Degrees That Pay You Back

Introduction

Are you wondering if that Philosophy major will help you pay the bills? Think you're set with an Engineering degree? Choosing a college major is a complex decision evaluating personal interest, difficulty, and career prospects. Your first paycheck right out of college might say a lot about your salary potential by mid-career. Whether you're in school or navigating the postgrad world, we are going to find out the short and long-term financial implications of this *major* decision.

The dataset

We'll be using data collected from a year-long survey of 1.2 million people with only a bachelor's degree available by the Wall Street Journal. It has the following column:

- College.Major
- Starting.Median.Salary
- Mid.Career.Median.Salary
- Career.Percent.Growth
- Percentile.10
- Percentile.25
- Percentile.75
- Percentile.90

```{r}
# Load relevant packages
library(tidyr)
library(dplyr)
library(readr)
library(ggplot2)
library(cluster)
library(factoextra)
# Read in the dataset
degrees <- read_csv("degrees-that-pay-back.csv", col_names=c("College.Major",
"Starting.Median.Salary", "Mid.Career.Median.Salary", "Career.Percent.Growth", "Percentile.10",
"Percentile.25", "Percentile.75", "Percentile.90"), skip=1)

Let's load the datasets and packages and take a look at the first few rows.
# Display the first few rows and a summary of the data frame
head(degrees)
summary(degrees)
```

```
College.Major      Starting.Median.Salary Mid.Career.Median.Salary Career.Percent.Growth
Length:50          Length:50              Length:50                Min.   : 23.40
Class :character   Class :character       Class :character         1st Qu.: 59.12
Mode  :character   Mode  :character       Mode  :character         Median : 67.80
                                                                   Mean   : 69.27
                                                                   3rd Qu.: 82.42
                                                                   Max.   :103.50
```

```
Percentile.10      Percentile.25      Percentile.75      Percentile.90
Length:50          Length:50          Length:50          Length:50
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

	College.Major	Starting.Median.Salary	Mid.Career.Median.Salary	Career.Percent.Growth	Percentile.10	Percentile.25	Percentile.75	Percentile.90
1	Accounting	$46,000.00	$77,100.00	67.6	$42,200.00	$56,100.00	$108,000.00	$152,000.00
2	Aerospace Engineering	$57,700.00	$101,000.00	75.0	$64,300.00	$82,100.00	$127,000.00	$161,000.00
3	Agriculture	$42,600.00	$71,900.00	68.8	$36,300.00	$52,100.00	$96,300.00	$150,000.00
4	Anthropology	$36,800.00	$61,500.00	67.1	$33,800.00	$45,500.00	$89,300.00	$138,000.00
5	Architecture	$41,600.00	$76,800.00	84.6	$50,600.00	$62,200.00	$97,000.00	$136,000.00
6	Art History	$35,800.00	$64,900.00	81.3	$28,800.00	$42,200.00	$87,400.00	$125,000.00
7	Biology	$38,800.00	$64,800.00	67.0	$36,900.00	$47,400.00	$94,500.00	$135,000.00
8	Business Management	$43,000.00	$72,100.00	67.7	$38,800.00	$51,500.00	$102,000.00	$147,000.00
9	Chemical Engineering	$63,200.00	$107,000.00	69.3	$71,900.00	$87,300.00	$143,000.00	$194,000.00
10	Chemistry	$42,600.00	$79,900.00	87.6	$45,300.00	$60,700.00	$108,000.00	$148,000.00
11	Civil Engineering	$53,900.00	$90,500.00	67.9	$63,400.00	$75,100.00	$115,000.00	$148,000.00
12	Communications	$38,100.00	$70,000.00	83.7	$37,500.00	$49,700.00	$98,800.00	$143,000.00
13	Computer Engineering	$61,400.00	$105,000.00	71.0	$66,100.00	$84,100.00	$135,000.00	$162,000.00
14	Computer Science	$55,900.00	$95,500.00	70.8	$56,000.00	$74,900.00	$122,000.00	$154,000.00
15	Construction	$53,700.00	$88,900.00	65.5	$56,300.00	$68,100.00	$118,000.00	$171,000.00

Notice that our salary data is in currency format, which R considers a string. We will remove special characters using the *gsub* function and convert all of our columns except *College.Major* to numeric, and we also convert the *Career.Percent.Growth* column to a decimal value.

