Relationship between Alcohol Consumption and Mortality Rate

The goal of this report is to find any relationship between alcohol consumption and mortality rate. In order to do this, I have chosen the dataset “Alcohol” and “Mortality”. Dataset Alcohol was composed of variables “Country”, “Year” and “Alcohol”(in Liters). The variable Alcohol represents the average amount of alcohol consumed among adult(15 years or older) population in the country. Dataset Mortality was composed of variables “Country”, “Year”, “Both sexes”, “Male” and “Female”. The variables Both sexes, Male and Female represents adult mortality rate, or probability of dying between 15 and 60 years per 1000 population. Dataset Alcohol was provided with R packages while dataset Mortality was acquired through kaggle.com. I have chosen these two dataset because I pursue a career in the dental field. I could not find a dataset on how alcohol consumption could affects one’s teeth, so I decided to find out how alcohol would affect someone’s health in general. I expect to find higher mortality rate in the countries that have higher alcohol consumption. My dataset was already tidy so I did not tidy them in R.

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readxl)
library(ggplot2)

Alcohol <- read_excel("C:/Users/okcij/Documents/R/Alcohol.xlsx")
Mortality<- read_excel("C:/Users/okcij/Documents/R/Mortality.xlsx")

Al_Mo <- inner_join(Alcohol, Mortality, by=c("Country","Year"))
#puts together two data by variables Country and Year
```

I joined dataset Alcohol and Mortality using inner join function. As a result, countries that were found in Alcohol but not in Mortality was dropped. Also, only the data from year 2005 to 2008 was kept because the dataset Alcohol only contained years 2005-2008 with a few exceptions(Sweden had dataset from years 2000-2004). Total of 2790 cases from Mortality and 90 cases from Alcohol were dropped. I decided to use inner join because dataset Mortality had total of 3111 cases while dataset Alcohol only had 411 cases. If I used any other types of join, I would’ve gotten too many N/A in the new dataset. It would’ve made many cases in the dataset useless because I can’t compare two variables if one of the variables is not available.

#Only year 2005 and 2008 was used for the purpose of this project, not years 2005 through 2008. Sweden was the only country that had data for years other than 2005 and 2008, so Sweden was considered as an outlier in this data.

```
library(dplyr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v tibble  3.0.5     v purrr   0.3.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#creates a new data with data from only 2005
Al_Mo_2005 <- Al_Mo %>%
  filter(Year=="2005")

#creates a new data with data from only 2008
Al_Mo_2008 <- Al_Mo %>%
  filter(Year=="2008")

Al_Mo_Mean_Alcohol <- Al_Mo %>%
  #groups the data by Country
  group_by(Country) %>%
  #summarizes the data by mean alcohol consumption
  summarize(mean_Alcohol = mean(Alcohol)) %>%
  #adds the mean alcohol consumption to the original data
  full_join(Al_Mo)
```

```
## Joining, by = "Country"
```

```
Al_Mo_Mean_Mortality <- Al_Mo %>%
  #groups the data by Country
  group_by(Country) %>%
  #summarizes the data by mean alcohol consumption
  summarize(mean_BothSexes = mean(BothSexes)) %>%
  #adds the mean mortality rate to the original data
  full_join(Al_Mo)
```

```
## Joining, by = "Country"
```

```
Al_Mo <- Al_Mo %>%
  #creates a variable that shows alcohol consumption per mortality rate
  mutate(Consumption = Alcohol/BothSexes)

Al_Mo_final <- full_join(Al_Mo_Mean_Mortality, Al_Mo_Mean_Alcohol,
                         by=c("Year", "Country", "Alcohol", "BothSexes",
                              "Male", "Female"))

Al_Mo %>%
  #groups by each year
  group_by(Year) %>%
  #selects the variable alcohol
  select(Alcohol) %>%
  #summarizes the minimum alcohol consumption per year
  summarize(min_Alcohol=min(Alcohol))
```

```
## Adding missing grouping variables: `Year`
```

```
## # A tibble: 9 x 2
##    Year min_Alcohol
##   <dbl>       <dbl>
## 1  2000        8.4
## 2  2001        9.1
## 3  2002        9.9
## 4  2003       10.2
## 5  2004       10.5
## 6  2005        0.02
## 7  2006        9.8
## 8  2007        9.8
## 9  2008        0.03
```

```
Al_Mo %>%
  group_by(Year) %>%
  select(Alcohol) %>%
  #summarizes the maximum alcohol consumption per year
  summarize(max_Alcohol=max(Alcohol))
```

```
## Adding missing grouping variables: `Year`
```

```
## # A tibble: 9 x 2
##    Year max_Alcohol
##   <dbl>       <dbl>
## 1  2000         8.4
## 2  2001         9.1
## 3  2002         9.9
## 4  2003        10.2
## 5  2004        10.5
## 6  2005        16.3
## 7  2006         9.8
## 8  2007         9.8
## 9  2008        18.8
```

```
Al_Mo %>%
  group_by(Year) %>%
  select(Alcohol) %>%
  #summarizes the mean of alcohol consumption per year
  summarize(mean_Alcohol=mean(Alcohol))
```

```
## Adding missing grouping variables: `Year`
```

```
## # A tibble: 9 x 2
##    Year mean_Alcohol
##   <dbl>        <dbl>
## 1  2000         8.4
## 2  2001         9.1
## 3  2002         9.9
## 4  2003        10.2
## 5  2004        10.5
## 6  2005         6.32
## 7  2006         9.8
## 8  2007         9.8
## 9  2008         6.48
```

