# STA 141 - Exploratory Data Analysis and Visualization

*Derek L. Sonderegger*

*September 02, 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.

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.
  - 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

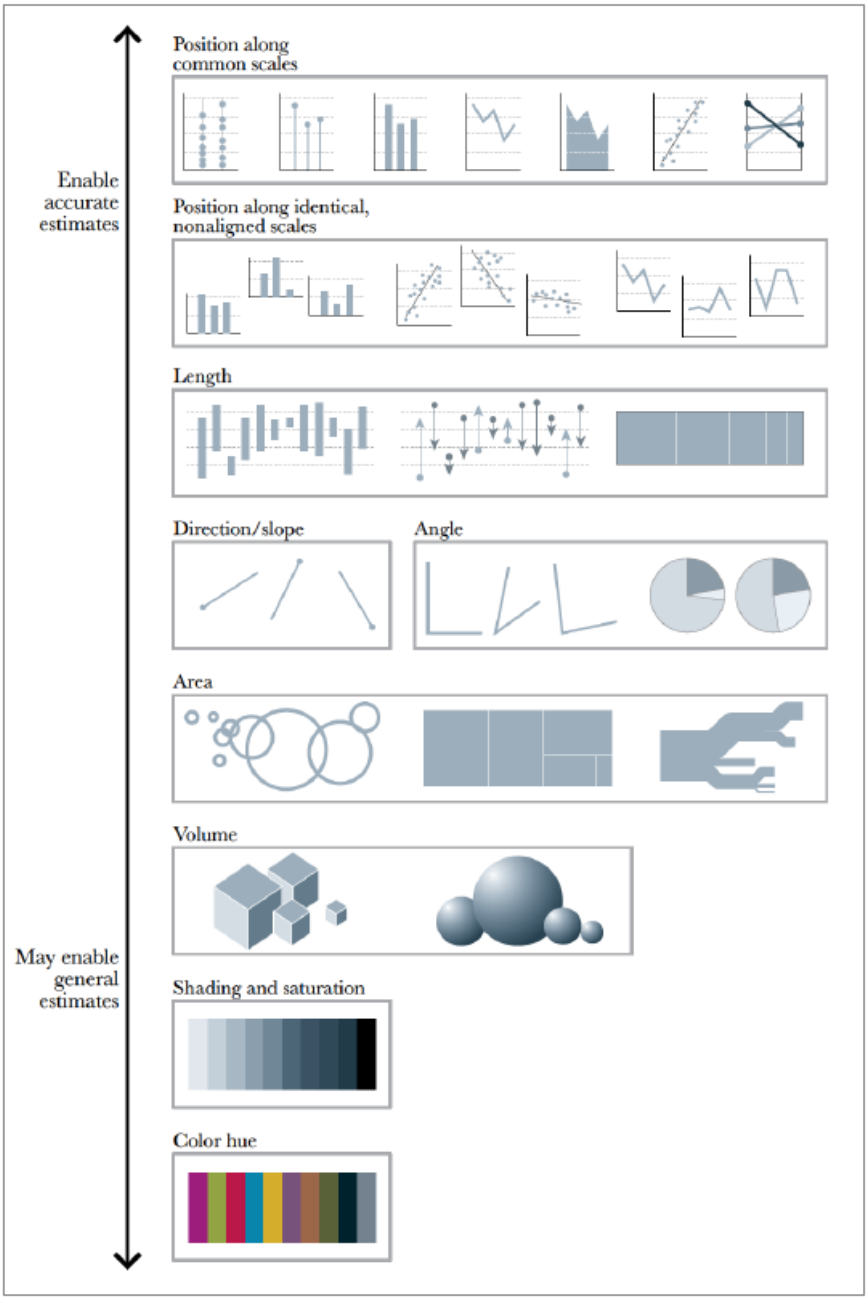
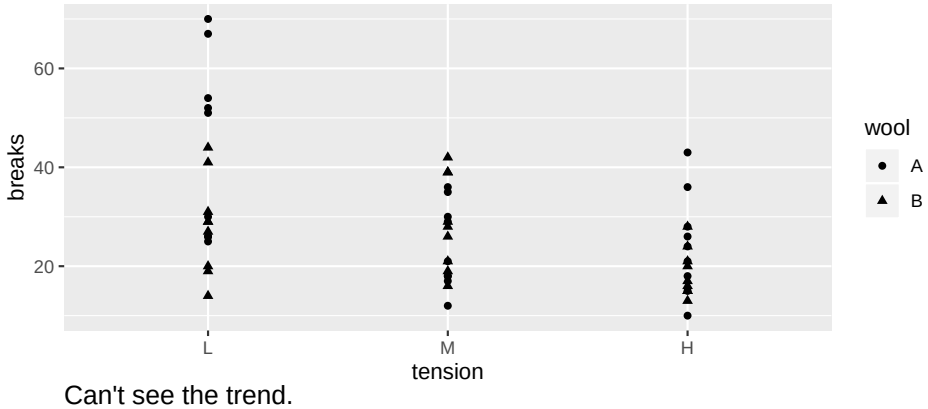


