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.

```r
# 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<-hdi%>%
  mutate(hdi_level=cut(hdi_rank, breaks=5, labels=rev(c("very low",
"low", "medium", "high", "very high"))))
```

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

#1-34 is very high, 35-61 is high, 62-87 is medium, 88-113 is low, 114-139 is very low

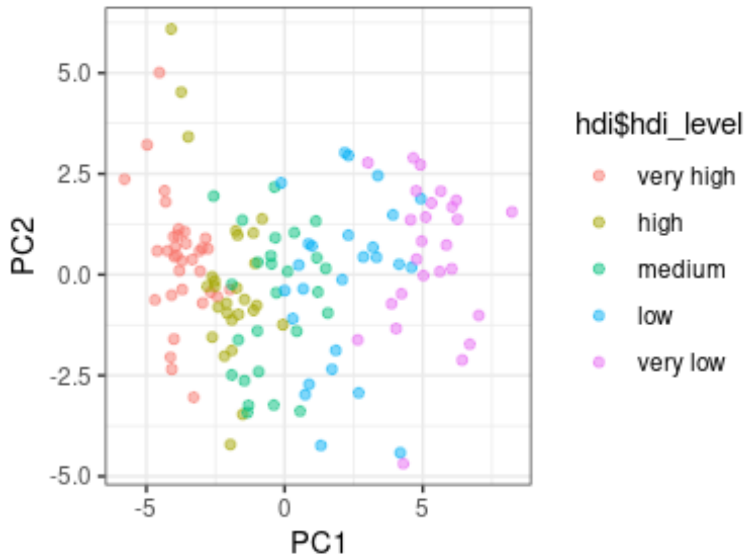
Question 1 (b). Exploratory analysis via PCA.

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.

```
# pca scatterplot
x_mx <- hdi %>%
  select(-c('hdi_rank', 'hdi_level',"country" )) %>%
  scale(center = T, scale = T)

x_svd <- svd(x_mx)
v_svd <- x_svd$v
z_mx <- x_mx %*% x_svd$v
pc_vars <- x_svd$d^2/(nrow(x_mx) - 1)
z_vars <- cov(z_mx) %>% diag()
cbind(z_vars, pc_vars)

##              z_vars     pc_vars
##  [1,] 11.520641018 11.520641018
##  [2,]  3.519292849  3.519292849
##  [3,]  2.335083050  2.335083050
##  [4,]  1.987995599  1.987995599
##  [5,]  1.618443836  1.618443836
```

```
##  [6,]  1.478049381  1.478049381
##  [7,]  1.243235880  1.243235880
##  [8,]  0.967035762  0.967035762
##  [9,]  0.874200000  0.874200000
## [10,]  0.753322439  0.753322439
## [11,]  0.657927467  0.657927467
## [12,]  0.612663838  0.612663838
## [13,]  0.480752066  0.480752066
## [14,]  0.449728154  0.449728154
## [15,]  0.438152176  0.438152176
## [16,]  0.399349784  0.399349784
## [17,]  0.327675071  0.327675071
## [18,]  0.280122273  0.280122273
## [19,]  0.222866810  0.222866810
## [20,]  0.186452296  0.186452296
## [21,]  0.124751355  0.124751355
## [22,]  0.115080423  0.115080423
## [23,]  0.095658545  0.095658545
## [24,]  0.088460806  0.088460806
## [25,]  0.074471889  0.074471889
## [26,]  0.067111684  0.067111684
## [27,]  0.039030387  0.039030387
## [28,]  0.022048932  0.022048932
## [29,]  0.011874849  0.011874849
## [30,]  0.005078442  0.005078442
## [31,]  0.003442939  0.003442939
```

```
z_mx[, 1:31] %>%
  as.data.frame() %>%
  rename(PC1 = V1, PC2 = V2) %>%
  bind_cols(select(hdi, hdi_rank, hdi_level, country)) %>%
  ggplot(aes(x = PC1, y = PC2)) +
  geom_point(aes(color = hdi$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?

"Very high" seems to be the most separated if anything. "Very low" seems to be a little separated more than the middle 3 ranks, which are not super separated at all.

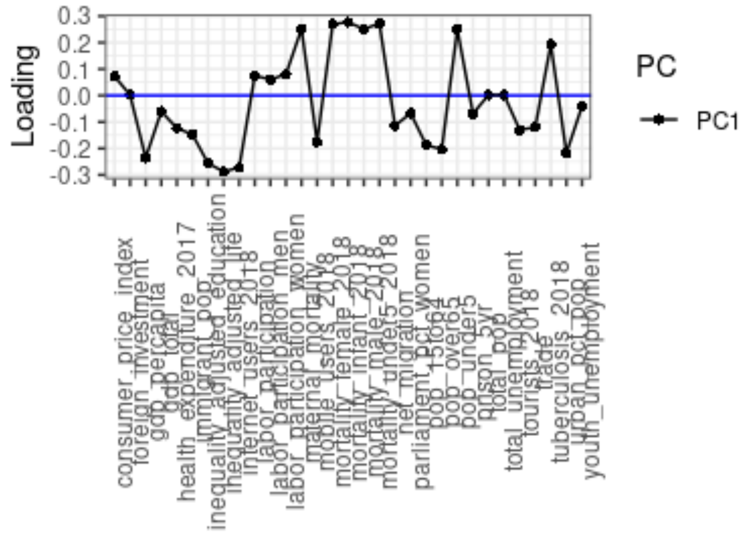
iii) Plot the loadings for the first four PCs. For each loading plot, comment on the following:

- 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 principal component?

PC1

