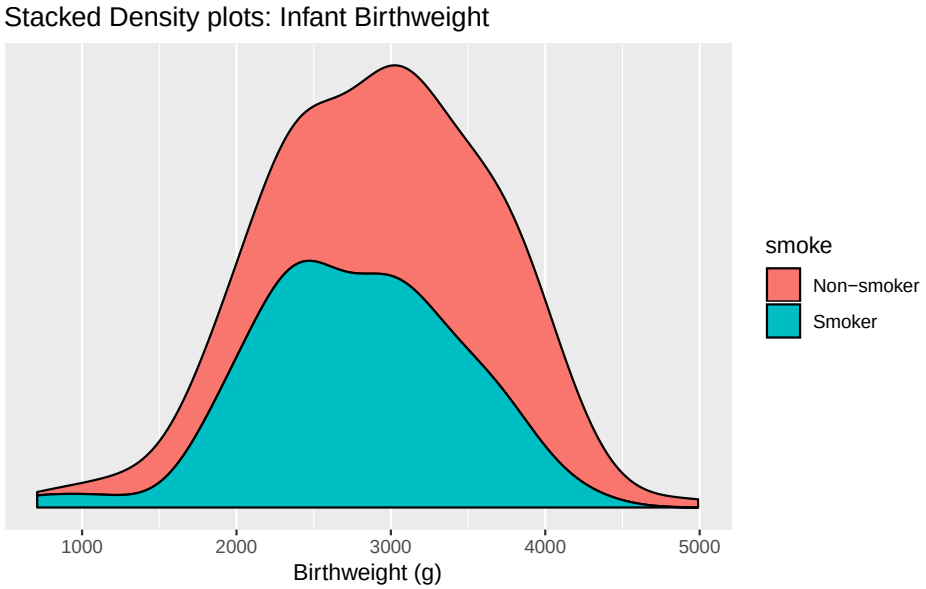
One of my favorite ways to display multiple distributions is to group each distribution into it's own plot in a process often referred to as faceting.



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

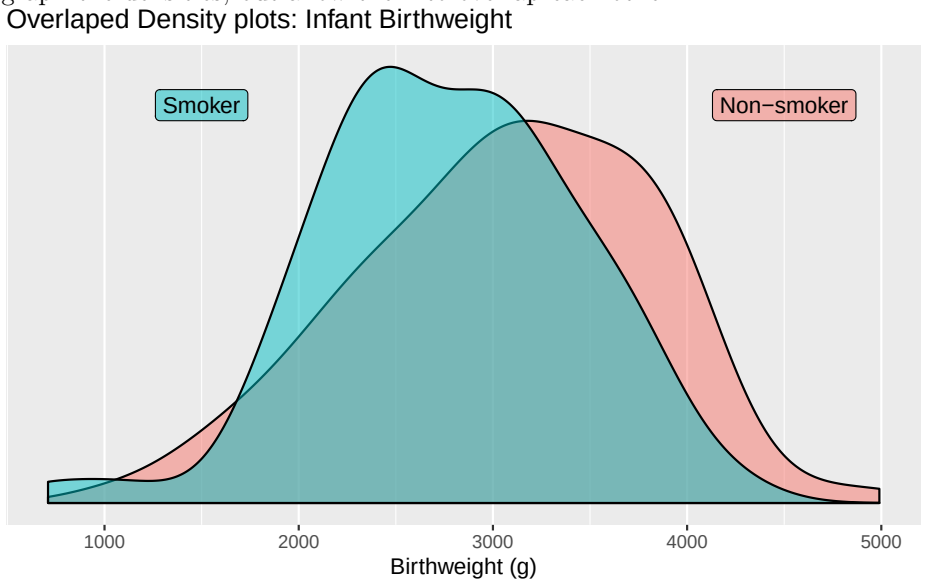
Stacking the distribution involes laying each distribution on top of each other, so that the zero of the top curve follows the curve on the bottom. You can visualize the B chart having the Non-smoker density graph just melt onto the smoker density.



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

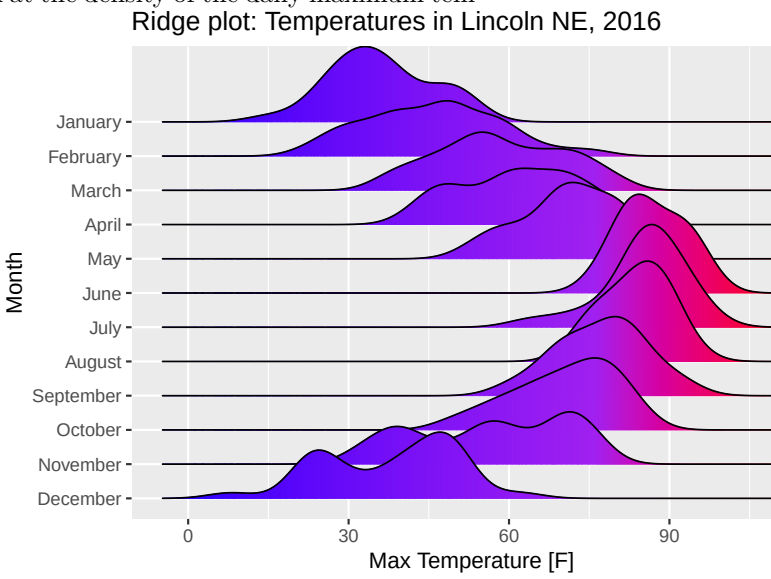
Another option is to graph the densities, but allow them to overlap each other



and be a bit see-thru.

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

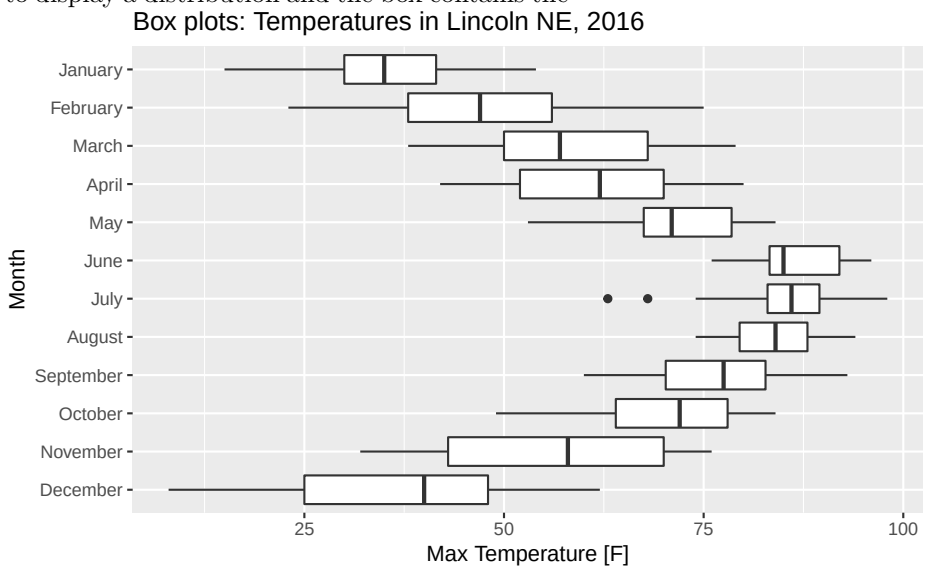
For another nice example, we can look at the density of the daily maximum tem-



perature in Lincoln Nebraska in 2016.

