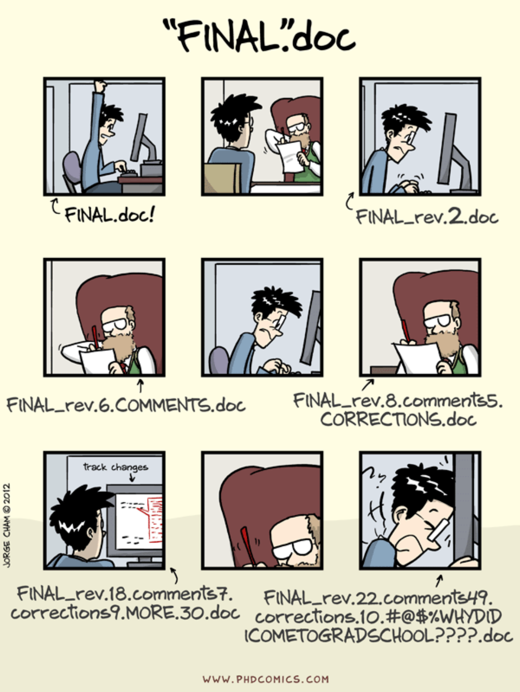
431 Class 05

github.com/THOMASELOVE/2019-431

2019-09-10

Today's Agenda

1. Course Project Instructions
   - We'll discuss further Thursday after you've had the chance to read them.
2. NHANES Example
   - See the related example in the Course Notes Chapters 3-6
3. Discussion of Jeff Leek's *Elements of Data Analytic Style*


"FINAL".doc
FINAL.doc!
FINAL_rev.2.doc
FINAL_rev.6.COMMENTS.doc
FINAL_rev.8.comments5.
CORRECTIONS.doc
track changes
FINAL_rev.18.comments7.
corrections9.MORE.30.doc
FINAL_rev.22.comments49.
corrections.10.#@$%WHYDID
ICOMETOGRADSCHOOL????.doc
JORGE CHAM © 2012
WWW.PHDCOMICS.COM

Today's Packages

The R packages we're using today are `NHANES`, `magrittr`, `janitor` and `tidyverse`.

```
library(NHANES); library(magrittr)
library(janitor); library(tidyverse)
```

I always load the `tidyverse` last.

- Also, I set the code chunk to `message = FALSE` when I want to hide several messages that come up when loading.

So my package loading code chunk header (inside the brackets) looks like:

`{r load_packages, message = FALSE}`

CWRU Colors

CWRU's color guide (see the README) specifies CWRU blue and CWRU gray

```
cwru.blue <- '#0a304e'
cwru.gray <- '#626262'
```

I'd like to use those later today.

Today's Example

We're going to work with subjects who participated in NHANES: National Health and Nutrition Examination Survey.

> *The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations.*

Use `?NHANES` to learn more about the data. The NHANES package contains 5000 observations from each of the 2009-10 and 2011-12 administrations.

- See the Course Notes, Chapters 3-6, for a related series of examples.
- Baumer, Kaplan and Horton (2017) *Modern Data Science with R* have developed similar examples.

A First Sample of NHANES data

To begin, we'll gather a random sample of 1,000 subjects participating in NHANES, and then select three variables of interest about those subjects.

```
set.seed(20190910)
# use set.seed to ensure that we all get the same random
# sample of 1,000 NHANES subjects in our nh1 data set

nh1 <- sample_n(NHANES, size = 1000) %>%
    select(ID, Age, Height)
```

The `sample_n` function, from the `dplyr` package

- `sample_n()` samples a fixed number of observations
- `sample_frac()` samples a fixed fraction of observations

can sample with or without replacement (default = without)

The `nh1` tibble

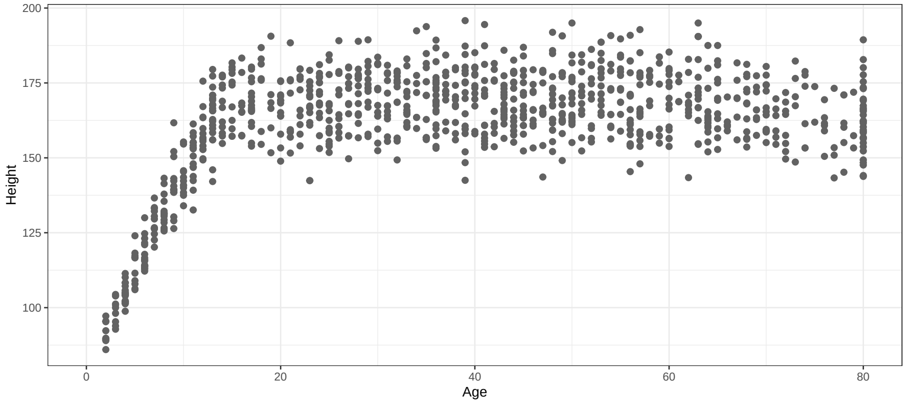
What are the units here?

```
# A tibble: 1,000 x 3
      ID   Age Height
   <int> <int>  <dbl>
 1 51781    29   174.
 2 53197    24   168.
 3 64940    13   180.
 4 59833    62   168.
 5 64485    32   177.
 6 61693    33   166.
 7 59005    20   149.
 8 61964    80   164.
 9 68222    36   165.
10 59741    60   175.
# ... with 990 more rows
```

Relationship of Height and Age - First Attempt

```
ggplot(data = nh1, mapping = aes(x = Age, y = Height)) +
    geom_point(size = 2, col = cwru.gray) + theme_bw()
```

```
Warning: Removed 31 rows containing missing values
(geom_point).
```



Interesting Results from Our First Attempt

1. Only 969 subjects are plotted, because the remaining 31 people have missing (NA) values for either Height, Age or both.
2. Unsurprisingly, the measured Heights of subjects grow from Age 0 to Age 20 or so, and we see that a typical Height increases rapidly across these Ages. The middle of the distribution at later Ages is pretty consistent at a Height somewhere between 150 and 175. The units aren't specified (must be cm). The Ages are in years.
3. No Age is reported over 80, and it appears that there is a large cluster of Ages at 80.

