# STAT UN1201 – Chapter 1

Prof. Joyce Robbins

# Waitlist

1. The waitlist moves in order as places open up.
2. Course materials are available here during change of program period: 1201.info
3. It is strongly advised to keep up with the material if you are trying to get in the class.

EVERYONE: Once you've made a decision not to take the class, please be considerate and drop it from your schedule.

# Very brief course outline

- Descriptive statistics (Chapter 1)
- Probability (Chapters 2 - 5)
- Inferential statistics (Chapters 6 - 9, 12 - 13)

# Descriptive statistics

- numerical summaries: mean, median, etc.
- graphical summaries

# Anscombe's Quartet

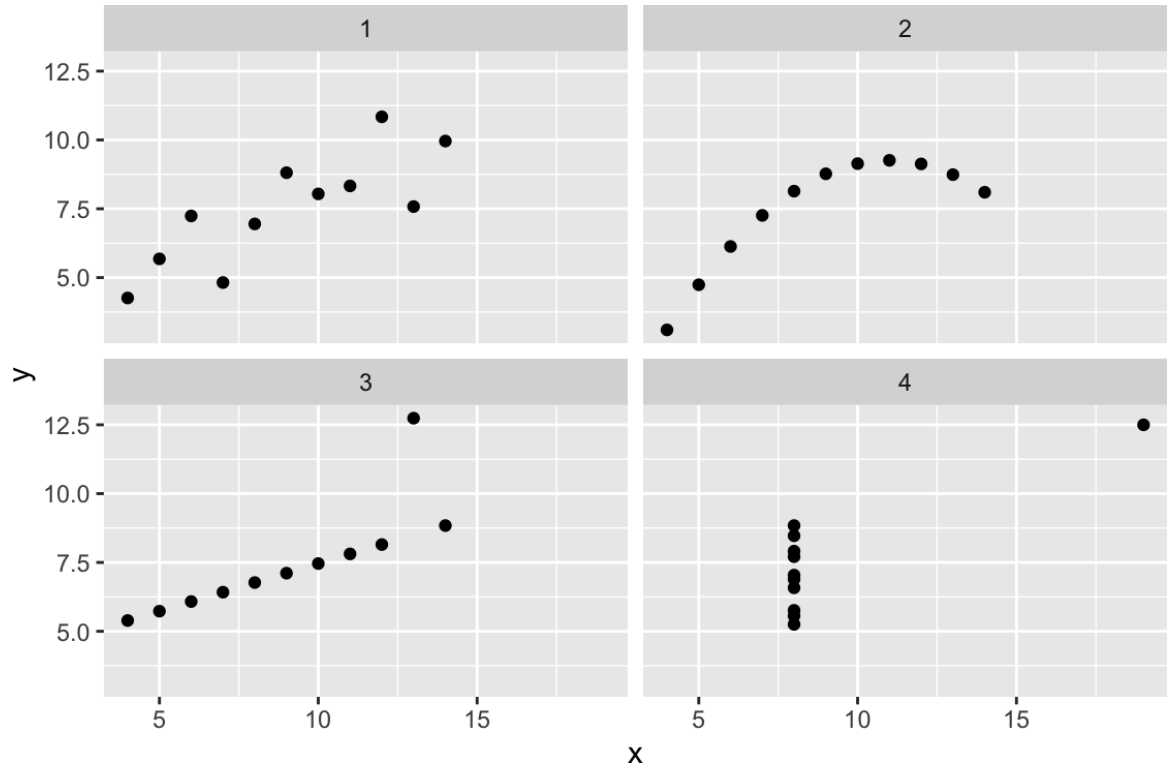
| x1 | x2 | x3 | x4 | y1 | y2 | y3 | y4 |
|---|---|---|---|---|---|---|---|
| 10 | 10 | 10 | 8 | 8.04 | 9.14 | 7.46 | 6.58 |
| 8 | 8 | 8 | 8 | 6.95 | 8.14 | 6.77 | 5.76 |
| 13 | 13 | 13 | 8 | 7.58 | 8.74 | 12.74 | 7.71 |
| 9 | 9 | 9 | 8 | 8.81 | 8.77 | 7.11 | 8.84 |
| 11 | 11 | 11 | 8 | 8.33 | 9.26 | 7.81 | 8.47 |
| 14 | 14 | 14 | 8 | 9.96 | 8.10 | 8.84 | 7.04 |
| 6 | 6 | 6 | 8 | 7.24 | 6.13 | 6.08 | 5.25 |
| 4 | 4 | 4 | 19 | 4.26 | 3.10 | 5.39 | 12.50 |
| 12 | 12 | 12 | 8 | 10.84 | 9.13 | 8.15 | 5.56 |
| 7 | 7 | 7 | 8 | 4.82 | 7.26 | 6.42 | 7.91 |
| 5 | 5 | 5 | 8 | 5.68 | 4.74 | 5.73 | 6.89 |

# Anscombe's Quartet, numerical summaries

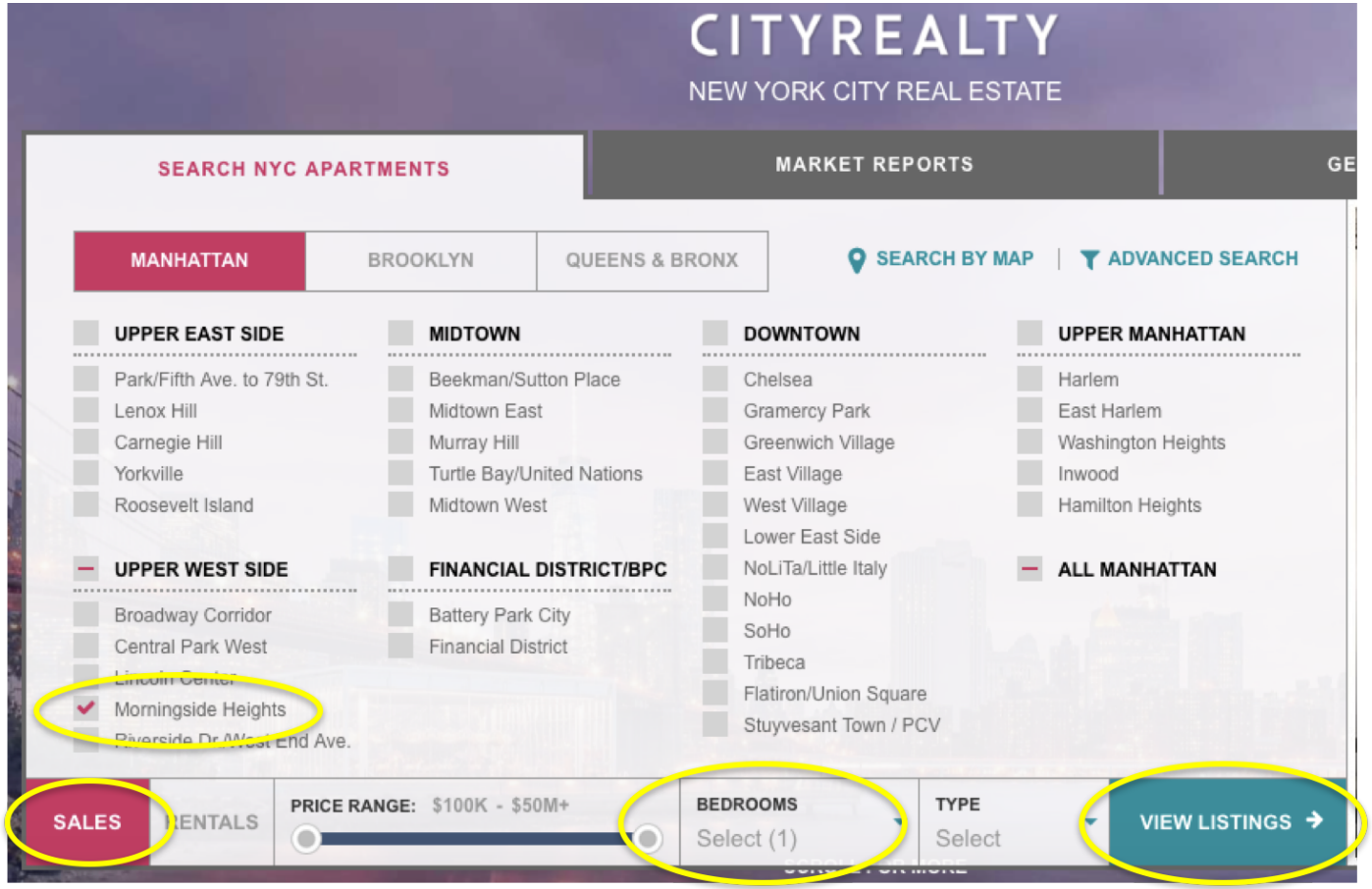
For each of the four data sets:

- Number of observations ($n$) = 11
- Mean of the $x$'s ($\bar{x}$) = 9.0
- Mean of the $y$'s ($\bar{y}$) = 7.6
- Equation of the regression line: $y = 3 + 0.5x$

# Anscombe’s Quartet, graphical analysis



# Collecting data



# Collecting data

HIDE ALL PHOTOS SHOW ALL PHOTOS

| ADDRESS | TYPE | NEIGHBORHOOD | BEDS / BATH / SIZE | PRICE | PRICE/FT² | LISTED |
|---|---|---|---|---|---|---|
| 395 Riverside Drive, #4E | CO-OP | Morningside Heights | 1 bd / 1 ba | $799K | - | Feb 22, 2016 |
| 390 Riverside Drive, #4G | CO-OP | Morningside Heights | 1 bd / 1 ba / 800 ft² | $745K | $931 | Jul 19, 2016 |
| 380 Riverside Drive, #8L | CO-OP | Morningside Heights | 1 bd / 1 ba / 700 ft² | $725K | $1,036 | Sep 06, 2016 |
| 535 West 110th Street, #6B | CO-OP | Morningside Heights | 1 bd / 1 ba / 750 ft² | $725K | $967 | Sep 07, 2016 |
| 535 West 110th Street, #6A | CO-OP | Morningside Heights | 1 bd / 1 ba / 750 ft² | $699K ( -4% ) | $933 | Apr 26, 2016 |
| 395 Riverside Drive, #2D | CO-OP | Morningside Heights | 1 bd / 1 ba / 900 ft² | $699K | $777 | Mar 03, 2016 |
| 528 West 111th Street, #24 | CO-OP | Morningside Heights | 1 bd / 1 ba | $675K ( -6.9% ) | - | Apr 15, 2015 |
| 545 West 111th Street, #9N | CO-OP | Morningside Heights | 1 bd / 1 ba / 650 ft² | $665K ( -9.5% ) | $1,023 | Mar 30, 2016 |
| 380 Riverside Drive, #6B | CO-OP | Morningside Heights | 1 bd / 1 ba | $599K | - | Oct 07, 2015 |
| 80 La Salle Street, #20C | CO-OP | Morningside Heights | 1 bd / 1 ba | $545K | - | May 19, 2016 |
| 609 West 114th Street, #36 | CO-OP | Morningside Heights | 1 bd / 1 ba / 600 ft² | $535K | $892 | Jun 22, 2016 |
| 80 La Salle Street, #21B | CO-OP | Morningside Heights | 1 bd / 1 ba / 725 ft² | $535K | $738 | May 04, 2016 |
| 80 La Salle Street, #14D | CO-OP | Morningside Heights | 1 bd / 1 ba | $529K | - | Apr 14, 2016 |
| 175 Claremont Avenue, #3 | CO-OP | Morningside Heights | 1 bd / 1 ba | $499K ( -9.3% ) | - | Mar 30, 2016 |
| 3117 Broadway, #56 | CO-OP | Morningside Heights | 1 bd / 1 ba / 500 ft² | $450K | $900 | Jun 28, 2016 |

