# Problem Set 03

**AQMSS II**

Polikanov Stepan

## Review of Key Theoretical Concepts

### Question 1: Missing Data

#### a

Censoring on the dependent variable means that there is a cutoff point after which the data is not observed directly (for example disease progression after death or population count under a certain threshold). On the other hand truncation occurs when some values on either side of the distribution are not included because of their specificity. For example if one records characteristics of days on which there was rain, then all characteristics of days when it wasn't raining would be truncated. In other words the difference is between whether we do not know the exact ommited value (censoring) or anything beyond some arbitrary threshold (truncation).

In linear regressions both censoring and truncation will lead to biased and inconsistent estimates. Moreover, type I errors might be more prevalent. There are model specifications designed to deal with either issues though, namely tobit models and truncated models. The former dichotomizes the data and assigns a constant to the censored bit and the latter relies on finding a threshold based on theory to correctly assign missing data label, which can then be modelled explicitly.

Since we would be dealing with different phenomena no fix is better than the other but in any case censored or truncated, the pattern should be addressed.

#### b

Endogenous selection on the dependent means that the data observed in the dependent variable is determined by some (un)observable process. The type of process and its relation to the predictors in the model dictates the missingness pattern in other variables. If selection is separate from the model then the pattern is MCAR, if it can be predicted with the variables in the model, then it is MAR and if both are driven by some unobservables then the pattern is Non-Ignorable (NI).

## Question 2: i.i.d

Drawing from Economic History, violations of i.i.d can be seen most clearly in the logic of path dependence. We often assume municipalities or other units are more or less homogeneous and thus can be treated as i.i.d. However (and this is amplified by the fact that the field likes to deal with population-level data and not samples from such) if we cannot isolate exogenous variation (which is what everyone is doing) or control for all sources of historical advantage (or other difference) between units (which is a daunting task considering that some advantages stem from selection processes), the assumption would be violated and one might miss heterogeneous effects. In turn this means that any effect would be blown out of proportion i.e policies having a more positive effect in higher human capital areas.

Independence is also an issue within the context of such research and the issue is again amplified by using all of the available data. It is nice when we can do matching (feasible in very specific designs) and ensure that there are no spillover effects or when spatial autocorrelation corrections (a cheap and popular solution but when data is (as it so often is) incomplete may be harder to implement) work but when either is not possible the units are almost surely dependent on their surroundings, violating the second part of the assumption.

# Working with Missing and Panel data (Why do I do this to myself)

```r
## Required packages
packages <- c("here", "kableExtra", "stargazer", "nnet",
              "nestedLogit", "sampleSelection", "mice", "miceadds", "micemd", "summarytools",
              "pan", "multilevel", "tidyverse", "modelsummary",
              "readxl", "here", "lubridate",  "countrycode")

## Install packages not yet installed
installed_packages <- packages %in% rownames(installed.packages())
  if (any(installed_packages == FALSE)) {
    install.packages(packages[!installed_packages])
  }

## Load package
invisible(lapply(packages, library, character.only = TRUE))

# Define global functions

## Fitting stargazer onto the page
resizebox.stargazer = function(..., tab.width = "!", tab.height = "!")
    {
  require(stringr)
  res = capture.output(stargazer::stargazer(...))
  tab.width = tab.width
  tab.height = tab.height
  res = prepend(res, "}", before = length(res))

  res = c(res[1:str_which(res, "^\\\\begin\\{tabular\\}.*")-1],
      paste0("\\resizebox*{",tab.width,"}{",tab.height,"}{%"),
      res[str_which(res, "^\\\\begin\\{tabular\\}.*"):length(res)]
      )
  cat(res, sep = "\n")
}
```

```r
options(scipen = 999)
```

```r
ep_raw <- read_excel(here("data", "exitpoll_rawdata.xlsx"),
                     sheet = 4, guess_max = 69262)
```

I first perform mechanical cleaning and translation operations as well as some very minor data homogenization (equalizing capitalized and non-capitalized exit poll items).

```r
ep_raw_clean <- ep_raw |>
  separate(col = "    ", into = c("countryname_ru", "countryname_en"),
           sep = " / ") |>
  separate(col = "   ", into = c("city_ru", "city_en"),
           sep = " / ") |>
  mutate(countryname_en = if_else(countryname_en == "Czech",
                                  "Czechia", countryname_en),
         countrycode_n = countrycode(countryname_en,
                                     origin = "country.name",
                                     destination = "iso3n"),
         countrycode_c = countrycode(countrycode_n,
```

```r
                                    origin = "iso3n",
                                    destination = "iso3c"),
        voting_station = as.character(`  `),
        city_en = if_else(str_detect(city_ru, "№") == T,
                          paste(city_en, str_sub(city_ru, -1,
                                                  nchar(city_ru))),
                          city_en)) |>
transmute(volunteer_id = `ID      `,
          data_source = case_when(
          `         ` == "                         QR    "
          ~ "One-off QR e-form",
          `         ` == "            " ~ "General e-form",
          `         ` == "           -    " ~ "Tokio e-form",
          `         ` == "           -      " ~ "Stockholm e-form",
          `         ` == "           -    " ~ "Prague e-form",
          `         ` == "           -    " ~ "Dubai e-form",
          `         ` == "           -      " ~ "Wellington e-form",
          `         ` == "           -     " ~ "Sydney e-form"),
          upload_time = ymd_hms(`         `),
          countryname_en, countryname_ru, city_ru, city_en, countrycode_c, countrycode_n,
          voting_station,
          vote = case_when(
            `                     ?` == " .  .      " ~ "Davankov",
            `                     ?` == " . .     " ~ "Slutsky",
            `                     ?`
            == "                (          ,          )"
            ~ "Spoiled ballot",
            `                     ?` == "         /    /    "
            ~ "Tore up/took",
            `                     ?` == " . .    " ~ "Putin",
            `                     ?` == "          "
            ~ "Declined to answer",
            `                     ?` == " . .      " ~ "Haritonov"
            ),
          sex = case_when(`    ` == "    " ~ "Male",
                          `    ` == "    " ~ "Female",
                          `    ` == "    " ~ "Other",
                          `    ` == "          " ~ "Declined to answer",
                          .default = NA),
          age_bin = case_when(
            `          ` == "          " ~ "Declined to answer",
            .default = `          `),
          out_of_Russia_time = case_when(
            `                   ?` == "    6     "~ "< 6 months",
            `                   ?` == "6      - 2    (         2022)"
            ~ "6 months - 2 years",
            `                   ?` == "    2   " ~ "< 2 years",
            `                   ?` == "2 - 5   " ~ "2 - 5 years",
            `                   ?` == "    5   " ~ "> 5 years",
            `                   ?` == "6-10   "~ "6 - 10 years",
            `                   ?` == "    10   " ~ "> 10 years",
            `                   ?` == "          "
            ~ "Declined to answer",
            `                   ?` == "    /  (      )"
            ~ "Tourist (lives in Russia)", .default = NA),
          time_to_vs = case_when(
            `                            ?` == "<30    "
            ~ "<30 minutes" ,
            `                            ?` == "30    - 1   "
            ~ "30 minutes - 1 hour" ,
```

```
  `                             ?` == "1 - 2   "
  ~ "1 - 2 hours",
  `                             ?` == "2 - 3   "
  ~ "2 - 3 hours",
  `                             ?` == "3 - 4   "
  ~ "3 - 4 hours",
  `                             ?` == "    2   "
  ~ "> 2 hours" ,
  `                             ?`
  == "    4    (                  )"
  ~ "> 4 hours (staying for the night)",
  `                             ?` == "          "
  ~ "Declined to answer", .default = NA),
result_trust = case_when(
  `                        ?` %in% c("     ", "     ")
  ~ "Definitely no",
  `                        ?` == "          "
  ~ "Probably no",
  `                        ?` == "        "
  ~ "Probably yes",
  `                        ?` %in% c("    ", "     ")
  ~ "Definitely yes",
  `                        ?` == "           "
  ~ "Don't know",
  `                        ?` == "          "
  ~ "Declined to answer", .default = NA),
uik_closetime = `   -               (CET)`,
uik_nvoters = `   -               `,
uik_didntvote_closed = `   -                         `,
id = row_number())
```

## 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 Commitee 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.

### Hypotheses

There are a number of interesting things to explore, differing in their intuitiveness. Firstly, I am interested in the demographic composition of incumbent electorate. It is documented that in Russia women vote for Putin more and older demographics also rarely engage with opposition. I wonder how that translates into eclectic russian diasporas abroad, especially considering that the exit poll was conducted selectively in countries where there were enough volunteers to do so. Gender and age are the target for the hypothesis one.

$H_1$: Older demographics and women were more likely to vote for the incumbent (or decline to answer)

$H_{1a}$: If voting for the oppositon those categories would prefer to vote for systemic (Slutsky or Haritonov) rather than non-systemic options. [1]

$H_{1b}$: People identifying as "Other" genders are more likely to vote for non-systemic oppostion and more likely to spoil the ballots rather than vote for Davankov.

Another exciting variable is the "time living outside of Russia". It is constructed in such a way that one can distinguish between theoretically interesting categories - namely those that left before 2014 annexation of Crimea, after 2014 but before 2019, between 2019 and 2022 and after the full-scale invasion in February 2022. The corresponding hypotheses are:

$H_{2a}$: Those that left before 2014 are the most likely group to not emigrate for political reasons and are expected to have the highest support for the incumbent

$H_{2b}$: Those that left after 2014 should exhibit more non-systemic opposition views and support appropriate options

$H_{2c}$: Those that left after the start of the full-scale invasion should exhibit more non-systemic opposition views and support appropriate options

It is hard to say what to expect from the remaining group of emigrants 2019-2022. On one hand, they have high temporal distance from the Crimea annexation and couldn't have predicted the war. On the other, if $H_1$ is confirmed they are not in the typical demographic of incumbent supporters, begging the question "Why emigrate?".

### Variables

The data has a challenging but not overwhelming levels of data missingness. Age and gender variables have the least missing values while other items show around 10 percent of NAs and Declined to answer.

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, [2] Prague has multiple deviations from the questionnaire and one question was reformulated. [3]. Moreover, I find at least one case of a volunteer misinterpreting items.

[1]Davankov is also a "systemic" option de facto, yet there is overwhelming evidence suggesting that he was the candidate of choice for the opposition. Overall he received 3.85% of votes, in the sample of voting stations where the exit poll was conducted the official result is 38% and the exit poll estimate is 46% of votes (those can be treated as lower and upper bounds).

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

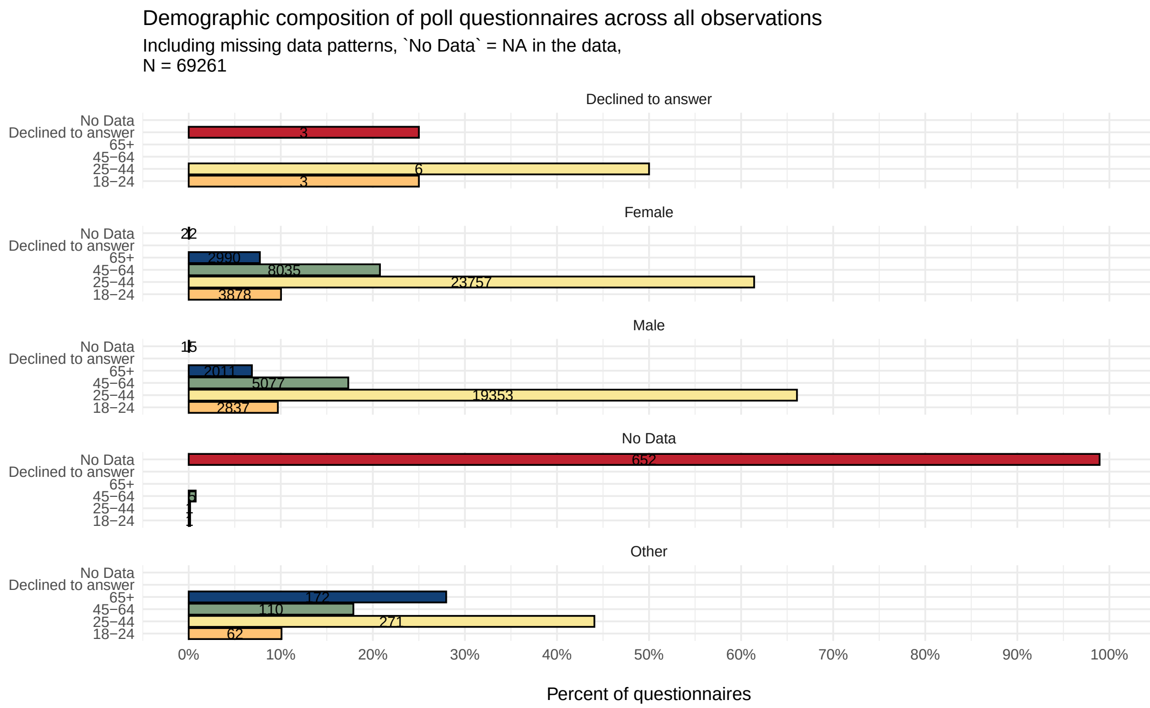
[3]Time living outside of Russia was asked as "How long do you live in the country you are now in?"

