431 Class 12

Thomas E. Love

2018-10-04

Today's Agenda

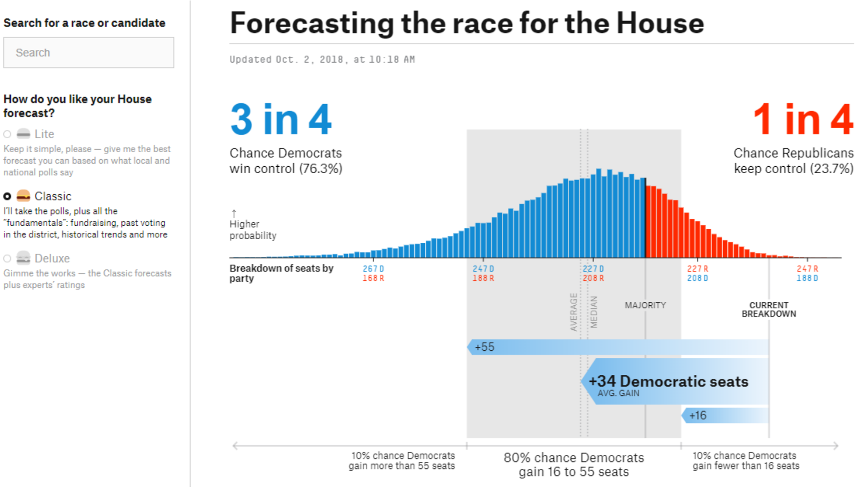
1. The FiveThirtyEight Election Forecast: U.S. House of Representatives
2. Some Thoughts on `dplyr` and its verbs
3. The Printer Case Study
4. Project Task A (the Group piece)
5. Setting up the first Quiz

FiveThirtyEight.com

Nate Silver, Founder and Editor in Chief

https://projects.fivethirtyeight.com/2018-midterm-election-forecast/house/

Status as of 10:30 AM on 2018-10-02



Search for a race or candidate

Search

How do you like your House forecast?

○ Lite
Keep it simple, please — give me the best forecast you can based on what local and national polls say

● Classic
I'll take the polls, plus all the "fundamentals": fundraising, past voting in the district, historical trends and more

○ Deluxe
Gimme the works — the Classic forecasts plus experts' ratings

Forecasting the race for the House

Updated Oct. 2, 2018, at 10:18 AM

3 in 4
Chance Democrats win control (76.3%)

1 in 4
Chance Republicans keep control (23.7%)

10% chance Democrats gain more than 55 seats

80% chance Democrats gain 16 to 55 seats

10% chance Democrats gain fewer than 16 seats

Today's R Starting Point

```
library(tidyverse)
```

Western Collaborative Group Study (wcgs)

See Notes, Chapter 13.

- Full data set has 3,154 observations on 22 variables.

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

Using the key verbs in `dplyr`

dplyr basics: The Key Verbs

Six key functions:

- Pick observations by their values (`filter()`).
- Reorder the rows (`arrange()`).
- Pick variables by their names (`select()`).
- Collapse many values down to a single summary (`summarise()`).
- Create new variables with functions of existing variables (`mutate()`).
- Change the scope of another function from operating on the whole data set to operating on it group-by-group (`group_by()`)

All of this comes from the Explore section of Grolemund and Wickham's R for Data Science, in particular the material on Data transformation.

http://r4ds.had.co.nz/transform.html

dplyr basics: How the verbs work

- The first argument is a data frame (or tibble).
- The second arguments describe what to do with the data frame. You can refer to columns in the data frame directly without using $.
- The result is a new data frame.

