Alcohol Consumption

Coursera Capstone by Matilda Ferrand

Introduction.

In this report going to examine the relationship between the age that alcohol consumption starts and the detrimental effects alcohol produces through life. This is a very difficult question to answer, because of its high correlation with other cultural factors. It's also rather difficult to measure: In countries where alcohol is banned it will be very difficult to get true values for alcohol consumption. This is also true for underage drinking, given that it is illegal. I have chosen to study this country by country rather than on an individual basis so that there will hopefully be a wider range of data, which is more even. My first step was deciding how to measure the age that alcohol consumption starts. I decided to go two different routes with this, and chose a dataset that gives the legal drinking age per countries and another dataset which gives the percentage of underage drinkers per country. Hopefully together, this will give a fuller picture of young drinking habits. For the effects of drinking I chose to examine average drinking patterns by looking at the number of people who had had a heavy drinking episode within the last 30 days and the number of people who had abstained within the last year. Again the heavy drinking can be subjective because of the unreliable nature of the word heavy. Finally I looked for data which would show the detrimental effects of drinking. Again I chose two datasets, showing the percentage of liver failure that was caused by drinking, and the percentage of road accidents caused by alcohol. These were perhaps the most subjective datasets, given that a country with a looser road policy, or with less regulated driving tests will have more accidents, therefore shifting the numbers. Liver failure can also be caused by diet or smoking, both of which are cultural. Despite these barriers, hopefully the number of data can lead to an interesting outcome, however the outcome can only be viewed as a correlation rather than causation.

Data.

To start with, I've had to load a number of datasets onto my Jupyter notebook. These datasets were Percentage of underage drinkers per country, legal ages for drinking beer, wine and spirits per country, percentage of people who've abstained from drinking for the past twelve months per country, percentage of people who've drunk heavily and episodically in the past month, percentage of liver cirrhosis cases caused by an abundance of alcohol and percentage of traffic accidents caused by alcohol. To do this I need to merge each dataset into a large Data Frame, using Country names as a common column.

Methodology: Data Cleaning

The first step was to ensure the data was loaded correctly and to look at the current state of the data frame. This was my initial data.

I immediately noticed a number of columns named year and that there was an unnecessary ID column. To fix this I re-indexed the data, so that the country was now the index. I then removed all but the first year column. I also considered the column names to be too long. While many people might not consider that too important, it was annoying me so I made them more concise.

Next I started to explore the dataset. I started by counting the number of null values in each column; there were none. Then I retrieved the statistics and the dataset information. I struggled graphing the data due to the large number of countries, which made it almost impossible to plot in a graph.

I decided that there were too many columns referring to the age which each alcohol was allowed per country, and decided I wanted to make this one column. I reviewed the different values, and decided to make a new column which states youngest age alcohol is permitted, which finds the youngest age any of these alcohols are allowed. After looking through the data, the youngest values were all on the sale of beer on a premise, so I renamed that column youngest total value and dropped the other columns.

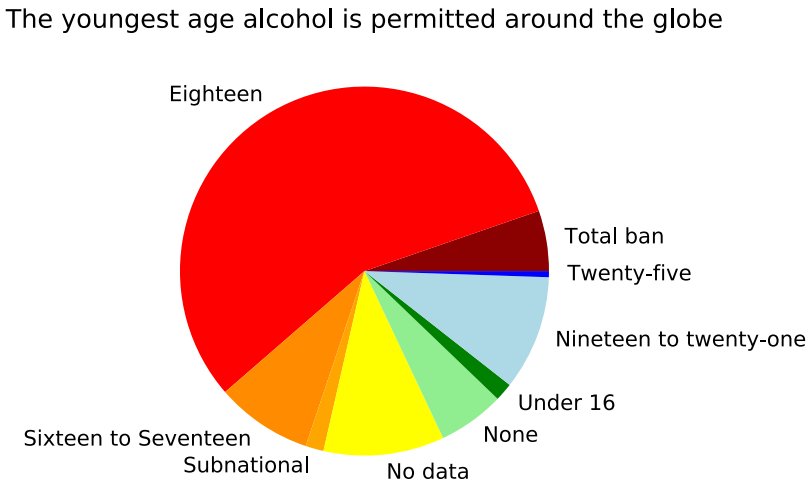
My column on alcohol use under 19 was in an odd format. The percentage was displayed with the values it was averaged from in square parentheses next to it. All useful information, but the model would be unable to identify this as numbers and instead would class it as an object, so I got rid of the square parentheses section.

