Title: “Group Assignment 1” Authors: “Shukla, Kunal(ks765)”, “Millen, Spencer(smm246)”, “Sohail, Ifrah(is172)”, “Siewny, Lauren(Les62)”, “Zheng, Peng Xi(pz68)”, “Manning, Felecia(fm150)” Date: “10/16/2022” Output: html_document: default

#Read the libraries and set working directory

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(readxl)
getwd()
```

```
## [1] "C:/Users/ifrah/Dropbox/PC/Downloads"
```

```
setwd("C:/Users/ifrah/Dropbox/PC/Downloads")
```

#Load in the CSV and Excel files

```
H2019 = read_xlsx("Happiness_2019-2021.xlsx", sheet = as.character(2019))
H2020 = read_xlsx("Happiness_2019-2021.xlsx", sheet = as.character(2020))
H2021 = read_xlsx("Happiness_2019-2021.xlsx", sheet = as.character(2021))
H2005_18 = read_csv("Happiness_2005-2018.csv")
```

```
## Rows: 1710 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): Country name
## dbl (11): year, Life Ladder, Log GDP per capita, Social support, Healthy lif...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
Regions = read_csv('Regions.csv')
```

```
## Rows: 153 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): Country name, Regional indicator
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

#Print the data frames

```
H2019
```

```
## # A tibble: 144 x 12
##    `Country name`  year `Life Ladder` `Log GDP per capita` `Social support`
##    <chr>          <dbl>         <dbl>                <dbl>            <dbl>
##  1 Afghanistan     2019          2.38                 7.63            0.420
```

```
##  2 Albania         2019          5.00                 9.52           0.686
##  3 Algeria         2019          4.74                 9.35           0.803
##  4 Argentina       2019          6.09                10.0            0.896
##  5 Armenia         2019          5.49                 9.52           0.782
##  6 Australia       2019          7.23                10.8            0.943
##  7 Austria         2019          7.20                10.9            0.964
##  8 Azerbaijan      2019          5.17                 9.58           0.887
##  9 Bahrain         2019          7.10                10.7            0.878
## 10 Bangladesh      2019          5.11                 8.47           0.673
## # i 134 more rows
## # i 7 more variables: `Healthy life expectancy at birth` <dbl>,
## #   `Freedom to make life choices` <dbl>, Generosity <dbl>,
## #   `Perceptions of corruption` <dbl>, `Positive affect` <dbl>,
## #   `Negative affect` <dbl>, `Confidence in national government` <dbl>
```

```
H2020
```

```
## # A tibble: 116 x 12
##    Country_Name  Year `Life Ladder` `Log GDP per capita` `Social support`
##    <chr>        <dbl>         <dbl>                <dbl>            <dbl>
##  1 Albania       2020          5.36                 9.49            0.710
##  2 Algeria       2020          5.44                 9.28            0.868
##  3 Argentina     2020          5.90                 9.89            0.897
##  4 Australia     2020          7.14                10.8             0.937
##  5 Austria       2020          7.21                10.9             0.925
##  6 Bahrain       2020          6.17                10.6             0.848
##  7 Bangladesh    2020          5.28                 8.49            0.739
##  8 Belgium       2020          6.84                10.8             0.904
##  9 Benin         2020          4.41                 8.11            0.507
## 10 Bolivia       2020          5.56                 8.97            0.805
## # i 106 more rows
## # i 7 more variables: `Healthy life expectancy at birth` <dbl>,
## #   `Freedom to make life choices` <dbl>, Generosity <dbl>,
## #   `Perceptions of corruption` <dbl>, `Positive affect` <dbl>,
## #   `Negative affect` <dbl>, `Confidence in national government` <dbl>
```

