Young people survey analysis using Python

By Krystian Plackowski

Abstract

The work aims to answer questions about social behaviour of young people (age 15-30). The dataset consists of survey responses of 1010 participants from Slovakia. Questions stated about preferences, hobbies, phobias and personality.

The analysis starts with searching for personality features which have the most interesting correlations with any of answers provided in survey. Next point, people are being divided into groups related to their fobias. Last part focuses on BMI (body mass index) distribution and faith comparision between genders.

K-means clustering algorithm finds relevant grouping of people. BMI distribution has gaussian form and average BMI differs between genders. It’s shown that women tend to be more religious than men.

Motivation

The work shows many data analysis methods available in Python in practice.

First thing analysis aims is to find out about phobias troubling young people. One can ask if arachnofobia is common? What other kind of phobias are likely to occure if a person is arachnofobic?

Secondly, we are interested if there are connections between personality features and other features, like i.a. specific music taste and hobbies.

Thirtly, we are looking for answers on gender reference questions: who and by how much is more breave - women or men? Are men more likely to became atheists than women?

The goal of the analysis is to fulfill autor’s own curiosity and to share the results with the world.

Dataset(s)

The exact link to the dataset:
https://www.kaggle.com/miroslavsabo/young-people-survey

The data file consists of 1010 rows and 150 columns (139 integer and 11 categorical). Each row describes one person.

The variables can be split into 8 groups: Music preferences (19 items), Movie preferences (12 items), Hobbies & interests (32 items), Phobias (10 items), Health habits (3 items), Personality traits, views on life, & opinions (57 items), Spending habits (7 items), Demographics (10 items).

The last group contains i.a. following informations about a person: Age, Height, Weight, Number of siblings, Gender, Left/Right handing, Education level.

Data Preparation and Cleaning

Many people didn’t answer all questions. It wasn’t possible to drop all rows containing any missing values, because it would remove some useful data.

Dropping NaN values was being coducted just before analysing a specifed feature, by discluding rows containing NaN in this specified feature’s column.

One value of ‘Weight’ was mistaken by +100 kg, resulting in abnormous BMI value, what should have been repaired.

Research Question(s)

1. Correlation analysis: Are there any connections between features of various aspects? (i.e. between music and hobbies)

2. Hierarchical clustering: Can we introduce a distance measure between types of music, basing only on probability of liking music B if music A is liked?

3. K-means clustering: Can we divide people into groups based on their phobias?

4. Outlier detection: Can we identify participants that answered questions randomly?

5. Histograms: Can we identify differences between women and men considering their faith in God?

Methods

Hierarchical clustering uses Ward’s method and reasonably identifies correlations between all types of music.

K-means clustering algorithm is used to divide people considering their phobias. The optimal number of clusters was chosen as 4 using „Elbow method”.

Outliers were detected as people who’s Euklidean distance from nearest cluster (result of K-means algorithm) was higher than 4.0.

To visualise differs between women and men, histogram plots were used. It shows how many people of each gender answered specified value (from closed range) in the survey.

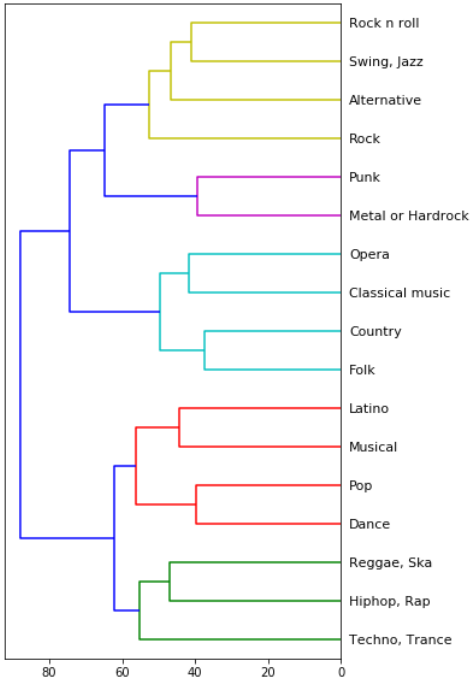
Findings

<Feel free to replicate this slide to show multiple findings>

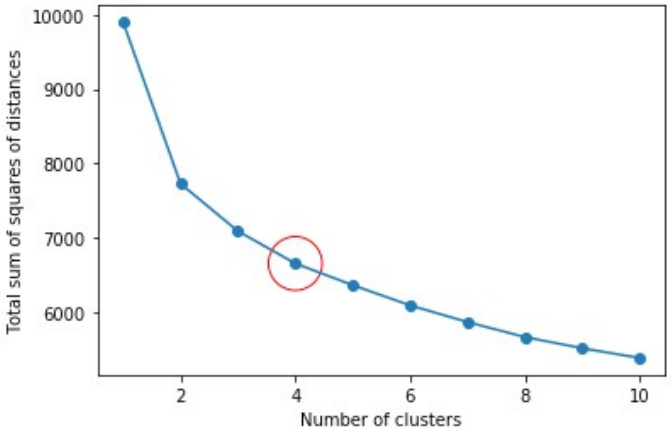
Present your findings. Include at least one visualization in your presentation (feel free to include more). The visualization should be honest, accessible, and elegant for a general audience.

