WIP DRAFT - Tidy Tuesday Freedom Dataset

M. Zhang

2022-02-23

Contents

Introduction

We will be working with the `freedom.csv` dataset via the Tidy Tuesday repository. This analysis will not be styled as a "final report", but rather as a quick walk through of some data analysis and wrangling that took place while exploring the data.

The MSDSO Discord group for the University of Texas Masters in Data Science Online program will be doing weekly explorations of TidyTuesday as an exercise for improving their data science skill sets in a collaborative environment.

Data Details

This dataset is pulled from the Tidy Tuesday Repository:

Thomas Mock (2022). Tidy Tuesday: A weekly data project aimed at the R ecosystem. https://github.com /rfordatascience/tidytuesday.

The original data is from Freedom House and the United Nationsvia Arthur Cheib.

Freedom House is a nonpartisan organization focused on producing research and reports on themes and trends related to democracy, political rights, and civil liberties. This data set `freedom` contains information in regards to various country's Civil Liberty `CL` and Political Rights `PR` index scores, as well as their Least Developed Country `LDC` indicator.

Data Loading

```
freedom <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2
```

Data Dictionary

The following tables contain information in regards to the columns available.

`freedom.csv`

variable	class	description
country	character	Country Name
year	double	Year
CL	double	Civil Liberties
PR	double	Political rights
Status	character	Status (Free `F`, Not Free `NF`, Partially Free `PF`)
Region_Code	double	UN Region code
Region_Name	character	UN Region Name
is_ldc	double	Is a least developed country (binary 0/1)

The definition for "Least Developed Country" is pulled from the United Nations. A country qualifies for LDC if it meets the criteria for Income, Human Assets, and Economic and Environmental Vulnerability. An important requirement for inclusion is that the country must agree to the classification to be added to the list.

The Civil Liberties and Political Rights score is generated by FreedomHouse, using a methodology inspired by the Universal Declaration of Human Rights which was adopted by the UN General Assembly in 1948, The Civil Liberties score is a combination of 15 separate indicators, and Political Rights score is a combination of 10 separate indicators. Each of these indicators scale from 0 to 4, with 4 representing the greatest amount of freedom. These scores are summarized into indexes for Civil Liberties `CL` and Political Rights `PR` on a scale from 1 to 7, with 1 representing the greatest freedom.

`Status` buckets the combined `CL` and `PR` scores into 3 general categories: Free, Partially Free, and Not Free.

Additional details are available here.

Data Exploration and wrangling

Examining the raw data

We will initially take a precursor inspection of the data, utilizing `summary` for numerical information and `table` for categorical information.

```
glimpse(freedom)
```

```
Rows: 4,979
Columns: 8
$ country     <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanistan"~
$ year        <dbl> 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004~
$ CL          <dbl> 7, 7, 7, 7, 7, 7, 7, 6, 6, 6, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6~
$ PR          <dbl> 7, 7, 7, 7, 7, 7, 7, 6, 6, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6~
$ Status      <chr> "NF", "NF", "NF", "NF", "NF", "NF", "NF", "NF", "NF", "NF"~
$ Region_Code <dbl> 142, 142, 142, 142, 142, 142, 142, 142, 142, 142, 142, 142~
$ Region_Name <chr> "Asia", "Asia", "Asia", "Asia", "Asia", "Asia", "Asia", "A~
$ is_ldc      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
```

```
names(freedom)
```

```
[1] "country"     "year"        "CL"          "PR"          "Status"
[6] "Region_Code" "Region_Name" "is_ldc"
```

```
freedom %>%
  select(CL, PR) %>%
  summary(freedom)
```

```
       CL              PR
 Min.   :1.000   Min.   :1.000
 1st Qu.:2.000   1st Qu.:1.000
 Median :3.000   Median :3.000
 Mean   :3.369   Mean   :3.411
 3rd Qu.:5.000   3rd Qu.:6.000
 Max.   :7.000   Max.   :7.000
```

```
freedom %>%
  count(country) %>%
  arrange(n)
```

```
# A tibble: 193 x 2
   country                 n
   <chr>               <int>
 1 South Sudan            10
 2 Montenegro             15
 3 Serbia                 18
 4 Timor-Leste            22
 5 Afghanistan            26
 6 Albania                26
 7 Algeria                26
 8 Andorra                26
 9 Angola                 26
10 Antigua and Barbuda    26
# ... with 183 more rows
```

```
freedom %>% select(year) %>% table(useNA = "ifany")
```

```
.
1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010
 189  189  189  189  190  190  190  190  191  191  191  192  192  192  192  192
2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
 193  193  193  193  193  193  193  193  193  193
```

```
freedom %>% select(Status) %>% table(useNA = "ifany")
```

```
.
   F   NF   PF
2219 1257 1503
```

```
freedom %>% select(Region_Code) %>% table(useNA = "ifany")
```

```
.
   2    9   19  142  150
1388  364  910 1218 1099
```

```
freedom %>% select(Region_Name) %>% table(useNA = "ifany")
```

