Clustering majors by socio-economic properties. A guide for students and governments

Erik-Cristian Seulean, 210001411, Group 5

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.

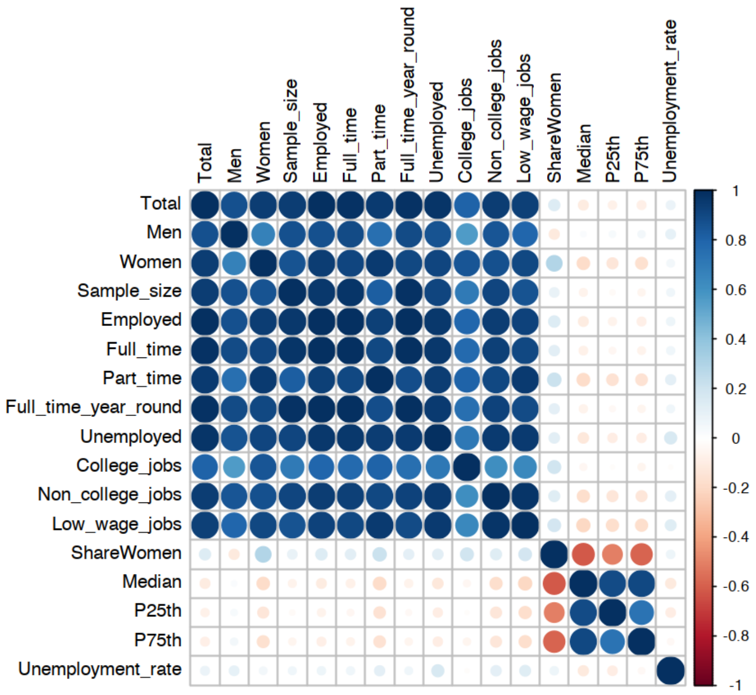


Figure 1: Correlation of features in the dataset

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 total variability in the dataset.

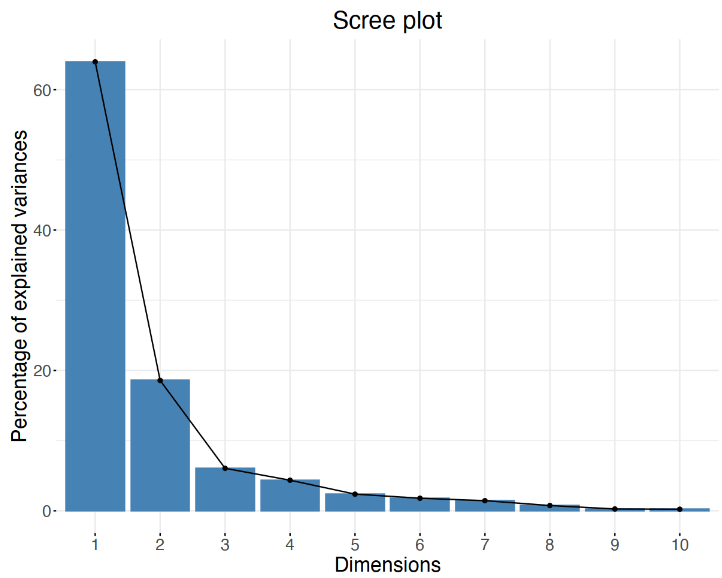


Figure 2: Percentage of variance explained by each dimension

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, denoted as Dim 1, on the x-axis, 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.

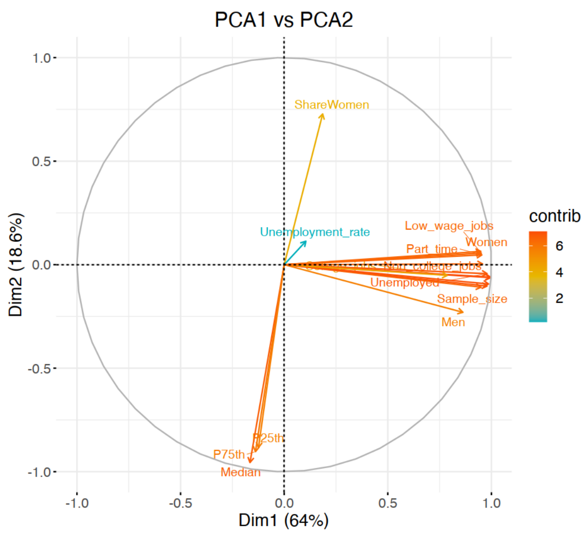


Figure 3: First two PCAs and feature direction

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. The reason that PCA is applied before using k-means is that it helps reduce the noise in the data and the dimension of the features, which could potentially speed up the clustering algorithm [4]. K-means works by allocating points to a cluster based on the distance to the center of each cluster and allocating 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 [6], 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.