Where is the missing data?

```r
summary(nh1)
```

```
       ID             Age            Height     
 Min.   :51671   Min.   : 0.00   Min.   : 86.0  
 1st Qu.:57266   1st Qu.:19.00   1st Qu.:156.9  
 Median :62127   Median :38.00   Median :165.9  
 Mean   :61918   Mean   :37.81   Mean   :162.4  
 3rd Qu.:66780   3rd Qu.:55.00   3rd Qu.:175.0  
 Max.   :71909   Max.   :80.00   Max.   :195.8  
                                 NA's   :31     
```

Subjects with Heights; Ages 21 to 79

```
nh1_rev <- nh1 %>%
    filter(complete.cases(Height)) %>%
    filter(Age > 20 & Age < 80)

dim(nh1_rev)
```

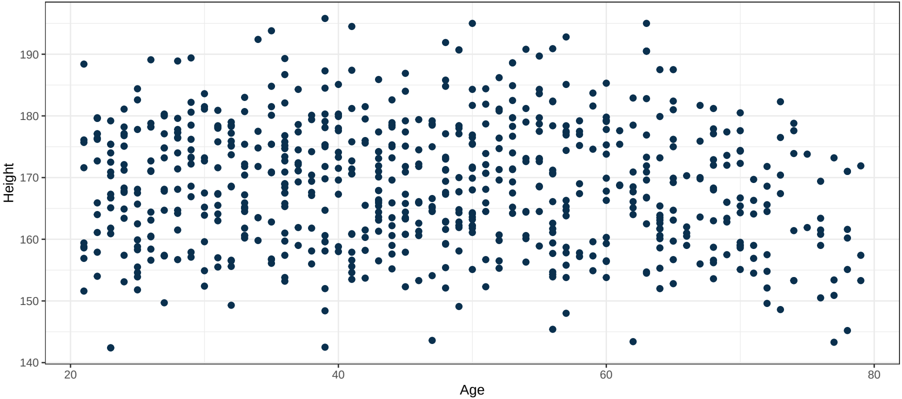
```
[1] 697   3
```

```
summary(nh1_rev)
```

```
       ID             Age            Height     
 Min.   :51678   Min.   :21.00   Min.   :142.4  
 1st Qu.:57204   1st Qu.:34.00   1st Qu.:161.3  
 Median :61663   Median :46.00   Median :168.8  
 Mean   :61773   Mean   :46.42   Mean   :168.9  
 3rd Qu.:66671   3rd Qu.:58.00   3rd Qu.:176.5  
 Max.   :71909   Max.   :79.00   Max.   :195.8  
```

Height/Age Scatterplot for `nh1_rev` sample

```
ggplot(data = nh1_rev, mapping = aes(x = Age, y = Height)) +
    geom_point(size = 2, col = cwru.blue) + theme_bw()
```



nh2: Let's Get Some More Data

We'll focus on data from the 2011_12 `SurveyYr`

Variables of interest to us include:

- `ID` as a code to index the rows (subjects) in the sample
- `SurveyYr` to make sure everyone comes from 2011-12.
- A few quantitative variables: `Age`, `Height`, `Weight`, `BMI`, `Pulse`, `SleepHrsNight` (we'll rename as `SleepHours`), `BPSysAve` and `BPDiaAve` (we'll rename these last two as `SBP` and `DBP`)
- Some binary variables: `Gender` (we'll rename as `Sex`), `PhysActive`, `SleepTrouble` and `Smoke100`
- Several multi-categorical variables: `Race1`, `HealthGen`, `Depressed`

For today, we'll make our life as easy as possible by sampling from the subjects who have complete data (no NA) on all of these variables.

Selecting our nh2 data set

```
set.seed(20190910) # so we can get the same sample again

nh2 <- NHANES %>%
    filter(SurveyYr == "2011_12") %>%
    select(ID, SurveyYr, Age, Height, Weight, BMI, Pulse,
           SleepHrsNight, BPSysAve, BPDiaAve, Gender,
           PhysActive, SleepTrouble, Smoke100,
           Race1, HealthGen, Depressed) %>%
    rename(SleepHours = SleepHrsNight, Sex = Gender,
           SBP = BPSysAve, DBP = BPDiaAve) %>%
    filter(Age > 20 & Age < 80) %>% ## ages 21-79 only
    drop_na() %>% # removes all rows with NA
    sample_n(., size = 1000) %>% # sample 1000 rows
    clean_names() # from the janitor package (snake case)
```

What's in nh2?

```r
dim(nh2)
```

```
[1] 1000   17
```

```r
names(nh2)
```

```
 [1] "id"             "survey_yr"      "age"           
 [4] "height"         "weight"         "bmi"           
 [7] "pulse"          "sleep_hours"    "sbp"           
[10] "dbp"            "sex"            "phys_active"   
[13] "sleep_trouble"  "smoke100"       "race1"         
[16] "health_gen"     "depressed"     
```

Codebook for `nh2` (ID and Quantitative Variables)

Name	Description
`id`	Identifying code for each subject
`survey_yr`	2011_12 for all, indicates administration date
`age`	Age in years at screening of subject (must be 21-79)
`height`	Standing height in cm
`weight`	Weight in kg
`bmi`	Body mass index ($\frac{weight}{height^2}$ in $\frac{kg}{m^2}$)
`pulse`	60 second pulse rate
`sleep_hrs`	Self-reported hours (usually gets) per night
`sbp`	Systolic Blood Pressure (mm Hg)
`dbp`	Diastolic Blood Pressure (mm Hg)

Codebook for `nh2` (Categorical Variables)

Binary Variables

Name	Levels	Description
`sex`	F, M	Sex of study subject
`phys_active`	No, Yes	Moderate or vigorous sports/recreation?
`sleep_trouble`	No, Yes	Has told a provider about trouble sleeping?
`smoke100`	No, Yes	Smoked at least 100 cigarettes in lifetime?

