Country Aid Mini Project

Rory Quinlan

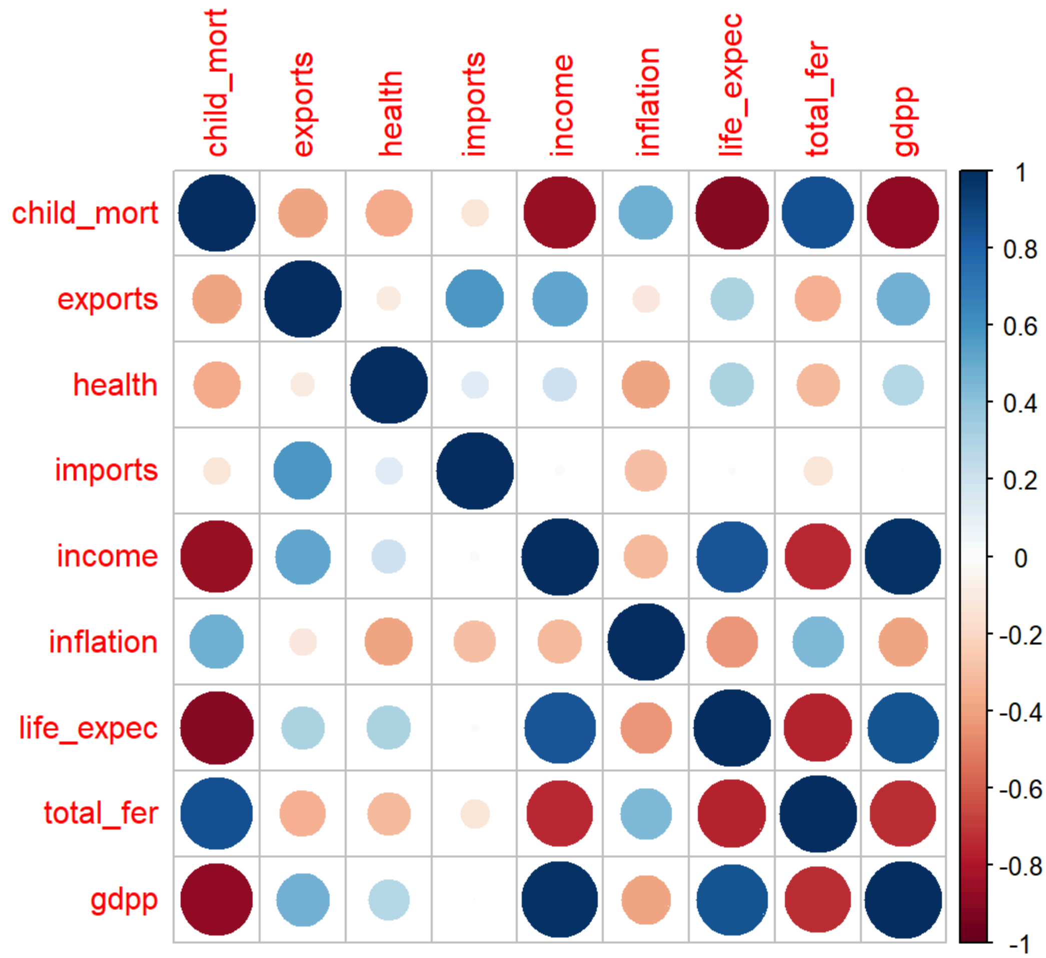
Setup

```
library(tidyverse)
library(factoextra)
library(cluster)
library(gridExtra)
library(kableExtra)
library(corrplot)
```

```
country = read_csv("Country-data.csv",show_col_types = FALSE)
```

