Describing Variables

Stat 120

January 13 2023

Distribution

"The distribution of the variable Y"

- ***describes its center, variability and shape***
- ***use both numbers and graphics***

Center: Mean or Average

Mean: average value in a sample or population

- $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ is an average of n values y_i in a sample
- μ is an average value of y in a population

Student Survey

Example: The data StudentSurvey.csv is a sample of student survey responses obtained by the textbook authors

```
survey <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/Stud
mean(survey$Pulse) # the command `mean` computes an average
[1] 69.57459
```

The mean pulse rate for this sample of students is $\bar{y} = 69.6$ beats per minute.

Center: Median

Median: the middle value when the data are ordered

- The median splits the data in half
- m is the median value in a sample
- M is the median value in a population

```
median(survey$Pulse) # the command `median` computes an median
[1] 70
```

The median pulse rate for this sample of students is $m = 70$ beats per minute.

Variability: Standard Deviation

Standard Devation (SD): average value in a sample or population

- $s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$ is the SD of n values y_i in a sample
- σ is the SD of values of y in a population

```
sd(survey$Pulse) # the command `sd` computes an average
[1] 12.20514
```

The SD of pulse rates for this sample of students is $s = 12.2$ beats per minute. The "average" deviation of individual pulse rates around the mean value is about 12.2 beats per minute.

Missing Data in R

- *Missing data values in R are coded as **NA** values*
- *Many basic statistic functions in R return an **NA** value if variable has any missing values*

```
movies <-read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/Holly
mean(movies$WorldGross)
[1] NA
sd(movies$WorldGross)
[1] NA
```

This lets the user (you) know that at least one value (maybe many, many values!) are missing

Missing Data in R

- Use the summary command to find how many are missing

```
summary(movies$WorldGross)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
   0.025   30.706   76.659  150.742  173.691 1328.111        2
```

There are 2 movies with missing world gross amounts.

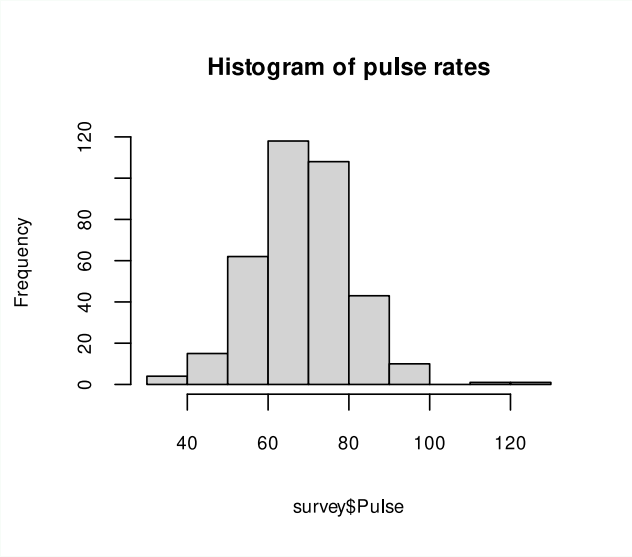
Add the argument na.rm = TRUE to remove missing values and get your summary stats:

```
mean(movies$WorldGross, na.rm = TRUE)
[1] 150.7423
sd(movies$WorldGross, na.rm = TRUE)
[1] 215.0186
```

