Adam Scerra

Final Project Report

**Introduction:**

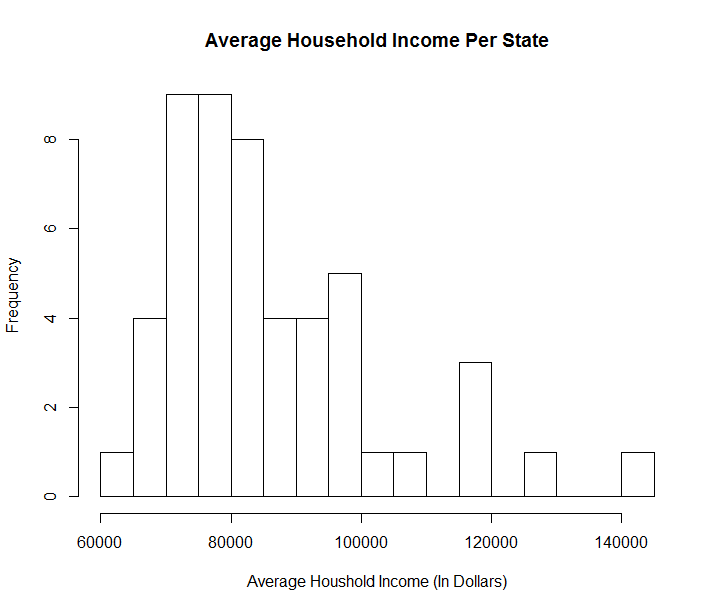
The goal of this project is to provide meaningful data analysis to a business about where in the United States would be best for them to start their business. The business is a high end gym that will specialize on working with youth athletes as they make their transition to the next level. Due to the nature of this gym and its specialized training techniques a membership and sessions to train there will be costly. They will be looking to start their business in an area that has a high amount of kids and also a high average income per household. This data can be found and mined with the dataset I chose and the data mining techniques I used.

The initial dataset that I used was from the 2011-2015 ACS 5-year Public Use Microdata Samples(PUMS) –CSV format. I chose to download the United States Housing Unit Records dataset. This came in a zipped folder with ACS 2011-2015\_PUMS\_README file, and the housing unit records dataset broken up into four .csv files. This data came from the United States Census Bureau, more specifically “factfinder.census.gov”. This dataset held information about all households in the United States. The raw data was 5.7 GB, after cleaning the subset of data that I used was 216,071 KB. The raw dataset as a whole consisted of 7,420,038 rows and 209 columns. Each row of data represents a household in the United States. There were 5 columns that I found to be most relative to my big data problem and they were ID(list of numbers displaying which number row it is), REGION(a column consisting of numbers that represent a region of the US that the household is in), ST(a column consisting of numbers that represent which state the household is in), FINCP(annual family income of the household), and NOC(number of children in the household).

I had to start cleaning the data; the first thing to do was to get rid of the columns that I did not need. The files were too large to open in excel, so to start cleaning the data I read the files into R. I had to read in the files 100,000 rows at a time in order to avoid memory issues. When I read in each row chunk into R I assigned it to an object I then took this object and combined only the 5 columns that I found relevant and created a new object with just these columns. Then I combined the 100,000 row chunks into a larger object and output it as a new file. I ended up with 8 files by the end of doing this because I could not call all of the rows and columns of the raw dataset into R without running out of memory, each of the four original files were broken up into two files that held the same number of rows but now consisted of only the 5 columns that I selected. I then could read in those eight files and combine them into one object. This object became my data subset and it now consisted of 7,420,038 rows with only 5 columns. There were still values that consisted of N/A’s I dealt with this by removing all N/A values with the na.rm=T parameter when I called on specific data. Once the data was cleaned of all missing values I could then start using R to evaluate rows of data. I first calculated the mean income per household for each state and stored it as a data frame, and then calculated the sum of all children in each state creating another data frame. I now had a data frame that gave me a list of each state with the value of the average income per household, and a data frame with a list of each state and its total amount of children. When I started to analyze this data I noticed a problem and it was that one state had so many children in it that no matter how I looked at it they were the obvious choice. But this is simply because the population for this state was so much greater than anywhere else. I decided to use percent of children in the state instead of total amount to give me a less skewed result. I did this by using information from U.S. population by State dataset to add a column of the total population for each state. I got this data from infoplease.com. In order to get the percentage of children that live in a household for each state I took the sum of number of children and divided that by the total population. I now had a data frame that listed all states and their percent of children living in households’ value. Then I combined the two data frames into one and output it as a file called Weka\_cut.csv.

