431 Class 11

Thomas E. Love

2017-10-03

Today's Agenda

1. Forming Scatterplot and Correlation Matrices
2. Building Tables Well: Ehrenberg paper
3. Project Task B
4. Highlights from Leek Chapters 1-4, 12

Please sit near the rest of your Task B Group today.

Today's R Setup

```
library(GGally); library(tidyverse)

source("Love-boost.R")
```

A new sample of 500 observations without NA

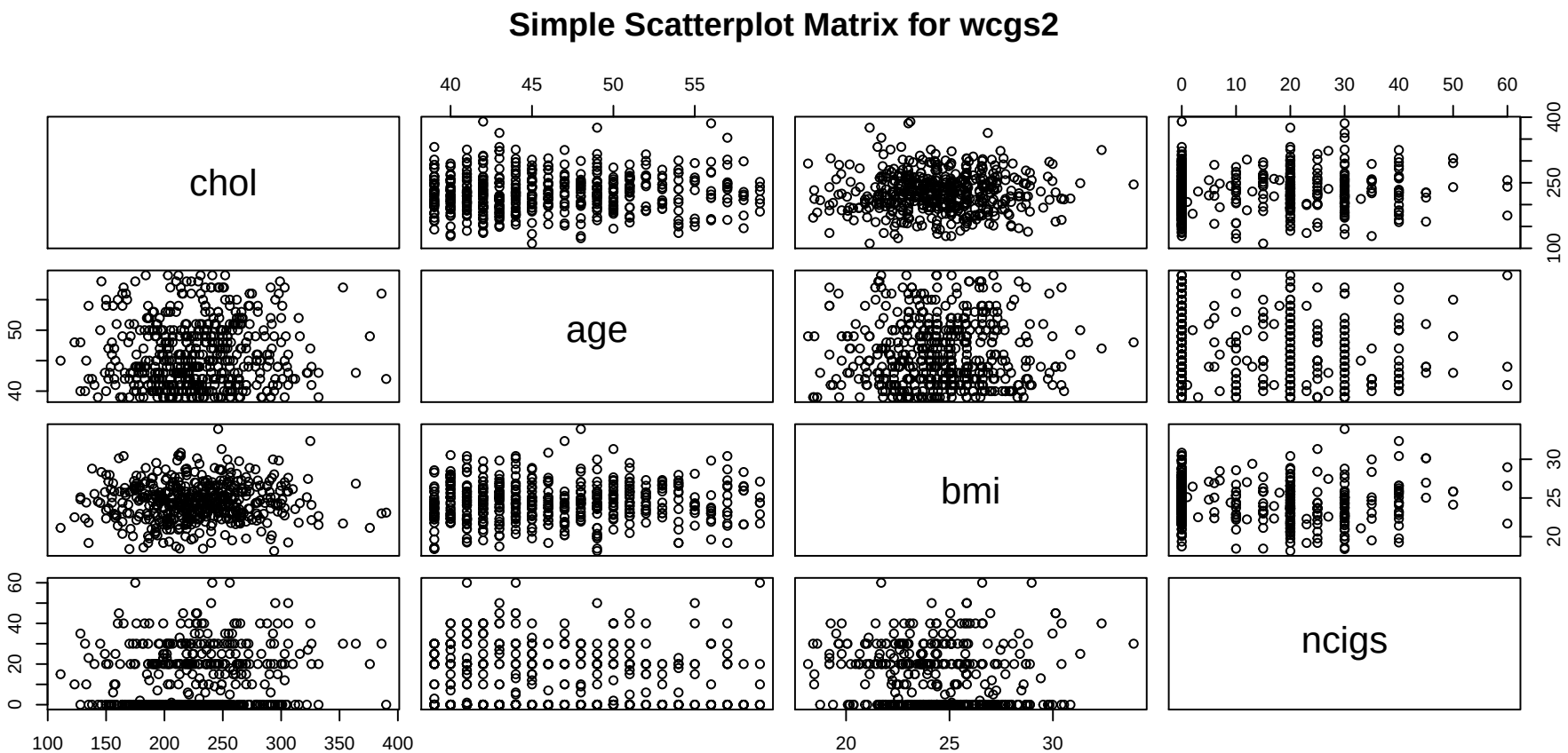
```
wcgs.full <- read.csv("wcgs.csv") %>% tbl_df()

set.seed(43102)
wcgs2 <- wcgs.full %>%
  filter(complete.cases(age, chol, bmi,
                        ncigs, arcus, dibpat)) %>%
  select(id, age, chol, bmi,
         ncigs, arcus, dibpat) %>%
  sample_n(500, replace = FALSE)
```

Codebook for wcgs2

Name	Stored As	Type	Details (units, levels, etc.)
id	integer	(nominal)	ID #, nominal and uninteresting
age	integer	quantitative	age, in years - no decimal places
chol	integer	quantitative	total cholesterol, mg/dL
arcus	integer	(nominal)	arcus senilis present (1) or absent (0)
dibpat	factor (2)	(binary)	behavioral pattern: A or B
bmi	number	quantitative	body-mass index
ncigs	integer	quantitative	number of cigarettes smoked per day

Multivariable Descriptions: A Scatterplot Matrix

```
pairs (~ chol + age + bmi + ncigs,
       data=wcgs2, main="Simple Scatterplot Matrix for wcgs2")
```



