DSC 424 Final Project Report

Group 5: Matthew Ghuneim, Andrew Fitzsimmons, Smit Patel

Non-Technical Summary

Exploratory Data Analysis

We developed initial regression models to get an idea of the different relationships between variables and how strongly correlated some variables are. We reviewed basic statistics within each variable in order to identify potential outliers and missing data 'NA' values that would need to be handled in our data cleaning. We also generated a correlation plot in order to get some initial insight to how the variables in our dataset were related to one another, and how they impacted our target variable of Life Expectancy.

Regularized Regression

We developed several regression models to try and predict which set of features in our dataset had an effect on the life expectancy the most. We wanted to compare two sets of features for two different years, 2000 and 2013. After creating our models, the model has predicted that for the year 2000, higher obesity levels for children 5-19 and more basic water services in the respective country lead to a higher life expectancy, while a higher HIV rate for a country leads to a lower life expectancy. For the year 2013, we have predicted those same findings, however it also states that a lower mortality for a child leads to a lower life expectancy and a higher domestic general government health expenditure leads to a higher life expectancy.

Factor Analysis

We tried to identify which factors have the most dominant effect on life expectancy. We wanted to look at the factors that determine life expectancy in 2013. After looking at the number of factors to consider, we finalized 3 factors. Although, I have included the results and analysis for 5 factors. We wanted to focus on the factors that have the most effect and made sense in deciding life factors. For 2013, we saw that (for 3 factors) the immunization coverage, white collar working class and condition of underdeveloped countries; these are the factors that white the collaboration of different features make the accounts of life expectancy.

Linear Discriminant Analysis

Our focus for Linear Discriminant Analysis was to perform an exploratory LDA. The goal here was to investigate how well LDA separated clusters and groupings of countries, with the target variable being region. The linear discriminants helped us identify some latent relationships between various groups of regions based on how + or - the regions appeared on the plot of LD1 and LD2. LD1 and LD2 could then be reviewed to determine which variables within the linear discriminants had the most significant positive and negative impact on the LDA scores for a given country. These

combinations of highly weighted variables ultimately give us some insight as to which latent factors are influencing some separation and distinction between different countries within regions that makes those regions unique from the rest of the dataset. When reviewing LD1, we identified for example, that infant mortality rates, rates of thinness in adolescents, rates of hepatitis, and rates of HIV were some of the primary factors that differentiate African countries from Europe and the Americas. Additionally, in reviewing LD2, we noticed that there were a few individual countries that seemed to separate themselves from the rest of the data. Bangladesh, Pakistan, Bhutan, Afghanistan, Sri Lanka, and Yemen all appeared to be grouped in their own cluster with respect to LD2 compared to the rest of the data set. We took a deeper dive into what differentiates these countries from the rest and noticed that these countries had some of the highest rates of thinness in adolescents, some of the lowest bmi, lowest gni per capita, and lowest government healthcare spend. Despite these apparent challenges with these countries' healthcare systems, they all were in the middle 50% of the distribution in terms of life expectancy. It appears that this may be due in part to the extremely low levels of alcohol use that is common amongst all these countries.

Technical Summary

Exploratory Data Analysis

Initial regressions took place, and we implemented 2 regression models for 2 different years: 2000 and 2016. Both adjusted R-squared values in each of the models were extremely high, and several variables had extreme cases of multicollinearity. The correlation plot below helps us to identify potential multicollinearity between variables, since this dataset merged WHO data and UNESCO data, there were effectively multiple instances of the similar variables that needed to be handled and removed. We also reviewed the descriptive statistics for each variable in order to resolve missing values, and identify variables that needed to be removed from the dataset due to a high amount of missing values.

Let's look at the summary and some plots for the understanding of data:

Raw Counts

Name	Value
Rows	1,349
Columns	26
Discrete columns	3
Continuous columns	23
All missing columns	0
Missing observations	0
Complete Rows	1,349
Total observations	35,074
Memory allocation	257.3 Kb

Figure 1: Basic Statistics for the data.

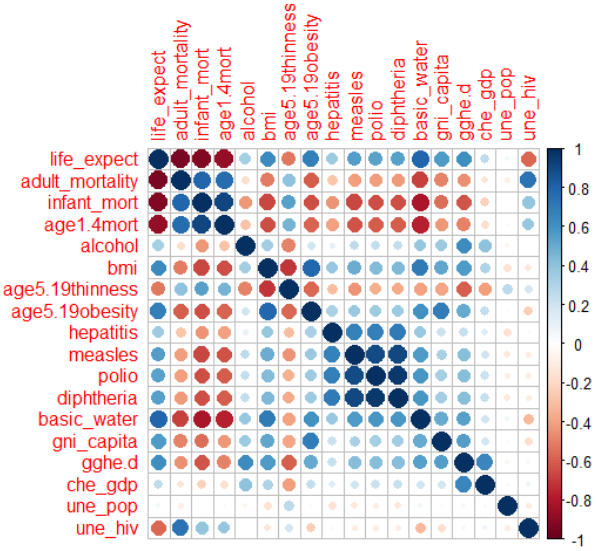


Figure 2: Corrplot of numeric variables of the data.

