Untitled

Problem 7.1

```
data = read.csv("data.csv", sep = ",")
summary(data)
```

```
##                Name          pos           Team           Gm
##  Aaron.Ekblad   :  1   D      :290   CGY    : 31   Min.   :  1.00
##  Aaron.Volpatti :  1   C      :160   EDM    : 31   1st Qu.: 28.00
##  Adam.Burish    :  1   L      : 73   N.J    : 29   Median : 65.00
##  Adam.Clendening:  1   R      : 71   CBJ    : 28   Mean   : 55.69
##  Adam.Cracknell :  1   RL     : 70   COL    : 28   3rd Qu.: 81.00
##  Adam.Henrique  :  1   LR     : 61   NYI    : 28   Max.   :108.00
##  (Other)        :830   (Other):111   (Other):661
##       Age            Salary             G               A
##  Min.   :18.00   Min.   : 0.550   Min.   : 0.000   Min.   : 0.00
##  1st Qu.:23.00   1st Qu.: 0.750   1st Qu.: 1.000   1st Qu.: 3.00
##  Median :26.00   Median : 1.000   Median : 5.000   Median :10.00
##  Mean   :26.24   Mean   : 2.217   Mean   : 8.219   Mean   :14.07
##  3rd Qu.:29.00   3rd Qu.: 3.362   3rd Qu.:13.000   3rd Qu.:21.00
##  Max.   :42.00   Max.   :14.000   Max.   :58.000   Max.   :65.00
##
##        P              G60              A60              P60
##  Min.   : 0.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.: 4.00   1st Qu.:0.1000   1st Qu.:0.4200   1st Qu.:0.630
##  Median :15.50   Median :0.3800   Median :0.7300   Median :1.140
##  Mean   :22.28   Mean   :0.4603   Mean   :0.7812   Mean   :1.242
##  3rd Qu.:35.00   3rd Qu.:0.7100   3rd Qu.:1.0800   3rd Qu.:1.780
##  Max.   :95.00   Max.   :5.0000   Max.   :5.5300   Max.   :5.530
##
##       PenD               CF.              PDO              PSh.
##  Min.   :-25.0000   Min.   :20.00   Min.   : 50.00   Min.   :  0.000
##  1st Qu.: -4.0000   1st Qu.:44.44   1st Qu.: 97.77   1st Qu.:  2.840
##  Median :  0.0000   Median :49.72   Median : 99.81   Median :  6.705
##  Mean   : -0.8433   Mean   :49.40   Mean   : 99.12   Mean   :  7.210
##  3rd Qu.:  2.0000   3rd Qu.:54.81   3rd Qu.:101.17   3rd Qu.: 10.530
##  Max.   : 27.0000   Max.   :72.22   Max.   :125.00   Max.   :100.000
##
##     ZSO.Rel            TOI.Gm
##  Min.   :-55.810   Min.   : 4.49
##  1st Qu.: -8.533   1st Qu.:12.23
##  Median :  1.785   Median :15.39
##  Mean   :  1.285   Mean   :15.26
##  3rd Qu.: 12.110   3rd Qu.:18.23
##  Max.   : 58.620   Max.   :28.77
##
```

Function helping to make right chouse about number of K

```
wssplot <- function(data, nc=15, seed=1234){
               wss <- (nrow(data)-1)*sum(apply(data,2,var))
               for (i in 2:nc){
                    set.seed(seed)
                    wss[i] <- sum(kmeans(data, centers=i)$withinss)}
                plot(1:nc, wss, type="b", xlab="Number of Clusters",
                     ylab="Within groups sum of squares")}
```