Filter rows with filter()

filter() allows you to subset observations based on their values.

```
wcgs.sub1 <- wcgs %>%
  filter(dibpat == "Type A" & age > 49)
wcgs.sub1
```

```
# A tibble: 522 x 22
     id   age agec  height weight lnwght wghtcat   bmi
  <int> <int> <fct>  <int>  <int>  <dbl> <fct>   <dbl>
1  2343    50 46-50     67    200   5.30 170-200  31.3
2  3656    51 51-55     73    192   5.26 170-200  25.3
3  3526    59 56-60     70    200   5.30 170-200  28.7
4 22057    51 51-55     69    150   5.01 140-170  22.1
5 12681    50 46-50     71    195   5.27 170-200  27.2
6  3284    59 56-60     72    206   5.33 > 200    27.9
7 21071    54 51-55     67    152   5.02 140-170  23.8
8 13371    55 51-55     72    185   5.22 170-200  25.1
```

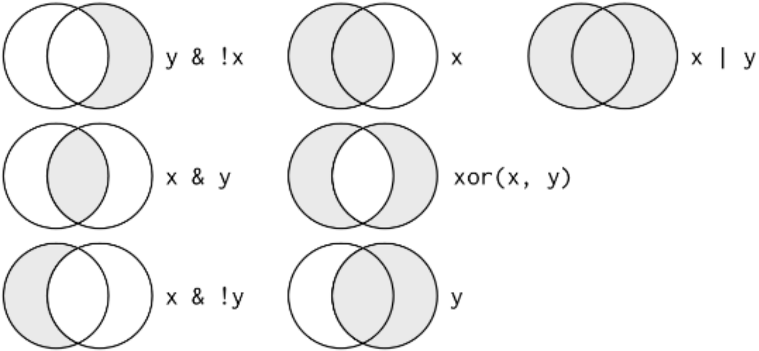
Comparison and Logical Operators

Comparison Operator	Meaning
>	is greater than
>=	is greater than or equal to
<	is less than
<=	is less than or equal to
!=	is not equal to
==	is equal to

Logical (Boolean) Operator	Meaning
&	and
|	or
!	not

Missing Values (`NA` in R) can make things tricky. They are contagious. Almost any operation involving an unknown value will also be unknown.

The complete set of Boolean Operators



Source: http://r4ds.had.co.nz/transform.html#logical-operators

Arrange rows with `arrange()`

`arrange()`, instead of selecting rows (like `filter()`), changes their order.

- Use `arrange(height)` to arrange in ascending order of height. Provide a second column name to break ties, if you like.
- Missing values are always sorted at the end.

```
wcgs %>%
  arrange(desc(height), desc(weight))
```

```
# A tibble: 3,154 x 22
      id   age agec  height weight lnwght wghtcat   bmi
   <int> <int> <fct>  <int>  <int>  <dbl> <fct>   <dbl>
 1 12012    47 46-50     78    250   5.52 > 200    28.9
 2  2145    41 41-45     78    220   5.39 > 200    25.4
 3 12680    43 41-45     78    190   5.25 170-200  22.0
 4 13512    42 41-45     77    220   5.39 > 200    26.1
 5 12620    49 46-50     77    210   5.35 > 200    24.9
 6 11209    45 41-45     77    195   5.27 170-200  23.1
```

Select columns with `select()`

`select()` lets you zoom in on the columns you actually want to use based on the names of the variables. R for Data Science lays out some helper functions within select() for use in bigger data sets.

```
wcgs.sub2 <- wcgs %>%
  select(id, age, height, weight, dibpat, smoke, behpat)
wcgs.sub2
```

```
# A tibble: 3,154 x 7
      id   age height weight dibpat smoke behpat
   <int> <int>  <int>  <int> <fct>  <fct> <fct>
 1  2343    50     67    200 Type A Yes   A1
 2  3656    51     73    192 Type A Yes   A1
 3  3526    59     70    200 Type A No    A1
 4 22057    51     69    150 Type A No    A1
 5 12927    44     71    160 Type A No    A1
 6 16029    47     64    158 Type A Yes   A1
```

Grouped summaries with `summarize()`

`summarise()` or `summarize()` collapses a data frame to a single row.

```r
wcgs.sub2 %>%
  summarize(mean.ht = mean(height, na.rm=TRUE),
            sd.ht = sd(height, na.rm=TRUE)) %>%
  round(digits = 2)
```

```
# A tibble: 1 x 2
  mean.ht sd.ht
    <dbl> <dbl>
1    69.8  2.53
```

