Assignment 2:
Working with statistical software (SPSS: Statistical Package for the Social Sciences)

Sociology 2206A -573 Fall 2019
Professor Don Kerr
Worth 15% of final grade (late penalty 10% of assignment grade per day)
Due December 4th, 2019 at 8:30am (at the beginning of class)

****

IMPORTANT!

DO NOT PUT THIS OFF UNTIL THE LAST MINUTE!!!

SPSS consultant to help YOU!!!

There is an SPSS consultant (David Bell) who is there specifically to help you (and students in other classes with similar assignments)! Please don't email myself or David Bell about SPSS difficulties – go to the lab in person during consulting hours.

David Bell
Wemple computer lab: W045
Wed Nov 13th 12-2, 7-10 p.m.
Thu Nov 14th, Fri Nov 15th, and Sat Nov 16th are all 1-6 p.m.
Tue Nov 19th 1-6 p.m.
Wed Nov 20th 12-2 p.m.
Thu Nov 21st & Fri Nov 22nd 1-6 p.m.
Sat Nov 23rd 1-6, 7-9 p.m.
There will also be hours available Tue Nov 26th and Wed Nov 27th, and potentially beyond, depending upon David Bell's availability (TBA)

****

Introduction:

The ability to work with SPSS (and other software packages) is a fundamental skill for sociologists and necessary in completing many of the assignments in more advanced courses in methods and statistics in Sociology. For this reason, we will be spending some time in the computing lab familiarizing ourselves with this software.

All of the computers in the computing lab (see details below) have an up to date version of SPSS (Statistical Package for the Social Sciences). You can also obtain a "SPSS Student version for Windows" from the University Computer Store" to install on your home computer (although I don't recommend it). The major disadvantage of the student version is that it does not allow you to easily work with the "syntax" language that we will be using in this course, nor does it permit you to work with more than 1,500 cases or 50 variables. This is a major limitation and subsequently, I recommend that you work with the version of SPSS as available in the computer

labs. In addition, there is a consultant (David Bell) available when working with SPSS in our computer labs (schedule is indicated earlier).

Remote Access: *While you can complete this assignment in the computing lab*, you do have an alternative if you feel quite confident working with computer software. Information Technology Services (ITS) at King's will now permit sociology students enrolled in Soc 2206 to have remote access to SPSS from their home computer. Subsequently, anyone with an internet connection at home should now be able to work with this software (24 hours a day, 7 days a week). In addition, you will have access many MS Office applications. Once you gain access to this network, each user can also store data securely in the 'My Documents' area of their desktop. For instructions, see below link:

https://www.kings.uwo.ca/its/support/remote-application-server-access/

The purpose of this assignment is to introduce you to this software and to some rather elementary data manipulations and statistical computations that are possible using SPSS. SPSS is probably the most widely used statistical package in sociology departments across Canada, largely due to its user friendly character. Once you become proficient on SPSS, you should not have too many difficulties in moving on to more powerful and complex statistical software, such as SAS (widely used outside of academia) or Stata (widely used by social scientists interested in apply more advanced statistical procedures). There are innumerable software packages used in neighboring social sciences. For example, the equivalent to SPSS in geography is GIS (Geographic Information Systems) which is particularly useful in manipulating data for various geographic units and mapping datasets.

Many of the examples provided in class (Soc 2205/2206) or in the textbook involve relatively small samples (or few cases) in the explanation of some basic statistical procedures. Yet obviously in reality, much social research involves virtually 1000's of cases. Consider a national survey of 20,000 persons, involving the collection of detailed information on a wide range of variables. It is clearly not feasible to analyze such information with a hand calculator; hence the utility of software such as SPSS. As an alternate, if you prefer to do the current assignment by hand calculator, you can (good luck, you'll need several years: late penalties apply).

For the purpose of the current introduction, I have selected a large dataset: Canada's NLSCY (National Longitudinal Survey of Children and Youth). You will also be working with a small dataset that I have created using the questionnaires completed in a previous year by students in my methods courses). Although the latter dataset is not based on a "probability" sampling plan, for the purposes of the current assignment, we will be treating it as though it were based on an appropriate sampling strategy.

There are several Parts (**1A, 1B, 2, 3A, 3B, 3C**) to this assignment. **For each part, you will be handing in a syntax file, an output file and a brief write-up. Please hand in all syntax together, all outputs together, and your write-up all together (probably a couple of double spaced pages all together). In other words you will be handing in 3 separate piles of paper with a cover page for each (syntax, output, write-up).**

It is strongly recommended that you begin the actual computing portion of this assignment in very short order – learning new software and new techniques can be time consuming, and problems can be unpredictable …

While I have included very detailed instructions here on how to use various features of SPSS, you may want to read ahead in your text. The Appendix of your textbook (or Chapter 13 if you are working with the 4th Canadian edition) includes a discussion on how to use SPSS as well. Further, unless you have already taken your 2205 class (only a few of you have), I highly recommend that you read Chapter 8 on Quantitative Data Analysis to complete the write-up portion of this assignment (or again, Chapter 13 in the 4th edition), as it provides an overview of the basic descriptive measures that can be used in SPSS (such as frequency distributions, means, contingency tables/crosstabs etc.).

How to access SPSS:
There are two ways that you can access the software you need to do this assignment. You can go to the student computing lab, or you can use remote access from your home computer. Each has plusses and minuses.

Using the student computing lab:
All of the computers in the lab have SPSS installed on them, and you can print there as well (for a price). This is also where you can find the SPSS consultant ('where to get help'), and where you might be able to help each other (you are *expected to do your own work*, but it is sometimes helpful to try to figure it all out with your fellow students, or get/give moral support).

The student lab is located in the basement of Wemple building (W045).
The lab is accessible *24 hours a day* with the help of your student card. You can access the lab pretty much anytime.

Where to get help:

Please don't email myself or David Bell about SPSS difficulties – go to the lab in person during consulting hours. The hours are listed on the first page of this assignment.

Using Remote Access:
I have signed the entire class up to be able to access SPSS (and some other basic programs like MS office) through a proxy from your home computer. Please click on the link below for instructions on how to access the system.

https://www.kings.uwo.ca/its/support/remote-application-server-access/

The instructions are all included in the link. You will need your uwo user name and password to enter the system (the same as your email etc.) **It is strongly recommended that you try to proxy into the system immediately** (even if you aren't sure you want to do it from home) so that we can fix any problems that might arise in a timely fashion.

Again, you should also be aware that there is a student version of SPSS available at the UWO bookstore, but you cannot use that software for this assignment! It doesn't allow you to work with enough cases, or enough variables – it isn't big enough to load the data I'm asking you to work with. You might also be able to pirate this off the internet, but it would be useless to you.

How to access your data:

The two datasets are available not just to you, but to students in other similar courses here at King's. **To find the data, go to the SPSS data folder once you've logged in via Scotty, and look for the folder that includes datasets for Soc 2206 sections 570 and 573.**

In there you should find a file called: **STUDENTSDATA.sav**, and another called: **nlscy2019data.sav.**

Note that you must have access to SPSS in order to open them (you must be in the computer lab, or you must access them through your proxy). If you click on the file, it should open SPSS automatically. Alternatively, you can open SPSS first (it will open to what looks like an empty spreadsheet), then click on the appropriate file.

****

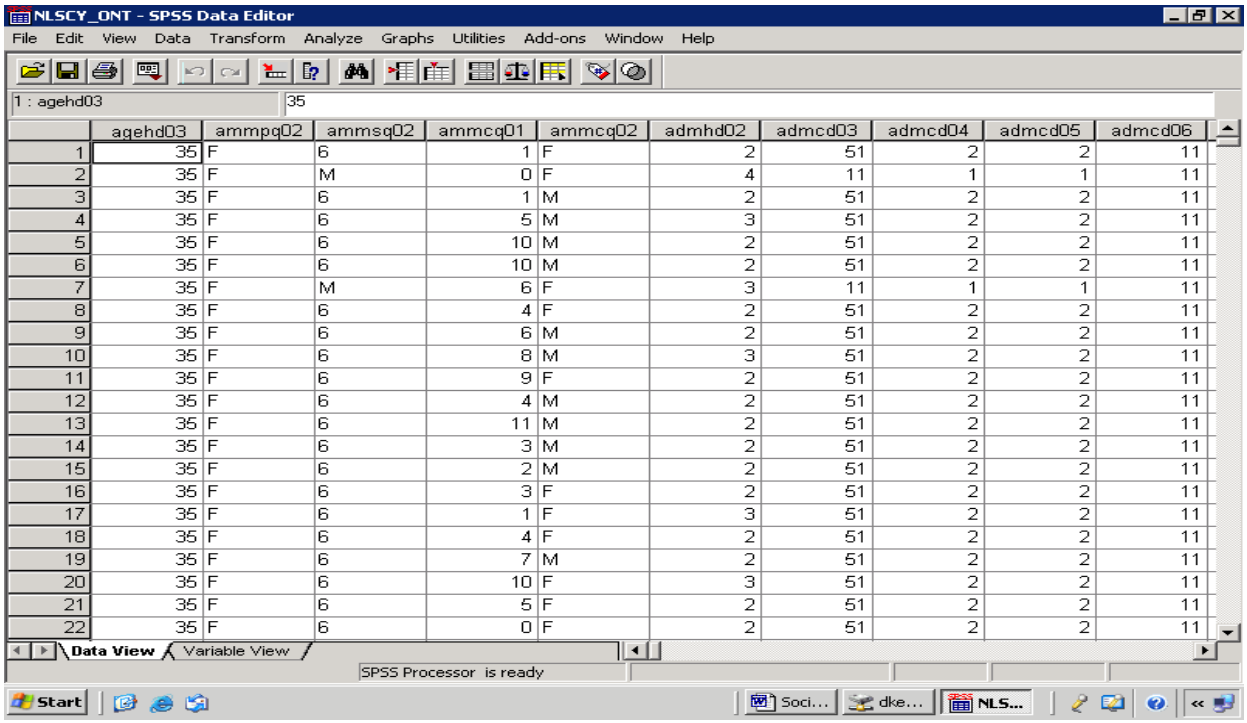
Part 1: Working with SPSS Files

SPSS files
There are 3 types of windows that SPSS handles in order to create 3 different types of files.

