431 Class 06

Thomas E. Love

2018-09-13

Today's Agenda

1. Announcements, including Minute Papers and Homework 1 Review
2. Who Wrote the Anti-Trump *New York Times* op-ed: R in Action
3. *Elements of Data Analytic Style* - What was the most useful thing?
4. Kidney Cancer Maps
5. Visualizing NHANES: **Course Notes** Chapters 3-6

From David Robinson at varianceexplained.org

VARIANCE EXPLAINED

ABOUT ME POSTS LEARN R TEXT MINING IN R INTRODUCTION TO EMPIRICAL BAYES

David Robinson

Chief Data Scientist at DataCamp, works in R and Python.

- Email
- Twitter
- Github
- Stack Overflow

Subscribe

Your email

Subscribe to this blog

Recommended Blogs

- DataCamp

Who wrote the anti-Trump New York Times op-ed? Using tidytext to find document similarity

Like a lot of people, I was intrigued by "I Am Part of the Resistance Inside the Trump Administration", an anonymous New York Times op-ed written by a "senior official in the Trump administration". And like many data scientists, I was curious about what role text mining could play.

> **Drew Conway** @drewconway
>
> Ok NLP people, now's your chance to shine. Just spitballing here but TF-IDF on "the op-ed" compared to the published writing of every senior Trump admin official? I want likelihood estimates with standard errors. GO!
>
> 7:01 PM - Sep 5, 2018
>
> 147 34 people are talking about this

This is a useful opportunity to demonstrate how to use the tidytext package that Julia Silge and I developed, and in particular to apply three methods:

- Using TF-IDF to find words specific to each document (examined in more detail in Chapter 3 of our book)
- Using widyr to compute pairwise cosine similarity
- How to make similarity interpretable by breaking it down by word

Since my goal is R education more than it is political analysis, I show all the code in the post.

Two Tasks To Start Class

Task 1: Elements of Data Analytic Style

Last time, I asked you to write down (so that you can share) the most important/interesting/surprising thing you learned from reading the four chapters of Jeff Leek's *Elements of Data Analytic Style*.

Form a group of about 5 people. We'll need 10 groups. Your group will be identified by the folder you receive, as Group A, B, . . . , J. Now, as a group,

1. Have everyone read out their statement
2. Identify **two** statements that you are willing to share, as a group.
3. Have one person from your group type those statements into the Google Form at http://bit.ly/431-2018-class6-leek.

When you finish this Task, move on to Task 2 (next slide)

Task 2: 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

Return to NHANES

Return to NHANES – As before. . .

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

set.seed(20180911) # note same seed as Class 5

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
```

Numerical Summary of BMI by Gender: 1/3

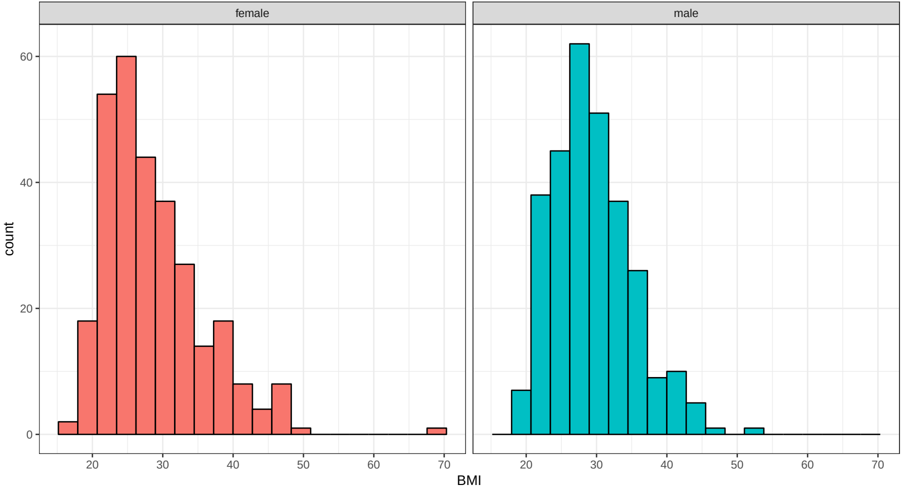
```
nh_3 %>%
    group_by(Gender) %>%
    summarize("Count" = n(), "Mean BMI" = mean(BMI),
              "Skew1" = (mean(BMI) - median(BMI))/sd(BMI))
```

```
# A tibble: 2 x 4
  Gender Count `Mean BMI` Skew1
  <fct>  <int>      <dbl> <dbl>
1 female   296       28.6 0.261
2 male     292       29.5 0.137
```

- If skew1 < 0.2, we rarely infer anything but symmetry.
- If skew1 > 0.2, we might infer substantial skew, but DTDP.

Skew1 and A Relevant Picture (code on next slide)

Do the female data appear skewed? Do the male data?



Skew1 and A Relevant Picture (code)

