Datasheet for Toronto Neighbourhood Profiles Data*

A deeper look into the 2021 Canadian Census Data

Julia Lee

December 14, 2024

The following contains information about the Toronto Neighbourhoods Profile data used within the present analysis. Extract of the questions are provided by Gebru and colleagues (2021) and can be found at https://doi.org/10.1145/3458723.

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
    - The Neighbourhood Profiles dataset was created to gain a better understanding of the populations living in the city of Toronto. By highlighting various demographic and socio-economic characteristics such as education, income, and employment of Toronto residents, this data can be used to understand how Toronto is changing and to better serve all Toronto residents.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - This dataset is a collection of data from the 2021 Canadian census conducted and provided by Statistics Canada. To organize population information for Toronto specifically, the Social Development, Finance & Administration of Open Data Toronto created this dataset.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
    - The creation of the dataset is funded by the city of Toronto, and it is made accessible to everyone through the city's open data portal.

*Code and data are available at: https://github.com/jjlee-lee/Toronto_Neighbourhood_Income.git

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

   - The instances that comprise the dataset represent Toronto residents and the various demographic and socio-economic characteristics they had at the time of the 2021 Census.

2. *How many instances are there in total (of each type, if appropriate)?*

   - There are over 2000 categories (e.g. education attainment, household size, etc.) of data provided for each of the 158 neighbourhoods in Toronto.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

   - The short-form census is sent out to all households in a given region, and the long-form census is sent out to only 25% of households in a given region. With this, the 2021 census data reflected in the Neighbourhood Profiles data can be a representative sample of all residents in a specific region.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

   - Each instance consists of unprocessed text in an excel file.

5. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

   - Missing data values can be imputed and the aggregation of individual household characteristics or experiences can lead to incorrect assumptions about certain regions and their populations.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

- The data for the Canadian census is collected through a combination of surveys (i.e. short and long form) where residents or households in Toronto answer questions related to demographic, social, economic, and cultural characteristics. The responses collected from the census are compared with data from prior censuses and government records to verify the data and ensure consistency along with validity within the data aggregation process.

2. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

   - The long form survey within the Canadian census employs a stratified systematic sampling design to collect data on 25% of households within a given region. For the Neighbourhoods Profiles data, 25% of private households in Toronto (who were sampled from a list of dwellings) would have received a long form survey.

3. *Who was involved in the data collection process (for example, students, crowdworkers, contractors)?*

   - The data collection process is completed by Statistics Canada every 5 years.

4. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

   - The census data was collected in 2021 and has a reference data of May 11, 2021. The instances that are found within the dataset are a reflection of the demographic, socio-economic, and cultural characteristics of Toronto households during 2021. Each instance further captures the change in these characteristics since the previous Canadian census in 2016.

5. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

   - Statistics Canada conducts formal reviews on the data that is collected through the census to ensure data quality alongside an ethical collection and use of the data. Further details on their exact procedures and guidelines can be accessed through the following links:
   - https://www12.statcan.gc.ca/census-recensement/2021/ref/98-26-0001/2020001/007-eng.cfm
   - https://www.statcan.gc.ca/en/trust/integrity

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of*

instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

- Once the census has been completed by households and returned to Statistics Canada, a process to match electronically submitted, phoned, and mailed survey responses with dwellings and geographic areas begins. While responses submitted online or provided through a telephone interview are automatically registered, mailed responses need to be manually registered to the Data Operations Centre (DOC). A more detailed process is outlined at the following link:
- https://www12.statcan.gc.ca/census-recensement/2021/ref/98-306/2021001/chap3-eng.cfm

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

    - Raw data for the Canadian census is not provided. Census data that has been aggregated across different geographical regions, like the Toronto Neighbourhood Profiles data, is available.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

    - As census data is widely available for anyone to view and analyze, the data has been used by several individual researchers and organizations to better understand social phenomenon and the locations in which social events occur.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

    - No, as the Toronto Neighbourhoods Profile data is open data, it would be challenging to maintain a record of all the ways in which it was used and analyzed.

3. *What (other) tasks could the dataset be used for?*

    - From education attainment to household size, the Canadian census includes a wide range of variables for different regions across Canada. With this, the Toronto Neighbourhoods Profile data could be used to inform decisions in urban planning, healthcare, education policies, and beyond.

Distribution

1. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

    - The Toronto Neighbourhoods Profile data is open data, meaning that it is accessible to all and can be easily downloaded by users. With this, the dataset is widely distributed through the city of Toronto's online open data portal.

2. *Any other comments?*

    - It is important to note that while the dataset is open to all and for any type of analysis, every user has a responsibility to use the data in ethical and respectful ways.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

    - The dataset found on the city of Toronto's open data portal will be supported and maintained by the city along with the Social Development, Finance & Administration division (the division that published the data to the portal).

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - The Social Development, Finance & Administration division can be contacted via email. An email address is provided on the open data portal's website where the dataset can be downloaded (https://open.toronto.ca/dataset/neighbourhood-profiles/).

3. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

    - As the dataset is a collection of information from the 2021 Canadian census, it will not be updated to add or remove new instances.

4. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

    - The data provided by residents and households who completed the 2021 census will be released to Library and Archives Canada (LAC) after 92 years. This release of information currently does not require their consent. For further information, please refer to the following link:
    - https://www12.statcan.gc.ca/census-recensement/2021/ref/personalinfo-renpersonnels-eng.cfm

5. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

    - Older versions of the census, including information for Toronto, are maintained by the Library and Archives Canada (LAC). Online or microfilm copies of census records dating back to 1926 can be accessed through the LAC.