Happiness and Democracy Visualizations

Giovanni Exume

Github: <https://github.com/giovanniexume>

Email: [gexume1@gmail.com](mailto:gexume1@gmail.com)

Table of Contents

[Importing Data 3](#_Toc122705415)

[Data Cleaning and Manipulation 4](#_Toc122705416)

[Joining Happiness and Democracy Index 4](#_Toc122705417)

[Adding in Continent Variable 5](#_Toc122705418)

[Data Visuals 5](#_Toc122705419)

[Distrubtions and Average Democracy Scores of Each Continent 5](#_Toc122705420)

[Happiness Metrics for Top 5 in Democracy 6](#_Toc122705421)

[Life Ladder Scores for the Top 10 and Bottom 10 in Democracy Index 7](#_Toc122705422)

[Correlations 8](#_Toc122705423)

[Checking for Normality 8](#_Toc122705424)

[ANOVA Differences in Positive Affect 10](#_Toc122705425)

Hello! This project uses datasets from the World Happiness Report, for 2020, using happiness indicators from the GallUp World Poll, and the Democracy Index for 2020 from The Economist, ranking democratic practices and institutions. As these datasets were provided to me through my classwork, I do not have links to the original sources. Listed below are the variables for both datasets

**World Happiness Variables:**

**country:** country name

**year:** year of report

**life\_ladder:** happiness score, or subjective well-being. Computed as national average response to the question of life evaluations

**log\_gdp\_per\_capita:** GDP in purchasing power parity (PPP) at constant 2011 international dollar prices, logged for normality.

**social\_support:** Computed as the national average of the binary responses (either 0 or 1) to the GWP question “If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?”

**healthy\_life\_expectancy\_at\_birth:** Healthy life expectancies at birth are based on the data extracted from the World Health Organization’s (WHO) Global Health Observatory data repository

**freedom\_to\_make\_life\_choices:** Computed as the national average of responses to the GWP question “Are you satisfied or dissatisfied with your freedom to choose what you do with your life?”

**generosity:** The residual of regressing national average of response to the GWP question “Have you donated money to a charity in the past month?” on GDP per capita.

**perceptions\_of\_corruption**: The measure is the national average of the survey responses to two questions in the GWP: “Is corruption widespread throughout the government or not” and “Is corruption widespread within businesses or not?” The overall perception is just the average of the two 0-or-1 responses.

**positive\_affect**: The average of three positive affect measures in GWP: happiness, laugh and enjoyment in the Gallup World Poll waves 3-7.

**negative\_affect**: The average of three negative affect measures in GWP: worry, sadness and anger.

**Democracy Index for 2020 Variables.**

**country:** country name

**overall\_score:** The average score of the 5 measures

**rank**: country’s overall rank on democracy index

**pluralism:** a measure of pluralism

**govfunc:** a measure of how well or badly the government functions

**poliparc:** a measure of political participation

**policul:** a measure of political culture

**civlib:** a measure of civil liberty

setwd("/Users/giovanniexume/Documents/Data and Text Mining")  
  
  
library(tidyverse)

library(ggplot2)  
library(broom)  
library(janitor)

library(SKTools)  
library(foreign)  
library(readxl)  
library(DataExplorer)  
library(maps)

library(plotly)

library(countrycode)  
library(scales)

library(ggpubr)

# Importing Data

happiness <- read\_xlsx("worldhappiness.xlsx") %>% filter(year == 2020)  
  
democracy <- read\_csv("democracy\_index\_2020.csv")

head(happiness)

## # A tibble: 6 × 11  
## `Country name` year `Life Ladder` `Log GDP per capita` `Social support`  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 Albania 2020 5.36 9.50 0.710  
## 2 Argentina 2020 5.90 9.85 0.897  
## 3 Australia 2020 7.14 10.8 0.937  
## 4 Austria 2020 7.21 10.9 0.925  
## 5 Bahrain 2020 6.17 10.6 0.848  
## 6 Bangladesh 2020 5.28 8.47 0.739  
## # … with 6 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>

head(democracy)

## # A tibble: 6 × 8  
## country `Overall score` rank pluralism govfunc poliparc policul civlib  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Norway 9.81 1 10 9.64 10 10 9.41  
## 2 Iceland 9.37 2 10 8.57 8.89 10 9.41  
## 3 Sweden 9.26 3 9.58 9.29 8.33 10 9.12  
## 4 New Zealand 9.25 4 10 8.93 8.89 8.75 9.71  
## 5 Canada 9.24 5 9.58 8.93 8.89 9.38 9.41  
## 6 Finland 9.2 6 10 8.93 8.89 8.75 9.41

dim(happiness)

## [1] 95 11

dim(democracy)

## [1] 167 8

The happiness data set, has 95 rows, 11 variables. The democracy dataset has 167 rows and 8 variables.

