Traffic Accidents Analysis

Homer Kay

4/20/2018

```
## EDA
ACC<- read.csv("~/Desktop/Data Mining/Accidents.csv")
str(ACC)

## 'data.frame':    42183 obs. of  21 variables:
##  $ RUSH_HR        : int  0 1 1 1 1 1 1 1 1 0 ...
##  $ ALCOHOL        : int  2 2 2 2 1 2 2 2 2 2 ...
##  $ ROAD_ALIGN     : int  2 1 1 1 1 1 1 1 1 1 ...
##  $ WORK_ZONE      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ WEEKDAY        : int  1 1 1 0 1 1 1 1 1 0 ...
##  $ INT_HWY        : int  0 1 0 0 0 0 1 0 0 0 ...
##  $ LIGHT_COND     : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ MANCOL         : int  0 2 2 2 2 0 0 0 0 0 ...
##  $ PED_CYCLIST    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ INTERSTATE     : int  1 1 1 1 0 1 0 0 1 1 ...
##  $ ROADWAY        : int  0 1 1 1 1 0 0 0 0 0 ...
##  $ ROAD_PROFILE   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ SPEED_LIMIT    : int  40 70 35 35 25 70 70 35 30 25 ...
##  $ SURFACE_COND   : int  4 4 4 4 4 4 4 4 4 4 ...
##  $ TRAFFIC_CTRL   : int  0 0 1 1 0 0 0 0 0 0 ...
##  $ TRAFFIC_WAY    : int  3 3 2 2 2 2 2 2 1 1 ...
##  $ VEH_INVL       : int  1 2 2 2 3 1 1 1 1 1 ...
##  $ ADVERSE_WEATHER: int  1 2 2 1 1 2 2 1 2 2 ...
##  $ INJURY_CRASH   : int  1 0 0 0 0 1 0 1 0 0 ...
##  $ PROPERTY_DMG   : int  0 1 1 1 1 0 1 0 1 1 ...
##  $ REGION         : int  1 2 4 4 2 3 2 1 3 2 ...

## STR shows which varaibles are numeric, integer, factor, etc.
library(corrplot)

## corrplot 0.84 loaded

## Make correlation matrix with numeric predictors
CORR_MATRIX <- cor(ACC[,])
CORR_MATRIX

##                       RUSH_HR      ALCOHOL     ROAD_ALIGN     WORK_ZONE
## RUSH_HR          1.0000000000  0.104196388 -2.471585e-02 -0.0204856215
## ALCOHOL          0.1041963878  1.000000000 -6.223688e-02  0.0057420002
## ROAD_ALIGN      -0.0247158511 -0.062236878  1.000000e+00 -0.0134424225
## WORK_ZONE       -0.0204856215  0.005742000 -1.344242e-02  1.0000000000
## WEEKDAY          0.1027125078  0.130572558 -4.723039e-02  0.0208451284
## INT_HWY         -0.0144574933  0.024419211  5.669899e-03  0.1010841244
```
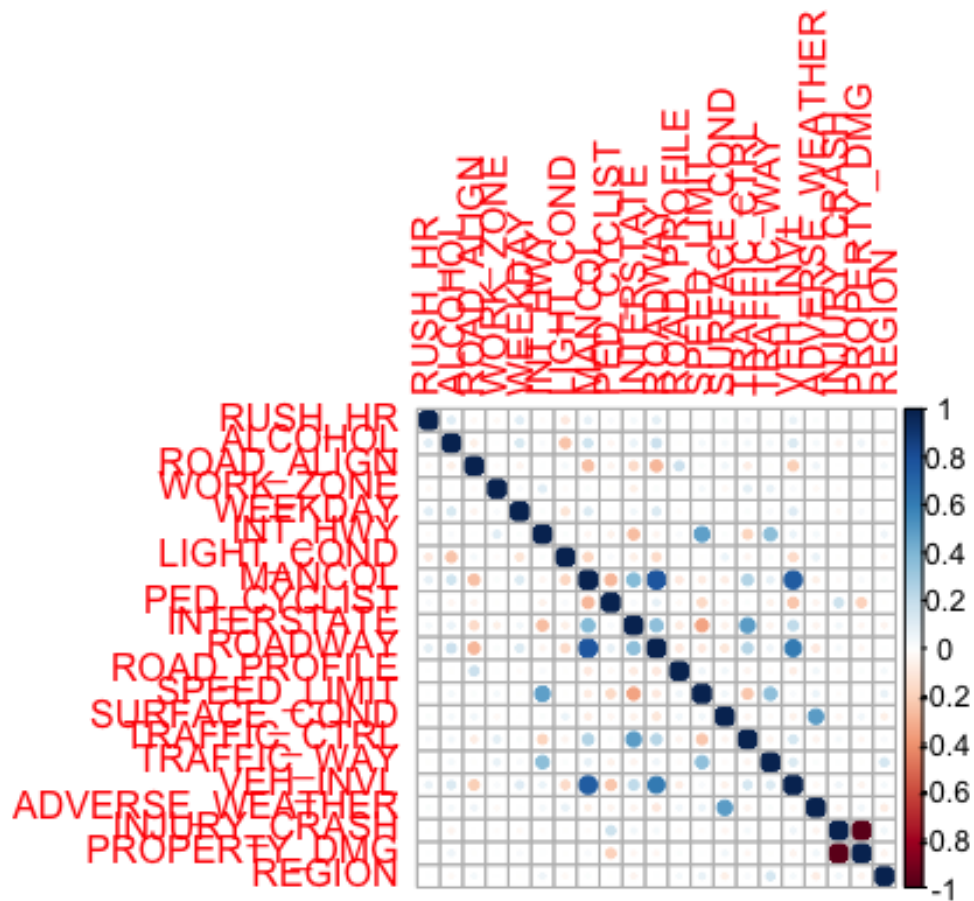
```
