Datasheet for Wages by Education Level*

Yuanyi (Leo) Liu

December 3, 2024

This datasheet provides details about the "Wages by Education Level" dataset, which documents hourly wage rates for Canadian workers from 2000 to 2019. The dataset includes demographic factors such as age, gender, and education level. It is intended to support research on wage disparities and the role of education in shaping economic outcomes. The dataset is available under the Open Government Licence - Ontario, allowing for broad use in research and policy discussions. It is maintained by the Government of Canada, ensuring its accuracy and applicability for continuous economic studies.

Extract of the questions from "Datasheets for datasets" (Gebru et al. 2021). The data is obtained from the Government of Ontario (Statistics Canada 2020).

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 developed to investigate the relationship between education level, age, gender, and average hourly wage rates in Canada from 2000 to 2019. Its purpose is to assess whether higher educational attainment leads to increased earnings and how this relationship varies across demographic factors and has evolved over time.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - Yuanyi (Leo) Liu, while studying at the University of Toronto.
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.*
    - No applicable.
4. *Any other comments?*

*Code and data are available at: Determinants of Wage Variation in Canada.

- No additional comments are provided.

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 represents aggregated hourly wage data categorized by education level, age group, and gender for individuals in Canada. Each instance corresponds to a unique combination of these factors, capturing variations in wages across demographic and educational categories from 2000 to 2019. The dataset does not include multiple types of instances beyond these categories.

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

    - There are 1080 instances in the dataset, each representing a unique combination of year, education level, age group, gender, and average hourly wages.

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 appears to be a complete set of instances from the available data during the specified timeframe. Representativeness is assumed but not specifically validated.

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 processed data including the year, education level, age group, gender, and the average hourly wage rate for the corresponding demographic and educational category.

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

    - The average hourly wage rate serves as the primary target variable, used to analyze wage differences across education levels, age groups, and genders.

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 is no information indicated as missing from the 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 instances are not explicitly stated but can be inferred through shared variables such as year, education level, age group, and gender, which allow for comparisons and trend analyses across these dimensions.

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.*

    - There are no recommended data splits mentioned, as the data appears to be used for temporal analysis rather than predictive modeling that would require such splits.

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

    - No specific errors, noise, or redundancies are noted. However, the potential for such issues exists as with any dataset and would be subject to further validation checks.

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 self-contained and does not rely on external resources.

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 dataset does not contain confidential information as it is aggregated data that does not identify individuals.

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.*

    - The dataset does not contain any content that might be offensive or otherwise cause anxiety.

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 dataset identifies sub-populations based on education level but does not provide further demographic breakdowns such as age, gender, etc.

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.*

    - It is not possible to identify individuals from the dataset as the data is aggregated and anonymized.

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 does not contain data that might be considered sensitive as it deals with aggregated economic figures without any direct personal identifiers.

16. *Any other comments?*

    - No additional comments are provided.

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 was acquired from a governmental database, ensuring that the data was directly observable and recorded through standardized bureaucratic processes. Given that the data originates from official government records, there is a high degree of validation inherent in the data collection method.

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 was collected using government administrative procedures and mechanisms. These include labor surveys and economic reports which are routinely validated through governmental audit and review processes.

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

- As the dataset was derived from an administrative record, the sampling strategy is not applicable. The data represents a census of the information available 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 collection of the data was performed by government employees as part of their standard duties. Compensation follows the governmental pay scale for such positions.

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 was collected annually, reflecting the earnings within that year, thus ensuring that the timeframe of data collection matches the creation timeframe of the data instances.

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.*

    - The data being from a government source and aggregated for public use, individual ethical review processes are typically not required beyond the government's standard data collection and dissemination policies.

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 was obtained directly from government sources and not through third parties, ensuring direct responsibility and accountability for the data accuracy.

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.*

    - Notification for data collection are not typically required for governmental data collection at this aggregate level, as it is a standard practice for economic and labor data reporting and does not involve individual-level 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.*

- Consent for data collection are not typically required for governmental data collection at this aggregate level, as it is a standard practice for economic and labor data reporting and does not involve individual-level data.

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).*

    - Given the nature of the data as aggregate and anonymized, mechanisms for individuals to revoke consent are not applicable.

