Group Assignment 1, MQSM:HA Group 1

Ashley Dickson, Simran Singh, Qi Tan, Jason Theiling

10/16/2022

World Happiness

The World Happiness Report is a landmark survey of the state of global happiness that ranks 156 countries by how happy their citizens perceive themselves to be.

More Information: https://worldhappiness.report/ed/2022/

Central to this survey is the "Life Evaluation" question. Life evaluation was measured by the individual answers to the Cantril ladder question: "Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?"

Reading the Data

0. ***Read in all the data provided to create a single data frame with Happiness metrics for the years from 2005 - 2021. Describe any changes you had to make to combine these data sets effectively.***

When reading the data in from multiple files a number of steps must be taken and several changes to the data were necessary. For full details, please see the following coding chunk with annotation.

In brief, multiple data frames must be created for the the data to be read into. This required reading both excel and csv files. For the excel, a new data frame was created for each sheet of data in the parent file. The csv was simply read in directly. These resulting data frames must then be combined in to a single working data frame. Unfortunately, the column names are inconsistent in the individual frames, so those must all be updated for identical syntax. Once this was accomplished the data frames are all joined using 'rbind'.

```
#Set the working directory to find the files.
#Then read each excel or csv into a named matrix

setwd('C:/Users/theil/OneDrive/Desktop/Coding Course/R files/class 2/Assignment 2')
happiness_data20052018 = read_csv("Happiness_2005-2018.csv", show_col_types = FALSE)
happiness_data2019 = read_xlsx("Happiness_2019-2021.xlsx", sheet = as.character(2019))
happiness_data2020 = read_xlsx("Happiness_2019-2021.xlsx", sheet = as.character(2020))
happiness_data2021 = read_xlsx("Happiness_2019-2021.xlsx", sheet = as.character(2021))

#Clean up column names before binding, making all names descriptive and identical

colnames(happiness_data20052018) = c('Country Name', 'Year', 'Life Ladder',
                                     'Log GDP per Capita', 'Social Support',
```

```
                                    'Healthy Life Expectancy at Birth',
                                    'Freedom to Make Life Choices', 'Generosity',
                                    'Perceptions of Corruption', 'Positive affect',
                                    'Negative Affect','Confidence in National Government')

colnames(happiness_data2019) = c('Country Name', 'Year', 'Life Ladder',
                                    'Log GDP per Capita', 'Social Support',
                                    'Healthy Life Expectancy at Birth',
                                    'Freedom to Make Life Choices', 'Generosity',
                                    'Perceptions of Corruption', 'Positive affect',
                                    'Negative Affect','Confidence in National Government')

colnames(happiness_data2020) = c('Country Name', 'Year', 'Life Ladder',
                                    'Log GDP per Capita', 'Social Support',
                                    'Healthy Life Expectancy at Birth',
                                    'Freedom to Make Life Choices', 'Generosity',
                                    'Perceptions of Corruption', 'Positive affect',
                                    'Negative Affect','Confidence in National Government')

colnames(happiness_data2021)= c('Country Name', 'Year', 'Life Ladder',
                                    'Log GDP per Capita', 'Social Support',
                                    'Healthy Life Expectancy at Birth',
                                    'Freedom to Make Life Choices', 'Generosity',
                                    'Perceptions of Corruption', 'Positive affect',
                                    'Negative Affect','Confidence in National Government')


happiness_data = rbind(happiness_data20052018, happiness_data2019, happiness_data2020,
                       happiness_data2021)
```

