IS5 in R: Displaying and Describing Data (Chapter 2)

Margaret Chien and Nicholas Horton (nhorton@amherst.edu)

June 10, 2018

Introduction and background

This document is intended to help describe how to undertake analyses introduced as examples in the Fifth Edition of *Intro Stats* (2018) by De Veaux, Velleman, and Bock. More information about the book can be found at http://wps.aw.com/aw_deveaux_stats_series. This file as well as the associated R Markdown reproducible analysis source file used to create it can be found at http://nhorton.people.amherst.edu/is5.

This work leverages initiatives undertaken by Project MOSAIC (http://www.mosaic-web.org), an NSF-funded effort to improve the teaching of statistics, calculus, science and computing in the undergraduate curriculum. In particular, we utilize the `mosaic` package, which was written to simplify the use of R for introductory statistics courses. A short summary of the R needed to teach introductory statistics can be found in the mosaic package vignettes (http://cran.r-project.org/web/packages/mosaic). A paper describing the mosaic approach was published in the *R Journal*: https://journal.r-project.org/archive/2017/RJ-2017-024.

Chapter 2: Displaying and Describing Data

Section 2.1: Summarizing and Displaying a Categorical Variable

See displays on page 19-23.

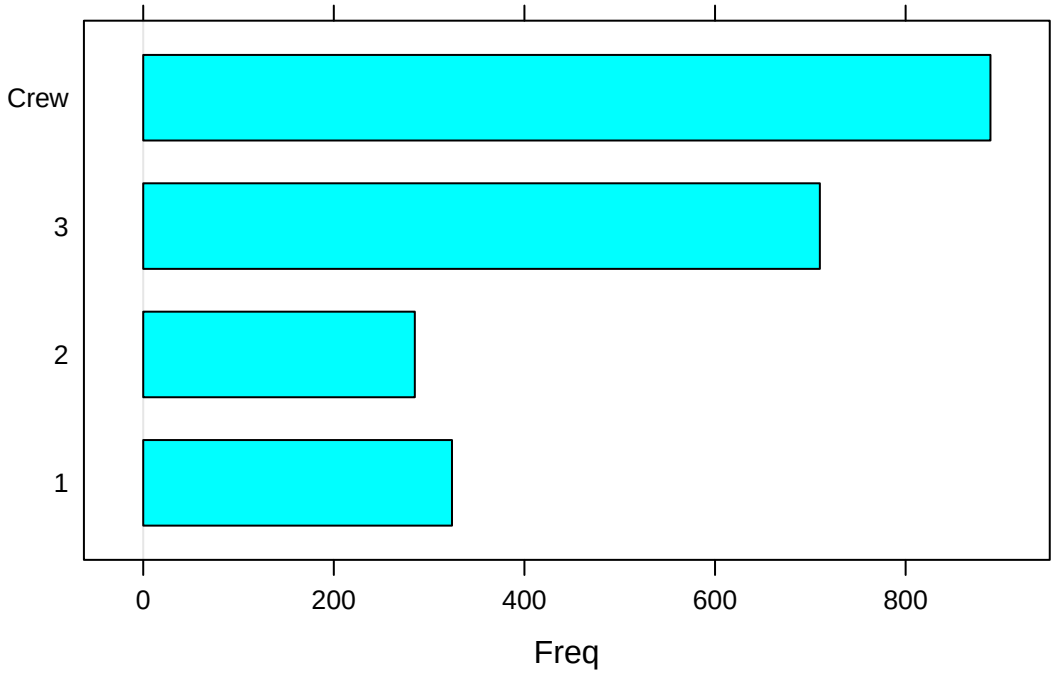
```
library(mosaic)
library(readr)
library(janitor) #for column names
options(digits = 3)
Titanic <- read_csv("http://nhorton.people.amherst.edu/is5/data/Titanic.csv")
tally(~ Class, data = Titanic)
```

```
## Class
##    1    2    3 Crew
##  324  285  710  889
```

```
tally(~ Class, format = "percent", data = Titanic)
```

```
## Class
##    1    2    3 Crew
## 14.7 12.9 32.2 40.3
```

```
barchart(tally(~ Class, data = Titanic))
```