1. **Data files** (*.sav) contain the data that any commands will manipulate and analyze. A data file must be open in order to perform an analysis.
2. **Output files** (*.spo) contain the output produced by SPSS, including any graphs, tables, or numbers. Results shown in output windows can generally be copied and pasted in to word processing documents.
3. **Syntax files** (*.sps) contain programs that can be run on SPSS, in its own programming language. These are the files that you must run on the relevant data files in order to get the respective output files.

Each type of file opens in its own specific type of window.

Once you have opened a dataset (*.sav), your screen should show the SPSS data editor, with all the appropriate variables and cases, as follows:



	agehd03	ammpq02	ammsq02	ammcq01	ammcq02	admhd02	admcd03	admcd04	admcd05	admcd06
1	35	F	6	1	F	2	51	2	2	11
2	35	F	M	0	F	4	11	1	1	11
3	35	F	6	1	M	2	51	2	2	11
4	35	F	6	5	M	3	51	2	2	11
5	35	F	6	10	M	2	51	2	2	11
6	35	F	6	10	M	2	51	2	2	11
7	35	F	M	6	F	3	11	1	1	11
8	35	F	6	4	F	2	51	2	2	11
9	35	F	6	6	M	2	51	2	2	11
10	35	F	6	8	M	3	51	2	2	11
11	35	F	6	9	F	2	51	2	2	11
12	35	F	6	4	M	2	51	2	2	11
13	35	F	6	11	M	2	51	2	2	11
14	35	F	6	3	M	2	51	2	2	11
15	35	F	6	2	M	2	51	2	2	11
16	35	F	6	3	F	2	51	2	2	11
17	35	F	6	1	F	3	51	2	2	11
18	35	F	6	4	F	2	51	2	2	11
19	35	F	6	7	M	2	51	2	2	11
20	35	F	6	10	F	3	51	2	2	11
21	35	F	6	5	F	2	51	2	2	11
22	35	F	6	0	F	2	51	2	2	11

Here we have the contents from the NLSCY in a spreadsheet format. A lot of work has already gone into setting this dataset up for you. Here there are responses for over 22,000 individuals across about 700 variables. Across the top of the dataset you will see the assigned variable names that SPSS uses in reading this data. Variables are called things like *agehd03, ammpq02, etc*. If you move your arrow with your mouse across the variables names, it is possible to see

the full name of each. You can also switch to 'variable view' (bottom left of the screen) for more details on the variables, rather than the scores for each case.

For your information, you can find documentation on the NLSCY at the following address:

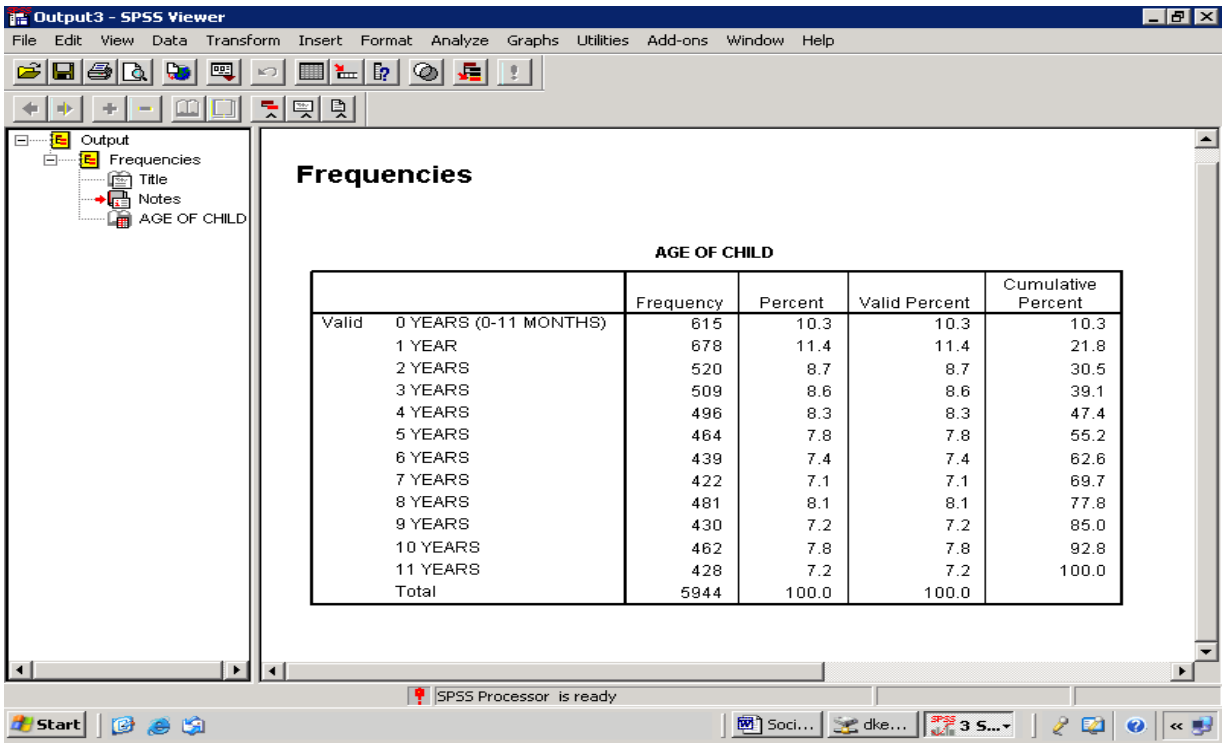
https://search1.odesi.ca/#/details?uri=%2Fodesi%2Fnlscy-89M0015-E-1994-1995-c-1-r-2-primary-file.xml

Here you can find the codebooks and a description of the dataset. Optionally, you can click on UTILITIES then VARIABLES then the variable you are interested in – if you require details on any single variable. Note: this UTILITIES feature has not been fully set up for the second dataset STUDENTS4.sav, so you will need to use the codebook attached to the end of this assignment outline for coding information.

Whereas each column in this dataset represents a variable, each line of this dataset represents a specific *case*. Our *unit of analysis* in working with this dataset is the *individual*, with each row representing the responses across variables for one respondent to the NLSCY. Theoretically, it is possible to make changes on any entry with the SPSS data editor in this spreadsheet (yet obviously, we should not be doing this unless we have a very good reason).

Output Files

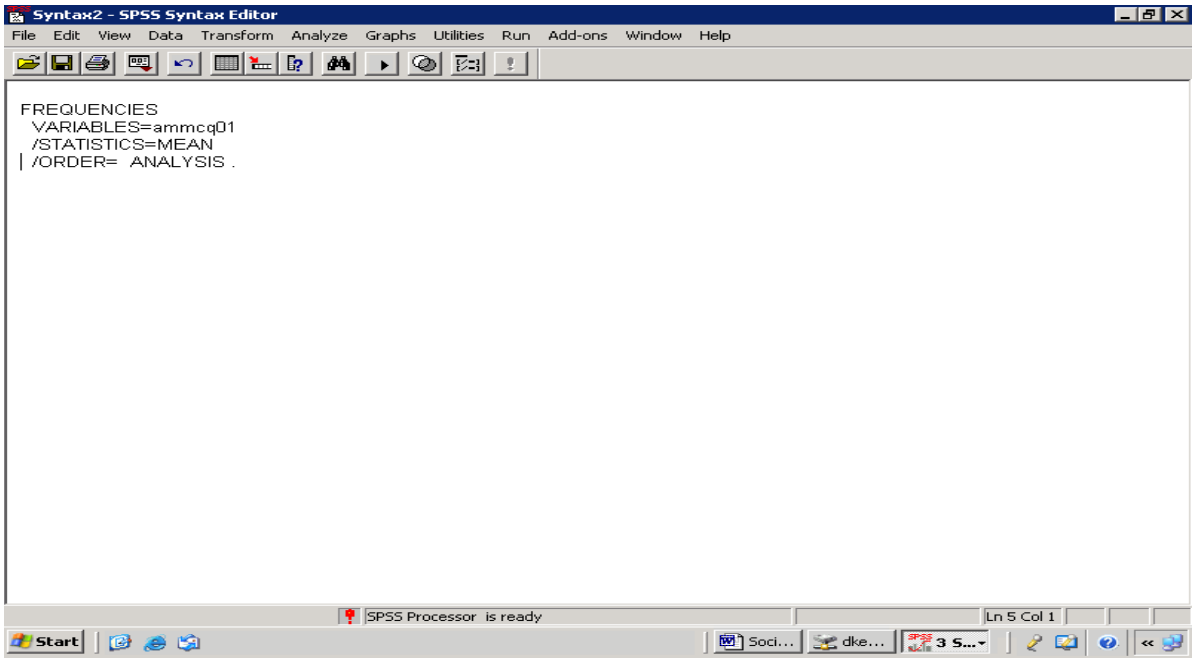
The output window (*.spo) looks like:



This output file *.spo gives us a frequency distribution on the fourth variable in our data set *age of child* (*ammcq01*). Note that in this example this variable has only 5944 cases with no missing values but in your dataset there may in fact be more cases This frequency distribution was run

exclusively on Ontario residents, and for this reason, is not identical to your dataset. In completing your assignments, you will be working with the full sample (except in Part 3 where you will be asked to choose a subsample) and be regularly printing up these output files. I will ask you to provide these when documenting your work.