```
Al_Mo %>%
  group_by(Year) %>%
  select(Alcohol) %>%
  #summarizes the median of alcohol consumption per year
  summarize(median_Alcohol=median(Alcohol))
```

```
## Adding missing grouping variables: `Year`
```

```
## # A tibble: 9 x 2
##    Year median_Alcohol
##   <dbl>          <dbl>
## 1  2000           8.4
## 2  2001           9.1
## 3  2002           9.9
## 4  2003          10.2
## 5  2004          10.5
## 6  2005           5.91
## 7  2006           9.8
## 8  2007           9.8
## 9  2008           5.92
```

```
Al_Mo %>%
  group_by(Year) %>%
  select(Alcohol) %>%
  #summarizes the standard deviation of alcohol consumption per year
  summarize(sd_Alcohol=sd(Alcohol))
```

```
## Adding missing grouping variables: `Year`
```

```
## # A tibble: 9 x 2
##    Year sd_Alcohol
##   <dbl>      <dbl>
## 1  2000      NA
## 2  2001      NA
## 3  2002      NA
## 4  2003      NA
## 5  2004      NA
## 6  2005       4.52
## 7  2006      NA
## 8  2007      NA
## 9  2008       4.77
```

```
Al_Mo %>%
  group_by(Year) %>%
  #summarizes the quantile of mortality rate per year
  summarize(quantile_Alcohol=quantile(Alcohol))
```

```
## `summarise()` has grouped output by 'Year'. You can override using the `.groups` argument.
```

```
## # A tibble: 45 x 2
## # Groups:   Year [9]
##     Year quantile_Alcohol
##    <dbl>            <dbl>
##  1  2000              8.4
##  2  2000              8.4
##  3  2000              8.4
##  4  2000              8.4
##  5  2000              8.4
##  6  2001              9.1
##  7  2001              9.1
##  8  2001              9.1
##  9  2001              9.1
## 10  2001              9.1
## # ... with 35 more rows
```

```
Al_Mo_final%>%
  group_by(Country) %>%
  select(Alcohol) %>%
  #shows countries with decreasing alcohol consumption
  arrange(desc(Alcohol))
```

```
## Adding missing grouping variables: `Country`
```

```
## # A tibble: 321 x 2
## # Groups:   Country [157]
##    Country   Alcohol
##    <chr>       <dbl>
##  1 Belarus      18.8
##  2 Ukraine      17.5
##  3 Estonia      17.2
##  4 Uganda       16.4
##  5 Lithuania    16.3
##  6 Hungary      16.3
##  7 Romania      16.2
##  8 Hungary      16.1
##  9 Ukraine      15.6
## 10 Estonia      15.6
## # ... with 311 more rows
```

```
Al_Mo %>%
  group_by(Year) %>%
  select(BothSexes) %>%
  #summarizes the minimum mortality rate per year
  summarize(min_BothSexes=min(BothSexes))
```

```
## Adding missing grouping variables: `Year`
```

```
## # A tibble: 9 x 2
##    Year min_BothSexes
##   <dbl>         <dbl>
## 1  2000            72
## 2  2001            72
## 3  2002            70
## 4  2003            68
## 5  2004            70
## 6  2005            61
## 7  2006            64
## 8  2007            63
## 9  2008            57
```

```
Al_Mo %>%
  group_by(Year) %>%
  select(BothSexes) %>%
  #summarizes the maximum mortality rate per year
  summarize(max_BothSexes=max(BothSexes))
```

```
## Adding missing grouping variables: `Year`
```

```
## # A tibble: 9 x 2
##    Year max_BothSexes
##   <dbl>         <dbl>
## 1  2000            72
## 2  2001            72
## 3  2002            70
## 4  2003            68
## 5  2004            70
## 6  2005           681
## 7  2006            64
## 8  2007            63
## 9  2008           596
```

```
Al_Mo %>%
  group_by(Year) %>%
  #summarizes the mean of mortality rate per year
  summarize(mean_BothSexes=mean(BothSexes))
```

```
## # A tibble: 9 x 2
##    Year mean_BothSexes
##   <dbl>          <dbl>
## 1  2000           72
## 2  2001           72
## 3  2002           70
## 4  2003           68
## 5  2004           70
## 6  2005          209.
## 7  2006           64
## 8  2007           63
## 9  2008          193.
```

```
Al_Mo %>%
  group_by(Year) %>%
  #summarizes the median of mortality rate per year
  summarize(median_BothSexes=median(BothSexes))
```

```
## # A tibble: 9 x 2
##    Year median_BothSexes
##   <dbl>            <dbl>
## 1  2000               72
## 2  2001               72
## 3  2002               70
## 4  2003               68
## 5  2004               70
## 6  2005              179
## 7  2006               64
## 8  2007               63
## 9  2008              165
```

```
Al_Mo %>%
  group_by(Year) %>%
  #summarizes the standard deviation of mortality rate per year
  summarize(sd_BothSexes=sd(BothSexes))
```

```
## # A tibble: 9 x 2
##    Year sd_BothSexes
##   <dbl>        <dbl>
## 1  2000           NA
## 2  2001           NA
## 3  2002           NA
## 4  2003           NA
## 5  2004           NA
## 6  2005         127.
## 7  2006           NA
## 8  2007           NA
## 9  2008         110.
```

```
Al_Mo %>%
  group_by(Year) %>%
  #summarizes the quantile of mortality rate per year
  summarize(quantile_BothSexes=quantile(BothSexes))
```

