EDA - Assignment 4

Jivitesh Poojary

February 14, 2017

Answer 1

Q1. Univariate analysis. Find the total score for each country (adding up the standardized scores for all four questions) and display the results in a table or graph.

The table is provided below.

```
##           Country  Birthplace    Language   Religion    Customs SumOfScore
## 1   United States -0.02846888 -0.51765024  0.4409230 -0.2549215 -0.3601177
## 2          Canada -0.64964149 -1.97338084 -0.4659513  0.6177590 -2.4712147
## 3          France -0.31082007  0.91732128 -0.7715873 -0.2100277 -0.3751138
## 4         Germany -0.90375756  0.93808037 -0.3322356 -1.3323728 -1.6302856
## 5          Greece  1.09910058  0.42429309  1.7602245  1.2454293  4.5290474
## 6         Hungary  1.29858169  0.93808037  1.0460204  1.3787639  4.6614464
## 7           Italy  0.99533652 -0.98559454  1.0980697  0.3909995  1.4988112
## 8     Netherlands -0.76215416  1.60929067 -0.9167644 -0.6140719 -0.6836998
## 9          Poland  1.08254025 -0.03413663  1.2729319  1.0021049  3.3234405
## 10          Spain  0.10296559 -0.94580630 -1.1020365 -0.6589657 -2.6038429
## 11         Sweden -1.76835111 -0.68891266 -1.1540105 -2.2657149 -5.8769892
## 12 United Kingdom  0.07035403  0.95842838 -0.2209023  0.3286979  1.1365780
## 13      Australia -1.30822563 -0.34638781 -0.6546815  0.2231975 -2.0860975
## 14          Japan  1.08254025 -0.29362515  0.0000000  0.1491227  0.9380378
```