Syntax Files
A syntax file looks like:



This syntax file runs a simple frequency distribution on the variable age of child and asks the computer to calculate the *mean* on this variable. At one time, the only way to run SPSS was in creating syntax files like this one. Now there are point and click options available that fill in the syntax for you. Each type of file can be saved using the file menu in Windows.

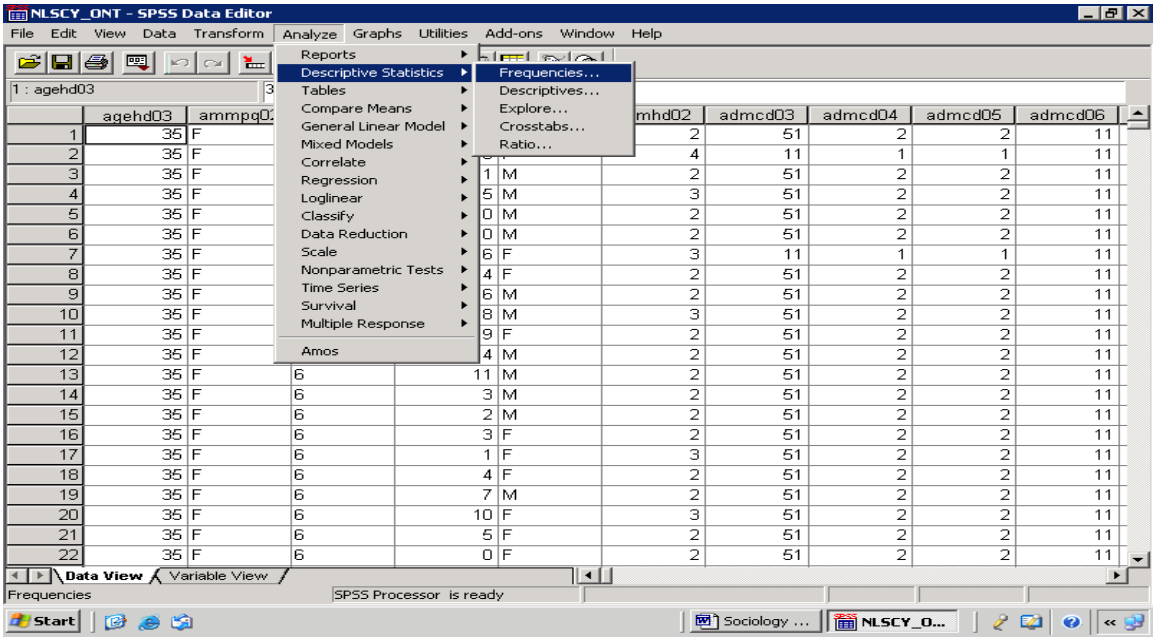
Syntax and the Menu System
There are two ways to execute a command in SPSS. On one hand, you can use the point and click Windows interface, and select the options you desire. This can easily be done from the data window. Unfortunately, if you use this option and fail to *paste* into the syntax file and run your program from there, it is easily possible to alter the data and to lose track of what you have actually done. You can also type directly into the syntax window. Either way, you need to make sure that your instructions end up in the syntax file, and that you run them from there .This way a program can be run over and over again, and you have a record of the analysis or data manipulations you have performed.

To run a piece of syntax in the syntax window, highlight it and select *run… selection*. For help writing a program in SPSS syntax, you can look at the *Syntax Guide* under the *help menu*.

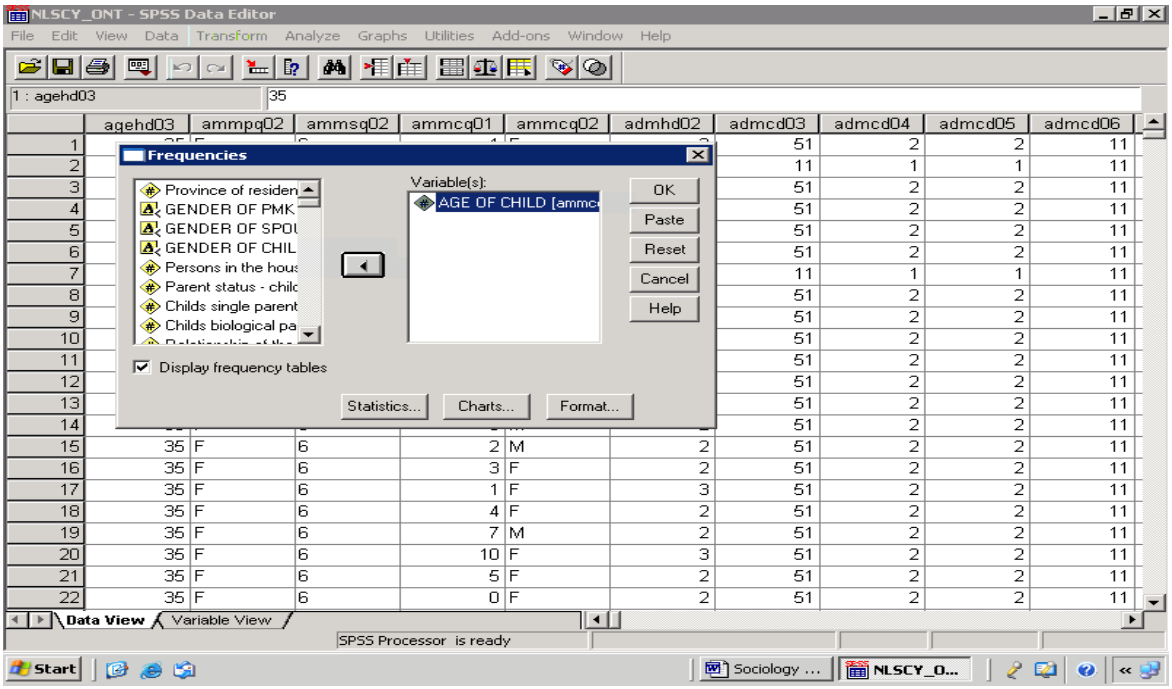
Obtaining Descriptive Statistics in SPSS

Frequencies

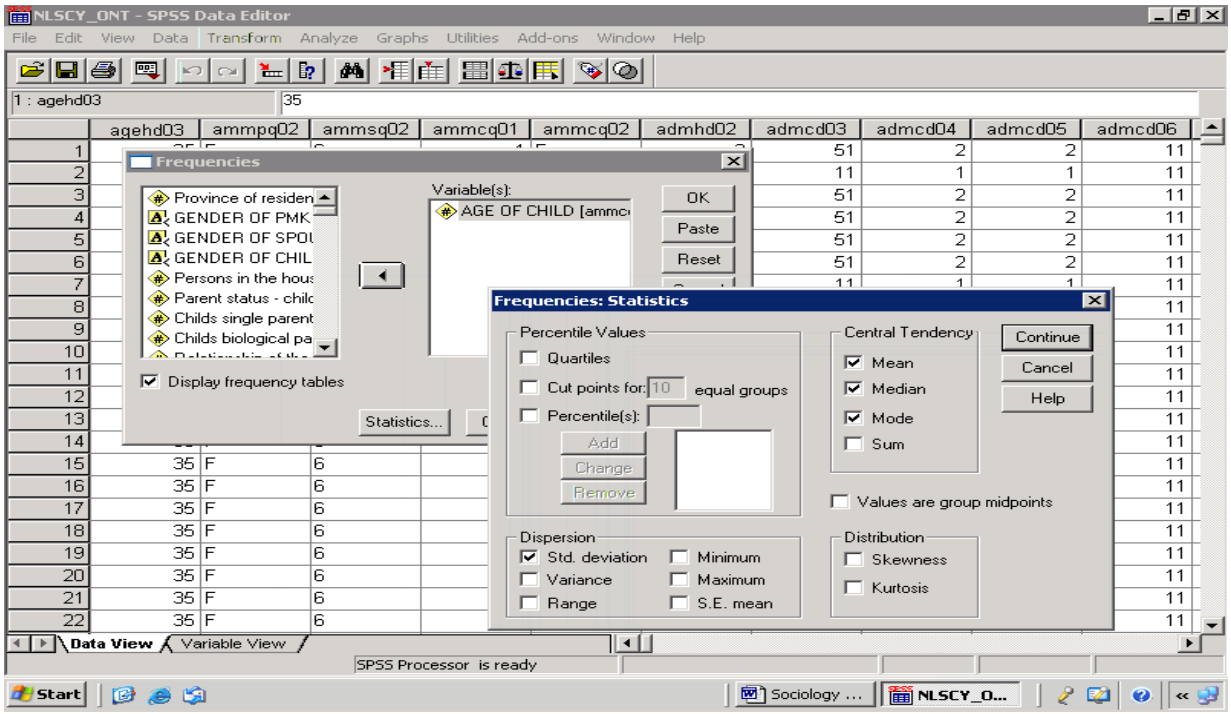
You can obtain a frequency distribution in different ways. In the menu system, you merely follow the hits: *analyze, descriptive statistics, frequencies*. For example in creating a frequency distribution and histogram for *ammcq01*:



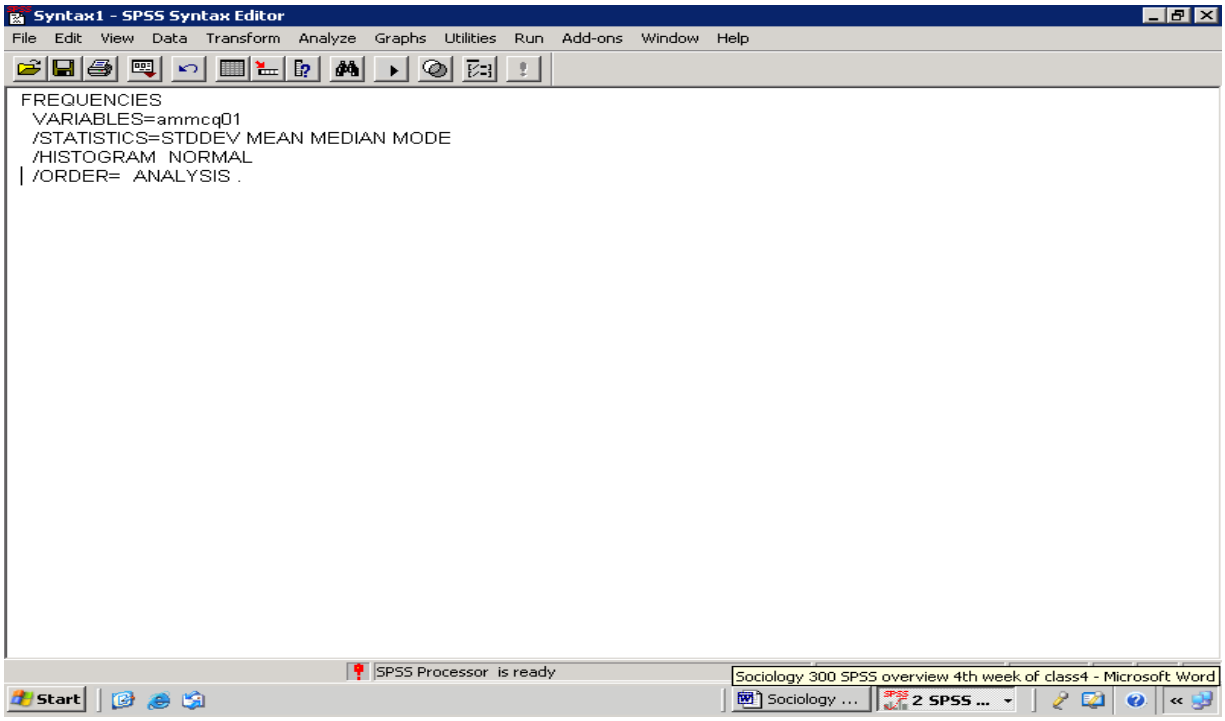
Specific variables or sets of variables can be moved over by merely highlighting the variable of interest and clicking on the arrow key. For example, the next figure demonstrates how we have moved over the variable of interest *ammcq01*.



By clicking on *Statistics* you can select whatever descriptive statistics you want (mean, mode, standard deviation, etc). If you click on *Charts*, you can specify that you want a histogram, etc.



By clicking on *paste* instead of *OK* you can create a SYNTAX file that you can work with:



If you then highlight and run (right click or the arrow button) these commands (this syntax), the software will produce a frequency distribution, standard deviation, median, mode and a histogram.

Descriptives

The menu commands *analyze, descriptive statistics, descriptives* will produce these same descriptive statistics, but not the frequency distribution or graphs. You must specify the statistics that you want under the *options* in the *descriptives* window.

The syntax:

```
DESCRIPTIVES
  VARIABLES= ammcq01
  /STATISTICS=MEAN STDDEV MIN MAX SEMEAM.
```

will produce the mean, standard deviation, minimum and maximum values and standard error for the variable *ammcq01*

Documenting your work

First, you should always use and save syntax files, even though it is possible to work without them. This allows you to go back to it at a later point in time, if need be, to make minor modifications to your work, and many researchers keep only their syntax files (rather than outputs) over the long term, because they can go back at any time and rerun or change things.

In the example syntax file below, I've specified a TITLE for documentation purposes. I've specified the date the program was last modified, the name of the program file (assign1.sps) as