Shape: histogram

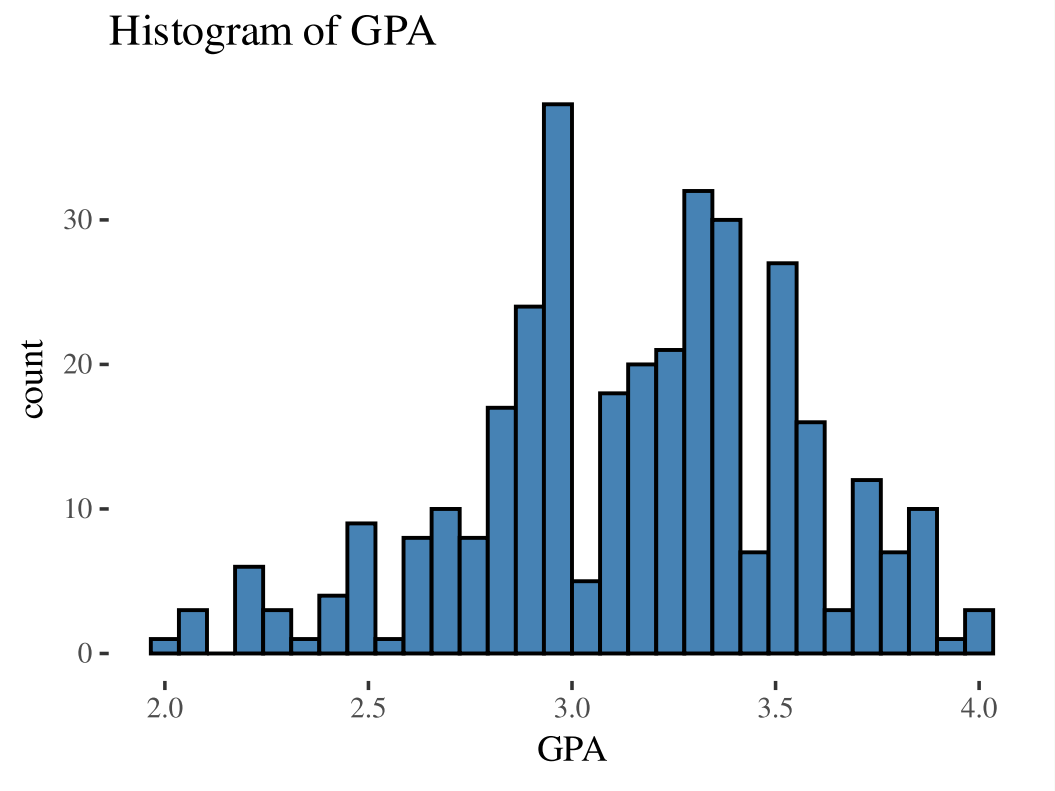
Histogram: aggregates values into bins and counts how many cases fall into each bin

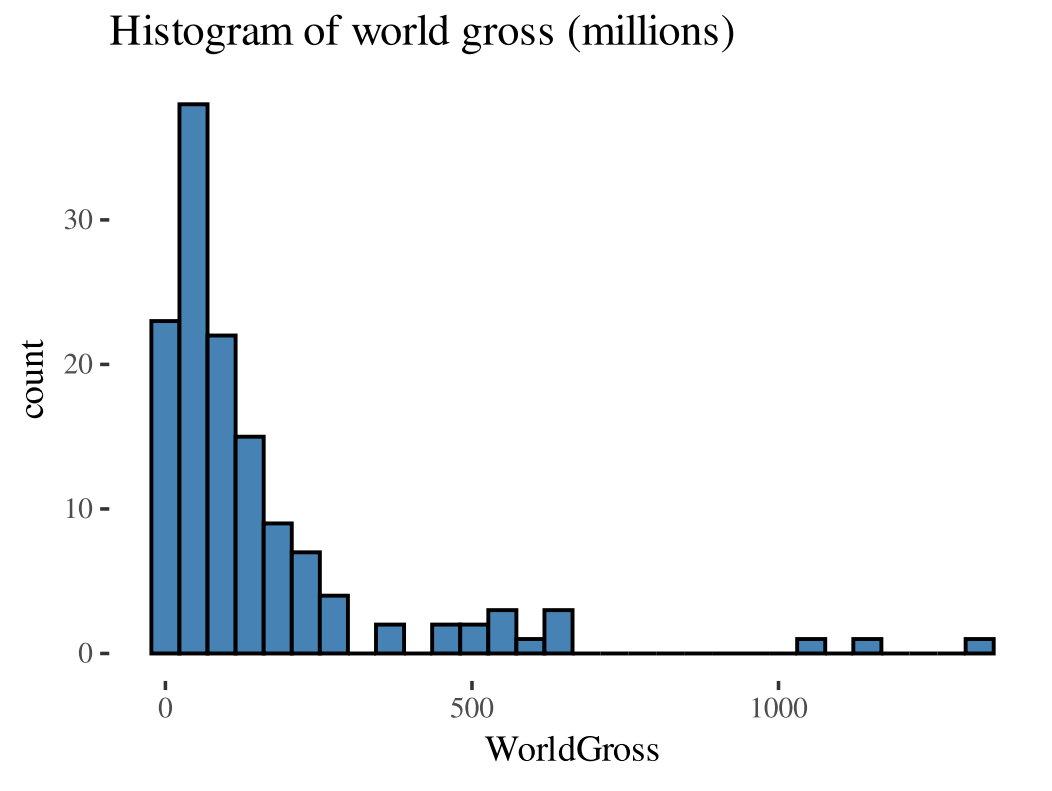
```
hist(survey$Pulse,
     main = "Histogram of pulse rates",
     cex.lab=0.7, cex.main = 0.9, cex.ax
```



