Polls Conducted for City Issues in Toronto from 2015 to Present*

A Simple Linear Regression Analysis on How Poll Pass Rate Related to Specific Poll Features

Yizhen Wang

27 April 2022

Abstract

Polls conducted by City Office establish residents' opinions on various city topics and issues. By obtaining the polls dataset provided on Toronto Open Data Portal, this report will explore and analyze how specific poll features impact the pass rate of the results. By modeling with simple linear regression model, the report has concluded that the main predictors were quantities of ballots in favour and opposed, numbers of ballots need proceed, numbers of ballots received and missing, and amount of final voters.

Keywords: Poll Engagement, Toronto Polls Conducted, City Clerk's Office, Simple Linear Regression, Open Data Toronto

Contents

*Code and data are available at: https://github.com/KSaxophone/Polls-conducted-Toronto.git

Introduction

With increasing size of cities, more and more urban issues such as lack of parking spots and traffic noises have appeared and influenced citizens' daily lives. Toronto City Office has began to conduct polls for residents to express their opinions on restricting these issues since 2015. Every owner, resident and tenant in the polling area is mailed a notice giving information about the poll and the deadline to submit a ballot, the ballot itself and a postage-paid return envelope. Poll results are available 10 business days after the final day of polling (*Polls Regarding Changes in a Neighbourhood* 2022).

For the proposal of the polls, if a poll meets the percentage benchmarks, the poll will pass, subject to Council approval and the proposal will proceed. However, if the poll not pass, some specific application types will be restricted. For example, the Front Yard Parking Polls cannot be conducted again for three years (*Polls Regarding Changes in a Neighbourhood* 2022). In this case, the interest of the paper is to find how the polls pass rate responses with recorded data.

This paper will focus on exploring the influential predictors on the pass rate of the polls conducted by Toronto through statistical model, Simple Linear Regression. The approach of the final results can be divided into questions as following:

- After the EDA processing, which variables have higher correlations with the response factor (polls pass rate)?
- After the statistical modeling, which variables significantly impact the response factor and how to interpret them?

For the first question, by plotting different graphs on both numerical and categorical variables, it is clear that variables BALLOTS_IN_FAVOUR, BALLOTS_NEEDED_TO_PROCEED, BALLOTS_OPPOSED, BALLOTS_RECEIVED_BY_VOTERS, BALLOTS_SPOILED, FINAL_VOTER_COUNT and MISSING have higher correlations by visualization.
The second question obtains the result of quantities of ballots in favour and opposed, numbers of ballots need proceed, numbers of ballots received and missing, and amount of final voters (BALLOTS_IN_FAVOUR, BALLOTS_NEEDED_TO_PROCEED, BALLOTS_OPPOSED, BALLOTS_RECEIVED_BY_VOTERS, FINAL_VOTER_COUNT and MISSING) are key determiners on polls pass rate.
The findings of the paper would be helpful to encourage residents of Toronto to vote more for the proposed city issues.

For the structure of the paper, firstly, the dataset will be obtained and cleaned, with relative variables selected and visualized. Basic statistics and explanatory description analysis will be displayed to help determine significant predictors. All of this process will appear in `Data` section. Then the data will be separated as training and testing samples to simulate the and conclude predictors with large impacts. The data in the main report would use training data, and the testing simulation will be added in scripts. In `Method` and `Model` section, statistic methodology such as simple linear regression or F-test will be applied to help construct and select the effective model. The final model will be displayed in `Result` section, along with its numerical explanation and diagnose plots. At last, the model will be interpreted in `Discussion` section, as well as its limitations and future possible improvements. Extra information such as datasheet enhancement will be provided in appendix and additional details if necessary.

Data

Data Source and Processing

Toronto's City Office has conducted over 1000 polls on behalf of City divisions, these polls are meant to establish the opinions of residents and businesses on various topics covered by a City by-law (*Polls Conducted by the City* 2022). The report is based on the data of poll results that published on the City of Toronto Open Data Portal, and the csv format of the data was downloaded into my R Studio from the website (Gelfand 2020). The results contain 1071 observations and 25 variables related to polls such as application topics, ballots distribution and casts, pass rate and so on. The collection of data starts from April, 2015, and the dataset was last refreshed on April 26th, 2022.

In order to constructing model on poll pass rate results, a data cleaning process was applied to firstly remove N/A values existing in the pass rate feature of the dataset. What's more, a new variable was created to record the quantity of missing ballots (difference between ballots distributed and returned). Finally, 14 variables of polls data were considered relevant to polls pass rate and were selected from the dataset, which are: APPLICATION_FOR, BALLOTS_BLANK, BALLOTS_IN_FAVOUR, BALLOTS_NEEDED_TO_PROCEED, BALLOTS_OPPOSED, BALLOTS_RECEIVED_BY_VOTERS, BALLOTS_RETURNED_TO_SENDER, BALLOTS_SPOILED, FINAL_VOTER_COUNT, PASS_RATE, POLL_RESULT, POTENTIAL_VOTERS, RESPONSE_RATE_MET and the newly mutated one, MISSING.

Data Characterisctics and Visualization (EDA)

Characterisctics

The dataset mainly focuses on various polls' application topics and their important features as well as results. Some applications includes traffic calming, front yard parking and Boulevard Cafe. The population of the dataset is the residents in Toronto City, and the data of residents' engagement for various polls is collected through mailing. Every resident in the polling area is mailed a notice of the poll's information and deadline, a ballot, and a postage-paid return envelope (*Polls Regarding Changes in a Neighbourhood* 2022). The problem of non-response may occur since some residents may miss the deadline to submit. If too many voters did not submit their ballots and non-response rate increased, the polls might not be able to meet the passing benchmarks and failed to proceed.

Specifically, the residents that were allowed to participate in voting include any owner, resident or tenant of a property in the polling area. All participants must be 18 years old on or before the final day of the polling (*Polls Regarding Changes in a Neighbourhood* 2022).

Since the data collection has began from 2015 and continues to get updated, the dataset has covered a relatively long time span of polls engagement, which allows the data to be more inclusive and comprehensive. Also, the dataset only collects poll results in Toronto City, so the geographical range of people's voting would be more specific. However, when polls are conducted, damaged or missing ballots will always exist and bring bias to the poll results. The consequences of ballots deficit may cause the final counts to be underestimated and influence the polls pass rate. The dataset has implicitly revealed this factor's data, which was reflected in the difference between number of ballots distributed and returned. The data processing section above has mutate the value into a new variable called MISSING, further analysis can be constructed with consideration of its negative impacts.

Note that there were no similar datasets that could have been used, since the results of this dataset were collected by Toronto City Office, and other non-official institution would not have authorities for residents voting collection. What's more, the dataset is updated frequently (almost daily), so the dataset is unique and specific to the city.

