This is Democracy: What would the Spotify playlist of a democracy look like?

Nadia Chan

2023-12-18

```r
knitr::include_graphics('ai_playlistcover.png')
```


```r
# out.width resizes image to 50% of the text width
# fig.align aligns the image to the center
```

```r
# libraries ---------------------------------------------------------------

library(tidyverse)
library(ggplot2)
library(stargazer)
library(readxl)
library(countrycode)
library(gridExtra)
library(lfe)


# importing data ----------------------------------------------------------

dat = read.csv('merged_data.csv') # cleaned and merged dataset
```

Introduction

Traditionally, academic assessments of democracy have relied on metrics such as "per capita GDP", "primary schooling", and the "gap between male and female primary attainment", among others, following the framework established by Barro (1999). However, in the midst of the technological revolution we are currently witnessing — recognizing the critical role that tools like the internet and artificial intelligence (AI) play in shaping communication, information access, and societal interactions — is becoming increasingly crucial. Even within this paper, the power of technology is demonstrated, by the AI-generated hypothetical playlist cover image above, courtesy of ChatGPT (OpenAI, 2023).

This shift calls for a reevaluation and adaptation of the metrics used to measure democracy, towards a quantification that more accurately reflects our dynamically evolving world.

My project aims to apply this perspective by investigating a relatively unexplored area: a potential link between how democratic a country is and the kind of music its people prefer. Diverse musical preferences in a society may point towards a cultural openness and tolerance to different stances, which by definition, are associated with the "tolerance" and "freedom of thought and expression" characteristics of a democracy (Massari, n.d.). **I therefore initially hypothesize that countries with higher levels of democracy likely exhibit a broader diversity in the genres of songs their citizens listen to.**

Literature Review

Volksten (2014)'s study delves into the social importance of music, tracing back to ancient times. He references timeless stories in which music played a symbolic role; like bringing down the walls of Jericho, and Orpheus' singing mesmerizing both predators (wolves) and prey (lamb) alike. Volksten emphasizes the continued relevance of music as a "rich and varied" form of cultural expression, in the 21st century. He poses that the expressive nature of music can inspire listeners to broaden their perspectives and appreciate different "sets of values".

In their 2023 publication, Adlington and Zubillaga investigate the phenomenon termed as the "third-wave of democratization", focusing on pivotal historical events that defined 'democratization', such as the Carnation Revolution in Portugal in 1974, the disintegration of the Soviet Bloc in the 90s, and the early days of the German Democratic Republic and South Korea. The authors examine how musical practices in these different places helped people imagine what democracy could look like.

In their 2023 study, Quevedo-Redondo and Navarro-Sierra conducted an in-depth analysis of Spotify data, focusing on playlists curated by Spain's major political parties during times of election campaigns. Their investigation aimed to determine whether these politicians used music strategically to influence voter decisions. Their findings revealed *significant* differences in the song choices between the different political parties.

Contradictorily, Gingerich (2023) stands as one of the few scholars to challenge the general consensus that artificial intelligence platforms promote cultural democracy. He argues that such AI-driven platforms — Spotify being a prime example — might actually diminish cultural democracy by limiting users' "direct and spontaneous engagement with a wide variety of cultural content".

Feeling very drawn to this concept of utilizing Spotify data in research, as seen in previous studies; I adapted the approach for my own project. However, the aforementioned literature primarily offer evidence of music impacting democracy level, whereas my focus is on the reverse of this relationship - how the level of democracy of a country might influence its citizens' music preferences - of which I found limited existing research on.

For example, Knowles (2012) conducted a survey in which various *music preferences* of a sample of US residents that self-reported as either "Democrat-aligned" or "Republican-aligned" were analyzed. It was concluded that the Republicans tended to have less diverse music tastes compared to the Democrats. However, nowhere in the article was the process of how the music data was collected to be utilized in the study, explained.

