

Project 5: Census Data Analysis



By: Aaron Dzaboff

1.

Problem Formulation

What Are We Solving For?



Develop a clustering model to help synthesize census data



Goal: Derive prototypical examples of “average Americans”



2.

Data Mining

- What sources were used
 - What was collected?

What was Mined?

Sources:



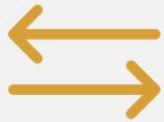
1990 US Census Data from UCI Machine Learning Repository

What was Collected?

- Comprised of census data of 2.5 million individuals represented by 68 categorical variables
- Data Collected Includes:
 - Income information
 - Citizen status
 - Age
 - Worker class
 - Disability information
 - Language proficiency
 - Work/Job information
 - Military service
 - Martial status
 - Place of birth
 - Weight
 - Family information
 - Education background
 - Gender
 - Transportation information

3. Data Preprocessing/Cleaning

Methods to preprocess data



The variables were mapped using SQL to make them into categorical variables



One hot encoded each of my variables to convert categorical data to integer data



Created a feature vector of the one hot encoded variables to fit the clustering methods



4.

Feature Selection

Correlation Analysis

- ▷ Performed a correlation analysis on all 68 variables
 - Wanted to identify variables that were highly correlated to justify reducing the number of variables in the clustering algorithm
- ▷ Identified variables which were highly correlated (.80 or greater)
 - Chose only one variable to represent all the other variables that it was highly correlated with
 - There were a few exceptions to this
 - Ex: place of birth and citizen
- ▷ Chose variables that tried to split people into distinct groups in order to best derive prototypical average American results



Features Selected From Correlation Analysis



CITIZEN
STATUS



WORKER
CLASS



WORK
DISABILITY
LIMITATION



ENGLISH
PROFICIENCY



OF
CHILDREN



AGE



OTHER
LANGUAGES
SPOKEN



MARITAL
STATUS



MEANS OF
TRANSPORTATION



MILITARY
STATUS



PLACE OF
BIRTH



POVERTY
STATUS



WEIGHT



PERSONAL
INCOME



SEX



TEMPORARY
WORK
ABSENCE



TRAVEL
TIME TO
WORK



WORKED IN
1989
STATUS



EDUCATIONAL
BACKGROUND

- Overall, I chose 19 variables that attempt to best place people into distinct groups



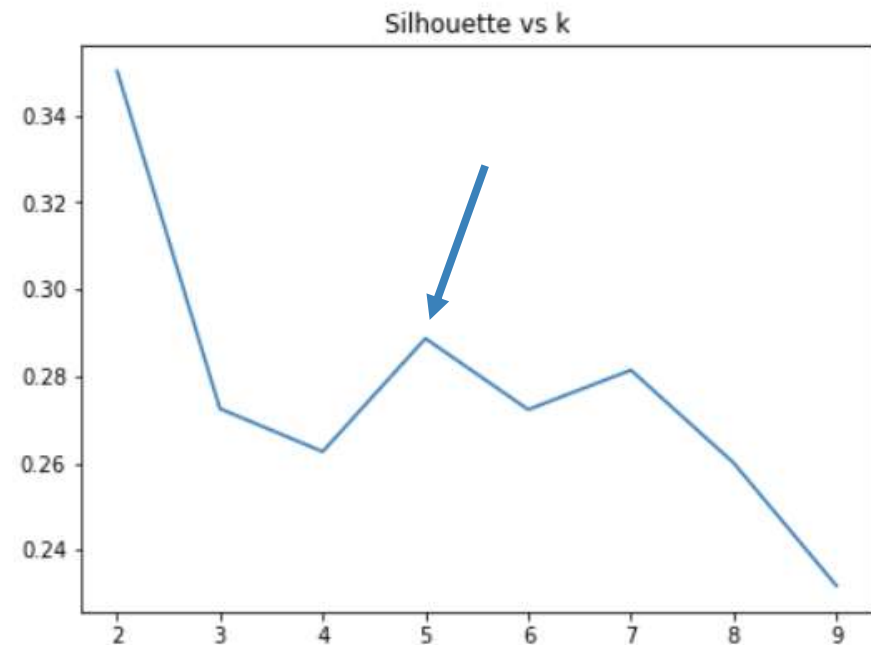
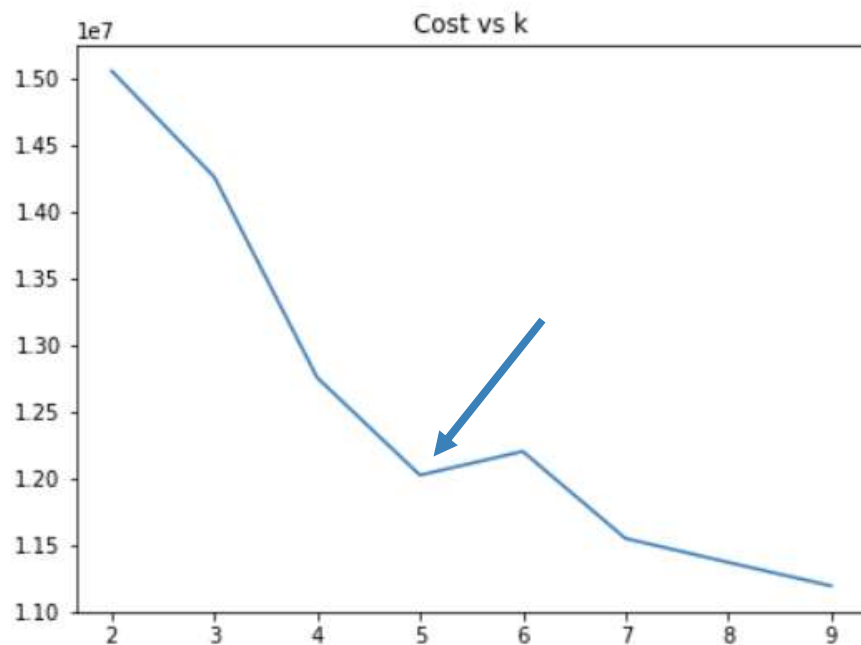
4.