There were 14 variables which could have potential influence on poll pass rates and they were selected by common sense and background knowledge for analyzing the core topic of the paper. Table 1 below has listed the real world meanings of these 14 variables. A dataframe called "df" was created for table 1, which shows variable names and their descriptions (Zhu 2021).

Table 1: Predictor Names and Descriptions

variables	descriptions
APPLICATION_FOR	Type of application
BALLOTS_BLANK	Number of ballots received with no mark to identify
BALLOTS_IN_FAVOUR	Number of ballots received and marked in favour
BALLOTS_NEEDED_TO_PROCEED	Number of ballots needed to proceed based on percentage of return
BALLOTS_OPPOSED	Number of ballots received and marked opposed
BALLOTS_RECEIVED_BY_VOTERS	Number of ballots returned to City Clerk's Office
BALLOTS_RETURNED_TO_SENDER	Number of ballot returned by Canada Post as not delivered
BALLOTS_SPOILED	Number of ballots received and were not clearly marked
FINAL_VOTER_COUNT	Number of total voters on the final poll list
PASS_RATE	Number of returned ballots needed for a positive poll result
POLL_RESULT	Final result of poll
POTENTIAL_VOTERS	Number of people residing within poll boundary range
RESPONSE_RATE_MET	If number of ballots returned has met the required response rate
MISSING	Number of missing ballots

Among these 14 predictors, three of them are categorical variables, which are type of applications (APPLICATION_FOR), the final result of poll (POLL_RESULT) and whether the results meet response rate (RESPONSE_RATE_MET). The rest 12 predictors are all numerical, with polls pass rate as the response variable of the report's analysis.

Visualizations

For these 14 variables, related visualizations were created to identify the variable characteristics (Wickham 2016). Scatterplots of numerical variables were created to explore whether they have certain correlation with response variable PASS_RATE. And box plots as well as bar plots were created to discover the features of categorical variables.

With results of scatter plots, we can generally identify the variables with relatively strong correlations to the response. From figure 1, the counts of blank ballots do not have a relatively clear pattern with the pass rate, and figure 9 of potential voters also indicate no strong correlation. Other scatter plots from figure 2 to figure 8 as well as figure 10 all show a relatively strong relationship with polls pass rate. And they will be selected into the starter model.

For categorical variables, figure 11 illustrate the application type's relationship with pass rates, and some of the elements have too large differences, so it will not be considered in the model. What's more, figure 12 has provided the bar plot of poll results towards each application type. The plot has shown that most of the applications were dominated by in favor options, especially for Front Yard Parking policy, most residents in the city would want the proposal to get approved. This figure could give an insight for the city government to improve its policy making.