Research Methodology

Independent variable

- Data for different countries' levels of democracy were compiled from the Economist Intelligence Unit (EIU)'s "Democracy Index" scores. These scores range from 0 (least democratic) to 10 (most).

Dependent variable

- Music diversity will be quantified in terms of the total number of 'popular' genres that countries, as a whole, listen to in a year. This number was extrapolated from the data set titled *"Weekly Top 200 songs on Spotify by country from 2021 to 2022"*, that was uploaded on "Kaggle", an online data science community platform, by user "Yelexa".
- The unit of analysis for this study of panel-data is therefore country-years in which the Spotify data was recorded. The highest possible number of genres that one observation can take assumes that each one of this country's 200 most listened-to tracks was of a different genre, for every week in that year = (200 genres per week) * (52 weeks per year) = 10,400 genres listened-to by a country in a specific year.

Control variable

- A wealthy country may be associated with both a higher level of democracy (Acemoğlu et al., 2017) *and* easier access to music streaming services, and consequently a more diverse music scene. To prevent the confounding effect of this on the main relationship that this study aims to investigate, countries' wealth, measure in terms of Gross Domestic Product per capita (obtained from the World Bank Development Indicators) will be accounted for in preceding analysis.

Clusters

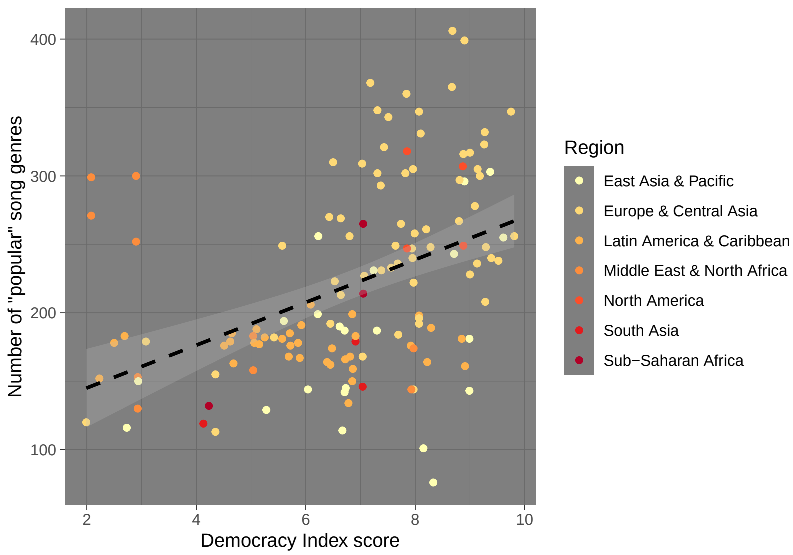
I have also chosen to cluster countries into 7 regions, as defined by the World Bank Indicator data, which comes included in R's 'countrycode' package. This will absorb factors that affect observations but are invariant with time.

Countries within the same region, may also tend to share political values, violating the assumption that each observation has to be Independent and Identically Distributed (IID), required for a linear regression analysis. Groupings of neighboring countries such as the EU and the USSR in the past, adopt similar political regimes, and are good examples of why their standard errors should be clustered at the region-level.

General Distribution

```r
dat %>% ggplot() +
  aes(x=dem_index, y=num_genres, color=as.factor(region)) +
  geom_point() +
  scale_color_manual(
    values = RColorBrewer::brewer.pal(7, 'YlOrRd'),
    name = 'Region') +
  geom_smooth(formula=y~x, method='lm',
              fill='grey', color='black', alpha=.25,
              linetype = 2) +
```

```
  xlab('Democracy Index score') +
  ylab('Number of "popular" song genres') +
  #ggtitle("") +
  theme_dark()
```

