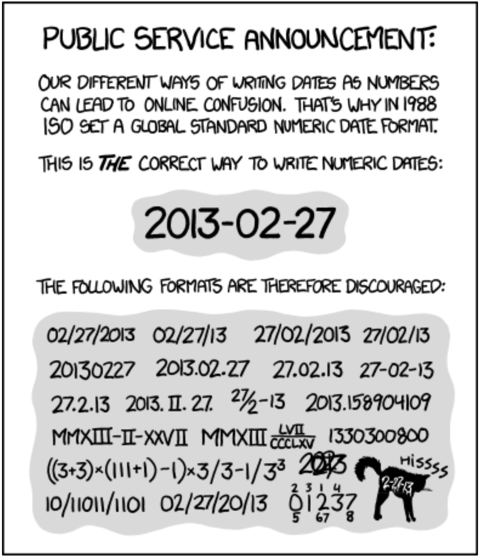
431 Class 04

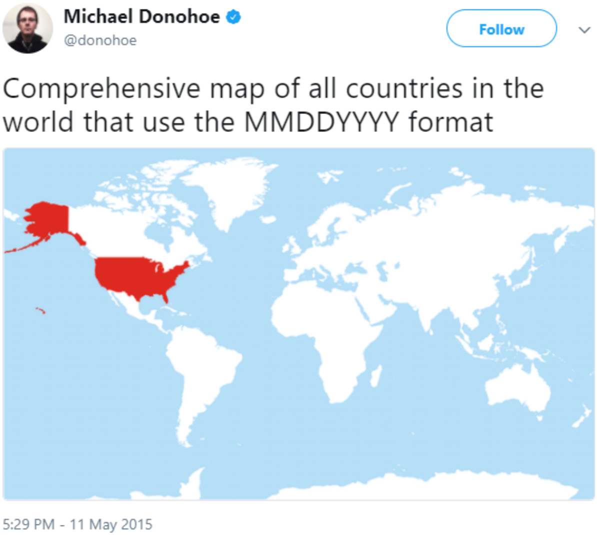
Thomas E. Love

2017-09-07

from XKCD


PUBLIC SERVICE ANNOUNCEMENT:
OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.
THIS IS THE CORRECT WAY TO WRITE NUMERIC DATES:
2013-02-27
THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:
02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013. II. 27. 27/2-13 2013.158904109
MMXIII-II-XXVII MMXIII LVII/CCCLXV 1330300800
((3+3)×(111+1)−1)×3/3−1/3³ 2013
HISSsss
2-27-13
10/11011/1101 02/27/20/13 01237
2 3 1 4
5 67 8

Dates



Michael Donohoe
@donohoe

Comprehensive map of all countries in the world that use the MMDDYYYY format

5:29 PM - 11 May 2015

https://twitter.com/donohoe/status/597876118688026624

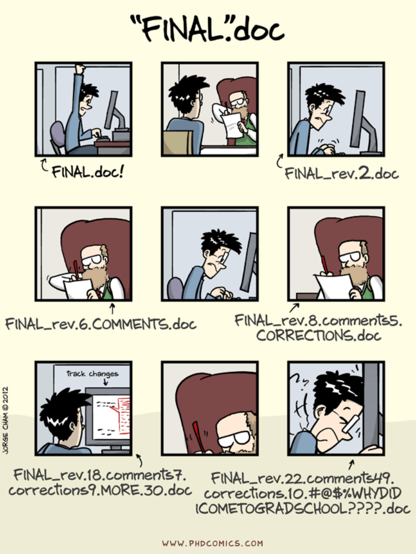
Today's Agenda

Visit the README file at

https://github.com/THOMASELOVE/431slides/tree/master/class_04

1. Highlights from Chapter 3 of the **Course Notes**
2. Jeff Leek *Elements of Data Analytic Style*
   - Chapter 5 is about Exploratory Analysis
   - Chapter 9 is about Written Analyses (keep this in mind for Assignments!)
   - Chapter 10 is about Creating Figures
   - Chapter 13 highlights a few matters of form
3. Kidney Cancer death rates

Choose good names for things.



http://www.phdcomics.com/comics/archive.php?comicid=1531

Great advice on machine-readable, human-readable names

https://speakerdeck.com/jennybc/how-to-name-files

- Jenny Bryan

“Go forward and use awesome filenames!”

Course Notes, Chapter 3 on Visualization

The packages we're using are NHANES, viridis and tidyverse.

```
library(NHANES)
library(viridis)
library(tidyverse)
```

The NHANES data: Collecting a Sample

To begin, we'll gather a random sample of 1,000 subjects participating in NHANES, and then identify several variables of interest about those subjects. See Baumer, Kaplan and Horton (2017) *Modern Data Science with R*. Use `?NHANES` to learn more about the data.

```r
set.seed(431001)
# use set.seed to ensure that we all
# get the same random sample
# of 1,000 NHANES subjects in nh_data

nh_data <- sample_n(NHANES, size = 1000) %>%
    select(ID, Gender, Age, Height, Weight, BMI,
           Pulse, Race1, HealthGen, Diabetes)
```

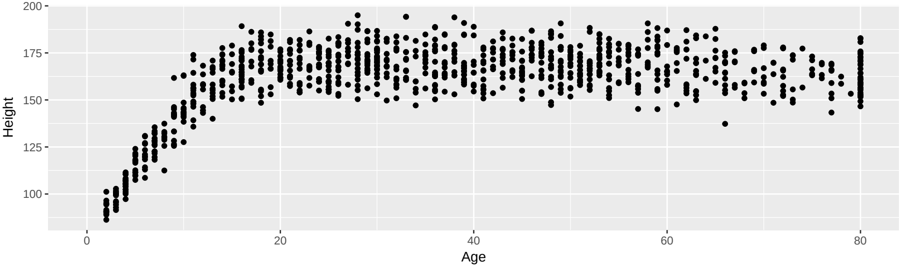
The nh_data tibble

```
# A tibble: 1,000 x 10
      ID Gender   Age Height Weight   BMI Pulse
   <int> <fctr> <int>  <dbl>  <dbl> <dbl> <int>
 1 59640   male    54  175.7  129.0 41.79    74
 2 59826 female    67  156.5   50.2 20.50    66
 3 56340   male     9  128.3   23.3 14.15    86
 4 56747   male    33  194.2  105.1 27.87    68
 5 51754 female    58  167.2  106.0 37.92    70
 6 52712   male     6  108.6   16.9 14.33    NA
 7 63908   male    55  168.6   90.6 31.90    62
 8 60865 female    25  155.5   55.0 22.75    58
 9 66642   male    41  177.9   89.3 28.20    72
10 59880 female    45  163.2   98.3 36.91    80
# ... with 990 more rows, and 3 more variables:
#   Race1 <fctr>, HealthGen <fctr>, Diabetes <fctr>
```