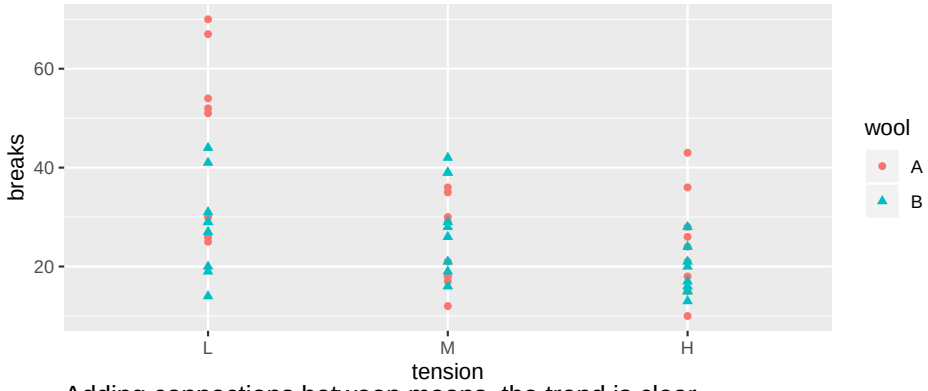

**Figure 5.5** Scale of elementary perceptual tasks, inspired by William Cleveland and Robert McGill.

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

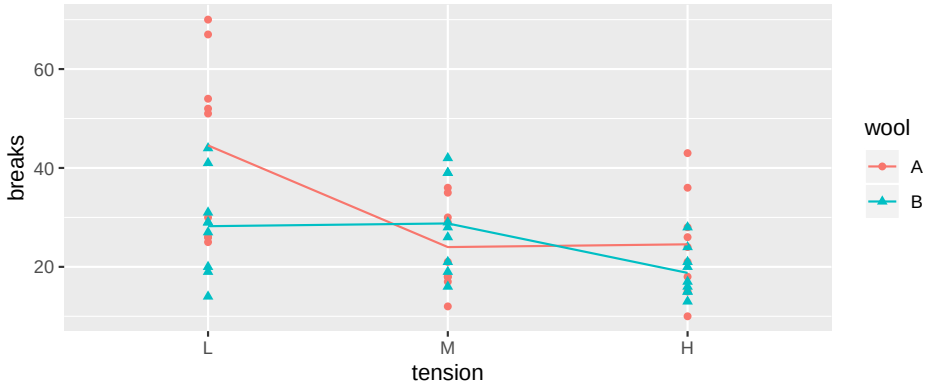
Can't easily distiguish wool types.

