Proposed structure

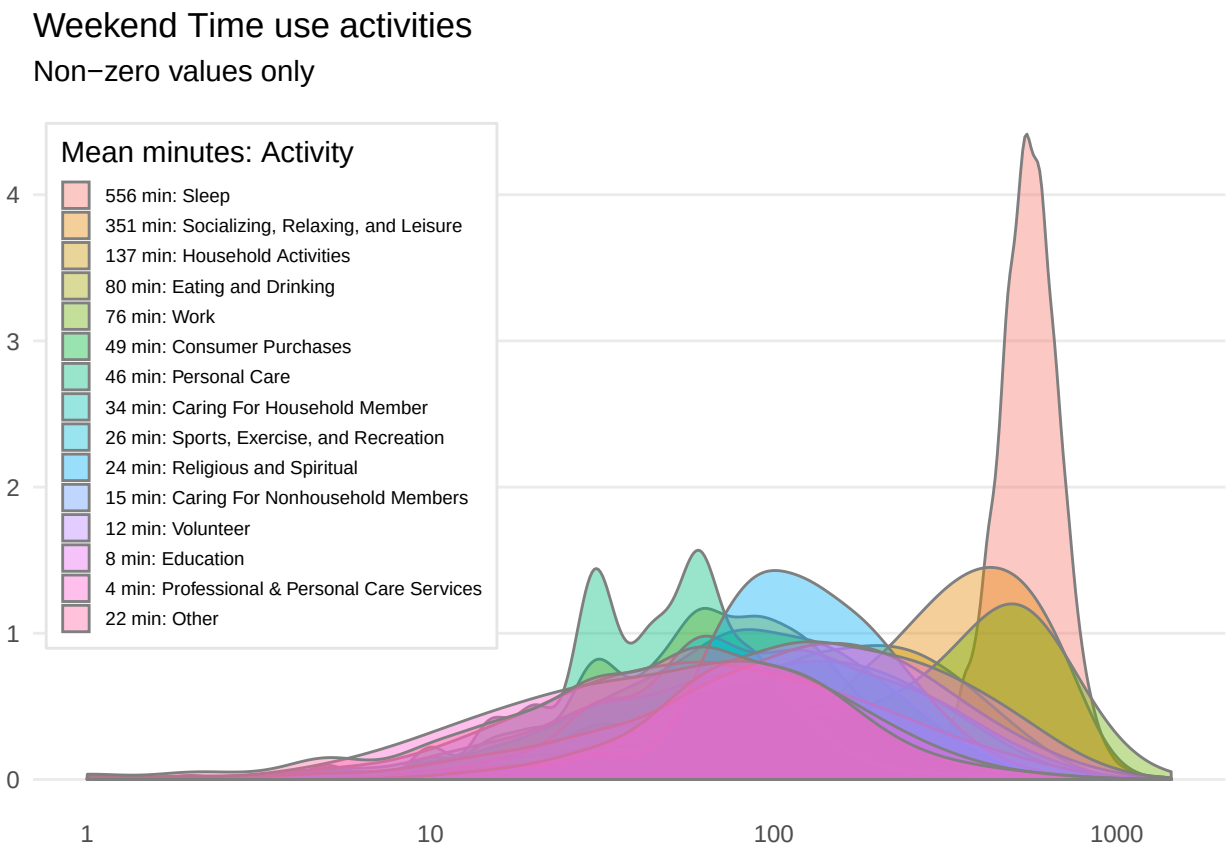
May 27 2020

Contents

Introduction and research question

What are your plans for the weekend? While it is easy to imagine that American's spend their weekends differently, this is a challenging question to empirically validate.

What are your plans for the weekend? Describing the average American's weekend plans in an easy exercise in descriptive statistics. On average, the typical American sleeps 10 hours sleeping, 6 hours socializing and relaxing, 2 hours on household activities and 1 hour eating and drinking. This picture of how the average weekend is spent is shown in the plot below.



While this plot speaks to the population average, it provides little relevant information about how actual individuals spend their time. Moreover, averages and traditional supervised methods are not designed to identify patterns of how different groupings of Americans fill their weekends.

