Pair Assignment 1

Tori Dykes and Mansi Wadhwa

Tuesday, October 04, 2016

The two datasets are: Swiss Fertility and Socio-economic indicators (1888) Data from R's built-in datasets and worldwide alcohol consumption (by country) from the fivethirtyeight datasets.

Swiss Fertility and Socio-economic indicators (1888) Data

```
data(swiss)
names(swiss)
```

```
## [1] "Fertility"        "Agriculture"      "Examination"
## [4] "Education"        "Catholic"         "Infant.Mortality"
```

The first dataset 'swiss' is the Swiss Fertility and Socio-economic indicators (1888) Data from R itself that provides standardized fertility measure and socioeconomic indicators for each of the 47 French-speaking provinces of Switzerland. It contains a total of 47 observations and 6 variables (all in percent).

Exploring the relationship between the variables Fertility and Catholic:

The variable Fertility gives the standardized fertility measure and the variable Catholic gives us the percentage of Catholics in the population of each province. Quite some research has been focussed on exploring the impact of socio-economic, cultural and religious conditions on fertility levels. The case of Switzerland is interesting due to its cultural diversity and the fact that its population declined greatly in 1885.

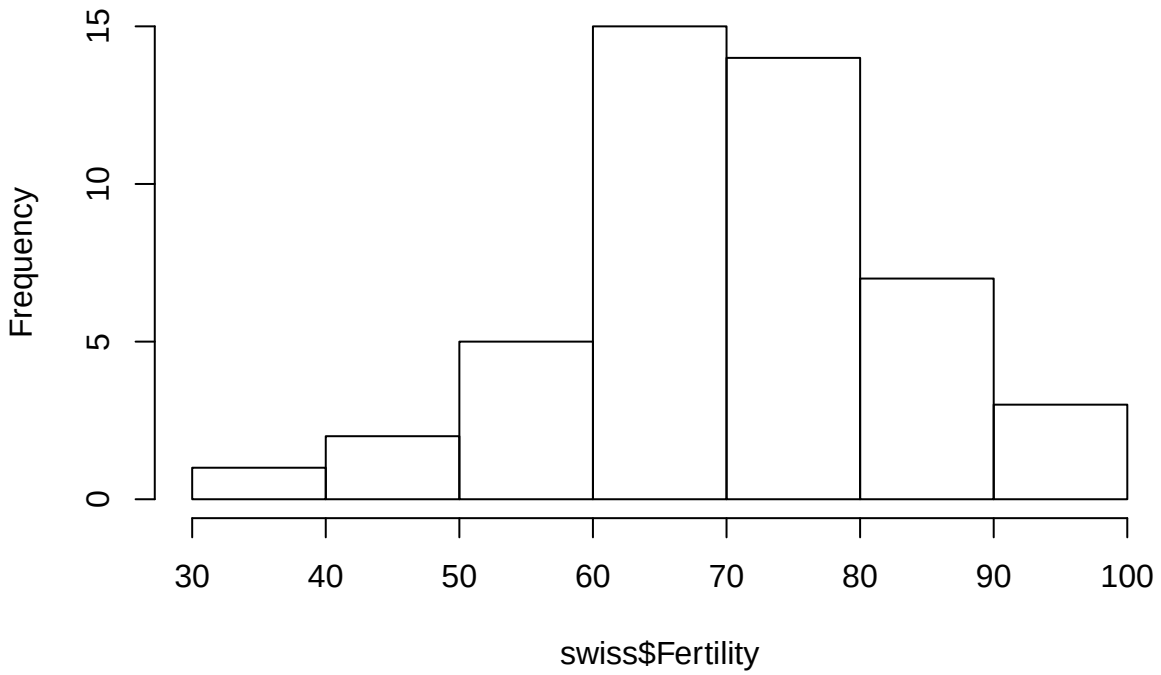
```
summary(swiss$Fertility)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   35.00   64.70   70.40   70.14   78.45   92.50
```

```
summary(swiss$Catholic)
```

