Sociable Cider Werks Taproom:
Craft Cider Market Analysis

Sarah Jordan
Springboard Capstone Project 1
January 21, 2018

Report Objective **3**

Recommendations **3**

The Data **4**

I. Datasets 4

II. Data Wrangling 4

Cluster Analysis **6**

I. Feature Selection 6

II. The Clusters 7

Exploratory Data Analysis **8**

I. Cluster Comparisons 8

II. Inferential Statistics 11

III. Cluster 2 Analysis 12

Final Recommendations **15**

Report Objective

Sociable Cider Werks is a cidery local to Minneapolis, Minnesota. Sociable has been rapidly growing, and its owners are looking to open a new taproom in 2018-2019. The owners would ideally like to open a new taproom in a metropolitan area outside of Minnesota with a population that is comparable to or larger than the population of Minneapolis (~400,000 people). The purpose of this report is to explore demographic features of large US cities to identify target cities for the new Sociable Cider Werks taproom.

Recommendations

I have narrowed down a long list of cities to three top recommendations for a new Sociable Cider Werks taproom:

1. Austin, Texas
2. Raleigh, North Carolina
3. Boston, Massachusetts

The Data

I. Datasets

Demographics:

- ACS_16_1YR_S1901_with_ann.csv: income data for 599 US cities, source
- ACS_16_1YR_S1901_metadata.csv: metadata the income data, source
- ACS_16_1YR_S0101_with_ann.csv: age data for 599 US cities, source
- ACS_16_1YR_S0101_metadata.csv: metadata for the age data, source
- ACS_16_1YR_B01003_with_ann.csv: population data for 599 US cities, source
- State_Regions.csv: region divisions for US states, source
- Price_Index.csv: cost of living index data for US cities, source

Cider Performance:

- cider_data.xlsx: cider shipment information for all 50 states, source
- Cidery_Locations.csv: all cideries in the United States by city and state, source

II. Data Wrangling

All code for data wrangling can be found **here**. The major data wrangling steps are outlined below:

US Census Bureau:

- *Renaming columns using the metadata*: all US Census Bureau data columns were labeled with codes. Using the metadata for each dataset, I renamed the columns based on the column descriptions.
- *Removing unnecessary columns:* After merging all of the demographic data into a single dataset, there were many unnecessary columns that made handling the dataset cumbersome. I systematically removed these columns for a more manageable dataset.
- *Removing null values:* there was only one row with any null values, and all of the entries were null. This city was The Villages CDP, Florida, and this does not seem like a potential target for a Sociable Cider Werks taproom so I eliminated this row.
- *Filtering based on size:* Sociable Cider Werks is only interested in cities with populations comparable to or larger than Minneapolis. I filtered the data to include only cities with populations greater than 400,000 people.
- *Merging dataframes:* US Census Bureau geographies include city and state in the same column. City and state are in separate columns in the cidery performance data, so I split

the US Census Bureau column into two so that I could merge all of my data into one dataframe.

Cider Performance Data:

- *Merging dataframes:* There was a leading space before city and state names in the cider performance data. I deleted this so that I could merge all of the data into one dataframe.
- *Rearranging cidery location data:* Cidery location data included all the names cideries in the United States and their cities. I had to rearrange this data by city with a count of the number of cideries in each data. I then had to convert all null values to 0's.

