17.806: Problem Set 1

Professor: In Song Kim

Please submit both your *write-up* and your *code* electronically to the Learning Modules.[1] This should be completed *before* class begins at 3 pm (late submissions will not be accepted).

Analyzing the impact of police stopping on politcal behavior

In the lecture, we have learned that an increase in the availability of (unstructured) data would advance empirical analysis in political science. In this problem set, you will explore how/whether policing against citizens and against racial minorities affects political behavior of citizens by leveraging a variety of data sources online. In recent years, micro-level administrative data on policing is readily available, which allows us to conduct more reliable analysis from observational data. This problem set particularly focuses on the stop-question-and-frisk (SQF) by the New York Police Department (NYPD).

NYPD discloses geo-located data on all SQF, which offers us a unique opportunity to study the relationship between policing and voting behavior at the fine-grained level. Some critics have argued that SQF is a racially discriminatory policy because people being stopped are overwhelmingly Black and Latino.[2] In particular, more than 50% of SQF are conducted against black people while they only account for a quarter of the population in NYC. Since policy toward policing has been one of the salient issues in recent elections, this problem set will examine how/whether SQF in a community affects voting behavior. Specifically, we will analyze the impact of police stopping on the electoral outcomes using the NYPD SQF data and election data from the 2016 and 2020 presidential elections. For the purpose of this course, these problems will help us practice various skills to collect new data and make use of `regex`.

1. As the first step, we will practice automatically downloading data and structuring the downloaded data. Although the number of files that we download in this question does not require us to develop a pipeline to download the data automatically, researchers often encounter websites with thousands or millions of target data that makes it impossible to download manually. Therefore, this exercise will get you started on this process while you think about your own data collection efforts.

   The NYPD discloses the SQF data at their website (`https://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page`) from years 2003 to 2019. Write codes to automatically download the SQF data from the website. In your codes, create a directory named `data` first, then create a subdirectory named `nypd_stop_data` inside `data`, and store the downloaded data into `nypd_stop_data` (i.e., write a script to automatically create the folders and subfolders and store the downloaded data rather than manually do so).

   Please follow these steps:

[1] Site is unavailable to OCW users.

[2] Milner, Adrienne N., Brandon J. George, and David B. Allison. 2016. "Black and Hispanic men perceived to be large are at increased risk for police frisk, search, and force." PloS One 11.1.

- Identify the URL that contains data
- Analyze the html by extracting tags and attributes that contain data
- Create folders/directories to store data
- Download data (Make sure to wait a bit between downloads because too much traffic may cause a website to crash, and/or the website could identify you as a "bot" and block you from accessing it. To do so, try using the `Sys.sleep` function.)

The directory to contain the data should look like this.

```
Pset1
├─ pset1_solution.Rmd
│
└─ data
   └─ nypd_stop_data
      ├─ sqf-2019.xlsx
      └─ ...
```

Hint

- Use the **rvest** package to analyze html
- Use `download.file` function in the base R to download files
- Use `regex` to identify files to download. Closely examine the html file of the website and make sure not to download irrelevant files. You can check the source code with the web browser (e.g., go to `View/Developer/View Source` in Google Chrome).

2. The downloaded SQF data span multiple years and contain more than one hundred variables that may not be relevant to our analysis. The natural next step is to clean and structure our data. In this problem, we will practice cleaning data programmatically, while the tools we learn can be applied to many settings. The goal is to construct one standard data frame `nypd_data` that contains four variables `year`, `race`, `xcoord`, and `ycoord` from years 2014 to 2019. This data frame will then be used to identify election districts for our empirical analysis.

   Hint: Use `regex` to identify targeted years and unzip corresponding files. For further information about the dataset, refer to the code books available at the NYPD website (`https://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page`).

3. **Extra Credit.** Follow the hint below to map the location of each SQF to corresponding election district based on `xcoord` and `ycoord`.[3] Since cleaning this data requires some knowledge in GIS system, we provide the cleaned version of the data (`nypd_sqf_2014_2019.csv`) for later analysis on the Canvas website.

   Hint

   - NYC publicizes GIS data of election district, which is also available on Canvas.
   - The geographic coordination system used in the SQF data is not based on longitude and latitude. `project` function in the **proj4** package will be useful to convert its geographic coordination system to longitude/latitude first. You may find the following website helpful: `https://spatialreference.org/ref/epsg/nad83-new-york-long-island-ftus/proj4/`.

[3] Note that "election district" is an administrative border created within each electoral district.

- To match the election district with each SQF location, `st_intersects` function in the **sf** package would be useful.

4. In this question, we will analyze the impact of SQF on the electoral outcomes using the collected NYPD data (`nypd_sqf_2014_2019.csv` on Canvas).

