Texas:

Mixed Beverage Analysis

Group 1

Betzy Guerra

Saatvi Rajgopal

Jorge Serrano

Kelli Smith

Introduction:

Alcohol is one of the most if not the most popular drink in the world next to water. This is for a good reason, it provides physical relaxation, diminishing stress, it reduces judgment allowing to stimulate your brain's reward system and it releases endorphins. However, is no secret that alcohol misuse will have a negative effect on a person's life. In fact, most studies, and life in general shows that abusing alcohol can lead to a ruined family, career, and many other important things in their life. Specifically in a job, alcohol misuse and abuse can impact several things such as one's job performance, a high chance of injuries, mistakes and overall, it has a negative effect on productivity. In a highly competitive job market such as one in the U.S, a lower job performance will most likely lead to termination and loss of earnings. Therefore, wouldn't it make sense for the alcohol misuse to continue to rise in a person if they had a job termination? Wouldn't it also make sense for there to be a bigger chance for a person to abuse alcohol if they lost their job in a manner that would be considered unfortunate, ill-fated, and unfair manner? Wouldn't anyone who lost their jobs, whether it was fairly or unfairly, and now face financial challenges want to reduce their stress, and release some type of endorphins? With unemployment wouldn't alcohol use rise up? That is what we want to determine, therefore a deeper study of the relationships between unemployment and alcohol use will be done to determine their associations.

Inspiration:

Our team wanted to take the skills we've learned so far during bootcamp to try to prove antidotal claims we've heard about the liquor industry, mainly that is had a positive correlation with unemployment growth.

As we continued to work on this project, we started to ask ourselves - who else would be interested in this dataset? Investors! So, we put on our investor hats to see what insights we might discover.

Hypotheses:

Unemployment and mixed beverage sales have a positive correlation, exception might be 2020 due to forced closures caused by Covid-19.

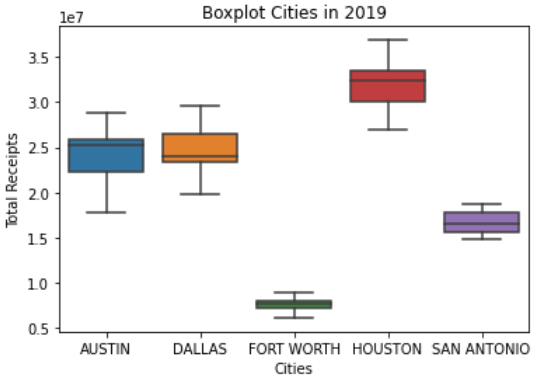
We expect to see mixed beverage sales increase around holidays, such as New Years, Easter, July 4th, Thanksgiving and Christmas.

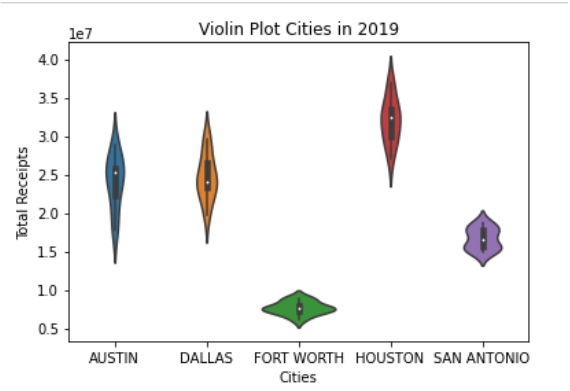
Sources of Data:

Our primary data set for this project was the Mixed Beverage Gross Receipts data made available by the Texas Comptroller of Public Accounts Agency. (https://data.texas.gov/dataset/Mixed-Beverage-Gross-Receipts/naix-2893). We downloaded the data using their API connection and then converted the results into a CSV file for the group to use for our project. Due to storage limitations on GitHub, we decided to filter the data set down to Dallas, Austin, Fort Worth, San Antonio, Houston for just 2019 and 2020. The data set included 24 columns of information and each row of the data set represented a specific location's information for a single month.

