Individual-level analysis: data preparation and descriptive statistics

Part of the final project for AQMSS II

Stepan Polikanov Vera Okisheva

Disclaimer: most of this part was done as an answer to Problem Set 03 by Stepan Polikanov.

Goals

The goal of this part is to provide an in-depth look into variables created for the individual-level analysis. This work includes substantive comments on particular variable choices as well as descriptive statistics that can be used in the final paper. It is dependent on the "raw_data_prep.R" script which gives the base dataframe. Data manipulations performed there are of purely technical nature as opposed to this notebook.

Preparing individual-level data

```
source(here::here("utilities", "check_packages.R"))

conflicts_prefer(dplyr::filter)
```

```
ep_raw_clean <- read_rds(here("data", "data_built", "ep_raw_clean.rds"))
```

Preparing Independent variables

Data

I intend on using most of the available poll questionnaire items to predict individual votes. The exit poll was conducted in 44 countries and 65 voting stations within them by the \href[https://voteabroad.info/#about-block]{Vote Abroad Initiative}, self-described as founded by "free people and independent activists from Russia living abroad". This means

that this endeavor is easily labelled within the "non-systemic opposition" in Russia by both descriptive signals (civic engagement, control of elections to avoid electoral fraud and simply "activism") and scope of operation (WEIRD countries, mostly within OECD and popular Russian tourist or immigration spots such as Vietnam or Kazakhstan).

One important contextual note is that this election was targeted by exactly one voting strategy proposition from the non-systemic opposition: "Afternoon against Putin". As the ballot did not offer any satisfactory alternative candidates (two anti-war candidates were not allowed to compete on bureaucratic grounds) Navalniy's Anti Corruption Foundation and other democratic forces such as the Anti War Committee of Russia called to show up at polls at 12 o'clock local time and cast a vote for anyone but Putin. This was meant to create a visual cue of opposition backers in the context of 3-day elections in Russia itself as a relatively safe way of passive protest. Later some activists semi-endorsed Davankov, a candidate from the new "New People" party, saying that while he was not anyone's choice he was the least deplorable of the bunch.

Exit poll authors uploaded raw data from volunteers where exit polls were conducted. These include in total 69261 questionnaires collected by 442 volunteers. The variables of interest to be used as predictors include:

- Gender of the respondent
  - Three responses: male, female, other or NA
- Age of the respondent
  - Four responses: 18-24, 25-44, 45-64, 65+ or NA
- Time traveled to the voting station
  - Six responses: <30 minutes, 30 minutes - 1 hour, 1 - 2 hours, 2 - 3 hours, 3 - 4 hours, > 4 hours (staying for the night), Declined to answer or NA
- Time living outside of Russia
  - Six responses: < 6 months, 6 months - 2 years, 2 - 5 years, /> 5 years, > 10 years, Tourist (lives in Russia), Declined to answer or NA
- Trust in the fairness of the election result
  - Five responses: definitely yes, definitely no, probably yes, probably no, struggle to answer and decline to answer or NA.

The data has an obvious nested structure with questionnaires (voters) nested within volunteers nested within voting stations nested within countries.

Variables

The following section attempts to resolve multiple issues. First is to homogenize items and responses. Second is to investigate patterns in data missingness. Lastly, I re-operationalize variables to make them more coherent and theoretically relevant.

Some volunteers, cities and countries reported slightly different items. For example, Sydney and Wellington didn't ask any questions other than the candidate choice, [1] Prague has multiple deviations from the questionnaire and one question was reformulated. [2]. Moreover, I find at least one case of a volunteer misinterpreting items.

The initiative that handled the poll reports that there were multiple interpretations for the "Time traveled to the voting station" question - some respondents may have included time spent in the queue to vote in the estimate and volunteers specified that only time to the location was meant only when asked.

Demographic structure

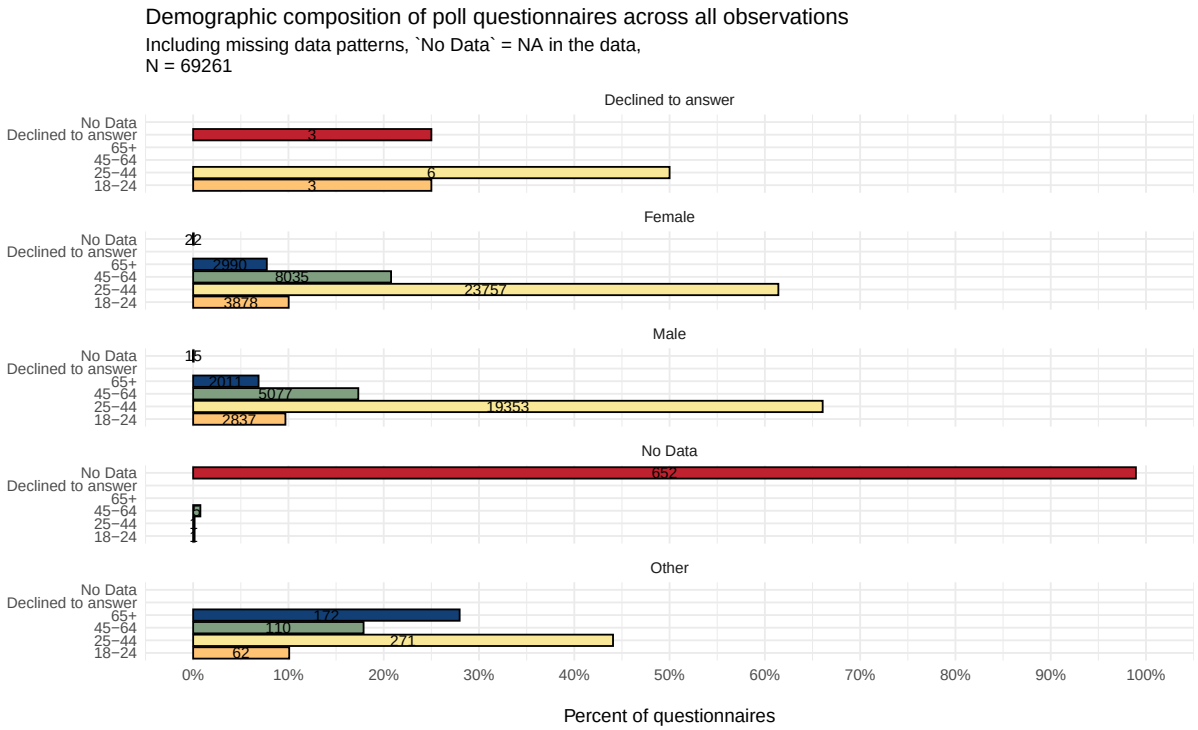
Starting with age and gender variables I first plot their counts and co-distributions.

```
# Demographic

## Original data
ep_raw_clean |>
  mutate(sex = if_else(is.na(sex), "No Data", sex),
         age_bin = if_else(is.na(age_bin), "No Data", age_bin)) |>
  group_by(sex, age_bin) |>
  summarize(n = n()) |>
  mutate(pct = n/sum(n)*100) |>
  ggplot(aes(x = pct, y = age_bin, fill = age_bin)) +
    geom_bar(stat = "identity", color = "black") +
    geom_text(aes(label = n), size = 3, position = position_stack(vjust = 0.5)) +
    facet_wrap(~ sex, nrow = 5) +
    scale_fill_manual(values = c("#FFC374", "#F9E897", "#7F9F80",
                                 "#124076", "#bf212f", "#bf212f")) +
    scale_x_continuous(limits = c(0, 100),
                       breaks = seq(0, 100, 10),
                       label = scales::label_number(suffix = "%")) +
    labs(x = "\nPercent of questionnaires",
         y = "",
         title = "Demographic composition of poll questionnaires across all observations",
         subtitle = paste("Including missing data patterns, `No Data` = NA in the data,\nN =", nrow(ep_raw_clean))) +
    theme_minimal() +
    theme(legend.position = "none")
```

