Project 5, U.S. Census- Understanding American Demographics

Elliott Newman: Big Data Analytics:
ISE:4172

Report Outline

Problem Statement

Data Preparation

Feature Selection

Exploratory Analysis

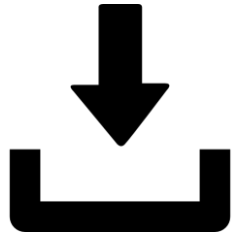
Cluster Modeling

Cluster Analysis

Problem Statement

- Every 10 years, the U.S. Census Bureau collects census data. This data is used to help policy makers evaluate the needs of their citizens and provide support to their communities. Billions of dollars go into annual resource funding.
- As a government data analyst, my job is to use clustering techniques to better understand Americans, based on their demographic traits, and cluster them into groups.

Data Description and Understanding

The UCI Machine Learning repository provides a 1990 U.S. Census participant dataset with 2.5 million records and 68 attributes.	 The data can be downloaded as a .txt file, with an accompanying data description.	 Some of the original data are mapped to new values.	 The UCI webpage offers T-SQL code identifying how certain attributes are created.

Main Data Mining Techniques for Data Collection

Loading in this data required a straight-forward process. Here is the programming procedure I used:

1. Saved the UCI '…Census.data.txt' url as a string variable.
2. Referencing the string variable, I used the '!wget' command to save a Census data .txt file to PySpark.
3. Read in the .txt file as a Spark Dataframe
4. Analyzed the data to select primary variables for analysis

- In the next slides, I describe the variable selection process.

Variable Selection

- To help myself get a better grasp of the dataset, I organized the attributes into 9 different categories. These categories help reflect variables that have impacted American policy making decisions in the past, and my own intuition of the data.
- These categories also helped me better visualize overlap between attributes and reduce dimensionality.
- The Marital and Spouse variables, for example, were similar, which I discovered through this categorization process.

Category	Examples of Attributes
Income	dIncome1-10, Earning, PIncome
Nationality	Citizen, Immigrant, Place-of-Birth
Military Status	Veteran Status, Active, Wars Served In
Age	Standardized Age Code
Employment	Employment status, Hours
Education	Education level, Languages Spoken
Family	Marital status, Relatives, Numb Children
Health	Weight, Disabilities, Sex
Misc	Size of Car

Variable Selection with Correlation Matrix

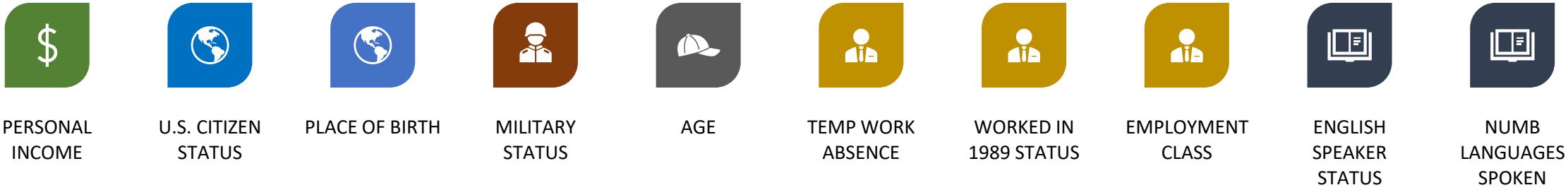
Sample of Variables on a Correlation Matrix