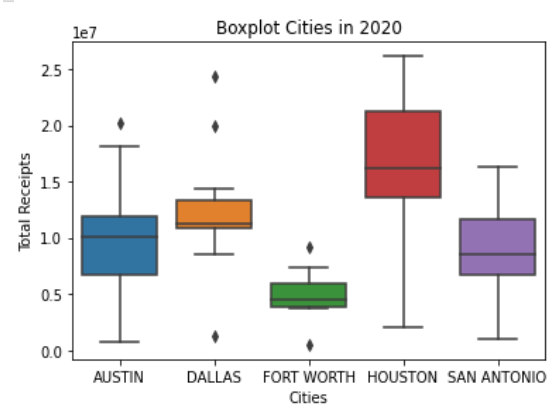
Data Cleaning and Exploration:

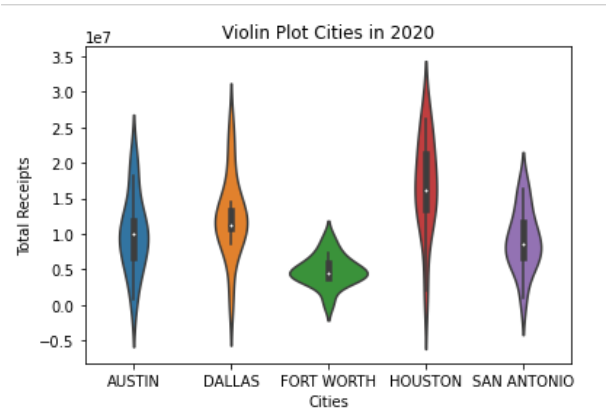
The data was not perfect, but we plowed through cleaning as much as we could so that we could have a data set to start with. Once we felt the data was manageable, we began our initial visualization of the spread of the data working on Box and Violin Plots for the years 2019 and 2020. Although each city, has a different distribution,





The violin plot of the distribution data emphasizes the distribution for each year in addition to showing the decrease in overall sales. In addition, upon a quick view of the plots, we can immediately determine that San Antonio and Ft. Worth have the lowest distribution. Dallas and Austin have very similar distribution sets where Houston towers all the other markets. Lastly, you will notice that there the outlines were eliminated. We realized that Mixed Beverages Sales included liquor sales at hotels, liquor stores, bars, restaurants, and venues that skewed the data.

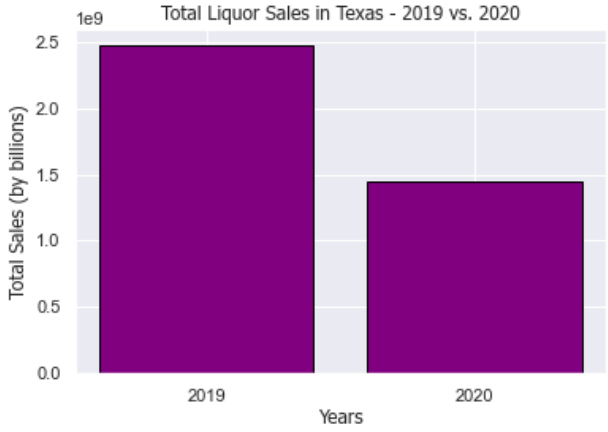




Analysis:

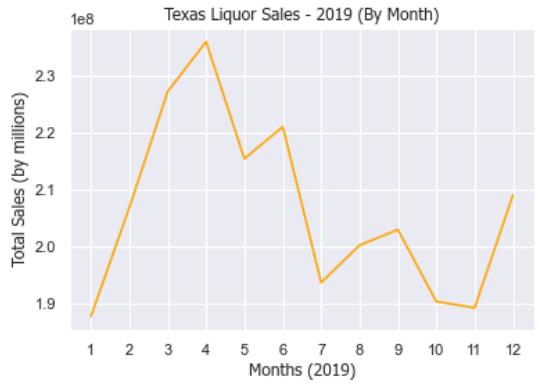
Total Sales – Yearly and Monthly

To limit the amount of data we were looking at due to the size of the data set, we decided to focus on the data 2019 and 2020 to narrow it down. The reason behind choosing these years was because we wanted to see how the data was skewed in 2020, and how a "normal" year of data would compare to it. To look at the data from these years, we made a bar graph depicting total sales from 2019 and from 2020 separately, allowing us to see the difference between the two.



