

## **Exploring Relationships between Education and Annual Income Across Age Cohorts**

by

Manny Avila, Teacher-candidate, Queen's University, Faculty of Education,  
1maa1@qmlink.queensu.ca

While on a practicum at Statistics Canada March, 2003

Working under the general direction of Joel Yan

### **INTRODUCTION**

What are the advantages of staying in school? From time to time, students wonder whether their academic efforts will be useful to them in the long term. Within this activity, we will attempt to determine if there is a relationship between a person's highest level of schooling and his or her annual income. We will examine data for a cross section of Ontario residents collected in the 1991 Census using Fathom.

### **RELATED EXPECTATIONS (for the MDM4U course)**

#### ***Overall Expectations***

- Organize data to facilitate manipulation and retrieval
- Solve problems involving complex relationships, with the aid of diagrams
- Demonstrate an understanding of standard techniques for collecting data
- Analyse data involving one variable, using a variety of techniques
- Describe the relationship between two variables by interpreting the correlation coefficient

#### ***Specific***

- Locate data to answer questions of significance or personal interest, by searching, well-organized databases
- Create a database
- Represent data in a matrix
- Compute using technology, measures of one-variable statistics (ie mean, mode, range, quartiles, variance, standard deviation, etc.)
- Interpret one-variable statistics to describe the data set
- Define the correlation coefficient for a set of data, using graphing calculators or statistical software
- Describe possible misuses of regression
- Access the validity of conclusions made on the basis of statistical studies, by analyzing possible sources of bias in the studies

## PROCEDURE

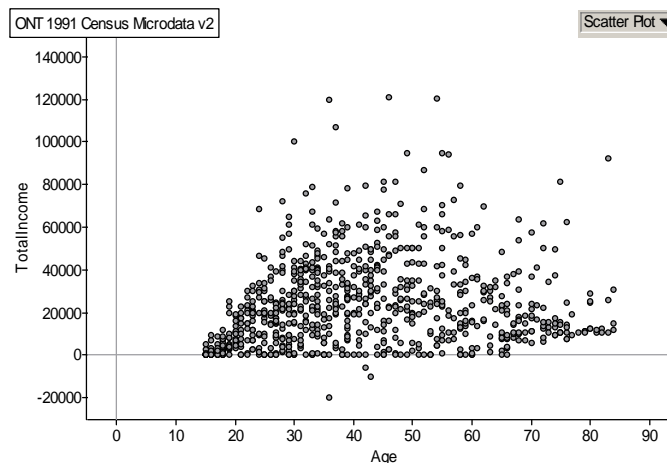
### Accessing the Ontario 1991 Microdata

1. Start Fathom.
2. Open file **ONT 1991 Census Microdata Master.ftm**
3. Double click on the collection “**Ont. 1991 Census Microdata**” and examine the data in the inspector. [**CRTL-I** Opens the inspector]
  - *How many cases are there? How many attributes are there?*
  - *Which of the attributes listed might have an impact a person’s earning potential?*
4. Create a **case table** by clicking on the collection once, and then dragging the case table icon onto the workspace.

### Creating a Total Income versus Age Scatter graph

*In general one can expect to earn more money as one gains more work experience. One way for us to examine one’s earning potential over a lifetime is to create a scatter graph of Total Income vs. Age.*

5. Create a new graph by dragging the graph icon onto the workspace (**CRTL-G** Creates a new graph).
6. To get Fathom to treat age entries as numeric instead of strings **press the CRTL key** and then **select the Age** attribute in the case table and then drag this onto the **X-axis** of the new graph.
7. Again press the **CRTL** key, select the **Total\_Income** attribute in the case table, and drag this onto Y-Axis of the graph.



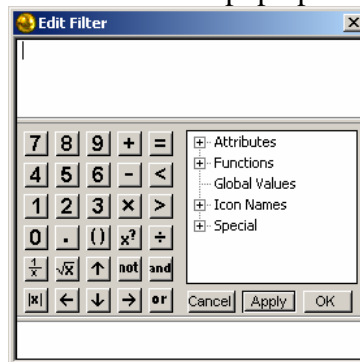
8. Take a moment to examine this scatter graph.
  - *Does this scatter graph provide you with any useful information?*
  - *What is the range of ages? What is the range of income?*
  - *After age 65, almost everyone has at least some income? Why?*

