Us College Majors Cluster Analysis

Introduction

Choosing a major is a difficult decision almost everyone has to make. This milestone will influence every student's career trajectory for good or bad.

Young people are encouraged to weight several factors before making a final decision such as:

- Employment rates in the field
- Advanced degree opportunities
- Salary expectations
- Overall program cost

Objetive

Since its a major life decision, we want to informe young people about the advantages or disadvantages they may encounter when selecting an specific major. Hence, we will compare the recommendations from three different methods for determining the optimal number of clusters, apply a k-means clustering analysis, and visualize the results.

Data

We will use a year-long survey of 1.2 million people with only a bachelor's degree by PayScale.Inc that can be found here.

```
library(ggplot2)
library(tidyr)
library(dplyr)
library(readr)
library(cluster)
library(factoextra)

majors <- read.csv('Us_college.csv')
col_names=c('College.Major', 'Starting.Median.Salary','Mid.Career.Median.Salary',
            'Career.Percent.Growth', 'Percentile.10','Percentile.25',
            'Percentile.75','Percentile.90')
names(majors) <- col_names
print(head(majors))
```

```
##           College.Major Starting.Median.Salary Mid.Career.Median.Salary
## 1            Accounting             $46,000.00              $77,100.00
## 2 Aerospace Engineering             $57,700.00             $101,000.00
## 3           Agriculture             $42,600.00              $71,900.00
## 4          Anthropology             $36,800.00              $61,500.00
## 5          Architecture             $41,600.00              $76,800.00
## 6           Art History             $35,800.00              $64,900.00
##   Career.Percent.Growth Percentile.10 Percentile.25 Percentile.75 Percentile.90
## 1                  67.6    $42,200.00    $56,100.00   $108,000.00   $152,000.00
## 2                  75.0    $64,300.00    $82,100.00   $127,000.00   $161,000.00
## 3                  68.8    $36,300.00    $52,100.00    $96,300.00   $150,000.00
```

```
## 4                  67.1    $33,800.00    $45,500.00    $89,300.00   $138,000.00
## 5                  84.6    $50,600.00    $62,200.00    $97,000.00   $136,000.00
## 6                  81.3    $28,800.00    $42,200.00    $87,400.00   $125,000.00
```

Data Cleaning

```
majors_clean <- majors %>%
    mutate_at(vars(Starting.Median.Salary:Percentile.90),
              function(x) as.numeric(gsub("[\\$,]","",x))) %>%
    mutate(Career.Percent.Growth = Career.Percent.Growth/100)
```

Clustering Analysis

First step, we have to determine the number of clusters we should model. There are several methods to get to an aproach, since one its not enough for the task. Hence we will use 3 techniques to optimize the number of clusters:

- Gap static method
- Silhouette method
- Elbow method

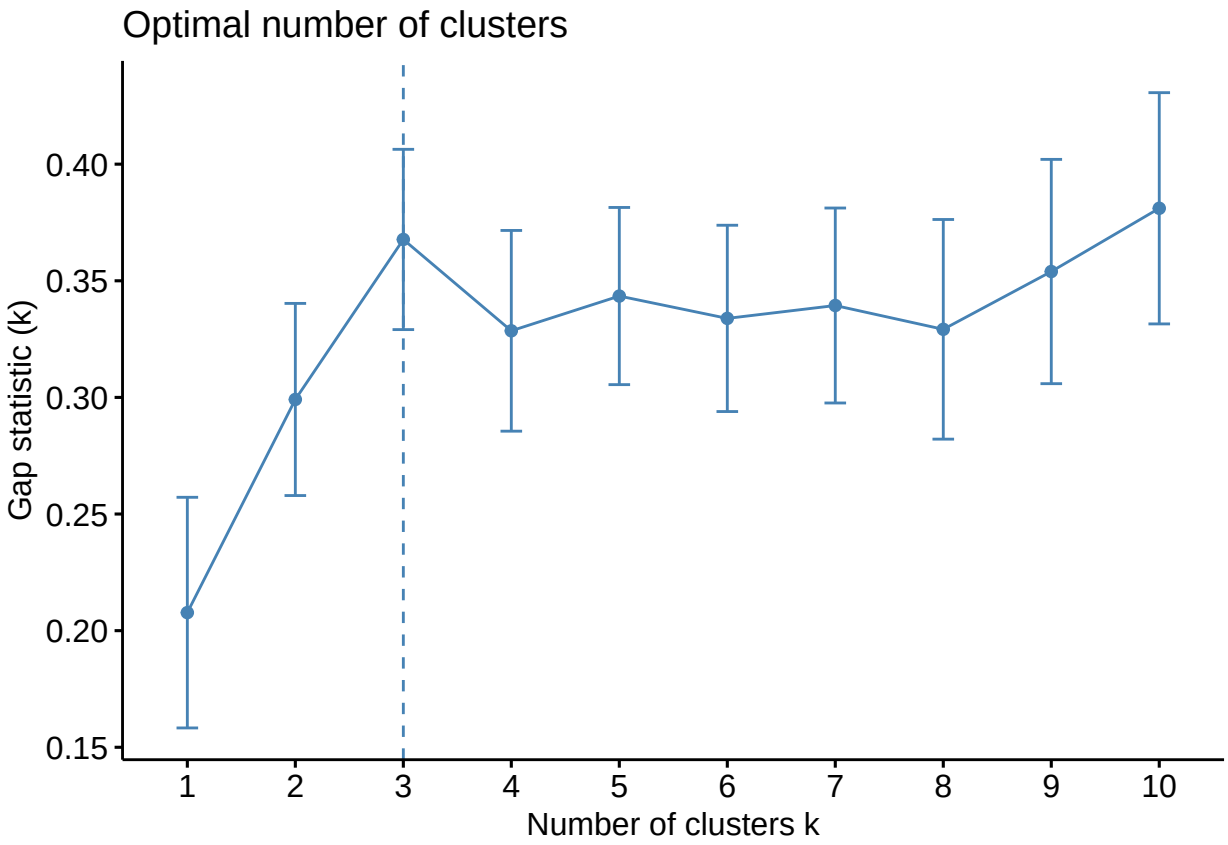
```
data <- majors_clean %>%
    select(Starting.Median.Salary, Mid.Career.Median.Salary,
           Percentile.10, Percentile.90) %>% scale()
```