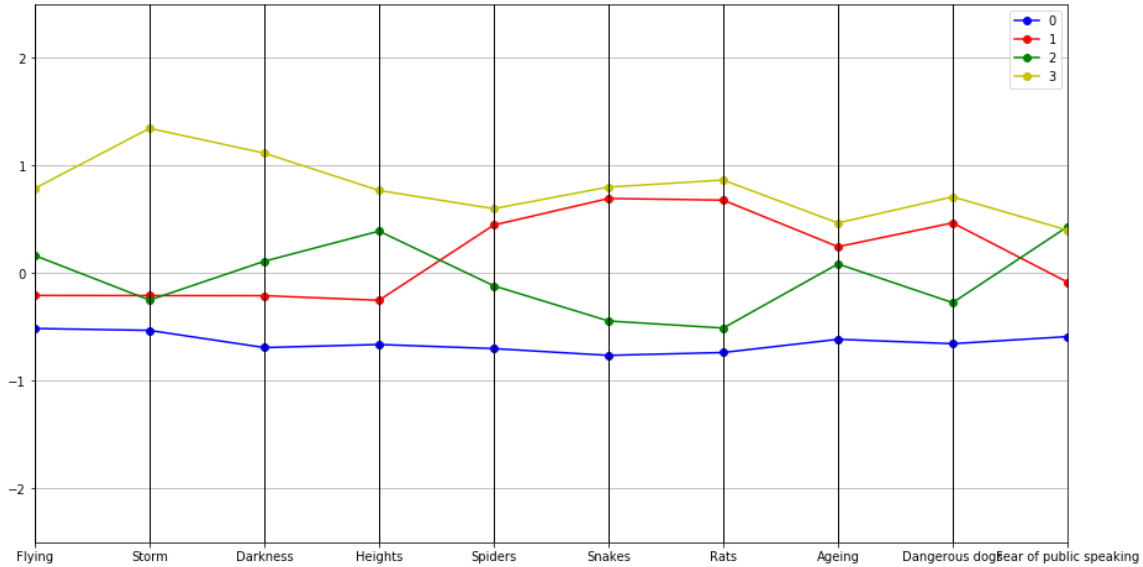
You need not come to a definitive conclusion, but you need to say how your findings relate back to your research question.

Hierarchical clustering



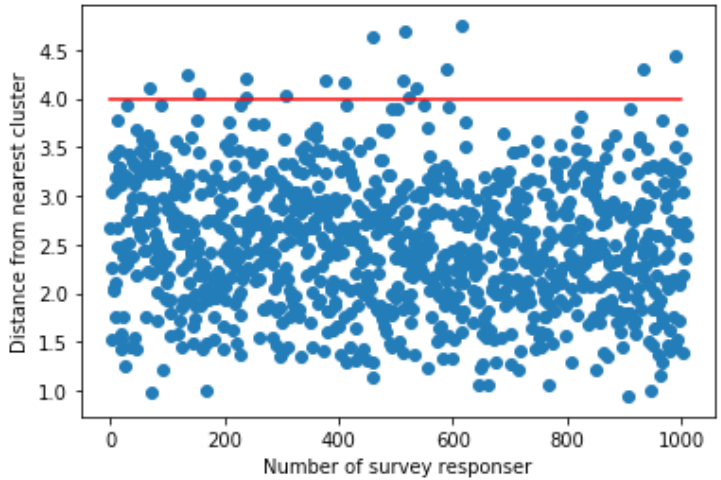
Findings


Total sum of squares of distances
10000
9000
8000
7000
6000
2
4
6
8
10
Number of clusters

