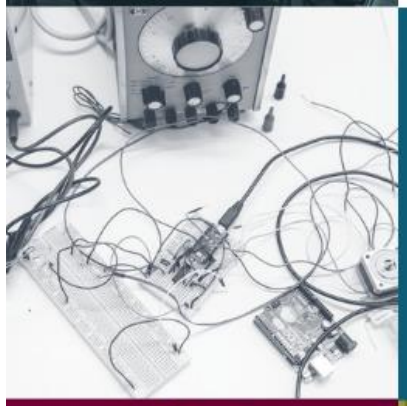
Sherman Centre for Digital Scholarship



Microdata Analysis using Python

Presenters: Vivek Jadon

Date: November 27, 2025

McMaster University Library

Land Acknowledgement

McMaster University is located on the traditional Territories of the Mississauga and Haudenosaunee Nations, and within the lands protected by the “Dish With One Spoon” wampum agreement.

Code of Conduct

The Sherman Centre and the McMaster University Library are committed to fostering a supportive and inclusive environment for its presenters and participants.

As a participant in this session, you agree to support and help cultivate an experience that is collaborative, respectful, and inclusive, as well as free of harassment, discrimination, and oppression. We reserve the right to remove participants who exhibit harassing, malicious, or persistently disruptive behaviour.

Please refer to our code of conduct webpage for more information: scds.ca/events/code-of-conduct

Session Recording and Privacy

This session is being recorded with the intention of being shared publicly via the web for future audiences. In respect of your privacy, participant lists will not be shared outside of this session, nor will question or chat transcripts.

Questions asked via the chat box will be read by the facilitator without identifying you. Note that you may be identifiable when asking a question during the session in an audio or visual format.

Certificate Programs

The Sherman Centre for Digital Scholarship Certificate of Attendance

The Sherman Centre's certificate program recognizes attendance at our workshops. It complements degree training, supports the development of critical competencies in data analysis, research data management, and digital scholarship, and formalizes core skills fostered by our workshops.

Participants are invited to attend seven workshops and receive a certificate of attendance. To verify your participation in today's workshop, we will provide a code and additional instructions at the end of the session.

You can learn more about the certificate program at **scds.ca/certificate-program**

The Canadian Certificate for Digital Humanities

This workshop is also eligible for the Canadian Certificate for Digital Humanities. To learn more about the certificate, visit **ccdhhn.ca**. You can also contact local liaison Alexis-Carlota Cochrane at scds@mcmaster.ca

Book an Appointment with the DASH Team

Receive help from a member of the DASH team! DASH can assist with the following topics:

- Creating data visualizations, including charts, graphs, and scatter plots
- Figuring out which statistical tests to run (e.g., t-test, chi-square, etc.).
- Analyzing data with software including SPSS, Python, R, SAS, ArcGIS, MATLAB, and Excel
- Choosing which software package to use, including free and open-source software
- Troubleshooting problems related to file formats, data retrieval, and download
- Selecting methodology and type of data analysis to use in a thesis project

Book an appointment: **https://library.mcmaster.ca/services/dash**

scds.ca
scds@mcmaster.ca

Winter 2025: Upcoming Workshops

Data Analysis Support Hub

January 15: Introduction to R Programming

January 20: Map Making for Absolute Beginners using QGIS

Digital Research

February 6: Create a Digital Exhibition with Omeka S

Research Data Management

November 27: Connecting the Research Ecosystem with Persistent Identifiers: Research Data Management Community of Practice

January 14: Best Practices for Managing Data in your Research

Do More with Digital Scholarship

January 22: Conducting Meta-Analysis for Systematic Reviews Using R

January 27: Streamline Your Research Materials Photos with Tropy

Register for Upcoming Workshops: https://u.mcmaster.ca/scds-workshops

Library | McMaster University | Sherman Centre for Digital Scholarship

Learning Objectives

By the end of this workshop, you will:

- Gain knowledge of the wide assortment of sample surveys available for research use
- Explore various statistical techniques using Python
- Gain practical experience in analyzing and interpreting data using Python

What is Microdata?

- In **statistics**, **microdata** refers to **individual-level data** collected from surveys, censuses, or administrative sources.
- It contains detailed records for each unit of observation (e.g., individuals, households, businesses) rather than aggregated summaries.