# Data Cleaning and Manipulation

## Joining Happiness and Democracy Index

happiness <- happiness %>% rename(country = "Country name")  
  
nations<-inner\_join(democracy, happiness, by= "country") %>%   
 rename(overalldemscore = "Overall score", demrank = "rank") %>% select(-year)  
  
nations <- nations %>% clean\_names()  
  
introduce(nations)

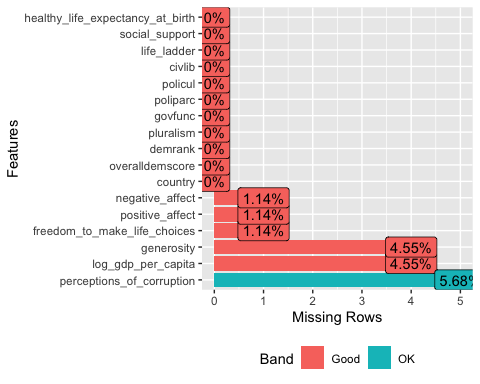
## # A tibble: 1 × 9  
## rows columns discrete\_columns continuous\_columns all\_missing\_columns  
## <int> <int> <int> <int> <int>  
## 1 88 17 1 16 0  
## # … with 4 more variables: total\_missing\_values <int>, complete\_rows <int>,  
## # total\_observations <int>, memory\_usage <dbl>

After joining, there are 88 countries left in the dataset and 17 feature variables. The output below shows some of the countries that have been excluded, as they were unavailable in the happiness dataset

anti\_join(democracy, happiness, by= "country")

## # A tibble: 79 × 8  
## country `Overall score` rank pluralism govfunc poliparc policul civlib  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Taiwan 8.94 11 10 9.64 7.22 8.13 9.71  
## 2 Luxembourg 8.68 13 10 8.57 6.67 8.75 9.41  
## 3 Costa Rica 8.16 18 9.58 6.79 7.22 7.5 9.71  
## 4 Cabo Verde 7.65 32 9.17 7 6.67 6.88 8.53  
## 5 Botswana 7.62 33 9.17 6.79 6.11 7.5 8.53  
## 6 Malaysia 7.19 39 9.58 7.86 6.67 6.25 5.59  
## 7 Panama 7.18 40 9.58 6.43 7.22 5 7.65  
## 8 Trinidad and… 7.16 41 9.58 7.14 6.11 5.63 7.35  
## 9 Jamaica 7.13 42 8.75 7.14 5 6.25 8.53  
## 10 Leste 7.06 44 9.58 5.93 5.56 6.88 7.35  
## # … with 69 more rows

plot\_missing(nations)



Most of the data is still in a good place as it comes to missing values.

## Adding in Continent Variable

This data doesn’t include a continent variable, but we can add that using the countrycode function

nations$continent <- countrycode(sourcevar = nations$country, origin = "country.name", destination = "region")

# Data Visuals

# Distrubtions and Average Democracy Scores of Each Continent

cont\_dem <- nations %>%  
 group\_by(continent) %>%  
 summarise(mean\_dem = mean(overalldemscore)) %>%  
 arrange(desc(mean\_dem))  
  
nations %>% ggplot(aes(x = overalldemscore, fill = continent)) +  
 geom\_histogram(show.legend = F) +  
 geom\_vline(cont\_dem, mapping = aes(xintercept = mean\_dem), color = "black", linewidth = 1) +  
 facet\_wrap(continent~.) +  
 scale\_fill\_discrete(type = RColorBrewer::brewer.pal(7, "Set1")) +  
 theme\_minimal()+  
 theme(legend.position = "bottom")+  
 labs(x = "Overall Democracy Score", y = "Count", title = "Distribution & Average Democracy Scores of Each Continent")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

Chart, waterfall chart

Description automatically generated

This visual shows that most of the countries in the dataset are based in Europe and Central Asia, with the fewest in North American and South America.

# Happiness Metrics for Top 5 in Democracy

top5dem <- nations %>%   
 head(n = 5) %>%  
 select(country, generosity, social\_support, freedom\_to\_make\_life\_choices, positive\_affect, negative\_affect, perceptions\_of\_corruption) %>%  
 rename(country = country,  
 "Generosity" = generosity,  
 "Social Support" = social\_support,  
 "Freedom to Make Life Choices" = freedom\_to\_make\_life\_choices,  
 "Positive Affect" = positive\_affect,  
 "Negative Affect" = negative\_affect,  
 "Perceptions of Corruption" = perceptions\_of\_corruption) %>%  
 gather(Happiness\_Metric, Score, Generosity:"Perceptions of Corruption")

top5dem %>%  
 ggplot(aes(x = Score, y = Happiness\_Metric, fill = country))+  
 geom\_col(show.legend = FALSE)+  
 facet\_wrap(country~.) +  
 scale\_fill\_discrete(type = c("#eee279", "#fac682", "#c48a3f", "#8a4f7d", "#7b0828")) +  
 theme\_minimal() +  
 labs(x = "Standardized Scores", y = "Happiness Metrics", title = "Happiness Metrics for Top 5 in Democracy")

