### Life Expectancy

## Team: Blake Freeman, Lijju Mathew, Aniketh Vankina

## Introduction

Over the course of many years, the World Health Organization collected data from various countries to depict the overall status. The data stored was valuable, as it ranged from financial, disease, to societal issues. The goal of this study is to predict life expectancy based on the many variables that were collected for this dataset.

## Data Description

As specified above the data collected from WHO is tailored to check the overall living status for multiple countries. As per this research, we are only interested in data from the year of 2014. The data consisted of 22 variables. Out of the 22 variables, two were categorical and 20 were numerical or continuous variables. The two categorical variables were “Country” and “Status”. The numerical variables consisted of a combination of areas that range from alcohol, GDP, HIV.AIDs, etc...

## Exploratory Data Analysis

In the Exploration of the Data we identified multiple trends and used this to shape our approach in modeling. Below bullets the observations that we identified from our exploratory data phase.

* Null Values and their treatment.
  + We Identified that there was a significant amount of null values in our data set. (Figure 1) After identifying the null values in the dataset we decided to use the Mice package to fill in the missing values with the means of those columns. We took this approach so that there is consistency when addressing the correlations and deciding on what factors we could use in our models.
* Out of total 183 countries, 151 countries are developing and 32 are developed.
  + Due to there being a very small sample size of developed versus developing countries we identified that developing countries had a significant correlation with life expectancy. We ultimately did not use this as a predictor value in our modeling since there is a small sample size of developed countries comparatively which was causing the data to be skewed.
* Correlations between variables.
  + From those factors we have found that Adult.Mortality, Total.expenditure, Income.composition.of.resources, Alcohol, HIV.AIDS, thinness..1.19.years affect life expectancy with a high amount of significance, therefore we used these variables in the models. That is shown in our correlation matrix below. See Appendix (Figure 2)
* The effects of healthcare expenditure.
  + From our findings we should increase the expenditure and it would have a positive impact on life expectancy.
* The effects of Infant and Adult mortality rates on life expectancy.
  + We found that both infant and adult mortality negatively impacted life expectancy however Adult mortality was more highly correlated than the infant mortality rate, therefore we excluded it from our feature selection. (Figure 3)
* The effects of Schooling on life expectancy
  + We found that there was a positive relationship with Schooling and life expectancy. (Figure 4)
* The relationship with drinking alcohol on life expectancy.
  + We found that there was a positive relationship with drinking alcohol and life expectancy. (Figure 4)

## Objective 1:

The overall goal that was required for this study is to check which variables are the key players that can affect and predict life expectancy. The particular direction on how we decided to obtain our goal was to dive deeper with the data to see which predictor variables had the greatest significance.

### Model Selection:

We took a custom approach for the model selection. As noted above in the Exploratory Data Analysis we did a correlation matrix to see the significance each variable had with life expectancy, excluding state and status. We also did a forward, backward, stepwise with all predictors. We then reduced the model using the LASSO selection technique, including all predictors that are statistically significant to life expectancy.

The results gave us six predictor variables we decided to use on our final model. However, we also decided to go with more of a simple model that reduced the number of predictor variables without sacrificing metrics for interpretation. A LASSO, Stepwise, and Forward selection techniques were used to evaluate the predictors in the final model. The results reduced the number down to three with just a slight decrease in the adjusted R2.

### Assumptions

(Appendix for Graphs)

After assessing the information and going through the data, the distribution of the residuals is fairly normal as shown on the residual plots. Additionally, the data is also normally distributed as per the QQ plot. The assumptions have been met for us to proceed. Independence is also assumed.

As far as the influential point analysis goes, the Cook’s D scale from 1 to 0.2. It does not seem that there are any significant outliers. Additionally, the leverage plots seem to be fairly uniform.

### Compare Competing Models

#### Possible Metrics

|  |  |  |  |
| --- | --- | --- | --- |
| Predictive Models | Predictors | ASE | Adj R2 |
| Final Model 1 | Adult Mortality, Alcohol, Total Expenditure, HIV.AIDS, Income Composition, Thinness. 1.19 years | 13.83 | 80.9 |
| Final Model 2 | Adult Mortality, Total Expenditure, Income Composition | 16.0 | 77.4 |

#### Parameter Interpretation

Holding all variables constant, for every count increase in Adult Mortality life expectancy will decrease by -0.02. For every percent increase in total expenditure, life expectancy will increase by 0.60. Finally, for every unit increase in income the life expectancy will increase by 31.48.

**Confidence Intervals:**

We are 95% confident that the true coefficient falls in the following for each of these variables:

Adult Mortality: [-0.032, -0.011]

Total Expenditure [0.319, 0.982]

### Conclusion

With the explanation, we decided that a model using Adult Mortality, Total Expenditure, and Income composition of resources to be the most accurate model with the least amount of variables. This has a result of an ASE of 16.0 and an Adjusted R-squared value of 77.4. Indicating that the variables we chose still have a high level of correlation and fit to the test set. We found that this trade off to be acceptable to try to avoid overfitting the model and still have a high level of accuracy.

## Objective 2:

### KNN

We settled on the use of K-nearest neighbors for our non-parametric model strategy. We modeled the full and reduced model noted above to get the results for this test.

### Results

After hyper-tuning the models (see figure 6,7,8 below)

* Full model had an accuracy of .031
* Reduced model had an accuracy of .037

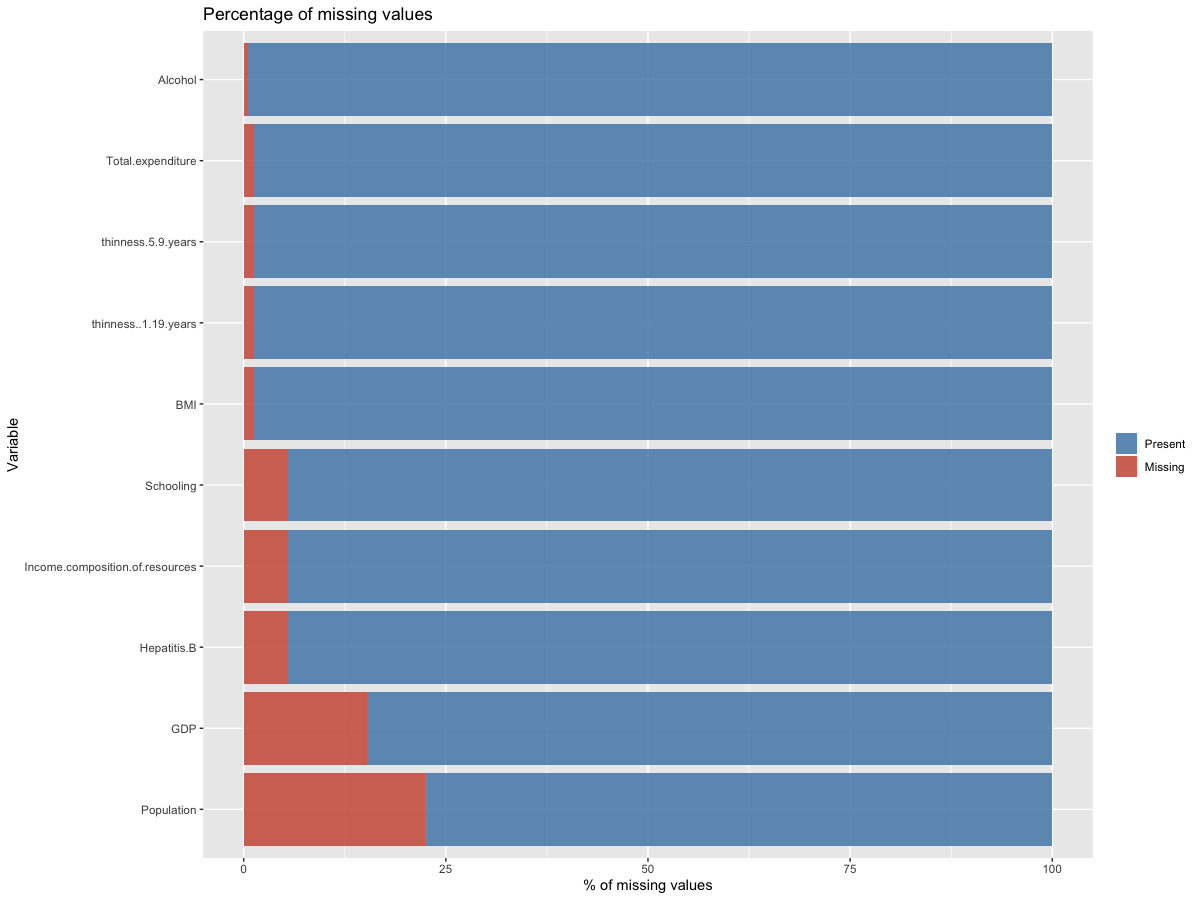
### Conclusion

From running both nonparametric models comparing it to the results that were established from the linear model. The data is more suited for a linear regression than a nonparametric model. This is clear from the low accuracy (fit) of the model which did not even achieve over 5% while we have a very close fit after running the models from linear regression models.