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.

```r
# 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")
```




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.

```
# Oh my god pdf with quarto table options is such a mess
# Reminder to self: see if this was reported on quarto github

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

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:

Table 1: `Other` gender in Argentina

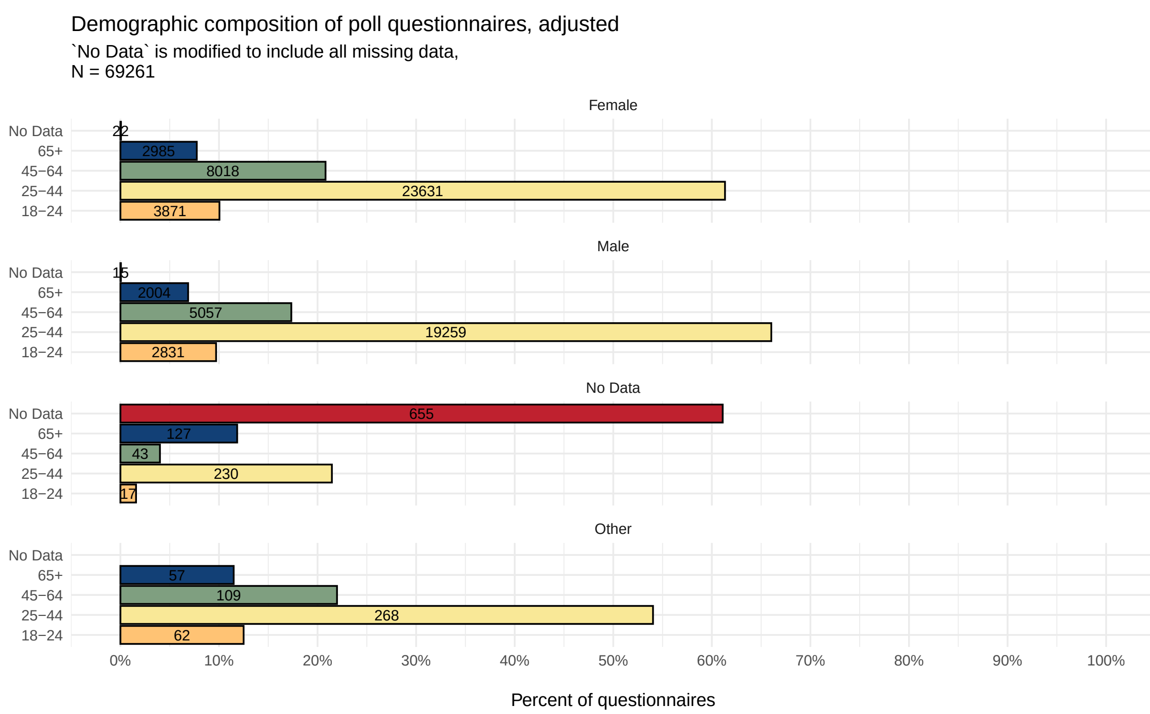
(a)

| | 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 |

(b)

| | 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 |

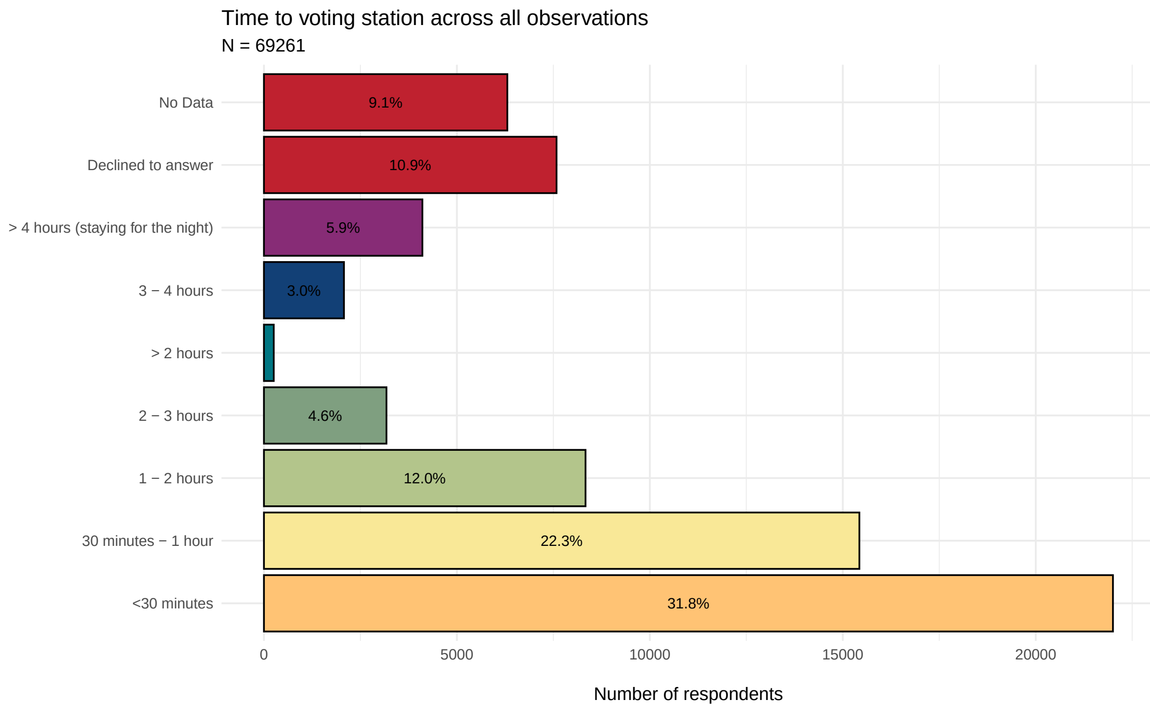
```r
ep_raw_dem <- ep_raw_clean |>
  mutate(across(c(age_bin, sex), ~ if_else(. == "Declined to answer" | is.na(.), "No Data", .)),
         sex = if_else(volunteer_id == "8023_2", "No Data", sex))

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")
```

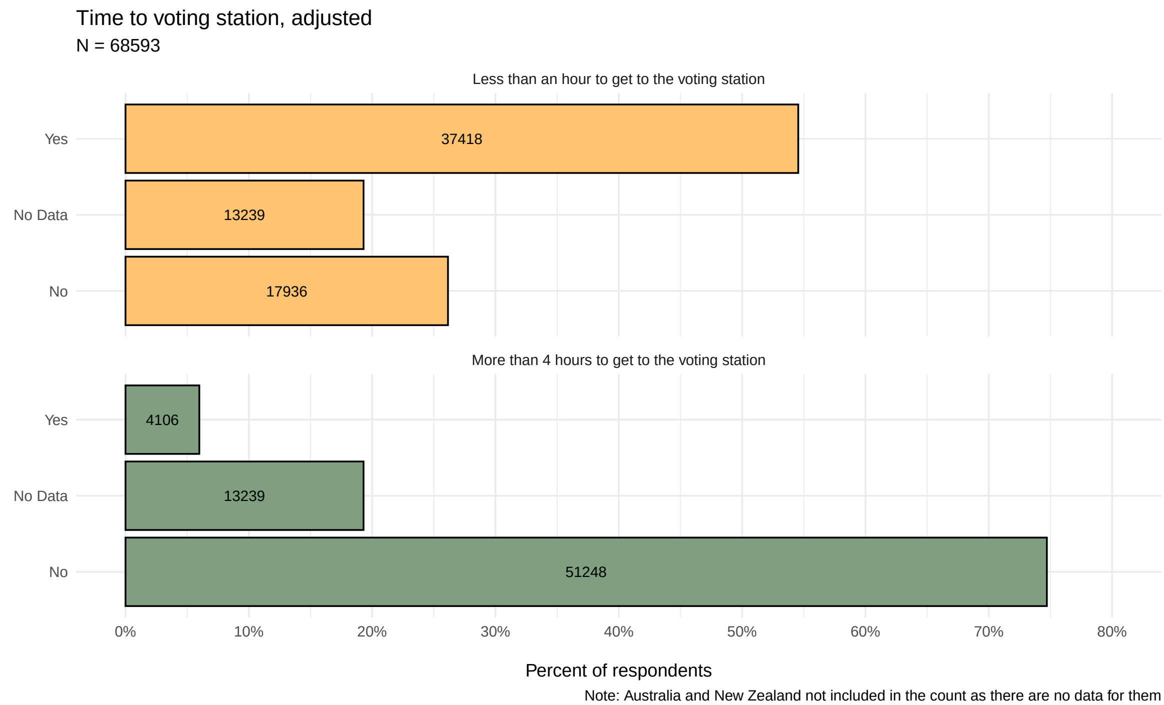



As mentioned before, people could interpret this question differently - some might include standing in line to vote in their estimate of time to the voting station. This would likely make categories in between $< 1$ hour and $> 4$ hours unreliable. I therefore think the safest way to still use this variable 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.

An alternative approach is to attribute differences in interpretations to volunteer- or city-level factors. This provides a possibility to factor in an important caveat - travel distances differ almost surely by city or by country. In Kazakhstan, a large land of steppes, distances are much larger than, say, in the island of Cyprus and both are different from Greece in mode of transportation needed to reach the voting station. Therefore the coefficient should be allowed to vary by either country or city. [4]

[4] I ended up not using this approach as models became increasingly hard to estimate and between-level interactions would have been too costly computationally. I leave this for the final project.

```r
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.

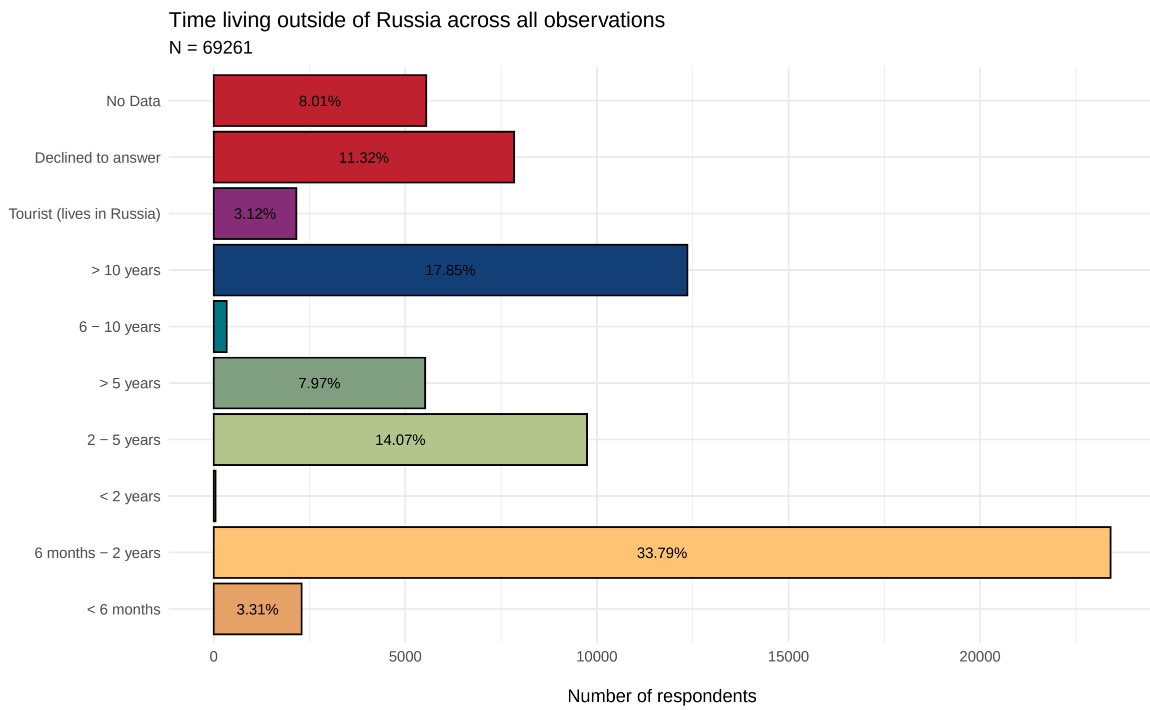
### Time out of Russia

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)) +
```

```r
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")
```

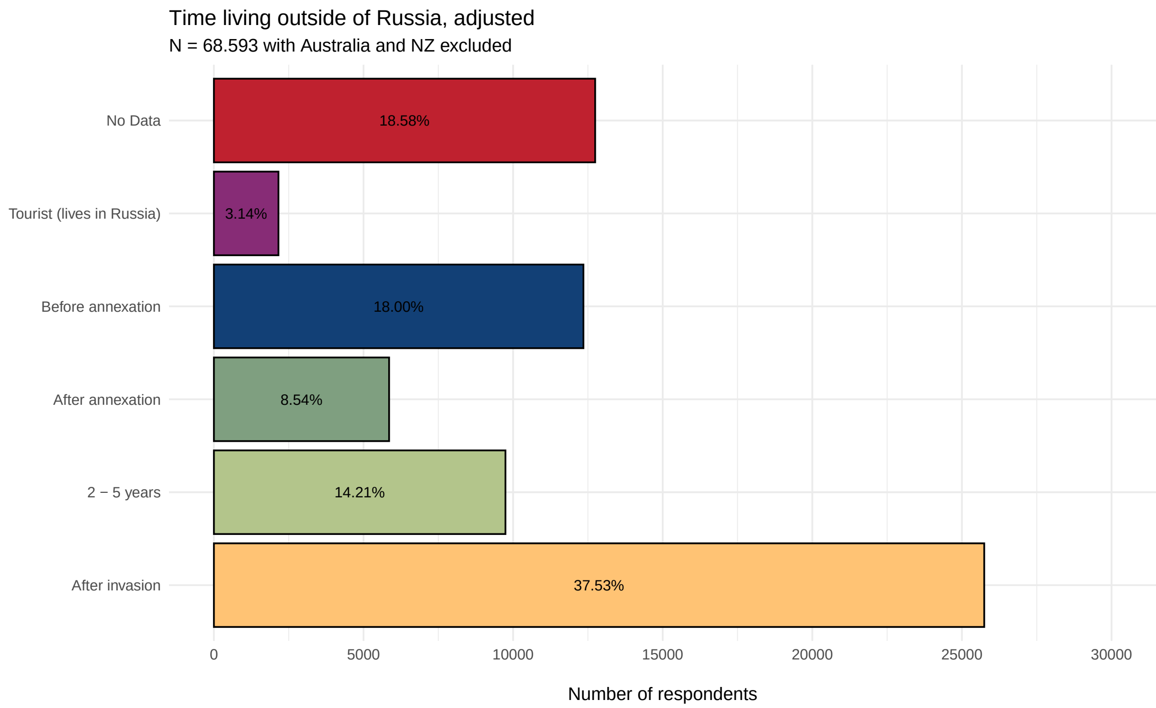



The options "6 - 10 years" and "< 2 years" were used only by Czechia, and the question was different, so it is really hard to justify including this country at all. However, the quantity of interest is not really when the person arrived but rather the timing. I can distinguish between three periods - 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")
```

