MTHM501: Working with Data

To what extent does a country's alcohol consumption effect its happiness level for both sexes in the years 2015 and 2019?

Marc Xu

Word Count: 2291

Contents

1. Introduction 3
2. Objectives 3
3. Data 4
Wrangling Datasets 4
Dealing With The Offset In Observations 5
4. Results 6
Handling Missing Data 6
Visualization of Data 7
Map Plots 7
Scatter Plots 8
Hypothesis Testing 9
Pearson's Correlation 9
T-Test 9
Covariance 9
5. Limitations 10
6. Conclusion 10
7. Appendix 11
Code: 11
..... 12
..... 14
8. Works Cited List 15

1. Introduction

Alcohol has been prevalent throughout the entirety of history, dating all the way back to as early as 10,000BC, where the discovery of jugs in the late Stone Age suggested that beverages were fermented with the intention for consumption (Durham, 26). Gradually, the concept of fermentation stretched through different civilizations resulting in the prominent role of alcohol consumption in modern day society.

We can categorize the population of alcohol users based on their primary purposes of consumption. First, we have the social drinker, in which the user consumes moderate levels of alcohol to enhance the social experience but does not create disturbances in their lives or create health issues. For example, social outgoings where many individuals are consuming alcohol making you feel compelled to join. Second, we have the business drinker, which is relatively similar to social drinking, however business drinking alcohol consumption may now range from moderate to excess. Business drinkers are often a by-product of peer pressure, as in many cultures it is socially rude to turn down drinks from a peer of higher level, usually referring to bosses and managers, which is why business drinking may result in excess drinking. Lastly, we have drinking as a coping mechanism. Drinkers of this category use the features of alcohol in excess to tackle an issue they face in real life. Through scientific research it is understood that alcohol is a nervous system depressant, when consumed, even in minimal quantities, it slows down the nervous system decreasing brain activity creating a sense of relaxation ("Alcohol and the Nervous System"). Often, the decrease in brain activity is enough for the consumer to forget about the burden of having to confront real life issues.

2. Objectives

This then raises the question, *to what extent does a country's alcohol consumption effect its happiness level*? Throughout this report, data will be gathered regarding a country's alcohol consumption and its respective happiness level to then come to a conclusion to whether the consumption of alcohol has any correlation to a country's happiness level. In order to reach the conclusion, it is necessary to first:

1. Find the appropriate data
2. Wrangle the data such that is in a presentable form for further analysis
3. Evaluate the possibility of missing values and how we deal with them
4. Use tools from R to analyse the data to give data meaning
5. Evaluate the results of the analysis to then contextualize against the context of this paper

3. Data

Data was extracted The Global Health Observatory by The World Health Organisation that recorded the yearly average alcohol consumption per capita by country and sex in litres of pure alcohol of people with age 15 or older. This data is recorded from the production, import, export and alcohol sale taxation numbers whilst taking into consideration of external consumption factors such as alcohol consumed by tourism. Hence, this data is viewed highly as an accurate indicator for a nation's alcohol consumption.

Secondly, data was extracted from the World Happiness Report to provide data on the happiness of a country. This data is surveyed in each country with a population sample ranging from 1000-3000 participants. Given the subjective nature of survey sampling and the small population, the accuracy and validity of this data set is to be questioned.

Wrangling Datasets

```
alcoholConsumptionByCapita = read_xlsx(path = "Alcohol Consumption by GDP Per Capita.xlsx")
newAlcoholConsumptionByCapita = rename(alcoholConsumptionByCapita,
                                       Region = "ParentLocation",
                                       CountryCode = "SpatialDimValueCode",
                                       Country = "Location",
                                       Year = "Period",
                                       Sex = "Dim1",
                                       MeanConsumption = "FactValueNumeric",
                                       CILowerBoundConsumption = "FactValueNumericLow",
                                       CIUpperBoundConsumption = "FactValueNumericHigh",
)
newAlcoholConsumptionByCapita = newAlcoholConsumptionByCapita[,c("Region",
                                                                 "CountryCode",
                                                                 "Country",
                                                                 "Year",
                                                                 "Sex",
                                                                 "MeanConsumption",
                                                                 "CILowerBoundConsumption",
                                                                 "CIUpperBoundConsumption")]
newAlcoholConsumptionByCapita = mutate(newAlcoholConsumptionByCapita,
                                       Region = as.factor(Region),
                                       CountryCode = as.factor(CountryCode),
                                       Country = as.factor(Country),
                                       Sex = as.factor(Sex))
newAlcoholConsumptionByCapita[newAlcoholConsumptionByCapita == 0] = NA
finalAlcoholConsumptionByCapita = subset(newAlcoholConsumptionByCapita,
                                         Year != 2000 &
                                         Year != 2005 &
                                         Year != 2010)
finalAlcoholConsumptionByCapita = subset(finalAlcoholConsumptionByCapita, Sex != "Female" &
Sex != "Male")
finalAlcoholConsumptionByCapita =
finalAlcoholConsumptionByCapita[order(finalAlcoholConsumptionByCapita$Country),]
```

The process to wrangle Alcohol Consumption data started through the first line of code reading the excel file. I then evaluated which of the variables were valuable to the analysis and renamed it such that it is easier to work with. After renaming variables we eliminated the unnecessary variables by creating a vector with the desired variables. Furthermore, 4 of the variables were changed to the data type factor as it would make it easier to interpret the data when a summary was called. For example, if the variable "Sex" was kept in the original form of a double, a summary of the data would show us statistical properties of the variable instead of listing the 3 possible values of "Sex" and informing us with the frequency of the 3 possible values for "Sex". Further on, we interpreted all values equating to exactly 0 as missing values. This is because there will always be a portion of the population consuming alcohol, meaning that a value of 0 litres of alcohol consumed per year is highly

unfeasible. Lastly, we removed the years 2000, 2005 and 2010 along with the sexes female and male and ordered the countries alphabetically, as this level of detailed data was not available in the Happiness Report.

