DSC424 - Final Project Milestone 3
Name: Andrew Fitzsimmons
Date: 04/30/2022

Data Cleaning and Handling Missing Data

R Code:

```
#retreive data from file location
setwd("C:/Users/apfit/OneDrive/Documents/Depaul/Spring 2022/Advanced Data Analysis/R datasources")
WHO = read.csv('who_life_exp.csv')

#display summary of data
summary(WHO)
WHO_Summary= WHO%>%
  group_by(country)%>%
  summarise(Mean_Life_Expectancy = mean(life_expect))%>%
  arrange(desc(Mean_Life_Expectancy))

#determine which columns have 'NA' values
colSums(is.na(WHO))
```

Description:
The first step in my data cleaning process was to pull the dataset that we are going to be using for our project into R. This was completed using the read.csv function. Once the dataset was in R, I looked at a summary of the data set, broken down by predictors.

```
> summary(WHO)
   country          country_code          region              year       life_expect      life_exp60    adult_mortality  infant_mort
 Length:3111        Length:3111        Length:3111        Min.   :2000   Min.   :36.23   Min.   :10.73   Min.   : 49.2   Min.   :0.001470
 Class :character   Class :character   Class :character   1st Qu.:2004   1st Qu.:63.20   1st Qu.:16.62   1st Qu.:108.3   1st Qu.:0.008255
 Mode  :character   Mode  :character   Mode  :character   Median :2008   Median :71.60   Median :18.51   Median :164.8   Median :0.019995
                                                          Mean   :2008   Mean   :69.15   Mean   :18.91   Mean   :193.5   Mean   :0.032496
                                                          3rd Qu.:2012   3rd Qu.:75.54   3rd Qu.:21.10   3rd Qu.:250.8   3rd Qu.:0.051720
                                                          Max.   :2016   Max.   :84.17   Max.   :26.39   Max.   :696.9   Max.   :0.164515

  age1.4mort          alcohol            bmi         age5.19thinness  age5.19obesity     hepatitis         measles          polio
 Min.   :0.000065   Min.   : 0.000   Min.   :19.80   Min.   : 0.100   Min.   : 0.100   Min.   : 2.00   Min.   :16.00   Min.   : 8.00
 1st Qu.:0.000355   1st Qu.: 1.198   1st Qu.:23.30   1st Qu.: 1.800   1st Qu.: 2.000   1st Qu.:81.00   1st Qu.:79.00   1st Qu.:81.00
 Median :0.000895   Median : 3.994   Median :25.50   Median : 3.800   Median : 5.200   Median :92.00   Median :92.00   Median :93.00
 Mean   :0.003489   Mean   : 4.835   Mean   :25.05   Mean   : 5.312   Mean   : 5.972   Mean   :85.44   Mean   :85.54   Mean   :86.61
 3rd Qu.:0.004877   3rd Qu.: 7.723   3rd Qu.:26.50   3rd Qu.: 7.800   3rd Qu.: 8.900   3rd Qu.:97.00   3rd Qu.:96.00   3rd Qu.:97.00
 Max.   :0.039095   Max.   :20.182   Max.   :32.20   Max.   :28.100   Max.   :26.700   Max.   :99.00   Max.   :99.00   Max.   :99.00
                    NA's   :50       NA's   :34      NA's   :34       NA's   :34       NA's   :569     NA's   :19      NA's   :19
   diphtheria     basic_water        doctors          hospitals         gni_capita         gghe.d           che_gdp           une_pop
 Min.   :19.00   Min.   : 18.70   Min.   : 0.128   Min.   : 0.0000   Min.   :   250   Min.   : 0.06236   Min.   : 1.025   Min.   :     76
 1st Qu.:82.00   1st Qu.: 71.66   1st Qu.: 6.391   1st Qu.: 0.5352   1st Qu.:  2540   1st Qu.: 1.53344   1st Qu.: 4.239   1st Qu.:   2195
 Median :93.00   Median : 91.99   Median :20.523   Median : 1.0727   Median :  7460   Median : 2.60130   Median : 5.758   Median :   8544
 Mean   :86.42   Mean   : 83.33   Mean   :19.866   Mean   : 2.0449   Mean   : 13397   Mean   : 3.12293   Mean   : 6.110   Mean   :  37076
 3rd Qu.:97.00   3rd Qu.: 98.55   3rd Qu.:30.982   3rd Qu.: 2.1048   3rd Qu.: 18250   3rd Qu.: 4.27811   3rd Qu.: 7.850   3rd Qu.:  25096
 Max.   :99.00   Max.   :100.00   Max.   :79.541   Max.   :56.4470   Max.   :123860   Max.   :12.06273   Max.   :20.413   Max.   :1414049
 NA's   :19      NA's   :32       NA's   :1331     NA's   :2981      NA's   :682      NA's   :100        NA's   :117      NA's   :37
   une_infant        une_life         une_hiv          une_gni         une_poverty    une_edu_spend     une_literacy      une_school
 Min.   :  1.60   Min.   :39.44   Min.   : 0.100   Min.   :   420   Min.   : 0.10   Min.   : 0.7874   Min.   : 14.38   Min.   : 0.5593
 1st Qu.:  8.00   1st Qu.:62.84   1st Qu.: 0.100   1st Qu.:  2970   1st Qu.: 0.60   1st Qu.: 3.2628   1st Qu.: 72.70   1st Qu.: 7.7359
 Median : 19.50   Median :71.41   Median : 0.400   Median :  8340   Median : 3.10   Median : 4.4254   Median : 90.95   Median :10.2704
 Mean   : 30.49   Mean   :68.96   Mean   : 2.038   Mean   : 14965   Mean   :10.85   Mean   : 4.5329   Mean   : 81.98   Mean   : 9.7122
 3rd Qu.: 48.05   3rd Qu.:75.57   3rd Qu.: 1.500   3rd Qu.: 20483   3rd Qu.:12.40   3rd Qu.: 5.4950   3rd Qu.: 95.79   3rd Qu.:12.0706
 Max.   :142.40   Max.   :83.98   Max.   :28.200   Max.   :122670   Max.   :94.10   Max.   :14.0591   Max.   :100.00   Max.   :14.3788
                                  NA's   :741      NA's   :117      NA's   :2198    NA's   :1286      NA's   :2540     NA's   :2306
```