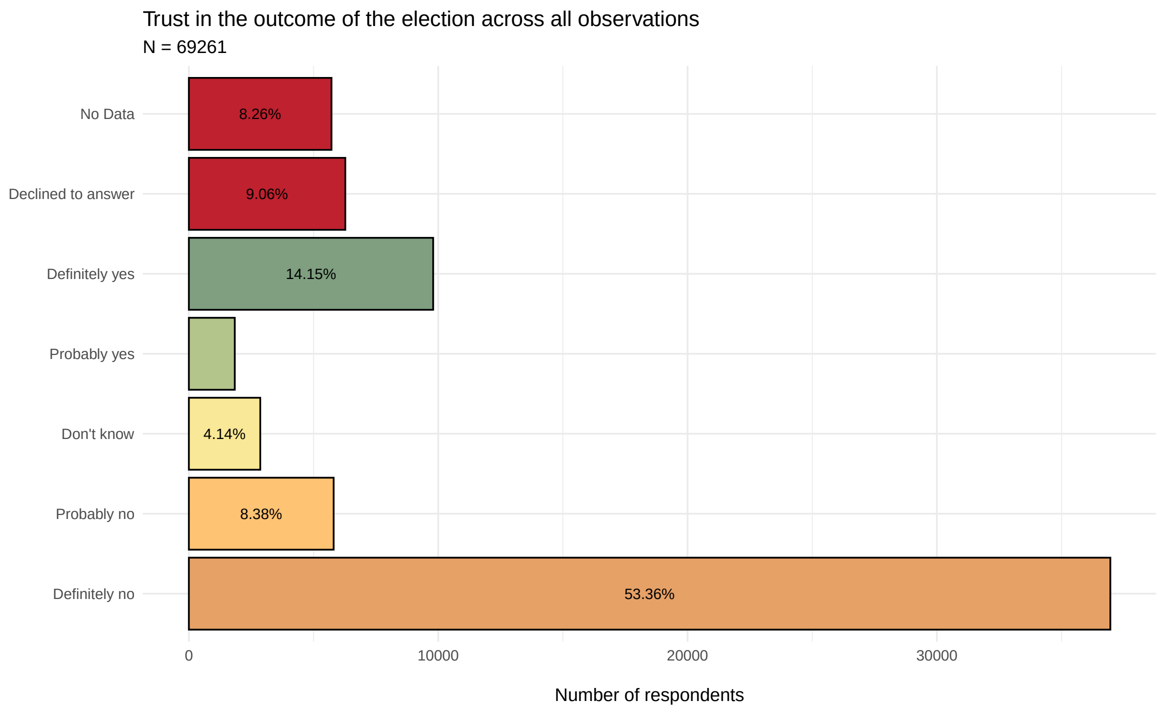



## Trust in the outcome

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")) +
```

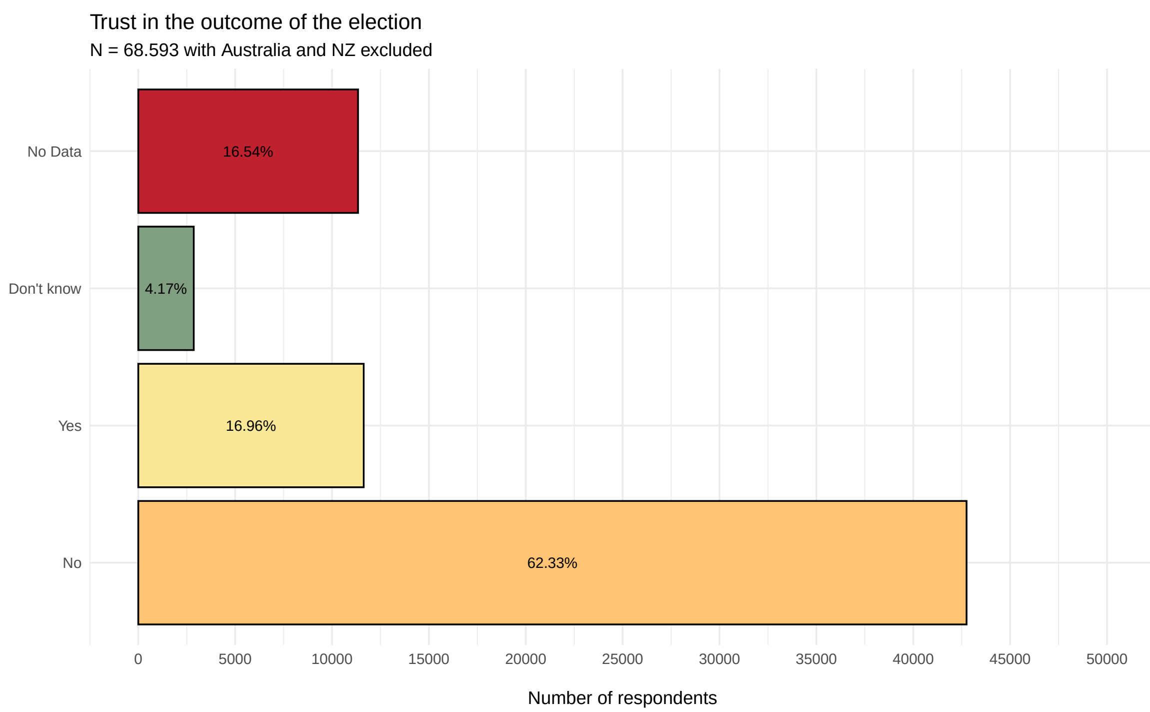
```r
  labs(x = "\nNumber of respondents",
       y = "",
       title = "Trust in the outcome of the election across all observations",
       subtitle = paste("N =", nrow(ep_raw_clean))) +
  theme_minimal() +
  theme(legend.position = "none")
```




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

```r
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") +
```

```r
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

The choice of candidate can be presented in multiple ways. First is to look at the binary choice between a candidate of interest, for example Putin as opposed to all other alternatives. The second one is to model everything at once using multinomial models.

The latter choice seems to be more reasonable, however there are caveats. First, by Arrow's theorem no non-dictatorial Pareto-efficient voting system with more than two outcomes can satisfy the independence of irrelevant alternatives. Secondly, this election is special - ideological salience of candidates is very much blurred and the choice mostly boils down to voting for or against the incumbent. The most-chosen candidate abroad, Davankov, by any available metric is not the actual choice of the opposition voters, but the least unattractive option in the what is essentially an election without alternatives or a "managed" election. This is supported by political weakness of contender candidates, their lack of campaigning and subsequent lack of support.

The points above lead me to believe that the choice for any voter abroad was actually nested. She first decides based on her ideological preferences whether to vote for the incumbent or not. Then, if she decides to vote yes, she does so, but if not she faces two choices - to vote for Davankov or to spoil the ballot by selecting multiple candidates. [5]

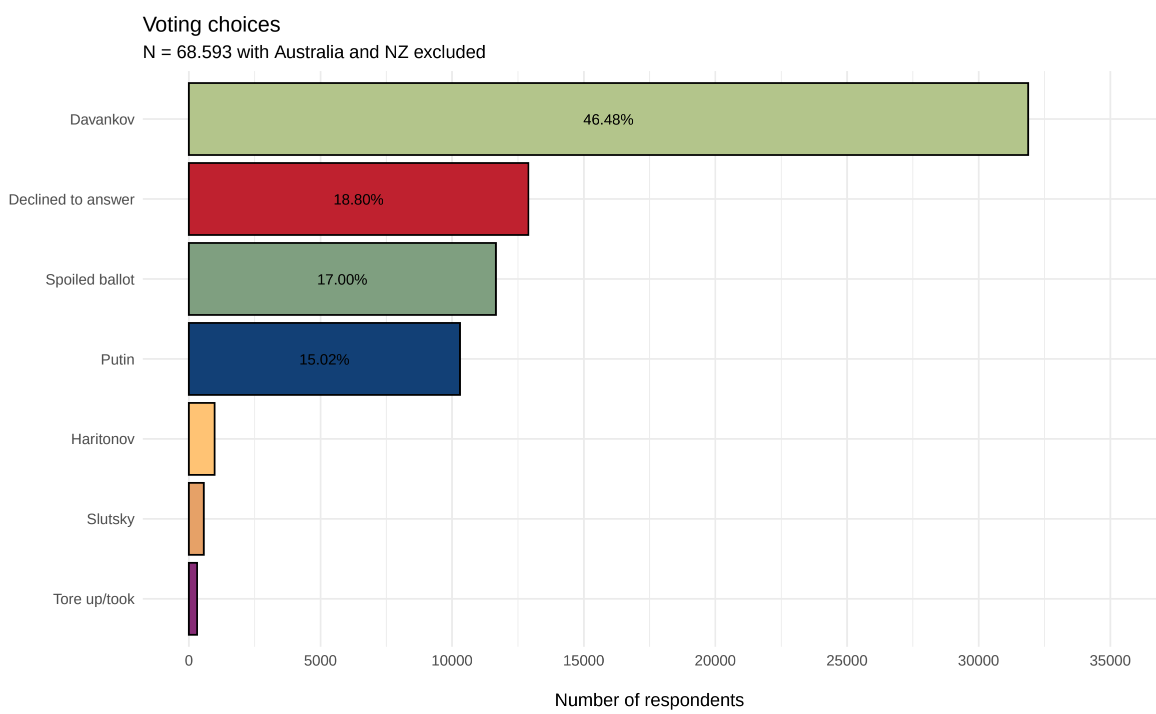
The question of two other candidates on the ballot remains. It could be that people genuinely support them and their platforms, however weak, and then the first choice is actually across four alternatives. On the other hand, the exact wording of many opposition figures (including, for example Navalny's Anti Corruption Foundation) called to vote for "any candidate except Putin", meaning that people might have voted for them simply because they oppose the incumbent. The latter makes the second-stage choice of the opposition voter across four options.

With three estimation options outlined: binary choice, multinomial and nested logit models I build the appropriate variables and present their distributions.

```
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),
         nest1_putin = vote_putin,
         nest1_nonsystemic = case_when(
           vote %in% c("Davankov", "Spoiled ballot") ~ 1,
           vote %in% c("Slutsky", "Haritonov") ~ 0,
           .default = NA),
         nest1_opposition = case_when(vote == "Davankov" ~ 1,
                                      vote == "Spoiled ballot" ~ 0,
                                      .default = NA),
         nest1_sysopposition = case_when(vote == "Slutsky" ~ 1,
                                         vote == "Haritonov" ~ 0,
                                         .default = NA))

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",
```

[5] In Russian election law this makes "the voter's will unidentifiable", meaning the vote doesn't go to any of the candidates but is still used in the calculations for turnout, essentially lowering everyone's result. It was proposed as an alternative to the "against all" option that was removed in 2006. However there is no threshold for spoiled votes after which the result of the election is deemed invalid.

```r
                                                  "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")
```




## Estimating Naive Regressions

The naive estimation strategies is pretty straightforward - perform list-wise NA deletion on all variables of interest and ignore the nested structure of the data. I model “Declined to answer” category for the outcome (which is the only missing data there is for the outcome) explicitly in binary models as a baseline but drop the category in multinomial and nested logit model.

```r
# Recode no data back to native NA
data_naive <- ep_raw_dep |>
  mutate(across(c(sex, age_bin, time_to_vs.less_than_hour,
                time_to_vs.less_than_hour, out_of_Russia_time,
                result_trust_bin, vote),
               ~ if_else(. %in% c("No Data", "Declined to answer"), NA, .)),
         vote = relevel(as.factor(vote), ref = "Putin"),
         sex = relevel(as.factor(sex), ref = "Male"),
         age_bin = relevel(as.factor(age_bin), ref = "25-44"),
         out_of_Russia_time = relevel(as.factor(out_of_Russia_time), ref = "Before annexation"),
         result_trust_bin = relevel(as.factor(result_trust_bin), ref = "Yes"))


m1a.naive <- lm(vote_putin ~ sex + age_bin + time_to_vs.less_than_hour
               + out_of_Russia_time + result_trust_bin, data = data_naive)

m1b.naive <- lm(vote_declined ~ sex + age_bin + time_to_vs.less_than_hour
               + out_of_Russia_time + result_trust_bin,
               data = data_naive)

m1c.naive <- lm(vote_putin_declined ~ sex + age_bin + time_to_vs.less_than_hour
               + out_of_Russia_time + result_trust_bin,
               data = data_naive)


m1d.naive <- lm(vote_davankov ~ sex + age_bin + time_to_vs.less_than_hour
               + out_of_Russia_time + result_trust_bin,
               data = data_naive)

m1e.naive <- lm(vote_spoiled ~ sex + age_bin + time_to_vs.less_than_hour
               + out_of_Russia_time + result_trust_bin,
               data = data_naive)

m1f.naive <- lm(vote_opposition ~ sex + age_bin + time_to_vs.less_than_hour
               + out_of_Russia_time + result_trust_bin,
               data = data_naive)

resizebox.stargazer(m1a.naive, m1b.naive, m1c.naive, m1d.naive, m1e.naive, m1f.naive,
          title = "Binary outcomes, naive approach", header = F,
          dep.var.labels = c("Vote Putin", "Decline to Answer",
                             "Putin or Declined","Vote Davankov", "Spoil the ballot",
                             "Vote Davankov or spoil"),
          covariate.labels = c(
            "Sex: Female", "Sex: Other", "Age: 18-24 (ref 25-44)",
            "Age: 45-65 (ref 25-44)", "Age: 65 + (ref 25-44)",
            "Took < 1 hour to get to the voting station",
            "Moved after March 2022 (ref before 2014)",
            "Moved after March 2019 but before March 2022 (ref before 2014)",
            "Moved after March 2014 but before March 2019 (ref before 2014)",
            "Didn't move - tourist, lives in Russia (ref before 2014)",
            "Trust in the result: Don't know (ref Yes)",
            "Trust in the result: No (ref Yes)", "Intercept"),
          tab.height = "\\textheight", tab.width= "\\textwidth"
          )
```

The baseline result from the binary models seems to support the hypotheses proposed. In particular, being a woman compared to a man on average leads to a a 1% increase in probability to vote for Putin against everyone else, holding everything else constant. The corresponding

Table 2: Binary outcomes, naive approach

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Vote Putin | Decline to Answer | Putin or Declined | Vote Davankov | Spoil the ballot | Vote Davankov or spoil |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Sex: Female | 0.010*** | 0.008*** | 0.018*** | −0.042*** | 0.028*** | −0.014*** |
| | (0.002) | (0.002) | (0.002) | (0.004) | (0.003) | (0.002) |
| Sex: Other | 0.005 | 0.003 | 0.008 | −0.087*** | 0.063*** | −0.024* |
| | (0.011) | (0.011) | (0.011) | (0.023) | (0.021) | (0.014) |
| Age: 18-24 (ref 25-44) | −0.003 | 0.0001 | −0.003 | 0.005 | −0.006 | −0.001 |
| | (0.003) | (0.003) | (0.003) | (0.006) | (0.005) | (0.003) |
| Age: 45-65 (ref 25-44) | 0.044*** | 0.015*** | 0.059*** | −0.154*** | 0.091*** | −0.063*** |
| | (0.003) | (0.003) | (0.003) | (0.006) | (0.005) | (0.003) |
| Age: 65 + (ref 25-44) | 0.063*** | 0.030*** | 0.093*** | −0.153*** | 0.054*** | −0.099*** |
| | (0.004) | (0.005) | (0.004) | (0.009) | (0.009) | (0.006) |
| Took < 1 hour to get to the voting station | 0.001 | 0.013*** | 0.014*** | −0.013*** | −0.004 | −0.017*** |
| | (0.002) | (0.002) | (0.002) | (0.004) | (0.004) | (0.002) |
| Moved after March 2022 (ref before 2014) | −0.042*** | −0.017*** | −0.059*** | 0.113*** | −0.052*** | 0.062*** |
| | (0.003) | (0.003) | (0.003) | (0.007) | (0.006) | (0.004) |
| Moved after March 2019 but before March 2022 (ref before 2014) | −0.034*** | −0.015*** | −0.049*** | 0.050*** | −0.001 | 0.049*** |
| | (0.003) | (0.004) | (0.003) | (0.007) | (0.007) | (0.004) |
| Moved after March 2014 but before March 2019 (ref before 2014) | −0.048*** | −0.020*** | −0.067*** | 0.170*** | −0.088*** | 0.081*** |
| | (0.003) | (0.003) | (0.003) | (0.006) | (0.005) | (0.003) |
| Didn't move - tourist, lives in Russia (ref before 2014) | 0.002 | −0.015*** | −0.013*** | 0.077*** | −0.069*** | 0.007 |
| | (0.005) | (0.005) | (0.005) | (0.010) | (0.009) | (0.006) |
| Trust in the result: Don't know (ref Yes) | −0.665*** | 0.031*** | −0.634*** | 0.435*** | 0.136*** | 0.571*** |
| | (0.005) | (0.005) | (0.005) | (0.011) | (0.010) | (0.006) |
| Trust in the result: No (ref Yes) | −0.731*** | −0.081*** | −0.812*** | 0.458*** | 0.328*** | 0.786*** |
| | (0.003) | (0.003) | (0.003) | (0.006) | (0.005) | (0.003) |
| Intercept | 0.762*** | 0.106*** | 0.868*** | 0.149*** | −0.033*** | 0.116*** |
| | (0.003) | (0.004) | (0.003) | (0.007) | (0.007) | (0.004) |
| Observations | 54,127 | 54,127 | 54,127 | 54,127 | 54,127 | 54,127 |
| $R^2$ | 0.726 | 0.047 | 0.762 | 0.291 | 0.082 | 0.668 |
| Adjusted $R^2$ | 0.726 | 0.047 | 0.762 | 0.291 | 0.082 | 0.668 |
| Residual Std. Error (df = 54114) | 0.193 | 0.210 | 0.199 | 0.418 | 0.384 | 0.247 |
| F Statistic (df = 12; 54114) | 11,955.090*** | 222.367*** | 14,476.830*** | 1,848.849*** | 402.713*** | 9,093.527*** |