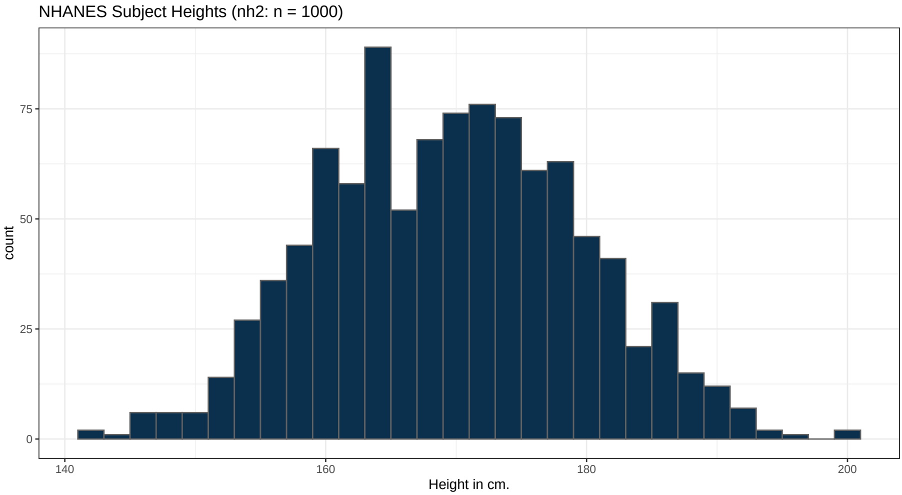
Multi-Categorical Variables

Name	Levels	Description
`race1`	5	Self-reported Race/Ethnicity
`health_gen`	5	Self-reported overall general health
`depressed`	3	How often subject felt depressed in last 30d

Distribution of Height in our nh2 Sample

```
ggplot(data = nh2, mapping = aes(x = height)) +
    geom_histogram(binwidth = 2, col = cwru.gray,
                   fill = cwru.blue) +
    theme_bw() +
    labs(title = "NHANES Subject Heights (nh2: n = 1000)",
         x = "Height in cm.")
```

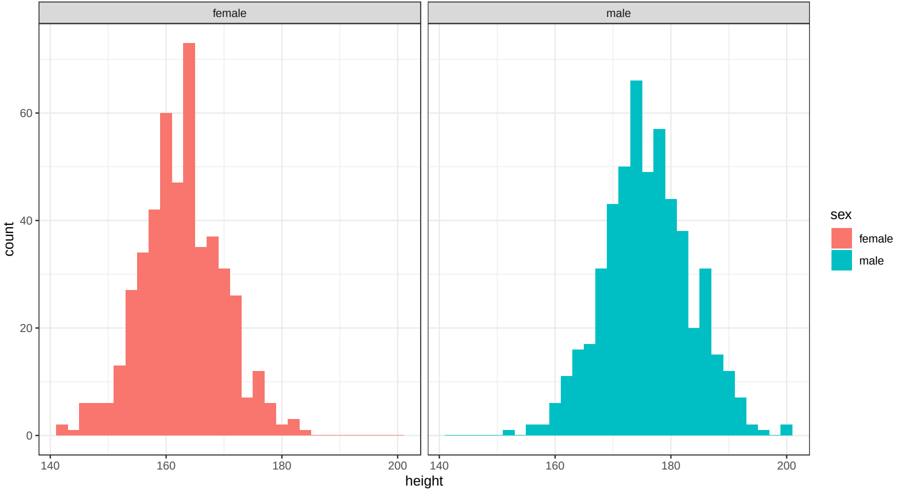
Distribution of Height in our `nh2` Sample



Comparing Height for Males vs. Females

```
ggplot(data = nh2, aes(x = height, fill = sex)) +
    geom_histogram(binwidth = 2) +
    theme_bw() +
    facet_wrap(~ sex)
```

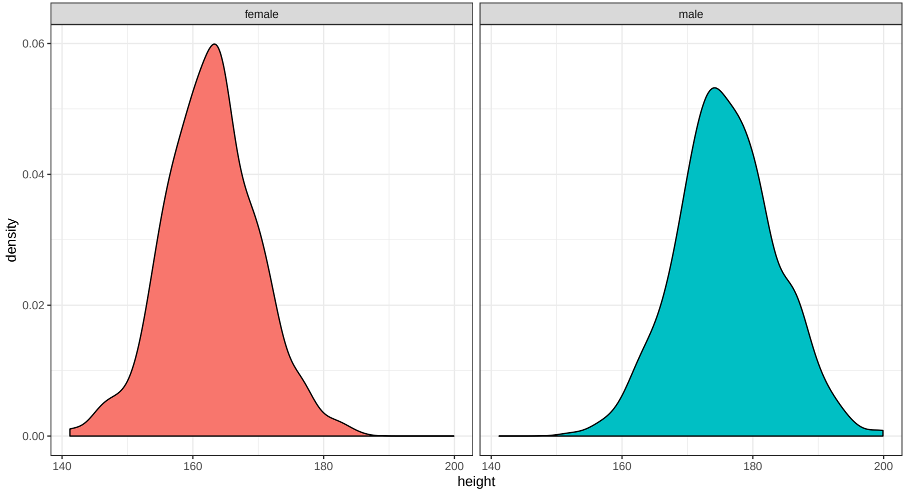
Comparing Height for Males vs. Females



Using geom_density instead of geom_histogram

```
ggplot(data = nh2, aes(x = height, fill = sex)) +
    geom_density(kernel = "gaussian") + # default choice
    theme_bw() +
    guides(fill = FALSE) +
    facet_wrap(~ sex)
```

