More Numerical and Graphical Summaries using Percentiles

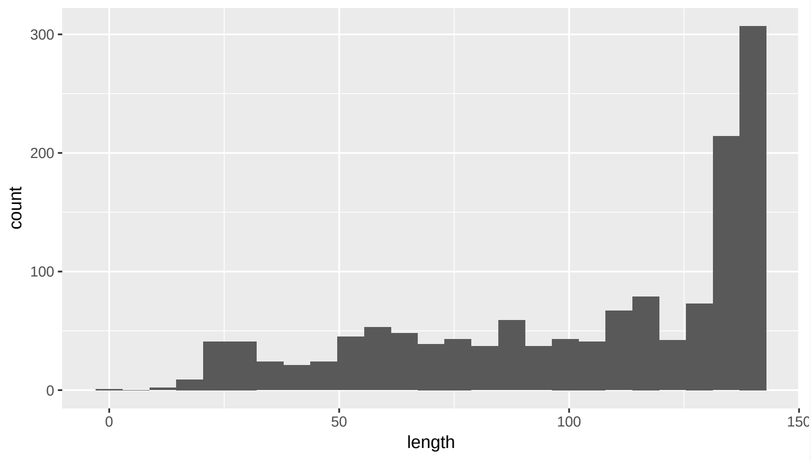
David Gerard

2017-09-18

Learning Objectives

- Percentiles
- Five Number Summary
- Boxplots to compare distributions.
- Sections 1.6.5 and 1.6.6 in DBC.

Trump's Tweet Length



- Mean = 102.7281, median = 114.5
- Standard deviation = 37.4711, MAD = 36.3237

Are these sufficient summaries?

- Tells us nothing about the left skew.
- Doesn't tell us that a fourth of all tweets are greater than 138 characters.
- Doesn't tell us that small tweets are quite rare.

Percentiles

percentile

The *p*th percentile of a distribution is the value that has *p* percent of the observations fall at or below it. To calculate the percentile, arrange the observations in increasing order and count up the required percent from the bottom of the list.

Why do we care?

- If we know a few percentiles, that gives us an idea of the shape of a distribution.
- Knowing the **same** percentiles of two distributions makes it easy to quickly compare them.
- It's usual to return the 0th (= minimum), 25th, 50th (= median), 75th, and 100th (= maximum) percentiles.

Quartiles

- The 25th and 75th percentiles have special names:

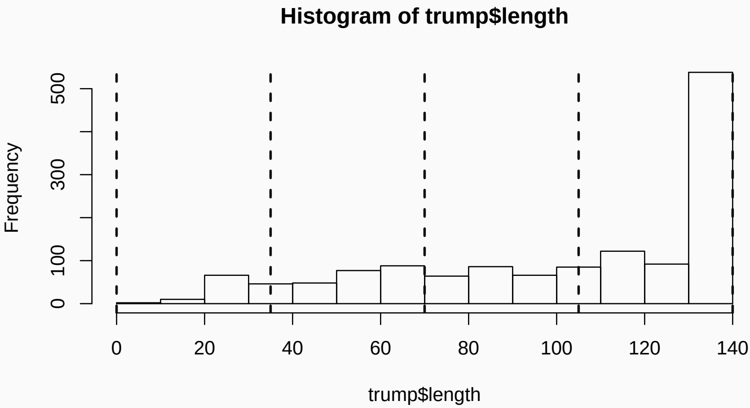
Quartiles

The first quartile Q_1 is the 25th percentile. It is the median of the lower half of the data.

The third quartile Q_3 is the 75th percentile. It is the median of the upper half of the data.