```
## `summarise()` has grouped output by 'Year'. You can override using the `.groups` argument.
```

```
## # A tibble: 45 x 2
## # Groups:   Year [9]
##     Year quantile_BothSexes
##    <dbl>              <dbl>
##  1  2000                 72
##  2  2000                 72
##  3  2000                 72
##  4  2000                 72
##  5  2000                 72
##  6  2001                 72
##  7  2001                 72
##  8  2001                 72
##  9  2001                 72
## 10  2001                 72
## # ... with 35 more rows
```

```
Al_Mo_final%>%
  group_by(Country) %>%
  select(mean_BothSexes) %>%
  #shows countries with decreasing mortality rate
  arrange(desc(mean_BothSexes))
```

```
## Adding missing grouping variables: `Country`
```

```
## # A tibble: 321 x 2
## # Groups:   Country [157]
##    Country                  mean_BothSexes
##    <chr>                             <dbl>
##  1 Zimbabwe                           638.
##  2 Zimbabwe                           638.
##  3 Lesotho                            579
##  4 Lesotho                            579
##  5 Central African Republic           517
##  6 Central African Republic           517
##  7 Zambia                             498.
##  8 Zambia                             498.
##  9 Malawi                             486.
## 10 Malawi                             486.
## # ... with 311 more rows
```

```
Al_Mo_final %>%
  #shows correlation coefficient between alcohol consumpiton and mortality    rate
  summarize(cor(Alcohol,BothSexes, use = "pairwise.complete.obs"))
```

```
## # A tibble: 1 x 1
##   `cor(Alcohol, BothSexes, use = "pairwise.complete.obs")`
##                                                      <dbl>
## 1                                                   -0.172
```