The first row was simply headings so that was then deleted.

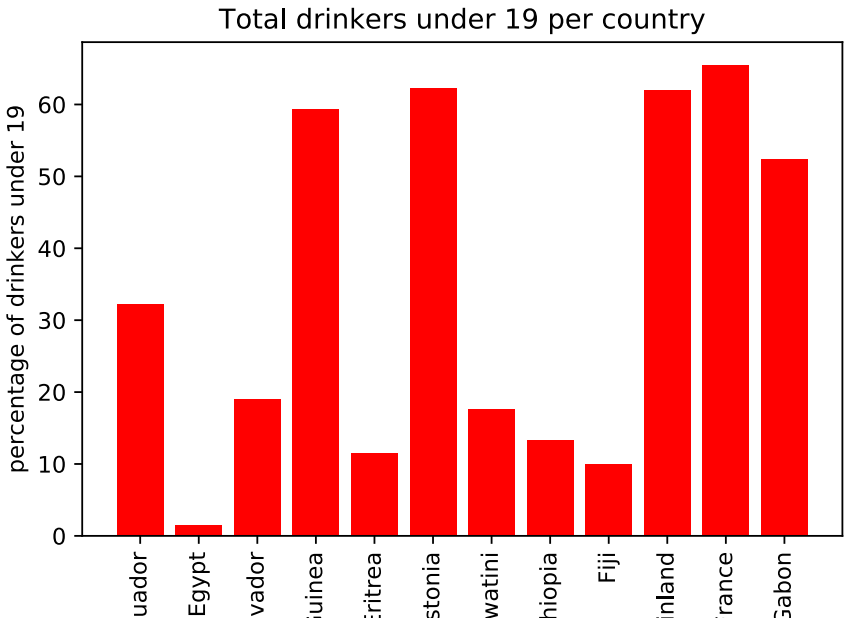
All the data types were objects, despite most of them being numbers. I as able to change all the data on recent heavy drinking and numbers of drinking under 19 easily from object form into float form. I then realised that on most of the columns there were null data points, but they were in the form of —- or *No Data.* I replaced the no data with a 50, and then changed the data types into floating points. I did this for each column except year, which I left as an object, and legal drinking age. I made the decision to make legal drinking age an object after consideration because a number of countries have subnational regulation, total bans and no age limit which are hard to quantify. I made a number of categories: subnational, total ban, none, eighteen, sixteen to seventeen, under sixteen, nineteen to twenty one and above twenty one.

