Social Statistics

Introducing Spread and Graphics

September 27, 2021

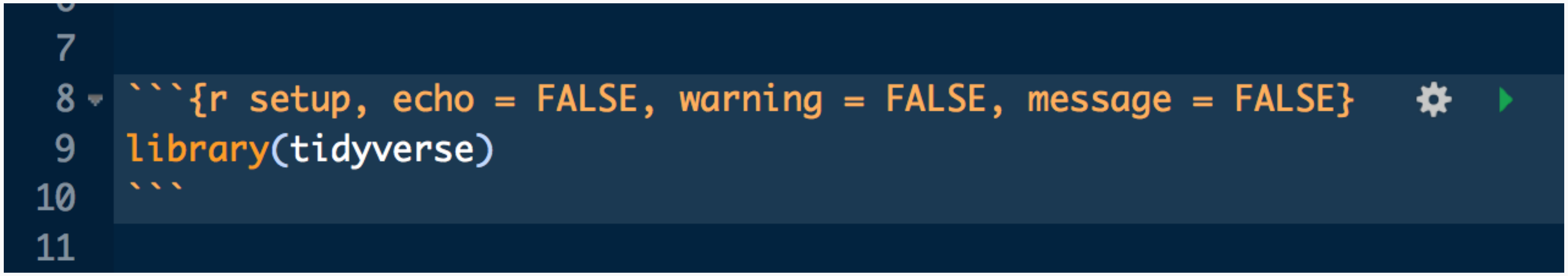
Assignment 2 General Thoughts

Include your Rmd file if you email me with questions

Remember to add your name and date to the header

Knit or Preview as you go so it's easier to identify where problems are

Load packages when you load your data. And when loading tidyverse, do not include echoes, warnings, and messages



~~~
```{r setup, echo = FALSE, warning = FALSE, message = FALSE}
library(tidyverse)
```
~~~

Review in-class notebooks before starting

Assignment 2 Recap

1. What are the mean and median of `agekdbrn`?

```
summary(assignment2$agekdbrn)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.00   20.00   23.00   24.23   28.00   57.00
```

This also works...

```
mean(assignment2$agekdbrn)
```

```
## [1] 24.22761
```

```
median(assignment2$agekdbrn)
```

```
## [1] 23
```

Assignment 2 Recap

2. Find the difference between 25th percentile and 75th percentile:

```r
pctle75 <- 28    # Don't use quotation marks!
pctle25 <- 20    # Or parentheses! Or curly brackets!
pctle75 - pctle25
```

```
## [1] 8
```

Quick Detour

IQR matters for definition of outliers

High outliers are values that are at least 1.5 times the IQR above the 75th percentile

```
pctle75 + 1.5*(pctle75 - pctle25) # cutoff for high outliers
```

```
## [1] 40
```

Low outliers are values that are at least 1.5 times the IQR below the 25th percentile

```
pctle25 - 1.5*(pctle75 - pctle25) # cutoff for low outliers
```

```
## [1] 8
```

Assignment 2 Recap

3. What is the mode of `agekdbrn` for respondents who completed 12 or fewer years of education?

```
table(assignment2$agekdbrn
      [assignment2$educ<=12])
```

```
## 
##  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  3
##   3  12  24  66 186 274 437 468 444 475 280 281 228 235 147 120 116  68 118  4
##  32  33  34  35  36  37  38  39  40  41  42  43  45  46  47  50  52 
##  61  39  29  34  17  15  14   9   9   5   5   2   3   2   5   2   2
```

Assignment 2 Recap

Want to sort?

```
sort(table(assignment2$agekdbrn
      [assignment2$educ<=12]), decreasing = TRUE)
```

```
##
##  21  19  20  18  23  22  17  25  24  16  26  27  30  28  29  15  32  31  33  3
## 475 468 444 437 281 280 274 235 228 186 147 120 118 116  68  66  61  47  39  3
##  34  14  36  37  38  13  39  40  41  42  47  12  45  43  46  50  52
##  29  24  17  15  14  12   9   9   5   5   5   3   3   2   2   2   2
```

Assignment 2 Recap

4. What proportion of respondents completed exactly 16 years of education?

```r
prop.table(table(assignment2$educ))
```

```
## 
##          8          9         10         11         12         13         14 
## 0.02278190 0.02884314 0.03877103 0.05434215 0.30274846 0.08224475 0.12948061 
##         15         16         17         18         19         20 
## 0.04786289 0.15936880 0.03114223 0.05099801 0.01891525 0.03250078 
```

Want to round?

```r
round(prop.table(table(assignment2$educ)),3) # 3 for 3 decimal places
```

```
## 
##     8     9    10    11    12    13    14    15    16    17    18    19    20 
## 0.023 0.029 0.039 0.054 0.303 0.082 0.129 0.048 0.159 0.031 0.051 0.019 0.033 
```

Assignment 2 Recap

5. Use `dplyr` to create a new data frame with only the `agekdbrn` and `educ` variables, and that is limited to respondents who have 16 or more years of education.

```
library(tidyverse)  # dplyr loads with tidyverse!
```

A Couple Options...

```
assignment2_q5a <- select(assignment2, agekdbrn, educ) # DF name but no $
assignment2_q5a <- filter(assignment2_q5a, educ>=16)
```

```
assignment2_q5b <- assignment2 |> # With pipe, need DF name in first line
     select(agekdbrn, educ) |> # But omit DF name from subsequent lines
     filter(educ>=16)
```