Correlation Matrix for Numeric Variables

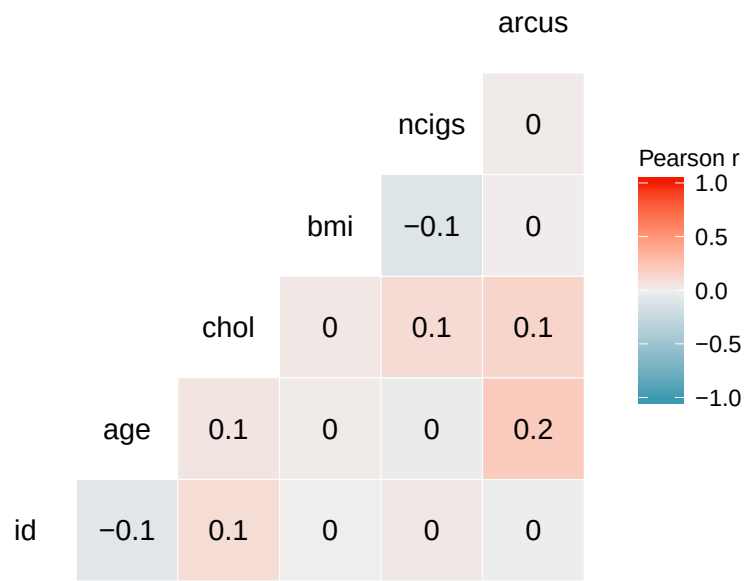
```
wcgs2 %>%
  select(chol, age, bmi, ncigs) %>%
  cor() %>%
  knitr::kable()
```

	chol	age	bmi	ncigs
chol	1.0000000	0.0543491	0.0334955	0.1076450
age	0.0543491	1.0000000	0.0196684	-0.0275027
bmi	0.0334955	0.0196684	1.0000000	-0.1053668
ncigs	0.1076450	-0.0275027	-0.1053668	1.0000000

Using GGally for a Correlation Matrix

```
ggcorr(wcgs2, name = "Pearson r", label = TRUE)
```

```
Warning in ggcorr(wcgs2, name = "Pearson r", label =
TRUE): data in column(s) 'dibpat' are not numeric and
were ignored
```

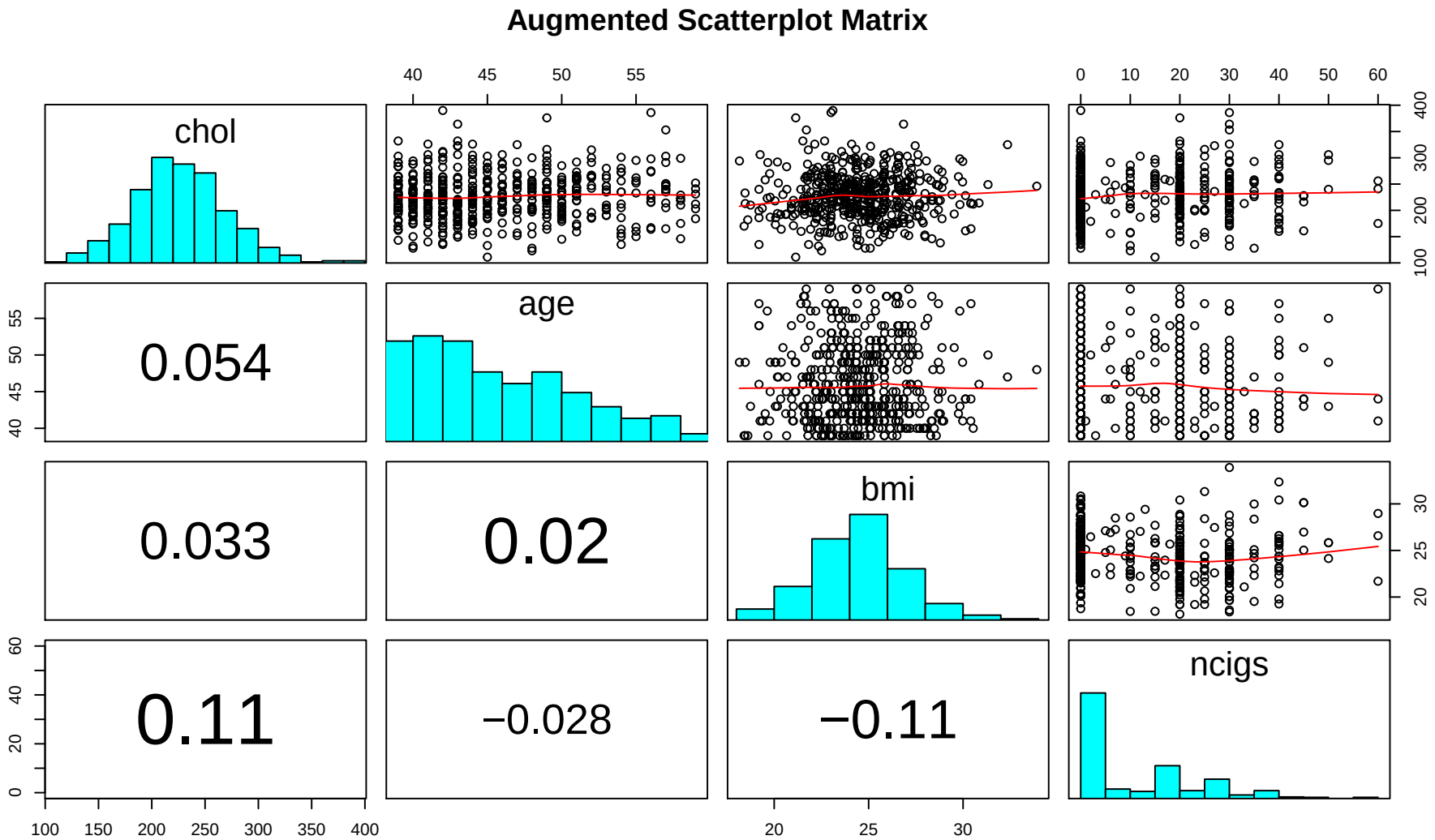


My Favorite Scatterplot Matrix

My favorite way to augment this plot adds smooths to the upper panel, and correlations in the lower panel, with histograms down the diagonal. To do this, we first create two functions (these modifications come from Chang's R Graphics Cookbook), called `panel.hist` and `panel.cor`.

