Dealing with Mickey Mouse degrees or what you shouldn't study if you care about a good life

Introduction

A few days ago, Nadhim Zahawi, the Secretary of State for Education in UK decided to take measures against universities due to concerns regarding employability and sustainability of possible graduates [1]. As society evolves, some degrees become obsolete while others that appear overnight don't offer favorable trajectories, demanding high education costs. For a future student, avoiding these traps might actually be the best thing to do, possibly even more important than having the highest grades possible.

With this idea in mind, both students and governments would benefit from statistical analysis done on university majors. This analysis would allow clustering university majors based on socio-economic characteristics, and enable both parties to take informed decision on what degree to pursue or what universities to cut from public funding due to low employability of graduates. Ultimately, both students and governments are interested in knowing what groups of majors have the highest unemployment rate and what universities offer the highest median income, in addition to preferable majors for each gender. As a result, the analysis in the following section will focus on answering these questions and offer scientific advice in this regard.

The data

The data used in this analysis consists of 172 observations and 17 numerical variables. Each observation represents one major, and includes socio-economic properties such as number of female graduates, unemployment rate after graduation, number of jobs that require college education or median income. The source for this data is the American Community Survey and represents majors surveyed between 2010-2012[2], and was initially used as part of an article appeared on FiveThirtyEight [3] that offered insights into picking college majors based on economic insights.

Technical details

For the purpose of this analysis, the two main methods that were used are principal component analysis (PCA) and k-means clustering. Principal component analysis is a method that helps reduce the number of features present in a dataset to a smaller dimension, while keeping as much variability in the data as possible. In general, PCA is useful for datasets that contain a high number of correlated features which can be combined into a subset of features that are uncorrelated, but maintain as much variability of initial features as possible.

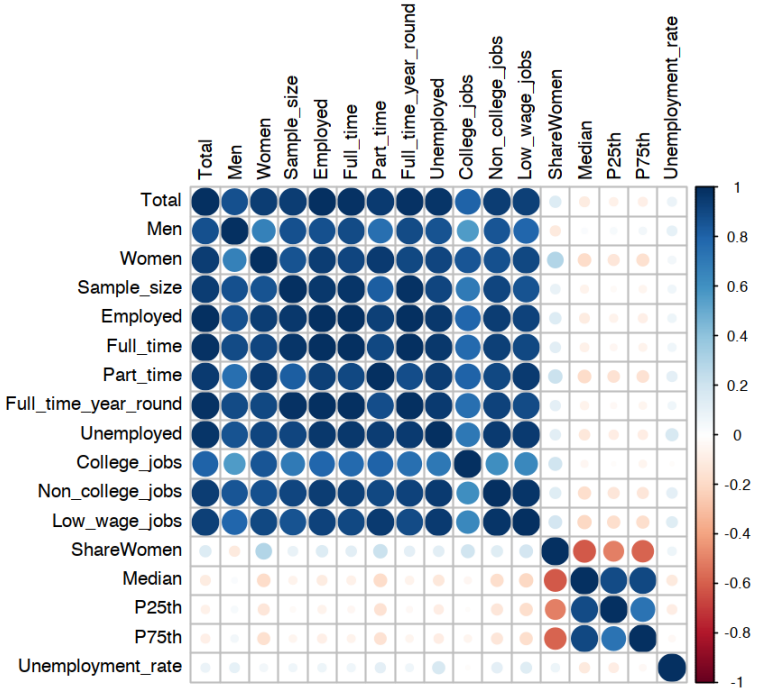

Figure 1: Correlation of features in the dataset

As we can see in Figure 1 there are two well-defined correlated groups of features, one containing 15 features that have a high correlation between them, and another one that contains 4 features. The only negative correlation that we can see is between the share of female graduates, and median, 25% and 75% quantiles, where higher the number of female graduates relates to lower incomes.

After applying PCA in this scenario, 83% of the variability in the initial data is represented by 2 components, while 3 components would represent 89% of variability. In order to pick the number of components visually, the scree plot in Figure 2 indicates that 2 components could be sufficient. From dimension 3 onwards, the percentage of variance explained is not dropping significantly anymore and these components together represent a small part of the variability

in the dataset.