```
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) %>%
  arrange(variable) %>%
  ggplot(aes(x = variable, y = Loading)) +
```

```
  geom_point(aes(shape = PC)) +
  theme_bw() +
  geom_hline(yintercept = 0, color = 'blue') +
  geom_path(aes(linetype = PC, group = PC)) +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(x = '')
```

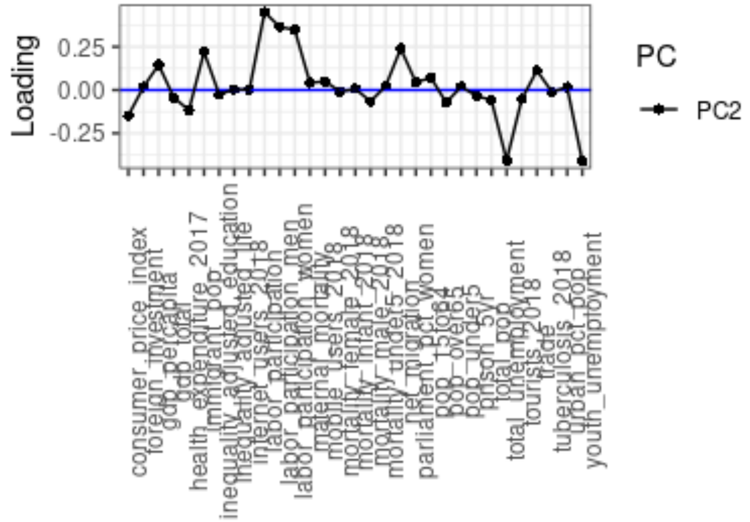


#mortality rates in general are the most influential (maternal, infants, male, under 5 years old) as well as population under 5 years old which is actually affected by 5-year old mortality rate from before. We can see that countries with lower inequality, mobile users, gdp, are where the mortality rates are higher.

PC2