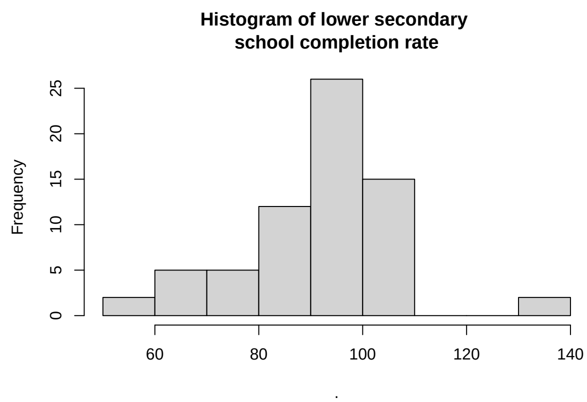


It appears from the scatter plot that there may be a positive relationship between democracy index scores and music diversity. However, it also shows us that we are indeed looking at clustered data, and so we cannot simply infer a relationship from this plot alone. A more finical representation of this is attached as Appendix 1.1.

Interaction term

I additionally want to investigate whether the potential positive relationship above is dependent on a country's education level. The intuition behind this is that in order for one to pick up on implicit political themes expressed in music, they must have gone through a certain threshold of education. I decided on using the median global school completion rate for that year, to act as this benchmark — the reason for this being represented visually through the histogram below. The left-tailed skew implies that the *majority* of values may actually be lower than what the mean value would be. The median would be more representative of the number of graduates in *most* countries, since the mean is more susceptible to outliers.

```
dat$sec_comp %>% hist(main="Histogram of lower secondary \nschool completion rate")
```




```r
# dichotomization ---------------------------------------------------------

dat = dat %>%
  group_by(year) %>%
  mutate(
    sec_comp = case_when(
      sec_comp > median(sec_comp, na.rm = TRUE) ~ 1,
      sec_comp <= median(sec_comp, na.rm = TRUE) ~ 0,
      is.na(sec_comp) ~ 0
    )
  )
```

Hypothesis and Model

The hypothesis I want to test claims that countries with higher democratic index scores, are associated, on average, with listening to a more diverse variety of song genres, given that its secondary school completion rate is above the global average.

$$\begin{aligned} NumGenres_{it} =& \alpha + \beta_1 DemIndex_{it} + \beta_2 GDPPC_{it} + \beta_3 SecCompRate_{it} + \\ & \beta_4 (DemIndex_{it} * SecCompRate_{it}) + \delta\tilde{X} + \varepsilon_{it} \end{aligned}$$

Findings

Summary Statistics

```r
dat %>% as.data.frame() %>% select(num_genres, dem_index, gdppc, sec_comp) %>%
  stargazer(type = 'latex',
            covariate.labels =
              c('Number of "popular" song genres',
                'Democracy Index score',
                'GDP per capita',
                'Lower secondary completion rate'
              ),
```

```r
          notes = c("GDPPC measured in current US Dollars",
                    "Completion rate measured as a percentage of relevant age group"),
          title = 'Summary Statistics',
          header = F)
```

Table 1: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
Number of "popular" song genres	139	220.935	69.435	76	406
Democracy Index score	139	6.854	1.929	1.990	9.810
GDP per capita	136	30,237.720	28,660.740	1,596.663	133,590.100
Lower secondary completion rate	139	0.237	0.427	0	1

GDPPC measured in current US Dollars
Completion rate measured as a percentage of relevant age group

The summary statistic table above tells us that even though the hypothetical highest number of genres a country can ‘achieve’ in a year is around 10 thousand genres, the minimum and maximum number of observations that appear in this data set are 76 and 406, respectively. There is neither a completely democratic nation (in which case, would take a DI score of 10), nor a completely autocratic (the opposite extreme) nation in the data set, and by extension, in the world; although there are countries in the data set that come close to this. The average DI score of all countries throughout both years captured in this data set is 6, which leans towards the democratic end of the extremes.