Relationship of Height and Age - First Attempt

```
ggplot(data = nh_data, aes(x = Age, y = Height)) +
    geom_point()
```

```
Warning: Removed 25 rows containing missing values
(geom_point).
```



Interesting Results from Our First Attempt

1. Only 975 subjects are plotted, because the remaining 25 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.

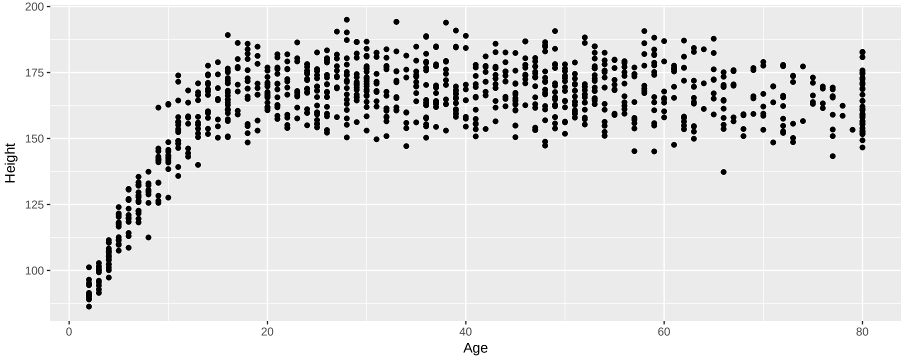
Subset of Subjects with Known Age and Height

Before we move on, let's manipulate the data set a bit, to focus on only those subjects who have complete data on both Age and Height. This will help us avoid that warning message.

```
nh_dat2 <- nh_data %>%
    filter(complete.cases(Age, Height))
```

Revised Height by Age plot (using nh_dat2)

```
ggplot(data = nh_dat2, aes(x = Age, y = Height)) +
    geom_point()
```



Adding Gender to the picture

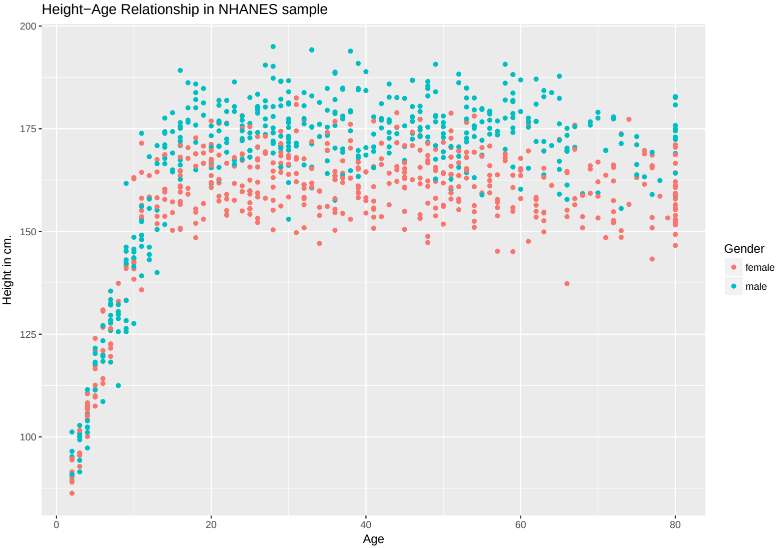
Goals

1. Add gender to the plot using color.
2. Adjust axis labels to show units of measurement.

```r
ggplot(data = nh_dat2,
       aes(x = Age, y = Height,
           color = Gender)) +
    geom_point() +
    labs(title = "Height-Age Relationship in NHANES sample",
         y = "Height in cm.")
```

Result on next slide. . .

Age-Height and Gender?



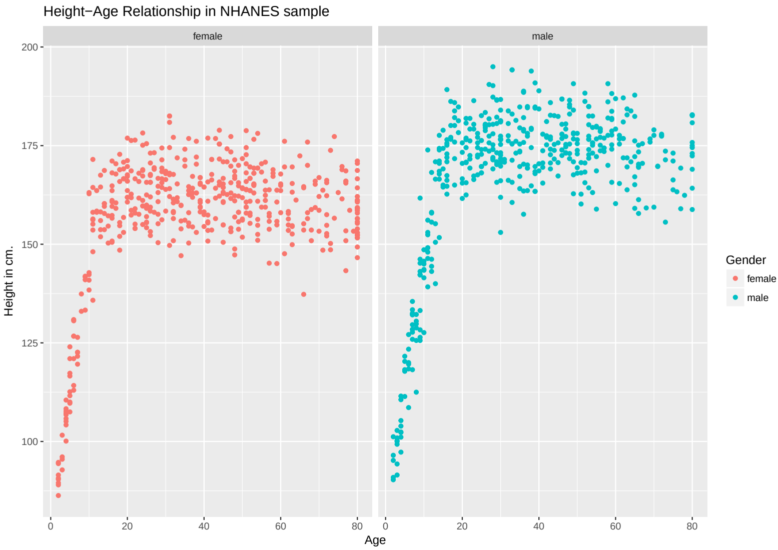
Can we show the Female and Male relationships in separate panels?

Sure. Add `facet_wrap(~ Gender)`

- Don't forget to add the + sign at the end of the preceding line, too.

```r
ggplot(data = nh_dat2,
       aes(x = Age, y = Height,
           color = Gender)) +
    geom_point() +
    labs(title = "Height-Age Relationship in NHANES sample",
         y = "Height in cm.") +
    facet_wrap(~ Gender)
```

Faceting by Gender

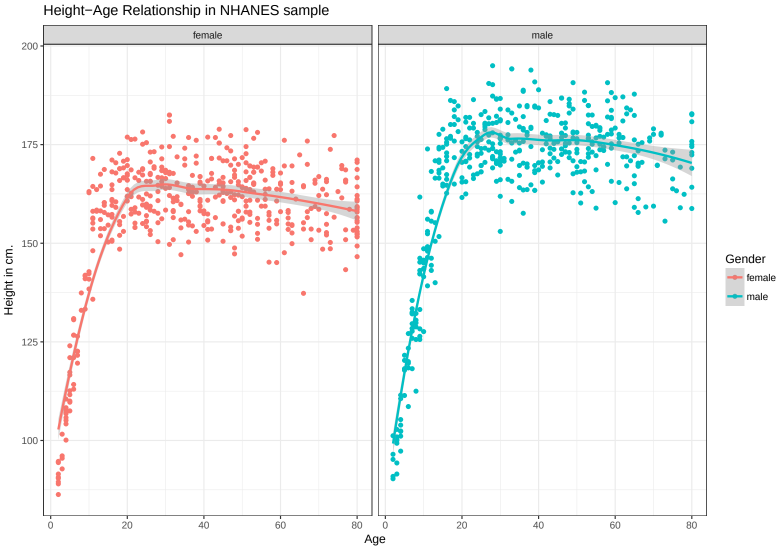


Can we add a smooth curve to show the relationship in each plot?

