Data Madness

Martin Gassner, Henry Mauranen, Matteo Maggiolo

17 March 2018

Introduction

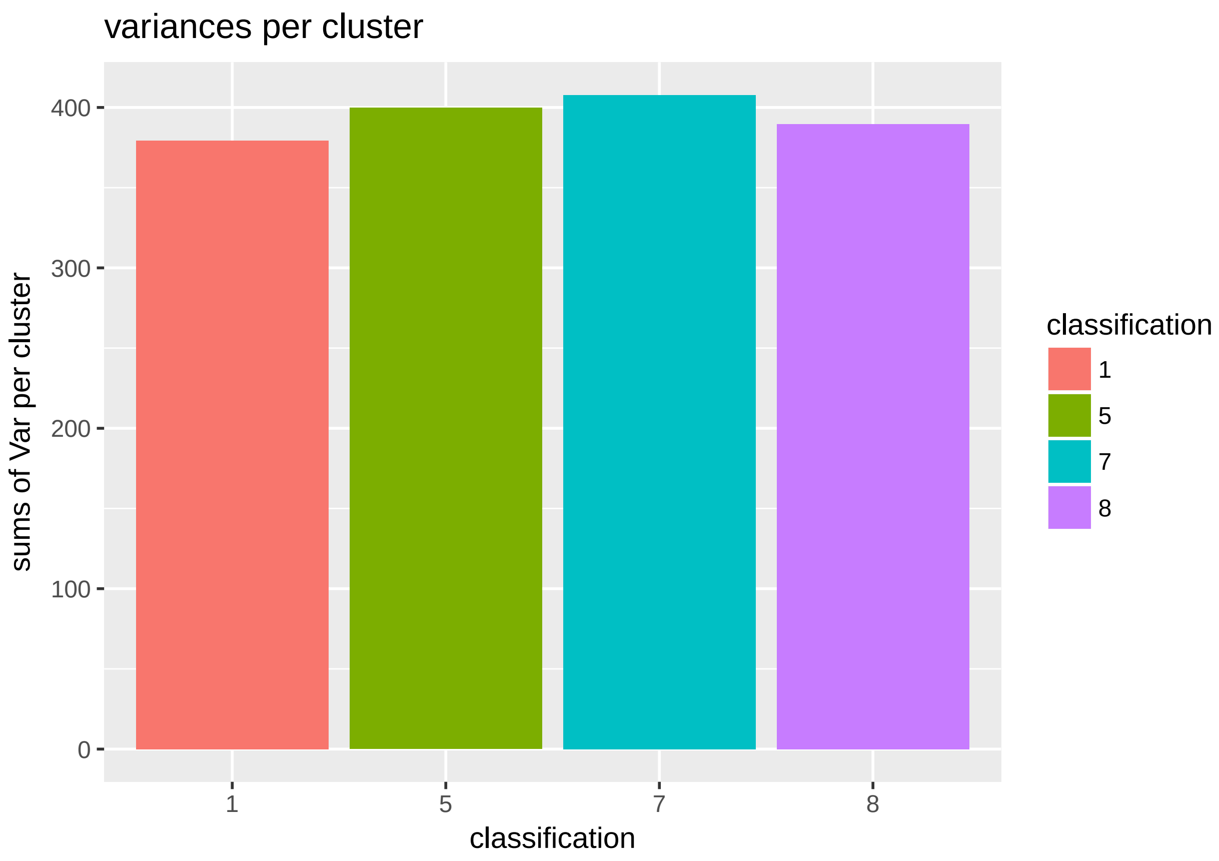
We worked on a dataset containing the recorded answers to a suvey about morality.

The questions we would like to adress are as follows:

1. Can we find out ways to interpret the answers of each block as one value?
2. Can we identify distinct clusters within different blocks of questions?
3. Based on these values and/or clusters, are there clear correlations between interpretable blocks?
4. Are there interpretable types of individuals?
5. Can we link the clusters to personal metrics?

EM clustering

The first approach for finding common groups is to cluster the dataset as whole. Here we decided on using EM clustering because of an existing R package. Even though EM is not specifically tailored for catecorigal data, which especially the refusal to answer is, we deemed it suitable. We could also contrast this with kmodes later during the analysis. A problem with this approach is how to deal with missing values and the mentioned refusals to answer a question. We decided to deal with this by imputing from a column-wise empirical distribution. This way the data set won't enforce existing relations, like imputation by bayesian inference would, and the distribution is maintained, so the mean and variance won't change.



After clustering, we observed 8 distinct groups. 4 of these consisted of 4 or fewer individuals, so we decided to focus on the 4 other major clusters. We then plotted the answer frequencies by cluster and question to observe distinctive features from each cluster. Major opinion divider here was religious identity. Observations for the 4 clusters are as follows:

1. Strongly non- or anti-religious. This was visible in all religious questions throughout the questionnaire. Interestingly, when compared to other clusters, this group showed a nearly equal compassion and adherence to ethics (when asked questions about things such as insurance fraud or cheating in an exam) as groups 2 and 3. This group showed some degree of preference towards moral relativism than other groups.

2. Strongly religious group. Similar to group 3, this group considers religion very important. To contrast, they are more relativistic about morals than group 3 and they tended to answer slightly weaker agreement within questions about sacredness of moral values.

3. Strongly religious group. As hilighted by cluster 2, this group appeared to be a more strict group of religious individuals than group 2.

4. Normal group. This group tended to answer neither agree or disagree for most of the questions, especially the ones about identity. They aren't characterized by any specific question in the entire questionnaire and won't stand out compared to the other 3 groups. However, the interesting observation was that this group considers moral questions, such as the ones in every day moralism section, more loosely than other groups. Their answers didn't show as uniform agreement about immoralities of things like lying or fraud as other groups did.

