K-Means_Clustering.R

Siri

2019-10-30

```
#K means Clustering

library(data.table)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Loading required package: RSQLite
```

```
library(MVA)
```

```
## Loading required package: HSAUR2
```

```
## Loading required package: tools
```

```
data<-read.csv("C:/Users/Siri/Downloads/dataset_final.csv", stringsAsFactors=FALSE)
View(data)

#Australian Tournament
AustralianOpen<-subset(data,data$Tournament=="Australian Open")
View(AustralianOpen)

AustralianOpen_Finalists<-subset(AustralianOpen,AustralianOpen$Round=='The Final')
View(AustralianOpen_Finalists)
AustralianOpen_Finalists<-subset(AustralianOpen_Finalists,select = c("PlayerName","Year"))
AustralianOpen_Finalists_allstats<-merge(x=AustralianOpen,y=AustralianOpen_Finalists, by=c("PlayerName"
View(AustralianOpen_Finalists_allstats)
setDT(AustralianOpen_Finalists_allstats)

#Win percentage calculation for finalists

total_matches=AustralianOpen_Finalists_allstats%>%
  group_by(PlayerName,Year)%>%
```

```
  summarize(total_matchs=n())
View(total_matches)


total_matches_Won=AustralianOpen_Finalists_allstats%>%
  filter(Winner=='TRUE')%>%
  group_by(PlayerName,Year)%>%
  summarise(total_matches_won=n())
View(total_matches_Won)

setDT(total_matches)
setDT(total_matches_Won)
total_matches$winpercentage=(winpercentage=(total_matches_Won$total_matches_won)/(total_matches$total_ma


#merging the win percentage in finalists data
AustralianOpen_Finalists_allstats=merge(x=total_matches,y=AustralianOpen_Finalists_allstats, by=c("Playe
View(AustralianOpen_Finalists_allstats)

#Factor Analysis  Analysis

head(AustralianOpen_Finalists_allstats)
```

```
##      PlayerName Year total_matchs winpercentage      MatchID         Round
## 1: Andre Agassi 2000            7             1 m_2000_A_114     4th Round
## 2: Andre Agassi 2000            7             1 m_2000_A_122 Quarterfinals
## 3: Andre Agassi 2000            7             1  m_2000_A_73     2nd Round
## 4: Andre Agassi 2000            7             1 m_2000_A_124    Semifinals
## 5: Andre Agassi 2000            7             1  m_2000_A_44     1st Round
## 6: Andre Agassi 2000            7             1  m_2000_A_97     3rd Round
##    AvgMinsPerGame AvgSecsPerPoint AvgMinsPerSet      Tournament
## 1:           3.84            37.9          41.3 Australian Open
## 2:           3.32            35.1          31.0 Australian Open
## 3:           3.44            37.2          31.0 Australian Open
## 4:           3.50            34.5          35.0 Australian Open
## 5:           3.48            37.3          29.0 Australian Open
## 6:           3.39            37.0          31.7 Australian Open
##    TotalMatchMins Points Age Rank Winner TotalSets avgOdds maxOdds
## 1:            165      0  30    1   TRUE         3       0       0
## 2:             93      0  30    1   TRUE         3       0       0
## 3:             93      0  30    1   TRUE         3       0       0
## 4:            175      0  30    1   TRUE         3       0       0
## 5:             87      0  30    1   TRUE         3       0       0
## 6:             95      0  30    1   TRUE         3       0       0
##    SP_Percent RP_Percent BP_Win_Percentage Aces firstServeReturnsWon
## 1:  0.7089552  0.2910448         0.7777778    8                   11
## 2:  0.5744681  0.4255319         0.5000000    6                   13
## 3:  0.5806452  0.4193548         0.0000000    8                   12
## 4:  0.6903226  0.3096774         0.8888889   13                   19
## 5:  0.5505618  0.4494382         1.0000000    6                   18
## 6:  0.5760870  0.4239130         0.0000000    8                   14
##    SecondServeReturnsWon FirstServesIn DoubleFaults FirstServePercentage
## 1:                    28            96            4            0.6906475
## 2:                    27            45            1            0.6617647
```

```