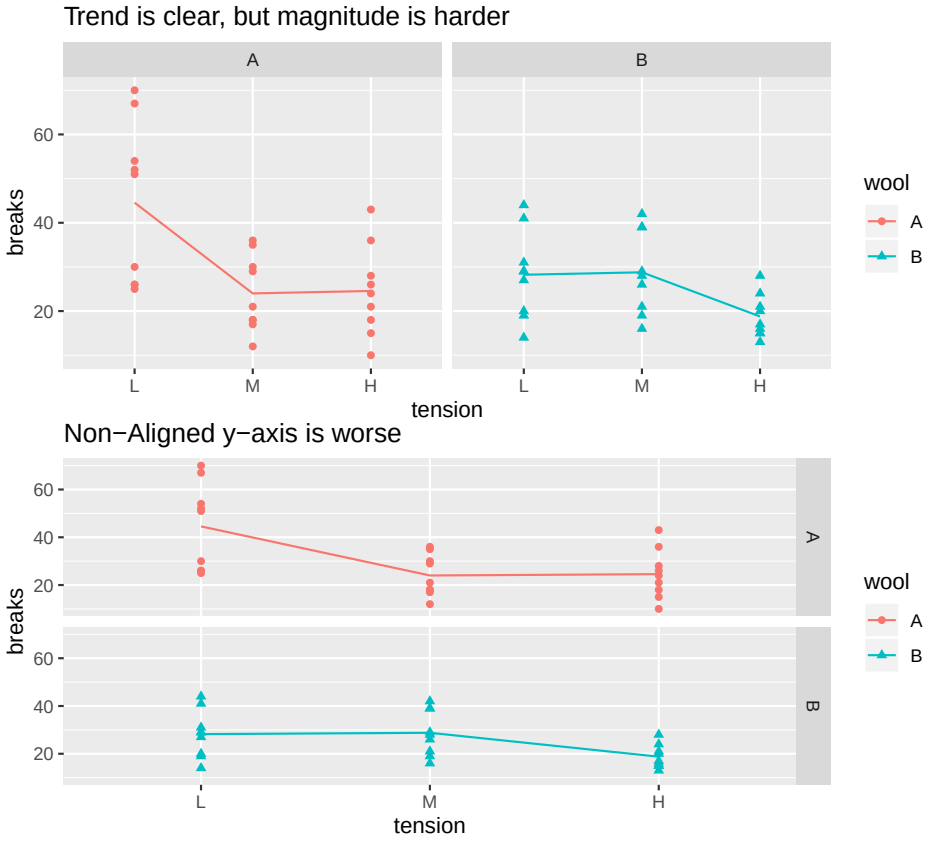



Can't see the trend.




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.

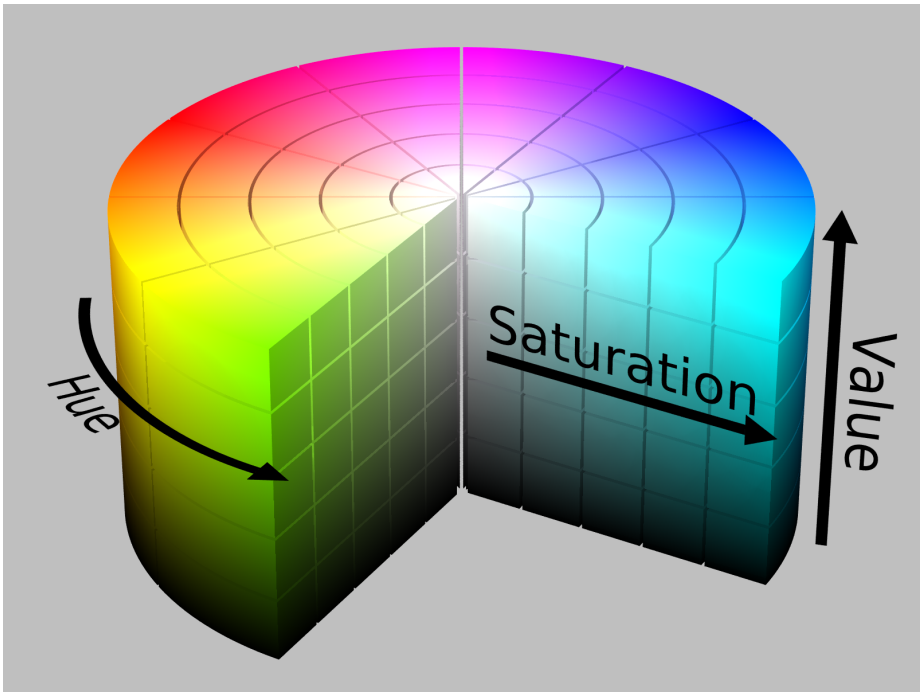



Figure 1.2: HSV Cylinder from Wikipedia

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.

