FINAL_PROJECT

rama krishna

12/9/2021

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse
1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(cluster)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa
```

```
setwd("C:/Users/krish/OneDrive/Desktop/R_MLCODES/64060_final project")
finaldata <- read.csv("ML_Project-Data.csv")
head(finaldata)
```

```
##     Undergraduate.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
##   Percent.change.from.Starting.to.Mid.Career.Salary
## 1                                              67.6
## 2                                              75.0
## 3                                              68.8
## 4                                              67.1
## 5                                              84.6
## 6                                              81.3
##   Mid.Career.10th.Percentile.Salary Mid.Career.25th.Percentile.Salary
## 1                      $42,200.00                       $56,100.00
## 2                      $64,300.00                       $82,100.00
## 3                      $36,300.00                       $52,100.00
```

```
## 4                        $33,800.00                        $45,500.00
## 5                        $50,600.00                        $62,200.00
## 6                        $28,800.00                        $42,200.00
##   Mid.Career.75th.Percentile.Salary Mid.Career.90th.Percentile.Salary
## 1                      $108,000.00                       $152,000.00
## 2                      $127,000.00                       $161,000.00
## 3                       $96,300.00                       $150,000.00
## 4                       $89,300.00                       $138,000.00
## 5                       $97,000.00                       $136,000.00
## 6                       $87,400.00                       $125,000.00
```

Modify the column names for easy reference

```
colnames(finaldata) <- c('major', 'starting_salary', 'midcareer_salary',
'career_growth_inpercentage' , 'percent10_salary' , 'percent25_salary',
'percent75_salary' , 'percent90_salary')
#View(finaldata)
```

#remove the dollar sign from the data .

```
majors <- finaldata['major']
majors
```

```
##                                 major
## 1                          Accounting
## 2                 Aerospace Engineering
## 3                         Agriculture
## 4                        Anthropology
## 5                        Architecture
## 6                         Art History
## 7                             Biology
## 8                  Business Management
## 9                  Chemical Engineering
## 10                            Chemistry
## 11                   Civil Engineering
## 12                      Communications
## 13                 Computer Engineering
## 14                     Computer Science
## 15                         Construction
## 16                     Criminal Justice
## 17                                Drama
## 18                            Economics
## 19                            Education
## 20               Electrical Engineering
## 21                              English
## 22                                 Film
## 23                              Finance
## 24                             Forestry
## 25                            Geography
## 26                              Geology
```

```
## 27                         Graphic Design
## 28           Health Care Administration
## 29                              History
## 30                Hospitality & Tourism
## 31              Industrial Engineering
## 32          Information Technology (IT)
## 33                        Interior Design
## 34               International Relations
## 35                             Journalism
## 36 Management Information Systems (MIS)
## 37                              Marketing
## 38                                   Math
## 39               Mechanical Engineering
## 40                                 Music
## 41                               Nursing
## 42                             Nutrition
## 43                            Philosophy
## 44                  Physician Assistant
## 45                               Physics
## 46                    Political Science
## 47                            Psychology
## 48                              Religion
## 49                             Sociology
## 50                               Spanish

salary <- finaldata %>%
          select(-major) %>%
          mutate_all(function(x) as.numeric(gsub("[\\$,]", "",x))) %>%
          mutate(career_growth_inpercentage = career_growth_inpercentage/100)
a<- bind_cols(majors, salary)
head(a)

