MACS 40800 - UML Project

Major Expectations in College

Cao, Pacheco, Shen, & Shi

December 9th, 2019

Contribution Statement

Each member of the group contributed equally to the work product of this project.

Introduction

Choosing a major in college is among the most formative and significant decisions that one can make of both education and one's life, because it strongly guides not only career prospects, but also the kinds of interactions and values that one comes to encounter along the way. It is in this sense that the deciding of one's major is significant in more than one way: it is an anticipation towards the future, a selection of one's vocation on the basis of economic promise and life prospects – and it is an indicator of the past, an expression of who one is, along with their attendant values and interests (Beggs & Bantham, 2008).

That this moment is so pivotal to one's "educational destiny" has subsequently attracted an equal measure of research over the matter, ranging from a search for general trends and outcomes such as the relationship between education and economic growth (Psacharopoulos, 1994), to an outline of the particulars in which individual attributes like uncertainty predispose the student towards particular kinds of majors (Porter & Umbach, 2006; McIlveen, Beccaria, & Burton, 2013; Balsamo, Lauriola, & Saggino, 2013). These lines of research look at past, present, and future aspects of deciding on one's major, including but not limited to what kind of effect one's demographic and background has on their choice (Dickson, 2010), what kinds of interests and personality traits drive the decision towards one major rather than another (Wiswall & Zafar, 2014), and what future considerations like satisfaction are relevant to the choice (Milsom & Coughlin, 2015). These lines of research, when taken in tandem, inform our understanding of what goes into and comes out of these majors while highlighting their significances to the student who is making the decision.

While much of this research has focused on exploring relationships between particular variables – for example, the relationship between demographics and choosing a STEM major (Moakler & Kim, 2014) – there is a paucity of high-level overviews which examine a diverse set of variables and their relations to one another all at once. Such an overview would be important for both a theoretical and a practical reason. The theoretical reason is that much literature has focused on the relationship between the choice of major and some independent variable, but not as much attention as been paid to potential covariation between those variables under study. Uncovering underlying patterns between such variables of interest can not only help point to explanations as to why they influence one's choice of major, but such interactions can help establish cross-talk between otherwise-disparate lines of research. The practical reason concerns the situation of the student: the uncertain student would find it helpful to get an initial grasp on what the landscape of majors is like, and what the affiliated tendencies are for each major – for instance, what the backgrounds of the people who go into it are, or what kinds of economic prospects they are likely to face.

In line with the student's situation of self-exploration, we propose an exploratory data analysis into the field of education as it concerns the individual: more specifically, we look for trends among many majors using various variables. These variables will include other education-related factors, in addition to occupational, income-based, and demographic factors. Our approach will involve a combination of k-means clustering and principal components analysis (PCA); the goal of our approach contains two objectives: an exploration into the majors themselves, along with the variables of interest. The first objective will be accomplished by means of clustering, in which we uncover natural groupings among majors in order to visualize whatever affinities they may have. The second objective will be accomplished by means of PCA, in which

we explore relationships among the variables of interest, such as job satisfaction and income. By uncovering potential trends, we hope to shed light on what kinds of expectations or prospects may be affiliated with a certain choice of major, along with how majors may be similar or dissimilar to one another. An example of a pattern that we may expect to see are relationships between demographic and socioeconomic standing for some given major, such as a high proportion of high-earning males within engineering.

Empirical Strategies

Data

For this project, we utilized the General Social Survey (GSS) data. GSS is a national-scale sociological survey that collects and examines the trends in the attitude, behavior and experience of the United States population across different aspects of human life. It is created by the National Opinion Research Center (NORC) at the University of Chicago in 1972 and has since been collecting data that enable countless important studies in the field of social science (NORC). To ensure a better representation of the broader population, GSS implements stratified multistage area probabilistic sampling to select households and individuals for face-to-face interviews.

The primary sampling unit is standard metropolitan statistical area from the NORC database, stratified by region, age and race. At the second stage of the sampling process, blocks are selected with probabilities proportional to size, of which researchers then go on canvas the houses and conduct interviews until quotas have been filled (cf. General Social Survey). We describe the GSS sampling process in detail because it is important for ensuring the validity of our data analysis

and the robustness of the project design in order to prepare for the following exploratory strategies which require further data manipulation.

The complete GSS dataset since 1972 contains 64814 rows of observations with 6108 variables derived from each survey question. Due to the fact that not all questions are asked every year, we filtered and created a new dataset that only includes observations from 2016 onwards in order to study the most recent pattern and keep consistency in our variables. Since the focus of our study is major, we dropped rows that either did not report the information or recorded degrees lower than college-level (e.g. anything below an associate's degree or equivalent). Based on previous literature, we identified four main aspects of socioeconomic features that are highly correlated to people's choice of specialization: broader education, income-specific measures, demographic background, as well as career/occupation (Moakler & Kim, 2014; Milsom & Coughlin, 2015). We selected 20 variables which were representative of our 4 categories using keyword search function on the GSS Data Explorer page under each aspect (for example, entering "education", "school", "study", "course", etc. to find education indicators) and cross-referenced each variable with the codebook to understand the text definition and its scale of measurement.

A full list of these variables is included in Table 1 below. One particularly interesting variable to note is the occupational prestige score, denoted as "PRESTG10". It is based on the 2010 census occupation classification. The prestige score refers to an index between 0 and 100 used by sociologists to describe the relative social class positions people have based on their reported occupation. The measurement considers the admiration and respect for particular occupations which people reported according to previous opinion surveys (Stevens & Featherman, 1981). This standard prestige score is a simple mean value of ratings for each occupation category,

transformed into a scale of 0 (the least prestigious) to 100 (the highest). Our initial stage of filtering resulted in 1970 observations with 20 variables.

