Data Analytics

Assignment 6

Probabilistic Approaches

Group members:

• Nanxun Xie

• Yifan Zhao

Using System: RStudio

Methods

In this assignment, we choose four categories (including chocolate) as our final variables to test.

A list of countries ranked in terms of Nobel laureates per capita was downloaded from Wikipedia (http://en.wikipedia.org/wiki/List_of_countries_by_Nobel_laureates_per_capita). Because the population of a country is substantially higher than its number of Nobel laureates, the numbers had to be multiplied by 10 million. Thus, the numbers must be read as the number of Nobel laureates for every 10 million persons in a given country.
Data on per capita yearly chocolate consumption in 23 countries was obtained from statistia (http://www.statista.com/statistics/263779/per-capita-consumption-of-chocolate-in-selected-countries-in-2007/), Hatena Blog (http://nbakki.hatenablog.com/entry/2014/02/15/000000).
Data were available from 2013 for 17 countries, from 2012 for 5 countries (US, Australia, Netherlands, Canada, Japan), from 2008 for 1 country(China).

Data on per capita yearly alcohol consumption in 26 countries was obtained from wikipedia (https://en.wikipedia.org/wiki/List_of_countries_by_alcohol_consumption_per_capita).
Separately, these data were available from 2011 for 16 countries, from 2012 for 1 countries(New Zealand), from 2010 for 2 country(United States,Australia),from 2009 for 5 countries(Italy, Greece, Chile, Netherlands, Spain),from 2008 for 1 countries(Belgium),from 2007 for 1 countries(Portugal)

Data on per capita yearly cigarette consumption per capita in 23 countries was obtained from wikipedia (https://en.wikipedia.org/wiki/List_of_countries_by_cigarette_consumption_per_capita)Data were available from 2014 for all countries.

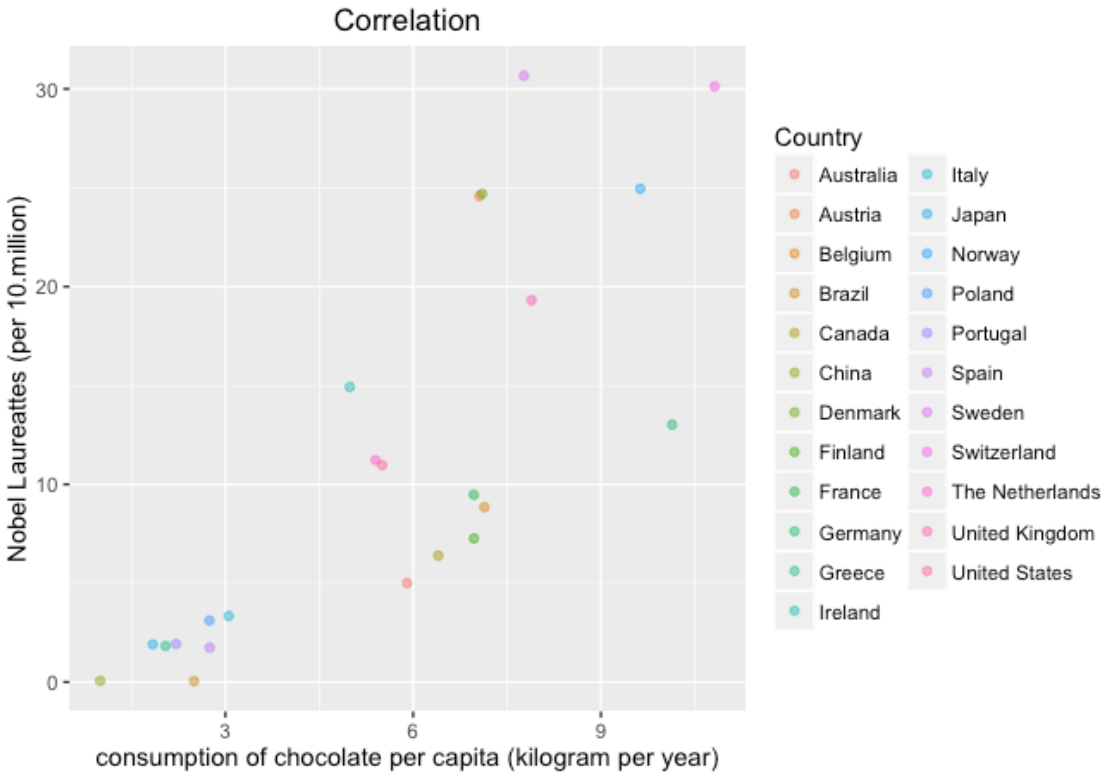
Data on per capita yearly tea consumption per capita in 22 countries was obtained from wikipedia (https://en.wikipedia.org/wiki/List_of_countries_by_tea_consumption_per_capita)Data were available as 2014 for all countries.

