Determinants of Canadian Politicians' Popularity on Wikipedia.[*]

Ethan Sansom

27 April 2022

Abstract

Voters internationally are increasingly using the internet as a primary tool to access information on politicians and political parties. Consequently, politcal organizers and researchers alike are using online platforms, such as Twitter, Facebook, and more recenlty Wikipedia, to analyze the sentiments of voters. This paper investigates the relationship between Canadian politicians' characteristics - including party affiliation, political position, and age - and politicians' Wikipedia page views in the months prior to the 2021 Federal Election. Wikipedia page views are found to be somewhat greater for politicians in more popular parties, and for Party Leaders, Ministers, and Prime Ministers in particular.

Keywords: Canada elections, election forecasting, Wikipedia, linear regression

Contents

[*] Code and data are available at: https://github.com/EthanSansom/wikipoliticians

1 Introduction

Online search engines are increasingly the avenue by which voters internationally access information on political candidates in their countries. A reported 68% of internet-using voters in Germany, France, Italy, Spain, Poland, the United Kingdom, and the United States, for instance, rated online searches as an important factor in their voting choices (Pradel 2021). Such internet usage provides political organizers, pundits, and researchers a wealth of relatively easy to access on the voting preferences and attitudes of the electorate. An increasingly large number of researchers have used name mentions of political parties or individual politicians on social media platforms such as Twitter and Facebook to measure the popularity of politicians and political parties (Smith and Gustafson 2017; Yasseri and Bright 2013). Such measures, however, are limited in their predictive abilities of voter behavior. Name mentions alone, for example, do not indicate a positive or negative attitude towards a politician or party (Yasseri and Bright 2013). Further, investigating social media platforms alone limits the sample of voters investigated to those who both participate in social media and use social media platforms as a venue for communicating political views (Yasseri and Bright 2013). Pressures to adhere to the political views of social circles, and potential negative outcomes at home or in the workplace of voicing divisive political views on a public platform, can additionally mask or misrepresent the intentions of voters on social media (Yasseri and Bright 2013). Authors Salem and Stephany (2021) argue that Wikipedia, the online encyclopedia, offers a better platform for accurately modelling voters' intentions. Specifically, the number of views of a given politician's or political party's Wikipedia page over a given time period can provide a strong measure of the number individuals searching for information about the politician or party (Yasseri and Bright 2013; Salem and Stephany 2021). Such information-seeking behavior has been shown to be reflective of the true public voting interest for certain candidates (Salem and Stephany 2021). Wikipedia page viewership, and other Wikipedia page metrics then, may offer a less ambiguous and more predictive measure of voting intent than that of other internet platforms. Thus, for researchers, pundits, pollsters, and political organizers alike, Wikipedia represents an important potential tool for predicting elections and describing the political attitude of the electorate.

In this paper, the Wikipedia pages of Canadian politicians are investigated to assess the viability of using Wikipedia for election forecasting and analyzing voter sentiment in Canada. Using monthly data on Wikipedia page views from January 1st 2021 to January 1st 2022, we compare the Wikipedia viewership of politicians from the five major Canadian Federal political parties, the Liberal Party, Conservative Party, New Democratic Party (NDP), Green Party, and the Bloc Québécois, to the results of the 2021 Canadian Federal Election. The page views of party leaders during this time, namely Justin Trudeau (Liberal), Erin O'Toole (Conservative), Jagmeet Singh (NDP), Annamie Paul (Green), and Yves-François Blanchet (Bloc Québécois), are similarly compared to the realized election outcomes. Wikipedia age views of both party members and leaders are shown to closely align with the results of the 2021 election. The parties with the first, second, and third most page views respectively, the Liberal, Conservative, and New Democratic Parties, likewise received the first, second, and third most votes in the federal election. To further investigate the correlation between party affiliation, political position, politician characteristics, and Wikipedia page views, we estimate a multiple linear regression model. In particular, we regress page views on politician indicators for party affiliation, political position, and politician birth year. The model results show that political positions, specifically Prime Minister, Minister, and Party Leader, are associated with a large increase in Wikipedia page views. Additionally, the model somewhat reaffirms the party rankings observed in overall page views. For instance affiliation with the Liberal and Conservative parties are associated with greater page views than the NDP politicians, while affiliation with the Bloc Québécois is associated with fewer page views. The linear models used in this paper, however, are shown to be substantially limited in terms of internal and external validity, warranting additional future analysis with different modelling techniques.

In the following Data Section 2 of this paper, the data collection and variable selection processes are discussed. The effects of politician birth year, party affiliation, and political position on Wikipedia page views are visualized. The Model Section 3 describes the multiple linear regression framework used in this analysis and proposes a causal justification for variables selected. In the Results Section 4, several linear models are compared and evaluated. Additionally, the results of the 2021 Federal election are compared with Wikipedia page viewership. Finally, in the Discussion Section 5, the results are analysed in the context of the 2021 federal election results and the political views of the Canadian electorate more broadly. The limitations of the sample and model specification are further discussed, including the challenges to the internal validity of the linear regression models used and the ability of Wikipedia page viewership to serve as a valid proxy for voting intention.

For this analysis, the R statistical software (R Core Team 2021) is used to import and clean Wikipedia page data, perform regression model estimation and data visualization, and compile the paper itself. Packages `tidyverse` (Wickham 2021b) and `dplyr` (Wickham, François, et al. 2021) are used to manipulate data, `ggplot2` (Wickham, Chang, et al. 2021), `gridExtra` (Auguie 2017), and `kableExtra` (Zhu 2021), are used for creating tables and plots, and `ggrepel` (Slowikowski 2021) is used to customize plot labels. A number of packages, `see` (Lüdecke, Patil, et al. 2021), `scales` (Wickham and Seidel 2020), `DiagrammeR` (Iannone 2022), `DiagrammeRsvg` (Iannone 2016), `rsvg` (Ooms 2022), `cowplot` (Wilke 2020), `ggcorrplot` (Kassambara 2019), and `gtsummary` (Sjoberg et al. 2021), provide additional aesthetic adjustments, summary tables, and visualizations. Packages `modelsummary` (Arel-Bundock 2022) and `performance` (Lüdecke, Ben-Shachar, et al. 2021) are used to display the results and assumptions of linear regression models. To access web APIs, the package `httr` (Wickham 2020) is used, and for scraping HTML data the package `rvest` (Wickham 2021a) is used. The package `here` (Müller 2020) is for file management. Lastly, packages `bookdown` (Xie 2021a), `knitr` (Xie 2021b), and `tinytex` (Xie 2021c) are utilized to format and compile the final paper.

2 Data

2.1 Collection

This paper analyses data from Wikipedia, which is a self-described free digital encyclopedia that hosts greater than six million English language entries and is currently the most popular reference website on the internet ("Wikipedia:about" 2022; Salem and Stephany 2021). To collect this data, the Wikidata database and associated API, which stores structured data on Wikipedia articles, was used to query the full set of biographies of Canadian politicians available on Wikipedia ("Wikidata:history Query Service" 2020). The set of articles was then filtered to include only Canadian politicians on Wikipedia affiliated with one of the five major federal political parties or their predecessors[1], resulting in a final sample of 7856 Wikipedia pages[2]. Table 1 displays a sample of the final dataset used for this analysis.

Table 1: Sample Wikipedia Data

Name	Party	Birth Year	Prime Minister	Party Leader	Minister	Page Views
Bud Cullen	Liberal	1927	0	0	1	782
Bud Olson	Liberal	1925	0	0	1	1266
Byron Ingemar Johnson	Liberal	1890	0	0	0	1990
Gordon Chown	CPC	1922	0	0	0	415
Louis Henry Davies	Liberal	1845	0	0	0	1210

The variables for politicians' party affiliation, birth year, and political position are available via the Wikidata API. Page views, which measure the number of distinct times a human Wikipedia user viewed each page during the year 2021, were collected separately using the Wikimedia REST API ("Wikimedia REST API" 2022). The monthly page views of each article in 2021 are recorded as well, although not displayed in Table 1 for brevity. Each of Prime Minister, Minister, and Party Leader are equal to 1 if the politician held that political position at some point, at the Federal or Provincial level, and 0 otherwise. Party is a categorical variable denoting the party affiliation of each politician. While the majority of politicians in the sample were affiliated with a single party, 81 politicians' Wikipedia pages recorded affiliation with multiple parties. For the purposes of this investigation, these politicians were manually assigned to the party with which they were most recently affiliated.

[1] Provincial parties, such as the Ontario Liberal Party, are also coded under their Federal counterparts.

[2] A shiny supplement to explore the Wikipedia dataset is linked here: https://7wouif-ethan-sansom.shinyapps.io/shiny/

Notably, the Wikidata query used in this paper was language agnostic, and included Wikipedia pages available in all languages. The page views, however, are those views from English language Wikipedia, with the exception of Sue Montgomery's biography, which was available only on French language Wikipedia. Consequently, while the sample of pages included is not limited by language, page views are biased towards politicians with primarily English speaking voters. This is particularly problematic for the analysis of pages associated with the Bloc Québécois, the most popular political party in Quebec, which contains the majority of Canada's French speaking population (Cross 2005).

2.2 Description

The majority of Wikipedia pages in the sample were viewed fewer than 1000 times in 2021, with some notable outliers in each party. Figure 1 depicts the page views of every politician in the sample, with the 2021 Federal election candidates from each party labelled by name.

