

How to use Census data on E-STAT with *Fathom* for a data analysis project

Is it worth completing university?

What the Census tells us about the correlation between education and income
(using municipal level data)

by Joel Yan, Statistics Canada, mdm4u@statcan.ca, 1-800-465-1222

Assignment

Is it worth completing university? Investigate the correlation between the proportion of people with a completed university degree and average income using census data for municipalities in an area of Canada. We do this on E-STAT by creating a scatter plot of the relationship between average incomes versus the percent of the population with a university degree for a group of municipalities within one economic region. Then we import the data into *Fathom* in order to compute the regression equation and correlation statistics.

E-STAT has over a thousand census characteristics for each municipality in Canada that can be manipulated within E-STAT and imported into *Fathom* to explore relationships. This activity is intended to provide an example of how this can be done.

Related Expectations for the Ontario Grade 12 Mathematics of Data Management Course:

- Solve problems involving complex relationships with the aid of diagrams (ODV.02, *Organization of Data for Analysis – Overall Expectations – page 49 in the Ontario curriculum*)
- Describe the relationship between 2 variables by interpreting the correlation coefficient (STV.04, *Statistics – Overall Expectations – page 52*)
- Organize and summarize data from secondary sources (e.g. the Internet) using technology (ST1.04, *Statistics – Specific Expectations on Collecting Data – page 52*)
- Calculate the correlation coefficient for a set of data using statistical software (ST4.02, *Statistics – Describing the relationship between 2 variables – page 52*)

Statistics Canada and Fathom

This lesson plan was prepared by Statistics Canada to facilitate the use of Statistics Canada data by teachers of the Ontario mathematics curriculum. This lesson requires the use of the *Fathom* software. *Fathom* is licensed by the Ontario Ministry of Education and used by schools across Ontario. Use of *Fathom* in this lesson is in no way an endorsement or recommendation of the *Fathom* software by Statistics Canada.

Procedure

Finding the right table using Search Census

1. Start up E-STAT - at <http://estat.statcan.ca>
2. From the left tool bar click **Search Census**.
3. Under **Census**, select “2001 Census” and click “Go!”
4. Select “2001 Census of Population (Provinces, Census Divisions, Municipalities)” and click “Go!”
5. Under **Profile selection**, select “2001 School Attendance, Education, Field of Study, Highest Level of Schooling and Earnings” and click “Go!”.

Geography and Characteristics selection

6. Under **Geography**, scroll way down and select “Census subdivisions in Ontario - 2001 - Kitchener, Waterloo, Barrie - 44 areas” or another group of municipalities anywhere in Canada.
7. Under **Characteristics**, click “View Checklist”
8. Select the following variables from the checklist:
 - “Total population 20 years and over by highest level of schooling”
 - “With bachelor’s degree of higher, university, population 20 years and over by highest level of schooling”
 - “Average employment income, \$, worked full year, full time, population 15 years and over with employment income”

Tip: Because the list of variables is long, use the “Find on this page” key from the Edit pull-down menu on your browser. For example, do a find on ‘schooling’, and then on ‘income’

9. Click the **Home** or **End** keys, and click on “Back to Main Selection Form”

Expressing Data as Percentages

10. E-STAT has basic computation and mapping capability. To see the geographic distribution of the first variable as a percent of the total population, we first click the “Table, Areas as rows” icon at the bottom of the screen. We then click the radio button under the table “Data as % of 1st characteristic” and click the “Redisplay As” tab. This produces a table showing the schooling level as a percentage of the total population. Note that when we compute the percentages of the total population by level of schooling, the income values in the table remain unaffected. This is because E-STAT recognizes which variables are absolute counts (e.g. the schooling level) and which are relative variables (e.g. income) and computes percentages based on the data type. For details, click on the Help button on the left side bar, under “Search Census”.