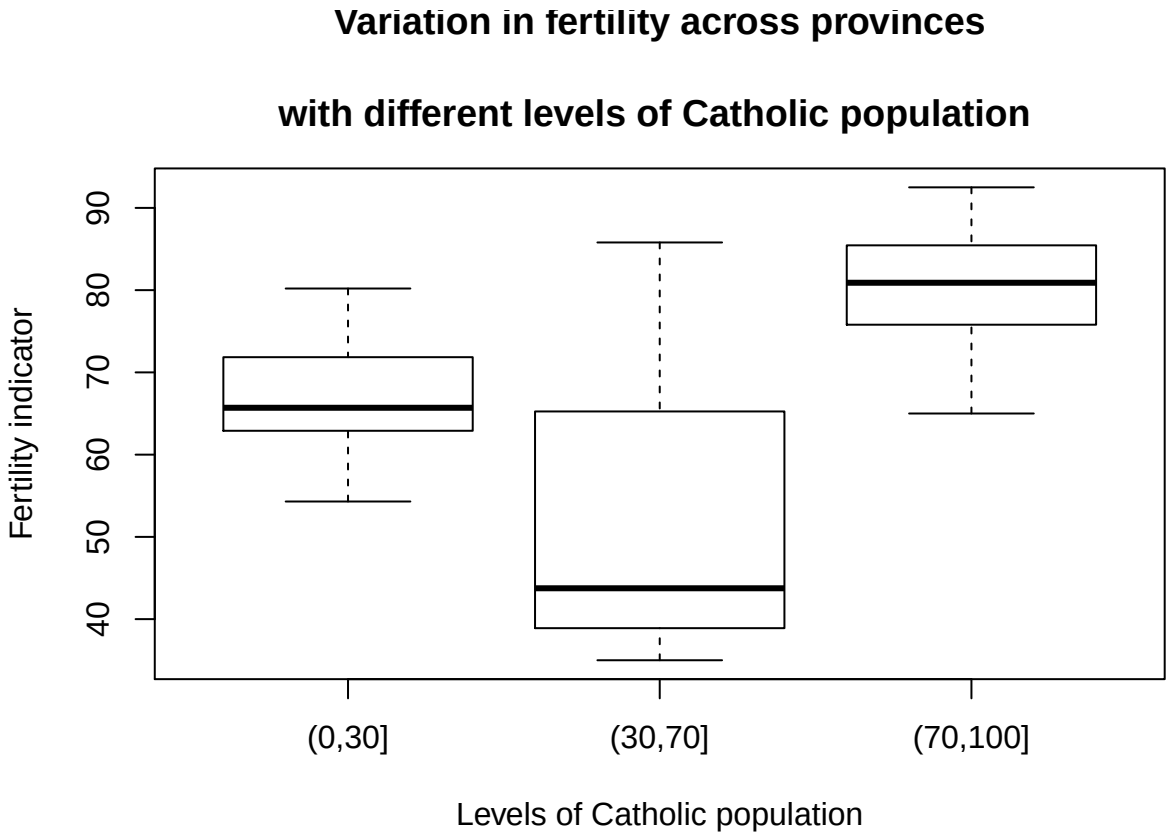
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.150   5.195  15.140  41.140  93.120 100.000
```

```
hist(swiss$Fertility, main = 'Standardized measure of fertility for 47 Swiss provinces in 1888')
```



To understand the relationship between fertility and Catholic better, we create a factor variable 'factor' that creates 3 categories for the variable Catholic.Variable factor takes 3 levels 0 to 30, 30 to 70 to 100. Level 0 to 30 represents those provinces where the Catholic population is less than or equal to 30% and so on.

```
swiss$factor <- cut(swiss$Catholic, c(0, 30, 70, 100))
boxplot(swiss$Fertility ~ swiss$factor, main = 'Variation in fertility across provinces
\nwith different levels of Catholic population \n', xlab = 'Levels of Catholic population',
ylab = 'Fertility indicator')
```



The boxplot shows that provinces with a Catholic population between 30 and 70 percent have a substantially lower ferility level than the other two categories.

