Young People Today

A Data Driven Portrait of the Next Generation

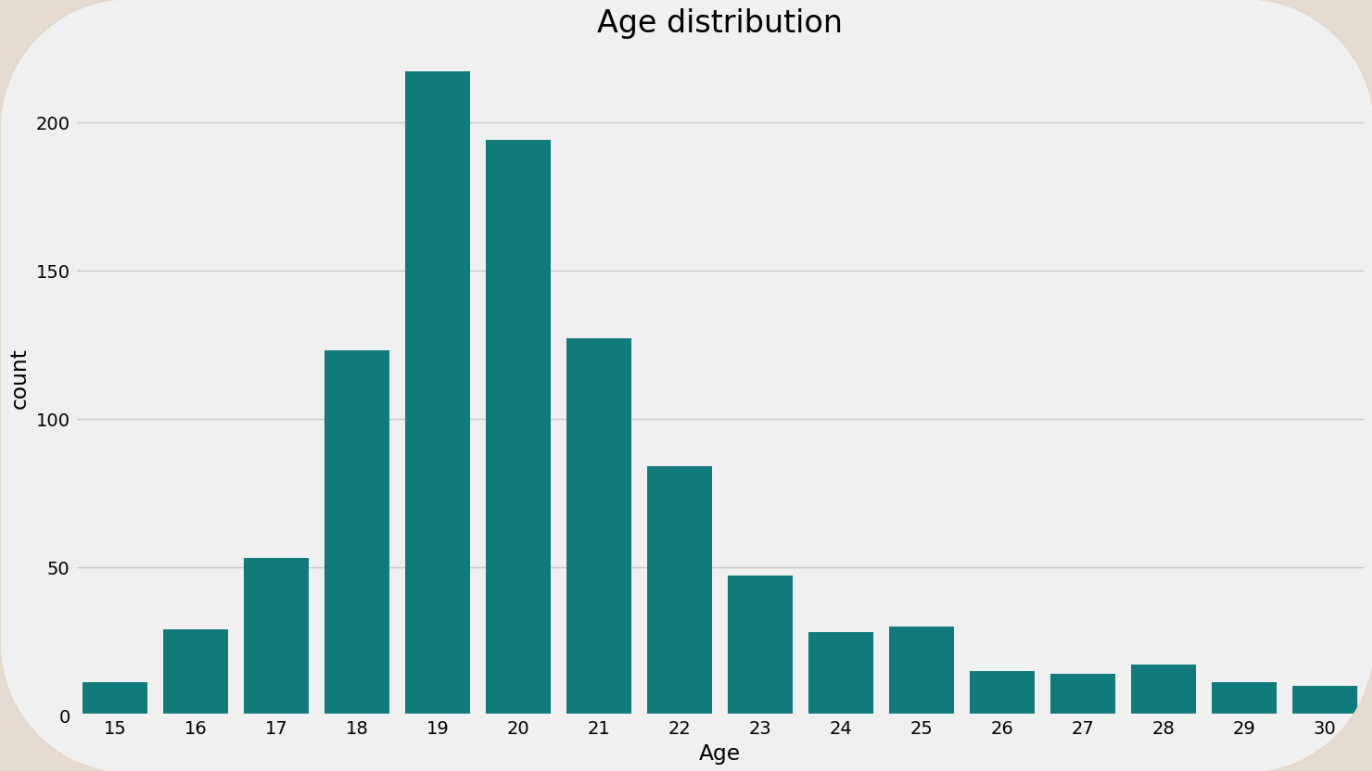
Burouj Armgaan

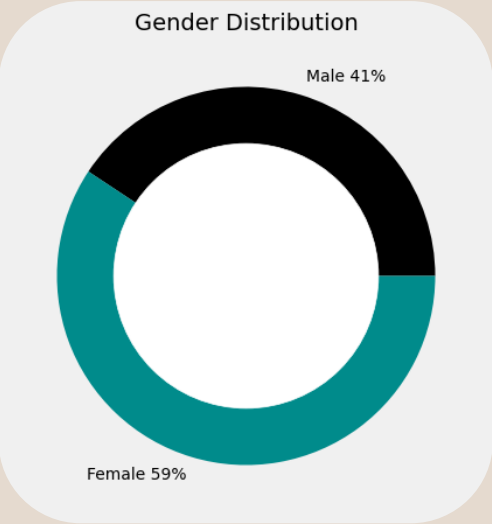
Young People Survey

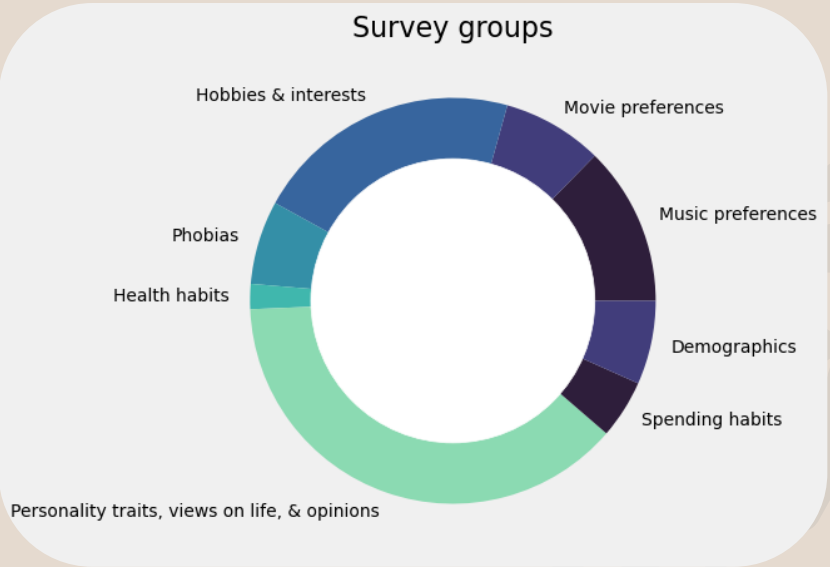
- In 2013, students of the Statistics class at FSEV UK were asked to invite their friends to participate in this survey.
- All participants were Slovakians aged 15-30.
- Dataset available at Kaggle.com

Data

1010 rows, 150 columns







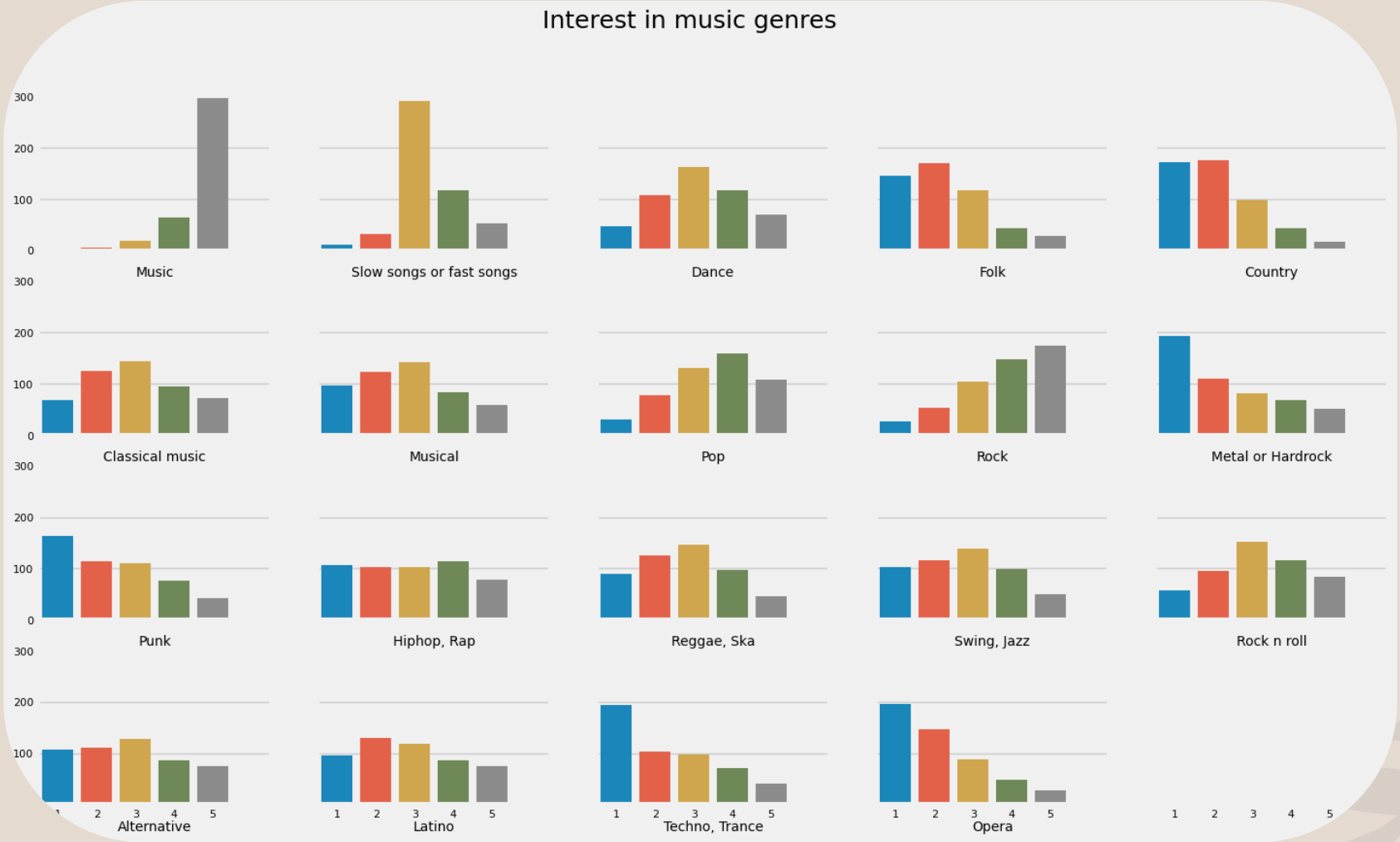
Research Questions

- Clustering
  - Given the music preferences, do people make up any clusters of similar behavior?
- Hypothesis Testing
  - Do women fear certain phenomena significantly more than men?
  - Do left-handed people have different interests than right-handed?
- Childhood's Impact
  - Analyze if there are differences between the participants based on where they lived most of their childhood: rural and urban area.
- Missing Data Analysis
  - Is there are trend between the questions students abstain from answering?

Clustering

Given the music preferences, do people make up any clusters of similar behavior?

Genres

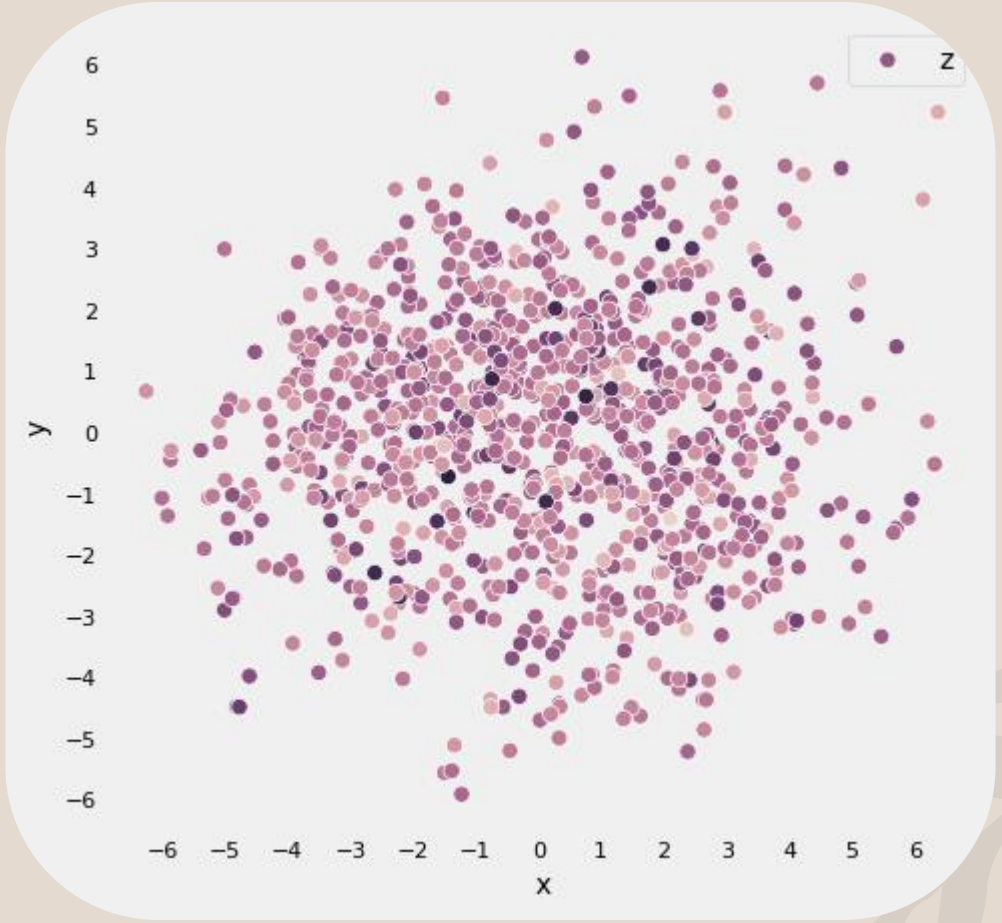

Interest in music genres
Music
Slow songs or fast songs
Dance
Folk
Country
Classical music
Musical
Pop
Rock
Metal or Hardrock
Punk
Hiphop, Rap
Reggae, Ska
Swing, Jazz
Rock n roll
Alternative
Latino
Techno, Trance
Opera
0
100
200
300
1
2
3
4
5

Insights

- Everyone likes music!
- Most genres show a gaussian behavior.
- Some genres are visibly popular: Pop, Rock.
- Some genres are visibly unpopular e.g., Folk, Country, Metal, Opera. Notably, these genres were quite popular in the previous generation.

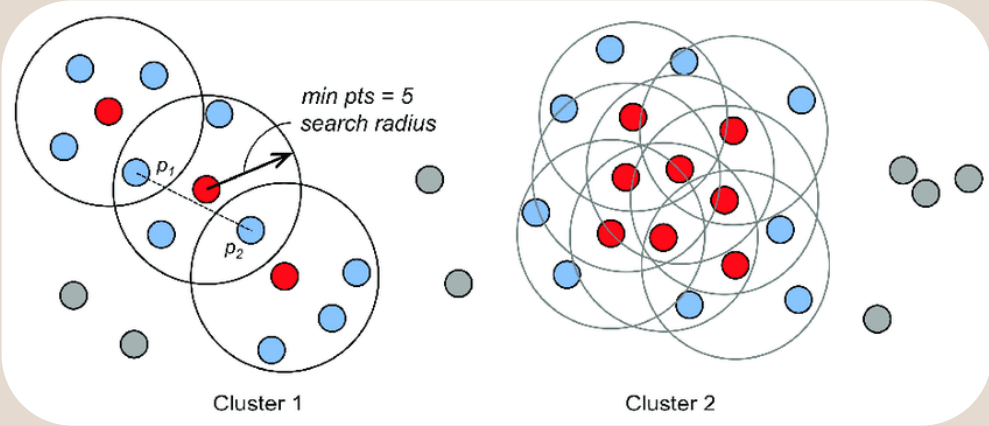
PCA

- No signs of clusters in reduced 3-dimensional data.