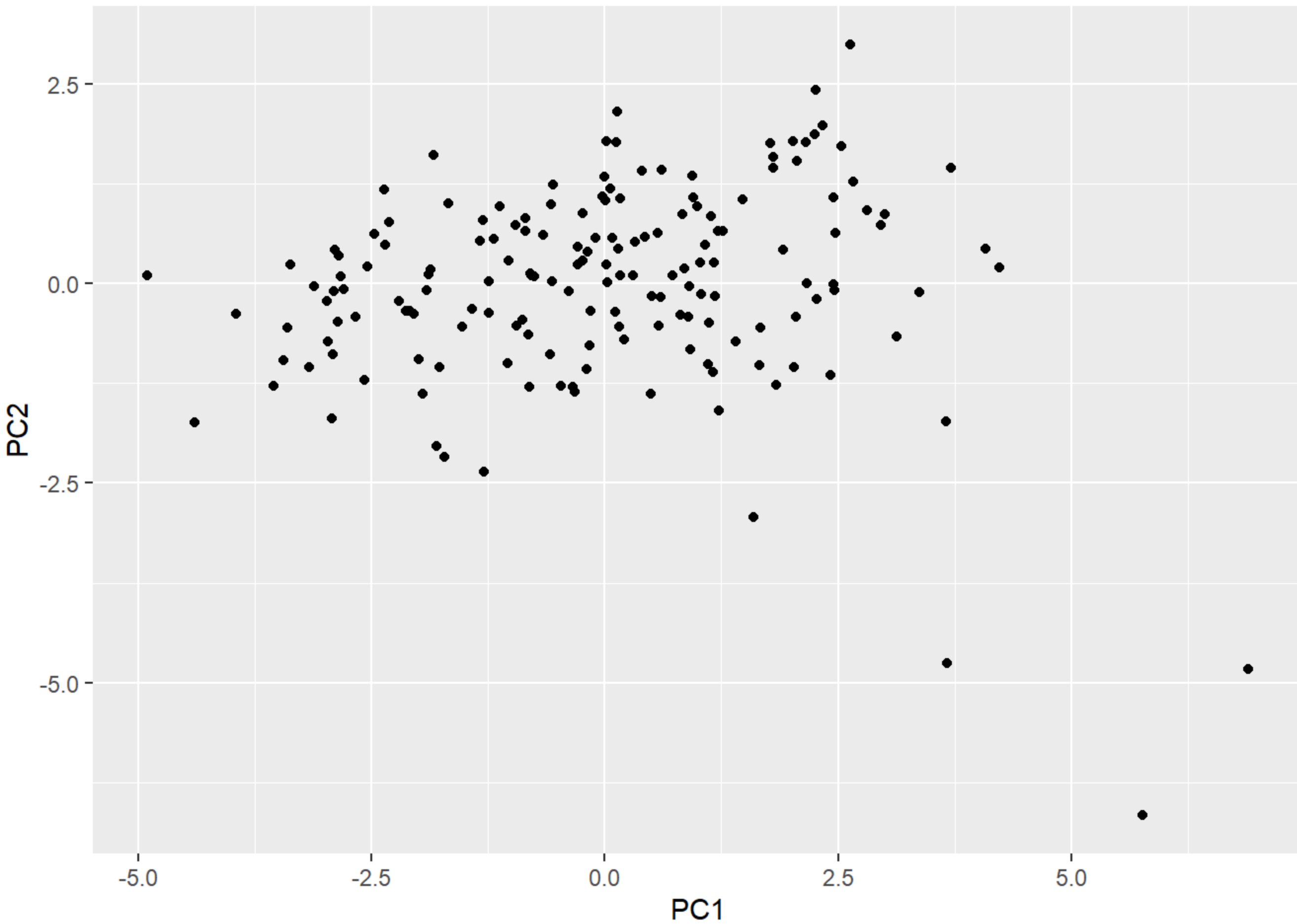
Data Exploration

```
key_predictors<- country %>% select("child_mort","exports", "health","imports","income","inflati
on", "life_expec", "total_fer", "gdpp")
cor<- cor(key_predictors,method = c("spearman"))

corrplot(cor)
```