A near identical process was applied to the data supplied from the Happiness Report.

Dealing With The Offset In Observations

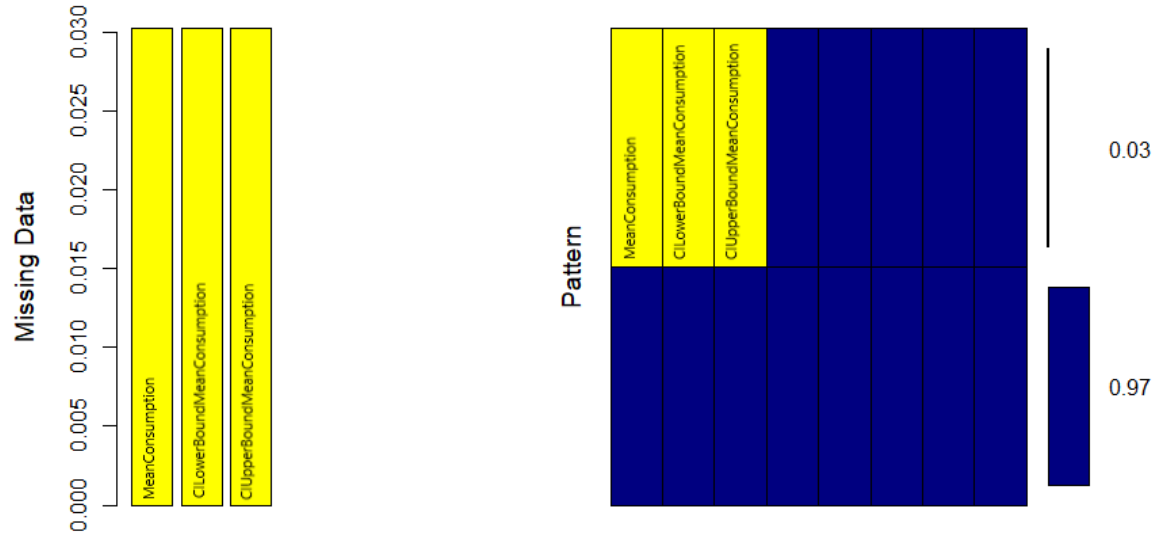
```
print(anti_join(newHappinessIndex2015, newHappinessIndex2019, by="Country"))
df1 = anti_join(newHappinessIndex2015, newHappinessIndex2019, by="Country")
finalNewHappinessIndex2015 = dplyr::setdiff(newHappinessIndex2015, df1)


print(anti_join(newHappinessIndex2019, newHappinessIndex2015, by="Country"))
df2 = anti_join(newHappinessIndex2019, newHappinessIndex2015, by="Country")
finalNewHappinessIndex2019 = dplyr::setdiff(newHappinessIndex2019, df2)


print(anti_join(finalAlcoholConsumptionByCapita, newHappinessIndex2015, by="Country"), n=112)
mismatchedCountries2015 = anti_join(finalAlcoholConsumptionByCapita, newHappinessIndex2015, by="Country")

print(anti_join(finalAlcoholConsumptionByCapita, newHappinessIndex2019, by="Country"), n=86)
mismatchedCountries2019 = anti_join(finalAlcoholConsumptionByCapita, newHappinessIndex2019, by="Country")


levels(finalAlcoholConsumptionByCapita$Country)[match("Democratic People's Republic of Korea",
                                                      levels(finalAlcoholConsumptionByCapita$Country))] = "North Korea"


finalFinalAlcoholComsumptionByCapita = dplyr::setdiff(finalAlcoholConsumptionByCapita, mismatchedCountries2015)
finalFinalAlcoholComsumptionByCapita = dplyr::setdiff(finalAlcoholConsumptionByCapita, mismatchedCountries2019)
```

The next step was hard to notice but I realized there was an offset between the number of observations between the Alcohol dataset and the Happiness Report dataset, meaning that within all datasets the number of countries were different. I tackled this issue by using the anti_join function. This function was used both ways to first test for differences within the 2015 and 2019 data on happiness and then setdiff from the dplyr package was used to remove the differences. When the test for differences was compiled against the alcohol data set, I noticed a discrepancy in the format the names were presented in. For example, Congo (Kinshasa) and Congo (Brazzaville) were recognized by their official names, Congo and the Democratic Republic of Congo instead of their simplified names, which was specifically difficult as I was not explicitly aware of the two geographical locations of Congo. The code snippet above only showed 1 example (North Korea) however there were a total of 20 country names to be replaced (to view the entire list of names changed, refer to appendix with code). This task was relatively difficult as it tested my geographical skills and understanding of country names globally, and if a mistake were to be made I would be omitting important data scrutinizing the validity of this paper. Finally, once all the datasets were changed to the same format setdiff was used to eliminate the differences.