Yep, and let's change the theme of the graph to remove the gray background, too.

```
ggplot(data = nh_dat2, aes(x = Age, y = Height, color = Gender
    geom_point() +
    geom_smooth(method = "loess") +
    labs(title = "Height-Age Relationship in NHANES sample",
         y = "Height in cm.") +
    theme_bw() +
    facet_wrap(~ Gender)
```

With Smooth Curves (from `loess`)

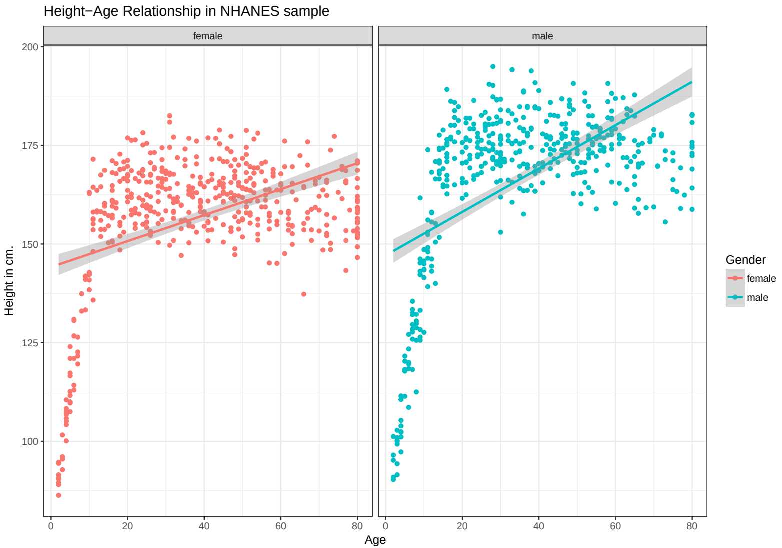


What if we want to assume straight line relationships?

We could look at a linear model in the plot. Does this make sense here?

```r
ggplot(data = nh_dat2,
       aes(x = Age, y = Height, color = Gender)) +
    geom_point() +
    geom_smooth(method = "lm") +
    labs(title = "Height-Age Relationship in NHANES sample",
         y = "Height in cm.") +
    theme_bw() +
    facet_wrap(~ Gender)
```

Linear Models for each Gender



A Subset: Ages 21-79

Suppose we wanted to look at a subset of our sample - those observations (subjects) whose Age is at least 21 and at most 79. We'll create that sample below, and also subset the variables to include nine of particular interest, and remove any observations with any missingness on *any* of the nine variables we're including here.

```r
nh_data_2179 <- nh_data %>%
    filter(Age > 20 & Age < 80) %>%
    select(ID, Gender, Age, Height, Weight, BMI,
           Pulse, Race1, HealthGen, Diabetes) %>%
    na.omit
```

The `nh_data_2179` tibble

```
nh_data_2179
```

```
# A tibble: 594 x 10
      ID Gender   Age Height Weight   BMI Pulse
   <int> <fctr> <int>  <dbl>  <dbl> <dbl> <int>
 1 59640   male    54  175.7  129.0 41.79    74
 2 59826 female    67  156.5   50.2 20.50    66
 3 56747   male    33  194.2  105.1 27.87    68
 4 63908   male    55  168.6   90.6 31.90    62
 5 60865 female    25  155.5   55.0 22.75    58
 6 66642   male    41  177.9   89.3 28.20    72
 7 59880 female    45  163.2   98.3 36.91    80
 8 71784 female    24  161.1   50.2 19.30    72
 9 67616   male    63  184.3   70.0 20.60    82
10 55391 female    32  161.4   69.2 26.56   114
# ... with 584 more rows, and 3 more variables:
#   Race1 <fctr>, HealthGen <fctr>, Diabetes <fctr>
```

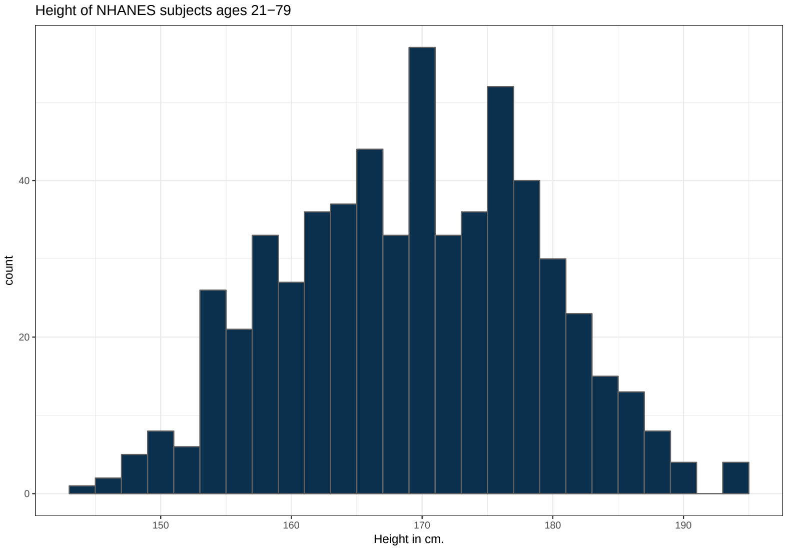
Distribution of Height

CWRU's color guide specifies CWRU blue and CWRU gray.

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

ggplot(data = nh_data_2179, aes(x = Height)) +
    geom_histogram(binwidth = 2,
                   col = cwru.gray, fill = cwru.blue) +
    labs(title = "Height of NHANES subjects ages 21-79",
         x = "Height in cm.") +
    theme_bw()
```

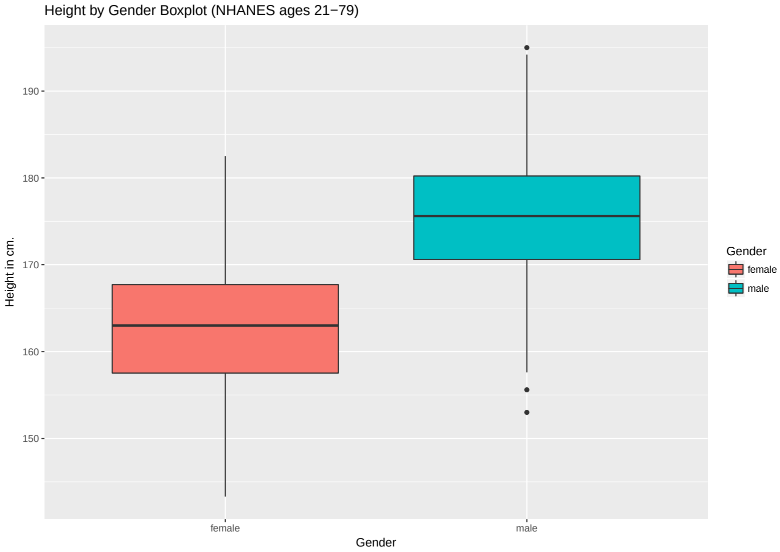
Distribution of Height (in CWRU colors)



A Boxplot of Height by Gender

