Datasheet for 'Canadian Grocery Prices'*

The Project Hammer Dataset

Ziheng Zhong

November 28, 2024

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 as part of Project Hammer to address issues of competition and collusion in the Canadian grocery sector. Its primary purpose is to enhance transparency by providing historical grocery price data from leading retailers, including Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart Canada, and Save-On-Foods. Covering the period from February 28, 2024, to the latest update, the dataset aims to fill the gap in publicly available, structured information on grocery pricing trends. It is designed to facilitate economic analysis by enabling detailed examination of price dynamics over time and across retailers. Additionally, the dataset supports legal and policy efforts by providing the data needed to analyze pricing behavior and advocate for fairer competition in the grocery market. By offering various formats, such as CSV files, a SQLite database, and Excel-friendly versions, it ensures accessibility for academics, policymakers, and analysts interested in studying the sector's market dynamics.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

    - The dataset was created by Jacob Filipp as part of Project Hammer. This project is not explicitly tied to a specific company, institution, or organization but appears to be driven by a personal or community-oriented goal to provide publicly accessible data for economic, academic, and policy analysis.

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.*

*Code and data are available at: https://jacobfilipp.com/hammer/.

- The creation of the dataset for Project Hammer was initiated by Jacob Filipp, a marketing operations professional based in Toronto. There is no publicly available information indicating that the project received external funding or was associated with any specific grants.

4. *Any other comments?*

   - NA.

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 the dataset represent grocery product prices and their associated metadata, comprising two main types of instances. The first type is product metadata, which includes information such as product name, vendor, brand, unit size, and category, providing descriptive details about individual grocery products. The second type is time-series price data, capturing historical price changes for these products across different vendors over time, with fields like current price, old price, and price per unit. These two types of instances are linked through unique identifiers (e.g., "id" or "product_id") and can be joined to enable comprehensive analysis of both product attributes and pricing trends, facilitating detailed studies of grocery price dynamics in Canada.

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

   - As mentioned in the previous question, this dataset comprises two primary components: product metadata and time-series price data. It contains a substantial number of instances in both components, with the exact number being too large to determine precisely.

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 is a sample of grocery prices rather than an exhaustive collection of all possible instances. It includes data from eight major Canadian grocery vendors—Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart Canada, and Save-On-Foods—spanning a specific time period from February 28, 2024, to November 12, 2024. While it captures a broad geographic coverage and diverse range of products from prominent retailers, it does not include all vendors or products available in

Canada, nor does it account for smaller, local retailers or specialty stores. The representativeness of the sample for the larger set of all Canadian grocery pricing data has not been explicitly validated or verified, as the focus appears to be on major players in the industry rather than creating a fully comprehensive dataset. This limitation may result from data availability or the project's scope, which prioritizes key market participants to analyze competition and pricing trends effectively.

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 in the dataset consists of processed features rather than raw data. For the product metadata, each instance includes structured fields such as product name, vendor, brand, category, unit size, and other descriptive attributes. For the time-series price data, each instance contains fields like current price, old price, and associated timestamps, representing historical pricing trends for each product across different vendors.

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

    - Each instance in the dataset is associated with an identifier that serves as its label or target for linking related data points. In the product metadata, the id or product_id field uniquely identifies each product, enabling cross-referencing between metadata and time-series price data.

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.*

    - Some information might be missing from individual instances in the dataset due to data unavailability or inconsistencies in the original sources. For example, certain products might lack fields such as price per unit, old price, or timestamps if the vendor did not provide complete historical pricing data or if specific updates were not recorded during the data collection period.

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, relationships between individual instances are made explicit through the use of unique identifiers, such as id or product_id, which link product metadata to time-series price data. These identifiers establish a direct relationship between the descriptive attributes of a product (e.g., name, brand, vendor) and its historical pricing trends.

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.*

- Time-based splits could be applied by dividing the data into training and testing periods (e.g., using earlier months for training and later months for testing) to analyze trends or predict future prices. Alternatively, vendor-based splits might be used to isolate pricing behaviors of specific retailers for comparative studies.

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

    - Potential errors could include incorrect or inconsistent entries in fields such as product names, prices, or units, resulting from variations in how data is recorded by different retailers. Noise might arise from temporary pricing anomalies, such as discounts, promotions, or data entry mistakes, which could skew analyses if not accounted for.

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 largely self-contained, comprising product metadata and time-series price data collected directly from eight Canadian grocery vendors. However, it implicitly relies on external resources, as the data originates from the vendors' publicly accessible platforms or systems, such as their websites or pricing APIs, during the collection period. There are no guarantees that these external sources, such as vendor websites or APIs, will remain accessible or constant over time, as vendors might change their platforms or restrict data access.There are no explicit restrictions mentioned for accessing or using the dataset itself, but users should be aware that the original data sources may be subject to terms of use, copyrights, or access limitations.