## 3:                    27            50            1            0.6578947
## 4:                    29           101            3            0.6824324
## 5:                    22            40            1            0.6557377
## 6:                    25            35            3            0.5303030
```

```
str(AustralianOpen_Finalists_allstats)
```

```
## Classes 'data.table' and 'data.frame':   277 obs. of  27 variables:
##  $ PlayerName           : chr  "Andre Agassi" "Andre Agassi" "Andre Agassi" "Andre Agassi" ...
##  $ Year                 : int  2000 2000 2000 2000 2000 2000 2000 2001 2001 2001 ...
##  $ total_matchs         : int  7 7 7 7 7 7 7 7 7 7 ...
##  $ winpercentage        : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ MatchID              : chr  "m_2000_A_114" "m_2000_A_122" "m_2000_A_73" "m_2000_A_124" ...
##  $ Round                : chr  "4th Round" "Quarterfinals" "2nd Round" "Semifinals" ...
##  $ AvgMinsPerGame       : num  3.84 3.32 3.44 3.5 3.48 3.39 3.86 3.81 4 3.75 ...
##  $ AvgSecsPerPoint      : num  37.9 35.1 37.2 34.5 37.3 37 35 38.3 32.6 33.3 ...
##  $ AvgMinsPerSet        : num  41.3 31 31 35 29 31.7 34.8 39.3 68 33.8 ...
##  $ Tournament           : chr  "Australian Open" "Australian Open" "Australian Open" "Australian Ope
##  $ TotalMatchMins       : int  165 93 93 175 87 95 139 118 68 135 ...
##  $ Points               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Age                  : int  30 30 30 30 30 30 30 31 31 31 ...
##  $ Rank                 : int  1 1 1 1 1 1 1 6 6 6 ...
##  $ Winner               : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ TotalSets            : int  3 3 3 3 3 3 3 3 1 3 ...
##  $ avgOdds              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ maxOdds              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SP_Percent           : num  0.709 0.574 0.581 0.69 0.551 ...
##  $ RP_Percent           : num  0.291 0.426 0.419 0.31 0.449 ...
##  $ BP_Win_Percentage    : num  0.778 0.5 0 0.889 1 ...
##  $ Aces                 : int  8 6 8 13 6 8 9 6 8 5 ...
##  $ firstServeReturnsWon : int  11 13 12 19 18 14 23 30 19 33 ...
##  $ SecondServeReturnsWon: int  28 27 27 29 22 25 27 18 16 32 ...
##  $ FirstServesIn        : int  96 45 50 101 40 35 77 55 40 77 ...
##  $ DoubleFaults         : int  4 1 1 3 1 3 5 0 2 2 ...
##  $ FirstServePercentage : num  0.691 0.662 0.658 0.682 0.656 ...
##  - attr(*, ".internal.selfref")=<externalptr>
##  - attr(*, "sorted")= chr  "PlayerName" "Year"
```

```
summary(AustralianOpen_Finalists_allstats)
```

```
##   PlayerName             Year       total_matchs   winpercentage
##  Length:277         Min.   :2000   Min.   :6.000   Min.   :0.8333
##  Class :character   1st Qu.:2005   1st Qu.:7.000   1st Qu.:0.8571
##  Mode  :character   Median :2009   Median :7.000   Median :0.8571
##                     Mean   :2009   Mean   :6.935   Mean   :0.9278
##                     3rd Qu.:2014   3rd Qu.:7.000   3rd Qu.:1.0000
##                     Max.   :2019   Max.   :7.000   Max.   :1.0000
##
##    MatchID             Round           AvgMinsPerGame  AvgSecsPerPoint
##  Length:277         Length:277         Min.   :2.930   Min.   :30.20
##  Class :character   Class :character   1st Qu.:3.860   1st Qu.:37.60
##  Mode  :character   Mode  :character   Median :4.280   Median :40.70
##                                        Mean   :4.361   Mean   :41.25
##                                        3rd Qu.:4.700   3rd Qu.:44.30
##                                        Max.   :9.030   Max.   :75.00
```

```