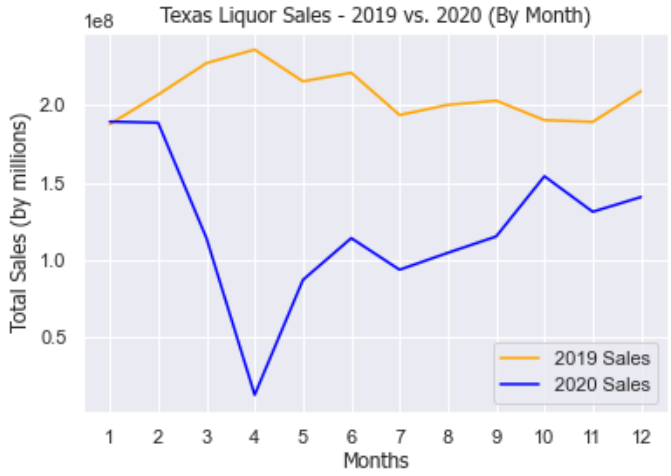
Looking at this graph, we can see that liquor sales in 2020 dropped significantly from 2019. Sales dropped by roughly $1 billion, which makes sense since everyone went into lockdown in 2020, including businesses. Many businesses shut down and lost a significant number of sales than they had in the prior years. One downfall of only focusing on these two years is that we can't see what the data was like during other years that were more "normal"- so we do not know if a fluctuation in sales is normal or not.

To look at the sales from these years a little deeper, we created a line chart depicting monthly sales in 2019 and 2020 separately, as well as in a combined line chart to compare the sales.



Looking at the sales over the months, rather than each year, allowed us to get a deeper look into how the sales fluctuated throughout the years. Since 2019 was our "normal" year of data, we expected to see trend in seasonality; for example, a spike in sales in July due to the Fourth of July,

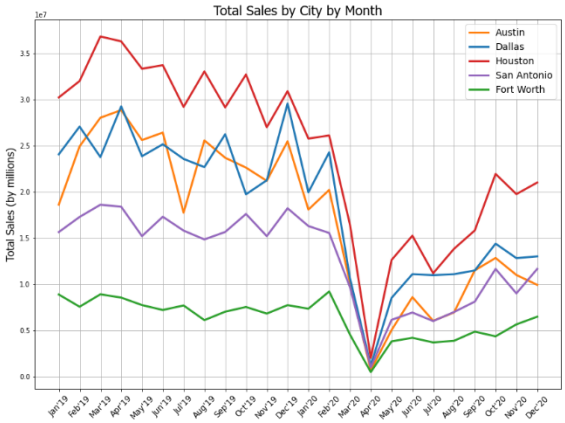
or a spike in October due to Halloween. As our line graph depicts, there was no distinct seasonality trend shown in the 2019 data. In 2019, its peak in liquor sales was in April. On the other hand, looking at the monthly sales from 2020, it tells a very different story. In 2020, sales dipped in April, and would slowly climb in the following months with a few, but less drastic dips. These dips in sales make sense, because as stated previously, the pandemic hit in 2020 and forced everyone to go into lockdown, specifically starting in April. In the following months, we would continue to have a spike in COVID cases every so often, causing people to go into temporary lockdowns a few times throughout the year. This is shown with the data, as the largest dip in sales starts in April, but dips a couple times again in July and November.

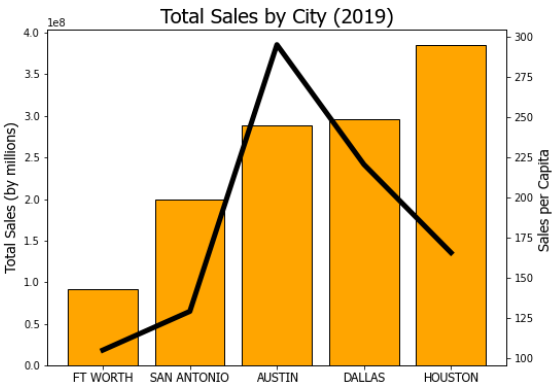


When comparing the two-line graphs side by side, there isn't any correlation in the sales to be found. This is to be expected though, because we intended for 2019 to be our standard year of data and for 2020's data to be more skewed. Overall, we should have looked at a few more years prior to 2019 as well, so we had a better idea of what a normal year of liquor sales in Texas really looks like.

