Schooling and Wages K-Means Analysis

Robert C Phillips

February 20, 2016

Introduction

The purpose of this assignment is to run a k-means analysis determine if groups of people exist that ultimately relate to wage. I am using R for this analysis as an additional challenge since Python and SAS solutions were already provided within the lectures.

For this analysis I am using the “Wages and Schooling” dataset provided here: http://vincentarelbundock.github.io/Rdatasets/doc/Ecdat/Schooling.html (http://vincentarelbundock.github.io/Rdatasets/doc/Ecdat/Schooling.html).

```
data <- read.csv("Schooling.csv")
```

The variables included in this dataset are:

- X: observation identifier
- smsa66: lived in smsa in 1966 ?
- smsa76: lived in smsa in 1976 ?
- nearc2: grew up near 2-yr college ?
- nearc4: grew up near 4-yr college ?
- nearc4a: grew up near 4-year public college ?
- nearc4b: grew up near 4-year private college ?
- ed76: education in 1976
- ed66: education in 1966
- age76: age in 1976
- daded: dads education (imputed avg if missing)
- nodaded: dads education imputed ?
- momed: mothers education
- nomomed: moms education imputed ?
- momdad14: lived with mom and dad at age 14 ?
- sinmom14: single mom at age 14 ?
- step14: step parent at age 14 ?
- south66: lived in south in 1966 ?
- south76: lived in south in 1976 ?
- lwage76: log wage in 1976 (outliers trimmed)
- famed: mom-dad education class (1-9)
- black: black ?
- wage76: wage in 1976 (raw, cents per hour)
- enroll76: enrolled in 1976 ?
- kww: the kww score
- iqscore: a normed IQ score
- mar76: married in 1976 ?

- libcrd14: library card in home at age 14 ?
- exp76: experience in 1976

The variable of interest for this analysis will be the **wage76** variable. It will be used to evaluate the established clusters.

Exploratory Analysis

Summarizing the data indcates the varables **kww**, **iqscore**, **mar76** and **libcrd14** contain missing values.

```
summary(data)
```

```
##        X          smsa66     smsa76     nearc2     nearc4     nearc4a
##  Min.   :   1.0   no :1055   no : 864   no :1683   no : 957   no :1527
##  1st Qu.: 753.2   yes:1955   yes:2146   yes:1327   yes:2053   yes:1483
##  Median :1505.5
##  Mean   :1505.5
##  3rd Qu.:2257.8
##  Max.   :3010.0
##
##  nearc4b         ed76            ed66           age76
##  no :2440   Min.   : 1.00   Min.   : 0.00   Min.   :24.00
##  yes: 570   1st Qu.:12.00   1st Qu.: 9.00   1st Qu.:25.00
##             Median :13.00   Median :11.00   Median :28.00
##             Mean   :13.26   Mean   :10.76   Mean   :28.12
##             3rd Qu.:16.00   3rd Qu.:12.00   3rd Qu.:31.00
##             Max.   :18.00   Max.   :18.00   Max.   :34.00
##
##      daded         nodaded        momed       nomomed    momdad14
##  Min.   : 0.000   no :2320   Min.   : 0.00   no :2320   no : 634
##  1st Qu.: 8.000   yes: 690   1st Qu.: 9.00   yes: 690   yes:2376
##  Median : 9.940              Median :11.00
##  Mean   : 9.989              Mean   :10.34
##  3rd Qu.:12.000              3rd Qu.:12.00
##  Max.   :18.000              Max.   :18.00
##
##  sinmom14   step14     south66    south76       lwage76
##  no :2707   no :2893   no :1763   no :1795   Min.   :4.605
##  yes: 303   yes: 117   yes:1247   yes:1215   1st Qu.:5.977
##                                             Median :6.287
##                                             Mean   :6.262
##                                             3rd Qu.:6.564
##                                             Max.   :7.785
##
##      famed       black          wage76       enroll76        kww
##  Min.   :1.000   no :2307   Min.   : 100.0   no :2732   Min.   : 4.00
##  1st Qu.:3.000   yes: 703   1st Qu.: 394.2   yes: 278   1st Qu.:28.00
##  Median :6.000              Median : 537.5              Median :34.00
##  Mean   :5.934              Mean   : 577.3              Mean   :33.54
##  3rd Qu.:8.000              3rd Qu.: 708.8              3rd Qu.:40.00
##  Max.   :9.000              Max.   :2404.0              Max.   :56.00
##                                                         NA's   :47
##     iqscore       mar76      libcrd14        exp76
##  Min.   : 50.0   2   :  14   no  : 976   Min.   : 0.000
##  1st Qu.: 93.0   3   :   3   yes :2021   1st Qu.: 6.000
##  Median :103.0   4   : 155   NA's:  13   Median : 8.000
##  Mean   :102.4   5   : 102               Mean   : 8.856
##  3rd Qu.:113.0   6   : 585               3rd Qu.:11.000
##  Max.   :149.0   yes :2144               Max.   :23.000
##  NA's   :949     NA's:   7
```

