Human Freedom Index and Suicides Project

Part 2: Data Joining and Modeling

Doug Cady

November 13, 2021

Contents

R version 4.1.1 "Kick Things"

```r
library(readr)
library(dplyr)
library(tidyr)
library(ggplot2)
library(psych)
```

Load datasets into R

```r
hfi <- read_csv("../data/clean_hfi_2018.csv")
suicides <- read_csv("../data/clean_suicides.csv")
```

Join datasets on `year` and `country` variables

```r
# How many countries in each dataset?
hfi_countries <- unique(hfi$country)
print(length(hfi_countries))  # 162
```

```
## [1] 162
```

```r
suic_countries <- unique(suicides$country)
print(length(suic_countries))  # 101
```

```
## [1] 94
```

```r
init_joined <- inner_join(hfi, suicides, by = c("year", "country"))
init_join_countries <- unique(init_joined$country)
print(length(init_join_countries))  # 78
```

```
## [1] 78
```

```
# 23 countries weren't joined together initially, why?
nonjoin_countries <- setdiff(suic_countries, init_join_countries)
print(nonjoin_countries)
```

```
##  [1] "Antigua and Barbuda"          "Aruba"
##  [3] "Cabo Verde"                   "Cuba"
##  [5] "Czech Republic"               "Grenada"
##  [7] "Kyrgyzstan"                   "Maldives"
##  [9] "Puerto Rico"                  "Republic of Korea"
## [11] "Russian Federation"           "Saint Lucia"
## [13] "Saint Vincent and Grenadines" "Slovakia"
## [15] "Turkmenistan"                 "Uzbekistan"
```

```
# print(setdiff(hfi_countries, init_join_countries))
```

Investigate country names that don't match

Use duckduckgo and wikipedia to sort out name mismatches (if possible)

```
# Antigua and Barbuda - only in suicides (ok)
# Aruba - only in suicides (ok)
# Azerbaijan - country in both, but years don't align (ok)
# Cabo Verde - in hfi named Cape Verde, change to Cape Verde in suicides data (!fix suicides)
# Cuba - only in suicides data (ok)
# Czech Republic - hfi abbreviated to Czech Rep. (!fix hfi)
# Dominica - only in suicides, different from Dominican Republic (in hfi) (ok)
# Grenada - only in suicides (ok)
# Kiribati - only in suicides (ok)
# Kyrgyzstan - hfi lists official name "Kyrgyz Republic" (!fix hfi)
# Macau - only in suicides (ok)
# Maldives - only in suicides (ok)
# Puerto Rico - only in suicides (ok)
# Republic of Korea - hfi lists as "Korea, South" (!fix hfi)
# Russian Federation - hfi lists as "Russia" (!fix suicides)
# Saint Kitts and Nevis - only in suicides (ok)
# Saint Lucia - only in suicides (ok)
# Saint Vincent and Grenadines - only in suicides (ok)
# San Marino - only in suicides (ok)
# Slovakia - hfi lists as "Slovak Rep." (!fix hfi)
# Sri Lanka - country in both, but years don't align (ok)
# Turkmenistan - only in suicides (ok)
# Uzbekistan - only in suicides (ok)

## Check existence of country and years available in each dataset:
# HFI
print(unique(filter(hfi, str_detect(country, "India"))$country))
```

```
## [1] "India"
```

```
print(unique(filter(hfi, str_detect(country, "India"))$year))
```

```
## [1] 2016 2015 2014 2013 2012 2011 2010 2009 2008
```

```
# Suicides
print(unique(filter(suicides, str_detect(country, "India"))$country))
```

```
## character(0)
```

```
print(unique(filter(suicides, str_detect(country, "India"))$year))
```

```
## numeric(0)
```

