

Datasheet for ‘Project Hammer Grocery Pricing Data’*

My subtitle if needed

Harsh M Pareek

2024-11-28

The Project Hammer Grocery Pricing Data is a comprehensive dataset created to drive competition and reduce collusion in the Canadian grocery sector. Compiled by Jacob Filipp as part of Project Hammer, the dataset encompasses historical pricing information from eight major Canadian grocery vendors: Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart Canada, and Save-On-Foods. The data collection commenced on February 28, 2024, and focuses on in-store pickup options within a Toronto neighborhood.

Extract of the questions from @gebru2021datasheets.

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 enhance competition in the Canadian grocery sector by providing a comprehensive database of historical grocery prices from major grocers. The specific task was to compile and make accessible detailed pricing information to support research, policy-making, and consumer awareness.
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 Project Hammer, by Jacob Filipp.
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 funding details are not publicly disclosed but likely by Filips.

*Code and data are available at: [LINK](#).

4. *Any other comments?*

- The dataset serves as a valuable resource for various stakeholders interested in analyzing grocery pricing trends and market dynamics in Canada.

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.*

- Each instance represents a specific grocery product offered by major Canadian supermarkets, including Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart Canada, and Save-On-Foods. The dataset includes product-level pricing information over time.

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

- Nearly 1.3 M in raw file and 130k in products file.

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 aims to be a comprehensive collection of products available from the included vendors during the data collection period, capturing as much available data as possible.

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 includes: `nowtime`: Timestamp of data collection. `vendor`: Name of the supermarket. `product_id`: Unique product identifier per vendor. `product_name`: Product description. `brand`: Brand name. `units`: Quantities or sizes. `current_price`: Current selling price after discounts. `old_price`: Previous price indicating discounts. `price_per_unit`: Price per unit measure. `other`: Additional details such as stock status or promotional indicators.

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

- The primary target variable is `current_price`, representing the price of the product at the time of data collection.

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.*
 - Yes, example generic apple, no brand.
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 are established through product_id within each vendor and temporal relationships via nowtime.
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.*
 - No
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
 - No
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.*
 - 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
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
 - No
13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

- No
14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
- No
15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*
- No
16. *Any other comments?*
- N/A

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 was collected via web scraping of major grocery retailers' websites, focusing on "in-store pickup" options within a Toronto neighborhood.
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 scripts were used to collect data at regular intervals. Validation included data cleaning steps to remove duplicates and inconsistencies.
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 dataset aimed to capture all available products listed online by the vendors, there is no sampling.
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 was conducted solely by Jacob Filipp. There was no external compensation.
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.*
 - Data collection began on February 28, 2024, and continues with regular updates. The timeframe matches the creation timeframe of the data.
 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.*
 - While formal ethical review was not conducted, ethical considerations were made to comply with website terms of service and avoid overloading servers.
 7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
 - Data was obtained directly from the public websites of the grocery vendors.
 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.*
 - N/A
 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.*
 - N/A
 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).*
 - N/A
 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.*
 - Since no personal data is involved, an impact analysis on data subjects was not necessary.

12. *Any other comments?*

- The project aims to benefit consumers by promoting transparency and competition.

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 included: Removing duplicates. Handling missing values. Standardizing product names and units. Identifying and addressing known issues (e.g., multiple entries).

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, both raw and processed data are available for users.

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 software is not publicly available but may be shared upon request.

4. *Any other comments?*

- Users are encouraged to apply additional preprocessing as needed for their specific analyses.

Uses

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

- Yes, it has been used for: Analyzing pricing trends over time. Investigating potential price collusion among retailers. Supporting academic research and legal actions to address competition issues.

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.*

- Related research and resources may be available on the Project Hammer website or upon contacting Jacob Filipp.

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

- Potential tasks include: Economic modeling of grocery pricing. Developing consumer price comparison tools. Policy analysis for market regulation. Studying the impact of promotions and sales strategies.

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?*
 - Users should be aware of data limitations, such as inconsistencies and focus on a specific geographic area. Careful preprocessing may be required to mitigate these issues.
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 commercial purposes without explicit permission. It should also not be used to make definitive legal conclusions without further validation.
6. *Any other comments?*
 - N/A

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
 - Yes, the dataset is publicly available for researchers, policymakers, and other interested parties.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
 - The dataset is distributed via: Zipped CSV files with full price data. An SQLite file with full price data. Excel-friendly CSV subsets based on search queries. The dataset does not currently have a DOI.
3. *When will the dataset be distributed?*
 - The dataset is already available and is updated regularly.
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.*

- Users should refer to the Project Hammer website for specific licensing and terms of use.
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 known restrictions have been imposed by third parties.
 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 regulatory restrictions are known to apply.
 7. *Any other comments?*
 - Users are encouraged to credit Project Hammer when utilizing the dataset in their work.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
 - The dataset is maintained by Jacob Filipp as part of Project Hammer.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
 - Contact can be made via email at “jacob” at jacobfilipp.com.
3. *Is there an erratum? If so, please provide a link or other access point.*
 - Known issues and updates are documented on the Project Hammer website.
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)?*
 - Yes, the dataset is updated regularly as new data is collected. Updates are communicated through the Project Hammer website.
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.*
 - Not applicable.

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.*
 - Previous versions may be archived and accessible for reference purposes.
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.*
 - Interested parties can reach out to Project Hammer to discuss collaboration and contributions.
8. *Any other comments?*
 - The project encourages community involvement to enhance the dataset and promote its goals.

1 References

- Filipp, J. (2024). Project Hammer: Comprehensive Grocery Pricing Data. Retrieved from <https://jacobfilipp.com/project-hammer>
- Competition Bureau Canada. (2023). Retail Grocery Market Study Report. Retrieved from <https://www.competitionbureau.gc.ca/>
- Statistics Canada. (2024). Monthly average retail prices for selected products. Retrieved from <https://www150.statcan.gc.ca/>
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daum{’e} III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92.