Gap static method

Our first method compares the total variation within clusters for different values of k to the null hypothesis, making this to maximize the gap. This hypothesis refers to a uniformly distributed simulated reference dataset with no observable clusters, generated by aligning with the principle components of our original dataset.

```
gap_stat <- clusGap(data, FUN = kmeans, nstart = 25, K.max = 10, B = 50)

gap_static <- fviz_gap_stat(gap_stat)
gap_static
```

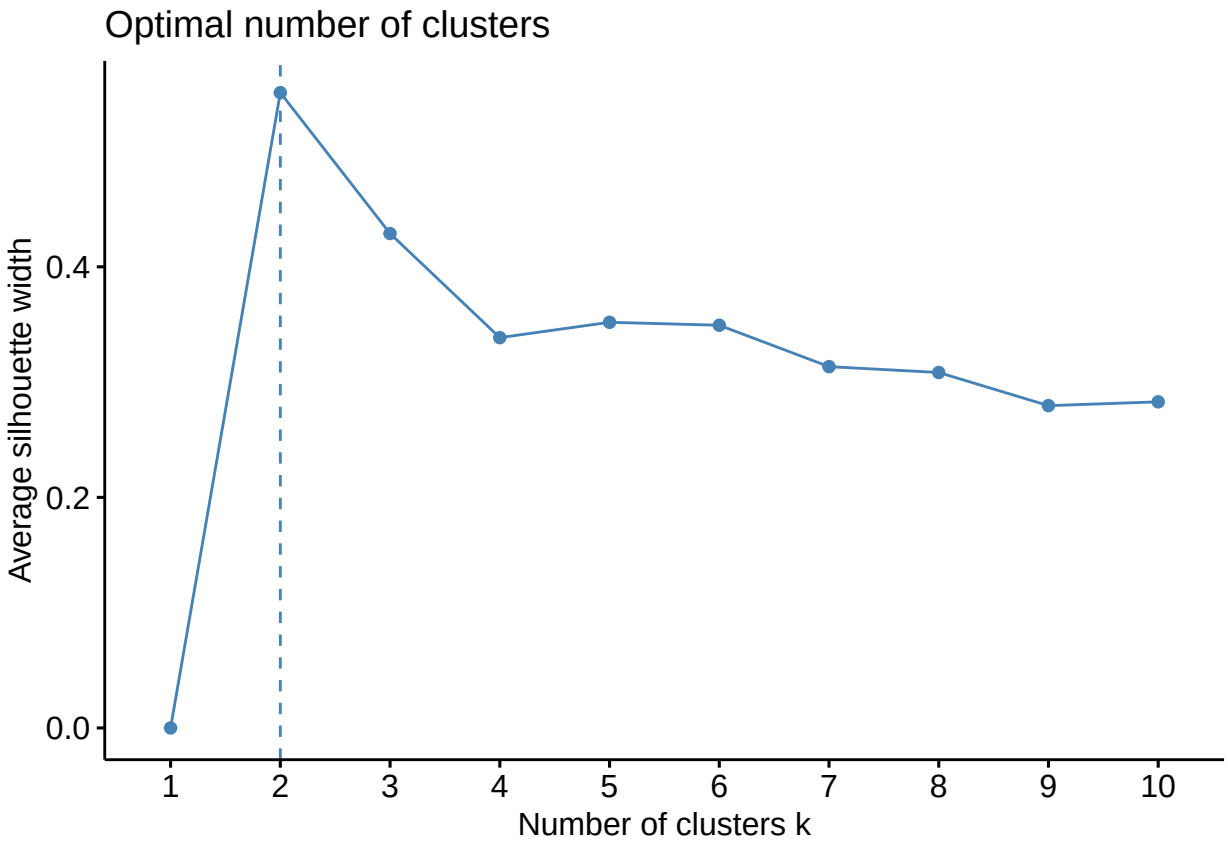


In the above visualization we see how much the variance is explained by k clusters in our dataset, making the ideal value 3 for our clusters.

Silhouette Method

