

Assignment 6: Multivariate Data Mining

Brendan Smith

April 25, 2016

Objective Statement:

Methods:

The function `is.na()` is the 'Not Available' function, which checks the dataframe to see where those elements are missing. Prepending the `!` operator to `is.na()` causes the function to return the element indices that do in fact contain the value or characters placed in the 'Not Available' function. Thus, in this instance, we search for `HGM_TYPE` and create a new dataframe that only contains the rows that contain a recorded hydrogeomorphic type.

We now add new columns to the newly created dataframe that are slightly more meaningful and easily discerned:

```
# Suggested additions
mdwhgm$area.sqkm = mdwhgm[,"Shape_Area"]/1000000 # m^2 to km^2
mdwhgm$catch.sqkm = mdwhgm[,"CATCHMENT_"]/1000000 # m^2 to km^2
mdwhgm$elev_m = mdwhgm[,"ELEV_MEAN"]
mdwhgm$elev_r = mdwhgm[,"ELEV_RANGE"]
mdwhgm$lat_dd = mdwhgm[,"LAT_DD"]
mdwhgm$lon_dd = mdwhgm[,"LONG_DD"]
mdwhgm$slope.pct = mdwhgm[,"FLOW_SLOPE"]
mdwhgm$edge.comp = mdwhgm[,"EDGE_COMPL"]
mdwhgm$clay = mdwhgm[,"ClayTot_r"]
mdwhgm$soil.kf = mdwhgm[,"Kf"]
```

```
# Additional meaningful columns
```

```
# EDA
summary(mdwhgm)
```

```
##      AREA_ACRE      STATE      ID      HUC12
## Min.      : 1.004    CA:431    UCDSNM000008: 1    180201220204: 10
## 1st Qu.: 6.037      NV: 7      UCDSNM000010: 1    180400061101: 8
## Median : 19.309                        UCDSNM000012: 1    180400100501: 8
## Mean    : 80.450                        UCDSNM000015: 1    160501010301: 7
## 3rd Qu.: 52.124                        UCDSNM000016: 1    160501010303: 6
## Max.    :4610.374                        UCDSNM000017: 1    180200030106: 6
##                                     (Other)      :432    (Other)      :393
##
##      OWNERSHIP      EDGE_COMPL
## Lassen National Forest : 60    Min.      :1.033
## Sierra National Forest : 58    1st Qu.:1.641
## Inyo National Forest   : 56    Median :2.062
## Private                 : 52    Mean    :2.340
## Stanislaus National Forest: 40    3rd Qu.:2.658
## Sequoia National Forest : 35    Max.    :9.642
## (Other)                 :137
##
##      DOM_ROCKTY      VEG_MAJORI
## granodiorite      :173    Riparian      :197
```

