Homework 4

PSTAT131-231

Background

From the United Nations Development Programme website:

> "The Human Development Index (HDI) was created to emphasize that people and their capabilities should be the ultimate criteria for assessing the development of a country, not economic growth alone ... The HDI is a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and have a decent standard of living. The HDI is the geometric mean of normalized indices for each of the three dimensions.
>
> The health dimension is assessed by life expectancy at birth, the education dimension is measured by mean of years of schooling for adults aged 25 years and more and expected years of schooling for children of school entering age. The standard of living dimension is measured by gross national income per capita. The HDI uses the logarithm of income, to reflect the diminishing importance of income with increasing GNI. The scores for the three HDI dimension indices are then aggregated into a composite index using geometric mean.
>
> A fuller picture of a country's level of human development requires analysis of other indicators and information presented in the statistical annex of the report."

For this assignment, the 2019 HDI rankings of 139 nations were merged with 34 variables from the statistical annex of the UNDP's HDI report in that year. These variables comprise various economic, demographic, public health, and education/technology/communication attributes of national populations.

You will use unsupervised learning techniques to identify structure in the data and leverage learned structure to account for drivers of differing human development outcomes between countries.

```
# import 2019 HDI data
load('data/hdi.RData')
```

Part 1: Exploratory analysis

Question 1 (a). Create an HDI factor.

i) Create a factor representing level of human development by dividing the HDI ranks evenly into 5 groups (hint: `?cut`) with labels "very low", "low", "medium", "high", and "very high". When you create the labels, remember that a rank of 1, 2, 3, etc. – a low numerical value – is a *high* rank. Store the result as `hdi_level`.

```
# create hdi factor
hdi_level <- cut(hdi$hdi_rank, breaks = 5, labels = c("very high", "high","medium","low","very low"))
```

ii) Which ranks are included in each category? Identify the cutoffs.

Answer The very high category contains ranks 1-34. The high category contains ranks 35-61. The medium category contains ranks 62-87. The low category contains ranks 88-113. The very low category contains ranks 114-139.

Question 1 (b). Exploratory analysis via PCA.

i) Compute the principal components and principal component loadings after centering and scaling the data (without the HDI, HDI level, and country variables). Construct a scatterplot showing the data projected onto the first two PC's, with color mapped to the HDI level factor you created.

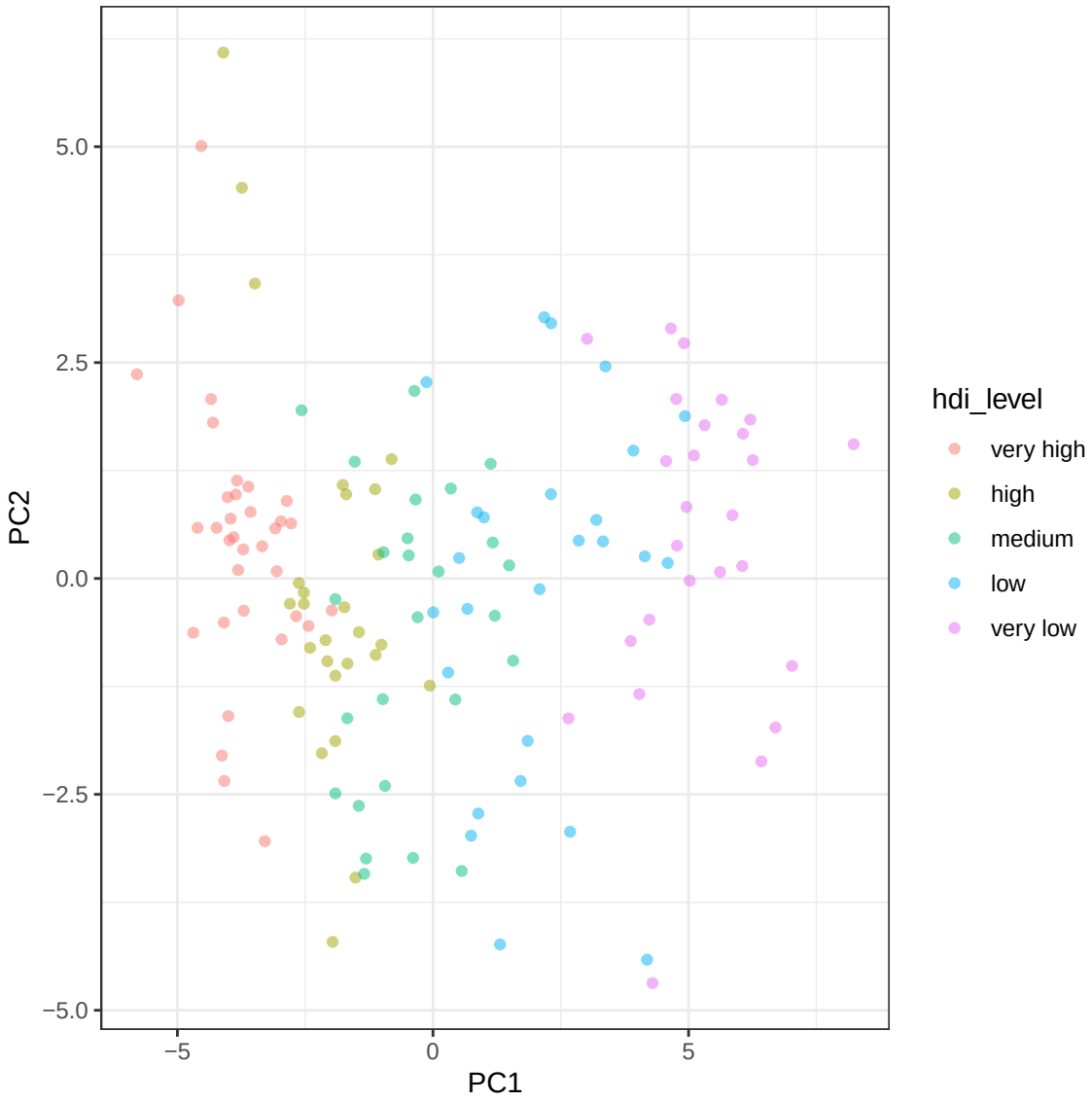
```
# extract features and center and scale
x_mx <- hdi %>%
  select(-c('hdi_rank', 'country')) %>%
  scale(center = T, scale = T)

# compute SVD
x_svd <- svd(x_mx)

# get loadings
v_svd <- x_svd$v

# compute PCs
z_mx <- x_mx %*% x_svd$v

# pca scatterplot
z_mx[, 1:2] %>%
  as.data.frame() %>%
  rename(PC1 = V1, PC2 = V2) %>%
  ggplot(aes(x = PC1, y = PC2)) +
  geom_point(aes(color = hdi_level), alpha = 0.5) +
  theme_bw()
```