Using the pipe (%>%) to filter and summarize

```r
wcgs.sub2 %>%
 filter(dibpat == "Type A") %>%
 summarize(pearson.r = cor(height, weight),
  spearman.r = cor(height, weight, method = "spearman")) %>%
 round(digits = 3) %>%
 knitr::kable()
```

pearson.r	spearman.r
0.534	0.542

Using group_by() with summarize to look group-by-group

```
wcgs.sub2 %>%
  group_by(behpat) %>%
  summarize(
    pearson.r = round(cor(height, weight),3) ) %>%
  knitr::kable()
```

behpat	pearson.r
A1	0.571
A2	0.526
B3	0.524
B4	0.557

Using group_by() to look at separated groups

You might have tried this approach instead, but it throws an error...

```
wcgs.sub2 %>%
  group_by(behpat) %>%
  summarize(
    pearson.r = cor(height, weight)) %>%
  round(digits = 3) %>%
  knitr::kable()
```

- Why doesn't this work?

Using group_by() to look at separated groups

You might have tried this approach instead, but it throws an error…

```r
wcgs.sub2 %>%
  group_by(behpat) %>%
  summarize(
    pearson.r = cor(height, weight)) %>%
  round(digits = 3) %>%
  knitr::kable()
```

- Why doesn't this work?
- When R sees the round command, it tries to apply it to every element of the table, including the behavior pattern labels, which aren't numbers. So it throws an error.

Add new variables with mutate()

mutate() adds new columns that are functions of existing columns to the end of your data set.

Suppose we want to calculate the weight/height ratio for each subject.

```
wcgs.sub3 <- wcgs.sub2 %>%
    select(id, weight, height) %>%
    mutate(wh.ratio = weight / height)
wcgs.sub3
```

```
# A tibble: 3,154 x 4
      id weight height wh.ratio
   <int>  <int>  <int>    <dbl>
 1  2343    200     67     2.99
 2  3656    192     73     2.63
 3  3526    200     70     2.86
 4 22057    150     69     2.17
 5 12927    160     71     2.25
```

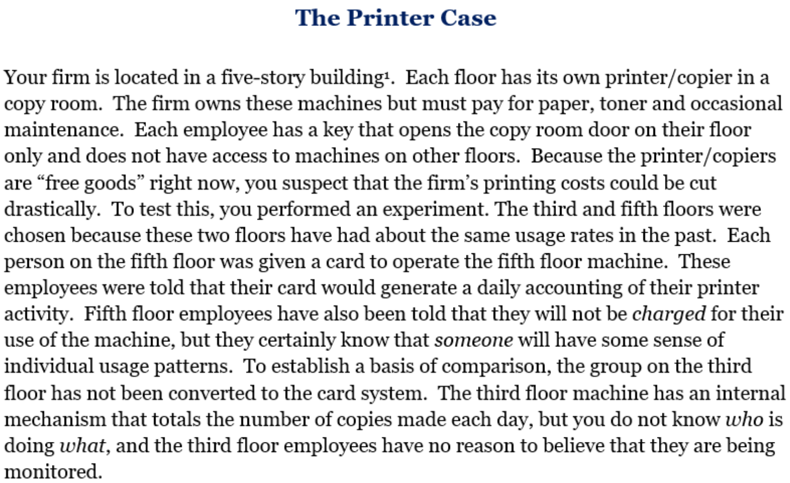
On Coding and dplyr

1. Learn `dplyr`, and use it to do most of your data management within R.
    - `dplyr` is mostly about these key verbs, and piping, for our purposes
    - some tasks produce results which be confusing, we're here to help
2. `dplyr` is most useful in combination with other elements of the `tidyverse`, most prominently `ggplot2`.

The Printer Case Study

The Printer Case, Setup

Get with your Project group.