City Level Trends

We then took our analysis one level deeper as we began to look at city level trends. In the line graph below, we plotted the total sales by city and month over 2019 and 2020. We were expecting to see consistency and seasonality across all the cities. When looking at 2019 specifically, our "normal" year (not impacted by forced closers and occupancy mandates), we found only 2 months





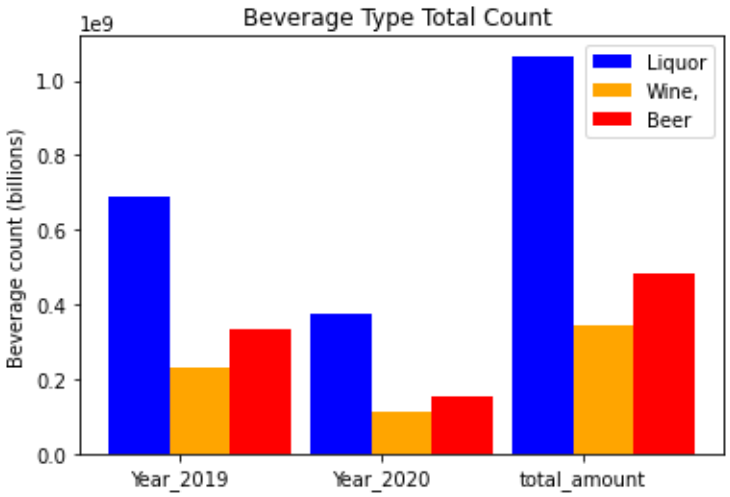
where all 5 cities there trending in the same direction. These months were April and December of 2019.

We also noticed how Houston always had the highest sales. Dallas and Austin typically battled back and forth each month fighting for 2nd and 3rd place. San Antonio and Fort Worth came in 4th and 5th place consistently. We guessed that population size was the main contributing factor to sales volume. We then normalized the city level data by calculating sales per capita. That is when we found that although Austin is 3rd in total sales volume, it is first in sales per capita with $295 per person. Dallas came in second at $221 per person and Houston in third with $166 per person.

Beverage Type Trends

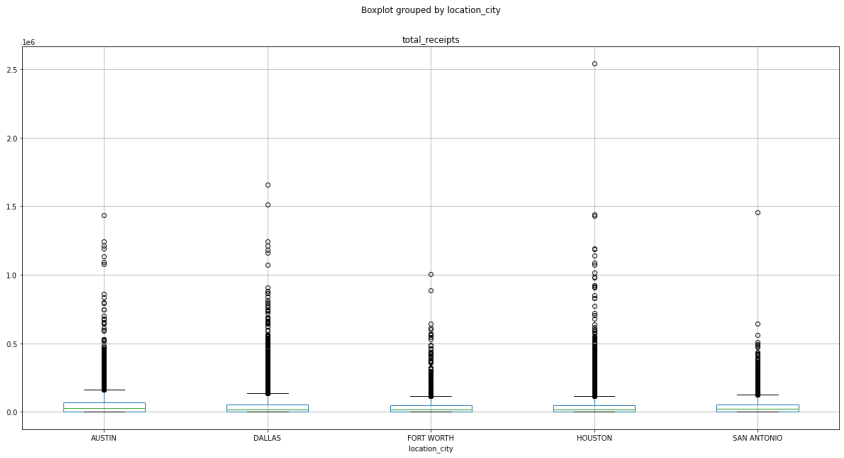
One of the things that we wanted to know through the data that was provided to us was, "What type of beverage sold the most in 2019 and what sold the most in 2020?" After cleaning and

analyzing the data, it was found that the beverage that sold the most was liquor, followed by beer, and lastly wine. This was the trend for both 2019 and 2020.