```r
ggplot(data = nh_data_2179,
       aes(x = Gender, y = Height, fill = Gender)) +
    geom_boxplot() +
    labs(title = "Boxplot of Height by Gender
         for NHANES subjects ages 21-79",
         y = "Height in cm.")
```

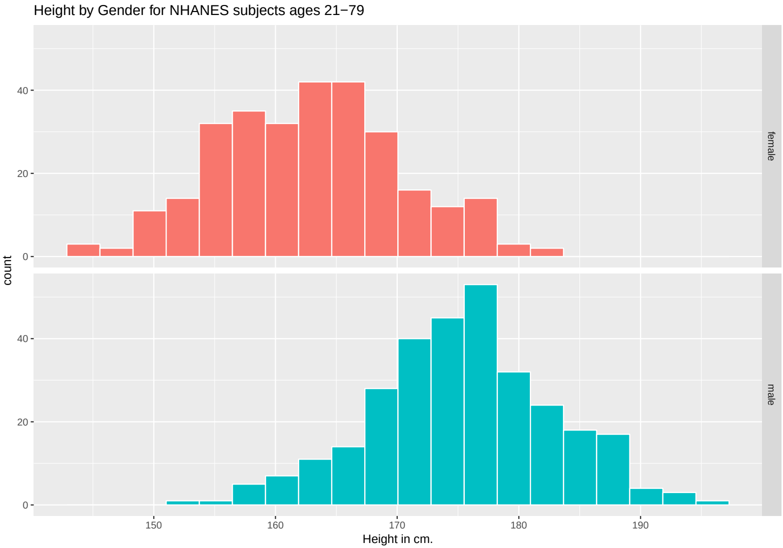
Boxplot: Height by Gender (ages 21-79)



Faceted Histograms of Height by Gender

```r
ggplot(data = nh_data_2179,
       aes(x = Height, fill = Gender)) +
    geom_histogram(color = "white", bins = 20) +
    labs(title = "Height by Gender Boxplot (NHANES ages 21-79)
         x = "Height in cm.") +
    guides(fill = FALSE) +
    facet_grid(Gender ~ .)
```

Faceted Histograms of Height by Gender



Quick Interlude: Assignment 1 (due Friday 2017-09-15)

1. Use the `YOURNAME-hw1.Rmd` template to your advantage.
2. Use the Getting Started in R document from our front page to help guide you.
3. The Course Notes contain all the code you might possibly need.
4. Grading will be very light on this assignment compared to later ones.
5. Submit the assignment (two files: R Markdown, plus either HTML or Word files) via canvas.case.edu
6. Apply the 15-minute rule.
    - If you can't solve a problem in 15 minutes, ask for help.
    - You are **absolutely supposed** to use Google and the TAs (and me) to improve your code.

A Look at Body-Mass Index

Let's look at a different outcome, 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*.

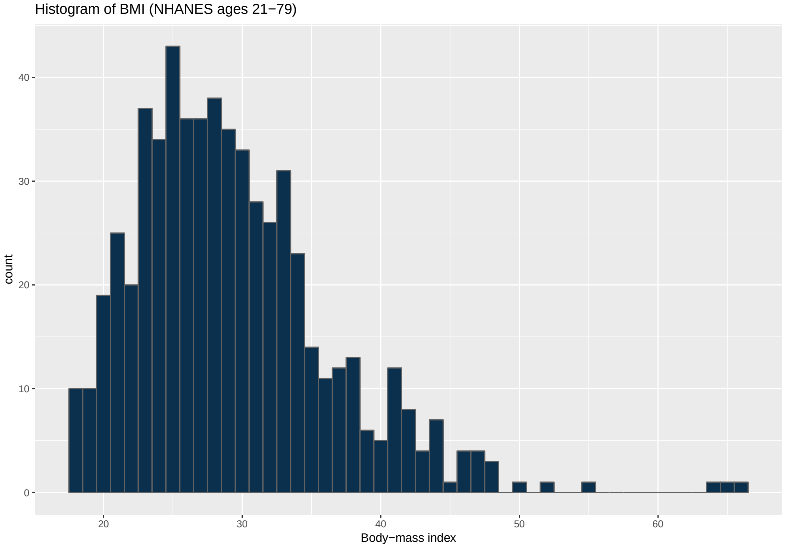
- BMI from 18.5 to 25 indicates optimal weight
- BMI below 18.5 suggests person is underweight
- BMI above 25 suggests overweight.
- BIM above 30 suggests obese.

Histogram of BMI

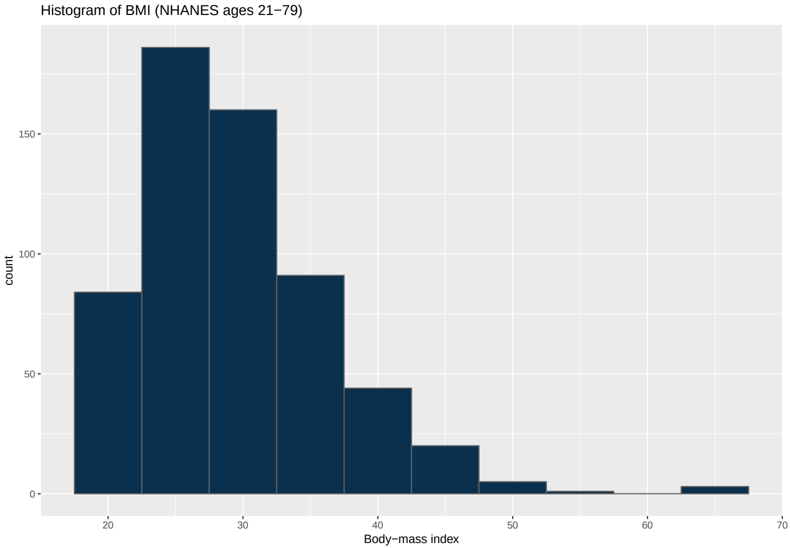
Here's a histogram, again with CWRU colors, for the BMI data.