Methodology: Data visualisation

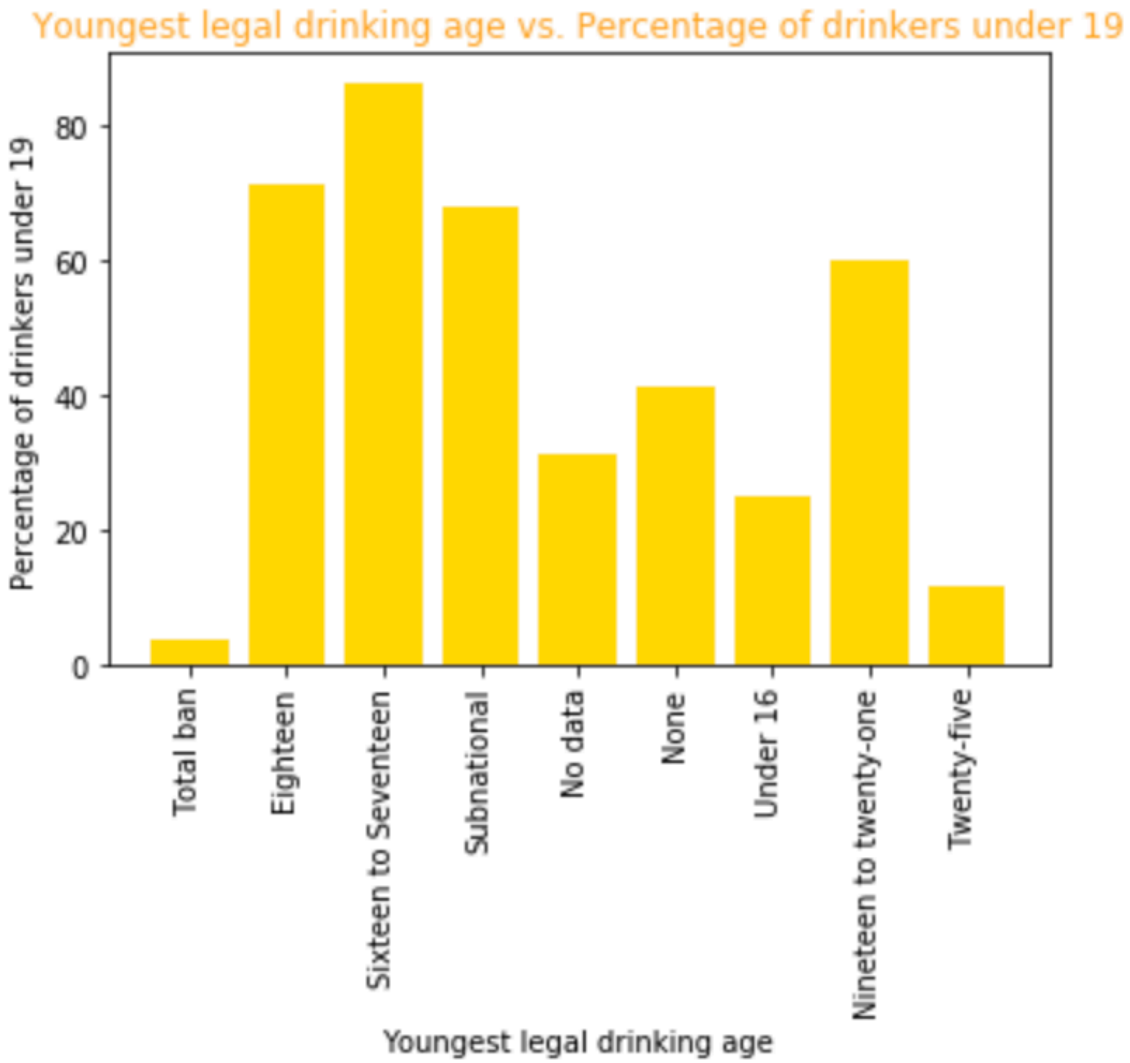
I started by looking at the differences between drinking ages. To view this I created a pie chart, as shown below to show the distribution of regulations. Over half of the values were eighteen with the next largest group being non teen to twenty-one and sixteen to seventeen. There was a large number of countries which didn't have data.



I then began to explore the different features in order to identify various correlations. The column I decided to focus on was the number of underage drinkers, so I plotted a number of features against this. In order to look at the range I created a bar graph of the percentage of drinkers under 19 with the different countries, however as there were so many different countries the bar graph was unreadable. To solve this issue, I selected a random number and plotted a 12 random countries. This was one of the resulting graphs. It shows a wide range, with Egypt having a percentage well under 10% and France having a percentage above 60%.



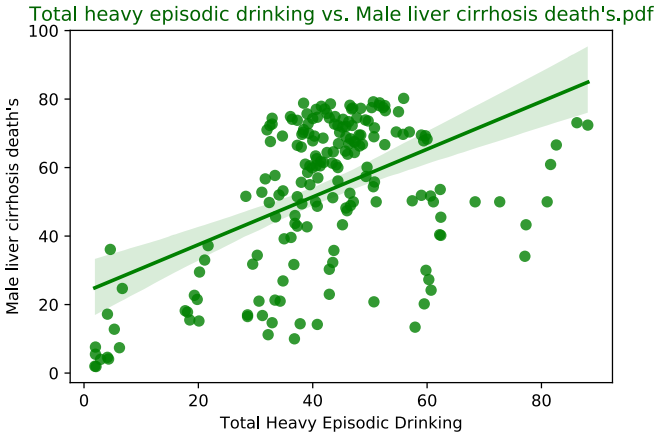
I then created a bar graph to show the relationship between the youngest legal drinking age and the percentage of drinkers under nineteen. The highest percentage was in the sixteen to seventeen category, with over 80% of the population drinking before nineteen, and, as expected, the lowest category was the Total Ban which had under 10% of the population drinking under nineteen.