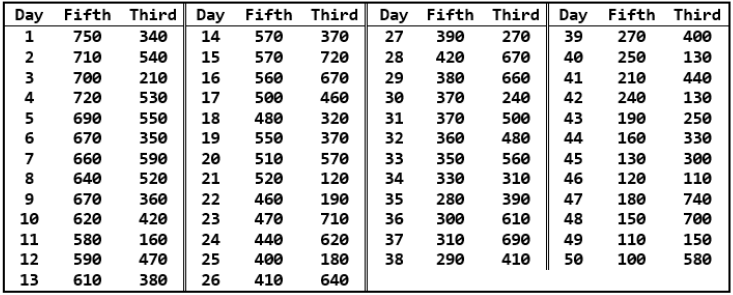
The Printer Case

Your firm is located in a five-story building[1]. Each floor has its own printer/copier in a copy room. The firm owns these machines but must pay for paper, toner and occasional maintenance. Each employee has a key that opens the copy room door on their floor only and does not have access to machines on other floors. Because the printer/copiers are “free goods” right now, you suspect that the firm’s printing costs could be cut drastically. To test this, you performed an experiment. The third and fifth floors were chosen because these two floors have had about the same usage rates in the past. Each person on the fifth floor was given a card to operate the fifth floor machine. These employees were told that their card would generate a daily accounting of their printer activity. Fifth floor employees have also been told that they will not be *charged* for their use of the machine, but they certainly know that *someone* will have some sense of individual usage patterns. To establish a basis of comparison, the group on the third floor has not been converted to the card system. The third floor machine has an internal mechanism that totals the number of copies made each day, but you do not know *who* is doing *what*, and the third floor employees have no reason to believe that they are being monitored.

The Printer Case, Main Table

You collected data from the machines over the last 50 working days. The data are in the table below and can be downloaded from the web in the **printer.csv** file. There are three variables: DAY, which indicates the day; FIFTH, the number of copies made on the 5th floor; and THIRD, the number made on the 3rd floor.

Will the card accounting system effectively lower usage if implemented across the firm?



Day	Fifth	Third	Day	Fifth	Third	Day	Fifth	Third	Day	Fifth	Third
1	750	340	14	570	370	27	390	270	39	270	400
2	710	540	15	570	720	28	420	670	40	250	130
3	700	210	16	560	670	29	380	660	41	210	440
4	720	530	17	500	460	30	370	240	42	240	130
5	690	550	18	480	320	31	370	500	43	190	250
6	670	350	19	550	370	32	360	480	44	160	330
7	660	590	20	510	570	33	350	560	45	130	300
8	640	520	21	520	120	34	330	310	46	120	110
9	670	360	22	460	190	35	280	390	47	180	740
10	620	420	23	470	710	36	300	610	48	150	700
11	580	160	24	440	620	37	310	690	49	110	150
12	590	470	25	400	180	38	290	410	50	100	580
13	610	380	26	410	640						

The Printer Case Discussion, Part 1

Fifty days of data. Fifth floor employees were given a card to operate their printer. Third floor employees were not.

1. Is this a randomized trial or an observational study?
2. What is the outcome we are studying?
3. What are the two treatments/exposures/interventions being compared?
4. What controls are in place as part of the study's design?
5. **Key Question**: Will the card accounting system effectively lower usage if implemented across the firm?

The Printer Case Discussion

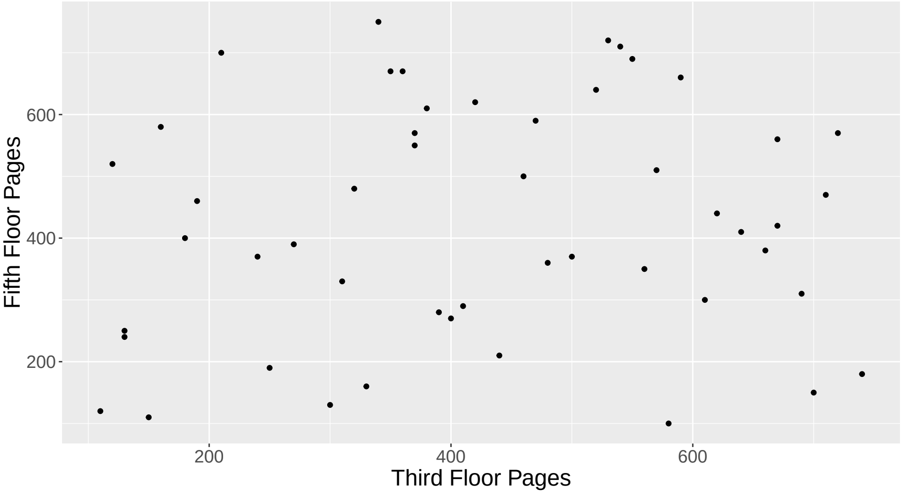
Go.