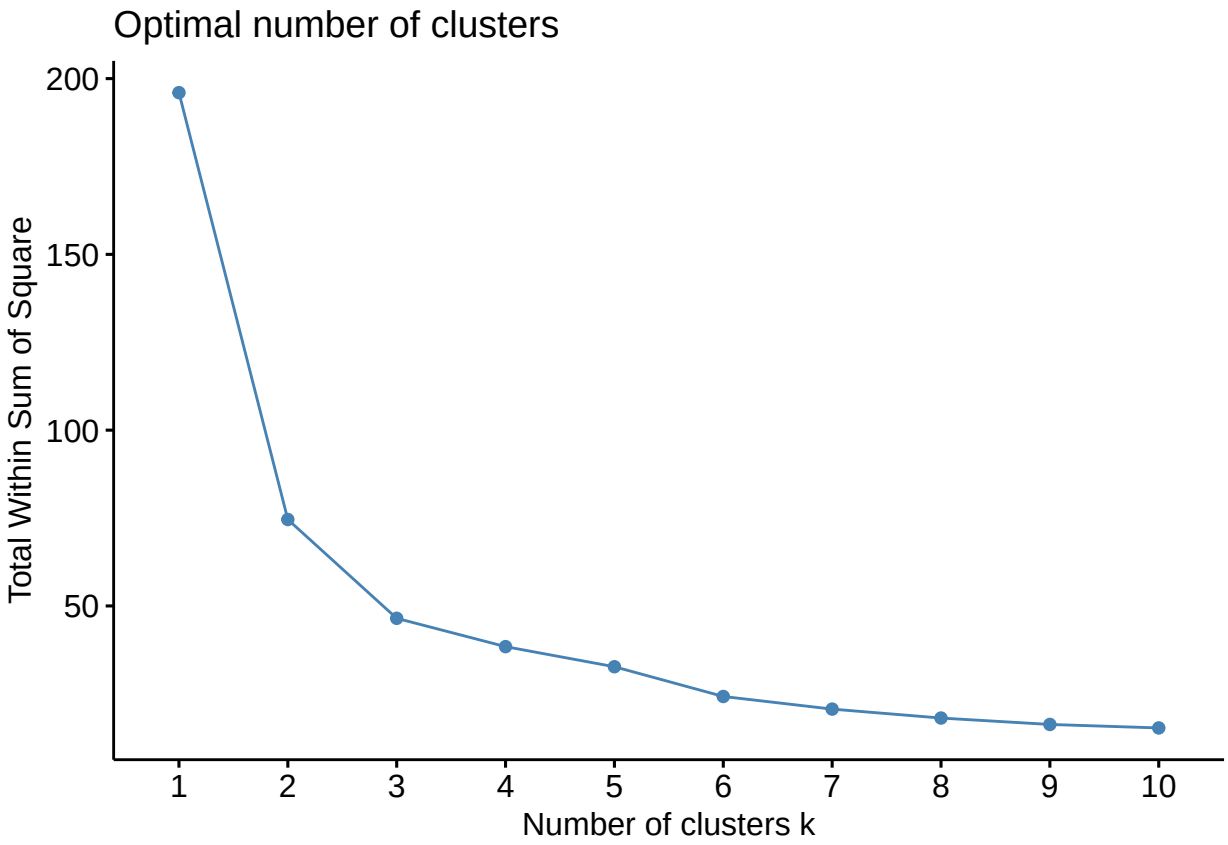
This method evaluates the clusters quality by how well each point fits within a cluster, maximizing average silhouette width.

```
silhouette <- fviz_nbclust(data, kmeans, method = "silhouette")
silhouette
```



Finally, the elbow method, which is the most common to use for this type of situations. This method plots the percent variance against the number of clusters.

```
elbow <- fviz_nbclust(data, kmeans, method = "wss")
elbow
```



As you can see there is a break point at the value 3, making the curve to bend like an elbow. This indicates the optimal point at which adding more clusters will no longer explain a significant amount of the variance.