well as the person who developed the program. **You must type this TITLE command directly into the syntax file (make a new syntax file, and then enter your TITLE command) a the top of the syntax file.** You should put it at the top of the file, before your other commands. You then select it and run it like any other command. This title is then found at the top of the resultant output file.

```
TITLE   November 20th  assign1.sps, D. Kerr.

EXAMINE
  VARIABLES=cmmcq01 BY cmmcq02r
  /PLOT BOXPLOT STEMLEAF HISTOGRAM
  /COMPARE GROUP
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

You can theoretically save your output file (in the "*my documents"* folder set aside for you) under whatever name you consider appropriate (for example: assign1.spo). You should also always save your syntax file which in this case was called assign1.sps. By properly documenting your work, you will have a good record of what you have done in the past, just in case you wanted to work with it again.

N.B. You can e-mail files to yourself as an attachment using UWO mail. But prior to doing this, it is necessary to convert your files into files that can be printed up using Word (or some other text editor). While in SPSS, go to "File>Export"…. In the "File Type" box, choose "Word/RTF" file. Click "Browse, go to "Save in" and choose "My Documents" to save all of your work.

You can now e-mail these files to your home computer by using explorer and your UWO mail account (merely attach the appropriate files to your e-mail). If you are having trouble doing this, see your SPSS consultant first.

You should be doing this with your final syntax files and output files. If you save stuff on the computer in the lab outside of the my documents space that has been allocated for you, they might not be there the next time you check (i.e. these computers are regularly cleaned up).

You can cut and paste from the word version, back into a syntax file in SPSS.

It is strongly recommended (no matter what your access or saving strategy) to save regularly and print each step as soon as you have completed it – lost work and entire lost files can be a frustrating part of trying to learn this and similar software if you aren't proactive and extremely careful.

****

Part 1A Requirements

Using the following data file as found in the SPSS data folder for this course: nlscy2019data.sav

Step 1: Create syntax and output files file1a.sps and file1a.spo.

This involves:

In your syntax file, first create a TITLE for your output that includes the syntax file name, date and your name. Do do so, you must go to File > New > Syntax. Once the syntax file opens up, type in the TITLE command at the top of your file, and then provide the corresponding information.

Next, run a FREQUENCY DISTRIBUTION on the two separate variables *adpps01* (depression score) and *alfpd02* (number of hours the parent most knowledgeable, PMK, works).
Then run the DESCRIPTIVES command on these two variable, including the mean, standard deviation, range and minimum and maximum values (note: while the FREQUENCY DISTRIBUTION potentially allows for the option of asking for these statistics (mean, standard deviation, etc), the DESCRIPTIVES COMMAND can potentially do so without requesting the frequency distribution (as you can well imagine, this be potentially useful with continuous variables that have far too many potential response categories (e.g. income as reported in dollars, or weight as reported in pounds).

In this same syntax file,

Step 2: Briefly interpret the various measures in your output for Part 1A. What do these measures tell us about parental depression and hours worked overall? Are most adults suffering from depression? Do they tend to work a lot of hours?

To Hand in: Printed Syntax and Output files file1a.sps and file1a.spo and write-up (a paragraph or two)

****

The variable adpps01 is meant to measure *depression* for the parents of a large sample of Canadian children. The NLSCY developed this scale by asking a whole series of questions to persons considered 'most knowledgeable' about the child selected in the NLSCY sample (usually their mother). This variable is a scale that tries to document *the level of depression experienced by the parent.* This is a scale that involved many questionnaire items in its construction. If interested, see the corresponding codebook on my assignment page for details on how this scale was created. The scale added up information as collected across several items, such that a high score suggests high levels of depression, whereas lower scores suggest that parents are doing well in terms of mental health.

The variable alfpd02 has "not applicable", as many parents do not have a job outside the home.

****

Part 1B Requirements

Using the other data file on STUDENTSDATA.sav

Step 1: Create syntax and output files file1b.sps and file1b.spo. This involves:

In your syntax file, create a TITLE for your output that includes the syntax file name, date and your name.
Choose any 3 variables in the data set and run frequency distributions on each of them (this can be done in 1 step).

Step 2: Report how many variables and how many cases are in this file in your write-up.

Step 3: Very briefly interpret the frequency distributions.

To Hand in: Printed Syntax and Output files file1b.sps and file1b.spo and write-up (a short paragraph or two, max)
****

A brief description of the variables in this second dataset is in the codebook attached to the end of this assignment outline. The names, content and coding of all of the variables are listed there. Please make reference to it when trying to select variables, and in making sense of your numbers.

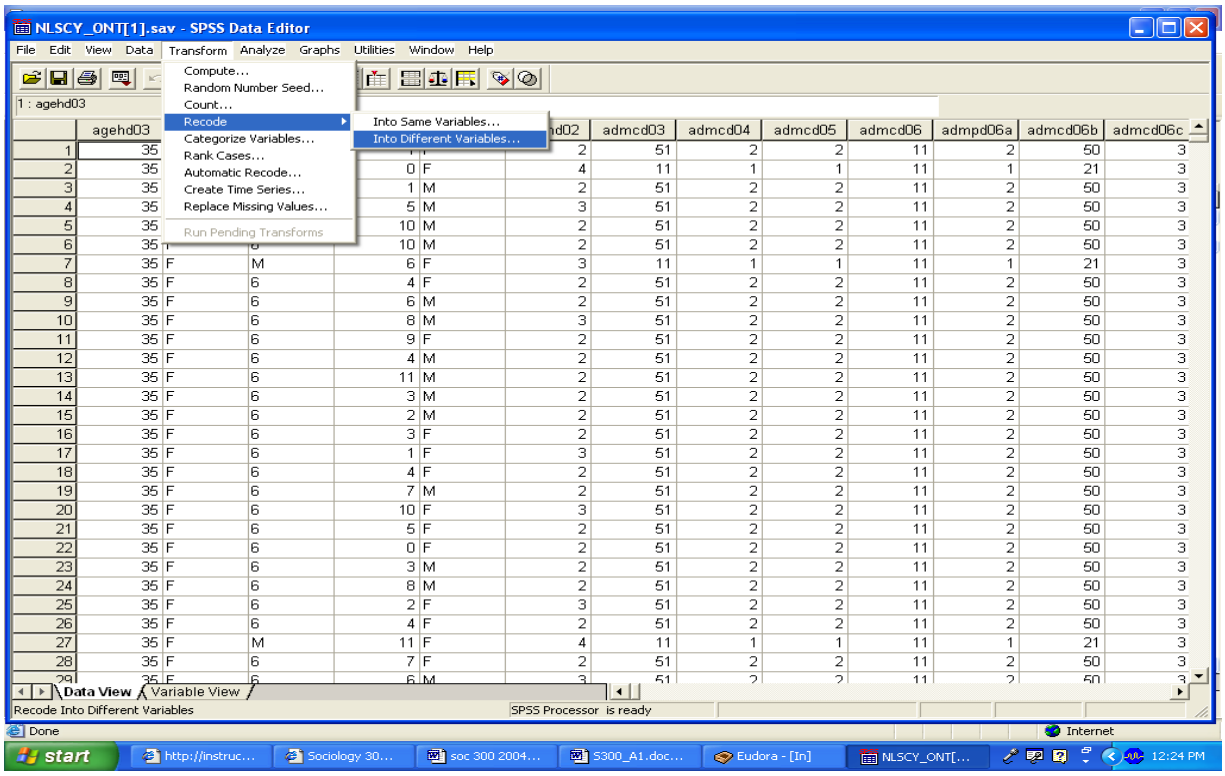
****

Part 2: Manipulating SPSS Files

In introducing SPSS a bit further, it is possible to recode variables into new categories, and to add new variable names and value labels to our recoded variables.

Recoding Variables in SPSS
It is often necessary to create new variables or to recode existing variables with new values. For example, we may have a variable indicating age in years, but may wish to create a new variable with age in five year intervals. It is generally a good idea to create a new variable, rather than changing the values in existing variables (the new variable that you create would have a new name in your dataset and be placed in a separate column).

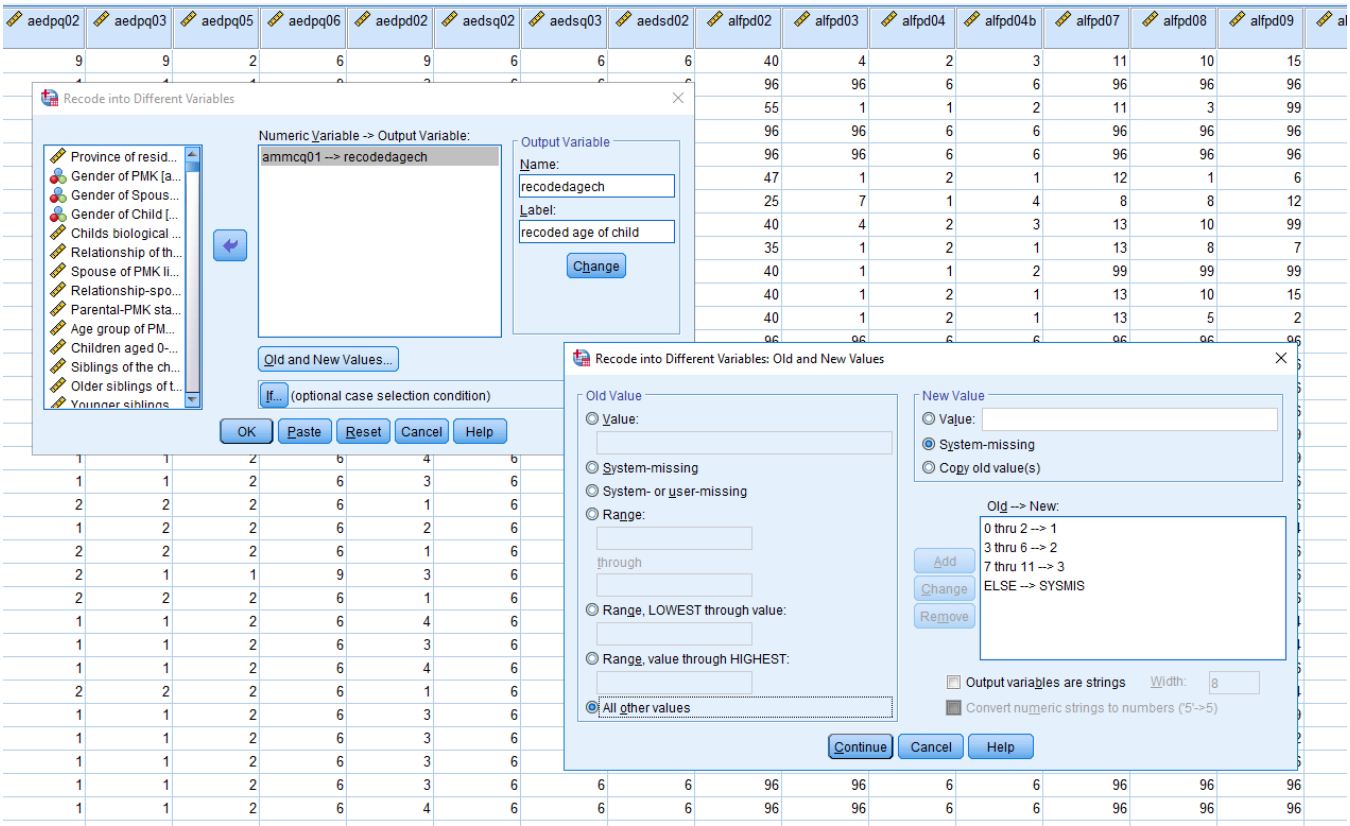
Example: re-categorizing a variable:
To recode a variable using the menu system, choose *transform, recode, into different variables*.



A box will open that will require you to specify a new target variable, and the rules for recoding the variable. For example, suppose that we wanted to recode the variable *ammcq01* (age of child) into a modified variable with grouped ages (*agegr*), whereby we collapse the original variable into fewer categories. The first step is to always consider how your variable was originally coded in the dataset. You can obtain this information from either the *utilities* function in SPSS or via the code book.

We then type in the name of the new variable *recodedagech* and give it an optional label (recoded age of child). Then click on the *old and new values* button in order to specify the rules for creating this new variable. According to the code book, ammcq01 is originally coded such that it ranges from 0 to 11, representing responses from less than one year of age through to 11 years of age. For the purpose of this exercise, assume that we are interested in recoding this variable, such that the new variable (*recodedagech*) has only 3 categories:

(1) aged 0-2 years, (2) aged 3-6, (3) aged 7-11 years



This variable can then be recoded using this procedure. In terms of the syntax for an SPSS program, to recode the age of child variable *ammcq01* into the variable *recodedagech*, you could type directly into the syntax file (alternatively, although certainly not necessary to do so):

RECODE ammcq01 (0 thru 2=1) (3 thru 6=2) (7 thru 11=3) (ELSE=SYSMIS) INTO recodedagech.

This syntax creates a new variable, *recodedagech.* If the value of the old age variable is greater than or equal to 0 *and* it is less than or equal to 2, the value of the new *recodedagech* variable is set to 1. Likewise, if the value of the old age variable was greater than or equal to 3 *and* it is less than or equal to 6, the new *recodedagech* variable is set to 2, and so on. The operators "le", "gt", "ge" "lt", "ne" and "=" can be used to specify "less than or equal to", "greater than", "greater than or equal to", "less than," "not equal to" and "equal to" respectively.

It is always necessary to specify *missing values* in the new variable that you are creating if the old variable has them. In this case, all of the cases which don't fit any of the 3 "IF lines" above will have a *recodedagech* value of zero. It is important to look carefully at the data to make sure that the transformations have occurred correctly, and to specify the missing values for the variable if they exist. We do so here, as a precaution on "missing values".

One important thing to note in creating a syntax file (SPSS program) is that your recode or compute procedures must always come before any specific statistical procedures, such as

frequencies, *descriptives* or *explore*. For example, the following syntax first creates the new variable *recodedagech* prior to running the frequencies on this variable (as well as the original variable ammcq01).