The variable **iqscore** would be interesting to include, but is missing for almost 1/3 of the observations. Therefore it will be removed. The other variables have a low number of missing values, therefore we'll filter out those observations so that we can include the variables in the cluster analysis.

```
#remove the iqscore variable
data$iqscore <- NULL

#remove observations with NA values
data <- na.omit(data)
```

K-Means Analysis

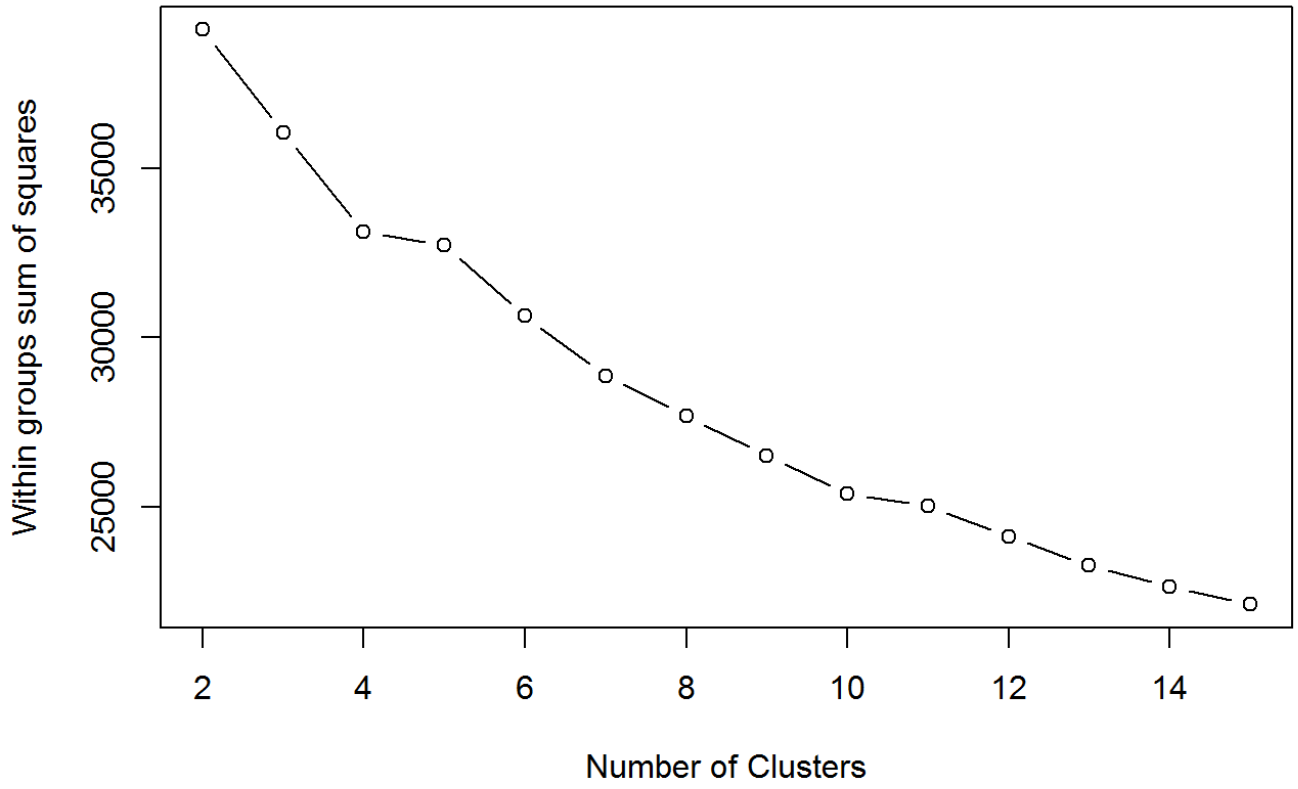
We will use the **kmeans** method in R for this analysis. We will convert our data frame to a matrix and retain the clustering variables within matrix and create a seperate vector for the target variable. Using a matrix has the convenience of converting categorical variables into binary variables (using dummy variables.) Once we have a numeric matrix, we can scale the values to ensure 1 column doesn't dominate.