Sorting the Data

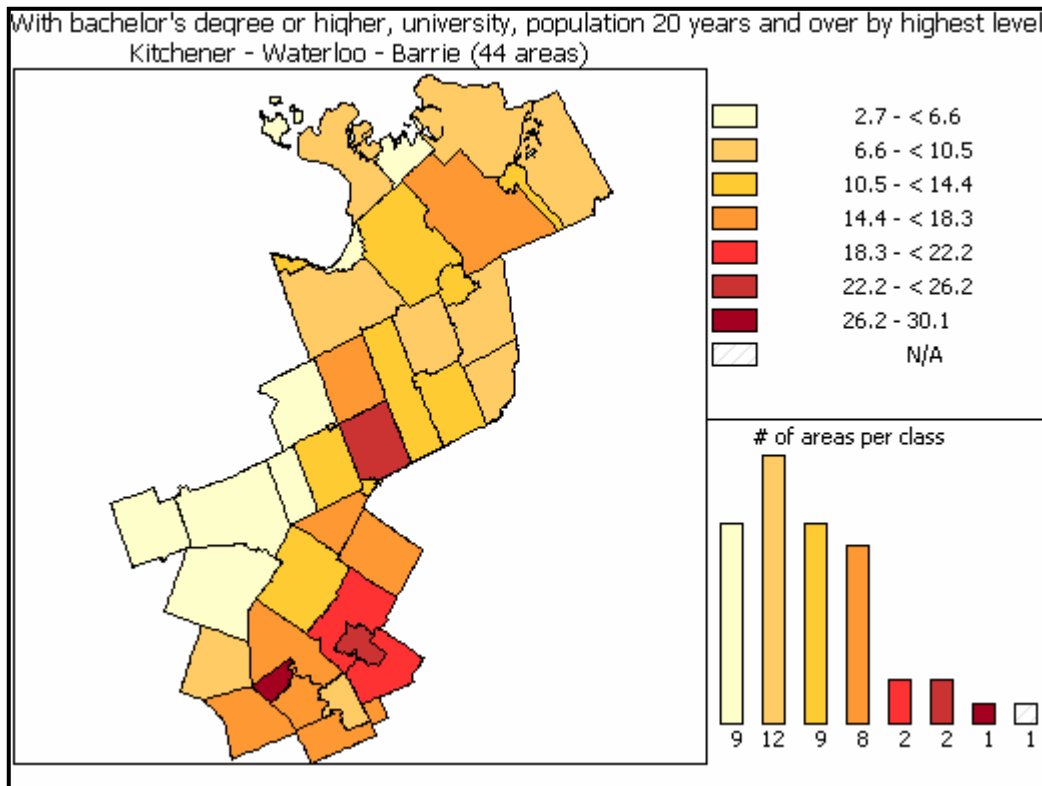
11. We can also sort the table in ascending order by any of the characteristics in the table. For example, to sort by average employment income, we click the radio button “Sort on data for the 2nd characteristic” and then click the “Redisplay As” tab.

Questions:

- (a) What are the minimum and maximum values for average employment income by census subdivision for the selected areas? _____
- (b) How might education level affect the average employment income?