DBSCAN

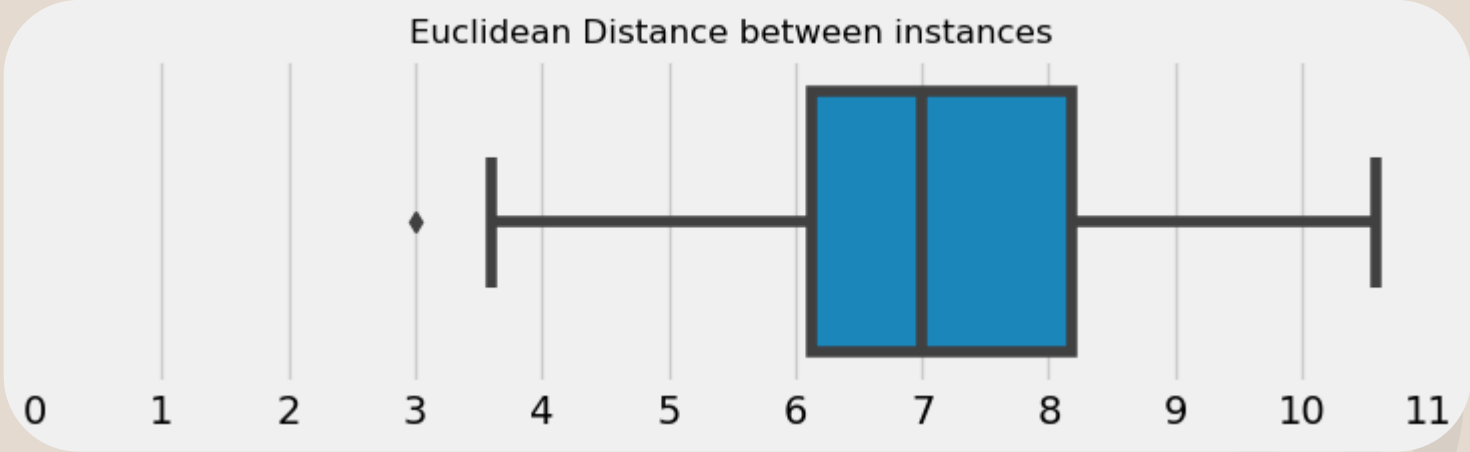
- DBSCAN is a density-based clustering algorithm.
- It has several merits over K-means.
- Only two parameters: Epsilon, Min-samples
- Higher Min-samples or lower Epsilon indicate higher density necessary to form a cluster.



D. Deng, "DBSCAN Clustering Algorithm Based on Density," 2020 7th International Forum on Electrical Engineering and Automation (IFEEA), Hefei, China, 2020, pp. 949-953, doi: 10.1109/IFEEA51475.2020.00199.

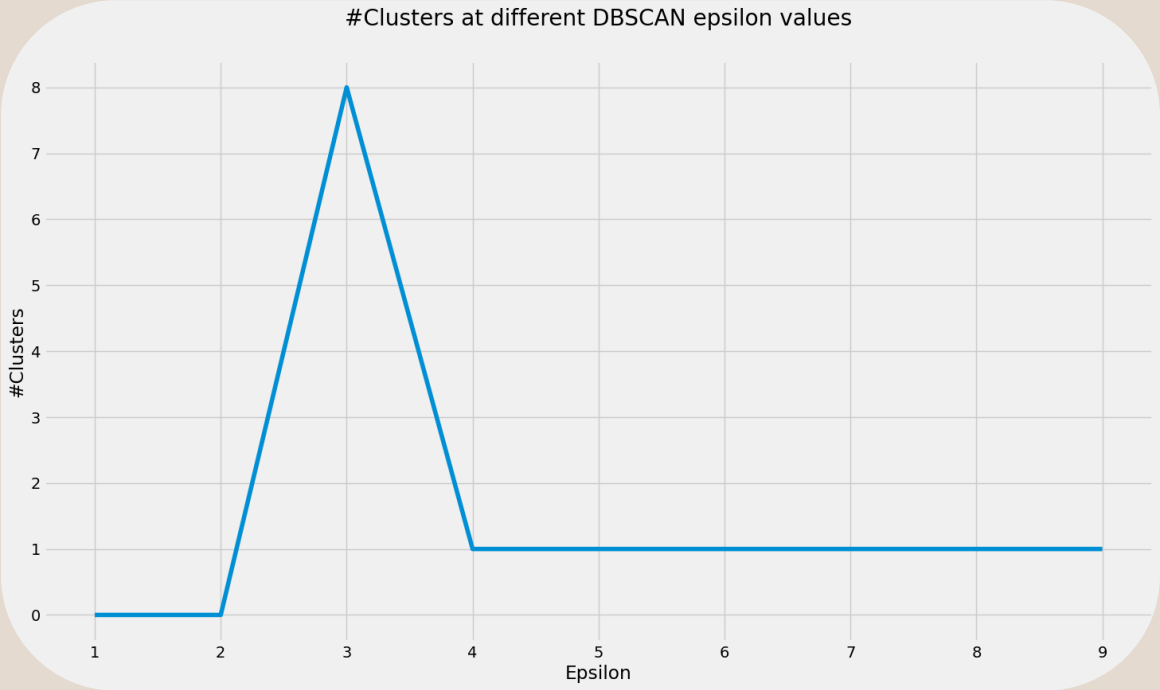
Hyperparameters

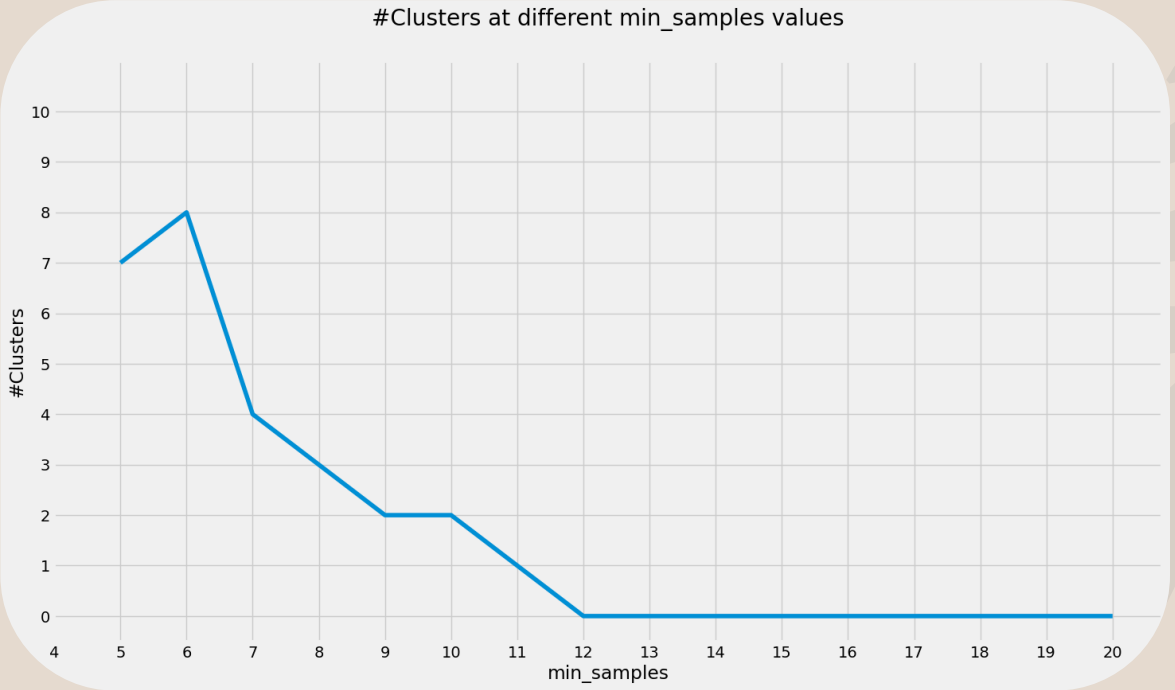
Get an idea of the distances between data points to make an educated guess about the hyperparameter search space.



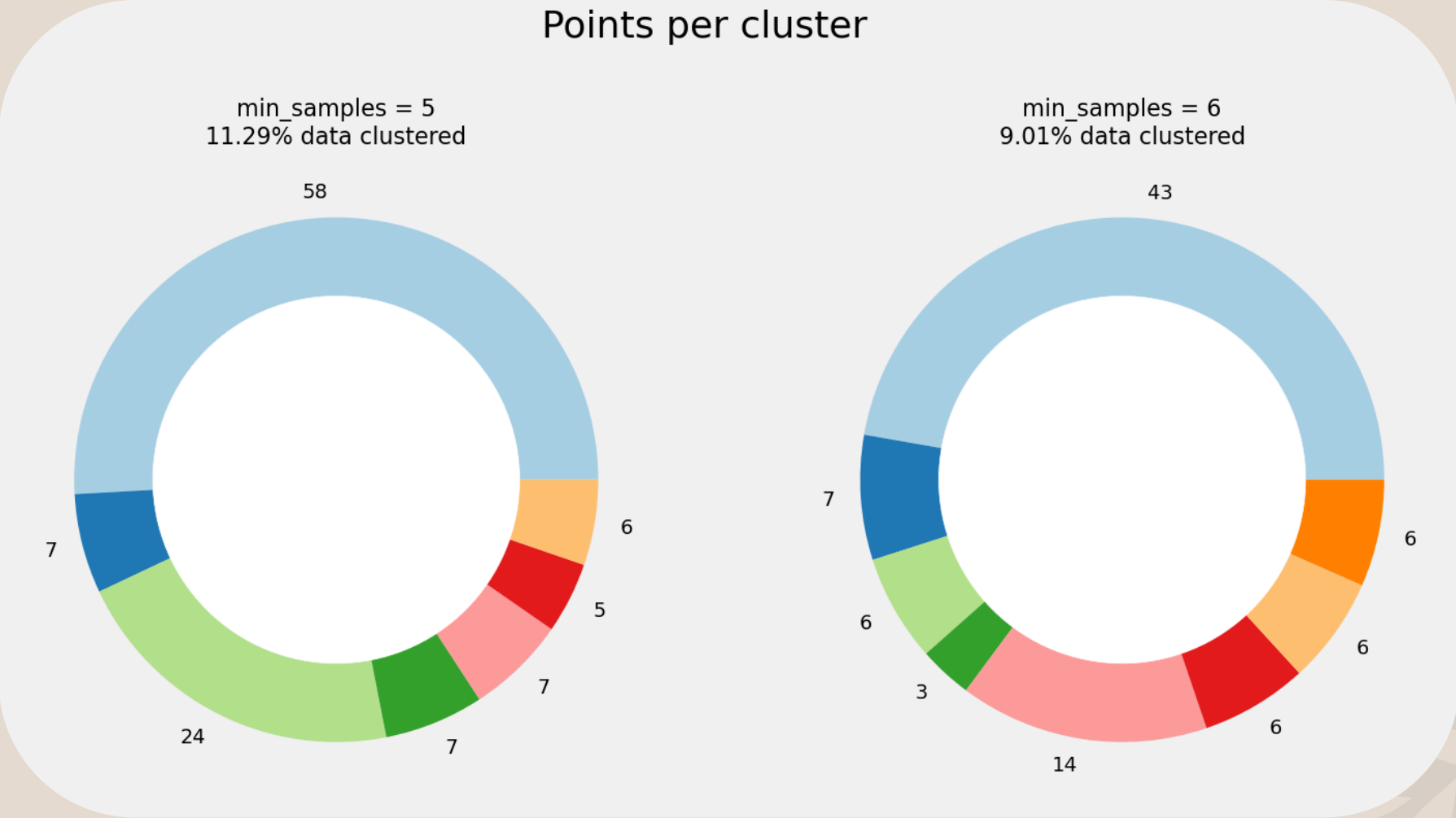
Hyperparameters

The best clusters are found at epsilon = 3 and Min-samples = {4, 5}





Clusters

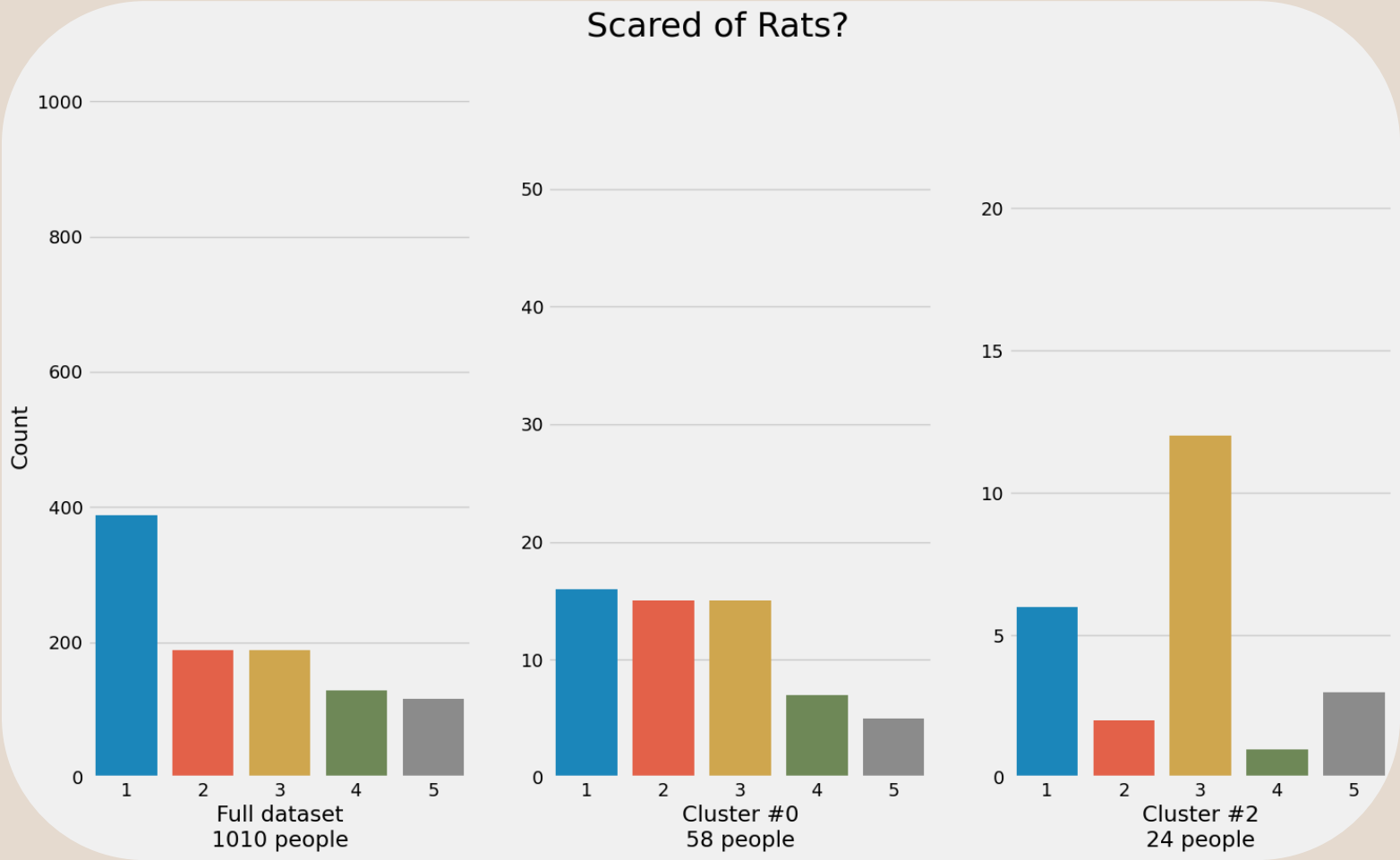

Points per cluster
min_samples = 5
11.29% data clustered
58
6
5
7
7
24
7
min_samples = 6
9.01% data clustered
43
6
6
6
14
3
6
7

What are we looking for?