```
H2021
```

```
## # A tibble: 119 x 12
##    COUNTRY                YEAR LIFE_LADDER `Log GDP per capita` `Social support`
##    <chr>                 <dbl>       <dbl>                <dbl>            <dbl>
##  1 Afghanistan            2021        2.44                NA                0.454
##  2 Albania                2021        5.26                 9.56             0.702
##  3 Algeria                2021        5.22                 9.30             0.841
##  4 Argentina              2021        5.91                 9.96             0.882
##  5 Armenia                2021        5.30                 9.50             0.762
##  6 Australia              2021        7.11                10.8              0.920
##  7 Austria                2021        7.08                10.9              0.863
##  8 Benin                  2021        4.49                 8.14             0.436
##  9 Bolivia                2021        5.57                 9.01             0.798
## 10 Bosnia and Herzegovi~  2021        5.75                 9.63             0.860
## # i 109 more rows
## # i 7 more variables: `Healthy life expectancy at birth` <dbl>,
## #   `Freedom to make life choices` <dbl>, Generosity <dbl>,
## #   `Perceptions of corruption` <dbl>, `Positive affect` <dbl>,
## #   `Negative affect` <dbl>, `Confidence in national government` <dbl>
```

```
H2005_18
```

```
## # A tibble: 1,710 x 12
##    `Country name`  year `Life Ladder` `Log GDP per capita` `Social support`
##    <chr>          <dbl>         <dbl>                <dbl>            <dbl>
##  1 Afghanistan     2008          3.72                 7.30            0.451
##  2 Afghanistan     2009          4.40                 7.47            0.552
##  3 Afghanistan     2010          4.76                 7.58            0.539
##  4 Afghanistan     2011          3.83                 7.55            0.521
##  5 Afghanistan     2012          3.78                 7.64            0.521
##  6 Afghanistan     2013          3.57                 7.66            0.484
##  7 Afghanistan     2014          3.13                 7.65            0.526
##  8 Afghanistan     2015          3.98                 7.63            0.529
##  9 Afghanistan     2016          4.22                 7.63            0.559
## 10 Afghanistan     2017          2.66                 7.63            0.491
## # i 1,700 more rows
## # i 7 more variables: `Healthy life expectancy at birth` <dbl>,
## #   `Freedom to make life choices` <dbl>, Generosity <dbl>,
## #   `Perceptions of corruption` <dbl>, `Positive affect` <dbl>,
## #   `Negative affect` <dbl>, `Confidence in national government` <dbl>
```

```
Regions
```

```
## # A tibble: 153 x 2
##    `Country name` `Regional indicator`
##    <chr>          <chr>
##  1 Finland        Western Europe
##  2 Denmark        Western Europe
##  3 Switzerland    Western Europe
##  4 Iceland        Western Europe
##  5 Norway         Western Europe
##  6 Netherlands    Western Europe
##  7 Sweden         Western Europe
##  8 New Zealand    North America and ANZ
##  9 Austria        Western Europe
## 10 Luxembourg     Western Europe
## # i 143 more rows
```