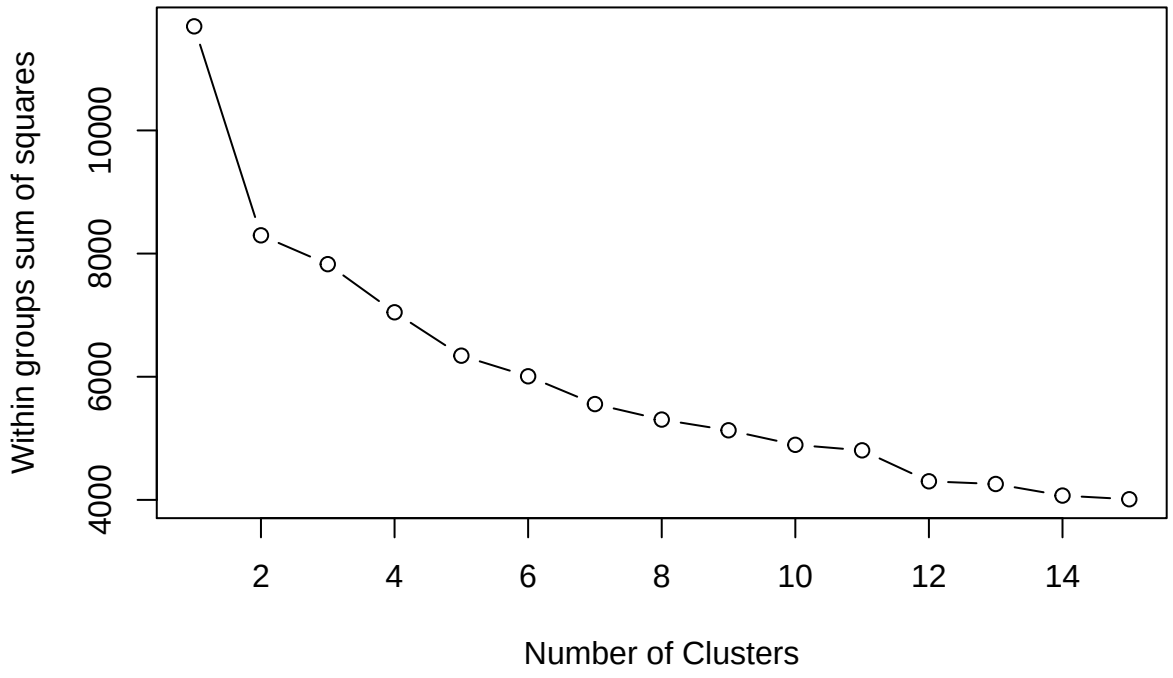
Data that we want to use for clusterization

```
kmeans.data <- data[4:18]
head(kmeans.data)
```

```
##    Gm Age Salary  G  A  P  G60  A60  P60 PenD   CF.    PDO  PSh. ZSO.Rel
## 1  74  28   9.50 28 42 70 1.19 1.79 2.98    3 61.43 101.02 12.56   24.54
## 2  77  29   9.25 23 31 54 0.94 1.27 2.22    7 61.89  97.75  9.43   12.99
## 3  83  29   9.00 43 30 73 1.69 1.18 2.87   -2 57.51 101.12 16.73   17.00
## 4  93  29   8.75 27 63 90 0.86 2.00 2.86  -14 54.74 100.20 11.30    3.43
## 5  81  29   8.75 23 32 55 0.92 1.28 2.20   -5 60.73  97.37  9.83   -5.62
## 6 108  24   8.00 50 40 90 1.44 1.15 2.60   11 57.82 102.05 15.02   24.30
##   TOI.Gm
## 1  19.07
## 2  18.97
## 3  18.36
## 4  20.30
## 5  18.53
## 6  19.25
```

Scaling the data

```
kmeans.data.scaled <- scale(kmeans.data[-1])
wssplot(kmeans.data.scaled)
```