Regularized Regression

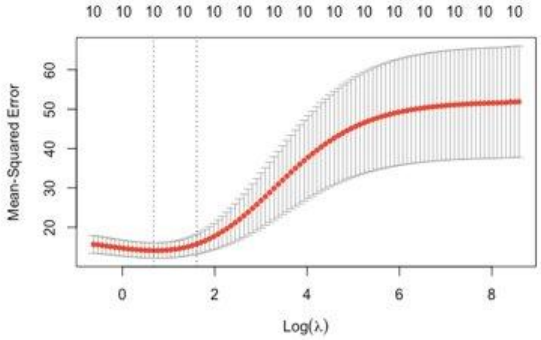
For our first line of analysis, we decided to look at multiple regularized regression models with life_expect as our parameter of interest. We wanted to compare 2 different models for 2 different years, 2000 and 2013, and see which variables contribute to the life expectancy variable. We started out by creating OLS models with all of the features, however many variables had extremely high multicollinearity, so we chose to manually remove each variable according to their VIF score. Each of the variables in the final models have VIF scores less than 5. Here are the final models for 2000 (left) and 2013 (right):

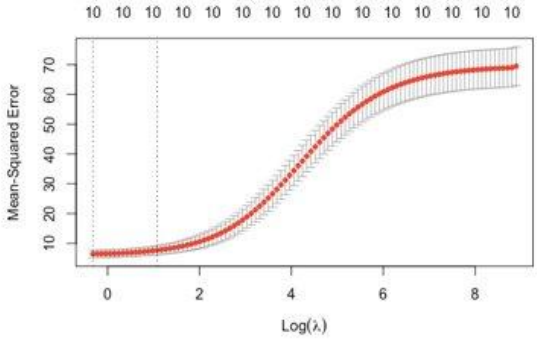
```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.166e+01  1.363e+01   3.056  0.00514 **
alcohol        1.912e-01  1.874e-01   1.020  0.31715
bmi            1.036e-01  6.365e-01   0.163  0.87196
age5.19obesity 7.711e-01  3.186e-01   2.421  0.02278 *
hepatitis     -5.057e-02  2.857e-02  -1.770  0.08843 .
measles        3.088e-02  9.704e-02   0.318  0.75286
polio          5.403e-02  8.181e-02   0.660  0.51477
basic_water    1.805e-01  5.844e-02   3.089  0.00473 **
che_gdp        3.985e-01  3.799e-01   1.049  0.30390
une_pop        9.845e-06  1.367e-05   0.720  0.47783
une_hiv       -5.919e-01  1.071e-01  -5.525 8.46e-06 ***
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.395e+01  2.734e+00  23.392  < 2e-16 ***
age1.4mort    -1.005e+03  1.153e+02  -8.716 6.35e-13 ***
alcohol       -2.594e-02  8.617e-02  -0.301   0.7642
age5.19thinness 9.235e-02 9.642e-02   0.958   0.3413
age5.19obesity 2.355e-01  1.015e-01   2.321   0.0231 *
hepatitis     -2.003e-03  1.917e-02  -0.104   0.9171
basic_water    5.782e-02  2.421e-02   2.389   0.0195 *
gghe.d         1.510e+00  2.290e-01   6.593 5.87e-09 ***
che_gdp       -1.507e-01  1.252e-01  -1.204   0.2324
une_pop       -4.910e-06  5.163e-06  -0.951   0.3448
une_hiv       -4.892e-01  5.253e-02  -9.312 4.85e-14 ***
```

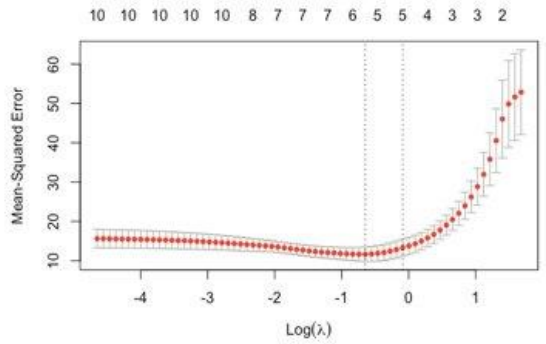
The adjusted R-squared values were 83.71% and 93.34% respectively. We evaluated this model and investigated for overfitting by splitting the data into training and test datasets (75% training and 25% test). The training and test RMSE values for 2000 were 2.50 and 4.64 respectively, and for 2013 the values were 2.01 and 2.68. These are symptoms of overfitting because for 2000, the test RMSE is much bigger than the training RMSE. While the test RMSE for 2013 is not as big as the training RMSE, the adjusted R-squared value is extremely high. However, the adjusted R-squared value for the 2000 model is high as well, so there seems to be more overfitting for this model.

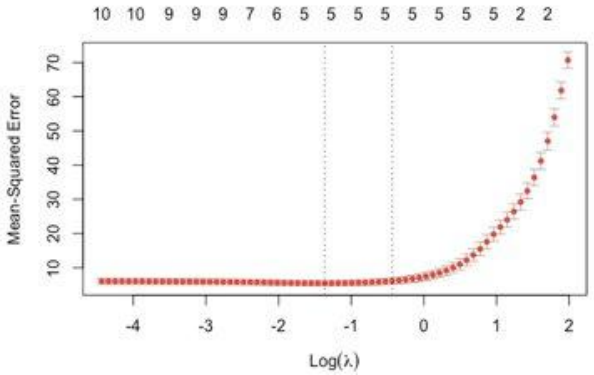
To see if we could improve overfitting in each of the models, we used several regularized regression techniques. First, we used ridge regression. For the 2000 model, the training RMSE and the test RMSE were 3.30 and 3.95 which was an improvement from the OLS model. As for the 2013 model, the training RMSE was 2.41 and the test RMSE was 3.08 which is not better or worse than the OLS model. The log-lambda plots for ridge regression are shown below for 2000 (left) and 2013 (right):





Next, we used lasso regression and we wanted to compare the feature selection using lasso with the features we chose. For the 2000 model, the training RMSE and the test RMSE were 3.11 and 4.12. Therefore, lasso did not help with overfitting for the 2000 model. However, in the 2013 model, the training RMSE was 2.27 and the test RMSE was 2.70 which is an improvement from the OLS and ridge models. The variables chosen in the 2000 model were age5.19obesity, basic_water, and une_hiv. For the 2013 model, it had those same three variables with the addition of age1.4mort and gghe.d. The log-lambda plots for lasso regression are shown below for 2000 (left) and 2013 (right):





Then, we tested out elastic net regression. For 2000, the best alpha value was 0.5 and for 2013, the best alpha value was 0.75. The training the test RMSE values for 2000 were 2.92 and 3.84. For 2013, they were 2.30 and 2.77. Finally, based on the professor's recommendation, we tried using the relaxed lasso technique. For the 2000 data, the only two variables selected were age5.19obesity and basic_water, and the R-squared value was 72.36% which is a big decrease. So, relaxed lasso did not do a good job here. For the 2013 data, it selected the same variables as lasso, however age5.19 obesity was removed. The R-squared value was 92.71% which is quite reasonable.

