

# **New York State Voter Removal Analysis**

By Cheul Hee (Jason) Yoo

## 1. Introduction

How accurately does the state of New York, the fourth populous state with 19.8 million residents, keep track of its voter list? In 2016 alone, over 100,000 New York state voters were improperly removed from the rolls.<sup>1</sup> In order to prevent such errors from transpiring again and to re-gain trust from New York state constituents, it is imperative to uncover patterns as to how the system actually removes voters. Therefore, the objective of this study was to identify purged voters and patterns thereof.

We hypothesized that an individual would be purged from the list based on the difference between the current year and the year the person has last voted. That being the case, we spearheaded our project with an assumption that, should the duration a voter has refrained from voting exceeds a certain limit, he/she is automatically removed the roll.

We were able to establish which individuals were more susceptible to the removal process. Thus, this paper paints probable criterions the system relies on to purge a balloter. Our findings can assuage the unpleasantness<sup>2</sup> that the New York state government faces by shining a light on current state of the New York electoral roll and voter removal patterns therein. The state government can also leverage the findings in order to correct the misidentification of purged voters that has been made up to this point and polish the accuracy of the system.

### 1.1 Related Work

No concrete academic work has been undertaken regarding the phenomenon in question. That said, a number of firms and journals investigated the issue. For instance, in 2010, an unnamed voter registration group hired a data analysis firm, TargetSmart, and discovered that Indiana's voter list was inundated with inaccurate information.<sup>3</sup> In fact, the state had the most number of voters who, albeit deceased, were still registered to vote.

## 2. Data Acquisition

In order to grasp the quality of the data, it is important to know how the data was collected by the New York Board of election in the first place. On that note, to participate in federal elections in New York State, a voter needs to register to vote with the New York Board of Elections at least 25 days prior to an upcoming election. A voter must also be a U.S. citizen, 18 years of age, and free of any felony conviction to be eligible to vote. In New York state, one can register to vote online, in person, or via mail.<sup>4</sup> Therefore, the New York state electoral roll, in essence, is an accumulation of voluntary input by individual New York State residents

1. NY AG Office U of Rochester Proposal 2017.pptx
2. "Justice Dept. Seeks to Join Suit Over 117,000 Purged Brooklyn Voters"  
<https://www.nytimes.com/2017/01/12/nyregion/board-of-elections-brooklyn-votes.html>
3. "Analysis by voter group finds dead people likely registered in Indiana"  
<http://www.cbsnews.com/news/analysis-by-voter-group-finds-dead-people-likely-registered-in-indiana/>
4. Voter Registration in New York by New York State DMV  
<http://www.dmv.org/ny-new-york/voter-registration.php>

It is equally worth noting that even though voters can register to vote through any one of the said methods, it is eventually the county board of elections that handle voter registrations or changes thereof.<sup>5</sup> Therefore, a complete list of New York State electoral roll is a composite of voter registration data sent from every county within the State of New York.

Such a dataset was provided as flat file by Lacey R. Keller, the Director of Research and Analytics at the office of the Attorney General of the State of the New York. She has given us a link to two different sets of compressed data, both of which had .txt extension. The only difference between the two was that one contained data pertaining to the latest 2016 presidential election whereas the other one did not. The link also contained a data dictionary that encapsulated the following about each individual attribute - Field Position, Field Size, Field Name, Description, and Notes which often contained information as to what values nominal attributes can take.

### 3. Data Preparation

We primarily worked on the file that included 2016 presidential election data, which had 16,636,983 distinct New York state voters with 45 attributes mostly nominal in nature. Out of this number, 5,754,556 voters were labelled as either 'inactive' or 'purged.' Given the voluntary nature of data collection method used for this particular electoral roll, the dataset was swamped with sporadic null values and inconsistencies. Before grouping informative features that represented gender, county, and etc. to attain a high level picture of how they correlate with the removal process, we leveraged the following steps in order to clean and process the data.

