Analysis factors affecting voter participation rate*

A Case Study of Toronto Poll Participation

Ruiying Li

December 3, 2024

This study investigates the factors influencing voter turnout in local elections using polling data from the Toronto municipal government. By analyzing over 1,000 records, we identify drivers of voter participation. The analysis finds that increased ballot distribution and a higher number of ballots cast are associated with higher turnout rates, while the number of eligible voters has a smaller impact. These findings suggest practical strategies to improve civic engagement and strengthen public decision-making on local issues.

1 Introduction

Overview

Municipal polls serve as a mechanism for engaging citizens in local decision-making on issues such as front yard parking, permit parking, and other community-specific topics. Unlike general elections, these polls focus on local concerns, giving residents a direct voice on matters that impact their daily lives. However, despite their importance, participation rates in these polls often remain low, raising questions about what drives or hinders voter engagement. This study examines the factors influencing voter participation in municipal polls, focusing on the relationship between logistical variables such as the number of ballots distributed, ballots cast, and participation rates.

Estimand

The estimand of interest is the municipal poll participation rate, defined as the proportion of ballots cast relative to ballots distributed. By analyzing patterns across different poll topics, we aim to understand how logistical and contextual factors shape voter behavior.

[*] Code and data are available at: https://github.com/Liruiying0414/Analysis-of-Toronto-polls-participating-rate

Improving participation in municipal polls ensures a more representative decision-making process, aligning with democratic principles. It can also inform strategies to optimize voter outreach and engagement, enhancing the effectiveness of local governance.

Results

Our analysis shows two key findings. First, logistical factors such as the number of pass rate and the final voter count significantly influence participation rates. Polls with higher ratios of pass rate to eligible voters tend to see higher engagement, indicating the influence of outreach efficiency and perceived relevance. Second, the type of issue being polled is important: polls addressing contentious or directly impactful issues, such as parking permits, tend to have higher participation rates compared to broader or less immediate topics. These findings highlight the interplay between logistical organization and issue salience in determining voter turnout.

1.1 software, data and package used

This paper use the statistical programming language R(R Core Team 2023), and data from Opendata Toronto(Open data Toronto 2015), and following packages: tidyverse(Wickham et al. 2019), dplyr(Wickham 2023a), ggplot(Wickham 2023b), rstanarm(Goodrich et al. 2022),knitr(Xie 2023), modelsummary(Arel-Bundock 2022), patchwork, arrow(Richardson et al. 2024) and bayeplot(Gabry and Stan Development Team 2023).

1.2 Paper structure

The remainder of this paper follows a structure Section 2 about the data and methodology, including details on cleaning and processing the dataset. Section 3 presents the results of our analysis, highlighting key factors influencing voter participation. Section 4 discusses these findings in the broader context of voter engagement research. Finally, Section 5 concludes with practical recommendations for policymakers and avenues for future research. Section A for additional information about survey and model information details.

2 Data

2.1 Overview

This analysis conducted in a statistical programming language R(R Core Team 2023), utilizes the dataset from Opendata Toronto on municipal-level voting records and application requests concerning various citizen issues, including parking permits, zoning, and other local concerns.(Open data Toronto 2015). The original dataset includes 25 variables, and 1296 observations in total, and this paper only interested in the following 7 variables from 2015,July to 2024,October. Through using tidyverse(Wickham et al. 2019) to clean data, such as type of application,potential voters, distributed ballots,ballots cast, final voters count, open date, end date, pass rate and participant rate, and only have 1069 observations, and use rstanarm(Goodrich et al. 2022) to build model function.

Municipal voting data is a rich source for understanding democratic engagement and decision-making at the local level. Through the dataset, this paper aim to analyze trends and factors affecting voter turnout in urban governance contexts,enabling a closer examination of how specific issues resonate with voters by modelsummary(Arel-Bundock 2022) and specify municipal contexts gives unique value in studying localized participation factors by sketching graph and using ggplot(Wickham 2023b) and bayesplot(Gabry and Stan Development Team 2023).

2.2 Measurement

The dataset captures the relationship between local issues and voter engagement. The variable application_for links each voting record to a primary issue, such as permit parking or zoning. Constructed variables, such as participation_rate, were derived to quantify voter engagement as the ratio of ballots_cast to ballots_distributed. To ensure data quality, extensive cleaning was performed to filter incomplete records and resolve discrepancies. The final dataset provides a foundation for analyzing the impact of local issues on voter turnout. Detailed cleaning data are shown Table 2 and Table 3

2.3 response variables

The primary response variable in this study is participation_rate, calculated as the ratio of ballots_cast to ballots_distributed. This variable quantifies the proportion of distributed ballots that were successfully returned, serving as a proxy for voter engagement. Table 1 are the key features of this variable, and Figure 1 shows its overall distribution.

Table 1 summarizes the descriptive statistics for voter participation rate, including mean (49.94%), median (49.12%), minimum (21.38%), and maximum (100%), and most areas have a voter participation rate close to 50%, indicating the central tendency of the data. Figure 1 shows many areas have participation rates mainly between 40% and 60%, forming a symmetrical bell-shaped distribution. This distribution suggests that a linear model may be appropriate.

Table 1: key variable for response rate

Metric	Participation_Rate
Mean	49.94
Median	49.12
Min	21.38
Max	100.00

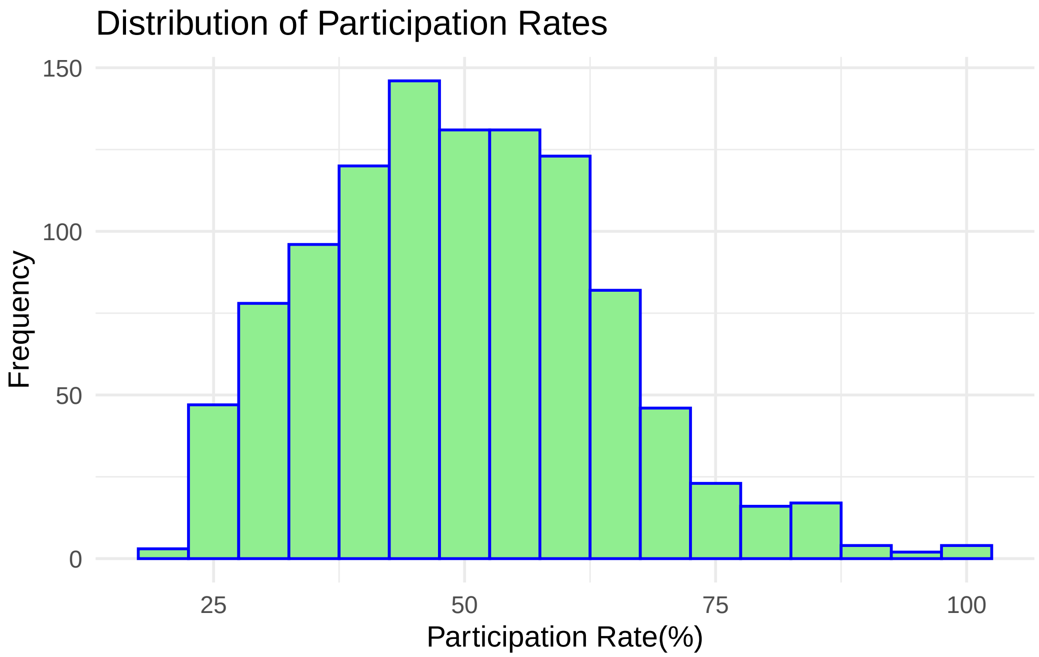


Figure 1: polls response rate

2.4 predictor variable

The model incorporates a set of key predictor variables, including 5 numeric variables and 1 categorical variables

2.4.1 Categorical variable:

application_for: A categorical variables and indicates the primary issue associated with a voting process and explores whether specific local issues influence voter turnout rates.

Figure 2 shows the frequency of issues related to various application types.”Front Yard Parking” is the most common issue, potentially affecting voters’ daily lives and indirectly influencing their likelihood of voting, this different public service issues have varying levels of impact across regions, which may correlate with voter behavior.