For our next pre-processing step, we recoded many of the categorical variables to make them amenable to quantitative analysis. For income, which was originally coded as brackets ranging from "under $1000" to "$170000 or above", we converted it into continuous variable by assigning numerical values to each nominal category. We used the median for each income bracket, e.g. all responses under "$20000 to $22499" were coded as 21250, since it was in line with prior sociological research that worked with GSS income data, and we treated everyone who reported income higher than $170000 as if their income was $170000 since few observations were at the top open interval. For variables that asked attitudinal questions, we converted those categories into ordinal likert scales. For instance, job satisfaction was recoded to a scale from 1 to 7, with a higher score indicating more current job satisfaction. In addition, we also converted categorical variables such as sex and whether the individual works in full-time position into binary variables. The detailed data cleaning procedure and code for the relevant variables can be found in the Appendix E. The last step of data preprocessing aggregated data on all individuals within their respective majors, making our unit of analysis the majors themselves (which each represent a population of individuals) rather than the individuals themselves. The final dataset thus contained 60 rows, i.e. 60 majors, with a feature space of 16 dimensions.

Table 1. Variables of Interest

Variable Name	***Variable Type***	***Description***	***Recode***
Educ	Integer	Highest year of school completed	N
Degree	Categorical	R's highest degree	Converted to ordinal
Major1	Categorical	College major	N
Colscinm	Integer	Number of college-level science courses R have taken	N

Rincom16	Categorical	R's income based on 2016 Census Bracket	Converted to numerical
Finrela	Categorical	Opinion of own family income level compared to American families in general	Converted to ordinal
Sex	Categorical	Sex	Converted to Boolean (isMale)
Age	Integer	Age	N
Race	Categorical	Race	Converted to Boolean (isWhite)
Sei10	Numerical	R's socioeconomic index based on 2010 census	N
Prestg10	Integer	R's occupational prestige score based on 2010 census	N
Wrkstat	Categorical	Labor force status (full-time, part-time, temporarily not working, unemployed laid off, etc.)	Converted to Boolean (isFulltime)
Hrs1	Integer	Number of hours worked last week	N
Wrkslf	Categorical	R self-employed or works for somebody	Converted to Boolean (isSelfEmp)
Jobsat	Categorical	How satisfied is R with his/her job	Converted to ordinal
Joblose	Categorical	How likely is R to lose current job	Converted to ordinal
Jobfind	Categorical	How likely is R to find an equally good job as the current one	Converted to ordinal

Note. R=Respondent.

Methodology

One of our primary objectives for the project, as indicated earlier, is to find patterns among college majors and observe any common characteristics for particular majors using variables related to socioeconomic status, occupation, and demographics. Although we had expectations for certain majors to be associated with one another based on preconceived notions - for example, that science, technology, engineering, and mathematics (STEM) would be closely associated due to their highly-related curricula - this still left open the possibility that there are other less-obvious affinities between majors which have not yet been considered, simply for the reason that there are a wide diversity of majors to choose from. Relatedly, uncovering the factors which render

variables similar (or dissimilar) to one another, or finding majors which are outliers in all respects, remains an interesting pursuit unto itself. Thus, clustering is a valuable approach for revealing such patterns among the 60 majors available in the preprocessed data.

Given the large set of variables in our feature space, it is likely that at least some of the variables are related to one another (e.g. perception of an above-average family income could be associated with higher occupational prestige). In order to add clarity and comprehensibility, we used principal component analysis (PCA) to reduce dimensionality for further empirical analysis. This method helped us identify the optimal number of input dimensions with the most important features chosen for each component.

After we have reduced the feature space down to several components, we used them to find some clustering. In order to generate the best cluster configuration, we considered three major methods: k-means, partitioning around medoids (PAM), and Gaussian mixture models (GMM). Since our major-based data was derived from an aggregation of individual-level data points, k-means, which calculates distances based on centroids, seemed to offer a good start. We tried to fit a PAM clustering as well, primarily due to its lower sensitivity to outliers. GMM was also included in the validation process because of its core function in probabilistic assignment, which might cluster differently from the other two methods.

Results & Discussion

PCA results

According to the scree plot (Figure 1), there is a huge drop in eigenvalue when the number of principal components increased from one to two and from five to six, before the curve flattened

out. We also looked at the summary of the PCA results (Figure 2), where the first six principal components cumulatively explained 72.82% of the total variance. The marginal improvement in capturing the rest of variance from the data diminished as we went above six. Additionally, we took into account the Kaiser rule, which suggests picking PCs with eigenvalues of at least 1. Combining all the evidence, we took PC=6 to be a good cut-off point.

