Are Billionaires Just Like Us? Probably Not.

A Cluster Analysis by Alex Rosenblum

Problem Statement

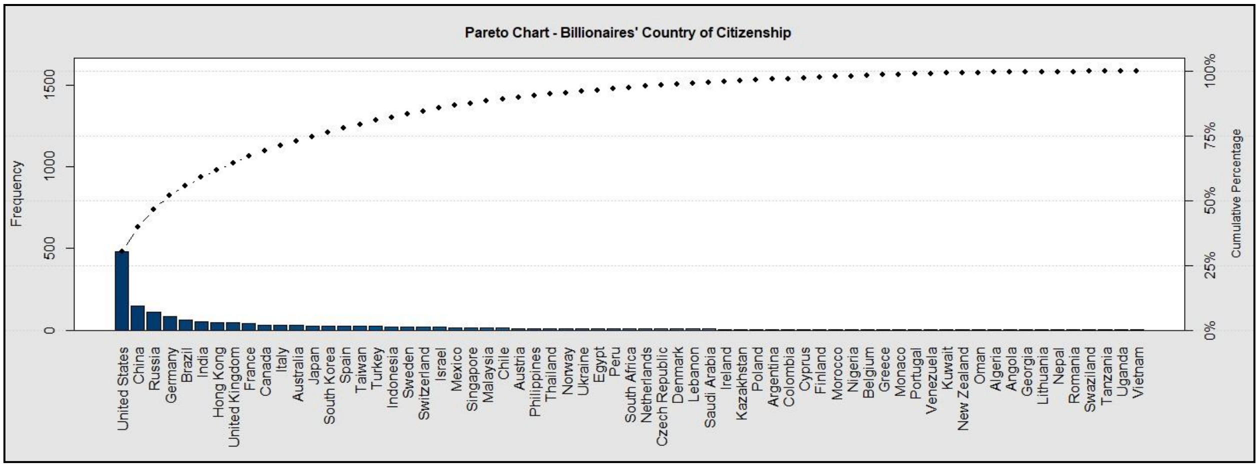
The billionaire class has been the subject of heated discussions for years, and for good reason. Depending on your point of view, they can either represent the greatest success stories of the modern age, having risen to the absolute top of global society, or be considered parasites on that same society hoarding their incomprehensible wealth while tens of millions suffer in poverty as the climate crisis many of their companies are responsible for slowly ravages the world around them. I chose this dataset for cluster analysis to better understand this strange, elusive group because "the billionaire class" is often discussed as an elite monolith comprising the wealthiest people in the world, but as there are over 1600 individuals represented in this dataset, I wanted to see what sorts of subclasses of billionaire could be discovered through a computational approach.

Data

The dataset, pulled from the CORGIS collection of datasets, contains information describing every known billionaire worldwide (in terms of USD) compiled at three points in time: 1996, 2001, and 2014. There are 2614 observations across the 3 time points, with at least 1600 individuals represented. Of the 22 attributes included, 2 are unique ID values and 3 are useless booleans for which the value is TRUE for every observation, leaving me with 17 usable attributes. These attributes can be summarized by a few main categories. There are demographic attributes which describe the individual's age (demographics.age) and gender (demographics.gender), geographical attributes like the country of which they are a citizen (location.citizenship), the GDP of that country for the year of interest (location.gdp), and the region of the world the country falls into (location.region), professional attributes like the year an individual's company was founded (company.founded) and their relationship to that company (company.relationship), and finally attributes that describe the individual's wealth, like their net worth (wealth.worth.in.billions), and various information about how their wealth came to be (wealth.type, wealth.how.category, wealth.how.industry, wealth.how.inherited).

As stated above, this dataset contains data collected at three points in time. Several billionaires – Bill Gates, for example – are represented in all three time points, indicating that they had been a billionaire at least as early as 1996 and held wealth above $1 billion until 2014. Filtering these rows, I found that the batch of data collected in 2014 alone included over 1600 individuals, more than half of the total dataset. To simplify this project, I decided to limit my analysis to the rows collected in 2014. I didn't want to lose the information present in the data from previous years, so I created a separate data-frame for each year, structured as df_XXXX where "XXXX" is the year in question (df_2014, for example). I will return to this subject in the data preparation section. The exploration below is limited to the 2014 data.

