Homework 09: Project Part II: Analysis write-up

Group Members: Andrew Tate and Jacob Benson

Title: Drunk Driving Deaths Compared to Alcohol Consumption and Weather History

Introduction

For our project we wanted to answer questions about DUI deaths and drinking in general nationally. We wanted to compare national drinking to DUI deaths to see if consumption increases deaths, compare weather to both drinking and deaths, along with other variables.
To accomplish these goals, we used several datasets, one on average monthly temperature by US state, one on alcohol consumption per state by liter of ethanol, and one on impaired driving death rates for each state between 2012 and 2014. We tackled this by both independently creating our own visualizations to try to come up with as many correlations and conclusions as possible. The main questions we had where:

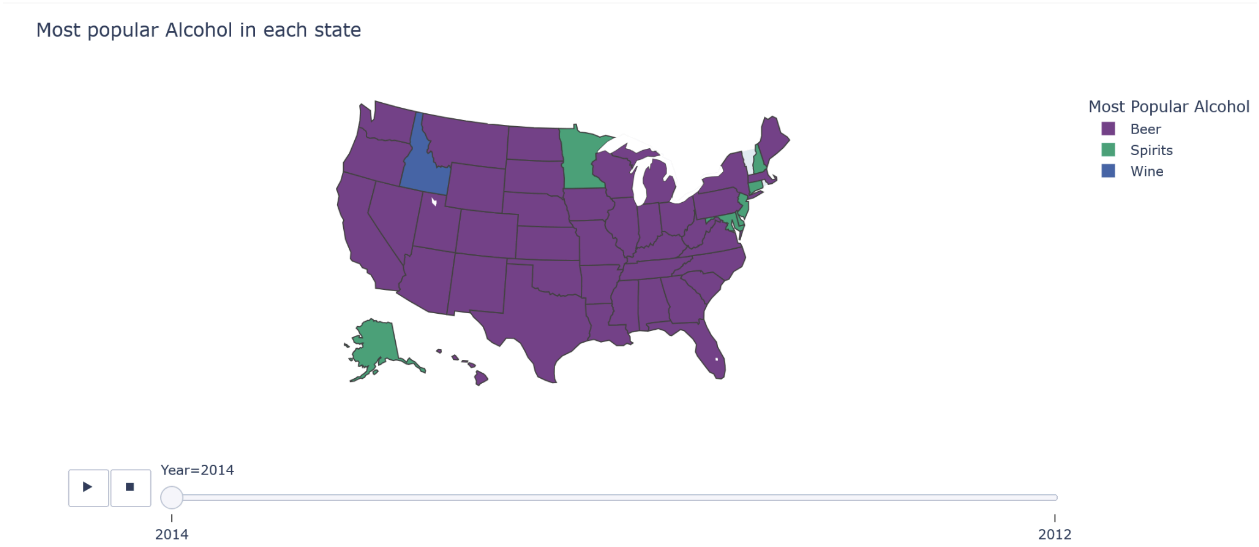
- Does a state consuming more alcohol result in larger DUI deaths?
- Does the temperature affect how much a state drinks?
- Does the type of alcohol a state consumes correlate to DUI deaths?
- Does the temperature of a state affect DUI deaths?

Methods and Results

For this project, there were two main difficulties with the data. The first is that the death dataset was clamped to 2012 and 2014, which caused a huge lack of accuracy as we can only extrapolate two years of data from every dataset. The other challenge is that we are trying to present 3 categories of data on one visualization. For each visualization we need to showcase the geographical information alongside 2 other columns, for example, alcohol consumption and average temperature in the state. We both came up with different ways of showing this each with their own merits. Now we will each touch up on our separate processes:

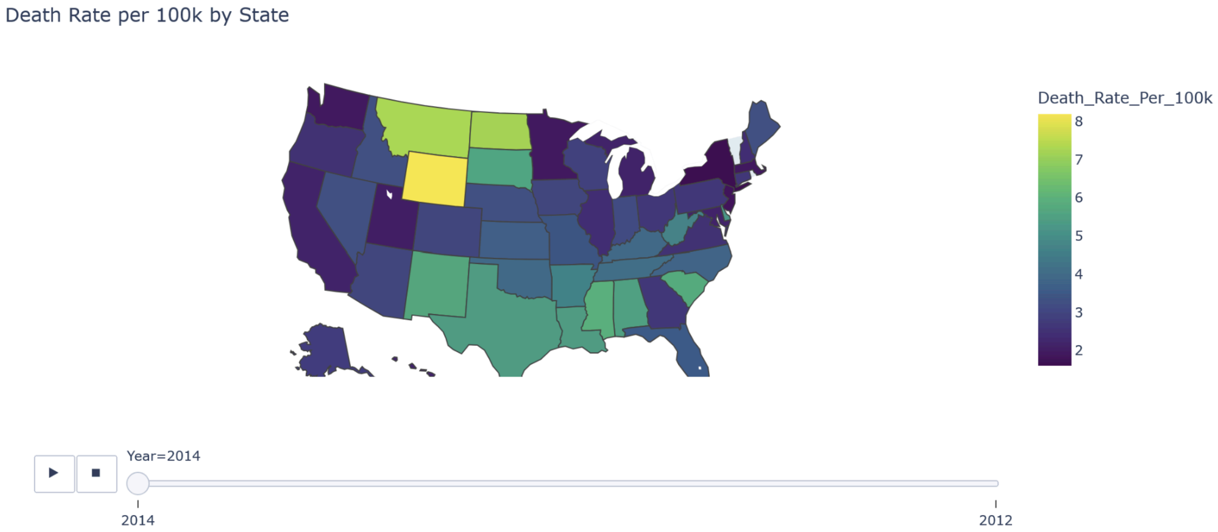
Andrew

I thought the best place to start was to create a large dataframe that aggregates all of the data from the 3 dataframes, discarding the ones I don’t need. I filtered the alcohol preference dataframe to 2012-2014 and created a new dataframe containing a “State” “Year” and “Death_Rate_Per_100k” columns from the dui dataframe. I inner-joined these dataframes together. I then created a lambda function to calculate a “Most Popular Alcohol” column for each state and each year. As shown in the visualization below utilizing plotly’s choropleth:



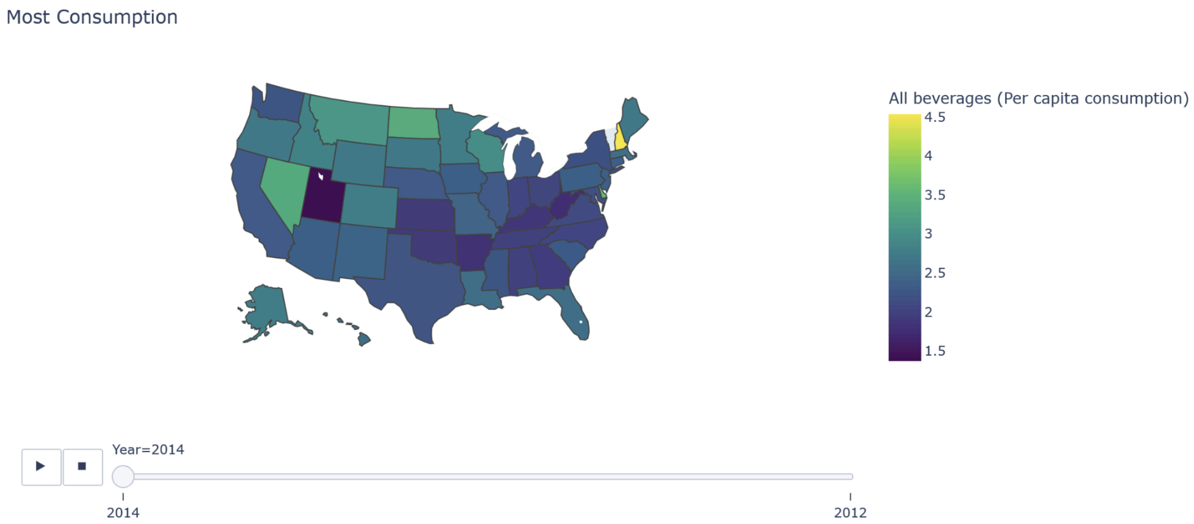
Here we can see that beer vastly outperforms all other types of alcoholic beverages, with wine being the least popular.

