# EDA: Voter Turnout and Race Competitiveness

Kamal Moravej Jahromi  Chad Neald  Rafael Pilliard Hellwig  Yuan Xiong

19/11/2020

## Intro

In this exploratory data analysis (EDA), we will take a first look at our elections data. Our research question of interest is whether more competitive elections are associated with greater voter turnout in British Columbia. To answer this, we will be using open data from Elections BC. Scripts for downloading these data are provided in the `src/` directory.

## Load the data

Let's start by loading the data. We are using two open data sources: the provincial voting results, and the provincial voter participation. We will load these in here and name them `pvr` and `pvp`, respectively.

```
# Load packages
library(tidyverse)

# Set defaults and seeds
theme_set(ggthemes::theme_fivethirtyeight() +
              theme(axis.title = element_text()))
set.seed(1)
```

```
# Read-in the elections results data
f1 <- here::here("data", "raw", "provincial_voting_results.csv")
pvr <- janitor::clean_names(read_csv(f1))
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   EVENT_NAME = col_character(),
##   EVENT_YEAR = col_double(),
##   ED_ABBREVIATION = col_character(),
##   ED_NAME = col_character(),
##   VA_CODE = col_character(),
##   EDVA_CODE = col_character(),
##   ADVANCE_VOTING_LOCATION = col_character(),
##   ADDRESS_STANDARD_ID = col_double(),
##   VOTING_OPPORTUNITY = col_character(),
##   CANDIDATE = col_character(),
##   ELECTED = col_character(),
##   AFFILIATION = col_character(),
##   VOTES_CONSIDERED = col_double(),
```

```
##   VOTE_CATEGORY = col_character(),
##   COMBINED_INDICATOR = col_character(),
##   RESULTS_REPORTED_UNDER = col_character()
## )
```

```
# Read-in the voter participation data
f2 <- here::here("data", "raw", "provincial_voter_participation_by_age_group.csv")
pvp <- janitor::clean_names(read_csv(f2))
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   EVENT_NAME = col_character(),
##   EVENT_YEAR = col_double(),
##   ED_ABBREVIATION = col_character(),
##   ED_NAME = col_character(),
##   AGE_GROUP = col_character(),
##   PARTICIPATION = col_number(),
##   REGISTERED_VOTERS = col_number(),
##   EVENT_DATE_TEXT = col_character()
## )
```

Let's take a look a sample of rows from our voter participation dataset by creating an exploratory data table:

```
sample_n(pvp, 10) %>%
  knitr::kable()
```

| event_name | event_year | ed_abbreviation | ed_name | age_group | participation | registered_voters | event_date_text |
|---|---|---|---|---|---|---|---|
| General Election 2013 | 2013 | FLA | Fort Langley-Aldergrove | 65-74 | 4223 | 5536 | 05/14/2013 |
| General Election 2017 | 2017 | MAP | Maple Ridge-Pitt Meadows | 75+ | 2411 | 3423 | 05/09/2017 |
| General Election 2005 | 2005 | PRN | Prince George North | 18-24 | 1223 | 2359 | 05/17/2005 |
| General Election 2017 | 2017 | BNN | Burnaby North | 75+ | 2497 | 3918 | 05/09/2017 |
| General Election 2013 | 2013 | SWH | Surrey-Whalley | 55-64 | 3542 | 6278 | 05/14/2013 |
| General Election 2017 | 2017 | ABM | Abbotsford-Mission | 65-74 | 3855 | 5181 | 05/09/2017 |
| General Election 2005 | 2005 | BNN | Burnaby North | 25-34 | 3241 | 6800 | 05/17/2005 |
| General Election 2009 | 2009 | WCA | West Vancouver-Capilano | 25-34 | 1163 | 3689 | 05/12/2009 |
| General Election 2009 | 2009 | KAS | Kamloops-South Thompson | 18-24 | 1666 | 4436 | 05/12/2009 |
| General Election 2009 | 2009 | PEN | Penticton | 35-44 | 2266 | 5325 | 05/12/2009 |