##Reading the Data: Using a for loop

The below r chunk walks through reading the multiple sheets of an excel file in to multiple data frames by making use of a 'for loop'.

```
# Reading the data for 2019... Can we write a for loop to make this faster? (ungraded)

#Set the working directory to find the files.

setwd('C:/Users/theil/OneDrive/Desktop/Coding Course/R files/class 2/Assignment 2')

# Create a list named 'happy' where each element corresponds
# to a distinct data frame for a specific year (2019, 2020, and 2021). Then access
# each year's data frame using double brackets, i.e. happy[["2019"]],
# happy[["2020"]], and happy[["2021"]]

# Create an empty list to store data frames

happy = list()

# Loop through the years reading each seet into a new matrix
for (i in 2019:2021) {
  sheet_name = as.character(i)
  happy_temp = read_xlsx("Happiness_2019-2021.xlsx", sheet = sheet_name)
```

```
  happy[[sheet_name]] = happy_temp
}
```

Initial Questions

1) **Do Canada and the United States have common happy/unhappy years?**

   Filter the combined dataset to view Happiness and associated variables for the United States and Canada over all available years. Find the 3 happiest and unhappiest years for each country in the data provided. Do they seem to align with one another? Are there any key features that differ over the years you selected?

When the data sets are analyzed the following were found:

1) The happiest years in the US were 2007, 2008 and 2013.
2) The happiest years in Canada were 2010, 2013 and 2009.
3) Of these only 2013 were in the top happiest years for both countries.
4) The least happy years in the US were 2016, 2015 and 2018.
5) The least happy years in Canada were 2019, 2020 and 2021.
6) None of the least happy years overlap.
7) By correlating the data within the data frames for each countries happiest and least happy years one can identify the factors for each country that are more strongly correlated to overall life ladder scores, ie what factor most influence happiness in each country.

   7a) For the US it appears that Social Support, Generosity, Positive Affect and Life Expectancy at Birth most strongly influenced global happiness.

   7b) For Canada it appears that Positive Affect, Freedom to Make Choices, Social Support and Generosity most strongly influenced global happiness.

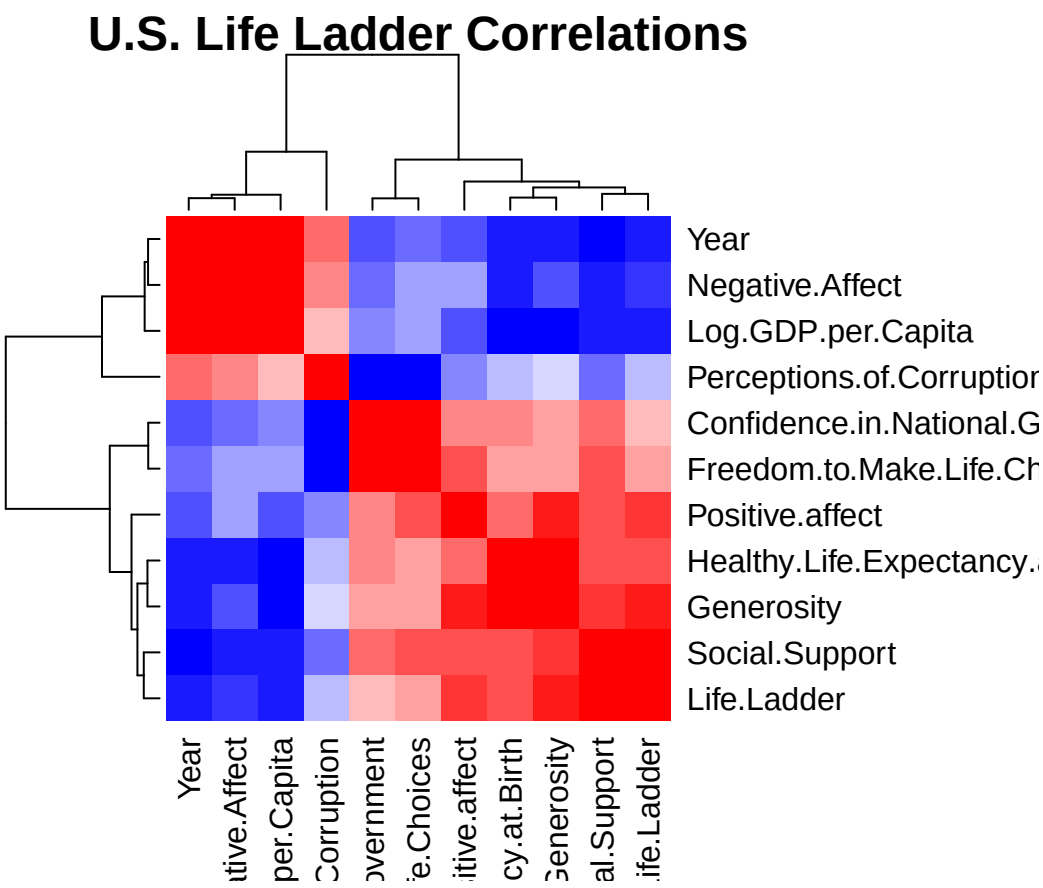
   7c) Remarkably enough, Generosity, Positive Affect and Social Support were central to both groups happiness!

```
#Set working Directory
setwd('C:/Users/theil/OneDrive/Desktop/Coding Course/R files/class 2/Assignment 2')

#Filter the data frame to only include the desired countries

happiness_data_US = happiness_data %>%
    filter(happiness_data[1]=='United States')
```

