

# 431 Class 12

Thomas E. Love

2017-10-05

# Today's Agenda

- ① Leek Chapters 1-4 and 12
- ② Some Thoughts on `dplyr` and its verbs
- ③ The Printer Case Study
- ④ Setting up the first Quiz

# Leek, Chapters 1-2

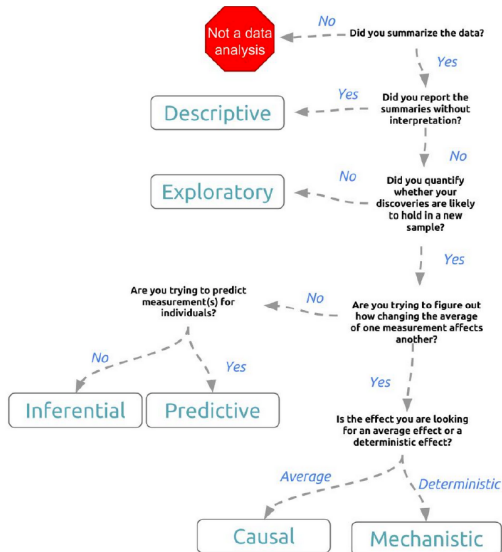
## Chapter 1 Introduction

## 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, 2017-10-03

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

```
> sessionInfo()
R version 3.4.2 (2017-09-28)
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_0.7.4      purrr_0.2.3      readr_1.1.1      tidyr_0.7.1      tibble_1.3.4      ggplot2_2.2.1    tidyverse_1.1.1
[9] GGally_1.3.2

loaded via a namespace (and not attached):
 [1] progress_1.1.2      reshape2_1.4.2      haven_1.1.0          lattice_0.20-35      colorspace_1.3-2     htmltools_0.3.6      yaml_2.1.14
 [8] rlang_0.1.2         foreign_0.8-69      glue_1.1.1           RColorBrewer_1.1-2  modelr_0.1.1         readxl_1.0.0         bindr_0.1
[15] plyr_1.8.4          stringr_1.2.0       munsell_0.4.3        gtable_0.2.0        cellranger_1.1.0     rvest_0.3.2         psych_1.7.8
[22] evaluate_0.10.1     labeling_0.3        knitr_1.17           forcats_0.2.0       parallel_3.4.2       highr_0.6            broom_0.4.2
[29] Rcpp_0.12.13        scales_0.5.0        backports_1.1.1      jsonlite_1.5         mnormt_1.5-5         hms_0.3             digest_0.6.12
[36] stringi_1.1.5       grid_3.4.2          rprojroot_1.2        tools_3.4.2         magrittr_1.5         lazyeval_0.2.0      pkgconfig_2.0.1
[43] prettyunits_1.0.2   xml2_1.1.1          lubridate_1.6.0      assertthat_0.2.0    rmarkdown_1.6        reshape_0.8.7       httr_1.3.1
[50] R6_2.2.2            nlme_3.1-131        compiler_3.4.2
```



# Today's R Setup

```
library(mice); library(tidyverse)  
  
source("Love-boost.R")
```

# 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 Wickham and Grolemund, R for Data Science, Chapter 5*

<http://r4ds.had.co.nz/transform.html#introduction-2>

# 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.

We'll work with the `wcgs` data.

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

```
# A tibble: 3,154 x 22
```

	id	age	agec	height	weight	lnwght	wghtcat
	<int>	<int>	<fctr>	<int>	<int>	<dbl>	<fctr>
1	2343	50	46-50	67	200	5.298317	170-200
2	3656	51	51-55	73	192	5.257495	170-200
3	3526	59	56-60	70	200	5.298317	170-200
4	22057	51	51-55	69	150	5.010635	140-170
5	12227	44	41-45	71	160	5.075174	140-170