##                    major starting_salary midcareer_salary
## 1             Accounting           46000            77100
## 2 Aerospace Engineering           57700           101000
## 3           Agriculture           42600            71900
## 4         Anthropology            36800            61500
## 5         Architecture            41600            76800
## 6          Art History            35800            64900
##   career_growth_inpercentage percent10_salary percent25_salary
percent75_salary
## 1                      0.676           42200           56100
108000
## 2                      0.750           64300           82100
127000
## 3                      0.688           36300           52100
96300
## 4                      0.671           33800           45500
89300
## 5                      0.846           50600           62200
```

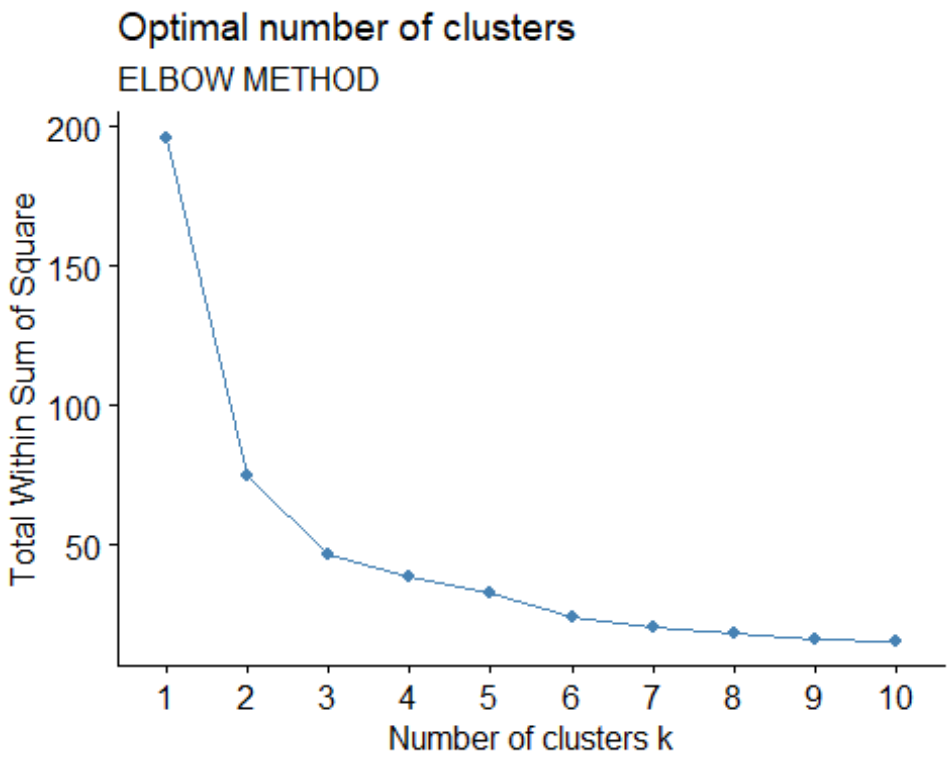
```
97000
## 6                     0.813            28800            42200
87400
##   percent90_salary
## 1           152000
## 2           161000
## 3           150000
## 4           138000
## 5           136000
## 6           125000
```

determining the optimal number of clusters based on starting_salary , midcareer_salary, 10percent_salary, 90percent_salary.

```
kdata <- a %>%
  select(starting_salary, midcareer_salary,percent10_salary,
percent90_salary) %>% scale()
head(kdata)