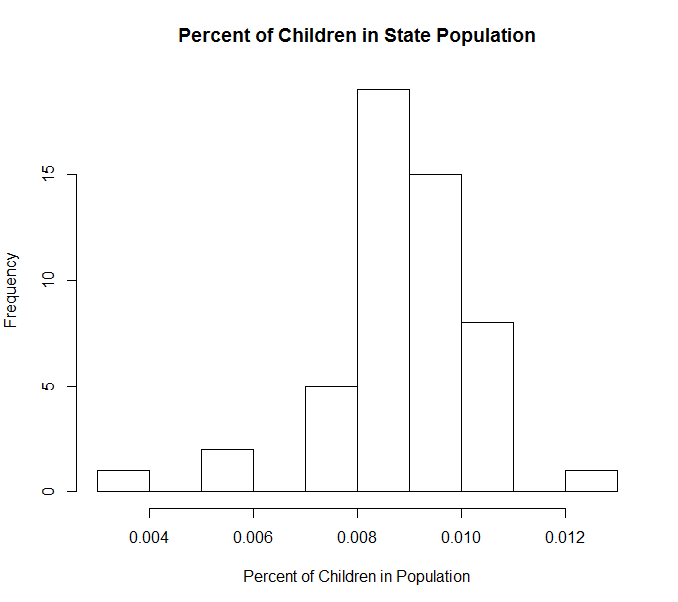
**Section 1** **– Data Visualization:**

The data is now in a position to visualize it and draw conclusions from the visualization. I used R as my visualization tool. First I visualized a graph of the relationship between Average Household Income and frequency. This graph helped to show how many states fell under which income category.



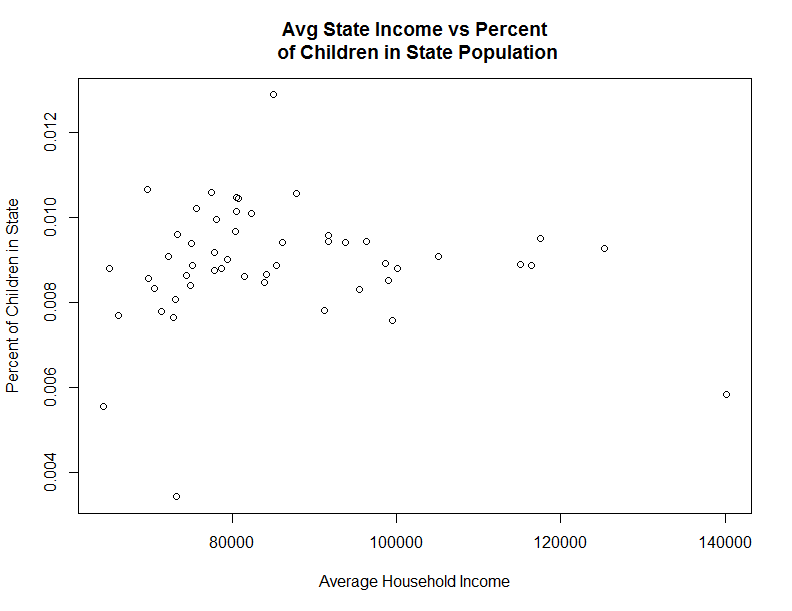
We can see here that there is a lot of states right around the $80,000 average income mark. There are few that are under the $70,000 mark then we can see another spike in the data right beneath $100,000. After this mark the results start to taper off. The handful of states on the right section of this graph all have a high potential for the market we are looking for in this big data problem.

I then created another histogram that shows the relationship between the percent of children in state population and frequency.

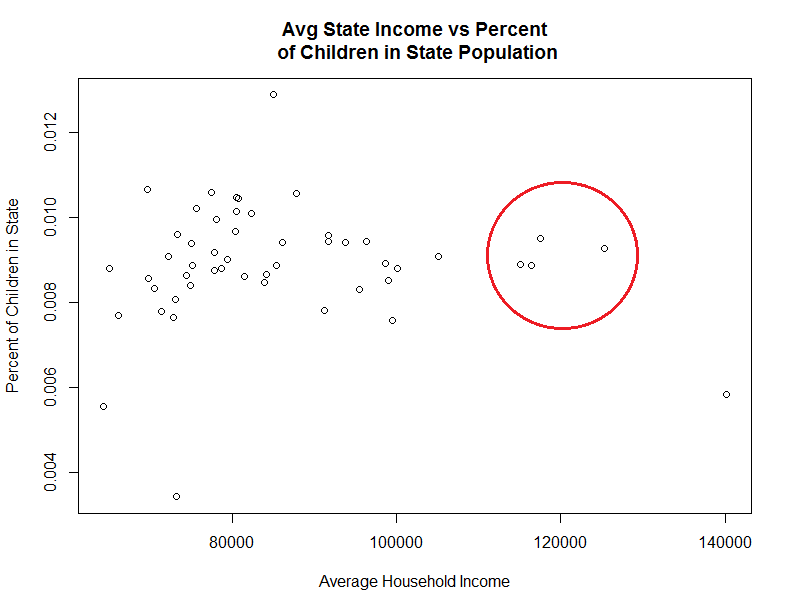


Here we can see that the large majority of states have a percent of population above .008 there are only a few that are underneath these marks and these states will most likely be no use to this project. Aside from the main grouping from .008-.011 there is one state that has a much higher percent of children in their population this state will supply the kids necessary to give the business more potential to succeed. These graphs are nice to get a feel for the data we have collected. It gives an idea of what the values are from state to state.