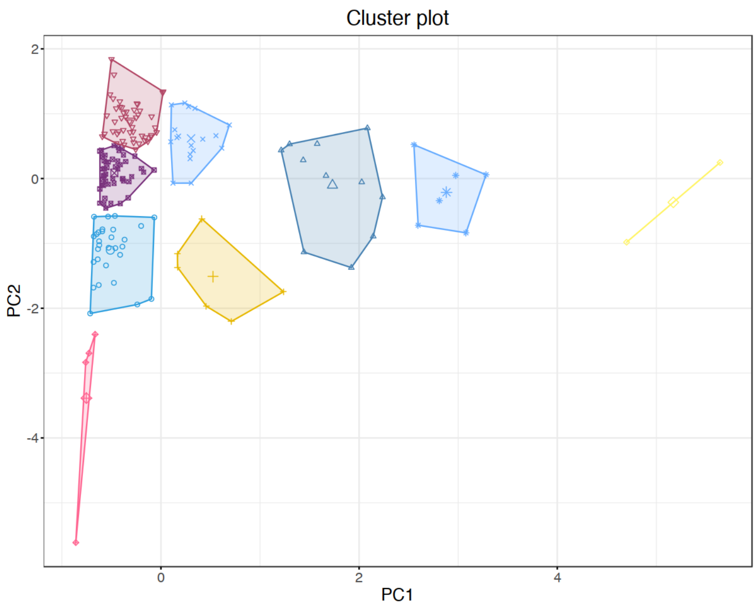


Figure 4: 9 polygons each representing one cluster

Results and summaries

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 probably should to be part of the same cluster.

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 to each other 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: As we can see in Figure 5 (Appendix) compared to the other two clusters, this cluster contains 29 majors across 6 different major categories.

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 share of women in this cluster is below 50% for almost all degrees, while the mean income is the second highest among all clusters.

The rest of the clusters have a Gini Index ranging between 0.78 and 0.9 and having between 2 and 48 majors. These are presented in the Appendix.

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 ? Cluster 8 in Table 11 represents the cluster with the highest percentage of female graduates, 71%. This cluster does not contain any science related degrees and also has the lowest median income, with $31,000, which is 25% lower than the median of all degrees. While the unemployment rate is lower than the mean unemployment among all majors, the wages are likely to lead to a salary gap between men and women due to a disproportionate number of men and women willing to study these topics. The only degree in this cluster that has a higher percentage of male students is Theology and religious vocations, which is expected.

Conclusions

The clustering strategy used separates clearly between STEM degrees and humanities in most clusters. If graduation income is a concern, students should pick degrees from clusters that have most majors from STEM. Similarly, governments should consider making these degrees more atractive for female students as the income disparity is driven by disproportionate percentages of male and female students in clusters of majors that have a higher than average income. Cluster 9 in Figure 13 offer the highest range of topics between majors while keeps the proportion of female to male students close to 1 and has median income close to the median of all majors. This represents the most balanced choice, that offers a bit of everything and students should consider if any of the majors in this cluster are interesting enough.

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] https://ranger.uta.edu/ chqding/papers/KmeansPCA1.pdf *K-means clustering via Principal Component Analysis - Chris Ding, Xiaofeng He*

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

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

A Clusters

Major	Major category
Metallurgical engineering	Engineering
Mining and mineral engineering	Engineering
Nuclear engineering	Engineering
Petroleum engineering	Engineering

Table 3: Cluster 1

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

Table 4: Cluster 2

Major	Major category
Actuarial science	Business
Management information systems and statistics	Business
Operations logistics and e-commerce	Business
Computer and information systems	Computers & mathematics
Information sciences	Computers & mathematics
Mathematics and computer science	Computers & mathematics
Aerospace engineering	Engineering
Architectural engineering	Engineering
Biological engineering	Engineering
Biomedical engineering	Engineering
Chemical engineering	Engineering
Computer engineering	Engineering
Electrical engineering technology	Engineering
Engineering mechanics physics and science	Engineering
Engineering technologies	Engineering
Environmental engineering	Engineering
Geological and geophysical engineering	Engineering
Industrial and manufacturing engineering	Engineering
Industrial production technologies	Engineering
Materials engineering and materials science	Engineering
Materials science	Engineering
Miscellaneous engineering	Engineering
Naval architecture and marine engineering	Engineering
Construction services	Industrial arts & consumer services
Military technologies	Industrial arts & consumer services
Court reporting	Law & public policy
Public policy	Law & public policy
Astronomy and astrophysics	Physical sciences
Physics	Physical sciences

Table 5: Cluster 3

Major	Major category
Business management and administration	Business
Psychology	Psychology & social work

Table 6: Cluster 4

Major	Major category
Biology	Biology & life science
General business	Business
Marketing and marketing research	Business
Communications	Communications & journalism
Nursing	Health
English language and literature	Humanities & liberal arts

Table 7: Cluster 5

Major	Major category
Commercial art and graphic design	Arts
Accounting	Business
Finance	Business
Elementary education	Education
General education	Education
History	Humanities & liberal arts
Physical fitness parks recreation and leisure	Industrial arts & consumer services
Criminal justice and fire protection	Law & public policy
Economics	Social science
Political science and government	Social science
Sociology	Social science