Regression Table

```r
## lm models ----
# simple cause-effect linear regression
model1 = lm(num_genres ~ dem_index,
            data = dat)
# interaction
model2 = lm(num_genres ~ dem_index + sec_comp + dem_index*sec_comp,
            dat)
# controlled interaction
model3 = lm(num_genres ~ dem_index + gdppc + sec_comp + dem_index*sec_comp,
            dat)

## felm models ----
# region fixed effects
model4 = felm(num_genres ~ dem_index + gdppc + sec_comp + dem_index*sec_comp | region,
              dat)
# with region clustered standard errors
model5 = felm(num_genres ~ dem_index + gdppc + sec_comp + dem_index*sec_comp | region | 0 | region,
              dat)
```

```r
stargazer(model1, model2, model3, model4, model5,
          type = 'latex',
          title = 'Core results',
          dep.var.labels = c('Number of song genres'),
          covariate.labels = c('Democracy Index score',
```

```
                              # 'GDP per capita',
                              'Lower secondary completion rate',
                              'Genres * Completion rate'),
          omit.stat = c('rsq', 'ser', 'f'),
          omit = c('gdppc'),
          style = 'ajps',
          model.numbers = F,
          #model.names = F,
          add.lines = list(
            c('Controlled for confounders', 'No', 'No', 'Yes', 'Yes', 'Yes'),
            c('Region Fixed Effects', 'No', 'No', 'No', 'Yes', 'Yes'),
            c('Region-level clustered S.E.', 'No', 'No', 'No', 'No', 'Yes')
          ),
          header = F)
```

Table 2: Core results

	Number of song genres				
		OLS		**felm**	
Democracy Index score	15.627***	9.133***	1.549	−0.163	−0.163
	(2.770)	(2.783)	(3.036)	(3.204)	(6.286)
Lower secondary completion rate		−168.714***	−139.695***	−120.447***	−120.447**
		(47.547)	(43.705)	(42.851)	(35.278)
Genres * Completion rate		29.346***	24.071***	21.173***	21.173***
		(6.456)	(5.966)	(5.757)	(3.502)
Constant	113.821***	147.896***	171.711***		
	(19.717)	(19.526)	(19.029)		
Controlled for confounders	No	No	Yes	Yes	Yes
Region Fixed Effects	No	No	No	Yes	Yes
Region-level clustered S.E.	No	No	No	No	Yes
N	139	139	136	136	136
Adj. R-squared	0.183	0.337	0.455	0.554	0.554

***p < .01; **p < .05; *p < .1

The basic linear regression model predicts that a one point increase in democratic index score, is correlated with listening to around 15 new song genres. This positive relationship is initially statistically significant, but does not remain so as the confounding variable is accounted for by the model.

The coefficient from the fixed effect regression model, on its own tells us that in a country with a below average secondary school completion rate, a one point increase in the DI scale would have very little impact on music diversity (a 0.163 less number of genres, if you want to be precise). In a country that inhabits more secondary graduates however, the coefficient becomes (-0.163+21.173 = +)21.01. I.e. a point increase on the DI scale predicts the country, on average, to now listen to 21 more genres of songs than before.