#Rename the columnes

```
# Rename specific columns in H2019
colnames(H2019) = c("Country", "Year", "Life Ladder", "Log GDP per capital", "Social Support", "Healthy

# Rename specific columns in H2020
colnames(H2020) = c("Country", "Year", "Life Ladder", "Log GDP per capital", "Social Support", "Healthy

# Rename specific columns in H2021
colnames(H2021) = c("Country", "Year", "Life Ladder", "Log GDP per capital", "Social Support", "Healthy

# Rename specific columns in H2005_18
colnames(H2005_18) = c("Country", "Year", "Life Ladder", "Log GDP per capital", "Social Support", "Heal

# Rename specific columns in Regions
colnames(Regions) = c("Country", "Region")  # Replace with your new column names
```

#Load the dplyr package

```r
# Load the dplyr package
library(dplyr)
```

#combine data frames using bind_rows

```r
combined_data = bind_rows(H2005_18, H2019, H2020, H2021)
combined_data
```

```
## # A tibble: 2,089 x 12
##    Country      Year `Life Ladder` `Log GDP per capital` `Social Support`
##    <chr>       <dbl>         <dbl>                 <dbl>            <dbl>
##  1 Afghanistan  2008          3.72                  7.30            0.451
##  2 Afghanistan  2009          4.40                  7.47            0.552
##  3 Afghanistan  2010          4.76                  7.58            0.539
##  4 Afghanistan  2011          3.83                  7.55            0.521
##  5 Afghanistan  2012          3.78                  7.64            0.521
##  6 Afghanistan  2013          3.57                  7.66            0.484
##  7 Afghanistan  2014          3.13                  7.65            0.526
##  8 Afghanistan  2015          3.98                  7.63            0.529
##  9 Afghanistan  2016          4.22                  7.63            0.559
## 10 Afghanistan  2017          2.66                  7.63            0.491
## # i 2,079 more rows
## # i 7 more variables: `Healthy life expectancy at birth` <dbl>,
## #   `Freedom to make life choices` <dbl>, Generosity <dbl>,
## #   `Perceptions of corruption` <dbl>, `Positive affect` <dbl>,
## #   `Negative affect` <dbl>, `Confidence in national government` <dbl>
```

```r
combined_data = inner_join(combined_data, Regions, by = "Country")
combined_data
```

```
## # A tibble: 2,018 x 13
##    Country      Year `Life Ladder` `Log GDP per capital` `Social Support`
##    <chr>       <dbl>         <dbl>                 <dbl>            <dbl>
##  1 Afghanistan  2008          3.72                  7.30            0.451
##  2 Afghanistan  2009          4.40                  7.47            0.552
##  3 Afghanistan  2010          4.76                  7.58            0.539
##  4 Afghanistan  2011          3.83                  7.55            0.521
##  5 Afghanistan  2012          3.78                  7.64            0.521
##  6 Afghanistan  2013          3.57                  7.66            0.484
##  7 Afghanistan  2014          3.13                  7.65            0.526
##  8 Afghanistan  2015          3.98                  7.63            0.529
##  9 Afghanistan  2016          4.22                  7.63            0.559
## 10 Afghanistan  2017          2.66                  7.63            0.491
## # i 2,008 more rows
## # i 8 more variables: `Healthy life expectancy at birth` <dbl>,
## #   `Freedom to make life choices` <dbl>, Generosity <dbl>,
## #   `Perceptions of corruption` <dbl>, `Positive affect` <dbl>,
## #   `Negative affect` <dbl>, `Confidence in national government` <dbl>,
## #   Region <chr>
```

#Create a vector for Canada and US to filter through the merged data

```r
Canada_and_US = c("Canada", "United States")
combined_data_USA_and_Canada = combined_data[combined_data$Country %in% Canada_and_US, ]
```

#Filter the data for Canada and the United States

```
Canada_and_US <- c("Canada", "United States")
filtered_data <- combined_data[combined_data$Country %in% Canada_and_US, ]
```

Find the 3 happiest years for each country. Assign filtered_data frame to happiest_years and then group the data with country and year and then summarize the data by looking for average happiness through the mean of the life ladder and then sort the data using the arrange function for country and further sorting by descending order using average_happinness and then group by country again to slice or subset the data from the first three top rows that are in the descending columns.

```
happiest_years <- filtered_data %>%
  group_by(Country, Year) %>%
  summarize(Average_Happiness = mean(`Life Ladder`)) %>%
  arrange(Country, desc(Average_Happiness)) %>%
  group_by(Country) %>%
  slice(1:3)
```

```
## `summarise()` has grouped output by 'Country'. You can override using the
## `.groups` argument.
```

Find the 3 unhappiest years for each country. Assign filtered_data frame to happiest_years and then group the data with country and year and then summarize the data by looking for average happiness through the mean of the life ladder and then sort the data using the arrange function for country and average_happiness then group by country again to slice or subset the data from the first three top rows that are in the descending columns.

```
unhappiest_years <- filtered_data %>%
  group_by(Country, Year) %>%
  summarize(Average_Happiness = mean(`Life Ladder`)) %>%
  arrange(Country, Average_Happiness) %>%
  group_by(Country) %>%
  slice(1:3)
```

```
## `summarise()` has grouped output by 'Country'. You can override using the
## `.groups` argument.
```

Print the results

```
cat("Happiest Years for Canada and the United States:\n")
```

```
## Happiest Years for Canada and the United States:
```

```
print(happiest_years)
```

```
## # A tibble: 6 x 3
## # Groups:   Country [2]
##   Country        Year Average_Happiness
##   <chr>         <dbl>             <dbl>
## 1 Canada         2010              7.65
## 2 Canada         2013              7.59
## 3 Canada         2009              7.49
## 4 United States  2007              7.51
## 5 United States  2008              7.28
## 6 United States  2013              7.25
```

Print the results

```
cat("\nUnhappiest Years for Canada and the United States:\n")
```

```
##
## Unhappiest Years for Canada and the United States:
```

```
print(unhappiest_years)
```

```
## # A tibble: 6 x 3
## # Groups:   Country [2]
##   Country        Year Average_Happiness
##   <chr>         <dbl>             <dbl>
## 1 Canada         2021              6.94
## 2 Canada         2020              7.02
## 3 Canada         2019              7.11
## 4 United States  2016              6.80
## 5 United States  2015              6.86
## 6 United States  2018              6.88
```

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?