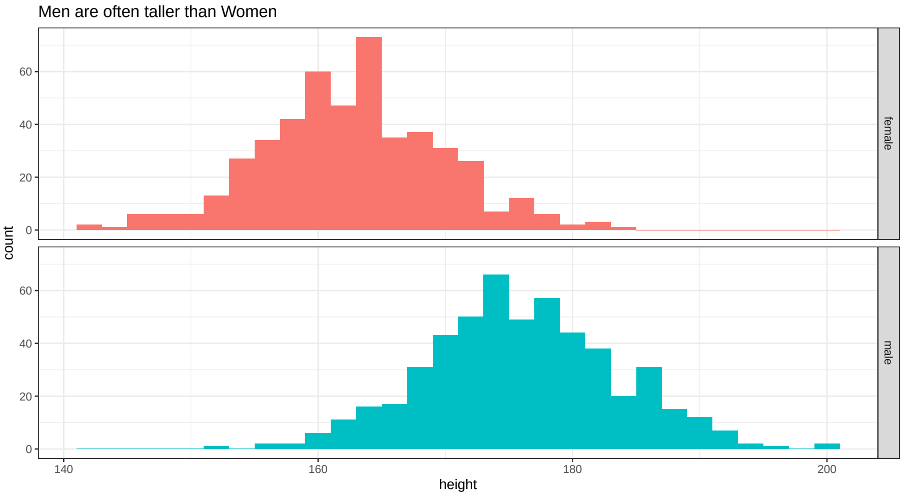
Using `geom_density` instead of `geom_histogram`



Histograms in a Single Column with `facet_grid`

```
ggplot(data = nh2, aes(x = height, fill = sex)) +
    geom_histogram(binwidth = 2) +
    theme_bw() +
    guides(fill = FALSE) +
    facet_grid(sex ~ .) +
    labs(title = "Men are often taller than Women")
```

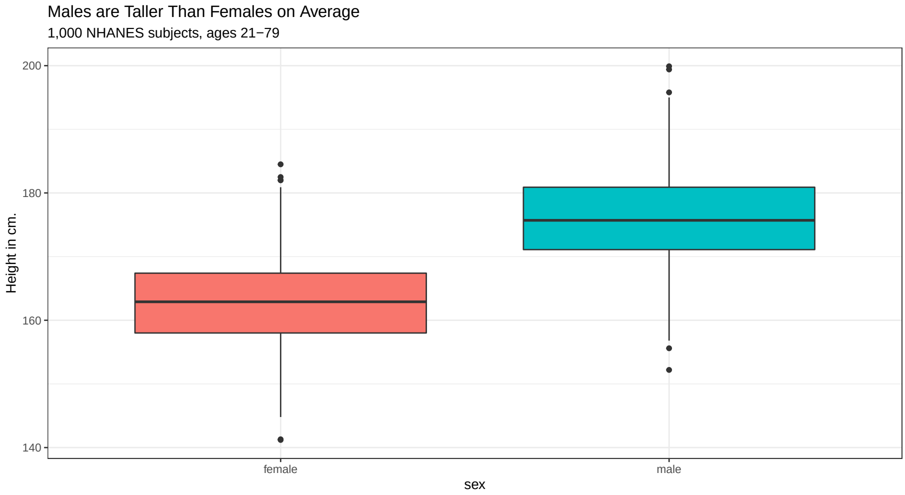
Histograms in a Single Column with `facet_grid`



Boxplot of Height for Males vs. Females

```
ggplot(data = nh2, aes(x = sex, y = height, fill = sex)) +
    geom_boxplot() +
    guides(fill = FALSE) +
    theme_bw() +
    labs(title = "Males are Taller Than Females on Average",
         subtitle = "1,000 NHANES subjects, ages 21-79",
         y = "Height in cm.")
```

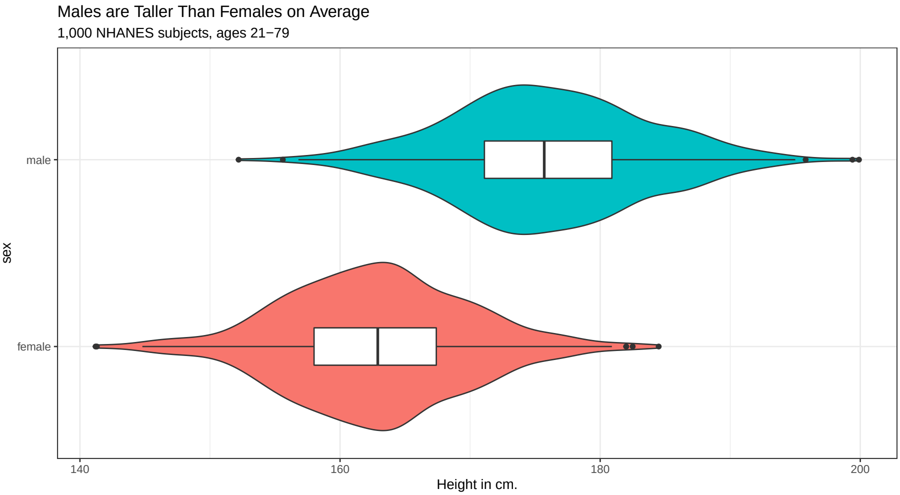
Boxplot of Height for Males vs. Females



Violin Plot of Height for Males vs. Females

```
ggplot(data = nh2, aes(x = sex, y = height, fill = sex)) +
    geom_violin() +
    geom_boxplot(fill = "white", width = 0.2) +
    guides(fill = FALSE) +
    coord_flip() +
    theme_bw() +
    labs(title = "Males are Taller Than Females on Average",
         subtitle = "1,000 NHANES subjects, ages 21-79",
         y = "Height in cm.")
```

