431 Class 05

Thomas E. Love

2017-09-12

I know of no person or group that is taking nearly adequate advantage of the graphical potentialities of the computer.

-- John Tukey

The Elements of Data Analytic Style

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

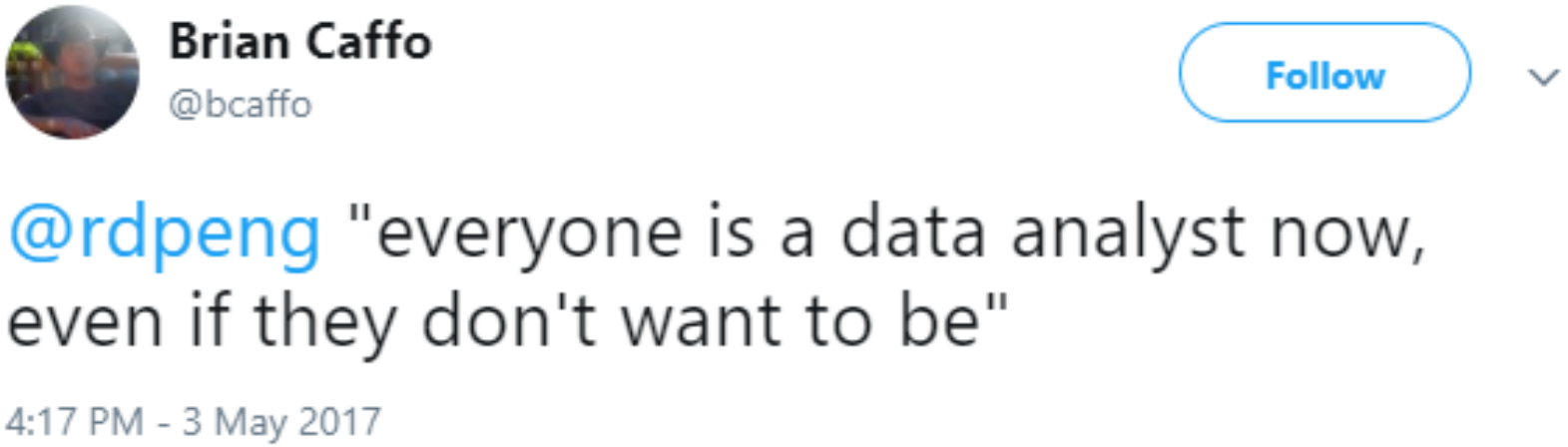
As a group, I'm looking for you to address:

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

Past Feedback on these chapters in Leek

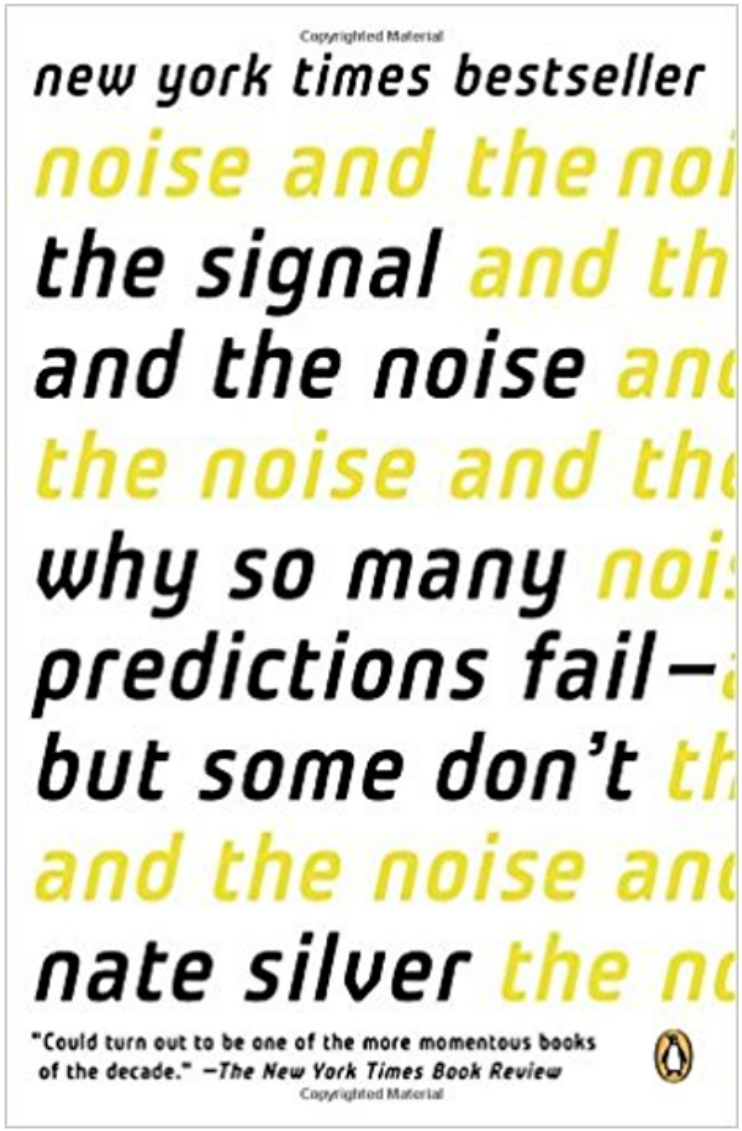
1. Plots are (always) better than (numerical) summaries for exploring data. Plotting more of the data allows you to identify outliers, confounders, missing data, relationships and correlations far more easily than summary measures do. Looking at the data is a much better way to understand what's really going on than just looking at numerical summaries, which can easily be deceiving. Explore the data before jumping to a statistical test. (endorsed x6)
2. Plot as much of the actual data as you can. Overlaying the full data on top of a summary (like a boxplot) is useful. (endorsed x4)
3. "Any strong pattern in a data set should be checked for confounders and alternative explanations." (endorsed x3)
4. "A common failure, particularly when using automated software, is to immediately apply statistical testing procedures and to look for statistical significance without first exploring the data." (endorsed x3)
5. Boxplots and bar charts are better than pie charts for making comparisons. Don't use pie charts. (endorsed x3)

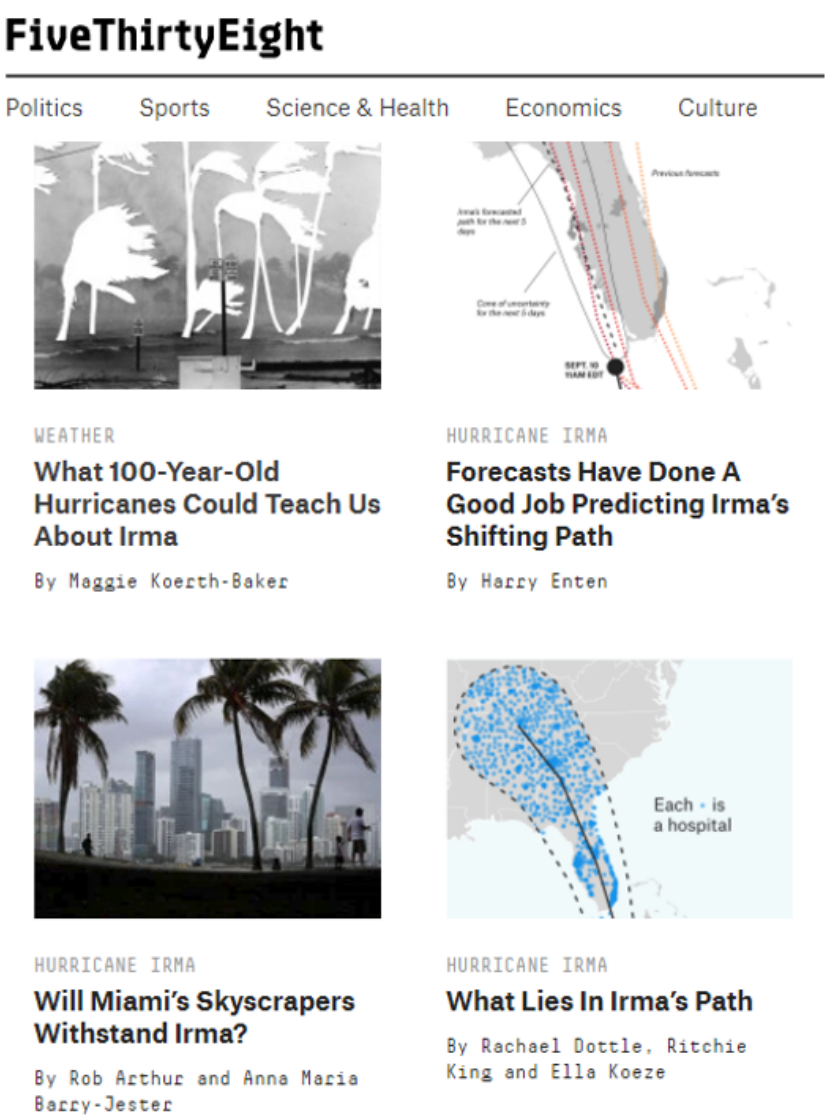
Remember. . .