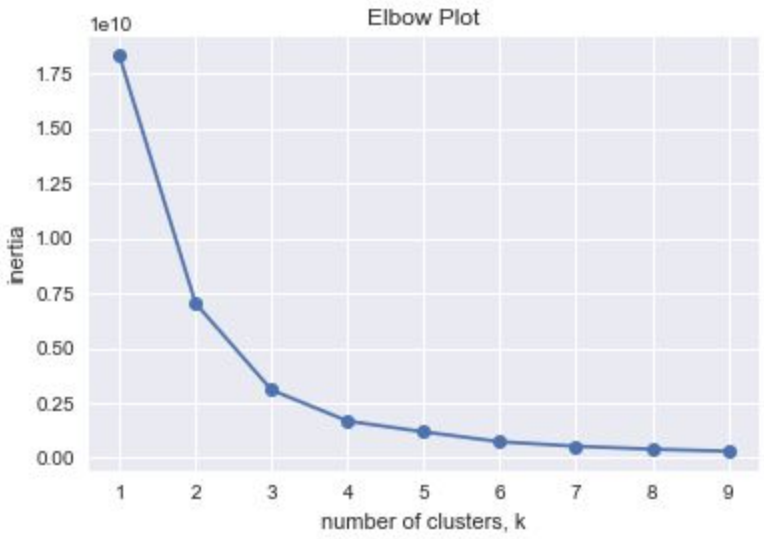
Cluster Analysis

I. Feature Selection

I performed a K-Means cluster analysis to cluster cities into groups based on demographics. The code for this analysis can be found here. The demographic features I included in the analysis are:

- *Log of population:* I wanted to include population in my cluster analysis, but there were such major outliers like New York City and Los Angeles that this skewed the clusters so they were based mainly on population. I decided to take the log of population so that population held some weight in the analysis, but it did not entirely control the cluster breakdown.
- *Age:* Sociable Cider Werk's consumers are typically in their 20s and 30s, so include city-level data on both median age and the percent of the population in their 20s and 30s in the analysis.
- *Median income:* Exploring the population's income is important when considering potential target cities for a new cider taproom. Craft cider is a luxury, and it is essential to consider whether a city's population have enough disposable income to support a new cidery.
- *Sex ratio:* According to Bart Watson, the economist for the Brewer's Association, Cider is typically more popular among females, so I include sex ratio data for all cities. This measures the number of males for every 100 females.

After clustering on these features, I decided to proceed with three clusters based on the inertia plot below:



II. The Clusters

I wound up with three clusters with the following cities:

Cluster 0:

- San Francisco, CA
- San Jose, CA
- Washington, D.C.
- Seattle, WA

Cluster 2:

- Long Beach, CA
- Los Angeles, CA
- Oakland, CA
- San Diego, CA
- Denver, CO
- Atlantic, GA
- Chicago, IL
- Boston, MA
- Minneapolis, MN
- New York, NY
- Charlotte, NC
- Raleigh, NC
- Portland, OR
- Austin, TX
- Virginia Beach, VA

Cluster 1:

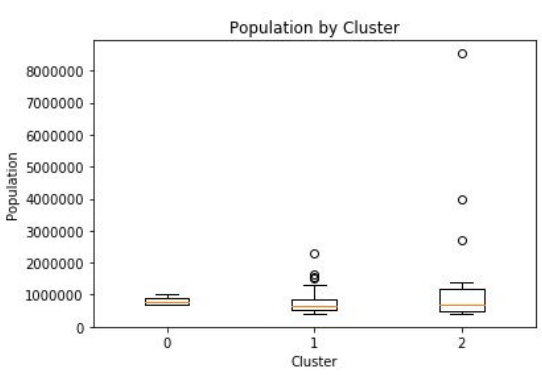
- Mesa, AZ
- Phoenix, AZ
- Tucson, AZ
- Fresno, CA
- Sacramento, CA
- Colorado Springs, CO
- Jacksonville, FL
- Miami, FL
- Indianapolis, IN
- Louisville, KY
- Baltimore, MD
- Detroit, MI
- Kansas City, MO
- Omaha, NE
- Las Vegas, Nevada
- Albequerque, NM
- Columbus, OH
- Oklahoma City, OK
- Tulsa, OK
- Philadelphia, PA
- Memphis, TN
- Nashville, TN
- Dallas, TX
- El Paso, TX
- Fort Worth, TX
- Houston, TX
- San Antonio, TX
- Milwaukee, TX

