A Diamond in the Rough (Part a)

Unit 1 - Lab 7a

Directions: Follow along with the slides and answer the questions in **BOLDED** font in your journal.

Messy data? Get used to it

- Since lab 1, we've been using data from the CDC.
- What you might not have noticed was how **clean** the data was:
  - Variables were named so we could understand what they were about.
  - There didn't seem to be any *typos* in the values.
  - Numerical variables were considered numbers.
  - Categorical variables were composed of categories.
- Unfortunately, more often than not, data is **messy** until YOU clean it.

Cleaning data takes a while

- *Munging* messy data takes a long time.
- We'll split this lab into 2 parts:
  - In this lab, we'll fix our numerical variables.
  - In the next lab, we'll fix our categorical variables.

Messy data?

- What do we mean by messy data?
- Variables might have **non-descriptive names**
  - *Var01*, *V2*, *a*, ...
- Categorical variables might have **misspelled categories**
  - *"blue"*, *"Blue"*, *"blu"*, ...
- Numerical variables might have been **input incorrectly**. For example, if we're talk about people's height in inches:
  - *64.7*, *68.6*, *676*, ...
- Numerical variables might be **incorrectly coded** as categorical variables (Or vice-versa)
  - "64.7", "68.6", "67.6"

The American Time Use Survey

- To show you what **messy** data looks like, we'll check out the *American Time Use Survey*, or *ATUS*.
- What is ATUS?
  - It's a survey conducted by the US government (Specifically the Bureau of Labor Statistics)
  - Survey thousands of people to find out exactly what activities they do throughout a single day.

  - Combine the thousands of people together to get an idea about how much time the typical person living in the US spends doing various activites.
- Let's take a look at the data before it's been properly cleaned.
  - Data scientists call the act of cleaning data **munging**
- Type the following commands into your console:

```
data(atus_dirty)
```

```
View(atus_dirty)
```

- **Write down as many problems with the data as you can find.**

Fixing Variable Names

- Let's start our cleaning by fixing the variables names.
- Currently the variables are named: caseid, V1, V2, V3, V4, V5, V6, V7
  - Verify this by typing `names(atus_dirty)` into the console.

Description of ATUS Variables

- The description of the actual variables:
  - `caseid`: Anonymous ID of survey taker.
  - `V1`: The age of the respondent.
  - `V2`: The gender of the respondent.
  - `V3`: Whether the person is employed full-time or part-time.
  - `V4`: Whether the person has a physical difficulty.
  - `V5`: How long the person sleeps, in minutes.
  - `V6`: How long the survey taker spent on homework, in minutes.
  - `V7`: How long the respondent spent socializing, in minutes.

New name, same old data

- To fix the variable names, we need to *assign* a new set of names in place of the old ones.
- Something like:

```
names(atus_dirty) <- c(new_name1, new_name2, ...)
```

- This would take the first variable (caseid) and rename it `new_name1`.
- It would then take the second variable (caseid) and rename it `new_name2`.
- And so on...

On your own

- **Come up with new variable names for each variable in the data**
  - Good names should be short and describe what the variable is related to.
  - Use an *underscore* '_' to combine 2 words or abbreviated words if you'd like.
- **Rename your variables using the method on the previous slide**
- **`View` your data when you're done. Make sure that the names are in the correct order**

Everyone together

- To keep everyone together for the rest of the lab, let's agree to adopt a common set of variable names.
  - (I'm sure your variable names were very good).
- Type the following into the console:

```
names(atus_dirty) <- c("caseid",
                       "age",
                       "gender",
                       "fulltime_emp",
                       "phys_challenge",
                       "sleep",
                       "homework",
                       "socializing")
```

Playing with Strings

- In programming, a **string** is sort of like a *word*.
  - It's a value made up of **characters** (i.e. letters)
- The following are example of strings. Notice that each **string** has quotes before and after.

```
"string"
```

```
"A1B2c3"
```

```
"Hot Cocoa"
```

Numbers are words? (Sometimes)

- Type the following commands into the console:

```
0015
```

```
"0015"
```

- **What's different about each output?**
- **What do you think would happen if we multiplied two *strings* together?**
  - **Create two strings and try it!**

Changing strings into numbers

- **strings** in R are called **character** objects.

- Click on the *Environment* pane and find the `atus_dirty`, data.
- Click on the blue arrow next to it.
  - Find the `age` variable under `atus_dirty`
- **Notice that R thinks, for the moment, that *age* is a *chr* or *character* object**
- **How many of the other variables you thought should be numerical variables were miss-specified as *character* objects?**
- To fix this problem, we need to tell R to think of our *numeric* variables **`as.numeric`** variables.
- When we use **`as.numeric`**...
  - We can turn *characters*: **3.14**
  - Back into *numbers*: **3.14**
- To fix our **age** variable then, we'd write:

```
atus_dirty <- transform(atus_dirty,
                        age = as.numeric(age))
```

- This code is telling R:
  - "Take the our current *atus_dirty* data..."
  - "... and over-write it with ..."
  - "... the *atus_dirty* data where the values of my *age* variable are numbers."

Transforming many variables at once

- We can also use the **`transform`** function to change many variables to numbers at once.
- For example:

```
atus_dirty <- transform(atus_dirty,
                        age = as.numeric(age),
                        sleep = as.numeric(sleep))
```

- **Translate what the above line of code is doing into words your great-grandmother could understand.**

On your own

- **Using the steps you just followed to change `age` into a numeric variable, do the same stapes for the following variables:**
  - `sleep`
  - `homework`
  - `socializing`
- **Why shouldn't we change the caseid variable into a *numerical* variable?**