TURED NEW DEVELOPMENTS



| | | | | | | |
|---|---|---|---|---|---|---|
| 510 West 123rd Street, #38 | CO-OP | Morningside Heights | 1 bd / 1 ba / 700 ft² | $450K | $643 | Jul 02, 2016 |
| 3115 Broadway, #43 | CO-OP | Morningside Heights | 1 bd / 1 ba / 650 ft² | $425K | $654 | Apr 18, 2016 |
| 3115 Broadway, #42 | CO-OP | Morningside Heights | 1 bd / 1 ba / 650 ft² | $379K ( -5% ) | $583 | Mar 28, 2016 |

# Cleaning / organizing data

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | | ADDRESS | TYPE | NEIGHBORH( | BEDS / BATH | PRICE | PRICE/FT2 | LISTED |
| 2 | + | | | | | | | |
| 3 | 3115 Broadv | CO-OP | | | | | | |
| 4 | Morningside | 1 bd / 1 ba / | $379K | $583 | 28-Mar-16 | | | |
| 5 | + | | | | | | | |
| 6 | 3115 Broadv | CO-OP | | | | | | |
| 7 | Morningside | 1 bd / 1 ba / | $425K | $654 | 18-Apr-16 | | | |
| 8 | + | | | | | | | |
| 9 | 510 West 12 | CO-OP | | | | | | |
| 10 | Morningside | 1 bd / 1 ba / | $450K | $643 | 2-Jul-16 | | | |
| 11 | + | | | | | | | |
| 12 | 3117 Broadv | CO-OP | | | | | | |
| 13 | Morningside | 1 bd / 1 ba / | $450K | $900 | 28-Jun-16 | | | |
| 14 | + | | | | | | | |
| 15 | 175 Claremo | CO-OP | | | | | | |
| 16 | Morningside | 1 bd / 1 ba | $499K | - | 30-Mar-16 | | | |
| 17 | + | | | | | | | |
| 18 | 80 La Salle St | CO-OP | | | | | | |
| 19 | Morningside | 1 bd / 1 ba | $529K | - | 14-Apr-16 | | | |

# Cleaning / organizing data

| | A | B | C | D |
|---|---|---|---|---|
| 1 | **Address** | | **Price** | **IntPrice** |
| 2 | 3115 Broadway, #42 | 1 bd / 1 ba / 650 ft2 | $379K | 379 |
| 3 | 3115 Broadway, #43 | 1 bd / 1 ba / 650 ft2 | $425K | 425 |
| 4 | 3117 Broadway, #56 | 1 bd / 1 ba / 500 ft2 | $450K | 450 |
| 5 | 510 West 123rd Street, #38 | 1 bd / 1 ba / 700 ft2 | $450K | 450 |
| 6 | 175 Claremont Avenue, #3 | 1 bd / 1 ba | $499K | 499 |
| 7 | 80 La Salle Street, #14D | 1 bd / 1 ba | $529K | 529 |
| 8 | 609 West 114th Street, #36 | 1 bd / 1 ba / 600 ft2 | $535K | 535 |
| 9 | 80 La Salle Street, #21B | 1 bd / 1 ba / 725 ft2 | $535K | 535 |
| 10 | 80 La Salle Street, #20C | 1 bd / 1 ba | $545K | 545 |
| 11 | 380 Riverside Drive, #6B | 1 bd / 1 ba | $599K | 599 |
| 12 | 545 West 111th Street, #9N | 1 bd / 1 ba / 650 ft2 | $665K | 665 |
| 13 | 528 West 111th Street, #24 | 1 bd / 1 ba | $675K | 675 |
| 14 | 535 West 110th Street, #6A | 1 bd / 1 ba / 750 ft2 | $699K | 699 |
| 15 | 395 Riverside Drive, #2D | 1 bd / 1 ba / 900 ft2 | $699K | 699 |
| 16 | 380 Riverside Drive, #8L | 1 bd / 1 ba / 700 ft2 | $725K | 725 |
| 17 | 535 West 110th Street, #6B | 1 bd / 1 ba / 750 ft2 | $725K | 725 |
| 18 | 390 Riverside Drive, #4G | 1 bd / 1 ba / 800 ft2 | $745K | 745 |
| 19 | 395 Riverside Drive, #4E | 1 bd / 1 ba | $799K | 799 |
| 20 | | | | |
| 21 | | | | |

# Types of data

- Quantitative
  - *Discrete (finite or can be listed in an infinite sequence)*
  - *Continuous (entire continuum of possible values)*
- Qualitative (categorical)

# Visualizing one-dimensional continuous data

- stem-and-leaf plot
- histogram
  - *frequency*
  - *relative frequency*
  - *density*
- boxplot

# Stem-and-leaf plot

Prices:

```
##  [1] 379 425 450 450 499 529 535 535 545 599 665
## [12] 675 699 699 725 725 745 799
```

Prices, rounded to nearest 10,000:

```
##  [1] 380 430 450 450 500 530 540 540 550 600 670
## [12] 680 700 700 730 730 750 800
```

# Stem-and-leaf plot

Prices:

```
##  [1] 379 425 450 450 499 529 535 535 545 599 665
## [12] 675 699 699 725 725 745 799
```

Prices, rounded to nearest 10,000:

```
##  [1] 380 430 450 450 500 530 540 540 550 600 670
## [12] 680 700 700 730 730 750 800
```

```
##
##   The decimal point is 2 digit(s) to the right of the |
##
##   3 | 8
##   4 | 355
##   5 | 03445
##   6 | 078
##   7 | 00335
##   8 | 0
```

# Stem and leaf plot

Total Fertility Rate (average births per woman)

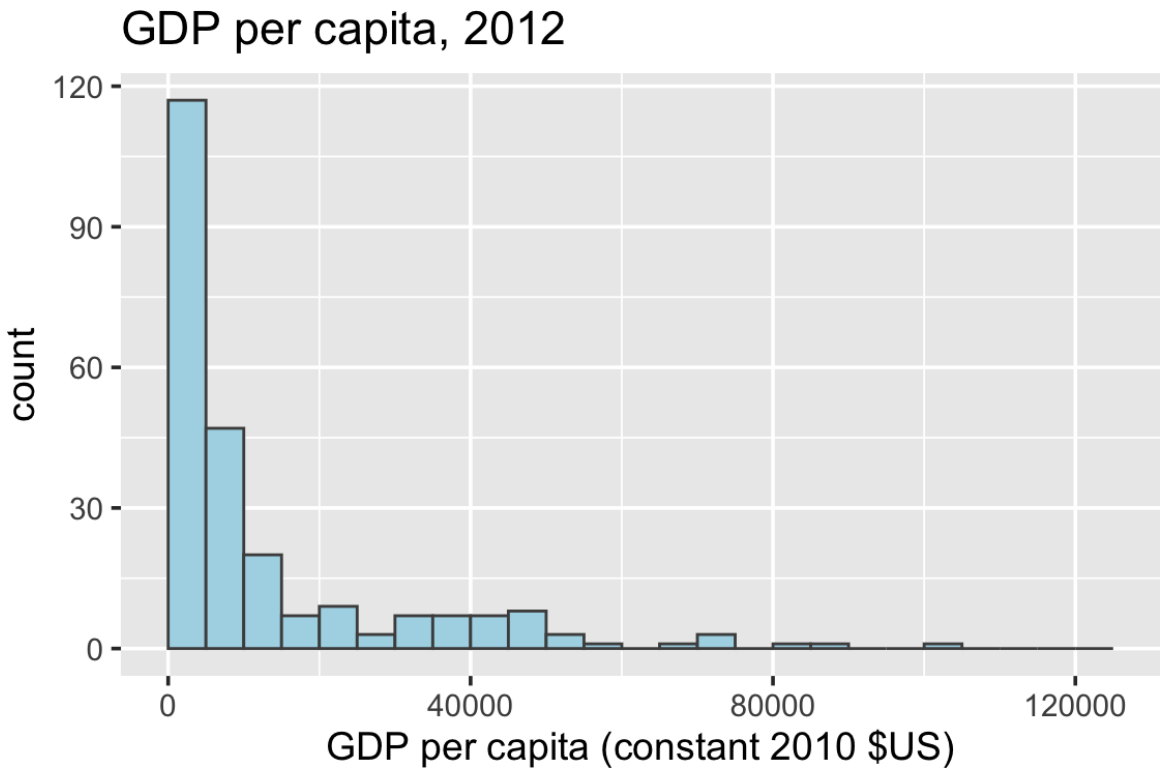
```
##
##   The decimal point is at the |
##
##   1 | 333333333444444555555555555566666666777777888888888899999999999
##   2 | 000000001111222223333333444444455555555666666667999 9
##   3 | 01112233333344567889
##   4 | 0011112245555667779999
##   5 | 0001222333355788
##   6 | 012344
##   7 | 6
```