Printer Case: Numerical Summary

```r
printer <- read.csv("printer.csv") %>% tbl_df
summary(printer)
```

```
      Day            Fifth           Third      
 Min.   : 1.00   Min.   :100.0   Min.   :110.0  
 1st Qu.:13.25   1st Qu.:282.5   1st Qu.:302.5  
 Median :25.50   Median :415.0   Median :415.0  
 Mean   :25.50   Mean   :426.2   Mean   :428.2  
 3rd Qu.:37.75   3rd Qu.:577.5   3rd Qu.:577.5  
 Max.   :50.00   Max.   :750.0   Max.   :740.0  
```

Printer Case: Scatterplot (r = 0.11)



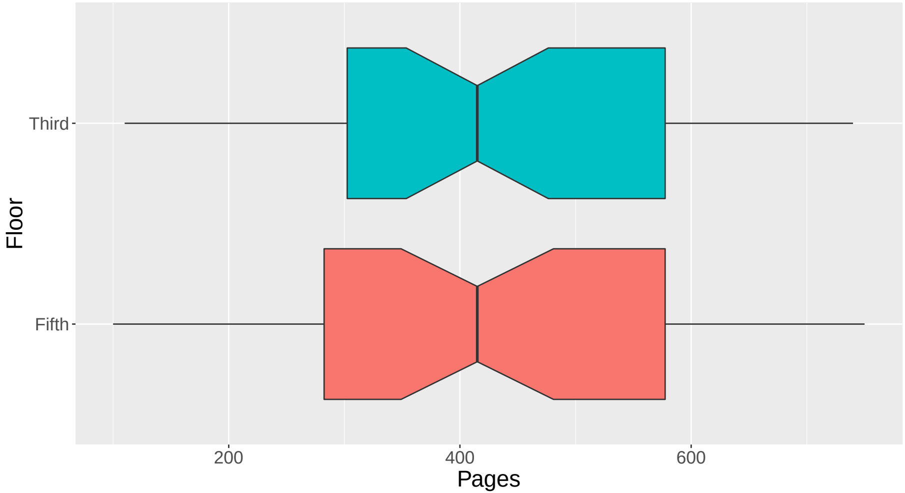
Printer Case: Gather the Columns

First, we'll gather up the data so that we can plot it more easily.

```
printer2 <- tidyr::gather(printer, Floor, Pages, -Day)
printer2
```

```
# A tibble: 100 x 3
     Day Floor Pages
   <int> <chr> <int>
 1     1 Fifth   750
 2     2 Fifth   710
 3     3 Fifth   700
 4     4 Fifth   720
 5     5 Fifth   690
 6     6 Fifth   670
 7     7 Fifth   660
 8     8 Fifth   640
 9     9 Fifth   670
```

Printer Case: Comparison Boxplot



Numerical Summary comparing the Two Floors

```
mosaic::favstats(Pages ~ Floor, data = printer2)
```

```
  Floor min    Q1 median    Q3 max  mean       sd  n
1 Fifth 100 282.5    415 577.5 750 426.2 188.8298 50
2 Third 110 302.5    415 577.5 740 428.2 186.0841 50
  missing
1       0
2       0
```

Statistical Inference comparing the Means?

```r
t.test(Pages ~ Floor, data = printer2)
```

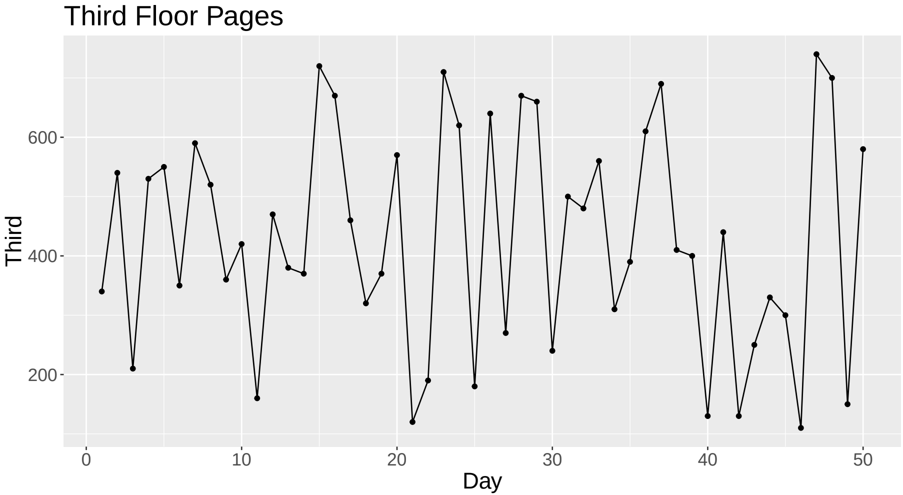
```
Welch Two Sample t-test (Edited Output)

data:  Pages by Floor
sample estimates:
mean in group Fifth mean in group Third
              426.2               428.2

alternative hypothesis:
true difference in means is not equal to 0

t = -0.053344, df = 97.979, p-value = 0.9576
95 percent confidence interval: -76.40  72.40
```