```
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', 2) %>%
  arrange(variable) %>%
  ggplot(aes(x = variable, y = Loading)) +
  geom_point(aes(shape = PC)) +
  theme_bw() +
  geom_hline(yintercept = 0, color = 'blue') +
  geom_path(aes(linetype = PC, group = PC)) +
```

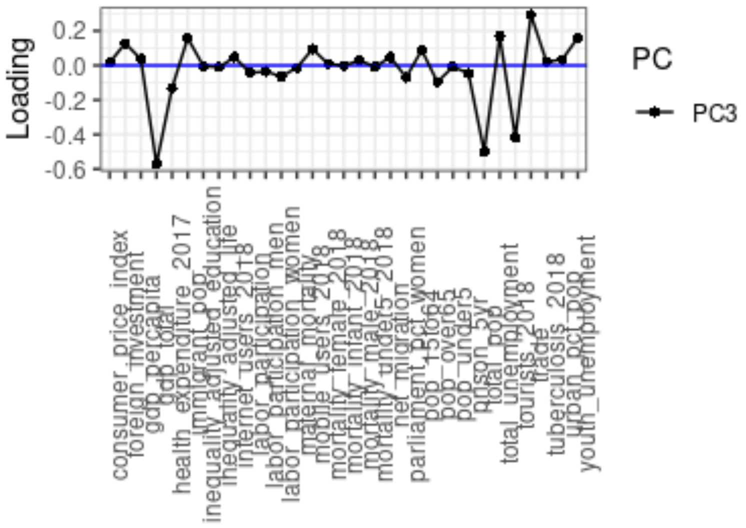
```
  theme(axis.text.x = element_text(angle = 90)) +
  labs(x = '')
```



#labor participation amond women, men and overall, as well as immigration/migration are most influential. Lower unemployment and youth unemployment affects immigration and net migration.

PC3

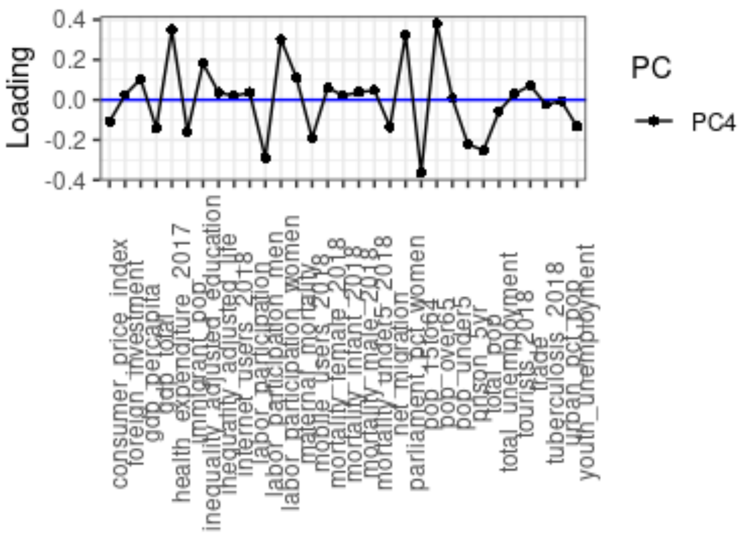
```
v_svd[, 1:4] %>%
  as.data.frame() %>%
  rename(PC1 = V1, PC2=V2,PC3=V3, PC4=V4) %>%
  mutate(variable = colnames(x_mx)) %>%
  gather(key = 'PC', value = 'Loading', 3) %>%
  arrange(variable) %>%
  ggplot(aes(x = variable, y = Loading)) +
  geom_point(aes(shape = PC)) +
  theme_bw() +
  geom_hline(yintercept = 0, color = 'blue') +
  geom_path(aes(linetype = PC, group = PC)) +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(x = '')
```



#trade and unemployment are most influential. The PC shows the relationship lower gdp and low populations having an influence higher trade and unemployment.

PC4

```
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',4) %>%
  arrange(variable) %>%
  ggplot(aes(x = variable, y = Loading)) +
  geom_point(aes(shape = PC)) +
  theme_bw() +
  geom_hline(yintercept = 0, color = 'blue') +
  geom_path(aes(linetype = PC, group = PC)) +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(x = '')
```



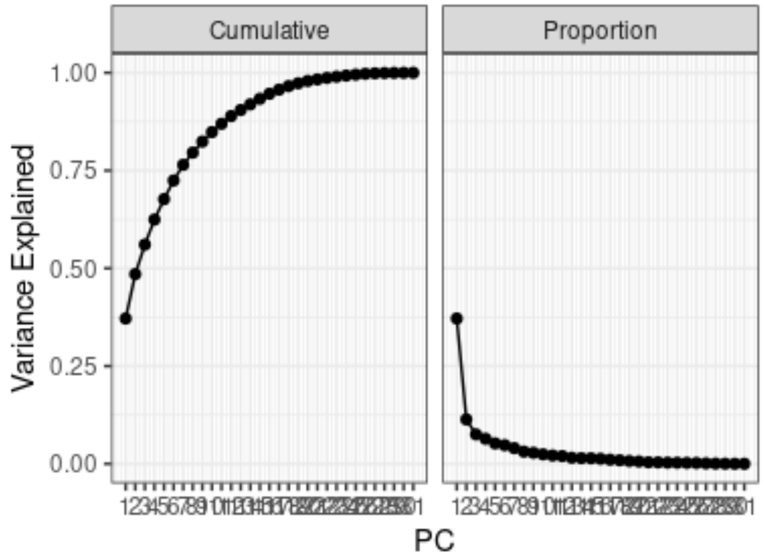
#pop over 65, women labor participation, health expenditure seem important for PC4. We see its relationship with women in parliament and gdp totals.

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?

#it seems like mortality rates, labor participation are some strong variables to determine hdi

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