These functions are in the Love-boost.R script.

Augmented Scatterplot Matrix



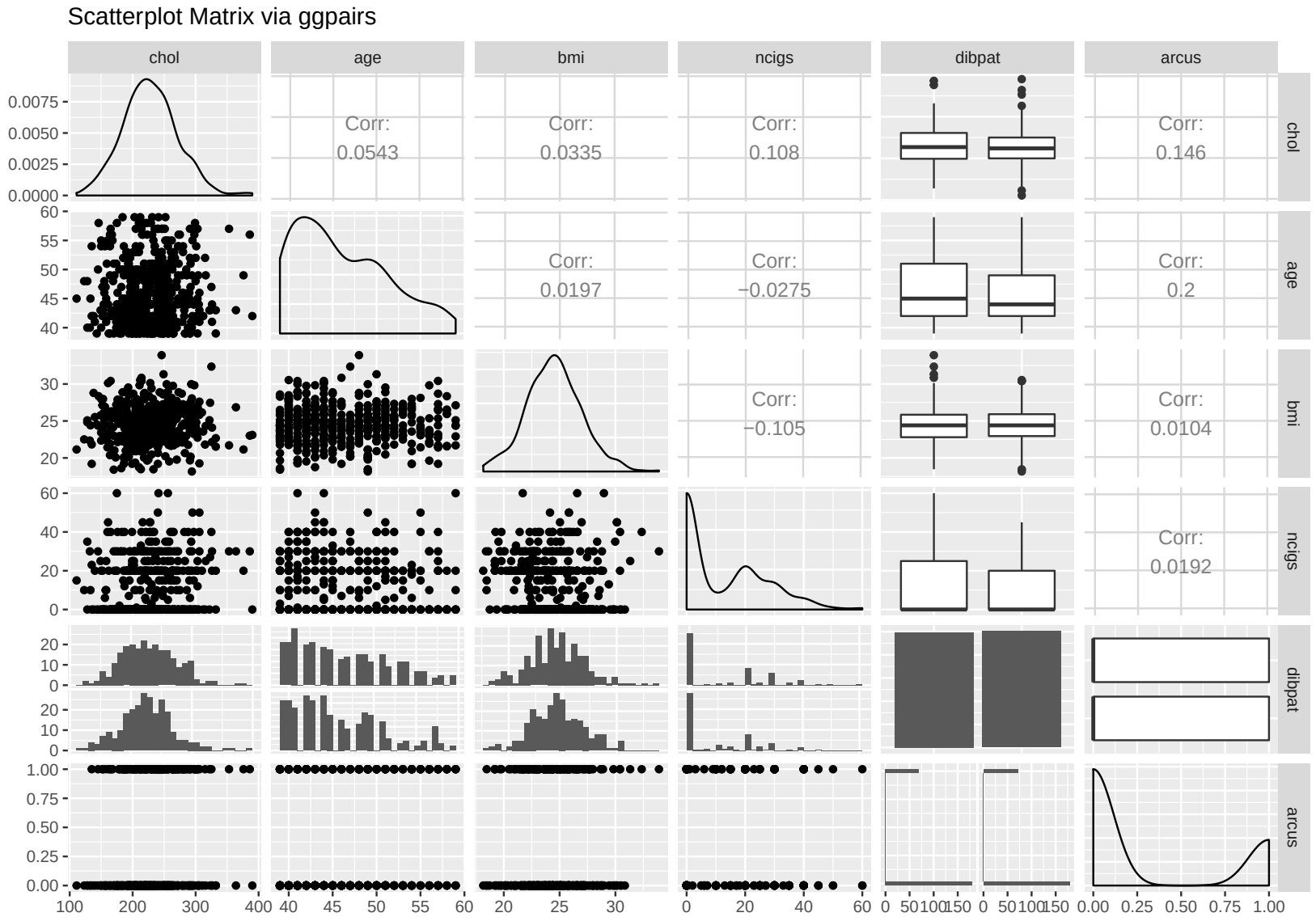
Code for Augmented Scatterplot Matrix

```
pairs (~ chol + age + bmi + ncigs, data=wcgs2,
       main="Augmented Scatterplot Matrix",
       upper.panel = panel.smooth,
       diag.panel = panel.hist,
       lower.panel = panel.cor)
```

Using GGally for a Scatterplot Matrix (Code)

```
tempdat <- wcgs2 %>%
  select(chol, age, bmi, ncigs, dibpat, arcus)

ggpairs(tempdat, title = "Scatterplot Matrix via ggpairs")
```