   We apply the "Difference-in-Differences" strategy to estimate the effect of the changes in SQF on the changes in the vote share for the Democratic candidate (We will learn more about the Difference-in-Differences strategy later in the course). In our regression analysis, the outcome variable is the proportion of votes for the Democratic candidate in the 2016 and 2020 presidential elections, Y_{it}, where i denotes the election district and t denotes one of the two time periods. The explanatory variable, D_{it}, is the average number of SQF in 2014–2016 and 2017–2019. We define T_t as the binary indicator that will be equal to one for the second period and zero for the first period. We will also include a vector of X_{it} as our control variables. Formally, we estimate the regression model:

   $$Y_{it} = \beta_0 + \beta_1 D_{it} + \beta_2 T_t + \beta_3 D_{it} T_t + \gamma' X_{it} + \epsilon_{it}$$

   Load the election result data from the 2016 and 2020 presidential election (`pres_res.csv`). The data is measured at election district level, which is much more granular than any other administrative district.[4]

   Following variables are contained in the data:

   - `year`: year of election, either 2016 or 2020
   - `election_type`: type of election
   - `district`: electoral district
   - `elect_dist`: election district
   - `vote_dem`: the proportion of votes for the Democratic candidate received in the election district

   The election district control variables are available in `demographic_df.csv` on Canvas. Those control variables are `total_pop`, `black`, `unemploy_rate`, and `median_income`.

   Estimate the model with an ordinary least square regression. Plot all the estimates along with 95% confidence intervals. In your plot, make sure to adjust for your standard error and highlight the quantity of interest with a different color. Finally, briefly interpret the results and explain the rationale of this modeling strategy and necessary assumptions for causal identification.

5. One potential confounder in our analysis is neighborhood safety. That is, whether the neighborhood is safe or not might affect both votes toward Democrats and SQF. To address this concern, we will collect crime report data in NYC while practicing using API. Follow the steps below to obtain the API token and download this dataset.

   - Sign up for API token at `https://data.cityofnewyork.us/signup` (Using "App Tokens" will be enough).
   - Access crime report data at `https://data.cityofnewyork.us/resource/qgea-i56i.json`.

[4] You can check election districts in NYC at this website: `https://vote.nyc/page/nyc-district-maps`

- Write codes to download the crime report data from years 2017 to 2019 and extract variables `rpt_dt`, `law_cat_cd`, `latitude`, `longitude`, `x_coord_cd`, `y_coord_cd`. You may find the **httr** package useful.
- Limit your download per query to 5000. Note that over-traffic can crash the server and other people might not be able to use the API. `$offset` and `$limit` would be useful to control the number of query.
- Construct the data frame and filter for "felony" and "violation" in the crime category.

Once you finish downloading the data, create a data frame to store the data. First five rows should look like this:

rpt_dt	law_cat_cd	latitude	longitude	x_coord_cd	y_coord_cd
2017-01-01T00:00:00.000	FELONY	40.872037533	-73.83784794	1029096	257023
2017-01-01T00:00:00.000	FELONY	40.830911443	-73.866137497	1021295	242026
2017-01-01T00:00:00.000	FELONY	40.67109684	-73.906209401	1010268	183786
2017-01-01T00:00:00.000	FELONY	40.837436665	-73.944159423	999702	244380
2017-01-01T00:00:00.000	VIOLATION	40.627248134	-73.942627681	1000176	167802

Although you can click and download the entire data from 2006, it is an extremely huge file (larger than 1GB) and would be difficult to load in R even if you directly download the data manually from the website. Therefore, it is necessary to download just relevant observations and columns by communicating with API. The New York City API is constructed based on the Socrata Open Data API, which is widely used in governments, non-profits, and NGOs around the world. **Once you complete this task, you will be able to deal with any datasets that use the Socrata's system at** `http://www.opendatanetwork.com/`**.**