9. Create a **least-squares line**, by right clicking on the graph and selecting Least-Squares line.
  - *What does the slope of the line represent?*
  - *What is the correlation coefficient?*
  - *From the data provided in this graph, can we conclude that there is a strong correlation between age and total income? What other factors may be skewing the data?*

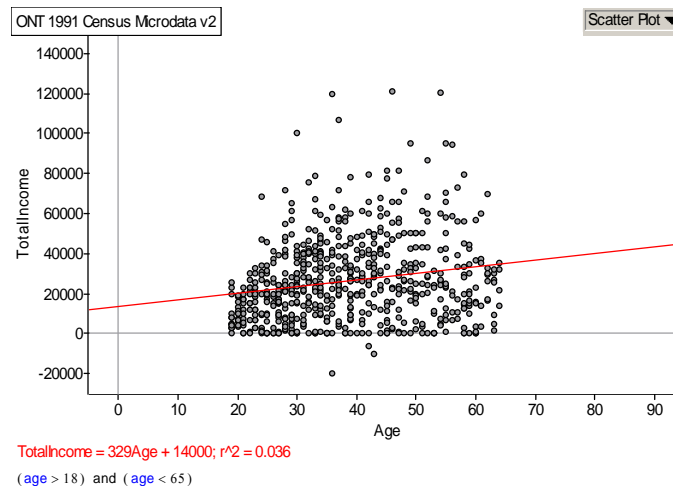
### Using Filters to Create More Homogeneous Groups

*Our analysis of correlations is most effective when we are examining a controlled set of data – that is, when the data are “free” from bias or when confounding factors which may affect the end result have been taken into account. For example, is it fair to compare a 16 year old working part time with a 55 year old experienced lawyer? In this case, the difference between their incomes probably cannot be explained by age alone. The difference in level of schooling, fulltime work, and work experience, amongst other factors are probably more important in explaining the income difference. Fathom allows us to apply filters to graphs and case tables. By employing filters, we can examine more uniform groups and identify trends within groups. For our analysis we will examine only individuals between the ages of 19 to 64 who work full-time.*

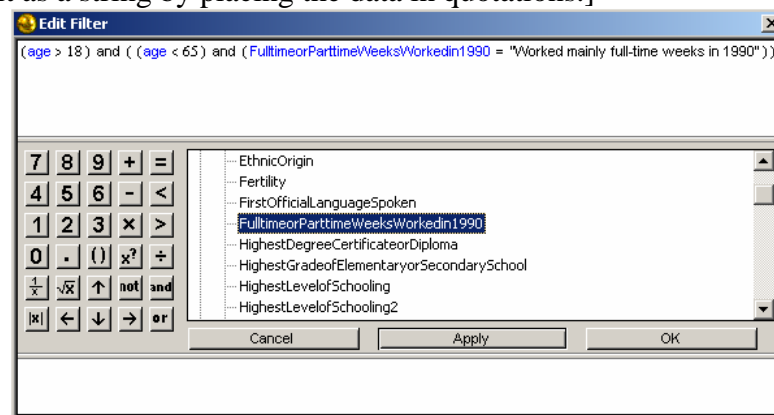
10. To add a filter, right click the graph and select **Add Filter** (Alternatively, press **CRTL-F**, OR click on the graph, then select Add Filter from the Data menu). A window like the one shown below should pop up.



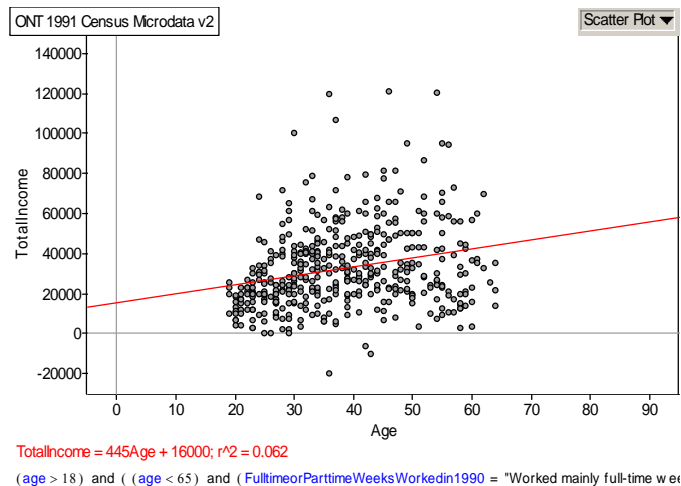
11. The filter accepts Boolean Expressions. In the filter box, type “(age>18) and (age<65)”, then press return [**A**ge should appear in blue]. This command tells Fathom to only display data for cases which satisfy the conditions specified in the filter. In this case, only individuals between age 19 and 64 are now displayed.



