Datasheet for the Canadian Grocery Pricing Dataset[*]

Jason Yang

December 3, 2024

This datasheet documents the dataset for the Project Hammer by Jacob Filipp

Motivation

1. *For what purpose was the dataset created?*

- The dataset was created as part of Project Hammer, which aims to foster greater competition and reduce potential collusion within the Canadian grocery sector by analyzing pricing dynamics across major vendors.

2. *How was this dataset collected?*

- The dataset was collected via a screen-scrape of the user interface (UI) of grocery vendor websites, capturing publicly available pricing data and product information.

3. *Who created the dataset?*

- The dataset was created by Jacob Filipp, who spearheaded the data collection and organization process as part of the broader Project Hammer initiative.

4. *Who funded the creation of the dataset?*

- Supported by Project Hammer's initiative to study grocery pricing dynamics
- https://jacobfilipp.com/hammer/ (Filipp (2024))

5. *How can this dataset be used?*

- The dataset can be used for predictive modeling, pricing trend analysis, and consumer research on grocery shopping behaviors.

[*] Code and data are available at: (https://github.com/jasonbot123/Canadian-Grocery-Price-Analyze).

6. *Who are the intended audience for the dataset?*

- The dataset is intended for researchers, policymakers, and consumers interested in grocery pricing trends.

7. *Any other comments?*

- N/A

Composition

8. *How many observations are there in total?*

- Number of Observations: 12,842,742 rows

9. *What does the instances that comprise the dataset represent?*

- Each row represents the price and metadata of a specific item listed on a grocery vendor's website.

10. *What types of variables are included in the dataset?*

- The dataset includes numeric variables (e.g., current_price), categorical variables (e.g., vendor), and text variables (e.g., product_name).

11. *Are there missing information from certain variables?*

- Yes. (e.g., old_price, other, brand, upc, sku, detail_url)

12. *Does the dataset contain data that may violate ethical rules for statistical analysis*

- No.

13. *Does the dataset contain any personally identifiable information (PII)?*

- No.

14. *Were there ethical concerns during data collection?*

- Ethical concerns were mitigated by adhering to the terms of service of the scraped websites and avoiding overloading servers.

Collection

15. *How was the data collected?*

- The dataset was collected via a screen-scrape of grocery vendors' website user interfaces.

16. *Over what time period was the data colelcted?*

- The data was collected periodically from February 28, 2024, to November 26, 2024.

17. *What sources were used to create the dataset?*

- The dataset was sourced from the publicly accessible websites of eight Canadian grocery vendors.

18. *How complete is the dataset?*

- The dataset covers the eight selected vendors but does not include in-store pricing, regional variations, or smaller grocery chains.

Limitation 19. *Are there known limitations to the dataset?* - Yes, the dataset does not include dynamic pricing changes, in-store promotions, or data from independent grocers.

Preprocessing.cleaning 20. *Was there any preprocessing* - No.

References

Filipp, Jacob. 2024. *Hammer: A Tool for Strategic Analysis.* https://jacobfilipp.com/hammer/.