J3 Iowa Liquor Consultancy

In order to meet Hawkeye Hooch’s goal of finding the best location for a new liquor store, J3 first took a look at the data given.

**Inspect the Data**

It was during this stage that we discovered that the state of Iowa acted as the alcohol wholesaler to its stores and that the prices listed in the database as retail were the prices stores paid to the state. There was no retail pricing available for the final transaction (store to consumer).

Several options presented themselves:

1. Assume a flat profit margin for all products, essentially making the retail price a proxy for the final transaction price.

2. Randomize profit margins.

3. Survey market for direction

We decided to use the first option. #2 would introduce unnecessary and inaccurate complexity and #3 would be costly in terms of time and might not be vital for outcome. #3 will, however, be listed in next steps, as it would be vital for a more accurate read of the market.

As we discussed how to use the liquor sales data given we decided that it was insufficient to answer our client’s needs. To suggest what markets should be targeted for new stores it was necessary to know more about the people in those markets.

We were able to draw demographic, housing, and business information from the Iowa Data Center website. These data were dated with some reflecting information as far back as 2004. Another next step for this project would be to secure current information. They would, however, serve for this initial model.

**Clean the Data**

After importing the liquor sales csv into Pandas we found that there was significant work to do in order to make it usable.

1. Convert “Date” to datetime format
2. To ensure accuracy, we imported a list of Iowa zip codes, cities, counties, county numbers, and zip code areas in square meters from unitedstateszipcodes.org, iowayouthsurvey.iowa.gov/images/iacountiesnumbers.pdf, and 2010 TIGER/ESRI shape files.
   1. Zip codes required considerable cleaning as there were some zip codes that were not in the state, some that did not exist, and one that was in an unreadable format ( ‘712-2’).
   2. Using the additional zip code data available to us we were able to identify and correct the majority of these.
3. All dollar-based metrics were then converted to floats.

For the demographic data cleaning was less onerous. Zip codes were converted from strings to integers and symbols and punctuation (',', '%', '$', '-') were removed from cells in order to allow them to be converted to floats. Once this data had been cleaned the two data sets were merged on zip code.

**Mine the Data**

We considered how to address the dates of the available data and determined that we should dismiss seasonality as a consideration. The question of how to optimize sales through the year is secondary to where to establish the new store.

For expediency sake we chose to set the 2016 data aside and concentrate on building the model based solely on 2015 data.

To achieve this we created a column reflecting the year and eliminated rows that held 2016 transactions.

Additional columns were added reflecting dollars per liter sold and stores per square km in each zip code.

Dollars per liter indicates where store in a zip code sit in terms of price and volume.

* A high-end store would have high price and low volume (and a high dollars per liter)
* a bargain bulk distributor would have low price and high volume (and a low dollars per liter).
* These data points will be useful if the client decides to focus on one of these formats for an expansion.

Stores per square km indicates market saturation. The model, utilizing demographic information about the zip code, will determine the extent to which an area’s saturation is a benefit or a drawback.

* A zip code that already has several liquor stores will either be:
  + oversaturated, in which case it would be a poor choice
  + a strong enough market that it might be able to support another store.
* A zip code that has no or very few liquor stores will either be:
  + a location that has no need for a liquor store
  + a location that is an untapped market

**Explore the Data**

Armed with the liquor sales and demographic data we looked at how each feature correlated with our targets of total sales and dollars per liter. For each we generated heat maps and histograms to visually understand what we could glean from our data.

[ INSERT HEAT MAPS and FIRST SET OF HISTOGRAMS ]

In doing so we found that there were a handful of stores that had extremely high sales. Investigation uncovered that at least one of these was a store that acted primarily as a distributor to bars. We chose to eliminate any store further than three SD from the median.

[ INSERT SECOND SET OF HISTOGRAMS ]

The total sales histogram described a right-skewed data set while dollars per liter resulted in a normal curve.

**Build Your Models**

Our model describes the viability of a zip code to host a new Hawkeye Hooch location using a broad set of features, primarily derived from demographic information.

We eliminated several features that we determined were either inter-dependent (if one went up or down the others would go up or down in unison) or wouldn’t be relevant to the final product.

We ran the remaining features through a model with total sales as the target. Ridge cross-validation of the model resulted in an r-squared of 0.865 and an alpha of 0.1.

[ INSERT PLOT ]

Running the features through a model with dollars per liter as the target resulted in a Ridge CV with an r-squared of 0.688 and an alpha of 0.1.

[INSERT PLOT]

**Conclusions**

According to our model, the best performing zip codes right now are:

[ INSERT TOP\_TEN TABLE]

You can see from the table that the predicted values are within X% of the actual sales from 2015 so we feel confident that this model will serve Hawkeye Hooch well as it searches for its next location.

**Next Steps**

Utilizing the demographic data was vital in providing this model’s strong results and before implementing it J3 urges Hawkeye Hooch to acquire current data as neighborhoods can undergo significant changes over time.

If Hawkeye Hooch is interested in focusing on a particular type of sales format, such as a high-end store or a bulk distributor, adding consumer retail price to these data is vital. The liquor sales data as given allows for no real insight into how profitable any store might be.

These data could also be utilized to determine the viability of service expansions, such as home delivery, or increased marketing focus utilizing ethnicity, age, or family status.

In order to promote a positive social image, Hawkeye Hooch may want to also use the data from this project to identify areas where drinking may be problematic and to introduce messaging promoting responsible consumption. The demographics data shows areas of high poverty and low employment, for example.