Clustering Analysis

- Clustering Methods
- Justification for number of clusters
- Prototypical American Examples

KMeans Clustering

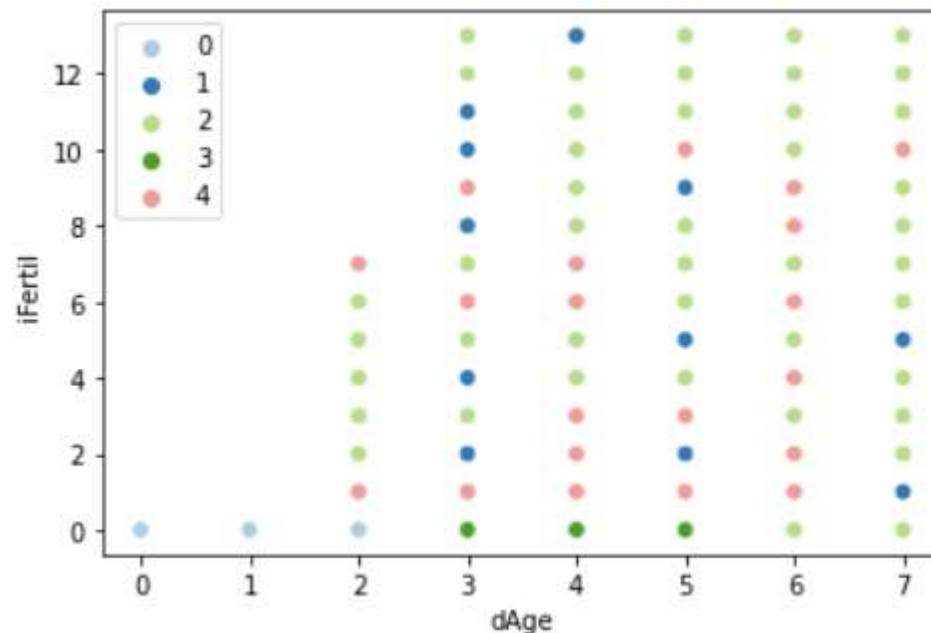
- ▷ Based on the Elbow method, the optimal number of clusters is 5



- ▷ This also coincides with the highest silhouette score. With only a .2886 silhouette score, there is some overlapping b/w clusters.