I also used this dataframe to create a DUI Death Rate choropleth:



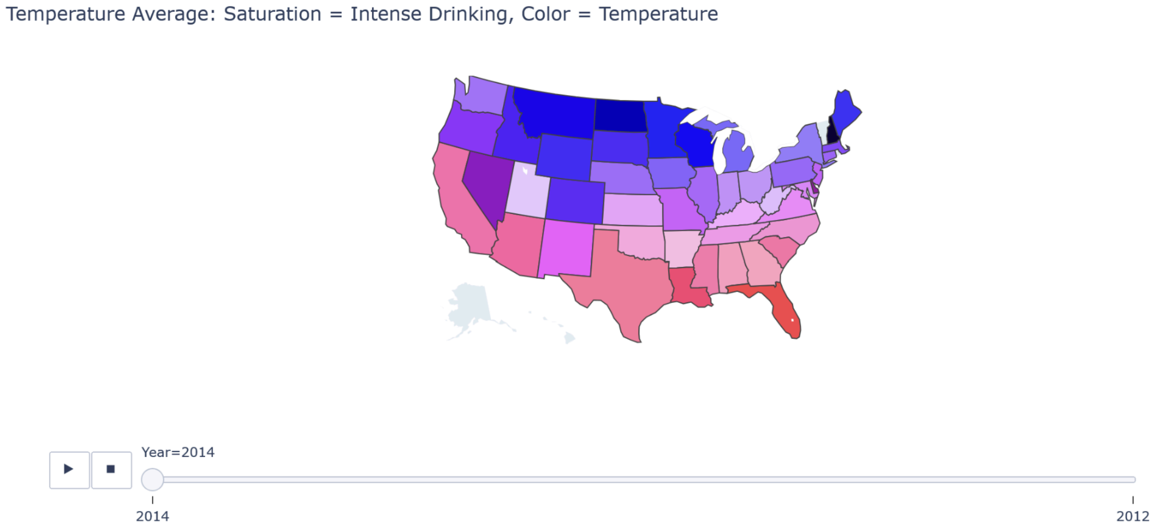
Here we can see that Montana, Wyoming, and North Dakota have large death rates for north states. Conversely New Mexico, Texas, Louisiana, and other southern bordering states also have large death rates. Meanwhile the middle states have relatively low rates.

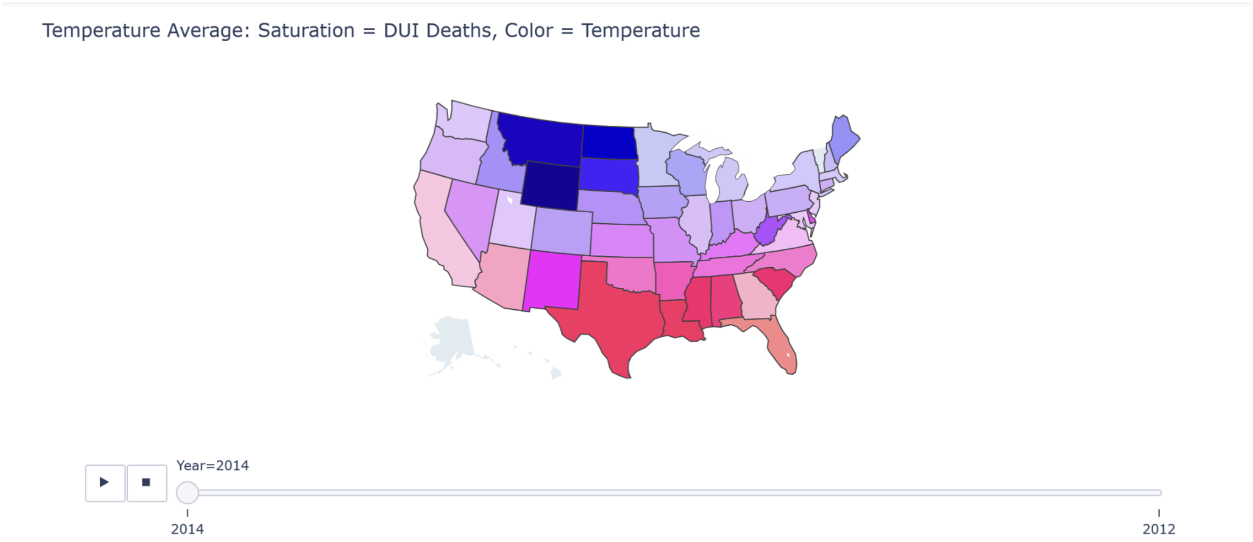
And Finally, Most Consumption:


Most Consumption

Above we can see that northern states have lots of consumption, with Utah having virtually none. New Hampshire has a ton of consumption, for some reason.