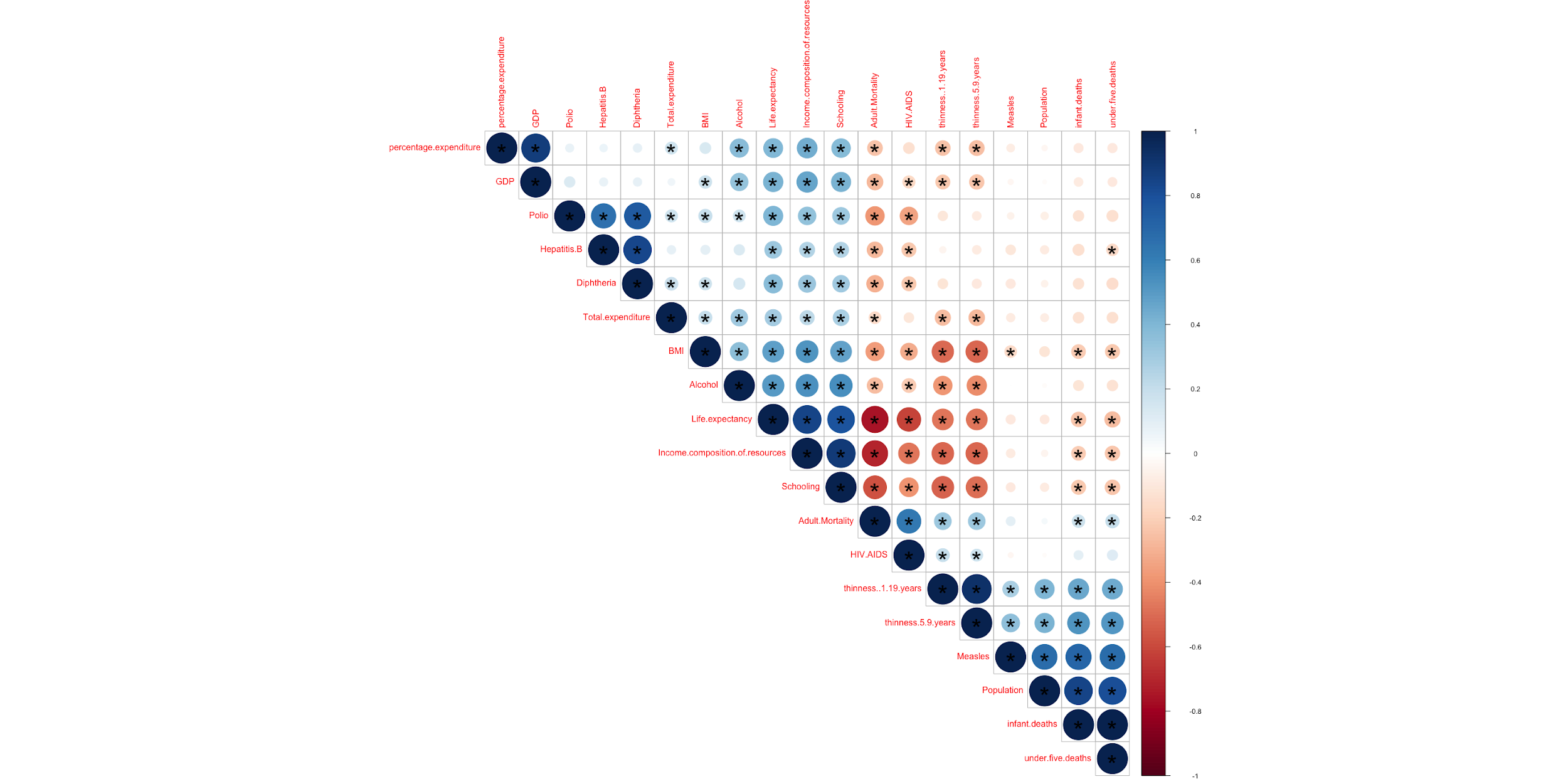
## Appendix:

### Charts

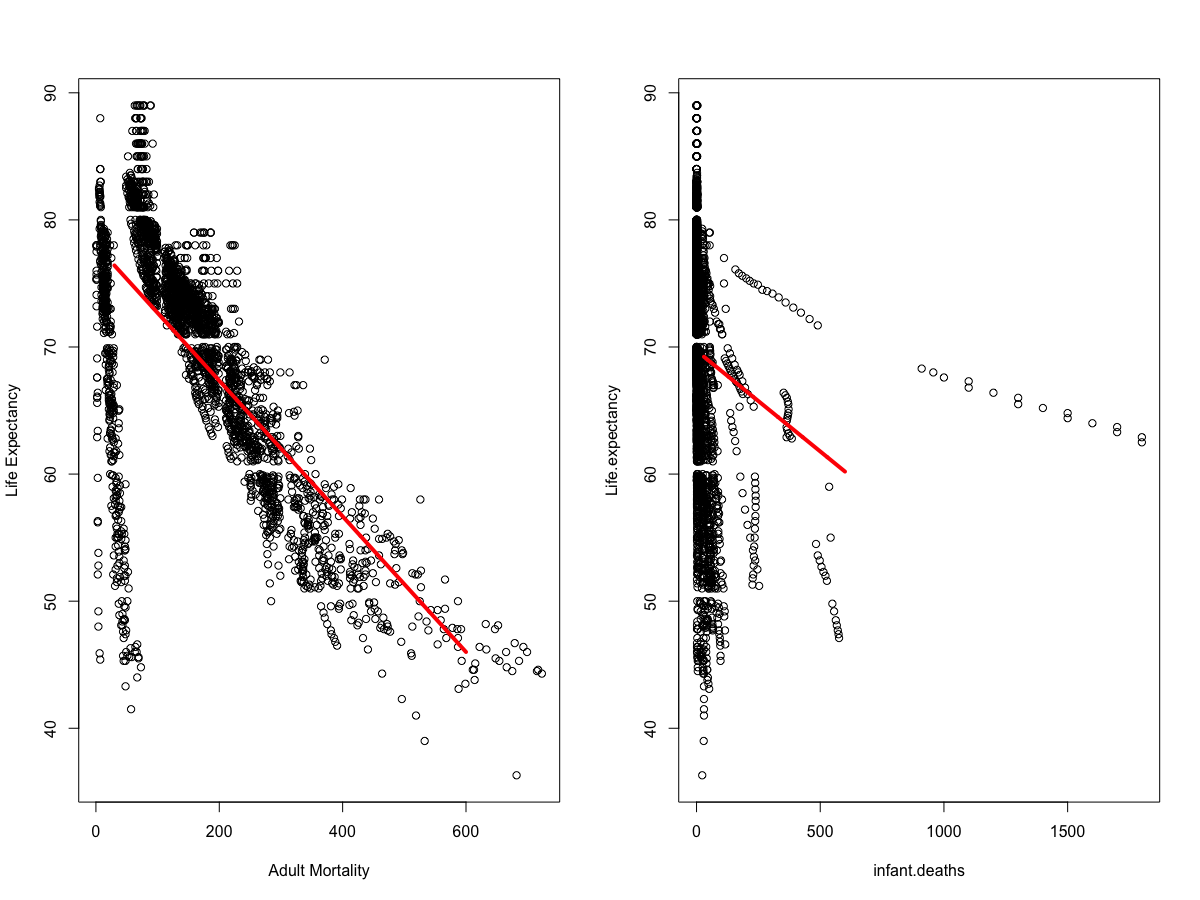
(Figure 1)