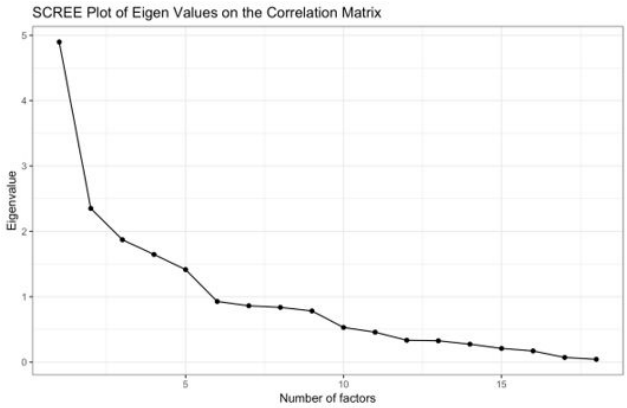


Figure 1: PCA Scree Plot

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11
Standard deviation     2.1993 1.4051 1.25270 1.18204 1.16695 0.95462 0.88562 0.88141 0.80159 0.71453 0.58912
Proportion of Variance 0.3023 0.1234 0.09808 0.08733 0.08511 0.05696 0.04902 0.04856 0.04016 0.03191 0.02169
Cumulative Proportion  0.3023 0.4257 0.52379 0.61111 0.69623 0.75318 0.80220 0.85076 0.89092 0.92283 0.94452
                          PC12    PC13    PC14    PC15    PC16
Standard deviation     0.56456 0.51340 0.43847 0.26580 0.20619
Proportion of Variance 0.01992 0.01647 0.01202 0.00442 0.00266
Cumulative Proportion  0.96444 0.98091 0.99293 0.99734 1.00000
```

Figure 2: PCA Summary Outcome

Thus, the result is the reduction of the 16 variables within our feature space down to 6 principal components. For the purposes of comprehensible interpretation, however, we will only focus on PC1 and PC2, which collectively explain 42.5% of the total variance. The loadings of each variable on PC1 and PC2 are shown in Table 2 and Figure 3.

The interpretations for the components were based on variables where the magnitude of their loadings on each principal component were larger than 0.30. From the results, we found that PC1 is highly associated with variables like years of education (educ), highest degree earned (degree_ord), individual income (income_cont), perceived family income (finrela_ord), socioeconomic status (sei10), and occupational prestige score (prestg10). These variables represent an individual's education level, financial status, and social status. While PC2 is very different, it loads strongly on variables related to gender (sex_isMale), working hours (wrkstas_isFulltime and hrs1), and job stability (joblose_ord), which means majors that have higher scores on PC2 lead to positions that require a longer working time and are more stable.

Table 2. Variable Loadings on PC1 and PC2

	sex_is Male	race_is Majority	age	wrksta t_isFull time	hrs1	wrkslf_isS elfEmp	educ	degree _ord	colscinm	incom e_cont	finrela _ord	sei10	prestg 10	jobsat _ord	joblose _ord	jobfind _lose
PC1	0.11	-0.03	0.19	0.15	0.26	0.00	**0.35**	**0.36**	0.25	**0.33**	**0.31**	**0.41**	**0.39**	-0.03	-0.06	-0.13
PC2	**0.33**	-0.12	-0.23	**0.42**	**0.46**	0.03	-0.30	-0.30	0.13	0.18	0.04	-0.09	-0.11	0.13	**-0.40**	0.09

Note: The magnitude of loadings larger than 0.30 are bolded.

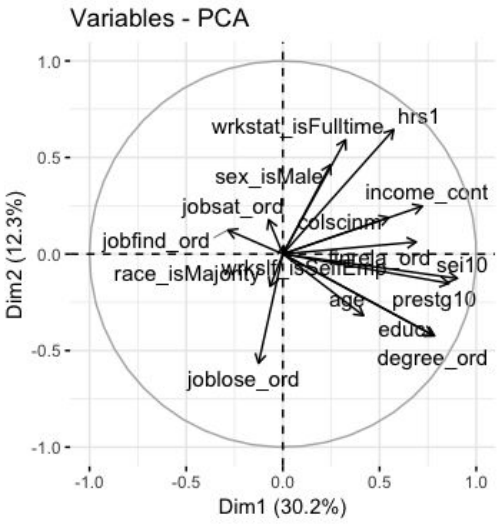


Figure 3: Biplot of PCA Outcome on PC1 and PC2.

The variables that loaded heavily on PC1 intuitively group together: higher education level is strongly correlated to higher income, more prestigious jobs, and a higher socioeconomic status. Such relationships have been documented in previous literature (Barringer et al, 1990; Griliches & Mason, 1972). As for PC2, longer working hours and job stability are highly affiliated with each other (while being orthogonal to PC1, which is mainly education and income). Being male has also been historically related with working longer hours outside of the household, which forms part of a feasible interpretation for the second component.

However, job satisfaction does not load heavily on either PC1 (-0.03) or PC2 (0.13). The low loading on PC1 suggests that job satisfaction does not necessarily indicate occupational prestige, income, or education levels. More broadly, this suggestion runs counter to the belief that pursuing a job with a higher income or greater prestige would lead to a higher job satisfaction, because it shows that they are not that closely covariant.

Clustering results

As per the first objective concerning clustering, we fit a clustering algorithm on the complete feature space of the data (i.e. all 16 input variables). We performed an internal validation test from k=1 to k=10 which suggested that k=2 with k-means was optimal on the connectivity score. Thus, we fit a k-means model with k=2. We also fit a PAM model to have something to compare the k-means model by, and the resulting clusters presented no significant change from k-means, which was expected because the dataset did not feature any prominent outliers.

Figure 4 shows the results of clustering on some of the variables from the data. One immediately observable result is that the clusters are more clearly separated when comparing the

income and education variables, as opposed to any other combination of variables. Additional notable observations from these initial results include the strong association between income and education, along with the strong association between income and occupational prestige. This is interesting because the variable "Occupation prestige" (prestg10) is not directly related to income, but to social perceptions from that type of occupation, as previously described. However, this high-dimensional feature space remained difficult for interpretation, so we made use of dimension reduction via principal components analysis in order to simplify the complexity of interactions between the numerous input variables.

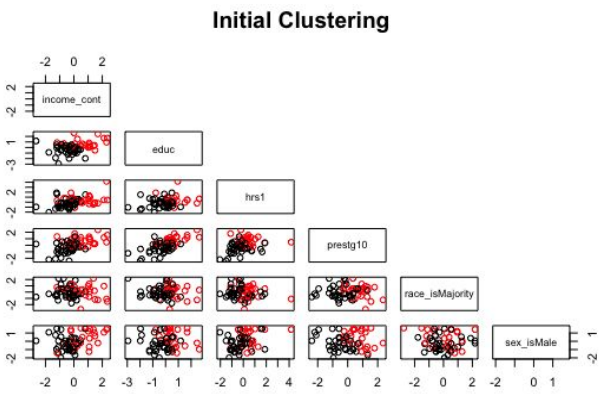


Figure 4: Cluster assignment using k-means with k=2 for selected variables of the complete feature space. Note: Colors represent the cluster assignment.

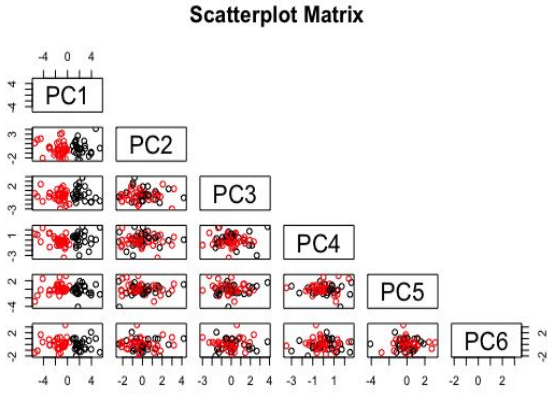


Figure 5: Cluster assignment using k-means with k=2 for the 6 PC feature space. Note: Colors represent the cluster assignment.

We implemented clustering once more, this time on the PCA model as described in the previous section. We repeated the same internal validation process as before with our 6-dimensional space (see Appendix B). Overall, k-means was shown as the most optimal method, so we based the rest of our clustering applications on k-means. Deciding on the number of clusters, however, was more complicated. While the lowest connectivity score was observed at k=2,

silhouette width suggested that k=6 yields the most well-defined clusters, while the Dunn index suggested k=10 as best for clustering. However, for all three clustering methods, the connectivity score was the only reliable metric, with the other two metrics returning noisy fluctuations for each value of k. Therefore, we used k=2 for our exploratory data analysis. The full clustering result for k=2 is presented in Figure 5. Within the scatterplot matrix, the two clusters are clearly separated when the first principal component (PC1) is approximately equal to zero. In other words, the clustering assignment is based primarily on what corresponds to education, income, and occupational prestige. The rest of the components have relatively little influence on cluster assignment. We do not see the separation of the clusters using any other PC as clearly as with PC1.

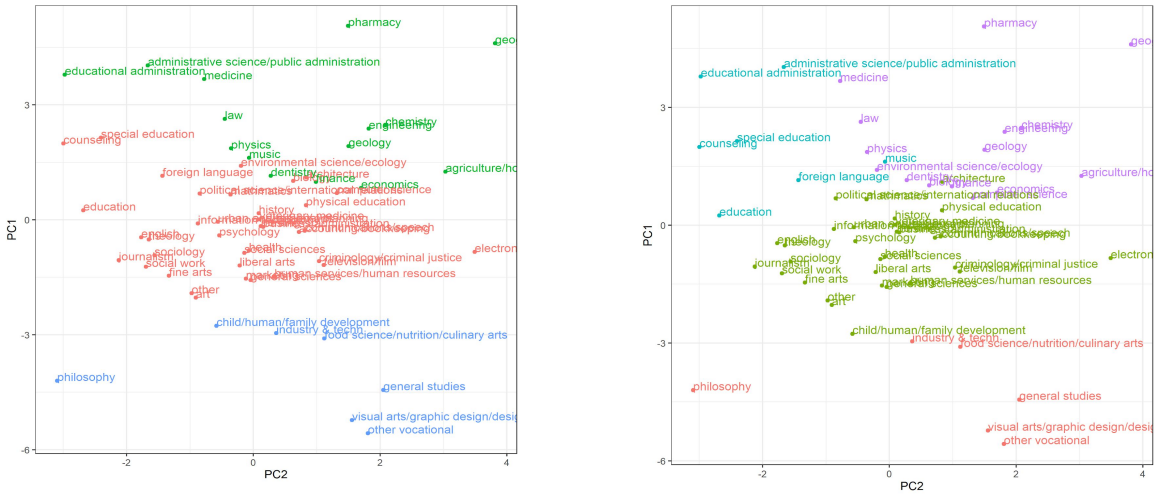
The resulting assignment is displayed in Appendix C along with a quick view of the two clusters in the first two PCs in Appendix D. Generally, we find majors affiliated with technical and scientific professions (e.g. STEM, medicine) to be located in the first cluster, as opposed to arts and humanities majors (e.g. sociology, philosophy) which are in the second. Given that the first cluster scores higher overall along the first PC dimension, which is associated with factors like income and occupational prestige, it is unsurprising that majors ranging from STEM to medicine can be found within this cluster, because these are the majors which are typically perceived as being "respectable". Likewise with the converse cluster - majors within cluster 2 are typically more on the side of liberal arts, the humanities, and socially-oriented work. These results partly echo the common perceptions of such majors.