.

```
  Africa Americas     Asia   Europe  Oceania
    1388      910     1218     1099      364
```

```
freedom %>% select(is_ldc) %>% table(useNA = "ifany")
```

.

```
   0    1
3803 1176
```

We can see that it appears that not all countries have full data for all 26 years. We will filter and examine a specific country to see if the missing years are sequential:

```
freedom %>%
  filter(country == "South Sudan") %>%
  select(year) %>% pull()
```

```
 [1] 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
```

A precursory glance suggest that not all countries have the full year range of data. There are no unexpected unique values from our initial look at the data. Additionally, the `PR` and `CL` scores are within the expected ranges given the definitions provided by FreedomHouse.

Checking for NA values

```
freedom %>%
  summarize(across(everything(), ~ sum(is.na(.)))) %>%
  tidyr::pivot_longer(everything()) %>%
  arrange(desc(value)) %>%
  deframe()
```

```
    country        year          CL          PR      Status Region_Code
          0           0           0           0           0           0
Region_Name      is_ldc
          0           0
```

NA counting methodology was taken from stackexchange

It appears that there are no NA values in this data set. This is reaffirming our findings from when we checked the summary and tables for the dataset earlier.

Recoding

Going forward, we will be using the `Region_Name` column in lieu of `Region_Code` for simplicity. Additionally, we will recode `is_ldc` into categorical values.

```
freedom = freedom %>%
  mutate(ldc = if_else(is_ldc == 1, "Yes", "No"))

freedom %>%
  select(is_ldc, ldc) %>% unique()
```

```
# A tibble: 2 x 2
  is_ldc ldc
   <dbl> <chr>
1      1 Yes
2      0 No
```

Grouping candidates

Let's examine the total number of unique entries per column to see good candidates for faceting or other categorization methods.

```
freedom %>%
  sapply(n_distinct)
```

```
    country        year          CL          PR      Status Region_Code
        193          26           7           7           3           5
Region_Name     is_ldc         ldc
          5           2           2
```

We can see that `LDC`, `Status` and `Region_Name` are all potential ways to cluster data.

Developing Questions

While examining the data set, the following potential questions arose for investigation:

1. How have the 5 of the best, and 5 of worst non LDC countries shifted in terms of freedom from 1995 to 2020?
2. How are proportions of Free, Partially Free, and Not Free countries shifting over time?
3. How are the distribution of Political Rights, and Civil Liberties fluctuating over time by region?
4. How are the distribution of Political Rights, and Civil Liberties fluctuating over time by LDC designator?