```{r}
# Clean up the data
degrees_clean <- degrees %>%
  mutate_at(vars(Starting.Median.Salary:Percentile.90),
            function(x) as.numeric(gsub("[\\$,]", "", x))) %>%
            mutate(Career.Percent.Growth = Career.Percent.Growth/100)
```

Empirical Analysis

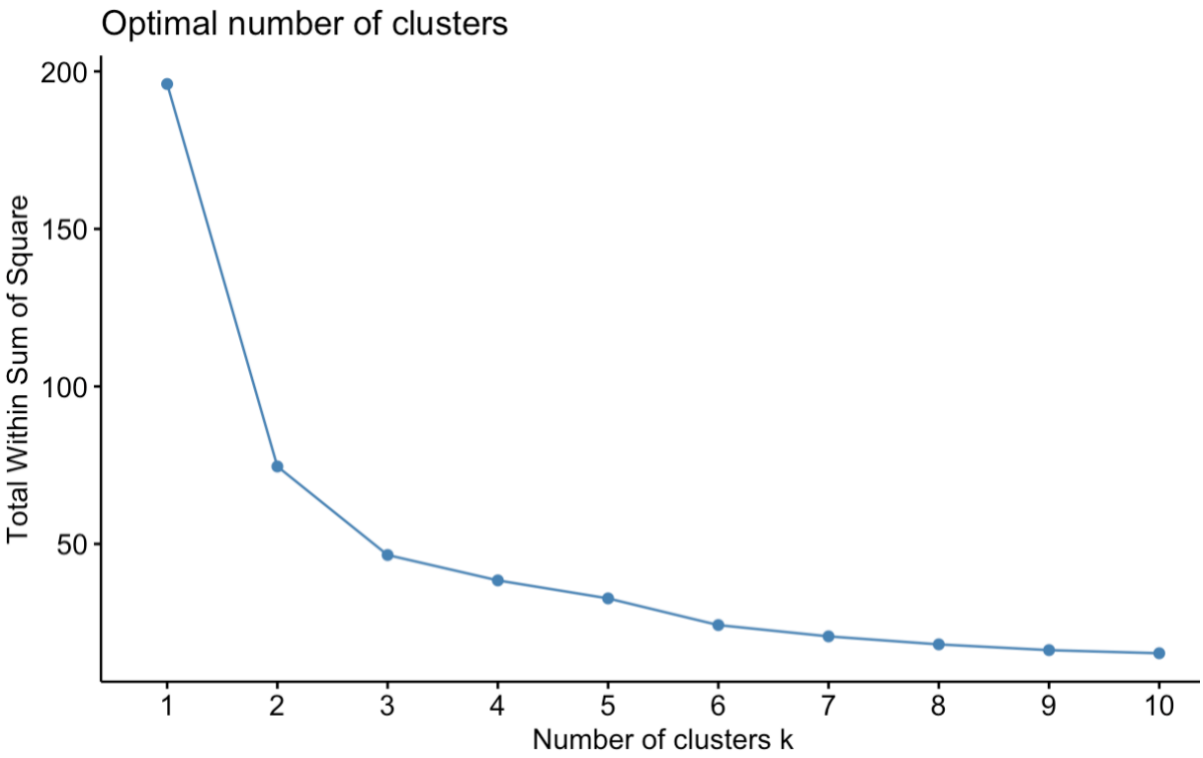
Now that we have a more manageable dataset, let's begin our clustering analysis by determining how many clusters we should be modeling. The best number of clusters for an unlabeled dataset is not always a clear-cut answer, but fortunately, several techniques help us optimize. We'll work with three different methods to compare recommendations:

- Elbow Method
- Silhouette Method
- Gap Statistic Method

First, up will be the **Elbow Method**. This method plots the percent variance against the number of clusters. The "elbow" bend of the curve indicates the optimal point at which adding more clusters will no longer explain a significant amount of the variance. To begin, let's select and scale the following features to base our clusters on*: Starting.Median.Salary, Mid.Career.Median.Salary, Perc.10, and Perc.90*. Then we'll use the fancy *fviz_nbclust* function from the factoextra library to determine and visualize the optimal number of clusters.

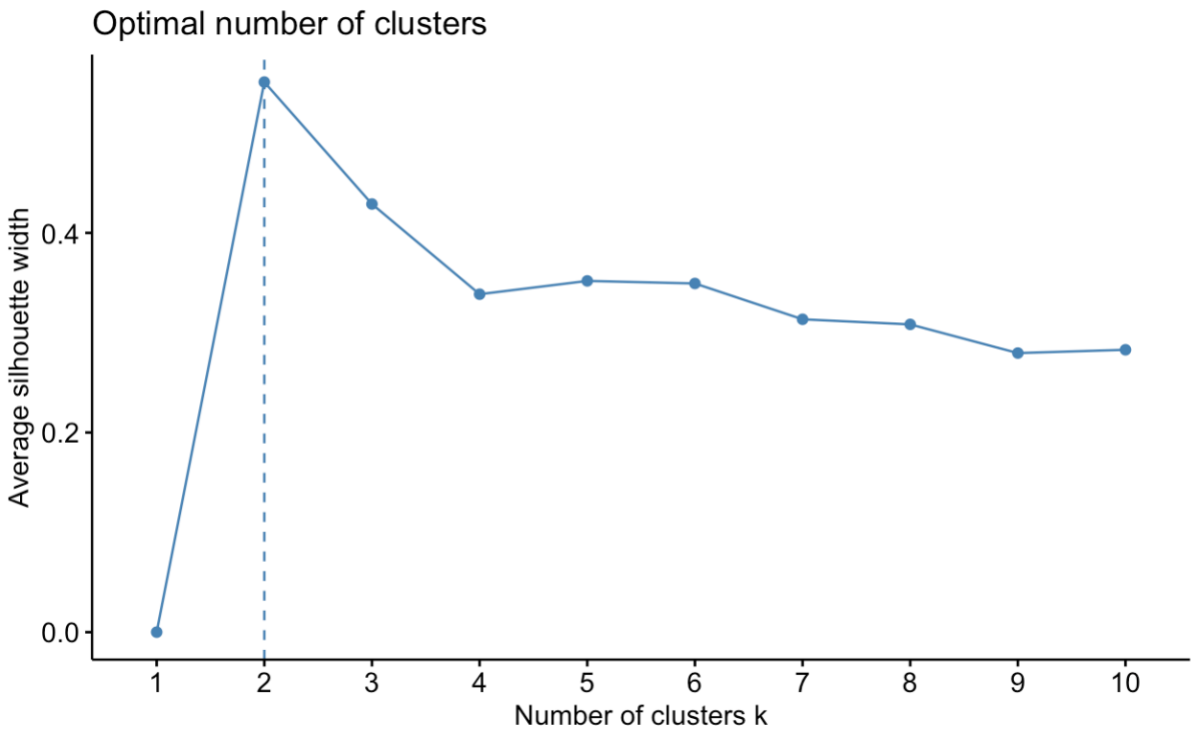
```{r}
# Select and scale the relevant features and store as k_means_data
k_means_data <- degrees_clean %>%
  select(Starting.Median.Salary, Mid.Career.Median.Salary,
```