Let's do some clusterization

3 clusters

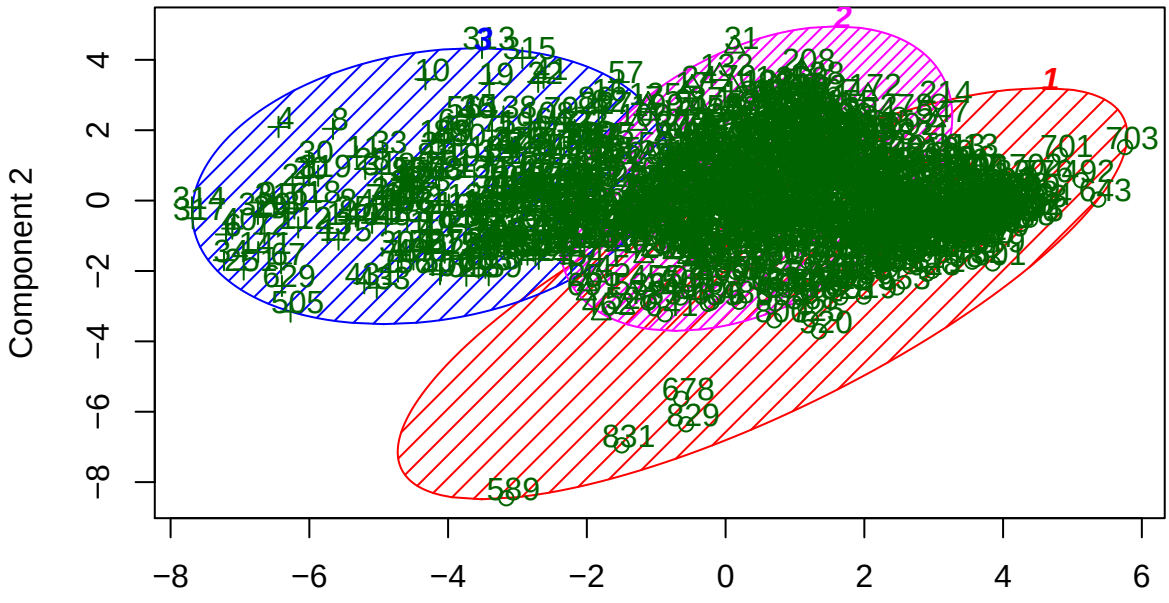
```
# K-Means Cluster Analysis
fit3 <- kmeans(kmeans.data, 3)
# get cluster means
aggregate(kmeans.data,by=list(fit3$cluster),FUN=mean)
```

```
##   Group.1       Gm      Age    Salary         G         A         P
## 1       1 16.96099 24.68440 0.9641135  1.234043  2.205674  3.439716
## 2       2 69.78274 27.20238 2.0880476  6.523810 12.110119 18.633929
## 3       3 84.05505 26.76147 4.0370275 19.866972 32.422018 52.288991
##          G60       A60       P60       PenD      CF.       PDO      PSh.
## 1 0.2882624 0.5518440 0.8401064 -0.6560284 47.93876  97.72191  5.141809
## 2 0.3896726 0.6703869 1.0602976 -2.4464286 46.70348  99.25080  6.923601
## 3 0.7918349 1.2486239 2.0405505  1.3853211 55.44344 100.74326 10.327615
##     ZSO.Rel   TOI.Gm
## 1  2.248298 12.22454
## 2 -6.948720 15.65595
## 3 12.729083 18.57009
```

```
# append cluster assignment
kmeans.data <- data.frame(kmeans.data, fit3$cluster)
```

Cluster Plot against 1st 2 principal components

```
# vary parameters for most readable graph
library(cluster)
clusplot(kmeans.data, fit3$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)
```




```
#Centroid Plot against 1st 2 discriminant functions

#library(fpc)
#plotcluster(kmeans.data, fit$cluster)
```

