Report

By: Victor Yuan, Derek Leung, and Laura Diao

Hypothesis:

Our null hypothesis states that higher taxes on alcoholic beverages, or taxes on those substances in general is unrelated to prevalence of drinking. Thus, we defined our testing hypothesis as: by observing the relationship between taxes on alcohol among all fifty states, there would be a significant and positive relationship between higher taxes on alcoholic beverages and lower rates and prevalence of binge drinking - based off of the given data. The main metrics we used for our observations include finding the mean of prevalence values, HeatMaps, regression, and charts to examine this relationship. The overall significance behind us examining this relationship, is to determine if the alcohol tax on binge drinking prevalence is effective enough, or requires modifications implemented. Throughout our investigation of this issue, we assumed that alcohol tax rates were meant to reduce the rate of Binge Drinking cases, and that other variables, such as Diabetes or Arthritis were not significant to our search.

Testing and Procedures:

In order to clean the data and make the given dataset more readable, we have removed all of the columns which yield no meaningful data. Without removing columns with irrelevant information to our goal, it would be harder to draw conclusions among various variables, to compare them and end with a reliable conclusion. In addition to removing empty columns, implementing pandas, we dealt with rows and deleted them, due to reasons such as no significant data available due to the lack of respondents. We did not face any significant challenges when cleaning the data. In the columns that we deleted, since what showed up in the code editor demonstrated that most of the values were NaN, we assumed that the rest of the values in those columns (which was significantly larger than what was visualized in the code editor) were also NaN. In order to confirm this assumption, we found the number of occurrences of NaN values and compared it to the overall length of the given dataset table, and demonstrated that the ratio of undefined data values made up the majority or entirety of the column, thus we could disregard those columns. After cleaning the data and focusing on Binge Drinking cases, the size of the dataset was reduced from 519718 rows × 34 columns to at least 4387 rows × 21 columns.

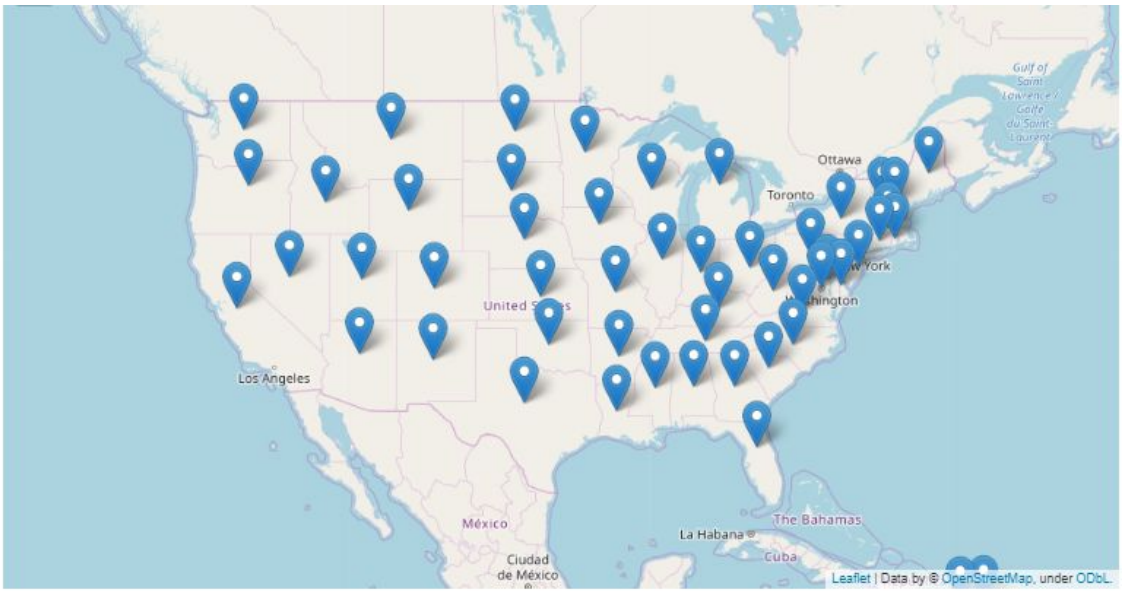
To further narrow the scope of our testing hypothesis, we first found that Beer had the only excise tax data consistently collected from all 50 unique locations, and proceeded to compare the excise tax on beer among those locations. Specifically, we drew the focus of our observations to two questions: Amount of alcohol excise tax by beverage type (beer), and Binge drinking prevalence among adults aged >= 18 years. In the former, we noticed that there were more than 50 cases covering Binge Drinking, we found that two states had binge drinking records for 2012 and 2015, and conducted further tests on the 2015 cases since they are more recent.