```
      Percentile.10, Percentile.90) %>%
  scale()
# Run the fviz_nbclust function with our selected data and methos "wss"
elbow_method <- fviz_nbclust(k_means_data, kmeans, method = "wss")
# View the plot
elbow_method
```



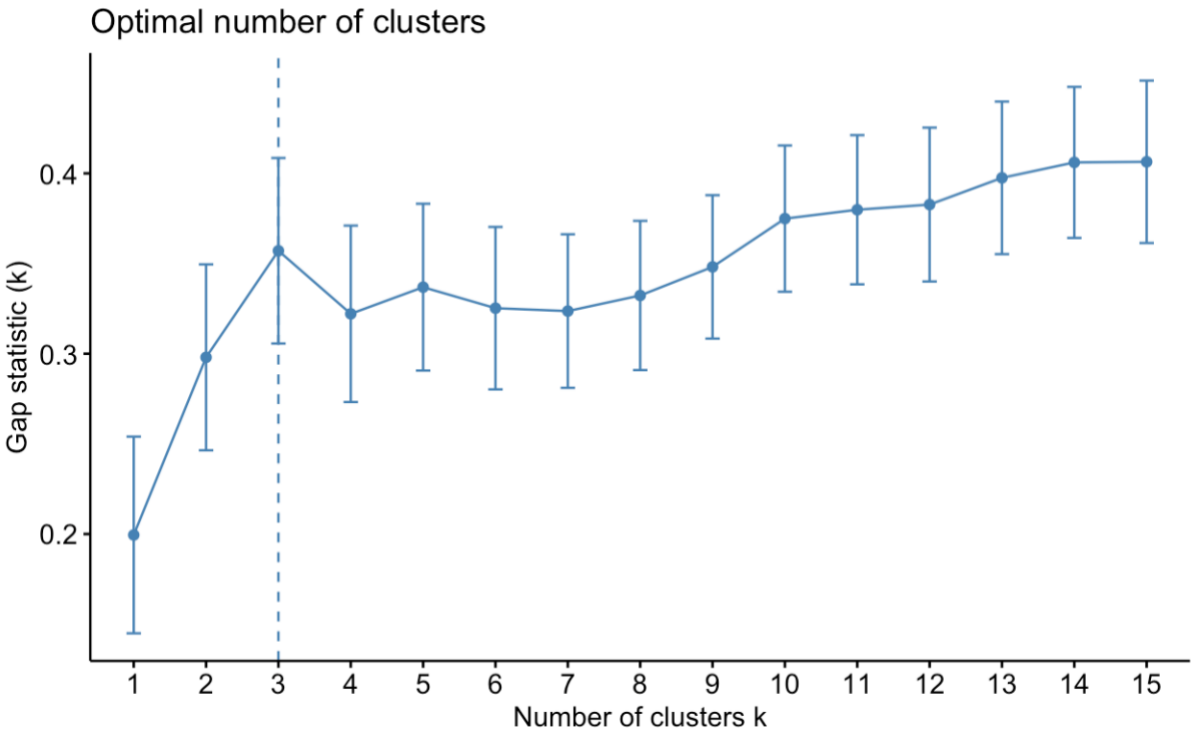
Wow, that fviz_nbclust function was pretty nifty. Instead of needing to "manually" apply the elbow method by running multiple k_means models and plotting the calculated the total within-cluster sum of squares for each potential value of k, fviz_nbclust handled all of this for us behind the scenes. Now let's try the Silhouette Method, The Silhouette Method will evaluate the quality of clusters by how well each point fits within a cluster, maximizing the average "silhouette" width.

