431 Class 05

Thomas E. Love

2018-09-11

Today's Agenda

1. Homework 1 review
2. Announcing the Course Project
3. NHANES Example: see **Course Notes** Chapters 3-6
4. Jeff Leek *Elements of Data Analytic Style*
   - Chapter 5 is about Exploratory Analysis
   - Chapter 9 is about Written Analyses
   - Chapter 10 is about Creating Figures
   - Chapter 13 highlights a few matters of form
5. Minute Paper after Class 5

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 2

Due Friday at Noon.

Minute Paper after Class 5

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

What You'll Need Later Today

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.

Course Notes, Chapters 3-6

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

```
library(NHANES)
library(magrittr)
library(tidyverse)
```

What We Did Previously

- Gathered a random sample of 1,000 NHANES subjects into `nh_data`, selecting 10 variables for further study
    - In the Class 4 slides, and in the Course Notes, I used `set.seed(431001)` so I obtained the same sample, and results.
    - Today, I'll switch the seed value, to obtain a new sample called `nh_2`.
- Variables we discuss today: BMI, Pulse, Race1, HealthGen, Diabetes, Gender (also collect ID, Age, Height, Weight).
- Built a subset of that sample who were ages 21-79 and had complete data on those 10 variables, and today, I'll do that again (with my new seed) and call it `nh_3`.
- Built a little code to specify the CWRU colors:

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

Code to create `nh_2` and `nh_3`

```r
set.seed(20180911) # note different seed than we've seen

nh_2 <- sample_n(NHANES, size = 1000) %>%
    select(ID, Gender, Age, Height, Weight, BMI,
           Pulse, Race1, HealthGen, Diabetes)

nh_3 <- nh_2 %>%
    filter(Age > 20 & Age < 80) %>%
    select(ID, Gender, Age, Height, Weight, BMI,
           Pulse, Race1, HealthGen, Diabetes) %>%
    na.omit
```

The nh_3 tibble

```
nh_3
```

```
# A tibble: 588 x 10
      ID Gender   Age Height Weight   BMI Pulse Race1
   <int> <fct>  <int>  <dbl>  <dbl> <dbl> <int> <fct>
 1 64042 male      39   175.   82.6  26.9    66 White
 2 68271 male      43   170.   74.1  25.6    68 White
 3 68630 male      32   166.   78.3  28.3    80 White
 4 70784 female    54   164.  123.   45.7    62 White
 5 66136 male      27   176.   70.6  22.8    72 White
 6 54560 female    72   162.   70.4  26.7    84 White
 7 65177 male      22   180.   73.5  22.8    70 White
 8 58384 male      36   189.  118    33.1    60 White
 9 55993 female    73   171.   65.9  22.5    60 White
10 71793 female    57   154.   82.4  34.6    64 White
# ... with 578 more rows, and 2 more variables:
#   HealthGen <fct>, Diabetes <fct>
```

Some Analyses related to Body-Mass Index in NHANES

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`, `Diabetes`, `Race1`, `Pulse`

1. What is the distribution of BMI in our `nh_3` sample of adults?
2. How does the distribution of BMI vary by whether the subject has been told that they have diabetes?
3. How does the distribution of BMI vary by the subject's Race?
4. What is the association between BMI and the subject's Pulse Rate?
5. Does that BMI-Pulse association differ in subjects who have been told they have diabetes, and those who have 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.

