

Overview of Project 1: Data processing

Resources and instructional videos:

Project 1, preparing for analysis; processing a dataset from Qualtrics;

- Already downloaded the data
 - Reproducible research: preserve data at key stages
 - Use sheets in a progression from left to right
 - Objective is to make a system that recodes values and documents decisions
 - Allows you to fix mistakes with minimum new work
 - Allows you to add additional lines of data if necessary with minimum new work
 - Provides documentation to allow others to inspect your decisions
 - Make decisions clearly documented in the process
 - Functions are better than using commands
 - If statements are self documenting; find and replace commands are not
 - This example is 9 years old, and shows some mistakes
 - Mix of upper and lower case
 - Downloaded data has choice text rather than numbers-- more work
 - Choice text is sloppy
-
- **Project 1 Part 1: Reproducible processing in Excel:** this 16 min video follows the outline in the slide above to explain what we are trying to accomplish by building an excel file that processes a raw dataset on the first sheet, to a clean and recoded dataset on the final sheet, while maintaining the in-between steps.
 - **Intro to some basic skills:** the following 4 videos use the same “made-up” dataset. I suggest that if you are new to excel you should create your own version of a simple dataset to use these functions. There is no substitute for hands-on practice.
 - **Fill down and across:** this ~4 min video demonstrates how to fill down and across in excel, and allows us to automatically increment across ranges, for both values and references.
 - **Using basic functions, rand, round, sum, average** This 5 min video demonstrates some basic functions that we will use in various stages of the analysis and recode process.
 - **Reference and recode -99 to NA, basic version** This 7 min video demonstrates how to reference another sheet on the same workbook and recode -99 to NA values. And how to fill that formula across.
 - <https://youtu.be/WfrHOPABBD4>
 - **View and format, paste special=values example** This 5 min video just demonstrates some helpful format and viewing options.
 - **Project 1 specific tasks:**

- [Efficient filling across of corrected search function in excel](#) This 4 min video provides the corrected search function and demonstrates how to quickly fill the corrected versions into our formula page of our processing sheet.
- [Descriptive stats intro and \\$operator](#) This 9 min video shows how to calculate basic descriptive statistics and demonstrates the use of the \$ operator to make cell references absolute.
- **Deliverable #1:** spreadsheet that processes our qualtrics data and results in two versions of the dataset-- one for use in Excel and one for use in R.
 - Ted's copy of [process.2021.welser.xlsx](#)
 - 1 final sheet with excel version: median substituted missing values
 - 1 final R version: rectangular dataset; NA preserved (for replacement in R), but symbols and spaces removed from qualitative variable answers.
- **Deliverable #2:** spreadsheet that starts with a copy of the Excel version of the data; then performs a series of exploratory analyses.
 - Ted's copy of a preliminary start of [explore.2021.welser.xlsx](#) This version just demonstrates a couple of tasks and will require in class explanation. Do not worry if you do not understand yet what I did here.
 - My updated version called. [new.explore.2021.welser.xlsx](#)
- **Deliverable #3:** google doc that follows the general format of a research article, but with greatly simplified sections.
 - See this example ([Template shared in our project 1 folder](#)), and I will demonstrate how to make a copy of this version and edit it into your own version.
 - See this example based on my analysis of our dataset, so far: Welser.P1.2021, this is also shared in our project 1 folder. You can use this file as your starting template or the one above. (my version: [Welser.P1.2021](#))
- [Template for making table 1](#)
- **Slidedeck #1** for start of course:
 - [Data analysis and project 1 slides](#)
- **Some more skill videos in excel**
 - [Making Histograms in Excel](#) Using Data Analysis add on for histograms
 - [Creating groups of variables](#) to study at the same time using correlation
 - Multiple indicators of related attributes
 - Combining indicators into a summed or averaged index
 - Using correlation matrices to identify indicators that should be combined or not.
 - [Creating indices and then using to scatter plots and correlation](#) to assess shape and degree of association.
 - What is the degree of association?
 - What is the shape of association?
 - <https://www.youtube.com/watch?v=XNmdP5vAuQI> how to jitter in excel
 - Jitter makes it so overlapping points are slightly "jittered" left right and up and down, so that they are not overlapping. Here is a video made by a student who was enrolled in my data analysis class 7 years ago.

- **Introduction chapter to functions in excel:**
 - <https://www.dropbox.com/s/4xe1tohztvwsku5/FunctionsIntro.pdf?dl=0>
- *Reproducible research with R and RStudio*, Christopher Gandrud. We will use the preface, and chapters 1 and 2 to
 - Introduction, [pages 1-7](#) The reason for including this section of the chapter is to demonstrate that there are wider opportunities for learning that could make your research process more effective and efficient.