The idea that different people spend their weekends in different ways is unsurprising but it is hard to quantify. Unsupervised machine learning, also referred to as cluster analysis, or clustering, offers a solution to these types of unstructured problems. The method applies distance metrics and linear algebra to recognize patterns and can quantify if different groups of Amerians spend their weekends in different ways.

Data on time use was obtained from the American Time Use Survey (ATUS), which measures how Americans spend their time doing various activities such as work, household activities, volunteering, leisure, and socializing. The ATUS is the most comprehensive measure of time use and obtains a representative sample of how Americans spend their day.

Key findings

Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

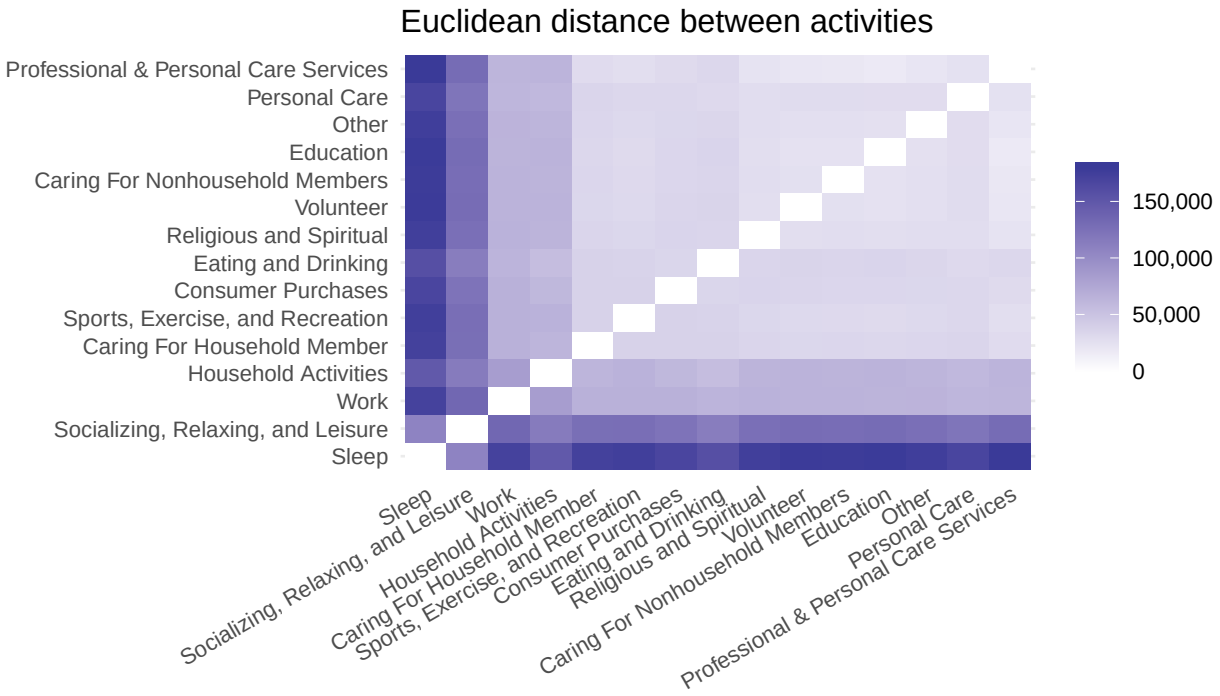
key points on data not yet mentioned - weekend only - going to examine a weighted samplen of 25k