2.2.7 Boxplots

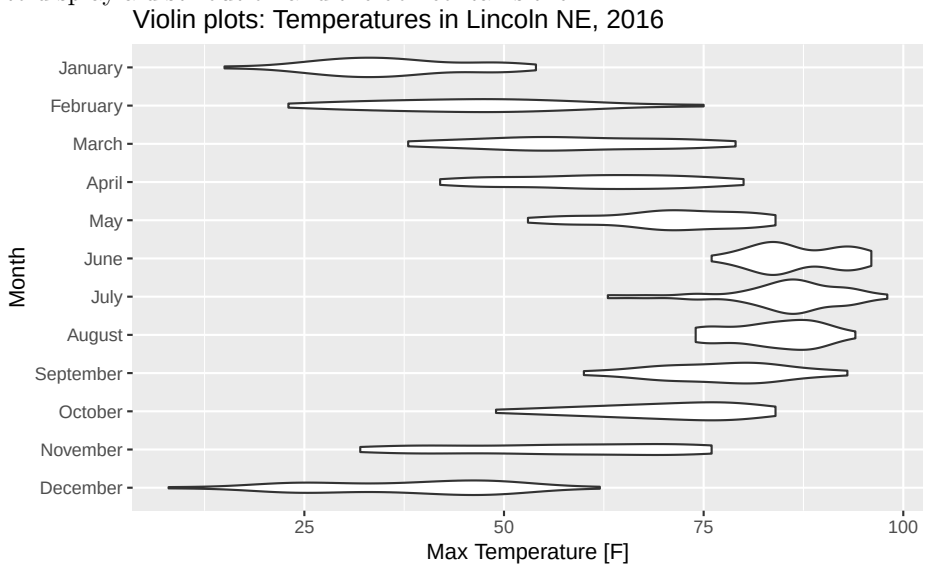
Boxplots are a traditional way to display a distribution and the box contains the



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.

2.2.8 Violin Plots

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



middle 50% of the data points.

Now we can see the two peaks in December, but the three peaks in November have been flattened out because the amount of space necessary to show it would require that the densities overlap.