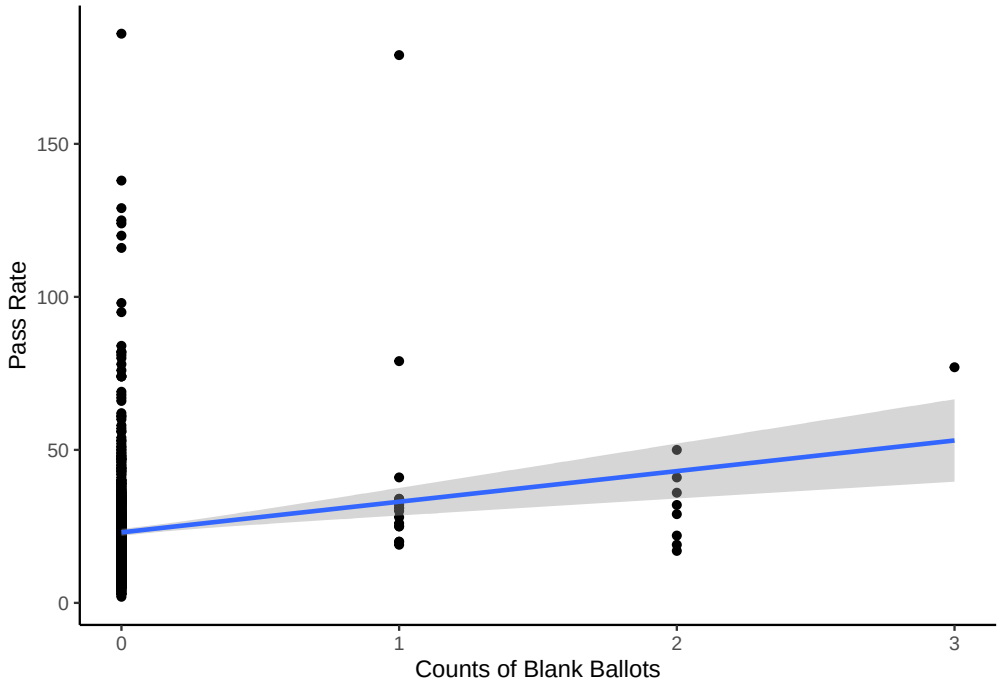

Figure 1: Scatterplots of Blank Ballots vs Pass Rate

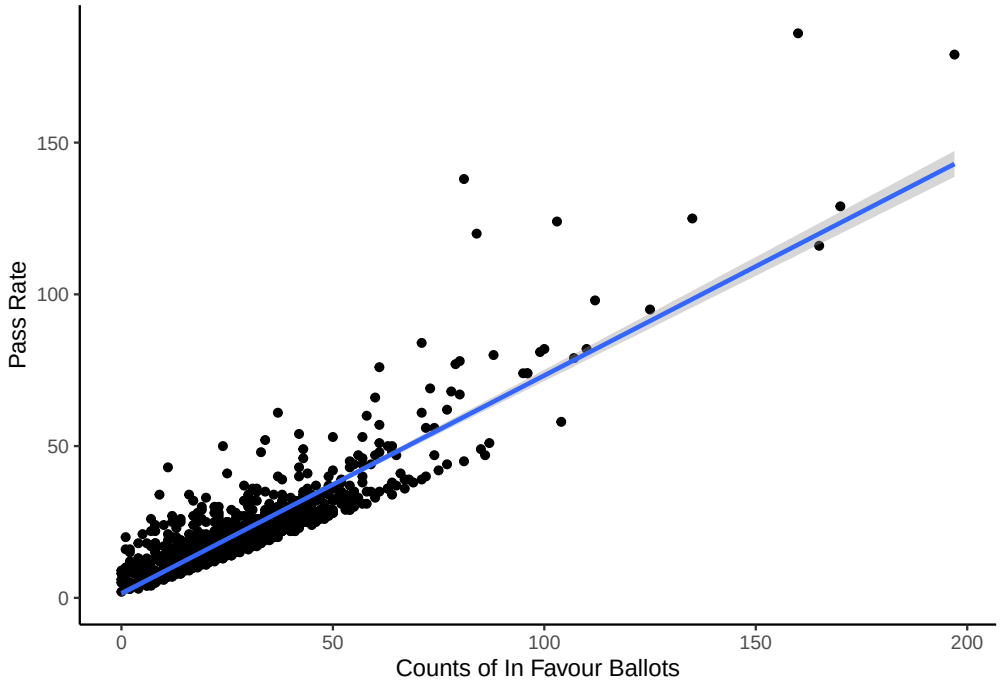

Figure 2: Scatterplots of In Favour Ballots vs Pass Rate

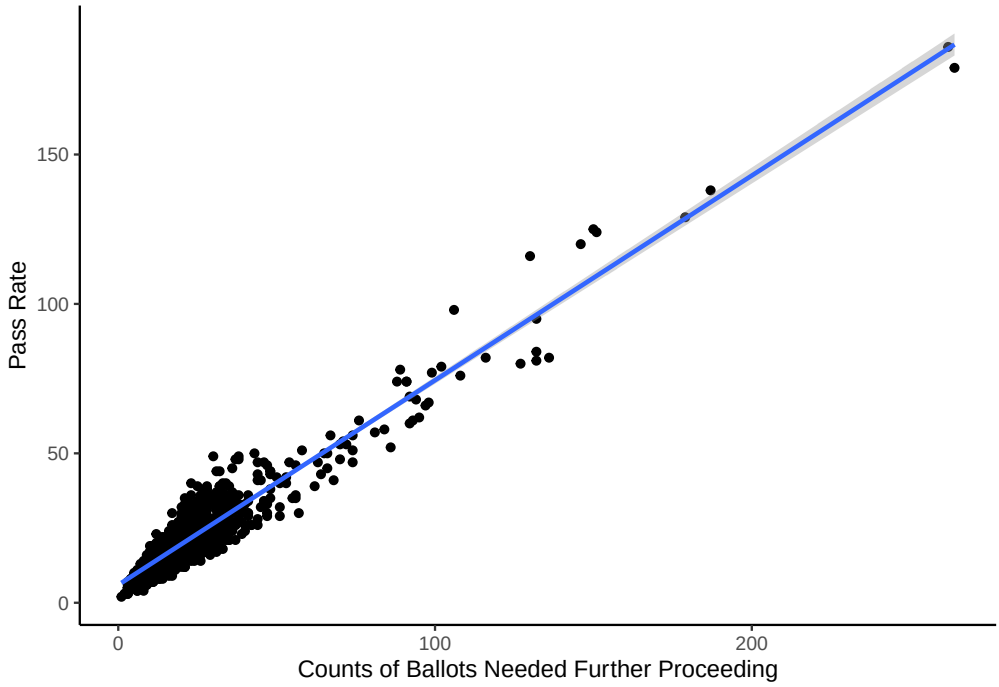

Figure 3: Scatterplots of Ballots Needed Further Proceeding vs Pass Rate

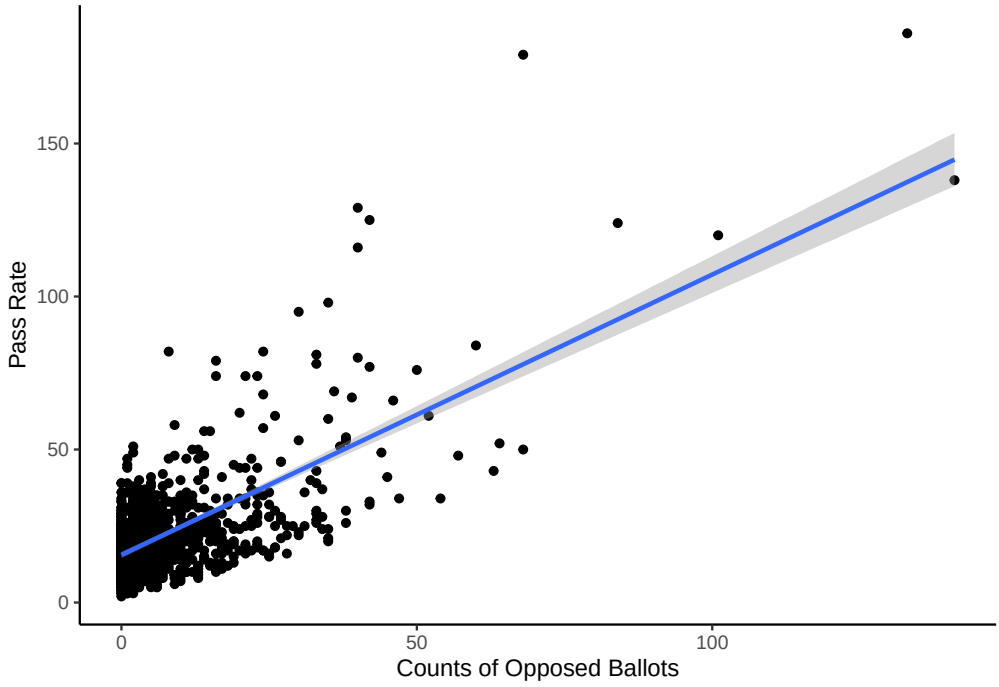

Figure 4: Scatterplots of Opposed Ballots vs Pass Rate

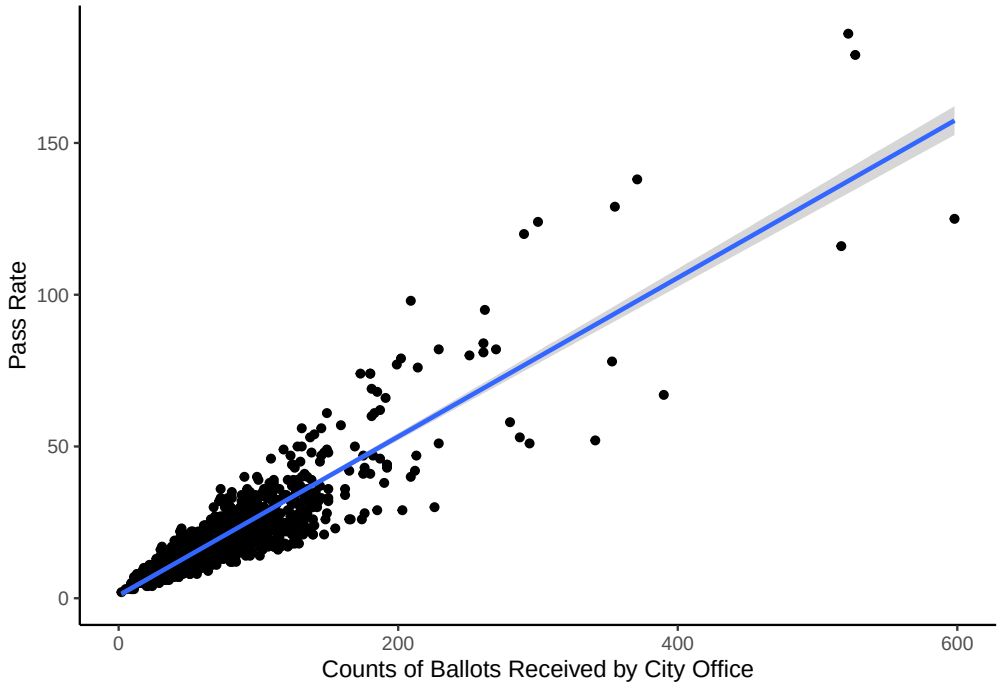

Figure 5: Scatterplots of Ballots Received by City Office vs Pass Rate

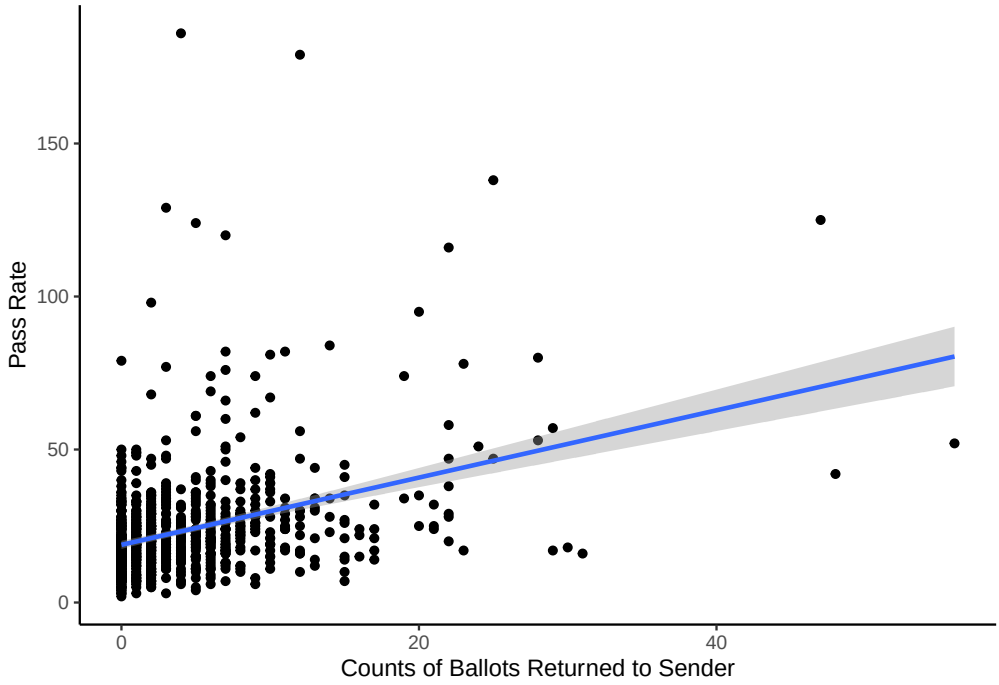

Figure 6: Scatterplots of Ballots Returned to Sender vs Pass Rate

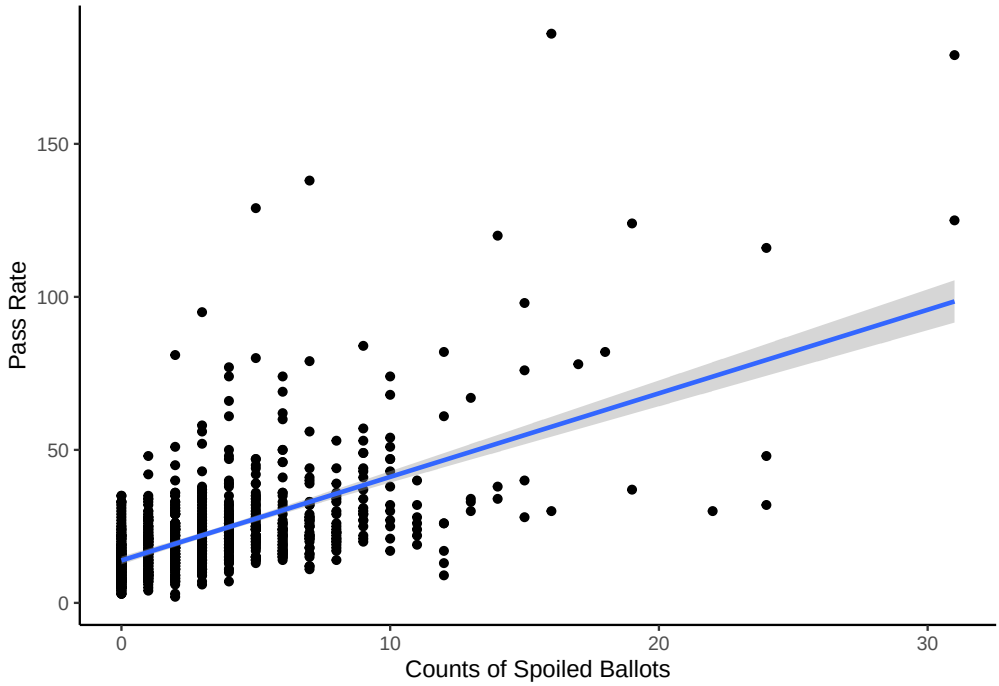

Figure 7: Scatterplots of Spoiled Ballots vs Pass Rate

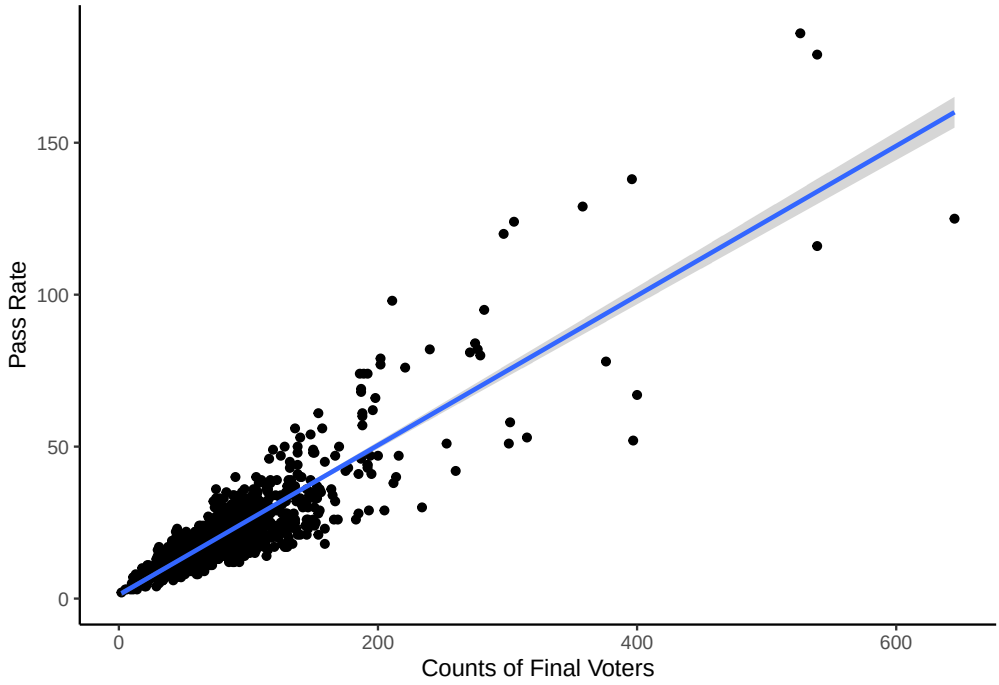

Figure 8: Scatterplots of Final Voters vs Pass Rate

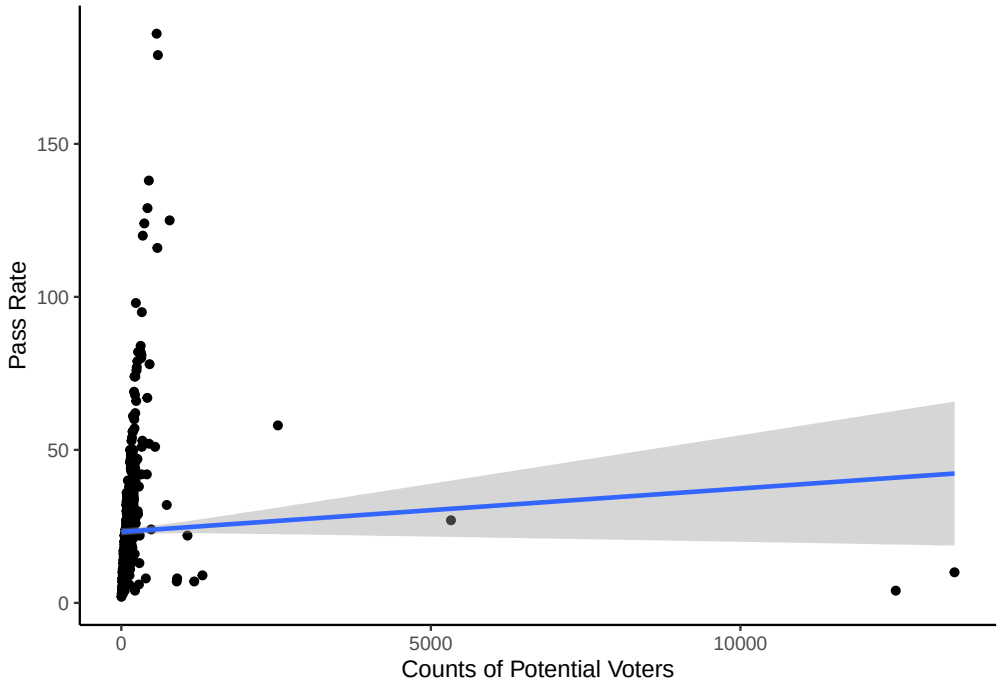

Figure 9: Scatterplots of Potential Voters vs Pass Rate

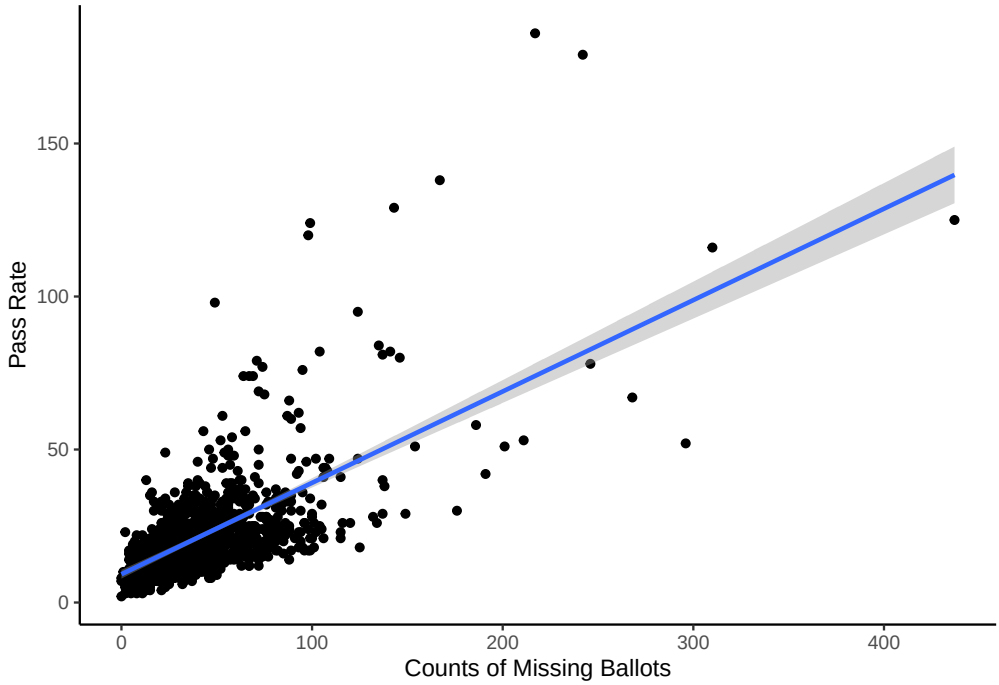

Figure 10: Scatterplots of Missing Ballots vs Pass Rate

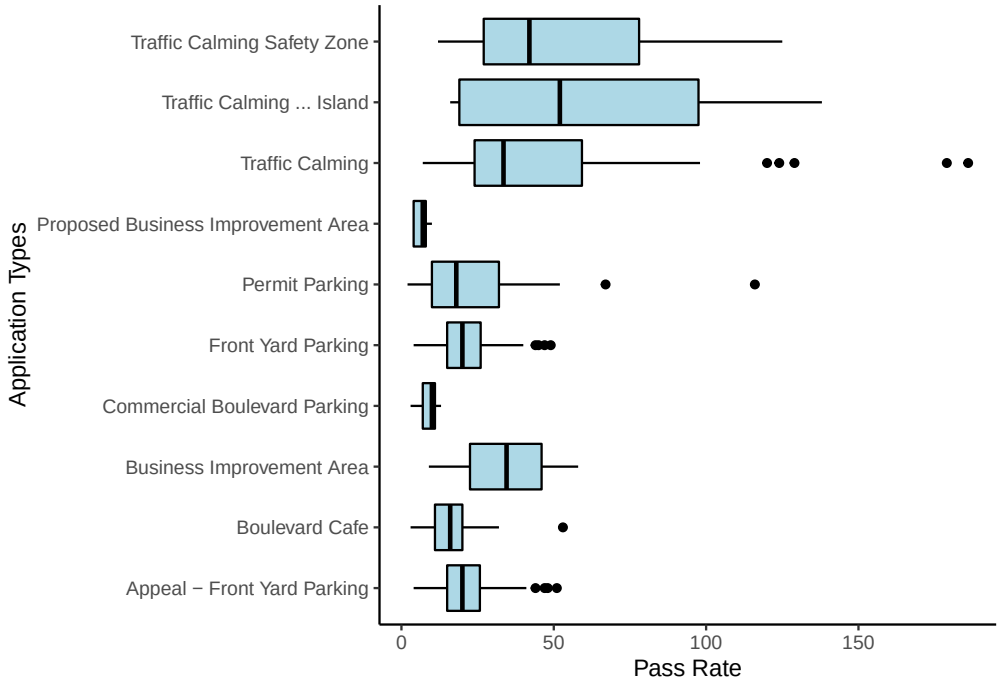

Figure 11: Side-by-side Boxplot of Application Type vs Pass Rate

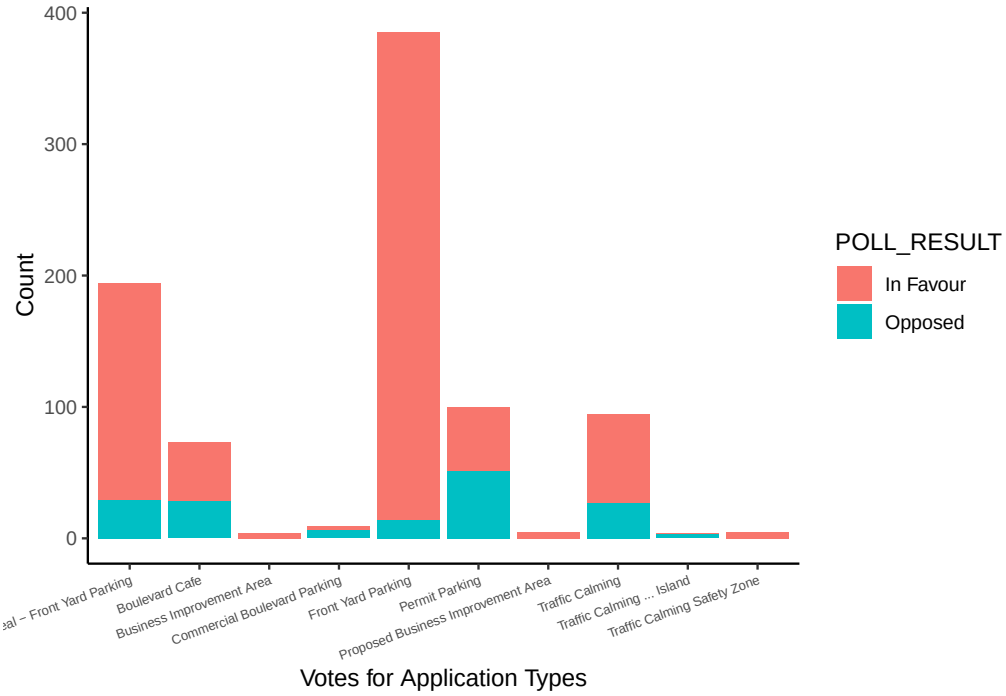

Figure 12: Poll Result towards Application Types

Table 2: ANOVA p-value Results

Model	ANOVA p-value