However, there were some important differences between our results and the common stereotypes associated with these kinds of majors. Music, for instance, was grouped into the first cluster (associated with STEM majors), while electronics and information technology were grouped into the second cluster (associated with liberal arts majors), suggesting that a perception

like “a technology-based majors such as IT is professional, respectable, and well-paying” may not be that accurate after all. These “surprises” are not readily attributable to a low population size for each major, because they each featured approximately 15 people in the survey.

It is also a worthwhile consideration that the first principal component - relating to “material success” (i.e. education, prestige, income) - heavily influenced the clustering results, in addition to the fact that many other variables (e.g. job satisfaction, demographics) did not load heavily onto this principal component. This indicates that although these majors were separated mainly by the “material success” principal component, that scoring high on this axis has relatively little to do with job satisfaction, hours worked, job stability, demographics, and so on. More broadly construed, this seems to raise the question of why do we see little correlation between higher societal esteem (as measured by the PRESTIG10 index) and the individual esteem of the major (as measured by the jobsat variable).

We also conducted clustering with k=3 through k=6 to see if we could find more meaningful groupings. The clustering assignments with k=2 and k=3 are shown in Figure 6. The three-cluster assignment also clearly divides the majors based purely on PC1, while the four-cluster assignment also takes into account variation in those features related to job stability (PC2) in order to cluster the majors.



K=3 **K=4**

Figure 6: Cluster assignments using k-means graphed over PC1 (income and education related feature) and PC2 (job stability related feature). Note: colors represent the cluster assignment.

For the four-cluster assignment, the two clusters located in the upper part of the graph are also separated by PC2. PC2 relates to variables that are indicative of "job stability"; majors that score higher along this axis are the ones that tend to be "professional" and likely require long years of study (medicine, law, physics), as opposed to the more "administrative" majors. As for clustering at k=6 and k=10 (since it was suggested by the Dunn and Silhouette indexes respectively), we could not find any feasible interpretations for the clustering results.

Conclusion

In our exploration of underlying patterns within both a wide array of majors, along with a number of features which we have considered, we have found both typical and expected patterns along with some interesting absences of otherwise-expected patterns in the data. As indicated through a process involving a combination of clustering and dimension reduction, we find that

majors tend to group along commonly-held perceptions regarding the professionality, respectability, and economical prospect of such majors, but not always, as is the case with music, electronics, and IT. Additionally, the results of dimension reduction show that certain features are not strongly related to one another, such as income/prestige and job satisfaction, suggesting that the general, societal-level esteem of a major does not necessarily reflect how individuals actually esteem that major.

Here, we suggest some potential lines for future research. At a broader level, our results might indicate some kind of reinforcing effect between societal-level variables: a major that is likely to lead to well-paying occupations will generally be deemed as more respectable, where in turn, firms would likely be willing to pay more for a major that is considered more respectable. An exploration into the "vicious/virtuous cycle" that underlies the relationship between "societal percept of a major" and "the actual outcomes and economic prospects of such majors" (and whether there can be confounding variables which underlie both) could be one future line of research. Our results could also be used to inform future research on how students come to choose their college majors. Such research has already been intimated, for example, by Wiswall & Zafar (2014) in which they report that heterogenous taste is the dominant factor in choice of major, more so than expected earnings. Not enough research pursues the individual-level considerations which go into deciding what major to pursue: despite the common agreement that education leads to higher income and is important for that reason in particular, in reality, most students do not choose their major mainly on the basis of expected earnings, nor how respectable their position is within society. The reasons are likely more personal than that, given that college is typically a time of self-exploration, and current literature would stand to benefit with an investigation beyond looking at the effects of several independent variables on college choice.

References

Balsamo, M., Lauriola, M., Saggino, A. (2013). Work values and college major choice. *Learning and Individual Differences, 24*, 110-116.

Barringer, H. R., Takeuchi, D. T., & Xenos, P. (1990). Education, occupational prestige, and income of Asian Americans. *Sociology of Education*, 27-43.

Beggs, J.M., Bantham, J.H., Taylor, S. (2008). Distinguishing the factors influencing college students' choice of major. *College Student Journal, 42*(2), 381-394.

Dickson, L. (2010). Race and gender differences in college major choice. *The Annals of the American Academy, 627*(1), 108-124.

Griliches, Z., & Mason, W. M. (1972). Education, income, and ability. *Journal of political Economy*, 80(3, Part 2), S74-S103.

Lin, T. (2004). The role of higher education in economic development: an empirical study of Taiwan case. *Journal of Asian Economics, 15*, 355-371.

McIlveen, P., Beccaria, G., Burton, L. J. (2013). Beyond conscientiousness: career optimism and satisfaction with academic major. *Journal of Vocational Behavior, 83*(3), 229-236.

Milsom, A., & Coughlin, J. (2015). Satisfaction with college major: a grounded theory study. *NACADA Journal, 35*(2), 5-14.

Moakler, M.W., Kim, M.M. (2014). College major choice in STEM: revisiting confidence and demographic factors. *The Career Development Quarterly, 62*(2), 128-142.

Porter, S. R., & Umbach, P. D. (2006). College major choice: an analysis of person-environment fit. *Research in Higher Education, 47*(4), 429-449.

Psacharopoulos, G. (1994). Returns to education: A global update. *World Development, 22*(9), 1325–1343.

Stephenson, B. C. (1979). Probability sampling with quotas: an experiment. GSS Methodological Report No.7, April, 1979. *Public Opinion Quarterly, 43*, 477-496.

Stevens, G., & Featherman, D. L. (1981). A Revised Socioeconomic Index of Occupational Status. *Social Science Research, 10*(4): 364–395. doi:10.1016/0049-089x(81)90011-9.

Wiswall, M., Zafar, B. (2014). Determinants of college major choice: identification using an information experiment. *The Review of Economic Studies, 82*(2), 791-842.

Appendices

Appendix A - Complete Loading of Six PCs

