2020FA MATH 1051H: Lecture #05

Statistics and Summary Plots

Today’s Review

In today’s lecture, we are going to review the ideas of Chapters 2.1 and 2.2 of our textbook, with several worked examples. The code demonstrated today will be gone over again in Workshops #03 and #04, in different ways, so that hopefully the combination will allow you to grasp how to work with things.

To Start … Data

To work with data in R, it must be organized in some fashion. The two objects we will use the most to do this are:

- **The Vector**: a comma-separated list of elements
- **The Data Frame**: a bunch of vectors, organized as columns in a spreadsheet-like object

The data.frame is more complicated, and I'll show you a quick demo of it today as we extract data, but we'll be **working** with the vector for your assignments for a while.

Example 1: Simulated Data

Our first example will use simulated data to explore the ideas. We will generate 100 randomly selected integers between 1 and 1000. This is done using the **sample()** function.

```
dat1 <- sample(x = seq(from = 1, to = 1000, by = 1),
               size = 100,
               replace = TRUE)
str(dat1)
```

```
##  num [1:100] 450 346 336 247 859 174 453 69 476 546 ...
```

Now that we have the data, let’s summarize it.

Example 1: Summary Statistics

```
summary(dat1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    13.0   174.0   451.5   473.2   754.2   999.0
```

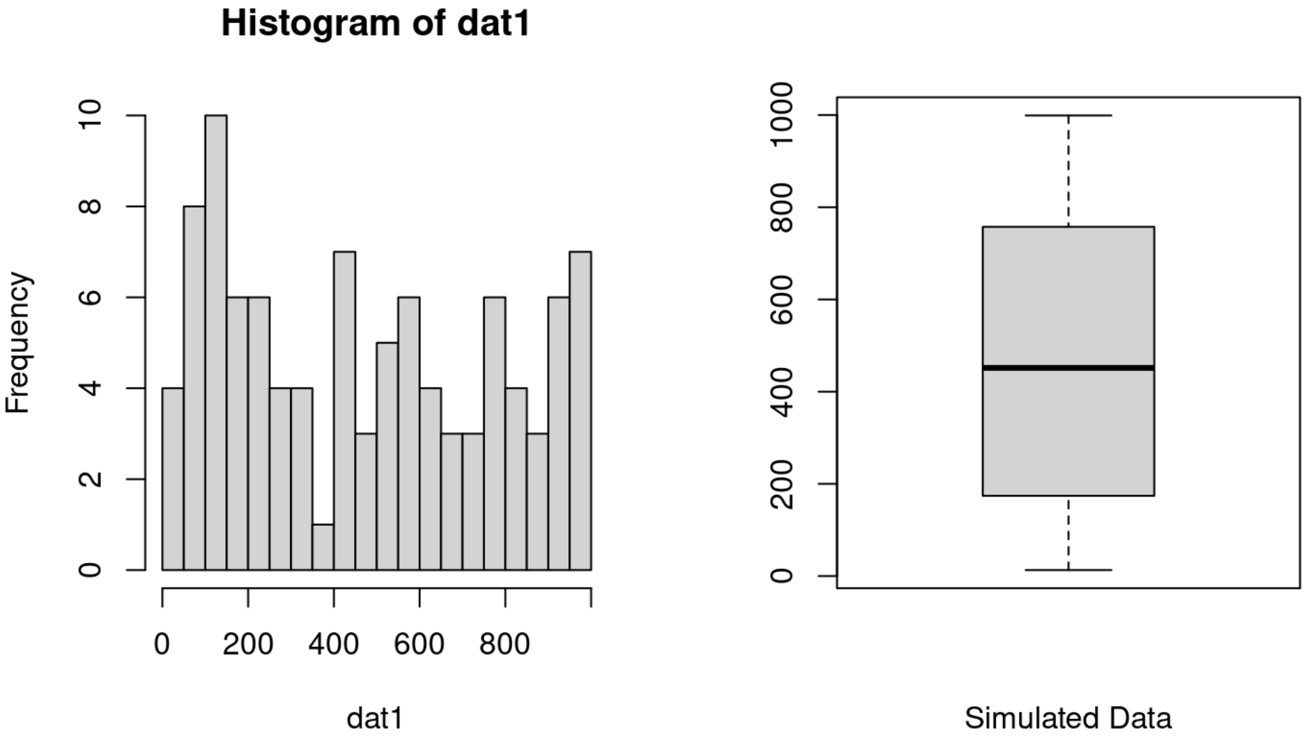
```
sd(dat1)
```

```
## [1] 308.5733
```

From this summary we see that the sample of 100 observations ranges from 13 to 999, with median 451.5 (there are an even number of observations) and mean 473.2. The mean is larger than the median: what does this imply might be happening?