- *Pulse rates are symmetrically distributed around a rate of about 70 beats per minute.*
- *Symmetric distributions are "centered" around a mean and median that are roughly the same in value.*

Shape: Left Skew & Right Skew






```
mean(survey$GPA, na.rm =T)
[1] 3.157942
median(survey$GPA, na.rm = T)
[1] 3.2
```

```
mean(movies$WorldGross, na.rm =T)
[1] 150.7423
median(movies$WorldGross, na.rm = T)
[1] 76.6585
```

Extreme values

outlier: *an observed value that is notably distinct from most other values in the dataset*

resistant: *a statistic is resistant to outliers if it is relatively unaffected by outliers*

- *Median is resistant to outliers*
- *Mean and SD are not resistant*

Movie world gross (millions of dollars) stats with and without Harry Potter movie (outlier!):

	Mean	SD	Median
with HP	150.7	215.0	76.7
without HP	141.9	189.7	75.0

Identifying extreme values in R

`which` identifies the **row number** of cases that satisfy a given criteria

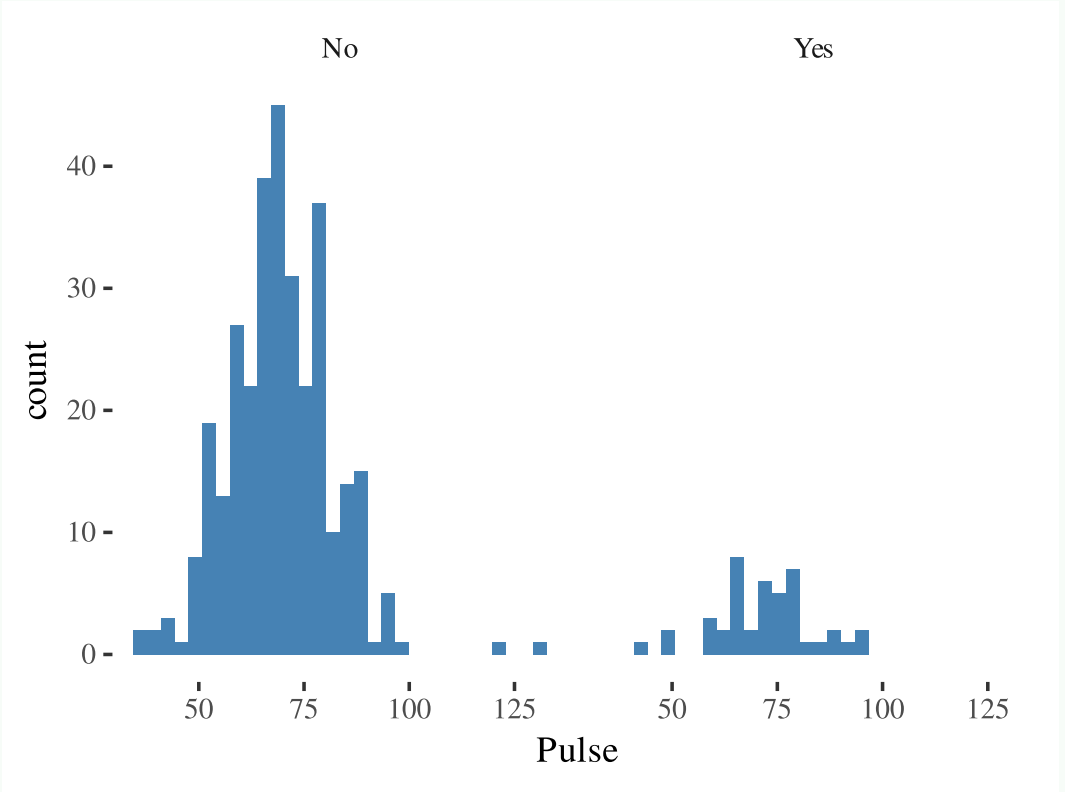
- Which movies had world gross bigger than 1200?

```
which(movies$WorldGross > 1200)  # gives row number
[1] 4
movies[4, c("Movie","WorldGross")]
                                                Movie WorldGross
4 Harry Potter and the Deathly Hallows Part 2   1328.111
```

Harry Potter (row number 4) had world gross of 1.328 billion dollars!