2.3 Proportions

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

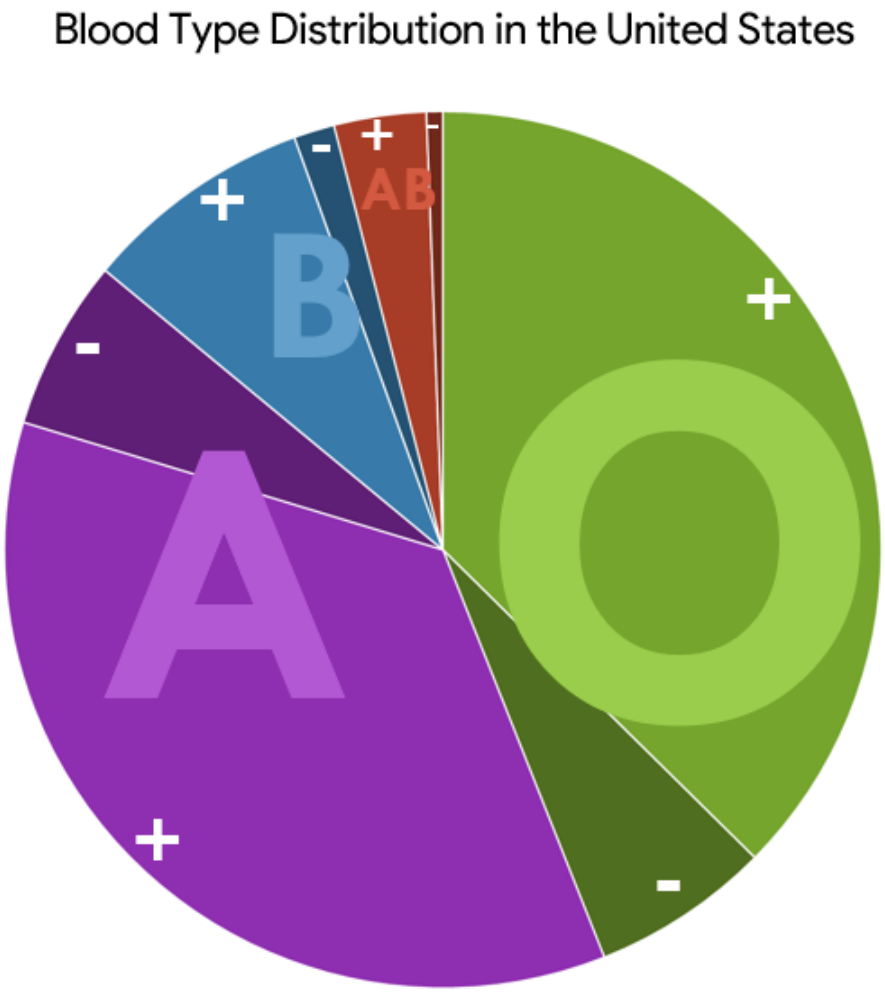


Figure 2.1: A Good Pie Chart

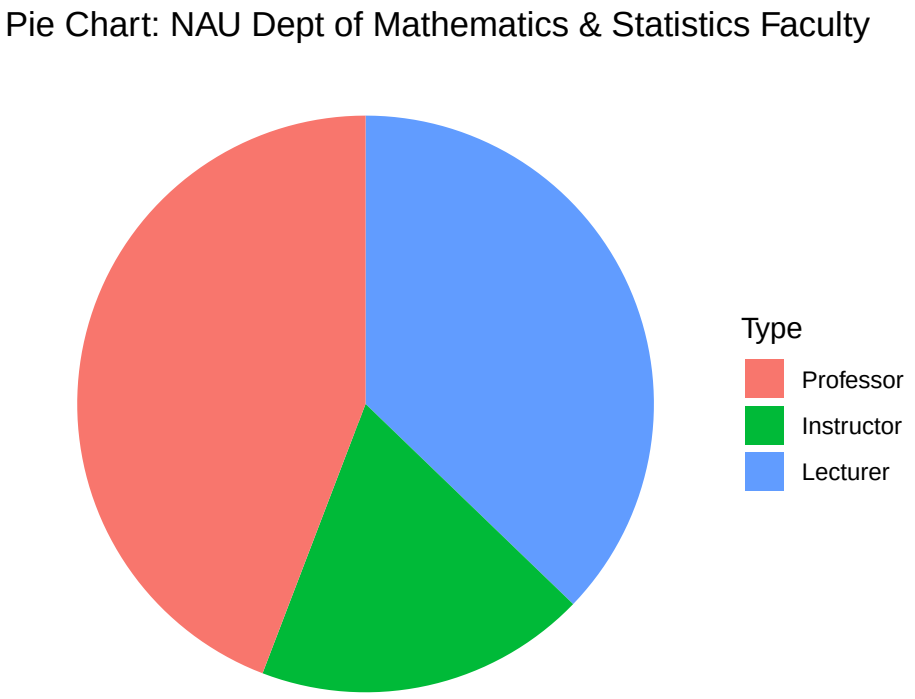
Chapter 3

Week 3

3.1 Proportions

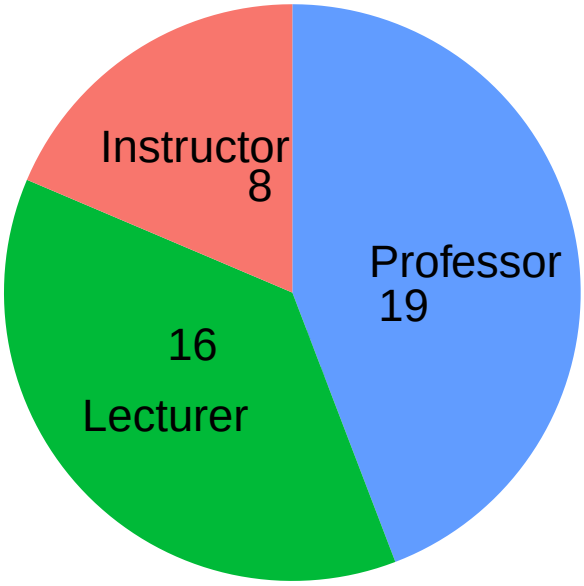
Conceptually graphing proportions is the same as graph raw values, but sum to 100%. This seemingly small difference means that our graphic can imply that our categories contain ALL possible categories.

3.1.1 Pie Charts

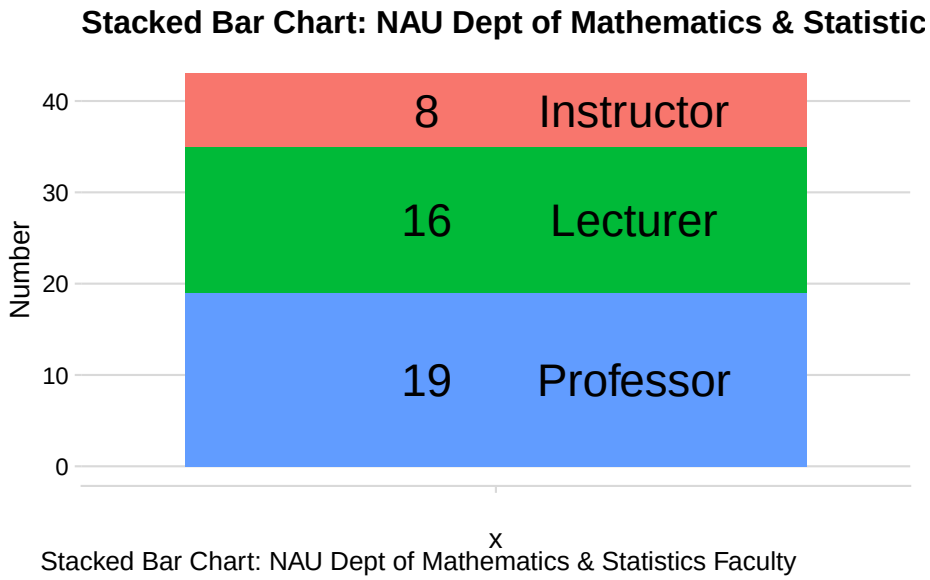


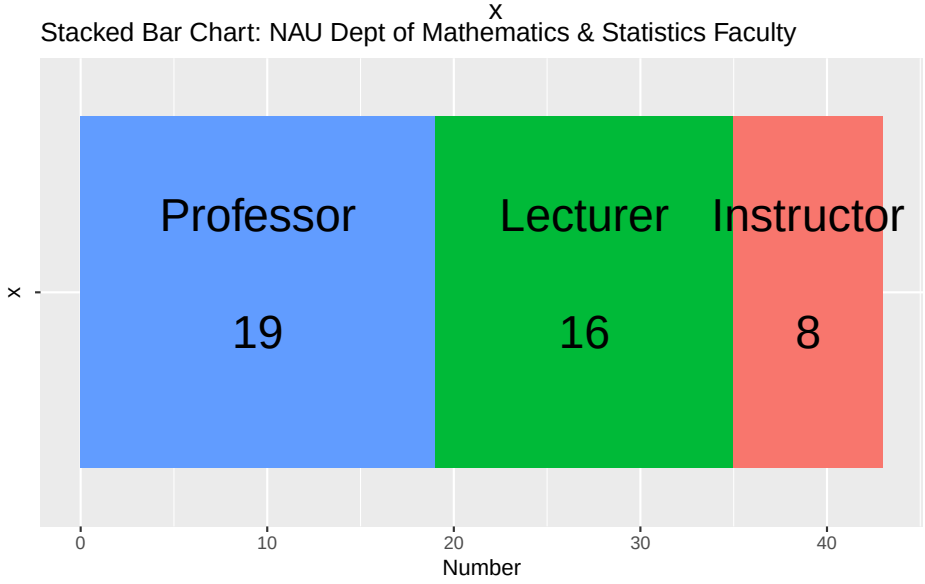
As typical, with a just a few categories, we should move the labels onto the graph and just annotate the graph. Also, we'll order the categories from the most temperary employees (instructors) to most permanent (professors)