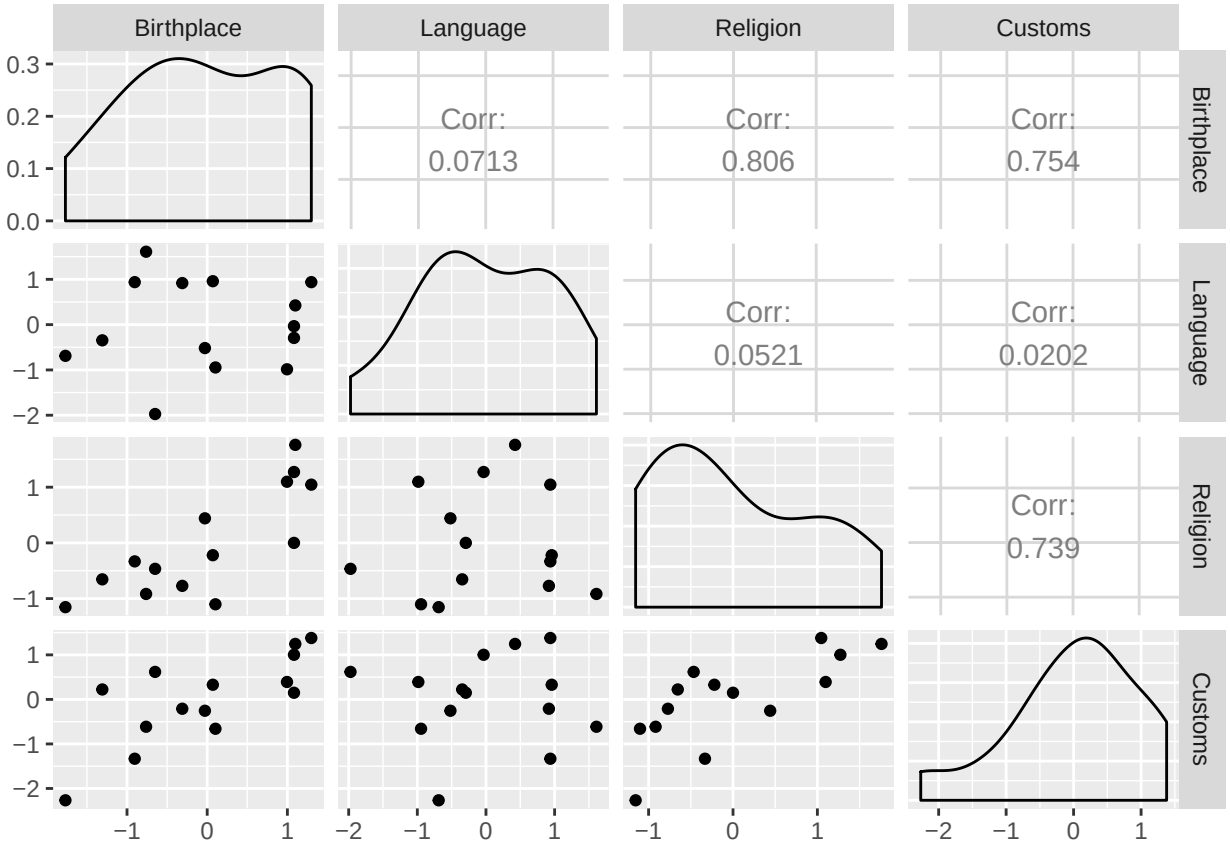
Answer 2

Q2. Bivariate analysis. Use ggpairs() in the GGally library to create a scatterplot matrix. There should be 4-choose-2 = 6 pairs of variables plotted. Identify which of these six pairs are strongly related, and which are weakly related.

From the plot below we can say the following:

Strongly related pairs: - 0.806 :Religion - Birthplace, 0.754 :Customs - Birthplace, 0.739 :Customs - Religion

Weakly related pairs: - 0.0202 :Customs - Language, 0.0521 :Religion - Language, 0.0713 :Language - Birthplace