Chart, timeline

Description automatically generated

# Life Ladder Scores for the Top 10 and Bottom 10 in Democracy Index

topbottom <- rbind(nations %>% head(n = 10), nations %>% tail(n = 10)) %>%   
 mutate(Place = ifelse(demrank <= 10, "Top", "Bottom"))

topbottom %>%  
 ggplot(aes(x = reorder(country, desc(demrank)), y= life\_ladder, fill = Place)) +  
 geom\_col(position = "dodge") +  
 geom\_text(label = topbottom$demrank, nudge\_y = 0.25) +  
 coord\_flip() +  
 theme\_minimal()+  
 labs(x = "Life Ladder", y = "Country", title = "Life Ladder Scores for the Top 10 and Bottom 10 in Democracy")+  
 scale\_fill\_manual(values = c("Bottom" = "indianred",  
 "Top" = "chartreuse4"))

Chart, bar chart

Description automatically generated

# Correlations

Let’s see what the relationship is between civil liberty scores and perceptions of corruption

## Checking for Normality

shapiro.test(nations$civlib)

##   
## Shapiro-Wilk normality test  
##   
## data: nations$civlib  
## W = 0.90993, p-value = 1.421e-05

shapiro.test(nations$perceptions\_of\_corruption)

##   
## Shapiro-Wilk normality test  
##   
## data: nations$perceptions\_of\_corruption  
## W = 0.87378, p-value = 7.659e-07

Both tests are significant, so each variable is not normally distributed and a Spearman’s correlation will be appropriate

cor.test(nations$civlib, nations$perceptions\_of\_corruption, method = "spearman")

## Warning in cor.test.default(nations$civlib, nations$perceptions\_of\_corruption, :  
## Cannot compute exact p-value with ties

##   
## Spearman's rank correlation rho  
##   
## data: nations$civlib and nations$perceptions\_of\_corruption  
## S = 135432, p-value = 7.282e-05  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## -0.4213546

There is a moderate negative relationship between civil liberty and perceptions of corruption

nations %>% ggplot(aes(x= civlib, y = perceptions\_of\_corruption, label = country)) +  
 geom\_text()+  
 geom\_point(alpha = 0.45) +  
 geom\_smooth(method = "loess", linewidth = 0.5) +  
 theme\_minimal() +  
 labs(x = "Civil Liberty", y = "Perceptions of Corruption", subtitle = "Spearman's Correlation = -0.42, p < 0.001")

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.

## `geom\_smooth()` using formula = 'y ~ x'

## Warning: Removed 5 rows containing non-finite values (`stat\_smooth()`).

Graphical user interface

Description automatically generated with medium confidence

This visual shows that as civil liberty increases in a country, the average perceptions of corruptions decrease, with a dropoff after 7.5 in civil liberty scores.

# ANOVA Differences in Positive Affect

Are there any meaningful differences in positive affect between the Middle East/North Africa, Europe/Central Asia, and Latin America/Caribbean?

positiveaffect <- nations %>% filter(continent == c("Middle East & North Africa", "Europe & Central Asia", "Latin America & Caribbean"))

anova(aov(positive\_affect~continent, data = positiveaffect))

## Analysis of Variance Table  
##   
## Response: positive\_affect  
## Df Sum Sq Mean Sq F value Pr(>F)   
## continent 2 0.070575 0.035288 9.411 0.001444 \*\*  
## Residuals 19 0.071243 0.003750   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anovaresults <- TukeyHSD(aov(positive\_affect~continent, data = positiveaffect))  
  
anovaresults

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = positive\_affect ~ continent, data = positiveaffect)  
##   
## $continent  
## diff lwr  
## Latin America & Caribbean-Europe & Central Asia 0.07173706 -0.009308906  
## Middle East & North Africa-Europe & Central Asia -0.12226656 -0.221236523  
## Middle East & North Africa-Latin America & Caribbean -0.19400361 -0.307610061  
## upr p adj  
## Latin America & Caribbean-Europe & Central Asia 0.15278302 0.0883979  
## Middle East & North Africa-Europe & Central Asia -0.02329659 0.0142780  
## Middle East & North Africa-Latin America & Caribbean -0.08039717 0.0009857

The output shows there are statistically significant differences in positive affect between the selected continents, with Europe/Central Asia and Latin America/Caribbean both having higher positive affect scores than the Middle East/North Africa, p < 0.05

positiveaffect %>% ggplot(aes(x = continent, y = positive\_affect, fill = continent))+  
 geom\_boxplot(show.legend = F)+  
 theme\_minimal()+  
 scale\_fill\_manual(values = c("indianred", "steelblue", "khaki"))+  
 labs(x = "Continent", y = "Positive Affect", title = "Differences in Positive Affect by Continent")

## Warning: Removed 1 rows containing non-finite values (`stat\_boxplot()`).

Chart, box and whisker chart

Description automatically generated

Github: <https://github.com/giovanniexume>

Email: [gexume1@gmail.com](mailto:gexume1@gmail.com)