K-means Clustering Analysis for Salary Increase By Major

Ramesh K Patil

13/06/2020

Executive Summary:

Does Salary Increase dependent on College Major? Does selecting college major has any sort of impact on salary increase by major while starting and mid career. Let's analyse salary increase potential based on college major using K-means Cluster Analysis.PayScale Inc. done a year long survey of 1.2 million people with bachelor degree only, this data helps us understand how's the major plays role in Salary increase. This data wa spublished by The Wall Street Journal and that is spanned across salaries for colleges by type, salaries for colleges by region and Degrees that pay you back. We will focus on Degrees that pay you back dataset to get answer to our questions.

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. It is popular for cluster analysis in data mining. k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, better Euclidean solutions can be found using k-medians and k-medoids.

K-Means clustering is a fast, robust, and simple algorithm that gives reliable results when data sets are distinct or well separated from each other in a linear fashion. It is best used when the number of cluster centers, is specified due to a well-defined list of types shown in the data.

Project Objectives:

1. Use K-means Clustering algorithm to perform analysis about Salary Increase by College Major.
2. Use different methods to have Optimal number of clusters which is needed for K-means Cluster Analysis.

Dataset Used for Analysis:

We will use data from an year-long survey of 1.2 million people with only a bachelor's degree by PayScale Inc., which is available at ,

http://online.wsj.com/public/resources/documents/info-Degrees_that_Pay_you_Back-sort.html?mod=article_inline

on The Wall Street Journal with title "Salary Increase By Major".

Data Wrangling:

Web Scraping

Let's start our analysis now by getting "Salary Increase By Major" data by web scraping it from the URL,

http://online.wsj.com/public/resources/documents/info-Degrees_that_Pay_you_Back-sort.html?mod=article_inline which is available on

The Wall Street Journal website public resources.

There are different set of data available on this article whereas we will be only focusing on "Salary Increase By Major" tab.

```
# Note: this process could take a couple of seconds

if(!require(rvest)) install.packages("rvest", repos = "http://cran.us.r-project.org")

# scraping our dataset from the wall street journal from article,
# We will be using Undergraduate major that pay you back dataset out of Salary Increase By Major
# Set from The Wall Street Journal

url<-"http://online.wsj.com/public/resources/documents/info-Degrees_that_Pay_you_Back-sort.html?mod=arti

pre_data <- read_html(url)

nodes <- pre_data %>% html_nodes('table')

# locate table of interest from nodes "xml_nodeset"
majortab <- nodes[[7]]

# create a raw data we scraped from WTJ page
raw_dataset <- html_table(majortab)

# Take a Look at few rows of this data
head(raw_dataset)
```

```
##                      X1                    X2                      X3
## 1   Undergraduate Major Starting Median Salary Mid-Career Median Salary
## 2            Accounting             $46,000.00               $77,100.00
## 3 Aerospace Engineering             $57,700.00              $101,000.00
## 4           Agriculture             $42,600.00               $71,900.00
## 5          Anthropology             $36,800.00               $61,500.00
## 6          Architecture             $41,600.00               $76,800.00
##                                                 X4
## 1 Percent change from Starting to Mid-Career Salary
## 2                                             67.6
## 3                                             75.0
## 4                                             68.8
## 5                                             67.1
## 6                                             84.6
##                                 X5                            X6
## 1 Mid-Career 10th Percentile Salary Mid-Career 25th Percentile Salary
## 2                         $42,200.00                        $56,100.00
## 3                         $64,300.00                        $82,100.00
## 4                         $36,300.00                        $52,100.00
## 5                         $33,800.00                        $45,500.00
```

```
## 6                        $50,600.00                        $62,200.00
##                                  X7                                X8
## 1 Mid-Career 75th Percentile Salary Mid-Career 90th Percentile Salary
## 2                        $108,000.00                       $152,000.00
## 3                        $127,000.00                       $161,000.00
## 4                         $96,300.00                       $150,000.00
## 5                         $89,300.00                       $138,000.00
## 6                         $97,000.00                       $136,000.00
```

```
# Take a look at summary of this data
summary(raw_dataset)
```

```
##       X1                 X2                 X3                 X4
##  Length:51          Length:51          Length:51          Length:51
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##       X5                 X6                 X7                 X8
##  Length:51          Length:51          Length:51          Length:51
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
#Remove the unwanted variables from the environment
rm(pre_data, nodes, majortab, url)
```

Data Exploration:

Let's check and confirm if the data scraped from our source is in tidy format or not by inspecting dataset.

First few rows of our dataset are viewed as,

```
head(raw_dataset)
```

```
##                      X1                    X2                      X3
## 1   Undergraduate Major Starting Median Salary Mid-Career Median Salary
## 2            Accounting             $46,000.00               $77,100.00
## 3 Aerospace Engineering             $57,700.00              $101,000.00
## 4           Agriculture             $42,600.00               $71,900.00
## 5          Anthropology             $36,800.00               $61,500.00
## 6          Architecture             $41,600.00               $76,800.00
##                                                 X4
## 1 Percent change from Starting to Mid-Career Salary
## 2                                             67.6
## 3                                             75.0
## 4                                             68.8
## 5                                             67.1
## 6                                             84.6
##                                 X5                            X6
## 1 Mid-Career 10th Percentile Salary Mid-Career 25th Percentile Salary
## 2                         $42,200.00                        $56,100.00
## 3                         $64,300.00                        $82,100.00
## 4                         $36,300.00                        $52,100.00
## 5                         $33,800.00                        $45,500.00
## 6                         $50,600.00                        $62,200.00
##                                 X7                            X8
## 1 Mid-Career 75th Percentile Salary Mid-Career 90th Percentile Salary
## 2                        $108,000.00                       $152,000.00
## 3                        $127,000.00                       $161,000.00
## 4                         $96,300.00                       $150,000.00
## 5                         $89,300.00                       $138,000.00
## 6                         $97,000.00                       $136,000.00
```