Visual representation

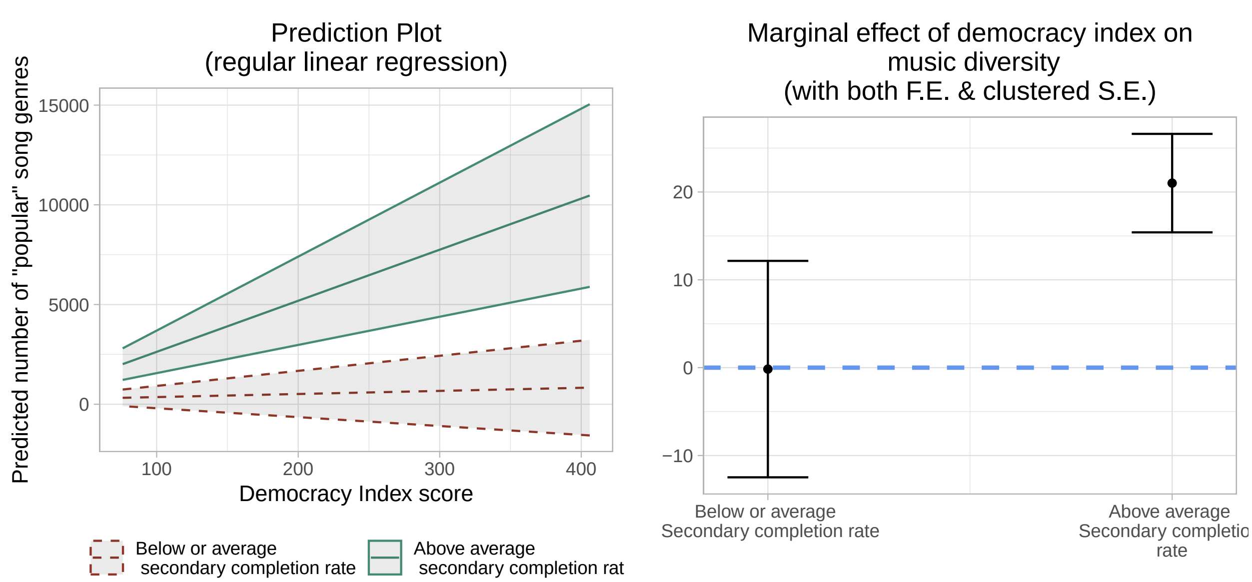
```
N = 50 # makes 50 predictions for each condition (i.e. 50 for young, 50 for not young)

predictions = data.frame(
```

```r
  dem_index =
    c(seq(min(dat$num_genres), max(dat$num_genres), length.out=N),
      seq(min(dat$num_genres), max(dat$num_genres), length.out=N)),

  sec_comp = c(rep(0,N), rep(1,N)),

  gdppc = mean(dat$gdppc, na.rm=T)

)

fits = predict(model3,
               newdata = predictions,
               interval = "confidence")

predictions = cbind(predictions, fits)

pred_plot = predictions %>% ggplot() +
  aes(x=dem_index, y=fit,
      ymin = lwr, ymax = upr,
      color = factor(sec_comp),
      linetype = factor(sec_comp)
  ) +
  geom_line() +
  geom_ribbon(alpha = 0.1) +
  scale_color_manual(name = "",
                     values = c("tomato4",
                                "aquamarine4"),
                     labels = c('Below or average \n secondary completion rate',
                                'Above average \n secondary completion rate')) +
  scale_linetype_manual(name = "",
                        values = c(2, 1),
                        labels = c('Below or average \n secondary completion rate',
                                   'Above average \n secondary completion rate')) +
  theme_light() +
  xlab("Democracy Index score") +
  ylab('Predicted number of "popular" song genres') +
  theme_light() +
  ggtitle("Prediction Plot\n(regular linear regression)") +
  theme(plot.title = element_text(hjust = 0.5),
        legend.position='bottom', legend.direction='horizontal')
```

```r
# simple vector containing binary indicators for conditional value
sec_comp = c(0, 1)

# b1 = coefficient of effect DI
# b4 = coefficient of effect (DI given SC)

# ME of DI on MD = b1 + (b4*SC)
effect = model5$coefficients[1] + (sec_comp*model5$coefficients[4])

# SE of DI on MV = sqrt( var(b1) + (SC)^2var(b4) + (2)(SC)*cov(b1,b4) )
margin_se = sqrt(
  vcov(model5)[1,1] +
```

```
    (sec_comp^2) * (vcov(model5)[4,4]) +
    2 * (sec_comp*vcov(model5)[1,4])
)

me_plot = cbind(sec_comp, effect, margin_se) %>%
  as.data.frame() %>%
  mutate(
    CI_L = effect - (1.96*margin_se),
    CI_H = effect + (1.96*margin_se)
  ) %>%
  ggplot() +
  aes(x=sec_comp, y=effect,
      ymin=CI_L, ymax=CI_H) +
  geom_point() +
  geom_errorbar(width = 0.2) +
  scale_x_continuous(breaks = c(0,1),
                     labels = c('Below or average \n Secondary completion rate',
                                'Above average \nSecondary completion \nrate')) +
  ggtitle('Marginal effect of democracy index on\n music diversity\n(with both F.E. & clustered S.E.)')
  xlab("") +
  ylab("") +
  geom_hline(yintercept=0, linetype=2, color='cornflowerblue', linewidth=1) +
  theme_light() +
  theme(plot.title = element_text(hjust = 0.5))
```

```
grid.arrange(
  pred_plot, me_plot,
  nrow = 1
)
```



