Alcohol Consumption and its Effects

Venkata Sarath Chelikani / vchelika,

Rohan Shukla / roshukla,

Madhav Jariwala / makejari

Table of Contents

TABLE OF FIGURES

ABSTRACT

The project aims to study alcohol consumption trends and their effect on the world. We also aim to investigate various social and economic factors which play a key role in the consumption of alcohol. Furthermore, we would like to analyze the data of after-effects of alcohol consumption. This includes health-related alcohol data, as well as other data analysis such as crime rates due to alcohol consumption, happiness index, and standard of living. Through this project, we would like to first check some of these trends on a global, country-wise scale, and then specifically focus on the data from the United States.

INTRODUCTION

Motivation

Consumption of alcohol can have adverse effects. Although moderate alcohol consumption can have sustaining and positive effects on the body, it is still a major cause of death and disability in the world [1]. Among the crimes happening, one-fourth of them occur under the influence of alcohol [2]. Furthermore, it has also been found that alcohol influence is more likely to be possible than any other drug-related substance abuse incident [2]. Thus, through this project, our aim is to visualize these crime-related statistics and show the reality in a much impactful way to make the audience aware. On the other hand, alcohol consumption is not always considered to have harmful effects. There is an inverse correlation between the liters of alcohol consumed and the mean average temperature at a place [3]. We know that alcohol acts as a vasodilator, to increase warmth, and it could give the feeling of warmth to people living in colder areas such as Siberia and Wisconsin [3]. Also, moderate, and controlled drinking of alcohol lowers the risk of chronic heart diseases [5]. Hence, we want to analyze the above-mentioned scenarios worldwide as well as in the United States exclusively and portray our findings through effective visualizations.

Background

Alcohol use is the 7th leading cause of death and disability in the world [1]. Almost one in four out of the 11.1 million crimes happening each year have consumed alcohol prior to committing a crime [2]. Alcohol is also termed to be related as a casual cause for nearly 60 illnesses [4]. The practice of fermenting sugary and starchy plants to make beer and wines have been happening since the prehistoric time [7]. These people used alcohol as a means to relieve fatigue and pain, as well as promote friendship and bravery. The current western culture is a mixture of inheritance from Greek, Roman, and Hebrew culture. People have been against alcoholism since the nineteenth century. The United States had also passed public sanctions, but they were revoked in the Great Depression [6]. Now, in the twentieth century, there has been a great increase in the treatment of alcoholism, with the rise of health and rehabilitation centres [6].

Objective and Contribution

Our aim is to analyze alcohol consumption and its effects in different countries. Each country differs in the employment rate, happiness index, crime rates, etc. Our project aims to analyze these trends, and create effective visualizations to answer these questions. Furthermore, we also aim to target the United States. We would like to analyze alcohol consumption with factors such as health index, poverty, climate changes, etc.

Existing Visualizations

- Visualization by The Economist: dataisbeautiful
    - This visualization portrays the number of drinkers and abstainers from each country, and by how much margin the consumption of drinkers varies from the average drinking of an adult. Countries such as France, Britain, and Canada have populations in which most of the adults are drinkers, whereas countries such as UAE and Chad are the ones which have the most number of abstainers, hence a more difference in the average consumption between regular drinkers and adults.

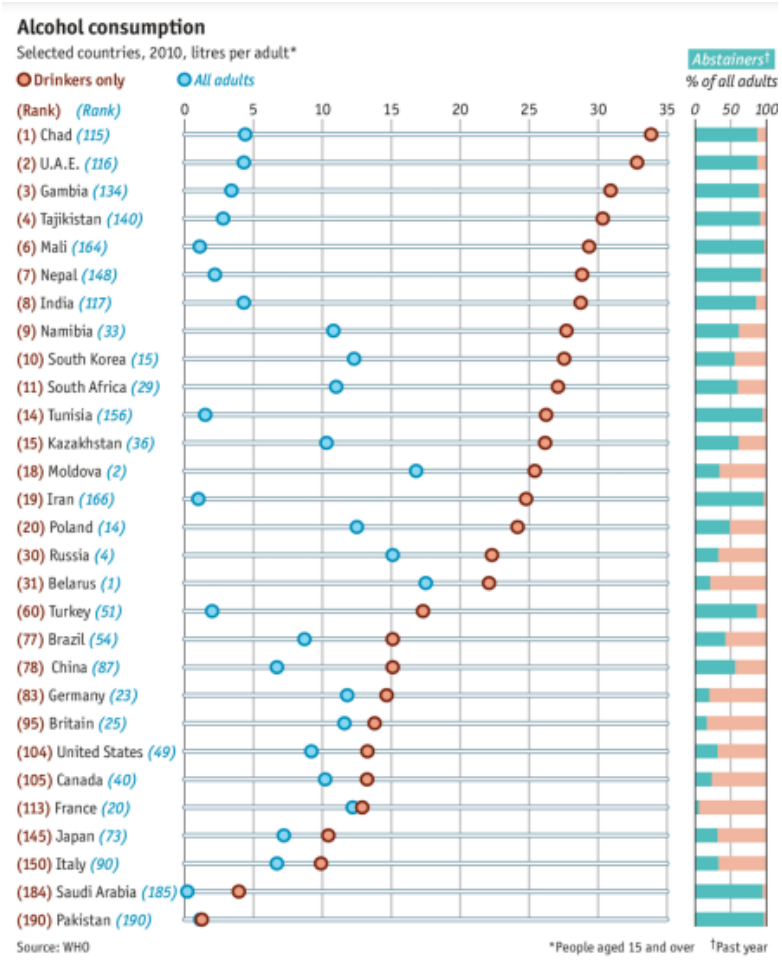

Figure 1: Existing Visualization of Alcohol Consumption Country-wise

- A look into alcohol consumption by Hannah Yan Han
    - This visualization showcases the types of alcohol consumed in different parts of the world. We notice that Asians usually prefer wine and spirits, while the rest of the world has Beer as the most popular drink. Although this visualization is insightful, using a bar chart might have been more effective.

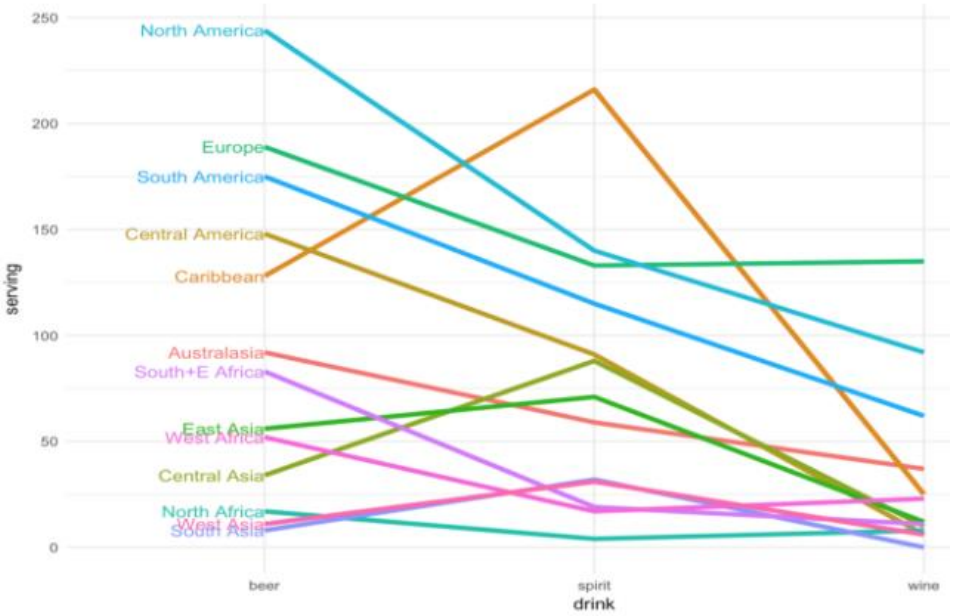

Figure 2: Existing Visualization Parallel Coordinates

- Visualizing World Alcohol Consumption: DataViz
    - This visualization is an excellent representation of the consumption of alcohol in the world, in a graph, which shows more than just the shades. One can easily note that once we move away from the equator, the amount of alcohol consumption by each person also increases. Also, the visual is clean and accurately categorized.

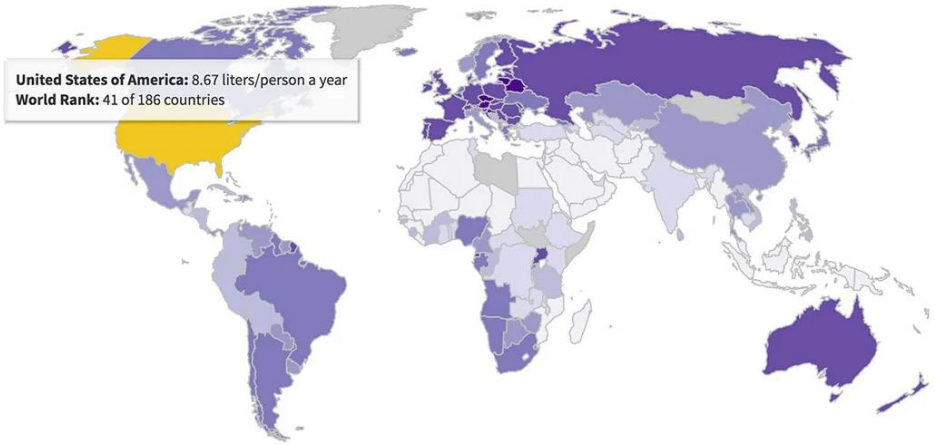

Figure 3: Existing Visualization Choropleth

PROCESS

Analysis of Data

The data of alcohol consumption was mainly taken from the World Health Organization repository. The Global Health Observatory was used to fetch these data. The National Highway Traffic Safety Administration data was also used for analysis of effects of alcohol through traffic accidents.

Candidate Visualization Methods

The candidate visualization methods were mostly map visualizations. These included using Choropleth modules, GeoScatter modules, Altair, and other map charts. Along with these, box plots were used to analyse the data and its distribution. It helped in outlier detection. Beeswarm plot was also used for analysing distribution of attributes. Furthermore, various bar charts and radial charts were considered to analyse the time data.

Failed Experiments

Many map chart visuals failed since they were lacking clarity and effectiveness. Also, bar charts failed to properly encapsulate the time analysis of alcohol driving accidents.

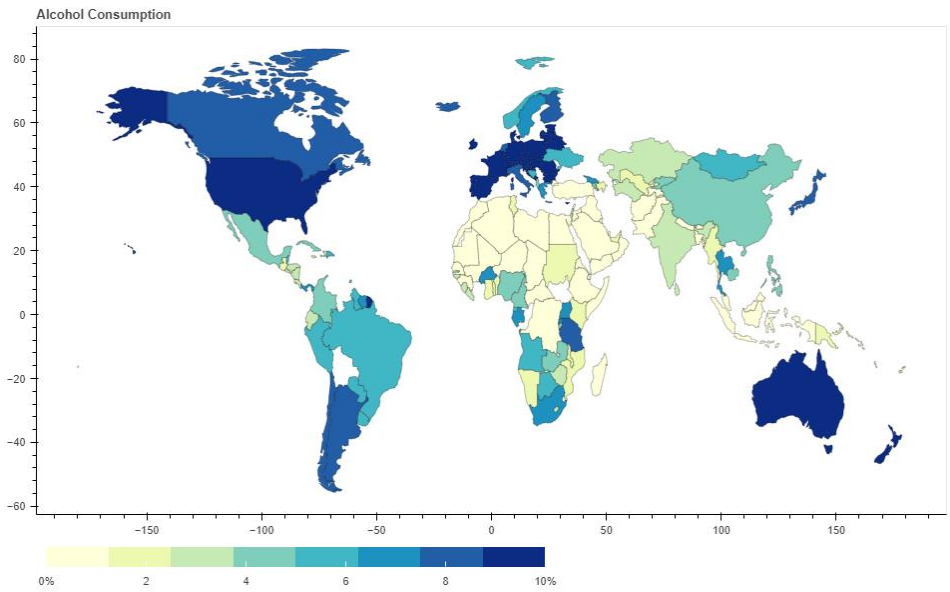

Figure 4: Failed Visualization - Choropleth Map

This is one of the examples of failed maps. It is a non-interactive chart. Although the colour variants are very effective, one cannot hover over it and get the details.

Why some methods worked but some didn't

Bar charts worked for visualizing the categories, where there were distinct ones. But, when we needed to plot hourly accidents and its occurrence, bar chart could not justify the periodicity. Hence, due to this reason, radial line chart was used to show the time data. One such example is shown in figure 5, where the bar chart isn't effective.