		dAge	iAvail	iCitizen	iDisabl1	iDisabl2	iEnglish	iFertil	dHispan	dHour8	iImmigr	dIncome	dIncome	dIncome	dIncome	dIncome	dIncome	d
2	dAge	1	-0.00574	0.05135	0.620457	0.643747	0.034701	0.412432	-0.02941	0.227468	0.129788	0.229731	0.108129	0.066955	0.339459	0.545814	0.105806	
3	iAvail	-0.00574	1	0.020889	0.099289	0.11016	0.028293	0.003883	0.011366	0.036484	0.008214	-0.00186	-0.00823	-0.00999	-0.03366	-0.04318	0.054675	
4	iCitizen	0.05135	0.020889	1	0.097313	0.093549	0.624663	0.037499	0.227487	0.030651	0.788207	0.01093	0.00139	-0.02047	-0.03203	-0.02714	0.019945	
5	iDisabl1	0.620457	0.099289	0.097313	1	0.977813	0.061324	0.298796	-0.0104	0.560735	0.107991	0.508482	0.133678	0.058604	0.252286	0.092953	0.023268	
6	iDisabl2	0.643747	0.11016	0.093549	0.977813	1	0.059268	0.302009	-0.01211	0.56681	0.106855	0.51027	0.140446	0.062528	0.260225	0.105773	0.033147	
7	iEnglish	0.034701	0.028293	0.624663	0.061324	0.059268	1	0.050559	0.265614	-0.00702	0.43953	-0.03709	-0.01315	-0.01742	-0.06409	-0.02068	0.054452	
8	iFertil	0.412432	0.003883	0.037499	0.298796	0.302009	0.050559	1	-0.00544	-0.06014	0.06484	-0.06982	-0.02065	-0.03369	-0.01832	0.185155	0.142863	
9	dHispanic	-0.02941	0.011366	0.227487	-0.0104	-0.01211	0.265614	-0.00544	1	-0.01426	0.18843	-0.02077	-0.01132	-0.01254	-0.04681	-0.03123	0.018646	
10	dHour89	0.227468	0.036484	0.030651	0.560735	0.56681	-0.00702	-0.06014	-0.01426	1	0.031516	0.775405	0.212482	0.118556	0.198039	-0.26265	-0.10352	
11	iImmigr	0.129788	0.008214	0.788207	0.107991	0.106855	0.43953	0.06484	0.18843	0.031516	1	0.028802	0.010848	-0.01554	0.015283	0.043165	0.029176	
12	dIncome1	0.229731	-0.00186	0.01093	0.508482	0.51027	-0.03709	-0.06982	-0.02077	0.775405	0.028802	1	-0.00938	0.015604	0.253647	-0.24761	-0.10459	
13	dIncome2	0.108129	-0.00823	0.00139	0.133678	0.140446	-0.01315	-0.02065	-0.01132	0.212482	0.010848	-0.00938	1	0.065457	0.110334	-0.03333	-0.0254	
14	dIncome3	0.066955	-0.00999	-0.02047	0.058604	0.062528	-0.01742	-0.03369	-0.01254	0.118556	-0.01554	0.015604	0.065457	1	0.070575	0.008687	-0.00704	
15	dIncome4	0.339459	-0.03366	-0.03203	0.252286	0.260225	-0.06409	-0.01832	-0.04681	0.198039	0.015283	0.253647	0.110334	0.070575	1	0.211127	-0.05575	
16	dIncome5	0.545814	-0.04318	-0.02714	0.092953	0.105773	-0.02068	0.185155	-0.03123	-0.26265	0.043165	-0.24761	-0.03333	0.008687	0.211127	1	0.053067	
17	dIncome6	0.105806	0.054675	0.019945	0.023268	0.033147	0.054452	0.142863	0.018646	-0.10352	0.029176	-0.10459	-0.0254	-0.00704	-0.05575	0.053067	1	

Note: Not all features are standardized in the same way/order. They are all considered categorical. My correlations were usually more "accurate" towards features ordered similarly. I checked the feature descriptions to ensure accuracy.

- Splitting the features into categories helped me reduce the data from 68 columns to 50. I found variables that repeated other ones, or ones that lacked enough documentation from UCI.
- I then used a correlation matrix on the next 50 variables as a statistical approach to variable selection. If two variables highly correlated, I would review their documentation and select one for my final variable list. There is a chance they can identify the same thing about a person.
- For example- some of the disability statuses have > .95 correlations, and don't need to all be included.

Final Variables Selected for Analysis



This author of the dataset was consistent and organized. I did not do any outlier removal.

All variables were saved as integers in the initial dataset. They were described as categorical. For analysis, I kept them as integers. For modeling, I converted them to one-hot encoding.

Data Dimensions:
Approx. 2.46 million rows
19 Features

Summary Statistics

- Feature Statistics

summary	Age	Numb Children	Weight
count	2458285	2458285	2458285
mean	3.8516429136572854	1.1815355013759592	1.1212890287334463
stddev	2.0484916579674715	1.8613088660891126	0.7107215771337748
min	0	0	0
25%	2	0	1
50%	4	0	1
75%	6	2	2
max	7	13	3

summary	Personal Income	dTravTime	Year in School
count	2458285	2458285	2458285
mean	1.8739283687611485	1.4835411679280475	8.446545864291569
stddev	1.4712726651551462	1.966292981803005	4.08057924792266
min	0	0	0
25%	0	0	5
50%	2	0	10
75%	3	3	11
max	5	6	17

