Problem set 3

Rui He

10/25/2019

Data Munging

1 & 2. Data loaded and munged.

```
library('tidyverse')
library('skimr')
library('seriation')

load("~/Documents/UML/legprof-components.v1.0.RData")
x_sub <- x %>%
  dplyr::select(stateabv,t_slength,sessid,slength,salary_real,expend) %>%
  filter(sessid=='2009/10')%>%
  drop_na()

x_con <-subset(x_sub,select =c(t_slength,slength,salary_real,expend))
x_scale <-scale(x_con)
state <- subset(x_sub,select = c(stateabv))
```

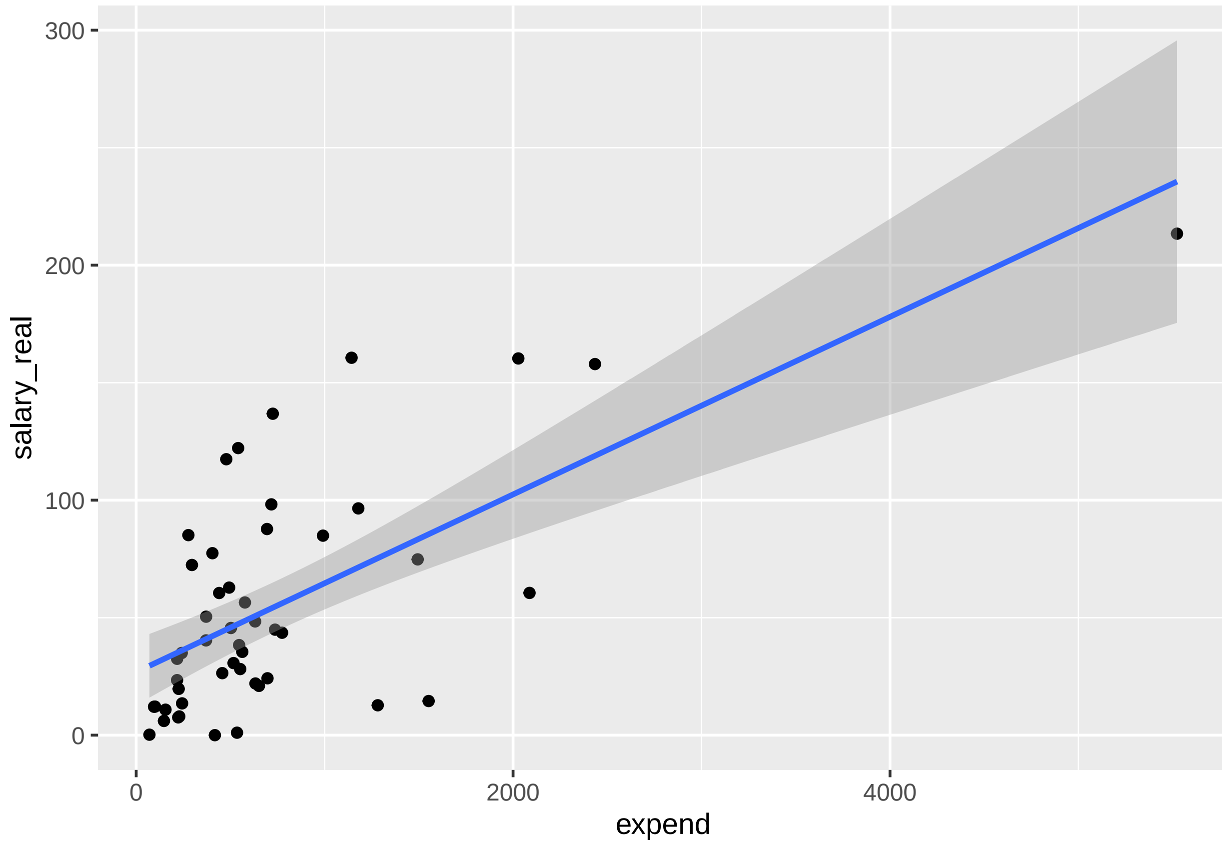
EDA

3 Mean and distribution of four variables are calculated. Cluster plot shows that there are correlations between the expenditure and salary, and between total session length and length. A couple outliers are salient in this case, for example, new york and california are higher in expenditure, salary and session length.