ii) Based on the plot, which, if any, HDI levels seem well separated along the first two PCs?

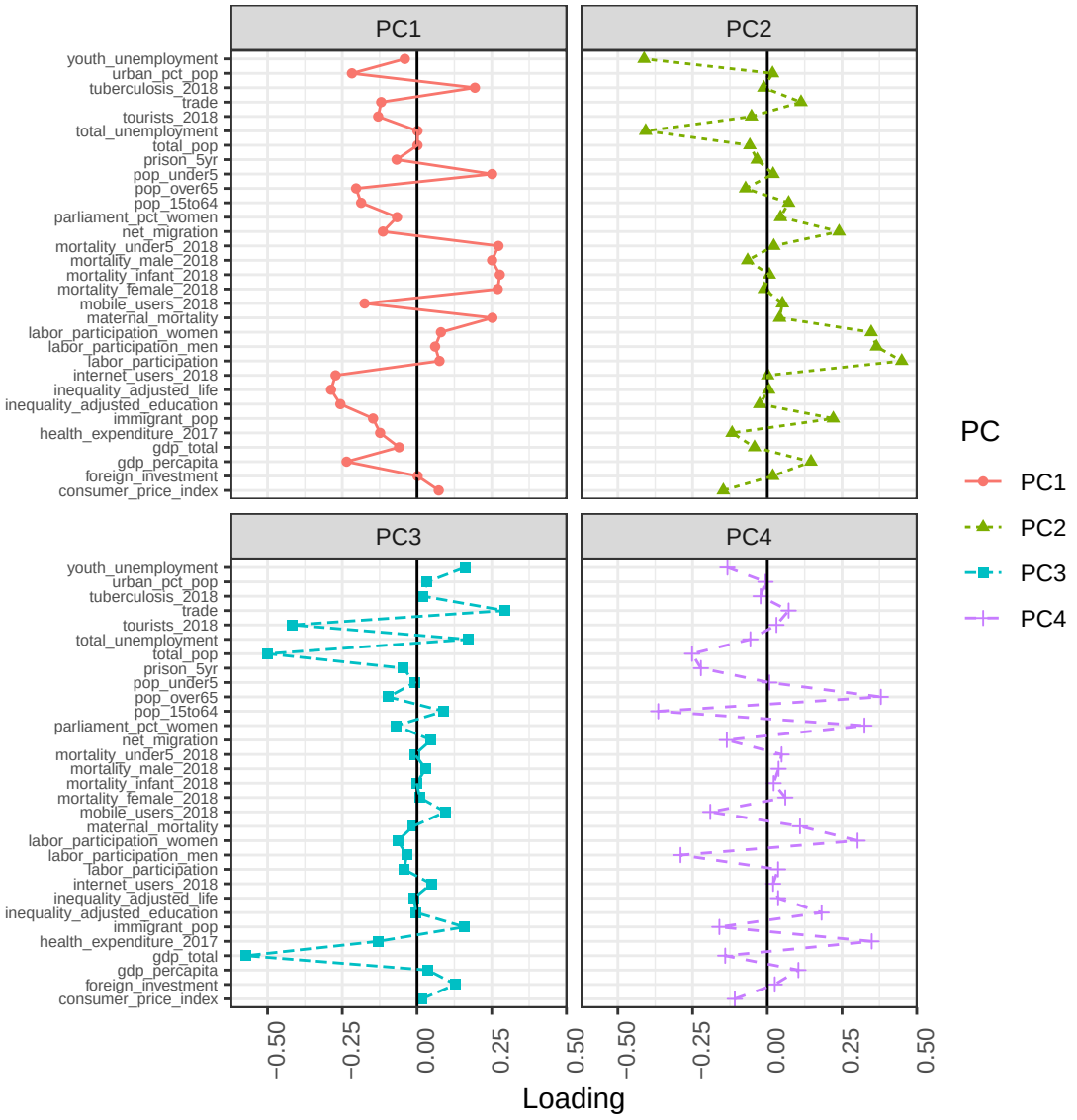
Answer Based on the plot all the HDI levels seem well separated along the first two PCs. The color of "medium" and "low" are very similar, but I can still discern that all the different hdi levels are pretty well separated.

iii) Plot the loadings for the first four PCs. For each loading plot, comment on the following:

- Which variables are most influential in determining the value of the principal component?
- Does the principal component seem to describe any interpretable attribute(s) of a country? If so, how would you interpret the pricipal component?

```
loading_plot <- v_svd[, 1:4] %>%
  as.data.frame() %>%
  rename(PC1 = V1, PC2 = V2, PC3 = V3, PC4 = V4) %>%
  mutate(variable = colnames(x_mx)) %>%
  gather(key = 'PC', value = 'Loading', 1:4) %>%
  arrange(variable) %>%
  ggplot(aes(x = variable, y = Loading)) +
```

```
  geom_point(aes(shape = PC, color = PC)) +
  theme_bw() +
  geom_hline(yintercept = 0, color = 'black') +
  geom_path(aes(linetype = PC, group = PC, color = PC)) +
  theme(axis.text.x = element_text(angle = 90), axis.text.y = element_text(size = 6)) +
  labs(x = '')
loading_plot+coord_flip()+facet_wrap(PC~.)
```



PC1 From the plot, I think the influential variables in determining the value of the principal component are the four mortality rate variables, 2 adjusted inequality variables, and the gdp per capital variable. I would interpret the principal component as a measure of mortality rates and life expectancy (life & death). In other words, as mortality increases, life expectancy decreases and vice versa.

PC2 Both of the labor participation variables and both of the unemployment variables are most influential variables in determining the value of the principal component. I would interpret the principal component as a measure of jobs and unemployment. Moreover, when labor participation increases, unemployment decreases.

PC3 Total GDP, total population, and tourists are the most influential variables in determining the value of the principal component. I think it's a bit more difficult to describe any interpretable attribute(s) of a country. If I were to interpret this, I would say that as the number of tourists in the country decrease, so does the GDP and population. Perhaps the country's economy is reliant on tourism, so when there is less tourism, there is less population and less GDP because of that.

PC4 Population 15-64, population above 65, and women in parliament are most influential variables in determining the value of the principal component. I would interpret the principal component as a measure of age and jobs. It seems that as a woman's age increases, so does the representation of women in the parliament. Furthermore, as people get older, the population of people between 15-64 decreases. Perhaps the parliament accepts older women because they have more life experiences?

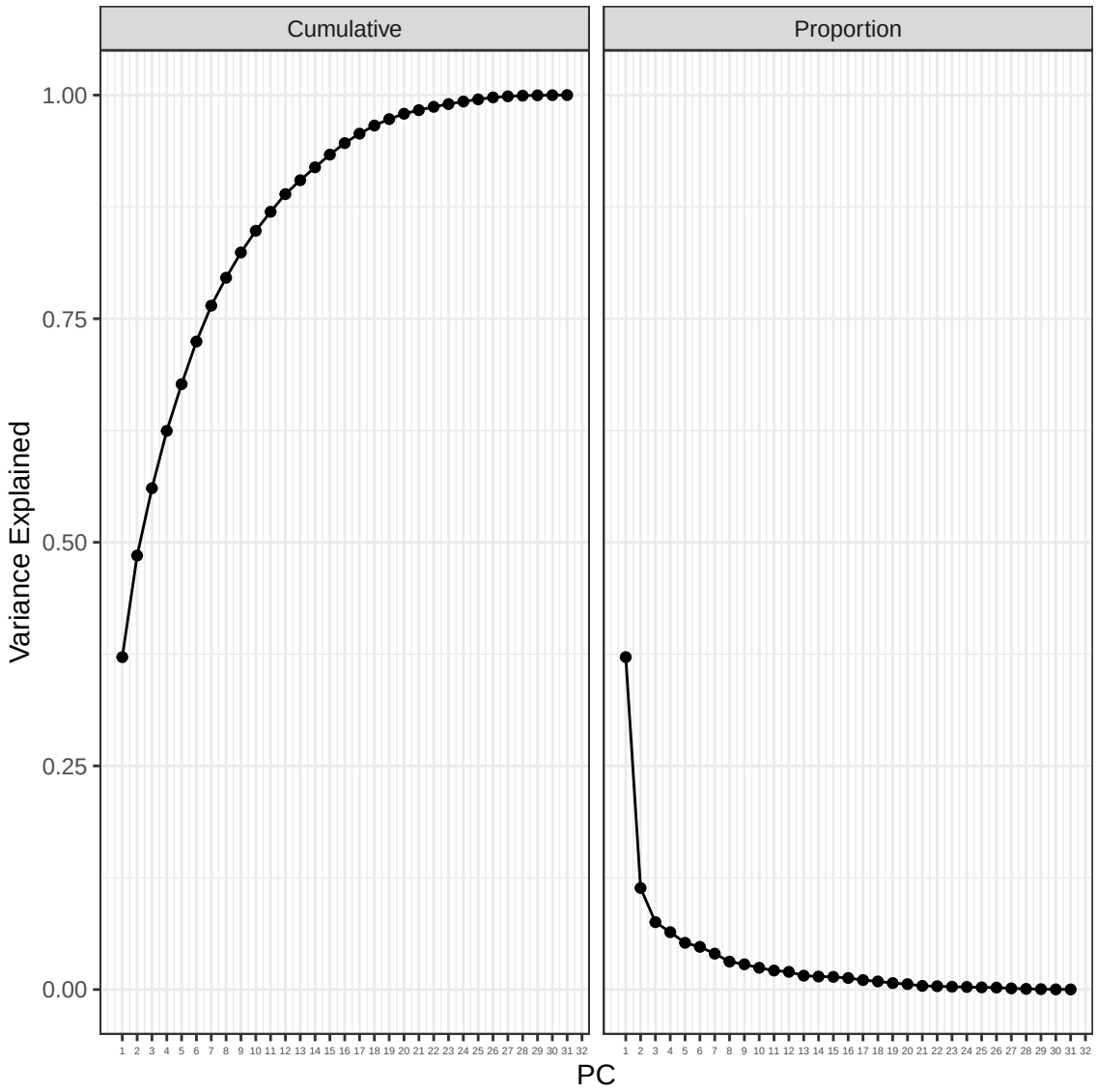
iv) Based on the loading plots for the first two PCs and the scatterplot, which variables seem to be the strongest correlates of human development?