Table 8: Cluster 6

Major	Major category
Drama and theater arts	Arts
Film video and photographic arts	Arts
Fine arts	Arts
Music	Arts
Hospitality management	Business
Advertising and public relations	Communications & journalism
Journalism	Communications & journalism
Mass media	Communications & journalism
Architecture	Engineering
Treatment therapy professions	Health
Anthropology and archeology	Humanities & liberal arts
Foreign language studies	Humanities & liberal arts
Liberal arts	Humanities & liberal arts
Philosophy and religious studies	Humanities & liberal arts
Family and consumer sciences	Industrial arts & consumer services
Chemistry	Physical sciences
Multi-disciplinary or general science	Physical sciences
Social work	Psychology & social work

Table 9: Cluster 7

Major	Major category
Animal sciences	Agriculture & natural resources
Miscellaneous agriculture	Agriculture & natural resources
Studio arts	Arts
Visual and performing arts	Arts
Ecology	Biology & life science
Environmental science	Biology & life science
Miscellaneous biology	Biology & life science
Physiology	Biology & life science
Zoology	Biology & life science
Human resources and personnel management	Business
Art and music education	Education

Early childhood education	Education
Educational administration and supervision	Education
Language and drama education	Education
Library science	Education
Mathematics teacher education	Education
Physical and health education teaching	Education
Science and computer teacher education	Education
Secondary teacher education	Education
Social science or history teacher education	Education
Special needs education	Education
Teacher education: multiple levels	Education
Communication disorders sciences and services	Health
Community and public health	Health
General medical and health services	Health
Health and medical administrative services	Health
Health and medical preparatory programs	Health
Miscellaneous health medical professions	Health
Nutrition sciences	Health
Area ethnic and civilization studies	Humanities & liberal arts
Art history and criticism	Humanities & liberal arts
Composition and rhetoric	Humanities & liberal arts
Humanities	Humanities & liberal arts
Intercultural and international studies	Humanities & liberal arts
Linguistics and comparative language and literature	Humanities & liberal arts
Other foreign languages	Humanities & liberal arts
Theology and religious vocations	Humanities & liberal arts
Cosmetology services and culinary arts	Industrial arts & consumer services
Multi/interdisciplinary studies	Interdisciplinary
Geosciences	Physical sciences
Clinical psychology	Psychology & social work
Counseling psychology	Psychology & social work
Educational psychology	Psychology & social work
Human services and community organization	Psychology & social work

Miscellaneous psychology	Psychology & social work
Criminology	Social science
General social sciences	Social science
Interdisciplinary social sciences	Social science

Table 11: Cluster 8

Major	Major category
Agricultural economics	Agriculture & natural resources
Agriculture production and management	Agriculture & natural resources
Forestry	Agriculture & natural resources
General agriculture	Agriculture & natural resources
Natural resources management	Agriculture & natural resources
Plant science and agronomy	Agriculture & natural resources
Soil science	Agriculture & natural resources
Miscellaneous fine arts	Arts
Biochemical sciences	Biology & life science
Botany	Biology & life science
Cognitive science and biopsychology	Biology & life science
Genetics	Biology & life science
Microbiology	Biology & life science
Molecular biology	Biology & life science
Neuroscience	Biology & life science
Pharmacology	Biology & life science
Business economics	Business
International business	Business
Miscellaneous business & medical administration	Business
Applied mathematics	Computers & mathematics
Communication technologies	Computers & mathematics
Computer administration management and security	Computers & mathematics
Computer networking and telecommunications	Computers & mathematics
Computer programming and data processing	Computers & mathematics

Statistics and decision science	Computers & mathematics
Miscellaneous education	Education
School student counseling	Education
Engineering and industrial management	Engineering
Mechanical engineering related technologies	Engineering
Miscellaneous engineering technologies	Engineering
Medical assisting services	Health
Medical technologies technicians	Health
Pharmacy pharmaceutical sciences and administration	Health
United states history	Humanities & liberal arts
Electrical, mechanical, and precision technologies	Industrial arts & consumer services
Transportation sciences and technologies	Industrial arts & consumer services
Pre-law and legal studies	Law & public policy
Public administration	Law & public policy
Atmospheric sciences and meteorology	Physical sciences
Geology and earth science	Physical sciences
Nuclear, industrial radiology, and biological technologies	Physical sciences
Oceanography	Physical sciences
Physical sciences	Physical sciences
Industrial and organizational psychology	Psychology & social work
Social psychology	Psychology & social work
Geography	Social science
International relations	Social science
Miscellaneous social sciences	Social science

Table 13: Cluster 9

[ ]:
```
require(ggplot)
require(tidyverse)
require(ggcorrplot)
require(ggthemes)
require(cluster)
require(corrplot)
require(factoextra)
require(cowplot)
```

[2]:
```
palette = c(
    "#2E9FDF", "#4782b3", "#E7B800",
    "#66acff", "#fff566", "#b34766",
    "#7a327d", "#66acff", "#ff6692",
    "#b3ab47", "#ffb3d7", "#66faff",
    "#7d7632", "#00AFBB", "#002db3",
    "#ff0000"
)
```