Pie Chart: NAU Dept of Mathematics & Statistics Faculty

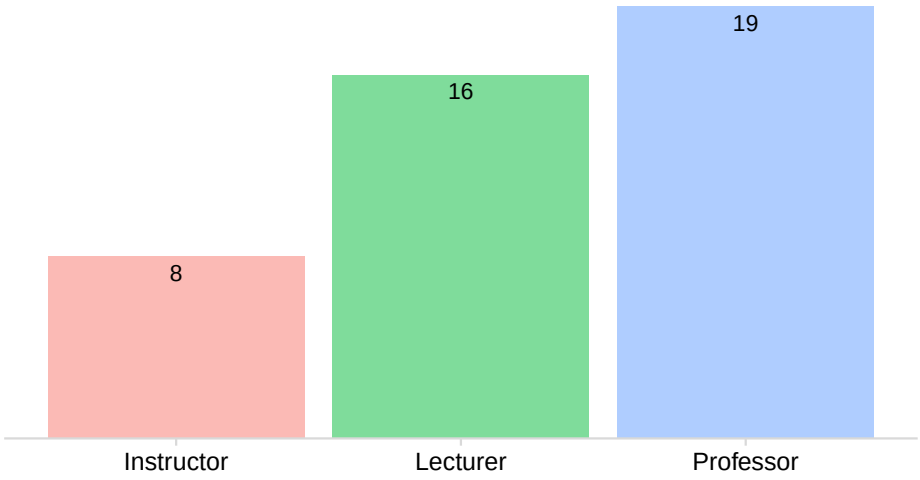


3.1.2 Stacked Bar





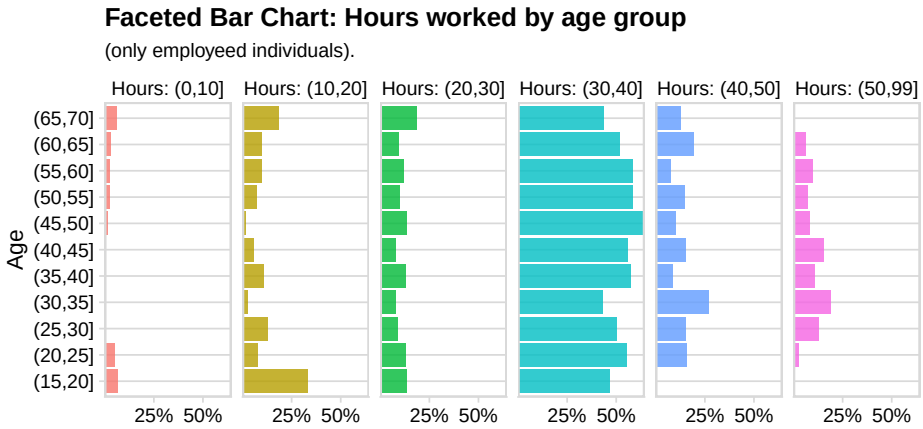
3.1.3 Side-by-side Barchart

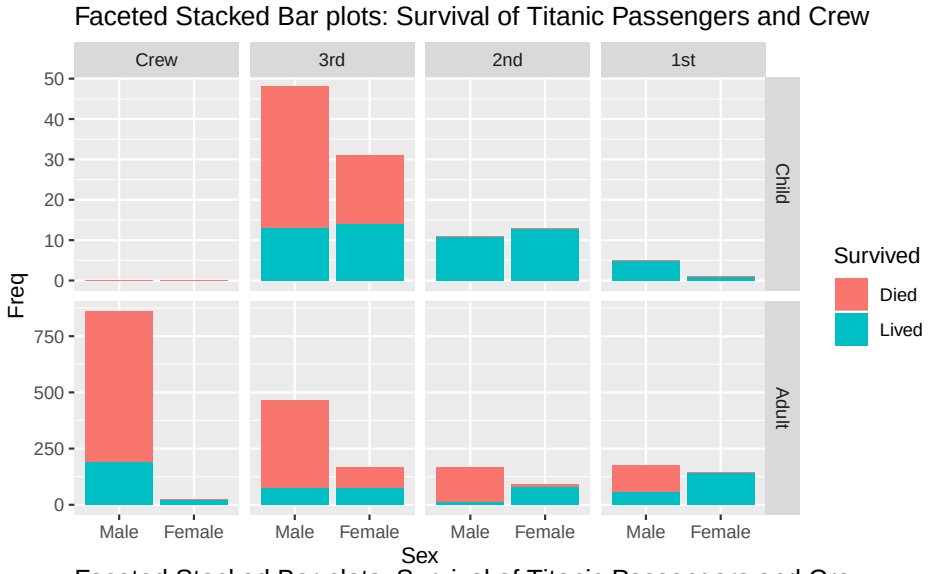


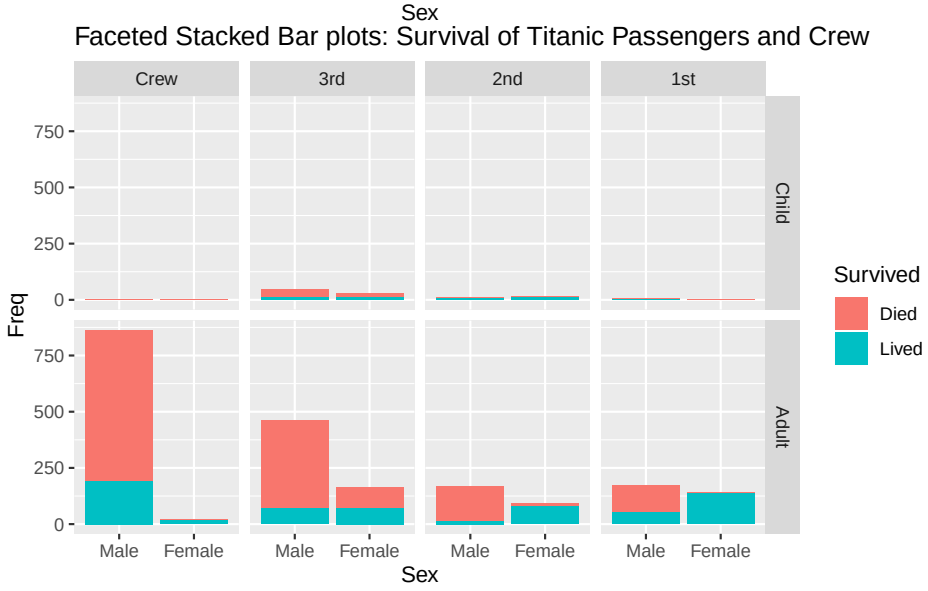
	Pie chart	Stacked bars	Side-by-side bars
Clear that data is proportions of a whole	**Yes**	**Yes**	*no*
Precise visual comparison of values	*no*	*no*	**Yes**
Visually appealing even in simple comparisons	**Yes**	*no*	**Yes**
Extendable to nested or multiple distributions or time series	*no*	**Yes**	*no*