Assignment 2 Recap

6. What are the mean and median of `agekdbrn` for respondents in this new data frame?

```
assignment2_q5b <- assignment2 |>
     select(agekdbrn, educ) |>
     filter(educ>=16) # No quotation marks

summary(assignment2_q5b$agekdbrn)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   14.00   24.00   27.00   27.75   31.00   57.00 
```

Assignment 2 Recap

7. How long did the assignment take?

```r
summary(time)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.670   1.500   2.000   2.038   2.750   3.000
```

Center, Spread, Shape

Range gives us the *minimum* and the *maximum* values

Mean and median give us the *center* of the distribution

Mode gives us the *most frequent* value

Also want information about the *spread* of distributions

- Variance
- Standard Deviation
- Skewness

Spread

Variance = how we measure *spread* but it has no common scale

Standard Deviation = measure of how far observations tend to be from the mean

Standard Deviation is the square root of the variance

$$s = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n}}$$

How do we find the variance and standard deviation in R?

Loading Files

We'll use the `gss_week3.csv` file on Canvas. Download this file, save it, and load it in `notebook_03_01.Rmd`.

```r
gss_week3 <- read.csv("gss_week3.csv")
```

Describing Spread

Start with a summary of the `agekdbrn` variable

```
summary(gss_week3$agekdbrn)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.00   20.00   23.00   24.22   28.00   57.00
```

For variance, use `var( )`:

```
var(gss_week3$agekdbrn)
```

```
## [1] 34.03922
```

For standard deviation, use `sd( )`:

```
sd(gss_week3$agekdbrn)
```

```
## [1] 5.834314
```

Describing Spread

We can show that the standard deviation is the square root of the variance:

```
var(gss_week3$agekdbrn) # Variance
```

```
## [1] 34.03922
```

```
sqrt(var(gss_week3$agekdbrn)) # Square Root of Variance
```

```
## [1] 5.834314
```

```
sd(gss_week3$agekdbrn) # Standard Deviation
```

```
## [1] 5.834314
```

```
sd(gss_week3$agekdbrn) ^ 2 # Standard Deviation Squared
```

```
## [1] 34.03922
```

Describing Spread

Would you expect more or less variation in the distribution of completed years of education (the `educ` variable)?

```
var(gss_week3$educ)
```

```
## [1] 7.25643
```

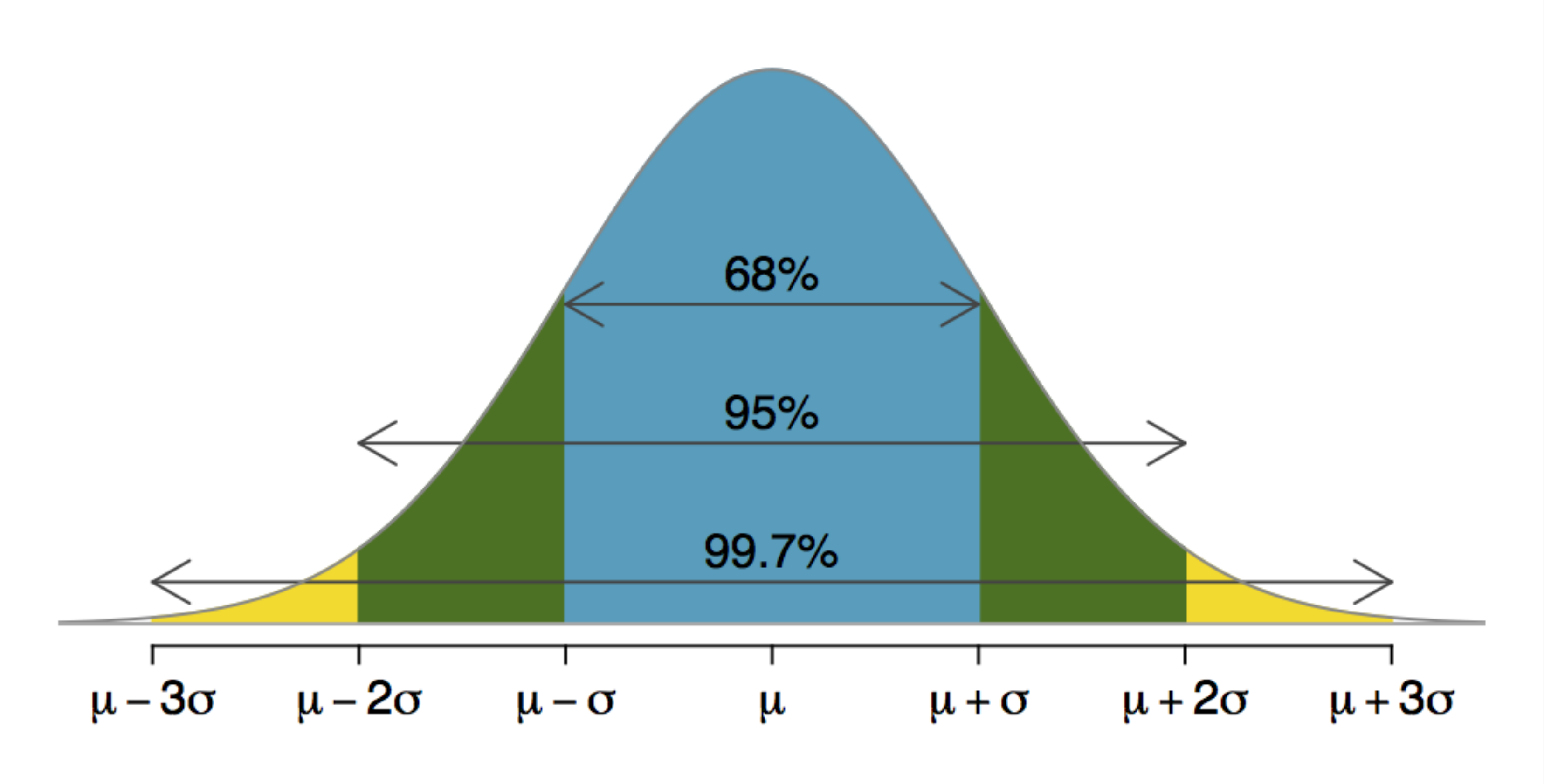
```
sd(gss_week3$educ)
```

```
## [1] 2.693776
```