K-means Algotithm

Once we have our results (2, 3 and 3) we can be sure that working with 3 clusters will be the best for our model.

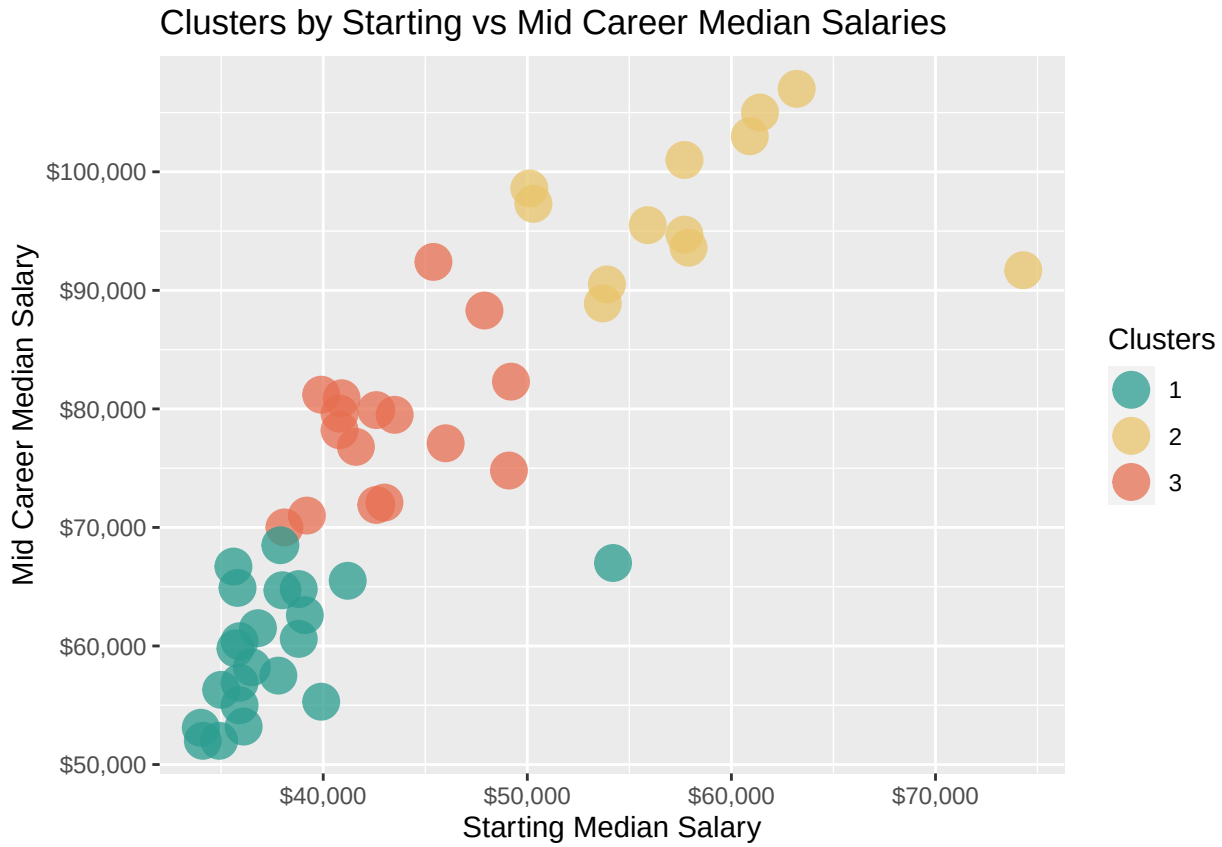
```
set.seed(354)
num <- 3
k_means <- kmeans(data, num, iter.max = 15, nstart = 25)

majors_labld <- majors_clean %>%
    mutate(clusters = k_means$cluster)
```

Accordingly to this, lets start by getting a view to how each cluster compares in terms of starting vs mid career salaries.

```
career <- ggplot(majors_labld,
                 aes(x=Starting.Median.Salary,y=Mid.Career.Median.Salary,
                     color=factor(clusters))) +
      geom_point(alpha=0.75,size=6) +
      scale_color_manual(name="Clusters",values=c("#2A9D8F","#E9C46A", "#E76F51")) +
      ggtitle('Clusters by Starting vs Mid Career Median Salaries') +
```

```
      scale_x_continuous(labels = scales::dollar) +
      scale_y_continuous(labels = scales::dollar) +
      labs(x='Starting Median Salary', y='Mid Career Median Salary')
career
```



As shown in the visualization the data follows a linear trend meaning that the higher your starting salary is, the higher your mid career salary will be. The clusters provide a level of delineation that also supports this.