1. The majority of optional features was completely null or almost completely null (>90% missing values) as voters could opt to not fill them out (Appendix A). Since they also happened to be heuristically least correlated with the attribute of interest – VoterHistory, coupled with the impracticality of accounting for completely null variables, we found it rational to drop them altogether. For instance, MIDDLENAME variable was thoroughly null. Additionally, we deemed it highly probable that accounting for the said attribute would yield a result that is of little interest for our specific project. Eventually, this effectively cut the number of variables to 16.
2. We needed to create several binary attributes that indicated whether a person had participated in an election held in a specific year. The rational was that VoterHistory variable enumerated voting history of each constituent using a string of election names with semicolons as delimiters. Since the range of VoterHistory was [2002, 2016], we made 15 binary variables for each year. Similarly, we were also able to derive the number of years that had been passed since a voter last voted.
3. We also obtained 'current age' of constituents using their date of birth. Ultimately, we were able to extract information as to their age and the number of years since they have last voted about 'active' and 'inactive' voters.

5. Voter Registration Information For Students by New Paltz State University of New York  
<https://www.newpaltz.edu/saus/voter-registration-information-for-students/>

### 3.1 Purged

A voter can be dropped from the electoral roll due to any of 10 possible reasons, out of which only four proved to be useful in our analysis – Moved, Death, and Other. In other words, voters are likely to be removed from electoral rolls because one of the followings has occurred to them: they have moved out of the county, were pronounced dead, or other circumstances not mentioned above.

## 4. Exploratory Analysis

As much as we were interested in investigating ‘purged’ voters, we also wanted to mine important data about ‘active’ voters since the differences between two voter sets could, in turn, reveal other thought-provoking patterns.

In that context, we dichotomized voters whose statuses fell under either ‘active’ or ‘pre-registered’ into two genders as show below (Figure 1). They were both bimodal; each had a peak at around 30 and another one around 55. This trend is further clarified if we plot the same information using line graphs (Figure 2).