##
##  AvgMinsPerSet    Tournament        TotalMatchMins      Points
##  Min.   :24.00   Length:277         Min.   : 28.0   Min.   :    0
##  1st Qu.:34.77   Class :character   1st Qu.:104.0   1st Qu.:    0
##  Median :40.65   Mode  :character   Median :135.0   Median : 4675
##  Mean   :41.44                      Mean   :144.3   Mean   : 5361
##  3rd Qu.:47.30                      3rd Qu.:174.0   3rd Qu.: 9595
##  Max.   :93.30                      Max.   :353.0   Max.   :16790
##  NA's   :1
##       Age             Rank          Winner          TotalSets
##  Min.   :21.0   Min.   : 1.000   Mode :logical   Min.   :0.000
##  1st Qu.:24.0   1st Qu.: 1.000   FALSE:20        1st Qu.:3.000
##  Median :26.0   Median : 3.000   TRUE :257       Median :3.000
##  Mean   :26.8   Mean   : 9.289                   Mean   :2.765
##  3rd Qu.:29.0   3rd Qu.: 8.000                   3rd Qu.:3.000
##  Max.   :36.0   Max.   :86.000                   Max.   :3.000
##
##     avgOdds          maxOdds         SP_Percent       RP_Percent
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.4000   Min.   :0.1828
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.5556   1st Qu.:0.3644
##  Median :0.0000   Median :0.0000   Median :0.5984   Median :0.4016
##  Mean   :0.6334   Mean   :0.6652   Mean   :0.5954   Mean   :0.4046
##  3rd Qu.:1.0700   3rd Qu.:1.1100   3rd Qu.:0.6356   3rd Qu.:0.4444
##  Max.   :7.5400   Max.   :9.9500   Max.   :0.8172   Max.   :0.6000
##
##  BP_Win_Percentage      Aces        firstServeReturnsWon
##  Min.   :0.0000    Min.   : 1.000   Min.   : 4.00
##  1st Qu.:0.4286    1st Qu.: 6.000   1st Qu.:17.00
##  Median :0.6471    Median : 9.000   Median :21.00
##  Mean   :0.5779    Mean   : 9.729   Mean   :22.15
##  3rd Qu.:0.8000    3rd Qu.:13.000   3rd Qu.:26.00
##  Max.   :1.0000    Max.   :33.000   Max.   :47.00
##
##  SecondServeReturnsWon FirstServesIn     DoubleFaults
##  Min.   : 3.00         Min.   : 12.00   Min.   :0.000
##  1st Qu.:18.00         1st Qu.: 47.00   1st Qu.:1.000
##  Median :22.00         Median : 57.00   Median :2.000
##  Mean   :23.31         Mean   : 62.08   Mean   :2.412
##  3rd Qu.:29.00         3rd Qu.: 77.00   3rd Qu.:4.000
##  Max.   :45.00         Max.   :135.00   Max.   :9.000
##
##  FirstServePercentage
##  Min.   :0.3692
##  1st Qu.:0.5806
##  Median :0.6316
##  Mean   :0.6267
##  3rd Qu.:0.6754
##  Max.   :0.8088
##
```

```
AustralianOpen_Finalists_allstats_Numeric<-subset(AustralianOpen_Finalists_allstats,select = c("Age","Ra
View(AustralianOpen_Finalists_allstats_Numeric)

#K-Means Clustering
```

```r
AustralianOpen_Finalists_allstats_Numeric_scale<-scale(AustralianOpen_Finalists_allstats_Numeric)
# K-means, k=2, 3, 4, 5, 6
# Centers (k's) are numbers thus, 10 random sets are chosen

(kmeans2<-kmeans(AustralianOpen_Finalists_allstats_Numeric,2,nstart = 10))
```

```
## K-means clustering with 2 clusters of sizes 186, 91
##
## Cluster means:
##        Age     Rank   avgOdds SP_Percent RP_Percent BP_Win_Percentage
## 1 26.85484  7.44086 0.5322581  0.5776930  0.4223070         0.5339920
## 2 26.69231 13.06593 0.8400000  0.6314527  0.3685473         0.6676714
##        Aces firstServeReturnsWon SecondServeReturnsWon FirstServesIn
## 1  8.419355             20.26344              20.93548      49.53226
## 2 12.406593             26.00000              28.16484      87.72527
##   DoubleFaults FirstServePercentage
## 1     1.887097            0.6237630
## 2     3.483516            0.6326372
##
## Clustering vector:
##   [1] 2 1 1 2 1 1 2 1 1 2 1 1 1 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 2 1 1 1 2
##  [36] 1 2 1 2 1 2 1 1 2 1 2 1 2 1 1 1 1 1 2 1 2 1 2 1 1 1 2 1 2 2 1 2 1 2 2
##  [71] 1 2 1 2 2 1 1 1 2 2 1 1 2 1 2 1 1 1 2 2 1 1 2 2 2 2 2 2 1 1 1 1 2 1 2
## [106] 2 1 2 1 2 1 1 2 2 1 1 1 2 2 1 1 1 1 2 1 1 1 2 1 1 1 2 1 1 2 2 1 1 1 1
## [141] 1 1 2 1 2 1 1 1 1 1 2 2 1 1 1 1 1 2 2 2 1 1 2 1 1 1 1 1 1 2 1 1 1 1 2 1
## [176] 1 1 2 1 2 2 1 1 1 2 2 1 2 2 2 1 2 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1 2 1
## [211] 1 1 1 1 1 2 2 1 1 2 1 2 1 1 1 1 1 1 1 2 2 1 1 1 2 1 2 1 2 1 1 1 1 1 2
## [246] 1 2 1 2 1 1 1 1 1 1 2 1 1 1 2 1 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1
##
## Within cluster sum of squares by cluster:
## [1] 75965.01 84238.40
##  (between_SS / total_SS =  37.8 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