##      starting_salary midcareer_salary percent10_salary percent90_salary
## [1,]       0.1805388        0.1438303       -0.1006601        0.3315471
## [2,]       1.4304232        1.6293723        1.7408869        0.6546924
## [3,]      -0.1826754       -0.1793839       -0.5922949        0.2597370
## [4,]      -0.8022762       -0.8258122       -0.8006147       -0.1711234
## [5,]      -0.2895031        0.1251833        0.5992944       -0.2429334
## [6,]      -0.9091039       -0.6144799       -1.2172543       -0.6378888
```

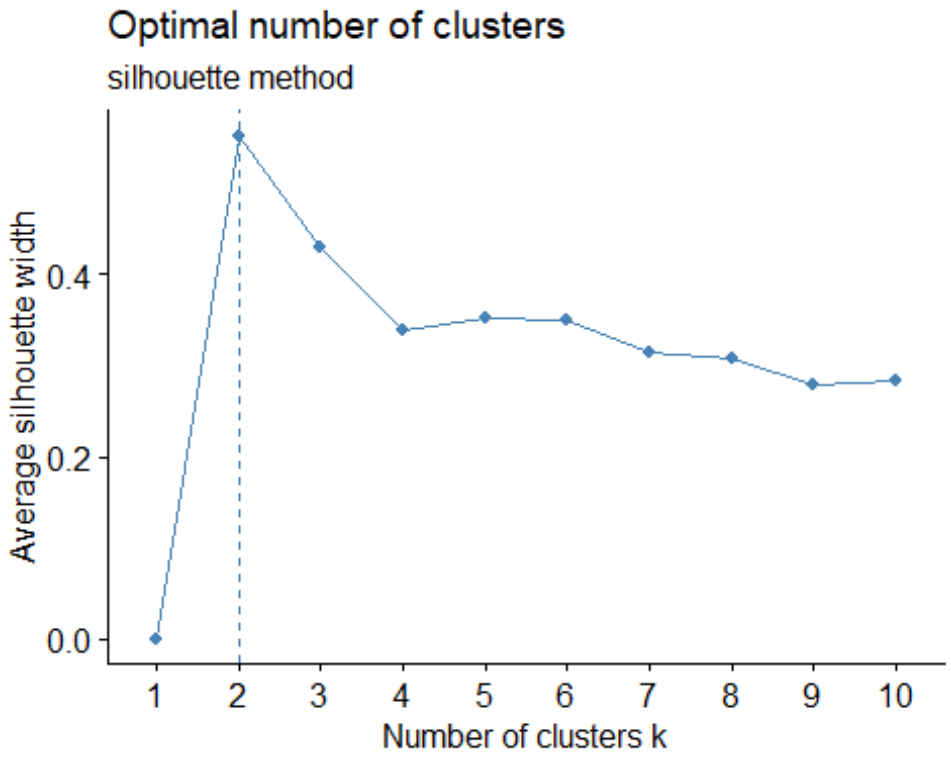
determine the number of clusters by using elbow method

```
fviz_nbclust(kdata,kmeans, method="wss") + labs(subtitle = "ELBOW METHOD")
```



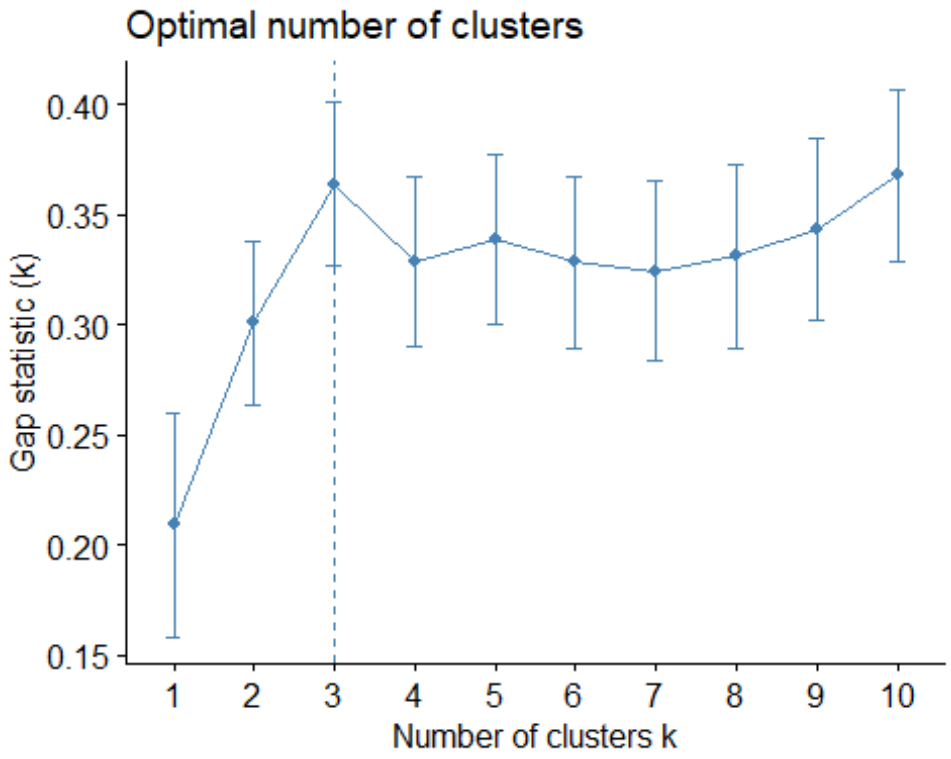
determine the number of clusters by using silhouette method

