
Iowa Liquor Distribution in 2021: Analysis and Product Recommendations

DataSci W200 Final Project - Spring 2022

Mike Varner, Christopher Ratsimbazafy Da Silva, and Ashkaan Moinzadeh

Github: https://github.com/UC-Berkeley-I-School/Project2_Moinzadeh_Ratsimbazafy_Varner

Consolidated iPynb File:

https://github.com/UC-Berkeley-I-School/Project2_Moinzadeh_Ratsimbazafy_Varner/blob/f3989c90ad14d2f7dffe6a114471d975787adc23/Project_2_Final.ipynb

Presentation:

https://github.com/UC-Berkeley-I-School/Project2_Moinzadeh_Ratsimbazafy_Varner/blob/f3989c90ad14d2f7dffe6a114471d975787adc23/Iowa%20Liquor%20Distribution%20in%202021.pdf

Target Audience:

Alcohol retailers in Iowa who are invested in maximizing their sales revenue by surveying the state's alcohol sales profiles.

Introduction:

The purpose of this report is to analyze liquor distribution data, from the Iowa Department of Commerce, to understand recent industry trends and to make product recommendations for end consumer retailers. To do this, we've acquired 2021 sales data from the state of Iowa logged by end consumer retailers. The data corresponds to 2.6M transactions that each account for 24 variables. In the following report, we will: outline our research questions, assess data quality, describe our assumptions, contextualize visualizations, and derive insights.

Questions Answered in This Report:

In order to make product recommendations to alcohol retailers, we must first understand some basic facts about retail liquor distribution. We will investigate the following questions:

- Are liquor sales seasonal in nature? If so, are these seasonal phenomena consistent across the various subcategories of liquor?
- Which geographic locations sold the most liquor? Do sales volumes align with population densities? We can address this by assessing total volumetric and dollar-wise sales within each geographic region.
- Which stores sold the most liquor? Do the highest selling stores demonstrate any patterns?
- Which products should retailers consider stocking in the future to increase sales? We will be inferring what is likely to be popular in the future from these data (trend following).

Questions for Future Research (Not Answered in This Report):

- Which types of liquors are the most profitable? This will be from the state of Iowa's perspective as our data reflect sales to retail locations and not to end-consumers.
- How much demand for a specific product should a retailer expect in a given timeframe? We can employ machine learning models to forecast demand of a specific product based on historical data.

Data Surveying and Preliminary Assessments:

To assess liquor sales on a state level, we examined data collected by the Iowa Department of Commerce, Alcoholic Beverages Division through a public dataset titled “Iowa Liquor Sales” (Iowa Liquor Sales, 1). All liquor sales logged in the dataset originate from sales made by establishments possessing class E liquor licenses, which grant the sale, and not consumption, of spirits on premises (Chapter 17 Class “E” Liquor Control Licenses - Iowa, 1).

The full dataset made available on *Iowa Data* is written in tabular format and is composed of 23.6 million row observations and 24 column variables. The data has been collected since 2012-01-01 and is contemporaneously logged on a monthly basis. A data dictionary is listed on the website which provides brief descriptions of each column variable, and a query tool can be used to filter and pull data by specified query parameters.

To provide insights that generalize to the entirety of Iowa and provide a recent snapshot of sales, we queried the raw data to lie between 2020-12-31 and 2021-12-31 and encoded it as a CSV file. The file was read through Python’s *Pandas* package and underwent a preliminary examination to determine efficacy of the query tool as well as fidelity of the dataset itself.

The data was successfully queried using a datetime criterion, yet the datetime format was not preserved upon conversion into CSV. We scripted a datetime-parsing lambda function to serve as an argument to Pandas’ csv reading function, and stored the queried dataset into a Pandas dataframe. To ensure the proper range of datetimes were queried, the minimum and maximum datetime values were assessed on the dataframe’s “Date” column. The minimum and maximum corresponded to the query dates specified above. The dataframe’s shape and column composition were likewise assessed and were found to correspond with the query expectations.

A column-wise percentage of present versus absent data was calculated using value counts, and all categories except for Store Location were missing only 0.01% of their data. Store Location, by contrast, was missing 11.87% of its entries.

The dataframe’s data types, as seen in *Tables 1* and *2*, are listed both by quantity as well as by type. It was found that the zip code, county number, and vendor number all needed to be converted to integers. Likewise, state bottle costs, retail costs, and sales in dollars needed to be reformatted into 2-decimal floats, which corresponds to USD currency. Volumes sold in both liters and gallons would likewise be converted to 2-decimal floats. The store location, which provides geographical coordinates for each store, was stored in a tuple format but was eventually destined for removal due to the presence of more relevant geographic variables.

