DSC424 - Final Project Milestone 5
Name: Andrew Fitzsimmons
Project Team: Group 5 - World Healthcare Organization Data
Date: 06/02/2022

Overview of The Data

Our group chose to use a dataset from Kaggle that focused on World Healthcare Organization data from the year 2000 to the year 2016. The original dataset contained 3112 rows and 32 columns. The dataset contained features such as obesity levels, BMI, mortality rates, water accessibility, vaccination availability, gdp, population, etc. The data was grouped by region, country, and year. Our group's primary goals were to identify which features in the dataset contribute to overall life expectancy (target variable) the most, and to model the contributions of the factors to the separation of the data by each region/continent.

Initial Data Cleaning

I handled the majority of the data cleaning for our group that ended up being used in our group members' individual analyses. To briefly summarize, I generated code in R to handle NA values for the 'Alcohol' variable by populating missing values with the average value for the other years present for a given country. I also handled sparsely populated variables by identifying variables with a significant amount of missing values and removing them from the dataset, as there was not enough information in these variables to interpolate a value or generate an aggregate value to fill into the NA cells. Finally, I pared down the remaining dataset to only rows that represented complete cases. With this data cleaning completed, our dataset shrunk from 3112 rows, down to 1349. Our group made the decision that the remaining ~1300 rows of complete case data would be sufficient for our analysis.

Example R Code:

```
#replace alcohol NAs with avg of 'good' values by country
#(remove south sudan from data set due to no data on many predictors)

x='Canada'
WHO$alcohol = ifelse(WHO$country== x & is.na(WHO$alcohol),
                     mean(WHO[WHO$country == x ,]$alcohol,na.rm=TRUE),WHO$alcohol)
```

```
> WHO_Reduced = WHO%>%
+   select(-une_school, -une_literacy, -une_edu_spend, -une_poverty, -hospitals,-doctors)
> #reduce selection further by only reviewing complete entries in the reduced data set
> WHO_Reduced_CompleteCases = WHO_Reduced[complete.cases(WHO_Reduced),]
> nrow(WHO_Reduced_CompleteCases)
[1] 1349
```

Individual Analysis

I opted to perform exploratory linear discriminant analysis as my analysis pipeline for the final report. I chose to focus my analysis on the year 2013, since 2013 had one of the higher counts of complete case data, and represented the most recent year (compared to current day) in the data set. I selected 'region' as my target variable for LDA so I could see how the linear discriminants generated by LDA would separate the data with respect to 'region'. The prior probabilities in Figure 1 show that the data I selected for LDA is predominantly comprised of the regions of Africa, Americas, and Europe (combining for ~79% of the entries in the selected data). As a result, I feel the analysis may be biased towards Africa, Americas, and Europe, and may not properly account for unique qualities of regions of Eastern Mediterranean, South East Asia, and Western Pacific.

Example R Code:

```
#LDA
library(MASS)
library(DAAG)
library(ggplot2)

#remove unused variables and select year from dataframe
WHO_LDA_Data = WHO_multicolinearity_removed%>%
  filter(year=='2013')%>%
  dplyr::select(-country,-year)

#generate eploratory LDA object
WHO.LDA = lda(region~., data=WHO_LDA_Data)
WHO.LDA.values = predict(WHO.LDA)
WHO.LDA.values

LD1 = WHO.LDA.values$x[,1]
LD2 = WHO.LDA.values$x[,2]
Region =WHO.LDA.values$class
WHO.LDA.Plot = data.frame(LD1,LD2,Region)

ggplot(WHO.LDA.Plot,aes(x = LD1,y=LD2,col=Region))+geom_point()+
  ggtitle('Plot of LD1 and LD2, Categorized by Region')+theme_bw()
```