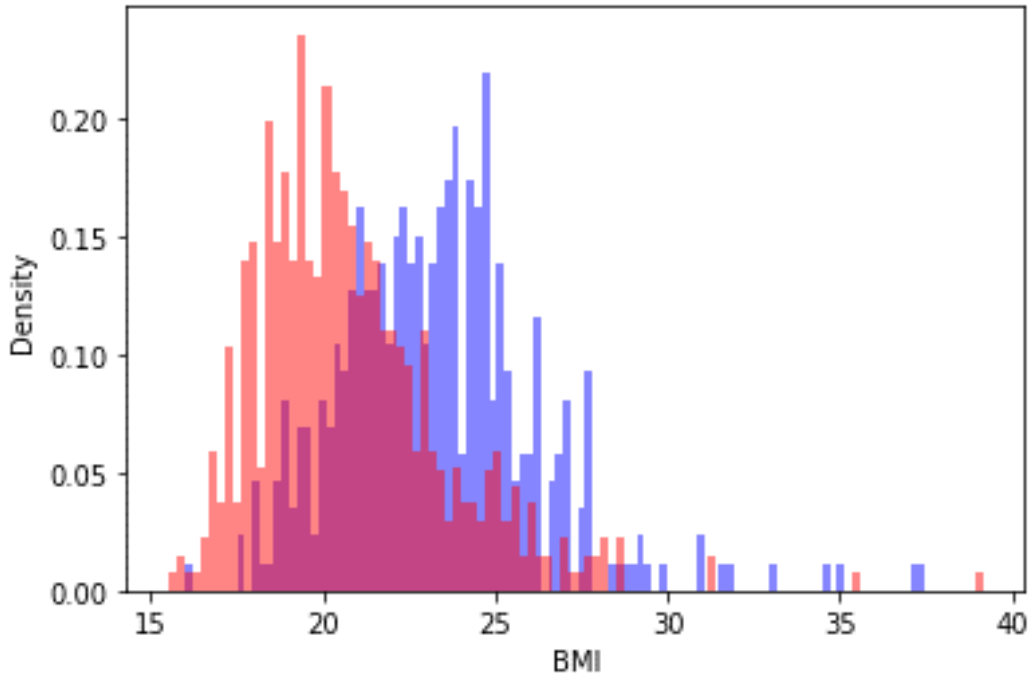

0
1
2
3
2
1
0
−1
−2
Flying
Storm
Darkness
Heights
Spiders
Snakes
Rats
Ageing
Dangerous dogs
Fear of public speaking

Outliers

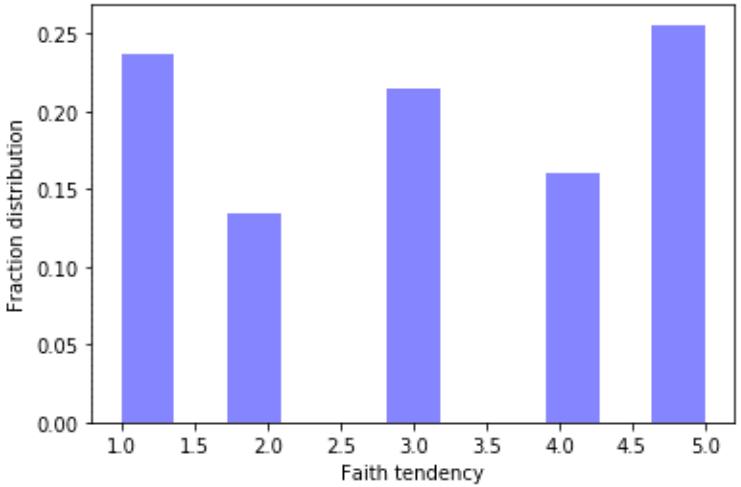
	Music	Slow songs or fast songs	Dance	Folk	Country	Classical music	Musical	Pop	Rock	Metal or Hardrock	...	Age	Height	Weight	Number of siblings	Gender	Left - right handed	Education
69	5.0	4.0	1.0	1.0	1.0	1.0	4.0	5.0	4.0	1.0	...	20.0	163.0	47.0	3.0	female	left handed	secondary school
135	5.0	3.0	2.0	1.0	3.0	5.0	1.0	3.0	5.0	4.0	...	18.0	185.0	62.0	1.0	male	right handed	secondary school
155	5.0	4.0	4.0	3.0	3.0	5.0	2.0	3.0	3.0	2.0	...	19.0	171.0	72.0	1.0	female	right handed	secondary school
236	1.0	3.0	3.0	3.0	2.0	3.0	5.0	4.0	4.0	3.0	...	19.0	171.0	57.0	0.0	female	right handed	secondary school
237	5.0	4.0	4.0	2.0	1.0	3.0	3.0	4.0	3.0	1.0	...	18.0	173.0	64.0	3.0	female	right handed	secondary school
307	5.0	3.0	1.0	4.0	3.0	2.0	5.0	2.0	5.0	4.0	...	19.0	176.0	68.0	2.0	female	right handed	secondary school
378	5.0	5.0	5.0	NaN	2.0	2.0	4.0	5.0	4.0	1.0	...	20.0	175.0	53.0	2.0	female	right handed	secondary school
409	5.0	4.0	5.0	1.0	5.0	1.0	5.0	4.0	4.0	1.0	...	16.0	165.0	50.0	0.0	female	right handed	currently a primary school pupil
458	5.0	NaN	5.0	5.0	5.0	5.0	5.0	5.0	1.0	1.0	...	18.0	168.0	52.0	1.0	female	right handed	secondary school
512	5.0	1.0	2.0	4.0	1.0	2.0	1.0	1.0	1.0	1.0	...	22.0	168.0	58.0	1.0	female	right handed	college/bachelor degree
516	5.0	5.0	5.0	3.0	1.0	2.0	3.0	5.0	4.0	4.0	...	22.0	173.0	80.0	2.0	male	right handed	secondary school
521	5.0	3.0	5.0	5.0	1.0	5.0	4.0	5.0	3.0	2.0	...	20.0	157.0	57.0	0.0	female	right handed	secondary school
536	5.0	5.0	2.0	1.0	1.0	1.0	2.0	4.0	2.0	2.0	...	19.0	168.0	57.0	1.0	female	right handed	college/bachelor degree
589	5.0	4.0	1.0	1.0	1.0	3.0	4.0	1.0	5.0	5.0	...	17.0	173.0	66.0	1.0	female	right handed	primary school
614	5.0	3.0	4.0	1.0	1.0	3.0	2.0	4.0	4.0	1.0	...	22.0	164.0	64.0	1.0	female	right handed	secondary school
932	5.0	3.0	4.0	2.0	2.0	1.0	3.0	5.0	3.0	2.0	...	20.0	155.0	45.0	1.0	female	right handed	secondary school
989	5.0	5.0	3.0	2.0	2.0	4.0	5.0	5.0	5.0	5.0	...	30.0	168.0	54.0	2.0	female	right handed	masters degree