After wrangling the data state beforehand, an educated decision has been made to only consider the years: 2015 and 2019 with no distinction between male and female due to the limitations of the dataset provided from the World Happiness Report. This narrows our question of interest to; *To what extent does a country's alcohol consumption effect its happiness level for both sexes in the years 2015 and 2019?*

4. Results

Handling Missing Data



To further understand the state of missing numbers in my dataset, I first plotted a diagram to visualize missing data. As shown above the alcohol data set only has 3% missing data, narrowing down to 1% if we focus on the only variable interested.

```
imputationData = finalFinalAlcoholComsumptionByCapita
imputationData$MeanConsumption[which(is.na(imputationData$MeanConsumption))] = mean(imputationData$MeanConsumption, na.rm =TRUE)
imputationData$CILowerBoundConsumption[which(is.na(imputationData$CILowerBoundConsumption))] = mean(imputationData$CILowerBoundConsumption, na.rm =TRUE)
imputationData$CIUpperBoundConsumption[which(is.na(imputationData$CIUpperBoundConsumption))] = mean(imputationData$CIUpperBoundConsumption, na.rm =TRUE)

my_imp = mice(finalFinalAlcoholComsumptionByCapita, m=5, method = "pmm", maxit = 20)

my_imp$imp$MeanConsumption
my_imp$imp$CILowerBoundConsumption
my_imp$imp$CIUpperBoundConsumption

summary(finalFinalAlcoholComsumptionByCapita$MeanConsumption)
summary(finalFinalAlcoholComsumptionByCapita$CILowerBoundConsumption)
cleanDataSet = complete(my_imp, 1)
```

To tackle the issue of missing data I've decided to use the MICE package. MICE is a package that can multiply impute data given that the data is continuous and not discrete. We are trying to impute mean consumption values for alcohol and not year nor sex, hence MICE can be applied to our dataset. The first step to multiply impute data using MICE is to carry out a simple imputation replacing all missing data with the mean of the respective columns. For example, we first replaced all the missing data in the mean consumption column by its mean. MICE then regress the mean consumption with the happiness level which is then iterated 20 times to give 5 sets a value for us to choose which most appropriate. The method of regression is specified by the method "pmm" which stands for predictive mean matching. It calculates the predicted value of desired variable and compares itself to the set of indices next to the missing data. For example, if the missing data is of index 10, it takes the mean and compares it to the indices +5 of 10 and -5. Note that the range may generally differ from 3 all the way to 10.

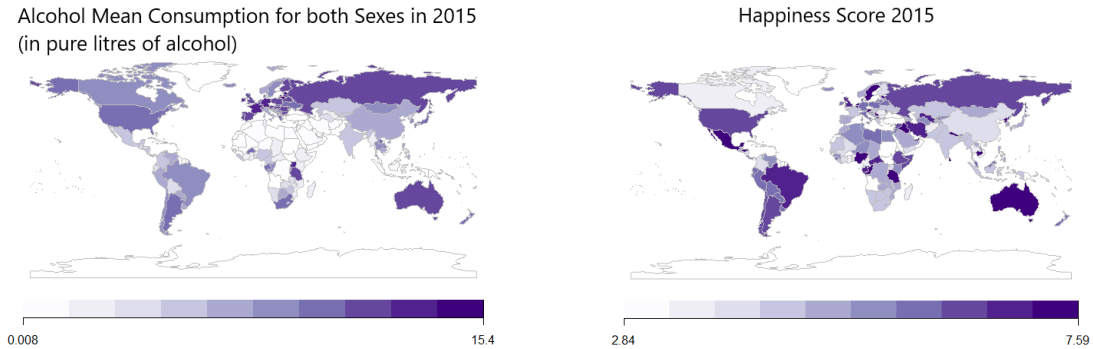
```
> my_imp$imp$MeanConsumption
        1     2     3     4     5
19  0.031 12.99 0.045 12.99 12.79
143 0.034 15.40 0.031 13.05 12.78
144 0.045 15.40 0.034 12.99 15.40
173 0.031 12.78 0.034 15.40 12.79
174 0.008 15.40 0.045 15.40 12.78
233 0.027 12.79 0.045 12.79 12.78
234 0.034 12.99 0.027 12.79 12.79
247 0.045 15.40 0.031 15.40 12.78
248 0.027 12.79 0.027 12.78 12.99
> my_imp$imp$CILowerBoundConsumption
        1     2      3     4      5
19  0.0110 10.90 0.0001 10.95 0.0150
143 0.0001 11.13 0.0150 10.95 0.0110
144 0.0001 10.95 0.0010 10.95 0.0010
173 0.0110 13.30 0.0150 11.13 0.0010
174 0.0010 13.30 0.0150 11.14 0.0010
233 0.0150 10.90 0.0001 11.13 0.0001
234 0.0001 11.14 0.0110 11.14 0.0150
247 0.0110 11.14 0.0001 11.14 0.0001
248 0.0010 10.90 0.0010 13.30 0.0110
```