3.2 Multiple Sets of Proportions

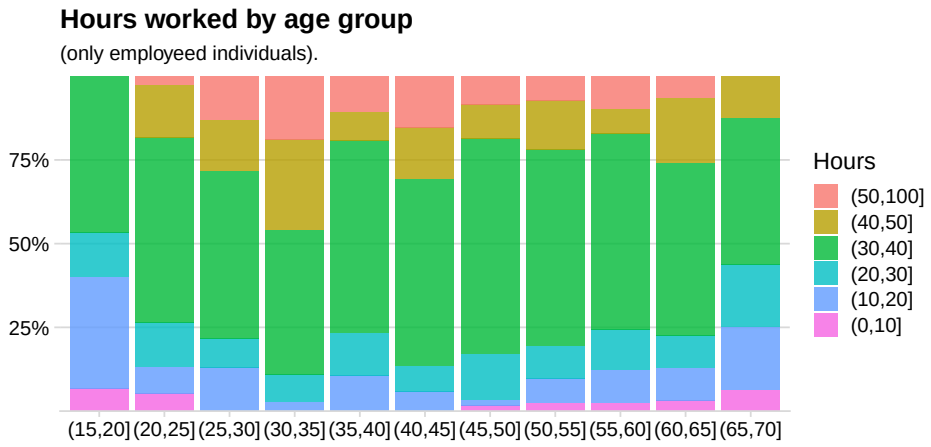
3.2.1 Faceted Bar charts




Faceted Stacked Bar plots: Survival of Titanic Passengers and Crew
Crew
3rd
2nd
1st
Child
Adult
Freq
50
40
30
20
10
0
750
500
250
0
Male
Female
Sex
Survived
Died
Lived

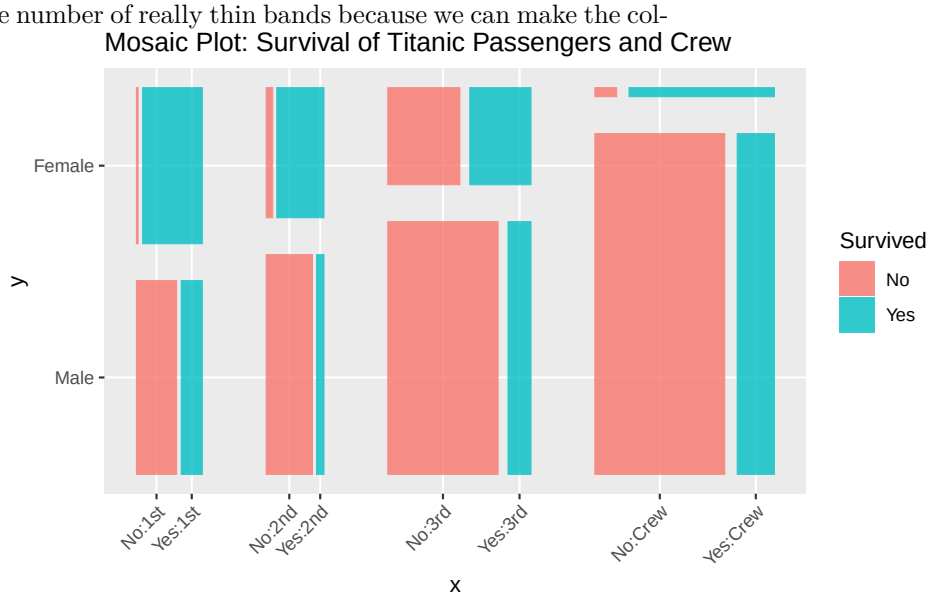

Faceted Stacked Bar plots: Survival of Titanic Passengers and Crew
Crew
3rd
2nd
1st
Child
Adult
Freq
750
500
250
0
Male
Female
Sex
Survived
Died
Lived

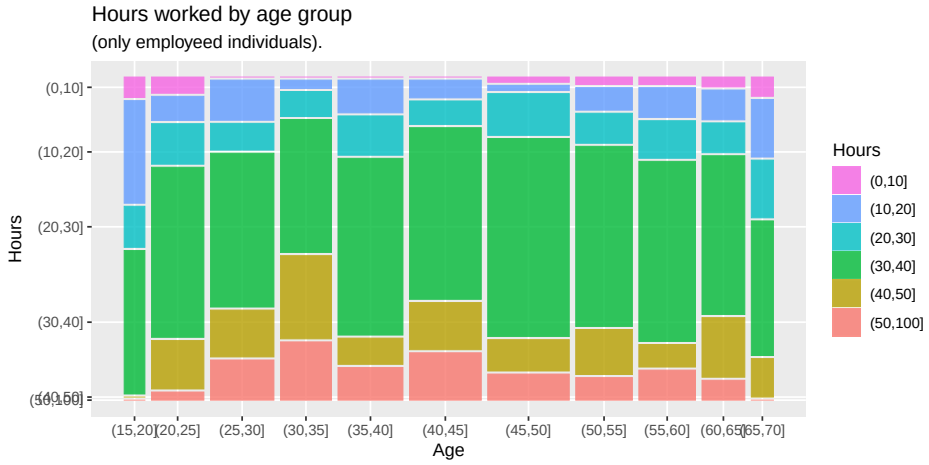
3.2.2 Side-by-Side Stacked Barcharts



3.2.3 Mosiac plots

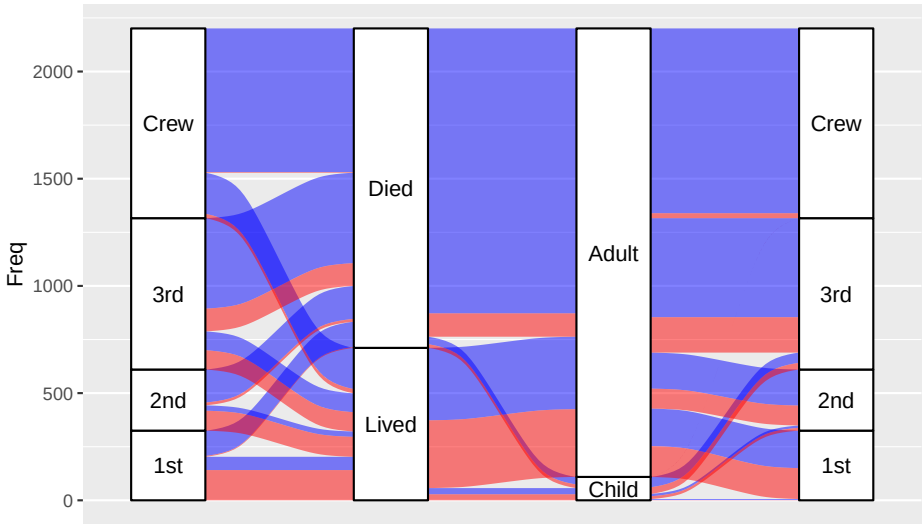
Sort like side-by-side stacked bar charts, but now we allow the column width to vary as well. The area is proportional the groups representation in the whole data. This reduces the number of really thin bands because we can make the column narrower as well.





3.2.4 Alluvial Plots

If we want to compare multiple categorical variables, another option is alluvial plots.



I find that alluvial plots work better for events that have a definate chronological order and there is less stream overlaps.. Here is an example from a Washington Post story about people graphing their online dating interactions.

3.2.5 Tree graphs

In mosaic plots, we had *crossed* variables where every category level of one factor could show up with all levels of another factor.