Conclusion

The visual aid of the plots aid the findings that we came to previously. The prediction plot alone cannot inform about the statistical significance of these relations, but the marginal effect plot on the right does, and it concurs with the results we found before. The slope of our linear fit for *above average secondary education level countries* is upward sloping, hence positive; while the slope of our linear fit for *below average secondary education* is from just looking at the plot perhaps very minimally positive, but for the most part, close enough to zero, that it is not noticeable to the naked eye without further calculations.

In *less secondary-level educated countries*, the estimated marginal effect of democracy on music diversity is very small, and not 'sufficiently far' enough away from the null, for the null hypothesis to be rejected. In *more secondary-level educated countries* however, the marginal effect is estimated to be positive **and** statistically significant at conventional levels.

All of the findings have been in support of our hypothesis, however, the research design still has many flaws that have to be acknowledged, in order to assess if the evidence found is strong enough to infer causality. Generally, a causal relationship can only be **establised** by means of a randomized experiment in which two identical groups undergo the study, and the only difference between them is the treatment (or control) condition. Since the data utilized in this study was obtained by observational means, although it did yield substantially and statistically significant results, causality is a very bold statement to make, and one that requires a higher level of justification than currently available.

Avenues for future improvement

The biggest limitation of this study is its lack of data to work with, limiting the resulting strength of evidence that the findings provided in testing the hypothesis. The Spotify statistics encompassed only data for the years 2021 and 2022. Additionally, the reliability of these observations may also be problematic, given that they were recorded in the time period directly following the Covid-19 crisis; and may not reflect people's *true* listening habits. In a follow-up study, the use of data extending across a longer time period, is insisted.

Relying solely on Spotify data to represent a country's music preferences can be misleading. Not only does Spotify lack an extensive database of non-English songs, but some tracks are also not universally available in all regions of the world (Quora, n.d.). This implies that in countries where the *most popular* songs are predominantly non-English, Spotify data may not provide an accurate representation of national music preferences. Taylor's 2017 study, analyzing the popularity of country music artists in various US states using Spotify data, encountered a similar issue. To address this, the research included a comparison of the most-listened-to artists on Spotify with the proportion of YouTube searches in the country that comprised these same artists. Discrepancy between these two statistics would be larger in countries whose music preferences are not reliably represented by Spotify data.

'Genre', as defined by the Spotify data set, reflects **artist** genre and not song genre. In the case of versatile artists who release songs of various genres, this may affect the accuracy of the 'true' genre of the **song**. Within the scope of this study, an assumption is made that an artist only releases tracks of the same genre; which tends to be true in *many* cases. After all, there is only one (unfortunately) Taylor Swift in the world.

Another weakness associated with how 'genre' is defined in the data, is that the categorization of genres is not standardized across countries. For instance, the same songs may be split into different categories of "Pop" and "Latin pop" in say, Argentina; but lumped into a single category "Pop" in say, China, where the "Latin pop" category is assumed to not be applicable and therefore doesn't exist.