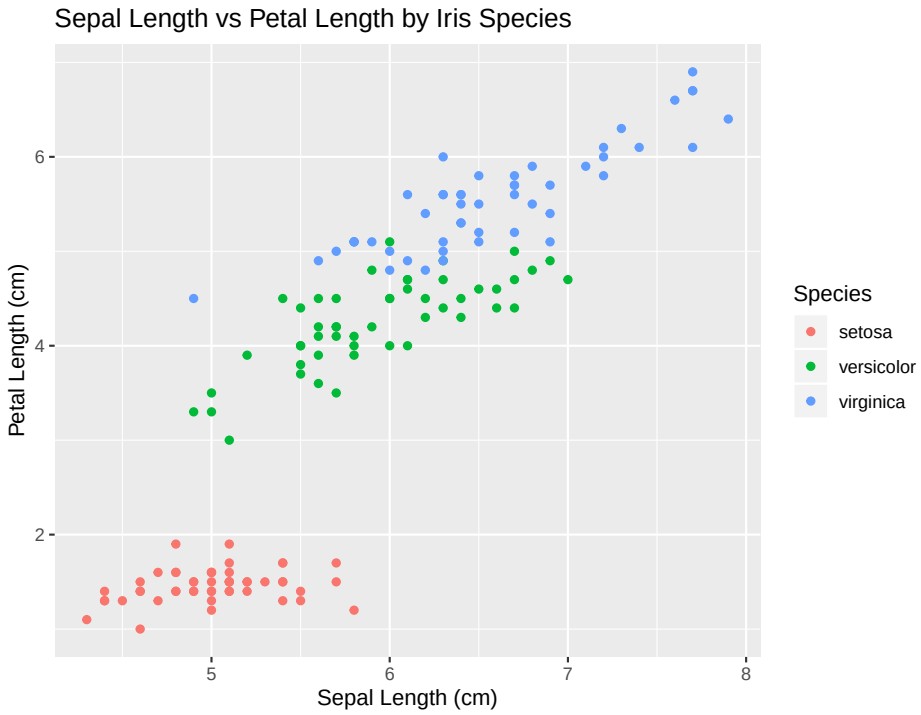
| StudentID | Assesment | Score |
|---|---|---|
| 1 | Exam 1 | 91 |

| 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

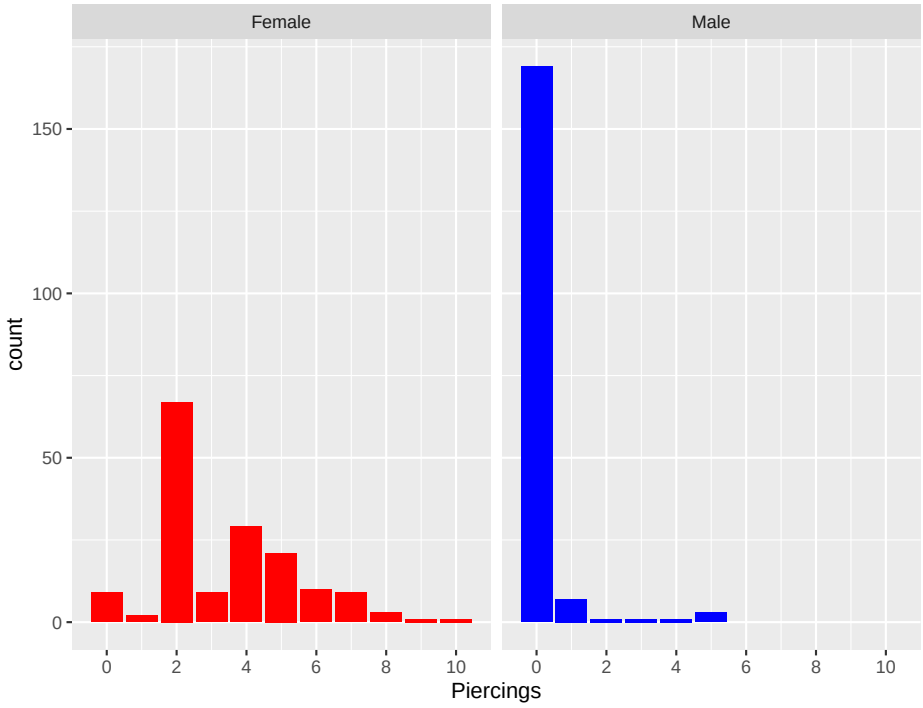
First we'll make a scatterplot with some colors.