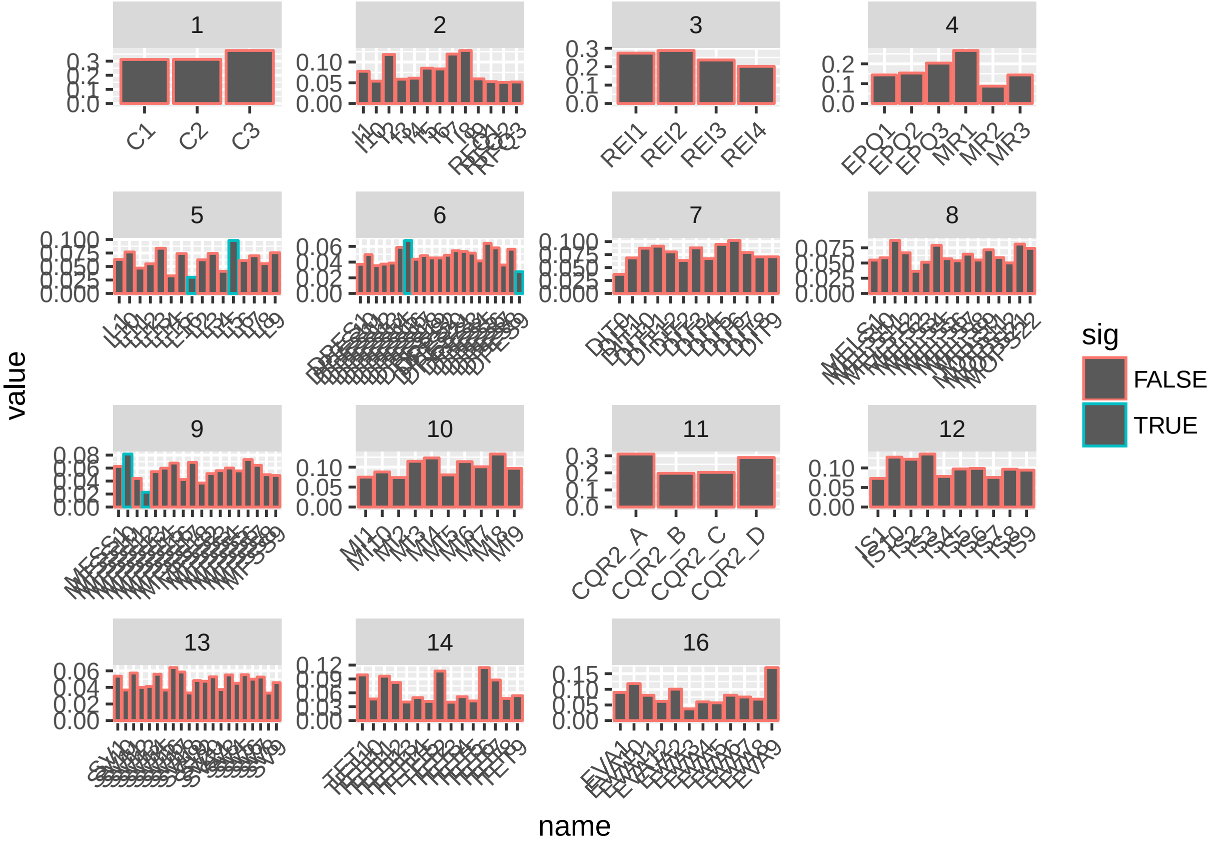
Finally, we were interested if some of the groups are more uniform in their values than other groups. We summed up variances from each question, producing the plot dispalyed after em_clustering block. The differences in variances are very small and practically insignificant. This is quite surprising as it could be epected that religious groups in USA, being mainly christian, would have a more uniform set of moral principles.

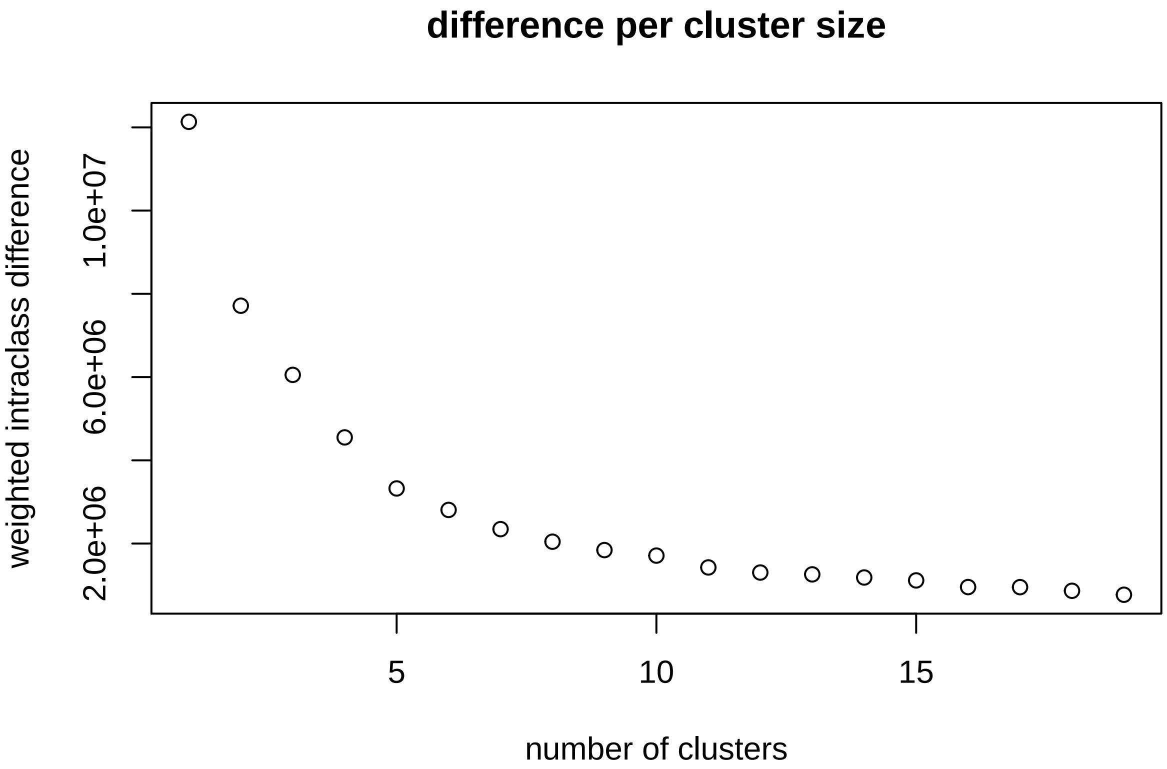
PCA weighting + kmeans

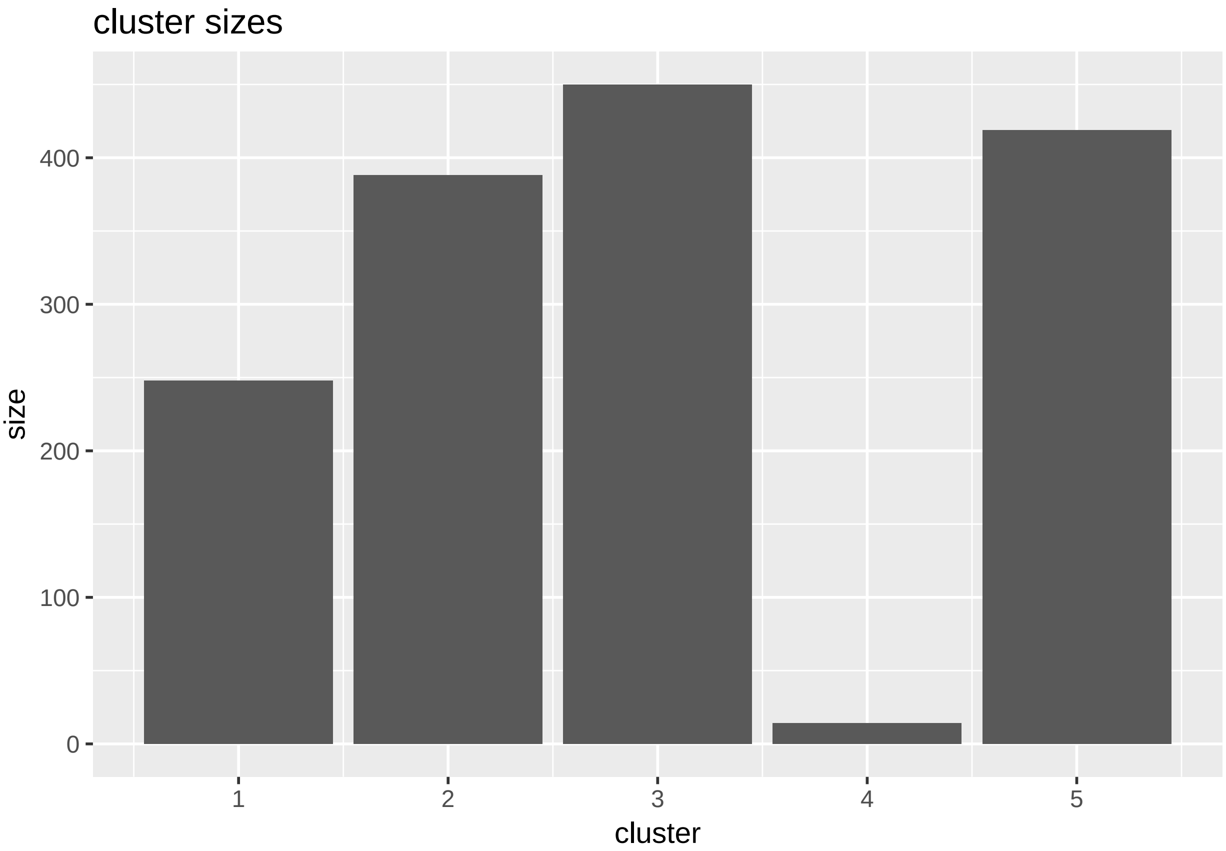
To obtain a weighting for each block of questions