[1] They represent Australia and New Zealand respectively as there were only one voting station per those countries.

[2] Time living outside of Russia was asked as "How long do you live in the country you are now in?"



There are a couple of problematic things emerging from this graph.

The share of 65+ year olds identifying as "Other" gender is higher than the share of 45-64 year olds. On first examination, most of this (125 out of 172) comes from one country - Argentina. There isn't really any reason to think that there are many older Russian emigrants identifying with non-binary genders relative to other ages. Going further and filtering data to see gender only for Argentina and by volunteer ID shows that the abnormal count comes from one volunteer - 8023_2, which must mean that this is an individual misinterpretation of answer categories. I modify the variable so that all "Other gender" answers from this volunteer are NA in the data.

```
kable(table(ep_raw_clean$countryname_en, ep_raw_clean$sex), booktabs = T,
      label = "t1a") |>
  kable_styling(latex_options = c("scale_down", "hold_position"))
```

	Declined to answer	Female	Male	Other
Argentina	0	676	555	125
Armenia	0	1695	2198	10
Australia	0	11	11	0
Austria	0	1845	1157	13
Belgium	0	437	198	4

Canada	0	1020	744	9
Costa Rica	0	55	37	0
Croatia	0	190	121	2
Cyprus	0	3063	2543	57
Czechia	12	805	484	9
Denmark	0	416	200	4
Estonia	0	705	783	20
Finland	0	1486	1122	78
France	0	1885	887	46
Germany	0	2878	2078	7
Great Britain	0	1069	653	4
Greece	0	802	273	1
Hungary	0	843	576	18
Ireland	0	333	295	1
Israel	0	1708	1301	26
Italy	0	1482	320	5
Japan	0	383	257	2
Kazakhstan	0	1118	1555	2
Kyrgyzstan	0	269	248	0
Lithuania	0	314	303	2
Luxembourg	0	536	340	1
Moldova	0	493	377	0
Montenegro	0	752	673	5
Netherlands	0	678	479	10
New Zealand	0	0	0	0
Norway	0	572	231	1
Poland	0	867	674	7
Portugal	0	420	323	10
Serbia	0	1957	1934	15
Slovakia	0	321	292	4
Spain	0	815	516	76
Sweden	0	631	373	5
Switzerland	0	1124	608	7
Thailand	0	285	292	2
Türkiye	0	713	559	0
UAE	0	839	703	1
USA	0	929	599	24
Uzbekistan	0	1144	1301	2
Vietnam	0	118	120	0

```
kable(table(ep_raw_clean$volunteer_id[ep_raw_clean$countryname_en == "Argentina"],
            ep_raw_clean$sex[ep_raw_clean$countryname_en == "Argentina"]), booktabs = T)|>
  kable_styling(latex_options = c("hold_position"))
```

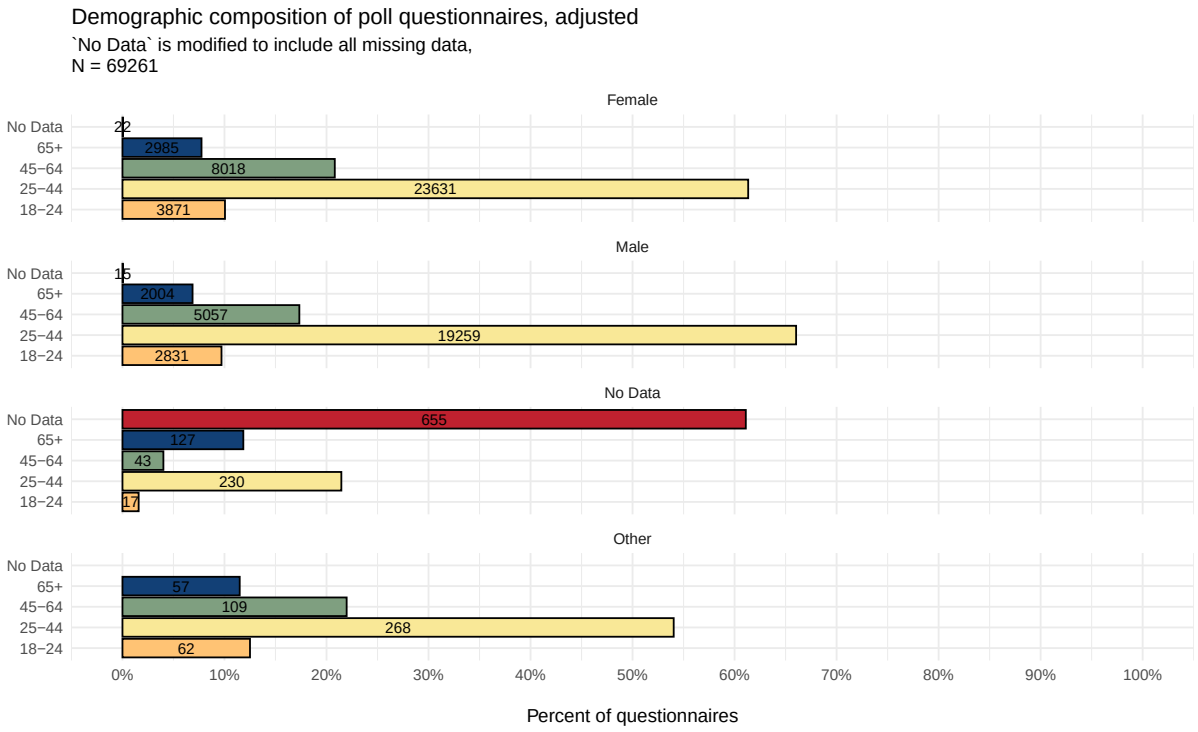
	Female	Male	Other
8023_1	33	26	0
8023_2	155	127	119
8023_3	4	4	0
8023_4	124	124	2
8023_5	171	136	1
8023_6	81	52	1
8023_7	103	82	2
8023_8	5	4	0

