Datasheet for 'Project Hammer by Jacob Filipp'*

Yiyi Feng

2024-12-02

The Hammer project dataset was created by Jacob Filipp, which provides historical grocery prices from major Canadian vendors, collected through automated web scraping of publicly available in-store pickup pricing for a Toronto neighborhood. The dataset supports academic research, economic analysis, and legal actions to improve competition and reduce collusion in the Canadian grocery sector. This datasheet will provide detailed information about the dataset from six aspects: Motivation, Composition, Collection Process, Preprocessing, Uses, and Distribution.

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 project compiles a historical database of grocery prices from major Canadian grocers to enable analysis of pricing trends. It fills the gap in publicly available data on detailed grocery prices, especially regarding concerns about price collusion and inflation. The data supports academic and policy research on price changes, sales patterns, retailer responses, and long-term pricing trends (Filipp 2024). It addresses the lack of pricing transparency among grocery retailers and provides tools for stakeholders to assess and promote fair competition.

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/kqlqkqlqF/Modeling-Price-Fluctuations-of-Seasonal-Fruits-in-Food-Products.git.

- The dataset was created by Jacob Filipp. The information provided does not explicitly mention a team or organization behind the project, implying it may have been independently initiated 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 source for this dataset is not explicitly mentioned. It appears that the project may have been self-funded or supported informally, as it emphasizes community collaboration. For further clarification, contacting Jacob Filipp directly would be recommended.

4. *Any other comments?*

   - No other comments.

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 historical grocery prices from major Canadian grocery retailers. Each instance corresponds to a specific grocery item and includes information such as: Product Details: Information about the product being tracked, including its category, brand, and packaging. Vendor Information: Details about the vendor which sell the corresponding product. Price Data: Historical pricing details, including regular prices, sale prices, and changes over time. Date and Time: The time period during which the price was recorded, enabling trend and temporal analysis.

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

   - As of November 27, 2024, the dataset contains 12,488,411 instances (data points), and this number is expected to increase as new data is added over 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 represents a sample of grocery product prices in Canada, not all possible instances. It includes pricing data from a mix of 8 major vendors across all provinces and territories, covering urban, suburban, and rural areas to ensure geographic diversity. While it aims to cover a wide range of grocery categories, it

may exclude niche or localized products. The dataset focuses on vendors with more readily available data, which may overrepresent larger retailers with a strong digital presence. It may not capture every price change or short-term promotion, and small, independent stores are underrepresented. Representativeness was assessed by comparing product categories with national sales data.

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.*

   - The dataset consists of two components, each representing different aspects of grocery product information: hammer-4-product dataset: This contains static information about the products. Each instance represents a unique product with attributes including a unique product identifier (id), a concatenated string with product details (concatted), vendor name, product name, unit of measurement, brand (if available), a URL for more details (if available), stock-keeping unit (sku), and Universal Product Code (upc), if applicable. hammer-4-raw dataset: This provides dynamic pricing information for products over time. Each instance represents a product's pricing and promotional details at a specific time, including the timestamp (nowtime), current price, previous price (if available), price per unit (price_per_unit), any promotional context (other), and a foreign key (product_id) linking to the product information in the hammer-4-product dataset.

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

   - Yes, the dataset contains implicit labels or targets associated with each instance, depending on its use. For the hammer-4-raw dataset: This dataset contains dynamic pricing information for grocery products over time. Each instance represents a product's price at a specific timestamp, with attributes including current_price, which tracks the price at the given moment, and price_per_unit, which normalizes the price based on the product's unit weight or quantity. The dataset also includes old_price, which shows the previous price if there was a change, and other, which can indicate promotional prices such as "SALE." The product_id links each pricing entry to its corresponding product details in the hammer-4-product dataset, enabling further analysis of price trends and promotional effectiveness. For the hammer-4-product dataset: There is no explicit target or label in this dataset. Instead, it provides the foundational metadata (product descriptions, units, vendor, etc.) necessary for analysis or modeling based on the hammer-4-raw dataset.