Finally, a weakness associated with my research design is to do with the decision to use music diversity as the determinant of democracy level. Countries like the UAE, where the number of expats (who are not the ones making political decisions in the country) exceeds that of UAE citizens, would present greater music diversity. This would incorrectly imply a higher democracy index score, when the UAE is not in actuality a democracy in nature. An alternative democratic indicator that could instead be considered to represent the dependent variable in a future similar study, is the human rights index.

Chicago-style References

Barro, Robert J. "Determinants of Democracy." *Journal of Political Economy* 107, no. S6 (1999). https://doi.org/10.1086/250107.

ChatGPT, response to "generate a 'this is:' Spotify playlist cover image that says 'this is democracy' " December 16, 2023, *OpenAI*, https://chat.openai.com/chat.

Massari, Maurizio. "Tolerance Is a Key Ingredient of Democracy." *Organization for Security and Co-operation in Europe*. https://www.osce.org/files/f/documents/2/8/13570.pdf.

Volgsten, Ulrik. "Music, Culture, Politics - Communicating Identity, Authenticity and Quality in the 21st Century." *Nordisk kulturpolitisk tidsskrift* 17, no. 1 (2014): 114–31. https://doi.org/10.18261/issn2000-8325-2014-01-07.

Adlington, Robert, and Igor Contreras Zubillaga. "Introduction to the Special Issue on Music and Democratic Transition." *Twentieth-Century Music* 20, no. 1 (2023): 2–5. doi:10.1017/S1478572222000445.

Quevedo-Redondo, Raquel, Marta Rebolledo, and Nuria Navarro-Sierra. "Music as Soft Power: The Electoral Use of Spotify." *Media and Communication* 11, no. 2 (2023). https://doi.org/10.17645/mac.v11i2.6344.

Gingerich, Jonathan. "Is Spotify Bad for Democracy? Artificial Intelligence, Cultural Democracy, and Law." *Yale Journal of Law and Technology* 24 (2022): 227–316.

Knowles, Jamillah. "The Echo Nest Data Reveals Political Leanings through Music." *TNW | Insider*, July 12, 2012. https://thenextweb.com/news/republicans-have-less-diverse-music-taste-than-democrats-how-music-can-predict-our-political-leanings.

Daron Acemoğlu, Suresh Naidu, James Robinson, Pascual Restrepo. "Democracy and Economic Growth: New Evidence." *VoxDev*, December 15, 2017. https://voxdev.org/topic/institutions-political-economy/democracy-and-economic-growth-new-evidence.

"Spotify: Why Don't You Have All the Songs in All Countries?" *Quora*. https://www.quora.com/Spotify-Why-dont-you-have-all-the-songs-in-all-countries. Taylor, Carson. "Mapping Country Music's Politics." *Medium*, September 23, 2017. https://medium.com/swlh/mapping-country-musics-politics-8270b19b63cc.

Data Sources

"Democracy Index 2021: The China Challenge." *Economist Intelligence Unit*, September 28, 2023. https://www.eiu.com/n/campaigns/democracy-index-2021/.

"Democracy Index 2022." *Economist Intelligence Unit*, September 28, 2023. https://www.eiu.com/n/campaigns/democracy-index-2021/.

Yelexa. "Spotify Weekly Top 200 Songs Streaming Data." *Kaggle*, July 25, 2022. https://www.kaggle.com/datasets/yelexa/spotify200.

"GDP per Capita (Current US Dollars)." *World Bank Open Data*. https://data.worldbank.org/indicator/NY.GDP.PCAP.CD.