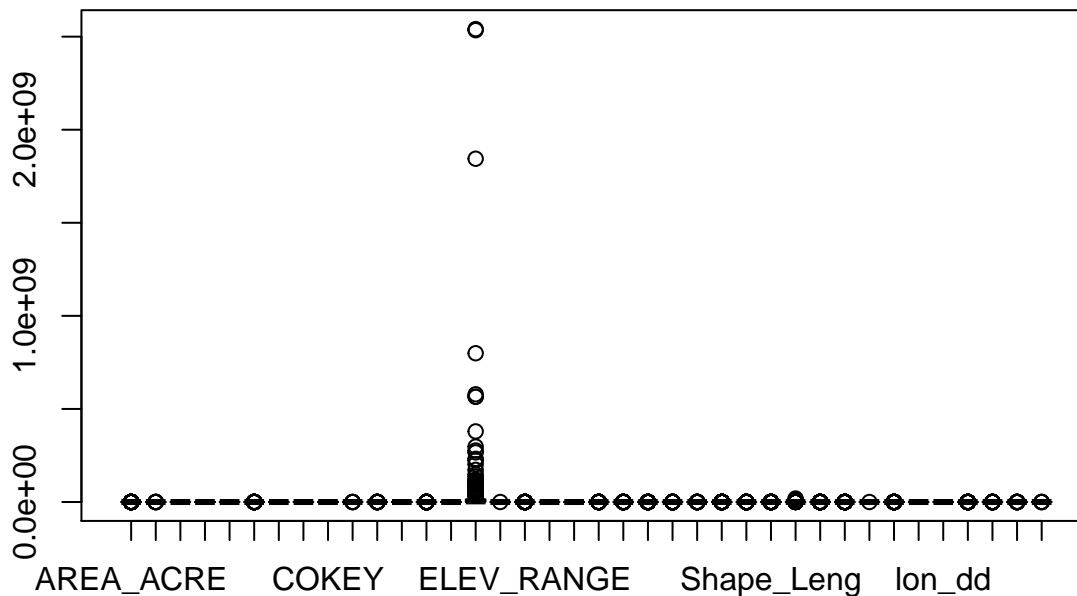
```

## andesite :154 Conifer :195
## glacial drift : 40 Shrubland : 32
## alluvium : 36 Hardwood : 9
## tephrite (basanite): 6 Barren-Rock/Sand/Clay: 2
## argillite : 5 Hardwood-Conifer : 1
## (Other) : 24 (Other) : 2
## COKEY Kf ClayTot_r MUKEY
## 470977:660084 : 15 Min. :0.0000 Min. : 1.00 470977 : 15
## 465178:642932 : 14 1st Qu.:0.2000 1st Qu.: 6.00 465178 : 14
## 464853:642321 : 12 Median :0.2400 Median :12.00 464853 : 12
## 1652104:1207250: 11 Mean :0.2718 Mean :12.06 1652104: 11
## 464983:642549 : 11 3rd Qu.:0.3200 3rd Qu.:15.00 464983 : 11
## 471192:666181 : 10 Max. :0.5500 Max. :50.00 471192 : 10
## (Other) :365 (Other):365
## SOIL_SURVE COMP_NAME CATCHMENT_ ELEV_MEAN
## SSURGO :379 Aquolls : 23 Min. :1.263e+03 Min. : 742.3
## STATSGO: 59 Monache variant: 21 1st Qu.:5.670e+05 1st Qu.:1728.9
## Cagwin family : 15 Median :3.350e+06 Median :2024.5
## Toem : 13 Mean :3.732e+07 Mean :2072.1
## AQUEPTS : 12 3rd Qu.:1.358e+07 3rd Qu.:2366.4
## Tahoe : 12 Max. :2.540e+09 Max. :3266.4
## (Other) :342
## ELEV_RANGE LAT_DD LONG_DD FLOW_RANGE
## Min. : 0.4037 Min. :35.45 Min. : -121.6 Min. : 42.43
## 1st Qu.: 9.7699 1st Qu.:37.45 1st Qu.: -120.6 1st Qu.: 1388.75
## Median : 19.9371 Median :38.78 Median : -120.1 Median : 3413.27
## Mean : 33.2681 Mean :38.77 Mean : -119.9 Mean : 7160.09
## 3rd Qu.: 36.6473 3rd Qu.:40.23 3rd Qu.: -119.1 3rd Qu.: 7277.69
## Max. :359.3870 Max. :41.98 Max. : -118.1 Max. :170870.00
##
## FLOW_SLOPE ED_MIN_LAK ED_MIN_FLO ED_MIN_SEE
## Min. :1.354e-05 Min. : 0 Min. : 0.0 Min. : 0.0
## 1st Qu.:2.870e-03 1st Qu.: 1553 1st Qu.: 0.0 1st Qu.: 642.6
## Median :7.199e-03 Median : 3535 Median : 0.0 Median : 2133.9
## Mean :1.278e-02 Mean : 5514 Mean : 928.9 Mean : 2990.9
## 3rd Qu.:1.624e-02 3rd Qu.: 7190 3rd Qu.: 311.7 3rd Qu.: 4430.1
## Max. :1.456e-01 Max. :32386 Max. :29463.1 Max. :15875.4
##
## HGM_TYPE ED_MIN_FSt Shape_Leng
## Riparian low gradient :181 Min. : 0.00 Min. : 242.4
## Riparian middle gradient : 72 1st Qu.: 0.00 1st Qu.: 991.6
## Subsurface low gradient : 51 Median : 0.00 Median : 1947.2
## Subsurface middle gradient: 35 Mean : 196.42 Mean : 4461.2
## Discharge slope : 24 3rd Qu.: 31.62 3rd Qu.: 4159.1
## Depressional perennial : 19 Max. :15389.20 Max. :147644.1
## (Other) : 56
## Shape_Area area.sqkm catch.sqkm
## Min. : 4063 Min. : 0.004063 Min. : 0.0013
## 1st Qu.: 24432 1st Qu.: 0.024432 1st Qu.: 0.5670
## Median : 78142 Median : 0.078142 Median : 3.3498
## Mean : 325573 Mean : 0.325573 Mean : 37.3219
## 3rd Qu.: 210937 3rd Qu.: 0.210937 3rd Qu.: 13.5770
## Max. :18657598 Max. :18.657598 Max. :2540.4858
##

