DATA 606 Data Project

Peter Gatica

Part 1 - Introduction

Abstract

The purpose of this project is to explore the possible reasons that citizens in countries around the world consume alcohol. After reviewing the alcohol consumption dataset, I did not see a meaningful way on how to run and interpret any linear models. I will admit that this may be due to my inexperience with data analytics. I found myself wondering what factors contribute to the consumption of alcohol in all countries. Is there a correlation between certain factors like economic freedom, religious freedom or freedom of expression, or overall happiness with one's life? I then thought that I would combine this dataset with a subset of data from the Human Freedom Index dataset for 2010. I downloaded both the datasets from FiveThirtyEight and the CATO Institute for 2010 respectively. I then manually combined variables from both datasets to use for my final project research. I matched the data from both sources by country, taking personal freedom, economic freedom, and the overall happiness scores from each country and, matching by country, I took beer, wine, spirit servings and total alcohol consumption in liters per person. I omitted any records that contained an N/A in fields to avoid inaccurate estimates in my linear models. I proceeded to run linear models on how happiness scores affect the amount of alcohol consumption for each country in the dataset. I did the same for each personal freedom and economic freedom scores. I also interpret the correlation coefficient and R-squared results for each model. I believe that this kind of research can be very useful around the world to identify countries that may have a propensity for alcohol abuse and create programs to help curtail alcoholism and perhaps other health and social issues that may result from alcohol abuse.

Part 2 - Data

Data Sources

FiveThirtyEight - The dataset that I am using as one of my sources was found that the FiveThirtyEight github link. This dataset on alcohol consumption by country for 2010 is the data behind the article Dear Mona Followup:Where Do People Drink The Most Beer, Wine and Spirits. The data was collected by the World Health Organization

The Human Freedom Index presents the state of human freedom in the world based on a broad measure that encompasses personal, civil, and economic freedom. Human freedom is a social concept that recognizes the dignity of individuals and is defined here as negative liberty or the absence of coercive constraint. Because freedom is inherently valuable and plays a role in human progress, it is worth measuring carefully. The Human Freedom Index is a resource that can help to more objectively observe relationships between freedom and other social and economic phenomena, as well as the ways in which the various dimensions of freedom interact with one another.

The report is co-published by the Cato Institute and the Fraser Institute.

Data collection

```
filename <- getURL("https://raw.githubusercontent.com/audiorunner13/Masters-Coursework/main/DATA606%20S
alc_hfi_2010 <- read.csv(text=filename)
```

```
alc_hfi_2010 <- na.omit(alc_hfi_2010)
```

Description of the dependent variable (what is being measured?)

The response variable is amount of alcohol (liters) consumed per person by country in 2010 and variable is numerical.

Description of the independent variable (what is being measured?, include at least 2 variables)

The explanatory variables are happiness score (hf_score), personal freedom score (pf_score), and economic freedom score (ef_score). Personal freedom and economic scores contribute to the overall happiness score of a country's citizens. All are numerical.

Research question

Does the happiness factor score of a country's citizens affect the amount of alcohol consumed by that country? How do personal expression and economic freedoms affect the amount of alcohol consumed by an individual in certain countries?