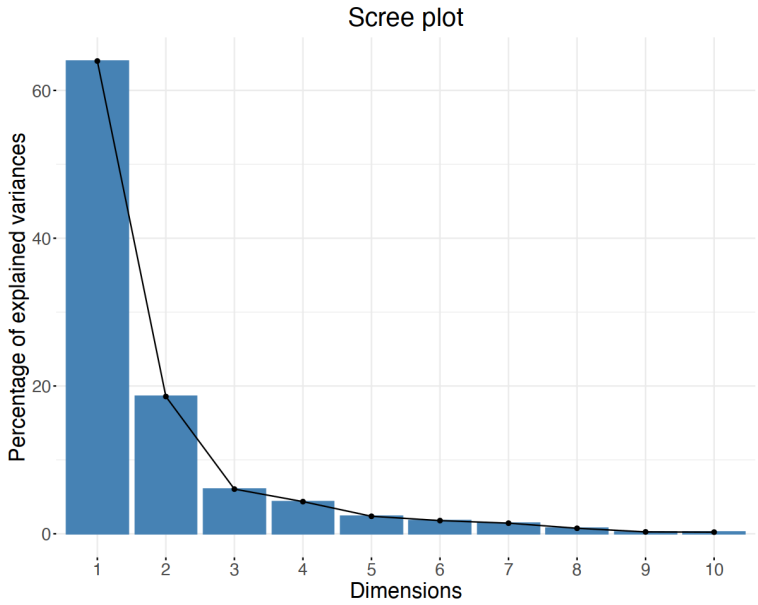

Figure 2: Percentage of variance explained by each dimension

Naturally, after selecting the first 2 PCAs, the question arises in understanding what these two components represent and how they relate to the initial features in the dataset. In Figure 3 the first component seems to be aligned with the first correlated group of features in figure 1, the higher the number of women, men, unemployment rate, the higher the value on the x-axis. The second PCA is correlated with the number of female graduates and inverse correlated with the three variables representing the income after university.

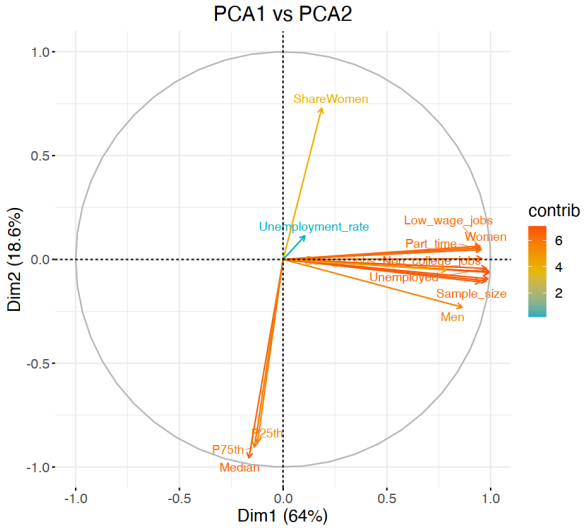

Figure 3: First two PCAs and feature direction

This is further evidence that the first two components are capable of representing the most important characteristics of the dataset, which were discovered during the initial exploratory analysis (via the correlation plot).

In order to be able to group the majors into clusters, the first two PCAs are used as input to a k-means clustering algorithm. K-means works by allocating points to a cluster based on the distance to the center of each cluster and selecting the cluster that yields the shortest distance. The algorithm works iteratively, by computing the cluster center and evaluates which cluster the points belong to, steps that are being repeated until no more significant changes occur. The algorithm requires the number of clusters to be specified beforehand or if not known, using visual tools such as silhouette or scree plots can help to find the appropriate number of clusters. In this particular scenario, neither scree plots nor the silhouette method gives satisfactory results, both suggesting that 2 clusters would be the best configuration. An alternative solution was considered in this scenario, using the fact that each major is part of one of the 16 major categories available. For a given number of clusters, from 2 to 16, the purity of each cluster was calculated using a Gini Index [5], where a pure cluster represents a cluster that contains only majors part of the same major category. From the 15 possible options for the number of clusters, 9 clusters represent the best mean Gini Index score.

