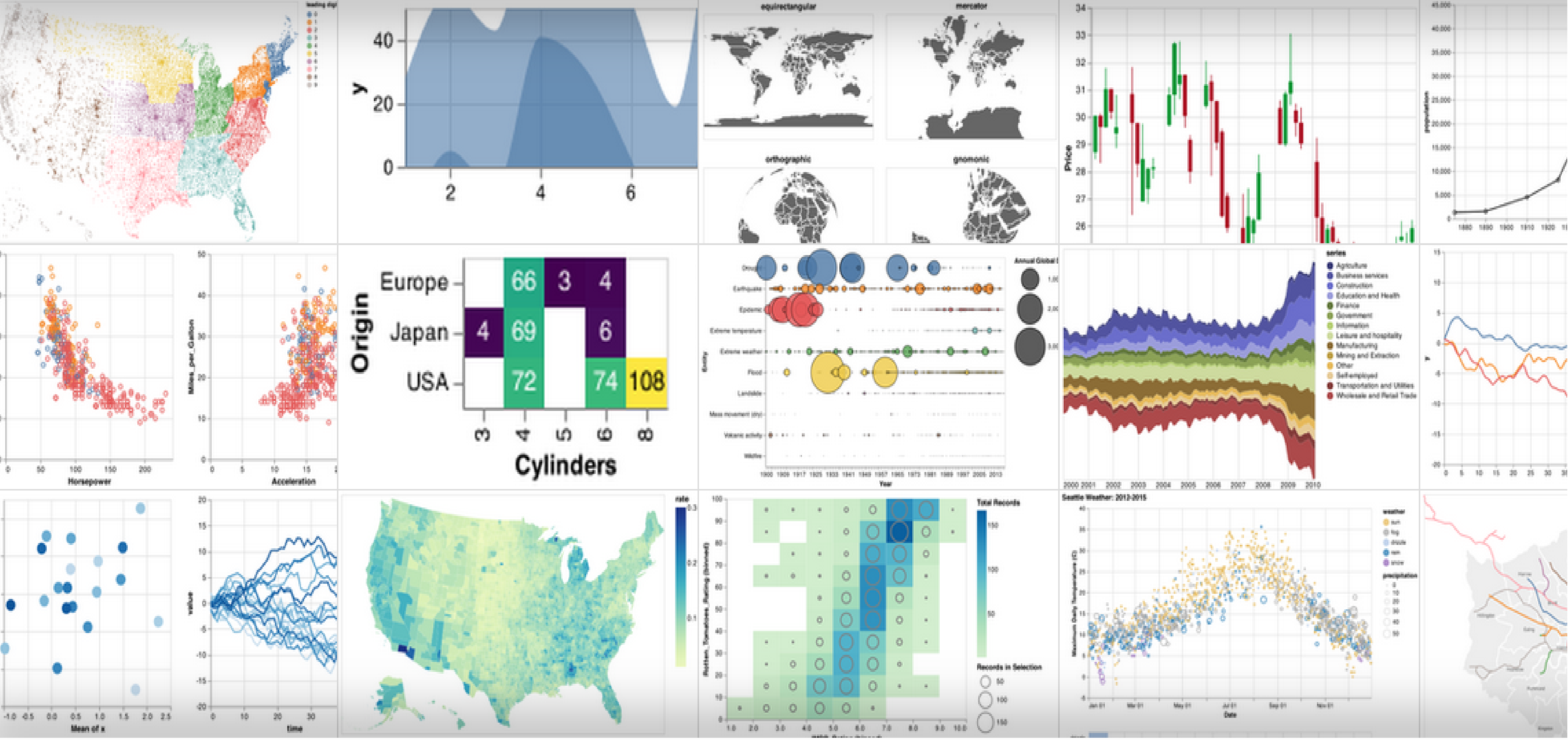


Data 227
Autumn, 2023
Tidy data, spreadsheet tips

Tidy Data

“Tidy datasets provide a standardized way to link the structure of a dataset (its physical layout) with its semantics (its meaning).”

No one will demand that our data be tidy; they will demand that the analysis get done. Tidying data is not always necessary or useful, but this is a tool that may help us on the path to avoid suffering.

Data Semantics

- A dataset is a collection of values, usually either numbers (if quantitative) or strings (if qualitative).
- Values are organized in two ways. Every value belongs to a variable and an observation.
- A variable contains all values that measure the same underlying attribute (like height, temperature, duration) across units.
- An observation contains all values measured on the same unit (like a person, or a day, or a race) across attributes.

Wickham, H. . (2014).

Tidy Data principles

There are three interrelated rules which make a dataset tidy:

- Each variable forms a column.
- Each observation forms a row.
- Each type of observational unit forms a table.

