Datasheet for 'Canadian Grocery Price Data'*

Tina Kim

December 2, 2024

The Canadian Grocery Price Dataset provides a comprehensive collection of product pricing and availability information from major grocery vendors across Canada. Designed to support research on affordability, convenience, and sustainability in grocery shopping, the dataset contains detailed entries on data collection dates, vendor names, product names, unit prices, sale statuses, stock statuses, and best-seller statuses. Data was collected through automated web scraping, ensuring consistency and temporal relevance, while preprocessing enhanced comparability across vendors. This dataset aims to inform consumer decision-making, vendor selection, and policy development by offering insights into pricing trends and vendor performance. The dataset is suitable for tasks such as predictive modeling, vendor scoring, and economic analysis, with potential applications in sustainability and food security studies. It is shared under an open-source license to encourage broad use and collaboration, with provisions for periodic updates and community contributions.

Extract of the questions from Gebru et al. (2021). Datset from Filipp (2024).

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 drive more competition and reduce collusion in the Canadian grocery sector. Some example tasks in mind were doing an economic analysis of pricing data, visualizing the price of making a standard sandwich at each of the grocers (200g white bread + 20g ham + 20g lettuce…) to find the cheapest vendor, and finding which grocer is generally the cheapest across all comparable product families.

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/thk421/Canadian-Grocery-Prices.git.

- This dataset was collected by Jacob Filipp for Project Hammer.

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 project does not have an explicit funder.

4. *Any other comments?*

   - The dataset is still in a developing stage and is actively looking for people to contribute in it.

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.*

   - Instances represent individual grocery products sold by Canadian vendors.

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

   - There are 14 columns and 12117221 rows across 8 distinct vendors.

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 products available from major Canadian grocery vendors with a focus on urban areas. Seasonal coverage is limited due to the timeframe of data collection.

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 the date and time of data collection, their current and previous prices, price per unit, sale/best-seller/stock status, distinct product id, vendor names, product names, units, and brand names.

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

   - No explicit label; the dataset is exploratory, though it supports predictions of affordability and vendor reliability.

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.*

    - Certain entries lack unit sizes or prices due to scraping limitations.

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.*

    - No explicit relationships, though vendor-product relationships and historical trends are implicit.

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.*

    - Splits for exploratory analysis, predictive modeling, and testing may be created based on temporal sequences.

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

    - Minor inconsistencies in product matching, redundant data due to a product being listed under multiple categories or being miscategorized, and missing data due to website scraping errors.

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 product information was scraped from vendor websites, but the dataset is self-contained.

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.*

    - No confidential data is present.

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 offensive data is present.

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.*

    - No direct identification of sub-populations.

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 personal data is present.

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, it contains only public product and pricing data.

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.*

    - Data was scraped directly from grocery vendors' websites.

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?*

    - Automated web scraping tools were used to collect product listings and prices.

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

    - Deterministic sampling, focusing on widely available products across major vendors.

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)?*

    - Jacob Filipp from Project Hammer, with no external contributors in the data collecting process.

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.*

    - From February 28, 2024 to the latest load.

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.*

    - No formal reviews, but scraping adhered to legal and ethical guidelines.

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

    - Directly through web scraping.

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.*

    - Not applicable, as data is public.

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.*

    - Not applicable.

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).*

    - 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.*

    - No formal impact analysis, but data usage focuses on public benefit.

12. *Any other comments?*

    - Data is limited to the publicly accessible product interface.

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.*

    - No.

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.*

    - Not applicable.

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

    - Not applicable.

Uses

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

    - It is not explicitly stated, but the original dataset collecter has encouraged people to play around with the data while suggesting some research questions.

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://jacobfilipp.com/hammer/

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

    - Predictive modeling, vendor comparison, and price optimization.

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?*

    - Care must be taken to avoid overgeneralizing findings to unrepresented vendors or regions.

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

    - Real-time decision-making or extrapolations to non-Canadian markets without validation.

6. *Any other comments?*
    - There is a lot of potential for using this data in sustainability and food security studies.

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 goal is for other people to use the dataset to make significant contributions to drive more competition and reduce collusion in the Canadian grocery sector.
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 originally distributed from the website https://jacobfilipp.com/hammer/, but can be distributed through GitHub repositories like https://github.com/thk421/Canadian-Grocery-Prices.git.
3. *When will the dataset be distributed?*
    - Upon publication of the related research.
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 in the GitHub repository.
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 restrictions.
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.*
    - None identified.
7. *Any other comments?*
    - This dataset is accessible to researchers and the public.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - Jacob Filipp from Project Hammer.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - They can be contacted through their website: https://jacobfilipp.com/hammer/
3. *Is there an erratum? If so, please provide a link or other access point.*
    - None currently.
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 constantly being updated with new instances (as of 2024-12-01 20:32:00.43 Eastern Time) and can be checked on the website https://jacobfilipp.com/hammer/.
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.*
    - No, dataset is intended for long-term availability.
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.*
    - Yes, it will be available on the website https://jacobfilipp.com/hammer/, and in case it is not maintained, an older version is archived on the repository https://github.com/thk421/Canadian-Grocery-Prices.git.
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 can be made by downloading the data from the website or Contributions via pull requests or issue submissions in the git repository. Validation can be done by contacting the original dataset collector at https://jacobfilipp.com/hammer/.

References

Filipp, Jacob. 2024. "Canadian Grocery Price Data." https://jacobfilipp.com/hammer/.

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.