Project 2

Yue Taira, Tushar Kohli, Raju Kakarlapudi

Data

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.6     ✓ dplyr   1.0.8
## ✓ tidyr   1.2.0     ✓ stringr 1.4.0
## ✓ readr   2.1.2     ✓ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
## Loading required package: airports
```

```
## Loading required package: cherryblossom
```

```
## Loading required package: usdata
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WB
a
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
## # A tibble: 6 × 11
##   country           area birth_rate death_rate infant_mortality_… internet_users
##   <fct>            <dbl>      <dbl>      <dbl>              <dbl>          <dbl>
## 1 Russia        17098242       11.9       13.8               7.08       40853000
## 2 Canada         9984670       10.3        8.31              4.71       26960000
## 3 United States  9826675       13.4        8.15              6.17      245000000
## 4 China          9596960       12.2        7.44             14.8       389000000
## 5 Brazil         8514877       14.7        6.54             19.2        75982000
## 6 Australia      7741220       12.2        7.07              4.43       15810000
## # … with 5 more variables: life_exp_at_birth <dbl>,
## #   maternal_mortality_rate <int>, net_migration_rate <dbl>, population <int>,
## #   population_growth_rate <dbl>
```

```
## [1] 70.65203
```

```
## # A tibble: 177 × 12
##    country            area birth_rate death_rate infant_mortalit… internet_users
##    <fct>             <dbl>      <dbl>      <dbl>             <dbl>          <dbl>
##  1 Russia           1.71e7       11.9       13.8              7.08       40853000
##  2 Canada           9.98e6       10.3        8.31             4.71       26960000
##  3 United States    9.83e6       13.4        8.15             6.17      245000000
##  4 China            9.60e6       12.2        7.44            14.8       389000000
##  5 Brazil           8.51e6       14.7        6.54            19.2        75982000
##  6 Australia        7.74e6       12.2        7.07             4.43       15810000
##  7 India            3.29e6       19.9        7.35            43.2        61338000
##  8 Argentina        2.78e6       16.9        7.34             9.96       13694000
##  9 Kazakhstan       2.72e6       19.6        8.31            21.6         5299000
## 10 Congo, Democrat… 2.34e6       35.6       10.3             73.2          290000
## # … with 167 more rows, and 6 more variables: life_exp_at_birth <dbl>,
## #   maternal_mortality_rate <int>, net_migration_rate <dbl>, population <int>,
## #   population_growth_rate <dbl>, country_type <fct>
```

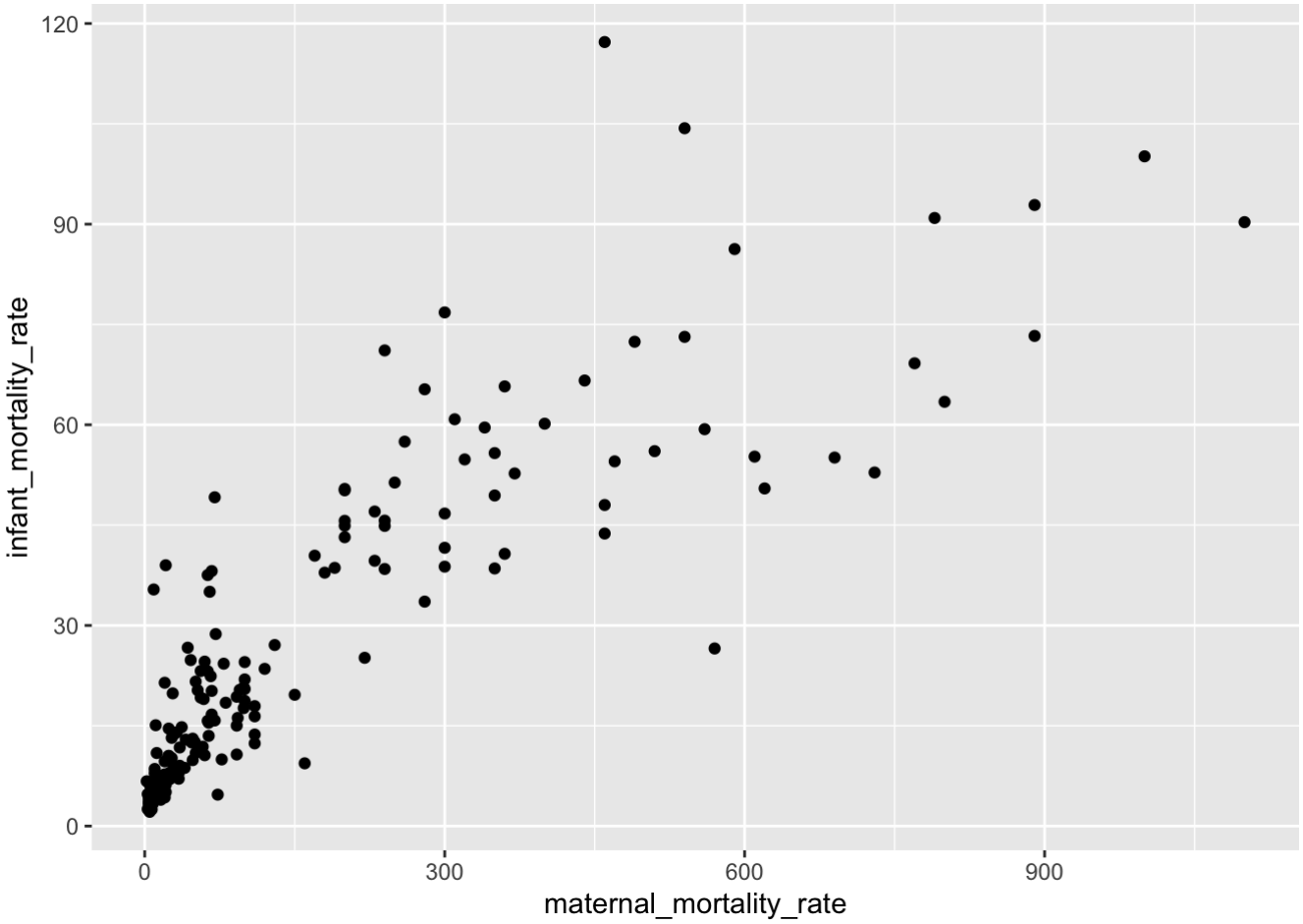
Title and Introduction

Perhaps one the most popular international news stories over the past couple weeks has been regarding birth rates and population change in countries across the globe. In countries like Japan, population growth rate has seen a steady decline, leading some to worry the nation may soon age and disappear. In contrast, Sub-Saharan African nations are set to see a six-fold increase in population. As a group we found the trends in global population and human development to be highly interesting, leading us to the CIA Factbook. This dataset is sourced from the openintro library in R. It contains columns for the country name, country area, population, population growth rate, birth rate, death rate, life expectancy, internet users, infant mortality rate, maternal mortality rate, and net migration rate for the year 2014. The original dataset has many rows with NA values so it requires cleaning by removing those rows. In addition, a categorical variable “country type” can be made based off of the existing variable for life expectancy. This dataset is interesting because we will be able to compare and contrast different developmental trends for countries based on their statistics. In particular, we would like to look at each country’s population and birth/death rates, as well as their life expectancy and mortality rates. We expect

to see higher life expectancy and lower mortality rates in more developed countries, and the opposite for less developed countries. As for birth and death rates, we expect to see lower rates for more developed countries, and higher ones for less developed countries.

Exploratory Data Analysis

```
# create scatter plot for maternal mortality rate and infant mortality rate
ggplot(cia_clean, aes(x = maternal_mortality_rate, y = infant_mortality_rate)) +
  geom_point()
```



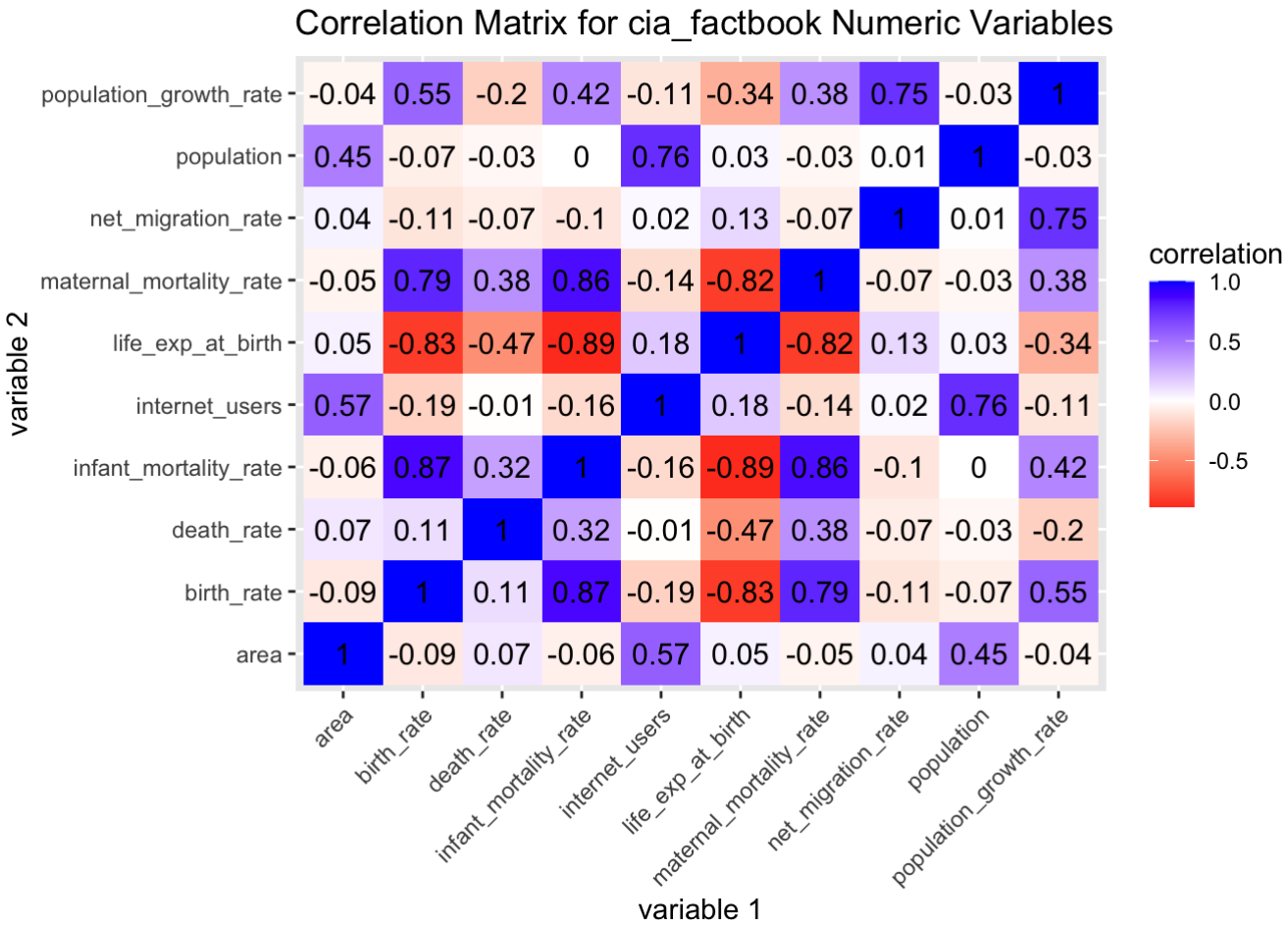