2.4.2 numerical variables:

potential_voters:it represents the total number of individuals eligible to participate in the vote for application. **ballots_distributed**: Total number of ballots distributed for voting and indicates administrative outreach. **ballots_cast**: The number of valid ballots returned and

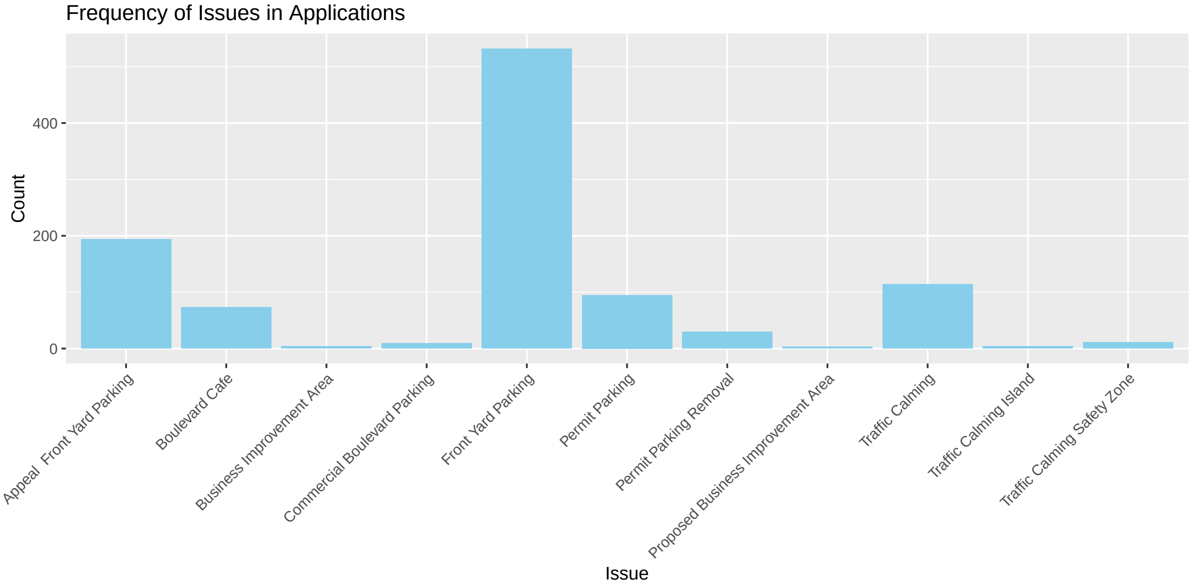


Figure 2: types of applications

measure the actual participation of voters, it is a key variable for calculating voter participate rates. **final_voter_count** Verified count of voters who participated by providing a refined measure of participation after accounting for errors or inconsistencies in ballots. **pass_rate**: it represents number of returned ballots needed for a positive poll result.

Figure 3 is a summary table for this four numeric variables, and shows potential voters, ballots distributed, and ballots cast exhibit right-skewed distributions, and pass rate is concentrated in a lower range. Final voter count density shows a unimodal distribution with the peak near lower values. (Figure 3) shows voter population and ballot numbers are unevenly distributed, with a few regions having significantly higher counts, and voter counts tend to fall within a common range. Pass rates are relatively low in most regions, indicating many ballots were either rejected or not successfully cast.

By sketching the graph, this is helpful to better understanding discrepancies,improving the accuracy of voter participation estimates. Together, these variables provide a foundation for analyzing voter participation patterns, thinking issue-specific engagement, administrative processes, and voter behaviors.

2.5 relation between vairables:

In this paper, ballots_cast is a main variable in analyzing relationship stems from its role as a direct measure of returned votes, it is a main factors influencing voter participation rate. As a core component of response rate (response_rate), ballots_cast also serves as a reference point for analyzing interactions with other variables, such as constituency size (potential_voters),

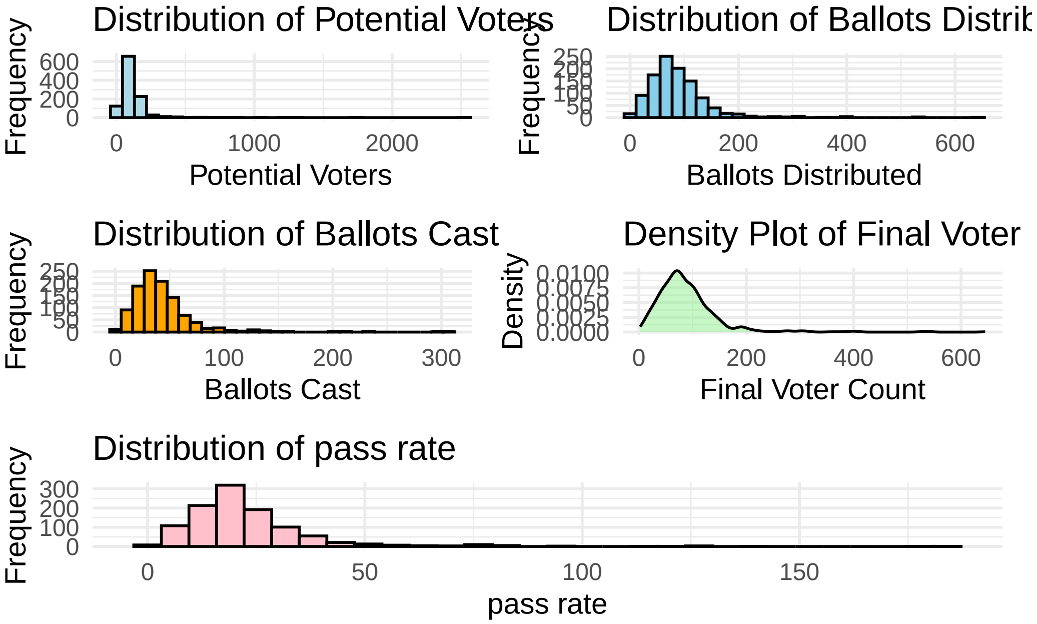


Figure 3: distribution of four numeric variables

final voter count (final_voter_count), positive result pass rate(pass_rate), and application type (application_for). Finding the relationships between ballots_cast and these variables allows us to break down the factors influencing voter participation, highlight which variables drives or hinder turnout

2.5.1 relation between ballot cast and final voter count

Figure 4 defined the positive linear relationship between ballots cast and the final voter count, and there is a strong positive correlation between the number of ballots cast and the final voter count, showing the number of ballots distributed and received reflects voter turnout. The clustered points suggest proportionality in most regions, but some outliers warrant further investigation.

2.5.2 relation between ballots cast and potential voters

Figure 5 illustrates the relationship between the number of ballots cast and the potential voters. The blue fitted line indicates the general trend in the data. There is a weak but positive correlation between ballots cast and the number of potential voters. Most data points are near the lower end, indicating a large number of areas with relatively few potential voters. Outliers with high potential voters but low ballots cast could reflect regions with barriers to voting or lack of voter participation.