```
# 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() +
  scale_x_continuous(breaks = 1:31, labels = as.character(1:31))
```



#it looks about 63% (anywhere from 60-65%)

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?

#The PC's with the highest proportion values will be the strongest drivers of variation which, similar to the previous question, seem to be labor participation rates and mortality rates, also among infants.

Part 2: Clustering with k-means

Question 2 (a). Choosing K.

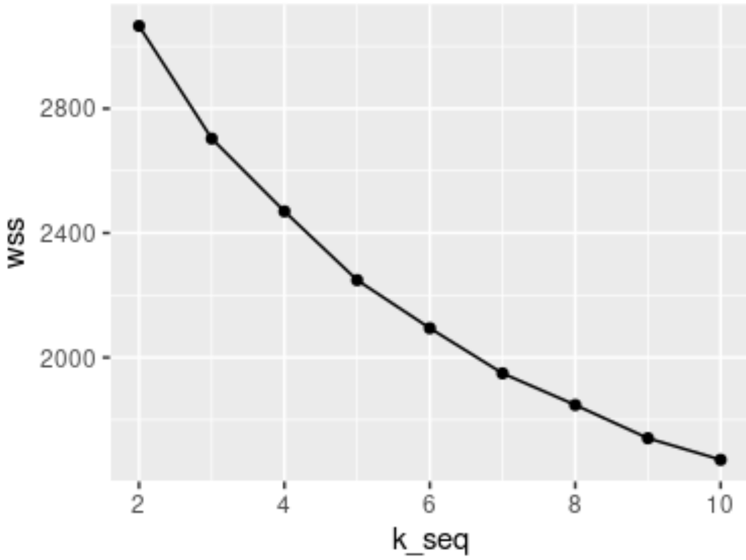
i) Compute SSE for k-means clustering of the centered and scaled data matrix for $k = 2, 3, \ldots, 10$ and plot SSE against the number of clusters k.

```
# sse vs k for k-means clustering
kmeans_out <- kmeans(x_mx, centers = 3, nstart = 5)
str(kmeans_out)
```

```
## List of 9
##  $ cluster     : int [1:139] 3 3 3 3 3 3 3 3 3 3 ...
##  $ centers     : num [1:3, 1:31] -1.354 0.139 1.029 -1.244
0.062 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:3] "1" "2" "3"
##   .. ..$ : chr [1:31] "inequality_adjusted_life"
"inequality_adjusted_education" "maternal_mortality"
"parliament_pct_women" ...
##  $ totss       : num 4278
##  $ withinss    : num [1:3] 716 1136 852
##  $ tot.withinss: num 2703
##  $ betweenss   : num 1575
##  $ size        : int [1:3] 38 59 42
##  $ iter        : int 3
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

```
clusters <- factor(kmeans_out$cluster,
                   labels = paste('cluster', 1:3))
centers <- kmeans_out$centers


k_seq <- 2:10
set.seed(22021)
wss <- sapply(k_seq, function(k){
  kmeans(x_mx,
         centers = k,
         nstart = 5,
```

```
         iter.max = 15)$tot.withinss
})

wss<-as.data.frame(wss)
wss%>%
  mutate(k=k_seq)%>%
  ggplot(aes(x=k_seq, y=wss))+
  geom_point()+
  geom_line()
```

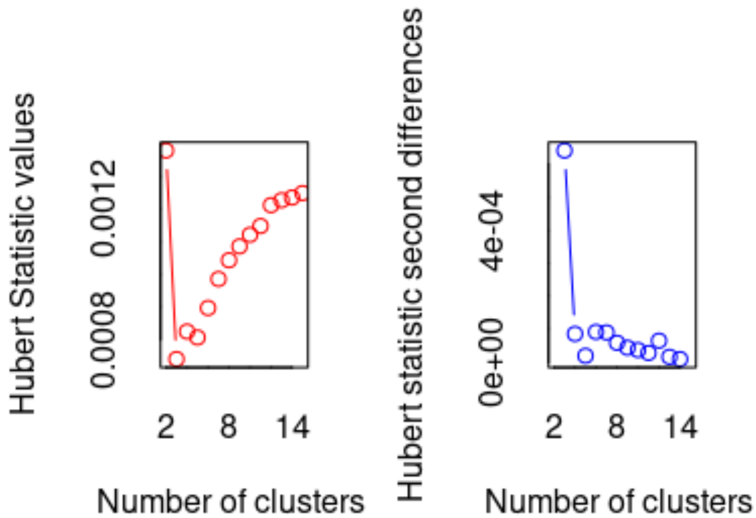


ii) How many clusters seem to be appropriate based on the plot?