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.*

    - The potential impact of the dataset and its use on data subjects is generally considered in the context of privacy and data protection laws under which the data is collected and released by the government. However, specific documentation of impact analysis would be under the purview of the governmental body responsible for the data.

12. *Any other comments?*

    - No additional comments are provided.

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.*

    - Preprocessing and cleaning were applied to the dataset, which included filtering specific subsets of data relevant to the study, such as data pertaining to the Canadian geography, average hourly wage rate, and specific age groups. The education levels were also ordered to facilitate analysis. Missing values or incomplete records were not reported; however, any that may have been present were likely excluded during the filtering process.

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.*

    - The raw data remains stored and is accessible as part of the government's open data portal. This ensures transparency and allows for replication or alternate analyses by other researchers.

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

    - The R scripts used for preprocessing, cleaning, and labeling are available within the project's repository. They provide a transparent methodological framework for replicating the dataset processing steps.

4. *Any other comments?*

    - No additional comments are provided.

Uses

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

    - The dataset has been used for an econometric analysis to examine the relationship between education level, age group, gender, and average hourly wage rates in Canada over the specified time frame. The analysis aimed to understand how these factors influence earnings and identify trends in wage disparities.

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.*

    - https://github.com/leoyliu/Determinants-of-Wage-Variation-in-Canada

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

    - Beyond the current econometric analysis, the dataset could be used for a range of studies in labor economics, education policy, wage disparity analysis, and longitudinal studies on wage progression over time.

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?*

    - The dataset's composition and preprocessing choices should be considered in future research. It focuses on specific age groups, excludes some education categories, and does not cover all demographics, which may limit its generalizability. Additionally, the aggregation of data might obscure variations within subpopulations. Users should align their research questions with the dataset's scope and limitations and contextualize findings to avoid biased or oversimplified interpretations. Incorporating supplementary data sources or validating results against broader datasets can mitigate these risks.

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

    - The dataset is not suited for analysis that requires individual-level detail or for making assumptions outside the scope of the variables provided, such as inferring reasons behind wage differences without considering other socioeconomic factors.

6. *Any other comments?*

    - No additional comments are provided.

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.*

    - The dataset will not be directly distributed by the creators but is available through the Open Government Canada platform. This ensures that the dataset remains accessible to the public and researchers without any restrictions imposed by the original creators.

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 is accessible via a direct download link provided by Open Government Canada. There is no digital object identifier (DOI) specifically assigned to this dataset.

3. *When will the dataset be distributed?*

    - The dataset is already available for public access and has been since its inclusion in the government's open data portal.

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 is distributed under the Open Government Licence – Ontario, which allows for free use and redistribution of the data with attribution. More details on the licensing terms can be found on the Open Government Canada website.

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.*

- There are no known third-party IP restrictions on this dataset, as it is provided by a governmental body with the intent for public use.

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 export controls or other regulatory restrictions are known to apply to this dataset. It is intended for public use within the legal frameworks governing public data in Canada.

7. *Any other comments?*

    - No additional comments are provided.

Maintenance

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

    - The dataset is maintained by the Government of Canada, specifically through its Open Government initiative.

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

    - The dataset can be inquired about through the contact mechanisms provided on the Open Government Canada platform.

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

    - Any corrections or updates to the dataset are handled by the Open Government Canada team and communicated through updates on the platform where the dataset is hosted.

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 is updated periodically as new wage data becomes available. These updates are documented and communicated through the Open Government Canada platform.

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.*

    - As the data is aggregated and anonymized, there are no specific retention limits applicable to this dataset. The data is intended for long-term use as part of governmental economic studies.

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 are archived and remain accessible via the Open Government Canada platform to ensure that historical data is available for longitudinal studies.

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.*

    - There are currently no mechanisms for external contributions to directly augment or extend the dataset. It is maintained exclusively by governmental statisticians and data clerks.

8. *Any other comments?*

    - No additional comments are provided.

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.

Statistics Canada. 2020. "Wages by education level." *Open Government Portal*. https://open.canada.ca/data/en/dataset/1f14addd-e4fc-4a07-9982-ad98db07ef86.