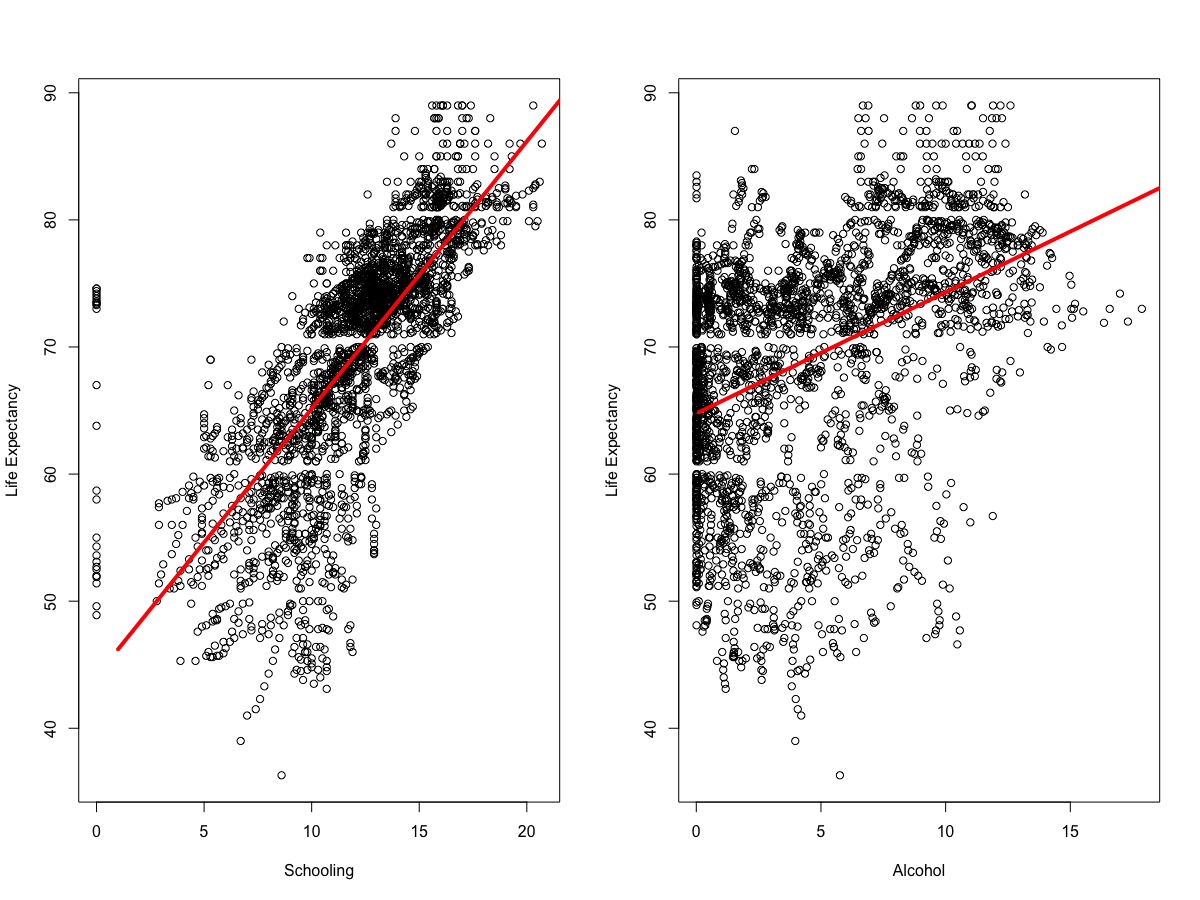
(Figure 2)



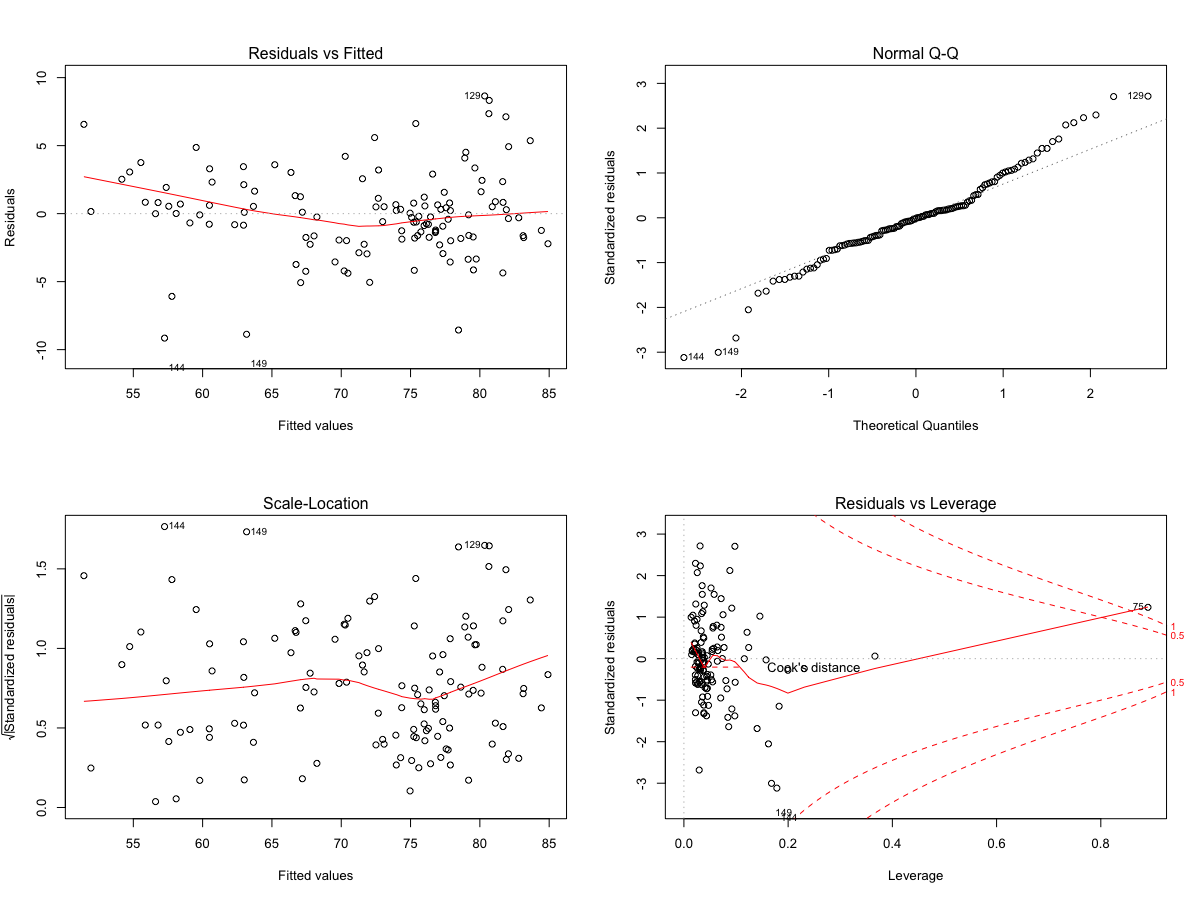
(Figure 3)



(Figure 4)



(Figure 5)



(Figure 6)

A close up of text on a white background

Description automatically generated

(Figure 7)