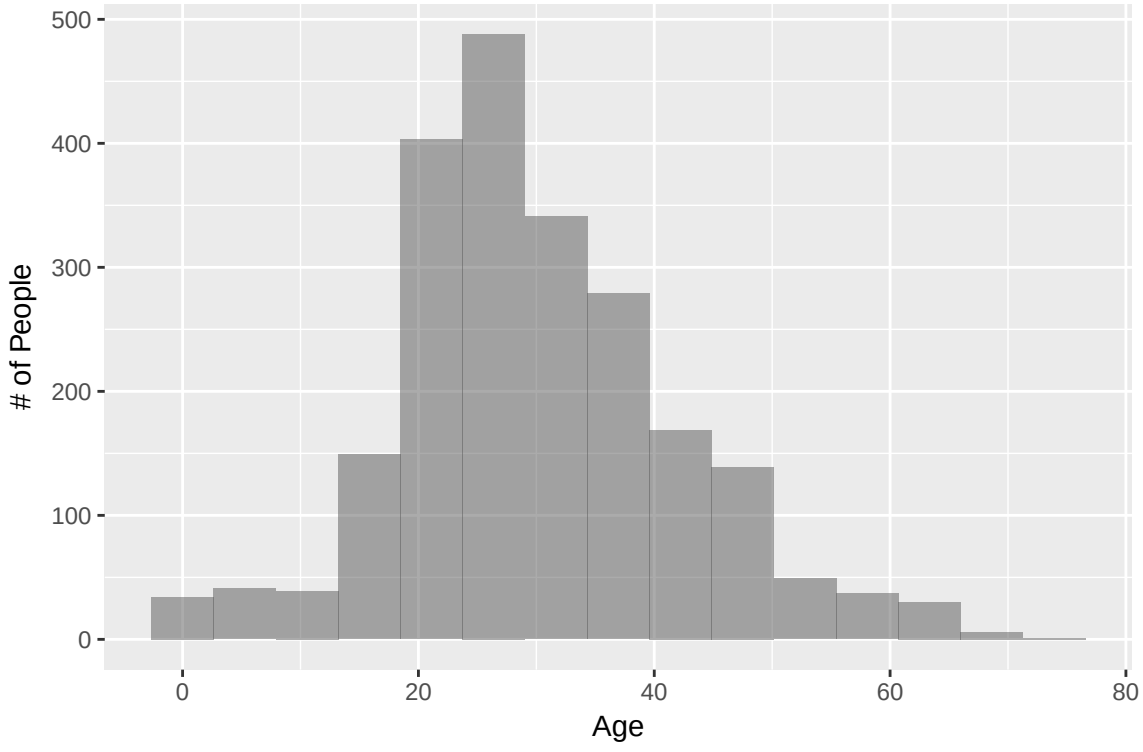


Section 2.2: Displaying a Quantitative Variable

Ages of Those Aboard the the Titanic

```
gf_histogram(~ Age, data = Titanic, bins = 15, ylab = "# of People")
```

```
## Warning: Removed 3 rows containing non-finite values (stat_bin).
```