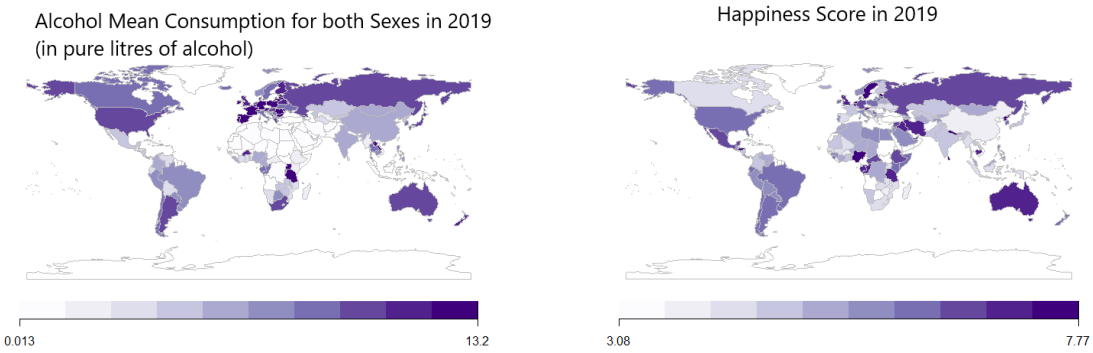
```
> summary(finalFinalAlcoholComsumptionByCapita$MeanConsumption)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.008   2.230   5.900   6.063   9.630  15.400       9

> summary(finalFinalAlcoholComsumptionByCapita$CILowerBoundConsumption)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 0.0001  1.6100  4.6800  4.9845  8.2900 13.3000       9
```

In the context of our imputation, MICE was only able to impute for 2 of the 3 columns, mean consumption and confidence interval lower bound consumption. Upon further research, users have claimed that if the data being imputed is too collinear MICE is unable to complete the imputation, hence missing values for confidence interval upper bound consumption, however this is not a big issue in our report as we are only interested in the mean consumption. As shown in the figure above, the 5 columns show the possible values for the missing data. We now have to pick one compared against the mean, at the first glance I considered columns 1, 3 and 5, as they were two extreme ends that were relatively equidistant from the mean. I ended choosing the values on the lower end, the first imputation column to be specific, as when I re-inspected the countries in question that we were imputing I noticed that countries such as Bangladesh, Mauritania, Somalia, etc were all under the Brandt Line. A line which conceptualizes the gap between the rich northern hemisphere and the poor southern hemisphere.

Visualization of Data

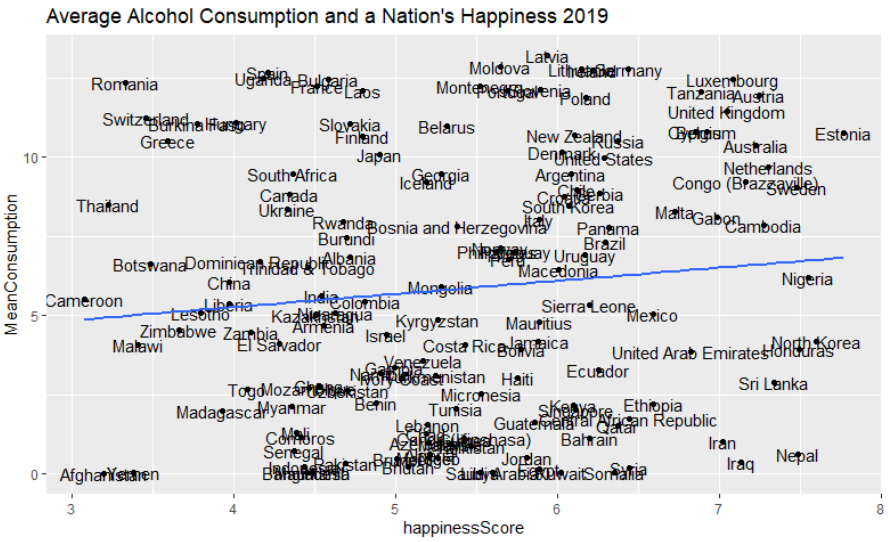
Map Plots

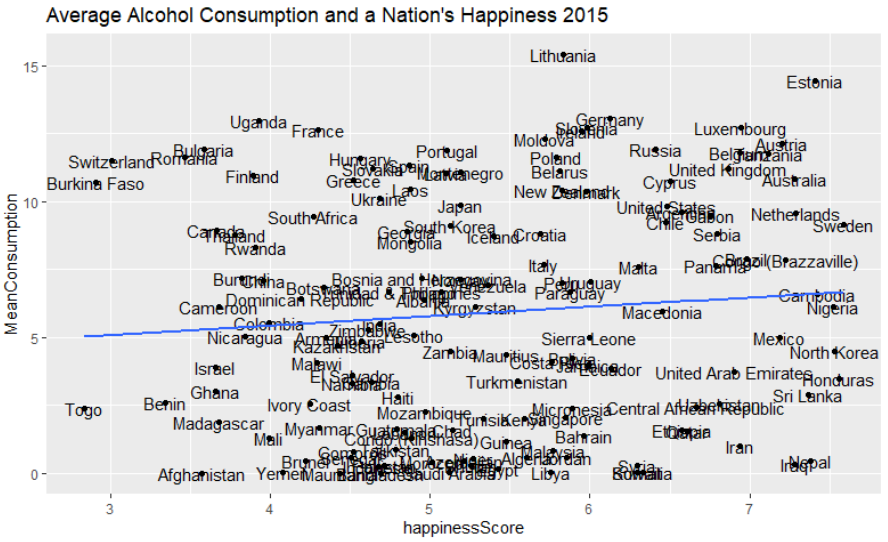




A map plot has been created using the RWorldMap package to gain better insights regarding the research question. From the 2015 data we can see that in some continents a higher average alcohol consumption correlated to a high happiness score, for example North America (with outliers in Canada and Mexico), South America, Asia (to a certain extent) and Oceania.