# Stem and leaf plot, change the length (scale)

Total Fertility Rate

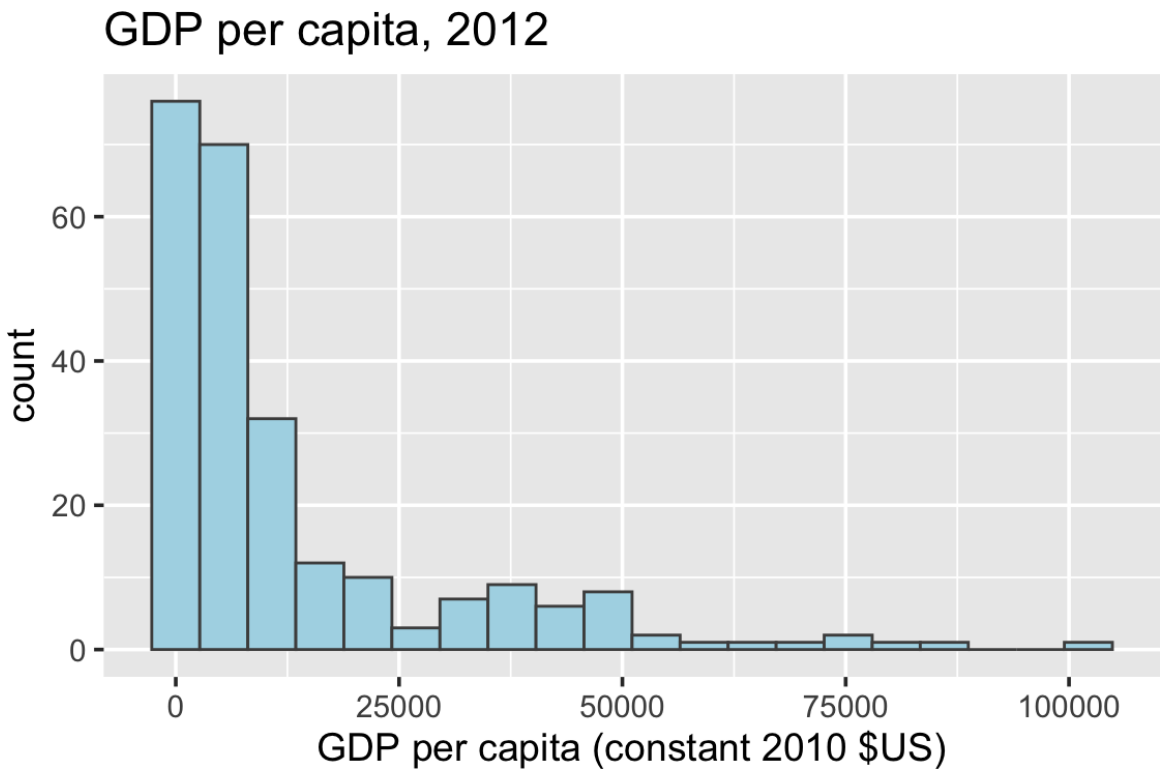
```
##
##   The decimal point is at the |
##
##   1 | 333333333444444
##   1 | 55555555555556666666677777788888888889999999999
##   2 | 00000000111122222333333344444444
##   2 | 555555556666666679999
##   3 | 01112233333344
##   3 | 567889
##   4 | 001111224
##   4 | 5555667779999
##   5 | 00012223333
##   5 | 55788
##   6 | 012344
##   6 |
##   7 |
##   7 | 6
```