For the first question, we will need to set criteria for "worst" and "best", and then identify these countries. We will use a combined `PR + CL` score as the gauge of measure. A score of 14 would represent "worst" and 2 would represent "best". We will also exclude countries that do not have data for all years.

```
 freedom = freedom %>%
  mutate(PR_CL = PR + CL)

freedom %>%
  filter(year == 1995, ldc == "No") %>%
  select(country, Region_Name, ldc, PR_CL) %>%
  arrange(PR_CL) %>% head(20)
```

```
# A tibble: 20 x 4
   country                  Region_Name ldc   PR_CL
   <chr>                    <chr>       <chr> <dbl>
 1 Andorra                  Europe      No        2
 2 Australia                Oceania     No        2
 3 Austria                  Europe      No        2
 4 Barbados                 Americas    No        2
 5 Belgium                  Europe      No        2
 6 Belize                   Americas    No        2
 7 Canada                   Americas    No        2
 8 Cyprus                   Asia        No        2
 9 Denmark                  Europe      No        2
10 Dominica                 Americas    No        2
11 Finland                  Europe      No        2
12 Iceland                  Europe      No        2
13 Ireland                  Europe      No        2
14 Liechtenstein            Europe      No        2
15 Luxembourg               Europe      No        2
16 Malta                    Europe      No        2
17 Marshall Islands         Oceania     No        2
```

```
18 Micronesia (Federated States of) Oceania     No        2
19 Netherlands                      Europe      No        2
20 New Zealand                      Oceania     No        2
```

```
freedom %>%
  filter(year == 1995, ldc == "No") %>%
  select(country, Region_Name, ldc, PR_CL) %>%
  arrange(desc(PR_CL)) %>% head(20)
```

```
# A tibble: 20 x 4
   country                              Region_Name ldc   PR_CL
   <chr>                                <chr>       <chr> <dbl>
 1 China                                Asia        No       14
 2 Cuba                                 Americas    No       14
 3 Equatorial Guinea                    Africa      No       14
 4 Iraq                                 Asia        No       14
 5 Libya                                Africa      No       14
 6 Nigeria                              Africa      No       14
 7 Democratic People's Republic of Korea Asia       No       14
 8 Saudi Arabia                         Asia        No       14
 9 Syrian Arab Republic                 Asia        No       14
10 Tajikistan                           Asia        No       14
11 Turkmenistan                         Asia        No       14
12 Uzbekistan                           Asia        No       14
13 Viet Nam                             Asia        No       14
14 Indonesia                            Asia        No       13
15 Iran (Islamic Republic of)           Asia        No       13
16 Kenya                                Africa      No       13
17 Qatar                                Asia        No       13
18 Algeria                              Africa      No       12
19 Azerbaijan                           Asia        No       12
20 Bahrain                              Asia        No       12
```

```
freedom %>%
  filter(year == 1995, ldc == "No", PR_CL == 2) %>%
  arrange(PR_CL) %>%
  select(country) %>%
  pull()
```

```
 [1] "Andorra"                          "Australia"
 [3] "Austria"                          "Barbados"
 [5] "Belgium"                          "Belize"
 [7] "Canada"                           "Cyprus"
 [9] "Denmark"                          "Dominica"
[11] "Finland"                          "Iceland"
[13] "Ireland"                          "Liechtenstein"
[15] "Luxembourg"                       "Malta"
[17] "Marshall Islands"                 "Micronesia (Federated States of)"
[19] "Netherlands"                      "New Zealand"
[21] "Norway"                           "Portugal"
[23] "San Marino"                       "Sweden"
[25] "Switzerland"                      "United States of America"
```

```
freedom %>%
  filter(year == 1995, ldc == "No", PR_CL == 14) %>%
  arrange(PR_CL) %>%
```

```
  select(country) %>%
  pull()
```

```
 [1] "China"
 [2] "Cuba"
 [3] "Equatorial Guinea"
 [4] "Iraq"
 [5] "Libya"
 [6] "Nigeria"
 [7] "Democratic People's Republic of Korea"
 [8] "Saudi Arabia"
 [9] "Syrian Arab Republic"
[10] "Tajikistan"
[11] "Turkmenistan"
[12] "Uzbekistan"
[13] "Viet Nam"
```

```
exclude_list <- freedom %>%
  group_by(country) %>%
  summarize(n = n()) %>%
  filter(n != 26) %>%
  select(country) %>% pull()
exclude_list
```

```
[1] "Montenegro"  "Serbia"      "South Sudan" "Timor-Leste"
```

A look at the data showed that there was not enough granularity in the scale to determine soley based off the data, thus we chose the following 5 countries based off interest and recent political events.

```
best_list <- c("Australia", "Canada", "Norway", "Sweden", "United States of America")
worst_list <- c("China", "Cuba", "Iraq", "Saudi Arabia", "Viet Nam")
best_list
```

```
[1] "Australia"                "Canada"
[3] "Norway"                   "Sweden"
[5] "United States of America"
```

```
worst_list
```

