Datasheet for ‘A dataset’[*]

Sirui Tan

2 April 2024

The dataset Upper income limit, income share and average income by economic family type and income decile was created by Statistics Canada to analyze income distribution across various economic family types and income deciles in Canada. It provides valuable insights into income disparities and trends over time, addressing the gap in comprehensive data on income distribution. The dataset comprises instances representing different socio-economic groups in Canada, with each instance corresponding to a specific combination of economic family type and income decile. It includes 183,043 instances and offers a detailed view of income distribution patterns. The dataset has been utilized for research on income dynamics, poverty analysis, labor market trends, and socioeconomic policy evaluation. It is openly accessible under the MIT License and is maintained by the Canadian government.

Extract of the questions from Gebru et al. (2021).

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 “Upper income limit, income share and average income by economic family type and income decile” was created to provide insights into income distribution across different economic family types and income deciles in Canada. The dataset aims to analyze the upper income limit, income share, and average income within each economic family type and income decile, offering a detailed understanding of income disparities and trends over time. The creation of this dataset likely aimed to address the gap in comprehensive data on income distribution, particularly at a granular level by economic family type and income decile.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

[*] Code and data are available at: https://github.com/siru1366/income.git.

- The dataset was created by the Statistics Canada, and it was also published by Statistics Canada on the Open Government Portal of 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.*

    - he creation of the dataset was funded by the government of Canada.

4. *Any other comments?*

    - The source of the dataset is the Canadian Income Survey.

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 in this dataset represent various economic family types and income deciles in Canada. Each instance corresponds to a specific combination of economic family type and income decile, providing insights into income distribution patterns across different socio-economic groups.

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

    - The original dataset has 183043 instances.

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 dataset likely contains a sample of instances rather than all possible instances within the population of economic family types and income deciles in Canada. The sample is likely representative of the larger population, ensuring coverage of diverse socio-economic groups and geographic regions.

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 is consist of the year, place, family type, income type, income value of the corresponding group of people of Canada.

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

    - NO

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.*

    - There are some information about income value from some instances because they are unavailable in the original dataset.

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.*

    - Relationships between individual instances may be implicit, with connections between economic family types and income deciles inferred from the dataset structure. For example, higher income deciles may be associated with specific economic family types, reflecting income disparities.

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.*

    - Recommended data splits may include training, validation, and testing sets to evaluate model performance in predicting income distribution patterns. These splits ensure robustness and generalizability of models trained on the dataset.

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

    - Errors, noise, or redundancies in the dataset may arise from data collection inconsistencies, reporting errors, or sampling biases. Data preprocessing techniques such as outlier detection and data cleaning may be employed to address these issues.

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.*

    - The dataset is linked to the Canadian Income Survey that is conducted annually.

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.*

    - The data is protected by confidentiality rules that prevent individuals from being identified.

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 contains sub-populations by geographical location, family type, etc. The distribution is even.

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.*

    - NO

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.*

    - The dataset contain income statistics of different groups of Canadians which might be considered sensitive.

16. *Any other comments?*

    - NO

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 Canadian Income Survey (CIS) is acquired through a combination of direct reporting by subjects and the merging of data from the Labour Force Survey (LFS) and tax records. Subjects provide information on various aspects such as labour market activity, income sources, household characteristics, and geographic details.

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?*

- Data collection mechanisms include survey questionnaires for direct reporting by subjects, administrative records for tax data, and automated data processing systems for merging and validating datasets. Validation processes ensure data accuracy and consistency across sources.

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 CIS dataset is a sample drawn from the larger population of Canadians. The sampling strategy is probabilistic, designed to ensure representativeness across demographic groups and geographic regions. Sampling methods are validated to ensure robustness and reliability of survey estimates.

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)?*

    - Data collection involves collaboration among various stakeholders including government agencies responsible for conducting surveys, data collection personnel, and statistical experts. Data collection staff are trained to administer surveys and ensure compliance with ethical standards.

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 timeframe for data collection typically spans from January to June of the year following the reference year. This timeframe aligns with the reference period for income and demographic information collected through surveys and tax records.