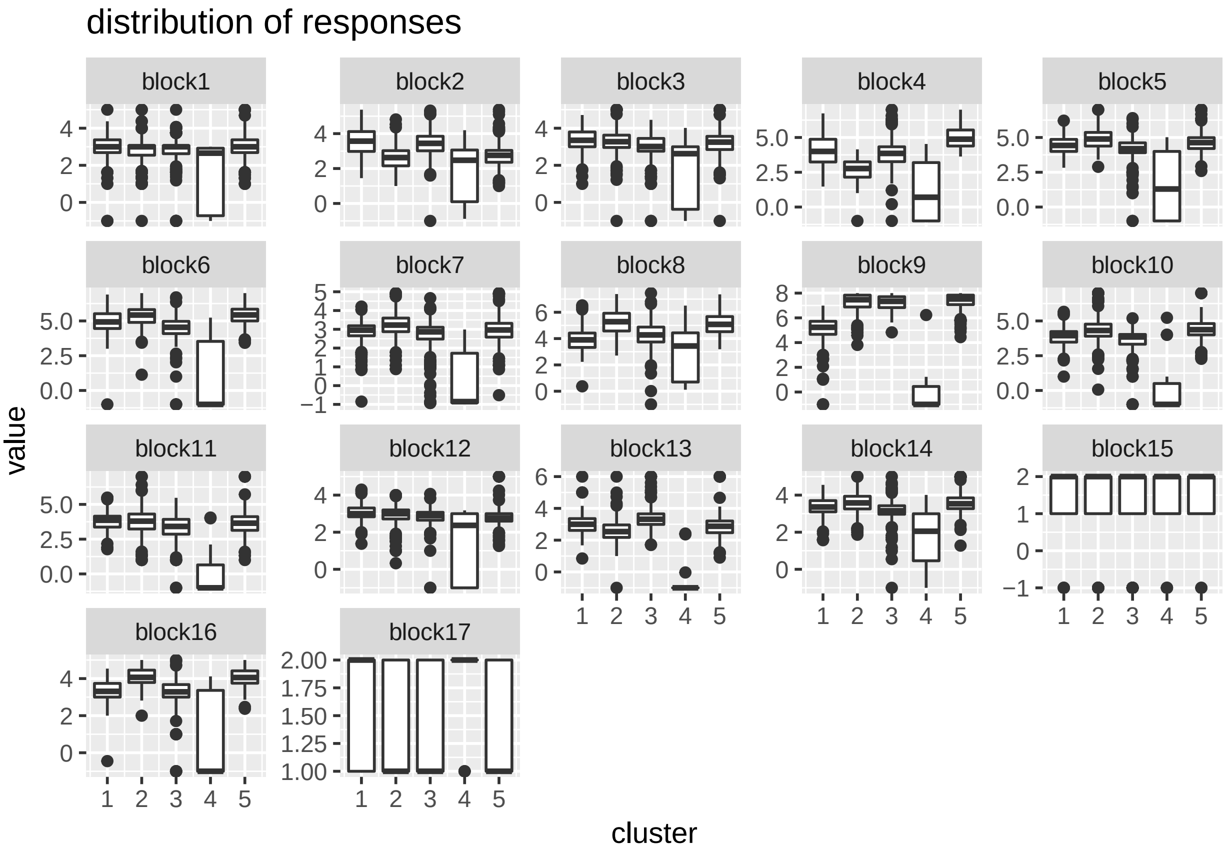
1. obtain the principal components for each block
2. keep the ones explaining 85% of the variance
3. for each variable: sum the loadings of the most important principal components, weighted by the variance explained.

Additionally we tested the weights for significance within each block.

