Eksplorativna analiza podataka (Exploratory Data Analysis - EDA) u R-u: Studija slucaja - UN Voting Dataset

Igor Hut

09 January, 2017

Contents

Uvodna razmatranja

- Izvor: *Erik Voeten, "Data and Analyses of Voting in the UN General Assembly"*
- Svi podaci su dostupni u okviru R paketa "unvote": https://cran.r-project.org/web/packages/unvotes/
- Kao i u GitHub repozitorujumu: https://github.com/dgrtwo/unvotes
- Sadrzani podaci se odnose na istoriju glasanja zemalja clanica na Generalnoj skupstini Ujedinjenih nacija. Date su informacije o datumu glasanja, temi o kojoj se glasalo i kako je svaka od zemalja glasala.

Setovi podataka

`unvotes` R paket sadrzi tri seta podataka u formi data frame-a (preciznije tbl_df, odn. tibble, sto obezbedjuje bolje formatiranje pri njihovom ispisivanju). Prvi set podataka, `un_votes` se odnosi na istoriju glasanja svake od zemalja. Svaka vrsta sadrzi country/vote par:

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(unvotes)
```

```
## Warning: package 'unvotes' was built under R version 3.3.2
```

```
## If you use data from the unvotes package, please cite the following:
##
## Erik Voeten "Data and Analyses of Voting in the UN General Assembly" Routledge Handbook of Internati
```

```
str(un_votes)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    711275 obs. of  3 variables:
##  $ rcid   : atomic  3 3 3 3 3 3 3 3 3 3 ...
##   ..- attr(*, "comment")= chr "rcid"
##  $ country: chr  "Egypt" "Honduras" "Costa Rica" "El Salvador" ...
##  $ vote   : Factor w/ 3 levels "abstain","no",..: 1 3 3 3 2 3 3 3 3 3 ...
```

```
un_votes
```

```
## # A tibble: 711,275 × 3
##     rcid     country    vote
##    <dbl>       <chr>  <fctr>
## 1      3       Egypt abstain
## 2      3    Honduras     yes
## 3      3  Costa Rica     yes
## 4      3 El Salvador     yes
## 5      3      France      no
## 6      3     Uruguay     yes
## 7      3       Chile     yes
## 8      3     Ecuador     yes
## 9      3   Argentina     yes
## 10     3       Haiti     yes
## # ... with 711,265 more rows
```

Paket takodje sadrzi i set podataka sa informacijama o svakom javnom glasanju, ukljucujuci datum, opis, i rezoluciju o kojoj je glasano:

```
un_roll_calls
```

```
## # A tibble: 5,356 × 9
##     rcid session importantvote       date   unres amend  para
##    <dbl>   <dbl>         <dbl>     <date>   <chr> <dbl> <dbl>
## 1      3       1             0 1946-01-01  R/1/66     1     0
## 2      4       1             0 1946-01-02  R/1/79     0     0
## 3      5       1             0 1946-01-04  R/1/98     0     0
## 4      6       1             0 1946-01-04 R/1/107     0     0
## 5      7       1             0 1946-01-02 R/1/295     1     0
## 6      8       1             0 1946-01-05 R/1/297     1     0
## 7      9       1             0 1946-02-05 R/1/329     0     0
## 8     10       1             0 1946-02-05 R/1/361     1     1
## 9     11       1             0 1946-02-05 R/1/376     0     0
## 10    12       1             0 1946-02-06 R/1/394     1     1
## # ... with 5,346 more rows, and 2 more variables: short <chr>, descr <chr>
```

Konacno `un_roll_call_issues` set podataka sadrzi informacije o medjusobnoj povezanosti razlicitih glasanja kao i o 6 generalnih problema o kojima je glasano u Gen. skupstini UN-a:

```
un_roll_call_issues
```

```
## # A tibble: 4,951 × 3
```

```
##     rcid short_name                issue
##    <dbl>      <chr>                <chr>
## 1     30         me Palestinian conflict
## 2     34         me Palestinian conflict
## 3     77         me Palestinian conflict
## 4   9002         me Palestinian conflict
## 5   9003         me Palestinian conflict
## 6   9004         me Palestinian conflict
## 7   9005         me Palestinian conflict
## 8   9006         me Palestinian conflict
## 9    128         me Palestinian conflict
## 10   129         me Palestinian conflict
## # ... with 4,941 more rows
```

```
count(un_roll_call_issues, issue, sort = TRUE)
```

```
## # A tibble: 6 × 2
##                                  issue     n
##                                  <chr> <int>
## 1                 Palestinian conflict  1047
## 2                           Colonialism   971
## 3                         Human rights   901
## 4         Arms control and disarmament   859
## 5 Nuclear weapons and nuclear material   712
## 6                 Economic development   461
```

Za vise informacija o svakom od pojedinacnih setova podataka koristite naredbu `help()`.

Priprema podataka za analizu

Za pocetak cemo da izvrsimo "inner join" za setove `un_votes` i `un_roll_calls` a na osnovu zajednicke kolone `rcid`. Na ovaj nacin dobijamo objedinjen set podataka sa vecinom informacija relevantnih za dalju eksplorativnu analizu.