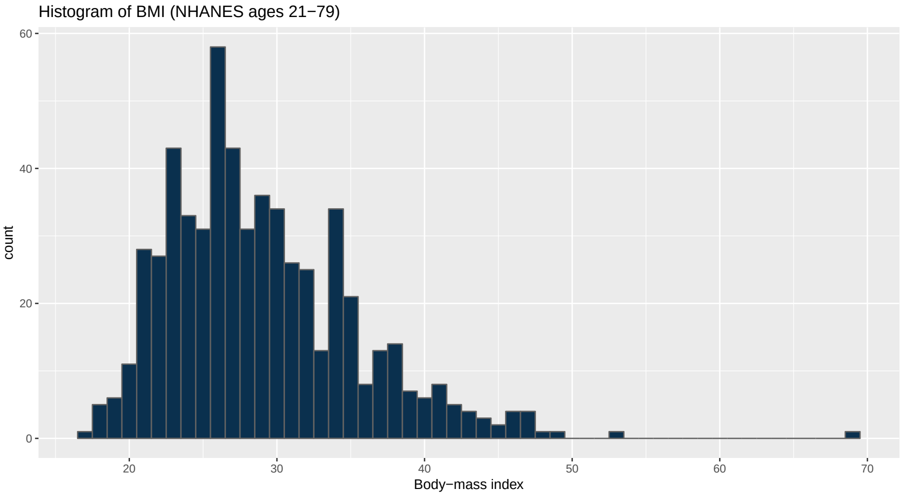
Histogram of BMI with binwidth = 1 (code)

Here's the code for a histogram, again with CWRU colors, for the BMI data in `nh_3`.

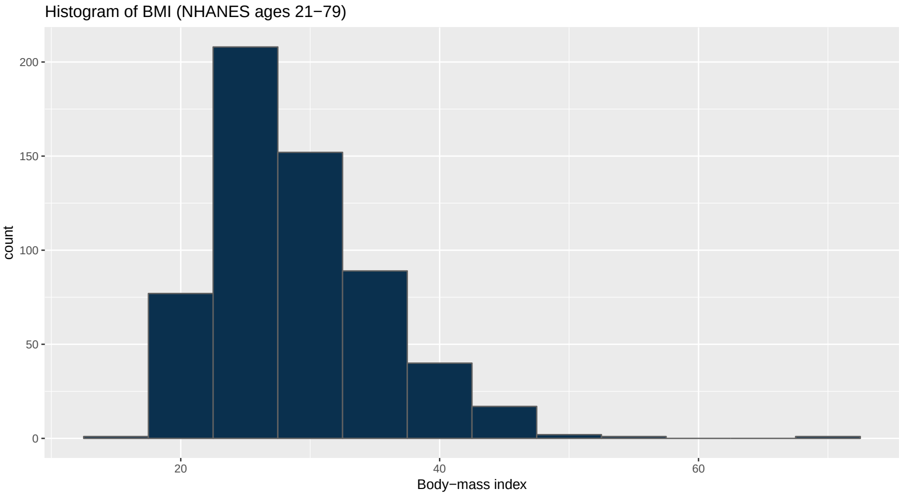
```
ggplot(data = nh_3, aes(x = BMI)) +
    geom_histogram(binwidth = 1,
                   fill = cwru.blue, col = cwru.gray) +
    labs(title = "Histogram of BMI (NHANES ages 21-79)",
         x = "Body-mass index")
```

- I'll set the `binwidth` to be 1 here.
- The `nh_3` data set contains 588 observations.

Histogram of BMI with binwidth = 1



Histogram of BMI with binwidth 5



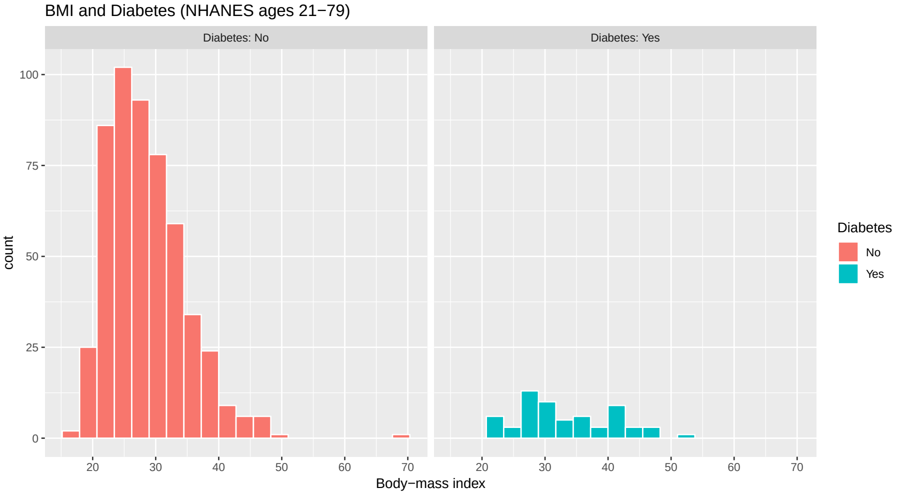
BMI Histograms faceted by Diabetes status (code)