The best models for the 2000 data were ridge regression and an elastic net with an alpha of 0.5. For 2013, the best models were the OLS model and lasso. Elastic net with an alpha of 0.75 also did well. The most significant predictors in 2000 for the life expectancy are obesity levels within kids and teenagers, country populations that have basic water services, and the prevalence of HIV from people ages 15-49. Obesity levels and water levels have positive coefficients which means these lead to higher life expectancy, while the rate of HIV has a negative beta, which leads to a lower life expectancy. For 2013, those same predictors were also significant and have the same signs for their betas, however the mortality of children 1-4 and the domestic general government health expenditure were also significant. For mortality, there is a negative beta, and for health expenditure, there is a positive beta. While most of these seem like obvious contributions, it does not make sense to us that a higher obesity level leads to a higher life expectancy. Maybe different countries have different definitions of what is considered obese or different obesity rates. Perhaps if we removed the feature completely, our models would have been much different.

Factor Analysis

For factor analysis, we have considered the year 2013 and for that we first wanted to confirm how many factors we wanted to focus on. Also, we did not include features that have too much multicollinearity and may blur the factor understanding. We look at the scree and plot and parallel analysis for determining the number of components and factors.

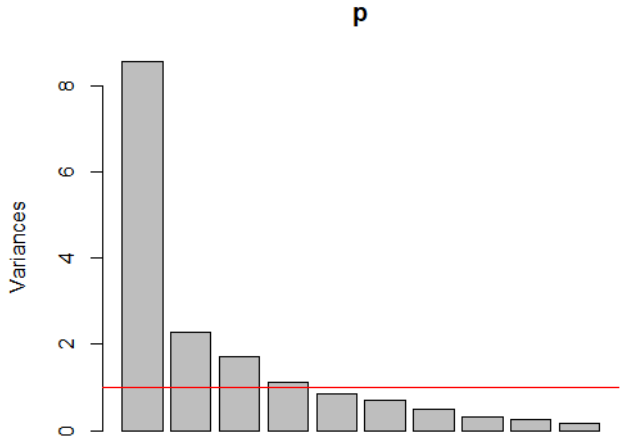


Figure 3: Scree Plot to determine number of Components.

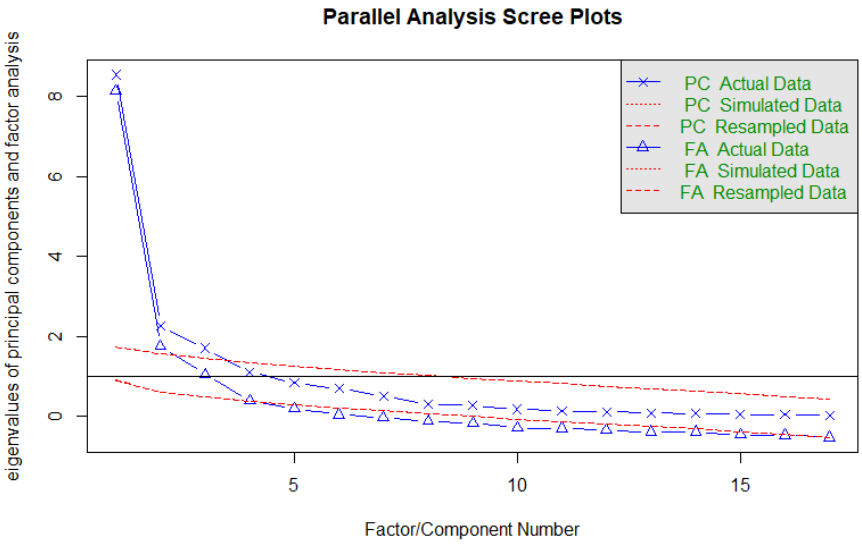
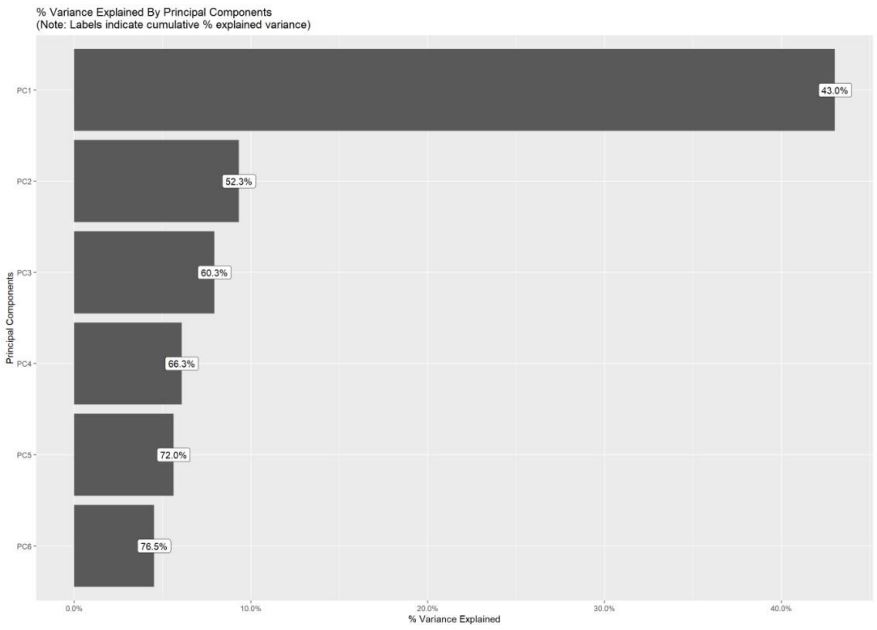


Figure 4: Parallel Analysis to determine how many components/ factors to emphasis on.