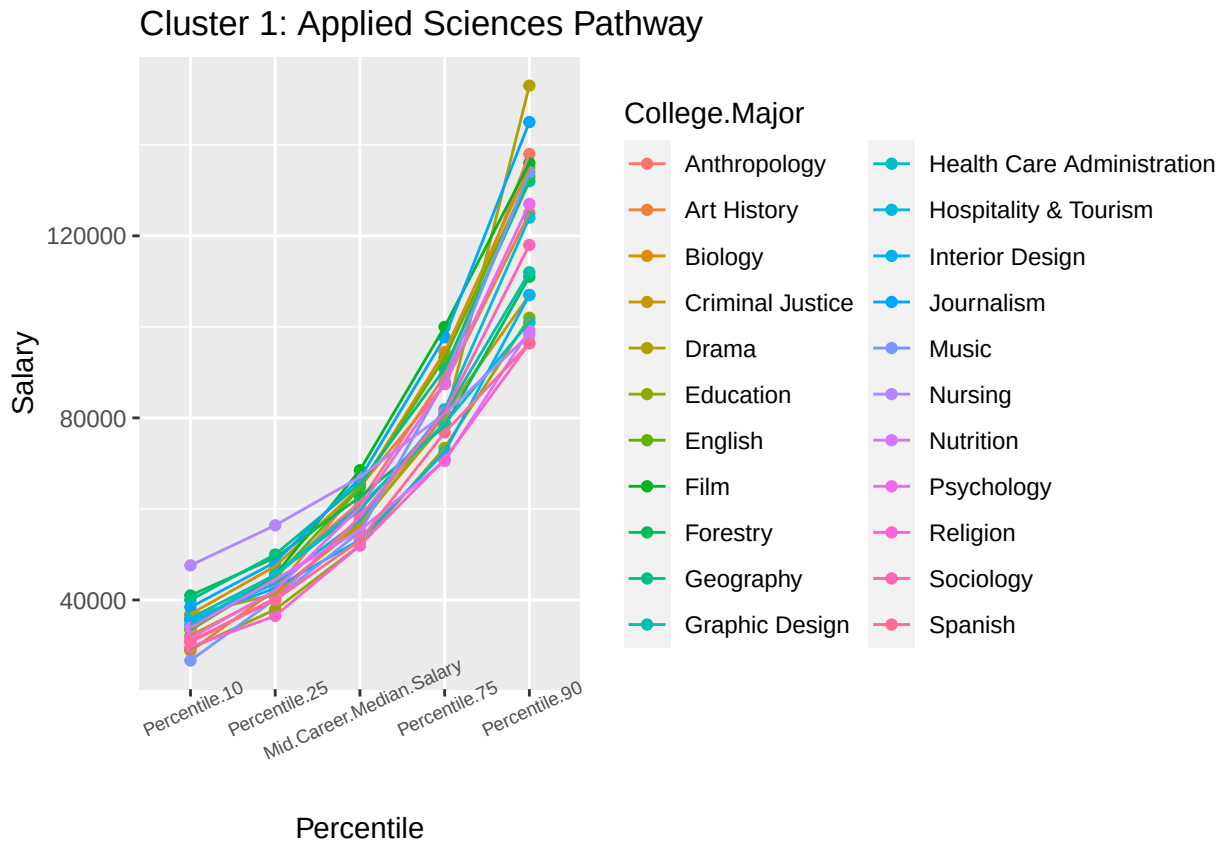
Furthermore we also can see two outliners from cluster 1 and 3 that may be explained if we go deeper into our analysis and get insights of the career salaries percentiles.

```
percentiles <- majors_labld %>%
    select(College.Major, Percentile.10, Percentile.25,
           Mid.Career.Median.Salary, Percentile.75,
           Percentile.90, clusters) %>%
    gather(key=percentile, value=salary, -c(College.Major, clusters)) %>%
    mutate(percentile=factor(percentile,levels=c('Percentile.10','Percentile.25',
            'Mid.Career.Median.Salary','Percentile.75','Percentile.90')))
```