Type of study

This is an observational study.

```
summary(alc_hfi_2010)
```

Summary Statistics of source dataset

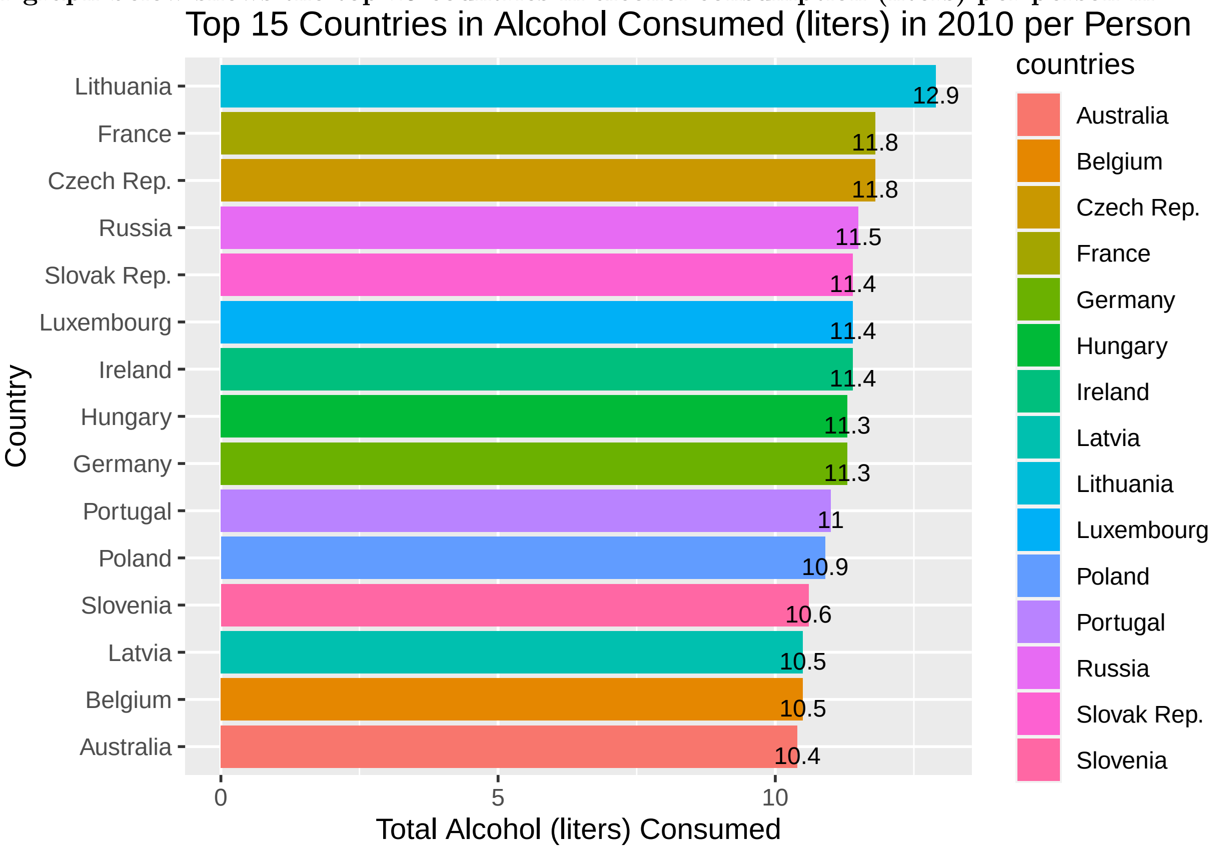
```
##       year       ISO_code          countries            region
##  Min.   :2010   Length:147         Length:147         Length:147
##  1st Qu.:2010   Class :character   Class :character   Class :character
##  Median :2010   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2010
##  3rd Qu.:2010
##  Max.   :2010
##      pf_rol      pf_religion_restrictions  pf_religion    pf_expression_control
##  Min.   :3.100   Min.   : 3.056           Min.   :4.291   Min.   :0.750
##  1st Qu.:4.309   1st Qu.: 6.250           1st Qu.:7.188   1st Qu.:3.750
##  Median :5.100   Median : 7.778           Median :8.096   Median :5.250
##  Mean   :5.465   Mean   : 7.455           Mean   :7.943   Mean   :5.325
##  3rd Qu.:6.517   3rd Qu.: 8.889           3rd Qu.:8.952   3rd Qu.:7.250
##  Max.   :8.700   Max.   :10.000           Max.   :9.944   Max.   :9.250
##  pf_expression      pf_score        pf_rank       ef_money_inflation
##  Min.   :3.269   Min.   :4.489   Min.   :  1.00   Min.   :4.188
##  1st Qu.:6.855   1st Qu.:6.353   1st Qu.: 38.50   1st Qu.:8.718
##  Median :8.138   Median :7.318   Median : 77.00   Median :9.238
##  Mean   :7.849   Mean   :7.311   Mean   : 76.92   Mean   :9.033
##  3rd Qu.:9.128   3rd Qu.:8.563   3rd Qu.:115.50   3rd Qu.:9.642
##  Max.   :9.750   Max.   :9.562   Max.   :153.00   Max.   :9.869
##  ef_money_currency    ef_money         ef_score       ef_rank
##  Min.   : 0.000    Min.   :1.972   Min.   :3.96   Min.   :  2.00
##  1st Qu.: 5.000    1st Qu.:6.947   1st Qu.:6.24   1st Qu.: 38.00
##  Median :10.000    Median :8.245   Median :6.85   Median : 77.00
##  Mean   : 6.531    Mean   :8.038   Mean   :6.75   Mean   : 77.18
##  3rd Qu.:10.000    3rd Qu.:9.305   3rd Qu.:7.35   3rd Qu.:115.50
##  Max.   :10.000    Max.   :9.887   Max.   :8.76   Max.   :153.00
##     hf_score        hf_rank        hf_quartile    beer_servings
##  Min.   :4.909   Min.   :  2.00   Min.   :1.000   Min.   :  0.0
##  1st Qu.:6.405   1st Qu.: 39.50   1st Qu.:1.500   1st Qu.: 27.0
##  Median :6.950   Median : 77.00   Median :2.000   Median : 85.0
##  Mean   :7.030   Mean   : 76.98   Mean   :2.497   Mean   :119.7
##  3rd Qu.:7.868   3rd Qu.:115.00   3rd Qu.:3.500   3rd Qu.:204.5
##  Max.   :8.879   Max.   :153.00   Max.   :4.000   Max.   :376.0
##  spirit_servings  wine_servings    total_litres_of_pure_alcohol
##  Min.   :  0.00   Min.   :  0.00   Min.   : 0.000
##  1st Qu.:  8.00   1st Qu.:  1.00   1st Qu.: 1.800
##  Median : 69.00   Median :  9.00   Median : 4.900
##  Mean   : 82.77   Mean   : 55.63   Mean   : 5.159
##  3rd Qu.:128.50   3rd Qu.: 82.50   3rd Qu.: 8.200
##  Max.   :326.00   Max.   :370.00   Max.   :12.900
```

Part 3 - Exploratory data analysis

Appropriate Visualizations

```
head(alc_hfi_2010[order(-alc_hfi_2010$total_litres_of_pure_alcohol),],15) %>%
    ggplot(aes(y=reorder(countries,total_litres_of_pure_alcohol),x=total_litres_of_pure_alcohol,fill=cou
      geom_bar(stat = 'identity',position=position_dodge()) +
      geom_text(aes(label=total_litres_of_pure_alcohol), vjust=1.0, color="black",
            position = position_dodge(0.9), size=3.0) +
      labs(x = ("Total Alcohol (liters) Consumed"),y = ("Country"),
      title = ("Top 15 Countries in Alcohol Consumed (liters) in 2010 per Person")  )
```