Here, parallel analysis suggests that we should focus on 3 factors and 3 components for an efficient result. Now, let's look at PCA visualization:



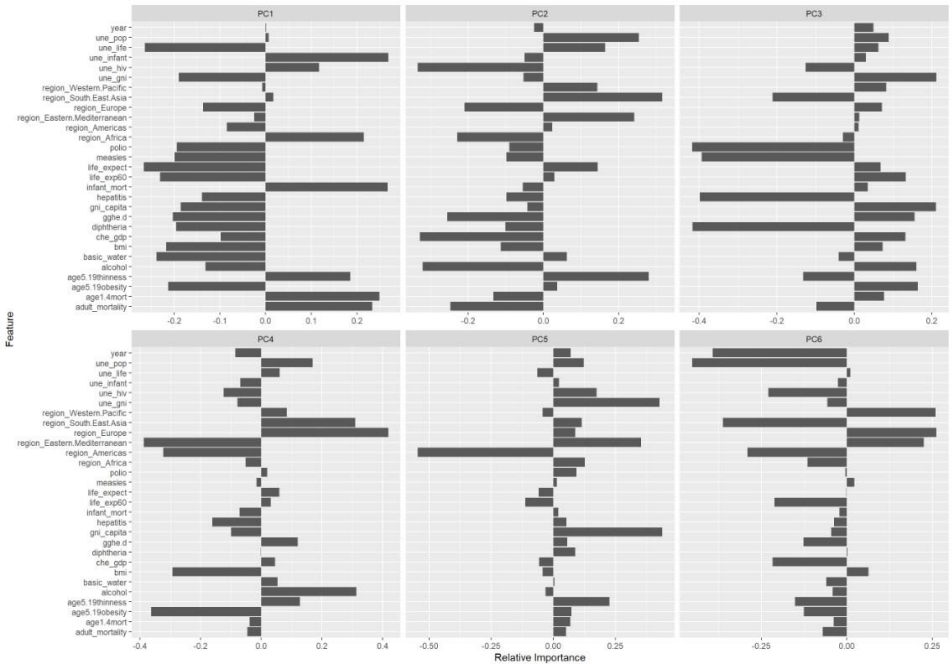


Figure 5: PCA factors and visualization.

```
Loadings:
                RC2    RC1    RC3
infant_mort     -0.581 -0.539  0.510
age1.4mort      -0.627 -0.446  0.476
hepatitis        0.939
measles          0.911
polio            0.938
diphtheria       0.940
basic_water      0.523  0.503 -0.506
alcohol                 0.686
bmi                     0.720
age5.19thinness        -0.796
age5.19obesity          0.677 -0.432
gni_capita              0.773
gghe.d                  0.860
che_gdp                 0.676
adult_mortality -0.434          0.748
une_hiv                 0.797
une_pop                -0.439

                RC2   RC1   RC3
SS loadings    5.008 4.953 2.570
Proportion Var 0.295 0.291 0.151
Cumulative Var 0.295 0.586 0.737
```

Figure 6: Loadings for factor analysis (3 factors)

In factor analysis, we have taken the cutoff as 0.4 for the loadings and also we have used varimax transformation for 3 factors. Almost 73% of the variance is captured by the data. In the loadings we can see that all the 3 factors have some different mixture of features. Let's look at each of the factor in brief:

Here, RC2 suggests that the immunization coverage for hepatitis, measles, polio and diphtheria have very high positive contributions and age1.4mort, infant_mort and adult_mortality have negative contributions. RC2 suggests that the mortality rate is very low and especially for children having high life expectancy.

Also, RC1 has negative contributions from infant_mort, age1.4mort along with age5.19thinness whereas positive contribution from gghe.d, gni_capita, che_gdp and bmi. RC1 suggests life expectancy for white collar working class.

RC3 has negative contributions from basic_water, age5.19obesity and une_pop while une_hiv and adult_mortality have slightly high positive contributions. RC3 suggests the likelihood of 3rd world countries, where basic necessities are not available and have less life expectancy.

Linear Discriminant Analysis

For our exploratory LDA, we chose to focus our 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. 'Region' was selected as the target variable for LDA so we could see how the linear discriminants generated by LDA would separate the data with respect to 'region'. The prior probabilities in Figure 7 below show that the data selected for LDA is predominantly composed of the regions of Africa, Americas, and Europe (combining for ~79% of the entries in the selected data). As a result, 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)

#normalize data
WHO_LDA_Data_Normalized = WHO_LDA_Data
WHO_LDA_Data_Normalized[,2:19]=scale(WHO_LDA_Data_Normalized[,2:19])


WHO_LDA_Data1 = WHO_multicolinearity_removed%>%
  filter(year=='2013')%>%
  dplyr::select(-year)



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

LD1 = WHO.LDA.Values$x[,1]
LD2 = WHO.LDA.Values$x[,2]
LD3 = WHO.LDA.Values$x[,3]
predRegion =WHO.LDA.Values$class
actualRegion = WHO_LDA_Data1$region
Country = WHO_LDA_Data1$country
WHO.LDA.Plot = data.frame(LD1,LD2,LD3,predRegion,actualRegion,Country)