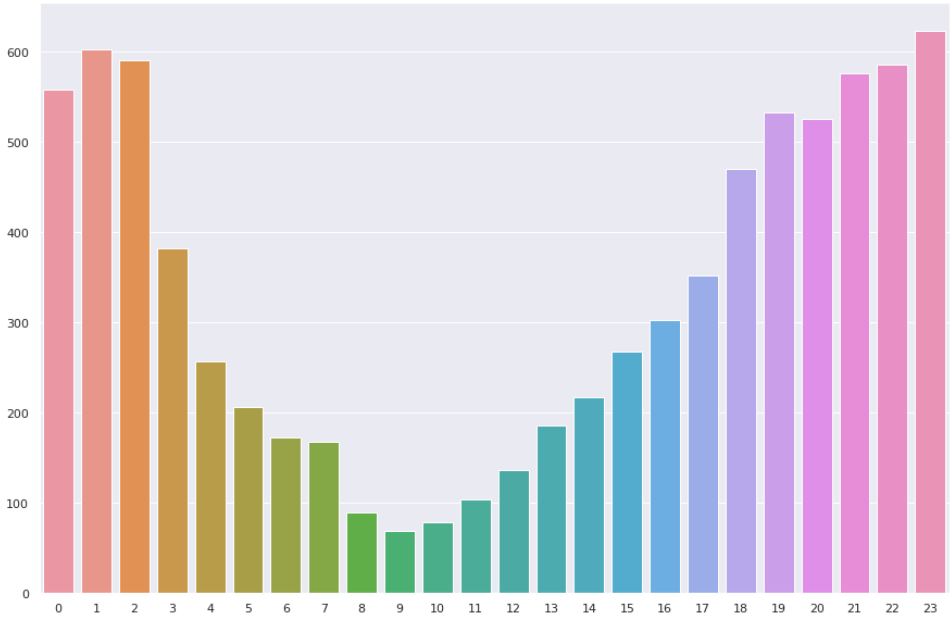

Figure 5: Why some method didn't work out? - Bar Charts

Instead of this, if we use radial chart, the periodicity is shown in figure 6.

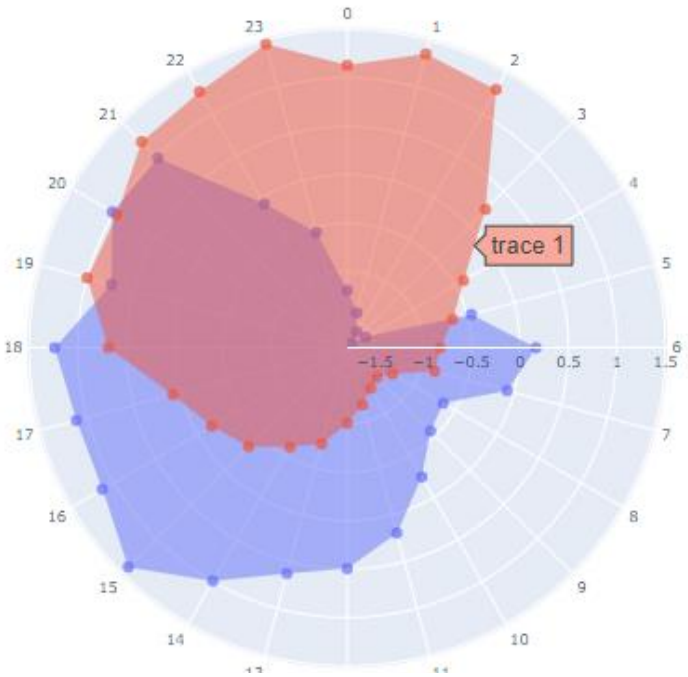

Figure 6: Why methods didn't work out? - Radial Chart

Another method we tried was using choropleth map to show alcohol consumption in U.S. states and a lot of states were in proximity and we couldn't get desired output.

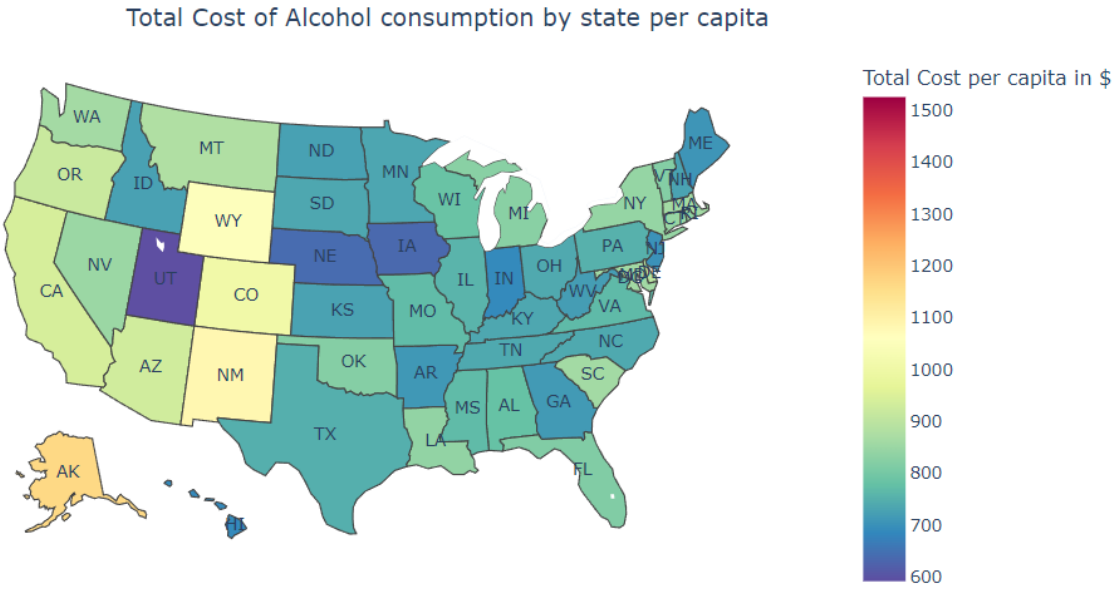

Figure 7: Why methods didn't work out? - Choropleth

Here, in figure 7, the highest cost of alcohol consumption by state per capita is by District of Columbia but because the region of District of Columbia is so small, we failed to show the result in an effective manner. Thus, we used bubble maps instead whose marker size depends on the value it receives.

Datasets

- Link: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/liver-cirrhosis-age-standardized-death-rates-(15-)-per-100-000-population

```
RangeIndex: 388 entries, 0 to 387
Data columns (total 34 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   IndicatorCode               388 non-null    object
 1   Indicator                   388 non-null    object
 2   ValueType                   388 non-null    object
 3   ParentLocationCode          388 non-null    object
 4   ParentLocation              388 non-null    object
 5   Location type               388 non-null    object
 6   SpatialDimValueCode         388 non-null    object
 7   Location                    388 non-null    object
 8   Period type                 388 non-null    object
 9   Period                      388 non-null    int64
 10  IsLatestYear                388 non-null    bool
 11  Dim1 type                   388 non-null    object
 12  Dim1                        388 non-null    object
 13  Dim1ValueCode               388 non-null    object
 14  Dim2 type                   0 non-null      float64
 15  Dim2                        0 non-null      float64
 16  Dim2ValueCode               0 non-null      float64
 17  Dim3 type                   0 non-null      float64
 18  Dim3                        0 non-null      float64
 19  Dim3ValueCode               0 non-null      float64
 20  DataSourceDimValueCode      0 non-null      float64
 21  DataSource                  0 non-null      float64
 22  FactValueNumericPrefix      0 non-null      float64
 23  FactValueNumeric            366 non-null    float64
 24  FactValueUoM                0 non-null      float64
 25  FactValueNumericLowPrefix   0 non-null      float64
 26  FactValueNumericLow         0 non-null      float64
 27  FactValueNumericHighPrefix  0 non-null      float64
 28  FactValueNumericHigh        0 non-null      float64
```

Figure 8: Datasets – 1

This dataset in figure 8 is from WHO, it consists of road traffic crash deaths occurred due to driving while under the influence of alcohol per 100,000 people (15+). This dataset has data gathered from different countries across the world. Here, the data is distributed between male and female, but we applied groupby on all the countries to get one sample from each country. The following is a brief description of the dataset before any operations were performed:

Dataset after groupby and other operations were performed:

country_code	country	road_traffic_deaths(per_100,000)
BIH	Bosnia and Herzegovina	32.6
RWA	Rwanda	115.9
SYC	Seychelles	18.6
CAN	Canada	13.2
CHE	Switzerland	6.0
VEN	Venezuela (Bolivarian Republic of)	106.5
IRL	Ireland	7.9
KOR	Republic of Korea	21.1
EGY	Egypt	41.0
SEN	Senegal	103.8

Figure 9: Datasets - 1 - groupby

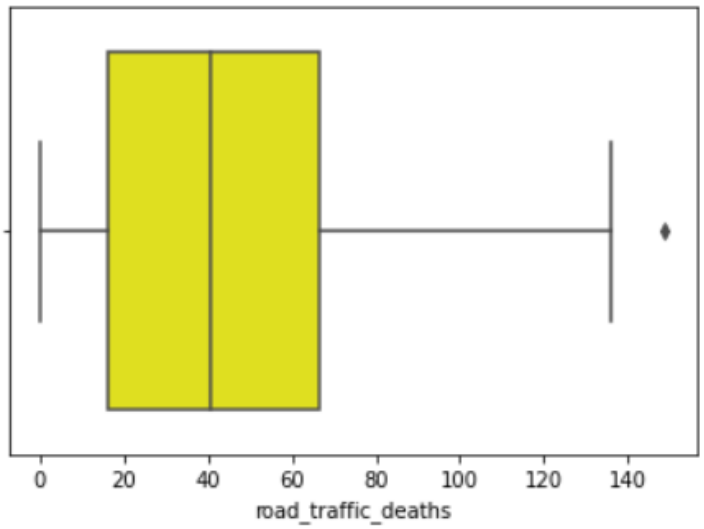

Figure 10: Datasets - 1 - Boxplot

Here, we can see that there is an outlier with a value of 148.8 road traffic deaths due to consumption of alcohol per 100,000 people. The outlier value belongs to Zimbabwe. The following graph in figure 11 is a beeswarm plot representing the same column as the above box plot, to get a better understanding of the data.

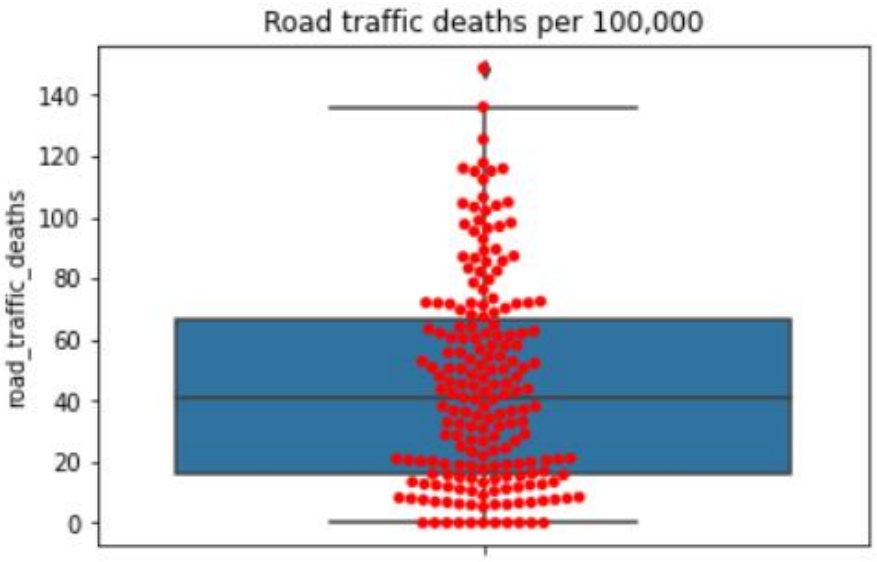

Figure 11: Datasets - 1 - Beeswarm

- Link: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/liver-cirrhosis-age-standardized-death-rates-(15-)-per-100-000-population

  This dataset is taken from WHO, it consists of deaths occurred from liver cirrhosis due to alcohol consumption per 100,000 people (15+). The data in this dataset is distributed across several countries. Operations like groupby were performed to get the data in a suitable format.

  The following is a brief description of the dataset before any operations were performed:

#	Column	Non-Null Count	Dtype
0	IndicatorCode	388 non-null	object
1	Indicator	388 non-null	object
2	ValueType	388 non-null	object
3	ParentLocationCode	388 non-null	object
4	ParentLocation	388 non-null	object
5	Location type	388 non-null	object
6	SpatialDimValueCode	388 non-null	object
7	Location	388 non-null	object
8	Period type	388 non-null	object
9	Period	388 non-null	int64
10	IsLatestYear	388 non-null	bool
11	Dim1 type	388 non-null	object
12	Dim1	388 non-null	object
13	Dim1ValueCode	388 non-null	object
14	Dim2 type	0 non-null	float64
15	Dim2	0 non-null	float64
16	Dim2ValueCode	0 non-null	float64
17	Dim3 type	0 non-null	float64
18	Dim3	0 non-null	float64
19	Dim3ValueCode	0 non-null	float64
20	DataSourceDimValueCode	0 non-null	float64
21	DataSource	0 non-null	float64
22	FactValueNumericPrefix	0 non-null	float64
23	FactValueNumeric	366 non-null	float64
24	FactValueUoM	0 non-null	float64
25	FactValueNumericLowPrefix	0 non-null	float64
26	FactValueNumericLow	0 non-null	float64
27	FactValueNumericHighPrefix	0 non-null	float64
28	FactValueNumericHigh	0 non-null	float64
29	Value	388 non-null	object
30	FactValueTranslationID	22 non-null	float64
31	FactComments	0 non-null	float64

Figure 12: Datasets - 2 - Info

Dataset after groupby and other operations were performed:

country_code	country	liver_cirrhosis_deaths
IDN	Indonesia	31.7
SSD	South Sudan	79.5
BTN	Bhutan	28.1
KNA	Saint Kitts and Nevis	0.0
GRC	Greece	131.0
KAZ	Kazakhstan	119.1
MNG	Mongolia	104.9
SWE	Sweden	126.7
IND	India	93.3
MHL	Marshall Islands	0.0

Figure 13: Datasets - 2 - Groupby

It can be clearly seen from the box plot in figure 14 and beeswarm plot in figure 15 from below that there are no outliers in the data.