```
joined <- inner_join(un_votes, un_roll_calls, by = "rcid")

joined
```

```
## # A tibble: 711,275 × 11
##     rcid     country    vote session importantvote       date  unres amend
##    <dbl>       <chr>  <fctr>   <dbl>         <dbl>     <date>  <chr> <dbl>
## 1      3       Egypt abstain       1             0 1946-01-01 R/1/66     1
## 2      3    Honduras     yes       1             0 1946-01-01 R/1/66     1
## 3      3  Costa Rica     yes       1             0 1946-01-01 R/1/66     1
## 4      3 El Salvador     yes       1             0 1946-01-01 R/1/66     1
## 5      3      France      no       1             0 1946-01-01 R/1/66     1
## 6      3     Uruguay     yes       1             0 1946-01-01 R/1/66     1
## 7      3       Chile     yes       1             0 1946-01-01 R/1/66     1
## 8      3     Ecuador     yes       1             0 1946-01-01 R/1/66     1
## 9      3   Argentina     yes       1             0 1946-01-01 R/1/66     1
## 10     3       Haiti     yes       1             0 1946-01-01 R/1/66     1
## # ... with 711,265 more rows, and 3 more variables: para <dbl>,
## #   short <chr>, descr <chr>
```

Eksplorativna analiza

Koliko ukupno ima glasova, za sva glasanja i sve godine, i koliko je procentualno bilo glasova "za" u odnosu na ukupan broj glasova?

```
summarise(joined, total = n(), percent_yes = mean(vote == "yes"))
```

```
## # A tibble: 1 × 2
##    total percent_yes
##    <int>       <dbl>
## 1 711275   0.7963952
```

Kako su glasale pojedinacne zemlje, u proseku, tokom istorije Gen. skupstine UN?

```
by_country <- joined %>%
    group_by(country) %>%
    summarise(n_votes = n(),
            percent_yes = mean(vote == "yes"))

# Print the by_country dataset
by_country
```

```
## # A tibble: 200 × 3
##                country n_votes percent_yes
##                  <chr>   <int>       <dbl>
## 1          Afghanistan    4824   0.8381012
## 2              Albania    3363   0.7204877
## 3              Algeria    4374   0.8978052
## 4              Andorra    1410   0.6510638
## 5               Angola    2950   0.9223729
## 6  Antigua and Barbuda    2521   0.9170964
## 7            Argentina    5207   0.7743422
## 8              Armenia    1479   0.7592968
## 9            Australia    5245   0.5542421
## 10             Austria    4786   0.6320518
## # ... with 190 more rows
```

```
arrange(by_country, percent_yes)
```

```
## # A tibble: 200 × 3
##                             country n_votes percent_yes
##                               <chr>   <int>       <dbl>
## 1                          Zanzibar       2   0.0000000
## 2                     United States    5237   0.2850869
## 3                             Palau     777   0.3063063
## 4                            Israel    4790   0.3503132
## 5      Federal Republic of Germany    2151   0.3984193
## 6  Micronesia, Federated States of    1341   0.4131245
## 7                     United Kingdom    5218   0.4269835
## 8                            France    5171   0.4320248
## 9                  Marshall Islands    1468   0.4788828
## 10                          Belgium    5238   0.4925544
## # ... with 190 more rows
```

```
arrange(by_country, desc(percent_yes))
```

```
## # A tibble: 200 × 3
##                   country n_votes percent_yes
##                     <chr>   <int>       <dbl>
## 1              Seychelles    1698   0.9770318
## 2             Timor-Leste     697   0.9670014
## 3  Sao Tome and Principe    2329   0.9665092
## 4                Djibouti    3193   0.9564673
## 5           Guinea-Bissau    2933   0.9546539
## 6                 Comoros    2435   0.9462012
## 7              Cabo Verde    3153   0.9454488
## 8              Mozambique    3306   0.9431337
## 9                   Yemen    1527   0.9423707
## 10              Zimbabwe    2766   0.9421547
## # ... with 190 more rows
```

```
arrange(by_country, n_votes)
```

```
## # A tibble: 200 × 3
##         country n_votes percent_yes
##           <chr>   <int>       <dbl>
## 1     Zanzibar       2   0.0000000
## 2     Kiribati      93   0.8172043
## 3  South Sudan      96   0.6979167
## 4   Montenegro     558   0.6433692
## 5       Tuvalu     576   0.8246528
## 6        Nauru     606   0.6089109
## 7  Timor-Leste     697   0.9670014
## 8        Tonga     775   0.7303226
## 9        Palau     777   0.3063063
## 10 Switzerland     857   0.6569428
## # ... with 190 more rows
```

Moze se primetiti da ima nekoliko zemalja koje su znacajno manje puta ucestvovale u glasanju od ostalih (Zanzibar, Kiribati, South Sudan). Ove zemlje i podatke vezane za njih cemo izostatiti iz buduce analize. Filterujemo set podataka tako da izostavimo sve zemlje koje su glasale manje od 100 puta:

```
by_country %>%
filter( n_votes > 100) %>%
  arrange(percent_yes)
```

```
## # A tibble: 197 × 3
##                             country n_votes percent_yes
##                               <chr>   <int>       <dbl>
## 1                     United States    5237   0.2850869
## 2                             Palau     777   0.3063063
## 3                            Israel    4790   0.3503132
## 4      Federal Republic of Germany    2151   0.3984193
## 5  Micronesia, Federated States of    1341   0.4131245
## 6                    United Kingdom    5218   0.4269835
## 7                            France    5171   0.4320248
## 8                 Marshall Islands    1468   0.4788828
## 9                           Belgium    5238   0.4925544
## 10                       Luxembourg    5169   0.5105436
## # ... with 187 more rows
```

Kako se menjao generalni trend glasanja kroz istoriju?

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.3.2
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

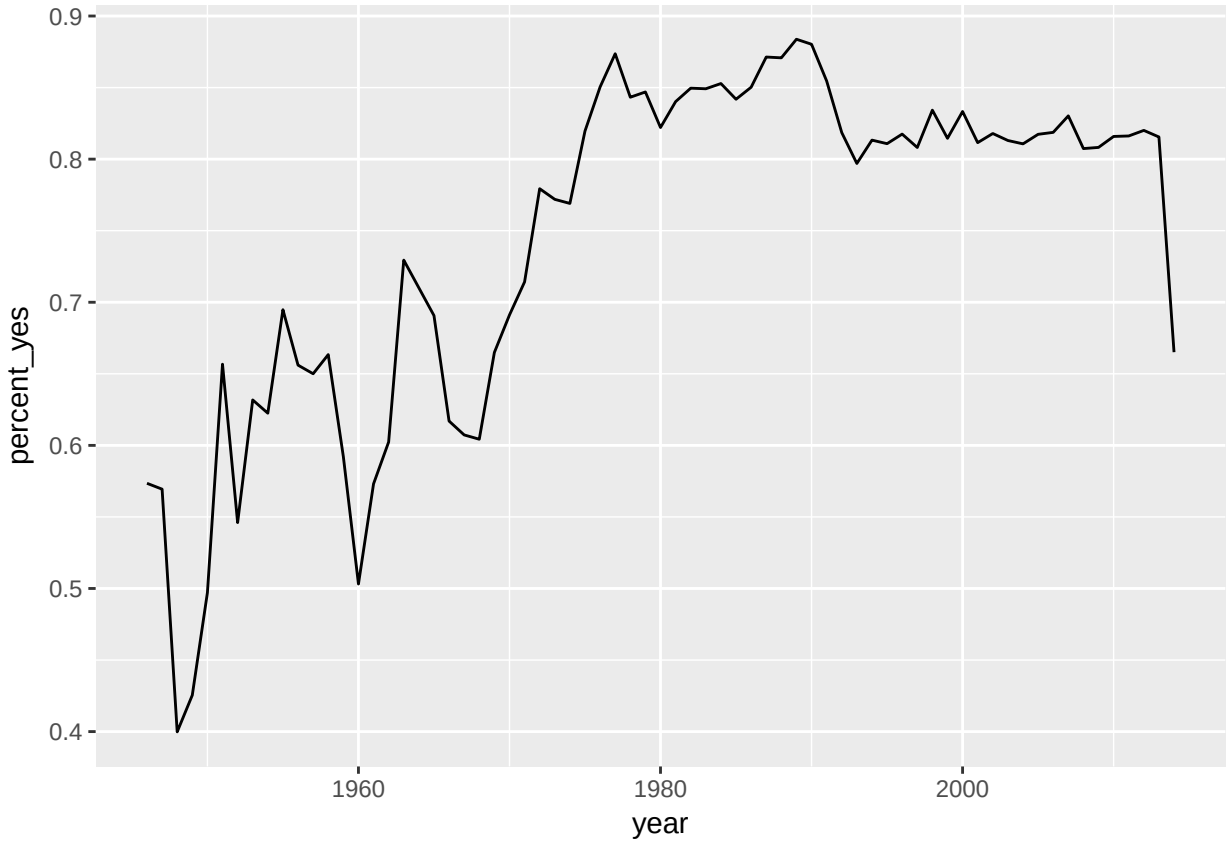
```
by_year <- joined %>%
  group_by(year = year(date)) %>%
  summarize(n_votes = n(),
            percent_yes = mean(vote == "yes")) %>%
  filter( n_votes > 100)

by_year
```

```
## # A tibble: 68 × 3
##     year n_votes percent_yes
##    <dbl>   <int>       <dbl>
## 1   1946    2143   0.5734951
## 2   1947    2039   0.5693968
## 3   1948    3454   0.3998263
## 4   1949    5700   0.4254386
## 5   1950    2911   0.4970800
## 6   1951     402   0.6567164
## 7   1952    4082   0.5460559
## 8   1953    1537   0.6317502
## 9   1954    1788   0.6224832
## 10  1955    2169   0.6947902
## # ... with 58 more rows
```