Learning objectives:

- cultivate data management skills and confidence in working with rectangular datasets ranging from small (20 variables by 40 number of cases) to large (300v by 100,000n)
- build work habits that automatically document data management steps and allow reversibility with minimum of lost effort
- build procedural familiarity with excel
- learn and use basic function programming in excel
- learn why and how to replace missing values with
 - means, medians, other values
- create indices that sum or average a set of indicators
- understand why and when to use different weighting rules for indices
- perform and interpret basic descriptive statistics and visualizations in excel
- prepare data sets in excel for use in better analysis programs
 - naming protocols for variables
 - rectangularization
- adopt standard documentation practices for variables, values, and codebooks
- Standardize norms for documenting work to allow replicability by self and others in the future
- Learn and practice common expectations for research report construction
 - clear, efficient prose
 - simple organization
 - standard report sections
 - how tables should look and what should be in them
- Link your group's research page here
 -
- Questions worth asking
 - Sociological
 - Interesting
 - Doable
 -

- Operationalization --> making better variablesexample

Tasks for the Missing values column

is zero a legit value?

is this variable reverse coded?

are there values that need to be recoded

for example "don't know"

Is the min value actually correct > 0 vrs 1

Tasks for project #1

1. Complete our collection of data through online survey. Yay!
 - a. as a class make sure that we get at least 40 participants
2. Dr. Welser makes a final version of our dataset available,
 - a. then download data to excel
 - b. --during computer lab -- save file on flashdrive to bring to class every day.
3. Clean up variable list
 - a. see in class demonstration
4. Rename variables according to helpful conventions
 - i. see examples
 1. Comp.Use.Research
 2. CU.Research
5. Create codebook
 - i. see example in excel sheet
6. Recode data

- a. learn to use functions
 - b. if statements with multiple parts
 - c. fill across
 - d. fill down
- 7. Clean data by solving missing cases
 - a. substituting values (means, medians, zeros)
 - i. Discuss in class: Why substitute something for a missing value?
 - ii. [Substituting mean or median](#) Video
 - iii. Which central tendency makes most sense?
 - 1. Mean
 - 2. Median
 - 3. Mode
 - iv. Fancier methods
 - 1. see discussion in class, but see also,
 - 2. multiple imputation for bigger and more important datasets
 - a. mice
 - b. documentation of simple substitutions in our dataset
 - i. create flag variable to note which cases have values substituted
- 8. Basic descriptive statistics in Excel--- strengths, weaknesses, hassles
- 9. Open content questions-- what should you do with them? [2021-- no open items]
 - a. find; count if; hand code
 - b. See formula in function exchange
 - c. What attributes should we measure by coding instances of key terms
 - d. Make at least 2 new variables measured from the qualitative responses
- 10. wordl; interesting but not especially helpful
 - a. still worth trying; tune to make image more informative
- 11. create indices; sums, averages.
 - a. [Constructing an index](#) a practical example, Video
 - b.
- 12. What are some descriptive questions worth asking?
 - a. ask two of them
- 13. Univariate distributions
 - a. histograms
 - b. [making histograms and scatter plots](#) Video
- 14. Bivariate relationships
 - a. scatter plots
 - b. correlations
 - c. try jittering points that otherwise overlap
 - i. here is the syntax $=(\text{RAND}()-0.5)*\text{AF\$1}$
 - ii. where "AF\$1" points at a weight term
 - iii. keep the weight term a small fraction of the magnitude of variation in your variables

Deliverables:

Links to shared files with:

1. processed data set in excel (with multiple sheets including)
 - a. raw data
 - b. recode formulas
 - c. values from recodes
 - d. codebook
 - e. ~~pre-processing of qualitative item~~ (insufficient data to perform this)
 - f. sheet formatted for R input
 - g. [example!](#)
2. From “f” above, single sheet of dataset saved as CSV file
3. Excel sheet with evidence of descriptive analysis work
 - a. should be much more extensive than final report
 - b. this sheet can be included in 1 above
4. Typed report on a google doc
 - a. consisting of 600-1000 words
 - b. tables, graphs as necessary
 - i. Describe sample
 - ii. Descriptive stats of key variables
 - iii. Histograms making a comparison
 - iv. Scatterplots showing a relationship
 - c. 2-3 notable findings worth mentioning
 - i. The more notable and interesting, the better
 - ii. The more clearly presented, the better
 - d. your assessment of strengths and weaknesses of using excel for the tasks in this assignment
 - i. This should be a concluding section of your report

Details on the report:

Example

How to video for stuff from the intro to the homework:

<http://www.youtube.com/watch?v=AlgiT4PwBr0>

Note: [see additional resources page](#) (need to be logged into google account) for example datasets and additional readings.