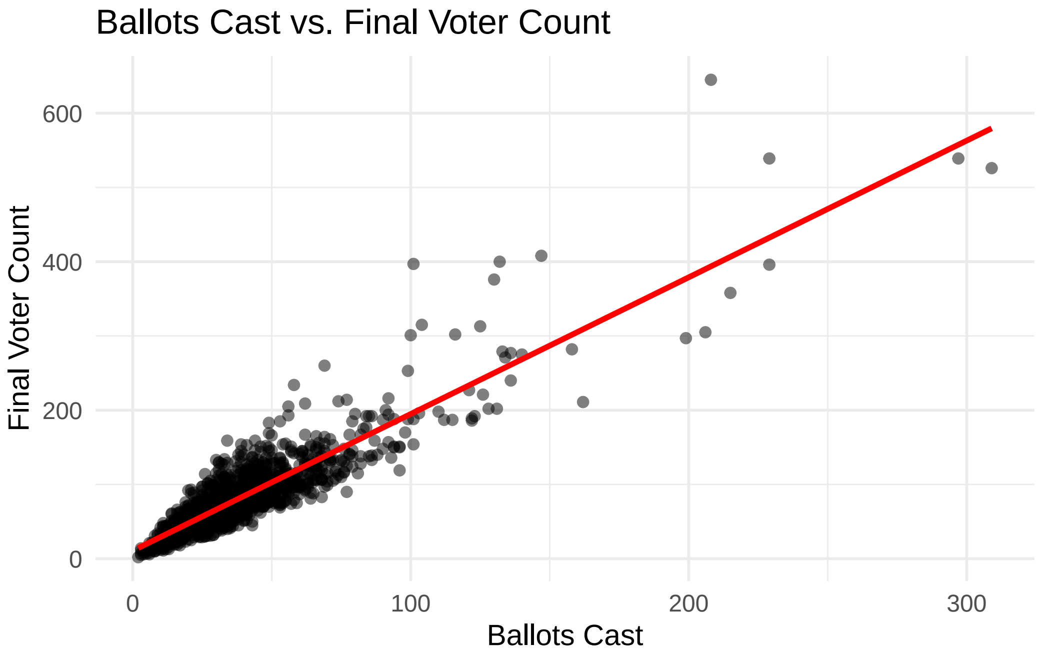


Figure 4: relationship between ballots cast and final voter count

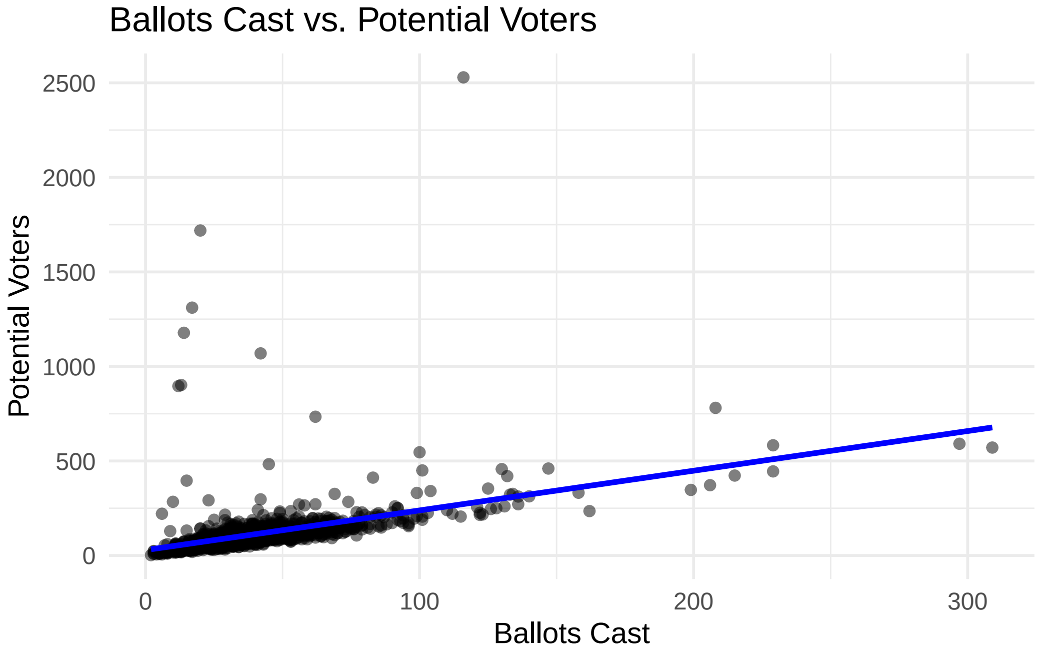


Figure 5: relationship between ballots cast and potential voters

2.5.3 relationship between ballost cast and pass rate

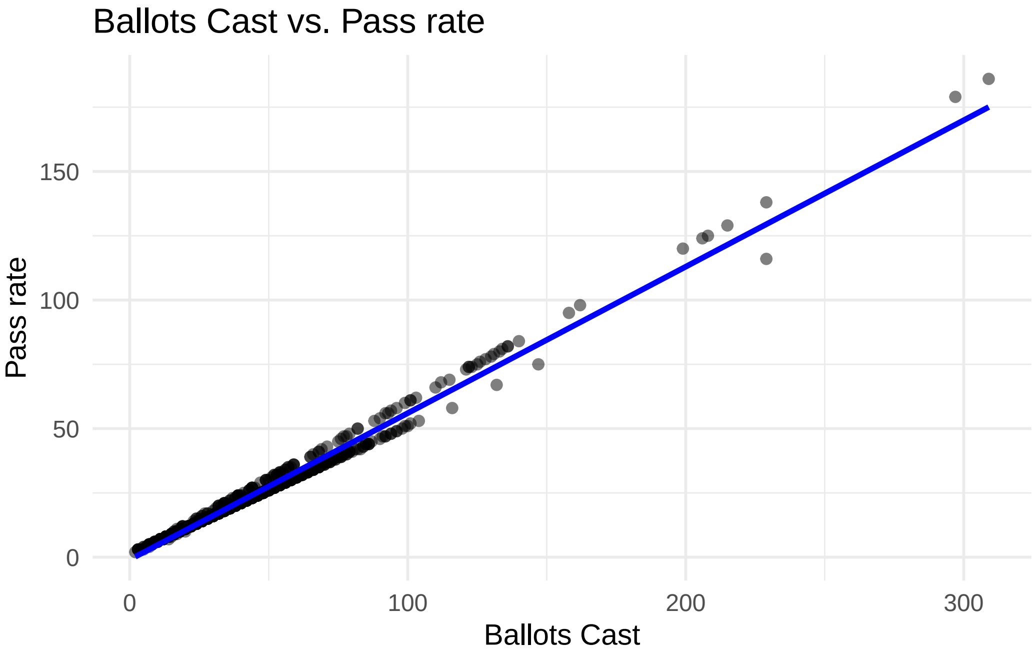


Figure 6: relationship between ballots cast and pass rate

Figure 6 illustrates the positive relationship between ballots cast and the pass rate. As the number of ballots cast increases, the pass rate also tends to rise, suggesting that a higher ballot count improves voting efficiency or ballot validity. The relationship could guide strategies to reduce invalid or spoiled ballot

2.6 application for with ballost cast

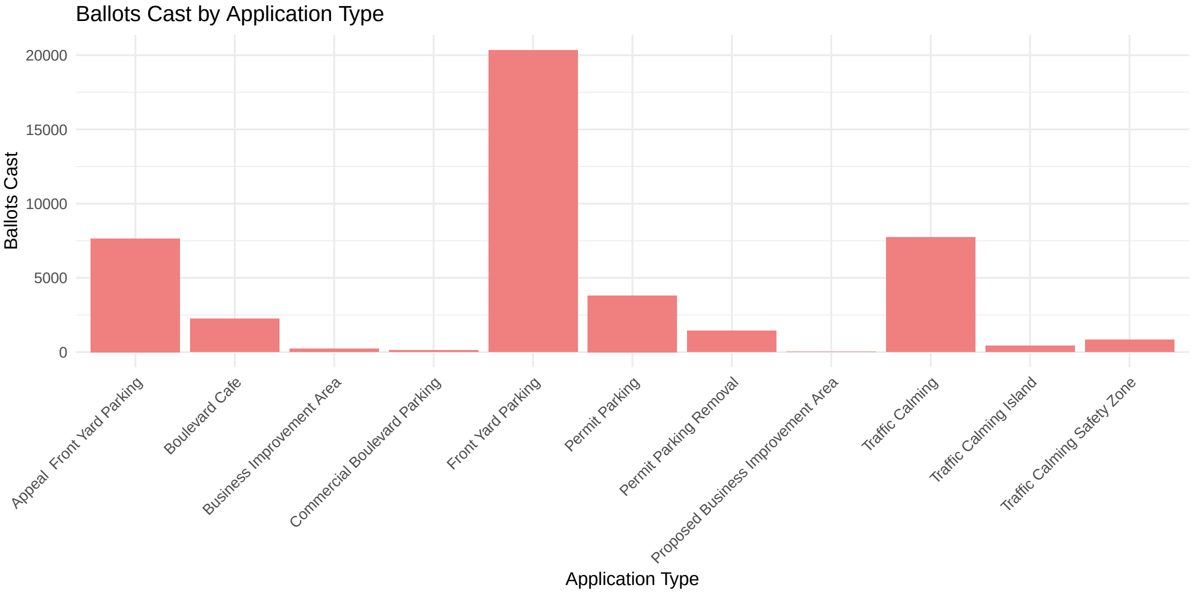


Figure 7: relationship between ballots cast and application types

Figure 7 displays the number of ballots cast for each application type. “Front Yard Parking” stands out with the highest number of ballots cast, application types like “Front Yard Parking” are associated with higher voter turnout, potentially reflecting voters’ concerns about specific local issues. The variation across application types suggests a need to examine their influence on voter behavior.

3 Model

3.1 Model Overview

The goal of this analysis is to examine the factors influencing voter participation, particularly focusing on the 'voter participation rate. The dependent variable in this study is the Response Rate, which is the ratio of Ballots Cast to Ballots Distributed, serving as an indicator of voter engagement. The analysis aims to assess how other factors, such as the final number of voters, the potential number of voters, application types and voting pass rate, impact the voter participation rate. We plan to employ a bayesian model to explore the relationship between the response rate and several predictor variables. Bayesian model is a straightforward and effective tool for modeling continuous dependent variables with multiple independent variables, helping to uncover the interactions between various factors. Model details, checking, and diagnostics are presented in Section A.

3.2 Bayesian Model Set-Up

In this model, we define y_i as the percentage of voting participation for unique poll$i$, calculated as the ratio of ballots cast to ballots distributed. Because response_rate already captures the relationship between ballot_cast and ballot_distributed, including both ballot_cast and ballot_distributed as separate predictors would introduce multicollinearity into the model,so ballots_cast and ballots_distributed would not be our predictors in model.

The predictors in the model are:

x_1 : final_voter_count,represent total number of final voters.

x_2 : potential_voters,the number of potential voters within the polling boundary.

x_3 : application_for, represents Categorical variable representing the type of applications.

x_4 : pass rate,represent the number of returned ballots needed for a positive poll result.