```
cor.test(swiss$Fertility, swiss$Catholic)
```

```
##
##  Pearson's product-moment correlation
##
## data:  swiss$Fertility and swiss$Catholic
## t = 3.5107, df = 45, p-value = 0.001029
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2036326 0.6626204
## sample estimates:
##       cor
## 0.4636847
```

The Pearson's correlation for this pair of variables is 0.46 and hence, shows a positive but moderately strong relationship.

We can also further explore the data by way of a simple regression exploring the relationship between fertility and the percentage of Catholics living in a province.

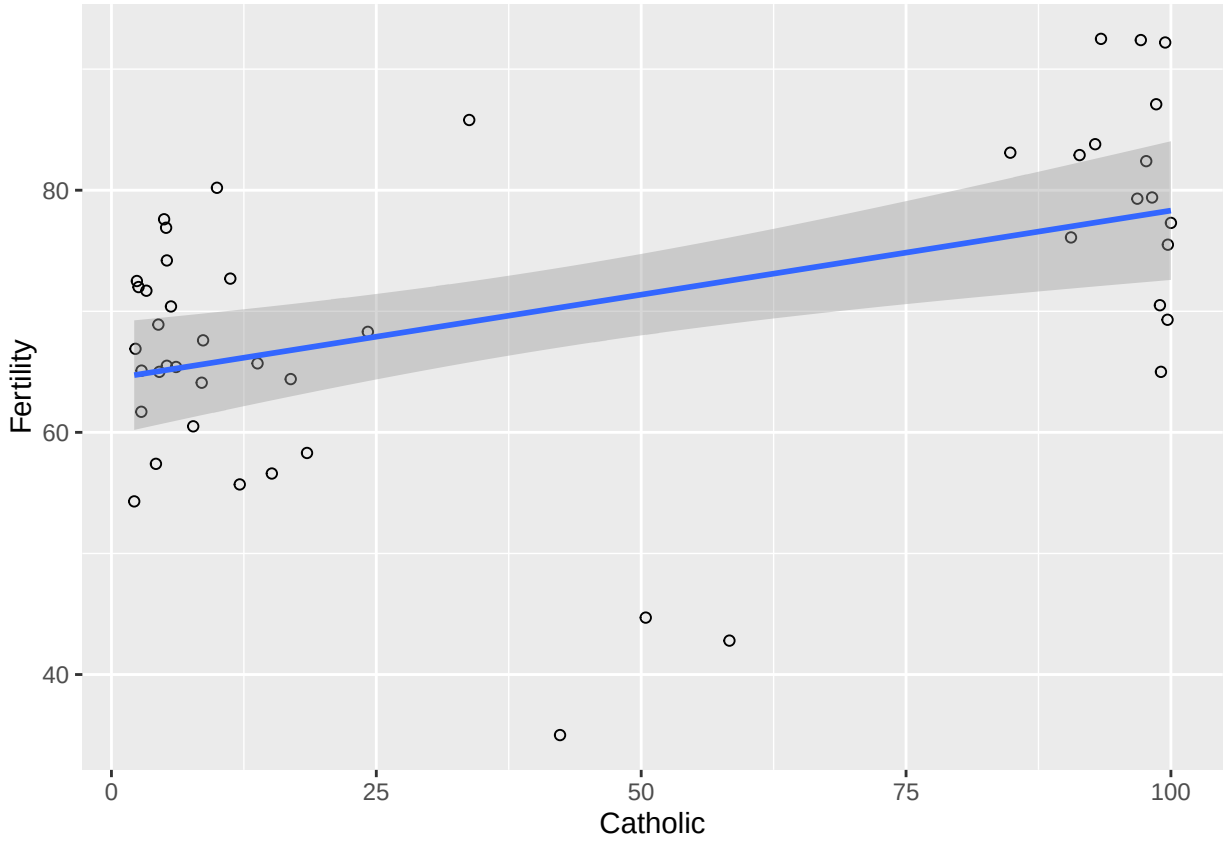
```
fertilreg <- lm(swiss$Fertility ~ swiss$Catholic)