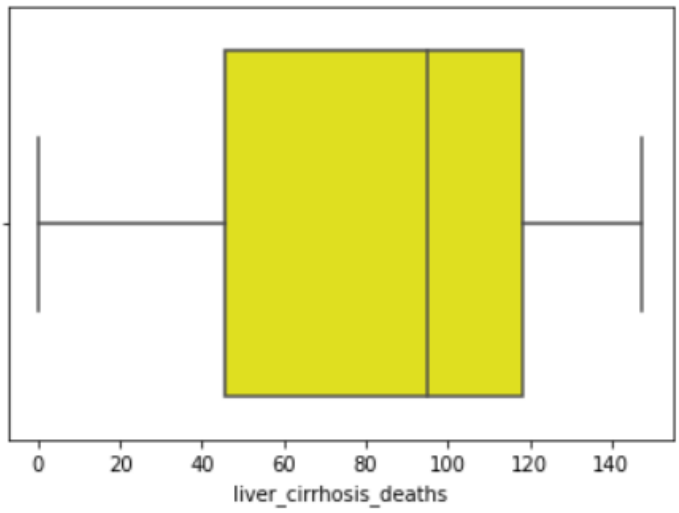

Figure 14: Datasets - 2 - Boxplot

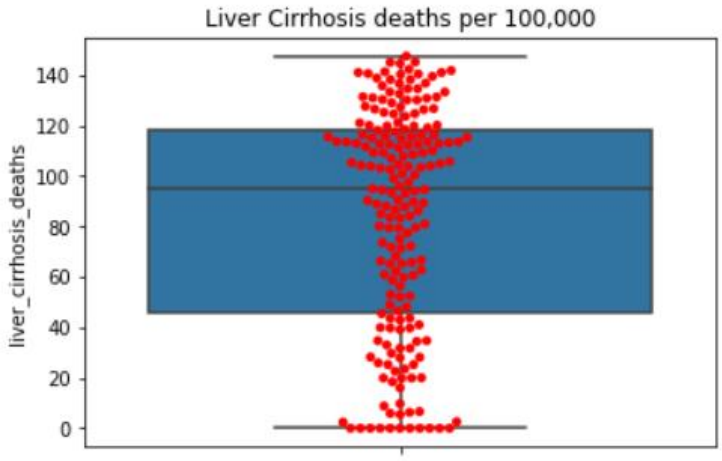

Figure 15: Datasets - 2 - Beeswarm

- Link:https://www.who.int/data/gho/data/indicators/indicator-details/GHO/alcohol-attributable-years-of-life-lost-(yll)-score

  This dataset is taken from WHO, it consists of a score of number of years lost due to consumption of alcohol, which is in a scale of ranging from 1 to 5. The score is covered across different countries. A groupby operation is performed to prepare the data in a suitable format. Figure 16 is a brief description of the dataset before any operations were performed:

#	Column	Non-Null Count	Dtype
0	IndicatorCode	194 non-null	object
1	Indicator	194 non-null	object
2	ValueType	194 non-null	object
3	ParentLocationCode	194 non-null	object
4	ParentLocation	194 non-null	object
5	Location type	194 non-null	object
6	SpatialDimValueCode	194 non-null	object
7	Location	194 non-null	object
8	Period type	194 non-null	object
9	Period	194 non-null	int64
10	IsLatestYear	194 non-null	bool
11	Dim1 type	0 non-null	float64
12	Dim1	0 non-null	float64
13	Dim1ValueCode	0 non-null	float64
14	Dim2 type	0 non-null	float64
15	Dim2	0 non-null	float64
16	Dim2ValueCode	0 non-null	float64
17	Dim3 type	0 non-null	float64
18	Dim3	0 non-null	float64
19	Dim3ValueCode	0 non-null	float64
20	DataSourceDimValueCode	0 non-null	float64
21	DataSource	0 non-null	float64
22	FactValueNumericPrefix	0 non-null	float64
23	FactValueNumeric	183 non-null	float64
24	FactValueUoM	0 non-null	float64
25	FactValueNumericLowPrefix	0 non-null	float64
26	FactValueNumericLow	0 non-null	float64
27	FactValueNumericHighPrefix	0 non-null	float64
28	FactValueNumericHigh	0 non-null	float64
29	Value	194 non-null	object
30	FactValueTranslationID	11 non-null	float64
31	FactComments	0 non-null	float64
32	Language	194 non-null	object
33	DateModified	194 non-null	object

Figure 16: Datasets - 3 - Info

Dataset after groupby and other operations were performed:

country_code	country	years_lost_score
GRD	Grenada	3.0
ZWE	Zimbabwe	4.0
ESP	Spain	2.0
CAN	Canada	2.0
CRI	Costa Rica	2.0
KHM	Cambodia	5.0
TTO	Trinidad and Tobago	4.0
CZE	Czechia	3.0
TUN	Tunisia	1.0
SOM	Somalia	1.0

Figure 17: Datasets - 3 - Groupby

It can be clearly seen from the box plot in figure 18 and beeswarm plot in figure 19 from below that there are no outliers in the data.

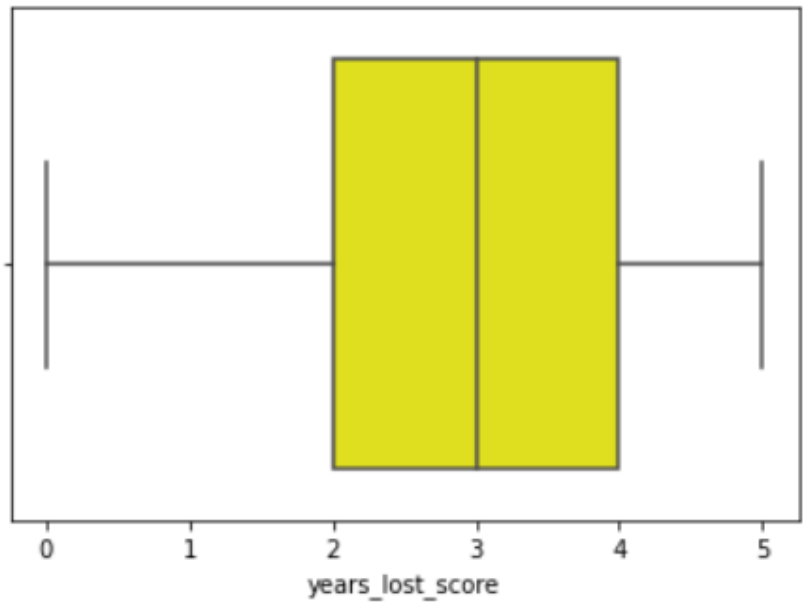

Figure 18: Datasets - 3 - Boxplot

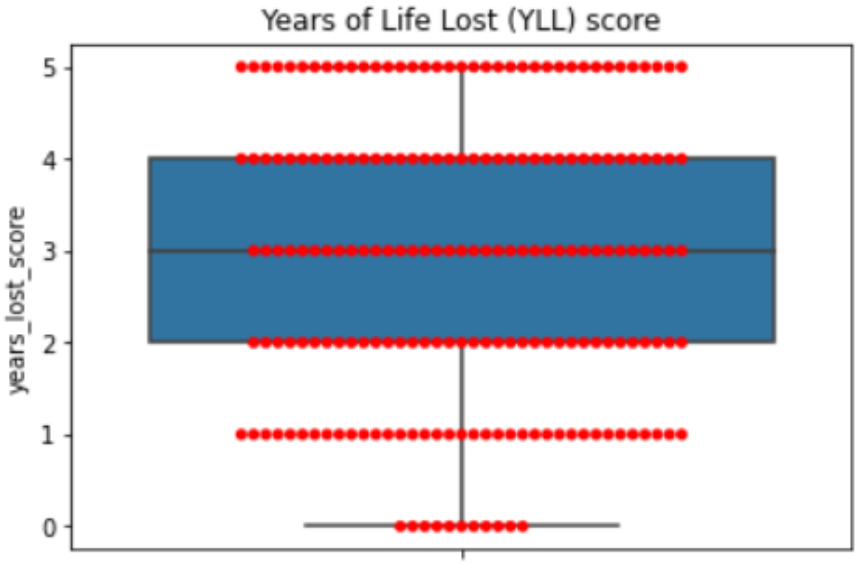

Figure 19: Datasets - 3 - Beeswarm plot

- Link: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/alcohol-recorded-per-capita-(15-)-consumption-(in-litres-of-pure-alcohol)

	Country	All types	Beer	Other alcoholic beverages	Spirits	Wine
0	Afghanistan	0.01	0.00	0.00	0.01	0.00
1	Albania	4.40	1.75	0.08	1.43	1.15
2	Algeria	0.59	0.31	0.00	0.08	0.20
3	Andorra	10.99	3.59	0.00	2.32	4.98
4	Angola	5.84	3.78	0.08	1.27	0.72
5	Antigua and Barbuda	11.88	2.97	0.41	4.55	3.95
6	Argentina	7.95	3.62	0.72	0.72	2.88
7	Armenia	3.77	0.52	0.01	2.78	0.46
8	Australia	9.51	3.71	0.81	1.32	3.67
9	Austria	11.90	6.30	0.00	1.90	3.70

Figure 20: Datasets - 4 – 10 records

Figure 20 is a dataset from WHO's global health observatory report, which provides the gala of the alcohol consumption per capita of people over 15+ years of age, in liters of pure alcohol. The dataset is showing the per capita consumption of alcohol of all the different types of alcohol consumed namely, spirits, beer, wine and other types. The main attribute is the 'All types' attribute, which will be used to analyse the world consumption.

We use box plot to analyse this one dimensional data.

```
sns.boxplot(data=alc_cons_who,x='All types',color='yellow')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f832547e590>
```

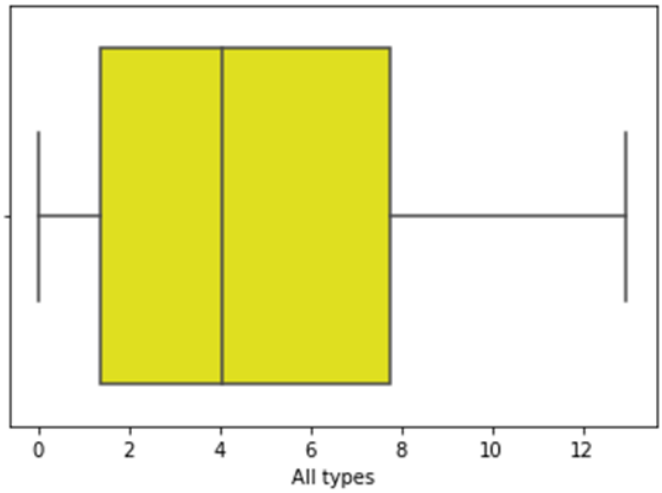

Figure 21: Datasets - 4 - Boxplot

As we can see in figure 21, there are no outliers in the attribute, and it is ready to be used in complex visualizations without causing any bias in the visualization. Next, we use the bee-swarm plot and the box plot to visualize the other attributes which are beer, spirits and wine.

```
sns.boxplot(y=alc_cons_who['Wine'],ax=axes[2]).set_title('Wine consumption per Capita')
sns.swarmplot(y=alc_cons_who['Wine'],ax=axes[2],color='red')
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/categorical.py:1296: UserWarning: 22.6% of the points cannot be placed; you
  warnings.warn(msg, UserWarning)
<matplotlib.axes._subplots.AxesSubplot at 0x7f8324ef4790>
```