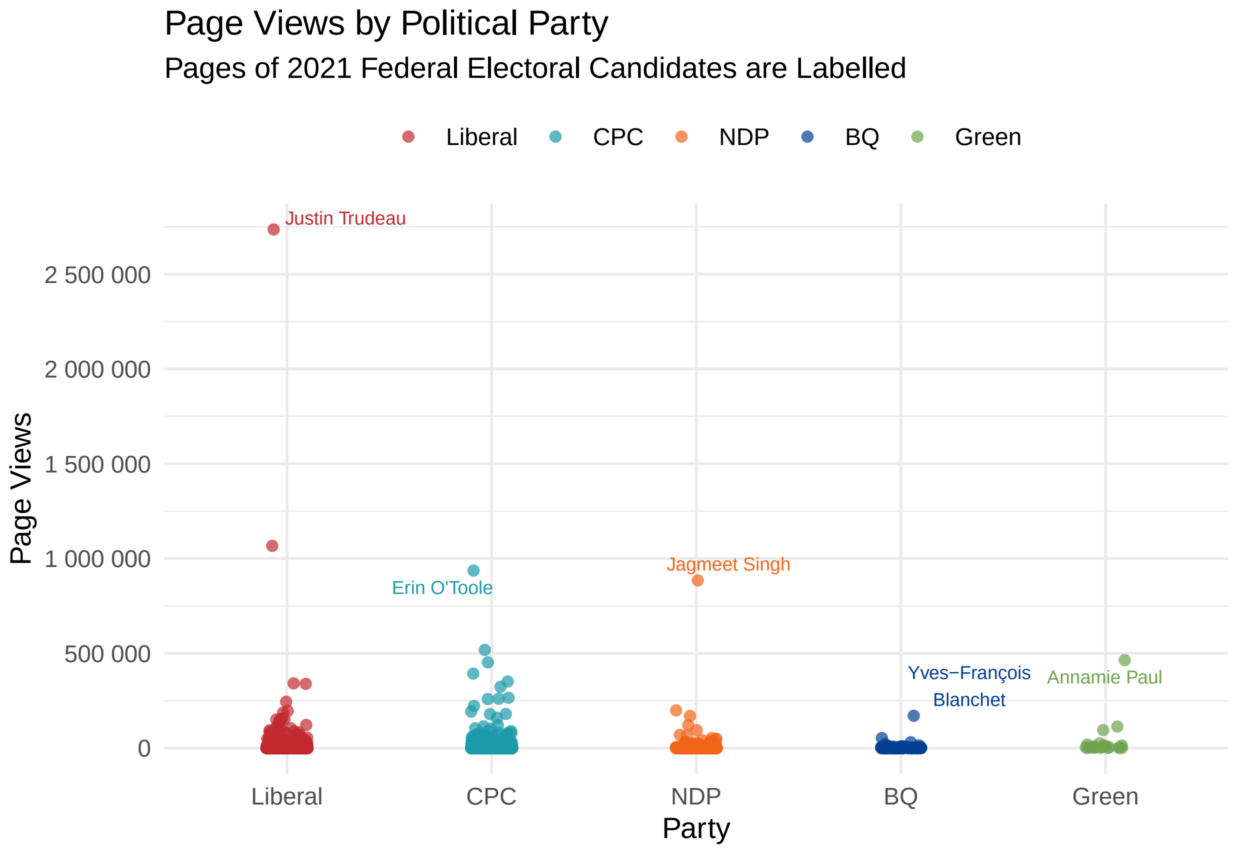

Figure 1: Wikipedia Page Views of Canadian Politicians by Party

The leaders of the Liberal, Conservative, New Democratic, Bloc Québécois, and Green parties are the most viewed pages within their respective parties. Justin Trudeau, who was reelected as Prime Minister in the 2021 Federal Election, has the most page views of any page by a wide margin, with greater than 2,500,000 page views. Erin O'Toole and Jagmeet Singh are the third and fourth most viewed pages, with near 1,000,000 page views each. Finally, Green and Bloc Québécois leaders Annamie Paul and Yves-François Blanchet have nearly 500,000 and 250,000 page views respectively.

Politicians which have held the position of Prime Minister, Party Leader, or Minister received substantially more page views in 2021 than those with no prior position. The median page views of non-position politicians, shown in Figure 2, is less than 1000 for all parties except Green. Median page views of Liberal, Conservative, Bloc Québécois, and Green party Prime Ministers, Party Leaders, and Ministers are all greater than 10,000. The distributions of non-position politicians page views have long left and right tails, with the exception of the Green party, indicating a number of outlying politicians with near zero or greater than 10,000 Wikipedia page views in 2021.

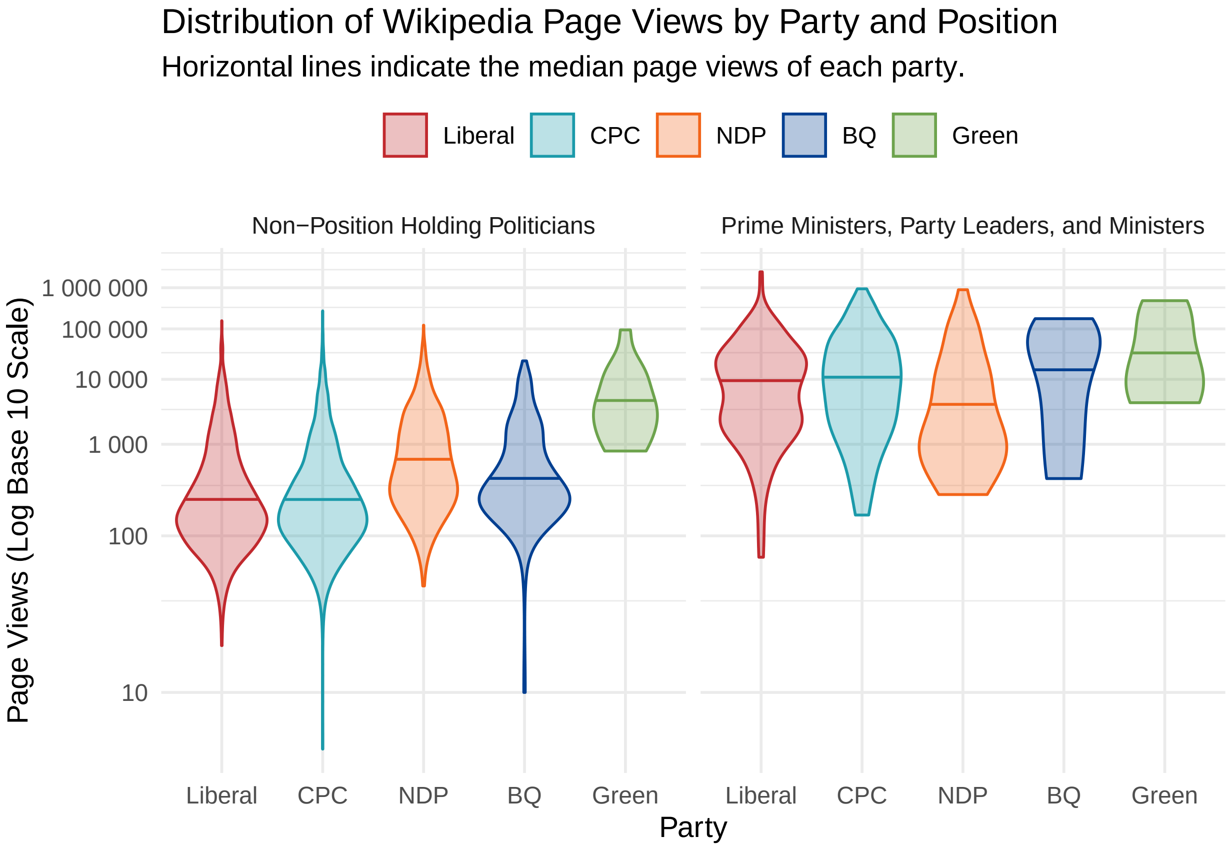

Figure 2: Distribution of Wikipedia Page Views by Party and Position

The Green party, which appears to receive more views on average than any other party, is particularly underrepresented on Wikipedia. The sample contains just 21 pages of Green party affiliated politicians, which have an average of 37,755.10 page views (Table 2). Notably, around 24% (2) of the Green party pages are party leaders, belonging to Annamie Paul and her predecessor Elizabeth May, which are the most viewed Green party pages. Page views within the sample of Green party are highly variable, with a sample standard deviation of 102,172.28, far greater than that of any other party. This indicates that the high average page views within the Green party is attributable to a few large outliers, which are highly influential in the small sample of 21 observations. Liberal party politicians had 3891.16 page views in 2021 on average (SD = 52,253.64), Conservatives 3167.43 (SD = 25,582.84), New Democrats 4499.04 (SD = 34,394.34), and Bloc Québécois politicians 3112.21 (SD = 14,774.53). Average page views do not differ substantially between parties, with the exception of the Green party, however the within party variation in page views is comparably very large. Only two of the five parties, Liberal and Conservative, have Prime Minister's among their members[3]. Four of the five parties, with the exception of the Green Party, are composed of between 1% and 4% Party Leaders and Ministers.

[3] The paper sample includes all 23 of Canada's Prime Ministers.

Table 2: Balance Table of Variables by Party

	Liberal (N=3420)		CPC (N=3504)		NDP (N=759)		BQ (N=152)		Green (N=21)	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Page Views	3891.16	52253.64	3167.43	25582.84	4499.04	34394.34	3112.21	14774.53	37755.10	102172.28
Birth Year	1895.13	46.65	1899.01	47.50	1947.02	20.70	1953.02	13.40	1964.65	11.76
Prime Minister (%)	0.32	5.66	0.34	5.84	0.00	0.00	0.00	0.00	0.00	0.00
Party Leader (%)	1.11	10.48	1.20	10.88	1.98	13.93	1.97	13.96	23.81	43.64
Minister (%)	3.92	19.41	2.08	14.28	2.64	16.03	1.32	11.43	0.00	0.00

Table 3 shows summary statistics of the full sample of politicians, which are similar to those of the Liberal, Conservative, New Democratic, and Bloc Québécois parties, indicating that there is little between party variation in the sample.

Table 3: Descriptive Statistics of Continuous Variables

	Mean	SD	Missing (%)
Page Views	3702.90	40 356.06	0
Birth Year	1902.85	47.73	3
Prime Minister (%)	0.29	5.40	0
Party Leader (%)	1.31	11.38	0
Minister (%)	2.91	16.82	0

Figure 3 depicts the density of page views by politician birth year. Page views appear to increase as politician birth year increases, with the majority of page views attributable to politicians born between 1930 and 1970. At nearly all birth years, from the early 1800's to 1980's, there remain outlying politicians with greater than 100,000 page views. Relatively few politicians born before 1925, however, received greater than 10,000 page views during 2021.

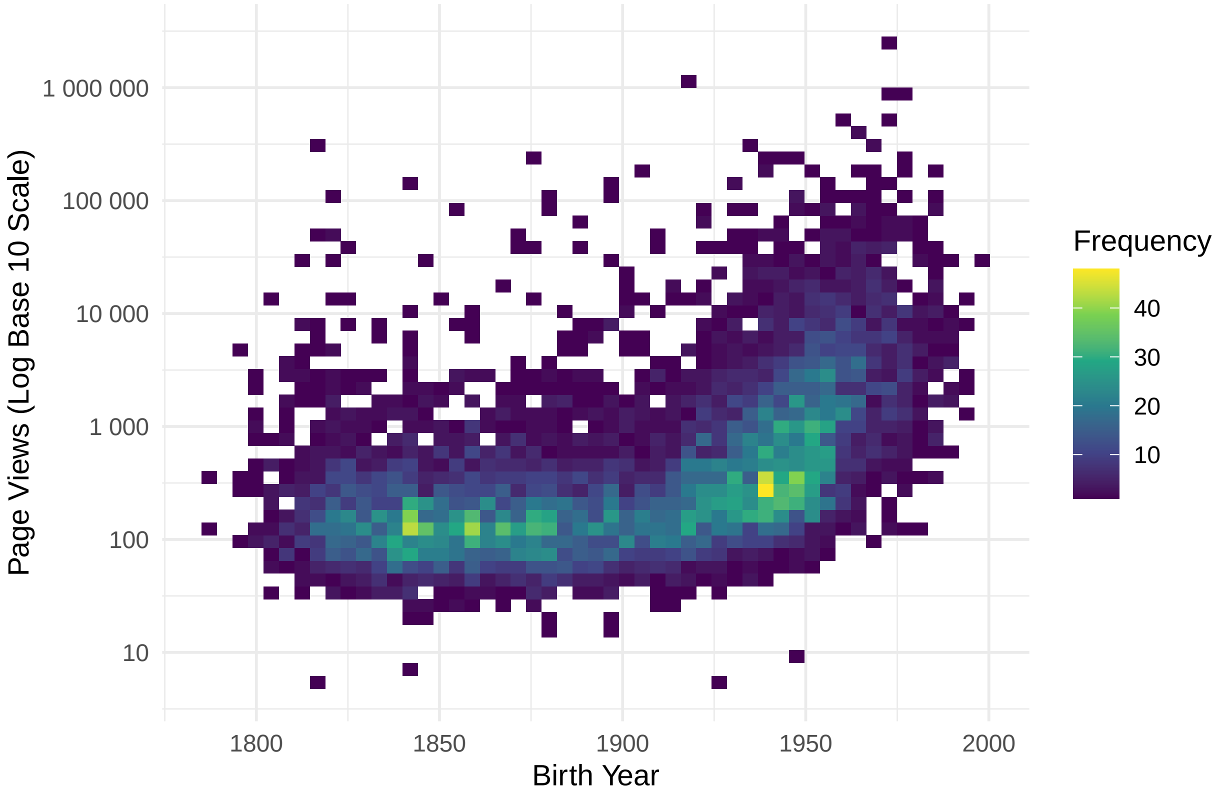

Figure 3: Distribution of Wikipedia Page Views by Politican Birth Year

3 Model

3.1 Causal Relationship

The utility of Wikipedia page views as a source of data for predicting elections and measuring the favourability of political parties is predicated on Figure 4, which depicts a causal relationship between a politicians' political affiliation and the Wikipedia page views which the politician receives. If a party's popularity causes the Wikipedia page views of its members to increase, then analysts, pundits, and political organizers alike can use Wikipedia as a tool to accurately assess or forecast election outcomes based on the page views and party affiliation of politicians.

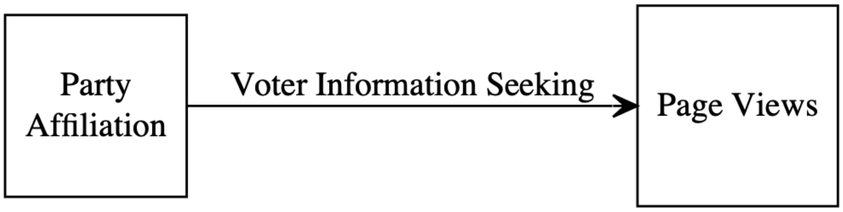

Figure 4: Casual Relationship Between Party Affiliation and Page Views

This simple relationship is premised on two assumptions. Firstly, voters' information seeking causes the Wikipedia page views of politicians to increase. This is an intuitive assumption, as Wikipedia is an encyclopedia and has little utility to most constituents other than as a source of information. Further, several studies have demonstrated that voters frequently use the internet broadly, and Wikipedia specifically, as a source of political information (Yasseri and Bright 2015, 2013; Salem and Stephany 2021). Table Figure 7 and Figure 10 demonstrate that Canadian politician's Wikipedia page views spike nearing a Federal election, suggesting that people access politician's Wikipedia pages when they become relevant for making voting decisions.

Secondly, it is assumed that a politicians' party affiliation causes voters to change their information seeking behavior regarding that politician. Specifically, that association with a popular political party causes an increase in voter information seeking, and consequently Wikipedia page views, compared to an association with a less favorable party. While less obvious than the previous claim, this assumption has been reaffirmed in similar contexts. In an analysis of European Parliamentary elections, Yasseri and Bright (2015) found theoretical and empirical evidence suggesting that Wikipedia is an important source of information for undecided swing-voters. These voters are more likely to seek information on parties with improving popularity, which tend to attract swing voters, and on small parties, for which less information is available elsewhere (Yasseri and Bright 2015). Utych and Kam (2014) assert that prior to elections, voters seek to both maximize on information on electoral candidates and minimize personal effort, resulting in voters limiting their information seeking towards a small number of strong candidates for election. Using both observational and experimental data, Utych and Kam (2014) found a positive relationship between a candidate's political viability and Google searchers for the candidate's name. Assuming that stronger political parties produce, or are produced by, stronger political candidates, this indicates that a party's political viability generates increased information seeking by the party's voter base.

To accurately assess the magnitude and direction of the relationship between party affiliation and Wikipedia page views, it is important to consider additional factors which are causally related to both party affiliation and page views. Figure 5 depicts a more complicated causal relationship, in which the effect of party affiliation on page views is confounded by politicians' birth year, and page views is additionally related to politician's political position and notoriety.

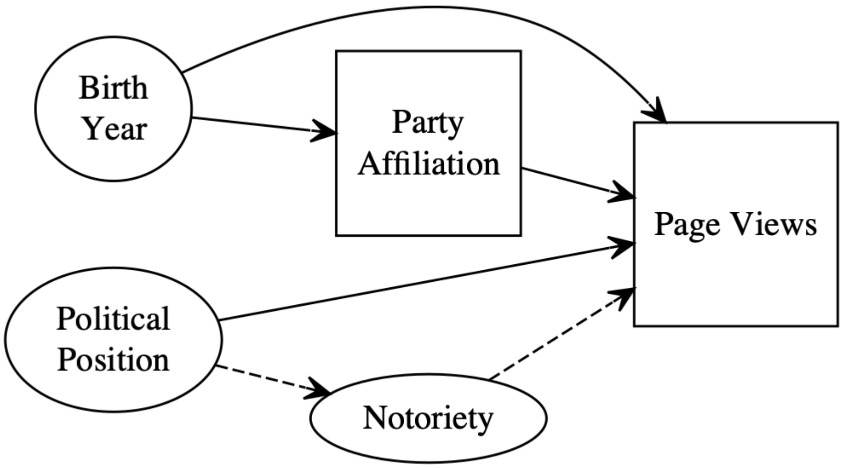

Figure 5: Proposed Causal Model

Only two of the five major political parties in Canada, the Liberal and Conservative parties, have existed in some form prior to 1950 (Cross 2005). Thus, politicians who died prior to 1950 were forced, within this sample, to either affiliate with the Liberal or Conservative parties. This is the main justification for the causal effect of politician birth year on party affiliation. Peterson, Smith, and Hibbing (2020) find in a longitudinal study of American adults that political leanings are also somewhat more likely to shift from liberal to conservative with age, although the authors caution against early work which reported stronger impacts of aging on political views. The mechanism by which politicians' birth year effect page views is similar to that of party affiliation, namely that politicians who are still alive and are young enough to continue seeking office remain relevant to voters and thus encourage information seeking behavior. Given that politicians' birth year is related to both the page views and party affiliation, it is considered a confounding variable and should be accounted for when modelling the causal effect of party affiliation on page views (Dablander 2020). Failure to do so will result in models which misrepresent the effect of party affiliation on page views, some of which is attributable to the effect of politicians' birth year.

Additionally, a politicians' political position, such as Prime Minister or Party Leader, is likely to causally effect their Wikipedia page views by the same mechanism as party affiliation. Specifically, politicians with more influential positions and which participate in larger, more consequential elections, are more relevant to voters and will thus encourage greater information seeking behavior relative to other politicians in less influential positions. This effect is further mediated by an unobserved variable in the sample of politicians' Wikipedia pages, notoriety. The assumption being that politicians' of greater political status have the power to make more consequential mistakes, and thus become controversial. Constituents then, are incentivised to seek further information on the politician's actions, which are relevant for making voting decisions.

3.2 Model Specification

The following linear model will be used to estimate the causal relationship depicted in Figure 5.

$$Pageviews_i = \beta_0 + \beta_1(Birthyear_i - 1950) + \alpha Party_i + \gamma Position_i + \epsilon_i, \tag{1}$$

The outcome variable, $Pageviews_i$, is the total number of Wikipedia page views received by politician i in 2021.

$Birthyear_i$ is the year in which politician i was born. Politician birth year is centered about the year 1950 in the model, and thus β_1 can be interpreted as the change in a politician's page views associated with being born an additional year after 1950.

$Party_i$ is a vector of indicator variables equal to 1 if politician i is affiliated with the named party and 0 otherwise. Specifically, $\alpha Party_i = \alpha_1 Liberal_i + \alpha_2 CPC_i + \alpha_3 Green_i + \alpha_4 BQ_i$. Note that an indicator for the NDP is omitted from the model and serves as a reference variable. As politicians in the sample are each affiliated with a single party, if a politician is in none of the parties included in the model ($Liberal_i = CPC_i = Green_i = BQ_i = 0$) then that politician must be affiliated with the NDP. Thus, politician i's inclusion in the NDP is determined by regressors in $Party_i$[4]. Including an indicator for NDP affiliation would render the model perfectly collinear and prevent estimation of the model using ordinary least squares (Wooldridge 2018). Estimates of the effect of party affiliation are consequently relative to the effect of NDP affiliation on Wikipedia page views. The estimate of α_1, for example, is interpreted as the difference in average page views of Liberal politicians and NDP politicians, holding other covariates constant. A positive α_1 would indicate that Liberal politicians receive more page views on average than NDP politicians, other variables held constant.

$Position_i$ is a vector of indicator variables equal to 1 if politician i is has attained the named position at some point in their lifetime and 0 otherwise. In particular, $\gamma Position_i = \gamma_1 PrimeMinister_i + \gamma_2 Minister_i + \gamma_3 PartyLeader_i$. Each coefficient can be interpreted as the change in Wikipedia page views which is associated with a politician having attained the given political position. Note that, unlike party affiliation, in which every politician in the sample is affiliated with a single party, many politicians in the sample have held multiple positions. Thus, the estimate of γ_1 for instance, measures the additional change in page views, on average and holding other covariates fixed, of attaining the position of Prime Minister.

Finally, ϵ_i is a random error term.

In addition to the full model, three partial models are estimated to compare the relevance of individual regressors for predicting Wikipedia page views and to reaffirm the role of birth year as a confounding variable. The remaining models have the following functional forms:

$$Pageviews_i = \beta_0 + \alpha Party_i + \epsilon_i \tag{2}$$

$$Pageviews_i = \beta_0 + \beta_1(Birthyear_i - 1950) + \alpha Party_i + \epsilon_i \tag{3}$$

$$Pageviews_i = \beta_0 + \gamma Position_i + \epsilon_i. \tag{4}$$

The application of these linear models serve several purposes in this paper. Firstly, the the estimates produced by these models provide a succinct description of how each politician characteristic is correlated with the outcome variable, page views, in the sample. In particular, the regression estimates provide a description of the partial association between page views and politician characteristics, such as birth year or political party, while holding the impact of other characteristics constant[5]. Secondly, to the extent that the models' correctly specify the causal relationship between politicians' characteristics and politicians' Wikipedia page

[4] In particular, an indicator NDP_i is a linear combination of the other party indicators, where $NDP_i = Liberal_i + CPC_i + Green_i + BQ_i$.

[5] This could be achieved, less succinctly, by blocking observations into homogeneous groups, politicians of the same party for example, and estimating the correlation between other factors, political position or birth year, and page views within each group.

views, these estimates provide a measure of the effect of belonging to a particular party on a politicians Wikipedia page views. Under the assumption that Wikipedia page views are themselves a valid proxy variable for voter information seeking, then the models provide a measure of voter interest in each of the five major Canadian Federal political parties. As will be demonstrated in the Model Assumptions Section 4.2, the linear models in this paper do not appear to correctly specify the relationship between politicians' characteristics and politicians' Wikipedia page views, nor do they satisfy the assumptions necessary for the internal validity of linear regression estimation. While falling short of causal inference, the results of these models can nonetheless be used to diagnose how and why linear regression analysis may fail to capture the causal relationship between politicians' characteristics and page views. The difficulties encountered here can subsequently be used to inform model and sample selection in future papers which analyze Wikipedia usage and information seeking behavior.

4 Results

Figure 6 compares the share of Wikipedia pages and page views of each party to the outcomes of the 2021 Federal election. The election gave the Liberal Party a minority government, with 159 Liberal candidates elected to the the House of Commons, with the Conservative Party forming the opposition with 119 seats despite narrowly taking the popular vote (Sam 2021; House of Commons Canada 2021). The New Democratic, Bloc Québécois, and Green parties won 25, 33, and 2 of 338 total House seats respectively (Sam 2021; House of Commons Canada 2021). The People's Party, not included in this analysis, gained around 4.9% of the popular vote, but failed to elect any candidates to the House of Commons (Sam 2021).

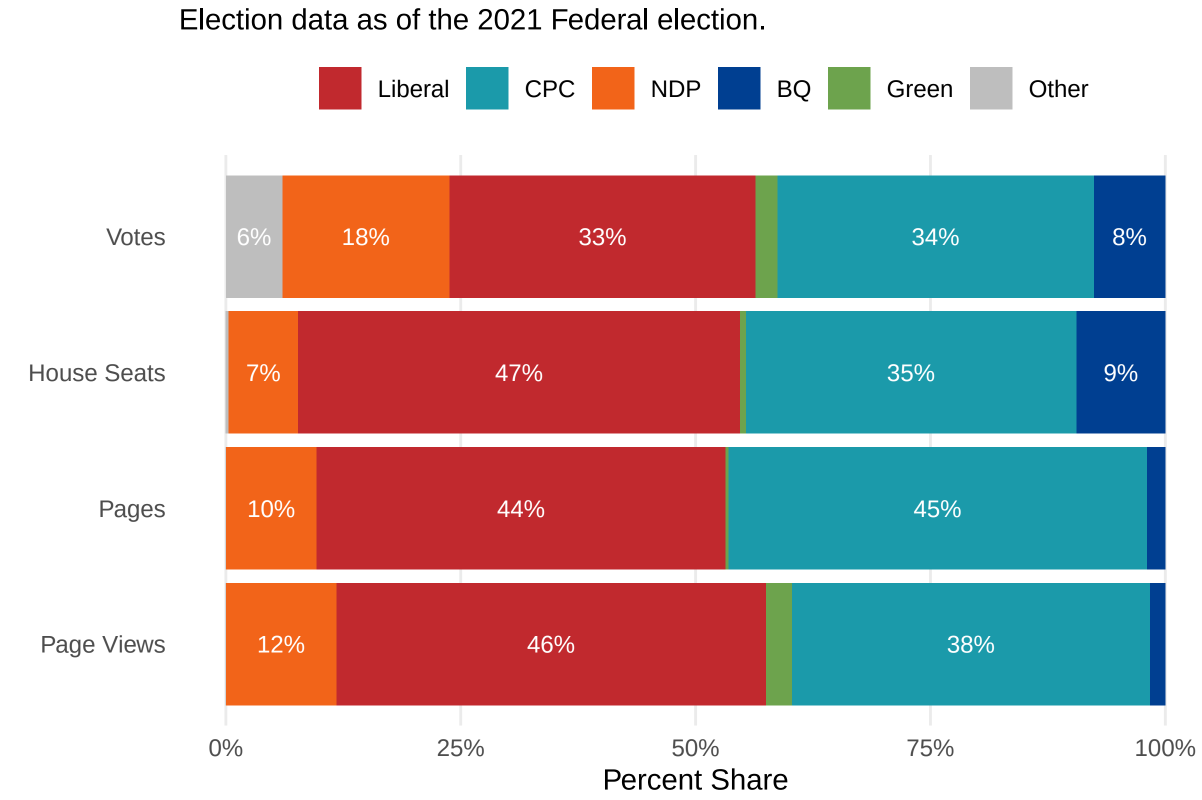

Figure 6: Share of Wikipedia Pages, House of Commons Seats, and Votes by Party

The share of Wikipedia page views in 2021 for the Liberal (46%), Conservative (38%), and New Democratic (12%) parties are very similar to the parties' share of House seats, 47%, 35%, and 7% respectively, and somewhat similar to the popular vote share. Green party page views and popular vote share appear similar as well. The share of page views of Bloc Québécois affiliated politicians, however, is far less than the share of Bloc Québécois votes and House seats. This is likely due to the downward bias on Bloc Québécois page views in the sample, as noted in the Data Section 2. The page views used in this analysis are those on English language Wikipedia, and thus fail to capture the Wikipedia activity of majority French speaking voters, many of which live in Quebec and disproportionately support the Bloc Québécois party (Cross 2005).