https://twitter.com/bcaffo/status/859864563218620420

The Signal and The Noise





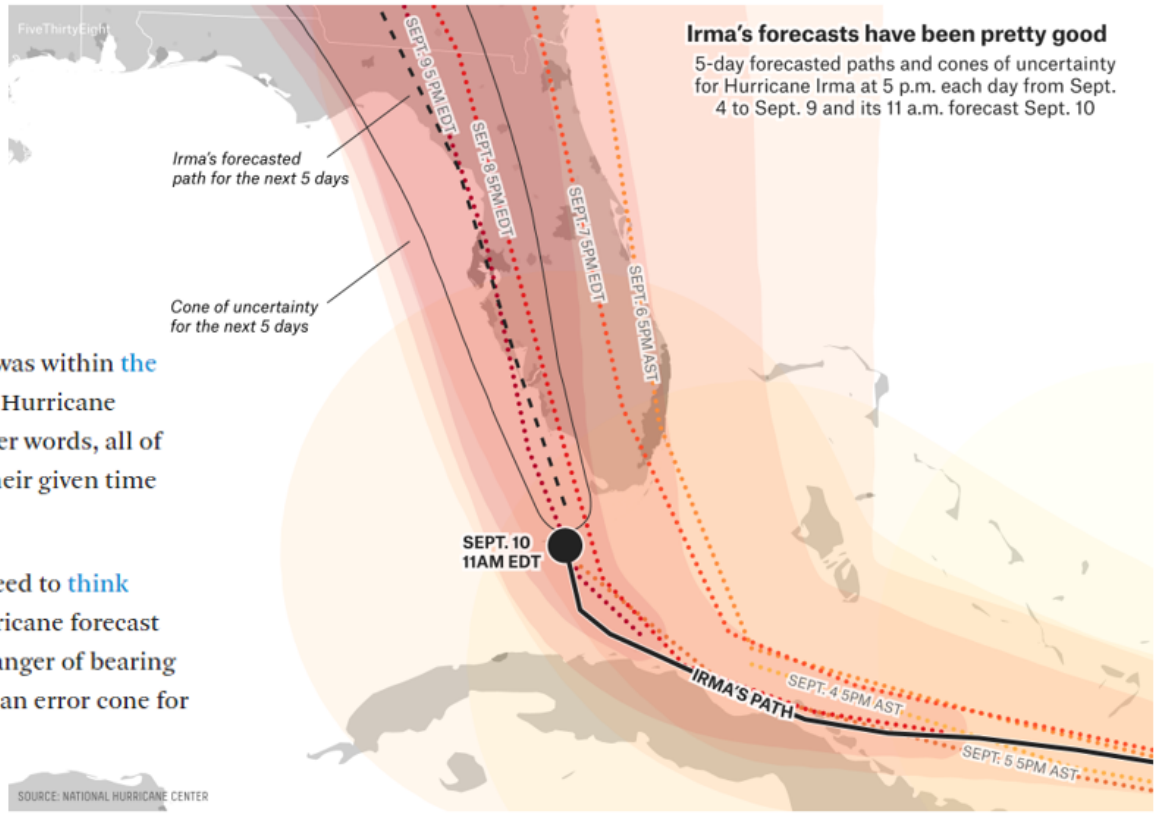
Why are we reading *The Signal and the Noise*?

- It provides fodder for Assignment 2, question 1, and many other assignments.
- It provides an excellent introduction to the Bayesian approach to thinking about combining probabilities and estimating uncertainty, topics I cover too lightly in this course.
- It also provides an excellent introduction to forecasting, and, especially how we might evaluate forecasts better, which is another topic I cover too lightly in the course notes, at least.
- It will (I hope) spark your interest in any number of issues you can tackle using tools from this course.
- Nate runs the fivethirtyeight.com web site, which you might want to start perusing in spare moments.

Forecasts of Irma's Shifting Path (2017-09-10)

Forecasts Have Done A Good Job Predicting Irma's Shifting Path

By Harry Enten
Filed under Hurricane Irma
Published Sep. 10, 2017



Despite these shifts, Irma's landfall in the continental U.S. was within the error cone of every single 5 p.m. forecast from the National Hurricane Center. (This cone is based on historical error rates.) In other words, all of these forecasts were within the expected range of error at their given time interval.

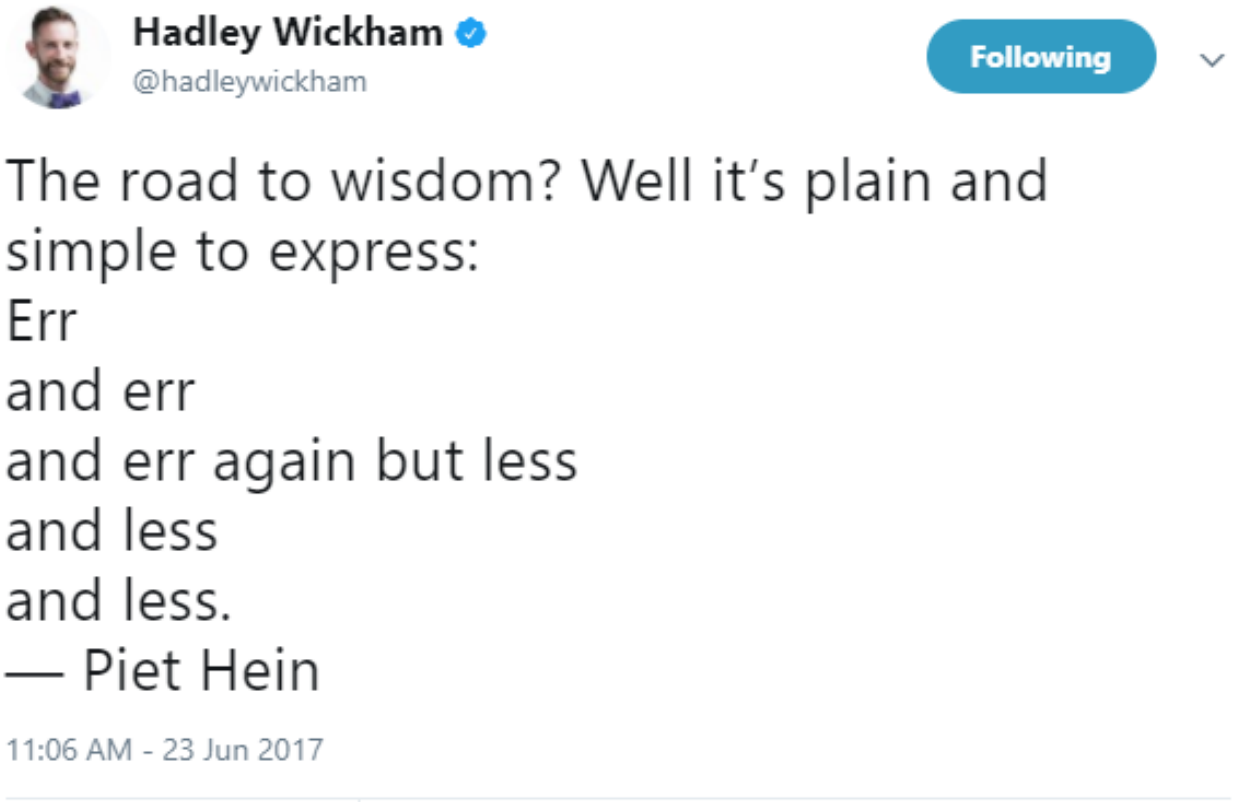
The case of Irma provides yet another example of why we need to think probabilistically about forecasting. The center line of a hurricane forecast cannot be taken as gospel. Western Florida was always in danger of bearing the brunt of Irma. The National Hurricane Center provides an error cone for a reason.

Using 431-help at case dot edu

To answer your R questions, we need to be able to replicate your work:

1. send your entire R Markdown (.Rmd) file as an attachment
2. in the body of your email, ask your question
3. if you're getting an error message, include it in the body of your email, or attach a screenshot.

The Road to Wisdom . . .



https://twitter.com/hadleywickham/status/878267930651140097

Data and Package Loading for New NHANES Example

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

nh_temp <- NHANES %>%
    filter(SurveyYr == "2011_12") %>%
    filter(Age >= 21 & Age < 65) %>%
    mutate(Sex = Gender, Race = Race3,
           SBP = BPSysAve, DBP = BPDiaAve) %>%
    select(ID, Sex, Age, Race, Education,
           BMI, SBP, DBP, Pulse, PhysActive,
           Smoke100, SleepTrouble, HealthGen)
```

Random Sample of 500 to get nh_adults

```
set.seed(431002)
# use set.seed to ensure that
# we all get the same random sample

nh_adults <- sample_n(nh_temp, size = 500)
```

Using dim and head (tail works, too)

```
dim(nh_adults)
```

```
[1] 500  13
```

```
head(nh_adults, 3)
```

```
# A tibble: 3 x 13
     ID    Sex   Age   Race     Education   BMI   SBP
  <int> <fctr> <int> <fctr>         <fctr> <dbl> <int>
1 64427   male    37  White College Grad  36.5   111
2 63788 female    40  White  High School  18.2   115
3 66874 female    31  White Some College  27.2    95
# ... with 6 more variables: DBP <int>, Pulse <int>,
#   PhysActive <fctr>, Smoke100 <fctr>,
#   SleepTrouble <fctr>, HealthGen <fctr>
```

Using str

```
str(nh_adults)
```

```
Classes 'tbl_df', 'tbl' and 'data.frame':   500 obs. of  13 va
 $ ID           : int  64427 63788 66874 69734 70409 68961 6261
 $ Sex          : Factor w/ 2 levels "female","male": 2 1 1 2 2
 $ Age          : int  37 40 31 26 44 64 37 42 33 37 ...
 $ Race         : Factor w/ 6 levels "Asian","Black",..: 5 5 5
 $ Education    : Factor w/ 5 levels "8th Grade","9 - 11th Grad
 $ BMI          : num  36.5 18.2 27.2 20.6 29.2 24.2 19.3 31.2
 $ SBP          : int  111 115 95 137 112 123 109 119 110 114 .
 $ DBP          : int  72 74 52 75 71 70 73 71 67 74 ...
 $ Pulse        : int  56 102 98 74 62 80 82 62 68 82 ...
 $ PhysActive   : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 1
 $ Smoke100     : Factor w/ 2 levels "No","Yes": 1 2 1 1 2 2 1
 $ SleepTrouble: Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1
 $ HealthGen    : Factor w/ 5 levels "Excellent","Vgood",..: 2
```

Also useful: names