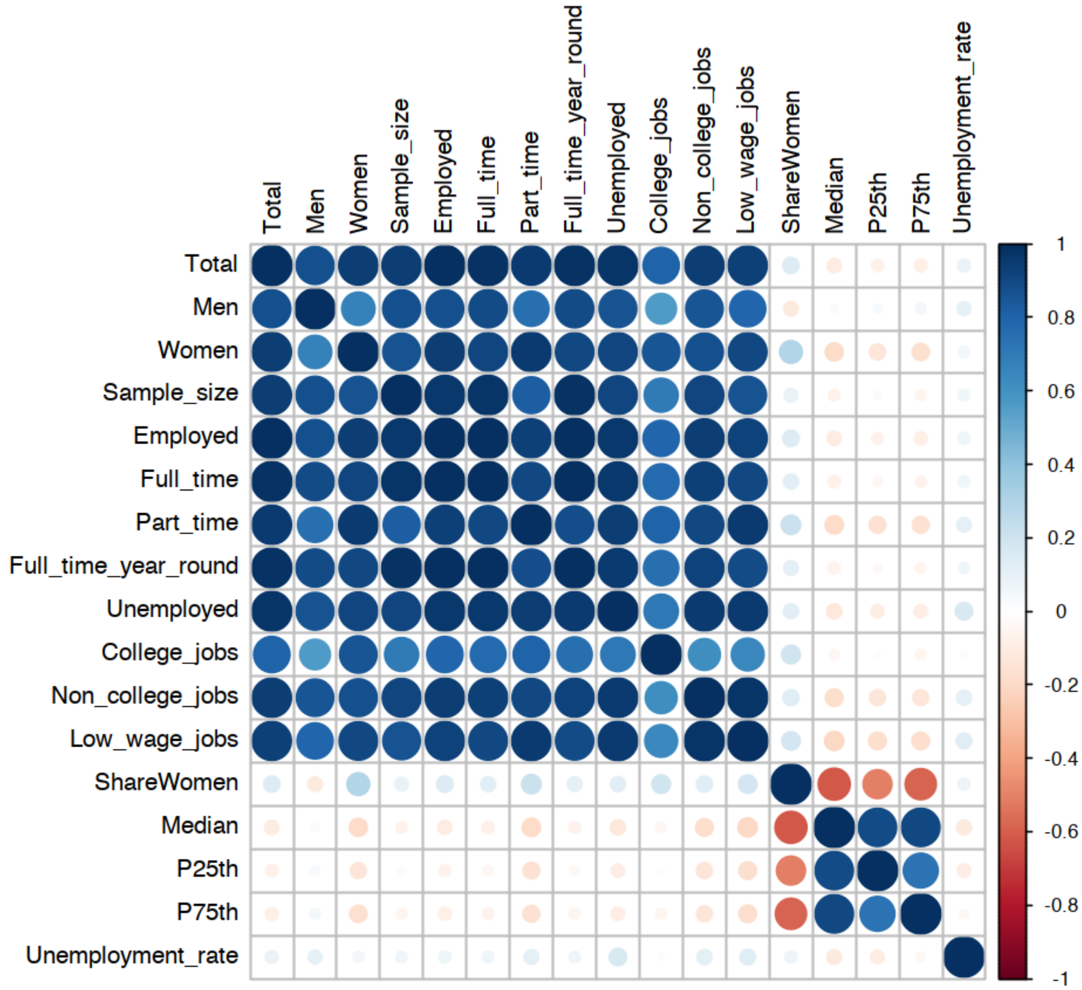
0.0.1 Drop categorical and index column from PCA

[3]:
```
data = read.csv("college.csv", header=TRUE, sep=",")
```

[4]:
```
data = data %>% drop_na()
subdata = data[-c(1, 2, 3, 7)]
subdata = subdata
```

0.0.2 Create correlation plot to check if PCA is worth applying

[5]:
```
options(repr.plot.width=8, repr.plot.height=8)
# set income related variables to the end, in order to improve on
# the visual aspect
subdata = subdata %>% relocate("ShareWomen", .after = last_col()) %>%
    relocate("Median", .after = last_col()) %>%
    relocate("P25th", .after = last_col()) %>%
    relocate("P75th", .after = last_col()) %>%
    relocate("Unemployment_rate", .after = last_col())
corrplot(cor(subdata), method="circle", tl.col = "black")
```



Groups of highly correlated variables that will be suitable for dimensionality reduction. Some of the existing features are computed from others. In this case the high correlation makes sense, but others, such as Share of women and median income have a negative correlation mostly as a result of socio-economic factors rather than feature engineering.

