*Learning Stata
*Using LSE online tutorial
http://www.lse.ac.uk/methodology/tutorials/Stata/typingInData.aspx
*All data is based on the file "MI_Statatutorial.dta" - this can be found at the above website

*Lesson One: Renaming Variables, Labels
*Open "Data Editor"
*Type in Data (var1, var2, var3)
rename var1 name //changes variable from 'var1' to 'name'
rename var2 gender // changes variable from 'var2' to 'gender'
rename var2 age // changes variable from 'var3' to 'age'
label variable age "age in years" // this command adds a label to the variable
*creating dummy variables and labeling them: (this is a two step process)
*Step One:
label define genderlabels 0 "female" 1 "male" // creates temporary label is the definition of the dummy variable
*Step Two:
label value gender genderlabels // associates the temporary definition with the variable that we want
*in this case the definition 0 = female 1 = male is now associated with the variable 'gender'
Browse // the variable 'gender' instead of containing zeros and ones now has female and male

*Lesson Two: The Do-File
*Can highlight specific do-commands and put run, this will tell stata to run only that command

*Lesson Three: Creating a Grouped Variable
1/5 //will create a grouping of 1 through 5
max // is the maximum value of a variable
min // is the minimum value of a variable
recode age (min/30=1) (31/45=2) (46/64=3) (65/max=4), gen(age_group) // this creates categories for ages
*in other words, the lowest age through 30 will be categorized as group 1... and so on
*gen(age_group) creates a new variable so that we do not mess witht the original variable

*Lesson Four: Mathematically Transforming a Variable
*let's say we want to have the log of one of our variables -
gen (educ_age2) = log(educ_age) // this creates a new variable "educ_age2" which will be the logarithmic value of the original variable "educ_age"

*Lesson Five: How to select sampling cases based on user-defined Criteria
*In other words, how to select the data we want from a larger data set
*Let's say we are interested in the number of people who have responded to the question "have you ever signed a petition"

tabulate petition // stata gives us a frequency table with those who responded yes and those that responded no to the question
*Now, we are interested in the respondents who are 30 years or younger
tabulate petition if age <=30 // now we have a frequency table that only records the responses of those 30 or younger
*We can also permanently drop data from the overall data set
drop if age >30 // this will remove all data points that are older that 30 years of age

*Lesson Six: Describing a Continuous Variable
*variable of interest will be "educ_age"
summarize educ_age, detail
*smallest = four smallest values for the variable
*50th percentile is the median (in this case the median age is 16)
Histogram educ_age // gives graphical representation of the variable values

*Lesson Seven: Describing a Categorical Variable
*Variable of interest will be "energy"
*Respondents are asked to respond on a four-point scale
Tabulate energy // creates frequency table
*to graphically represent this we want to create a bar graph
*Bar graphs are a two step process:
*Step One:
generate tempvar=1 // this creates a holding variable where all values are equal to one
*Step Two:
graph bar (count) tempvar, over(energy) // creates a bar graph with an overlay of the variable of interest

*Lesson Eight: 2- and 3- Way Contingency Tables
*"pol_interest_2cat" as explantory variable and "demo" as response variable
tabulate demo pol_interest_2cat, chi column
*the general formula is tabulate y-variable x-variable, chi column
*column gives the percentages
*The bottom of the table gives our chi2 statistic in this case is it 15.5. Next to the chi2 statistic is the probability value (P = 0.000)
*Since our chi2 statistic is large and Pr-value is nearly zero we can reject this null hypothesis
*in this case the null hypothese would be: there is NO relationship in the population. By rejecting this we are saying there is a relationship in the population
*Now let's say we want to know if there is a difference between men and women in the population
codebook gender // this tells us how the variable 'gender' is coded - in this case 0 = female 1 = male
tabulate demo pol_interest_2cat if gender==0, chi column // this will give contingency table data for female respondents

tabulate demo pol_interest_2cat if gender==1, chi column // this will give contingency table data for male respondents

*Lesson Nine: Correlation Coefficients
*How to find correlation metric for continuous variables
*variables of interest: 'tax' 'resources' 'age' 'income'
correlate tax resources age income income

*Lesson Ten: Drawing a Scatterplot
*looking at age and income
scatter income age // gives a scatter plot
*there are two ways to add a line of best fit
twoway(scatter income age)(lfit income age)
scatter income age || lfit income age

*Lesson Eleven: Box-plot and Stem-and-leaf Plots
*vol_groups and petition
graph box vol_group, over(petition)
stem vol_groups if petition==0
stem vol_groups if petition==1

*Lesson Twelve: Independent Samples t-test
*How to compare the means of two groups (continuous response variable, binary explanatory variable)
*DV = 'resources' grouped by 'gender' looking to see if there is a difference between males and females with regards to the 'resources' question
*command in stata is: ttest dv, by(iv)
ttest resources, by(gender) // the data output gives both the null and alternative hypothesis. The null is that there is no difference between male and female response to the question "we are using up the earth's resources too quickly"

*Lesson Thirteen: Simple Linear Regression
*How to test the linear association between two continuous variables (continuous response variable, continuous explanatory variable)
*y-variable = resources x-variable = age
*command in stata: regress y-variable x-variable
regress resources age
*R-squared: the percentage of variance which is explained by our response (y) variable
*We now want to know if there is a linear relationship between the y and x variable
*to do this we conduct a hypothesis test: the null = there is no linear relation between resources and age in the population
*in other words the beta-coefficient for age would be equal to zero
*to check our null we conduct a t-test (which was already given in the regression printout) the t-value is 2.61 and the p-value is 0.01 or near zero, so we can reject the null
*further, the beta-coefficient is positive meaning there is a positive linear relation between the two variables

*We can also say: for an expected increase in age of one year there is an expected increase in the value of resources by 0.025

*Lesson Fourteen: Multiple Regression
*How to test the linear association between a continuous response variable and more than one explanatory variable (continuous response variable, explanatory variables various levels of measurement)
*y=resources, x1=age, x2=gender, x3=income
regress resources age gender income
*R-squared says that 13.66% of the y-variable variance is explained by the regression model
*We do the same test as in a bivariate regression, create a series of hypotheses
*1 - there is no linear relation between resources and age, controlling for income and gender
*2 - there is no linear relation between resources and income, controlling for age and gender
*The variable gender is a dummy variable so our hypothesis will needs to be phrased differently
*3 - There is no difference in the population between males and females in theor average values of the variable resources
*now we can run the ttest and corresponding p-values:
*income has a rather high p-value so we fail to reject the null
*age proves to have a low p-value, we can reject the null, there is a liner relationship. Further the beta-coeff is positive so there is a positive linear relationship between age and resources
*gender p-value shows that we can reject the null, so there is a difference between male and female response in the population, in this case on average the value of the variable resources is 1.5 points lower for males than for females

*Lesson Fifteen: Binary Logistic Regression
*How to test the association between a binary response variable and explanatory variables with various levels of measurement
*In this case the response variable (y) is a binary variable (aka a dummy variable)
*y=petition, x1=gender, x2=age, x3=vote
logit petition gender age vote
*if we wanted odds ratios instead of coefficients the commands is
logistic petition gender age vote
*This will probably be covered with Brandon in 8102

*Lesson Sixteen: Contour Plots in Stata
clear
*click file --> Example Datasets --> Stata 12 Manual Datasets --> Graphics Reference Manual --> sandstone.dta (click 'use')
*command formula is twoway contour z x y
twoway contour depth northing easting