Assignment_3

Syam Sundar Herle

September 27, 2017

Getting Data

The data are read into the R,

```
setwd("C:/Studies/Semester3/EDA/prob3/data")
dir_path =getwd()
dir_path
```

```
## [1] "C:/Studies/Semester3/EDA/prob3/data"
```

```
bp=read.csv("Q1.csv",header = FALSE,fileEncoding="UTF-8-BOM")
ln=read.csv("Q2.csv",header = FALSE,fileEncoding="UTF-8-BOM")
rn = read.csv("Q3.csv",header = FALSE,fileEncoding="UTF-8-BOM")
ni =read.csv("Q4.csv",header = FALSE,fileEncoding="UTF-8-BOM")
```

Score standardization

```
bp <- transform(bp, score = (4 * bp[2] + 3*bp[3]+2*bp[4]+1*bp[5])/(bp[2]+bp[3]+bp[4]+bp[5]) )
ln <- transform(ln, score = (4 * ln[2] + 3*ln[3]+2*ln[4]+1*ln[5])/(ln[2]+ln[3]+ln[4]+ln[5]) )
rn <- transform(rn, score = (4 * rn[2] + 3*rn[3]+2*rn[4]+1*rn[5])/(rn[2]+rn[3]+rn[4]+rn[5]) )

ni <- transform(ni, score = (4 * ni[2] + 3*ni[3]+2*ni[4]+1*ni[5])/(ni[2]+ni[3]+ni[4]+ni[5]) )
countries <- bp[1]

bp$V2.1 <- round((bp$V2.1 - mean(bp$V2.1))/sd(bp$V2.1),digits = 3)

ln$V2.1 <- round((ln$V2.1 - mean(ln$V2.1))/sd(ln$V2.1),digits = 3)


rn$V2.1 <- round((rn$V2.1 - mean(rn[1:13,"V2.1"]))/sd(rn[1:13,"V2.1"]),digits = 3)

ni$V2.1 <- round((ni$V2.1 - mean(ni$V2.1))/sd(ni$V2.1),digits = 3)

#Replacing value of japan
rn[14,8] <- 0
```