[6]:
```
# Apply pca to the data and specify that the features should
# be scaled and centered
pca = prcomp(subdata, scale=TRUE, center=TRUE)
```

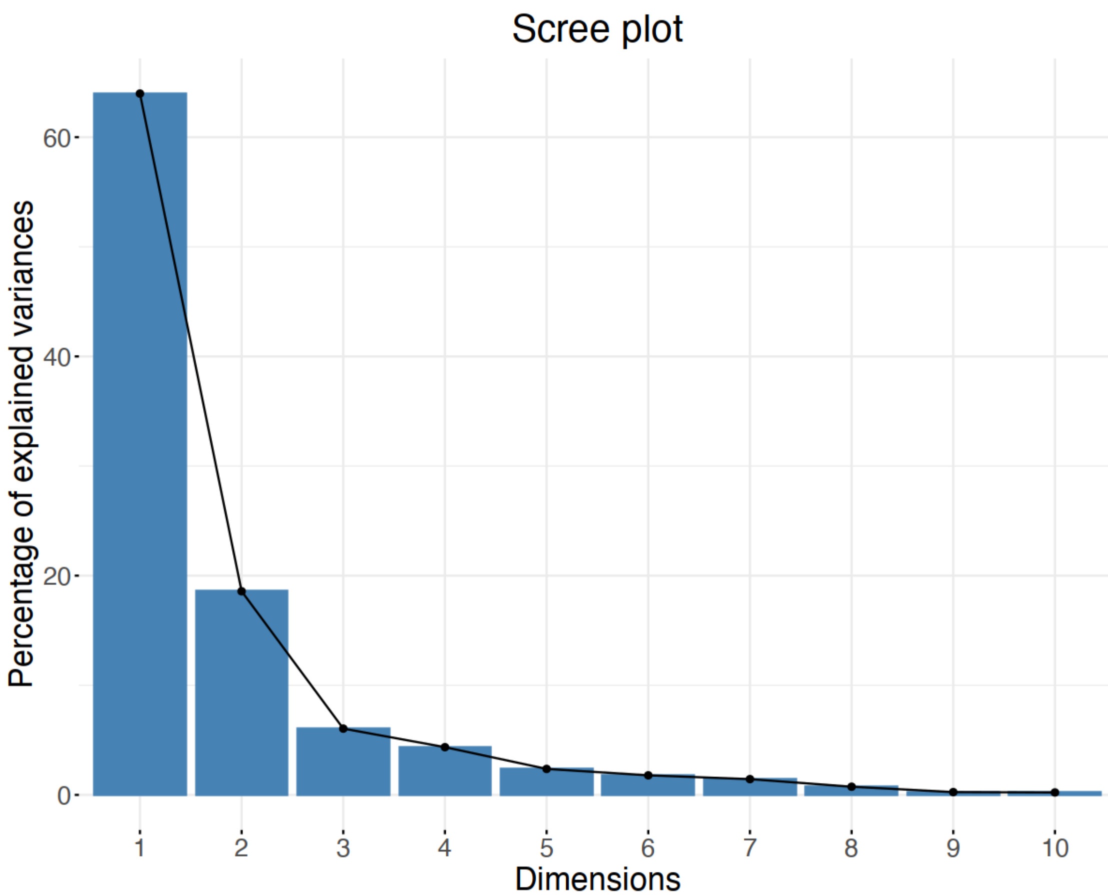
0.0.3 Variance explained by the first 3 principal components

[7]:
```
lambdas = pca$sdev^2
print(paste("Variance explained by first 2 components", round(sum(lambdas[1:2])/
  ↪sum(lambdas), 2)))
```

```
[1] "Variance explained by first 2 components 0.83"
```

[8]:
```
# Plot screeplot to help select how many PCAs to keep
options(repr.plot.width=10, repr.plot.height=8)
```

```
fviz_eig(pca) + theme(
    plot.title = element_text(hjust = 0.5),
    text = element_text(size = 20)
)
```

