Name: Delaney Edwards\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date: 02/16/21\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Goal: learn how to create good spreadsheets!**

**Purpose:**

1. **What is a spreadsheet and how do they work**
2. **What separates a good dataset from a bad one**
3. **How can we apply best practices to create a spreadsheet that is useful to other**

What is clean data? Clean data are in a format that is ready to analyze. Characteristics of clean data include data that are:

Common symptoms of messy data include data that contain:

* Special characters (e.g. commas in numeric values)
* Numeric values stored as text/character data types
* Duplicate rows
* White space
* Missing data
* Zeros instead of null values
* “highlights” and colors used to denote information
* More than one type of info stored in a cell

# Part A: For each spreadsheet identify at least one thing that makes it problematic. Then describe how you might go about fixing it (i.e., what could you do to make it ‘clean.’

**For help see:** [**https://datacarpentry.org/spreadsheets-socialsci/02-common-mistakes/index.html**](https://datacarpentry.org/spreadsheets-socialsci/02-common-mistakes/index.html) **or watch the online video I filmed and put on AsULearn for Week 4**

1. **Sheet 1**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| **Sample** | **Date excavated** | **Date analyzed** | **Element** |
| **1** | **5/9/2019** | **1/5/20** | **Radius** |
| **2** | **5/1/2019** | **1/5/20** | **Radius** |
| **3** | **5/3/2019** | **1/6/20** | **Ulna** |
| **4** | **W** | **1/6** | **femur** |

**Problem:** Dates are written in a format that will make them difficult for the computer to understand. There is also a date labeled “w.”

**Better version:** Writing the dates in the format 2019-05-09, 2019-05-01, 2019-05-03, etc. If “W” is to represent an unknown value, it should be represented with “NA”

1. **Sheet 2**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Pit A |  | Pit B |  |  |
|  |  |  |  |  |  |
| Munsell | weight | Date | Munsell | weight | Date |
| 10 yr 4/3 | 1.5 | 2020-02-03 | 10 yr 2/3 | 7.5 | 2020-02-05 |
| 10 yr 4/3 | 1.7 |  | 10 yr 2/3 | 3 |  |
| 10 yr 4/3 | 2.0 |  | 10 yr 2/3 |  |  |
| 10 yr 4/3 | 2.1 | 2020-02-04 | 10 yr 2/3 | 3.3 |  |
| 10 yr 4/4 | 2 |  | 10 yr 2/3 | 9 |  |

**Problem:** There are cells that lack any information, which the computer will have trouble interpreting. I also assume that having both Pit A and Pit B in one singular table could potentially cause problems.

**Better version:** Using NA or zeros instead of blank cells. To be safe, put pit A and pit B in two different files/tables.

1. **Sheet 3**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  |  | Site | elements | Biological sex | Analysist |  |
|  | Skeleton UW-503 |  |  |  |  |  |
| Transect A |  | Oblion | Pelvis | Male | GP |  |
|  |  |  | femur | male | GP |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  | Skeleton UW-504 |  |  |  |  |  |
| Transect B |  | PineTree | Cranium | Female | SL |  |
|  |  |  | Pelvis | Female | SL |  |
|  |  |  | Tibia | NA | SL |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

**Problem:** There are essentially two different tables in one, which will severely confuse the computer. The way the data table is formatted is highly confusing, as there are rows with data but no label to indicate relations. Names of some columns and rows are written with spaces.

**Better version:** use different files for each table of information to avoid confusion and loss of data. Names of columns and rows should be written in camel case or snake case with no spaces or special characters.

1. **Sheet 4**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Week 1 |  | Week 2 |  | Week 3 |  |  |
| ID | SEX | date | weight | date | weight | Date | weight |  |
| 002 | M |  |  |  |  |  |  |  |
| 005 | F |  |  |  |  |  |  |  |
| 006 | F |  |  |  |  |  |  |  |
| 010 | M |  |  |  |  |  |  |  |
| 012 | M |  |  |  |  |  |  |  |
| 015 | M |  |  |  |  |  |  |  |

**Problem:** This may not be what you were looking for or correct, as I am still working to understand all of the standards and details for spreadsheets, but I feel as though Week 1, Week 2, and Week 3 headers could potentially confuse the computer and make data hard to read.