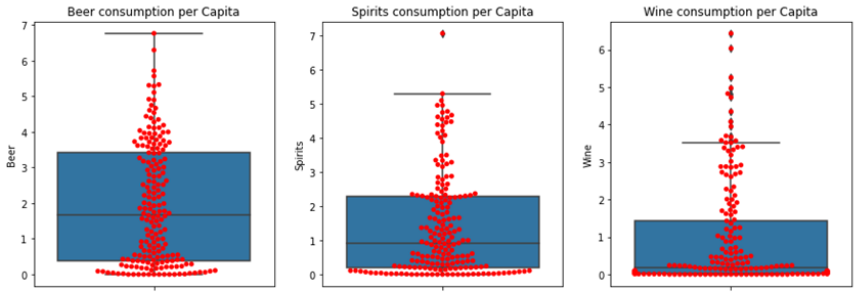

Figure 22: Datasets - 4 - Beeswarm

Also, we use the “geopandas.shp” file and join it with the countries in the WHO alcohol consumption dataset. Once we do this, the data integrity is also maintained, and the dataset is complete to be used on the various map plots.

```
[51] merged.head()
```

	country	country_code	geometry	Country	All types	Beer	Other alcoholic beverages	Spirits	Wine
0	Fiji	FJI	MULTIPOLYGON (((180.00000 -16.06713, 180.00000...	Fiji	2.71	1.64	0.0	0.79	0.29
1	United Republic of Tanzania	TZA	POLYGON ((33.90371 -0.95000, 34.07262 -1.05982...	United Republic of Tanzania	7.81	0.74	6.6	0.38	0.09
2	Canada	CAN	MULTIPOLYGON (((-122.84000 49.00000, -122.9742...	Canada	8.00	3.50	0.4	2.10	2.00
3	United States of America	USA	MULTIPOLYGON (((-122.84000 49.00000, -120.0000...	United States of America	8.93	3.97	0.0	3.29	1.67

Figure 23: Datasets - 4 - First 5 rows

- Link: https://www.nhtsa.gov/file-downloads?p=nhtsa/downloads/FARS/2019/National/

The dataset is the record of all the accidents which occurred in the United States. The dataset has the attribute, 'Drunk_Dr', which tells whether the accident was occurred under the influence of alcohol or not.
The dataset seen in figure 24 has 33244 instances with 90 attributes.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 33244 entries, 0 to 33243
Data columns (total 91 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   STATE         33244 non-null  int64
 1   STATENAME     33244 non-null  object
 2   ST_CASE       33244 non-null  int64
 3   VE_TOTAL      33244 non-null  int64
 4   VE_FORMS      33244 non-null  int64
 5   PVH_INVL      33244 non-null  int64
 6   PEDS          33244 non-null  int64
 7   PERSONS       33244 non-null  int64
 8   PERMVIT       33244 non-null  int64
 9   PERNOTMVIT    33244 non-null  int64
 10  COUNTY        33244 non-null  int64
 11  COUNTYNAME    33244 non-null  object
 12  CITY          33244 non-null  int64
 13  CITYNAME      33244 non-null  object
 14  DAY           33244 non-null  int64
 15  DAYNAME       33244 non-null  int64
 16  MONTH         33244 non-null  int64
 17  MONTHNAME     33244 non-null  object
 18  YEAR          33244 non-null  int64
 19  DAY_WEEK      33244 non-null  int64
 20  DAY_WEEKNAME  33244 non-null  object
```

Figure 24: Datasets - 5 - Info

Since, we don't need so many attributes to analyse or min the data, we removed most of the rows and kept the 10 important attributes to analyse the dataset.

```
accid_df.head(29)
```

	STATE	STATENAME	MONTH	MONTHNAME	HOUR	ROUTENAME	LATITUD	LONGITUD	WEATHER1NAME	DRUNK_DR
0	1	Alabama	2	February	12	Interstate	32.666222	-85.336658	Clear	1
1	1	Alabama	1	January	18	Interstate	33.997828	-86.053997	Rain	0
2	1	Alabama	1	January	19	Interstate	33.660842	-85.391011	Cloudy	0
3	1	Alabama	1	January	3	Interstate	33.956472	-86.140522	Clear	0
4	1	Alabama	1	January	5	Interstate	30.656269	-87.809461	Fog, Smog, Smoke	1
5	1	Alabama	1	January	12	Interstate	32.183306	-86.424683	Clear	0
6	1	Alabama	1	January	9	Interstate	34.012775	-86.074814	Cloudy	0
7	1	Alabama	2	February	21	County Road	31.068686	-85.333658	Clear	1
8	1	Alabama	2	February	7	Local Street - Municipality	30.279125	-87.677108	Clear	0
9	1	Alabama	2	February	5	Local Street - Municipality	32.344300	-86.317081	Clear	0
10	1	Alabama	1	January	1	State Highway	33.104300	-85.773344	Rain	1
11	1	Alabama	1	January	18	County Road	34.026453	-85.955872	Rain	0

Figure 25: Datasets - 5 - First few Rows

We check for outliers in the target attribute, 'Drunk_Dr', using boxplot.

```
sns.boxplot(data=accid_df,x='DRUNK_DR',color='yellow')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8310fc3510>
```

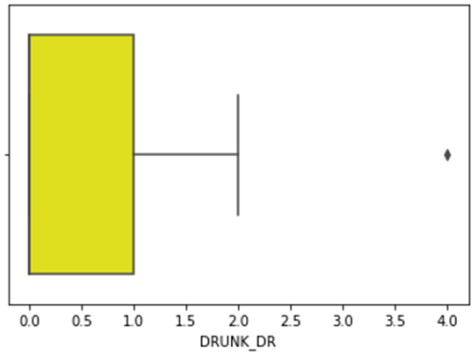

Figure 26: Datasets - 5 - Boxplot

Here, we notice that there are some data noises as values of 2 and 4, where actually it should have only zeroes and ones.

Thus, we count these, and since they are very small in number when compared to the whole dataset, we remove them.

```
accid_df['DRUNK_DR'].value_counts()
```

```
0    24833
1     8161
2      249
4        1
Name: DRUNK_DR, dtype: int64
```

Since there are very less number of unknown data, we can consider it as outliers and drop from the data

```
[23] accid_df.drop(accid_df.loc[accid_df['DRUNK_DR']==4].index, inplace=True)
     accid_df.drop(accid_df.loc[accid_df['DRUNK_DR']==2].index, inplace=True)
```

Figure 27: Datasets - 5 - Pre-processing

Now, on plotting the graph, we see that the data has only zeroes and ones.

```
sns.histplot(data=accid_df,x='DRUNK_DR',color='red')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8310f5f7d0>
```

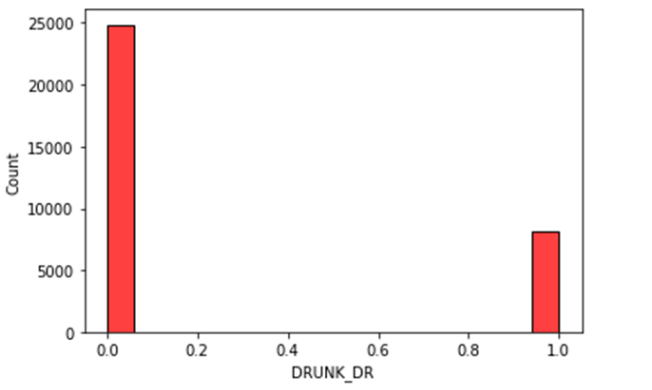

Figure 28: Datasets - 5 – Hist plot

Also, we group by this data on the countries, for some of the visual analysis, so that the visuals can be plotted.
Also, since the population varies from each state, the numbers aren't the actual representation of the scenario. Hence, we create another attribute which shows the percentage of drunk driving accidents. This makes the analysis fair for all the states.

```
[25] y1 = accid_df[accid_df['DRUNK_DR']==0].groupby('STATENAME').size()
     y2 = accid_df[accid_df['DRUNK_DR']==1].groupby('STATENAME').size()
     x = accid_df['STATENAME'].unique()
     sn = accid_df['STATE'].unique()
```

Figure 29: Datasets - 5 - Pre-processing 2

```
[27] df_piv['perc_drunk'] = df_piv['drunk']/(df_piv['drunk']+df_piv['not_drunk'])
     df_piv.head()
```

	STATENAME	STATE	not_drunk	drunk	perc_drunk
0	Alabama	1	631	220	0.258519
1	Alaska	2	40	20	0.333333
2	Arizona	4	714	190	0.210177
3	Arkansas	5	345	120	0.258065
4	California	6	2493	802	0.243399

Figure 30: Datasets - 5 - Pre-processing 3

- Link: https://www.cdc.gov/alcohol/data-stats.htm

The dataset is taken from CDC (Centre for Disease Control and Prevention). The dataset includes state-wise Excessive alcohol consumption by State. It has Total cost to states because of excessive drinking, Cost to states per drink, and Cost per Capita. Excessive alcohol use/binge drinking is defined as consuming 4 or more drinks for women and 5 or more for men.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51 entries, 0 to 50
Data columns (total 4 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Location             51 non-null     object
 1   Total Cost ($)       51 non-null     int64
 2   Cost per drink ($)   51 non-null     float64
 3   Cost per capita ($)  51 non-null     int64
dtypes: float64(1), int64(2), object(1)
memory usage: 1.7+ KB
```

Figure 31: Datasets - 6 - Info

	Location	Total Cost ($)	Cost per drink ($)	Cost per capita ($)
0	Alabama	3724300000	2.27	779
1	Alaska	827200000	2.25	1165
2	Arizona	5946400000	2.27	930
3	Arkansas	2073300000	2.27	711
4	California	35010600000	2.44	940

Figure 33: Datasets - 6 - Rows

Location is U.S. states names. The dataset is already cleaned so we don't have to clean the data. But, in order to plot this data on a choropleth map, we need to add another column with ISO codes for states. Also, we need to convert the Total Cost into Billions because that would be easier to work with. So eventually we get,

	Location	Total Cost ($)	Cost per drink ($)	Cost per capita ($)	code
0	Alabama	3.7243	2.27	779	AL
1	Alaska	0.8272	2.25	1165	AK
2	Arizona	5.9464	2.27	930	AZ
3	Arkansas	2.0733	2.27	711	AR
4	California	35.0106	2.44	940	CA

Figure 32: Datasets - 6 - Pre-processing

So, this is the dataset that we would be using for visualization.

- Link: https://www.newsweek.com/usa-states-ranked-happiness-hawaii-1533784

This dataset contains state-wise happiness score which includes "Emotional & Physical Well-being" rank, "Work Environment" rank, "Community & Environment" rank, and total happiness score.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 50 entries, 0 to 49
Data columns (total 8 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Overall Rank (1 = Happiest)           50 non-null     int64
 1   State                                 50 non-null     object
 2   Total Score                           50 non-null     float64
 3   "Emotional & Physical Well-Being" Rank  50 non-null     int64
 4   "Work Environment" Rank               50 non-null     int64
 5   "Community & Environment" Rank        50 non-null     int64
 6   code                                  50 non-null     object
 7   alcoholConsumptionGallons             50 non-null     float64
dtypes: float64(2), int64(4), object(2)
memory usage: 3.5+ KB
```

Figure 34: Datasets - 7 - Info

We added the “code” column here, to plot the values on choropleth. We also added “alcoholConsumptionGallons” to compare these variables. Here, the dataset is already cleaned so we would go right into the visualization. So, this is what our dataset looks like,

	Overall Rank (1 = Happiest)	State	Total Score	"Emotional & Physical Well-Being" Rank	"Work Environment" Rank	"Community & Environment" Rank	code	alcoholConsumptionGallons
0	1	Hawaii	69.58	2	16	3	HI	2.66
1	2	Utah	69.42	14	1	1	UT	1.35
2	3	Minnesota	65.87	4	4	10	MN	2.79
3	4	New Jersey	64.10	1	30	22	NJ	2.36
4	5	Maryland	61.78	3	33	9	MD	2.08

Figure 35: Datasets - 7 - Rows

- Link: https://www.bls.gov/web/laus/laumstrk.htm

This dataset is taken from the U.S Bureau of Labour Statistics. It includes Unemployment rate and we’ve added “code” column and “alcoholConsumptionGallons” column to make further visualizations.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 51 entries, 0 to 50
Data columns (total 5 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   State                     51 non-null     object 
 1   October 2021(p)rate       51 non-null     float64
 2   Rank                      51 non-null     int64  
 3   code                      51 non-null     object 
 4   alcoholConsumptionGallons  50 non-null     float64
dtypes: float64(2), int64(1), object(2)
memory usage: 2.4+ KB
```

Figure 37: Datasets - 8 - Info

	State	October 2021(p)rate	Rank	code	alcoholConsumptionGallons
0	Nebraska	1.9	1	NE	2.16
1	Utah	2.2	2	UT	1.35
2	Oklahoma	2.7	3	OK	1.85
3	Idaho	2.8	4	ID	2.94
4	South Dakota	2.8	4	SD	2.87

Figure 36: Datasets - 8 - Rows

This dataset is cleaned, we only have one record missing in “alcoholConsumptionGallons” but that we don’t need to clean it to make visualizations.