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

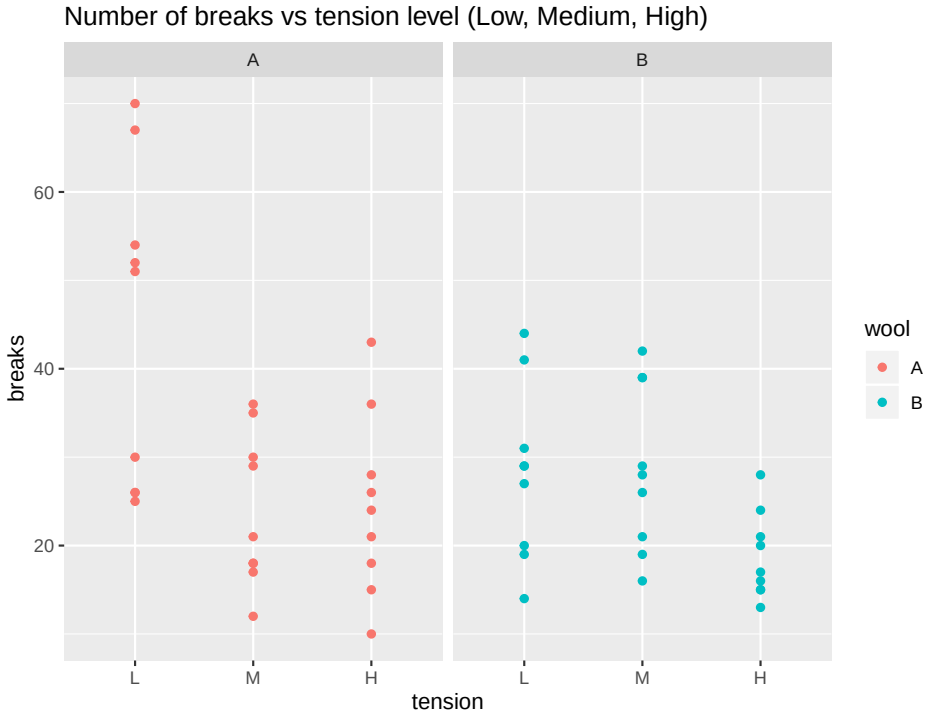
### 1.3.2.2 Task 2: Modifying how a variable is displayed