```r
# Computing the percentage of variation accounted for Two clusters
perc.var.2 <- round(100*(1 - kmeans2$betweenss/kmeans2$totss),1)
names(perc.var.2) <- "Perc. 2 clus"
perc.var.2
```

```
## Perc. 2 clus
##         62.2
```

```r
# Computing the percentage of variation accounted for three clusters
(kmeans3<-kmeans(AustralianOpen_Finalists_allstats_Numeric,3,nstart = 10))
```

```
## K-means clustering with 3 clusters of sizes 83, 170, 24
##
## Cluster means:
##        Age      Rank   avgOdds SP_Percent RP_Percent BP_Win_Percentage
## 1 27.18072  5.361446 0.9939759  0.6313981  0.3686019         0.6731333
```

```
## 2 27.10000  4.564706 0.5467059  0.5759143  0.4240857         0.5342501
## 3 23.37500 56.333333 0.0000000  0.6084010  0.3915990         0.5578347
##        Aces firstServeReturnsWon SecondServeReturnsWon FirstServesIn
## 1 11.963855             25.84337              27.93976      88.26506
## 2  8.311765             20.10588              20.88235      48.97059
## 3 12.041667             23.83333              24.50000      64.37500
##   DoubleFaults FirstServePercentage
## 1     3.566265            0.6419375
## 2     1.776471            0.6286210
## 3     2.916667            0.5601471
##
## Clustering vector:
##   [1] 1 2 2 1 2 2 1 2 2 1 2 2 2 1 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 1
##  [36] 2 1 2 1 2 1 2 2 1 2 1 2 1 2 2 2 2 2 1 1 1 2 1 2 2 2 1 2 1 1 2 1 2 1 1
##  [71] 3 3 3 3 1 3 3 2 1 1 2 2 1 2 1 2 2 2 1 1 2 3 3 3 3 3 3 2 2 2 2 1 2 1
## [106] 3 3 3 3 3 3 3 1 1 2 2 1 1 1 2 2 2 2 1 2 2 2 1 2 2 2 1 2 2 1 1 2 2 2 2
## [141] 2 2 1 2 1 2 2 2 2 1 1 2 2 2 2 2 1 1 1 2 2 1 2 2 2 2 2 2 1 2 2 2 2 1 2
## [176] 2 2 1 2 1 1 2 2 2 1 1 2 1 1 1 2 1 2 1 1 2 2 2 2 2 2 2 2 3 3 3 3 2 1 2
## [211] 1 2 2 2 2 1 1 2 2 1 2 1 2 2 2 2 2 2 2 1 1 2 2 2 1 2 1 2 1 2 2 2 2 2 1
## [246] 2 1 2 1 2 2 2 2 2 2 1 2 2 2 1 2 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##
## Within cluster sum of squares by cluster:
## [1] 39824.09 48446.93 18770.48
##  (between_SS / total_SS =  58.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

```
perc.var.3 <- round(100*(1 - kmeans3$betweenss/kmeans3$totss),1)
names(perc.var.3) <- "Perc. 3 clus"
perc.var.3
```

```
## Perc. 3 clus
##         41.6
```

```
# Computing the percentage of variation accounted for three clusters
(kmeans4<-kmeans(AustralianOpen_Finalists_allstats_Numeric,4,nstart = 10))
```

```
## K-means clustering with 4 clusters of sizes 25, 32, 127, 93
##
## Cluster means:
##        Age      Rank   avgOdds SP_Percent RP_Percent BP_Win_Percentage
## 1 23.36000 55.600000 0.0000000  0.6125895  0.3874105         0.5621880
## 2 27.43750  4.843750 1.1346875  0.6251454  0.3748546         0.6459919
## 3 27.15748  4.070866 0.5826772  0.5692632  0.4307368         0.4953919
## 4 27.02151  5.494624 0.7003226  0.6160998  0.3839002         0.6713913
##        Aces firstServeReturnsWon SecondServeReturnsWon FirstServesIn
## 1 12.160000             23.48000              24.32000      65.68000
## 2 11.593750             30.12500              32.46875     104.00000
## 3  7.440945             19.25984              19.52756      44.45669
## 4 11.559140             22.98925              25.05376      70.75269
##   DoubleFaults FirstServePercentage
```

```
## 1     2.880000            0.5662707
## 2     3.593750            0.6578367
## 3     1.519685            0.6301298
## 4     3.096774            0.6274827
##
## Clustering vector:
##   [1] 2 3 3 2 3 3 4 3 3 4 3 4 3 2 3 3 3 2 3 3 3 3 4 3 3 3 3 4 3 3 2 3 3 3 4
##  [36] 3 4 3 4 3 2 3 3 4 4 2 3 4 4 3 4 3 4 4 4 2 3 2 3 4 4 2 3 4 4 3 4 3 4 2
##  [71] 1 1 1 1 1 1 1 4 4 4 4 3 4 4 4 3 3 3 4 4 3 1 1 1 1 1 1 1 4 3 3 4 2 3 2
## [106] 1 1 1 1 1 1 1 2 4 3 3 4 4 4 3 4 4 3 4 4 3 3 4 3 3 3 4 4 3 2 2 3 3 4 3
## [141] 3 3 2 4 4 3 3 3 3 2 4 4 4 3 3 3 2 4 4 3 4 4 3 3 3 3 3 4 2 3 3 3 3 2 3
## [176] 3 3 2 3 2 2 3 3 3 4 4 3 2 4 2 3 4 4 2 2 3 3 3 3 4 3 3 3 1 1 1 1 3 4 3
## [211] 4 3 3 3 3 4 4 3 3 4 3 4 3 3 4 4 3 3 3 2 4 3 3 3 2 4 4 3 4 4 3 3 3 4 4
## [246] 4 4 3 2 3 3 3 3 4 3 4 4 4 3 4 3 4 2 4 4 4 4 4 3 4 4 3 3 4 3 4 3
##
## Within cluster sum of squares by cluster:
## [1] 20218.71 11938.03 29348.46 24266.53
##  (between_SS / total_SS =  66.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

```
perc.var.4 <- round(100*(1 - kmeans4$betweenss/kmeans4$totss),1)
names(perc.var.4) <- "Perc. 4 clus"
perc.var.4
```

```
## Perc. 4 clus
##         33.3
```

```
## Computing the percentage of variation accounted for three clusters
(kmeans5<-kmeans(AustralianOpen_Finalists_allstats_Numeric,5,nstart = 10))
```

```
## K-means clustering with 5 clusters of sizes 34, 17, 8, 92, 126
##
## Cluster means:
##        Age      Rank   avgOdds SP_Percent RP_Percent BP_Win_Percentage
## 1 27.17647  6.794118 1.0679412  0.6276807  0.3723193         0.6533355
## 2 23.70588 42.941176 0.0000000  0.5810561  0.4189439         0.4482886
## 3 23.62500 82.000000 0.0000000  0.6536542  0.3463458         0.6595927
## 4 27.08696  5.521739 0.7079348  0.6164983  0.3835017         0.6690272
## 5 27.11111  3.555556 0.5873016  0.5694201  0.4305799         0.5033258
##        Aces firstServeReturnsWon SecondServeReturnsWon FirstServesIn
## 1 11.911765             29.61765              32.02941     103.32353
## 2  8.058824             22.58824              22.88235      53.23529
## 3 17.625000             24.00000              24.87500      77.62500
## 4 11.608696             22.98913              25.08696      70.89130
## 5  7.492063             19.34127              19.61905      44.72222
##   DoubleFaults FirstServePercentage
## 1     3.500000            0.6594331
## 2     3.176471            0.5572020
## 3     2.375000            0.5503397
## 4     3.097826            0.6275243
```