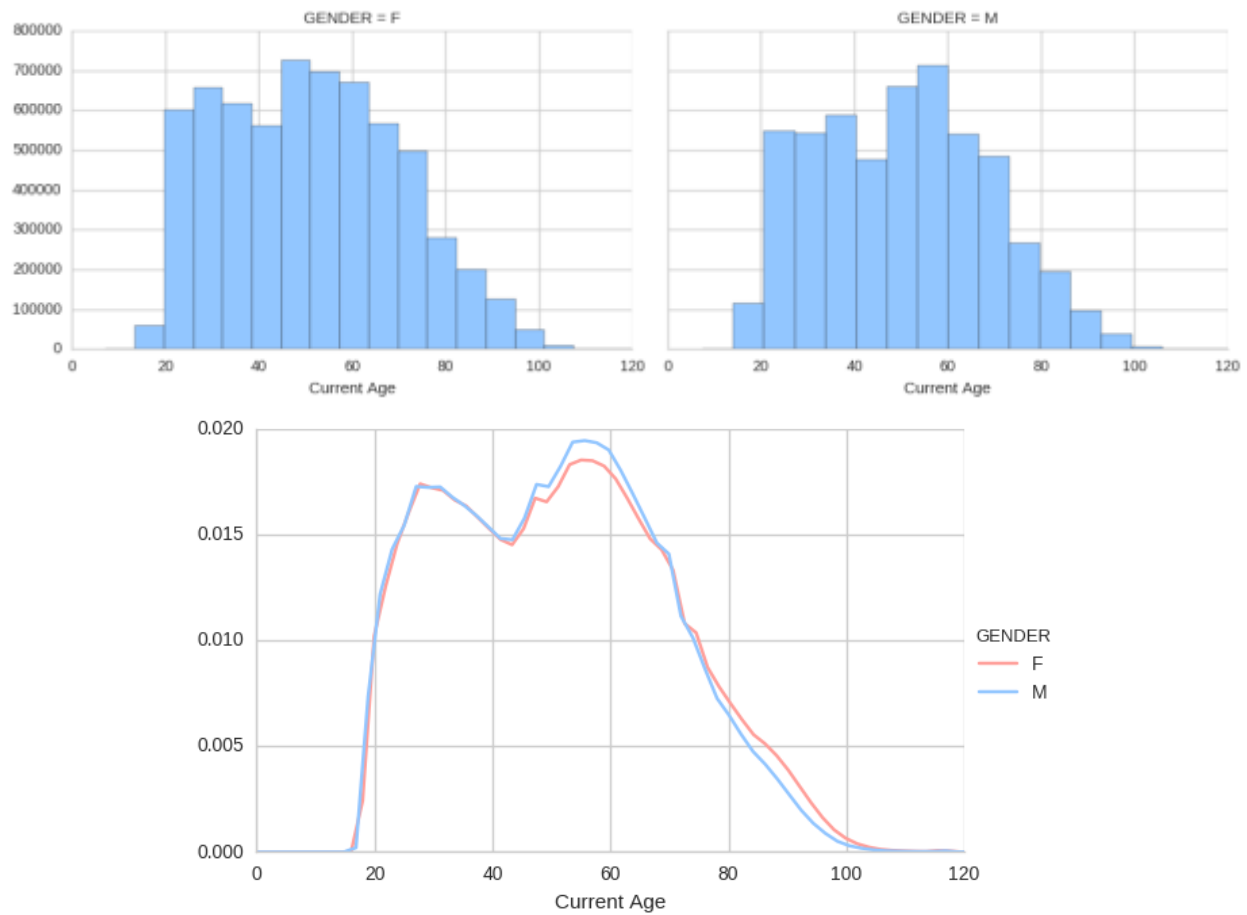


Figure 1 & 2. Above) Count of Current age partitioned into two genders. Below) Density graph of figure 1.

## 4.1 Purged Voters

We first plotted a heat map that shows how many voters were purged in each year (ranging from 1990 – 2016) and reasons they were dropped for the specific year. Put differently, the darker the color, the more people the system purged. In that regard, it is apparent that three reason codes – Moved, Death, and Other – dominate the map. Additionally, we can note the number of dropped voters increases significantly as the year progresses, which could potentially be attributed to the fact that an increasing number of residents are more willing to report their statuses to their respective county election boards. Automation of data collection could also have played a role.



Figure 3. Number of Purged Voters (Reason and Year)

Next, we plotted multivariate density graphs using age which voters were purged against the densities of each reason code, which then were further dichotomized into two genders as shown below. A notable pattern is that for both genders, Moved and Other are similar in terms of their distributions. Furthermore, Death is the only one that displays right-skewed distribution. Such distributions can be ascribed to the fact that seniors are more likely to stay at one place and that their chance of death is naturally higher than that of younger population.