#2D Plot
ggplot(WHO.LDA.Plot,aes(x = LD1,y=LD2,col=actualRegion))+geom_point()+
  ggtitle('Plot of LD1 and LD2, Categorized by Region')+theme_bw()+geom_text(label=Country)
#3D Plot
library(plotly)
plot_ly(x=WHO.LDA.Plot$LD1, y=WHO.LDA.Plot$LD2, z=WHO.LDA.Plot$LD3,
        type='scatter3d',mode='markers',color = WHO.LDA.Plot$actualRegion)
```

Figure 7

```
> WHO.LDA
Call:
lda(region ~ ., data = WHO_LDA_Data_Normalized)

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      1.90379349  2.29864954  0.16626469  3.96183158  3.770431938
adult_mortality  0.89800507  1.43300448  0.49578315  3.12125428  2.221409627
infant_mort      0.95658944  0.30871583 -0.39094188  0.72585604 -0.474726669
age1.4mort      -1.05642437  0.01724741 -0.52904084  0.45758692  2.177695610
alcohol          0.30525009 -0.32145435  0.20115875  0.19224972 -0.245847537
bmi              0.72156217  0.54158481 -0.49604884 -1.89494230  0.951350087
age5.19thinness -0.11437764  2.09530647  0.38523128 -0.87445933  0.312195331
age5.19obesity   0.34566171  1.27873338 -1.00458773  0.87967308 -0.746238685
hepatitis       -0.19935684 -0.24684418  0.15883267  0.33780363 -0.461901457
measles          0.66193738  0.04549329  0.03108698  0.55243478 -1.687130106
polio           -0.95261611 -0.60666795  0.24419453 -0.72352056  0.088403328
diphtheria       0.08624413  1.21890017 -0.16874611  0.04525501  2.353398600
basic_water      0.25570974 -0.51138609  0.30978359  1.00566080  0.407410628
gni_capita      -0.80768690 -0.53487090  0.82593273 -1.02916235 -1.041382148
gghe.d           0.08786713 -0.21203302  0.06648310 -0.11211553  0.410307479
che_gdp          0.30799611 -0.21787795 -0.28003111 -0.25913320  0.031033477
une_pop          0.07696046  0.19356011 -0.04585381  0.26446485  0.478353561
une_hiv         -0.75980940 -0.21606769 -0.32148697 -0.16781122  0.005891331

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, plotting the LDA values against LD1 and LD2 (figure 8) should provide a good visual representation of the separation and clustering of regions in the data (LD3 was added for a 3D plot as well in figure 9, but did not seem to present any significantly different insights from the first two discriminants). Figure 8 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 in figure 7, 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 polio, and high rates of HIV, indicating that these are all challenges within Africa's healthcare systems.

Figure 8

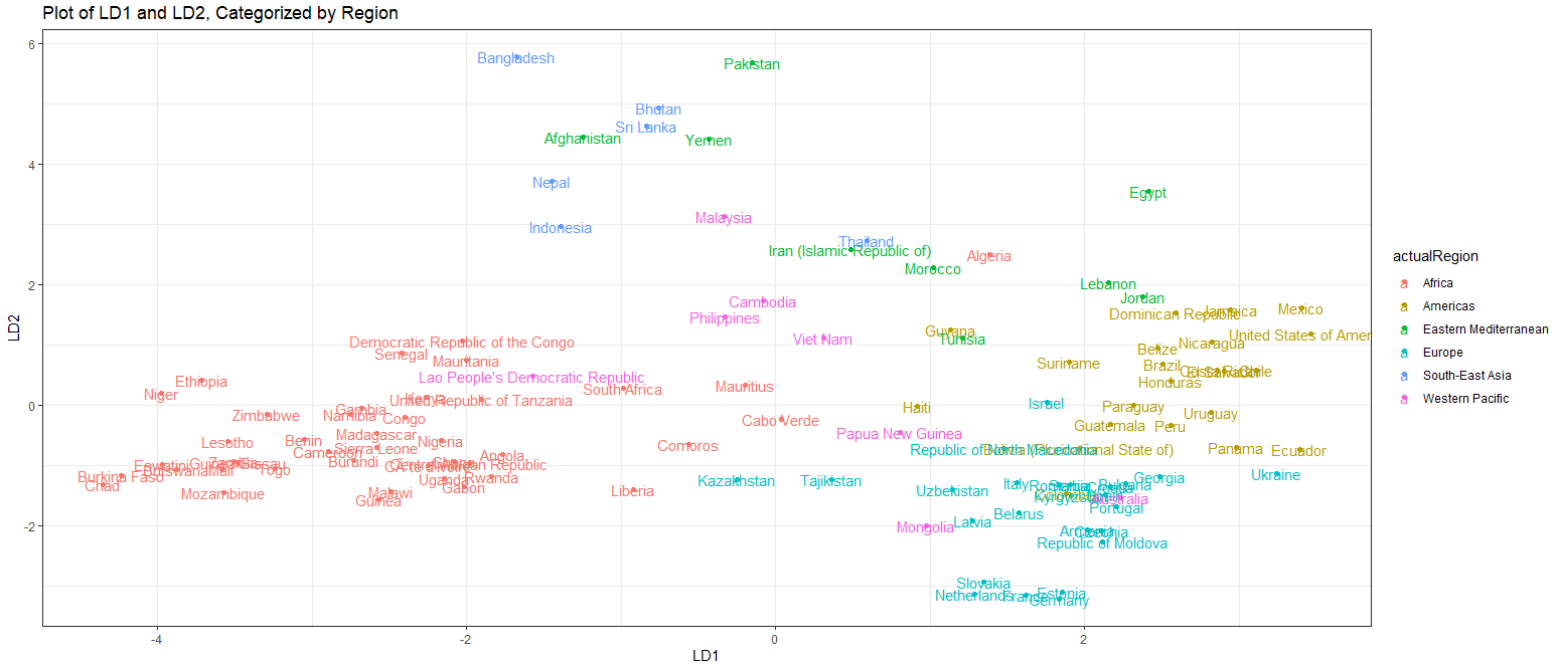
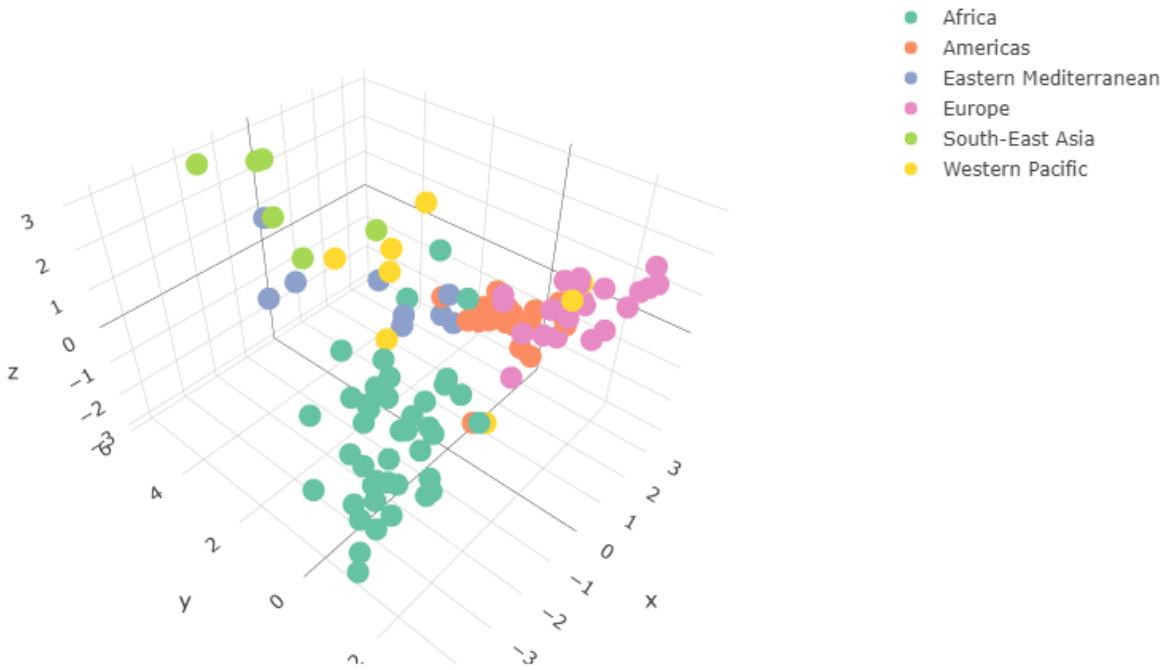