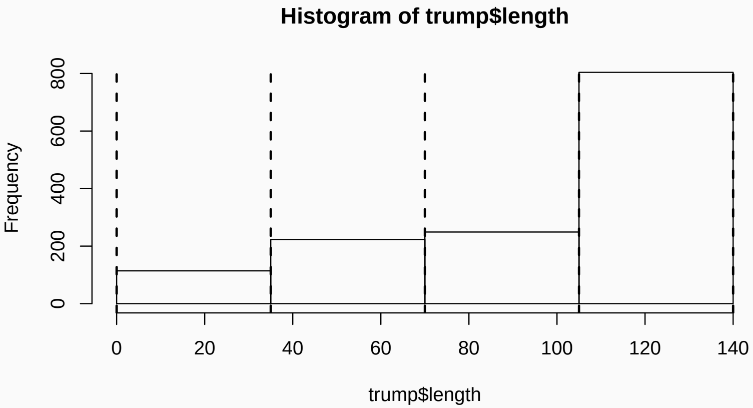
Example: Trump's Tweet Length

These ARE NOT the qartiles of Trump's tweet length



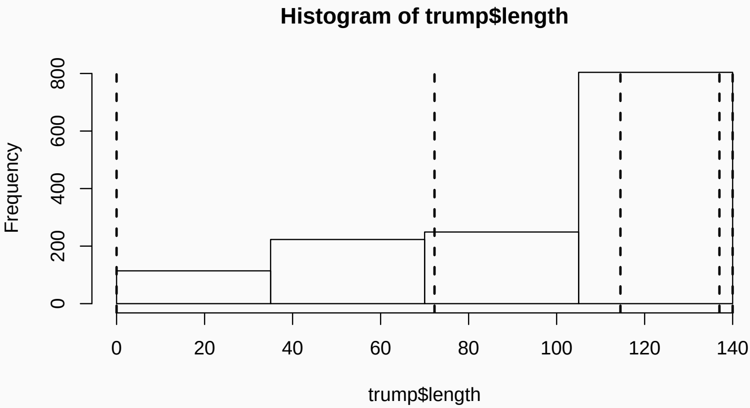
Example: Trump's Tweet Length

These ARE NOT the qaurtiles of Trump's tweet length



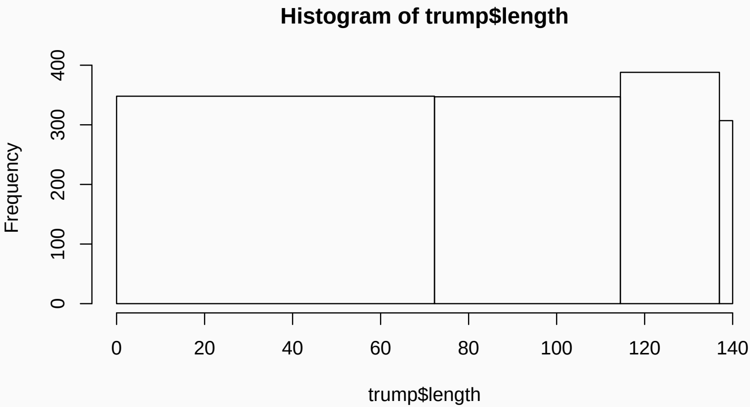
Example: Trump's Tweet Length

These ARE the qaurtiles of Trump's tweet length



Example: Trump's Tweet Length

These ARE the qaurtiles of Trump's tweet length



Boxplot

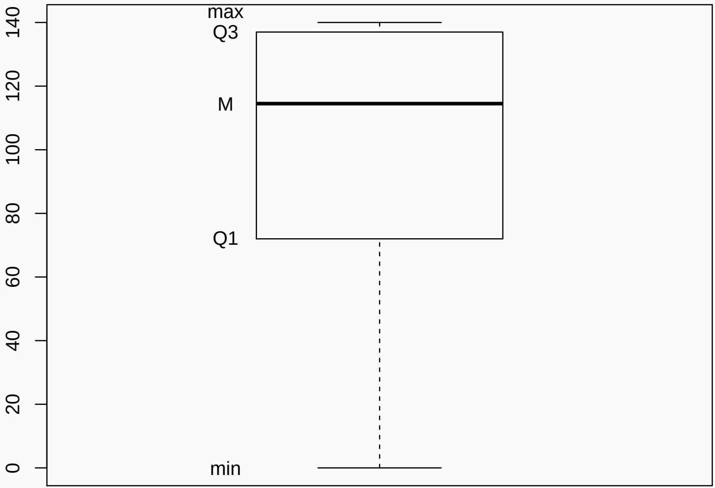
- It's *very* useful to plot these quantiles in what is called a boxplot.