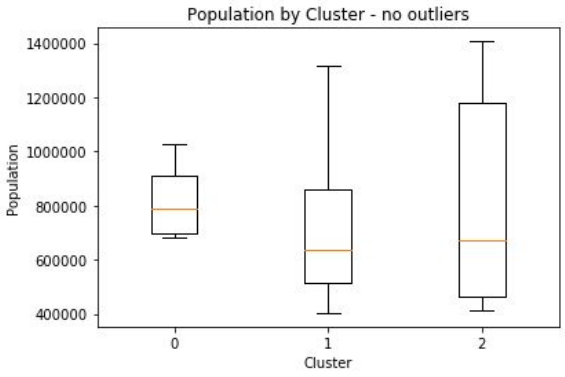
Exploratory Data Analysis

I. Cluster Comparisons

Population:

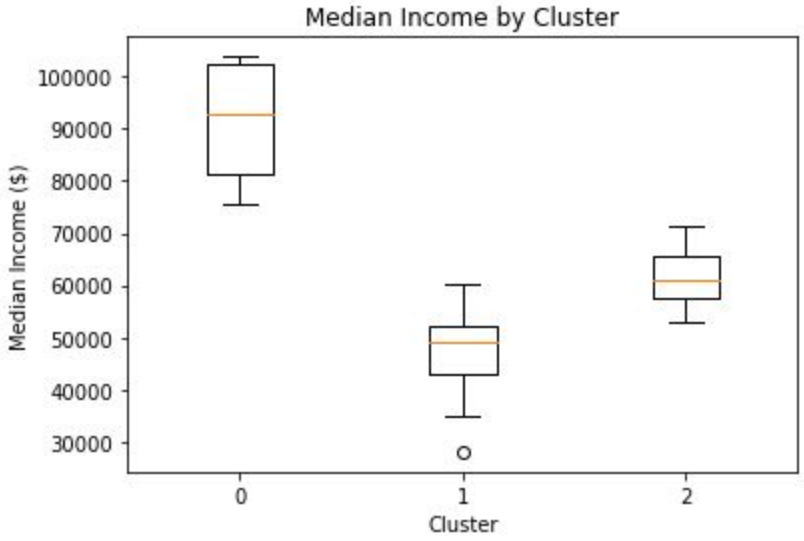
There are some major outliers that make the overall cluster trends difficult to see. After removing the outliers, we see that three clusters are fairly similar. Cluster 0 has the highest median population, and Cluster 2 has the widest range of populations.





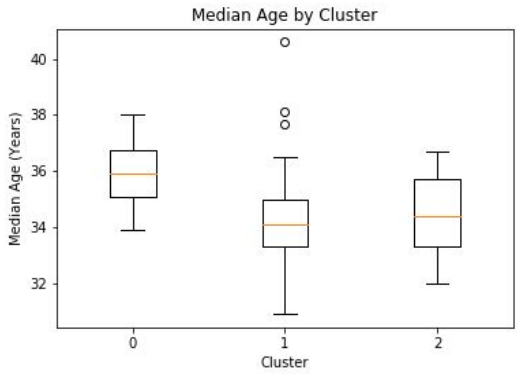
Income:

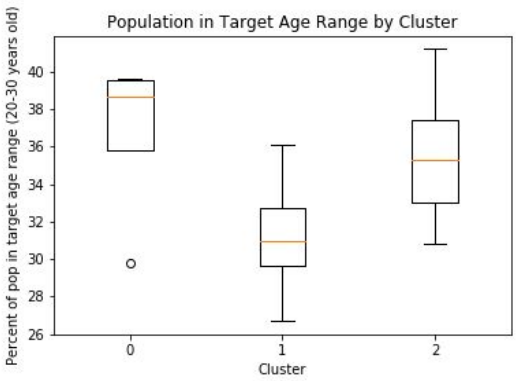
With a median median income of about $90,000, Cluster 0 has the most compelling demographic spread in terms of income. Cluster 1 clearly has cities with lower median incomes than the other two clusters, which may make it a less desirable cluster of cities to focus on when picking a target city.



