Investigation on Potential Variables Affects Polling in Toronto

Yiying Chen

2 February 2022

Abstract

As a significant measure of collecting advice from Toronto residents, polling was always an important reference for city’ s development, and connects people in Toronto with the government. The dataset provides historical data for conducted polling in Toronto with exhaustive variables. This report utilizes the dataset as foundations, aiming to find potential variables that might affect the efficiency of polling in Toronto. This analysis could help us to find potential variables that affects the efficiency, give targeted advice to the polling procedure.

1. Introduction

Toronto City Government committed to solve specific life problems for residents in Toronto. So, they conduct polls regularly for different area and communities in Toronto throughout these years. This data records the performances of each poll from 2015 until now, and that is also the reason why this data was very valuable for data analysis. By investigating on those historical data, we could see what the variables could affect the result of the poll and get a deeper understanding of polling.

The design of polling requires efforts from various department city governments and the community. There are specific criteria of assessing the reliability of poll, the response rate, it may varies based on different sample size. If the poll does not meet the response rate, then the polling cannot be used as a reference for the urban development. Hence, improve the efficiency of polls was significant.

This report will analyze variables from the dataset, investigating what are the potential factors that might affect the efficiency of polling. After the data cleaning procedure, we select data that directly link to the poll result to analyze. By utilizing bar plot, scatterplot and summary table, we could visualize how different factors affect the response rate. Based on the result, we are able to come out advice that were particularly suitable for polling in Toronto, including design of the polling, distribution of polling. Correspondingly, by those improvements, we expect a higher respond rate, in other words, a more efficient polling system.

2. Data

2.1 Data source

Initially, the report utilizes data from City of Toronto open data, mainly display the result of polling that focuses on the opinions of residents and businesses on various topics. The rules of polling were based on the 190th chapter on City by-law, about polling and notification. Residents within the polling area are eligible for the polling, giving their opinions as whether in favor or opposed. They could also do not show their perspectives by leaving it blank, which would be not calculated in the result. We access this data by using the R package **opendatatoronto**. The data starts from April 1, 2015, last updates on February 2, 2022.

2.2 Methodology and Data Collection

The dataset we used in this report contains responses from polling from 2015 to now. It records the specific details of the polling: the results of the poll, the specific date that polls open and pass, the data of each type of response that includes neutral/blank results… Undoubtedly, this data was meaningful, it contains primary data, and updated regularly just after the closing of each poll. The questions of the poll all closely relate to residents' daily life, trying to solve or optimize existing situation, including Boulevard Café, Traffic Calming, Front Yard Parking, Permit parking. This long-lasting record of data will give us a solid background for statistic study.

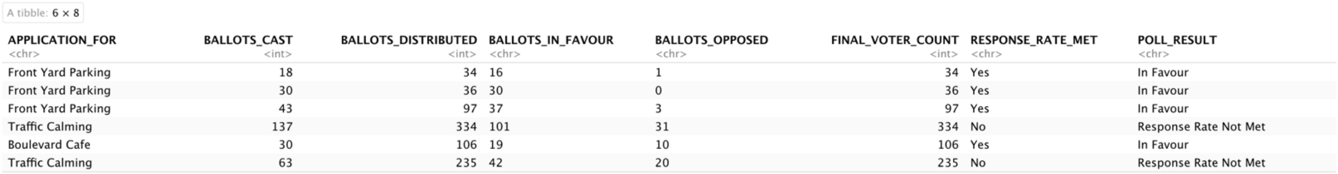
Under the pandemic situation caused by COVID-19, many people worked from home. Correspondingly, they will more focus on problems that exist in their daily life such as front yard parking and respond to the polling. They will pay attention to issues of community and infrastructure that they might not realize before. The dataset has its own pros and cons. It is reliable that the government record the data from each polling and update them regularly regarding to various topics throughout these years. However, polling usually take place in specific area and address, which we not sure whether it included all streets and avenues in Toronto. And as we inspect the data, there are also many N/A (not available) signs under the record of address, which is harder for us to determine whether the address of polling was exhaustive or not.