```
## 5     1.515873            0.6314428
##
## Clustering vector:
##   [1] 1 5 5 1 5 5 4 5 5 4 5 4 5 1 5 5 5 1 5 5 5 5 4 5 5 5 5 4 5 5 1 5 5 5 4
##  [36] 5 4 5 4 5 1 5 5 4 4 1 5 4 4 5 4 5 4 4 4 1 5 1 5 4 4 1 5 4 4 5 4 5 4 1
##  [71] 2 2 2 1 1 2 2 4 4 4 4 5 4 4 4 5 5 5 4 4 5 3 3 3 3 3 3 3 4 5 5 4 1 5 1
## [106] 3 2 2 2 2 2 2 1 4 5 5 4 4 4 5 4 4 5 4 5 5 5 4 5 5 5 4 4 5 1 1 5 5 4 5
## [141] 5 5 1 4 4 5 5 5 5 1 4 4 4 5 5 5 1 4 4 5 4 4 5 5 5 5 5 4 1 5 5 5 5 1 5
## [176] 5 5 1 5 1 1 5 5 5 4 4 5 1 4 1 5 4 4 1 1 5 5 5 5 4 5 5 2 2 2 2 2 2 4 5
## [211] 4 5 5 5 5 4 4 5 5 4 5 4 5 5 4 4 5 5 5 1 4 5 5 5 1 4 4 5 4 4 5 5 5 4 4
## [246] 4 4 5 1 5 5 5 5 4 5 4 4 4 5 4 5 4 1 4 4 4 4 4 5 4 4 5 5 4 5 4 5
##
## Within cluster sum of squares by cluster:
## [1] 14783.697  5952.405  2748.490 24028.195 26941.983
##  (between_SS / total_SS =  71.1 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

```
perc.var.5 <- round(100*(1 - kmeans5$betweenss/kmeans5$totss),1)
names(perc.var.5) <- "Perc. 5 clus"
perc.var.5
```

```
## Perc. 5 clus
##         28.9
```

```
## Computing the percentage of variation accounted for three clusters
(kmeans6<-kmeans(AustralianOpen_Finalists_allstats_Numeric,6,nstart = 10))
```

```
## K-means clustering with 6 clusters of sizes 65, 20, 17, 8, 103, 64
##
## Cluster means:
##        Age      Rank   avgOdds SP_Percent RP_Percent BP_Win_Percentage
## 1 26.66154  3.830769 0.5084615  0.5492819  0.4507181         0.3494872
## 2 27.40000  4.600000 0.9935000  0.6323234  0.3676766         0.6780857
## 3 23.70588 42.941176 0.0000000  0.5810561  0.4189439         0.4482886
## 4 23.62500 82.000000 0.0000000  0.6536542  0.3463458         0.6595927
## 5 27.37864  4.417476 0.5814563  0.5929434  0.4070566         0.6574877
## 6 27.04688  6.109375 0.9785938  0.6309836  0.3690164         0.6747399
##        Aces firstServeReturnsWon SecondServeReturnsWon FirstServesIn
## 1  7.384615             18.53846              18.78462      37.32308
## 2 10.950000             30.10000              34.20000     111.55000
## 3  8.058824             22.58824              22.88235      53.23529
## 4 17.625000             24.00000              24.87500      77.62500
## 5  8.980583             21.17476              22.32039      56.60194
## 6 12.390625             24.54688              26.01562      80.98438
##   DoubleFaults FirstServePercentage
## 1     1.430769            0.6134388
## 2     3.150000            0.6728199
## 3     3.176471            0.5572020
## 4     2.375000            0.5503397
## 5     1.980583            0.6398418
```

```
## 6     3.671875            0.6325176
##
## Clustering vector:
##   [1] 6 1 5 2 1 1 6 5 1 6 5 5 1 2 5 5 1 6 1 5 1 5 5 5 1 5 1 5 5 1 6 5 1 1 6
##  [36] 5 6 1 6 1 2 5 1 6 5 6 1 6 5 1 5 1 5 6 6 2 1 6 1 5 5 2 5 6 6 1 6 5 6 6
##  [71] 3 3 3 6 6 3 3 5 6 6 5 5 6 5 6 1 1 1 6 6 1 4 4 4 4 4 4 4 5 1 1 5 2 1 6
## [106] 4 3 3 3 3 3 3 6 6 1 5 6 6 6 5 5 5 1 6 5 1 5 6 5 1 1 6 5 1 2 2 1 1 5 1
## [141] 5 1 2 5 6 5 1 1 5 6 6 5 5 5 5 1 2 6 6 1 5 6 5 1 5 1 5 5 2 1 5 5 1 2 5
## [176] 1 5 2 5 2 2 5 5 1 6 6 5 6 6 2 5 6 5 2 2 1 5 1 5 5 5 5 3 3 3 3 3 3 6 5
## [211] 6 1 5 5 1 6 6 1 1 6 5 6 1 5 5 5 5 5 1 2 6 5 1 1 6 5 6 5 6 5 5 5 1 5 6
## [246] 5 6 1 2 5 5 5 5 5 1 6 5 5 1 6 5 6 6 6 5 5 5 5 5 5 5 5 1 5 5 5 1
##
## Within cluster sum of squares by cluster:
## [1] 11790.704  6504.073  5952.405  2748.490 18068.963 18651.769
##  (between_SS / total_SS =  75.3 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

```r
perc.var.6<- round(100*(1 - kmeans6$betweenss/kmeans6$totss),1)
names(perc.var.6) <- "Perc. 6 clus"
perc.var.6
```