```
# find correlation between infant and maternal mortality rates
cor(cia_clean$maternal_mortality_rate,cia_clean$infant_mortality_rate, use = "pairwise.c
omplete.obs")
```

```
## [1] 0.857046
```

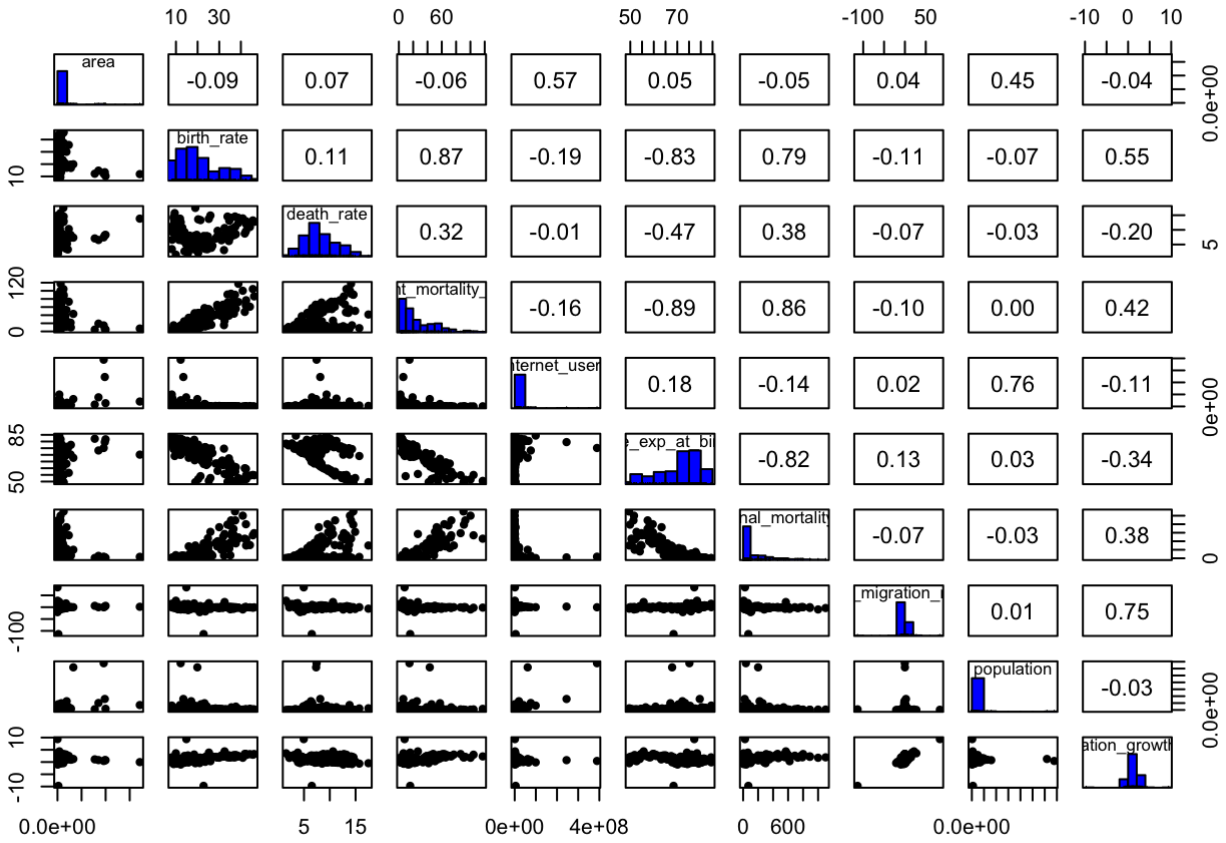
```
# create dataframe from cia_clean with only numeric variables
cia_num <- cia_clean %>%
  select_if(is.numeric)

# build a correlation matrix between all numeric variables
cor(cia_num, use = "pairwise.complete.obs")
```

```
##                                area  birth_rate   death_rate
## area                     1.00000000 -0.09300600  0.074589447
## birth_rate              -0.09300600  1.00000000  0.109784153
## death_rate               0.07458945  0.10978415  1.000000000
## infant_mortality_rate   -0.06061531  0.86536598  0.322897930
## internet_users           0.56992897 -0.19440771 -0.005720992
## life_exp_at_birth        0.04728153 -0.82766058 -0.474209922
## maternal_mortality_rate -0.04838870  0.79056586  0.383781491
## net_migration_rate       0.04219178 -0.11099481 -0.071592852
## population               0.45346935 -0.06904815 -0.030836269
## population_growth_rate  -0.04338162  0.54618294 -0.197333986
##                         infant_mortality_rate internet_users life_exp_at_birth
## area                            -0.060615312    0.569928967        0.04728153
## birth_rate                       0.865365979   -0.194407707       -0.82766058
## death_rate                       0.322897930   -0.005720992       -0.47420992
## infant_mortality_rate            1.000000000   -0.155193196       -0.88750000
## internet_users                  -0.155193196    1.000000000        0.18115597
## life_exp_at_birth               -0.887499996    0.181155970        1.00000000
## maternal_mortality_rate          0.857046028   -0.141888829       -0.82081171
## net_migration_rate              -0.102949745    0.022248502        0.13368824
## population                       0.001136847    0.760296108        0.03127375
## population_growth_rate           0.418880039   -0.109103614       -0.33713188
##                         maternal_mortality_rate net_migration_rate   population
## area                                -0.04838870        0.042191778  0.453469351
## birth_rate                           0.79056586       -0.110994813 -0.069048150
## death_rate                           0.38378149       -0.071592852 -0.030836269
## infant_mortality_rate                0.85704603       -0.102949745  0.001136847
## internet_users                      -0.14188883        0.022248502  0.760296108
## life_exp_at_birth                   -0.82081171        0.133688242  0.031273746
## maternal_mortality_rate              1.00000000       -0.069349092 -0.033936150
## net_migration_rate                  -0.06934909        1.000000000  0.005274375
## population                          -0.03393615        0.005274375  1.000000000
## population_growth_rate               0.38368855        0.745155206 -0.034746764
##                         population_growth_rate
## area                                -0.04338162
## birth_rate                           0.54618294
## death_rate                          -0.19733399
## infant_mortality_rate                0.41888004
## internet_users                      -0.10910361
## life_exp_at_birth                   -0.33713188
## maternal_mortality_rate              0.38368855
## net_migration_rate                   0.74515521
## population                          -0.03474676
## population_growth_rate               1.00000000
```

