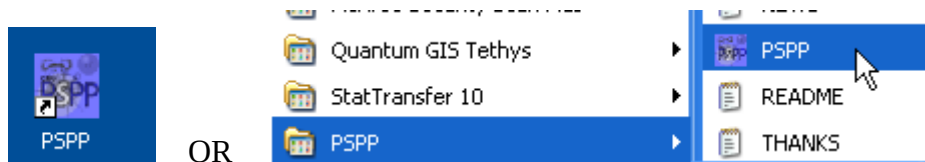


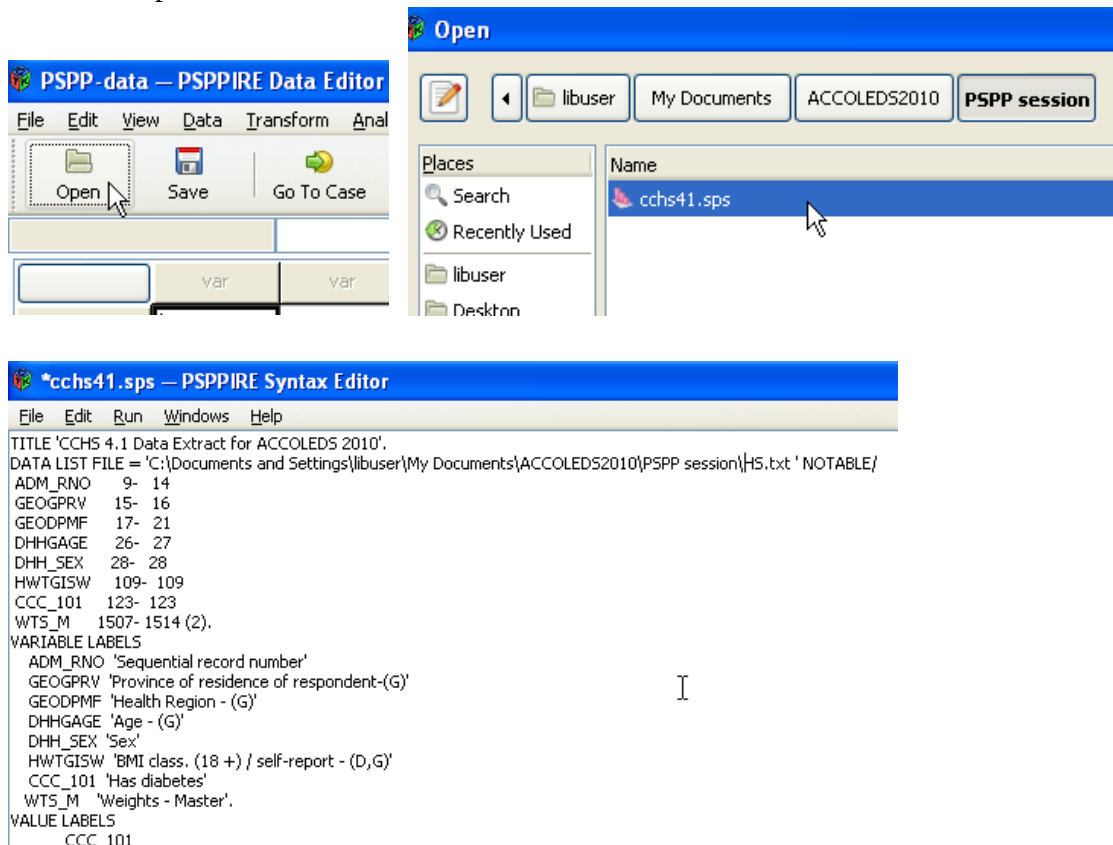
Using Free, Open-Source Tools to Extract and Map DLI Data: Extracting the Data

In this exercise we will be using the open source statistical software called **PSPP** to extract data from the *Canadian Community Health Survey 4.1*. We will also use **PSPP** to do some data manipulation (weighting, transforming), some basic descriptive statistics (frequencies and cross-tabulation) and finally aggregate some statistics by health region for use in the mapping part.