```
## Perc. 6 clus
##         24.7
```

```r
## Computing the percentage of variation accounted for three clusters
(kmeans9<-kmeans(AustralianOpen_Finalists_allstats_Numeric,9,nstart = 10))
```

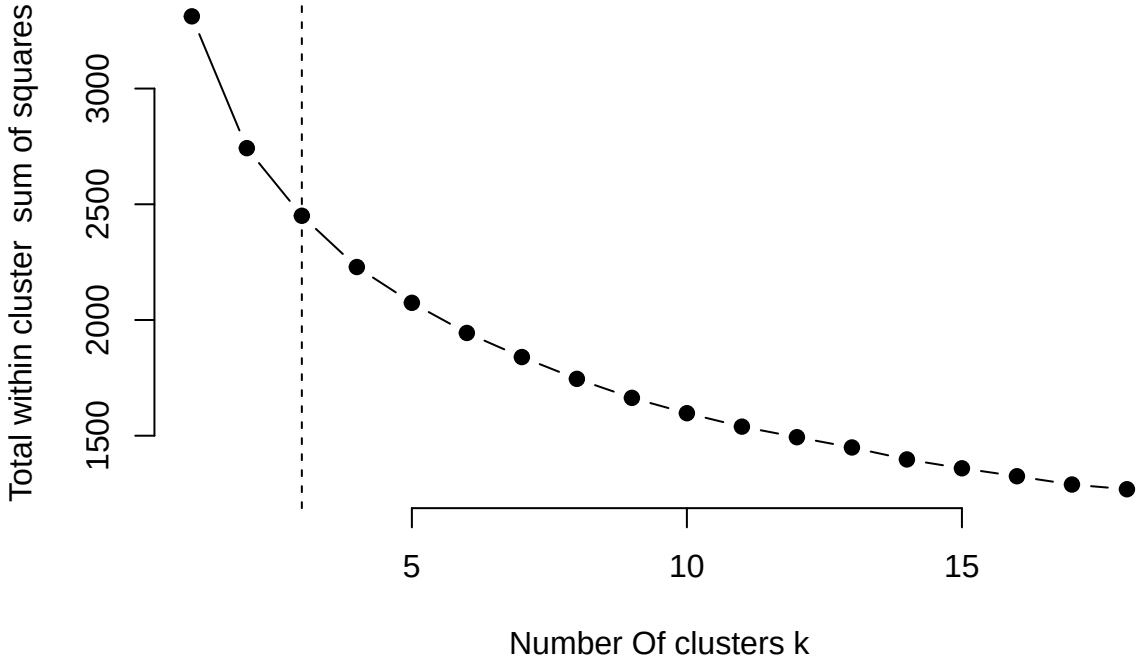
```
## K-means clustering with 9 clusters of sizes 8, 20, 7, 6, 45, 14, 54, 60, 63
##
## Cluster means:
##        Age      Rank   avgOdds SP_Percent RP_Percent BP_Win_Percentage
## 1 28.37500  4.000000 0.6837500  0.5046217  0.4953783         0.4062500
## 2 27.40000  4.600000 0.9935000  0.6323234  0.3676766         0.6780857
## 3 24.00000 86.000000 0.0000000  0.6586875  0.3413125         0.6697866
## 4 22.00000 46.000000 0.0000000  0.6289503  0.3710497         0.6437042
## 5 27.42222  5.133333 1.1042222  0.6377472  0.3622528         0.6763246
## 6 24.14286 41.714286 0.0000000  0.5756546  0.4243454         0.4206131
## 7 26.61111  6.018519 0.5281481  0.5952129  0.4047871         0.6479549
## 8 26.13333  4.033333 0.4476667  0.5498040  0.4501960         0.3605357
## 9 28.12698  3.000000 0.7147619  0.6025018  0.3974982         0.6630678
##        Aces firstServeReturnsWon SecondServeReturnsWon FirstServesIn
## 1  4.250000             13.50000              10.62500      18.75000
## 2 10.950000             30.10000              34.20000     111.55000
## 3 18.000000             22.42857              25.14286      76.42857
## 4 11.666667             26.33333              26.50000      81.00000
## 5 12.111111             26.06667              24.42222      84.08889
## 6  8.285714             21.71429              21.64286      49.28571
## 7 11.351852             22.37037              27.25926      65.37037
## 8  8.166667             19.43333              20.90000      40.30000
## 9  7.650794             19.98413              19.44444      53.52381
```

```
##   DoubleFaults FirstServePercentage
## 1     0.625000            0.6063924
## 2     3.150000            0.6728199
## 3     2.428571            0.5491823
## 4     3.666667            0.5930853
## 5     3.822222            0.6354181
## 6     2.714286            0.5601654
## 7     2.870370            0.6226843
## 8     1.600000            0.6134588
## 9     1.587302            0.6509678
##
## Clustering vector:
##   [1] 5 8 9 2 8 8 5 9 8 5 9 7 8 2 9 9 1 5 8 9 8 8 7 9 1 9 8 7 9 8 5 9 8 8 7
##  [36] 9 5 8 5 8 2 9 8 5 7 5 8 5 9 8 7 8 7 5 7 2 8 5 8 7 7 2 9 5 7 8 5 9 7 5
##  [71] 6 4 6 4 4 6 6 7 5 5 9 9 5 7 7 8 1 8 7 5 8 3 3 3 3 3 3 3 7 8 8 7 2 8 5
## [106] 4 6 4 6 4 6 6 5 7 8 9 7 5 5 9 7 7 8 5 9 8 9 5 8 8 1 5 7 9 2 2 8 8 7 8
## [141] 9 8 2 7 5 9 8 8 9 5 5 7 7 9 9 8 2 7 7 8 7 5 9 1 9 8 9 7 2 8 9 9 8 2 9
## [176] 8 9 2 9 2 2 9 9 1 5 5 9 5 7 2 9 5 7 2 2 8 9 9 9 7 9 9 6 6 6 6 6 5 8
## [211] 7 8 9 9 8 7 7 8 8 5 8 5 8 9 7 7 9 9 8 2 5 9 8 8 5 9 7 9 7 7 9 9 8 7 5
## [246] 7 5 8 2 9 9 9 9 9 1 5 9 7 1 5 9 7 5 5 7 7 7 7 8 7 7 9 8 7 9 7 8
##
## Within cluster sum of squares by cluster:
## [1]   866.5893  6504.0732  1614.1058  1591.6823 10508.1946  4038.6627
## [7] 10250.4368  7240.8805  7800.9030
##  (between_SS / total_SS =  80.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

