Datasheets for Datasets: Enhancing Transparency and Accountability in Data Collection[*]

A Comprehensive Examination of the Project Hammer Grocery Dataset

Dingshuo Li

December 3, 2024

This datasheet documents the Project Hammer Grocery Dataset, a comprehensive compilation of historical grocery prices from Canadian retailers, with a focus on organic products. The dataset was created to address gaps in publicly available structured data, enabling the analysis of price trends, vendor strategies, and seasonal variations. By capturing detailed pricing information across vendors and regions, the dataset provides insights into market dynamics and competitive behavior in Canada's grocery sector. This resource is critical for fostering transparency, informing policy decisions, and supporting research on consumer economics and anti-competitive practices.

Extract of the questions from Gebru et al. (2021). And the data came 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 enable a detailed analysis of grocery price trends across Canada, focusing specifically on organic products. It fills a gap in publicly available datasets by providing a structured format for analyzing price evolution and vendor-specific pricing strategies.
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/dawsonlll/Organic-Product-Pricing-Analysis.git

- The dataset was assembled by Jacob Filipp as part of Project Hammer, an initiative dedicated to promoting competition and reducing collusion in Canada's grocery sector.

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 creation of the dataset was independently funded, with no specific grants associated.

4. *Any other comments?*

    - This dataset serves as a valuable resource for examining market dynamics, pricing models, and economic behaviors within the Canadian grocery industry.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

    - The dataset comprises records of grocery prices, including fields such as product id, product name, old price, current price, vendor, and date.

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

    - It includes data from eight vendors: Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart, and Save-On-Foods, covering dates from February 28, 2024, current time.

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 designed to cover all available instances within a specified timeframe, providing a comprehensive view of the market during that period.

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 entry contains raw price data, supplier details and time information, allowing analysis of price changes over time.

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

    - The primary target variable of interest is the 'current_price' of each product.

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 price are missing becuase the data is not recorded. Any missing values were addressed through data cleaning and imputation processes to ensure completeness.

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.*

    - The dataset explicitly captures relationships between old and current prices, as well as vendor-specific pricing strategies.

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.*

    - Not really. But for predictive modeling, a common approach is to split the data into training (70%), validation (15%), and testing (15%) sets.

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

    - Potential errors from web scraping were mitigated through validation checks and data cleaning procedures.

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, with data sourced directly from publicly accessible grocery websites.

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 include confidential information.

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 data is factual and does not contain offensive content.

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.*

    - Sub-populations are identified by vendor, facilitating analysis of different pricing strategies.

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 individuals can be identified from the dataset.

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 sensitive personal information.

16. *Any other comments?*

    - No

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 collected through web scraping and API utilization from top grocers' 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 scripts and available APIs were employed for data collection, with validation through cross-referencing and in-store price checks.

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

- A stratified sampling approach was used to ensure geographic and vendor diversity.

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)?*

    - Data collection was conducted by the project's creator, Jacob Filipp.

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 spanned from February 28, 2024 and up-to-date

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 collection process adhered to ethical guidelines, focusing on publicly available information.

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 publicly accessible websites.

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 was collected from public sources.

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.*

    - Regular audits ensured compliance with data protection standards.

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.*

   - The dataset was analyzed for potential impact on data subjects. As the dataset does not contain personal, sensitive or identifiable information, the risk to individual privacy or data protection is negligible. The analysis focused on ensuring compliance with ethical guidelines for the use of publicly available data, confirming that the content of the dataset meets legal standards and does not infringe intellectual property rights.

12. *Any other comments?*

    - No

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.*

   - Data underwent cleaning to correct anomalies and impute missing values, ensuring high quality for analysis.

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 is archived and available at https://jacobfilipp.com/hammer/

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

   - Preprocessing was performed using R scripts, with code available at https://github.com/dawsonlll/Org
     Product-Pricing-Analysis.git

4. *Any other comments?*

   - No

Uses

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

   - The dataset has been utilized for analyzing price trends and vendor pricing strategies.

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/ The dataset is available for download in CSV and SQLite formats.

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

   - The dataset can be used for economic analysis, market competition studies, and legal assessments of pricing practices.

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?*

   - Should be cautious of potential biases due to the data collection methods and ensure appropriate analytical techniques are applied.

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

   - The dataset is not intended for analyses requiring personal or sensitive information.

6. *Any other comments?*

   - No

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 publicly available for download.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

   - It is distributed in CSV and SQLite formats. No DOI found.

3. *When will the dataset be distributed?*

   - The dataset was made available on February 28, 2024

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 provided under MIT, encouraging academic and legal 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.*

    - There are no known restrictions 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 apply.

7. *Any other comments?*

    - No

Maintenance

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

    - The dataset is maintained by Jacob Filipp.

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

    - linkedin.com/in/jacobfilipp or jacobfilipp.com

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

    - There is no erratum currently available. Any errors identified in the dataset will be documented and corrected in subsequent updates, with notifications provided on the GitHub repository.

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 will be updated periodically to include new price records, correct labeling errors, and remove obsolete entries. Updates are managed by the Project Hammer team and communicated through the GitHub repository and the official 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.*

    - The dataset does not relate to personal data or individuals; therefore, no retention limits are 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.*

    - Older versions of the dataset will be archived and remain accessible for reproducibility and transparency.

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.*

    - A mechanism for community contributions is available through GitHub. Contributors can propose extensions or augmentations to the dataset via pull requests, which will be reviewed and validated by the Project Hammer team before inclusion. Approved contributions will be documented in the repository's change log.

8. *Any other comments?*

    - The dataset serves as a vital tool for academic and industry research. Its ongoing maintenance and open-source nature aim to ensure its relevance and usability over time.

References

Filipp, Jacob. 2024. *Hammer Project: Comprehensive Dataset on Organic Food Pricing*. 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. https://doi.org/10.1145/3458723.