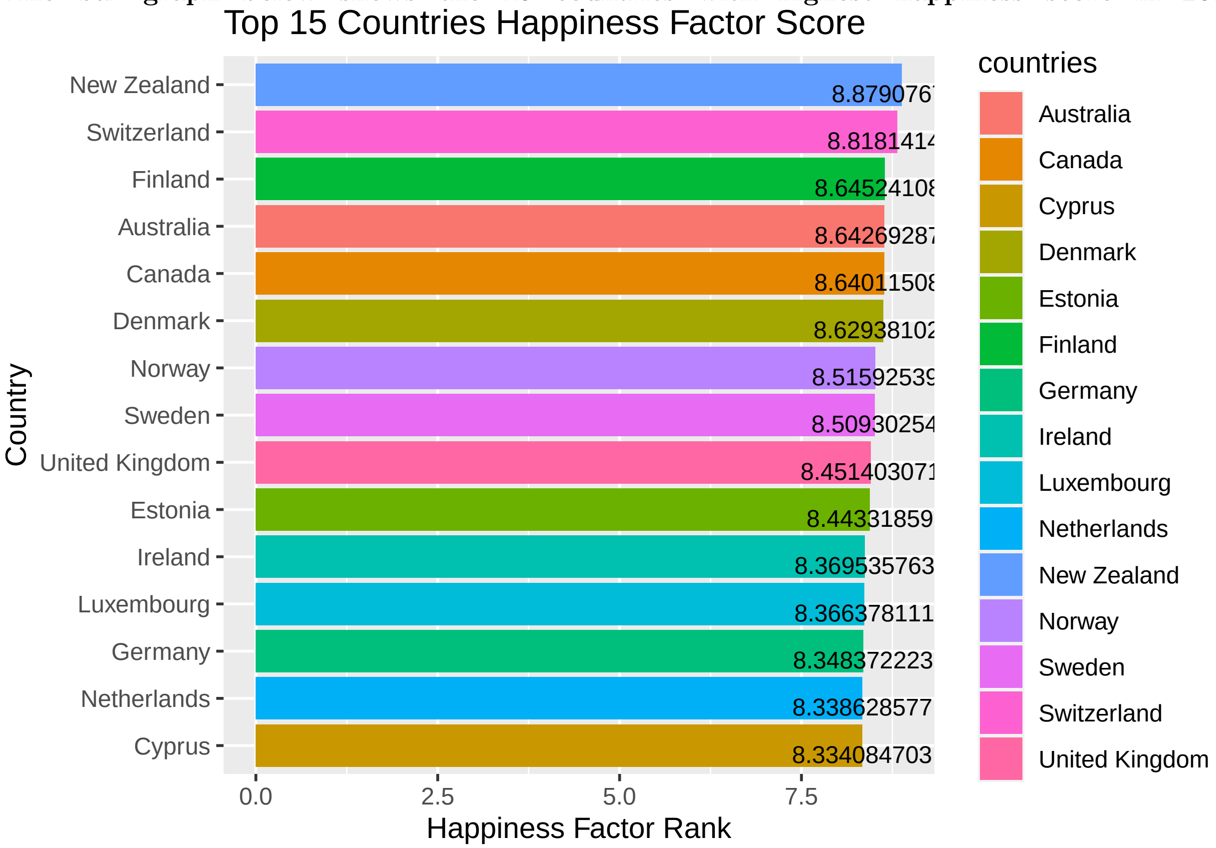
The bar graph below shows the top 15 countries in alcohol consumption (liters) per person in 2010.



```
      theme_minimal()
```

```
head(alc_hfi_2010[order(-alc_hfi_2010$hf_score),],15) %>%
    ggplot(aes(y=reorder(countries,hf_score),x=hf_score,fill=countries)) +
      geom_bar(stat = 'identity',position=position_dodge()) +
      geom_text(aes(label=hf_score), vjust=1.0, color="black",
            position = position_dodge(0.9), size=3.0) +
      labs(x = ("Happiness Factor Rank"),y = ("Country"),
      title = ("Top 15 Countries Happiness Factor Score")  )
```

The bar graph below shows the 15 countries with highest happiness score in 2010.

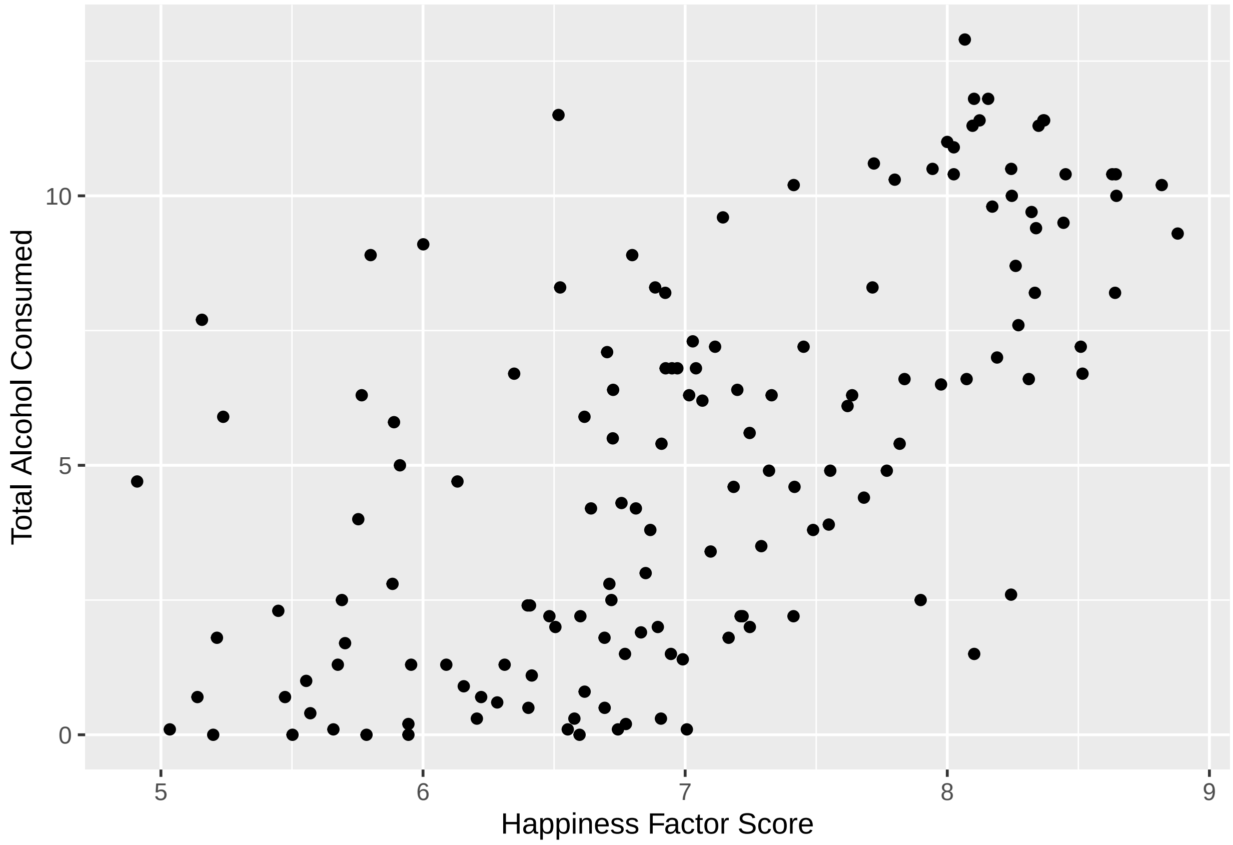


```
      theme_minimal()
```

Statistical Output

```
ggplot(data = alc_hfi_2010, aes(x = hf_score, y = total_litres_of_pure_alcohol)) + geom_point() +
  labs(x = ("Happiness Factor Score"),y = ("Total Alcohol Consumed"))
```

An initial glance at the scatterplot below does shows a possible linear correlation between the happiness factor and the total alcohol consumed (liters) per person. The plots have a wide spread and are not tightly packed. The relationship looks linear, we can quantify the strength of the relationship with the correlation coefficient.



```
    theme_minimal()
```

```
alc_hfi_2010 %>%
  summarise(cor(hf_score, total_litres_of_pure_alcohol, use = "complete.obs"))
```

```
##   cor(hf_score, total_litres_of_pure_alcohol, use = "complete.obs")
## 1                                                         0.6465383
```

The correlation coefficient is moderately strong at 65%. The calculated R^2 which is a more reliable indicator of the correlation is a moderate 42%. Next let's see if these values change when we run a linear model on the dataset.