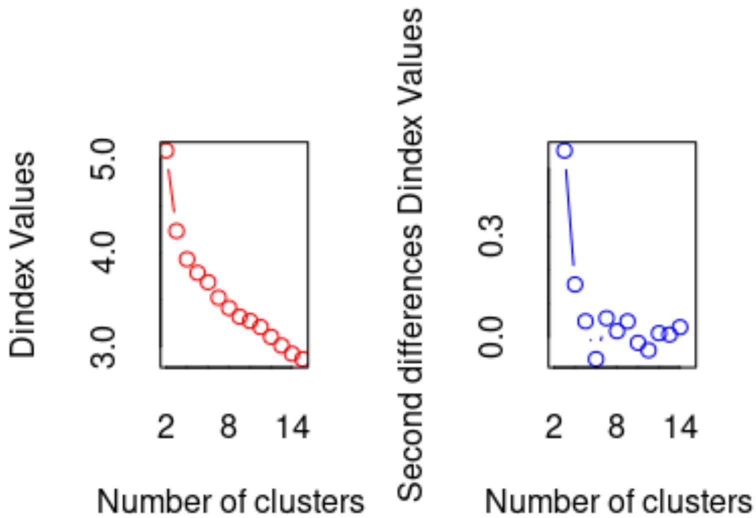
#I would say that 3 clusters seem appropriate.

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?

```
nb_out <- NbClust(x_mx, method = 'kmeans')
```




```
## *** : 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
##
```

```
##
## ********************************************************************
```

#it saays that the majority rules determines 3 as the best number of clusters, which is in line with what i said on the previous part

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.

```
d_mx <- dist(x_mx, method = 'euclidean')
hclust_out <- hclust(d_mx, method = 'complete')

clusters <- cutree(hclust_out, k = 3) %>%
  factor(labels = paste('cluster', 1:3))

tibble(clusters) %>% count(clusters)
```