Mapping the Data

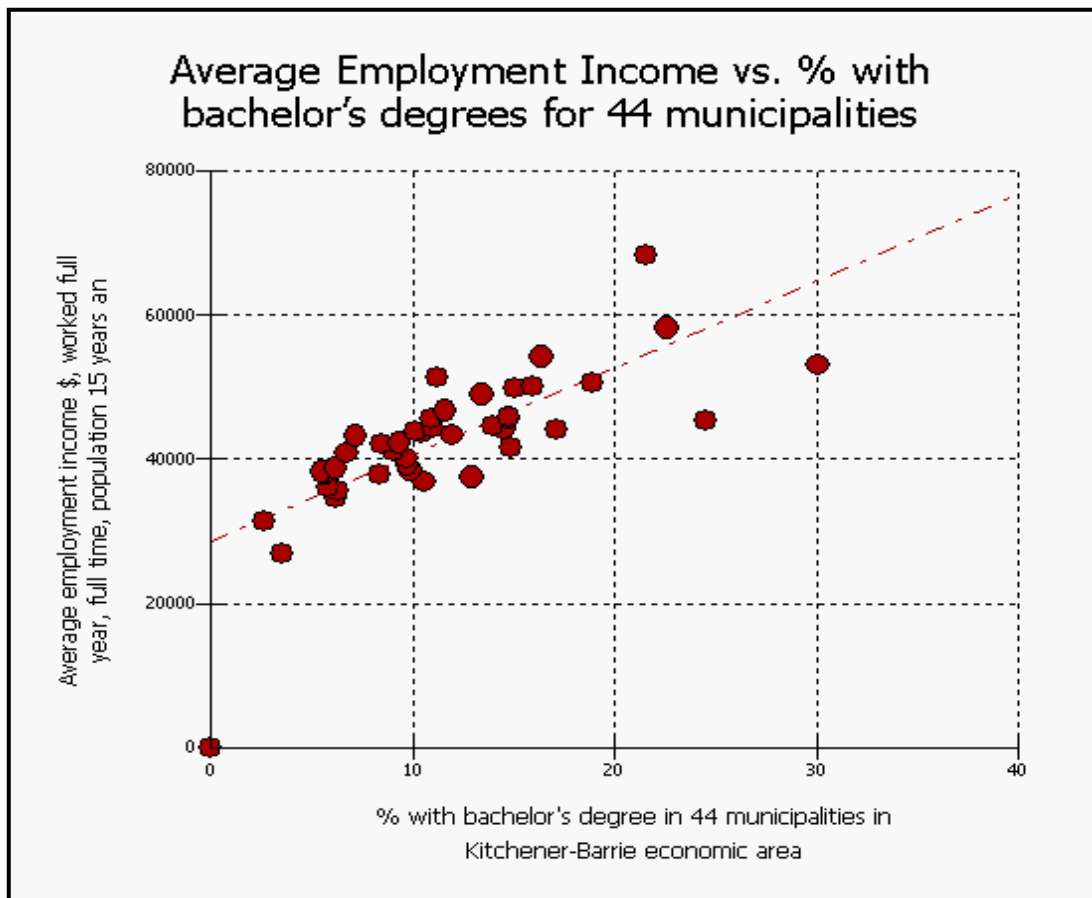
12. Maps help us visualize and better understand the areas for which we are looking at data. To produce a statistical map, we click the Map icon under the graph. By default this produces a colour shaded map of the first characteristic using 7 classes, as shown below. This type of map can be useful in identifying census subdivision areas with the selected characteristics. Furthermore, the frequency histogram to the lower right of the map can be used to assess the distribution of the data for the selected areas (e.g. normal, skewed, bimodal). E-STAT offers three different types of maps (proportional circle, dot, and shaded) and several options for selecting class intervals and colours. For details, check the Help on the left side bar under “Search Census”.



Analysis question: After looking at the frequency histogram to the right of the map, describe the distribution of census subdivision values for schooling level within the selected urban centre.

Producing a scatter graph

13. E-STAT can also produce a scatter graph, showing the relationship between any two variables. In this exercise, we want to see if there is a relationship between schooling levels and average employment income for census subdivisions. To produce a scatter graph with the line of best fit shown, click the “Scatter Graph with Line of Best Fit” icon at the bottom of the screen.
14. If desired, we can modify the graphs labels, by clicking ‘Modify Graphic’ under the graph. For example, enter “Average Employment Income vs. % with bachelor’s degrees for 44 municipalities” in the Title box. Then enter “% with bachelor's degree in 44 municipalities in Kitchener-Barrie economic area” in the Subtitle box. Check the box “Show grid on chart”. Click on ‘Redraw’. The resulting scatter graph is shown below.



15. Questions involving analysis of this graph:

- (a) Is the line of best fit a reasonable model for predicting the behaviour of average income as a function of % of university graduates within an area? _____
- (b) What does the slope of this line of best fit represent? _____
- (c) What is the approximate slope of the line for your selected centre? _____

Hint: Additional details on [Scatterplots](#) are available in [Statistics: Power from Data!](#).