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, there is some missing information in individual instances, particularly in the hammer-4-raw dataset. Missing Information in hammer-4-raw: The old_price column may be empty when no sale or discount is applied to a product, such as when

a product's price remains unchanged or there is no previous price history available. The price_per_unit field is also missing in some rows, especially when the price remains consistent over time or when the product is sold in bulk without a specific unit price. Additionally, the other column, which indicates if the product is on sale (e.g., marked as "SALE"), may be blank when no promotion is applied to the product during that period. Missing Information in hammer-4-product: certain product details such as detail_url, sku, and upc are missing for some products, like "Cherries Red" and "Peaches." This missing data likely results from the absence of a standardized detail URL, SKU, or UPC at the time of data collection, or it may indicate that the specific product did not have these identifiers available. These gaps in data reflect inconsistencies in the product information across different sources or periods.

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 in the dataset are made explicit, though the structure of these relationships varies between the two datasets. In the hammer-4-product dataset, relationships between instances are made explicit through the product_id column. This column links different attributes (such as product name, brand, units, SKU, and UPC) to the same product. For example, multiple rows can exist for the product "Apples Ambrosia" (with product_id 3), each row representing different attributes like the product's name, packaging units, or SKU. The product_id ensures that these various attributes are correctly associated with the same product, establishing clear relationships between the instances. In the hammer-4-raw dataset, relationships are also established through the product_id column. Each row represents a price change or sale event for a specific product, with the product_id linking these price changes to the corresponding product in the hammer-4-product dataset. For instance, multiple pricing records for "Apples Ambrosia" (product_id 3) are recorded over time, each with a different nowtime timestamp. This establishes a temporal relationship, where each row in hammer-4-raw represents a price event tied to a specific product at a particular moment in time. The shared product_id between the two datasets ensures that pricing changes are linked to the correct product, while the nowtime field in hammer-4-raw specifies when the price change occurred.

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, there's no recommended data splits.

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

- The hammer-4-raw and hammer-4-product datasets contain several issues that could impact the quality and reliability of analyses. One major concern is missing values, particularly in the old_price column of hammer-4-raw. Many rows lack this information, especially when there is no price change, which can result in incomplete insights into historical pricing trends. Additionally, the nowtime column may contain duplicates or inaccuracies, which could affect time-series analyses if not properly verified. Another critical issue is the inconsistent formatting of the price_per_unit column in hammer-4-raw, where varying units like "0.40/100g" and "0.48/100g" complicate price comparisons across products. Without standardization, it becomes challenging to draw reliable conclusions about pricing. The dataset also contains repeated entries, where the current_price and other fields remain unchanged across several consecutive timestamps. This redundancy inflates the dataset size without adding new information, as seen in rows from June 14th to June 17th, which only differ in timestamps. In hammer-4-product, redundancies are also present. The product_name and concatted columns frequently overlap in the information they provide, such as product attributes like brand, unit size, and vendor. While this may aid human readability, it complicates data processing tasks. Additionally, fields like detail_url, sku, and upc are missing values, limiting the ability to uniquely identify products or link records across datasets. Addressing these issues would improve the dataset's clarity and usefulness for analysis.

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 not fully self-contained, as it relies on external resources through the detail_url, sku, and upc columns in the hammer-4-product dataset. The detail_url field links to external websites for additional product information, such as specifications or real-time pricing. However, there is no guarantee that these URLs will remain valid or that the linked content will stay accessible or unchanged. Websites often restructure or remove outdated pages, which could result in broken links or outdated data if the dataset is revisited in the future. Furthermore, there are no archived versions of these external resources, meaning the state of the dataset at the time of creation cannot be fully reconstructed without relying on the current status of those links. The sku and upc fields also create dependencies on external databases, such as retailer inventory systems or barcode repositories. Accessing these systems may require paid services or have usage restrictions, such as licensing agreements or API limits. For example, looking up product details using the upc field might need integration with proprietary barcode lookup services, which could

involve fees or usage limitations. Additionally, the dataset does not clarify whether these external resources are subject to intellectual property or legal restrictions.

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 information that would be considered confidential under privacy or legal laws. It mainly includes product pricing, price change records, and product details like vendor, product name, unit size, and brand. These are public commercial data with no personally identifiable information (PII) or sensitive data that would require special legal protections.

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 offensive, insulting, or threatening content. As mentioned before, it focuses on product pricing and product details like price changes, product names, brands, and pricing per unit. This information is neutral, factual, and commercial, with no content intended to provoke emotional reactions or cause harm.

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 include any demographic information such as age, gender, ethnicity, or other personal characteristics. The information centers on product attributes (e.g., product names, prices, and promotions) and logistical identifiers (e.g., timestamps and product IDs), without any reference to consumer demographics. Therefore, there is no segmentation or distribution of sub-populations within the dataset. Any demographic analysis would require external data, making the dataset neutral and free from demographic identifiers.