Answer Based on the loading plots for the first two PCs and the scatterplot, I think that the PC1 variables that I listed earlier seem to be the strongest correlates of human development since they have the highest hdi values.

v) Construct the scree and cumulative variance plots. How much total variation is captured by the first four PCs?

```
# compute PC variances
pc_vars <- x_svd$d^2/(nrow(x_mx) - 1)

# scree and cumulative variance plots
tibble(PC = 1:min(dim(x_mx)),
       Proportion = pc_vars/sum(pc_vars),
       Cumulative = cumsum(Proportion)) %>%
  gather(key = 'measure', value = 'Variance Explained', 2:3) %>%
  ggplot(aes(x = PC, y = `Variance Explained`)) +
  geom_point() +
  geom_path() +
  facet_wrap(~ measure) +
  theme_bw() + theme(axis.text.x = element_text(size = 4))+
  scale_x_continuous(breaks = 1:32, labels = as.character(1:32))
```



Answer I would say that about 70 percent of the total variation is captured by the first four PCs.

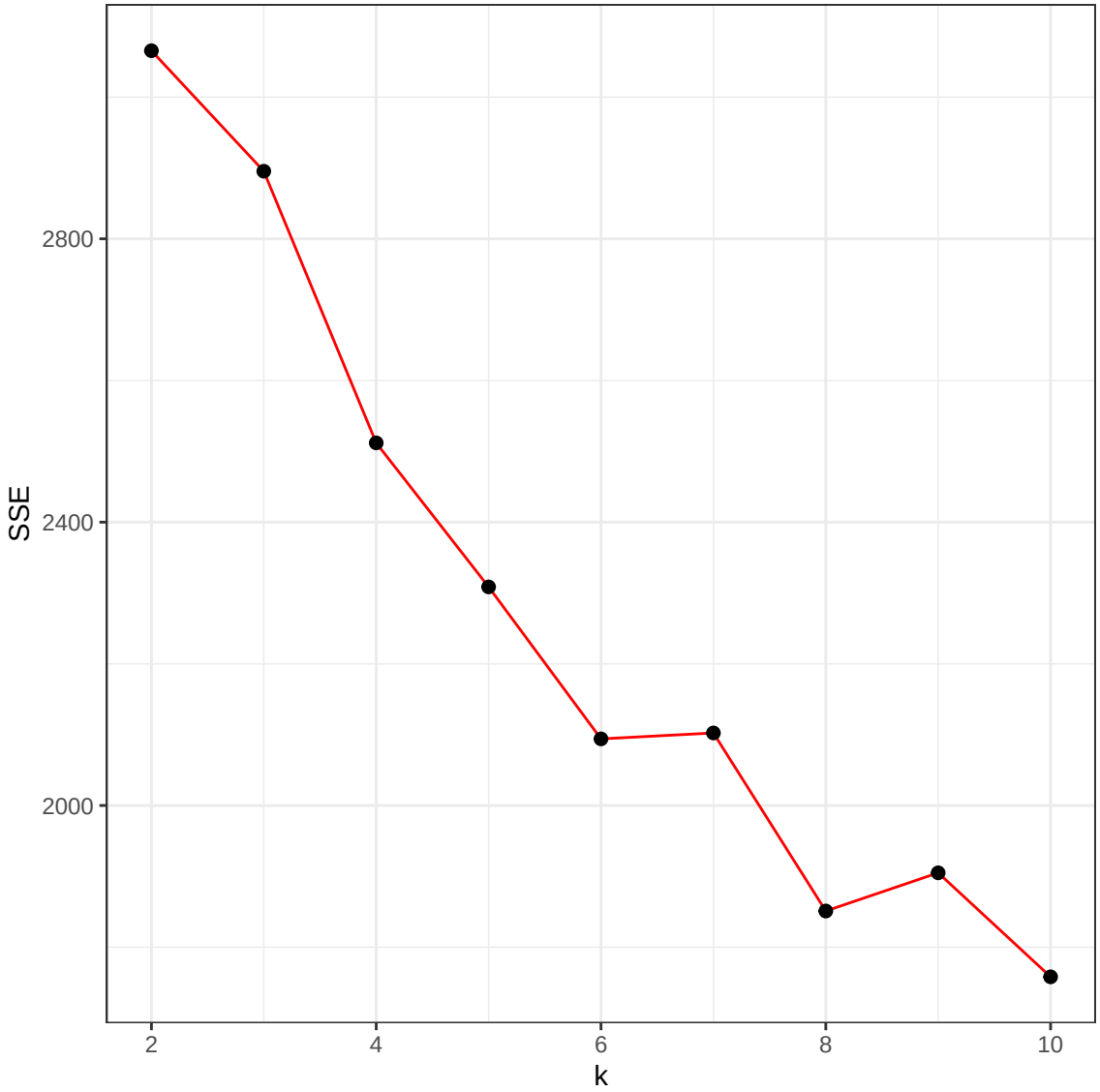
vi) Based on the loading plots for the first four PCs and the scree and cumulative variance plots, which variables seem to be the strongest drivers of total variation in the data?

Answer The PC's with the highest proportion values will be the strongest drivers of variation. So from our plot, we can see that PC1 has the highest proportion value. So, we can say that the four mortality rate variables, 2 adjusted inequality variables, and the gdp per capital variable seem to be the strongest drivers of total variation in the data.

Part 2: Clustering with k-means

Question 2 (a). Choosing K.