Scatter Plots





Scatter plots with regression lines were also plotted to get a better idea of the relationship between a nation's average alcohol consumption and its happiness score. From the regression line alone we notice that even though the line is relatively flat, the slope is minimally positive, suggesting a very weak relationship between mean consumption and happiness score.

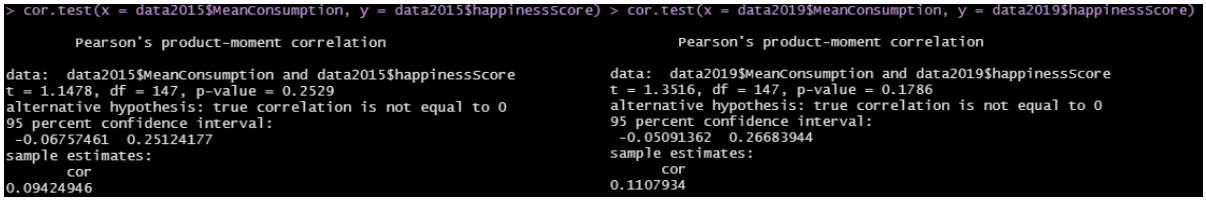
Hypothesis Testing

Hypothesis: average consumption does have a positive relationship between happiness

Null hypothesis: average consumption does *not* have a positive relationship between happiness.

Pearson's Correlation

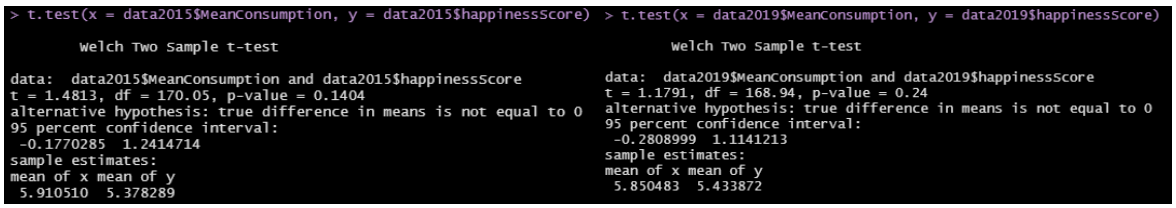
2015 Data (left) 2019 Data (right):



As stated earlier it could be told that the Pearson's correlation for both data sets would end up positive. This means that there is a positive relationship between the two datasets, meaning that if there was an increase in consumption then there will be an increase in happiness. However, this increase is very minimal as the Pearson's correlation is extremely close to 0.

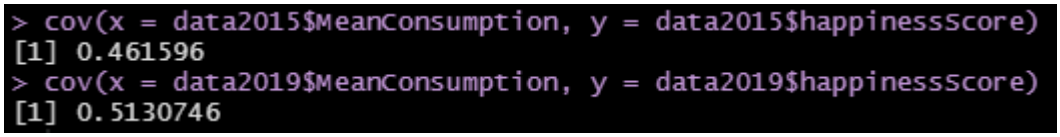
T-Test

2015 Data (left) 2019 Data (right):



The 2015 dataset scored a t score of 1.4813 whilst the 2019 dataset scored a t score of 1.1791. Given that the t scores are relatively small, this means that our hypothesis is true in which average consumption does have a positive relationship between happiness and we can reject our null hypothesis that average consumption does *not* have a positive relationship between happiness.

Covariance



As shown in the figure above the covariance for both datasets are positive meaning that they will change together in a similar fashion. This also suggests that an increase in one variable will lead to an increase in the other variable.

After our analysis we can say that average alcohol consumption does have a positive relationship with the happiness score.

5. Limitations

When extracting my data I noticed the inconsistency of the happiness report, in the sense that it did not provide information explicitly compared to the alcohol data. Upon reading the happiness report, there was no information regarding the population sampled, and as my alcohol data only applied to people of age 15 or older, if the sampled population in the happiness report contained people with ages 14 or younger, this would highly invalidate my research as my control variable is no longer controlled.

Furthermore, when imputing the missing data, given the large range of values given for the average consumption of alcohol, I believe the first imputation with the mean highly invalidates this piece of work. This is because with intuition on the nation's GDP you can tell some values imputed were very inconsiderate given the context of the country. For example, in places such as Bangladesh and Somalia, two countries with missing data where we know as a fact have relatively low GDP and low wealth, it is unexpected for the average alcohol consumption to be a high value. To avoid this issue, if the first imputed value could be weighed against a country's wealth or GDP to level out the multiple imputation, I believe it would highly increase the validity of our results.

6. Conclusion

To conclude, I can say that there is a positive relationship between the average alcohol consumption and a country's happiness, but the extent is relatively small, as proven by the hypothesis tests. The Pearson's correlation value was too close to 0 to conclude that the relationship is strong. The T-test scores were relatively small too in the context of our dataset; however, it couldn't prove the relationship was strong. And finally given the positive covariance it proved there was a positive relationship between average alcohol consumption and a country's happiness, but the value was too small to claim the relationship is strong.