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.*

    - Ethical review processes are conducted by institutional review boards to ensure compliance with privacy regulations and ethical guidelines. These processes evaluate the survey methodology, data collection procedures, and potential risks to participants.

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

    - Data collection is primarily conducted directly from individuals through survey responses. However, tax data is obtained indirectly from government agencies responsible for tax administration and compliance.

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.*

    - Participants are notified about data collection through informed consent processes outlined in survey questionnaires and tax forms. Information about data usage, confidentiality, and rights is provided to participants prior to data collection.

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.*

    - Participants consent to the collection and use of their data by voluntarily completing survey questionnaires and tax forms. Consent language explicitly outlines the purposes of data collection, confidentiality measures, and participant rights.

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).*

    - Consent mechanisms allow participants to revoke their consent at any time by contacting survey administrators or tax authorities. Participants are informed about their rights to withdraw consent and the implications of doing so.

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.*

    - An analysis of the potential impact of the dataset on data subjects is conducted to assess risks and mitigate harms. This includes evaluating privacy risks, confidentiality protections, and potential biases in data collection and analysis.

12. *Any other comments?*

    - NO

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.*

    - Yes. Instances that are not related to the objective of this report is removed. Missing values are removed. Variables were renamed to better analyze the data.

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.* -Yes, the raw data is in the inputs folder called "income.csv". It can be obtained from the repository's inputs folder or the link:

https://open.canada.ca/data/en/dataset/b06716c0-eea7-4267-87b6-4faaa2679f22.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - R studio software and packages associated with it were used to clean the data.
4. *Any other comments?*
    - NO

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - The Canadian Income Survey (CIS) dataset has been utilized for various research and analytical tasks related to income distribution, poverty analysis, labor market dynamics, and socioeconomic policy evaluation. Researchers have employed the dataset to investigate trends in income inequality, assess the effectiveness of social welfare programs, and examine the impact of economic policies on household income levels.
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
3. *What (other) tasks could the dataset be used for?*
    - The dataset holds potential for a wide range of tasks beyond its current applications. Future research endeavors could explore topics such as intergenerational income mobility, the gender pay gap, regional disparities in income distribution, and the relationship between education and income attainment. Additionally, the dataset could be leveraged for comparative studies across countries or regions to gain insights into global income inequality trends.
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?*

- While the CIS dataset provides valuable insights into income dynamics and socioeconomic trends, users should be mindful of certain considerations. For instance, the dataset's sampling design and weighting procedures may impact the generalizability of findings to the entire Canadian population. Users should also be cautious about drawing causal inferences from observational data and consider potential confounding variables that could influence the relationships observed in the dataset.

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

   - NO

6. *Any other comments?*

   - NO

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 made available to third parties outside of the entity. It will be openly accessible to individuals, organizations, researchers, policymakers, and any other interested parties.

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

   - DOI: https://doi.org/10.25318/1110019201-eng

3. *When will the dataset be distributed?*

   - he dataset will be distributed on May 5th, 2025.

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.*

   - The dataset will be distributed under the MIT License.

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.*

   - The Open Government License of Canada(https://open.canada.ca/en/open-government-licence-canada)

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

7. *Any other comments?*

   - NO

Maintenance

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

   - Canadian government

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

   - The email address is provided in the github. The email address of the original dataset is infostats@statcan.gc.ca.

3. *Is there an erratum? If so, please provide a link or other access point.*

   - 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)?*

   - The dataset will be updated annually to incorporate new data and correct any errors.

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.*

   - There are no specific limits on the retention of individual data within the dataset.

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.*

   - Older versions of the dataset will be retained for reference purposes. However, ongoing updates will be provided annually to ensure the dataset remains current. Any obsolete versions will be clearly marked and archived in the repository.

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.*

    - Contributions to the dataset are not currently solicited or accepted. The dataset is maintained by the original owner, and any updates or modifications will be managed internally.

8. *Any other comments?*

    - NO

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. “Datasheets for Datasets.” *Communications of the ACM* 64 (12): 86–92.