CSE332 Spring 2018 Homework 2

Doeun Kim

4/15/2018

Task

I would like to analyze the data of life expectancy, air pollution, homicide, and assault by continent and conclude with several facts for each continent and overall.

Gather the Data

If I read the data from excel file, the data type becomes a string although it must be a float. To resolve this problem, I copied the content of 'ScoreData' sheet of the 'BetterLifeIndex_Data_2011V6.xls' file into empty csv file. Therefore, as I read the csv file, all the data were properly read.

```r
# raw data
raw_data <- read.csv("BetterLifeIndex_Data_2011V6.csv", header=F)
```

Preprocess the Data

Since there were some headings above the table in the csv file, I excluded first 6 rows that are corresponding to the headings. Then, I collected 5 columns of data and assigned to each corresponding variable. Also, I manually modified the value 10 to 9.99 because R didn't indicate the value 10 properly in a graph.

```r
# pre-processing raw data
raw_data <- raw_data[7:40,]
# gather country name, air pollution rate, life expectancy rate, homicide
rate, and assault rate
country_name <- raw_data[,2]
air_pollution <- raw_data[,12]
life_exp <- raw_data[,15]
homicide <- raw_data[,18]
assault <- raw_data[,19]
```

Since I want to analyze the data by continent, I manually labeled the continent name for each country. The continent array will be used for the color of the graph lines later.

```r
# label continent name for each country for coloring the graph lines later
continent<-c('Oceania','Western Europe','Western Europe','North
America','South America','Eastern Europe','Northern Europe','Northern
Europe', 'Northern Europe','Western Europe','Western Europe','Southern
Europe','Eastern Europe','Northern Europe','Western Europe','Middle East
Asia','Southern Europe','East Asia','East Asia','Western Europe','South
America','Western Europe','Oceania','Northern Europe','Eastern
Europe','Southern Europe','Middle Europe','Middle Europe','Southern
Europe','Northern Europe','Middle Europe','West Asia',"Western Europe","North
America")
```

Then, I put all the data arrays together into the data frame which will be used for plotting a parallel coordinate plot.

```
# create data frame for graph
df <- data.frame(country=country_name,life_expectancy=life_exp,
                 air_pollution = air_pollution, homicide=homicide,
                 assault=assault, continent=continent)
```

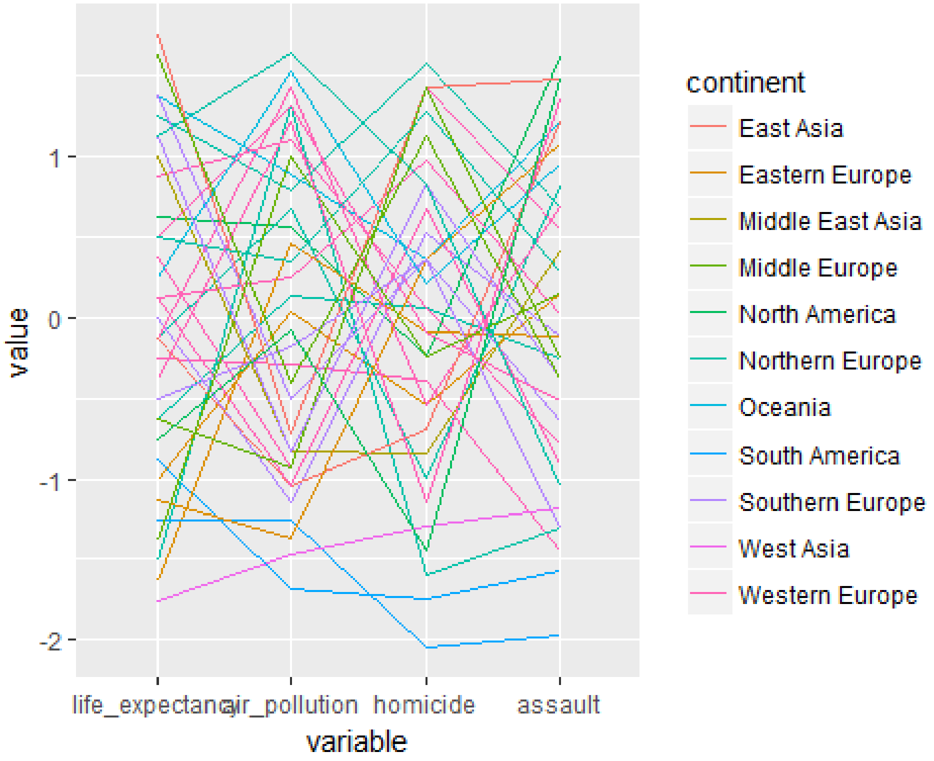
Plot a Parallel Coordinate Plot And Analysis