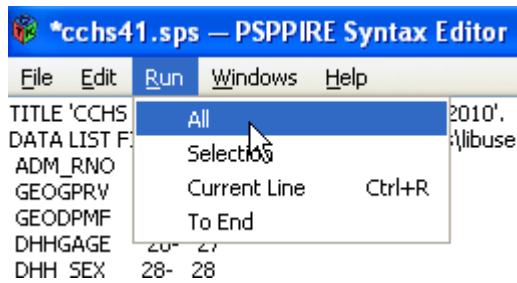
- 1) Start up the **PSPP** program:



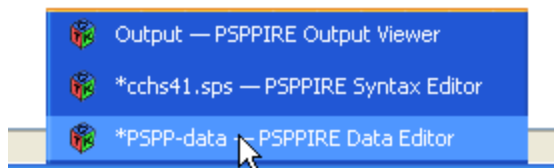
- 2) Open the **cchs41.sps** syntax file: Click on **Open** and then select the **cchs41.sps** file. The syntax file will extract some selected variables from the full CCHS raw data file (**HS.txt**). You will need to edit the **DataList File** command to the path being used on the current computers.



- 3) Run the cchs41.sps syntax file by selecting **Run** and clicking on **All**.



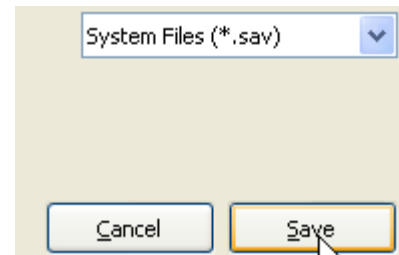
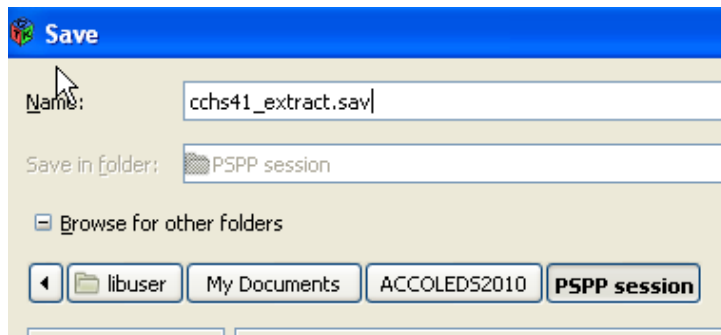
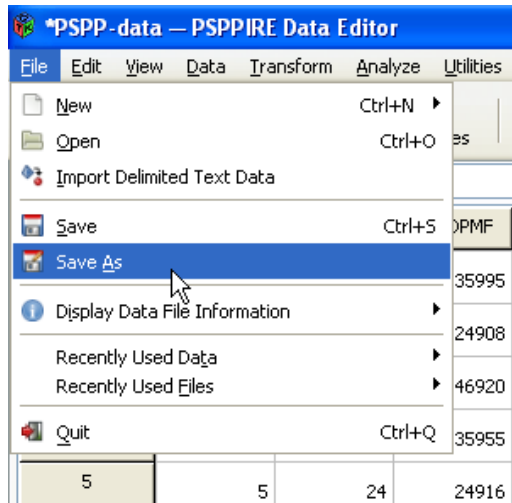
- 4) Have a look at the data file that has been created by selecting the PSPPIRE Data Editor window.



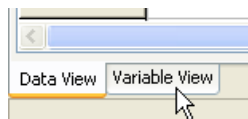
The screenshot shows the PSPPIRE Data Editor window titled '*PSPP-data — PSPPIRE Data Editor'. The window has a menu bar (File, Edit, View, Data, Transform, Analyze, Utilities, Windows, Help) and a toolbar with icons for Open, Save, Go To Case, Variables, Find, Insert Cases, Insert Variable, Split File, Weight Cases, and Value Labels. Below the toolbar is a data table with 10 rows and 10 columns. The first column contains case numbers 1 through 10. The subsequent columns contain numerical data for variables ADM_RNO, GEOGPRV, GEODPMF, DHHGAGE, DHH_SEX, HWTGISW, CCC_101, WTS_M, and var.

	ADM_RNO	GEOGPRV	GEODPMF	DHHGAGE	DHH_SEX	HWTGISW	CCC_101	WTS_M	var
1	1	35	35995	14	1	2	2	336.77	
2	2	24	24908	8	1	9	2	58.92	
3	3	46	46920	7	1	2	2	77.15	
4	4	35	35955	10	2	2	2	48.96	
5	5	24	24916	7	1	2	2	481.27	
6	6	35	35995	6	1	2	2	946.31	
7	7	48	48924	9	1	2	2	45.39	
8	8	35	35947	13	2	3	2	25.40	
9	9	24	24915	12	2	2	2	342.71	
10	10	46	46915	5	1	3	2	61.33	

5) Save the data file created and name it **cchs41_extract.sav**.