Answer: We deduced that there's a certain synchrony in the happiest years for both countries: the United States experienced the happiest years from 2007 to 2013, closely mirrored by Canada's happiest years between 2008 and 2013. However, when it comes to the unhappiest years, the two nations diverge. The U.S. faced its unhappiest years from 2015 to 2018, whereas Canada's was from 2019 to 2021. A notable observation is the overall decline in happiness over the years for both nations. The U.S. experienced a dip in pre-pandemic years, while Canada's decline coincided with the pandemic period. Despite this decline and the challenges of recent times, Canada has generally maintained a higher happiness index, although both countries show a downward trend.

Combined_data is assigned to the happiest_region variable as a new data frame and then we group by region and year. Then, we summarize by average_happiness for the mean of the life ladder and then we sort the data using the arrange function with region and in descending order, the average_happiness. Then, print happiest_regions.

```
happiest_regions <- combined_data %>%
  group_by(Region, Year) %>%
  summarize(Average_Happiness = mean(`Life Ladder`)) %>%
  arrange(Region, desc(Average_Happiness))
```

```
## `summarise()` has grouped output by 'Region'. You can override using the
## `.groups` argument.
```

```
print(happiest_regions)
```

```
## # A tibble: 167 x 3
## # Groups:   Region [10]
##    Region                       Year Average_Happiness
##    <chr>                       <dbl>             <dbl>
##  1 Central and Eastern Europe  2021              6.24
```

```
##  2 Central and Eastern Europe  2020              6.14
##  3 Central and Eastern Europe  2019              5.94
##  4 Central and Eastern Europe  2018              5.90
##  5 Central and Eastern Europe  2017              5.74
##  6 Central and Eastern Europe  2016              5.58
##  7 Central and Eastern Europe  2015              5.44
##  8 Central and Eastern Europe  2006              5.42
##  9 Central and Eastern Europe  2008              5.42
## 10 Central and Eastern Europe  2014              5.41
## # i 157 more rows
```

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.

#Answer: a dataset (combined_data) related to happiness metrics, specifically grouping the data by Year and Region and then calculating key summary statistics for each group: the mean, 25th percentile, and 75th percentile of the "Life Ladder" score, excluding missing values. The resultant summarized data, happiness_summary_by_region_year.

```r
# Summarizing happiness by region for each year
happiness_summary_by_region_year <- combined_data %>%
  group_by(`Year`, `Region`) %>%
  summarise(
    Mean_Happiness = mean(`Life Ladder`, na.rm = TRUE),
    Q25_Happiness = quantile(`Life Ladder`, 0.25, na.rm = TRUE),
    Q75_Happiness = quantile(`Life Ladder`, 0.75, na.rm = TRUE)
  )
```

```
## `summarise()` has grouped output by 'Year'. You can override using the
## `.groups` argument.
```

```r
# Viewing the summarized data
head(happiness_summary_by_region_year)
```

```
## # A tibble: 6 x 5
## # Groups:   Year [1]
##    Year Region                        Mean_Happiness Q25_Happiness Q75_Happiness
##   <dbl> <chr>                                  <dbl>         <dbl>         <dbl>
## 1  2005 Central and Eastern Europe              5.28          5.12          5.39
## 2  2005 East Asia                               6.52          6.52          6.52
## 3  2005 Latin America and Caribbean             6.80          6.61          6.90
## 4  2005 Middle East and North Africa            5.68          5.20          6.09
## 5  2005 North America and ANZ                   7.38          7.36          7.40
## 6  2005 South Asia                              5.22          5.22          5.22
```

```r
print(happiness_summary_by_region_year)
```

```
## # A tibble: 167 x 5
## # Groups:   Year [17]
##     Year Region                        Mean_Happiness Q25_Happiness Q75_Happiness
##    <dbl> <chr>                                  <dbl>         <dbl>         <dbl>
##  1  2005 Central and Eastern Europe              5.28          5.12          5.39
##  2  2005 East Asia                               6.52          6.52          6.52
##  3  2005 Latin America and Caribbean             6.80          6.61          6.90
##  4  2005 Middle East and North Africa            5.68          5.20          6.09
```

```
##  5  2005 North America and ANZ                7.38          7.36          7.40
##  6  2005 South Asia                           5.22          5.22          5.22
##  7  2005 Western Europe                       7.08          6.89          7.35
##  8  2006 Central and Eastern Europe           5.42          5.26          5.81
##  9  2006 Commonwealth of Independent~         4.83          4.63          5.17
## 10  2006 East Asia                            5.40          5.14          5.68
## # i 157 more rows
```