Figure 1

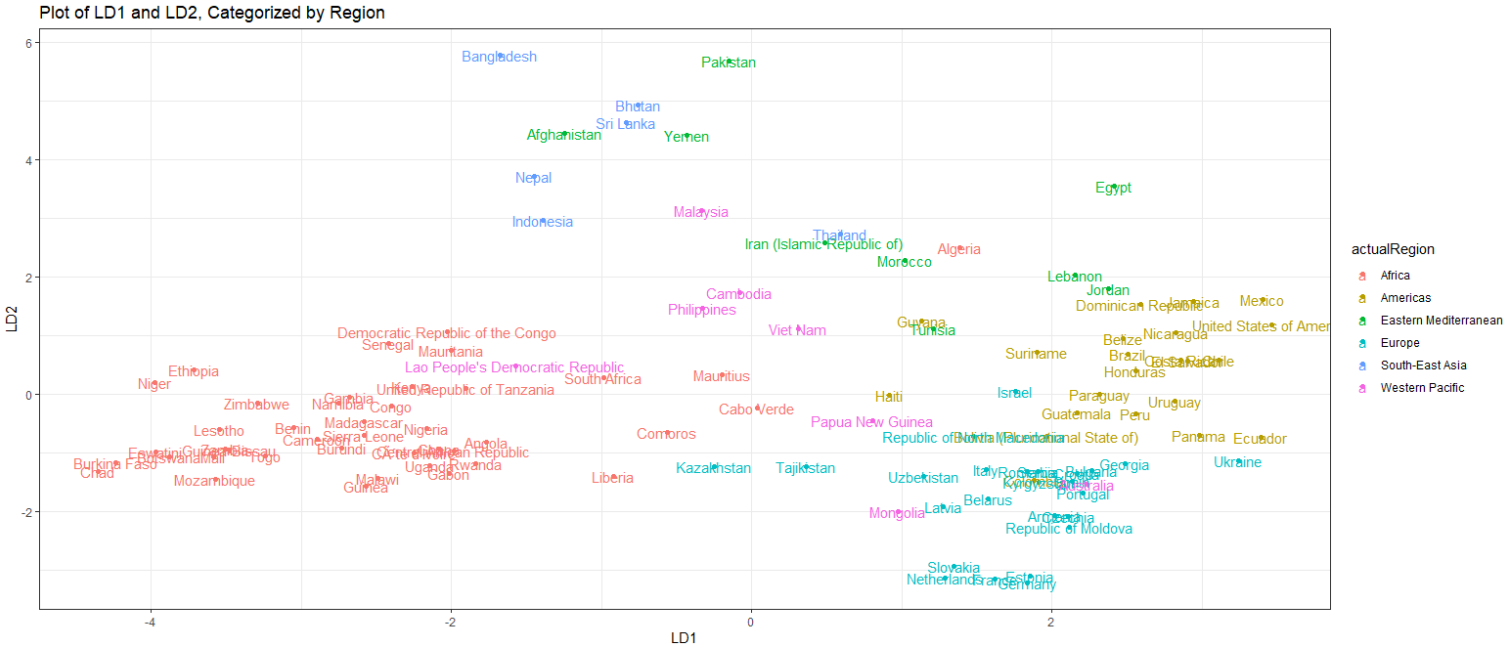
```
Prior probabilities of groups:
                Africa               Americas Eastern Mediterranean                 Europe
            0.37500000             0.19642857             0.08035714             0.22321429
      South-East Asia        Western Pacific
            0.05357143             0.07142857
```

```
Coefficients of linear discriminants:
                              LD1           LD2           LD3           LD4           LD5
life_expect          2.347627e-01  2.834536e-01  2.050261e-02  4.885458e-01  4.649437e-01
adult_mortality      9.364577e-03  1.494366e-02  5.170126e-03  3.254906e-02  2.316530e-02
infant_mort          3.821904e+01  1.233426e+01 -1.561947e+01  2.900045e+01 -1.896697e+01
age1.4mort          -2.844302e+02  4.643668e+00 -1.424382e+02  1.232000e+02  5.863196e+02
alcohol              8.262547e-02 -8.701166e-02  5.444990e-02  5.203839e-02 -6.654631e-02
bmi                  3.590068e-01  2.694606e-01 -2.468046e-01 -9.428115e-01  4.733357e-01
age5.19thinness     -2.656756e-02  4.866964e-01  8.948127e-02 -2.031188e-01  7.251652e-02
age5.19obesity       8.289647e-02  3.066654e-01 -2.409199e-01  2.109629e-01 -1.789627e-01
hepatitis           -1.351307e-02 -1.673191e-02  1.076620e-02  2.289745e-02 -3.130921e-02
measles              4.903422e-02  3.369998e-03  2.302825e-03  4.092261e-02 -1.249772e-01
polio               -6.649344e-02 -4.234596e-02  1.704499e-02 -5.050237e-02  6.170630e-03
diphtheria           6.135469e-03  8.671343e-02 -1.200472e-02  3.219474e-03  1.674225e-01
basic_water          1.382794e-02 -2.765408e-02  1.675208e-02  5.438283e-02  2.203143e-02
gni_capita          -7.507127e-05 -4.971412e-05  7.676715e-05 -9.565653e-05 -9.679231e-05
gghe.d               4.579813e-02 -1.105159e-01  3.465234e-02 -5.843688e-02  2.138605e-01
che_gdp              1.303245e-01 -9.219218e-02 -1.184915e-01 -1.096488e-01  1.313141e-02
une_pop              1.489286e-06  3.745641e-06 -8.873312e-07  5.117741e-06  9.256767e-06
une_hiv             -1.530435e-01 -4.352112e-02 -6.475504e-02 -3.380113e-02  1.186653e-03

Proportion of trace:
   LD1    LD2    LD3    LD4    LD5
0.5097 0.2976 0.1037 0.0606 0.0284
```