4 clusters

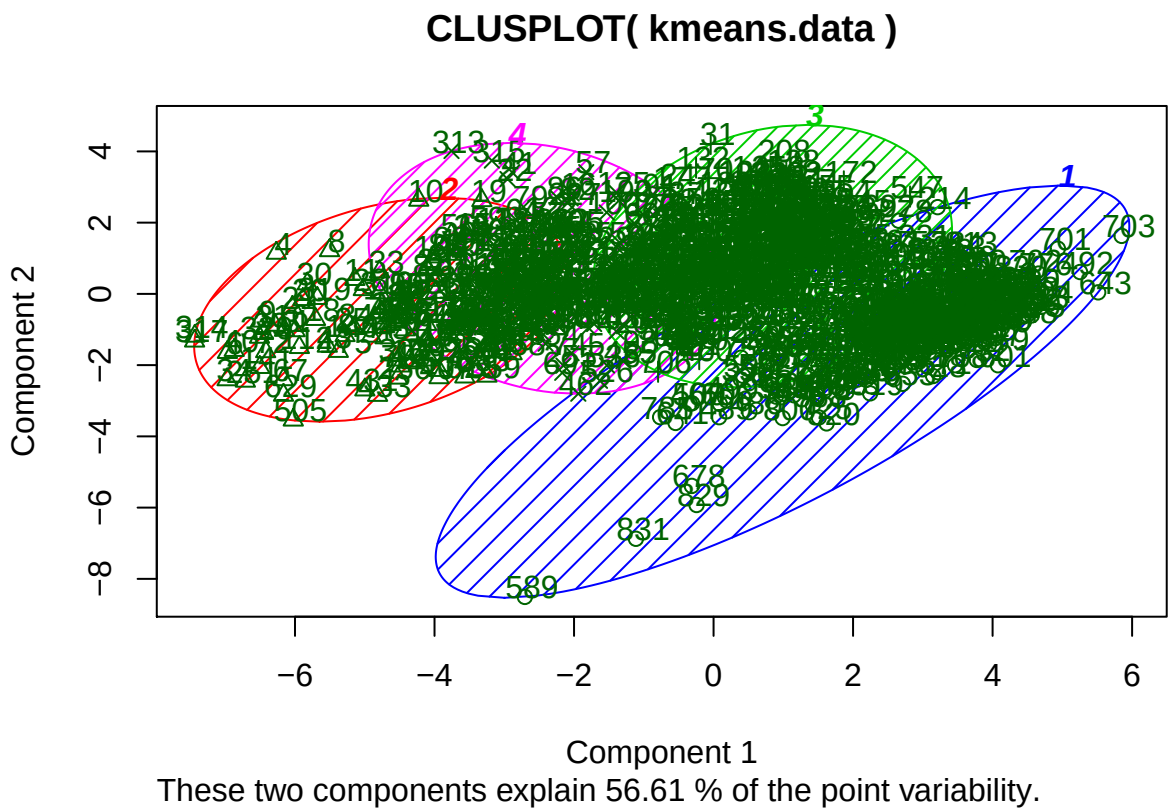
```
# K-Means Cluster Analysis
fit4 <- kmeans(kmeans.data, 4)
# get cluster means
aggregate(kmeans.data,by=list(fit4$cluster),FUN=mean)
```

```
##   Group.1       Gm      Age   Salary         G         A         P
## 1       1 16.15018 24.56044 0.947326  1.175824  2.131868  3.307692
## 2       2 87.42593 26.84259 4.811852 24.620370 39.861111 64.481481
## 3       3 67.60700 27.58755 1.951023  4.941634 10.000000 14.941634
## 4       4 77.41414 26.46970 2.898096 13.237374 21.727273 34.964646
##         G60       A60       P60       PenD      CF.       PDO      PSh.
## 1 0.2865568 0.5535165 0.8400733 -0.6923077 48.12015  97.71839  5.107363
## 2 0.9154630 1.4550926 2.3703704  2.8611111 56.58500 101.32426 11.273426
## 3 0.3185603 0.5859533 0.9047860 -2.9105058 44.98681  98.90074  6.092685
## 4 0.6357071 0.9809091 1.6168687 -0.3888889 52.97061 100.15278  9.343838
##      ZSO.Rel   TOI.Gm fit3.cluster
## 1   2.749414 12.17300     1.000000
## 2  15.622870 19.00231     3.000000
## 3 -10.380506 15.32840     1.964981
## 4   6.586465 17.37939     2.555556
```

```
# append cluster assignment
kmeans.data <- data.frame(kmeans.data, fit4$cluster)
```

Cluster Plot against 1st 2 principal components

```
# vary parameters for most readable graph
clusplot(kmeans.data, fit4$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)
```



7 clusters

```
# K-Means Cluster Analysis
fit7 <- kmeans(kmeans.data, 7)
# get cluster means
aggregate(kmeans.data,by=list(fit7$cluster),FUN=mean)
```