```
RECODE ammcq01 (0 thru 2=1) (3 thru 6=2) (7 thru 11=3) (ELSE=SYSMIS) INTO
recodedagech.

FREQUENCIES
VARIABLES= ammcq01 recodedagech
/ORDER= ANALYSIS
```

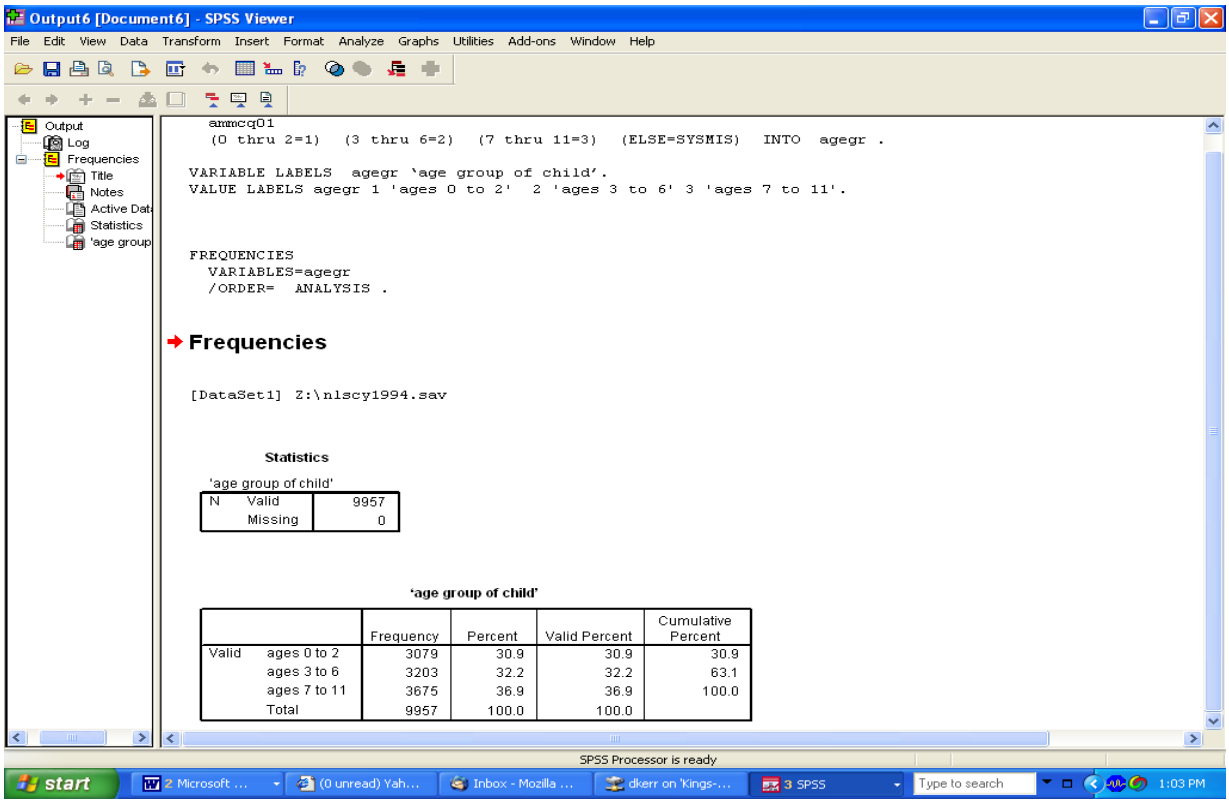
Variable and Value labels

An important part of documenting your work is adding variable and value labels whenever you create new variables. This can be done relatively easily with SPSS syntax.

Returning to the previous example, after creating the new variable, we can specify the variable label (what we want to call the new variable *recodedagech*) as well as identify the corresponding value labels (what we want to call each category of the variable we just created). The variable name must be 7 characters or less, Variable and value labels can be longer, but you should also try to keep them relatively short and as descriptive as possible.