12. Take a moment to examine the graph.
  - What is the range of ages?
  - What is the range of income?
  - What is the slope? How does this compare with the previous slope?
  - In comparison to the previous graph, how has the value of the r-squared (correlation coefficient) changed?
13. Fathom will recognize,  $\geq$ ,  $\leq$ ,  $\neq$ ,  $>$ ,  $<$ ,  $=$ , as well as AND, OR, NOT, NOR statements in its filters. We are going to add another condition to the filter – fulltime work. To alter the filter, double click on “(age>18) and (age<65)”.
14. To the current filter type an additional **and**, then type (. This should create a set of parenthesis. Single click inside the empty parenthesis, and then Double click on **Attribute** and then double click **FulltimeorParttimeWeeksWorkedin1990** [you may need to adjust the size of the filter menu]. Then type, = “**Worked mainly full-time**” [NOTE: Fathom’s filters automatically treat data as numeric, unless we specify it as a string by placing the data in quotations.]



15. Then click **Apply**.



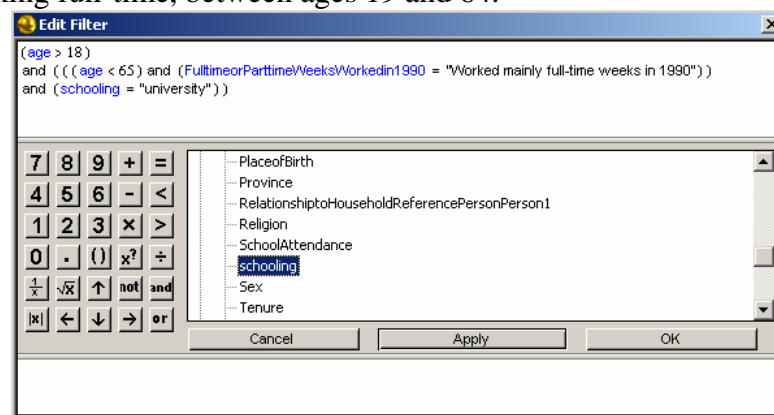
16. Take a moment to examine the graph.

- *How has the correlation coefficient changed? What does this mean?*

17. At the outset of this activity we wanted to determine whether there was a relationship between one's level of income, and one's level of schooling. The census collected data on one's Highest Level of Schooling. The results were assigned amongst 14 different sub categories. Examine the "HighestLevelofSchooling" attribute. Under the "schooling" attribute, the data has been regrouped from 14 into 4 groups: university, less than university, high school, less than high school.

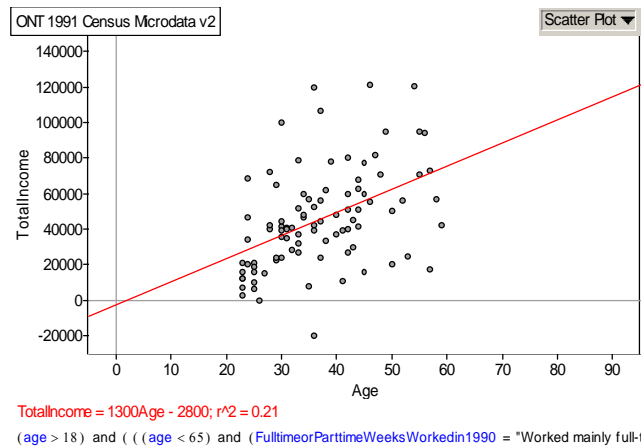
- *What are the advantages to grouping the data this way?*
- *What are the disadvantages to grouping the data this way?*
- *How might the size of our sample affect what we can say about the general population?*

18. Add another condition to the filter. Let's examine, only university graduates, who are working full-time, between ages 19 and 64.



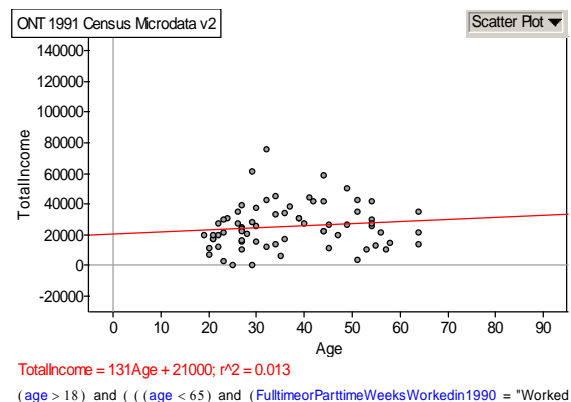
19. Examine the graph.