Example 1: Summary Plots

```r
par(mfrow = c(1, 2))
hist(dat1, breaks = 20)
boxplot(dat1, xlab = "Simulated Data")
```



Example 1: What Can We Say?

From the two plots and the summary statistics, we might say:

- the data is either uniform or unimodal
- there is a very slight rightward skew, but not a strong one
- there are no obvious outliers
- 50% of the data is between 174 and 754.2

Example 2: Global Population

The World Health Organization Global Tuberculosis Report contains information on the cases of tuberculosis worldwide. It also include countries’ names and their populations, which is what we will look at.

```
library("tidyr")
data(who)
data(population)
dat2 <- population$population
str(dat2)
```

```
##  int [1:4060] 17586073 18415307 19021226 19496836 19987071 20595360 21347782 22202806 23116
```

Example 2: Global Population

The population data.frame consists of 4,060 observations of country+year+population. We extracted only the population numbers, so we now have 4,060 populations for these countries over time.

Example 2: Summary Statistics

```
summary(dat2)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
## 1.129e+03 6.029e+05 5.319e+06 3.003e+07 1.855e+07 1.386e+09
```

```
sd(dat2)
```

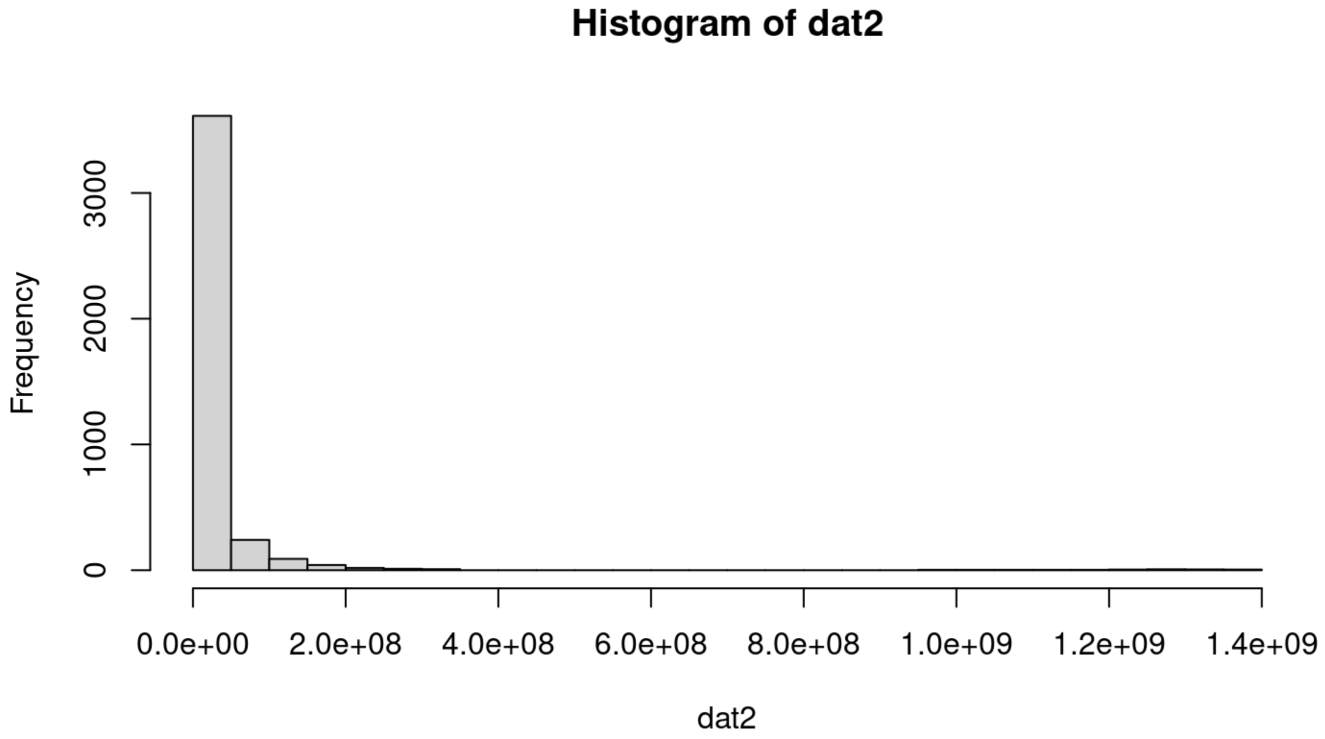
```
## [1] 121347349
```

In this summary, things are presented in scientific notation: any big number (or small number) in R will be presented this way. We see that the 4,060 observations range from 1.129e+03 (1,129) to 1.386e+09 (1,386,000,000). Do you know which countries these might be?

The mean is clearly **much** greater than the median here (almost 10x larger).

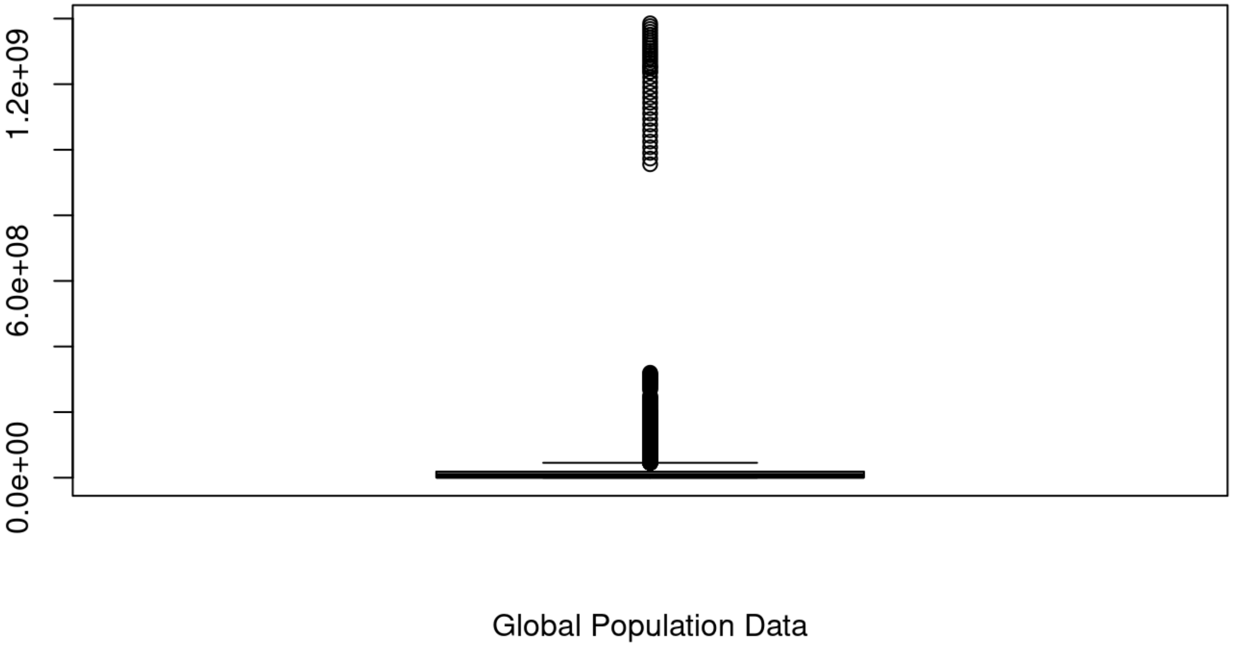
Example 2: Summary Plots

```
hist(dat2, breaks = 40)
```



Example 2: Summary Plots

```
boxplot(dat2, xlab = "Global Population Data")
```



Example 2: What Can We Say?