## LIGHT_COND       -0.0940467095 -0.241099598  3.605178e-02 -0.0151129418
## MANCOL            0.0861365800  0.160393303 -2.571246e-01  0.0168800420
## PED_CYCLIST       0.0211946191 -0.035818117 -4.582314e-02 -0.0061870237
## INTERSTATE        0.0342730773  0.067771927 -1.580224e-01 -0.0646174045
## ROADWAY           0.0934213295  0.180300432 -2.862550e-01  0.0194943515
## ROAD_PROFILE      0.0008857834 -0.008943752  1.811904e-01  0.0070474301
## SPEED_LIMIT       0.0018843235  0.034532934  1.542608e-02  0.0352848146
## SURFACE_COND      0.0188966785  0.032618418  4.932024e-02 -0.0247876921
## TRAFFIC_CTRL      0.0238949498  0.040653367 -7.160326e-02  0.0932835891
## TRAFFIC_WAY      -0.0095679262  0.038041615 -7.147204e-03  0.0742863300
## VEH_INVL          0.0758754378  0.127131540 -2.095758e-01  0.0269188864
## ADVERSE_WEATHER   0.0124662140  0.033873582  5.346738e-02 -0.0219745569
## INJURY_CRASH     -0.0014844791 -0.046697541  8.785779e-06 -0.0101539401
## PROPERTY_DMG      0.0057070331  0.055845898 -6.560590e-03  0.0093704492
## REGION            0.0052985899 -0.025775215  3.152734e-02  0.0006819275
##                       WEEKDAY      INT_HWY    LIGHT_COND        MANCOL
## RUSH_HR           0.1027125078 -0.014457493 -0.094046709   0.08613658
## ALCOHOL           0.1305725576  0.024419211 -0.241099598   0.16039330
## ROAD_ALIGN       -0.0472303929  0.005669899  0.036051782  -0.25712461
## WORK_ZONE         0.0208451284  0.101084124 -0.015112942   0.01688004
## WEEKDAY           1.0000000000  0.013909962 -0.115682763   0.12192755
## INT_HWY           0.0139099622  1.000000000  0.013247980  -0.03642033
## LIGHT_COND       -0.1156827626  0.013247980  1.000000000  -0.17587138
## MANCOL            0.1219275546 -0.036420332 -0.175871381   1.00000000
## PED_CYCLIST       0.0046719436 -0.056510005  0.022724593  -0.28801194
## INTERSTATE        0.0400322283 -0.263744314 -0.066142355   0.38946373
## ROADWAY           0.1265311133 -0.042219973 -0.152725503   0.77858483
## ROAD_PROFILE     -0.0120781142  0.026866782 -0.005474871  -0.09600606
## SPEED_LIMIT      -0.0098337731  0.488045198 -0.004938672  -0.08495529
## SURFACE_COND     -0.0000220246  0.017008677  0.062109790  -0.07584111
## TRAFFIC_CTRL      0.0185555744 -0.182650231 -0.044618578   0.25672234
## TRAFFIC_WAY       0.0238640586  0.364103025  0.013735184   0.08290255
## VEH_INVL          0.1173099351  0.061608564 -0.156773615   0.74346423
## ADVERSE_WEATHER   0.0036428090  0.037461710  0.075797833  -0.06138246
## INJURY_CRASH     -0.0214379538 -0.013596942 -0.005460283  -0.01414155
## PROPERTY_DMG      0.0250074393  0.009756132 -0.001352061   0.02771440
## REGION           -0.0052714849  0.057276401 -0.026635224   0.01724555
##                    PED_CYCLIST   INTERSTATE       ROADWAY   ROAD_PROFILE
## RUSH_HR           0.021194619  0.03427308  0.093421330   0.0008857834
## ALCOHOL          -0.035818117  0.06777193  0.180300432  -0.0089437517
## ROAD_ALIGN       -0.045823143 -0.15802236 -0.286255046   0.1811903686
## WORK_ZONE        -0.006187024 -0.06461740  0.019494351   0.0070474301
## WEEKDAY           0.004671944  0.04003223  0.126531113  -0.0120781142
## INT_HWY          -0.056510005 -0.26374431 -0.042219973   0.0268667819
## LIGHT_COND        0.022724593 -0.06614235 -0.152725503  -0.0054748714
## MANCOL           -0.288011938  0.38946373  0.778584830  -0.0960060641
## PED_CYCLIST       1.000000000  0.02941686  0.083842175  -0.0248510600
## INTERSTATE        0.029416858  1.00000000  0.354925769  -0.0873749207
## ROADWAY           0.083842175  0.35492577  1.000000000  -0.0998846525
## ROAD_PROFILE     -0.024851060 -0.08737492 -0.099884652   1.0000000000
```
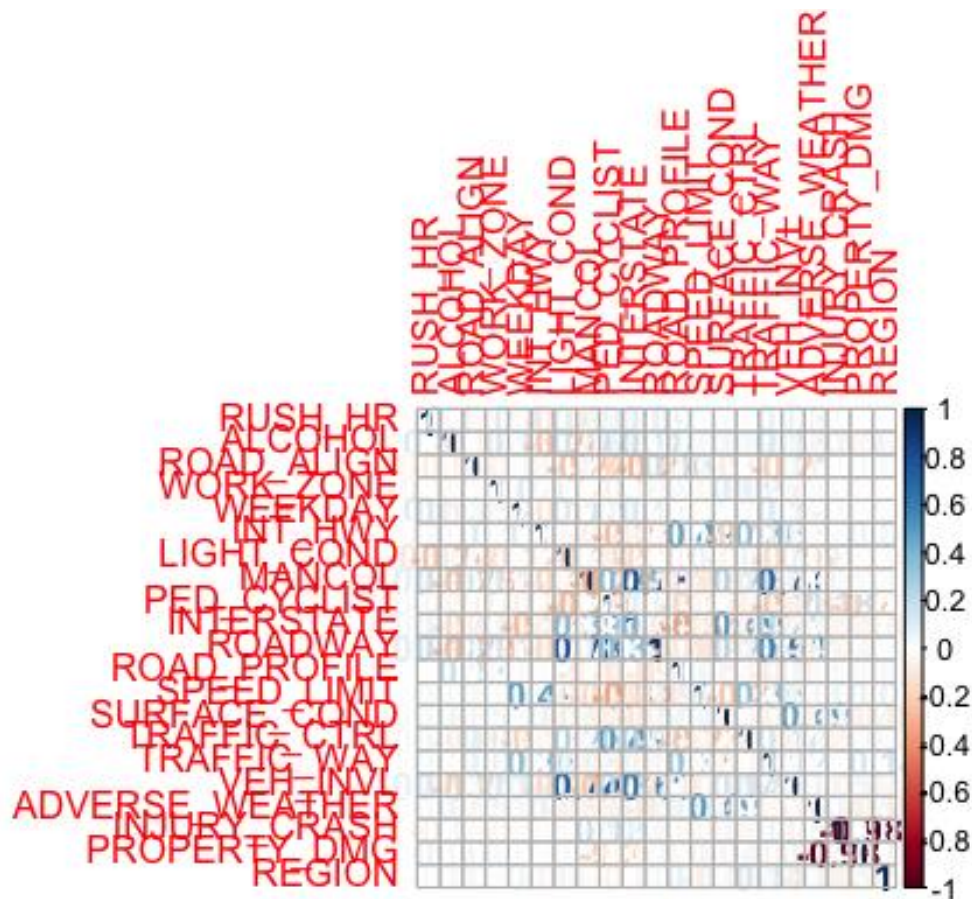
```
## SPEED_LIMIT      -0.160725553 -0.34380150 -0.094740348  0.0549377114
## SURFACE_COND     -0.037406649 -0.05570470 -0.109641189  0.0418616213
## TRAFFIC_CTRL      0.014216289  0.49149736  0.247119996 -0.0430934791
## TRAFFIC_WAY      -0.057366044 -0.04862394  0.052923519 -0.0161612600
## VEH_INVL         -0.233719863  0.21466243  0.610640834 -0.0721359201
## ADVERSE_WEATHER  -0.042965679 -0.04780733 -0.096888703  0.0466001285
## INJURY_CRASH      0.179725823  0.04198543  0.037737501 -0.0186513003
## PROPERTY_DMG     -0.198300171 -0.03225288 -0.032899520  0.0146024023
## REGION           -0.002612402 -0.04477315  0.003026648  0.0332729632
##                    SPEED_LIMIT  SURFACE_COND TRAFFIC_CTRL   TRAFFIC_WAY
## RUSH_HR            0.001884323  0.0188966785   0.02389495 -0.009567926
## ALCOHOL            0.034532934  0.0326184182   0.04065337  0.038041615
## ROAD_ALIGN         0.015426084  0.0493202421  -0.07160326 -0.007147204
## WORK_ZONE          0.035284815 -0.0247876921   0.09328359  0.074286330
## WEEKDAY           -0.009833773 -0.0000220246   0.01855557  0.023864059
## INT_HWY            0.488045198  0.0170086767  -0.18265023  0.364103025
## LIGHT_COND        -0.004938672  0.0621097895  -0.04461858  0.013735184
## MANCOL            -0.084955290 -0.0758411060   0.25672234  0.082902546
## PED_CYCLIST       -0.160725553 -0.0374066488   0.01421629 -0.057366044
## INTERSTATE        -0.343801496 -0.0557047012   0.49149736 -0.048623939
## ROADWAY           -0.094740348 -0.1096411888   0.24712000  0.052923519
## ROAD_PROFILE       0.054937711  0.0418616213  -0.04309348 -0.016161260
## SPEED_LIMIT        1.000000000  0.0430607885  -0.23317177  0.352645129
## SURFACE_COND       0.043060788  1.0000000000  -0.04193318 -0.017525814
## TRAFFIC_CTRL      -0.233171771 -0.0419331801   1.00000000 -0.043744118
## TRAFFIC_WAY        0.352645129 -0.0175258137  -0.04374412  1.000000000
## VEH_INVL           0.026060790 -0.0538159239   0.14187684  0.123627257
## ADVERSE_WEATHER    0.060920635  0.4900794310  -0.04519207  0.010551756
## INJURY_CRASH      -0.013168227 -0.0392257848   0.03424708  0.003858935
## PROPERTY_DMG       0.003147664  0.0415079457  -0.03003743 -0.004061320
## REGION             0.070444717 -0.0641548852   0.02765528  0.143267126
##                     VEH_INVL ADVERSE_WEATHER   INJURY_CRASH PROPERTY_DMG
## RUSH_HR           0.07587544    0.012466214 -1.484479e-03  0.005707033
## ALCOHOL           0.12713154    0.033873582 -4.669754e-02  0.055845898
## ROAD_ALIGN       -0.20957580    0.053467381  8.785779e-06 -0.006560590
## WORK_ZONE         0.02691889   -0.021974557 -1.015394e-02  