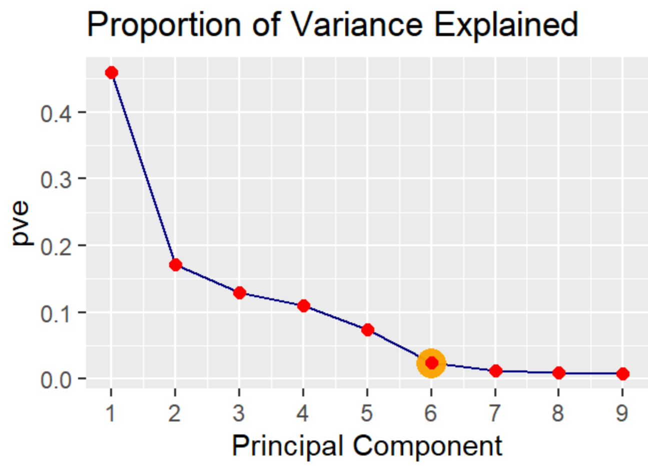
Visualize with PCA

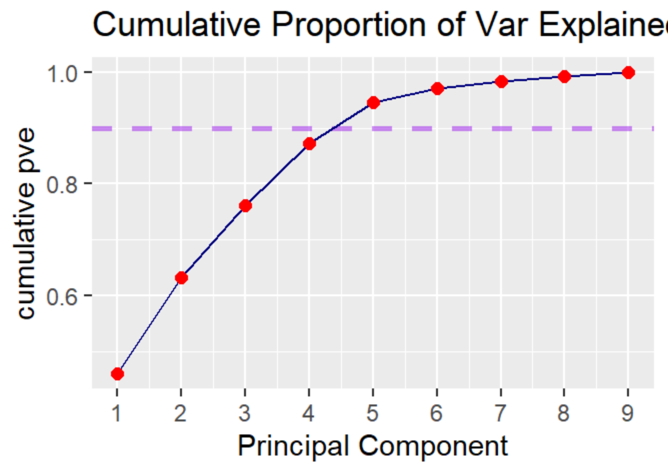
```
# Visualize with PCA
country_pc = country %>%
  select(-country) %>%
  prcomp(scale=TRUE)

country_pc$x %>%
  as_tibble() %>%
  select(PC1, PC2) %>%
  ggplot(aes(x= PC1, y=PC2)) +
  geom_point()
```

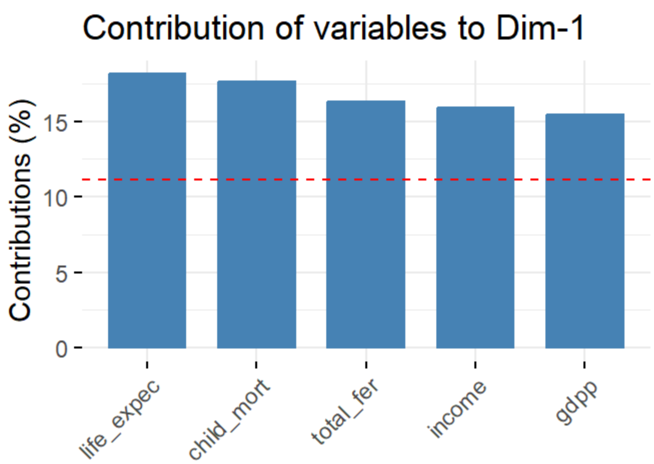


2.5
0.0
-2.5
-5.0
PC2
-5.0
-2.5
0.0
2.5
5.0
PC1

```
# How many Components should we have?
PRVar<- country_pc$sdev^2
PVE<- PRVar[1:9]/sum(PRVar)

PC=1:9
data=data.frame(PC, PVE)

p1<-ggplot(data=data, aes(x=PC, y=PVE))+
  geom_line(color="navy")+
  geom_point(aes(x=6,y=0.023127004),cex=5,color="orange",alpha=0.3)+
  geom_point(color="red",cex=2)+
  labs(title="Proportion of Variance Explained", x="Principal Component",y="pve")+
  scale_x_continuous(breaks = 1:9)

p2<-ggplot(data=data, aes(x=PC, y=cumsum(PVE)))+
  geom_hline(aes(yintercept=0.9),lty=2,color="purple",linewidth=1, alpha=0.5)+
  geom_line(color="navy")+
  geom_point(color="red",cex=2)+
  labs(title="Cumulative Proportion of Var Explained",
       x="Principal Component",
       y="cumulative pve")+
  scale_x_continuous(breaks = 1:9)

p3<-fviz_contrib(country_pc, choice = "var", axes = 1, top = 5)

grid.arrange(p1, p2,p3, ncol = 2)
```