To properly address the above research questions, a column-wise exploration was required to identify inconsistencies in data entry and trends between columns. This step was vital to ensuring that variables were reformatted to more accurately reflect trends in logged data when feasible. Given that our research questions required information about temporal and geographic distributions of sales, and likewise demanded proper identification of stores, products and vendors, we conducted seven major column explorations by generating hypotheses about the data. All seven explorations are detailed in *Table 3*.

Reformatting Strategy:

Variables requiring integer-based typification were converted into integer format using the *Pandas astype* method. For monetary variables requiring conversion to dollar-based formats like “State Bottle Cost,” “State Bottle Retail” and “Sale (Dollars),” the *Pandas round* method was applied uniformly on these variables to round them to two decimal places. “Volume Sold (Liters)” and “Volume Sold (Gallons)” underwent the same two decimal formatting.

All “Item Number” and “Item Description” were stored in a dictionary. “Item Number” was assigned to the dictionary’s keys, and “Item Description” to its values. This was achieved by grouping values by “Item Number,” selecting by “Item Description,” and applying the *Pandas to_dict()* method. To ensure that all items of a corresponding “Item Number” possessed a single “Item Description,” the dictionary was modified to include only one value per key. A lambda function and item dictionary were then applied back on the master dataset to reassign the dataset’s “Item Description” values with the proper “Item Number” values.

Finally, given that data entry errors for “County” corresponded to differences in upper- and lowercase attributions, all “County” values were converted into lowercase to ensure proper grouping.

Rationales Against Various Reformatting Cases:

When examining the relationship between “Store Name” and “Store Number,” we found that 1.1% of “Store Number” values had more than one “Store Name”. *Table 4* illustrates the nature of this issue. The majority of these misattributions are due to some commonality between the “Store Name” values, such as the location referenced in the store name of the broader store chain. These 1.1% appear to be data entry errors, but the pure cause is unclear. It is equally possible that these stores could have moved locations, changed names, or merged. Since there is no clear explanation for why the “Store Name” values are conjoined, “Store Name” will not be modified. For similar reasons, “Vendor Name” will remain unmodified given that only 2.7% of “Vendor Number” values have more than one “Vendor Name,” which is an insignificant relative to the total sample size.

Upon examining the variable “Volume Sold (Liters)”, we concluded that the variable “Pack” does not accurately describe each transaction. We attempted to calculate the “Volume Sold (Liters)” using “Pack” and found that “Pack” and “Bottles Sold” were not the same in a given transaction. This is because retailers can, and do, purchase partial packs of liquor. Ultimately, we care about the volume, bottle type, and type of liquor sold, and not about the “Pack” size that retailers purchased. For these reasons, we have removed this variable from the data.

To reiterate, these data are well populated with all variables but “Store Location” having fewer than 0.01% missing values. Given the magnitude of these missing values, we did not remove observations from the data as they would have minimal impact on our calculations. By contrast, we did remove, as seen in *Table 5*, a number of variables that were irrelevant in the final analysis. In brief:

- Overly-specific geographic variables were dropped in preference for broader ones.
- All numerical keys were dropped, since they only served as basis to measure fidelity in data entry.
- The remaining variables were dropped due to irrelevance in the final analysis.

Graphical Interpretations of Final Data:

Seasonal Trends within Iowa Liquor Sales

We can see how the daily total sales in dollars changed throughout the year (*Figure 1A*). Daily liquor sales ranged from \$277 to \$2,903,828, and fluctuated on what appears to be a

weekly basis. The day with the lowest sales was Nov 14th and the highest sales day was Oct 10th. On first glance, we suspect that the sales on Nov 14th were likely miscoded, given daily sales averaged \$1,427,352. Visually, there does not appear to be a strong seasonality effect at least when observed at the daily level. We could have also measured total daily liquor sales by the total amount of liquor sold in liters. However, we found an extremely strong positive correlation of 0.99 between these two measures. While there would be little added benefit of comparing these variables visually, since both variables would demonstrate equivalent distributions irrespective of normalization, there is still value in knowing both variables scale with each other consistently, with one variable's changes not deviating drastically from changes in the other.

Monthly plots of volume-wise liquor sales (*Figure 1B*) likewise reveal consistent volumetric outputs, with a monthly volumetric average of \$2.06M +/- \$174,534.34. This metric was achieved by grouping by month, summing "Volume Sold (Liters)" across each month, and calculating the mean and standard deviation across all 12 months. With further inspection, the months of January and February yielded volumetric sales that were 1.94 and 1.66 sd below the average, respectively. Similarly, June and December yielded rates that were 1.17 and 1.62 sd above the average, respectively. Hence, we argue that vendors should expect a dip in sales in the former two months, and a spike in sales in the latter two.