The population of this study was all Toronto residence, aiming to indicate their ideas toward various topics in the community. The dataset contains the perspectives of residents who respond to the polling. By the information provided in Toronto city government, the polling usually delivered by the e-polling website and mailed to residents live in the polling area. Including owner, resident and tenant all have a chance to fill out their ballot. Normally, if respond rate (which is also known as completion rate or return rate) of polling exceeds 50%, then the result would be an excellent reference for scientific study. If the sample size was not enough for the study, then the statistical result for the dataset will be easily underestimated.

2.3 Data characteristic

The dataset Polls Conducted by the City includes polling results and detail data from 2015-2022. There are 1054 observations in the dataset and 25 variables as the polling output. After the data cleaning procedure, we only select variables that are useful and meaningful for our data analysis for polling. The application for includes the topic of polling, ballots cast and ballots distributed refer to the ballots returned and distributed during the polling process. The next two attributes, ballots in favor and ballots opposed show residents' attitude toward the question discussed. And for final vote count refers the total number of voters per polling. There are corresponding response rate depends on the number of ballots distributed, and the variable indicates whether it was a useful result for analysis, recorded as "yes" or "no".

The final column shows the result of polling: whether voters satisfied with the polling result or not, and record the situation that the response rate not met, which the result was not useful for study.



A tibble: 6 × 8

APPLICATION_FOR <chr>	BALLOTS_CAST <int>	BALLOTS_DISTRIBUTED <int>	BALLOTS_IN_FAVOUR <chr>	BALLOTS_OPPOSED <chr>	FINAL_VOTER_COUNT <int>	RESPONSE_RATE_MET <chr>	POLL_RESULT <chr>
Front Yard Parking	18	34	16	1	34	Yes	In Favour
Front Yard Parking	30	36	30	0	36	Yes	In Favour
Front Yard Parking	43	97	37	3	97	Yes	In Favour
Traffic Calming	137	334	101	31	334	No	Response Rate Not Met
Boulevard Cafe	30	106	19	10	106	Yes	In Favour
Traffic Calming	63	235	42	20	235	No	Response Rate Not Met

2.3.1 Conducted poll results and correspond topic

By utilizing a titling graph, we could put the agreed ballots, the polling topic and poll result together to make a compare the polling result between various polling topics. As figure 1 shown below, we could analyze the result from various aspects:

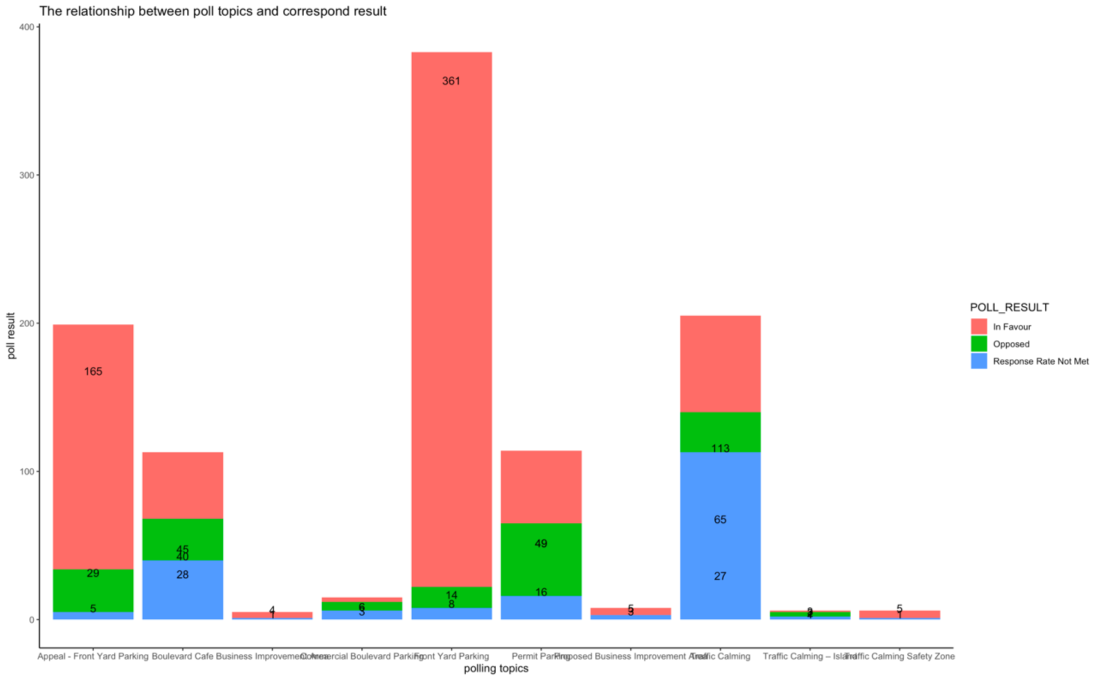


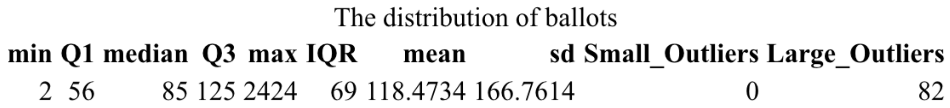
Figure 1

As we can observed from the graph, initially we could see that people have obvious tendency for polling topics. For front yard parking, traffic calming, appeal-front yard parking, people have an obvious tendency of giving their comment on the polling. We could analyze that these topics are mostly relate to voters' daily life, and they are willing to participate in the voting. The more feedback a topic received, can show the more significant this topic was.

We could also analyze this graph through the poll result. The pink bar means voters satisfied with the current situation, the green bar shows voters opposed to the current arrangement, and the blue bar shows that the response rate not met. The traffic calming section has the most response rate not met, which means many voters are reluctant to show their opinions toward the traffic calming in their section. For the front yard parking, it received the most agreements, indicate that voters are very satisfied with the arrangement for front yard parking section. Meanwhile, permit parking received most opposed ballot. Voters are not satisfied with the permit parking regulation and willing to make some changes on that. Undoubtedly, the government should pay more attention on the topics that show obvious tendency on poll result, that would reflect the true thing that residence truly needed.

2.3.2 The distribution of ballots

As a significant variable, the distribution of ballots needs a detailed analysis. It could shows the problems that would exist in the distribution of the ballots, which involves the design section in polling. We did a summary table for this variable particularly, to analyze the performance of data between different quantiles and any outliers.



The distribution of ballots

min	Q1	median	Q3	max	IQR	mean	sd	Small_Outliers	Large_Outliers
2	56	85	125	2424	69	118.4734	166.7614	0	82

Figure 2

From the distribution, the mean number of ballots distributed for each polling was approximately 118.47 (in 2 decimals). The dataset shows that, for specific topic, the polling would be held within specific area not a wide range. So, there is no need to distribute the ballot to unrelated people. However, there was a large stand deviation for this dataset. It represents that the number of ballots distribution varies a lot for different polling. And this variable also closely relates to the response rate. If the response rate not met, then it cannot be used as a reference.

The minimum number of ballots distributed was 2, which was rare in polling. A large sample size was the foundation of data analysis, the more people involved, the more precise result we would get. Although the aim of conducting polling was to solve problems in specific area or community, we still need to collect result from residents as much as possible. The opinions from 2 people cannot represent the opinions of residents in that area. Anyway, the sample size in the polling should large enough.

Although outliers exist in this dataset, the dataset was still a reliable resource for data analysis. We could see that the value for Q1 and Q3, which was the 25th and 75th percentage in the distribution of data are keep in the same level. And those values are also reasonable for the numbers of residents live in a specific area. When the government conduct a polling, they should distribute ballots as much as possible to make sure that the response rate could meet. We should try our best to avoid the low response rate, hence, to make every polling valuable.

2.3.3 The relationship between lost ballots and response rate.

In data analysis for this dataset, we could refer to the result of polling only if it passes the response rate. From figure 1, we could also see that the low response rate exists a lot in various topics of polling. To improve the efficiency of polling, we should investigate what are the possible factors that might the response rate.