We can facet two histograms of our BMI data based on whether the subjects have been told they have diabetes.

```
ggplot(data = nh_3, aes(x = BMI, fill = Diabetes)) +
    geom_histogram(bins = 20, col = "white") +
    labs(title = "BMI and Diabetes (NHANES ages 21-79)",
         x = "Body-mass index") +
    facet_wrap(~ Diabetes, labeller = "label_both")
```

- We've let the fill of the bars change depending on diabetes status.
- We've set the number of bins to be 20 in each plot, rather than specifying the binwidth.
- We added an argument `labeller = "label_both"` to our `facet_wrap` request which will get the machine to specify the name of the variable we're using to facet the data as well as its values.

BMI Histograms faceted by Diabetes status



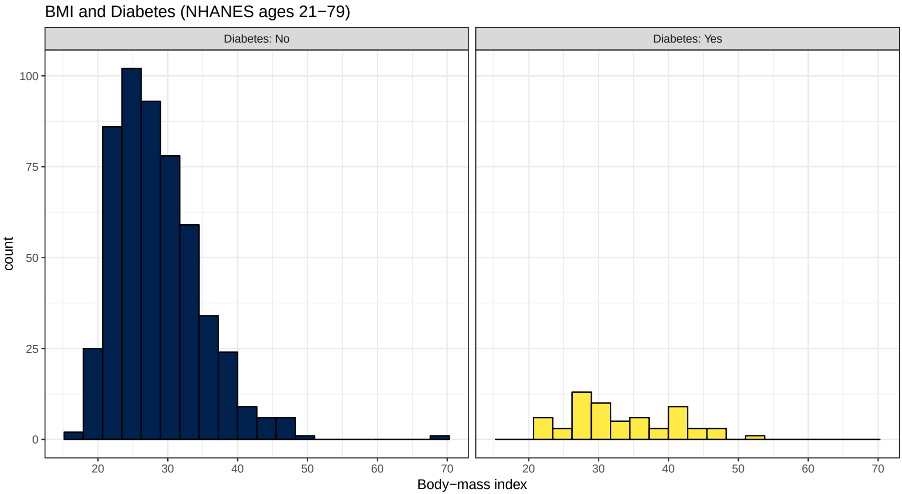
- Do we need this legend for the fill?
- To drop it, we would add `guides(fill = FALSE)`.

BMI Histograms by Diabetes status, II. (code)

Let's drop the legend and also change the fill scheme to something more appropriate for color-blind folks.

```r
ggplot(data = nh_3, aes(x = BMI, fill = Diabetes)) +
    geom_histogram(bins = 20, col = "black") +
    scale_fill_viridis_d(option = "cividis") +
    guides(fill = FALSE) +
    labs(title = "BMI and Diabetes (NHANES ages 21-79)",
         x = "Body-mass index") +
    facet_wrap(~ Diabetes, labeller = "label_both") +
    theme_bw()
```

BMI Histograms faceted by Diabetes status, II.