Analyzing

Here using the Chocolate Consumption to show the procedure.

```
#using ggplot2 to draw pictures
library(ggplot2)
#getting data from Chocolate.csv to T1
T1<-read.csv(file="~/Desktop/DA/Assign6/Data/Chocolate.csv", header =
T)
#drawing picture between chocolate consumption and Nobel

ggplot(T1,aes(x=Chocolate,y=Nobel))
+geom_point(aes(colour=Country),alpha=0.5,position =
'jitter')+labs(x='consumption of chocolate per capita (kilogram per
year)',y="Nobel Laureattes (per 10.million)",title='Correlation')
```




```
# calling lm
Retention.reg <- lm(T1$Nobel ~ T1$Chocolate, data = T1)
```

# calling summary to produce the result summaries
summary(Retention.reg)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -4.5228     2.8361  -1.595    0.126
T1$Chocolate   2.8124     0.4565   6.161 4.11e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.083 on 21 degrees of freedom
Multiple R-squared:  0.6438,    Adjusted R-squared:  0.6268
F-statistic: 37.96 on 1 and 21 DF,  p-value: 4.112e-06
```

# calling cor to calculate the correlation
cor(T1$Chocolate,T1$Nobel)

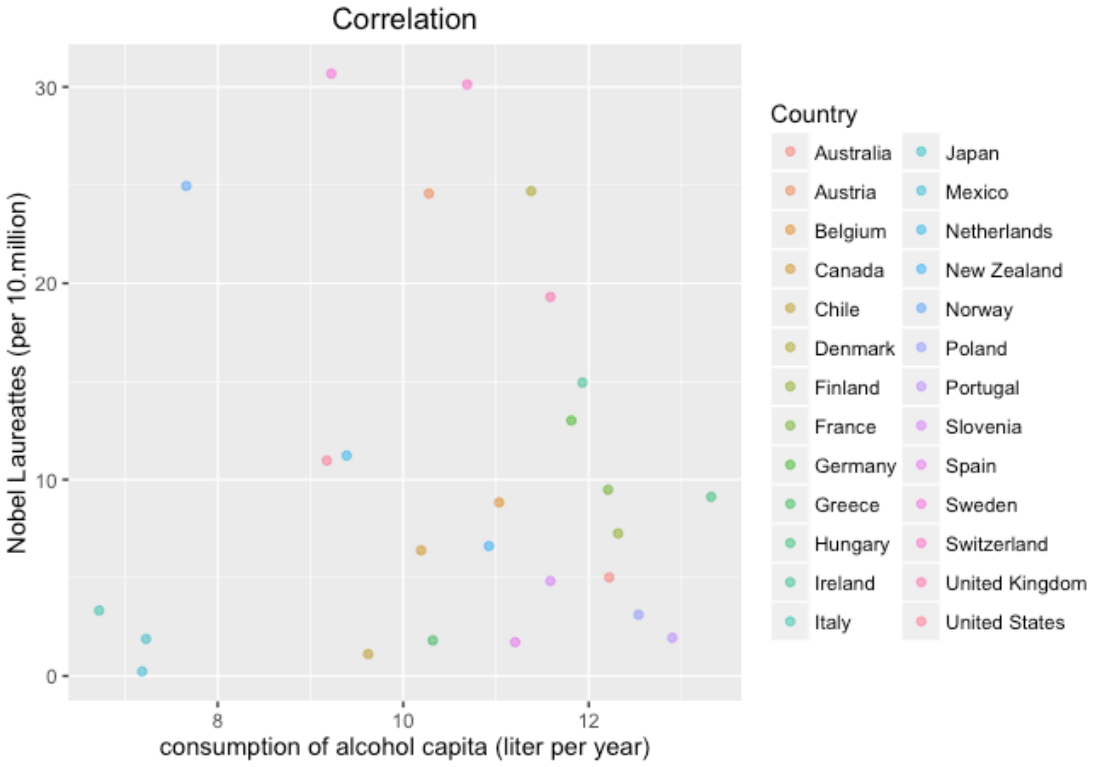
```
> cor(T1$Chocolate,T1$Nobel)
[1] 0.8023754
```

#deleting the data in Switzerland, then calling cor to calculate the new correlation
T1<-T1[T1$Country!="Switzerland",]
cor(T1$Chocolate,T1$Nobel)

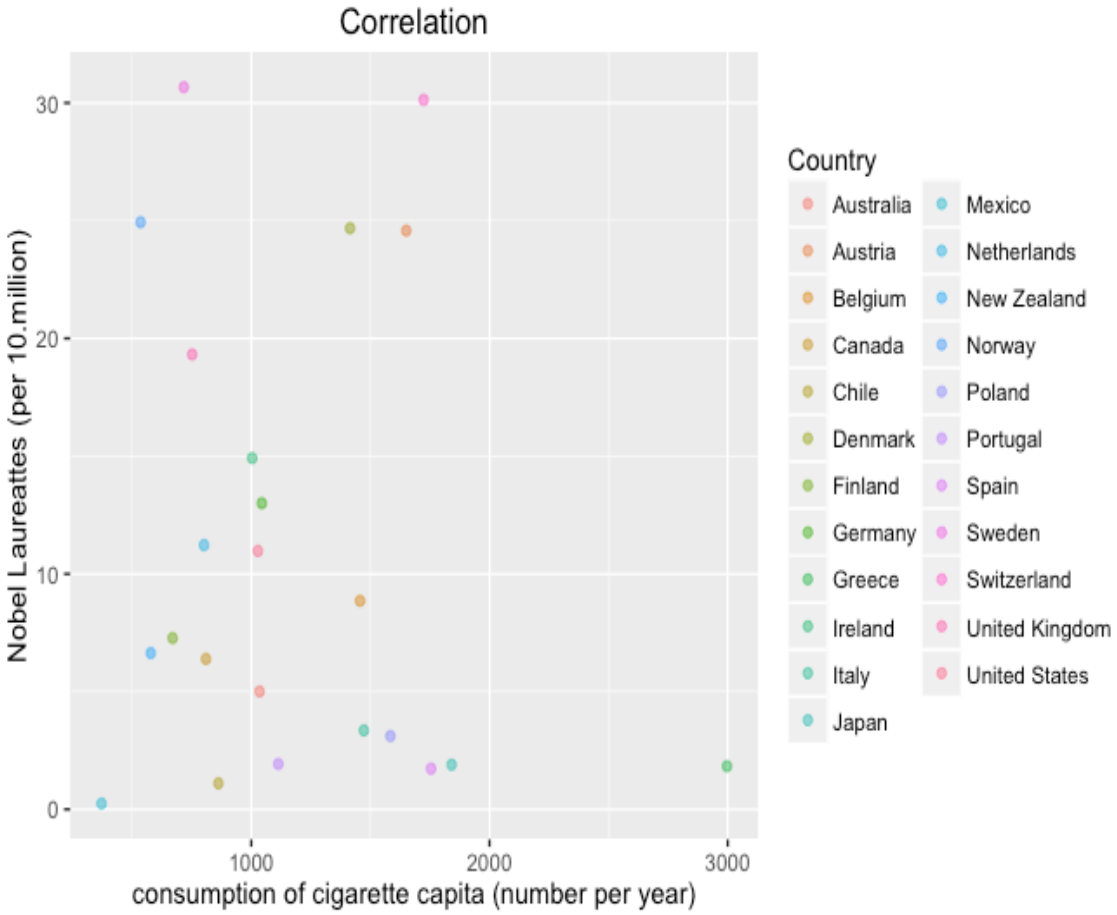
```
> T1<-T1[T1$Country!="Switzerland",]
> cor(T1$Chocolate,T1$Nobel)
[1] 0.7625107
```

Plots

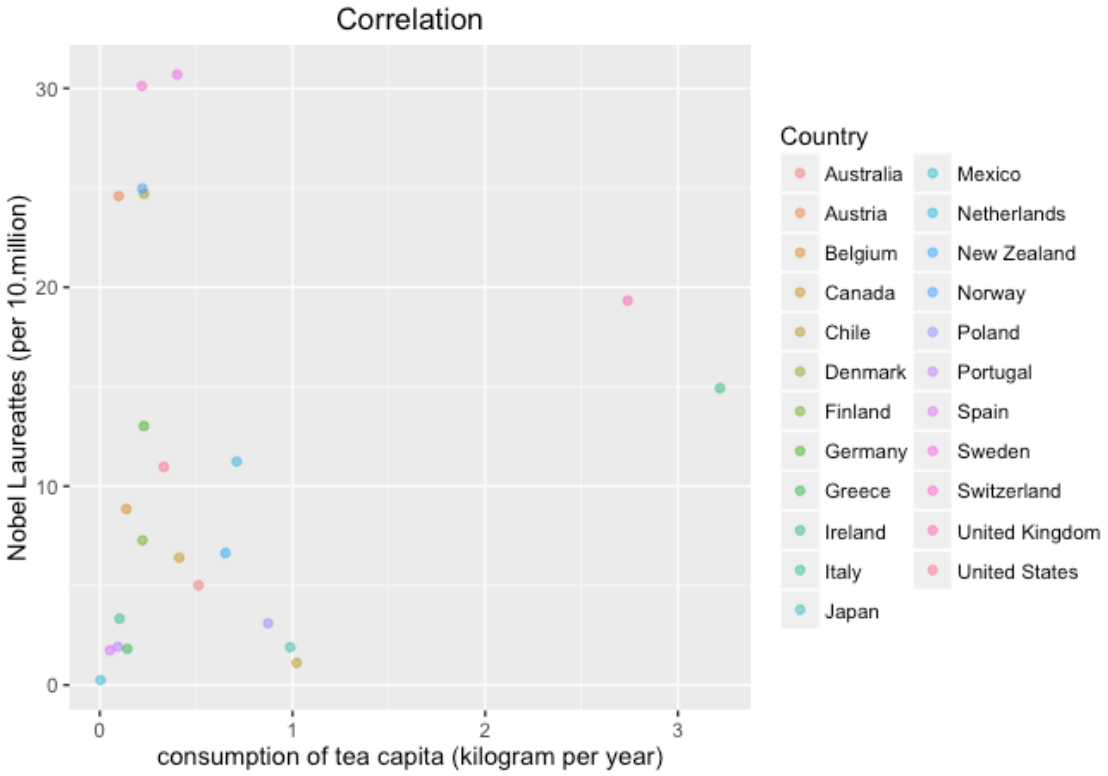
#Alcohol



#Cigarette


Correlation
Nobel Laureattes (per 10.million)
30
20
10
0
1000
2000
3000
consumption of cigarette capita (number per year)
Country
Australia
Austria
Belgium
Canada
Chile
Denmark
Finland
Germany
Greece
Ireland
Italy
Japan
Mexico
Netherlands
New Zealand
Norway
Poland
Portugal
Spain
Sweden
Switzerland
United Kingdom
United States

#Tea



Results

By comparing the correlations between four variables with Noble, we find that only between chocolate consumption and the number of Noble has a significant linear correlation in these total 23 countries. And when we delete the value under “Switzerland”, which has the max value of the chocolate consumption, the new correlation coefficient immediately decreases from 0.8023754 to 0.7625107.