Title: Voting and vola**tea**lity: Explanation of electoral behavior in British elections 2015 and 2017

Author: Bogdan Romanov, Team A1

Link to the repository:
https://github.com/RomanovBogdan/IDS_project

Business understanding: The primary rationale for this project stems from the political science and electoral studies areas rather than business-oriented endeavours. So, I would assume that it would make more sense to project CRISP-DM structure onto the academic narrative.

Background: The necessary theoretical background for the project is that electoral volatility — "the net change within the electoral party system resulting from individual vote transfers" (Pedersen, 1979, p. 3) — is perceived as an indicator of democracy stability. Low volatility signalizes that decision-makers implement optimal public policies, which are welcomed by the population-electorate (Casal Bértoa et al., 2017). However, by contrast, high volatility might be an omen of institutional crisis, which the cabinet change can overcome.

Within this context, the United Kingdom constitutes a curios case, which was characterized by relatively low electoral volatility, especially at the turn of the XXth century, followed by a drastic, unexpected growth of volatility in years 2015 and 2017; this trend is illustrated in Figure 1.

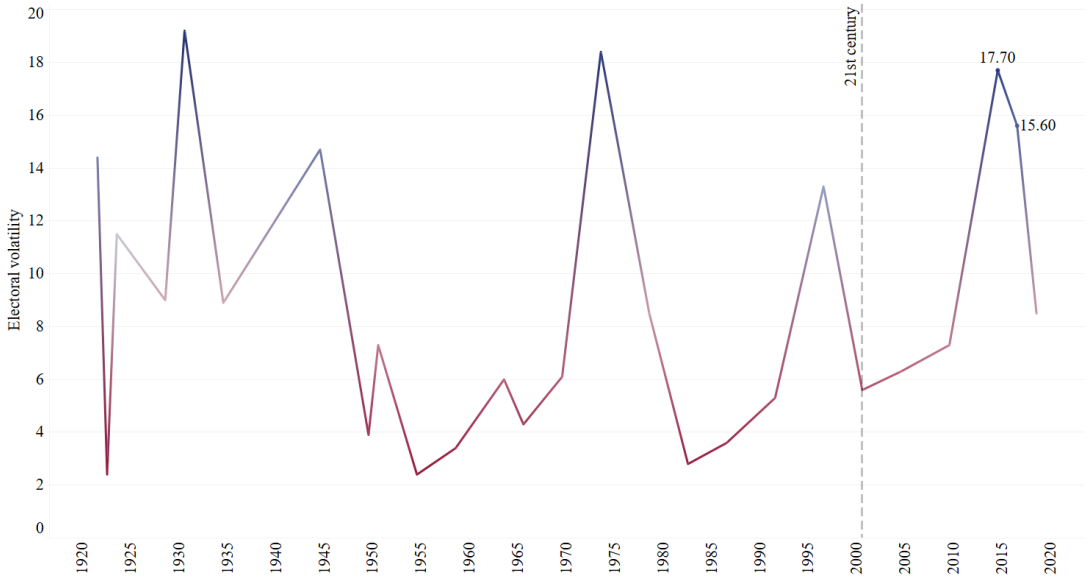

Fig. 1. The evolution of electoral volatility in the UK from 1918 to 2020
Source: Party Systems and Governments Observatory. (n.d.). Electoral Volatility. Who Governs Europe. Retrieved the 25th of November 2021, from https://whogoverns.eu/party-systems/electoral-volatility/

The current state-of-the-art does not have a substantial explanation of why the loyal electorate has turned away from their party; this project strives to elaborate on potential causal mechanisms.

Business goals: Thus, the "business" goal is to answer the research question: What mechanisms explain voting against the preferences observed during UK 2015, 2017 general elections? The answer to this question would shed light on and give us an insight into what could cause regime crisis manifested in high electoral volatility. What is more crucial, not only do we get the list of factors affecting the volatility, but we also will comprehend how this impact can be reversed to stabilize the democracy. Also, the project is another illustration that social sciences can be accompanied by data science techniques and still have this academia-centric scholar contribution without going all-in into the data science area.