```
#           Alcohol(2005) |  Alcohol(2008)  |Mortality(2005)|Mortality(2008)
#-------------------------------------------------------------------------
#Minimum|0.02(Afghanistan)|0.03(Afghanistan)|  61(Iceland)  | 57(Iceland)
#-------------------------------------------------------------------------
#Maximum| 16.27(Hungary)  | 18.85(Belarus)  | 681(Zimbabwe) | 596(Zimbabwe)
#-------------------------------------------------------------------------
#Mean   |      6.319      |      6.479      |    208.987    |    192.731
#-------------------------------------------------------------------------
#Median |       5.91      |       5.92      |      179      |      165
#-------------------------------------------------------------------------
#SD     |      4.522      |      4.773      |    127.253    |    110.126
```

Years 2000, 2001, 2002, 2003, 2004, 2006 and 2007 were not included because only data from Swiss was available. Since no other countries had data on those years it was determined that it would be more sufficient to exclude the data.

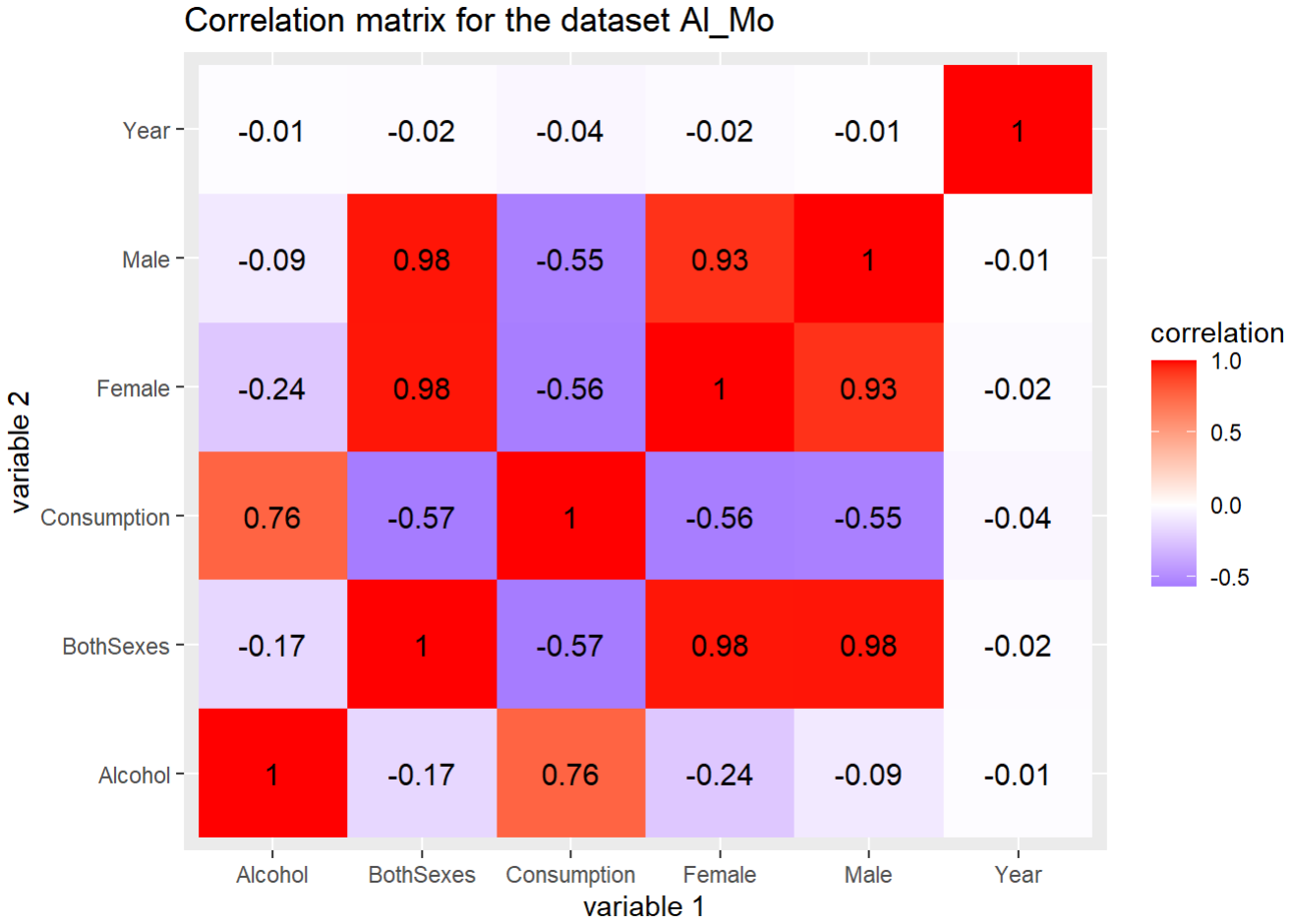
Afghanistan consumed the least amount of alcohol in both 2005 and 2008. Hungray consumed the most amount of alcohol in 2005 and Belarus consumed the most amount of alcohol in 2008. Iceland had the lowest mortality rate in both 2005 and 2008 while Zimbabwe had the highest mortality rate in both 2005 and 2008.

Mean of Alcohol in 2008 was slightly higher than mean of alcohol in 2005 while mean of Mortality in 2008 was lower than mean of 2005. Median and SD for variable Alcohol did not show much change from year 2005 to 2008, while the variable Mortality showed greater amount of decrease. This could indicate that alcohol consumption increased over the years while the mortality rate decreased.

Variable Alcohol and Mortality(written as BothSexes in data) had correlation coefficient of -0.17, which means that mortality rate increases as the alcohol consumption decreases.

#Adjustments were made to show both the code and the outputs.

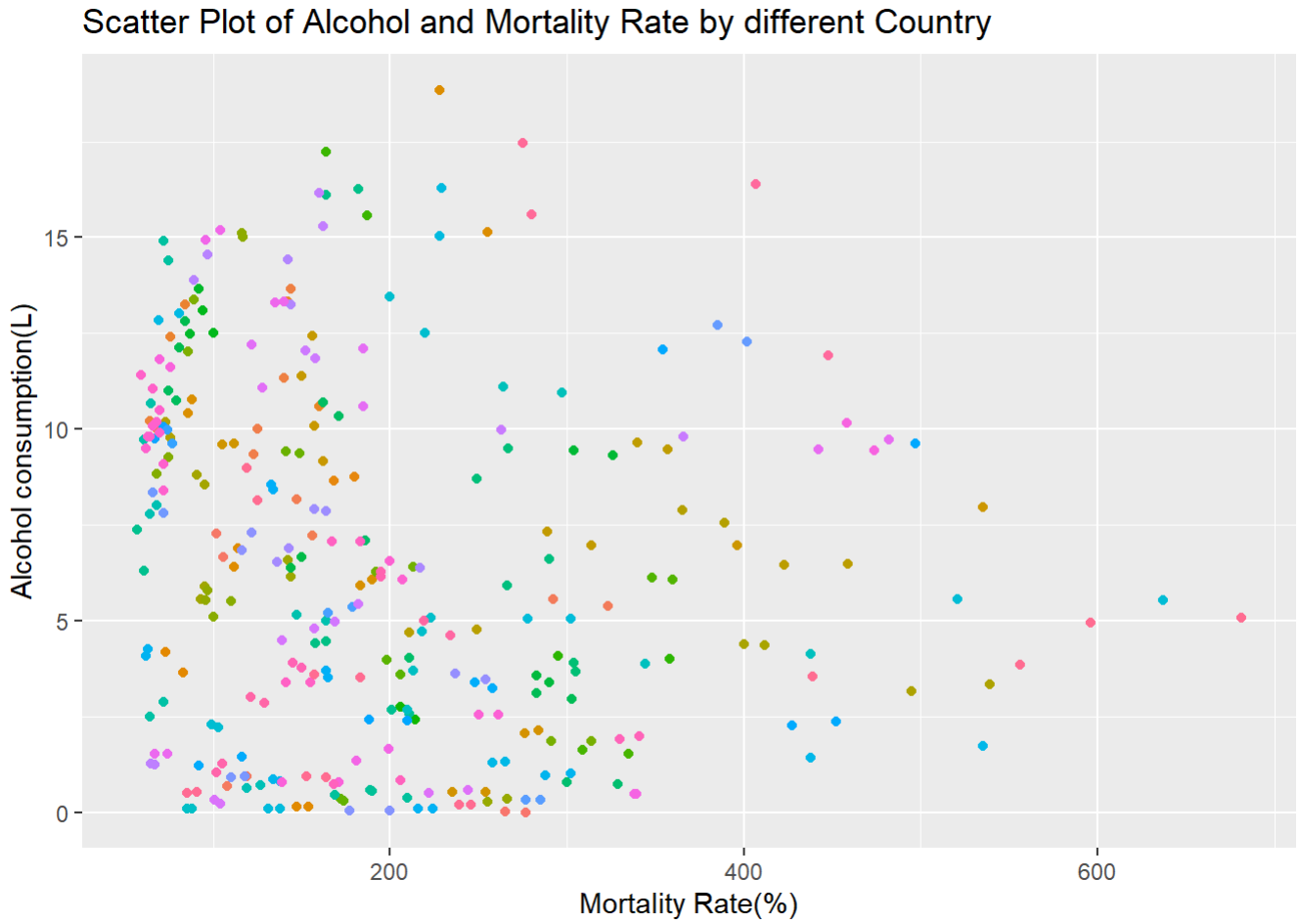
```
Al_Mo_num <- Al_Mo %>%
  #creates a new data with only numeric variables
  select_if(is.numeric)