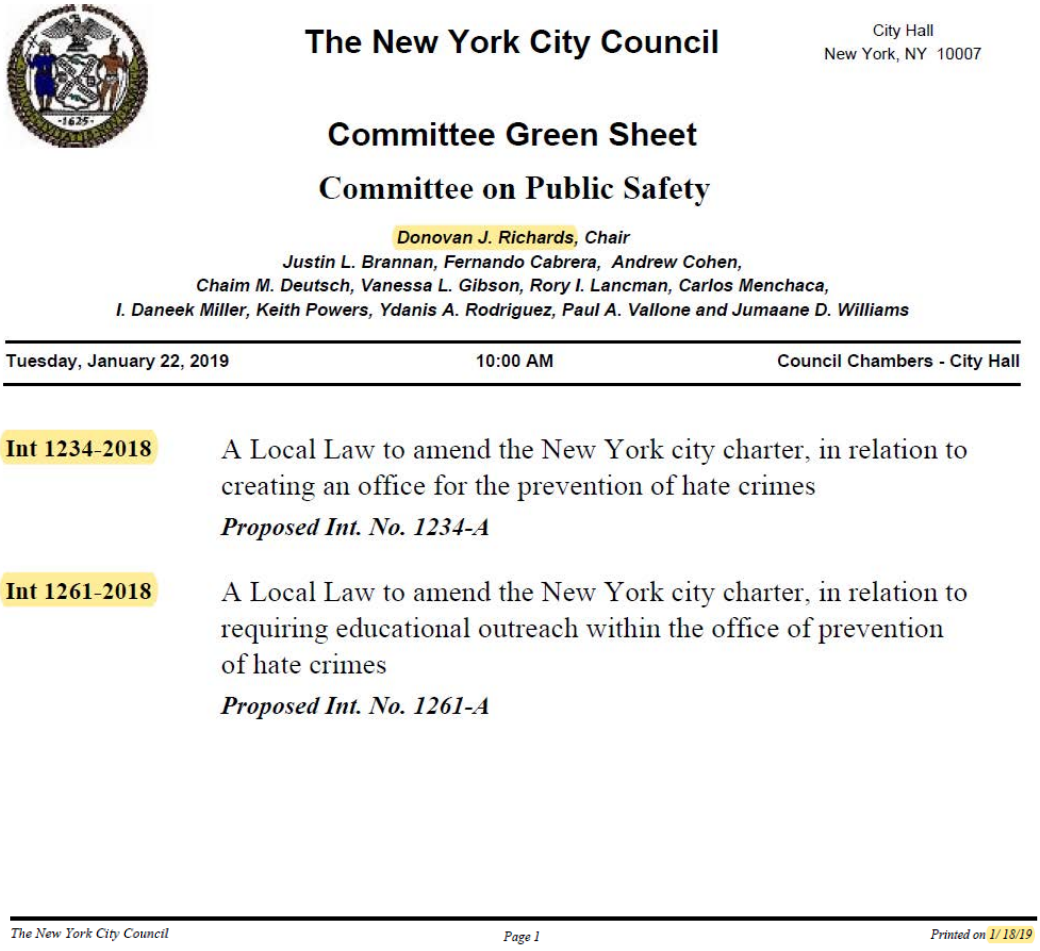
6. **Extra Credit.** Now follow the same steps in Question 3 to map the crime report data to election districts. Calculate the number of crime reports from 2014 to 2016 and from 2017 to 2019 for each election district. Include this additional control variable in our "Difference-in-Differences" analysis. Are our results robust?

7. Another way to strengthen our analysis is to test the mechanism through which SQF affects voting behaviors. One possible mechanism is that Democratic politicians have introduced more or fewer bills on police reforms as a response to increasing concerns over law enforcement. To start on this analysis and practice parsing PDF files, download the meeting agenda for NYC Committee on Public Safety from years 2014 to 2019 (available on Canvas).[5] Read in PDF files and apply `regex` to extract bill, bill number, committee chair, and date of introduction. For this question, you should focus on bills with "Int" as the prefix and can approximate the date of introduction as the printed date (See Figure below). Construct a dataset with the first few rows as:

bill	bill number	chair	date
Int 1234-2018	1234-2018	Donovan J. Richards	1/ 18/19
Int 1261-2018	1261-2018	Donovan J. Richards	1/ 18/19
Int 1105-2018	1105-2018	Donovan J. Richards	2/ 6/19
Int 1309-2018	1309-2018	Donovan J. Richards	2/ 6/19

[5]The original files are available at `https://legistar.council.nyc.gov/DepartmentDetail.aspx?ID=6913&GUID=BCE87221-FD8F-40B5-94D4-66C5F4F643E7`

What do you observe about the number of reform bills across two periods of study (i.e., 2014-2016 and 2017-2019)? Given our goal, can you think of a way to improve this data collection strategy?



The New York City Council

City Hall
New York, NY 10007

Committee Green Sheet

Committee on Public Safety

Donovan J. Richards, Chair
Justin L. Brannan, Fernando Cabrera, Andrew Cohen,
Chaim M. Deutsch, Vanessa L. Gibson, Rory I. Lancman, Carlos Menchaca,
I. Daneek Miller, Keith Powers, Ydanis A. Rodriguez, Paul A. Vallone and Jumaane D. Williams

Tuesday, January 22, 2019	10:00 AM	Council Chambers - City Hall

Int 1234-2018 A Local Law to amend the New York city charter, in relation to creating an office for the prevention of hate crimes

Proposed Int. No. 1234-A

Int 1261-2018 A Local Law to amend the New York city charter, in relation to requiring educational outreach within the office of prevention of hate crimes

Proposed Int. No. 1261-A

The New York City Council *Page 1* *Printed on 1/18/19*

Figure 1: **Sample Meeting Agenda for NYC Committee on Public Safety.** The highlighted parts represent the key information we would like to extract for our analysis.

RES.TLL-008 Social and Ethical Responsibilities of Computing (SERC)

Fall 2022