Feature Notes

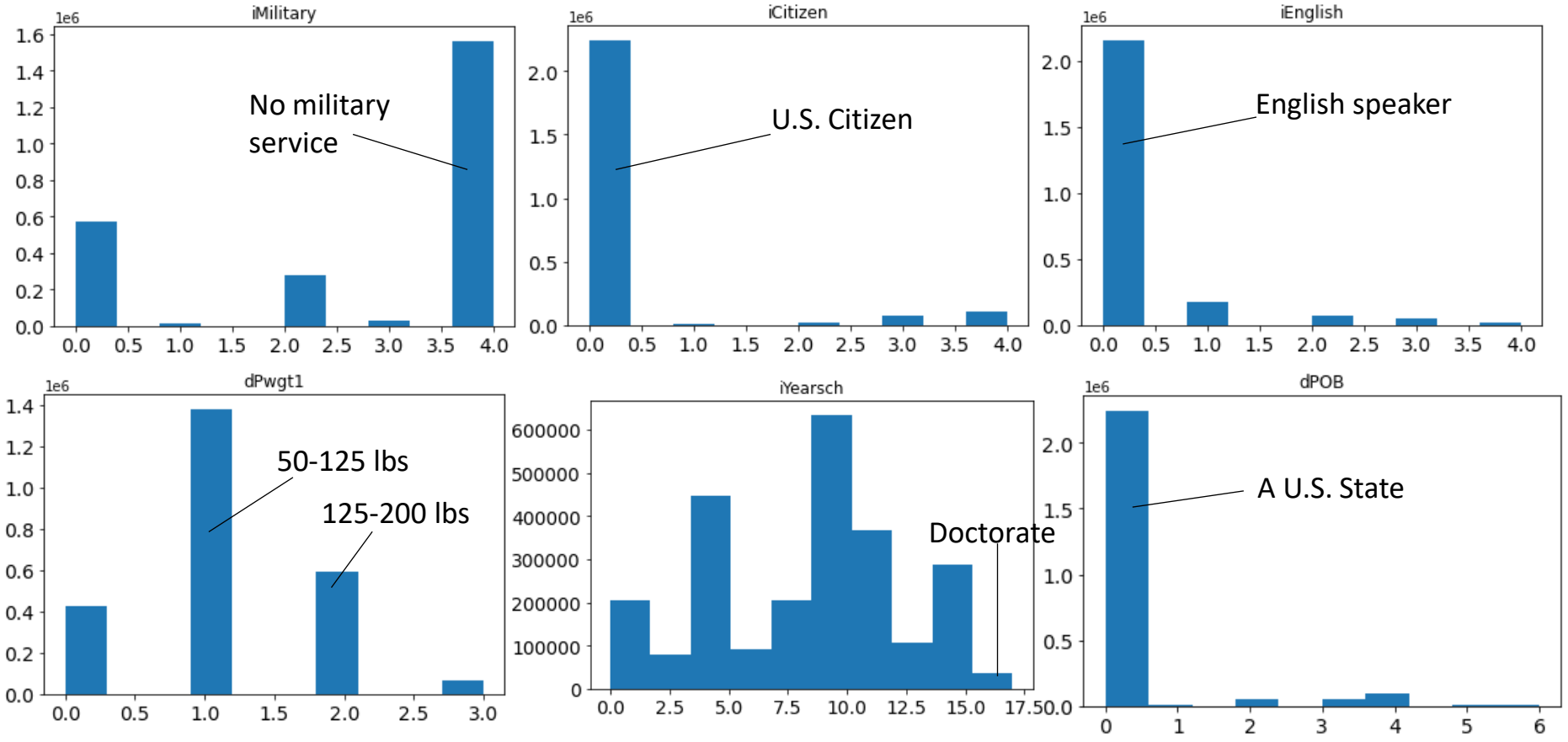
- Many of these variables are standardized based on quantitative variables. I analyzed them like I would a quantitative variable, due to their ordinal/interval like nature.
- Example- The mean for Age is 2.04. The SQL documentation describes anyone labeled as a “2” between 20-30 years old.
- Year in School goes from no school at 0 to a doctorate degree at 17.
- Income is in increments of 15000, with its highest category (5) representing those with an income over $60,000.
- Numb Children is almost exact in showing the number of children. A 13 indicates 13 children or more.

Exploratory Analysis-Histograms

I used histograms to look at count distributions of features in my dataset. On the right are six example histograms.

With more time, I would label them better. At a high level, most Americans:

- Have no military experience
- Are citizens
- Speak English
- Are in a 50-200 pound weight range.
- Have a variety of school levels
- Are born in the U.S.