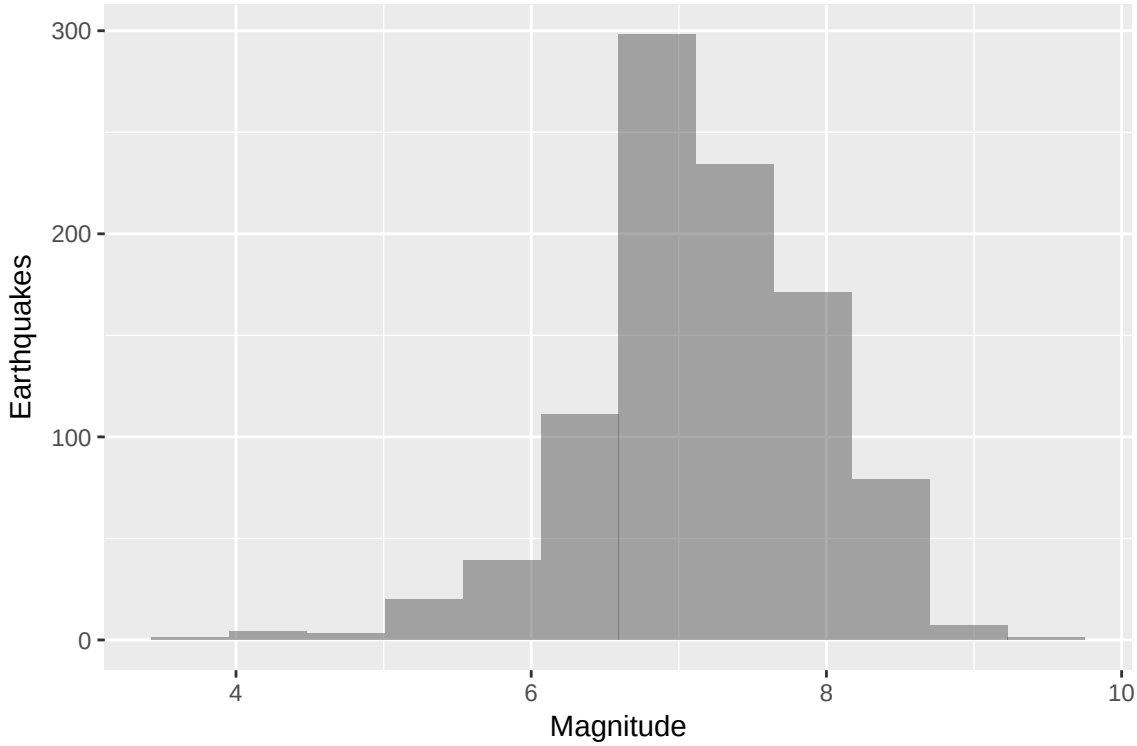
Earthquakes and Tsunamis

```
Earthquakes <- read_csv("http://nhorton.people.amherst.edu/is5/data/Tsunamis_2016.csv")
```

```
## Parsed with column specification:
## cols(
##   Year = col_integer(),
##   Focal_Depth = col_integer(),
##   Primary_Magnitude = col_double(),
##   Country = col_character(),
##   Latitude = col_double(),
##   Longitude = col_double(),
##   Deaths = col_integer(),
##   Missing = col_integer(),
##   Injuriez = col_integer(),
##   `Damage($M)` = col_double()
## )
```

```
gf_histogram(~Primary_Magnitude, data = Earthquakes, bins = 12, ylab = "Earthquakes", xlab = "Magnitude
```

```
## Warning: Removed 119 rows containing non-finite values (stat_bin).
```