*Note:* *p<0.1; **p<0.05; ***p<0.01

increases for being 45-64 and 65+ compared to the reference category of 25-44 are 4% and 6% respectively. With regards to $H_{1b}$ people who do not identify as either male or female are indeed more likely to spoil the ballot. The second set of hypothesis also get preliminary support with all categories of emigrants being less likely to vote for Putin compared to those who left before 2014 and with the strength of the effect going from those who left after 2014 (on average 4.8% holding everything else equal) to those who left in 2019-2022, 3.4% under the same conditions.

```
m2.naive <- multinom(vote ~ sex + age_bin + time_to_vs.less_than_hour
                     + out_of_Russia_time + result_trust_bin,
                     data = data_naive)
```

```
resizebox.stargazer(m2.naive,
          title = "Multinomial regression, naive approach", header = F,
          covariate.labels = c(
            "Sex: Female", "Sex: Other", "Age: 18-24 (ref 25-44)",
            "Age: 45-65 (ref 25-44)", "Age: 65 + (ref 25-44)",
            "Took < 1 hour to get to the voting station",
            "Moved after March 2022 (ref before 2014)",
            "Moved after March 2019 but before March 2022 (ref before 2014)",
            "Moved after March 2014 but before March 2019 (ref before 2014)",
            "Didn't move - tourist, lives in Russia (ref before 2014)",
            "Trust in the result: Don't know (ref Yes)",
            "Trust in the result: No (ref Yes)", "Intercept"),
          tab.height = "\\textheight", tab.width= "\\textwidth"
          )
```

The multinomial results are also in line with hypotheses. Here the log odds of choosing a candidate against a baseline (Putin) are presented. For example for women on average the relative risk ratio to choose Davankov over Putin is $exp(-0.555) = 0.574$ which is a pretty substantial effect. Those effects are even higher when looking at older demographics. All other hypotheses are also mirrored from the binary models.

Keep in mind that the main assumption of this model is violated, if we were to kick Putin or Davankov out of the choice options the results would be unpredictable compared to kicking out the irrelevant candidates like Slutsky or Haritonov.

```
comparisons <- logits(not_putin = dichotomy(opposition = c("Davankov",
                                                           "Spoiled ballot",
                                                           "Slutsky",
                                                           "Haritonov"),
                                            "Putin"),
                      opposition = dichotomy(
                        systemic = c("Slutsky", "Haritonov"),
                        nonsystemic = c("Davankov", "Spoiled ballot")),
                      nonsystemic = c("Spoiled ballot", "Davankov"),
                      systemic = c("Haritonov", "Slutsky"))

wlf.nested <- nestedLogit(vote ~ sex + age_bin + time_to_vs.less_than_hour
                          + out_of_Russia_time + result_trust_bin,
                          dichotomies = comparisons,
                          data = data_naive,
```

Table 3: Multinomial regression, naive approach

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | Davankov | Haritonov | Slutsky | Spoiled ballot | Tore up/took |
| | (1) | (2) | (3) | (4) | (5) |
| Sex: Female | $-0.555^{***}$ | $-0.584^{***}$ | $-0.460^{***}$ | $-0.333^{***}$ | $-0.684^{***}$ |
| | (0.069) | (0.092) | (0.109) | (0.071) | (0.138) |
| Sex: Other | $-0.080$ | 0.249 | 0.690 | 0.465 | 0.833 |
| | (0.369) | (0.490) | (0.511) | (0.379) | (0.616) |
| Age: 18-24 (ref 25-44) | 0.121 | 0.079 | $0.395^{**}$ | 0.095 | $0.432^{**}$ |
| | (0.119) | (0.157) | (0.169) | (0.123) | (0.210) |
| Age: 45-65 (ref 25-44) | $-1.607^{***}$ | $-1.027^{***}$ | $-0.967^{***}$ | $-0.758^{***}$ | $-0.470^{**}$ |
| | (0.086) | (0.127) | (0.152) | (0.089) | (0.185) |
| Age: 65 + (ref 25-44) | $-2.004^{***}$ | $-0.584^{***}$ | $-0.884^{***}$ | $-1.020^{***}$ | $-0.565$ |
| | (0.140) | (0.189) | (0.252) | (0.150) | (0.389) |
| Took < 1 hour to get to the voting station | $-0.390^{***}$ | $-0.298^{***}$ | $-0.230^{*}$ | $-0.387^{***}$ | $-0.273^{*}$ |
| | (0.076) | (0.100) | (0.118) | (0.078) | (0.146) |
| Moved after March 2022 (ref before 2014) | $1.624^{***}$ | $0.997^{***}$ | $1.149^{***}$ | $1.108^{***}$ | $1.378^{***}$ |
| | (0.105) | (0.141) | (0.171) | (0.109) | (0.218) |
| Moved after March 2019 but before March 2022 (ref before 2014) | $0.937^{***}$ | $0.411^{**}$ | $0.783^{***}$ | $0.719^{***}$ | $0.883^{***}$ |
| | (0.118) | (0.163) | (0.188) | (0.122) | (0.243) |
| Moved after March 2014 but before March 2019 (ref before 2014) | $2.287^{***}$ | $1.259^{***}$ | $1.423^{***}$ | $1.541^{***}$ | $1.488^{***}$ |
| | (0.090) | (0.123) | (0.150) | (0.093) | (0.198) |
| Didn't move - tourist, lives in Russia (ref before 2014) | $0.725^{***}$ | 0.241 | $0.805^{***}$ | $-0.125$ | 0.166 |
| | (0.134) | (0.209) | (0.223) | (0.149) | (0.401) |
| Trust in the result: Don't know (ref Yes) | $4.212^{***}$ | $3.360^{***}$ | $3.642^{***}$ | $3.880^{***}$ | $2.611^{***}$ |
| | (0.111) | (0.187) | (0.193) | (0.158) | (0.666) |
| Trust in the result: No (ref Yes) | $7.312^{***}$ | $5.818^{***}$ | $5.340^{***}$ | $8.203^{***}$ | $7.112^{***}$ |
| | (0.103) | (0.144) | (0.160) | (0.130) | (0.339) |
| Intercept | $-2.359^{***}$ | $-3.888^{***}$ | $-4.431^{***}$ | $-3.906^{***}$ | $-6.517^{***}$ |
| | (0.103) | (0.161) | (0.194) | (0.134) | (0.378) |
| Akaike Inf. Crit. | 69,701.630 | 69,701.630 | 69,701.630 | 69,701.630 | 69,701.630 |
| *Note:* | | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

```
                          subset = data_naive$vote != "Tore up/took"
                          | !is.na(data_naive$vote))

resizebox.stargazer(models(wlf.nested),
          title = "Nested Logit models, naive approach", header = F,
          dep.var.labels = c("Putin vs everyone", "Non-systemic vs systemic opposition",
                             "Spoiled vs Davankov", "Slutsky vs Haritonov"),
          covariate.labels = c(
            "Sex: Female", "Sex: Other", "Age: 18-24 (ref 25-44)",
            "Age: 45-65 (ref 25-44)", "Age: 65 + (ref 25-44)",
            "Took < 1 hour to get to the voting station",
            "Moved after March 2022 (ref before 2014)",
            "Moved after March 2019 but before March 2022 (ref before 2014)",
            "Moved after March 2014 but before March 2019 (ref before 2014)",
            "Didn't move - tourist, lives in Russia (ref before 2014)",
            "Trust in the result: Don't know (ref Yes)",
            "Trust in the result: No (ref Yes)", "Intercept"),
          tab.height = "\\textheight", tab.width= "\\textwidth"
          )
```

To finish up "naive" estimations the nested logit models provide the clearest picture of the voter choices. I propose a scenario when a voter going in the voting booth (actually, way before that but let's keep it simple) has to first decide whether he will vote for the incumbent or not. This is reflected in the first column, "Putin vs everyone" with Putin being 1 and 0 being everyone else. Women are on average $exp(0.513) = 1.670295$ times more likely to choose Putin here with 45-65 category being more than 3 times more likely to do so than those aged 25-44 (all of this keeping everything else constant of course)!

The coefficient for the "Other" gender is not significant meaning that their choices among opposition candidates are probably different from men (reference). On the other hand everyone who moved after 2014 are less likely to choose the incumbent with those left after annexation being 86% less likely to do so, and those who left after the start of the full-scale invasion being 75% less likely to with those in the middle being a less decisive 55%, all else equal.

Putting Putin aside let's look at the determinants of choosing non-systemic (Davankov or spoil) vs systemic (Haritonov vs Slutsky) options. Gender doesn't appear to matter but as predicted by $H_{1a}$ older demographics are less likely to vote for non-systemic opposition in this comparison. The pattern by migrant arrival time persists. People who do not believe in the fairness of the election result are likely to pick non-systemic options.

Within those women and other genders are more likely to vote for Davankov (20% and 40% respectively) then men. Groups who presumably left Russia for political reasons are more likely to spoil the ballot. There is a very interesting result in the "trust" variable. Those who do not know if they believe the election at this point are more likely to spoil the ballot compared to those that believe in the result while those who do not believe with any degree of certainty are more likely to vote for Davankov.

In the last branch (1.418 obsservations) there is one result that is significant - that is, people who do not trust in the result choose Haritonov. I think I have an explanation for this. If not

Table 4: Nested Logit models, naive approach

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | Putin vs everyone | Non-systemic vs systemic opposition | Spoiled vs Davankov | Slutsky vs Haritonov |
| | (1) | (2) | (3) | (4) |
| Sex: Female | 0.513*** | 0.022 | −0.219*** | 0.156 |
| | (0.068) | (0.056) | (0.024) | (0.115) |
| Sex: Other | −0.134 | −0.383 | −0.516*** | 0.426 |
| | (0.372) | (0.298) | (0.146) | (0.573) |
| Age: 18-24 (ref 25-44) | −0.134 | −0.079 | 0.025 | 0.327* |
| | (0.118) | (0.085) | (0.036) | (0.170) |
| Age: 45-65 (ref 25-44) | 1.334*** | −0.307*** | −0.838*** | 0.008 |
| | (0.083) | (0.087) | (0.039) | (0.179) |
| Age: 65 + (ref 25-44) | 1.527*** | −0.946*** | −0.968*** | −0.445 |
| | (0.126) | (0.153) | (0.106) | (0.304) |
| Took < 1 hour to get to the voting station | 0.367*** | −0.107* | −0.001 | 0.033 |
| | (0.075) | (0.059) | (0.024) | (0.122) |
| Moved after March 2022 (ref before 2014) | −1.413*** | 0.402*** | 0.513*** | 0.150 |
| | (0.103) | (0.090) | (0.041) | (0.187) |
| Moved after March 2019 but before March 2022 (ref before 2014) | −0.816*** | 0.292*** | 0.217*** | 0.390* |
| | (0.116) | (0.102) | (0.045) | (0.208) |
| Moved after March 2014 but before March 2019 (ref before 2014) | −1.990*** | 0.741*** | 0.742*** | 0.165 |
| | (0.087) | (0.080) | (0.036) | (0.166) |
| Didn't move - tourist, lives in Russia (ref before 2014) | −0.529*** | 0.0005 | 0.855*** | 0.547** |
| | (0.130) | (0.141) | (0.084) | (0.273) |
| Trust in the result: Don't know (ref Yes) | −4.032*** | 0.712*** | 0.402*** | 0.259 |
| | (0.105) | (0.127) | (0.140) | (0.239) |
| Trust in the result: No (ref Yes) | −7.366*** | 1.917*** | −0.811*** | −0.544*** |
| | (0.099) | (0.094) | (0.106) | (0.183) |
| Intercept | 1.880*** | 1.272*** | 1.467*** | −0.472** |
| | (0.098) | (0.119) | (0.111) | (0.229) |
| Observations | 51,215 | 42,375 | 40,957 | 1,418 |
| Log Likelihood | −3,535.670 | −5,870.536 | −22,764.710 | −911.511 |
| Akaike Inf. Crit. | 7,097.340 | 11,767.070 | 45,555.420 | 1,849.023 |
| *Note:* | | | | *p<0.1; **p<0.05; ***p<0.01 |

voting for non-systemic opposition options, voters who are opposition-minded (and following the “vote for anyone but Putin” formula) are more likely to support a communist candidate Haritonov for two reasons: one being that the communist party in local elections was the one who benefited from Navalny’s endorcements the most and the second is that Slutsky was implicated in sexually harassing journalists a couple of years ago.

Overall even without accounting for the nested structure and data missingness all models converge in their predictions and support the hypotheses (to my big surprise).

### Treating missing data seriously

I first impute independent variables' values whenever possible, then talk about ways to account for non-responses in vote choice and lastly introduce multiple levels of analysis present in the data.