# Frequency histogram

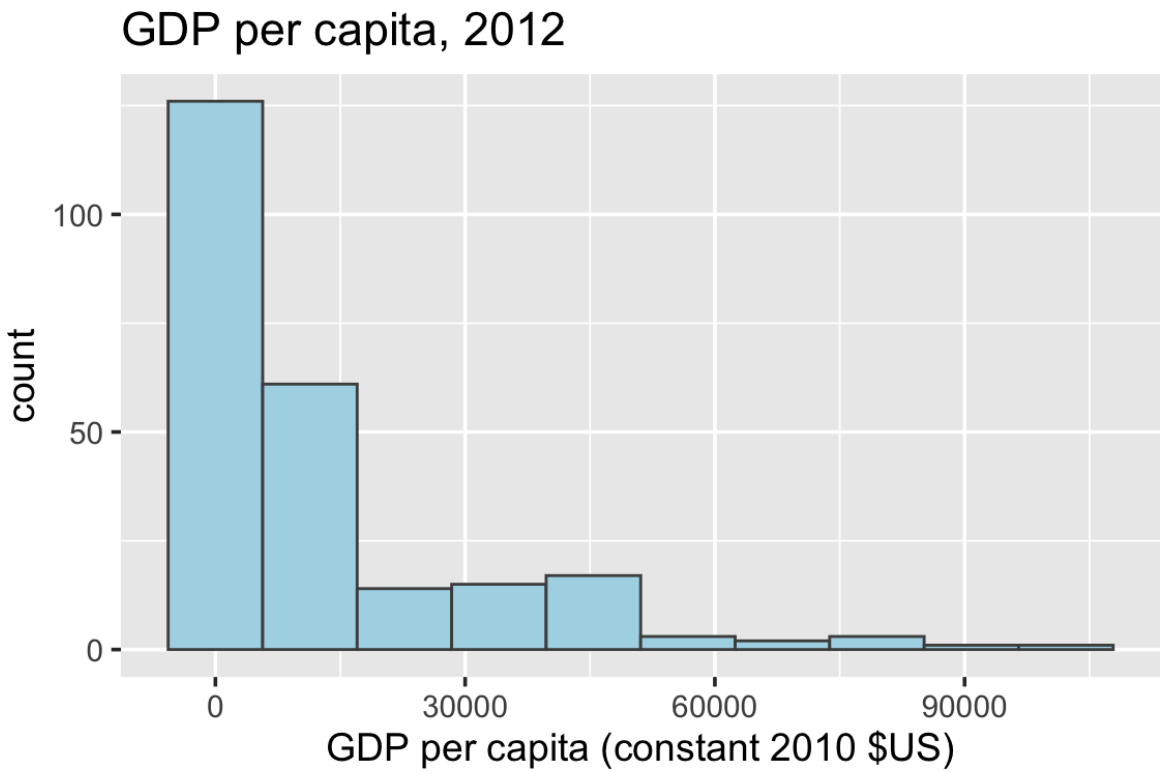



Source: World Bank

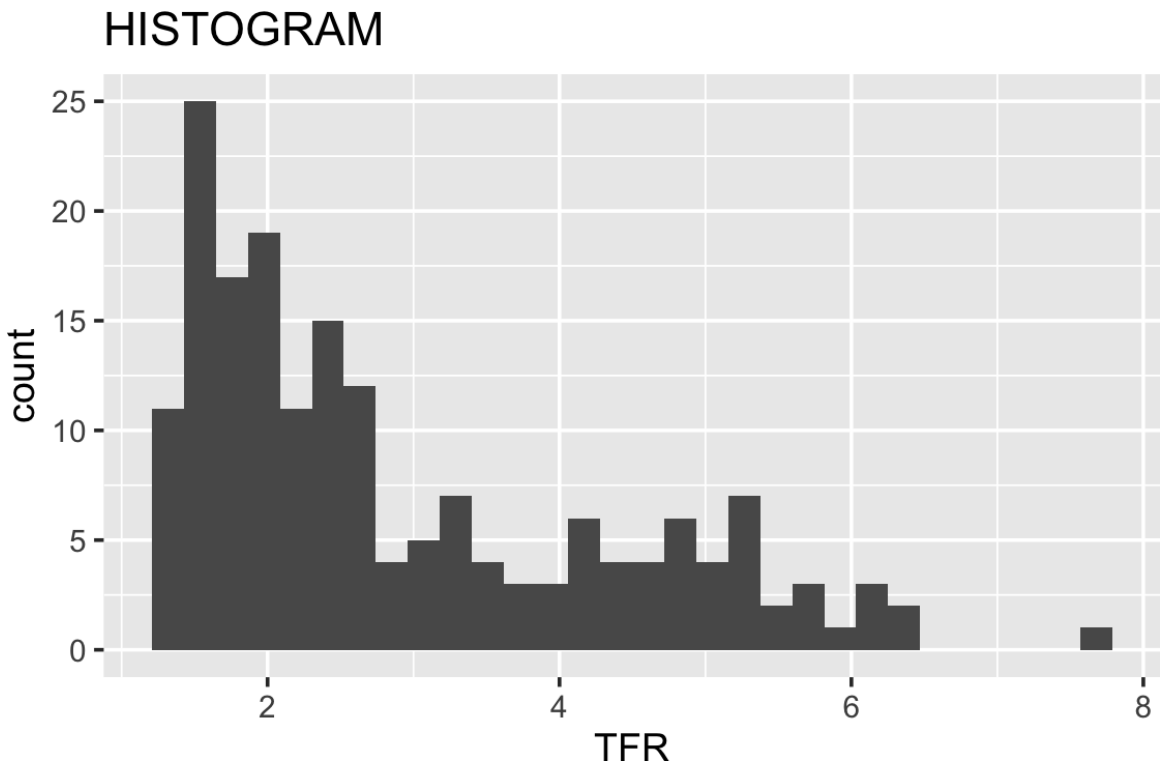
# Change number of bins to 20



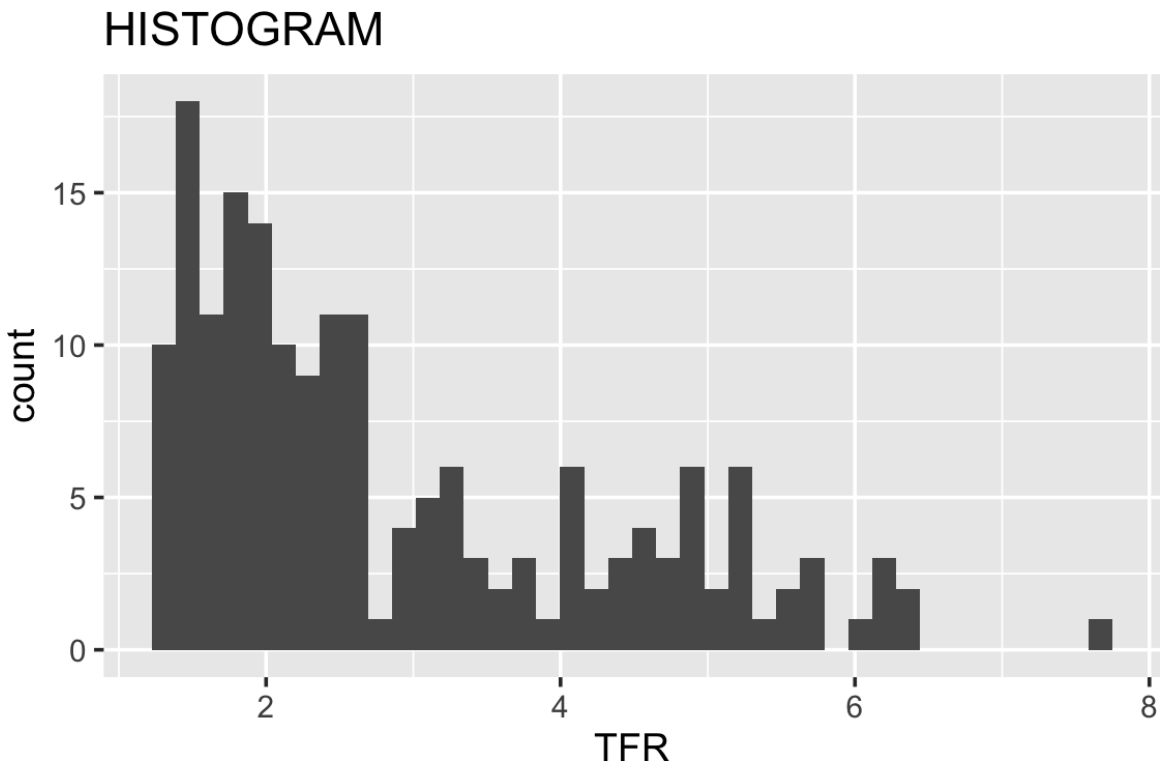
# Change number of bins to 10



# Total fertility rate



# Change the number of bins to 40



# Shapes of histograms

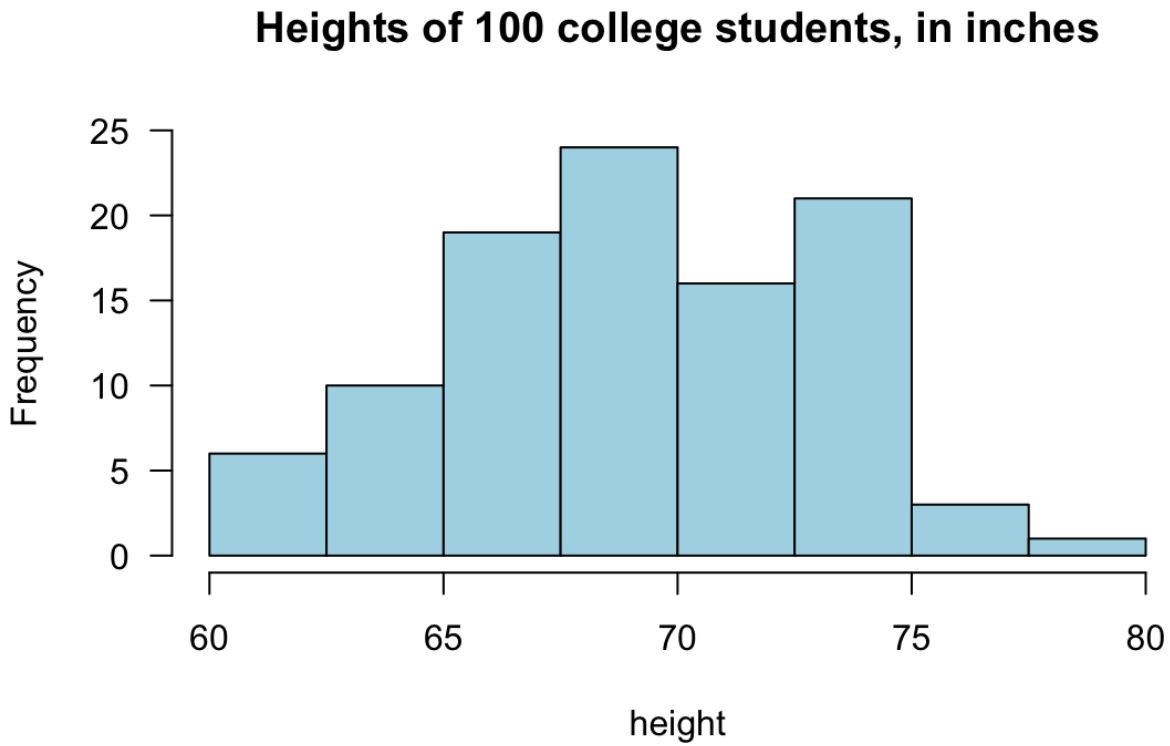
- Unimodal
- Bimodal
- Skewed left
- Skewed right

# Discrete data

```
## [1] "Heights of 100 college students, in inches"
```

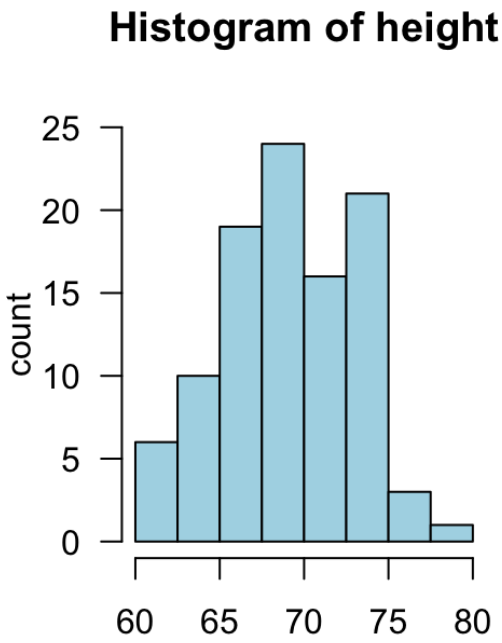
```
##   [1] 60 60 61 61 61 62 63 63 64 64 64 64 65 65 65
##  [16] 65 66 66 66 66 66 66 67 67 67 67 67 67 67 67
##  [31] 67 67 67 67 67 68 68 68 68 68 68 68 69 69 69
##  [46] 69 69 69 69 69 69 70 70 70 70 70 70 70 70 71
##  [61] 71 71 71 71 71 72 72 72 72 72 72 72 72 72 72
##  [76] 73 73 73 74 74 74 74 74 74 74 74 74 74 74 74
##  [91] 74 75 75 75 75 75 76 76 77 79
```

# Discrete data

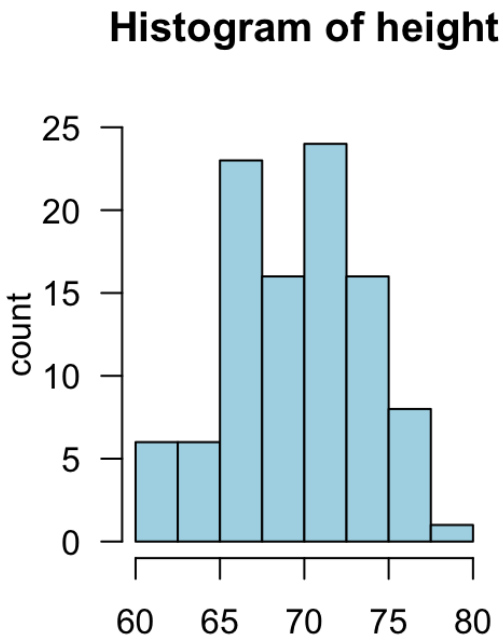



```
##   [1] 60 60 61 61 61 62 63 63 64 64 64 64 65 65 65
##  [16] 65 66 66 66 66 66 66 67 67 67 67 67 67 67 67
##  [31] 67 67 67 67 67 68 68 68 68 68 68 68 69 69 69
##  [46] 69 69 69 69 69 69 70 70 70 70 70 70 70 70 71
##  [61] 71 71 71 71 71 72 72 72 72 72 72 72 72 72 72
##  [76] 73 73 73 74 74 74 74 74 74 74 74 74 74 74 74
##  [91] 74 75 75 75 75 75 76 76 77 79
```

# Discrete data histogram




RIGHT CLOSED, LEFT OPEN




RIGHT OPEN, LEFT CLOSED

```
##   [1] 60 60 61 61 61 62 63 63 64 64 64 64 65 65 65
##  [16] 65 66 66 66 66 66 66 67 67 67 67 67 67 67 67
##  [31] 67 67 67 67 67 68 68 68 68 68 68 68 69 69 69
##  [46] 69 69 69 69 69 69 70 70 70 70 70 70 70 70 71
##  [61] 71 71 71 71 71 72 72 72 72 72 72 72 72 72 72
##  [76] 73 73 73 74 74 74 74 74 74 74 74 74 74 74 74
##  [91] 74 75 75 75 75 75 76 76 77 79
```

# EXERCISE

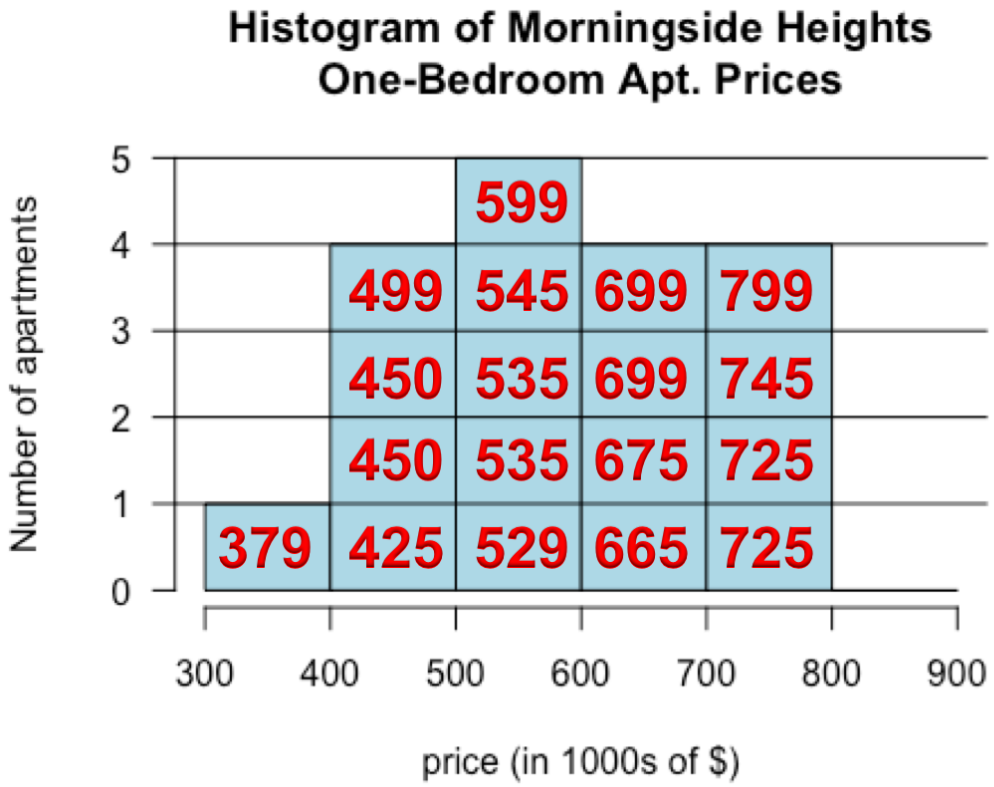
Draw a histogram of the asking prices for one-bedroom apartments in Morningside Heights (prices in thousands of $)
Data source: cityrealty.com, 9/13/2016