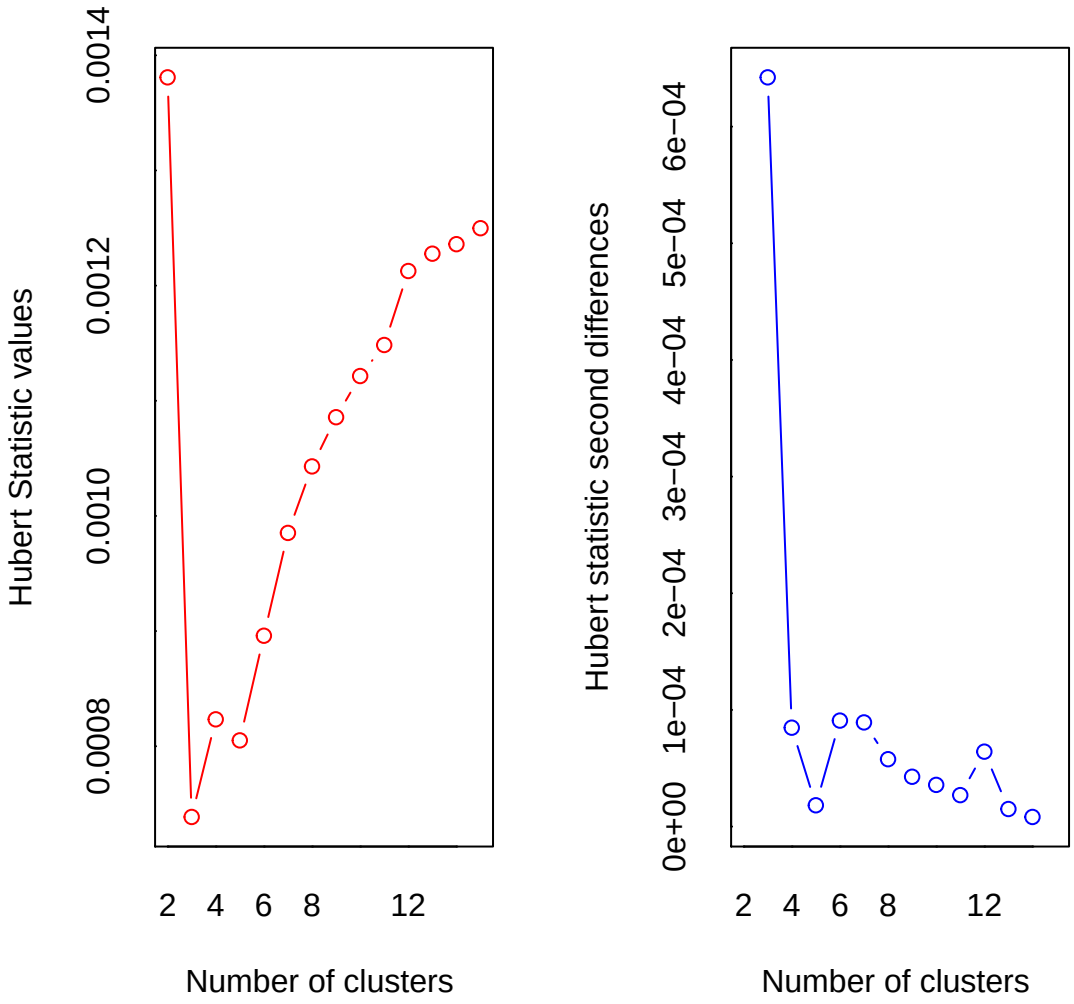
i) Compute SSE for k-means clustering of the centered and scaled data matrix for $k = 2, 3, \ldots, 10$ and plot SSE against the number of clusters k.



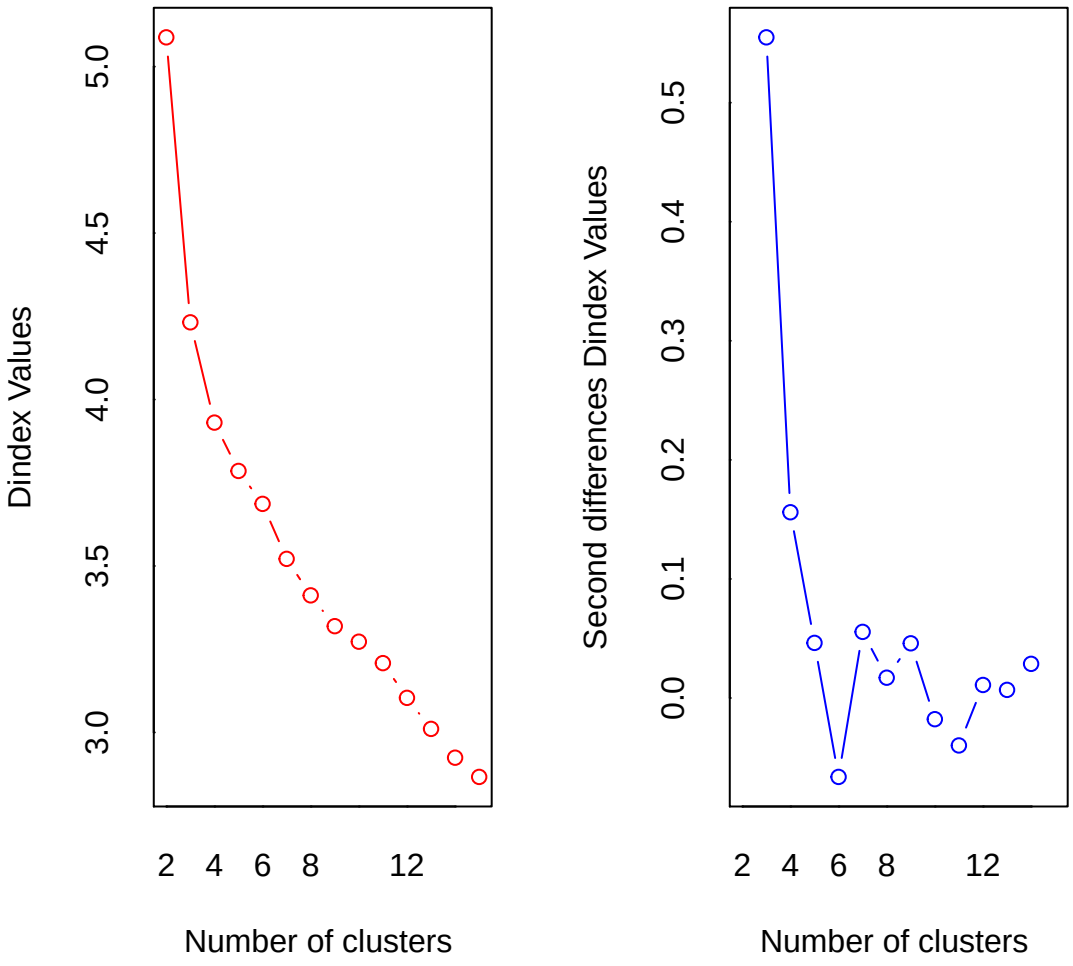
ii) How many clusters seem to be appropriate based on the plot?