Numerical Summaries: BMI, by Diabetes Status

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

```
nh_3 %>%
    group_by(Diabetes) %>%
    summarize(count = n(), mean(BMI), median(BMI))
```

```
# A tibble: 2 x 4
  Diabetes count `mean(BMI)` `median(BMI)`
  <fct>    <int>       <dbl>         <dbl>
1 No         526        28.5          27.4
2 Yes         62        33.1          31.4
```

Numerical Summaries: BMI, by Diabetes Status, II

Neatening up the presentation a little bit, with some rounding and the `kable` function from the `knitr` package…

```
nh_3 %>%
    group_by(Diabetes) %>%
    summarize("Count" = n(),
              "Mean(BMI)" = round(mean(BMI),2),
              "Median(BMI)" = median(BMI)) %>%
    knitr::kable()
```

Diabetes	Count	Mean(BMI)	Median(BMI)
No	526	28.54	27.40
Yes	62	33.10	31.44

BMI by Race

How many people fall into each of the available Race1 categories, and what can we learn about "average" BMI in those groups?

```
nh_3 %>%
    group_by(Race1) %>%
    summarize(count = n(), mean(BMI), median(BMI)) %>%
    knitr::kable()
```

Race1	count	mean(BMI)	median(BMI)
Black	63	32.15159	31.860
Hispanic	39	29.37205	29.360
Mexican	34	29.91235	29.225
White	412	28.53252	27.380
Other	40	28.03725	26.515

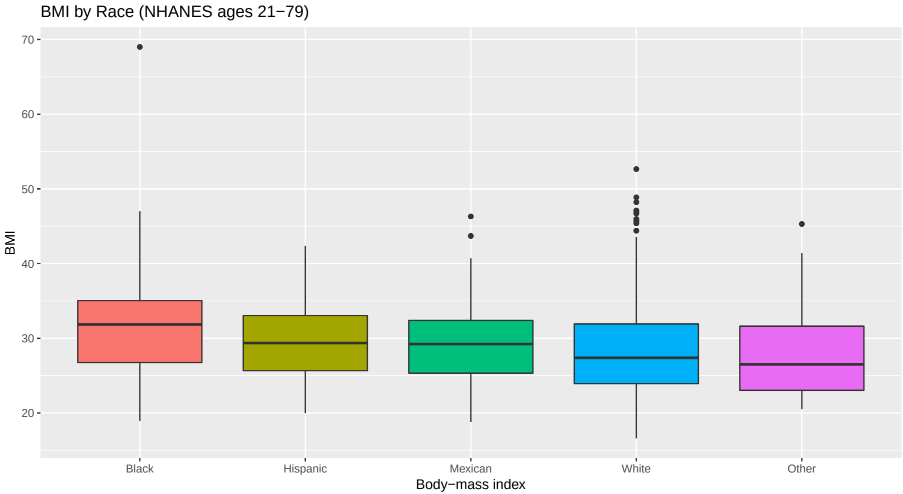
BMI and Race Comparison Boxplot (code)

Let's consider a plot to compare the distribution of BMI across the five available levels of Race1.

- It would be helpful to think in advance about what you expect to see here...

```
ggplot(data = nh_3,
       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 Comparison Boxplot



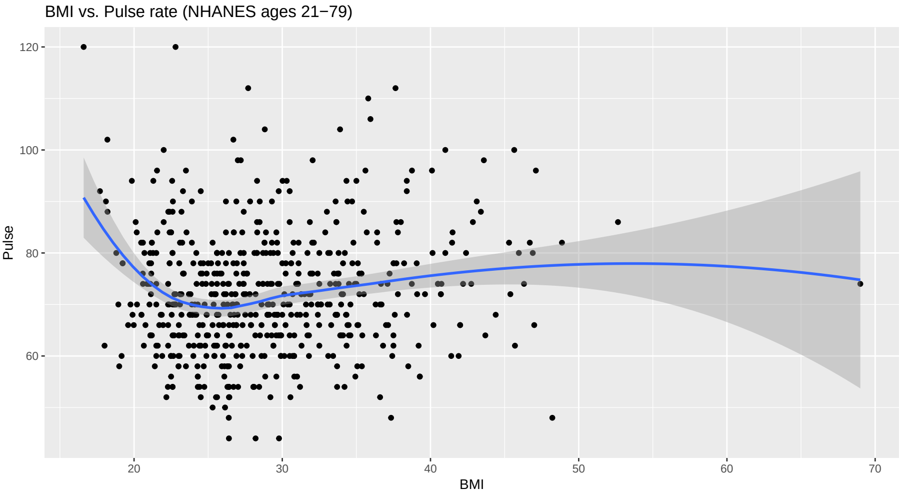
BMI and Pulse Rate Scatterplot (code)

Now, we'll look at the association between Pulse rate and BMI, and we'll fit a "loess" smooth curve to help us see the "center" of that association.

- Do you think people with higher BMI will have lower or higher Pulse rates?