‘Other‘ gender in Argentina

From Table Ia it is also clear that only one country actually used the “Declined to answer” category - Czechia for gender. In appendix I also find that this is the case for the age variable. I therefore recode “Declined to answer” as “No Data”, which means we now have one missing data category for both gender and age. The distribution is now as follows:

```
ep_raw_dem <- ep_raw_clean |>
  mutate(across(c(age_bin, sex), ~ if_else(. == "Declined to answer" | is.na(.), "No Data", .)),
         sex = if_else(volunteer_id == "8023_2", "No Data", sex))

ep_raw_dem |>
  group_by(sex, age_bin) |>
  summarize(n = n()) |>
  mutate(pct = n/sum(n)*100) |>
  ggplot(aes(x = pct, y = age_bin, fill = age_bin)) +
    geom_bar(stat = "identity", color = "black") +
    geom_text(aes(label = n), size = 3, position = position_stack(vjust = 0.5)) +
    facet_wrap(~ sex, nrow = 5) +
    scale_fill_manual(values = c("#FFC374", "#F9E897", "#7F9F80",
                                 "#124076", "#bf212f")) +
    scale_x_continuous(limits = c(0, 100),
                       breaks = seq(0, 100, 10),
                       label = scales::label_number(suffix = "%")) +
    labs(x = "\nPercent of questionnaires",
         y = "",
         title = "Demographic composition of poll questionnaires, adjusted",
         subtitle = paste("`No Data` is modified to include all missing data,\nN =", nrow(ep_raw_clean))) +
    theme_minimal() +
    theme(legend.position = "none")
```

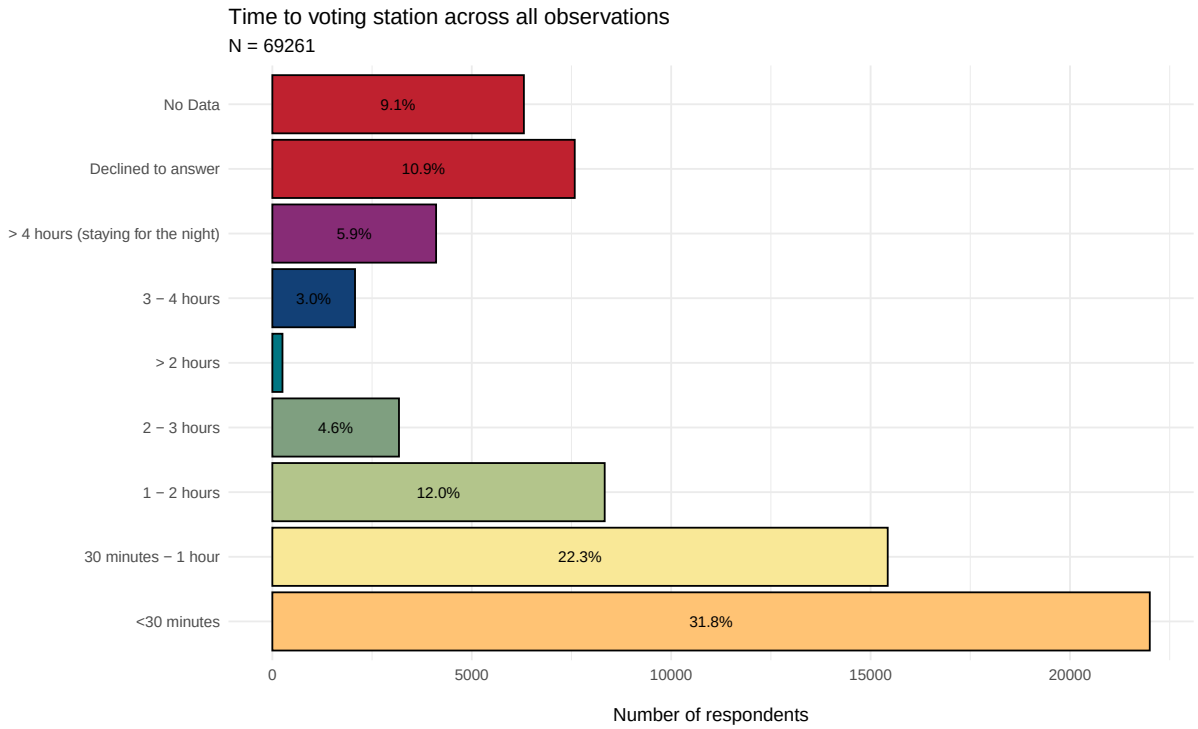


Time to get to the voting station

The "time to the voting station" variable also has inconsistent categories - some volunteers have used a "> 2 hours" category with very low counts, and Czechia had scale that ended in "> 2 hours" and not in "> 4 hours". This means that there are no recorded observations there for "2-3 hours", "3-4 hours" and "> 4 hours" categories.

```
ep_raw_clean |>
  group_by(time_to_vs, .drop = F) |>
  summarise(n = length(time_to_vs)) |>
  mutate(pct = (n/sum(n)),
         lbl = if_else(pct > 0.02, scales::percent(pct), NA),
         time_to_vs = if_else(is.na(time_to_vs), "No Data", time_to_vs),
         time_to_vs = factor(time_to_vs,
                             levels = c("<30 minutes", "30 minutes - 1 hour",
                                        "1 - 2 hours", "2 - 3 hours", "> 2 hours",
                                        "3 - 4 hours", "> 4 hours (staying for the night)",
                                        "Declined to answer", "No Data"))) |>
  ggplot(aes(x = n, y = time_to_vs,
             fill = time_to_vs)) +
    geom_bar(stat = "identity", color = "black") +
    geom_text(aes(label = lbl), size = 3, position = position_stack(vjust = 0.5)) +
    scale_fill_manual(values = c("#FFC374", "#F9E897", "#b3c58b", "#7F9F80", "#007682",
                                 "#124076", "#872c76", "#bf212f", "#bf212f")) +
    labs(x = "\nNumber of respondents",
         y = "",
```

```
        title = "Time to voting station across all observations",
        subtitle = paste("N =", nrow(ep_raw_clean))) +
    theme_minimal() +
    theme(legend.position = "none")
```