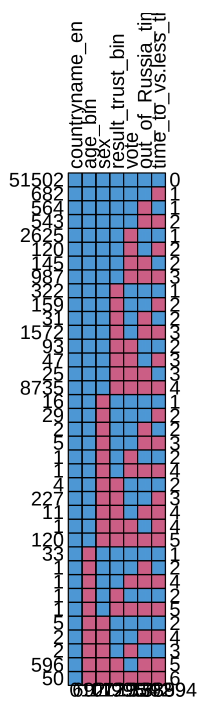
```r
# Create table of missing data
data_naive |>
  select(vote, sex, age_bin, time_to_vs.less_than_hour, out_of_Russia_time,
         result_trust_bin, countryname_en) |>
  dfSummary(graph.col = F, varnumbers = F, split.cells = 20,
            plain.ascii = FALSE, headings = F, valid.col = F,
            style = "grid")
```

| Variable | Stats / Values | Freqs (% of Valid) | Missing |
|---|---|---|---|
| vote<br>[factor] | 1. Putin<br>2. Davankov<br>3. Haritonov<br>4. Slutsky<br>5. Spoiled ballot<br>6. Tore up/took | 10410 (18.5%)<br>32202 (57.2%)<br>994 ( 1.8%)<br>573 ( 1.0%)<br>11813 (21.0%)<br>315 ( 0.6%) | 12954<br>(18.7%) |
| sex<br>[factor] | 1. Male<br>2. Female<br>3. Other | 29166 (42.8%)<br>38527 (56.5%)<br>496 ( 0.7%) | 1072<br>(1.5%) |
| age_bin<br>[factor] | 1. 25-44<br>2. 18-24<br>3. 45-64<br>4. 65+ | 43388 (63.3%)<br>6781 ( 9.9%)<br>13227 (19.3%)<br>5173 ( 7.5%) | 692<br>(1.0%) |
| time_to_vs.less_than_hour<br>[character] | 1. No<br>2. Yes | 17941 (32.4%)<br>37431 (67.6%) | 13889<br>(20.1%) |
| out_of_Russia_time<br>[factor] | 1. Before annexation<br>2. 2 - 5 years<br>3. After annexation<br>4. After invasion<br>5. Tourist (lives in Russia) | 12362 (22.1%)<br>9745 (17.4%)<br>5858 (10.5%)<br>25746 (46.1%)<br>2158 ( 3.9%) | 13392<br>(19.3%) |
| result_trust_bin<br>[factor] | 1. Yes<br>2. Don't know<br>3. No | 11639 (20.3%)<br>2864 ( 5.0%)<br>42763 (74.7%) | 11995<br>(17.3%) |

| Variable | Stats / Values | Freqs (% of Valid) | Missing |
|---|---|---|---|
| countryname_en [character] | 1. Cyprus | 5663 ( 8.2%) | 0 (0.0%) |
| | 2. Germany | 4963 ( 7.2%) | |
| | 3. Serbia | 3906 ( 5.6%) | |
| | 4. Armenia | 3903 ( 5.6%) | |
| | 5. Israel | 3035 ( 4.4%) | |
| | 6. Austria | 3015 ( 4.4%) | |
| | 7. France | 2818 ( 4.1%) | |
| | 8. Finland | 2686 ( 3.9%) | |
| | 9. Kazakhstan | 2675 ( 3.9%) | |
| | 10. Uzbekistan | 2447 ( 3.5%) | |
| | [ 34 others ] | 34150 (49.3%) | |

Examine missing data patterns:

```
# Create graph
data_naive |>
  select(vote, sex, age_bin, time_to_vs.less_than_hour, out_of_Russia_time,
         result_trust_bin, countryname_en) |>
  md.pattern(rotate.names = T)
```



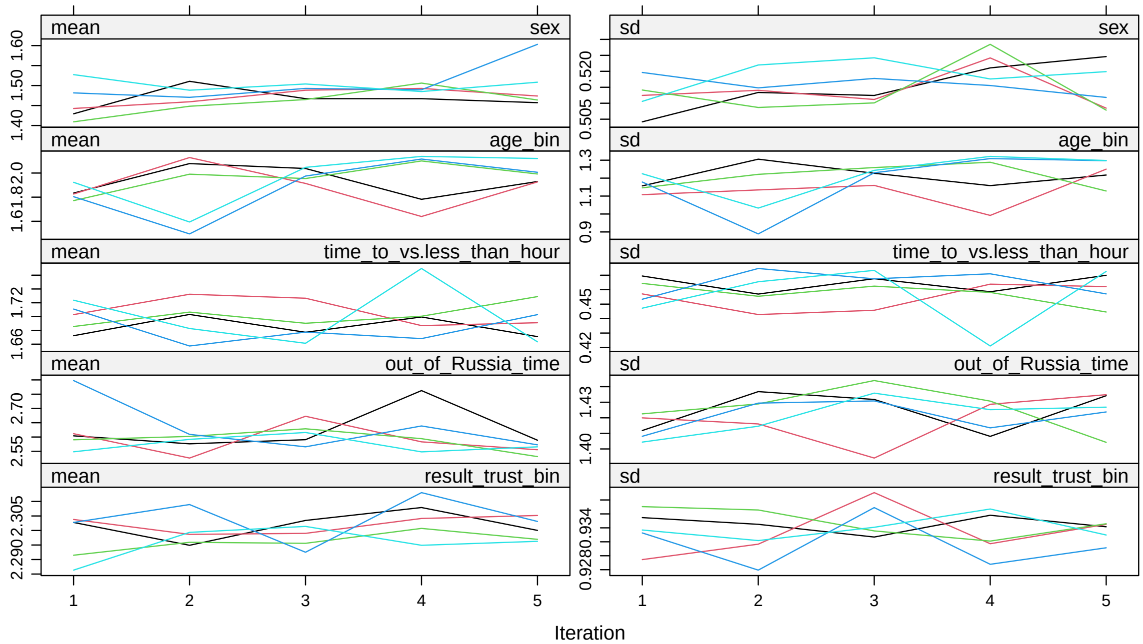
### Independent variables

Missingness in data in exit polls perceived to be conducted by opposition backers likely has to do with the candidate choice, namely, the incumbent. However the answer to the vote choice question also suffers from missingness.

The issue is inherently formed by the country where the voting station is situated. The expectation is, for example, that voters in countries closer to Russia and within the Collective Security Treaty Organization (CSTO) could still fear monitoring from the Russian state and behave closer to how voters behave in Russia itself - avoid sharing political opinions that diverge from the "party line" and project higher social desirability bias. This would lead to outcome missing data in those countries underappreciating the opposition vote (namely Davankov and spoiled ballots). On the other hand in countries farther from home, incumbent supporters would have reasons to conceal their choices. Many voting lines within the "Afternoon against Putin" campaign were accompanied by demonstrations denouncing Putin, sometimes in collaboration with Ukranian activists. This would mean that Putin-backers could conceal their preferences in other variables as well when outnumbered.

Therefore, I use imputation methods suitable for multi-level data structures. I first fit an appropriate pooled imputation using correct imputation methods. I then estimate a multi-level imputation model, subject to model constraints described further.

```
# First try, ignore the groups, do best method for polytonomous data

# Data
impdata <- data_naive |>
  select(vote, sex, age_bin, time_to_vs.less_than_hour, out_of_Russia_time,
         result_trust_bin, countryname_en, city_en) |>
  mutate(across(c(time_to_vs.less_than_hour, countryname_en, city_en), ~ as.factor(.)))

# Dry run
imp1dry <- mice(impdata, seed = 1, maxit = 0)

pred <- imp1dry$predictorMatrix
# imp1dry$method

# Impute everything except vote
# imp1 <- mice(impdata, method = c("", "polyreg", "polyreg", "logreg", "polyreg", "polyreg", "", ""))

# save(imp1, file = "imp1.RData")
load(here("scripts", "imp1.RData"))

# Check conversion
plot(imp1, layout = c(2, 5))
```




```
fit1 <- with(imp1, multinom(as.factor(vote) ~ as.factor(sex)
                             + as.factor(age_bin)
                             + as.factor(time_to_vs.less_than_hour)
                             + as.factor(out_of_Russia_time)
                             + as.factor(result_trust_bin)))


ms_fit1 <- pool(fit1)

save(ms_fit1, file = "ms_fit1.RData")


load("ms_fit1.RData")

modelsummary(ms_fit1, shape = term + statistic ~ response,
             col.names = c("", "Davankov", "Haritonov", "Slutsky",
                           "Spoiled ballot", "Tore up/took"),
             title = "Pooled multinomial model with pooled imputation",
             output = "kableExtra", stars = T,
             coef_map = c("(Intercept)" = "Intercept",
                          "as.factor(sex)Female" = "Sex: Female",
                          "as.factor(sex)Other" = "Sex: Other",
                          "as.factor(age_bin)18-24" = "Age: 18-24 (ref 25-44)",
                          "as.factor(age_bin)45-64" = "Age: 45-65 (ref 25-44)",
                          "as.factor(age_bin)65+" = "Age: 65 + (ref 25-44)",
                          "as.factor(time_to_vs.less_than_hour)Yes"
                          = "Took < 1 hour to get to the voting station",
                          "as.factor(out_of_Russia_time)2 - 5 years"
                          = "Moved after March 2019 but before March 2022 (ref before 2014)",
                          "as.factor(out_of_Russia_time)After annexation"
                          = "Moved after March 2014 but before March 2019 (ref before 2014)",
```

```
                    "as.factor(out_of_Russia_time)After invasion"
                    = "Moved after March 2022 (ref before 2014)",
                    "as.factor(out_of_Russia_time)Tourist (lives in Russia)"
                    = "Didn't move - tourist, lives in Russia (ref before 2014)",
                    "as.factor(result_trust_bin)Don't know"
                    = "Trust in the result: Don't know (ref Yes)",
                    "as.factor(result_trust_bin)No"
                    = "Trust in the result: No (ref Yes)"),
           gof_map = NA, booktabs = T,
           add_rows = data.frame(term = c("Observations", "Imputations", "Imputation Model"),
                                 Davankov = c("56 307", "5", "Polytonomous regression"),
                                 Haritonov = c("56 307", "5", "Polytonomous regression"),
                                 Slutsky = c("56 307", "5", "Polytonomous regression"),
                                 `Spoiled ballot` = c("56 307", "5", "Polytonomous regression"),
                                 `Tore up/Took` = c("56 307", "5", "Polytonomous regression"))) |>
  kable_styling(latex_options = c("scale_down", "hold_position"))
```

Table 6: Pooled multinomial model with pooled imputation

| | (1) | | | | |
|---|---|---|---|---|---|
| | Davankov | Haritonov | Slutsky | Spoiled ballot | Tore up/took |
| Intercept | −2.381*** | −3.853*** | −4.457*** | −3.897*** | −6.558*** |
| | (0.099) | (0.155) | (0.188) | (0.128) | (0.371) |
| Sex: Female | −0.558*** | −0.600*** | −0.453*** | −0.333*** | −0.697*** |
| | (0.067) | (0.091) | (0.104) | (0.069) | (0.131) |
| Sex: Other | −0.037 | 0.389 | 0.463 | 0.685+ | 0.617 |
| | (0.359) | (0.445) | (0.505) | (0.359) | (0.607) |
| Age: 18-24 (ref 25-44) | 0.176 | 0.141 | 0.407* | 0.132 | 0.491* |
| | (0.112) | (0.148) | (0.161) | (0.115) | (0.200) |
| Age: 45-65 (ref 25-44) | −1.645*** | −1.033*** | −1.051*** | −0.810*** | −0.458** |
| | (0.089) | (0.123) | (0.146) | (0.090) | (0.176) |
| Age: 65 + (ref 25-44) | −2.053*** | −0.549** | −1.027*** | −1.084*** | −0.832* |
| | (0.135) | (0.172) | (0.238) | (0.138) | (0.389) |
| Took < 1 hour to get to the voting station | −0.384*** | −0.306** | −0.224+ | −0.382*** | −0.289* |
| | (0.074) | (0.097) | (0.124) | (0.075) | (0.142) |
| Moved after March 2019 but before March 2022 (ref before 2014) | 1.585*** | 0.986*** | 1.110*** | 1.074*** | 1.339*** |
| | (0.107) | (0.137) | (0.175) | (0.112) | (0.216) |
| Moved after March 2014 but before March 2019 (ref before 2014) | 0.907*** | 0.391* | 0.723*** | 0.681*** | 0.819*** |
| | (0.119) | (0.160) | (0.186) | (0.124) | (0.245) |
| Moved after March 2022 (ref before 2014) | 2.242*** | 1.229*** | 1.434*** | 1.500*** | 1.510*** |
| | (0.087) | (0.117) | (0.143) | (0.091) | (0.200) |
| Didn't move - tourist, lives in Russia (ref before 2014) | 0.717*** | 0.264 | 0.858*** | −0.110 | 0.208 |
| | (0.127) | (0.198) | (0.212) | (0.142) | (0.379) |
| Trust in the result: Don't know (ref Yes) | 4.028*** | 3.078*** | 3.441*** | 3.723*** | 2.875*** |
| | (0.098) | (0.172) | (0.183) | (0.142) | (0.563) |
| Trust in the result: No (ref Yes) | 7.247*** | 5.698*** | 5.296*** | 8.117*** | 7.074*** |
| | (0.100) | (0.137) | (0.155) | (0.126) | (0.331) |
| Observations | 56 307 | 56 307 | 56 307 | 56 307 | 56 307 |
| Imputations | 5 | 5 | 5 | 5 | 5 |
| Imputation Model | Polytonomous regression | Polytonomous regression | Polytonomous regression | Polytonomous regression | Polytonomous regression |

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

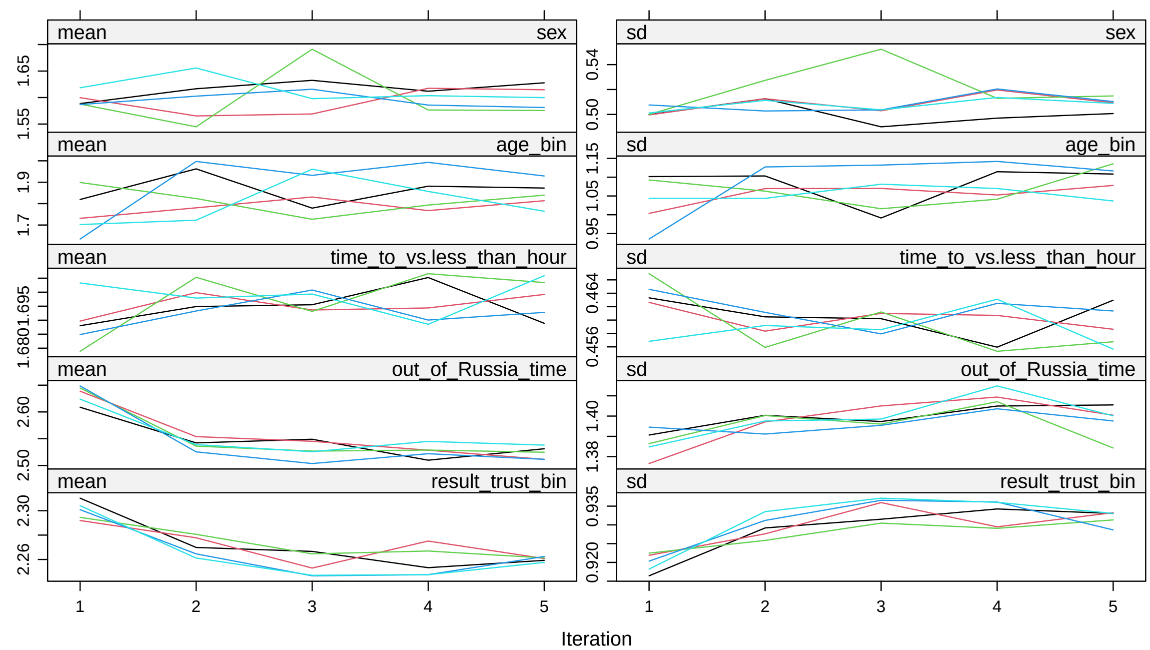
Results here mimic the multinomial naive model closely.

There are no polytonomous methods embedded within mice or its expansion packages. However the need to integrate higher levels of analysis is pertinent, so I use predictive mean matching methods to impute values while considering the structure of the data. I use city (voting station) level fixed effects first and then estimate with country fixed effects.

