DATA 606 Data Project: Worldwide Alcohol Use

Thomas Hill

Data Preparation

```
library(stringr)
library(dplyr)
library(lubridate)

# load data from the fivethirtyeight github
etoh_consumption <- read.csv('https://raw.githubusercontent.com/fivethirtyeight/data/master/alcohol-cons

etoh_use_disorders <- read.csv('https://raw.githubusercontent.com/hillt5/DATA606_Final_Project/master/Pr

head(etoh_consumption)
names(etoh_consumption)
```

I also looked at the prevalence of alcohol use disorders (AUD's) in 2016. The UN defines an AUD, or problem drinking, as meeting the criteria set forth in the Diagnostic and Statistical Manual (DSM). This includes the admissions of frequent cravinngs, withdrawals and patterns of drinking more than is acceptable for social or health reasons. It also asks about the consequences of drinking - interfering with work, school, and social interactions.

```
names(etoh_use_disorders) <- c('country', 'year', 'both', 'male', 'female')
etoh_use_disorders <- etoh_use_disorders[-1,]
etoh_use_disorders$male <- as.numeric(word(etoh_use_disorders$male, 1))
```

```
## Warning: NAs introduced by coercion
```

```
etoh_use_disorders$female <- as.numeric(word(etoh_use_disorders$male, 1))
etoh_use_disorders$both <- as.numeric(word(etoh_use_disorders$male, 1))
head(etoh_use_disorders)
```

```
##               country year both male female
## 2         Afghanistan 2016  0.6  0.6    0.6
## 3             Albania 2016  9.9  9.9    9.9
## 4             Algeria 2016  1.5  1.5    1.5
## 5             Andorra 2016  9.8  9.8    9.8
## 6              Angola 2016 10.6 10.6   10.6
## 7 Antigua and Barbuda 2016  9.8  9.8    9.8
```

Data for alcohol use disorders is the 12-month prevalence in people older than the age of 15 in the country as a percent. The percent is given as for both genders, as well as separately for each gender (with the denominator being the total amont of women for AUD% in females for example.)

Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.

What effect does alcohol type have on total alcohol consumption?

Cases

What are the cases, and how many are there?

The cases are average serving sizes per person for each country in 2010, distinguished by one of three alcohol types: beer, wine, and spirit.

Data collection

Describe the method of data collection.

These data were collected by the World Health Organization as part of its Global Information System on Alcohol and Health (GISAH). These are based on government records and metrics from alcohol companies to estimate consumption per country in 2010.

Type of study

What type of study is this (observational/experiment)?

This study is observational as it aims to estimate the real consumption of alcohol in 2010. There is no randomization or treatment arm, the metrics are derived but taken as representative of the actual alcohol consumption. This data is usually collected by a division of the UN or national sources.

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

The raw data is available on GitHub https://raw.githubusercontent.com/fivethirtyeight/data/master/alcohol-consumption/drinks.csv, and the accompanying article was originally made available on https://fivethirtyeight.com/features/dear-mona-followup-where-do-people-drink-the-most-beer-wine-and-spirits/. More information is made available about the GISAH and its estimation methods on https://www.who.int/substance_abuse/activities/gisah/en/.

Dependent Variable

What is the response variable? Is it quantitative or qualitative?

The response variables are the total alcohol consumption in liters per capita. It is a quantitative variable.

Independent Variable

You should have two independent variables, one quantitative and one qualitative.

I'm looking primarily at the principle alcohol type (beer, wine, or spirit) to predict total alcohol consumption. I will also use the percent of total of each alcohol type as a quantitative measure of each country's trends.

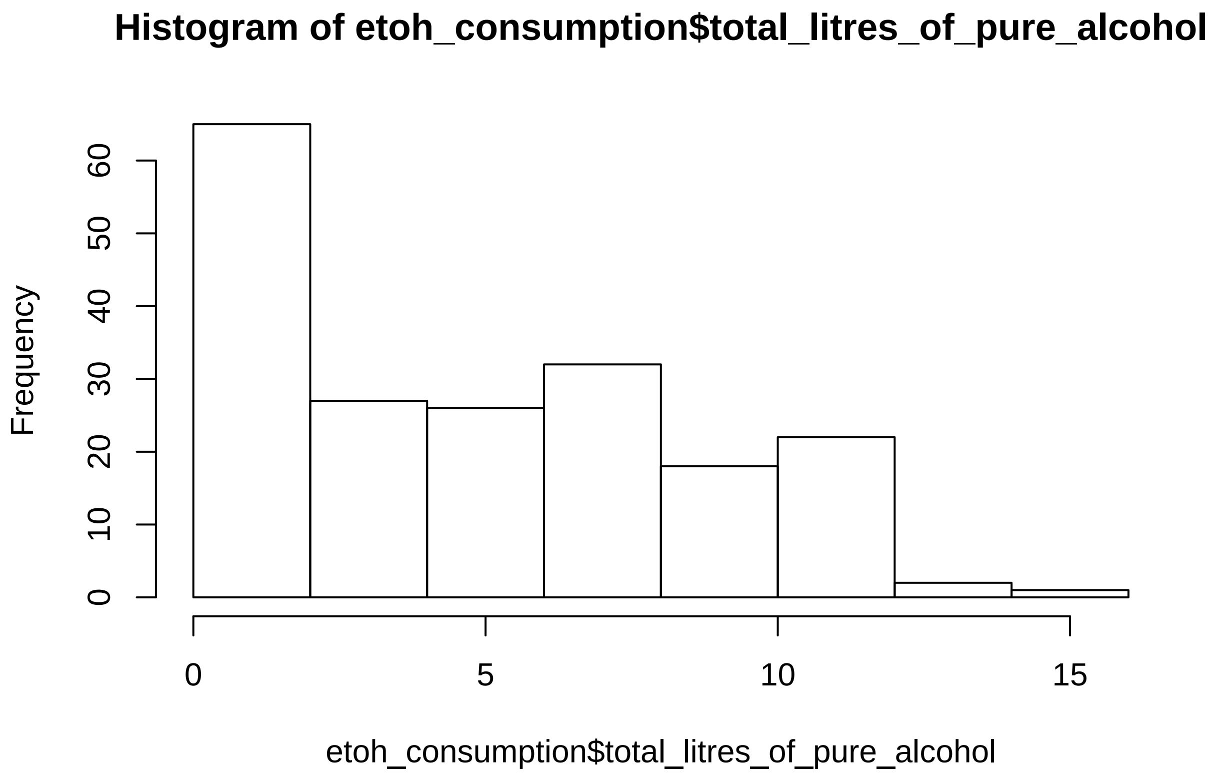
Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

```
summary(etoh_consumption$total_litres_of_pure_alcohol)
```