```
Rotation (n x k) = (16 x 16):
                          PC1         PC2         PC3         PC4         PC5         PC6
sex_isMale          0.112005138  0.32916663 -0.10722333  0.33324358 -0.29900609  0.15864595
race_isMajority    -0.029922587 -0.11928479  0.53266867 -0.02391411 -0.17738824  0.62331781
wrkstat_isFulltime  0.147536962  0.42097595  0.18209271  0.03461176  0.44377844 -0.14517521
wrkslf_isSelfEmp    0.002158139  0.03137706 -0.10887190 -0.17026585 -0.69556288 -0.36639343
educ                0.351328162 -0.29859236 -0.12074515 -0.14657758  0.07642600  0.05188049
degree_ord          0.356728997 -0.30169906 -0.13464339 -0.12295450  0.08433960  0.06864300
colscinm            0.249080502  0.13445655  0.08088956  0.42623982 -0.16271239 -0.05262610
income_cont         0.327996690  0.17639127 -0.02291820  0.21098059 -0.21782020 -0.08043582
finrela_ord         0.313501797  0.04439030 -0.15977885 -0.04212778 -0.17946926  0.49771548
age                 0.189563569 -0.22716954  0.44786705  0.11450183 -0.01183238 -0.22983575
sei10               0.410461169 -0.09008762 -0.04983687 -0.15107663  0.01442573 -0.05666931
prestg10            0.390808829 -0.10811694  0.06107024 -0.20738747  0.04845667 -0.17959128
hrs1                0.260610857  0.45868419 -0.02097375 -0.04388533  0.18159137  0.13226597
jobsat_ord         -0.033507205  0.12651458  0.44162716 -0.38213901 -0.19952972 -0.02139606
joblose_ord        -0.057273446 -0.40281731 -0.10989182  0.51983487  0.05919994  0.06343100
jobfind_ord        -0.129340870  0.08833212 -0.42438388 -0.31062653 -0.01758261  0.23739911
```

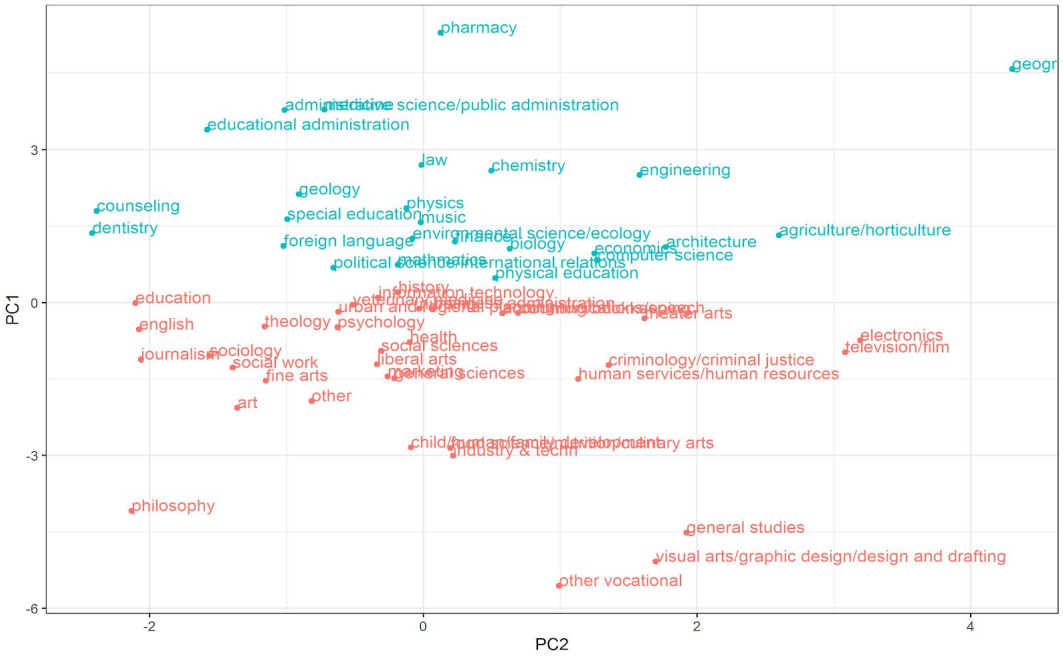
Appendix B - Summary of Internal Validation