Given the music preferences, do people make up any clusters of similar behavior?

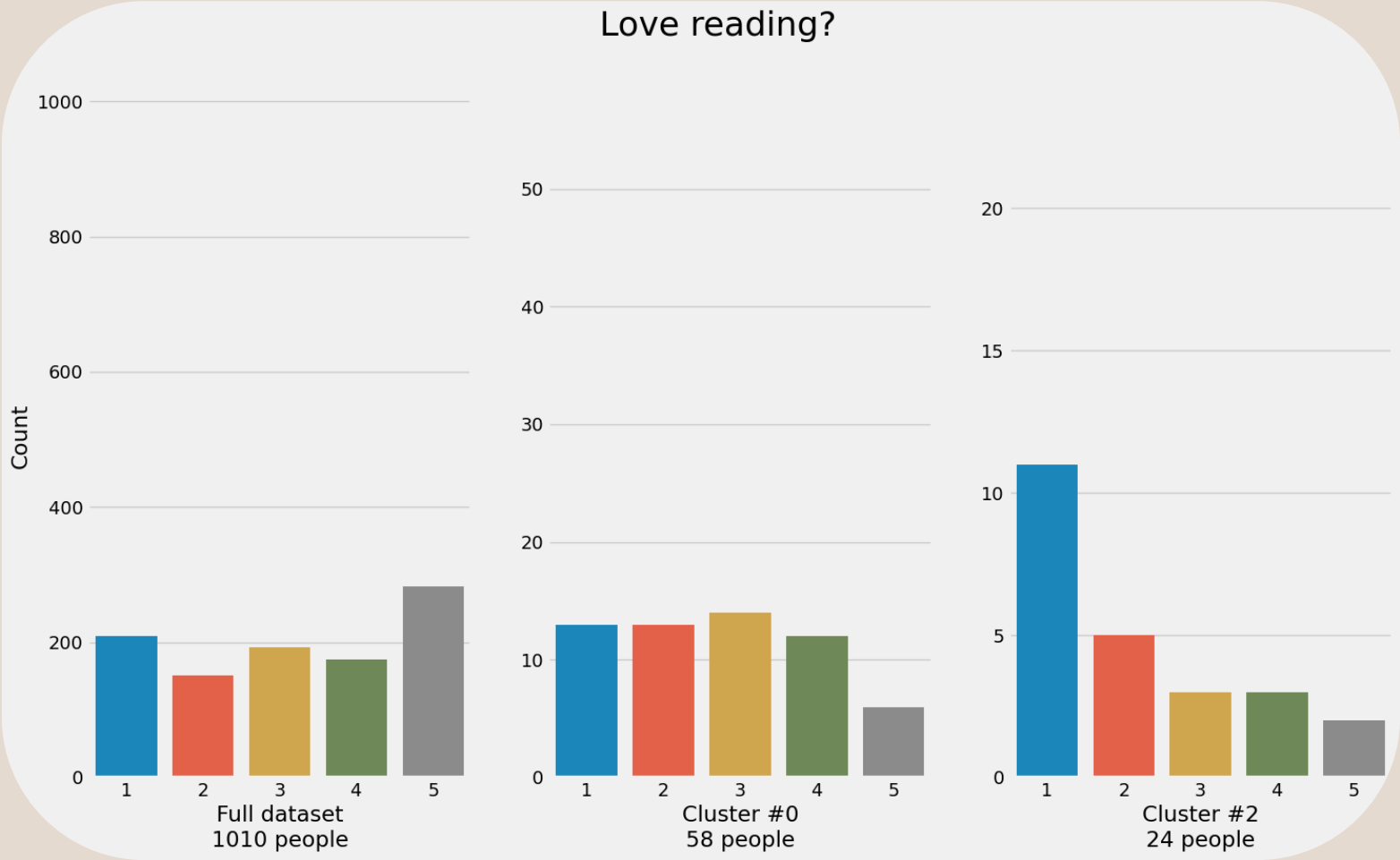
Scared of rats!

More than half of cluster #2 fears rats!



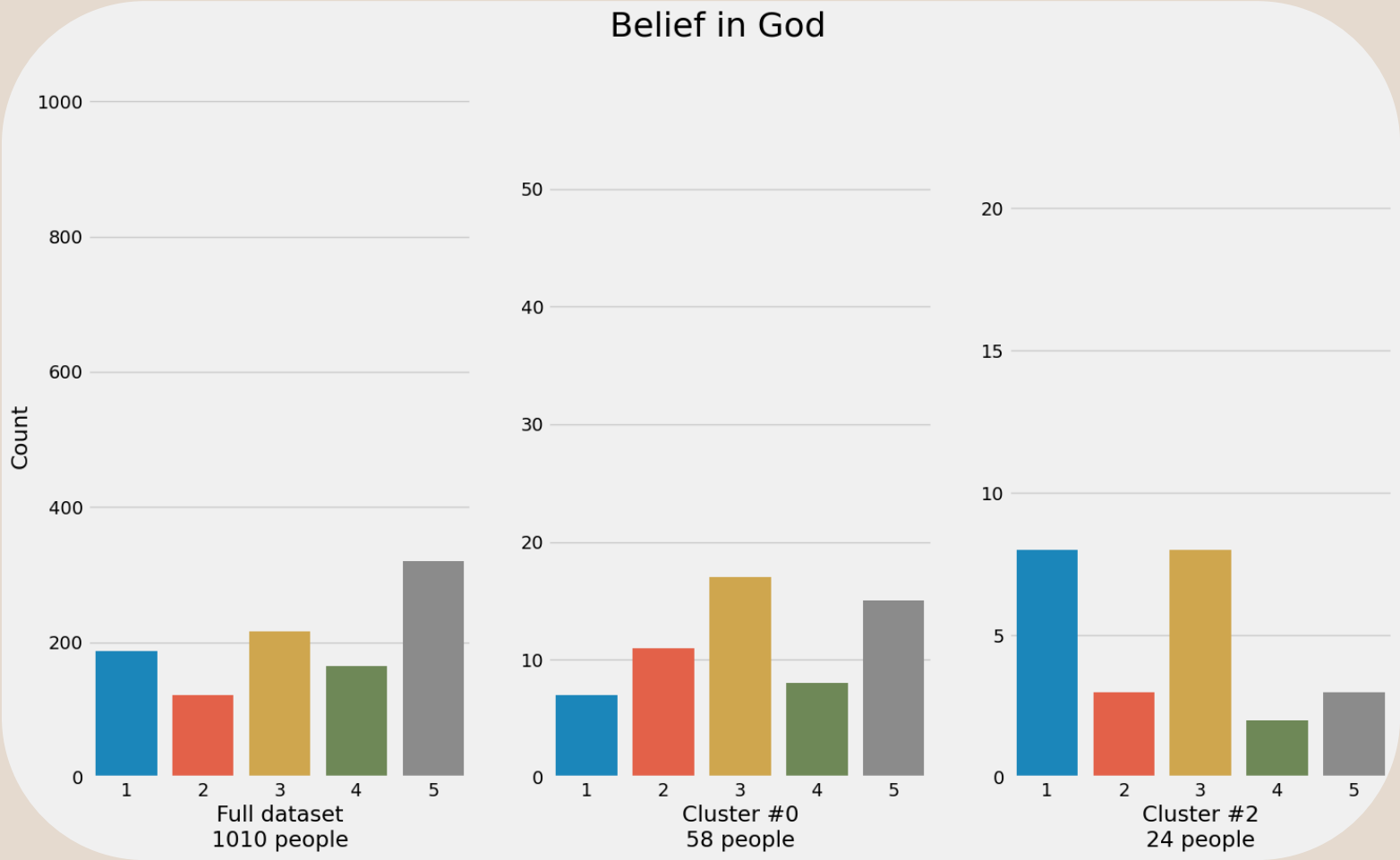
Hate Reading

If you're in cluster #2, you hate reading.



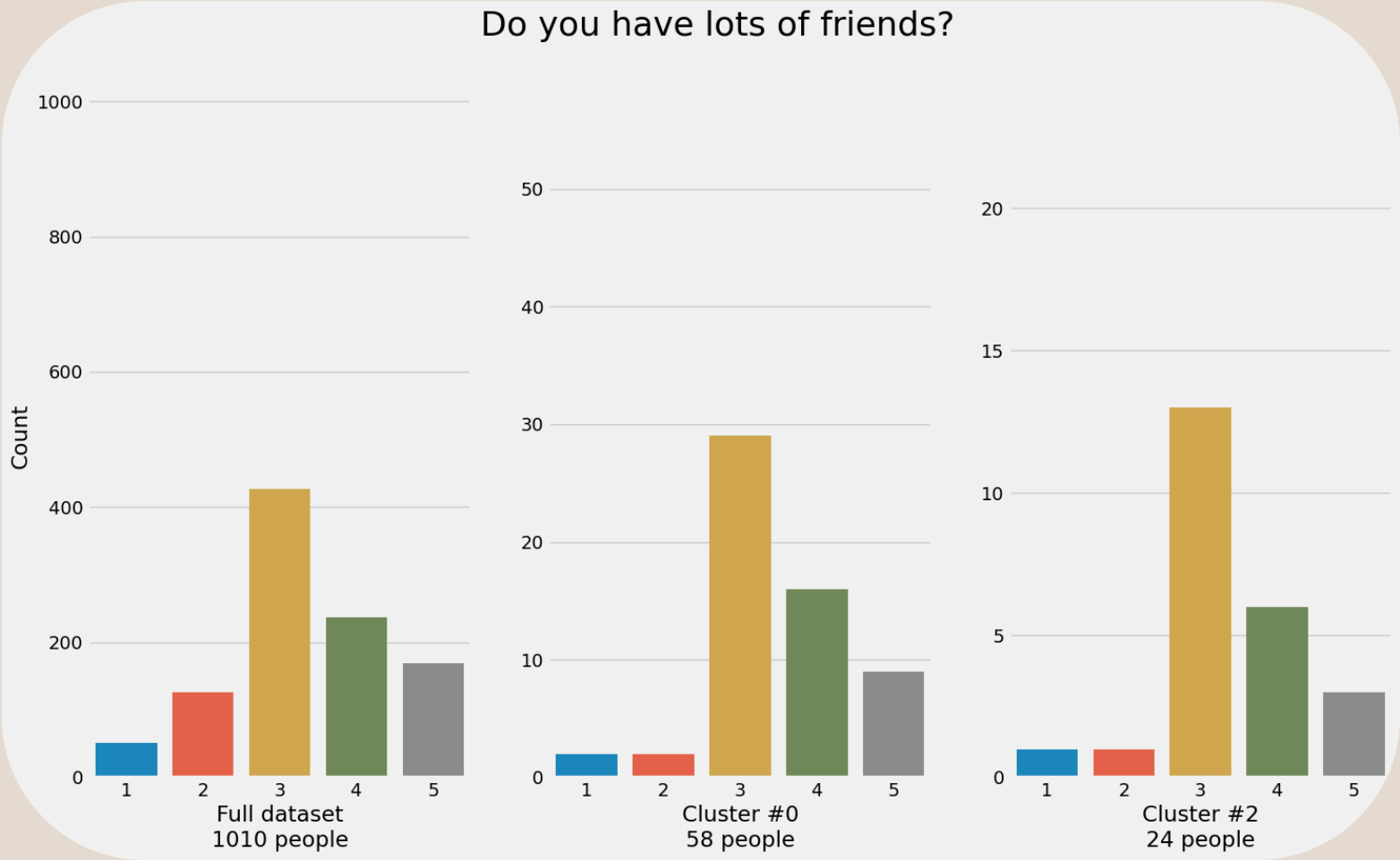
Atheistic

The young population, in general, has mixed religious beliefs. But cluster #2 is relatively atheistic.



Lots of friends?

We see similar behavior across the clusters.



Insights

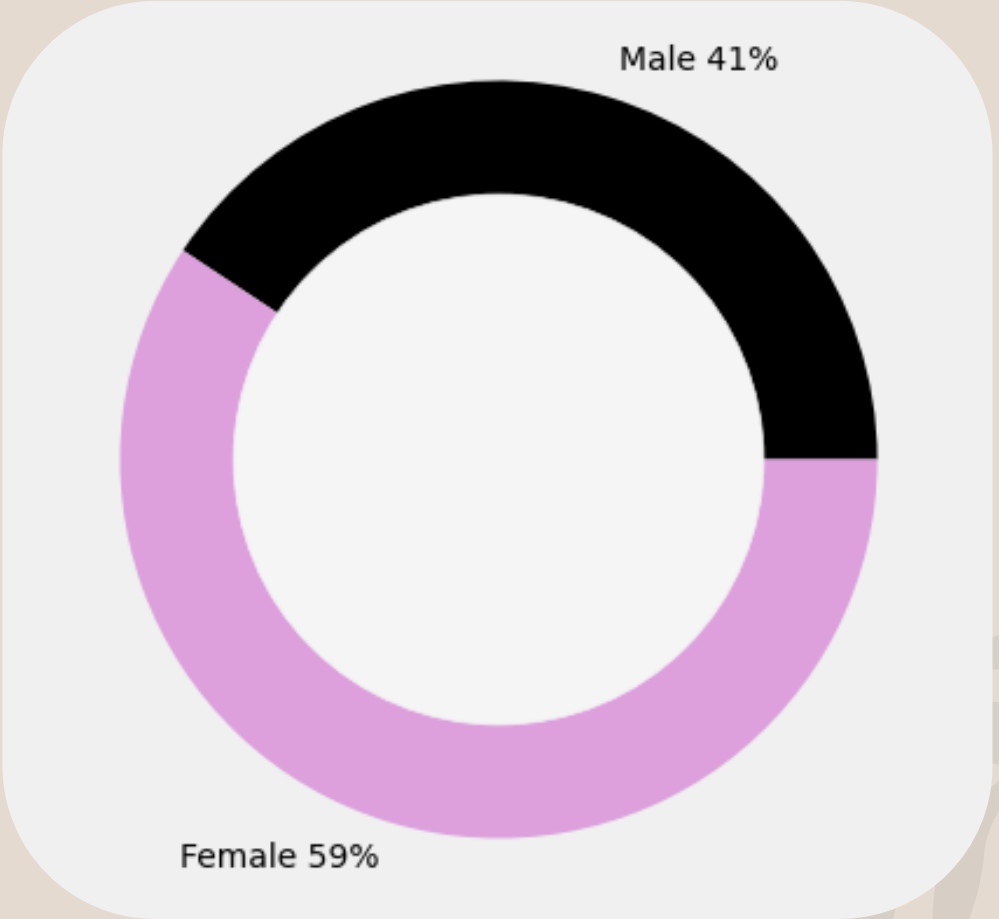
- Some behaviors are surprisingly linked, like music preference and fear of rats.
- While others, like how many friends one has, isn't related to a person's music preference.
- However, we must keep in mind that almost 90% of the population remained unclustered. This implies that people have varied music preferences, and only for a select few people can we derive their behavior from their musical taste.

