Data Preparation and Check-In

Andrew Kerr, Isabella McCarty, Martin Hsu,
Matteo Shafer, Erik Luu

Client A

Data Collection

We took a subset of the data filtering for only the franchise “Casey’s” from 2017 to 2020 since Casey’s is a franchise with multiple locations, which fit the description of a potential client A. The years 2017-2020 were selected in order to predict 2020 sales using 2017-2019 sales. The data was pulled directly from the Iowa state government’s data website, data.iowa.gov, by applying the appropriate filters in our data cleaning step to their in-website query interface and downloading the result as a CSV.

```
SELECT * WHERE date >= '{year}-01-01' AND date <= '{year}-12-31' AND
CONTAINS(name, “CASEY’S GENERAL STORE”)
```

Data Cleaning

We selected specific columns that we were interested in: Date, County, Type of Liquor, Size of Liquor, Sale Price of Singular Bottle, Sale(Dollars), State Bottle Retail, State Bottle Cost, and Pack size. Additionally, we filtered for a specific store franchise (Casey’s) from a specific timeframe (2017-2020) and then aggregated the data so that instead of each observation being a transaction, each observation is instead a store-year-month combination.

Feature Engineering

We categorized the size of liquors into < 750 liters as small and > 750 liters as large, and similarly categorized the Sale Price of a Singular Bottle categorizing them by < $10 as cheap, $10 - $25 as mid price, and > $25 as expensive. Using the existing columns, we also determined the year, month, number of full packs sold, and number of single bottles sold in excess of whole packs for each transaction. Liquor classification was performed on the liquor’s “Category Name” to create easily accessible and generalizable categories of alcohol such as Brandy, Gin, Rum, Tequila, Vodka, and more. This will allow analysis of alcohol transactions for each type of alcohol at the storefront in the corresponding year/month. These features were created to provide more insights into what drives the most value in monthly sales.

Link to Colab Notebook

Client B

Data Collection

We selected a random sample of 50k observations from the dataset since Client B is interested in all of Iowa, not any particular store or location. Some feature engineering was done in the query to the API as seen below, such as extracting the day of the week from the date column.

```
SELECT Date, date_extract_m(Date) as Month, date_extract_dow(Date) as
DayOfWeek, NAMES AS StoreName, Zipcode, City, County, category_name,
sale_liters, sale_dollars
LIMIT 50000
```

Data was also scrapped from Wikipedia on colleges in Iowa and their 2012 enrollment numbers to be used during feature creation.

Data Cleaning

The majority of the data cleaning was done to the initial Iowa Liquor Sales dataset in the query to the API, as stated above. We selected the features listed above to serve as possible parameters in our model for predicting what drives a high or low alcohol purchase. When location features were missing or “nan”, those rows were dropped from the dataset, as some were observed when pulling the random samples. Additional cleaning was performed on the scraped data on Iowa colleges, namely renaming columns from obscure titles to those of more accurate descriptors such as “City”, and “Enrollment”. To increase the cohesiveness of the data, “City” was capitalized to match the format of the Iowa Liquor Sales observations. Other modifications included changing “Enrollment” into a numeric variable, as it was collected in such a way that characters were included in the initial scrape, and to use in later feature engineering. Initial exploration was performed on `sale_liters` in which negative values were present, thus dropped from the data as these were interpreted as “Returns” not relevant to our Client’s needs. However subsequent samples showed no trace of negative `sale_liters`, a finding we have noted for future use of the data. Furthermore, later exploratory data analysis showed issues with multiple colleges being in the same city. When attempting to map and merge the college data with Iowa Liquor Sales, additional rows were unintentionally added to the dataset that would incorrectly inflate sales data. To remedy this, a more creative approach to feature engineering was utilized.

Feature Engineering

Engaging yet intuitive features we thought to engineer for our data included: What general alcohol type was purchased, whether or not the sale was within 2 weeks prior to a government-observed holiday (information obtained through ‘holidays’ python package), the count of colleges in the city the sale took place, and the numeric and categorical student population of the city (small, medium, large), based on the typical enrollments of Iowa colleges. In preparation for model classification of a purchase, we engineered the feature `CostPerLiter`, dividing sale_dollars by sale_liters to observe expensive or cheap liquors. For each purchase, when a particular store did not have any colleges near it, the corresponding default value of choice for `Institution` and `Student Pop` was 0, while the categorical classification, `Size`, was ‘None’.

Link to Colab Notebook