Analysis

Distance

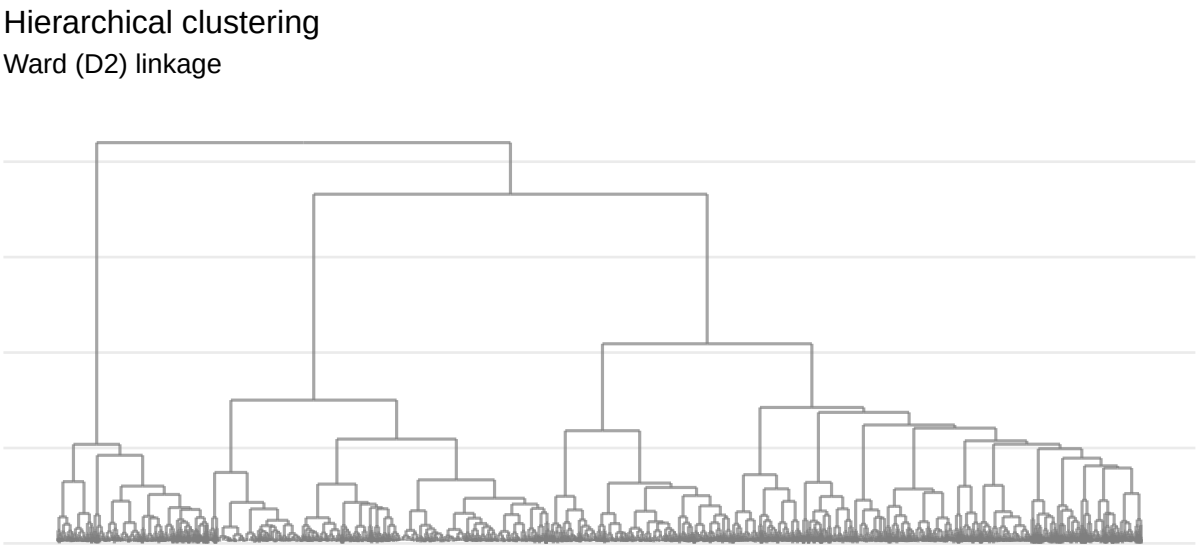
Clustering algorithms rely on calculating measures of distance between data points and variables. There are several established metrics such as the euclidean, manhattan, Jacquard and many others. As a starting point, it is useful to examine the distances between variables to understand if an underlying relationship exists.

The distance matrix confirms some simple assumptions. For example, those who work more on the weekend spend less time sleeping. Similarly, those who work more on weekends have less social and leisure time. While these simple comparisons confirm the assumption that patterns exist they do not reveal any underlying complex structure.



Clustering

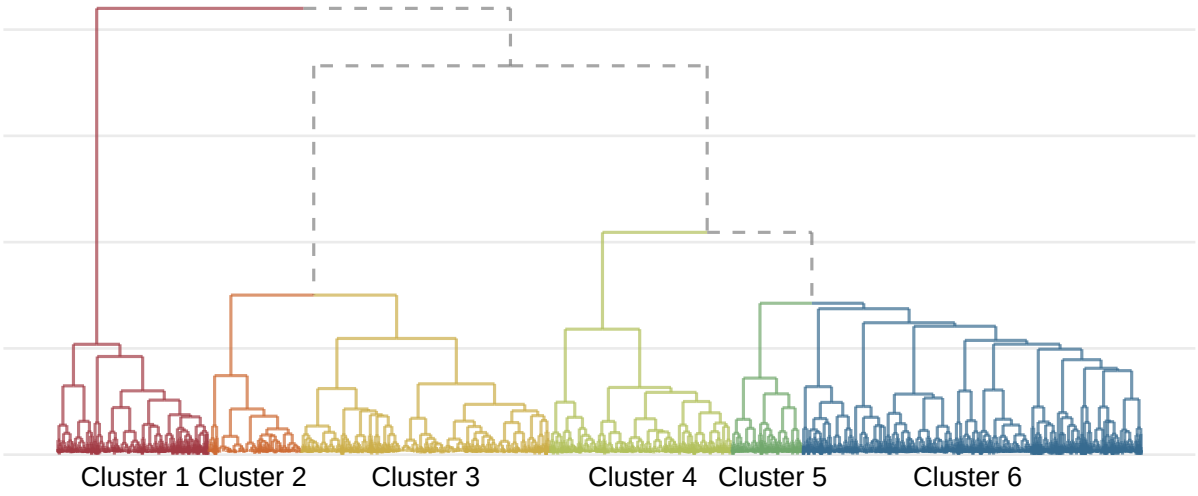
Hierarchical clustering extends the use of distance metrics to understand the similarities and differences between data points (i.e. the people), rather than the distance between variables. A powerful way to visualize this distance is by plotting a dendrogram of the clusters. Hierarchical clustering groups observations by iteratively reducing the number of clusters by combining data points. Dendrograms are a way of visualizing these relationships and identifying existing groupings. In this context, branches that are closer to one another represent people with more similar weekends.