0.009370449
## WEEKDAY           0.11730994    0.003642809 -2.143795e-02  0.025007439
## INT_HWY           0.06160856    0.037461710 -1.359694e-02  0.009756132
## LIGHT_COND       -0.15677362    0.075797833 -5.460283e-03 -0.001352061
## MANCOL            0.74346423   -0.061382462 -1.414155e-02  0.027714402
## PED_CYCLIST      -0.23371986   -0.042965679  1.797258e-01 -0.198300171
## INTERSTATE        0.21466243   -0.047807333  4.198543e-02 -0.032252876
## ROADWAY           0.61064083   -0.096888703  3.773750e-02 -0.032899520
## ROAD_PROFILE     -0.07213592    0.046600128 -1.865130e-02  0.014602402
## SPEED_LIMIT       0.02606079    0.060920635 -1.316823e-02  0.003147664
## SURFACE_COND     -0.05381592    0.490079431 -3.922578e-02  0.041507946
## TRAFFIC_CTRL      0.14187684   -0.045192066  3.424708e-02 -0.030037435
## TRAFFIC_WAY       0.12362726    0.010551756  3.858935e-03 -0.004061320
## VEH_INVL          1.00000000   -0.029740391  5.123357e-02 -0.044129359
## ADVERSE_WEATHER  -0.02974039    1.000000000 -4.172276e-02  0.045867806
```

```
## INJURY_CRASH        0.05123357      -0.041722765  1.000000e+00 -0.978146283
## PROPERTY_DMG        -0.04412936       0.045867806 -9.781463e-01  1.000000000
## REGION               0.03011401      -0.079357914  4.782097e-02 -0.049356991
##                            REGION
## RUSH_HR             0.0052985899
## ALCOHOL            -0.0257752150
## ROAD_ALIGN          0.0315273363
## WORK_ZONE           0.0006819275
## WEEKDAY            -0.0052714849
## INT_HWY             0.0572764014
## LIGHT_COND         -0.0266352243
## MANCOL              0.0172455487
## PED_CYCLIST        -0.0026124016
## INTERSTATE         -0.0447731496
## ROADWAY             0.0030266483
## ROAD_PROFILE        0.0332729632
## SPEED_LIMIT         0.0704447169
## SURFACE_COND       -0.0641548852
## TRAFFIC_CTRL        0.0276552822
## TRAFFIC_WAY         0.1432671265
## VEH_INVL            0.0301140079
## ADVERSE_WEATHER    -0.0793579145
## INJURY_CRASH        0.0478209699
## PROPERTY_DMG       -0.0493569911
## REGION              1.0000000000
```

```r
corrplot(CORR_MATRIX)
```