The next graph that I produced was a scatter plot this graph shows some very important results. The plot is of the relationship between a states average income per household and percent of children in the population.



Here we have average household income on the x-axis and percent if children per state in the y-axis. Any point that falls in the lower left corner of this plot will be excluded from the states that are believed to be a good place for this startup. Any data point that is near the top right of this plot will be a great place to use for this startup, because they will be states that have a high average income per household and a high percentage of children in their population. There are a couple outliers on this plot we see that there is one state that has far more percent of children in their population then all the others and also we see a state that has a much higher average household income then the rest. Based on looking at this plot there is a cluster of four states that seem to be a good fit to answer this big data question.



These four states highlighted in the red circle all produce high average household incomes and have a high percentage of children in their population. These points all represent a different state. It is possible to figure out which state is which and to find out the exact values of all points and outliers with statistical analysis.

**Section 2 – Statistical Analysis**

Using statistical analysis we can better understand our data and find out exactly what the values are behind our graphs; R is being used to produce these statistics. The statistical measures that need to be calculated to make better sense of this data are highest average income, and which state it represents, highest percentage of children and which state that represents. These measures will help us uncover what and where our outliers are coming from. The average household income for the entire country will give us perspective on how much more or less the measures we have seen in our graphs are compared to the rest of the country. The location with the highest overall amount of children and what amount that is will give us a good idea of what to consider a high amount of children. The original dataset came with a list of numbers that assigned a state to each number, the code I wrote produced a number and I then went to the list provided by the dataset to find out which state it was assigned.

Highest Average Income\State: $140,201.30 \ 11 = District of Colombia(DC)

Highest Percentage of Children\State: .012897 \ 45 = South Carolina

Average Household Income for Country : $87,542.23

Mean Number of Children Per State: 56,520.64

Correlation between Average Household Income and Percent of Children in State: -0.02211

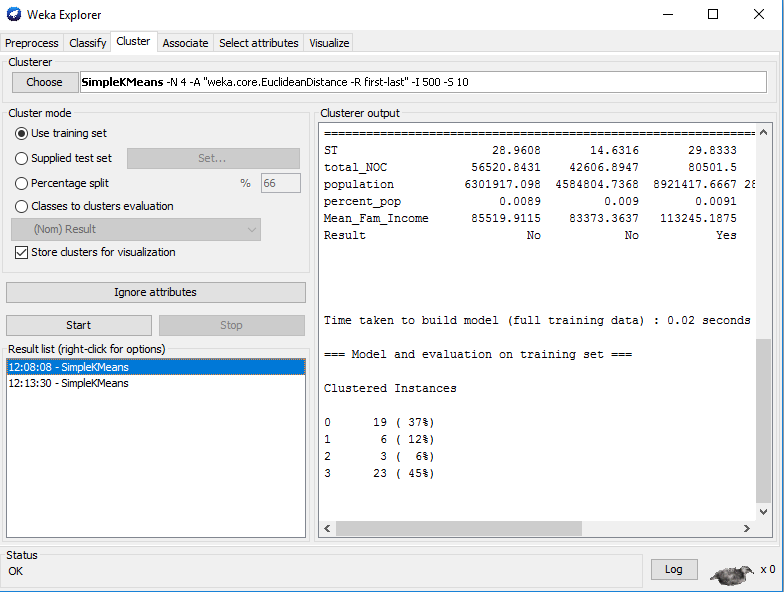
These statistics give us some pretty good insight as to what we are looking at in the graphs. The outlier that was found in regards to average household income was the district of Columbia, with a population of only 667,228 people this most likely will not be a good place for the business startup. South Carolina seems to have the most children per its population, yet is far below the average household income for the country, coming in at $73,042.04. This will not be a good place to start a gym that is difficult to pay for since the average income of the people living there is not very high. In the results of the R code that is at the end of this report there is a list containing every states average income, total number of children, and percentage of children. Using these lists the four states that were highlighted in the scatterplot above can be found. They are Massachusetts, Maryland, New Jersey, and Connecticut. Based on this statistical analysis I would recommend that this business starts in anyone of these four states.