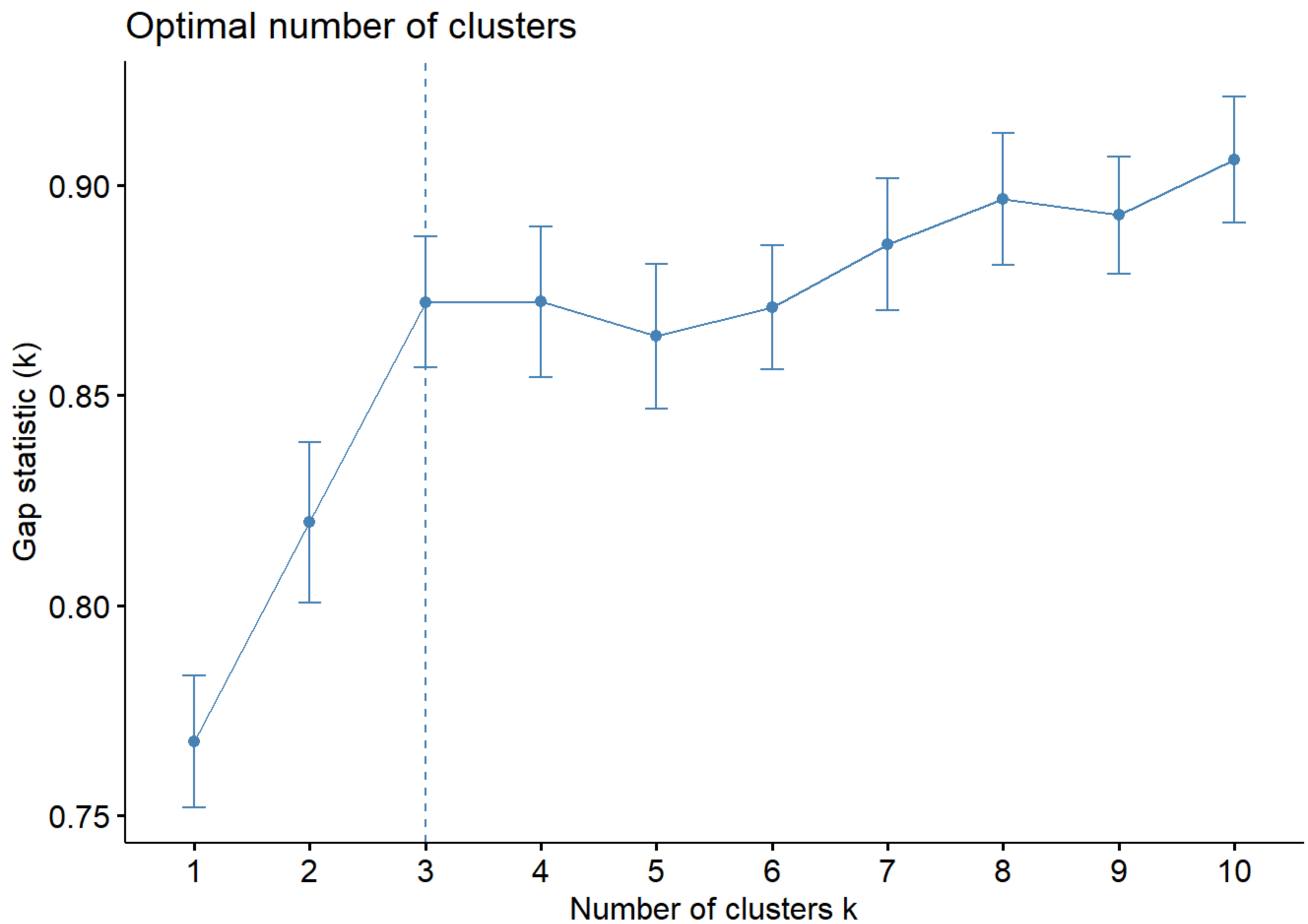






Partitioning Data

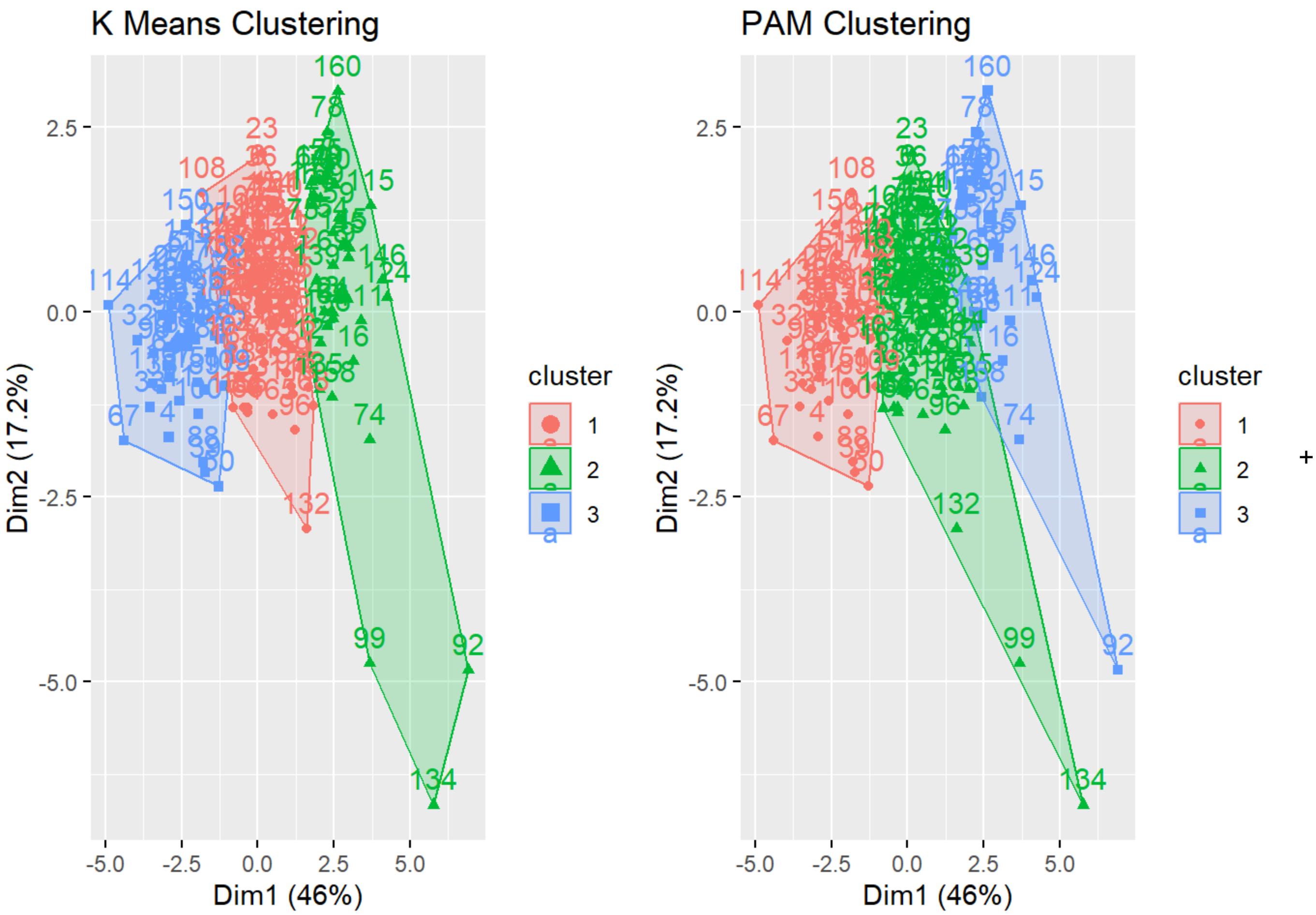
```
# Scale data
country_s = scale(country[,-1])
# Graph to find the optimal number of clusters
fviz_nbclust(country_s, kmeans, method="gap_stat")
```