```
names(nh_adults)
```

```
 [1] "ID"           "Sex"          "Age"
 [4] "Race"         "Education"    "BMI"
 [7] "SBP"          "DBP"          "Pulse"
[10] "PhysActive"   "Smoke100"     "SleepTrouble"
[13] "HealthGen"
```

nh_adults variables

Variable	Description	Sample Values
ID	a numerical code identifying the subject	64427, 63788
Sex	sex of subject (2 levels)	male, female
Age	age (years) at screening of subject	37, 40
Race	reported race of subject	6 levels
Education	educational level of subject	5 levels
BMI	body-mass index, in kg/m^2	36.5, 18.2
SBP	systolic blood pressure in mm Hg	111, 115
DBP	diastolic blood pressure in mm Hg	72, 74
Pulse	60 second pulse rate in beats per minute	56, 102
PhysActive	Moderate or vigorous-intensity sports?	Yes, No
Smoke100	Smoked at least 100 cigarettes lifetime?	Yes, No
SleepTrouble	Told a doctor they have trouble sleeping?	Yes, No
HealthGen	Self-report general health rating	5 levels

Multi-categorical variable levels on the next slide.

`nh_adults` levels for multi-categorical variables

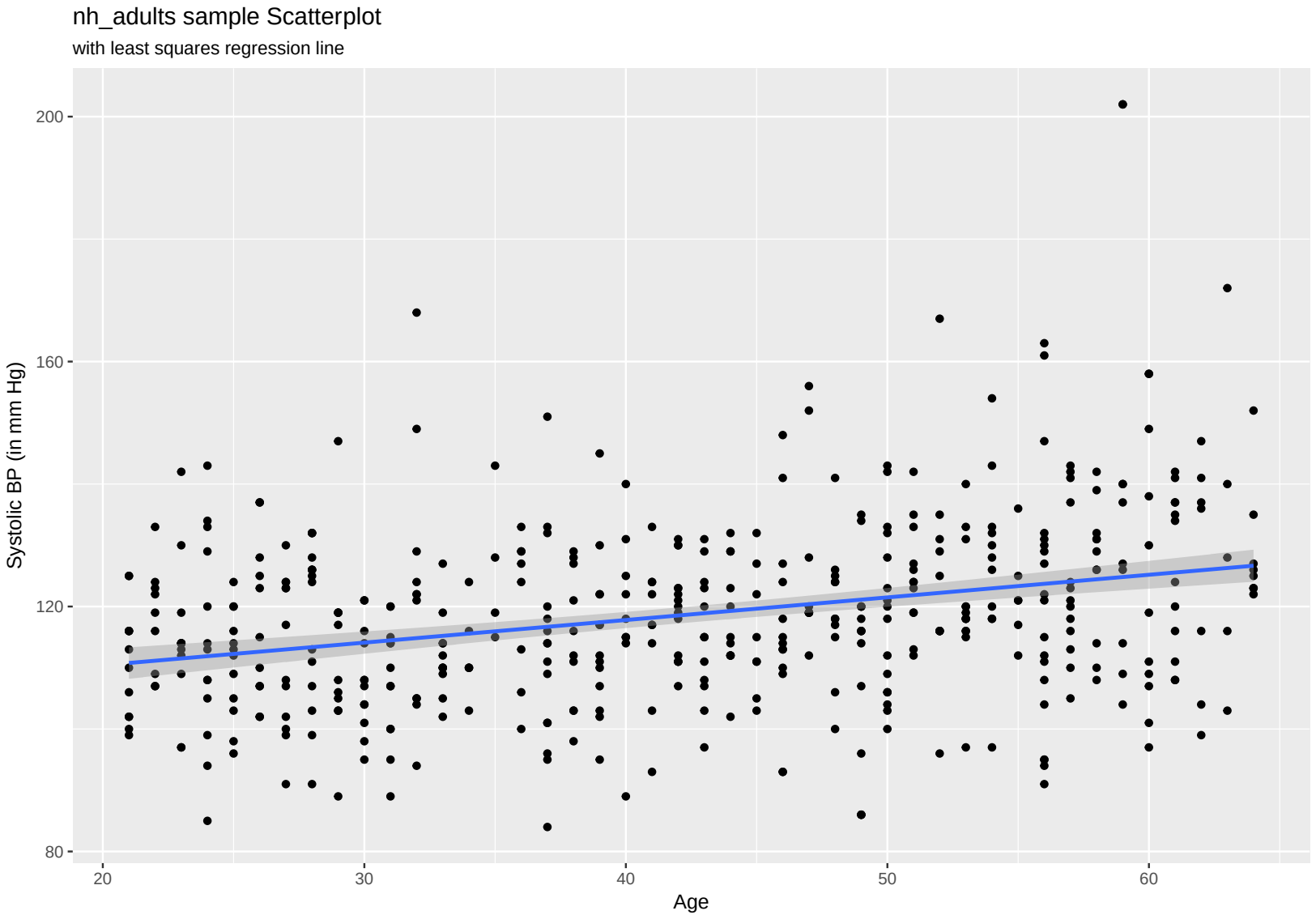
- **Race**
  - Mexican
  - Hispanic
  - White
  - Black
  - Asian
  - Other
- **Education**
  - 8th Grade
  - 9 - 11th Grade
  - High School
  - Some College
  - College Grad
- **HealthGen**:
  - Excellent, Vgood, Good, Fair, Poor.

Summarizing Quantitative Data

```
nh_adults %>%
  select(Age, BMI, SBP) %>%
  summary()
```

```
      Age              BMI              SBP        
 Min.   :21.0   Min.   :17.80   Min.   : 84.0  
 1st Qu.:31.0   1st Qu.:24.20   1st Qu.:109.0  
 Median :42.0   Median :27.70   Median :118.0  
 Mean   :42.1   Mean   :28.73   Mean   :118.6  
 3rd Qu.:53.0   3rd Qu.:32.10   3rd Qu.:127.0  
 Max.   :64.0   Max.   :69.00   Max.   :202.0  
                NA's   :3       NA's   :15     
```

Scatterplot with Least Squares regression line



Scatterplot code

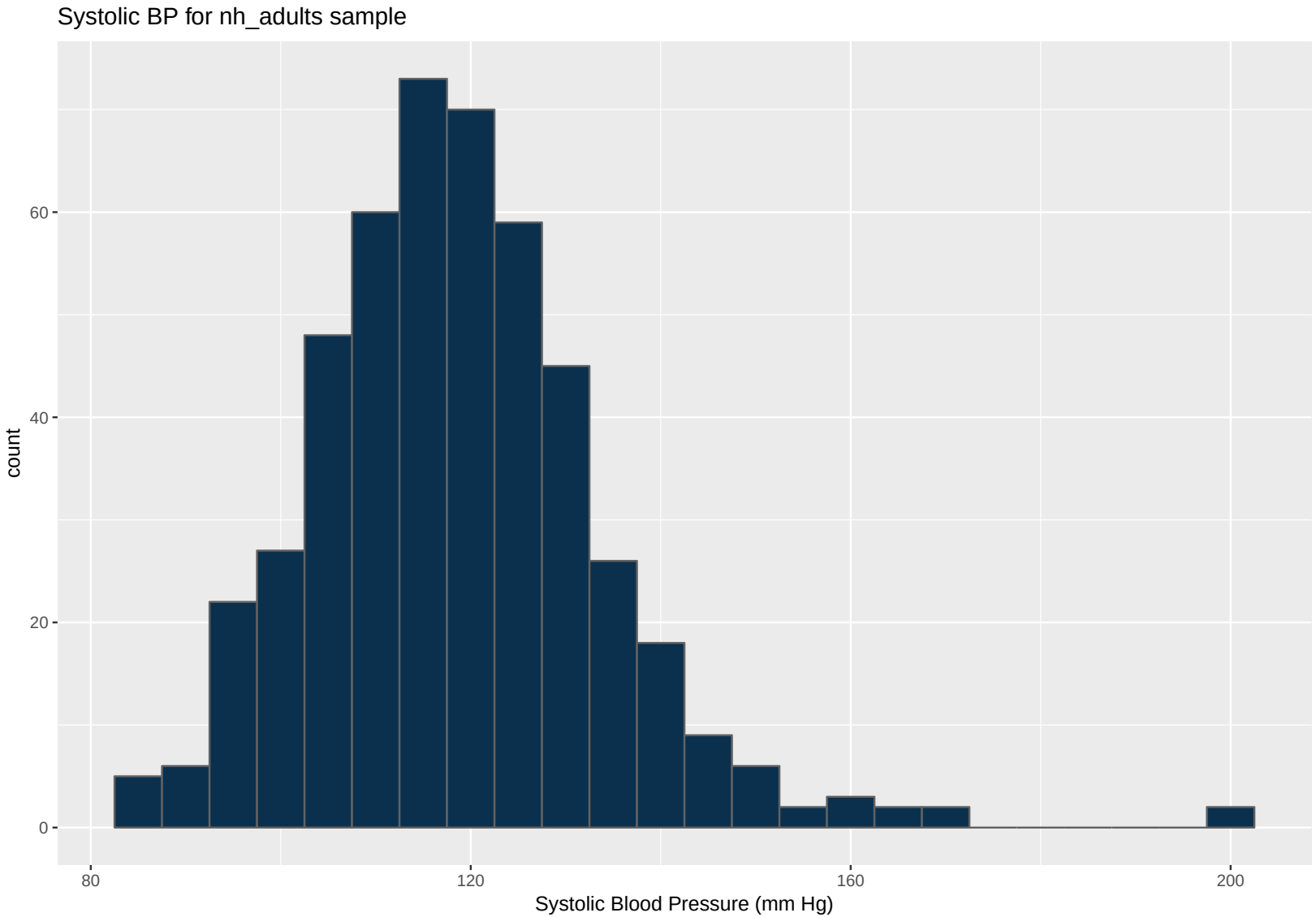
```
nh_adults %>%
    filter(complete.cases(Age, SBP)) %>%
ggplot(data = ., aes(x = Age, y = SBP)) +
    geom_point() +
    geom_smooth(method = "lm") +
    labs(y = "Systolic BP (in mm Hg)",
         title = "nh_adults sample Scatterplot",
         subtitle = "with least squares regression line")
```

Histogram of SBP (code)

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

ggplot(data = filter(nh_adults, complete.cases(SBP)),
       aes(x = SBP)) +
    geom_histogram(binwidth = 5, col = cwru.gray,
                   fill = cwru.blue) +
    labs(title = "Systolic BP for nh_adults sample",
         x = "Systolic Blood Pressure (mm Hg)")
```