Now the challenge was merging these values to make some meaningful connections between our data. At this point I also imported the temperature dataset. I resampled the average temperature to years instead of months, and filtered it to 2012 and 2014. I then merged it on state. I now needed a way to represent this 2D data. My solution was to have a color gradient system. The temperature is represented by a gradient between red and blue. Blue means a state is cold, while red means a state is warm, purple would be somewhere in between. Then the color's saturation represents some other statistic, either alcohol consumption, or death rate. For example in a choropleth that compares alcohol consumption to temperature, a dark blue state would be a heavy drinking, cold state, while a light blue state is a light drinking cold state. Conversely, a dark red state is a heavy drinking warm state and a pink state is a light drinking warm state. This allows for our edge cases to really jump out. These visualizations are shown here:


Temperature Average: Saturation = Intense Drinking, Color = Temperature

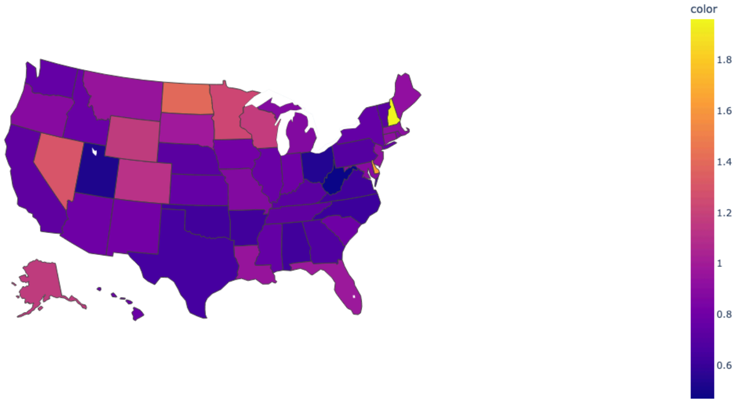

Temperature Average: Saturation = DUI Deaths, Color = Temperature
▶
■
Year=2014
2014
2012

Here we can see a few interesting things. With the first visualization, we can see that people in the northern states vastly drink more than people in warm, southern states, with two exceptions in Louisiana and Florida. The bottom graph shows us equally useful info. Here we can see that there are pretty equal numbers of saturated red and blue states, telling us that the temperature of the state doesn't seem to have an effect on the DUI death rates in that state. Here you are just as likely to die of a substance-influenced death in Montana as you are in Texas.

https://colab.research.google.com/drive/1fl3uQZIWGrrQOJR5gxRvc70_AdeTk6UK?usp=sharing

Jacob

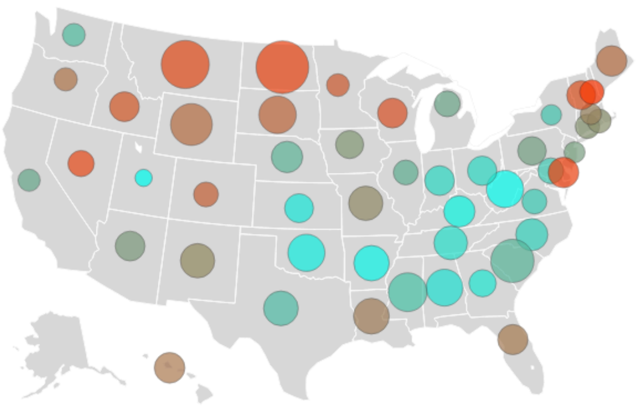
I first took the data from the two datasets, merged them together, and filtered our drinking datasets for the years 2012, and 2014 (the only years provided for driving death statistics. I then created a simple choropleth for the prevalence of drinking in 2014 by states.



We see some general trends of drinking, with lesser drinking in the "Bible Belt" and especially in Utah.