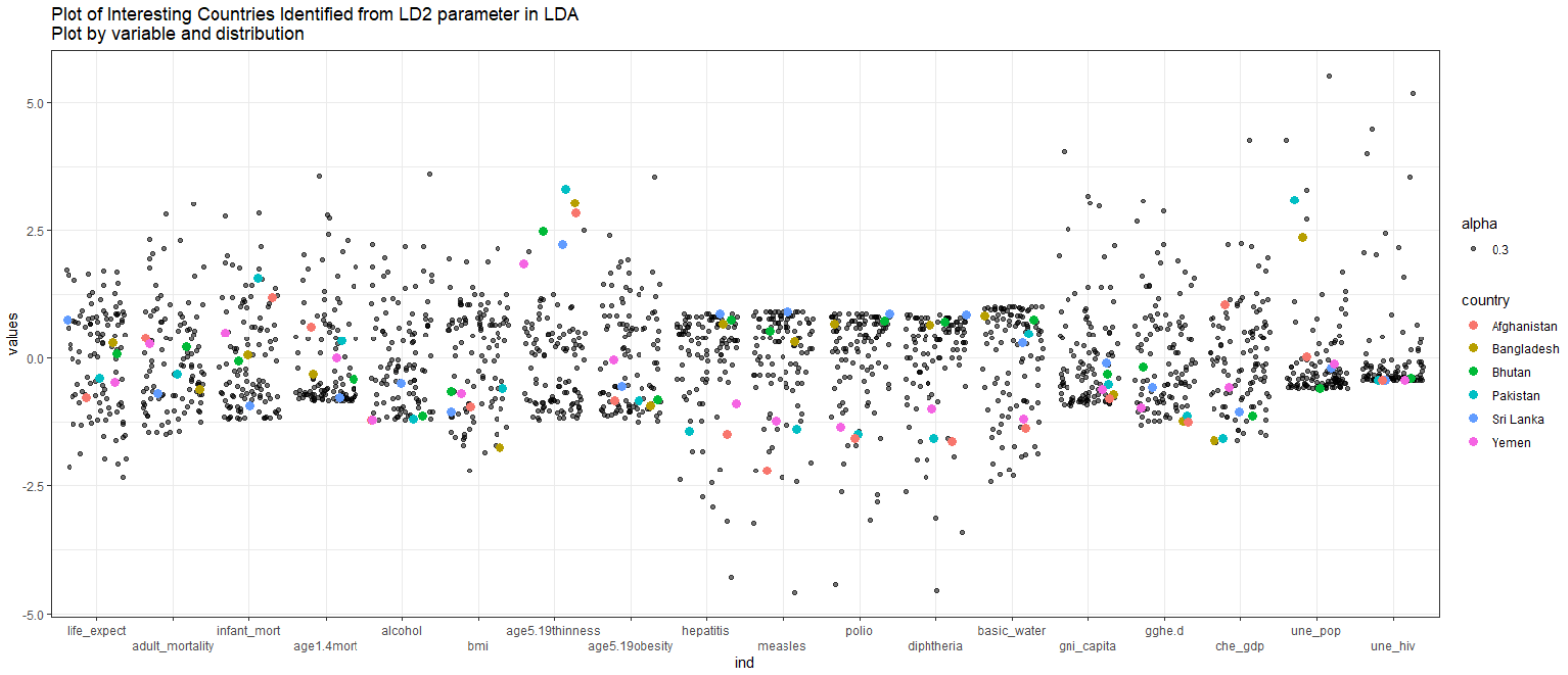
After running LDA on the dataset, it appears that LD1 and LD2 have accounted for 80% of the variation in the data. As a result, I felt that plotting the LDA values against LD1 and LD2 should provide a good visual representation of the separation and clustering of regions in the data. Figure 2 below shows a plot of LDA values with LD1 as the x axis, LD2 as the y axis, colored by region, and labeled by country. Although the data labels are a bit difficult to read in some cases, they are useful for this exploratory visualization for identifying countries that may be outliers that are worth looking into further. Africa, Europe, and the Americas all generated points that formed fairly clear clusters within the visual. Based on the visualization, it appears that there is some separation between Africa from Europe and the Americas in terms of the LD1 axis. Europe and the Americas are generally positive, and African countries are generally negative. After reviewing the coefficients of the linear discriminants, it appears that a more negative LD1 value is driven primarily by high age 1-4 mortality rates, high rates of thinness in ages 5-19, high rates of hepatitis, and high rates of HIV, indicating that these are all challenges within Africa's healthcare systems.