**Better version:** To avoid any issues, each week should be separated into their own respective spreadsheets.

1. **Sheet 5**

|  |  |  |
| --- | --- | --- |
| Sample | date | Glucose level |
| 101 | 2019-01-14 | 150.5 |
| 102 | 2019-01-14 | 93.6 |
| 103 | 2019-01-14 | 99.5 |
| 104 | 2019-01-15 | 108.0 |
| 105 | 2019-01-15 | 7.9 |
| 106 | 2019-01-15 | 102.7 |

**Problem:** highlighted cell, which I assume to mean that this value should be left out or something special should be done with it

**Better version:** create another column that would convey the information you are trying to convey by highlighting the cell

# Part B: fixing/data cleaning examples

Sadly, most folks don’t follow the above suggestions. Here is an example, modified from a colleague’s assignment, that lets you see how to work with messy data

You can download the data for this part on [this google drive folder](https://drive.google.com/drive/u/1/folders/1zQtdbwQuVrxtHrFG7lXdW02ZV0r5cKnu): (https://drive.google.com/drive/u/1/folders/1zQtdbwQuVrxtHrFG7lXdW02ZV0r5cKnu)

The file are from New York City and give info on real estate sales for different boroughs ( administrative divisions that make up New York City) …Let’s say you wanted to help a friend who was trying to learn more about this and they gave you these files and asked you what you can learn. Put the data into your current project folder so that you can access it easily (if you don’t have Excel on your machine don’t worry about that. You should be able to open the file just to view it on the Drive)

## Setup

To do this part of the lab make sure you run the following code as we will need these R libraries.

library(tidyverse)

install.packages("janitor") #this helps to clean our data

library(janitor)

library(readxl) # this package comes when you install the tidyverse

#package but is only loaded individually

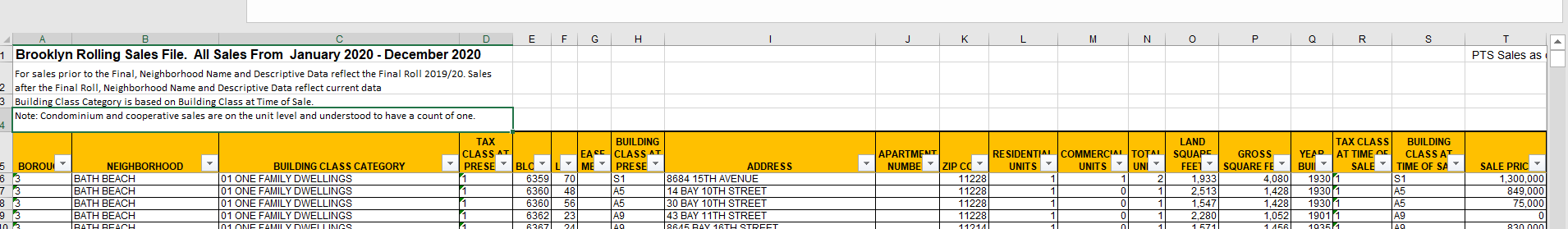
## Step 1: take a look at the raw data

* + Take a look at the Excel sheet either in the Google folder or after downloading and make some notes about what problems you see that you may want to edit/fix as you go

## Step 2: get the data into R

In the folder you will see multiple versions of essentially the same file (one is an excel sheet and the other is in a CSV (with some of the text edited out…not ideal but c’est la vie).

* + - If you have Excel on your computer open it and take a look…you will see something like I copied below. Note that the first few lines are info that we don’t need in our data table,,,



* + - We are going to read the data in with read\_xls(), a function in the readxl package
    - The function read\_xls takes a number of arguments (the stuff that goes in between the parenthesis)
* path = where on your machine the file is stored
* sheet = the name of the sheet (some excel files have multiple sheets)
* na = what to count as missing value (by default it counts blank cells as missing
* skip = Minimum number of rows to skip before reading anything, be it column names or data. Leading empty rows are automatically skipped, so this is a lower bound. Ignored if range is give
  + - looking at these data it seems that all we need here is to tell it to skip the first 4 line
    - on my computer, I have the data in my working directory in a folder called “datasets.” You will have to change this to wherever you put the data. If you are having trouble PLEASE reach out. Getting a handle on this is one of the biggest hurdles and I am here to help!