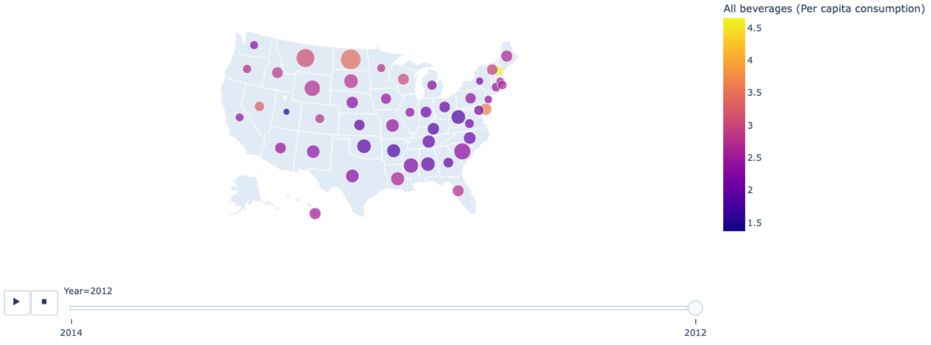
Next I wanted a good way to display two quantitative variables on a locations map by state, which I first thought to do opacity of color, but then opted to go with a bubble map, with size of the bubble in each state relating to driving deaths, and the color pertaining to the level of alcohol drinking in that particular state. My first challenge was West Virginia, which had a comma separating the state name followed by a latitude longitude pair, so I made a function to remove that comma, which I could then obtain latitude and longitude pairs for the central location for each state.

Next I used a gradient function that I found on kaggle, and edited it to include the bounds of my data length, and plotted a bubble plot for alcohol consumption vs deaths in 2014.



Impaired Driving Deaths per 100,000
With color corresponding to level of drinking (Liters of Ethanol per Year)

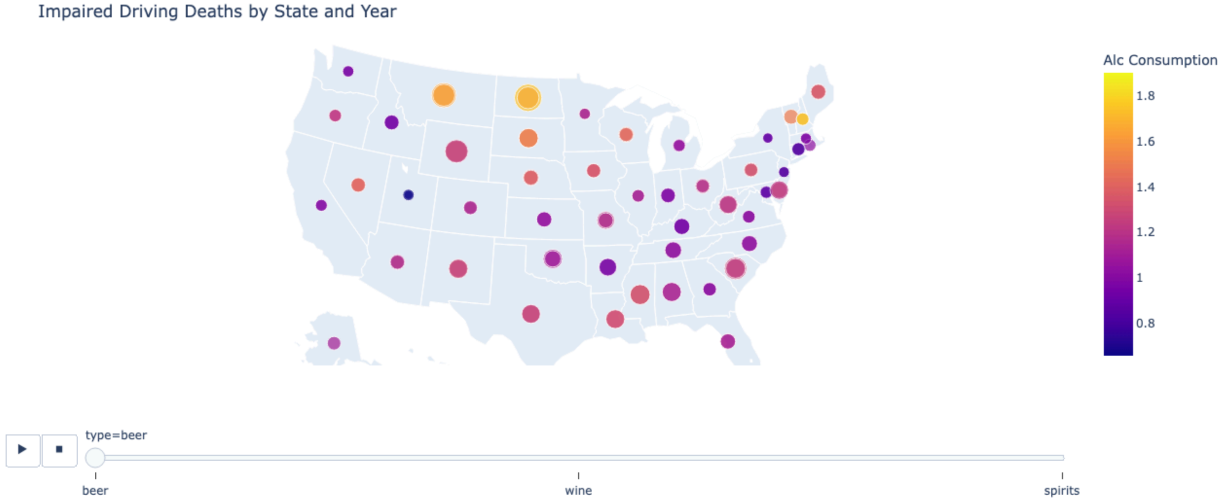
This was fairly interesting as we didn't see the trends that we expected to. Despite there being some areas of higher level of drinking associated with higher levels of drunk driving fatalities, in areas such as West Virginia, South Carolina, and Louisiana, we see a fairly high level of drunk driving fatalities with a fairly low level of drinking. Similarly, we see a fairly low rate of drunk driving fatalities with a high level of drinking in Arizona and New Hampshire.