cor(Al_Mo_num, use = "pairwise.complete.obs") %>%
  #saves as data
  as.data.frame %>%
  #converts each row to a variable
  rownames_to_column %>%
  #all correlations appear in the same column
  pivot_longer(-1, names_to = "other_var", values_to = "correlation") %>%
  ggplot(aes(rowname, other_var, fill=correlation)) +
  geom_tile() +
  #changes the scale for a neutral appeal
  scale_fill_gradient2(low="blue",mid="white",high="red") +
  #overlays the values
  geom_text(aes(label = round(correlation,2)), color = "black", size = 4) +
  #writes title and labels the axis
  labs(title = "Correlation matrix for the dataset Al_Mo", x = "variable 1", y = "variable 2")
```



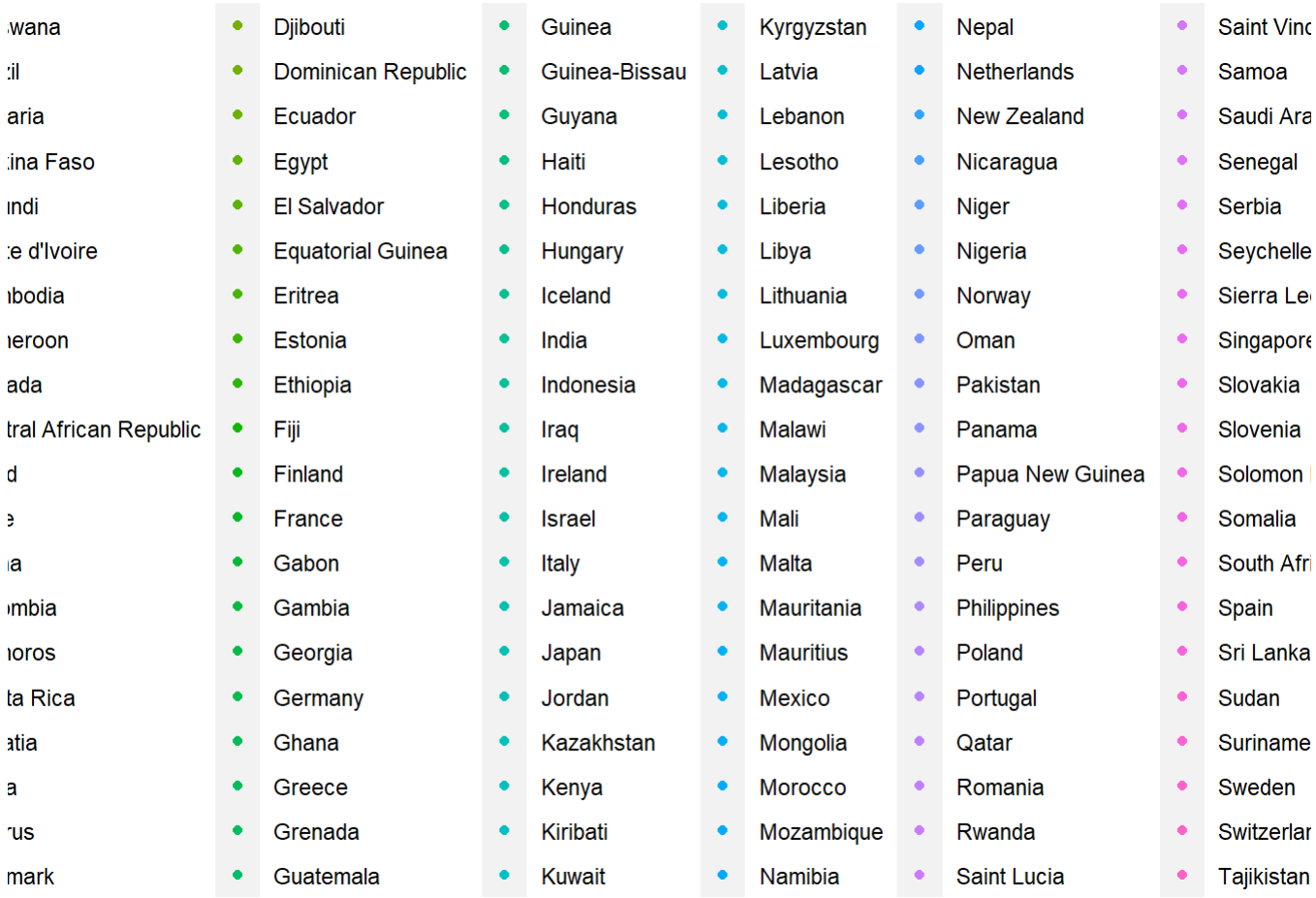
This correlation heatmap shows that mortality rate and alcohol consumption has very weak relationship. Males have almost no relationship with alcohol consumption because the correlation coefficient is -0.09, which is very close to 0. Although females have a slightly stronger correlation coefficient than males, the value is only -0.24 which still indicates a weak relationship between alcohol consumption and mortality rate of females.

#Due to some adjustments made in the above section regarding summary statistics, the output of the correlation heatmap was changed as well. From this new correlation heatmap, we can observe that males and females both have moderate relationship regarding consumption, which is the amount of alcohol consumed(in L) per one percent of mortality rate.