boxplot

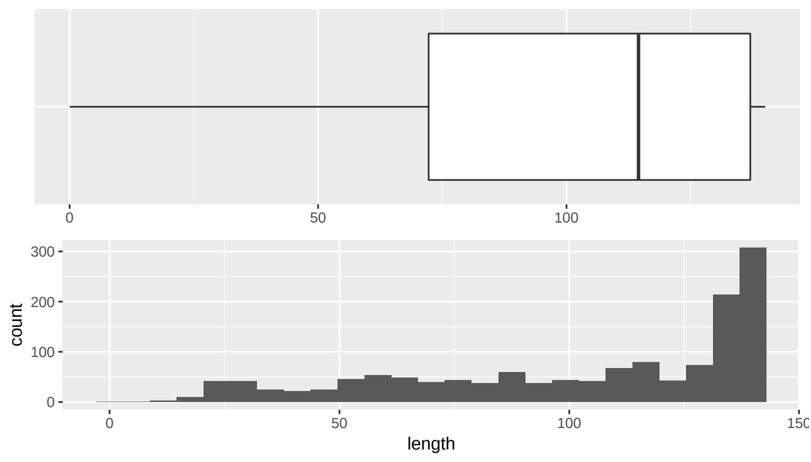
A boxplot is a graph of the five number summary. A central box spans the quartiles Q_1 and Q_3. A line in the box marks the median M. Lines (the "whiskers") extend from the box out to the smallest and largest observations.

Trump's Tweets

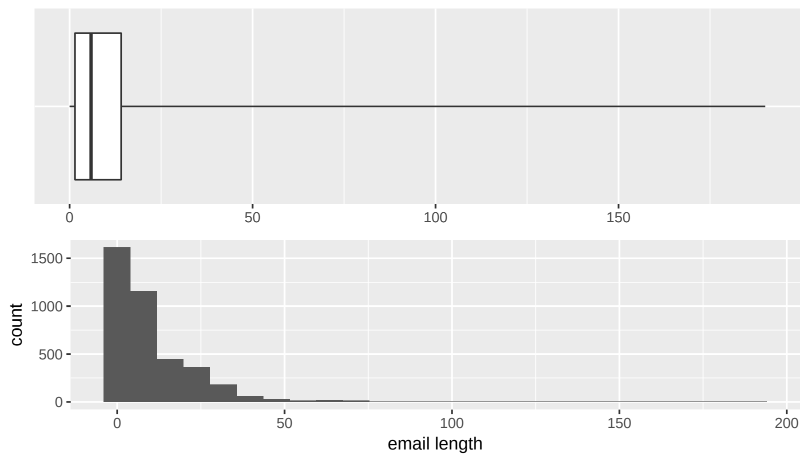
```r
boxplot(trump$length, range = 0)
```



