

Mining Iowa Liquor Sales

Team members:
Nicholas Stafford
Alexander Sueppel
Alex Thomas
Kevin Vick

Description

We intend to examine trends in alcohol sales in the state of Iowa over the past several years. We will identify longer-term trends, such as changes in preferences for certain brands or types of liquor, as well as seasonal trends in alcohol consumption. Finally, we will identify local trends and preferences in certain cities around Iowa.

Prior Work

- Weekdays/months/holidays with largest liquor sales:
 - <https://www.womply.com/blog/analysis-liquor-store-sales-data-reveals-the-biggest-drinking-days-of-the-year/>
- Types of alcohol bought compared to other beverages bought:
 - <https://www.parkstreet.com/alcoholic-beverage-market-overview/>
- Alcoholic Sales Per County (Iowa):
 - <https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales-by-Year-and-County/ahiv-u4uz>
- Preferred alcoholic beverages in Iowa:
 - <https://www.desmoinesregister.com/story/opinion/columnists/iowa-view/2020/12/18/alcohol-sales-increase-costs-for-iowans/3943579001/>
- Alcohol and college football:
 - <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3101303/>

Dataset(s)

We chose to use the Iowa Liquor Sales dataset provided by the state of Iowa (<https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales/m3tr-qhgy>). This dataset contains information on liquor purchases in Iowa liquor stores by product and date of purchase from January 1, 2012 to current. The dataset contains 22.3 million rows of data and 22 attributes, primarily containing information on date of sale, type of product, and location information.

Proposed Work: Data Preprocessing - Summary

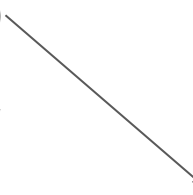
Raw dataset is mostly clean & rich

- No need for data integration
 - Many attributes and millions of data points = rich dataset to be mined
- Data transformation may be unnecessary
 - TBD during outlier analysis and data distribution exercises
 - No apparent redundancies or errors in data
- Initial data preprocessing (expected to be minimal)
 - Data Reduction (ignore several attributes)
 - Data Cleaning (outlier analysis, resolve inconsistencies)

Proposed Work: Data Preprocessing - Reduction & Cleaning

- Data Reduction:


- Ignore attribute “Store Location”
 - Missing values (lat/long coordinates of some stores)
 - Similar information already in “city” location
 - Unnecessary to be this precise geographically
- Ignore attribute “Invoice/Item Number”
 - Specific to one transaction - no relationship to others



Store Lo...
POINT (-91.4...
POINT (-93.8...
POINT (-91.7...
POINT (-93.6...
POINT (-93.3...
POINT (-96.3...

- Data Cleaning:

- Outlier analysis on numerical attributes, if necessary
 - e.g. “Bottles Sold”, “Volume Sold”
- Resolving inconsistencies
 - Varying capitalization



Polk
POLK

Proposed Work: Data Mining

- Associations/Correlations between attributes
 - e.g. time (“Date”) and quantity (“Bottle Volume”)
- Trend Analysis
 - e.g. trends over time in one location
- Clustering
 - e.g. using locations to find hot spots in data
 - Urban vs rural city differences
 - College towns

List of Tools - Python Framework

- NumPy
 - Matrix operations, base for other libraries
- pandas
 - DataFrame as core idiom, translated from source csv
 - Ease of filtering/aggregation
- SciPy
 - Descriptive statistical inquiries, distribution-specific approaches
- scikit-learn
 - Core package for data mining methods
 - Regression, classification, and clustering
- matplotlib/seaborn
 - Inspection of abstract ideas for interpretation
 - Visualization as sanity check

List of Tools - SQL (Backup Approach)

- Database development
 - csvkit: creates PostgreSQL database from source csv
 - Direct extractions from queries to csv saves intermediate work
 - Ability to bypass most of the Python framework if needed
- Query-driven analytics
 - Single-step filtering and integration
 - More supervised guidance needed for investigation
 - Less open-ended inquiry
 - Greater potential for error as math mostly written manually
 - Dramatic inefficiencies as complexity increases

Evaluation

- Bayesian Classification
 - e.g. compute probability a certain type/amount of liquor will be purchased under a specific set of conditions
- Market Basket Analysis
 - Use of Apriori Algorithm on applicable attributes
- Information Visualization
 - Assess conclusions through human-centric lens