We will establish those items as the variables X and Y. Furthermore, we need to split the data into a training and test set.

```
#define the model with wage76 as the target
model1 <- wage76 ~ smsa66 + smsa76 + nearc2 + nearc4 + nearc4a + nearc4b + ed76 + ed66
+ age76 +
                   daded + nodaded + momed + nomomed + momdad14 + sinmom14 + step14 + s
outh66 + south76 +
                   famed + black + enroll76 + kww + mar76 + libcrd14 + exp76

#build the matrix and vector
X <- scale(model.matrix(model1, data)[,-1])
Y <- data$wage76

#center and scale our matrix values
X <- scale(X)

#split into training and test sets
set.seed(1972)
train.rows <- sample(1:nrow(X), nrow(X) / 2)
test.rows <- (-train.rows)

X.train <- X[train.rows,]
Y.train <- Y[train.rows]

X.test <- X[test.rows,]
Y.test <- Y[test.rows]
```

Now that we have our training and test data we can run the k-means analysis. We will run the analysis mutliple times, each time adding an additional cluster. The **kmeans** method in R gives us the squared Euclidean distance in the resulting object. We can use that to plot the change as the number of clusters increases.

```
#store the values
kms <- list()
set.seed(100)
for (i in 2:15) kms[[i]] <- kmeans(X.train, centers=i, nstart=50)

kms.withinss <- sapply(sapply(kms, "[[", "withinss"), sum)

plot(2:15, kms.withinss[2:15], type="b",
     xlab="Number of Clusters",
     ylab="Within groups sum of squares")
```



The plot shows a rapid decrease in the Within Sum of Squares (WSS) between 2 and 4 clusters, and then further decreses beyond 5 clusters. Since this data is concerned with poeople and education levels, a large number of clusters may be difficult to interpet. Therefore, for the subsequent analysis, we choose to use 4 clusters. The following shows the counts per cluster.