One obvious trait of this dataset is the imbalance of represented countries of citizenship. The pareto chart below illustrates that the United States dominates the global billionaire scene, representing over 30% of observations. This will become important as we will be using GDP data as an input for our model. An even stronger imbalance is present among gender demographics, with 89.5% of billionaires being men. However, gender will not be considered when building the model, so this imbalance will have less impact on analysis. The other issue I will face is the presence of missing values – demographics.age has several values equal to zero and unless there are infant billionaires out there, that needs to be taken care of in some way. Additionally, the year 2014 is completely devoid of GDP information – this will have to come from a second dataset and be incorporated with the billionaires dataset.



Data Preparation

The next step in the data preparation process was to address missing values. Many had been taken care of by filtering out the rows from 2001 and 1996, but there were several remaining. The first attribute that had missing values to be addressed was demographics.age. For this attribute, there were 63 missing values. Many of these were married couples or siblings ("Thomas and Raymond Kwok", for example), for which a single age value would not apply. Several others were individuals whose ages were simply absent. One option for this issue would be to remove the pairs and use rnorm() to assign ages to the individuals randomly within the distribution of the rest of the demographics.age attribute for the dataset. However, since this was the only attribute with a sizeable portion of missing values, I decided it would be acceptable to simply remove all rows with the demographics.age value missing.

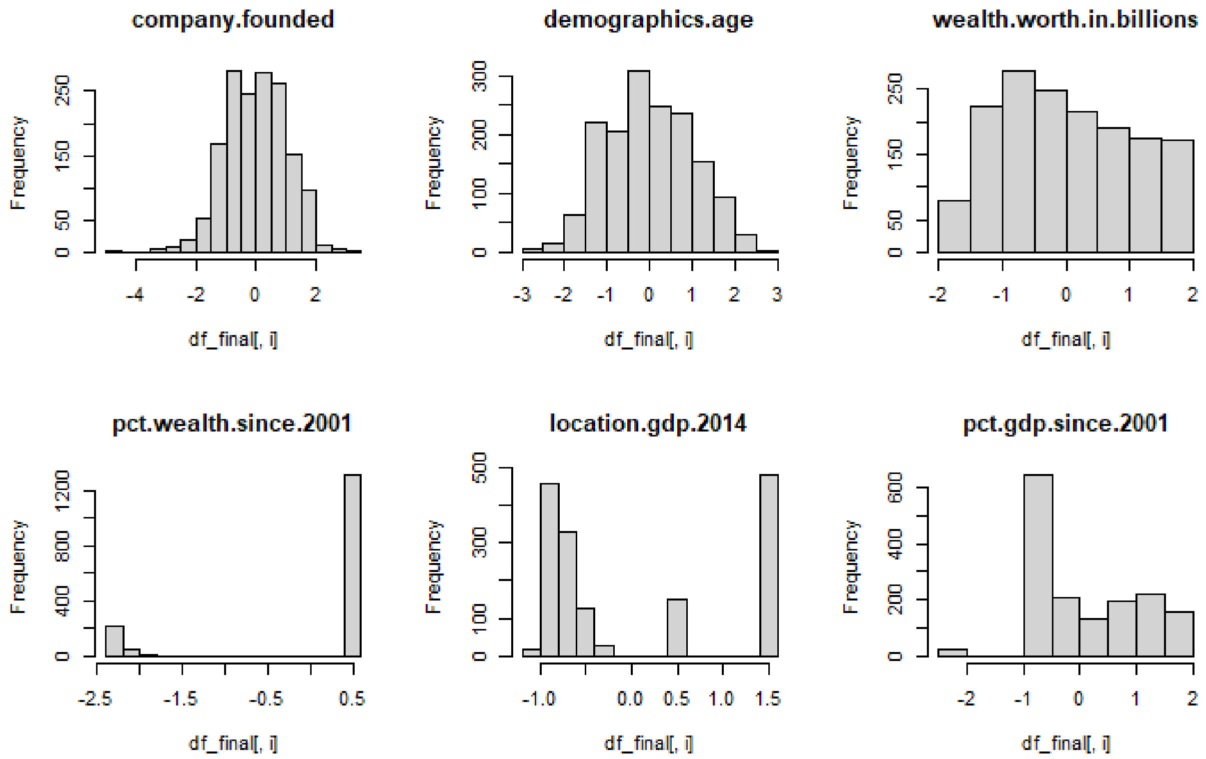
The other attributes which had missing values was company.name and company.founded. There were only 3 rows missing company.name (company.founded was also missing), so those rows were removed. Having only 9 remaining rows with company.founded missing I was able to look up these values individually, with the exception of Miao Shouliang, for whose company Fuyuan Group I was unable to find a founding year. This row was removed.