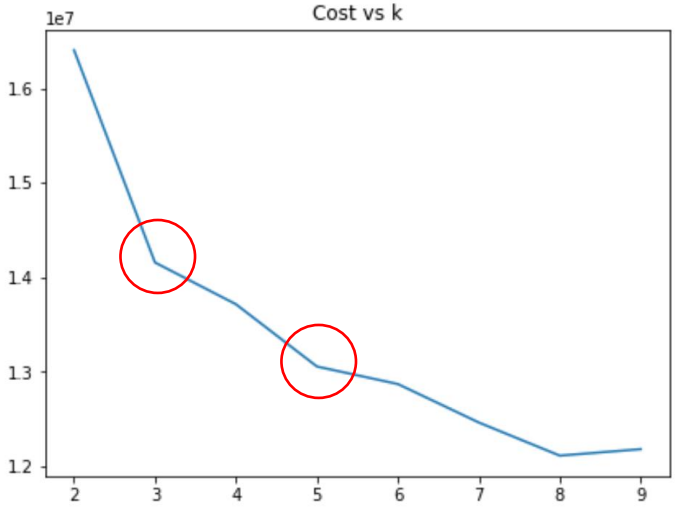


Other notable qualities not shown in the histograms: Most Americans make between 15-60 thousand dollars, a plurality are married, and about half worked last year.

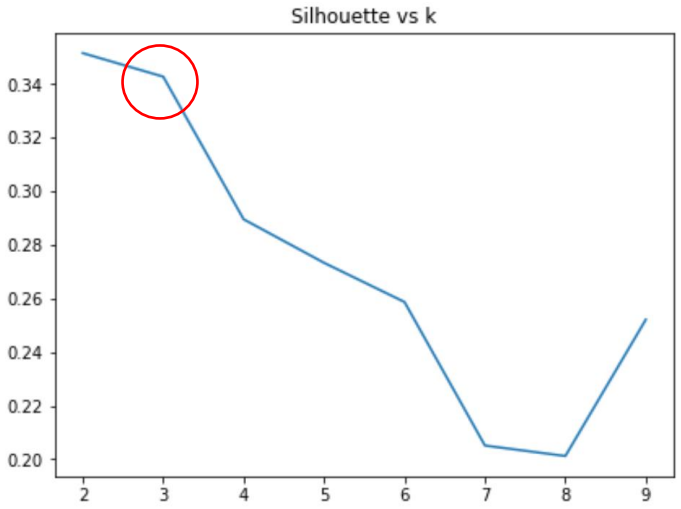
Cluster Number Selection

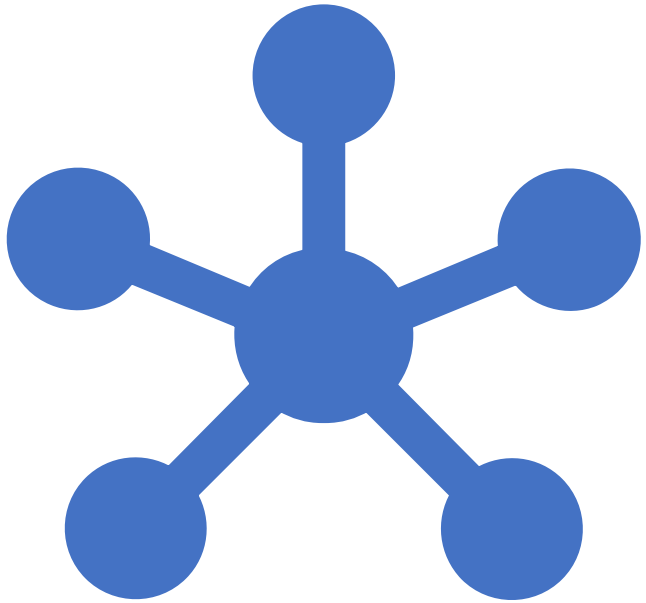
After prepping/exploring the data, I began the unsupervised modeling process. I ran a k-means model using the Spark API. I needed to select the ideal number of K clusters to separate the data into clusters.
Two techniques for selecting k come from sum of squared distance costs, and silhouette scores.

- Elbow Method: Multiple k-means models were ran to gather sum of squared distance cost per cluster sizes 2-9. The goal was to find where the cost began to level off. There were two slight leveling off “elbows”. Cost slightly went up again at 9.



- Silhouette Scores: Silhouettes for k 2 and 3 had the highest scores, with the 8-9 range having a slight increase. The elbow method gives a guide for the number of clusters, while silhouette scores can be a metric for the quality of the cluster size.