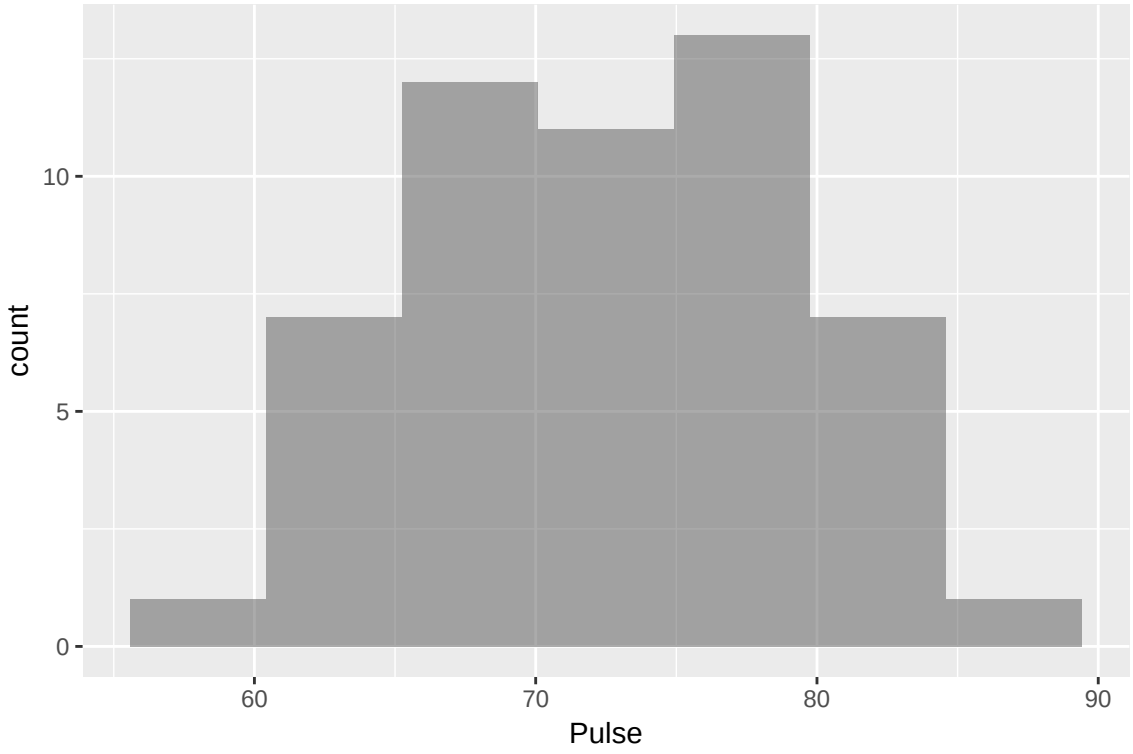
Stem-and-Leaf Displays

See examples in textbook.

```
Pulse_rates <- read_csv("http://nhorton.people.amherst.edu/is5/data/Pulse_rates.csv")
```

```
## Parsed with column specification:
## cols(
##   Pulse = col_integer()
## )
```

```
gf_histogram(~ Pulse, data = Pulse_rates, bins = 7)
```




```
with(Pulse_rates, stem(Pulse))
```

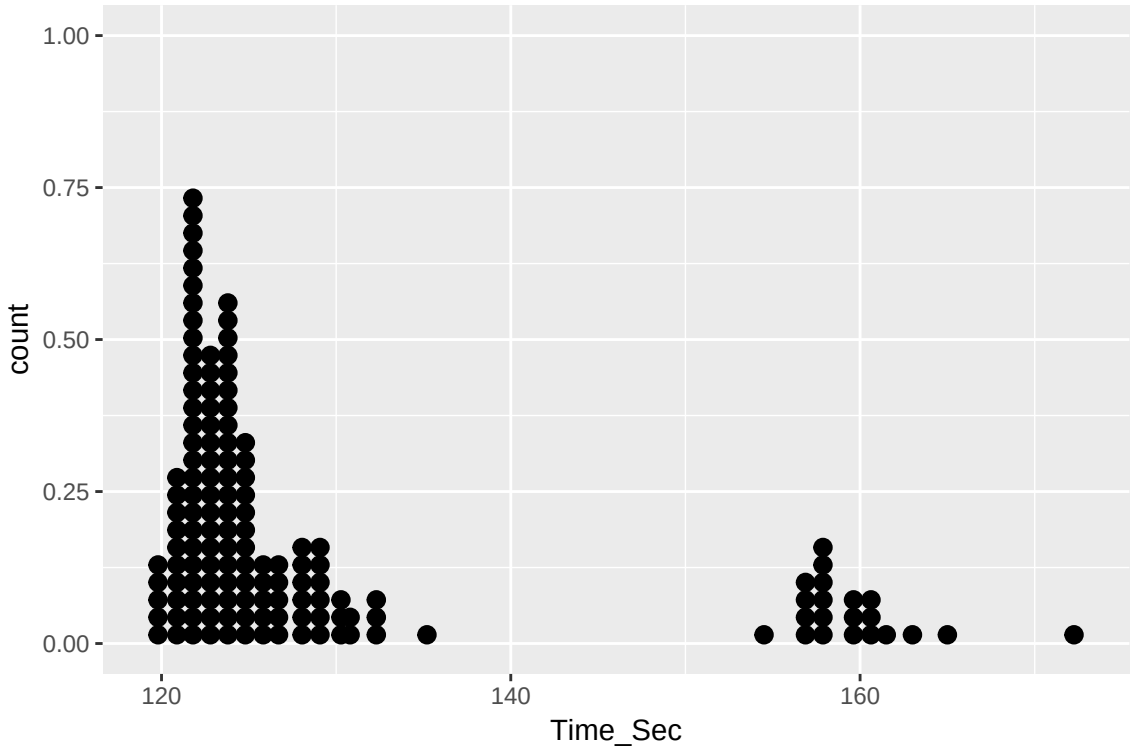
```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   5 | 7
##   6 | 13444
##   6 | 556668888899
##   7 | 0012223333444
##   7 | 5557777888889
##   8 | 0112233
##   8 | 6
```

Dotplot

```
Derby <- read_csv("http://nhorton.people.amherst.edu/is5/data/Kentucky_Derby_2016.csv")
```

```
## Parsed with column specification:
## cols(
##   Year = col_integer(),
##   Year_no = col_integer(),
##   Date = col_character(),
##   Winner = col_character(),
##   Mins = col_integer(),
##   Secs = col_double(),
##   Time_Sec = col_double(),
##   Distance = col_double(),
##   Speed_mph = col_double()
## )
```