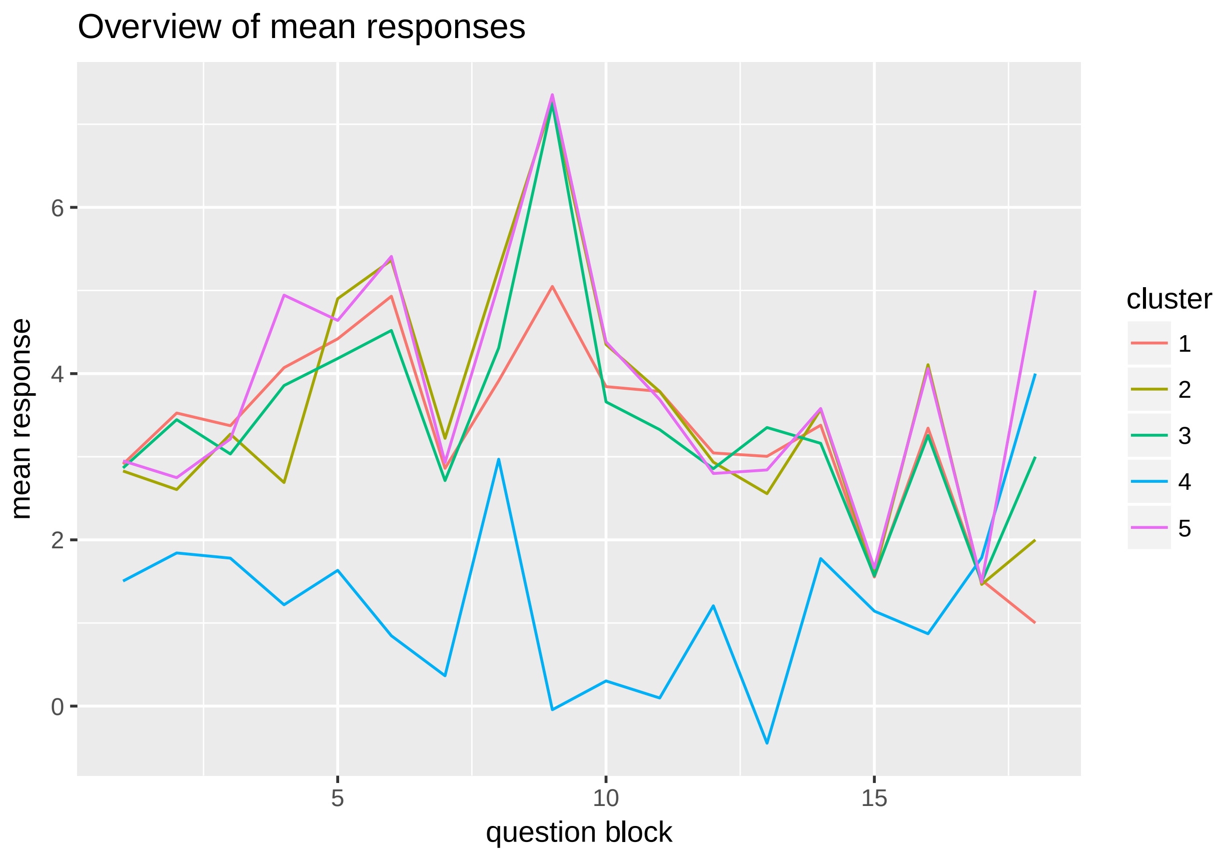



difference per cluster size
weighted intraclass difference
2.0e+06
6.0e+06
1.0e+07
5
10
15
number of clusters


cluster sizes
size
400
300
200
100
0
1
2
3
4
5
cluster



distribution of responses
block1
block2
block3
block4
block5
block6
block7
block8
block9
block10
block11
block12
block13
block14
block15
block16
block17
value
cluster
1 2 3 4 5
0
2
4
0.0
2.5
5.0
−1
1
3
6
8
1.00
1.25
1.50
1.75
2.00



big differences Block 2: Identity measures - clusters 2,5 (less important) vs. 1,3 (very important)

highly weighted: job, political party, favorite sports team

Block 4: Moral relativism - cluster 2 (relative) vs 5 (disagreeing)

higly weighted: What is right and wrong is up to each person to decide

Block 9: Sacredness of moral principles - cluster 1 (less) vs 2,3,5 (more)

higly weighted: Sign a piece of paper that says "I hereby sell my soul, after my death, to whoever has this piece of paper" (SIGNIFICANT)

low weighted: Attend a performance art piece in which all participants (including you) have to act like animals for 30 minutes, including crawling around naked and urinating on stage

Block 17: all different; rank 1, 2, 3, 4, 5

```
## [1] "Block 2: Identity measures - clusters 2,5 (less important) vs. 1,3 (very important)"

## [1] 0.2268667

## [1] "Block 4: Moral relativism - cluster 2 (relative) vs 5 (disagreeing)"

## [1] 8.384556e-214

## [1] "Block 9: Sacredness of moral principles - cluster 1 (less) vs 2,3,5 (more)"

## [1] 0.004285064

## [1] "Block 17: Dictator game - cluster 1 vs 2, 3, 4, 5"

## [1] 0.001281282
## [1] "Block 17: Dictator game - cluster 2 vs 1, 3, 4, 5"
## [1] 5.086156e-153
## [1] "Block 17: Dictator game - cluster 3 vs 1, 2, 4, 5"
## [1] 0.7656305
## [1] "Block 17: Dictator game - cluster 4 vs 1, 2, 3, 5"
## [1] 0.0008385006
## [1] "Block 17: Dictator game - cluster 5 vs 1, 2, 3, 4"
## [1] 1.022094e-149
```

Cluster 1

- find many things (country, religion etc) relevant to their identity
- do not hold moral values high/more pragmatic
- give the least in the dictator game

Cluster 2

- find less things relevant to their identity
- moral relativists
- do think of their moral values sacred/would take a lot of money
- second to last in dictator game