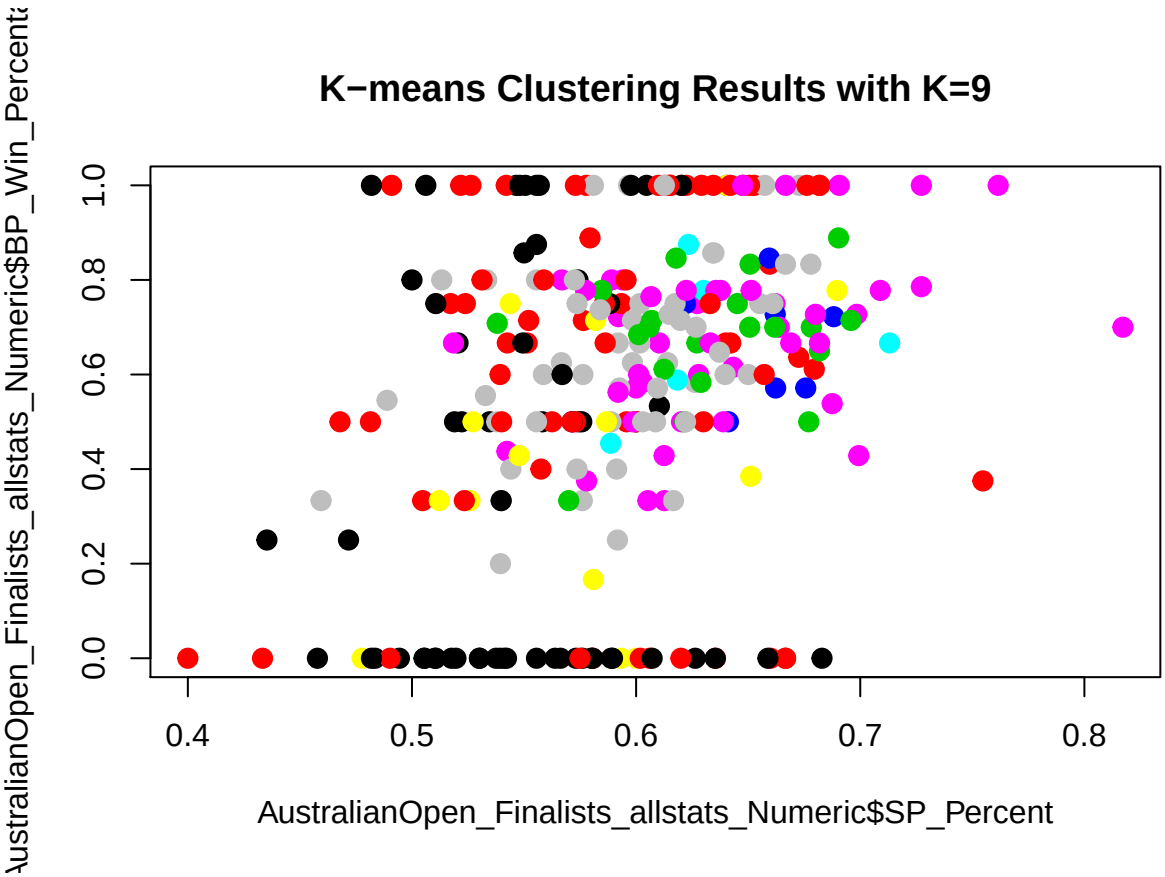
```
perc.var.9<- round(100*(1 - kmeans9$betweenss/kmeans9$totss),1)
names(perc.var.9) <- "Perc. 9 clus"
perc.var.9
```

```
## Perc. 9 clus
##         19.6
```

