Alcohol Consumption in Utah

Project Process Book

Azita Zeyghami, u1240546@utah.edu, u1240546
Rey Johnson, rey.johnson@utah.edu, u01178192
Zach Shepelak, zachary.shepelak@utah.edu, u1121796

GitHub Repository: https://github.com/mtnbkazco/ACU

Overview and Motivation (Same as Proposal):
Utah has a unique population when it comes to the consumption of alcohol, this is largely because of the Utah-based faith, The Church of Jesus Christ Latter-day Saints. In 2010, the population of Utah was 2.7 million. According to the estimates released by the Census Bureau in January 2020, Utah's population has increased 16% over the last decade. According to the census, net migration had a bigger impact on population growth than high fertility and reproduction rates. Simply put, more people not from Utah are moving to Utah. In fact, people moving into the state outpaced new births in 2019 (Richards 2020). This unremitting demographic shift has led to a major change in Utah. As of 2018 48.91% of the Salt Lake County population (less than half) was Latter-day Saint. That is the lowest since the 1930s (Canham 2018). This means that slowly overtime the change in demographics has also caused a change in lifestyle. Although we are unable to look back ninety years at alcohol consumption, we can look back over the last couple years and see trends in alcohol consumption in Salt Lake County.

In addition to these gradual trends of alcohol consumption due to the evolving demographic profile of Utah, the possibility of more abrupt shifts in alcohol consumption is looming with the emergence of COVID-19.

With regard to COVID-19, it is clear that this virus has had a profound impact on our daily lives. In the last year we, as a society, face new challenges concerning mental health and dealing with anxiety and stress. In April it was reported that the pandemic had given way to an increase in domestic violence in countries abroad (Taub 2020). That same month it was also reported that there was an increase in alcohol sales abroad as well (Troianovski 2020). As COVID-19 continues to play a role in structuring our everyday lives, it's important to gauge the impact it is having closer to home, and within our communities. This led to our second motivation with this project. We want to visualize how alcohol trends have shifted in Utah during COVID-19 in Utah.

PEER REVIEW:

Jairon Terrero	Grant Keller

- Circle Squiggle Chart:
    - How many years are being encoded?
        - 3
    - Maybe just do all in aggregate or filter based on year choice
        - We agree this is a good choice
    - Or just do three lines instead of filter, or 2019/2020 average
- 2nd graph (Stacked Bar Chart):
    - Do a stacked area chart instead
        - We agree this is a good idea too
- Gapminder financial chart:
    - Put vertically adjacent to stacked area chart
    - Maybe have dot plot filter stacked bar chart based on month selection
- Organizationally put it all together of same page
    - We think this could work but also maybe tabs for different insights.. We also considered maybe even doing a article formatted storytelling visualization
- Any forms of interaction with stuff?
    - We are going heavy on tool tips
    - Filtering by year or category
- County sales location and include a geo map?
    - After further digging we couldn't find any data
- Thinks questions and project goals are good
- Breakdown by kind, what about specific brands?
    - Maybe after question for after data aggregation
    - Maybe add to tooltip or click event to show top 10 brands
        - When clicking on area chart or stacked bar chart
- Any other ideas that lead us to this?
    - We thought about doing analysis of domestic violence during covid, etc. and this was the best dataset other than putting in a lot of work into the SLC police reports
    - We have overall interest in a "Non-covid covid report"
    - Maybe geospatial trail data but too many options there, and our alcohol viz is unique

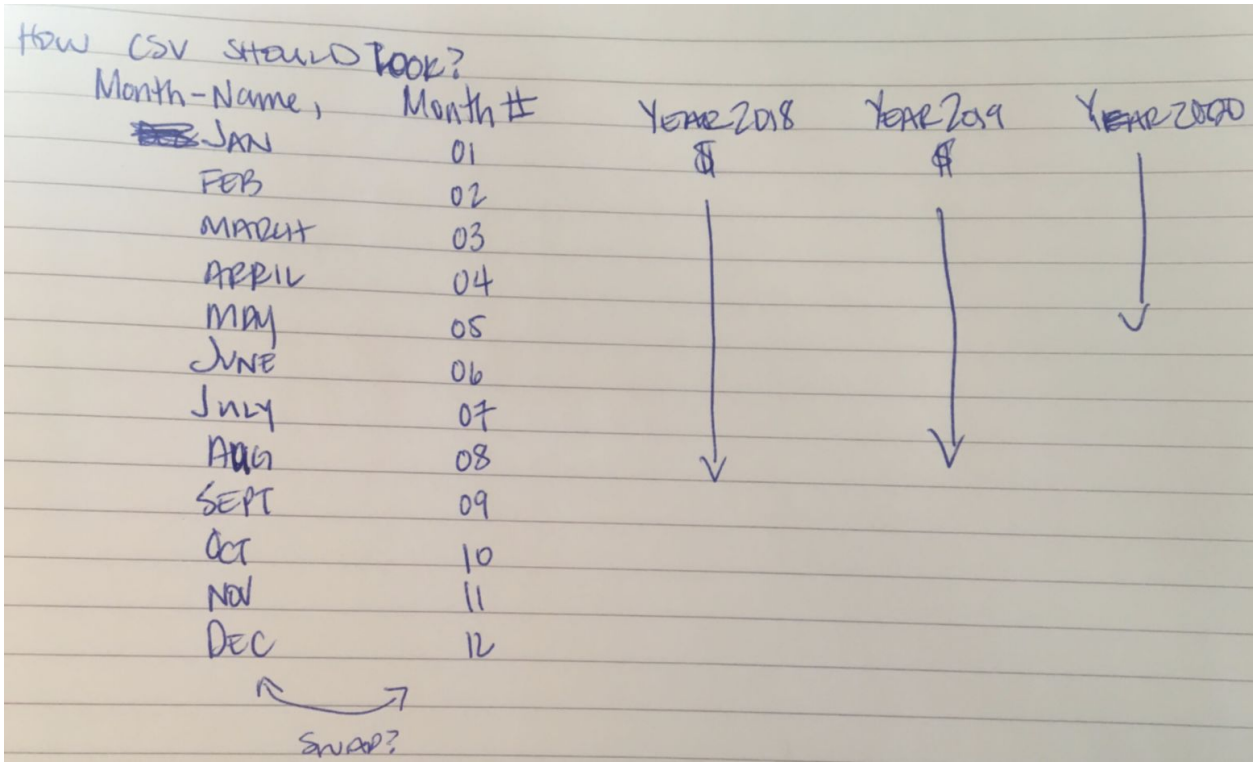
Data Exploration, Aggregation Explanation and Issues:

As explained in our proposal the data is available for all of the months going back to April of 2018, and is in a fairly similar format. There was still significant manipulation required, however, we eventually figured out what to do. The notebook that covers this process is DataVIZ_ACU_Data_Processing_and_Cleaning.ipynb, and is available in our GitHub repository.

Regarding data accuracy, November 2019 has no class codes and there is no data for December 2019 for some reason, but other than that our data is accurate. We were thinking of doing string matching for November 2019 to match the missing class codes to codes from previous months based on Item ID, which in theory should be the same. However, there were some issues we ran into with this so we didn't do it. Specifically, from the master data there were only ~370 unique Item ID's that had multiple unique Class codes so there was not enough data to match up to the ~6000 rows of Data for November 2019.

On our git repository, the file ACU_DATA_FINAL.xlsx is our final data that we will be using.

Radial chart data wrangling:



Created month dictionary for data:

```
data = pd.concat(data_build, axis=1, keys=headers)
columns = ["MonthDate", "SalesTotal"]
monthSales = pd.DataFrame(columns=columns)

month_dict = [{"month_str": "April", "month_dt": "2018-04-01"},
              {"month_str": "May", "month_dt": "2018-05-01"},
              {"month_str": "June", "month_dt": "2018-06-01"},
              {"month_str": "July", "month_dt": "2018-07-01"},
              {"month_str": "Aug", "month_dt": "2018-08-01"},
              {"month_str": "Sept", "month_dt": "2018-09-01"},
              {"month_str": "Oct", "month_dt": "2018-10-01"},
              {"month_str": "Nov", "month_dt": "2018-11-01"},
              {"month_str": "Dec", "month_dt": "2018-12-01"},
              {"month_str": "Jan", "month_dt": "2019-01-01"},
              {"month_str": "Feb", "month_dt": "2019-02-01"},
              {"month_str": "March", "month_dt": "2019-03-01"},
              {"month_str": "April", "month_dt": "2019-04-01"},
              {"month_str": "May", "month_dt": "2019-05-01"},
              {"month_str": "June", "month_dt": "2019-06-01"},
              {"month_str": "July", "month_dt": "2019-07-01"},
              {"month_str": "Aug", "month_dt": "2019-08-01"},
              {"month_str": "Sept", "month_dt": "2019-09-01"},
              {"month_str": "Oct", "month_dt": "2019-10-01"},
              {"month_str": "Nov", "month_dt": "2019-11-01"},
              {"month_str": "Dec", "month_dt": "2019-12-01"},
              {"month_str": "Jan", "month_dt": "2020-01-01"},
              {"month_str": "Feb", "month_dt": "2020-02-01"},
              {"month_str": "March", "month_dt": "2020-03-01"},
              {"month_str": "April", "month_dt": "2020-04-01"},
              {"month_str": "May", "month_dt": "2020-05-01"},
              {"month_str": "June", "month_dt": "2020-06-01"},
              {"month_str": "July", "month_dt": "2020-07-01"},
              {"month_str": "Aug", "month_dt": "2020-08-01"},
              {"month_str": "Sept", "month_dt": "2020-09-01"}]
```

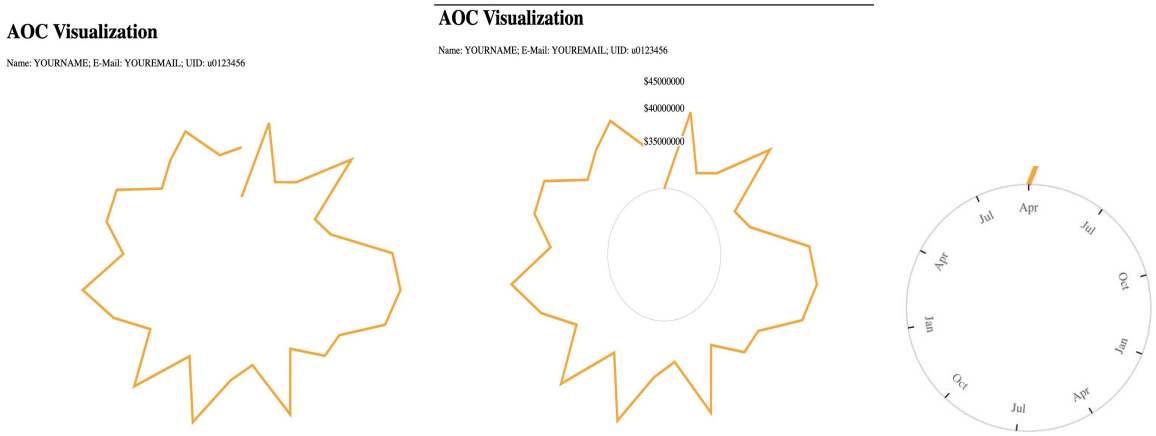
Iterated through the data and fill in blanks where needed.