Model Selection

- I chose k-means for my final model with 3 clusters.
- 3 showed a slight bend for sum of squared distance cost and had the highest silhouette score. Good silhouette scores are closest to a value of 1.
- 2 clusters seemed too low. Higher cluster numbers might also work, especially with variables likes age/POB having multiple values.
- I looked at clusters beyond 9, and some of the distance costs fluctuated. Overall, silhouette scores beyond 9 went down.
- I ran a Bisecting K-Means Hierarchical clustering model as well. Running 2-10 clusters took 3 hours and shut down my program.
- The hierarchical model silhouette scores I was able to get through 4 models were almost identical. I had to run each one separately. I tried setting up a dendrogram in both Pandas and the Anaconda Orange data mining software.

Final Clusters and Descriptions

Cluster 1- Young Americans, Students: Number of instances- 574,468

- These people are younger. They have up to high school or early college experience and don't have information about work or income. They are most likely healthy, not having disability statuses. They don't have children. Many are listed as attending a college, and few are listed as not being in school.

Cluster 2- Older Americans: Number of instances- 771,305

- This cluster represents older Americans. These Americans have a variety of education levels and may have children. This cluster has the most cases of work disabilities. These Americans may be retired, since they are reported as having lower current income levels. Most of them didn't work last year. There are more women in this group, possibly because women live longer.

Cluster 3- Working Adults: Number of instances- 1,112,512

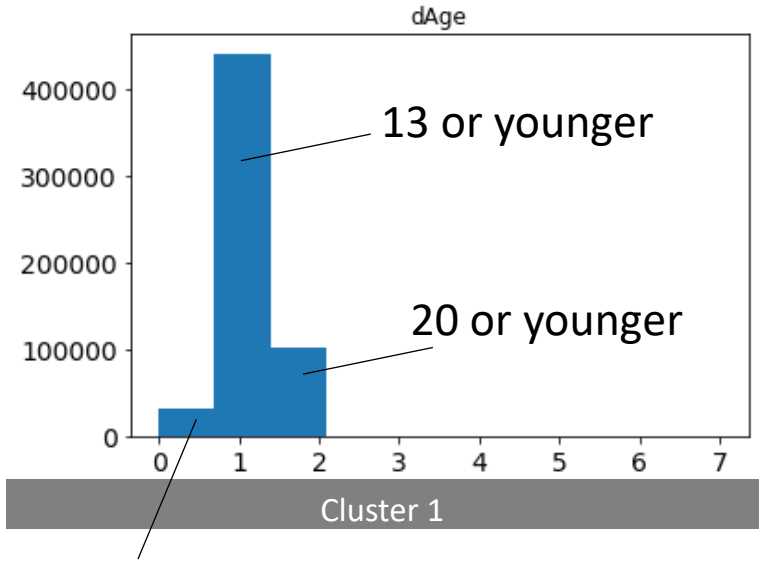
- These are working adults, or parents. They also have spouses or children. They have higher income levels compared to cluster 2, and more school experience. Almost all were listed as working last year. They have varying work travel times, indicating their ability to work/travel to work. They are the largest cluster.