The model and the prior distribution expressed as:

$$
\begin{aligned}
\text{response\_rate}_i | \mu_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \beta_0 + \beta_1 \cdot \text{potential\_voters}_i + \beta_2 \cdot \text{final\_voter\_count}_i \\
&\quad + \beta_3 \cdot \text{application\_for}_i + \beta_4 \cdot \text{pass\_rate}_i \\
\beta_0 &\sim \text{Normal}(50, 20) \\
\beta_1, \beta_2, \beta_3, \beta_4 &\sim \text{Normal}(0, 5) \\
\sigma &\sim \text{Exponential}(1)
\end{aligned}
$$

where β_0 is intercept, $\beta_1, \beta_2, \beta_3, \beta_4$ are the coefficients for each predictor, and ϵ_i represents the standard error term, assumed to be normally distributed with mean 0 and standard deviation. Diagnostic plots checks were performed to assess whether the model is under a good prior assumptions.

The model is applied in R (R Core Team 2023) with using the 'rstanarm' package (Goodrich et al. 2022) and modelsummary package from(Arel-Bundock 2022), with glm() baysian model function to create our bayesian linear regression, using default priors for ease of interpretation and computational efficiency.

3.3 Model justification

In this analysis, we used a Bayesian framework to estimate the relationship between response rate and predictor variables. Bayesian methods enable us to combine prior information while better quantifying uncertainty.

In the model, the response rate is the observed response rate for the i-th poll, modeled as a Normal distribution with mean μ_i and standard deviation σ, mu_i is a linear combination of the intercept β_0. Since response_rate already captures the relationship between ballot_cast and ballot_distributed, including both ballot_cast and ballot_distributed as separate predictors would introduce multicollinearity into the model, and can lead to unstable estimates and difficulties in interpreting the individual effects of these variables.

Therefore, to avoid multicollinearity and maintain a model that focuses on the key determinants of voter participation, we have chosen to include response_rate as a response variable instead of both ballot_cast and ballot_distributed. This approach simplifies the model while still reflecting the underlying voting behavior. When response rate is expressed as a percentage, The intercept is centered at 50%, representing an assumption that the average response rate is near the midpoint, and standard deviation of 20 reflects reasonable variation, allowing the response rate to fall between 10% and 90% under typical situation.

The predictor potential voters, final voter count and application for,weighted by their respective coefficients β_1, β_2, and β_3. Priors for β_0,β_1, β_2, and β_3 are specified as Normal distributions with chosen means and standard deviations, the priors can remain consistent but should match the influence expected on the response rate, and assumes a moderate effect of predictors on the response rate, allowing larger deviations if necessary. This prior distribution provides a weakly informative priors balance prior knowledge with flexibility, preventing overfitting in small or noisy datasets and centered priors for coefficients reflect an initial assumption of no large effects, allowing the data to drive inference. The prior for σ regularizes variance estimates and avoids implausible error ranges.

3.4 Model limitations and weakness

The bayesian model also has many limitations and weakness.Bayesian models rely on prior distributions, which introduce a level of subjectivity. While priors can incorporate domain knowledge, inappropriate or overly restrictive priors can bias the result, the choice of priors may affect posterior estimates, particularly if the data is sparse or the priors are overly informative, When the model struggles to account for extreme values (outliers) in the data. This could result in the model underperforming in situations involving rare or extreme voter participation patterns. Also,with numerous predictors such as application_for_* variables, there is a risk of overfitting, especially if the dataset is small or noise, and this model assumes a linear relationship between predictors and the response variable, and might oversimplify complex real-world dynamics, the model assumes additive effects and does not include interactions between predictors such as how potential_voters might interact with specific application_for_ variables,This simplification might miss important relationships, reducing the model's ability to explain complex voter behavior patterns.

4 Results

4.1 Overview

the results of the regression and variable analysis for voter participation are presented. The goal of this analysis was to examine how various factors, such as the number of potential voters, final voter count, application types, and pass rate, influence the response rate in the context of voting participation. Table 4 shows the summary statistics of the key variables used in the analysis. For instance, the variable response_rate has a mean of 49.9% with a standard deviation of 14.4%. The distribution of potential_voters is skewed due to a few extreme observations, with a maximum value of 2529

4.2 Regression results

Through looking the summarized regression table in Table 5, it analyzes the regression coefficients, standard errors, and significance levels for the model examining the relationship between response rate and several predictors. The positive coefficient for potential voters (0.010) suggests a small but statistically significant positive relationship between the number of potential voters and the response rate. However, the small effect size indicates that the increase in the response rate is marginal as the number of potential voters increases, and the negative coefficient for final voter count is -0.380 suggests that a higher final voter count is associated with a lower response rate. This may reflect the fact that larger electorates tend to experience lower individual participation, possibly due to a perceived dilution of individual impact. Several application types show extremely huge negative effects on the response rate, such as Boulevard cafe and Business improvement area applications, Traffic calming and Commercial boulevard parking show significant negative effects, with coefficients of -4.533, -11.022, -2.940 and -8.467 respectively, suggesting that these applications are associated with lower voter participation. Others such as Traffic Calming Safety Zone, Front Yard Parking shows high positive effects on the response rate, with coefficients of 1.421, 6.897 respectively. The pass rate variable is statistically significant and has a positive coefficient 1.298 and indicates a higher pass rate is associated with an increased response rate, this suggests voters may be more likely to participate when the likelihood of the poll succeeding is higher.

4.3 Variable results

For Figure 8 shows the relationship between the number of potential voters and the response rate. The blue regression line indicates a slight negative trend. The number of potential voters increases, the response rate tends to decrease. This suggests that larger populations may face systemic challenges, such as administrative inefficiencies or voter disengagement. most of the points near the left side suggests that most areas have smaller voter populations with varying response rates. The shaded confidence interval around the regression line highlights the uncertainty in the trend for regions with higher potential voters.

Figure 9 shows the relationship between the final voter count and response rate. The negative slope suggests an inverse relationship. Areas with higher voter counts tend to have lower response rates, potentially reflecting challenges like resource allocation in larger populations or voters group. This trend indicates the importance of focusing voter mobilization efforts on high-population areas.

Figure 10 demonstrate a positive relationship between pass rate and response rate. Regions with higher pass rates tend to have higher response rates, suggesting that ensuring ballot validity can enhance voter engagement. Improvements in ballot processing or handling could indirectly increase voter response rates.

Figure 11 shows voter participation rates vary significantly depending on the type of application. issues that directly affect residents' lives, such as "Front Yard Parking," may stimulate higher voter participation rates, while other relatively less related application types have relatively lower participation rates. This helps to formulate useful policies that encourage voter participation, especially by focusing on issues that directly affect residents' lives, in order to increase public voting enthusiasm.