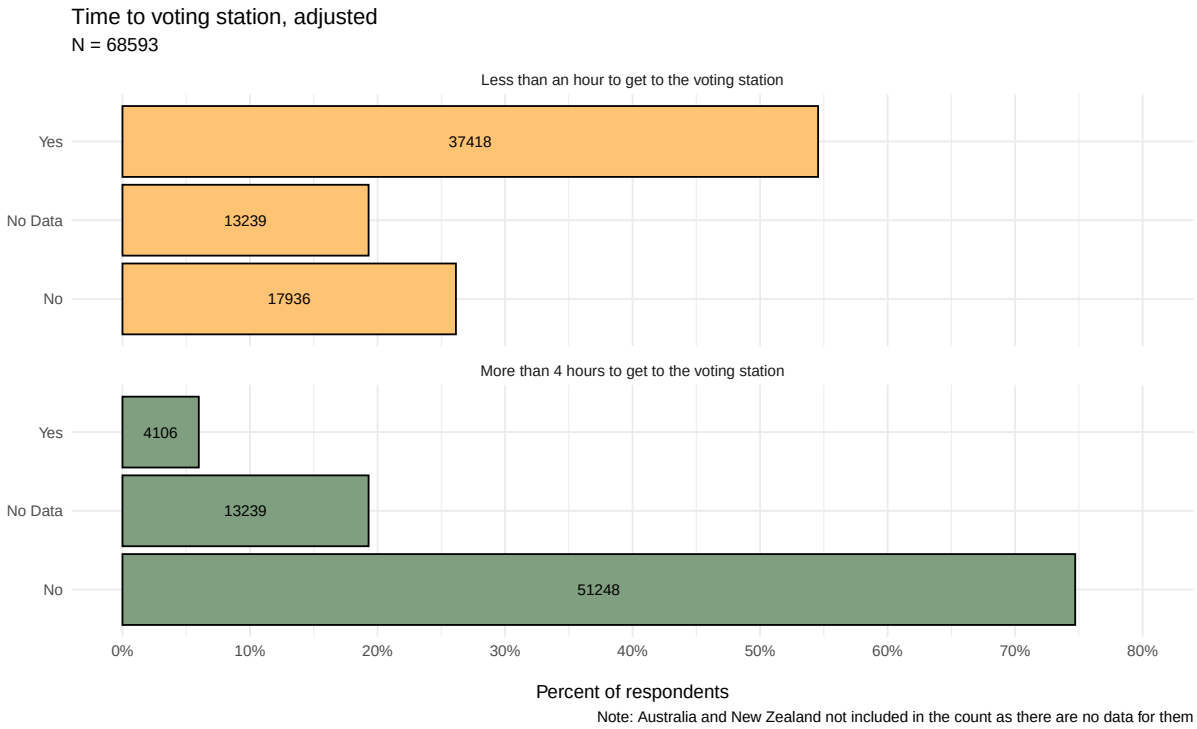


As mentioned before, people could interpret this question differently - some might include standing in line to vote in their estimate of time to the voting station. This would likely make categories in between $<$ 1 hour and $>$ 4 hours unreliable. I therefore think the safest way to still use this variable is to divide it into two dummies, the first indicating if a person spent less than an hour to get to the voting station (as that would mean they are local) and the second indicating if a person spent more than 4 hours to get to the voting station and is planning on staying overnight - which would decidedly indicate they are not local and traveled to vote. For Czechia the second variable would be impossible to build and the country would be dropped from the analysis. I preserve missing values so as not to conflate not wanting to answer with answering differently. As in the demographic variables I assume no data means that a person didn't answer the question.

```
ep_raw_tvs <- ep_raw_dem |>
  mutate(time_to_vs.less_than_hour = case_when(
    time_to_vs %in% c("<30 minutes", "30 minutes - 1 hour") ~ "Yes",
    time_to_vs %in% c("1 - 2 hours", "2 - 3 hours", "> 2 hours",
                      "3 - 4 hours", "> 4 hours (staying for the night)") ~ "No",
    .default = "No Data"),
        time_to_vs.more_than_4hours = case_when(
```

```
    time_to_vs == "> 4 hours (staying for the night)" ~ "Yes",
    time_to_vs %in% c("<30 minutes", "30 minutes - 1 hour",
                      "1 - 2 hours", "2 - 3 hours", "> 2 hours",
                      "3 - 4 hours") ~ "No",
    .default = "No Data"))

ep_raw_tvs |>
  filter(!countryname_en %in% c("Australia", "New Zealand")) |>
  pivot_longer(cols = c(time_to_vs.less_than_hour,
                        time_to_vs.more_than_4hours)) |>
  group_by(name, value) |>
  summarise(n = length(value)) |>
  mutate(pct = n/sum(n)*100,
         name = case_when(name == "time_to_vs.less_than_hour"
                          ~ "Less than an hour to get to the voting station",
                          name == "time_to_vs.more_than_4hours"
                          ~ "More than 4 hours to get to the voting station")) |>
  ggplot(aes(x = pct, y = value,
             fill = name)) +
      geom_bar(stat = "identity", color = "black") +
      geom_text(aes(label = n), size = 3, position = position_stack(vjust = 0.5)) +
      scale_fill_manual(values = c("#FFC374", "#7F9F80")) +
      scale_x_continuous(limits = c(0, 80),
                         breaks = seq(0, 80, 10),
                         label = scales::label_number(suffix = "%")) +
      facet_wrap(~ name, nrow = 2) +
      labs(x = "\nPercent of respondents",
           y = "",
           title = "Time to voting station, adjusted",
           subtitle = paste(
             "N =", nrow(filter(ep_raw_clean,!countryname_en %in% c("Australia",
                                                                    "New Zealand")))
                             ),
           caption =
             "Note: Australia and New Zealand not included in the count as there are no data for them") +
      theme_minimal() +
      theme(legend.position = "none")
```



As those variables are highly correlated, it makes sense to only use them one at a time.