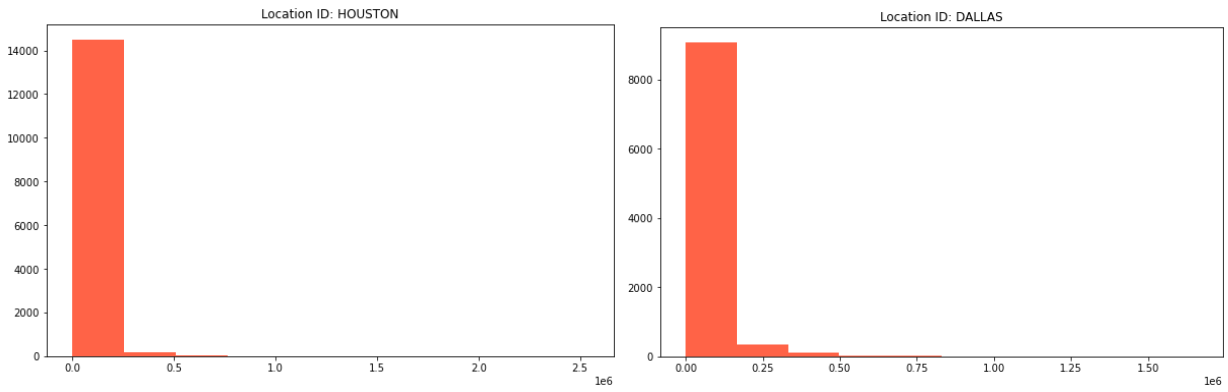
ANOVA:

Next an ANOVA test was ran with first with the two cities of our t-test and their total number of receipts. The test output was statistic=-6.39, pvalue=2.60, the same as our t-test, and then the same test was ran with all of the five cities and their total number of receipts together, the result of the test being f-statistic=51.085 with the pvalue=5.517, indicating that there is a significant difference between each group. Several tests were ran after pairing each city with each other. The p-values for each pairwise t-test seemed to suggest that the mean of total receipts in both the cities of Fort Worth and San Antonio were most likely different than that of the other cities, with each p-value of both cities was below 0.05. Just like in in the t-test it is important to note that the outliers for each city were not removed before the tests were performed.



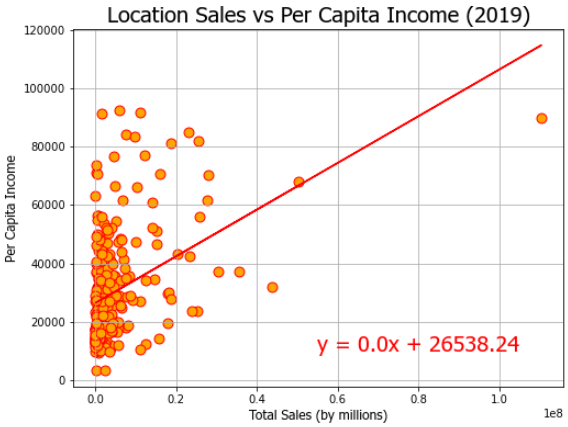
T-tests:

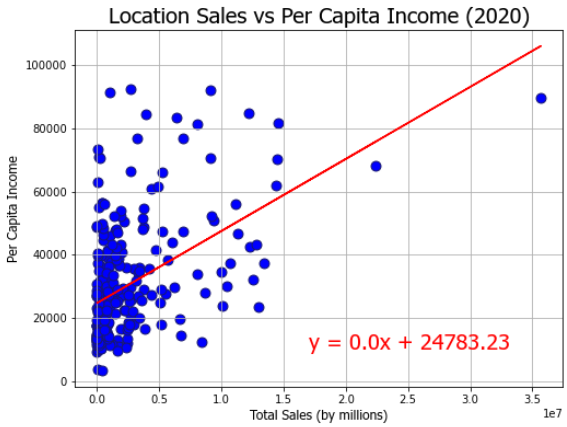
Then a t-test was performed using two of the cities in our data: Houston and Dallas, and their total receipts. The test was ran and it was found that the average pvalue=2.60 of total receipts of Dallas is higher than that of Houston's, showing that there is a large difference in the data groups. However, it is important to note that the outliers were not taken out.