Hypothesis Testing

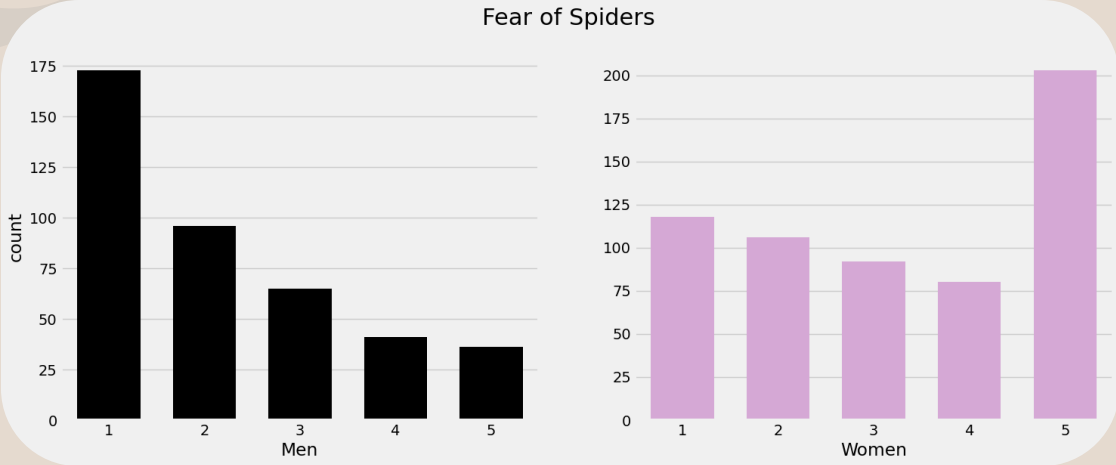
Do women fear certain phenomena significantly more than men?

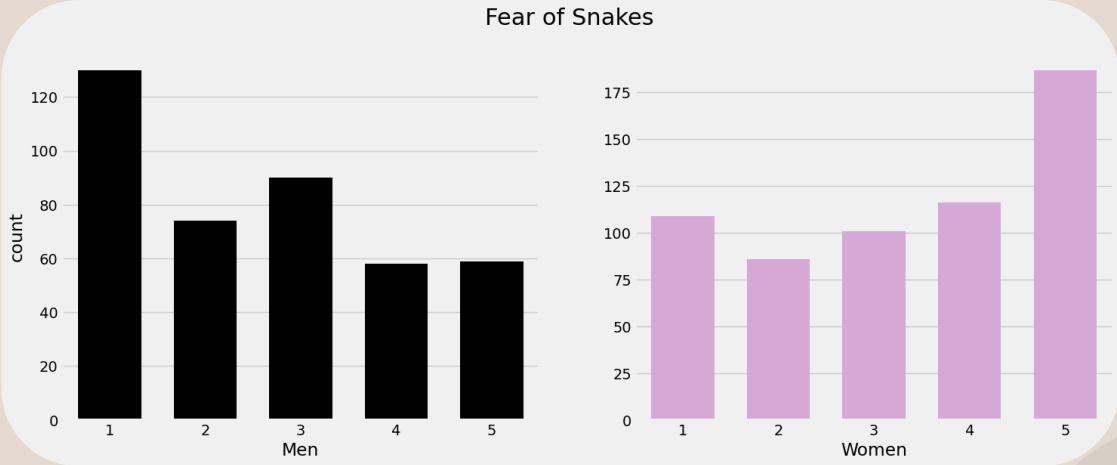
Male to Female Ratio

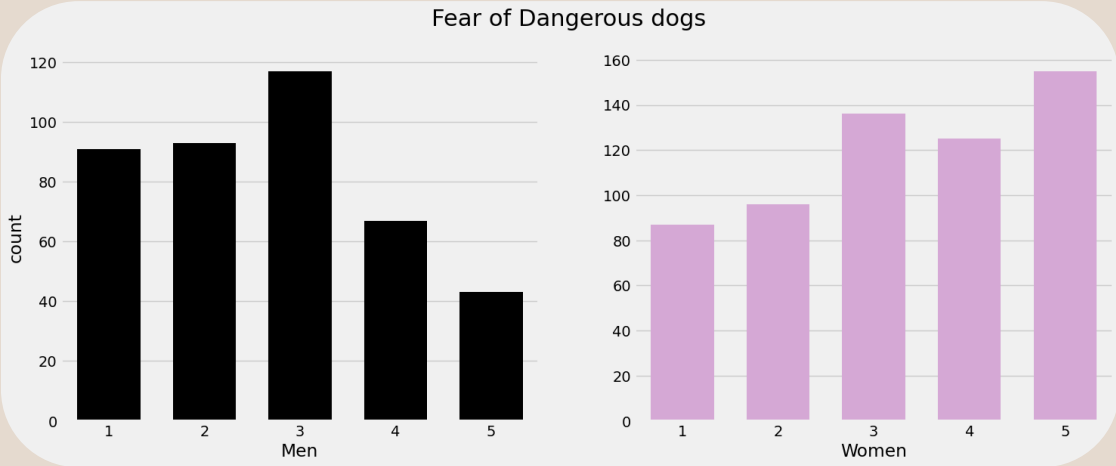
There are more women than men, but the split is close enough for us to use a shared y-axis while plotting the distributions.

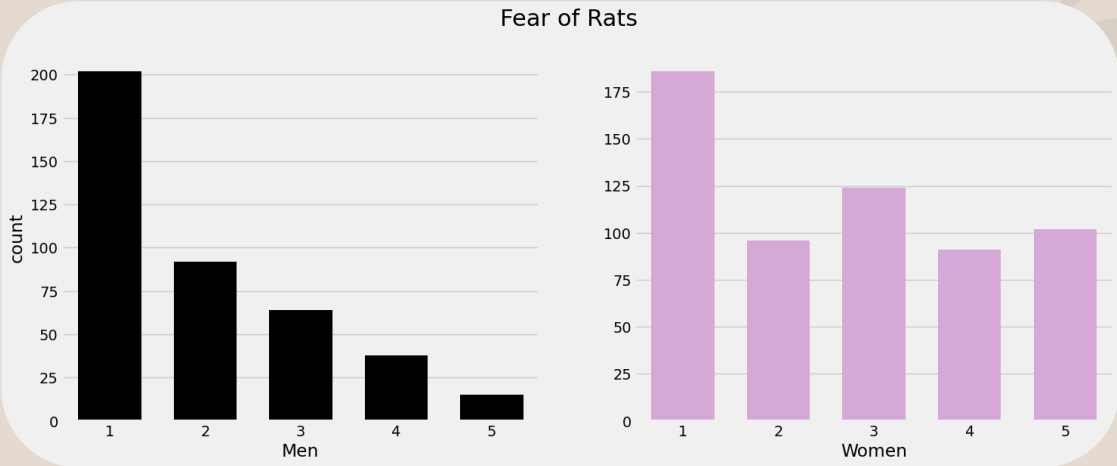


Spiders, Snakes, Rats & Dogs!