7. Appendix

Code:

```
require(lme4)
require(lmerTest)
require(tidyverse)
require(readxl)
require(magrittr)
require(sjPlot)
require(sjmisc)
require(sjstats)
require(arm)
require(compare)
require(mice)
require(VIM)
require(dplyr)
require(rworldmap)
require(RColorBrewer)

#DATA WRANGLING

#Read excel file
alcoholConsumptionByCapita = read_xlsx(path = "Alcohol Consumption by GDP Per Capita.xlsx")

#Rename appropriate Variable Names
newAlcoholConsumptionByCapita = rename(alcoholConsumptionByCapita,
                                      Region = "ParentLocation",
                                      CountryCode = "SpatialDimValueCode",
                                      Country = "Location",
                                      Year = "Period",
                                      Sex = "Dim1",
                                      MeanConsumption = "FactValueNumeric",
                                      CILowerBoundConsumption = "FactValueNumericLow",
                                      CIUpperBoundConsumption = "FactValueNumericHigh",
                                      )
#Eliminating unnecessary columns by selecting the ones we wish to keep
newAlcoholConsumptionByCapita = newAlcoholConsumptionByCapita[,c("Region",
                                                                 "CountryCode",
                                                                 "Country",
                                                                 "Year",
                                                                 "Sex",
                                                                 "MeanConsumption",
                                                                 "CILowerBoundConsumption",
                                                                 "CIUpperBoundConsumption")]
#Changing variable types
newAlcoholConsumptionByCapita = mutate(newAlcoholConsumptionByCapita, Region = as.factor(Region),
                                                                      CountryCode = as.factor(CountryCode),
                                                                      Country = as.factor(Country),
                                                                      Sex = as.factor(Sex))
#Made assumption that all values of 0 meant missing values
newAlcoholConsumptionByCapita[newAlcoholConsumptionByCapita == 0] = NA

#Removed the years 2000, 2005 and 2010 and female and male from sex
#as this data isn't applicable to happiness report
#Then sorted Country column alphabetically
finalAlcoholConsumptionByCapita = subset(newAlcoholConsumptionByCapita,
                                         Year != 2000 &
                                         Year != 2005 &
                                         Year != 2010)
finalAlcoholConsumptionByCapita = subset(finalAlcoholConsumptionByCapita, Sex != "Female" & Sex != "Male")
finalAlcoholConsumptionByCapita = finalAlcoholConsumptionByCapita[order(finalAlcoholConsumptionByCapita$Country),]



#Read excel file
happinessIndex2015 = read_csv(file = "2015.csv")

#Rename appropriate Variable Names
newHappinessIndex2015 = rename(happinessIndex2015,
                               Rank = "Happiness Rank",
                               GDPperCapita = "Economy (GDP per Capita)",
```

```
                              lifeExpectancy = "Health (Life Expectancy)",
                              happinessScore = "Happiness Score",
                              )

#Changing variable types
newHappinessIndex2015 = mutate(newHappinessIndex2015, Country = as.factor(Country))


#Eliminating unnecessary columns by selecting the ones we wish to keep
newHappinessIndex2015 = newHappinessIndex2015[,c("Country",
                                                 "happinessScore",
                                                 "GDPperCapita",
                                                 "lifeExpectancy",
                                                 "Freedom",
                                                 "Generosity"
                                                 )]



#Read excel file
happinessIndex2019 = read_csv(file = "2019.csv")

#Rename appropriate Variable Names
newHappinessIndex2019 = rename(happinessIndex2019,
                              Rank = "Overall rank",
                              happinessScore = "Score",
                              lifeExpectancy = "Healthy life expectancy",
                              Freedom = "Freedom to make life choices",
                              GDPperCapita = "GDP per capita",
                              Country = "Country or region",
                              )

#Eliminating unnecessary columns by selecting the ones we wish to keep
newHappinessIndex2019 = newHappinessIndex2019[,c("Country",
                                                 "happinessScore",
                                                 "GDPperCapita",
                                                 "lifeExpectancy",
                                                 "Freedom",
                                                 "Generosity"
                                                  )]
#Changing variable types
newHappinessIndex2019 = mutate(newHappinessIndex2019, Country = as.factor(Country))


#Used anti_join to find if there were any discrepancies between datasets and removed them using dplpyr's setdiff