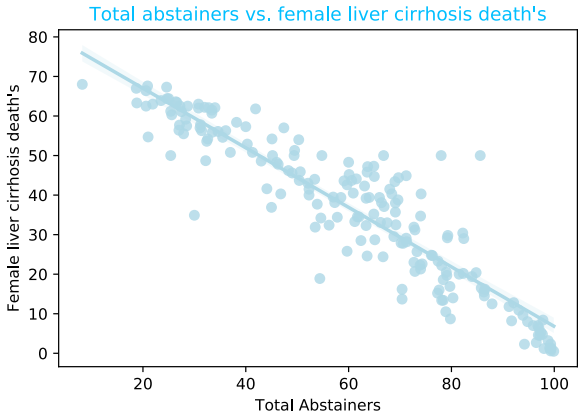


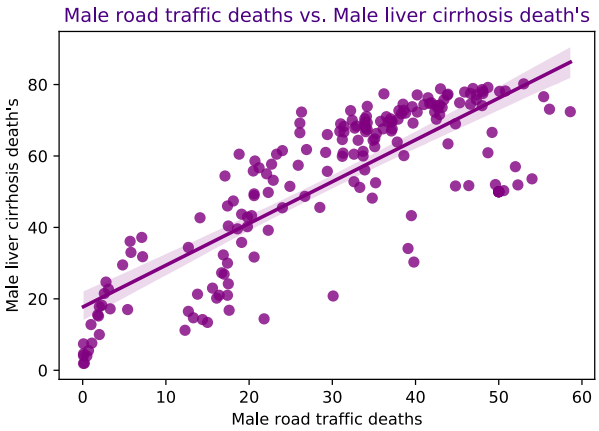
I then created a few plots to show how different features correlated in the below graphs. The correlation was strongest between female liver cirrhosis deaths and total alcohol abstainers, showing a strong inverse relationship. There was some relationship between total heavy episodic drinking and male liver cirrhosis death's and a stronger positive correlation between male road traffic deaths and male liver cirrhosis death's. When I graphed youngest legal drinking age vs. Total heavy episodic drinking it generally tended to show the younger legal drinking age producing higher heavy episodic drinking later in life, but I was surprised to see that the under sixteen column showed fairly low

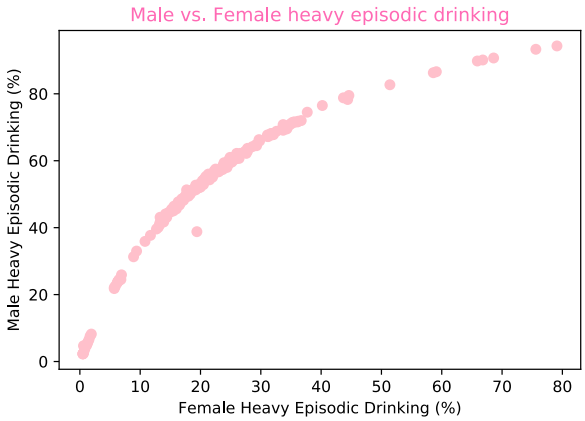
youngest legal drinking ages. This could be, however, down to the small size of the under sixteen section: there weren't many countries so therefore there would have been less data.

I also plotted male vs. Female heavy episodic drinking.