Violin Plot of Height for Males vs. Females



A Look at Body-Mass Index

Let's look at the *body-mass index*, or BMI. The definition of BMI for adult subjects (which is expressed in units of kg/m^2) is:

$$\text{BMI} = \frac{\text{weight in kg}}{(\text{height in meters})^2} = 703 \times \frac{\text{weight in pounds}}{(\text{height in inches})^2}$$

BMI is, essentially, a measure of a person's *thinnness* or *thickness*.

- BMI from 18.5 to 25 indicates optimal weight
- BMI below 18.5 suggests person is underweight
- BMI above 25 suggests overweight.
- BMI above 30 suggests obese.

A First Set of Exploratory Questions

Variables of Interest: `bmi`, `phys_active`, `health_gen`, `pulse`

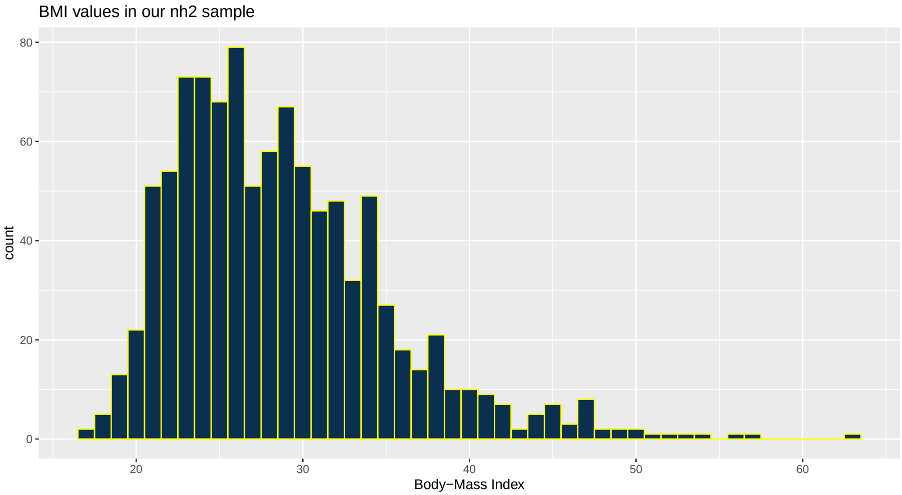
1. What is the distribution of BMI in our `nh2` sample of adults?
2. How does the distribution of BMI vary by whether the subject is physically active?
3. How does the distribution of BMI vary by the subject's self-reported general health?
4. What is the association between BMI and the subject's pulse rate?
5. Does that BMI-Pulse association differ in subjects who are physically active, and those who are not?

Note: These are NOT what anyone would call research questions, which involve generating scientific hypotheses, among other things. These are merely triggers for visualizations and (small) analyses.

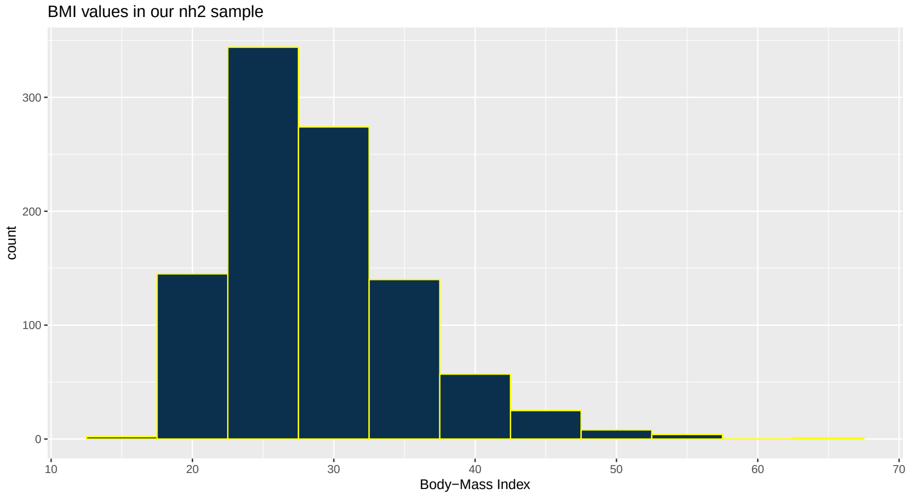
Histogram of BMI in `nh2` with binwidth = 1

```
ggplot(nh2, aes(x = bmi)) +
    geom_histogram(binwidth = 1, fill = cwru.blue,
                   col = "yellow") +
    labs(title = "BMI values in our nh2 sample",
         x = "Body-Mass Index")
```