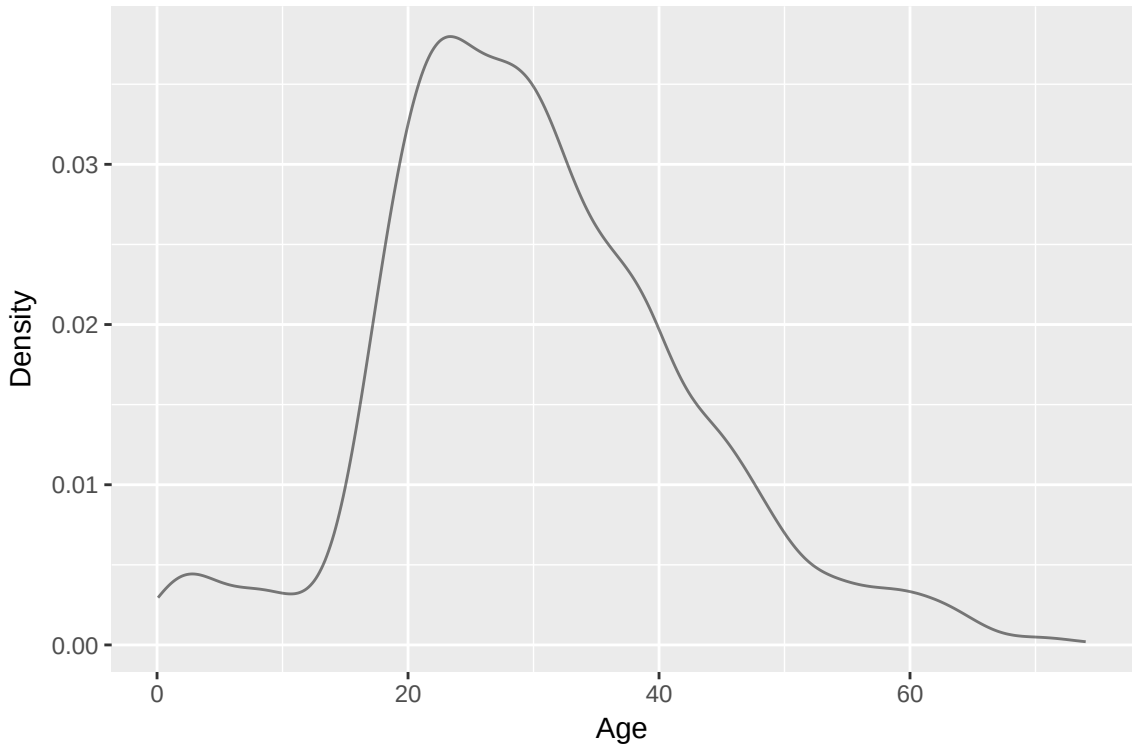
```
gf_dotplot(~ Time_Sec, data = Derby, binwidth = 1)
```



Density Plots

```
gf_dens(~ Age, data = Titanic, ylab = "Density")
```

```
## Warning: Removed 3 rows containing non-finite values (stat_density).
```