From the two plots and the summary statistics, we might say:

- the data has a strong right skew
- there are obvious outliers
- 50% of the data is between 602,900 and 18,550,000
- 25% of the data is greater than 18,550,000

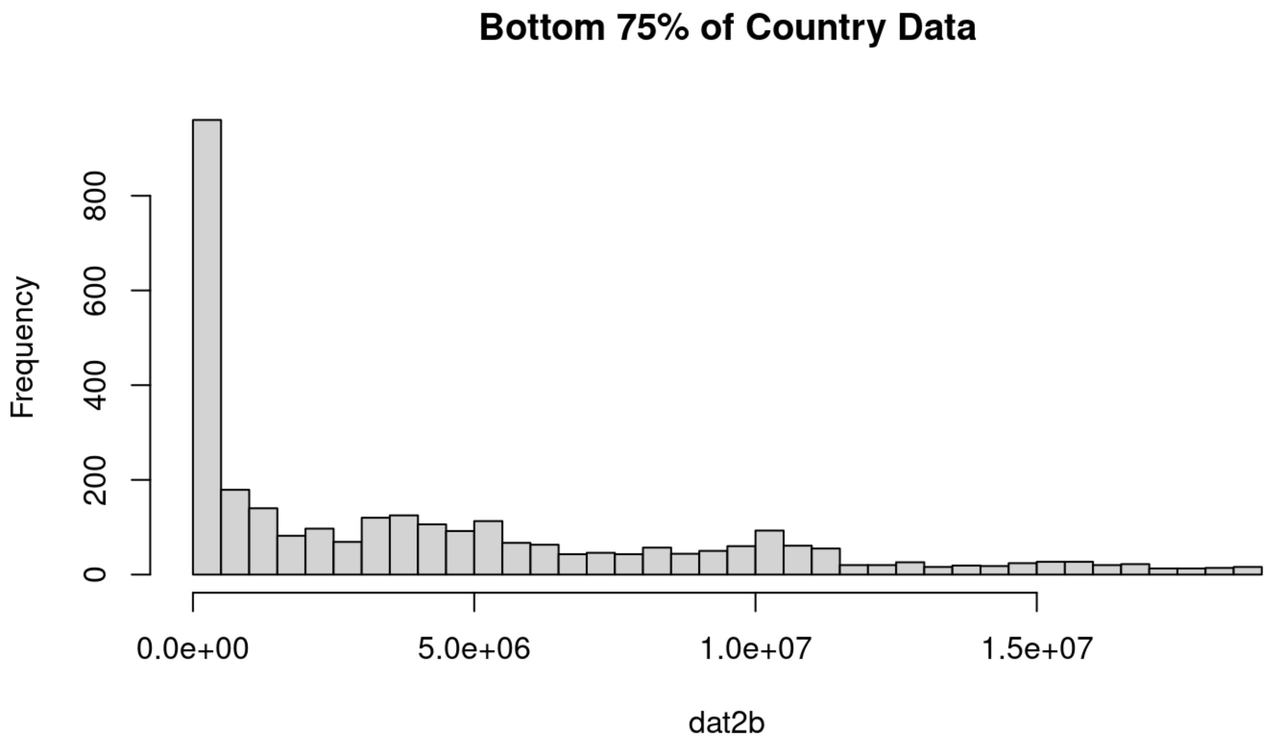
Example 2: What’s Happening?

There’s too big a range in the data! 25% of the data is above 18 million, meaning 75% of it is below 18 million … but above 18 million are many of the countries you know of (e.g., Canada, USA, Australia, Brazil, Russia, China, India, Germany, France, the UK, etc.).