The Happiness Factor

Is the amount of a country's alcohol consumption affected by how happy it's citizens are?

In my dataset is a happiness factor score of countries around the world. This happiness factor score is based on certain variables such as freedom of expression, freedom of religion and economic freedom to name a few. There many more variables that go into the happiness factor score but for purposes of this project I will only consider the ones I just mentioned and extracted.

```
(mod_hf_alc <- lm(total_litres_of_pure_alcohol ~ hf_score, data = alc_hfi_2010))
```

The happiness factor to alcohol linear model

```
##
## Call:
## lm(formula = total_litres_of_pure_alcohol ~ hf_score, data = alc_hfi_2010)
##
## Coefficients:
## (Intercept)     hf_score
##     -12.191        2.468
```

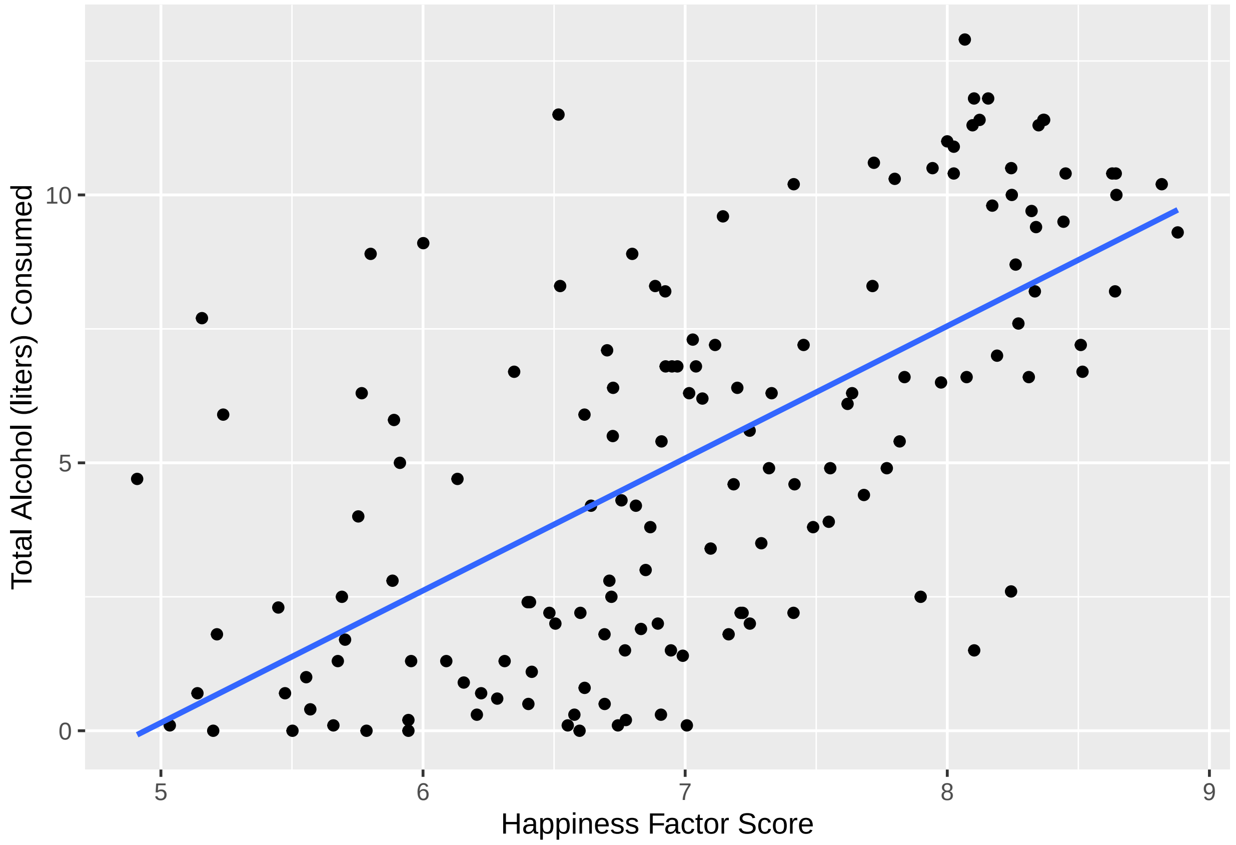
```
summary(mod_hf_alc)
```

```
##
## Call:
## lm(formula = total_litres_of_pure_alcohol ~ hf_score, data = alc_hfi_2010)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3058 -2.0845 -0.4203  1.8400  7.6080
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12.1911     1.7161  -7.104 5.05e-11 ***
## hf_score      2.4679     0.2418  10.205  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.831 on 145 degrees of freedom
## Multiple R-squared:  0.418,  Adjusted R-squared:  0.414
## F-statistic: 104.1 on 1 and 145 DF,  p-value: < 2.2e-16
```

```
ggplot(data = alc_hfi_2010, aes(x = hf_score, y = total_litres_of_pure_alcohol)) +
  geom_point() + stat_smooth(method = "lm", se = FALSE) +
  labs(x = ("Happiness Factor Score"),y = ("Total Alcohol (liters) Consumed"))
```

Running a linear model verifies that there is a slightly moderate correlation of 41.4% between a country's happiness score and the amount of alcohol that its citizens consume. The scatterplot with the least sum of the squares line can illustrate that correlation. It does show a positive linear relationship, however, based on how the points are not tightly packed along the blue line suggests that there is not a overly strong correlation. Let's next look at the correlation between alcohol consumption and personal freedom.

```
## `geom_smooth()` using formula 'y ~ x'
```



