Preparing Secondary Datasets for Analysis

for class projects using SPSS

#	Description of Step	SPSS Syntax
1	Create a **project folder** and set it as your **Working Directory.** Gather and examine the documentation to identify what each case represents and how data was collected.	CD '*path to your project folder*'. ** CD = Change Directory ex CD 'E:/my_class/my_project/'.
2	Identify **variables of interest**. Select up to twice the number required as some may be problematic Create a new data file with only your selected variables, plus the unique ID variable.	SAVE OUTFILE='my_data.sav' /KEEP= *id_var var1 var2 var3 var4 var5.*
3	Use descriptive statistics, value labels, and the codebook to classify each **variable** as categorical or numeric and find the meaning of each **value**. If most values are unlabeled, it is probably numeric.	FREQUENCIES *var1 var2 var3 var4 var5*. FREQUENCIES *categorical* / BARCHART. FREQUENCIES *numeric* / HISTOGRAM. DESCRIPTIVES *numeric*.
4	Determine if any values mean **non-answers** like “Not Applicable” or “Refused””. Check the **smallest** (negative) or **largest** (9. 99, or 999) values. Make sure these are **treated as missing** by the software.	_**In SPSS frequency tables, values treated as missing are at the bottom next to ‘Missing’, and not included in the column “Valid %”._ MISSING VALUES *var* (7,9).
5	Re-run descriptive statistics. **Drop variables** if they: a. have many more missing values than others b. are categorical with over 90% of cases in one group c. have nonsensical values or response patterns	FREQUENCIES *var1 var2 var3 var4 var5*. SAVE OUTFILE='my_data.sav' /DROP = *problem_vars.*
6	See how many cases have **no missing values**: preferably 20 for each variable (5 vars→100 cases) and representative of the original cases. Drop variables with missing values and/or keep only cases with none, so analyses have the same n.	COMPUTE **miss** = NMISS(*var1, var2, etc)*. CROSSTABS *varlist* BY **miss** /CELLS=COUNT ROW BPROP. SELECT IF **miss = 0.**
7	Use frequencies from Step 3 to identify **ordinal** vars and numeric vars with 7 or less values. Carefully consider whether to treat as ordinal (if appropriate analyses were taught), numeric (if mean and sd make sense), or else categorical.	_** If an ordinal variable is not linearly related to other values, try treating as nominal._
8	Look at descriptive statistics and histograms for **numeric variables** and consider whether to **group values** if the distribution is neither normal nor flat. Numeric vars store more information, but may not best represent the responses.	DESCRIPTIVES *numeric*. RECODE *var* (1 THRU 10 =1)(11 THRU 20=2) INTO **in2**. VALUE LABELS **in2** 1 “Low” 2 “High”.
9	Look at frequencies for **categorical variables** and consider whether to **combine groups**. Have 2 to 5 groups, none with less than 10% of cases. If appropriate, compare one group to “all others” (recode to 1 vs 0).	RECODE *var* (1 2 **= 1**)(3 4 5 **= 2**) INTO **in2**. VALUE LABELS **in2** **1** ‘Low’ **2** ‘High’. COMPUTE **is3** = (*var* **= 3**). VALUE LABELS **is3** **1** ‘Group 3’ **0** ‘Others’.
10	Before running your final analysis, examine the relationship between each pair of vars (bivariate). The choice for analysis depends on whether your X and Y is categorical or numeric and which is the predictor vs response.	CROSSTABS *categorical* BY *categorical* /CELLS=COUNT ROW BPROP. MEANS *numeric* BY *categorical.* GRAPH SCATTER *numeric* WITH *numeric* .
		EXECUTE.