```
fviz_nbclust(kdata,kmeans, method="silhouette") + labs(subtitle = "silhouette method")
```



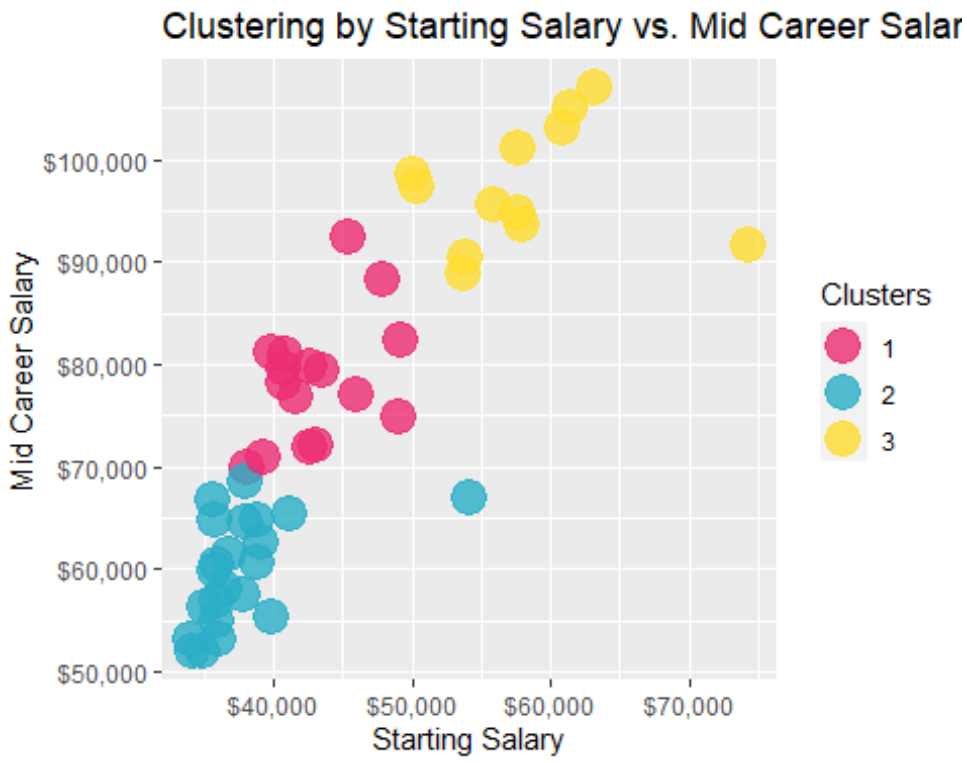
determine the number of clusters of gap statistic method

```
gap <- clusGap(kdata, FUN = kmeans, nstart = 25,
                    K.max = 10, B = 50)
fviz_gap_stat(gap)
```

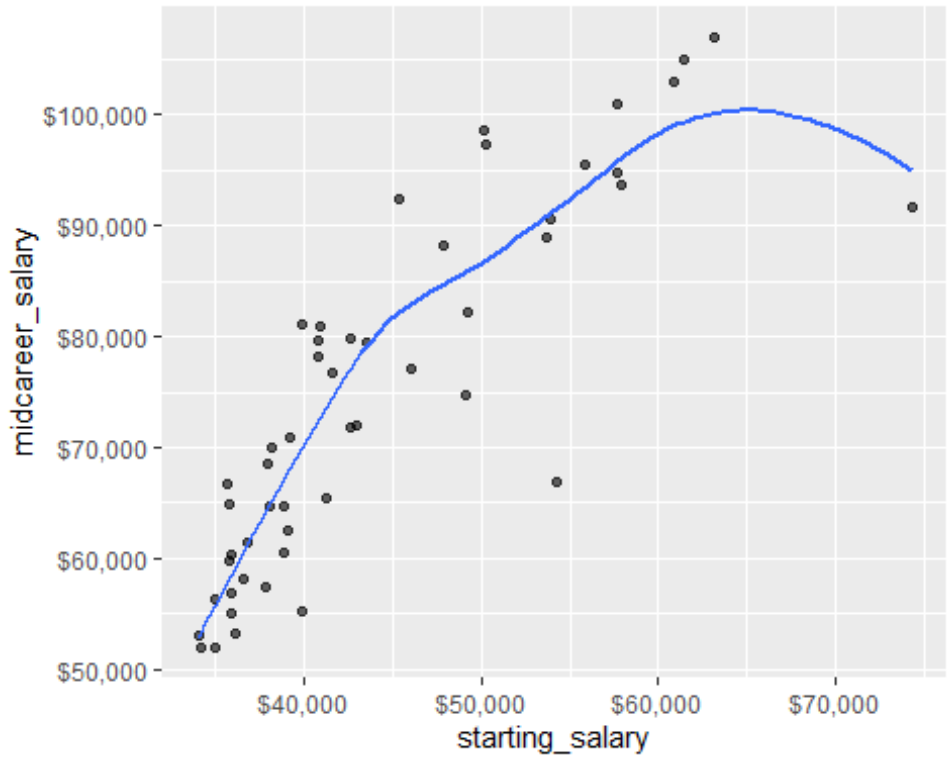


Set k equal to the optimal number of clusters which is 3 since k=3 from elbow method and gap statistics and run kmeans algorithm and visualize the clusters.