Let's do the same for our election results data. Here, we only show a sub-selection of the columns.

```
pvr %>%
  sample_n(10) %>%
  select(ed_name, event_year, event_name, affiliation, vote_category,
         votes_considered) %>%
  knitr::kable()
```

| ed_name | event_year | event_name | affiliation | vote_category | votes_considered |
|---|---|---|---|---|---|
| West Kootenay-Boundary | 2005 | General Election 2005 | NA | Valid | 0 |
| Saanich South | 2009 | General Election 2009 | BC Green Party | Valid | 10 |
| Shuswap | 2009 | General Election 2009 | Conservative | Valid | 8 |
| Vancouver-Mount Pleasant | 2005 | General Election 2005 | BC Marijuana Party | Valid | 2 |
| Victoria-Beacon Hill | 2013 | General Election 2013 | BC Green Party | Valid | 33 |
| Westside-Kelowna | 2013 | 2013 Westside-Kelowna By-election | BC NDP | Valid | 4 |
| Westside-Kelowna | 2009 | General Election 2009 | Conservative | Valid | 4 |
| Nanaimo | 2017 | General Election 2017 | BC NDP | Valid | 97 |
| Surrey-Cloverdale | 2013 | General Election 2013 | NA | Rejected | 1 |
| Okanagan-Vernon | 2005 | General Election 2005 | NA | Valid | 8 |

Let's create some EDA profile reports. These will be created as PDFs in the `eda` directory, and will include marginal plots, basic descriptive statistics, and information about missing data. We'll use the `dataMaid` package for this.

```
# Create PDF profile reports
dataMaid::makeDataReport(pvr, file = here::here("eda", "profile_pvr.Rmd"),
                         replace = TRUE)
dataMaid::makeDataReport(pvp, file = here::here("eda", "profile_pvp.Rmd"),
                         replace = TRUE)
```

## Data Cleaning and Transformation

The data is relatively clean, but too granular for our research question. Let's start by aggregating the voter participation so that each row (unit of analysis) represents an Electoral District (ED) for a given electoral event. We'll add a new column for the `turnout` by dividing the number of electors who participated by the total number of registered voters:

```
# Aggregate participation by event and electoral district
pvp_agg <- pvp %>%
    group_by(event_name, ed_name) %>%
    summarise(across(participation:registered_voters, sum),
              .groups = "drop") %>%
    mutate(turnout = participation / registered_voters)
```

We can also aggregate the voting results data. As we do this, we will also compute some variables for each ED and electoral event, such as `competitiveness`. We operationalized the latter as the point difference in vote share between the runner-up and the winner. For example, if in a given district, a party wins with 42% of the votes, and the runner up has 30%, this would be a 12-point difference.