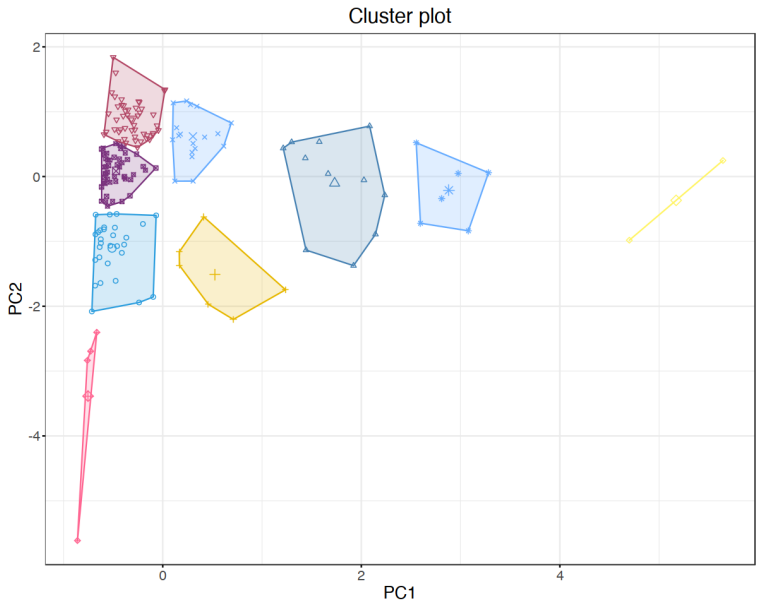

Figure 4: 9 polygons each representing one cluster

Results

The socio-economic features presented in the initial dataset only partially recover the 16 major categories existing in the initial dataset and the majority of the clusters are impure. Nevertheless, the results are worth exploring further, given that there are majors that are across major categories that are worth to be part of the same cluster. The clusters will be considered in order of increasing Gini Index (lower the Gini Index higher the purity).

Gini Index 0 - maximum purity

Major	Major category
Petroleum Engineering	Engineering
Mining and material Engineering	Engineering
Metallurgical Engineering	Engineering
Nuclear Engineering	Engineering

Table 1: Cluster with the highest purity

This cluster represents a subset of engineering majors that are probably the closest related together even if we don't consider socio-economic features. All the 4 majors are related by the fact that they all deal with highly valuable commodities, the extraction and processing of these commodities. In this regard, the clustering is so good that if somebody would be asked to handpick 4 majors that are related to each other from the entire dataset, there would be a high probability to pick these 4 majors.

Gini Index 0.44

Major	Major category
Computer science	Computers & Mathematics
Mathematics	Computers & Mathematics
Mechanical Engineering	Engineering
Electrical Engineering	Engineering
General Engineering	Engineering
Civil Engineering	Engineering

Table 2: Cluster with the highest purity

While this cluster has majors coming from two major categories, resulting in a higher Gini Index, the majors are all related to each other. The job prospects between them are generally

the same, all requiring a high level of mathematical understanding to be successful. In many situations, somebody studying mechanical engineering or electrical engineering can easily do a postgraduate in computer science or vice versa, and furthermore the job offerings for graduates from one major are likely accessible to candidates from any of the other degrees. As a conclusion, this cluster is expected as well, maybe not as closely related as the previous one, but not far apart either.

Gini Index 0.62 Compared to the other two clusters, this cluster contains 29 majors across 6 different major categories:

- **Engineering:** Naval Architecture, Chemical, Computer, Aerospace, Biomedical, Materials, Mechanics, Biological, Industrial and Manufacturing, Architectural, Electrical, Materials, Miscellaneous, Environmental, Engineering Technologies, Geological and Geophysical, Industrial production, Construction
- **Computer & Mathematics:** Computer and information systems, Information sciences, Mathematics and computer science
- **Business:** Actuarial science, Management Information Systems and Statistics, Operations, Logistics and E-commence
- **Industrial Arts:** Construction services, Military Technologies
- **Law & Public policy:** Court reporting, Public policy
- **Physical sciences**: Astronomy and astrophysics

Clearly these majors have much higher dissimilarities in terms of topics than the previous two clusters and contains some majors that are STEM while others that are humanities. While the business majors and the industrial arts ones are somehow related to the engineering or the mathematics ones, the majors from the law and public policy category aren't related by area of study. In this case, socio-economic factors are the ones that drive the clustering of unrelated majors together.

The rest of the clusters have a Gini Index ranging between 0.78 and 0.9 and having between 2 and 48 majors.

Results and summaries

Which cluster has the highest median income ? The cluster that contains the 4 engineering degrees: petroleum, mining and materials, metallurgical and nuclear engineering has a

median income of $80,000, which is $25,000 more than the second-highest cluster, but it also has the lowest percentage of female graduates, with only 13% and also the highest unemployment rate of 8%. This is definitely a high-risk high reward degree, where you can have a very good wage early in your career but also have higher chance than other majors to be unemployed. The median income of this cluster twice the median of all 172 majors available in the dataset.

Which cluster has the highest percentage of female graduates ?

References

[1] dailymail.co.uk/news/article-10631871/Education-Secretary-Nadhim-Zahawi-plans-crackdown-Mickey-Mouse-degrees.html *Education Secretary Nadhim Zahawi plans crackdown on 'Mickey Mouse' degrees - with universities required to publish drop-out rate and graduate job outcomes on every advert*

[2] census.gov/programs-surveys/acs/microdata.html *United States Census Bureau*

[3] FiveThirtyEight.com *American website that focuses on opinion poll analysis, politics, economics, and sports blogging*

[4] fivethirtyeight.com/features/the-economic-guide-to-picking-a-college-major *The Economic Guide To Picking A College Major*

[5] *An Introduction to Statistical Learning, Gini Index - James, Witten, Hastie, Tibshirani*

A Clusters