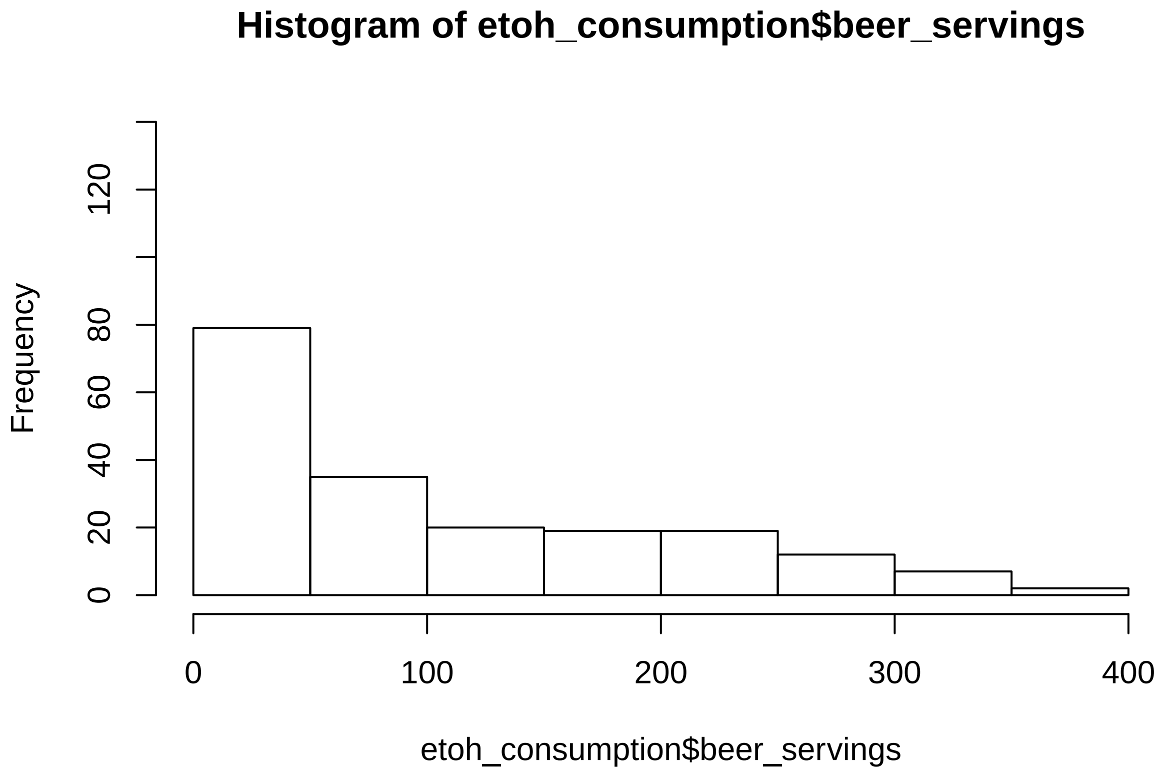
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.300   4.200   4.717   7.200  14.400
```