Histogram of SBP (result)



Measure the Center: Mean and Median

```
nh_adults %>%
    summarize(count = n(), mean = mean(SBP),
              median = median(SBP))
```

```
# A tibble: 1 x 3
  count  mean median
  <int> <dbl>  <int>
1   500    NA     NA
```

What happened?

Mean and the Median (without NAs)

```
nh_adults %>%
    filter(complete.cases(SBP)) %>%
    summarize(count = n(), mean = mean(SBP),
              median = median(SBP))
```

```
# A tibble: 1 x 3
  count     mean median
  <int>    <dbl>  <int>
1   485 118.5918    118
```

Or, use . . .

```
nh_adults %>%
    summarize(mean = mean(SBP, na.rm=TRUE))
```

The Trimmed Mean

The 90% trimmed mean is the mean of the middle 90% of the data.

```
nh_adults %>%
    filter(!is.na(SBP)) %>%
    summarize(mean = mean(SBP),
              trim90 = mean(SBP, trim = 0.05),
              trim80 = mean(SBP, trim = 0.1))
```

```
# A tibble: 1 x 3
      mean   trim90   trim80
     <dbl>    <dbl>    <dbl>
1 118.5918 117.9542 117.7866
```

What I've called `trim90` here is called both a 90% trimmed mean, and a 10% trimmed mean by some people.

The Mode of a Quantitative Variable

The mode is the most common value.

```
nh_adults %>%
    group_by(Age) %>%
    summarize(count = n()) %>%
    arrange(desc(count))
```

```
# A tibble: 44 x 2
     Age count
   <int> <int>
 1    56    19
 2    50    18
 3    28    16
 4    37    16
 5    42    16
 6    49    15
 7    24    13
```

Measuring Spread: The Range

The **range** spans the minimum and maximum of the data.

```
nh_adults %>%
    select(SBP) %>%
    range(., na.rm=TRUE)
```

```
[1]  84 202
```

Often, we'll take the difference (max - min) and call that the range. Here, that's 202 - 84 = 118.

The Inter-Quartile Range (IQR)

The **inter-quartile range** is the difference between the third and first quartiles (75^{th} and 25^{th} percentiles) of the data.

```
IQR(nh_adults$SBP, na.rm=TRUE)
```

```
[1] 18
```

Range and IQR are easy to see in the `summary` ...

```
summary(nh_adults$SBP)
```

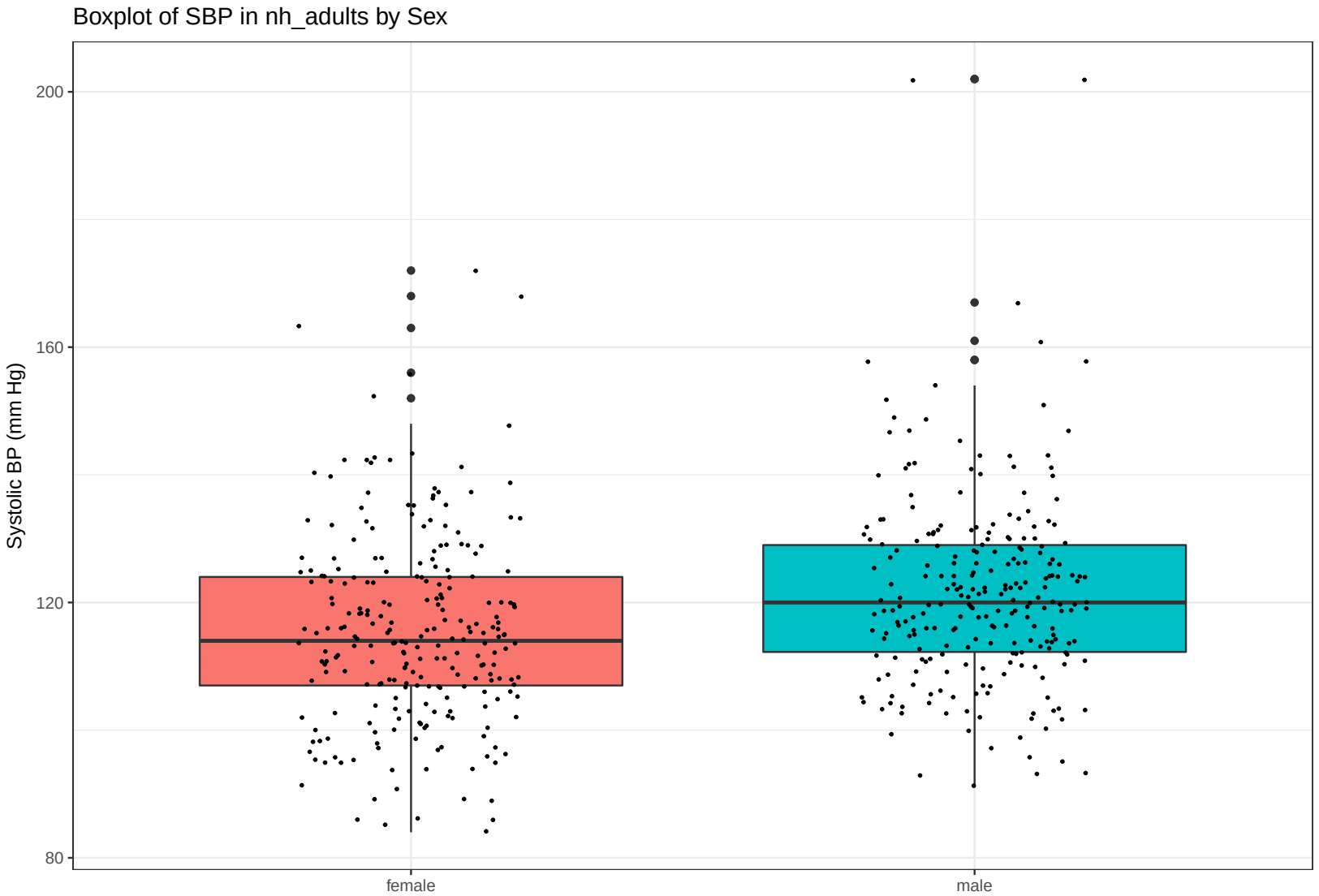
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   84.0   109.0   118.0   118.6   127.0   202.0
   NA's
     15
```

Boxplot with Points Jittered In

The boxplot displays the five-number summary.

```
ggplot(data = filter(nh_adults, complete.cases(Sex, SBP)),
       aes(x = Sex, y = SBP, fill = Sex)) +
    geom_boxplot() +
    geom_jitter(width = 0.2, size = 0.5) +
    guides(fill = FALSE) +
    theme_bw() +
    labs(title = "Boxplot of SBP in nh_adults by Sex",
         x = "", y = "Systolic BP (mm Hg)")
```

Boxplot with Points Jittered In, Results



The Variance and the Standard Deviation

The IQR is always a reasonable summary of spread, just as the median is always a reasonable summary of the center of a distribution. Yet, most people are inclined to summarize a batch of data using two numbers: the **mean** and the **standard deviation**. This is really only a sensible thing to do if you are willing to assume the data follow a Normal distribution: a bell-shaped, symmetric distribution without substantial outliers.

But **most data do not (even approximately) follow a Normal distribution**. Summarizing by the median and quartiles (25th and 75th percentiles) is much more robust, explaining R's emphasis on them.

We use `var` in R to calculate the variance and `sd` to calculate a standard deviation.

Normal Summaries for Several `nh_adults` variables

```
nh_adults %>%
    select(Age, BMI, SBP, DBP, Pulse) %>%
    summarize_all(sd, na.rm = TRUE)
```