```{r}
# Run the fviz_nbclust function with the method "silhouette"
sillhouette_method <- fviz_nbclust(k_means_data, kmeans,
                                   method = "silhouette")
# View the plot
sillhouette_method
```



Both plots above it don't show the same optimal number of clusters. Let's try our final method. For our last method, let's see what the **Gap Statistic Method** has to say about this. The Gap Statistic Method will compare the total variation within clusters for different values of *k* to the null hypothesis, maximizing the "gap." The "null hypothesis" refers to a uniformly distributed *simulated reference* dataset with no observable clusters, generated by aligning with the principle components of our original dataset. In other words, how much k clusters explain more variance in our dataset than in a fake dataset where all majors have equal salary potential?

```{r}
# Use the clusGap function to apply the Gap Statistic Method
gap_stat <- clusGap(k_means_data, FUN = kmeans,
nstart = 25, K.max = 15, B = 50)
# Use the fviz_gap_stat function to visualize the results
gap_stat_method <- fviz_gap_stat(gap_stat)
# View the plot
gap_stat_method
```



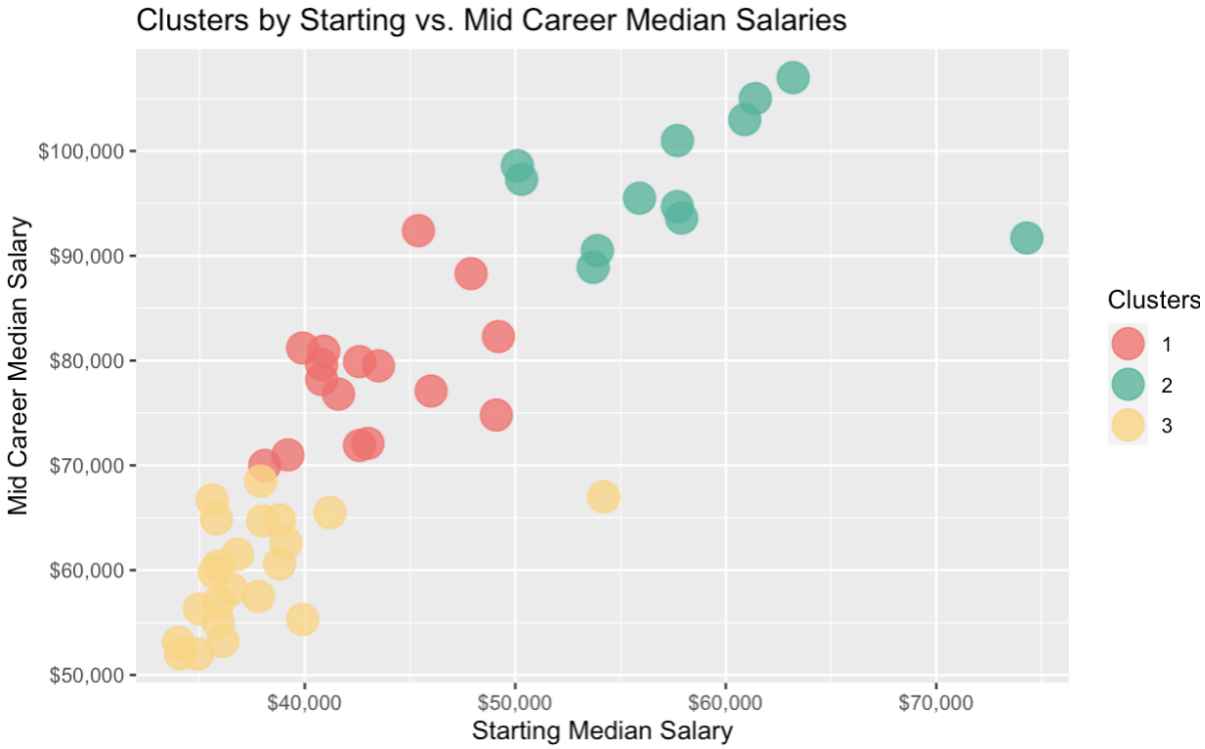
Looks like the Gap Statistic Method agreed with the Elbow Method! According to majority rule, let's use 3 for our optimal number of clusters. With this information, we can now run our k-means algorithm on the selected data. We will then add the resulting cluster information to label our original dataframe.