**Section 4:**

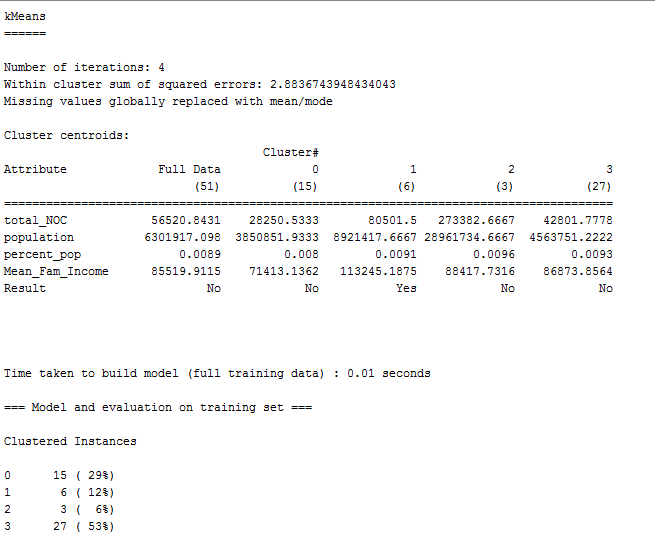
**Data Mining Question:**  Where in the US, by state, would a gym that offers high end specialized training for youth athletes be most successful?

In order to do this the data that was cleaned and analyzed with R needed to be prepared even further to use for data mining. This is where the use of Python came into play. The file I had created called Weka\_cut.csv had all of the attributes that were used to make statistical decisions. After looking at the data and results from the previous sections. A state will be classified as Yes it will be a good place to start this business if the percent of children in the population is above .008 and the average income per household is greater than $100,000. Using this I imported the Weka\_cut.csv file into python and ran it through an if statement where if it met those standards it would return the result Yes in a new column names Results. If it did not meet the standards then it classified the state as No and added it to the Result column. Once I had run this Python code I had the Weka\_cut.csv file with an additional column named Result with all Yes or No values.

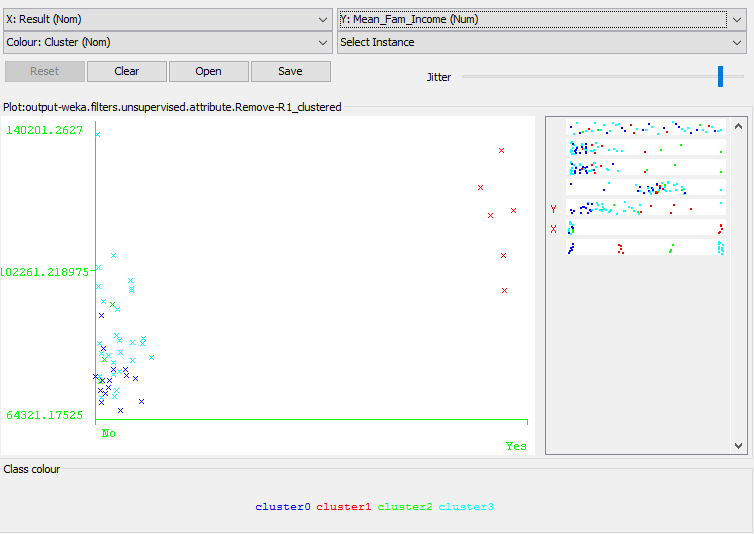
Before I was able to use this Weka\_cut.csv file in WEKA it needed to be converted into an .arff file format. This was done with the use of Java. I first needed to implement the weka.api package and imprt weka.core.Instances. I created a class CSV2Arff and loaded the Weka\_cut .csv file as input and saved it as a .arff file named Real\_Data that could be opened in WEKA.

Once the file was in the correct format it was ready to be opened in WEKA(Waikato Environment for Knowledge Analysis). In the Explorer application under the preprocess tab the .arff file was able to open and be brought into WEKA. The Data Mining model used to answer this question was clustering. The SimpleKMeans algorithm was used to cluster the data. Four clusters were chosen to show specific results. This algorithm was run on the training set of the data. 

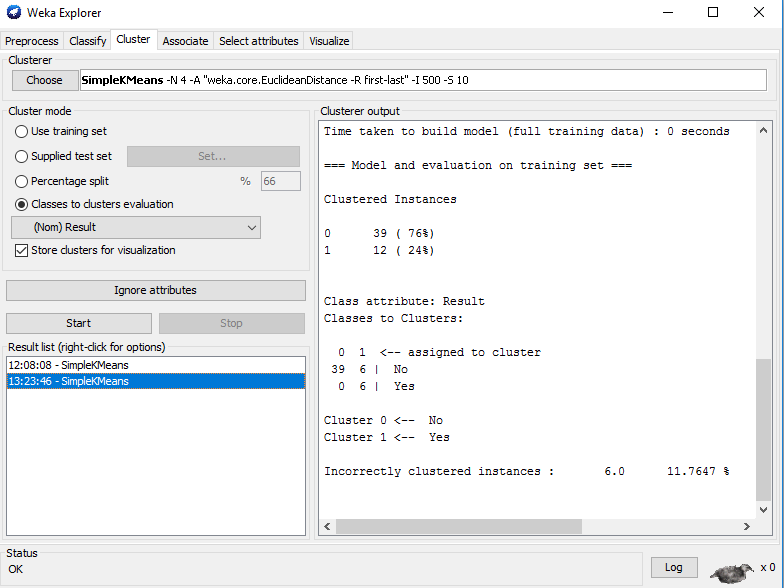
Analyzing results:



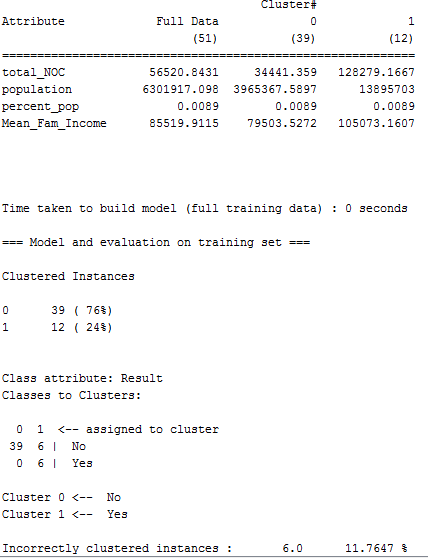
Here are the results shown of the four clusters that were produced using the SimpleKMeans algorithm. The results were based off of four different attributes. Total number of children in state(total\_NOC), total population of the state(population), percent of children in the population(percent\_pop), and the average family income of all households in the state(Mean\_Fam\_Income). Looking at the results cluster by cluster we see that 15 instances were grouped in cluster 0, 6 in cluster 1, 4 in cluster 2, and 27 in cluster 3. First cluster number 2 should be excluded for analysis because there are too few instances to make an informed business decision on what it is representing. Starting off by looking at cluster number 0 this cluster is full of instances that have just barely made the cut for the percent\_pop they are well under the average Mean\_Fam\_Income, and have a relatively low population in general. This cluster of states does not interest us as a potentially lucrative location for this business startup. Cluster 3 holds the most instances and its results are improved from what we see in cluster 0. There are far more people in these states and the percent of children based on population is higher than any other cluster we have here. Although the average income per household does not give us a promising result. In these states we will find no shortage of customers available for this type of service but it will be difficult to get them to spend a good amount of money on a consistent basis. Next we look at cluster 1; this cluster holds 6 instances. We see that total\_NOC is above 80,000 which is far greater than any other cluster. The percent\_pop is at .0091 which is a pretty strong result; based on the entire population there are a relatively high amount of children there, and then we see the Mean\_Fam\_Income attribute. This is far greater than any other cluster these are the states that will have the most money per household to spend. The results attribute yields a yes for this cluster of instances only. States in this cluster will have a large amount of children in its population giving a good customer base for this startup, and since the average amount of income is so much higher there will be more exposable income that can be used on their children .This cluster gives us the results of the best locations in the US for this type of business startup. To make these cluster stand out better they needed to be visualized. Here is a graph based on the results of the four clusters. The y axis represents average total income and the x axis displays no for those that do not meet the standards and yes for those that do. As you can see cluster 1 shows up in red on the right hand side showing states that were predicted to be more suitable for this business startup. Average family income plays a large part in this decision; most of the states that were labeled as yes are in the upper end of the average family income.



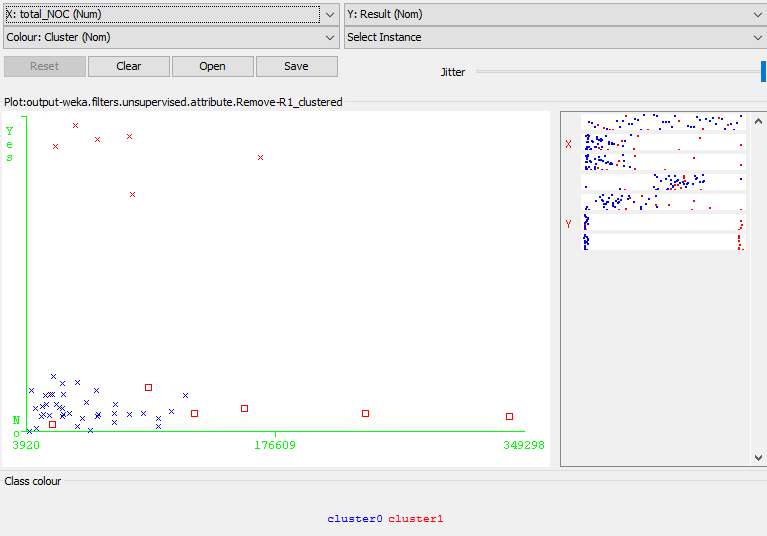
After exploring this data using the training set for evaluation, another SimpleKMeans algorithm was ran using classes to clusters evaluation. This yielded some interesting results.



Here is the output.



This was done using just two clustered instances; cluster 0 was to hold no values and it consisted of 39 instances and cluster 1 was meant to hold yes values and had 12. Of the 39 instances in cluster 0 all 39 were clustered correctly whereas cluster 1 had 6 correct instances and 6 incorrect, leading to an incorrectly clustered instance percentage of 11.7647. Looking at the attribute values for these two clusters we see some similarities and some differences. The percent\_pop attribute stands out right away because they both have the same amount of children in their population. Yet all of the clustered instances in cluster 0 were no’s, this is due to the Mean\_Fam\_Income attribute, it is much larger in cluster 1 giving those 12 instances higher potential to spend money. It can also be seen that in the first two attributes cluster 1 is much greater then in cluster 0. In order to make more sense of this output the data must be visualized.