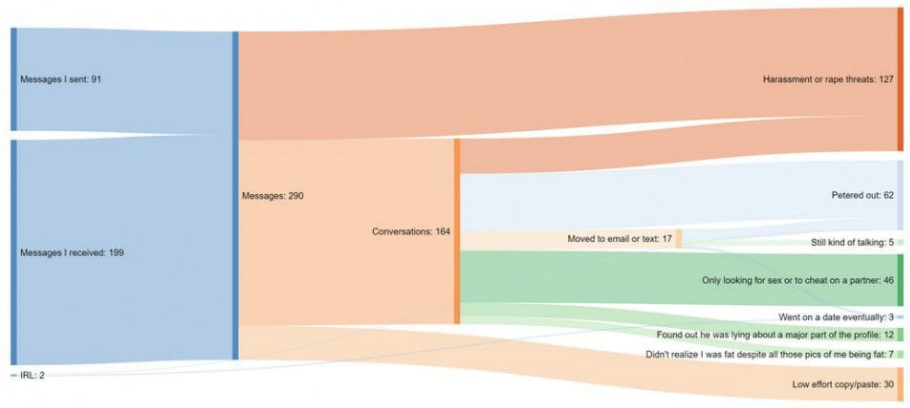


Figure 3.1: These are the results of 6.5 weeks of online dating by a 37 year old woman.

Table 3.2: Crossed Factors Suitable for a Mosaic Plot.

Factor.1	Factor.2	value
A	w	4
A	x	6
B	w	15
B	x	25

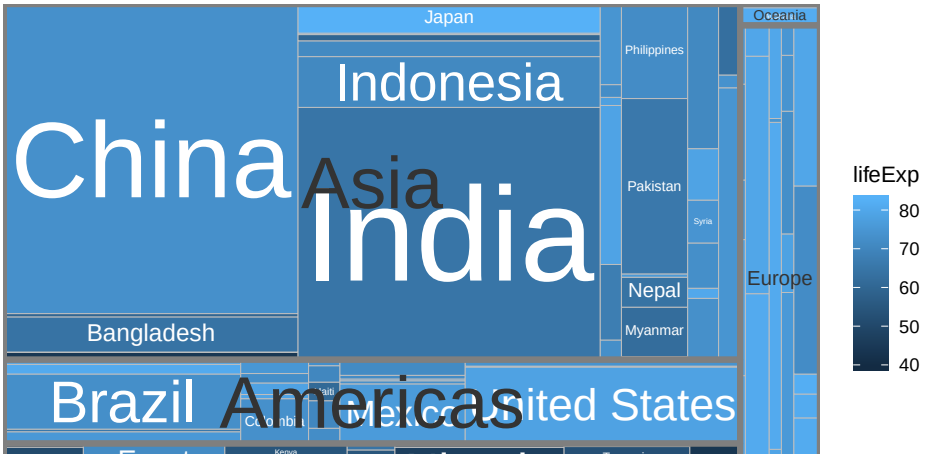
Another possibility is that the variables are nested such that a category level of the second factor only ever occurs within a single level of the first factor.

Table 3.3: Nested Factors Suitable for a Tree Map.

Factor.1	Factor.2	value
A	w	4
A	x	6
B	y	15
B	z	25

When we have a hierarchical structure of categories, then mosaic plots aren't quite right. Instead we'll hierarchically subdivide the graph area up.

Tree map: Population vs Life Expectancy in 2007



This differs from a moasic plot in that a country only occurs withing one continent whereas in a mosaic plot, a category level will occur in multiple "containers".

3.3 Exercises

1. Alluvial plots are a particular type of *Sankey* graphs which show flow rates and amounts and have been around for quite some time. In 1869, Charles Minard created a graphic that details the size of Napaleon's army as they marched on Russia and subsequently returned. You can find the original or the modern English translation on Wikipedia.

    a) How many men did the army start marching with?
    b) How many men arrived in Moscow?
    c) How many men died crossing the Berezina River on the return trip? (approximately from the map information provided)
    d) How cold was it when they cross the Berezina River on the retern trip?

2. Read Chapter 10 and 11 in Claus Wilke's Fundamentals of Data Visualiztion book. In chapter 10 he presents several different graphics that visualize the bridge construction era, bridge material, and which river they cross for bridges near Pittsburgh, Pennsylvania. Discuss three of them and explain which graph you prefer and why.

3. Download data about the Titanic disaster at the GitHub site for this class. Save the file as a Titanic.csv and open it in Tableau.

    a) In Tableau, a faceted stacked barchart just as we did in these notes.
    b) In a new worksheet, copy your faceted stacked barchart and then turn it into faceted pie charts.
    c) Comment on which you prefer and why.
    d) Finally create a mosaic plot of the Titatic dataset.