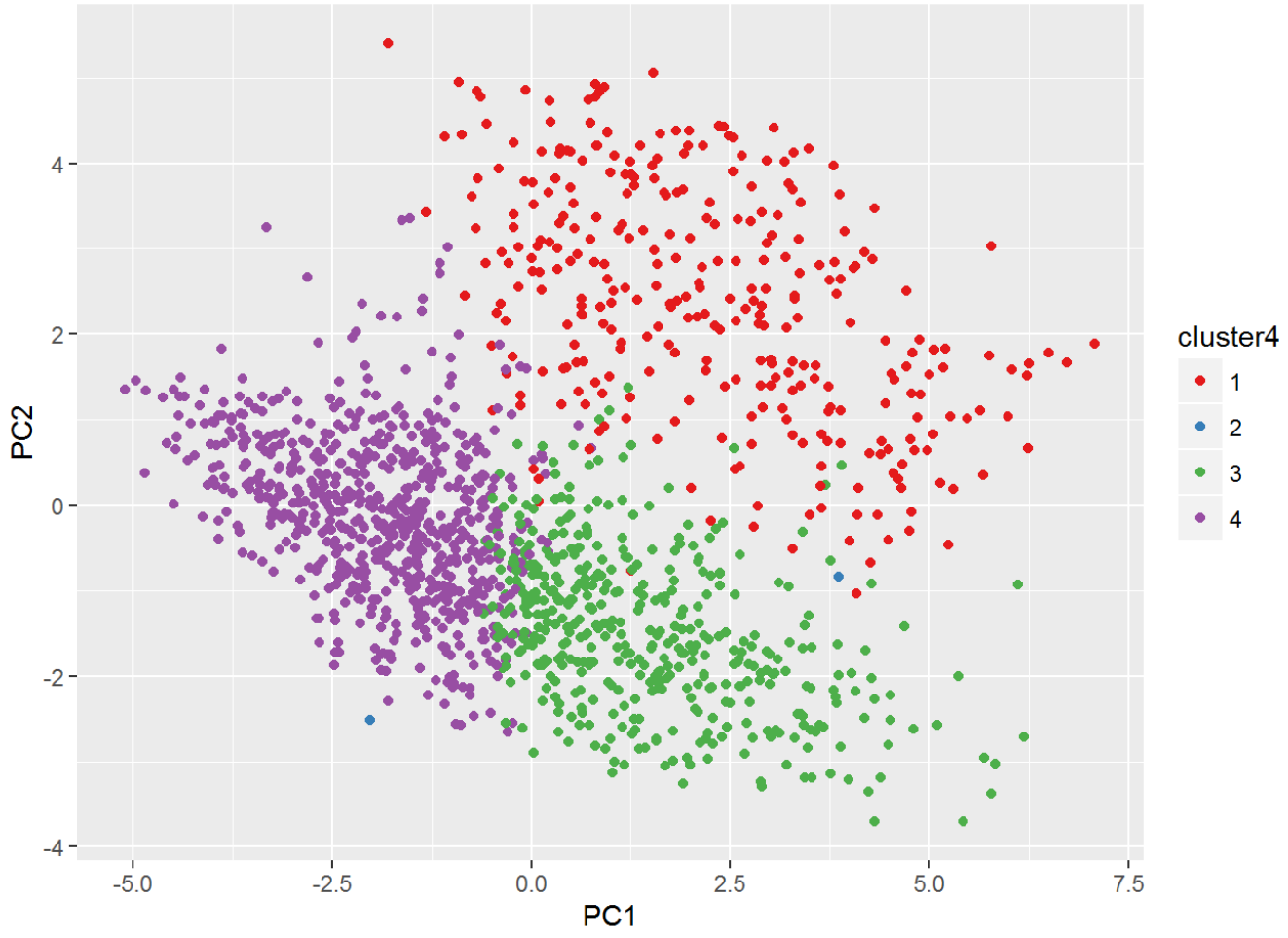
```
#store the 4-cluster model in a new variable
kms.4 <- kms[[4]]

#display the counts per cluster
table(kms.4$cluster)
```

```
##
##   1   2   3   4
## 317   3 438 714
```

Cluster 2 has just 3 observations assigned to it. This may not be desirable in general, but we'll stick with 4 clusters to finish this part of the analysis.

We will use principal component analsys to build new features that capture the variance, and use the first 2 components to visualize the clusters.

```
#perform pca
X.train.pca <- prcomp(X.train)

#get the first 2 principal components
X.train.pca.2 <- as.data.frame(X.train.pca$x[,1:2])

#4-cluster assignment
X.train.pca.2$cluster4 <- as.factor(kms.4$cluster)

#plot the 4-cluster assignment
require(ggplot2)
require(RColorBrewer)
ggplot(X.train.pca.2, aes(x=PC1, y=PC2, color=cluster4)) + geom_point(shape=16) +
  scale_colour_brewer(palette="Set1")
```