In comparison with the polytonomous models here vote cannot be a predictor due to model specificity. I therefore drop it as a predictor.

```
impdata2 <- impdata |>
  mutate(across(c(vote, sex, age_bin, time_to_vs.less_than_hour, out_of_Russia_time,
         result_trust_bin, countryname_en, city_en), ~ as.integer(.)))
```

```r
pred2 <- imp1$predictorMatrix
pred2[, "city_en"] <- -2
pred2[, "countryname_en"] <- 1
pred2[, "vote"] <- 0

#imp2 <- mice(impdata2, seed = 1, predictorMatrix = pred2, method = c("", "2l.pmm", "2l.pmm", "2l.pmm", "2l.pmm", "2l.pmm", "

# save(imp2, file = "imp2.RData")
load(here("scripts", "imp2.RData"))

# Converges OK!
plot(imp2, layout = c(2, 5))
```




```r
fit2 <- with(imp2, multinom(as.factor(vote) ~ as.factor(sex)
                            + as.factor(age_bin)
                            + as.factor(time_to_vs.less_than_hour)
                            + as.factor(out_of_Russia_time)
                            + as.factor(result_trust_bin)))


ms_fit2 <- pool(fit2)

save(ms_fit2, file = "ms_fit2.RData")


load("ms_fit2.RData")

modelsummary(ms_fit2, shape = term + statistic ~ response,
```

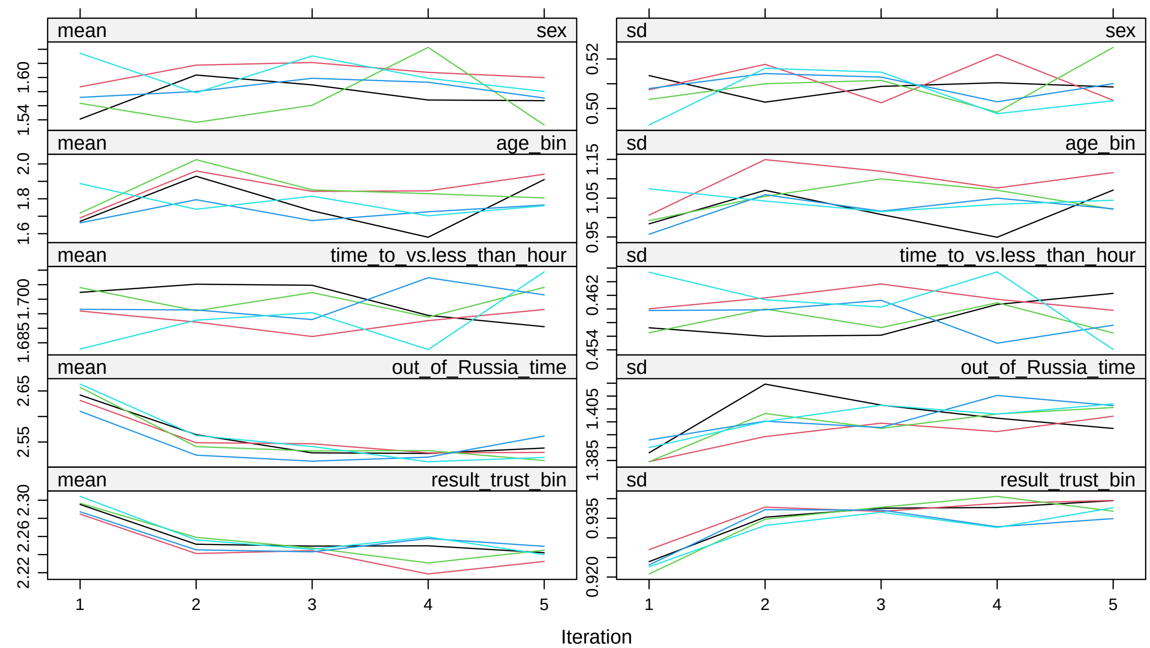
```r
            col.names = c("", "Davankov", "Haritonov", "Slutsky",
                          "Spoiled ballot", "Tore up/took"),
            title = "Pooled multinomial model with city fixed effect imputation",
            output = "kableExtra", stars = T,
            coef_map = c("(Intercept)" = "Intercept",
                         "as.factor(sex)2" = "Sex: Female",
                         "as.factor(sex)3" = "Sex: Other",
                         "as.factor(age_bin)2" = "Age: 18-24 (ref 25-44)",
                         "as.factor(age_bin)3" = "Age: 45-65 (ref 25-44)",
                         "as.factor(age_bin)4" = "Age: 65 + (ref 25-44)",
                         "as.factor(time_to_vs.less_than_hour)2"
                         = "Took < 1 hour to get to the voting station",
                         "as.factor(out_of_Russia_time)2"
                         = "Moved after March 2019 but before March 2022 (ref before 2014)",
                         "as.factor(out_of_Russia_time)3"
                         = "Moved after March 2014 but before March 2019 (ref before 2014)",
                         "as.factor(out_of_Russia_time)4"
                         = "Moved after March 2022 (ref before 2014)",
                         "as.factor(out_of_Russia_time)5"
                         = "Didn't move - tourist, lives in Russia (ref before 2014)",
                         "as.factor(result_trust_bin)2"
                         = "Trust in the result: Don't know (ref Yes)",
                         "as.factor(result_trust_bin)3"
                         = "Trust in the result: No (ref Yes)"),
            gof_map = "none",
            booktabs = T,
            add_rows = data.frame(term = c("Observations", "Imputations", "Imputation Model"),
                                  Davankov = c("56 307", "5", "PMM with city FE"),
                                  Haritonov = c("56 307", "5", "PMM with city FE"),
                                  Slutsky = c("56 307", "5", "PMM with city FE"),
                                  `Spoiled ballot` = c("56 307", "5", "PMM with city FE"),
                                  `Tore up/Took` = c("56 307", "5", "PMM with city FE"))
           ) |>
  kable_styling(latex_options = c("scale_down", "hold_position"))
```

```r
pred3 <- imp1$predictorMatrix
pred3[, "countryname_en"] <- -2
pred3[, "city_en"] <- 1
pred3[, "vote"] <- 0

#imp3 <- mice(impdata2, seed = 1, predictorMatrix = pred3, method = c("", "2l.pmm", "2l.pmm", "2l.pmm", "2l.pmm", "2l.pmm", "

# save(imp3, file = "imp3.RData")
load(here("scripts", "imp3.RData"))

# Converges OK!
plot(imp3, layout = c(2, 5))
```

Table 7: Pooled multinomial model with city fixed effect imputation

| | (1) | | | | |
|---|---|---|---|---|---|
| | Davankov | Haritonov | Slutsky | Spoiled ballot | Tore up/took |
| Intercept | −1.947*** | −3.709*** | −4.317*** | −3.159*** | −6.172*** |
| | (0.127) | (0.151) | (0.208) | (0.122) | (0.358) |
| Sex: Female | −0.409*** | −0.459*** | −0.339*** | −0.187* | −0.548*** |
| | (0.067) | (0.084) | (0.101) | (0.069) | (0.130) |
| Sex: Other | 0.303 | 0.606 | 0.672 | 1.036* | 0.935 |
| | (0.363) | (0.439) | (0.481) | (0.363) | (0.606) |
| Age: 18-24 (ref 25-44) | 0.303** | 0.259+ | 0.513** | 0.254* | 0.607** |
| | (0.111) | (0.145) | (0.161) | (0.114) | (0.201) |
| Age: 45-65 (ref 25-44) | −1.800*** | −1.165*** | −1.191*** | −0.976*** | −0.630*** |
| | (0.065) | (0.112) | (0.142) | (0.069) | (0.166) |
| Age: 65 + (ref 25-44) | −2.136*** | −0.717*** | −1.139*** | −1.208*** | −0.961* |
| | (0.124) | (0.171) | (0.236) | (0.142) | (0.382) |
| Took < 1 hour to get to the voting station | −0.237** | −0.157 | −0.073 | −0.233* | −0.125 |
| | (0.079) | (0.098) | (0.124) | (0.081) | (0.151) |
| Moved after March 2019 but before March 2022 (ref before 2014) | 1.268*** | 0.694*** | 0.829*** | 0.772*** | 1.009*** |
| | (0.105) | (0.131) | (0.172) | (0.113) | (0.223) |
| Moved after March 2014 but before March 2019 (ref before 2014) | 0.725*** | 0.241+ | 0.551** | 0.507*** | 0.614** |
| | (0.103) | (0.146) | (0.177) | (0.108) | (0.236) |
| Moved after March 2022 (ref before 2014) | 1.700*** | 0.745*** | 0.947*** | 0.968*** | 0.934*** |
| | (0.091) | (0.115) | (0.158) | (0.098) | (0.198) |
| Didn't move - tourist, lives in Russia (ref before 2014) | 0.364* | −0.065 | 0.575** | −0.434** | −0.128 |
| | (0.136) | (0.195) | (0.217) | (0.148) | (0.391) |
| Trust in the result: Don't know (ref Yes) | 3.517*** | 2.821*** | 3.195*** | 2.999*** | 2.585*** |
| | (0.096) | (0.176) | (0.177) | (0.155) | (0.530) |
| Trust in the result: No (ref Yes) | 5.504*** | 4.221*** | 3.837*** | 6.059*** | 5.387*** |
| | (0.066) | (0.120) | (0.139) | (0.091) | (0.323) |
| Observations | 56 307 | 56 307 | 56 307 | 56 307 | 56 307 |
| Imputations | 5 | 5 | 5 | 5 | 5 |
| Imputation Model | PMM with city FE | PMM with city FE | PMM with city FE | PMM with city FE | PMM with city FE |

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001



```r
fit3 <- with(imp3, multinom(as.factor(vote) ~ as.factor(sex)
                              + as.factor(age_bin)
                              + as.factor(time_to_vs.less_than_hour)
                              + as.factor(out_of_Russia_time)
                              + as.factor(result_trust_bin)))


ms_fit3 <- pool(fit3)

save(ms_fit3, file = "ms_fit3.RData")


load("ms_fit3.RData")

modelsummary(ms_fit3, shape = term + statistic ~ response,
             col.names = c("", "Davankov", "Haritonov", "Slutsky",
                           "Spoiled ballot", "Tore up/took"),
             title = "Pooled multinomial model with country fixed effects imputation model",
             output = "kableExtra", stars = T,
             coef_map = c("(Intercept)" = "Intercept",
                          "as.factor(sex)2" = "Sex: Female",
                          "as.factor(sex)3" = "Sex: Other",
                          "as.factor(age_bin)2" = "Age: 18-24 (ref 25-44)",
                          "as.factor(age_bin)3" = "Age: 45-65 (ref 25-44)",
                          "as.factor(age_bin)4" = "Age: 65 + (ref 25-44)",
                          "as.factor(time_to_vs.less_than_hour)2"
                          = "Took < 1 hour to get to the voting station",
                          "as.factor(out_of_Russia_time)2"
                          = "Moved after March 2019 but before March 2022 (ref before 2014)",
                          "as.factor(out_of_Russia_time)3"
                          = "Moved after March 2014 but before March 2019 (ref before 2014)",
                          "as.factor(out_of_Russia_time)4"
                          = "Moved after March 2022 (ref before 2014)",
                          "as.factor(out_of_Russia_time)5"
                          = "Didn't move - tourist, lives in Russia (ref before 2014)",
                          "as.factor(result_trust_bin)2"
                          = "Trust in the result: Don't know (ref Yes)",
                          "as.factor(result_trust_bin)3"
                          = "Trust in the result: No (ref Yes)"),
             gof_map = "none",
             booktabs = T,
             add_rows = data.frame(term = c("Observations", "Imputations", "Imputation model"),
                                   Davankov = c("56 307", "5", "PMM with country FE"),
                                   Haritonov = c("56 307", "5", "PMM with country FE"),
                                   Slutsky = c("56 307", "5", "PMM with country FE"),
                                   `Spoiled ballot` = c("56 307", "5", "PMM with country FE"),
                                   `Tore up/Took` = c("56 307", "5", "PMM with country FE"))
            ) |>
  kable_styling(latex_options = c("scale_down", "hold_position"))
```

### Missingness in the dependent variable

The dependent variable, the vote choice, is missing according to two latent variables. Both of them can be ascribed to the individual decision not to disclose the vote but probably operate on the city/country level. The first latent variable is the perception of support during the voting

Table 8: Pooled multinomial model with country fixed effects imputation model

| | (1) | | | | |
|---|---|---|---|---|---|
| | Davankov | Haritonov | Slutsky | Spoiled ballot | Tore up/took |
| Intercept | −1.932*** | −3.675*** | −4.263*** | −3.127*** | −6.168*** |
| | (0.095) | (0.145) | (0.176) | (0.099) | (0.367) |
| Sex: Female | −0.413*** | −0.469*** | −0.355** | −0.193** | −0.565*** |
| | (0.065) | (0.088) | (0.107) | (0.068) | (0.132) |
| Sex: Other | −0.283 | 0.136 | 0.290 | 0.424 | 0.360 |
| | (0.316) | (0.415) | (0.479) | (0.310) | (0.587) |
| Age: 18-24 (ref 25-44) | 0.280** | 0.234+ | 0.494** | 0.233* | 0.586** |
| | (0.096) | (0.139) | (0.153) | (0.099) | (0.191) |
| Age: 45-65 (ref 25-44) | −1.741*** | −1.128*** | −1.155*** | −0.914*** | −0.579*** |
| | (0.075) | (0.112) | (0.142) | (0.078) | (0.172) |
| Age: 65 + (ref 25-44) | −2.189*** | −0.696*** | −1.131*** | −1.210*** | −0.951* |
| | (0.132) | (0.167) | (0.236) | (0.140) | (0.381) |
| Took < 1 hour to get to the voting station | −0.258*** | −0.180+ | −0.119 | −0.253** | −0.172 |
| | (0.070) | (0.093) | (0.114) | (0.070) | (0.145) |
| Moved after March 2019 but before March 2022 (ref before 2014) | 1.223*** | 0.672*** | 0.807*** | 0.720*** | 0.939*** |
| | (0.109) | (0.138) | (0.163) | (0.110) | (0.222) |
| Moved after March 2014 but before March 2019 (ref before 2014) | 0.704*** | 0.239 | 0.562** | 0.486*** | 0.613* |
| | (0.112) | (0.162) | (0.188) | (0.117) | (0.237) |
| Moved after March 2022 (ref before 2014) | 1.729*** | 0.797*** | 0.980*** | 0.994*** | 0.966*** |
| | (0.097) | (0.121) | (0.148) | (0.099) | (0.208) |
| Didn't move - tourist, lives in Russia (ref before 2014) | 0.438* | 0.025 | 0.599** | −0.349* | −0.011 |
| | (0.147) | (0.188) | (0.214) | (0.152) | (0.370) |
| Trust in the result: Don't know (ref Yes) | 3.497*** | 2.763*** | 3.139*** | 2.955*** | 2.719*** |
| | (0.101) | (0.181) | (0.188) | (0.134) | (0.658) |
| Trust in the result: No (ref Yes) | 5.522*** | 4.205*** | 3.831*** | 6.064*** | 5.440*** |
| | (0.064) | (0.114) | (0.136) | (0.088) | (0.341) |
| Observations | 56 307 | 56 307 | 56 307 | 56 307 | 56 307 |
| Imputations | 5 | 5 | 5 | 5 | 5 |
| Imputation model | PMM with country FE | PMM with country FE | PMM with country FE | PMM with country FE | PMM with country FE |

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

process. This explains why people that are voting for Putin in Europe for example, might refuse to disclose their vote - larger shares of opposition backers and different demonstrations occurring on election day make people voting for the incumbent feel like a minority.