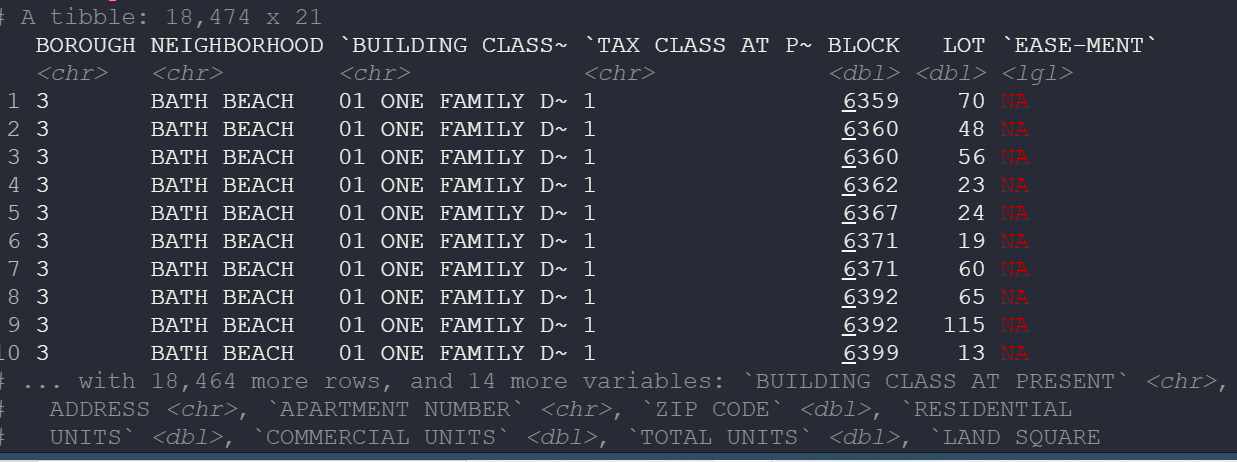
brooklyn <- read\_xls(path ="datasets/rollingsales\_brooklyn.xls", skip = 4)

Note: the ‘path’ is where the data sheet is located on the computer. If you write path = “” and hit the tab key between the quotation marks you should see the files in your current working directory pop up. That will let you choose the file you want

* + We now have the dataset stored as an object called “brooklyn.” To examine it we can just type the function’s name into R

brooklyn

you will see something like this:



* + Note that it tells us there are 11,874 rows and 21 columns of data…that is a lot of data!
  + It also shows the first few rows and data.
  + Now, enter the following into R to create two new objects that have the same data source but use different *functions* for reading in the data (read.csv and read\_csv). Note that this time we are using the .csv file rather than the .xls file

brooklyn\_csv\_base <- read.csv("datasets/rollingsales\_brooklyn\_csv.csv", skip = 4)

brooklyn\_csv\_readr <- read\_csv("datasets/rollingsales\_brooklyn\_csv.csv", skip = 4)

Now compare the three different datasets….what differences do you see in how the data are stored? One way to do this is with the Tidyverse’s glimpse() function, which gives a nice summary of the object. You will notice that with glimpse you can see the name of each column and then the class of that column. Pay attention to each object and when the class of a column is different.

glimpse(brooklyn\_csv\_base)

glimpse(brooklyn\_csv\_readr)

glimpse(brooklyn)

jot down some notes about how these objects differ:

The excel file I used only comes up with 6,056 rows, whereas both csv objects have 8,146 rows.

In each object, the columns are labeled differently. In the excel object, the first column is BOROUGH, in the csv\_base is it X2, and in the csv\_readr it is ‘2’.

This may be incorrect (it likely is) but I assume that these labels - <int>, <chr>, <lgl> - stand for the class of each set of data values, and this differ based on the object. In the excel object, the first row is labeled characters, but in csv\_base, it is labeled integer.

I can see that there is variation and trouble with the interpretation of dates in the objects. In the excel file the column reads “SALE DATE”, in csv\_base it is labeled “X7.23.2020” and in csv\_readr it is labeled “’7/23/2020’”.