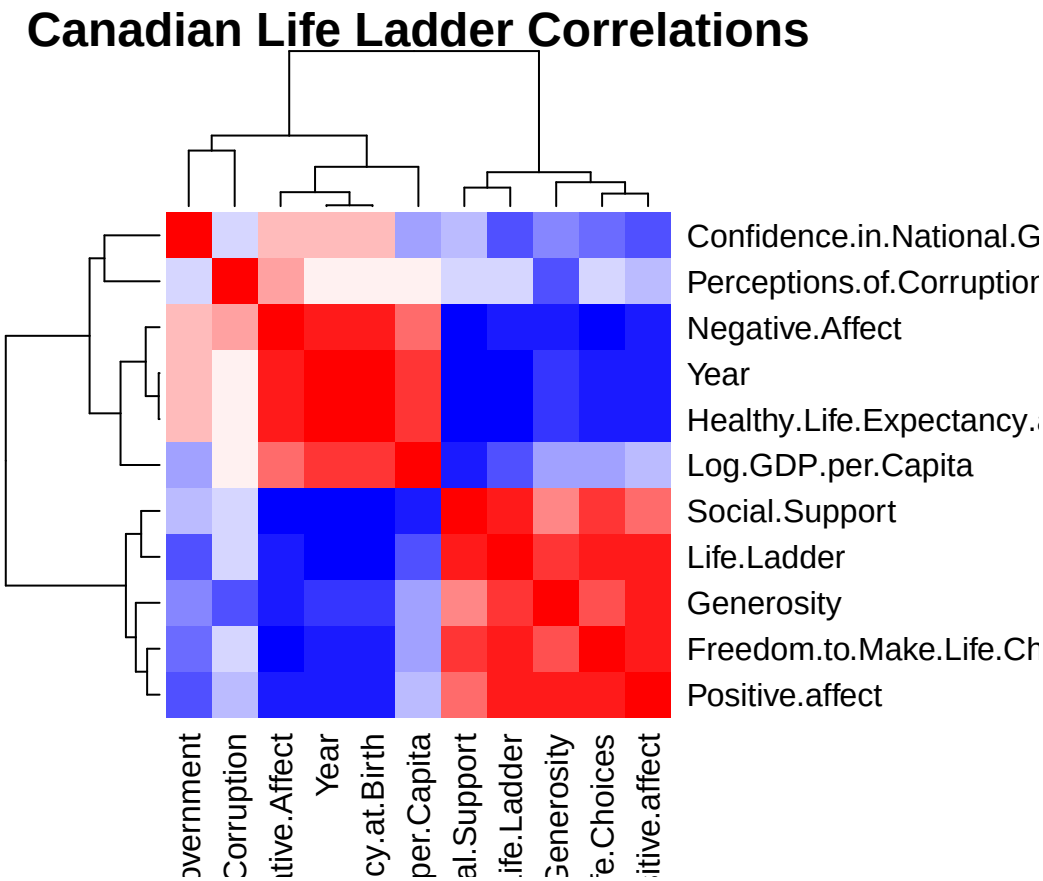
```
## Warning: Using one column matrices in ‘filter()‘ was deprecated in dplyr 1.1.0.
## i Please use one dimensional logical vectors instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```r
happiness_data_Canada = happiness_data %>%
    filter(happiness_data[1]=='Canada')


# Arrange the data in descending 'Life Ladder' scores find the top years
#Use the head function to select the top 3 years for both the US and Canada
#Do the same for the bottom 3 years for both the US and Canada

happiness_data_US_happiest_threeyears = head(happiness_data_US %>%
                                        arrange(desc(happiness_data_US[3])),3) %>%
                                        select(Year)
                                        #2007, 2008, 2013
happiness_data_US_unhappiest_threeyears = head(happiness_data_US %>%
                                        arrange(happiness_data_US[3]),3) %>%
                                        select(Year)
                                        #2016, 2015, 2018

happiness_data_Canada_happiest_threeyears = head(happiness_data_Canada %>%
                                            arrange(desc(happiness_data_Canada[3])),3) %>%
                                            select(Year)
                                        #2010, 2013, 2009
happiness_data_Canada_unhappiest_threeyears = head(happiness_data_Canada %>%
                                            arrange(happiness_data_Canada[3]),3) %>%
                                            select(Year)
                                        #2021, 2020, 2019

#Create 2 new data frames that combine the 3 happiest and 3 least happy
#data frames into one for both the US and Canada. Do this by filtering
#the data frame by country name and joining by year.  Then remove the
#column 'County Name' from the frame so that all columns are numeric
#we have only used data from a single country so that column is redundant.

happiness_data_filtered_canada= rbind(happiness_data %>%
                    filter(happiness_data[1]=='Canada') %>%