```
set.seed(7)
num_clusters <- 3
k_means <- kmeans(kdata , num_clusters , iter.max = 15, nstart = 25)
a$clusters <- k_means[[1]]

salary_increment<- ggplot(a,
                          aes(x=starting_salary,
                              y=midcareer_salary,color=factor(clusters))) +
                   scale_x_continuous(labels=scales::dollar)+
                   scale_y_continuous(labels=scales::dollar)+
                   geom_point(alpha=4/5,size=6)+
                   labs(x="Starting Salary",y="Mid Career Salary",
                        title="Clustering by Starting Salary vs. Mid Career
Salary",
                       colour="Clusters")+
                   scale_colour_manual(values=c("#EC2C73", "#29AEC7",
"#FFDD30"))
# visualize the output
salary_increment
```



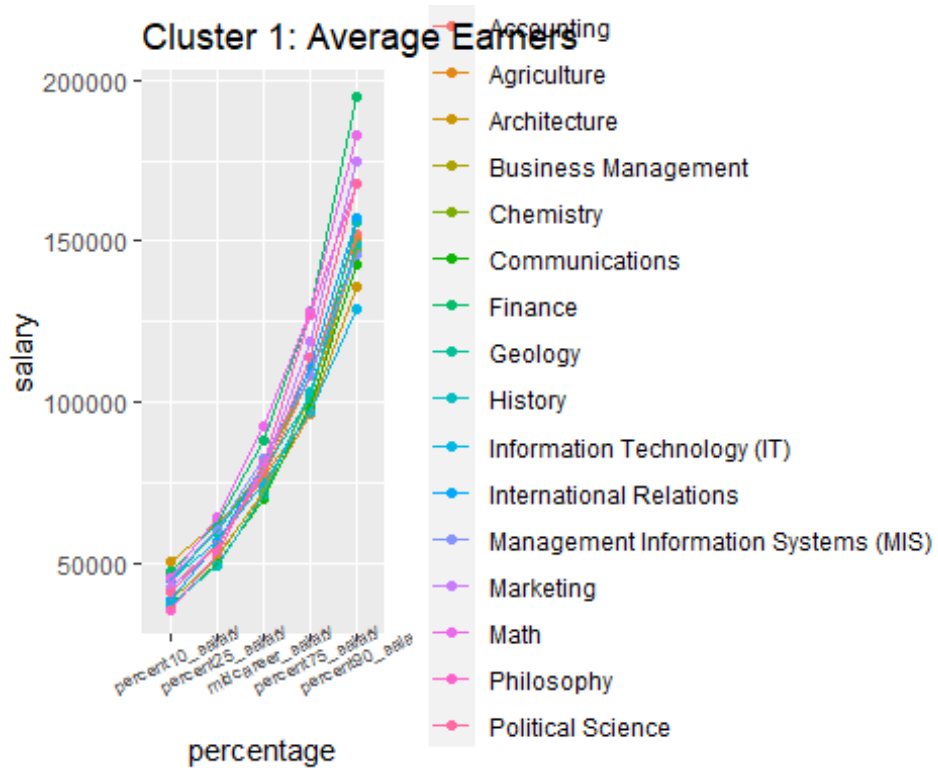