```{r}
# Set a random seed
set.seed(111)
# Set k equal to the optimal number of clusters
num_clusters <- 3
# Run the k-means algorithm
k_means <- kmeans(k_means_data, centers = num_clusters,
iter.max = 15, nstart = 25)
# Label the clusters of degrees_clean
degrees_labeled <- degrees_clean %>%
mutate(clusters = k_means$cluster)
```

Let's take a look at how each cluster compares in Starting vs. Mid Career Median Salaries. What do the clusters say about the relationship between Starting and Mid Career salaries?

```{r}
# Graph the clusters by Starting and Mid Career Median Salaries
career_growth <- ggplot(degrees_labeled,
aes(x=Starting.Median.Salary,y=Mid.Career.Median.Salary, color = factor(clusters))) +
  geom_point(alpha = 4/5, size = 6) +
  scale_x_continuous(labels = scales::dollar) +
  scale_y_continuous(labels = scales::dollar) +
  xlab("Starting Median Salary") +
  ylab("Mid Career Median Salary") +
  scale_color_manual(name = "Clusters", values = c("#EE716D","#55B49E",
                "#F9D585")) +
  ggtitle("Clusters by Starting vs. Mid Career Median Salaries")
```

```
# View the plot
career_growth
```



Unsurprisingly, most data points are hovering in the bottom left corner, with a relatively linear relationship. In other words, the higher your starting salary, the higher your mid-career salary. The three clusters provide a level of delineation that intuitively supports this.

How might the clusters reflect potential mid-career growth? There are also a couple of curious outliers from clusters 1 and 3, and perhaps this can be explained by investigating mid-career career percentiles further and exploring which majors fall in each cluster.