Based on the above summary, it looks like the dataset contains some categorical variables in the way of country and country_code, an ordinal variable in year, and the remaining variables appear to be continuous and numeric. I noticed that some of the variables have an incredibly large gap between the 3rd quartile and Max value (gni_capita, une_pop, uni_gni for example), this indicates that these variables may be good candidates for log or exponential scaling, I made a note to take a look at the distributions of these variables later in the process when I begin to scale and transform my data for use in a regression model. I also noticed that I have quite a few predictors that are very sparsely populated. Une_school, une_literacy, une_poverty, and hospitals all have over 2000 NA entries out of a total of 3111 rows of data. I will likely need to either determine how to merge outside data with these variables, or remove them from the dataset, since the combination of those 5 variables are unlikely to yield many complete entries.

I then generated a data frame called WHO_Summary. This data frame groups the data by the categorical variable ‘country’, and generates a mean value for the life expectancy target variable by country. The resulting table is then sorted by descending order on mean life expectancy. This table gives me some good initial insight into the dataset. I now have a feel for the range of life expectancy values that we’ll be working with (from 82 years to 46 years). I am also able to see that the top countries in terms of life expectancy, tend to be more developed countries, while the bottom countries on the list tend to be less developed countries. This makes sense, as more developed countries may have a higher GDP per capita, more access to clean water, better healthcare, and more access to vaccines/immunizations. These are all predictors that exist within the data set, so we should hopefully be able to determine whether there are strong correlations between these predictors (among others) and our target variable of life expectancy.

	country	Mean_Life_Expectancy
1	Japan	82.75357
2	Switzerland	81.84819
3	Australia	81.54399
4	Iceland	81.50785
5	Italy	81.49052
6	Spain	81.39119

...

	country	Mean_Life_Expectancy
179	Chad	50.83169
180	Eswatini	50.26028
181	Lesotho	49.92741
182	Central African Republic	48.21520
183	Sierra Leone	46.79126

Dealing with NA values in Alcohol predictor

R Code:

```
#determine which columns have 'NA' values
colSums(is.na(WHO))

#review alcohol NAs
WHO%>%
  filter(is.na(alcohol))%>%
  group_by(country)%>%
  summarise()

WHO%>%
  filter(country=='Canada')%>%
  select(country, alcohol)

WHO%>%
  filter(country=='Afghanistan')%>%
  select(country, alcohol)

WHO%>%
  filter(country=='Montenegro')%>%
  select(country, alcohol)
```

```
WHO%>%
  filter(country=='Serbia')%>%
  select(country, alcohol)

WHO%>%
  filter(country=='Sudan')%>%
  select(country, alcohol)

WHO%>%
  filter(country=='South Sudan')%>%
  select(country, alcohol)

#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)

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

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

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

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

#remove south sudan
WHO = WHO%>%
  filter(country!='South Sudan')

#REVIEW BMI NAs
WHO%>%
  filter(is.na(bmi))%>%
  group_by(country)%>%
  summarise()
```

```
WHO%>%
  filter(country=='Sudan')%>%
  select(country, bmi)

#remove Sudan due to no data present in bmi

WHO = WHO%>%
  filter(country!='Sudan')
```