Figure 1: General Profile of Liquor Sales in Iowa Throughout 2021

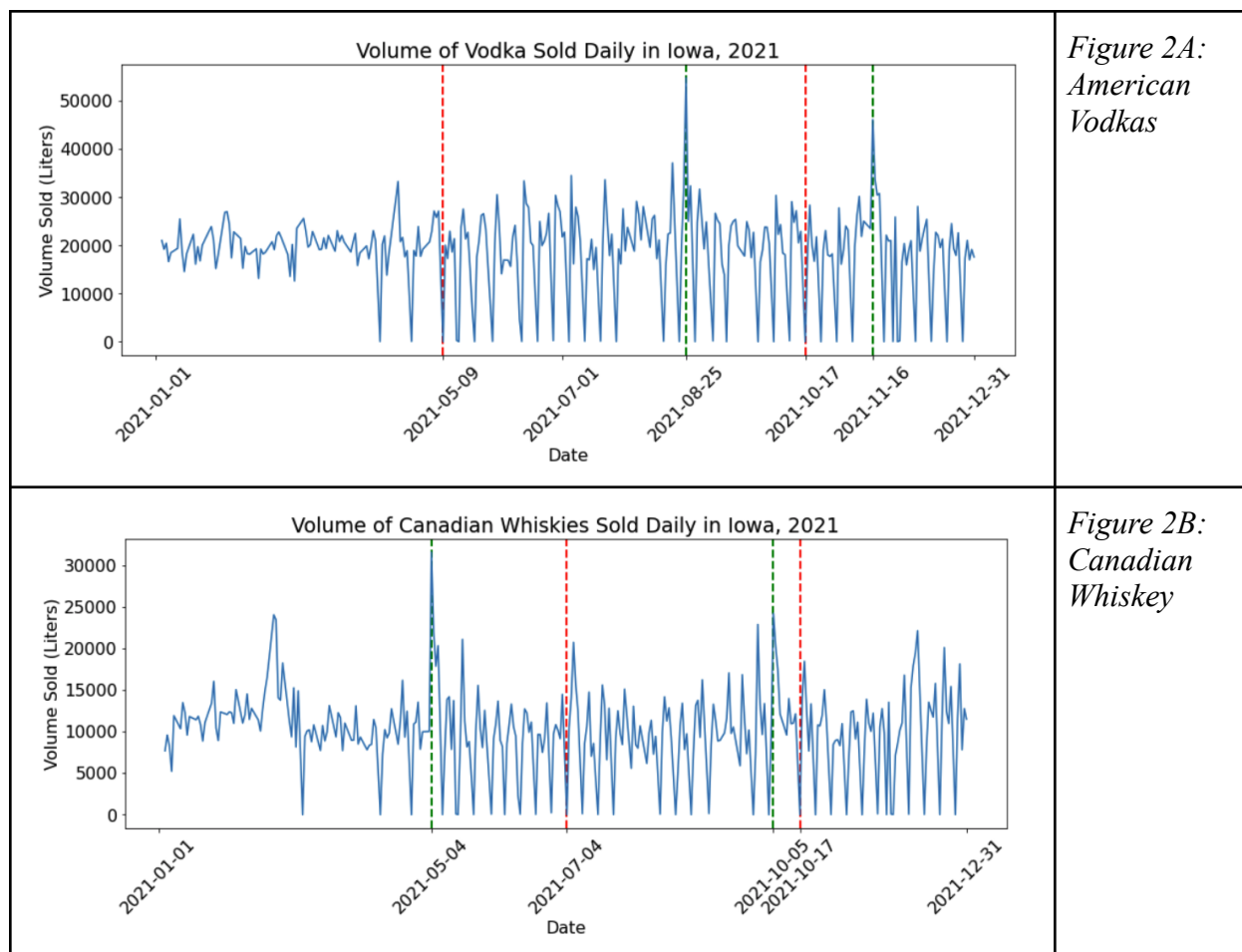


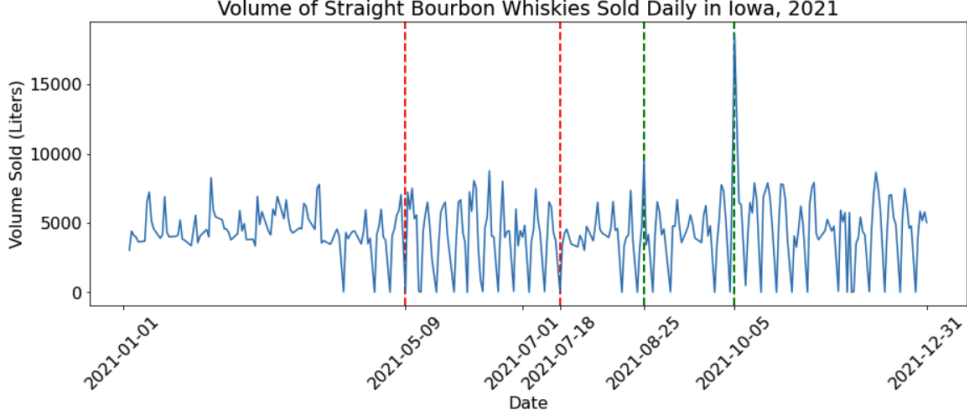
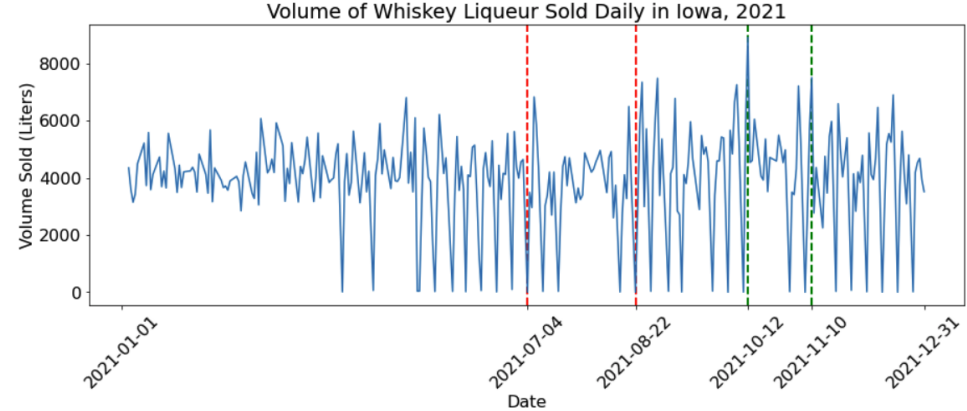
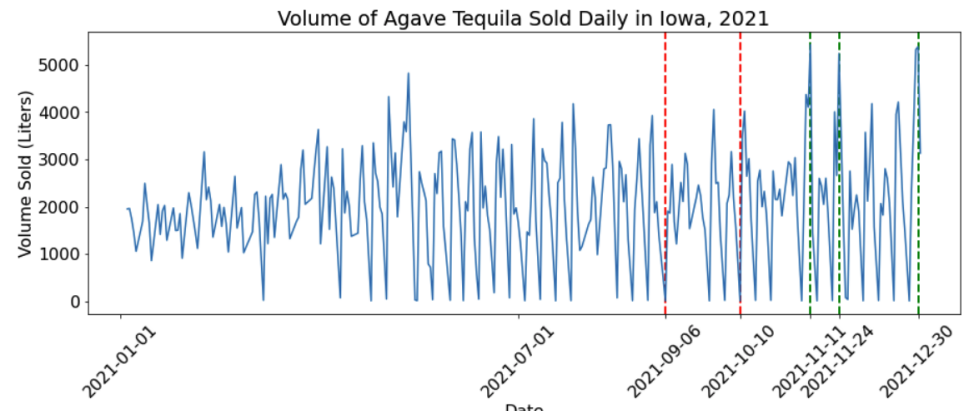
Daily Trends for The Top 5 Best Selling Liquors

Figure 2A depicts the volumes (in liters) sold daily for the top liquor category (in terms of annual sales). The dashed vertical lines indicate peaks and troughs within daily volumes sold, with red lines corresponding to troughs and green lines to peaks. By way of example, the daily sales volumes for all American vodka brands peaked on 08/25 with 54,846.18 liters sold state-wide, and fell to their lowest levels on 10/17 at 5 liters. Across all popular liquor types, we can see large swings between high and low sales volumes. These swings follow a weekly cadence, and coincide with the fact that all of the “worst” liquor sales dates fell on Sundays.

By supplement, Figures 2B-2E show volumes sold for the remaining top liquors. Amongst various whiskies (i.e. Canadian Whiskey, Straight Bourbon Whiskey, Whiskey Liqueur), many of their “best” sales dates coincided with one other on Columbus Day Weekend (10/8-10/11) and Veterans Day (11/11). Lastly, agave tequila sales tended to peak around all major U.S. holidays, which indicates that retailers should overstock their shelves with tequila products ahead of holidays.

Figure 2: Monthly Sales for The Top Five B Liquor Categories (Sold in Liters)