```r
ggplot(data = nh_data_2179, aes(x = BMI)) +
    geom_histogram(binwidth = 1,
                   fill = cwru.blue, col = cwru.gray) +
    labs(title = "Histogram of BMI (NHANES ages 21-79)",
         x = "Body-mass index")
```

Histogram of BMI with binwidth 1



Histogram of BMI with binwidth 5

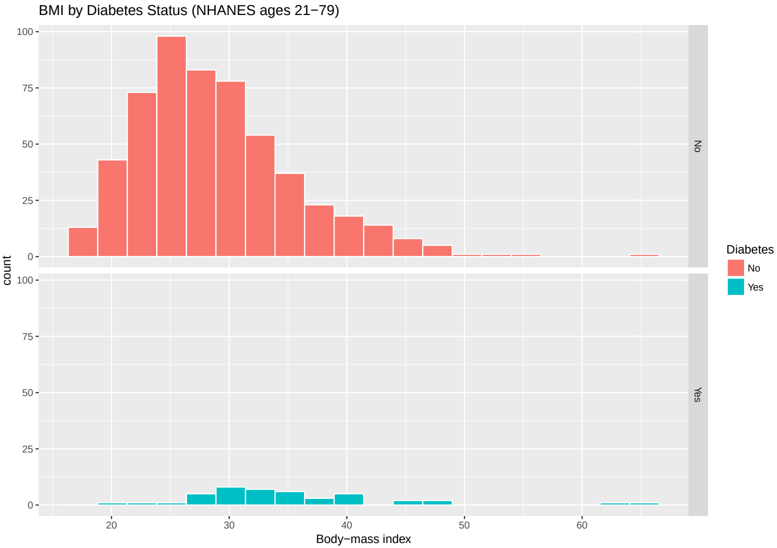


BMI and Diabetes

We can split up our histogram into groups based on whether the subjects have been told they have diabetes.

```
ggplot(data = nh_data_2179,
       aes(x = BMI, fill = Diabetes)) +
    geom_histogram(color = "white", bins = 20) +
    labs(title = "BMI by Diabetes Status (NHANES ages 21-79)",
         x = "Body-mass index") +
    facet_grid(Diabetes ~ .)
```

BMI and Diabetes



Numerical Summary: BMI and Diabetes

How many people fall into each of these Diabetes categories, and what is their "average" BMI?

```
nh_data_2179 %>%
    group_by(Diabetes) %>%
    summarize(count = n(), mean(BMI), median(BMI)) %>%
    knitr::kable()
```

Diabetes	count	mean(BMI)	median(BMI)
No	551	28.89544	27.89
Yes	43	35.26209	33.43

BMI by Race

How many people fall into each of the five Race1 categories, and what is their "average" BMI?

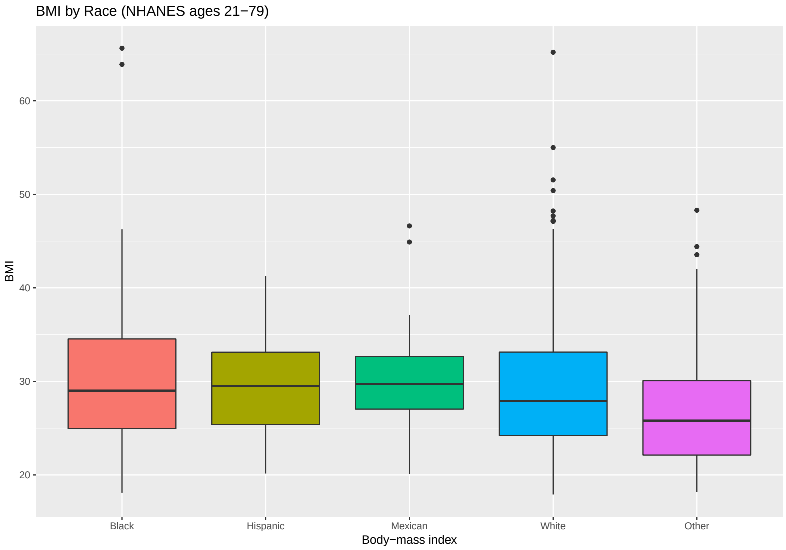
```
nh_data_2179 %>%
    group_by(Race1) %>%
    summarize(count = n(), mean(BMI), median(BMI)) %>%
    knitr::kable()
```

Race1	count	mean(BMI)	median(BMI)
Black	63	31.04444	29.010
Hispanic	44	29.36227	29.505
Mexican	50	29.97040	29.730
White	387	29.27326	27.900
Other	50	27.25300	25.805

BMI and Race Boxplot

```r
ggplot(data = nh_data_2179,
       aes(x = Race1, y = BMI, fill = Race1)) +
    geom_boxplot() +
    guides(fill = FALSE) +
    labs(title = "BMI by Race (NHANES ages 21-79)",
         x = "Body-mass index")
```

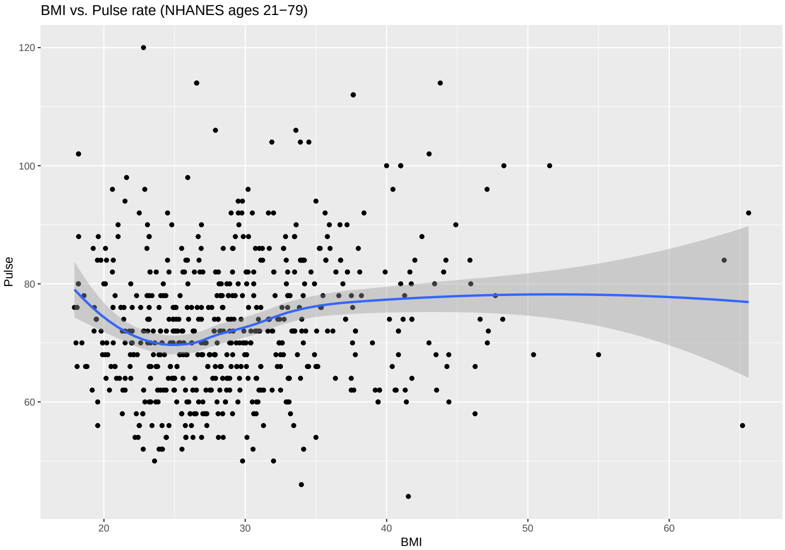
BMI and Race Boxplot



BMI and Pulse Rate