Wickham, H. . (2014).

Common problems 1

Column headers are values, not variable names (shows up frequently in tabular data).

```
## Row Col Hgt90 Hgt96 ... Deer97 Cover95 Fert
Spacing

## 0 1 1 NaN NaN ... -2147483648 0 0 15

## 1 1 2 14.0 284.0... 12015

## 2 13 17.0 387.0... 01015

## 3 1 4 NaN NaN ... -2147483648 0 0 15

##4 1 5 24.0 294.0...

##5 1 6 22.0 310.0...

##6 1 7 18.0 318.0...
```

Common problems 2

```
##                                       Name  ...
Value
## 0  ’C’ A CATERING AND EVENT COMPANY  ...  4300-4304 N CENTRAL
AVE
## 1  ’C’ A CATERING AND EVENT COMPANY  ...  60634
## 2  ’C’ A CATERING AND EVENT COMPANY  ...  Restaurant
## 3  ’C’ A CATERING AND EVENT COMPANY  ...  No Entry
## 4  ’C’ A CATERING AND EVENT COMPANY  ...  4300-4304 N CENTRAL
AVE
## 5  ’C’ A CATERING AND EVENT COMPANY  ...  60634
## 6  ’C’ A CATERING AND EVENT COMPANY  ...  Restaurant
## 7  ’C’ A CATERING AND EVENT COMPANY  ...  Pass
```

Common problems 3

```
##                       V1   V2   V3   V4   V5  ... V59 V60 V61
## 0  MX000017004195504TMAX  310  310  310  320  ...             -214
## 1  MX000017004195504TMIN  150  150  160  150  ...             -214
## 2  MX000017004195504PRCP    0    0    0    0  ...             -214
## 3  MX000017004195505TMAX  310  310  310  300  ...             -214
## 4  MX000017004195505TMIN  200  160  160  150  ...             -214
## 5  MX000017004195505PRCP    0    0    0    0  ...             -214
## 6  MX000017004195506TMAX  300  290  280  270  ...             -214
## 7  MX000017004195506TMIN  160  160  150  140  ...             -214
```