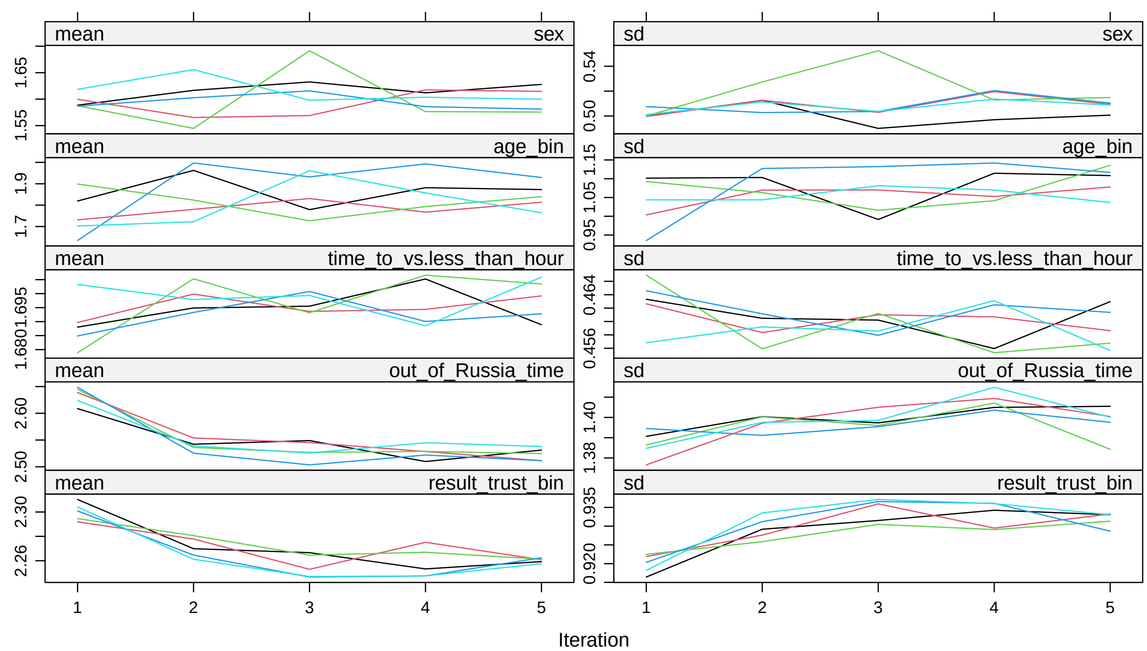
The other latent variable is closeness to Russia. This entails cultural, geographical and ideological distance. In countries more approving of Russian policy those that are closer to it or share borders, there would be more people that are not opposition backers, leading to an opposite effect of the one mentioned above. Moreover, capabilities of the state might be perceived as extending to those areas. In those cases, similar to actual authoritarian contexts, people will censor themselves against disclosing disdain for incumbent.

One of the options with regards to the dependent variable is to model it explicitly. Since missingness in the predictors is dealt with, we can directly compare the demographic composition of those refusing to disclose the vote vis-a-vis other candidates. Considering the geography of exit poll locations - rich industrialized countries, mostly OECD members, I expect the first latent variable to prevail. That is, coefficients for the "Declined to answer" should be more or less in line with those voting for Putin. I test this using binary outcome models, dichotomizing the responses by candidates and declined to answer. I also estimate the Putin + Declined to answer category, representing an extreme counterfactual where everyone who censored themselves actually voted for Putin.

I use imputed (city-level fixed effect) data and estimate first pooled linear models.

```
impbindata <- data_naive |>
  select(vote_putin, vote_davankov, vote_spoiled, vote_opposition, vote_declined, vote_putin_declined,
         sex, age_bin, time_to_vs.less_than_hour, out_of_Russia_time,
```

```
        result_trust_bin, countryname_en, city_en) |>
  mutate(across(c(time_to_vs.less_than_hour, countryname_en, city_en), ~ as.factor(.)),
        across(everything(), ~ as.integer(.)))

imp4dry <- mice(impbindata, seed = 1, maxit = 0)
pred4 <- imp4dry$predictorMatrix

pred4[, "city_en"] <- -2
pred4[, "countryname_en"] <- 1
pred4[, "vote_putin"] <- 0
pred4[, "vote_davankov"] <- 0
pred4[, "vote_spoiled"] <- 0
pred4[, "vote_opposition"] <- 0
pred4[, "vote_declined"] <- 0
pred4[, "vote_putin_declined"] <- 0

# imp4 <- mice(impbindata, seed = 1, predictorMatrix = pred4, method = c("", "", "", "", "", "", "2l.pmm", "2l.pmm", "2l.pmm"

# save(imp4, file = "imp4.RData")
load(here("scripts", "imp4.RData"))

# Converges OK!
plot(imp4, layout = c(2, 5))
```




```
m1a.imp <- with(imp4, lm(vote_putin ~ as.factor(sex)
                        + as.factor(age_bin) + as.factor(time_to_vs.less_than_hour)
                        + as.factor(out_of_Russia_time)
                        + as.factor(result_trust_bin)))
```

```r
m1b.imp <- with(imp4, lm(vote_declined ~ as.factor(sex)
                         + as.factor(age_bin) + as.factor(time_to_vs.less_than_hour)
                         + as.factor(out_of_Russia_time)
                         + as.factor(result_trust_bin)))

m1c.imp <- with(imp4, lm(vote_putin_declined ~ as.factor(sex)
                         + as.factor(age_bin) + as.factor(time_to_vs.less_than_hour)
                         + as.factor(out_of_Russia_time)
                         + as.factor(result_trust_bin)))

m1d.imp <- with(imp4, lm(vote_davankov ~ as.factor(sex)
                         + as.factor(age_bin) + as.factor(time_to_vs.less_than_hour)
                         + as.factor(out_of_Russia_time)
                         + as.factor(result_trust_bin)))

m1e.imp <- with(imp4, lm(vote_spoiled ~ as.factor(sex)
                         + as.factor(age_bin) + as.factor(time_to_vs.less_than_hour)
                         + as.factor(out_of_Russia_time)
                         + as.factor(result_trust_bin)))

m1f.imp <- with(imp4, lm(vote_opposition ~ as.factor(sex)
                         + as.factor(age_bin) + as.factor(time_to_vs.less_than_hour)
                         + as.factor(out_of_Russia_time)
                         + as.factor(result_trust_bin)))


modelsummary(list(pool(m1a.imp), pool(m1b.imp), pool(m1c.imp),
                  pool(m1d.imp), pool(m1e.imp), pool(m1f.imp)),
             title = "Binary outcomes pooled, imputed data", header = F,
             col.names = c("", "Vote Putin", "Decline to Answer",
                           "Vote Putin or Declined", "Vote Davankov", "Spoil the ballot",
                           "Vote Davankov or spoil"), output = "kableExtra", stars = T,
             coef_map = c("(Intercept)" = "Intercept",
                          "as.factor(sex)2" = "Sex: Female",
                          "as.factor(sex)3" = "Sex: Other",
                          "as.factor(age_bin)2" = "Age: 18-24 (ref 25-44)",
                          "as.factor(age_bin)3" = "Age: 45-65 (ref 25-44)",
                          "as.factor(age_bin)4" = "Age: 65 + (ref 25-44)",
                          "as.factor(time_to_vs.less_than_hour)2"
                          = "Took < 1 hour to get to the voting station",
                          "as.factor(out_of_Russia_time)2"
                          = "Moved after March 2019 but before March 2022 (ref before 2014)",
                          "as.factor(out_of_Russia_time)3"
                          = "Moved after March 2014 but before March 2019 (ref before 2014)",
                          "as.factor(out_of_Russia_time)4"
                          = "Moved after March 2022 (ref before 2014)",
                          "as.factor(out_of_Russia_time)5"
                          = "Didn't move - tourist, lives in Russia (ref before 2014)",
                          "as.factor(result_trust_bin)2"
                          = "Trust in the result: Don't know (ref Yes)",
                          "as.factor(result_trust_bin)3"
                          = "Trust in the result: No (ref Yes)"),
             booktabs = T)|>
  kable_styling(latex_options = c("scale_down", "hold_position"))
```

The results are not completely in line with list-wise deletion binary models. Namely the model gives a highly unexpected result of older voters being less likely to choose Putin, effect sizes and significance levels also vary.

Table 9: Binary outcomes pooled, imputed data

| | Vote Putin | Decline to Answer | Vote Putin or Declined | Vote Davankov | Spoil the ballot | Vote Davankov or spoil |
|---|---|---|---|---|---|---|
| Intercept | 0.599*** | 0.162*** | 0.761*** | 0.209*** | 0.010+ | 0.220*** |
| | (0.006) | (0.008) | (0.006) | (0.007) | (0.005) | (0.006) |
| Sex: Female | 0.015*** | 0.005 | 0.019*** | −0.040*** | 0.024*** | −0.016*** |
| | (0.002) | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) |
| Sex: Other | 0.007 | −0.047** | −0.040* | −0.075*** | 0.103*** | 0.028 |
| | (0.015) | (0.017) | (0.019) | (0.021) | (0.017) | (0.019) |
| Age: 18-24 (ref 25-44) | −0.009* | −0.014** | −0.023*** | 0.022*** | −0.004 | 0.018*** |
| | (0.004) | (0.005) | (0.004) | (0.006) | (0.005) | (0.005) |
| Age: 45-65 (ref 25-44) | 0.022*** | 0.185*** | 0.207*** | −0.224*** | 0.022*** | −0.202*** |
| | (0.003) | (0.005) | (0.004) | (0.005) | (0.004) | (0.004) |
| Age: 65 + (ref 25-44) | −0.022*** | 0.280*** | 0.258*** | −0.238*** | −0.013* | −0.251*** |
| | (0.004) | (0.006) | (0.006) | (0.007) | (0.006) | (0.006) |
| Took < 1 hour to get to the voting station | 0.004 | 0.013* | 0.017*** | −0.015*** | −0.005 | −0.019*** |
| | (0.004) | (0.005) | (0.003) | (0.004) | (0.003) | (0.003) |
| Moved after March 2019 but before March 2022 (ref before 2014) | −0.039*** | −0.033*** | −0.072*** | 0.101*** | −0.028*** | 0.073*** |
| | (0.005) | (0.007) | (0.005) | (0.006) | (0.005) | (0.006) |
| Moved after March 2014 but before March 2019 (ref before 2014) | −0.029*** | −0.015 | −0.044*** | 0.037*** | 0.007 | 0.044*** |
| | (0.005) | (0.009) | (0.006) | (0.007) | (0.005) | (0.007) |
| Moved after March 2022 (ref before 2014) | −0.044*** | −0.046*** | −0.090*** | 0.158*** | −0.058*** | 0.100*** |
| | (0.004) | (0.006) | (0.004) | (0.005) | (0.004) | (0.004) |
| Didn't move - tourist, lives in Russia (ref before 2014) | 0.042*** | −0.047*** | −0.005 | 0.053*** | −0.053*** | −0.001 |
| | (0.008) | (0.010) | (0.009) | (0.010) | (0.008) | (0.009) |
| Trust in the result: Don't know (ref Yes) | −0.519*** | 0.246*** | −0.272*** | 0.187*** | 0.056*** | 0.243*** |
| | (0.005) | (0.010) | (0.010) | (0.009) | (0.007) | (0.009) |
| Trust in the result: No (ref Yes) | −0.565*** | −0.036*** | −0.601*** | 0.341*** | 0.242*** | 0.584*** |
| | (0.003) | (0.005) | (0.004) | (0.005) | (0.004) | (0.004) |
| Num.Obs. | 69 261 | 69 261 | 69 261 | 69 261 | 69 261 | 69 261 |
| Num.Imp. | 5 | 5 | 5 | 5 | 5 | 5 |
| R2 | 0.486 | 0.119 | 0.523 | 0.270 | 0.069 | 0.487 |
| R2 Adj. | 0.486 | 0.119 | 0.523 | 0.269 | 0.069 | 0.486 |

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

## Introducing fixed effects

I now estimate the same binary models but with city-level fixed effects. I think a minimalist multi-level approach with a single fixed effect is probably enough in this case considering the complexities of imputation.

```
m1a.ml <- with(imp4, lm(vote_putin ~ as.factor(sex)
                         + as.factor(age_bin) + as.factor(time_to_vs.less_than_hour)
                         + as.factor(out_of_Russia_time)
                         + as.factor(result_trust_bin) + as.factor(city_en)))

m1b.ml <- with(imp4, lm(vote_declined ~ as.factor(sex)
                         + as.factor(age_bin) + as.factor(time_to_vs.less_than_hour)
                         + as.factor(out_of_Russia_time)
                         + as.factor(result_trust_bin) + as.factor(city_en)))

m1c.ml <- with(imp4, lm(vote_putin_declined ~ as.factor(sex)
                         + as.factor(age_bin) + as.factor(time_to_vs.less_than_hour)
                         + as.factor(out_of_Russia_time)
                         + as.factor(result_trust_bin) + as.factor(city_en)))

m1d.ml <- with(imp4, lm(vote_davankov ~ as.factor(sex)
                         + as.factor(age_bin) + as.factor(time_to_vs.less_than_hour)
                         + as.factor(out_of_Russia_time)
                         + as.factor(result_trust_bin) + as.factor(city_en)))

m1e.ml <- with(imp4, lm(vote_spoiled ~ as.factor(sex)
                         + as.factor(age_bin) + as.factor(time_to_vs.less_than_hour)
                         + as.factor(out_of_Russia_time)
                         + as.factor(result_trust_bin) + as.factor(city_en)))
```

```
m1f.ml <- with(imp4, lm(vote_opposition ~ as.factor(sex)
                        + as.factor(age_bin) + as.factor(time_to_vs.less_than_hour)
                        + as.factor(out_of_Russia_time)
                        + as.factor(result_trust_bin) + as.factor(city_en)))

modelsummary(list(pool(m1a.ml), pool(m1b.ml), pool(m1c.ml),
                  pool(m1d.ml), pool(m1e.ml), pool(m1f.ml)),
             title = "Binary outcomes multi-level, imputed data", header = F,
             col.names = c("", "Vote Putin", "Decline to Answer",
                           "Vote Putin or Declined", "Vote Davankov", "Spoil the ballot",
                           "Vote Davankov or spoil"), output = "kableExtra", stars = T,
             coef_map = c("(Intercept)" = "Intercept",
                          "as.factor(sex)2" = "Sex: Female",
                          "as.factor(sex)3" = "Sex: Other",
                          "as.factor(age_bin)2" = "Age: 18-24 (ref 25-44)",
                          "as.factor(age_bin)3" = "Age: 45-65 (ref 25-44)",
                          "as.factor(age_bin)4" = "Age: 65 + (ref 25-44)",
                          "as.factor(time_to_vs.less_than_hour)2"
                          = "Took < 1 hour to get to the voting station",
                          "as.factor(out_of_Russia_time)2"
                          = "Moved after March 2019 but before March 2022 (ref before 2014)",
                          "as.factor(out_of_Russia_time)3"
                          = "Moved after March 2014 but before March 2019 (ref before 2014)",
                          "as.factor(out_of_Russia_time)4"
                          = "Moved after March 2022 (ref before 2014)",
                          "as.factor(out_of_Russia_time)5"
                          = "Didn't move - tourist, lives in Russia (ref before 2014)",
                          "as.factor(result_trust_bin)2"
                          = "Trust in the result: Don't know (ref Yes)",
                          "as.factor(result_trust_bin)3"
                          = "Trust in the result: No (ref Yes)"),
             booktabs = T)|>
  kable_styling(latex_options = c("scale_down", "hold_position"))
```