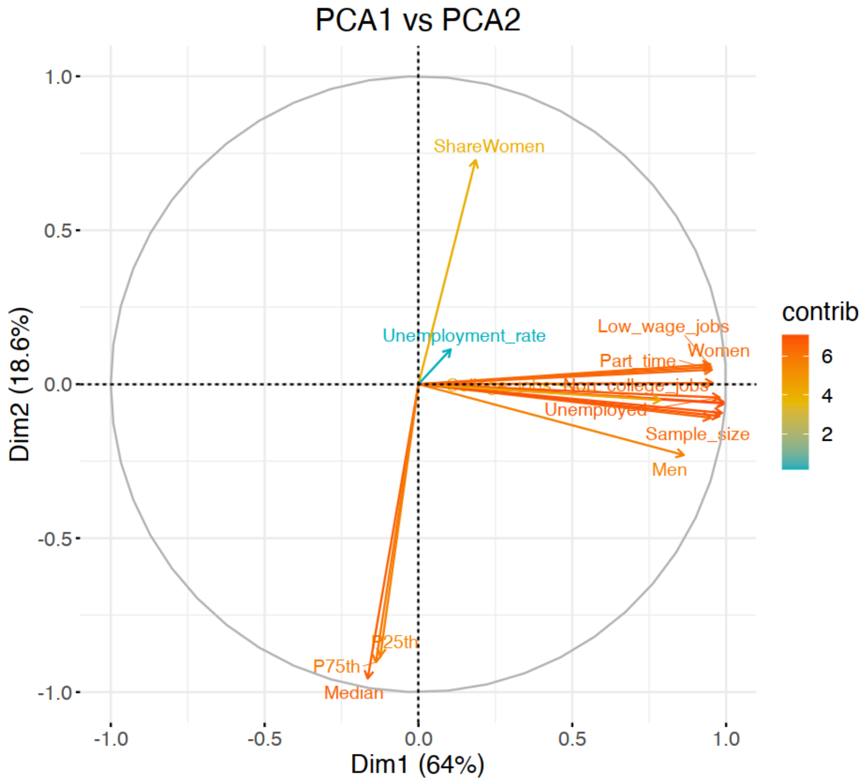


This suggests that first 2 PCA components should probably be kept. Third one has an extra 6% of explained variance. Some additional analysis can be done to verify if clustering has more informative results with 3 PCAs instead of 2.

0.0.4 Plot first two PCAs and existing features

[9]:
```
options(repr.plot.width=10, repr.plot.height=7)
factoextra::fviz_pca_var(pca,
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE,
             title = "PCA1 vs PCA2") + theme(
    plot.title = element_text(hjust = 0.5),
    text = element_text(size = 16)
)
```

```
Warning message:
“ggrepel: 7 unlabeled data points (too many overlaps). Consider increasing
max.overlaps”
```



[10]:
```
# Function to compute the Gini Index of a cluster
get_gini = function(clusters) {
    grouped = clusters %>% group_by(category) %>% count()
    grouped["percentage"] = grouped["n"] / sum(grouped["n"])

    return(sum(grouped["percentage"]  * (1 - grouped["percentage"])))
}
```

[11]:
```
# This is my St Andrews ID, use it for reproductibility
set.seed(210001411)

# Store best configuration for the clusters
best_gini = 1
best_configuration = -1
best_km = NA
ginies_per_cluster = c()
best_clusters = NA

# For each number of clusters selected
# Calculate the gini index of each cluster
# and find the mean gini index for a particular
# number of clusters between all clusters

for(nr_clusters in seq(2, 16)) {
    res.km = kmeans(pca$x[1:172, 1:2], nr_clusters, nstart=20, iter.max=500)
    clusters = data.frame(cluster=res.km$cluster, major=data$Major, category =␣
↪data$Major_category)
```

```
    clusters = clusters[order(clusters$category), ]

    ginies = c()
    total_gini = 0

    for(i in seq(1:nr_clusters)) {
        # Calculate gini index for each cluster
        cluster_gini = get_gini(clusters[clusters["cluster"] == i, ])
        ginies = c(ginies, cluster_gini)
        total_gini = total_gini + cluster_gini
    }

    mean_gini = total_gini/nr_clusters
    if(mean_gini < best_gini - 0.05) {
        # If this gini is significantly improving the
        # best configuration so far, store it.
        # If the improvement is not large enough,
        # avoid storing a very high number of clusters

        best_gini = mean_gini
        best_configuration = nr_clusters
        best_km = res.km
        ginies_per_cluster = ginies
        best_clusters = clusters
    }
}
print(paste("Best gini index", round(best_gini, 2), "and number of clusters",␣
 ↪best_configuration))
gini_by_cluster = data.frame(cluster = seq(1, best_configuration), gini =␣
 ↪ginies_per_cluster)
gini_by_cluster = gini_by_cluster[order(gini_by_cluster$gini),]
```

```
[1] "Best gini index 0.64 and number of clusters 9"
```

[13]:
```
# Attach a column for the cluster index in the initial dataset
names(best_clusters)[names(best_clusters) == 'major'] = 'Major'
data = merge(x=data,y=best_clusters[-c(3)],by="Major")
```