379, 425, 450, 450, 499, 529, 535, 535, 545,
599, 665, 675, 699, 699, 725, 725, 745, 799

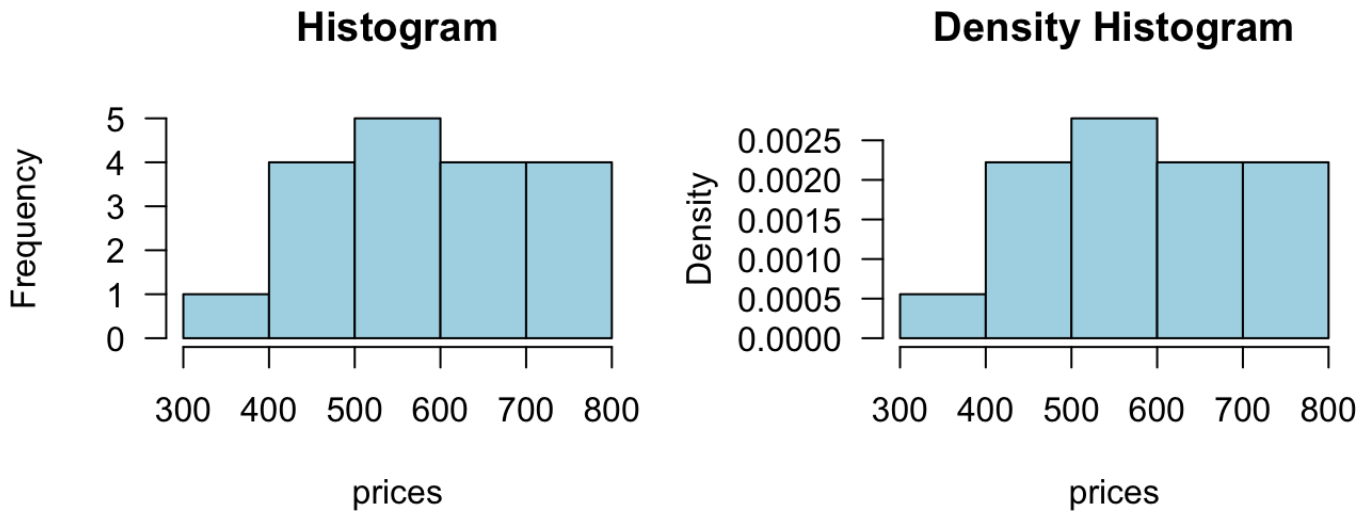
# Drawing histograms

Asking prices for one-bedroom apartments in Morningside Heights, in thousands
Data source: cityrealty.com, 9/13/2016

379, 425, 450, 450, 499, 529, 535, 535, 545,
599, 665, 675, 699, 699, 725, 725, 745, 799

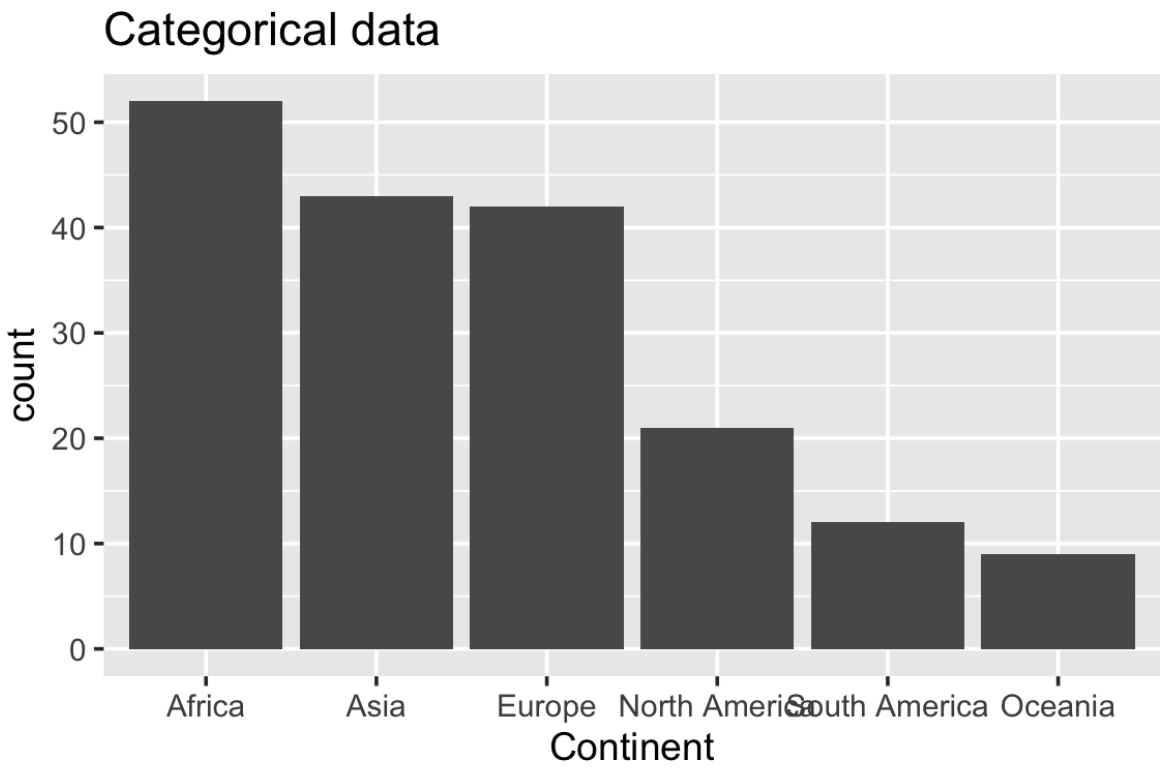

Histogram of Morningside Heights
One-Bedroom Apt. Prices
Number of apartments
0
1
2
3
4
5
379
425
529
665
725
450
535
675
725
450
535
699
745
499
545
699
799
599
300
400
500
600
700
800
900
price (in 1000s of $)

# Density histogram

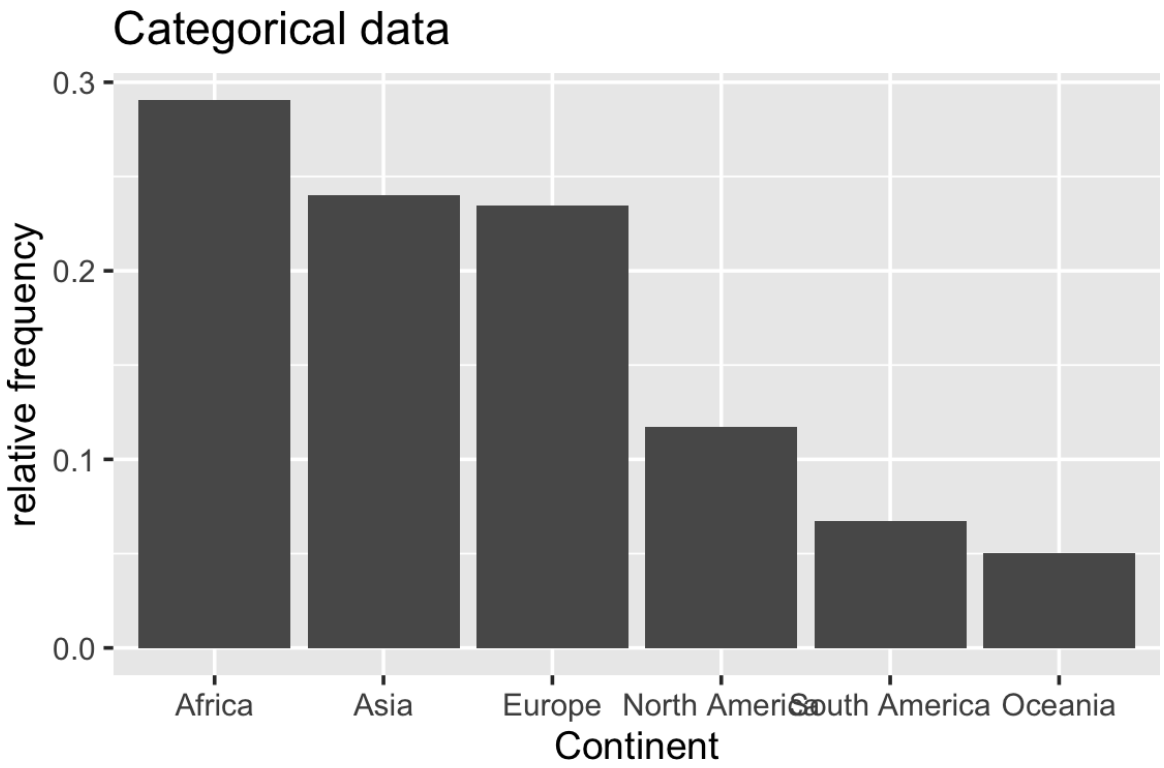



| Class | Frequency | Rel. Frequency | Density |
|---|---|---|---|
| (300, 400] | 1 | .056 | .00056 |
| (400, 500] | 4 | .222 | .00222 |
| (500, 600] | 5 | .278 | .00278 |
| (600, 700] | 4 | .222 | .00222 |
| (700, 800] | 4 | .222 | .00222 |