Cluster 3

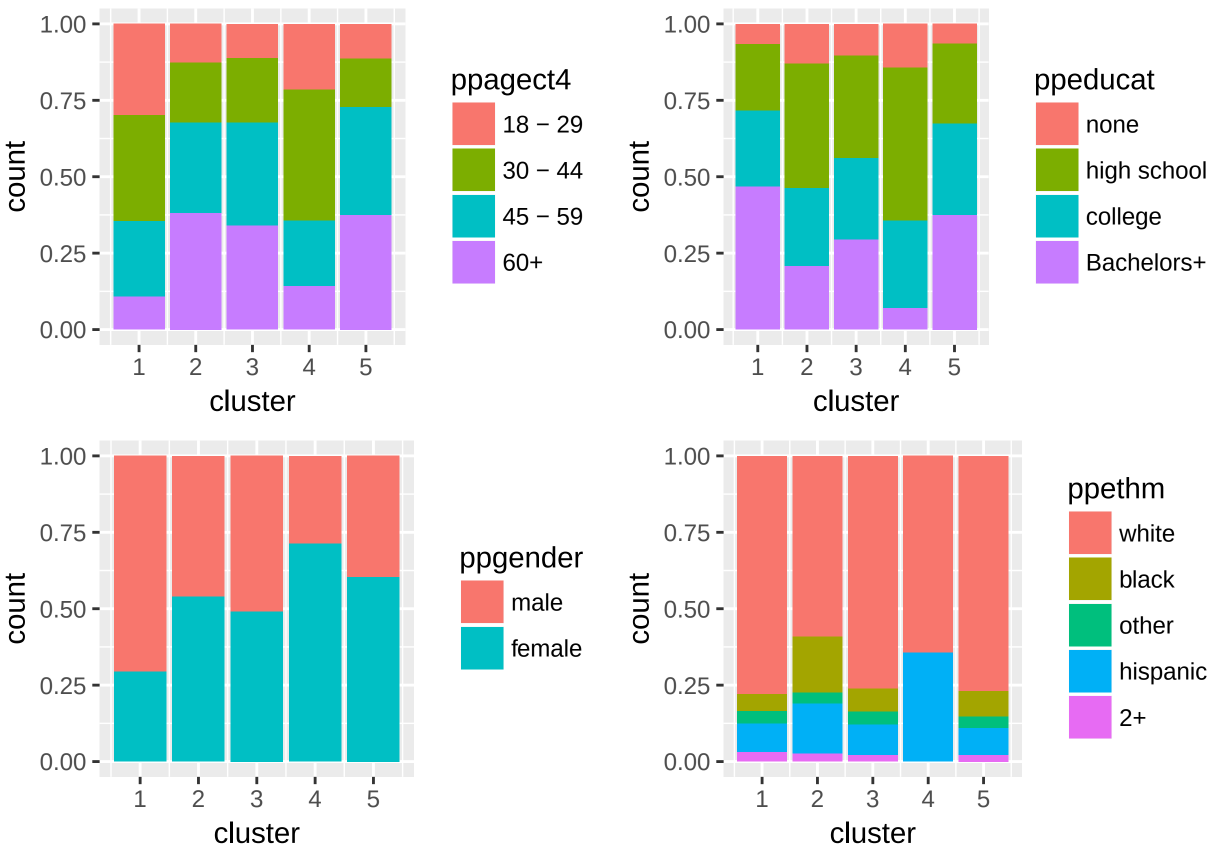
- average, really everywhere

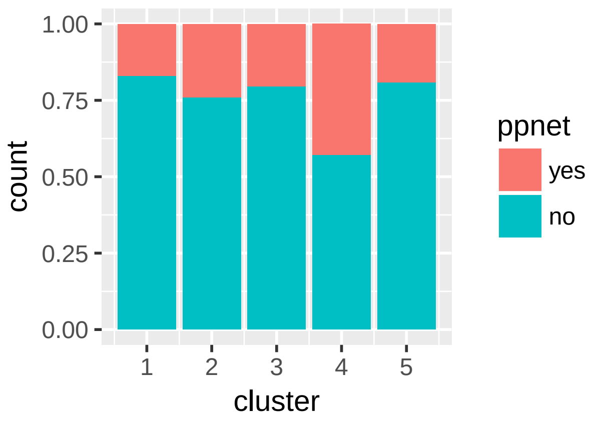
Cluster 4

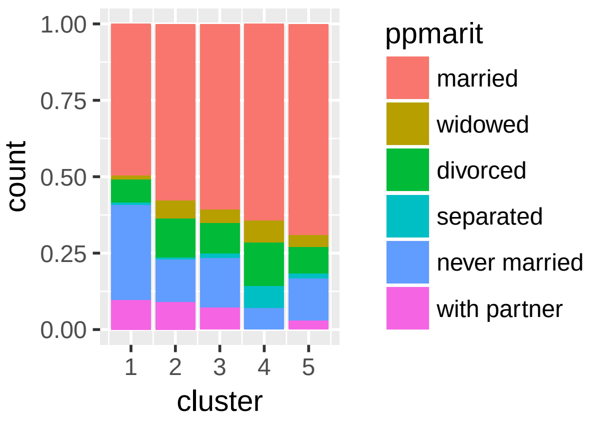
- irrelevant, too small, collects refused answers

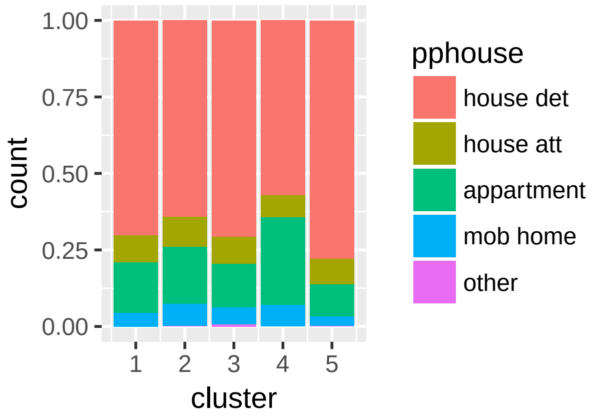
Cluster 5

- find less things relevant to their identity
- oppose moral relativism
- do hold their morals dear
- give the most in the dictator game (exactly average)

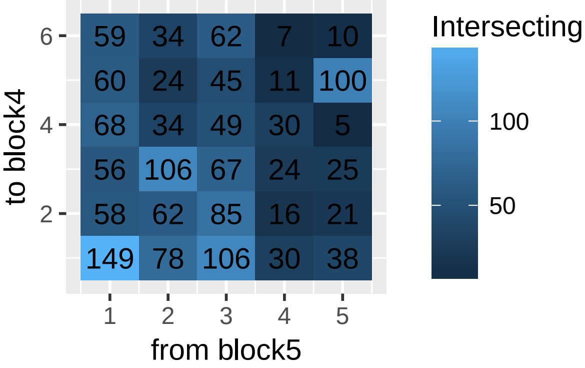

count
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
cluster
ppagect4
18 − 29
30 − 44
45 − 59
60+
count
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
cluster
ppeducat
none
high school
college
Bachelors+
count
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
cluster
ppgender
male
female
count
1.00
0.75
0.50
0.25
0.00
1
2
3
4
5
cluster
ppethm
white
black
other
hispanic
2+

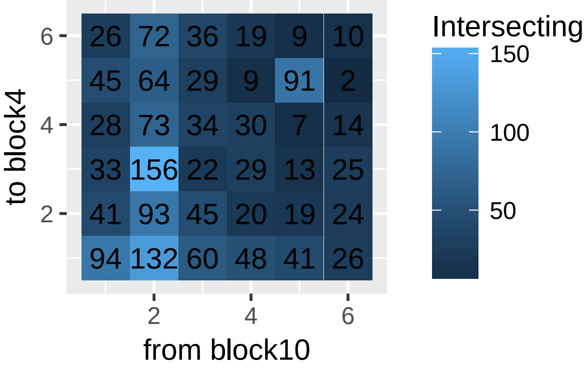

1.00
0.75
0.50
0.25
0.00
count
1
2
3
4
5
cluster
ppnet
yes
no

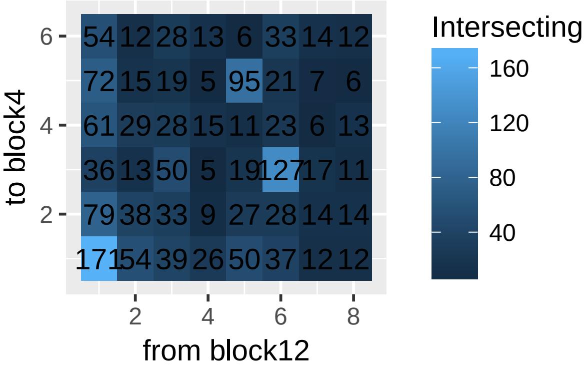

1.00
0.75
0.50
0.25
0.00
count
1
2
3
4
5
cluster
ppmarit
married
widowed
divorced
separated
never married
with partner


1.00
0.75
0.50
0.25
0.00
count
1
2
3
4
5
cluster
pphouse
house det
house att
appartment
mob home
other