Design Evolution

First Try at “Circle Squiggle Chart”

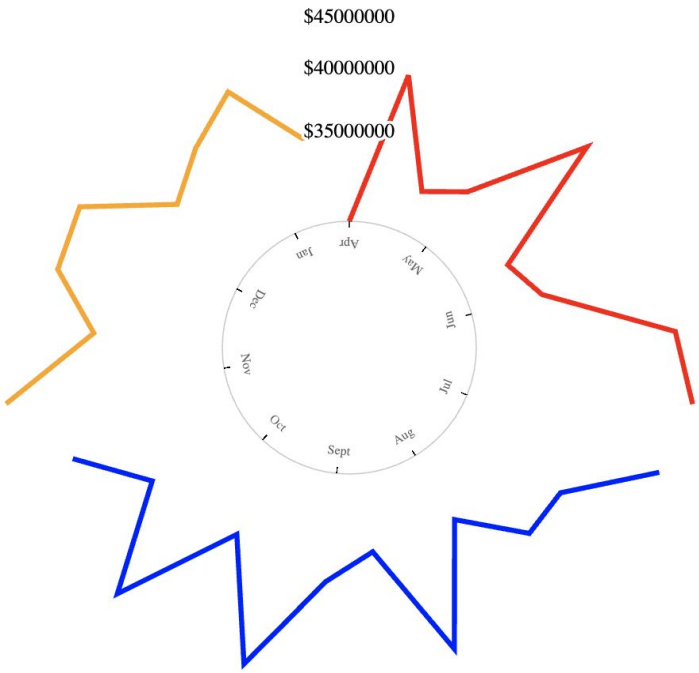
Data set creation. Parsed the whole dataset by month and year and calculated the total sales for each month in the year. The csv creation holds month(date), total sales for that month, and month(string)

Need to radius lines and multiple lines per year still...

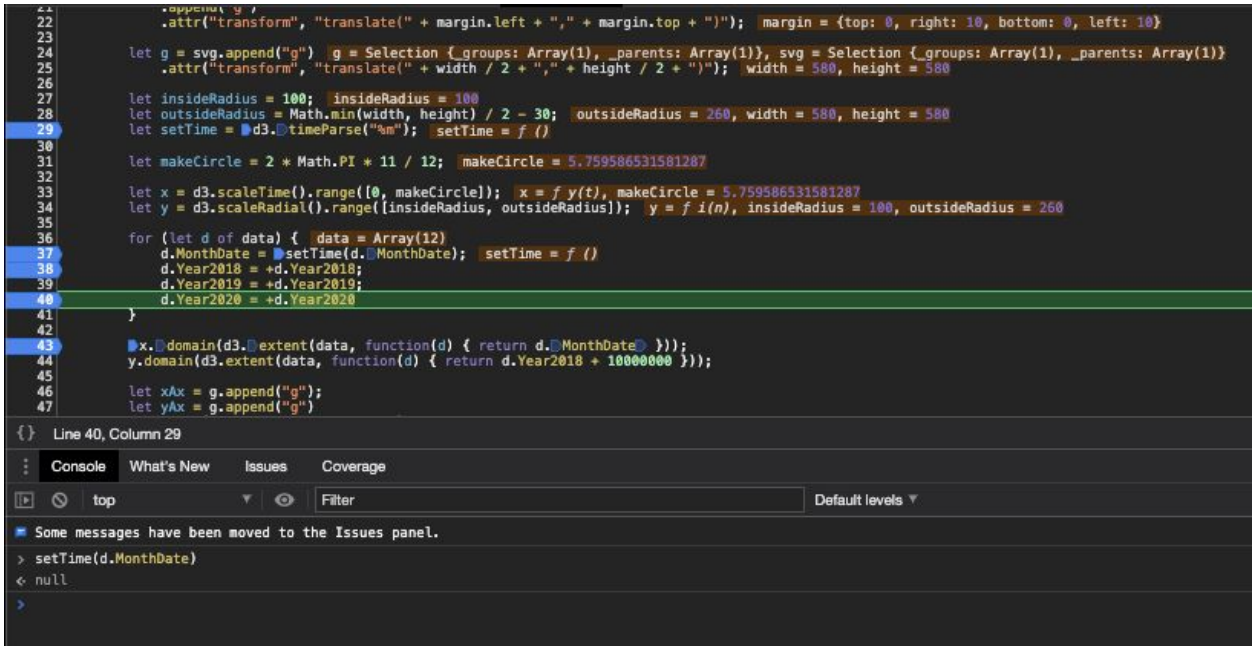


Got labels to show: (path is reading ALL data, needs to be separated by year (2018,2019,2020)

Three lines: but they will not plot in the same area. Each line seems to have its own area and they should not, they should plot together in the same area and at the same time.

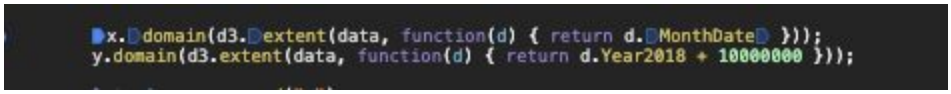


Possible time issue with the line related to time?

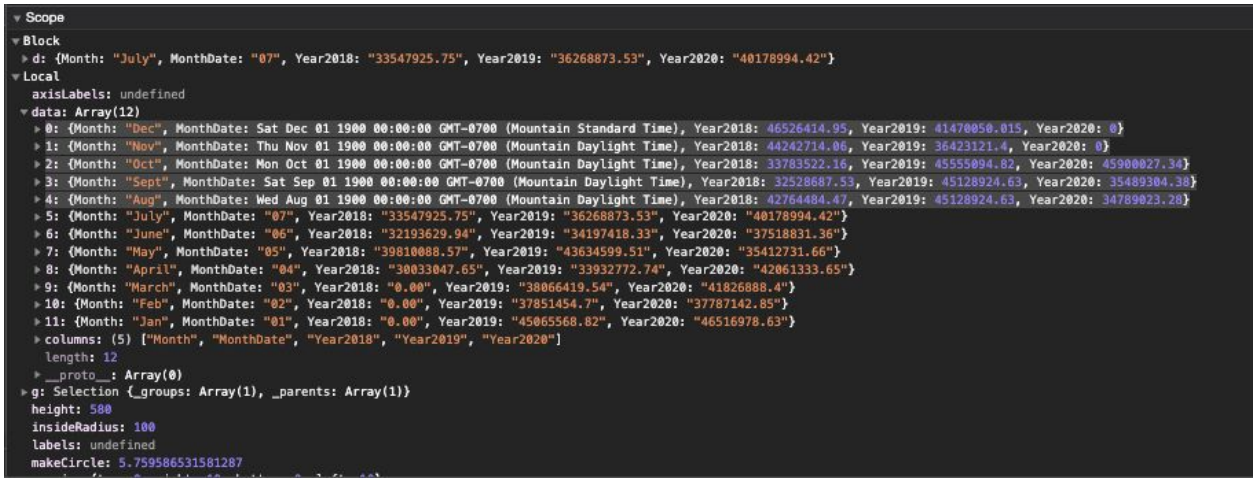


Returning null…

Fixing with x.domain:

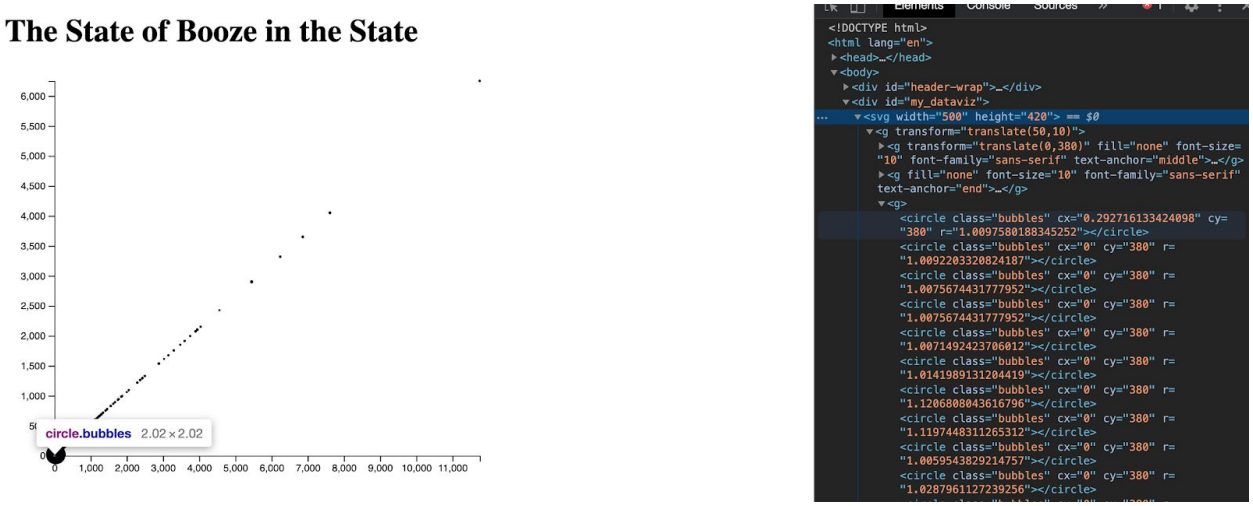


Return:

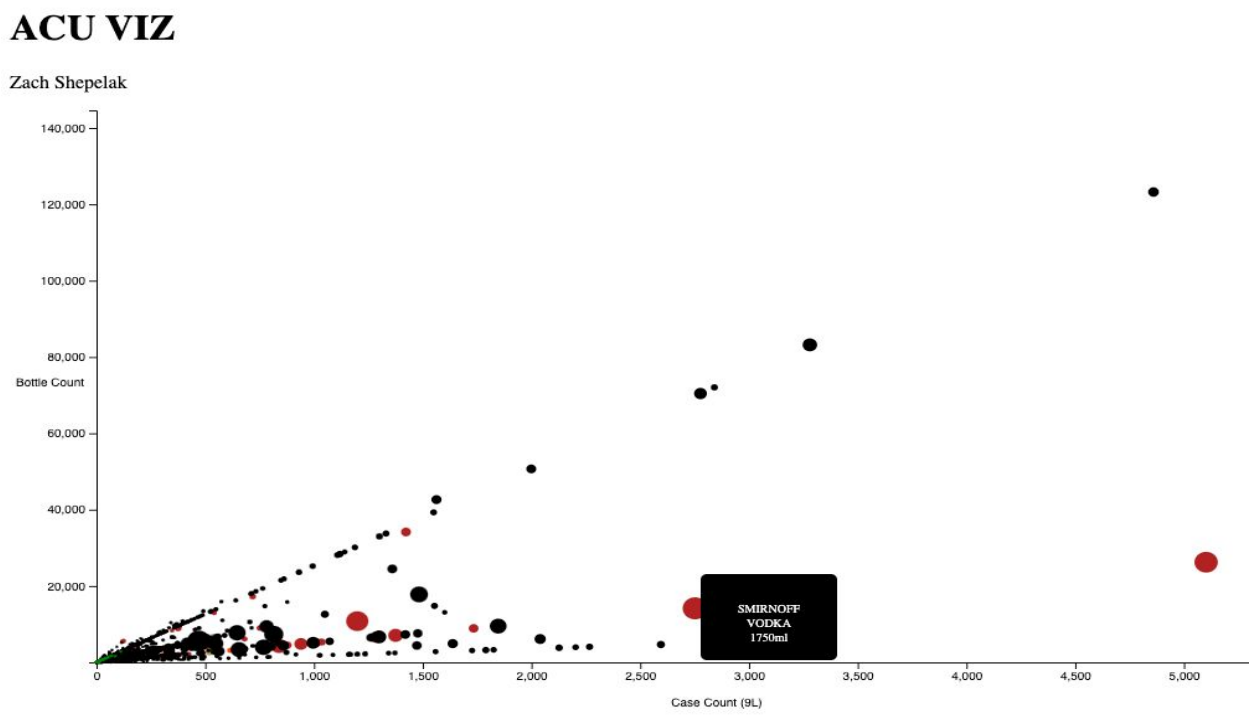


First Try at Bubble Chart:

This is using all of the ~170,000 rows as a bubble which is incredibly slow. Even if I separate by year, it will still be too slow to use. The only option would be aggregation by class type, but after looking at the chart we can see almost everything has a similar gross margin so I don't know how interesting this visualization will be. I may keep the same idea of chart design but change the variables slightly to add a more interesting and granular view.



Replacement Option for Financial Margin View



Rather than using the Cost and Price as axes to try and visualize margin, this approach visualizes Case Count versus Bottle Count, and encodes the Sales figures for that month by the size of the circle. This allows the viewer to view what size of alcohol is sold most, along with

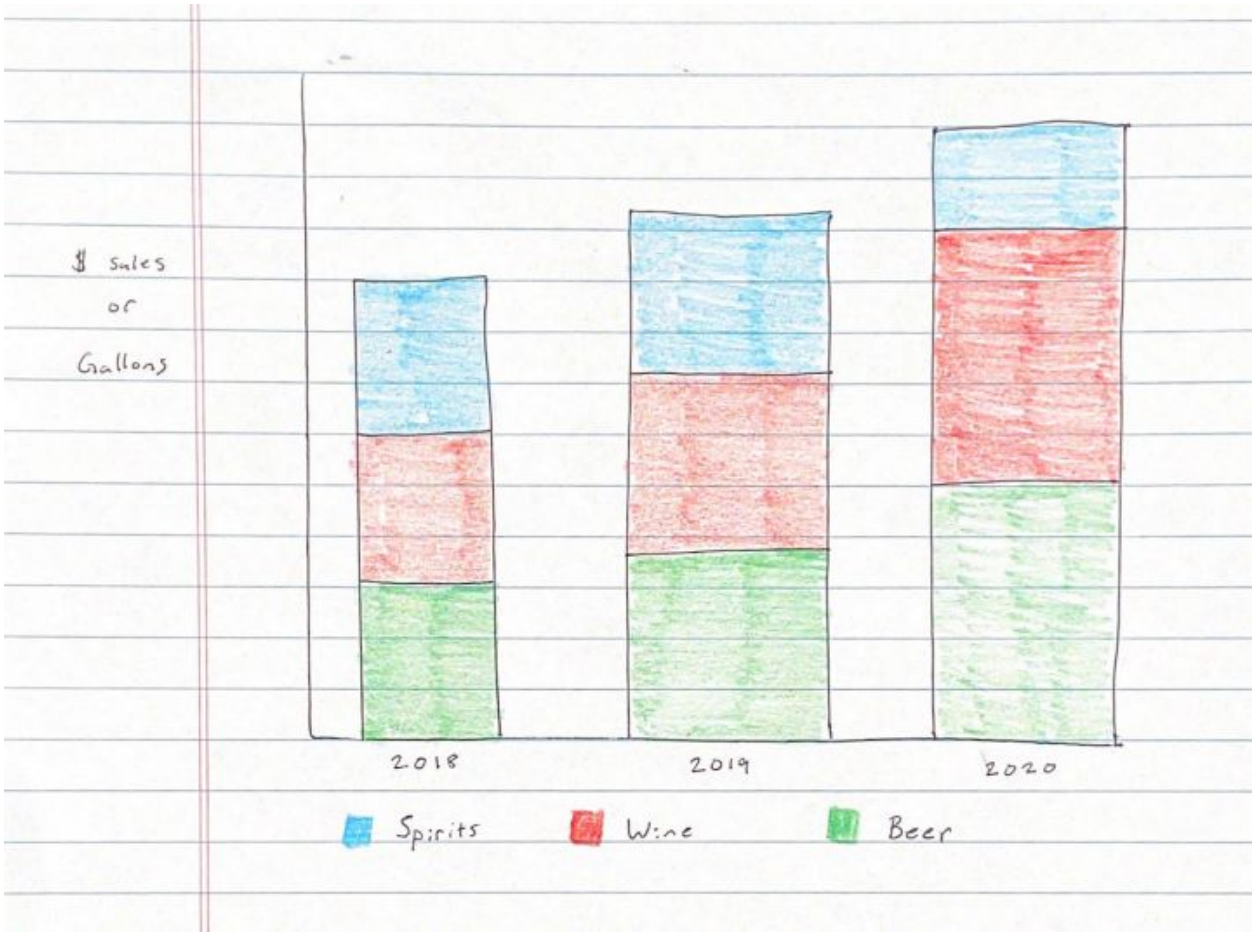
seeing which specific products are the most popular via a tooltip. Additionally, type of alcohol will be encoded by color, however, I am not yet sure if I want to do all the different subtypes that the website uses, or simply encode as “wine”, “beer”, or “spirits”.

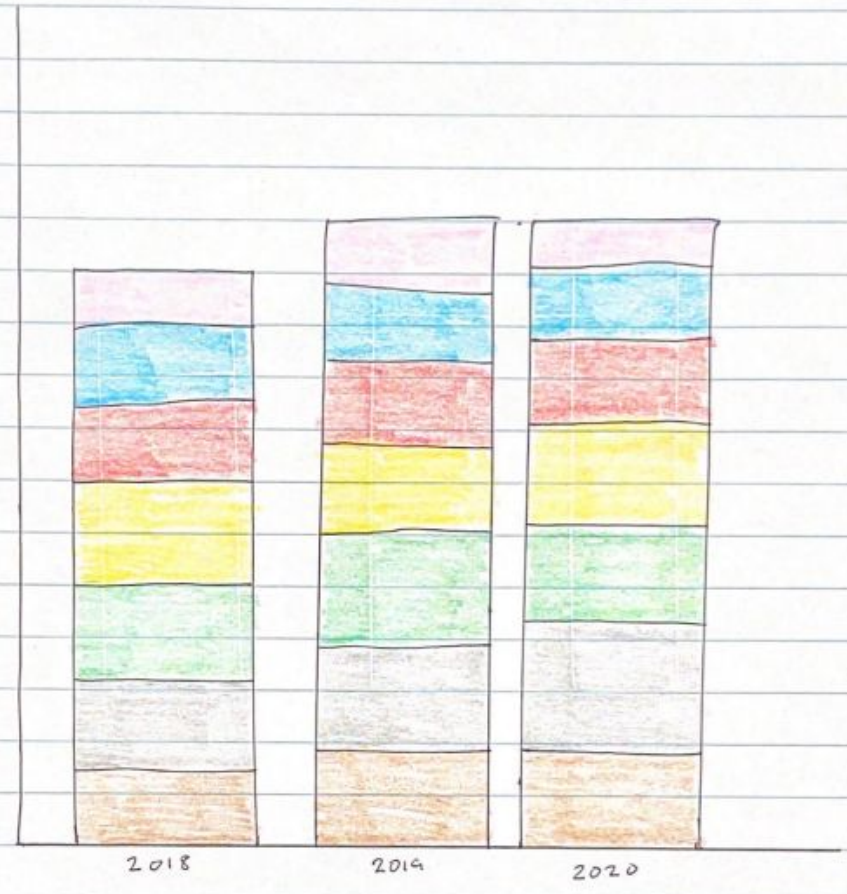
Additionally, to get around the time issue with loading ~170000 dots, I am planning on adding a slider to the top that will allow the viewer to select the different months, and filter the data accordingly. The picture shown above is of just one month’s data

I will also potentially add a search bar so the viewer can search for a specific product of interest.