```
# create a heatmap with geom_tile
cor(cia_num, use = "pairwise.complete.obs") %>%
  # save as a data frame
  as.data.frame %>%
  # convert row names to an explicit variable
  rownames_to_column %>%
  # pivot so that all correlations appear in the same column
  pivot_longer(-1, names_to = "other_var", values_to = "correlation") %>%
  ggplot(aes(rowname, other_var, fill = correlation)) +
  # create heatmap with geom_tile
  geom_tile() +
  # change the scale to make the middle appear neutral
  scale_fill_gradient2(low="red",mid="white",high="blue") +
  # overlay values
  geom_text(aes(label = round(correlation,2)), color = "black", size = 4) +
  # add title and axis labels
  labs(title = "Correlation Matrix for cia_factbook Numeric Variables", x = "variable 1"
, y = "variable 2") +
  # rotate x-axis labels to make them more readable
  theme(axis.text.x=element_text(angle=45,hjust=1))
```



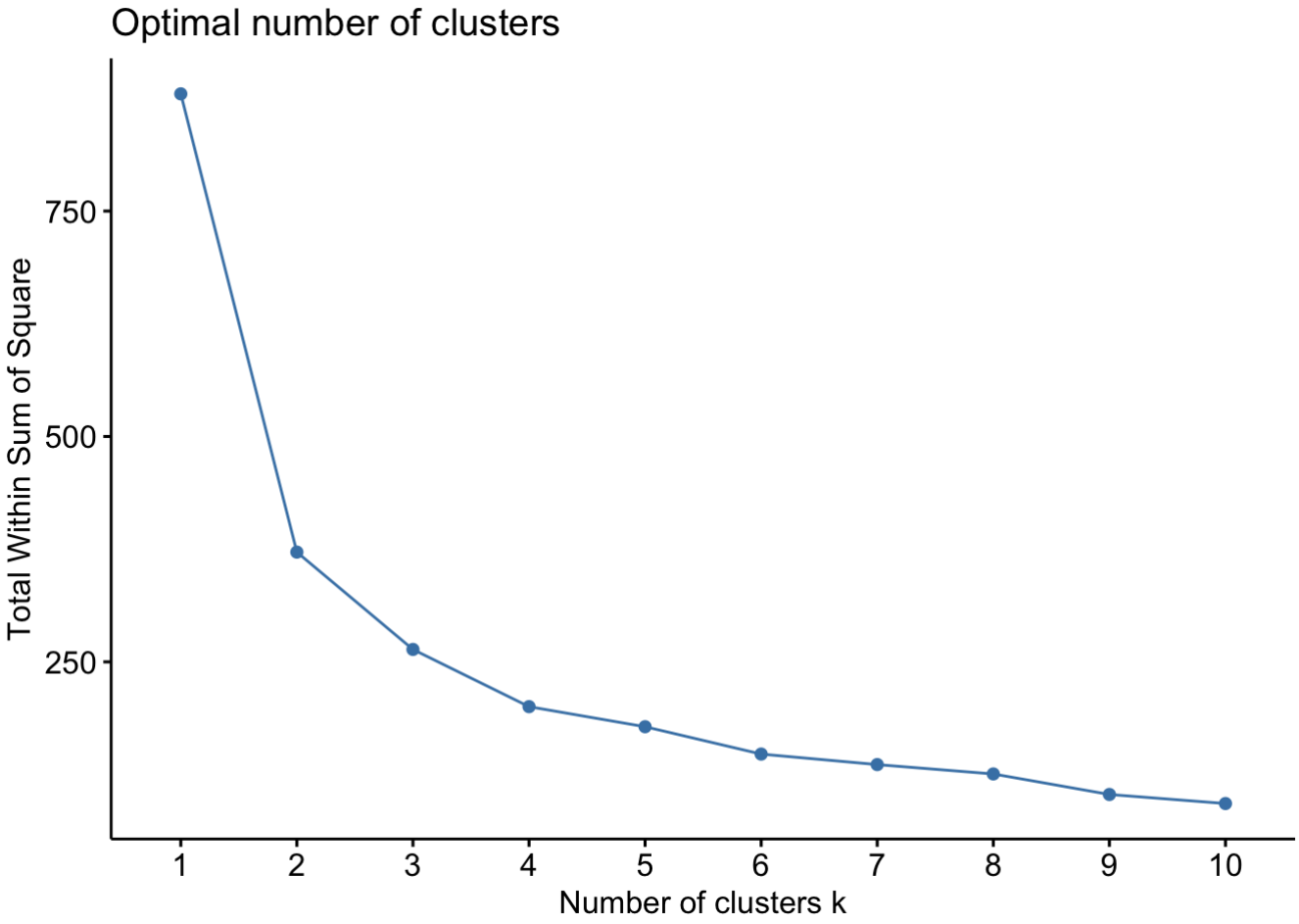
```
# create correlation matrix with univariate and bivariate graphs
pairs.panels(cia_num,
             method = "pearson", # correlation coefficient method
             hist.col = "blue", # color of histogram
             smooth = FALSE, density = FALSE, ellipses = FALSE)
```



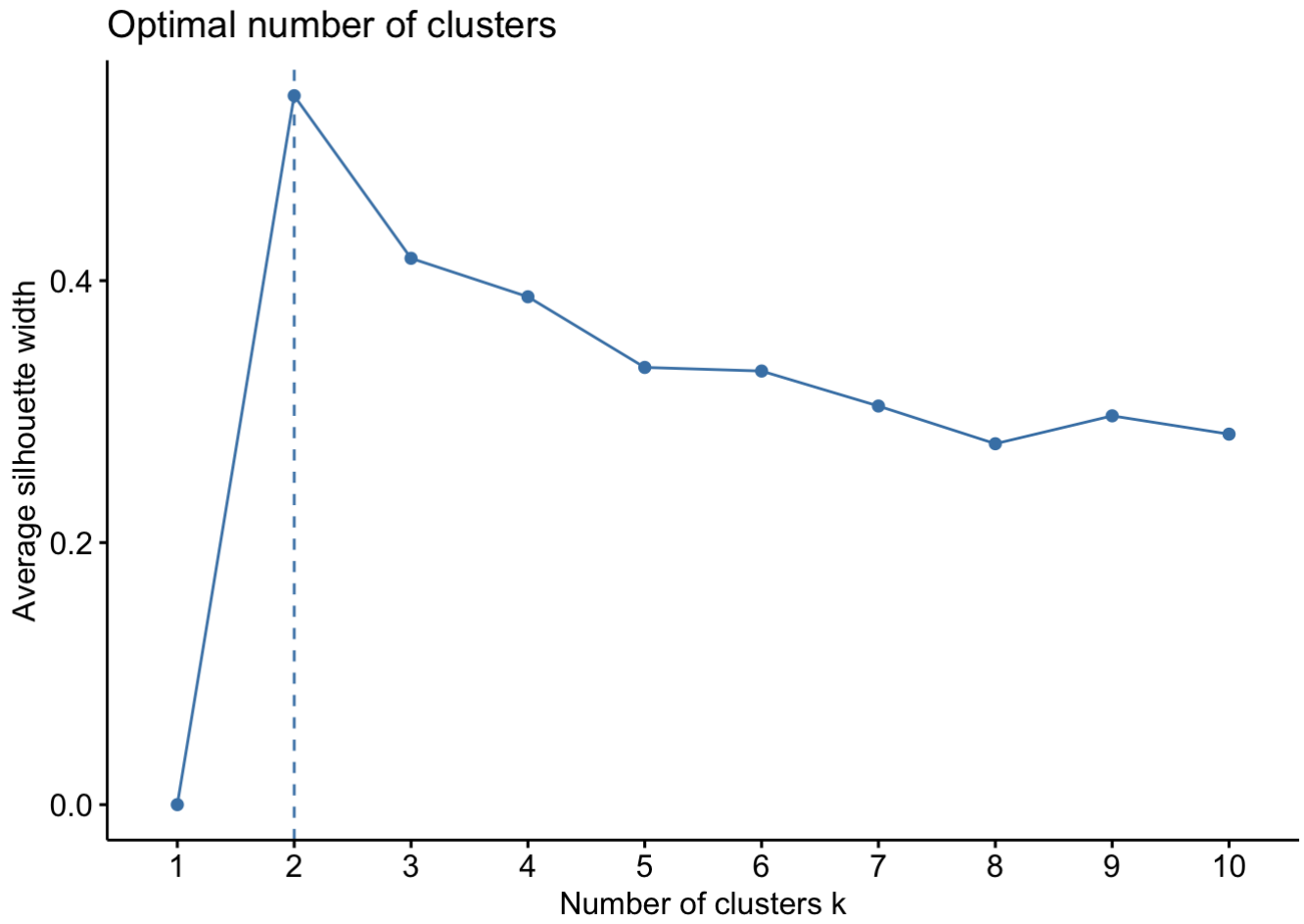
When analyzing the CIA Factbook dataset, we first wanted to verify which columns had the greatest relationships. Using the correlation matrix and heat map, we determined the variables with the strongest relationships. We thought these findings would be useful to decide which variables to use for further analysis in the project. Looking at the heatmap and correlation coefficients between the variables, we found that life expectancy at birth and infant mortality rate have the strongest negative correlation with a correlation coefficient of -0.89. This makes sense because if infant mortality is high, it is likely that the life expectancy of these infants would be low. The next highest correlation found was between maternal mortality rate and infant mortality rate with a correlation coefficient of 0.86. This also makes sense as generally, the factors that drive maternal mortality also drive infant mortality. There was also a strong negative correlation between maternal mortality rate and life expectancy at birth. Again, this aligns with previous observations as women generally have children at younger ages. Another reason these two correlations make sense is healthcare. Nations with good healthcare can generally support older populations with less mortality during childbirth. There were also high correlations found between internet users and area and internet users and population. However, these variables were not chosen for further analysis because these correlations and the fact that a higher population/country area leads to a higher number of internet users do not indicate anything about how developed a country is, which is the main goal of the clustering and classification portions of this project.

Clustering

```
# STEP 1: choose number of clusters

# select variables, scale, and create new dataframe
cia_var <- cia_clean %>%
  select(-country_type, -country, -area, -internet_users, -net_migration_rate, -populati
on, -population_growth_rate) %>%
  scale

# find optimal number of clusters using within sum-of-squares (wss) method
# minimize WSS while keeping a small number of clusters
fviz_nbclust(cia_var, pam, method = "wss")
```




```
# find optimal number of clusters using silhouette method
# check silhouette width
fviz_nbclust(cia_var, pam, method = "silhouette")
```

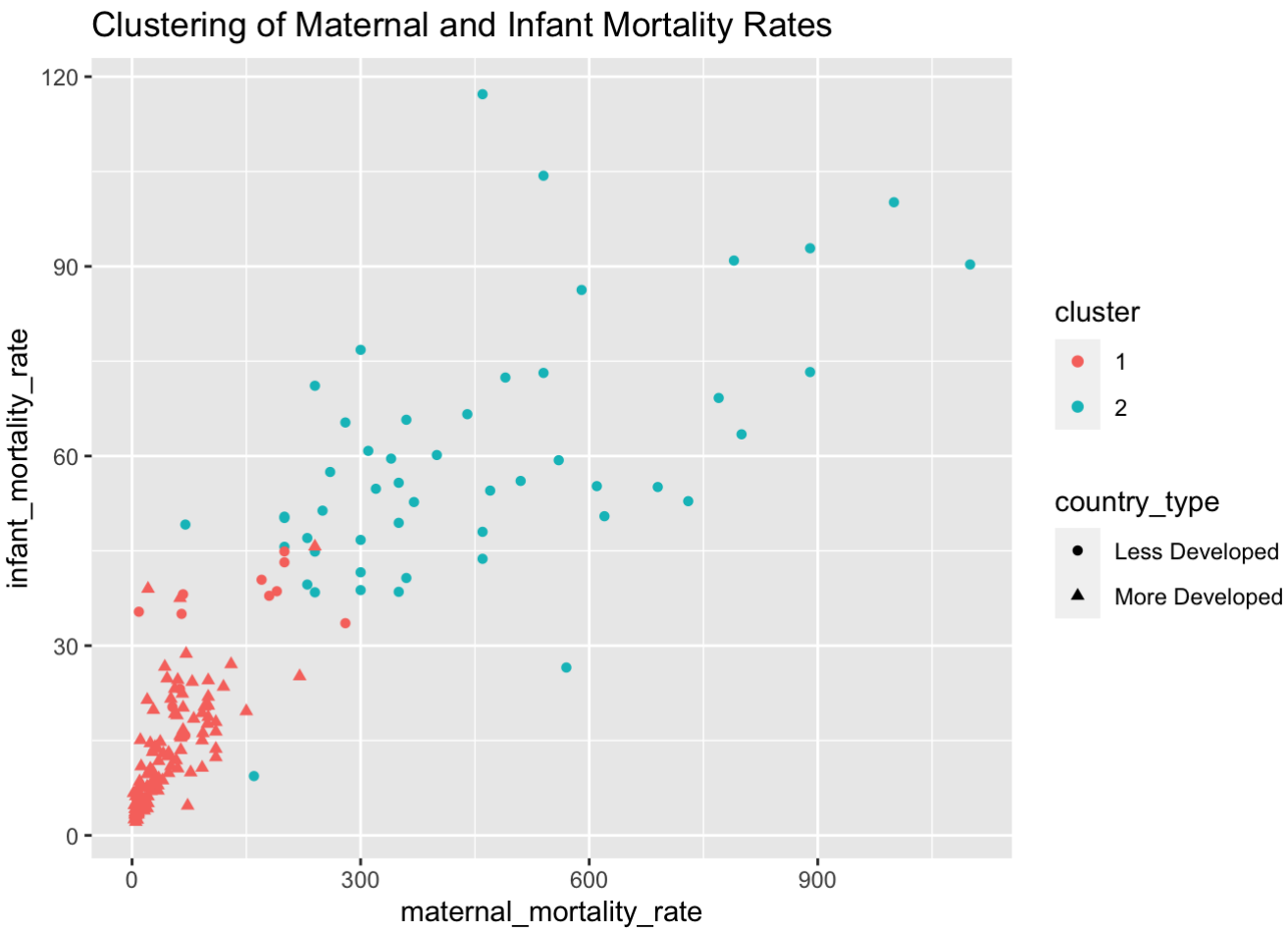


After finding our variables of interest, the next step in our project was to cluster our data. To do that, we searched for the best number of clusters. Using the WSS method (which measures the “compactness” of the clustering and minimizes it), it seems that the optimal number of clusters for the PAM method is around 2 or 3. We then implemented the silhouette method (which measures the quality of a clustering and determines how well each object lies within its cluster by considering both the WSS and the between-sum-of-squares), determining that the optimal number of clusters was indeed 2. Having found a number of clusters, we then moved on to clustering the data using the PAM method.

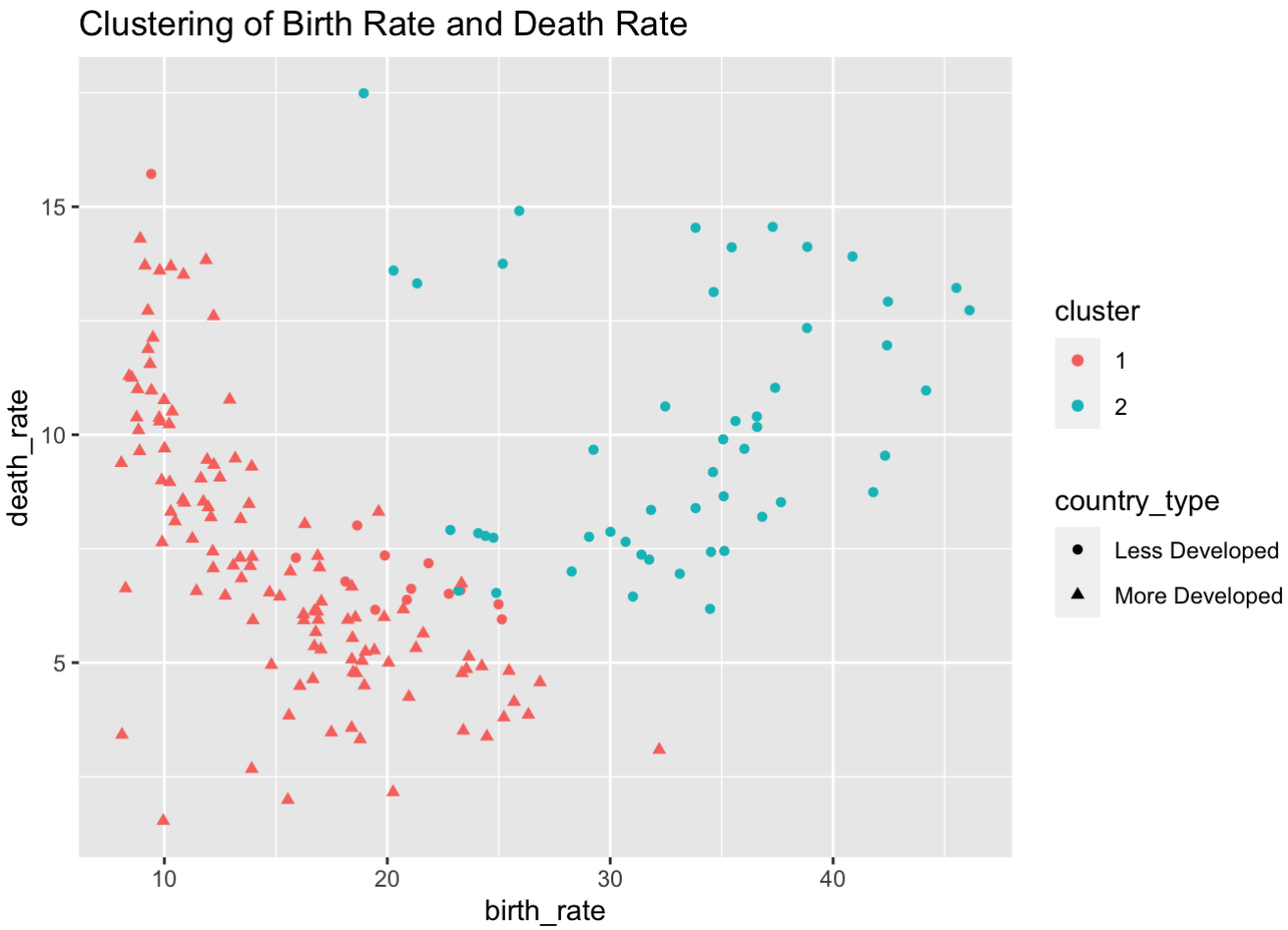
PAM Clustering