Cluster 1: Applied Sciences Pathway

It seems this cluster is characterized by job stability and sets in the middle of the road in our dataset starting off not too low and not too high in the lowest percentile. It also represents the majors with the greatest difference between the lowest and highest percentiles.

```
cluster_1 <-  ggplot(percentiles[percentiles$clusters==1,],
                     aes(x=percentile,y=salary,
                     group=College.Major, color=College.Major, order=salary)) +
                     geom_point() +
                     geom_line() +
                     theme(axis.text.x = element_text(size=7, angle=25)) +
                     ggtitle('Cluster 1: Applied Sciences Pathway') +
                     labs(x='Percentile', y='Salary')
cluster_1
```

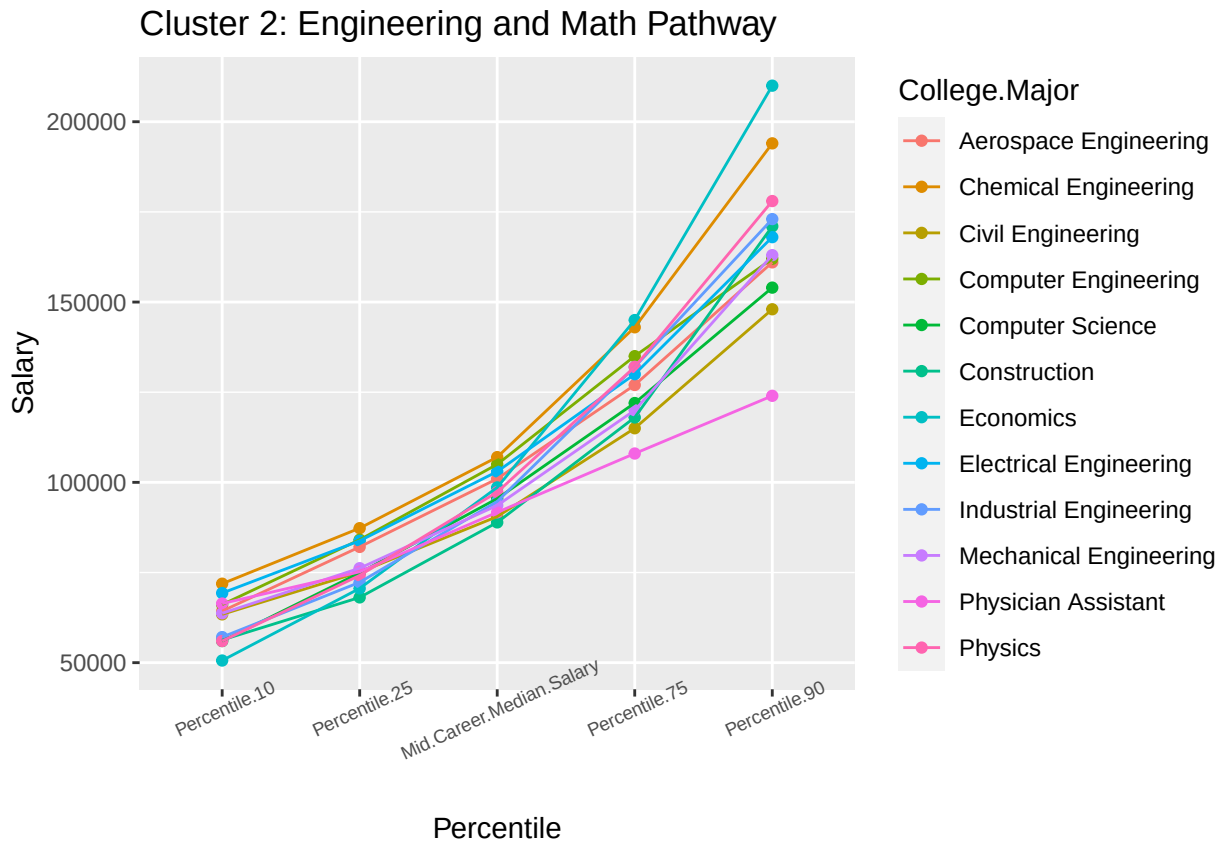


Cluster 2: Engineering and math pathway

If you are good at math and want financial security you should consider one of this mayors. These engineering majors represent the highest growth potential in the 90th percentile, as well as the best security in the 10th percentile rankings. We see one of the outliers, now identifiable as Physician Assistant lagging in the highest percentiles.

```
cluster_2 <- ggplot(percentiles[percentiles$clusters==2,],
    aes(x=percentile,y=salary,
    group=College.Major, color=College.Major)) +
    geom_point() +
    geom_line() +
    ggtitle('Cluster 2: Engineering and Math Pathway') +
    theme(axis.text.x = element_text(size=7, angle=25)) +
```

```
    labs(x='Percentile', y='Salary')
cluster_2
```