```r
ggplot(data = nh_data_2179, aes(x = BMI, y = Pulse)) +
    geom_point() +
    geom_smooth(method = "loess") +
    labs(title = "BMI vs. Pulse rate (NHANES ages 21-79)")
```

BMI and Pulse Rate



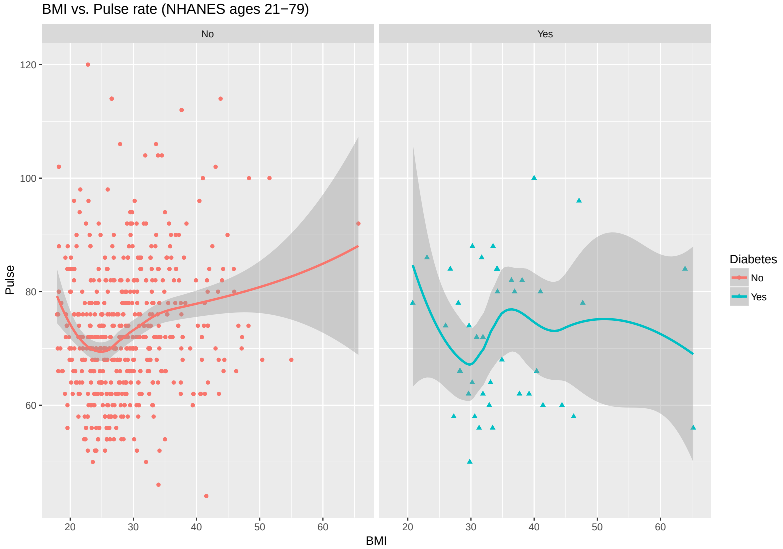
Diabetes vs. No Diabetes

Could we see whether subjects who have been told they have diabetes show different BMI-pulse rate patterns than the subjects who haven't?

- Let's try doing this by changing the **shape** *and* the **color** of the points based on diabetes status.

```
ggplot(data = nh_data_2179,
       aes(x = BMI, y = Pulse,
           color = Diabetes, shape = Diabetes)) +
    geom_point() +
    geom_smooth(method = "loess") +
    labs(title = "BMI vs. Pulse rate (NHANES ages 21-79)") +
    facet_wrap(~ Diabetes)
```

Does Diabetes status affect Pulse-BMI relationship?



The Leek text

Bring at least one (written down) question and/or comment about something in the text that is meaningful to you.

1. What was the most important thing you learned in reading the Leek materials?
2. What was the muddiest, least clear thing?

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

Karl Broman's **excellent presentation on displaying data badly** and related issues.

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 $p = 1.6 \times 10^{-25}$)

General Health Status

Here's a Table of the General Health Status results. This is a self-reported rating of each subject's health on a five point scale (Excellent, Very Good, Good, Fair, Poor.)

```
nh_data_2179 %>%
    select(HealthGen) %>%
    table() %>%
    addmargins()
```

```
.
Excellent      Vgood       Good       Fair       Poor
       67        213        221         80         13
      Sum
      594
```

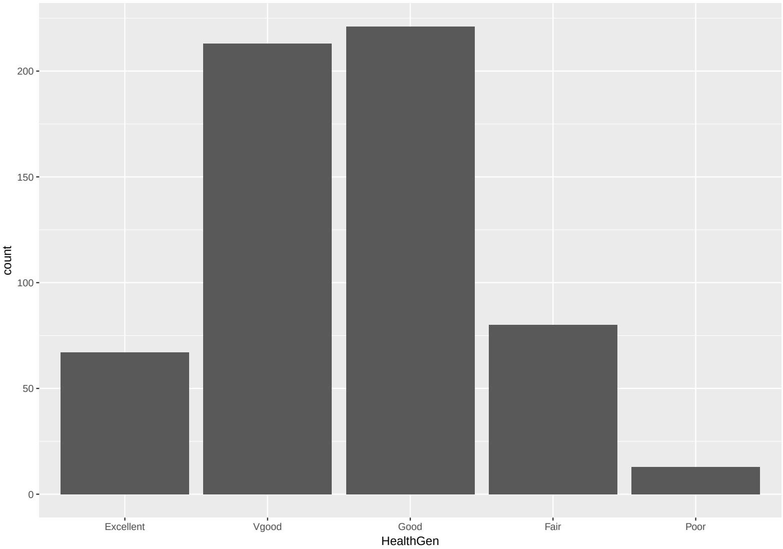
The HealthGen data are categorical, which means that summarizing them with averages isn't as appealing as looking at percentages, proportions and

Bar Chart for Categorical Data

Usually, a **bar chart** is the best choice for a graphing a variable made up of categories.

```
ggplot(data = nh_data_2179, aes(x = HealthGen)) +
    geom_bar()
```

Original Bar Chart of General Health

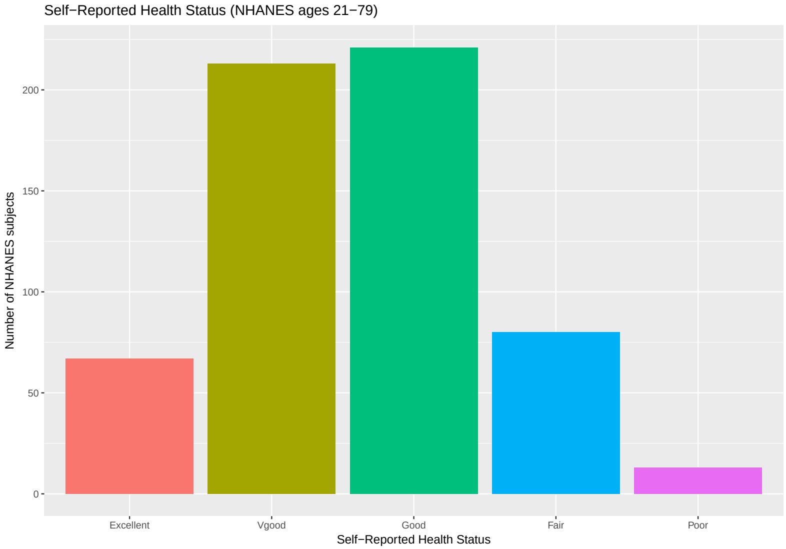


Improving the Bar Chart

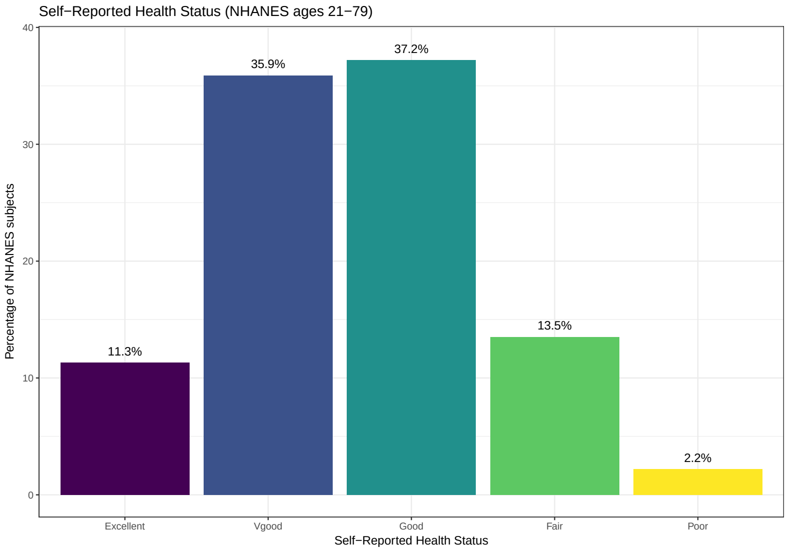
There are lots of things we can do to make this plot fancier.