Age:

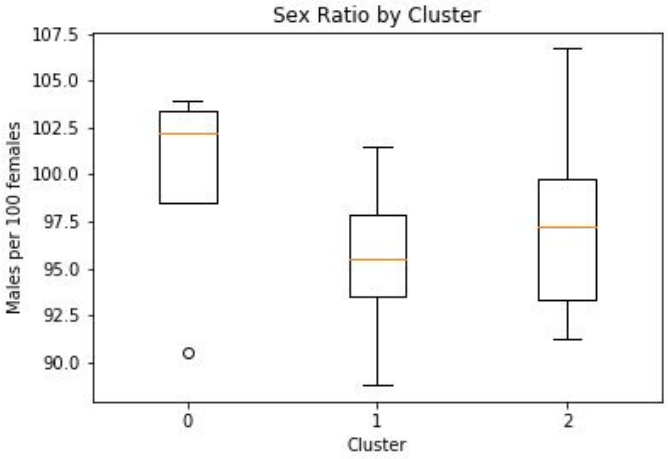
All of these clusters have fairly similar median ages for their populations. However, when we explore the percent of these populations in our target age range (20-39 years old), we see that Cluster 0 and Cluster 2 both have a high percentage of their populations in their 20s and 30s. Once again, Cluster 1 falls flat, with a far smaller proportion of its population in a prime age range, which could further affect these cities' abilities to support a cider taproom.





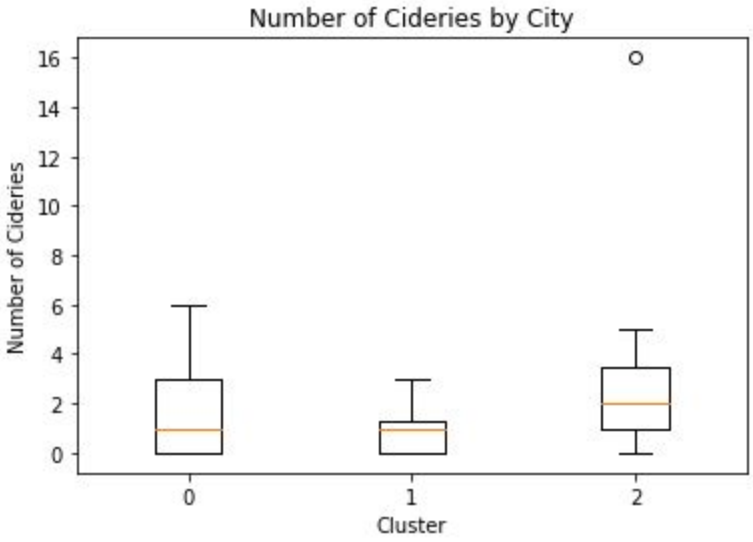
Sex Ratio:

Cider is typically more popular among females than males. Cluster 0 appears to include a larger share of cities that have a higher number of males than females than the other two clusters. While this may not be ideal for a cidery, the population of Cluster 0 is also young and high-earning, so I would not discount Cluster 0 for this reason.



Cidery Prevalence:

Based on the boxplot below, Cluster 0 and Cluster 2 continue to like strong candidates to focus on. Cluster 2 has a major outlier: one of its cities has 16 cideries (unsurprisingly, it's Portland, Oregon).



Code Source:

All of the code to generate the graphs above can be found here.

II. Inferential Statistics

After exploring the data graphically, I am drawn to both Cluster 0 and Cluster 2, but there are ultimately only subtle differences in the demographic distributions of these clusters. I need to explore whether there are statistically significant differences in the prevalence of cideries between these three clusters to help decide which city or cities Sociable Cider Werks should consider for their new taproom. See my code here.

Summary Statistics:

	Cluster 0	**Cluster 1**	**Cluster 2**
Mean # of Cideries	2	0.89	3
Median # of Cideries	1	1	2
Percent of Cities with 1 or more Cideries	50%	53.6%	80%

The summary statistics above make Cluster 2 an immediate standout. However, I took some precautions to make sure that the differences in cidery prevalence between these clusters was statistically significant.