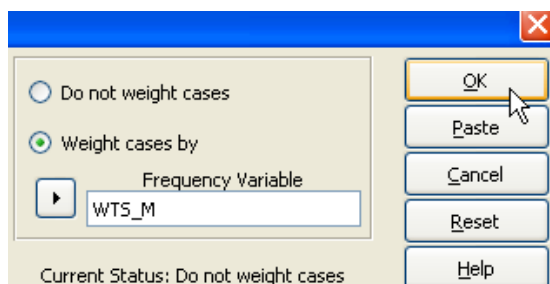
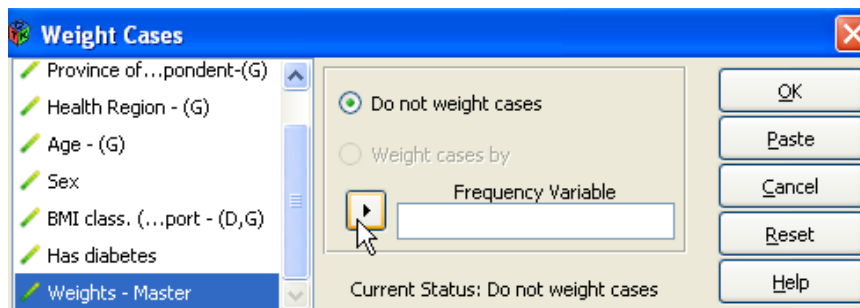
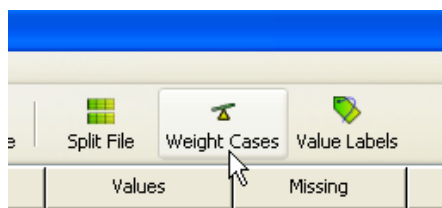
6) Have a look at the **Variable View**.



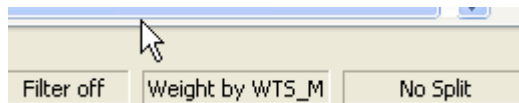
***cchs41_extract.sav – PSPPIRE Data Editor**

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	ADM_RNO	Numeric	6	0	Sequential record number	None	None	8	Right	Scale
2	GEOGPRV	Numeric	2	0	Province of residence	{10,"INFLD & LAB."}_	96 - 99	8	Right	Scale
3	GEOOPMF	Numeric	5	0	Health Region - (G)	{10911,"EASTERN RE"	99996 - 99999	8	Right	Scale
4	DHHGAGE	Numeric	2	0	Age - (G)	{1,"12 TO 14 YEARS"}_	None	8	Right	Scale
5	DHH_SEX	Numeric	1	0	Sex	{1,"MALE"}_	6 - 9	8	Right	Scale
6	HWTGISW	Numeric	1	0	BMI class. (18 +) / s	{1,"UNDERWEIGHT"}_	6 - 9	8	Right	Scale
7	CCC_101	Numeric	1	0	Has diabetes	{1,"YES"}_	6 - 9	8	Right	Scale
8	WTS_M	Numeric	9	0	Weights - Master	None	None	8	Right	Scale

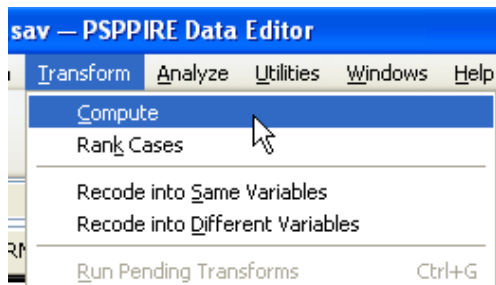
- 7) Apply weighting to the cases by selecting the **Weight Cases** graphic tool at the top of the page. Use the **Weights-Master (WTS_M)** variable as the weighting factor. Move the variable over to the **Frequency Variable** box and click **OK**.



