Datasheet for 'Canadian Census 2021'*

Aviral Bhardwaj

April 19, 2024

Extract of the questions from @gebru2021datasheets.

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 dataset was part of the Canadian Census 2021. The dataset was created to provide information about the population of Canada, and report other population related metrics. Table 43-10-0010-01 was created as a part of Special Interest Tables, which are released as part of the Census Program Data. This combined Canada Census data as well as Longitudinal immigration database (imdb).
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - Statistics Canada
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.*
    - Canadian Government
4. *Any other comments?*
    - TBD

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.*

*Code and data are available at: https://github.com/Aviral-03/canada_immigrant_income_distribution/tree/main

- The instances in this dataset represent demographic information, such as population counts, immigration data, and other related metrics for various regions within Canada. It includes details about age, gender, ethnicity, education, employment, and other demographic factors.

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

   - In total 63 census subdivisions defined.

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).*

   - This dataset is a sample of the larger set of Canadian Census data. Subdivisions were selected based on the appropriateness of data, such as provinces, territories, and census metropolitan areas subdivision were removed, and other census not important for the analysis was also removed. The selection was based on the need to provide a representative sample of the population which was 25%.

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 likely consists of demographic features such as population counts, age distributions, immigration statistics, education levels, employment rates, and other relevant demographic indicators. These may be presented in tabular format or as structured data. These are all

5. *Is there a label or target associated with each instance? If so, please provide a description.*

   - Yes, a unique identifier is associated with each instance, which is the census subdivision code. This code is unique to each census subdivision and is used to identify the region to which the data pertains.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

   - No information is missing from the instances. The dataset is complete and contains all the relevant demographic information for each census subdivision.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

- Yes, a new README file was created to explain the relationships between the instances. The README file provides information about the structure of the dataset, the meaning of each column, and how the data is organized. This helps users understand the relationships between the instances and how they can be used for analysis.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

    - No

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

    - No

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - No, the dataset is self-contained and does not rely on external resources. All the data is included in the dataset itself, and there are no external links or dependencies. This makes it easy to use and analyze the data without any additional requirements.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - Yes, the dataset contained information regarding the Income of the population, which is considered confidential. Therefore, only the aggregated data was included in the dataset, and no individual-level information was provided. This ensures that the privacy of individuals is protected and that the data is used in a responsible manner.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - No

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

- Yes, the dataset likely identifies subpopulations based on various demographic factors such as age, gender, ethnicity, education level, employment status, and immigration status. These subpopulations are typically identified through the breakdown of demographic statistics for different demographic categories within each geographic region (i.e Provinces in Canada)

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - Direct identification of individuals from this dataset is unlikely since it deals with aggregated demographic statistics at a regional level.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - No

16. *Any other comments?*

    - TBD

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 associated with each instance in the Census Canada dataset was acquired through self-reported survey responses from individuals residing in Canada. Participants provided information about various demographic factors, including age, gender, ethnicity, education, occupation, and household composition. The data collection process involved direct reporting by subjects, and efforts were made to ensure data accuracy through validation and verification procedures by Statistics Canada.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

- The data collection for the Census Canada dataset primarily relied on manual human curation through survey forms distributed to Canadian residents. Statistics Canada employed various mechanisms, including online surveys, paper questionnaires, and assisted interviews, to collect demographic information from respondents across the country.

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

   - The Census 2021 Canada dataset is not a sample but represents the entire population of Canada. The census aims to collect information from every individual residing in the country, making it a complete enumeration rather than a sample.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

   - The data collection process for the Census Canada involved various individuals, including enumerators, field supervisors, and support staff hired by Statistics Canada.

5. *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 data is aggregated every 5 years. The data is collected in the year of the census.