Right now, we have a column for each percentile salary value. To visualize the clusters and majors by mid-career percentiles, we'll need to reshape the degrees_labeled data using tidyr's gather function to make a percentile *key* column and a salary *value* column to use for the axes of our following graphs. We'll then be able to examine the contents of each cluster to see what stories they might be telling us about the majors.

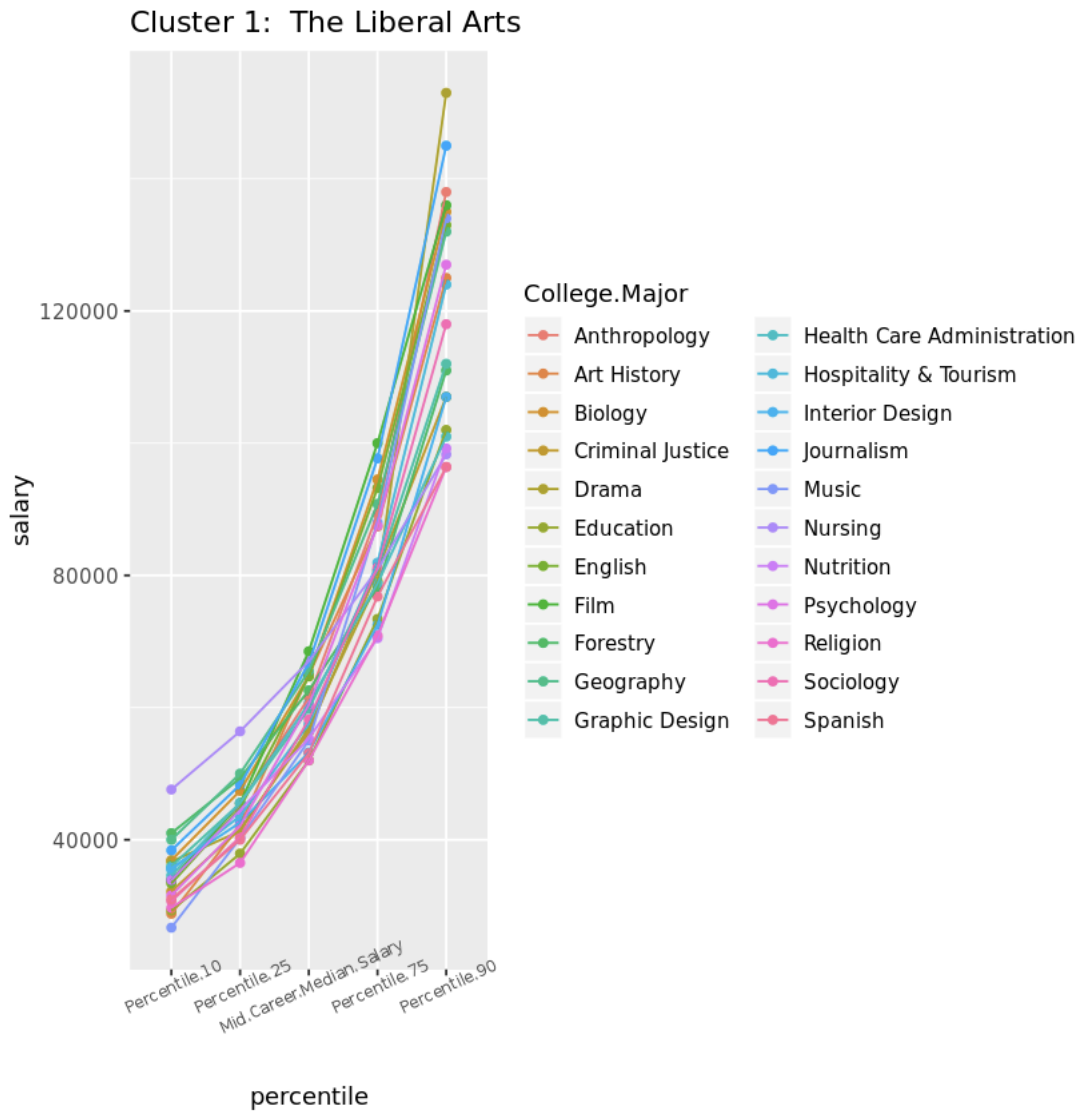
```
# Use the gather function to reshape degrees and
# use mutate() to reorder the new percentile column
degrees_perc <- degrees_labeled %>%
select(College.Major, Percentile.10, Percentile.25, Mid.Career.Median.Salary,
Percentile.75, Percentile.90, clusters) %>%
gather(key =percentile, value =salary, -c(College.Major, clusters)) %>%
mutate(percentile = factor(percentile, levels=c('Percentile.10','Percentile.25',
'Mid.Career.Median.Salary','Percentile.75','Percentile.90')))
```

Let's graph Cluster 1 and examine the results. These Liberal Arts majors may represent the lowest percentiles with limited growth opportunities, but there is hope for those who make it! Music is our

riskiest major with the lowest 10th percentile salary, but Drama wins the highest growth potential in the 90th percentile for this cluster. Nursing is the outlier culprit of cluster number 1, with a higher safety net in the lowest percentile to the median. Otherwise, this cluster does represent the majors with limited growth opportunities.

```{r}
# Graph the majors of Cluster 1 by percentile
cluster_1 <-  ggplot(degrees_perc[degrees_perc$clusters==1,],
              aes(x=percentile,y=salary,
              group=College.Major, color=College.Major, order=salary)) +
              geom_point() +
              geom_line() +
              ggtitle('Cluster 1:  The Liberal Arts') +
              theme(axis.text.x = element_text(size=7, angle=25))