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, it is not possible to identify individuals directly or indirectly from this dataset. The dataset does not contain any personal data, such as names, contact details, or other identifying information, that could be used to identify individuals.

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, the dataset does not contain any sensitive data as defined by privacy standards. There are no personal attributes like race, ethnicity, sexual orientation, religious beliefs, political opinions, union memberships, or other personal characteristics that could be considered sensitive. Additionally, the dataset does not include financial data, health information, biometric or genetic data, government IDs, or criminal history.

16. *Any other comments?*

    - No other comments.

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 dataset includes price information from several grocery vendors, such as Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart Canada, and Save-On-Foods. Prices were collected through screen scraping from these vendors' websites, specifically for the "in-store pickup" option in a Toronto neighborhood. As a result, the dataset may lack details typically available through internal APIs. It doesn't cover every day or every vendor; from February 28 till now, it focuses on a small basket of products, and later expands to a larger range. There are also gaps, especially for some vendors or days, where data extraction failed, leading to missing prices. Consequently, the dataset is incomplete due to the limitations of the screen-scraping process and possible technical issues.

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 a screen-scraping method, where software programs were used to extract price information directly from the websites of grocery vendors. This process targeted the "in-store pickup" option in a specific Toronto neighborhood, retrieving pricing details from vendors like Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart Canada, and Save-On-Foods. There was no formal validation system in place. Limitations in the extraction process led to missing data for certain vendors or days, particularly when specific scraping tasks failed.

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 appears to be a sample from a larger set, with the sampling strategy based on screen-scraping prices from specific grocery vendor websites over a defined period. This approach is deterministic, focusing on a fixed set of grocery vendors and a specific product selection. Initially, the sample included a small basket of products, which was later expanded to cover a larger variety. The scraping process was not probabilistic, meaning the data was not randomly selected from all available products across the vendors. Instead, it extracted prices for a predefined set of products available at the time of scraping. While this deterministic approach ensures consistent tracking of the same set of products and vendors, it also results in gaps, such as missing prices for certain vendors on specific days or incomplete product information due to scraping failures. These gaps are more due to technical limitations or vendor-specific issues rather than a random sampling method.

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 for this dataset was primarily automated through a screen-scraping procedure, using software programs instead of human labor. As a result, no individuals (such as students, crowdworkers, or contractors) were directly involved in the collection process. A script or program was responsible for scraping prices from grocery vendor websites on specified dates. Since the process was automated, there was no compensation involved, as no human workers participated. If manual curation or validation had been necessary, there could have been compensation for workers, depending on the scope of the work. However, in this case, no human involvement means compensation details are not applicable.

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 from 2024 February 28 till now, starting with a smaller basket of products and later expanding to include a larger variety. The data collection period closely aligns with the dataset's creation timeframe, as the data was gathered continuously during this period. However, there were gaps in the data for certain vendors on specific days due to failed data extractions. While the timeframe of data creation matches the collection period, some data may be missing for certain dates and vendors due to these issues.

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.*

    - Based on the provided information, there does not appear to have been a formal ethical review, such as an institutional review board (IRB) process, for this dataset.

The data was collected via screen scraping from publicly accessible websites, and no personal or sensitive information seems to be involved. While there is no explicit mention of an ethical review, the primary ethical consideration would be ensuring compliance with the websites' terms of service or usage policies, rather than issues typically addressed by an IRB, like privacy or informed consent.

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 collected indirectly from third-party sources, specifically through screen scraping of publicly accessible grocery vendor websites. The vendors included Voila, T&T, Loblaws, No Frills, Metro, Galleria, Walmart Canada, and Save-On-Foods. The scraping focused on prices for the "in-store pickup" option in a neighborhood in Toronto. The process used automated tools to extract pricing information from the user interfaces of these websites, rather than collecting data directly from individuals. As such, the data is not tied to any specific individual but reflects publicly available product pricing information.

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 were not notified about the data collection because it was conducted through automated screen scraping of publicly accessible grocery vendor websites. Since the data collected does not include personal or identifiable information and only consists of publicly available product pricing, there was no need to notify individuals about the data collection.

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.*

   - The individuals did not provide explicit consent for the collection and use of their data because the dataset contains only non-personal information. The data consists of product pricing gathered through automated web scraping from publicly accessible vendor websites. This information, such as prices for in-store pickup options, was not tied to any specific individuals and was collected from the public sections of these websites. Since the data is not personal or private, individual consent was not required for its collection.