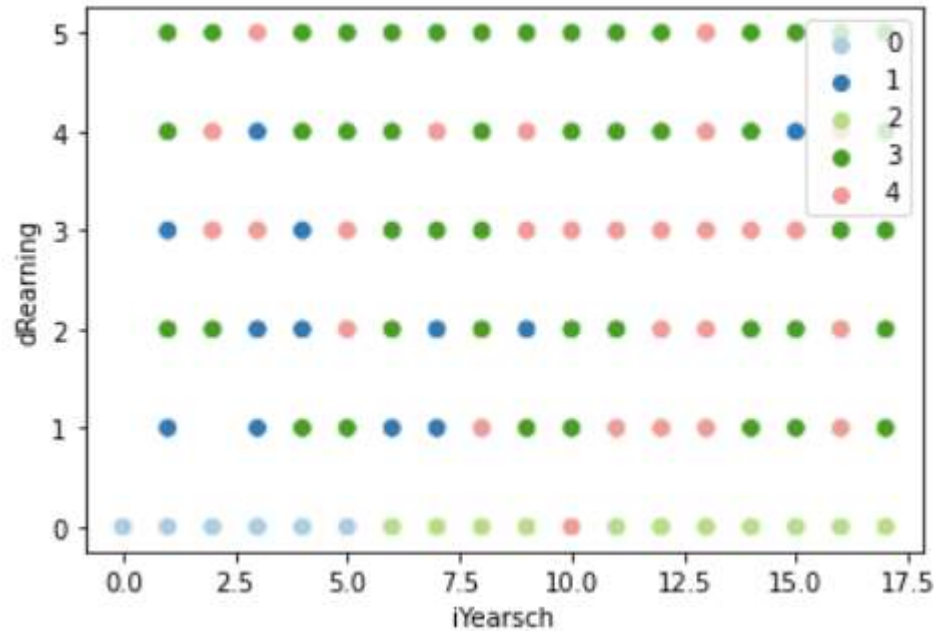
Variables that Stratified the Clusters

Number of Children and Age



- ▷ Can see that the people in cluster 0 are younger with no children. Also, people in cluster 3 tend to be middle aged people with no children

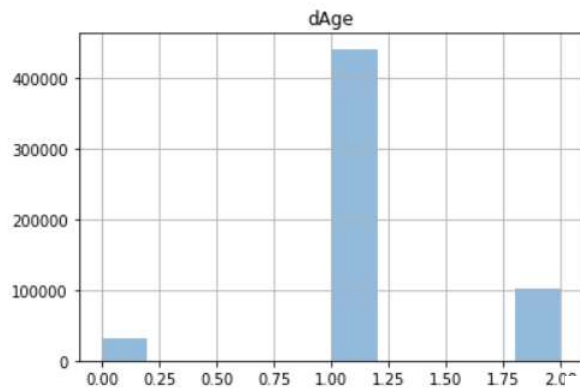
Income and Education



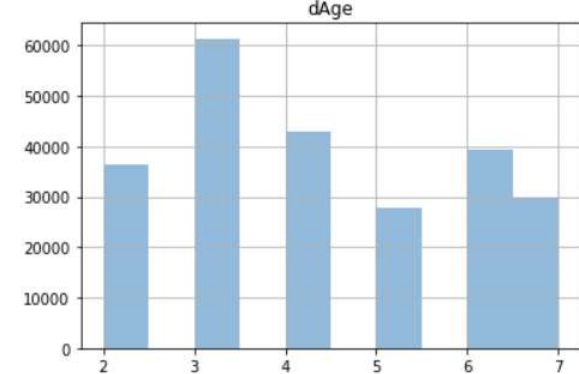
- ▷ People in clusters 3 and 4 tend to have more education and have higher income

Age Distribution

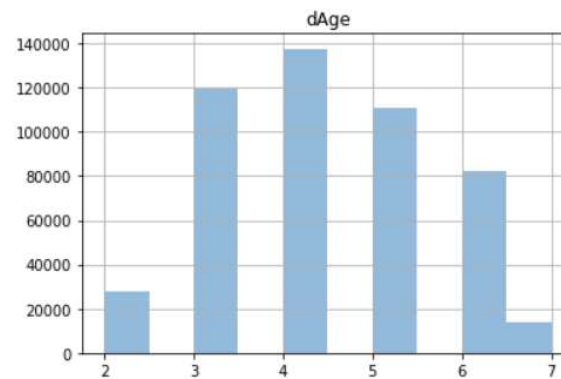
Cluster 0:



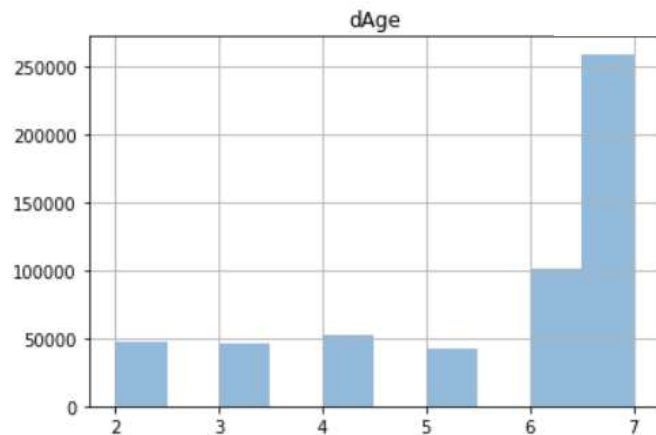
Cluster 1:



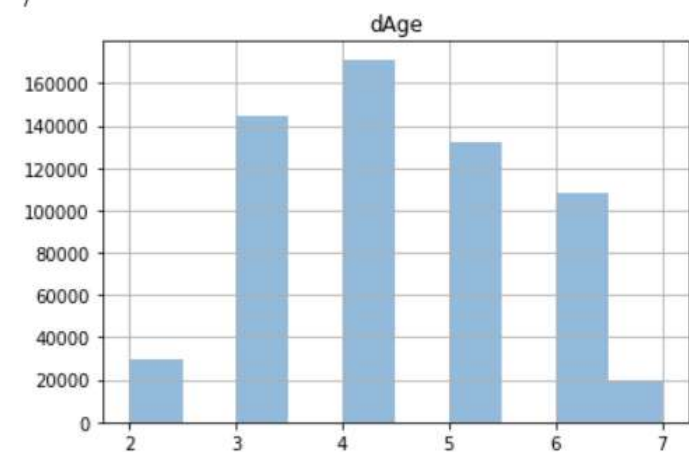
Cluster 4:



Cluster 2:



Cluster 3:



Cluster 0 Prototype: Teenagers who live with their parents

These are young people (less than 20) with no children, little income, and the majority aren't in the workforce.

Summary Statistics:

	iCitizen	iClass	iDisabl1	iEnglish	iFertil	dAge	iLang1	iMarital	iMeans	iMilitary	dPOB	dPoverty	dPwgt1	dRearning	iSex	iTmpabsnt	dTravtime	iWork89	iYearsch
prediction																			
0	0.131005	0.000000	0.000000	0.138906	0.028778	1.119947	1.271771	3.998386	0.000000	0.000000	0.138005	1.786332	1.095614	0.000000	0.487858	0.000000	0.000000	0.000000	2.985326

Cluster 1 Prototype: Middle class workers

These people are roughly middle aged and younger. They have less years in school and relatively less income.

Summary Statistics:

	iCitizen	iClass	iDisabl1	iEnglish	iFertil	dAge	iLang1	iMarital	iMeans	iMilitary	dPOB	dPoverty	dPwgt1	dRearning	iSex	iTmpabsnt	dTravtime	iWork89	iYearsch
prediction																			
1	0.306247	2.092894	1.866642	0.226392	1.470664	4.259263	1.867071	1.676365	0.885809	3.696622	0.301645	1.742205	1.123104	2.256054	0.529673	2.217584	0.000000	1.000597	9.864808

Cluster 2 Prototype: Retired people

These people make up the oldest cluster, there are more widows, served time in the military, and are the second lowest earning group.

Summary Statistics:

	iCitizen	iClass	iDisabl1	iEnglish	iFertil	dAge	iLang1	iMarital	iMeans	iMilitary	dPOB	dPoverty	dPwgt1	dRearning	iSex	iTmpabsnt	dTravtime	iWork89	iYearsch
prediction																			
2	0.365935	0.539719	1.683315	0.284936	2.282567	5.605569	1.848047	1.107139	0.055086	3.721451	0.340714	1.718288	1.094472	0.000657	0.671302	2.593181	0.000930	1.996781	8.692499

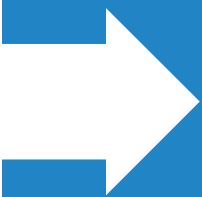
Cluster 3 Prototype: Middle aged people without kids

Roughly the same age as clusters 1 and 4. These people are the highest earners, they are highly educated, and they don't have children

Summary Statistics:

	iCitizen	iClass	iDisabl1	iEnglish	iFertil	dAge	iLang1	iMarital	iMeans	iMilitary	dPOB	dPoverty	dPwgt1	dRearning	iSex	iTmpabsnt	dTravtime	iWork89	iYearsch
prediction																			
3	0.352554	2.162364	1.956531	0.224244	0.000000	4.337167	1.868646	1.183215	1.743269	3.367175	0.347706	1.927095	1.148833	3.104361	0.000000	0.000000	3.435581	1.020954	10.833830

Cluster 4 Prototype: Middle aged people with children



Roughly the same age as clusters 1 and 3, second highest earners, they have larger families, are highly educated, and spend the most time driving to work.

Summary Statistics:

	iCitizen	iClass	iDisabl1	iEnglish	iFertil	dAge	iLang1	iMarital	iMeans	iMilitary	dPOB	dPoverty	dPwgt1	dRearning	iSex	iTmpabsnt	dTravtime	iWork89	iYearsch
prediction																			
4	0.296595	1.904189	1.969162	0.188257	2.612279	4.289386	1.881561	1.314202	1.651902	3.965038	0.291318	1.921674	1.146474	2.526427	1.000000	0.000000	3.188505	1.032344	10.929245

5. Limitations

How Can This Project be Improved?

- ▷ Also tried hierarchical clustering, but did not produce any better results
- ▷ By using KMeans, the solution is a local optimum, and not a global
- ▷ KMeans is very dependent on the initial values chosen. As increases, you need advanced versions of k-means to pick better values of the initial centroids
- ▷ Even though the dataset was trimmed to 19 variables, it is still highly dimensional which causes a distance-based similarity measure, like KMeans, to converge to a constant value between any given examples.



Thanks!

Any questions?