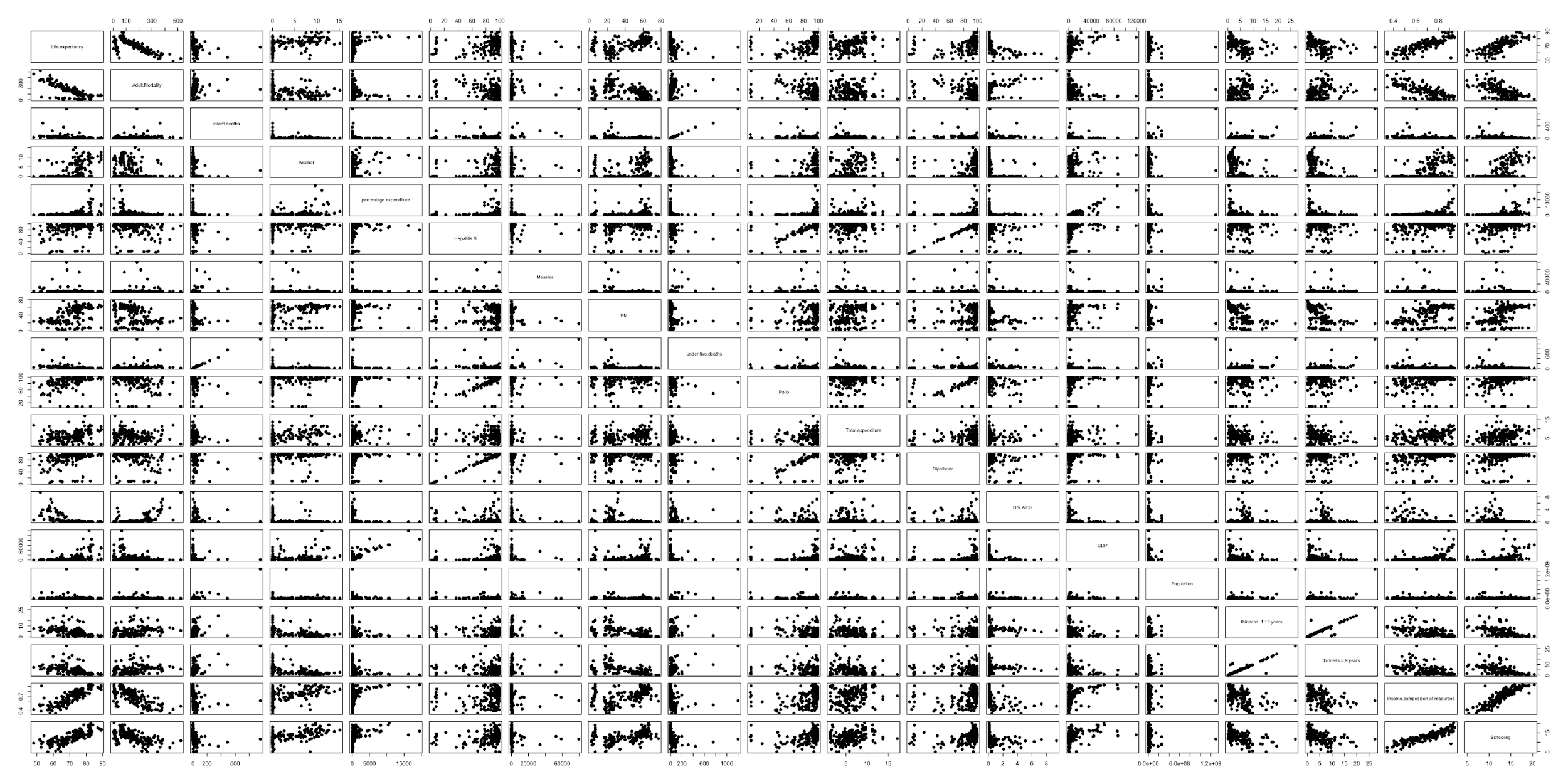
A screenshot of a cell phone

Description automatically generated

(Figure 8)