```
ggplot(data = nh_data_2179,
       aes(x = HealthGen, fill = HealthGen)) +
    geom_bar() +
    guides(fill = FALSE) +
    labs(x = "Self-Reported Health Status",
         y = "Number of NHANES subjects",
         title = "Self-Reported Health Status (NHANES ages 21-
```

The Improved Bar Chart



Or, we can really go crazy. . . (code on next slide)



What crazy looks like. . .

```r
nh_data_2179 %>%
    count(HealthGen) %>%
    ungroup() %>%
    mutate(pct = round(prop.table(n) * 100, 1)) %>%
    ggplot(aes(x = HealthGen, y = pct, fill = HealthGen)) +
    geom_bar(stat = "identity", position = "dodge") +
    scale_fill_viridis(discrete = TRUE) +
    guides(fill = FALSE) +
    geom_text(aes(y = pct + 1,    # nudge above top of bar
                  label = paste0(pct, '%')),  # prettify
              position = position_dodge(width = .9),
              size = 4) +
    labs(x = "Self-Reported Health Status",
         y = "Percentage of NHANES subjects",
         title = "Self-Reported Health Status (NHANES ages 21-
    theme_bw()
```

Working with Tables

We can add a marginal total, and compare subjects by Gender, as follows...

```
nh_data_2179 %>%
    select(Gender, HealthGen) %>%
    table() %>%
    addmargins() %>%
    knitr::kable()
```

	Excellent	Vgood	Good	Fair	Poor	Sum
female	34	107	107	34	8	290
male	33	106	114	46	5	304
Sum	67	213	221	80	13	594

Getting Row Proportions

We'll use prop.table and get the row proportions by feeding it a 1.

```
nh_data_2179 %>%
    select(Gender, HealthGen) %>%
    table() %>%
    prop.table(.,1) %>%
    round(.,2) %>%
    knitr::kable()
```

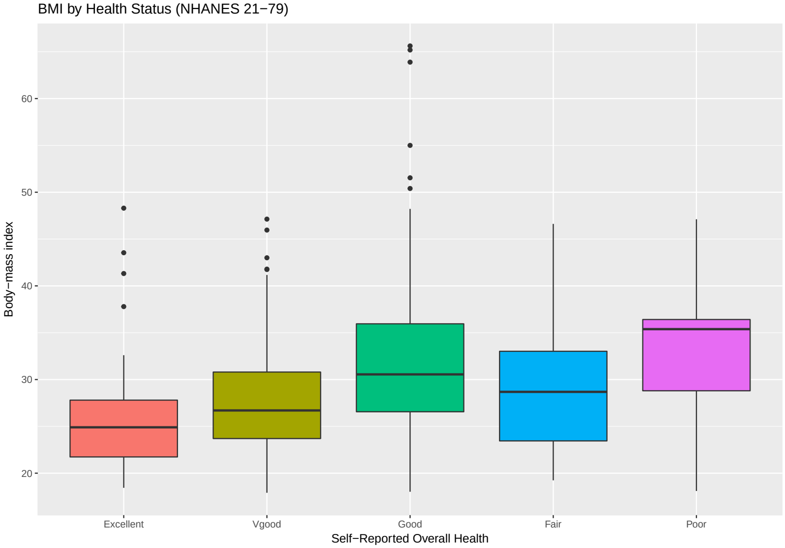
	Excellent	Vgood	Good	Fair	Poor
female	0.12	0.37	0.37	0.12	0.03
male	0.11	0.35	0.38	0.15	0.02

BMI by General Health Status

Let's consider now the relationship between self-reported overall health and body-mass index.

```r
ggplot(data = nh_data_2179,
       aes(x = HealthGen, y = BMI, fill = HealthGen)) +
    geom_boxplot() +
    labs(title = "BMI by Health Status (NHANES 21-79)",
         y = "Body-mass index",
         x = "Self-Reported Overall Health") +
    guides(fill = FALSE)
```

What happens with the `Poor` category?



Not many people self-identify with the Poor category

```
nh_data_2179 %>%
    group_by(HealthGen) %>%
    summarize(count = n(), mean(BMI), median(BMI)) %>%
    knitr::kable()
```

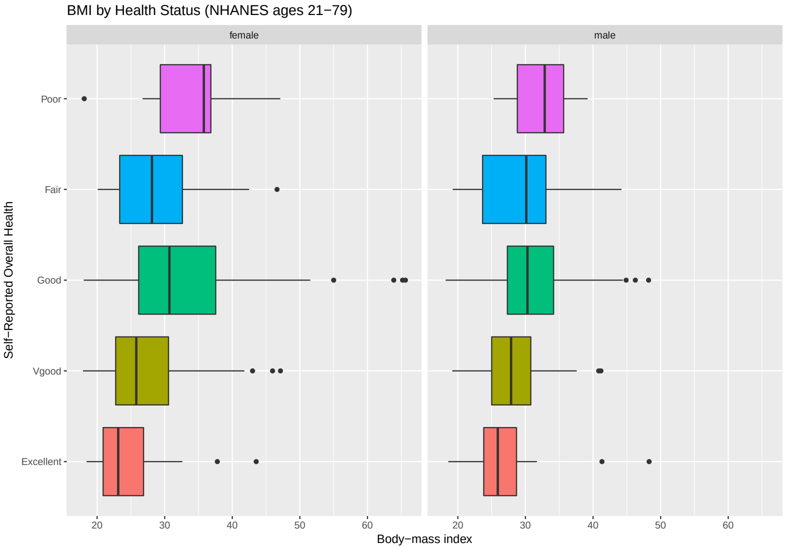
HealthGen	count	mean(BMI)	median(BMI)
Excellent	67	25.70060	24.900
Vgood	213	27.55878	26.700
Good	221	32.00321	30.550
Fair	80	29.28663	28.685
Poor	13	33.08154	35.380

BMI by Gender and General Health Status

We'll start with two panels of boxplots to try to understand the relationships between BMI, General Health Status and Gender. Note the use of coord_flip to rotate the graph 90 degrees.