```

```
##      elev_m      elev_r      lat_dd      lon_dd
## Min.   : 742.3   Min.   : 0.4037   Min.   :35.45   Min.   : -121.6
## 1st Qu.:1728.9   1st Qu.: 9.7699   1st Qu.:37.45   1st Qu.: -120.6
## Median :2024.5   Median : 19.9371   Median :38.78   Median : -120.1
## Mean   :2072.1   Mean   : 33.2681   Mean   :38.77   Mean   : -119.9
## 3rd Qu.:2366.4   3rd Qu.: 36.6473   3rd Qu.:40.23   3rd Qu.: -119.1
## Max.   :3266.4   Max.   :359.3870   Max.   :41.98   Max.   : -118.1
##
##      slope.pct      edge.comp      clay      soil.kf
## Min.   :1.354e-05   Min.   :1.033   Min.   : 1.00   Min.   :0.0000
## 1st Qu.:2.870e-03   1st Qu.:1.641   1st Qu.: 6.00   1st Qu.:0.2000
## Median :7.199e-03   Median :2.062   Median :12.00   Median :0.2400
## Mean   :1.278e-02   Mean   :2.340   Mean   :12.06   Mean   :0.2718
## 3rd Qu.:1.624e-02   3rd Qu.:2.658   3rd Qu.:15.00   3rd Qu.:0.3200
## Max.   :1.456e-01   Max.   :9.642   Max.   :50.00   Max.   :0.5500
##
```

```
boxplot(mdwghgm)
```

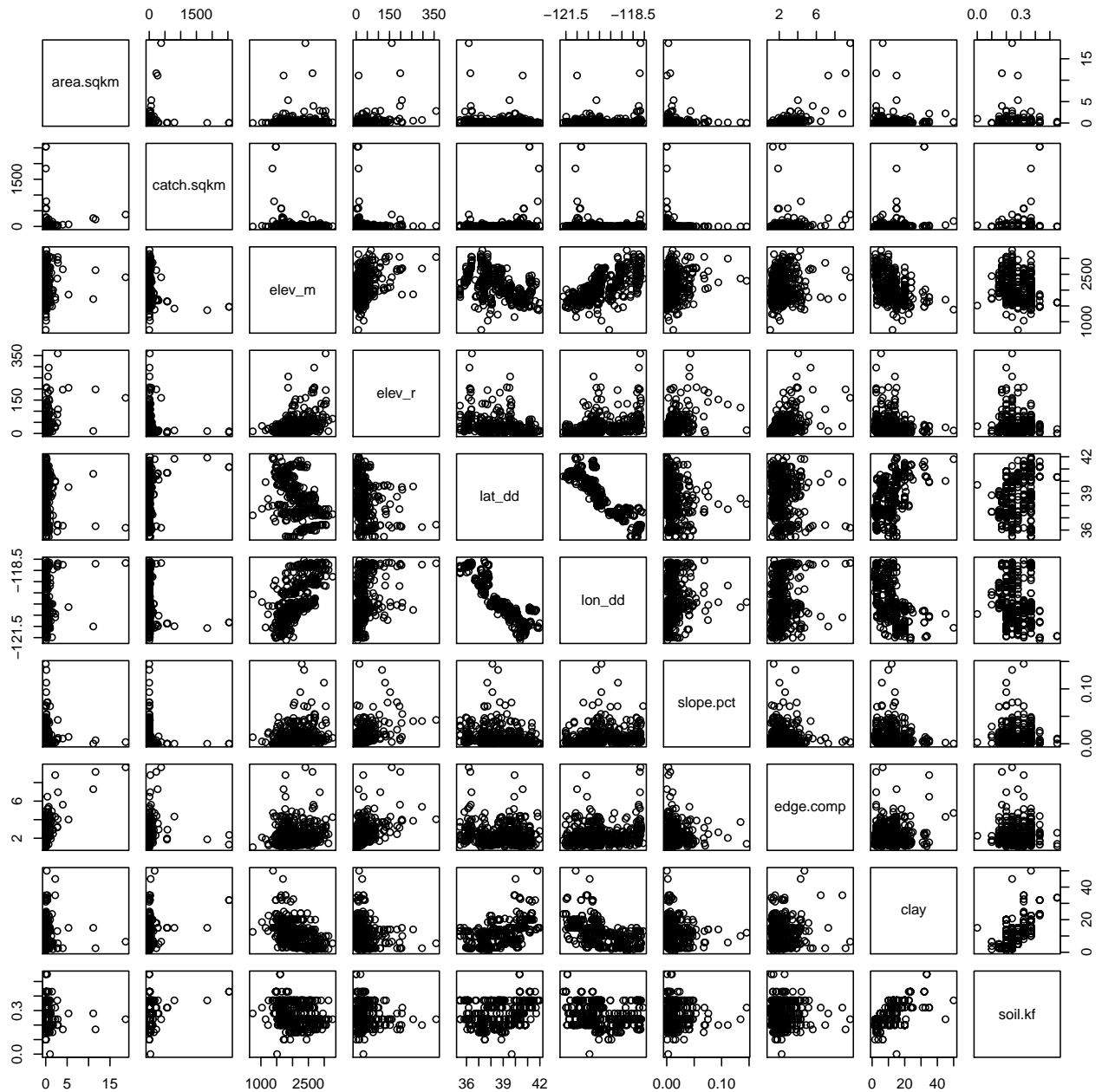


```
#Optional method for keeping track of the relevant variables
```

```
rel_cols = c("area.sqkm", "catch.sqkm", "elev_m", "elev_r", "lat_dd", "lon_dd", "slope.pct", "edge.comp")
```

```
rmdwghgm <-mdwghgm[,rel_cols]
```

```
plot(rmdwghgm)
```