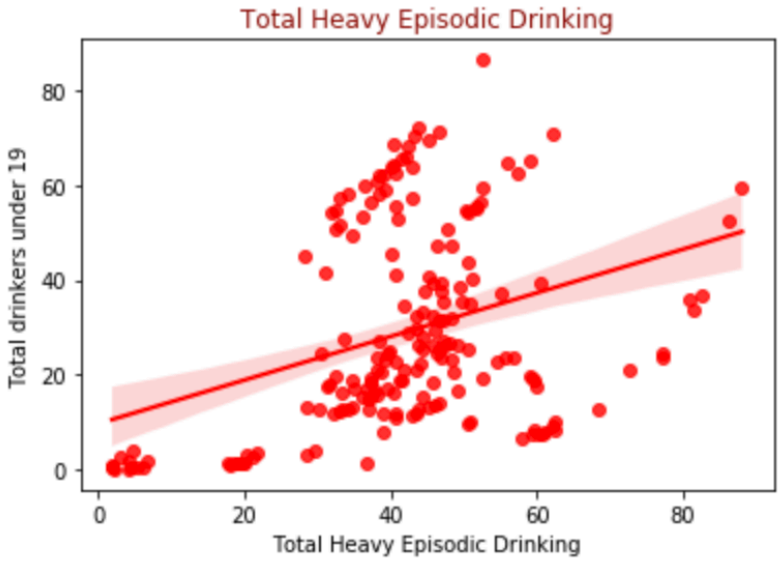


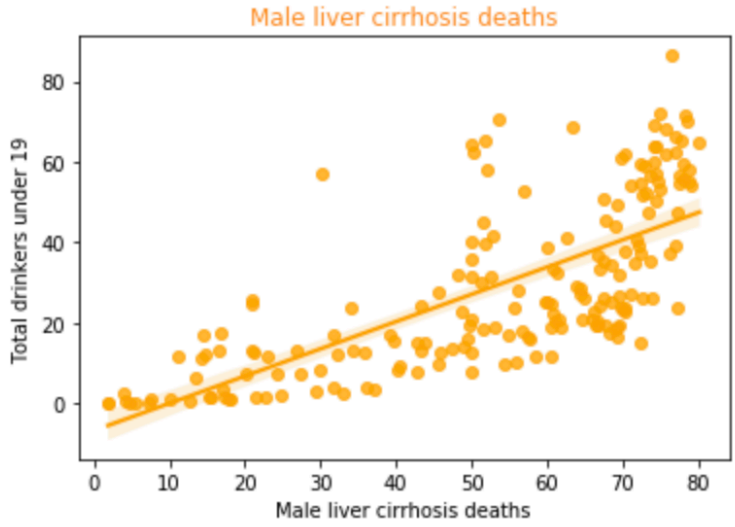


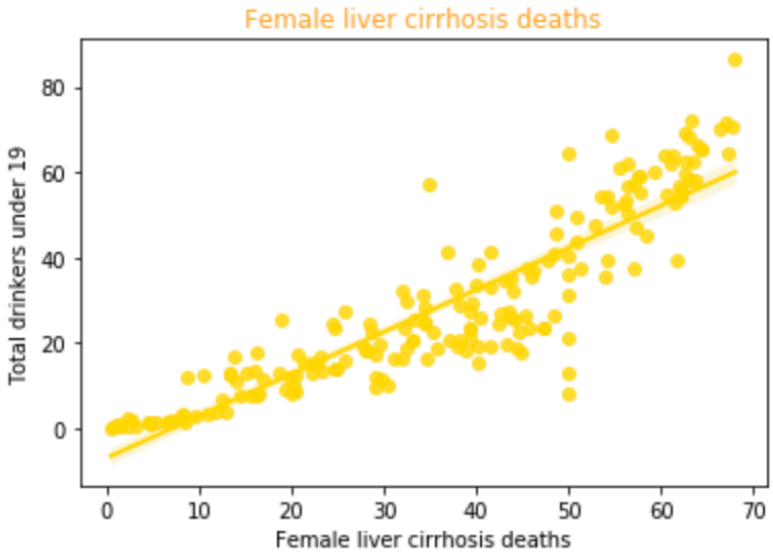


Results: Correlations

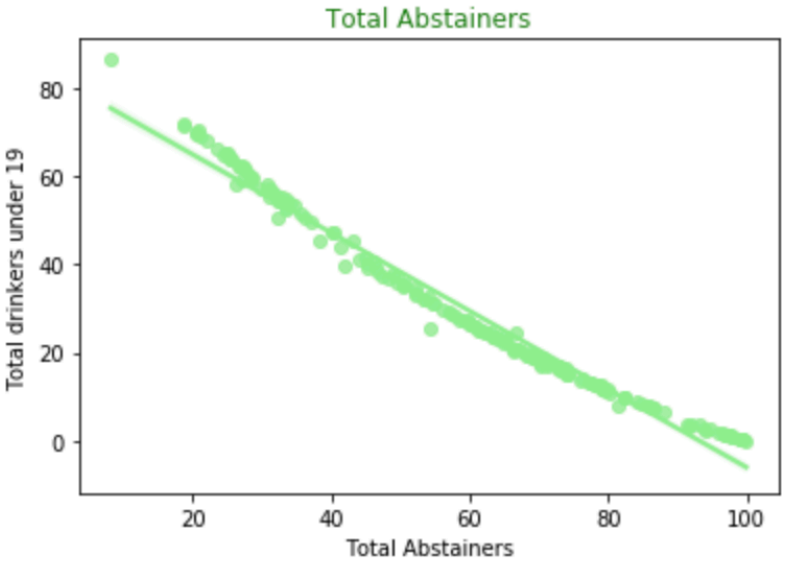
I then graphed all the values against number of drinkers under nineteen, and calculated Pearson R values to see the correlations. The lowest correlation was seen from heavy episodic drinking; something which was unexpected. It was shown only to have ~0.3 correlation. Female liver cirrhosis deaths were more correlated than male liver cirrhosis deaths, which are more common. Female liver cirrhosis seems to range more between countries than male liver deaths and female liver deaths. The total abstainers show an incredibly high negative correlation with the percentage of drinkers under nineteen. This was surprising; but it shows that drinking under the age of nineteen is more correlated with normal drinking patterns or the lack thereof than of binge drinking, as shown through the total heavy episodic drinking graphs.