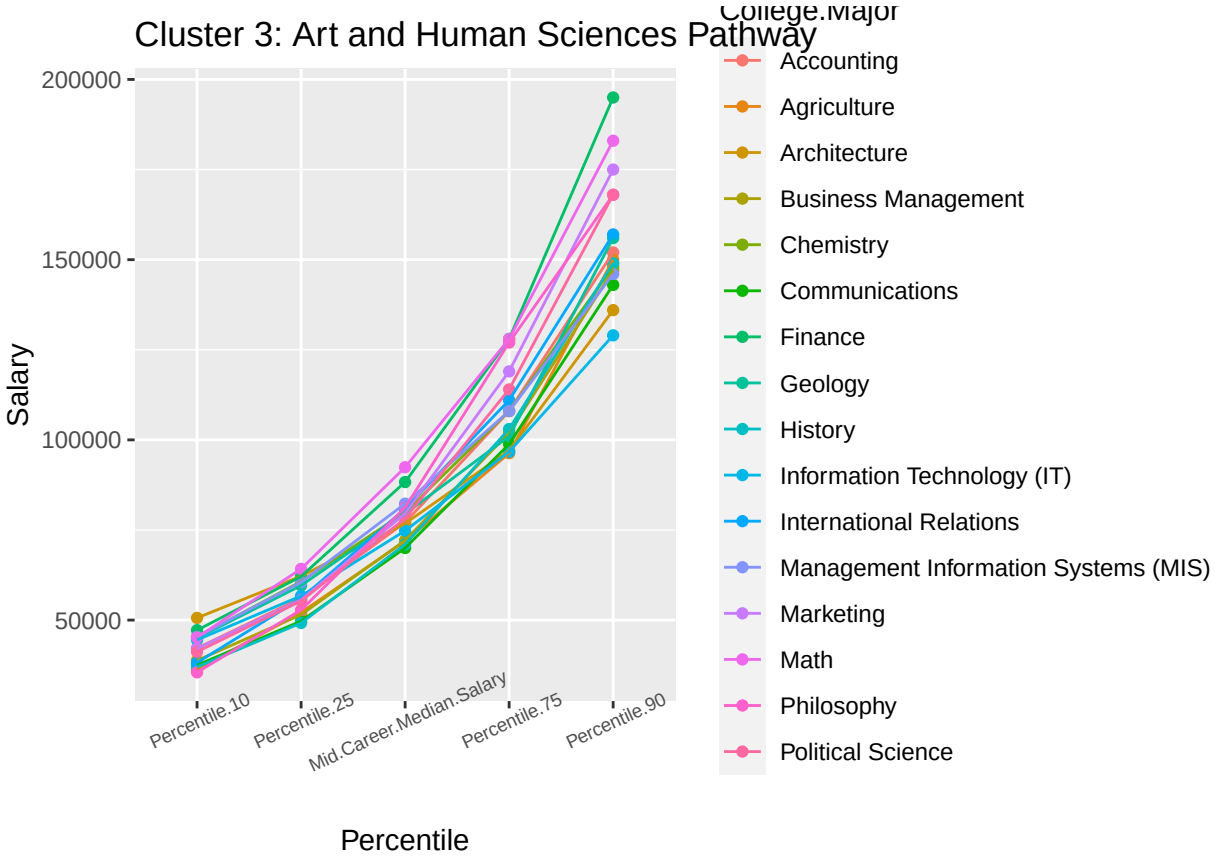
Cluster 1: Art and Human Sciences Pathway

Not all is science and math, there are also good majors with a lot of opportunities. If you are passionate about art or have the hearth to help others these are good options. However the majors surrounding this cluster represent the lowest percentiles with limited growth opportunity.

Music major is the riskiest with the lowest 10th percentile salary, but Drama has the highest growth potential in the 90th percentile for this cluster. Nursing is the outlier culprit of this cluster, with a higher safety net in the lowest percentile to the median.

```
cluster_3 <- ggplot(percentiles[percentiles$clusters==3,],
    aes(x=percentile,y=salary,
    group=College.Major, color=College.Major)) +
    geom_point() +
    geom_line() +
    ggtitle('Cluster 3: Art and Human Sciences Pathway') +
    theme(axis.text.x = element_text(size=7, angle=25)) +
    labs(x='Percentile', y='Salary')

cluster_3
```



Conclusions and recommendations

Dealing with unsupervised data always requires skill and some creativity. It is recommended the use of not just one method to look after the ideal number of clusters but to use more to be certain and get a good model.

Yes, it is important to focus on high starting career salaries when choosing a major and to consider the growth potential. However, keep in mind that whether a major falls to the sciences, engineering or arts and human sciences cluster, one's financial destiny is influenced by numerous other factors including the school attended, passion or talent for the subject.