Next I wanted to incorporate back in the information from 2001 and 1996 previously filtered out. To start, I made two new data-frames called df_2001 and df_1996 to contain these subsets separated by year. In the df_2014 data-frame I created the new variables wealth.in.2001 and wealth.in.1996, as well as the calculated variables pct.wealth.since.2001 and pct.wealth.since.1996, which measures how much of a billionaire's wealth in 2014 was acquired since 2001 or 1996, respectively. For example, Bill Gates' wealth in billions was 18.5 in 1996, 58.7 in 2001, and 76 in 2014. So for Bill Gates, pct.wealth.since.2001 = ((76 – 58.7) / 76) * 100% = 22.8%, meaning that 22.8% of Bill Gates' wealth in 2014 was newly acquired since 2001. Similarly, someone who became a billionaire between the years 2001 and 2014 would have a pct.wealth.since.2001 equal to 100%, since all of their 2014 wealth would be newly acquired. My motivation for creating this variable is to measure how aggressive and successful a billionaire may have been in growing their wealth without excluding billionaires not represented in 2001 or 1996.

The next step in the data preparation process was to address the fact that the entire attribute of location.gdp was missing for 2014 data. To fix this, I downloaded a second dataset containing yearly GDP data for every country for several decades and imported it as a new data-frame. To merge this information with the existing data-frame, I wrote script to iterate over each billionaire and pull the GDP for 2014, as well as 2001 and 1996 for the country listed in the location.citizenship attribute and save them as new attributes in df_2014 (location.gdp.2014, for example). However, this was complicated due to some mismatching country names which were addressed individually. There were two values in the location.citizenship list which were not represented at all in the GDP dataset, those being Taiwan and Guernsey. I was able to pull the GDP for Taiwan from a different source than the rest of the GDP data and add it to the GDP data-frame manually. The location Guernsey, being a British Crown Dependency, was replaced with the value "United Kingdom", already present in the location.citizenship list. In addition to creating the GDP attributes, I also created the attributes pct.gdp.since.2001 and pct.gdp.since.1996, which were calculated in exactly the same way as the pct.wealth.since.XXXX variables described above. The idea for this attribute is to measure how fast the economy surrounding each billionaire grew during the time period of interest.

At this point, I also created some binary dummy variables that I wanted to incorporate into the model but ultimately gave that up due to the complications of using binary variables in a clustering model. The four variables measure whether a billionaire inherited their wealth (wealth.is.inherited), if they are a founder of their company (company.is.founder), if their wealth was gained through the field of Finance (wealth.how.is.financial), and whether they are female (is.female).

Now satisfied with my calculated variables and having no more missing values, I moved on to adjusting the numerical attributes to make them fit for clustering. At first, all I did was z-score normalization for every numerical variable. However, after going on to do some initial clustering, it became clear that more adjustments were needed. I plotted some histograms and saw that a few input variables for my model were very highly skewed. In particular, wealth.worth.in.billions, pct.wealth.since.2001, and company.founded were in need of transformation. I tinkered with different transformations, trying to get a skew closer to zero, using the skewness() function from the "moments" package to evaluate each transformation. Wealth.worth.in.billions was inverse transformed, pct.wealth.since.2001 was negative inverse transformed, and company founded was negative log transformed. These transformations improved the respective skews considerably, and these transformed attributes were z-score scaled before returning to the model. Their final scaled distributions can be seen in the histogram below – pct.wealth.since.2001 obviously still did not resemble anything close to a normal distribution due to the huge number of new billionaires for whom 100% of wealth was acquired since 2001, but the model was able to produce good clusters regardless.

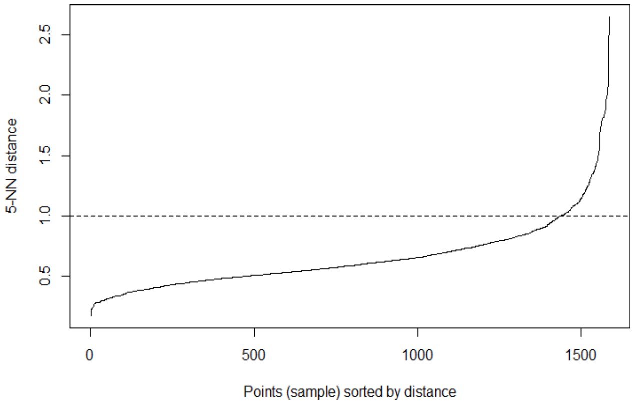


Modeling

In the end, 6 attributes made it into the final model: wealth.worth.in.billions, pct.wealth.since.2001, company.founded, demographics.age, location.gdp.2014, and pct.gdp.since.2001. I decided to exclude the 1996 data, as I was more interested in the recent developments and had concerns about multicollinearity between that and the 2001 data.