- *How has the correlational coefficient changed?*
- *What is the range of income? What is the slope of the line, and what does it represent?*



(age > 18) and ((age < 65) and (FulltimeorParttimeWeeksWorkedin1990 = "Worked mainly full-"

20. Create a similar scatter graph with a least-squares line, that examines, people working full-time, between ages 19 and 64, who's highest level of schooling is high school. [Hint, in your filter, (schooling="highschool") should be one of your conditions].



(age > 18) and ((age < 65) and (FulltimeorParttimeWeeksWorkedin1990 = "Worked

21. Compare the two graphs that you have created.
- What trends do you notice?
  - Explain the meaning of the differences in the two slopes of the two regression lines.
  - What other types of diagrams could you create to more firmly establish your conjectures?

## Grouping the Data According to Age Cohorts

If you examine the Age attribute, you'll notice that the individual ages are recorded. Sometimes the trends in data are more visible to us after the data have been grouped. For our next analysis, we will regroup the data into different age groups. The cohorts will be as follows:

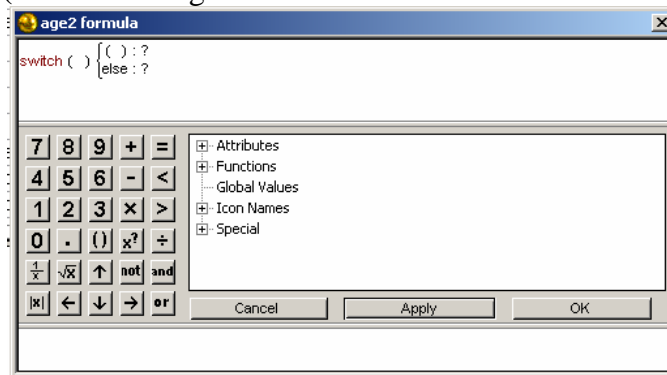
18 and under	45-54
19-24	55-64
25-34	65-74
35-44	75+

The **switch** command allows us to regroup our data.

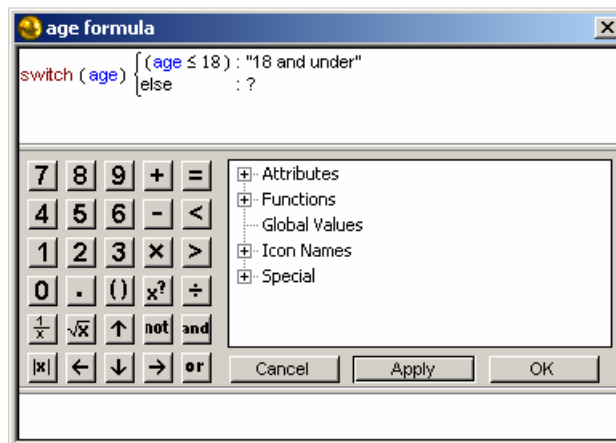
- Why is it useful to group the data this way?

- What is unique about the 18 and under age cohort? The 19-24 cohort? The 65-74 & 75+ cohorts?

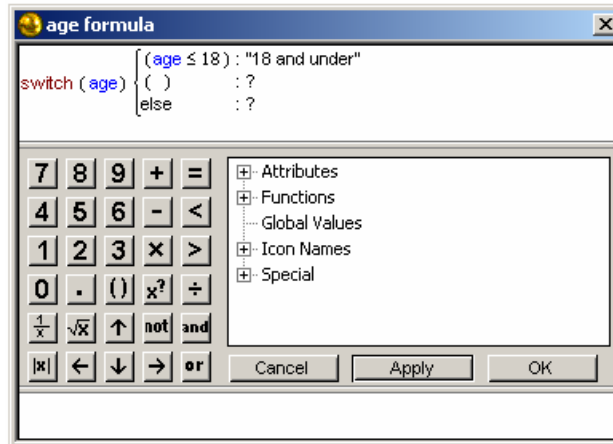
22. In your case table scroll far to the left and create a new attribute called “age2.”
23. Right click on the “age2” attribute and then select **Edit Formula**. (**CRTL-E** brings up the formula box). We are going to use the “switch” command to reorganize the age data.
24. Type **switch**( into the dialogue box. A window like below should appear.