                    inner_join(happiness_data_Canada_happiest_threeyears,
                               by = "Year"),
                    happiness_data %>%
                    filter(happiness_data[1]=='Canada') %>%
                    inner_join(happiness_data_Canada_unhappiest_threeyears,
                               by = "Year")) %>%
                    select (-"Country Name")


happiness_data_filtered_us= rbind(happiness_data %>%
                    filter(happiness_data[1]=='United States') %>%
                    inner_join(happiness_data_US_happiest_threeyears, by = "Year"),
                    happiness_data %>%
                    filter(happiness_data[1]=='United States') %>%
                    inner_join(happiness_data_US_unhappiest_threeyears, by = "Year")) %>%
                    select (-"Country Name")

happiness_data_filtered_cleaned_us =
          happiness_data_filtered_us[complete.cases(happiness_data_filtered_us), ]
```

```
happiness_data_filtered_cleaned_canada =
          happiness_data_filtered_canada[complete.cases(happiness_data_filtered_canada),]


#Make a correlation data frame for the US data only.
#Make a heat map to visualize the data for better interpretation

happiness_data_filtered_cleaned_us_numeric =
  as.data.frame(lapply(happiness_data_filtered_cleaned_us, as.numeric))

correlation_matrix_us = cor(happiness_data_filtered_cleaned_us_numeric)
heatmap(correlation_matrix_us, col=colorRampPalette(c("blue", "white", "red"))(20),
        symm=TRUE, main = "U.S. Life Ladder Correlations")
```




```
#Make a correlation data for the Canada data only
#Make a heat map to visualize the data for better interpretation

happiness_data_filtered_cleaned_canada_numeric =
  as.data.frame(lapply(happiness_data_filtered_cleaned_canada, as.numeric))

correlation_matrix_canada = cor(happiness_data_filtered_cleaned_canada_numeric)
heatmap(correlation_matrix_canada, col=colorRampPalette(c("blue", "white", "red"))(20),
        symm=TRUE, main = "Canadian Life Ladder Correlations")
```