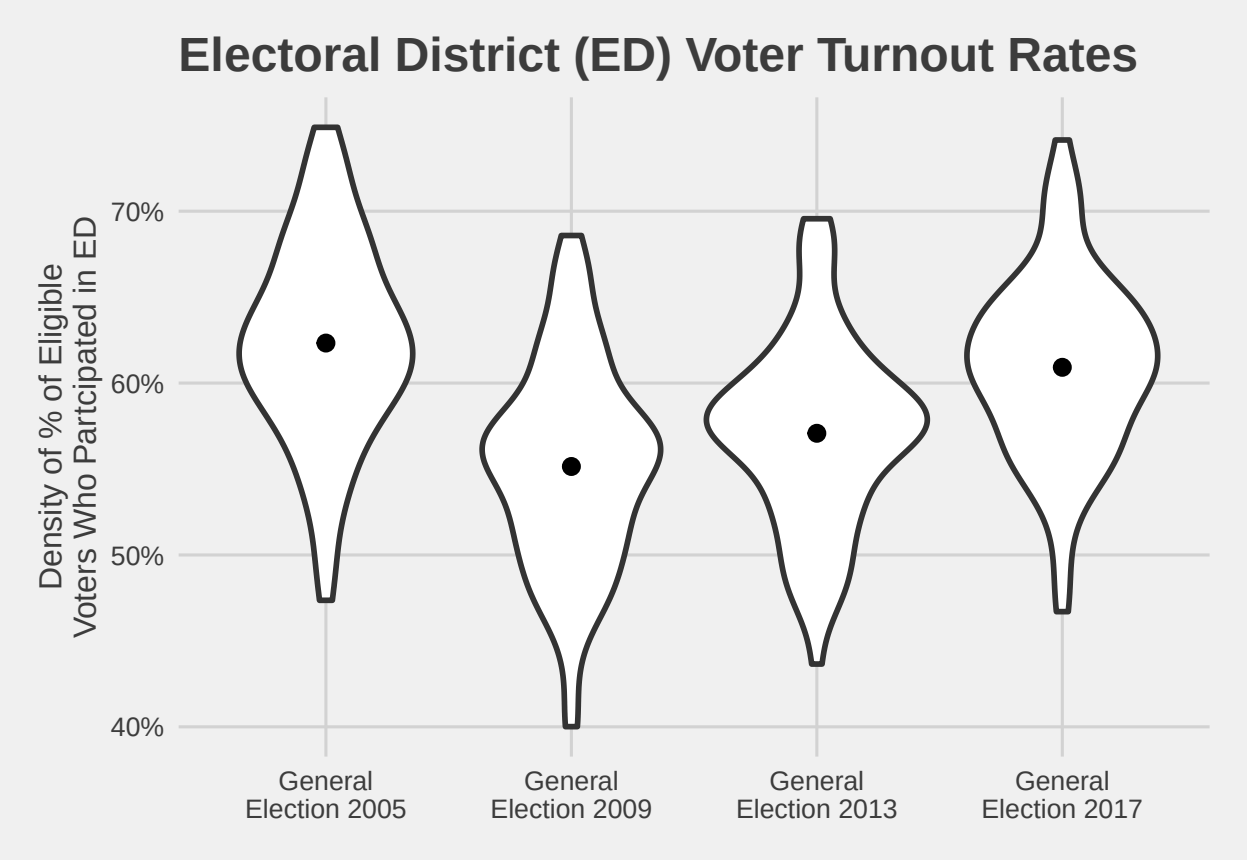
Finally, we will also join-in our voter turnout data.

```
# Aggregate election results by event and electoral district.
pvr_agg <- pvr %>%
    filter(vote_category == "Valid") %>%
    group_by(event_name, ed_name, affiliation) %>%
    summarise(votes = sum(votes_considered),
              .groups = "drop_last") %>%
    arrange(event_name, ed_name, desc(votes)) %>%
    mutate(vote_share = votes / sum(votes),
           rank = row_number(),
           vote_trail = votes - first(votes) ,
           share_trail = vote_share - first(vote_share),
           vote_diff = nth(vote_trail, 2),
           competitiveness = nth(share_trail, 2),
           winning_party = nth(affiliation, 1)) %>%
    nest(candidates = c(affiliation, votes, vote_share, vote_trail,
                        share_trail, rank)) %>%
    ungroup %>%
    left_join(pvp_agg, by = c("event_name", "ed_name"))
```

## Analysis

Now, let's plot our dependent variable: voter turnout. It appears that we have data on this at the electoral district for the General Elections held in 2005, 2009, 2013, and 2017 (but not sufficient amounts of data for by-elections).