```
# Da vizualizujemo ovaj trend pomocu linijskog grafika
library(ggplot2)
```

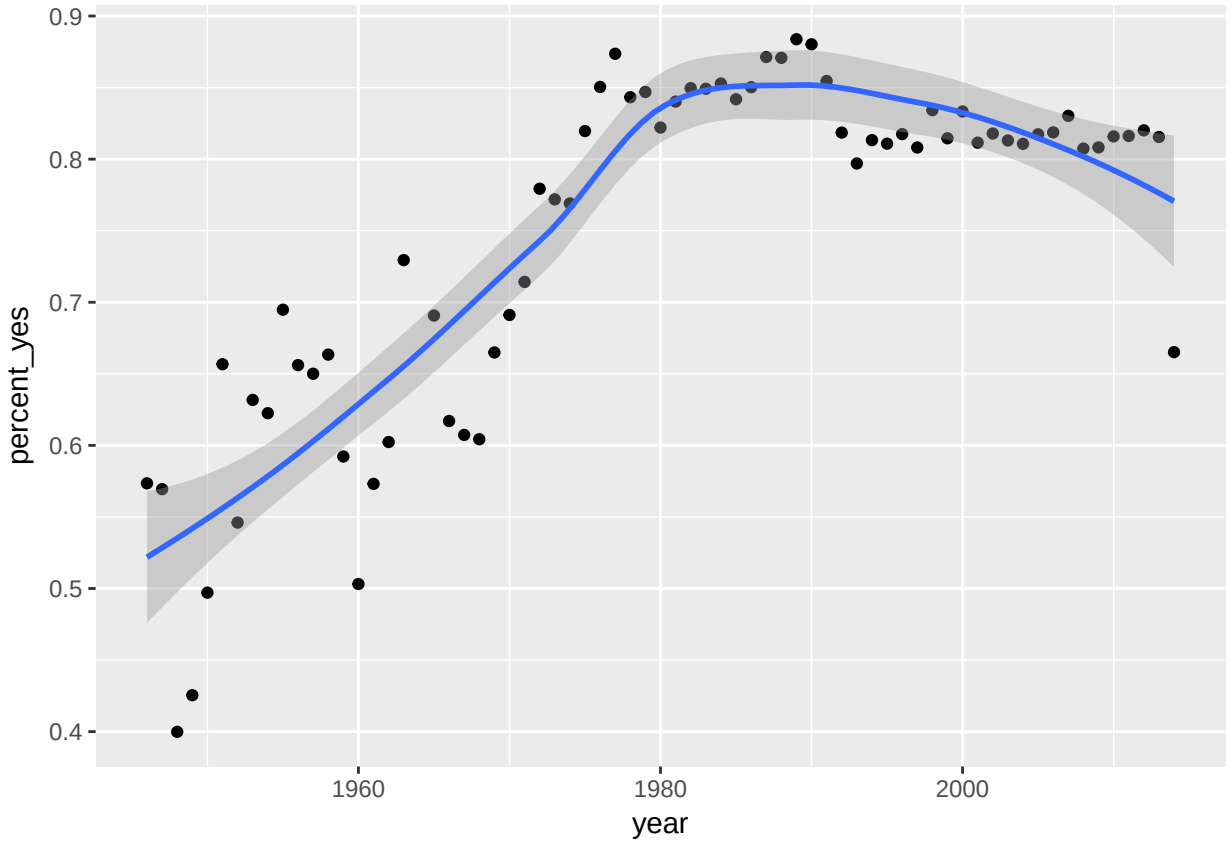
```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
ggplot(by_year, aes(x = year, y = percent_yes)) +
  geom_line()
```




```
# Scatter plot + geom_smooth

ggplot(by_year, aes( year, percent_yes)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess'
```



Zanimljiv trend. Neposredno nakon nakon zavrsetka Drugog svetskog rata su evidentno postojale ostre nesuglasice po mnogim pitanjima. U periodu od 80-tih godina proslog veka, pa sve do skoro, deluje kao da je postojao, gotovo, koncenzus po mnogim pitanjima, medju vecinom clanica Gen. skupstine UN.

Koliko cesto je svaka od zemalja glasala "za" po godinama?

```
library(lubridate)

by_year_country <- joined %>%
  group_by( year = year(date), country) %>%
  summarise(n_votes = n(), percent_yes = mean(vote == "yes"))

by_year_country
```

```
## Source: local data frame [9,496 x 4]
## Groups: year [?]
##
##     year                          country n_votes percent_yes
##    <dbl>                            <chr>   <int>       <dbl>
## 1   1946                      Afghanistan      17   0.4117647
## 2   1946                        Argentina      43   0.6976744
## 3   1946                        Australia      43   0.5581395
## 4   1946                          Belarus      43   0.4418605
## 5   1946                          Belgium      43   0.6046512
## 6   1946 Bolivia, Plurinational State of      43   0.6976744
## 7   1946                           Brazil      43   0.6046512
```

```
## 8   1946                          Canada      42   0.6428571
## 9   1946                           Chile      43   0.6046512
## 10  1946                        Colombia      42   0.3095238
## # ... with 9,486 more rows
```

Kako su kroz istoriju glasale, SAD, SSSR, Rusija, Jugoslavija, Srbija i Hrvatska?

Proverimo prvo da li ima svih ovih zemalja u setu:

```
country <- distinct(by_country, country) %>%
  arrange(country)
View(country)

inner_join(country, data.frame(country = c("Serbia", "Croatia", "United States", "Russian Federation",
```

```
## Joining, by = "country"
```

```
## Warning in inner_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## character vector and factor, coercing into character vector
```