Answer 6 clusters, since the elbow of the plot seems to be at the 6th cluster.

iii) Now use `NbClust` to take a majority vote on the best number of clusters by examining a multitude of index criteria. Does the majority vote match your answer in the previous part?




```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##                 In the plot of Hubert index, we seek a significant knee that corresponds to a
##                 significant increase of the value of the measure i.e the significant peak in Hubert
##                 index second differences plot.
##
```

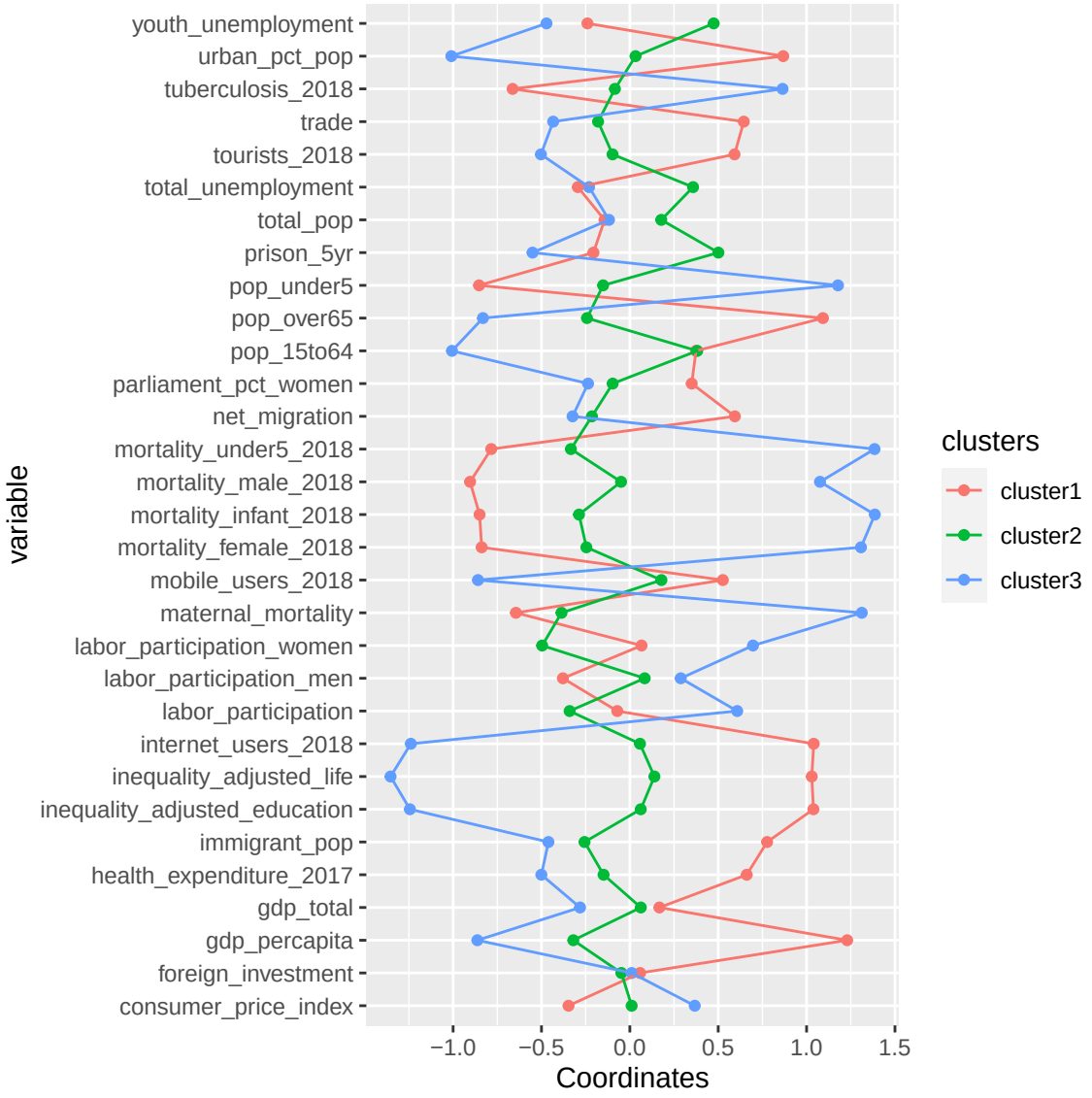



```
## *** : The D index is a graphical method of determining the number of clusters.
##                 In the plot of D index, we seek a significant knee (the significant peak in Dindex
##                 second differences plot) that corresponds to a significant increase of the value of
##                 the measure.
##
## *******************************************************************
## * Among all indices:
## * 4 proposed 2 as the best number of clusters
## * 13 proposed 3 as the best number of clusters
## * 2 proposed 4 as the best number of clusters
## * 1 proposed 10 as the best number of clusters
## * 1 proposed 13 as the best number of clusters
## * 1 proposed 14 as the best number of clusters
## * 2 proposed 15 as the best number of clusters
##
##                     ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  3
##
##
## *******************************************************************
```

No, the majority vote of 3 clusters did not support my previous answer for 6 clusters.