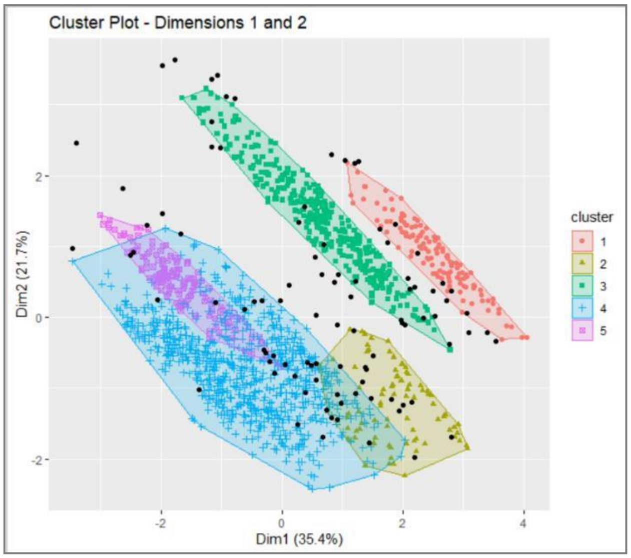
The two models I ended up trying were K-Means and DBSCAN. K-Means was chosen as more of a personal preference due to the concrete, unambiguous process of tuning hyperparameters. NbClust() tells you how many clusters to develop (4, in this case), and $avg.silwidth is a hard and fast number that tells you exactly how well or poorly your clusters separate the data. However, because many of the attributes included a relatively high number of outliers (in particular, pct.wealth.since.2001 and company.founded), I decided that a density-based clustering algorithm was more appropriate.

I used the KNNdistplot() function to determine the optimal value for eps, which turned out to be 1. Playing with the parameters led me to a minPts value of 10. This model produced 5 clusters, with 95 observations determined to be noise.