```
ggplot(nh_3, aes(x = BMI, fill = Gender)) +
  geom_histogram(bins = 20, col = "black") +
  guides(fill = FALSE) +
  theme_bw() +
  facet_wrap(~ Gender)
```

Numerical Summary of BMI by Gender: 2/3

```
nh_3 %>%
    filter(Gender == "female") %$%
    summary(BMI)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.60   23.48   26.73   28.56   32.04   69.00
```

```
nh_3 %>%
    filter(Gender == "male") %$%
    psych::describe(BMI)
```

```
   vars   n  mean   sd median trimmed  mad   min   max
X1    1 292 29.49 5.74   28.7   29.06 5.49 19.23 52.65
   range skew kurtosis   se
X1 33.42 0.74     0.61 0.34
```

Numerical Summary of BMI by Gender: 3/3

```
mosaic::favstats(BMI ~ Gender, data = nh_3)
```

```
  Gender   min     Q1 median      Q3   max     mean
1 female 16.60 23.475  26.73 32.0425 69.00 28.56182
2   male 19.23 25.600  28.70 33.2950 52.65 29.48860
        sd   n missing
1 7.011839 296       0
2 5.736194 292       0
```

BMI by Gender and Diabetes status? 1/2

```
nh_3 %>%
    group_by(Gender, Diabetes) %>%
    summarize("Count" = n(),
              "skew1" = (mean(BMI) - median(BMI))/sd(BMI),
              mean(BMI), median(BMI))
```

```
# A tibble: 4 x 6
# Groups:   Gender [?]
  Gender Diabetes Count skew1 `mean(BMI)` `median(BMI)`
  <fct>  <fct>    <int> <dbl>       <dbl>         <dbl>
1 female No         269 0.254        28.1          26.4
2 female Yes         27 0.123        32.8          31.9
3 male   No         257 0.109        29.0          28.4
4 male   Yes         35 0.259        33.3          31.3
```

BMI by Gender and Diabetes status? 2/2