Most Variability: Elev Mean & Lat Soil Lat Lon Elev mean

Correlated: Soil & Clay Lat & Clay Lat & Elev Mean elev mean & Lon Edge & Elev Range Elev Mean & Lon Elev Mean & Clay

Step 2 - Clustering and Clustering Output

```
# Heirarchical Clustering
#dist using euclidean
plot.new()
rmdwhgm.dist<- dist(x = rmdwhgm[,rel_cols],method = "euclidean") #hclust using ward.D
rmdwhgm.hc<- hclust(rmdwhgm.dist,method="ward.D")
rect.hclust(rmdwhgm.hc,k=6)
```

```

# k-means Clustering
rmdwhgm$hc6 <- cutree(rmdwhgm.hc, k=6) #store group # in hc6
rmdwhgm.km6 <- kmeans(rmdwhgm[,rel_cols],centers = 6)
rmdwhgm$km6 <- rmdwhgm.km6$cluster #store group # in km6
table(rmdwhgm$hc6, rmdwhgm$km6)

```

```

##
##      1  2  3  4  5  6
##  1 103  0  0  0  2  0
##  2  12  0  0 60  0  0
##  3   0  0  0  0 69  0
##  4   0 23  0  0  4 72
##  5   0  0  0 90  0  0
##  6   0  0  3  0  0  0

```

```

# Load the DEM
gdal_grid = readGDAL("DEM.tif")