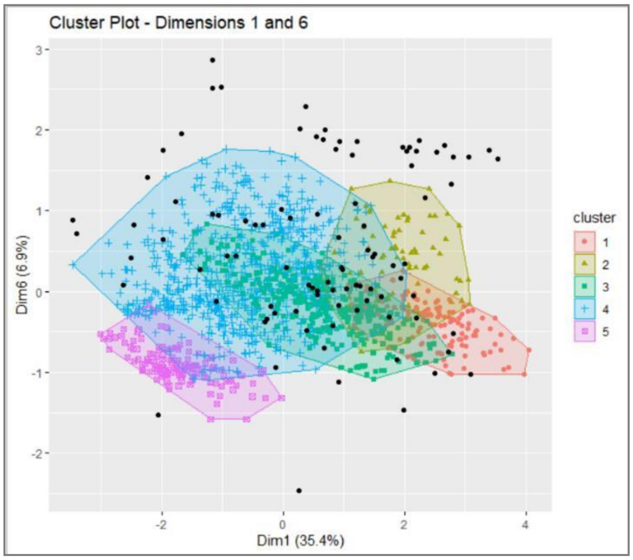


Evaluation

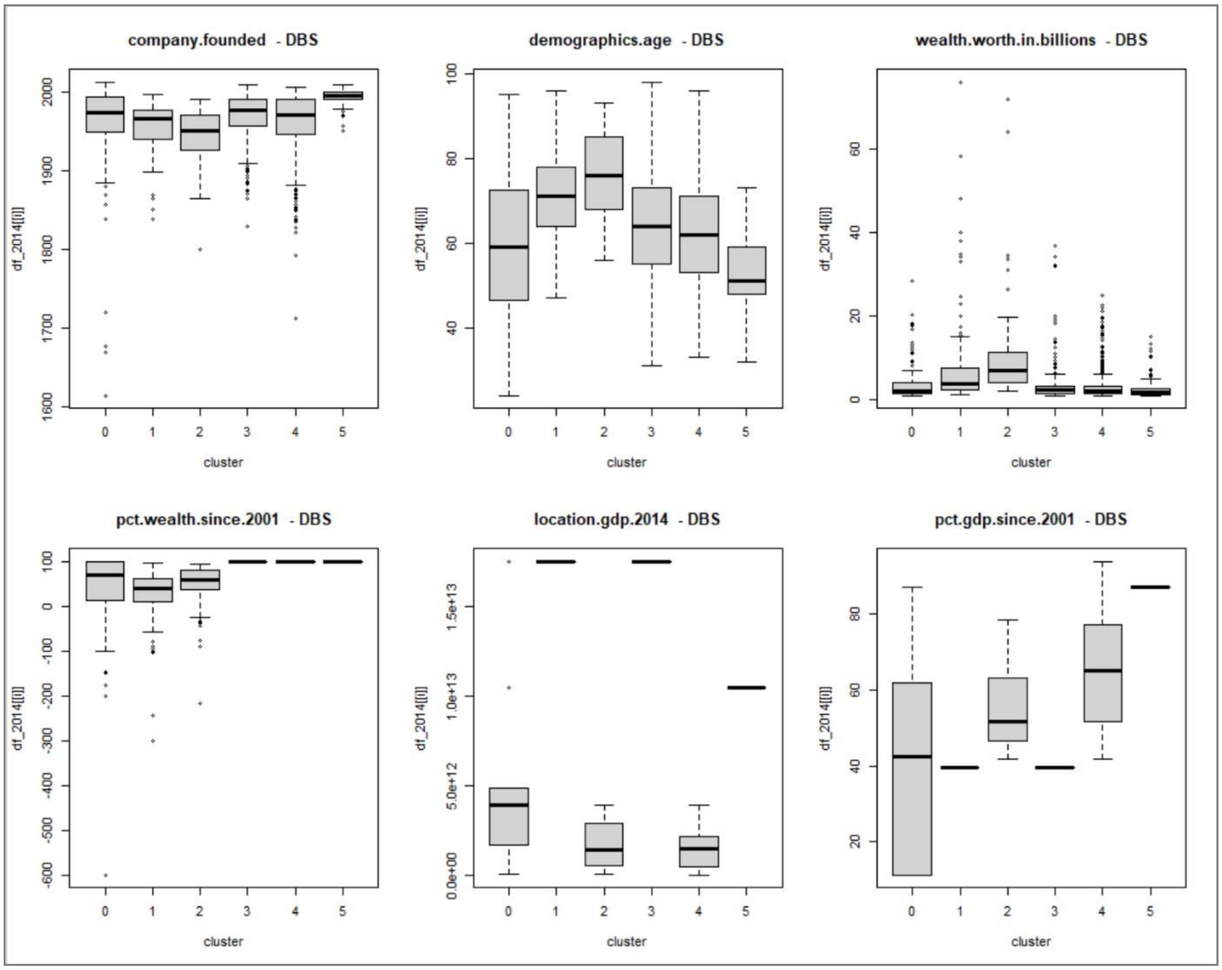
Unlike K-means clustering for which many metrics are calculated, is no concrete way to evaluate a DBSCAN model. The best we can do is a visual examination of the cluster plot output from the fvizcluster() function, seen to the right.



Before discussing the clusters themselves, it should be noted that a relatively large amount of variance is captured by the two principal components visualized on the x- and y-axes – a total of 57.1% - which indicates that this plot is a relatively good 2-dimensional representation of the data in n-dimensional space. In looking at the clusters themselves, clusters 1 and 3 stand out strongly for their tight density and being the only ones with no overlap. Cluster 4 is another obvious standout for its size alone. Capturing 800 rows, this cluster represents fully half of the dataset and can be considered characteristic of the “common billionaire”. Cluster 5 has significant visual overlap with the colossal cluster 4 but replacing Dim2 on the y-axis with Dim6 shows that this dimension is where the separation occurs between these two clusters. Unfortunately, Dim6 captures only 6.9% of the variance in the data, so this separation should be taken with some skepticism.



Examining the boxplots on the next page for each variable can tell us that the variables the model considered to be most important were wealth.worth.in.billions, pct.wealth.since.2001, location.gdp.2014, and pct.gdp.since.2001. Next, we will take a closer look at these variables to determine how the model separated these clusters.