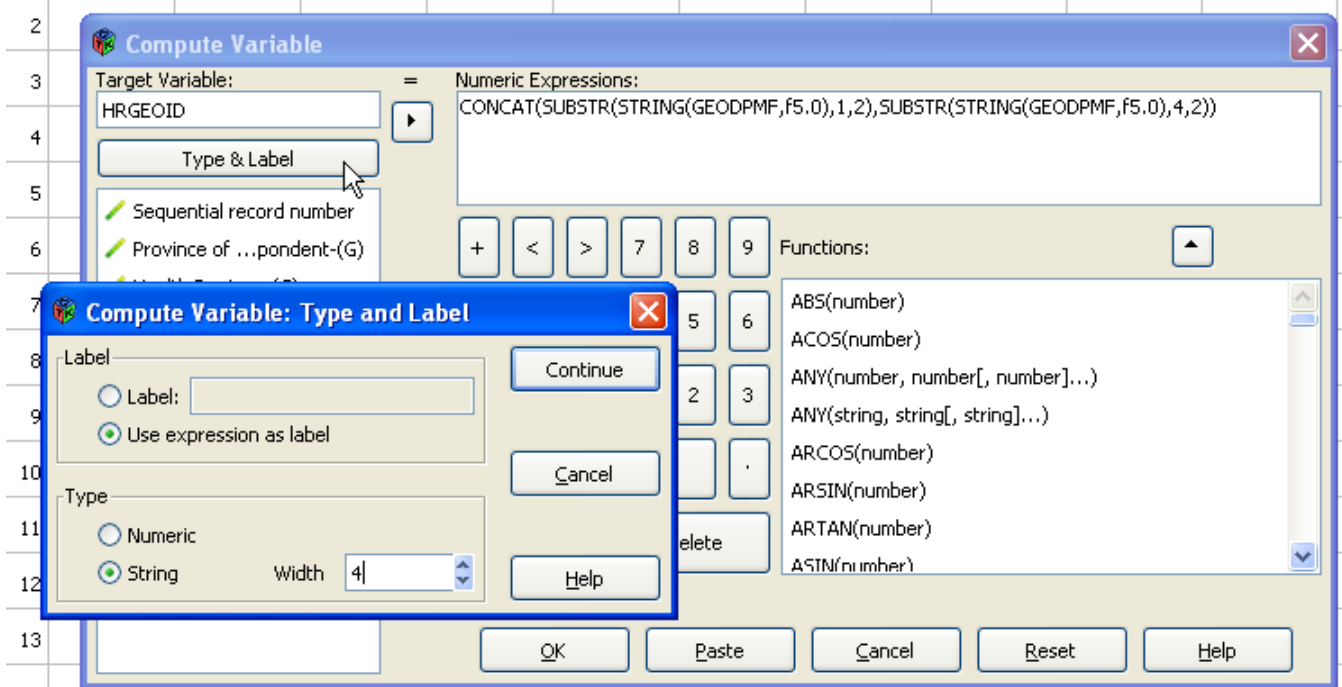
Note the change in status at the bottom right showing that weighting is on.



- 8) We will now transform the Health Region (**GEODPMF**) variable so that it will be compatible with the Health Region geocode in the boundary file in the mapping exercise. At the top menu select **Transform → Compute**

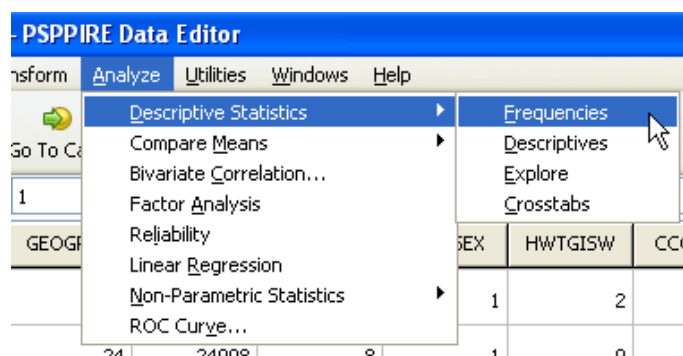


- 9) Create the **Target Variable**. Name it **HRGEOID**. Click on **Type & Label** and give it a **Type String** and **Width 4**. Click on **Continue**. Enter the following **Numeric Expression**:
CONCAT(SUBSTR(STRING(GEODPMF,f5.0),1,2),SUBSTR(STRING(GEODPMF,f5.0),4,2)).
 Click on **OK**. This expression will transform the 5-digit numeric **GEODPMF** variable into a 4-digit string variable by removing the “9” in the third position of the **GEODPMF** number. There should now be another column (variable) named **HRGEOID**.