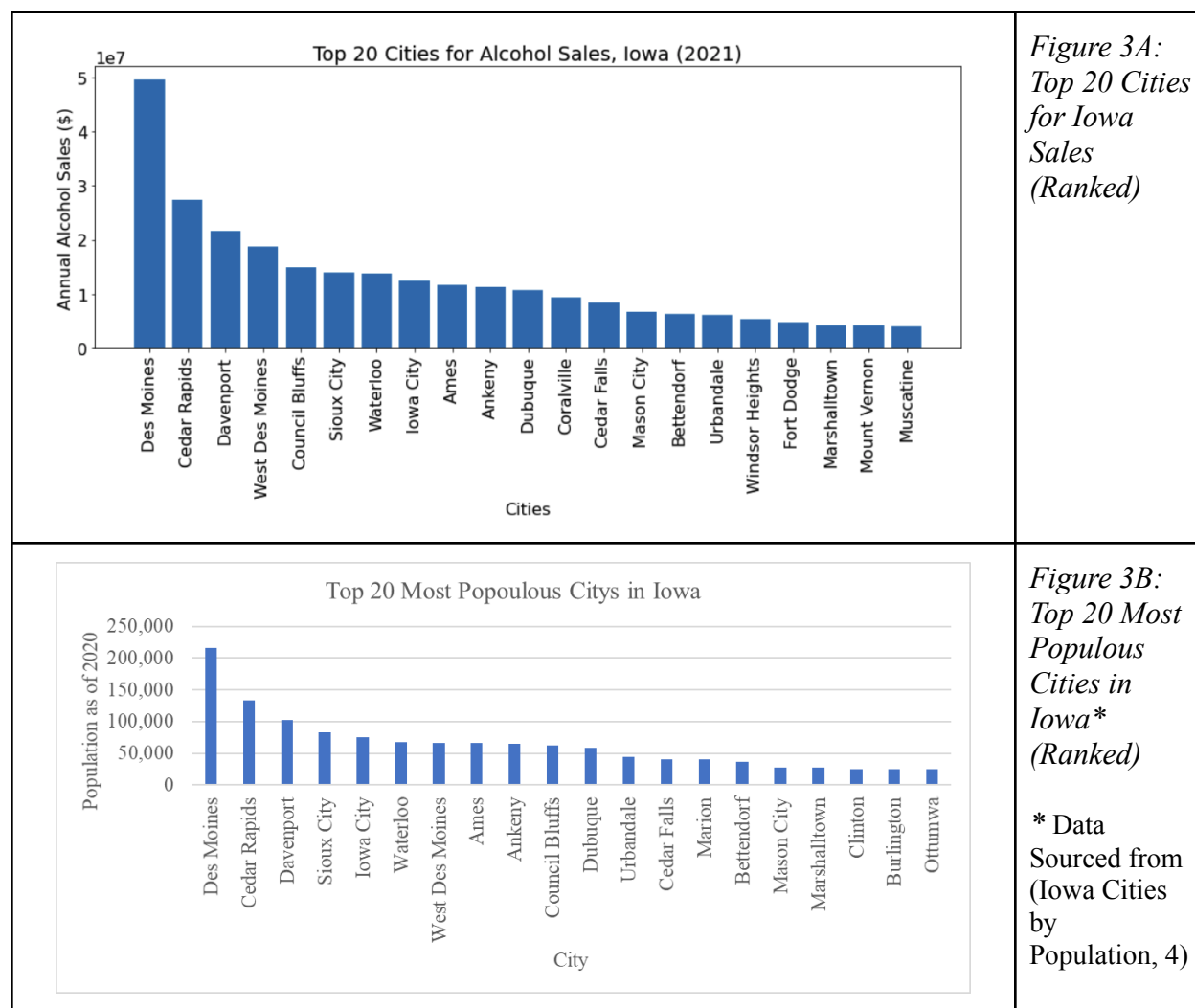
	<p><i>Figure 2C: Straight Bourbon Whiskey</i></p>
	<p><i>Figure 2D: Monthly Whiskey Liqueur Sales (Volume Sold in Liters)</i></p>
	<p><i>Figure 2E: Agave Tequila</i></p>

Graphical Comparison Between Top 20 Populous, and Liquor-Selling, Cities

When comparing the sales by city barchart in *Figure 3A* with the city population barchart in *Figure 3B*, we can confirm that the most populous cities generally have the highest sales (Iowa

Cities by Population, 1). West Des Moines has comparatively high sales relative to its population (7th in population and 4th in sales).

