Report of the Big 5 personality test data analysis

Kristian Smid

Untouched Big 5 personality dataset consists of 19719 observations of 57 variables. After first look at basic statistics I found some uncertainties in age and country column, which I corrected and ended up with only 18 missing values (10 of age, 8 of country). I still do not know what does "(nu" value of country variable mean, but there is a lot of it, so I kept it there. Also, I connected another variable called "performance" to the dataset. Which means that after all the corrections, the final dataset consists of **19701** observations of **58** variables.

Variables:	
performance	1 = low performer – 5 = top performer
race	1-12 different races, 13 = other, 0 = missed
age	13 – 100
engnat – English native speaker	1 = yes, 2 = no, 0 = missed
gender	1 = male, 2 = female, 3 = other, 0 = missed
hand	1 = right, 2 = left, 3 = both, 0 = missed
source – how respondent came to the test – http	1 = another page on the test website, 2 = Google, 3 = Facebook, 4 = .edu url, 6 = other source
country	ISO code
set of 50 questions about personality	1 = disagree – 3 = neutral – 5 = agree

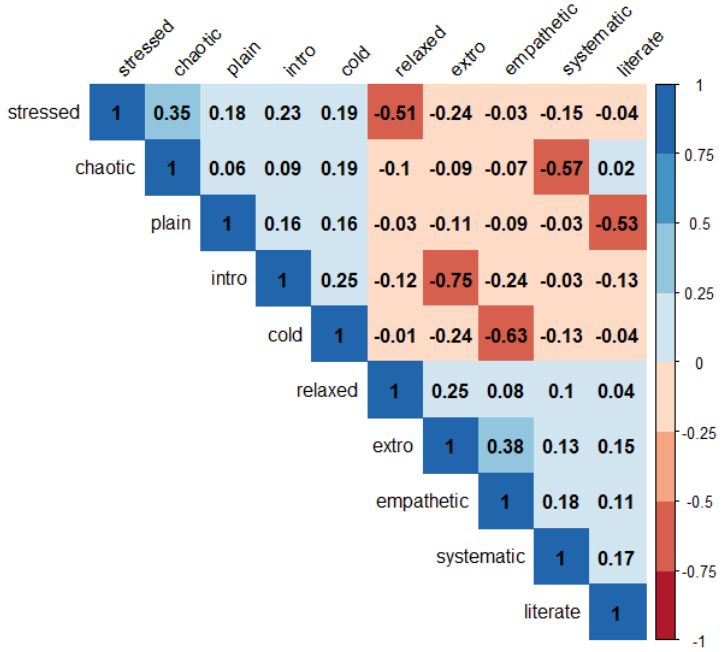
Set of fifty questions is a lot and the value of each question means something different. So, I connect them (summed up their values) based on their meaning into 10 indices: extrovert, introvert, relaxed, stressed, empathetic, cold, systematic, chaotic, literate, plain. And then divided them with the number of their questions to keep them in the same range (1-5), cause not every index had the same amount of questions. I knew I was losing exactness and working with abstract terms here, but I thought it did make a sense… and I did it for a good reason, too.

In the next step, I made a correlation test between indices and performance, to find out if there is some connection. I used Kendall's tau-b, for I think it is the best way to test two ordinal values.

Index	Corr. w/ performance	Sig
Extrovert ('I feel comfortable around people.', 'I don't mind being the center of attention.', ...)	0.00758	0.15584
Introvert ('I keep in the background.', 'I am quiet around strangers.', ...)	-0.01081	**0.04332**
Relaxed ('I am relaxed most of the time.', 'I seldom feel blue.', ...)	-0.00273	0.62366
Stressed ('I worry about things.', 'I am easily disturbed.', ...)	-0.00100	0.84983
Empathetic ('I am interested in people.', 'I feel others' emotions.', ...)	0.00809	0.13243
Cold ('I feel little concern for others.', 'I insult people.', ...)	-0.00445	0.41169
Systematic ('I am always prepared.', 'I like order.', ...)	-0.00067	0.89991
Chaotic ('I leave my belongings around.', 'I make a mess of things.', ...)	-0.00808	0.13354
Literate ('I have a rich vocabulary.', 'I have a vivid imagination.', ...)	0.00025	0.96279
Plain ('I have difficulty understanding abstract ideas.', 'I do not have a good imagination.', ...)	0.00595	0.28142