```
## # A tibble: 4 × 1
##               country
##                 <chr>
## 1            Croatia
## 2 Russian Federation
## 3      United States
## 4         Yugoslavia
```

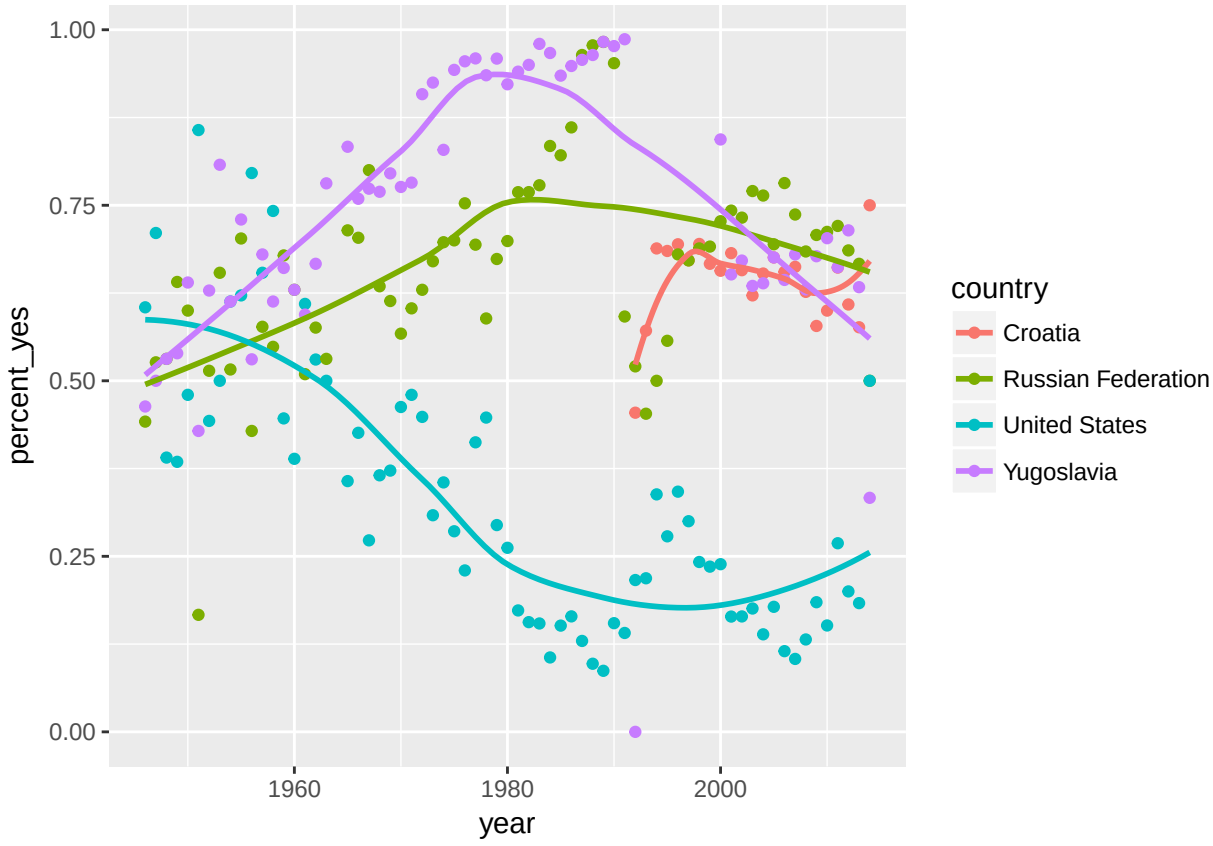
Izgleda da nema ni Srbije ni SSSR-a. Posmatracemo Jugoslaviju, Rusiju, Hrvatsku i SAD:

```
filt_countries <- filter(by_year_country, country %in% c("Croatia", "United States", "Russian Federatio

filt_countries
```

```
## Source: local data frame [220 x 4]
## Groups: year [68]
##
##     year            country n_votes percent_yes
##    <dbl>              <chr>   <int>       <dbl>
## 1   1946 Russian Federation      43   0.4418605
## 2   1946      United States      43   0.6046512
## 3   1946         Yugoslavia      41   0.4634146
## 4   1947 Russian Federation      38   0.5263158
## 5   1947      United States      38   0.7105263
## 6   1947         Yugoslavia      38   0.5000000
## 7   1948 Russian Federation      64   0.5312500
## 8   1948      United States      64   0.3906250
## 9   1948         Yugoslavia      64   0.5312500
## 10  1949 Russian Federation     103   0.6407767
## # ... with 210 more rows
```

```
ggplot(filt_countries, aes(x = year, y = percent_yes, col = country)) +
  geom_point() +
  geom_smooth(se = FALSE)
```

```
## `geom_smooth()` using method = 'loess'
```



Evidentno da su opservacije za Srbiju, Srbiju i Crnu Goru i Jugoslavuju deklarisane kao "Yugoslavia" a za Rusiju i SSSR kao "Russian Federation". Zanimljivo, a donekle i ocekivano, generalni trend glasanja Ruske Federacije i Jugoslavije/ Srbije su prilicno pozitivno korelisani tokom istorije, dok je situacija sa SAD upravo suprotna. Zarad podrobnije komparacije tokom poslednje dve decenije posmatracemo samo period od 1990-te do 2014-te godine.