print(anti_join(newHappinessIndex2015, newHappinessIndex2019, by="Country"))
df1 = anti_join(newHappinessIndex2015, newHappinessIndex2019, by="Country")
finalNewHappinessIndex2015 = dplyr::setdiff(newHappinessIndex2015, df1)


print(anti_join(newHappinessIndex2019, newHappinessIndex2015, by="Country"))
df2 = anti_join(newHappinessIndex2019, newHappinessIndex2015, by="Country")
finalNewHappinessIndex2019 = dplyr::setdiff(newHappinessIndex2019, df2)

#Used anti_join to find if there were any discrepancies between datasets and removed them using dplpyr's setdiff
#Then renamed country names such that format became standard throughout all datasets
print(anti_join(finalAlcoholConsumptionByCapita, newHappinessIndex2015, by="Country"), n=112)
mismatchedCountries2015 = anti_join(finalAlcoholConsumptionByCapita, newHappinessIndex2015, by="Country")

print(anti_join(finalAlcoholConsumptionByCapita, newHappinessIndex2019, by="Country"), n=86)
mismatchedCountries2019 = anti_join(finalAlcoholConsumptionByCapita, newHappinessIndex2019, by="Country")


levels(finalAlcoholConsumptionByCapita$Country)[match("Bolivia (Plurinational State of)",
                                                      levels(finalAlcoholConsumptionByCapita$Country))] = "Bolivia"
levels(finalAlcoholConsumptionByCapita$Country)[match("Brunei Darussalam",
                                                      levels(finalAlcoholConsumptionByCapita$Country))] = "Brunei"
levels(finalAlcoholConsumptionByCapita$Country)[match("CÃ´te dâ€™Ivoire",
                                                      levels(finalAlcoholConsumptionByCapita$Country))] = "Ivory Coast"
levels(finalAlcoholConsumptionByCapita$Country)[match("Democratic People's Republic of Korea",
```

```
                                                    levels(finalAlcoholConsumptionByCapita$Country))] = "North Korea"
levels(finalAlcoholConsumptionByCapita$Country)[match("Democratic Republic of the Congo",
                                                    levels(finalAlcoholConsumptionByCapita$Country))] = "Congo (Kinshasa)"
levels(finalAlcoholConsumptionByCapita$Country)[match("Congo",
                                                    levels(finalAlcoholConsumptionByCapita$Country))] = "Congo (Brazzaville)"
levels(finalAlcoholConsumptionByCapita$Country)[match("Iran (Islamic Republic of)",
                                                    levels(finalAlcoholConsumptionByCapita$Country))] = "Iran"
levels(finalAlcoholConsumptionByCapita$Country)[match("Lao People's Democratic Republic",
                                                    levels(finalAlcoholConsumptionByCapita$Country))] = "Laos"
levels(finalAlcoholConsumptionByCapita$Country)[match("Micronesia (Federated States of)",
                                                    levels(finalAlcoholConsumptionByCapita$Country))] = "Micronesia"
levels(finalAlcoholConsumptionByCapita$Country)[match("Republic of Korea",
                                                    levels(finalAlcoholConsumptionByCapita$Country))] = "South Korea"
levels(finalAlcoholConsumptionByCapita$Country)[match("Republic of Moldova",
                                                    levels(finalAlcoholConsumptionByCapita$Country))] = "Moldova"
levels(finalAlcoholConsumptionByCapita$Country)[match("Russian Federation",
                                                    levels(finalAlcoholConsumptionByCapita$Country))] = "Russia"
levels(finalAlcoholConsumptionByCapita$Country)[match("Syrian Arab Republic",
                                                    levels(finalAlcoholConsumptionByCapita$Country))] = "Syria"
levels(finalAlcoholConsumptionByCapita$Country)[match("The former Yugoslav Republic of Macedonia",
                                                    levels(finalAlcoholConsumptionByCapita$Country))] = "Macedonia"
levels(finalAlcoholConsumptionByCapita$Country)[match("Trinidad and Tobago",
                                                    levels(finalAlcoholConsumptionByCapita$Country))] = "Trinidad & Tobago"
levels(finalAlcoholConsumptionByCapita$Country)[match("United Kingdom of Great Britain and Northern Ireland",
                                                    levels(finalAlcoholConsumptionByCapita$Country))] = "United Kingdom"
levels(finalAlcoholConsumptionByCapita$Country)[match("United Republic of Tanzania",
                                                    levels(finalAlcoholConsumptionByCapita$Country))] = "Tanzania"
levels(finalAlcoholConsumptionByCapita$Country)[match("United States of America",
                                                    levels(finalAlcoholConsumptionByCapita$Country))] = "United States"
levels(finalAlcoholConsumptionByCapita$Country)[match("Venezuela (Bolivarian Republic of)",
                                                    levels(finalAlcoholConsumptionByCapita$Country))] = "Venezuela"
levels(finalAlcoholConsumptionByCapita$Country)[match("Viet Nam",
                                                    levels(finalAlcoholConsumptionByCapita$Country))] = "Viet Nam"

finalFinalAlcoholComsumptionByCapita = dplyr::setdiff(finalAlcoholConsumptionByCapita, mismatchedCountries2015)
finalFinalAlcoholComsumptionByCapita = dplyr::setdiff(finalAlcoholConsumptionByCapita, mismatchedCountries2019)