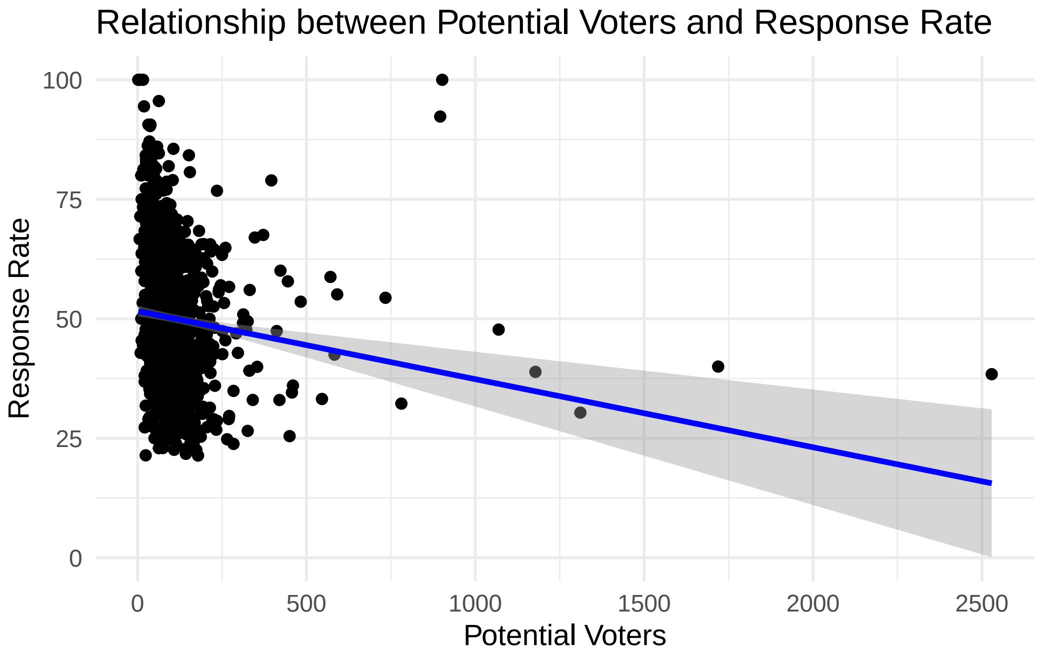


Figure 8: Relationship between Potential Voters and Response Rate

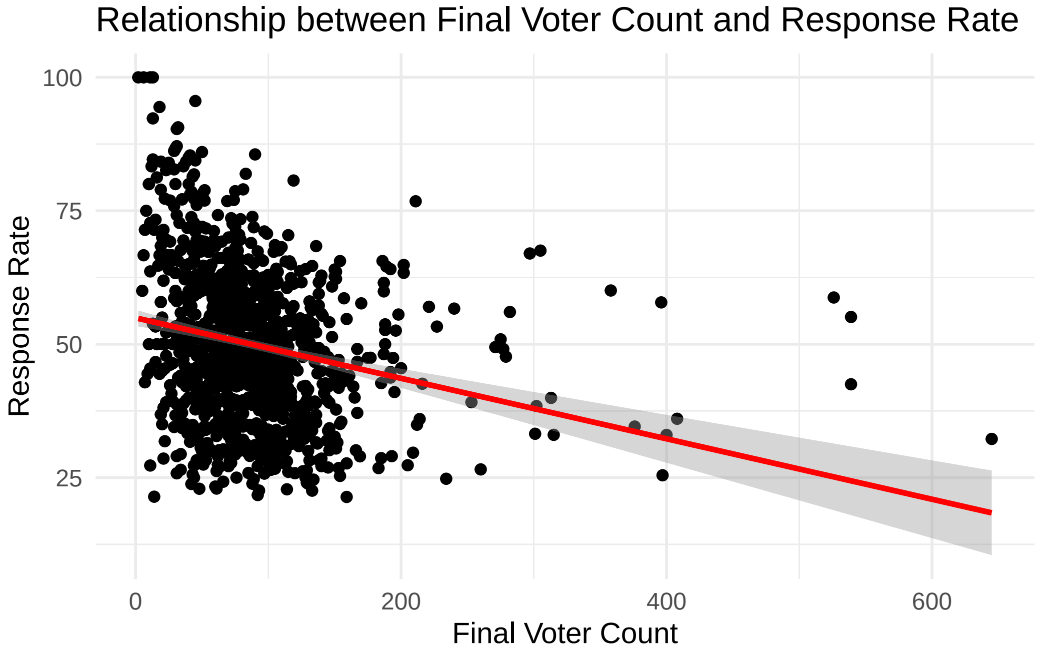


Figure 9: Relationship between Final Voters Count and Response Rate

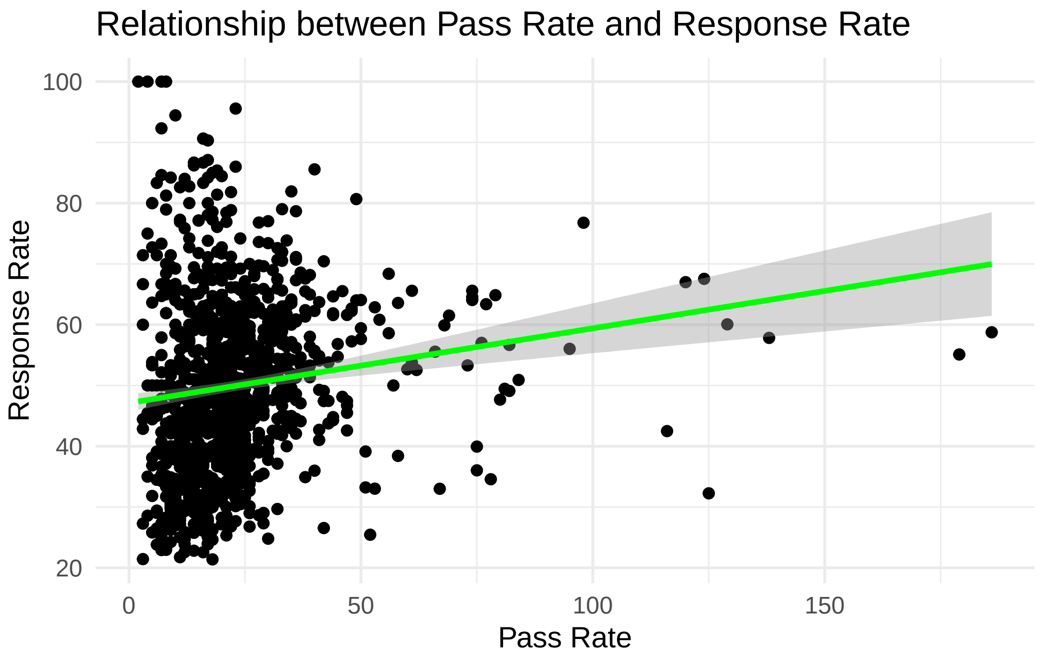


Figure 10: Relationship between Pass rate and Response Rate

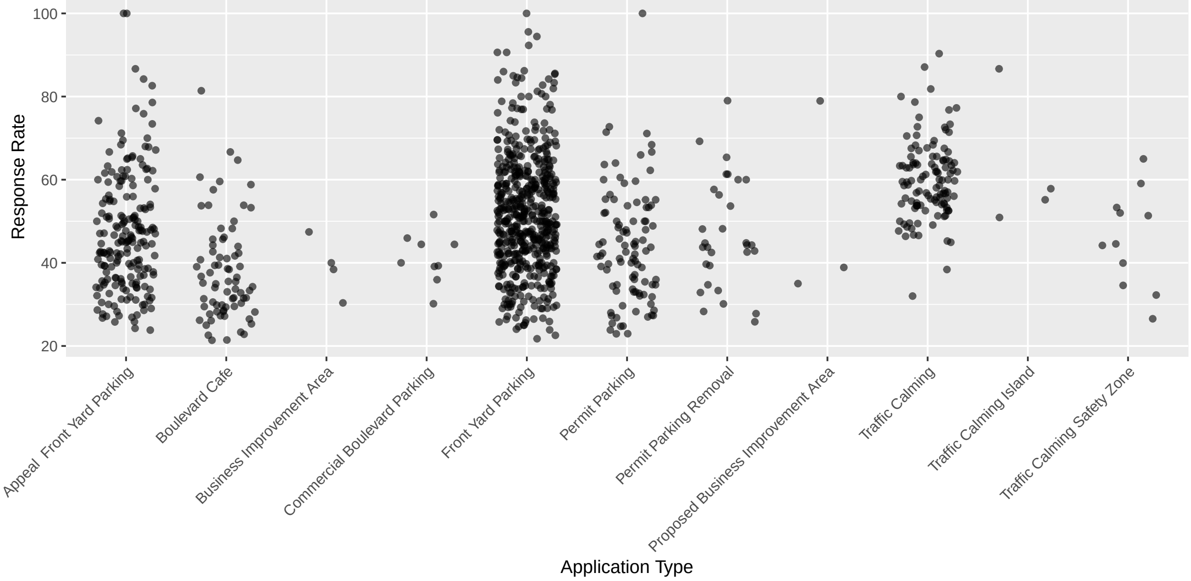


Figure 11: Relationship between Application types and Response Rate

5 Discussion

5.1 Overview

This paper examines factors influencing voter participation in municipal polls. By analyzing over 1,000 records from Toronto's local government, we gain insights into how logistical and contextual elements affect voter engagement. Specifically, the study demonstrates that factors like ballot distribution, application type, and pass rate play important roles in shaping voter turnout. This analysis provides a clearer understanding of the logistical dynamics that can either encourage or hinder voter participation.

5.2 Thoughts of voter bahavior

One key conclusion is that logistical efficiency, we learn that factors like the final voter count and pass rates play significant roles in shaping voter participation rates. Specifically, the results show a negative relationship between the final voter count and response rates, suggesting that higher final voter counts might indicate challenges in getting voters to participate. The pass rate shows a positive correlation with response rates, implying that when the pass rate is higher, voters are more likely to engage in the process. Additionally, the type of issue being voted on plays a major role. Issues that are more directly impactful on residents' lives—such as permit parking or front yard parking—tend to receive higher participation rates, Understanding these relationships can help policymakers better design and promote initiatives that encourage voter engagement.

5.3 Weakness and limitations

Although bayesian regression models are valuable tools, there are several limitations in this study that must be addressed. Firstly, the variable for potential voters did not significantly affect response rates, suggesting it may not be a strong predictor of voter participation. This could be because the number of potential voters is a broad measure and might not fully capture the factors that actually influence voter engagement.

Secondly, bayesian regression model assumes linear relationships between variables without consider complex interactions. In reality, the factors influencing voter turnout are likely more intricate, such as the interaction effects of socio-economic factors, which may not be fully reflected in the model. Additionally, there may be omitted variables, and even when multiple explanatory variables are included, the model might still miss key factors. Omitting important variables can introduce bias into the model. For example, factors like socio-economic background, media coverage, and political climate could affect voting behavior, but if these are not considered in the model, the results may be incomplete or inaccurate.

Thirdly, different models have varying assumptions, and these differences can influence the interpretation of the results. The variables related to applications are broad, and more specific data could be necessary to understand why certain applications have a greater impact on participation. Furthermore, the regression analysis was based on existing data, and the lack of external data for validation could undermine the findings. Data from different regions or time periods might be needed to assess the generalizability and stability of the model, as well as to examine how sensitive the model parameters are to different choices.