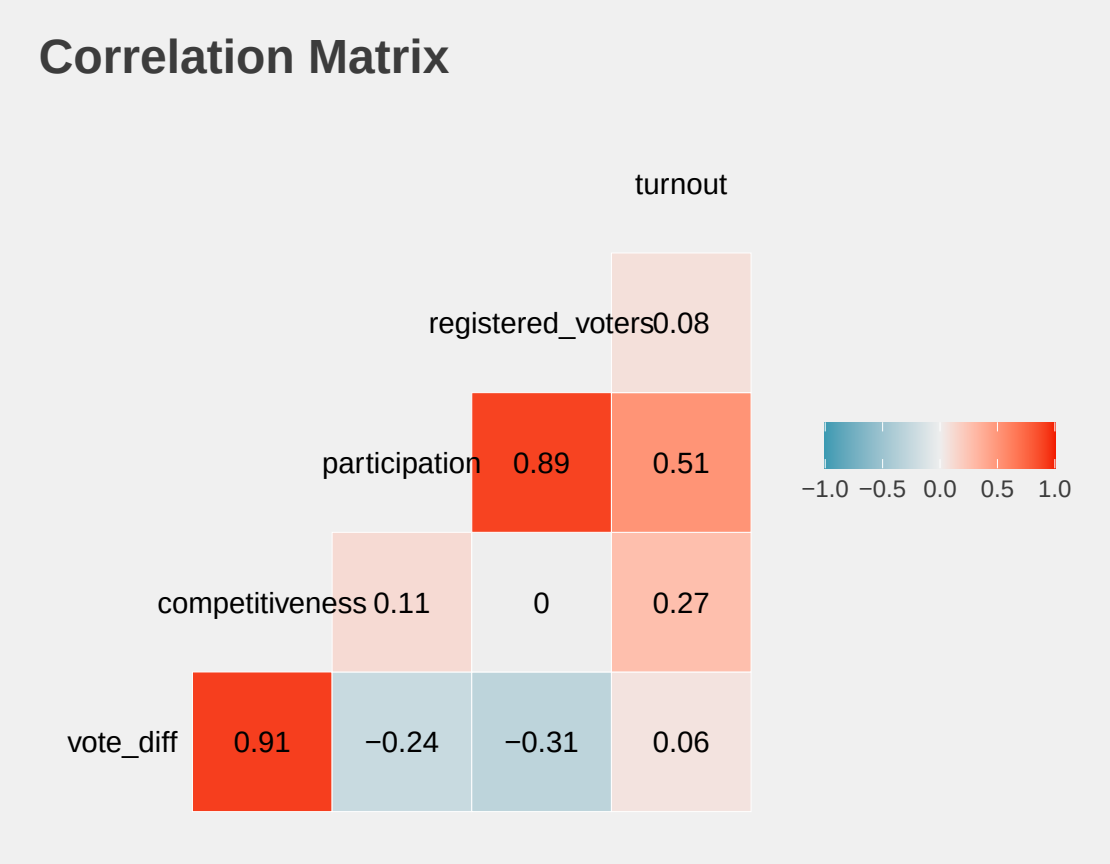
```
# Violin plots of voter turnout
pvp_agg %>%
    ggplot(aes(y = turnout, x = factor(str_wrap(event_name, 15)))) +
    geom_violin(size = 1) +
    scale_y_continuous(labels = scales::percent_format(1)) +
    labs(title = "Electoral District (ED) Voter Turnout Rates",
         y = "Density of % of Eligible\nVoters Who Partcipated in ED",
         x = NULL) +
    stat_summary(fun = mean)
```




Turnout seems to vary quite a bit from one election to another. That might be something to keep in mind for subsequent analyses, as we may want to control for this factor.

Let's look at some of the other correlations between numeric variables. We are particularly interested in `turnout`:

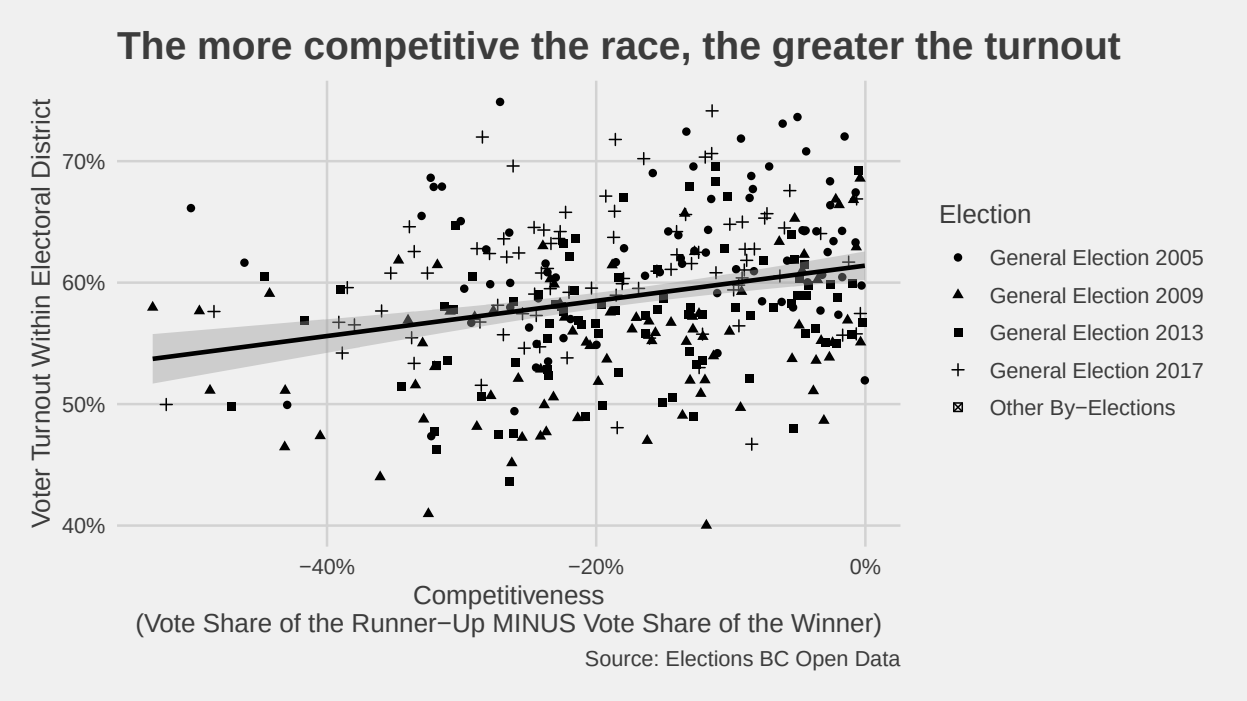
```
pvr_agg %>%
  select(where(is.numeric)) %>%
  GGally::ggcorr(label = TRUE, label_round = 2) +
  labs(title = "Correlation Matrix")
```




We see that there is 0.27 correlation between `turnout` and `competitiveness`. In subsequent analysis, we will test if this correlation is just spurious or statistically significant.

Let's plot these two variables against one another in a scatter plot, and add a trendline:

```
# Scatter plot relating the voter turnout to the competitiveness of a race
pvr_agg %>%
    drop_na(competitiveness) %>%
    mutate(across(event_name, fct_lump, n = 4,
                  other_level = "Other By-Elections")) %>%
    ggplot(aes(x = competitiveness, y = turnout)) +
    geom_point(aes(shape = event_name)) +
    geom_smooth(method = "lm", formula = y ~ x, colour = "black") +
    scale_y_continuous(labels = scales::percent_format(1)) +
    scale_x_continuous(labels = scales::percent_format(1)) +
    labs(title = "The more competitive the race, the greater the turnout",
         caption = "Source: Elections BC Open Data",
         y = "Voter Turnout Within Electoral District",
         x = "Competitiveness\n(Vote Share of the Runner-Up MINUS Vote Share of the Winner)",
         shape = "Election") +
    theme(legend.position = "right", legend.direction = "vertical")
```




As hypothesized, the more competitive a race is, the greater the associated turnout. This is reflected visually in the positive sloping trendline.

## Conclusion

This exploratory data analysis has given us some nice visuals that support our hypothesis. In subsequent analyses, we will test this more formally using regression and/or other statistical tests.