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 data that might be considered confidential. It consists of publicly accessible grocery price information and product metadata collected from major Canadian grocery vendors, which are typically available on their websites or pricing systems. The dataset does not include any private or sensitive information, such as personal data, non-public communications, or legally protected details. Its content is limited to market data intended for transparency and analysis of grocery pricing trends.

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 data that, if viewed directly, might be offensive, insulting, threatening, or anxiety-inducing. It is focused entirely on grocery pricing information and product metadata, which are neutral and factual in nature.

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 does not identify any sub-populations, such as by age, gender, or other demographic characteristics. Its focus is entirely on grocery pricing information and product metadata, with no inclusion of data related to individuals or specific demographic groups. The dataset is structured around products, vendors, and time-series price trends, without any reference to sub-populations or their distributions.

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 directly or indirectly from the dataset.There is no inclusion of customer information, transaction details, or other data points that could be used to deduce individual identities, either in isolation or through combination with external datasets.

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 any data that might be considered sensitive. It strictly includes grocery pricing information and product metadata, with no references to race, ethnic origins, sexual orientations, religious beliefs, political opinions, union memberships, locations, financial or health data, biometric or genetic data, government identification, or criminal history. Its content is focused solely on market data, which is neutral and devoid of any personally sensitive information.

16. *Any other comments?*

    - NA.

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 through direct observation of publicly available information from the platforms of major Canadian grocery vendors, such as their websites or pricing APIs. The data represents observable attributes like product names, brands, categories, and prices, which were recorded without relying on reported responses from subjects or inferences derived from other data. There is no explicit mention of validation or verification processes in the documentation, but the structured nature of the dataset suggests an effort to standardize and clean the collected information for consistency. However, since the data originates directly from vendor sources, its accuracy likely depends on the reliability of the vendors' published information at the time of collection.

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 software-based mechanisms, such as web scraping tools, software programs, or APIs that accessed publicly available information from the platforms of major Canadian grocery vendors. These tools automated the extraction of product metadata and pricing information over time. While the exact validation procedures are not detailed, the reliability of the collected data likely depends on the accuracy and consistency of the vendors' publicly available systems at the time of data collection. Manual curation may have been involved to standardize and clean the data, ensuring it is structured and free from obvious errors or inconsistencies. The mechanisms used would have been tested for functionality to ensure proper extraction of relevant fields, but the ultimate validation of the data relies on the integrity of the vendors' platforms.

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

   - This is a independent dataset, not a sub portion of any other data.

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 was led by Jacob Filipp. There is no indication that students, crowdworkers, or contractors were involved in the data collection.

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 over the timeframe from February 28, 2024, to November 12, 2024. This collection timeframe matches the creation timeframe of the data associated with the instances, as the dataset captures grocery price information and product metadata that were contemporaneously available on the vendors' platforms during this period. The data reflects real-time observations of prices and product details as they were updated on the vendors' systems, ensuring that the dataset accurately represents the state of grocery pricing during the specified collection window.

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.*

    - There is no indication that any ethical review processes.The data collection involved publicly available information from grocery vendor platforms, which generally does not require ethical review as it does not involve private or sensitive data or interactions with individuals.

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 not collected directly from individuals but was obtained from third-party sources, specifically the publicly available platforms of major Canadian grocery vendors. It was accessed and compiled through automated tools or software programs by Jacob Filipp as part of Project Hammer. The dataset itself is provided directly through the Hammer website, serving as the consolidated repository of the collected data for public access and analysis.

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.*

    - The individuals in question were not notified about the data collection because the dataset does not involve personal data or interactions with individuals.

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 from individuals was not applicable to this dataset because it does not involve personal data or information about individuals.

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).*

    - Since the dataset does not involve personal data or require consent from individuals, no mechanisms for revoking consent were necessary or provided.

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 analysis of the potential impact of the dataset and its use on data subjects, such as a data protection impact analysis, has been conducted. This is because the dataset does not involve data subjects or personal information.

12. *Any other comments?*

    - NA.

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 were performed on the dataset to prepare it for analysis. The process began with merging the raw data and product metadata using an inner join on the `product_id` field, selecting relevant columns such as `nowtime`, `vendor`, `product_id`, `product_name`, `brand`, `current_price`, `old_price`, `units`, and `price_per_unit`. The dataset was then filtered to include only entries from "Walmart" and "TandT," and non-essential columns were removed, keeping only `nowtime`, `vendor`, `current_price`, `old_price`, and `product_name`. Date information was transformed by extracting the year, month, and day from the `nowtime` field, while `current_price` and `old_price` were cleaned by parsing numeric values to ensure proper formatting. Products were filtered by retaining those with the keyword "beef" in their names while excluding irrelevant products based on a list of terms such as "flavour," "vermicelli," "rice," and others, using string matching. Rows with missing values in critical columns were dropped, and the `nowtime` column was removed after extracting its components. This comprehensive preprocessing step ensured the dataset was clean, consistent, and tailored for analyzing beef-related products from the selected vendors.