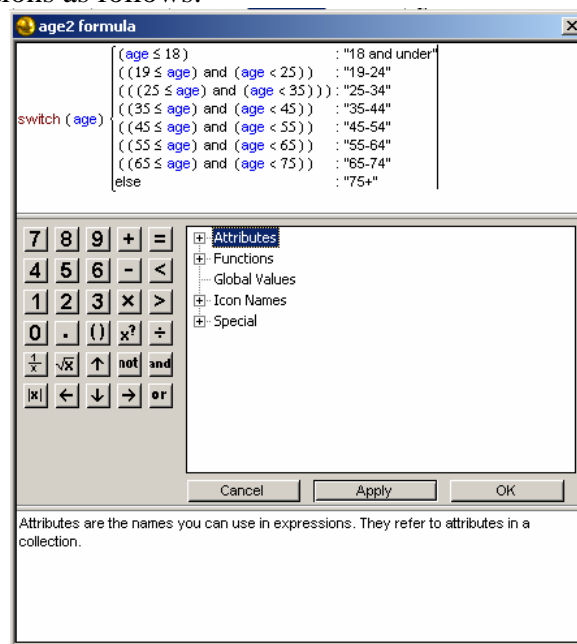
25. Type **age** into the ( ) beside switch. On the right hand side type **(age ≤ 18): “18 and under” and under”**. To create ≤, depress CRTL key, and then select < from the keypad displayed on the screen.



26. To create another condition, place the cursor at the end of the line and press **CRTL+Enter**.



27. Enter the conditions as follows:



28. Click **APPLY**, and then **OK**. In the age2 attribute, you should now find that each “case” has now been assigned to an age cohort.

## Using Histograms and Box & Whisker Plots to Analyse Trends

- *What information can histograms provide? Why might a histogram be useful for analyzing this set of data?*
- *What information do box & whisker plots provide? Why might this type of diagram be useful to analyse this set of data?*

### Histograms

29. Create a new graph by dragging the graphing icon onto the workspace.

30. Drag the age2 attribute onto the X-Axis.

- *What information does this Histogram provide?*

31. While depressing the **CRTL** key, drag the **TotalIncome** attribute onto the Y-axis.

- *What information does this Histogram provide?*
32. Apply a filter to this histogram to so that it only represents, university graduates, aged 19-64, who were employed full-time in 1990. [Big Hint: You can copy and paste the filter from the corresponding scatter graph, by highlighting it, pressing CTRL-C, and then pasting it (CTRL-V) into the “new filter” window].
  33. Adjust the bin widths, vertical and horizontal axis if necessary. Double click on the x-Axis and then adjusting the values – double click on the blue letters and respecify them.

Information about this graph:

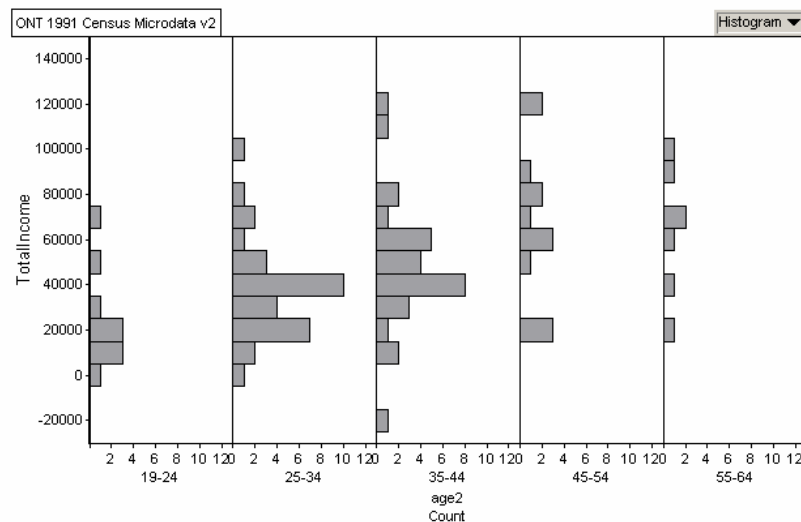
Histogram: Bin width: **10000** starting at: **-25000**

The **TotalIncome** axis is vertical from **-50000** to **150000**

The **Count** axis is horizontal from 0 to **73.522**

#### 34. Examine the Histogram.

- *What trends can you identify?*
- *As people get older, in general what seems to be happening to their income?*
- *What type of diagram might show this trend more explicitly?*