****

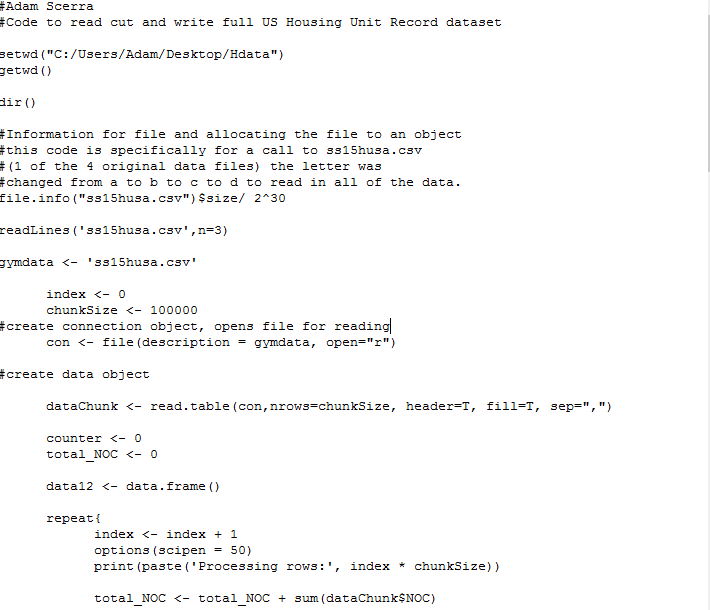
Here is an interesting display of the cluster by classes data. Cluster 0 is in blue and cluster 1 is in red. The red x’s are correctly clustered as yes where the red squares are incorrectly clustered. These are instances whose results were listed as no but were classified in the yes cluster. With the exception of one instance the other five do have something interesting in common. They all have a large amount of the attribute total\_NOC, which is total number of children. They were clustered as no in the first run of the algorithm but here the algorithm clustered them into the yes category. It appears that the total number of children in the state is playing a major role here. Although these states do not have the necessary average income to predict families will spend money they have enough volume of children to make a case that these areas will be successful for this business startup.

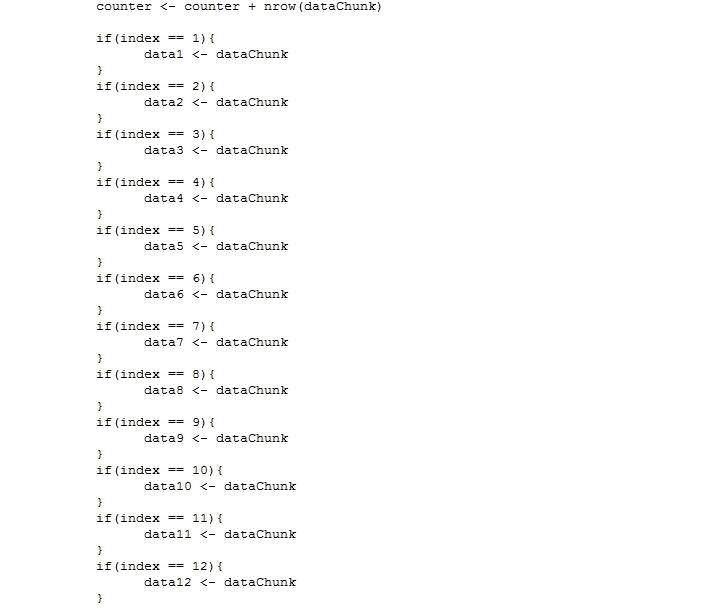
**Section 5:**

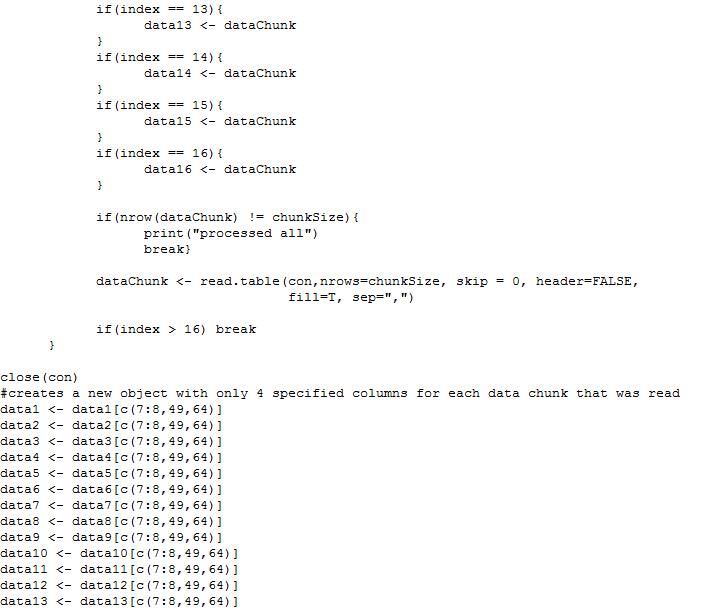
Based on what was done here there is a lot that could be asked and accomplished for future use. The model that was created here can be used anytime new household and population data is released. When new data is released I would like to see if there had been any spikes in number of children in any state, even if this state had a lower number of children to start with; a lot of population growth of children would give reason to believe that the gym will succeed for years to come. Also for future use I would like to get a dataset that had more specific data on the individual state locations, such as city and town. Unfortunately the dataset I was working with did not supply that, it would be nice to narrow down my results even further in the future.