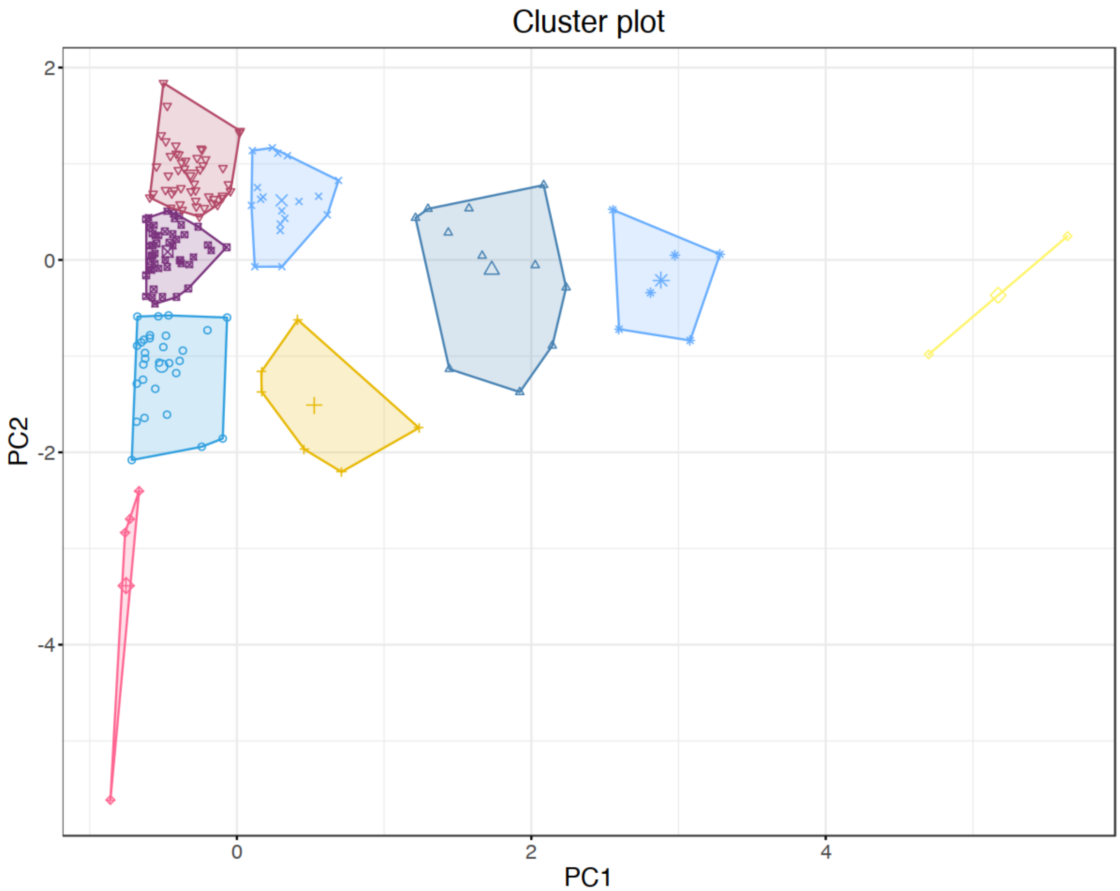
[ ]:
```
# Find summaries for each cluster (in this case the mean)
data %>% select(-Major, -Major_category) %>% group_by(data$cluster) %>%␣
 ↪summarise(across(everything(), mean))
```

0.0.5 Visualise polygons of clusters in 2D using the first two PCAs

Alternatively the data can be the initial datapoints and pairs of features from it.

[15]:
```
options(repr.plot.width=10, repr.plot.height=8)
fviz_cluster(best_km, data = pca$x[1:172, 1:2],
             palette = palette,
             geom = "point",
             ellipse.type = "convex",
             ggtheme = theme_bw()
             ) + theme(
```

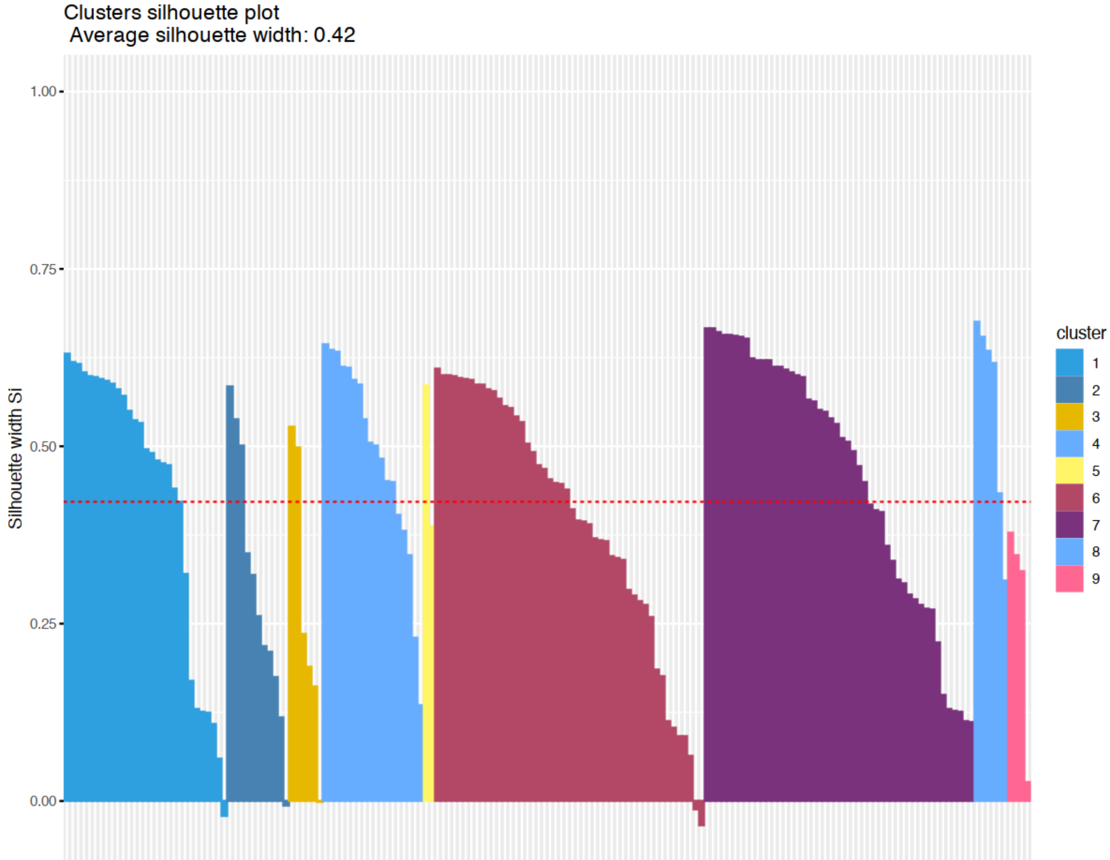
```
    plot.title = element_text(hjust = 0.5),
    text = element_text(size = 16),
    legend.position="none"
)
```