Histogram of BMI in `nh2` with binwidth = 1



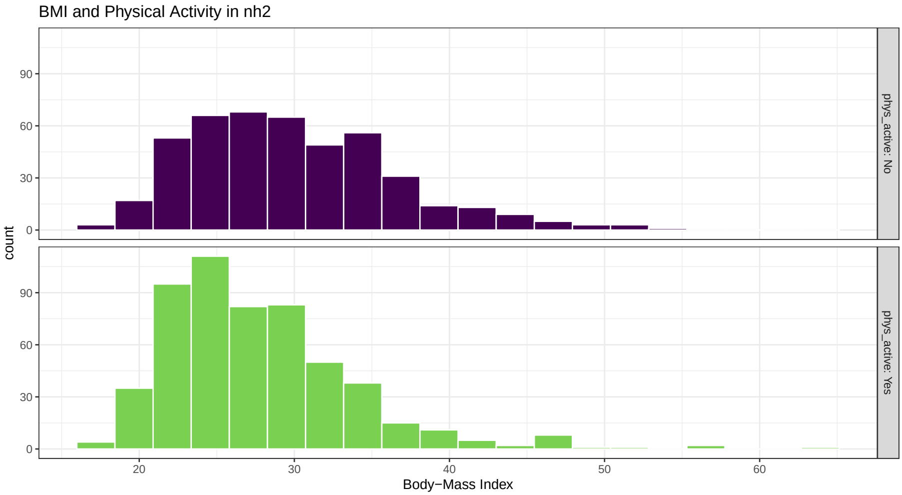
Histogram of BMI in `nh2` with binwidth = 5



BMI Histograms faceted by Physical Activity Status

```
ggplot(nh2, aes(x = bmi, fill = phys_active)) +
    geom_histogram(bins = 20, col = "white") +
    labs(title = "BMI and Physical Activity in nh2",
         x = "Body-Mass Index") +
    scale_fill_viridis_d(end = 0.8) +
    guides(fill = FALSE) +
    theme_bw() +
    facet_grid(phys_active ~ ., labeller = "label_both")
```

BMI Histograms faceted by Physical Activity Status



Average BMI by Physical Activity Status, I

Create a tibble that helps us answer:

- What is the "average" BMI in each activity group?
- How many people fall into each activity group?

```
nh2 %>%
    group_by(phys_active) %>%
    summarize(count = n(), mean(bmi), median(bmi))
```

```
# A tibble: 2 x 4
  phys_active count `mean(bmi)` `median(bmi)`
  <fct>       <int>       <dbl>         <dbl>
1 No            456        30.0          28.9
2 Yes           544        27.7          26.4
```

Average BMI by Physical Activity Status, II

Making this look a bit more presentable as a table...

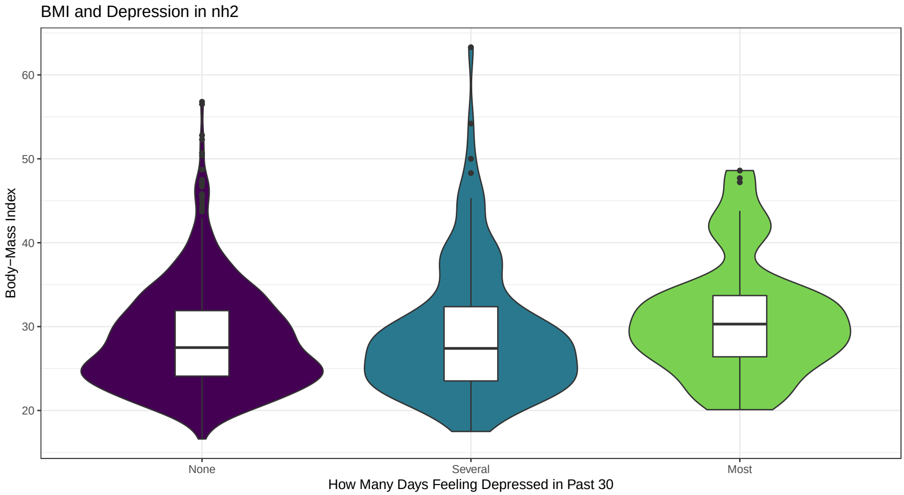
```
nh2 %>%
    group_by(phys_active) %>%
    summarize("Count" = n(),
              "Mean(BMI)" = round(mean(bmi),2),
              "Median(BMI)" = median(bmi)) %>%
    knitr::kable()
```

phys_active	Count	Mean(BMI)	Median(BMI)
No	456	29.98	28.90
Yes	544	27.73	26.45

BMI by Depression Status: Violin Plot

```
ggplot(nh2, aes(x = depressed, y = bmi, fill = depressed)) +
    geom_violin() +
    geom_boxplot(width = 0.2, fill = "white") +
    labs(title = "BMI and Depression in nh2",
         y = "Body-Mass Index",
         x = "How Many Days Feeling Depressed in Past 30") +
    scale_fill_viridis_d(end = 0.8) +
    guides(fill = FALSE) +
    theme_bw()
```

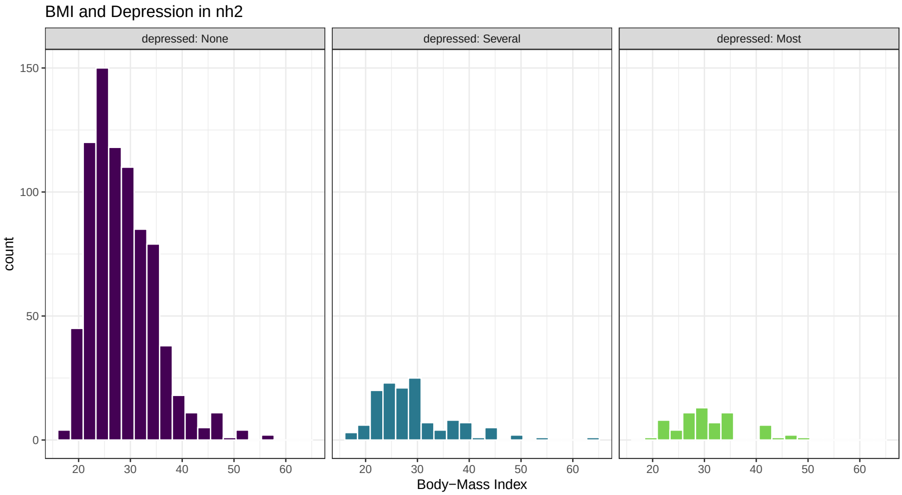
BMI by Depression Status: Violin Plot



BMI by Depression Status, Faceted Histograms

```
ggplot(nh2, aes(x = bmi, fill = depressed)) +
    geom_histogram(bins = 20, col = "white") +
    labs(title = "BMI and Depression in nh2",
         x = "Body-Mass Index") +
    scale_fill_viridis_d(end = 0.8) +
    guides(fill = FALSE) +
    theme_bw() +
    facet_wrap(~ depressed, labeller = "label_both")
```