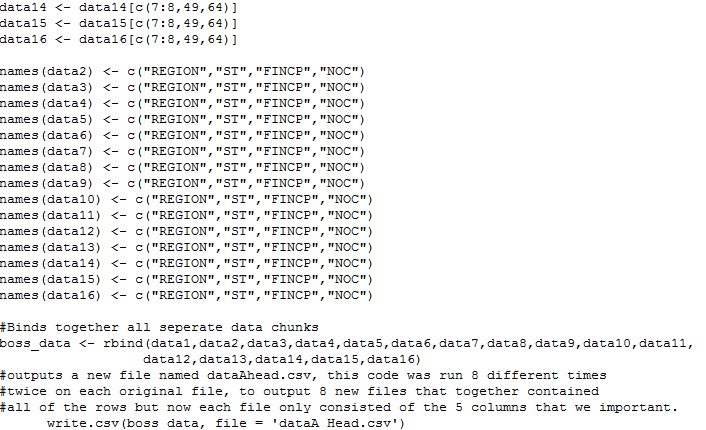
**Appendix:**

R Code to create files that only contain the 5 columns of interest.

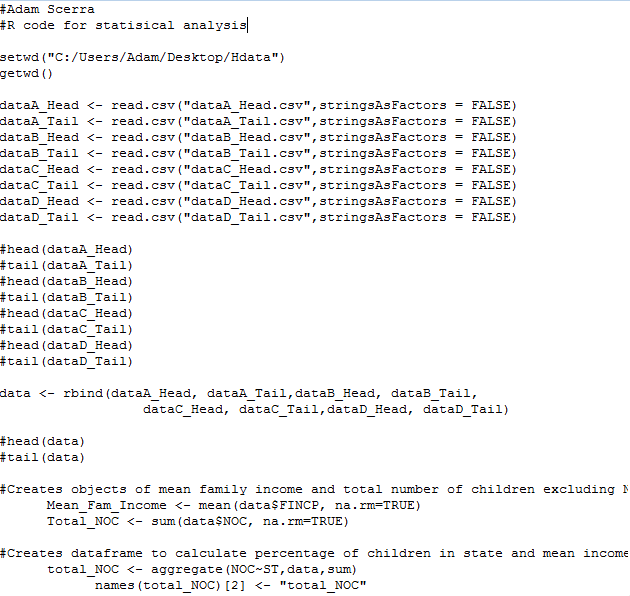




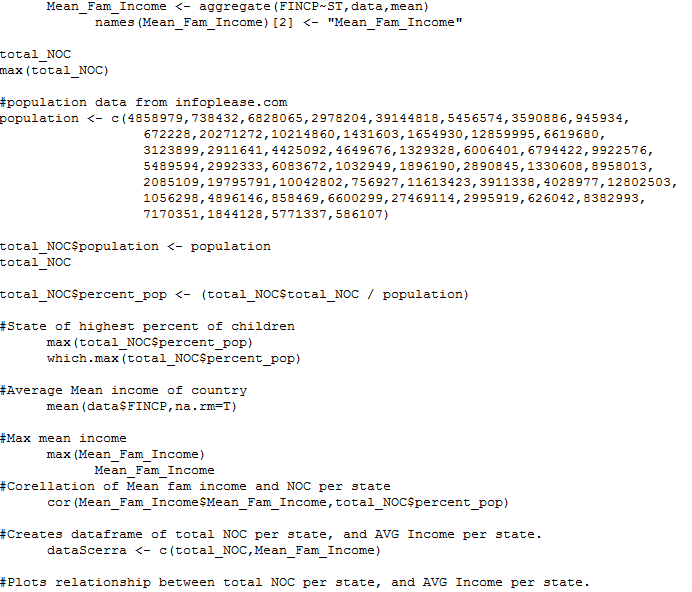


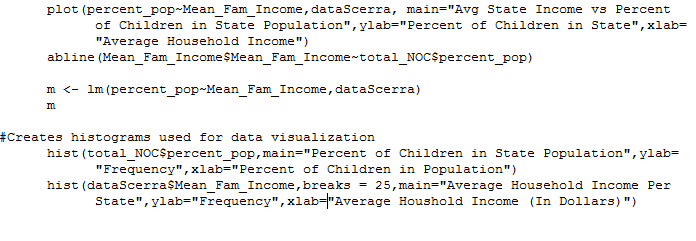


R code for to read 8 output files and join them together to run statistical analysis.