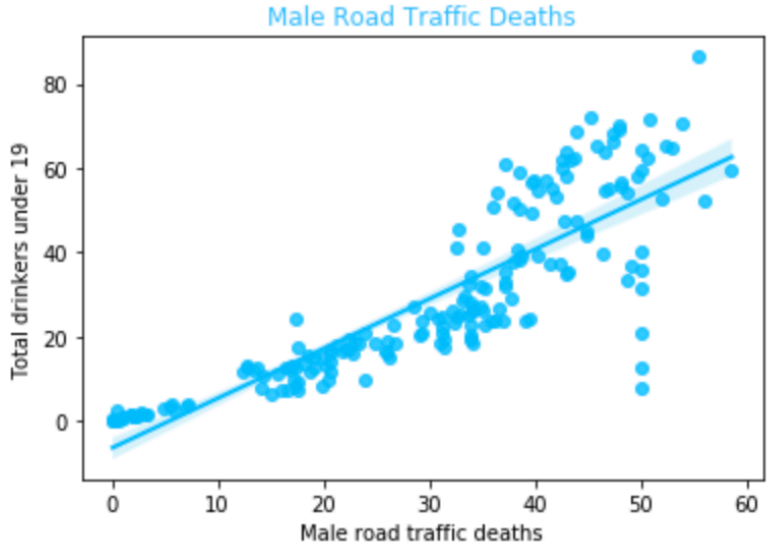


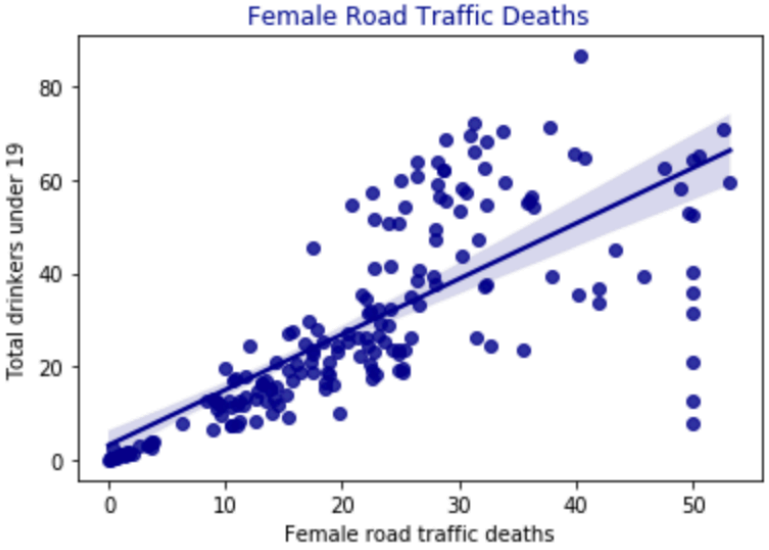




Male and female road traffic deaths both







showed correlation with young deaths, though less definitively than the total abstainers. Male road traffic deaths show more correlation than female road traffic deaths, but it has a large number of countries which have had 50% of road traffic deaths due to alcohol and a range of total drinkers under nineteen. The