Details of dataset can be viewed as summary,

```
summary(raw_dataset)
```

```
##       X1                 X2                 X3                 X4
##  Length:51          Length:51          Length:51          Length:51
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##       X5                 X6                 X7                 X8
##  Length:51          Length:51          Length:51          Length:51
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

Data Cleaning:

During data exploration we can see that ,first row of our dataset is a header row i.e. it contains column names of the table and rest will be data rows.

Also we can observe that data table columns in scraped data have column names as "X1" through "X8" and does not make any sense of data hold by respective column, so we need to provide meaningful names to these columns.

Let's work on resolving mentioned observations to make more sense.

```
if(!require(dplyr)) install.packages("dplyr", repos = "http://cran.us.r-project.org")

#As we see while exploring the dataset that column names are marked as X1 to X8 and not
#much informative
#Row one of dataset is contains info of stored data in respective column

colnames(raw_dataset) <- c("Undergraduate_Major", "Starting_Median_Salary", "Mid_Career_Median_Salary",
                           "Career_Percent_Growth", "Percentile_10", "Percentile_25", "Percentile_75",
                           "Percentile_90" )

raw_dataset <- raw_dataset[-1,]

rownames(raw_dataset) <- 1:nrow(raw_dataset)
```

Another observation that we need to take care of are,

1. All columns except undergraduade.major are in currency format and considered as String in R. So we need to remove currency symbol using the "gsub" function and convert them to numeric.
2. Convert "Career.Percent.Growth" column to a decimal value.

```
# strip '$' sign and convert Career_Percent_Growth to decimal value
degrees <- raw_dataset %>%
  mutate_at(vars(Starting_Median_Salary: Percentile_90), function(x)
    as.numeric(gsub('[\\$,]',"",x))) %>%
  mutate(Career_Percent_Growth = Career_Percent_Growth / 100)
```

Salary Increase By Major DataSet:

Please remember we had 51 rows when we got that data using web scraping from The Wall Street Journal web site.Final dataset is having 50 rows after cleaning of the data.

Let's inspect the first few rows of *degrees* data frame and some *summary* statistics:

```
degrees %>% as_tibble()
```

```
## # A tibble: 50 x 8
##    Undergraduate_M~ Starting_Median~ Mid_Career_Medi~ Career_Percent_~
##    <chr>                      <dbl>            <dbl>            <dbl>
##  1 Accounting                 46000            77100            0.676
##  2 Aerospace Engin~           57700           101000            0.75
##  3 Agriculture                42600            71900            0.688
##  4 Anthropology               36800            61500            0.671
##  5 Architecture               41600            76800            0.846
##  6 Art History                35800            64900            0.813
##  7 Biology                    38800            64800            0.67
##  8 Business Manage~           43000            72100            0.677
##  9 Chemical Engine~           63200           107000            0.693
```

```
## 10 Chemistry                    42600            79900            0.876
## # ... with 40 more rows, and 4 more variables: Percentile_10 <dbl>,
## #   Percentile_25 <dbl>, Percentile_75 <dbl>, Percentile_90 <dbl>
```

```
summary(degrees)
```

```
##  Undergraduate_Major Starting_Median_Salary Mid_Career_Median_Salary
##  Length:50           Min.   :34000          Min.   : 52000
##  Class :character    1st Qu.:37050          1st Qu.: 60825
##  Mode  :character    Median :40850          Median : 72000
##                      Mean   :44310          Mean   : 74786
##                      3rd Qu.:49875          3rd Qu.: 88750
##                      Max.   :74300          Max.   :107000
##  Career_Percent_Growth Percentile_10   Percentile_25   Percentile_75
##  Min.   :0.2340        Min.   :26700   Min.   :36500   Min.   : 70500
##  1st Qu.:0.5913        1st Qu.:34825   1st Qu.:44975   1st Qu.: 83275
##  Median :0.6780        Median :39400   Median :52450   Median : 99400
##  Mean   :0.6927        Mean   :43408   Mean   :55988   Mean   :102138
##  3rd Qu.:0.8243        3rd Qu.:49850   3rd Qu.:63700   3rd Qu.:118750
##  Max.   :1.0350        Max.   :71900   Max.   :87300   Max.   :145000
##  Percentile_90
##  Min.   : 96400
##  1st Qu.:124250
##  Median :145500
##  Mean   :142766
##  3rd Qu.:161750
##  Max.   :210000
```

Methods/Analysis:

Optimal number of clusters:

K-means clustering is the most commonly used unsupervised machine learning algorithm for partitioning a given data set into a set of k groups (i.e. k clusters), where k represents the number of groups pre-specified. Please be noted that, k needs to be specified and should be passed as an input.

Since we need to provide k i.e. number of clusters to our k-means clustering algorithm. In Normal situation, number of clusters i.e. value of k is passed one by one to find out optimal number of clusters. There are many different methods available to find out optimal number of clusters and then that optimal value of k will be used for partitioning our dataset.

We will use below mentioned three different methods to compare value of k to decide and get optimal number of k.

- Elbow Method
- Silhouette Method
- Gap Statistic Method

To begin, let's prepare by loading the following packages:

```
# Note: this process could take a couple of minutes

url <- "http://cran.us.r-project.org"
if(!require(tidyverse)) install.packages("tidyverse", repos = url)
if(!require(cluster)) install.packages("cluster", repos = url)
if(!require(factoextra)) install.packages("factoextra", repos = url)
if(!require(ggthemes)) install.packages("ggthemes", repos = url)
```

1. The elbow method:

Elbow Method is used to define clusters such that the total intra-cluster variation i.e. total within-cluster variation or total within-cluster sum of square, is minimized. that means this method plots the percent variance against the number of clusters. The bend in the knee or elbow of the curve indicates the optimal point at which adding more clusters will no longer explain a significant amount of the variance.

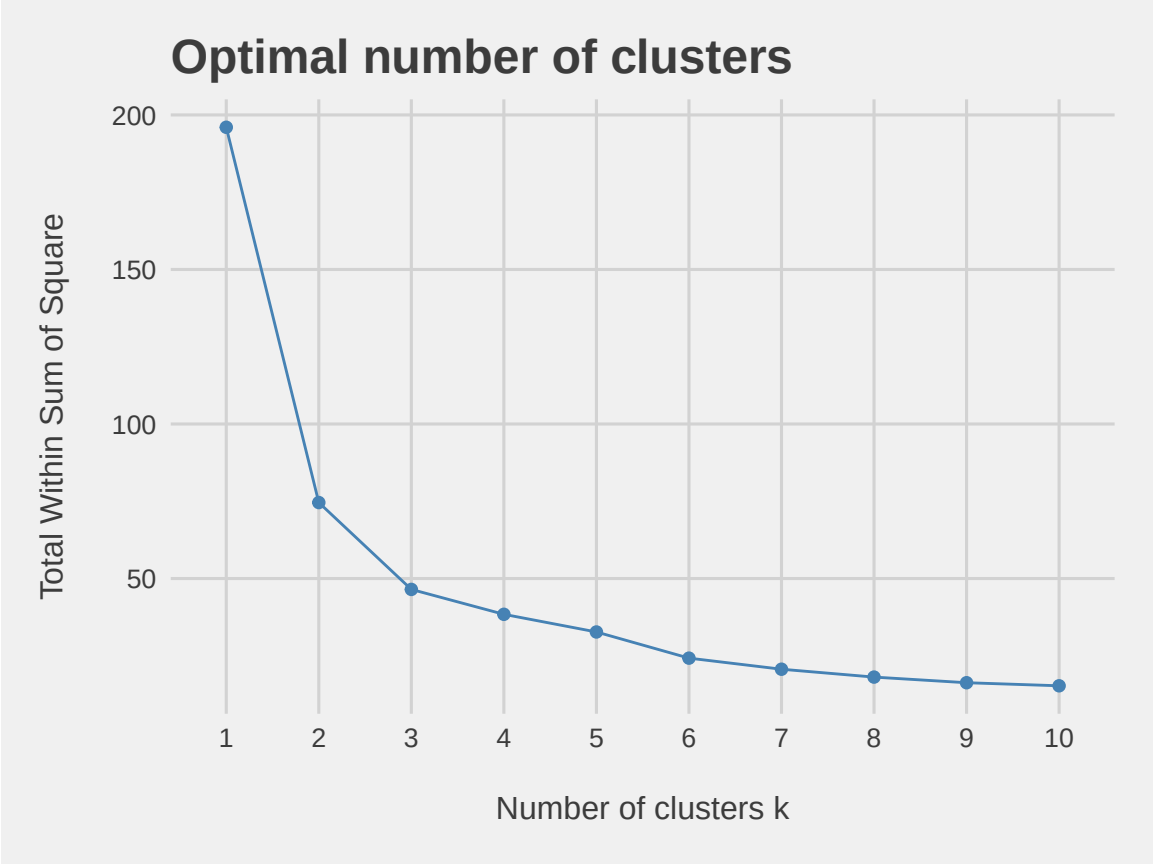
To begin, let's select and scale the following features to base our clusters on: *Starting.Median.Salary*, *Mid.Career.Median.Salary*, *Percentile.10*, and *Percentile.90.* Then we'll use the fancy *fviz_nbclust* function from the *factoextra* library to determine and visualize the optimal number of clusters.

```
# select and scale the relevant features and store as k_means_data

k_means_data <- degrees %>%
  select(Starting_Median_Salary, Mid_Career_Median_Salary, Percentile_10, Percentile_90) %>%
  scale()