Pam or Kmeans

```
# Preform clustering on data
km_mod = kmeans(country_s, centers=3)

# Preform clustering on the data
pam_mod = pam(country_s, 3)
```

```
# Scatter plot with clusters
p1<-fviz_cluster(km_mod, data = country_s,title = "K Means Clustering")

# Scatter plot with clusters
p2<-fviz_cluster(pam_mod, data = country_s,title = "PAM Clustering")

grid.arrange(p1, p2, ncol = 2)
```



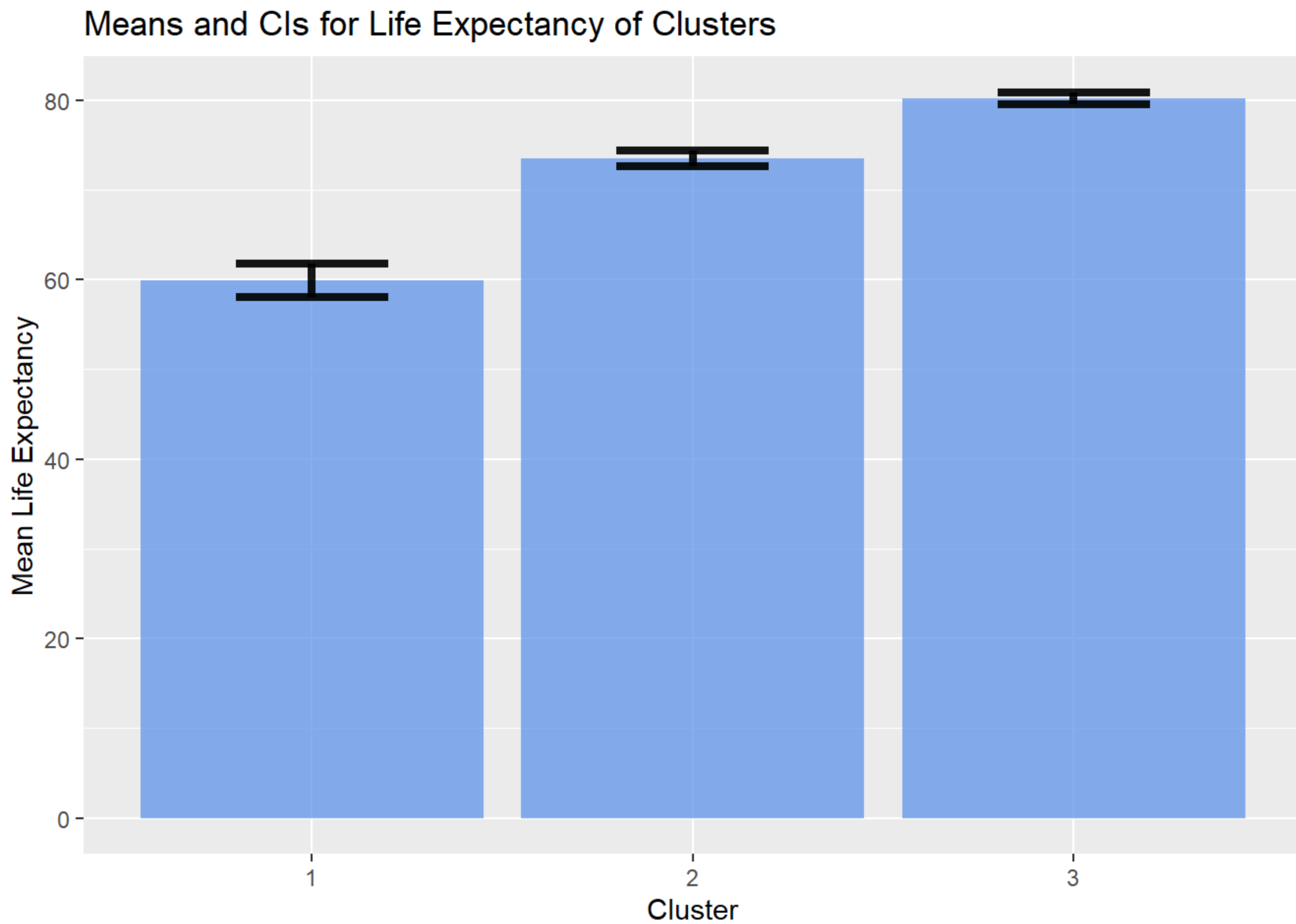
From scree plot it appears 3 is the optimal number of clusters + It appears that PAM will be better, as it is more resistent to outliers in the dataset