ANOVA and Tukey's Test:

I performed a one-way ANOVA on the following hypothesis:

$$H_0 : \mu_{cluster0} = \mu_{cluster1} = \mu_{cluster2}$$

Ultimately I was able to reject the null hypothesis: with a p-value of 0.03, there is a statistically significant difference in the mean number of cideries between at least two of the clusters. After performing a Tukey test, I found that there was a statistically significant difference between Cluster 2 and Cluster 1.

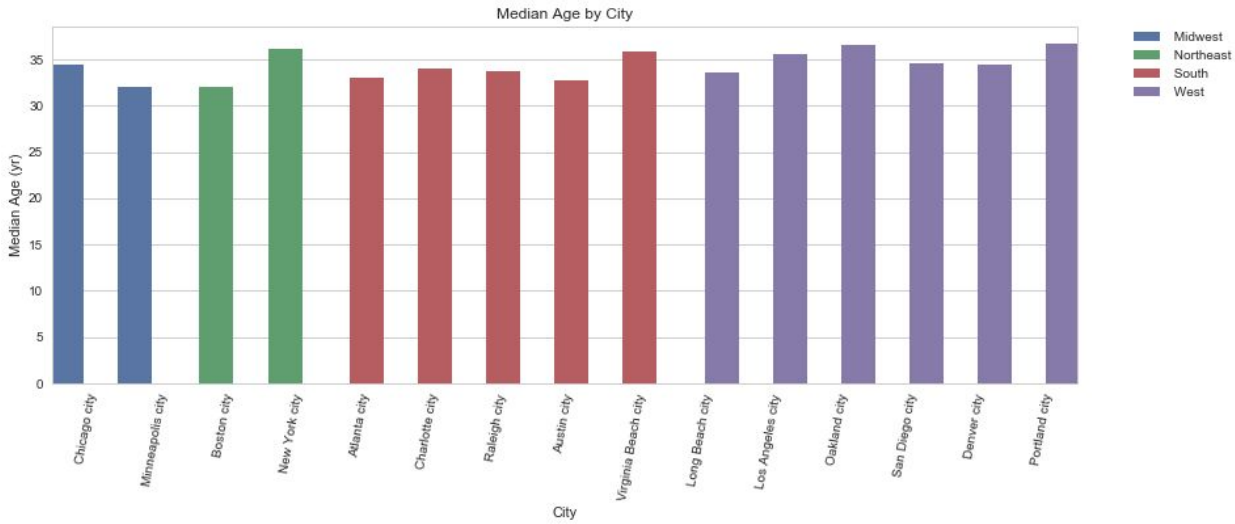
When comparing Cluster 0 and Cluster 2, the higher percentage of cities with a cidery in Cluster 2 (80%) than in Cluster 0 (50%), compelled me to recommend a city from Cluster 2. However, I decided to proceed with permutation sampling to see how likely that this great of a difference in cidery prevalence could be observed just by chance.

Permutation Sampling:

I made 10,000 random permutations of two clusters of the same lengths as Cluster 2 and Cluster 0 to find the probability that you would find a 30% difference in the number of cities with cideries in Cluster 2 and Cluster 0 simply by chance. With a p-value of 0.0001, I was able to reject this hypothesis. This test confirms that somehow the mix of demographic features that define Cluster 2 create an strong market for craft cider.