```
Clustering Methods:
 kmeans pam model

Cluster sizes:
 2 3 4 5 6 7 8 9 10

Validation Measures:
                            2       3       4       5       6       7       8       9      10

kmeans Connectivity   25.1532 41.7147 46.3115 51.8468 56.4544 69.8603 61.8218 62.3552 66.3837
       Dunn            0.1365  0.1298  0.1381  0.1598  0.1672  0.1220  0.2016  0.2161  0.2170
       Silhouette      0.1902  0.1906  0.1825  0.2044  0.2066  0.1515  0.1965  0.2009  0.1872
pam    Connectivity   27.0861 49.8857 50.4484 56.0091 65.4778 65.2635 68.9536 72.6913 74.5702
       Dunn            0.1019  0.1298  0.1328  0.1353  0.1353  0.1353  0.1428  0.0963  0.1213
       Silhouette      0.1686  0.0892  0.1341  0.1354  0.1230  0.1440  0.1559  0.1535  0.1528
model  Connectivity   51.9937 72.1190 83.4135 84.7837 87.4599 75.9417 78.5456 65.0413 78.1853
       Dunn            0.1056  0.1038  0.0926  0.0926  0.0982  0.1113  0.1193  0.1767  0.1640
       Silhouette      0.0368 -0.0139 -0.0500 -0.0391 -0.0247  0.0364  0.0620  0.1738  0.1284

