Exploratory Data Analysis

Rafael Pilliard Hellwig

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. To answer this, we will be using open data from Election British Columbia.

Load the data

Let's start by loading the data. We are using two pen data sources from Elections BC: the provincial voting results, and the provincial voter participation. We'll name these `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 the sample sample or rows from our voter participation dataset:

```
sample_n(pvp, 10)
```

```
## # A tibble: 10 x 8
##    event_name event_year ed_abbreviation ed_name age_group participation
##    <chr>           <dbl> <chr>           <chr>   <chr>             <dbl>
##  1 General E~       2013 FLA             Fort L~ 65-74              4223
##  2 General E~       2017 MAP             Maple ~ 75+                2411
##  3 General E~       2005 PRN             Prince~ 18-24              1223
##  4 General E~       2017 BNN             Burnab~ 75+                2497
##  5 General E~       2013 SWH             Surrey~ 55-64              3542
##  6 General E~       2017 ABM             Abbots~ 65-74              3855
##  7 General E~       2005 BNN             Burnab~ 25-34              3241
##  8 General E~       2009 WCA             West V~ 25-34              1163
##  9 General E~       2009 KAS             Kamloo~ 18-24              1666
## 10 General E~       2009 PEN             Pentic~ 35-44              2266
## # ... with 2 more variables: registered_voters <dbl>, event_date_text <chr>
```

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)
```

```
## # A tibble: 10 x 6
##    ed_name   event_year event_name    affiliation vote_category votes_considered
##    <chr>          <dbl> <chr>         <chr>       <chr>                    <dbl>
##  1 West Koo~       2005 General Elec~ <NA>        Valid                        0
```

```
##  2 Saanich ~       2009 General Elec~ BC Green P~ Valid                    10
##  3 Shuswap         2009 General Elec~ Conservati~ Valid                     8
##  4 Vancouve~       2005 General Elec~ BC Marijua~ Valid                     2
##  5 Victoria~       2013 General Elec~ BC Green P~ Valid                    33
##  6 Westside~       2013 2013 Westsid~ BC NDP      Valid                     4
##  7 Westside~       2009 General Elec~ Conservati~ Valid                     4
##  8 Nanaimo         2017 General Elec~ BC NDP      Valid                    97
##  9 Surrey-C~       2013 General Elec~ <NA>        Rejected                  1
## 10 Okanagan~       2005 General Elec~ <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, replace = TRUE)
dataMaid::makeDataReport(pvp, 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 `share_diff`, the difference in share of the vote between the winner and the runner up for a particular race in an electoral district. We'll 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 = first(votes) - votes,
           share_trail = first(vote_share) - vote_share,
           vote_diff = nth(vote_trail, 2),
           share_diff = nth(share_trail, 2),
           winning_party = nth(affiliation, 1)) %>%
    nest(candidates = c(affiliation, votes, vote_share, vote_trail,
                        share_trail, rank)) %>%
```

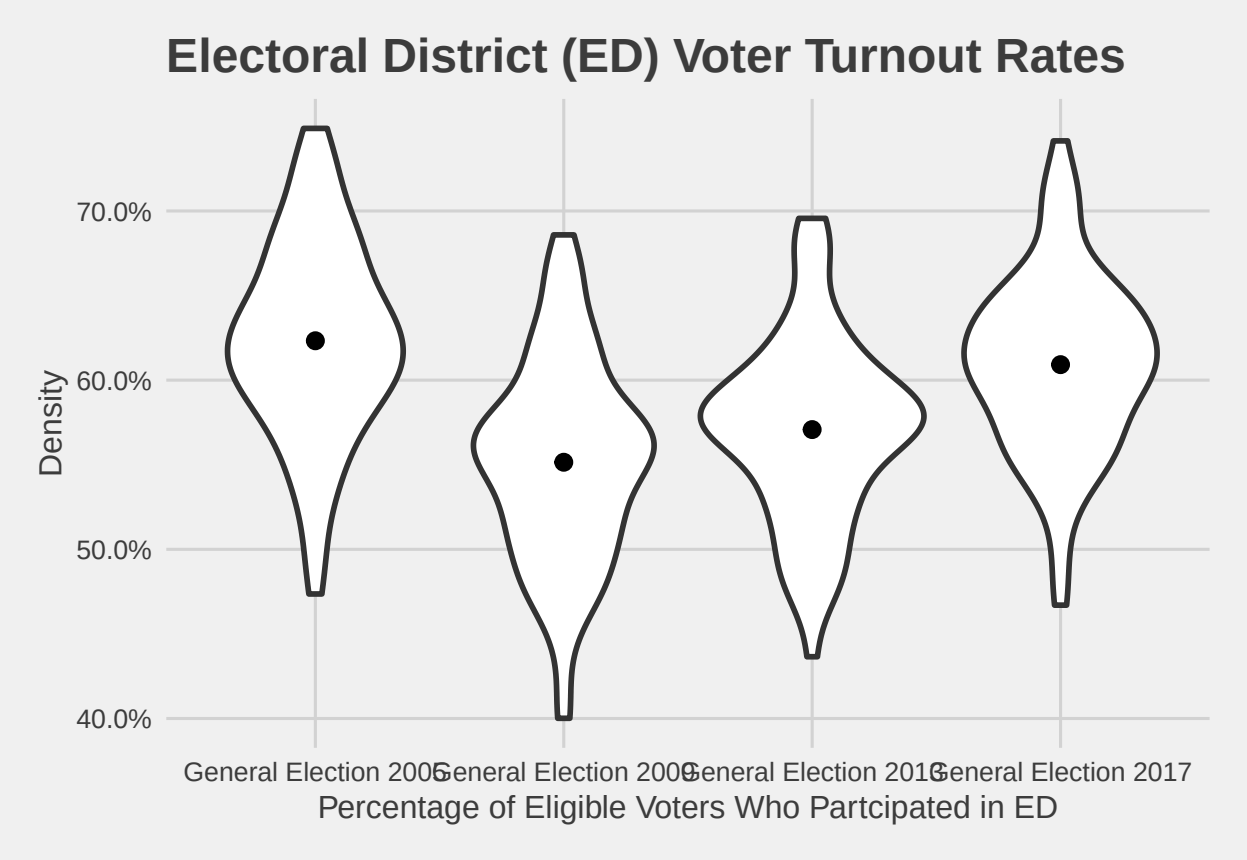
```
    ungroup %>%
    left_join(pvp_agg, by = c("event_name", "ed_name"))
```

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 for by-elections).

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