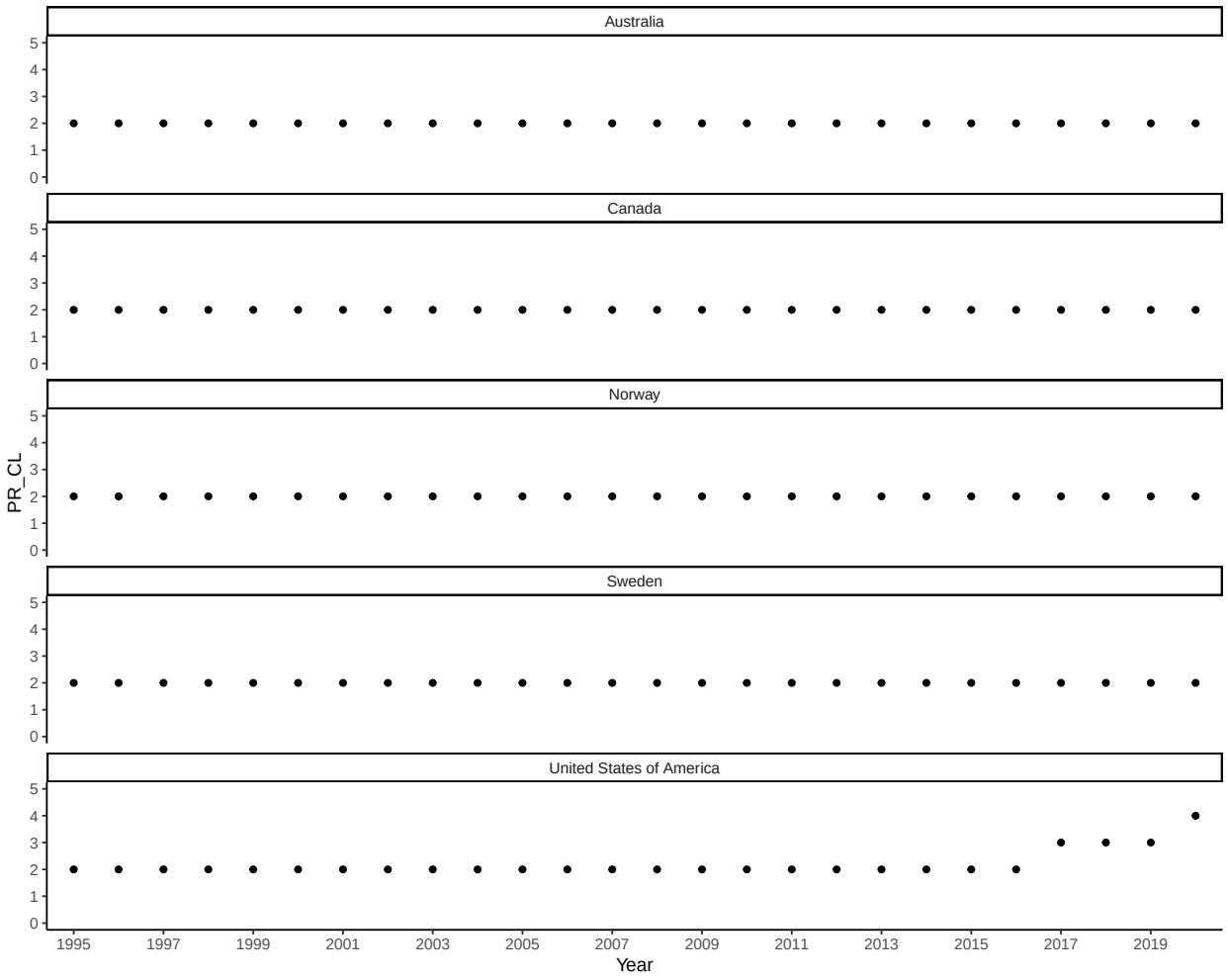
```
[1] "China"        "Cuba"         "Iraq"         "Saudi Arabia" "Viet Nam"
```

Data Visualizations (WORK IN PROGRESS)