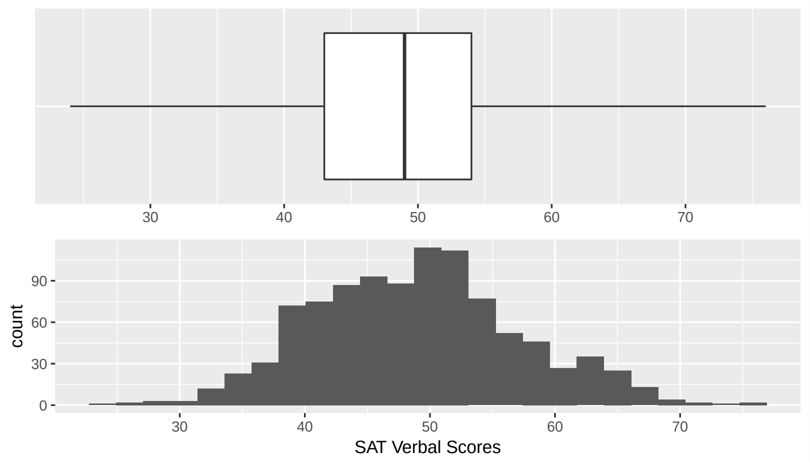
Boxplots tell us about skew: trump



Boxplots tell us about skew: email

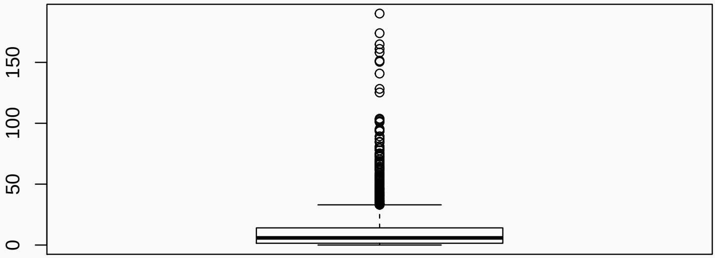


Boxplots tell us about skew: satGPA



Most boxplots you see will actually have more in them

```
boxplot(email$num_char)
```



What are those points?

IQR

To answer that, we first need to introduce the interquartile range (IQR).

IQR

The interquartile range IQR is the distance between the first and third quartiles,

$$IQR = Q_3 - Q_1,$$

and is a measure of spread.

Like MAD, IQR is a robust measure of spread

```
IQR(c(1, 2, 2, 3, 3))

[1] 1

IQR(c(1, 2, 2, 3, 10))

[1] 1

IQR(c(1, 2, 2, 3, 20))

[1] 1

IQR(c(1, 2, 2, 3, 100))

[1] 1
```

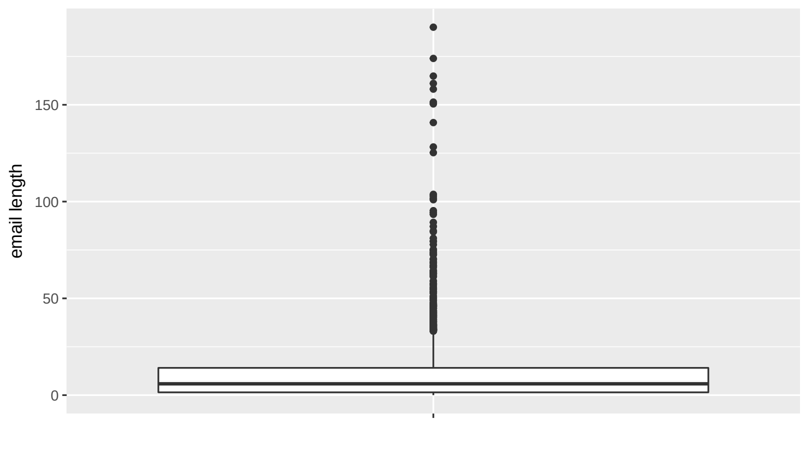
$1.5 \times IQR$ Rule

$1.5 \times IQR$ Rule

People will often call an observation a suspected outlier if it falls more than $1.5 \times IQR$ above the third quartile or below the first quartile.

- In most boxplots, the upper whisker extends to the largest observation within $1.5 \times IQR$ of Q_3.
- In most boxplots, the lower whisker extends to the smallest observation within $1.5 \times IQR$ of Q_1.
- Points outside of $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$ are labelled "suspsected outliers" and are plotted individually.

Sometimes, be suspicious of this rule



5.25 percent of all emails are “outliers”?

Recall Movie Scores Dataset

Observational units: Movies that sold tickets in 2015.

Variables:

- `rt` Rotten tomatoes score normalized to a 5 point scale.
- `meta` Metacritic score normalized to a 5 point scale.
- `imdb` IMDB score normalized to a 5 point scale.
- `fan` Fandango score.