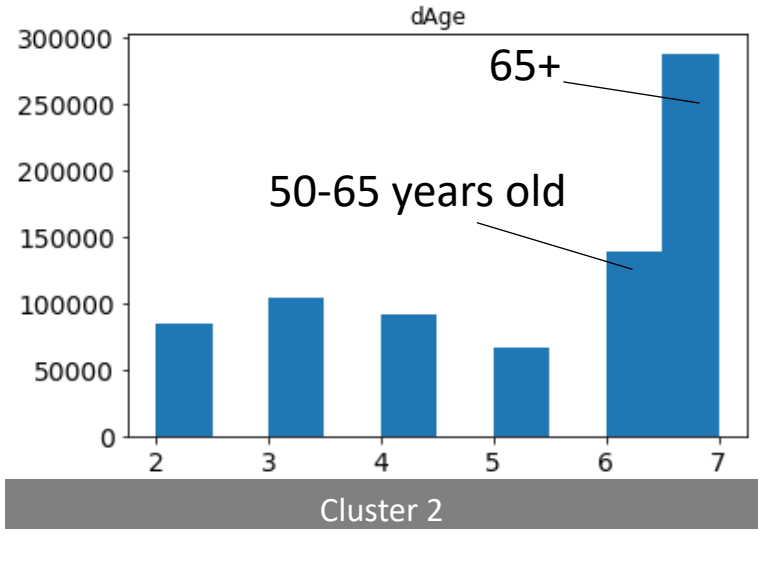
Info about all clusters

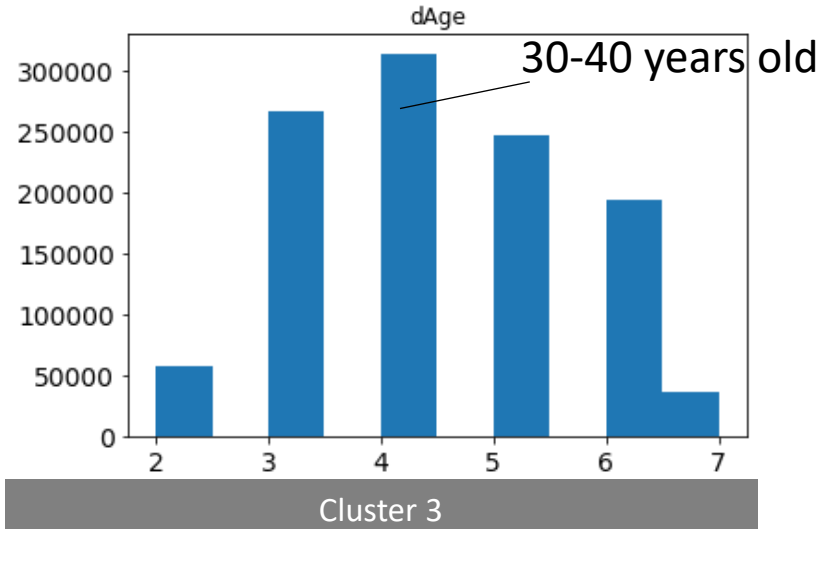
- Variables consistent in all clusters: Place of Birth, Weight, Numb Languages Spoken, English Speaking, Citizen Status
- Variables consistent in clusters 2 and 3: Military, Employment class, Spouse, Numb Children
- Variables consistent in clusters 1 and 3: sex

Final Cluster Analysis

One of the main describers for each cluster was age. Cluster 1 contained people 20 years old or younger. Clusters 2 and 3 had a mixture of ages. Cluster 2 had more people above the age of 65; Cluster 3 had a normal distribution distribution among ages >20 to >65.

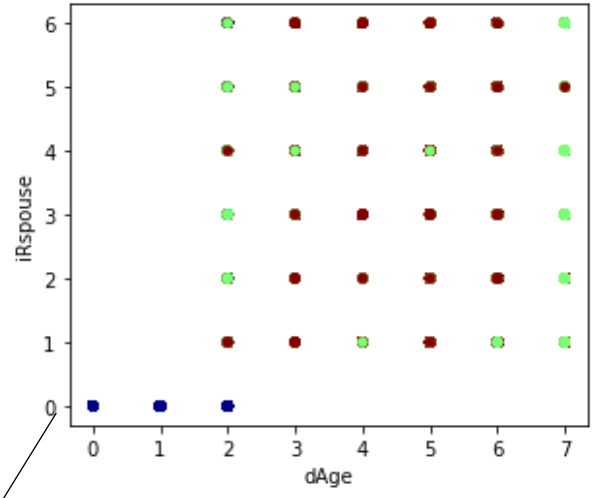




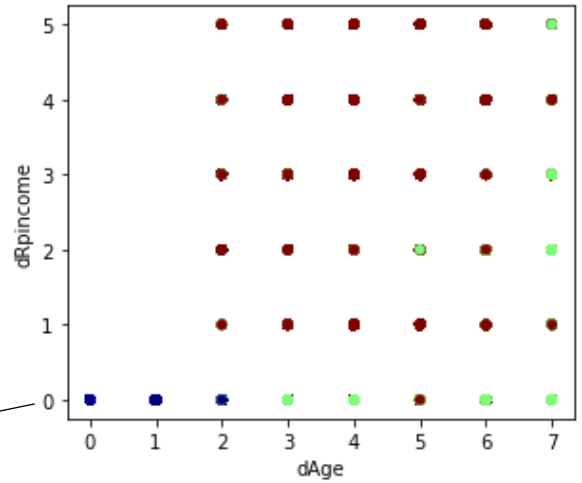


Final Clustering Analysis

I visualized age in comparison to other variables within scatter plots. Young Americans tend to not have spouses or income information.