A screenshot of a cell phone

Description automatically generated



### Code

|  |
| --- |
| --- |
|  | title: "Life Expectancy Analysis" |
|  | author: "Lijju Mathew, Blake Freeman, Aniketh Vankina" |
|  | date: "June 4, 2020" |
|  | output: html\_document |
|  | editor\_options: |
|  | chunk\_output\_type: console |
|  | --- |
|  |  |
|  | ```{r setup, include=FALSE} |
|  | knitr::opts\_chunk$set(echo = TRUE) |
|  | ``` |
|  |  |
|  | **## R Markdown** |
|  |  |
|  | install.packages("model4you") |
|  |  |
|  | ```{r library} |
|  | library(dvmisc) |
|  | library(rgl) |
|  | library(Metrics) |
|  | library(tidyverse) |
|  | library(ggplot2) |
|  | library(caTools) |
|  | library(dplyr) |
|  | library(magrittr) |
|  | library(readr) |
|  | library(survival) |
|  | library(nlme) |
|  | library(gridExtra) #grid.arrange() |
|  | library(class) |
|  | library(forcats) |
|  | library(MASS) |
|  | library(GGally) |
|  | library(tidyr) |
|  | library(maps) |
|  | library(mapproj) |
|  | library(stringr) |
|  | library(rmarkdown) |
|  | library(knitr) |
|  | library(jsonlite) |
|  | library(RCurl) |
|  | library(class) |
|  | library(httr) |
|  | library(mice) |
|  | library(corrplot) |
|  | library(GoodmanKruskal) # GKtauDataframe() |
|  | library(glmnet) |
|  | library(car) |
|  | library(moderndive) # get\_regression\_table() |
|  | library(faraway) |
|  | library(randomForest) |
|  | library(e1071) |
|  | library(naivebayes) |
|  | library(caret) |
|  | library(SuperLearner) |
|  | library(psych) |
|  | library(tm) |
|  | library(VIM) |
|  | library(model4you) |
|  |  |
|  |  |
|  | ``` |
|  |  |
|  | ============ |
|  | OBJECTIVE 1 |
|  | ============ |
|  |  |
|  | ```{r datacleanup} |
|  | who\_data <- read.csv(file.choose()) |
|  | #who\_data <- read.csv("/Users/lijjumathew/Library/Mobile Documents/com~apple~CloudDocs/Lijju/SMU/Courses/Applied Statistics/Project/Life Expectancy Data.csv") |
|  | who\_data <- who\_data %>% filter(Year == 2014) |
|  | head(who\_data) |
|  | summary(who\_data) |
|  | str(who\_data) |
|  |  |
|  | # Find the no of attributes with missing values |
|  | sort(sapply(who\_data, function(x) sum(is.na(x))), decreasing = T) |
|  |  |
|  | #Missing data and percentage plot |
|  | missing.values <- who\_data %>% |
|  | gather(key = "key", value = "val") %>% |
|  | dplyr::mutate(isna = is.na(val)) %>% |
|  | dplyr::group\_by(key) %>% |
|  | dplyr::mutate(total = n()) %>% |
|  | dplyr::group\_by(key, total, isna)%>% |
|  | dplyr::summarise(num.isna = n())%>% |
|  | dplyr::mutate(pct = num.isna / total \* 100) |
|  |  |
|  | levels <- (missing.values %>% filter(isna == T) %>% arrange(desc(pct)))$key |
|  |  |
|  | percentage.plot <- missing.values %>% |
|  | ggplot() + geom\_bar(aes(x = reorder(key, desc(pct)), y = pct, fill=isna), stat = 'identity', alpha=0.8) + |
|  | scale\_x\_discrete(limits = levels) + |
|  | scale\_fill\_manual(name = "", values = c('steelblue', 'tomato3'), labels = c("Present", "Missing")) + |
|  | coord\_flip() + |
|  | labs(title = "Percentage of missing values", x = 'Variable', y = "% of missing values") |
|  | percentage.plot |
|  |  |
|  | tempData <- mice(who\_data,m=1,maxit=0,meth='fastpmm',seed=500) |
|  | who\_imp <- complete(tempData,1) |
|  |  |
|  | sort(sapply(who\_imp, function(x) sum(is.na(x))), decreasing = T) |
|  | glimpse(who\_imp) |
|  | ``` |
|  |  |
|  | ```{r split} |
|  | set.seed(100) |
|  | split\_percent = .70 |
|  | trainIndices = sample(1:dim(who\_imp)[1],round(split\_percent \* dim(who\_imp)[1])) |
|  | train = who\_imp[trainIndices,] |
|  | test = who\_imp[-trainIndices,] |
|  | summary(train) |
|  | ``` |
|  |  |
|  | <<<<<<< HEAD |
|  | dev.off() |
|  | ======= |
|  | ```{r} |
|  | #For Git |
|  | #EDA Aniketh f |
|  | library(ISLR) |
|  | library(dplyr) |
|  | library(ggplot2) |
|  | library(ggcorrplot) |
|  | library(tidyr) |
|  | library(magrittr) |
|  | library(plyr) |
|  |  |
|  | life = read.csv(file.choose(), header = TRUE) |
|  | summary(life) |
|  | dplyr::glimpse(life) |
|  | attach(life) |
|  | #Life expectancy based on developing and developed countries |
|  | t(aggregate(Life.expectancy ~ Status, data = life, summary)) |
|  |  |
|  |  |
|  | #Does various predicting factors which has been chosen initially really affect the Life expectancy? What are the predicting variables actually affecting the life expectancy? |
|  | library(car) |
|  | who\_imp = who\_imp[,-3] |
|  | who\_imp = who\_imp[,-1] |
|  | full.model = lm(Life.expectancy~., data = who\_imp) |
|  | vif(full.model)[,3]^2 |
|  |  |
|  | alias(lm(Life.expectancy~., data = who\_imp)) |
|  |  |
|  | #Should a country having a lower life expectancy value(<65) increase its healthcare expenditure in order to improve its average lifespan? |
|  | life %>% select(Country, Life.expectancy, Total.expenditure, percentage.expenditure) %>% filter(Life.expectancy < 65) |
|  |  |
|  | #How does Infant and Adult mortality rates affect life expectancy? |
|  | par(mfrow = c(1,2)) |
|  | plot(Adult.Mortality, Life.expectancy, xlab = "Adult Mortality", ylab = "Life Expectancy") |
|  | new = data.frame(Adult.Mortality = seq(30, 600,1)) |
|  | lines(seq(30,600,1), predict(lm(Life.expectancy~Adult.Mortality), newdata = new), col = "red", lwd = 4) |
|  |  |
|  | plot(infant.deaths, Life.expectancy) |
|  | new2 = data.frame(infant.deaths = seq(30,600,1)) |
|  | lines(seq(30,600,1), predict(lm(Life.expectancy~infant.deaths), newdata = new2), col = "red", lwd = 4) |
|  | #Does Life Expectancy has positive or negative correlation with eating habits, lifestyle, exercise, smoking, drinking alcohol etc. |
|  |  |
|  | #What is the impact of schooling on the lifespan of humans? |
|  | plot(Schooling, Life.expectancy, xlab = "Schooling", ylab = "Life Expectancy") |
|  | new3 = data\_frame(Schooling = seq(1, 30,1)) |
|  | lines(seq(1,30,1), predict(lm(Life.expectancy~Schooling), newdata = new3), col = "red", lwd = 4) |
|  | #There is a postive corelation between life expectancy and schooling |
|  |  |
|  | #Does Life Expectancy have positive or negative relationship with drinking alcohol? |
|  | plot(Alcohol, Life.expectancy, xlab = "Alcohol", ylab = "Life Expectancy") |
|  | new3 = data\_frame(Alcohol = seq(0.1, 30,.1)) |
|  | lines(seq(0.1,30,.1), predict(lm(Life.expectancy~Alcohol), newdata = new3), col = "red", lwd = 4) |
|  | #Do densely populated countries tend to have lower life expectancy? |
|  |  |
|  | #What is the impact of Immunization coverage on life Expectancy? |
|  |  |
|  |  |
|  | ``` |
|  | >>>>>>> 665aeb5e73cd46f0488d0dadf9155b4c1026b357 |
|  |  |
|  | ```{r EDA EDA EDA} |
|  | summary(who\_data) |
|  | who\_imp\_conti <- who\_imp[, !sapply(who\_imp, is.factor)] |
|  | who\_imp\_categ <- who\_imp[, sapply(who\_imp, is.factor)] |
|  |  |
|  | # Box plots to find outliers |
|  | boxplot(who\_imp\_conti[,2:20]) |
|  | boxplot(who\_imp\_conti$Population) |
|  | who\_imp[which.max(who\_imp$Population),] |
|  |  |
|  | # ScatterPlot |
|  | head(who\_imp\_conti) |
|  | pairs(who\_imp\_conti[,2:20], pch=19) |
|  |  |
|  | # Computing the p value of correlations |
|  | cor.mtest <- function(mat, ...) { |
|  | mat <- as.matrix(mat) |
|  | n <- ncol(mat) |
|  | p.mat<- matrix(NA, n, n) |