Outputs:

> getwd()

[1] "C:/Users/Adam/Desktop/Hdata"

> max(total\_NOC)

[1] 349298

> total\_NOC$population <- population

> total\_NOC

#Total Number of Children per State and Total Population

ST total\_NOC population

1 1 40466 4858979

2 2 7811 738432

3 4 62683 6828065

4 5 26207 2978204

5 6 349298 39144818

6 8 51470 5456574

7 9 33348 3590886

8 10 7403 945934

9 11 3920 672228

10 12 212201 20271272

11 13 91995 10214860

12 15 10867 1431603

13 16 17655 1654930

14 17 121374 12859995

15 18 63575 6619680

16 19 31133 3123899

17 20 29570 2911641

18 21 40200 4425092

19 22 39095 4649676

20 23 10166 1329328

21 24 53323 6006401

22 25 60436 6794422

23 26 105173 9922576

24 27 48768 5489594

25 28 16618 2992333

26 29 57082 6083672

27 30 9167 1032949

28 31 19813 1896190

29 32 25441 2890845

30 33 11348 1330608

31 34 85312 8958013

32 35 17857 2085109

33 36 174337 19795791

34 37 87928 10042802

35 38 7256 756927

36 39 39746 11613423

37 40 30435 3911338

38 41 34725 4028977

39 42 110899 12802503

40 44 8790 1056298

41 45 39482 4896146

42 46 8770 858469

43 47 57007 6600299

44 48 258649 27469114

45 49 38638 2995919

46 50 5301 626042

47 51 76253 8382993

48 53 67550 7170351

49 54 14216 1844128

50 55 55887 5771337

51 56 5919 586107

> total\_NOC$percent\_pop <- (total\_NOC$total\_NOC / population)

> #State of highest percent of children

> max(total\_NOC$percent\_pop)

[1] 0.01289688

> which.max(total\_NOC$percent\_pop)

[1] 45

> #Average Mean income of country

> mean(data$FINCP,na.rm=T)

[1] 87542.23

> #Max mean income

> max(Mean\_Fam\_Income)

[1] 140201.3

> Mean\_Fam\_Income

ST Mean\_Fam\_Income

1 1 70495.34

2 2 87857.48

3 4 77841.36

4 5 65046.42

5 6 98619.70

6 8 96390.49

7 9 125298.30

8 10 91255.10

9 11 140201.26

10 12 80479.59

11 13 79365.24

12 15 99546.80

13 16 69679.22

14 17 91758.36

15 18 73341.23

16 19 78006.36

17 20 80459.92

18 21 72221.57

19 22 74877.45

20 23 72886.45

21 24 116465.33

22 25 115030.36

23 26 77466.45

24 27 85397.40

25 28 64321.18

26 29 75079.19

27 30 75158.70

28 31 80762.53

29 32 78618.22

30 33 99002.35

31 34 117474.11

32 35 69738.78

33 36 100083.27

34 37 77849.43

35 38 91682.54

36 39 73237.99

37 40 71323.22

38 41 81473.76

39 42 84193.98

40 44 95508.89

41 45 73042.04

42 46 75626.87

43 47 74432.73

44 48 86153.91

45 49 85060.51

46 50 83941.51

47 51 105119.76

48 53 93832.67

49 54 66117.74

50 55 80361.70

51 56 82330.73

> #Corellation of Mean fam income and NOC per state

> cor(Mean\_Fam\_Income$Mean\_Fam\_Income,total\_NOC$percent\_pop)

[1] -0.02210945

Call:

lm(formula = percent\_pop ~ Mean\_Fam\_Income, data = dataScerra)

Coefficients:

(Intercept) Mean\_Fam\_Income

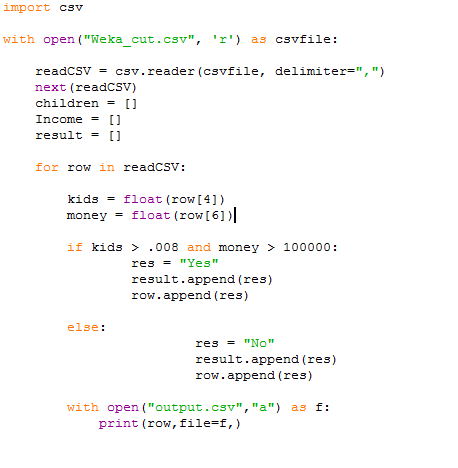
0.009083461379 -0.000000001936

> country\_mean <- mean(data$FINCP,na.rm=T)

> country\_mean

[1] 87542.23

Python Code:



Java Code:

