Provincial Voter Participation by Age Group: A Study on Younger and Older Voters in British Columbia

Aimen Arif, Jacob Lomax, Gurnoor Singh

University of the Fraser Valley

COMP 381

Professor. Carl Jenzen

June 19, 2023

Git Repository URL: https://sc-gitlab.ufv.ca/group-3/project-proposal1

Summary

This executive summary provides an overview of the proposal titled “Provincial Voter Participation by Age Group: A Study on Younger and Older Voters in British Columbia”. The proposal analyzed provincial voter participation by age group in British Columbia using clustering, classification, and regression techniques to identify patterns and groups, providing insights into factors influencing voter participation across different age groups. Expected results include understanding voting trends, changes in registered voter numbers, and the probability of voting based on age, contributing to strategies for increasing voter engagement, all of these are expected to be answered after running and analyzing our dataset.

Expected Deviations

Throughout the proposal evaluation process, a few deviations from the original proposal were identified. These are the following deviations: 1) In addition to the outlines statistical tests and techniques, we could use analytical methods like data visualization techniques to present our findings more effectively. 2) Our dataset might have missing values, data quality issues or sampling biases, so we may need a better approach to solve these limitations. 3) We could add data collection methods other then surveys to gather a more comprehensive understanding of voting behavior. 4) Choose the most fitting algorithm out of several algorithms for an effective analysis.

Research Questions

1. Do younger people in British Columbia vote more or less than older people in BC?
    a. According to Elections.bc.ca (2009), their study on the voter turnout among neighborhoods found out that the neighborhoods that were most likely to vote had a larger number of older individuals and a greater proportion of university educated citizens. This implies that the presence of older individuals suggests a higher level of political interest, engagement and habitual voting behavior. It is very clear that the university educated citizens are adults and have more knowledge on the system thus the voter turnout is higher for these individuals. Elections.bc.ca (2009) also mentioned that “The exception to the “younger people don’t vote” trend was the 18 or 19 year old cohort. Three out of every five registered voters in this age group cast a ballot…most if these individuals may have had to consciously make an effort to register, unlike older individuals who would already be on the voters list”. This implies that the younger generation has had to exert additional effort to cast their votes, which could contribute to the lower voter turnout among this demographic across British Columbia. After conducting this research, we decided to follow a different approach and find the turnout rates for all age groups using data manipulation and clustering using k-means. While our original plan was to identify just one age group with the highest voting turnout, examining turnout rates for each group provides insights into the

varying levels of engagement and participation among different generations.

To answer this research question, we ran the dataset in JupyterHub using the code provided to analyze the voter turnout among different age groups in British Columbia. After preprocessing the data and aggregating the total number of registered voters and votes cast for each age group, we calculated the voter turnout percentage for each age group by dividing the votes cast by the registered voters and multiplying by 100. Utilizing k-means clustering, we grouped the age groups into two clusters based on their turnout percentages. Moreover, we determined the turnout status of each age group by comparing their turnout with the overall average. The analysis revealed interesting insights into the voting behavior of different age groups. The scatter plot visualized the turnout percentages, with each point representing an age group and color-coded and shaped according to the cluster label.

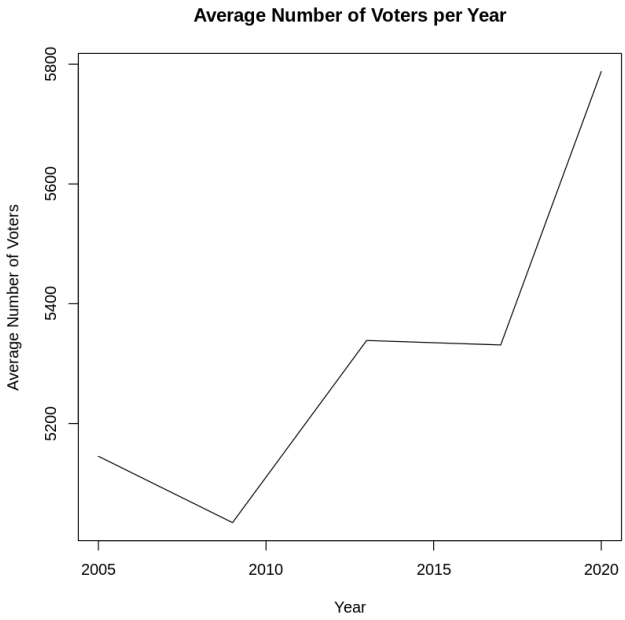
The number of people registered to vote in British Columbia increased consistently from 2005 to 2020. In 2005, there were 2,845,284 registered voters, and by 2020, the count had grown to 3,524,812. The most significant increases happened between 2009 and 2013, as well as between 2017 and 2020, with growth rates of 5.28% and 8.57%. Overall,

this shows that more people in British Columbia were registering to vote during this time.

2. How have the number of registered voters in British Columbia changed from 2005 to 2020?
    a. After conducting our research on the number of registered voters in British Columbia from 2005 to 2020, several key findings emerged. We used the provided code to analyze a dataset and understand how the number of registered voters in British Columbia changed from 2005 to 2020. Our original plan was just to find out the turnout status for each age group but realized that wasn't enough to thoroughly understand the data. Therefore, our approach involved analyzing a dataset to understand the voter turnout patterns across different age groups. The code performed various operations like converting data types, filtering the data for the desired years, calculating the total number of registered voters per year, determining the percentage change between consecutive years, and organizing the results. By running the code, we obtained a clear summary that shows the year, the number of registered voters, and the percentage change for each year during the specified period.

Firstly, there had been an overall upward trend in the number of registered voters over the analyzed period. The data showed a consistent

increase in voter registration, indicating a growing participation in the electoral process. The following graph shows this by plotting the average number of registered voters across all areas and age groups against the year.



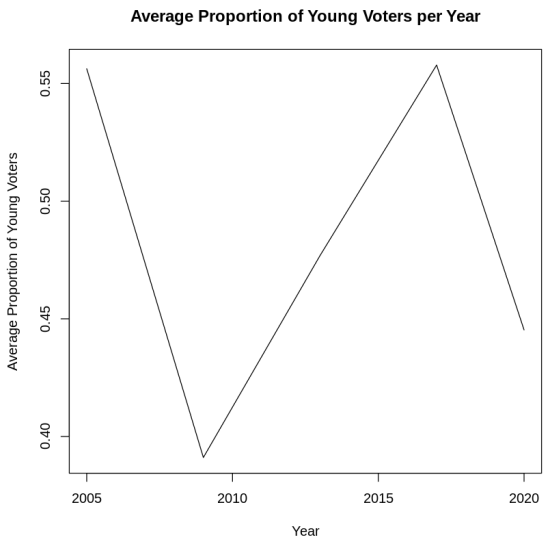
Secondly, while the overall trend was positive, there were notable variations in the growth rate throughout the years. Some years experienced significant increases in voter registration, indicating possible influences such as political events, policy changes etc. These findings highlight the importance of ongoing efforts to encourage civic participation and ensure accessible and inclusive voter registration processes.

The output gave a summary of voter turnout data based on different age groups. Our expected results showed the percentage of voter turnout for each age group, assigns numeric labels to the groups, and categorizes their turnout status as "Low" or "High", rather then just

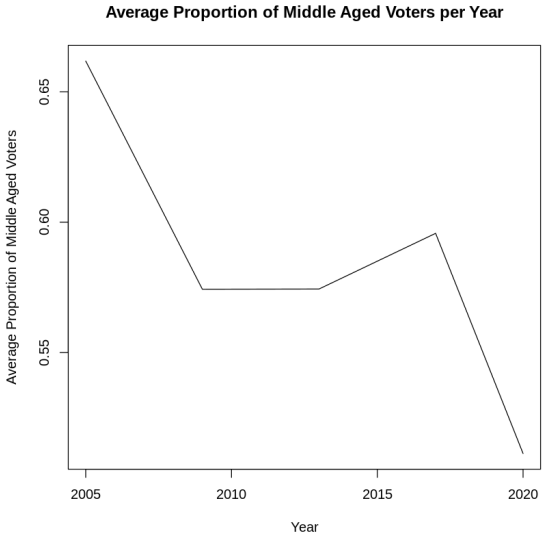
showing the turnout status. For example, the voter turnout for ages 18-24 were 48.57429% whereas for ages 65-74 were 73.45227%. It is important to acknowledge that this analysis solely focuses on the number of registered voters and does not account for potential changes in the eligible voting population or voter turnout rates.

3. Did a bigger or smaller proportion of young people vote in 2020 compared to 2005?
    a. For the purpose of this question, we will consider “young people” to be the smallest age group in our dataset, which is 18-24. In order to answer this question, the first thing we had to do was filter the data to only include the age group 18-24. This was done using the subset function to create a subset of the original dataset where only the rows with the age group 18-24 are included. From there we could create a new column called proportion, that is the voter participation divided by the total registered voters. This gives us data on the proportion of people who voted for every row in the dataset and allows us to learn a lot about voter trends.

When we plot the proportion of young voters against the year we get the following graph, which shows



the trend of young voter proportions since 2005. We can see that a lower proportion of young people voted in 2020 compared to 2005, although the graph does vary quite a bit between these two dates. We can compare this graph to the same graph but with the age group of 45-54, which shows that they also have a lower proportion of voters in 2020, although the trend is much more clear and less bumpy.



We can use regression to better understand these trends. If we create a linear regression model of the proportion of young voters, we find that the correlation between the voting proportion and the year is -0.0009901, which is a very slightly downward slope. This is because the downward voting trend is so slight and it goes up and down between 2005 and 2020. Building a linear regression model for the 45-54 age group produces a correlation of -0.007131. This is a more clearly downward trend, which is evident when looking at the above graphs.

This data shows that overall voting turnout has decreased for young people since 2005, and the same is true for other age groups also. Although there may be more people registered to vote, this is not translating to a larger proportion of those people actually voting.

4. Can we predict the probability that an individual will vote, given only their age?
    a. Yes, predicting the probability that an individual will vote based solely on their age can be approached using various classification algorithms. These algorithms learn from historical data, capturing the relationship between age and voting behavior, and then utilize this learned information to make predictions.

        One commonly used algorithm for classification tasks is logistic regression. Logistic regression models the probability of an event occurring, in this case, an individual voting, based on the input variables, such as age. By training a logistic regression model on a dataset that includes age and corresponding voting outcomes, the model can estimate the probability of an individual voting given their age.

        We predicted it by, using the logistic regression model described in the provided code.

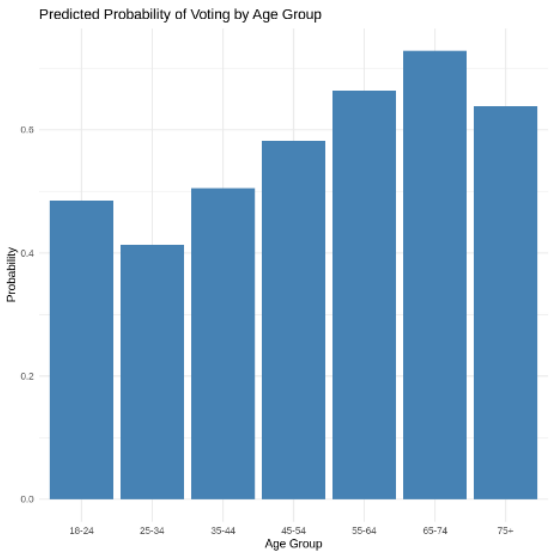
        The logistic regression model is built using the formula "PARTICIPATION / REGISTERED_VOTERS - AGE GROUP". This formula suggests that the

probability of an individual voting is being modeled as a function of the ratio of participation to registered voters, along with the age group they belong to.

The logistic regression model is trained using the data read from the CSV file "voterdata.csv". The "PARTICIPATION" and "REGISTERED_VOTERS" columns in the data are converted to numeric format, indicating that these variables represent numerical values.

After training the model, the unique age groups present in the data are extracted. These age groups are stored in the "age groups" variable.

A new dataframe named "new_data" is created, which consists of the "AGE_GROUP" column derived from the unique age groups. This allowed us to input specific age groups into the model for prediction.



The logistic regression model is then used to predict the probability of voting for each age group. The "predict" function is applied to the model, using the "new_data" dataframe as the input. The predicted probabilities are calculated using the type "*response".

After doing our analysis and obtaining results, we found out that for example age groups like 18-24 had a probability of 0.48%, 25-34 had a probability of 0.41% and so on. One age group of 65-74 had the highest probability of approximately 0.73% which indicates that the older age group is more likely to vote in British Columbia. There can be a lot of factors contributing to this like pension concerns, old homes etc. Therefore, by providing an individual's age and using the logistic regression model, we can estimate the probability that the individual will vote.

Data Description

We planned to use the "provincial_voter_participation_by_age_group" dataset as the primary source of data for this project, and can be accessed from Provincial Voter Participation by Age Group. It provides detailed information about voter participation among various age groups in BC. The dataset's structure, column names, and general information can be found in the Provincial Voter Participation by Age Group Data Dictionary document (Data.gov BC, 2018).

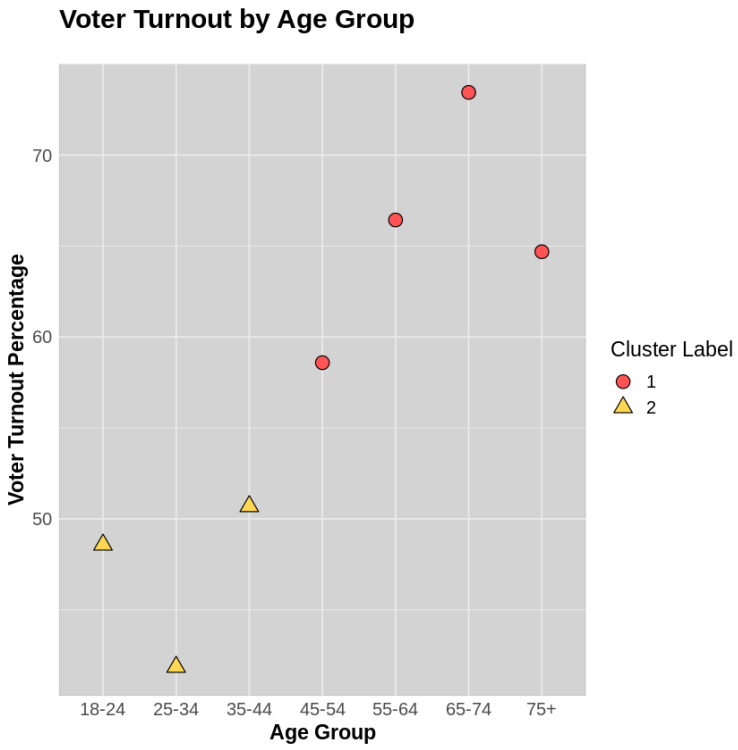
Application of Machine Learning

I. **Clustering:** Clustering is a unsupervised technique used to identify potential patterns and group similar data points together. In our first proposal, it was mentioned that clustering can be used for data examination and making judgments.

In this project, clustering will help analyze voter turnout data by age group. First our data will be converted into a numeric format and handle any missing data if needed. Then the voter turnout will be represented in percentage to show the proportion of voters. Specifically, the k-means clustering algorithm will be utilized to divide the data of different age groups into distinct clusters. Towards Data Science (2018) states that “To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids”. This means that K-means algorithm begins with random cluster centers and iteratively adjusts them to optimize the clustering results. K-means works by grouping data points into clusters based on their similarity so it will be easy to identify if younger people in British Columbia vote more or less than older people in BC.

Once the clustering is performed, we can compare the voting patterns of younger and older age groups. If the clustering analysis shows that younger age groups tend to have lower voter turnout percentages and form a separate cluster from older age groups, it suggests that younger people in BC vote less than older people.

The scatter plot provides information on the distribution of voter turnout percentages across different age groups. It helps identify patterns and groupings among the data points by differentiating them based on cluster labels By examining the plot, we can analyze the relationship between age group and voter turnout where the older age group has a high voter turnout compared to young people, and it also provide insights into trends and correlations between age and voter turnout.



II. **Classification:** classification refers to the application of supervised machine learning techniques to predict the probability that an individual will vote based solely on their age. The goal is to build a classification model that can categorize individuals into "likely to vote" or "unlikely to vote" groups, given their age as the input feature. This involves training the model on labeled data, where the target variable is the voting behavior

(whether an individual voted or not), and the input feature is the age of the individual.

The classification process involves several steps. First, we need labeled data, which consists of historical records that indicate whether an individual voted or not, along with their corresponding age. This labeled data is used to train the classification model. The model learns from the patterns and relationships in the training data, enabling it to make predictions on new, unseen data.

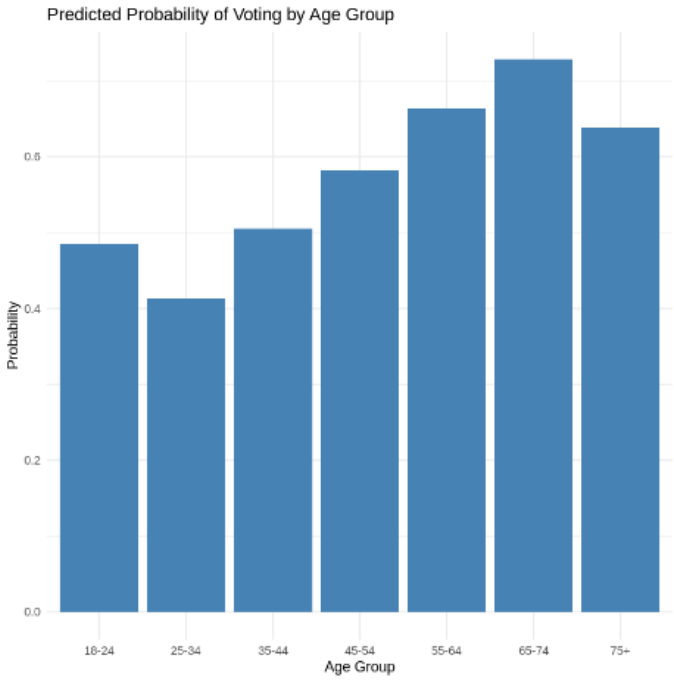
During the training phase, the classification algorithm examines the relationship between the input feature (age) and the target variable (voting behavior) to identify patterns and create decision boundaries. The algorithm uses various mathematical techniques to optimize its performance and make accurate predictions.

Once the model is trained, it can be applied to new data, where the age of an individual is known, but their voting behavior is unknown. The model will analyze the age of the individual and assign a probability or confidence score indicating the likelihood of that individual voting. For example, the model might predict that a 30-year-old individual has a 75% probability of voting based on the patterns it learned during training.

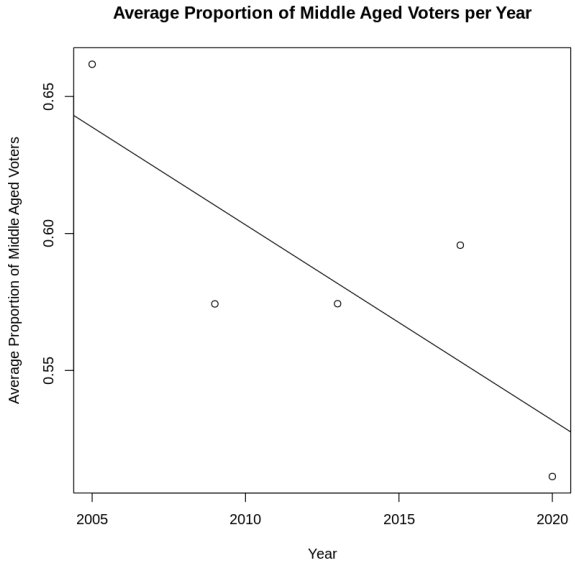
The accuracy of the classification model can be evaluated by comparing its predictions with the actual voting behavior of individuals in

a separate test dataset. Metrics such as accuracy, precision, recall, and F1-score can be used to assess the model's performance and determine its effectiveness in predicting voting behavior based on age.

Predicting the probability of an individual voting based solely on their age can be accomplished using classification algorithms such as logistic regression. These algorithms learn from historical data to establish the relationship between age and voting behavior. By training a logistic regression model on a dataset that includes age and voting outcomes, the model can estimate the probability of an individual voting given their age. The provided code demonstrates the implementation of logistic regression for this purpose, where the model is trained using data from "voterdata.csv" and predicts voting probabilities for different age groups. The formula used in the logistic regression model incorporates the ratio of participation to registered voters along with age groups. This approach provided a means to estimate the probability of an individual voting based on their age.

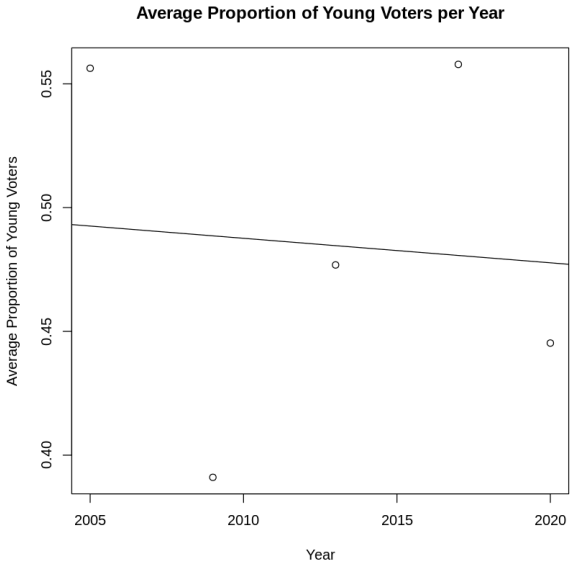


III. **Regression:** Regression is a supervised machine learning technique that allows us to relate a dependent variable to the independent variables. This allows us to predict continuous output data based on the input data. In this project, regression was used to better understand the relationships between variables in our dataset. It can allow us to model the correlation between two variables and graph this correlation in order to easily see trends in our data. Linear regression is an algorithm that “shows how closely the two values are linked but not if one variable caused the other” (Master’s in Data Science, 2022). This is because, while regression is very good at showing the correlation between variables, it can not reveal the causation.



When studying the trend of the proportion of people who vote in different age groups, we used linear regression to better understand how this proportion correlates to time. When we created a linear regression model of the proportion of middle aged voters per year, we found the correlation to be -0.007131 and the intercept to be 14.936451. If we use these numbers to create a line instead of just connecting the points, we

get the following graph. This produces a straight line that best fits the data, and shows the trend in voter proportions over the years and shows how it is trending down. If we do the same thing with the 18-24 group, we get this graph. It shows a much less steep line, and is very far off of some of the points. This is because the values of those points are so bumpy, therefore it is very hard to create a straight line that best fits the data. Even so, the line does still show a downward trend in voter proportions over time, although it is much less clear than the previous example.



Linear regression has allowed us to better understand trends and correlations in our dataset and it could be applied to many more things that we did not explore in this project. Although it can not explain the causation behind these trends, it can be useful in determining if the trends exist or not.

Concerns

During the project, we encountered and addressed some ethical concerns. One concern was data and privacy. It is important to handle and processing voter data to ensure privacy and data protection to maintain confidentiality and prevent misuse, so we followed the privacy protocols and avoided any manipulation or deletion of data.

To maintain fairness and unbiased analysis, methodologies were followed using the provided data, and reliable data sources were used, appropriate statistical techniques were employed, and it was peer reviewed.

Other ethical concern is the representation and potential bias in the sample used for analysis. It is important to ensure that the sample accurately represents the diverse age groups in British Columbia, avoiding any biased selection that could skew the results or exclude underrepresented groups.

One important ethical concern is the potential impact on participants and stakeholders. Analyzing voter turnout rates and demographic data could potentially reveal sensitive information about individuals and communities. So we handled this information with care and ensured that it is used responsibly and ethically.

Conclusion

In conclusion, the analysis of voter turnout in British Columbia revealed some important findings. The study showed that neighborhoods with more older individuals and university-educated citizens had higher voter turnout, indicating their higher political

interest and engagement. However, 18 or 19-year-olds stood out as an exception, with a higher proportion of them making the effort to register and vote because of all the registration process they may have to go through. By using data manipulation and clustering techniques, the analysis provided insights into the turnout patterns across different age groups. The analysis also found a consistent increase in the number of registered voters in British Columbia from 2005 to 2020. Significant growth rates were observed between 2009-2013 and 2017-2020, indicating a rising interest in voter registration. Further research can explore why older individuals and university-educated citizens have higher voter turnout and how to engage younger and less educated demographics. Investigating the impact of political campaigns, policy issues and digital technologies/social media on voter behavior is important. We can also conduct comparative studies with other regions or countries which can provide a broader understanding of political participation.

For our research on the number of registered voters in British Columbia from 2005 to 2020 revealed an overall increasing trend in voter registration. However, there were variations in the growth rate across different years, indicating possible influences such as political events or policy changes. These findings emphasize the importance of promoting civic participation and ensuring accessible voter registration processes. Our analysis provided insights into voter turnout among different age groups. The data showed varying turnout percentages, such as 48.57% for ages 18-24 and 73.45% for ages 65-74. We can consider exploring how social media platforms influence voter

attitude and behavior. By understanding the impact of social media on voter registration and turnout, we can determine its potential as a tool to increase political participation. This research would provide valuable insights into the role of social media in shaping the way people engage with the electoral process.

We found that the number of people actually voting, relative to the amount of registered voters, has been generally decreasing since 2005. This is true among all age groups, although this trend is much more clear for middle aged people in particular. Among young people aged 18-24, the data is much more varied by year. There is still a smaller proportion of them voting, but it has gone up and down between 2005 and 2020. The cause behind these trends could be an entire project by itself and our data will not reveal any of the causes, but learning the reasons behind these trends could be very important for future elections.

Logistic regression is a widely used classification algorithm that models the probability of an event, such as an individual voting, based on input variables like age. By training a logistic regression model on a dataset containing age and corresponding voting outcomes, the model can estimate the probability of an individual voting given their age. After training the model on data from "voterdata.csv," specific age groups are extracted and used to predict the probability of voting using the logistic regression model. By inputting an individual's age into the model, the probability that they will vote can be estimated. Using the logistic regression model described in the provided code employs the formula "PARTICIPATION / REGISTERED_VOTERS - AGE GROUP" which

considers the ratio of participation to registered voters along with age groups to model the probability of voting, we found out that the 25-34 age groups in British Columbia are less likely to vote with a probability of only 0.41% whereas the 65-74 has a voting probability of approximately 0.73% which differs from the other age group immensely.

References

1. Government of British Columbia. (2018). Provincial Voter Participation by Age Group [Data file]. Data.gov.bc.ca. Retrieved from https://catalogue.data.gov.bc.ca/dataset/6d9db663-8c30-43ec-922b-d541d22e634f/resource/646530d4-078c-4815-8452-c75639962bb4/download/provincial_voter_participation_by_age_group.csv
2. Government of British Columbia. (2018). Provincial Voter Participation by Age Group Data Dictionary (Publication No. a31fec32-e9e6-45ca-bab0-b5ddc1236bcf) [PDF file]. Data.gov.bc.ca. Retrieved from https://catalogue.data.gov.bc.ca/dataset/6d9db663-8c30-43ec-922b-d541d22e634f/resource/a31fec32-e9e6-45ca-bab0-b5ddc1236bcf/download/provincial-voter-participation-by-age-group-data-dictionary.pdf
3. Government of British Columbia. (2018). Provincial Voter Participation by Age Group. Data.gov.bc.ca. Retrieved from https://catalogue.data.gov.bc.ca/dataset/provincial-voter-participation-by-age-group
4. Elections BC. (2009). Who heads to the polls? [PDF document]. Retrieved from https://elections.bc.ca/docs/stats/Who-heads-to-the-polls.pdf
5. Towards Data Science. (n.d.). Understanding K-means Clustering in Machine Learning. Towards Data Science. Retrieved June 15, 2023, from

https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

6. Master's in Data Science. (2022). Linear Regression. Retrieved June 16, 2023, from https://www.mastersindatascience.org/learning/machine-learning-algorithms/linear-regression/