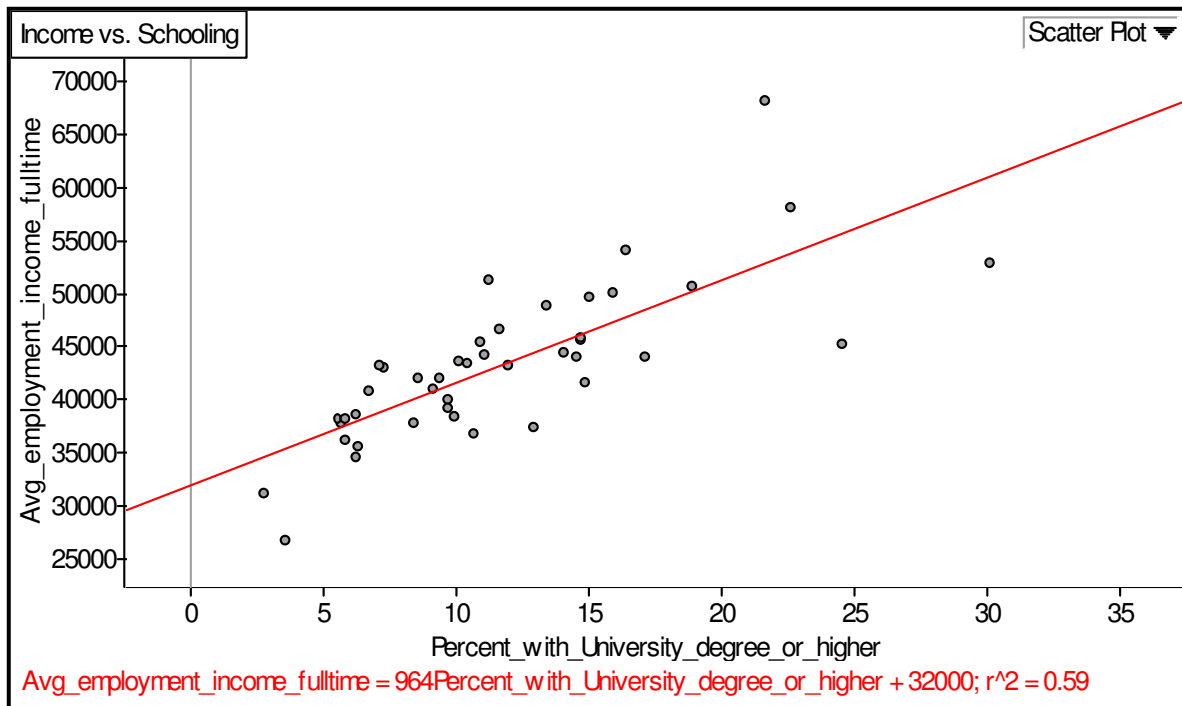
16. E-STAT will let us easily produce scatter plots with lines of best fit for assessing potential relationships between any two variables. However, E-STAT can not currently output the equation of the line of best fit or the r^2 value. To obtain these statistics we can export the data from E-STAT to another package, such as a spreadsheet, graphing calculator, or analytical software such as *Fathom*.
17. Scroll down and select from **Screen outputs**: “Spreadsheet.WK1 Geo=Rows”
18. Within the **Downloading** box select “Open”.
19. When the data opens within your spreadsheet program, highlight columns D & E. From the **Edit** menu, select *Copy*.

Load the Data into Fathom

20. Switch to *Fathom*. (If *Fathom* isn't already running, you will need to launch it.)
21. In a new document, make a new empty collection.
22. With the collection selected, chose ***Paste Cases*** from the **Edit** menu.
23. Make a case table for the collection (for example, by choosing ***Case table*** from the **Insert** menu)
24. Since the attribute names are quite long, it is helpful to shorten them. In this case, we shortened the names to Percent_with_University_degree_or_higher, and Avg_employment_income_fulltime.
25. Change the name of the collection to ‘Income vs. Schooling’
26. Save your *Fathom* document by choosing ***Save*** from the **File** menu.

Graphing the Data

27. Make a graph showing the variation in *Average income vs. the schooling level*.
28. Using the **Graph** pull-down menu overlay a ***Least-Squares Line*** (as shown on the next page). This gives you the equation for the line of best fit, as well as the coefficient of determination (r^2).



Write a description of the pattern you see on this graph. Write a possible explanation for the shape observed in this distribution.

29. You can also add a residual plot to help you assess the quality of the fit of the least squares line.

30. Repeat the entire process and analysis using the percent of the population with less than grade 9 education. This will involve going back to E-STAT to retrieve new data.

Note: This analysis could be repeated at any geographic level. Using E-STAT for example, you could extract the same variables for the 49 counties, districts, and regional municipalities in Ontario and repeat the analysis. A separate teaching activity (*lesson 3a*) is available which walks through the steps using census tract level data for a large urban centre of over 50,000 population (e.g. Oshawa).

For further analysis of the relationship, look at the 2001 Census results on the relationship between education and income. You will find this at <http://www12.statcan.ca/english/census01/products/highlight/Earnings/Index.cfm?Lang=E>. Select Census on the top blue bar, the Data on the left bar, then Highlight tables.