Visualization of Sales by Category

Concept Drawings




$
2018
2019
2020
Vodka
Rum
Gin
Tequila
Mezcal
Brandy
Whiskey

Vodka

		0	50%	100%
Basic	40%			
Flavored	30%			
Domestic	20%			
Imported	10%			

Rum

Gin

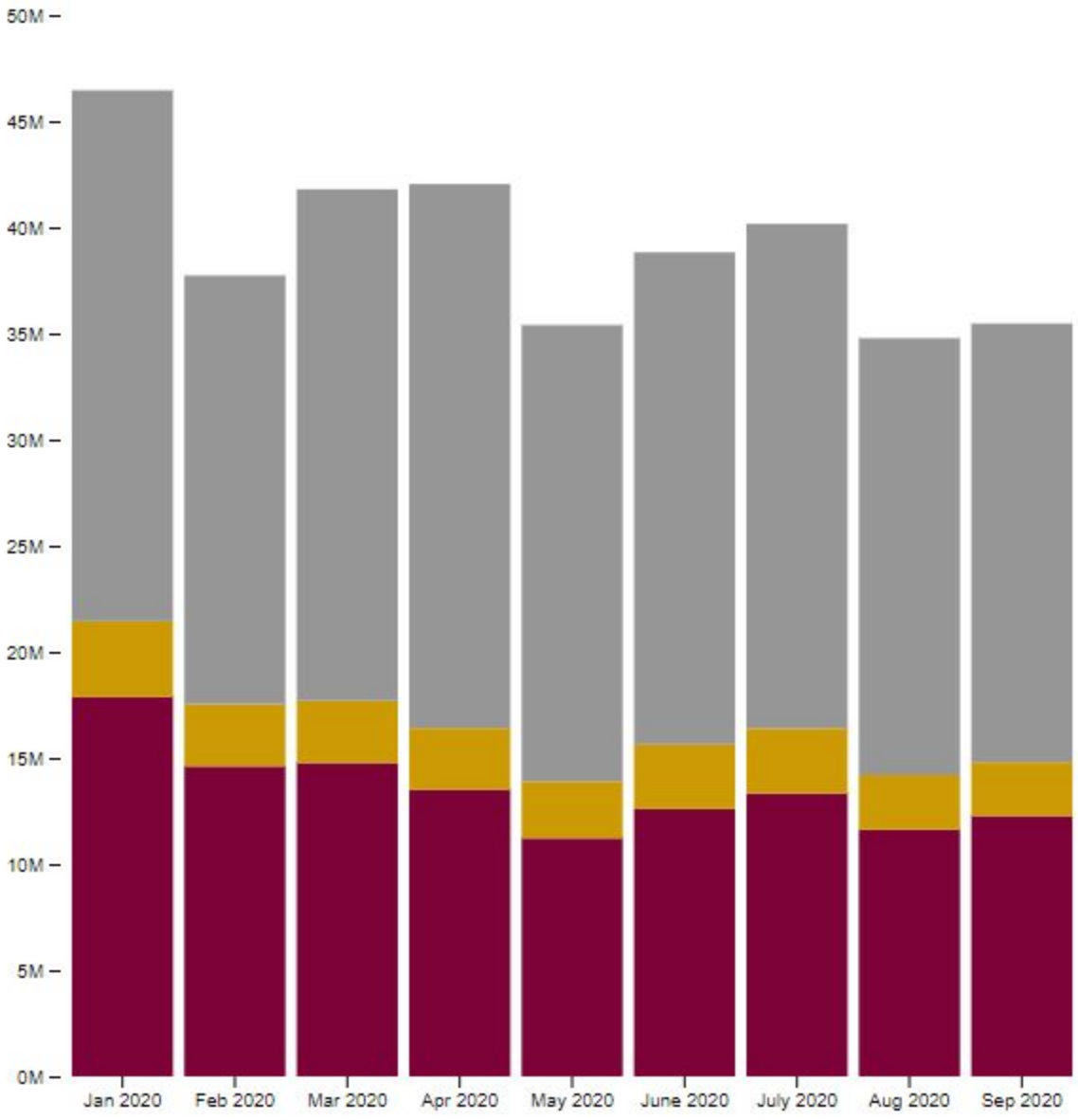
Tequila

Mezcal

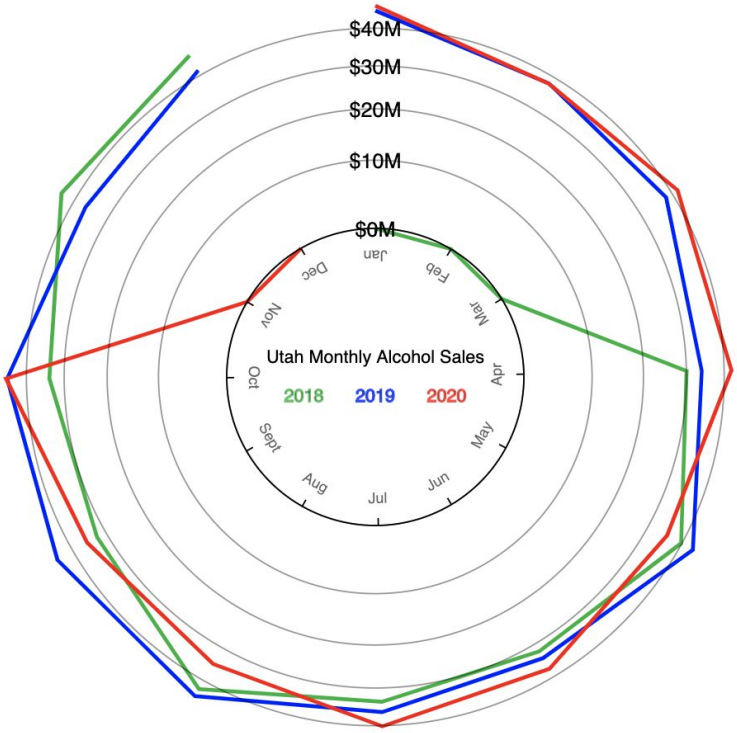
Brandy

Whiskey

Rough Draft in D3 of Categories Visualization

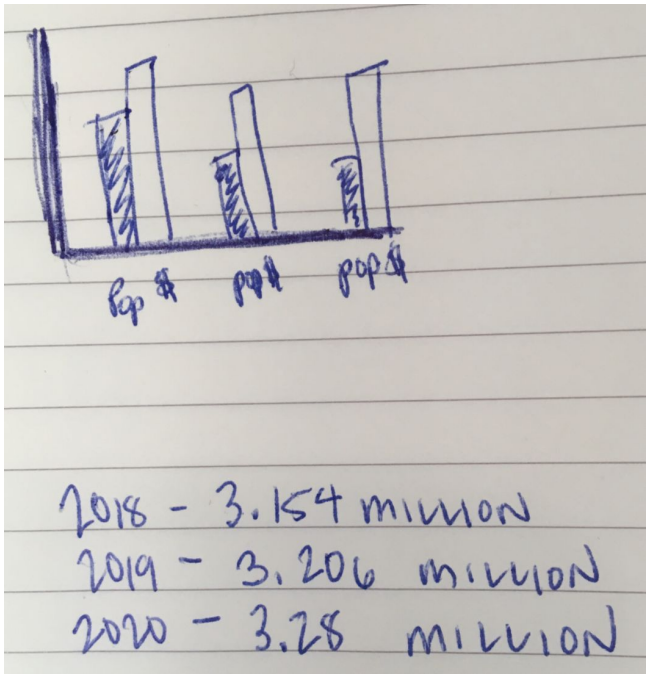


Update on "Circle Squiggle Chart"

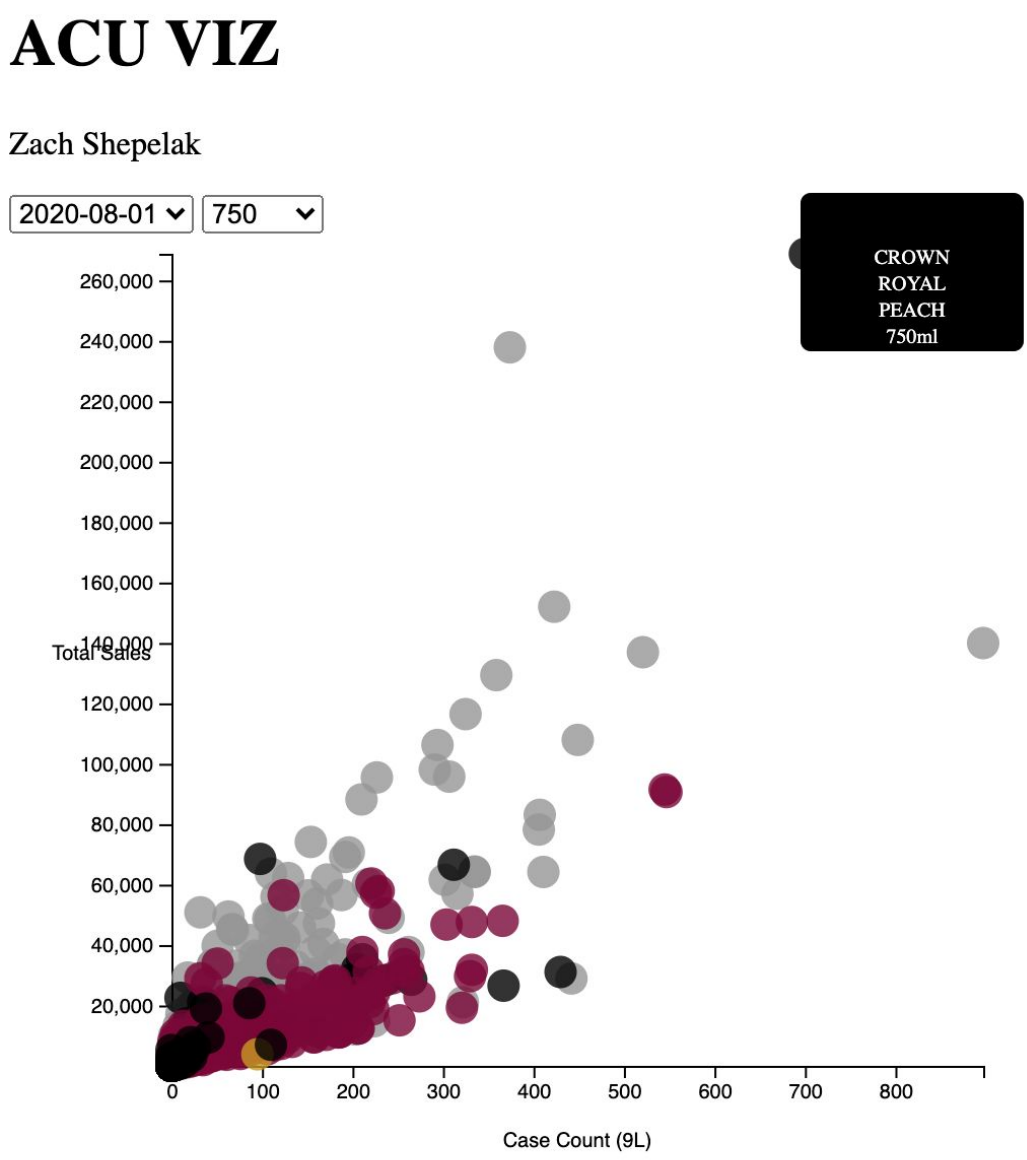


Added legend. I am not sure about the colors for each line. Still trying to sort out a hover option to make the chart interactive. Thoughts on full display in the end?...