# Frequency bar chart



# Relative frequency bar chart



# Five number summary

1. min
2. lower fourth
3. median
4. upper fourth
5. max

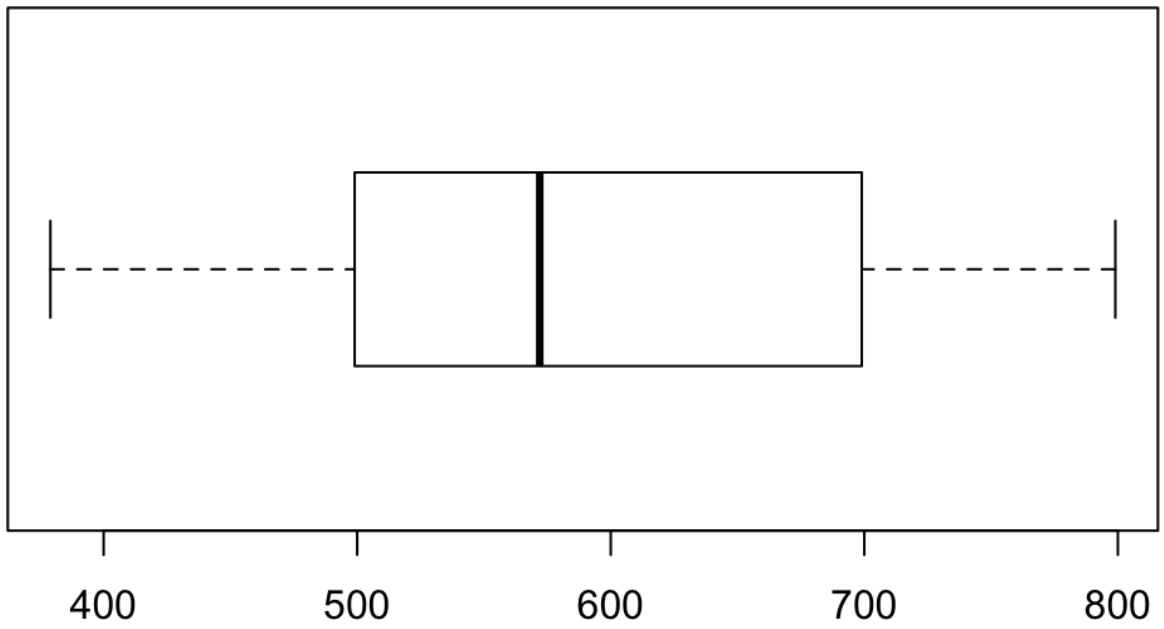
```
fivenum(prices)
```

```
## [1] 379 499 572 699 799
```

# Boxplot (no outliers)

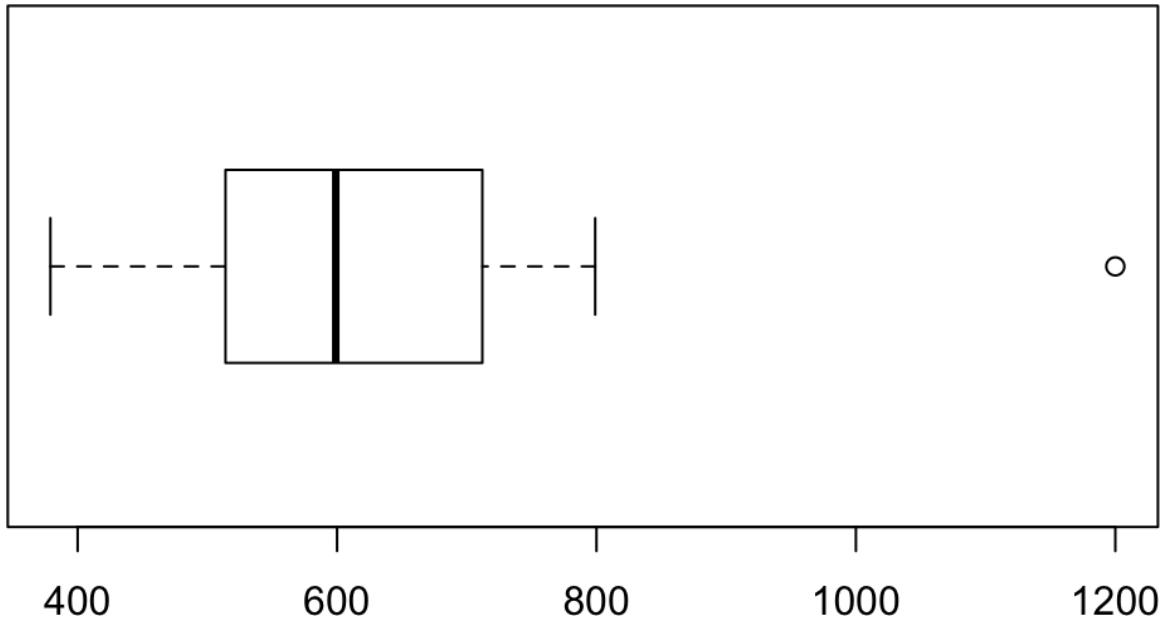
379, 425, 450, 450, 499, 529, 535, 535, 545,
599, 665, 675, 699, 699, 725, 725, 745, 799

```
## [1] 379 499 572 699 799
```



# Boxplot with outlier

```
##  [1]  379  425  450  450  499  529  535  535  545
## [10]  599  665  675  699  699  725  725  745  799
## [19] 1200
```




Box shows lower fourth, median, upper fourth

Whiskers show data within 1.5 times the *fourth spread* ($f_s$) of the nearest fourth

$$f_s = 712 - 514 = 198$$

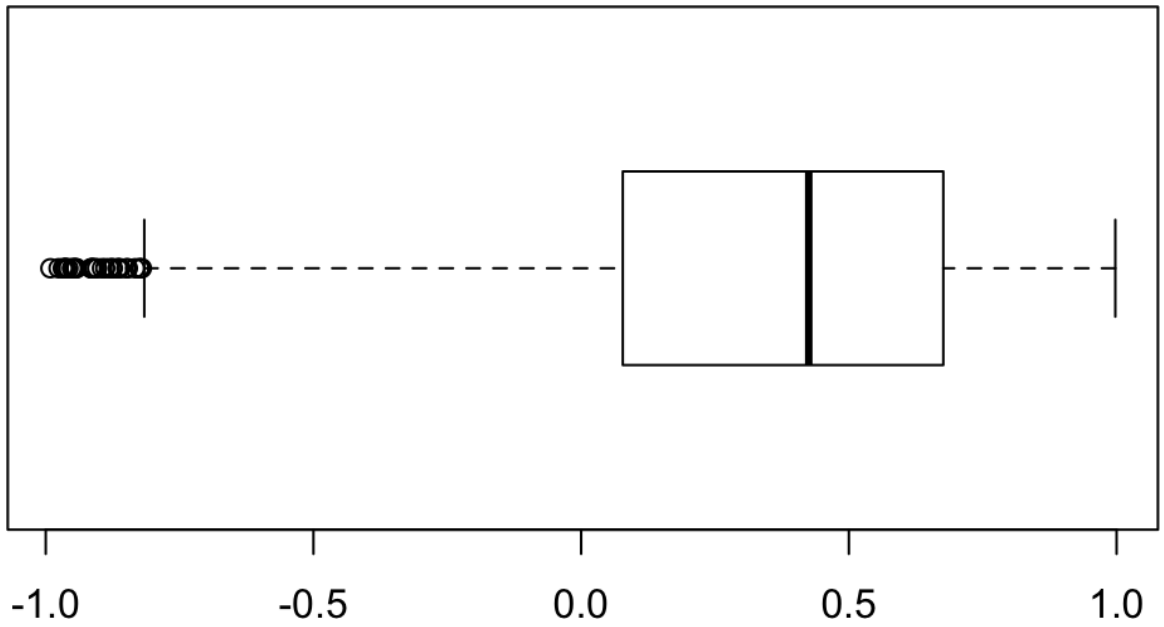
$$1.5f_s = 297$$

lower outlier boundary: lower fourth - $1.5f_s$ = 514 - 297 = 217

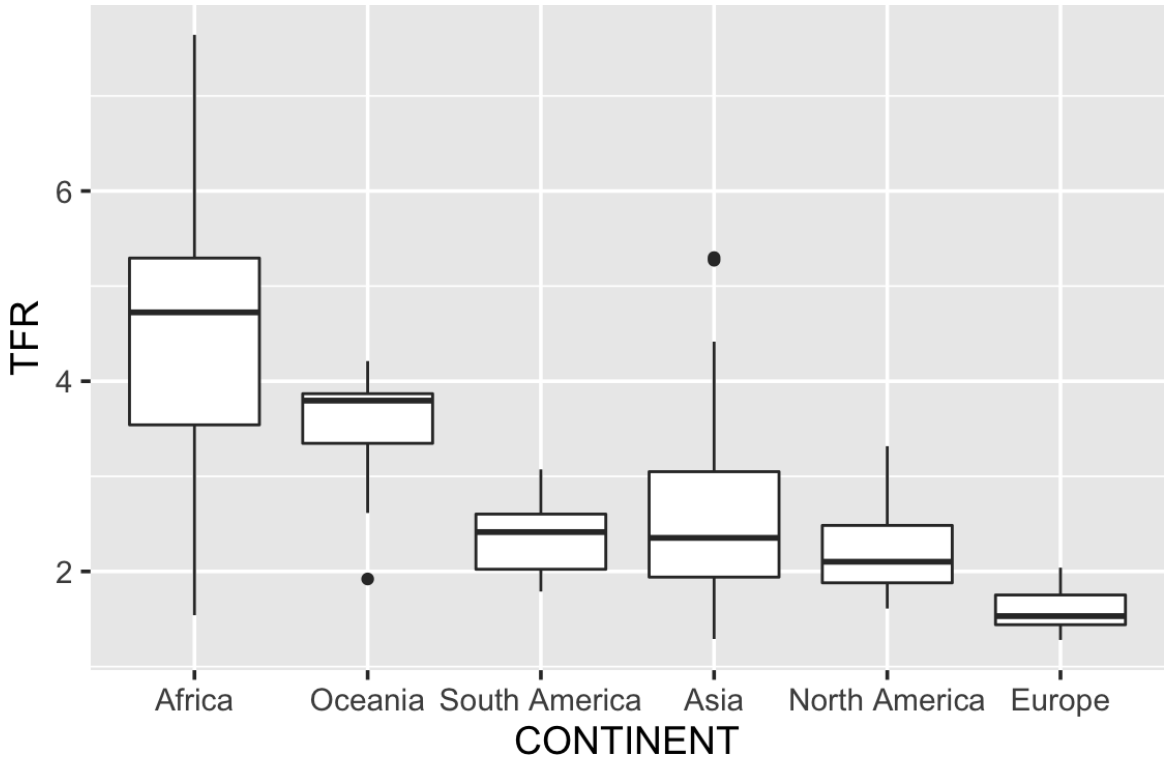
upper outlier boundary: upper fourth + $1.5f_s$ = 712 + 297 = 1009

Add outliers and then draw whiskers to the lowest and highest data values that are *not* outliers.