Choosing number of clusters

Identifying the final number of clusters is a nuanced and subjective exercise. That being said, mathematical and quantitative guidelines exist to aid this decision. To determine the number of groups the Calinski-Harabasz index, Silhouette width and Gap-statistic were consulted.

These metrics, combined with the visualization provided by the dendrogram suggest that American's weekend activities can be divided into 6 distinct clusters. Of the 6 final clusters, 5 represent distinct groupings while the 6th appears to be more general.



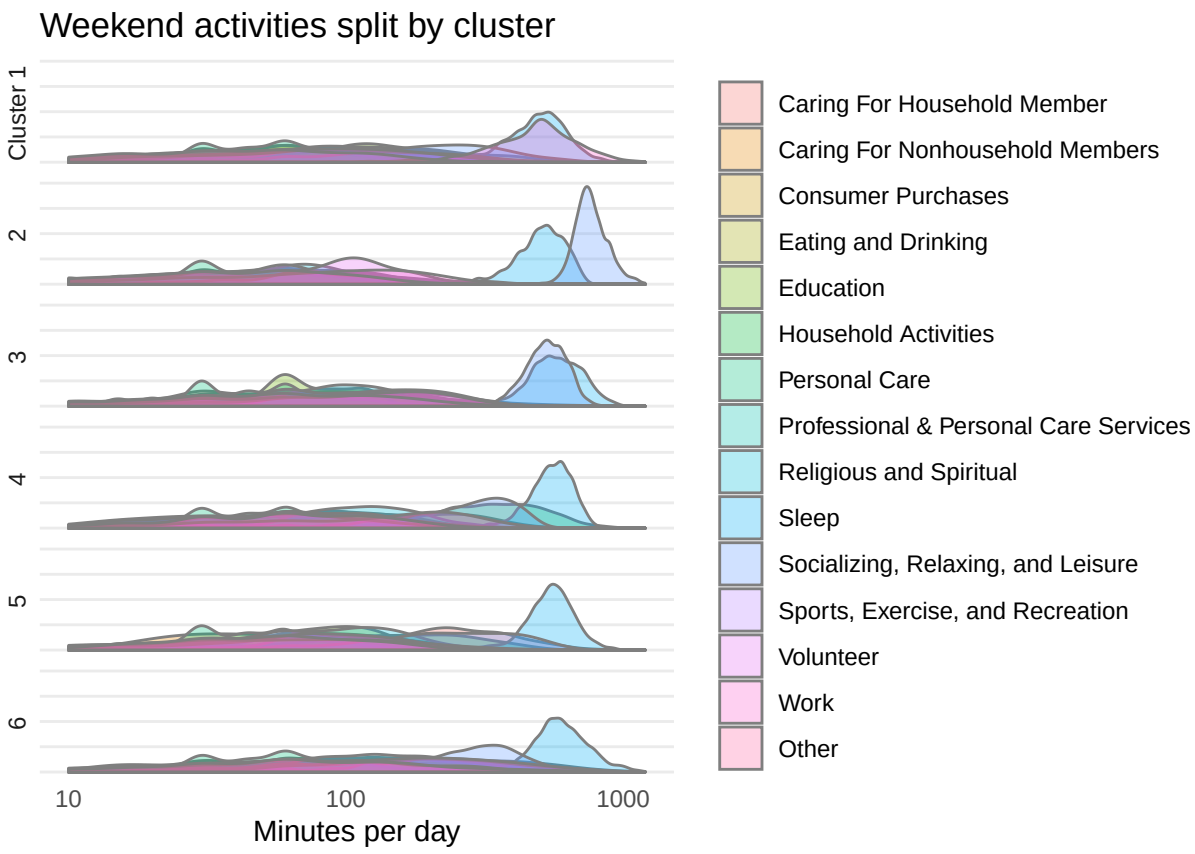
Hierarchical clustering is only one of many clustering methods. We considered the popular K-means algorithm as well as model based clustering. A hierarchical approach offered the best fit to this particular data set. Within the time use survey, many respondents indicate 0 minutes spent on several activities. The K-means algorithm assumes all variables have an equal variance and does not handle zero inflated variables well. These assumptions do not fit the specifics of time use. Simmialty, model based clustering has more rigid assumptions that variables come from a multivariate normal data generating process. Hierarchical clustering avoids these assumptions and is why it is a good fit for this particular clustering problem.