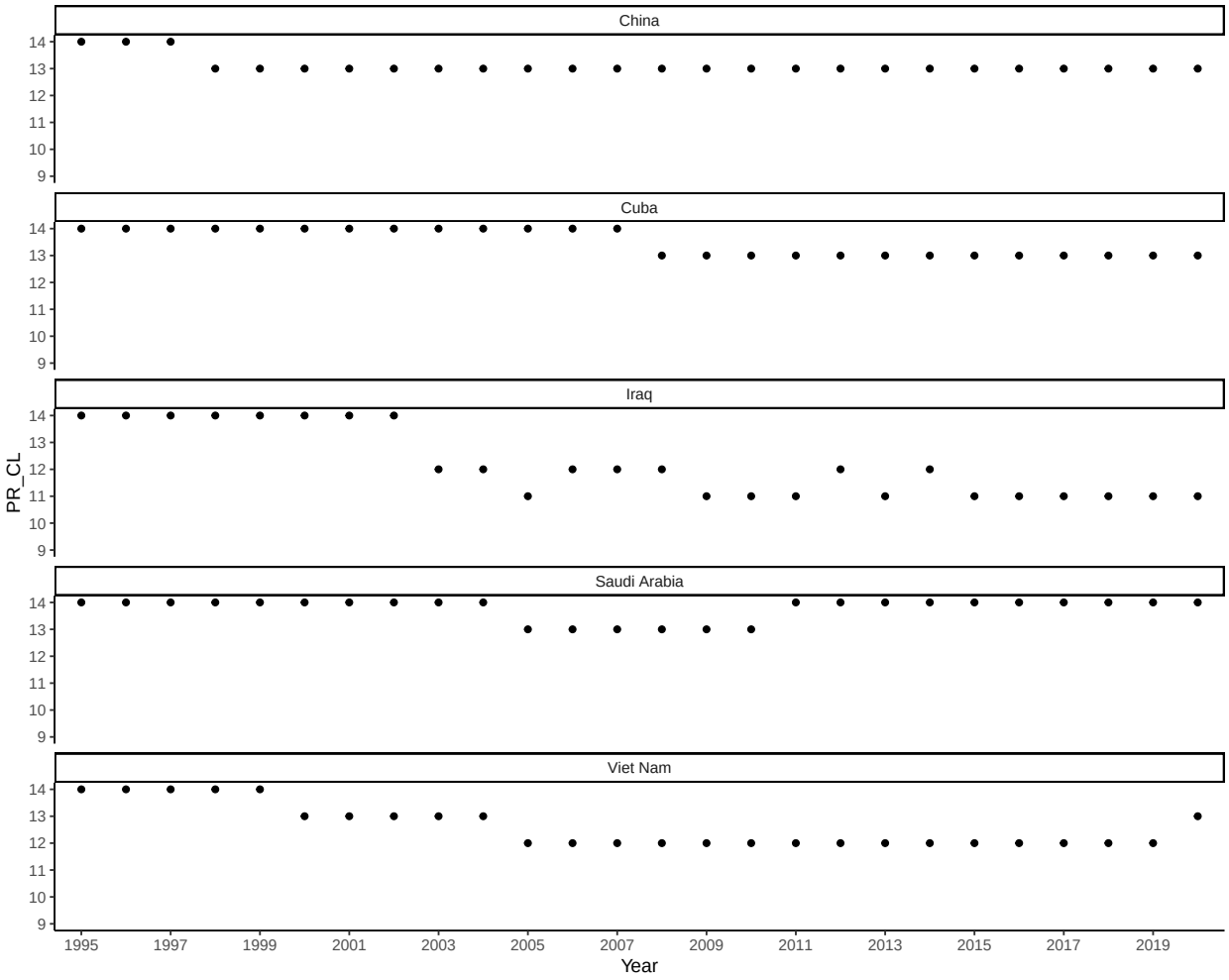
Trends for top and worst 5

```
freedom %>%
  filter(country %in% best_list) %>%
  ggplot() +
  aes(x = factor(year), y = PR_CL) +
  facet_wrap(
    vars(country),
    ncol = 1) +
  geom_point() +
  scale_y_continuous(
    limit = c(0, 5)) +
  scale_x_discrete(
```

```
    name = "Year",
    labels = factor(seq(from = 1995, to = 2020, by = 2)),
    breaks = factor(seq(from = 1995, to = 2020, by = 2))) +
  theme_classic()
```