```
# STEP 2: Use PAM method for clustering select variables (life expectancy, birth rate, d
eath rate, maternal mortality rate, infant mortality rate)

# apply a clustering algorithm
pam_results <- cia_var %>%
  pam(k = 2)

# save cluster assignment as a column in dataset
cia_pam <- cia_clean %>%
  mutate(cluster = as.factor(pam_results$clustering))
```

```
# STEP 3: visualize clusters
# make a plot of data colored by final cluster assignment (infant vs maternal mortality
 rates)
cia_pam %>%
  ggplot(aes(maternal_mortality_rate, infant_mortality_rate, color = cluster)) +
  geom_point(aes(shape = country_type)) +
  labs(title="Clustering of Maternal and Infant Mortality Rates")
```




```
# make a plot of data colored by final cluster assignment (death rate vs birth rate)
cia_pam %>%
  ggplot(aes(birth_rate, death_rate, color = cluster)) +
  geom_point(aes(shape = country_type)) +
  labs(title="Clustering of Birth Rate and Death Rate")
```



Once we had clustered the data into 2 clusters, we then created two plots, one displaying the relationship between maternal mortality and infant mortality and the other displaying birth rate vs death rate. In each plot, data points are shaped by country_type and colored by cluster. When viewing the scatter plots for infant vs maternal mortality rates, the clustering looks pretty accurate in that it clusters more developed countries into cluster 1 and less developed countries into cluster 2. This pattern with the clustering is also evident in the scatter plot for death rate vs birth rate. Some countries were less developed in cluster 1 but they mainly existed on the border between clusters. This was an exciting observation as it displays how the two clusters represent developed and underdeveloped nations.

```
# STEP 4: evaluate clustering by calculating accuracy from cluster and country type
# compare the cluster and species
table(cia_pam$cluster, cia_pam$country_type)
```

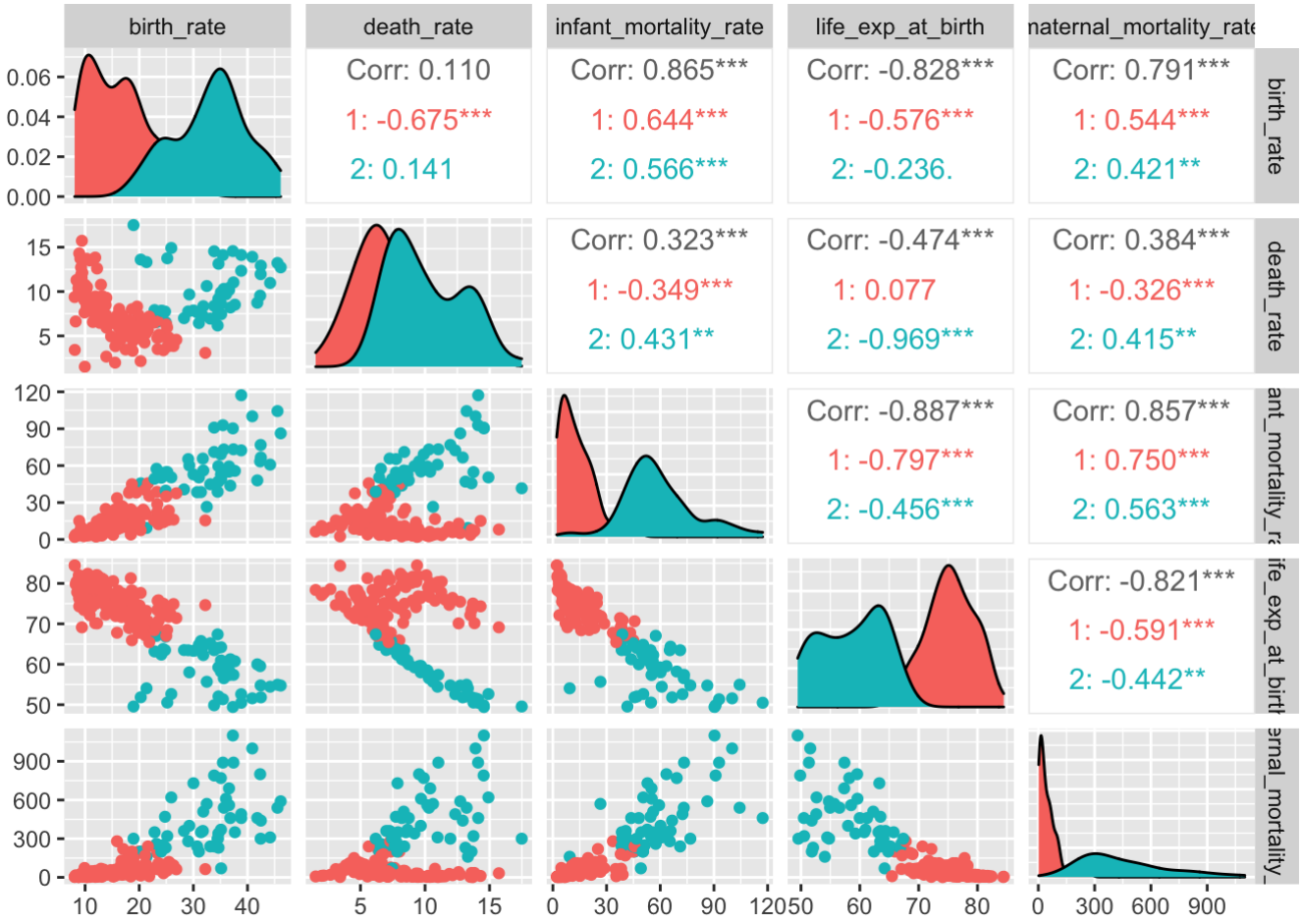
```
##
##     Less Developed More Developed
##   1             13            114
##   2             50              0
```

```
# calculate percentage of accuracy
(47+42+83/177)
```

```
## [1] 89.46893
```

The accuracy percentage of this clustering using the pam method is around 89.47%, when comparing the clustering with the actual country type variable from the dataset.

```
# STEP 5.1: interpret clustering by visualizing clusters by showing pairwise combination
s of variables

# show all pairwise combinations of variables colored by cluster assignment using ggpair
s
ggpairs(cia_pam, columns = c(3,4,5,7,8), aes(color = cluster))
```



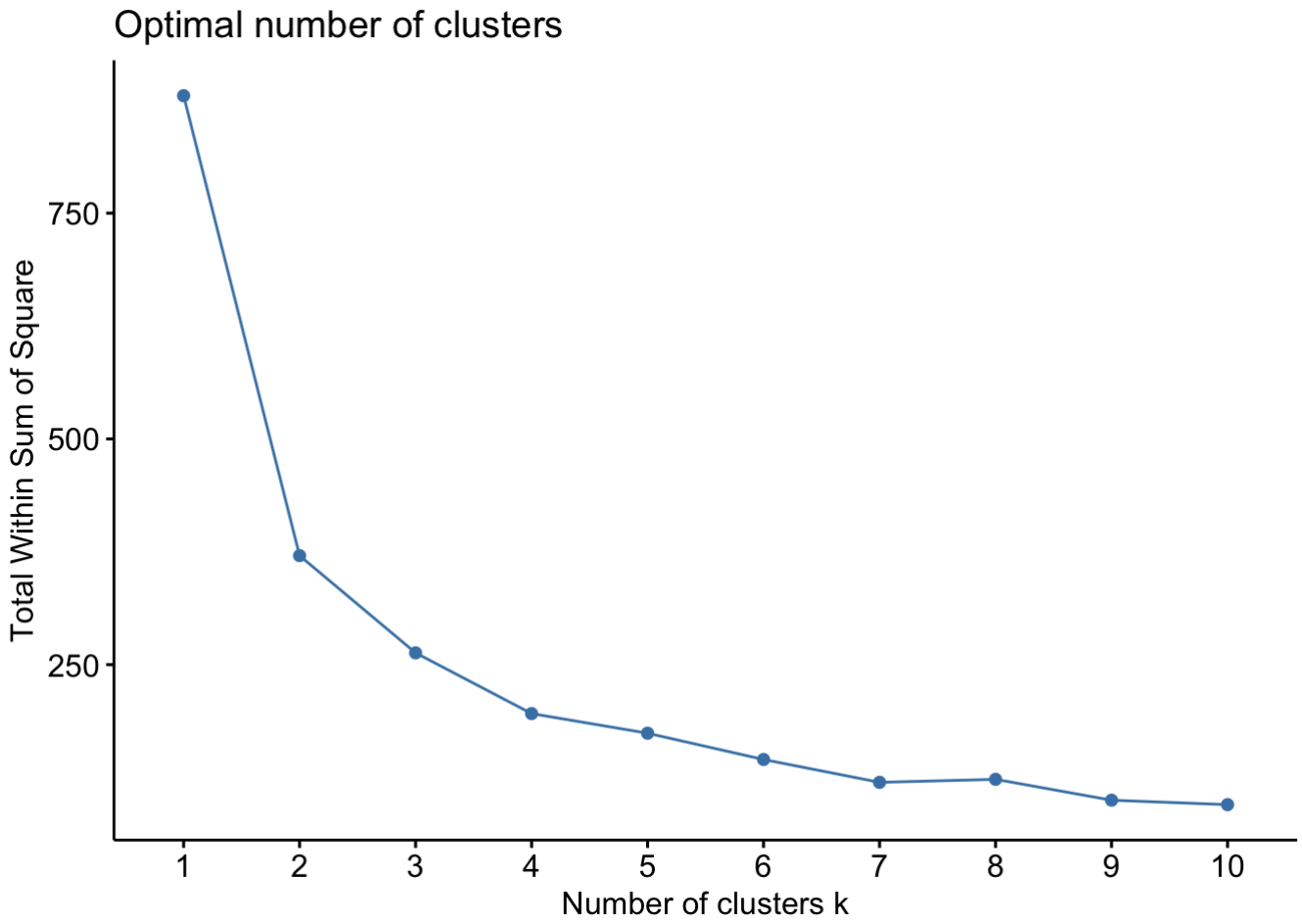