```
ggplot(data = nh_3, 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 Scatterplot



Correlation Coefficient: Summarizing 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.

```
nh_3 %$% cor(BMI, Pulse)
```

```
[1] 0.09169249
```

- 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 very close to zero. This implies we have almost no linear association in this case, across the entire sample.

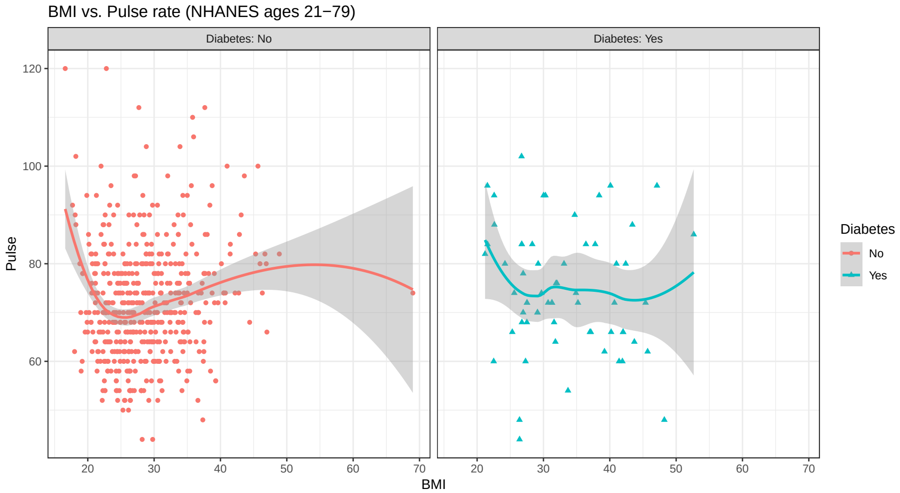
Does Diabetes affect Pulse-BMI association? (code)

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_3,
       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, labeller = "label_both") +
    theme_bw()
```

Does Diabetes status affect Pulse-BMI association?



Correlation of BMI and Pulse by Diabetes?

- Recall that the correlation coefficient for the relationship between BMI and Pulse in the full sample was quite close to zero.
  - Specifically, it was 0.0917
- Grouped by diabetes status, do we get a different story?

```
nh_3 %>%
  group_by(Diabetes) %>%
  summarize(cor(BMI, Pulse))
```

```
# A tibble: 2 x 2
  Diabetes `cor(BMI, Pulse)`
  <fct>                 <dbl>
1 No                    0.108
2 Yes                  -0.113
```

Working with a Categorical Outcome (Self-Reported General Health) in NHANES

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.)

```r
nh_3 %>%
    select(HealthGen) %>%
    table() %>%
    addmargins()
```

```
.
Excellent      Vgood       Good       Fair       Poor
       69        206        223         76         14
      Sum
      588
```

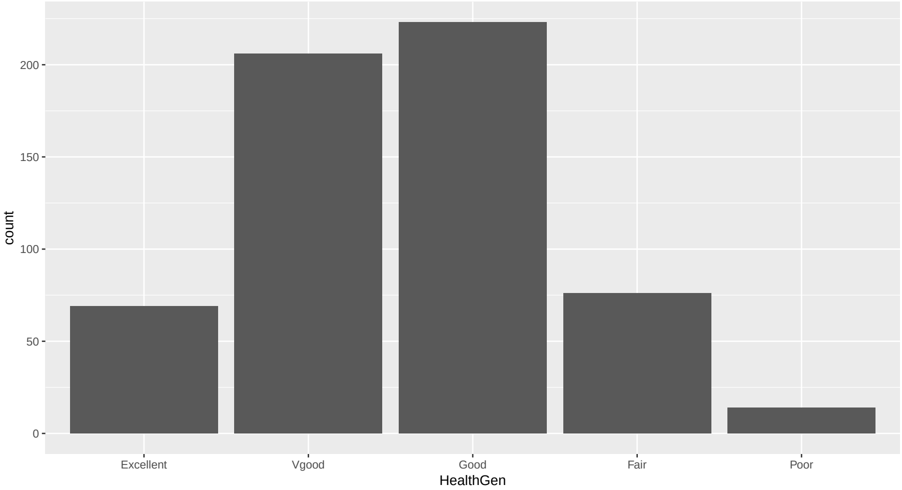
The HealthGen data are categorical, which means that summarizing them with averages isn't as appealing as looking at percentages, proportions and rates.

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_3, aes(x = HealthGen)) +
    geom_bar()
```