# We will the fviz_nbclust function with selected data and method "wss"
# wss = total within-cluster sume of square

elbow_method <- fviz_nbclust(k_means_data, FUNcluster = kmeans, method = "wss")
elbow_method + theme_fivethirtyeight() +
  theme(axis.title = element_text()) +
  xlab('\nNumber of clusters k') +
  ylab('Total Within Sum of Square\n')
```


Optimal number of clusters
Total Within Sum of Square
200
150
100
50
1
2
3
4
5
6
7
8
9
10
Number of clusters k

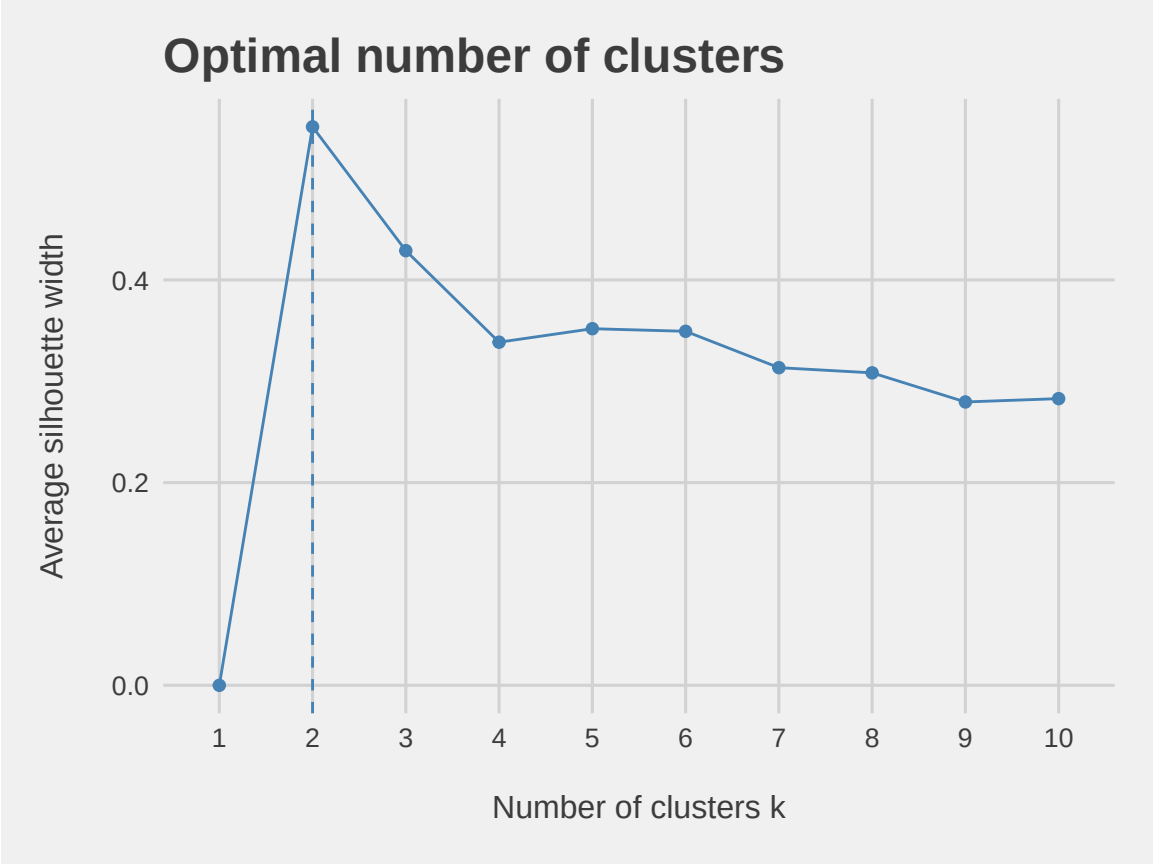
2. The Silhouette method:

The average silhouette approach measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering.

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 total within cluster sum of squares for each potential value of k, *fviz_nbclust* handled all of this for us behind the scenes. The *fviz_nbclust* can be used for the **Silhouette Method** as well.

```
# Use the fviz_nbclust function with the method "silhouette"