Recall Movie Scores Dataset

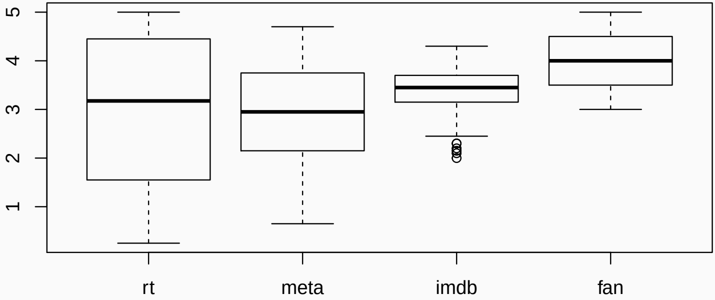
```
read_csv("../../data/movie.csv") %>%
  select(FILM, RT_norm, Metacritic_norm,
         IMDB_norm, Fandango_Stars) %>%
  transmute(film = FILM, rt = RT_norm, meta = Metacritic_norm,
            imdb = IMDB_norm, fan = Fandango_Stars) ->
  movie
head(movie)
```

```
# A tibble: 6 x 5
                            film    rt  meta  imdb   fan
                           <chr> <dbl> <dbl> <dbl> <dbl>
1 Avengers: Age of Ultron (2015)  3.70  3.30  3.90   5.0
2             Cinderella (2015)  4.25  3.35  3.55   5.0
3                Ant-Man (2015)  4.00  3.20  3.90   5.0
4        Do You Believe? (2015)  0.90  1.10  2.70   5.0
5  Hot Tub Time Machine 2 (2015)  0.70  1.45  2.55   3.5
6      The Water Diviner (2015)  3.15  2.50  3.60   4.5
```

How to compare these distributions?

Side-by-side boxplots!

```
boxplot(movie[, 2:5])
```

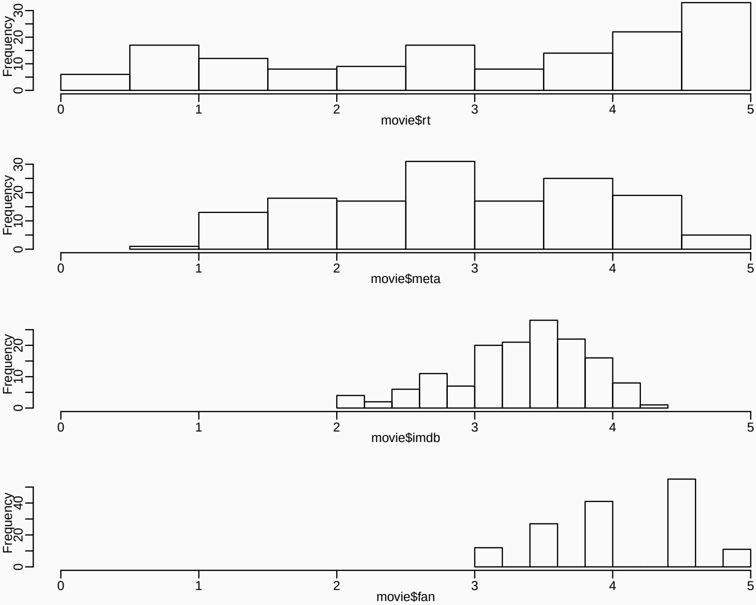


Another Option: stacked histograms

```r
old_parameters <- par(mfrow = c(4, 1))
hist(movie$rt, xlim = c(0, 5))
hist(movie$meta, xlim = c(0, 5))
hist(movie$imdb, xlim = c(0, 5))
hist(movie$fan, xlim = c(0, 5))
par(old_parameters)
```

IMPORTANT: Same *x*-limits for all plots when stacking vertically.

Another Option: stacked histograms


Frequency
movie$rt
Frequency
movie$meta
Frequency
movie$imdb
Frequency
movie$fan