# View the plot
cluster_1
```

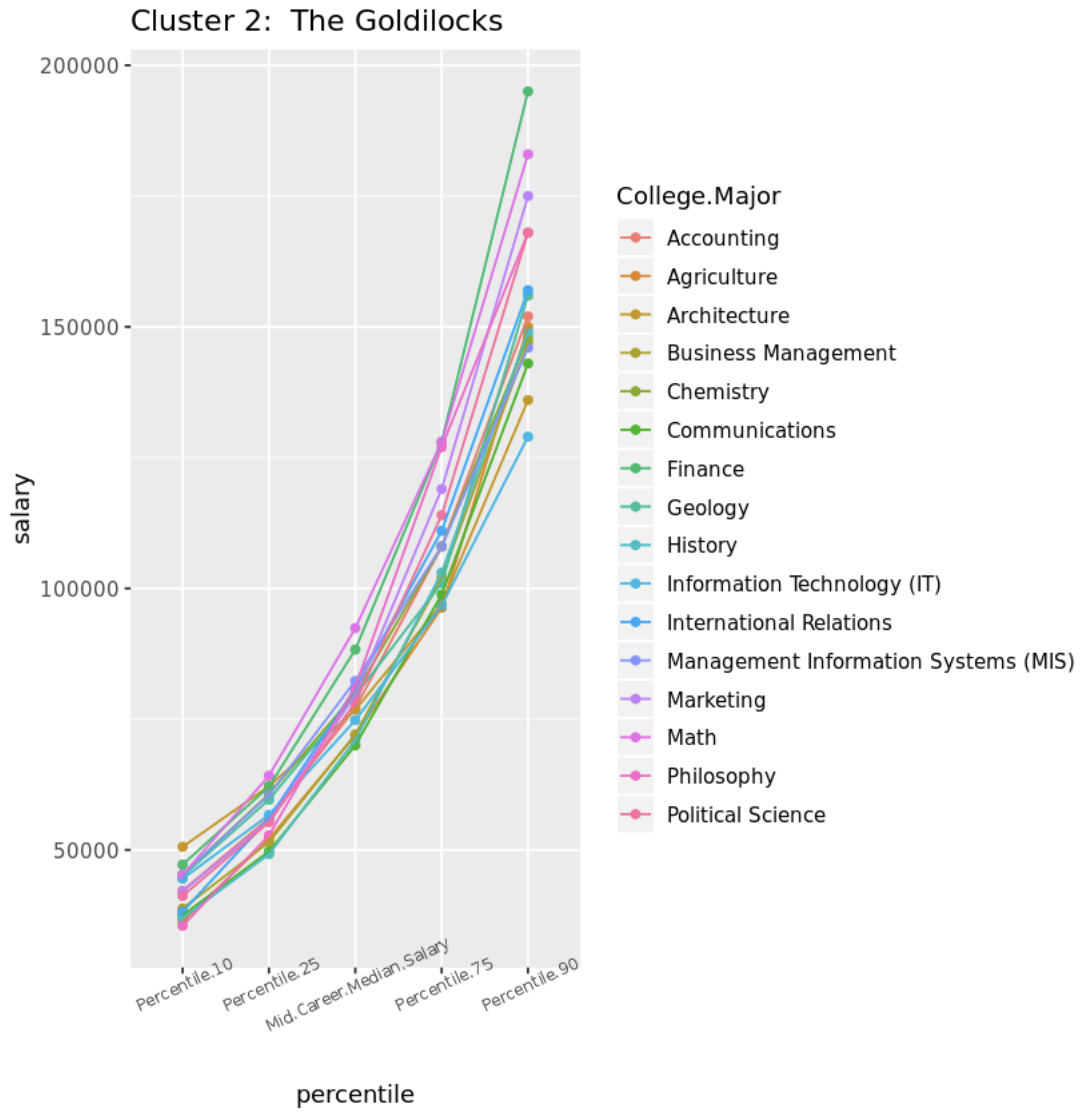


On to Cluster 2, right in the middle! Accountants are known for having stable job security, but once you're in the big leagues, you may be surprised to find that Marketing or Philosophy can ultimately

result in higher salaries. The majors of this cluster are pretty middle of the road in our dataset, starting not too low and not too high in the lowest percentile. However, this cluster also represents the majors with the most significant difference between the lowest and highest percentiles.

```
# Modify the previous plot to display Cluster 2
cluster_2 <-  ggplot(degrees_perc[degrees_perc$clusters==2,],
              aes(x=percentile,y=salary,
              group=College.Major, color=College.Major, order=salary)) +
              geom_point() +
              geom_line() +
              ggtitle('Cluster 2:  The Goldilocks') +
              theme(axis.text.x = element_text(size=7, angle=25))