```
hist(etoh_consumption$total_litres_of_pure_alcohol)
```

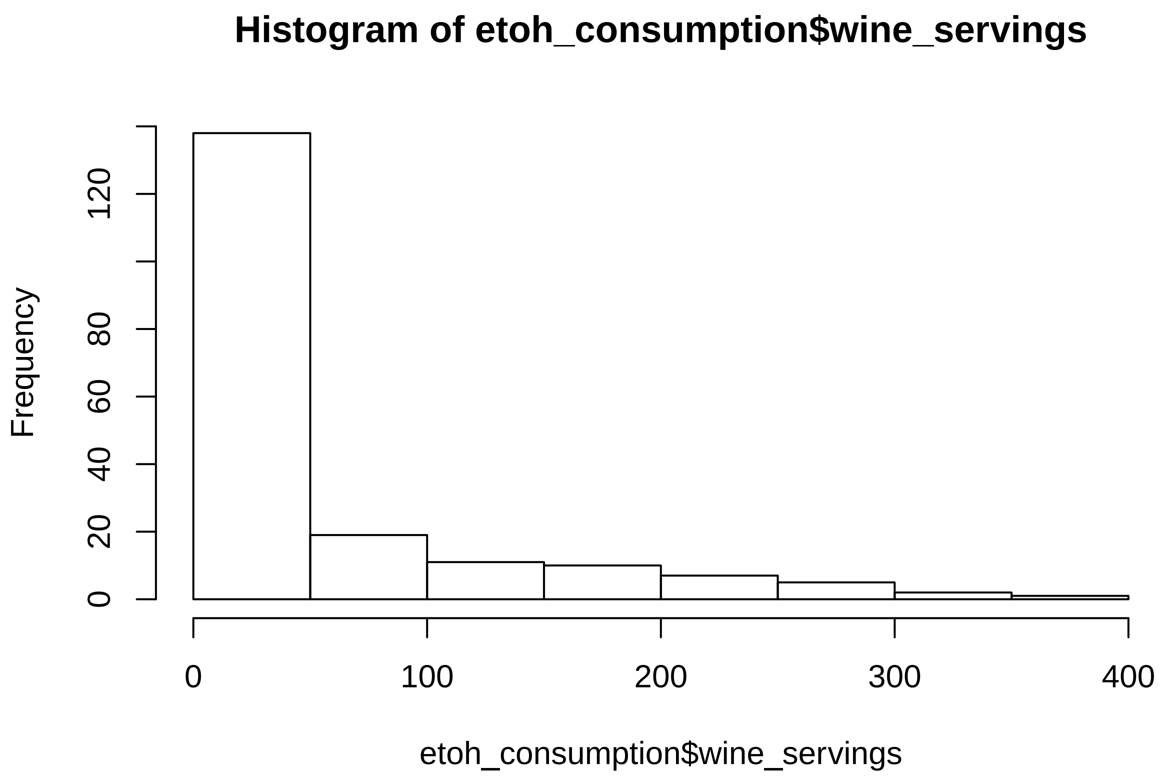


The total alcohol consumption per country is skewed to the right, likely because there are many countries where effective alcohol consumption is zero for religious reasons. This is likely an underestimator of alcohol consumption in this country as alcohol per the GISAH methodology is based off governement and commercial resources, and alcohol has a long history of personal production.

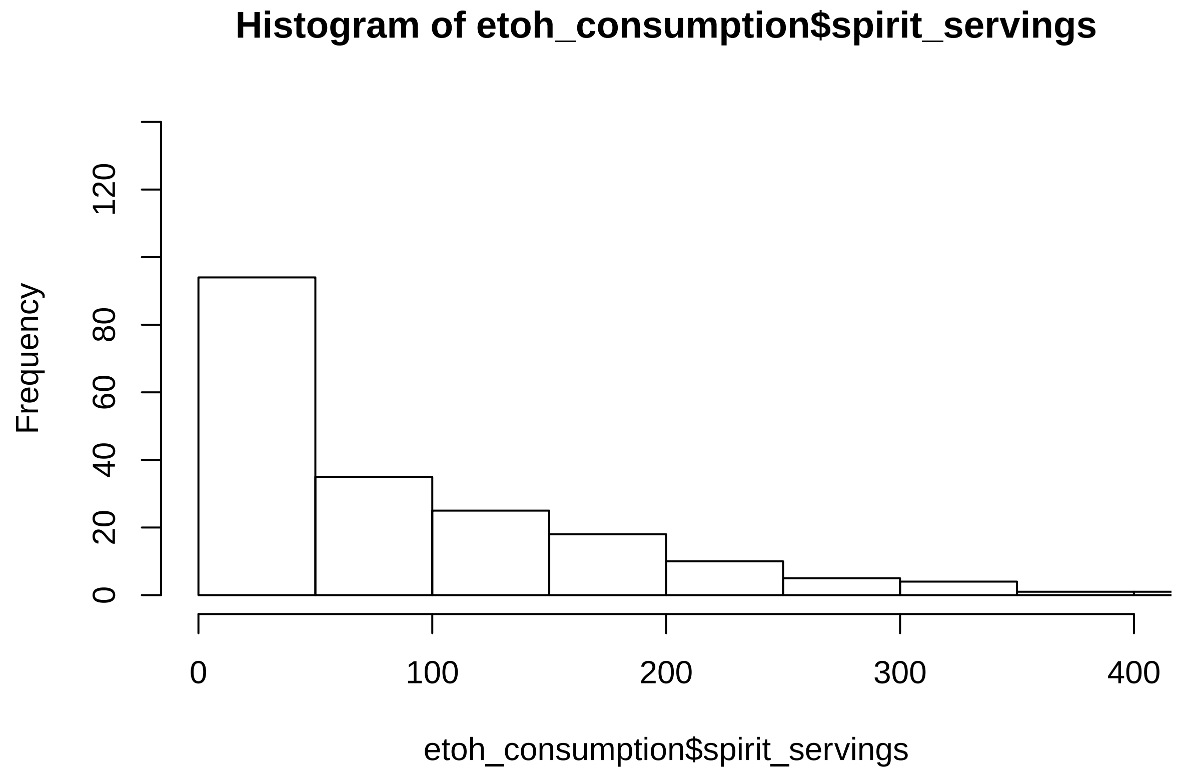
```
hist(etoh_consumption$beer_servings, xlim = c(0, 400), ylim = c(0,140))
```



```r
hist(etoh_consumption$wine_servings, xlim = c(0, 400), ylim = c(0,140))
```



```
hist(etoh_consumption$spirit_servings, xlim = c(0, 400), ylim = c(0,140))
```



By serving, wine appears to be the most exclusive alcohol type. Beer and spirits seem to have small differences in consumption.

Correlation of Alcohol Type to Total Consumption

```
consumption_by_type <- lm(total_litres_of_pure_alcohol ~ beer_servings + spirit_servings + wine_serving
summary(consumption_by_type)
```

```
##
## Call:
## lm(formula = total_litres_of_pure_alcohol ~ beer_servings + spirit_servings +
##     wine_servings, data = etoh_consumption)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8670 -0.6865 -0.4010 -0.0392  7.4990
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.722450   0.149223   4.841 2.67e-06 ***
## beer_servings   0.018303   0.001253  14.608  < 2e-16 ***
## spirit_servings 0.015558   0.001244  12.511  < 2e-16 ***
## wine_servings   0.016005   0.001440  11.112  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.349 on 189 degrees of freedom
## Multiple R-squared:  0.8742, Adjusted R-squared:  0.8722
## F-statistic: 437.7 on 3 and 189 DF,  p-value: < 2.2e-16
```