Description:

Now that I've identified the need to deal with the NA values in the dataset. I'll first take a look at the countries that have missing values in the alcohol predictor.

```
> #review alcohol NAs
> WHO%>%
+   filter(is.na(alcohol))%>%
+   group_by(country)%>%
+   summarise()
# A tibble: 6 x 1
  country
  <chr>
1 Afghanistan
2 Canada
3 Montenegro
4 Serbia
5 South Sudan
6 Sudan
```

With a list of the 6 countries that I need to look into to handle the NA values in the alcohol predictor, I'll now start to look a step deeper and investigate the distribution of values within each country.

```
> WHO%>%
+   filter(country=='Canada')%>%
+   select(country, alcohol)
   country alcohol
1   Canada      NA
2   Canada      NA
3   Canada      NA
4   Canada      NA
5   Canada      NA
6   Canada     8.0
7   Canada     8.2
8   Canada     8.3
9   Canada     8.4
10  Canada     8.4
11  Canada     8.3
12  Canada     8.2
13  Canada     8.3
14  Canada     8.2
15  Canada     8.0
16  Canada     8.0
17  Canada     8.1
```

Above is an example of the alcohol values when filtering the dataset by country == 'Canada'. It appears that there is a fairly normal distribution of values between 8 and 8.4. This leads me to believe that in general, alcohol use in Canada has been fairly consistent over this time span. In order to handle the NA values here, I think it is appropriate to average the values that I do have for canada, and apply the average value to the NA columns. This is done by executing the below code to find NA values where country == 'Canada', and apply the mean value to those cells. The resulting table shows the NAs are now filled in with a value of 8.2. I repeated this exercise for the other 5 countries that I identified as needing cleaning for the alcohol variable, and determined that 4 could be handled the same way, but 'South Sudan' did not have any alcohol values present for any entry. I chose to remove 'South Sudan' from the dataset as a result. It had many other columns of predictors with a significant number of missing values as well.

```
> x='Canada'
> WHO$alcohol = ifelse(WHO$country== x & is.na(WHO$alcohol),
+                      mean(WHO[WHO$country == x ,]$alcohol,na.rm=TRUE),WHO$alcohol)
> WHO%>%
+   filter(country=='Canada')%>%
+   select(country, alcohol)
   country alcohol
1   Canada     8.2
2   Canada     8.2
3   Canada     8.2
4   Canada     8.2
5   Canada     8.2
6   Canada     8.0
7   Canada     8.2
8   Canada     8.3
9   Canada     8.4
10  Canada     8.4
11  Canada     8.3
12  Canada     8.2
13  Canada     8.3
14  Canada     8.2
15  Canada     8.0
16  Canada     8.0
17  Canada     8.1
```

Dealing with sparsely populated predictors

R Code:

```
#reduce data set to remove variables with significant missing values across all entries
colSums(is.na(WHO))

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)

WHO_Summary_Reduced= WHO_Reduced_CompleteCases%>%
  group_by(country)%>%
  summarise(Mean_Life_Expectancy = mean(life_expect))%>%
  arrange(desc(Mean_Life_Expectancy))
```

Description:

After reviewing the predictors with significant counts of NA values (over 1000), I decided it would be best to delete them from my dataset, since trying to maintain complete cases, with those predictors in, left with with just 2 rows of data out of my now 3077 rows (after removing sudan and south sudan data). With the sparsely populated predictors removed from my data

set, I am now able to search for entries that represent complete cases only. This filtering leaves me with a dataset that contains 1349 rows of complete entries. This is still a large data set, and since it only contains complete entries, I can now run a correlation plot on all predictors to see what additional inferences I can make about the data. I also rerun my summary table of average life expectancy by country, and notice that while I did lose ~60 countries from my dataset, I still have 121 countries represented, and my range in target variable is 81.4 - 48.7. I think this range is still representative of my original dataset, so I am happy with the result of the reduced, complete case set.

```
> 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
```

Generating and Reviewing Correlation Plot

R Code:

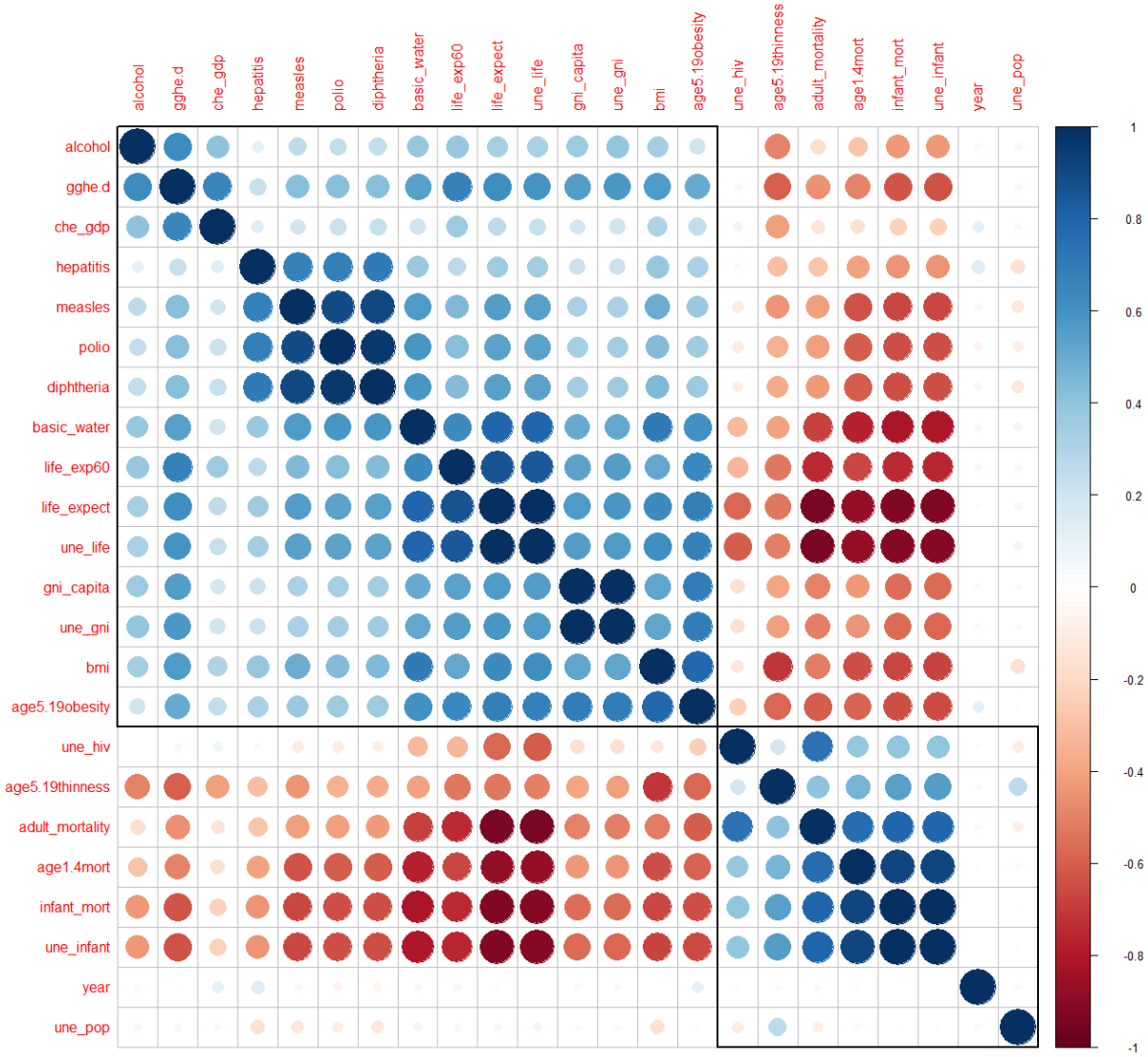
```
#correlation plot on reduced, complete case data set
library(corrplot)

WHOcor = cor(WHO_Reduced_CompleteCases%>%
             select(-country,-country_code,-region))
corrplot(WHOcor,order='hclust',addrect=2)

pairs(WHO_Reduced_CompleteCases%>%
      select(-country,-country_code,-region))
```

Description:

Now that I have a complete dataset, I can run a correlation plot on the numerical variables. I order the variables on the plot by ‘hclust’, which tries to group correlated predictors together when generating the plot. From the resulting plot, we can see that there are quite a few predictors with strong positive and negative correlation to life expectancy. Measles, polio, and diphtheria all represent % of 1 year old children that have received the vaccines. These all show a strong positive correlation with life expectancy. Basic_water also shows strong positive correlation to life expectancy as it represents a population with access to clean drinking water. BMI is also a strong positive correlation since it is a measure of malnourishment amongst the population. As less of the population is malnourished, the overall life expectancy increases. Une_hiv shows a strong negative correlation with life expectancy since it represents the % prevalence of HIV in the population between age 15 and 49.


alcohol
gghe.d
che_gdp
hepatitis
measles
polio
diphtheria
basic_water
life_exp60
life_expect
une_life
gni_capita
une_gni
bmi
age5.19obesity
une_hiv
age5.19thinness
adult_mortality
age1.4mort
infant_mort
une_infant
year
une_pop
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1