```
RECODE ammcq01 (0 thru 2=1) (3 thru 6=2) (7 thru 11=3) (ELSE=SYSMIS) INTO
recodedagech.
VARIABLE LABELS  recodedagech ‘age group of child’.
VALUE LABELS recodedagech 1 'ages 0 to 2' 2 'ages 3 to 6' 3 'ages 7 to 11'.
EXECUTE.
```

If we then run a frequency distribution on *recodedagech*, we should observe the newly specified labels.



Note: in working with the NLSCY, someone has in fact gone through the trouble of setting up a database that has all the variable names and value labels already allocated. This is not the case with the second dataset "STUDENTSDATA.sav". When you create new variables, you should subsequently specify these new names and labels.

Part 2 Requirements

Using data file: STUDENTSDATA.sav

Step 1: Create syntax and output files file2a.sps and file2a.spo. This involves:

In your syntax file, create a TITLE for your output that includes the syntax file name, date and your name.

Run frequency distributions on the variables *pulse*, *haircut* and *income*, and use the results to decide on the details for the following "step 2".

Step 2: Recode *pulse*, *haircut*, *income* so that they each have 4 or 5 categories, where each category has roughly equal numbers of cases. Remember to name your new variables something different than the old names

Step 3. Give Variable Labels to each new variable. Give Value Labels to the categories in each new variable.

To Hand in: Printed Syntax and Output files file2a.sps and file2a.spo, no write-up for this section. This output should include the three initial variables and the 3 recoded variables (with corresponding variable and value labels).

****

Part 3: Selecting Subsamples and Running Contingency tables (crosstabs)

Selecting subsamples is an important part of using very large datasets with many cases – often we are only interested in people with specific characteristics. Part 3A deals with selecting subsamples.

Crosstabs (contingency tables) allow us to look at how the characteristics of one variable are distributed along another. In other words, contingency tables form one of the simplest ways to explore the possible *relationship* between two variables. Parts 3B and 3C ask you to run and describe crosstabs using the subsample created in part 3A.

Here you are asked to work with 4 variables from the NLSCY. The first is ammcq01 (child's age), which ranges from 0 to 11. The second is afnhq01i (one of the indicators that make up the family functioning scale used in part 1A). This variable comes from the question 'To what extent do you agree or disagree with the following statement: Making decisions is a problem for your family': (1) strongly agree (2) agree (3) disagree, and (4) strongly disagree. The third is admcd04

(Child's single parent status) which is coded as (1) dual parent families (2) lone parent family, and (3) child does not live with parent(s). The fourth is poverty (low income status), which is divided simply into (1) poor (2) not poor.

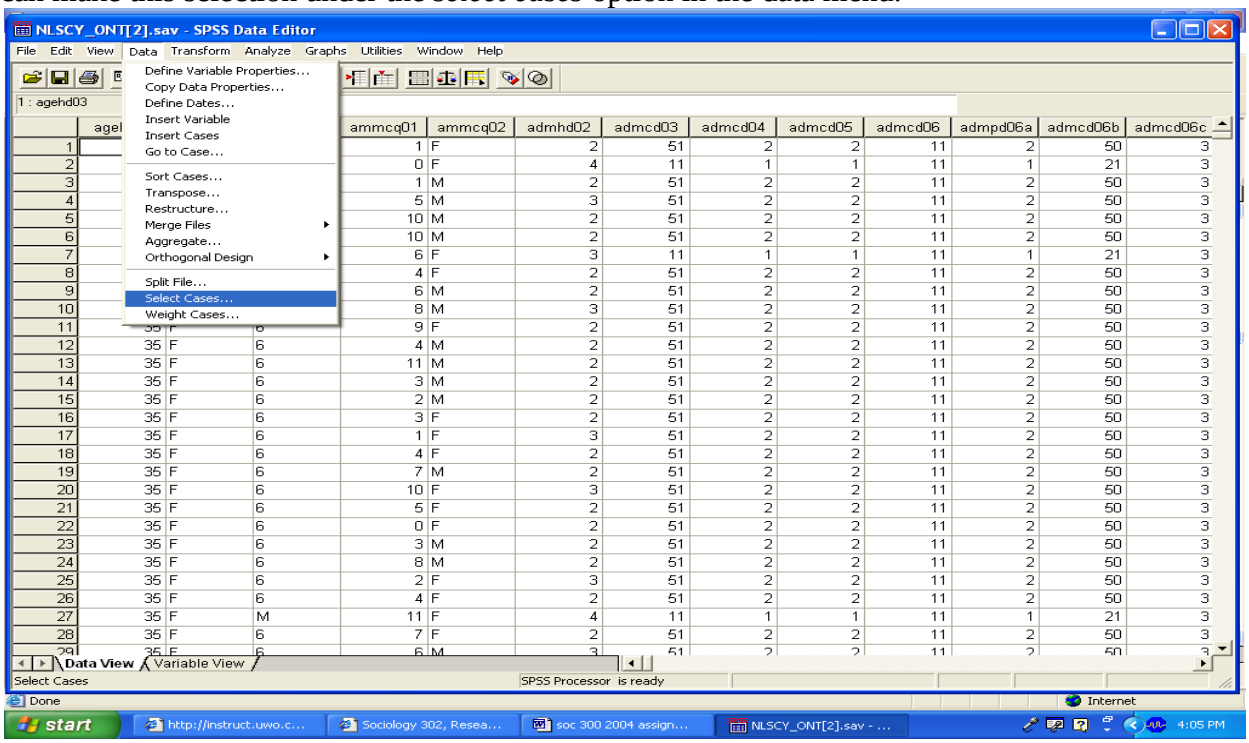
Part 3A is asking you to create a subsample for age of child.

Part 3B is asking you to run and describe crosstabs with child's single parent status as an independent variable (potential cause) and family functioning (afnhq01i) as the dependent variable (potential effect)

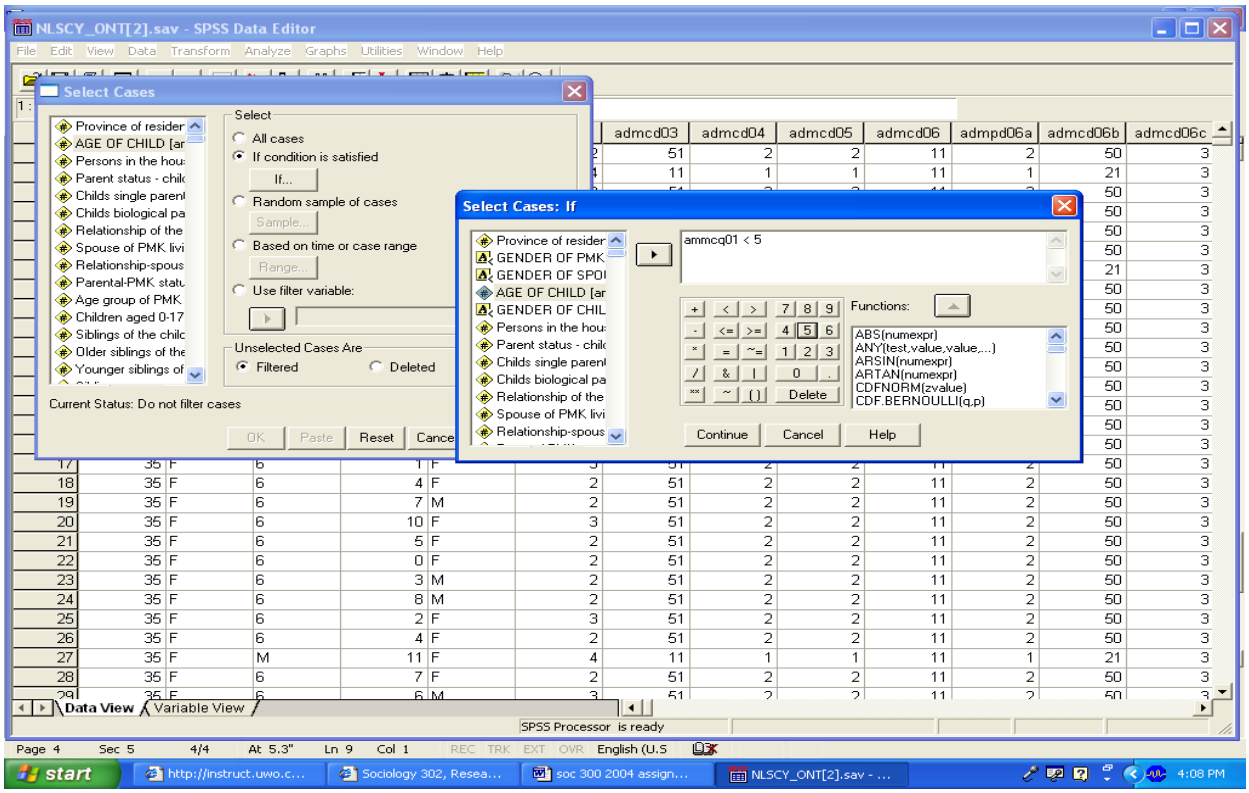
Part 3C asks you to consider a different potential independent variable (poverty status), and to compare the findings of part B and C. Which do you think has the larger effect on family functioning?

Selecting Cases

Sometimes you need to perform an analysis on only some of the cases in a dataset. For example, suppose that you wanted to do an analysis involving exclusively children under the age of 5. You can make this selection under the *select cases* option in the data menu.



Once you have opened up the *select cases* box, you can highlight the variable of interest (which in this case is age of child, *ammcq01*), and then use the *if* button to open another dialogue box that allows you to specify the rules by which the cases will be selected.



In this example, only cases meeting the condition that the value on variable *ammcq01* is 'less than 5' are selected. It is important to check whether unselected cases will be *deleted* (permanently) or merely *filtered* temporarily. You want them filtered, not deleted most of the time. Cases that are not selected will not be included in any subsequent analysis, until you specifically tell the computer to do so (or if you close your dataset). In order to use all of the cases again, you must select *all cases* in the select cases dialogue box, which overrides the previous command.

The syntax that achieves the above selection on *children under the age of 5* is as follows:

```
USE ALL.
COMPUTE filter_$=(ammcq01 < 5).
VARIABLE LABEL filter_$ 'ammcq01 < 5 (FILTER)'.
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .
```

Again, the ordering of procedures in your syntax file is crucial. You must select your cases prior to running any statistical procedures. To move back to the full sample, you can specify this again using the *data, select cases option*. The following example is a program file which contains the syntax to:

(1) select exclusively children aged under 5
(2) run a frequency distribution on the variable ammcq01 (age of child)
(3) remove this filter and return to the full sample
(4) run the same frequency

```
USE ALL.
COMPUTE filter_$=(ammcq01 < 5).
VARIABLE LABEL filter_$ 'ammcq01 < 5 (FILTER)'.
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .

FREQUENCIES
 VARIABLES=ammcq01
 /ORDER=  ANALYSIS .

FILTER OFF.
USE ALL.
EXECUTE .