Clean up country names that didn't match initially

```
hfi_clean_cntry <- hfi %>%
    mutate(country = str_replace(country, "Czech Rep.", "Czech Republic"),
           country = str_replace(country, "Kyrgyz Republic", "Kyrgyzstan"),
           country = str_replace(country, "Korea, South", "Republic of Korea"),
           country = str_replace(country, "Slovak Rep.", "Slovakia"),)


suic_clean_cntry <- suicides %>%
    mutate(country = str_replace(country, "Cabo Verde", "Cape Verde"),
           country = str_replace(country, "Russian Federation", "Russia"))

joined <- inner_join(hfi_clean_cntry, suic_clean_cntry, by = c("year", "country"))

print(length(unique(joined$country))) # 78 init, 84 final
```

```
## [1] 84
```

```
# 2016 doesn't seem to have many records, maybe it should be excluded from this analysis?
# Yes, it has far fewer (< 20%) records than other years have
joined %>%
    group_by(year) %>%
    count()
```

```
## # A tibble: 9 x 2
## # Groups:   year [9]
##    year     n
##   <dbl> <int>
## 1  2008   912
## 2  2009   960
## 3  2010   948
## 4  2011   924
## 5  2012   876
## 6  2013   864
## 7  2014   840
## 8  2015   672
## 9  2016   150
```

```
joined_no2016 <- filter(joined, year != 2016)
print(paste(nrow(joined_no2016), ncol(joined_no2016)))
```

```
## [1] "6996 17"
```

```
write_csv(joined_no2016, "../output/clean_hfi_suicides_joined_no2016.csv")
```