```
## Warning: Removed 4 rows containing missing values (geom_segment).
```



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.

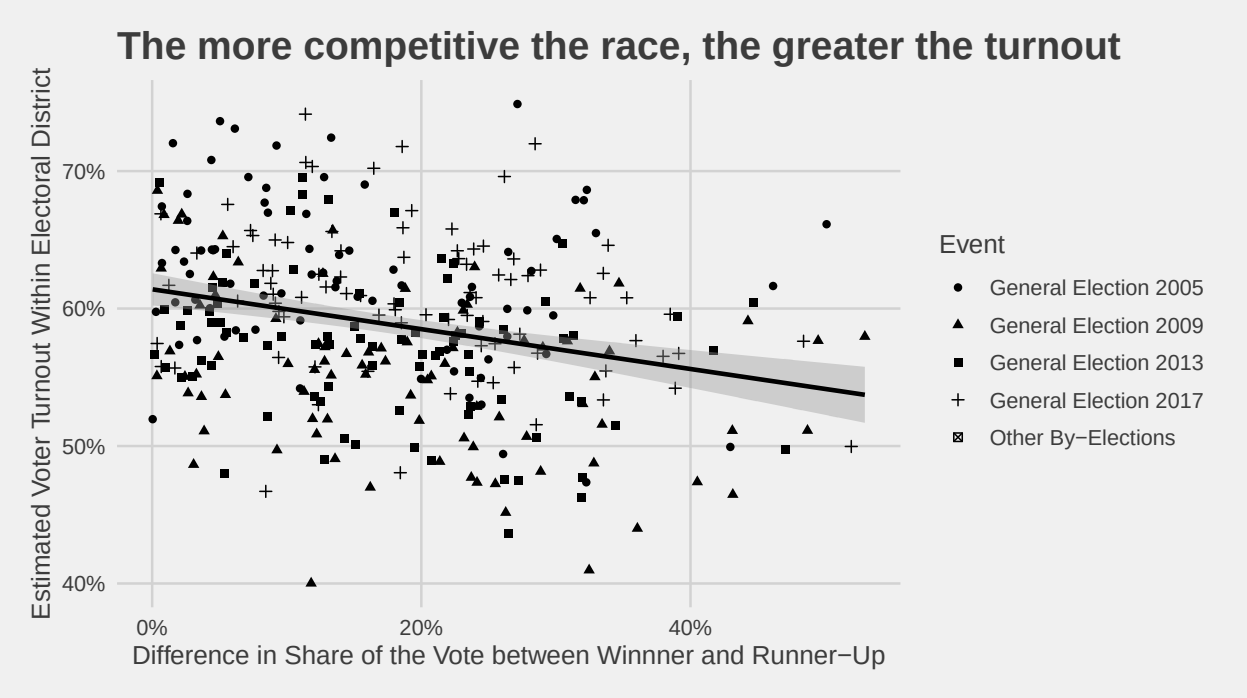
Let's continue to transform our data and visualize the relationship between our two variables of interest: voter turn out (as a percentage of total eligible voters in a district) vs. the competitiveness of a race. Here, we

operationalize the latter concept as the point difference in vote share between the winner and the runner-up. 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.

```
# Scatter plot relating the voter turnout to the competitiveness of a race
pvr_agg %>%
    drop_na(share_diff) %>%
    mutate(across(event_name, fct_lump, n = 4,
                  other_level = "Other By-Elections")) %>%
    ggplot(aes(x = share_diff, 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 = "Estimated Voter Turnout Within Electoral District",
         x = "Difference in Share of the Vote between Winnner and Runner-Up",
         shape = "Event") +
    theme(legend.position = "right", legend.direction = "vertical")
```

```
## Warning: Removed 9 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 9 rows containing missing values (geom_point).
```



As hypothesised, the more competitive a race is, the greater the associated turnout. On the chart, we see a negative sloping trendline, because competitiveness is the additive inverse of the difference in vote share (the x-axis).

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 methods.