To interpret the visualization, we can look at the correlation of PC1 and PC2 with the original variables.

```
X.train.cor <- cor(X.train, X.train.pca.2[1:2])
X.train.cor
```

```
##                     PC1         PC2
## smsa66yes   -0.41216943  0.42566000
## smsa76yes   -0.31912148  0.34686860
## nearc2yes   -0.19311286  0.24899676
## nearc4yes   -0.36321378  0.39988110
## nearc4ayes  -0.37079780  0.34866026
## nearc4byes   0.04008747  0.03338634
## ed76        -0.70078139  0.05636691
## ed66        -0.53791475  0.26590156
## age76       -0.01324134  0.32006059
## daded       -0.57792168  0.20581867
## nodadedyes   0.48177417  0.73471480
## momed       -0.62189882  0.06399015
## nomomedyes   0.48177417  0.73471480
## momdad14yes -0.41117991 -0.63630531
## sinmom14yes  0.32980251  0.58677491
## step14yes    0.05056537  0.14879846
## south66yes   0.56753961 -0.22000857
## south76yes   0.51860601 -0.22985259
## famed        0.74317685  0.15501099
## blackyes     0.55663751  0.09607213
## enroll76yes -0.10593669 -0.00406151
## kww         -0.62194524  0.21098000
## mar763       0.01567949 -0.03434817
## mar764       0.04046214  0.10294816
## mar765       0.05969257  0.11387667
## mar766      -0.08950027 -0.05073117
## mar76yes     0.01988210 -0.04921286
## libcrd14yes -0.57489316  0.15931451
## exp76        0.43617402  0.20564100
```

This data shows that PC1 represents a negative correlation with **ed76 (-0.70078139)** and **momed (-0.62189882)**. PC1 also shows a positive correlation with **famed (0.74317685).** PC2 represents a positive correlation with **nodadedyes (0.73471480)** and **nomomedyes (0.73471480)**. PC2 also has a negative correlation with **momdad14yes (-0.63630531).** However, none of the correlations are strong, therefore further analysis would be needed for further interpretation.

Using this information, we can summarize the visualized clusters as follows.

1. The red cluster represents observations where mom and dad education is more prominent. 2 The green cluster represents observations where family education is more prominent.
2. The purple cluster represents observations where mom, dad, and family eduation are less promiment.

Note that the fourth cluster is diffcult to find on the visualization since it only contains 3 points.

Regarding the wage variable that is the target variable of interest in this analysis, we can apply the clusters to that data and look at the per group means.

```
#add cluster to the wage values
Y.train.4 <- data.frame(wage=Y, cluster=as.factor(kms.4$cluster))

#compute means per cluster
aggregate(wage ~ cluster, data = Y.train.4, FUN = mean)
```