For the latter question we made some assumptions while dealing with binge drinking prevalence, assuming that the Data Values to be percent prevalence among Crude Prevalence and Age-adjusted Prevalence, and made our problem more tractable by also assuming Data Values within the given confidence interval are accurate. In the next step, we checked what kind of prevalence is measured, isolated crude prevalence from data, found the mean (of years) prevalence of binge drinking by state. The same was conducted but instead isolating age adjusted prevalence from data, and mean (of years) prevalence of binge drinking by state. After computing the mean so there was one data value per state rather than dealing with cases from multiple years, so that we could graph each type of prevalence by increasing tax rates on Beer. We have provided figures below with Prevalence graphed on the Y axis and tax rates on the X axis.

To generate the Geospatial Map, we used the latitudes and longitudinal data provided to see where each survey was conducted, and we found that each location was at the center of the state in which the survey was conducted.

- Data Visualizations and Reported Findings

Figure 1:



The GeoLocation Data isn't unique for each observation. Instead, each state has only one GeoLocation data. The dataset contains such information on 54 U.S. territories

Figure 2:

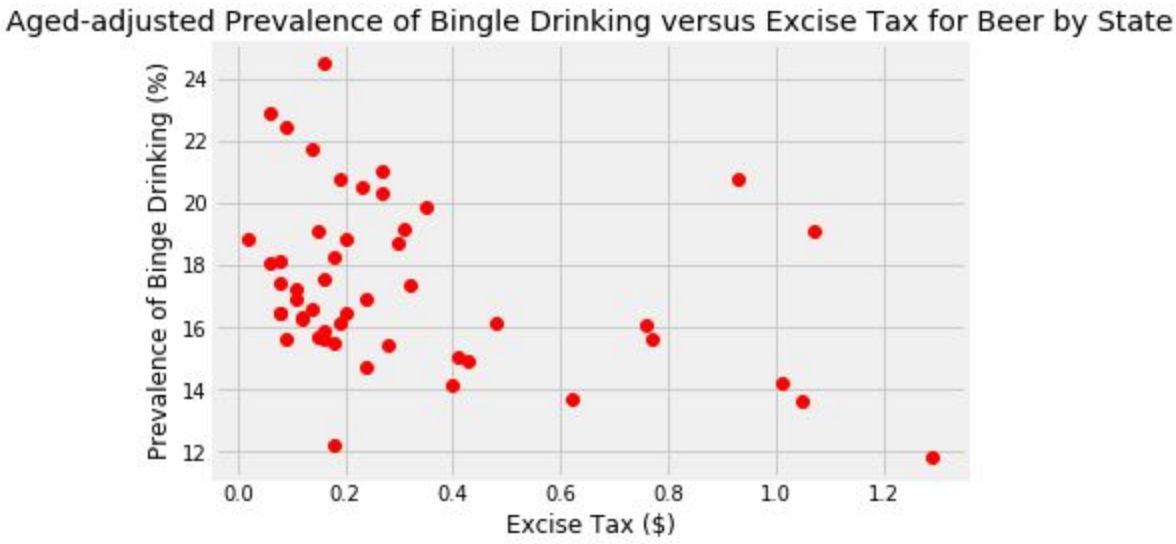


Figure 3:

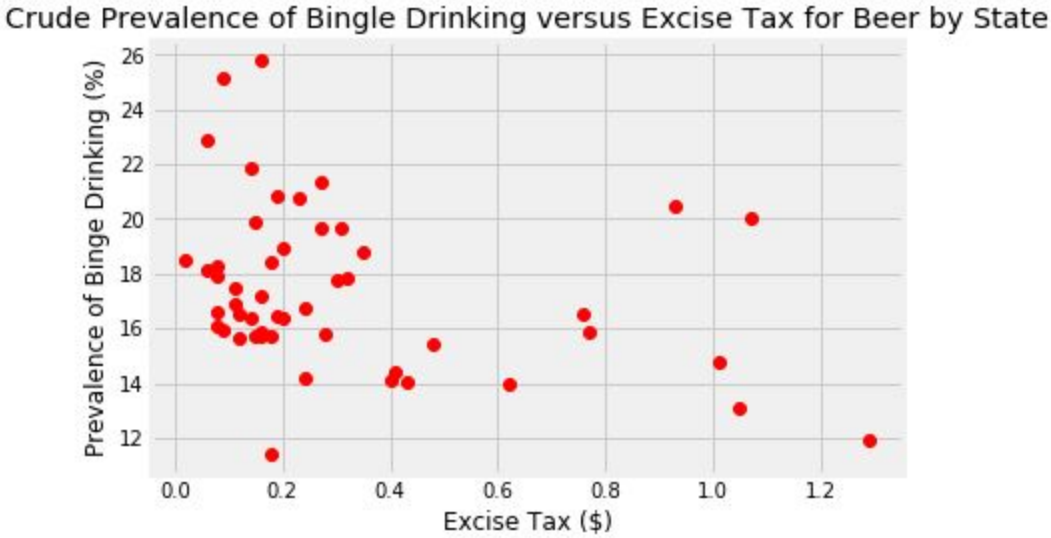


Figure 4:

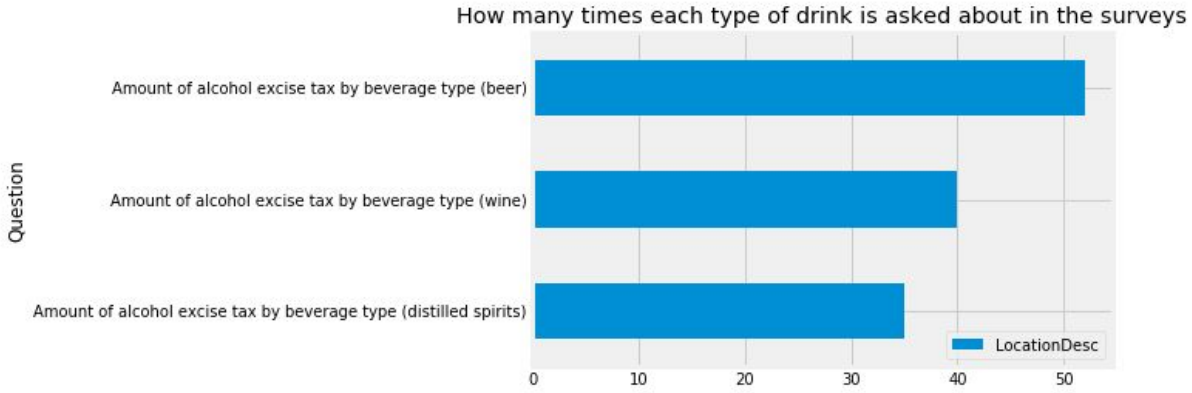
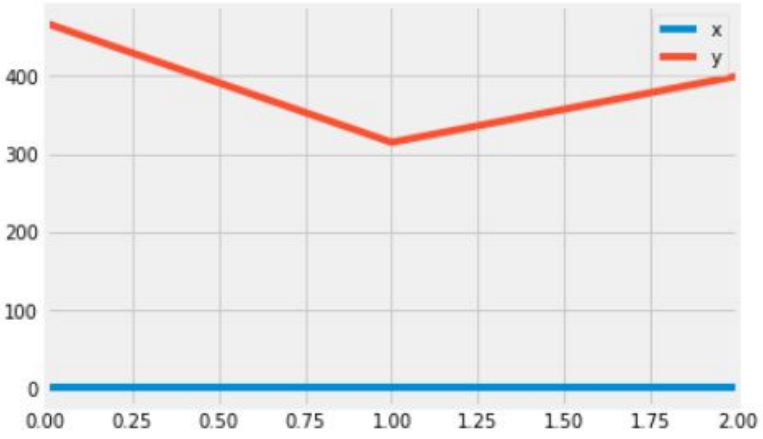


Figure 5:

NLP vectors for each drink



	x	y
0	0	468
1	0	315
2	0	400

NLP and other Analysis

Since there was no Response data, we could not apply any Natural Language Processing on the sentiment or common language of the patients of chronic diseases. However, we applied NLP to the Question, from the given data set, focusing on each alcoholic drink type, and computing a vector scaled by how many times it appears in the data. The vectors for each alcoholic drink, are graphed on Figure 5 to demonstrate that Beer is the most consistent for all 50 states, to support our focus on only Beer throughout our observations on the relationship between tax rates and their potential efficacy against binge drinking prevalence.

Observing the Geospatial map, in figure 1 we found that each marker in the states were located at the center, this means we cannot differentiate between different regions within state, counties, etc, and could only conduct comparisons between states only. Which is a limit to our findings.

Proposal & Conclusion

Based off of our observations and experiments, we have rejected the null hypothesis, which states that there is no significant, positive relationship between increasing tax rates and lower binge drinking prevalence values. Because we found a -0.33253456154382516 correlation coefficient using the age adjusted drinking prevalence data, there is a weak negative correlation between excise tax on beer and drinking prevalence. This suggests higher excise taxes on beer will reduce drinking prevalence. The P-value found with this data was 0.018300411590148696 < .05 which indicates that the results were statistically significant. Therefore, we reject the null hypothesis in favor of the alternative which states that excise tax does impact drinking prevalence.

There is a possible type-1 error in our conclusion. This would be the case if the null hypothesis is actually true when we rejected it. Another issue we ran into was the lack of data for the Response to the Question column provided, and no explicit confidence interval for the reported prevalence values. Specifically, it would have been useful since applying Natural Language Processing to the former could have allowed for more analysis on the natural language in the dataset. Since we found a significant relationship between higher tax rates and lower prevalence values of Binge Drinking, in states where drinking prevalence is an issue, we propose an increase in excise taxes to reduce Binge Drinking prevalence.