Personal Freedom Factor

```
alc_hfi_2010 %>%
  summarise(cor(pf_score, total_litres_of_pure_alcohol, use = "complete.obs"))
```

When considering the personal freedom score alone, the correlation coefficient is stronger at 69.5%. The calculated R^2 which is a more reliable indicator of the correlation increases to 48.3%.

```
##   cor(pf_score, total_litres_of_pure_alcohol, use = "complete.obs")
## 1                                                          0.6954629
```

```
(mod_pf_alc <- lm(total_litres_of_pure_alcohol ~ pf_score, data = alc_hfi_2010))
```

The personal freedom factor alcohol linear model

```
##
## Call:
## lm(formula = total_litres_of_pure_alcohol ~ pf_score, data = alc_hfi_2010)
##
```

```
## Coefficients:
## (Intercept)     pf_score  
##      -9.344        1.984
```

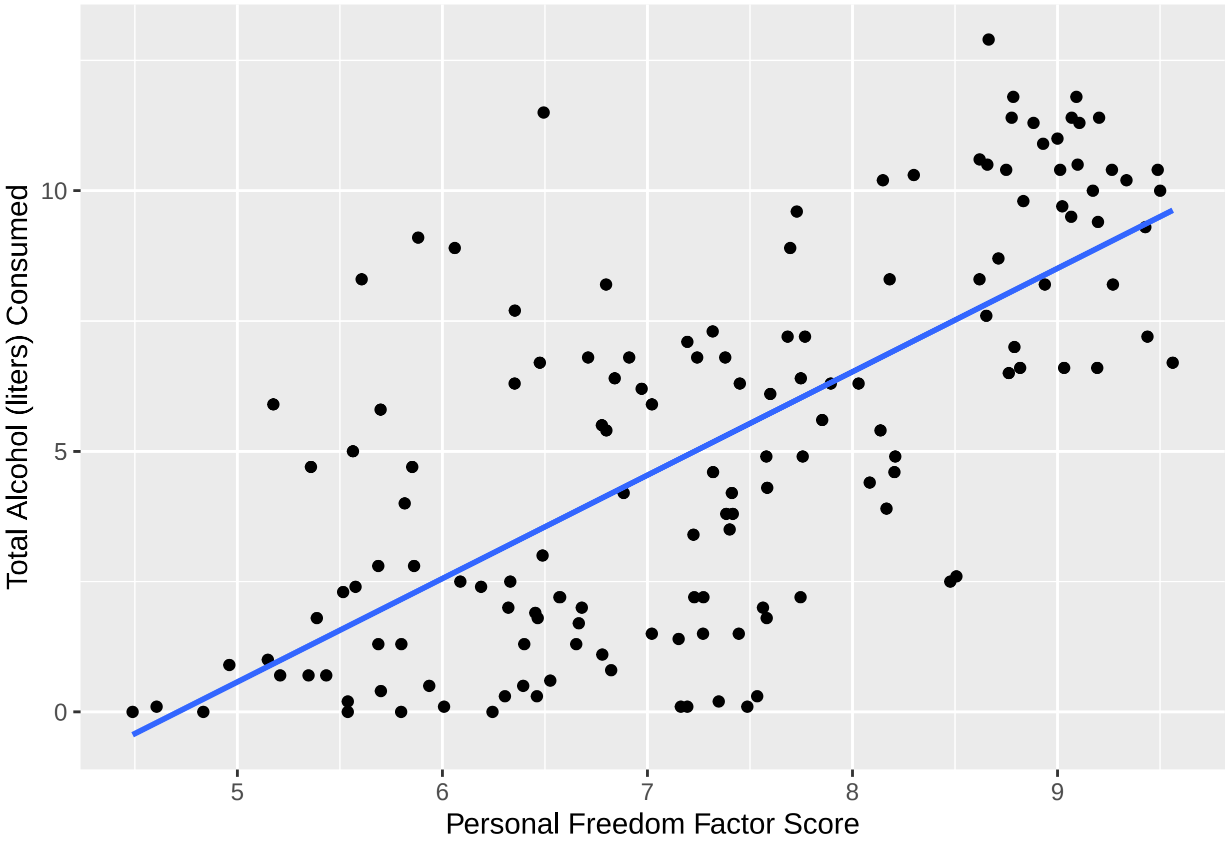
```
summary(mod_pf_alc)
```

```
## 
## Call:
## lm(formula = total_litres_of_pure_alcohol ~ pf_score, data = alc_hfi_2010)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5.4087 -1.9192 -0.1873  1.8010  7.9620 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -9.3442     1.2637  -7.394 1.05e-11 ***
## pf_score      1.9838     0.1702  11.655  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.666 on 145 degrees of freedom
## Multiple R-squared:  0.4837, Adjusted R-squared:  0.4801 
## F-statistic: 135.8 on 1 and 145 DF,  p-value: < 2.2e-16
```

```
ggplot(data = alc_hfi_2010, aes(y = total_litres_of_pure_alcohol, x = pf_score)) +
  geom_point() +  stat_smooth(method = "lm", se = FALSE) +
  labs(x = ("Personal Freedom Factor Score"),y = ("Total Alcohol (liters) Consumed"))
```

Running a linear model on the personal freedom score alone verifies that there is a more moderate R^2 of 48% between a country's personal freedom score and the amount of alcohol that its citizens consume. The scatterplot with the least sum of the squares line does show the points are little more tightly packed along the blue line suggests a more constant variability than the happiness factor correlation. Finally, let's explore the correlation between alcohol consumption and economic freedom.

```
## `geom_smooth()` using formula 'y ~ x'
```



Economic Freedom Factor

```
alc_hfi_2010 %>%
  summarise(cor(ef_score, total_litres_of_pure_alcohol, use = "complete.obs"))
```

Considering the economic freedom score alone, the correlation coefficient is stronger at 41.6%. The calculated R^2 which is a more reliable indicator of the correlation drops to a very 17% suggesting a very low correlation between economic freedom and the amount of alcohol that a person consumes.

```
##   cor(ef_score, total_litres_of_pure_alcohol, use = "complete.obs")
## 1                                                          0.415568
```

```
(mod_ef_alc <- lm(total_litres_of_pure_alcohol ~ ef_score, data = alc_hfi_2010))
```