Section 2.3: Shape

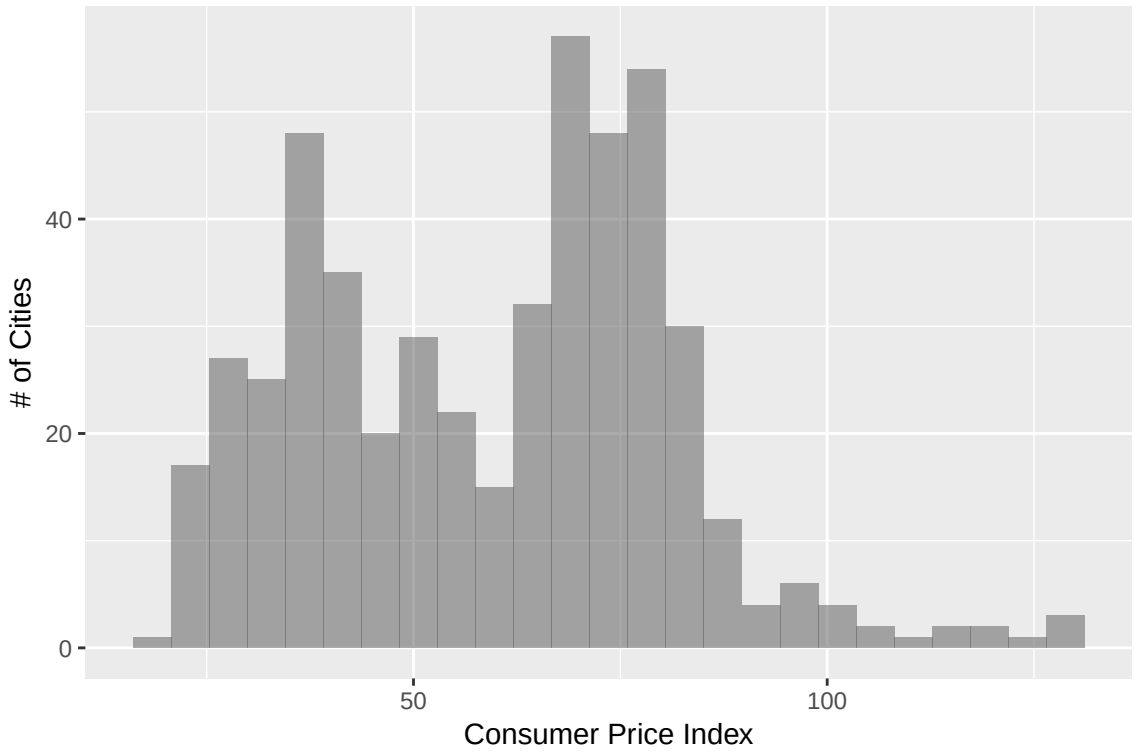
See displays on pages 28-29.

Consumer Price Index

```
CPI <- read_csv("http://nhorton.people.amherst.edu/is5/data/CPI_Worldwide.csv") %>%
  clean_names()
```

```
## Parsed with column specification:
## cols(
##   City = col_character(),
##   Consumer.Price.Index = col_double(),
##   Rent.Index = col_double(),
##   Consumer.Price.Plus.Rent.Index = col_double(),
##   Groceries.Index = col_double(),
##   Restaurant.Price.Index = col_double(),
##   Local.Purchasing.Power.Index = col_double()
## )
```

```
gf_histogram(~ consumer_price_index, data = CPI, ylab = "# of Cities",
             xlab = "Consumer Price Index")
```



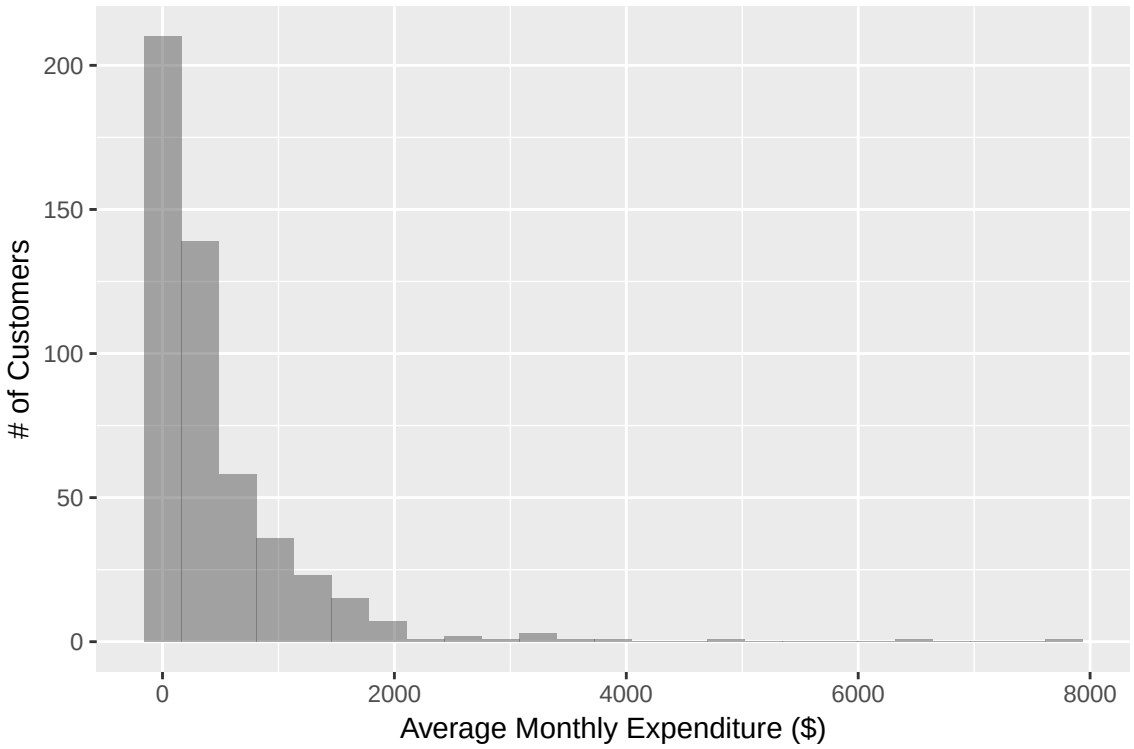
We can use `clean_names()` from the `janitor` package to format the names of the columns when necessary.