Spouse Info N/A, or <15 years old

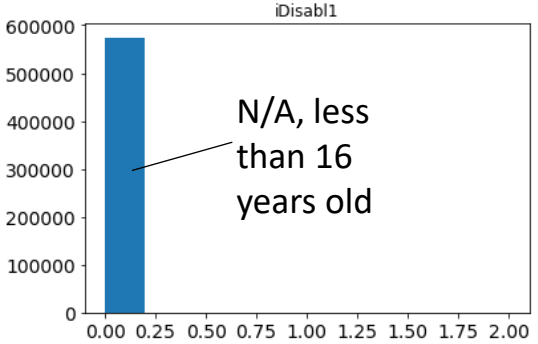


Income info N/A

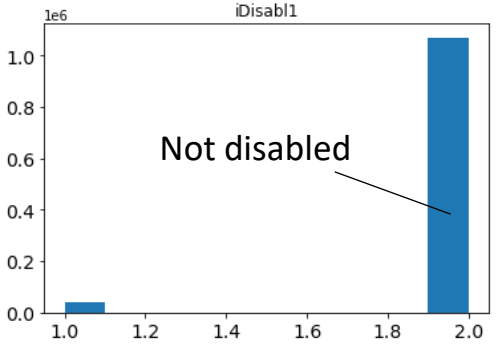
Blue- Cluster 1, Green-Cluster 2, Red-Cluster 3

I also looked at disability distribution. Many of the people in cluster 2 are listed as having a disability.

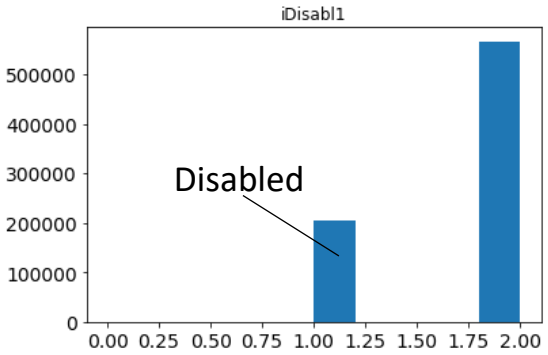
Cluster 1



Cluster 2



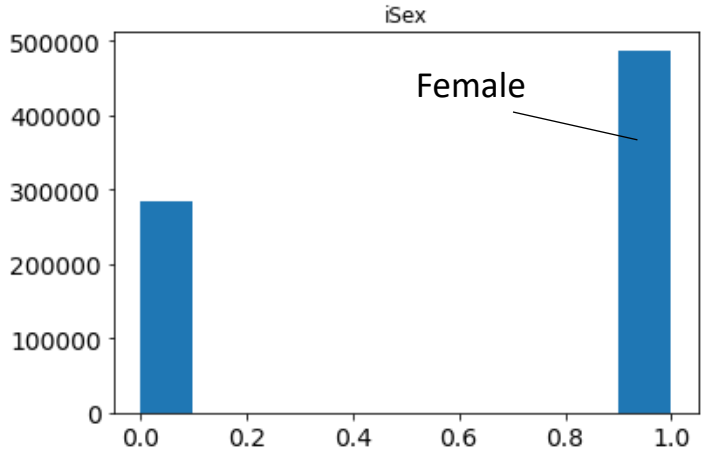
Cluster 3



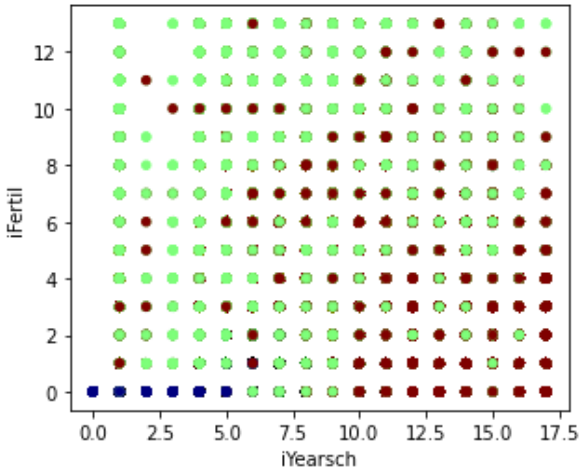
Final Clustering Analysis

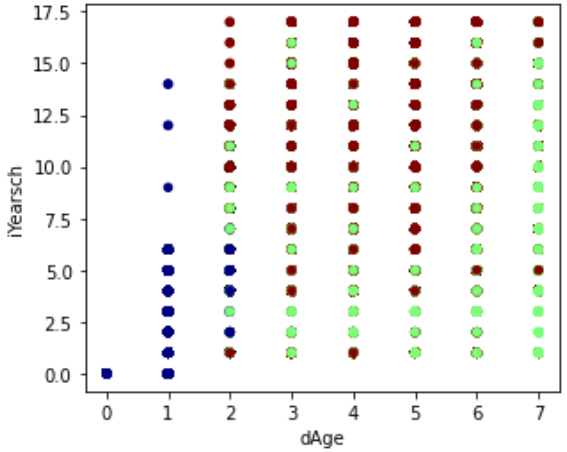
- Blue- Cluster 1
- Green-Cluster 2
- Red-Cluster 3