I decided to use parallel coordinate plot because multiple features are needed for my analysis, the number of objects are not that many, and this kind of plot helps me to distinguish the lines by color. In order to plot a parallel coordinate plot, I used GGally library which will help me to draw a parallel coordinate plot easily. In the function ggparcoord, the attribute 'groupColumn' assigns the color for each continent, and the value 'std' of the attribute 'scale' subtracts mean and divide by standard deviation.

Overall

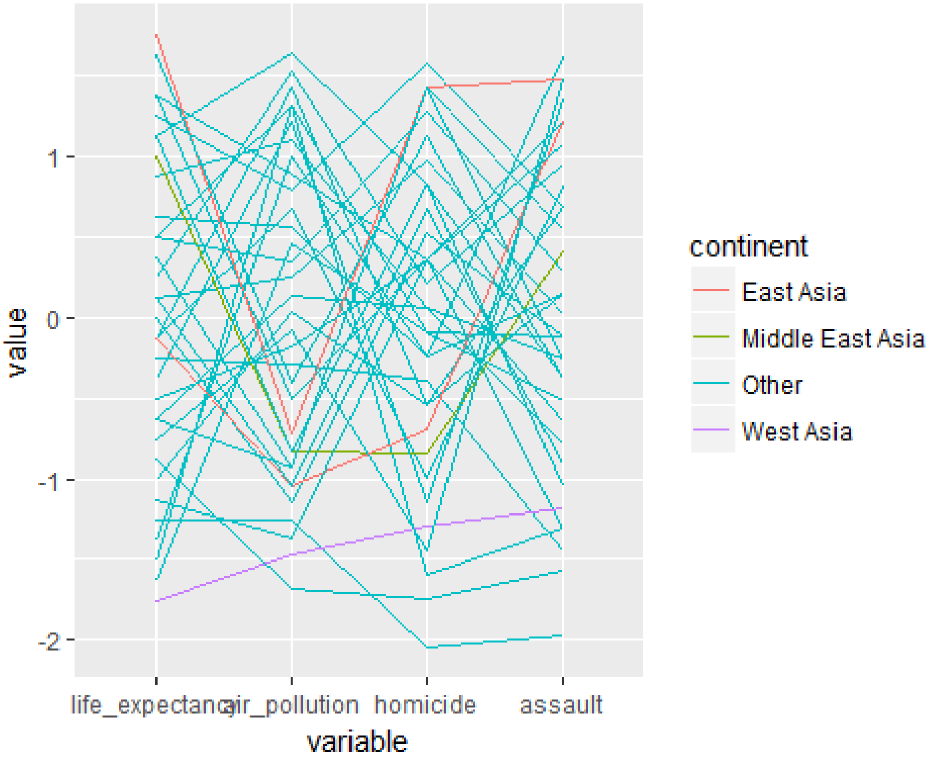
Since there are so many European countries in the list, I decided to specify the continent deeper. So, I used specific region names of continents as lables. But this makes hard to recognize and distinguish the color clearly. Therefore, I decided to plot the graph and specify the regions for each continent.

```
# plot parallel coordinate graph (scale min is 0 and max is 1)
library(GGally)
ggparcoord(df, columns = 2:5, groupColumn = 'continent', scale = 'std')
```