|  | diag(p.mat) <- 0 |
|  | for (i in 1:(n - 1)) { |
|  | for (j in (i + 1):n) { |
|  | tmp <- cor.test(mat[, i], mat[, j], ...) |
|  | p.mat[i, j] <- p.mat[j, i] <- tmp$p.value |
|  | } |
|  | } |
|  | colnames(p.mat) <- rownames(p.mat) <- colnames(mat) |
|  | p.mat |
|  | } |
|  | p.mat <- cor.mtest(who\_imp\_conti[,2:20]) |
|  | correlation <- cor(who\_imp\_conti[,2:20]) |
|  | # Correlation plot with significance level of 0.05 |
|  | corrplot(correlation, type="upper", order="hclust", p.mat = p.mat, sig.level = 0.05) |
|  |  |
|  | cat\_cor<- GKtauDataframe(who\_imp\_categ) |
|  | plot(cat\_cor, corrColors = "blue") |
|  | ``` |
|  |  |
|  | EDA |
|  | 1. Out of total 183 countries, 151 countries are developing and 32 are developed |
|  | 2. Outliers - Population of India seemed as an outlier, but that is a fact. |
|  | 3. Correlation |
|  | Percentage.expenditure - GDP |
|  | Polio - Hepatitis.B, Diptheria |
|  | Life.expectancy - Income.composition.of.resources, Schooling, Adult.Mortality, HIV.AIDS |
|  | Income.composition.of.resources - Schooling, Adult.Mortality, HIV.AIDS |
|  | Schooling - Adult.Mortality |
|  | thinness..1.19.years - thinness..5.9.years |
|  | Population - infant.deaths, under.five.deaths |
|  | infant.deaths - under.five.deaths |
|  |  |
|  | dev.off() |
|  |  |
|  | ```{r Full Model with all predictors} |
|  | par(mfrow=c(2,2)) |
|  | model\_full = lm(Life.expectancy ~ Adult.Mortality + infant.deaths + Alcohol + percentage.expenditure + Hepatitis.B + |
|  | Measles + BMI + under.five.deaths + Polio + Total.expenditure + Diphtheria + HIV.AIDS + GDP + |
|  | Population + thinness..1.19.years + thinness.5.9.years + Income.composition.of.resources + |
|  | Schooling,data=train) |
|  | scope\_formula = formula(model\_full) |
|  | scope\_formula |
|  | model\_full\_selection = step(object=model\_full, scope=scope\_formula, direction="backward") |
|  | summary(model\_full\_selection) |
|  | plot(model\_full\_selection) |
|  |  |
|  | model\_full\_selection\_aic = AIC(model\_full\_selection) |
|  | model\_full\_selection\_bic = BIC(model\_full\_selection) |
|  | model\_full\_selection\_aic # AIC of the model using forward selection method |
|  | model\_full\_selection\_bic |
|  | ASE=rss(model\_full\_selection)/nrow(train) |
|  | ASE |
|  | ``` |
|  | ================================================================================== |
|  | From the Full Model the below are the significant predictors - Forward Selection |
|  | ================================================================================== |
|  |  |
|  | Adult.Mortality -1.730e-02 4.586e-03 -3.772 0.000263 \*\*\* |
|  | infant.deaths 3.647e-01 1.102e-01 3.309 0.001270 \*\* |
|  | Alcohol 1.928e-01 9.368e-02 2.058 0.041997 \* |
|  | under.five.deaths -2.948e-01 8.244e-02 -3.576 0.000521 \*\*\* |
|  | HIV.AIDS -5.697e-01 2.959e-01 -1.926 0.056758 . |
|  | Income.composition.of.resources 3.141e+01 5.572e+00 5.638 1.37e-07 \*\*\* |
|  |  |
|  | Residual standard error: 3.323 on 109 degrees of freedom |
|  | Multiple R-squared: 0.8817, Adjusted R-squared: 0.8622 |
|  | AIC=690.1365 BIC=747.1771 |
|  | ================================================================================== |
|  | From the Full Model the below are the significant predictors - Backward Selection |
|  | ================================================================================== |
|  |  |
|  | Adult.Mortality -0.019037 0.004314 -4.413 2.24e-05 \*\*\* |
|  | infant.deaths 0.346787 0.091338 3.797 0.000231 \*\*\* |
|  | Alcohol 0.190732 0.084952 2.245 0.026588 \* |
|  | under.five.deaths -0.278161 0.072770 -3.822 0.000211 \*\*\* |
|  | Total.expenditure 0.208568 0.106121 1.965 0.051680 . |
|  | HIV.AIDS -0.542359 0.271797 -1.995 0.048258 \* |
|  | Income.composition.of.resources 32.084721 3.092474 10.375 < 2e-16 \*\*\* |
|  |  |
|  | Residual standard error: 3.238 on 120 degrees of freedom |
|  | Multiple R-squared: 0.8764, Adjusted R-squared: 0.8692 |
|  | AIC=673.8125 BIC=699.4808 |
|  | ================================================================================== |
|  | From the Full Model the below are the significant predictors - Stepwise Selection |
|  | ================================================================================== |
|  |  |
|  | Adult.Mortality -0.019037 0.004314 -4.413 2.24e-05 \*\*\* |
|  | infant.deaths 0.346787 0.091338 3.797 0.000231 \*\*\* |
|  | Alcohol 0.190732 0.084952 2.245 0.026588 \* |
|  | under.five.deaths -0.278161 0.072770 -3.822 0.000211 \*\*\* |
|  | Total.expenditure 0.208568 0.106121 1.965 0.051680 . |
|  | HIV.AIDS -0.542359 0.271797 -1.995 0.048258 \* |
|  | Income.composition.of.resources 32.084721 3.092474 10.375 < 2e-16 \*\*\* |
|  |  |
|  | Residual standard error: 3.238 on 120 degrees of freedom |
|  | Multiple R-squared: 0.8764, Adjusted R-squared: 0.8692 |
|  | AIC=673.8125 BIC=699.4808 |
|  | ================================================================================== |
|  |  |
|  | dev.off() |
|  | ```{r reduced model1- from manual selection, echo=FALSE} |
|  | par(mfrow=c(2,2)) |
|  | reduced\_model\_1 <- lm(Life.expectancy~Schooling+Income.composition.of.resources+Alcohol+ HIV.AIDS, data=train) |
|  | get\_regression\_table(reduced\_model\_1) |
|  | summary(reduced\_model\_1) |
|  | # Model validation using test set |
|  | test$Life.expectancy.Predicted <- predict(reduced\_model\_1, subset(test,select=-c(Country,Life.expectancy))) |
|  | test.correlation <- round(cor(test$Life.expectancy, test$Life.expectancy.Predicted),4) |
|  | test.RMSE <- round(sqrt(mean(test$Life.expectancy-test$Life.expectancy.Predicted)^2),4) |
|  | c(correlation = test.correlation, RMSE = test.RMSE) |
|  | vif(reduced\_model\_1) |
|  |  |
|  | reduced\_model\_1\_aic = AIC(reduced\_model\_1) |
|  | reduced\_model\_1\_bic = BIC(reduced\_model\_1) |
|  | reduced\_model\_1\_aic # AIC of the model using forward selection method |
|  | reduced\_model\_1\_bic |
|  |  |
|  | plot(reduced\_model\_1) |
|  | # Scatterplot |
|  | pairs(data.frame(train$Life.expectancy,train$Schooling,train$Alcohol,train$Income.composition.of.resources,train$HIV.AIDS), pch=19) |
|  | ASE=rss(reduced\_model\_1)/nrow(train) |
|  | ASE |
|  | ``` |
|  |  |
|  | From the Reduced Model (manual selection) the below are the significant predictors |
|  | 1.Income.composition.of.resources |
|  | 2.HIV.AIDS |
|  | Residual standard error: 3.708 on 123 degrees of freedom |
|  | Multiple R-squared: 0.8159, Adjusted R-squared: 0.81 |
|  | correlation RMSE |
|  | 0.8378 0.5866 |
|  | VIF |
|  | Schooling = 6.275441 |
|  | Income.composition.of.resources = 6.636833 |
|  | Alcohol = 1.571182 |
|  | HIV.AIDS = 1.254619 |
|  | AIC=705.623 BIC=722.7352 |
|  |  |
|  | dev.off() |
|  |  |
|  | ```{r reduced model2 - from stepwise selection, echo=FALSE} |
|  | par(mfrow=c(2,2)) |
|  | reduced\_model\_2 <- lm(Life.expectancy~Adult.Mortality+infant.deaths+Alcohol+under.five.deaths+Total.expenditure+ HIV.AIDS+Income.composition.of.resources, data=train) |
|  | get\_regression\_table(reduced\_model\_2) |
|  | summary(reduced\_model\_2) |
|  | # Model validation using test set |
|  | test$Life.expectancy.Predicted <- predict(reduced\_model\_2, subset(test,select=-c(Country,Life.expectancy))) |