BMI by Depression Status, Faceted Histograms



BMI by Depression Status, Numerically

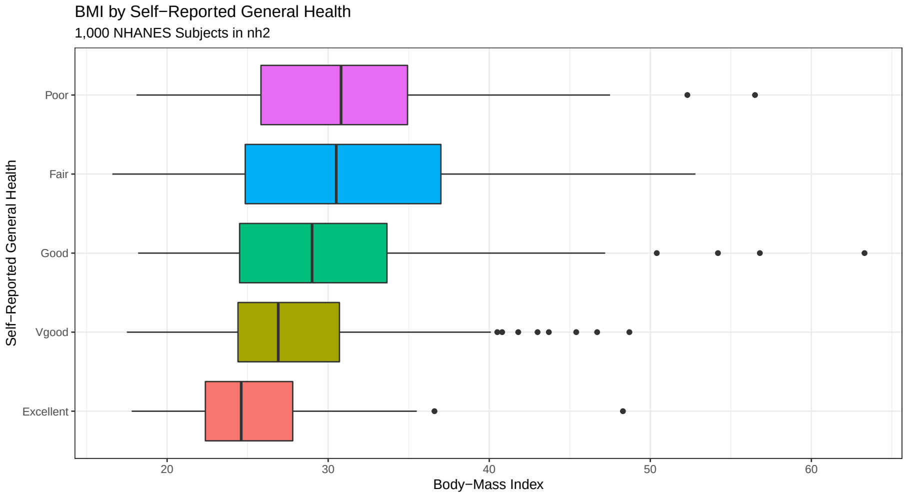
```
nh2 %>%
    group_by(depressed) %>%
    summarize("Count" = n(),
              "Mean(BMI)" = round(mean(bmi),2),
              "Median(BMI)" = median(bmi)) %>%
    knitr::kable()
```

depressed	Count	Mean(BMI)	Median(BMI)
None	801	28.53	27.5
Several	134	29.12	27.4
Most	65	30.89	30.3

BMI by Self-Reported Health Status

```
ggplot(nh2, aes(x = health_gen, y = bmi,
                fill = health_gen)) +
    geom_boxplot() +
    theme_bw() +
    coord_flip() +
    guides(fill = FALSE) +
    labs(title = "BMI by Self-Reported General Health",
         subtitle = "1,000 NHANES Subjects in nh2",
         x = "Self-Reported General Health",
         y = "Body-Mass Index")
```

BMI by Self-Reported Health Status



BMI by Self-Reported Health Status

```
nh2 %>%
    group_by(health_gen) %>%
    summarize(count = n(), mean(bmi),
              median(bmi), sd(bmi)) %>%
    knitr::kable(digits = 2)
```

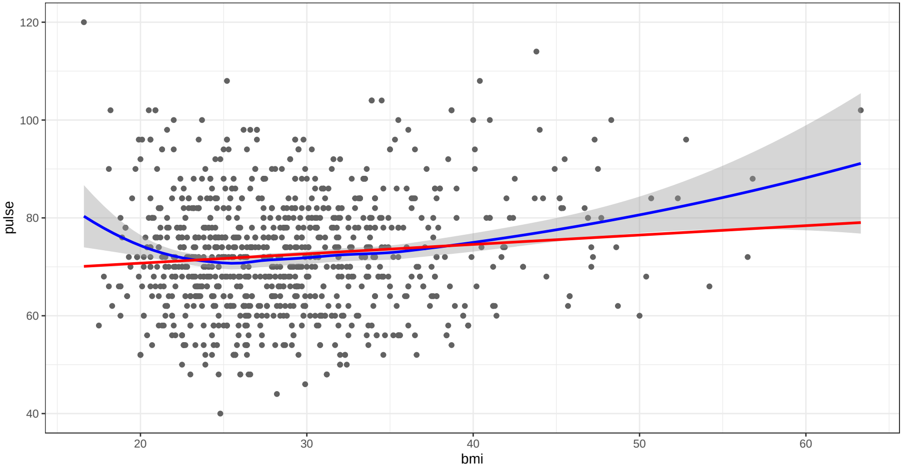
health_gen	count	mean(bmi)	median(bmi)	sd(bmi)
Excellent	144	25.47	24.6	4.51
Vgood	329	27.86	26.9	5.14
Good	383	29.62	29.0	6.76
Fair	124	31.69	30.5	7.83
Poor	20	32.56	30.8	9.80

Association of BMI and Pulse Rate

```
ggplot(nh2, aes(x = bmi, y = pulse)) +
    geom_point(col = cwru.gray) +
    geom_smooth(method = "loess", se = TRUE, col = "blue") +
    geom_smooth(method = "lm", se = FALSE, col = "red") +
    theme_bw() +
    labs(title = "BMI and Pulse Rate in 1,000 nh2 Subjects")
```

Association of BMI and Pulse Rate

BMI and Pulse Rate in 1,000 nh2 Subjects



Correlation Coefficient to Summarize Association?

