iClicker Question

How many of your professors know your name?

- Less than 1
- Between 1 and 2
- Between 2 and 5
- Between 5 and 10
- More than 10
Do you wish Amy knew your name?

- Yes
- No
- Don’t care
Data Cleaning
“This is not what I meant when I said ‘we need better data cleansing!’”
Garbage in, garbage out!

USE THE CRS DATABASE TO SIZE THE MARKET.

THAT DATA IS WRONG.

THEN USE THE SIBS DATABASE.

THAT DATA IS ALSO WRONG.

CAN YOU AVERAGE THEM?

SURE. I CAN MULTIPLY THEM TOO.
Children’s ages, reported in months

- In areas of the world that receive lots of international aid, caretakers are asked to report the age of their infants in months
- Often caretakers are unable to report an exact age, so instead report a nice round number
- Data are contaminated due to this recall bias
- Rounding up leads to over reporting of malnutrition, and rounding down leads to underreporting (missed cases of malnutrition)
Spring Weekend Poll Results

- Let’s say we want to know which artists the most Brown students voted for to come and perform during Spring weekend
- The poll asks students to write in an artist’s name
  - Pros: candidates are not limited to a fixed set of artists (anyone is fair game)
  - Cons: it’s hard to tally the results
    - How many people voted for Waka Flocka Flame?
    - Some people might write Waka Flocka; some might waka flacka, or waka floka, etc.
- We can easily change the case of a text to all upper or all lower
- Parsing out extra words and correcting spelling is more difficult. You will do this manually for a small data set in studio today.
What is data cleaning?

To produce technically correct data:

1. **Type checking**: verify that data values are stored correctly: e.g., numbers as numerics not characters; categories as factors; etc.

2. **Normalizing**: are data values comparable: e.g., is the gender value (a factor) for males stored as $M$ or $m$ or Male or male, etc.
What is data cleaning?

To produce consistent data:

1. **Correcting incorrect values** (e.g., negative ages, pregnant males, etc.)
   - But what if a very young child is reported married? Which is wrong, status or age?

2. **Handling extreme and missing values:**
   - Detect outliers, and possibly remove them
   - Possibly impute (i.e., infer) missing values
   - Use sound judgment, and always document and defend your decisions!
Type checking

Use type coercion functions:

```r
> as.character(2017)
"2017"

> as.numeric(TRUE)
1

> as.logical(0)
FALSE

> as.factor("Male")
Male
Levels: Male
```
String Manipulation

- `toupper` & `tolower`: changes the case of strings
- Good style and makes processing easier
  - E.g., `==` is case sensitive

```r
> toupper(c("Green", "Red"))
"GREEN" "RED"

> tolower(c("Green", "Red"))
"green" "red"
```
Normalization: strings

Use stringr library:

> str_trim("     Hello World!")
"Hello World!"

> str_pad("Hello World!", 5, "left")
"     Hello World!"

> str_detect("Hello World!", "ello")
TRUE

> str_replace("Hello World!", "Hello", "Yellow")
"Yellow World!"
Normalization: dates

Use `lubridate` library:

```r
> ymd("20110630")
"2011-06-30 UTC"

> dmy("30/06/2011")
"2011-06-30 UTC"

> mdy_hms("06302011111111")
"2011-06-30 11:11:11 UTC"
```
Normalization: dates

Use lubridate library:

> ymd("20110630")
"2011-06-30 UTC"

> dmy("30/06/2011")
"2011-06-30 UTC"

> mdy_hms("06302011111111")
"2011-06-30 11:11:11 UTC"

Use lubridate library:

> day(ymd("20110630"))
30

> year(dmy("30/06/2011"))
30

> hour(mdy_hms("06302100111111"))
11
Libraries for cleaning data

- `stringr (str_detect, str_replace, ...)
- `lubridate (ymd, mdy, dmy, hms, ymd_hms, ...)
- `tidyr (gather, spread, unite, separate, ...)
- `...
Ways to clean data

- Edit text input for a single variable to be consistent (spelling, capitalization, spacing)
- Standardize units of measurement for a single variable
- Remove duplicate rows (common) or columns (less frequent)