The economic factor alcohol linear model

```
##
## Call:
## lm(formula = total_litres_of_pure_alcohol ~ ef_score, data = alc_hfi_2010)
```

```
##
## Coefficients:
## (Intercept)     ef_score
##      -7.118        1.819
```

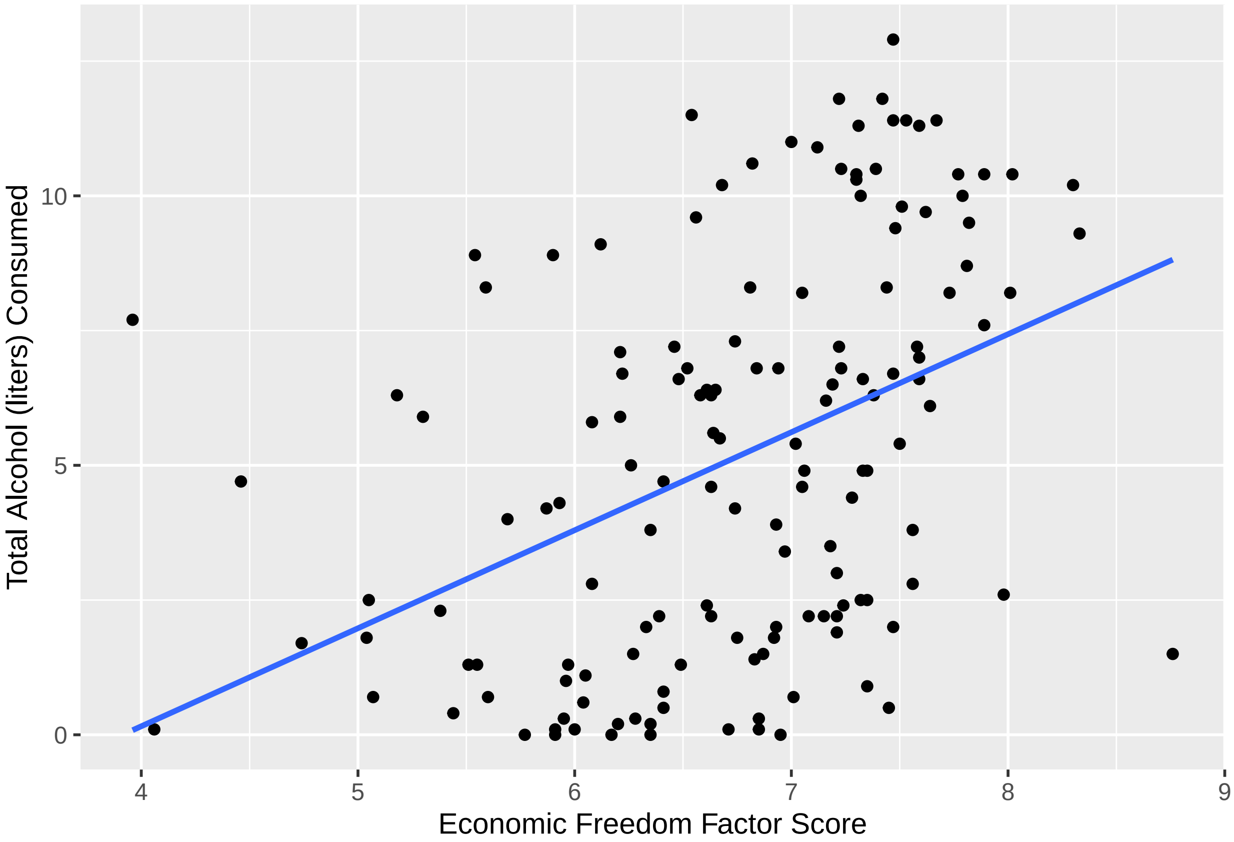
```
summary(mod_ef_alc)
```

```
##
## Call:
## lm(formula = total_litres_of_pure_alcohol ~ ef_score, data = alc_hfi_2010)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3156 -3.1323  0.1963  2.5361  7.6150
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.1177     2.2488  -3.165  0.00189 **
## ef_score      1.8189     0.3306   5.502 1.66e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.375 on 145 degrees of freedom
## Multiple R-squared:  0.1727, Adjusted R-squared:  0.167
## F-statistic: 30.27 on 1 and 145 DF,  p-value: 1.657e-07
```

```
ggplot(data = alc_hfi_2010, aes(y = total_litres_of_pure_alcohol, x = ef_score)) +
  geom_point() +   stat_smooth(method = "lm", se = FALSE) +
  labs(x = ("Economic Freedom Factor Score"),y = ("Total Alcohol (liters) Consumed"))
```

Running a linear model on the economic freedom score alone verifies that there is a very weak R^2 of 16.7% between a country's economic freedom score and the amount of alcohol that its citizens consume. The scatterplot with the least sum of the squares line show this weak correlation. One can see that the points are widely scattered and do not run a long the length of the blue as do the prior two models.

```
## `geom_smooth()` using formula 'y ~ x'
```



Part 4 - Conclusion

I was surprised to find that the happiness factor for each country to did not have a much stronger correlation to the amount of alcohol consumed by its citizens. It was only a moderate correlation. I had expected to see a tighter gathering around the sum of the least squares line showing that people with a lower happiness factor drank more. This was not the case in any of the scenarios as shown by the higher negative residuals at the lower ends of the plots. I also expected that the higher a happiness factor would decrease the amount of alcohol consumed by persons, however, it was the opposite. The higher the happiness factor the more positive residuals are.

Why is the analysis important?

I believe that this kind of research can be very useful around the world to identify countries that may have a propensity for alcohol abuse and create programs to help curtail alcoholism and perhaps other health and social issues that may result from alcohol abuse.

Limitations of the analysis?

I have to say that the limitations of the analysis is reflective of my experience with data analystics and visualization. With more experience, I could easily apply the mutiple variable method of linear modeling for more accurate determination of the correlation between the

different happiness index factors and a country's alcohol consumption by its citizens. I certainly would delve more into which of the different factors are the greater contributors to alcohol consumption. My attempt is a very elementary one at best.