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

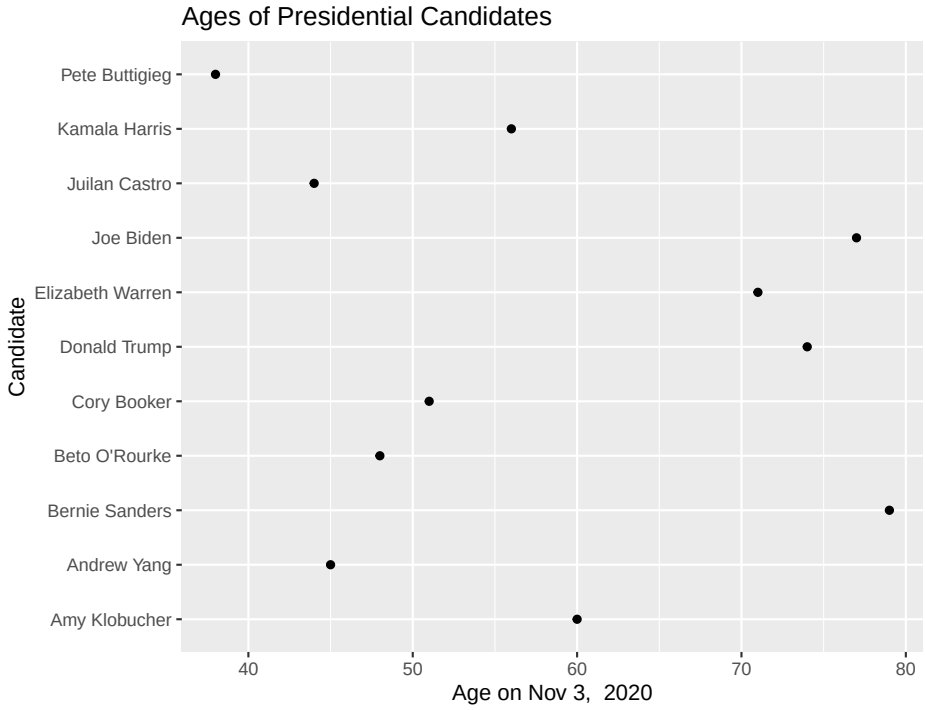
# Week 2

We will now start systematically working through various graph types and design decisions pertainent to each.

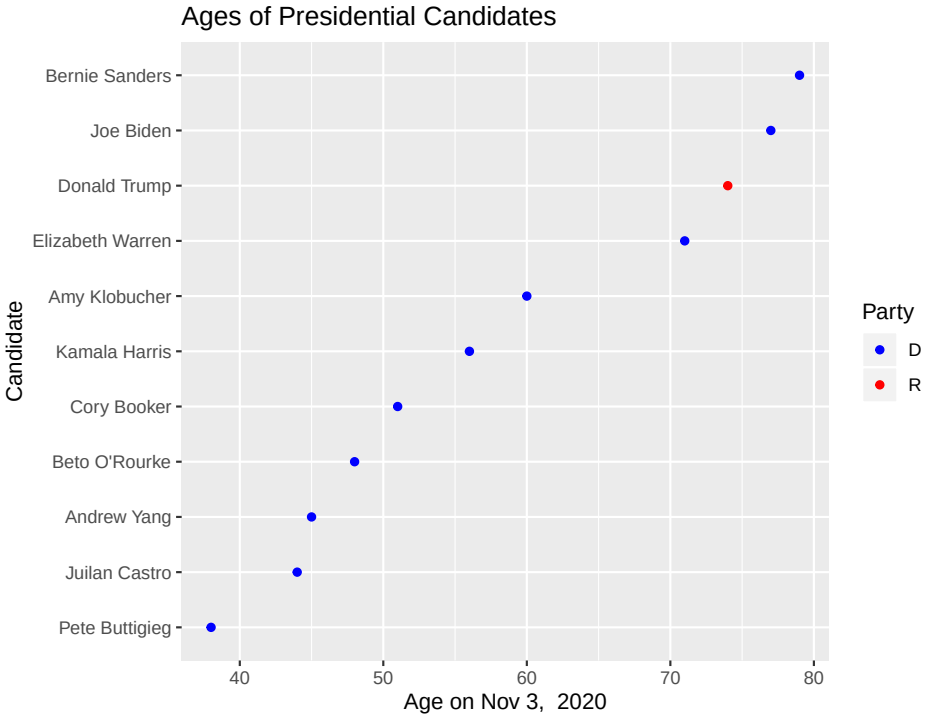
## 2.1 Visualizing Amounts

Read Chapter 6 from Wilke's Fundamentals of Data Visualiztion book.