Statistical Inference comparing the Means?

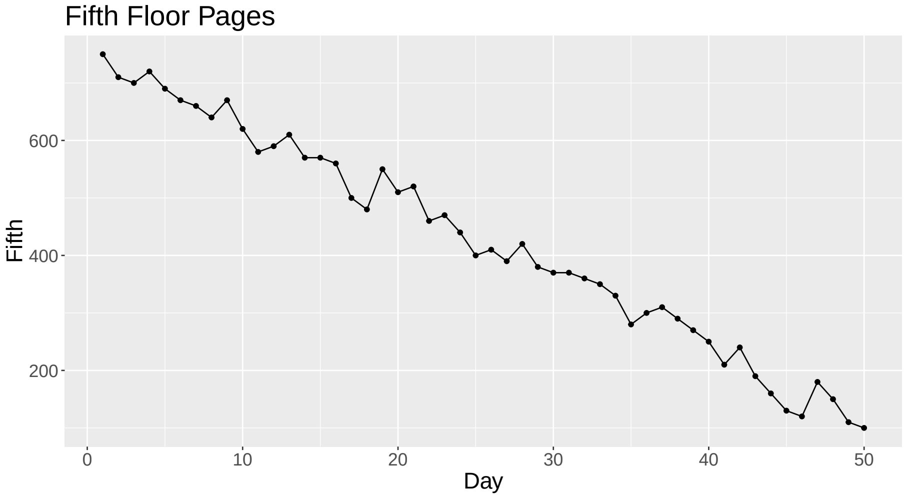
```r
t.test(Pages ~ Floor, data = printer2,
       conf.level = 0.90, var.equal = TRUE)
```

```
Two Sample t-test (Edited output)

data:  Pages by Floor
sample estimates:
mean in group Fifth mean in group Third
              426.2               428.2

alternative hypothesis:
true difference in means is not equal to 0

t = -0.053344, df = 98, p-value = 0.9576
90 percent confidence interval: -64.258  60.258
```

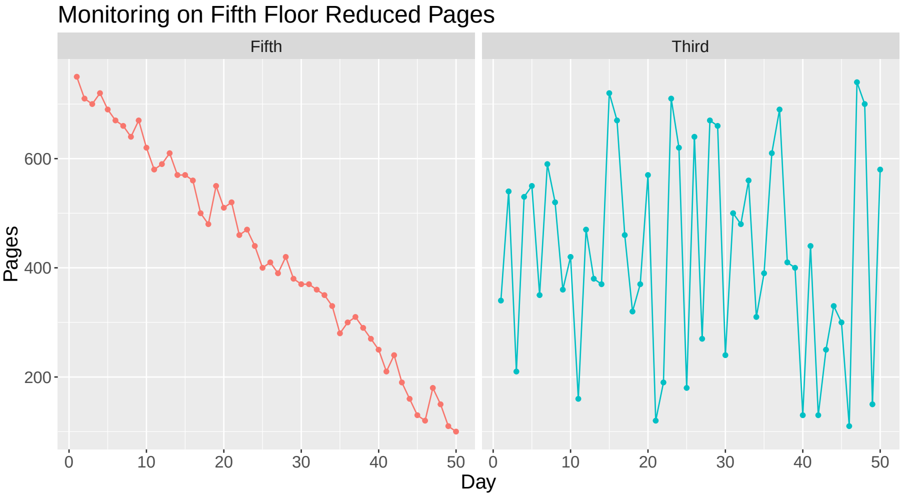
Printer Case: Third Floor



Printer Case: Fifth Floor



Comparing the Patterns over Time



Project Task A Discussion Time

Task A requires three things of your group

1. Develop and propose 2-3 research questions for Study 1
2. Propose 6-10 "homemade" survey questions for Study 1
3. Propose a "scale" for Study 1 that relates to (at least one of) your research questions.