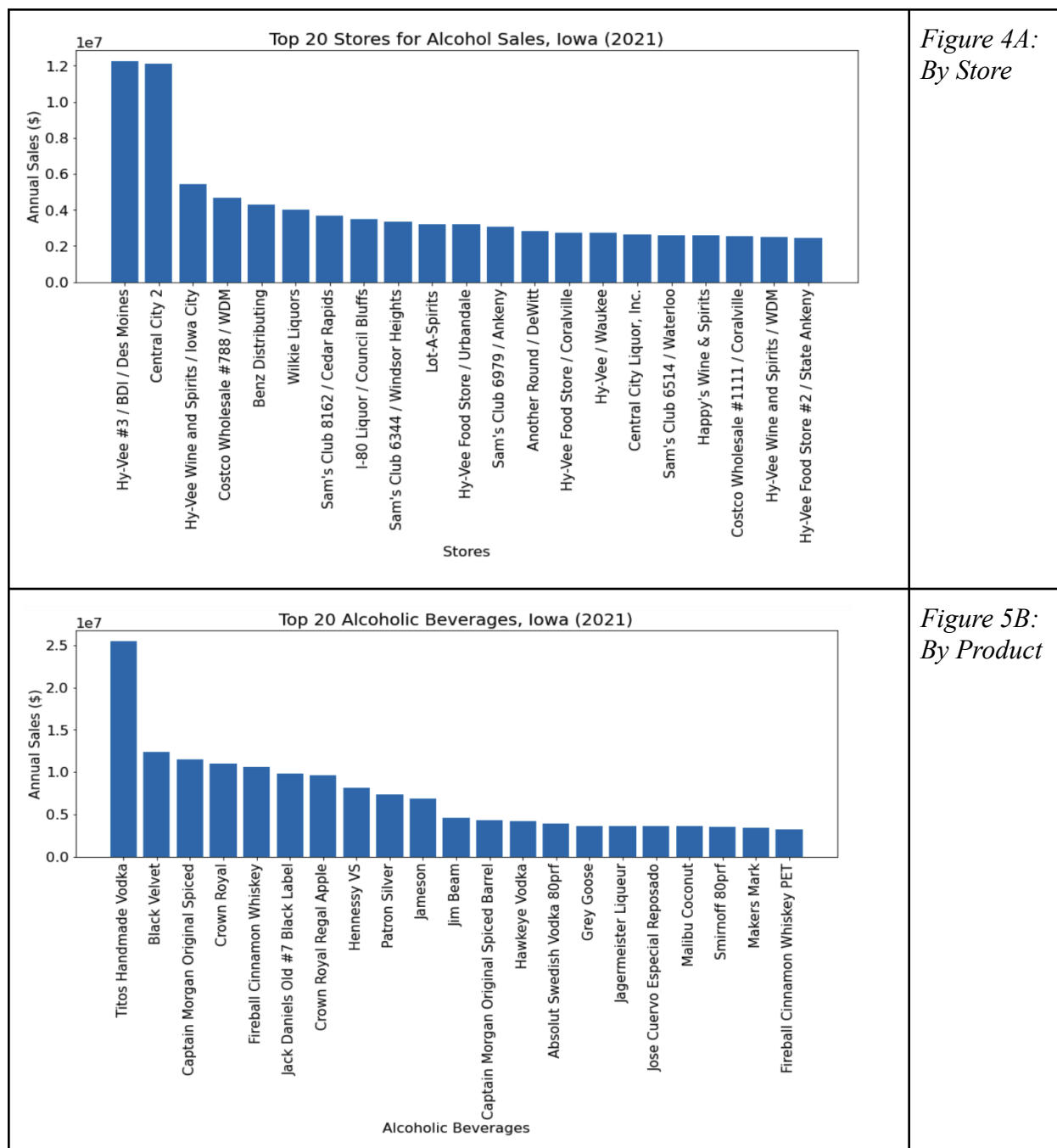
Figure 3: Geographic Trends within Iowa Liquor Sales



Store-Level Analysis of Liquor Sales

Figure 4A demonstrates that the majority of liquor sales in Iowa occur in grocery chains and wholesalers within Iowa’s largest cities, with Hy-Vee, Costco and Sam’s Club locations in Des Moines and West Des Moines leading the way. *Figure 4B* adds complexity to the trend seen in *Figure 2*, whereby Tito’s Homemade Vodka placed first, in dollarwise sales, amongst all vodka brands as well as all liquor brands (roughly \$25.5 million). From this, we can determine that Tito’s Homemade Vodka makes up the majority of annual vodka sales in Iowa. After Tito’s Vodka, Black Velvet, Captain Morgan, Crown Royal and other dark liquors were the next most popular choices, ranging between \$9-\$12 million in total annual sales.

Figures 4A and B: Top 20 Liquor Stores and Products Ranked By Annual Alcohol Sales



Product Recommendations

Figure 5 corroborates the trends outlined in Figure 4B, where the overwhelming favorites amongst liquor types and liquor products were Vodka, and specifically Tito's Handmade Vodka, respectively. Figure 5A outlines the most popular liquor type for each month based on monthly sales - American Vodkas ranked #1 11 out of 12 months of the year. Similarly, Figure 5B

highlights how Tito's Handmade Vodka ranked #1 all year-round with higher monthly sales seen especially in June, August and November.

Figure 5: Most Popular Liquor Categories and Products Ranked By Annual Alcohol Sales

Top Liquor Category Monthly Sales (USD)			Top Liquor Product Monthly Sales (USD)		
Month			Month		
January	American Vodkas	3918381.03	January	Titos Handmade Vodka	1362248.98
February	Canadian Whiskies	4369127.18	February	Titos Handmade Vodka	1651086.39
March	American Vodkas	4855906.92	March	Titos Handmade Vodka	1695243.29
April	American Vodkas	4787015.31	April	Titos Handmade Vodka	1959135.02
May	American Vodkas	4990454.34	May	Titos Handmade Vodka	2276565.64
June	American Vodkas	5957270.68	June	Titos Handmade Vodka	2745259.29
July	American Vodkas	5203116.24	July	Titos Handmade Vodka	2109023.14
August	American Vodkas	6491278.80	August	Titos Handmade Vodka	3384701.91
September	American Vodkas	5038699.86	September	Titos Handmade Vodka	1652227.39
October	American Vodkas	5088933.01	October	Titos Handmade Vodka	2283440.83
November	American Vodkas	5905344.07	November	Titos Handmade Vodka	2606016.70
December	American Vodkas	5019749.08	December	Titos Handmade Vodka	1757915.52
<i>Figure 5a: Most Popular Categories</i>			<i>Figure 5b: Most Popular Products</i>		

Insights:

What are the most popular liquor categories and products?