Auto Selected	0.6139
Manual Selected	0.2259

Table 3: Model Selection Data

Model	Adj. R^2	AIC
Start Model	0.9974608	1539.012
Auto Selected	0.9974639	1537.27
Manual Selected	0.997455	1537.418

Methodology and Model Selection

The paper tries to find the most influential factors among potential predictors selected before and then construct a simple linear regression model on those factors. To achieve this, from the previous explanatory description analysis, seven variables were considered to have higher correlation to the pass rate of poll results. They were: BALLOTS_IN_FAVOUR, BALLOTS_NEEDED_TO_PROCEED, BALLOTS_OPPOSED, BALLOTS_RECEIVED_BY_VOTERS, BALLOTS_SPOILED, FINAL_VOTER_COUNT and MISSING. In this case, the whole dataset was divided into training and testing for simulation, and a starter model was constructed with these seven predictors based on training data, and ready to be examined with model selection methods.

After checking the multicollinearity of the starter model, an auto reduced method was applied to select the significant variables. At the same time, by checking the significance of each variables p-value and common knowledge, a manual reduced selection was also done by hand. The two reduced regression models are listed below: -Auto

$$PassRate = \beta_0+\beta_1 InFavour+\beta_2 BallotsNeededProceed+\beta_3 Opposed+\beta_4 BallotsReceived+\beta_5 FinalVoter+\beta_6 Missing+\epsilon$$

-Manual

$$PassRate = \beta_0 + \beta_1 BallotsNeededProceed + \beta_2 BallotsReceived + \beta_3 FinalVoter + \beta_4 Missing + \epsilon$$

In order to compare both models and make a selection, an ANOVA Partial-F test was constructed to compare the p-value of model reduction. Also, an adjust R-squared summary and an AIC summary were applied to both model, in order to show the goodness of each reduction.

As a result, table 2 and table 3 have concluded the numerical results of each test. From table 2, both auto and manual selected model have very large p-value (0.6 and 0.2), in this case, they all do not reject the null hypothesis that at least one of the coefficients removed is significant. Thus, both models indicate that reduced is more efficient than starter model. As we look at table 3, both adjust R-squared and AIC values are quite similar for all three models, so they are considered to be equally efficient with this summary. Notice that the AIC values are extremely large and this will be discussed in the weakness section.

To conclude, since table 5 offers an equal comparison, the final model would be selected as the auto reduced, because of its larger p-value and more comprehensive variables.

P-value

A p-value is a statistical measurement used to validate a hypothesis against observed data. A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true. The lower the

p-value, the more likely the null hypothesis will be rejected. Usually, we consider the P-value as low when it is less than 0.05, which we say there are some evidence to reject the null hypothesis.

Partial F-test

A partial F-test is used to determine whether or not there is a statistically significant difference between a regression model and some nested version of the same model. If the p-value corresponding to the F test-statistic is below a certain significance level (e.g. 0.05), then we can reject the null hypothesis and conclude that at least one of the coefficients removed from the full model is significant.

Adjusted R-squared

Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new variable improves the model more than expectation, and decreases when a predictor improves the model by less than expected.

AIC

The AIC is designed to find the model that explains the most variation in the data, while penalizing for models that use an excessive number of parameters. The smaller the AIC, the better the model fits.