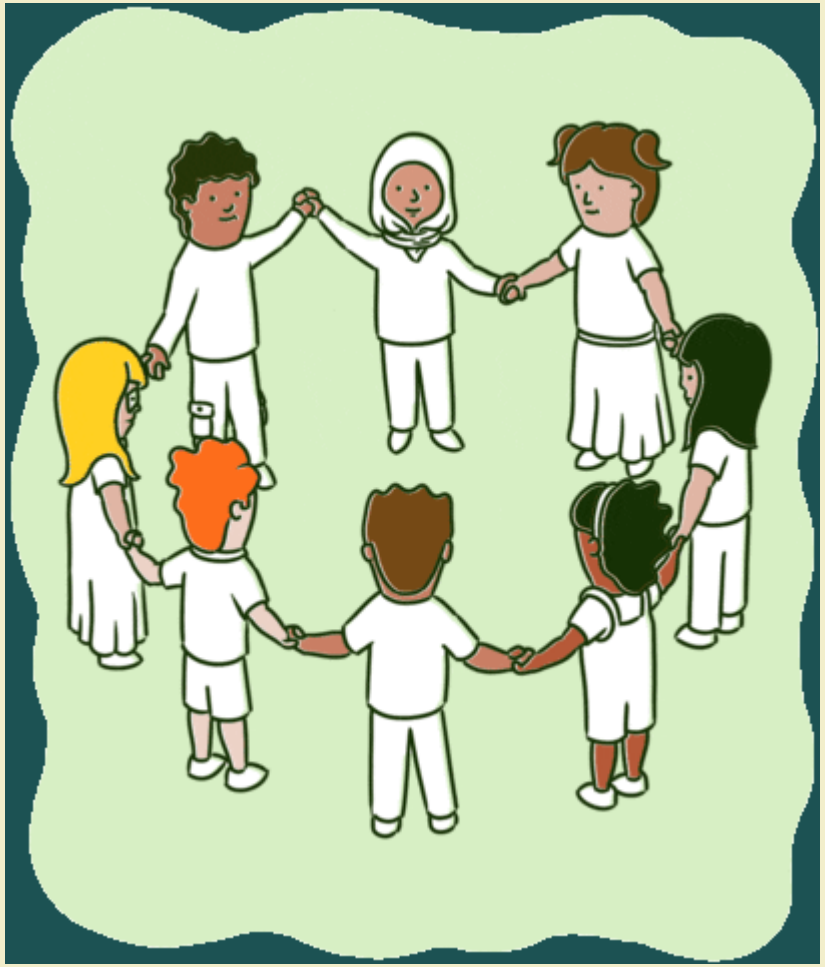
Adding a categorical variable: graphics

```
library(ggplot2)
ggplot(survey, aes(x=Pulse)) +
  geom_histogram(fill="steelblue") +
  facet_wrap(~Smoke)
```