```
# Samo da jos jednom proverimo raspone vrednosti, pre svega za var. "year"
summary(filt_countries)
```

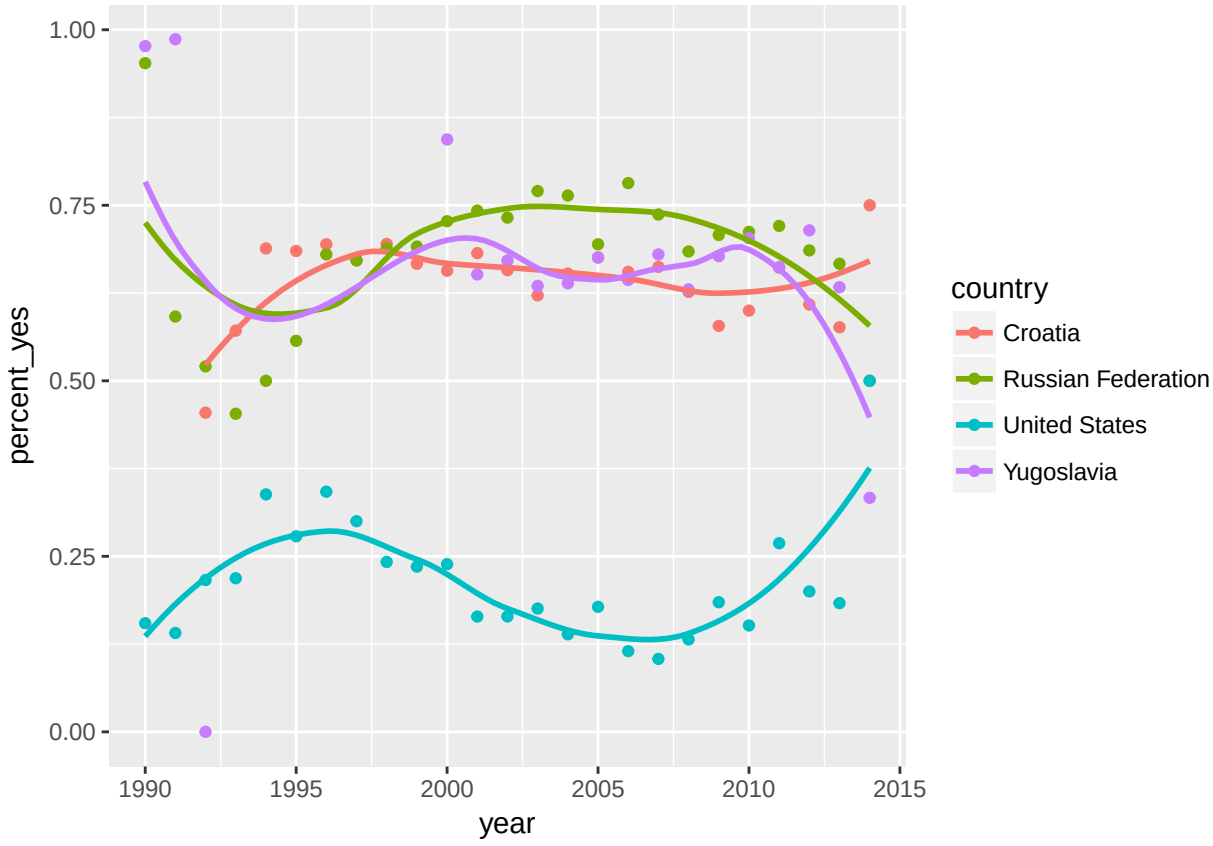
```
##       year        country             n_votes        percent_yes
##  Min.   :1946   Length:220         Min.   :  2.00   Min.   :0.0000
##  1st Qu.:1965   Class :character   1st Qu.: 54.00   1st Qu.:0.4455
##  Median :1983   Mode  :character   Median : 70.00   Median :0.6317
##  Mean   :1982                      Mean   : 75.39   Mean   :0.5812
##  3rd Qu.:2001                      3rd Qu.: 89.00   3rd Qu.:0.7159
##  Max.   :2014                      Max.   :160.00   Max.   :0.9865
```

```
ggplot(filt_countries, aes(x = year, y = percent_yes, col = country)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  scale_x_continuous(limits = c(1990, 2014))
```

```
## `geom_smooth()` using method = 'loess'
```

```
## Warning: Removed 129 rows containing non-finite values (stat_smooth).
```

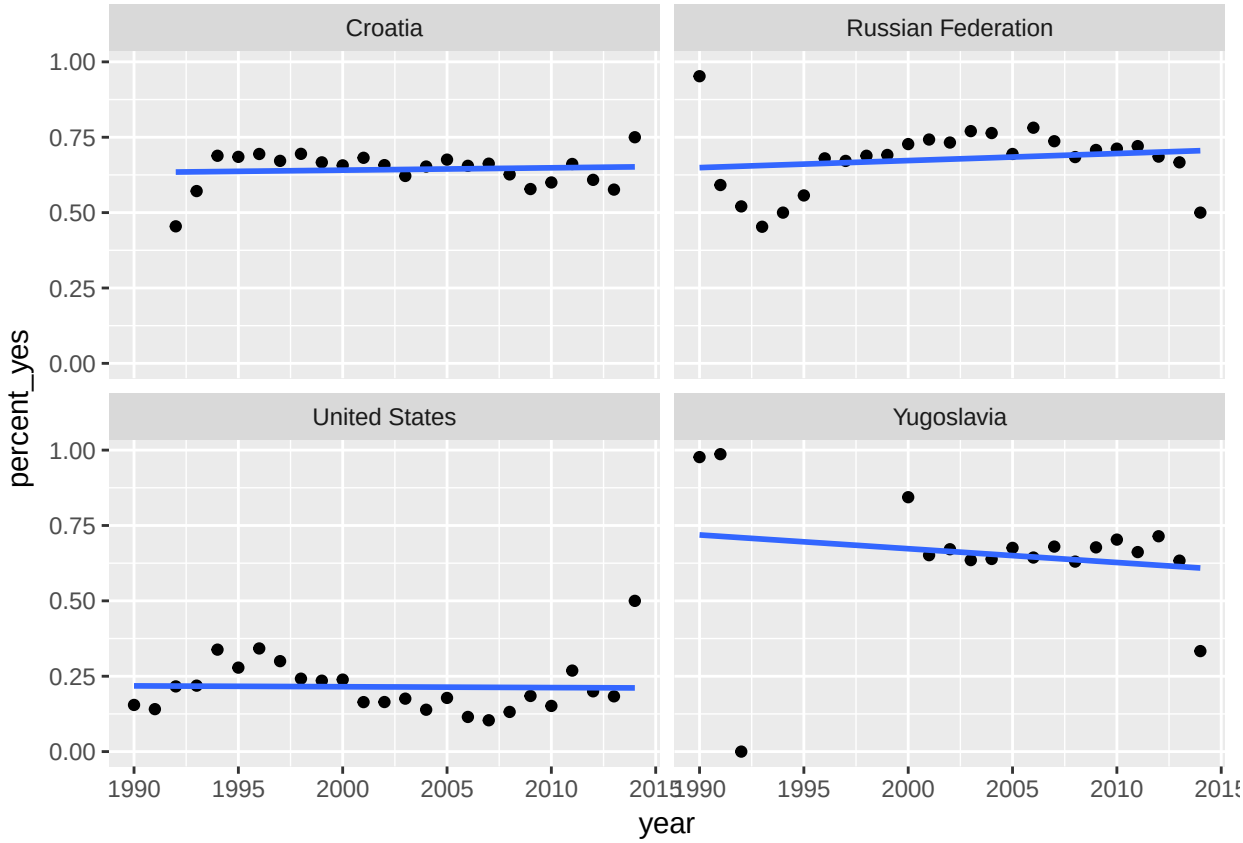
```
## Warning: Removed 129 rows containing missing values (geom_point).
```