silhouette_method <- fviz_nbclust(k_means_data, FUNcluster = kmeans, method = "silhouette")
silhouette_method + theme_fivethirtyeight() +
  theme(axis.title = element_text()) +
  xlab('\nNumber of clusters k') +
  ylab('Average silhouette width\n')
```



3. The Gap Statistic method:

The gap statistic compares the total intracluster variation for different values of k with their expected values under null reference distribution of the data (i.e. a distribution with no obvious clustering). Here the "null reference distribution or 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 more variance is explained by k clusters in our dataset than in a fake dataset where all majors have equal salary potential?

To compute the gap statistic method, we can use the clusGap function which provides the gap statistic and standard error for an output.

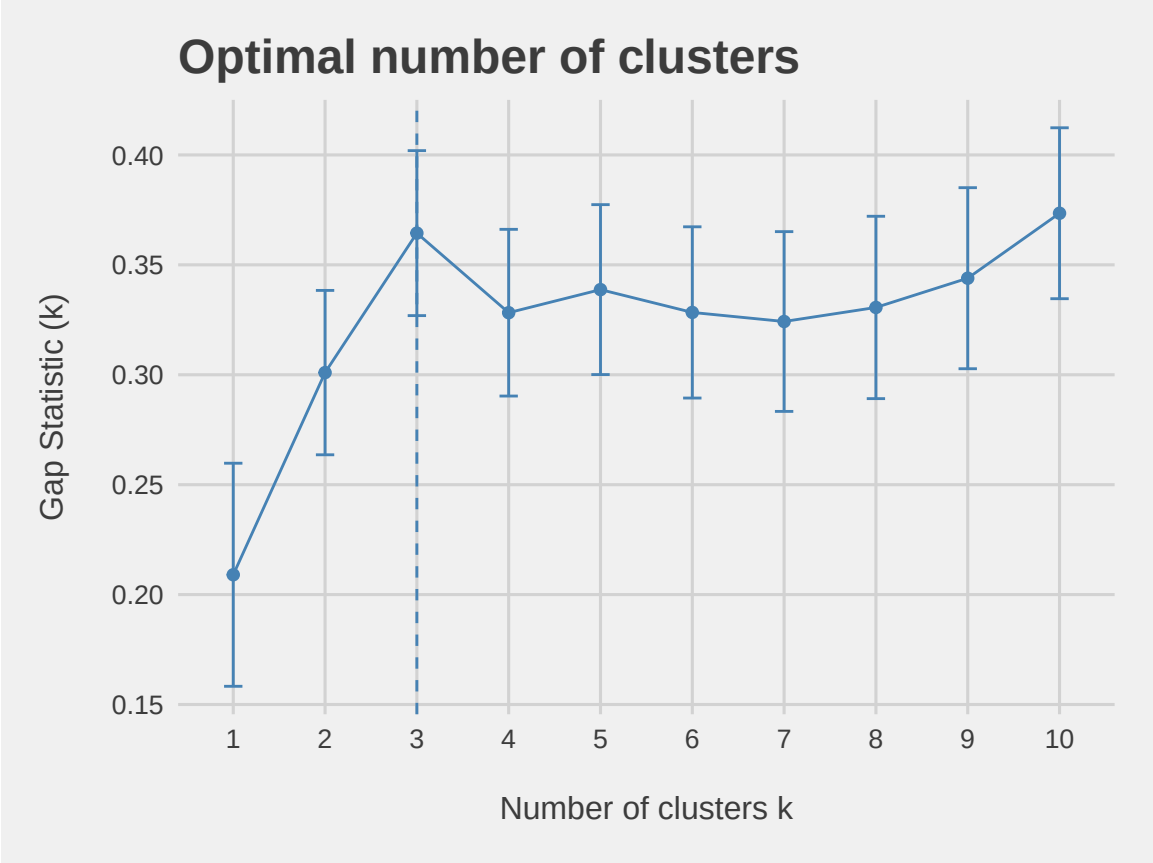
As last two methods gives us different optimal number of clusters, we need another method for majority to decide on optimal number of clusters.

```
# We will use the clusGap function to apply the Gap Statistic Method

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

# use the fviz_gap_stat function to vizualize the results

gap_stat_method <- fviz_gap_stat(gap_stat)
gap_stat_method + theme_fivethirtyeight() +
  theme(axis.title = element_text()) +
  xlab('\nNumber of clusters k') +
  ylab('Gap Statistic (k)\n')
```



K-means Algorithm:

As discussed earlier, k-means cluster algorithm needs k i.e. number of clusters as an input and we are using different methods to find that optimal number of cluster. Looking at outcome of three methods, we can confirm that we get same outcome i.e. optimal number of clusters for two methods, Elbow Method and Gap Statistic Method. Elbow Method and Gap Statistic methods gives us value of k i.e. number of clusters as 3.

With this k, 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.

```
# set a random seed
suppressWarnings(set.seed(111, sample.kind = 'Rounding'))

# set k equal to the optimal number of clusters as per our observarion
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 major to show which cluster major belongs to
degrees_labeled <- degrees %>%
  mutate(clusters = k_means$cluster)
```

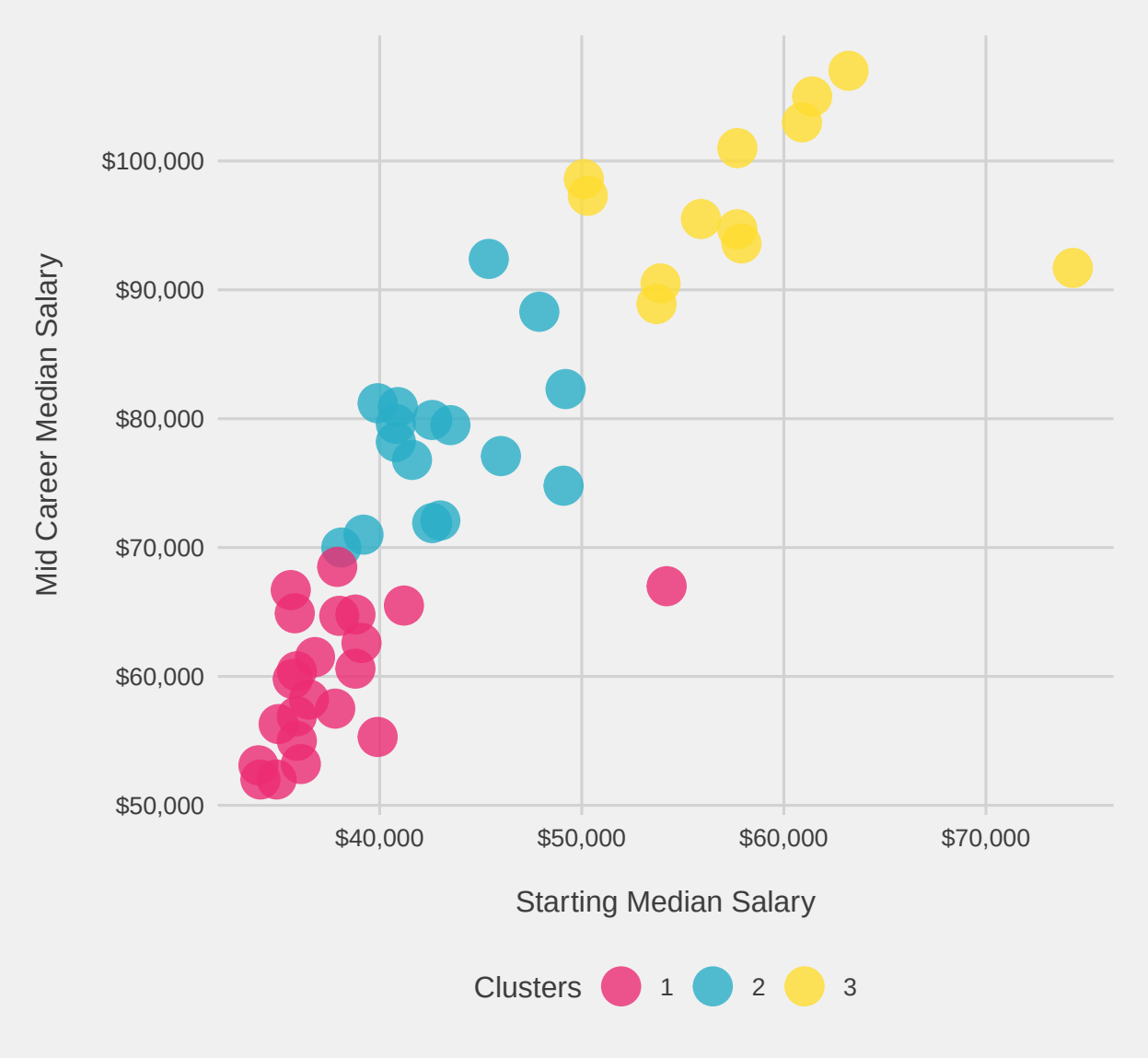
Data Visualization:

As we are now in the position to take our analysis to next level, lets take a look by visualizing our results. We will be first having 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?