```
# A tibble: 1 x 5
       Age      BMI      SBP      DBP    Pulse
     <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
1 12.53706 6.487816 15.30267 10.83152 11.47438
```

How Variance (and SD) are calculated...

In thinking about spread, we might consider how far each data value is from the mean. Such a difference is called a *deviation*. We could just average the deviations, but the positive and negative differences always cancel out, leaving an average deviation of zero, so that's not helpful. Instead, we *square* each deviation to obtain non-negative values, and to emphasize larger differences. When we add up these squared deviations and find their mean (almost), this yields the **variance**.

$$\text{Variance} = s^2 = \frac{\Sigma(y - \bar{y})^2}{n - 1}$$

Why almost? It would be the mean of the squared deviations only if we divided the sum by n, but instead we divide by $n - 1$ because doing so produces an estimate of the true (population) variance that is *unbiased*.

The Standard Deviation

To return to the original units of measurement, we take the square root of s^2, and instead work with s, the **standard deviation**.

$$\text{Standard Deviation} = s = \sqrt{\frac{\Sigma(y - \bar{y})^2}{n-1}}$$

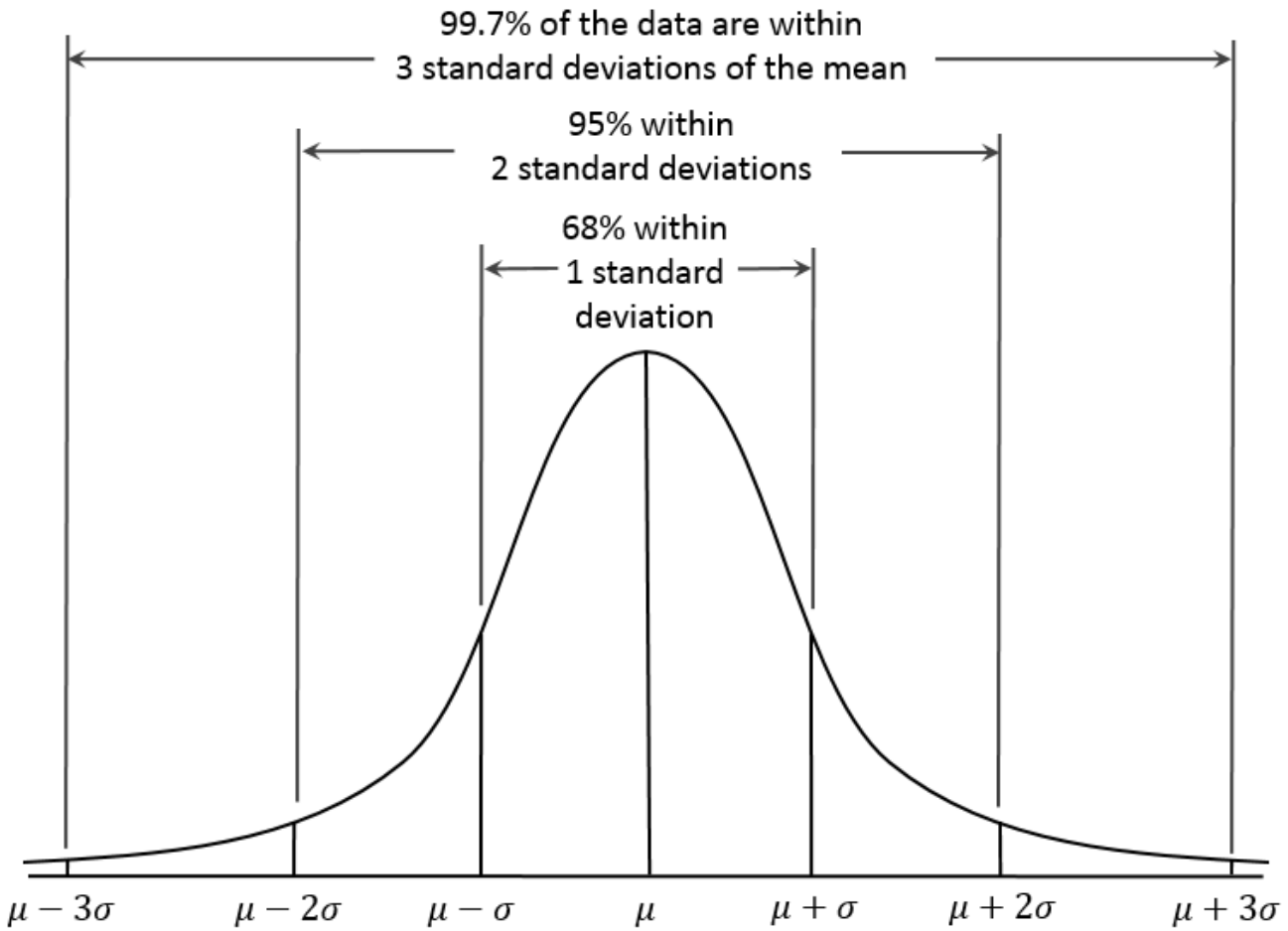
Interpretation 1: Chebyshev's Inequality

Chebyshev's Inequality tells us that for any distribution, regardless of its relationship to a Normal distribution, no more than $1/k^2$ of the distribution's values can lie more than k standard deviations from the mean. This implies, for instance, that for **any** distribution,

- at least 75% of the values must lie within two standard deviations of the mean, and
- at least 89% must lie within three standard deviations of the mean.

The "Empirical Rule"

We often refer to the population or process mean of a distribution with μ and the standard deviation with σ, leading to the Figure below.



Interpretation 2: The "Empirical Rule"

For a set of measurements that follow a Normal distribution, the interval:

- Mean $\pm$ Standard Deviation contains approximately 68% of the measurements;
- Mean $\pm$ 2(Standard Deviation) contains approximately 95% of the measurements;
- Mean $\pm$ 3(Standard Deviation) contains approximately all (99.7%) of the measurements.

But if the data are not from an approximately Normal distribution, then this Empirical Rule is less helpful.

Checking the Empirical Rule for the SBP data, 1

```
nh_adults %$%
    mosaic::favstats(SBP)
```

```
 min  Q1 median  Q3 max     mean       sd   n missing
  84 109    118 127 202 118.5918 15.30267 485      15
```

So the Empirical Rule suggests for the SBP data:

- $118.6 \pm 15.3 = (103.3, 133.9)$ contains about 68% of the SBP values,
- $118.6 \pm 2(15.3) = (88.0, 149.2)$ holds about 95% of the SBP values,
- $118.6 \pm 3(15.3) = (72.7, 164.5)$ holds about all (99.7%) of the SBP values.

Checking the Empirical Rule for the SBP data, 2

```
nh_adults %>%
    filter(!is.na(SBP)) %>%
    summarize(n = sum(!is.na(SBP)),
              within1sd = sum(SBP >= 103.3 & SBP <= 133.9),
              percent = 100 * within1sd / n)
```

```
# A tibble: 1 x 3
      n within1sd  percent
  <int>     <int>    <dbl>
1   485       351 72.37113
```

The Emp_Rule function in Love-boost.R can help

```
source("Love-boost.R")