Original Bar Chart of General Health

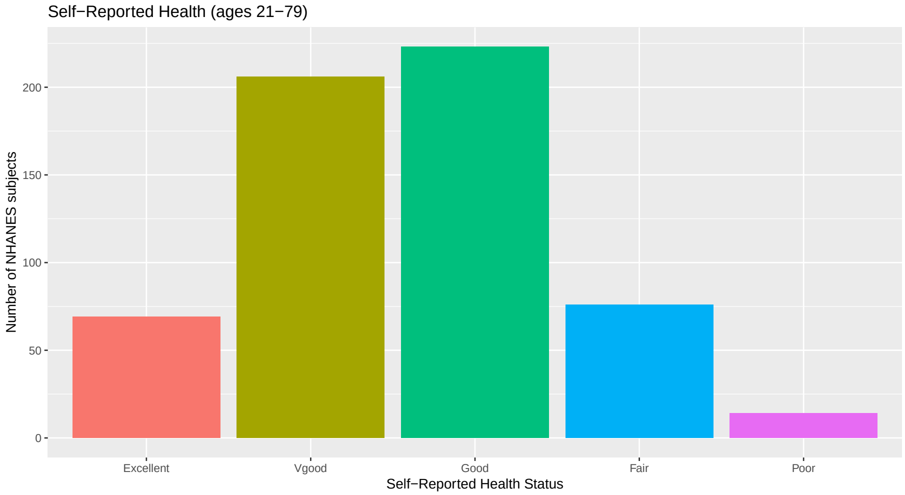


Improving the Bar Chart

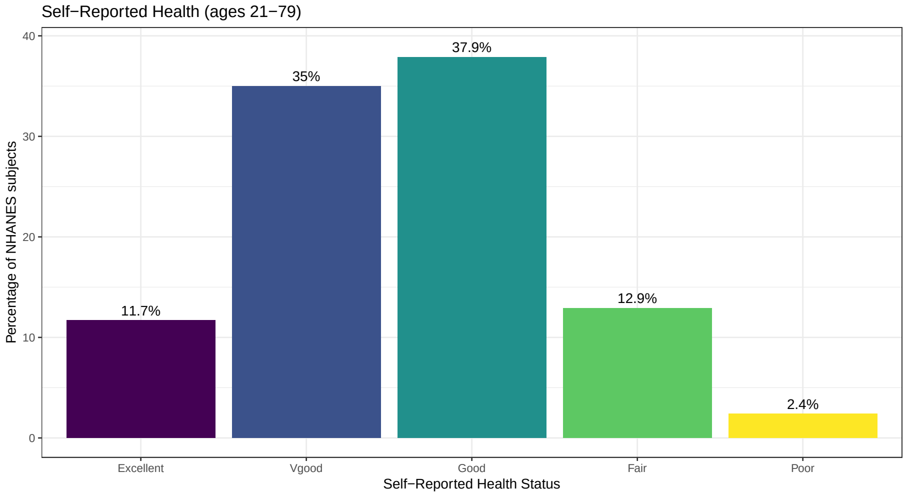
There are lots of things we can do to make this plot fancier.

```r
ggplot(data = nh_3,
       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 (ages 21-79)")
```

