Preparing Secondary Datasets for Analysis

for class projects using STATA

#	Description of Step	STATA 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.	keep *id_var var1 var2 var3 var4 var5* compress save *my_data*
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.	tab1 *var1 var2 var3 var4 var5* tabulate *categorical*, plot histogram *numeric* summarize *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 Stata, missing values are (or start with) periods and are hidden unless requested.* mvdecode *var1 var2,* mv(*8 9*) tab1 *var1 var2*, missing
5	Re-run descriptive statistics. **Drop variables** if they: a. have far fewer Obs than others (i.e., more missing values) b. are categorical with over 90% of cases in one group c. have nonsensical values or response patterns	codebook, compact *** Obs = valid values* drop *problem_vars* save *my_data*, replace
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.	misstable pattern egen **miss** = rowmiss(*) keep 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.	tabstat *var*, stat(mean sd med min max n) recode *var* (1/10=**1** Lo) (11/20=**2** Hi), g(**in2**)
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** 'Lo') (3 4 5=**2** 'Hi'), g(**in2***)* generate **is3** = (*var* == **3**) label define **is3** **1** 'Group 3' **0** 'Others' label values **is3** **is3**
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.	tab *categorical categorical*, row chi2 tab *categorical*, sum(*numeric*) scatter *numeric numeric* help fvvarlist // *info to use cat vars in regress*