stargazer(list(fertilreg), header = F, float = F, single.row = T)
```

	Dependent variable:
	Fertility
Catholic	0.139*** (0.040)
Constant	64.428*** (2.305)
Observations	47
R^2	0.215
Adjusted R^2	0.198
Residual Std. Error	11.190 (df = 45)
F Statistic	12.325*** (df = 1; 45)
Note:	*p<0.1; **p<0.05; ***p<0.01

Finally, we can plot the values and apply a regression line (as well as shaded confidence intervals) to get a further visual perspective:

```
ggplot(swiss, aes(x=Catholic, y=Fertility)) +
    geom_point(shape=1) +    # Use hollow circles
    geom_smooth(method=lm)
```



We see that while there is a positive relationship between the two variables, it's not very strong. Moreover, we see a strong bimodal distribution in the datapoints, with there being significant clustering around both very low rates and very high rates of Catholicism and few data points between them.

Worldwide Alcohol Consumption (by country)

This dataset, published by the WHO in 2010, looks at the amount of beer, spirits, or wine were consumed per person in a given country in 2010. The units (according to this post: http://fivethirtyeight.com/datalab/dear-mona-followup-where-do-people-drink-the-most-beer-wine-and-spirits/) are individual servings – so, a can of beer, a glass of wine, or a shot of a spirit. The dataset also looks at the total amount of alcohol consumed in liters.