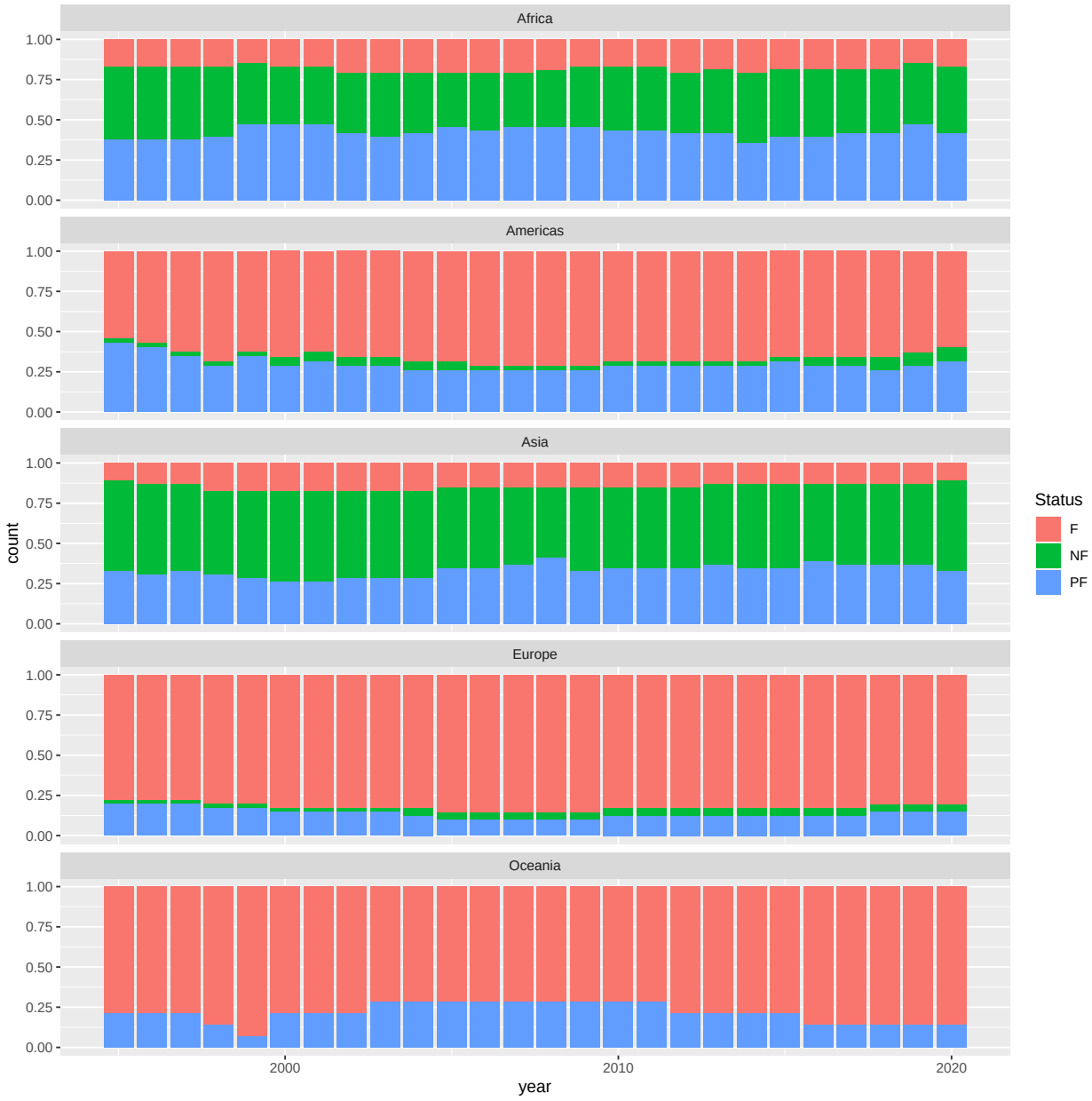

```
freedom %>%
  filter(country %in% worst_list) %>%
  ggplot() +
  aes(x = factor(year), y = PR_CL) +
  facet_wrap(
    vars(country),
    ncol = 1) +
  geom_point() +
  scale_y_continuous(
    limit = c(9, 14)) +
  scale_x_discrete(
    name = "Year",
    labels = factor(seq(from = 1995, to = 2020, by = 2)),
    breaks = factor(seq(from = 1995, to = 2020, by = 2))) +
  theme_classic()
```



Unfortunately, these plots ended up poor visually. There was also limited movement in the metrics. We could speculate that maybe those with the highest and lowest freedom scores are most "stable" in regards to the operations of their current regime form. A notable exception among those countries scoring "high" in freedom is the U.S., in which this metric shows a decline in freedom starting in 2017.