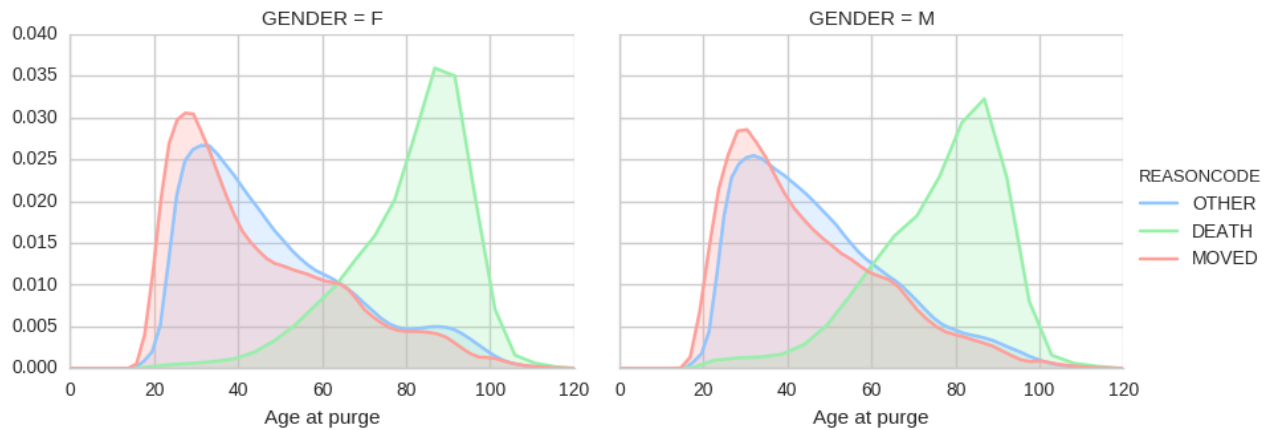


Figure 4. Multivariate Density Distribution of Age, and Reason code Separated by Gender

Additionally, we plotted a bar graph using year and the number of purged voters. Each year is then further divided into three genders. We can observe that across all years, the number of female purged voters dominate that of male voters.

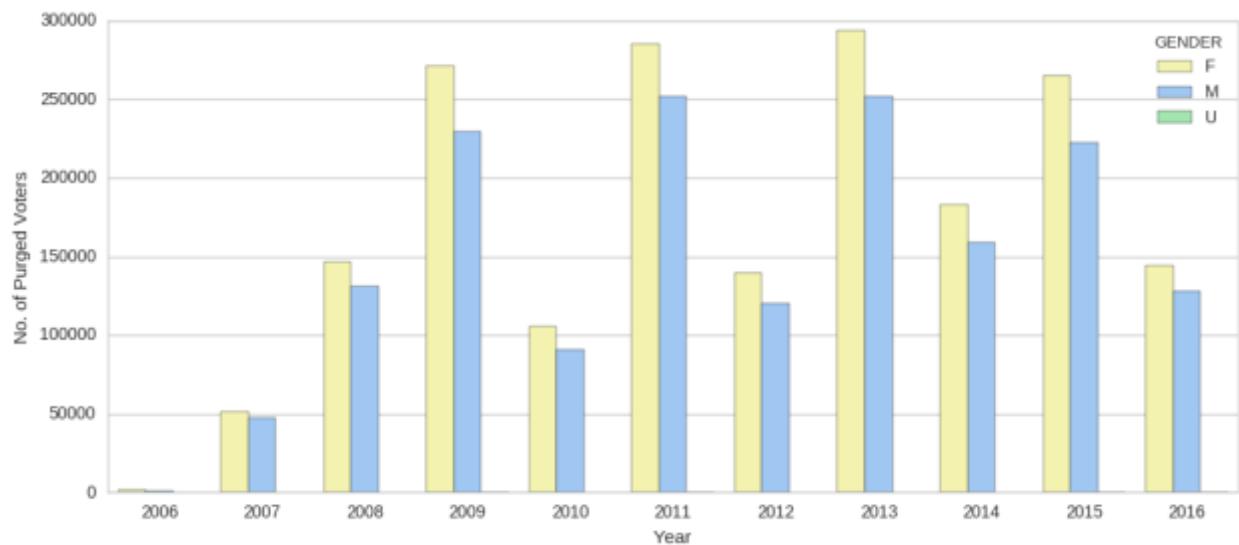


Figure 5. Bar Graph (Year vs No. of Purged Voters) with Each Year Partitioned Into Three Genders.

Lastly, we plotted a heat map using ‘years since the last voted’ and age at which voters were purged. Interesting point about this map is that it shows two major groups that most dropped from the voter roll. The first of two is those who are below the age of 35 and the years passed since they last voted were below 8. The second group consisted of people who were above 55 in terms of age and their years passed since last voted were below 4.

