

Datasheet for ‘Labour Force Survey - Statistics Canada’*

Datasheet

Sehar Bajwa

March 27, 2024

According to Gebru, Datasheets for datasets are essential for enhancing transparency, accountability, and reproducibility by documenting key information such as dataset motivation, composition, biases, and recommended uses. This is a datasheet summarizing the Labour Force Survey (LFS) dataset provided by Statistics Canada. The dataset offers detailed labor force statistics for the city of Toronto and other regions in Canada.

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 was created to provide detailed labor force statistics for different geographic areas within Canada, focusing particularly on the city of Toronto. The aim is to track employment trends, unemployment rates, and participation rates over time to facilitate informed decision-making regarding economic policies, workforce development, and resource allocation.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
 - The original dataset is published by Statistics Canada, and published on Open Data Toronto by Economic Development and Culture Canada.

*Code and data are available at: <https://github.com/SEHB2012/lfsdatasheet>

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.*

- Govt of Canada

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.*

- The dataset comprises monthly and annual tables containing labor force statistics for different geographic areas, including the city of Toronto, Toronto Census Metropolitan Area (CMA), Ontario, and Canada. The total number of instances varies across the tables, depending on the specific breakdown of data by geography, time period, and labor force indicators.

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

- LFS is based on a monthly sample of approximately 2,800 households in the Toronto CMA, about half of the sample is from the city of 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).*

- LFS is based on a monthly sample of approximately 2,800 households in the Toronto CMA, about half of the sample is from the city of Toronto.

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.*

- TBD

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

- TBD

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.*

- TBD
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.*
 - TBD
 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.*
 - TBD
 9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
 - TBD
 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.*
 - TBD
 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 LFS Survey questionnaire has confidential data, including names, telephone numbers, ages, email addresses. Statistics Canada is prohibited by law from releasing any data which would divulge information obtained under the Statistics Act that relates to any identifiable person, business or organization without the prior knowledge or the consent in writing of that person, business or organization. Various confidentiality rules are applied to all data that are released or published to prevent the publication or disclosure of any information deemed confidential. If necessary, data are suppressed to prevent direct or residual disclosure of identifiable data.
 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.*
 - TBD

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.*
 - The LFS uses standard population totals like age, gender and geography to calibrate the weights in their tables.
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.*
 - TBD
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.*
 - In January 2006, the LFS added 5 questions to identify the immigrant population, and in 2003-04 they added questions to recognise Indigenous participants.
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.*
 - Through survey responses from knowledgeable household members. There was extensive training of interviewers and quality assurance checks are implemented to ensure data accuracy.
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 for the Labour Force Survey was collected using computerized questionnaires on laptops or desktop computers. Interviewers directly captured responses during interviews. Validation mechanisms included built-in range checks and prompts for inconsistent data entry. Manual editing and imputation were performed for out-of-range or inconsistent entries.

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 Labour Force Survey (LFS) utilizes a monthly sample of approximately 2,800 households in the Toronto Census Metropolitan Area (CMA), with half of the sample representing the city of Toronto. This sampling approach employs a rotating panel design, where households are surveyed for six consecutive months before being replaced by new households. This design ensures efficient estimation of month-to-month changes, with a five-sixths overlap between consecutive months' samples. The six-month rotation minimizes respondent burden while maintaining sample representativeness.
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 Labour Force Survey involved trained interviewers. Interviewers were likely employees of Statistics Canada or contracted survey professionals. Compensation for interviewers typically includes a salary or hourly wage. Specific details on compensation amounts for interviewers or contractors are not provided in the information available.
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 Labour Force Survey data is collected monthly. Data collection typically covers the reference period of the previous month. The creation timeframe of the data associated with the instances aligns with the monthly data collection schedule. The data is processed and analyzed over approximately eight and a half weeks after the reference period.
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.*
 - TBD
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
 - The data in question is collected directly from individuals through interviews conducted by Statistics Canada interviewers.
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.

- TBD

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.*

- TBD

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).*

- TBD

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.*

- TBD

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.*

- Yes, preprocessing and cleaning of the data were conducted. The process involved editing and imputation to resolve missing or erroneous data entries. Imputation methods such as carry-forward, deterministic, and donor (hot-deck) imputation were used to replace missing values with logically consistent ones. The guide to LFS also mentions the use of industry and occupation coding based on job descriptions provided by respondents.

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.*

- TBD

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

- TBD

4. *Any other comments?*

- TBD

Uses

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

- TBD

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.*

- TBD

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

- TBD

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?*

- TBD

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

- Tasks that require precise industry and occupation coding may be impacted by potential errors in the dataset.

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.*

- TBD

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
 - TBD
3. *When will the dataset be distributed?*
 - TBD
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 has an Open Government license - Toronto.
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.*
 - TBD
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.*
 - TBD
7. *Any other comments?*
 - TBD

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
 - Statistics Canada is responsible for maintaining the dataset.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
 - edcresearch@toronto.ca
3. *Is there an erratum? If so, please provide a link or other access point.*
 - There is a format error while exporting that has been noticed and added to the dataset. It can be accessed here: <https://ckan0.cf.opendata.inter.prod-toronto.ca/dataset/97c25f5b-7542-40d4-b7fa-ef46e9442016/resource/68317412-497b-4f90-a834-37d347b4615e/download/lfs-data-format-error-docx.docx>

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)?*
 - TBD
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.*
 - TBD
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.*
 - TBD
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.*
 - TBD
8. *Any other comments?*
 - TBD

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.