Data Visualization with Clusters

```
# Add column of clusters by factor
country_pam = country %>% mutate(cluster=factor(pam_mod$cluster))
```

```
p1<-ggplot(country_pam,aes(x=cluster,y=income,color=cluster)) +
  geom_boxplot()+ labs(title="Child Mortality by Cluster",y="Child Mortality (per 1000 births)")

# Boxplot
p2<-ggplot(country_pam,aes(x=cluster,y=income,color=cluster)) +
  geom_boxplot()+ labs(title="Income Summary by Cluster",y="Income (USD $)")

grid.arrange(p1, p2, ncol = 2)
```


```
# See if clusters have significant difference in life expectancy
my_sum<-country_pam %>% group_by(cluster) %>% summarise( n=n(),
    mean=mean(life_expec),
    sd=sd(life_expec)
  ) %>%
  mutate( se=sd/sqrt(n))  %>%
  mutate( ic=se * qt((1-0.05)/2 + .5, n-1))


ggplot(my_sum) +
  geom_bar( aes(x=factor(cluster), y=mean), stat="identity", fill="cornflowerblue", alpha=0.75)
+
  geom_errorbar( aes(x=cluster, ymin=mean-ic, ymax=mean+ic), width=0.4, colour="black", alpha=0.
9, size=1.5) +
  ggtitle("Means and CIs for Life Expectancy of Clusters")+ labs(y="Mean Life Expectancy",x="Clu
ster")
```



Selection

```
# All countries in cluster 1
df<-as.data.frame(country_pam[country_pam$cluster==1,]) %>% arrange(life_expec)

df$country
```

```
##  [1] "Haiti"                    "Lesotho"                 
##  [3] "Central African Republic" "Zambia"                  
##  [5] "Malawi"                   "South Africa"            
##  [7] "Mozambique"               "Sierra Leone"            
##  [9] "Guinea-Bissau"            "Afghanistan"             
## [11] "Cote d'Ivoire"            "Chad"                    
## [13] "Uganda"                   "Botswana"                
## [15] "Cameroon"                 "Congo, Dem. Rep."        
## [17] "Burundi"                  "Burkina Faso"            
## [19] "Guinea"                   "Namibia"                 
## [21] "Togo"                     "Niger"                   
## [23] "Tanzania"                 "Mali"                    
## [25] "Angola"                   "Congo, Rep."             
## [27] "Nigeria"                  "Kiribati"                
## [29] "Liberia"                  "Madagascar"              
## [31] "Equatorial Guinea"        "Eritrea"                 
## [33] "Benin"                    "Ghana"                   
## [35] "Kenya"                    "Gabon"                   
## [37] "Lao"                      "Senegal"                 
## [39] "Rwanda"                   "Pakistan"                
## [41] "Gambia"                   "Comoros"                 
## [43] "India"                    "Sudan"                   
## [45] "Myanmar"                  "Iraq"                    
## [47] "Yemen"                    "Mauritania"              
## [49] "Nepal"                    "Tajikistan"              
## [51] "Timor-Leste"
```