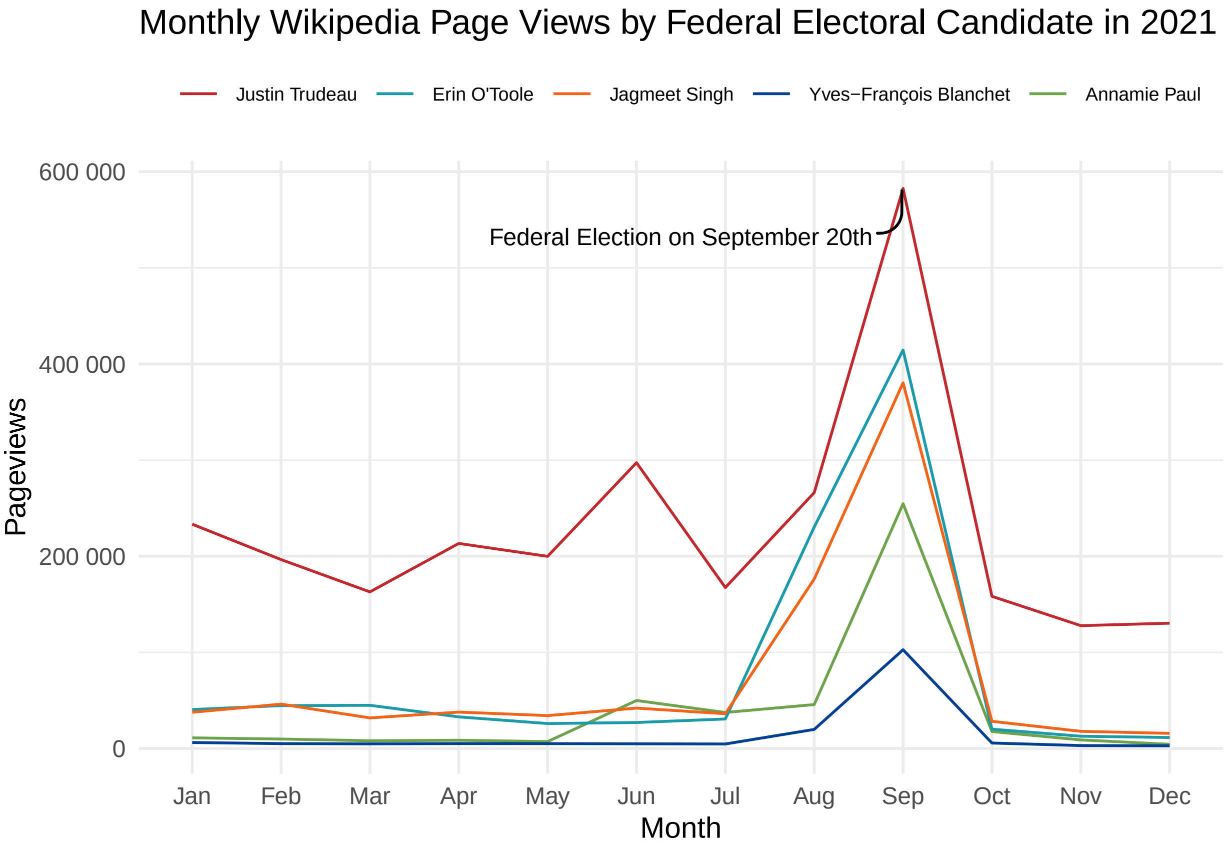

Figure 7: Monthly Wikipedia Page Views by Federal Electoral Candidate in 2021

Figure 7, shows the page views of the 2021 Federal election candidates in the months surrounding the election which occurred September 21st, 2021. Notably, every party leader's page views increased substantially in the months preceding the election, supporting the causal connection made between information seeking and party affiliation in Model Section 3.1. Justin Trudeau, received the greatest number of page views in the months immediately preceding the election, and received substantially more page views than other candidates for the majority of the year. Trudeau's pre-election page views are likely a result of his position as incumbent Prime Minister preceding the election granting greater name recognition and interest, nationally as well as internationally. compared to the other candidates.

The cumulative monthly page views of the five political parties, excluding the counts of the electoral candidates, are shown in Appendix Table 10. The Green and Bloc Québécois parties received very few page views for the majority of the year, with a small spike into the thousands of page views during September 2021. The New Democratic Party consistently received around 200,000 page views per month and nearly 500,000 in September. Interestingly, the Liberal and Conservative parties received near identical page views for the majority of the year, both climbing to around 1,500,000 page views in September 2021. It should be noted when interpreting the cumulative page views of each party, that the Liberal and Conservative parties comprise the majority of Wikipedia pages in the sample with 3420 and 3504 pages respectively. The New Democratic, Bloc Québécois, and Green parties have a far smaller presence on Wikipedia, with 759, 152, and 21 pages respectively.

4.1 Model Results

In Table 4, the *Full* model, which regresses Wikipedia page views on politician birth year, political position, and party affiliations, estimates that on average, relative to New Democratic politicians, and holding birth year constant, Liberal Party affiliation increases page views by 1156.28 (95% CI [-1960.50, 4273.07]), Conservative Party affiliation by 742.20 (95% CI [-2342.60, 3827.00]), Bloc Québécois affiliation by -1616.58 (95% CI [-1960.50, 4273.07]), and Green Party affiliation by 18,465.17 (95% CI [2396.33, 34,534.01]). These estimates are somewhat indicative of trends observed in the data, primarily Liberal politicians receive the greatest number of page views in the sample, followed by Conservatives and New Democratic, with Bloc Québécois politicians receiving the fewest page views. With the exception of the Green Party, the confidence intervals for party effect are relatively wide, containing a range of -1500 to 1500. There is thus little statistical justification to indicate that in the aggregate, party affiliation is a significant indicator of page views. The Green Party, however, is associated with a very substantial increase in page views relative to the mean page views in the sample of 3702.90 (Table 3). This effect estimate should be interpreted with skepticism, as multiple linear regression model estimates are easily biased to influential outliers in small samples, such as the case with the sample of Green Party politicians.

The *Full* model adjusted R-Squared ($Adj.R^2 = 0.234$), which provides a measure of the proportion of variability (movement) in the outcome variable, which is explained by variability (movement) in the model regressors, is very slightly greater than that of the *Position* model ($Adj.R^2 = 0.230$), which includes only political position covariates. This indicates that political position, far more so than political affiliation, predicts politicians' Wikipedia page views in the sample. The *Full* model estimates that, on average, and holding birth year and party affiliation constant, the effect of having held the title of Prime Minister on page views is 262,779.40 (95% CI [247,377.26, 278,181.55]), Party Leader is 69,042.35 (95% CI [61,608.87, 76,475.82]), and Minister is 30,722.04 (95% CI [25,846.14, 35,597.93]). These effect estimates and confidence intervals are similar in the *Position* model. Relative to mean and median page views observed in the full sample, these effect estimates are very large. As with the estimated effect of Green Party affiliation, these estimates should be interpreted very cautiously. Prime Ministers, Party Leaders, and Ministers make up less than 5% of the total sample, shown in Table 3, and the page views of positioned politicians are highly variable, as depicted in Figure 2. Further, the electoral candidates from each of the five political parties are all party leaders, while Justin Trudeau is also a Prime Minister, and are among the largest outlying observations in respect to page views. This combination of high within group variability, large outliers, and small sample sizes are likely to bias the linear regression estimates of political positions' effect on page views.

The *Party* model, which includes only party indicator variables, estimates effects of Liberal and Conservative party affiliation of -607.88 and -1331.61, contrary to expectations and previous observations. In the *Party and Birth Year* model, which adds politicians birth year as a covariate, estimates effects of Liberal and Conservative party affiliation of 2998.03 and 1990.82 respectively, and a modest effect of 75.72 (95% CI [55.19, 96.24]) for birth year. This provides some indication that, as proposed in Model Section 3.1, birth year confounds the effect of party affiliation on page views. The adjusted R-Squared of the *Party* and *Party and Birth Year* models are 0.002 and 0.008 respectively, further affirming that party affiliation and birth year explain very little variation in politicians' Wikipedia page views.

A large potential source of variability and bias in the model are the large number of politicians in the sample who were not alive in 2021. Under the assumption that voter information seeking, for gaining information relevant to elections and to access party viability, drives increased politician page viewership on Wikipedia, there is little reason to believe that voters would seek information on political candidate who are no longer active or living. Thus, there is not a causal justification, in the models presented here, which relates the party affiliation or position of deceased politicians to Wikipedia page views. To test this claim, the regression models were re-estimated using the sub-sample Wikipedia pages of politicians born after 1920 and with no recorded year of death on Wikipedia. Appendix Table 5 displays the results of the restricted models. While model fit is slightly increased by restricting the sample in this way (restricted *Full* model $Adj.R^2 = 0.319$), effect estimates and confidence intervals are not substantially different, with the exception of politician birth year for which the restricted *Full* model effect estimate is 477.37 (95% CI [310.55, 644.19]).