Figure 2



I also chose to look into Bangladesh, Pakistan, Bhutan, Afghanistan, Sri Lanka, and Yemen individually as they represented relative outliers on the high end of the LD2 axis. For this, I took a more visual analysis approach and plotted the z scores of each variable for each country. I then highlighted my specific countries of interest as colored dots to differentiate them from the remainder of the data. It appears that these countries of interest that showed high LD2 values have the highest rates of thinness in ages 5 - 19, among the lowest bmi, the lowest rates of HIV, among the lowest gni per capita and government healthcare spend, and some of the lowest rates of alcohol use. Life expectancy for these 6 countries were all within the middle 50% of the distribution for the life expectancy variable as well. The above relative rates of each variable align directly with the LD2 coefficients previously calculated in Figure 1. However, the visual representation of the data helped me better identify the similarities between these countries that resulted in their high LD2 values.

Figure 3


Plot of Interesting Countries Identified from LD2 parameter in LDA
Plot by variable and distribution
values
5.0
2.5
0.0
-2.5
-5.0
life_expect
adult_mortality
infant_mort
age1.4mort
alcohol
bmi
age5.19thinness
age5.19obesity
hepatitis
measles
polio
diphtheria
basic_water
gni_capita
gghe.d
che_gdp
une_pop
une_hiv
ind
alpha
0.3
country
Afghanistan
Bangladesh
Bhutan
Pakistan
Sri Lanka
Yemen