Question 2 (b). Cluster centers.

i) Compute k-means clusters using the value of k identified in (ii). Plot the centroid coordinate per variable for each cluster centroid. (The 'centroid coordinate' for a variable is the value of that variable at the cluster center.) This should look very much like a loading plot.

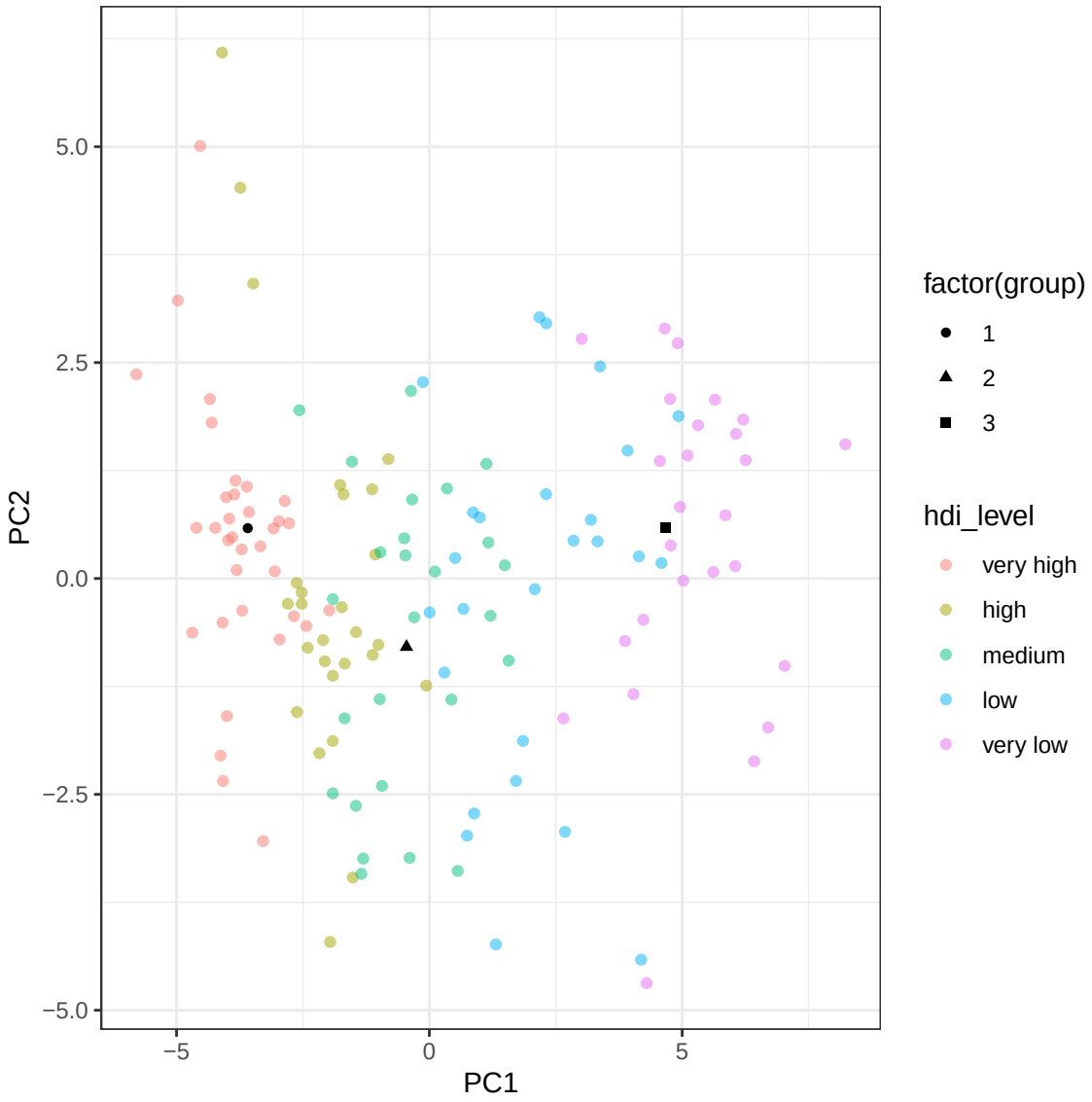


ii) Visually, which variables seem to be dimensions along which the cluster centers are most separated?

Answer The variables that seem to be the most separated along the cluster centers are mortality_under5_2018, internet_users_2018,inequality_adjusted_life, and inequaility_adjusted_education.

Question 2 (c). Cluster visualization.

i) Project the k-means cluster centers onto the first two principal components and plot the data together with the cluster centers. Display the HDI level using the color aesthetic, as before, and add a shape aesthetic to show the cluster assignment.



ii) Based on their approximate correspondence to the HDI levels, how would you interpret each cluster in terms of HDI?

Cluster 1 has very high hdi levels, Cluster 2 has high and medium hdi levels, and Cluster 3 has low and very low hdi levels.