Table 4: Model Results

	Party	Party and Birth Year	Position	Full
Liberal	−607.88	2998.03		1156.28
	(−3782.07, 2566.30)	(−537.55, 6533.62)		(−1960.50, 4273.07)
CPC	−1331.61	1990.82		742.20
	(−4498.94, 1835.71)	(−1514.69, 5496.33)		(−2342.60, 3827.00)
BQ	−1386.83	−2249.92		−1616.58
	(−8411.99, 5638.33)	(−9496.35, 4996.50)		(−7987.80, 4754.65)
Green	33 256.05	33 009.79		18 465.17
	(15 769.01, 50 743.09)	(14 833.85, 51 185.74)		(2396.33, 34 534.01)
Prime Minister			260 325.47	262 779.40
			(245 167.44, 275 483.51)	(247 377.26, 278 181.55)
Party Leader			71 423.51	69 042.35
			(64 157.37, 78 689.65)	(61 608.87, 76 475.82)
Minster			32 064.42	30 722.04
			(27 389.52, 36 739.32)	(25 846.14, 35 597.93)
Birthyear - 1950		75.72		51.63
		(55.19, 96.24)		(33.31, 69.94)
Num.Obs.	7836	7577	7836	7577
R2	0.002	0.009	0.230	0.234
R2 Adj.	0.002	0.008	0.230	0.234
F	3.998	13.707	780.152	289.496

Estimate 95 percent confidence interverals appears in brackets. Birthyear - 1950, measures the number of years between the politician's year of birth and the year 1950. Liberal, CPC (Conservative Party of Canada), Green, and BQ (Bloc Québécois) are party indicators equal to 1 if the politician is a member of the party, and 0 otherwise. The NDP (New Democratic Party) is excluded from the set of party indicators and is used as the reference group. Prime Minister, Party Leader, and Minister are position indicators equal to 1 if the politician has held that position and 0 otherwise.

4.2 Model Assumptions

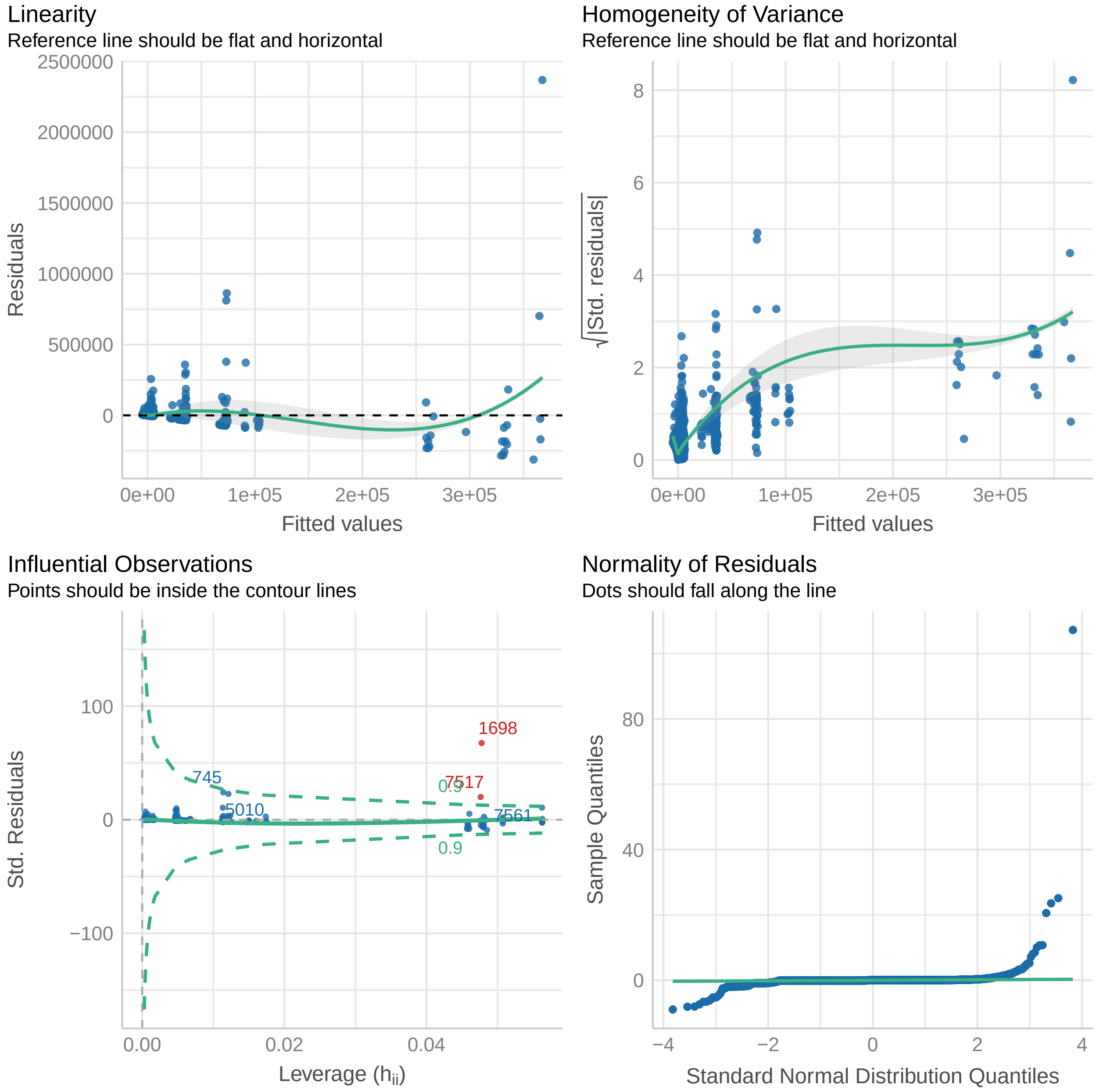

Figure 8: Evaluation of Full Model Assuptions

To perform valid statistical and causal inference, regression models must satisfy a number of important assumptions. Critically, the functional form of the regression models used in this paper are explicitly linear and thus the underlying relationship between Wikipedia page views and politicians' characteristics is assumed to be linear as well. The error term ϵ is also assumed to be random and independent of regressors in the model. In particular, ϵ is often assumed to be distributed normally, with 0 mean and constant variance. Systematic deviations of ϵ from this mean indicate non-linearity in the model and non-constant variance of the error term jeopardizes the validity of statistical inference, particularly in small sample settings.

Figure 8 depicts a number of graphical checks for the linearity and error assumptions of the unrestricted

Full model discussed in the Model Results Section 4.1. In Figure 8, the model residuals, the difference between the actual page views of each Wikipedia page and those which are predicted by the model, are used to estimate ϵ. The *Full* model fails the visual diagnostic tests for normality, linearity, and constant (homogeneous) error variance. In particular, error variance increases systematically as page views increase, indicating a non-linear relationship between regressors and page views in the model. Additionally, Figure 8 provides evidence of influential outliers in the sample which, as noted previously, bias regression estimates.

Appendix Figure 11 displays the same set of diagnostic tests for the restricted sample *Full* model, which appears less linear than the unrestricted model and contains more influential outlying observations.

5 Discussion

As demonstrated by several authors, Wikipedia provides a strong tool for assessing the information seeking behavior, and political intentions, of the voters in North America and Europe (Yasseri and Bright 2013, 2015; Smith and Gustafson 2017). In the single Federal election observed in this paper in Figure 7 and Figure 6, the share of Wikipedia page views of political parties and electoral candidates were shown to resemble the popular vote and House of Commons seats share in the 2021 Canadian Federal Election. Further, as demonstrated in Figure 7 and Appendix Figure 10, the Wikipedia page views of Canadian politicians and electoral candidates spiked in the month of and months preceding the 2021 election. This suggests that Canadian voters use Wikipedia to inform themselves on relevant political actors preceding an election, as described by Salem and Stephany (2021) and Utych and Kam (2014). The results of this initial investigation were not replicated in the subsequent linear regression analysis, the estimates of which suggest that the effect of party affiliation on Wikipedia page views is small or nonexistent. These models, however, were shown in Model Section 4.2 to be flawed in a number of ways, and subsequent analysis is needed to determine a use case, if any, for using linear models to analyze characteristics of Canadian politicians' Wikipedia pages.

5.1 Limitations

As noted in Model Section 4.2, the linear regression models estimated in this paper have severe limitations, pertaining to both causal and statistical inference. The expected causal relationship between Canadian politicians' party affiliation and Wikipedia page views in 2021 is not evidenced by the regression models, which estimate very moderate effects of party affiliation for all but the Green Party. Several variables, including the set of political position indicators and the indicator for Green Party affiliation, have small underlying samples of observations. Within the small sub-samples of positioned politicians and Green party members, there are high within group variability of politician page views. Small sample sizes, coupled with high within group variance, likely bias the regression estimates of these parameters and inhibit valid statistical inference. Further, the *Full* regression model specified appears to have non-normally distributed error terms as well as heterogeneous error variance, as indicated by Figure 8. Consequently, small sample inference, which relies on a strong assumption of identically distributed normal error terms, is likely invalid for constructing regression estimates of these parameters (Wooldridge 2018). Finally, both the unrestricted and restricted models estimated appear non-linear in their parameters and contain influential outlying observations, both of which bias model estimates.

Several sampling factors additionally limit the results of this paper. The failure to collect Wikipedia page views from both French and English sources, as noted in the Data Section 2, creates a downward bias on the page views of Bloc Québécois affiliated politicians. This problem is difficult to overcome, as introducing additional non-English Wikipedia pages will potentially account for page views of non-Canadian citizens. Including such page views may bias upwards the page views of high ranking and incumbent politicians, who are more likely to appear in foreign press and impact foreign affairs, thus encouraging increased page views from a population outside of the Canadian electorate. The issue of foreign viewership remains in the current sample as well, as there is no way to differentiate between the page views of Canadians and foreign citizens on English language Wikipedia. Thus, several mechanisms by which individuals seek information internationally may be more influential than domestic voter information seeking behavior in determining the page views of Canadian politicians.