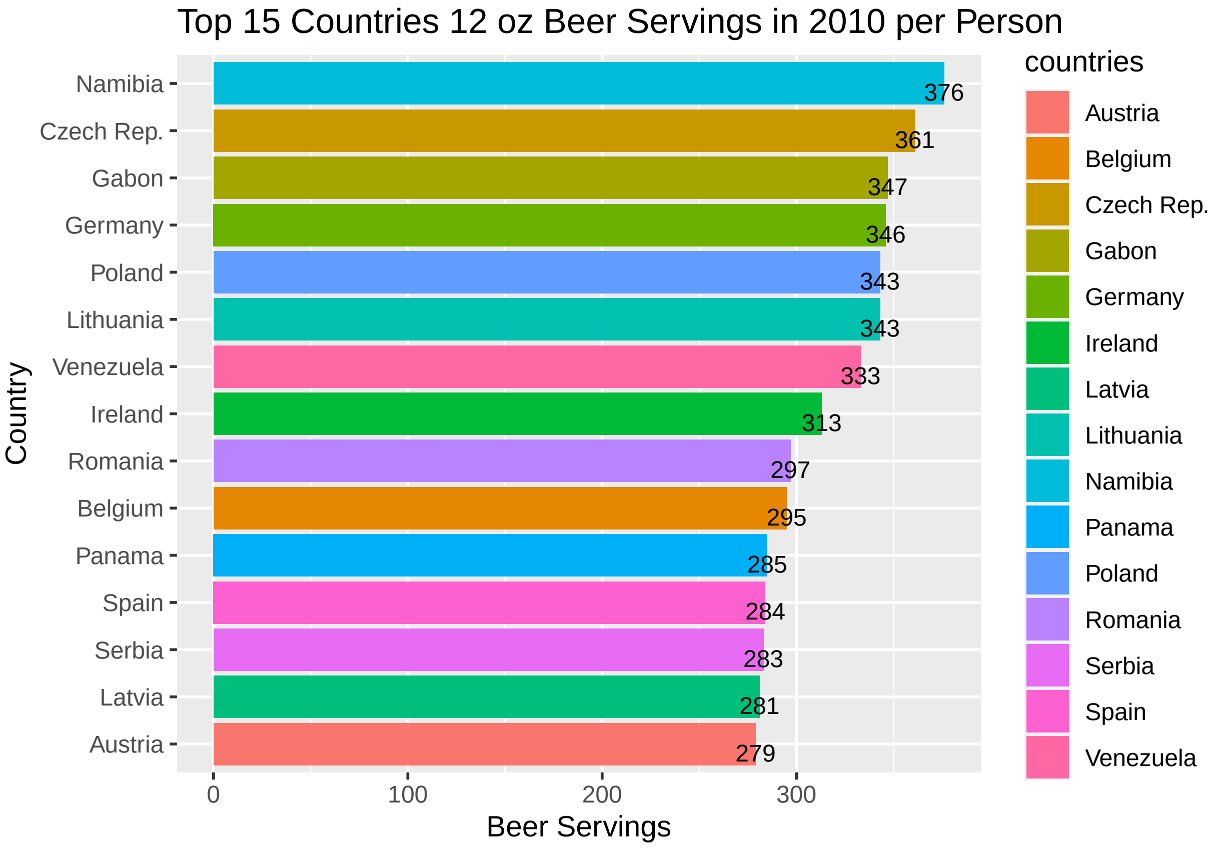
References

Dear Mona is Mona Chalabi, a former contributor on FiveThirtyEight posting articles that answer readers' questions as well as postings regarding data and data analytics.

Appendix

#####Bar graph of the top 15 countries with the highest beer consumption per person in 2010.

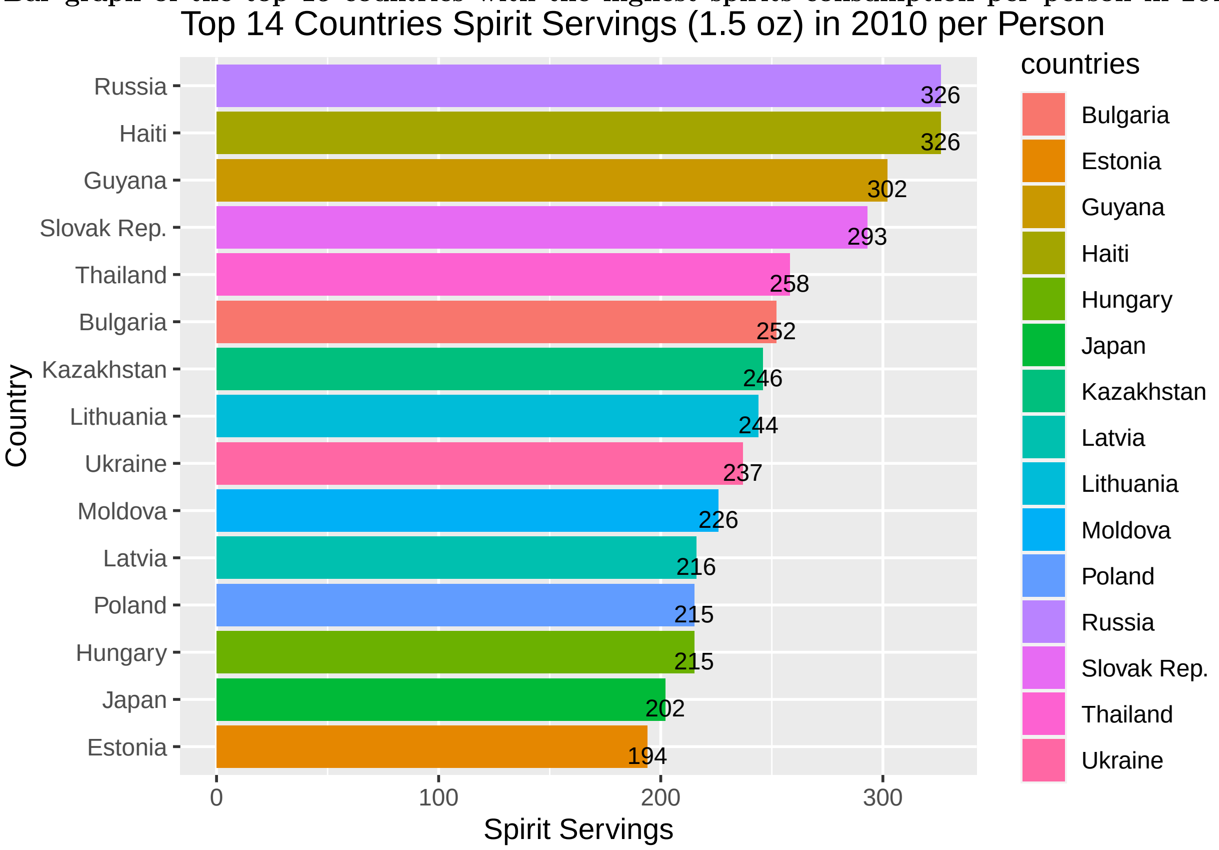
```
head(alc_hfi_2010[order(-alc_hfi_2010$beer_servings),],15) %>%
    ggplot(aes(y=reorder(countries,beer_servings),x=beer_servings,fill=countries)) +
      geom_bar(stat = 'identity',position=position_dodge()) +
      geom_text(aes(label=beer_servings), vjust=1.0, color="black",
            position = position_dodge(0.9), size=3.0) +
      labs(x = ("Beer Servings"),y = ("Country"),
      title = ("Top 15 Countries 12 oz Beer Servings in 2010 per Person")  )
```



```
      theme_minimal()
```

```
head(alc_hfi_2010[order(-alc_hfi_2010$spirit_servings),],15) %>%
    ggplot(aes(y=reorder(countries,spirit_servings),x=spirit_servings,fill=countries)) +
      geom_bar(stat = 'identity',position=position_dodge()) +
      geom_text(aes(label=spirit_servings), vjust=1.0, color="black",
            position = position_dodge(0.9), size=3.0) +
      labs(x = ("Spirit Servings"),y = ("Country"),
      title = ("Top 14 Countries Spirit Servings (1.5 oz) in 2010 per Person")  )
```