```
ggplot(Al_Mo_final, aes(x=BothSexes, y=Alcohol, color=Country)) +
  #creates a scatter plot
  geom_point(show.legend=FALSE) +
  #adds titles
  labs(title="Scatter Plot of Alcohol and Mortality Rate by different Country", x="Mortality Rat
e(%)", y="Alcohol consumption(L)")
```



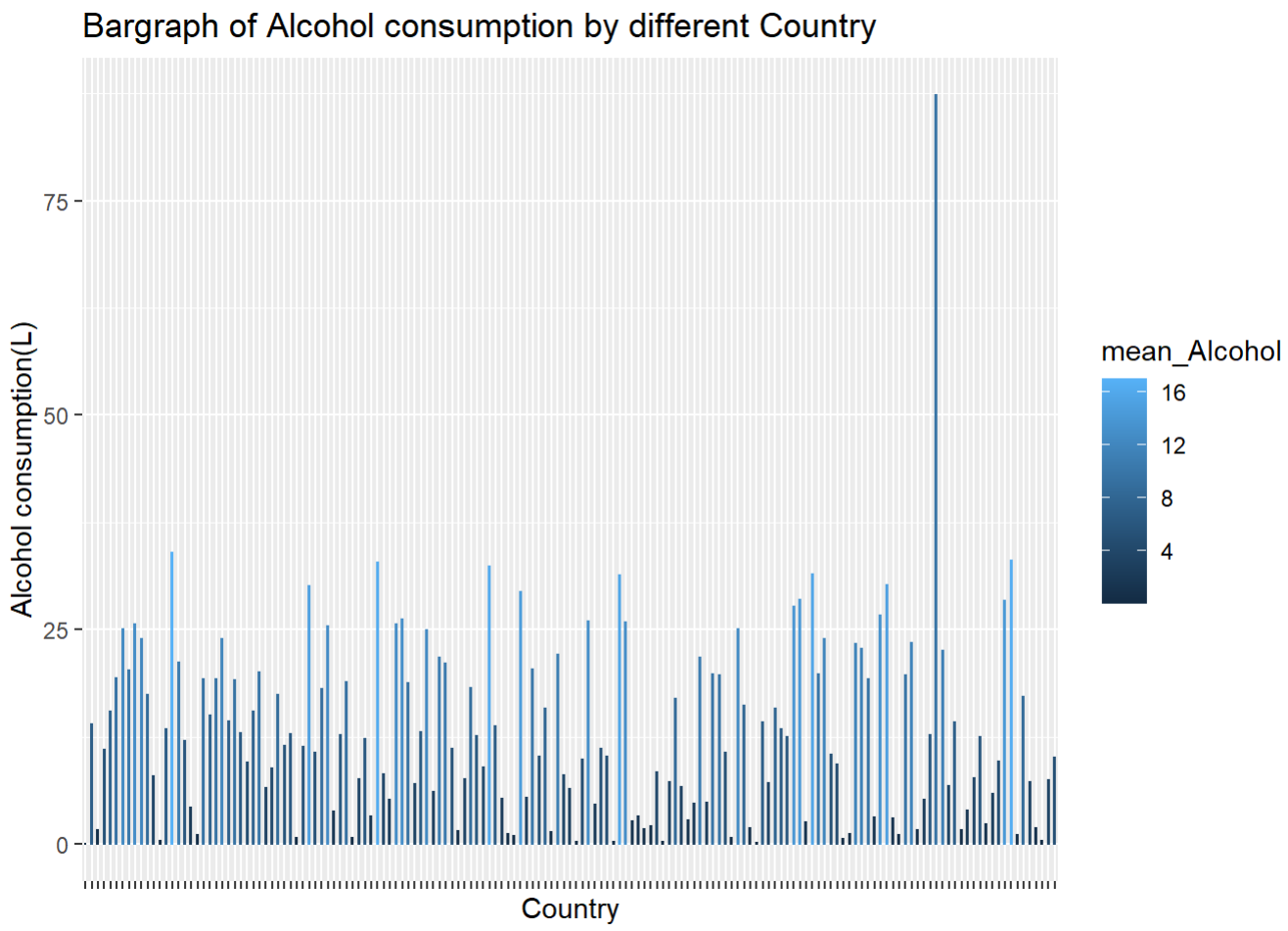