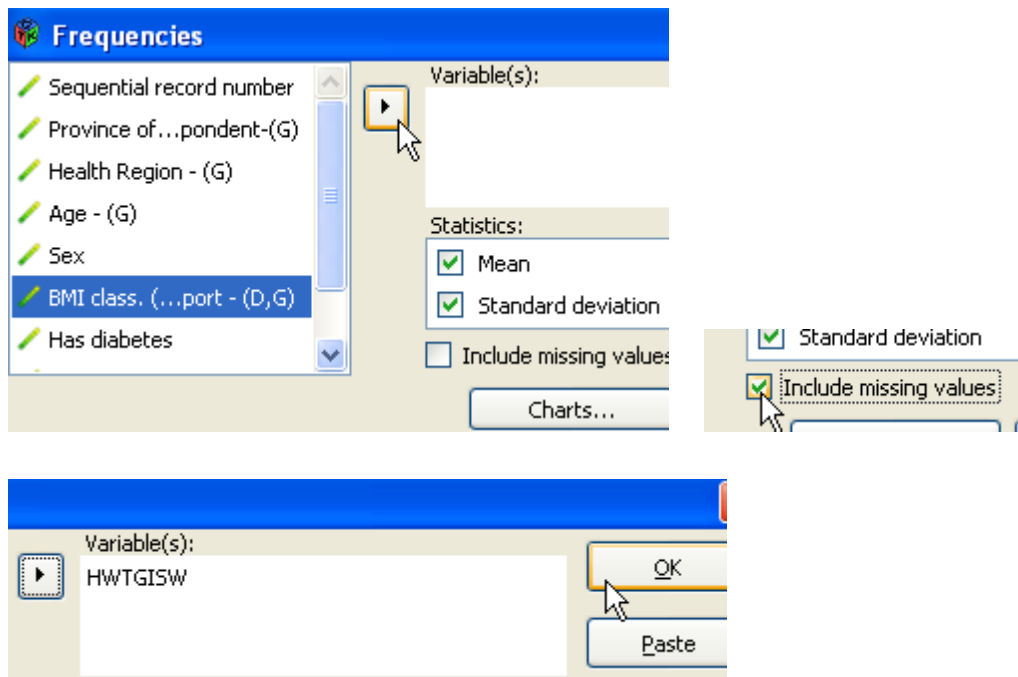


	HRGEOID
7	3595
9	2408
7	4620

10) We will now create some descriptive statistics. Click on **Analyze → Descriptive Statistics → Frequencies** at the top menu..



- 11) In the resulting dialogue box, move the **BMI class (HWTGISW)** variable to the **Variable** box, check off **Include missing values** and click **OK**.



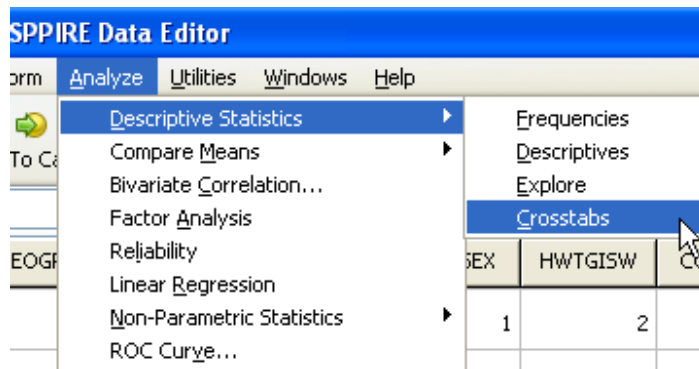
- 12) Open the PSPP output window to see the resulting tables showing the number of weighted cases falling into each Body Mass Index category. What percentage of Canadians (12 years and older) are overweight or obese?

FREQUENCIES
FREQUENCIES
/VARIABLES= HWTGISW
/FORMAT=AVALUE TABLE.

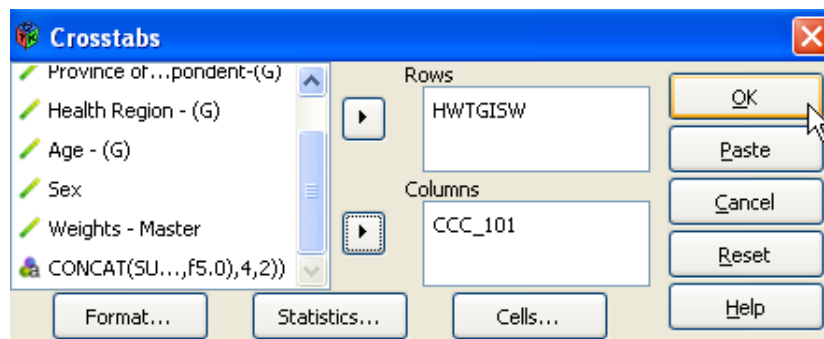
Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
UNDERWEIGHT	1	637776	2.28	2.65	2.65
NORMAL WEIGHT	2	11170641	39.87	46.42	49.07
OVERWEIGHT	3	8157794	29.12	33.90	82.97
OBESE	4	4097596	14.63	17.03	100.00
NOT APPLICABLE	6	2558832	9.13	Missing	
NOT STATED	9	1394732	4.98	Missing	
Total		28017372	100.0	100.0	

N	Valid	24063808
	Missing	3953564
Mean		2.65
Std Dev		.79
Minimum		1.00
Maximum		4.00

13) Click on **Analyze** → **Descriptive Statistics** → **Crosstabs** at the top menu.