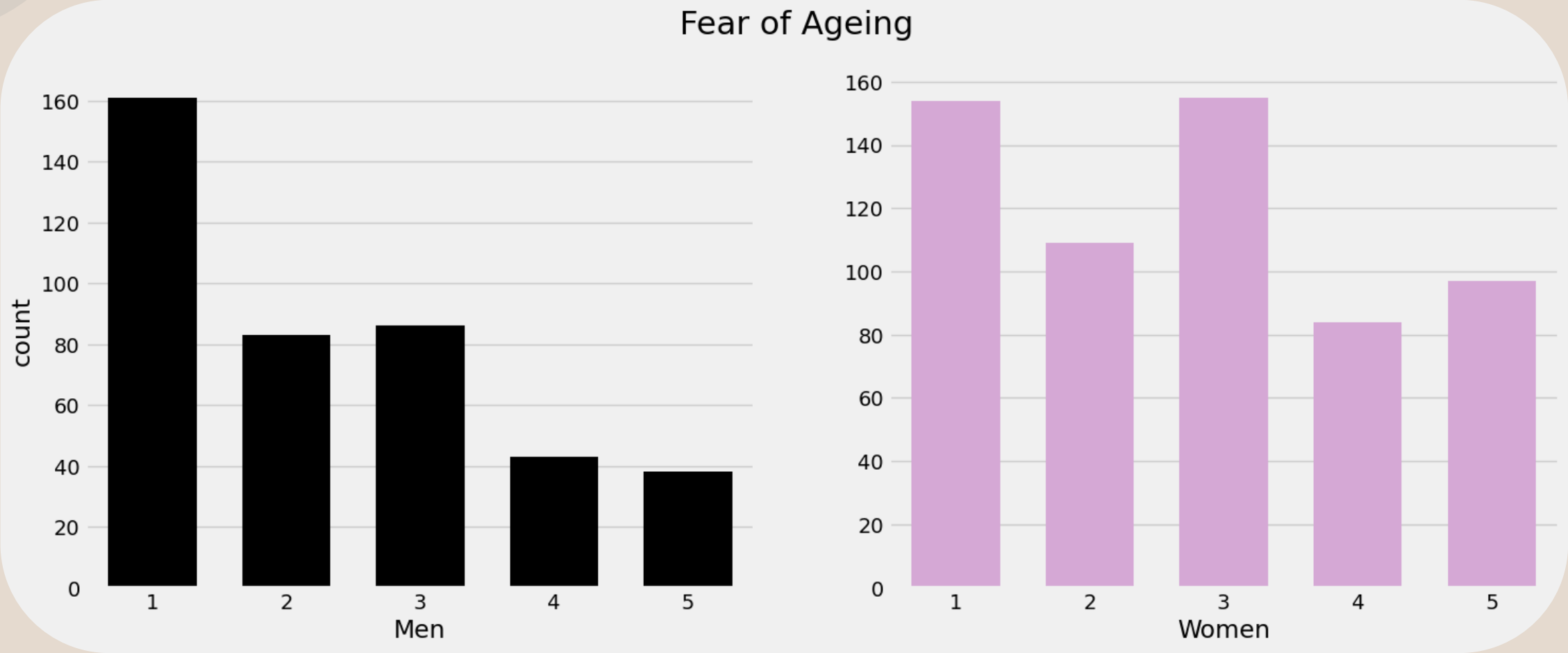




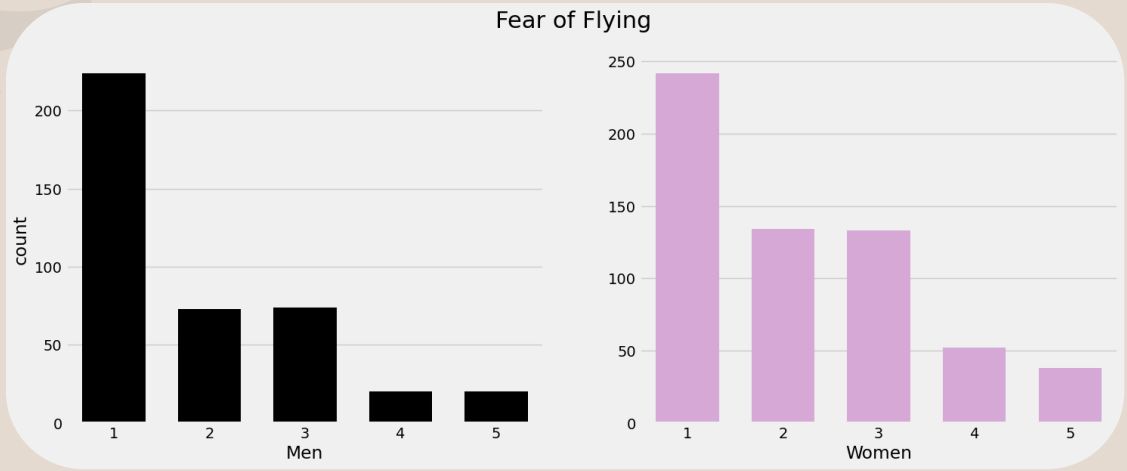


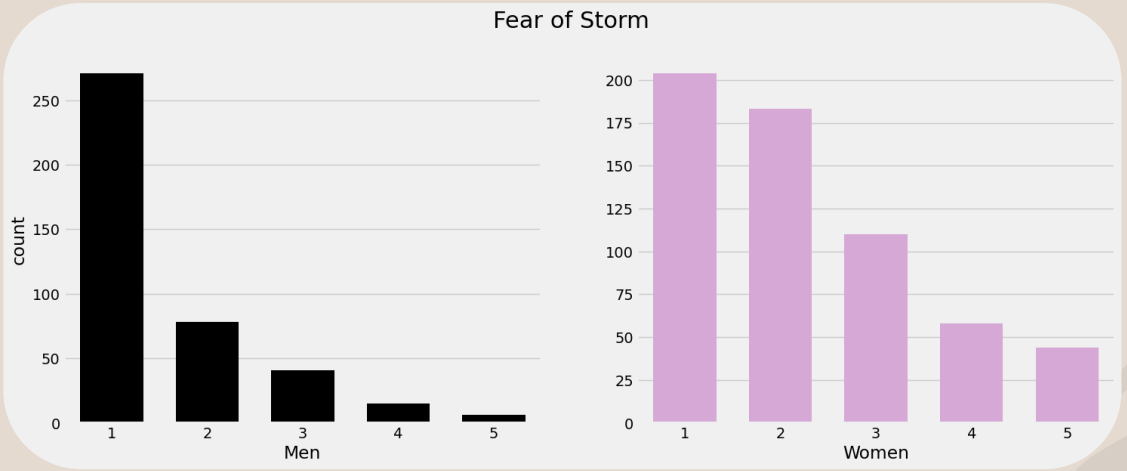


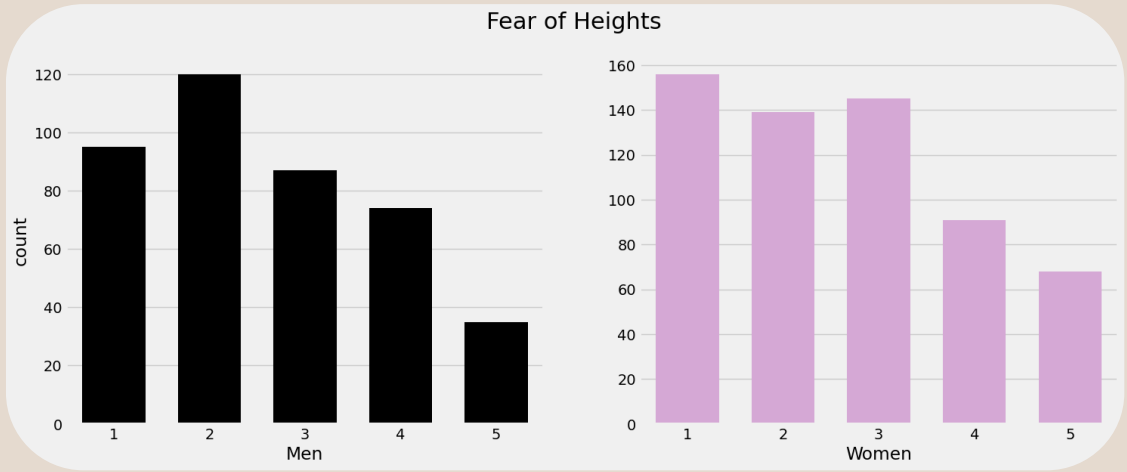
Ageing, of course

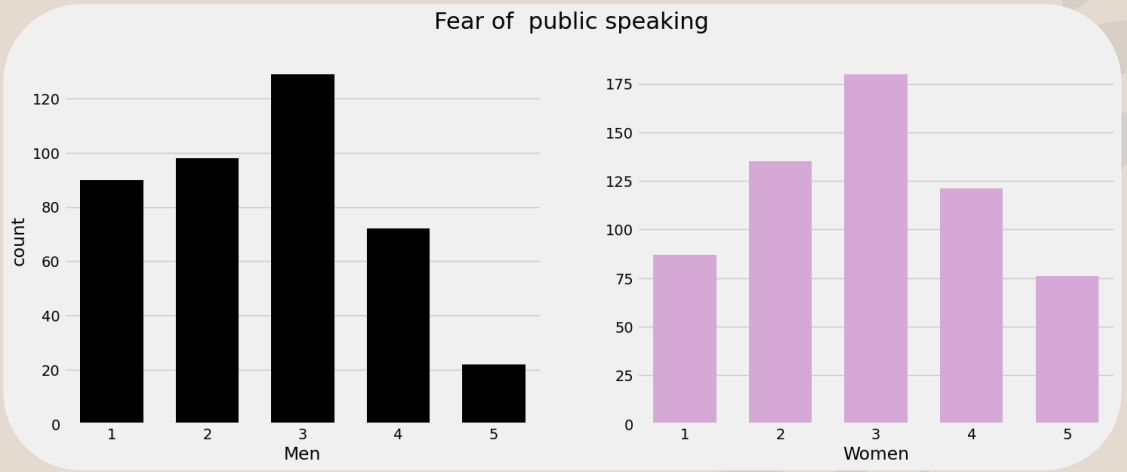


For the rest, both genders show similar response







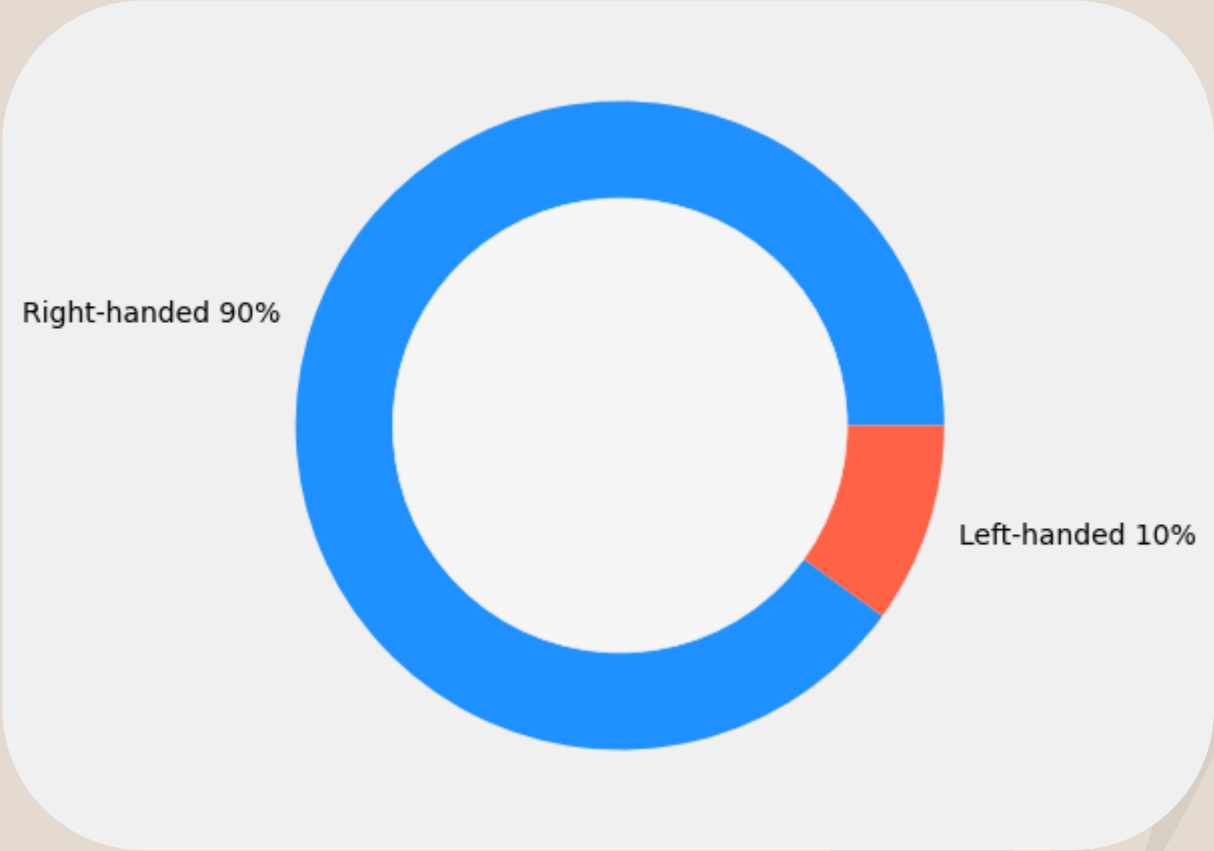


Insights

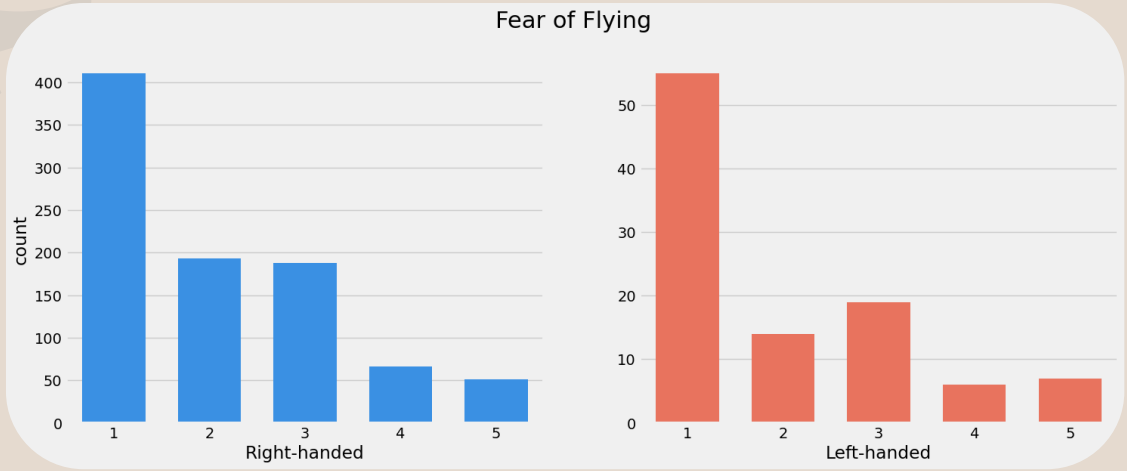
- 10 phobias listed in the survey.
- In all cases where the distributions diverge, women display a more pronounced fear response than men.
- Women generally demonstrate a stronger fear response to certain stimuli, such as spiders, snakes, rats, dogs, and ageing, than men.
- The prevalence and intensity of other phobias do not seem to differ significantly between genders.

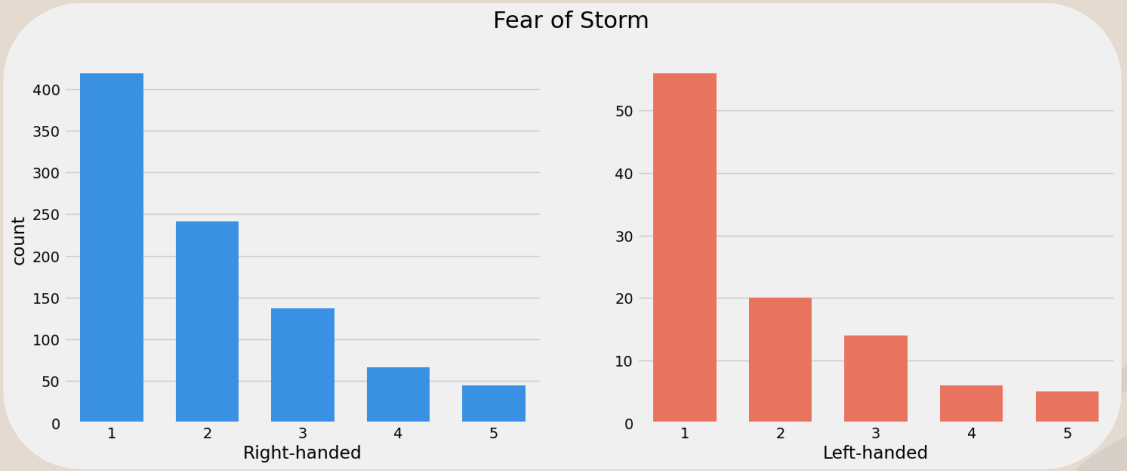
Right – Left Handedness Ratio

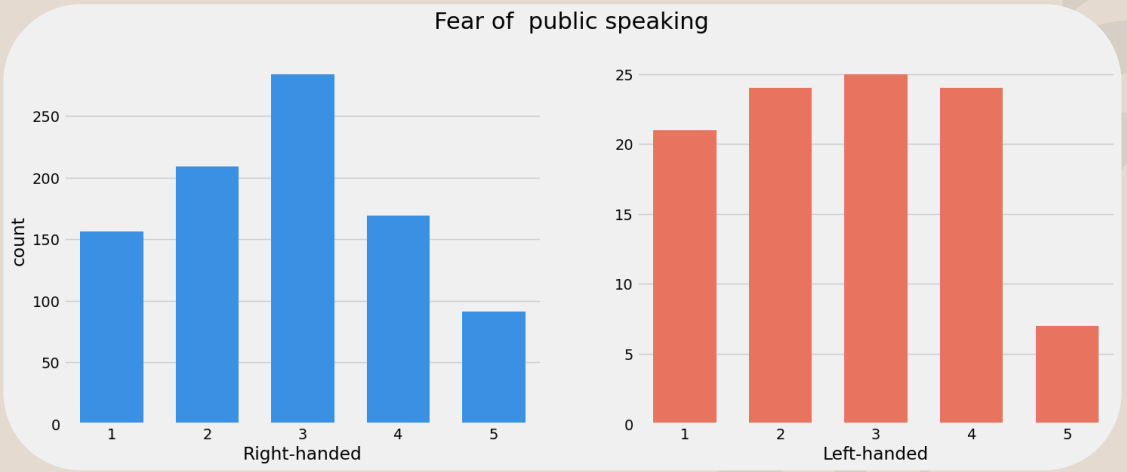
- A similar study for right-left handedness.
- There are overwhelmingly more right-handed people than left-handed people. Hence, we should use separate y-axes for the two.



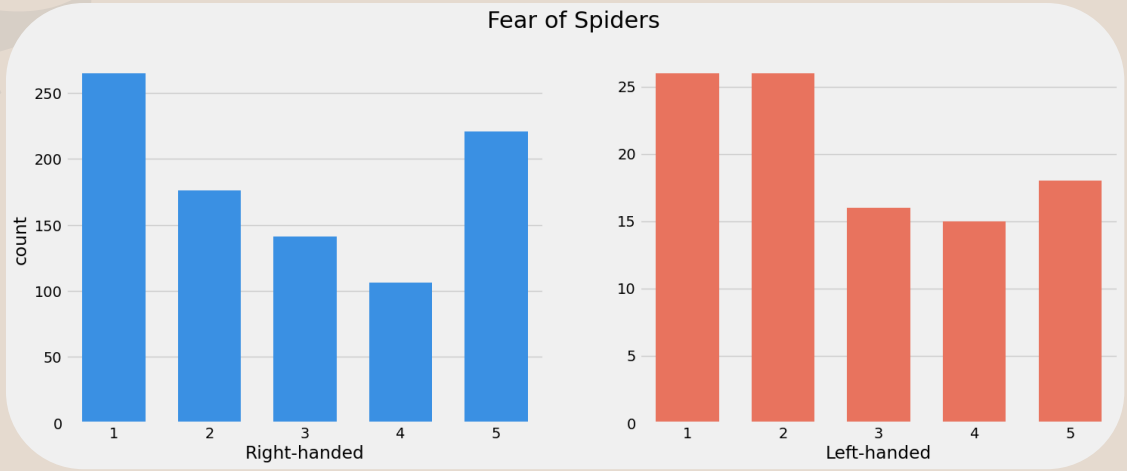
No variation: Flying, Storm, Heights, P-Speaking