```r
ggplot(data = nh_data_2179,
       aes(x = HealthGen, y = BMI, fill = HealthGen)) +
    geom_boxplot() +
    labs(title = "BMI by Health Status (NHANES ages 21-79)",
         y = "Body-mass index",
         x = "Self-Reported Overall Health") +
    guides(fill = FALSE) +
    facet_wrap(~ Gender) +
    coord_flip()
```

BMI by Gender and General Health Status Boxplots



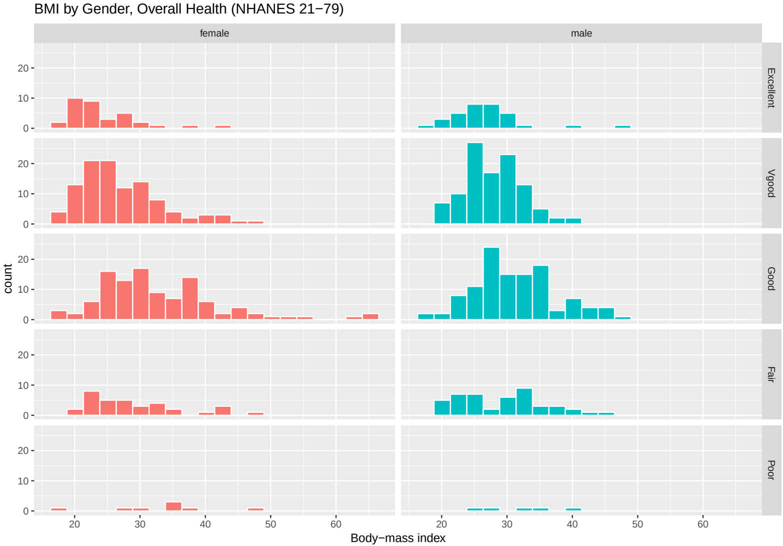
Histograms of BMI by Health and Gender

Here's a plot of faceted histograms, which might be used to address similar questions.

```
ggplot(data = nh_data_2179,
       aes(x = BMI, fill = Gender)) +
    geom_histogram(color = "white", bins = 20) +
    labs(title = "BMI by Gender, Overall Health (NHANES 21-79)
         x = "Body-mass index") +
    guides(fill = FALSE) +
    facet_grid(HealthGen ~ Gender)
```

- Note the new approach with `facet_grid`...

Histograms of BMI by Health and Gender



Conclusions

This is just a small piece of the toolbox for visualizations that we'll create in this class. Many additional tools are on the way, but the main idea won't change. Using the `ggplot2` package, we can accomplish several critical tasks in creating a visualization, including:

- Identifying (and labeling) the axes and titles
- Identifying a type of `geom` to use, like a point, bar or histogram
- Changing fill, color, shape, size to facilitate comparisons
- Building "small multiples" of plots with faceting

Good data visualizations make it easy to see the data, and `ggplot2`'s tools make it relatively difficult to make a really bad graph.

Kidney Cancer Death Rates

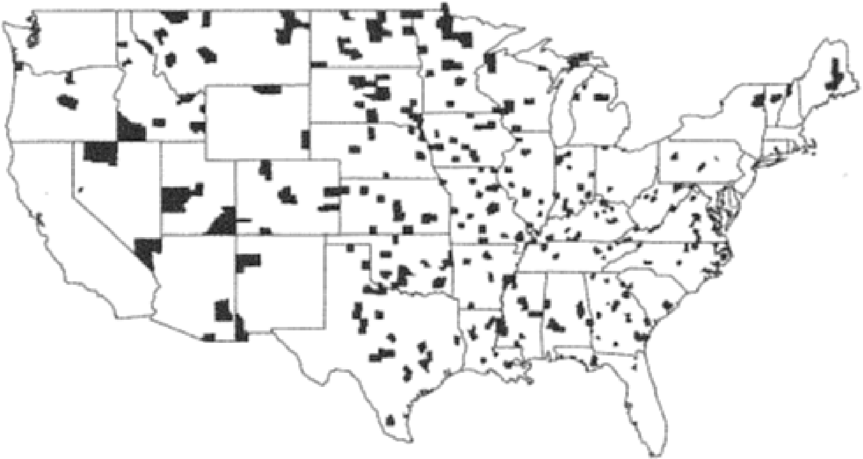
Your map shows U.S. counties.

- The shaded counties are in the top 10% of age-standardized rates for death due to cancer of the kidney/ureter for white males, in 1980-1989.

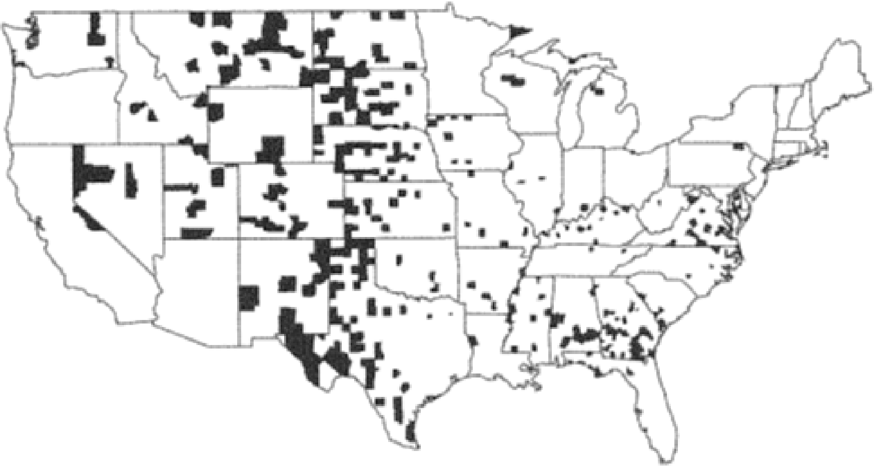
Your Tasks

1. Describe the patterns you see in the map.
2. Speculate as to the cause of these patterns.

Highest kidney cancer death rates



Lowest kidney cancer death rates



What's next?

- More on the NHANES example (Notes Chapters 4-6)
- Read Silver Introduction and Chapter 1 (about 50 denser pages) by 2017-09-12
- Assignment 1 is due 2017-09-15 at noon
- I'll tell you more about the project after Assignment 1.

Notes on the Kidney Cancer example, 1

I first asked you what you noticed about the map, in the hope that someone would point out the obvious pattern, which is that many of the countries in the Great Plains but relatively few near the coasts are shaded.

- Why might that be? Could these be the counties with more old people? Ah, but these rates are age-adjusted.
- They're mostly in rural areas: could the health care there be worse than in major cities? Or perhaps people living in rural areas have less healthy diets, or are exposed to more harmful chemicals? Maybe, but the confusing fact is that the highest 10% and the lowest 10% each show disproportionately higher rates in those Great Plains counties.

Notes on the Kidney Cancer example, 2

- Consider a county with 100 white males. If it has even one kidney death in the 1980s, its rate is 1 per thousand per year, which is among the highest in the nation. If it has no such deaths, its rate will be 0, which is the lowest in the nation.
- The observed rates for smaller counties are *much* more variable, and hence they are more likely to be shaded, even if nothing special is truly going on.
- If a small county has an observed rate of 1 per thousand per year, it's probably random fluctuation. But if a large county (like Cuyahoga) has a very high rate, it is probably a real phenomenon.

Source

My source for this example was Andrew Gelman and Deborah Nolan's book *Teaching Statistics: a bag of tricks* which is the source of a number of things we'll see in the course, including some of the "age guessing" example we've previously done.