# 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
	<int>	<int>	<fctr>	<int>	<int>	<dbl>	<fctr>
1	2343	50	46-50	67	200	5.298317	170-200
2	3656	51	51-55	73	192	5.257495	170-200
3	3526	59	56-60	70	200	5.298317	170-200
4	22057	51	51-55	69	150	5.010635	140-170
5	12681	50	46-50	71	195	5.273000	170-200
6	3284	59	56-60	72	206	5.327876	> 200
7	21071	54	51-55	67	152	5.023880	140-170
8	13371	55	51-55	72	185	5.220356	170-200

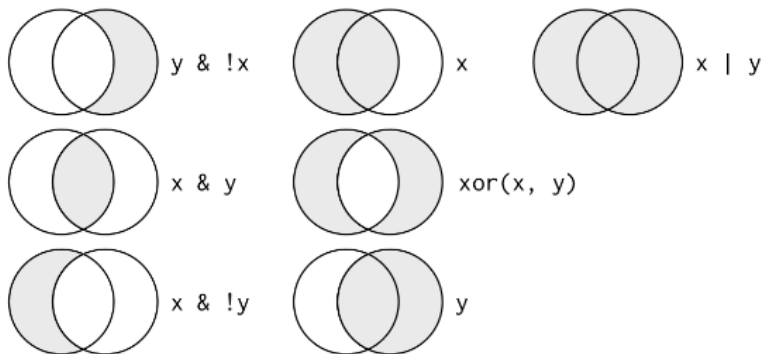
# Comparison and Logical Operators

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
	<int>	<int>	<fctr>	<int>	<int>	<dbl>	<fctr>
1	12012	47	46-50	78	250	5.521461	> 200
2	2145	41	41-45	78	220	5.393628	> 200
3	12680	43	41-45	78	190	5.247024	170-200
4	13512	42	41-45	77	220	5.393628	> 200
5	12620	49	46-50	77	210	5.347107	> 200
6	11209	45	41-45	77	195	5.273000	170-200

# 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	behp
	<int>	<int>	<int>	<int>	<fctr>	<fctr>	<fctr>
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.

```
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   69.78   2.53
```

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

```
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...

```
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 %>%  
  mutate(wh.ratio = weight / height)  
wcgs.sub3
```

```
# A tibble: 3,154 x 8
```

	id	age	height	weight	dibpat	smoke	behpat
	<int>	<int>	<int>	<int>	<fctr>	<fctr>	<fctr>
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

# 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.
- ③ Hmisc doesn't play nicely with dplyr, so don't load the whole Hmisc library, just call individual functions you need with, for example, `Hmisc::describe` or `Hmisc::smean.cl.boot`

# The Printer Case, Setup

## The Printer Case

Your firm is located in a five-story building<sup>1</sup>. 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 5<sup>th</sup> floor; and **THIRD**, the number made on the 3<sup>rd</sup> 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

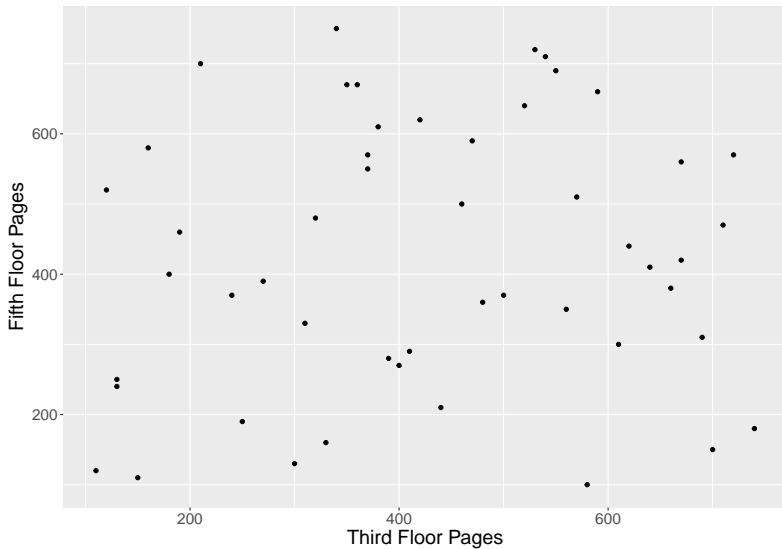
Go.

# Printer Case: Numerical Summary