Your Turn 1

10:00



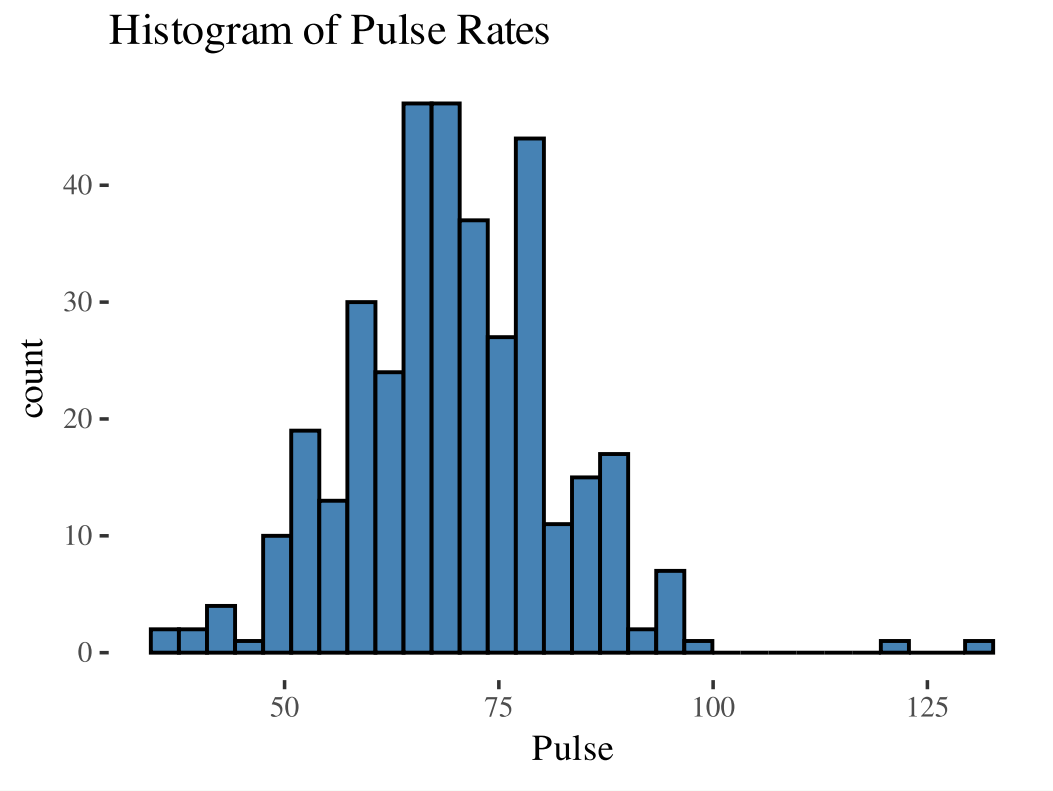
Go to our class activity .Rmd for today and skim through your turn 1

Feel free to talk to your neighbor

Shape and Stats

Mean and standard deviation are good summary stats of a symmetric distribution.

Similar variation to the left and right of the mean so one measure of SD is fine.




```
# mean
mean(survey$Pulse)
[1] 69.57459
```

```
# standard deviation
sd(survey$Pulse)
[1] 12.20514
```