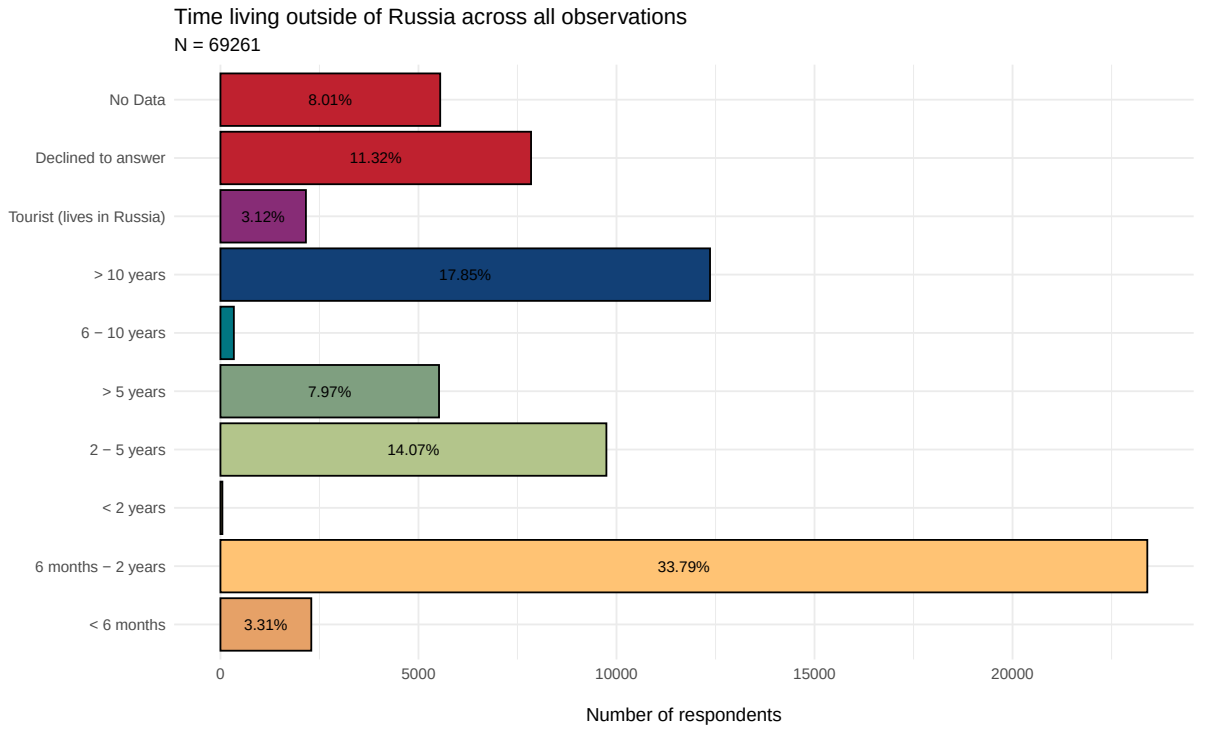
Time out of Russia

The time lived outside of Russia shares the three problematic countries with the rest of the sample, those being Australia and New Zealand with almost no observations and Czechia with a different question - "How long do you live in the country you are now in?" as opposed to "How long have you been living outside of Russia?".

Regardless, I first present the given distribution:

```
ep_raw_clean |>
  group_by(out_of_Russia_time) |>
  summarise(n = n()) |>
  mutate(pct = round(n/sum(n), 4),
         lbl = if_else(pct < 0.01, NA, scales::percent(pct)),
         out_of_Russia_time = if_else(is.na(out_of_Russia_time), "No Data", out_of_Russia_time),
         out_of_Russia_time = factor(out_of_Russia_time,
                                     levels = c("< 6 months", "6 months - 2 years", "< 2 years",
                                                "2 - 5 years", "> 5 years", "6 - 10 years", "> 10 years",
                                                "Tourist (lives in Russia)", "Declined to answer", "No Data"))) |>
  ggplot(aes(x = n, y = out_of_Russia_time,
             fill = out_of_Russia_time)) +
    geom_bar(stat = "identity", color = "black") +
    geom_text(aes(label = lbl), size = 3, position = position_stack(vjust = 0.5)) +
    scale_fill_manual(values = c("#E6A167", "#FFC374", "#F9E897", "#b3c58b", "#7F9F80", "#007682",
```

```
                                "#124076", "#872c76", "#bf212f", "#bf212f")) +
  labs(x = "\nNumber of respondents",
       y = "",
       title = "Time living outside of Russia across all observations",
       subtitle = paste("N =", nrow(ep_raw_clean))) +
  theme_minimal() +
  theme(legend.position = "none")
```



The options "6 - 10 years" and "< 2 years" were used only by Czechia, and the question was different, so it is really hard to justify including this country at all. However, the quantity of interest is not really when the person arrived but rather the timing. I can distinguish between three periods - more than 10 years means that people immigrated before the annexation of Crimea. 2-10 years ago means that people immigrated after the annexation but before the full scale invasion of Ukraine. Lastly, people that immigrated less than 2 years before did so after the start of the invasion. Those thresholds *on average* capture reasons for immigration out of Russia. The expected relationship is that those that came before the annexation of Crimea did not do so for political motives and on average do not necessarily oppose the Russian government. Those that came in the aftermath of the annexation were the first wave of political immigrants, but since 10 years from 2014 is a long time, probably only those that were in immigration from 2014 to 2019 (between 5 and 10 years) did so out of political reasons. The remaining group of those that immigrated between 2019 and 2022 cannot be generalized. Lastly, those that immigrated after 2022 probably did so out of one of two reasons - disagreement with the war or fear for own life in light of mobilization efforts. The threshold for mobilization fear would