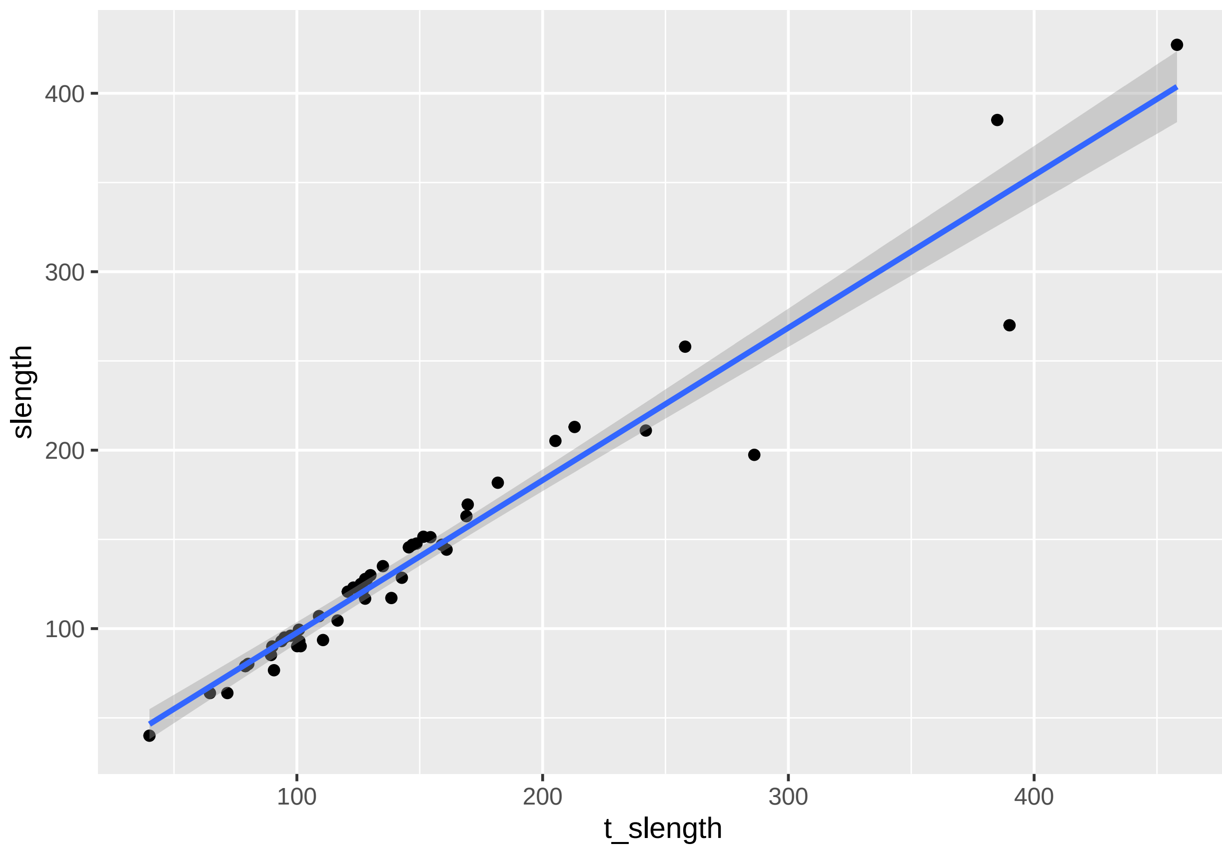
```
summary(x_con)
```

```
##    t_slength         slength        salary_real         expend
##  Min.   : 40.00   Min.   : 40.0   Min.   :  0.00   Min.   :  70.43
##  1st Qu.: 97.42   1st Qu.: 93.0   1st Qu.: 19.69   1st Qu.: 277.08
##  Median :127.77   Median :123.0   Median : 40.33   Median : 535.14
##  Mean   :147.80   Mean   :138.5   Mean   : 54.99   Mean   : 744.47
##  3rd Qu.:159.00   3rd Qu.:151.2   3rd Qu.: 77.43   3rd Qu.: 724.91
##  Max.   :458.15   Max.   :427.1   Max.   :213.41   Max.   :5523.10
```

```
ggplot(x_con,aes(x=expend,y=salary_real))+
  geom_point()+
  geom_smooth(method=lm)
```



