Web-scraping for Health Policy Research

Cynthia Huang, PhD Candidate

16 Aug, 2023

Monash EBS Industry Engagement Workshop

Industry Data for Social Good

Company A
Retail price analytics

MNHS
Health Policy
Research

Reduce harms from alcohol via Public Health Policy
informed by insights from
web-scraped price data

MONASH
University

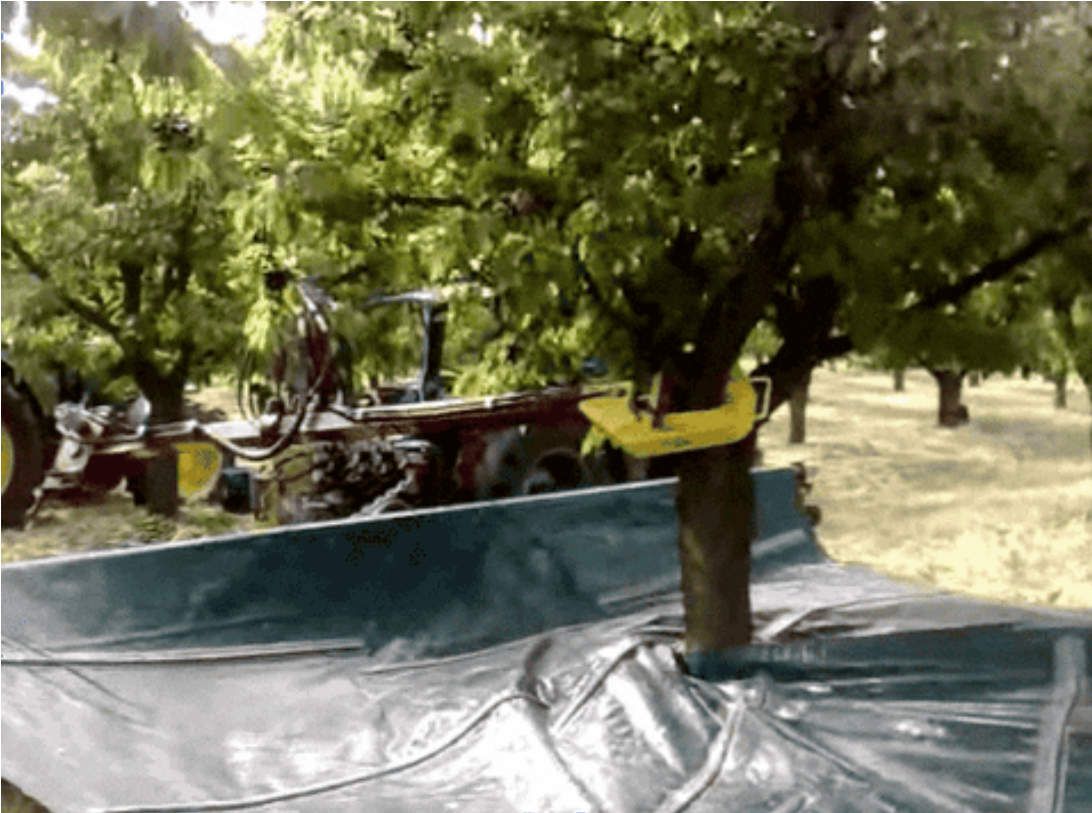# Before Online Shopping…

Traditional Data Sources

- **Examples**
  - Consumer survey data (e.g. Nielsen Homescan)
  - Manual price audits
- **Characteristics**
  - Quality controlled
  - Small scale, limited sampling frame
  - Costly to collect