temp <- filter(nh_adults, complete.cases(SBP))
Emp_Rule(temp$SBP)
```

	count	proportion
Mean +/- 1 SD	351	0.7237
Mean +/- 2 SD	466	0.9608
Mean +/- 3 SD	480	0.9897
Entire Data Set	485	1

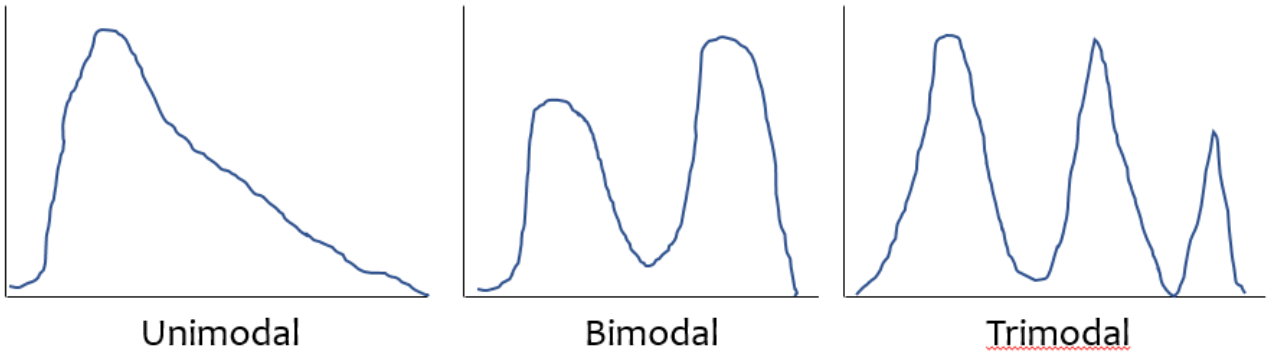
Measuring the Shape of a Distribution

When considering the shape of a distribution, one is often interested in three key points.

- The number of modes in the distribution, which I always assess through plotting the data.
- The **skewness**, or symmetry that is present, which I typically assess by looking at a plot of the distribution of the data, but if required to, will summarize with a non-parametric measure of **skewness**.
- The **kurtosis**, or heavy-tailedness (outlier-proneness) that is present, usually in comparison to a Normal distribution. Again, this is something I nearly inevitably assess graphically, but there are measures.

A Normal distribution has a single mode, is symmetric and, naturally, is neither heavy-tailed or light-tailed as compared to a Normal distribution (we call this mesokurtic).

Multimodal vs. Unimodal distributions



Truly multimodal distributions are usually described that way in terms of shape.

Skew

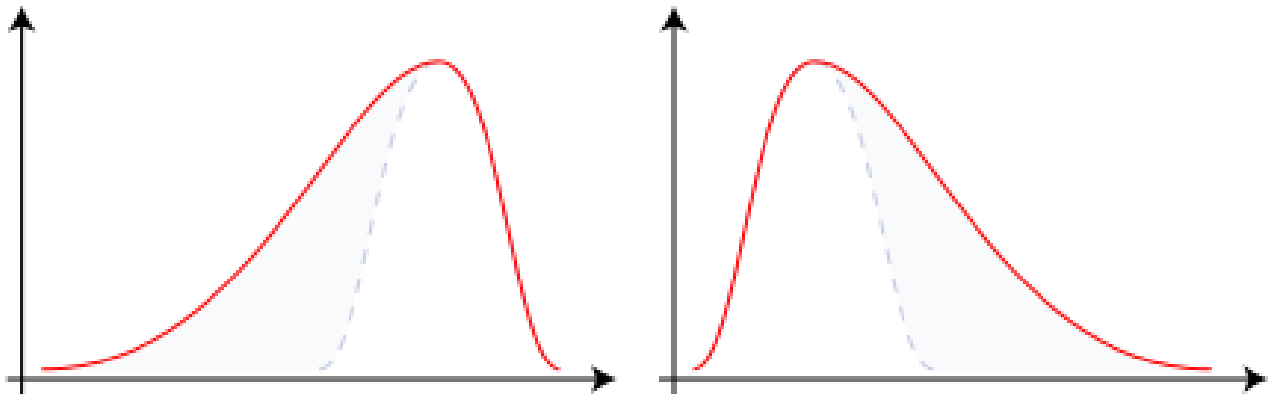
For unimodal distributions, skewness and kurtosis become useful ideas. Whether or not a distribution is approximately symmetric is an important consideration in describing its shape. Graphical assessments are always most useful in this setting, particularly for unimodal data. My favorite measure of skew, or skewness if the data have a single mode, is:

$$skew_1 = \frac{\text{mean} - \text{median}}{\text{standard deviation}}$$

- Symmetric distributions generally show values of $skew_1$ near zero. If the distribution is actually symmetric, the mean should be equal to the median.
- Distributions with $skew_1$ values above 0.2 in absolute value generally indicate meaningful skew.

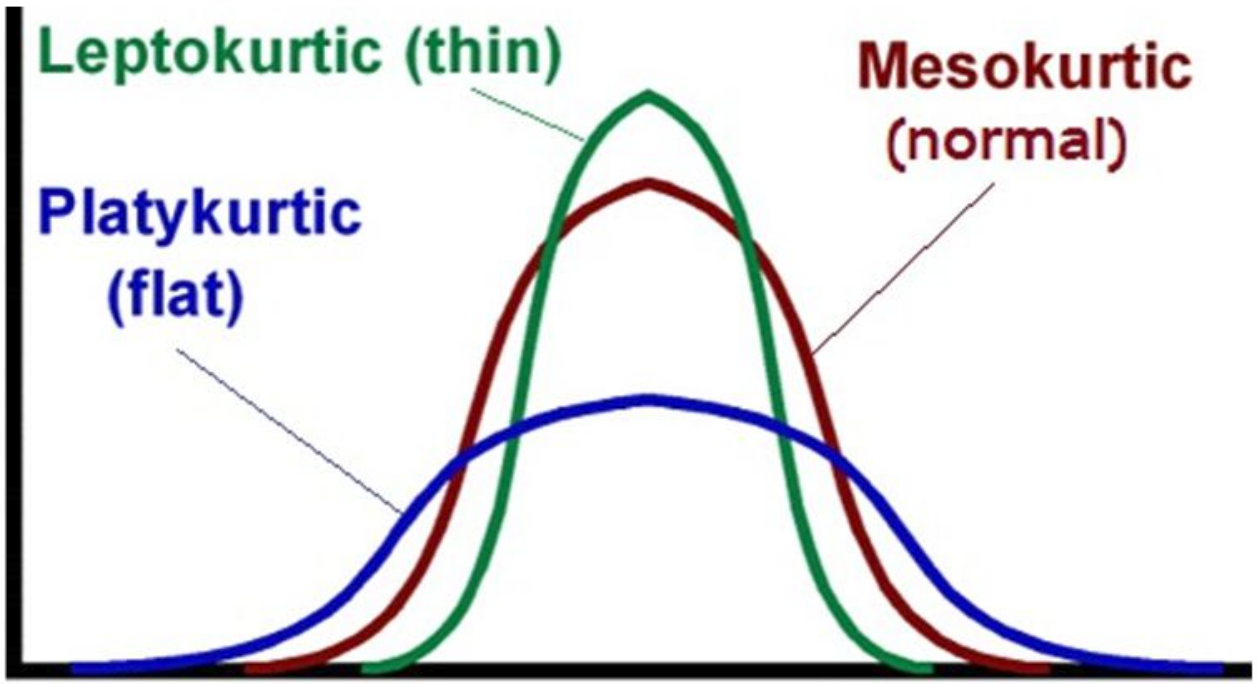
Measuring Skew

- Positive skew (mean $>$ median if the data are unimodal) is also referred to as *right skew*.
- Negative skew (mean $<$ median if the data are unimodal) is referred to as *left skew*.



Kurtosis

When we have a unimodal distribution that is symmetric, we will often be interested in the behavior of the tails of the distribution, as compared to a Normal distribution with the same mean and standard deviation.



The describe function in the psych package

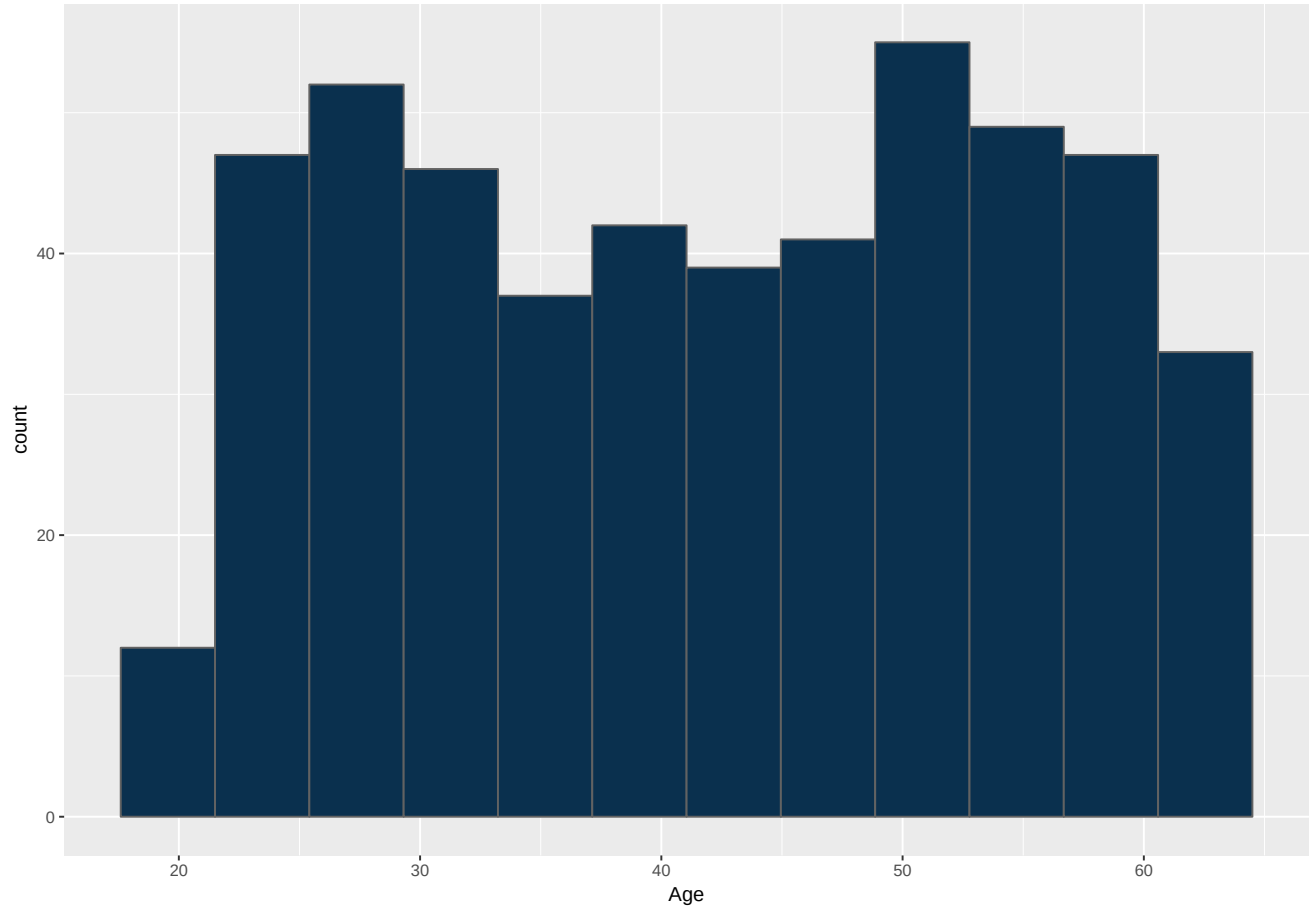
```
psych::describe(nh_adults %>% select(Age, SBP))
```