|  | test.correlation <- round(cor(test$Life.expectancy, test$Life.expectancy.Predicted),4) |
|  | test.RMSE <- round(sqrt(mean(test$Life.expectancy-test$Life.expectancy.Predicted)^2),4) |
|  | c(correlation = test.correlation, RMSE = test.RMSE) |
|  | vif(reduced\_model\_2) |
|  | plot(reduced\_model\_2) |
|  | reduced\_model\_2\_aic = AIC(reduced\_model\_2) |
|  | reduced\_model\_2\_bic = BIC(reduced\_model\_2) |
|  | reduced\_model\_2\_aic # AIC of the model using forward selection method |
|  | reduced\_model\_2\_bic |
|  | # Scatterplot |
|  | pairs(data.frame(train$Life.expectancy, train$Adult.Mortality,train$Total.expenditure,train$Income.composition.of.resources,train$HIV.AIDS), pch=19) |
|  | ASE=rss(reduced\_model\_2)/nrow(train) |
|  | ASE |
|  | ``` |
|  | ================================================================================== |
|  | From the Reduced Model (stepwise selection) the below are the significant predictors |
|  | ================================================================================== |
|  |  |
|  | Adult.Mortality -0.019037 0.004314 -4.413 2.24e-05 \*\*\* |
|  | infant.deaths 0.346787 0.091338 3.797 0.000231 \*\*\* |
|  | Alcohol 0.190732 0.084952 2.245 0.026588 \* |
|  | under.five.deaths -0.278161 0.072770 -3.822 0.000211 \*\*\* |
|  | Total.expenditure 0.208568 0.106121 1.965 0.051680 . |
|  | HIV.AIDS -0.542359 0.271797 -1.995 0.048258 \* |
|  | Income.composition.of.resources 32.084721 3.092474 10.375 < 2e-16 \*\*\* |
|  |  |
|  | Residual standard error: 3.238 on 120 degrees of freedom |
|  | Multiple R-squared: 0.8764, Adjusted R-squared: 0.8692 |
|  | correlation RMSE |
|  | 0.7630 0.4363 |
|  | VIF |
|  | Adult.Mortality - 2.823806 |
|  | infant.deaths - 800.207011 |
|  | Alcohol - 1.497704 |
|  | under.five.deaths - 805.376827 |
|  | Total.expenditure - 1.116503 |
|  | HIV.AIDS - 1.811741 |
|  | Income.composition.of.resources - 2.760631 |
|  | AIC=673.8125 BIC=699.4808 |
|  | ================================================================================== |
|  |  |
|  | dev.off() |
|  |  |
|  | ```{r LASSO, echo=FALSE} |
|  | set.seed(100) |
|  | #Formatting data for GLM net |
|  | x=model.matrix(Life.expectancy ~ Adult.Mortality + infant.deaths + Alcohol + percentage.expenditure + Hepatitis.B + |
|  | Measles + BMI + under.five.deaths + Polio + Total.expenditure + Diphtheria + HIV.AIDS + GDP + |
|  | Population + thinness..1.19.years + thinness.5.9.years + Income.composition.of.resources + |
|  | Schooling, train)[,-1] |
|  | Life.expectancy~Schooling+Income.composition.of.resources+Total.expenditure+percentage.expenditure+BMI+Alcohol |
|  | y=train$Life.expectancy |
|  |  |
|  | xtest<-model.matrix(Life.expectancy ~ Adult.Mortality + infant.deaths + Alcohol + percentage.expenditure + Hepatitis.B + |
|  | Measles + BMI + under.five.deaths + Polio + Total.expenditure + Diphtheria + HIV.AIDS + GDP + |
|  | Population + thinness..1.19.years + thinness.5.9.years + Income.composition.of.resources + |
|  | Schooling,test)[,-1] |
|  | ytest<-test$Life.expectancy |
|  |  |
|  | grid=10^seq(10,-2, length =100) |
|  | lasso.mod=glmnet(x,y,alpha=1, lambda =grid) |
|  |  |
|  | cv.out=cv.glmnet(x,y,alpha=1) #alpha=1 performs LASSO |
|  | plot(cv.out) |
|  | bestlambda<-cv.out$lambda.min #Optimal penalty parameter. You can make this call visually. |
|  | lasso.pred=predict (lasso.mod ,s=bestlambda ,newx=xtest) |
|  |  |
|  | testMSE\_LASSO<-mean((ytest-lasso.pred)^2) |
|  | testMSE\_LASSO |
|  | coef(lasso.mod,s=bestlambda) |
|  | ``` |
|  | ================ |
|  | Lasso Predictors |
|  | ================ |
|  | Adult.Mortality -0.0191246691 |
|  | Alcohol 0.1290622227 |
|  | Total.expenditure 0.0559652197 |
|  | Diphtheria 0.0002095938 |
|  | HIV.AIDS -0.4152214620 |
|  | thinness..1.19.years -0.0494267615 |
|  | Income.composition.of.resources 34.2057499766 |
|  |  |
|  |  |
|  | dev.off() |
|  |  |
|  | ```{r reduced model3 - from Lasso selection, echo=FALSE} |
|  | par(mfrow=c(2,2)) |
|  | reduced\_model\_3 <- lm(Life.expectancy ~ Adult.Mortality + Alcohol + Total.expenditure + Diphtheria + HIV.AIDS + thinness..1.19.years + Income.composition.of.resources,data=train) |
|  | get\_regression\_table(reduced\_model\_3) |
|  | summary(reduced\_model\_3) |
|  | # Model validation using test set |
|  | test$Life.expectancy.Predicted <- predict(reduced\_model\_3, subset(test,select=-c(Country,Life.expectancy))) |
|  | test.correlation <- round(cor(test$Life.expectancy, test$Life.expectancy.Predicted),4) |
|  | test.RMSE <- round(sqrt(mean(test$Life.expectancy-test$Life.expectancy.Predicted)^2),4) |
|  | c(correlation = test.correlation, RMSE = test.RMSE) |
|  | vif(reduced\_model\_3) |
|  | plot(reduced\_model\_3) |
|  | reduced\_model\_3\_aic = AIC(reduced\_model\_3) |
|  | reduced\_model\_3\_bic = BIC(reduced\_model\_3) |
|  | reduced\_model\_3\_aic # AIC of the model using forward selection method |
|  | reduced\_model\_3\_bic |
|  | summary(reduced\_model\_3) |
|  | ASE=rss(reduced\_model\_3)/nrow(train) |
|  | ASE |
|  | ``` |
|  |  |
|  | dev.off() |
|  |  |
|  |  |
|  | ==================================================================================== |
|  | From the Reduced Model (lasso selection) the below are the significant predictors |
|  | ==================================================================================== |
|  |  |
|  | Adult.Mortality -0.019658 0.004567 -4.305 3.43e-05 \*\*\* |
|  | Alcohol 0.190417 0.089385 2.130 0.0352 \* |
|  | Total.expenditure 0.158346 0.114060 1.388 0.1676 |
|  | Diphtheria 0.014960 0.014128 1.059 0.2918 |
|  | HIV.AIDS -0.631963 0.284766 -2.219 0.0284 \* |
|  | thinness..1.19.years -0.112665 0.096625 -1.166 0.2459 |
|  | Income.composition.of.resources 33.338972 3.397006 9.814 < 2e-16 \*\*\* |
|  |  |
|  | Residual standard error: 3.402 on 120 degrees of freedom |
|  | Multiple R-squared: 0.8636, Adjusted R-squared: 0.8556 |
|  | correlation RMSE |
|  | 0.8026 0.2577 |
|  | VIF |
|  | Adult.Mortality - 2.867694 |
|  | Alcohol - 1.502603 |
|  | Total.expenditure - 1.168843 |
|  | Diphtheria - 1.149265 |
|  | HIV.AIDS - 1.802256 |
|  | thinness..1.19.years - 1.544284 |
|  | Income.composition.of.resources - 3.018706 |
|  | AIC=686.4176 BIC=712.0859 |
|  | ================================================================================== |
|  |  |
|  | ```{r model4 - Life Expectancy differences between developed and devloping countries, echo=FALSE} |
|  | par(mfrow=c(2,2)) |
|  | reduced\_model\_4 <- lm(Life.expectancy ~ Status ,data=who\_data) |
|  | get\_regression\_table(reduced\_model\_4) |
|  | summary(reduced\_model\_4) |
|  | ``` |
|  | ==================================================================================== |
|  | From the Linear model Life Expectancy for Developing vs Developed |
|  | ==================================================================================== |
|  | (Intercept) 81.138 1.298 62.493 < 2e-16 \*\*\* |
|  | StatusDeveloping -11.636 1.429 -8.141 6.17e-14 \*\*\* |
|  |  |
|  | 1. The P value of dummy variable StatusDeveloping is very significant suggesting that |