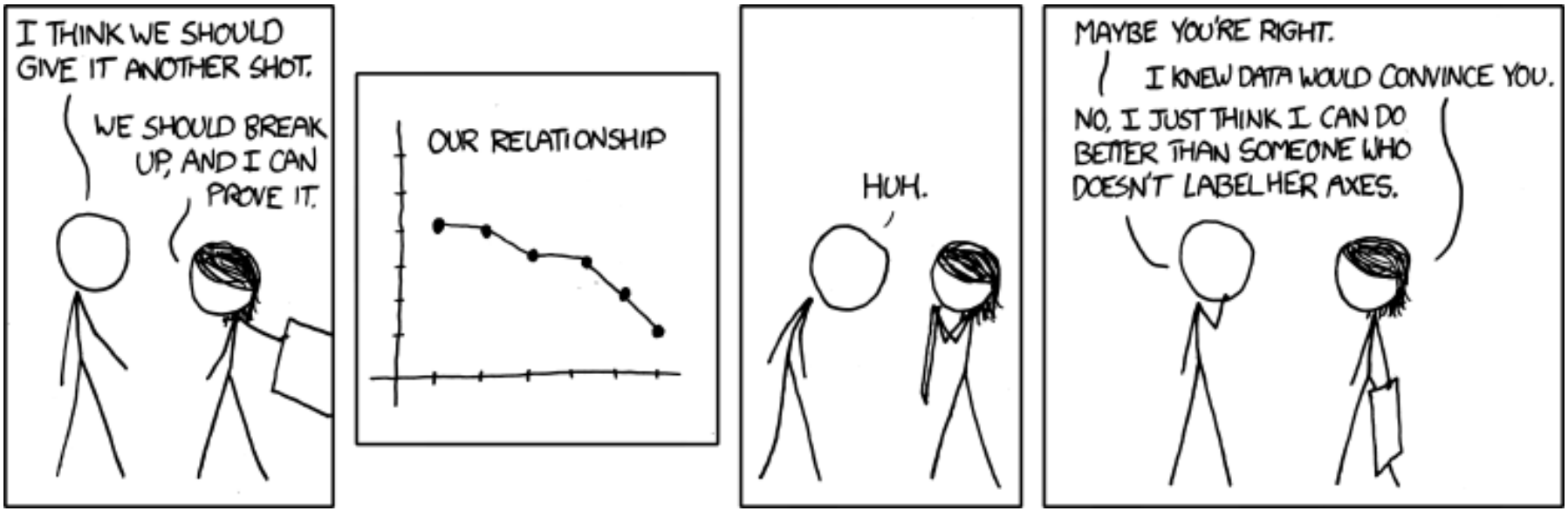
Using GGally for a Scatterplot Matrix (Result)



What Makes a Good Graph? (Tufte, lightly edited)

1. During the discovery stage of your work use any style or type of graph you wish. Design becomes important as soon as you want to convey information. At that point you have to create graphs that communicate ideas to others.
2. Graphs communicate most easily when they have a specific message – for instance, "coffee production up!" They lose impact and are less successful when their point is vague – for example, "The number of students in public high schools, 1993-2003."
3. Graphs are powerful when you use the title to reinforce your specific message – "The number of students in public high schools has fallen by a third in ten years." Such transparent messages will be understood and remembered by readers. If you don't tell readers what the graph is saying, some will never know.
4. After years of hard thinking, I concluded that graphs are like jokes: if you have to explain them they have failed.

This doesn't apply to axis labels



Building Tables Well

Getting information from a table is like extracting sunlight from a cucumber.

Farquhar AB and Farquhar H

Building Tables Well

There are three key tips related to the development of tables, in practice, as described by Ehrenberg, and also by Howard Wainer[1] who concisely states them as:

1. Order the rows and columns in a way that makes sense.
2. Round - a lot!
3. ALL is different and important.

[1] Visual Revelations (1997), Chapter 10.

Now HERE's a Contingency Table

TABLE 1

Deaths Due to Unexpected Events, by Type of Event, Selected Countries: Mid-1970's

(Rate per 100,000 population)

Country	Year[1]	Deaths due to all causes	Deaths due to unexpected events					
			Total	Transport accidents	Natural factors[2]	Accidents occurring mainly in industry[3]	Homicides and injuries caused intentionally[4]	Other causes[5]
Austria	1975	1,277.2	75.2	34.8	29.7	4.3	1.6	4.8
Belgium	1975	1,218.5	62.6	25.0	25.8	1.5	9	9.4
Canada	1974	742.0	62.1	30.9	18.0	3.9	2.5	6.8
Denmark	1976	1,059.5	41.1	18.3	15.6	1.0	7	5.5
Finland	1974	952.5	62.3	23.7	26.0	2.9	2.6	7.1
France	1974	1,049.5	77.8	23.8	31.0	1.0	9	21.1
Germany (Fed. Rep.)	1975	1,211.8	66.4	24.8	31.6	1.8	1.2	7.0
Ireland	1975	1,060.7	48.6	19.8	20.1	1.9	1.0	5.8
Italy	1974	957.8	47.2	22.8	19.2	1.9	1.1	2.2
Japan	1976	625.6	30.5	13.2	9.7	2.1	1.3	4.2
Netherlands	1975	832.2	40.3	17.8	18.2	1.0	7	2.6
Norway	1976	998.9	48.4	17.3	25.1	1.9	7	3.4
Sweden	1975	1,076.6	55.8	17.2	27.9	1.3	1.1	8.3
Switzerland	1976	904.1	48.4	20.6	20.4	2.1	9	4.4
United Kingdom	1976	1,217.9	34.8	13.0	13.9	1.3	1.1	5.5
United States	1975	888.5	60.6	23.4	15.8	2.6	10.0	8.8

[1]Most current year data available.
[2]Includes fatal accidents due to poisoning, falls, fire, and drowning.
[3]For some countries data relate to accidents caused by machines only.
[4]By another person, including police.
[5]Includes accidents caused by firearms, war injuries, injuries of undetermined causes, and all other accidental causes.