(age > 18) and ((age < 65) and (FulltimeorParttimeWeeksWorkedin1990 = "Worked mainly full-time weeks in 1990")) an

Information about this graph:

Histogram: Bin width: **10000** starting at: **-25000**

The **TotalIncome** axis is vertical from **-30000** to **150000**

The **Count** axis is horizontal from 0 to **13.000**

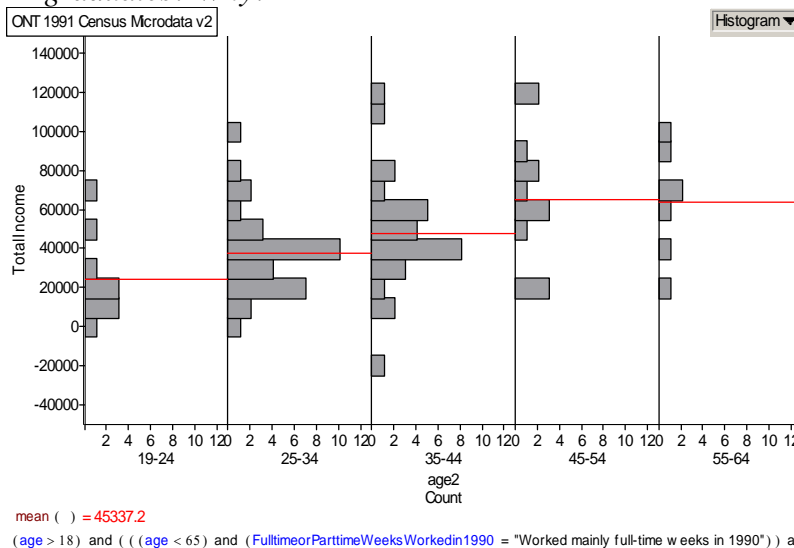
### Plotting the Mean Value On the Histogram

*Sometimes we may want to summarize our data by using a single statistic – the mean, the median, etc. By using the ‘Plot value’ option in Fathom we can plot these values.*

35. Right click on the histogram and select **PLOT VALUE**.
36. Double click on **Functions**, double click on **Statistical**, double click on **One Attribute**, and then double click on **mean**, click on **apply**, and then click on **OK**.

The red line on your histogram represents the mean. (You may also want to add in the median).

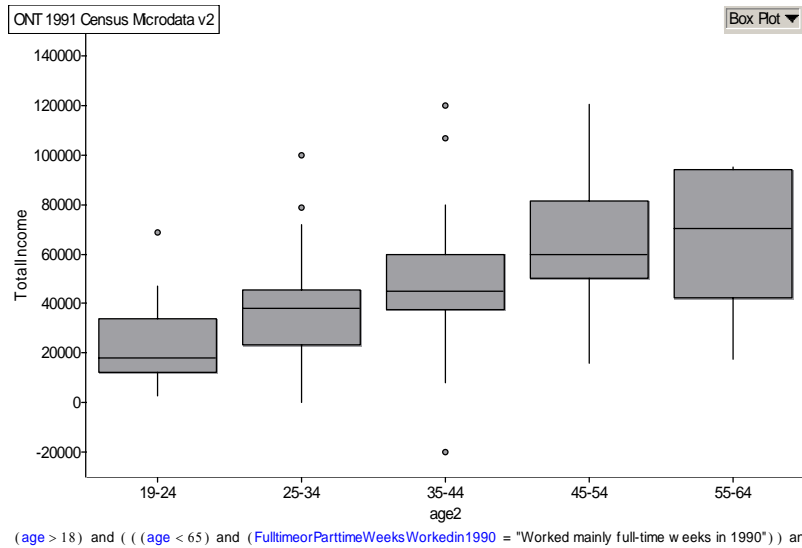
- *In general what happens to the mean across the age cohorts?*
- *On average, how much can university graduates expect to make in the decade before they retire?*
- *Do you believe that the same trend will exist for high school graduates? Why?*