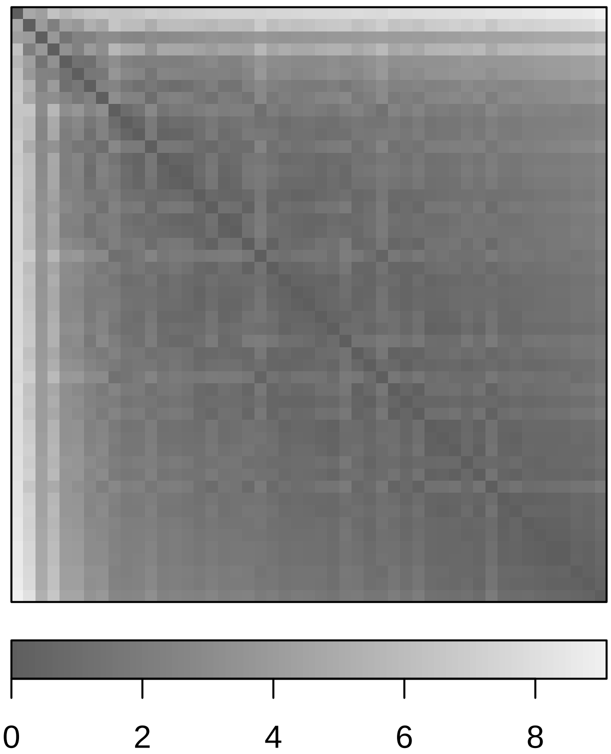

```r
ggplot(x_con,aes(x=t_slength,y=slength))+
  geom_point()+
  geom_smooth(method=lm)
```



Diagnosing clusterability

4. Using ODI shows that the data may be cluster as seen in the upper left, there is a small lighter square. In the lower right, the square is darker, suggesting there may be some clusterability in the observations.

```
x_dist <- dist(x_scale,method = "euclidean")
dissplot(x_dist)
```



K-mean

5. k-mean: The algorithm was initiated at k=2 and there are two clusters. The summary of centers showed the separation of two clusters and cluster 1 are higher in session length, expenditure, and salary. We can further investigate the states in the first cluster as displayed below.

```
library(cluster)
# fit k-mean, k = 2
set.seed(335)
kmeans <- kmeans(x_scale, center= 2,nstart = 15)
x_con$Cluster_km <- as.factor(kmeans$cluster)
kmeans$cluster
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
##  2  2  2  2  1  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  1  1  2  2  2
## 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 50
##  2  2  2  2  2  2  1  2  2  1  2  2  1  2  2  2  2  2  2  2  2  2  2  2
```

```
kmeans$centers
```

```
##    t_slength    slength salary_real     expend
## 1  2.1000302  2.1014710   2.0307585  1.4677087
## 2 -0.2930275 -0.2932285  -0.2833616 -0.2047966
```

```r
kmeans$size
```

```
## [1]  6 43
```

```r
# find states within cluster 1
state[which(x_con$Cluster_km==1),]
```

```
## [1] "CA" "MA" "MI" "NY" "OH" "PA"
```

GMM Model

6. The Gaussian Mixture models shows two distributions one centered around ~0 with narrower variance whereas one with much wider variance.