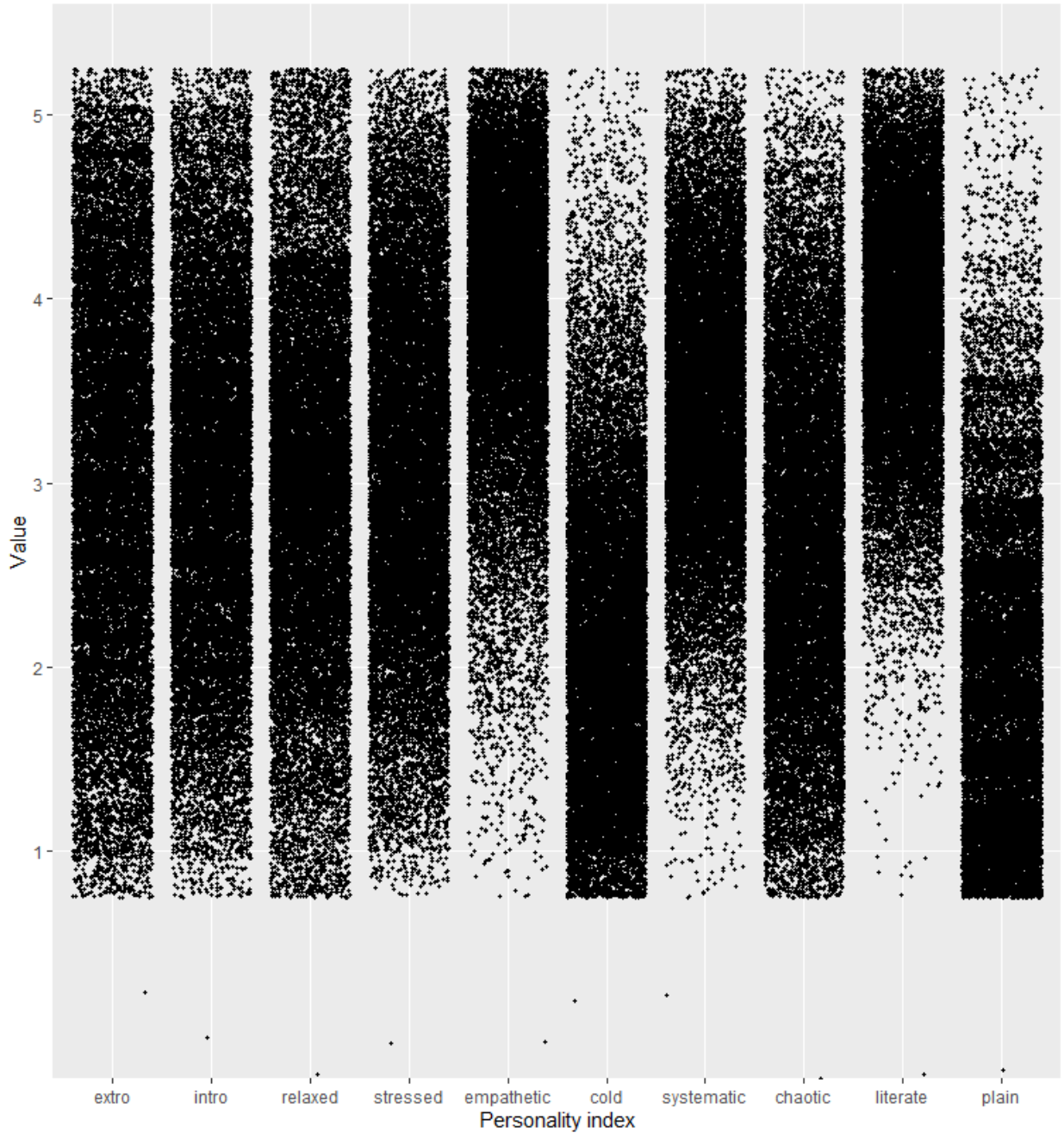
So far, I have always worked with samples only, so I cannot judge a real data correlation situation, but from my point of view it looks like there is almost none. We can only see one statistically significant super weak negative correlation at introvert index, that is all.

After this disappointment, I wanted to do something with reasonable p-value.



I do like matrixes. I think there is a lot to interpret. For example, strong positive correlation 0.35 between extro and empathetic indices, meaning that the more people tend to be extrovert, the more they are inclined to be empathetic. We can find the same relationship between stressed and chaotic indices with almost the same positive correlation value 0.38. It is interesting.

Then, I created this plot to visualize the distribution of the values of the indices.



It is important to say that I used the jitter function to create this plot, because the values, even though I basically made the mean of them, are still discrete. The jitter function just puts some randomness into them so that we can see the pattern of the distribution better. Now to the interpretation.

At first look, we can see the people are not sure or decided if they are rather extroverts or introverts. Their answers mostly spread equally from 2 to 4. Even more neutral is the relax index with most of its values around 3. On the other hand, people are decisive about their tendencies to be cold or empathetic. And confident that they are not plain but literate but that is not surprising.

Like a next step, I added little comparison between the means of indices of personality and gender. But first, it would be great to know what the distribution of men and women and other genders in this dataset looks like.

Gender	Freq
Female	11974
Male	7603
Other	100

Personality	Mean_of_Women	Mean_of_Men	Mean_of_Others
extro	3.108	3.065	2.384
intro	3.004	3.156	3.554
relaxed	2.901	3.148	2.650
stressed	3.238	2.931	3.439
empathetic	3.992	3.694	3.535
cold	2.069	2.404	2.487
systematic	3.445	3.379	3.297
chaotic	2.755	2.762	3.115
literate	3.820	3.953	4.119
plain	2.057	1.934	1.787

Let us focus on the column of the mean of other genders. We can see much lower value in extro, higher in intro, lower in relaxed, higher in stressed, then higher in chaotic. It looks like people who are uncertain about their gender tend to be more introvert, stressed and chaotic. Which I would not call a positive effect. There is only hundred of them, I know, but we can still read some tendency of it.

Finally, there is a little comparison between gender and its mean of performance and the mean of age. We can see that men have a slightly higher mean of performance and that the mean of age is unbelievably equal.

Gender	Mean of Performance	Mean of Age
Female	26.02255	2.974612
Male	26.70130	2.966461
Other	23.04000	2.950000