```
ggplot(a, aes(starting_salary,midcareer_salary))+ geom_point(alpha = 0.6)+
  geom_smooth(se = F) +  #loses fit
  scale_x_continuous(labels=scales::dollar)+
                  scale_y_continuous(labels=scales::dollar)

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



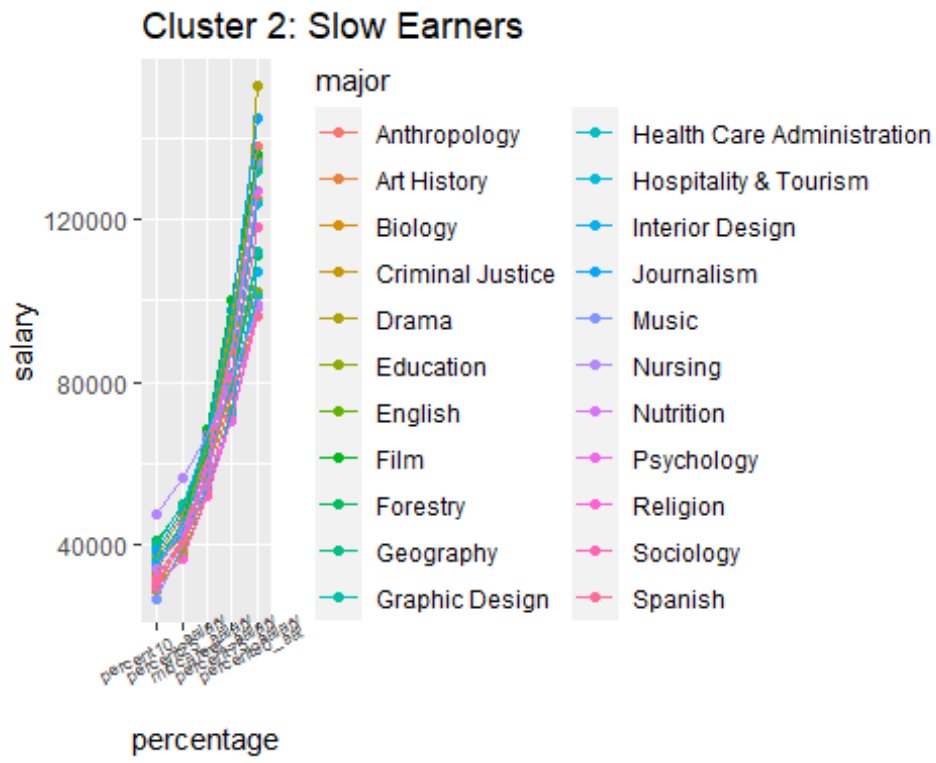

```
paste('correlation coefficient',
      round(with(a, cor(starting_salary,midcareer_salary)), 4))
```

```
## [1] "correlation coefficient 0.8485"
```