Credit Card Expenditures

```
CreditCardEx <- read_csv("http://nhorton.people.amherst.edu/is5/data/Credit_card_charges.csv") %>%
  clean_names()
```

```
## Parsed with column specification:
## cols(
##   `Charges($)` = col_double()
## )
```

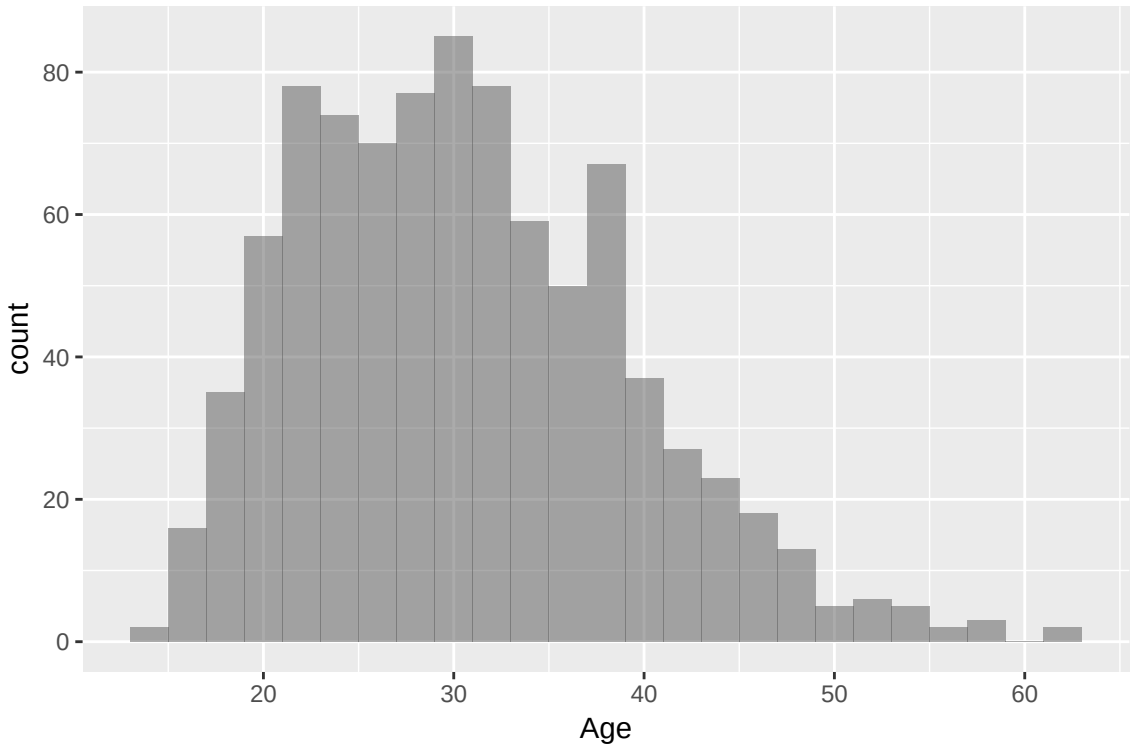
```
gf_histogram(~ charges, data = CreditCardEx, ylab = "# of Customers",
             xlab = "Average Monthly Expenditure ($)")
```



Section 2.4: Center

Finding Median and Mean

```
TitanicCrew <- filter(Titanic, Class == "Crew")
gf_histogram(~ Age, data = TitanicCrew)
```




```
favstats(~ Age, data = TitanicCrew)
```

```
##  min Q1 median Q3 max mean   sd   n missing
##   14 24     30 37  62 31.1 8.55 889       0
```

Section 2.5: Spread

The Range

```
range(~ Age, data = TitanicCrew)
```

```
## [1] 14 62
```

```
diff(range(~ Age, data = TitanicCrew))
```

```
## [1] 48
```

The Interquartile Range

```
favstats(~ Age, data = TitanicCrew)
```

```
##  min Q1 median Q3 max mean   sd   n missing
##   14 24     30 37  62 31.1 8.55 889       0
```

```
IQR(~ Age, data = TitanicCrew)
```

```
## [1] 13
```

We can find IQR by subtracting Q1 from Q3.