5.4 Future step

Future research could focus on integrating additional influencing factors into the model, particularly those not fully addressed in this study. Advanced methods such as non-linear models, mixed-effects models, or machine learning techniques could be employed to improve the accuracy of predictions regarding voter behavior. Moreover, expanding data collection to include more diverse and geographically representative samples would help test the model's broader applicability and external validity. Incorporating more detailed information on voters' socioeconomic backgrounds, cultural differences, and political attitudes could enhance the model's precision. Additionally, future studies could move beyond correlational analysis and explore causal relationships, using experimental designs or natural experiments to better understand the direct effects of various factors on voter turnout and participation.

A Appendix

A.1 Appendix A: Idealized methodology for participate rate

The following study utilizes data from municipal voter participation polls, which capture both individual and collective voting behaviors by Opendata Toronto(Open data Toronto 2015). In this study, voter participation is analyzed using data on various factors influencing response rates. One of the critical components of our analysis is understanding how the survey data was collected and how the sampling methodology impacts the findings. Voter participation data is often collect from local government polls or surveys, where the population of interest consists of residents within specific boundaries. The survey design shows administrative records rather than self-reported surveys, minimizing certain biases such as social desirability bias, but introducing challenges such as incomplete records or non-uniform data collection practices.

A.1.1 Survey Objectives

The primary objective of this survey is to understand the factors influencing voter participation rates, including socio-demographic characteristics, voting motivations,voting preferences and perceived barriers. By collecting data on these dimensions, this study seeks to identify actionable findings that can inform policy interventions aimed at increasing voter turnout and engagement.

Specifically, the survey aims to:

- Examine the relationship between voter turnout and socio-economic variables such as age, education, and income.
- Investigate potential obstacles that discourage participation, such as time constraints or lack of trust in the electoral system.
- Analyze how motivations, such as civic responsibility or political alignment, impact voter turnout.

Provide data for model validation and testing the accuracy of our regression findings.

A.1.2 Sampling approach

The idealized sampling approach ensures representativeness and reliability of the data collected. A stratified random sampling method will be employed to capture diverse perspectives across different demographic and geographic segments. Strata will include:

- Age groups: Ensuring representation from young voters (18–30), middle-aged voters (31–55), and senior citizens (56+).

- Geographic regions: Urban, suburban, and rural voters to account for geographic variations in voter participation.
- Socio-economic status: Income and education level will be used to ensure balanced representation across socio-economic classes.

A minimum sample size of 1,000 respondents is proposed, ensuring sufficient statistical power to detect meaningful relationships between variables while accounting for non-responses.

A.1.3 Respondent recruitment

Respondents will be recruited through:

- Online Panels: Partnering with survey platforms to access a broad pool of respondents.
- Community Outreach: Distributing survey links through community organizations, local newsletters, and public libraries.
- Social Media Campaigns: Targeted advertisements on social media platforms to reach younger and technologically savvy voters.
- Randomized Mail Invitations: Sending invitations to participate in the survey via mail to ensure inclusivity of offline populations.

To incentivize participation, respondents will be entered into a raffle for gift cards or receive small monetary rewards. This approach minimizes selection bias while enhancing response rates.

A.1.4 Data validation

The validity of collected data will be ensured through pre-testing the Survey, conducting a pilot survey with 50 respondents to identify and correct unclear or ambiguous questions. Response Consistency Checks,Employing built-in validation questions to identify inconsistent responses such as “Have you voted in the last election?” vs. “In which year did you last vote?” IP and Geolocation Checks is also considerable, and verify the authenticity of responses by detecting duplicate entries or geographically inconsistent data, and post-collection review, using automated tools to clean and validate data for missing or extreme values.

A.1.5 Weighting and Data Adjustments

Post-survey weighting will be applied to align the sample distribution with population demographics, ensuring representativeness. Adjustments will be made for underrepresented Groups to apply weights to correct for over- or under-representation of specific age groups, geographic areas, or socio-economic categories.For Non-response Bias, using statistical techniques like raking or post-stratification to adjust for differential response rates among subgroups. Survey Mode Effects,Incorporating adjustments for potential differences between online and offline respondents. These adjustments ensure the findings can be generalized to the broader population while maintaining data integrity.

A.1.6 Survey design

The survey is structured to maximize respondent engagement and data quality. It includes the following sections:

Background Information: Collect basic demographic data of participants (such as age, gender, education level, income level) and whether they are registered voters.

Community Participation: Understand residents' participation in community activities, including whether they have voted or participated in public policy discussions.

Type of Application Cognition and Attitude: Design specific questions for application types to explore whether respondents have been exposed to or submitted such applications, as well as their satisfaction and transparency with the approval process.

Voters' Willingness to Participate: Explore the specific impact of different application types on voter participation rates, such as whether the application process is perceived to affect their trust in the government.

Thank you and contact information: Thank you to the respondents for completing the survey and providing their contact information for future inquiries.

A.1.7 Tradeoffs and limitations

Despite careful planning, this methodology has certain tradeoffs and limitations:

- Self-reported Data: Responses are subject to recall bias or social desirability bias, particularly in questions about voting behavior or motivations.
- Non-response Bias: While weighting adjustments can address some biases, they may not fully eliminate the impact of underrepresented groups.

- Limited Coverage of Offline Voters: Although mail invitations and community outreach are included, there is still a risk of under-representation among individuals without internet access or those less engaged in civic matters.
- Sample Generalizability: Findings may not be fully generalizable to populations outside the sampled geographic or demographic groups.
- Static Data: The survey provides a snapshot in time, limiting the ability to capture dynamic changes in voter attitudes or behaviors over multiple election cycles.

A.1.8 Idealized survey questions

Thank you for your interest in our Toronto polls participate Survey. Your participation is 100% voluntary and you can withdraw at any time, for any reason, with no questions asked. This survey collects information about voters' views and reason for attending Toronto polls voting . The data collected will not be shared with any external parties and will strictly be used for research purposes only. This survey is completely anonymous and your data will be protected. Any published material regarding the results drawn from this survey will not be traceable back to you. The goal of this survey is to conduct research about the Toronto polls. Please answer as accurately as possible.

Contact Information

If you have any questions or concerns about this survey or its methodology, please reach out via email to the following individual:

Ruiying Li: ruiying.li@.mail.utoronto.ca

Correspondence will not be shared with any external parties.

Screening and consent

By checking the box below, I consent to this survey collecting personal information and information about my personal views, my preferred polls attended in Toronto, for research purposes only. - I consent

Demographic Information

1. What is your age range? 18-25 26-35 36-45 46-55 56 and above
2. What is your highest level of education? High school or less Associate's/Bachelor's degree Master's degree Doctorate
3. What is your annual total household income? Less than 20,000/20,000 - 39,999/40,000 - 59,999/ 60,000 - 79,999/80,000 - 99,999/100,000 or more

Community Participation

4. Have you participated in a local election in the past?Yes/No

5. Have you ever participated in community activities (e.g., public discussions or meetings)? Yes/No

Perception Disorders

6. Have you ever encountered difficulties during the voting process, such as obtaining channels or time constraints? Yes/No

7.If so, what are the main challenges?

Awareness and Attitudes Toward Application Types

8. Have you heard of the following types of applications? (Multiple choice) Boulevard Cafe/Business Improvement Area...
9. Have you ever submitted any of the following types of applications? (Multiple choice) Boulevard Cafe/Business Improvement Area/Commercial Boulevard Parking...
10. How transparent do you find the approval process for applications? (Scale: 1-5, 1 = Very non-transparent, 5 = Very transparent)
11. Do you believe the applications types affect your willingness to participate? (Single choice) Yes/No

Voter Participation Intent

12. Do you believe that community policies impact your willingness to vote? (Single choice) Yes/No

Suggestions

13. What changes do you think can encourage more people to participate in voting? [answer]
14. Thank you for your response! Would you be willing to participate in the follow-up investigation? Yes/No

Final section

15.Thank you for participating in this survey! If you have any questions, please contact [Survey Contact Name] (phone or email).

Our idealized survey can be found at this link

A.2 Appendix B: Pollester Methodology

Our polling methodology follows a structured and standardized approach that is commonly used in survey-based research and includes several key elements

A.2.1 Sampling Approach

The sampling frame is based on geographic location (Toronto), eligibility (registered voters or residents within certain boundaries), and voter history or behavior. This allows for the selection of a representative sample, which helps ensure that the survey results are reflective of the broader population.

A.2.2 Survey design

The survey is designed with a clear set of questions that capture essential information about voter participation. These questions are structured in a way that minimizes bias and allows respondents to provide accurate reflections of their views and behaviors. Response options are carefully crafted to cover the full spectrum of potential answers, ensuring that each individual's response can be accurately categorized. Data is collected through an online survey tool, ensuring that the process is both efficient and scalable. Online surveys also reduce the potential for interviewer bias and provide respondents with the flexibility to answer at their convenience, increasing the likelihood of more accurate responses.