```
ggplot(Al_Mo_final, aes(x=BothSexes, y=Alcohol, color=Country)) +
  geom_point()
```



A legend was deleted because there were too many countries and the graph did not appear if a legend was present.There is no correlation between alcohol consumption and mortality rate because the data is too scattered without a pattern and thus a slope cannot be drawn.

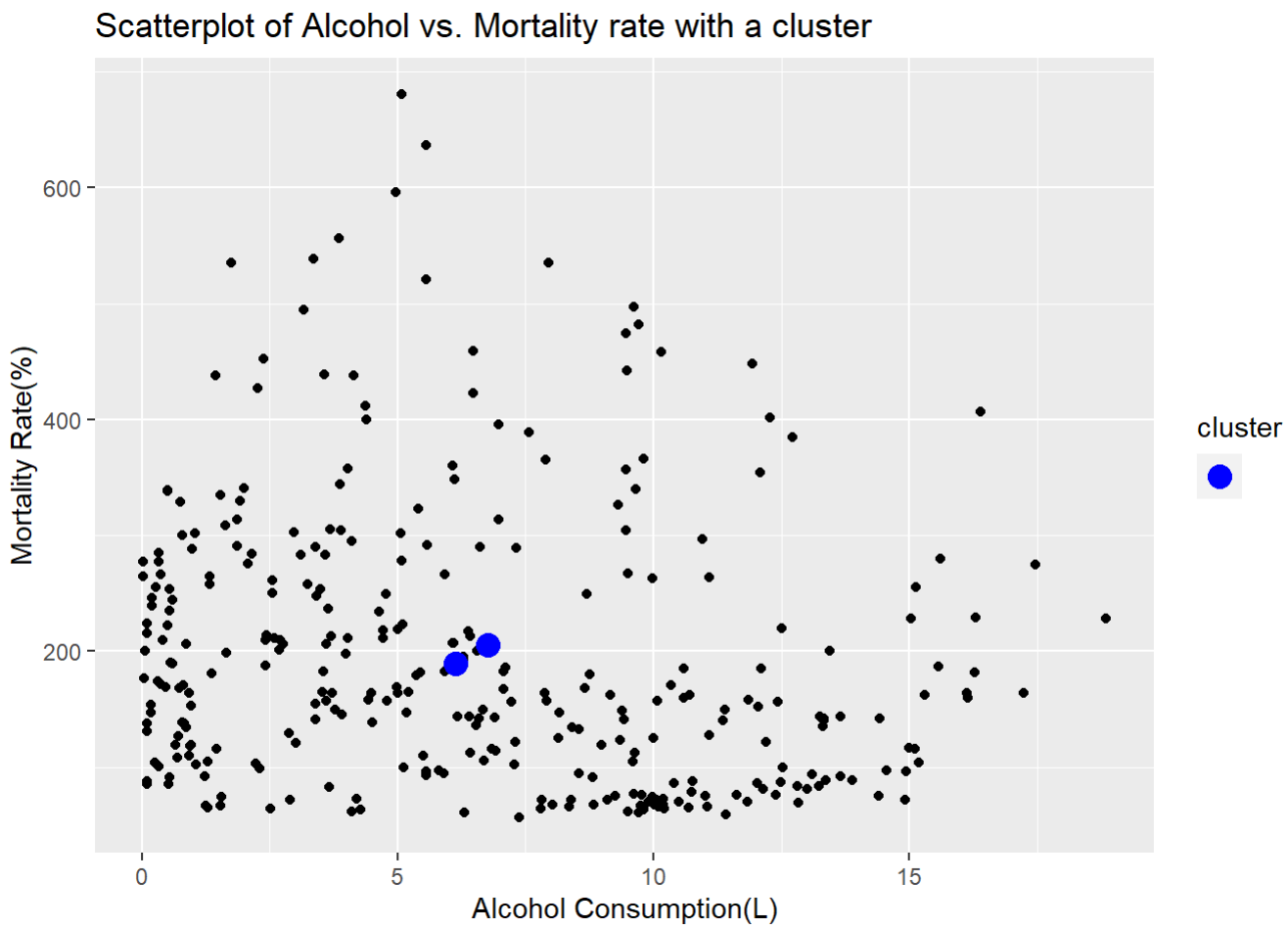
#A legend was added on a separate graph so that both graph and the legend is visible.