Frequently, variables are scaled prior to cluster analysis. In this particular context, variables are already recorded on the same scale so no transformations were included and variables were left in their original non-transformed format.

Breaking down the clusters

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

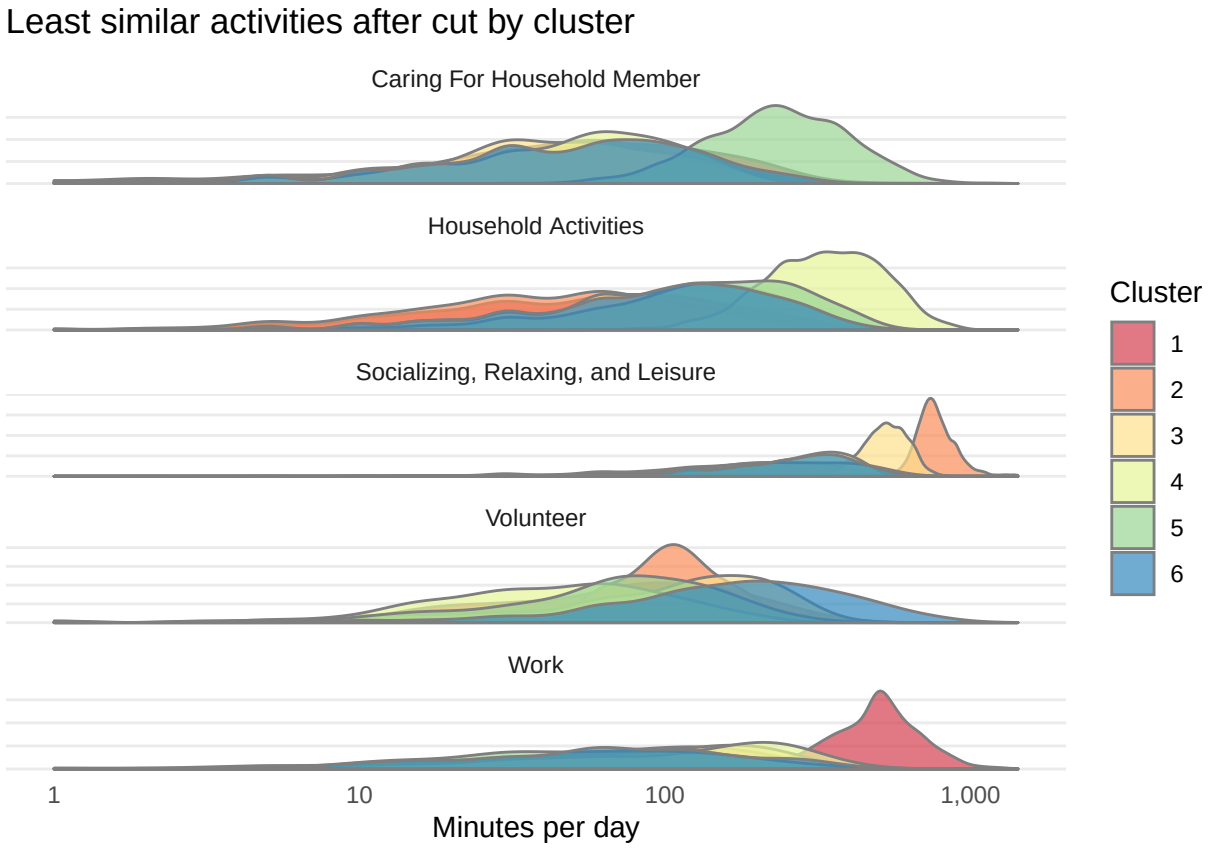
- Cluster 1:
- Cluster 2
- Cluster 3
- Cluster 4
- Cluster 5
- Cluster 6: Catch all ‘Other’