```
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 ( $r = 0.11$ )



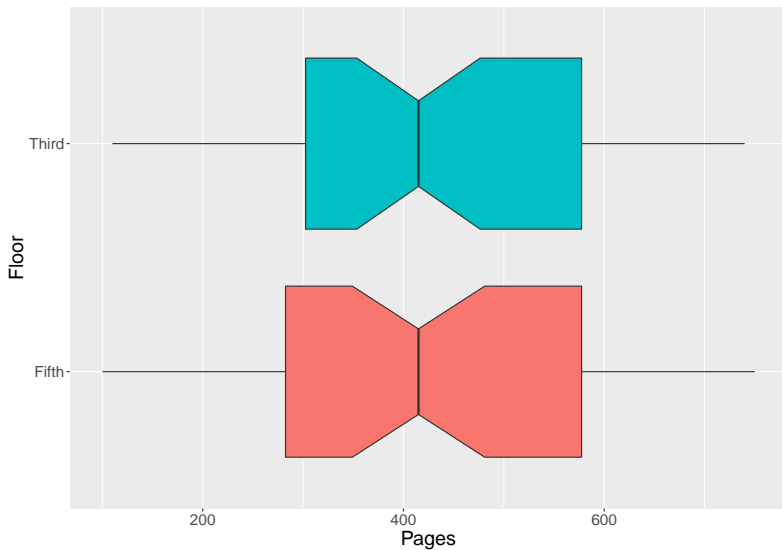
# Printer Case: Gather the Columns

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

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

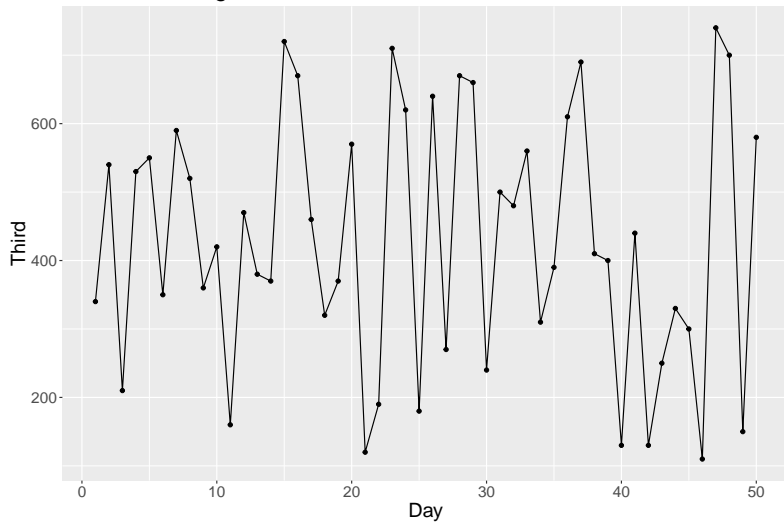
```
# A tibble: 100 x 3
   Day Floor Pages
<int> <chr> <int>
1     1 1 Fifth    750
2     2 2 Fifth    710
3     3 3 Fifth    700
4     4 4 Fifth    720
5     5 5 Fifth    690
6     6 6 Fifth    670
7     7 7 Fifth    660
8     8 8 Fifth    640
9     9 9 Fifth    670
```