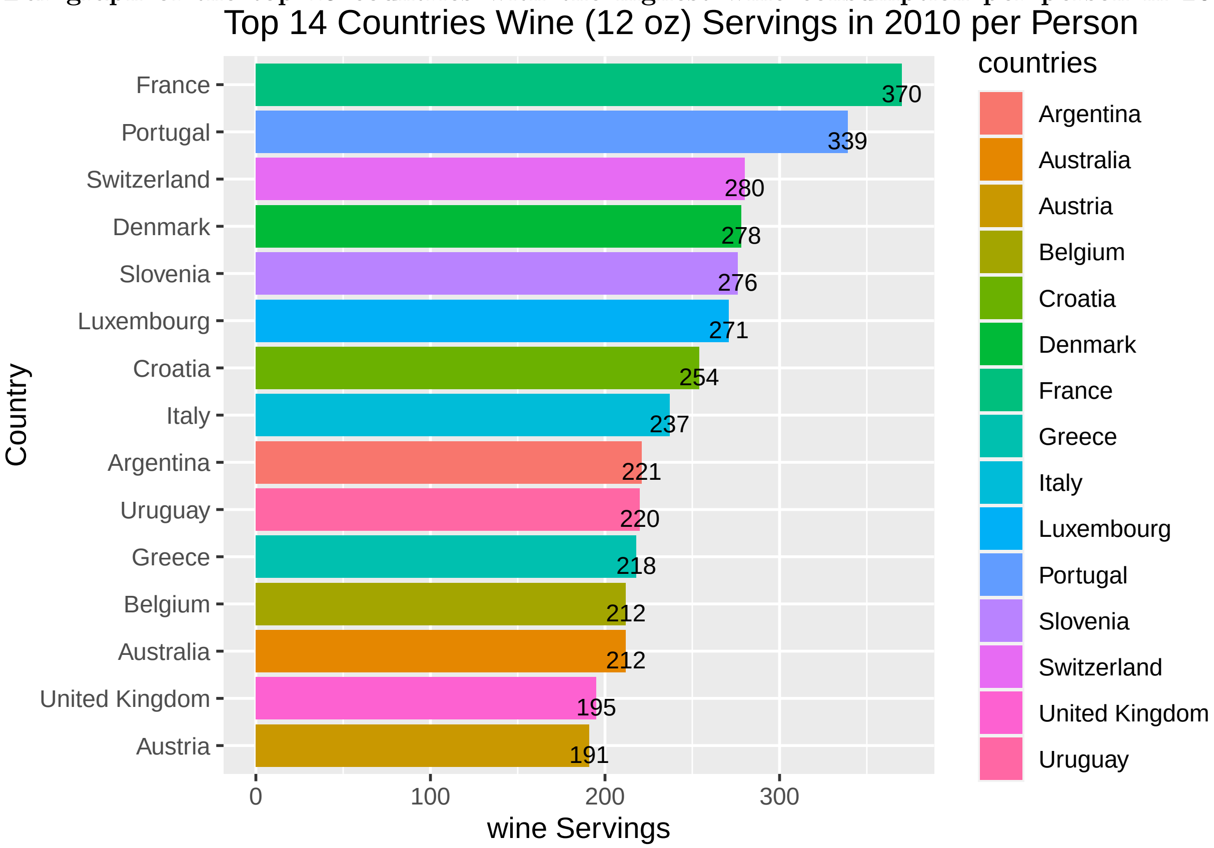
Bar graph of the top 15 countries with the highest spirits consumption per person in 2010.



```
      theme_minimal()
```

```
head(alc_hfi_2010[order(-alc_hfi_2010$wine_servings),],15) %>%
    ggplot(aes(y=reorder(countries,wine_servings),x=wine_servings,fill=countries)) +
      geom_bar(stat = 'identity',position=position_dodge()) +
      geom_text(aes(label=wine_servings), vjust=1.0, color="black",
            position = position_dodge(0.9), size=3.0) +
      labs(x = ("wine Servings"),y = ("Country"),
      title = ("Top 14 Countries Wine (12 oz) Servings in 2010 per Person")  )
```

Bar graph of the top 15 countries with the highest wine consumption per person in 2010.



```
    theme_minimal()
```