```
## Output plot with numeric values.
corrplot(CORR_MATRIX, method ="number")
```

```r
library(rpart)
library(rpart.plot)
set.seed(123)
# 80% train; 20% test
data_train <- ACC[1:33746,]
data_test <- ACC[33747:42183,]

library(ggplot2)
library(readr)

library(clustMixType)
# apply k prototyps


x1 <- as.factor(data_train$RUSH_HR)
x2 <- as.factor(data_train$ALCOHOL)
x3 <- as.factor(data_train$ROAD_ALIGN)
x4 <- as.factor(data_train$WORK_ZONE)
x5 <- as.factor(data_train$WEEKDAY)
x6 <- as.factor(data_train$INT_HWY)
x7 <- as.factor(data_train$LIGHT_COND)
x8 <- as.factor(data_train$MANCOL)
x9 <- as.factor(data_train$PED_CYCLIST)
```
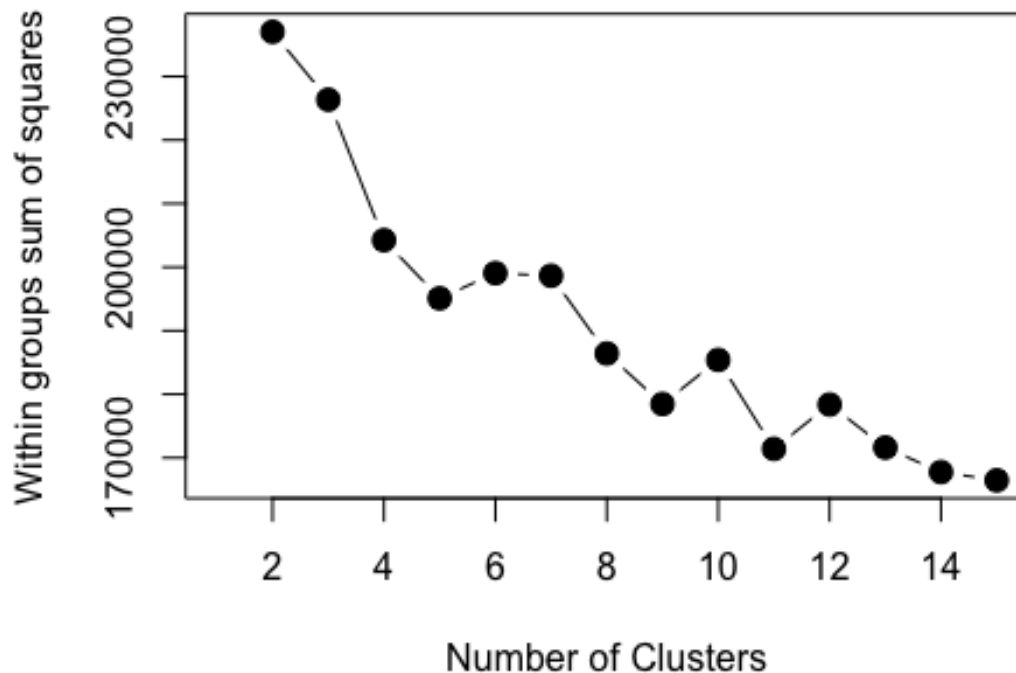
```
x10 <- as.factor(data_train$INTERSTATE)
x11 <- as.factor(data_train$ROADWAY)
x12 <- as.factor(data_train$ROAD_PROFILE)
x13 <- as.factor(data_train$SPEED_LIMIT)
x14 <- as.factor(data_train$SURFACE_COND)
