Fairness in Clustering:
A Socially Conscious Approach to Data Science

Armira Nance, Brie Sloves, Sophie Boileau, Muno Siyakurima, Jeremiah Mensah, Victor Huang, Avery Hall
Advised by Layla Oesper, Professor of Computer Science, Carleton College

A Brief Breakdown of K-means Clustering

Clustering is used in data science to group similar data points. K-means clustering is an offshoot of clustering that groups similar data points into k clusters using arbitrarily determined data points as "centroids." This method of clustering recalculates with each iteration in order to determine the most accurate centers around which groups of data points are clustered. As a group we aim to explore how fairly it clusters data points' protected attributes, though I worked most extensively with the data. There was a lot to consider with preparing and analyzing the data.

Data Wrangling: R/RStudio vs Python

My focus on this project was the data handling, which meant deciding which tool to use for data cleaning and analysis. There were a number of deterministic factors in making this decision.

- We were already using Python for the implementation.
- Operations such as find/replace, filtering out values, and looking at the frequency of specific data attributes were significantly quicker using R
- R allows us to look at data as we're performing operations on it
- R has quicker and more intuitive plotting power
- It is easier to correct mistakes made in R, making it a tool that's easy to experiment with

Re-working the Data

We pulled data from the National Center for Education Statistics' High School Longitudinal Study (HSLS) which gathered education data for 9th grade students from different schools across the country and kept track of them over the course of 7 years (2009 - 2016). In order to prepare the data, we had to do some housekeeping, such as

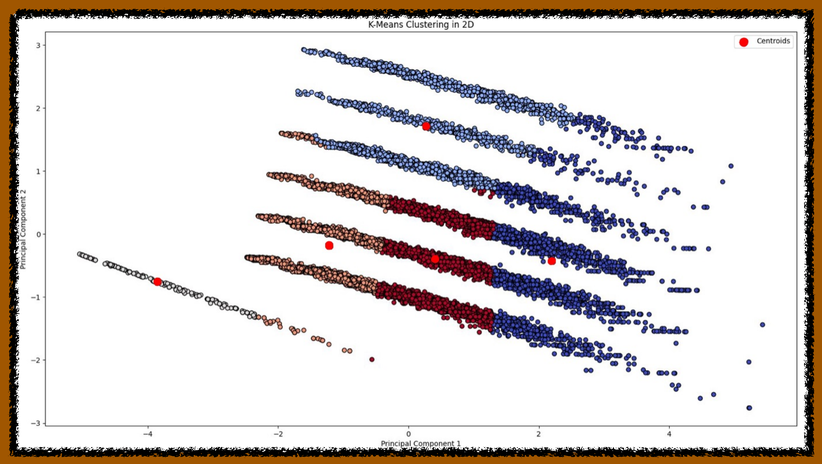
- Filtering out suppressed data
- Filtering out missing values
- Column conversions from strings to integers
- Standardizing numerical data
- Combining relevant data for clarity
- Encoding categorical data with a socially-conscious approach

Clustering the Data

We were able to successfully implement the basic k-means clustering algorithm and visualize the clusters. We clustered based on:

- Highest parent's level of education
- Income per person in the household
- Hours of extracurricular activity during a typical week

The algorithm produced the five color-coded clusters found below with centroids located at the positions marked by a red dot.



Analysis and Conclusion

The data revealed that, while the k-means algorithm did choose logical centers for clustered data points, the clustering was not fair. Across the five clusters, we examined race as a protected attribute among the data points, using prevalence within the population to determine fair representation within the clusters. The racial distribution of the full dataset gives us a basis for comparison.

Amer. Indian/Alaska Native, non-Hispanic	0.71
Asian, non-Hispanic	7.58
Black/African-American, non-Hispanic	9.65
Hispanic, no race specified	0.75
Hispanic, race specified	15.16
Missing	0.02
More than one race, non-Hispanic	8.64
Native Hawaiian/Pacific Islander, non-Hispanic	0.49
White, non-Hispanic	57.01

Amer. Indian/Alaska Native, non-Hispanic	0.57
Asian, non-Hispanic	11.38
Black/African-American, non-Hispanic	8.07
Hispanic, no race specified	0.21
Hispanic, race specified	9.96
More than one race, non-Hispanic	8.14
Native Hawaiian/Pacific Islander, non-Hispanic	0.57
White, non-Hispanic	61.09

Racial Distribution of Cluster 0

Amer. Indian/Alaska Native, non-Hispanic	0.28
Asian, non-Hispanic	11.62
Black/African-American, non-Hispanic	5.03
Hispanic, no race specified	0.05
Hispanic, race specified	7.28
More than one race, non-Hispanic	6.59
Native Hawaiian/Pacific Islander, non-Hispanic	0.14
White, non-Hispanic	69.02

Racial Distribution of Cluster 2

Amer. Indian/Alaska Native, non-Hispanic	0.63
Asian, non-Hispanic	4.11
Black/African-American, non-Hispanic	12.20
Hispanic, no race specified	0.31
Hispanic, race specified	13.96
More than one race, non-Hispanic	10.52
Native Hawaiian/Pacific Islander, non-Hispanic	0.74
White, non-Hispanic	57.53

Racial Distribution of Cluster 3

Amer. Indian/Alaska Native, non-Hispanic	1.00
Asian, non-Hispanic	4.94
Black/African-American, non-Hispanic	11.13
Hispanic, no race specified	1.30
Hispanic, race specified	21.61
More than one race, non-Hispanic	8.93
Native Hawaiian/Pacific Islander, non-Hispanic	0.47
White, non-Hispanic	50.62

Racial Distribution of Cluster 4

References and Acknowledgments

1. Chierichetti, Flavio, et al. "Fair clustering through fairlets." Advances in neural information processing systems 30 (2017).
2. Ghadiri, Mehrdad, Samira Samadi, and Santosh Vempala. "Socially fair k-means clustering." Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021.

Thank you to Brie, Muno, Sophie, Avery, Victor, Jeremiah for persevering to the end of comps and thank you to Layla for advising us.