Explore correlations between freedom, suicides, and GDP

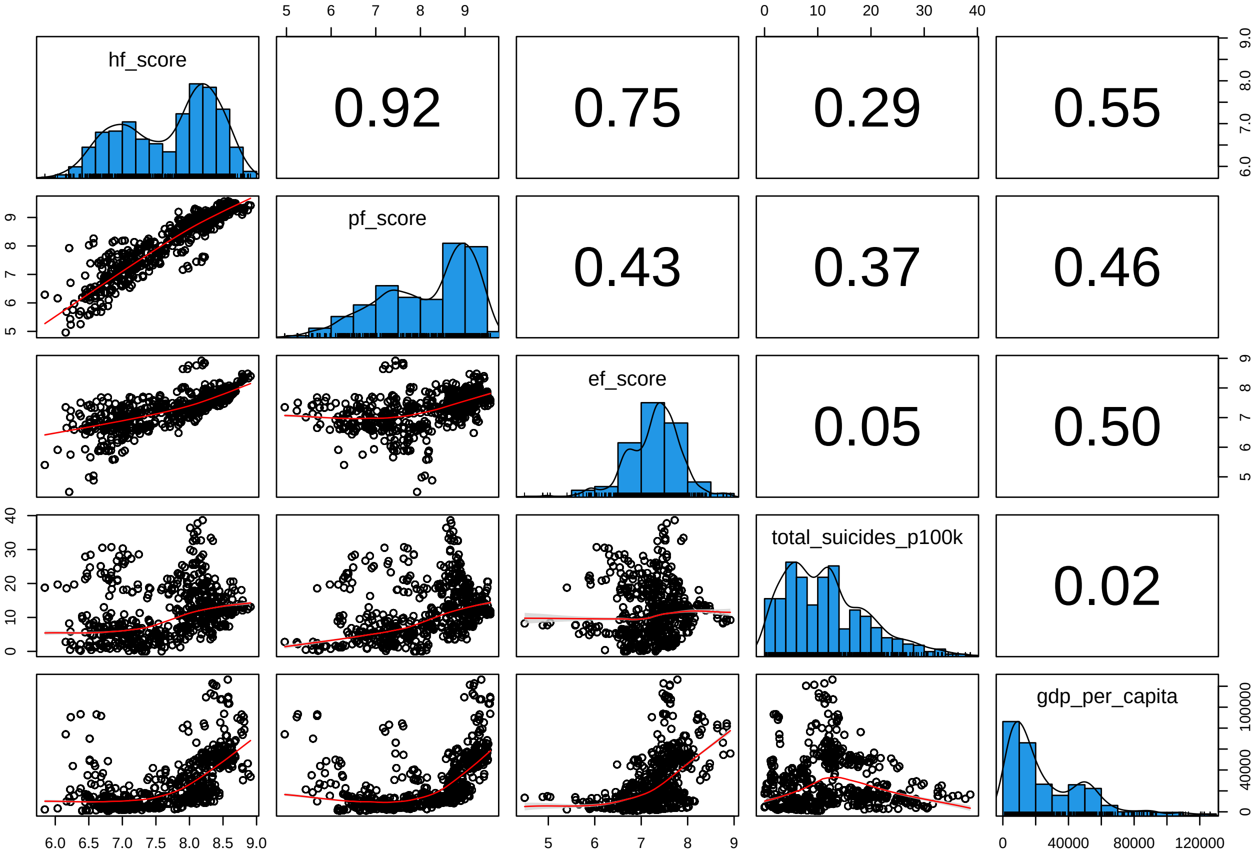
Overall, personal, and economic freedom scores, number of suicides per 100k, and GDP per capita

```
# Suicides are currently broken down by age and sex, we need to aggregate to total counts
corr_data <- joined_no2016 %>%
    group_by(year, country) %>%
    mutate(total_suicides = sum(suicides_no),
           total_pop = sum(population),
           total_suicides_p100k = total_suicides / total_pop * 100000) %>%
    ungroup() %>%
    select(hf_score, pf_score, ef_score, total_suicides_p100k, gdp_per_capita)

summary(corr_data)
```

```
##     hf_score        pf_score        ef_score     total_suicides_p100k
##  Min.   :5.84   Min.   :4.96   Min.   :4.49   Min.   : 0.00
##  1st Qu.:7.08   1st Qu.:7.35   1st Qu.:6.91   1st Qu.: 5.43
##  Median :7.90   Median :8.48   Median :7.34   Median :10.79
##  Mean   :7.70   Mean   :8.13   Mean   :7.27   Mean   :11.63
##  3rd Qu.:8.26   3rd Qu.:9.00   3rd Qu.:7.65   3rd Qu.:16.46
##  Max.   :8.91   Max.   :9.57   Max.   :8.92   Max.   :38.66
##  NA's   :180    NA's   :180    NA's   :180
##  gdp_per_capita
##  Min.   :   977
##  1st Qu.:  8347
##  Median : 17052
##  Mean   : 27493
##  3rd Qu.: 44747
##  Max.   :126352
##
```

```
# Correlation pairs plot with histograms, density, and scatter plots
pairs.panels(corr_data,
             smooth = TRUE,
             scale = FALSE,
             density = TRUE,
             ellipses = FALSE,
             method = "pearson",
             pch = 21,
             lm = FALSE,
             cor = TRUE,
             jiggle = FALSE,
             hist.col = 4,
             stars = FALSE,
             ci = TRUE)
```