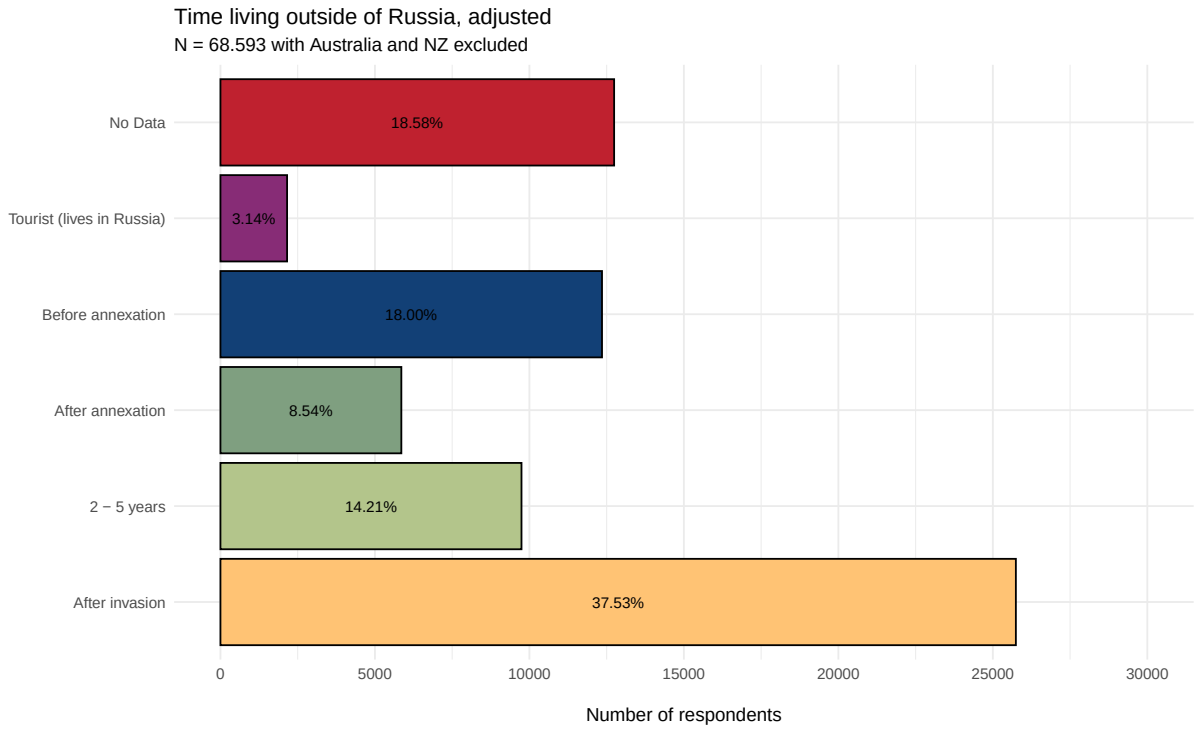
have been 1.5 years ago but there is no detailed data, so I can only hypothesize an interaction effect between being the second wave immigrant and being male.

Note that Czech scale is much more comparable with this definition under the assumption that people moved to the Czech Republic directly and not after immigrating first to another country.

With this variable there is also a need to describe missing data more precisely. Both "Declined to answer" and missing cells are present in the data. Most volunteers used predominantly one or the other, with some of the country-level results being coherent with using one of those two options. However, some volunteers used both with similar counts. In this case one has to ask whether the two options mean the same thing. One option is that when a person left without finishing the questionnaire, blank cells are used, but if verbally declining then "Declined to answer" is used. There is unfortunately no way to tell, yet even this interpretation means that the question was left unanswered. It is highly unlikely that blank cells appear when information was lost or due to some failure on the part of the volunteer. Hence, I combine NA and Declined to answer into one category of missing data.

```
ep_raw_out <- ep_raw_tvs |>
  mutate(out_of_Russia_time = case_when(
    out_of_Russia_time %in% c("< 6 months", "6 months - 2 years", "< 2 years") ~ "After invasion",
    out_of_Russia_time %in% c("> 5 years", "6 - 10 years") ~ "After annexation",
    out_of_Russia_time == "> 10 years" ~ "Before annexation",
    out_of_Russia_time == "Declined to answer" | is.na(out_of_Russia_time) ~ "No Data",
    .default = out_of_Russia_time))

ep_raw_out |>
  filter(!countryname_en %in% c("Australia", "New Zealand")) |>
  group_by(out_of_Russia_time) |>
  summarise(n = n()) |>
  mutate(pct = round(n/sum(n), 4),
         lbl = if_else(pct < 0.01, NA, scales::percent(pct)),
         out_of_Russia_time = factor(out_of_Russia_time,
           levels = c("After invasion", "2 - 5 years", "After annexation",
                      "Before annexation", "Tourist (lives in Russia)", "No Data"))) |>
  ggplot(aes(x = n, y = out_of_Russia_time,
             fill = out_of_Russia_time)) +
    geom_bar(stat = "identity", color = "black") +
    geom_text(aes(label = lbl), size = 3, position = position_stack(vjust = 0.5)) +
    scale_fill_manual(values = c("#FFC374", "#b3c58b", "#7F9F80",
                                 "#124076", "#872c76", "#bf212f", "#bf212f")) +
    scale_x_continuous(limits = c(0, 30000),
                       breaks = seq(0, 30000, 5000)) +
    labs(x = "\nNumber of respondents",
         y = "",
         title = "Time living outside of Russia, adjusted",
         subtitle = "N = 68.593 with Australia and NZ excluded") +
    theme_minimal() +
    theme(legend.position = "none")
```