Part 3: Interpretation

Question 3 (a). Clusters and HDI level.

i) Re-examine the plot of centroid coordinates with the approximate HDI level for each cluster in mind. Describe the characteristics of the average high-HDI country relative to the global average based on the centroid coordinates for the highest-HDI cluster: which variables are above or below average?

Answer From the plot of centroid coordinates, we can see that the high HDI level country is within Cluster 1. Moreover, we see that for Cluster 1, there are low female, male, infant, under 5 years mortality rates. On the other hand, there are high internet users, inequality adjusted life, and inequality adjusted education.

ii) Describe the characteristics of the average low-HDI country relative to the global average. Which variables are above or below average?

Answer From the plot of centroid coordinates, we can see that the low HDI level country is within Cluster 3. Moreover, we see that for Cluster 3, there are low internet users, inequality adjusted life, and inequality adjusted education. On the other hand, there are high under 5 and infant mortality with high maternal morality.

Question 3 (b). Summary.

Reflect on your results. Overall, which variables seem to be the strongest drivers of human development? You can reference any of the results above that strike you as important in answering the question. Answer in 2-4 sentences.

Answer I think the variables in PC1, specifically the four mortality rate variables, 2 adjusted inequality variables, and the gdp per capital variable, seem to be the strongest drivers of human development. I think this is a reasonable assumption because our loading plot in 1b(iii) showed this significance, which was then supported again by the centroid plot in 2b(ii) and the cluster plot in 2c(i). Moreover, I think this it is important to note that our findings highlight the connection between inequality, money, and death.

Codes

```
library(tidyverse)
library(NbClust)
knitr::opts_chunk$set(echo=F,
                      eval=T,
                      cache=T,
                      results='markup',
                      message=F,
                      warning=F,
                      fig.height=6,
                      fig.width=6,
                      fig.align='center')
# import 2019 HDI data
load('data/hdi.RData')
# create hdi factor
hdi_level <- cut(hdi$hdi_rank, breaks = 5, labels = c("very high", "high","medium","low","very low"))
# extract features and center and scale
x_mx <- hdi %>%
  select(-c('hdi_rank', 'country')) %>%
  scale(center = T, scale = T)

# compute SVD
x_svd <- svd(x_mx)

# get loadings
v_svd <- x_svd$v

# compute PCs
z_mx <- x_mx %*% x_svd$v

# pca scatterplot
z_mx[, 1:2] %>%
  as.data.frame() %>%
  rename(PC1 = V1, PC2 = V2) %>%
  ggplot(aes(x = PC1, y = PC2)) +
  geom_point(aes(color = hdi_level), alpha = 0.5) +
  theme_bw()
loading_plot <- v_svd[, 1:4] %>%
  as.data.frame() %>%
  rename(PC1 = V1, PC2 = V2, PC3 = V3, PC4 = V4) %>%
  mutate(variable = colnames(x_mx)) %>%
  gather(key = 'PC', value = 'Loading', 1:4) %>%
  arrange(variable) %>%
  ggplot(aes(x = variable, y = Loading)) +
  geom_point(aes(shape = PC, color = PC)) +
  theme_bw() +
  geom_hline(yintercept = 0, color = 'black') +
  geom_path(aes(linetype = PC, group = PC, color = PC)) +
  theme(axis.text.x = element_text(angle = 90), axis.text.y = element_text(size = 6)) +
  labs(x = '')
loading_plot+coord_flip()+facet_wrap(PC~.)
# compute PC variances
pc_vars <- x_svd$d^2/(nrow(x_mx) - 1)
```

```
# scree and cumulative variance plots
tibble(PC = 1:min(dim(x_mx)),
       Proportion = pc_vars/sum(pc_vars),
       Cumulative = cumsum(Proportion)) %>%
  gather(key = 'measure', value = 'Variance Explained', 2:3) %>%
  ggplot(aes(x = PC, y = `Variance Explained`)) +
  geom_point() +
  geom_path() +
  facet_wrap(~ measure) +
  theme_bw() + theme(axis.text.x = element_text(size = 4))+
  scale_x_continuous(breaks = 1:32, labels = as.character(1:32))
# sse vs k for k-means clustering
k_seq = 2:10
sse = sapply(k_seq, function(k){
  kmeans(x_mx, centers = k)$tot.withinss
}
)

data.frame(k=k_seq, SSE = sse) %>%
  ggplot(aes(x = k, y = SSE)) +
  geom_line(color = 'red') +
  geom_point(size = 2) +
  theme_bw()
nb_out = NbClust(x_mx, method ='kmeans')
# compute clusters
kmeans_out <- kmeans(x_mx, centers = 3, nstart = 5)
clusters <- factor(kmeans_out$cluster,
                   labels = paste('cluster', 1:3))

# obtain centriods
centers <- kmeans_out$centers

# plot centroid coordinates against variable
centers %>%
  t() %>%
  as.data.frame() %>%
  rename(cluster1 = 1, cluster2 = 2, cluster3 = 3) %>%
  mutate(variable = colnames(centers))%>%
  pivot_longer(cluster1:cluster3, names_to = 'clusters', values_to = 'Coordinates') %>%
  ggplot(aes(x = Coordinates, y = variable, colour = clusters)) +
  geom_point() +
  geom_line(aes(group = clusters), orientation ='y')
p = z_mx[, 1:2] %>%
  as.data.frame() %>%
  rename(PC1 = V1, PC2 = V2) %>%
  ggplot(aes(x = PC1, y = PC2)) +
  geom_point(aes(color = hdi_level), alpha = 0.5) +
  theme_bw()
center = centers %*% v_svd[, 1:2] %>% as.data.frame()
colnames(center) <- paste('PC', 1:2, sep = '')
center$group = (1:3)
p+geom_point(aes(shape = factor(group)),data=center)
```