youngest legal drinking age showed to be less linearly correlated to the percentage of people under nineteen.

Feature	Pearson R Correlation score
Total Heavy Episodic Drinking	**0.34796**
Male liver cirrhosis deaths	**0.70956**
Female liver cirrhosis deaths	**0.8975**
Total Abstainers	**-0.98884**
Male road traffic deaths	**0.85391**
Female road traffic deaths	**0.75527**

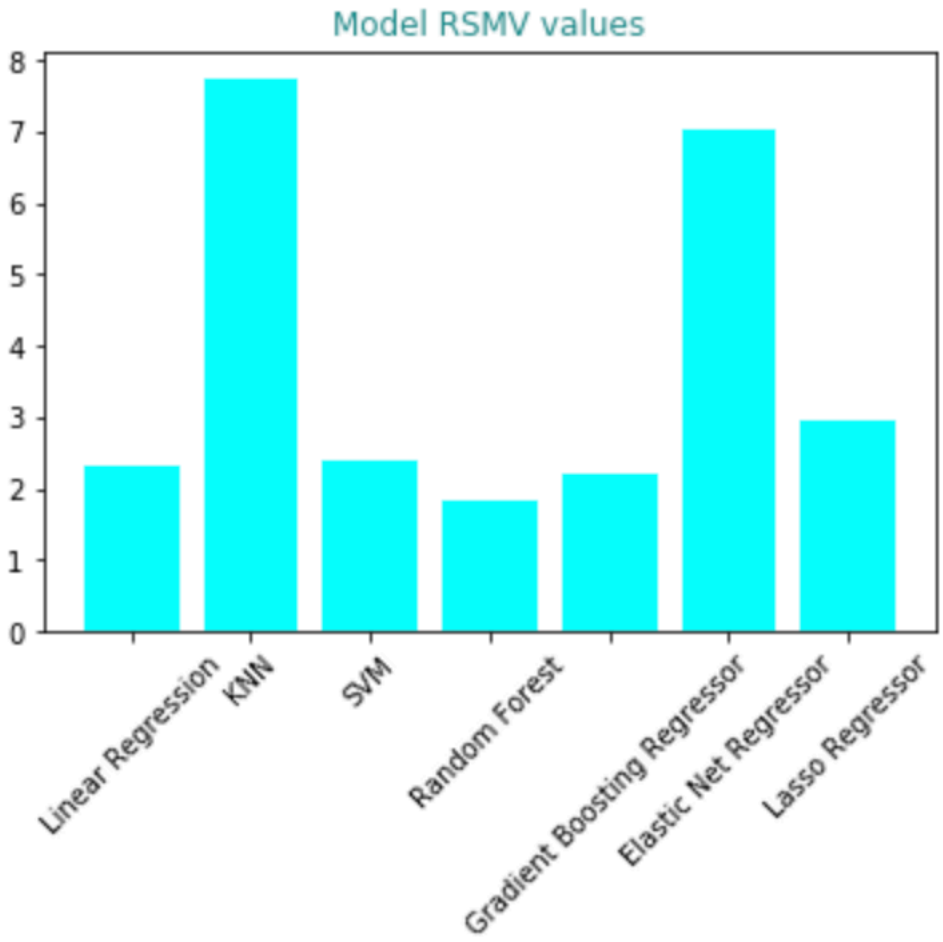
Results: Data Modelling.