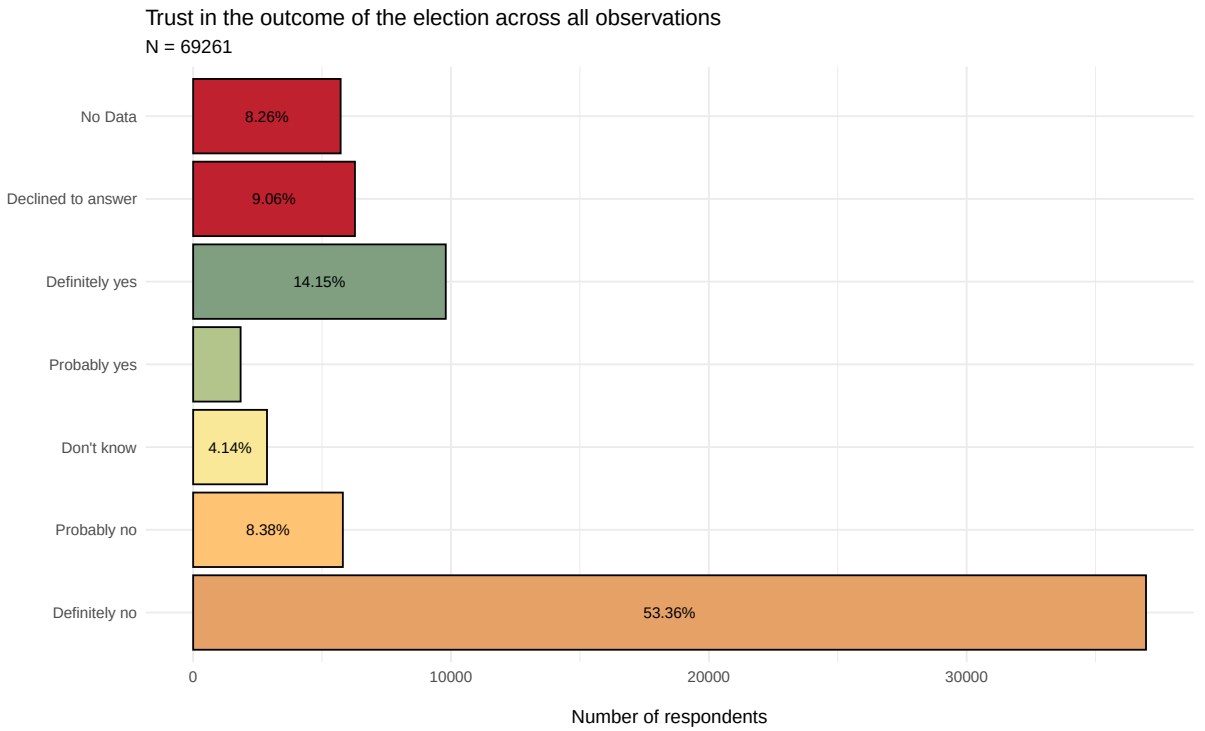


Trust in the outcome

Lastly, and probably the easiest-to-deal-with variable is trust in the outcome of the election. The original scale allows for different levels of confidence within the "yes" and "no" answers. There is no strong reason for this aside from distinguishing between people with strong and less strong feelings. However, the outcomes of election are polarized by default as illustrated by low percentages given to two candidates of the parliamentary opposition.

```
ep_raw_clean |>
  group_by(result_trust) |>
  summarise(n = n()) |>
  mutate(pct = round(n/sum(n), 4),
         lbl = if_else(pct < 0.03, NA, scales::percent(pct)),
         result_trust = if_else(is.na(result_trust), "No Data",
                                result_trust),
         result_trust = factor(result_trust,
                               levels = c("Definitely no", "Probably no",
                                          "Don't know", "Probably yes",
                                          "Definitely yes", "> 10 years",
                                          "Declined to answer", "No Data"))) |>
  ggplot(aes(x = n, y = result_trust,
             fill = result_trust)) +
    geom_bar(stat = "identity", color = "black") +
    geom_text(aes(label = lbl), size = 3, position = position_stack(vjust = 0.5)) +
    scale_fill_manual(values = c("#E6A167", "#FFC374", "#F9E897", "#b3c58b", "#7F9F80",
                                 "#bf212f", "#bf212f")) +
```

```
    labs(x = "\nNumber of respondents",
         y = "",
         title = "Trust in the outcome of the election across all observations",
         subtitle = paste("N =", nrow(ep_raw_clean))) +
    theme_minimal() +
    theme(legend.position = "none")
```