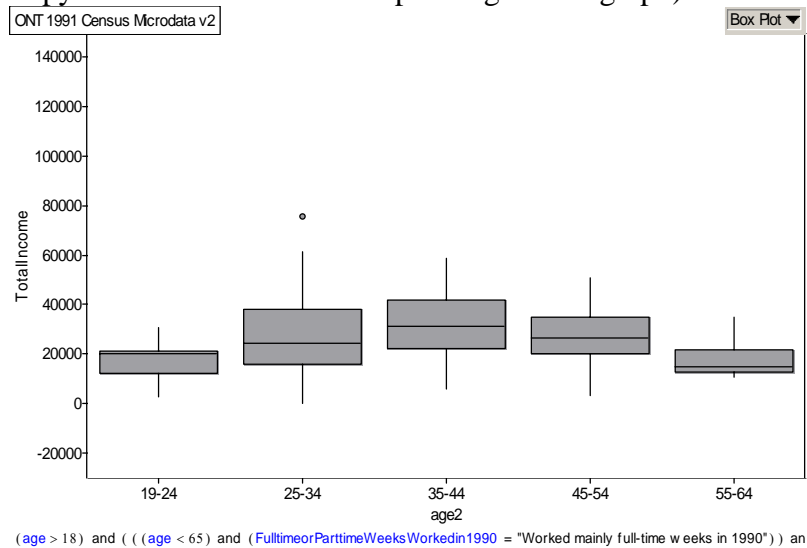
### Box & Whisker Plots

37. In the top right hand corner of the graph select **BOX PLOT** from the pull down menu.

- *What are the 5 pieces of information that a box plot represents?*
- *What line represents the median? What happens to the median across the age cohorts? How is the median different than the mean? Which value is more useful to represent our data and why?*
- *In 1990, what might a university graduate aged 51 expect to earn if she were employed full-time?*



38. By using your knowledge of box plots and data filters, create a box plot to examine the trends between age and total income, for individuals employed full-time, aged 19-64, but who's highest level of schooling is "high school" (again, you can copy the filter from the corresponding scatter graph).



- *What happens to the median income across age cohorts? Speculate on factors that may contribute to this trend.*
- *How does this median income compare with the median income for university graduates?*
- *Identify 3 other types of analysis that you could perform using the microdata provided, that may lead you to a better understanding on why we observe this trend.*

## Using a Summary Table to Identify Relationships Between Schooling & Income

39. Drag a new summary table onto the workspace
40. Drag the **Age2** attribute onto the Row. Drag the **schooling** attribute onto the column.

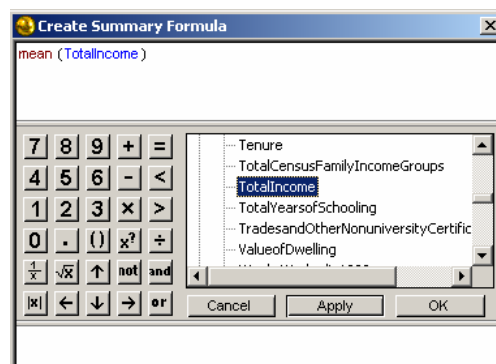
ONT 1991 Census Microdata v2		Summary Table				Row Summary
		schooling				
		highschool	less than high school	less than university	university	
age2	18 and under	5	256	2	0	263
	19-24	19	15	33	12	79
	25-34	31	37	80	41	189
	35-44	20	38	59	37	154
	45-54	19	34	40	17	110
	55-64	9	41	26	9	85
	65-74	12	42	16	5	75
75+	4	26	12	4	46	
Column Summary		119	489	268	125	1001

S1 = count ( )

- What information does this table provide?
- How many university graduates are in the 55-64 age cohort? How might this affect our confidence in the trends we observed in the box plot?
- Summary tables cannot be filtered. Compare the numbers provided in the summary table with the numbers provided by the filtered histogram. Provide an explanation for the differences you observe.

## Calculating Mean Income

41. Right click on the summary table, and select **ADD FORMULA**.
42. In the dialogue window, double click on **functions**, then double click on **statistical**, double click **One Attribute**, double click **mean**. “mean ()” with mean in burgundy should appear in your dialogue box. Between the parenthesis type, TotalIncome. (Alternatively, you can select TotalIncome under the Attribute heading).