# Boxplot with outliers



# Multiple box plots



# EXERCISE

(based on #72, p. 49)

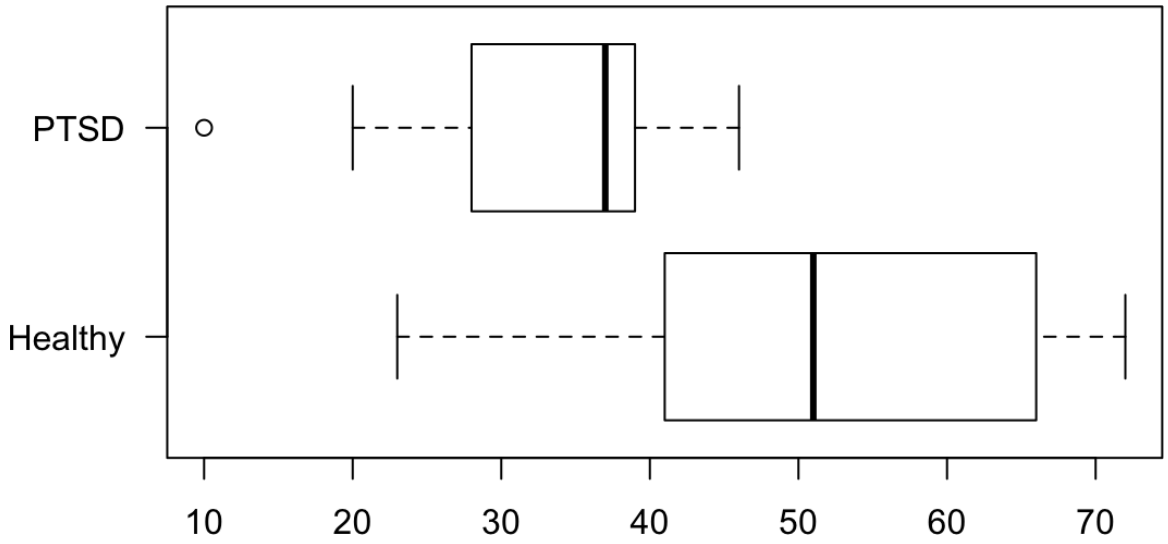
Data on a receptor binding measure:

PTSD: 10, 20, 25, 28, 31, 35, 37, 38, 38, 39, 39, 42, 46

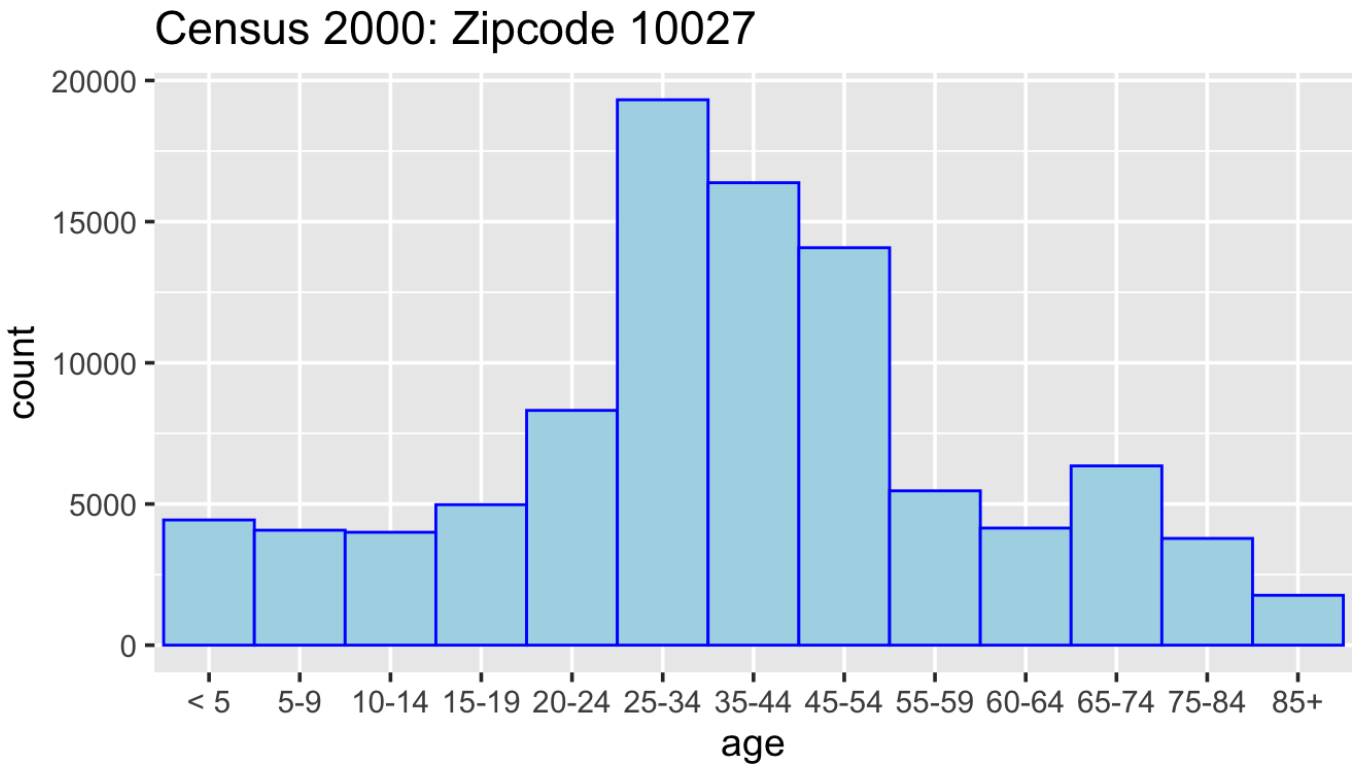
Healthy: 23, 39, 40, 41, 43, 47, 51, 58, 63, 66, 67, 69, 72

Draw a comparative boxplot.

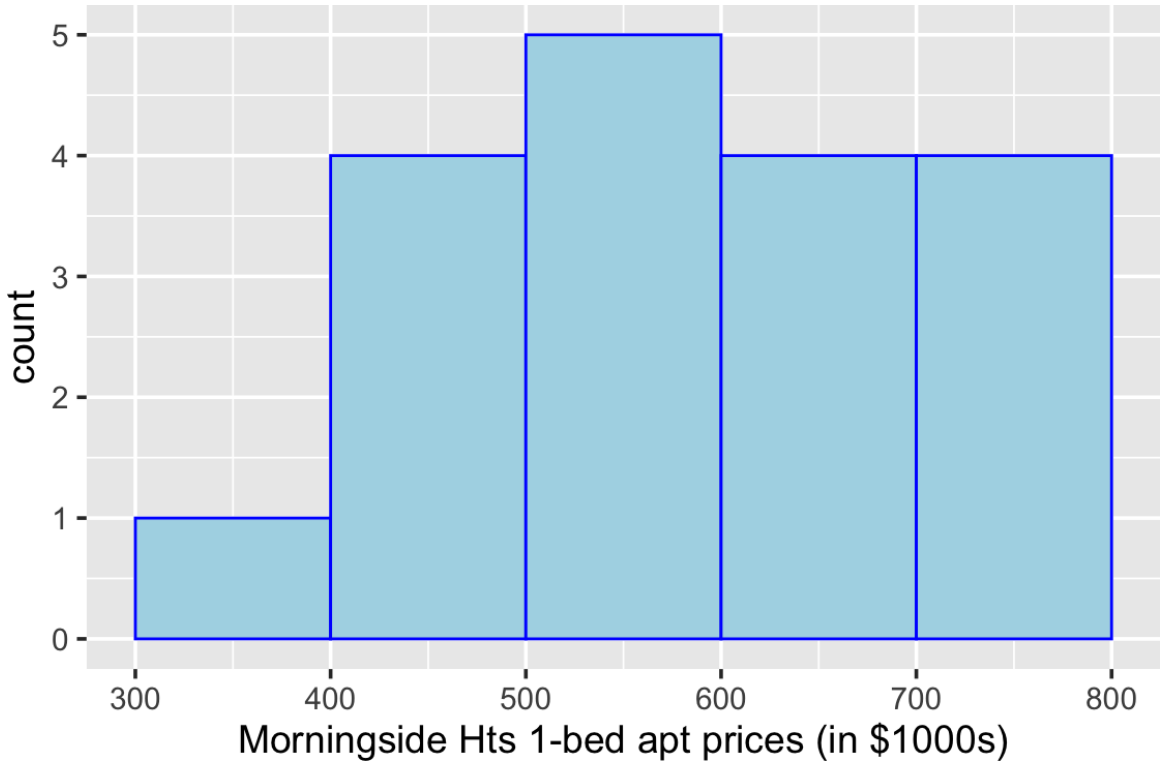
# Comparative boxplot



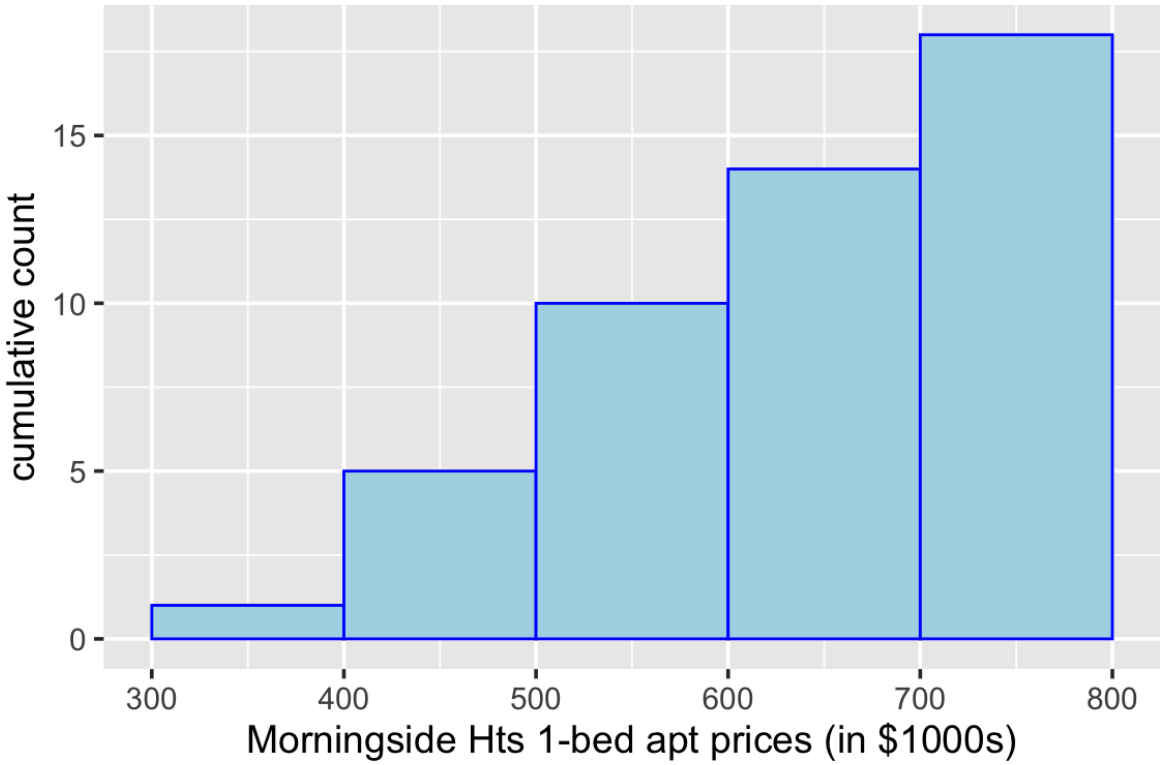
# Histogram: what's wrong?