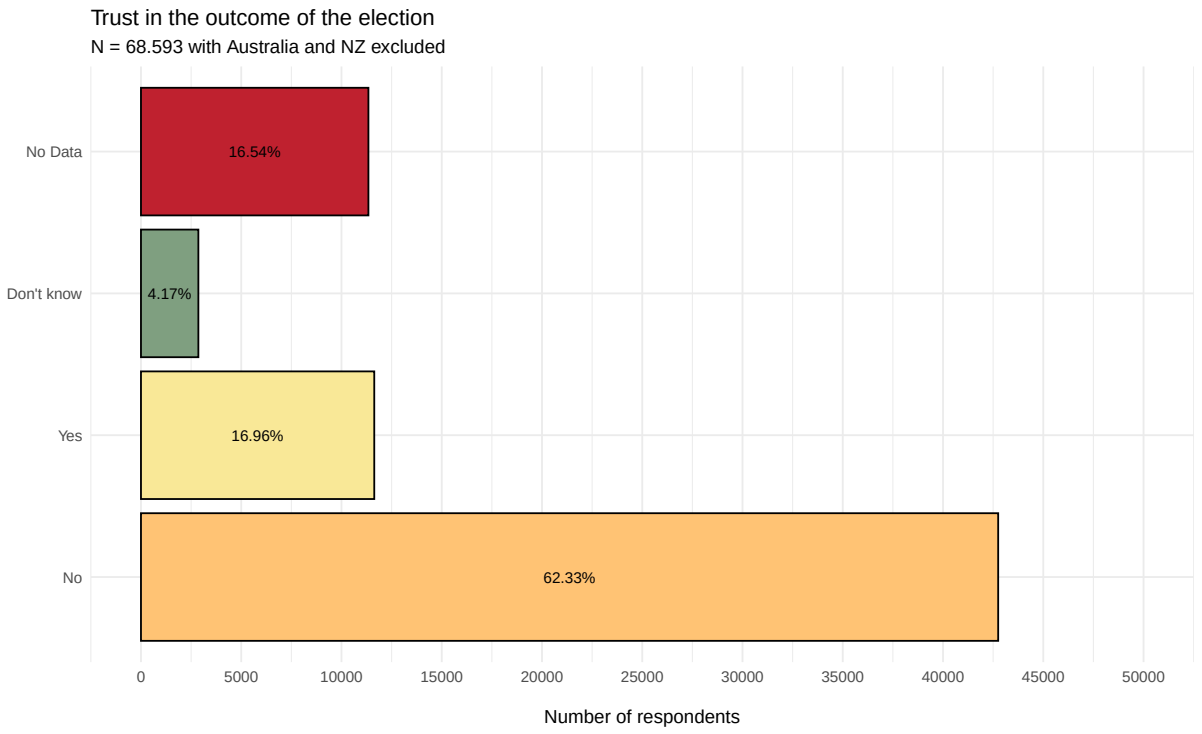


I collapse the two probably/definitely categories to get three categories: yes, no, don't know and declined to answer or NA.

```
ep_raw_trst <- ep_raw_out |>
  mutate(result_trust_bin = case_when(result_trust %in% c("Probably no", "Definitely no") ~ "No",
                                      result_trust %in% c("Probably yes", "Definitely yes") ~ "Yes",
                                      result_trust == "Declined to answer" | is.na(result_trust) ~ "No Data",
                                      .default = result_trust))

ep_raw_trst |>
  filter(!countryname_en %in% c("Australia", "New Zealand")) |>
  group_by(result_trust_bin) |>
  summarise(n = n()) |>
  mutate(pct = round(n/sum(n), 4),
         lbl = if_else(pct < 0.01, NA, scales::percent(pct)),
         result_trust_bin = factor(result_trust_bin,
                                   levels = c("No", "Yes", "Don't know", "No Data"))) |>
  ggplot(aes(x = n, y = result_trust_bin,
             fill = result_trust_bin)) +
    geom_bar(stat = "identity", color = "black") +
    geom_text(aes(label = lbl), size = 3, position = position_stack(vjust = 0.5)) +
```

```
    scale_fill_manual(values = c("#FFC374", "#F9E897", "#7F9F80",
                                 "#bf212f", "#bf212f")) +
    scale_x_continuous(limits = c(0, 50000),
                       breaks = seq(0, 50000, 5000)) +
    labs(x = "\nNumber of respondents",
         y = "",
         title = "Trust in the outcome of the election",
         subtitle = "N = 68.593 with Australia and NZ excluded") +
    theme_minimal() +
    theme(legend.position = "none")
```



Preparing the dependent variables

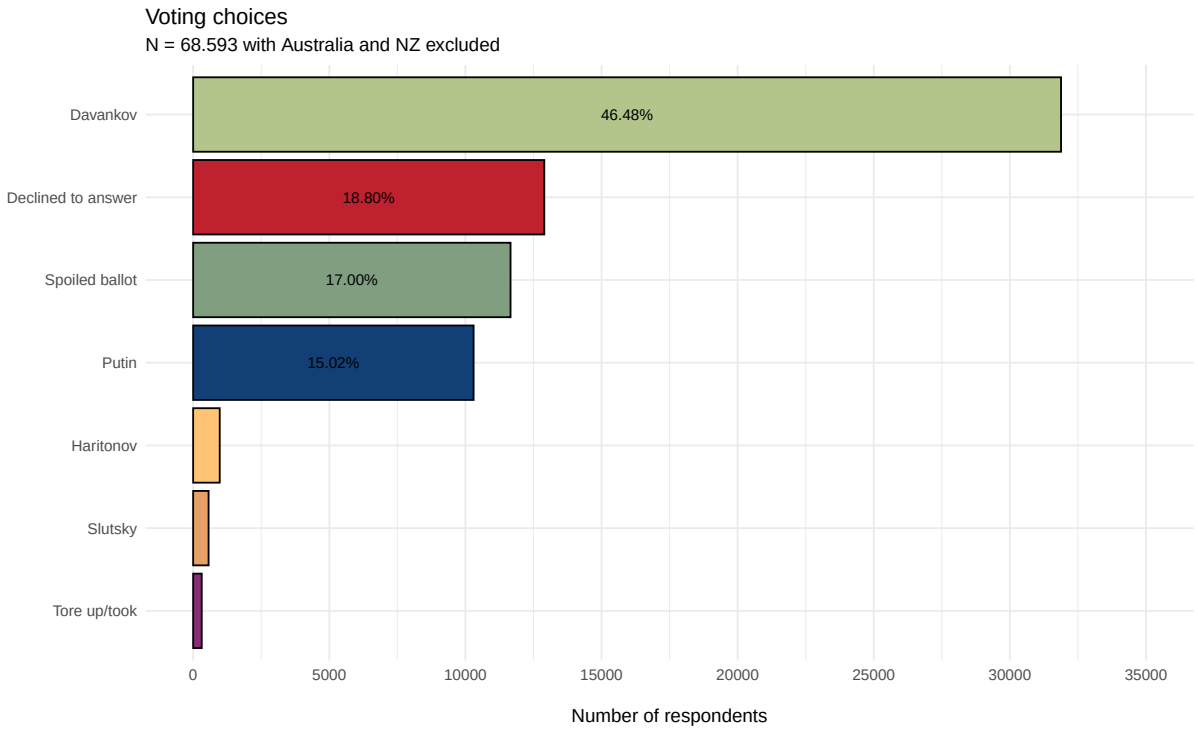
Models that can be used to model individual choice need different operationalizations of dependent variables to work. Those are: each candidate and specific combinations against everything else; multinomial choice; dichotomized variables for nested logit models. I create variables for the first and second approaches; nested logit package used handles creation of its variables by itself.

```
ep_raw_dep <- ep_raw_trst |>
  mutate(vote_putin = if_else(vote == "Putin", 1, 0),
         vote_davankov = if_else(vote == "Davankov", 1, 0),
         vote_spoiled = if_else(vote == "Spoiled ballot", 1, 0),
         vote_opposition = if_else(
```

```
          vote %in% c("Davankov", "Spoiled ballot"), 1, 0),
        vote_declined = if_else(vote == "Declined to answer", 1, 0),
        vote_putin_declined = if_else(
          vote %in% c("Putin", "Declined to answer"), 1, 0),
        vote = factor(vote))

results1 <- ep_raw_dep |>
  filter(!countryname_en %in% c("Australia", "New Zealand")) |>
  group_by(vote) |>
  summarise(n = n()) |>
  mutate(pct = round(n/sum(n), 4),
        lbl = if_else(pct < 0.02, NA, scales::percent(pct)),
        result_trust_bin = factor(vote,
                                  levels = c("Putin", "Davankov",
                                             "Spoiled ballot", "Slutsky",
                                             "Haritonov", "Tore up/took",
                                             "Declined to answer"))) |>
  ggplot(aes(x = n, y = reorder(vote, n),
             fill = vote)) +
    geom_bar(stat = "identity", color = "black") +
    geom_text(aes(label = lbl), size = 3, position = position_stack(vjust = 0.5)) +
    scale_fill_manual(values = c("#b3c58b", "#bf212f",
                                 "#FFC374","#124076", "#E6A167", "#7F9F80","#872c76")) +
    scale_x_continuous(limits = c(0, 35000),
                       breaks = seq(0, 35000, 5000)) +
    labs(x = "\nNumber of respondents",
         y = "",
         title = "Voting choices",
         subtitle = "N = 68.593 with Australia and NZ excluded") +
    theme_minimal() +
    theme(legend.position = "none")

results1
```




```
# Save data

  ## Save the final dataset
  write_rds(ep_raw_dep, here("data", "data_built", "ep_raw_dep.rds"))
```