```
ggplot(Al_Mo_final, aes(x=Country, y=Alcohol, color=mean_Alcohol)) +
  #creates a barplot
  geom_bar(stat="identity", width=0.1) +
  theme(axis.text.x=element_blank()) +
  labs(title="Bargraph of Alcohol consumption by different Country", x="Country", y="Alcohol con
sumption(L)")
```



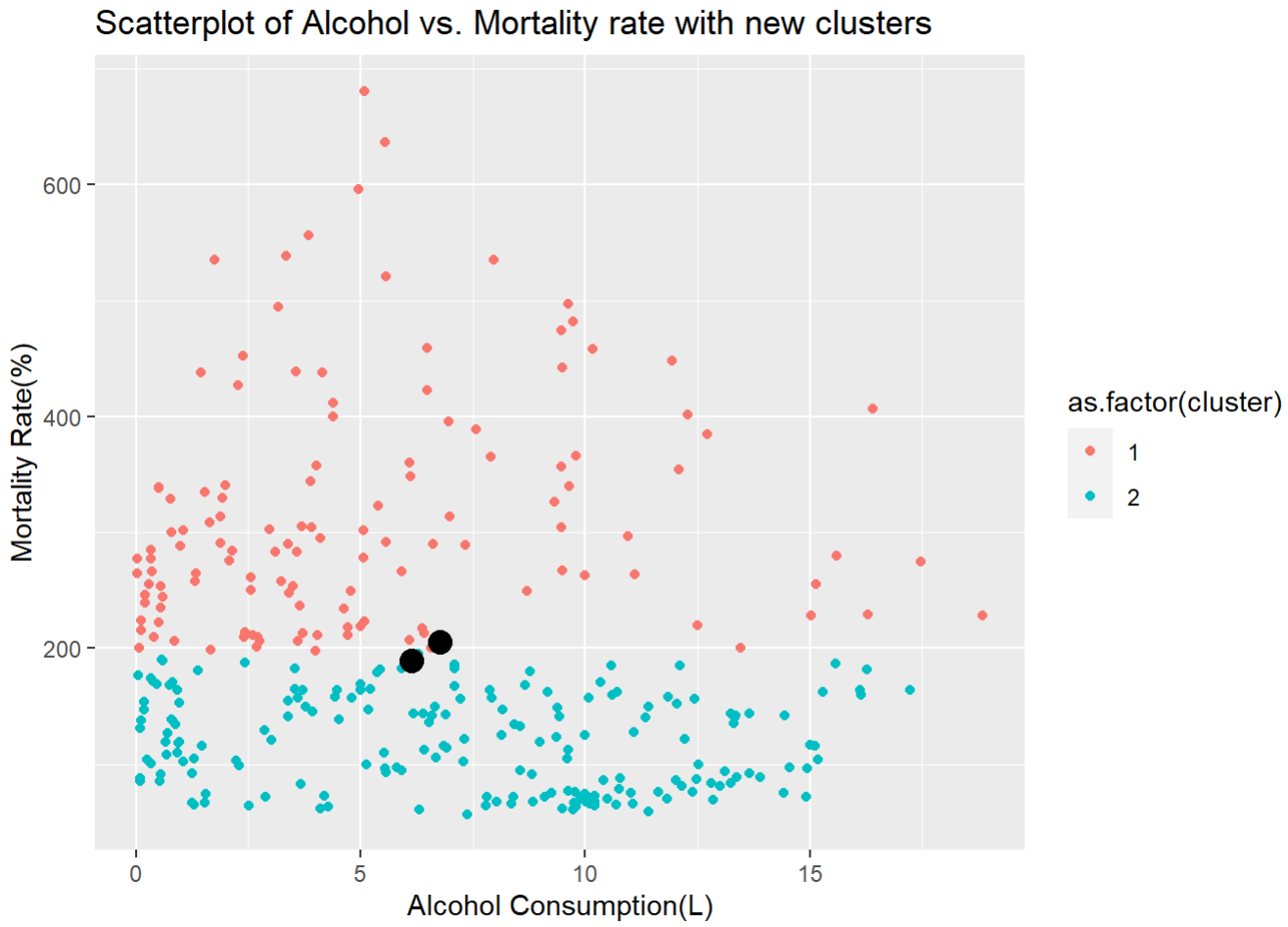
Countries in x-axis was not labeled because there were too many countries and labeling them made the x-axis not readable. The bars were colored according to their mean alcohol consumption, with lighter blue having a higher mean while darker blue having a lower mean. Belarus had the highest mean alcohol consumption while Afghanistan had the lowest mean alcohol consumption.

```
library(cluster)

Al_Mo_k <- Al_Mo %>%
  #randomly assigns 1 of 2 clusters to the 321 observations
  mutate(cluster=sample(c('1','2'), 321, replace=T)) %>%
  group_by(cluster) %>%
  #finds centers with means
  summarize(Alcohol=mean(Alcohol), BothSexes=mean(BothSexes))

ggplot(Al_Mo) +
  geom_point(aes(Alcohol, BothSexes)) +
  geom_point(data = Al_Mo_k, aes(Alcohol, BothSexes,fill=""), color="blue", size=4) +
  scale_fill_manual(name="cluster", values = "black") +
  labs(title="Scatterplot of Alcohol vs. Mortality rate with a cluster",
       x="Alcohol Consumption(L)", y="Mortality Rate(%)")
```



Final number of clusters, which was two, was determined considering that only two variable are being compared. There were no groups found using the clusters, which means there is no relationship between mortality rate and alcohol consumption.

```
#attributes observation to cluster in terms of distances
Al_Mo_distance <- Al_Mo %>%
  #calculates distances between each observation and the center of each cluster
  mutate(dist1 = sqrt((Alcohol - Al_Mo_k$Alcohol[1])^2 + (BothSexes - Al_Mo_k$BothSexes[1])^2),
         dist2 = sqrt((Alcohol - Al_Mo_k$Alcohol[2])^2 + (BothSexes - Al_Mo_k$BothSexes[2])^2))
 %>%
  #caculates by row
  rowwise() %>%
  #chooses the cluster with the minimum distance
  mutate(cluster = which.min(c(dist1,dist2))) %>%
  #stops the calculations by rows
  ungroup()

Al_Mo_Centers <- Al_Mo_distance %>%
  group_by(cluster) %>%
  #calculates the new centers
  summarize(Alcohol = mean(Alcohol), BothSexes = mean(BothSexes))

ggplot(Al_Mo_distance) +
  #plots new centers with clusters
  geom_point(aes(Alcohol, BothSexes, color = as.factor(cluster))) +
  geom_point(data = Al_Mo_k, aes(Alcohol,BothSexes), color="black", size=4)+
  labs(title="Scatterplot of Alcohol vs. Mortality rate with new clusters",
       x="Alcohol Consumption(L)", y="Mortality Rate(%)")
```



From this scatterplot, we can conclude that there is no considerable relationship between alcohol consumption and mortality rate because there was no significant improvement. Although women have slightly higher correlation coefficient compared to men, the value is not considerably high.

#All adjustments were made regarding the comments. I got points off from Join/Merge section for adressing the wrong sets of data. Although the inner_join function did include years other than 2005 and 2008, new codes were written in the Summary Statistics part so that only data from year 2005 and 2008 were used. I also got points off from Visualizations part for not including a legend, which was included as I went over this project. Finally, I got points taken off from Formatting because some codes were not showing up properly. This problem was fixed after some adjustments to the order of the codes. I did not get any points off from any other sections, so those sections were not mentioned as I made changes.