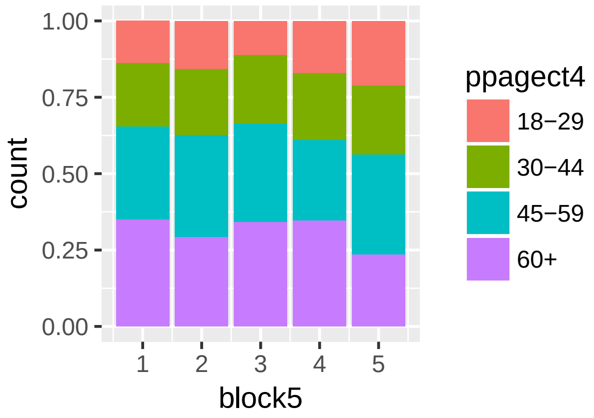
k-modes

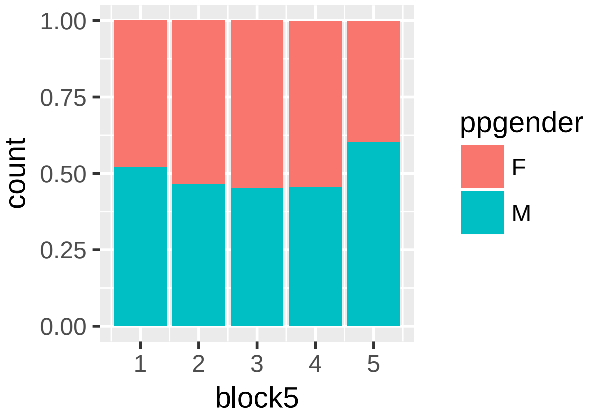

Transitions from block5 to block4

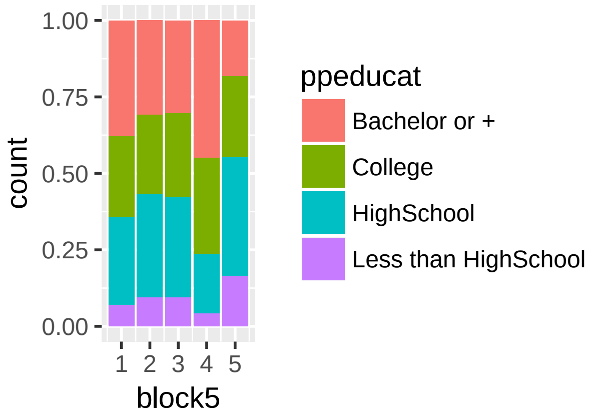

Transitions from block10 to block4

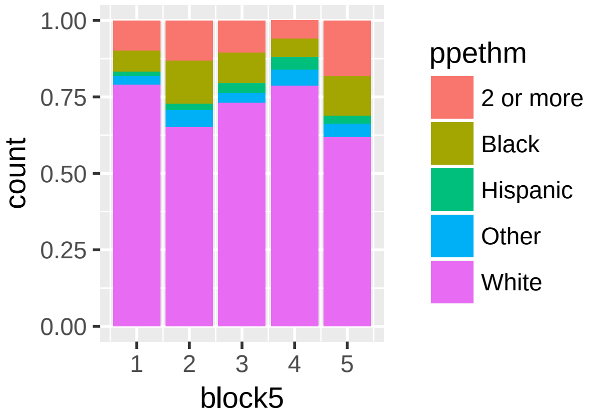

Transitions from block12 to block4

```
## [1] "Interesting transition from block5(5) to block4(5), block12(5) to block4(5), block10(5) to bloc
```

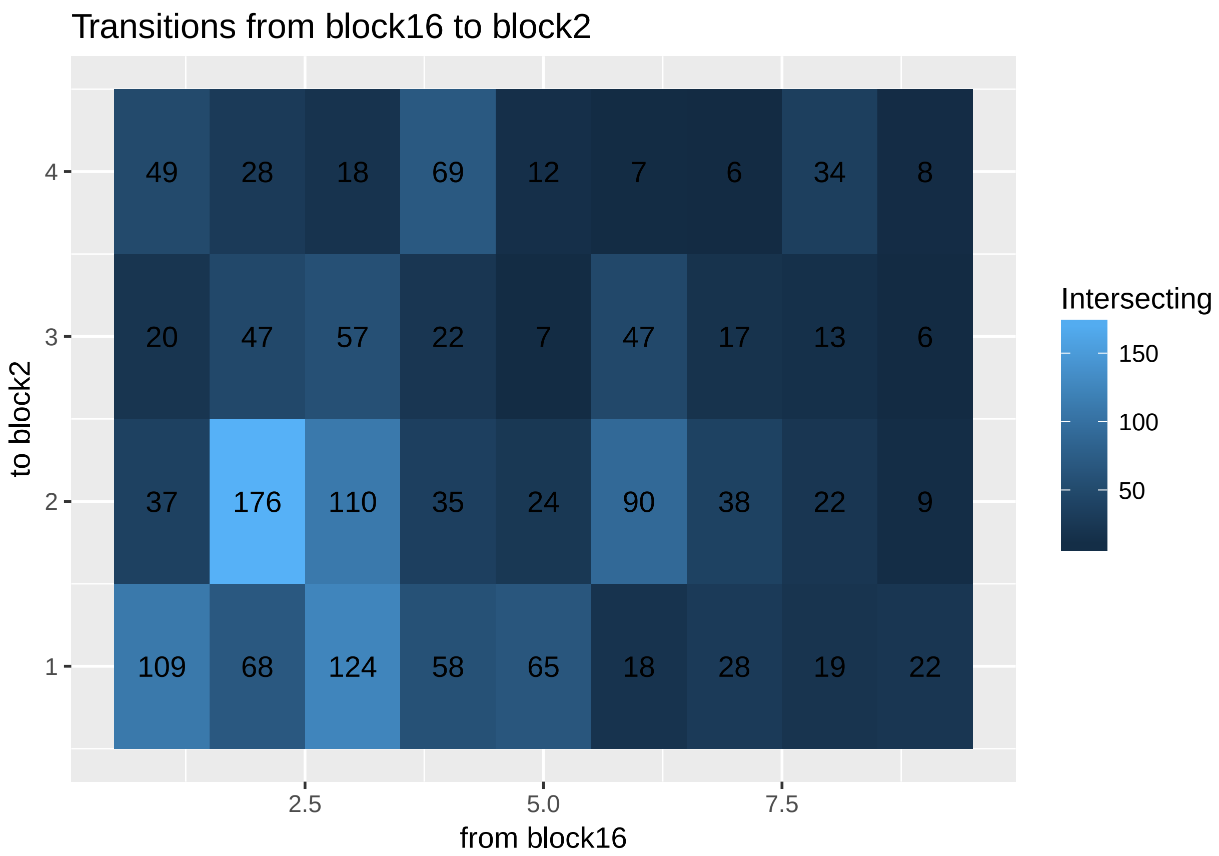




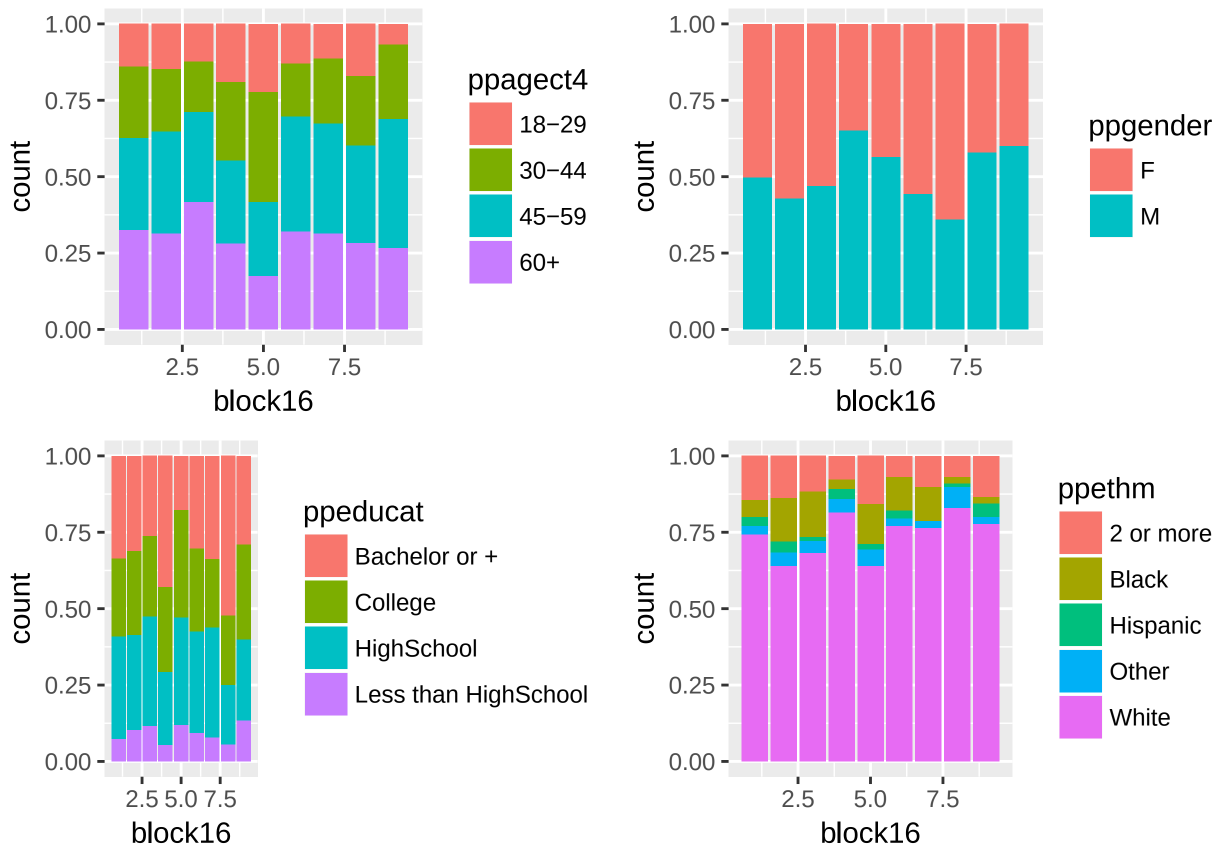




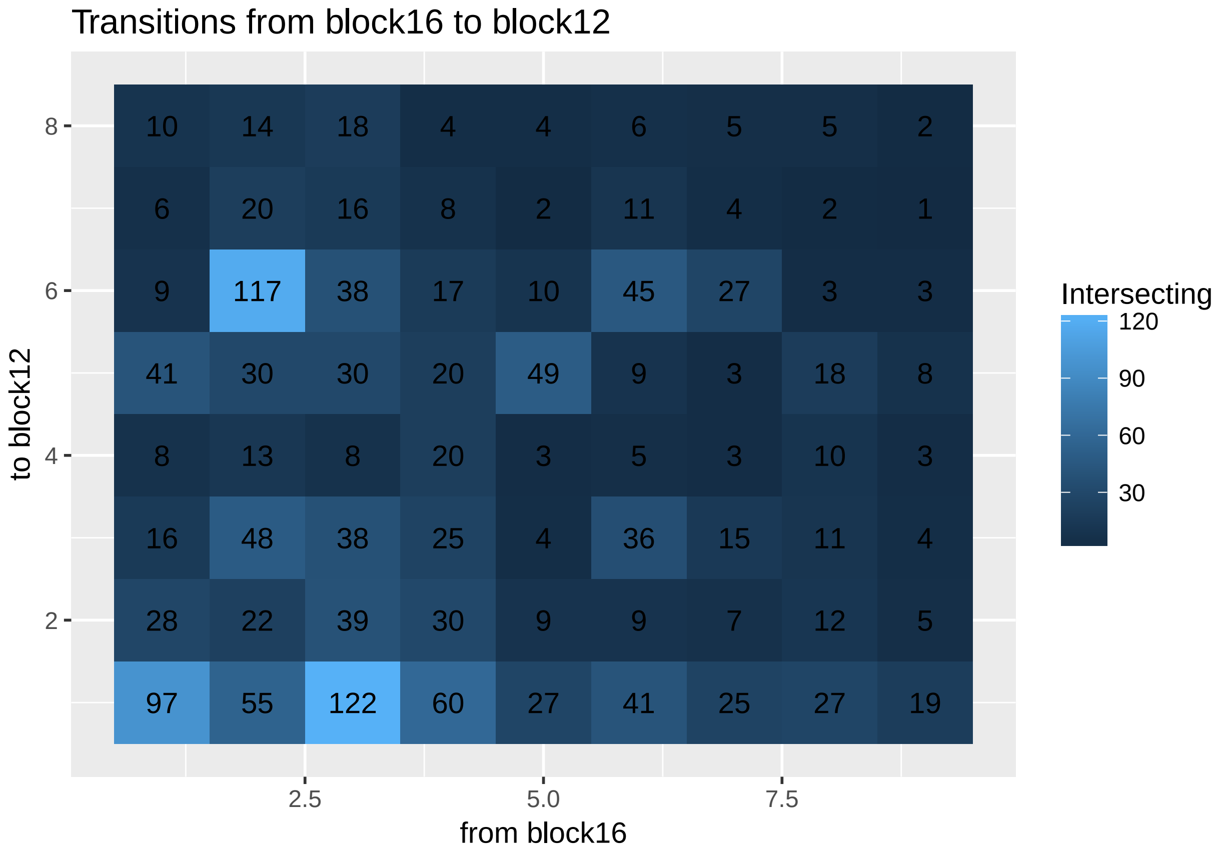
```
## [1] "Plotting only block 5 info [check cluster 5]"
```



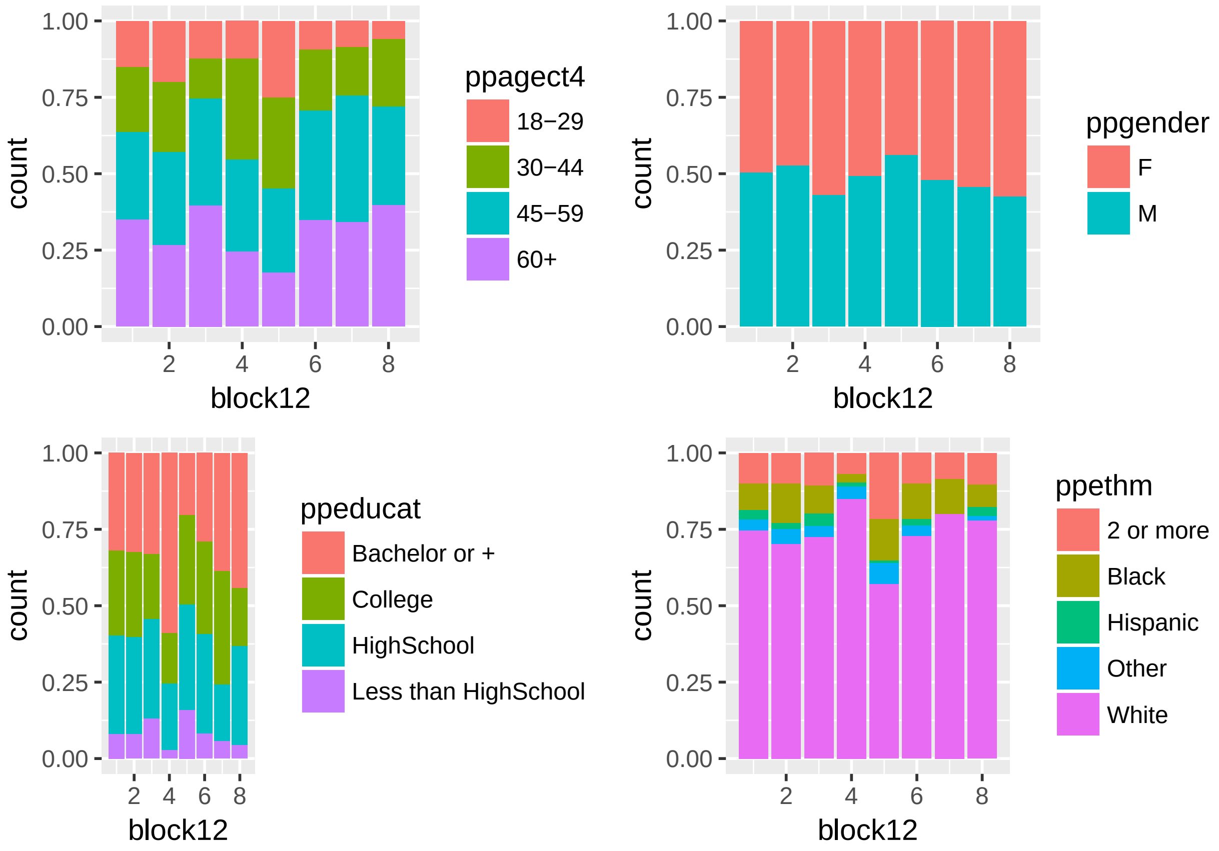
```
## [1] "Interesting transition from block16(2) and block2(2) [The non religious] [about 275 people in bo
```



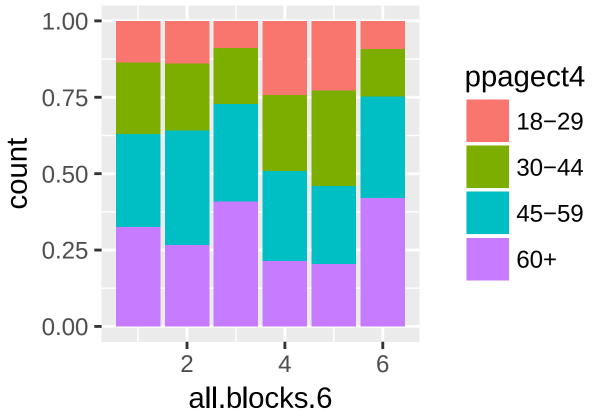
```
## [1] "Plotting only block 16 info [check cluster 2]"
```

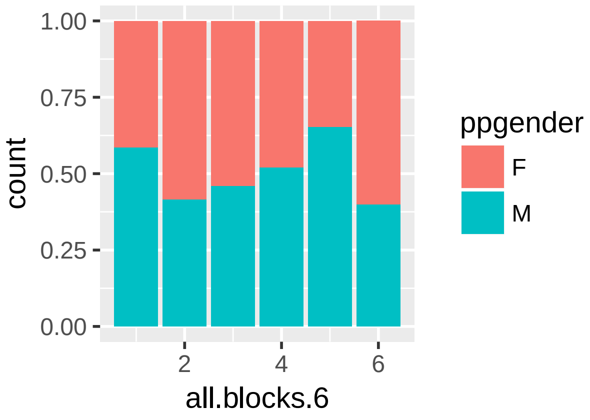


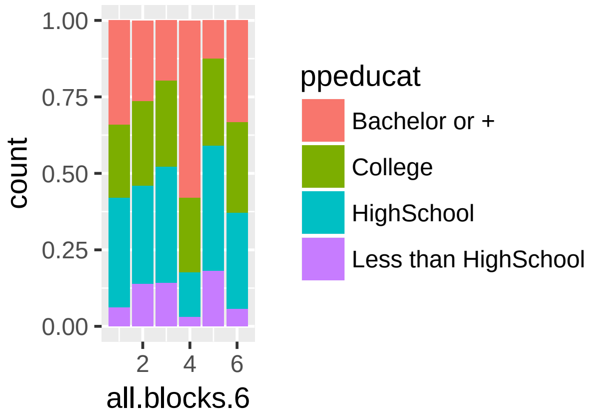
```
## [1] "Interesting transition between block16(3) and block12(1) [High integrity and moral values (chri
```

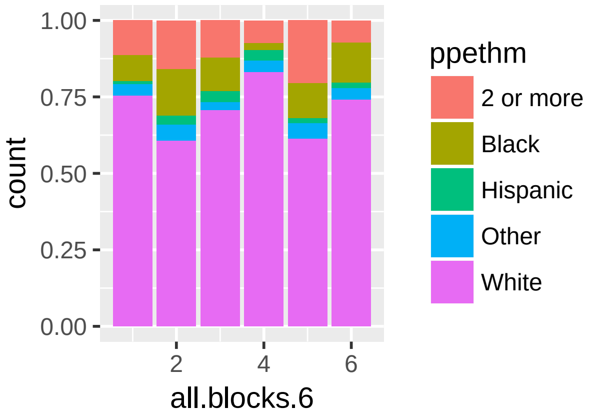


```
## [1] "Plotting only block 12 info [check cluster 6]"
```









```
## [1] "Clusters in all.blocks.6 => {6: [Non religious], 4: [Hippies/High integrity and moral], 1: [Alwa
```

Conclusions

1. Can we find out ways to interpret the answers of each block as one value?

Yes we can! We used PCA to obtain weights from the sum of loadings per variable weighted by the variance explained by each principal component.

2. Can we identify distinct clusters within different blocks of questions?

Yes we can! We performed k-modes clustering on a per block basis and obtained the cetroids of each cluster. We then related the results with the questions presented in the questionnaire.

3. Based on these values and/or clusters, are there clear correlations between interpretable blocks?

Sometimes. Our results show that the correlations between variables are not particularly strong. However we could find correlations between individual clusters for specific blocks. Additionally, there was often overlap between the clusters and only few variables clearly separated a cluster from the others.

4. Are there interpretable types of individuals?

Based on the per-block basis clustering we could find pairs of clusters correlated over multiple blocks. This established a link between blocks of questions for a specific cluster, such as:

- a group of individuals never taking a side
- a group of individuals with high answers to integrity and ethical values questions

Given our clustering on the entire dataset we could find 4 types

- Materialistic pragmatists
- Individualists
- Average people
- Saints
- Can we link the clusters to personal metrics?

Yes, we can! Corresponding to the previous 4 types

- Materialistic pragmatists
  - mostly young
  - educated
  - male
  - abstain from marriage
- Individualists
  - ethnically more diverse
  - less tertiary education
  - fewer living in detached houses
- Average people
  - guess what… average - everywhere!
- Saints
  - most likely to be married
  - least likely to have an appartment
  - most likely to be living in detached houses
  - more women