# View the plot
cluster_2
```

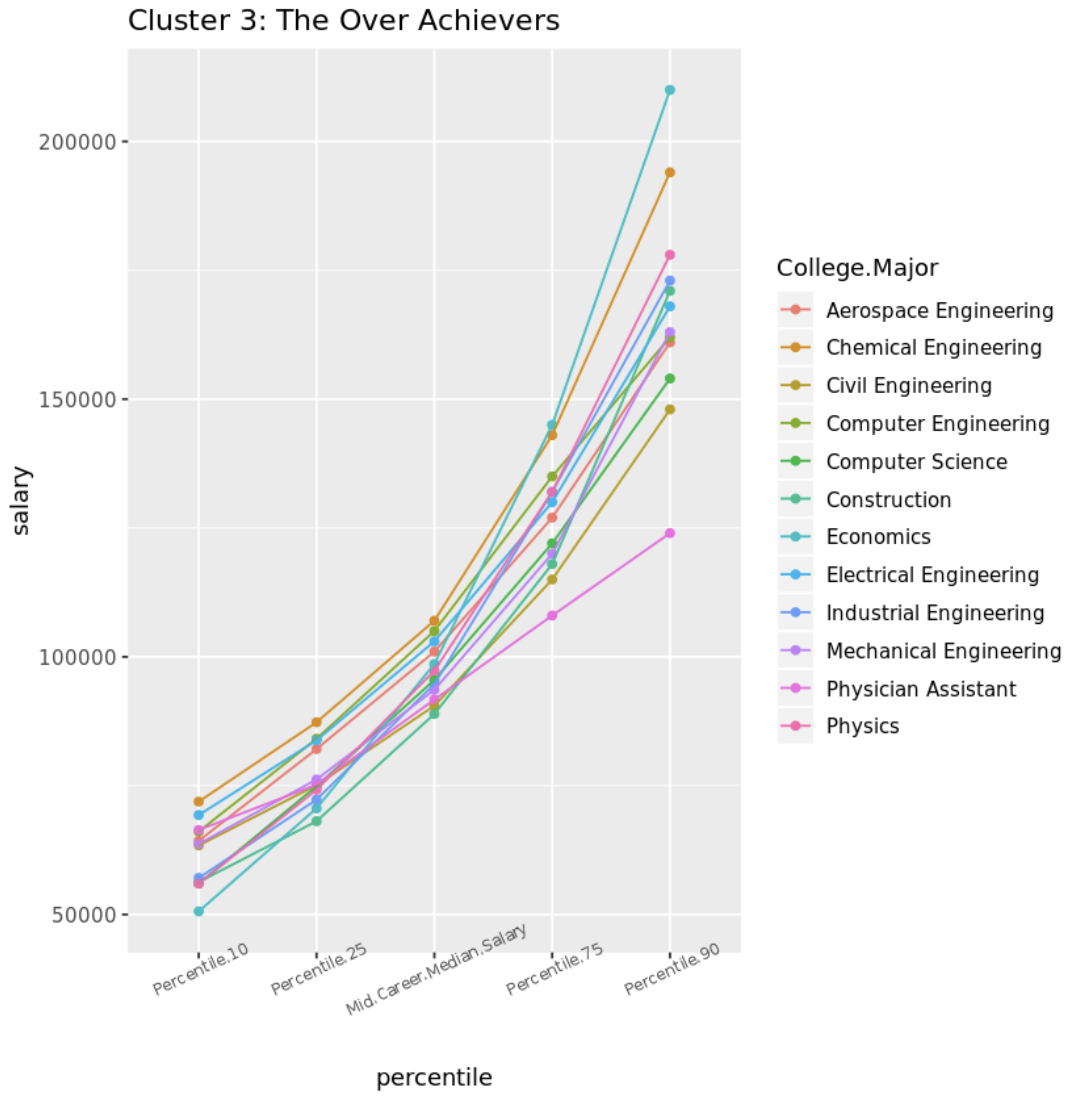


Finally, let's visualize Cluster 3. If you want financial security, these are the majors to choose from. Besides our one previously observed outlier now identifiable as Physician Assistant lagging in the highest percentiles, these heavy hitters and solid engineers represent the highest growth potential in the 90th percentile, as well as the best security in the 10th percentile rankings.

```{r}
```

```
# Modify the previous plot to display Cluster 3
degrees_perc3 <-subset(degrees_perc, clusters==3)
cluster_3 <-  ggplot(degrees_perc2,
              aes(x=percentile,y=salary,
              group=College.Major, color=College.Major, order=salary)) +
              geom_point() +
              geom_line() +
              ggtitle("Cluster 3: The Over Achievers") +
              theme(axis.text.x = element_text(size=7, angle=25))

# View the plot
cluster_3
```



Conclusion

Which two careers tied for the highest career percent growth? While it's tempting to focus on starting career salaries when choosing a major, it's important to also consider the growth potential down the road. Keep in mind that whether a major fall into the Liberal Arts, Goldilocks, or Over Achievers cluster, one's financial destiny will undoubtedly be influenced by numerous other factors, including the school attended, location, passion or talent for the subject, and of course the actual

career(s) pursued. Surprisingly the highest career percent growth jobs do not come from cluster 3 but cluster 2(Philosophy and Math).

```{r}
arrange(degrees_labeled, desc(Career.Percent.Growth)) %>%
  select(College.Major, Starting.Median.Salary, Mid.Career.Median.Salary, Career.Percent.Growth, clusters)
```

	College.Major	Starting.Median.Salary	Mid.Career.Median.Salary	Career.Percent.Growth	clusters
1	Math	45400	92400	1.035	1
2	Philosophy	39900	81200	1.035	1
3	International Relations	40900	80900	0.978	1
4	Economics	50100	98600	0.968	2
5	Marketing	40800	79600	0.951	1
6	Physics	50300	97300	0.934	2

Source:

- http://online.wsj.com/public/resources/documents/info-Degrees_that_Pay_you_Back-sort.html?mod=article_inline
- https://www.wsj.com/articles/SB121746658635199271
- https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/