Complete details at https://thomaselove.github.io/431-2018-project/

- Sample research questions at the Class 11 README and Class 12 README page.
- Old surveys available at https://github.com/THOMASELOVE/431-2018-project/tree/master/oldsurveys
- Examples of "scales" at the Class 11 README and Class 12 README page, as well.

Don't forget that the rest of Task A (related to Study 2) is for you to do, by yourself.

Setting Up Quiz 1

Setting Up Quiz 1

There are a total of 40 questions, each worth 2, 2.5 or 3 points, plus an affirmation that your work is yours alone.

- Please select or type in your best response for each question. The questions are not arranged in any particular order, and you should answer all of them.
- You must complete this quiz by 7 PM Monday. You will have the opportunity to edit your responses after completing the quiz, but this must be completed by the deadline.
- If you wish to complete part of the quiz and then return to it later, please scroll down to the end of the quiz and complete the **affirmation** (final question). The affirmation is required, and you will have to complete it in order to exit the quiz and save your progress. You will then be presented with (and emailed) a link to "Edit your progress" which you will want to bookmark, so you can return to it easily.

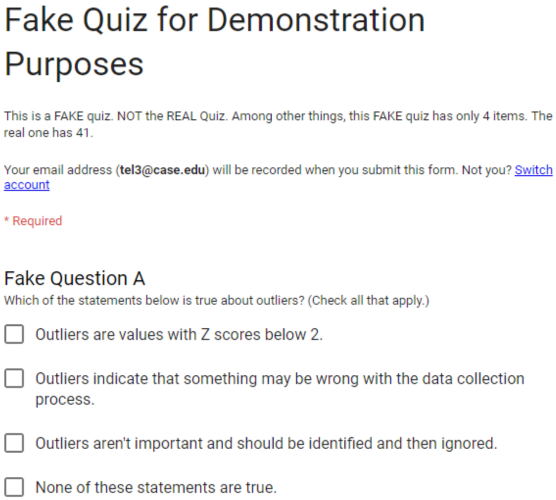
Quiz 1: Main item types.

Fake Quiz is at https://goo.gl/forms/hw37w3BrpibPDGQ03

1. Short Answer Questions
2. Multiple Choice
3. Checkboxes
4. Matching

- You are welcome to consult the materials provided on the course website, but you are not allowed to discuss the questions on this quiz with anyone other than Professor Love or the Teaching Assistants, who may be reached at `431-help at case dot edu`.

Fake Quiz: Question A



Fake Quiz for Demonstration Purposes

This is a FAKE quiz. NOT the REAL Quiz. Among other things, this FAKE quiz has only 4 items. The real one has 41.

Your email address (**tel3@case.edu**) will be recorded when you submit this form. Not you? Switch account

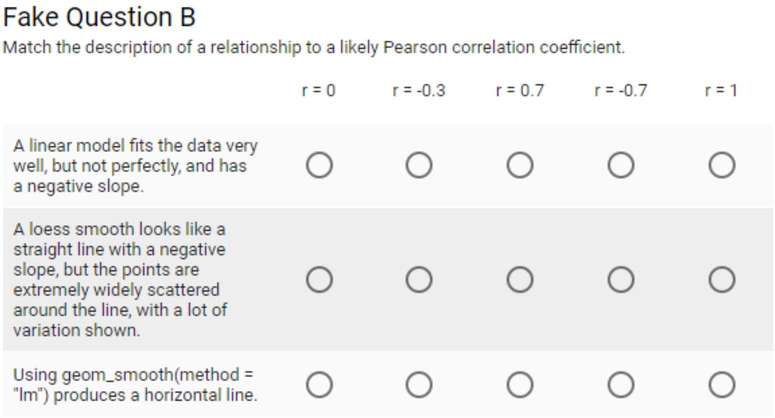
* Required

Fake Question A

Which of the statements below is true about outliers? (Check all that apply.)