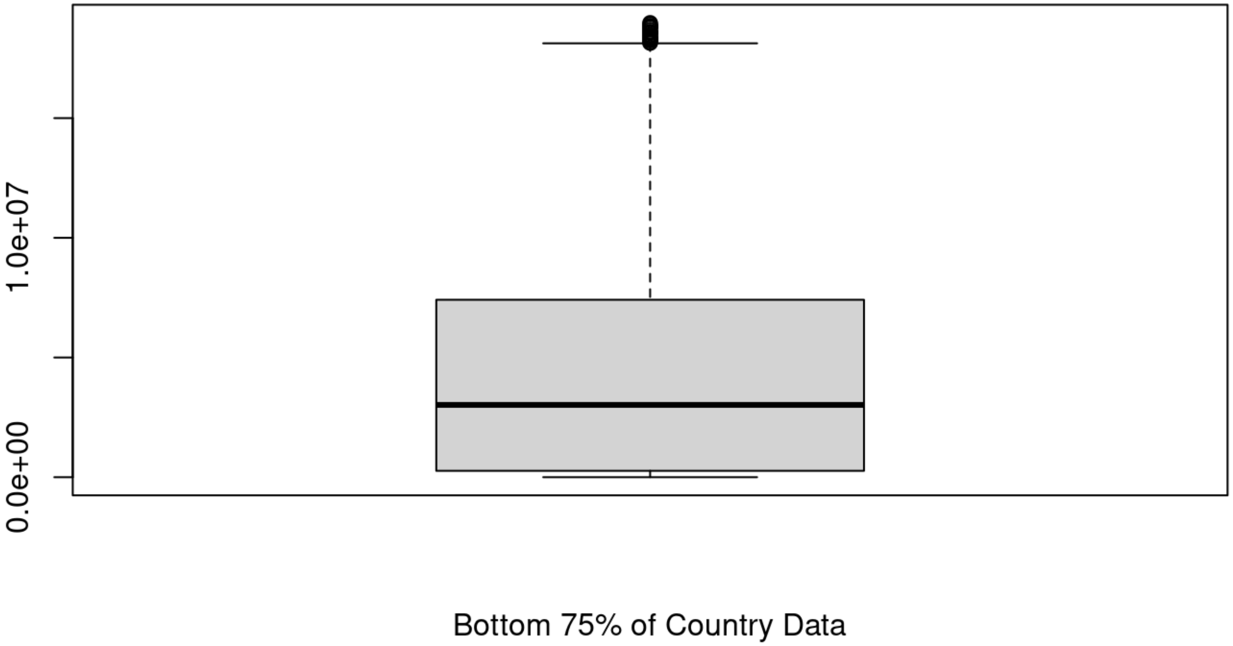
So this data is very odd … let’s take a closer look.

Example 2: Advanced (Bottom 75%)

I don't expect you to follow the code, so I've hidden it here. What I'm going to do is look at only the observations that are less than 19 million (the bottom 75%) alone, and then the upper 25%.