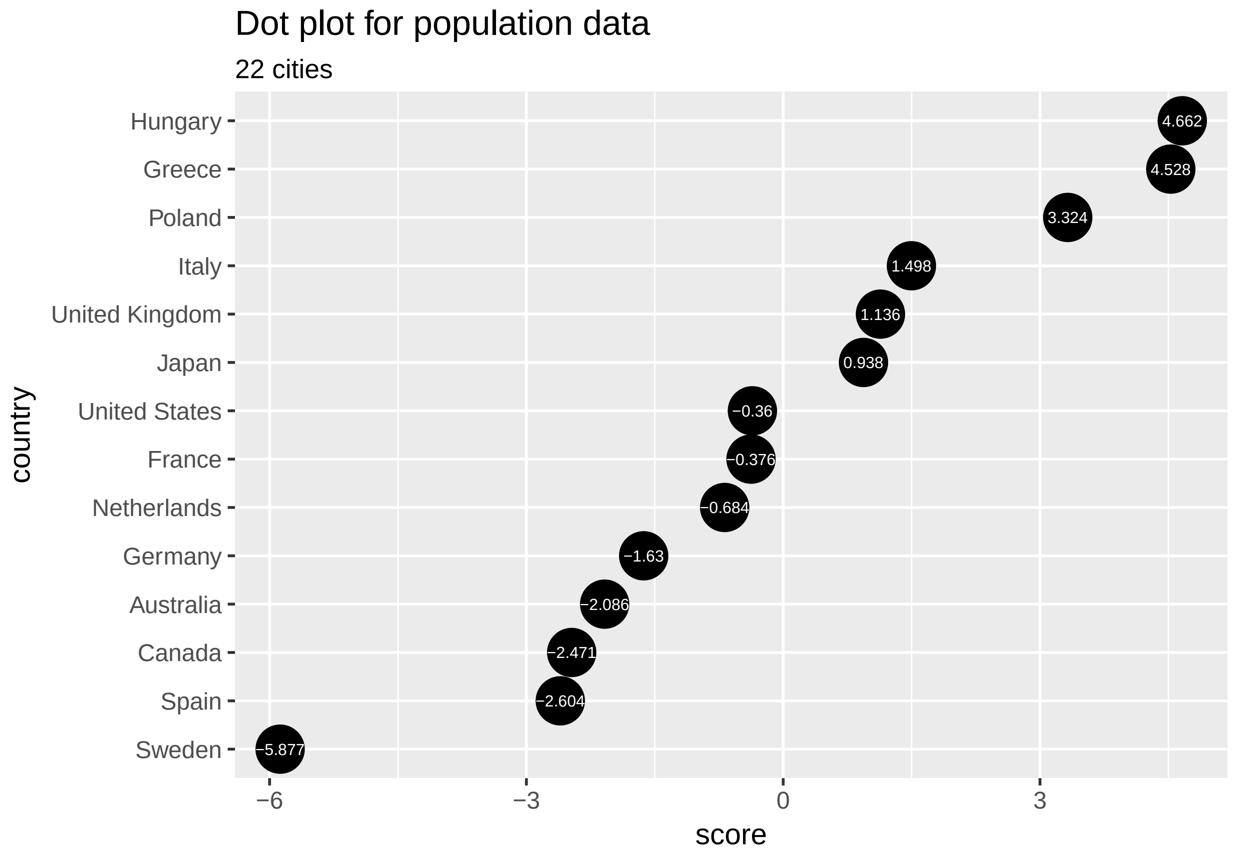
Solution 1

Univariate Analysis

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```
library(grid)
#Rounding all the values to three digit decimal place
vec <- round(bp$V2.1 + ln$V2.1 +rn$V2.1 +ni$V2.1,digits = 3)
univariate <- cbind(countries,vec)
colnames(univariate) <- c("country", "score")
univariate <- univariate[order(univariate$score), ]
univariate$`country` <- factor(univariate$`country`, levels = univariate$`country`)
ggplot(univariate, aes(x=`country`, y=score, label=score)) +
geom_point(stat='identity', fill="black", size=8) +
geom_text(color="white", size=2) +
labs(title="Dot plot for population data",subtitle="22 cities")+coord_flip()
```



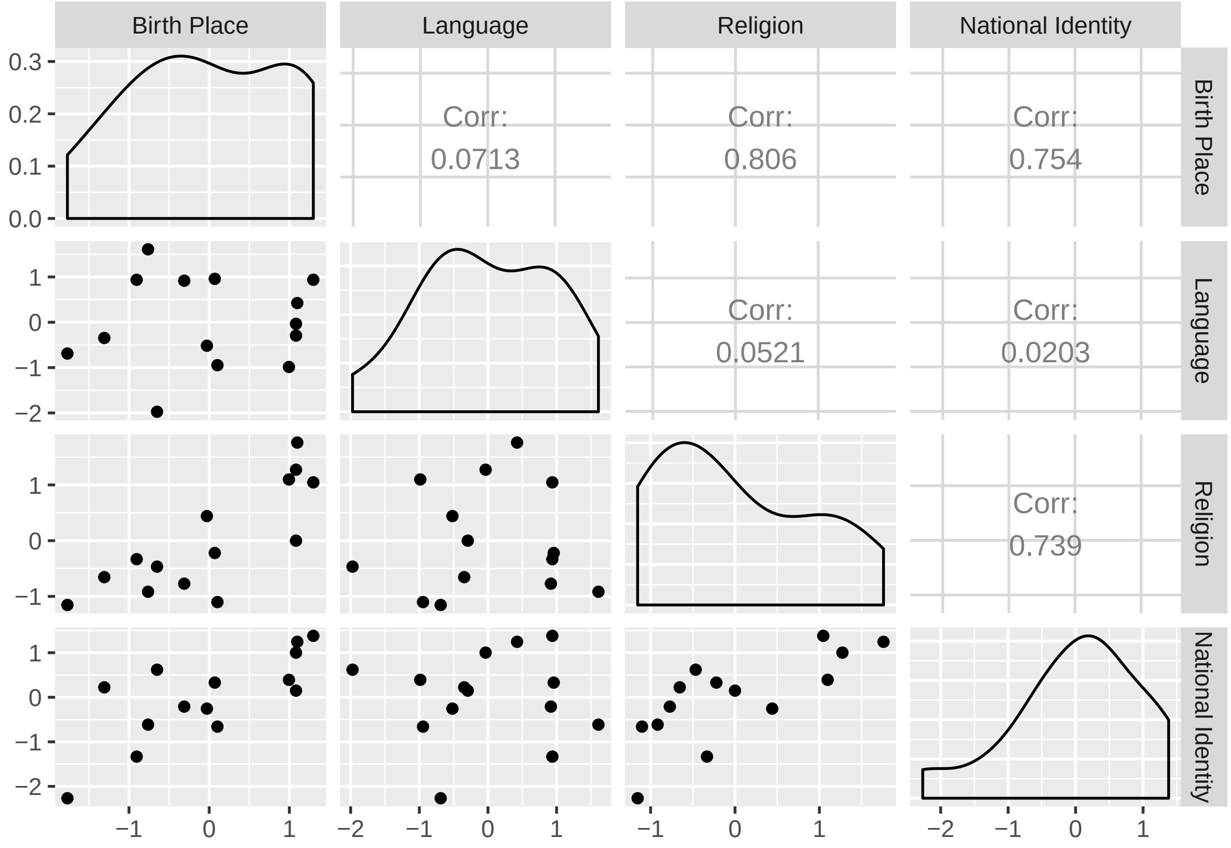
Solution 2

Bivariate Analysis