If you are curious about the factors mentioned above, a similar analysis to evaluate them can be conducted on the additional data provided by the Wall Street Journal article, comparing salary potential by type and region of college attended.

Regardless of the salaries, job opportunities and other factors, I want to remind you the most important thing: Follow your passions!

Table 1: Top College Majors Sorted by Career Percent Growth

College Major	Start Median Salary	Mid Median Salary	Percent Growth	Percntl 10	Percntl 25	Percntl 75	Perctnl 90	Cluster
Accounting	46000	77100	0.676	42200	56100	108000	152000	3
Aerospace Engineering	57700	101000	0.750	64300	82100	127000	161000	2

College Major	Start Median Salary	Mid Median Salary	Percent Growth	Percntl 10	Percntl 25	Percntl 75	Perctnl 90	Cluster
Agriculture	42600	71900	0.688	36300	52100	96300	150000	3
Anthropology	36800	61500	0.671	33800	45500	89300	138000	1
Architecture	41600	76800	0.846	50600	62200	97000	136000	3
Art History	35800	64900	0.813	28800	42200	87400	125000	1
Biology	38800	64800	0.670	36900	47400	94500	135000	1
Business Management	43000	72100	0.677	38800	51500	102000	147000	3
Chemical Engineering	63200	107000	0.693	71900	87300	143000	194000	2
Chemistry	42600	79900	0.876	45300	60700	108000	148000	3
Civil Engineering	53900	90500	0.679	63400	75100	115000	148000	2
Communications	38100	70000	0.837	37500	49700	98800	143000	3
Computer Engineering	61400	105000	0.710	66100	84100	135000	162000	2
Computer Science	55900	95500	0.708	56000	74900	122000	154000	2
Construction	53700	88900	0.655	56300	68100	118000	171000	2
Criminal Justice	35000	56300	0.609	32200	41600	80700	107000	1
Drama	35900	56900	0.585	36700	41300	79100	153000	1
Economics	50100	98600	0.968	50600	70600	145000	210000	2
Education	34900	52000	0.490	29300	37900	73400	102000	1
Electrical Engineering	60900	103000	0.691	69300	83800	130000	168000	2
English	38000	64700	0.703	33400	44800	93200	133000	1
Film	37900	68500	0.807	33900	45500	100000	136000	1
Finance	47900	88300	0.843	47200	62100	128000	195000	3
Forestry	39100	62600	0.601	41000	49300	78200	111000	1
Geography	41200	65500	0.590	40000	50000	90800	132000	1
Geology	43500	79500	0.828	45000	59600	101000	156000	3
Graphic Design	35700	59800	0.675	36000	45500	80800	112000	1
Health Care Administration	38800	60600	0.562	34600	45600	78800	101000	1
History	39200	71000	0.811	37000	49200	103000	149000	3
Hospitality & Tourism	37800	57500	0.521	35500	43600	81900	124000	1
Industrial Engineering	57700	94700	0.641	57100	72300	132000	173000	2
Information Technology (IT)	49100	74800	0.523	44500	56700	96700	129000	3
Interior Design	36100	53200	0.474	35700	42600	72500	107000	1
International Relations	40900	80900	0.978	38200	56000	111000	157000	3
Journalism	35600	66700	0.874	38400	48300	97700	145000	1
Management Information Systems (MIS)	49200	82300	0.673	45300	60500	108000	146000	3
Marketing	40800	79600	0.951	42100	55600	119000	175000	3
Math	45400	92400	1.035	45200	64200	128000	183000	3
Mechanical Engineering	57900	93600	0.617	63700	76200	120000	163000	2
Music	35900	55000	0.532	26700	40200	88000	134000	1

College Major	Start Median Salary	Mid Median Salary	Percent Growth	Percntl 10	Percntl 25	Percntl 75	Perctnl 90	Cluster
Nursing	54200	67000	0.236	47600	56400	80900	98300	1
Nutrition	39900	55300	0.386	33900	44500	70500	99200	1
Philosophy	39900	81200	1.035	35500	52800	127000	168000	3
Physician Assistant	74300	91700	0.234	66400	75200	108000	124000	2
Physics	50300	97300	0.934	56000	74200	132000	178000	2
Political Science	40800	78200	0.917	41200	55300	114000	168000	3
Psychology	35900	60400	0.682	31600	42100	87500	127000	1
Religion	34100	52000	0.525	29700	36500	70900	96400	1
Sociology	36500	58200	0.595	30700	40400	81200	118000	1
Spanish	34000	53100	0.562	31000	40000	76800	96400	1