43. Click **APPLY**, and then **OK**.

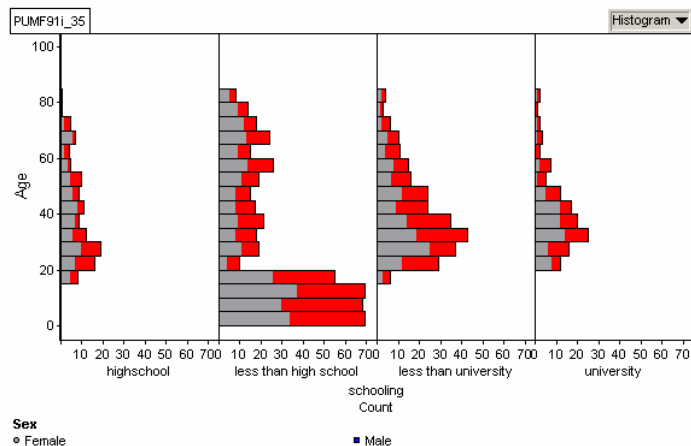
ONT 1991 Census Microdata v2		Summary Table				
		schooling				Row Summary
age2		highschool	less than high school	less than university	university	
	18 and under	5 3284.8	256 1723.82	2 9750	0	263 2142.3684
	19-24	19 15125.263	15 13247.533	33 11901.424	12 20222.417	79 14196.316
	25-34	31 21657.355	37 17210.378	80 24585.775	41 33973.61	189 24698.106
	35-44	20 23840.85	38 20581.842	59 32943.424	37 40818.973	154 30603.188
	45-54	19 20235.421	34 22925.882	40 33929.325	17 51751	110 30917.209
	55-64	9 21254.778	41 19644.854	26 32268.192	9 56278.111	85 27555.388
	65-74	12 16246.833	42 17977.31	16 24896.188	5 29066.2	75 19915.72
	75+	4 33566	26 13474.308	12 22044.444	4 59422	46 21411.256
Column Summary		119 19806.647	489 15526.841	268 26851.532	125 39321.424	1001 23714.547

S1 = count ( )  
S2 = mean (TotalIncome)

- *What information does the table provide?*
- *This information is not filtered. How might factors such as part-time work be affecting the numbers reported?*
- *With these biases in mind, what trends do we observe? Do your observations support or discredit the trends we observed in the box-plots and scatter graphs?*
- *Provide an explanation why university graduates aged 75+ have a higher annual income than their colleagues in the 65-74 age cohort.*

## Extensions

1. Using the skills you've learned in this activity, perform similar analysis with the "less than university" and "less than high school" groups. [schooling attribute]. Make comparisons among the 4 different groups.
2. Using 1990 figures, estimate the financial cost of a university education. Make a statistically informed estimate on a university graduate's total income over a lifetime. Clearly state what assumptions you are making in your estimate. Compare the university graduate's lifetime earnings with the lifetime earnings of a person who's highest level of schooling is less than university, high school and less than high school. Create a poster that would explain to grade 10 students the "costs" of dropping out of school.
3. Statistics Canada conducted a similar type of analysis using information from the 1996 Census. Instead of only examining Ontario citizens, this study examined the whole Canadian population. A description of the study is available online at <http://www.statcan.ca/Daily/English/980512/d980512.htm>
  - a. Examine the tables in this study. Do they support the conclusions you made in your analysis?
  - b. This study also examined specific populations – visible minorities, aboriginal people, occupations, women, families and regional. Using the 1991 Census microdata available to you, perform a similar type of analysis on one or more of these unique populations. (You can choose another population if you choose, for example Chinese speakers only).
4. Examine whether there are any significant differences between men and women.
  - a. Create a split-Histogram, which represents age on the y axis, and schooling on x-axis, and is split with between the sexes.
    - i. Create a new graph. While depressing the CTRL key drag the "age" attribute onto the y-axis.
    - ii. Drag the "sex" attribute on top of the plot → you should see a BOX highlighted around the plot.
    - iii. Drag the "schooling" attribute onto the x-axis.
    - iv. Double click on **Male**. Parts of your histogram should be highlighted in red. Examine your other diagrams. All the male cases will be represented in red. *Do you observe any significant differences between the sexes? What factors may account for these differences?*
    - v. You may apply filters to this histogram if you desire.



5. Evaluate the validity of the analysis conducted in this activity. Identify sources of bias and suggest ways to improve the sample. How confident can we be that the trends we observe in the 1990 data are applicable today in today's economy? Why?
6. Examine the 40 attributes provided in the 1991 Census microdata file. Find a relationship between two or more of these attributes. Use filters, and/or the switch function to examine the relationships between different groups. You can use scatter graphs, box plots, histograms, tables or other methods to communicate the relationship that you have discovered.

File: U:/yanjoel/Queens\_Manny\_Avila/MDM4U income and education activity v5.doc  
 And D: Fathom/1991 Census/  
 J: 1991 Census Microdata/  
 Updated: 6 March 2003 by Manny Avila  
 Updated: April 28, 2003 by Joel  
 September 18 by Joel to reflect updated microdata