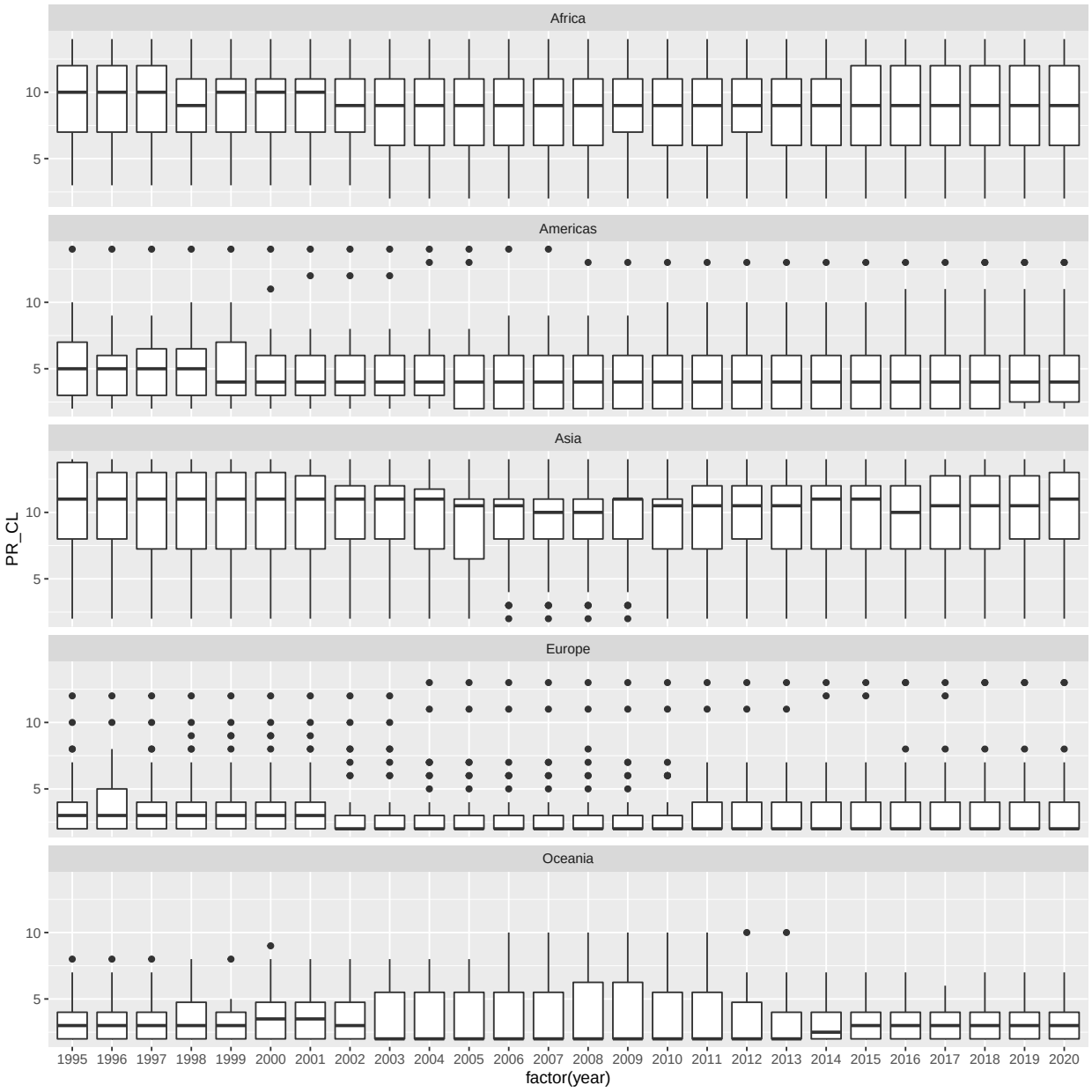
Overall trends in freedom over time

```
freedom %>%
  filter(!country %in% exclude_list) %>%
  ggplot() +
  aes(x = year, fill = Status) +
  geom_bar(position = "fill") +
  facet_wrap(
  vars(Region_Name),
  ncol = 1)
```



Overall trends in freedom over time by region

```
freedom %>%
  filter(!country %in% exclude_list) %>%
  ggplot() +
  aes(x = factor(year), y = PR_CL) +
  geom_boxplot() +
  facet_wrap(
    vars(Region_Name),
    ncol = 1)
```


Africa
Americas
Asia
Europe
Oceania
PR_CL
10
5
1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
factor(year)

Conclusion