```
ggplot(filt_countries, aes(x = year, y = percent_yes)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  scale_x_continuous(limits = c(1990, 2014)) +
  facet_wrap( ~ country)
```

```
## Warning: Removed 129 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 129 rows containing missing values (geom_point).
```



Posmatrajuci izolovan samo period 1990-2014 god. mozemo uociti, mozda malo neocekivano, da su Srbija, Ruska Federacija i Hrvatska imale procentualno mnogo vise glasova "za" od SAD, te da se trend glasanja ove tri zemlje prilicno poklapa u datom periodu. Sta vise ovo pogotovo vazi za Hrvatsku i Rusku Federaciju. Naravno potrebno je dataljnije utvrditi kako je glasano za pojedinacne rezolucije pre nego sto izvucemo bilo kakve zakljucke.

Kako je svaka od izabranih zemalja glasala tokom vremena o generalnim problemima sadrzanim u 'un_roll_call_issues?

Prvo da proverimo o kojim se problemima tacno radi:

```
distinct(un_roll_call_issues, issue)
```

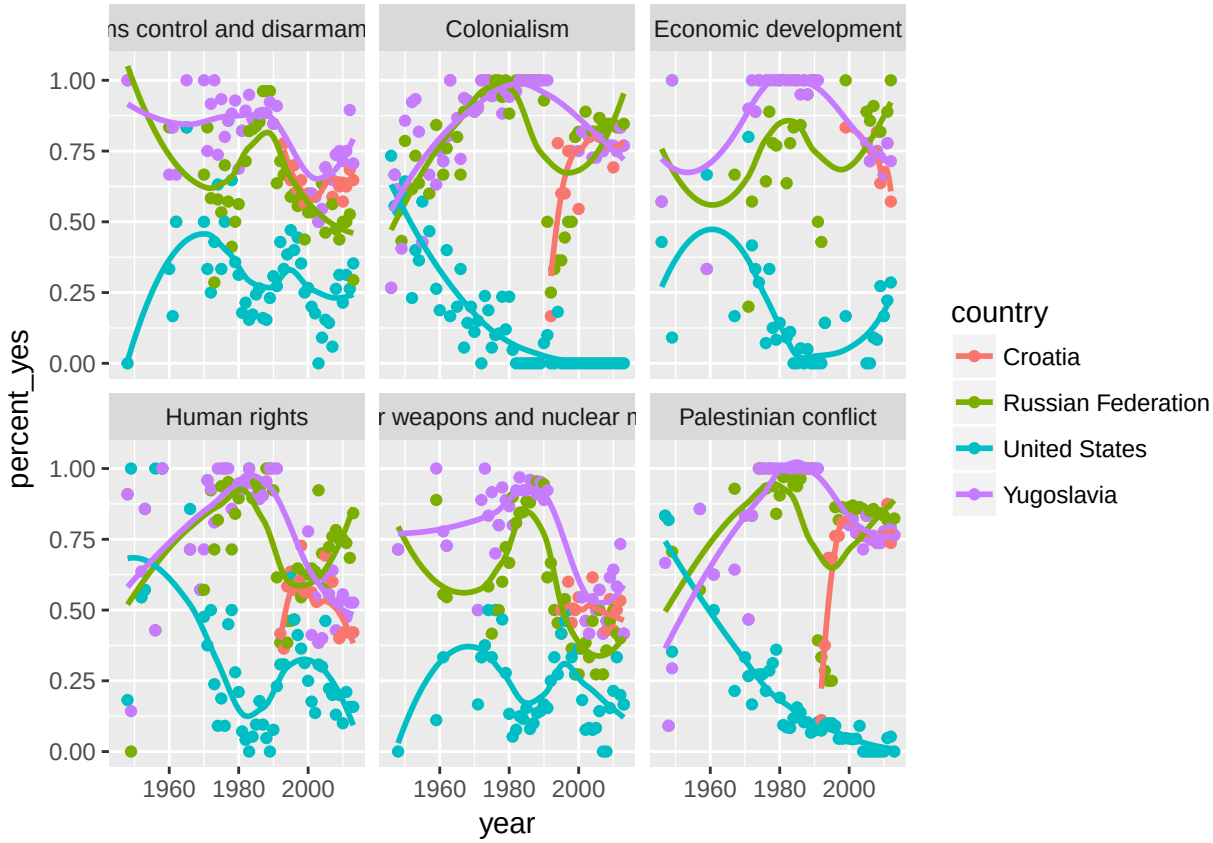
```
## # A tibble: 6 × 1
##                                 issue
##                                 <chr>
## 1                Palestinian conflict
## 2 Nuclear weapons and nuclear material
## 3         Arms control and disarmament
## 4                        Human rights
## 5                         Colonialism
## 6                Economic development
```

Nazalost u ovom setu nema podataka o glasanjima koja su se direktno ticala politickih desavanaja na prostoru bivse Jugoslavije, sto bi nama bilo posebno zanimljivo. Elem, nastavimo sa analizom:

```
joined %>%
  filter(country %in% c("Croatia", "United States", "Russian Federation", "Yugoslavia")) %>%
```

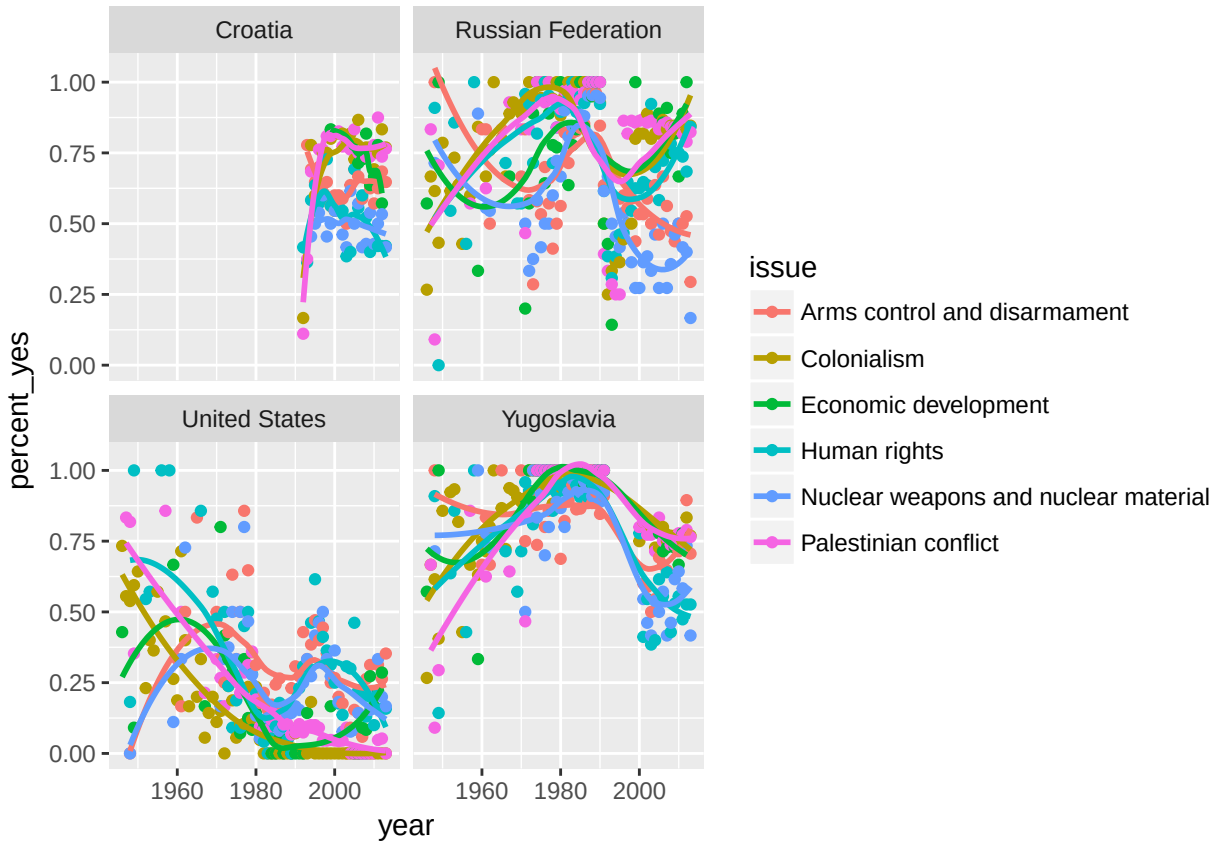
```
  inner_join(un_roll_call_issues, by = "rcid") %>%
  group_by(year = year(date), country, issue) %>%
  summarize(votes = n(),
            percent_yes = mean(vote == "yes")) %>%
  filter(votes > 5) %>%
  ggplot(aes(year, percent_yes, col = country)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  facet_wrap(~ issue)
```

```
## `geom_smooth()` using method = 'loess'
```




```
#Da probamo i sa drugacijim rasporedom
joined %>%
  filter(country %in% c("Croatia", "United States", "Russian Federation", "Yugoslavia")) %>%
  inner_join(un_roll_call_issues, by = "rcid") %>%
  group_by(year = year(date), country, issue) %>%
  summarize(votes = n(),
            percent_yes = mean(vote == "yes")) %>%
  filter(votes > 5) %>%
  ggplot(aes(year, percent_yes, col = issue)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  facet_wrap(~ country)
```

```
## `geom_smooth()` using method = 'loess'
```



U nastavku ce biti sprovedena korelaciona analiza, regresiona analiza, kao i detaljna graficka analiza za odabrane zemlje i njihovo glasanje u Gen. skupstini UN a u svetlu razlicitih kriterijuma...