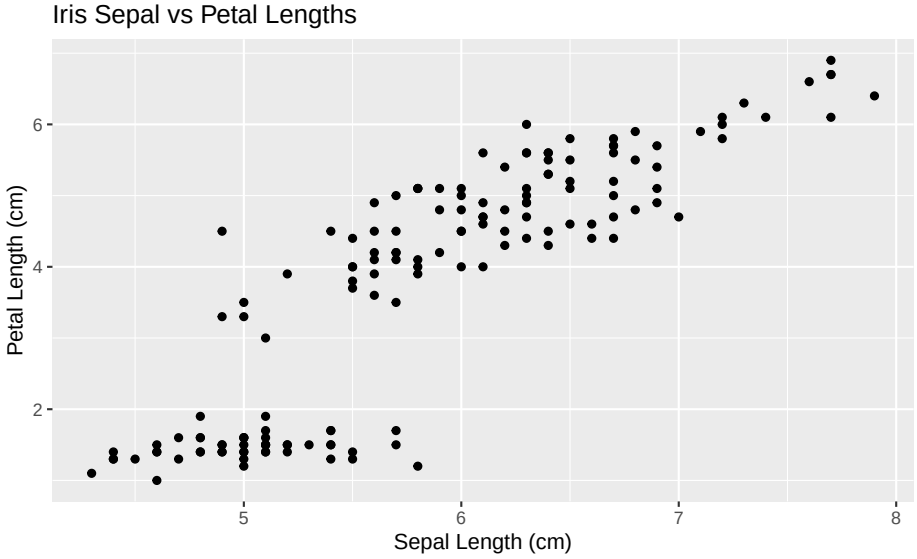
Chapter 4

Week 4

4.1 Visualizing 2 or more Continuous variables

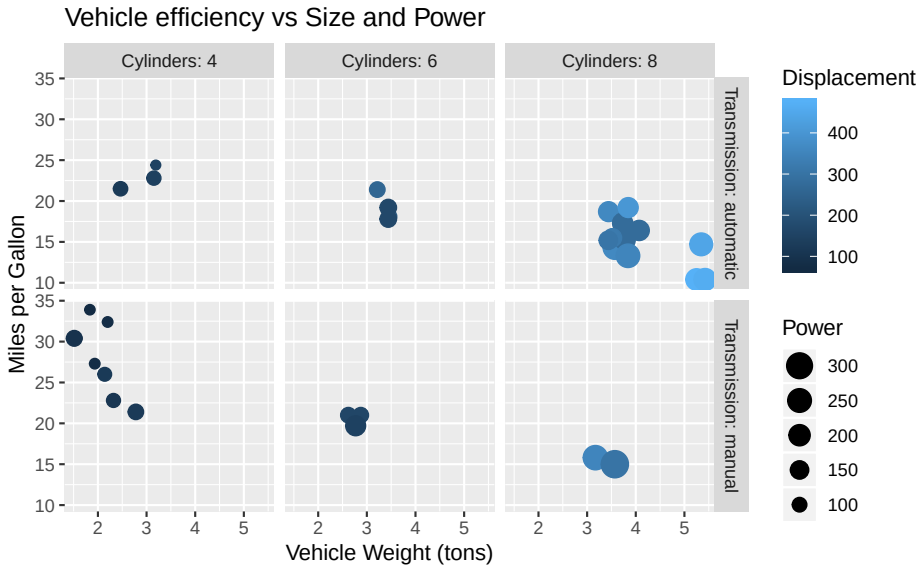
4.1.1 Scatter plots

Basic idea is to build off of a scatter plot. This visualizes the relationship between two continuous variables.



In a scatter plot we can see the relationship between two variables. We can see the relationship among more variables (either continuous or discrete) by adding Size, Color, and Shape.

We could also add other categorical variables by adding faceting. With this combination we can visualize the relationship between up to 6 different variables.



4.1.2 Pairs plots (All-vs-all scatterplots)

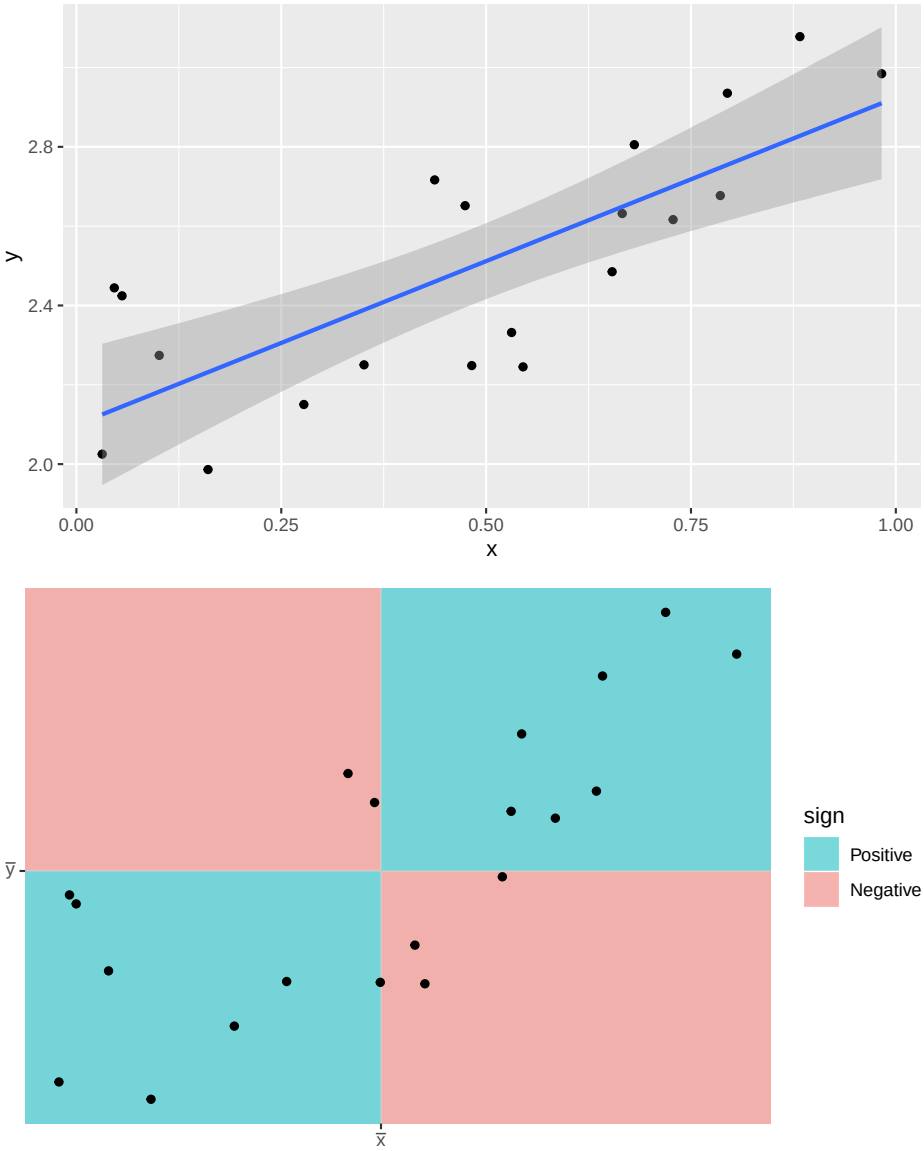
4.1.3 Correlation Plots

4.2 Pearson's Correlation Coefficient

We first consider Pearson's correlation coefficient, which is a statistics that measures the strength of the linear relationship between the predictor and response. Consider the following Pearson's correlation statistic