- Link: https://en.wikipedia.org/wiki/Political_party_strength_in_U.S._states

  This dataset is taken from Wikipedia. It includes information on political parties ruling those states.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 2 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   State                     50 non-null     object
 1   2020 presidentialelection  50 non-null     object
dtypes: object(2)
memory usage: 928.0+ bytes
```

Figure 39: Datasets - 9 - Info

	State	2020 presidentialelection
0	Alabama	Republican
1	Alaska	Republican
2	Arizona	Democratic
3	Arkansas	Republican
4	California	Democratic

Figure 38: Dataset - 9 - Rows

Here, we can analyze that data is pretty cleaned, so we don't need to do any data cleaning.

- Link: https://www.federalreserve.gov/releases/z1/dataviz/household_debt/state/map/#year:2020

  This dataset was taken from the Federal Reserve System, and it includes the debt-to-income ratio, which means, higher the ratio, more debt-ridden the households are in these states.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51 entries, 0 to 50
Data columns (total 6 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Year                      51 non-null     int64
 1   Quarter                   51 non-null     int64
 2   State                     51 non-null     object
 3   Debt-to-Income Ratio Low  51 non-null     float64
 4   Debt-to-Income Ratio High  50 non-null     float64
 5   Debt_to_income            50 non-null     float64
dtypes: float64(3), int64(2), object(1)
memory usage: 2.5+ KB
```

Figure 40: Datasets - 10 - Info

This dataset has Year, Quarter, Debt-to-income ratio Low and High and one column we added was "Debt_to_income" ratio which is the average of high and low ratios. We opted to use 2019 data and all quarters combined to compare it with the Alcohol consumption data we had from 2019. So, our final data was,

	Year	Quarter	State	Debt-to-Income Ratio Low	Debt-to-Income Ratio High	Debt_to_income
0	2021	1	Alabama	1.41	1.51	1.460
1	2021	1	Alaska	1.75	1.89	1.820
2	2021	1	Arizona	1.75	1.89	1.820
3	2021	1	Arkansas	1.24	1.33	1.285
4	2021	1	California	1.62	1.75	1.685

Figure 41: Dataset - 10 - Rows

Again, we only have one data missing, but we didn't find it necessary to remove the rows with empty columns as they were still useful for us.