```
# Plot 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 = 7) +
  scale_x_continuous(labels = scales::dollar) +
  scale_y_continuous(labels = scales::dollar) +
  scale_color_manual(name = "Clusters", values = c("#EC2C73", "#29AEC7", "#FFDD30"))
career_growth + theme_fivethirtyeight() +
  theme(axis.title = element_text()) +
  xlab('\nStarting Median Salary') +
  ylab('Mid Career Median Salary\n')
```



As we can visualize that most of the data points are floating in the top left corner showcasing relatively linear relationship. This explains that higher starting salary for an individual shows higher mid career salary.The three clusters provide a level of delineation that intuitively supports this.

How does it explains potential mid career growth in clusters?There are couple of odd outliers we see and that can be explained by investigating the mid career percentile further and exploring which majors fall in each cluster.

As of now, we have a column for each *percentile salary* value. In order 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 plots. We'll then be able to examine the contents of each cluster to see what stories they might be telling us about the majors.

```
# We will use the gather function to reshape degrees and use mutate() to reorder
# the new percentile column

degrees_perc <- degrees_labeled %>%
  select(Undergraduate_Major, Percentile_10, Percentile_25, Mid_Career_Median_Salary, Percentile_75,
  Percentile_90, clusters) %>%
  gather(key=percentile, value=salary, -c(Undergraduate_Major, clusters)) %>%
  mutate(percentile = factor(percentile, levels = c("Percentile_10", "Percentile_25",
  "Mid_Career_Median_Salary", "Percentile_75", "Percentile_90")))
```

1. The liberal arts cluster:

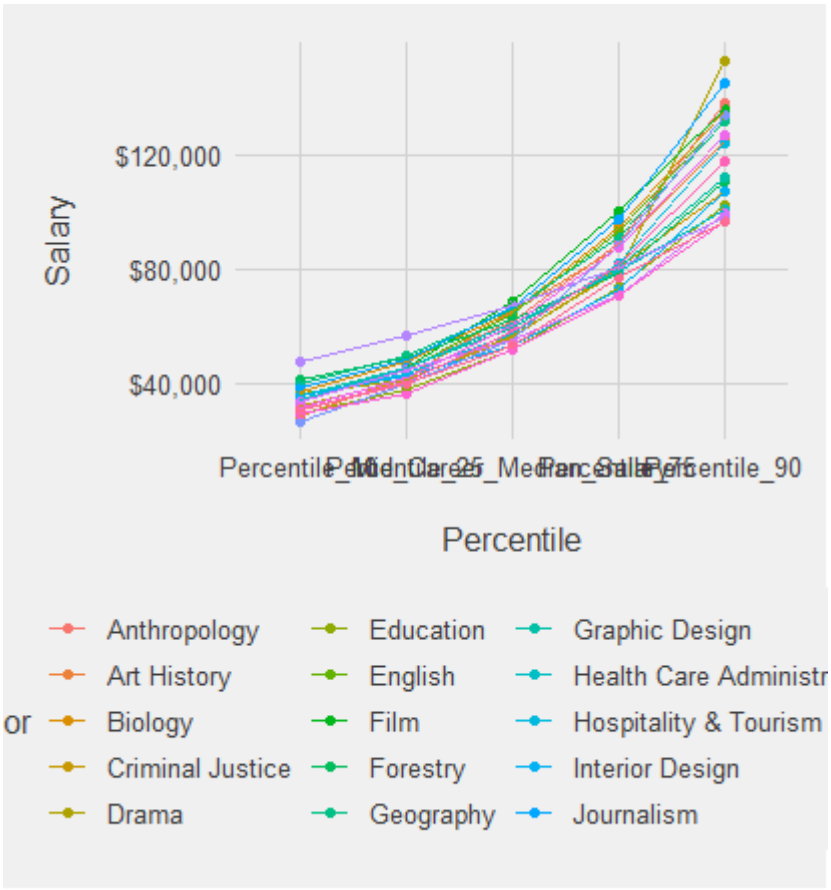
Let's plot Cluster I and analyse the results. These Liberal Arts majors may represent the lowest percentiles with limited growth opportunity, but there is hope for those who make it! Music is our riskiest major with the lowest 10th percentile salary, but Drama wins the highest growth potential in the 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 opportunity.

An aside: It's worth noting that most of these majors leading to lower-paying jobs are women-dominated, according to this **Glassdoor study**. According to the research:

"*The single biggest cause of the gender pay gap is occupation and industry sorting of men and women into jobs that pay differently throughout the economy. In the U.S., occupation and industry sorting explains 54 percent of the overall pay gap—by far the largest factor.*"

Does this imply that women are statistically choosing majors with lower pay potential, or do certain jobs pay less because women choose them?