```
# STEP 5.2: interpret clustering by creating summary statistic for each variable

# find means of each variable for each cluster
cia_pam %>%
  group_by(cluster) %>%
  summarise_at(c("birth_rate", "death_rate", "infant_mortality_rate", "life_exp_at_birt
h", "maternal_mortality_rate"), mean, na.rm = T)
```

```
## # A tibble: 2 × 6
##   cluster birth_rate death_rate infant_mortality_rate life_exp_at_birth
##   <fct>        <dbl>      <dbl>                <dbl>             <dbl>
## 1 1             15.6       7.32                 14.3              75.4
## 2 2             33.2      10.1                  59.3              58.6
## # … with 1 more variable: maternal_mortality_rate <dbl>
```

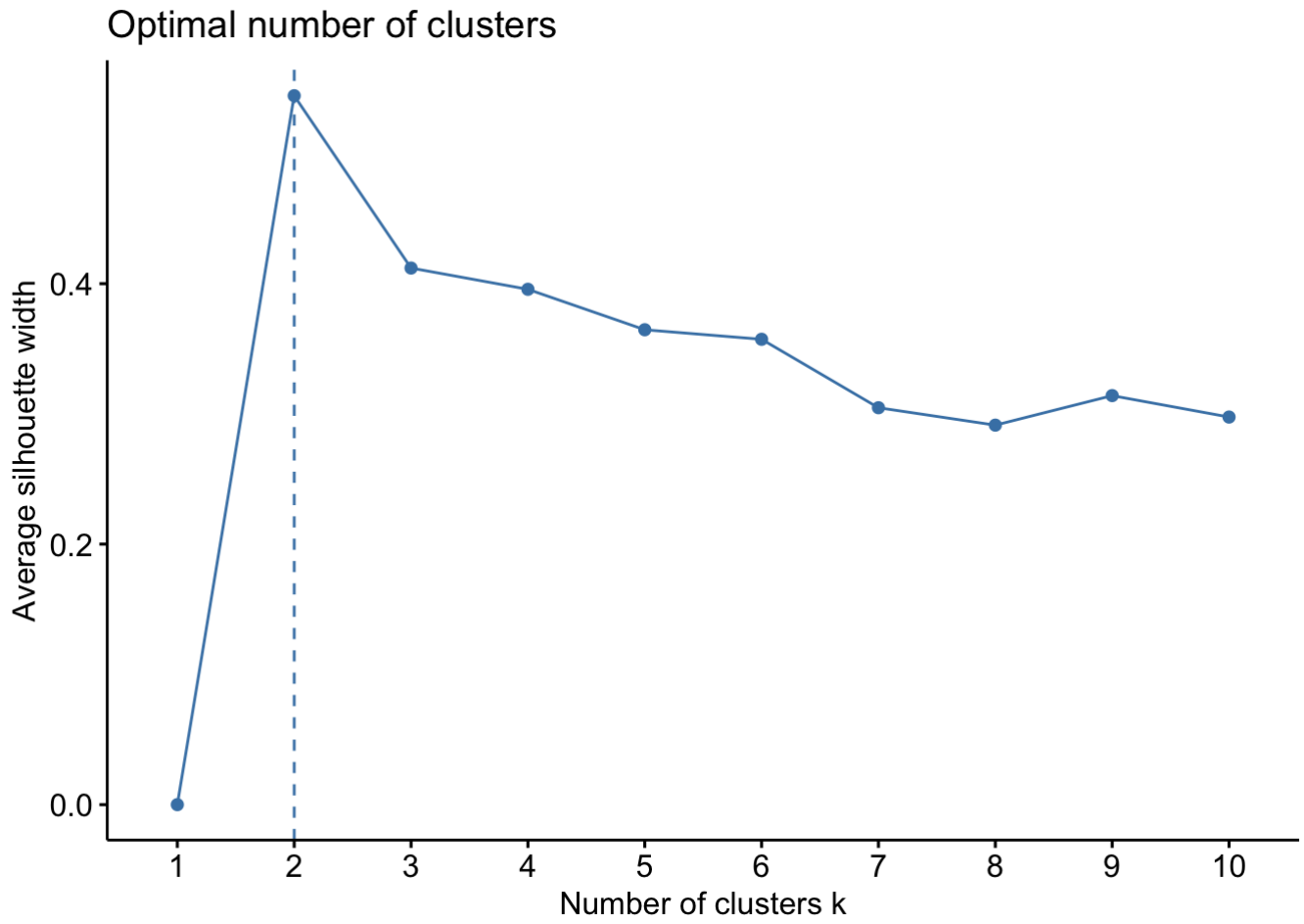
Upon interpreting the clusters created using the PAM method, one can see that the countries clustered in cluster 1 are mostly those determined to be More Developed based on life expectancy at birth. Similarly, the countries in cluster 2 are mostly those in the Less Developed category. Looking at the summary statistics for each variable and the ggpairs scatterplot matrix, one can see that the average values for each variable is very different for each cluster. This is expected, as developed countries tend to have lower birth rates, death rates, and mortality rates, but higher life expectancy and vice versa for developing countries. For example, the average life expectancy at birth for countries in cluster 1 is around 75.38 while the average for countries in cluster 2 is around 58.64. The density plots in the ggpairs matrix show this trend, as countries in cluster 1 (indicated in red) tend to have life expectancy at birth higher than the median, while countries in cluster 2 (indicated in blue) have life expectancies less than the median. This divide is also evident for birth rate, as the density plots for each cluster are on either side of the median. The cluster 1 plot is left-skewed, meaning it has lower average birth rates than the median, while the cluster 2 plot is right-skewed, meaning it has higher average birth rates than the median. This trend is also true, in the sense that cluster 2 is more right-skewed than cluster 1, for death rate and infant and maternal mortality rates. The variables that showed strong negative or positive correlations with each other also showed those trends in the scatter plots and correlation coefficients in the ggpairs matrix. For next steps, it could be interesting to view the clustering when 3 clusters are created.

K-Means Clustering

```
# STEP 1: choose number of clusters

# find optimal number of clusters using within sum-of-squares (wss) method
# minimize WSS while keeping a small number of clusters
fviz_nbclust(cia_var, kmeans, method = "wss")
```




```
# find optimal number of clusters using silhouette method
# check silhouette width
fviz_nbclust(cia_var, kmeans, method = "silhouette")
```

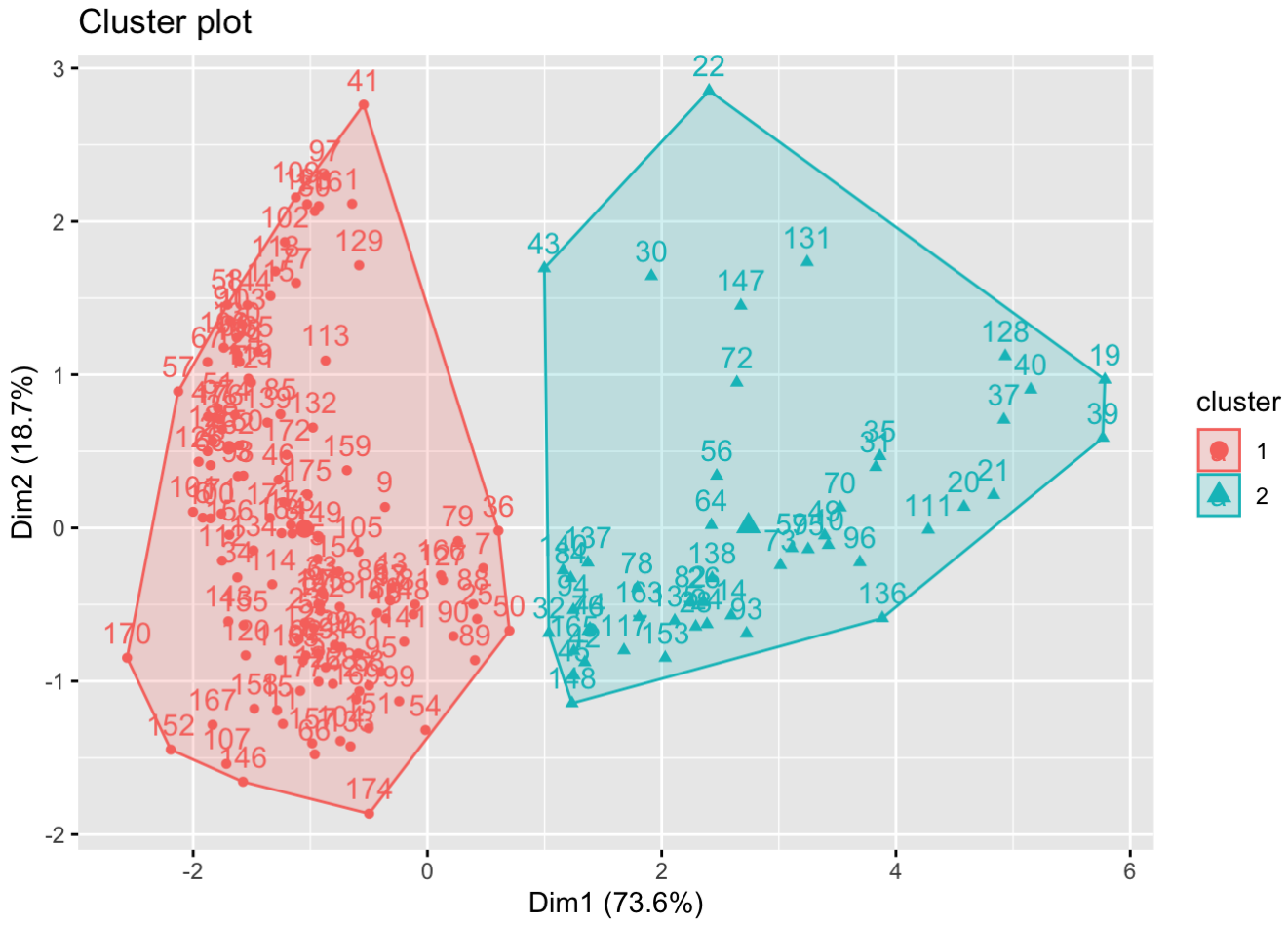


Using the WSS method, it seems that the optimal number of clusters for the kmeans method is around 2 or 3. Using the silhouette method, it is determined that the optimal number of clusters is 2.