Idea for population/bar chart: no-go

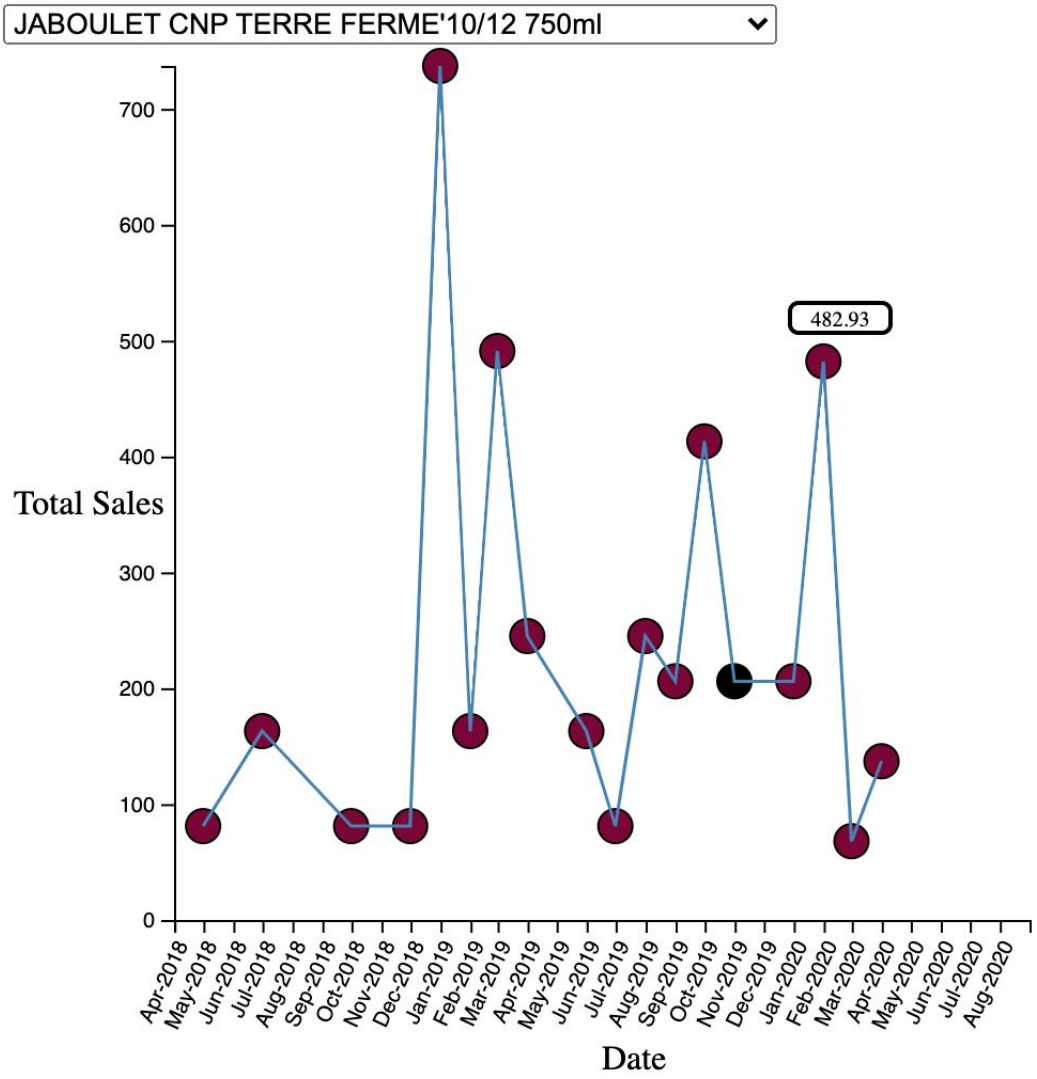


UPDATE ON Bubble Chart:



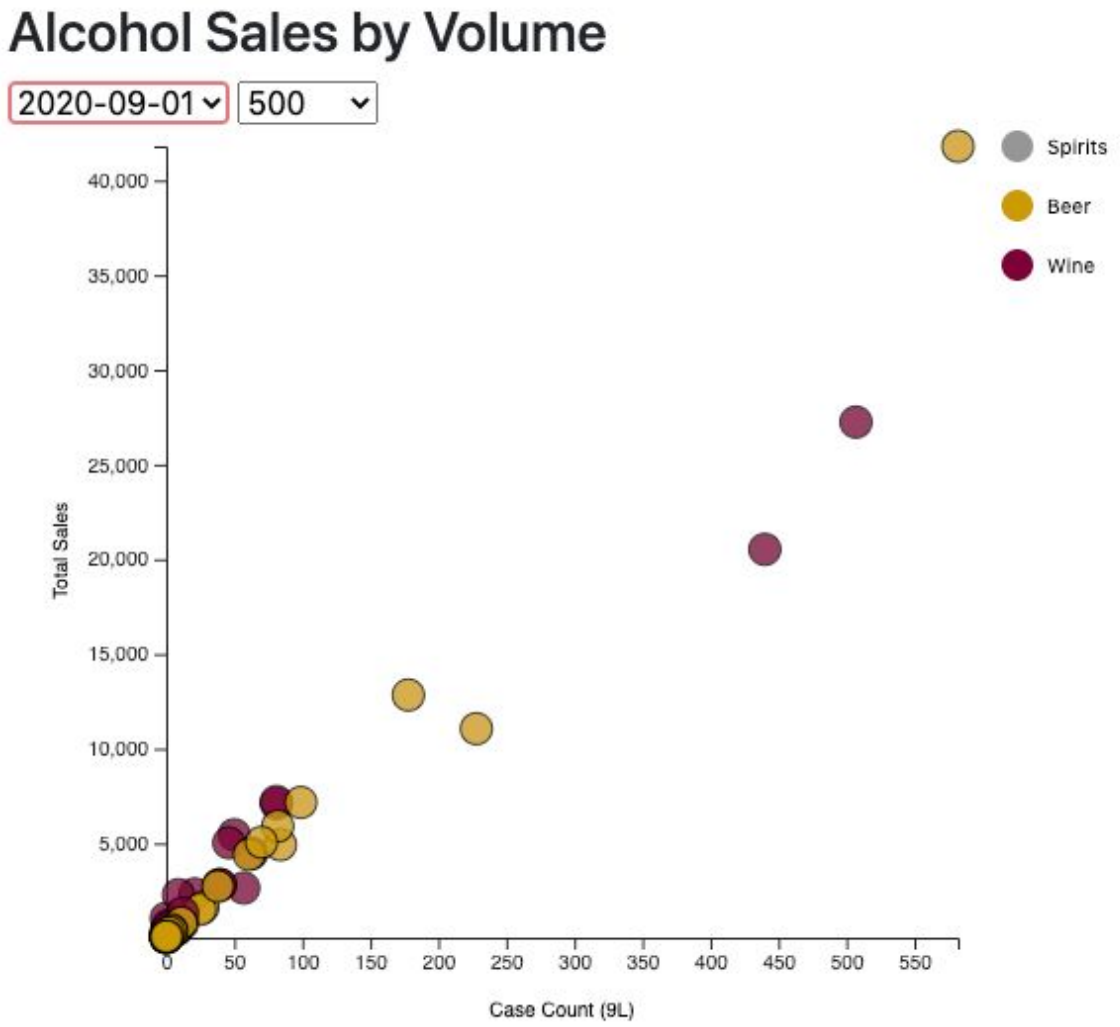
The multiple dropdowns have been figured out, and now filter the data accordingly. Axes also change dynamically. Color matches other category plot to make for a consistent color scheme. Next steps are really just polishing up the visualization