5.2 Future Analysis

Future analyses of Canadian politicians' Wikipedia pages may avoid many of the shortcomings of this paper by using more selective sampling strategies. This analysis, which cast a wide net to include all Wikipedia pages of Canadian politicians affiliated with the five major Federal political parties, was hampered by high within group variability, influential outlying observations, and relatively small sub-sample blocks of observations. A more targeted approach, which investigated the Wikipedia page views of a smaller homogeneous set of Canadian politicians, would be subject to fewer such concerns. The sample could, for instance, be limited only to the Wikipedia pages of candidates for prior or subsequent provincial parliament elections. These candidates are more likely to be of similar ages, are certainty still living and practicing politicians, and are likely of similar political positions. Controlling for variability in these factors would allow for a more precise estimate of the effect of political party and of politicians political viability on Wikipedia page views.

Additionally, future work could incorporate time series models into their analysis. Much of the variation in monthly page views of the federal electoral candidates and of politicians in the sample generally is attributable to the spike in page views preceding the 2021 election. It is during this period that Wikipedia page views appear most closely aligned with the proposed model of voter seeking behavior. Thus, analyzing changes in page views over time, and particularly during consequential elections, may provide stronger insight into when and to what extent Wikipedia page views are an accurate indication of voter behavior and attitudes. Such analysis would add to the expanding literature using Wikipedia to predict political and other behaviours, benefitting researchers, pundits, and political organizers in future elections.

6 Appendix

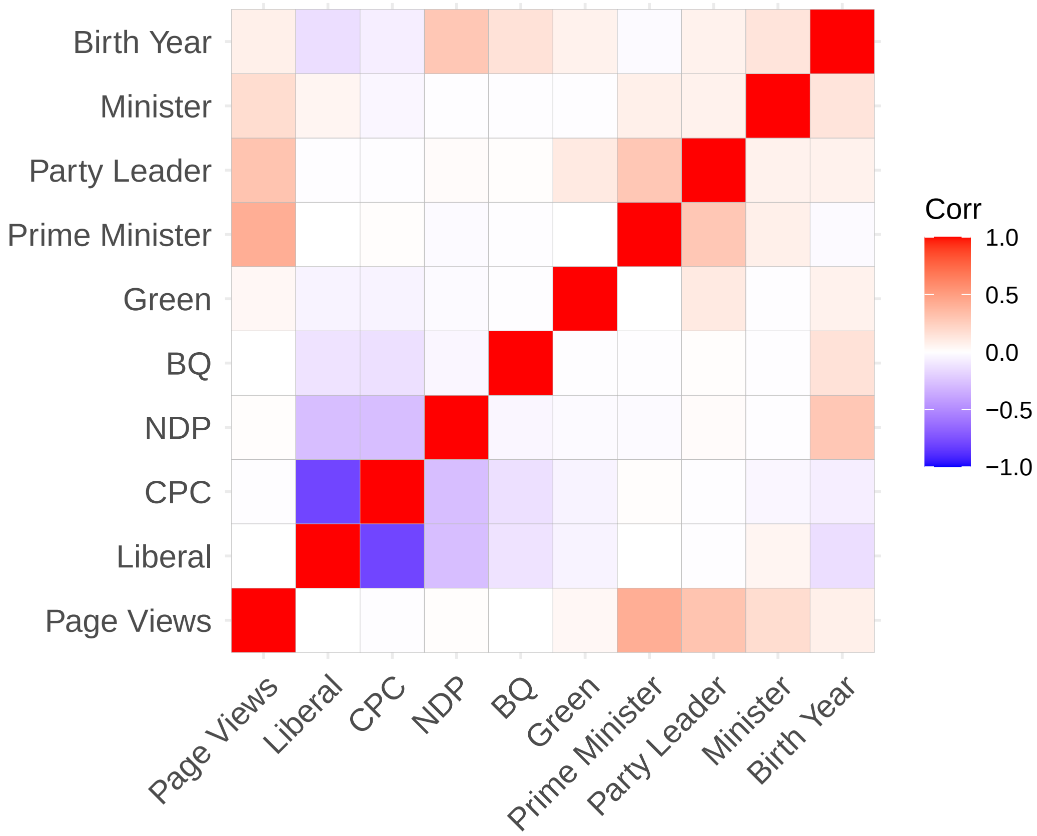

Figure 9: Correlation of Regression Variables

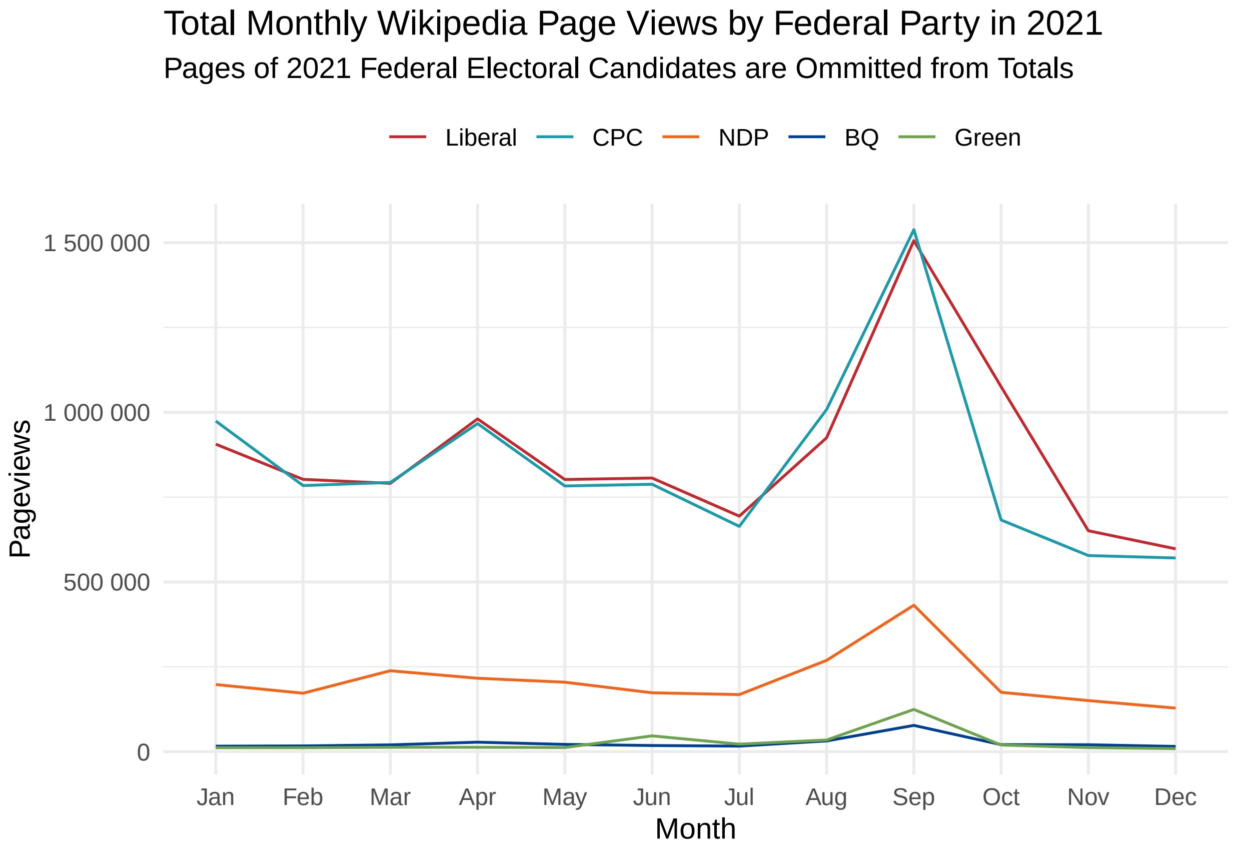

Figure 10: Total Monthly Wikipedia Page Views by Federal Party in 2021

Table 5: Restricted Model Results

	Party	Party and Birth Year	Position	Full
Liberal	5596.14	7083.53		2649.81
	(−1840.35, 13 032.63)	(−342.36, 14 509.43)		(−3557.52, 8857.14)
CPC	2762.49	4852.55		1579.46
	(−4462.37, 9987.34)	(−2387.25, 12 092.35)		(−4433.08, 7592.00)
BQ	−2061.07	−1593.94		−792.10
	(−15 069.35, 10 947.22)	(−14 543.27, 11 355.40)		(−11 535.72, 9951.52)
Green	33 727.25	29 184.35		9141.32
	(3811.75, 63 642.75)	(−647.88, 59 016.59)		(−15 838.24, 34 120.88)
Prime Minister			555 308.61	560 614.33
			(512 650.60, 597 966.62)	(518 129.63, 603 099.03)
Party Leader			93 825.61	91 718.66
			(79 153.76, 108 497.46)	(76 981.75, 106 455.57)
Minster			29 775.36	27 108.15
			(20 783.51, 38 767.20)	(18 039.13, 36 177.16)
Birthyear - 1950		500.14		477.37
		(300.08, 700.20)		(310.55, 644.19)
Num.Obs.	2472	2472	2472	2472
R2	0.003	0.013	0.312	0.321
R2 Adj.	0.001	0.011	0.311	0.319
F	1.802	6.262	373.271	145.693

Politicians born before year 1920 or which have a recorded year of death are omitted from the specification. Estimate 95 percent confidence interverals appears in brackets. Birthyear - 1950, measures the number of years between the politician's year of birth and the year 1950. Liberal, CPC (Conservative Party of Canada), Green, and BQ (Bloc Québécois) are party indicators equal to 1 if the politician is a member of the party, and 0 otherwise. The NDP (New Democratic Party) is excluded from the set of party indicators and is used as the reference group. Prime Minister, Party Leader, and Minister are position indicators equal to 1 if the politician has held that position and 0 otherwise.