x15 <- as.factor(data_train$TRAFFIC_CTRL)
x16 <- as.factor(data_train$TRAFFIC_WAY)
x17 <- data_train$VEH_INVL
x18 <- as.factor(data_train$ADVERSE_WEATHER)
x19 <- as.factor(data_train$INJURY_CRASH)
x20 <- as.factor(data_train$PROPERTY_DMG)
x21 <- as.factor(data_train$REGION)

data_frame <-
data.frame(x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,x12,x13,x14,x15,x16,x17,x18,x19
,x20,x21)


kpro_ACC <- kproto(data_frame, 7)

## Estimated lambda: 1.209163

data_train$cluster = kpro_ACC$cluster

# Check for the optimal number of clusters given the data


wss<-vector()
for (i in 2:15){ wss[i] <- sum(kproto(data_frame, i)$withinss)}

## Estimated lambda: 1.209163
##
## Estimated lambda: 1.209163
##
## Estimated lambda: 1.209163
##
## Estimated lambda: 1.209163
##
## Estimated lambda: 1.209163
##
## Estimated lambda: 1.209163
##
## Estimated lambda: 1.209163
##
## Estimated lambda: 1.209163
##
## Estimated lambda: 1.209163
##
## Estimated lambda: 1.209163
##
```

```
## Estimated lambda: 1.209163
##
## Estimated lambda: 1.209163
##
## Estimated lambda: 1.209163
##
## Estimated lambda: 1.209163

par(mfrow=c(1,1))
plot(1:15, wss, type="b", xlab="Number of Clusters",
     ylab="Within groups sum of squares",
     main="Assessing the Optimal Number of Clusters with the Elbow Method",
     pch=20, cex=2)
```
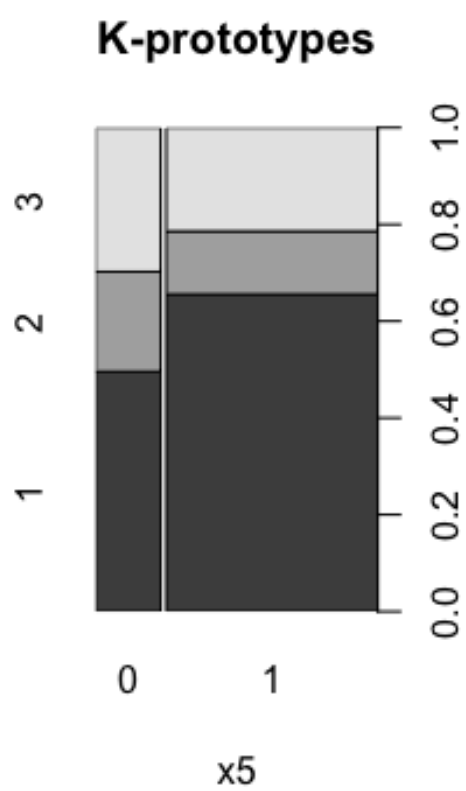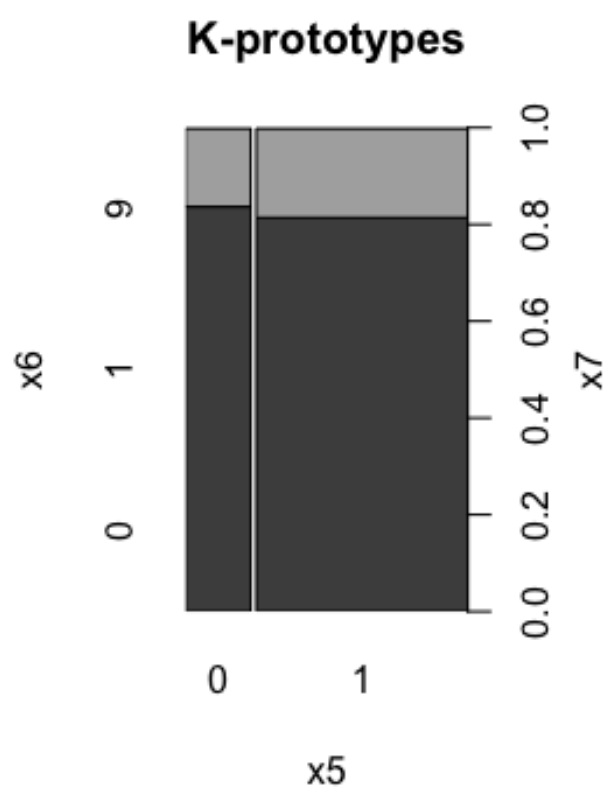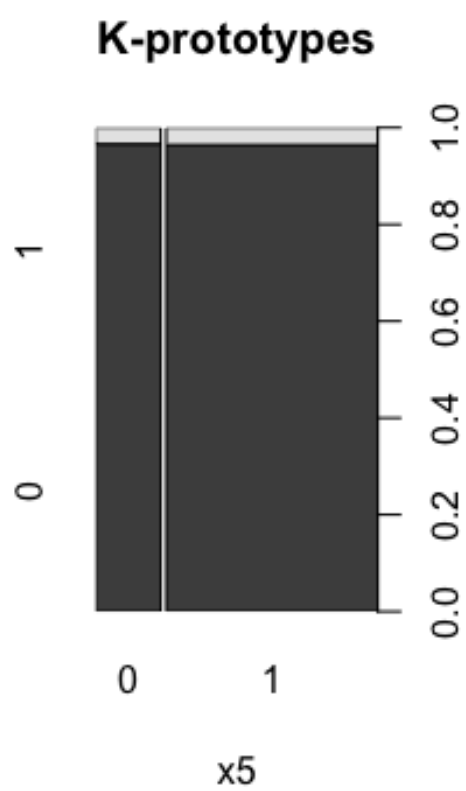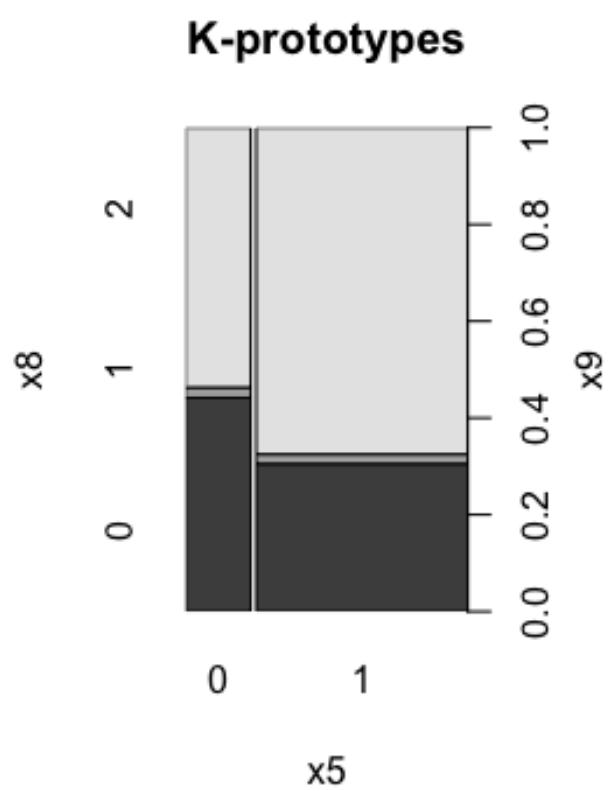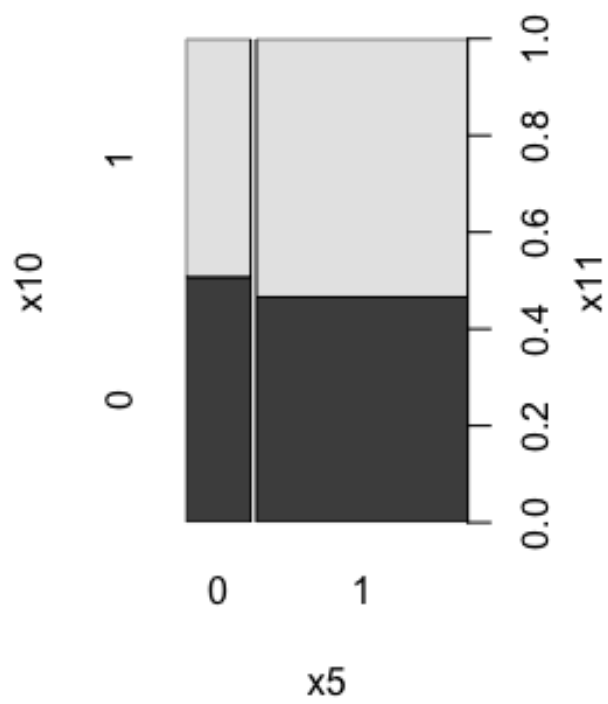


```
# plots between Total and other numerical Attributes with clusters:
par(mfrow=c(1,2))

for(i in 1: 1:6){
  plot(data_frame[,c(5,5+i)], col=data_frame$cluster, main="K-prototypes")
}
```

K-prototypes

K-prototypes