Describing The Shape of the Spread

For now, keep in mind that the shape we like the most is a *normal distribution* (or bell curve)

The Normal Distribution



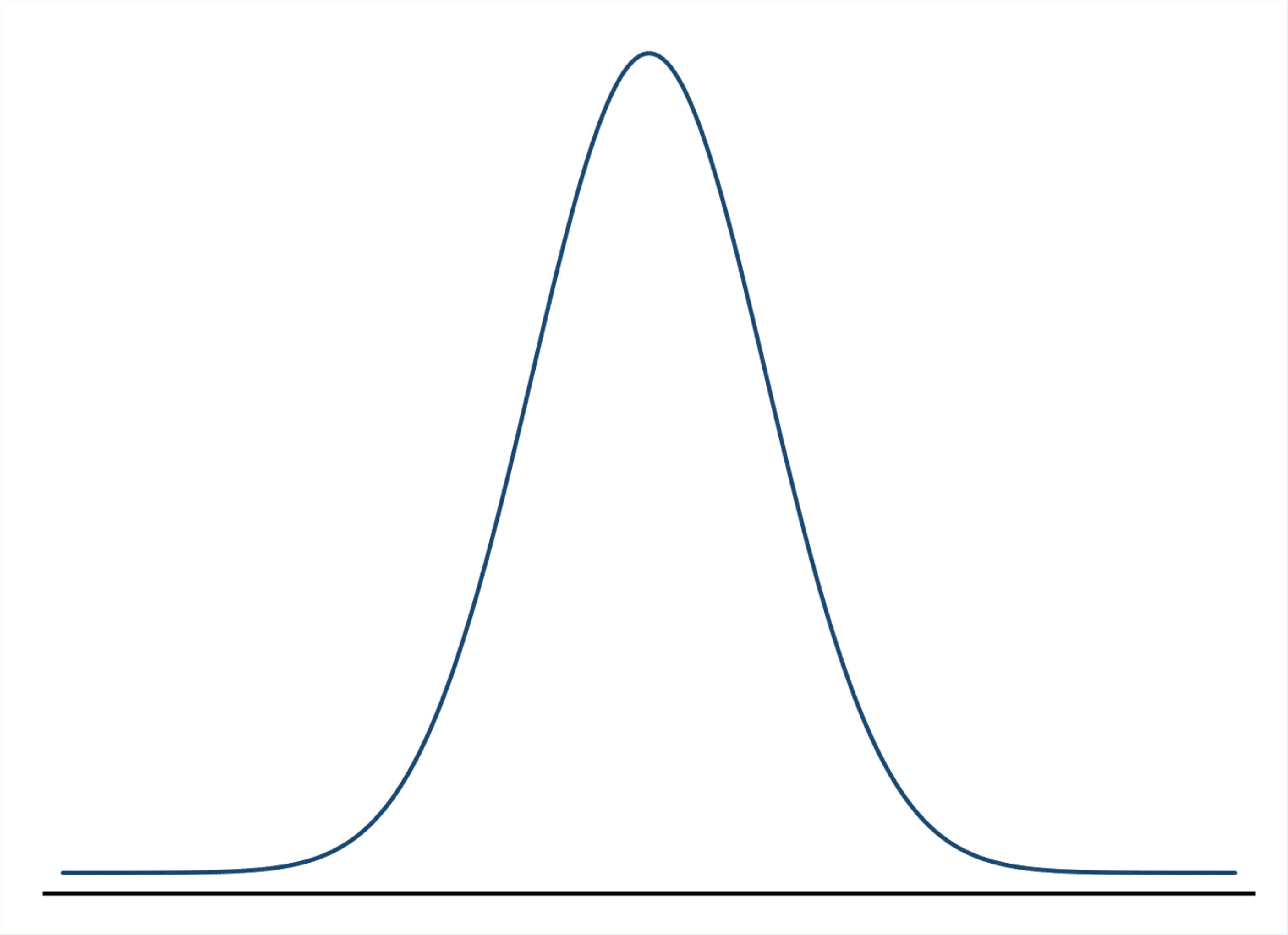
Describing The Shape of the Spread