2) **How is happiness distributed by region?**

Summarize happiness by finding the average (mean), 25th percentile (quantile(x, .25)), and 75th percentile (quantile(x, .75)) by region for each year.

In an attempt to characterize how happiness is distributed globally by geographical region, several computations were performed. First, a new data frame must be created that houses the individual countries and classifies them into geographical regions. Once done, columns names were again relabeled to match those in the existing happiness data frames. The regional data and happiness data were joined on the Country Name to ensure the regions and happiness data aligned.
That data frames was sorted by to select out just the Regional Indicator, the Life Ladder global happiness score and the year. The mean, 25th percentile and 75th percentile happiness scores were calculated and displayed for each region and in each year available.

```
#Read in the Regions file
#set the column names to match those in the happiness data frames
my_regions = read_csv("Regions.csv", show_col_types = FALSE)
colnames(my_regions)= c('Country Name', 'Regional Indicator')

#left join the regions data frame and the happiness data frames
#doing as a left join will allow countries without a region assigned to
#be kept in the data frame

happiness_data_with_region = happiness_data %>%
    left_join(my_regions, by ='Country Name')
```

```r
#Pull out only the columns of interest, specifically the
#"Regional Indicator","Life Ladder", and "Year"

hd_short = happiness_data_with_region %>%
  select ("Regional Indicator","Life Ladder", "Year")

#remove null values and assign the columns to be numeric in class

hd_short <- hd_short %>% replace(.=="NULL", NA) # replace with NA
hd_short <- hd_short[!hd_short$`Life Ladder` == "null", ]
hd_short[2] = as.numeric(hd_short$`Life Ladder`)
hd_short[3] = as.numeric(hd_short$`Year`)


#Summarize the resulting data to generate a summary of the mean, 25th percentile and
#75th percentile for each region by year
summary(hd_short %>% group_by(`Regional Indicator`, `Year`))
```

```
##  Regional Indicator  Life Ladder         Year
##  Length:2089        Min.   :2.179   Min.   :2005
##  Class :character   1st Qu.:4.652   1st Qu.:2010
##  Mode  :character   Median :5.405   Median :2014
##                     Mean   :5.474   Mean   :2014
##                     3rd Qu.:6.294   3rd Qu.:2017
##                     Max.   :8.019   Max.   :2021
```

```r
df_stat = hd_short %>%
  group_by(`Regional Indicator`, `Year`)  %>%
  summarize(
    mean = mean(`Life Ladder`, na.rm = TRUE),
    qs25 = quantile(`Life Ladder`, probs = c(0.25), na.rm = TRUE),
    qs75 = quantile(`Life Ladder`, probs = c(0.75), na.rm = TRUE),
    .groups = "drop"  # Ungroup the result
    )
print(df_stat)
```

```
## # A tibble: 184 x 5
##    `Regional Indicator`        Year  mean  qs25  qs75
##    <chr>                      <dbl> <dbl> <dbl> <dbl>
##  1 Central and Eastern Europe  2005  5.28  5.12  5.39
##  2 Central and Eastern Europe  2006  5.42  5.26  5.81
##  3 Central and Eastern Europe  2007  5.10  4.75  5.39
##  4 Central and Eastern Europe  2008  5.42  5.38  5.53
##  5 Central and Eastern Europe  2009  5.24  4.89  5.49
##  6 Central and Eastern Europe  2010  5.17  4.73  5.60
##  7 Central and Eastern Europe  2011  5.23  4.94  5.57
##  8 Central and Eastern Europe  2012  5.37  5.14  5.82
##  9 Central and Eastern Europe  2013  5.30  5.07  5.82
## 10 Central and Eastern Europe  2014  5.41  5.15  5.73
## # i 174 more rows
```

3. **Team-Generated Question**

Consult with your team to formulate a question similar to those in the previous questions. Explain in words how you can use the provided data to answer that question, then write code to format the data to answer the question.