- Link: https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_poverty_rate

  This dataset was taken from Wikipedia. This dataset includes 2019 poverty rates, 2014 poverty rates and Supplemental Poverty Measure. This dataset included records for Puerto Rio and American Samoa, so we had to remove those rows as we wanted consistent data throughout the project. Apart from that, we didn't have to clean the data as much.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 54 entries, 0 to 53
Data columns (total 5 columns):
 #   Column                                                                Non-Null Count  Dtype
---  ------                                                                --------------  -----
 0   Rank                                                                  54 non-null     object
 1   State                                                                 54 non-null     object
 2   2019 Poverty rate(percent of persons in poverty)[note 2][7]           54 non-null     object
 3   2014 Poverty Rates (includes unrelated children)                      53 non-null     object
 4   Supplemental Poverty Measure (2017-2019 average) (Geographically Adjusted)  54 non-null     object
dtypes: object(5)
memory usage: 2.2+ KB
```

Figure 42: Datasets - 11 - Info

RESULTS AND INSIGHTS

Visualizations

The alcohol consumption levels are visualized at a worldwide scale using the choropleth map. First, we analyze the consumption of all types of alcohol in the world.

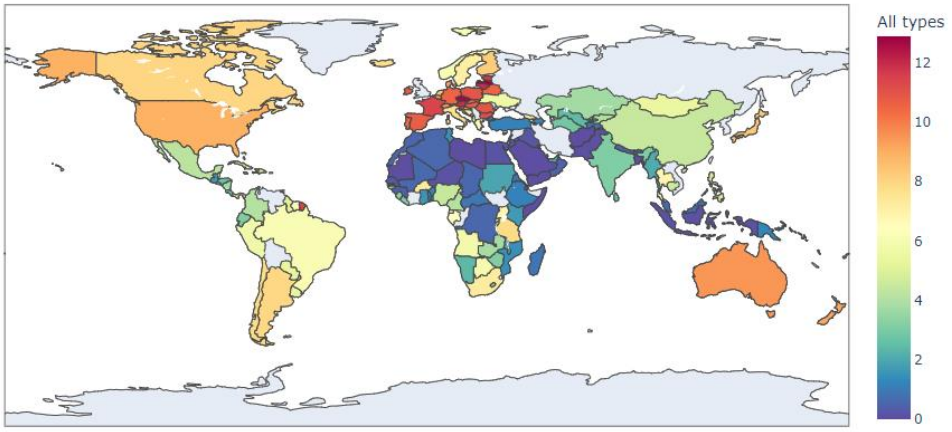

Figure 43: Choropleth of worldwide alcohol consumption (All types)

In this plot show in figure 43, we can see that Europe is the continent, where the alcohol consumption per capita is the highest. Also, places where the capita of the country is higher, the value decreases. It could be possible that the people of United States would be consuming a lot of alcohol, but due to a higher capita, the value would be shown as a lower one. Analysing alcohol consumption per capita, thus, gives a better and fair comparison.

Next, we analyze the alcohol consumption of beer on world-wide scale:

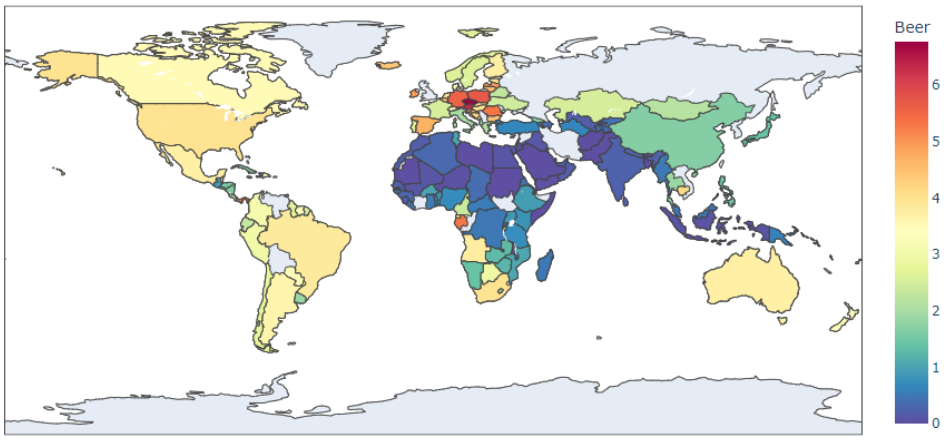

Figure 44: Choropleth - Beer consumption world-wide

Here, we can notice in figure 44 that Germany, which is famous for its breweries, consumes the highest beer. After this we plot the alcohol consumption of wine in all the countries:

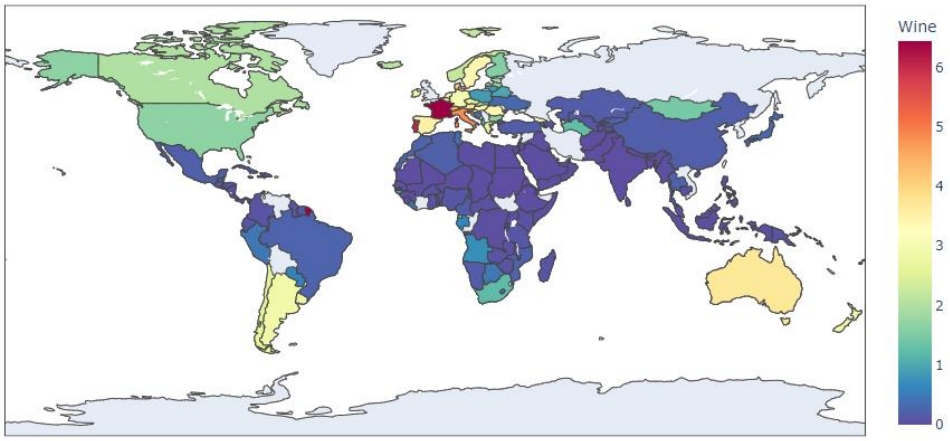

Figure 45: Choropleth - Wine Consumption

In this plot shows in figure 45, we observe Europe to be leading the wine consumption too. It goes along with the fact that most vineyards are present in France and nearby countries, where the consumption is also high.

Lastly, we plot the alcohol consumption of spirits in the world:

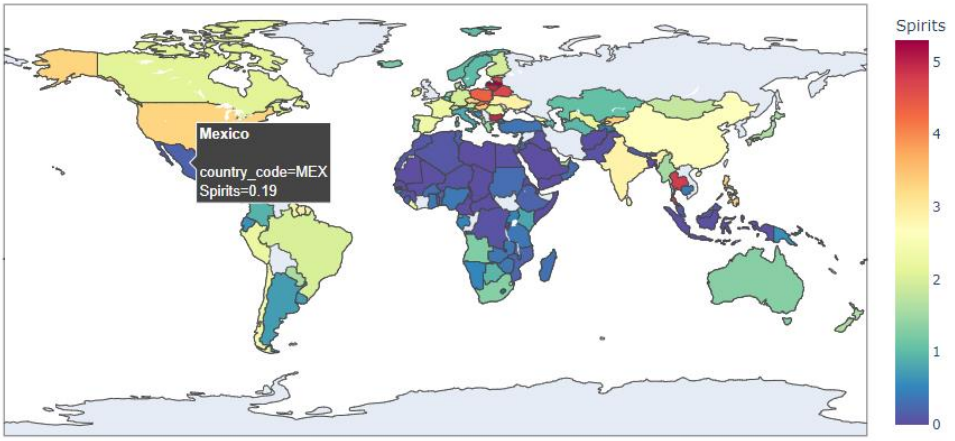

Figure 46: Choropleth - Spirit Consumption

Here in figure 46, we see the eastern Europe countries like Bulgaria, Hungary and Czech Republic to be leading in the spirits consumption.

The following plot in figure 47 represents the deaths that occurred by driving under the influence of alcohol per 100,000 people in various countries in the world. It can be seen that the death rate is higher at countries located in the African continent.

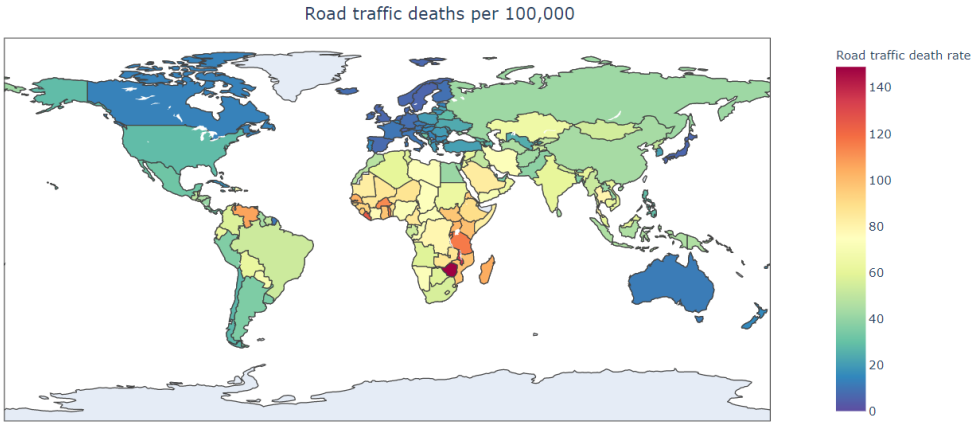

Figure 47: Choropleth of Road Traffic deaths per 1,00,000 population

The following plot in figure 48 represents the deaths from liver cirrhosis due to alcohol consumption per 100,000 people in different countries across the world. It can be seen from the graph that the rate of death by liver cirrhosis is higher at countries in the west.

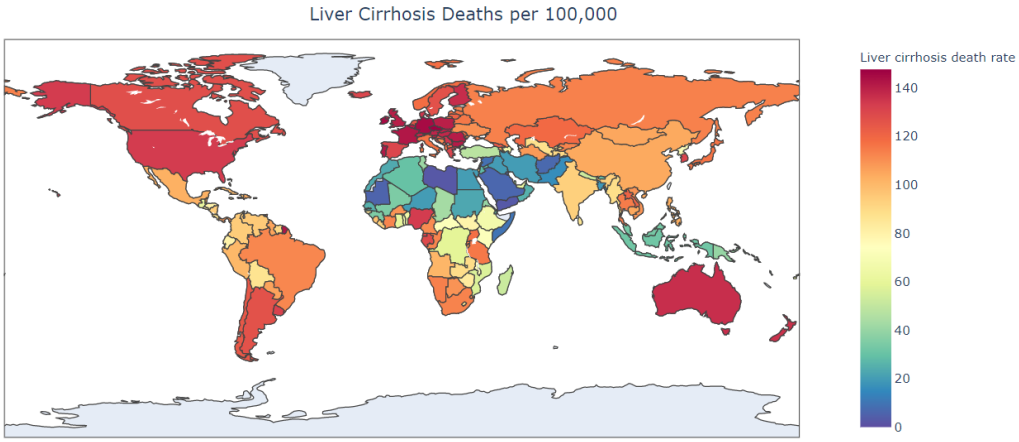

Figure 48: Choropleth - Liver Cirrhosis Deaths

The following plot shown in figure 49 represents a score which represents years lost due to consumption of alcohol across different countries in the world. It can be seen from the graph that the score is relatively higher in countries at Africa and Russia.

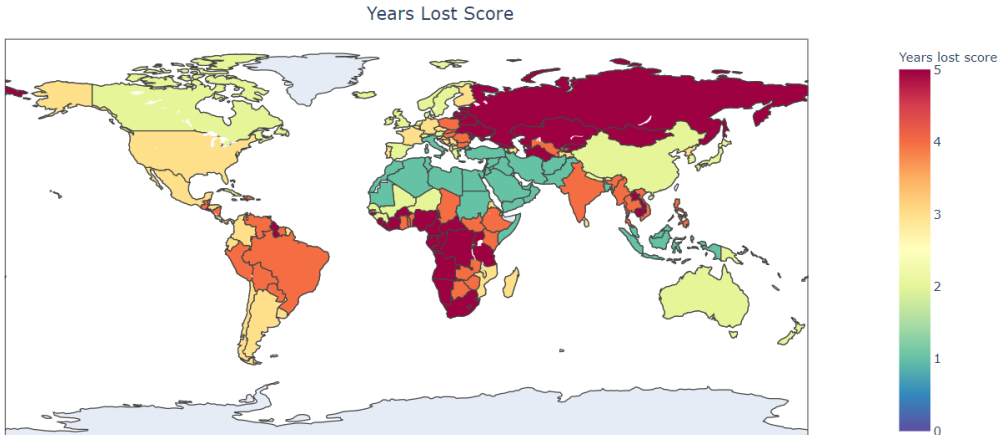

Figure 49: Choropleth - Lost years

We analyse the accidents caused in the United States under the influence of alcohol. To eliminate the population factor of each states, we calculate the percentage of accidents occurring in the state which are under the influence of alcohol, out of the total number of accidents.

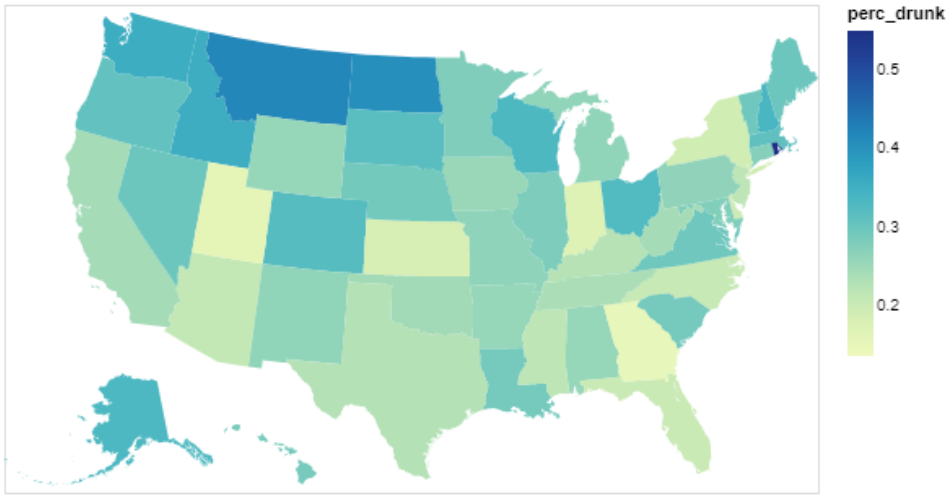

Figure 50: Choropleth - accident under influence of alcohol

We first start by observing the top-ten states in the US where the percentage of alcohol-influenced accidents are the highest.

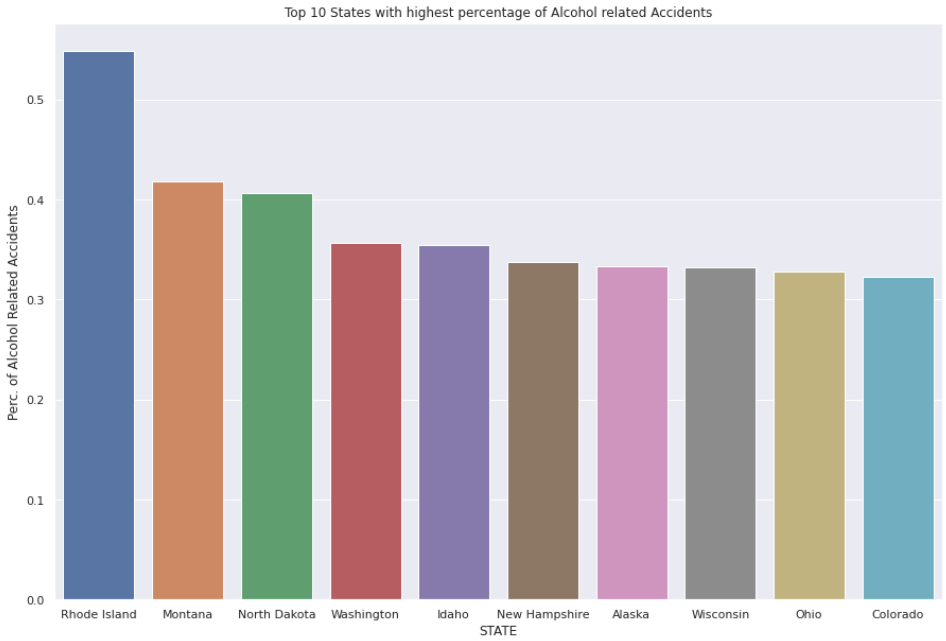

Figure 51: Bar plot - Top ten states with alcohol related accidents

This chart shows that most of the top states have a similar drunk driving accident percentage. Thus, we explore the other factors which influence drunk driving accidents.

We check the location of the accident occurring in the United States.

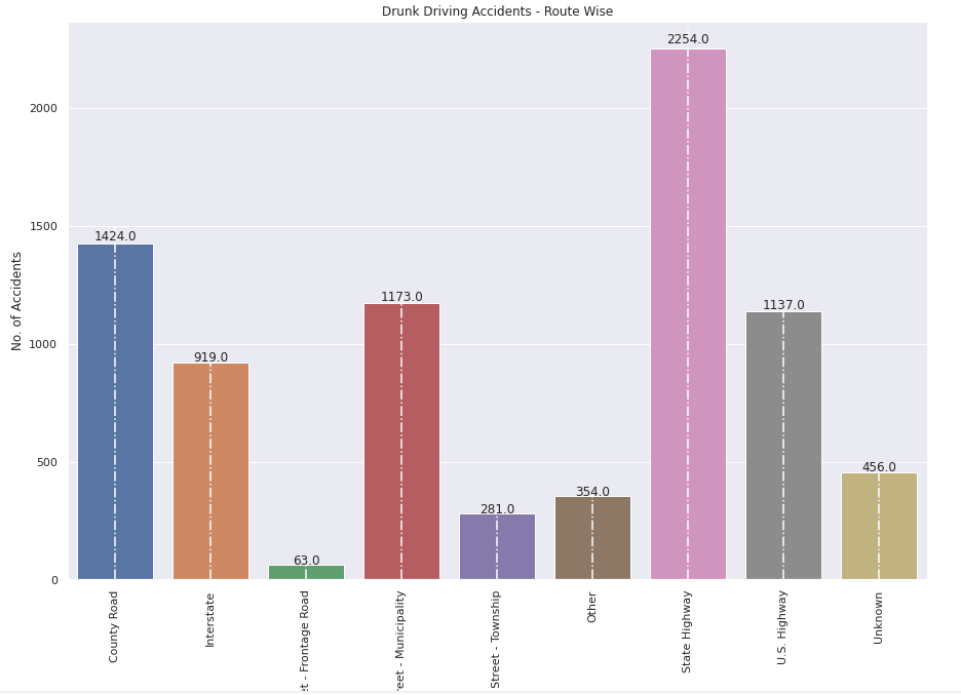

Figure 52: Bar plot - Route wise accidents

This visual shows that state highways are the places where most of the accidents occurring involve the alcohol use. County Road and US Highways, which are again places which are not monitored by the officials for alcohol checks while driving, are the places where the rate accidents are mostly drink driving accidents.

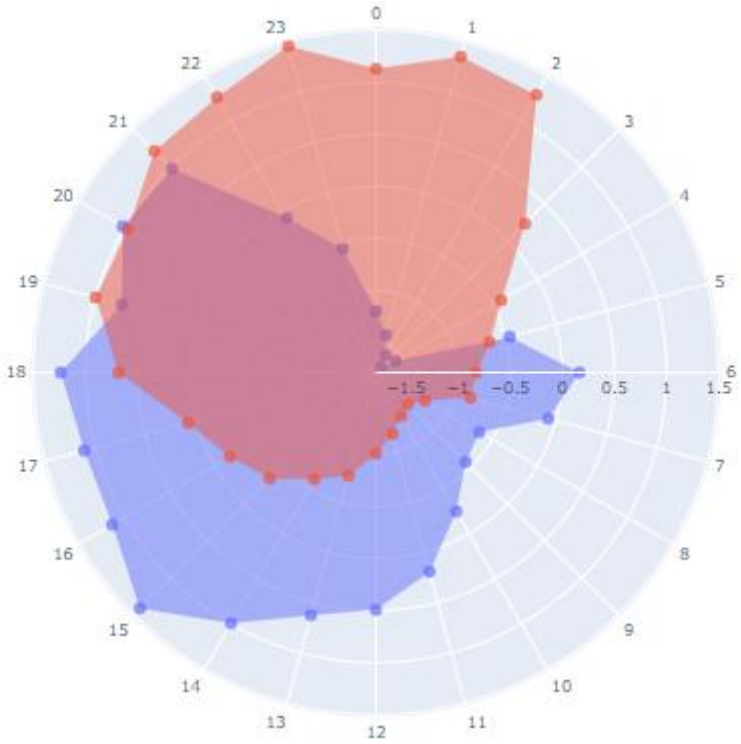

Figure 53: Radial Chart - Timings of Accidents

Next, we analyse the time of the accidents shown in figure 53. Here, we see that we have plotted the hours of the day in the radial chart. It helps for the user to have continuity in the data. We notice that non-drunk driving percentages are higher in the morning, when people are usually not fresh, I.e. at 6 am. After this, the accident rates decrease drastically, when people are mostly working. Then, the rates increase when people are out in the evening, and mostly tired after the whole day.