Results

Since the final model was determined through auto selected method, the predictors BALLOTS_IN_FAVOUR, BALLOTS_NEEDED_TO_PROCEED, BALLOTS_OPPOSED, BALLOTS_RECEIVED_BY_VOTERS, FINAL_VOTER_COUNT and MISSING were seem to have more significant linear relation with the response variable PASS_RATE. So the six variables out of 14 from the beginning are going to be applied in the simple linear regression model. After summarizing the output, the linear regression model equation is listed as follows:

$$Y = 0.491 - 0.023X_1 + 0.196X_2 - 0.021X_3 - 0.059X_4 + 0.539X_5 - 0.521X_6 + \epsilon$$

The variables and their representation:

- Y: Pass rate of polls
- X_1: Number of ballots received and marked "in favour"
- X_2: The number of ballots needed to proceed based on percentage of return
- X_3: Number of ballots received and marked "opposed"
- X_4: Number of ballots returned to City Clerk's Office
- X_5: Number of total voters on the final poll list
- X_6: Number of missing ballots

As we can see, the overall impact of these predictors are:

- If number of in favour ballots raises by 1 person, the average pass rate is expected to decrease 0.023 by quantity.
- If number of ballots needed to proceed raises by 1 person, the average pass rate is expected to increase 0.196 by quantity.
- If number of opposed ballots raises by 1 person, the average pass rate is expected to decrease 0.021 by quantity, which is quite similar to in favour effects.
- If number of ballots returned to the City Office raises by 1 person, the average pass rate is expected to decrease 0.059 by quantity.
- If number of total voters raises by 1 person, the average pass rate is expected to increase 0.539 by quantity.
- If number of missing ballots raises by 1 person, the average pass rate is expected to decrease 0.521 by quantity.

In addition, figure 13 has shown the goodness of fit for the final simple linear regression model. From the residual versus fitted we can see that there is no fanning pattern existed, which means the assumption of constant variance holds. The normal qq plot shows that the normality are satisfied except for one outlier point (543). This is also shown in the leverage plot, where most points are equally lying around 0.5 besides one leverage point (543). Also, the points in residual versus fitted are bit clustered on the left side, which shows independence needed improvement. These limitations will be further discussed in weakness and futrue steps.

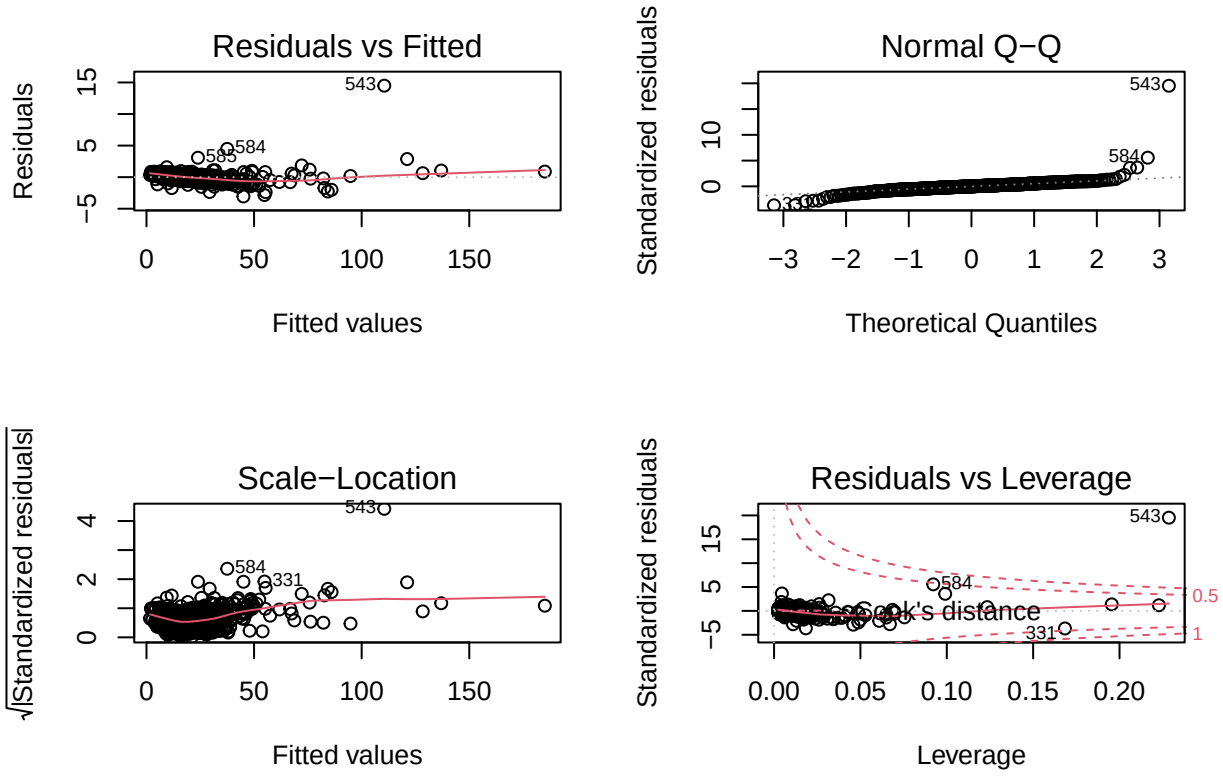

Figure 13: Goodness of Model

Discussion

The discussion section will be approached from two aspects, the first part will interpret the result more in detail and illustrate areas of interest. The second part will talk about limitations and future improvements for the analysis.

Conclusions

The paper has used the 14 chosen variables to examine their relationship with Toronto polls pass rate. Six predictors were modeled to have significant relationships with the response factor, they were number of in favour ballots, number of ballots needed to proceed, number of opposed ballots, number of received ballots, number of total voters and number of missing ballots. The final model has mostly captured the information provided from the dataset, and it is simple enough to interpret at the same time.

For instance, the pass rate of polls appears to be negatively correlated with ballots opposed and in favour quantities, and their values were very close (0.021 and 0.023). This may be interpreted as when both results increase at same value, the pass rate would be drawn down because of their equal distribution of voting. The number of missing ballots also obtain a negative relationship, since if the missing quantities increases, final counts could be underestimated and polls pass rate will decrease as well. What's more, both number of ballots needed to proceed and number of total voters have positive relation with polls pass rate. As more ballots come to proceeding based on percentage of return, the required pass rate would increase, and thus drives up the pass rates. And if total voters increase, the basic number of voting will simply go up and raise the polls pass rate.

In conclusion, the pass rate of a certain poll proposal will be more likely to achieve if more residents participate in voting the proposal. So the citizens in Toronto City with permission to vote should be more encouraged to do so. This could help the City better determine the next step policies and improve the potential issues

existing in each neighborhood.