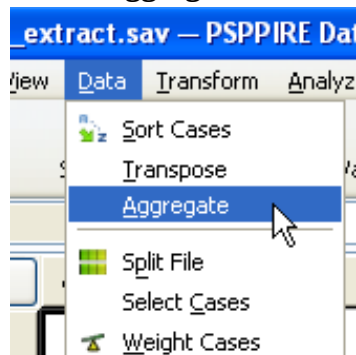
14) Move the **BMI class (HWTGISW)** variable into the **Rows** box and the **Has diabetes (CCC_101)** variable into the **Columns** box. Click **OK**.



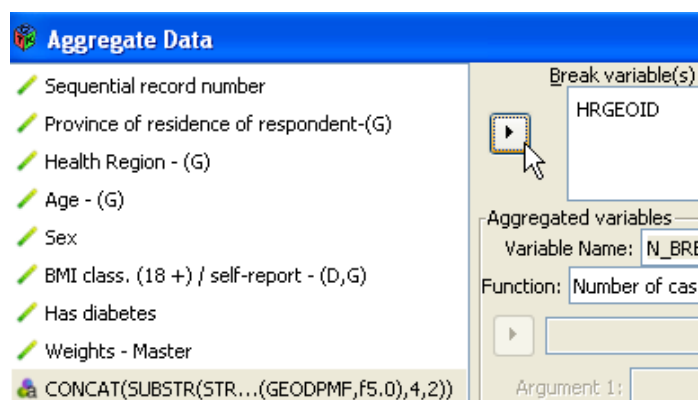
15) Open the PSPP output window to see the resulting table showing the diabetes variable cross-tabulated with the Body Mass Index variable. Is there any relationship between having diabetes and being overweight or obese?

HWTG/ISW	CCC_101		Total
	YES	NO	
UNDERWEIGHT	12342.9 1.9% .8% .1%	625185.7 98.1% 2.8% 2.6%	637528.6 100.0% 2.7% 2.7%
NORMAL WEIGHT	356872.4 3.2% 23.7% 1.5%	10809689.9 96.8% 48.0% 45.0%	11166562.3 100.0% 46.4% 46.4%
OVERWEIGHT	532973.3 6.5% 35.4% 2.2%	7617102.8 93.5% 33.8% 31.7%	8150076.1 100.0% 33.9% 33.9%
OBESE	603492.5 14.7% 40.1% 2.5%	3488538.9 85.3% 15.5% 14.5%	4092031.4 100.0% 17.0% 17.0%
Total	1505681.1 6.3% 100.0% 6.3%	22540517.2 93.7% 100.0% 93.7%	24046198.3 100.0% 100.0% 100.0%

- 16) We will now group (**Aggregate**) all the cases by Health Region so that we can prepare the diabetes incidence data file that will be used in the mapping part. Select **Data** → **Aggregate** from the top menu.



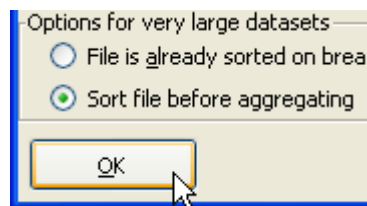
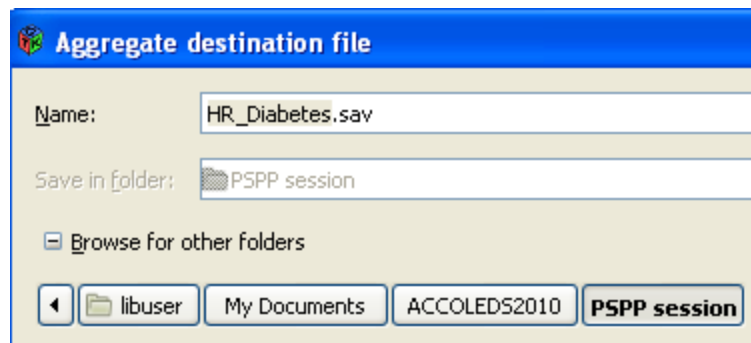
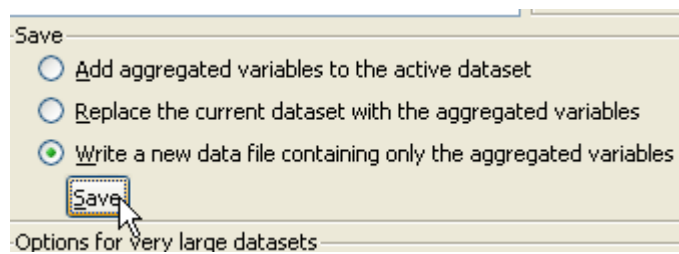
- 17) Move the newly computed **HRGEOID** variable (4 digit string) into the **Break variable(s)** box.