The drunk-driving percentages on the other hand start rising when it's the night time. It starts occurring from 7pm when most of the recreational places open, and peaks at 11pm to 1 am, when these places start closing and people have to return home.

Here, we try to analyze state-wise alcohol consumption of U.S. States and try to compare socio-economic effects of alcohol use. Our aim with these visualizations is trying to analyze some relation between alcohol use and its socio effects on factors such as Poverty, Unemployment, Study, Debt, etc. However, whatever we conclude is not based on intensive research and we have not considered all factors while making these assumptions. However, we would be citing published papers whenever we make any assumptions.

Here, we would begin with visualizing the state-wise alcohol consumption data as we would be comparing these results with other socio-economic factors as well. Using boxplot, choropleth and diverging plot would give us some insight.

First, let's analyze how much alcohol is consumed on average among all U.S. states using boxplot shown in figure 54 and plot all points because combining these two, makes the visualization more comprehensible.

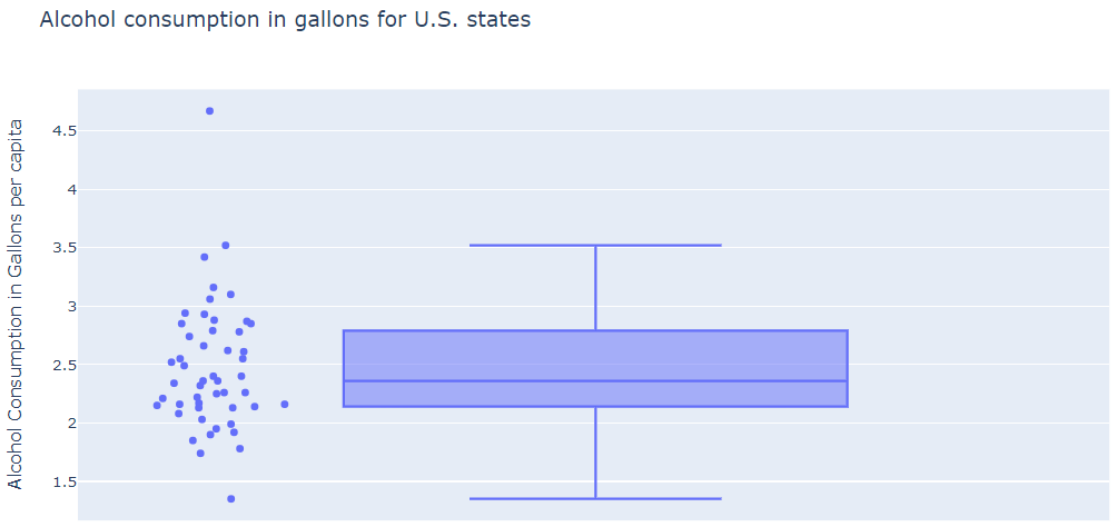

Figure 54: Boxplot and BeeSwarm plot

Second, we would use choropleth.

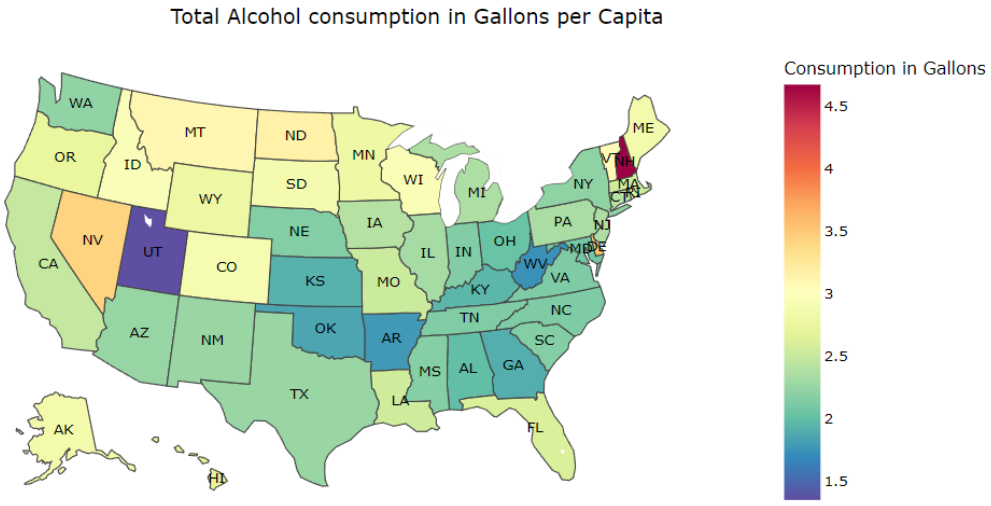

Figure 55: Choropleth - Total Alcohol Consumption in Gallons per capita

So here, we can analyze that the lowest alcohol consumption per capita is in Utah which is shown in purple colour and the highest alcohol consumption is in New Hampshire which is coloured red. New Hampshire consumed 4.67 gallons of alcohol in 2019 and Utah consumed only 1.35 gallons of alcohol.

We can also analyze those northwest states consume more alcohol compared to other states. Another graph we can use is a diverging chart, which would give us relative values of consumption in all states.

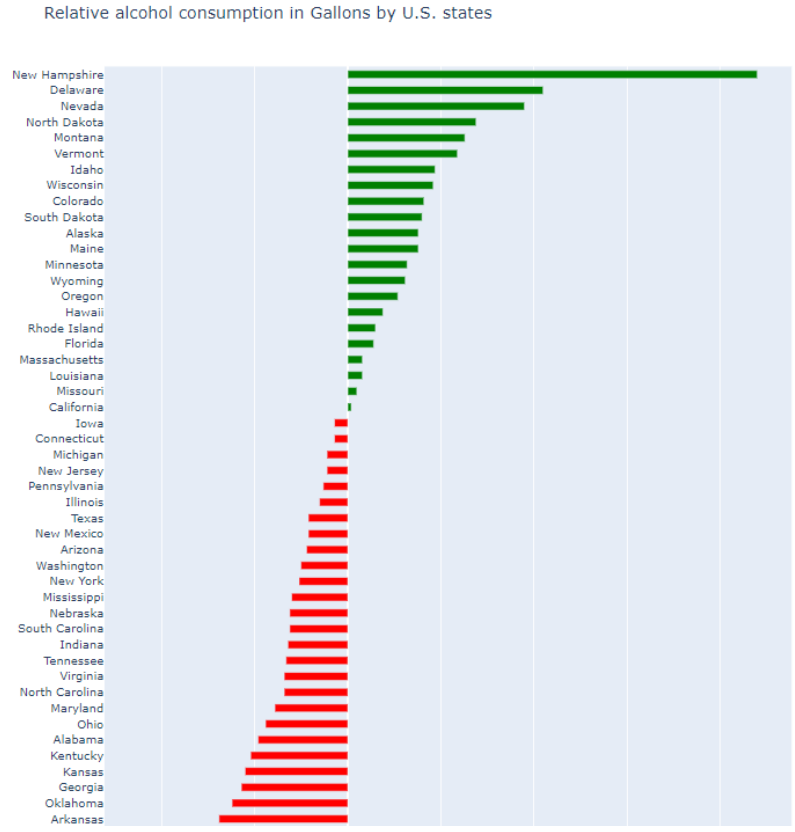

Figure 56: Diverging plot

This gives us clear knowledge of how each state compares to other states in alcohol consumption. Next, let's analyze alcohol consumption based on government parties that control those states.

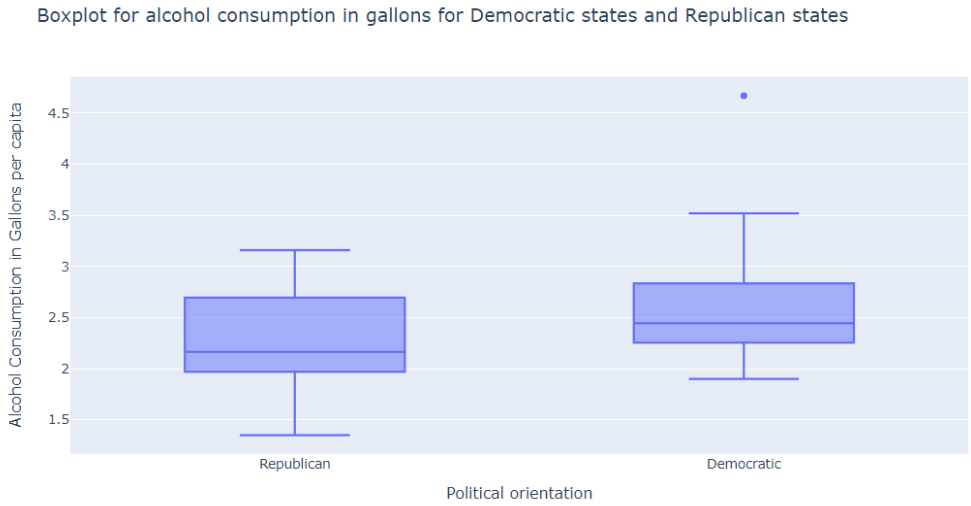

Figure 57: Boxplot comparing Alcohol consumption among Republican and Democrat states

From the above visualization, we can analyze those republican states on average consume less alcohol than democrat states. This confirms that democrat states drink more alcohol compared to republican states [8].

Now, let's try to analyze how alcohol consumption affects states and their expenditures. According to CDC, Excessive alcohol consumption cost the United States $249 billion in 2010, which amounts to about $2.05 per drink or about $807 per person. Of all costs, 72% of the cost is due to workplace productivity, 11% is health care expense, and other costs are due to a combination of criminal justice expense, motor vehicle crash costs and property damage.

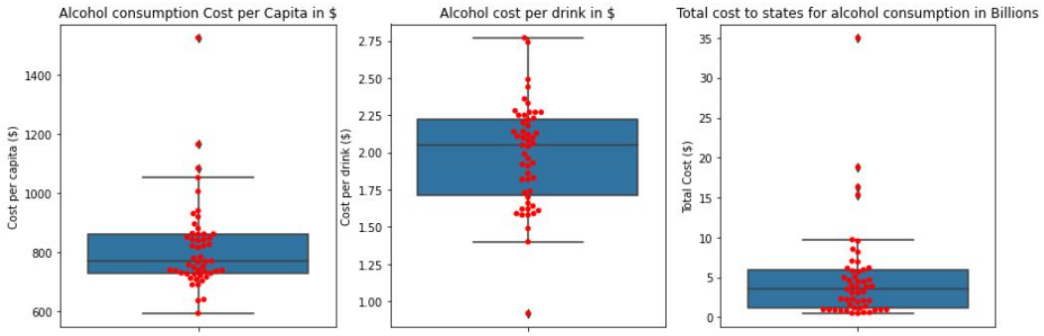

Figure 58: Boxplots to compare state-wise expenditure behind alcohol consumption

Here, we can depict that information given in the paragraph above can be confirmed using the visuals. Few states are there which are outliers in Alcohol Consumption cost per Capita and Total cost so let's try to find which states have extremely high total cost and cost per capita using choropleth. First, we will plot the choropleth of Total cost to states:

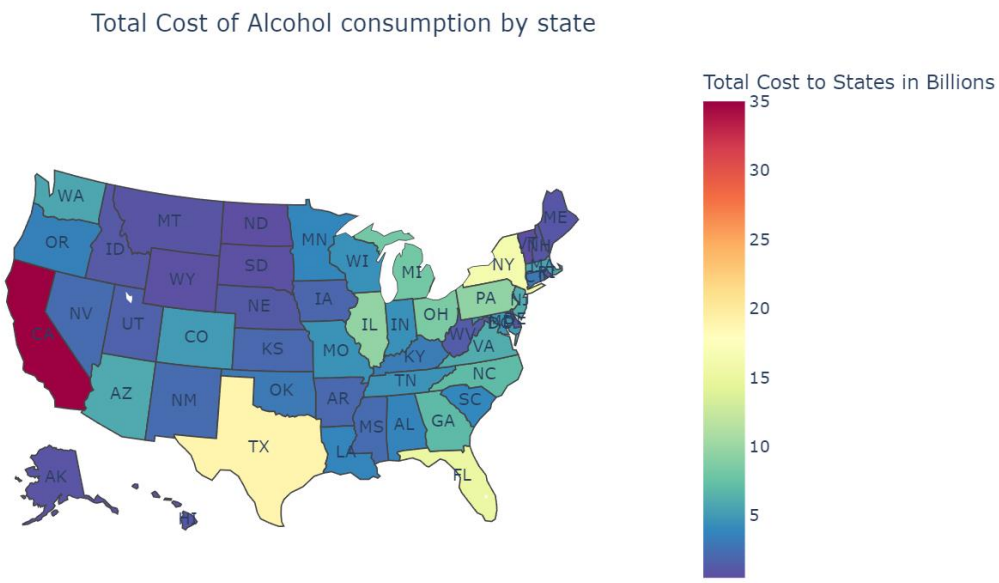

Figure 59: Choropleth - Total cost to states because of alcohol consumption

Here, it's clear that California has the highest total cost of around $35 billion. The reason why we chose this colormap is that it gives distinct colours for outliers. But one thing we need to consider here is that Total cost is not an accurate measure to compare all states as California has the highest population so it is obvious that it would have a higher total cost. For accurate comparison, let's compare the Total cost per capita then.

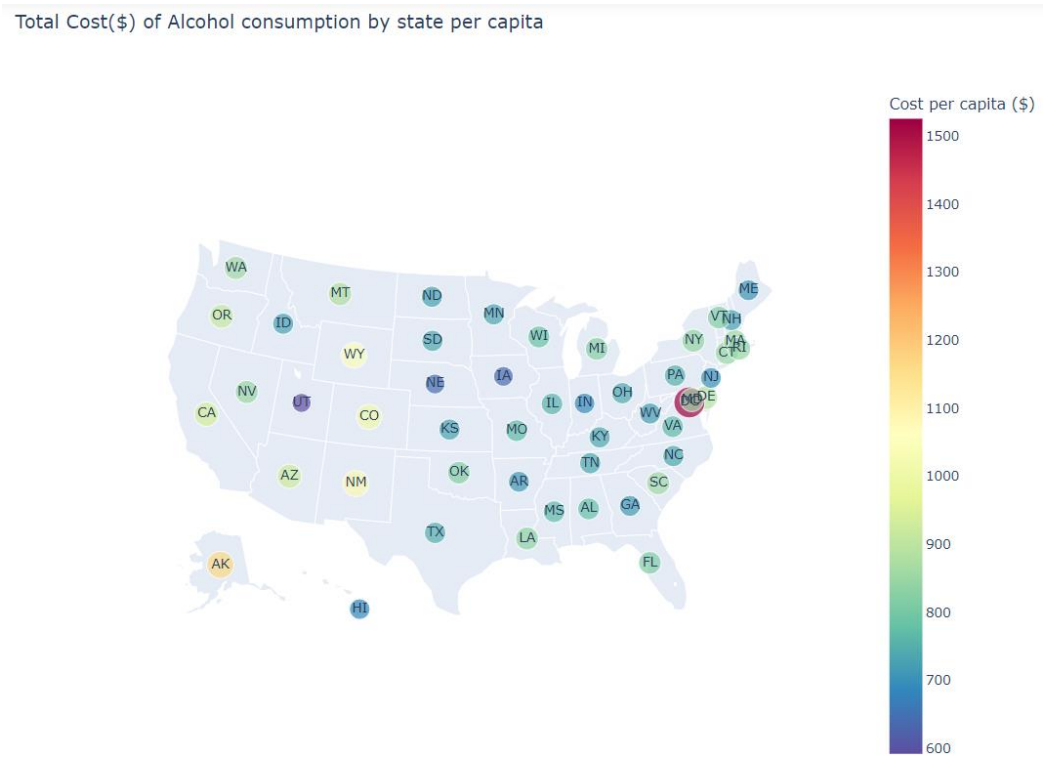

Figure 60: Bubble Map for alcohol consumption in gallons per capita

Here, we have used bubble maps because in our case, each line of the dataframe is represented as a marker point. The bubble size depends on the value of the total cost. Also, in our case, the state with the highest cost per capita is the District of Columbia and it is quite small to use choropleth. Thus, we decided to opt for a bubble map. Now, let's try to find the relation between happiness and alcohol consumption.

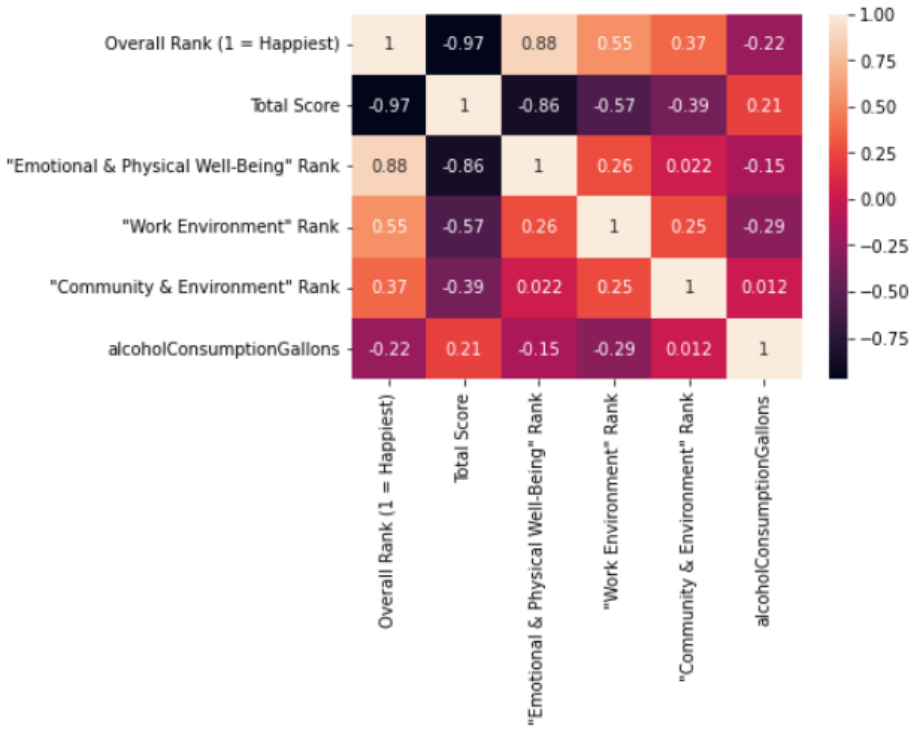

Figure 61: Heatmap - Comparing Correlations

Here is the correlation plot between columns "Total Score" and "alcoholConsumptionGallons". As we can see, there is a 0.21 correlation. But let's try to visualize this relation with scatter plot with trendline.

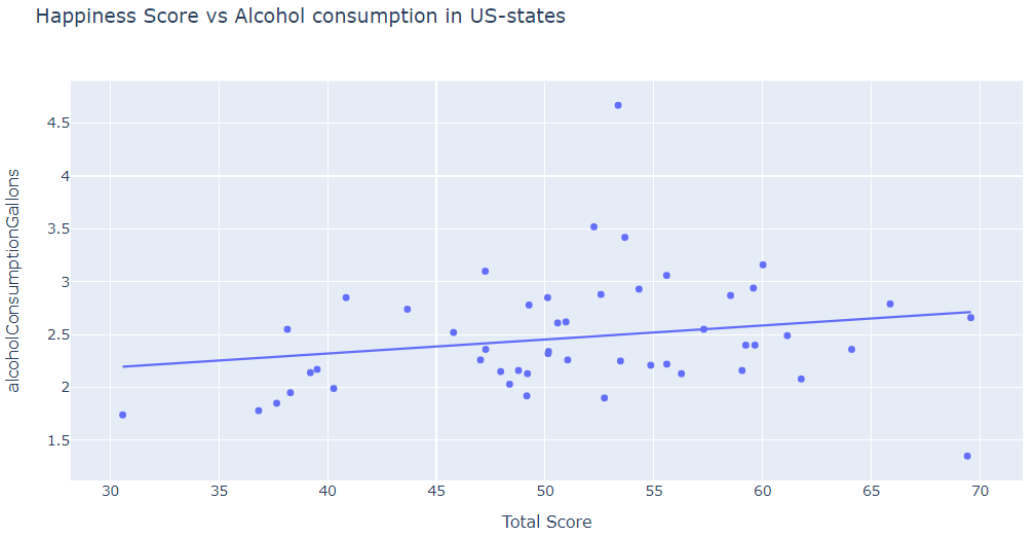

Figure 62: Happiness Score vs Alcohol Consumption

We can clearly see that there is an upward trend in this relationship, but this can also be a coincidence. Also, the correlation is quite high enough to make viable judgments. But there are a lot of reports that cites that there is some correlation between happiness and alcohol consumption [9].

Let's try to change our focus to Unemployment and alcohol consumption. From few reports [10] we came across, there is a positive correlation between alcohol consumption and unemployment rate, but we couldn't find any significant relation between these factors from the dataset we used. The results are,

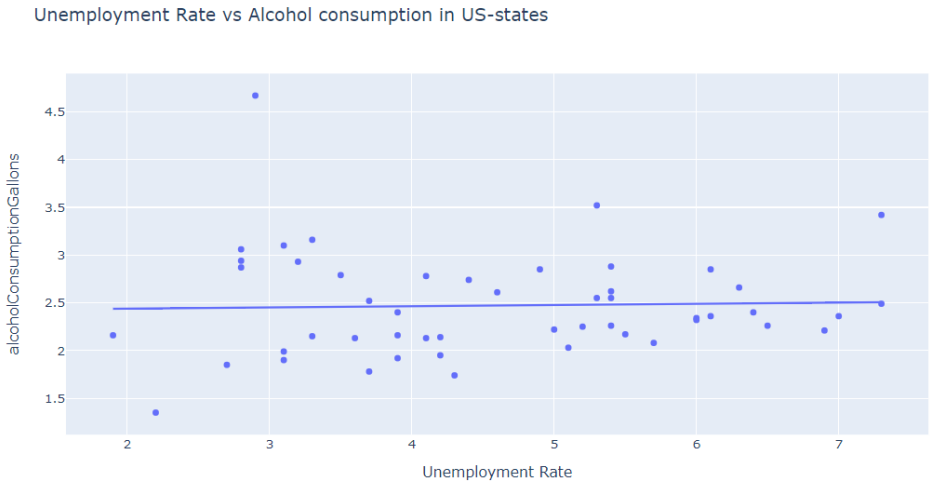

Figure 63: Scatter plot - Unemployment Rate vs Alcohol Consumption with trendline