```
consumption_by_type_wo_beer <- lm(total_litres_of_pure_alcohol ~ spirit_servings + wine_servings, data =
summary(consumption_by_type_wo_beer) #omit beer
```

```
##
## Call:
## lm(formula = total_litres_of_pure_alcohol ~ spirit_servings +
##     wine_servings, data = etoh_consumption)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5040 -1.4235 -0.6563  0.8593  7.6736
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.514105   0.202338   7.483 2.63e-12 ***
## spirit_servings 0.023320   0.001636  14.252  < 2e-16 ***
## wine_servings   0.026575   0.001813  14.662  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.963 on 190 degrees of freedom
## Multiple R-squared:  0.7321, Adjusted R-squared:  0.7293
## F-statistic: 259.6 on 2 and 190 DF,  p-value: < 2.2e-16
```

```
consumption_by_type_wo_wine <- lm(total_litres_of_pure_alcohol ~ spirit_servings + beer_servings, data =
summary(consumption_by_type_wo_wine) #omit wine
```

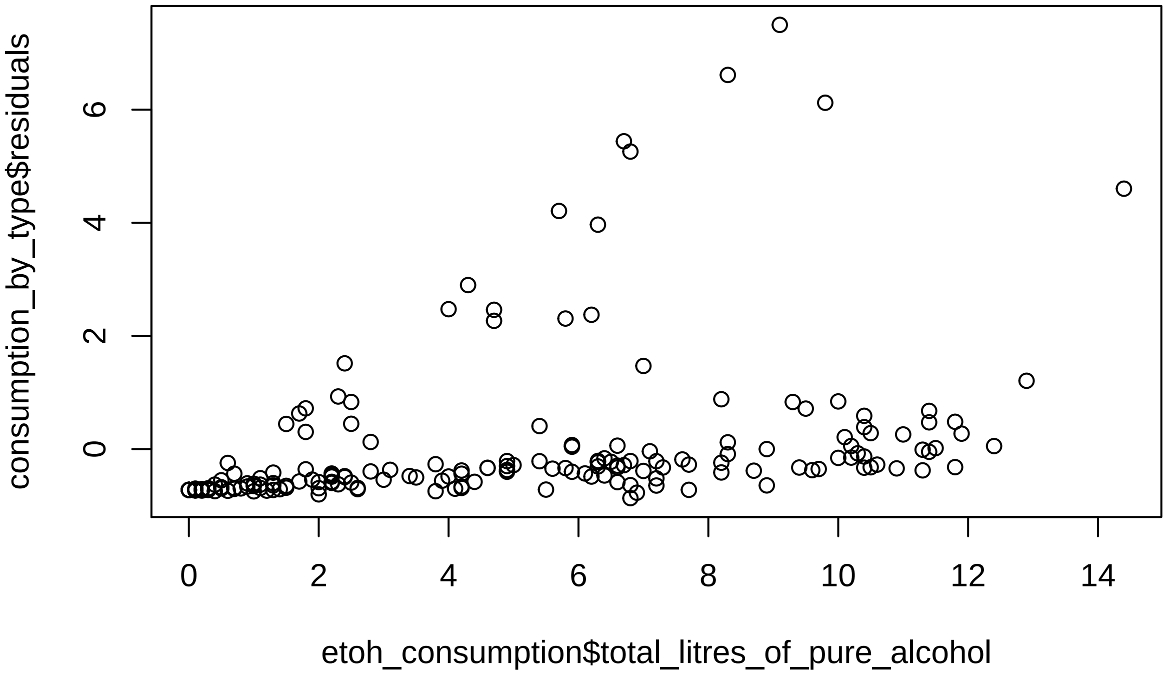
```
##
## Call:
## lm(formula = total_litres_of_pure_alcohol ~ spirit_servings +
##     beer_servings, data = etoh_consumption)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5968 -0.9130 -0.6667  0.5402  7.1228
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.841267   0.190873   4.407 1.75e-05 ***
## spirit_servings 0.014697   0.001592   9.233  < 2e-16 ***
## beer_servings   0.025296   0.001389  18.208  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.73 on 190 degrees of freedom
## Multiple R-squared:  0.792,  Adjusted R-squared:  0.7898
## F-statistic: 361.7 on 2 and 190 DF,  p-value: < 2.2e-16
```