Key Characteristics of Microdata

- **Unit-level Data** – Each record represents a single entity (e.g., person, household, company).
- **Raw & Unaggregated** – Unlike summarized statistics (macrodata), microdata retains detailed information.
- **Used for Analysis** – Researchers use microdata to explore relationships, trends, and correlations.

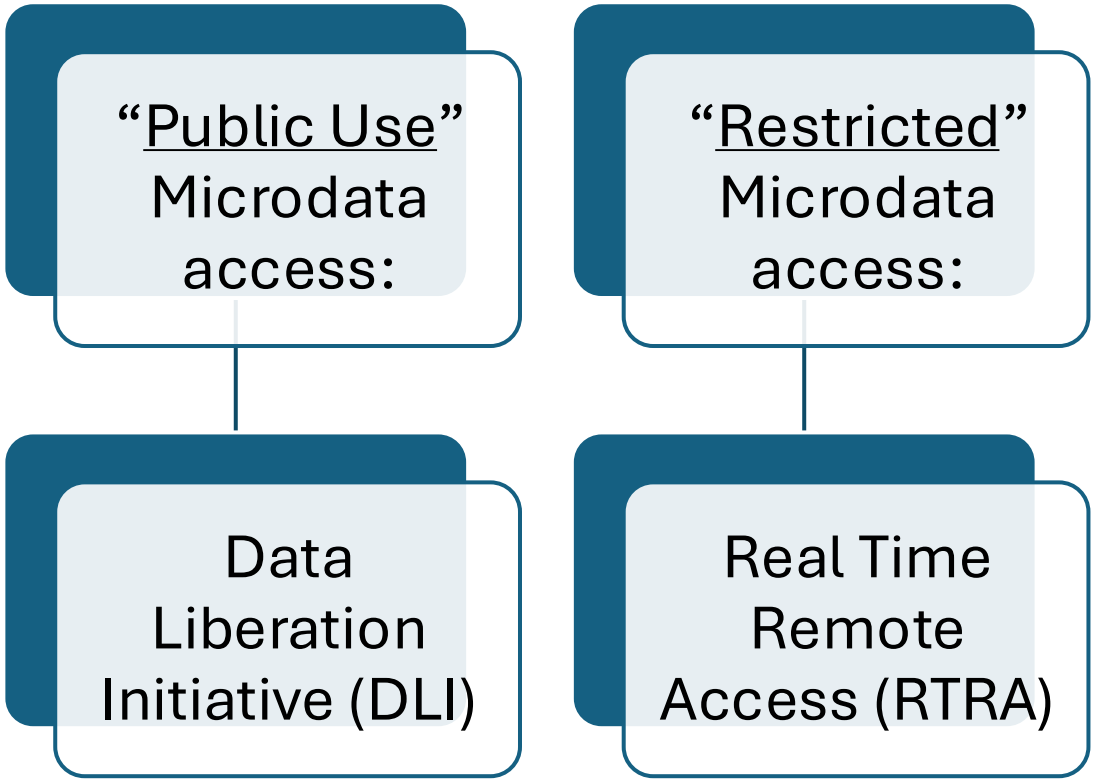
Example of Microdata (Survey Data on Employment)

ID	Age	Gender	Occupation	Income ($)	Education Level
001	30	Male	Engineer	75,000	Bachelor's
002	45	Female	Teacher	50,000	Master's
003	29	Male	Doctor	100,000	PhD

Uses of Microdata

- **Policy Analysis** – Governments use it to make data-driven decisions.
- **Economic & Social Research** – Helps study trends in income, employment, and health.
- **Machine Learning & AI** – Used for predictive modeling and trend analysis.

Microdata Access Programs



Data Liberation Initiative (DLI)

- DLI provides access to Stat Can's Standard products, Databases, Public use Microdata files and Geographic information files.
  - Main focus is on Socio-Economic data: Health, Education/Literacy, Labor Market, Income, Travel, Justice, Census of Population etc.
  - Databases such as the Small Area Business and Labour Database, Inter-Corporate Ownership, Financial Performance Indicators, Trade data etc.
  - An enhanced line of Census products
  - Aggregated data on subject such as Justice and Education
  - $5,000 worth of custom tabulations from Stat Can's Data Service Centres
  - All standard geographic files and databases including PCCF & PCCF+

Data Liberation Initiative (DLI)

- Metadata from DLI surveys are marked-up in DDI/XML format for discoverability at the variable level from Odesi Data Portal.
- DLI members have support through a very active listserv.
- Currently over 77 subscribing institutions -- McMaster University Library is part of Stat Can DLI program.