```
alcoholconsump <- read.csv('drinks.csv', skip = 0, header = T, sep = ',',
              fileEncoding = 'latin1', encoding = 'UTF-8',
              stringsAsFactors = F)
```

Here are some general overview statistics for the data set:

Beer Consumption

Average consumption per person per year (across all countries): 106.1606218

Max consumption per person per year: 376 by Namibia

Wine Consumption

Average consumption per person per year (across all countries): 49.4507772

Max consumption per person per year: 370 by France

Spirits Consumption

Average consumption per person per year (across all countries): 80.9948187

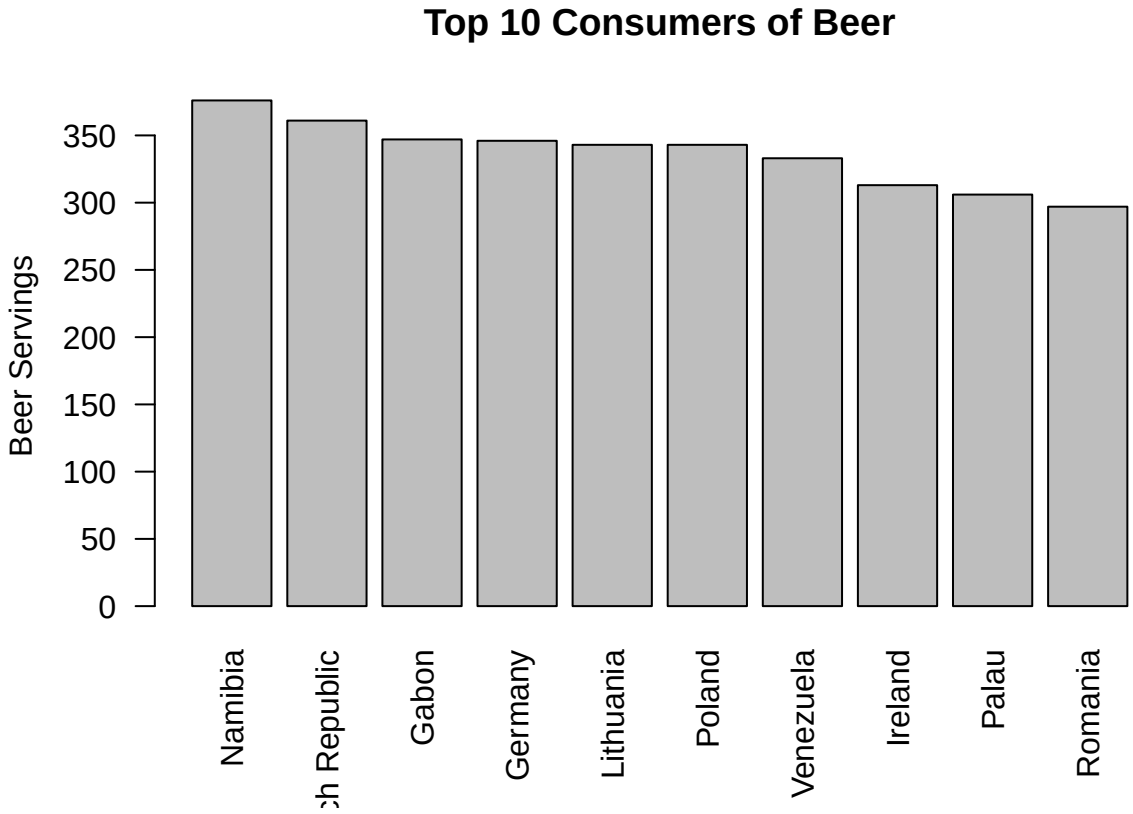
Max consumption per person per year: 438 by Grenada

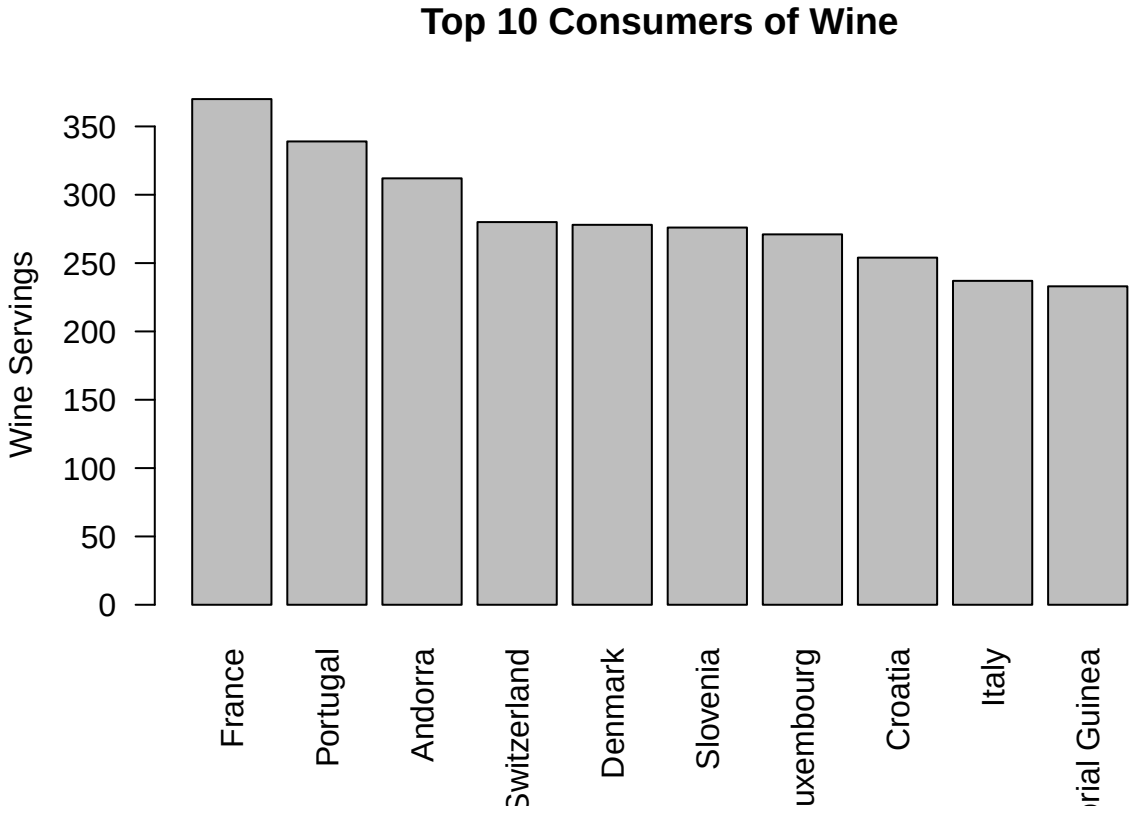
Overall Alcohol Consumption (pure)

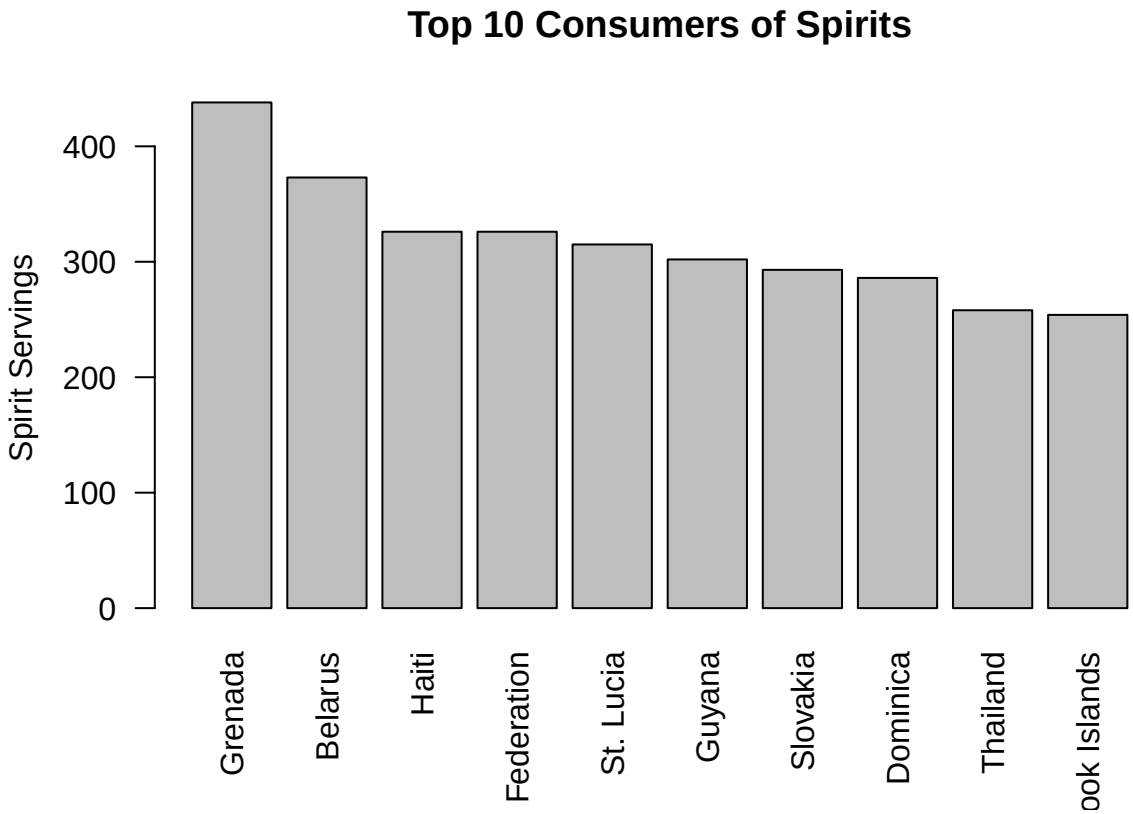
Average consumption per person per year (across all countries): 4.7170984

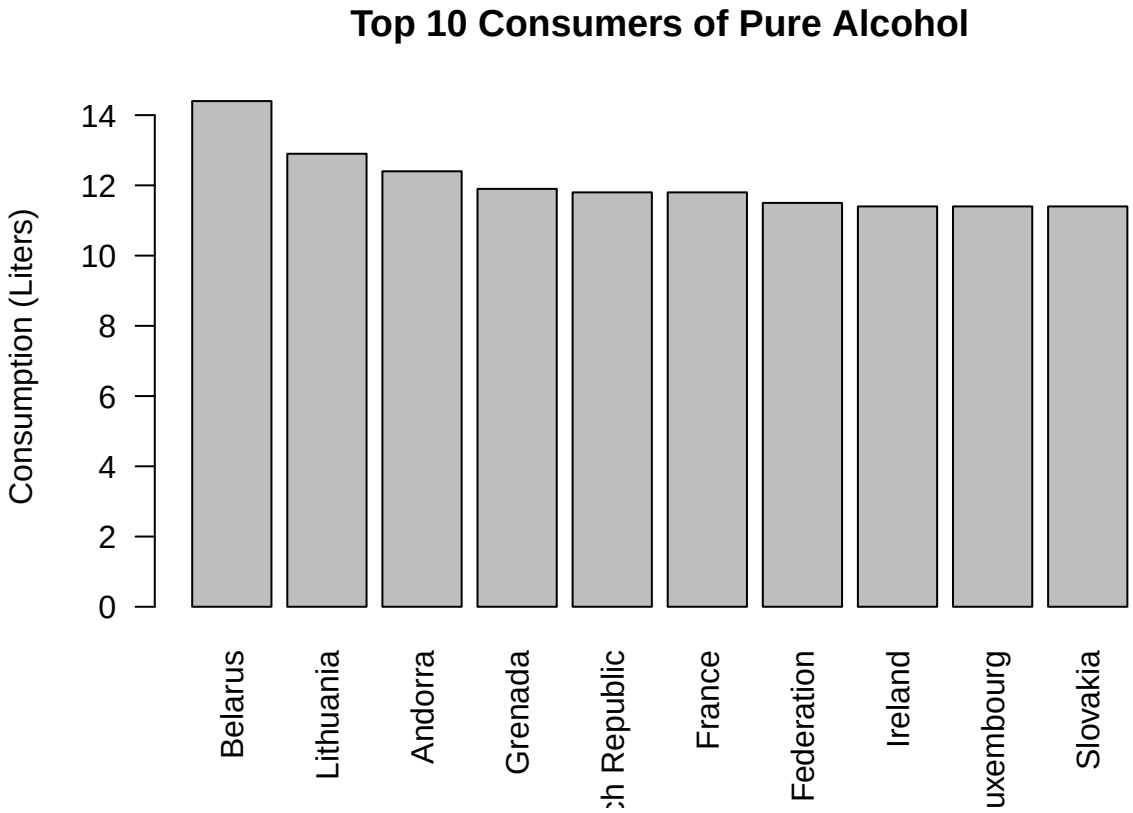
Max consumption per person per year: 14.4 by Belarus

Since there are 193 countries in this list, we can't look at everyone at once. To give a bit more perspective, here are some visualizations just for the top 10 countries in each category:


Top 10 Consumers of Beer
Beer Servings
0
50
100
150
200
250
300
350
Namibia
ch Republic
Gabon
Germany
Lithuania
Poland
Venezuela
Ireland
Palau
Romania


Top 10 Consumers of Wine
Wine Servings
0
50
100
150
200
250
300
350
France
Portugal
Andorra
Switzerland
Denmark
Slovenia
uxembourg
Croatia
Italy
orial Guinea


Top 10 Consumers of Spirits
Spirit Servings
0
100
200
300
400
Grenada
Belarus
Haiti
Federation
St. Lucia
Guyana
Slovakia
Dominica
Thailand
ook Islands


Top 10 Consumers of Pure Alcohol
Consumption (Liters)
0
2
4
6
8
10
12
14
Belarus
Lithuania
Andorra
Grenada
ch Republic
France
Federation
Ireland
uxembourg
Slovakia