A.2.3 Weighting and Data Adjustments

Given that certain groups of people may be underrepresented or overrepresented in the sample, weighting is applied to adjust for demographic imbalances. This helps to improve the accuracy of the survey results and ensures that the final dataset more accurately represents the general population.

A.2.4 Strengths and limitations

The use of geographic and eligibility criteria ensures that the sample is representative of the broader voter population. By targeting specific areas within Toronto and focusing on eligible voters, the survey captures data from individuals who are most likely to participate in polls. The structured nature of the survey ensures that responses are consistent across respondents. This consistency helps to minimize bias and improves the reliability of the results. Moreover, the survey questions are designed to capture key factors related to voter participation, making the data highly relevant to the research question. The decision to use an online survey platform provides scalability and flexibility. Online surveys also allow for easy data collection and processing, which can save time and reduce potential errors compared to other methods like phone interviews or face-to-face surveys.

The limitations also very obvious, when efforts are made to ensure a representative sample, online surveys may not capture the opinions of individuals who lack internet access or who are less likely to participate in digital surveys. This could introduce a bias towards certain demographic groups, such as younger individuals or those with higher socioeconomic status, who are more likely to be online. Despite efforts to increase response rates, some individuals may choose not to participate in the survey, potentially leading to non-response bias. This is particularly concerning when certain demographic groups are underrepresented in the sample. Since the survey is voluntary, those who choose to participate may have stronger opinions or different characteristics compared to those who do not. This self-selection bias could affect the generalizability of the results to the broader population. If the survey is structured to minimize bias, the wording of certain questions or response options could still influence how respondents answer. For example, leading questions or questions with limited response options may not capture the full range of respondents' views.

A.2.5 Literature

The methodology used in this study is based on widely accepted practices within the field of survey research. The sampling approach and survey design align with best practices recommended in the literature, such as those survey methodology outlined by Groves et al.(Groves et al. 2009) and Fowler (2014).(Groves et al. 2014) These authors emphasize the importance of careful survey design, representative sampling, and data weighting to ensure that survey results are valid and reliable.

However, while the methodology used in this study is sound, it is important to recognize that no methodology is perfect. Other studies have pointed out the limitations of online surveys and non-random sampling, which can lead to biases that affect the accuracy of results (Dillman, Smyth, and Christian 2014). Future research could explore alternative data collection methods, such as mixed-mode surveys or face-to-face interviews, to address these concerns.

A.3 Appendix C: Data and model details

A.3.1 Data detail

The key variables from dataset are used in the paper are:

ballot cast: Number of ballots returned

ballost distributed: Number of ballots distributed

potential voters: Number of people residing within poll boundary range

final voter count: Number of total voters on the final poll list

pass rate: Number of returned ballots needed for a positive poll result

response rate: the ratio of ballots_cast to ballots_distributed

application for: Type of application

Table 2: Cleaned Data of Polls Variables Part 1

application_for	potential_voters	ballots_distributed	ballots_cast
Front Yard Parking	41	34	18
Front Yard Parking	40	36	30
Front Yard Parking	135	97	43
Boulevard Cafe	120	106	30
Appeal Front Yard Parking	144	109	35
Front Yard Parking	100	80	48
Front Yard Parking	93	73	36
Traffic Calming	48	43	24
Traffic Calming	93	78	55
Traffic Calming	106	90	54

Table 3: Cleaned Data of Polls Variables Part 2

final_voter_count	pass_rate	response_rate
34	10	52.94
36	16	83.33
97	23	44.33
106	16	28.30
109	19	32.11
80	25	60.00
73	19	49.32
43	15	55.81
78	33	70.51
90	33	60.00

A.3.2 Model details

the results of the regression analysis for voter participation are presented. The goal of this analysis was to examine how various factors, such as the number of potential voters, final voter count, application types, and pass rate, influence the response rate in the context of voting participation. This paper includes summary statistics, tables, and visualizations to present and explain the findings clearly. The following table (Table 4) shows the summary statistics of the key variables used in the analysis. For instance, the variable response_rate has a mean of 49.9% with a standard deviation of 14.4%. The distribution of potential_voters is skewed due to a few extreme observations, with a maximum value of 2529

The model summary is shown in Table 5 to quantify the impact of potential voters, final voter count, and different application types on voter participation rates, we use a Bayesian linear regression model.

Table 4: summary of model varaibles

x

	Unique	Missing Pct.	Mean	SD	Min	Median	Max
potential_voters	246	0	117.0	134.0	2.0	95.0	2529.0
ballots_distributed	210	0	88.1	59.5	2.0	78.0	645.0
ballots_cast	126	0	42.0	28.7	2.0	37.0	309.0
final_voter_count	210	0	88.1	59.5	2.0	78.0	645.0
pass_rate	84	0	23.0	16.5	2.0	20.0	186.0
response_rate	762	0	49.9	14.4	21.4	49.1	100.0

Table 5: Regression estimates for Voter Participation Analysis

	polls
(Intercept)	52.361
potential_voters	0.010
final_voter_count	−0.380
application_forBoulevard Cafe	−4.533
application_forBusiness Improvement Area	−11.022
application_forCommercial Boulevard Parking	−8.467
application_forFront Yard Parking	1.421
application_forPermit Parking	0.751
application_forPermit Parking Removal	0.442
application_forProposed Business Improvement Area	−5.511
application_forTraffic Calming	−2.940
application_forTraffic Calming – Island	−3.371
application_forTraffic Calming Safety Zone	6.897
pass_rate	1.298
Num.Obs.	1069
R2	0.572
R2 Adj.	0.538
Log.Lik.	−3915.450
ELPD	−3945.1
ELPD s.e.	43.4
LOOIC	7890.1
LOOIC s.e.	86.7
WAIC	7887.1
RMSE	9.41

A.3.3 Posterior predictive check

Figure 12 and Figure 13 captures the general trend in the data and provides predictions consistent with the observed values. Figure 12 compares the posterior predictive distribution to the observed data (y_obs). The blue curves represent the posterior predictive distribution, with multiple lines showing the uncertainty in predictions, and Figure 13 compares the prior and posterior distributions of the model parameters. The dots represent posterior means, and the horizontal lines indicate the spread of the distributions for parameters like sigma and application types.

For Figure 12, it aligns well with the observed data, indicating that the model can replicate the overall patterns in voter participation. The model has reasonable predictive power under the current data.

For Figure 13, the clear shift from prior to posterior distributions suggests that the data has had a significant impact on updating the parameters. This indicates that the priors were reasonable and not overly restrictive.