We can see in figure 63 above that the trendline is quite flat. Even the correlation score we got was only 0.033 which is very insignificant to make any assumptions. However, the Debt-to-Income ratio and alcohol consumption have a high correlation according to our visualizations. Alcohol is quite expensive and the more you spend on alcohol, the more debt you have can be a viable assumption in our opinion. Let's try to back this assumption with some proper visualization.

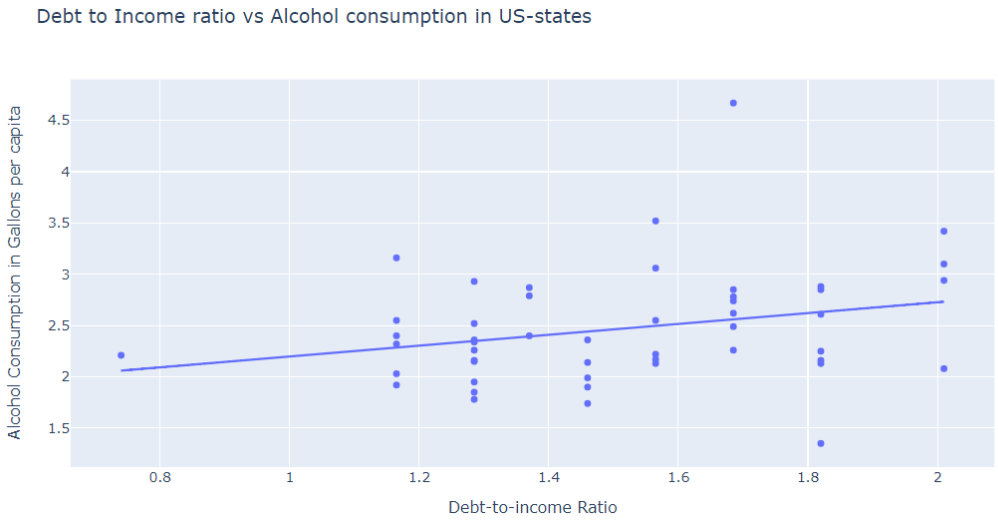

Figure 64: Scatter plot with trendline – Debt to Income ration Vs Alcohol Consumption

The correlation between Alcohol consumption and the debt-to-income ratio is 0.26, which is slightly high. Here, we can depict that there is an obvious positive relation between the debt-to-income ratio and alcohol consumption. As some reports suggest, people with binge drinking problems have higher debt compared to people who don't [11]. Though we are not taking other factors such as income, cost of living, etc. into consideration, this can be a valid assumption.

DISCUSSION

Alcohol Consumption is a very common practice in most parts of the world. With varying alcohol types, through this project, we capture the distribution of consumption of these types through effective visualizations. Map visuals are the best wat to capture these. One can use various types of maps, and through interactive and creative methods, we can show the trends better. Alcohol consumption can be correlated with many social and economic factors too. We also visualized these factors such as happiness index, income per person, unemployment rate, and political factors such as the state being a democratic or a republican one. Furthermore, we also analyzed the effects of alcohol consumption, especially in the United States. United States accidents on the roads were analyzed which were caused under the influence of alcohol. We, capture these mentioned trends through various plots, and aim to portray a much clearer picture of the current scenario.

CONCLUSION

The consumption of alcohol per capita is the highest in Europe and the value decreases where the per capita of that country is high. Analyzing alcohol consumption per capita gives a better and fair comparison than directly comparing the alcohol intake. After the effective analysis on different types of alcohol consumed, we can understand from the visualizations that consumption of beer is high in Germany this might be due to the presence of more breweries in Germany. In countries like France and Italy consumption of wine is high as there are more wineries located in those countries. Road traffic deaths under the influence of alcohol are higher in some of the countries like Zimbabwe, Malawi, and Liberia which are in Africa. This might be due to not having more patrol cars or more testing for drunk driving. After analysis on the number of gallons consumed by different states, we came across these insights: the consumption of alcohol is the highest in New Hampshire and the least in Utah. Also, consumption of alcohol is more in Northwest states compared to the rest of the states. After plotting different kinds of alcohol consumption data, we understood that just comparing the consumption of alcohol won't be enough. As states with higher populations will consume more alcohol compared to states with lower populations. We used various visualizations to represent the data in various datasets based on alcohol consumption to gain an understanding of the current scenario for alcohol consumption. We used different visualizations like choropleth maps for representing data of different countries and at the country level for the USA, box plot, beeswarm plot, radial chart, diverging chart, and heatmap.

FUTURE WORK

The current analysis is portraying the current scenario in the world and in the United States about the alcohol consumption and its effects. In the future, one can aim to visualize and convey the effects of alcohol on the social life of the person, and the surrounding community. Since there already exists economic and historic data about the alcohol and its consumption and correlating it with the social life of the person could be an interesting scenario to visualize.

REFERENCES

[1] Griswold, M. G., Fullman, N., Hawley, C., Arian, N., Zimsen, S. R., Tymeson, H. D., ... & Farioli, A. (2018). Alcohol use and burden for 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. The Lancet, 392(10152), 1015-1035.

[2] US Department of Health and Human Services. (2000). 10th special report to the US Congress on alcohol and health. Highlights from Current Research from the Secretary of Health and Human Services. Washington, DC: US Department of Health and Human Services.

[3] Ventura-Cots, M., Watts, A. E., Cruz-Lemini, M., Shah, N. D., Ndugga, N., McCann, P., ... & Bataller, R. (2019). Colder weather and fewer sunlight hours increase alcohol consumption and alcoholic cirrhosis worldwide. Hepatology, 69(5), 1916-1930.

[4] Room, R., Babor, T., & Rehm, J. (2005). Alcohol and public health. The lancet, 365(9458), 519-530.

[5] Health risks and benefits of alcohol consumption. Alcohol Res Health. 2000;24(1):5-11.

[6] Mann, K., Hermann, D., & Heinz, A. (2000). One hundred years of alcoholism: the twentieth century. Alcohol and alcoholism, 35(1), 10-15.

[7]Keller, M. (1979). A historical overview of alcohol and alcoholism. Cancer Research, 39(7 Part 2), 2822-2829.

[8] Yakovlev, P., & Guessford, W. (2013). Alcohol Consumption and Political Ideology: What's Party Got to Do with It? Journal of Wine Economics, 8(3), 335-354. doi:10.1017/jwe.2013.23

[9] Richard Florida, (2011), The Drunkenness of Nations: [link]

[10] WHO, Global report of 2004, [link]

[11] Hayley Hudson ,Alcohol Rehab Guide, (2021), [link]