```
#Binding the data to form a dataframe
bivariate <- cbind.data.frame(bp$V2.1,ln$V2.1,rn$V2.1,ni$V2.1)
colnames(bivariate) <- c("Birth Place","Language","Religion","National Identity")
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 3.3.3
```

```
ggpairs(bivariate)
```



From above scatter plot we can see that the following pairs are strong related,

- Religion and Birth place
- National Identity and Birth place
- National Identity and Religion

The following pairs are weakly related,

- Language and Birth place
- Language and Religion
- Language and National Identity

So it is clear that the variable ‘Language’ is weakly realted to other three.

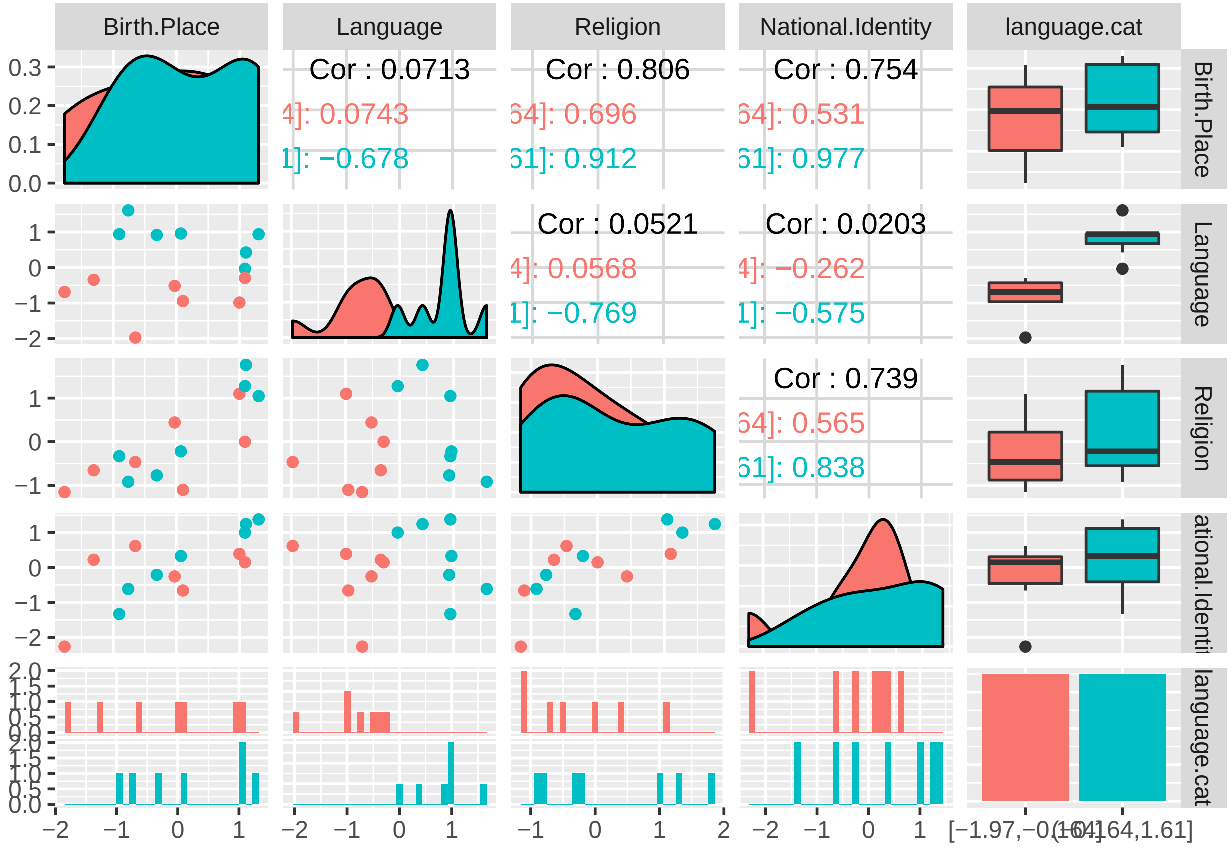
Solution 3

Trivariate Analysis

The Religion,National Identity and Birth Place are strongly related variables while languge is the weak variable

```
# the language is divided into two with respect to its mean
language.cat = cut_number(bivariate$Language, n = 2)
ggpairs(data.frame(bivariate, language.cat), aes(color = language.cat))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From the above trivariate plot we can see the inner data of the Religion,National Identity and Birth Place variables but we were not able to see the distribution clearly in the Bivariate anlysis and bivariate may be misleading.

From the above scatter plot we can see that the mean of Religion,National Identity and Birth Place with respect to language is same (From the whisker plot).But from the density plot we couldn’t infer anything about the pattern with respect to language, this may be due to the less availibity of the data.