```r
mosaic::favstats(BMI ~ Gender + Diabetes, data = nh_3)
```

```
  Gender.Diabetes   min     Q1 median    Q3   max
1       female.No 16.60 23.300  26.40 31.10 69.00
2         male.No 19.23 25.500  28.40 32.60 45.30
3      female.Yes 21.20 26.730  31.90 39.00 47.11
4        male.Yes 21.55 27.515  31.28 39.65 52.65
      mean       sd   n missing
1 28.13249 6.812739 269       0
2 28.96914 5.200491 257       0
3 32.83926 7.649917  27       0
4 33.30286 7.808263  35       0
```

This is clearly the easiest approach, in my view.

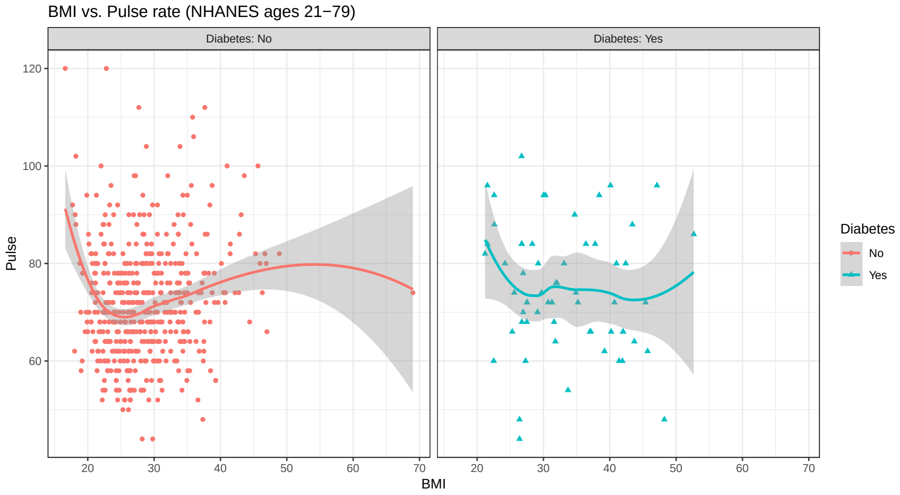
Does Diabetes affect Pulse-BMI association? (code)

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.)

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

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

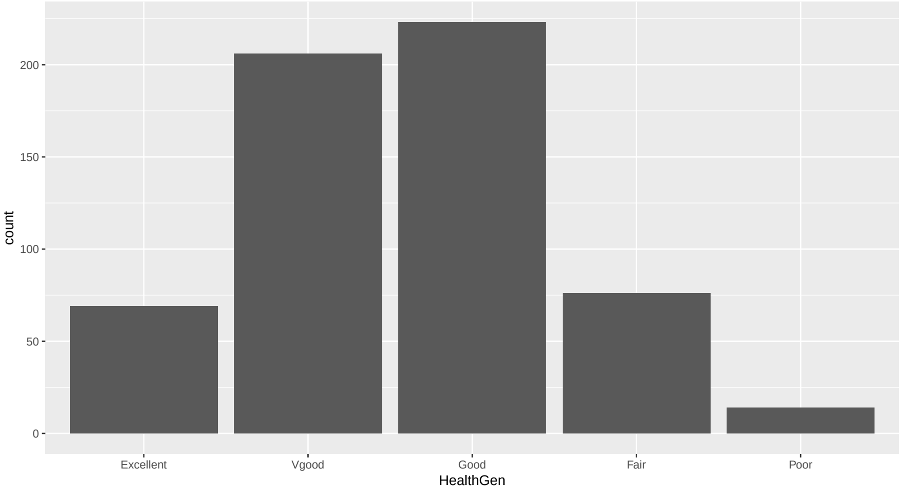
The `HealthGen` data are categorical, which means that summarizing them with averages isn't as appealing as looking at percentages, proportions and rates.

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

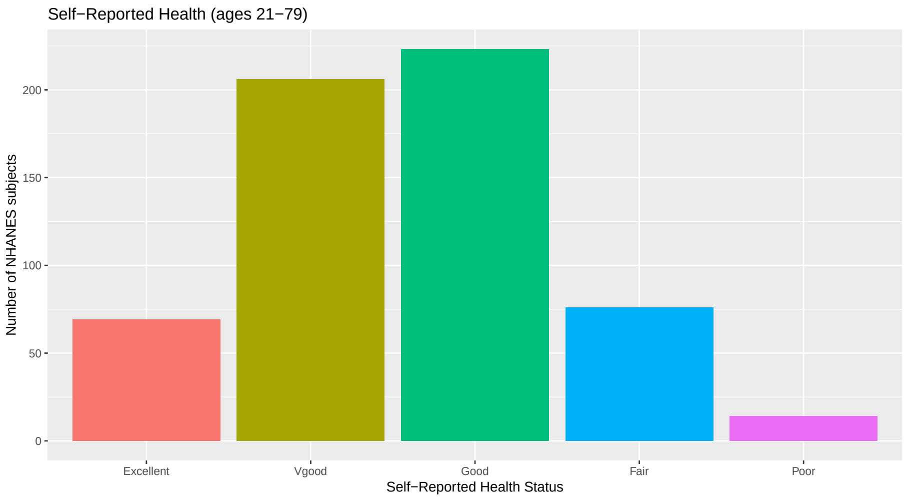


Improving the Bar Chart

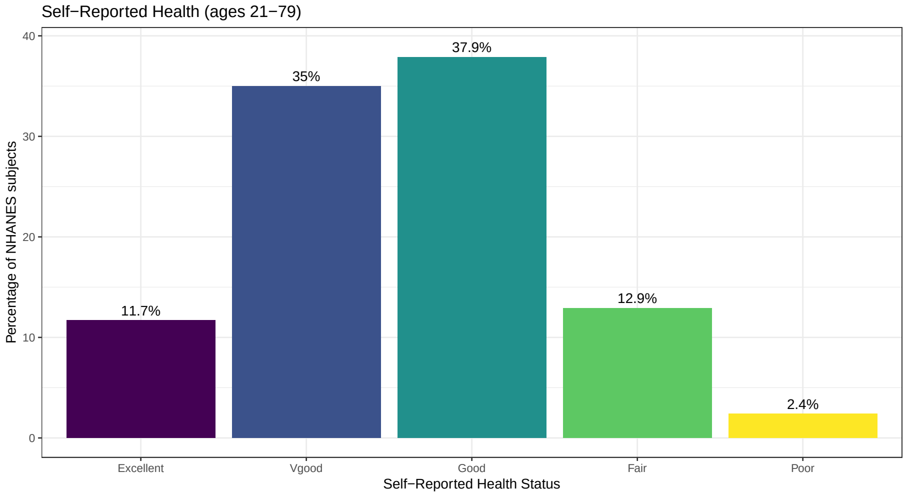
There are lots of things we can do to make this plot fancier.