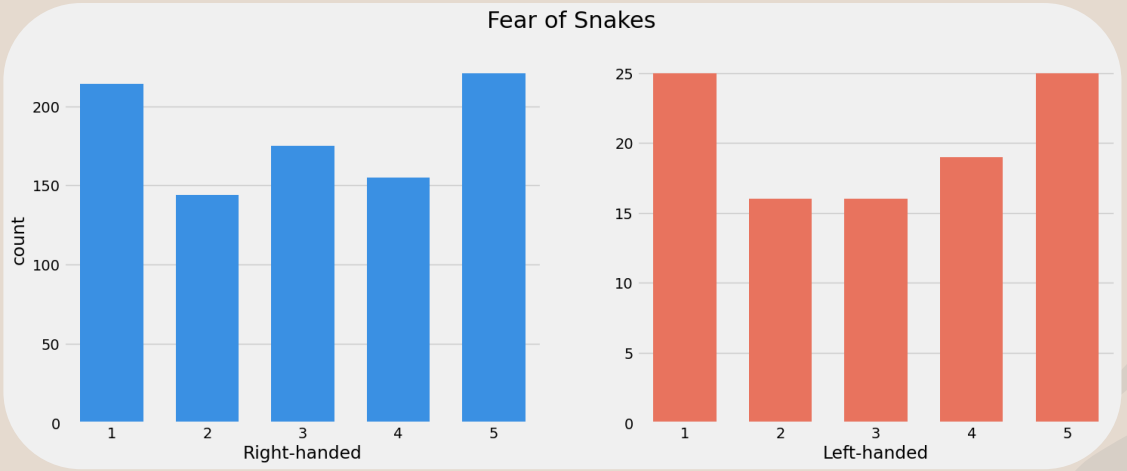


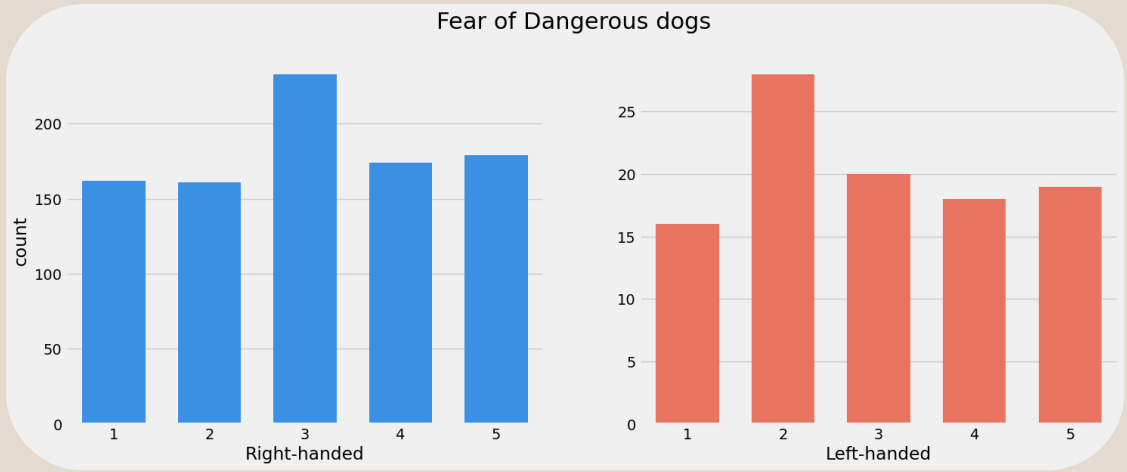


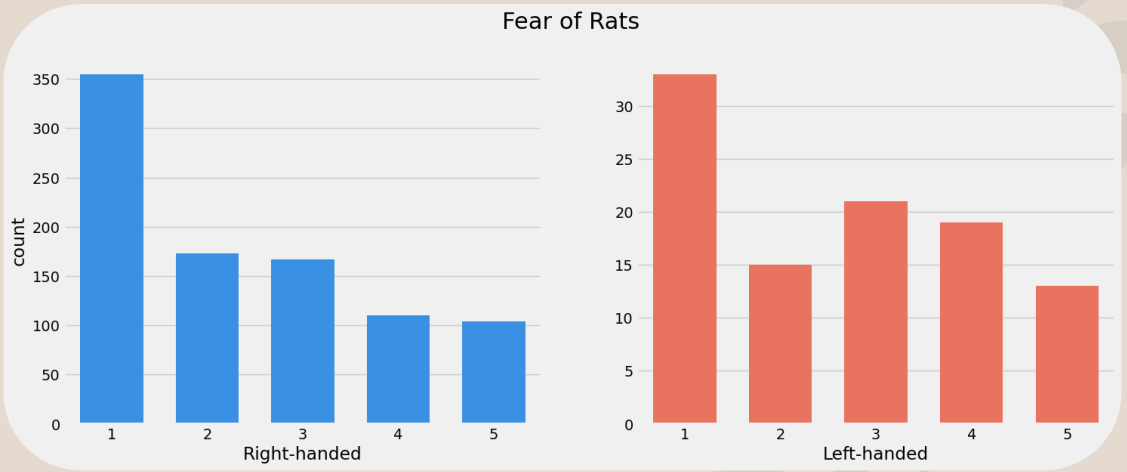


No variation: Spiders, Snakes, Rats & Dogs







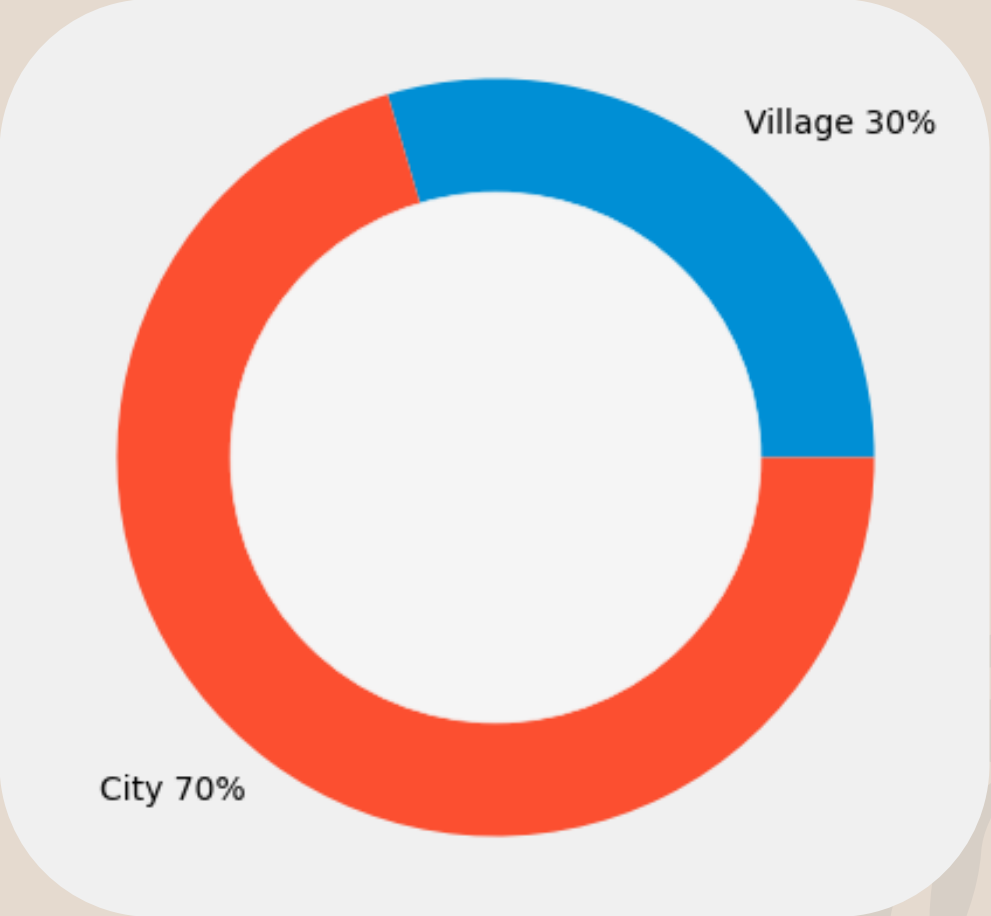


Childhood's Impact

Analyze if there are differences between the participants based on where they lived most of their childhood: rural and urban area.

Rural to Urban Ratio

Most students who took the survey have lived most of their childhood in cities.

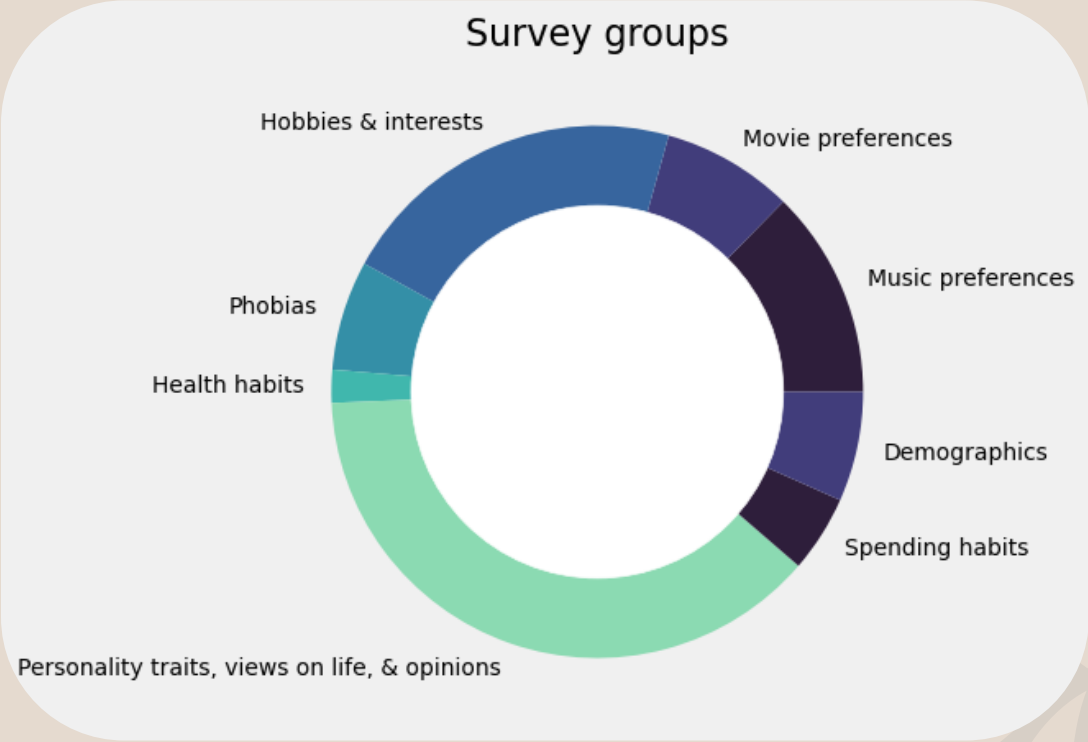


Correlation with the survey groups

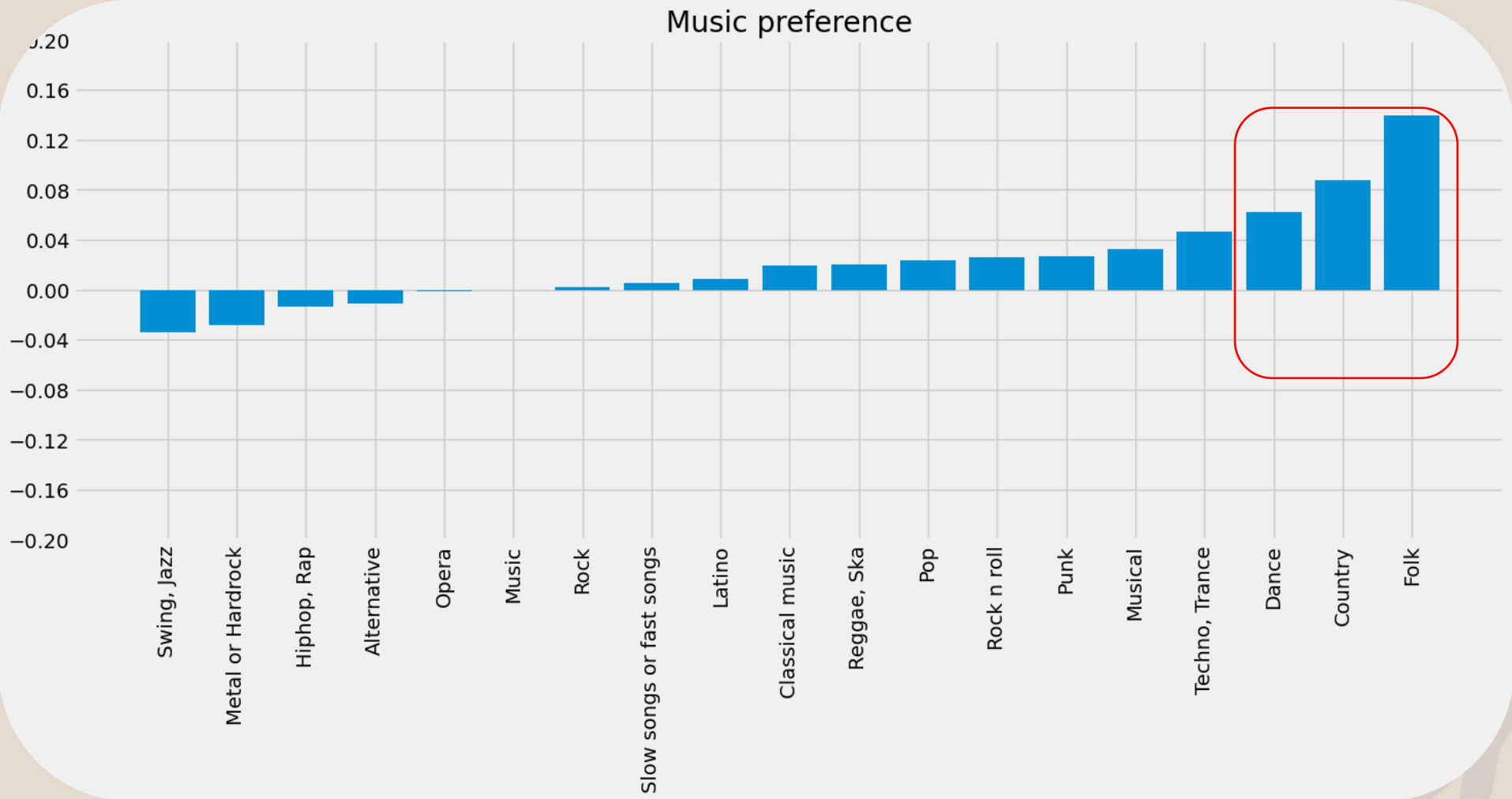
- Compute Pearson's coefficient for each features against a boolean feature that's True if the student has lived most of his/her childhood in a village.
- Range: -1 to 1