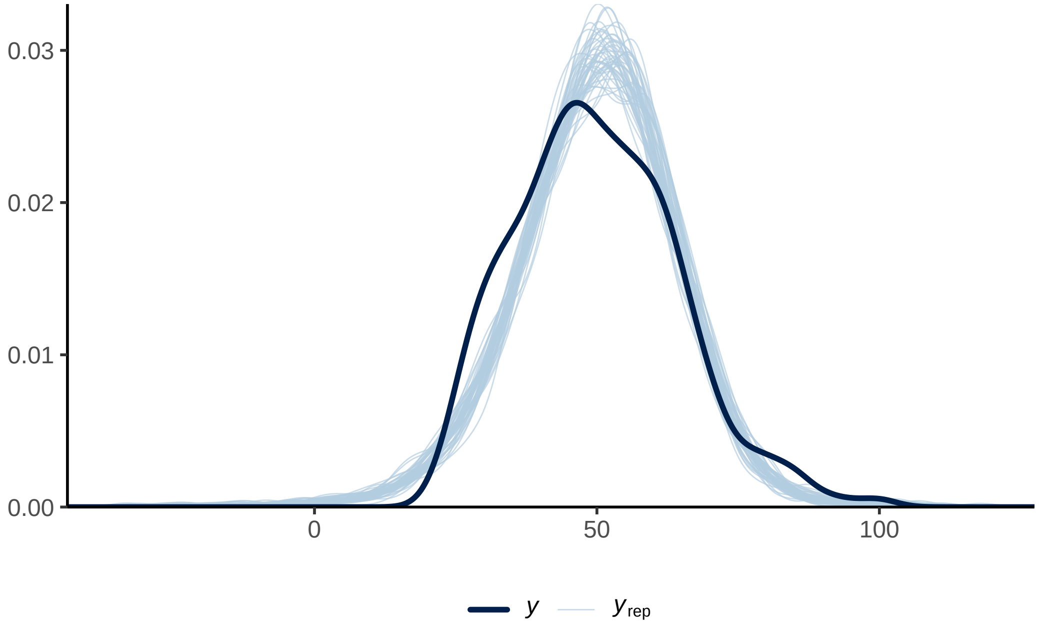


Figure 12: Examining how the Bayesian linear regression model fits

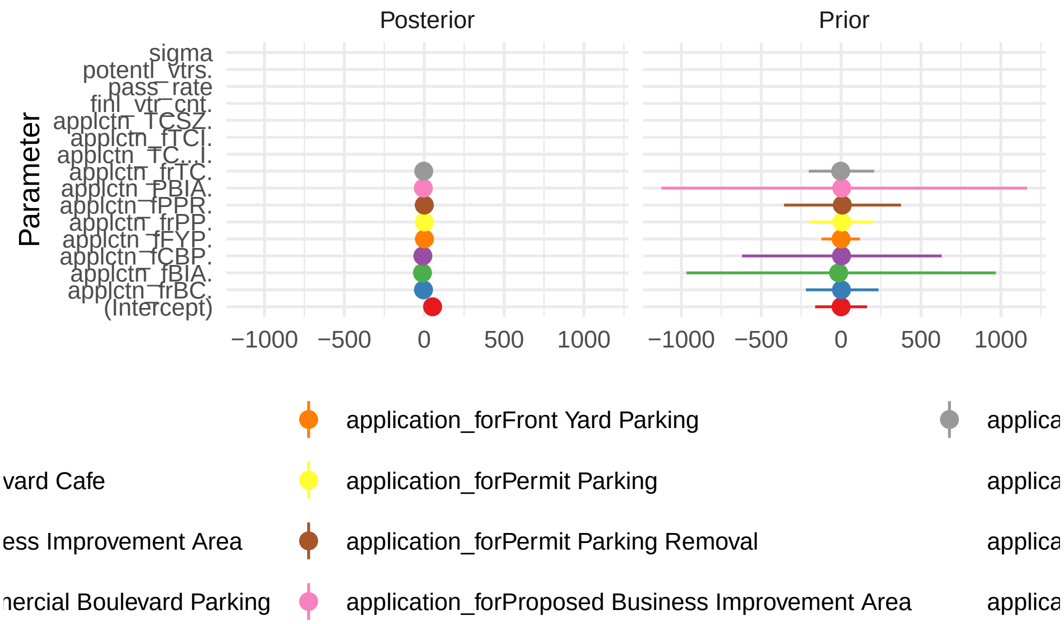


Figure 13: Examining how the Bayesian linear regression model is affected by data

A.3.4 95% Credible Intervals Plot

Figure 14 identifies the key information of voter participation and quantifies the uncertainty of the effects. It displays the 95% Bayesian credible intervals for all model parameters, including quantitative variables (e.g., potential_voters and final_voter_count) and categorical variables (e.g., levels of application types). Horizontal lines indicate the credible interval, and dots represent the posterior mean estimates.Parameters like potential_voters and final_voter_count have narrow credible intervals that do not include zero, suggesting they significantly affect voter participation. the categorical variables (e.g., certain application types) have wide credible intervals that include zero, indicating they may not significantly impact voter participation. Narrow intervals indicate precise estimates, while wider intervals suggest greater uncertainty, possibly due to limited data or weak relationships with the response variable.

A.3.5 Model Diagnostics

Figure 15a is a trace plot, and show convergence ensures that the posterior estimates are stable and trustworthy. It displays the Markov Chain Monte Carlo (MCMC) sampling trajectories for all parameters. Each plot corresponds to a parameter, showing the sampling iterations for multiple chains. Horizontal and stable traces with overlapping chains indicate that the MCMC algorithm has converged, suggesting reliable posterior estimates. The absence of irregularities or non-converging chains confirms that the sampling process is strong. Figure 15b is a Rhat (Gelman-Rubin) diagnostic for assessing convergence in Bayesian sampling.The points are all

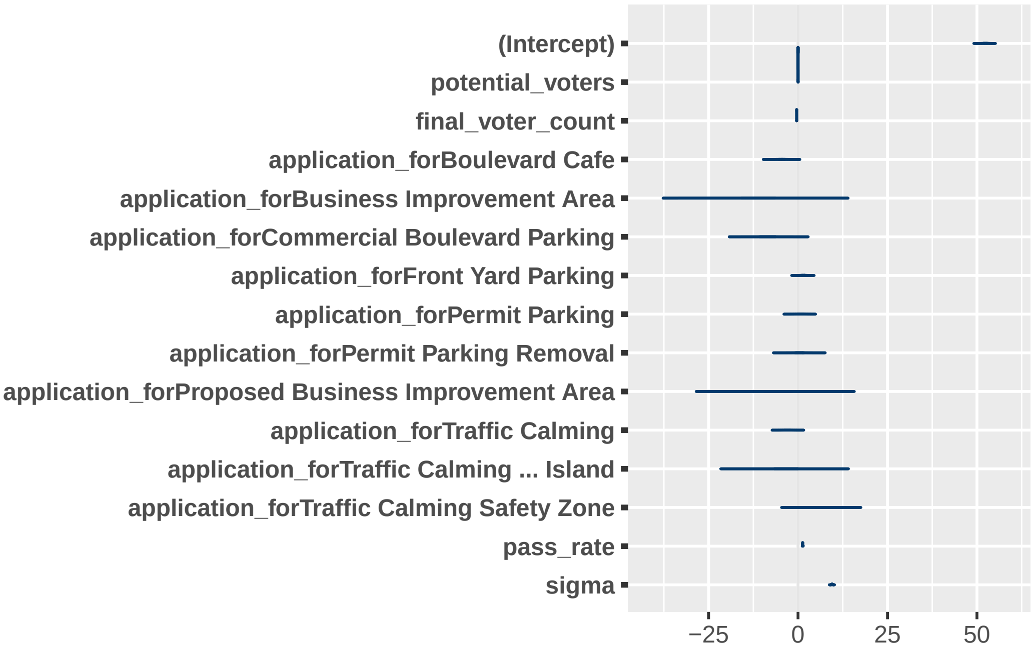
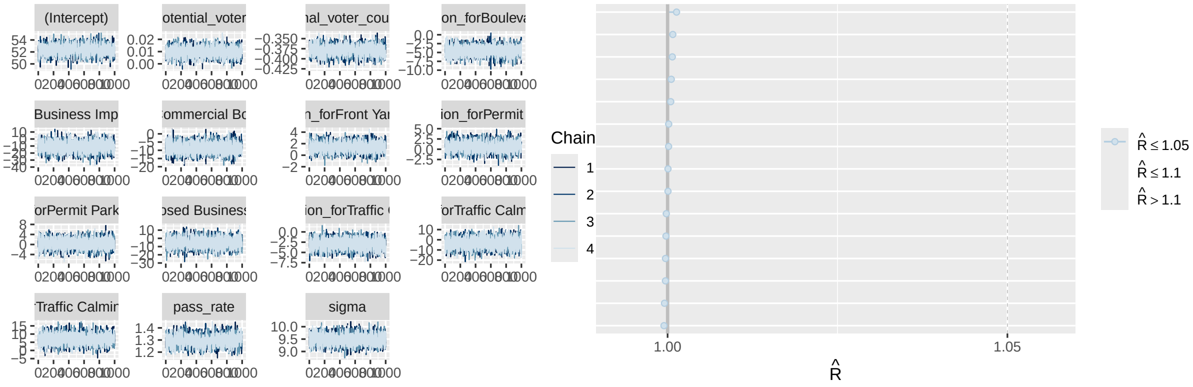


Figure 14: 95% credibility intervals for the polls model

clustered near 1.0, indicating that the Rhat values for all parameters are very close to 1. This suggests that the MCMC sampling chains of the model have converged well and that the parameter sampling process is stable.



(a) Trace plot

(b) Rhat plot

Figure 15: Checking the convergence of the MCMC algorithm by trace plot and rhat plot on linear model

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Dillman, D. A., J. D. Smyth, and L. M. Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method, 4th Edition*. Wiley. https://www.wiley.com/en-us/Internet-Phone-Mail-and-Mixed-Mode-Surveys.

Gabry, Jonah, and Stan Development Team. 2023. "Bayesplot: Plotting for Bayesian Models." https://mc-stan.org/bayesplot/.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

Groves, R. M., F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Progress in Physical Organic Chemistry, Volume 1*. Wiley-Interscience. https://www.wiley.com/en-us/Progress+in+Physical+Organic+Chemistry%2C+Volume+1-p-9780470171806.

———. 2014. *Survey Research Methods (5th Ed.)*. Sage Publications. https://books.google.ca/books/about/Survey_Research_Methods.html?id=WM11AwAAQBAJ/&redir_esc=y.

Open data Toronto. 2015. *Polls Conducted by the City*. City Clerk's Office. https://open.toronto.ca/dataset/polls-conducted-by-the-city/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Wickham, Hadley. 2023a. "Dplyr: A Grammar of Data Manipulation." https://cran.r-project.

org/web/packages/dplyr/index.html.

———. 2023b. "Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics." https://cran.r-project.org/web/packages/ggplot2/index.html.

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.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r.* https://cran.r-project.org/package=knitr.