Standard Deviation

```
sd(~ Age, data = TitanicCrew)
```

```
## [1] 8.55
```

```
var(~ Age, data = TitanicCrew)
```

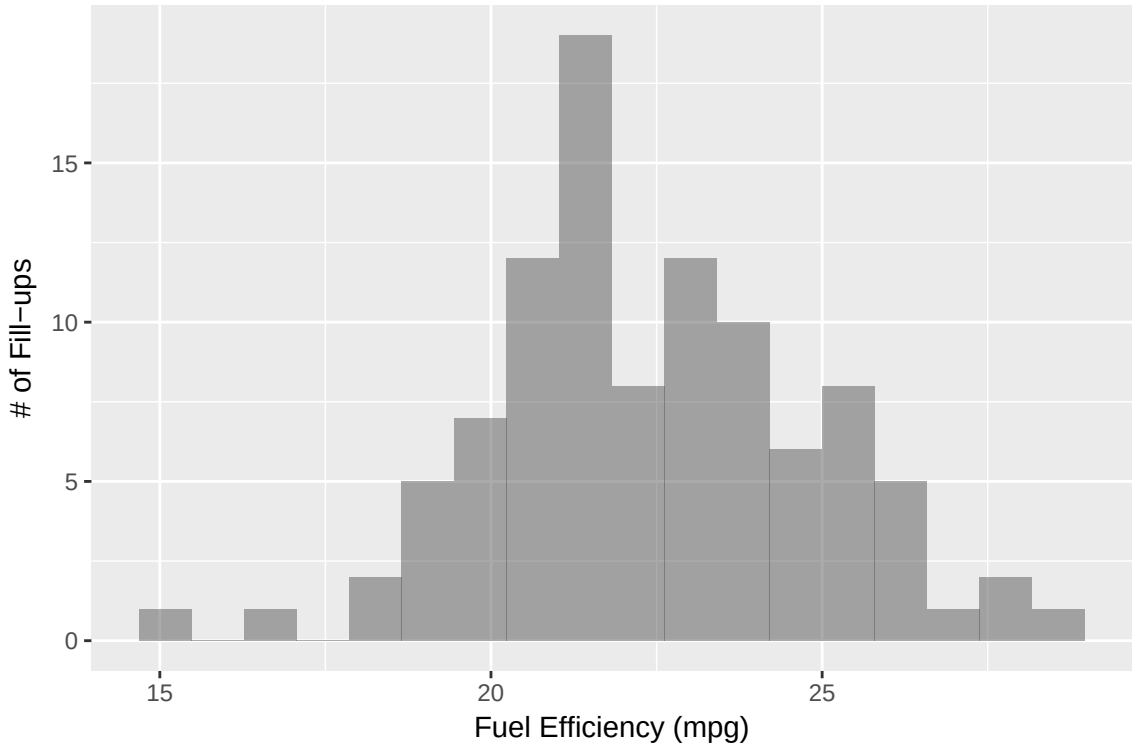
```
## [1] 73.1
```

Summarizing a Distribution

```
Nissan <- read_csv("http://nhorton.people.amherst.edu/is5/data/Nissan.csv")
```

```
## Parsed with column specification:
## cols(
##   mpg = col_double()
## )
```

```
gf_histogram(~ mpg, data = Nissan, bins = 18, xlab = "Fuel Efficiency (mpg)",
             ylab = "# of Fill-ups")
```



```
favstats(~ mpg, data = Nissan)
```

```
##   min   Q1 median Q3  max mean   sd   n missing
##  14.7 20.8   22.1 24 28.2 22.4 2.45 100       0
```

Random Matters

```
Commute <- read_csv("http://nhorton.people.amherst.edu/is5/data/Population_Commute_Times.csv") %>%
  clean_names()
```

```
## Parsed with column specification:
## cols(
##   Commute.Time = col_integer()
## )
```

```
gf_histogram(~ commute_time, data = Commute, binwidth = 10, xlab = "Commute Time (min)",
             ylab = "# of Employees")
```