Datasheet for Toronto Sourdough Bread Price Dataset'*

Price Variations in Toronto's Sourdough Bread Market: Evidence from Major Retailers

Grace Nguyen

December 14, 2024

This datasheet documents a comprehensive dataset of sourdough bread prices from major retailers in Toronto. The dataset spans February to July 2024, covering five major retail chains. It enables analysis of pricing patterns, market segmentation, and retail strategy in specialty food markets.

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.*
    - To analyze pricing patterns in Toronto's sourdough bread market through the major retailers, the dataset was created. This helps close a gap in knowledge about how retailers price specialty bread products in urban markets and how retailers position themselves in that market through pricing strategy.
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 through Project Hammer (Filipp 2024), with daily price tracking across major Toronto retailers.
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.*
    - This dataset was compiled as part of academic research with no external funding.
4. *Any other comments?*

*Code and data are available at: https://github.com/gracenguyen133/Sourdough-Bread-Pricing.git

- The dataset specifically focuses on sourdough bread to enable focused analysis of a specialty product category.

Composition

1. *What do the instances represent?*
    - Each instance represents a daily price observation for a sourdough bread product at a specific retailer in Toronto.
2. *How many instances are there in total?*
    - 1,485 price observations after cleaning
    - Covering 5 major retailers
    - Including multiple brands and product variants
3. *Does the dataset contain all possible instances?*
    - The dataset covers all major retail chains in Toronto
    - Some days may have missing observations due to stock unavailability
    - Geographic coverage limited to Toronto area
4. *What data does each instance consist of?*
    - Date of observation
    - Vendor name
    - Product description
    - Brand name
    - Price in CAD
    - Package size in grams
    - Calculated price per 100g

Collection process

1. *How was the data acquired?*
    - Direct observation of retail prices
    - Daily monitoring of both in-store and online prices
    - Manual verification of price accuracy
2. *What mechanisms were used?*
    - Standardized data collection protocol
    - Web scraping for online prices
    - Manual recording for in-store verification

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling done?*

- Standardization of product names and brands
- Conversion of all prices to price per 100g for comparison
- Removal of invalid entries and out-of-stock items
- Classification of product types (artisan, sliced, regular)
- Treatment of missing values and outliers (>3 SD from mean)

2. *Was the "raw" data saved?*

   - Original data preserved in CSV format
   - Raw data stored in data/01-raw_data/raw_sourdough_prices.csv
   - Cleaned data saved as parquet file in data/02-analysis_data/

3. *Is the preprocessing software available?*

   - All cleaning scripts available in project repository
   - R scripts numbered 00-07 document complete workflow
   - Uses tidyverse (Wickham et al. 2019) and arrow (Developers 2023) packages

Uses

1. *Current uses*

   - Analysis of retail pricing strategies
   - Investigation of market segmentation
   - Study of temporal price patterns

2. *Repository information*

   - Code and data available on GitHub
   - Analysis reproducible through R scripts
   - Results documented in academic paper

3. *Potential future uses*

   - Consumer behavior analysis
   - Market competition studies
   - Retail strategy research
   - Price trend forecasting

4. *Impact considerations*

   - Price data may influence market competition
   - Temporal limitations (6-month period)
   - Geographic specificity to Toronto market

5. *Usage restrictions*

   - Not suitable for individual store pricing decisions
   - Should not be used for price-fixing purposes

- Limited applicability outside Toronto market

Distribution

1. *Third-party distribution*
    - Dataset publicly available through GitHub
    - Open access for research purposes
    - No restrictions on academic use
2. *Distribution method*
    - GitHub repository
    - Parquet file format for efficient storage
    - Accompanying R scripts for analysis
3. *Distribution timeline*
    - Available immediately upon publication
    - Regular updates during study period
    - Final version released July 2024
4. *Licensing*
    - Requires attribution to original authors
    - No commercial restrictions
5. *Export controls*
    - No export restrictions apply
    - Publicly available retail information
    - No sensitive data included

Maintenance

1. *Support/hosting*
    - Maintained by original authors
    - Hosted on GitHub
    - Regular updates during active study
2. *Contact information*
    - Primary author email: [EMAIL]
    - GitHub issues for technical questions
    - Academic department contact available
3. *Erratum*
    - Updates tracked through GitHub

- Version control maintained

4. *Update schedule*
    - Monthly validation checks
    - Immediate correction of identified errors
    - Final version marked with release tag
5. *Data retention*
    - Public retail data retained indefinitely
    - No personal information included
    - Historical versions maintained
6. *Version support*
    - All versions available through GitHub
    - Version numbering follows semantic versioning
    - Deprecated versions clearly marked
7. *Contribution mechanism*
    - GitHub pull requests welcomed
    - Issue tracker for problems/suggestions
    - Clear contribution guidelines provided
8. *Additional notes*
    - Regular backup procedures in place
    - Automated testing for data integrity
    - Documentation maintained with code

This dataset provides a comprehensive view of Toronto's sourdough bread market while maintaining high standards of data quality and accessibility. The documentation and maintenance procedures ensure long-term usefulness for research and analysis.

References

Developers, Apache Arrow. 2023. "Arrow r Package." Apache.org. https://arrow.apache.org/docs/r/.

Filipp, Jacob. 2024. "Project Hammer – Jacob Filipp." Jacobfilipp.com. https://jacobfilipp.com/hammer/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4 (November): 1686. https://doi.org/10.21105/joss.01686.