```
consumption_by_type_wo_spirits <- lm(total_litres_of_pure_alcohol ~ wine_servings + beer_servings, data
summary(consumption_by_type_wo_wine) #omit spirits
```

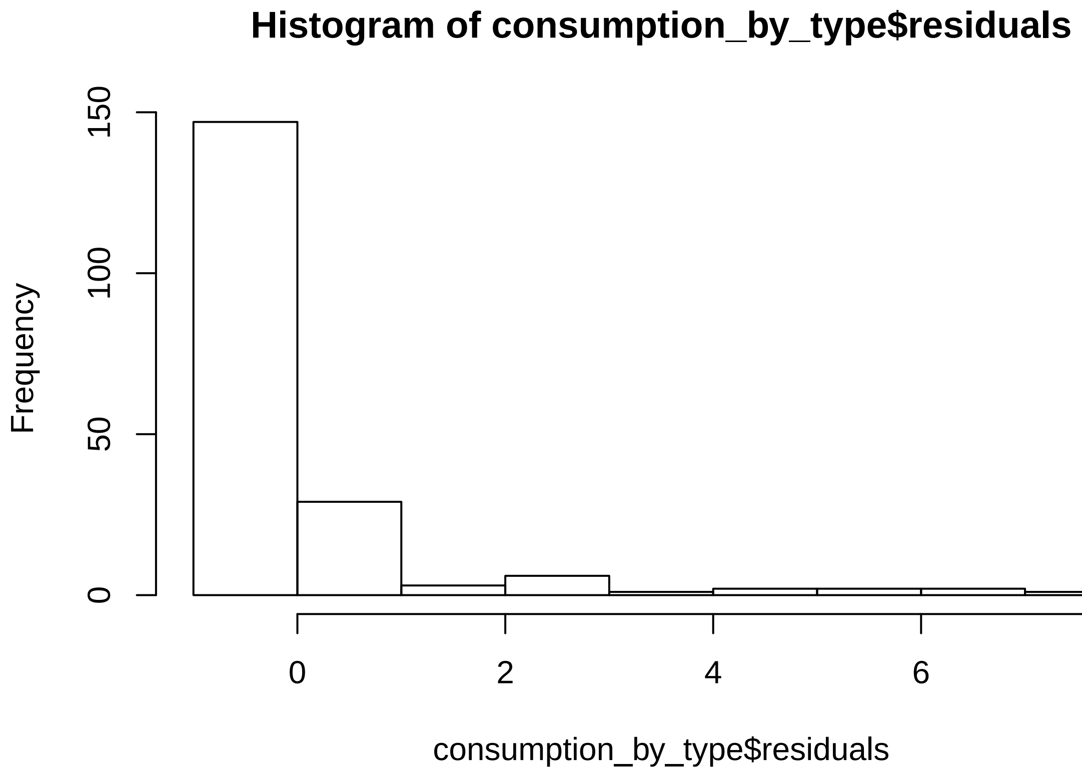
```
##
## Call:
## lm(formula = total_litres_of_pure_alcohol ~ spirit_servings +
##     beer_servings, data = etoh_consumption)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5968 -0.9130 -0.6667  0.5402  7.1228
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.841267   0.190873   4.407 1.75e-05 ***
## spirit_servings 0.014697   0.001592   9.233  < 2e-16 ***
## beer_servings   0.025296   0.001389  18.208  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.73 on 190 degrees of freedom
## Multiple R-squared:  0.792,  Adjusted R-squared:  0.7898
## F-statistic: 361.7 on 2 and 190 DF,  p-value: < 2.2e-16
```

After running a multivariate regression, the three variables are all similarly correlated to predicting overall alcohol consumption. As a second option, I omitted one variable from three different multivariate regressions to see if omitting a random variable improved the correlation coefficient. This did not improve the strength of the prediction as measured by adjusted R-squared. Next, I'll take a look at the residuals to find any reasons to question the underlying assumptions.

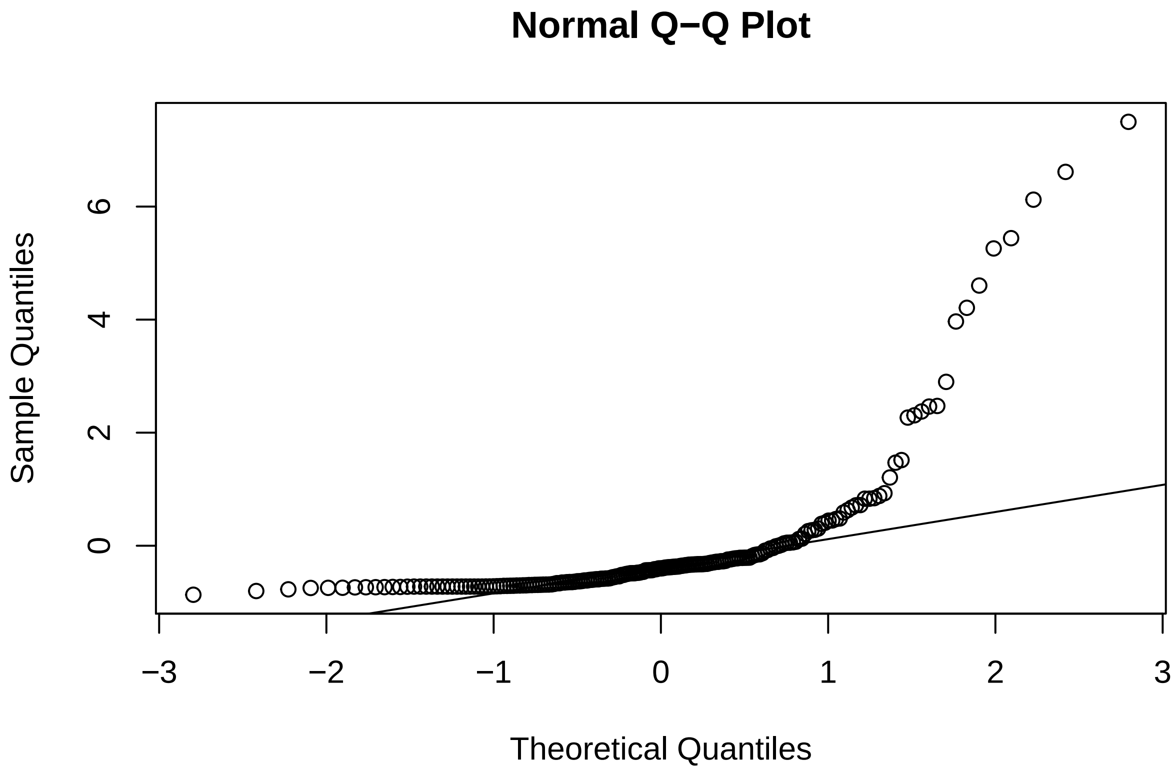
```
plot(consumption_by_type$residuals ~ etoh_consumption$total_litres_of_pure_alcohol)
```