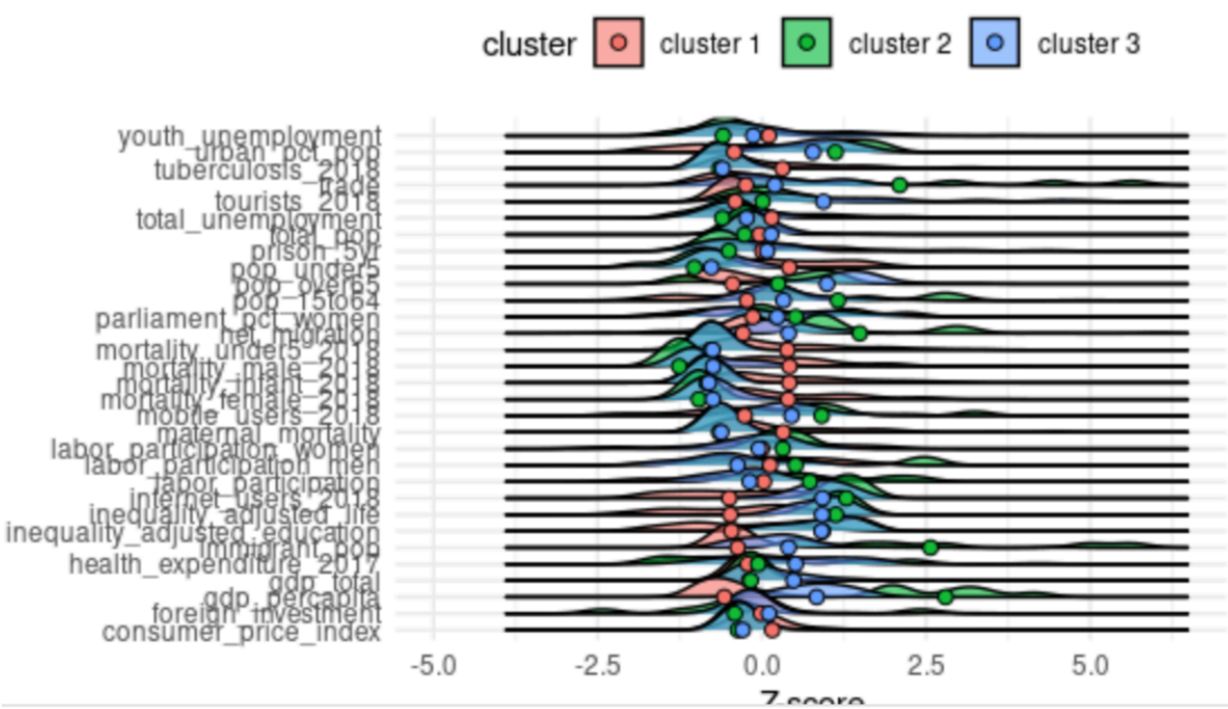
```
## # A tibble: 3 x 2
##   clusters      n
##   <fct>     <int>
## 1 cluster 1    90
## 2 cluster 2     3
## 3 cluster 3    46
```

```
hclust_out <- hclust(d_mx, method = 'ward.D')

# obtain centroids
centers<-kmeans_out$centers
centers
```

```
##   inequality_adjusted_life inequality_adjusted_education
maternal_mortality
## 1               -1.3537697                   -1.24435557
1.3123076
## 2                0.1390938                    0.06195072
-0.3864583
## 3                1.0294456                    1.03881951
-0.6444440
```

```
##   parliament_pct_women labor_participation_women
labor_participation_men
## 1           -0.2363592                 0.6968554
0.28804702
## 2           -0.0970397                -0.4960288
0.08407314
## 3            0.3501665                 0.0663142
-0.37871672
##    total_pop urban_pct_pop pop_under5 pop_15to64 pop_over65
## 1 -0.1179695   -1.00982121  1.1776326 -1.0062187 -0.8311210
## 2  0.1778184    0.03254266 -0.1511207  0.3815607 -0.2424578
## 3 -0.1430583    0.86793306 -0.8531885  0.3743864  1.0925621
##   mortality_infant_2018 mortality_under5_2018 mortality_female_2018
## 1             1.3862520             1.3845078             1.3078530
## 2            -0.2876487            -0.3334371            -0.2457109
## 3            -0.8501501            -0.7842502            -0.8381303
##   mortality_male_2018 tuberculosis_2018 health_expenditure_2017
gdp_total
## 1          1.07600615        0.86415226              -0.5004562
-0.28165484
## 2         -0.04961685       -0.08433544              -0.1485769
0.06227201
## 3         -0.90382951       -0.66338083               0.6615089
0.16735323
##   gdp_percapita consumer_price_index labor_participation
total_unemployment
## 1    -0.8626266           0.36732290          0.60753639
-0.2305256
## 2    -0.3197559           0.01032677         -0.34070255
0.3578218
## 3     1.2296526          -0.34684642         -0.07106981
-0.2940837
##   youth_unemployment prison_5yr      trade foreign_investment
net_migration
## 1         -0.4715060 -0.5505621 -0.4338087         0.01026396
-0.3239292
## 2          0.4744619  0.5011165 -0.1795821        -0.04779579
-0.2143893
## 3         -0.2399054 -0.2058217  0.6447636         0.05785526
```

```
0.5942447
##   immigrant_pop tourists_2018 internet_users_2018 mobile_users_2018
## 1    -0.4605910   -0.50307703         -1.23898726        -0.8590820
## 2    -0.2562589   -0.09768065          0.05741199         0.1786764
## 3     0.7767079    0.59238298          1.04033829         0.5262668

# plot centroid coordinates against variable
x_mx %>%
  scale() %>%
  as_data_frame() %>%
  mutate(state = rownames(x_mx),
         cluster = = factor(clusters,
                            labels = paste('cluster', 1:3))) %>%
  gather(key = 'variable', value = 'value', 1:31) %>%
  ggplot(aes(x = value,  y = variable)) +
  geom_density_ridges(aes(fill = cluster), alpha = 0.6) +
  theme_minimal() +
  labs(x = 'Z score', y = '')
```



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

#I plotted the ridges, but i dont know how to single out the actual centers. But this still allows me to see the clusters and the centers regardless. Total pop, gdp total and per capita, internet users, mortality rates and inequality rates were most separated.