```
    vars   n   mean    sd median trimmed   mad min max
Age    1 500  42.10 12.54     42   42.11 16.31  21  64
SBP    2 485 118.59 15.30    118  117.79 13.34  84 202
    range  skew kurtosis   se
Age    43 -0.03    -1.23 0.56
SBP   118  1.00     3.44 0.69
```

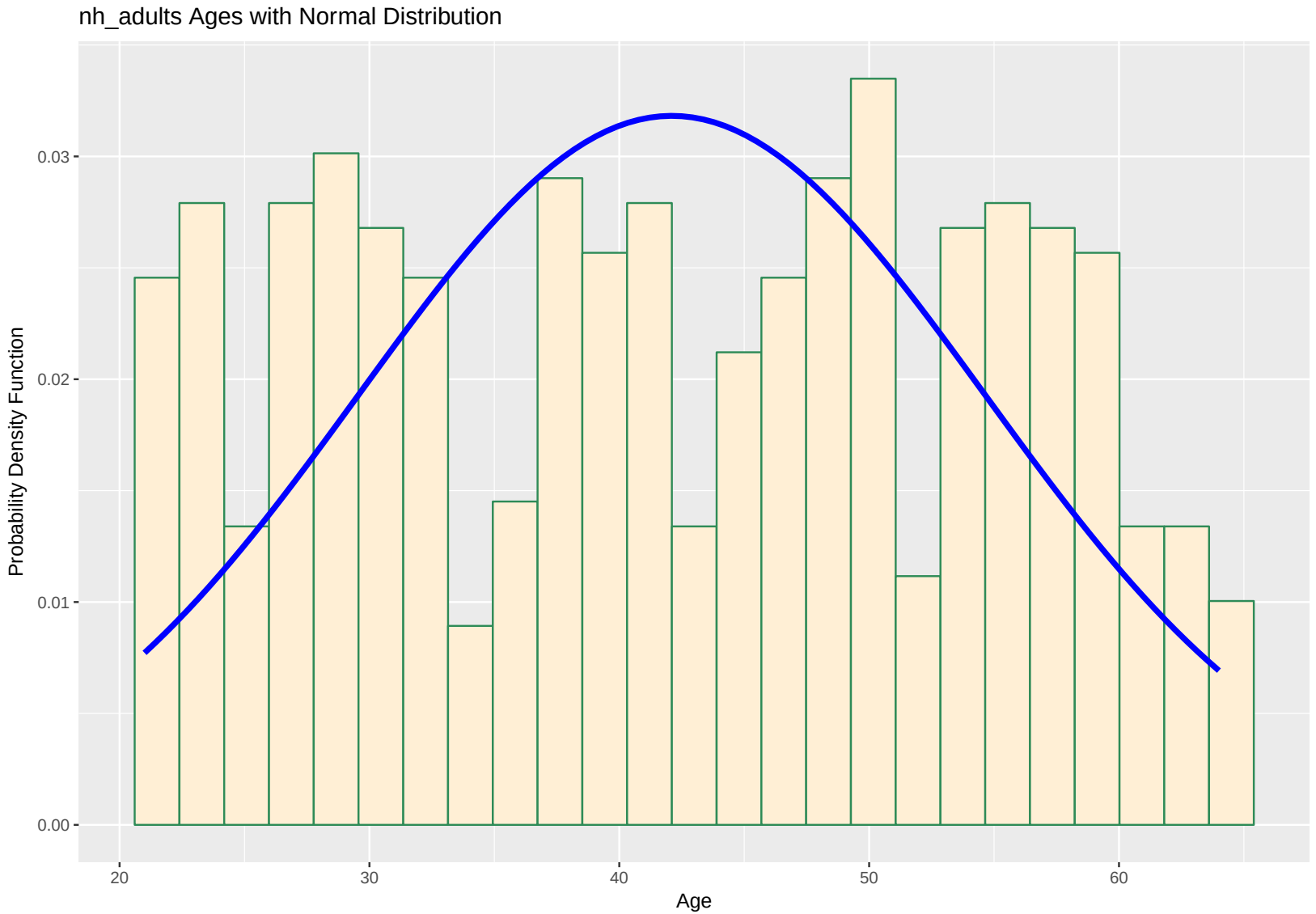
1. This `skew` is not our $skew_1$ measure.
2. Interpret `kurtosis` with care. A plot is more useful.
3. Meaning of `trimmed`, `mad` in Course Notes, Section 5
4. `se` = standard error of the mean = $\frac{sd}{\sqrt{n}}$

.

Do the Age data appear skewed? Outlier-prone?

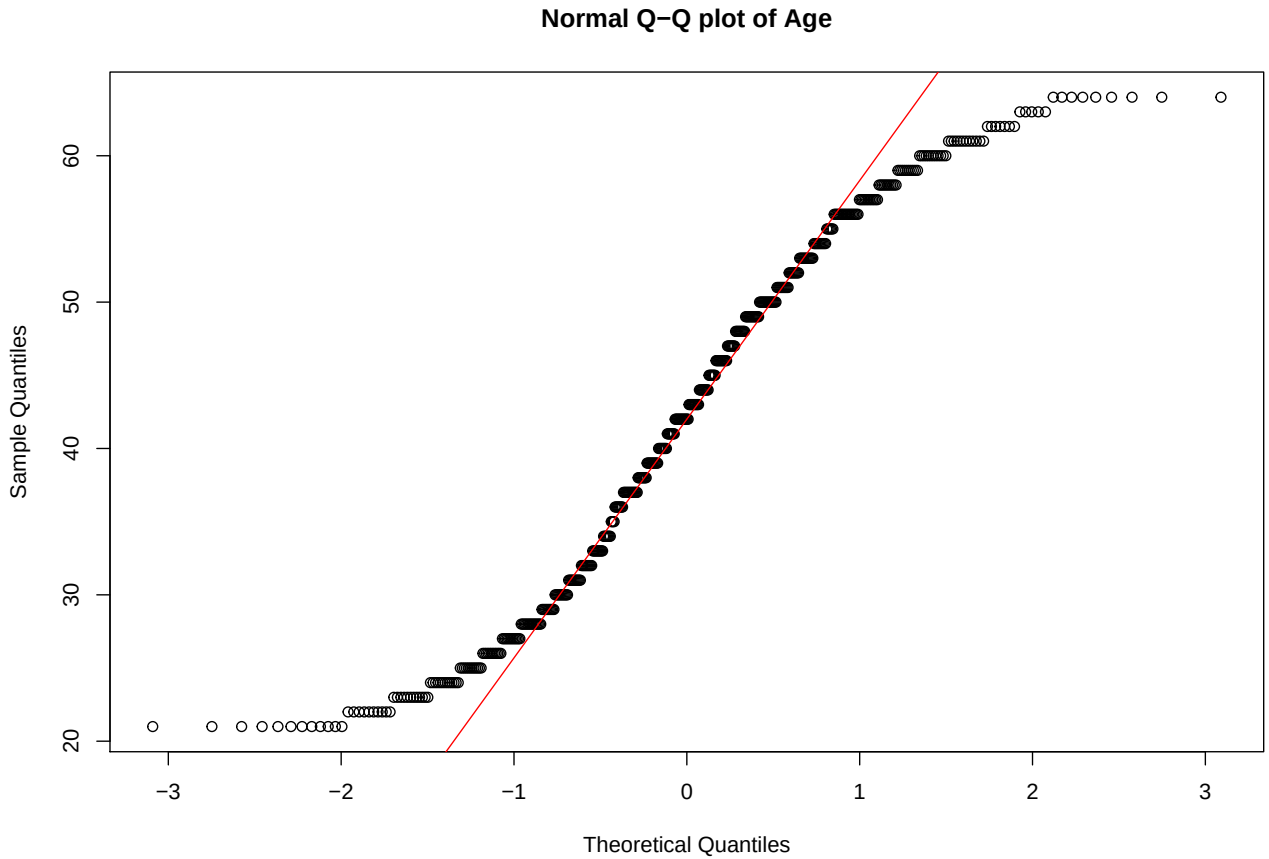


Ages + Normal Model (Section 8.3)

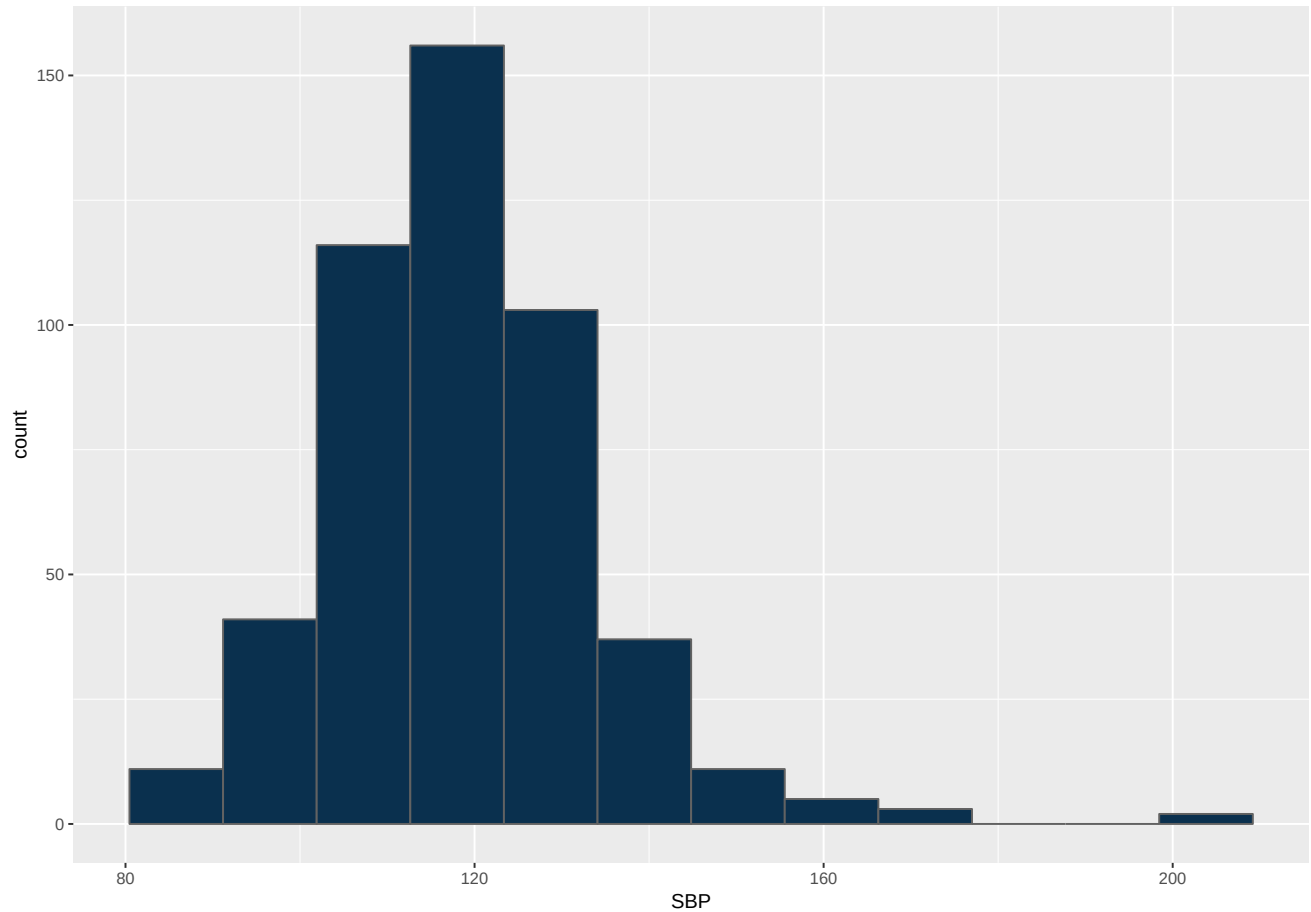


Normal Q-Q plot for Age (See Notes, section 8)

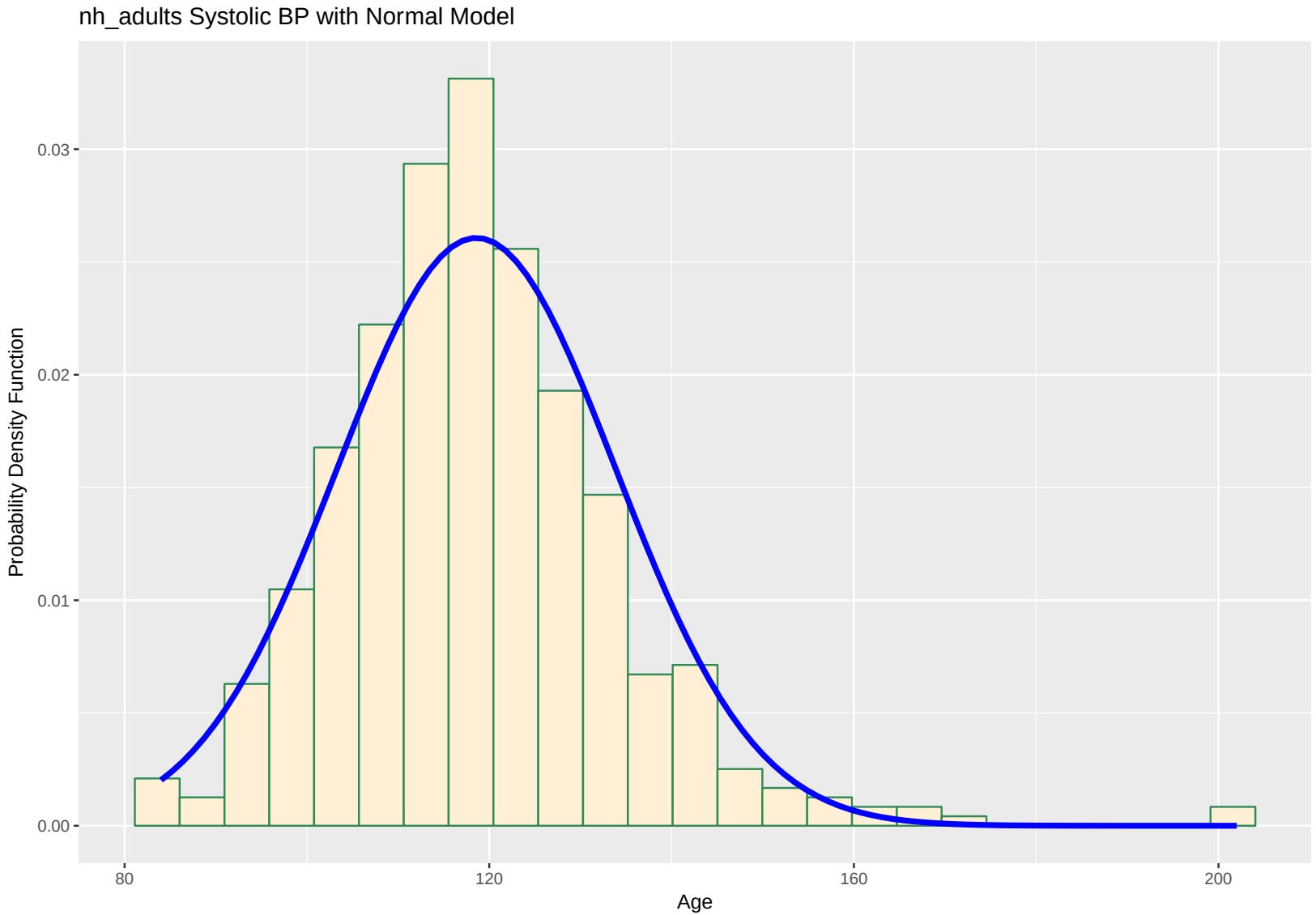
```
qqnorm(nh_adults$Age, main = "Normal Q-Q plot of Age")
qqline(nh_adults$Age, col = "red")
```



Do the SBP data appear skewed? Outlier-prone?