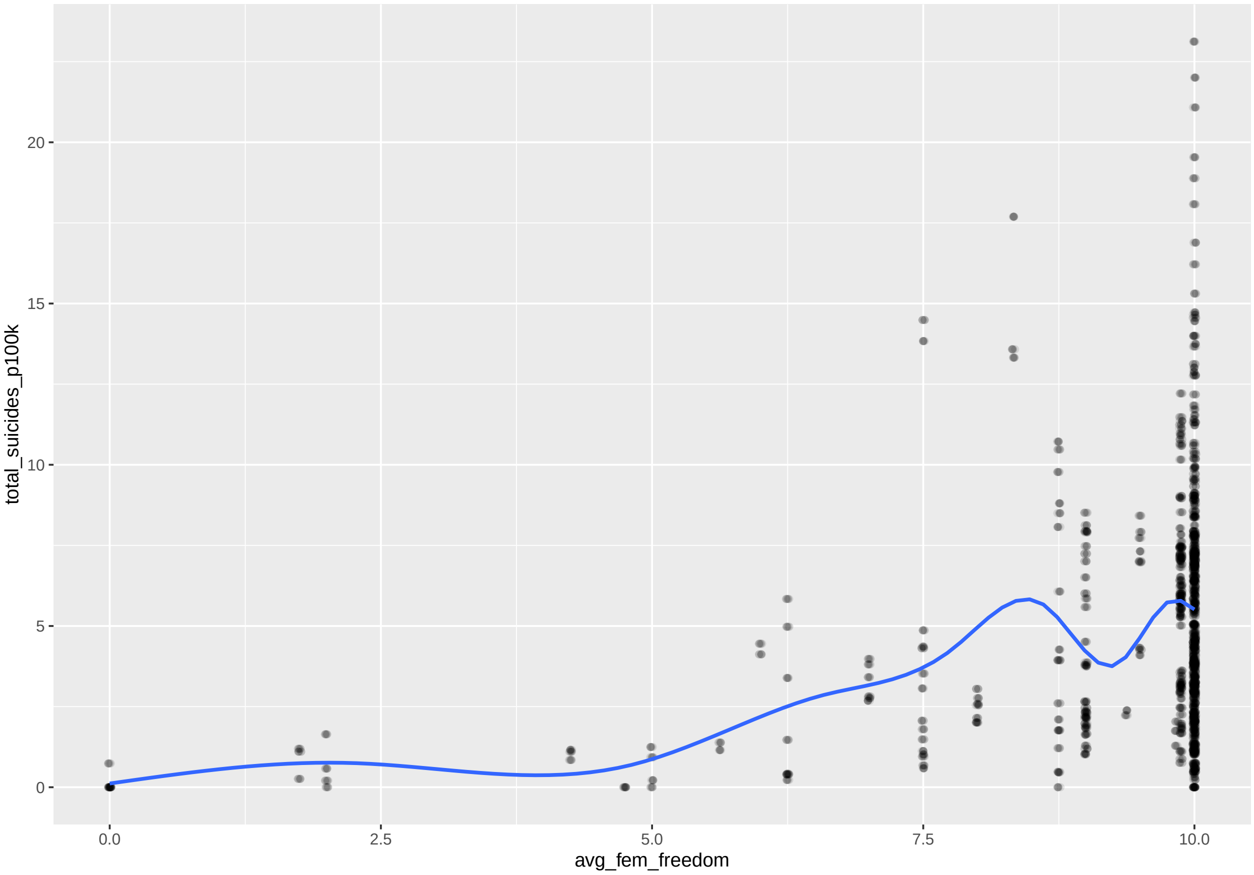
Overall human freedom score seems to be most highly correlated with GDP per capita, though this is not too surprising. Suicide rates do seem to have a weak correlation to personal and human freedom, but not economic freedom or gdp per capita. This makes me wonder if they would be more correlated to a country's happiness rather than it's freedom.

Do female freedom scores relate to female suicide rates?

```
fem_freedom_suic <- joined_no2016 %>%
    filter(sex == 'female') %>%
    group_by(year, country) %>%
    mutate(total_suicides = sum(suicides_no),
           total_pop = sum(population),
           total_suicides_p100k = total_suicides / total_pop * 100000) %>%
    ungroup() %>%
    select(year, country, pf_ss_women_fgm, pf_ss_women_inheritance, pf_movement_women,
           pf_identity_sex_female, pf_identity_divorce, total_suicides_p100k)

fem_freedom_suic$avg_fem_freedom <- rowMeans(select(fem_freedom_suic,
                                                    pf_ss_women_fgm, pf_ss_women_inheritance,
                                                    pf_movement_women, pf_identity_sex_female,
                                                    pf_identity_divorce), na.rm = TRUE)

ggplot(fem_freedom_suic, aes(x = avg_fem_freedom, y = total_suicides_p100k)) +
    geom_jitter(alpha = 0.1) +
    geom_smooth(se = FALSE)
```



```
# Hypothesis: Countries with high female freedom scores will have less than average female suicide rate
sorted_fem_free_scores <- fem_freedom_suic %>%
    filter(year == 2015) %>%
    unique() %>%
    select(avg_fem_freedom, total_suicides_p100k) %>%
    arrange(desc(avg_fem_freedom))

print(nrow(sorted_fem_free_scores)) # 56
```

```
## [1] 56
```

```
hi_fem_free <- sorted_fem_free_scores[1:28, ]
lo_fem_free <- sorted_fem_free_scores[29:56, ]
# Perform a Welch's t-test to determine if two populations are signifcantly different
# p-value of 0.01, significant at 98% confidence level
# Reject null hypothesis and conclude that there may exist a relationship between higher
# BUT if we look back at the scatter plot of suicide rates vs freedom scores, more than half
# (almost 65%) of the countries have a freedom score of 10. So which countries should be classified
# as lo vs hi freedom if there are 10 scores in both groups??
t.test(hi_fem_free$total_suicides_p100k, lo_fem_free$total_suicides_p100k)
```

```
##
##  Welch Two Sample t-test
##
## data:  hi_fem_free$total_suicides_p100k and lo_fem_free$total_suicides_p100k
## t = 2.7, df = 49, p-value = 0.01
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.5406 3.7646
## sample estimates:
## mean of x mean of y
##     6.326     4.173
```