Product Search Plot Update



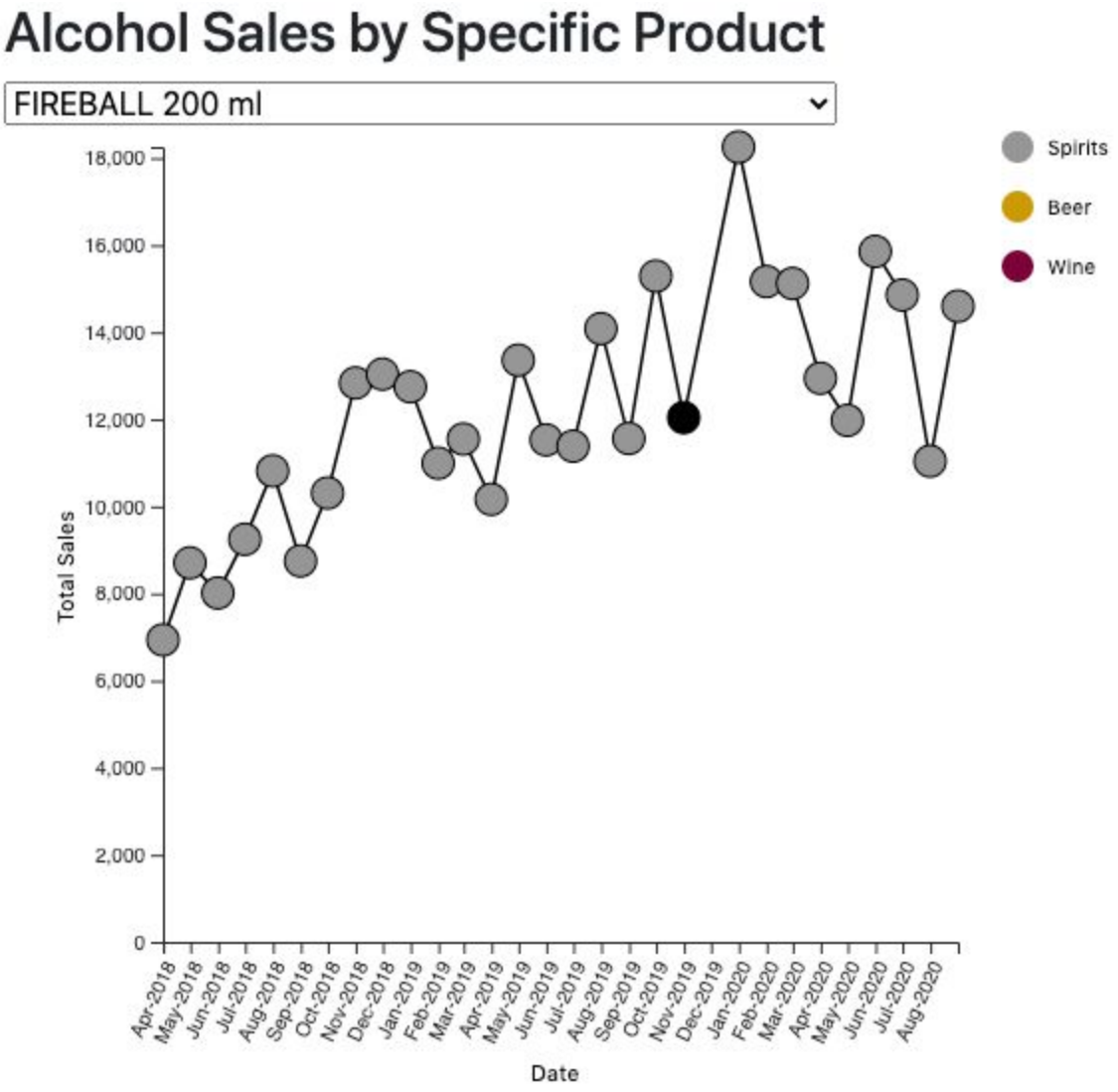
Another visual has been added that allows the user to see sales totals over time for specific products. The dropdown has a lot of options, but it is easy to search for specific products of interest so the functionality should be fine.

Bubble Chart Update:



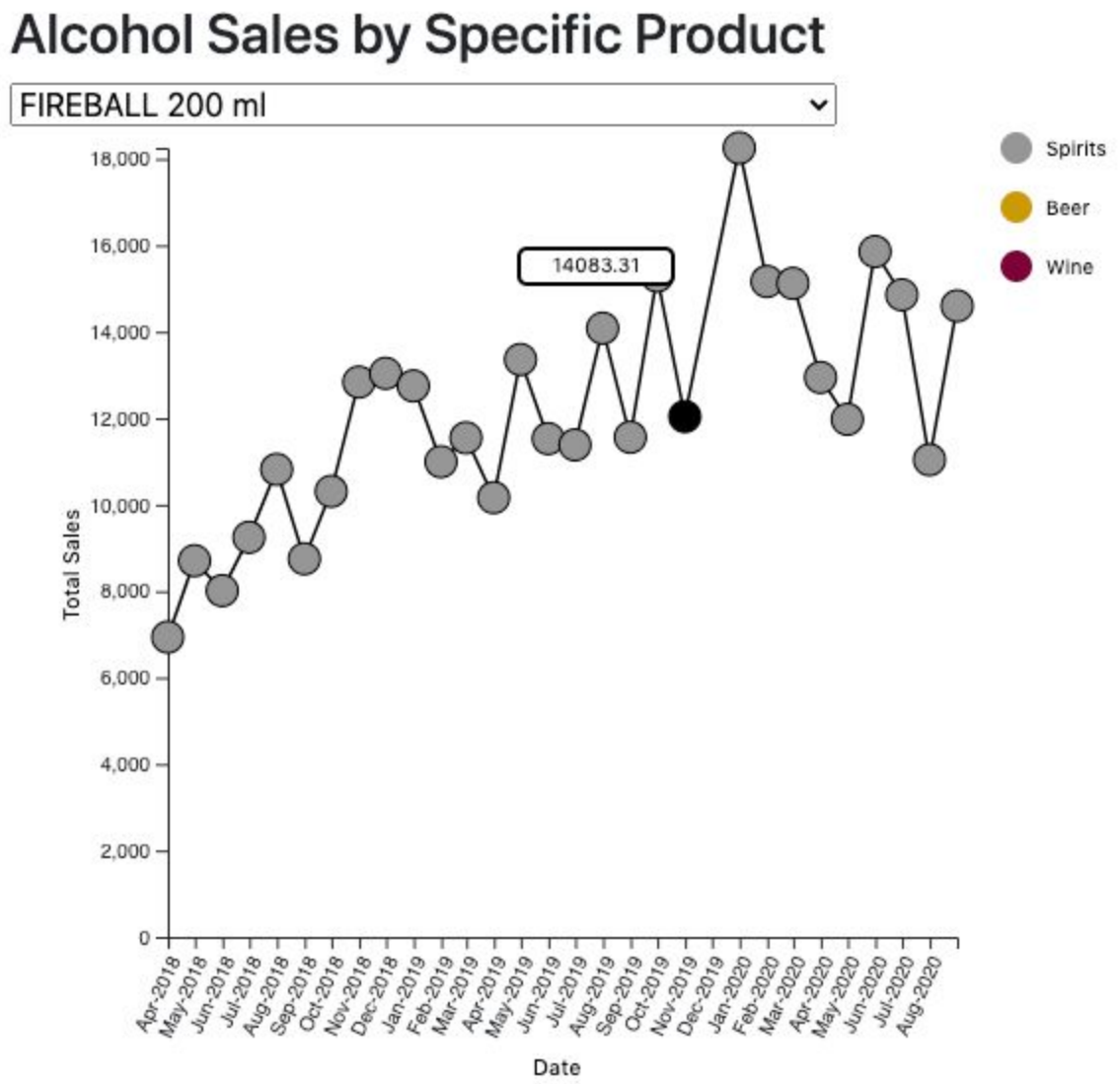
Now with a legend, and all of the circles are colored properly

Update on Search Plot:



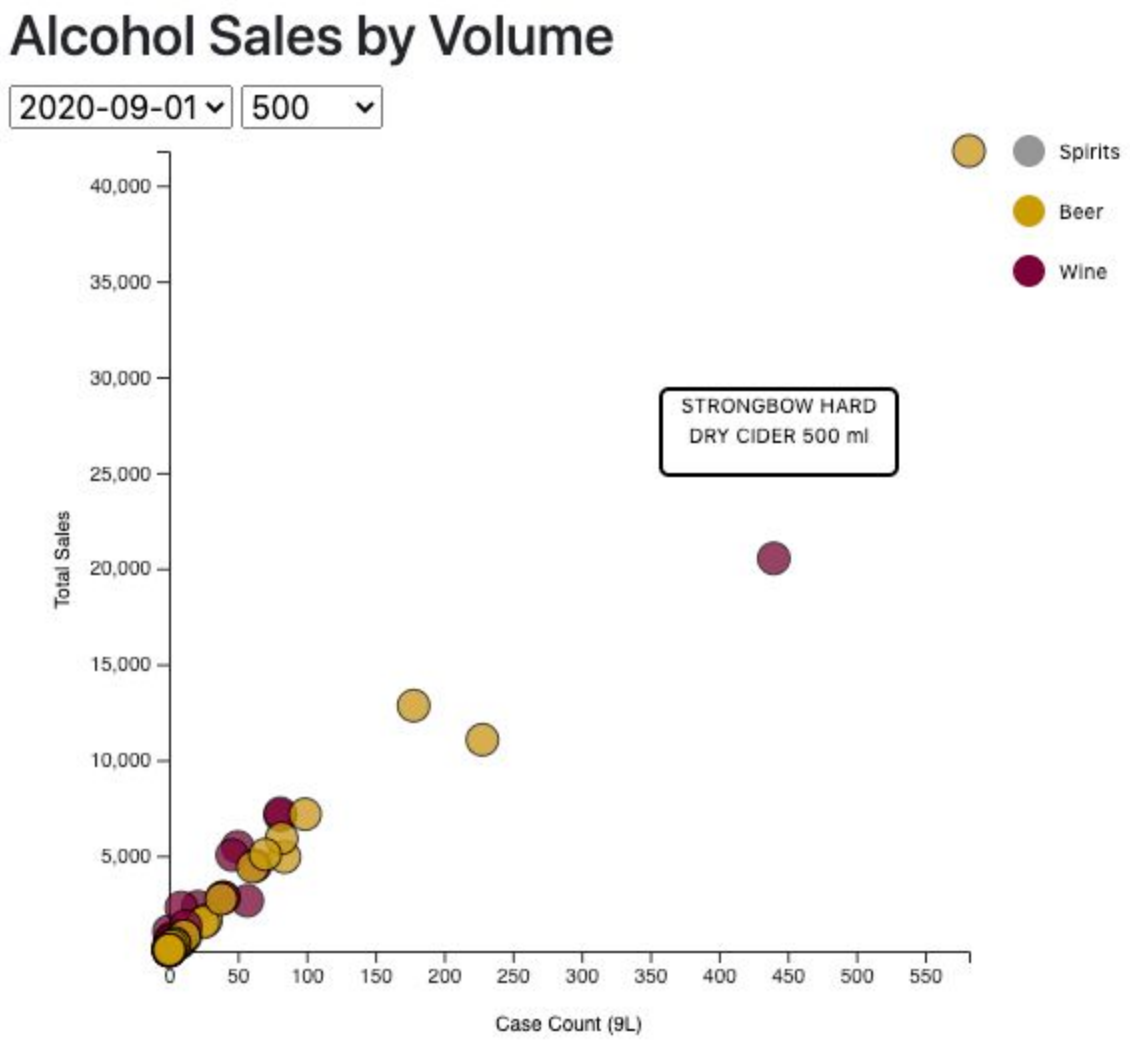
Now with legend, and all bubbles colored properly. November 2019 is missing class codes so those circles will be black unfortunately