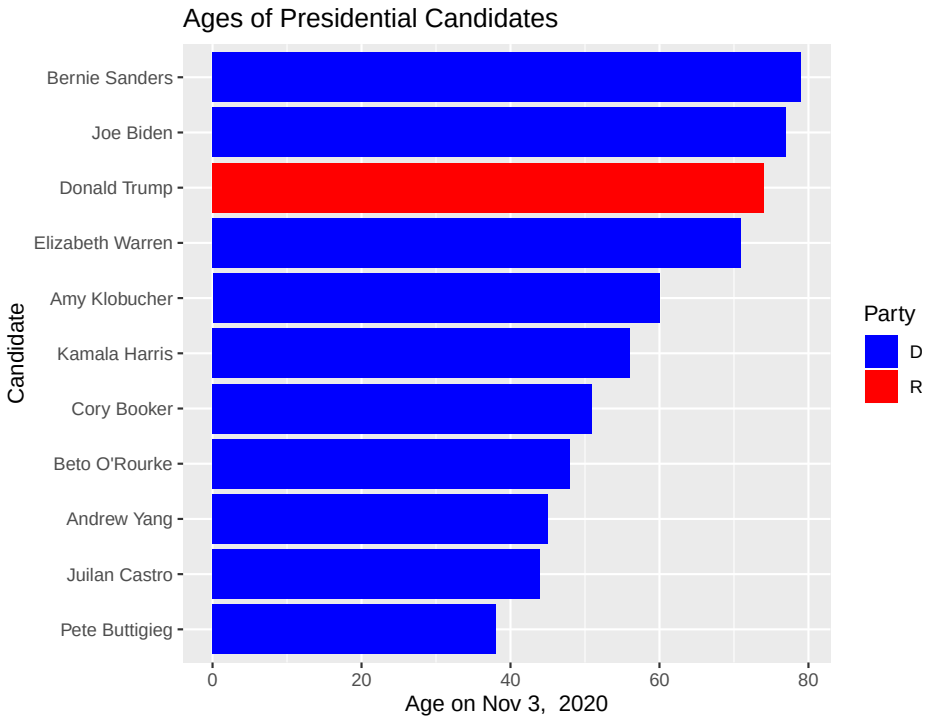
The 2020 presidential candidate field has a wide range of ages. The New York Times has a nice article showing the candidate ages. I grabbed a few of the most prominent candidates and pulled their birthdays from Wikipedia and then calculated their age on election day.



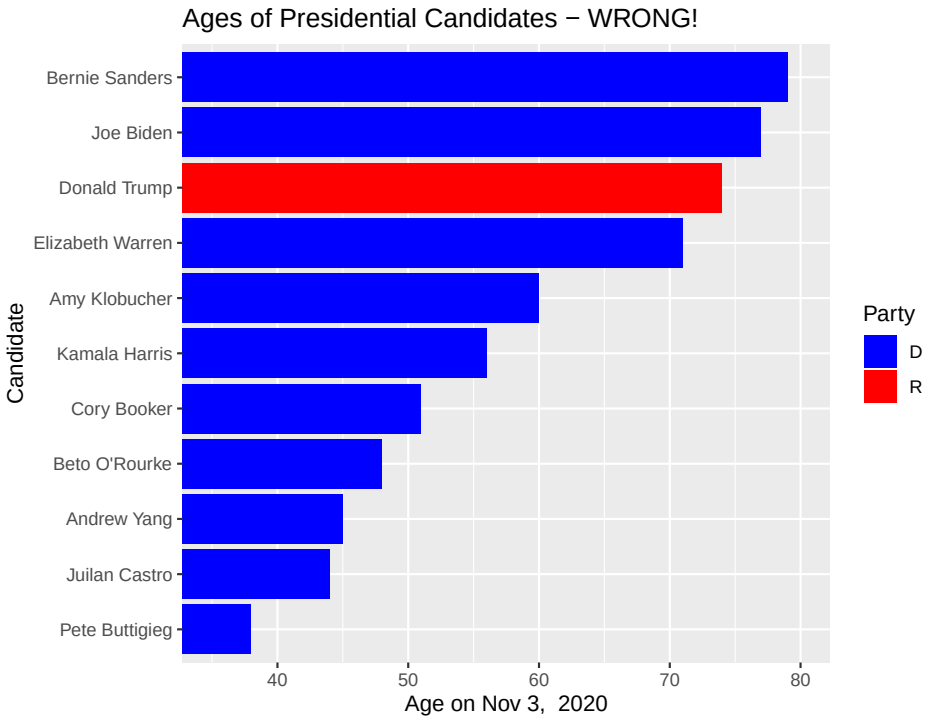

The order of the candidates is useless. Here we have ordered them alphabetically when we should try to think about an ordering that improves clarity. Lets switch to sorting the candidates by age.




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.




## 2.2 Visualizing Distributions