But values are often not normally distributed

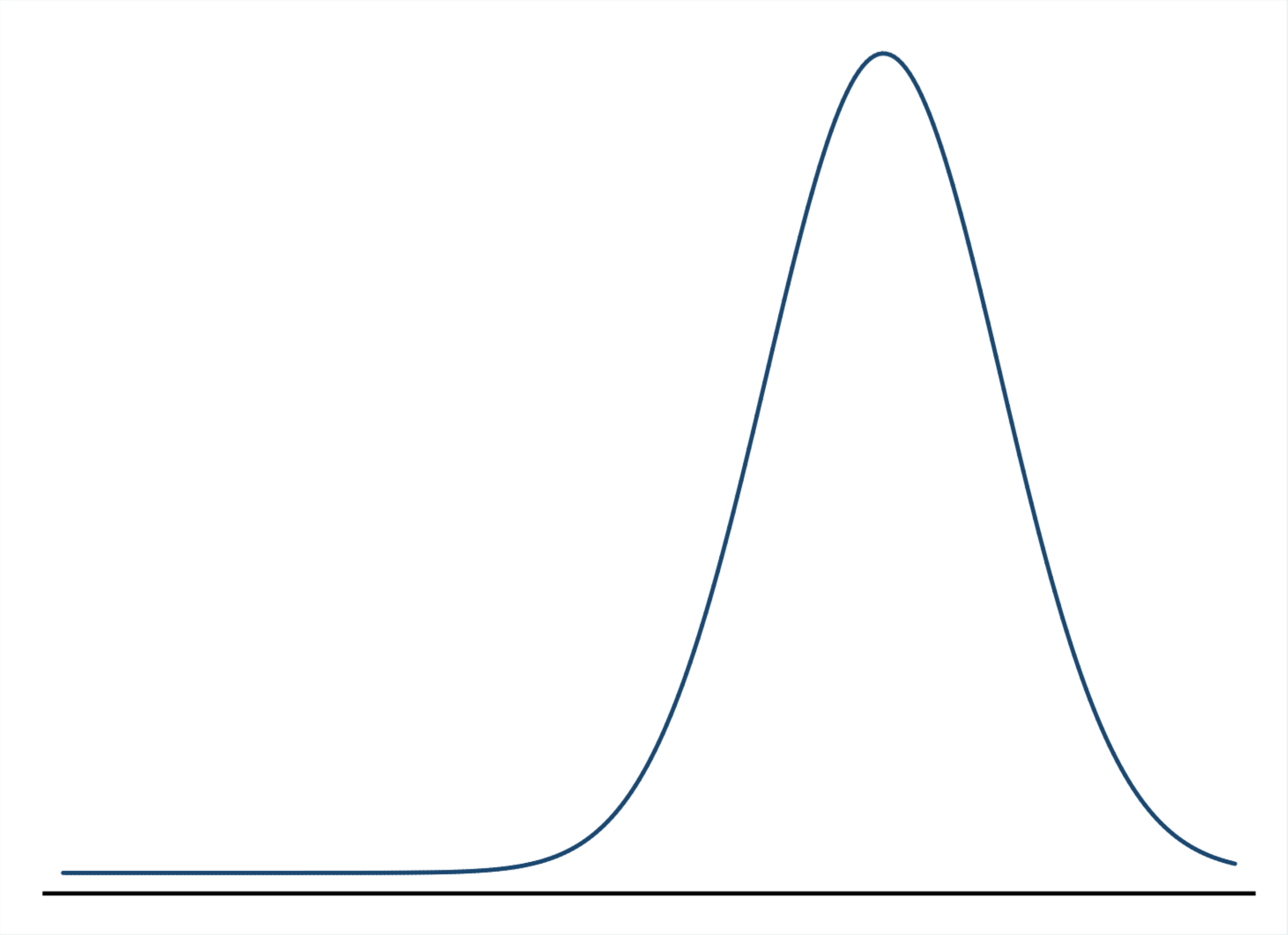
The measure of `skewness` tells us where the "long tail" extends

- Right skewed distributions extend to higher distributions
- Left skewed distributions extend to lower distributions