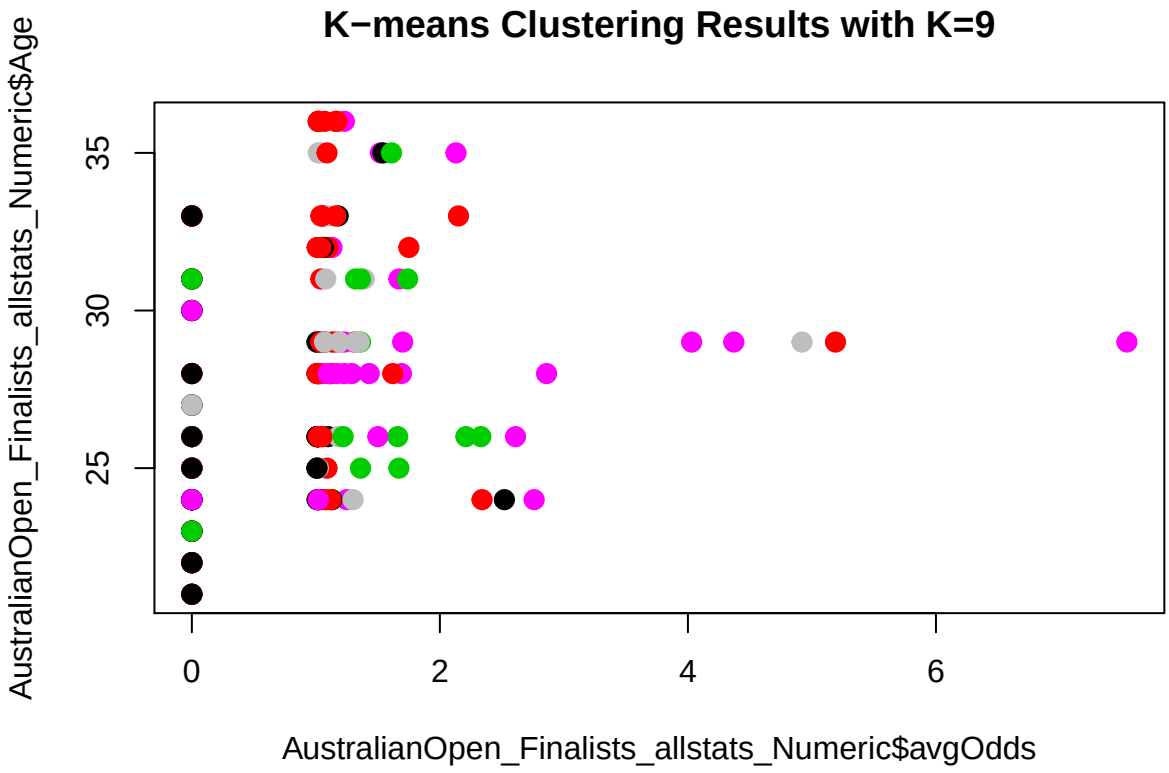
```
AustralianOpen_Finalists_allstats_Numeric_Scale<-scale(AustralianOpen_Finalists_allstats_Numeric)


k.max<-18
wss<-sapply(1:k.max,function(k){kmeans(AustralianOpen_Finalists_allstats_Numeric_Scale,k,nstart=50)$tot

plot(1:k.max,wss, type='b',pch =19,frame = FALSE, xlab ="Number Of clusters k",ylab ="Total within clust
abline(v=3, lty=2)
```



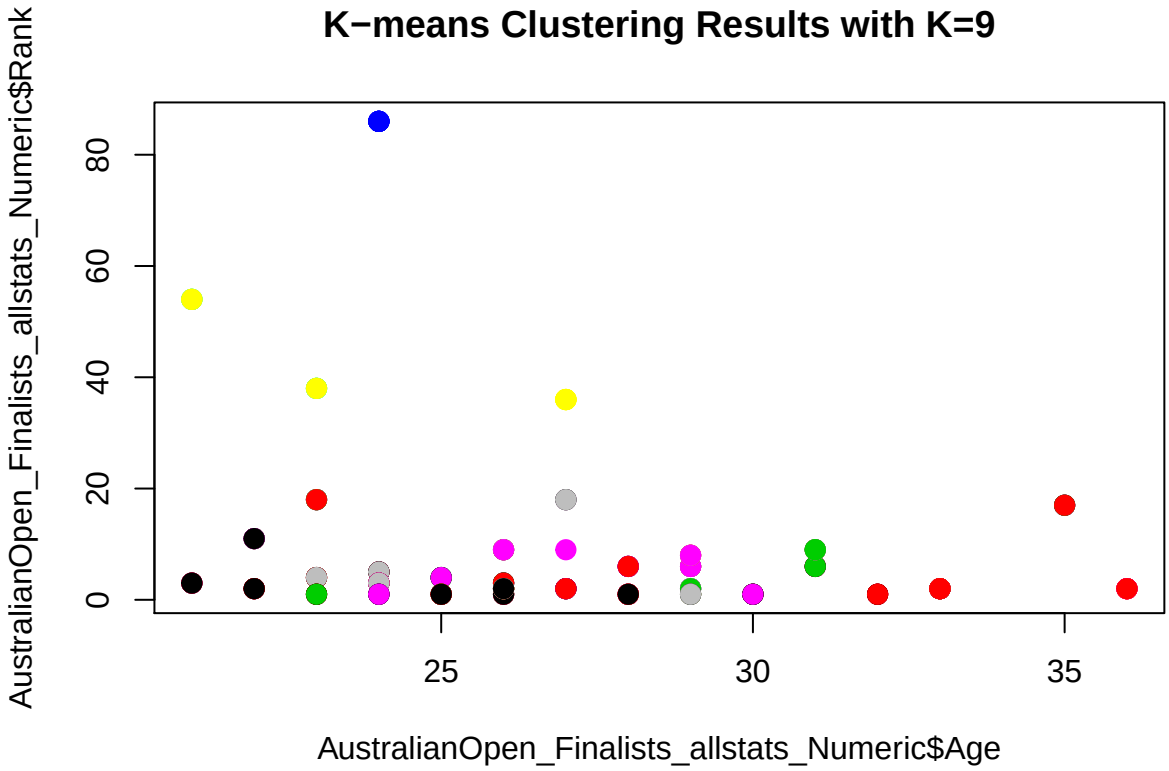

```
plot(AustralianOpen_Finalists_allstats_Numeric$SP_Percent,AustralianOpen_Finalists_allstats_Numeric$BP_
```




```
plot(AustralianOpen_Finalists_allstats_Numeric$avgOdds,AustralianOpen_Finalists_allstats_Numeric$Age,col
```




```
plot(AustralianOpen_Finalists_allstats_Numeric$Age,AustralianOpen_Finalists_allstats_Numeric$Rank,col=(
```


K−means Clustering Results with K=9
AustralianOpen_Finalists_allstats_Numeric$Rank
0
20
40
60
80
25
30
35
AustralianOpen_Finalists_allstats_Numeric$Age