```
hist(consumption_by_type$residuals)
```



```
qqnorm(consumption_by_type$residuals)
qqline(consumption_by_type$residuals)
```



Based on the various plots for residuals, the residuals are not random with respect to the dependent variable, and the residual histogram is very skewed to the right. This multivariate model does not hold up to the necessary underlying assumptions.

Single variable correlation - total alcohol consumption versus alcohol use disorders

As a final digression, I'll be looking at the diagnosis of alcohol use disorders as a dependent variable versus total alcohol consumption. The assumption of my own that I'm testing is that the higher the alcohol consmumption in a country, the more likely it is to interfere with the fabric of society.

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.6.3
```

```
library(lubridate)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
consumption_vs_disorder <- etoh_consumption %>%
  inner_join(etoh_use_disorders)
```

```
## Joining, by = "country"
```

```
## Warning: Column `country` joining factors with different levels, coercing to
## character vector
```

```
consumption_vs_disorder$both <- replace_na(consumption_vs_disorder$both, replace = 0)

#I felt justified in replacing NA's with '0' as the two countries in question, Monaco and Saudi Arabia,


sum(is.na(consumption_vs_disorder$both))
```

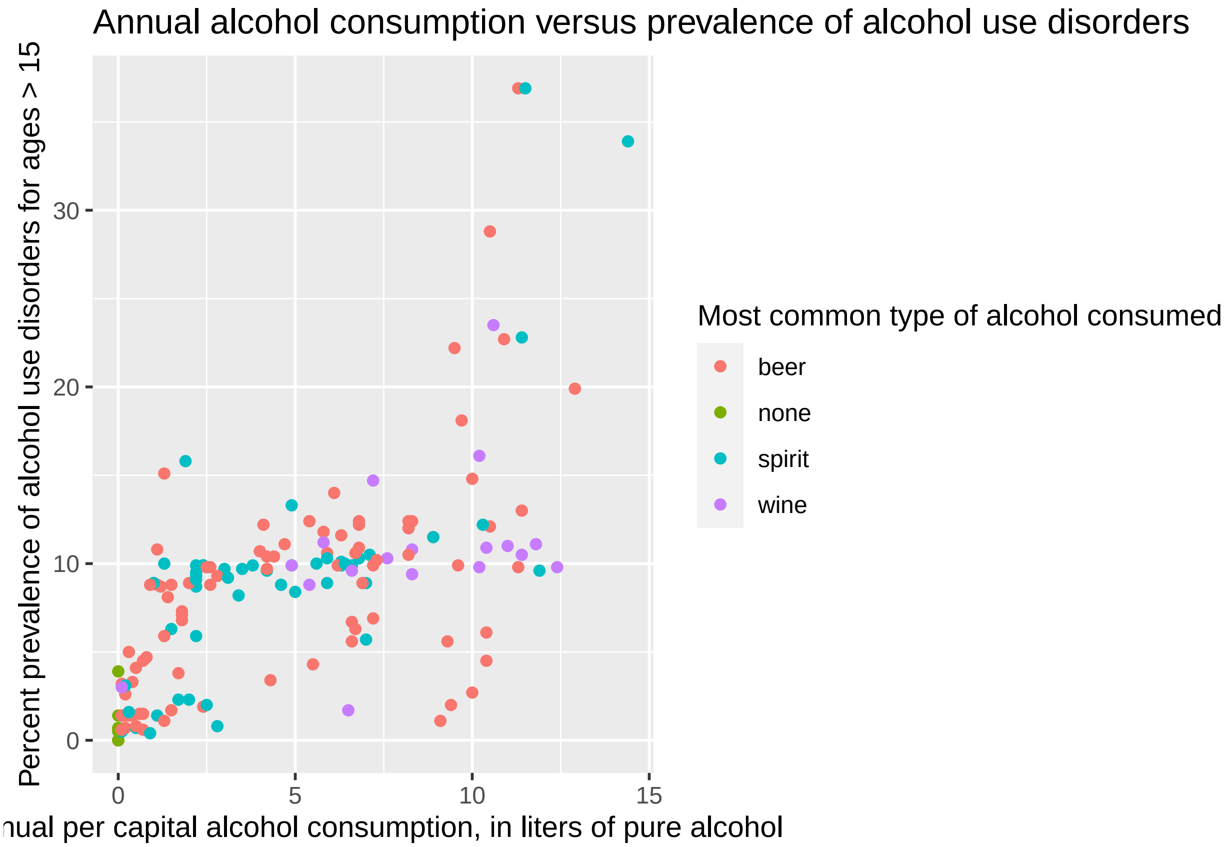
```
## [1] 0
```

```
consumption_vs_disorder <- consumption_vs_disorder %>%
  mutate(highest_type = case_when(beer_servings > spirit_servings & beer_servings > wine_servings ~ "be
                                  spirit_servings > beer_servings & spirit_servings > wine_servings ~ "
                                  wine_servings > beer_servings & wine_servings > spirit_servings ~ "wi
                                  TRUE ~ "none"))


consumption_vs_disorder_lm <- lm(both ~ total_litres_of_pure_alcohol, data = consumption_vs_disorder)
summary(consumption_vs_disorder_lm)
```

```
##
## Call:
## lm(formula = both ~ total_litres_of_pure_alcohol, data = consumption_vs_disorder)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2141  -2.8773  -0.3552   2.5453  21.1824
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.37267    0.58759    5.74 4.44e-08 ***
## total_litres_of_pure_alcohol  1.09247    0.09762   11.19  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.836 on 165 degrees of freedom
## Multiple R-squared:  0.4315, Adjusted R-squared:  0.4281
## F-statistic: 125.2 on 1 and 165 DF,  p-value: < 2.2e-16
```

```
ggplot(consumption_vs_disorder, aes(x = total_litres_of_pure_alcohol, y = both, color = highest_type)) +
         geom_point() +
  labs(title = "Annual alcohol consumption versus prevalence of alcohol use disorders", x = "Annual per
```



Increased alcohol use is mildly correlated with alcohol use disorders, with an R-squared value of approximately 0.43. I interpret the slope as meaning that on average, every additional liter of pure alcohol consumpmtion per capita is associated with an increase in 1% prevalence of alcohol use disorders in the population. Per the original fivethrityeight article, this is a little over 30 standard drinks - whether it is a can of beer, a glass of wine or shot of spirit.