# Printer Case: Comparison Boxplot



# Printer Case: Third Floor

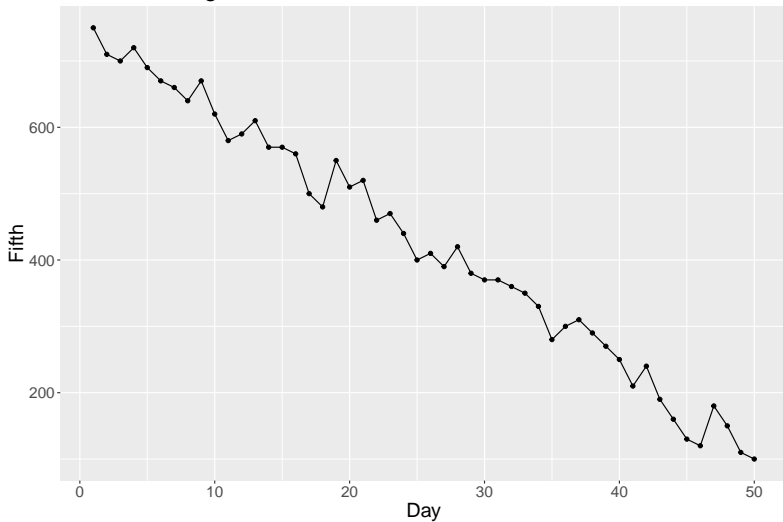
Third Floor Pages





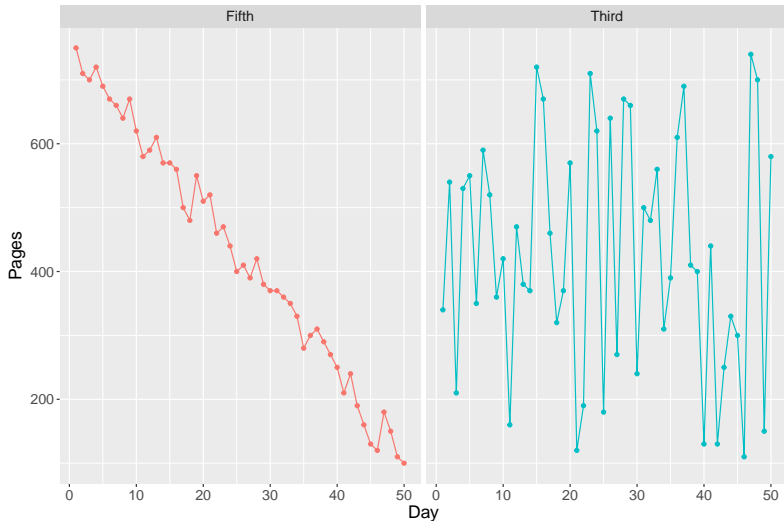
# Printer Case: Fifth Floor

Fifth Floor Pages



# Comparing the Patterns over Time

Monitoring on Fifth Floor Reduced Pages



# Project Task A, due 2017-10-13 at Noon

I must read lots of these + REDOs - make it painless for me.

- 1 Make one submission, via Canvas, as requested. I'll review it then, and get back to you. I cannot review your materials in advance, or I'd go insane. I process this work in batches.
- 2 The only sorts of questions I'm going to answer in advance are highly specific ones about the instructions for the Task. If you want clarification of what I'm looking for, no problem. If you want us to evaluate your work, you'll have to wait.
- 3 If your response to Task A is more than three pages long, cut it to three pages. For the proposal summary, I specified a word limit. Obey this in your first submission. One page is more than enough for the proposal in *every* case. Some data descriptions can easily be done in a page, as well, but some can't. Try to keep that to two pages. One is better if you answer my questions.
- 4 Don't make me search for things - label what you're doing using my labels from the instructions.

# Setting Up Quiz 1

There are a total of 41 questions, 18 worth 2 points, 18 worth 3 points, 4 worth 2.5 points, and 1 that affirms 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 Noon on Monday, 2016-10-09. 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** (Question 41). 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 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.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
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.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Using <code>geom_smooth(method = "lm")</code> produces a horizontal line.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

# 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.

**SUBMIT**

## Fake Quiz for Demonstration Purposes

Your response has been recorded.

[Edit your response](#)

# Link to the Quiz

will be provided by 3 PM Thursday 2017-10-05.