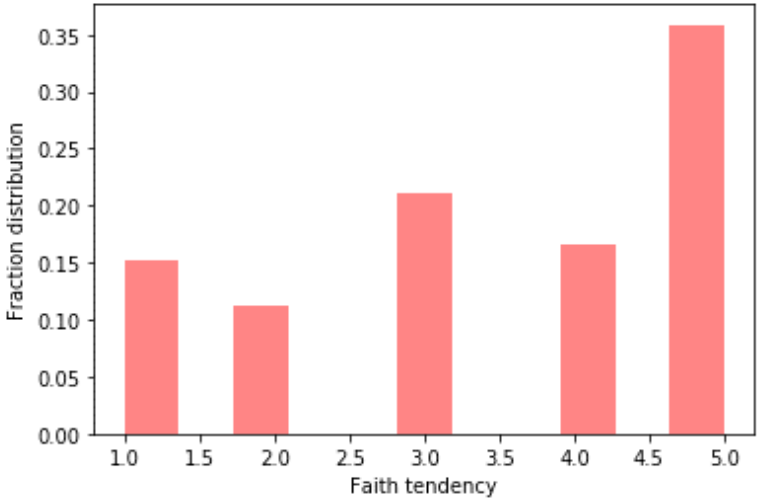


BMI distribution



Faith in God comparision





Limitations

The dataset consist of informations from one country in Europe and all people surveyed are aged 15-30 years. Though the data is gathered from various places in Slovakia, there is a big chance that features like BMI and faith will vary if compared with another country. Phobias analysis is likely to agree amoung well rich countries. Old people are likely of be outliers aswell.

Conclusions

There seem to be tree kinds of correlated fobias: 1-animals (Rats, Snakes, Dangerous dogs, Spiders), 2-nature (Darkness, Storm, Flying), 3-social (Fear of public speaking, Heights, Ageing).

People seem to divide in 4 different types considering phobias: 1-scared of nothing, 2-scared of various animals, 3-scared of everything except of animals, 4-scared of everything.

Outliners are mostly woman, who enjoy pop music.

In Slovakia women’s average BMI is 20, where men’s is 24.

Woman are more likely to be religious. 36% of woman and 25% of men are strongly religious, where 15% of woman and 24% men are atheists.

References

Two custom functions for convenient ploting of clustering analysis algorithm results were borrowed from EDX course „Python for Data Science” prepared by University of San Diego in California.