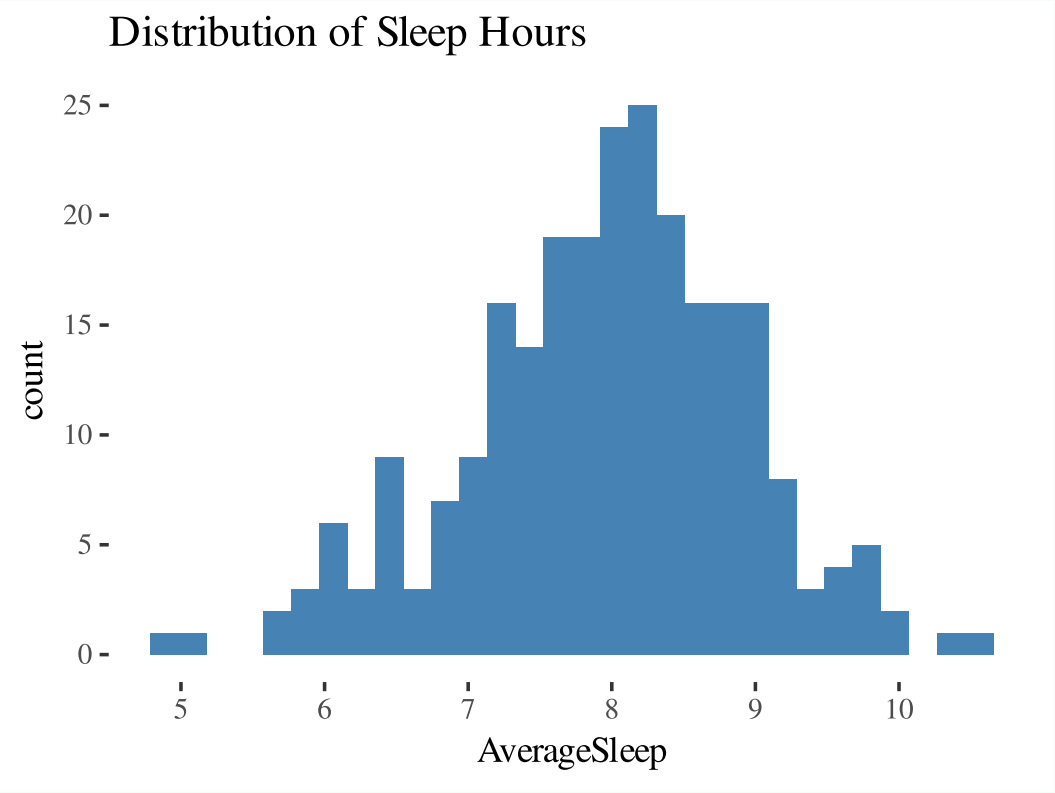
Shape: data distribution

If a distribution of data is approximately bell-shaped, about 95% of the data should fall within two standard deviations of the sample mean.

- *for a sample: 95% of values between* $\bar{y} - 2s$ *and* $\bar{y} + 2s$
- *for a population: 95% of values between* $\mu - 2\sigma$ *and* $\mu + 2\sigma$

Shape: Right skew

```
sleep <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/Sleep
```



Question The standard deviation for hours of sleep per night is closest to

(a) 0.5

(b) 1

(c) 2

(d) 4

Standardizing data: z-score

The z-score of a data value, x, tells us how many standard deviations the value is above or below the mean:

$$z = \frac{x - \text{mean}}{\text{SD}}$$

- E.g. if a value x has $z = -1.5$ then the value x is **1.5 standard deviations below** the mean.

Question: ***If we standardize all values in a bell-shaped distribution, 95% of all z-scores fall between what values?***

Shape and Stats: Quartiles

Quartiles divide values in to quarters

- 1st Quartile: Q_1 is the 25th percentile
- 2nd Quartile: Q_2 is the 50th percentile (median)
- 3rd Quartile: Q_3 is the 75th percentile

5-number summary is quartiles along with min and max: $\mathbf{min}, Q_1, \mathbf{m}, Q_3, \mathbf{max}$

Interquartile Range (IQR) is the range of the middle 50% of values:

- $IQR = Q_3 - Q_1$
- the **range** is just $\mathbf{max} - \mathbf{min}$

Shape and Stats: Quartiles

```
summary(movies$WorldGross)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
   0.025   30.706   76.659  150.742  173.691 1328.111        2
```

The 5-number summary is

- $\min = 0.025, Q_1 = 30.7, \mathrm{m} = 76.7, Q_3 = 173.7, \max = 1328.1$
- right skewed: variation is upper 25 of movies is much larger than lower 25%
  - upper range: $\max - Q_3 = 1328.1 - 173.7 = 1154.4$
  - lower range: $Q_1 - \min = 30.7 - 0.025 = 30.675$