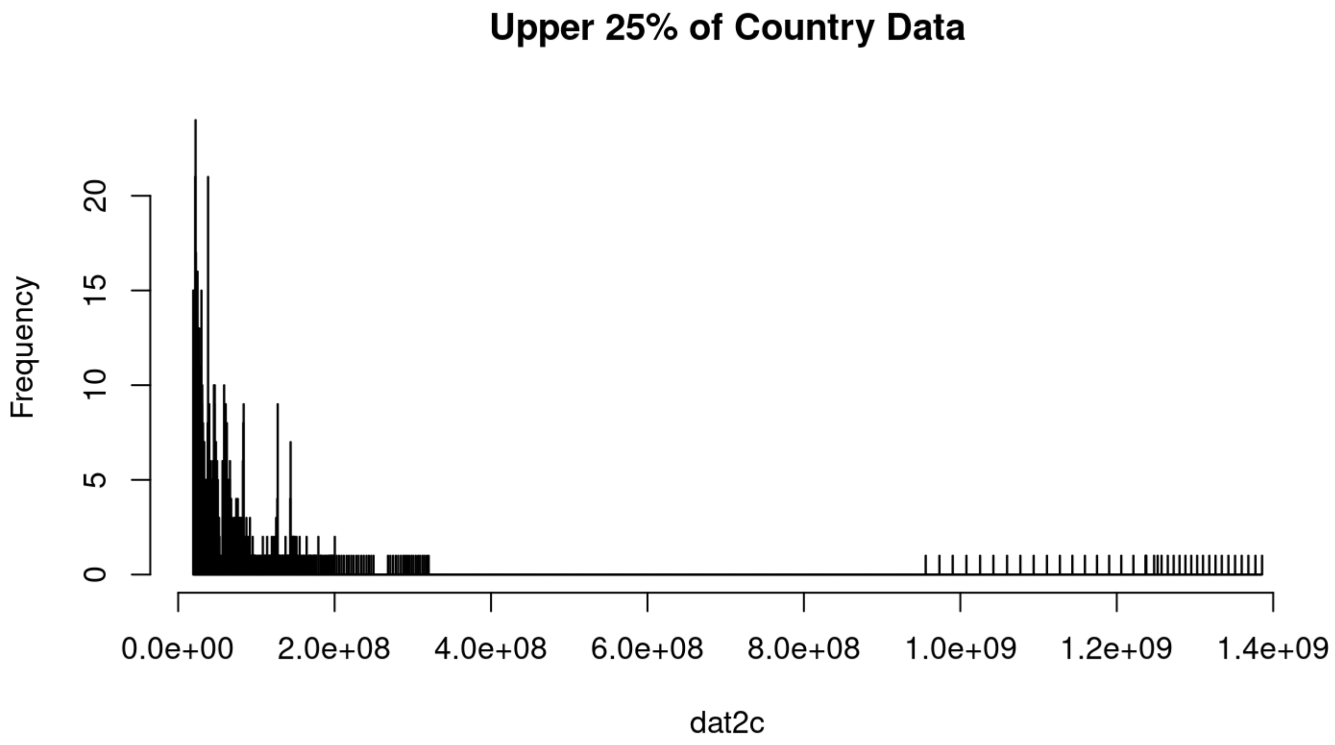


Example 2: Advanced (Bottom 75%)



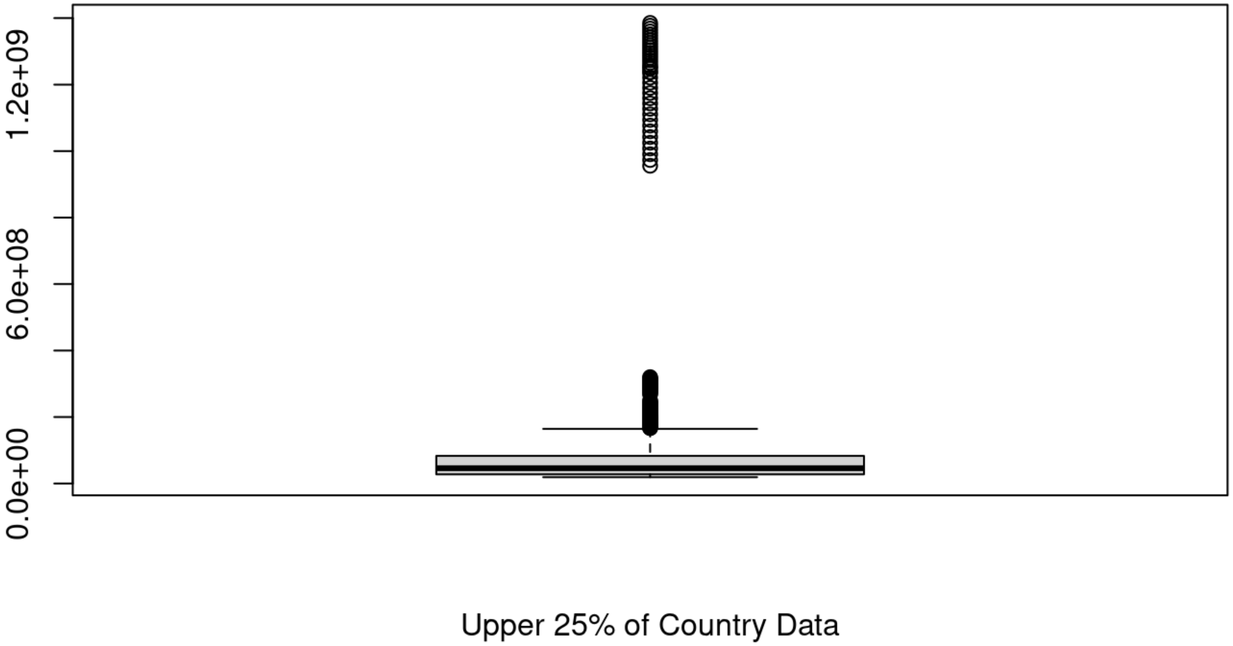
Example 2: Advanced (Top 25%)

So what does this show? We **still** have a huge number of really small observations, easily 1/4 of the entire data set, and then a very long right tail. What about the upper 25%?