Step 2: connecting datasets together

Note how each borough is setup the same in the sheets. This is a happy thing for us. Lets say we want to look at all the data at one go!

brooklyn <- read\_excel("rollingsales\_brooklyn.xls", skip = 4)

bronx <- read\_excel("rollingsales\_bronx.xls", skip = 4)

manhattan <- read\_excel("rollingsales\_manhattan.xls", skip = 4)

staten\_island <- read\_excel("rollingsales\_statenisland.xls", skip = 4)

queens <- read\_excel("rollingsales\_queens.xls", skip = 4)

# Bind all dataframes into one, save as "NYC\_property\_sales"

NYC\_property\_sales <- bind\_rows(brooklyn, bronx, manhattan, staten\_island, queens)

glimpse(NYC\_property\_sales)

Now we have all the data in one dataset. Note the column “BOROUGH” is a numerical code for the borough so that info is still present. We could rename them if we wanted to…

Step 3: Clean up the data: Let’s circle back to Brooklyn dataset from above

* Take a look at the col names :

colnames(brooklyn)

take a moment and jot down some thoughts about what we might want to fix about the column names

I would we write them in lowercase letters instead of all caps and put them in snake case to get rid of the spaces in the names

To me, a few things stick out about the names of the columns that may make data analysis tricky

* 1. All capitalized
  2. Spaces in the col names

We can fix this in a number of ways:

* **To get the capitalization fixed:**

Take a look at the function below. A few things are happening here. The “%>%” is a function from the tidyverse that makes coding a bit easier. It is called a pipe and basically it says “then”. So you can read the code as “take the column names from brooklyn then use the str\_to\_title function. This function, also part of the Tidyverse, take a string (an ordered sequences of characters) and changes the case of the letters. Play around and see what other str\_to\_ functions exist and what they do

colnames(brooklyn) %>% str\_to\_title()

* But this code doesn’t change anything. to save this we just assign it to colname(brooklyn)

colnames(brooklyn) <- colnames(brooklyn) %>% str\_to\_title()

**Now, we want to remove the spaces**

Thankfully Tidyverse gives us that with the useful str\_replace\_all() function, Here, the str\_replace\_all() takes two arguments. The first one is what we want to replace and the second it what we want to put in its place.

colnames(brooklyn) %>% str\_replace\_all(" ", "\_")

In this instance, we are replacing a space with a underscore. How would you replace the space with it with a . ?

Remember to save it:

colnames(brooklyn) <- colnames(brooklyn) %>% str\_replace\_all(" ", "\_")

fun fact: you could do this all in a single line like this: colnames(Brooklyn) <- colnames(brooklyn) %>% str\_to\_title() %>% str\_replace\_all(" ", "\_")

**Protip:**

Now that you did that I’m going to teach you a trick using the janitor package we installed at the beginning. In that package is a function called clean\_names() which will change the case and remove spaces in one go!

brooklyn %>% clean\_names()

# Part 3: Try it on your dataset

Take a look at the data you are examining for this course. Thinking about what we talked about this week and in this lab is there anything in your dataset that needs to be ‘cleaned to that it can work well. If so, what steps might you think about here?

Being very honest, the spreadsheet that I downloaded for my project is formatted in a way that is strange and almost unreadable to me, so I am still working on trying to understand the information that it is providing. So far, as far as I can tell, the names for the columns are actually okay based on the standards from the last couple weeks, so there is nothing to change there. I can tell that there are some missing values that were left blank rather than being labeled ‘NA’, so changing that will help to clean up the data and make it more user friendly.

# Part 4: assessment

1. How can you apply ‘clean’ spreadsheet techniques to other work you do?

I have already started labeling any document I create on my laptop (word, excel, ppt, etc) using the same methods as have been described in the past couple of weeks regarding spreadsheets (snake case is my favorite).

1. After doing this lab, what sorts of reasons might you give for the importance of good data organization?

Good data organization is imperative for having R studio, and presumably every other stats software, be able to read and understand the data it has been given. In the examples that have been shown as well as the bit of work we have done in R I can see that even the most tiny errors in spacing or labeling can prevent an entire chunk of code from being able to run.

1. “Reading in” data to R, especially *messy* data, is one of the most frustrating parts of data analysis. Do you think you are getting the hang of it? What could make this easier?

I think after talking to you I am beginning to understand the process more. I honestly don’t think anything else could have made this easier, it is just one of those things I have to struggle with a bit before understanding.

1. Next week we are starting to think about data visualization. For your dataset what sorts of figures might you want to make?

I believe scatter plots may be useful, especially looking at the differences in the male and female measurements. Doing a scatter plot for each measurement of the radius (maximum length, transverse diameter of the shaft, etc,) to compare male versus female numbers could be an interesting way to help me and other understand the data (with stuff like this I’m a visual learner).