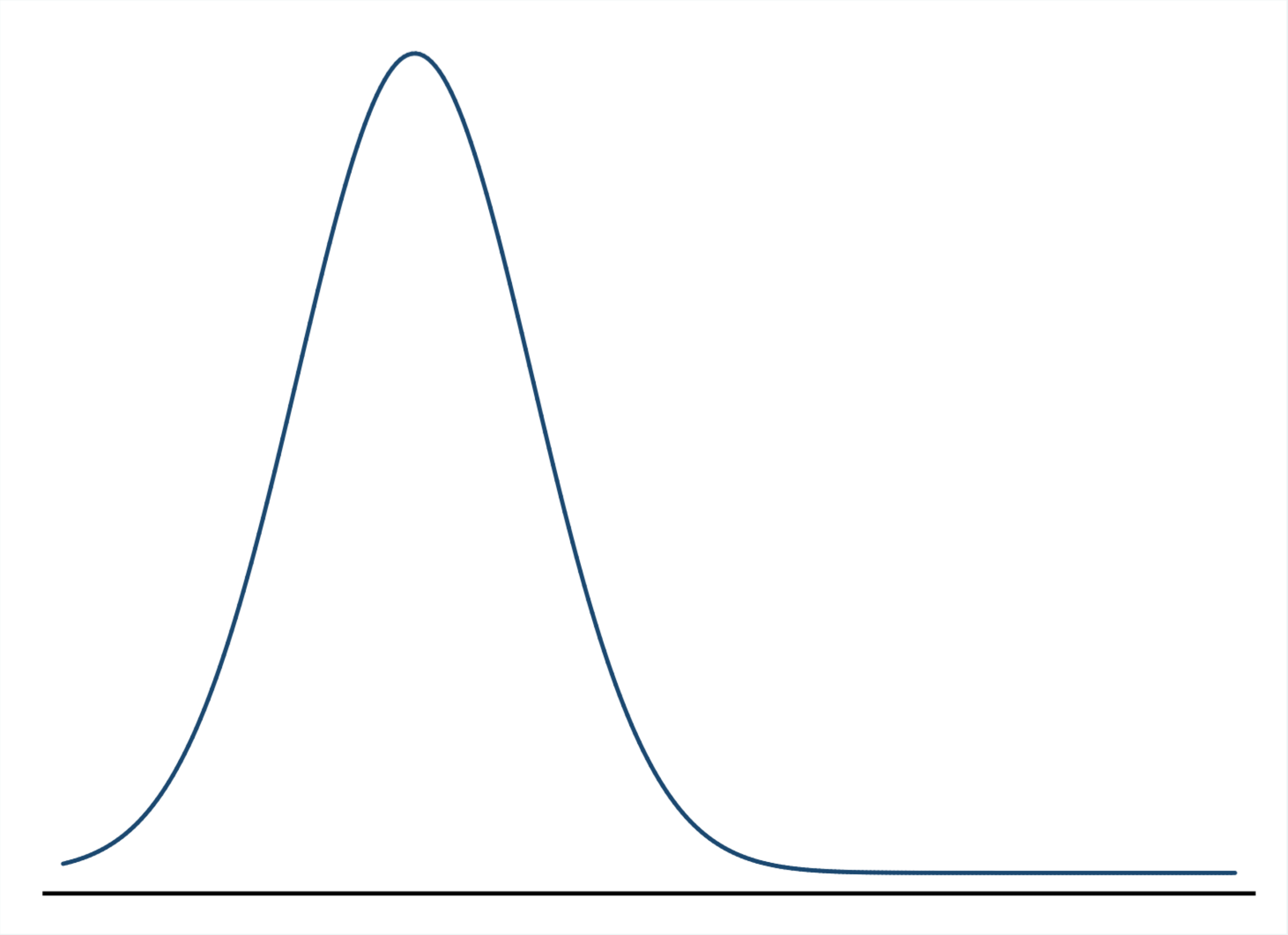
Describing Shape - Normal Distribution



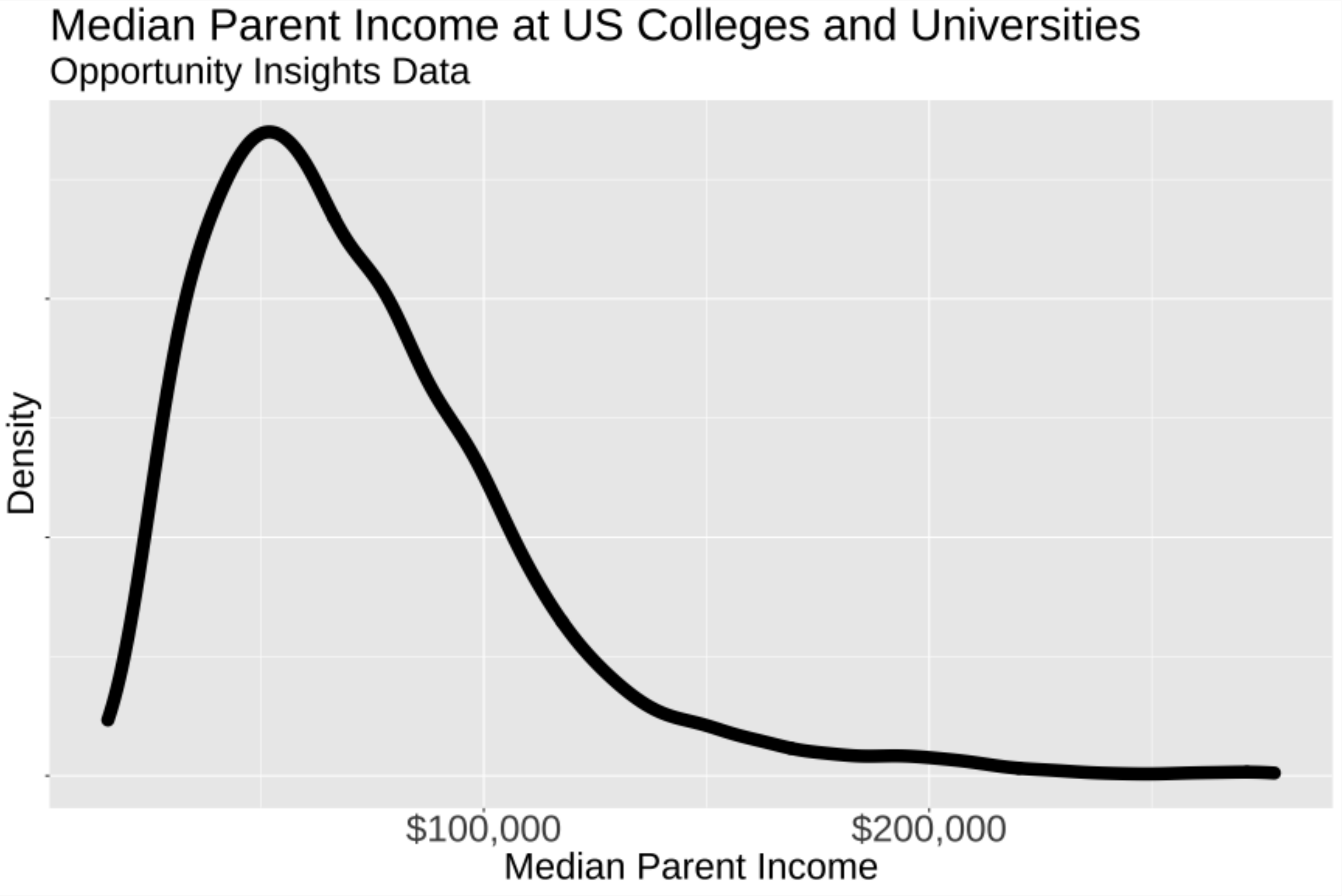
Describing Shape - Left Skew