Example 2: Advanced (Top 25%)

So the histogram shows that after the bottom 75%, things start to be very few cases in each category - the bins of the two plots are the same width, but things are far more spaced out on the upper end.



Example 2: What Can We Say?

From our expanded analysis:

- the data has a strong right skew, with lots of outliers
- 25% of the data is greater than 18,550,000, but is very spread out
- even the 75% of the data that is less than 18,550,000 is very strongly right-skewed

Example 3: Traffic Fatalities

There is an Applied Econometrics package in R which contains some interesting data. One of the sets is traffic fatalities for the “lower 48” United States states for the 7 years from 1982 to 1988 ($48 \times 7 = 336$).

```
library("AER")
data("Fatalities")
dat3 <- Fatalities$fatal
str(dat3)
```

```
##  int [1:336] 839 930 932 882 1081 1110 1023 724 675 869 ...
```

Example 3: Summary Statistics

```
summary(dat3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    79.0   293.8   701.0   928.7  1063.5  5504.0
```

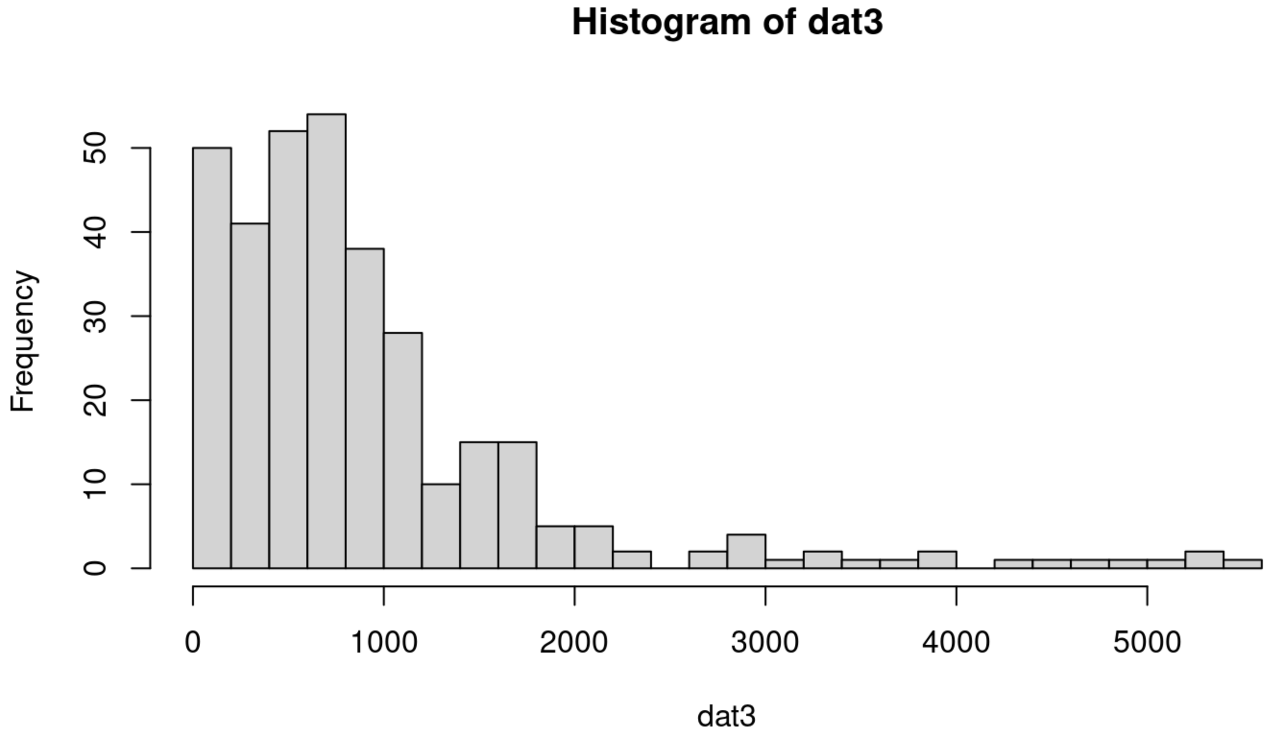
```
sd(dat3)
```

```
## [1] 934.0515
```

In this summary, the mean is clearly much greater than the median, so we expect a right skew. The range is quite large as well: a minimum of 79 fatalities in a given year and state, and a maximum of 5504 fatalities in another state and a (not necessarily the same) year.

