Jeff Levesque

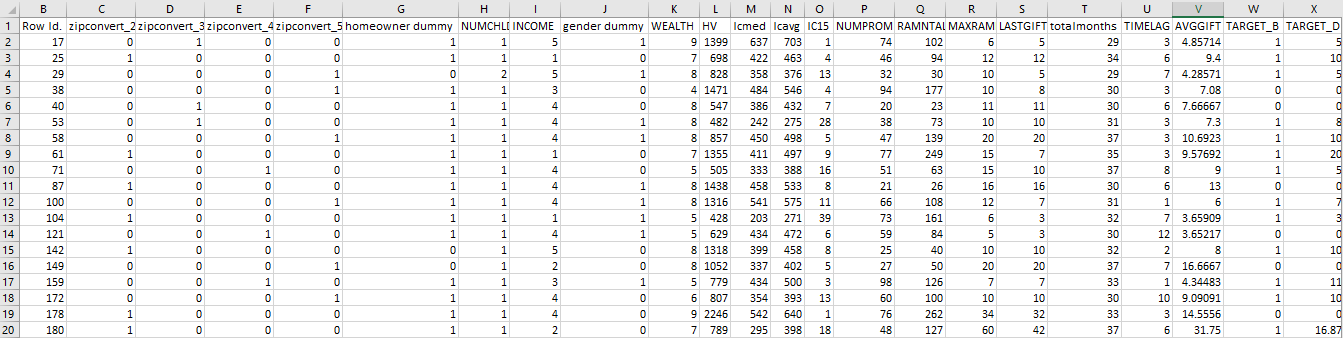
Professor Landowski

IST 652

Homework #1

**Data / Source**

The provided csv dataset represents historical donor activity. Specifically, each of the 3121 rows, characterizes a unique individual, along with various related attributes:



**Analysis**

Data Preparation:

Initially the following columns were removed, since for the context of clustering, they do not provide any additional exploratory benefits:

* Row Id
* Row Id.

Additionally, all zip related columns were removed for similar reasons, in addition to being very sparse in value:

* zipconvert\_2
* zipconvert\_3
* zipconvert\_4
* zipconvert\_5

Exploratory Analysis:

Two analysis techniques were implemented, the kmeans clustering, and the aggregated income-hv mean. However, the kmeans was further split into two additional study:

* wealth versus average donation
* number of children versus average donation

The comparison between wealth versus average donation required several intermediary steps before the kmeans clusters were generated. Specifically, both dimensions were normalized, followed by applying principal component analysis (pca). Once the pca was computed, the corresponding values were provided to generate the kmeans scatterplot.

For the number of children versus average donation, the normalization step was skipped. This was because more thought was required to develop a reasonable normalization function.

Lastly, the aggregated income-hv mean entailed aggregated the income column from the provided dataset, then computing the mean on the home value (HV) column.

**Program Description**

Dependencies needed by analysis.py, to generate the kmeans, and aggregated sum include various python packages:

* python-pip
* sklearn
* matplotlib
* numpy
* pandas

As well as some linux (ubuntu) dependencies:

* python-dev
* python-pip

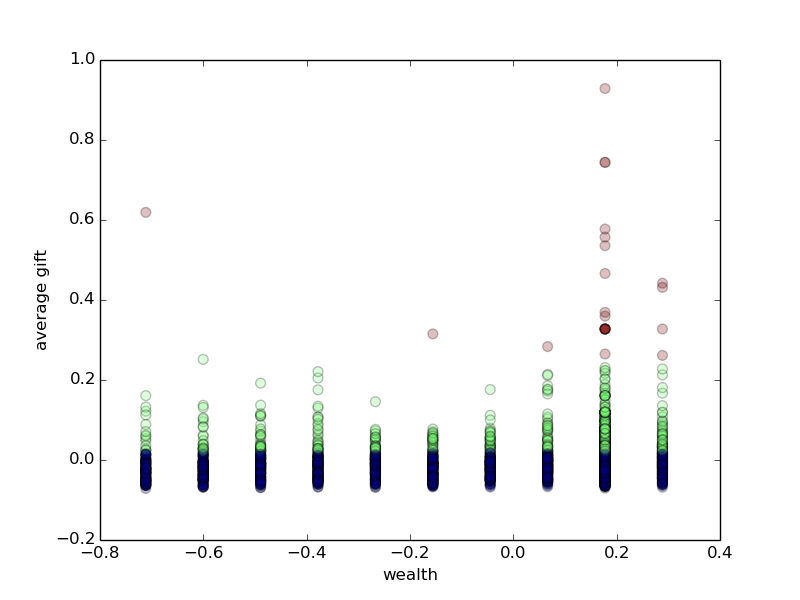
**Result / Output**

Once the above dependencies have been installed, the corresponding python script can be executed as a standalone script in any directory. Upon successful completion, the following files will be generated in the current working directory:

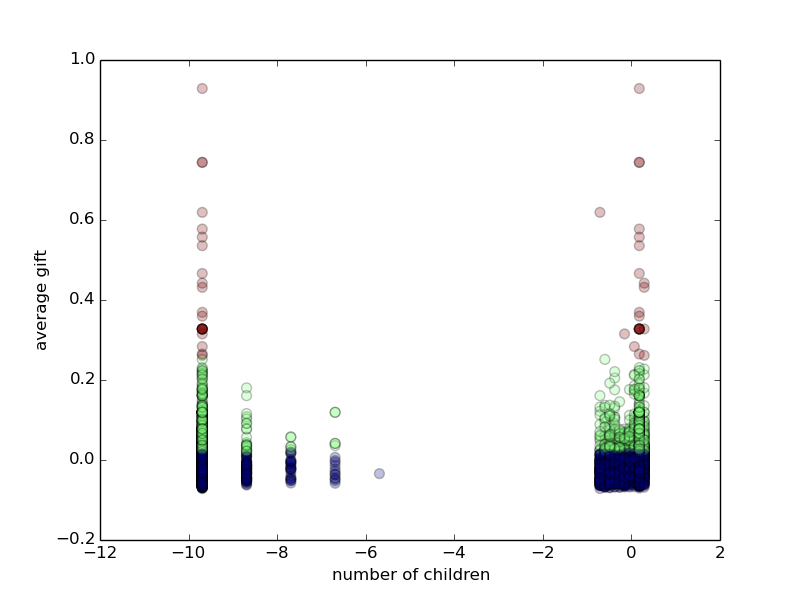
* kmeans\_wealth\_gift.png
* kmeans\_child\_gift.png
* summary.txt

The png files are visualizations of the corresponding kmeans clusters, while the summary.txt contains the aggregated mean home values, grouped by income.

Upon inspection of the kmeans\_wealth\_gift.png, as donor’s wealth increases, the average gift value tends to increase as well. More specifically, the red cluster tends to be denser at higher wealth and average gift values, while the second cluster (visually colored green) follows the same trend. The last, third cluster appears to be consistent, and unchanged across the wealth dimension.



The result of the kmeans\_child\_gift.png, appears less intuitive. More generally, the results almost seem to be a binomial condition, while favoring more to the end with more children. These results could potentially be improved with an implemented normalization step.



Finally, it was found that as donor’s income group is related to their home value. Specifically, as the income group increase, the home values similarly increase. When viewing the corresponding summary.txt, it is important to note that each successive line increments in the income group, which is provided at the top most of the file.