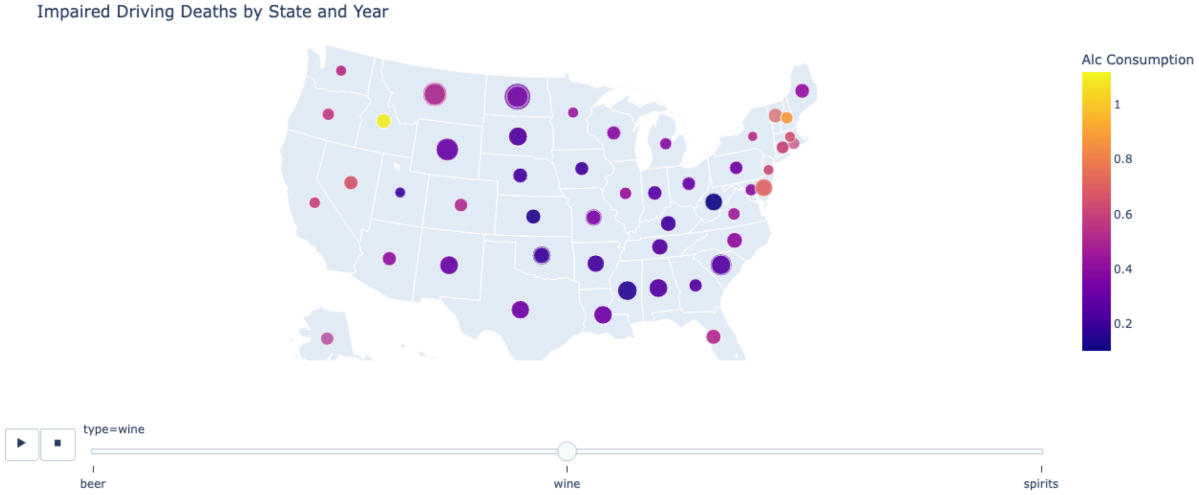
Impaired Driving Deaths by State and Year

Similar results were found pertaining to 2012, with no obvious correlations between level of drinking or alcohol consumption and drunk driving deaths.

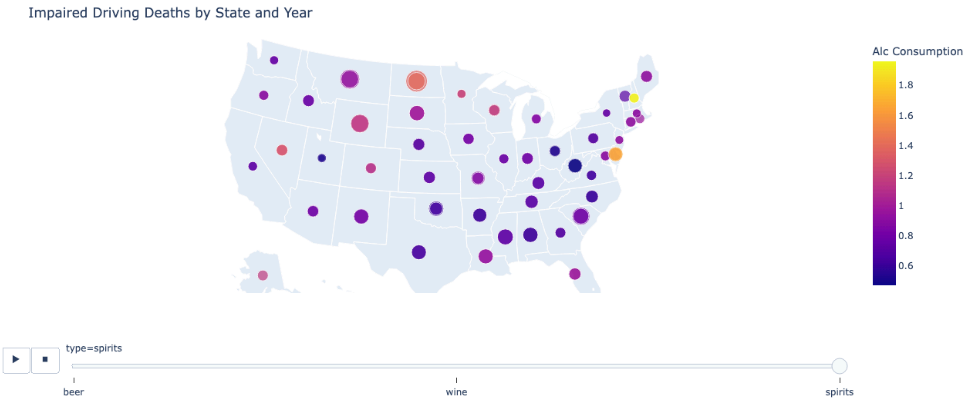
Next I wanted to make bubble plots with a slider for the type of alcohol (beer, spirits, wine) for the color, and the alcohol related driving fatalities as the size of the bubble, to see if there are any obvious correlations between a specific kind of alcohol and driving deaths.



In this plot we can see a correlation between alcohol level and driving deaths with the aforementioned Louisiana and South Carolina having a fairly high level of beer drinking compared to the rest of the United States, which appears to be in greater relation to the higher level of driving deaths. This trend is visible around the country, with high levels of deaths in states such as North Dakota, Montana, South Dakota, all of which record high levels of beer drinking. There are, however, some exceptions like New Hampshire, Nevada, Arkansas, which have either high levels of beer drinking or high levels of deaths, but not both.



In this plot we see little to no correlation to high levels of wine drinking and driving fatalities. The only exception to this could be Delaware, which has a fairly high level of wine drinking in addition to higher levels of deaths in the northeast region (but in the original combined drinking levels Delaware is also an outlier in the east coast).



Finally from the spirits graph, we don't see any obvious correlation between consumption and driving deaths, even though North Dakota, Wyoming and Delaware do present this correlation. The reason I say that there isn't any obvious correlation is because of states such as South Carolina, West Virginia, Texas, Oklahoma, Alabama, or Arkansas, which have relatively low levels of spirit consumption with higher levels of alcohol related driving fatalities.

https://colab.research.google.com/drive/1yx-C8qmOBYUf44L6IYUxCCQr0ITAfu5X?usp=sharing

Conclusions:

After our data analysis and produced diagrams, we can revisit our questions and provide more insight into our questions.

- Does a state consuming more alcohol result in larger DUI deaths?

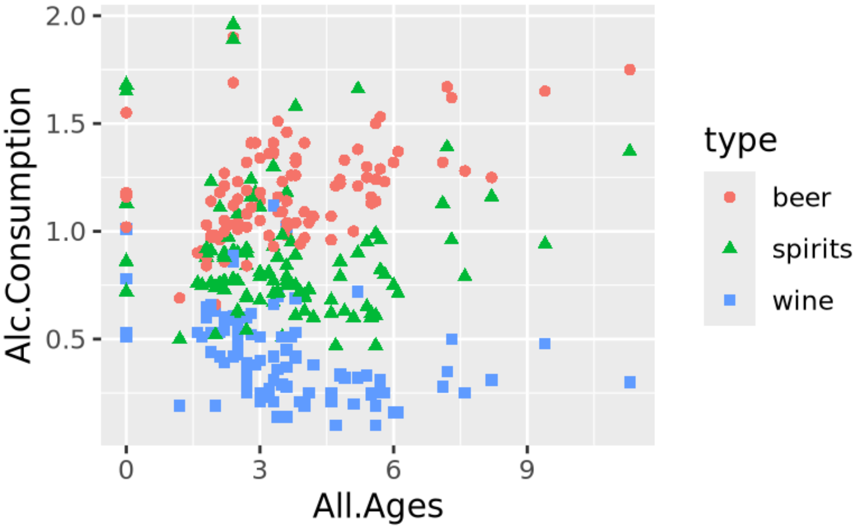
From our visualizations we could not see any obvious trends to support the notion that consuming more alcohol results in higher DUI deaths. Additionally, it appeared that there could be other factors that could contribute to a higher prevalence of DUI deaths.

- Does the temperature affect how much a state drinks?

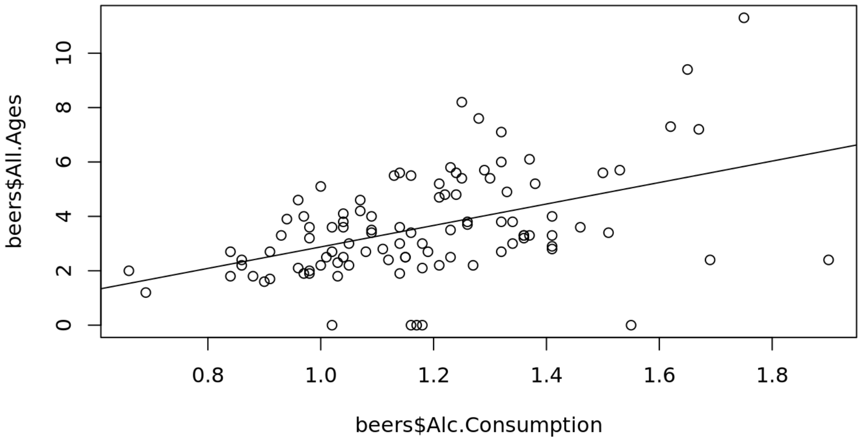
In the temperature vs drinking visualization we can see some trends of higher drinking with lower average temperatures, such as midwest states, with all blue states being fairly highly saturated. Also with higher temperatures in states like Florida and Georgia also show higher levels of drinking. Overall, this could show a slight positive correlation between more extreme temperatures and level of drinking.

- Does the type of alcohol a state consumes correlate to DUI deaths?

From the beer consumption vs dui deaths plot, we can see some correlation between beer consumption and DUI deaths, which could explain some of the higher levels of DUI deaths in the south, such as Alabama, Arkansas, South Carolina, or Louisiana. That being said, there are states with high levels of beer drinking, such as Minnesota, Wisconsin, or Nevada, yet they do not have high levels of DUI related deaths, indicating a lower correlation.



A simple scatter plot with groups of beer spirits and wine show that wine and spirits have little to no correlation, but beer appears to have a slightly higher positive correlation.



Here is the fitted line for regression from the filtered beer data, and it has a correlation coefficient of about 0.46, which indicates a medium strong correlation between beer consumption and drunk driving related deaths. The overall R^2 value for the model is about 0.21, which means that about 21% of the variability in the alcohol related driving deaths are explained by the variability in the beer alcohol consumption by state.

- Does the temperature of a state affect DUI deaths?

From our temperature vs DUI deaths plot, we can see that there may be some trends with higher DUI deaths in areas with more extreme temperatures, such as the Midwest or the South. That being said, areas such as Florida, Georgia, and New Mexico all have fairly high or similar average temperatures as other states in the South, but are not saturated, indicating that they have lower DUI deaths, despite the more extreme temperatures. Similar exceptions can be found in the Northern states as well, with Minnesota, Wisconsin, and Michigan being some of the coldest states, yet lack a high level of DUI deaths.

Discussion:

Some potential errors and confounding variables, and potential future questions:

- Sampling and scope of the data:
    - First the data only came from a couple years (2012 and 2014). A larger sample size for our analysis could show different trends and could even add in some time effect, where we could see how the trend of DUI deaths vs alcohol changes over time.
    - How the data for alcohol consumption was recorded could also be a potential error, as it records where alcohol is bought, and the quantity bought, rather than the alcohol actually drunk in that state. We suspect that New Hampshire has a very large consumption level for this exact reason, as New Hampshire has no sales tax on wine, spirits, and a low tax on beer. For this reason, surrounding states (such as Maine, Vermont, and Massachusetts) will drive to New Hampshire to buy alcohol at the discounted price and then take it back to their homes.
    - Finally, if the alcohol consumption was only recorded from personal sales of alcohol rather than alcohol purchased in bars or restaurants, our data for the consumption of alcohol could be skewed.
- Potential Confounding Variables (non-exhaustive list, but some potential errors):
    - Driver ability - If drivers in particular regions are on average worse than drivers in other states, then this would produce an effect on the drunk driving deaths metric.
    - Road conditions - Road conditions could also affect likelihood of drunk driving deaths, thinking about weather conditions, how often the road is repaved or maintained, or wildlife abundance.
    - Bar prevalence - If individuals in the area are more likely to drink at a bar, pub, or restaurant, then they are more likely to drive drunk than if someone is drinking in their own home. This could show higher DUI death rates in areas of the country where more individuals go to the bar, with lower rates for individuals who stay home and drink (for the same level of alcohol consumption).
    - Religion - We may have mentioned this a little bit, but from our alcohol consumption choropleth, we can see that areas with larger religious populations have a lower alcohol consumption (see Utah or potentially states throughout the bible belt).

- Potential Future Questions:
    - It would be interesting to grab a dataset on general driver performance across the United States. From there we could answer some questions by comparing the DUI rate to the death rate from general vehicular collisions to determine the “best” drunk drivers, or the state that consumes the most alcohol with the best driving rates. Conversely, we could also determine the worst sober drivers, or the state that consumes the least alcohol with the most accidents
    - We could also figure out if there is a direct bible belt correlation by grabbing a data set on religion participation across states and determining if what religion somebody is lowers their chances of dying by DUI, we could also determine if certain types of alcohol are more or less popular across religions.
    - It might be interesting to see if some of those northern states that are not very heavy on DUI deaths are heavy in marijuana usage. It could be that DUI deaths are not accounted for on marijuana, and those northern states are simply getting stoned and then dying by vehicular collision
    - Finally, it would be informative to determine why these northern states are drinking more. We could grab a mental health dataset and see if these states are linked with higher rates of depression. Maybe the cold, mental health, and heavy drinking are all linked?

References:

Linze.yu (2022) Alcohol consumption us, Kaggle. Available at: https://www.kaggle.com/datasets/linzey/alcohol-consumption-us (Accessed: 05 December 2024).

U.S. Department of Health & Human Services - impaired driving death rate, by age and Gender, 2012 & 2014, all States (2021) Catalog. Available at: https://catalog.data.gov/dataset/impaired-driving-death-rate-by-age-and-gender-2012-2014-all-states (Accessed: 05 December 2024).

Wong, J. (2022) Average monthly temperature by US state, Kaggle. Available at: https://www.kaggle.com/datasets/justinrwong/average-monthly-temperature-by-us-state (Accessed: 05 December 2024)