Given that we found that generosity was a key driver to happiness for both Canada and the United States, we sought to see if it correlated to happiness in the top 10 happiest countries and if its was lower for the 10 least happy country for the year 2020, as noted by life ladder scores.

The results of the two-sample t-test does not provide strong enough evidence to conclude that there is a significant difference in generosity levels between the top 10 happiest countries and the 10 unhappiest countries in 2020.

The p-value of 0.9493 is much greater than the significance level of 0.05. This suggests that we fail to reject the null hypothesis and we do not have enough evidence to conclude that the means of the two groups are significantly different.

Therefore, based on the analysis, we cannot assert that the top 10 happiest countries in 2020 were significantly more generous than the 10 unhappiest countries and we conclude that there is no strong evidence in these results to suggest that happiness directly affects the generosity level of a country.

It is also important to note that the relationship between happiness and generosity may not be straightforward and may require further investigation or a more detailed analysis to draw meaningful conclusions about the impact of happiness on generosity.

```
# Create a new data frame that filters the combined data set to only have 2020
# data and GDP and generosity variable

Happy_and_generosity = happiness_data[c("Country Name", "Year", "Life Ladder",
                                        "Generosity")] %>%
                       filter(Year == 2020)

# Data Preparation:
# Sort the data by Life Ladder in ascending order to identify the top 10 happiest
# countries and 10 unhappiest countries

sorted_LifeLadder = Happy_and_generosity[order(Happy_and_generosity$'Life Ladder'), ]

# Get the bottom 10 countries (lowest to highest)
unhappiest_2005 = sorted_LifeLadder[1:10, ]

# Get the top 10 values (lowest to highest)
happiest_2005 = sorted_LifeLadder[107:116, ]
print (happiest_2005)
```

```
## # A tibble: 10 x 4
##    `Country Name`  Year `Life Ladder` Generosity
##    <chr>          <dbl>         <dbl>      <dbl>
##  1 Austria         2020          7.21    0.00804
##  2 New Zealand     2020          7.26    0.118
##  3 Norway          2020          7.29    0.0711
##  4 Germany         2020          7.31   -0.0636
##  5 Sweden          2020          7.31    0.0882
##  6 Netherlands     2020          7.50    0.148
##  7 Switzerland     2020          7.51   -0.0705
##  8 Denmark         2020          7.51    0.0482
##  9 Iceland         2020          7.58    0.155
## 10 Finland         2020          7.89   -0.119
```

```r
# 2. Data Analysis:

# Calculate summary statistics for the generosity levels of the two groups
summary_stats_happy = summary(happiest_2005$Generosity)
summary_stats_unhappy = summary(unhappiest_2005$Generosity)

# Perform a t-test to compare the two groups
t_test = t.test(happiest_2005$Generosity, unhappiest_2005$Generosity)

# Print the results
print("Summary Statistics for Top 10 Happiest Countries:")
```

```
## [1] "Summary Statistics for Top 10 Happiest Countries:"
```

```r
print(summary_stats_happy)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.11885 -0.04570  0.05962  0.03839  0.11078  0.15464
```

```r
cat("\nSummary Statistics for 10 Unhappyiest Countries:")
```

```
##
## Summary Statistics for 10 Unhappyiest Countries:
```

```r
print(summary_stats_unhappy)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.15387 -0.09856 -0.04807  0.03378  0.06774  0.47454
```

```r
cat("\nT-test Results:")
```

```
##
## T-test Results:
```

```r
print(t_test)
```

```
##
##  Welch Two Sample t-test
##
## data:  happiest_2005$Generosity and unhappiest_2005$Generosity
## t = 0.0649, df = 12.86, p-value = 0.9493
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1490608  0.1582835
## sample estimates:
##  mean of x  mean of y
## 0.03839241 0.03378105
```