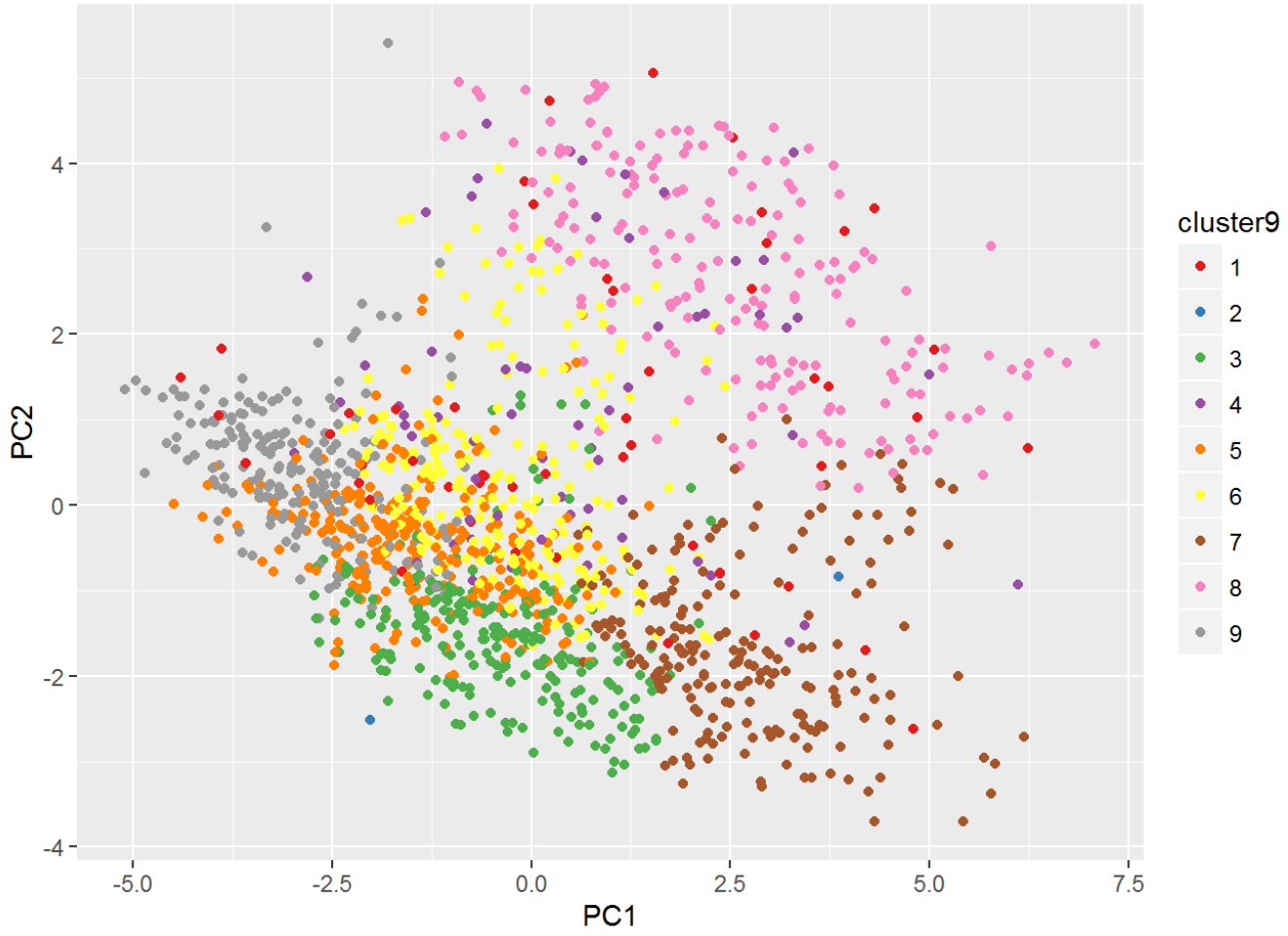
```
##   cluster     wage
## 1       1 593.8044
## 2       2 675.1667
## 3       3 587.1598
## 4       4 565.7507
```

There does not appear to be much difference in the means between the clusters. This may indicate that this particular clustering of the data is not effective. Since we have the clustering data readily avialable, we could look at a different number of clusters and make additional conclusions. The following is the output for 9 clusters.

```
#store the 9-cluster model in a new variable
kms.9 <- kms[[9]]

#display the counts per cluster
table(kms.9$cluster)
```

```
##
##   1   2   3   4   5   6   7   8   9
##  52   3 227  63 282 231 210 196 208
```

```
#10-cluster assignment (also remove 4)
X.train.pca.2$cluster4 <- NULL
X.train.pca.2$cluster9 <- as.factor(kms.9$cluster)

#plot the clustering
ggplot(X.train.pca.2, aes(x=PC1, y=PC2, color=cluster9)) + geom_point(shape=16) +
  scale_colour_brewer(palette="Set1")
```




```
#add cluster to the wage values
Y.train.9 <- data.frame(wage=Y, cluster=as.factor(kms.9$cluster))

#compute means per cluster
aggregate(wage ~ cluster, data = Y.train.9, FUN = mean)
```

```
##   cluster     wage
## 1       1 536.1250
## 2       2 675.1667
## 3       3 577.4493
## 4       4 588.0000
## 5       5 565.0071
## 6       6 576.7186
## 7       7 598.8190
## 8       8 591.7423
## 9       9 572.4375
```

As expected, numerous clusters will be difficult to interpret, which is evident given the overlap shown in the visualization. Additionally, the means of the wages per cluster are more uniform than not.

It's possible this dataset is not adequate for showing difference in wages. Other datasets could be located and analyzed in a similar fashion. Perhaps thsi will be done in a future assignment.