Source: United Nations, World Health Organization, World Health Statistics Annual, 1978, vol. I, Vital Statistics and Cause of Death.

Four Questions

1. What is the general level (per 100,000 population) of accidental death in the countries chosen?
2. How do the countries differ with respect to their rates of accidental death?
3. What are the principal causes of accidental death? Which are the most frequent? The least frequent?
4. Are there any unusual interactions between country and cause of accidental death?

See the Supplementary Table on the Class 11 page.

Wainer H (1997) *Visual Revelations*, Chapter 10

Deaths due to Unexpected Events, by Type of Event, Selected Countries: Mid-1970's (Rate per 100,000 population)

Country	Year[1]	Deaths due to all causes	Deaths due to unexpected events					
			Total	Transport Accidents	Natural Factors[2]	Accidents Occurring Mainly in Industry[3]	Homicides and Injuries caused intentionally[4]	Other Causes[5]
Austria	1975	1,277.2	75.2	34.8	29.7	4.3	1.6	4.8
Belgium	1975	1,218.5	62.6	25.0	25.8	1.5	9	9.4
Canada	1974	742.0	62.1	30.9	18.0	3.9	2.5	6.8
Denmark	1976	1,059.5	41.1	18.3	15.6	1.0	7	5.5
Finland	1974	952.5	62.3	23.7	26.0	2.9	2.6	7.1
France	1974	1,049.5	77.8	23.8	31.0	1.0	9	21.1
Germany (Fed. Rep.)	1975	1,211.8	66.4	24.8	31.6	1.8	1.2	7.0
Ireland	1975	1,060.7	48.6	19.8	20.1	1.9	1.0	5.8
Italy	1974	957.8	47.2	22.8	19.2	1.9	1.1	2.2
Japan	1976	625.6	30.5	13.2	9.7	2.1	1.3	4.2
Netherlands	1975	832.2	40.3	17.8	18.2	1.0	7	2.6
Norway	1976	998.9	48.4	17.3	25.1	1.9	7	3.4
Sweden	1975	1,076.6	55.8	17.2	27.9	1.3	1.1	8.3
Switzerland	1976	904.1	48.4	20.6	20.4	2.1	9	4.4
United Kingdom	1976	1,217.9	34.8	13.0	13.9	1.3	1.1	5.5
United States	1975	888.5	60.6	23.4	15.8	2.6	10.0	8.8

Source: United Nations, World Health Organization, World Health Statistics Annual, 1978, vol. I, Vital Statistics and Cause of Death.

Ehrenberg's Main Ideas?

Wainer's Three Rules for Table Construction

1. Order the rows and columns in a way that makes sense.
2. Round, a lot!
3. ALL is different and important

- Wainer H (1997) *Visual Revelations* Chapter 10.

Alabama First!

Which is more useful to you?

2013 Percent of Students in grades 9-12 who are obese

State	% Obese	95% CI	Sample Size
Alabama	17.1	(14.6 - 19.9)	1,499
Alaska	12.4	(10.5-14.6)	1,167
Arizona	10.7	(8.3-13.6)	1,520
Arkansas	17.8	(15.7-20.1)	1,470
Connecticut	12.3	(10.2-14.7)	2,270
Delaware	14.2	(12.9-15.6)	2,475
Florida	11.6	(10.5-12.8)	5,491
...			
Wisconsin	11.6	(9.7-13.9)	2,771
Wyoming	10.7	(9.4-12.2)	2,910

or ...

Alabama First!

State	% Obese	95% CI	Sample Size
Kentucky	18.0	(15.7 - 20.6)	1,537
Arkansas	17.8	(15.7 - 20.1)	1,470
Alabama	17.1	(14.6 - 19.9)	1,499
Tennessee	16.9	(15.1 - 18.8)	1,831
Texas	15.7	(13.9 - 17.6)	3,039
...			
Massachusetts	10.2	(8.5 - 12.1)	2,547
Idaho	9.6	(8.2 - 11.1)	1,841
Montana	9.4	(8.4 - 10.5)	4,679
New Jersey	8.7	(6.8 - 11.2)	1,644
Utah	6.4	(4.8 - 8.5)	2,136

It is a rare event when Alabama first is the best choice.

Archiving Data: Sortable Online Tables

2013: Percent of students in grades 9-12 who are obese†

Location Type	Location	Value	95% CI	Sample Size
National	National	13.7	(12.6-14.9)	12580
	Kentucky	18.0	(15.7-20.6)	1537
	Arkansas	17.8	(15.7-20.1)	1470
	Alabama	17.1	(14.6-19.9)	1499
	Tennessee	16.9	(15.1-18.8)	1831
	Texas	15.7	(13.9-17.6)	3039
	West Virginia	15.6	(13.5-18.0)	1561
	Mississippi	15.4	(13.1-17.9)	1446
	Missouri	14.9	(12.3-17.8)	1539
	Delaware	14.2	(12.9-15.6)	2475
	South Carolina	13.9	(11.6-16.5)	1555
	Louisiana	13.5	(11.0-16.4)	1034
	North Dakota	13.5	(11.8-15.3)	1931
	Hawaii	13.4	(11.6-15.4)	4405
	Vermont	13.2	(11.3-15.4)	5853
	Michigan	13.0	(11.4-14.9)	4110
	Ohio	13.0	(10.8-15.5)	1404

Notes on the Data in the previous slides

Source: Estimates from the National Youth Risk Behavior Surveillance System (YRBSS). Available at http://www.cdc.gov/nccdphp/DNPAO/index.html.