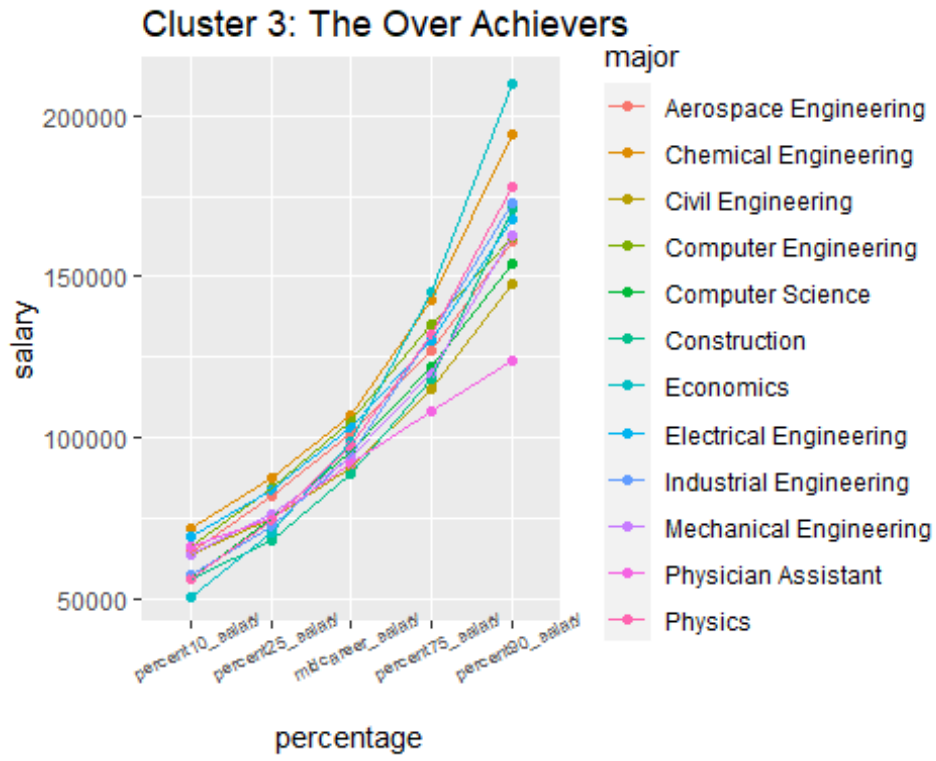
```
salary_variation <- a %>% select(major,percent10_salary, percent25_salary,
midcareer_salary, percent75_salary, percent90_salary, clusters) %>%
gather(key = percentage, value = salary, -c(major, clusters))

salary_variation$percentage = factor(salary_variation$percentage,levels =
c('percent10_salary', 'percent25_salary', 'midcareer_salary',
'percent75_salary', 'percent90_salary'))

cluster_1 <- salary_variation %>% filter(clusters==1) %>%
          ggplot(aes(x=percentage,y=salary,group=major,color=major,
order=salary))+
          geom_point()+
          geom_line()+
          labs(title="Cluster 1: Average Earners")+
          theme(axis.text.x = element_text(size=7,angle=25))
cluster_1
```




```
cluster_2 <- salary_variation %>% filter(clusters==2) %>%
          ggplot(aes(x=percentage,y=salary,group=major,color=major,
order=salary))+
          geom_point()+
          geom_line()+
          labs(title="Cluster 2: Slow Earners")+
          theme(axis.text.x = element_text(size=7,angle=25))
cluster_2
```




```
cluster_3 <- salary_variation %>% filter(clusters==3) %>%
          ggplot(aes(x=percentage,y=salary,group=major,color=major,
order=salary))+
          geom_point()+
          geom_line()+
          labs(title="Cluster 3: The Over Achievers")+
          theme(axis.text.x = element_text(size=7,angle=25))
cluster_3
```



arranging the majors in the descending order of the career percentage growth.

```
a <- a %>% arrange(desc(career_growth_inpercentage))
head(a,8)
```

```
##                     major starting_salary midcareer_salary
## 1                    Math           45400            92400
## 2              Philosophy           39900            81200
## 3 International Relations           40900            80900
## 4               Economics           50100            98600
## 5               Marketing           40800            79600
## 6                 Physics           50300            97300
## 7       Political Science           40800            78200
## 8               Chemistry           42600            79900
##   career_growth_inpercentage percent10_salary percent25_salary
percent75_salary
## 1                      1.035            45200            64200
128000
## 2                      1.035            35500            52800
127000
## 3                      0.978            38200            56000
111000
## 4                      0.968            50600            70600
145000
## 5                      0.951            42100            55600
119000
## 6                      0.934            56000            74200
```

```
132000
## 7                     0.917            41200            55300
114000
## 8                     0.876            45300            60700
108000
##   percent90_salary clusters
## 1           183000        1
## 2           168000        1
## 3           157000        1
## 4           210000        3
## 5           175000        1
## 6           178000        3
## 7           168000        1
## 8           148000        1
```