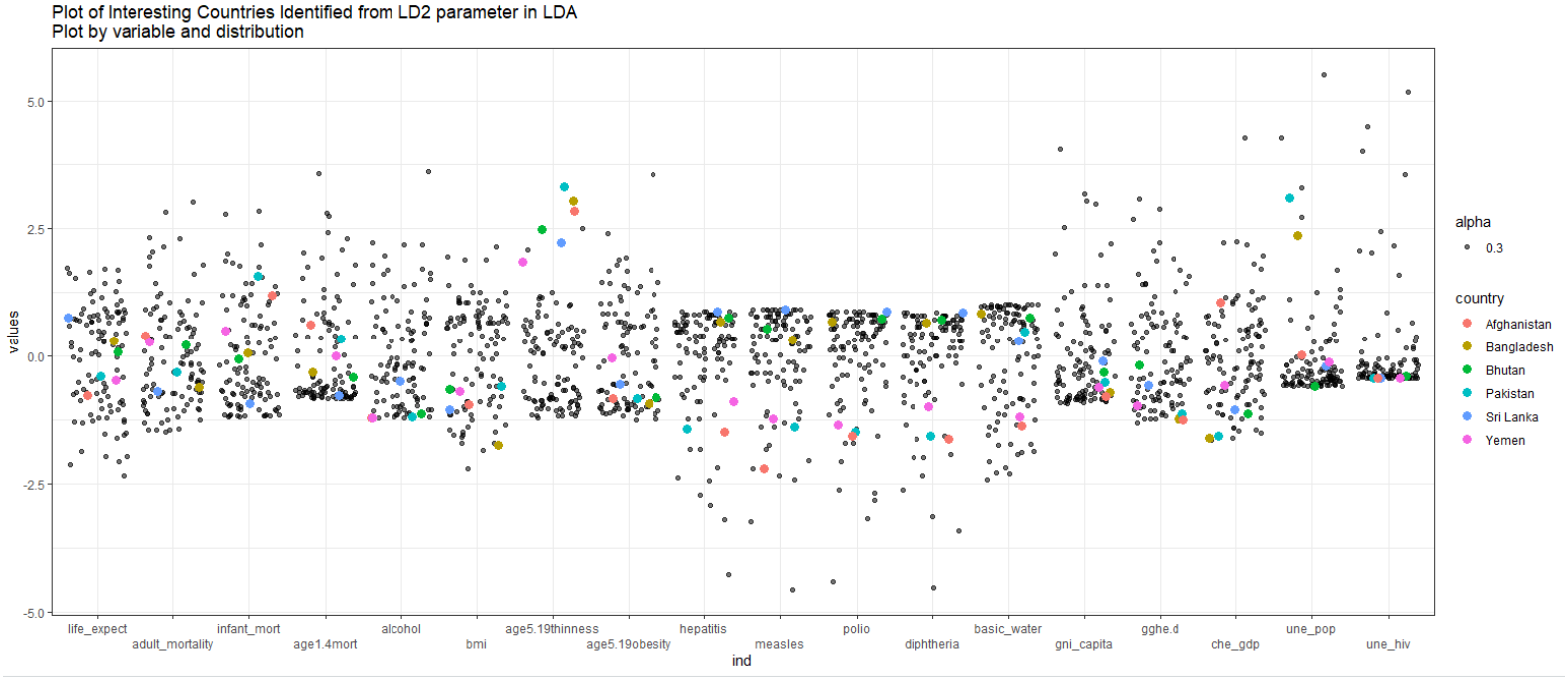
Figure 9



We also looked into Bangladesh, Pakistan, Bhutan, Afghanistan, Sri Lanka, and Yemen individually as they represented relative outliers on the high end of the LD2 axis in figure 8. For this, an alternative visual analysis approach was performed (figure 10). Z scores were plotted for each country, within each variable category. Specific countries of

interest noted above are highlighted 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 7. However, the visual representation of the data helped identify the similarities between these countries that resulted in their high LD2 values.