Correlations:

Next, we were hoping to find correlations between the trends we identified in our mixed beverage data set verses census data. As you can see by the heatmap below, we found no strong correlations between total sales or store count vs the datapoints available in our census data set. Per Capita

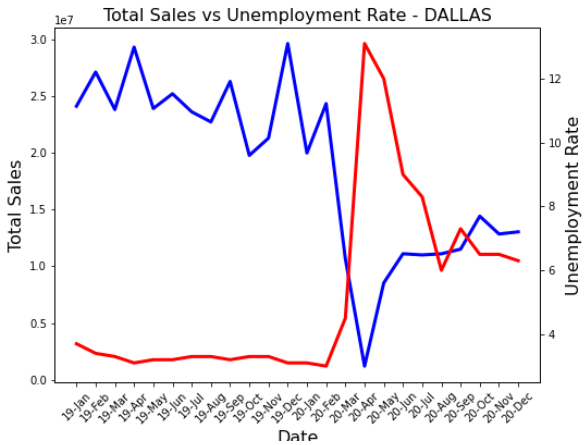




Income was the most strongly correlated with total sales at the zip code level. In 2019 the R-Squared was 0.17 and 2020 the R-Squared increased to .24.

For the multiple regression we decided to see what impact Per Capita Income and Population had on the number of stores per zip code. They were only slightly positively correlated at 17%.

Lastly, we needed to test our hypothesis – are mixed beverage sales positively correlated to unemployment rates? Using Dallas as our sample, we plotted monthly total sales vs monthly unemployment rates over the two-year time frame. Briefly, it was easy to spot that these two trends are negatively correlated, mostly due to the impact of Covid-19. Covid-19 had a major impact on the hospitality, entertainment and food service industries leading to forced closures,



Dep. Variable:	store_count	**R-squared:**	0.170
Model:	OLS	**Adj. R-squared:**	0.164
Method:	Least Squares	**F-statistic:**	26.48
Date:	Wed, 04 Aug 2021	**Prob (F-statistic):**	3.46e-11
Time:	12:05:28	**Log-Likelihood:**	-1174.1
No. Observations:	261	**AIC:**	2354.
Df Residuals:	258	**BIC:**	2365.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2390	3.877	0.062	0.951	-7.396	7.874
Per Capita Income	0.0005	7.51e-05	7.259	0.000	0.000	0.001
Population	8.868e-05	8.25e-05	1.075	0.283	-7.37e-05	0.000

Omnibus:	182.833	**Durbin-Watson:**	1.585
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2108.752
Skew:	2.700	**Prob(JB):**	0.00
Kurtosis:	15.836	**Cond. No.**	1.28e+05

occupancy mandates resulting in massive layoffs and decreased sales volume. When calculating the OSL regression results, we found that total sales and unemployment are 77% negatively correlated. To further test this hypothesis, a longer time frame should be analyzed and 2020 should be excluded as an outlier.

Conclusion:

Overall, we found our two hypotheses to be false. Firstly, we expected unemployment and mixed beverage sales to have a positive correlation, with one exception being due to COVID and forced closures. As we saw in the regression models, we found unemployment and mixed beverage sales to have a negative correlation.

In addition, our second hypothesis was that we expected to see some sort of trend in seasonality in the two years of data, for example, a rise in sales in July due to the Fourth of July. But based off the graphs formulated from the data, we did not see any distinct trend in seasonality.

Limitations:

We had to limit the size of our data set to make it more manageable for the short timeframe we had to work on the project. We filtered the data set down to 5 cities for 2019 and 2020. We also capped the number of rows per month when we downloaded the data through the API connection.

As we began to dig into the data, we soon discovered that location types included large venues, hotels, fast food restaurants, not just liquor stores as we originally anticipated. These location types skewed our analysis considerably.

Finally, that data was reported as monthly sales, so we didn't have the ability to look at weekly or daily sales, which impacted us being able to thoroughly analyze sales around specific holidays.

Future Work:

If we had more time to work on this project, we would like to analyze the entire dataset, for all years and cities available. This would allow us to truly test our hypothesis regarding the correlation between unemployment and sales over longer periods of time.

Secondly, we would like to use a google API connection to segment our data into the different location types – venues, hotels, restaurants, etc. Our guess is that when analyzing the data by location category it would decrease the number of outliers we experienced when looking at the data set as a whole. For the outliers that remained, we would exclude those from our trend analysis.

Finally, there were data points included in our data set that we didn't have time to explore, such as inner/outer city designations and responsibility start and end dates. And for the data points we did explore, we would like to take our analysis further. Examples being beverage trends at a city level and store level trends, including average store sales, what percentage of stores had increases or decreases in sales from one year to the next, binning locations based on total sales for more apples-to-apples comparisons.