MONASH University

User Beware: Web-Scraped Data
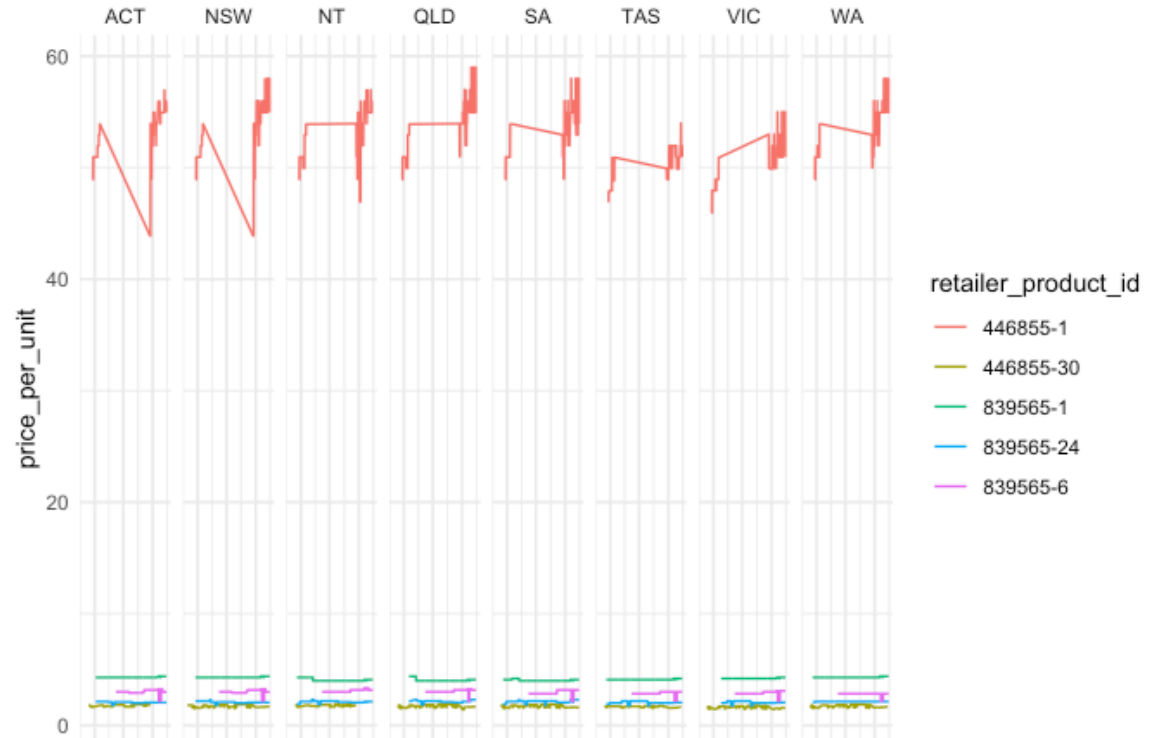
Webscraped Data

- **More coverage! More data!**
  - 4 major retailers → 25+ retailers
  - ~1 month → ~3 years of daily prices
  - 1 state → stores in all states
  - ~30,000 unique products
- **Challenges**
  - Need for cross-retailer harmonisation
  - Incomplete product information
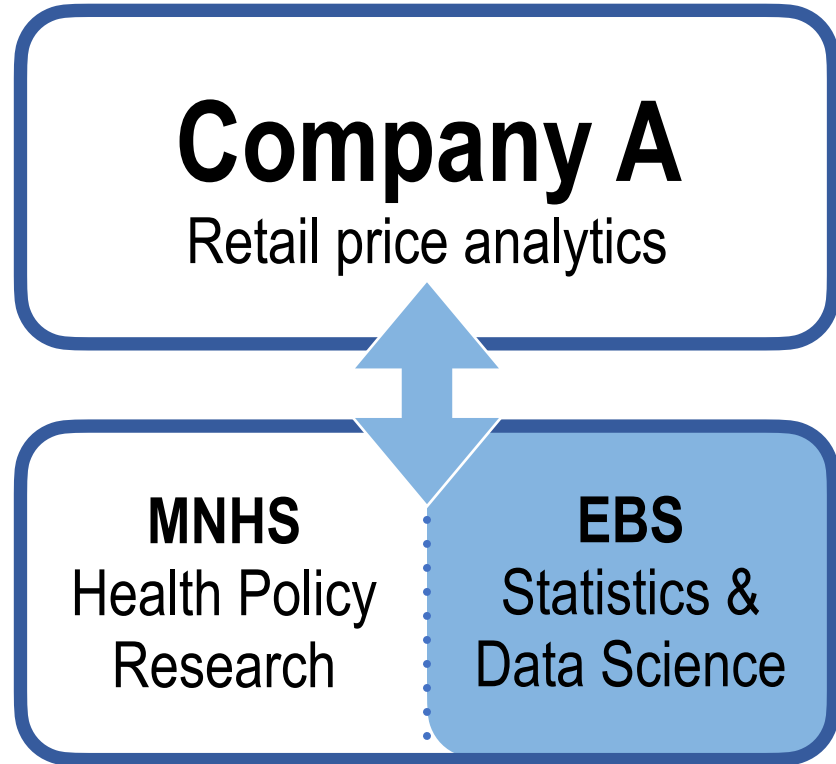 - **Overwhelming possibilities for analysis?**

MONASH University

Statistical insights for complex data

My contributions:

- **Translate and clarify** data between contexts

- Summarise and highlight key **data characteristics**

- **Align research questions** with data quality and scale

- Feed **data improvements** back to industry partner

A Multi-disciplinary Collaboration

Company A
Retail price analytics

MNHS
Health Policy
Research

EBS
Statistics &
Data Science

Reduce harms from alcohol via Public Health Policy
informed by insights from
web-scraped price data

MONASH
University

Web-scraping Expectations vs. Reality

Expectation

Reality