The Improved Bar Chart



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



What crazy looks like. . .

```r
nh_3 %>%
    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_d() +
    guides(fill = FALSE, col = 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 (ages 21-79)") +
    theme_bw()
```

Working with Tables

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

```
nh_3 %>%
    select(Gender, HealthGen) %>%
    table() %>%
    addmargins() %>%
    knitr::kable()
```

	Excellent	Vgood	Good	Fair	Poor	Sum
female	39	116	100	33	8	296
male	30	90	123	43	6	292
Sum	69	206	223	76	14	588

Getting Row Proportions

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

```
nh_3 %>%
    select(Gender, HealthGen) %>%
    table() %>%
    prop.table(.,1) %>%
    round(.,2) %>%
    knitr::kable()
```

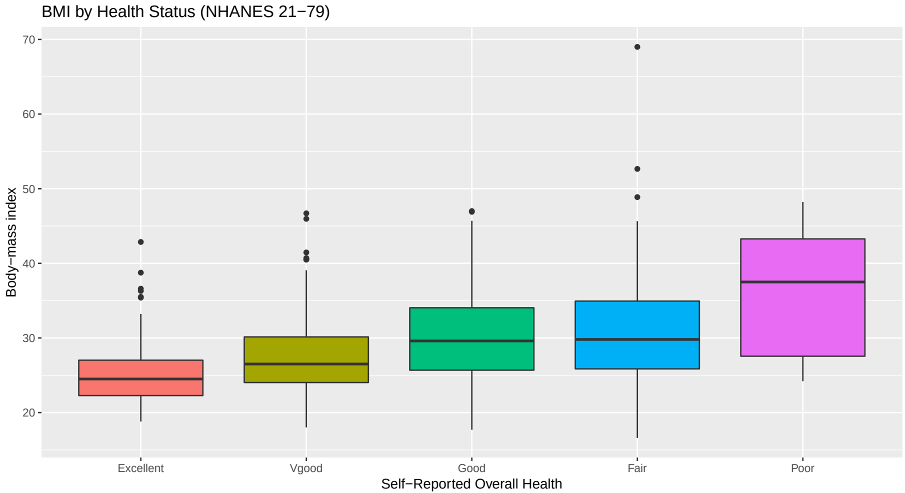
	Excellent	Vgood	Good	Fair	Poor
female	0.13	0.39	0.34	0.11	0.03
male	0.10	0.31	0.42	0.15	0.02

BMI by General Health Status

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

```
ggplot(data = nh_3,
       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?



Summary Table of BMI distribution by HealthGen

```
nh_3 %>%
    group_by(HealthGen) %>%
    summarize("BMI n" = n(),
              "Mean" = round(mean(BMI),1),
              "SD" = round(sd(BMI),1),
              "min" = round(min(BMI),1),
              "Q25" = round(quantile(BMI, 0.25),1),
              "median" = round(median(BMI),1),
              "Q75" = round(quantile(BMI, 0.75),1),
              "max" = round(max(BMI),1)) %>%
    knitr::kable()
```

- Resulting table is shown in the next slide.

Not many self-identify in the `Poor` category

HealthGen	BMI n	Mean	SD	min	Q25	median	Q75	max
Excellent	69	25.5	4.9	18.8	22.3	24.5	27.0	42.9
Vgood	206	27.7	5.2	18.0	24.0	26.5	30.1	46.7
Good	223	30.1	5.9	17.7	25.7	29.6	34.0	47.0
Fair	76	31.2	8.7	16.6	25.9	29.8	34.9	69.0
Poor	14	36.7	8.5	24.2	27.6	37.5	43.3	48.2

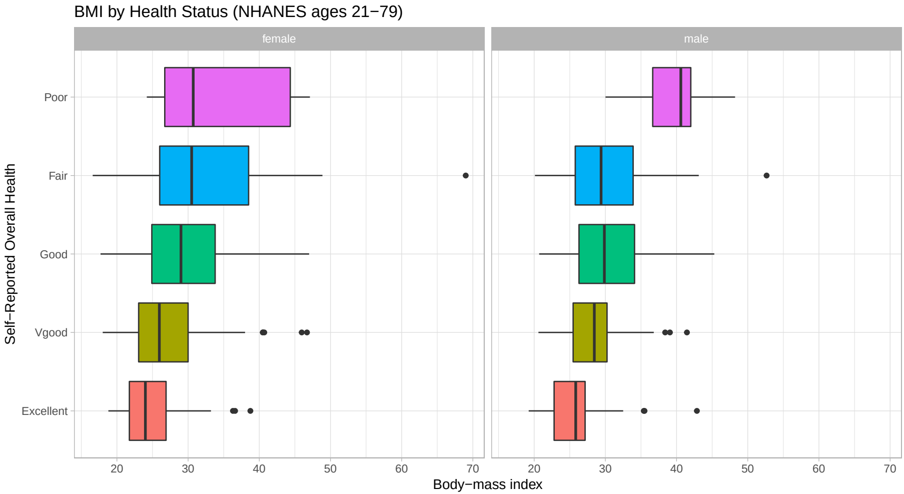
BMI by Gender and General Health Status

We'll start with two panels of boxplots to try to understand the relationships between BMI, General Health Status and Gender.