Shape and Stats: Boxplot

Boxplot: *Visualization of 5-number summary*

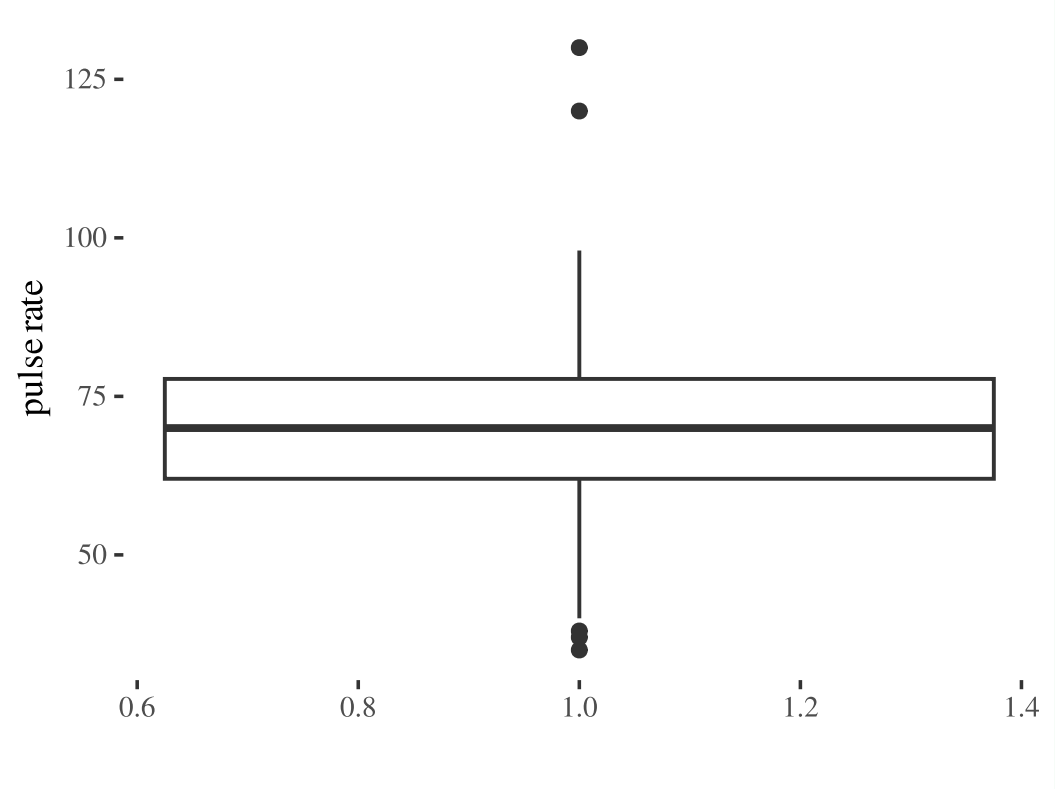
- *Draw a numerical scale appropriate for the data*
- *Draw a box stretching from Q_1 to Q_3*
- *Divide the box with a line at the median*
- *Draw a line from each quartile to the most extreme data value that is not an outlier*
- *Identify each outlier individually by plotting with a symbol such as an asterisk or dot*

Outlier rule of thumb: cases that are more extreme than

$$Q_1 - 1.5(IQR) \quad \text{or} \quad Q_3 + 1.5(IQR)$$

Shape and Stats: Boxplot

```
ggplot(data = survey, aes(x = 1, y = Pu
  geom_boxplot() +
  labs(x = "", y = "pulse rate")
```



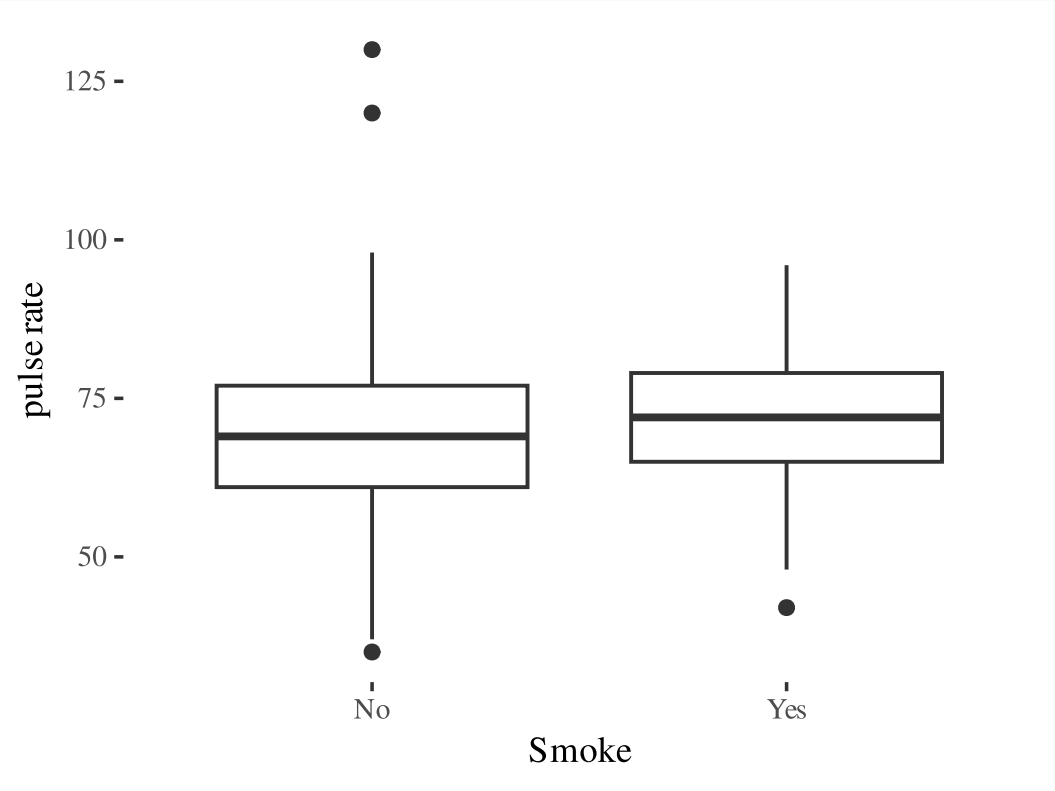

```
summary(survey$Pulse)
   Min. 1st Qu.  Median    Mean 3rd Qu.
  35.00   62.00   70.00   69.57   77.75
```