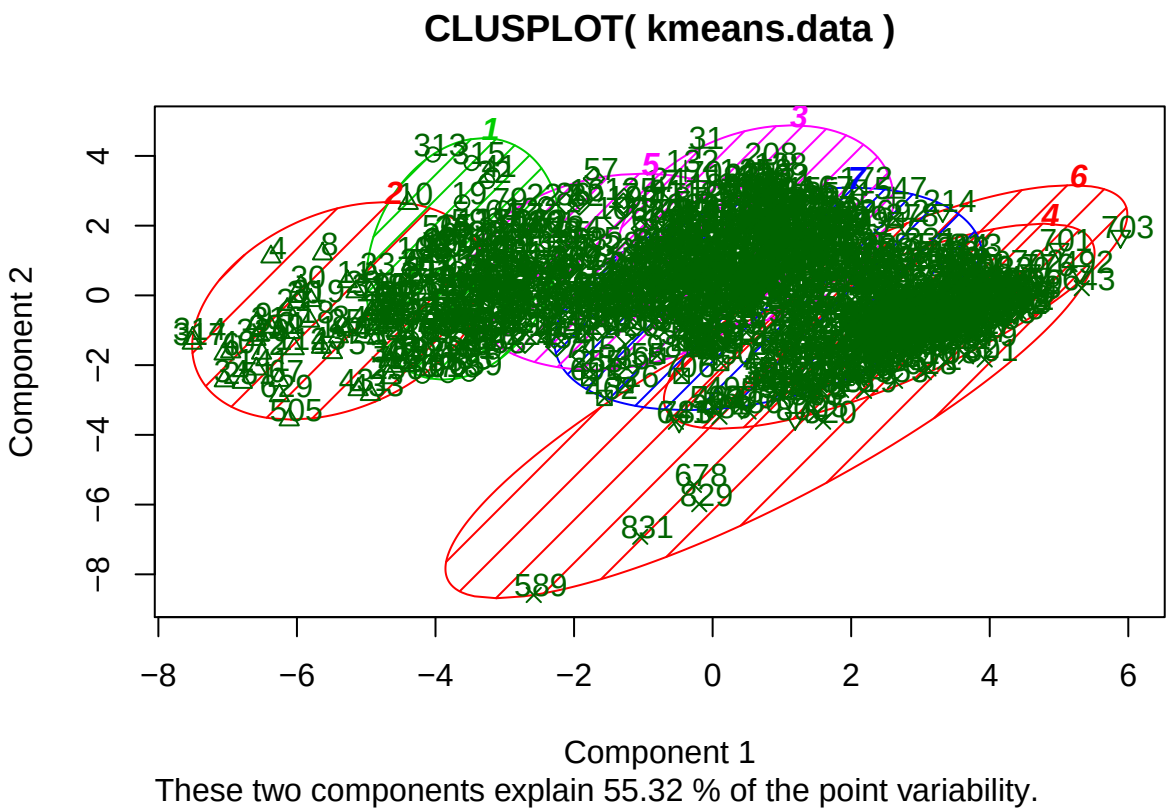
```
##   Group.1       Gm      Age    Salary          G         A         P
## 1       1 83.29508 26.67213 3.6924426 17.5163934 29.918033 47.434426
## 2       2 89.05172 27.03448 5.5573448 28.7413793 44.448276 73.189655
## 3       3 77.09231 28.43846 2.0957692  5.0307692 10.761538 15.792308
## 4       4 10.96552 23.83448 0.8535379  0.7862069  1.179310  1.965517
## 5       5 77.05839 26.19708 2.3741752 11.0583942 18.394161 29.452555
## 6       6 15.54444 25.73333 0.9842333  0.5777778  1.655556  2.233333
## 7       7 49.73377 26.33117 1.7576364  4.7532468  8.389610 13.142857
##         G60       A60       P60       PenD      CF.       PDO      PSh.
## 1 0.7157377 1.1627049 1.8781967 -0.9836066 55.29533 100.69180  9.731639
## 2 1.0337931 1.5789655 2.6132759  5.2586207 57.50759 101.54672 12.263793
## 3 0.2782308 0.5385385 0.8169231 -3.6000000 42.47015  98.74685  5.865846
```

```
## 4 0.3068966 0.4417931 0.7486897 -0.1931034 50.29945  97.48807  5.955655
## 5 0.5616058 0.8817518 1.4434307  0.1240876 50.74401  99.69584  8.726058
## 6 0.1715556 0.6385556 0.8100000 -0.9888889 42.69233  97.05000  2.661222
## 7 0.4188961 0.6967532 1.1161688 -2.0909091 49.39948  99.53299  6.935519
##      ZSO.Rel   TOI.Gm fit3.cluster fit4.cluster
## 1  12.522377 18.81697     3.000000     3.180328
## 2  17.352931 19.20241     3.000000     2.000000
## 3 -18.630923 15.66492     2.000000     3.007692
## 4  12.098069 11.59269     1.000000     1.000000
## 5   1.511971 16.50547     2.277372     3.817518
## 6 -14.980111 12.29289     1.011111     1.111111
## 7   2.265325 14.68571     1.688312     2.525974
```

```
# append cluster assignment
kmeans.data <- data.frame(kmeans.data, fit7$cluster)
```