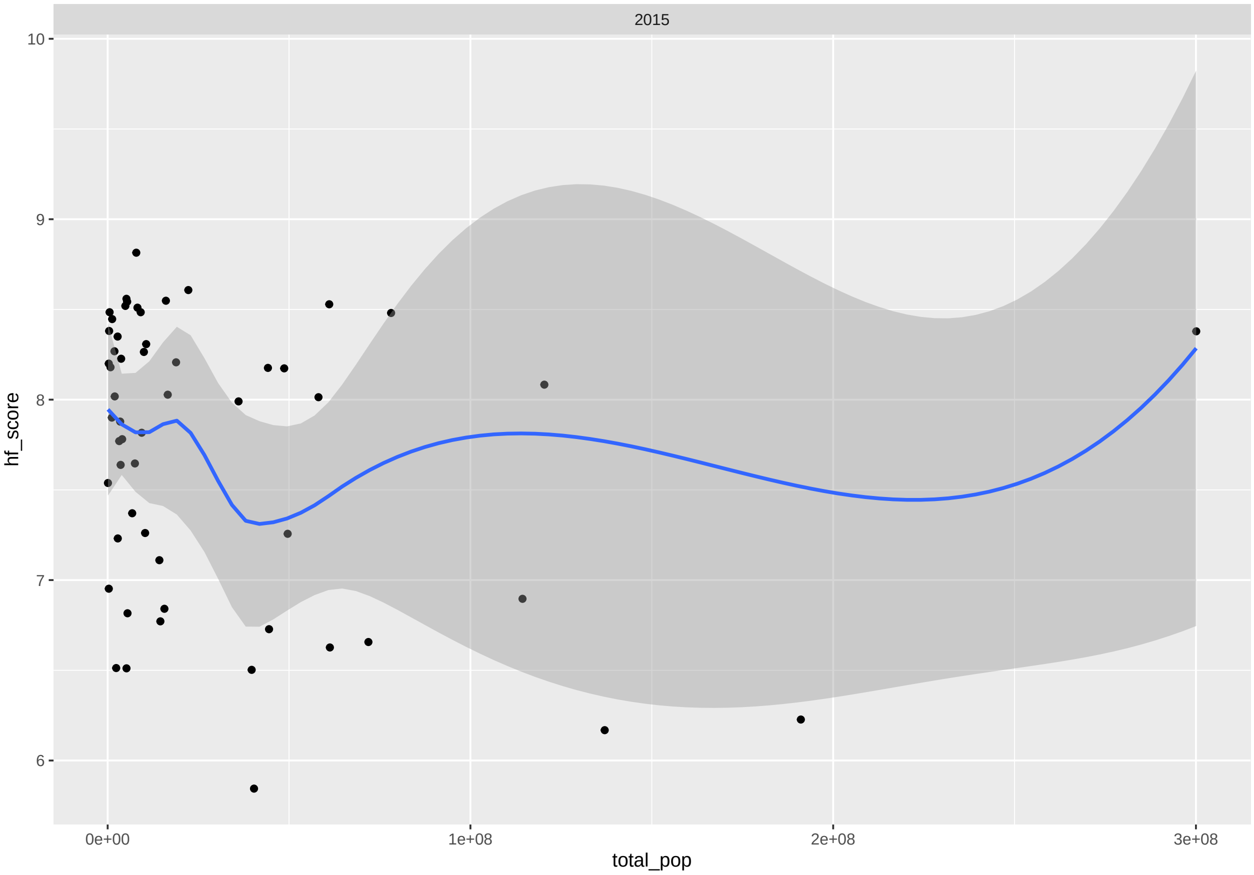
The relationship between female freedom in a country and female suicide rates is non-existent to weak in strength. Many countries have top freedom scores and have a wide range of suicide rates per one hundred thousand persons.