```

```

## DEM.tif has GDAL driver GTiff
## and has 1137 rows and 1233 columns

```

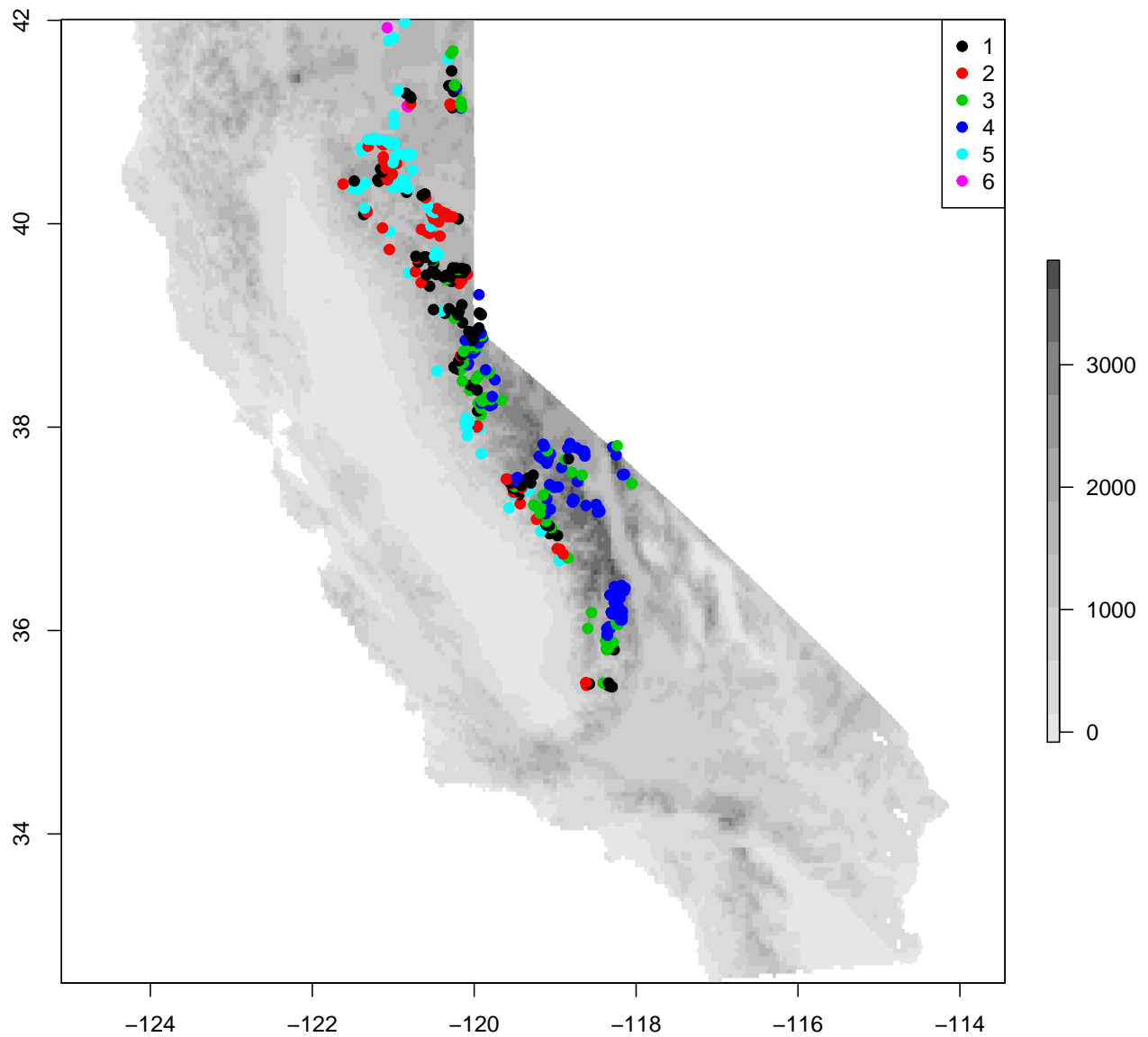
```

dem = raster(gdal_grid) #use data as a projected raster
plot(dem,col=gray.colors(10, start=0.9, end=0.3))

# Create a vector to aid in plotting text for ProjLoc$ProjCode
xtext = rmdwhgm$lon_dd
ytext = rmdwhgm$lat_dd

# Plot the ProjLoc over the DEM
points(rmdwhgm$lon_dd,rmdwhgm$lat_dd,pch=19,col=rmdwhgm$hc6)
legend("topright",legend=levels(as.factor(rmdwhgm$hc6)),col=1:length(rmdwhgm$hc6),pch=19)

```



```
plot(dem,col=gray.colors(10, start=0.9, end=0.3))

# Create a vector to aid in plotting text for ProjLoc$ProjCode
xtext = rmdwhgm$lon_dd
ytext = rmdwhgm$lat_dd

# Plot the ProjLoc over the DEM
points(rmdwhgm$lon_dd,rmdwhgm$lat_dd,pch=19,col=rmdwhgm$km6)
legend("topright",legend=levels(as.factor(rmdwhgm$km6)),col=1:length(rmdwhgm$km6),pch=19)
```