"Lower secondary completion rate, total (% of relevant age group)." *World Bank Open Data*. https://data.worldbank.org/indicator/SE.SEC.CMPT.LO.ZS

Appendix 1.1

Region-isolated scatterplots

```
EastAsia = dat %>% subset(region == 'East Asia & Pacific') %>%
  ggplot() +
  geom_point() +
  aes(x=dem_index, y=num_genres) +
  geom_smooth(formula=y~x, method='lm',
              fill='grey', color='maroon', alpha=.25,
              linetype = 2) +
  xlab('DI score') +
  ylab('Number of genres') +
  xlim(1, 10) +
  ylim(-200, 600) +
  theme_minimal() +
  ggtitle('East Asia & Pacific') +
  theme(plot.title = element_text(hjust = 0.5))

Europe = dat %>% subset(region == 'Europe & Central Asia') %>%
  ggplot() +
  aes(x=dem_index, y=num_genres) +
  geom_point() +
  geom_smooth(formula=y~x, method='lm',
              fill='grey', color='maroon', alpha=.25,
              linetype = 2) +
  xlab('DI score') +
  ylab('Number of genres') +
  xlim(1, 10) +
  ylim(-200, 600) +
  theme_minimal() +
  ggtitle('Europe & Central Asia') +
  theme(plot.title = element_text(hjust = 0.5))

LatinAmerica = dat %>% subset(region == 'Latin America & Caribbean') %>%
  ggplot() +
  aes(x=dem_index, y=num_genres) +
  geom_point() +
  geom_smooth(formula=y~x, method='lm',
              fill='grey', color='gold3', alpha=.25,
              linetype = 2) +
  xlab('DI score') +
  ylab('Number of genres') +
  xlim(1, 10) +
  ylim(-200, 600) +
  theme_minimal() +
  ggtitle('Latin America & Caribbean') +
  theme(plot.title = element_text(hjust = 0.5))

MiddleEast = dat %>% subset(region == 'Middle East & North Africa') %>%
  ggplot() +
  aes(x=dem_index, y=num_genres) +
```

```
  geom_point() +
  geom_smooth(formula=y~x, method='lm',
              fill='grey', color='gold3', alpha=.25,
              linetype = 2) +
  xlab('DI score') +
  ylab('Number of genres') +
  xlim(1, 10) +
  ylim(-200, 600) +
  theme_minimal() +
  ggtitle('Middle East & North Africa') +
  theme(plot.title = element_text(hjust = 0.5))

NorthAmerica = dat %>% subset(region == 'North America') %>%
  ggplot() +
  aes(x=dem_index, y=num_genres) +
  geom_point() +
  geom_smooth(formula=y~x, method='lm',
              fill='grey', color='gold3', alpha=.25,
              linetype = 2) +
  xlab('DI score') +
  ylab('Number of genres') +
  xlim(1, 10) +
  ylim(-200, 600) +
  theme_minimal() +
  ggtitle('North America') +
  theme(plot.title = element_text(hjust = 0.5))

SouthAsia = dat %>% subset(region == 'South Asia') %>%
  ggplot() +
  aes(x=dem_index, y=num_genres) +
  geom_point() +
  geom_smooth(formula=y~x, method='lm',
              fill='grey', color='maroon', alpha=.25,
              linetype = 2) +
  xlab('DI score') +
  ylab('Number of genres') +
  xlim(1, 10) +
  ylim(-200, 600) +
  theme_minimal() +
  ggtitle('South Asia') +
  theme(plot.title = element_text(hjust = 0.5))

SubSaharanAfrica = dat %>% subset(region == 'Sub-Saharan Africa') %>%
  ggplot() +
  aes(x=dem_index, y=num_genres) +
  geom_point() +
  geom_smooth(formula=y~x, method='lm',
              fill='grey', color='maroon', alpha=.25,
              linetype = 2) +
  xlab('DI score') +
  ylab('Number of genres') +
  xlim(1, 10) +
  ylim(-200, 600) +
```

```
  theme_minimal() +
  ggtitle('Sub-Saharan Africa') +
  theme(plot.title = element_text(hjust = 0.5))
```

```
grid.arrange(EastAsia, Europe, MiddleEast, LatinAmerica,
             NorthAmerica, SouthAsia, SubSaharanAfrica,
             nrow = 4)
```