Figure 10



Individual Reports

Individual Report - Matthew Ghuneim

Main Responsibilities

For the project, I was responsible for making the initial regressions for Milestone 3 and also performing regularized regression for the years 2000 and 2013, with the variable life_expect as our parameter of interest. Each of us had preprocessed our data in different ways, but I ended up using the dataset that Andrew cleaned, because a lot of the missing data was removed and replaced. However, the data from 2014-2016 was removed since there were many missing values, so 2013 was the most recent year. I was also the “timekeeper” to make sure everyone met deadlines and I created the PowerPoint we are using for our presentation.

Data Preprocessing/OLS

I started out by computing two OLS regression models (for 2000 and 2013) without the categorical variables and while most variables were not significant, the R-squared value was 99%. I used 75% of the data as training data and the rest as testing data. There were several variables in the dataset that seemed like obvious contributions to the life expectancy variable so I manually removed some of them from the training and testing datasets. I also removed the variables with high multicollinearity, and in each training and test set, the variables have a VIF score less than 5.

Regularized Regression for the data from 2000

For the 2000 data, there are 10 variables in the OLS model, and the training and testing RMSEs are 2.50 and 4.64 respectively. Since the test RMSE is much bigger than the training RMSE, this would suggest overfitting. Starting with ridge regression, I plotted the MSEs vs the log lambdas and chose to select the lambda.1se value. The training RMSE was 3.30 and the test RMSE was 3.95 which is a massive improvement from OLS. Next, I used lasso regression and I also plotted the MSEs vs the log lambdas. Using lambda.1se again, I computed the RMSEs and the training RMSE was 3.11 and the test RMSE was 4.12. Lasso still performs better than OLS, but not as well as ridge. The variables chosen by lasso regression were age5.19obesity, basic_water, and une_hiv. Then, I computed elastic net regression to see if combining ridge and lasso would improve the model more since there was a massive improvement with both methods, and I found that an alpha value of 0.5 gave me the best result. The training RMSE was 2.92 and the test RMSE was 3.84. For relaxed lasso, there were only 2 variables selected and the R-squared value dropped drastically. The results of each of the regressions suggest that ridge regression and elastic net performs the best on the data from 2000.

Regularized Regression for the data from 2013

There are also 10 variables in the OLS model for the 2013 data, and the training and testing RMSEs are 2.01 and 2.68 respectively. The test RMSE is a bit bigger than the training RMSE so overfitting may be an issue, even though the test RMSE is not that much bigger. Starting with ridge regression, I plotted the MSEs vs the log lambdas and chose to select the lambda.1se value. The training RMSE was 2.41 and the test RMSE was 3.08. Therefore, ridge regression does not perform better or worse than the OLS model. Next, I used lasso regression and I also plotted the MSEs vs the log lambdas. Using lambda.1se again, I computed the RMSEs and the training RMSE was 2.27 and the test RMSE was 2.70. So, lasso performs better than OLS and ridge. The variables chosen by lasso regression were age1.4mort, age5.19obesity, basic_water, gghe.d, and une_hiv. Then, I computed elastic net regression, and I found that an alpha value of 0.75 gave me the best result. The training RMSE was 2.30 and the test RMSE was 2.77. Therefore, out of all of the regressions used, lasso regression and OLS perform the best on the 2013 data, although elastic net with an alpha of 0.75 performs well.

Conclusions/Takeaways

I have drawn a few conclusions from the models I computed. I have found that for 2000, the model predicts that higher obesity levels for children 5-19 and more basic water services lead to a higher life expectancy, while a higher HIV rate for a country leads to a lower life expectancy. For the year 2013, the model predicted those same variables, however it also predicted that a lower mortality for a child leads to a lower life expectancy and a higher domestic general government health expenditure leads to a higher life expectancy. However, a higher obesity rate leading to a higher life expectancy does not make sense, so I definitely learned that in that scenario, you have to use your best judgment when it comes to which variables to keep. As far as data analysis, I learned that it is important to try out different approaches beyond standard model building techniques such as forward selection/backward elimination. Regularized regression helped with overfitting, and it also gave the best models for each of the years. Also, the models that I created are just one set of variables. If you run the models more than once, you will not get the same RMSEs each time, so the performance of each regression changes. My goal was to create more parsimonious models and keep the variables that have obvious contributions out of the model. However, I learned that you always need to keep assumptions in mind, and in order to make the most accurate prediction, something may need to be used beyond regularized regression.

Individual Report - Smit Patel

Main Responsibilities

In this group project, I was mainly responsible for doing some summary of data, visualization of variables, some plotting, initial regressions and for the year 2013, doing factor analysis. Like Matthew, I also used Andrew's version of cleaned data as it had the most structured data and also wanted to keep continuity to our analysis. I was also responsible for helping the teammates if they were to be stuck somewhere or needed an opinion on any analysis.

Data Overview

I had to focus on variables that might have the most effect on life expectancy and can possibly be the most important features in determining it. Also, to give different plots, correlation and overall summary of the data. I had removed some variables that seemed to have multicollinearity that affected the efficiency of the model apart from the cleaned data, but kept the necessary variables. Also, for the PCA visualization, I have used an inbuilt library and so it has the variable region divided into the values and not considered as one.