```
cluster_I <-  ggplot(degrees_perc %>% filter(clusters == 1), aes(x=percentile, y=salary,
  group=Undergraduate.Major, color=Undergraduate.Major)) +
  geom_point() +
  geom_line() +
  theme(axis.text.x = element_text(size=7)) +
  scale_y_continuous(labels = scales::dollar)
cluster_I + theme_fivethirtyeight() + labs(color = "Undergraduate Major") +
  theme(axis.title = element_text()) +
  xlab('\nPercentile') +
  ylab('Salary\n')
```


$120,000
$80,000
$40,000
Salary
Percentile_90
Percentile
Anthropology
Art History
Biology
Criminal Justice
Drama
Education
English
Film
Forestry
Geography
Graphic Design
Health Care Administr
Hospitality & Tourism
Interior Design
Journalism

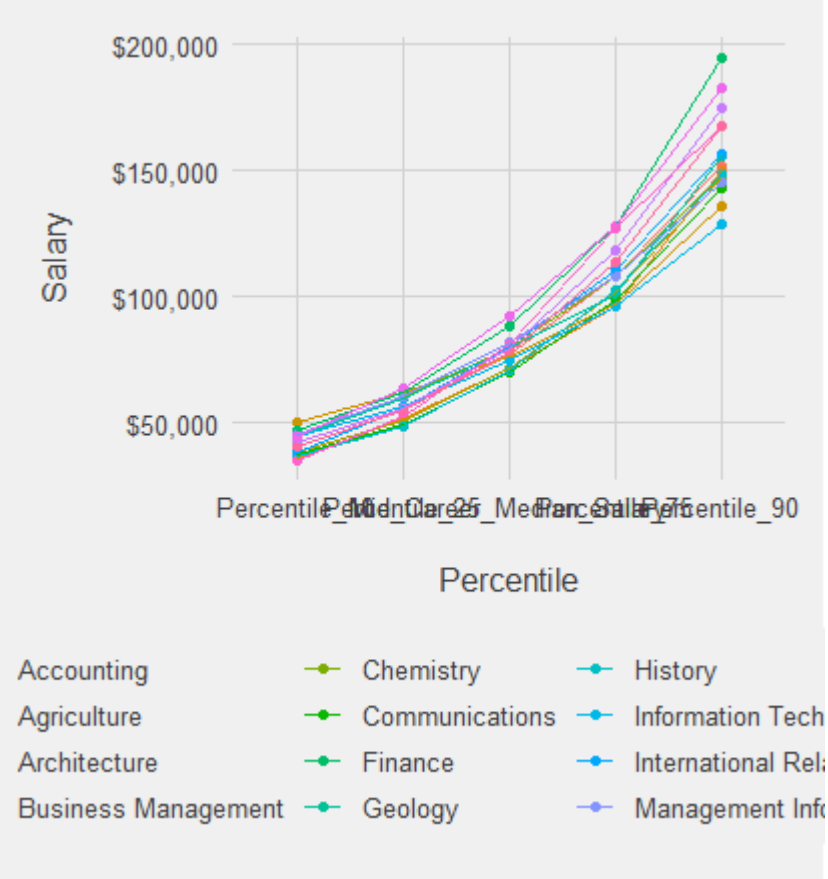
or

2. The goldilocks cluster:

As on Cluster II, 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 fairly middle of the road in our dataset, starting off not too low and not too high in the lowest percentile. However, this cluster also represents the majors with the greatest differential between the lowest and highest percentiles.

