GRD 610A Data Visualization II

Data Manipulation

Jenn Schilling

February 17, 2021

Today

- Data Visualization of the Week
- Discussion of Chapter 3: The Truth Continuum of *The Truthful Art* (Cairo)
- Lab on Summary statistics, grouped analysis, manipulating rows, columns and tables - Chapter 5 of *Data Visualization* (Healy)
- Homework Assignment #2

Discussion - The Truth Continuum

Axiom: Any visualization is a model

(Cairo, 2016, p. 69)

Coda: The more adequately a model fits whatever it stands for without being needlessley complex, and the easier it is for is intended audience to interpret correctly, the better it will be.

(Cairo, 2016, p. 70)

- How do you define a model?
- What does it mean for one model to be "better" than another?
- Do you agree or disagree with this axiom and coda? Why or why not?
- Why is it important to consider complexity, controls, and models in data visualization?

Discussion - The Truth Continuum

"It's more complicated than that."

Good visualizations shouldn't over-simplify information. They need to clarify it. In many cases, clarifying a subject requires *increasing* the amount of information, not *reducing* it.

(Cairo, 2016, p. 78)

- What does it mean to present nuance and context in data visualization? What are the implications of this on design?
- How does this relate to considering your audience when creating a data visualization?

Discussion - The Truth Continuum

Mind Bugs

Patternicity

We look for and see patterns in everything, even when they are not there.

Storytelling

We like to generalize and see cause and effect, even when it's not there.

Confirmation

We look for information and interpret evidence as confirming our views, avoiding dissonance and ignoring alternate viewpoints.

What does this mean for data visualization design?

Discussion - The Truth Continuum

- What do you think of Cairo's truth continuum?
- How can we judge or evaluate the data visualizations we create on the truth continuum?
- What should we aim for on the truth continuum?
- What are some strategies that you will use to make truer data visualization?
- Is there a conflict between simplifying and making true models/visualizations? How can a designer approach such a conflict?

15 Minute Break

15:00

Chapter 5: Show the Right Numbers

The pipe operator: %>%

1. Individual-Level GSS Data on Region and Religion

id	bigregion	religion
1014	Midwest	Protestant
1544	South	Protestant
665	Northeast	None
1618	South	None
2115	West	Catholic
417	South	Protestant
2045	West	Protestant
1863	Northeast	Other
1884	Midwest	Christian
1628	South	Protestant

2. Summary Count of Religious Preferences by Census Region

bigregion	religion	N
Northeast	Protestant	123
Northeast	Catholic	149
Northeast	Jewish	15
Northeast	None	97
Northeast	Christian	14
Northeast	Other	31

3. Percent Religious Preferences by Census Region

bigregion	religion	N	pct
Northeast	Protestant	123	28.3
Northeast	Catholic	149	34.3
Northeast	Jewish	15	3.4
Northeast	None	97	22.3
Northeast	Christian	14	3.2
Northeast	Other	31	7.1

Figure 5.1 (Healy, 2019, p.95)

Dataset (2016 General Social Survey)

	year	id	ballot	age	childs	sibs	degree	race	sex	region	income16	rel
1	2016	1	1	47	3	2	Bachelor	White	Male	New England	$170000 or over	Non
2	2016	2	2	61	0	3	High School	White	Male	New England	$50000 to 59999	Non
3	2016	3	3	72	2	3	Bachelor	White	Male	New England	$75000 to $89999	Cath
4	2016	4	1	43	4	3	High School	White	Female	New England	$170000 or over	Cath
5	2016	5	3	55	2	2	Graduate	White	Female	New England	$170000 or over	Non
6	2016	6	2	53	2	2	Junior College	White	Female	New England	$60000 to 74999	Non

Pipe Operator - Step by Step

```
gss_sm %>%
  group_by(bigregion, religion) %>%
  summarize(N = n()) %>%
  mutate(freq = N / sum(N),
         pct = round((freq * 100), 0))
```

```
## # A tibble: 24 x 5
## # Groups:   bigregion [4]
##    bigregion religion       N    freq   pct
##    <fct>     <fct>      <int>   <dbl> <dbl>
##  1 Northeast Protestant   158 0.324      32
##  2 Northeast Catholic     162 0.332      33
##  3 Northeast Jewish        27 0.0553      6
##  4 Northeast None         112 0.230      23
##  5 Northeast Other         28 0.0574      6
##  6 Northeast <NA>           1 0.00205     0
##  7 Midwest   Protestant   325 0.468      47
##  8 Midwest   Catholic     172 0.247      25
##  9 Midwest   Jewish         3 0.00432     0
## 10 Midwest   None         157 0.226      23
## # ... with 14 more rows
```

Assignment / Equals

Before

```
p <- ggplot(data = gapminder,
            mapping = aes(x = year,
                          y = gdpPercap))
```

Now

```
gss_sm %>%
  group_by(bigregion, religion) %>%
  summarize(N = n()) %>%
  mutate(freq = N / sum(N),
         pct = round((freq * 100), 0))
```

Creating Columns / Variables

```
gss_sm %>%
  group_by(bigregion, religion) %>%
  summarize(N = n()) %>%
  mutate(freq = N / sum(N),
         pct = round((freq * 100), 0))
```

Lab Time

Pages 91 - 101, 110-113, 132 (bullet points 2-3), 133 (bullet points 1-3)

Homework Assignment

Task: Create 3 calculated fields and plot them.

Due: February 24, 2020

Rubric

Notes

- You should explore a dataset other than `gapminder` (some ideas: `babynames`, `palmerpenguins`, a CSV file you found; you may also use `gss_sm` or `organdata`, but you must create something different than the book/lab)
- To use an R package dataset that you have not used before, remember to run `install.packages("package_name")` once in the console and add `library(package_name)` to the `setup` portion of the .Rmd file
- Use your resources: Healy, Google, Student Community BUT cite where you get code from if you copy it directly
- See Blackboard assignment for a template .Rmd file

Tasks to Complete

- Reading (Cairo - Chapter 4: Of Conjuectures and Uncertainty)
- Homework #2
- Prepare for your Data Visualization of the Week