Sex for Cluster 3

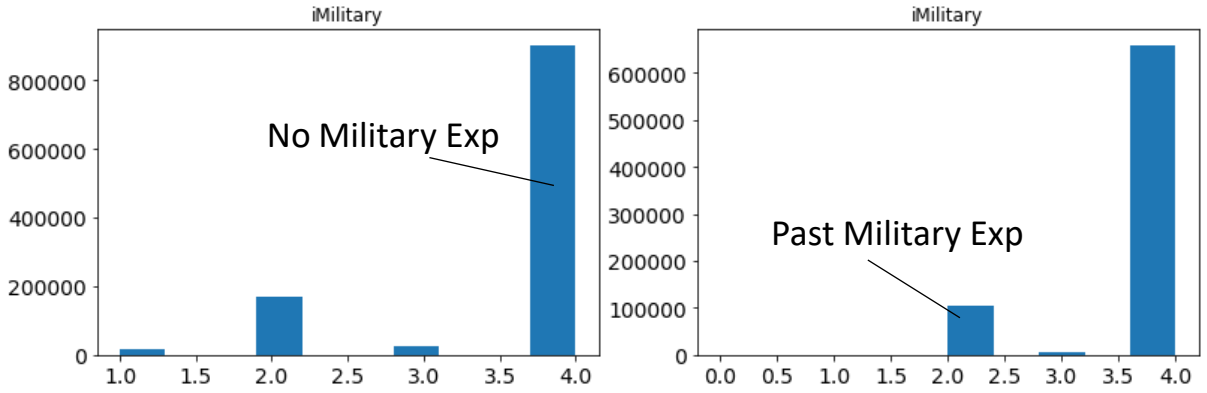


Years in School vs Numb Children, Age vs Years in School

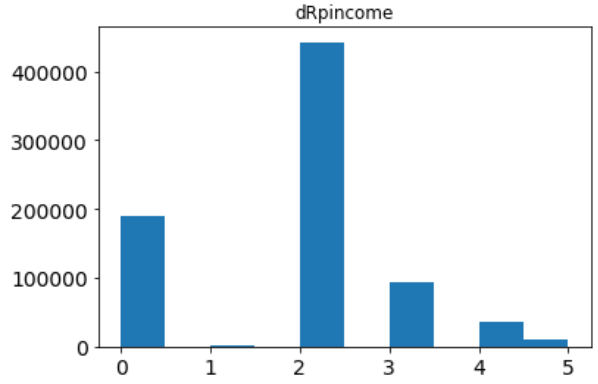


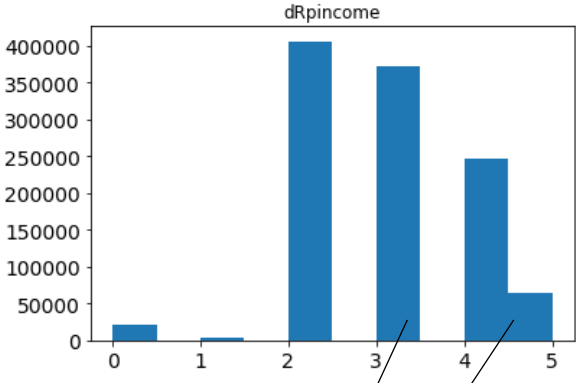


Military Status for Cluster 2 and 3



Incomes for Cluster 2 and 3





Highest income levels

Final Thoughts and Considerations

- The final clustering model was able to group the sample of Americans into clusters. Within the limits of this data sample, law-makers can understand information about 3 groups of Americans. One of their most important identifying traits is age.
- In 2020, the data would look different. There might be more diversity. Instead of just sex there could be sex and identified gender.
- Incomes would be different in 2020. To make a comparison to future data, the 1990 data incomes would need to account for inflation.
- With more time I would use Tableau or a visualization software to improve the exploratory analysis/final clusters visuals. Each category integer value could be labeled as a string. I could create stacked histograms.
- Create a Tableau dashboard to cover more charts and visualizations on the high level cluster descriptions.
- With more instances and features, the data could be divided even further.

Thank You

References

- Classmates
- Lecture Slides
- Medium.com
- Pandas documentation
- Spark API Documentation
- Stack exchange
- Stackoverflow
- Towardsdatascience.com