```
# Plot the majors of Cluster II by percentile

cluster_II <-  ggplot(degrees_perc %>% filter(clusters == 2), aes(x=percentile, y=salary,
  group=Undergraduate.Major, color=Undergraduate.Major)) +
  geom_point() +
  geom_line() +
  theme(axis.text.x = element_text(size=7)) +
  scale_y_continuous(labels = scales::dollar)
cluster_II + theme_fivethirtyeight() + labs(color = "Undergraduate Major") +
  theme(axis.title = element_text()) +
  xlab('\nPercentile') +
  ylab('Salary\n')
```

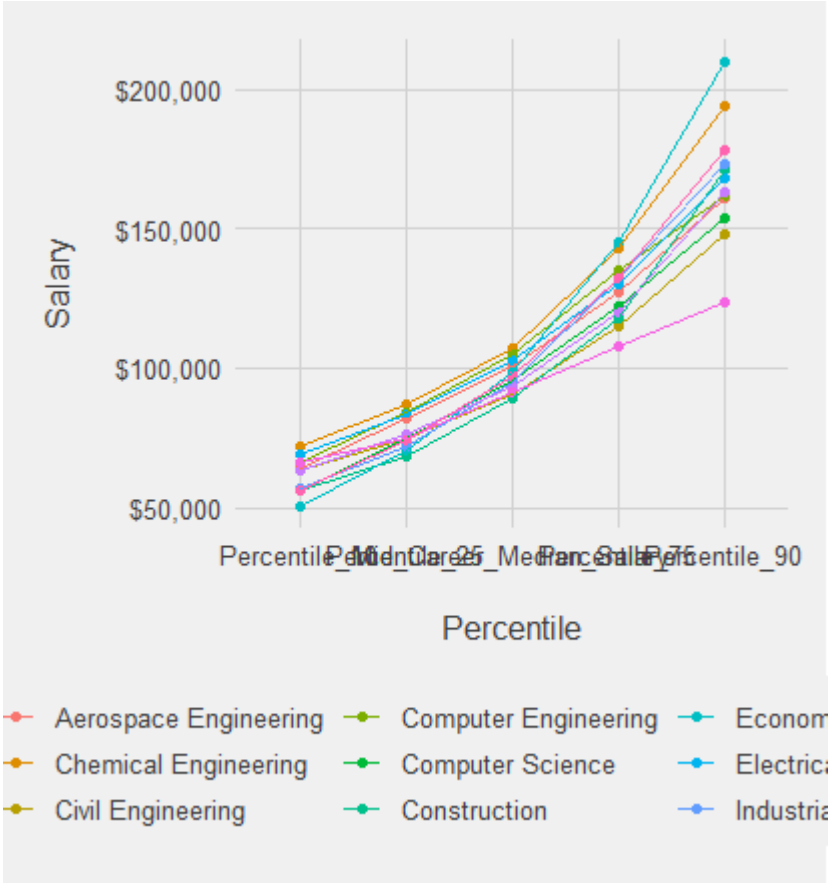


3. The over achiever cluster:

Finally, let's plot and analyse Cluster III. 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.

```
# plot the majors of Cluster III by percentile

cluster_III <-  ggplot(degrees_perc %>% filter(clusters == 3), aes(x=percentile, y=salary,
  group=Undergraduate.Major, color=Undergraduate.Major)) +
  geom_point() +
  geom_line() +
  theme(axis.text.x = element_text(size=7)) +
  scale_y_continuous(labels = scales::dollar)
  cluster_III + theme_fivethirtyeight() + labs(color = "Undergraduate Major") +
  theme(axis.title = element_text()) +
  xlab('\nPercentile') +
  ylab('Salary\n')
```



Data Analysis Results:

As k-means cluster analysis needs number of clusters k to be modelled, getting its value is not straight forward and hence optimized value of k needs to be used for analysis. We got our optimal number of clusters i.e. k by using Elbow Method, Silhouette Method and Gap Statistic Method.

The value of k according to each method are as follows:

method	k
Elbow method	3
Silhouette method	2
Gap Statistic method	3

According to majority rule, running K-means with $k = 3$, assigned each major to one of the three clusters. After visualizing each cluster, we obtain the following results:

- **Cluster I** majors may represent the lowest percentiles with limited growth opportunity.
  - Music is the riskiest major with lowest 10th percentile salary.
  - Drama has highest growth potential in the 90th percentile for this cluster.
  - Nursing is an outlier for this cluster with higher safety net in the lowest percentile to the median.

- **Cluster II** majors start off not too low and not too high in the lowest percentile, but majors in this cluster represent greatest differential between the lowest and highest percentiles.
  - Accountants have stable job security.
  - Marketing or Philosophy ultimately result in higher salaries.

- **Cluster III** majors are characterized by financial security and highest growth potential in the 90th percentile as well as best security in the 10th percentile rankings.
  - Physician Assistant is an outlier in this cluster lagging in the highest percentiles.

Conclusion:

Clustering is a broad set of techniques for finding subgroups of observations within a data set. When we cluster observations, we want observations in the same group to be similar and observations in different groups to be dissimilar. Because there isn't a response variable, this is an unsupervised method, which implies that it seeks to find relationships between the n observations without being trained by a response variable.

K-means clustering is the most commonly used unsupervised machine learning algorithm for partitioning a given data set into a set of k groups (i.e. k clusters), where k represents the number of groups pre-specified.

From the analysis of our data, we observed that **Math** and **Philosophy** has highest career percent growth. Since most of the times, starting career salary is primary focus, it is important to consider salary growth potential down the career path.Also remember that wether a major falls into the Liberal Arts, Goldilocks, or Over Achievers cluster, there are many other factors affecting salary increase including the school attended, location, passion or talent for the subject, and of course the actual career(s) pursued

This concludes our analysis, exploring salary projections by college majors via k-means clustering analysis. Dealing with unsupervized data always requires a bit of creativity, such as our usage of three popular methods to determine the optimal number of clusters. We also used visualizations to interpret the patterns revealed by our three clusters.

Reference:

- 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://en.wikipedia.org/wiki/K-means_clustering