```
# STEP 2: Use kmeans to cluster selected variables

# use kmeans function to find 2 clusters
kmeans_results <- kmeans(na.omit(cia_var),2)

# show available components
names(kmeans_results)
```

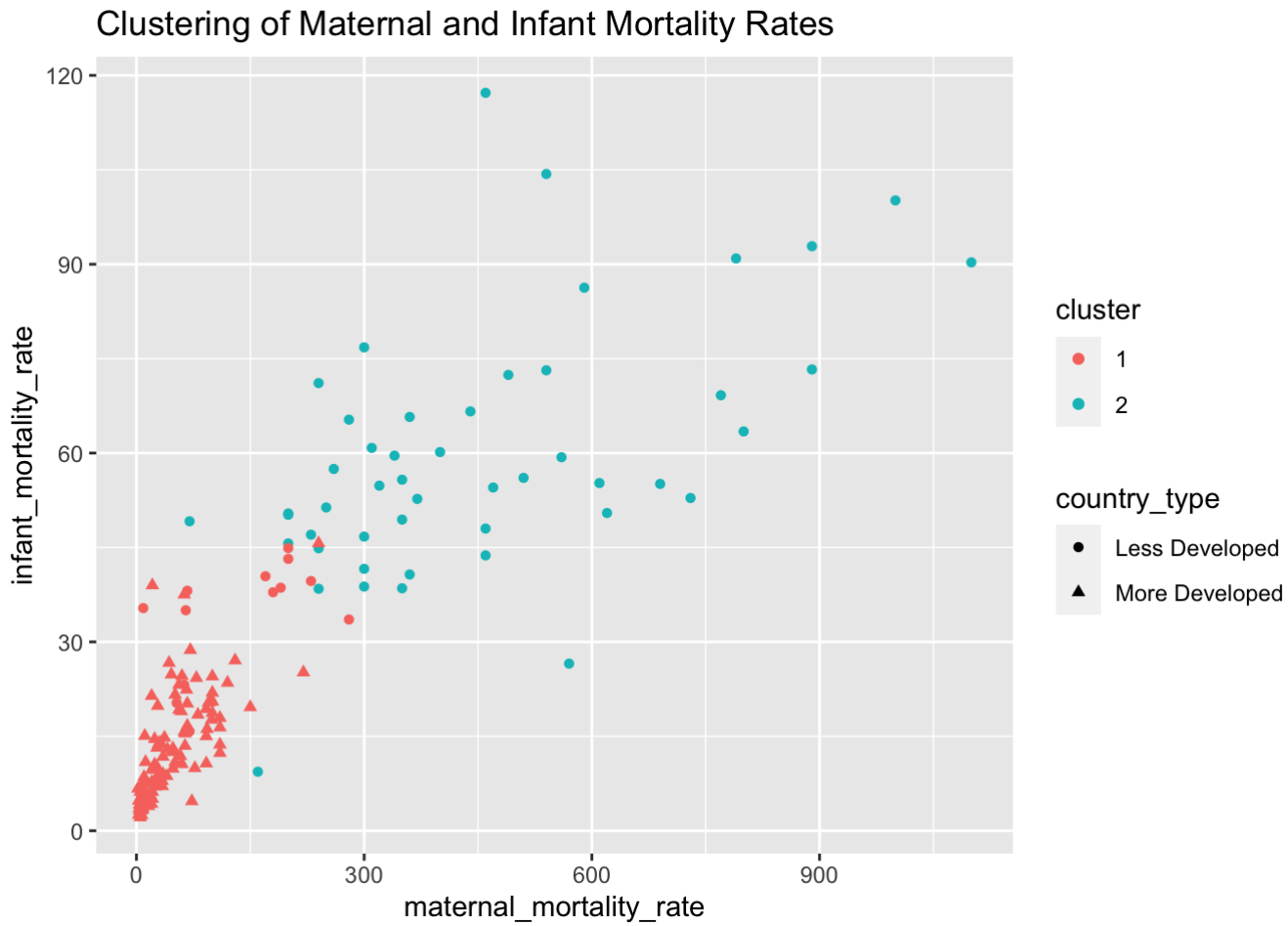
```
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
# visualize data by final cluster assignment
fviz_cluster(kmeans_results, data = cia_var)
```

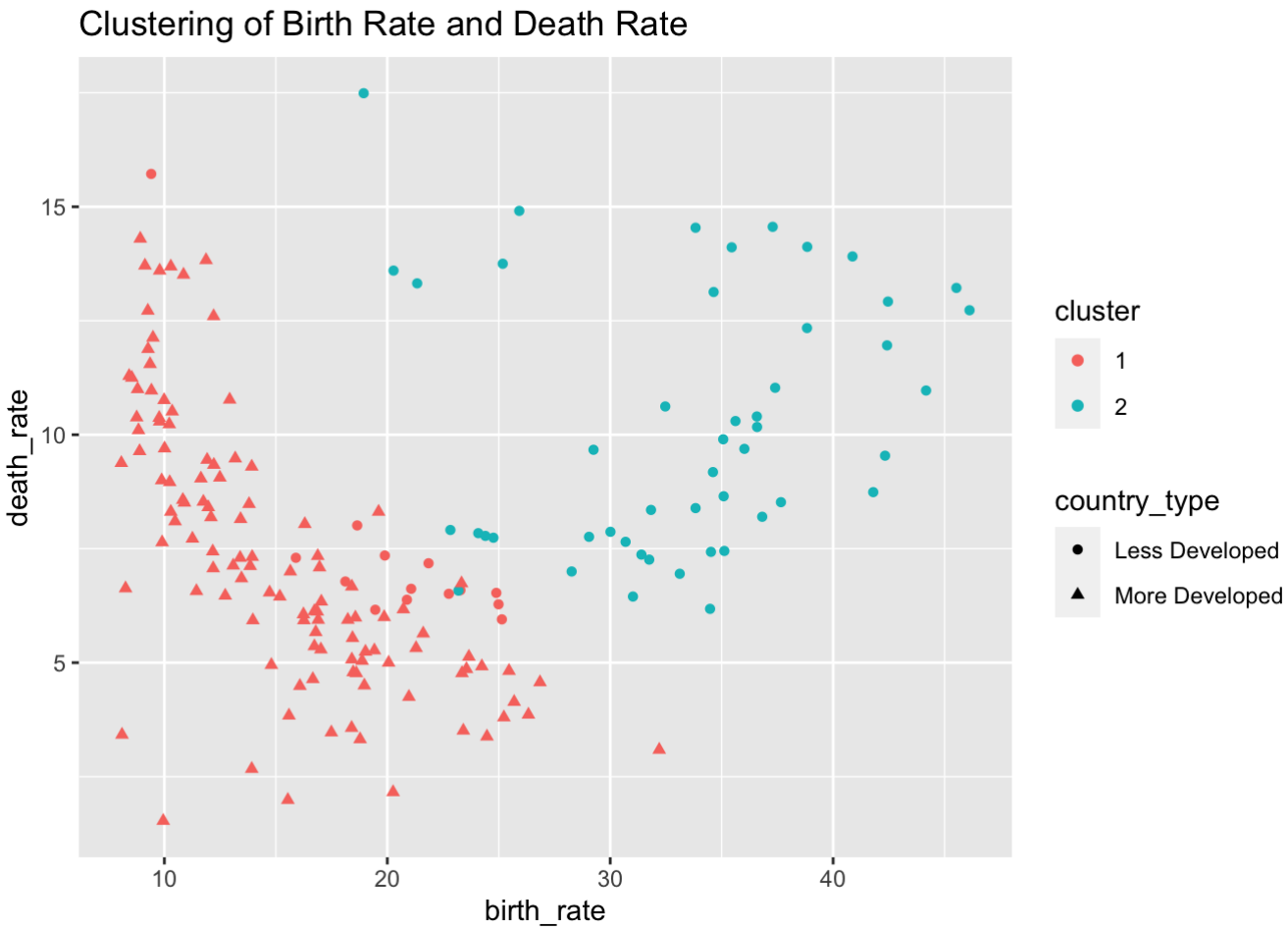



```
# save cluster assignment as a column in dataset
cia_kmeans <- cia_clean %>%
  mutate(cluster = as.factor(kmeans_results$cluster))
```

```
# STEP 3: visualize clusters

# visualize data by final cluster assignment (death rate vs birth rate)
cia_kmeans %>%
  ggplot(aes(maternal_mortality_rate, infant_mortality_rate, color = cluster)) +
  geom_point(aes(shape = country_type)) +
  labs(title="Clustering of Maternal and Infant Mortality Rates")
```




```
# visualize data by final cluster assignment (infant vs maternal mortality rates)
cia_kmeans %>%
  ggplot(aes(birth_rate, death_rate, color = cluster)) +
  geom_point(aes(shape = country_type)) +
  labs(title="Clustering of Birth Rate and Death Rate")
```



When viewing the scatter plot for infant vs maternal mortality rates, the kmeans method mostly clusters more developed countries into cluster 2 and less developed countries into cluster 1. This pattern with the clustering is also evident when the scatter plot displays death rate vs birth rate. This is the opposite clustering (1 vs 2) as the pam method but still clusters similarly.

```
# STEP 4: evaluate clustering by calculating accuracy from cluster and country type

# compare the cluster and country type
table(cia_kmeans$cluster, cia_kmeans$country_type)
```

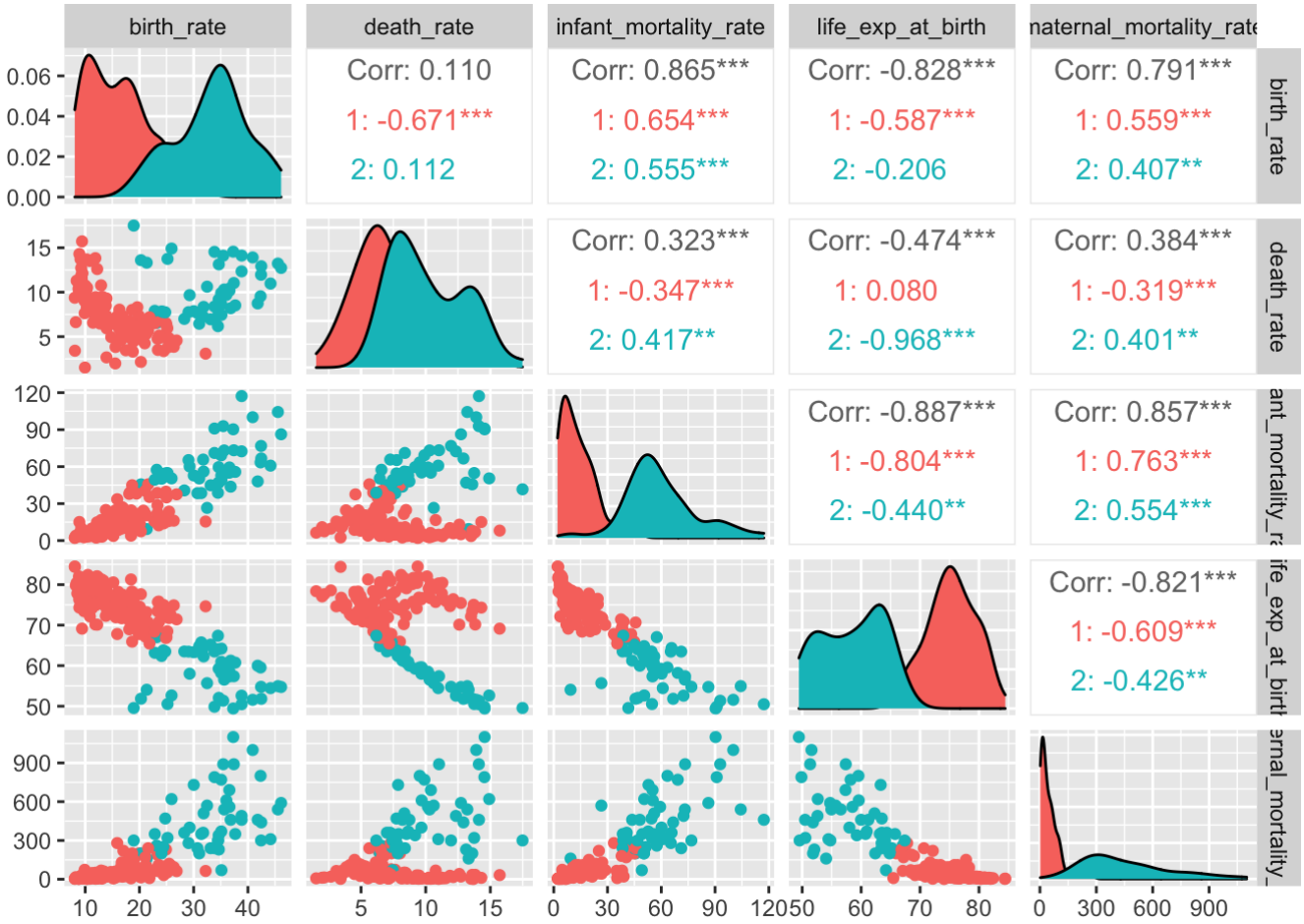
```
##
##     Less Developed More Developed
##   1             14            114
##   2             49              0
```

```
# calculate percentage of accuracy
(47+43+83/177)
```

```
## [1] 90.46893
```

The accuracy percentage of this clustering using the pam method is around 90.47%, when comparing the clustering with the actual country type variable from the dataset. This accuracy is slightly higher than the one provided by the pam method.

```
# STEP 5.1: interpret clustering by visualizing clusters by showing pairwise combination
s of variables

# show all pairwise combinations of variables colored by cluster assignment using ggpair
s
ggpairs(cia_kmeans, columns = c(3,4,5,7,8), aes(color = cluster))
```



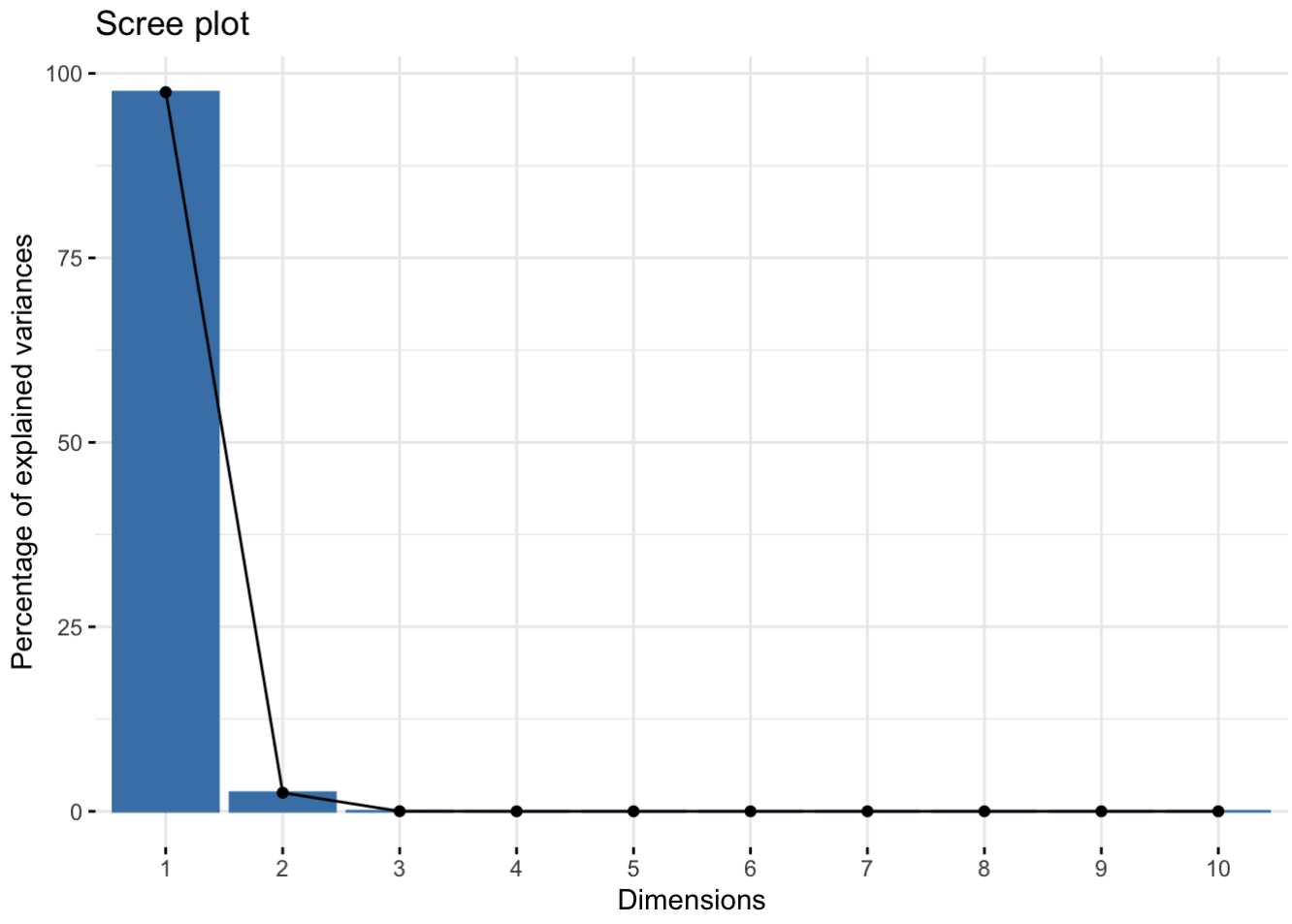

```
# STEP 5.2: interpret clustering by creating summary statistic for each variable

# find means of each variable for each cluster
cia_kmeans %>%
  group_by(cluster) %>%
  summarise_at(c("birth_rate", "death_rate", "infant_mortality_rate", "life_exp_at_birt
h", "maternal_mortality_rate"), mean, na.rm = T)
```

```
## # A tibble: 2 × 6
##   cluster birth_rate death_rate infant_mortality_rate life_exp_at_birth
##   <fct>        <dbl>      <dbl>                <dbl>             <dbl>
## 1 1             15.7       7.31                 14.5              75.3
## 2 2             33.3      10.2                  59.7              58.5
## # … with 1 more variable: maternal_mortality_rate <dbl>
```

Using the K-Means method for clustering, countries categorized as More Developed based on life expectancy at birth were mostly clustered into cluster 2 and countries categorized as Less Developed were mostly clustered into cluster 1. These results are similar to those of the PAM method in the sense that the More Developed countries were clustered together and the Less Developed countries were clustered together. The only difference between the clusters of the two methods was the cluster numbers were switched. The K-Means method produced similar results in that countries in cluster 2 (More Developed countries) tend to have lower birth and death rates and lower maternal and infant mortality rates and higher life expectancies and the opposite for those in cluster 1 (Less Developed countries). The accuracy of the K-Means model was slightly higher than that of the PAM model. However, the results confirmed each other that these highly correlated variables are predictive of the how developed a country is based on their life expectancy. Because it is difficult to categorize countries by how developed they are, we have used these selected variables to cluster them and see trends that are similar for each cluster. The development of a country is difficult to quantify as so many factors affect it, but these clustering methods help to visualize the similarities between countries based on the variables that we selected.

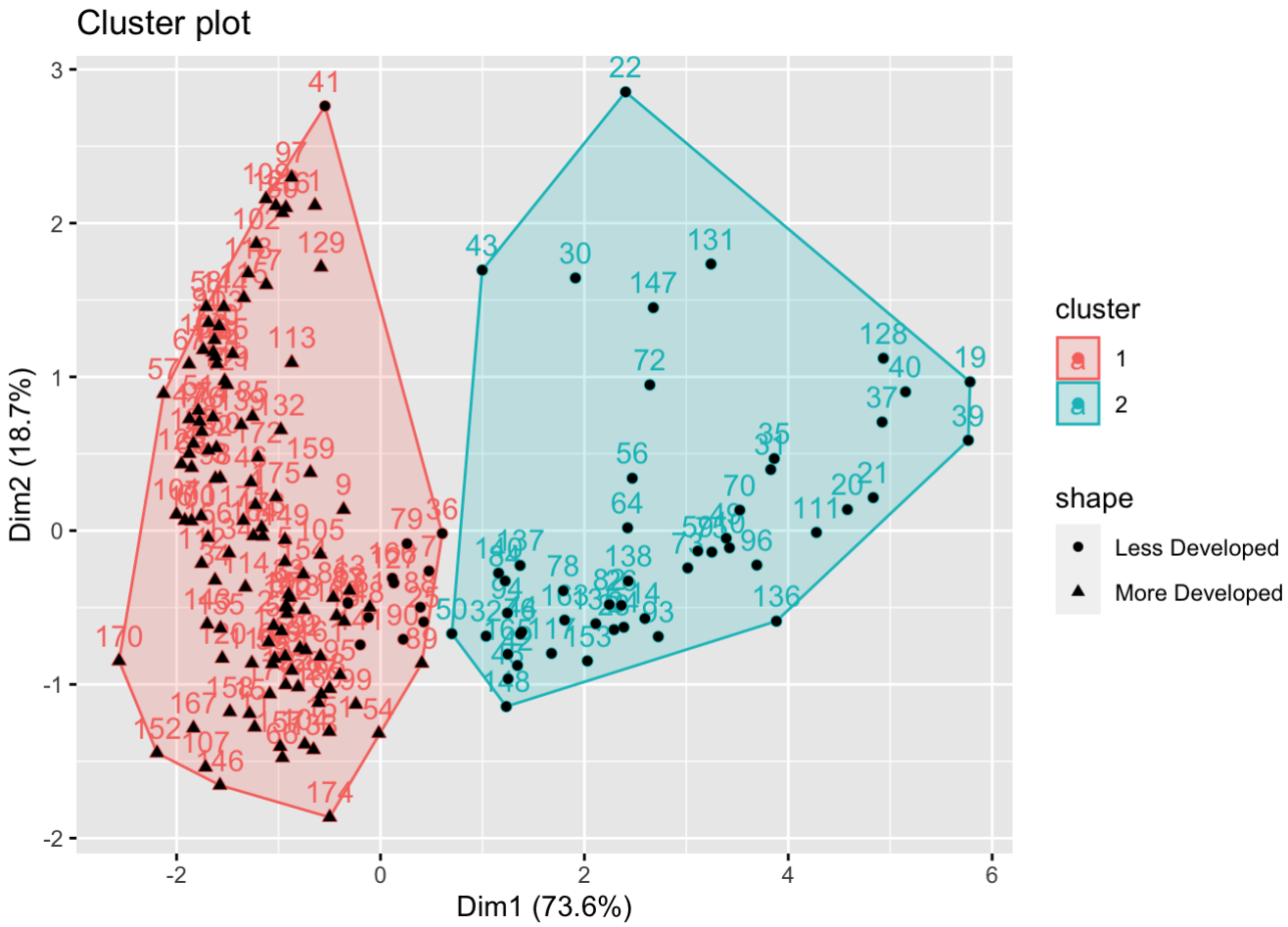
4. Dimensionality Reduction