Public Use Microdata File (PUMF) Access

- Each Public Use Microdata File is based on a corresponding master data file. The modifications performed by Statistics Canada before the PUMF is released ensure that the risk of breaching confidentiality has been removed. Since the results of any analysis performed do not have to be scrutinized before they are released, the file is considered “Public”.

- Modifications made to the Master files for conversion to PUMFs may include **collapsing of variables** (e.g., age groups instead of individual years of age); **collapsing variables into one variable** (e.g., multiple language questions collapsed into one language variable for analysis); **suppressing variables** (although the variable is part of the master file, it will not show up in the public file); and **removing outliers** (removing cases that are extremes - often used with income).

- By using these techniques to anonymise the files, combining variables will not result in the user identifying a respondent from any given survey.

Public Use Microdata File (PUMF) Access

- **Benefits**
  - Free
  - Very few restrictions on access & use of the data
  - No approval process to access the data
- **Limitations**
  - Content is limited (screened and grouped for confidentiality)
  - Not all surveys have a PUMF
  - PUMFs are cross-sectional, i.e., represent data collected at one point in time

Most Popular PUMF Collections

- Canadian Community Health Survey (CCHS)
- Labor Force Survey (LFS)
- Census of Population
- Discharge Abstract Database (DAD) -- CIHI
- General Social Survey (GSS)
- Survey of Household Spending (SHS)
- Canadian Tobacco, Alcohol and Drugs Survey (CTADS)
- National Household Survey (NHS)
- PCCF & PCCF+

Inter-University Consortium for Social and Political Research (ICPSR)

- **ICPSR**: http://www.icpsr.umich.edu/
  - ICPSR maintains and provides access to a vast archive of social science data for research and instruction.
  - The ICPSR data contain mainly US survey statistics with some international content. The thematic categories lists data holdings on 17 broad subject areas such as education, health, justice etc.
  - The ICPSR data holdings can be searched, downloaded and analyzed online. Users can either create a MyData account on ICPSR website or can login using Google/LinkedIn/ORCID accounts to download or analyze data online. Only authorized McMaster Users (current Faculty, Staff and Students) are allowed access to the data.
  - ICPSR supplies data files for use with statistical software, such as SAS, SPSS, and Stata.

Odesi Data Portal

- **Odesi…**
  - is a web-based data extraction system delivered through Scholars Portal, provides access to diverse, quality, numeric data sets including microdata survey collection from DLI, demographic data from Statistics Canada and polling data from Gallup and other organizations.
  - facilitates the exploration (searching, data manipulation, creation of summary statistics, graphing and export) of multiple, sophisticated data sets.
  - Access is open to all OCUL institutions and is controlled by IP address
  - https://odesi.ca

Other Microdata Resources

- **Data and Statistics Collection:**
- https://library.mcmaster.ca/data-statistics-collection

Jupyter Notebook

- McMaster has access to Jupyter notebook via Compute Canada

- https://mcmaster.syzygy.ca/

Download Exercise File

- http://bit.ly/2MVaTmv

DASH GitHub Repository

- https://library.mcmaster.ca/services/dash#tab-dash-github-repository

Sherman Centre for Digital Scholarship

SCDS Links

Send SCDS an Email:

scds@mcmaster.ca

Register for a Workshop:

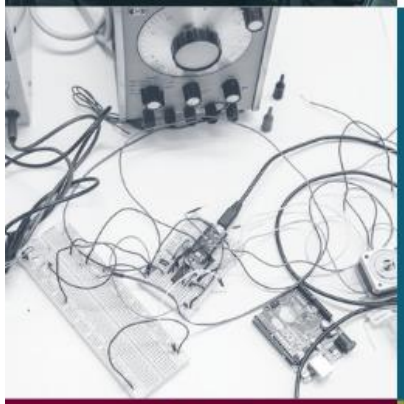
https://u.mcmaster.ca/scds-workshops

Subscribe to our Newsletter:

https://u.mcmaster.ca/sign-up

Schedule a Consultation:

https://libcal.mcmaster.ca/appointments



McMaster University | Library