Discussion

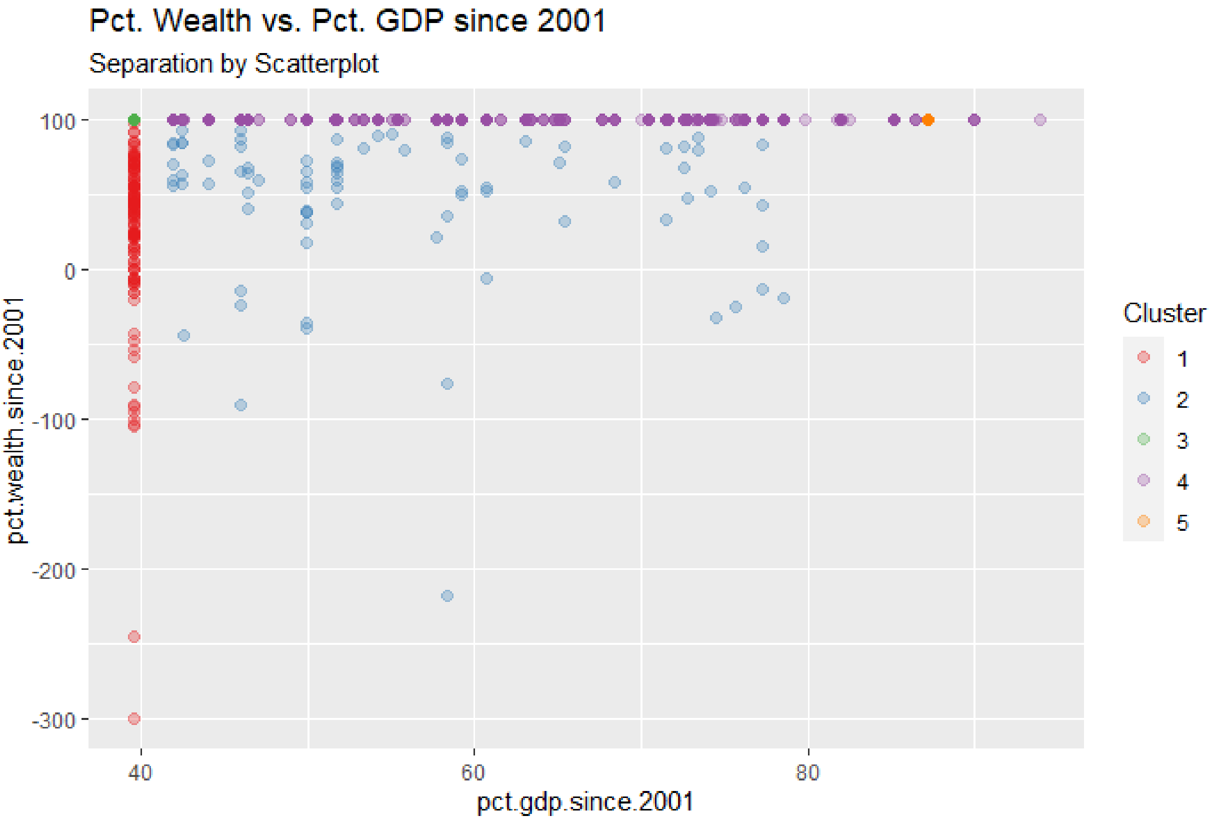
Right away, it is clear that location.gdp.2014 and pct.gdp.since.2001 (which are inherently the same for every billionaire from a given country) were critical for separating the most largely represented nationalities from countries with fewer billionaires. Most notably, clusters 1 and 3 are separated from the other clusters by the GDP attributes, with 100% of billionaires in these groupings being from the United States. Cluster 5 also shows a similar trait containing only billionaires from China, and all of the billionaires from China (aside from those designated as noise). That leaves the rest of the world to be represented by cluster 2 and the gigantic cluster 4.

Since clusters 1 and 3 are grouped together by geography, and 2 and 4 as well, the next step to characterize each cluster is to find out what the model uses to distinguish clusters 1 from 3 and 2 from 4. In both cases, the answer can be seen in the personal wealth attributes wealth.worth.in.billions and, more clearly, pct.wealth.since.2001. Clusters 3 and 4 are evidently characterized by the fact that 100% of their wealth in 2014 was acquired since 2001. In other words, they are new billionaires. Additionally, examining the wealth.worth.in.billions boxplot shows that they have far less wealth on average than clusters 1 and 2, with their respective Q_3 (3.30 and 3.23) being comparable to the Q_1 of clusters 1 and 2 (2.2 and 4.1). Consequently, while there are some new billionaires in clusters 1 and 2, far more of them seem to have had billionaire status in 2001 and their wealth is far higher on average.