Zoom on on specific activities

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Discuss how this cluster split this activities the most

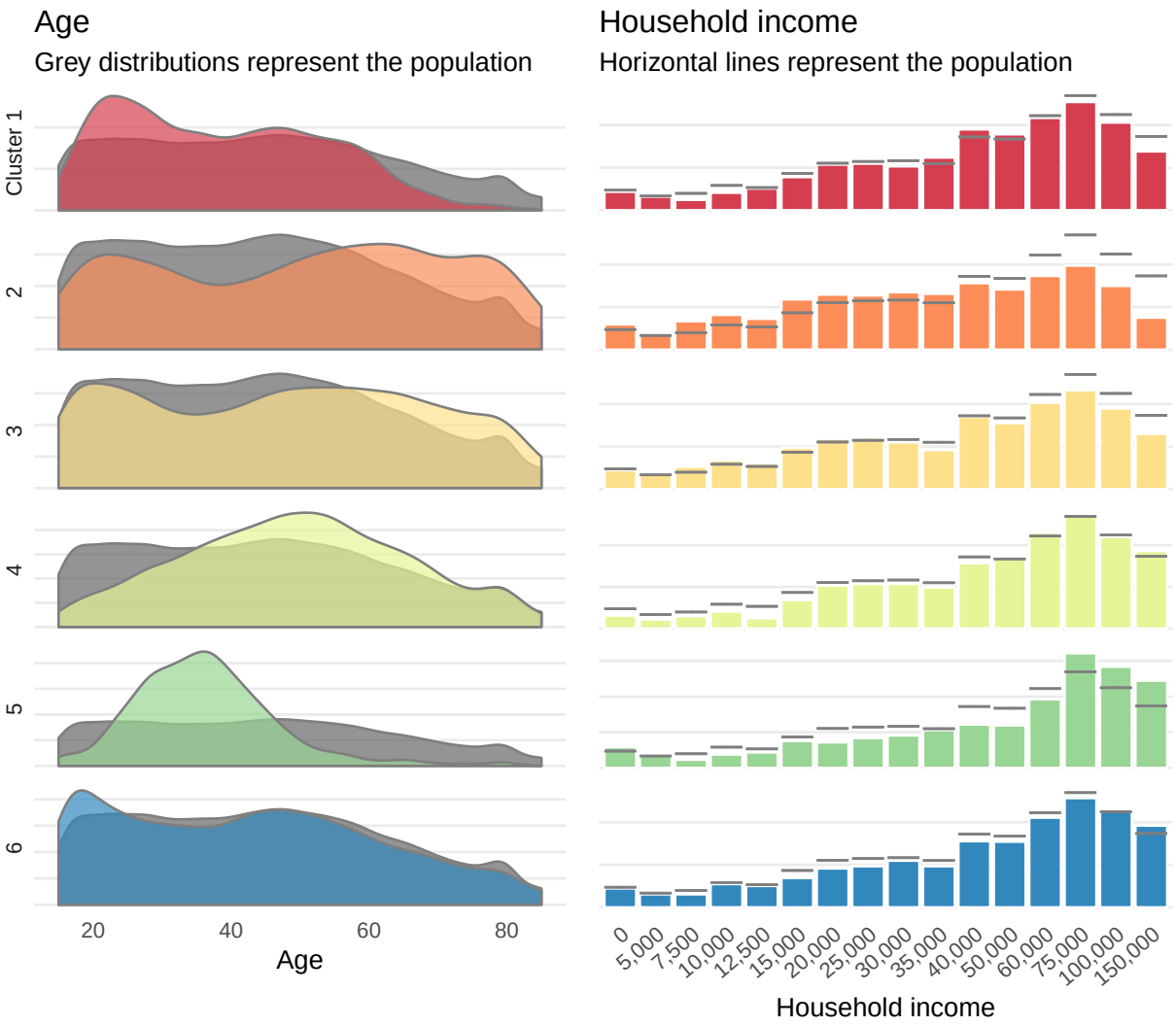


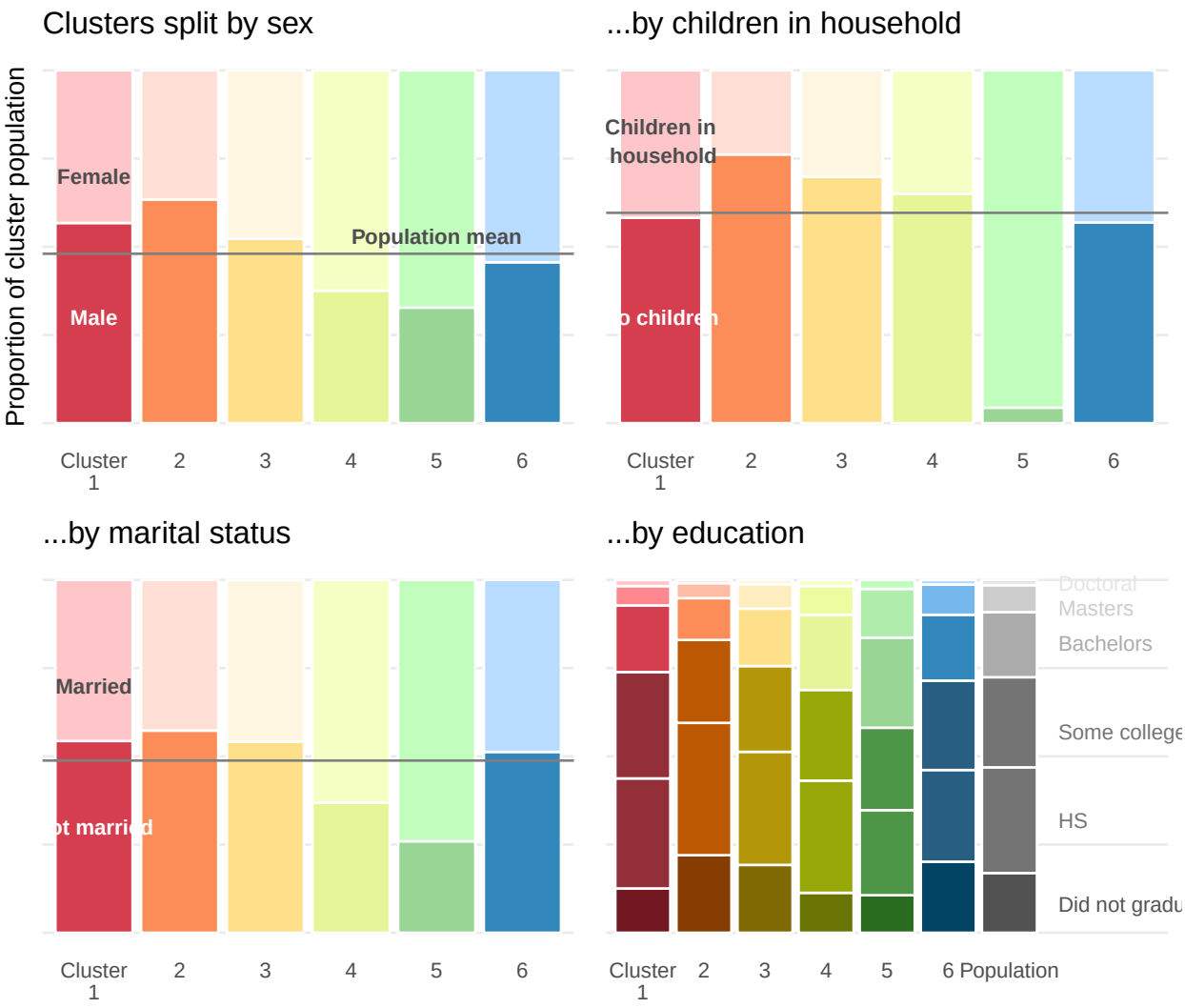
Discuss demographics

Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

discuss how some of the cluster difference are associated with different demographics

Age, income, married, child, educate, gender


Age
Grey distributions represent the population
Household income
Horizontal lines represent the population
Cluster 1
2
3
4
5
6
20
40
60
80
Age
0
5,000
7,500
10,000
12,500
15,000
20,000
25,000
30,000
35,000
40,000
50,000
60,000
75,000
100,000
150,000
Household income



Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.