I also attached a scatterplot that colored by the most popular seving of alcohol - beer, wine, spirits, or none for negligible consumption - and it appearsthat there exist some outliers with heavy drinking, which are preferential drinkers of beer and spirits.

Below is the residual analysis of this linear regression from the previous section.

```
plot(consumption_vs_disorder_lm$residuals ~ consumption_vs_disorder$total_litres_of_pure_alcohol, color
```

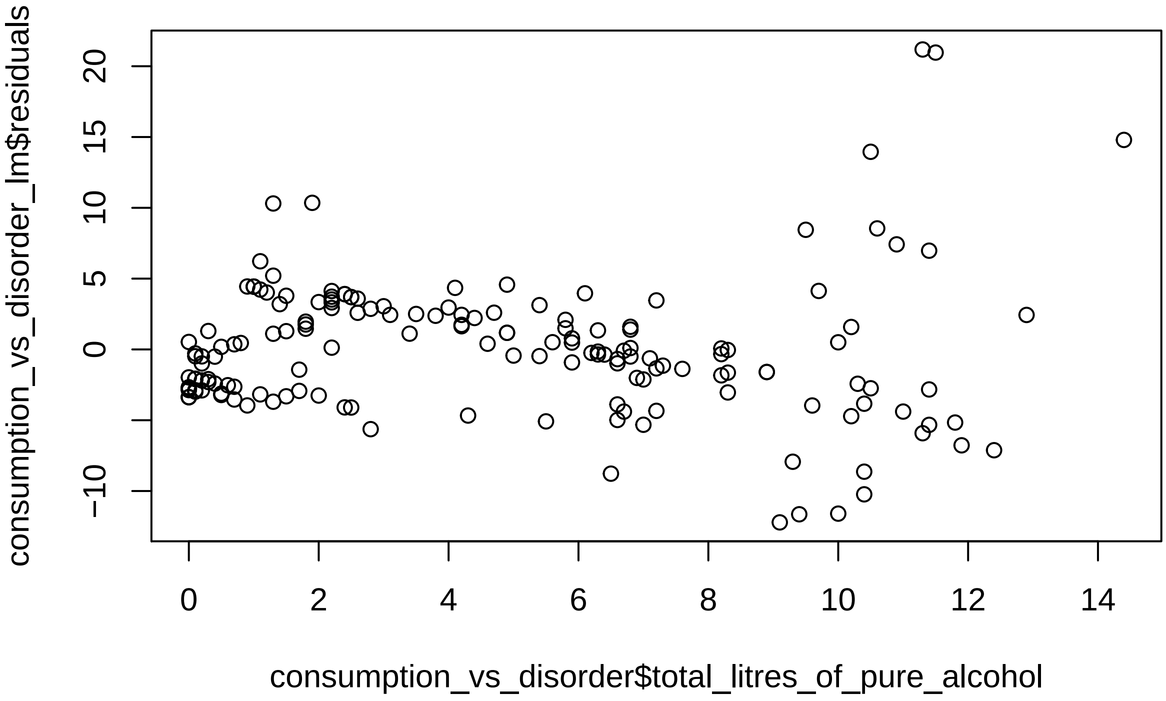
```
## Warning in plot.window(...): "color" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "color" is not a graphical parameter
```

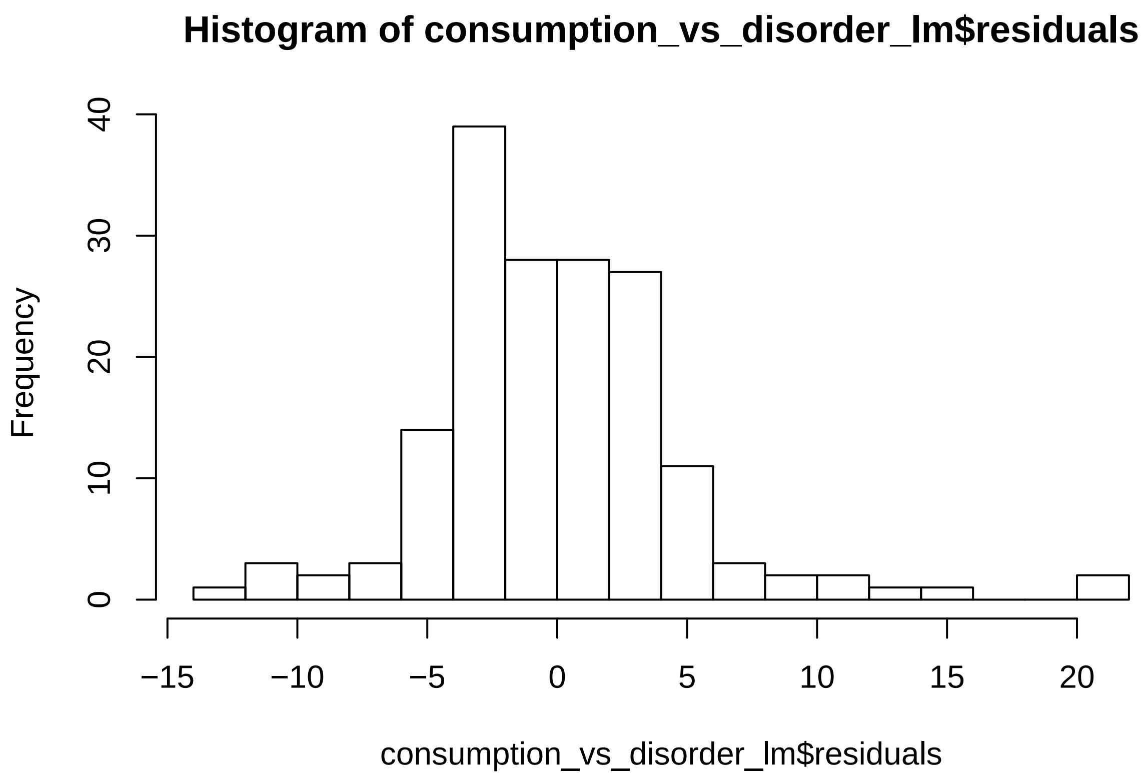
```
## Warning in axis(side = side, at = at, labels = labels, ...): "color" is not a
## graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "color" is not a
## graphical parameter
```