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

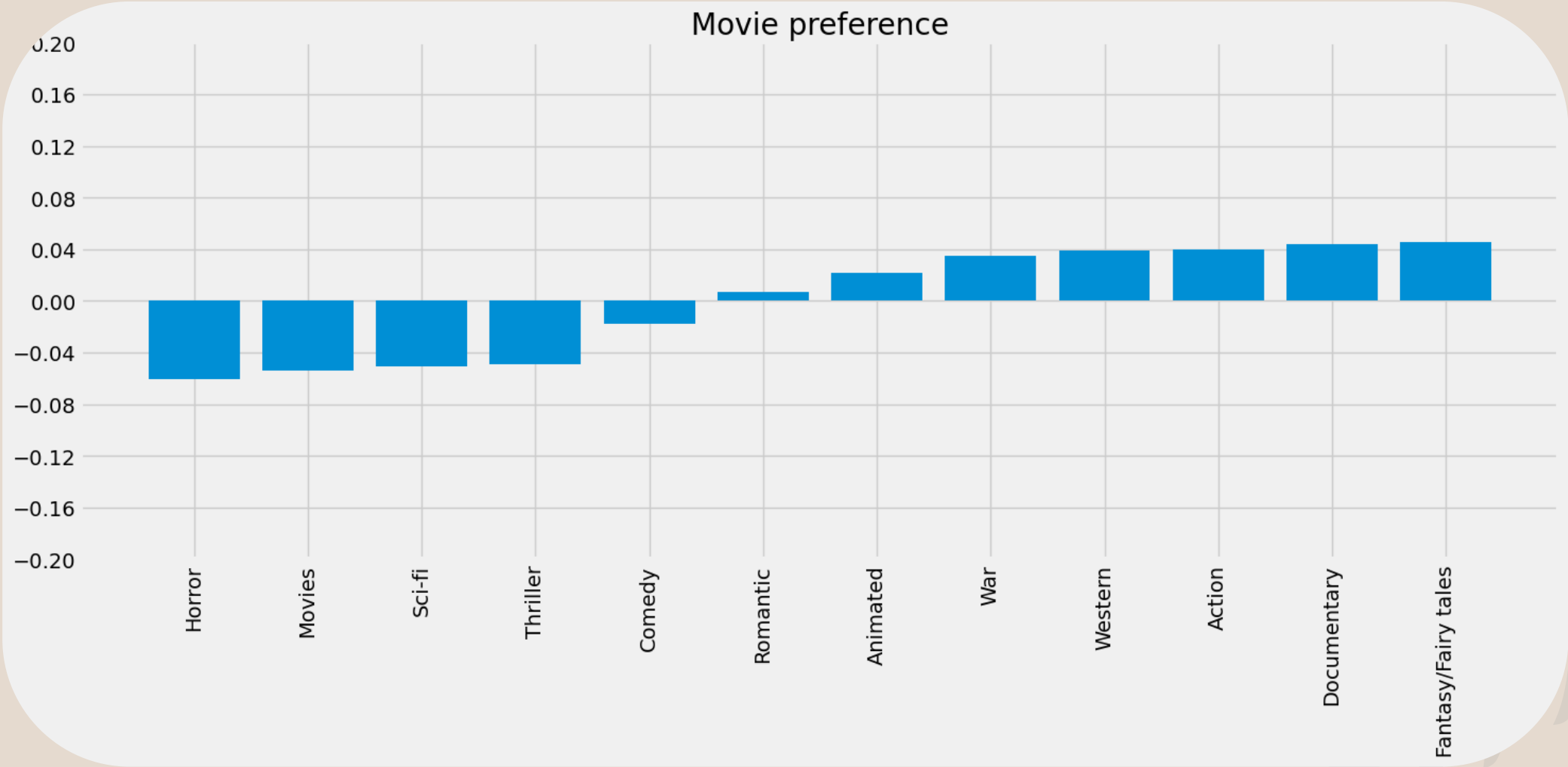
Where, $\overline{X}$ = mean of X variable
$\overline{Y}$ = mean of Y variable



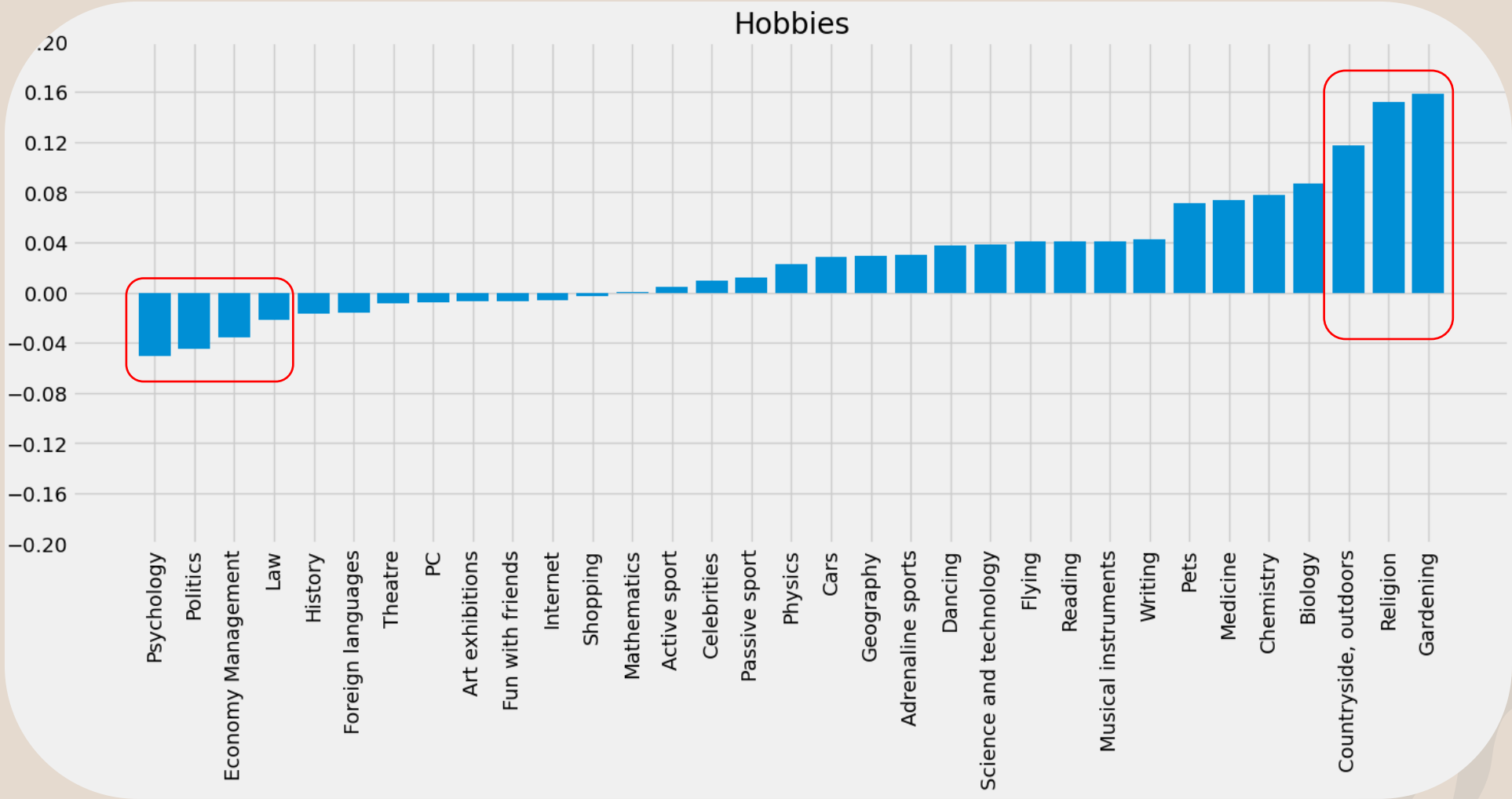
Music


Music preference
0.20
0.16
0.12
0.08
0.04
0.00
−0.04
−0.08
−0.12
−0.16
−0.20
Swing, Jazz
Metal or Hardrock
Hiphop, Rap
Alternative
Opera
Music
Rock
Slow songs or fast songs
Latino
Classical music
Reggae, Ska
Pop
Rock n roll
Punk
Musical
Techno, Trance
Dance
Country
Folk

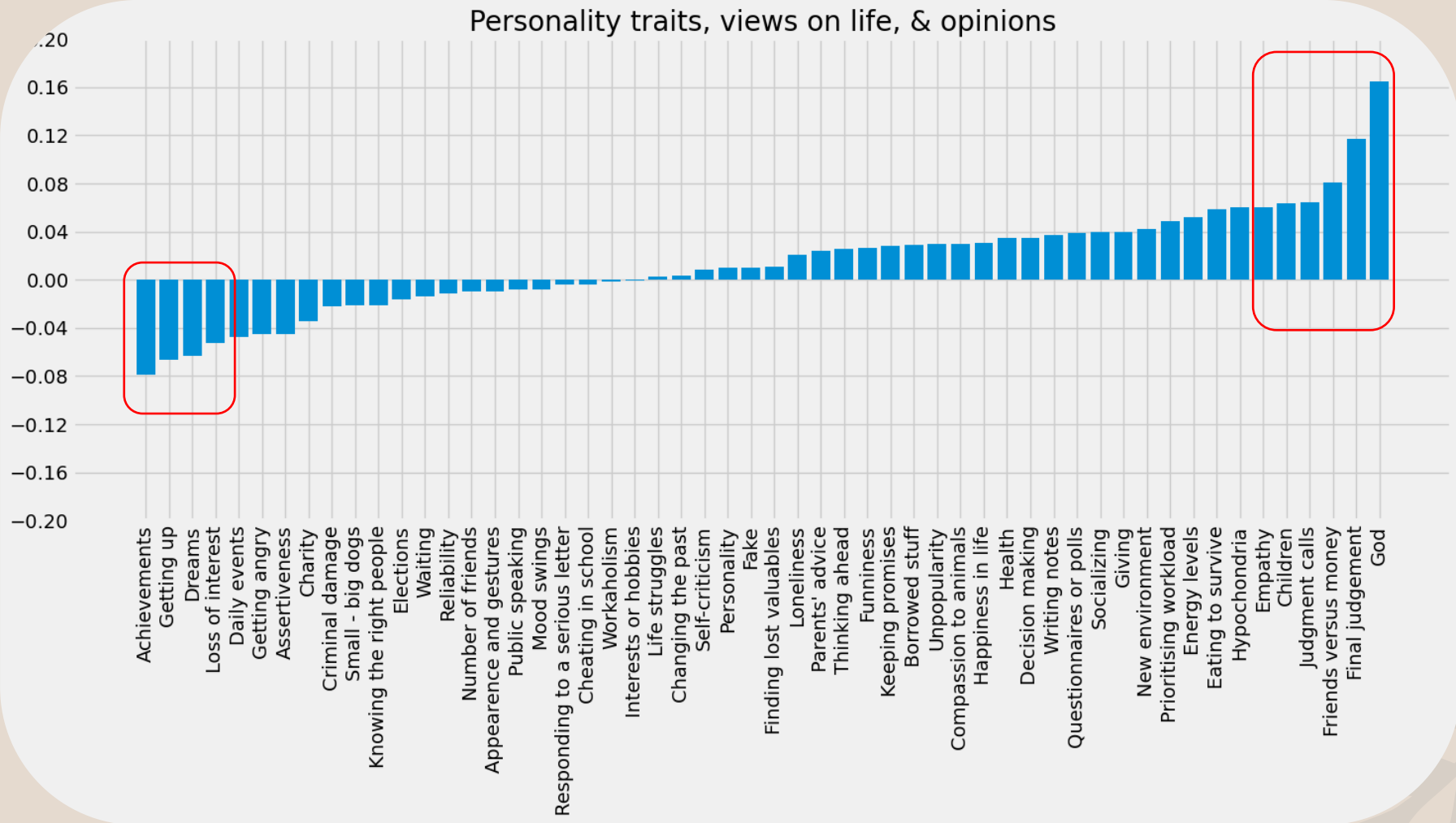
Movies


Movie preference
0.20
0.16
0.12
0.08
0.04
0.00
−0.04
−0.08
−0.12
−0.16
−0.20
Horror
Movies
Sci-fi
Thriller
Comedy
Romantic
Animated
War
Western
Action
Documentary
Fantasy/Fairy tales

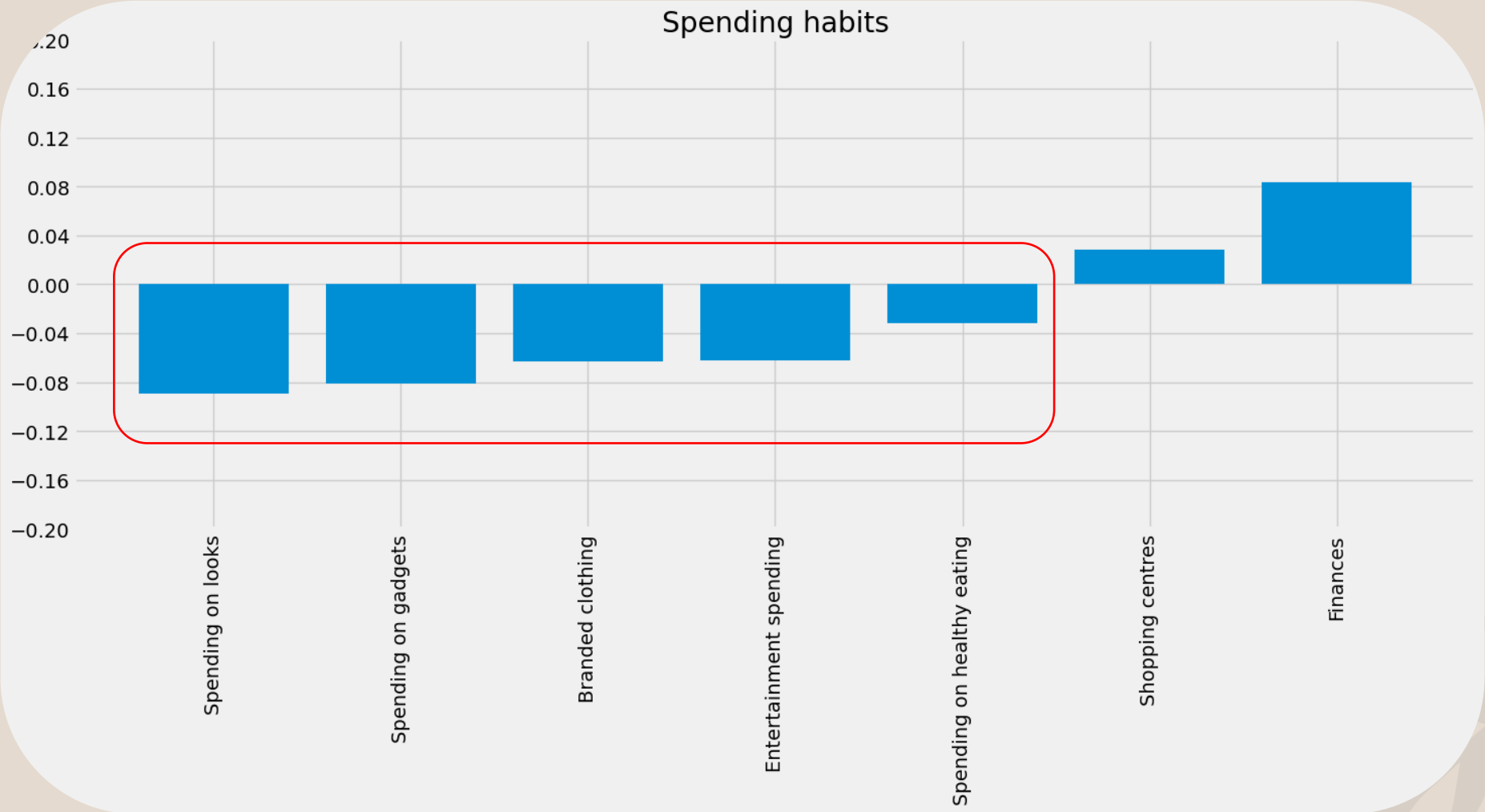
Hobbies


Hobbies
.20
0.16
0.12
0.08
0.04
0.00
−0.04
−0.08
−0.12
−0.16
−0.20
Psychology
Politics
Economy Management
Law
History
Foreign languages
Theatre
PC
Art exhibitions
Fun with friends
Internet
Shopping
Mathematics
Active sport
Celebrities
Passive sport
Physics
Cars
Geography
Adrenaline sports
Dancing
Science and technology
Flying
Reading
Musical instruments
Writing
Pets
Medicine
Chemistry
Biology
Countryside, outdoors
Religion
Gardening

Personality, Life Opinions



Spending Habits



Insights

Music	Tend to like country/folk music and dance.
Hobbies	Like to engage in gardening, religious, and outdoor activities compared to dwelling on psychology, politics and the economy.
Life	People of faith; strong belief in judgement after death, value friends over money; not lured by achievements and dreams.
Spending habits	Tend to save money due to their tough upbringing.