```
library(factoextra)
# we use all numeric variables
pca <- cia_num %>%
  # use prcomp to find the principal components
  prcomp()
# percent of variances explained! choose how many pcs to retain
fviz_eig(pca)
```



```
# get the coefficients/loadings of each variable for each principal component (or dimens
ion)
pca_df <- get_pca_var(pca)$coord %>% as.data.frame
# use fviz_cluster to to visualize the observations using the first two pcs
fviz_cluster(pam_results, data = cia_clean, shape = cia_clean$country_type) +
  geom_point(aes(shape = cia_clean$country_type)) +
  guides(shape = guide_legend(title = "shape"))
```

```
## Warning in if (shape %in% colnames(data)) {: the condition has length > 1 and
## only the first element will be used
```




```
pam_results
```

```
## Medoids:
##       ID birth_rate death_rate infant_mortality_rate life_exp_at_birth
## [1,] 173 -0.6882518 -0.3156842            -0.5647079         0.4773806
## [2,]  82  1.4830264  0.1701950             1.0452221        -1.1006654
##      maternal_mortality_rate
## [1,]              -0.5228299
## [2,]               0.9250185
## Clustering vector:
##   [1] 1 1 1 1 1 1 1 1 2 1 1 1 2 1 1 1 1 2 2 2 2 1 2 1 2 1 2 1 2 2 2 1 1 2 1 2
##  [38] 1 2 2 1 2 2 2 2 1 1 1 2 2 1 1 1 1 1 2 1 1 2 1 1 1 1 2 1 1 1 1 1 2 1 2 2 1
##  [75] 2 2 1 2 1 1 1 2 1 2 1 1 1 1 1 1 1 1 2 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2
## [112] 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 1 2 2 2 2 1 2 1 1 1 1 1 1 2 2
## [149] 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1
## Objective function:
##    build     swap
## 1.368294 1.326185
##
## Available components:
##  [1] "medoids"    "id.med"     "clustering" "objective"  "isolation"
##  [6] "clusinfo"   "silinfo"    "diss"       "call"       "data"
```

We use `fviz_eig` to plot pcs vs percent of variance explained and we see negligible explanation after two pcs, so we pick two pcs! The two pcs explain 73.6 + 18.7 = 92.3% of the total variation in our dataset. High scores on PC1 mean the country has a high value for maternal mortality rate, very low values for area, internet users and population, and moderate values for birth rate, death rate, infant mortality rate, life expectancy at birth, net migration rate and population growth rate. Low scores on PC1 mean the country has a low value for maternal mortality rate, very high values for area, internet users and population, and moderate values for everything else. High scores on PC2 mean the country has a high value for population but very low values for area and internet users and moderate values for everything else. Low scores on PC2 mean the country has a very low value for population but very high values for area and internet users, and moderate values for everything else. Thus the dimension of greatest variability distinguishes high maternal mortality rate sites from the others. Given the shape or category we assign to the data points, it seems that high maternal mortality rate also distinguishes less developed countries from more developed countries quite well. We perform pca to reduce dimenionality of our dataset because two dimensions explain almost all the variation in our dataset. This narrows down the number of variables we need to consider to only two that are most predictive of how developed a country is: maternal mortality rate and population. Population less so, to a less degree because it seems that there are more less developed countries with high populations, but this division is not as stark as low and high maternal mortality rate.

Classification and Cross-Validation

For classification, and cross validation, the binary response variable we chose to predict was country_type. The model is trained on all variables except country name. The classifier we utilized to accomplish this is k-Nearest Neighbors.

```
# KNN

# Step 1: Prepare Dataset for training
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'lattice'
```

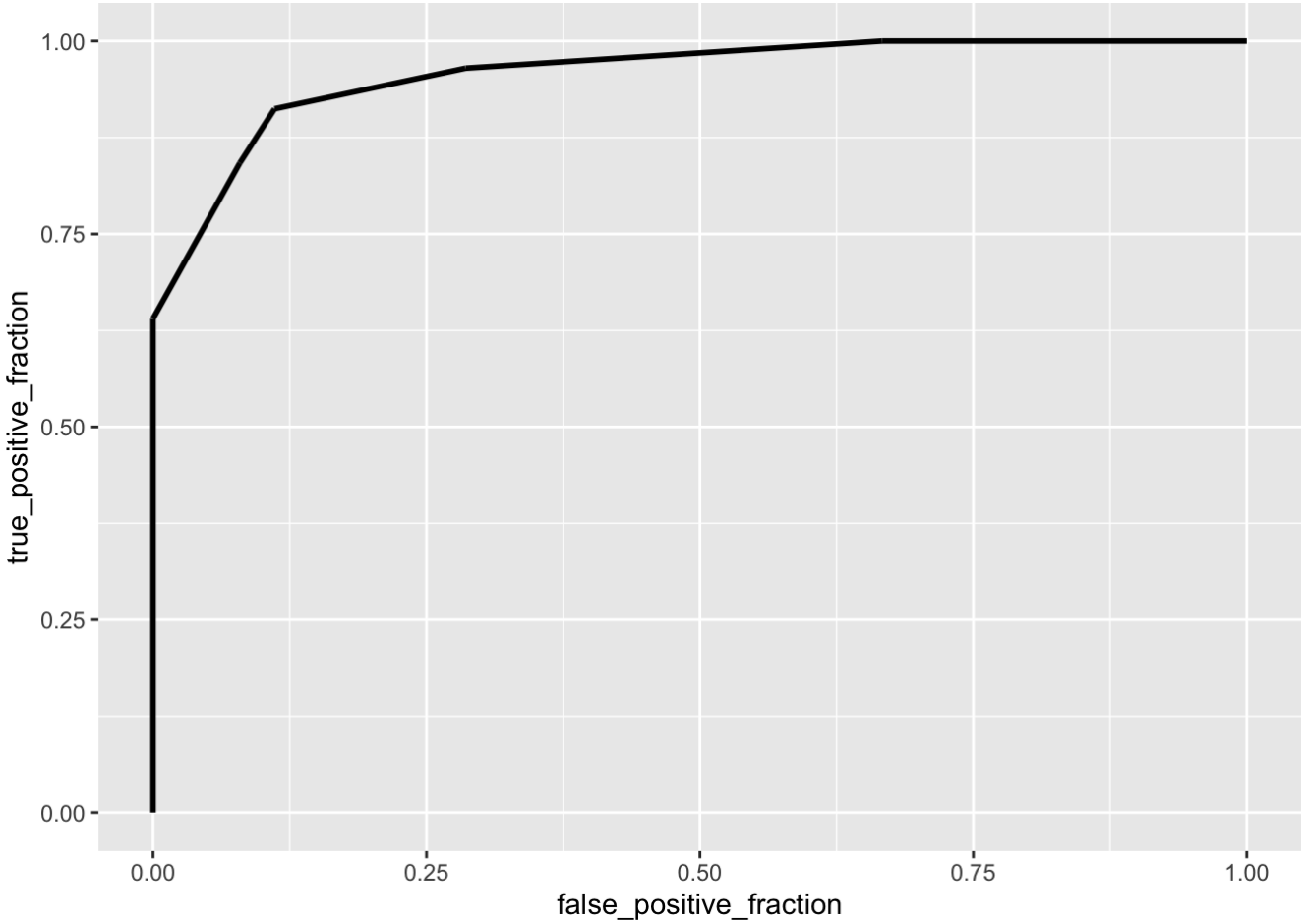
```
## The following objects are masked from 'package:openintro':
##
##     ethanol, lsegments
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:openintro':
##
##     dotPlot
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(plotROC)
cia_clean <- cia_clean %>% mutate(country_type = ifelse(country_type == 'Less Developed'
, 0, 1))

# Step 1: Train the Model
knn_fit <- knn3(factor(country_type == 1,
                       levels = c("TRUE","FALSE")) ~ birth_rate + death_rate + infant_mo
rtality_rate + internet_users + life_exp_at_birth + maternal_mortality_rate + net_migrat
ion_rate + population + population_growth_rate,
                data = cia_clean,
                k = 5)

kNN_cia <- cia_clean %>%
  mutate(proportion = predict(knn_fit, cia_clean)[,1])

# Step 2: Build ROC curve
ROC <- kNN_cia %>% ggplot() +
  geom_roc(aes(d = country_type, m = proportion), n.cuts = 0)
ROC
```




```
# calculate ROC
calc_auc(ROC)
```

```
##   PANEL group       AUC
## 1     1    -1 0.9581593
```

After training our classifier model and testing it on the whole dataset, we received an AUC of %95.81. This was a great accuracy as it was close to perfect. Our model accuracy was not surprising as all of our previous graphs displayed each Less and More developed nations in close proximity.

```
# Step 3: Perform k-fold Cross-Validation
set.seed(322)

# your code goes below this line (make sure to edit comment)
k = 10

# Randomly order rows in the dataset
data <- cia_clean[sample(nrow(cia_clean)), ]

# Create k folds from the dataset
folds <- cut(seq(1:nrow(data)), breaks = k, labels = FALSE)

# Use a for loop to get diagnostics for each test set
diags_k <- NULL

for(i in 1:k){
  # Create training and test sets
  train <- data[folds != i, ] # all observations except in fold i
  test <- data[folds == i, ]  # observations in fold i

  # Train model on training set (all but fold i)
  knn_fit <- knn3(factor(country_type == 1,
                         levels = c("TRUE","FALSE")) ~ birth_rate + death_rate + infant_mo
rtality_rate + internet_users + life_exp_at_birth + maternal_mortality_rate + net_migrat
ion_rate + population + population_growth_rate,
                  data = train,
                  k = 5)

  kNN_cia <- test %>%
    mutate(proportion = predict(knn_fit, test)[,1])

  # Step 2: Build ROC curve
  ROC <- kNN_cia %>% ggplot() +
    geom_roc(aes(d = country_type, m = proportion), n.cuts = 0)
  ROC
  calc_auc(ROC)

  # Get diagnostics for fold i (AUC)
  diags_k[i] <- calc_auc(ROC)$AUC
}
mean(diags_k)
```

```
## [1] 0.883212
```

Though model accuracy was satisfactory, we wanted to also ensure that our KNN classifier could effectively applied to potential nations not in our data or new data in later years. In order to verify this, we performed a 10-fold cross validation, choosing different portions of data to train and test in each fold. At the end of each fold, we saved the AUC. The final mean performance was calculated by averaging these AUC values across all the folds. Mean performance came out to %88.32. While this was still good, it was slightly worse than the original AUC

calculated. This indicates that our model is likely over-fitting to the training data and cannot classify new observations as well as previously trained observations. Nevertheless, our overall accuracy was still very good and it proves the interconnected nature of life expectancy and all other variables.

Overall, it was highly informative to analyze, cluster, and classify the CIA Factbook dataset. While no country is the same, it seems humanity as a whole follows similar trends as we develop and modernize.

Formatting.

Create the report using R Markdown, with headers for each section; include comments to the R code; include references (datasets, context). The final report is less than 20 pages. If working in a group, acknowledge how each member contributed to the project.

Yue - cleaning dataset, exploratory data analysis and clustering

Tushar - dimensionality reduction using pca

Raju - narrative introduction, classification and cross-validation

dataset - openintro