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).*

- Consent was not obtained from individuals for this dataset, as it does not involve personal or identifiable information. Since the dataset contains no personal data, there was no need for individuals to provide consent or a mechanism to revoke it.

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 the dataset consists solely of publicly available product pricing information scraped from grocery vendor websites, and does not contain personal or identifiable data, there was no need for a formal Data Protection Impact Assessment (DPIA). A DPIA is typically required when processing personal data that could impact individuals' rights and freedoms. As this dataset does not involve personal data, there is no direct impact on data subjects.

12. *Any other comments?*

    - No other comments.

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, the raw data has not undergone any preprocessing, cleaning, or labeling. The dataset was directly captured from various grocery vendors' websites via screen-scraping without any prior processing. As such, the data is in its raw form and has not yet been cleaned or modified in any way.

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.*

    - Since raw data has not undergone any preprocessing, this question was skipped.

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

    - Since raw data has not undergone any preprocessing, this question was skipped.

4. *Any other comments?*

    - No other comments.

Uses

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

   - The dataset, titled "Modeling Price Change of Seasonal Fruits-Flavored Food Products: A Predictive Analysis," was used in the study authored by Yiyi Feng. The study aims to predict price changes for seasonal fruit-flavored products, specifically strawberry and banana-flavored items, using data from the Hammer project. The analysis examines price fluctuations between in-season and off-season periods, emphasizing the expected price increases for these products after November, with a more significant rise observed for strawberry-flavored products.

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.*

   - Code and data are available at: https://github.com/kqlqkqlqF/Modeling-Price-Fluctuations-of-Seasonal-Fruits-in-Food-Products.git.

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

   - The dataset can be used for several valuable analyses. First, it could support price sensitivity and elasticity analysis, examining how price changes impact consumer demand for seasonal fruit-flavored products. This could help identify consumer price sensitivity and guide optimal pricing strategies for vendors. Second, the dataset could assist in demand forecasting, particularly during peak and off-peak seasons. By combining price data with sales volume data, businesses could forecast demand and optimize inventory management to prevent stockouts or overstocking during key months. Lastly, the dataset could be used to compare pricing strategies across vendors. By analyzing how different vendors price similar products, analysts could identify trends, inconsistencies, and strategies, offering insights into vendor positioning and potential opportunities for competitive advantage in pricing seasonal fruit-flavored 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 dataset, comprising product pricing information scraped from grocery vendor websites, contains no personal data and poses no apparent risks of harm or unfair treatment to individuals or groups. However, its future use has some limitations. First, the data represents a specific set of grocery vendors in Toronto, making it less applicable for analyses of national or international pricing trends, which could lead to skewed conclusions when applied outside this context. Additionally, the focus on "in-store pickup" prices excludes delivery options and some promotional variations, potentially limiting the generalizability of findings to other sales scenarios.

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 requiring broad or generalized conclusions about pricing trends across all grocery vendors, locations, or product categories, as it focuses on a specific subset of vendors in Toronto and a narrow selection of products. It is limited to in-store pickup prices, excluding delivery options and promotional pricing variations, making it unsuitable for analyses aiming to generalize across different contexts or markets. Additionally, the dataset is inappropriate for inferring customer behavior, such as demand forecasting or consumer preference modeling, as it lacks customer-specific information, including purchase history, demographics, or behavioral patterns. Finally, the dataset should not be used for legal or regulatory purposes, such as evaluating pricing compliance or setting pricing policies, as it does not provide comprehensive or legally verified data necessary for such assessments. Misuse in these contexts could lead to inaccurate or misleading conclusions.

6. *Any other comments?*

    - No other comments.

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 Hammer Project's dataset, authored by Jacob Filipp, is publicly available and focuses on enhancing competition and mitigating collusion within the Canadian grocery sector (Filipp 2024). The dataset includes historical grocery prices from top grocers across Canada and serves multiple purposes, such as academic research, economic analysis, and legal action. It aims to facilitate a deeper understanding of pricing trends, uncover patterns in vendor strategies, and provide evidence for addressing anti-competitive practices. By improving transparency, the project seeks to foster fairer competition and benefit consumers in the grocery market. Accessible through the Hammer Project's official website, the dataset is available in various formats, including CSV and SQLite. For convenience, an Excel-friendly version can also be requested using specific search phrases. This open-access resource allows researchers, analysts, and legal professionals to explore the data and generate insights into pricing dynamics. With its focus on promoting fair competition and reducing potential market abuses, the Hammer Project's dataset is a vital tool for driving informed economic and legal discussions within Canada's grocery industry.