III. Cluster 2 Analysis

Age:



Excluding Minneapolis, the five cities with the lowest median age and the five cities with the highest percentage of the population in their 20s and 30s are:

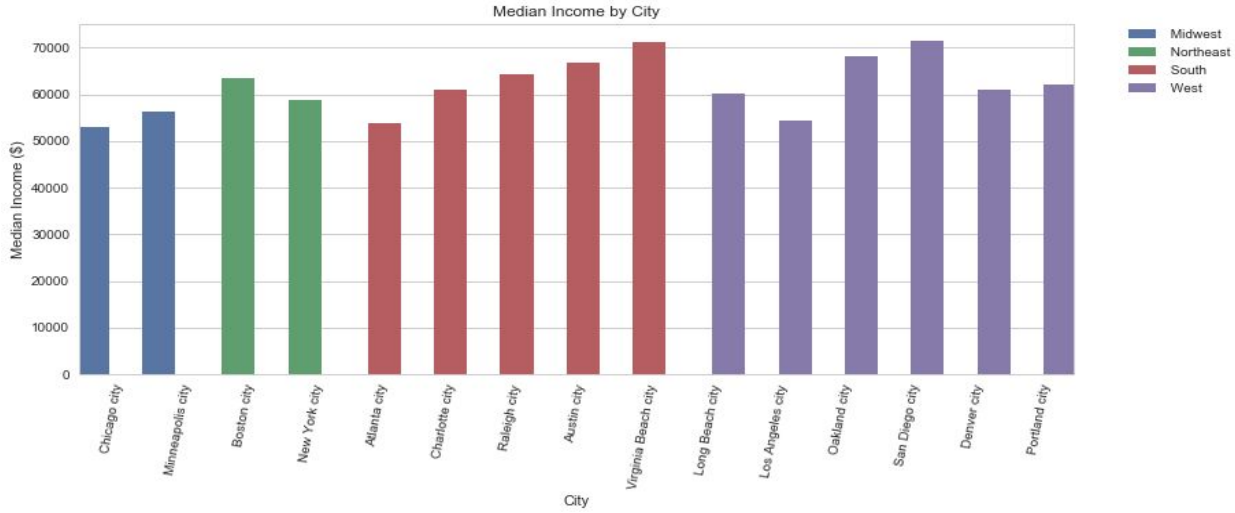
Top 5 Candidates Based on Median Age:

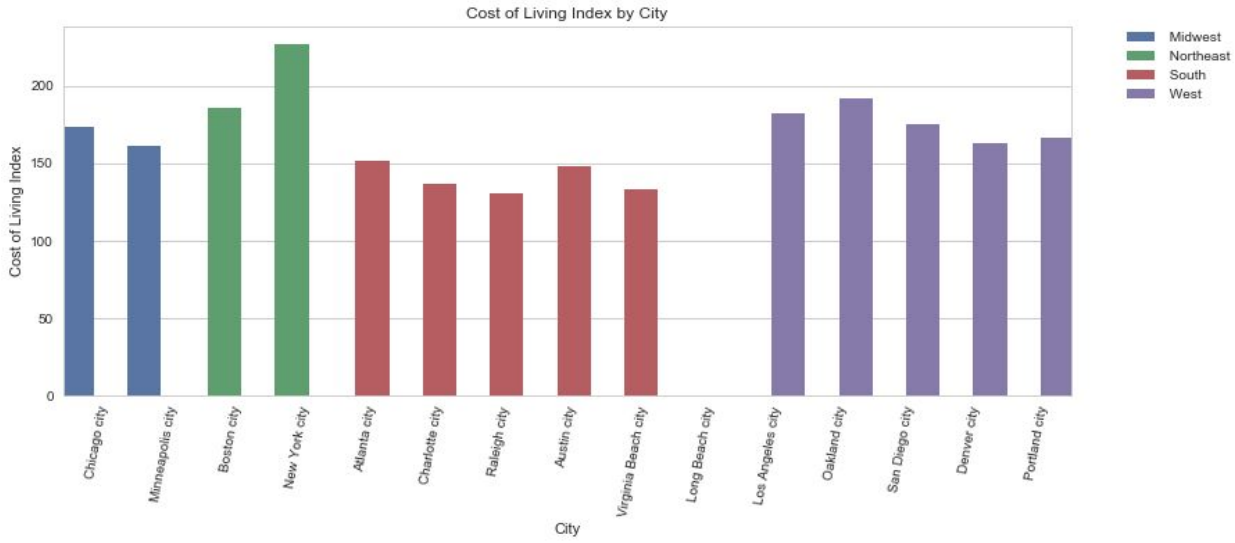
- Boston
- Austin
- Atlanta
- Long Beach
- Raleigh