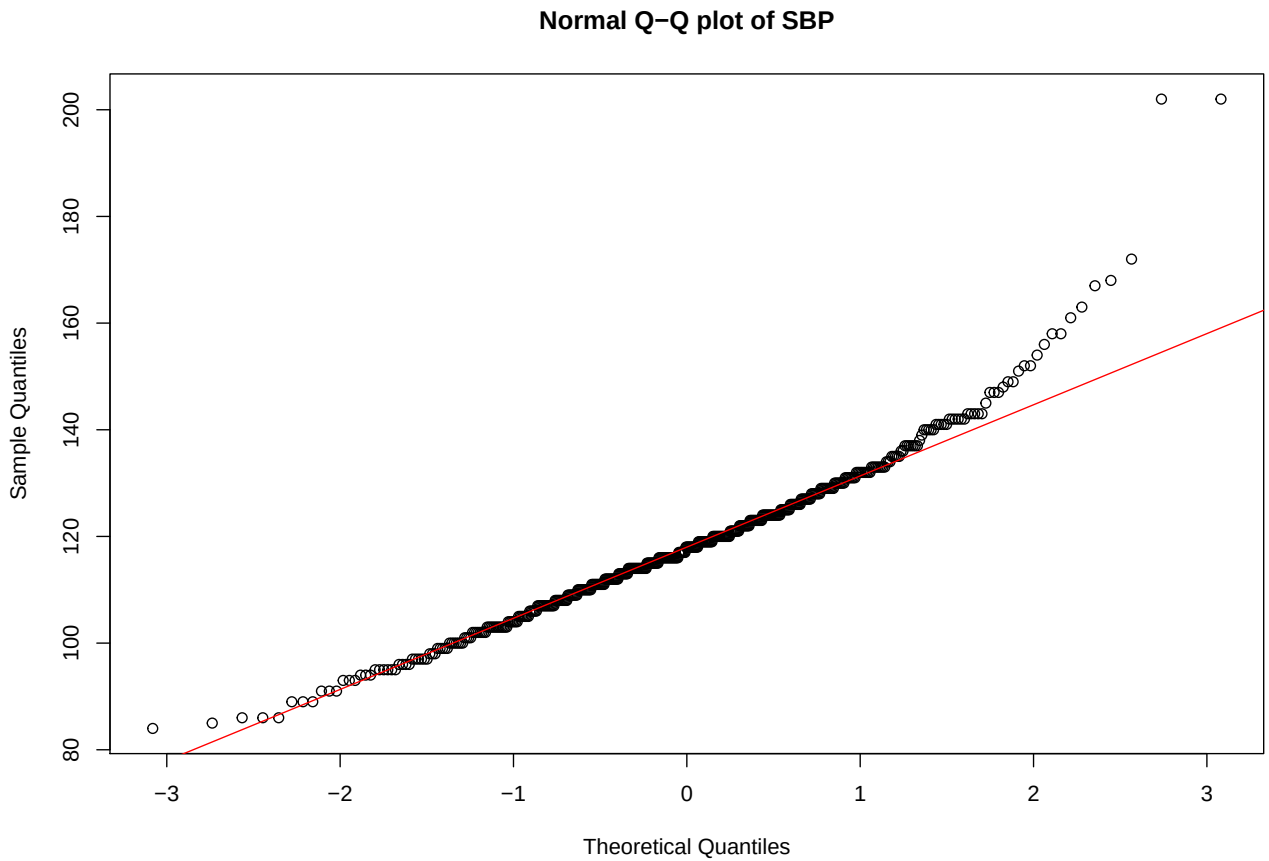


SBP + Normal Model (Section 8.3)



Normal Q-Q plot for SBP

```
qqnorm(nh_adults$SBP, main = "Normal Q-Q plot of SBP")
qqline(nh_adults$SBP, col = "red")
```



Visit Class 5 form by noon Thursday 2017-09-14

The address is https://goo.gl/forms/U3a9r3qNRI5XPNQg2

- Doing this will help your class participation grade. Not doing it will hurt.
  - Your goals for the course.
  - Read *The Signal and The Noise* (Intro and Chapter 1) first.
- Remember to **log in to Google via CWRU** to use the form.

If you missed class and are reading this later, please fill out the form by noon Thursday.

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?

- We'll look at Course Notes Sections
- Assignment 1 is due 2017-09-15 at noon.
- I'll tell you more about the project on 2017-09-19.
- Read Silver Chapters 2-3 for 2017-09-26.

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.