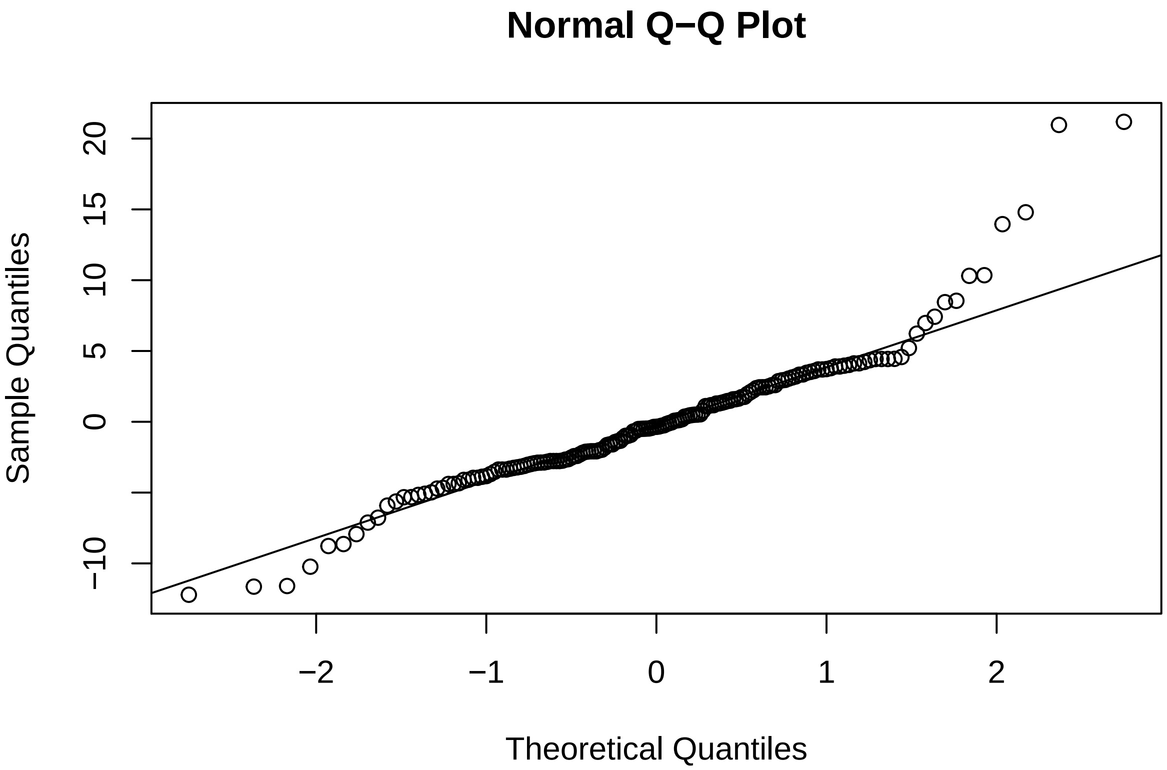
```
## Warning in box(...): "color" is not a graphical parameter
## Warning in title(...): "color" is not a graphical parameter
```



```r
hist(consumption_vs_disorder_lm$residuals, breaks = 12)
```



```
qqnorm(consumption_vs_disorder_lm$residuals)
qqline(consumption_vs_disorder_lm$residuals)
```



Based on the histogram of residuals, I would say the residuals are normally distributed and the residual plot has no major patterns. There are some outlying countries with high positive residuals, which means that the predicted values underestimate when compared to the observed values. Since these outliers exist on the higher side of the total litres, it's possible that these countries lack a public health framework which would naturally seek to minimize the negative health impacts on increased alcohol use.

Conclusions

My multivariate analysis of the type of alcohol favored in a given country and its impact on overall consumption did not meet the underlying assumptions for the residuals. Even after looking at variables to investigate collinearity, this did not improve the original model. Instead, I performed an additional analysis looking at alcohol consumption versus alcohol use disorders, and a linear relationship can be found between these two variables. When filtering for different alcohol type preferences, it's easier to visualize that most countries prefer beer, countries who prefer wine tend to drink middling amounts - neither low or high - and that spirits and beer represent the maximum values of consumption as well as alcohol use disorders. This model did meet the assumptions to accept the correlation between total consumption and disorders as valid. It is also worth mentioning that there are a handful of countries with low or no recorded alcohol consumption that may impact the model negatively.

Limitations and Further Research

In the future, it would be interesting to look at the patterns in drinking relative to gender, as well as looking at alcohol trends over time - specifically timed to economic downturns or major changes in alcohol and health regulations. It is also plausible to express alcohol not as total servings, but rather relative cost to consumer as a funciton of purchasing-power-parity, which the UN has records to support. On the one hand, this would indicate what percent of a household's income is being spent on alcohol, but would also confound the cost with governmental efforts to discourage alcohol overuse by taxation. Finally, identifying cultural norms that decrease overall alcohol consumption could help to separate countries where consumption is close to zero.