Tooltip Update on Search Line Chart:



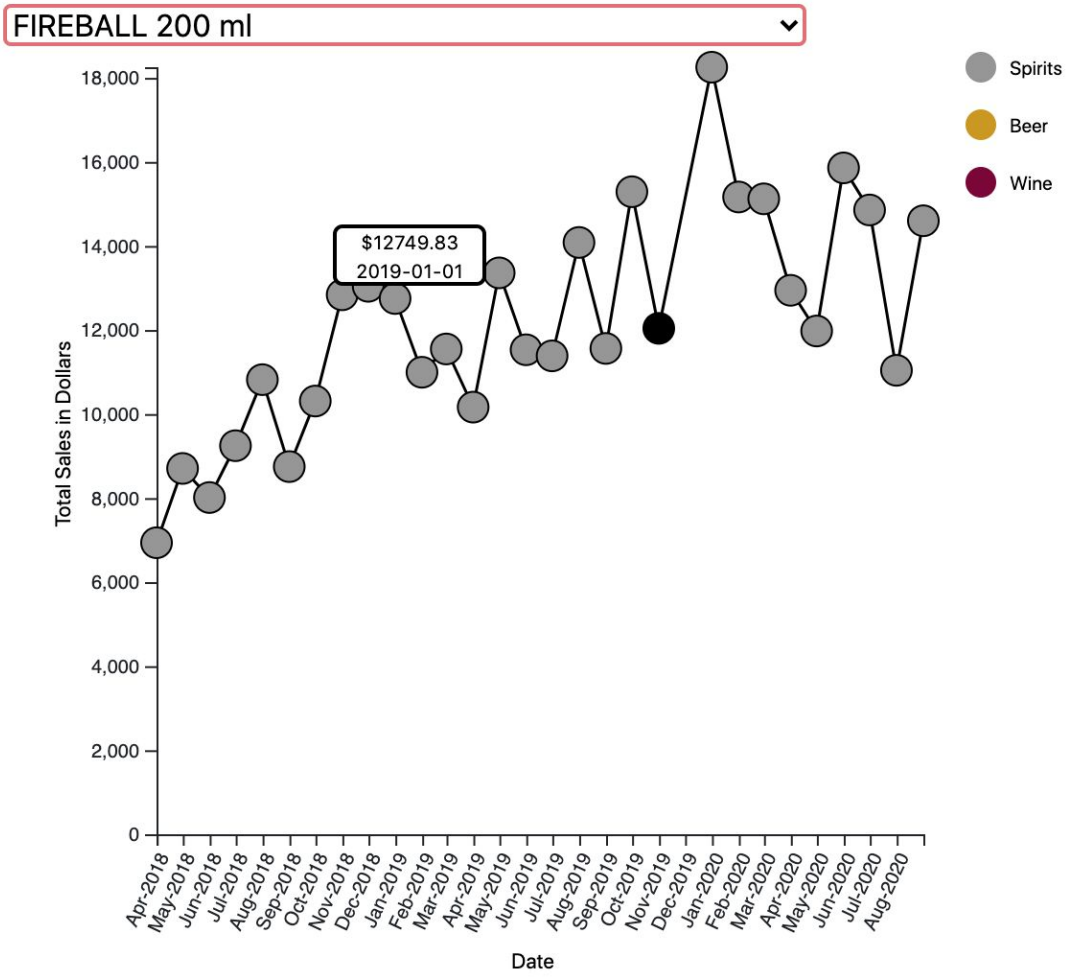
Tooltips have been fixed and now show the sales total for that month.

Tooltip Update on Bubble Chart:



Tooltips have been fixed and show the product name

Update of Product Line Chart:



I added more information to the tooltip to make it easier to read the date form the X axis below, also I added units.

Data Preparation for Categories Visualization

Data published by the Utah Department of Alcoholic Beverage Control includes the monthly sales figures of alcohol in utah divided by categories with an alphanumeric category code (ex. “ADF”). We researched this and determined that these codes come from a terminology called the “Control States Codes”.

In their raw form, the sales figures are highly granular, divided into categories including:

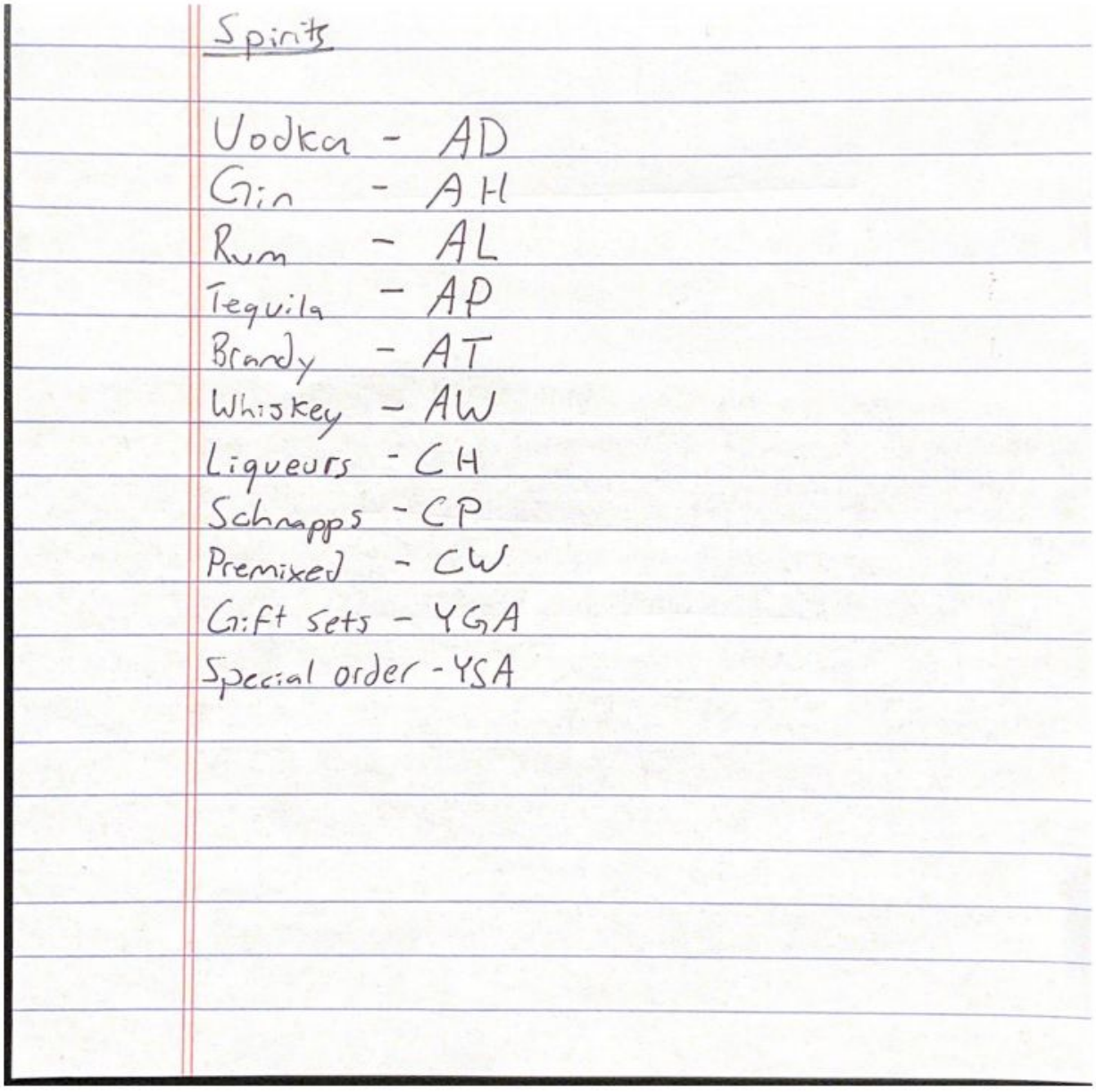
- ALU - Rum, misc
- ALP - Rum, spiced and flavored
- ALJ - Rum, dark
- GLC - Cider, basic
- IHL - Sparking wine, extra dry
- PLP - Red varietal - Barbera

In order to create the visualization we want, with multiple levels of granularity, we had to decide how to categorize these.

In some cases, we could break apart the CSC codes into categories. For example, codes beginning with “A” are spirits and codes beginning with “AL” are rum. However, this pattern does not always hold. “YGA” represents “gift sets, spirits” which we also want to categorize as spirits.

We decided to categorize them into a system with three distinct levels of granularity. At the highest level, everything is categorized as a “wine”, “beer”, or “spirit”. At the next level, we include a slightly more specific categorization, such as “rum”, “gin”, or “vodka”.