Asia

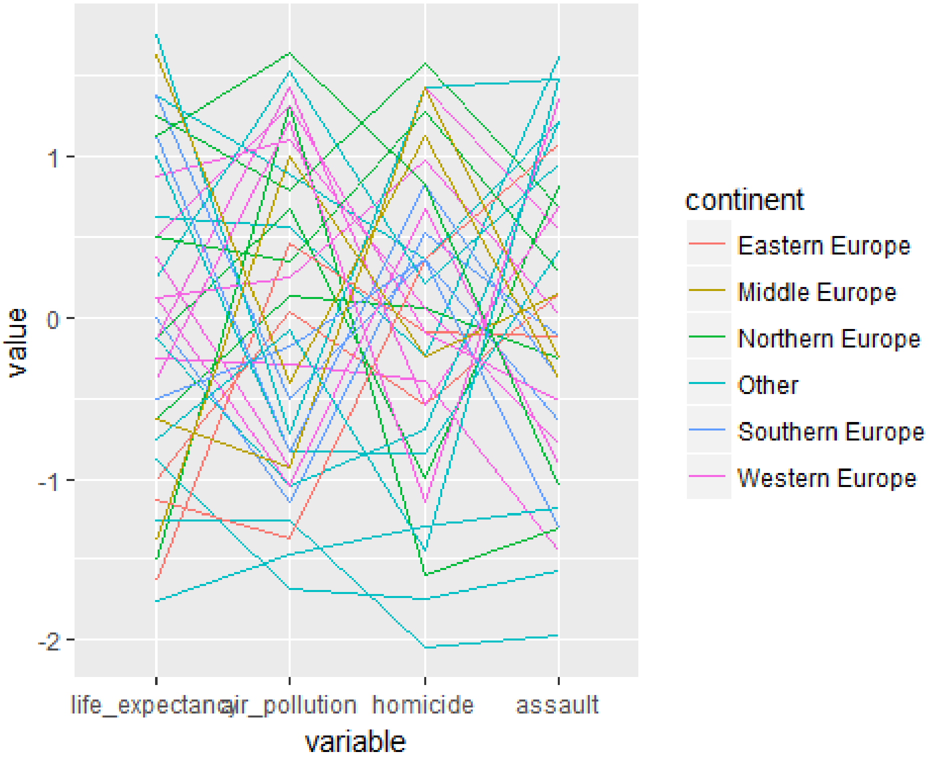
```
# plot parallel coordinate graph for Asia
continent<-c('Other','Other','Other','Other','Other','Other','Other','Other',
'Other','Other','Other','Other','Other','Other','Other','Middle East
Asia','Other','East Asia','East
Asia','Other','Other','Other','Other','Other','Other','Other','Other','Other'
,'Other','Other','Other','West Asia','Other','Other')
df <- data.frame(country=country_name,life_expectancy=life_exp,
                 air_pollution = air_pollution, homicide=homicide,
                 assault=assault, continent=continent)
ggparcoord(df, columns = 2:5, groupColumn = 'continent', scale = 'std')
```



- East Asian countries have worse air pollution than average, but are usually safe to live.
- East Asians usually live longer than the bottom 50% countries people.
- Middle East Asians live longer than their unsafety and bad air quality.
- West Asians live the shortest in consequence of the second most severe air pollution and the relatively low public security.