# Frequency histogram



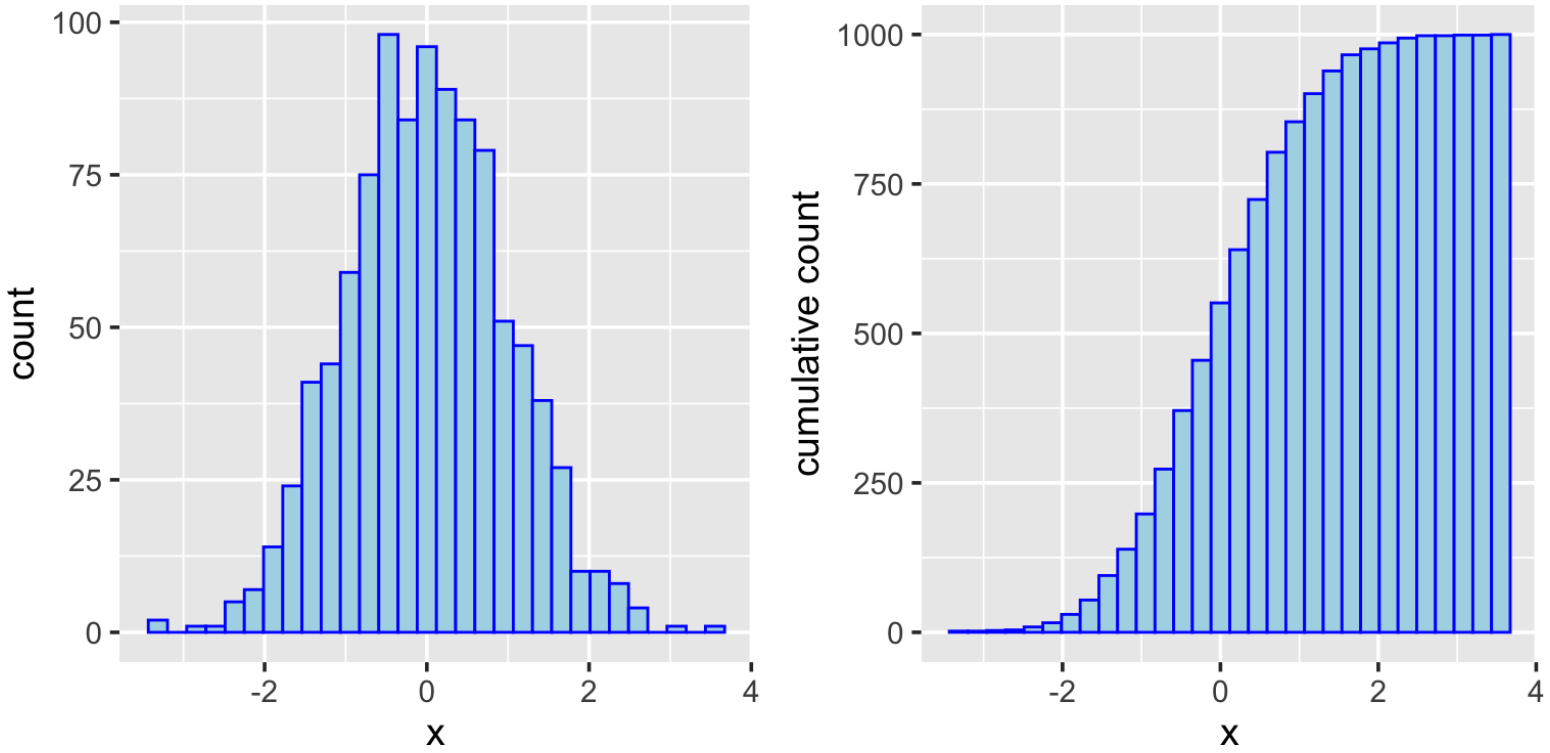
# Cumulative frequency histogram



# Cumulative frequency histogram

| Class | Freq | CumulativeFreq |
|---|---|---|
| 300-400 | 1 | 1 |
| 400-500 | 4 | 5 |
| 500-600 | 5 | 10 |
| 600-700 | 4 | 14 |
| 700-800 | 4 | 18 |

# Cumulative frequency histogram



# EXERCISE

(based on #17, p. 26)
Construction industry data:

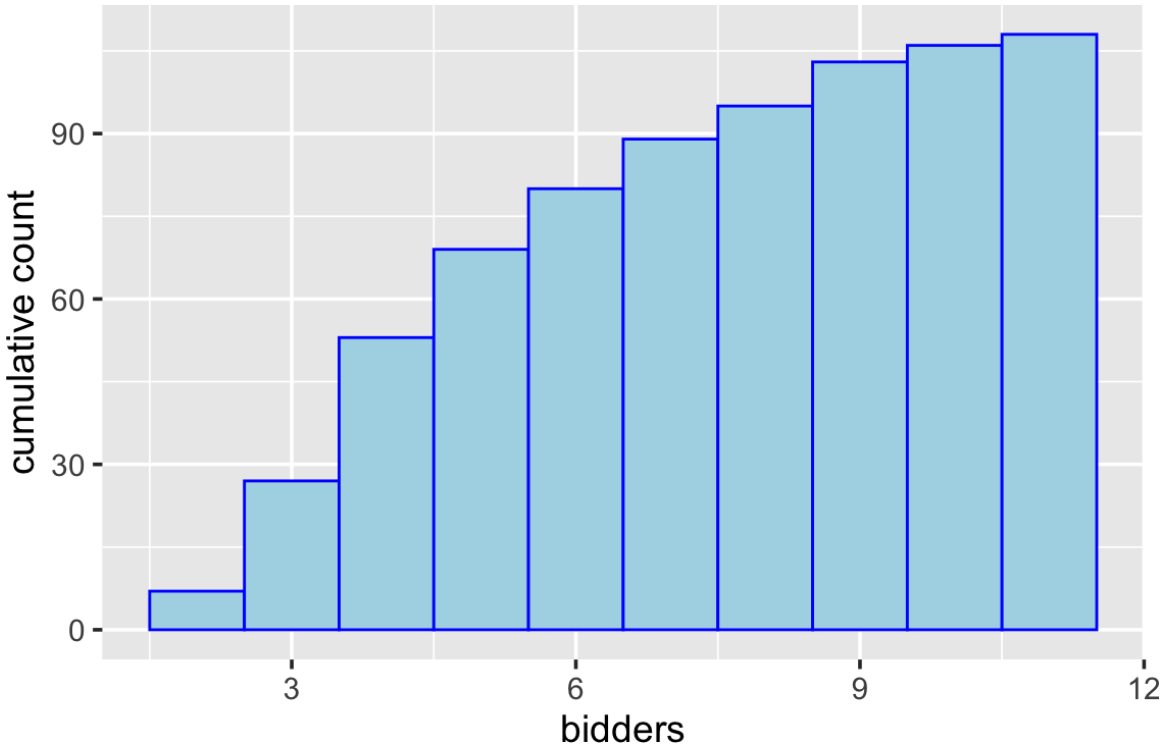
| bidders | contracts |
|---|---|
| 2 | 7 |
| 3 | 20 |
| 4 | 26 |
| 5 | 16 |
| 6 | 11 |
| 7 | 9 |
| 8 | 6 |
| 9 | 8 |
| 10 | 3 |
| 11 | 2 |

$a$) What proportion of the contracts involved at most five bidders?

$b$) What proportion of the contracts involved between five and ten bidders, inclusive?

$c$) Draw a cumulative frequency histogram.

# Cumulative frequency histogram



# Sample and population means

population mean: $\mu$ = sum of N population values / N

sample mean: $\bar{x} = \frac{x_1+x_2+\ldots+x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$

population median: $\tilde{\mu}$

sample median: $\tilde{x}$

# Measures of variability

**deviations from the mean**

$x_1 - \bar{x},\ x_2 - \bar{x},\ etc.$

Data: 3, 8, 11, 14
Mean: 9

| $value$ | $deviation$ | $deviation^2$ |
|---|---|---|
| 3 | -6 | 36 |
| 8 | -1 | 1 |
| 11 | 2 | 4 |
| 14 | 5 | 25 |

**Sum of squared deviations**

$S_{xx}$: 36 + 1 + 4 + 25 = 66

**Population variance**

$\sigma^2$ = 66/4 = 16.5

$\sigma^2 = \sum_{i=1}^{N} (x_i - \mu)^2 / N$

# Sample variance

**Sum of squared deviations**:

$S_{xx}$: 36 + 1 + 4 + 25 = 66

**Sample variance**:

$s^2$ = 66 / **3** = 22

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

**Why n-1?**

Short answer: using **n** would result in an underestimation, since the values in the sample are closer to the sample mean than to the true population mean (which we don't know)

# Standard deviation

**Square root of variance**

- Population s.d. = $\sqrt{\sigma^2}$
- Sample s.d. = $\sqrt{s^2}$
- *same units as original values*

# EXERCISE (p. 47, #62)

Consider the following information on ultimate tensile strength ($lb/in^2$) for a sample of $n$ = 4 hard zirconium copper wire specimens:

$\bar{x}$ = 76,831
$s$ = 180
smallest $x_i$ = 76,683
largest $x_i$ = 77,048

Set up equations to determine the values of the two middle sample observations. *Do not solve.*