Now let's update the multinomial model by also introducing city fixed effects. Due to high computational costs I do not do any more multinomial models in this problem set.

```
fit2_ml <- with(imp2, multinom(as.factor(vote) ~ as.factor(sex)
                                + as.factor(age_bin)
                                + as.factor(time_to_vs.less_than_hour)
                                + as.factor(out_of_Russia_time)
                                + as.factor(result_trust_bin)
                                + as.factor(city_en)))


ms_fit2.ml <- pool(fit2_ml)

save(ms_fit2.ml, file = "ms_fit2.ml.RData")


load("ms_fit2.ml.RData")

modelsummary(ms_fit2.ml, shape = term + statistic ~ response,
             col.names = c("", "Davankov", "Haritonov", "Slutsky",
                           "Spoiled ballot", "Tore up/took"),
             title = "Multi-level multinomial model with city fixed effect imputation",
             output = "kableExtra", stars = T,
```

Table 10: Binary outcomes multi-level, imputed data

| | Vote Putin | Decline to Answer | Vote Putin or Declined | Vote Davankov | Spoil the ballot | Vote Davankov or spoil |
|---|---|---|---|---|---|---|
| Intercept | 0.613*** | 0.080*** | 0.694*** | 0.327*** | −0.038*** | 0.289*** |
| | (0.008) | (0.011) | (0.010) | (0.012) | (0.010) | (0.010) |
| Sex: Female | 0.012*** | 0.007* | 0.019*** | −0.036*** | 0.021*** | −0.015*** |
| | (0.002) | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) |
| Sex: Other | 0.014 | −0.026 | −0.011 | −0.087*** | 0.086*** | 0.000 |
| | (0.015) | (0.017) | (0.018) | (0.020) | (0.017) | (0.018) |
| Age: 18-24 (ref 25-44) | −0.006 | −0.016*** | −0.022*** | 0.028*** | −0.011* | 0.017*** |
| | (0.004) | (0.005) | (0.004) | (0.006) | (0.005) | (0.005) |
| Age: 45-65 (ref 25-44) | 0.028*** | 0.164*** | 0.192*** | −0.209*** | 0.022*** | −0.188*** |
| | (0.003) | (0.004) | (0.004) | (0.005) | (0.004) | (0.004) |
| Age: 65 + (ref 25-44) | −0.005 | 0.248*** | 0.243*** | −0.222*** | −0.015* | −0.237*** |
| | (0.004) | (0.006) | (0.006) | (0.007) | (0.007) | (0.006) |
| Took < 1 hour to get to the voting station | −0.004 | 0.008+ | 0.004 | −0.007+ | 0.001 | −0.006 |
| | (0.004) | (0.004) | (0.004) | (0.004) | (0.003) | (0.004) |
| Moved after March 2019 but before March 2022 (ref before 2014) | −0.046*** | −0.010 | −0.057*** | 0.081*** | −0.023*** | 0.058*** |
| | (0.005) | (0.007) | (0.006) | (0.007) | (0.005) | (0.006) |
| Moved after March 2014 but before March 2019 (ref before 2014) | −0.028*** | −0.009 | −0.037*** | 0.032*** | 0.005 | 0.038*** |
| | (0.005) | (0.008) | (0.006) | (0.007) | (0.005) | (0.007) |
| Moved after March 2022 (ref before 2014) | −0.066*** | −0.011* | −0.077*** | 0.124*** | −0.037*** | 0.087*** |
| | (0.004) | (0.005) | (0.004) | (0.005) | (0.005) | (0.004) |
| Didn't move - tourist, lives in Russia (ref before 2014) | 0.010 | −0.004 | 0.005 | 0.024* | −0.034*** | −0.010 |
| | (0.007) | (0.010) | (0.009) | (0.010) | (0.008) | (0.009) |
| Trust in the result: Don't know (ref Yes) | −0.482*** | 0.196*** | −0.285*** | 0.206*** | 0.049*** | 0.255*** |
| | (0.005) | (0.010) | (0.009) | (0.009) | (0.008) | (0.009) |
| Trust in the result: No (ref Yes) | −0.539*** | −0.043*** | −0.581*** | 0.336*** | 0.229*** | 0.564*** |
| | (0.004) | (0.005) | (0.004) | (0.005) | (0.005) | (0.004) |
| Num.Obs. | 69 261 | 69 261 | 69 261 | 69 261 | 69 261 | 69 261 |
| Num.Imp. | 5 | 5 | 5 | 5 | 5 | 5 |
| R2 | 0.502 | 0.167 | 0.543 | 0.283 | 0.076 | 0.505 |
| R2 Adj. | 0.501 | 0.166 | 0.543 | 0.282 | 0.075 | 0.504 |

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

```
          coef_map = c("(Intercept)" = "Intercept",
                       "as.factor(sex)2" = "Sex: Female",
                       "as.factor(sex)3" = "Sex: Other",
                       "as.factor(age_bin)2" = "Age: 18-24 (ref 25-44)",
                       "as.factor(age_bin)3" = "Age: 45-65 (ref 25-44)",
                       "as.factor(age_bin)4" = "Age: 65 + (ref 25-44)",
                       "as.factor(time_to_vs.less_than_hour)2"
                       = "Took < 1 hour to get to the voting station",
                       "as.factor(out_of_Russia_time)2"
                       = "Moved after March 2019 but before March 2022 (ref before 2014)",
                       "as.factor(out_of_Russia_time)3"
                       = "Moved after March 2014 but before March 2019 (ref before 2014)",
                       "as.factor(out_of_Russia_time)4"
                       = "Moved after March 2022 (ref before 2014)",
                       "as.factor(out_of_Russia_time)5"
                       = "Didn't move - tourist, lives in Russia (ref before 2014)",
                       "as.factor(result_trust_bin)2"
                       = "Trust in the result: Don't know (ref Yes)",
                       "as.factor(result_trust_bin)3"
                       = "Trust in the result: No (ref Yes)"),
          gof_map = "none",
          booktabs = T,
          add_rows = data.frame(term = c("Observations", "Imputations", "Imputation Model"),
                                Davankov = c("56 307", "5", "PMM with city FE"),
                                Haritonov = c("56 307", "5", "PMM with city FE"),
                                Slutsky = c("56 307", "5", "PMM with city FE"),
                                `Spoiled ballot` = c("56 307", "5", "PMM with city FE"),
                                `Tore up/Took` = c("56 307", "5", "PMM with city FE"))
         ) |>
  kable_styling(latex_options = c("scale_down", "hold_position"))
```

Table 11: Multi-level multinomial model with city fixed effect imputation

| | (1) | | | | |
|---|---|---|---|---|---|
| | Davankov | Haritonov | Slutsky | Spoiled ballot | Tore up/took |
| Intercept | −2.077*** | −3.743*** | −5.301*** | −3.967*** | −12.443*** |
| | (0.271) | (0.317) | (0.539) | (0.286) | (2.926) |
| Sex: Female | −0.316* | −0.390** | −0.201+ | −0.119 | −0.348+ |
| | (0.109) | (0.121) | (0.111) | (0.113) | (0.190) |
| Sex: Other | −0.242 | 0.246 | 0.251 | 0.521 | 0.613 |
| | (0.419) | (0.447) | (0.663) | (0.338) | (0.663) |
| Age: 18-24 (ref 25-44) | 0.321* | 0.256 | 0.383 | 0.220 | 0.562* |
| | (0.141) | (0.267) | (0.296) | (0.150) | (0.244) |
| Age: 45-65 (ref 25-44) | −1.730*** | −1.034*** | −1.440*** | −0.934*** | −0.642* |
| | (0.101) | (0.197) | (0.278) | (0.096) | (0.278) |
| Age: 65 + (ref 25-44) | −2.199*** | −0.761*** | −1.250*** | −1.307*** | −1.370+ |
| | (0.184) | (0.198) | (0.269) | (0.202) | (0.681) |
| Took < 1 hour to get to the voting station | −0.065 | 0.066 | 0.058 | −0.027 | −0.009 |
| | (0.153) | (0.174) | (0.253) | (0.152) | (0.235) |
| Moved after March 2019 but before March 2022 (ref before 2014) | 1.307*** | 0.788*** | 0.890** | 0.917*** | 1.219+ |
| | (0.136) | (0.177) | (0.223) | (0.163) | (0.533) |
| Moved after March 2014 but before March 2019 (ref before 2014) | 0.649** | 0.220 | 0.554** | 0.459* | 0.725 |
| | (0.147) | (0.185) | (0.189) | (0.157) | (0.426) |
| Moved after March 2022 (ref before 2014) | 1.892*** | 0.917*** | 1.034*** | 1.382*** | 1.462** |
| | (0.111) | (0.147) | (0.177) | (0.123) | (0.373) |
| Didn't move - tourist, lives in Russia (ref before 2014) | 0.586*** | 0.110 | 0.792*** | 0.062 | 0.739 |
| | (0.120) | (0.278) | (0.224) | (0.151) | (0.580) |
| Trust in the result: Don't know (ref Yes) | 3.547*** | 2.762*** | 3.125*** | 3.054*** | 5.714 |
| | (0.143) | (0.238) | (0.227) | (0.281) | (3.702) |
| Trust in the result: No (ref Yes) | 5.474*** | 4.196*** | 3.742*** | 6.140*** | 9.495** |
| | (0.078) | (0.175) | (0.225) | (0.131) | (2.872) |
| Observations | 56 307 | 56 307 | 56 307 | 56 307 | 56 307 |
| Imputations | 5 | 5 | 5 | 5 | 5 |
| Imputation Model | PMM with city FE | PMM with city FE | PMM with city FE | PMM with city FE | PMM with city FE |

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

The final fixed effects model produces more moderate estimates with significance falling for gender but remaining everywherer else.

Overall it is hard to say whether data imputation did the models good. On one hand, I am confident that effective imputation in this case should be both polytonomous and multi-level, something that is not quite possible. On the other hand the data almost certainly is not MAR or MCAR so much more nuance has to go into trying to recover the censored data and in particular vote choice.

I would like to try and fit the Heckman model which would basically try and categorize "Declined to answer" based on country characteristics and demographics and then estimate the main model. I also leave for future work the imputation into ordered logit (didn't figure out how to do it efficiently).

I would say that the huge sample size and clear hierarchies are beneficial to this research project while lack of variables and the need to aggregate if one wants to have more of them are detrimental. In particular the already severe NI missing data problems are amplified by having a lot of potential determinants unobserved. I read a couple of papers on estimation of latent traits for imputation in those situations however no clear practical approach emerged.

# Appendix

```
kable(table(ep_raw_clean$countryname_en, ep_raw_clean$age_bin)[1:22, ], booktabs = T,
      label = "at1") |>
  kable_styling(latex_options = c("scale_down", "hold_position"),
                bootstrap_options = "striped")
kable(table(ep_raw_clean$countryname_en, ep_raw_clean$age_bin)[23:44, ], booktabs = T) |>
  kable_styling(latex_options = c("scale_down", "hold_position"),
                bootstrap_options = "striped")
```

| | 18-24 | 25-44 | 45-64 | 65+ | Declined to answer |
|---|---|---|---|---|---|
| Argentina | 85 | 1009 | 118 | 144 | 0 |
| Armenia | 408 | 2864 | 400 | 231 | 0 |
| Australia | 0 | 10 | 6 | 6 | 0 |
| Austria | 595 | 1760 | 481 | 179 | 0 |
| Belgium | 56 | 383 | 158 | 42 | 0 |
| Canada | 139 | 779 | 589 | 266 | 0 |
| Costa Rica | 3 | 45 | 26 | 18 | 0 |
| Croatia | 21 | 201 | 72 | 19 | 0 |
| Cyprus | 312 | 3898 | 1096 | 357 | 0 |
| Czechia | 389 | 828 | 72 | 18 | 3 |
| Denmark | 35 | 436 | 126 | 23 | 0 |
| Estonia | 45 | 468 | 537 | 458 | 0 |
| Finland | 248 | 1615 | 577 | 246 | 0 |
| France | 381 | 1795 | 533 | 109 | 0 |
| Germany | 397 | 2508 | 1465 | 593 | 0 |
| Great Britain | 249 | 1228 | 219 | 30 | 0 |
| Greece | 99 | 364 | 394 | 219 | 0 |
| Hungary | 207 | 851 | 267 | 112 | 0 |
| Ireland | 51 | 413 | 155 | 10 | 0 |
| Israel | 370 | 1483 | 559 | 623 | 0 |
| Italy | 270 | 858 | 558 | 121 | 0 |
| Japan | 75 | 431 | 118 | 18 | 0 |

| | 18-24 | 25-44 | 45-64 | 65+ | Declined to answer |
|---|---|---|---|---|---|
| Kazakhstan | 233 | 2063 | 281 | 98 | 0 |
| Kyrgyzstan | 56 | 291 | 102 | 68 | 0 |
| Lithuania | 45 | 408 | 85 | 81 | 0 |
| Luxembourg | 31 | 524 | 276 | 46 | 0 |
| Moldova | 62 | 295 | 334 | 179 | 0 |
| Montenegro | 67 | 1143 | 202 | 18 | 0 |
| Netherlands | 118 | 941 | 89 | 19 | 0 |
| New Zealand | 0 | 0 | 0 | 0 | 0 |
| Norway | 61 | 453 | 220 | 70 | 0 |
| Poland | 191 | 1117 | 207 | 33 | 0 |
| Portugal | 21 | 617 | 98 | 17 | 0 |
| Serbia | 278 | 3322 | 255 | 51 | 0 |
| Slovakia | 204 | 308 | 80 | 25 | 0 |
| Spain | 141 | 1069 | 163 | 34 | 0 |
| Sweden | 79 | 634 | 235 | 61 | 0 |
| Switzerland | 144 | 978 | 504 | 113 | 0 |
| Thailand | 71 | 458 | 46 | 4 | 0 |
| Türkiye | 164 | 940 | 149 | 19 | 0 |
| UAE | 112 | 1099 | 264 | 38 | 0 |
| USA | 126 | 947 | 388 | 91 | 0 |
| Uzbekistan | 114 | 1368 | 703 | 262 | 0 |
| Vietnam | 28 | 186 | 20 | 4 | 0 |