The Pearson correlation coefficient is a very limited measure. It only describes the degree to which a **linear** relationship is present in the data. But we can look at it.

```
nh2 %$% cor(bmi, pulse)
```

```
[1] 0.1076127
```

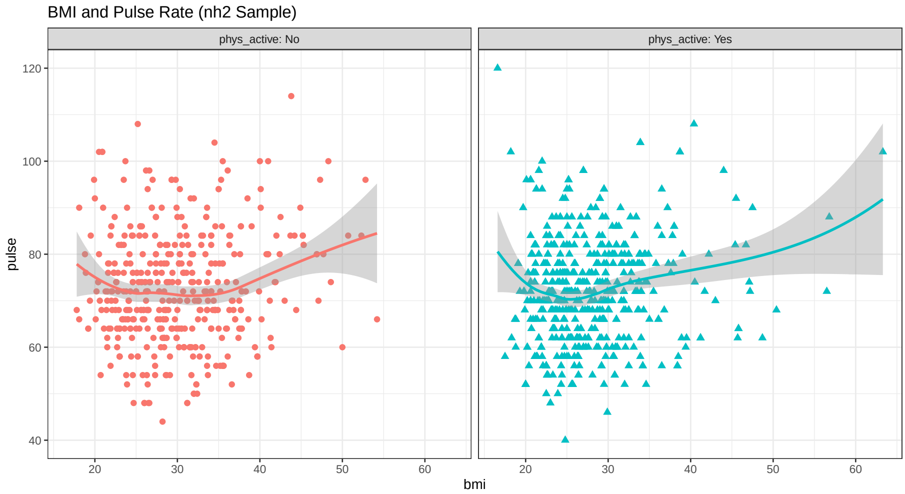
- The Pearson correlation ranges from -1 (perfect negative [as x rises, y falls] linear relationship) to +1 (perfect positive [as x rises, y rises] linear relationship.)
- Our correlation is pretty close to zero. This implies we have a very weak linear association in this case, across the entire sample.

Does Physical Activity affect the Pulse-BMI Association?

Let's change the shape and color of the points based on physical activity status.

```
ggplot(data = nh2, aes(x = bmi, y = pulse,
                       color = phys_active,
                       shape = phys_active)) +
    geom_point(size = 2) +
    geom_smooth(method = "loess") +
    guides(color = FALSE, shape = FALSE) +
    labs(title = "BMI and Pulse Rate (nh2 Sample)") +
    facet_wrap(~ phys_active, labeller = "label_both") +
    theme_bw()
```

Does Physical Activity affect the Pulse-BMI Association?



Correlation(bmi, pulse) by Physical Activity?

- The Pearson correlation coefficient for the relationship between `bmi` and `pulse` in the full sample was quite weak, specifically, it was 0.108.
- Grouped by physical activity status, do we get a different story?

```
nh2 %>%
    group_by(phys_active) %>%
    summarize(cor(bmi, pulse)) %>%
    knitr::kable(digits = 3)
```

phys_active	cor(bmi, pulse)
No	0.101
Yes	0.114

The Elements of Data Analytic Style

What I Asked You To Do

Write down (so that someone else can read it) the most important/interesting/surprising thing you learned from reading the four chapters of Jeff Leek's *Elements of Data Analytic Style*.

- One sentence is plenty.
- If you cannot limit yourself to one thing, try to keep it to two.
- Later in today's class (about 2 PM), you'll share these with a colleague.

Now, as a group of 4-5 people. . .

Share your "interesting things" with your group. Identify one of the things to represent the "most" interesting thing mentioned by your group, and make sure that person is ready to share that with the class. You have 3 minutes.

What Did You Come Up With?

Dr. Love's list of interesting items is on the next few slides. You'll probably get more out of this if you wait to review those slides.

Leek Chapter 5: Exploratory Analysis

- EDA To understand properties of the data and discover new patterns
- Visualize and inspect qualitative features rather than a huge table of raw data

1. Make big data as small as possible as quickly as possible
2. Plot as much of the actual data as you can
3. For large data sets, subsample before plotting
4. Use log transforms for ratio measurements
5. Missing values can have a mighty impact on conclusions

Leek: Chapter 9 Written Analyses

Elements: title, introduction/motivation, description of statistical tools used, results with measures of uncertainty, conclusions indicating potential problems, references

1. What is the question you are answering?
2. Lead with a table summarizing your tidy data set (critical to identify data versioning issues)
3. For each parameter of interest report an estimate and measure of uncertainty on the scientific scale of interest
4. Summarize the importance of reported estimates
5. Do not report every analysis you performed

Leek: Chapter 10 Creating Figures

Communicating effectively with figures is non-trivial. The goal is clarity.

> *When viewed with an appropriately detailed caption, (a figure should) stand alone without any further explanation as a unit of information.*

1. Humans are best at perceiving position along a single axis with a common scale
2. Avoid chartjunk (gratuitous flourishes) in favor of high-density displays
3. Axis labels should be large, easy to read, in plain language
4. Figure titles should communicate the plot's message
5. Use a palette (like `viridis`) that color-blind people can see (and distinguish) well

Check out Karl Broman's excellent presentation on displaying data badly at https://github.com/kbroman/Talk_Graphs

Leek Chapter 13: A Few Matters of Form

- Variable names should always be reported in plain language.
- If measurements are only accurate to the tenths digit, don't report estimates with more digits.
- Report estimates followed by parentheses that hold a 95% CI or other measure of uncertainty.
- When reporting p values, censor small values ($p < 0.0001$, not $p = 0$ or $p = 1.6 \times 10^{-25}$)

Reminders

The Course Project

Take a look at the web site. We'll start taking questions about the Project at `431-help` after class today.

Homework C

Due Friday at Noon.

Minute Paper after Class 5

Please complete today's Minute Paper (by noon Wednesday).