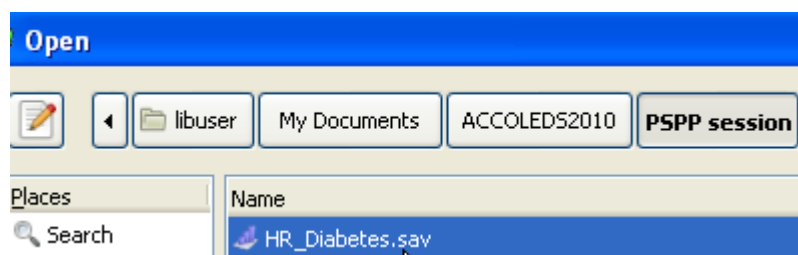
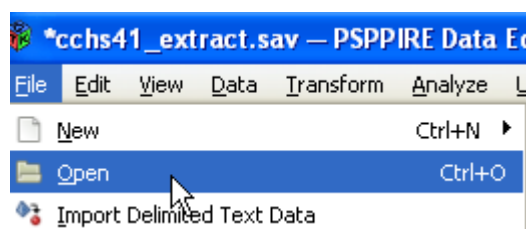
- 18) Under **Aggregated Variables** give the **Variable Name** as *HR_Diabetes* and select the **Function** *Percentage less than*. What we are doing is basically creating a new variable that will contain the aggregated data (percentage of individuals who have been diagnosed with diabetes). Since the value for Yes=1 that is why we are using the percentage of individuals with a value less than 2; the missing values will be greater than 2.

- 19) Move the **Has diabetes (CCC_101)** variable into the box under **Function** and enter 2 as **Argument 1**. Add the Function to the box below.

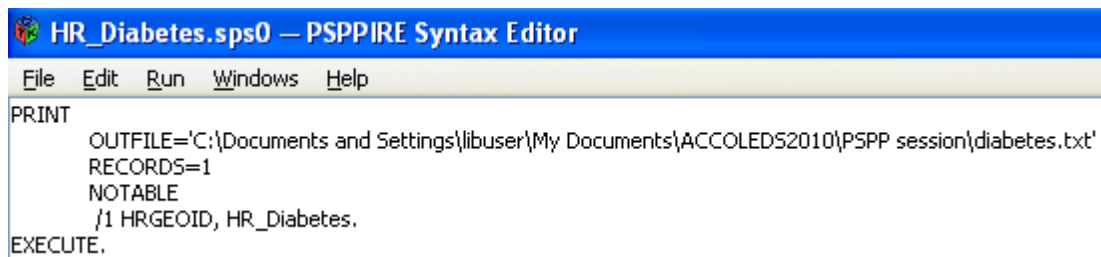
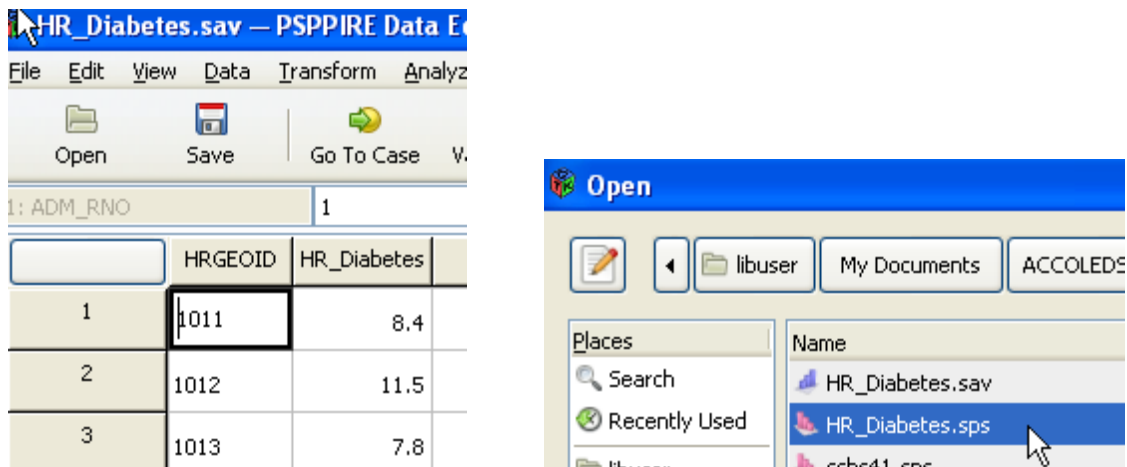
- 20) Save the resulting file as a **new data file containing only the aggregated variables** and name it **HR_Diabetes.sav**. Click **OK**.