|  | there is statistical evidence of difference in average life expectancy between developing and developed countries. |
|  | 2. The average life expectancy in developed countries is estimated to be 81.138 |
|  | 3. The average life expectancy in developing countries is 11.638 years less than developed countries. |
|  |  |
|  |  |
|  | dev.off() |
|  |  |
|  | ```{r Final Model1 , echo=FALSE} |
|  | set.seed(100) |
|  | par(mfrow=c(2,2)) |
|  | final\_model1 <- lm(Life.expectancy ~ Adult.Mortality + Alcohol + Total.expenditure + HIV.AIDS + Income.composition.of.resources + thinness..1.19.years,data=train) |
|  | get\_regression\_table(final\_model1) |
|  | summary(final\_model1) |
|  | # Model validation using test set |
|  | test$Life.expectancy.Predicted <- predict(final\_model1, subset(test,select=-c(Life.expectancy))) |
|  | test.correlation <- round(cor(test$Life.expectancy, test$Life.expectancy.Predicted),4) |
|  | test.RMSE <- round(sqrt(mean(test$Life.expectancy-test$Life.expectancy.Predicted)^2),4) |
|  | c(correlation = test.correlation, RMSE = test.RMSE) |
|  | vif(final\_model1) |
|  | plot(final\_model1) |
|  | final\_model1\_aic = AIC(final\_model1) |
|  | final\_model1\_bic = BIC(final\_model1) |
|  | final\_model1\_aic # AIC of the model using forward selection method |
|  | final\_model1\_bic |
|  | summary(final\_model1) |
|  | ASE=rss(final\_model1)/nrow(train) |
|  | ASE |
|  |  |
|  | ``` |
|  |  |
|  |  |
|  |  |
|  | ```{r Final Model2 , echo=FALSE} |
|  | par(mfrow=c(2,2)) |
|  | final\_model2 <- lm(Life.expectancy ~ Adult.Mortality + Total.expenditure + Income.composition.of.resources ,data=train) |
|  | get\_regression\_table(final\_model2) |
|  | summary(final\_model2) |
|  | # Model validation using test set |
|  | test$Life.expectancy.Predicted <- predict(final\_model2, subset(test,select=-c(Life.expectancy)), interval="confidence") |
|  | test.correlation <- round(cor(test$Life.expectancy, test$Life.expectancy.Predicted),4) |
|  | test.RMSE <- round(sqrt(mean(test$Life.expectancy-test$Life.expectancy.Predicted)^2),4) |
|  | c(correlation = test.correlation, RMSE = test.RMSE) |
|  | vif(final\_model2) |
|  | plot(final\_model2) |
|  | final\_model2\_aic = AIC(final\_model2) |
|  | final\_model2\_bic = BIC(final\_model2) |
|  | final\_model2\_aic # AIC of the model using forward selection method |
|  | final\_model2\_bic |
|  | summary(final\_model2) |
|  | ASE=rss(final\_model2)/nrow(train) |
|  | ASE |
|  | ``` |
|  |  |
|  |  |
|  | dev.off() |
|  |  |
|  | =========== |
|  | OBJECTIVE 2 |
|  | =========== |
|  |  |
|  | ```{r knn with all the predictors} |
|  | # Train model with knn and get importance of predictors |
|  | control <- trainControl(method="repeatedcv", number=10, repeats=3) |
|  | set.seed(100) |
|  | model.knn <- train( Life.expectancy ~ ., data=who\_imp[,-c(2)], method="knn", trControl=control) |
|  | #Top 10 predictor ranking |
|  | importance.knn <- varImp(model.knn, scale=FALSE) |
|  | rank.knn <- importance.knn$importance |
|  | write.csv(rank.knn, "rank.knn.csv") |
|  | rank.knn <- read.csv("rank.knn.csv", header=TRUE) |
|  | colnames(rank.knn) <- c("Predictors", "Importance") |
|  | rank.knn <- rank.knn[order(rank.knn$Importance, decreasing = TRUE),] |
|  | ggplot(rank.knn[1:20,], aes(x=reorder(Predictors, Importance),y=Importance)) + geom\_bar(stat = "identity") + coord\_flip() + labs(title="Importance of Predictors", x="Predictors", y="Importance") +theme(axis.text.x=element\_text(hjust=0.5, vjust=0.5, size = 12))+theme(axis.text.y=element\_text(size = 12)) |
|  | ``` |
|  |  |
|  |  |
|  | ```{r knn with all the predictors} |
|  | # Train model with knn and get importance of predictors |
|  | control <- trainControl(method="repeatedcv", number=10, repeats=3) |
|  | set.seed(100) |
|  | model.knn <- train(Life.expectancy ~ Adult.Mortality + infant.deaths + Alcohol + percentage.expenditure + Hepatitis.B + |
|  | Measles + BMI + under.five.deaths + Polio + Total.expenditure + Diphtheria + HIV.AIDS + GDP + |
|  | Population + thinness..1.19.years + thinness.5.9.years + Income.composition.of.resources + |
|  | Schooling, data=who\_imp[,-c(2)], method="knn", trControl=control) |
|  | #Top 10 predictor ranking |
|  | importance.knn <- varImp(model.knn, scale=FALSE) |
|  | rank.knn <- importance.knn$importance |
|  | write.csv(rank.knn, "rank.knn.csv") |
|  | rank.knn <- read.csv("rank.knn.csv", header=TRUE) |
|  | colnames(rank.knn) <- c("Predictors", "Importance") |
|  | rank.knn <- rank.knn[order(rank.knn$Importance, decreasing = TRUE),] |
|  | ggplot(rank.knn[1:20,], aes(x=reorder(Predictors, Importance),y=Importance)) + geom\_bar(stat = "identity") + coord\_flip() + labs(title="Importance of Predictors", x="Predictors", y="Importance") +theme(axis.text.x=element\_text(hjust=0.5, vjust=0.5, size = 12))+theme(axis.text.y=element\_text(size = 12)) |
|  | ``` |
|  |  |
|  | Life.expectancy ~ Adult.Mortality + Total.expenditure + Income.composition.of.resources |
|  | Adult.Mortality + Alcohol + Total.expenditure + HIV.AIDS + Income.composition.of.resources + thinness..1.19.years |
|  |  |
|  | ```{r} |
|  | library(dplyr) |
|  | lifeReduced = who\_imp %>% select(Life.expectancy, Adult.Mortality, Total.expenditure, Income.composition.of.resources) |
|  | lifeFull = who\_imp %>% select(Life.expectancy, Adult.Mortality, Alcohol, Total.expenditure, HIV.AIDS, Income.composition.of.resources, thinness..1.19.years) |
|  | ``` |
|  |  |
|  | ```{r knn life reduced} |
|  | set.seed(100) |
|  | iterations = 10 |
|  | numks = 30 |
|  | splitPerc = .70 |
|  | masterAcc = matrix(nrow = iterations, ncol = numks) |
|  | for(j in 1:iterations) |
|  | { |
|  | trainIndices = sample(1:dim(lifeReduced)[1],round(splitPerc \* dim(lifeReduced)[1])) |
|  | train = lifeReduced[trainIndices,] |
|  | test = lifeReduced[-trainIndices,] |
|  | for(i in 1:numks) |
|  | { |
|  | classifications = knn(train[,c(2,3,4)],test[,c(2,3,4)],train$Life.expectancy, prob = TRUE, k = i) |
|  | u <- union(classifications,test$Life.expectancy) |
|  | t <- table(factor(classifications, u), factor(test$Life.expectancy, u)) |
|  | CM = confusionMatrix(t) |
|  | masterAcc[j,i] = CM$overall[1] |
|  | } |
|  | } |
|  | MeanAcc = colMeans(masterAcc) |
|  | plot(seq(1,numks,1),MeanAcc, type = "l") |
|  | which.max(MeanAcc) |
|  | max(MeanAcc) |
|  | ``` |
|  |  |
|  |  |
|  | ```{r knn life full} |
|  | set.seed(100) |
|  | iterations = 10 |
|  | numks = 30 |
|  | splitPerc = .70 |
|  | masterAcc = matrix(nrow = iterations, ncol = numks) |
|  | for(j in 1:iterations) |
|  | { |
|  | trainIndices = sample(1:dim(lifeFull)[1],round(splitPerc \* dim(lifeFull)[1])) |
|  | train = lifeFull[trainIndices,] |
|  | test = lifeFull[-trainIndices,] |
|  | for(i in 1:numks) |
|  | { |
|  | classifications = knn(train[,c(2,3,4,5,6,7)],test[,c(2,3,4,5,6,7)],train$Life.expectancy, prob = TRUE, k = i) |
|  | u <- union(classifications,test$Life.expectancy) |
|  | t <- table(factor(classifications, u), factor(test$Life.expectancy, u)) |
|  | CM = confusionMatrix(t) |
|  | masterAcc[j,i] = CM$overall[1] |
|  | } |
|  | } |
|  | MeanAcc = colMeans(masterAcc) |
|  | plot(seq(1,numks,1),MeanAcc, type = "l") |
|  | which.max(MeanAcc) |
|  | max(MeanAcc) |
|  | ``` |