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 from the Hammer project will be distributed through the project's website. It is available for download in various formats, including: CSV Files: The full dataset is offered in zipped CSV format, containing both the "product" file (metadata and product details) and the "raw" file (time-series price data). SQLite File: A full SQLite database containing all the data, which can be viewed using software like DB Browser and exported to CSV for further analysis. Excel-Friendly CSV Subset: A more user-friendly subset of the data is available in an Excel-compatible CSV format, which includes data filtered based on specific search phrases. There is no mention of a Digital Object Identifier (DOI) for the dataset on the project's website. However, the dataset is openly accessible and can be downloaded directly from the Hammer project's webpage:https://jacobfilipp.com/hammer/.

3. *When will the dataset be distributed?*

   - The dataset from the Hammer project is already available for distribution. It can be accessed at any time on the project's website.

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 Hammer Project dataset does not appear to be governed by a traditional copyright or intellectual property (IP) license. Instead, its availability is defined by specific terms of use (ToU) outlined on the project website. The dataset is primarily intended to enhance competition and address collusion within the Canadian grocery sector by compiling historical grocery prices and providing this data for academic research, economic analysis, and legal action. Its purpose underscores its role in fostering transparency and enabling actionable insights in the market. The dataset is freely accessible through the project's website, offered in multiple formats such as CSV and SQLite, with no explicit fees or usage restrictions for academic, research, or advocacy work. The Hammer Project also emphasizes community involvement, encouraging users to collaborate and contribute to its application in driving meaningful change. Users are urged to align their efforts with changemakers in economic analysis, legal action, or policy advocacy to advance the dataset's goals of improving fairness and competition in the grocery industry.

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.*

   - Based on the information provided on the Hammer Project website, there are no indications that third parties have imposed intellectual property (IP)-based or other restrictions on the dataset. It is openly accessible to users, aligning with its purpose of fostering competition and addressing collusion in the Canadian grocery sector.

The dataset is intended for use in academic analysis and legal action, supporting transparency and actionable insights. Its unrestricted access reflects the project's commitment to enabling widespread use and collaboration toward improving fairness and competition in the grocery market (Filipp 2024).

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.*

    - Based on the information available from the Hammer project website, there are no export controls or other regulatory restrictions explicitly mentioned that apply to the dataset or individual instances.

7. *Any other comments?*

    - No other comments.

Maintenance

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

    - The Hammer project dataset will be supported, hosted, and maintained by Jacob Filipp, the creator of the project.

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

    - The owner and curator of the Hammer project dataset, Jacob Filipp, can be contacted via email. The email address provided on the Hammer project website is: Email: jacob@jacobfilipp.com, since on the website they mentioned : Reach out to me (email "jacob" at this website) (Filipp 2024).

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

    - There doesn't appear to be any publicly available erratum or correction document for the Hammer project dataset on the official website or the dataset 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 Hammer project dataset is updated once per day with the latest pricing data from the Canadian grocery sector. However, there is no specific mention on the website about the correction of labeling errors or other issues. Therefore, while updates to the dataset are made regularly, it is unclear if or how labeling errors will be fixed or if any data cleaning will occur.

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 Hammer project dataset does not contain personal data or information related to individuals. It consists of publicly available grocery pricing data scraped from vendor websites. Since the data does not involve any personal or identifiable information, there are no applicable limits on data retention related to individuals, nor were individuals informed about retention periods or deletion policies. The dataset is focused on product pricing and vendor information, which does not require specific retention policies for individuals or any enforcement regarding the deletion of data related to people.

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.*

    - The Hammer project dataset will not continue to support or host older versions. The dataset is continuously updated, with the latest data being available on the project's website.

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.*

    - Yes, others can contribute to the Hammer project dataset by providing feedback and collaborating. Jacob Filipp, the project's creator, welcomes help with data processing, analysis, and identifying errors or improvements. Those interested can contact him via email to offer support or suggestions. If contributions are accepted, they will be added to the dataset and reflected in updates on the project's website. However, there is no formal process for notifying contributors or making their contributions public.

8. *Any other comments?*

    - No other comments.

References

Filipp, Jacob. 2024. "Project Hammer." 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.