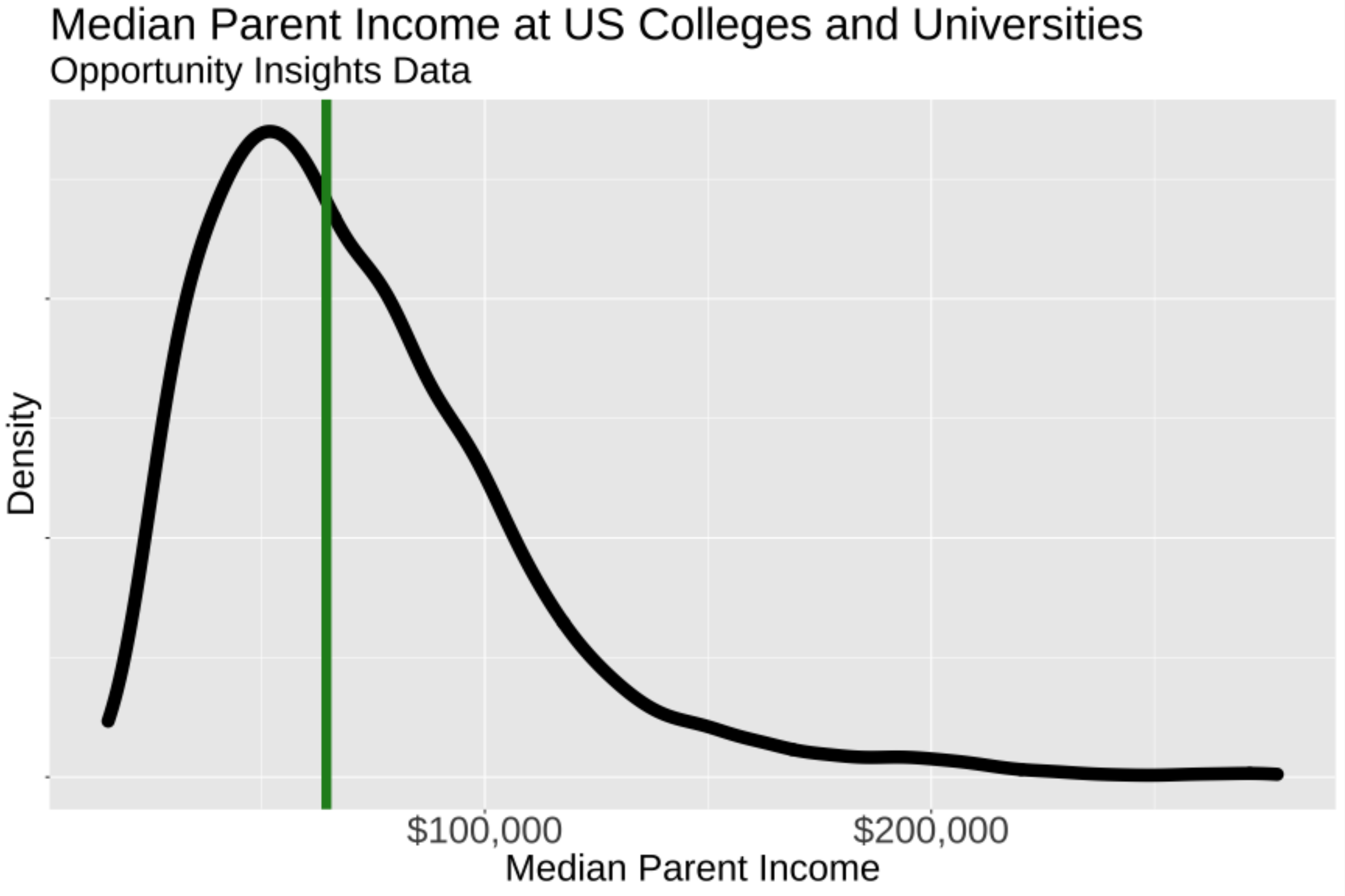
Describing Shape - Right Skew



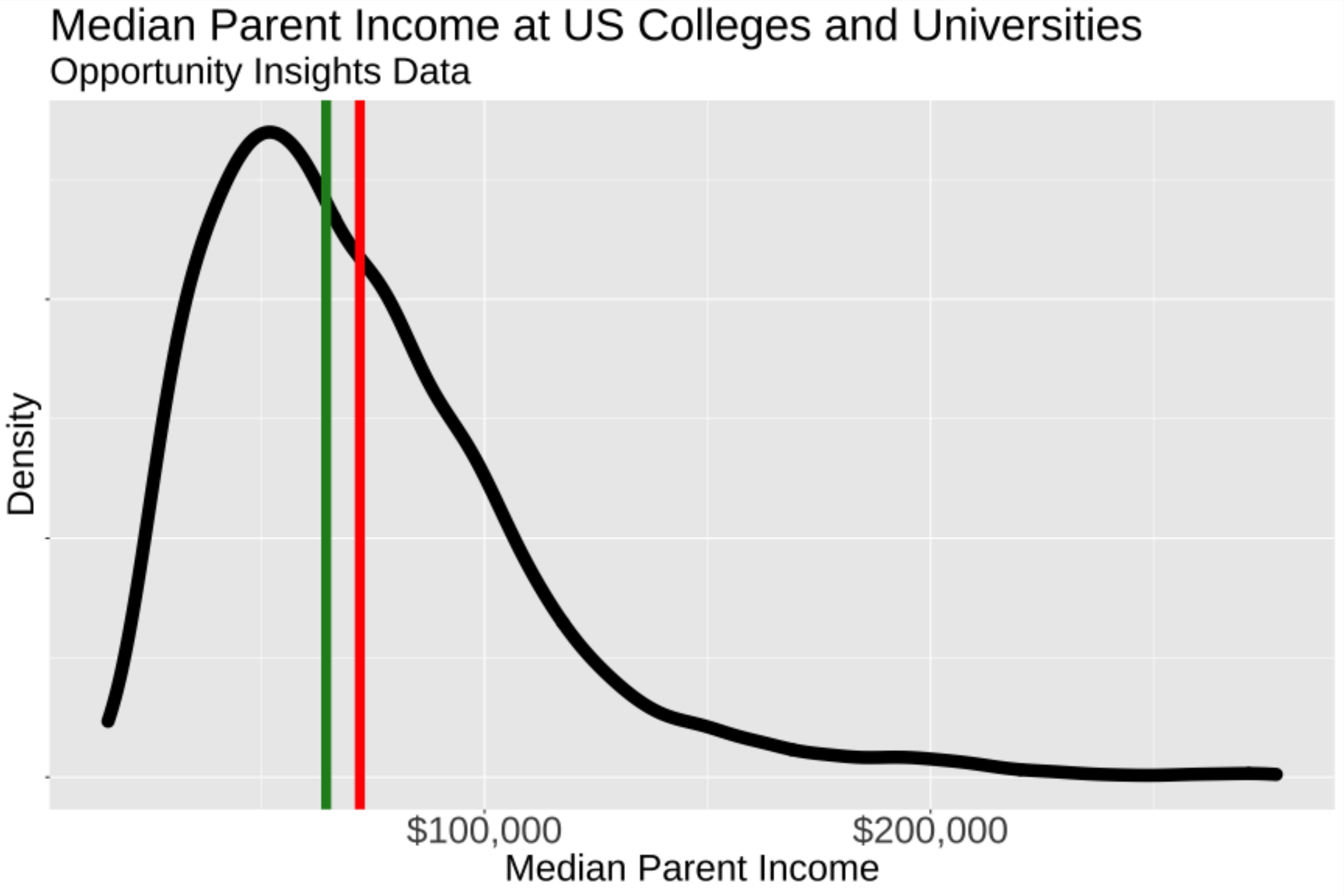
Income Is Often Right Skewed