Europe

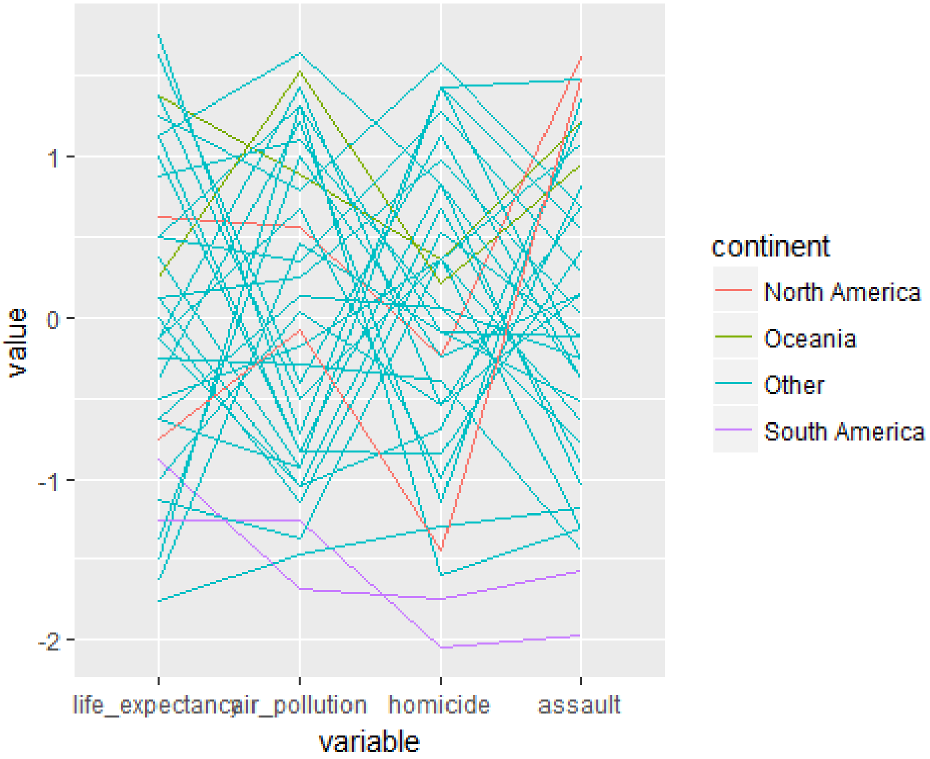
```
# plot parallel coordinate graph for Europe
continent<-c('Other','Western Europe','Western
Europe','Other','Other','Eastern Europe','Northern Europe','Northern Europe',
'Northern Europe','Western Europe','Western Europe','Southern
Europe','Eastern Europe','Northern Europe','Western Europe','Other','Southern
Europe','Other','Other','Western Europe','Other','Western
Europe','Other','Northern Europe','Eastern Europe','Southern Europe','Middle
Europe','Middle Europe','Southern Europe','Northern Europe','Middle
Europe','Other','Western Europe','Other')
df <- data.frame(country=country_name,life_expectancy=life_exp,
                 air_pollution = air_pollution, homicide=homicide,
                 assault=assault, continent=continent)
ggparcoord(df, columns = 2:5, groupColumn = 'continent', scale = 'std')
```



- Eastern Europeans live shorter than other regions Europeans.
- Compared to good security, Eastern Europeans live short.
- Middle European countries have good security and good air quality in average, but have low life expectancy.
- Northern European countries have good air quality, but half of them have high security and life expectancy and half of them are not.
- Southern European countries have a lower air quality, more assaults, fewer homicides, and higher life expectancy than the average.
- Western European countries have a higher life expectancy, but half of them have worse air quality, more homicides, and fewer assaults than the average, and vice versa.

North And South America, And Oceania

```
# plot parallel coordinate graph for Europe
continent<-c('Oceania','Other','Other','North America','South
America','Other','Other','Other',
'Other','Other','Other','Other','Other','Other','Other','Other','Other','Othe
r','Other','Other','South
America','Other','Oceania','Other','Other','Other','Other','Other','Other','O
ther','Other','Other','Other','North America')
df <- data.frame(country=country_name,life_expectancy=life_exp,
                 air_pollution = air_pollution, homicide=homicide,
                 assault=assault, continent=continent)
ggparcoord(df, columns = 2:5, groupColumn = 'continent', scale = 'std')
```



- North American countries have the fewest assaults.
- Oceanian countries have high rate for every features.
- South American countries have the most assaults and homicides.

Conclusion

- The countries where the air quality is bad mostly have high security, and vice versa.
- If the country’s life expectancy is remarkably low, its environment and security are remarkably low also.
- Proportional relationships: life expectancy and homicide, air pollution and assult
- Inversely proportional relationships: life expectancy and air pollution, air pollution and homicide, homicide and assault
- South America and West Asia are the worst continents to live and need to resolve environmental and security problems in order to increase the life expectancy.
- Oceania is the best continent to live.