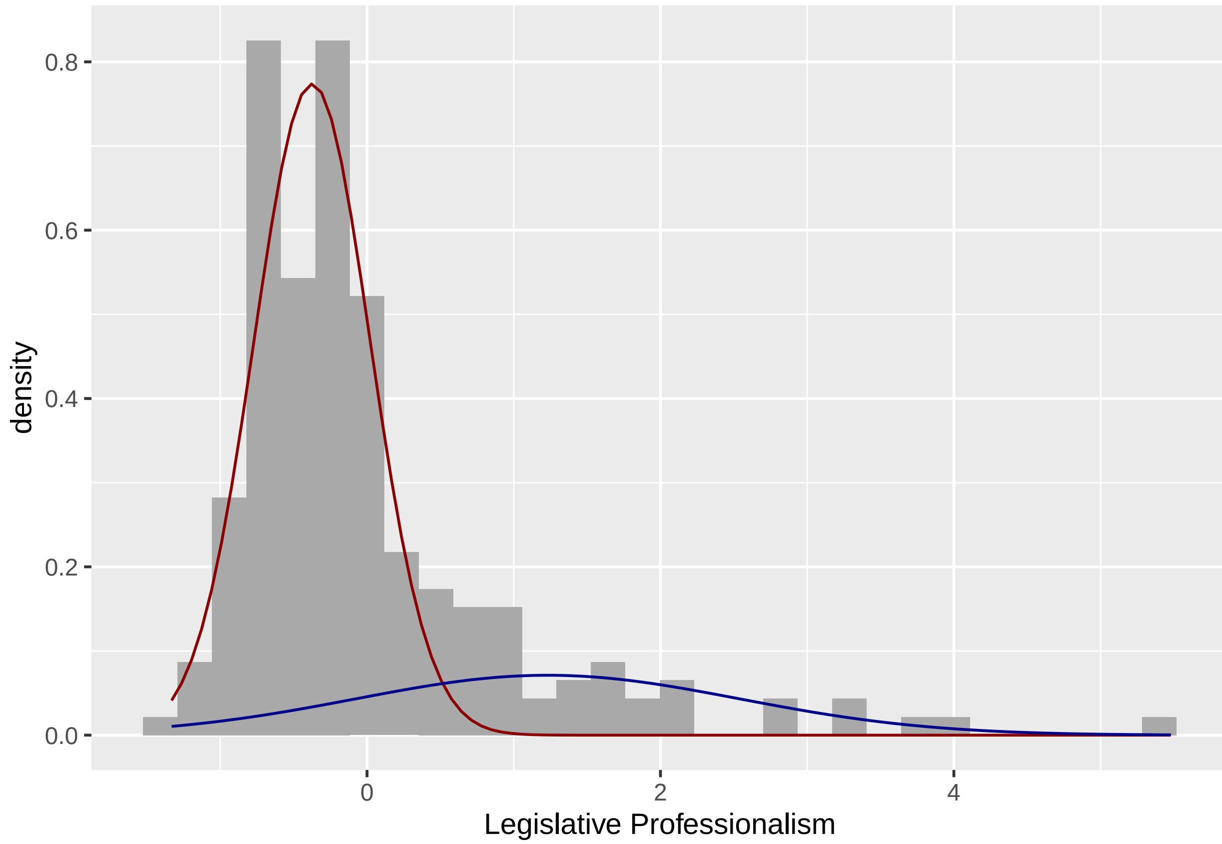
```r
library(mixtools) # fitting GMMs via EM
library(plotGMM) # customizing GMM plot

gmm <- normalmixEM(x_scale, k = 2)
summary(gmm)

posterior <- data.frame(cbind(gmm$x, gmm$posterior))
t_gmm<-data.frame(as.table(ifelse(posterior$comp.1>0.3,1,2)))
colnames(t_gmm)[colnames(t_gmm)=='Freq'] <- 'cluster'

state_ind <- which(t_gmm$cluster==1)
x_con$cluster_gmm <- t_gmm$cluster[1:49]
state[state_ind[1:9],]

ggplot(data.frame(x = gmm$x)) +
  geom_histogram(aes(x, ..density..), fill = "darkgray") +
  stat_function(geom = "line", fun = plot_mix_comps,
                args = list(gmm$mu[1], gmm$sigma[1], lam = gmm$lambda[1]),
                colour = "darkred") +
  stat_function(geom = "line", fun = plot_mix_comps,
                args = list(gmm$mu[2], gmm$sigma[2], lam = gmm$lambda[2]),
                colour = "darkblue") +
  xlab("Legislative Professionalism")
```




```
  ylab("Density") +
  theme_bw()
```

PAM

7. For the additional method, I used parition around medoids (PAM), which instead using the mediud instead of centroid in K-means.

```
pam <- pam(x_scale, 2)
x_con$Cluster_pam <- pam$clustering
state[which(pam$clustering==2),]
```

```
## [1] "CA" "IL" "MA" "MI" "NY" "OH" "PA"
```