Missing Data Analysis

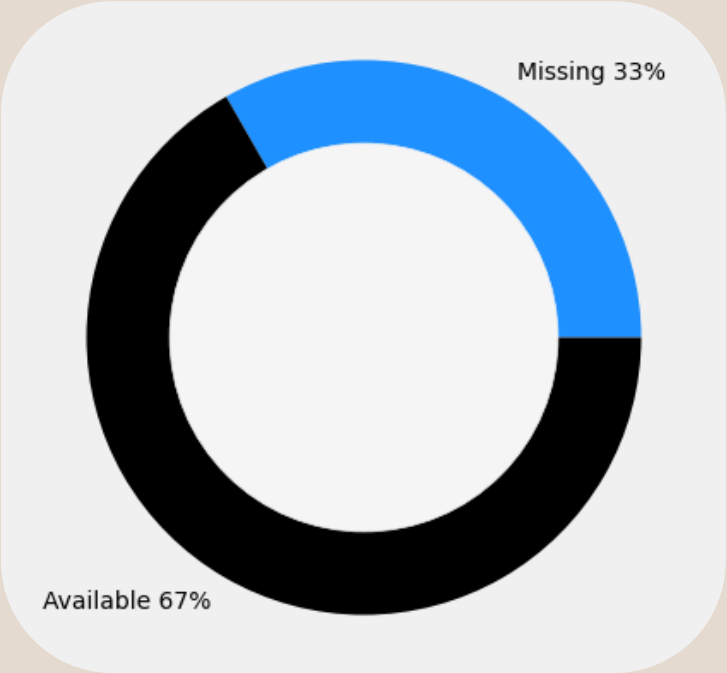
Is there are trend between the questions students abstain from answering?

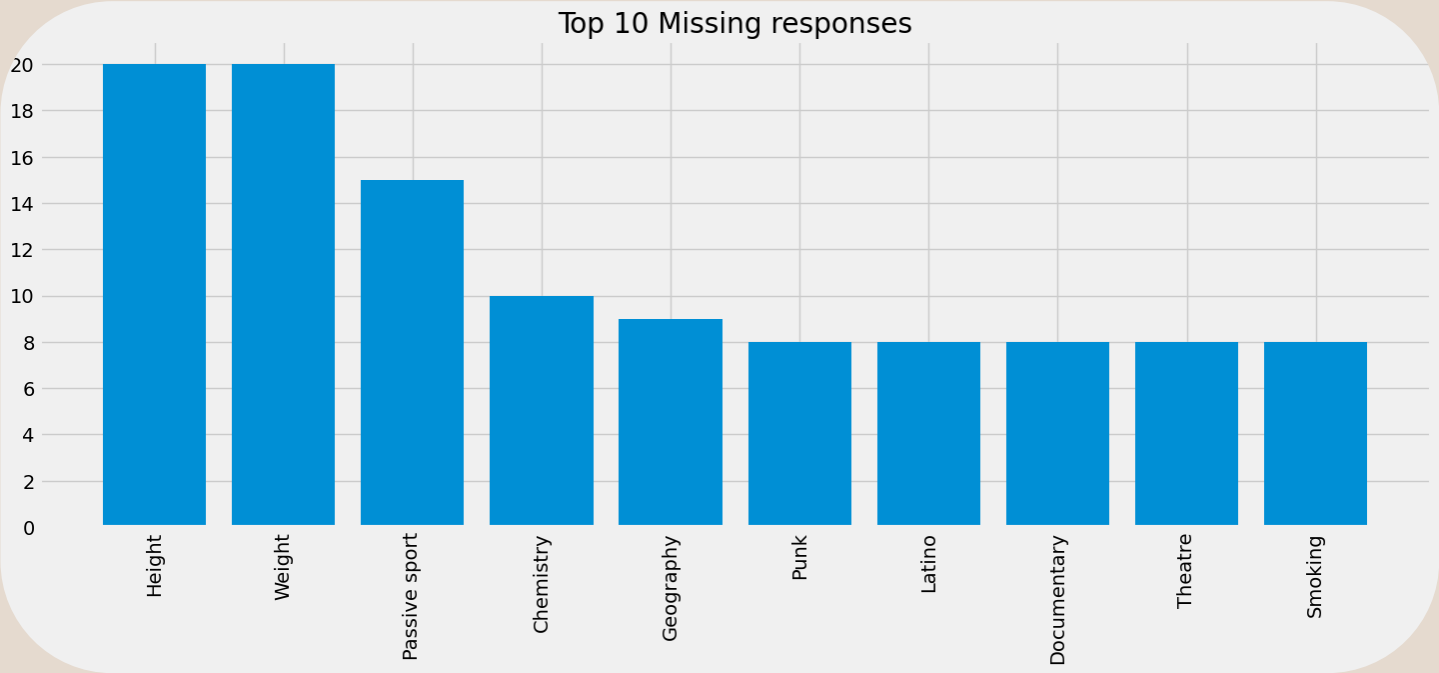
Analyze what?

- A missing response indicates that the student chose not to answer that survey question.
- Is there are trend between the questions students abstain from answering?
- Let's check if there's a correlation between unanswered questions. We're trying to answer, whether people who did not answer Q1 did not answer Q2 as well.

Unanswered Survey Questions

- Seemingly, one third of the data has at least one instance missing.
- However, of the 1010 instances, the maximum missing values from a single attribute are only 20.





Pearson's Coefficient

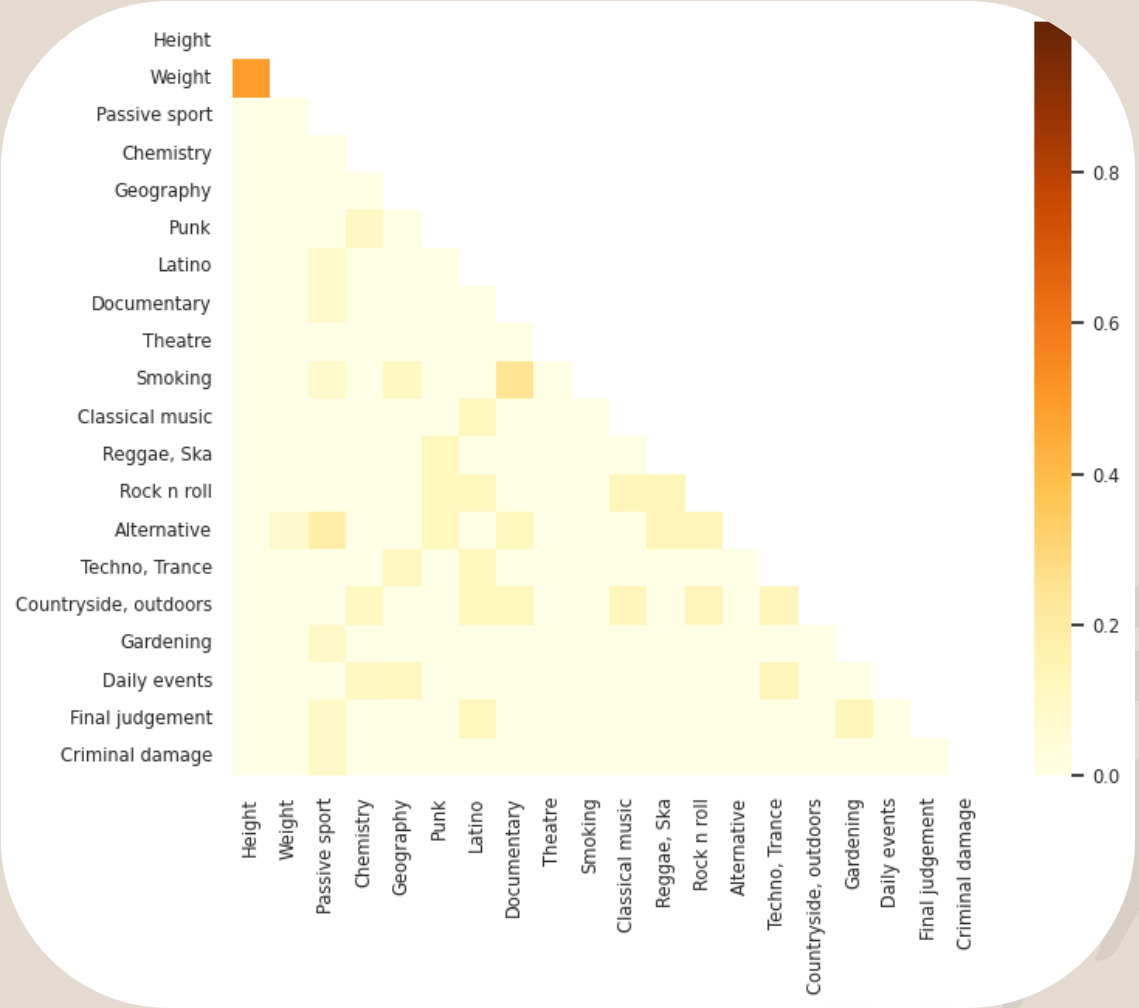
- Compute Pearson's coefficient pairwise between top k features.
- Top k based on #missing entries.
- This results in a k * k matrix.

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable
$\overline{Y}$ = mean of Y variable

Pearson's Coefficient

- Compute Pearson's coefficients between the most frequently unanswered survey questions.
- Mostly uncorrelated.
- Correlations:
  - Height & Weight: Body consciousness.
  - Smoking & Documentary?

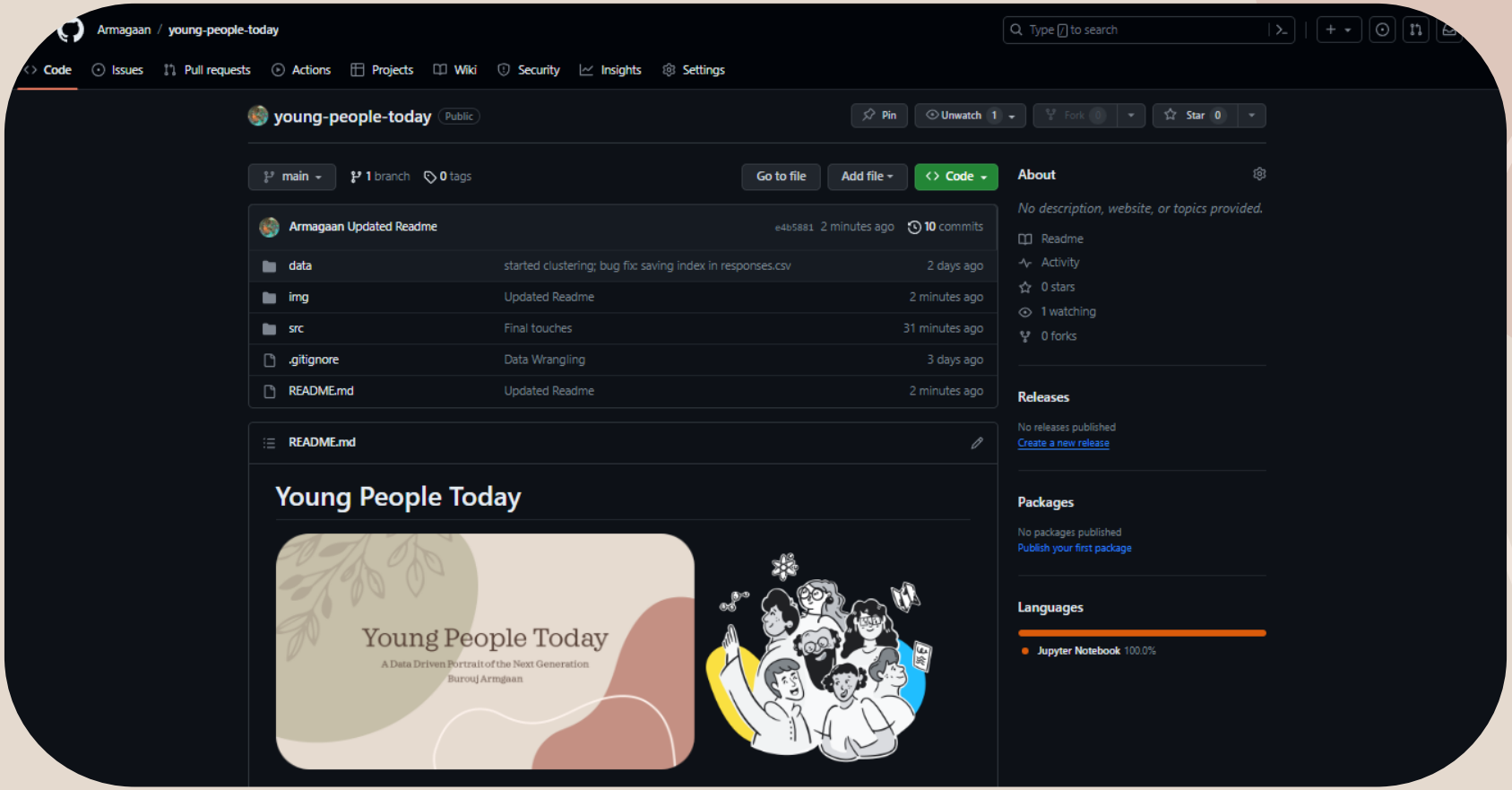


Conclusion

- Clustering
  - Given the music preferences, do people make up any clusters of similar behavior?
- Hypothesis Testing
  - Do women fear certain phenomena significantly more than men?
  - Do left-handed people have different interests than right-handed?
- Childhood's Impact
  - Analyze if there are differences between the participants based on where they lived most of their childhood: rural and urban area.
- Missing Data Analysis
  - Is there are trend between the questions students abstain from answering?

Code is up

SCAN ME



Thank you

Burouj Armgaan