Top 5 Candidates Based on Target Age:

- Boston
- Austin
- Atlanta
- Denver
- Raleigh

Income and Cost of Living Index:





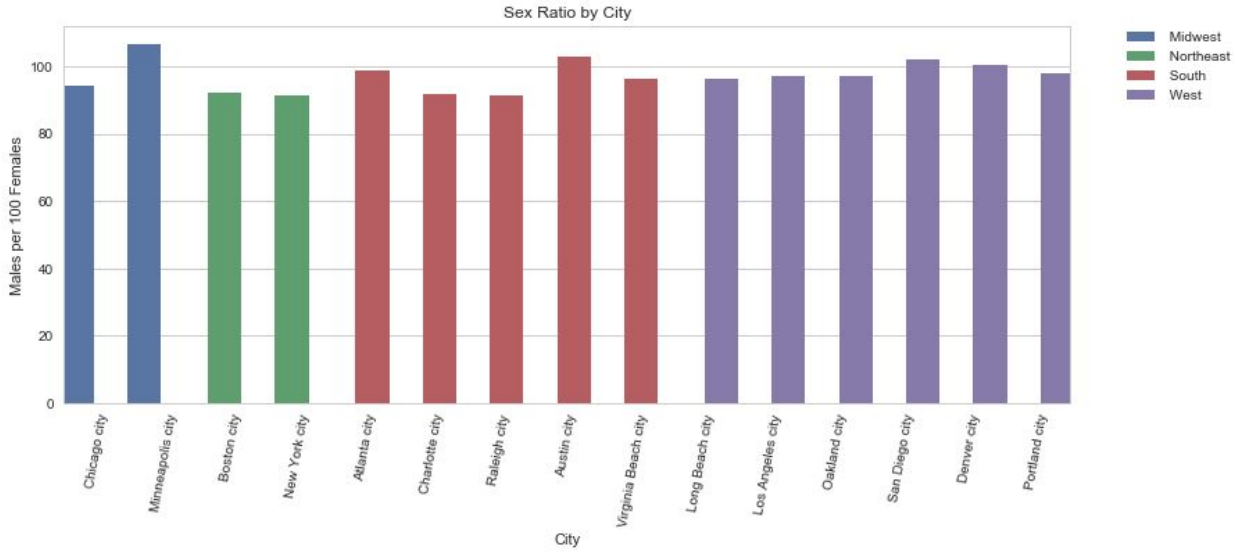
Top 5 Candidates, Median Income:

- San Diego
- Virginia Beach
- Oakland
- Austin
- Raleigh

Top 5 Candidates, Cost of Living Index:

- Raleigh
- Virginia Beach
- Charlotte
- Austin
- Atlanta

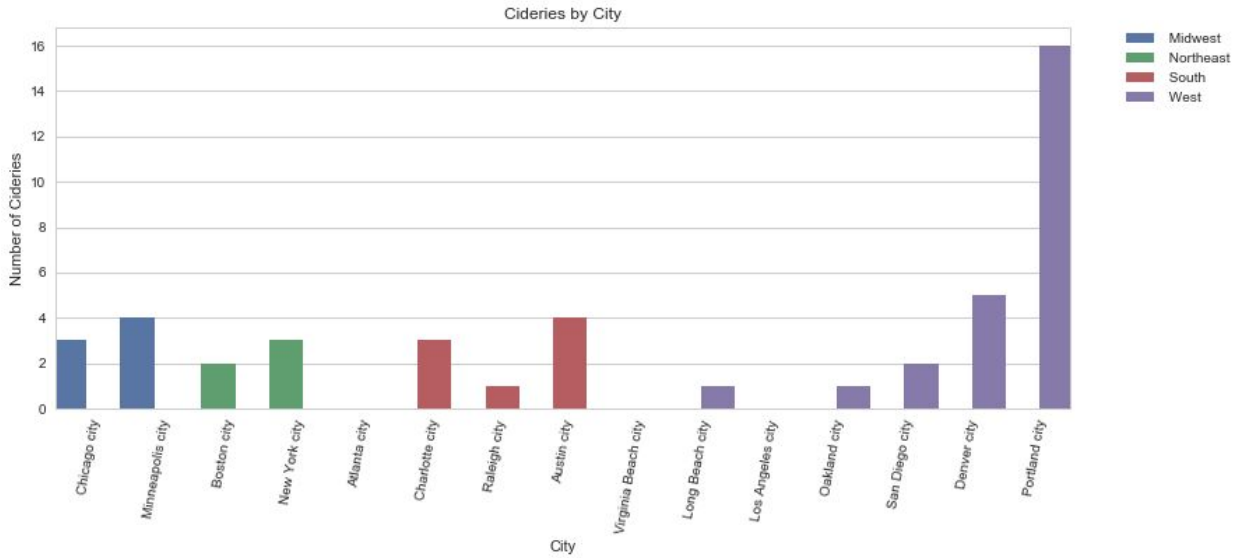
Sex Ratio:



Top 5 Candidates Based on Sex Ratio:

- New York City
- Raleigh
- Charlotte
- Boston
- Chicago

Cidery Prevalence:



Top 5 Candidates Based on Cidery Prevalence:

- Portland
- Denver
- Austin
- Charlotte
- Chicago

Final Recommendations

Considering the cities above, Austin, Raleigh, and Boston are my top three recommendations for the new Sociable Cider Werks taproom.

Austin, Texas:

Austin has a young population with a high median income, while the price of living remains low, which means these high-earning individuals have lots of disposable income. Furthermore, Austin has demonstrated success in the craft cider market with four cideries currently in business. Overall, Austin looks highly similar to Minneapolis in terms of age distribution, cidery prevalence, and the ratio of males to females. However, its population has a higher overall median income with a lower price of living, which could help make a cidery even more successful in Austin than in Minneapolis.

Breakdown:

- Median Age: 32.7
- % of Population in 20s and 30s: 39.1%
- Median Income: $66,697
- Males per 100 Females: 102.9
- Price of Living Index: 163 (4th lowest in this cluster)
- Number of Cideries: 4

Raleigh, North Carolina

Similar to Austin, Raleigh has a young population and a high median income. Raleigh has the lowest cost of living index of all of these cities, so its population has plenty of disposable income for craft cider. Cider is typically more popular among females, and Raleigh has more females than males. Raleigh has only one cidery so far -- this could be seen as a warning or an exciting opportunity. Bart Watson, the economist for the US Brewer's Association, notes that the best place to open a new cidery has typically been in cities with other cideries. While there aren't many cideries here yet, Raleigh has the right demographic distribution, so it could be a great place to start carving out the craft cider industry alongside the other existing cidery.

Breakdown:

- Median Age: 33.8
- % of Population in 20s and 30s: 35.8%
- Median Income: $64,456
- Males per 100 Females: 91.3
- Price of Living Index: 131 (the lowest in this cluster)
- Number of Cideries: 1

Boston, Massachusetts

Boston tops these rankings in the age category, and a young population is important for the craft cider scene. While Boston doesn't make the Top 5 list for median income, it does come in sixth for median income, and it is ranked second in this cluster for mean income. However, the high cost of living in Boston is extremely high, which could affect the population's overall disposable income for craft cider. This could also indicate potential high startup costs for a new business.

Breakdown:

- Median Age: 32.1
- % of Population in 20s and 30s: 41.2%
- Median Income: $63,621
- Males per 100 Females: 92.4
- Price of Living Index: 186 (the second highest in this cluster)
- Number of Cideries: 2