Optimal Scores:

             Score   Method Clusters
Connectivity 25.1532 kmeans 2
Dunn          0.2170 kmeans 10
Silhouette    0.2066 kmeans 6
```

Appendix C - Majors clustering assignment for k=2

Cluster 1 majors		Cluster 2 majors	
• Pharmacy • Geography • Admin science /public admin • Educational admin • Medicine • Law • Chemistry • Engineering • Special education • Counseling • Geology • Physics • Music • Environmental science/ecology	• Agriculture/horticulture • Dentistry • Foreign language • Architecture • Biology • Finance • Economics • Computer science • Political science/international relations • Mathematics • Physical education • History • Veterinary medicine	• Education • Urban and regional planning • Theater arts • Information technology • Nursing • Business administration • Communication/speech • Accounting/bookkeeping • Psychology • English • Theology • Health • Electronics • Social sciences	• Journalism • Criminology/criminal justice • Television/film • Liberal arts • Social work • Fine arts • Human services/human resources • Marketing • General sciences • Art • Child/human/family development • Other

Appendix D - Clustering Assignment for k=2



Appendix figure D. Cluster assignments using k-means with k=2 graphed over PC1 (income and education related feature) and PC2 (job stability related feature). Note: colors represent the cluster assignment.

Appendix E - Inline Code

```
################################################
#        PART 1: Data Wrangling            #
################################################
#rm(list=ls())
library(tidyverse)
library(foreign)

### 1. PREPARE THE DATASET
{
#GSS <- read.dta("C:/Users/Rosenberg Lab/Documents/UML_Group_Project/GSS7218_R1.DTA")
#gss_c <- read.dta("/Users/JesusPacheco/Downloads/GSS7218_R1.dta")
gss_c <- read.dta(file.choose())

# get only 2016 and 2018
gss16 <- gss_c %>% filter(year>=2016)
names(gss16) <- tolower(names(gss16))

list_varName <- c("wtss",
                  "educ", "degree", "major1", "colscinm",
```

```
            "rincom16", "finrela",
            "sex", "age", "race",
            "sei10", "prestg10", "wrkstat", "hrs1", "wrkslf", "occ10", "jobsat", "joblose", "jobfind")

# extract only our variables of interest
gss <- gss16 %>% select(list_varName)
gss <- gss %>% drop_na(major1)

# remove people who somehow have degrees from high school; we are only interested in people who have attended
college
gss <- filter(gss, degree != "high school")
}

### 2. RECODING
{
#We converted the string variables to lower in all cses
gss$major1 <- tolower(gss$major1)

# convert degree into ordinal levels
gss$degree <- tolower(gss$degree)
gss <- gss %>% mutate(degree_ord = recode(degree,
                                  "junior college"=1,
                                  "bachelor"=2,
                                  "graduate"=3))

# convert income brackets to numerical values so we can treat as a continuous variable
gss$rincom16 <- tolower(gss$rincom16)
gss <- gss %>% mutate(income_cont = recode(rincom16,
                                   "under $1 000"=500,
                                   "$1 000 to 2 999"=2000,
                                   "$3 000 to 3 999"=3500,
                                   "$4 000 to 4 999"=4500,
                                   "$5 000 to 5 999"=5500,
                                   "$6 000 to 6 999"=6500,
                                   "$7 000 to 7 999"=7500,
                                   "$8 000 to 9 999"=9000,
                                   "$10000 to 12499"=11250,
                                   "$12500 to 14999"=13750,
                                   "$15000 to 17499"=16250,
                                   "$17500 to 19999"=18750,
                                   "$20000 to 22499"=21250,
                                   "$22500 to 24999"=23750,
                                   "$25000 to 29999"=27500,
                                   "$30000 to 34999"=32500,
                                   "$35000 to 39999"=37500,
                                   "$40000 to 49999"=45000,
                                   "$50000 to 59999"=55000,
                                   "$60000 to 74999"=67500,
                                   "$75000 to $89999"=82500,
                                   "$90000 to $109999"=100000,
                                   "$110000 to $129999"=120000,
```

```
                                "$130000 to $149999"=140000,
                                "$150000 to $169999"=160000,
                                "$170000 or over"=170000))

# convert finrela into ordinal (likert) levels
gss$finrela <- tolower(gss$finrela)
gss <- gss %>% mutate(finrela_ord = recode(finrela,
                                "far below average"=1,
                                "below average"=2,
                                "average"=3,
                                "above average"=4,
                                "far above average"=5))

# convert sex into boolean (isMale)
gss <- gss %>% mutate(sex_isMale = recode(sex,
                                "female"=FALSE,
                                "male"=TRUE))

# convert race into boolean (isMajority, i.e. isWhite)
gss <- gss %>% mutate(race_isMajority = recode(race,
                                  "other"=FALSE,
                                  "black"=FALSE,
                                  "white"=TRUE))

# convert wrkstat into boolean (isFulltime)
gss$wrkstat <- tolower(gss$wrkstat)
gss <- gss %>% mutate(wrkstat_isFulltime = recode(wrkstat,
                                    "working parttime"=FALSE,
                                    "working fulltime"=TRUE))

# convert wrkslf into boolean (isSelfEmp)
gss$wrkslf <- tolower(gss$wrkslf)
gss <- gss %>% mutate(wrkslf_isSelfEmp = recode(wrkslf,
                                  "someone else"=FALSE,
                                  "self-employed"=TRUE))