```
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. . .

```
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()
```

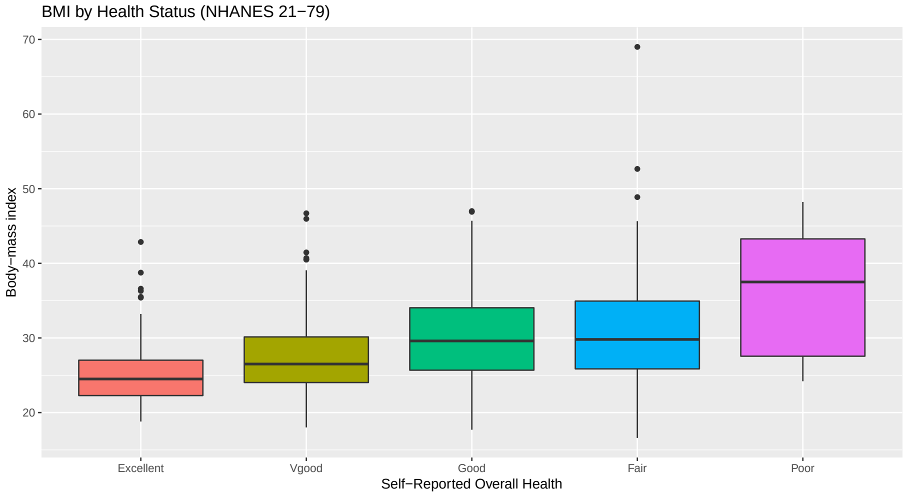
	Excellent	Vgood	Good	Fair	Poor
female	0.13	0.39	0.34	0.11	0.03
male	0.10	0.31	0.42	0.15	0.02

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

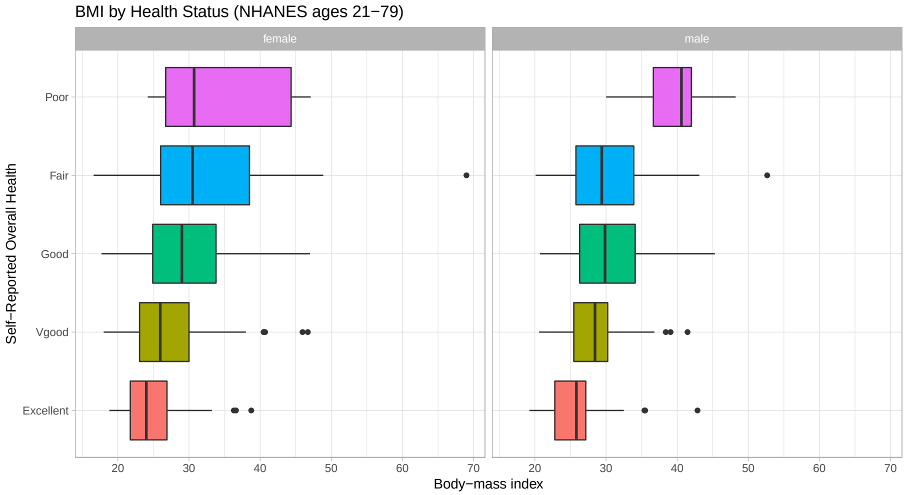
BMI by Gender and General Health Status

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



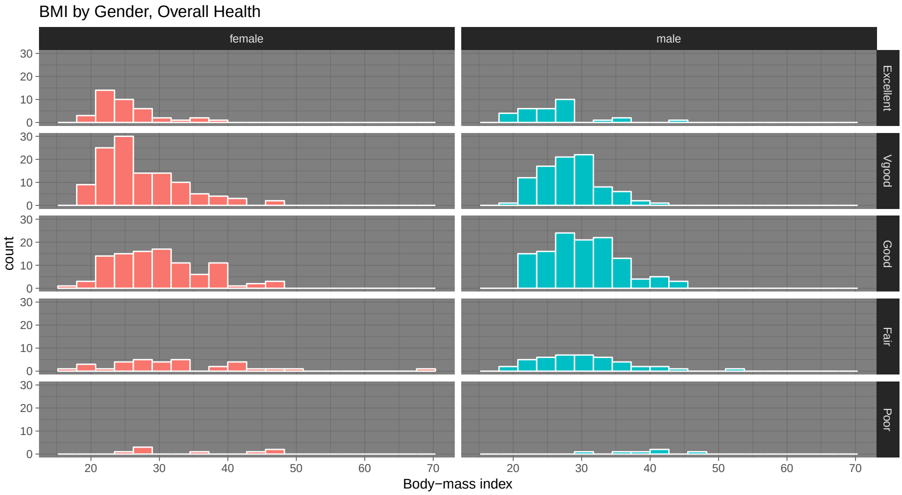
Histograms of BMI by Health and Gender

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.

Reminders

The Course Project

Take a look at the web site. We'll start working on the project in class 2018-09-25.

Homework 2

Due Friday at Noon.

The Signal and the Noise

Please read the Introduction and Chapter 1 before Tuesday's class

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.