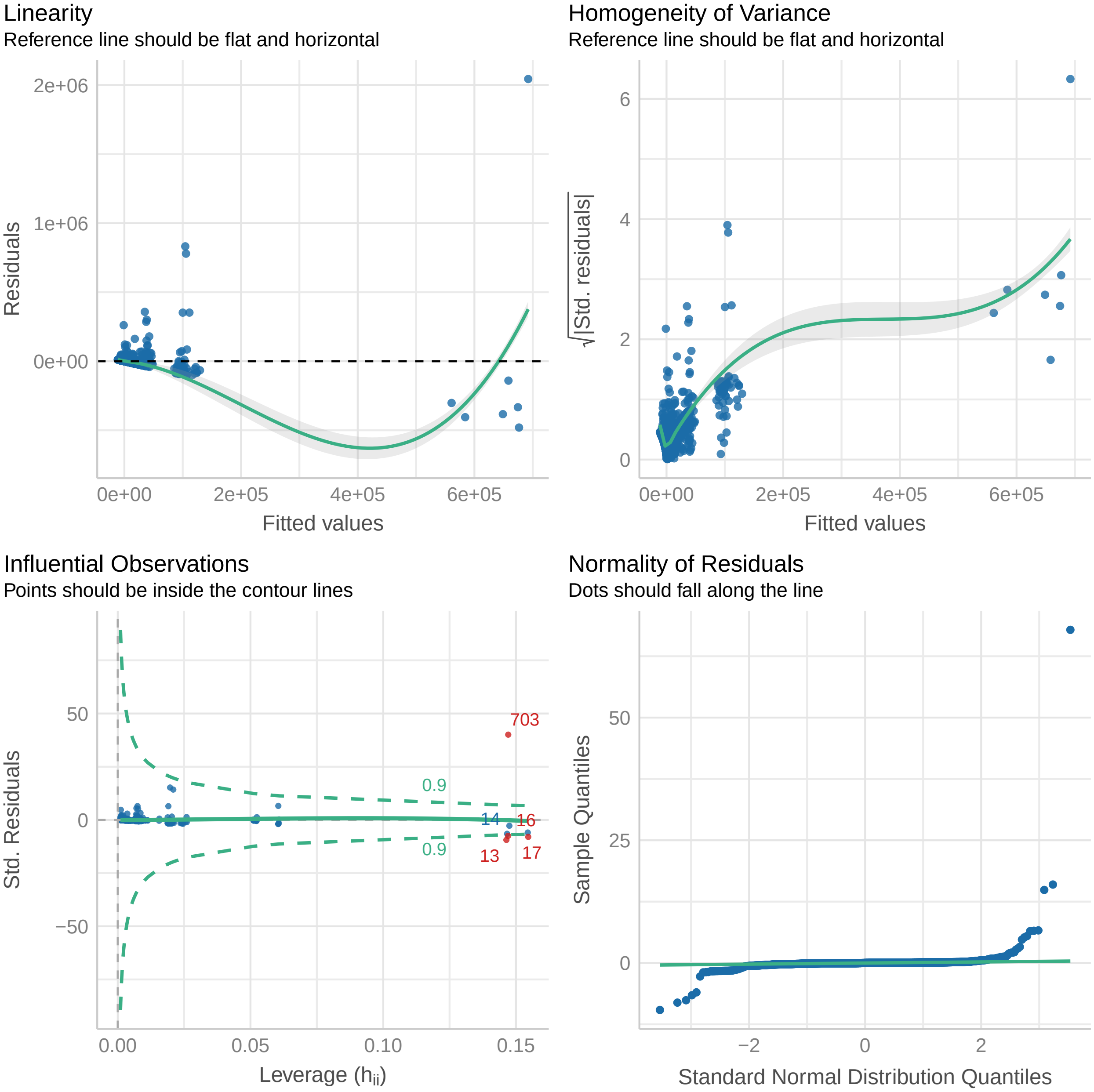

Figure 11: Evaluation of Restricted Full Model Assuptions

References

Arel-Bundock, Vincent. 2022. *Modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready*. https://CRAN.R-project.org/package=modelsummary.

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

Cross, William. 2005. *Political Parties*. UBC Press.

Dablander, Fabian. 2020. "An Introduction to Causal Inference." PsyArXiv. https://doi.org/10.31234/osf.io/b3fkw.

House of Commons Canada. 2021. "Member of Parliment." https://www.ourcommons.ca/members/en/party-standings.

Iannone, Richard. 2016. *DiagrammeRsvg: Export DiagrammeR Graphviz Graphs as SVG*. https://CRAN.R-project.org/package=DiagrammeRsvg.

———. 2022. *DiagrammeR: Graph/Network Visualization*. https://CRAN.R-project.org/package=DiagrammeR.

Kassambara, Alboukadel. 2019. *Ggcorrplot: Visualization of a Correlation Matrix Using 'Ggplot2'*. https://CRAN.R-project.org/package=ggcorrplot.

Lüdecke, Daniel, Mattan S. Ben-Shachar, Indrajeet Patil, Philip Waggoner, and Dominique Makowski. 2021. "performance: An R Package for Assessment, Comparison and Testing of Statistical Models." *Journal of Open Source Software* 6 (60): 3139. https://doi.org/10.21105/joss.03139.

Lüdecke, Daniel, Indrajeet Patil, Mattan S. Ben-Shachar, Brenton M. Wiernik, Philip Waggoner, and Dominique Makowski. 2021. "see: An R Package for Visualizing Statistical Models." *Journal of Open Source Software* 6 (64): 3393. https://doi.org/10.21105/joss.03393.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Ooms, Jeroen. 2022. *Rsvg: Render SVG Images into PDF, PNG, (Encapsulated) PostScript, or Bitmap Arrays*. https://CRAN.R-project.org/package=rsvg.

Peterson, Johnathan C., Kevin B. Smith, and John R. Hibbing. 2020. "Do People Really Become More Conservative as They Age?" *The Journal of Politics* 82 (2): 600–611. https://doi.org/10.1086/706889.

Pradel, Franziska. 2021. "Biased Representation of Politicians in Google and Wikipedia Search? The Joint Effect of Party Identity, Gender Identity and Elections." *Political Communication* 38 (4): 447–78. https://doi.org/10.1080/10584609.2020.1793846.

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

Salem, Hamza, and Fabian Stephany. 2021. "Wikipedia: A Challenger's Best Friend? Utilizing Information-Seeking Behaviour Patterns to Predict US Congressional Elections." *Information, Communication & Society* 0 (0): 1–27. https://doi.org/10.1080/1369118X.2021.1942953.

Sam, Cedric. 2021. "Canadian Election Results: Trudeau Wins Third Mandate—with Another Minority." https://www.bloomberg.com/graphics/2021-canada-election-live-results/.

Sjoberg, Daniel D., Karissa Whiting, Michael Curry, Jessica A. Lavery, and Joseph Larmarange. 2021. "Reproducible Summary Tables with the Gtsummary Package." *The R Journal* 13: 570–80. https://doi.org/10.32614/RJ-2021-053.

Slowikowski, Kamil. 2021. *Ggrepel: Automatically Position Non-Overlapping Text Labels with 'Ggplot2'*. https://CRAN.R-project.org/package=ggrepel.

Smith, Benjamin K., and Abel Gustafson. 2017. "Using Wikipedia to Predict Election Outcomes: Online Behavior as a Predictor of Voting." *Public Opinion Quarterly* 81 (3): 714–35. https://doi.org/10.1093/poq/nfx007.

Utych, Stephen M., and Cindy D. Kam. 2014. "Viability, Information Seeking, and Vote Choice." *The Journal of Politics* 76 (1): 152–66. https://doi.org/10.1017/S0022381613001126.

Wickham, Hadley. 2020. *Httr: Tools for Working with URLs and HTTP*. https://CRAN.R-project.org/package=httr.

———. 2021a. *Rvest: Easily Harvest (Scrape) Web Pages*. https://CRAN.R-project.org/package=rvest.

———. 2021b. *Tidyverse: Easily Install and Load the Tidyverse*. https://CRAN.R-project.org/package=tidyverse.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2021. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.

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

Wickham, Hadley, and Dana Seidel. 2020. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.

"Wikidata:history Query Service." 2020. Wikidata. https://www.wikidata.org/wiki/Wikidata:History_Query_Service.

"Wikimedia REST API." 2022. Wikimedia. https://wikimedia.org/api/rest_v1/#/.

"Wikipedia:about." 2022. https://en.wikipedia.org/wiki/Wikipedia:About.

Wilke, Claus O. 2020. *Cowplot: Streamlined Plot Theme and Plot Annotations for 'Ggplot2'*. https://CRAN.R-project.org/package=cowplot.

Wooldridge, Jeffrey M. 2018. *Introductory Econometrics: A Modern Approach*. 7th ed. Cengage.

Xie, Yihui. 2021a. *Bookdown: Authoring Books and Technical Documents with r Markdown*. https://CRAN.R-project.org/package=bookdown.

———. 2021b. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

———. 2021c. *Tinytex: Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents*. https://github.com/yihui/tinytex.

Yasseri, Taha, and Jonathan Bright. 2013. "Can Electoral Popularity Be Predicted Using Socially Generated Big Data?" *It - Information Technology* 56 (December). https://doi.org/10.1515/itit-2014-1046.

———. 2015. "Predicting Elections from Online Information Flows: Towards Theoretically Informed Models," May.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with Kable and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.