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 was saved in addition to the preprocessed/cleaned data to support unanticipated future uses. It is accessible via the Hammer project's official website at https://jacobfilipp.com/hammer/. This ensures that users can reference or reuse the original unprocessed data for alternative analyses or tasks.

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

   - Yes, the software used to preprocess and clean the data is available. The data cleaning process was implemented using R (R Core Team 2023), a statistical programming language, with specific packages such as `dplyr` (Wickham et al. 2023) and `tidyr` (Wickham, Henry, and Vaughan 2023).

4. *Any other comments?*

   - NA.

Uses

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

   - The dataset has been primarily used as part of Project Hammer to analyze grocery pricing trends across major Canadian retailers. Additionally, this dataset was used as the fundation of this paper. Specifically, it was processed to focus on specific product categories, beef-related products, for detailed analysis. This work supports economic analysis, policy research, and advocacy for fairer pricing practices in the grocery sector. Beyond these uses, the dataset is designed to be versatile and available for broader tasks by researchers and analysts in the future.

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/iJustinn/Canadian_Grocery_Analysis.git

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

   - The dataset could be used for a wide range of tasks beyond its current applications, including comparative pricing analysis to identify market inefficiencies or disparities across vendors and regions, and time-series forecasting to predict future grocery prices based on historical trends. It could support consumer behavior studies by examining how pricing trends influence purchasing decisions for specific products or categories, and market competition analysis to investigate potential anti-competitive behaviors or collusion among vendors. The dataset also holds potential for policy development, providing insights to craft regulations promoting fair competition, and sustainability research by analyzing the pricing and accessibility of sustainable or locally sourced products.

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 composition and preprocessing of the dataset could impact its future uses in certain ways. Since the dataset focuses on specific vendors ("Walmart" and "TandT") and includes only beef-related products while excluding other items through targeted filtering, it is not representative of the entire grocery market or all product categories. This selective focus may limit the generalizability of analyses conducted with the dataset and could lead to biased conclusions if used to make broader market claims. Additionally, the exclusion of certain products (e.g., plant-based alternatives, soups) and the reliance on publicly available data from specific vendors might inadvertently overlook diverse consumer preferences or regional variations.

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 tasks that require a comprehensive or representative view of the entire grocery market, as it focuses on specific vendors ("Walmart" and "TandT") and beef-related products while excluding other categories.

6. *Any other comments?*

    - NA.

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 is openly distributed to third parties outside of the entity on behalf of which it was created. It is accessible to the public through the official Project Hammer website, enabling researchers, analysts, and policymakers to use the data for economic analysis, policy research, and advocacy. Additionally, the processed dataset is openly available on GitHub, providing an accessible resource for those seeking a similar insight about Canadian grocery market.

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 distributed through multiple channels, including a tarball download on the official Project Hammer website and a processed version available on GitHub.

3. *When will the dataset be distributed?*

    - The dataset is already being distributed and is currently available to the public. It can be accessed through the official Project Hammer website and the processed version is available on GitHub.

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.*

    - There are no fees or restrictive terms of use explicitly mentioned on the Project Hammer website or 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 third parties have imposed IP-based or other restrictions on the data associated with the instances.

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.*

    - The dataset comprises publicly available grocery pricing information and product metadata from Canadian vendors, which are not subject to export controls or other regulatory restrictions.

7. *Any other comments?*

    - NA.

Maintenance

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

    - The dataset is supported, hosted, and maintained by Jacob Filipp as part of Project Hammer. The processed version of this dataset will be maintained via the GitHub repository mentioned in previous sections.

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

    - You can contact the owner of the dataset, Jacob Filipp, by emailing him at jacob@jacobfilipp.com, or leave message in the GitHub repository if there's any queston about the processed dataset.

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

   - NA.

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 expected to be updated periodically. Updates may include corrections to errors, the addition of new instances, or the deletion of outdated or inaccurate entries. Jacob Filipp, the creator of Project Hammer, is responsible for managing these updates.

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.*

   - The dataset does not relate to people and does not contain personal or sensitive data. It consists solely of publicly available product and pricing information from grocery vendors.

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 not explicitly mentioned as being supported, hosted, or maintained.

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.*

   - Original dataset related: Contact owner of the dataset, Jacob Filipp, via the email mentiond earlier.
   - Processed dataset related: By adding GitHub issue in the repository mentioned earlier.

8. *Any other comments?*

   - NA.

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.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023. *dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Lionel Henry, and Davis Vaughan. 2023. *tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.