```
ggplot(data = nh_3,
       aes(x = HealthGen, y = BMI, fill = HealthGen)) +
    geom_boxplot() +
    guides(fill = FALSE) +
    facet_wrap(~ Gender) +
    coord_flip() +
    theme_light() +
    labs(title = "BMI by Health Status (NHANES ages 21-79)",
         y = "Body-mass index",
         x = "Self-Reported Overall Health")
```

- Note the use of `coord_flip` to rotate the graph 90 degrees.
- Note the use of a new theme, called `theme_light()`.

BMI by Gender and General Health Status Boxplots



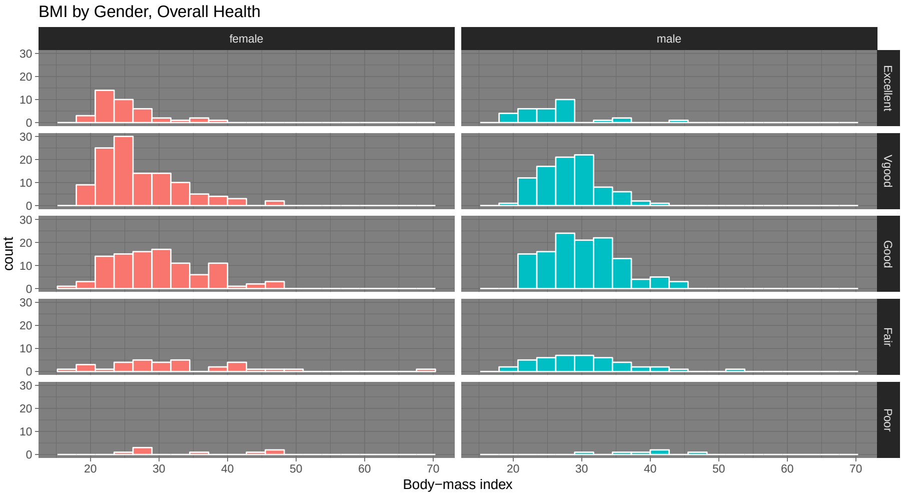
Histograms of BMI by Health and Gender

Here are doubly faceted histograms, which can help address similar questions.

```
ggplot(data = nh_3,
       aes(x = BMI, fill = Gender)) +
    geom_histogram(color = "white", bins = 20) +
    labs(title = "BMI by Gender, Overall Health",
         x = "Body-mass index") +
    guides(fill = FALSE) +
    facet_grid(HealthGen ~ Gender) +
    theme_dark()
```

- Note the use of `facet_grid` to specify rows and columns.
- Note the use of a new theme, called `theme_dark()`.

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.

The Elements of Data Analytic Style

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 at https://github.com/kbroman/Talk_Graphs may be helpful…

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 2

Due Friday at Noon.

Minute Paper after Class 5

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