0.0.6 Find if number of clusters can be picked using sillhouette plots

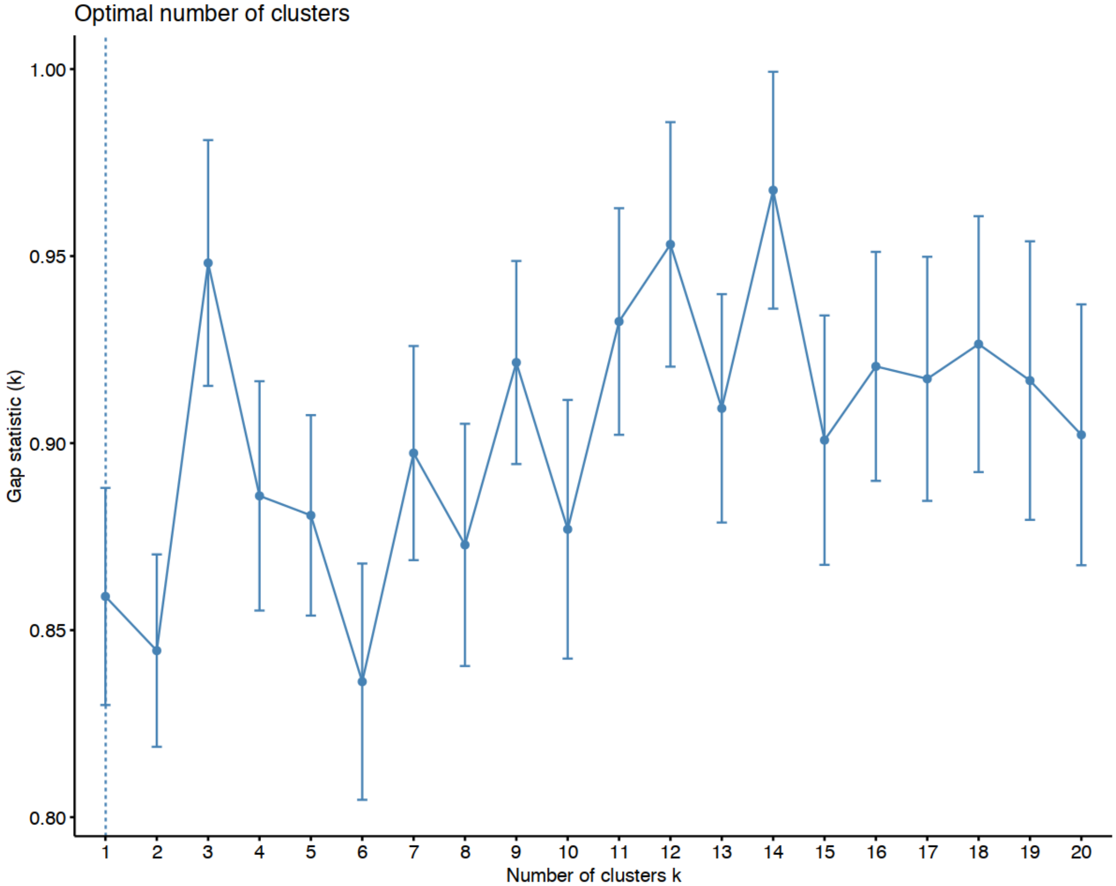
[16]:
```
sil <- silhouette(x = best_km$cluster, dist = dist(pca$x[1:172, 1:2]))
fviz_silhouette(sil) +
scale_fill_manual(values = palette) +
scale_color_manual(values = palette)
```

```
  cluster size ave.sil.width
1       1   29          0.43
2       2   11          0.30
3       3    6          0.27
4       4   18          0.49
5       5    2          0.49
6       6   48          0.39
7       7   48          0.46
8       8    6          0.55
9       9    4          0.27
```



0.0.7 Does gap metric suggest a better number of clusters ?

[18]:
```
fviz_nbclust(x = pca$x[1:172, 1:2], FUNcluster = kmeans, method = "gap", k.max =
 ↪20)
```


Optimal number of clusters
Gap statistic (k)
1.00
0.95
0.90
0.85
0.80
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Number of clusters k