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?

```
#Z <- scale(x_mx) %*% svd(scale(x_mx))$v[, 1:2]
#colnames(Z) <- paste('PC', 1:2, sep = '')
#as.data.frame(Z) %>%
 # mutate(hdi_level = rownames(Z),
         #cluster = factor(clusters,
                          #labels = paste('cluster', 1:3))) %>%
  #ggplot(aes(x = PC1, y = PC2)) +
  #geom_point(aes(color = cluster)) +
  #theme_bw() +
  #geom_text(aes(label = state), size = 2.5, alpha = 0.5)

# add to plot
```

#i didn’t figure out this part but here’s what I put so far

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?

```
centers

##   inequality_adjusted_life inequality_adjusted_education
maternal_mortality
## 1               -1.3537697                   -1.24435557
1.3123076
## 2                0.1390938                    0.06195072
-0.3864583
## 3                1.0294456                    1.03881951
-0.6444440
##   parliament_pct_women labor_participation_women
labor_participation_men
## 1           -0.2363592                 0.6968554
0.28804702
## 2           -0.0970397                -0.4960288
0.08407314
## 3            0.3501665                 0.0663142
-0.37871672
##    total_pop urban_pct_pop pop_under5 pop_15to64 pop_over65
## 1 -0.1179695   -1.00982121  1.1776326 -1.0062187 -0.8311210
## 2  0.1778184    0.03254266 -0.1511207  0.3815607 -0.2424578
## 3 -0.1430583    0.86793306 -0.8531885  0.3743864  1.0925621
##   mortality_infant_2018 mortality_under5_2018 mortality_female_2018
## 1             1.3862520             1.3845078             1.3078530
## 2            -0.2876487            -0.3334371            -0.2457109
## 3            -0.8501501            -0.7842502            -0.8381303
##   mortality_male_2018 tuberculosis_2018 health_expenditure_2017
gdp_total
## 1          1.07600615        0.86415226              -0.5004562
-0.28165484
## 2         -0.04961685       -0.08433544              -0.1485769
0.06227201
```

```
## 3         -0.90382951       -0.66338083               0.6615089
0.16735323
##   gdp_percapita consumer_price_index labor_participation
total_unemployment
## 1    -0.8626266           0.36732290          0.60753639
-0.2305256
## 2    -0.3197559           0.01032677         -0.34070255
0.3578218
## 3     1.2296526          -0.34684642         -0.07106981
-0.2940837
##   youth_unemployment prison_5yr      trade foreign_investment
net_migration
## 1         -0.4715060 -0.5505621 -0.4338087         0.01026396
-0.3239292
## 2          0.4744619  0.5011165 -0.1795821        -0.04779579
-0.2143893
## 3         -0.2399054 -0.2058217  0.6447636         0.05785526
0.5942447
##   immigrant_pop tourists_2018 internet_users_2018 mobile_users_2018
## 1    -0.4605910   -0.50307703         -1.23898726        -0.8590820
## 2    -0.2562589   -0.09768065          0.05741199         0.1786764
## 3     0.7767079    0.59238298          1.04033829         0.5262668
```

#High level hdi countries experience lowest adjusted inequality factors for life and for education, lowest mortality for all applicable groups, as well as higher than average internet users and pop over 65.

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

#low hdi compared to average have significantly lower rates internet users, health expenditures and gdp per capita. Higher female mortality probably from birth, higher inequality, as well as more labor participation especially with women were observed.

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.

#I would say that mortality rates and inequality rates are the strongest drivers of human development. We see this by looking at the PC loadings from question 1, particularly PC1 but also the others. It also shows when graphing the clusters (and centroids) in question 2 as well, where we saw that highest countries experiences lower rates of those factors, in every sense, and for low countries it was the opposite.