In 2021, Iowa alcohol retailers saw \$428,123,535.40 in alcohol sales with American Vodkas (14%) and Canadian Whiskies (11%) accounting for roughly 25% of sales. American Vodkas saw \$61,410,334.48 in sales with Tito's Handmade Vodka accounting for 41% of all vodka sales. Canadian Whiskies sales totalled \$47,392,655.67 with Black Velvet and Crown Royal products accounting for 29% and 47% of Canadian Whiskey sales respectively.

When should retailers stock the most popular items?

American Vodkas were the most popular liquor, in terms of monthly sales, 11 out of 12 months. Similarly, Tito's Handmade Vodka was the most popular liquor product, ranking #1 in sales every month of 2021. As such, we recommend that all Iowa alcohol retailers have American Vodkas and Canadian Whiskies stocked on their shelves all year round. While American Vodka sales spiked during the warmer months (May-August) and Canadian Whiskies during the fall and

winter, spikes in Agave Tequila sales coincided with many major U.S. holidays throughout the year. As such, we also recommend that Agave Tequila brands be consistently stocked.

Which retailer locations are selling the most?

Unsurprisingly, alcohol retailers in Iowa's major cities (Des Moines, Cedar Rapids, Davenport, West Des Moines) drove the most liquor sales, accounting for 27% of annual sales. Separately, 13 of the top 20 retailers (in terms of annual liquor sales) were either big-box retailers (Sam's Club, Costco Wholesale) or regional, supermarket chains (i.e. Hy-Vee).

Appendix:

Table 1: Data Dictionary with Shorthand Descriptions:

Invoice/Item Number -Unique Key for Logging Ordered Transactions	Date - Timestamp for transaction date	Store Number -Unique Key for Store Names	Store Name - Name of Store Who Ordered Liquor	Address - Street Addresses of Stores	City - Cities in Iowa
Zip Code - Zip Code of Purchasing Store	Store Location - Latitude and Longitude of Store Coordinates	County Number - Unique Key for County Names	County - County Name of Purchasing Store	Item Number - Unique Key for Item Product Ordered.	Item Description - Item Name for Product Ordered.
Pack - Number of Bottles Associated with Ordered Product	Bottle Volume (ml) - Volume of Bottle Associated with Ordered Product	State Bottle Cost - Amount Alcoholic Beverages Division Paid for Each Bottle Ordered	State Bottle Retail - Amount Stores Paid for Each Bottle Ordered	Bottles Sold - Number of Bottles Sold in Transaction	Sale (Dollars) - Total USD Sales Amount for Ordered Product
Volume Sold (Liters) - Total Volume of Alcohol Sold Per Order, in Liters	Volume Sold (Gallons) - Total Volume of Alcohol Sold Per Order, in Gallons	Category - Unique Key for Category Names	Category Name -Name of Broader Category a Given Product Belongs To.	Vendor Number - Unique Key for Vendor Names	Vendor Name - Name of Vendor Supplying Alcohol to Stores

Table 2: Data Types and Counts: Includes Color-Coded Remapping Plans

dtype (count)	object (9)	Float64 (8) *	Int64 (6)	Datetime64[ns] (1)
Variables *reformat floats to 2 decimal places	"Invoice/Item Number" "Store Name" "Address" "City" "Store Location" "County" "Category Name" "Vendor Name" "Item Description"	"Zip Code" "County Number" "Vendor Number" "State Bottle Cost" "State Bottle Retail" "Sale (Dollars)" "Volume Sold (Liters)" "Volume Sold (Gallons)"	"Store Number" "Category" "Item Number" "Pack" "Bottle Volume (ml)" "Bottles Sold"	"Date"

Table 3: Preliminary Assessment Chart:

Hypothesis	Procedure	Findings and Conclusion
1. All sales invoices were unique and unrepeatd in the dataset.	<ul style="list-style-type: none"> Select "Invoice/Item Number" Calculate value counts, sort by descending order. 	<ul style="list-style-type: none"> No value counts > 1. Data requires no reformatting, no duplicate invoices logged.
2. All item numbers would uniquely map to respective item descriptions.	<ul style="list-style-type: none"> Select "Item Description", "Item Number", "Pack" and "Bottle Volume (ml)." 	<ul style="list-style-type: none"> 7.6% of "Item Number" values have multiple "Item Descriptions" "Item Description" changes over time. Note that "Item Number" corresponds to unique combinations of "Item