#Sort Countries Alphabetically 2015 and 2019 Happiness Data
finalNewHappinessIndex2015 = finalNewHappinessIndex2015[order(finalNewHappinessIndex2015$Country),]
finalNewHappinessIndex2019 = finalNewHappinessIndex2019[order(finalNewHappinessIndex2019$Country),]

#HANDLING MISSING DATA
#Visualizing Missing Data
missingDataPlot = aggr(finalFinalAlcoholComsumptionByCapita, col=c("navyblue", "yellow"),
                       numbers=TRUE, sortVars=TRUE,
                       label=names(finalFinalAlcoholComsumptionByCapita), cex.axis=1,
                       gap=3, ylab=c("Missing Data", "Pattern"))
#Simple Imputation
imputationData = finalFinalAlcoholComsumptionByCapita
imputationData$MeanConsumption[which(is.na(imputationData$MeanConsumption))] = mean(imputationData$MeanConsumption, na.rm =TRUE)
imputationData$CILowerBoundConsumption[which(is.na(imputationData$CILowerBoundConsumption))] = mean(imputationData$CILowerBoundConsumption, na.rm =TRUE)
imputationData$CIUpperBoundConsumption[which(is.na(imputationData$CIUpperBoundConsumption))] = mean(imputationData$CIUpperBoundConsumption, na.rm =TRUE)

#MICE Imputation
#CIUpperbound is collinear to CILowerbound hence MICE cant impute Upperbound
my_imp = mice(finalFinalAlcoholComsumptionByCapita, m=5, method = "pmm", maxit = 20)

my_imp$imp$MeanConsumption
my_imp$imp$CILowerBoundConsumption
my_imp$imp$CIUpperBoundConsumption

summary(finalFinalAlcoholComsumptionByCapita$MeanConsumption)
summary(finalFinalAlcoholComsumptionByCapita$CILowerBoundConsumption)
cleanDataSet = complete(my_imp, 1)


#HYPOTHESIS TESTING
data2015 = cleanDataSet[,c("Country", "Year", "MeanConsumption")]
data2015 = subset(data2015, Year != 2019)
data2015$happinessScore = finalNewHappinessIndex2015$happinessScore
```

```
data2019 = cleanDataSet[,c("Country", "Year", "MeanConsumption")]
data2019 = subset(data2019, Year != 2015)
data2019$happinessScore = finalNewHappinessIndex2019$happinessScore

cor.test(x = data2015$MeanConsumption, y = data2015$happinessScore)
cov(x = data2015$MeanConsumption, y = data2015$happinessScore)
t.test(x = data2015$MeanConsumption, y = data2015$happinessScore)

cor.test(x = data2019$MeanConsumption, y = data2019$happinessScore)
cov(x = data2019$MeanConsumption, y = data2019$happinessScore)
t.test(x = data2019$MeanConsumption, y = data2019$happinessScore)

#PLOTS
missingDataPlot = aggr(finalFinalAlcoholComsumptionByCapita, col =c ("navyblue", "yellow"),
                       numbers = TRUE, sortVars = TRUE,
                       label = names(finalFinalAlcoholComsumptionByCapita), cex.axis = 1,
                       gap = 3, ylab = c("Missing Data", "Pattern"))

#Map Plot
map1 = aggregate(data2015$MeanConsumption, by = list(data2015$Country), FUN=sum)
colnames(map1)[colnames(map1)=="x"] = "Mean Consumption"
map1 = joinCountryData2Map(data2015, nameJoinColumn = "Country", joinCode = "NAME")
ColorPalette = RColorBrewer::brewer.pal(9, "Purples")
mapCountryData(map1, nameColumnToPlot = "MeanConsumption",
               catMethod = "fixedwidth",
               colourPalette = ColorPalette,
               numCats = 10)

map2 = aggregate(data2019$MeanConsumption, by = list(data2019$Country), FUN=sum)
colnames(map2)[colnames(map2)=="x"] = "Mean Consumption"
map2 = joinCountryData2Map(data2019, nameJoinColumn = "Country", joinCode = "NAME")
ColorPalette = RColorBrewer::brewer.pal(9, "Purples")
mapCountryData(map2, nameColumnToPlot = "MeanConsumption",
               catMethod = "fixedwidth",
               colourPalette = ColorPalette,
               numCats = 10)

map3 = aggregate(data2015$happinessScore, by = list(data2015$Country), FUN=sum)
colnames(map3)[colnames(map3)=="x"] = "Happiness Score"
map3 = joinCountryData2Map(data2015, nameJoinColumn = "Country", joinCode = "NAME")
ColorPalette = RColorBrewer::brewer.pal(9, "Purples")
mapCountryData(map3, nameColumnToPlot = "happinessScore",
               catMethod = "fixedwidth",
               colourPalette = ColorPalette,
               numCats = 10)

map4 = aggregate(data2019$happinessScore, by = list(data2019$Country), FUN=sum)
colnames(map4)[colnames(map4)=="x"] = "Happiness Score"
map4 = joinCountryData2Map(data2019, nameJoinColumn = "Country", joinCode = "NAME")
ColorPalette = RColorBrewer::brewer.pal(9, "Purples")
mapCountryData(map4, nameColumnToPlot = "happinessScore",
               catMethod = "fixedwidth",
               colourPalette = ColorPalette,
               numCats = 10)
#Scatter Plots
Scatter1 = ggplot(data2015, aes(x=happinessScore, y=MeanConsumption)) +
  geom_point() +
  geom_text(label=data2015$Country) +
  geom_smooth(method=lm, se=FALSE) +
  ggtitle("Average Alcohol Consumption and a Nation's Happiness 2015")
Scatter1

Scatter2 = ggplot(data2019, aes(x=happinessScore, y=MeanConsumption)) +
  geom_point() +
  geom_text(label=data2019$Country) +
  geom_smooth(method=lm, se=FALSE) +
  ggtitle("Average Alcohol Consumption and a Nation's Happiness 2019")
Scatter2
```

8. Works Cited List

Charles H, Patrick; Durham, NC (1952). Alcohol, Culture, and Society. Duke University Press (reprint edition by AMS Press, New York, 1970). pp. 26–27

"Alcohol and the Nervous System" *Transformations Treatment Center* https://www.transformationstreatment.center/resources/addiction-articles/how-does-alcohol-affect-the-nervous-system/