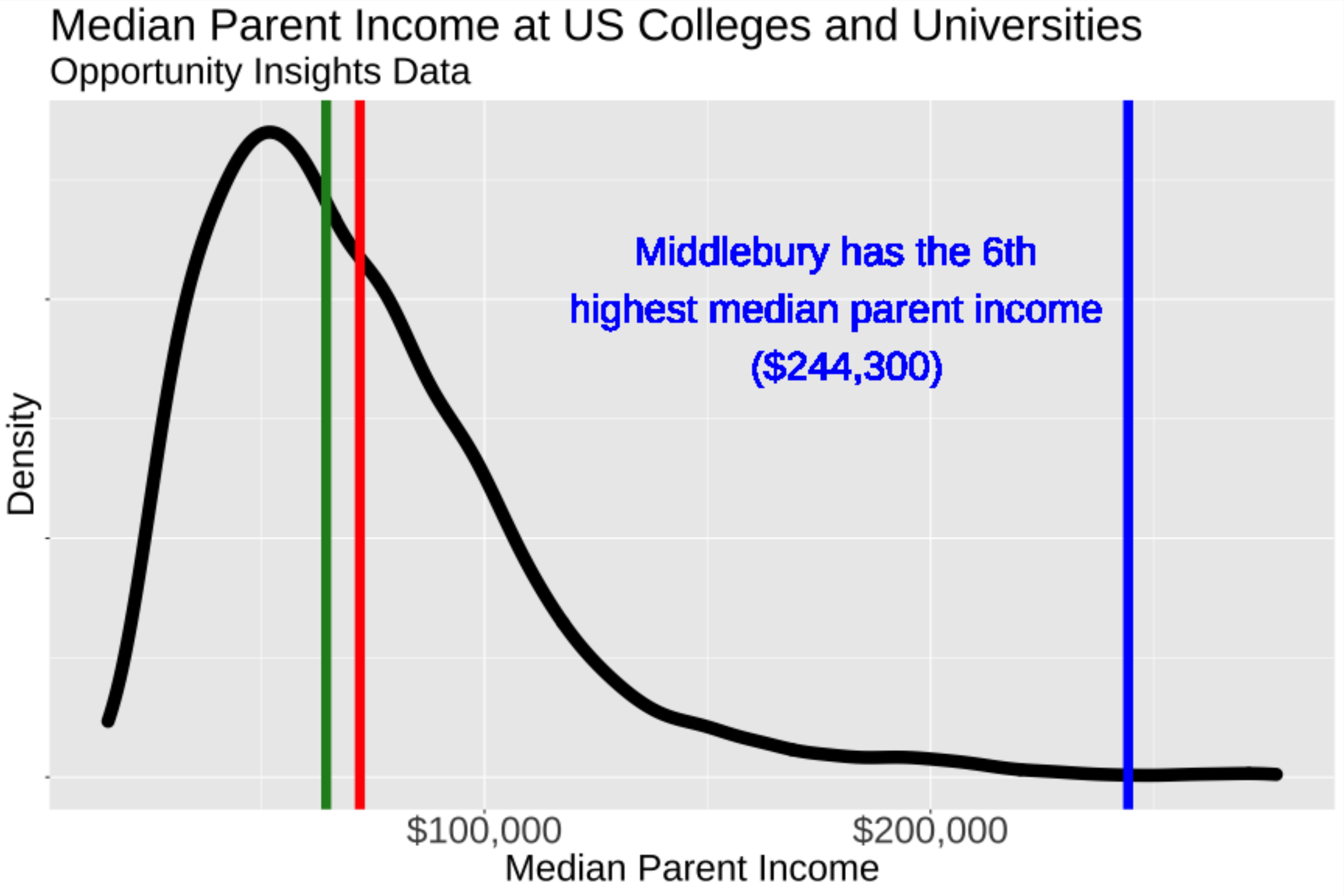
Median Not Centered



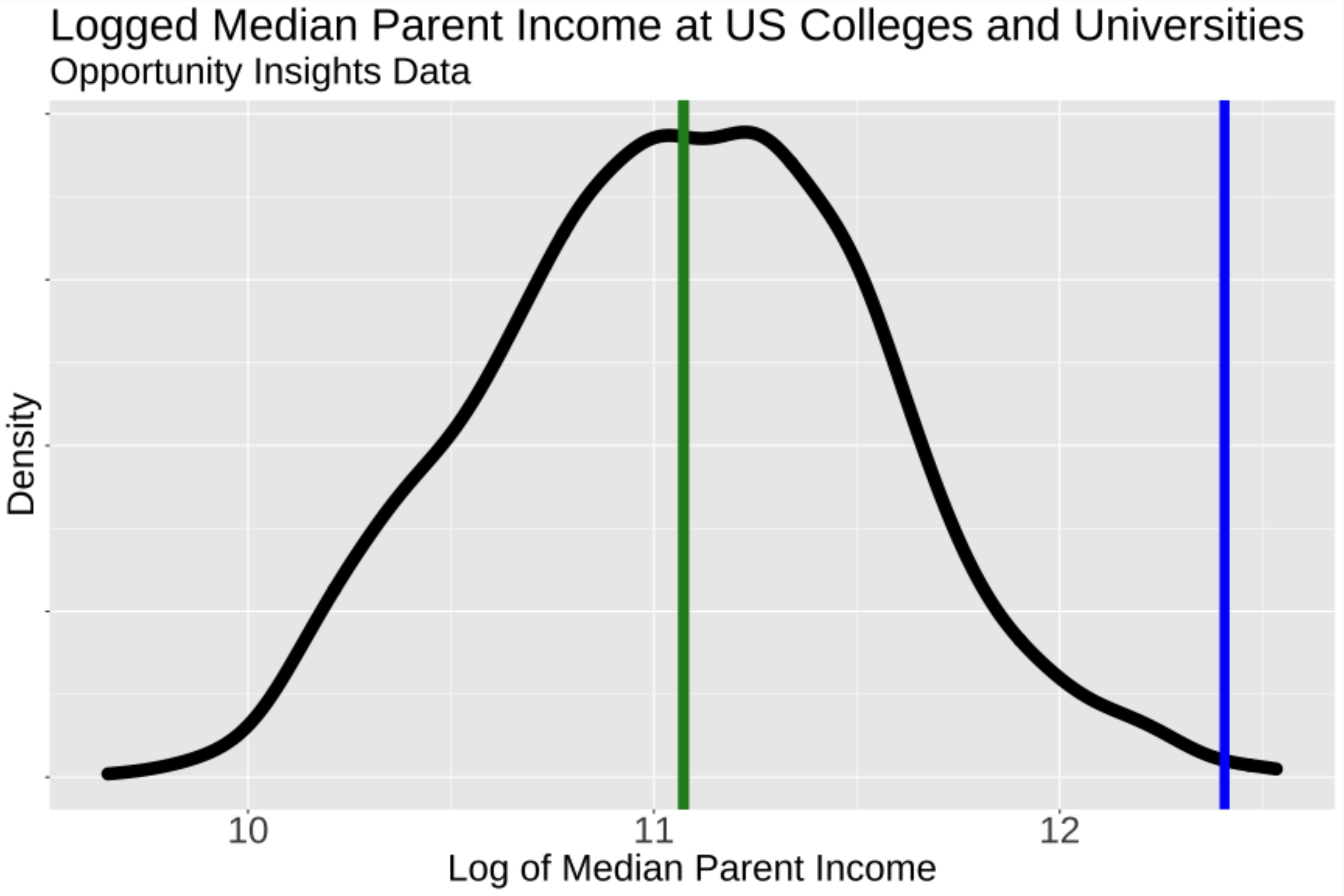
Mean Pulls To Tail



And Pulls To Highest Values



Transforming Skewed Distributions


Logged Median Parent Income at US Colleges and Universities
Opportunity Insights Data
Density
Log of Median Parent Income