We did some mutations in this step, based on the raw data we got. Since we have the number of ballots that was distributed and returned. We subtract the ballots distributed with the ballots cast to get a new variable, "lost ballots" . Basically, we investigate whether the number of ballots that were not returned could affect the final response rate. We made a scatterplot for those two variables.

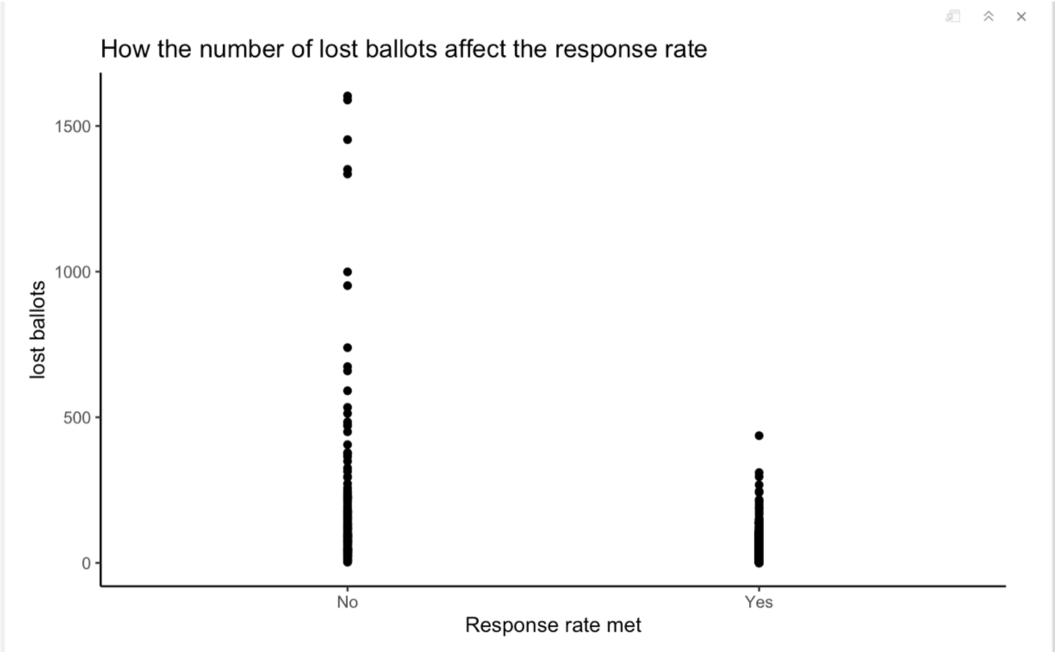


Figure 3

This step confirms that the number of ballots that were not returned could largely affect the response rate, which directly relate to the efficiency of our polling. It is normal that a small number of polls might lost during the distribution process due to various reasons, however, the large number of lost polls should be investigated. It means that the distribution procedure probably exist some problems that many ballots are not received by residents, and we need improvements on that to promote the efficiency.

Apparently, for the low response rate, which is the polls that could not be used for analysis, tend to have large numbers on lost ballots. There are two possible explanations: the distribution step exist systematic errors that many voters not receive their ballots; or voters not interested in specific topic, so they choose not to fill the ballots. On the other side, for polls that met the response rate, the number of lost ballots was all lower than 500 without any outliers. So we should aim to lower the number of lost ballots from each poll to about 500, then we would get more useful polling results accordingly.

3. Reference

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software,
4(43), 1686, https://doi.org/10.21105/joss.01686

Yihui Xie (2021). knitr: A General-Purpose Package for Dynamic Report Generation
in R. R package version 1.36.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and
Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In
Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing
Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

Sharla Gelfand (2020). opendatatoronto: Access the City of Toronto Open Data
Portal. R package version 0.1.4.
https://CRAN.R-project.org/package=opendatatoronto

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2021). dplyr: A
Grammar of Data Manipulation. R package version 1.0.7.
https://CRAN.R-project.org/package=dplyr