Principal Component Analysis

For the PCA, we can clearly abridge that the 1st component is trying to address the situation of Africa, where adult mortality rate is high along with high infant mortality and the number of HIV cases. Some parts of SouthEast Asia also have the same problem. We can also see that for this particular component, no government related support or immunization is provided and suffer largely from that.

Second component does not give much explanation and has no clear interpretation as it says that certain areas have high life expectancy where population is quite high and some other factors which do not add up. Also, these areas do not have access to immunization either and also the support from government is not good along with no economic growth.

The third component suggests that the regions which receive government aid/ support tend to have acceptable life expectancy even though no access to basic water and no immunization coverage. Although, these regions still have a high infant mortality rate.

I think that the PCA used with the default R package gives us a somewhat good interpretation as we are looking at the outcome for the data over the years.

Factor Analysis

For factor analysis I consider the year 2013. I consider 3 factors with varimax transformation. For the loadings, the cutoff is set to 0.4. About 73% of the variance in the data is captured by the 3 factors.

The formulae for the 3 factors are as follows:

RC2: – 0.581*infant_mort – 0.627*age1.4mort + 0.939*hepatitis + 0.911*measles + 0.938*polio + 0.940*diphtheria + 0.523 – 0.434*adult_mortality.
RC2 suggests that the mortality rate is very low and especially for children having high life expectancy.

RC1: – 0.539*infant_mort – 0.446*age1.4mort + 0.503*basic_water + 0.686*alcohol + 0.720*bmi – 0.796*age5.19thinness + 0.677* age5.19obesity + 0.773*gni_capita + 0.860*gghe.d + 0.676*che_gdp.
RC1 has negative contributions from infant_mort, age1.4mort along with age5.19thinness whereas positive contribution from gghe.d, gni_capita, che_gdp and bmi. RC1 suggests life expectancy for white collar working class.

RC3: 0.510*infant_mort + 0.476*age1.4mort – 0.506*basic_water – 0.432*age5.19obesity _ 0.748*adult_mortality + 0.797*une_hiv – 0.439*une_pop.
RC3 suggests the likelihood of 3rd world countries, where basic necessities are not available and have less life expectancy.

Let's look at PFA with 4 factors:

```
Loadings:
                RC2    RC1    RC3    RC4
age1.4mort      -0.578 -0.413  0.559
hepatitis        0.934
measles          0.906
polio            0.946
diphtheria       0.943
alcohol                0.683
bmi                    0.685 -0.404
age5.19thinness       -0.766         0.415
age5.19obesity         0.654 -0.483
gni_capita             0.771
gghe.d                 0.856
che_gdp                0.696
adult_mortality                0.801
infant_mort     -0.524 -0.503  0.604
basic_water      0.477  0.473 -0.568
une_hiv                        0.832
une_pop                                0.912

                 RC2   RC1   RC3   RC4
SS loadings    4.732 4.697 2.994 1.220
Proportion Var 0.278 0.276 0.176 0.072
Cumulative Var 0.278 0.555 0.731 0.803
```

We capture 80% of the variance with 4 factors. Here, the first 3 factors still have somewhat the same interpretation and the 4th factor has contribution from population and thinness amongst young. This is a little similar to one of the components in PCA.

Conclusion

As seen in both factor analysis and component analysis, both suggest that 3rd world countries have a high mortality rate and very low life expectancy. This indicates that more resources and contribution should focus on improvement of those regions. We also see that immunization coverage also plays an important role in having a good life expectancy. Adding to that, economy, population and government support also influence life expectancy. We can see that it was having a good contribution for the white collar working class. Hence, we can say that all the variables here give an insightful meaning to the data when composed in some fashion.

Individual Report - Andrew Fitzsimmons

Main Responsibilities

My primary responsibilities for this project were to clean the dataset to get it into a structure that was useful for our analysis, this involved resolving ‘NA’ values within variables, removing variables that were incomplete, and reducing the number of rows to ensure that we were left with only complete cases on which to perform our analysis. After cleaning, we were still left with over ~1300 rows of data for our analysis. I also performed an exploratory Linear Discriminant Analysis on the dataset to cluster the data by ‘region’. This process included scaling the data, setting up the LDA pipeline, interpreting the coefficients of the linear discriminants, creating visualizations to represent the analysis, and interpreting the results/drawing conclusions from the groupings of the LDA clusters.

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 was careful not to oversimplify this process in a way that obscured the message of the data. I reviewed each country with missing ‘Alcohol’ values and identified that the missing values appeared to be random with respect to each country, and the values that were present, were all tightly grouped within each 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.

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 7 show that the data I selected for LDA is predominantly composed 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.

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 (figure 8) should provide a good visual representation of the separation and clustering of regions in the data (LD3 was added for a 3D plot as well in figure 9, but did not seem to present any significantly different insights from the first two discriminants). Figure 8 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 polio, and high rates of HIV, indicating that these are all challenges within Africa's healthcare systems.

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 (figure 10) 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 7. However, the visual representation of the data helped me better identify the similarities between these countries that resulted in their high LD2 values.

Conclusion

After performing my exploratory LDA, it appears that the state of healthcare in Africa is different from that of the Americas and Europe. This is primarily driven by high infant mortality rates in Africa, high rates of adolescent thinness in Africa, high rates of polio and hepatitis in Africa, and high rates of HIV. These challenges do not appear to be present in Europe and the Americas. Another interesting finding was that Bangladesh, Pakistan, Bhutan, Afghanistan, Sri Lanka, and Yemen all appear to have an average life expectancy despite having some of the highest rates of adolescent thinness, lowest BMI, lowest GNI per capita, lowest government healthcare spend. The one factor that stands out that is similar across these 6 countries is that they all have some of the lowest alcohol usage of all countries in the dataset