Full Classification of CSC Codes to Categories



Spirits

Vodka - AD
Gin - AH
Rum - AL
Tequila - AP
Brandy - AT
Whiskey - AW
Liqueurs - CH
Schnapps - CP
Premixed - CW
Gift sets - YGA
Special order - YSA

R, T, YST

Beer

RCP - Flavored Malt Beverages

TCP - ~~Gluten free, wheat, bef, fruit, etc~~ Other

TH - Lager

TN - Ale

TUB - Domestic

TUD
TVF
TVH
TVN
TVP
TVU
} Imported

YST - Special Orders

E, G, I, K, L, M, P

Wine

ED - Vermouth
EH - Sherry
EL - Marsala E - Fortified
EP - Madiera
ES - Port
EV - Late Harvest
GF - Fruit Wine
GL - Cider
GP - Sake
IP, IH - Sparkling
K - White
L - Wine Offer
M - Rose
P - Red
YGE - Gift Sets
YSE - Special Order

Data Wrangling for Categories Visualization

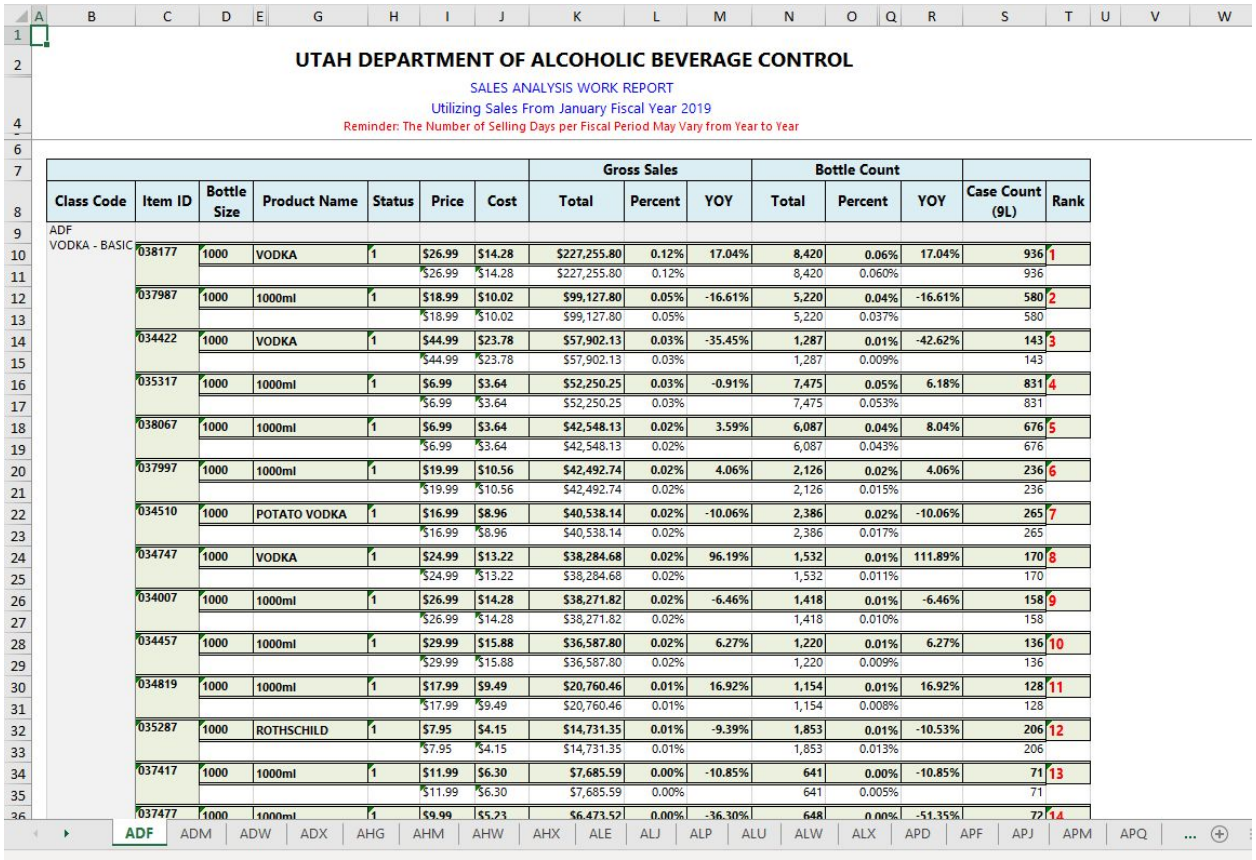
The data used for this visualization came in the format of Excel files posted to the DABC website, with one Excel file representing a month of sales data. A screenshot of a sample Excel file can be seen below.

Unfortunately, these files did not all have a standard naming convention, so it took some work to adjust them before we could run them through our data processing workbook in Jupyter.
Each Excel file contained many sheets, with each sheet representing sales data for a particular sub-category, designated by a three-character code from the Control States Codes specification (ex. “YSE for Wine, Special Order”).

We created a Jupyter notebook which would open an Excel file, iterate through the sheets, and extract the relevant sales data for each CSC code. After creating a classification system which allowed us to aggregate these into useful categories for visualization, as explained in the previous section, we were able to write code in the Jupyter notebook which would combine the sales data extracted from each CSC code into the categorical data we need for our visualization.

This processed data is stored in JSON and CSV files in the “data” folder of our repository. It is these cleaned data files which are loaded by D3 to generate the visualization. Attempting to create a visualization from the raw data would have consumed unnecessary time, computing power, and bandwidth from our site’s visitors since the source files were quite large to start with. In the interest of creating a dynamic visualization which loads quickly and performs well, we decided to use only the processed data.

Sample Raw Excel file.



UTAH DEPARTMENT OF ALCOHOLIC BEVERAGE CONTROL

SALES ANALYSIS WORK REPORT

Utilizing Sales From January Fiscal Year 2019

Reminder: The Number of Selling Days per Fiscal Period May Vary from Year to Year

							Gross Sales			Bottle Count				
Class Code	Item ID	Bottle Size	Product Name	Status	Price	Cost	Total	Percent	YOY	Total	Percent	YOY	Case Count (9L)	Rank
ADF														
VODKA - BASIC	038177	1000	VODKA	1	$26.99	$14.28	$227,255.80	0.12%	17.04%	8,420	0.06%	17.04%	936	1
					$26.99	$14.28	$227,255.80	0.12%		8,420	0.060%		936	
	037987	1000	1000ml	1	$18.99	$10.02	$99,127.80	0.05%	-16.61%	5,220	0.04%	-16.61%	580	2
					$18.99	$10.02	$99,127.80	0.05%		5,220	0.037%		580	
	034422	1000	VODKA	1	$44.99	$23.78	$57,902.13	0.03%	-35.45%	1,287	0.01%	-42.62%	143	3
					$44.99	$23.78	$57,902.13	0.03%		1,287	0.009%		143	
	035317	1000	1000ml	1	$6.99	$3.64	$52,250.25	0.03%	-0.91%	7,475	0.05%	6.18%	831	4
					$6.99	$3.64	$52,250.25	0.03%		7,475	0.053%		831	
	038067	1000	1000ml	1	$6.99	$3.64	$42,548.13	0.02%	3.59%	6,087	0.04%	8.04%	676	5
					$6.99	$3.64	$42,548.13	0.02%		6,087	0.043%		676	
	037997	1000	1000ml	1	$19.99	$10.56	$42,492.74	0.02%	4.06%	2,126	0.02%	4.06%	236	6
					$19.99	$10.56	$42,492.74	0.02%		2,126	0.015%		236	
	034510	1000	POTATO VODKA	1	$16.99	$8.96	$40,538.14	0.02%	-10.06%	2,386	0.02%	-10.06%	265	7
					$16.99	$8.96	$40,538.14	0.02%		2,386	0.017%		265	
	034747	1000	VODKA	1	$24.99	$13.22	$38,284.68	0.02%	96.19%	1,532	0.01%	111.89%	170	8
					$24.99	$13.22	$38,284.68	0.02%		1,532	0.011%		170	
	034007	1000	1000ml	1	$26.99	$14.28	$38,271.82	0.02%	-6.46%	1,418	0.01%	-6.46%	158	9
					$26.99	$14.28	$38,271.82	0.02%		1,418	0.010%		158	
	034457	1000	1000ml	1	$29.99	$15.88	$36,587.80	0.02%	6.27%	1,220	0.01%	6.27%	136	10
					$29.99	$15.88	$36,587.80	0.02%		1,220	0.009%		136	
	034819	1000	1000ml	1	$17.99	$9.49	$20,760.46	0.01%	16.92%	1,154	0.01%	16.92%	128	11
					$17.99	$9.49	$20,760.46	0.01%		1,154	0.008%		128	
	035287	1000	ROTHSCHILD	1	$7.95	$4.15	$14,731.35	0.01%	-9.39%	1,853	0.01%	-10.53%	206	12
					$7.95	$4.15	$14,731.35	0.01%		1,853	0.013%		206	
	037417	1000	1000ml	1	$11.99	$6.30	$7,685.59	0.00%	-10.85%	641	0.00%	-10.85%	71	13
					$11.99	$6.30	$7,685.59	0.00%		641	0.005%		71	
	037477	1000	1000ml	1	$9.99	$5.23	$6,473.52	0.00%	-36.30%	648	0.00%	-51.35%	72	14

ADF | ADM | ADW | ADX | AHG | AHM | AHW | AHX | ALE | ALJ | ALP | ALU | ALW | ALX | APD | APF | APJ | APM | APQ

Merging Process

Since we decided to create three separate visualizations for the data, we split up the project and had each group member create one of the visualizations. At the time of the project milestone, these visualizations were each separate and had not yet been combined.

When it came time to combine, we had to standardize to all use the same version of D3 and adjust some of the code, especially with the tooltips, to make this version work for all the visualizations. We then verified that all of our code used unique names which would not conflict with each other's visualizations. When this was done, we merged our visualizations.

Our combined structure includes one JavaScript file for each visualization, containing the code necessary to generate the visualization. These individual files are called from a unified "script.js" file which uses JavaScript Promises to load all of the data and then creates the component visualizations. In turn, this "script.js" file is invoked by the shared "index.html" file where we set up our page's structure, including both the divs for the

visualizations as well as our storytelling narrative. A common CSS file is also shared, containing the custom styling info for the project.

The end result is a single web page which integrates all three distinct visualizations into a unified narrative format. It tells a coherent story, while allowing viewers to interact with the data in multiple ways and explore it further to answer their own questions.

Overview and Motivation Works Cited

Canham, Matt. 2018. "Salt Lake County is now minority Mormon, and the impacts are far reaching." *The Salt Lake Tribune.* December 09. https://www.sltrib.com/religion/2018/12/09/salt-lake-county-is-now/.

Richards, Conner. 2020. "Utah's population grew more in the last decade than any other state, census data shows." *Daily Herald.* January 02. https://www.heraldextra.com/news/local/utah-s-population-grew-more-in-the-last-decade-than-any-other-state-census-data/article_563d5ee2-9baf-5761-a6cc-7f88eb4d3fb9.html

Taub, Amanda. 2020. "A New Covid-19 Crisis: Domestic Abuse Rises Worldwide." *New York Times*, April 06: https://www.nytimes.com/2020/04/06/world/coronavirus-domestic-violence.html.

Troianovski, Anton. 2020. "In Pandemic's Grip, Russia Sees Spike in Age-Old Bane: Drinking." *New York Times*, April 16: https://www.nytimes.com/2020/04/14/world/europe/russia-coronavirus-alcoholism.html.