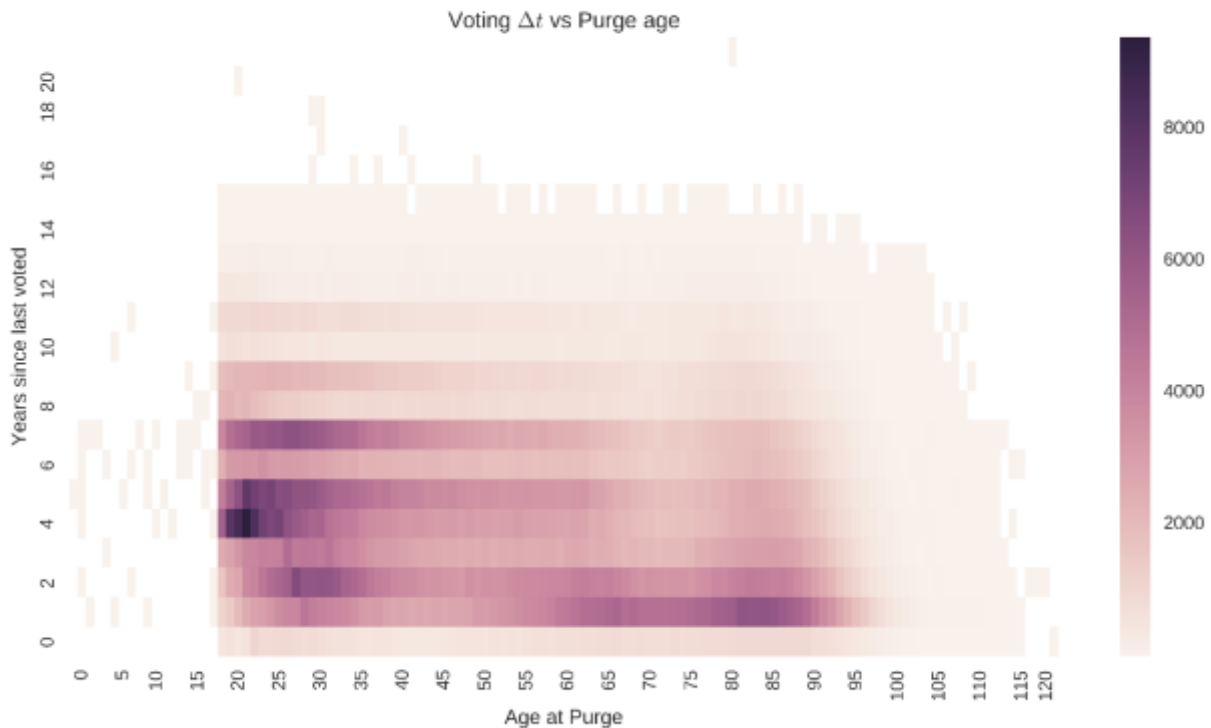


Figure 6. Years Since Last Voted vs Age At Purged

## 5. Conclusion

Our analysis of the New York State voter data gives a complete picture of active and ‘purged’ New York constituents. What with numerous errors and inconsistencies contained in the dataset, we were still able to generate insightful information about New York state voters, which were further crystallized with visualizations shown above. All in all, we narrowed our focus on purged voters and patterns therein, producing a result that show there are two major groups that were purged from the state’s electoral voter roll as shown in figure 6.

Appendix

Missing Values

LASTNAME	
FIRSTNAME	
MIDDLENAME	
NAMESUFFIX	
RADDNUMBER	
RHALFCODE	
RAPARTMENT	
RPREDIRECTION	
RSTREETNAME	
RPOSTDIRECTION	
RCITY	
RZIP5	
RZIP4	
MAILADD1	
MAILADD2	
MAILADD3	
MAILADD4	
DOB	
GENDER	
ENROLLMENT	
OTHERPARTY	
COUNTYCODE NUMBER	
ED NUMBER	
LD NUMBER	
TOWNCITY	
WARD	
CD NUMBER	
SD NUMBER	
AD NUMBER	
LASTVOTEDDATE	
PREVYEARVOTED	
PREVCOUNTY	
PREVADDRESS	
PREVNAME	
COUNTYVRNUMBER	
REGDATE	
VRSOURCE	
IDREQUIRED	
IDMET	
STATUS	
REASONCODE	
INACT_DATE	
PURGE_DATE	
SBOEID	
VoterHistory	