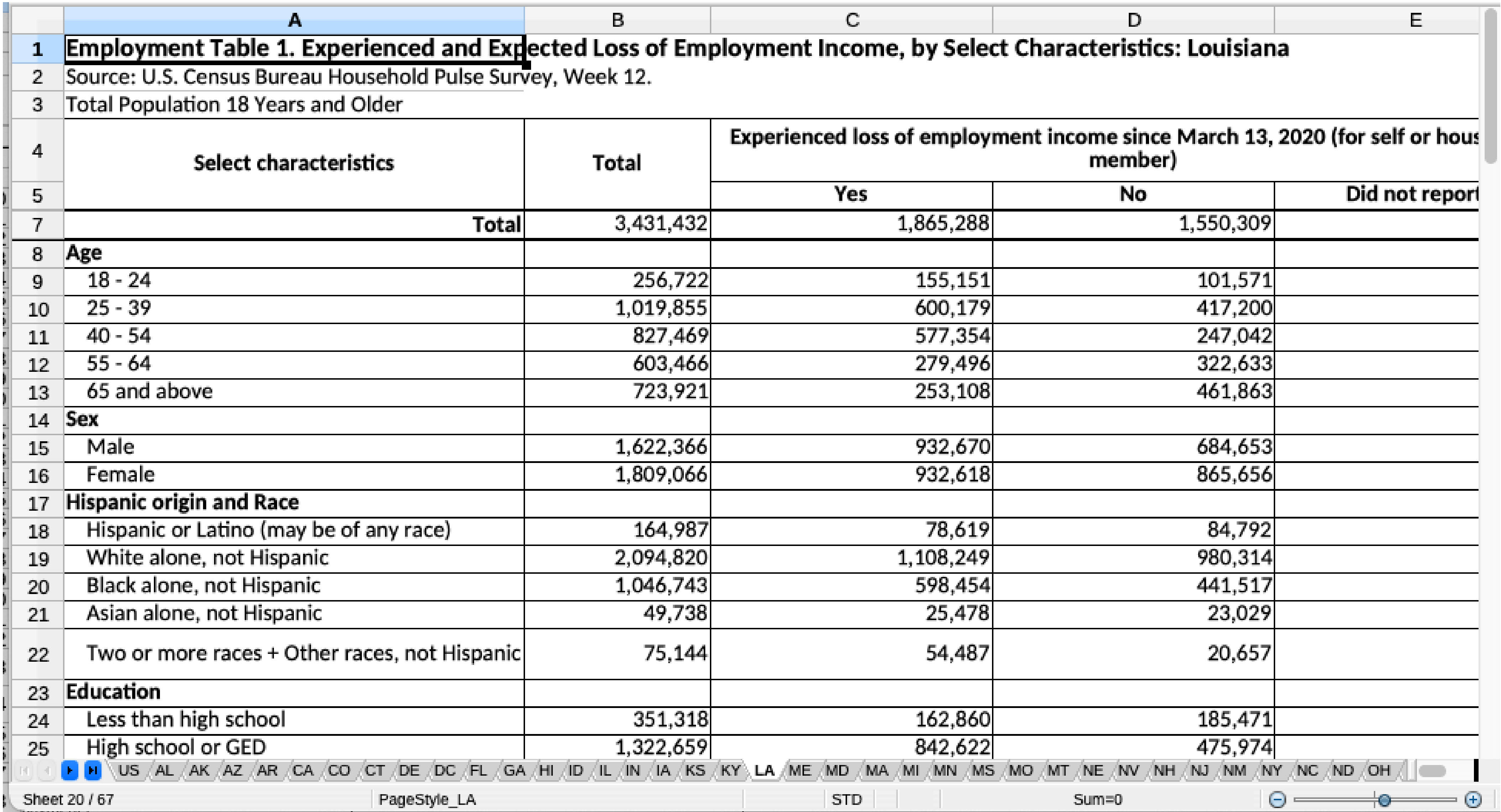
Common problems 4

```
## [’artist’, ’track’, ’date.entered’, ’wk1’, ’wk2’, ’wk3’, ’wk4’]
## [’wk6’, ’wk7’, ’wk8’, ’wk9’, ’wk10’, ’wk11’, ’wk12’, ’wk13’]
## [’wk14’, ’wk15’, ’wk16’, ’wk17’, ’wk18’, ’wk19’, ’wk20’, ’wk21’]
## [’wk22’, ’wk23’, ’wk24’, ’wk25’, ’wk26’, ’wk27’, ’wk28’, ’wk29’]
## [’wk30’, ’wk31’, ’wk32’, ’wk33’, ’wk34’, ’wk35’, ’wk36’, ’wk37’]
## [’wk38’, ’wk39’, ’wk40’, ’wk41’, ’wk42’, ’wk43’, ’wk44’, ’wk45’]
## [’wk46’, ’wk47’, ’wk48’, ’wk49’, ’wk50’, ’wk51’, ’wk52’, ’wk53’]
## [’wk54’, ’wk55’, ’wk56’, ’wk57’, ’wk58’, ’wk59’, ’wk60’, ’wk61’]
```

Common Problems...

- A single observational unit is stored in multiple tables or files– these tables and files are often split up by another variable, so that each represents a single year, person, or location, etc.
- As long as the format for individual records is consistent, this is an easy problem to fix.
- For example, consider the Household Pulse Survey run by the US Census Bureau.
- “Designed to quickly and efficiently deploy data collected on how people’s lives have been impacted by the coronavirus pandemic.”
- You can visit the Household Pulse Survey website to download multiple files, one for every survey collection period, that could be combined to examine trends over time.

Household pulse survey reports



Employment Table 1. Experienced and Expected Loss of Employment Income, by Select Characteristics: Louisiana

Source: U.S. Census Bureau Household Pulse Survey, Week 12.

Total Population 18 Years and Older

	A	B	C	D	E
4	Select characteristics	Total	Experienced loss of employment income since March 13, 2020 (for self or hous member)		
5			Yes	No	Did not report
7	Total	3,431,432	1,865,288	1,550,309	
8	**Age**				
9	18 - 24	256,722	155,151	101,571	
10	25 - 39	1,019,855	600,179	417,200	
11	40 - 54	827,469	577,354	247,042	
12	55 - 64	603,466	279,496	322,633	
13	65 and above	723,921	253,108	461,863	
14	**Sex**				
15	Male	1,622,366	932,670	684,653	
16	Female	1,809,066	932,618	865,656	
17	**Hispanic origin and Race**				
18	Hispanic or Latino (may be of any race)	164,987	78,619	84,792	
19	White alone, not Hispanic	2,094,820	1,108,249	980,314	
20	Black alone, not Hispanic	1,046,743	598,454	441,517	
21	Asian alone, not Hispanic	49,738	25,478	23,029	
22	Two or more races + Other races, not Hispanic	75,144	54,487	20,657	
23	**Education**				
24	Less than high school	351,318	162,860	185,471	
25	High school or GED	1,322,659	842,622	475,974	

US AL AK AZ AR CA CO CT DE DC FL GA HI ID IL IN IA KS KY LA ME MD MA MI MN MS MO MT NE NV NH NJ NM NY NC ND OH
Sheet 20 / 67
PageStyle_LA
STD
Sum=0

Data Cleaning

- Familiarize yourself with the data set.
- Check for structural errors (more on this in a bit).
- Check for data irregularities (use plots, summary statistics, etc.).
- Decide how to deal with missing values (beyond the scope of this class).
- Document data versions and changes made.