	<ul style="list-style-type: none"> • Drop duplicates. • Group By “Item Description” and “Item Number.” 	Description,” “Pack” and “Bottle Volume (ml)” for all downstream graphing.
3. All store numbers would uniquely map to respective store names.	<ul style="list-style-type: none"> • Select “Store Name” and “Store Number.” • Drop duplicates. • Group By “Store Number” 	<ul style="list-style-type: none"> • 98.9% of “Store Number” values map to a unique “Store Name” value. • Remaining 1.1% based on improper assignment of “Store Name” values to correct “Store Number.” • Data reformatting is discouraged due to miniscule proportion of improper assignments. This fractionally constitutes only 52585/2.6M transactions, so reformatting will be ignored.
4. All category ids would uniquely map to respective category names.	<ul style="list-style-type: none"> • Select “Category” and “Category Name.” • Drop duplicates. • Group By “Category.” 	<ul style="list-style-type: none"> • All “Category” numerical values correspond to a unique “Category Name.” • Data requires no reformatting.
5. All vendor numbers would uniquely map to respective vendor names.	<ul style="list-style-type: none"> • Select “Vendor Name” and “Vendor Number.” • Drop duplicates. • Group By “Vendor Number.” 	<ul style="list-style-type: none"> • 97.3% of “Vendor Number” maps to a unique “Vendor Name”. • Remaining 2.7% of “Vendor Names” are coding errors. Ex. “BAD BEAR ENTERPRISES LLC” and “BAD BEAR ENTERPRISES LLC / Legendary Rye” • Data requires no reformatting, percentage of vendor name deviants insignificant relative to total population.
6. All county numbers would uniquely map to respective county names.	<ul style="list-style-type: none"> • Select “County” and “County Number.” • Drop duplicates. • Group By “County Number.” 	<ul style="list-style-type: none"> • 79.8% of “County Number” maps to a unique “County”. • 20.2% of “County” are case issues. Ex. “ADAIR” and “Adair” • Convert “County” to lowercase
7. “Volume Sold (Liters)” should equal the arithmetic product of the “Bottles Sold” and “Bottle Volume (ml),” scaled to liters. Intuitively, “Bottles Sold” should also be equal to “Pack” for a given transaction.	<ul style="list-style-type: none"> • Calculate volume per order as “Bottles Sold” x “Bottle Volume (ml)”/1000 • Compare to “Volume Sold (Liters)” • Compare “Bottles Sold” with “Pack” 	<ul style="list-style-type: none"> • “Volume Sold (Liters)” was correctly calculated by “Bottles Sold” x “Bottle Volume (ml)”/1000. • “Pack” provides description for where bottles are <i>coming</i> from. “Pack” provides no information on how much liquor was sold in the transaction. • Therefore, use “Volume Sold (Liters)” when performing sales calculations. Drop “Pack” due to irrelevance in research questions.

Table 4: Example of Store Misattributions

Store Number	Associated Store Names
6245	“Ankeney Wind And Spirits / Ankeny” <i>and</i> “Neighborhood Liquor House / Ankeny”
5107	“Kum & GO #206 / Clive” <i>and</i> “Kum & Go #4098 / Windsor Heights.”

Table 5: Dropped Variables Chart

Variable(s)	Justification for Variable Removal
Store Location (Tuple) Address	Too Specific. “Zip” and “County Name” Preferred.
County Number Vendor Number Store Number Item Number Category	Duplicative References to Corresponding Variable Names. Only Used for Data Fidelity.
Volume Sold (Gallons)	Duplicative. Liters Preferred.
Pack	Provides No Information on Sales
Invoice/Item Number	No Application To Graphing. Only Used for Data Fidelity.
State Bottle Cost/State Bottle Retail	Irrelevant for This Analysis. Consider for Future EDAs.

Citations:

1. *Chapter 17 Class “E” Liquor Control Licenses - Iowa.*
[https://www.legis.iowa.gov/docs/ACO/GNAC/iacpdf\(8-8-01\)/iac/185iac/18517/18517.pdf](https://www.legis.iowa.gov/docs/ACO/GNAC/iacpdf(8-8-01)/iac/185iac/18517/18517.pdf)
2. Iowa Department of Commerce, Alcoholic Beverages Division. *Iowa Liquor Sales*, 1 Apr. 2022, <https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales/m3tr-qhgy>.
3. Link to 2021 data:
<https://drive.google.com/file/d/1FmhlMA6rGmKcSLUefp-7G2l2Q14OW5hj/view?usp=sharing>
4. “Iowa Cities by Population.” *Iowa Outline*,
https://www.iowa-demographics.com/cities_by_population.