DegreesThatPayYouBack

Which college majors will pay the bills?

Wondering if that Philosophy major will really 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, join me as we explore the short and long term financial implications of this major decision.

In this notebook, we'll be using data collected from a year-long survey of 1.2 million people with only a bachelor's degree by PayScale Inc., made available here by the Wall Street Journal for their article Ivy League's Big Edge: Starting Pay. After doing some data clean up, we'll compare the recommendations from three different methods for determining the optimal number of clusters, apply a k-means clustering analysis, and visualize the results.

To begin, let's prepare by loading the following packages: tidyverse, dplyr, readr, ggplot2, cluster, and factoextra. We'll then import the data from degrees-that-pay-back.csv (which is stored in a folder called datasets), and take a quick look at what we're working with.

Read in the dataset

```
## Read in the dataset
degrees <- read_csv("./data/degrees-that-pay-back.csv",
col_names= c("College.Major", "Starting.Median.Salary", "Mid.Career.Median.Salary", "Career.Percent.Grou
```

Display the first few rows and a summary of the data frame

```
head(degrees)
```

```
## # A tibble: 6 x 8
##   College.Major Starting.Median~ Mid.Career.Medi~ Career.Percent.~ Percentile.10
##   <chr>         <chr>            <chr>                       <dbl> <chr>
## 1 Accounting    $46,000.00       $77,100.00                   67.6 $42,200.00
## 2 Aerospace En~ $57,700.00       $101,000.00                  75   $64,300.00
## 3 Agriculture   $42,600.00       $71,900.00                   68.8 $36,300.00
## 4 Anthropology  $36,800.00       $61,500.00                   67.1 $33,800.00
## 5 Architecture  $41,600.00       $76,800.00                   84.6 $50,600.00
## 6 Art History   $35,800.00       $64,900.00                   81.3 $28,800.00
## # ... with 3 more variables: Percentile.25 <chr>, Percentile.75 <chr>,
## #   Percentile.90 <chr>
```

```
summary(degrees)
```

```
##  College.Major      Starting.Median.Salary Mid.Career.Median.Salary
##  Length:50          Length:50              Length:50
##  Class :character   Class :character       Class :character
```

```
##  Mode  :character   Mode  :character       Mode  :character
##
##
##
##  Career.Percent.Growth Percentile.10      Percentile.25      Percentile.75
##  Min.   : 23.40        Length:50          Length:50          Length:50
##  1st Qu.: 59.12        Class :character   Class :character   Class :character
##  Median : 67.80        Mode  :character   Mode  :character   Mode  :character
##  Mean   : 69.27
##  3rd Qu.: 82.42
##  Max.   :103.50
##  Percentile.90
##  Length:50
##  Class :character
##  Mode  :character
##
##
##
```

Currency and strings and percents, oh my!

Notice that our salary data is in currency format, which R considers a string. Let's strip those special characters using the gsub function and convert all of our columns except College.Major to numeric.

While we're at it, we can also convert the Career.Percent.Growth column to a decimal value.

```
# 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)
```

The elbow method

Great! 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 there are several techniques to 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.