Familiarize yourself with the dataset

- Domain knowledge, domain knowledge, domain knowledge! Have a good grasp on what your variables mean, which are important, and which need cleaning.
- Ask as many questions of your collaborator(s) as you need. Do this sooner, rather than later!
- General steps:
- Check the file size.
- Check the dimensions of your dataset.
- Check the first few rows of your dataset and the storage types of the variables to make sure you have read it in correctly.

Check for structural errors

Structural Errors: faulty data types, non-unique ID numbers, mislabeled variables, string inconsistencies, etc. It’s hard to give you a full list, since there are so many ways a dataset can be messy!

General steps

- Check the names of the columns in your dataset and rename them if necessary.
- Check the storage formats of the variables.
- Look for duplicate rows.
- Check variations in a column (e.g., "female", "Female", "F", etc.)

Check for data irregularities

- Data Irregularities: accuracy concerns like invalid values and outliers. Invalid Values: Values that don’t make sense.
- Individual Values: E.g., 19 pounds for a woman over 18, ages under 15 for a dataset of employees
- Across Columns: E.g., Sexual orientation, gender, and household partners
- Outliers:
- Repeat analysis with and without the point, and see if the results change.

Missing Values

- There are two types of missing values:
- Explicitly, i.e. flagged with None, NaN, NA, etc. Strategizing about missing values is beyond the scope of this class.
- Implicitly, i.e. simply not present in the data.
- One bad habit is entering a year in an early row and leaving later rows empty because they should be filled with the value (and just are not, for some reason).
- We like explicit values, a lot! It’s much more helpful to know if something was left missing intentionally.

Don't contaminate a stock solution

If you discover any of these errors in a data file, never modify it directly, but instead write code to correct the value and explain why you made the fix.

You will be forgiven for having two nearly-identical copies of the data.

Creating data

- Be consistent.
- Use a consistent data layout in multiple files (make it tidy!).
- Use consistent codes for categorical variables.
- Use a consistent fixed code for any missing values. Use consistent variable names.
- Use consistent subject identifiers.
- Use consistent file names.
- Use a consistent format for all dates.
- Use consistent phrases in your notes.
- Be careful about extra spaces within cells.

Creating data

- Choose good names for things.
- As a general rule, don’t use spaces, either in variable names or file names
- Where you might uses paces, use underscores or perhaps hyphens.
- Be careful about extraneous spaces at the beginning or end of a variable name. Avoid special characters ($,@,%,#,&,*,(,),!,/, etc.), except for underscores and hyphens.
- The main principle in choosing names, whether for variables or for file names, is short, but meaningful.
- Finally, never include “final” in a file name.

Shoutout to ISO-8601

Write dates like YYYY-MM-DD.

“When entering dates, we strongly recommend using the global ISO 8601 standard, YYYY-MM-DD, such as 2013-02-27.”

If you like, you can read about some horror stories coming from spreadsheet errors. (http://web.archive.org/web/20220810000430/www.eusprig.org/horror-stories.htm)   Many of them involve incorrect date formatting.

PUBLIC SERVICE ANNOUNCEMENT:
OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.
THIS IS *THE* CORRECT WAY TO WRITE NUMERIC DATES:
2013-02-27
THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:
02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013. II. 27. 27/2-13 2013.158904109
MMXIII-II-XXVII MMXIII LVII/CCCLXV 1330300800
((3+3)×(111+1)-1)×3/3-1/3³ 2013
HISSSSS
2-27-13
10/11011/1101 02/27/20/13 01237

Creating data 4

- Create a data dictionary. It is helpful to have a separate file that explains what all of the variables are.
- Such a “data dictionary” might contain:
- The exact variable name as in the data file.
- A version of the variable name that might be used in data visualizations.
- A longer explanation of what the variable means.
- The measurement units.
- Expected minimum and maximum values (helpful for data cleaning or identifying outliers).
- This is part of the metadata that you will want to prepare: information about the data.

Creating data 5

- Don’t leave any cells empty, and use some common code for missing data. Don’t include calculations in the raw data.
- Don’t use font color or highlighting as data.
- Make backups.
- Use data validation to avoid data entry errors. Save the data in plain text (usually, a .csv file).

These tips come from Broman, K. W., & Woo, K. H. (2018). Data organization in spreadsheets. The American Statistician, 72(1), 2-10.