Cluster Plot against 1st 2 principal components

```
# vary parameters for most readable graph
clusplot(kmeans.data, fit7$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)
```



comparing 2 cluster solutions 3-means and 4-means

```
library(fpc)
d <- dist(kmeans.data)
cluster.stats(d, fit3$cluster, fit4$cluster)
```

```
## $n
```

```
## [1] 836
##
## $cluster.number
## [1] 3
##
## $cluster.size
## [1] 282 336 218
##
## $min.cluster.size
## [1] 218
##
## $noisen
## [1] 0
##
## $diameter
## [1] 123.32413  99.51566  93.59876
##
## $average.distance
## [1] 35.32876 34.00133 37.26233
##
## $median.distance
## [1] 32.47629 32.73185 35.24955
##
## $separation
## [1] 8.635514 8.579261 8.579261
##
## $average.toother
## [1] 77.96508 62.46557 76.41253
##
## $separation.matrix
##           [,1]     [,2]      [,3]
## [1,]  0.000000 8.635514 30.041718
## [2,]  8.635514 0.000000  8.579261
## [3,] 30.041718 8.579261  0.000000
##
## $ave.between.matrix
##          [,1]     [,2]     [,3]
## [1,]  0.00000 65.32813 97.44222
## [2,] 65.32813  0.00000 58.76262
## [3,] 97.44222 58.76262  0.00000
##
## $average.between
## [1] 71.83572
##
## $average.within
## [1] 35.08642
##
## $n.between
## [1] 229476
##
## $n.within
## [1] 119554
##
## $max.diameter
```

```
## [1] 123.3241
##
## $min.separation
## [1] 8.579261
##
## $within.cluster.ss
## [1] 601179.3
##
## $clus.avg.silwidths
##         1         2         3
## 0.4374800 0.3488071 0.3340740
##
## $avg.silwidth
## [1] 0.3748764
##
## $g2
## NULL
##
## $g3
## NULL
##
## $pearsongamma
## [1] 0.611356
##
## $dunn
## [1] 0.06956677
##
## $dunn2
## [1] 1.576998
##
## $entropy
## [1] 1.083429
##
## $wb.ratio
## [1] 0.4884258
##
## $ch
## [1] 834.2156
##
## $cwidegap
## [1] 52.04894 28.18832 26.44934
##
## $widestgap
## [1] 52.04894
##
## $sindex
## [1] 11.77328
##
## $corrected.rand
## [1] 0.6785897
##
## $vi
## [1] 0.668891
```

comparing 2 cluster solutions 3-means and 7-means

```
#cluster.stats(d, fit3$cluster, fit7$cluster)
```

comparing 2 cluster solutions 4-means and 7-means

```
#cluster.stats(d, fit4$cluster, fit7$cluster)
```