# convert jobsat to jobsat_ord (likert)
gss$jobsat <- tolower(gss$jobsat)
gss <- gss %>% mutate(jobsat_ord = recode(jobsat,
                                  "completely sat"=7,
                                  "very sat"=6,
                                  "fairly sat"=5,
                                  "neither sat"=4,
                                  "fairly dissat"=3,
                                  "very dissat"=2,
                                  "completely dissat"=1))
# convert joblose to joblose_ord (likert)
gss$joblose <-tolower(gss$joblose)
gss <- gss %>% mutate(joblose_ord = recode(joblose,
                                "very likely"=4,
                                "fairly likely"=3,
```

```
                                "not too likely"=2,
                                "not likely"=1))

# convert jobfind to jobfind_ord (likert)
gss$jobfind <-tolower(gss$jobfind)
gss <- gss %>% mutate(jobfind_ord = recode(jobfind,
                                "very easy"=3,
                                "somewhat easy"=2,
                                "not easy"=1))


### End of Recoding
}

### 3. AGGREGATE THE DATA BY MAJOR (FINAL DATASET gss_f)
{
gss_f <- gss %>% select(final_vars) %>% group_by(major1) %>% mutate(num = n()) %>%
   summarise_all(funs(mean(.,na.rm=TRUE)))
   #summarise_all(funs = weighted.mean(.,w=wtss))
gss_f <- gss_f %>% drop_na()
write.csv(gss_f, "gss_f.csv")
}

################################################
#    PART 2: Analysis (PCA + clustering)      #
################################################
rm(list=ls())
library(tidyverse)
library(seriation)
library(factoextra)
library(clValid)
library(ggfortify)

### 1. GET THE DATASET FROM THE WRANGLING PROCESS
{
#setwd("/Users/JesusPacheco/GitHub/UML-project/FINAL")
gss_f <- read_csv(file.choose()) %>% select(-c(X1)) %>% data.frame()
row.names(gss_f) <- gss_f$major1
gss_scaled <- gss_f %>% select(-c(major1, num)) %>% scale()

}

### 2. INITIAL CLUSTERING
{
## 2.1. Looking at clusterability
dist_mat<- dist(gss_pca, method = "euclidean")
png("initial_clusterability. png")
dissplot(dist_mat)
dev.off()

# 2.2. Check internal validation
```

```
initial_intval <- clValid(gss_scaled, nClust = 2:10,
                      clMethods = c("kmeans", "pam", "model"),
                      validation = "internal")
summary(initial_intval)

# 2.3. K-means with k=2
initial_km <- kmeans(gss_scaled, centers = 2, nstart=10)
gss_scaled <- data.frame(cbind(gss_scaled,as.matrix(initial_km$cluster)))

png("initial_clustering.png")
pairs(~ income_cont + educ + hrs1 + prestg10 + race_isMajority + sex_isMale, data=gss_scaled,
    col = gss_scaled$V17, bg = c("blue", "yellow", "red"),
    main="Initial Clustering", upper.panel=NULL)
dev.off()
}

### 3. PCA
{
## 3.1. Fit the PCA model
pca_outcome <- prcomp(gss_scaled[,1:16], scale = TRUE)
summary(pca_outcome)

## 3.2 SCREEPLOT & POV
qplot(y=pca_outcome$sdev) + geom_line() +
 labs(title="SCREE plot of PCA analysis", y="Eigen values", x= "Number of factors") +
 theme_bw()
ggsave("screeplot.png")

POV <- pca_outcome$sdev^2/sum(pca_outcome$sdev^2)
round(POV[1:10],4)

## 3.3 BIPLOTS

autoplot(pca_outcome,
        shape = F,
        loadings.label = T,
        repel=T) +
 theme_bw()

#Looking only at the components
fviz_pca_var(pca_outcome,repel = TRUE)
ggsave("biplot_1.png")

#This is the reduced data set
gss_pca <- pca_outcome$x[, 1:6]

}

### 4. PCA + CLUSTERING
{
## 4.1. CLUSTERABILITY
```

```
png("eda.png")
pairs(~PC1 + PC2 + PC3 + PC4 + PCA5 + PCA6, data=gss_pca,
     main="Principal component matrix")
dev.off()

dist_mat<- dist(gss_pca, method = "euclidean")
png("clusterability. png")
dissplot(dist_mat)
dev.off()


## 4.2 INTERNAL VALIDATION
internal <- clValid(gss_pca, nClust = 2:10,
                    clMethods = c("kmeans", "pam", "model"),
                    validation = "internal")
summary(internal)

## 4.3 CLUSTERING
#kmeans
for (i in 1:6) {
  assign(paste0("km",i),kmeans(gss_pca, centers = i, nstart=10) )
}
#pam
pam2 <- pam(gss_pca, k=2)

#adding the clusters to the dataset
gss_pca <- data.frame(cbind(gss_pca,as.matrix(km2$cluster)))
names(gss_pca)[7] <- "clust"
gss_pca <-  data.frame(cbind(gss_pca, gss_f$major1))
names(gss_pca)[8] = "major1"

#plotting the clusters in the complete set of PCs
png("final_clustering.png")
pairs(~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6, data=gss_pca, col = gss_pca$clust, upper.panel = NULL,
     main="Scatterplot Matrix")
dev.off()

#plotting the clusters only in PC1 and PC2
gss_pca %>% ggplot(aes(PC2, PC1, col=as.factor(clust))) + geom_point() +
  geom_text(aes(label=row.names(gss_pca)),hjust=0, vjust=0) + theme_bw() +
  theme(legend.position = "none")
ggsave("cluster_pc1_pc2.png")

}
```