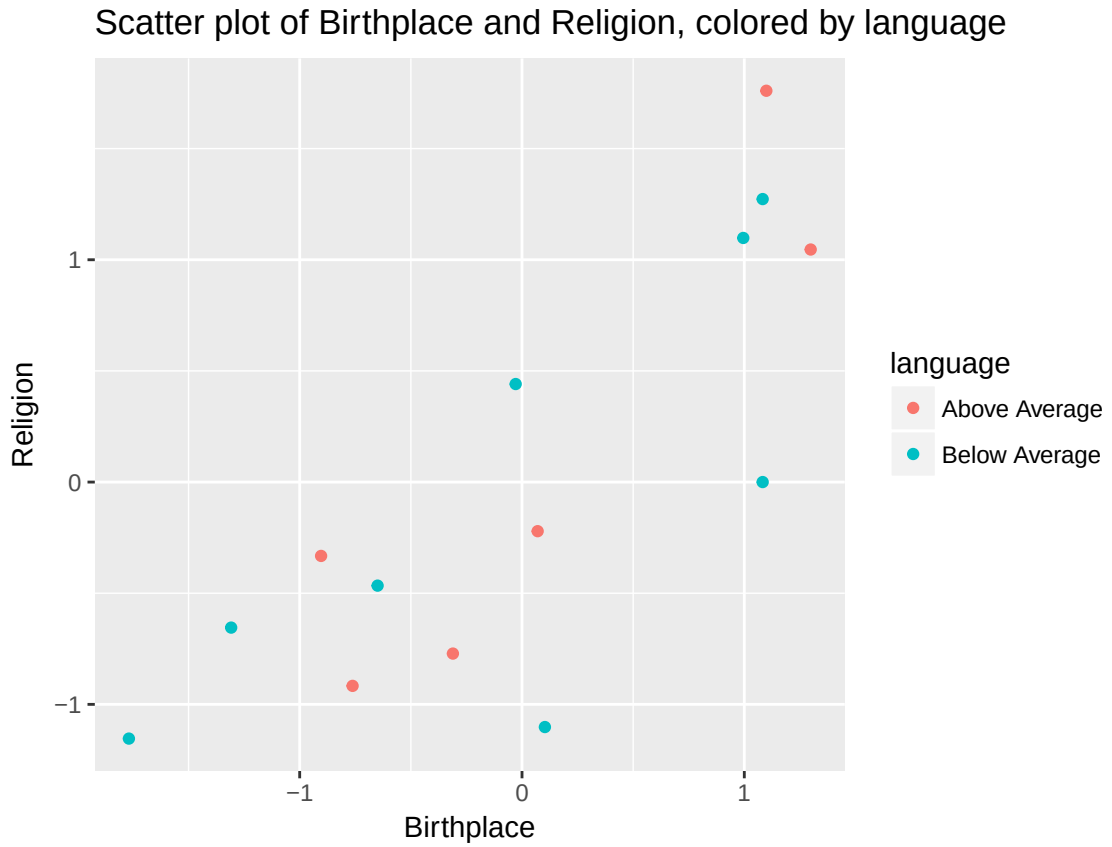
Answer 3

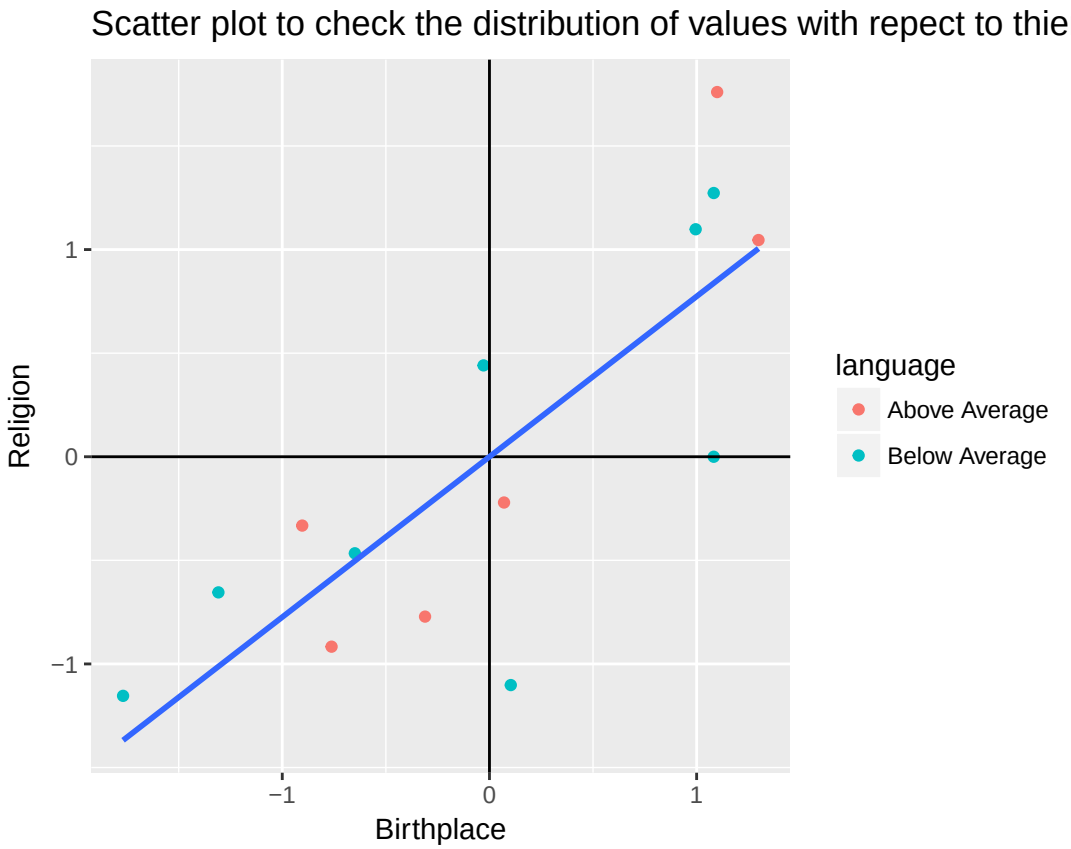
Q3. Trivariate analysis. You should find that three of the variables are quite strongly related, while the other variable is more weakly related. Draw a scatterplot of the two most strongly correlated variables. Then color the points according to the weakly related variable (e.g. make the points where the weakly related variable is above average one color, and the weakly related variable is below average another color.) What does this tell you that was not obvious from the bivariate analysis?

The two most strongly related pairs were Religion - Birthplace 0.806. And the weakly related variable was language. The mean value of standardized scores for the weakly related variable is 2.142857e-10. On an average the standardized scores of the birthplace are slightly higher than that calculated for religion.

The trivariate analysis shows that 'teal' colored points on the scatter plot are the ones below average on the standardized language score. while the 'red' colored points on the scatter plot are the ones above average on the standardized scores. From the plot we can see that the points cannot be linearly seperated.

Another observation is that most of the countries which score below average in language tend to score higher in birthplace than in religion. However for countries which score above average in language tend to score higher in religion than in birthplace. We can also observe Given the size of the data we cannot be completely sure of our inference




Scatter plot to check the distribution of values with repect to thier mea
Religion
1
0
−1
−1
0
1
Birthplace
language
Above Average
Below Average