Does a country's population size impact freedom?

There exists only a weak correlation between population size and human freedom for a country

```
joined_2015 <- filter(joined, year == 2015) %>%
    group_by(year, country) %>%
    mutate(total_suicides = sum(suicides_no),
           total_pop = sum(population),
           total_suicides_p100k = total_suicides / total_pop * 100000) %>%
    ungroup() %>%
    select(year, country, total_pop, hf_score) %>%
    unique() %>%
    arrange(total_pop)

ggplot(joined_2015, aes(x = total_pop, y = hf_score)) +
    geom_point() +
    geom_smooth() +
    facet_wrap( ~ year, nrow = 2)
```



```
# Hypothesis: Countries with smaller populations will have more freedom
total_rows <- nrow(joined_2015)  # 56
row_split <- total_rows / 2  # 28
small_pop <- joined_2015[1:28, ]
large_pop <- joined_2015[29:56, ]

# Perform a Welch's t-test to determine if two populations are signifcantly different
# p-value of 0.06, significant at 90% confidence level, but I am using 95% confidence
# Fail to reject null hypothesis at 95% confidence level, though it's close
t.test(small_pop$hf_score, large_pop$hf_score)
```

```
##
##  Welch Two Sample t-test
##
## data:  small_pop$hf_score and large_pop$hf_score
## t = 1.9, df = 50, p-value = 0.06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01253  0.80089
## sample estimates:
## mean of x mean of y
##     7.912     7.518
```

What are important factors that contribute to human freedom?

Factors that can positively impact human freedom include the free movement of women, women's inheritance rights, and divorce rights. A country's population seems to have a negative relationship with human freedom, as we saw a bit earlier.

```r
no_countries <- joined_no2016 %>%
    group_by(year, country) %>%
    mutate(total_suicides = sum(suicides_no),
           total_pop = sum(population),
           total_suicides_p100k = total_suicides / total_pop * 100000) %>%
    ungroup() %>%
    select(pf_ss_women_inheritance, pf_movement_women, pf_identity_sex_female,
           pf_identity_divorce, hf_score, total_pop, gdp_per_capita) %>%
    unique()

lm1 <- lm(hf_score ~ ., data = no_countries)
summary(lm1)
```

```
##
## Call:
## lm(formula = hf_score ~ ., data = no_countries)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6352 -0.2135  0.0786  0.3422  0.9207
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              5.62e+00   2.08e-01   26.99  < 2e-16 ***
## pf_ss_women_inheritance  3.96e-02   1.84e-02    2.15    0.033 *
## pf_movement_women        1.48e-01   2.17e-02    6.79  1.6e-10 ***
## pf_identity_sex_female  -3.58e-02   2.35e-02   -1.52    0.129
## pf_identity_divorce      4.49e-02   1.89e-02    2.37    0.019 *
## total_pop               -3.36e-09   6.89e-10   -4.88  2.4e-06 ***
## gdp_per_capita           1.47e-05   1.30e-06   11.27  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.465 on 175 degrees of freedom
##   (401 observations deleted due to missingness)
## Multiple R-squared:  0.624,  Adjusted R-squared:  0.611
## F-statistic: 48.5 on 6 and 175 DF,  p-value: <2e-16
```