Codes

```
library(tidyverse)
library(NbClust)
knitr::opts_chunk$set(echo=T,
                      eval=T,
                      cache=T,
                      results='markup',
                      message=F,
                      warning=F,
                      fig.height=3,
                      fig.width=4,
                      fig.align='center')
# import 2019 HDI data
load('data/hdi.RData')
# create hdi factor
hdi<-hdi%>%
  mutate(hdi_level=cut(hdi_rank, breaks=5, labels=rev(c("very low",
"low", "medium", "high", "very high"))))

# pca scatterplot
x_mx <- hdi %>%
  select(-c('hdi_rank', 'hdi_level',"country" )) %>%
  scale(center = T, scale = T)

x_svd <- svd(x_mx)
v_svd <- x_svd$v
z_mx <- x_mx %*% x_svd$v
pc_vars <- x_svd$d^2/(nrow(x_mx) - 1)
z_vars <- cov(z_mx) %>% diag()
cbind(z_vars, pc_vars)

z_mx[, 1:31] %>%
  as.data.frame() %>%
  rename(PC1 = V1, PC2 = V2) %>%
  bind_cols(select(hdi, hdi_rank, hdi_level, country)) %>%
  ggplot(aes(x = PC1, y = PC2)) +
  geom_point(aes(color = hdi$hdi_level), alpha = 0.5) +
```

```r
  theme_bw()


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) %>%
  arrange(variable) %>%
  ggplot(aes(x = variable, y = Loading)) +
  geom_point(aes(shape = PC)) +
  theme_bw() +
  geom_hline(yintercept = 0, color = 'blue') +
  geom_path(aes(linetype = PC, group = PC)) +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(x = '')
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', 2) %>%
  arrange(variable) %>%
  ggplot(aes(x = variable, y = Loading)) +
  geom_point(aes(shape = PC)) +
  theme_bw() +
  geom_hline(yintercept = 0, color = 'blue') +
  geom_path(aes(linetype = PC, group = PC)) +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(x = '')
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', 3) %>%
  arrange(variable) %>%
  ggplot(aes(x = variable, y = Loading)) +
  geom_point(aes(shape = PC)) +
```

```
  theme_bw() +
  geom_hline(yintercept = 0, color = 'blue') +
  geom_path(aes(linetype = PC, group = PC)) +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(x = '')
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',4) %>%
  arrange(variable) %>%
  ggplot(aes(x = variable, y = Loading)) +
  geom_point(aes(shape = PC)) +
  theme_bw() +
  geom_hline(yintercept = 0, color = 'blue') +
  geom_path(aes(linetype = PC, group = PC)) +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(x = '')
# 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() +
  scale_x_continuous(breaks = 1:31, labels = as.character(1:31))
# sse vs k for k-means clustering
kmeans_out <- kmeans(x_mx, centers = 3, nstart = 5)
str(kmeans_out)
clusters <- factor(kmeans_out$cluster,
                   labels = paste('cluster', 1:3))
centers <- kmeans_out$centers
```

```r
k_seq <- 2:10
set.seed(22021)
wss <- sapply(k_seq, function(k){
  kmeans(x_mx,
         centers = k,
         nstart = 5,
         iter.max = 15)$tot.withinss
})

wss<-as.data.frame(wss)
wss%>%
  mutate(k=k_seq)%>%
  ggplot(aes(x=k_seq, y=wss))+
  geom_point()+
  geom_line()

nb_out <- NbClust(x_mx, method = 'kmeans')
d_mx <- dist(x_mx, method = 'euclidean')
hclust_out <- hclust(d_mx, method = 'complete')

clusters <- cutree(hclust_out, k = 3) %>%
  factor(labels = paste('cluster', 1:3))

tibble(clusters) %>% count(clusters)
hclust_out <- hclust(d_mx, method = 'ward.D')

# obtain centroids
centers<-kmeans_out$centers
centers
# plot centroid coordinates against variable
x_mx %>%
  scale() %>%
  as_data_frame() %>%
  mutate(state = rownames(x_mx),
         cluster = = factor(clusters,
                                labels = paste('cluster', 1:3))) %>%
  gather(key = 'variable', value = 'value', 1:31) %>%
```

```
  ggplot(aes(x = value, y = variable)) +
  geom_density_ridges(aes(fill = cluster), alpha = 0.6) +
  theme_minimal() +
  labs(x = 'Z score', y = '')


#Z <- scale(x_mx) %*% svd(scale(x_mx))$v[, 1:2]
#colnames(Z) <- paste('PC', 1:2, sep = '')
#as.data.frame(Z) %>%
 # mutate(hdi_level = rownames(Z),
         #cluster = factor(clusters,
                          #labels = paste('cluster', 1:3))) %>%
  #ggplot(aes(x = PC1, y = PC2)) +
  #geom_point(aes(color = cluster)) +
  #theme_bw() +
  #geom_text(aes(label = state), size = 2.5, alpha = 0.5)

# add to plot

centers
```