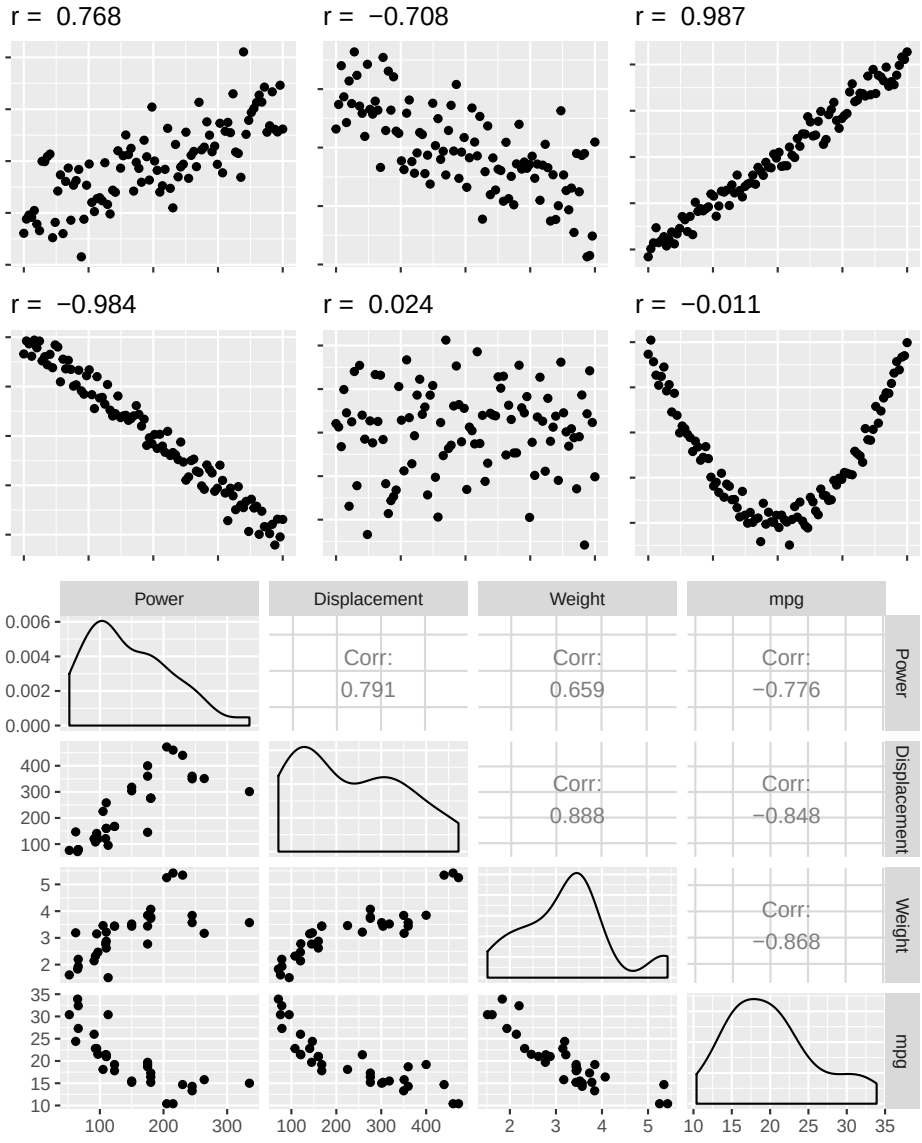
$$r = \frac{\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}{n-1}$$

where x_i and y_i are the x and y coordinate of the ith observation. Notice that each parenthesis value is the standardized value of each observation. If the x-value is big (greater than $\bar{x}$) and the y-value is large (greater than $\bar{y}$), then after multiplication, the result is positive. Likewise if the x-value is small and the y-value is small, both standardized values are negative and therefore after multiplication the result is positive. If a large x-value is paired with a small y-value, then the first value is positive, but the second is negative and so the multiplication result is negative.



The following are true about Pearson's correlation coefficient:

1. r is unit-less because we have standardized the x and y values.
2. $-1 \leq r \leq 1$ because of the scaling by $n-1$
3. A negative r denotes a negative relationship between x and y, while a positive value of r represents a positive relationship.
4. r measures the strength of the *linear* relationship between the predictor and response.

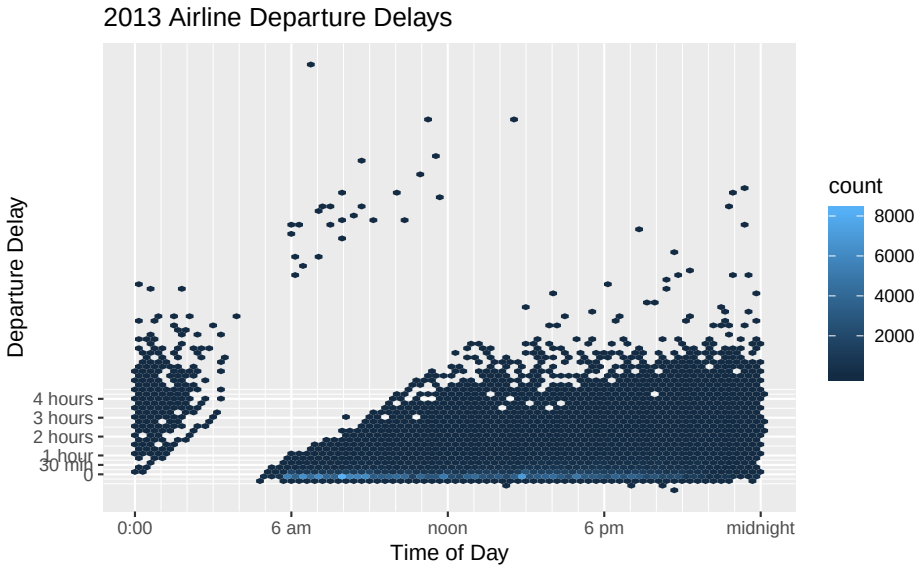


4.3 Overplotting

4.3.1 Transparency

4.3.2 Intensity Maps

delay	n	proportion
(-30,0]	200085	0.6091
(0,30]	80141	0.2439
(30,60]	21710	0.06608
(60,120]	16858	0.05132
(120,180]	5830	0.01775
(180,Inf]	3893	0.01185



This graph leads me to think that MOST flights are quite late, when if fact, they aren't. This is due to the problem of "proportional pixels". There is so much space and color devoted to flights that are more than 30 minutes late that the viewer can't help but have that impression.

4.3.3 Contour Plots

4.4 Exercises

1. Read Chapters 12 and 18 from Wilke's book. Feel free to skip section 12.3.

2. We will use a smaller version of the diamonds dataset that Wilke uses in his Chapter 18. You can download it from my GitHub site in a .csv file. We will examine price, carats, cut (Fair, Good, Very Good, Premium, and Ideal), color (J - D), and clarity (I1, SI2, SI1, VS2, VS1, VVS2, VVS1, IF), where the category labels I've given are from worst to best.

a) Had I not told you that clarity level IF is the best clarity, what graphs could you make to figure that out? Create and show a graph that demonstrates this and explain your graph.
b) If the diamond clarity isn't good, the dimond cutter won't worry too much about the quality of cut. Make a graph that demonstrates that and explain your reasoning.
c) While in principle it is possible for the dimond carats to be any number, they are often cut to be some common carat size. Create a visualization that shows this and discuss how the carat size changes as the cut and clarity improve.

3. From the Gapminder.com website, I've downloaded a bunch of interesting covariates about countries. You can find my dataset at my GitHub site in a .csv file. The variables I've included include the country region, year, population size, population growth, percent of population with basic sanitation, GDP per capita, Total GDP, life expectancy, adult male and female labor force participation rates. *Fertility is the number of children per woman, so a fertility rate of 2 children per woman is a stable population.*

a) For all the following questions, only consider the year 2015.
b) Investigate the relationship between GDP and GDP_per_capita. Why should we prefer to work with GDP_per_capita when comparing standards of living between, say, the United States and Canada?
c) Investigate the relationship between life expectancy, fertility, and GDP_per_capita. Do these relationships seem to vary by region? Comment on your graphs and relationships that you observe.
d) Investigate the relationship relationship between life expectancy, adult female labor force participation and fertility. Does this vary by region? Comment on your graphs and relationships that you observe.