Weaknesses and next steps

The limitation of the model mainly comes from two aspects. The first one would be that during the AIC test of model selection, all compared models have quite large AIC values, which shows that the model is not well fitted with parameter quantities. This may because of the relative high correlations within factors. More advanced model should help solve this limitation, and I will practice more statistic models in my future study.

Another limitation would be that the goodness of model assumptions may not hold too well (i.e. not independent and leverage point exists). The reason may come from that I have not applied transformation approaches to the model, since I feel it will reshape the model on purpose. Thus, some part of the plots may not look too neat, such as clusters on the left corner. Next time, I will do more investigation on different variables to exclude their correlation with each other.

Appendix

Datasheet

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
    - The dataset was created to explore what ballots predictors could have influence on the pass rate of polls conducted in Toronto, and how they impact it. The task is to use simple linear regression model to simulate a conclusion.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The polls data was established by Toronto Clerk's Office on behalf of Toronto residents, and it can help for monitoring government policy.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
    - The creation was funded by the City Government of Toronto.
4. *Any other comments?*
    - When an application for a proposal of the City is submitted, the City conducts a poll of people in the affected area. In order for a poll to be considered positive it must meet benchmarks as determined by specific by-law or city policy. If the result of the poll is positive, the application may proceed through the approval process. Depending upon the type of poll, final approval by City Council may be required.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
    - The instances represent the documented results of Toronto residents' voting. The types can mainly divide into: addresses, dates, application types, recorded ballots categories, poll results.
2. *How many instances are there in total (of each type, if appropriate)?*
    - There are 1071 instances
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
    - The dataset does not contain all possible instances, some instances such as detailed address may not be avaliable. Since it is conducted through the city and no larger set is represented. It is refreshed daily and the missing instances are not large.
4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
    - Each instance consists of 25 features in the dataset. There are 13 numerical features and others are categorical.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
    - The first column of each instance corresponds to its unique row identifier, which is the label associated with it.
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*
    - No individual instances are missing. Yet some of the instances may have certain missing features such as addresses.
7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

- Some instances may have same categorical features, such as same application types or polls closing date.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*
   - Recommended data splits were 70% training and 30% testing.
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
   - There are no errors, sources of noise, or redundancies in the dataset.
10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
    - The dataset is self-contained.
11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*
    - There is no confidential data, and the dataset is publicly available.
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
    - There are no offensive, insulting, threatening, or ansiety causing data.
13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
    - Only owners, residents or tenants of a property in the polling area can participate in polls voting. All participants must be 18 years old on or before the final day of the polling. Other than this, the dataset itself has no population divides.
14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
    - It is not possible to identify individuals.
15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*
    - There are no sensitive features in the dataset.
16. *Any other comments?*
    - None.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*
   - The data is collected by the City Office from established poll results. The data is directly observable with poll result counts.
2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*
   - The data was collected manually. The poll information was recorded and ballots results were counted.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

- The dataset is not a sample from a larger set.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*
   - The City Clerk's Office is responsible for conducting polls on behalf of City divisions. No compensation recorded.
5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*
   - The data was collected from April 1, 2015 to present. The recent timefram matches.
6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*
   - Ethical review processes were not conducted.
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
   - I obtained the data via Open Toronto Data website: https://www.toronto.ca
8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*
   - The data collection is mandatory for conducting polls, and participants in the polling area are mailed with a notice giving information about the poll and the deadline to submit a ballot.
9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*
   - Since participants' ballot voting results are necessary for polls construction. The individuals are consented to the collection and use of their data. The exact language to which consent was granted is not available.
10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
    - A mechanism to revoke consent was not provided.
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - An analysis of the potential impact of the dataset and its use on data subjects was not available.
12. *Any other comments?*
    - None.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
   - The data was obtained and loaded into R in CSV format.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
   - None.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
   - No preprocessing.
4. *Any other comments?*
   - None.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

- The part of the dataset was used for another paper to analyzing residents voting on traffic calming policy.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
   - https://github.com/KSaxophone/STA304-Paper1-code
3. *What (other) tasks could the dataset be used for?*
   - The dataset can be used for exploring voting on different application types.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
   - None. The dataset is public available and updated frequently.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
   - The dataset should not be used for policy decisions in other cities or countries, since the data is Toronto specific.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
   - No, the dataset is fully personal using.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
   - The dataset will be distributed through Github.
3. *When will the dataset be distributed?*
   - The dataset will be distributed on April 27, 2022.
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
   - The dataset will be released under the MIT license
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
   - There are no restrictions
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
   - No such controls or restrictions are applicable.
7. *Any other comments?*

- None.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
   - The City Clerk's Office
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
   - Can be contacted via email: opendata@toronto.ca
3. *Is there an erratum? If so, please provide a link or other access point.*
   - There is no erratum available currently.
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

- The dataset can be updated daily by the City Office. The dataset faces the public so consumers can check for updates freely.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
   - The dataset records Toronto residents polls voting results for policy improving, and there are no applicable limits.
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
   - The older versions would not be hosted.
7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*
   - There is no mechanism for other users to contribute to the dataset at the moment. If they have questions, they can contact with City Office through email opendata@toronto.ca.

Additional details

Code and data are available at: https://github.com/KSaxophone/Polls-conducted-Toronto.git

You are welcome to use the repository to reproduce the paper if needed. The entire paper was produced through RMD files (Allaire et al. 2022). R statistical programming language (R Core Team 2021) is used for analyzing, and the packages `tidyverse`(Wickham et al. 2019), `dplyr` (Wickham et al. 2021) are used for data visualizing and data manipulating in this project. The package `knitr`(Xie 2014) is used to knit the R markdown file to pdf form, figures and tables will be produced and generated through packages `ggplot2` (Wickham 2016) and `ggpubr` (Kassambara 2020), `patchwork` (Pedersen 2020) as well as kableExtra (Zhu 2021). What's more, the package `car` (Fox and Weisberg 2019) is used to creating regression models of the variables, and package `broom` (Robinson, Hayes, and Couch 2022) is used to generate tidy tibbles.

References

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2022. *Rmarkdown: Dynamic Documents for r*. https://github.com/rstudio/rmarkdown.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal*.

Kassambara, Alboukadel. 2020. *Ggpubr: 'Ggplot2' Based Publication Ready Plots*. https://rpkgs.datanovia.com/ggpubr/.

Pedersen, Thomas Lin. 2020. *Patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork.

Polls Conducted by the City. 2022. City Clerk's Office. https://open.toronto.ca/dataset/polls-conducted-by-the-city/.

Polls Regarding Changes in a Neighbourhood. 2022. https://www.toronto.ca/city-government/planning-development/polls-regarding-changes-in-a-neighbourhood/.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, Alex Hayes, and Simon Couch. 2022. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*.