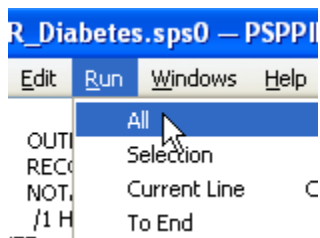
21) Open the newly created **HR_Diabetes.sav** file.



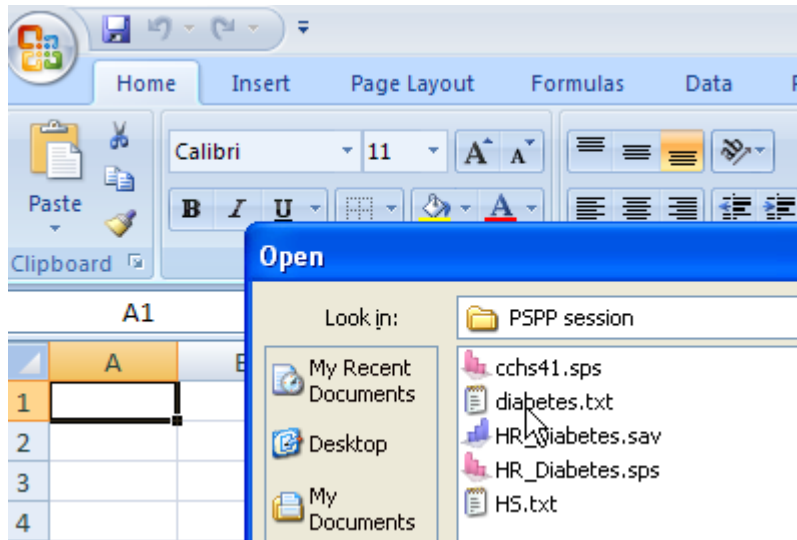
- 22) We will now need to export the aggregated file in a format that will be compatible with the GIS software used in the mapping part. Open the **HR_Diabetes.sps** file. This syntax will create the output required. You will need to edit the **OUTFILE** command line to the path being used on the current computers.



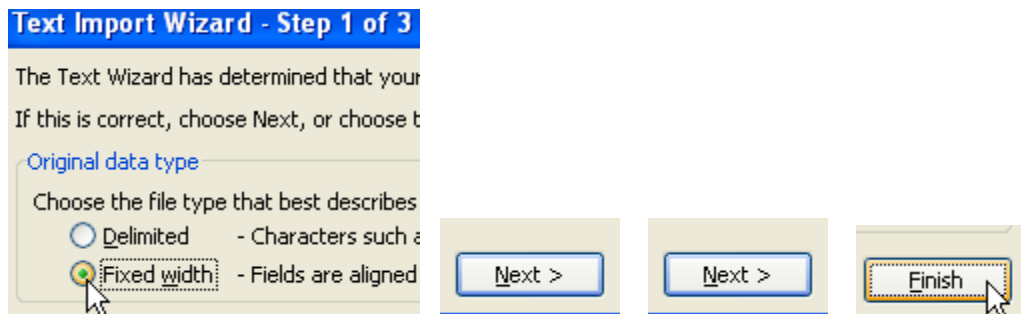
- 23) Run the syntax from the **HR_Diabetes.sps** file: **Run → All**.



- 24) Close down all the **PSPP** windows and start up Microsoft **Excel**. Open the newly created **diabetes.txt** file. You will need to change the **Files of type:** to **All Files (*.*)** to see the **diabetes.txt** file.



- 25) Since you are importing a txt file you will be presented with a **Text Import Wizard**. Select **Fixed Width** as the **Original data type** in Step 1. Click on **Next**, then at Step 2 click on **Next** again and finally at Step 3 click on **Finish**.



- 26) Insert a line at the top and enter the following as column headers: **HRGEOID** (health region geocode) into cell A1 and **DIABERC** (percent with diabetes variable) into cell B1. Save the file as **diabetes.csv**. When asked about keeping the format just answer Yes; you will be prompted with this question twice. This file will serve as the data input table for the mapping part.

Clipboard Font			
A1			
	A	B	C
1			
2	1011	8.4	
3	1012	11.5	
4	1013	7.8	

	A	B
1	HRGEOID	DIABPERC
2	1011	8.4
3	1012	11.5
4	1013	7.8

File name:	diabetes.csv
Save as type:	CSV (MS-DOS) (*.csv)
Save	

Good job! Time for a well-deserved break.

For more information on PSPP and to download the program, go to <http://www.gnu.org/software/pspp/>.