Now we have enough information to give a basic characterization of each cluster. To summarize, we have:

- Cluster 1: Long-time American billionaires
- Cluster 2: Long-time non-American billionaires (wealthiest on average)
- Cluster 3: New American billionaires
- Cluster 4: New non-American billionaires (most common)
- Cluster 5: Chinese billionaires

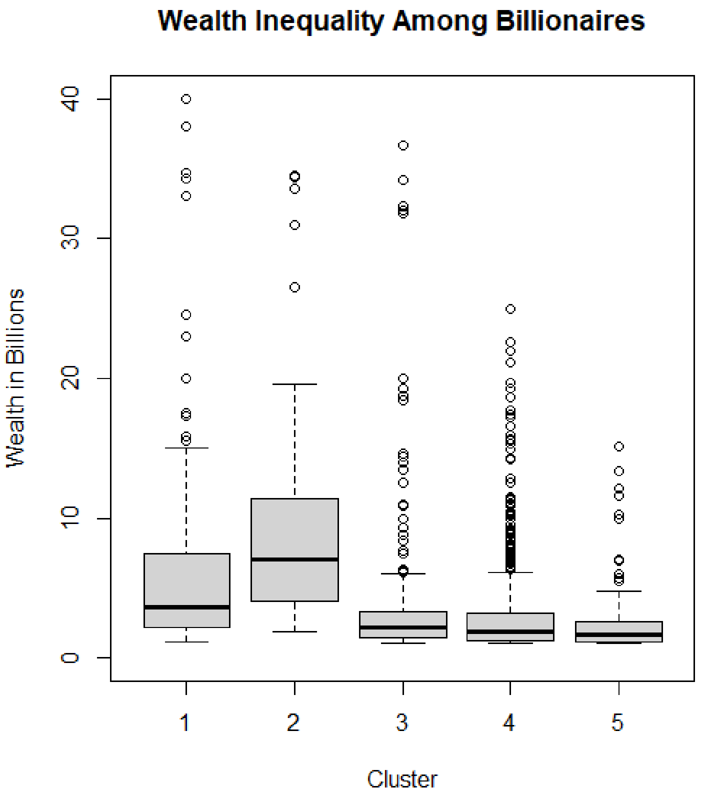
This separation can also be visualized in the scatterplot below with pct.wealth.since.2001 on the y-axis and pct.gdp.since.2001 on the x-axis. The United States, having the lowest GDP growth of this dataset, falls vertically along the far left edge of the plot, with the rest of the globe occupying the remaining space to the right. The old guard American billionaires of cluster 1 is represented by the column of red along the left side, with the American new money all existing inside of the opaque green dot at the top left corner. Global new billionaires can be found along the top line of purple, the new Chinese billionaires among them represented by the orange speck near the right side of the plot. And finally, the global veteran billionaires, the wealthiest people in the world, can be seen spread throughout the rest of the plot in blue.



Conclusions

I am satisfied with the way the data was clustered, although I do wish I could have gleaned more valuable insights. Certainly missing was an analysis of all the categorical variables regarding wealth type, company type, and others. Though I pored over histograms and bar charts of these variables by cluster, I was unable to learn anything terribly interesting. If I were to do this analysis again, the first thing I would do is find a way to incorporate these categoricals as inputs to the model – in my research I found a concept called Gower's distance which is a distance measure that can include binary and categorical variables, but I decided not to put it into practice so as to stay within the scope of the course and not bite off more than I can chew. I would also be interested in trying a hierarchical clustering algorithm, just to see if the model would prioritize grouping billionaires from the same geographical location or if it would consider the wealth value to be important enough to traverse the US border first.

What struck me most when doing this analysis is how, even among billionaires, there is an elite class. Clusters 1 and 2, representing the wealthiest people in the world, comprise only 13% of the billionaires represented in this dataset. This seems obvious in retrospect – considering the old saying that the difference between a million dollars and a billion dollars is about a billion dollars, it makes sense that that trend continues as billionaires become more numerous. This analysis proves Obi-Wan Kenobi right: there's always a bigger fish.


Wealth Inequality Among Billionaires
Wealth in Billions
Cluster