To go directly to this table visit this link

- Obese is defined as body mass index (BMI)-for-age and sex $\geq$ 95th percentile based on the 2000 CDC growth chart; BMI was calculated from self-reported weight and height (weight [kg]/ height [m^2]).

Order rows and columns sensibly

- Alabama First!
- Size places - put the largest first. We often look most carefully at the top.
- Order time from the past to the future to help the viewer.
- If there is a clear predictor-outcome relationship, put the predictors in the rows and the outcomes in the columns.

Order the rows and columns sensibly.

Country	Total unexpected deaths	Transport accidents	Natural factors	Industrial accidents	Homicides	Other Causes
France	77.8	23.8	31.0	1.0	0.9	21.1
Austria	75.2	34.8	29.7	4.3	1.6	4.8
Germany	66.4	24.8	31.6	1.8	1.2	7.0
Belgium	62.6	25.0	25.8	1.5	0.9	9.4
Finland	62.3	23.7	26.0	2.9	2.6	7.1
Canada	62.1	30.9	18.0	3.9	2.5	6.8
United States	60.6	23.4	15.8	2.6	10.0	8.8
Sweden	55.8	17.2	27.9	1.3	1.1	8.3
Ireland	48.6	19.8	20.1	1.9	1.0	5.8
Norway	48.4	17.3	25.1	1.9	0.7	3.4
Switzerland	48.4	20.6	20.4	2.1	0.9	4.4
Italy	47.2	22.8	19.2	1.9	1.1	2.2
Denmark	41.1	18.3	15.6	1.0	0.7	5.5
Netherlands	40.3	17.8	18.2	1.0	0.7	2.6
United Kingdom	34.8	13.0	13.9	1.3	1.1	5.5
Japan	30.5	13.2	9.7	2.1	1.3	4.2

Round - a lot!

- Humans cannot understand more than two digits very easily.
- We almost never care about accuracy of more than two digits.
- We can almost never justify more than two digits of accuracy statistically.

Suppose we want to report a correlation coefficient of 0.25

- How many observations do you think you would need to justify such a choice?
- To report 0.25 meaningfully, we should know the second digit isn't 4 or 6, right?

Reporting a correlation coefficient of 0.25

To report 0.25 meaningfully, we desire to be sure that the second digit isn't 4 or 6.

- That requires a standard error less than 0.005
- The *standard error* of any statistic is proportional to 1 over the square root of the sample size, n.

So $\frac{1}{\sqrt{n}} \sim 0.005$, but that means $\sqrt{n} = \frac{1}{0.005} = 200$.

And if $\sqrt{n} = 200$, then $n = (200)^2 = 40{,}000$.

Do we usually have 40,000 observations?

Round, a lot!

Country	Total unexpected deaths	Transport accidents	Natural factors	Industrial accidents	Homicides	Other Causes
France	78	24	31	1	1	21
Austria	75	35	30	4	2	5
Germany	66	25	32	2	1	7
Belgium	63	25	26	2	1	9
Finland	62	24	26	3	3	7
Canada	62	31	18	4	3	7
United States	61	23	16	3	10	9
Sweden	56	17	28	1	1	8
Ireland	49	20	20	2	1	6
Norway	48	17	25	2	1	3
Switzerland	48	21	20	2	1	4
Italy	47	23	19	2	1	2
Denmark	41	18	16	1	1	6
Netherlands	40	18	18	1	1	3
United Kingdom	35	13	14	1	1	6
Japan	31	13	10	2	1	4

ALL is different and important

Country	**Total unexpected deaths**	Transport accidents	Natural factors	Industrial accidents	Homicides	Other Causes
France	**78**	24	31	1	1	21
Austria	**75**	35	30	4	2	5
Germany	**66**	25	32	2	1	7
Belgium	**63**	25	26	2	1	9
Finland	**62**	24	26	3	3	7
Canada	**62**	31	18	4	3	7
United States	**61**	23	16	3	10	9
Sweden	**56**	17	28	1	1	8
Ireland	**49**	20	20	2	1	6
Norway	**48**	17	25	2	1	3
Switzerland	**48**	21	20	2	1	4
Italy	**47**	23	19	2	1	2
Denmark	**41**	18	16	1	1	6
Netherlands	**40**	18	18	1	1	3
United Kingdom	**35**	13	14	1	1	6
Japan	**31**	13	10	2	1	4

Cluster when you can, and highlight outliers.

Country	Total unexpected deaths	Transport accidents	Natural factors	Industrial accidents	Homicides	Other Causes
France	78	24	31	1	1	21
Austria	75	35	30	4	2	5
Germany	66	25	32	2	1	7
Belgium	63	25	26	2	1	9
Finland	62	24	26	3	3	7
Canada	62	31	18	4	3	7
United States	61	23	16	3	10	9
Sweden	56	17	28	1	1	8
Ireland	49	20	20	2	1	6
Norway	48	17	25	2	1	3
Switzerland	48	21	20	2	1	4
Italy	47	23	19	2	1	2
Denmark	41	18	16	1	1	6
Netherlands	40	18	18	1	1	3
United Kingdom	35	13	14	1	1	6
Japan	31	13	10	2	1	4

Visualizing Categories

http://flowingdata.com/projects/2016/alcohol-world/

Recorded APC is defined as the recorded amount of alcohol consumed per capita (15+ years) over a calendar year in a country, in litres of pure alcohol. The indicator only takes into account the consumption which is recorded from production, import, export, and sales data often via taxation.

- Numerator: The amount of recorded alcohol consumed per capita (15+ years) during a calendar year, in litres of pure alcohol.
- Denominator: Midyear resident population (15+ years) for the same calendar year, UN World Population Prospects, medium variant.

http://apps.who.int/gho/indicatorregistry/App_Main/view_indicator.aspx?iid=462

Project Task B Groups

1. Club Tukey (Estee, Chaim, Caroline, Hyung Chul, Vinh, Frances)
2. 4 brains 1 heart (Laura Baldassari, Kedar, Sarah Planchon, Sneha, Xin Xin)
3. The Foxy Hedgehogs (Sriram, Laura Cremer, Gavin, Adam, Arshna, Connor)
4. The Outliers (Ruke, Ashlei, Brianna, Grace, Jon, Pavel)
5. Pearson Project (Albar, Abhishek, C.W., M.K., Sandra)
6. The Ridiculous Six (Gwen, Sarah Frischmann, Ryan, Nik, Roberto, Elina)
7. Shakalaka (Todd, Dongze, Dannielle, Ruipeng, Xueyi)
8. Super 6 (Sophia, Vishali, J.J., Preeti, Andrew Shan, Bilal)
9. The Two Keys (Imad, Jack, Neel, Kaylee, Andrew Tang, Peter)

The Four Parts of Project Task B

1. Develop and propose **two** research questions for Study 1.
2. Develop and propose 4 **quantitative** survey items for Study 1, each explicitly and directly linked to a research question.
3. Develop and propose 4 **categorical** (2-5 levels) survey items for Study 1, explicitly and directly linked to a research question.
4. Identify (from the published literature, the internet, or it can be one you've made up) an interesting *scale* related to one of your two Study 1 research questions that [a] includes no more than 10 survey items, and [b] is available for public use.

Details, resources at
https://github.com/THOMASELOVE/431project/tree/master/TaskB

15 Questions Dr. Love plans to include in the Survey

The following items will be included in the survey. As a result, you will not want to ask these questions in your Task B, although you should consider these groupings as candidates for application in your research questions.

These 8 items will be provided in groups after the application of cutpoints we will identify together after the survey is complete.

1. In what year were you born?
2. How would you rate your current health overall (Excellent, Very Good, Good, Fair, Poor)
3. For how long, in months, have you lived in Northeast Ohio?
4. What is your height in inches? (If you are five feet, eight inches tall, please write 68 inches. To convert from centimeters to inches, multiply your height in centimeters by 0.3937, and then round the result to the nearest inch.)
5. What is your weight in pounds? (To convert from kilograms to pounds, multiply your weight in kilograms by 2.2046, and then round the result to the nearest pound.)
6. What is your pulse rate, in beats per minute? (Please either use a tracking device, or count your pulse for 15 seconds then multiply by 4)
7. Last week, on how many days did you exercise? (0 - 7)
8. Last night, how many hours of sleep did you get?

The following 7 items will have yes/no responses, and thus produce binary groups for analysis.

1. Were you born in the United States?
2. Is English the language you speak better than any other?
3. Do you identify as female?
4. Do you wear prescription glasses or contact lenses?
5. Before taking 431, had you ever used R before?
6. Are you currently married or in a stable domestic relationship?
7. Have you smoked 100 cigarettes or more in your entire life?

Leek, Chapters 1-2

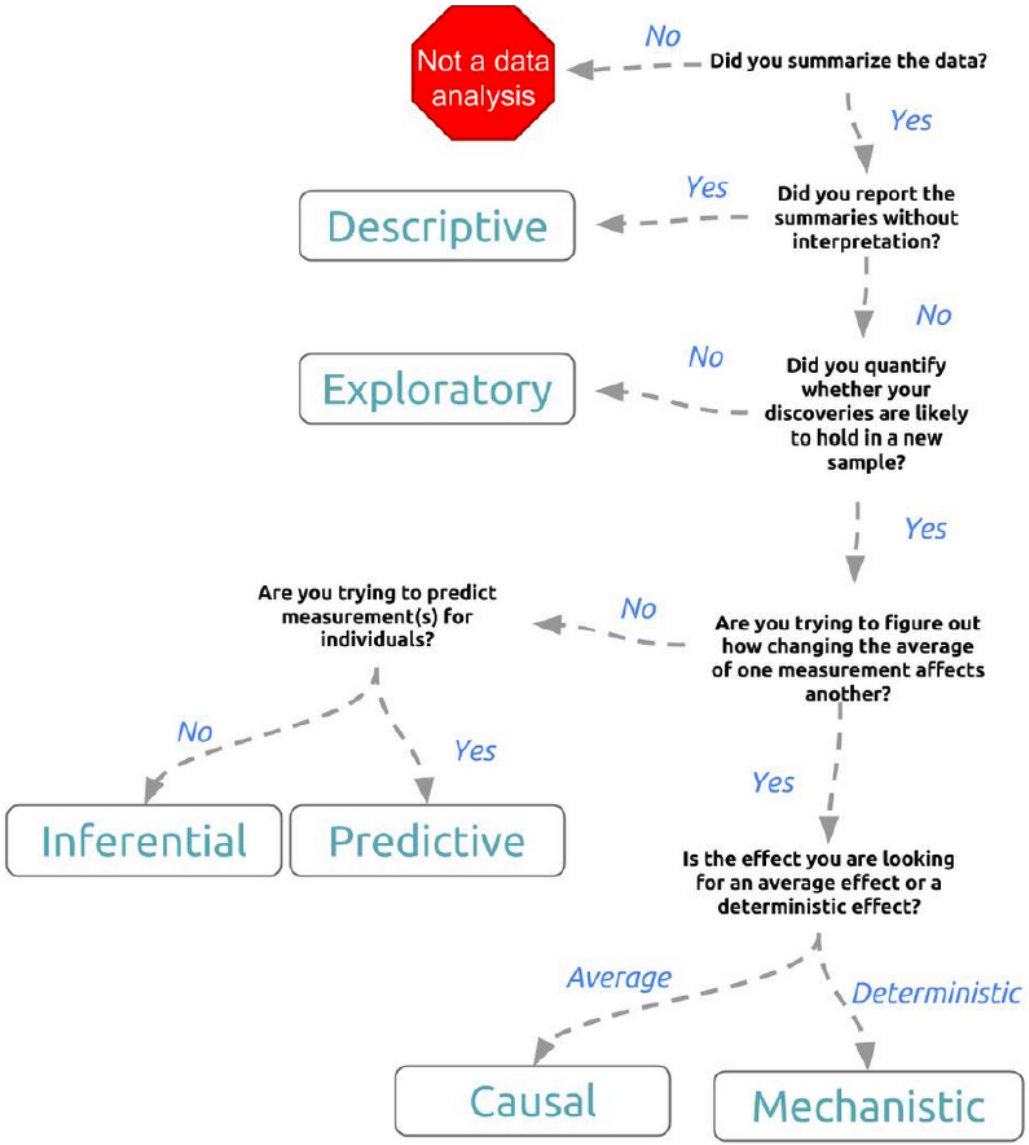
Chapter 1 Introduction

Chapter 2 The Data Analytic Question

See next slide.

Type	Strongest Coverage
Descriptive & Exploratory	Part A
Inferential	Part B
Predictive	Part C
Causal & Mechanistic	432

Leek, Chapter 2



Leek, Chapter 3 (Tidying the Data)

Components of a Processed Data Set

1. The raw data.
2. A tidy data set.
3. A code book describing each variable and its values in the tidy data set.
4. An explicit and exact recipe you used to go from 1 to 2 to 3.

See https://github.com/jtleek/datasharing for a guide for your project.

Tidy Data Video from Hadley Wickham https://vimeo.com/33727555

Leek, Chapter 4 (Checking the Data)

- Coding variables appropriately
    - Continuous, Ordinal, Categorical, Missing, Censored
- Code categorical / ordinal variables so that R will read them as factors.
- Encode everything using text, not with colors on the spreadsheet.
- Identify the missing value indicator, and use `NA` whenever you can.
- Check for coding errors, particularly label switching.

Leek, Chapter 12 (Reproducibility)

Reproducibility of workflow is what we're aiming for.

- Everything in a script. (R Markdown)
- Everything stored in a plain text file (future-proof: .csv, .Rmd)
- Organize your data analysis in subfolders of the project directory
- Use version control (something I should do more of)
- Add `sessionInfo()` command to final version of work when you need to preserve the details on software and parameters - see next slide.

My session info, at home, 2017-10-02

Include this information in your project submissions, but not probably in your other assignments, unless we ask you for it.

```
> sessionInfo()
R version 3.4.1 (2017-06-30)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows >= 8 x64 (build 9200)

Matrix products: default

locale:
[1] LC_COLLATE=English_United States.1252  LC_CTYPE=English_United States.1252    LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C                           LC_TIME=English_United States.1252

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] bindrcpp_0.2      dplyr_0.7.3       purrr_0.2.3       readr_1.1.1       tidyr_0.7.1       tibble_1.3.4      ggplot2_2.2.1
 [8] tidyverse_1.1.1   forcats_0.2.0     mice_2.30         GGally_1.3.2      broom_0.4.2       viridis_0.4.0     viridisLite_0.2.0
[15] Epi_2.19

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.12       lubridate_1.6.0    lattice_0.20-35    prettyunits_1.0.2  zoo_1.8-0          rprojroot_1.2      digest_0.6.12
 [8] assertthat_0.2.0   psych_1.7.8        R6_2.2.2           cellranger_1.1.0   plyr_1.8.4         backports_1.1.0    evaluate_0.10.1
[15] etm_0.6-2          highr_0.6          httr_1.3.1         progress_1.1.2     rlang_0.1.2        lazyeval_0.2.0     readxl_1.0.0
[22] rpart_4.1-11       Matrix_1.2-10      rmarkdown_1.6      labeling_0.3       splines_3.4.1      stringr_1.2.0      foreign_0.8-69
[29] munsell_0.4.3      compiler_3.4.1     numDeriv_2016.8-1  modelr_0.1.1       pkgconfig_2.0.1    mnormt_1.5-5       htmltools_0.3.6
[36] nnet_7.3-12        gridExtra_2.3      mosaicCore_0.4.0   reshape_0.8.7      MASS_7.3-47        grid_3.4.1         nlme_3.1-131
[43] mosaicData_0.14.0  jsonlite_1.5       gtable_0.2.0       ggformula_0.6      magrittr_1.5       scales_0.5.0       stringi_1.1.5
[50] reshape2_1.4.2     xml2_1.1.1         ggdendro_0.1-20    RColorBrewer_1.1-2 tools_3.4.1        cmprsk_2.2-7       glue_1.1.1
[57] hms_0.3            parallel_3.4.1     survival_2.41-3    yaml_2.1.14        colorspace_1.3-2   mosaic_1.1.0       rvest_0.3.2
[64] knitr_1.17         bindr_0.1          haven_1.1.0
```