- $IQR = 77.75 - 62 = 15.75$
- $1.5(15.75) = 23.625$
- lower "fence" = $62 - 23.625 = 38.375$
- upper "fence" = $77.75 + 23.625 = 101.375$

```
which(survey$Pulse < 38.375)
[1]  55 106 200
which(survey$Pulse > 101.375)
[1]   3 171
```

Shape and Stats: Side-by-side Boxplots

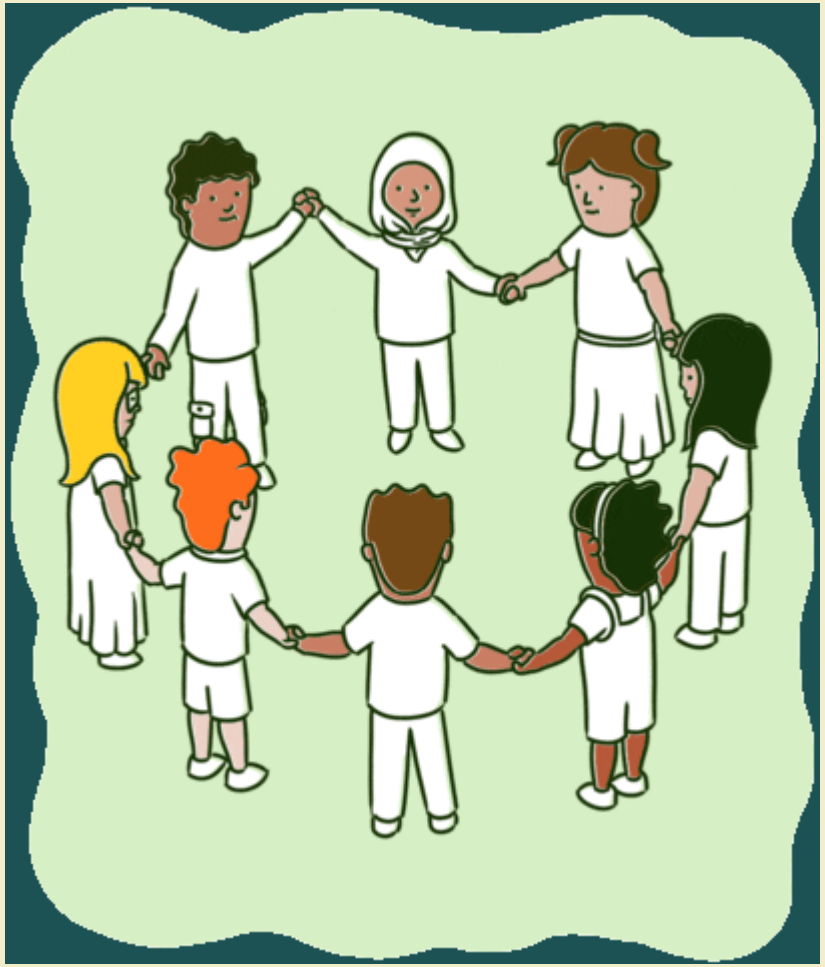
```
ggplot(data = survey, aes(x = Smoke, y
  geom_boxplot() +
  ylab("pulse rate")
```



- Median pulse rates are slightly higher for smokers than non-smokers (72 vs. 69 beats per minute) but variation is slightly lower (IQR 14 vs 16 beats per minute).
- Both distributions are roughly symmetric.
- Overall, just a slight association between smoking status and pulse rates.

Your Turn 2

10:00



Go over the remaining portion of in-class activity and let me know if you have any questions!