FREQUENCIES
 VARIABLES=ammcq01
 /ORDER=  ANALYSIS .
```

****

Part 3A Requirements

Using data file: nlscy2019data.sav

Step 1: Create syntax and output files file3a.sps and file3a.spo.
This involves:
In your syntax file, create a TITLE for your output that includes the syntax file name, date and your name
Using *select cases,* filter out all cases where the PMK's child is under 4 so that we are left only with cases where the child is aged 4 to 11.
Run a frequency distribution on variables afnhq01i (one indicator in the family functioning scale used in Part 1A) and admcd04 using our newly filtered sample
Step 2: Very briefly summarize these distributions

To Hand in: Printed Syntax and Output files file3a.sps and file3a.spo and write-up (a sentence or two)
****

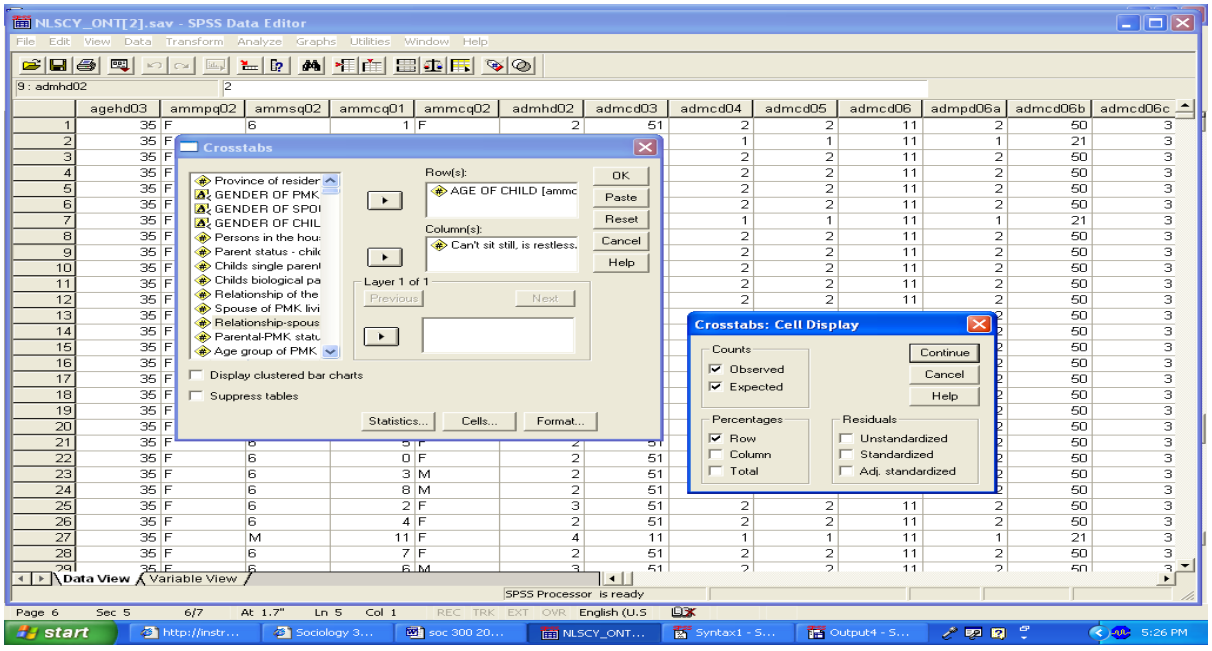
Crosstabs
The crosstabs procedure in SPSS allows you to create a contingency table (also called the crosstabs procedure with SPSS). You can review what the contingency table is all about in your required readings for this course (the chapter on Quantitative analysis). This procedure can be found in the menu system under *analyze, descriptive statistics, crosstabs*. The dialogue box asks you to specify the row and column variables. Obviously, when working with this procedure,

variables that have a very large number of categories become unmanageable. It is sometimes necessary to recode variables if you want to work with the crosstabs procedure – something you just did in Part 2.

As an example, assume that we wanted to examine the relationship between *ammcq02* (gender of the child) and *abecq6b* (can't sit still, is restless?). In this assignment, we will place the dependent variable in the columns and the explanatory (independent) variable in the rows. The text book provides an example which does the opposite (dependent in the rows and independent in the columns). This decision has important implications in terms of interpretation (please read the text carefully on this).

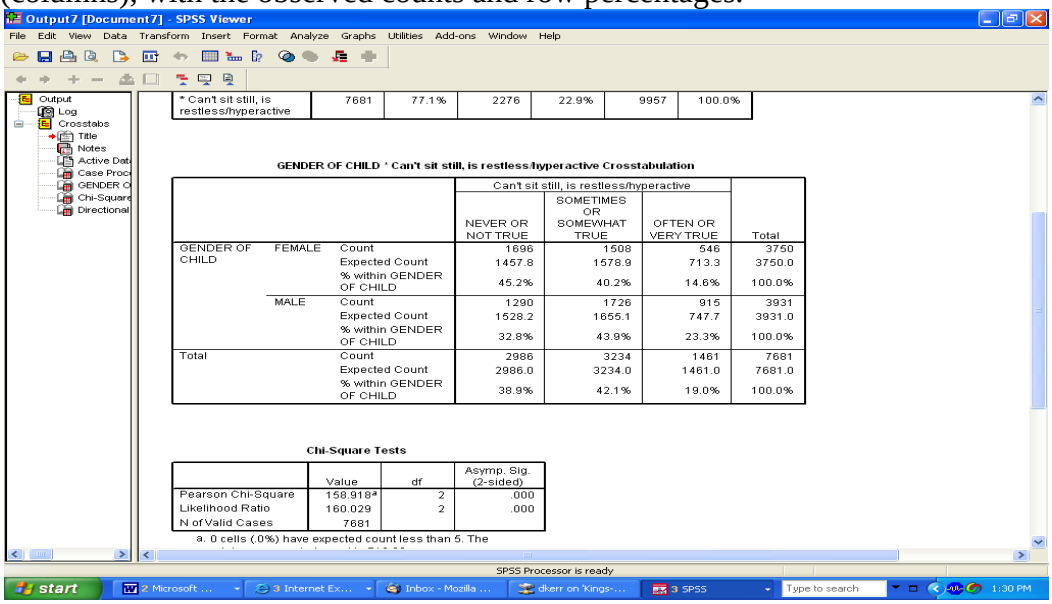
In this example, we can assume that gender is an independent variable placed in our rows and that our indicator of behavioral problems is our dependent variable (placed in the columns). In other words, we assume that gender causes differences in behavioural problems, rather than assuming that behavioural problems cause differences in gender.

The *cells* box allows you to specify the contents of the cells in your output (frequencies, percentages, type of percentage: row, column or total). The *statistics* box allows you to produce the chi square, or other indicators of association. Certainly it is useful, at a minimum, to ask for the row percentages as well as the observed counts. As specified above, when one sets up a contingency table with the dependent variable in the columns, then it makes sense to examine the row percentages as representing the conditional distributions. In other words, you will be able to see how the distribution on the dependent variable varies by category of your independent. If you placed the dependent variable in the rows, then it would make sense to examine the column percentages.



The syntax is as follows:
CROSSTABS
/TABLES= ammcq02 BY abecq6b
/FORMAT= AVALUE TABLES
/CELLS= COUNT EXPECTED ROW .

This procedure produces a cross tabulation of the variable *ammcq02* (rows) BY *abecq6b* (columns), with the observed counts and row percentages.



This output provides you with row %'s that assist in interpretation. For example, we can see that a much larger percentage of boys often can't sit still compared to girls.

Note that this table does not include the same sample that you are working with earlier (in this case, the n=7691). In this example, the sample involves only children aged 2-11 (as the question was not asked of younger children). The other cases (infants) are excluded from this crosstab. If you were to run this example yourself, you'd have to filter out children under 3, as in Part 3A (to avoid having a lot of missing cases).

Part 3B Requirements

Using data file: nlscy2019data.sav
Step 1: Create syntax and output files file3b.sps and file3b.spo
This involves:
In your syntax file, create a TITLE for your output that includes the syntax file name, date and your name
Using 'select cases', filter out all cases where the PMK's child is under 4 so that we are left only with cases where the child is aged 4 to 11.

Using the same filtered sample, run a crosstab on afnhq01h and admcd04,(including row percentages), where admcd04 is the independent variable.
Step 2: What do the row percentages and bivariate tables tell us about the way that afnhq01i is affected by single parent status (admcd04)? This variable asks respondents whether "family members feel accepted as they are?"

To Hand in: Printed Syntax and Output files file3b.sps and file3b.spo and write-up (a paragraph or two)

****

Part 3C Requirements

Using data file: nlscy2019data.sav
Step 1: Create syntax and output files file3c.sps and file3c.spo
This involves: In your syntax file, create a TITLE for your output that includes the syntax file name, date and your name
Using 'select cases', filter out all cases where the PMK's child is under 4 so that we are left only with cases where the child is aged 4 to 11.
Run a second crosstab, this time replacing admcd04 with a variable measuring low income (poverty).
Step 2:
What do the row percentages and bivariate tables tell us about the way that afnhq01h is affected by low income (poverty)? Briefly compare the two crosstabs you've run.
Which do you think has the larger effect on family functioning, based on what you've seen in Part3B and C?
To Hand in: Printed Syntax and Output files file3c.sps and file3c.spo and write-up (a paragraph or two)
****

CODEBOOK for STUDENTSDATA.sav
DETAILED DESCRIPTION OF VARIABLES

soc **Course Number**

CONTENT	*CODE*
Soc 3306	1
Soc 2206 Section 572	2
Soc 2206 Section 574	3

height **Student's height in feet**

CONTENT	*VALUE*
Reported in feet (5.5 feet = 5 feet 6 inches)	

birthyr **Year of birth**

CONTENT	*VALUE*
Year (1981, 1982, …etc)	
Not Stated	9999

Sex **Sex of respondent**

CONTENT	*CODE*
Male	1
Female	2
Not Stated	9

pulse **Pulse of respondent (beats per minute)**

CONTENT	*VALUE*
Beats per minute	
Not Stated/Error	-99

brkfast **Respondent ate a healthy breakfast**

CONTENT	*CODE*
Yes	1
No	2
Not Stated	9

pet **Respondent has a pet living with them**

CONTENT	*CODE*

No	1
Yes	2
Not Stated	9

rmmates **Number of roommates during the school year**

CONTENT	*VALUE*
# of roommates (excluding self)	
Not Stated	-9

movies **Summer movies attended in the theatre**

CONTENT	*VALUE*
# of movies	
Not Stated	-9

job **Enjoyment of current (or last) job (ranges from 1 to 7)**

CONTENT	*VALUE*
Hated it	1
. .	
. .	
. .	
Loved it	7
Not Stated	9

haircut **Cost of last haircut (including tip)**

CONTENT	*VALUE*
In dollars and cents	
Not Stated	999

money **Money on hand (in change only)**

CONTENT	*VALUE*
In dollars and cents	
Not Stated	999

cds **Number of CDs owned**

CONTENT	*VALUE*
# of CDs reported	
Not Stated	999

smoke **Does respondent smoke?**

CONTENT	*CODE*
Yes	1
No	2
Not Stated	9

highsal **Importance of a high salary (ranging from 1 to 7)**

CONTENT	*CODE*
Not at all important	1
. .	
. .	
. .	
Very important	7
Not Stated	9

polpart **Vote for a Federal political party**

CONTENT	*CODE*
Liberals	1
Conservatives	2
NDP	3
Green	4
Other	5
Do not plan on voting	6

Not Stated	9

study **Average number of study hours a week (not including class-time)**

CONTENT	*VALUE*
# of hours	
Not Stated	-9

grade **Grade in Sociology 020**

CONTENT	*CODE*
< 50	1
50-54	2
55-59	3
60-64	4
65-69	5
70-74	6
75-79	7
80-84	8
85-89	9
90+	10
no grade	11
Not Stated	99

profage **Students' guesses at Don Kerr's age**

CONTENT	*VALUE*
Age guessed	
Not stated	-99

income **Parents' annual income**

CONTENT	*VALUE*
In dollars	
Not stated	999999