8. Visualization of results

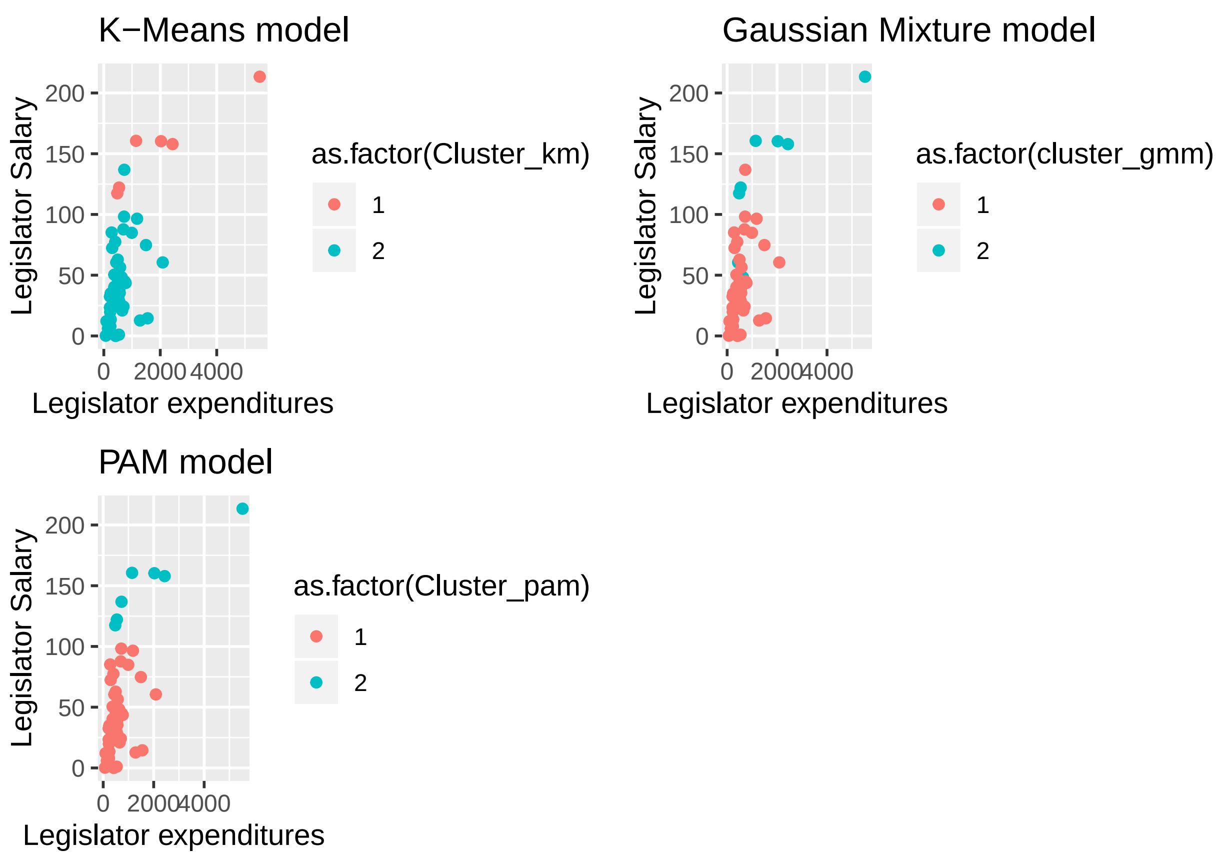
```
kmeans_plot <-x_con %>%
  ggplot(aes(x=expend, y = salary_real, color = as.factor(Cluster_km))) +
  geom_point() +
  labs(x = 'Legislator expenditures',
      y = 'Legislator Salary',
      title = 'K-Means model')
```

```
  theme_bw()

gmm_plot <-x_con %>%
  ggplot(aes(x=expend, y = salary_real, color = as.factor(cluster_gmm))) +
  geom_point() +
  labs(x = 'Legislator expenditures',
      y = 'Legislator Salary',
      title = 'Gaussian Mixture model')
  theme_bw()

pam_plot <-x_con %>%
  ggplot(aes(x=expend, y = salary_real, color = as.factor(Cluster_pam))) +
  geom_point() +
  labs(x = 'Legislator expenditures',
      y = 'Legislator Salary',
      title = 'PAM model')
  theme_bw()

  library(gridExtra)
grid.arrange(kmeans_plot,gmm_plot,pam_plot,ncol=2)
```



9. Validation of methods

```r
library(clValid)
  internal <- clValid(x_scale, 2:10,
                     clMethods = c("kmeans","model","pam"),
                     validation = "internal"); summary(internal)
```

10. Disucssion of validation output

Validation measures looked at the connectivity, Dunn's index and average silhouette width. Based on the result, it seems that GMM yield highest number of connectivity when k = 2. For K-means, k=2 or 3 generate a similar average silhouette width but the dunn's index is higher for k = 3, which may suggests that when k = 3 the result may be optimal. Whereas for GMM and PAM, k = 2 has higher dunn's index and silhouette width, suggesting that the two-cluster configuration is more valid with these two methods.

It may sugguest that the GMM model serve as a optimal approach for this data set according to the connectivity. The sub-optimal partitioning method gives comparable results with the other methods, and in this case of looking at legislator professionalism, some states are consistently clustered together (CA, NY, NJ, MI), showing a coherent result from the three methods.