- □ Outliers are values with Z scores below 2.
- □ Outliers indicate that something may be wrong with the data collection process.
- □ Outliers aren't important and should be identified and then ignored.
- □ None of these statements are true.

Fake Quiz: Question B

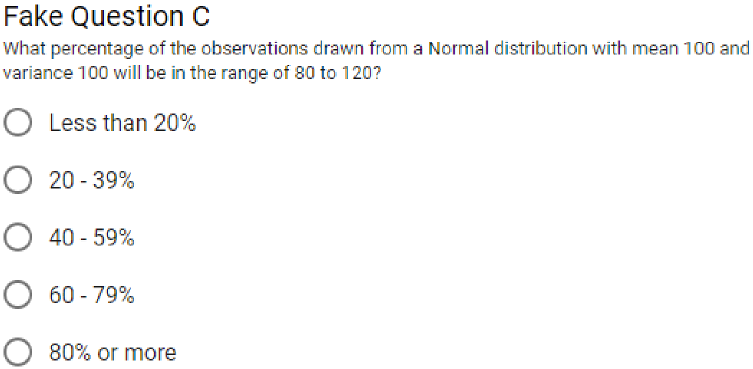


Fake Question B

Match the description of a relationship to a likely Pearson correlation coefficient.

	r = 0	r = -0.3	r = 0.7	r = -0.7	r = 1
A linear model fits the data very well, but not perfectly, and has a negative slope.	○	○	○	○	○
A loess smooth looks like a straight line with a negative slope, but the points are extremely widely scattered around the line, with a lot of variation shown.	○	○	○	○	○
Using geom_smooth(method = "lm") produces a horizontal line.	○	○	○	○	○

Fake Quiz: Question C



Fake Question C

What percentage of the observations drawn from a Normal distribution with mean 100 and variance 100 will be in the range of 80 to 120?

○ Less than 20%

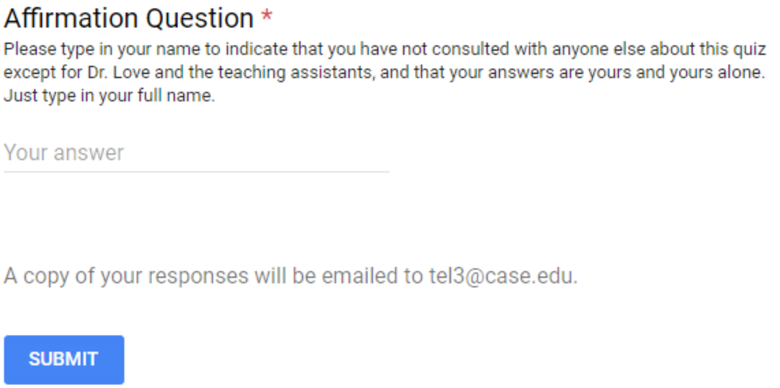
○ 20 - 39%

○ 40 - 59%

○ 60 - 79%

○ 80% or more

Fake Quiz: Affirmation



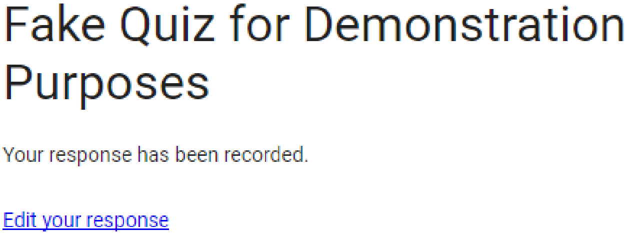
Affirmation Question *

Please type in your name to indicate that you have not consulted with anyone else about this quiz except for Dr. Love and the teaching assistants, and that your answers are yours and yours alone. Just type in your full name.

Your answer

A copy of your responses will be emailed to tel3@case.edu.

Fake Quiz: Completion



Fake Quiz for Demonstration Purposes

Your response has been recorded.

Edit your response

Link to the Quiz

will be provided by noon Friday 2018-10-05. The link will appear:

1. On the Class 12 README page, and
2. On the Quizzes page.