6. *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 adheres to strict ethical standards and undergoes regular reviews to ensure compliance with legal and ethical guidelines governing data collection, storage, and usage.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

   - No, the data was collected directly from individuals residing in Canada through self-reported survey responses. Statistics Canada is the primary agency responsible for collecting census data in Canada. Data was downloaded from the Statistics Canada website.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

- Yes, individuals residing in Canada were notified about the data collection process through various communication channels, including mail, online announcements, and public awareness campaigns. Statistics Canada provides detailed information about the census process, the purpose of data collection, and the importance of participation to ensure accurate and representative data.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

    - Yes, individuals were asked to provide consent for the collection and use of their data as part of the census process. Statistics Canada ensures that participants are informed about the purpose of data collection, the confidentiality of their information, and the importance of their participation in the census. Consent is obtained through various means, including online forms, paper questionnaires, and assisted interviews.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - Individuals were assured of confidentiality and privacy protections regarding their data, as outlined by Statistics Canada.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - Statistics Canada conducts thorough assessments of the potential impact of census data collection on data subjects, including privacy implications and data protection measures, which can be found in their official documentation.

12. *Any other comments?*

    - TBD

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.*

    - Refer to https://www12.statcan.gc.ca/census-recensement/2021/ref/98-304/2021001/chap8-eng.cfm for information regarding the data processing and cleaning.

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.*

    - NA

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - A few Structured Query Languages (SQL) and statistical analysis system (SAS) modules are also part of the census edit and imputation processing flow.

4. *Any other comments?*

    - TBD

Uses

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

    - As of now, there is no specific information available regarding the specific tasks for which this dataset has been used. However, given its comprehensive nature and the involvement of Statistics Canada, it's likely that researchers, policymakers, and analysts have utilized this dataset for various purposes such as demographic research, population projections, immigration policy analysis, socioeconomic studies, and urban planning.

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.*

    - Stats Canada: https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/about-apropos/about-apropos.cfm?Lang=E

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

    - The dataset could be utilized for a wide range of tasks and analyses, including:
    - Studying population trends and dynamics over time.
    - Assessing the impact of immigration on demographic composition and economic development.
    - Identifying disparities in education, employment, and income across different demographic groups and regions.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

- No. The data is structured and aggregated in a way that protects individual privacy and confidentiality. However, users should be aware of the limitations of the dataset and the potential biases that may arise from the sampling strategy or data collection methods. It is important to interpret the data in context and avoid making generalizations or assumptions that could lead to unfair treatment or misinterpretation of the results.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

    - The dataset should not be used for individual-level analysis or identification. It is not suitable for making decisions about specific individuals or groups, as it does not contain personal information or detailed demographic data at the individual level. The dataset is intended for aggregate analysis and population-level studies.

6. *Any other comments?*

    - TBD

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

    - Yes, the dataset will be distributed to third parties outside of Statistics Canada for research, analysis, and policy development purposes. Researchers, policymakers, analysts, and other stakeholders may access the dataset through official channels and data repositories.

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

    - The dataset will likely be distributed through official channels such as the Statistics Canada website, data repositories, and other authorized platforms. StatsCan r library is available for R users to access the data.

3. *When will the dataset be distributed?*

    - NA

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

    - Yes, the dataset will likely be distributed under a specific license or terms of use that govern its usage, distribution, and sharing. Users may be required to comply with

certain conditions, restrictions, or fees associated with the dataset. The specific licensing terms and conditions will be provided by Statistics Canada.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
   - No
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
   - No, general census is available to the public.
7. *Any other comments?*
   - TBD

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
   - Statistics Canada
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
   - Official contact information for Statistics Canada can be found on their website: https://www.statcan.gc.ca/eng/start
3. *Is there an erratum? If so, please provide a link or other access point.*
   - N0. There is no erratum.
4. *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)?*
   - Done by Statistics Canada
5. *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.*
   - Any applicable limits on data retention and associated enforcement mechanisms will be specified in accordance with relevant privacy regulations and guidelines.

6. *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.*

    - Old versions of the dataset may be archived and maintained for historical reference, but they may not be actively supported or updated. Any obsolescence of older versions will be communicated to dataset consumers through official channels and documentation.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

    - No

8. *Any other comments?*

    - TBD