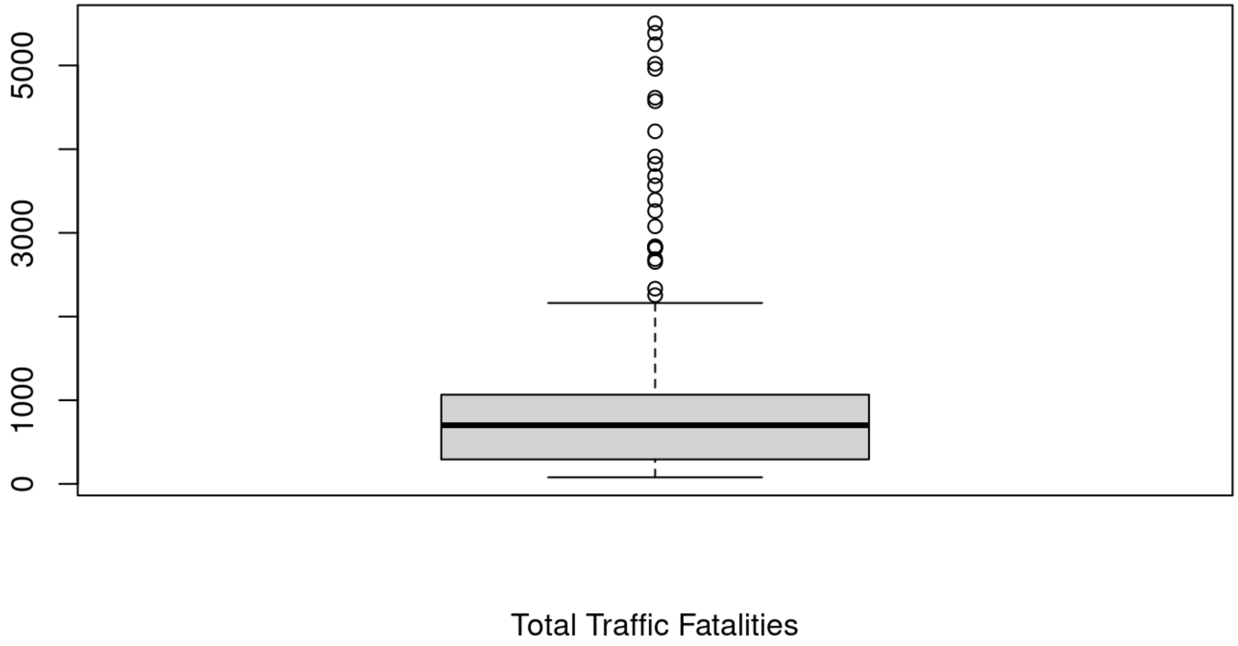
Example 3: Summary Plots

```
hist(dat3, breaks = 20)
```



Example 3: Summary Plots

```
boxplot(dat3, xlab = "Total Traffic Fatalities")
```



Example 3: What Can We Say?

- left truncated - definitely not symmetric
- clear right skew
- many outliers (above about 2500)
- unimodal

Example 3: Comparing IQR and SD

```
sd(dat3)
```

```
## [1] 934.0515
```

```
IQR(dat3)
```

```
## [1] 769.75
```

Notice how the IQR is quite a bit smaller than the standard deviation? This is the **robustness** discussion from Lecture 04.

Summary

What is the point?

When you get a data set, and you’ve never seen it before, you need to explore it. The demonstrations today (and on Tuesday and Monday) should give you a hint as to how to start doing that. There are many possible visualizations, especially when there are multiple variables, but for single variables, the key things to check are:

- summary statistics: mean, median, quantiles, IQR, sd
- the histogram (possibly experiment with multiple breakpoints)
- the boxplot (to see outliers and skew)

Next?

In Workshops #03 and #04, we’ll start unpacking these skills, and giving you the chance to learn them well, and practice practice practice. You should expect to use these skills on R Assignments #2, 3 and 4.