This code takes a dataset with information about happiness by region, groups the data by region, and calculates summary statistics (mean, Q25, and Q75) for each region. The results are stored in a new data frame and displayed for review.

```
# Overall summary of happiness by region without considering years
overall_happiness_summary_by_region <- combined_data %>%
  group_by(`Region`) %>%
  summarise(
    Overall_Mean_Happiness = mean(`Life Ladder`, na.rm = TRUE),
    Overall_Q25_Happiness = quantile(`Life Ladder`, 0.25, na.rm = TRUE),
    Overall_Q75_Happiness = quantile(`Life Ladder`, 0.75, na.rm = TRUE)
  )

print(overall_happiness_summary_by_region)
```

```
## # A tibble: 10 x 4
##    Region     Overall_Mean_Happiness Overall_Q25_Happiness Overall_Q75_Happiness
##    <chr>                       <dbl>                 <dbl>                 <dbl>
##  1 Central a~                   5.56                  5.12                  6.01
##  2 Commonwea~                   5.22                  4.73                  5.72
##  3 East Asia                    5.65                  5.28                  6.02
##  4 Latin Ame~                   5.99                  5.61                  6.47
##  5 Middle Ea~                   5.37                  4.70                  6.17
##  6 North Ame~                   7.25                  7.15                  7.38
##  7 South Asia                   4.53                  4.22                  4.99
##  8 Southeast~                   5.35                  4.88                  5.90
##  9 Sub-Sahar~                   4.30                  3.82                  4.76
## 10 Western E~                   6.83                  6.44                  7.42
```

```
# Viewing the overall summarized data
head(overall_happiness_summary_by_region)
```

```
## # A tibble: 6 x 4
##   Region      Overall_Mean_Happiness Overall_Q25_Happiness Overall_Q75_Happiness
##   <chr>                        <dbl>                 <dbl>                 <dbl>
## 1 Central an~                   5.56                  5.12                  6.01
## 2 Commonweal~                   5.22                  4.73                  5.72
## 3 East Asia                     5.65                  5.28                  6.02
## 4 Latin Amer~                   5.99                  5.61                  6.47
## 5 Middle Eas~                   5.37                  4.70                  6.17
## 6 North Amer~                   7.25                  7.15                  7.38
```

3. Team-Generated Question: What are the top 3 unhappiest regions in 2021 so that we can best direct charity/funds and donations?

Combined_data is assigned to the region_21_top3_unhappiest and then the pipe operator is used to add the filter function to filter fo the year 2021 adn then group by region and then summarize by the mean of the

life ladder. We then create a vector to rename the columns to Region and life ladder, and then we sort the columns using the arrange function and then get the top 3 rows using the head function with n = 3.

```
region_21_top3_unhappiest =
combined_data %>%
  filter(`Year`=='2021') %>% group_by(`Region`) %>% summarize(mean(`Life Ladder`))

colnames(region_21_top3_unhappiest) <- c('Region','Life_Ladder')

region_21_top3_unhappiest %>% arrange(`Life_Ladder`) %>% head(n=3)
```

```
## # A tibble: 3 x 2
##   Region                       Life_Ladder
##   <chr>                              <dbl>
## 1 South Asia                          3.84
## 2 Sub-Saharan Africa                  4.49
## 3 Middle East and North Africa        5.01
```