I then began to create a data model to predict the number of drinkers under 19. I selected *Total heavy episodic drinking* and *Total abstainers* ignoring the separate male and female sections. I also selected the *Youngest legal drinking age*, *Female liver cirrhosis deaths*, *Male liver cirrhosis deaths*, *Female road traffic deaths* and *Male road traffic deaths*.

I was looking at regression models so I decided to import a Linear regression model, a Random forest regression, a Support vector regression modelling, a Gradient boosting regressor model and a K neighbours regressor model. I started by splitting my data into 80% test set and 20% train set, put the data into dummies, normalised the data and then started building the models. I then made predictions for each of these models based off the validation data (20% of the dataset set aside when I calculated the train test split), and Evaluated them using mean squared error and root mean squared error. The results of these evaluations are shown in a table and a graph, with Random forest and Gradient boosting regressor performing the best and the KNN and Elastic Net Regressor performing the worst.

4]:

	Models	RMSE
0	Linear Regression	2.326989
1	KNN	7.745028
2	SVM	2.395010
3	Random Forest	1.850992
4	Gradient Boosting Regressor	2.200069
5	Elastic Net Regressor	7.018678
6	Lasso Regressor	2.954112



Discussion.

The results have shown a very clear inverse correlation between the number of people who've abstained from drinking for a year and the number of people who drink between the ages of 15 and 19. Personally, I found this result unexpected. Using this result, I can hypothesise that the number of people who drink under the age of nineteen is related to a steady alcohol culture rather than a culture of excessive alcohol drinking, as the correlation was very low in that sector. I would hypothesise that the number of drinkers between 15 and 19 would vary based upon cultural values and availability of alcohol. This would imply that where there are more abstainers, there are probably fewer sources of alcohol and it's probably less prominent in culture than it is in countries where drinking is more common. I had thought that heavy drinking would be highly cultural and based upon availability, but these results have made me rethink this hypothesis and I would now suggest that heavy drinking was more a subject of alcoholism and potentially mental health, both of which are not as affected by culture or availability of supply. These results, however, in no way prove this hypothesis and another study would have to be completed with more data to show this. The data showed another clear correlation between the total abstainers from alcohol and female liver cirrhosis deaths from alcohol. With a -0.92

Pearson R score, the data showed a clear negative linear correlation between these two features. This makes logical sense, but when compared with the 0.503 Pearson R score between male liver cirrhosis deaths due to alcohol and total heavy episodic drinking, it once again shows us that the percentage of the population that abstain from alcohol is a much more reliable measure of alcohol related subjects than the number of people who abuse it. The models proved rather accurate when calculating the percentage of drinkers between 15 and 19, with Random Forest Regressor, Gradient boosting Regressor and the Linear regressor producing the best results. This implies that the relationship between the points is more linear than otherwise and the success of the random forest regressor suggests that some features hold more weight than the others, and that splitting the data points into branches is highly effective. The worst model was the K nearest Neighbours, perhaps because it wasn't providing enough of a weighted view: all the features were assumed to have the same weight as each other.

Conclusion.

The results of this exploration imply that Alcoholism and destructively heavy drinking are separate from culture and alcohol regulations. This could change the way we view addicts, and regulate the legal drinking ages. Of course, a lot of research would have to be done into the way that over drinking was measured and, of course, alcoholism itself would have to be defined in a quantitive way. The results shown were promising, and with more data and more time, I am confident that a model could be created with enough accuracy to predict the number of drinkers between the age of 15-19 with a very high accuracy. Moving onwards from this topic, as a follow on project I would be very interested to look at country wide instances of heavy episodic drinking in relation to the number of people abstaining from drinking. This study was intended as a broad stroke approach because of the large numbers of data, but in the future I would like to use more data; in particular other years in relation to abstinence and heavy episodic drinking.