There are four main hypotheses, which could be possible answers:

> *H0.* British voters have voted against their preferences due to the *strategic voting* — their former party had little success chances (Franklin et al., 1994, pp. 547–550);
>
> *H1.* British voters have voted against their preferences due to the *expressive voting* — the electorate does not lose interest in the former party altogether but tries to draw attention that the other party has more beneficial policies (Franklin et al., 1994, pp. 552–554);
>
> *H2.* British voters have voted against their preferences due to the *economic voting* — turning away because of ineffective economic development (Tavits, 2005);
>
> *H3.* British voters have voted against their preferences due to *external shock* — Brexit affected the secondary beliefs of the electorate (Dekalchuk, 2014; Sabatier & Weible, 2007)

Business success criteria: The number of correct hypotheses measures "business" success because these statements are not exclusive, they could be occurring simultaneously but to different extents.

Situation assessment: I am the only person working on the project; despite that, I have a sufficient understanding of data science, statistics, and political science. I cannot pinpoint any "legal and security obligations" and not causes of delay; all I can say is that project will be finished before the deadline of 13th of December. All crucial terms and notions were already explained in previous sections, but a more detailed description will follow in the full-scale project. Finally, since the project does not have a vivid business-oriented undertone, the cost-benefit analysis is not applicable.

Data-mining goals: since the main analytical tool used in the project is logistic regression, due to the binary nature of survey data, the lowest bar for the deliverables is four models, each devoted to different hypotheses. Additionally, basic visualization tools (e.g., bar charts, line charts) will also be presented in the project. Finally, I will refer to the notions of 'accuracy',

'recall', and 'precision' in the context of the model predicting the behavior of a voter — the trade-off between precision and recall will also be addressed and resolved with the use of threshold value.

Consequently, the data-mining success criteria are also straightforward: logistic regression model accuracy, the significance of variables, and Pseudo R-squared. The same approach is used for accuracy, recall, and precision — higher values would be preferable if we describe it in the simplest terms. Finally, about visualization, I cannot quantify them, but I would say that readability by the general audience is the desired outcome.

Gathering data: The project is devoted solely and exclusively to British elections, and, luckily, there is an agency, which conducts surveys in the United Kingdom from 2014 to 2020, i.e., British Election Study. As a result, I have access to several waves of these voter online surveys. The primary consideration behind the data selection was the proximity of data to the date of the elections. Since the general elections of 2015 were held on the 7th of May and elections of 2017 – on the 8th of June, I need waves 4[1] and 12[2], respectively. The former wave covers the March 2015 period (two months before the elections); the latter wave focuses on time from May to June 2017; hence, both waves are as close as possible to the election days[3].

Regarding the formats, British Election Study offers two options: SPSS file and STATA file. So, it does not matter which file will be imported into the Jupyter notebook, but I will use the .dta format used in STATA. Additional research and analysis of the dataset showed that this project does not require complementary datasets. All necessary operationalizations of variables mentioned in "business" understanding can be found in datasets for both waves.

Describing data: the structure of two datasets does vary to some extent. First of all, the number of variables in wave 4 is 660, while in wave 12 is almost two times less — 391. Secondly, regarding the number of observations, N for wave 4 is 31,545, and wave 12 has more observations — 34,394. Finally, about the format of variables, since the data is a result of surveys, which usually operate with nominal/categorical data, then the format is the same here: most of the variables have ordinal string values or still nominal ordinal, but discrete values, e.g., Likert scale from 1 to 10 for attitude; another type is date-format, which is used to track the duration of survey; also, there is continuous age and turnout. I might say that data format does not cause an issue, especially after the class on dummy-fication of categorical variables.

[1] "The March 2015 wave of the 2014-2018 British Election Study Internet Panel. Version 3.9."
[2] "2017 campaign wave of the 2014-2018 British Election Study Internet Panel. Update version 1.5."
[3] Both datasets are available after the registration on the website. Now they are downloaded and uploaded to the Github repository, I had to archive them because of the size limitations.

Exploring data: preliminary exploration allowed me to pinpoint that these variables can be used for operationalization of dependent and independent indicators:

Category	Concept	Operationalization
Dependent variable	The voter will cast his/her vote for party X	*generalElectionVote*
Independent variables	*H1.* Party has or does not have chances to win elections	*winConstituencyLab winConstituencyCon* (and similar variables for other parties) *noChanceCoalitionLab* *noChanceCoalitionCon* (and similar variables for other parties)
	H2. Other parties do have better policies implemented	*mii* *bestOnMII* (and similar variables for particular parties most important issues)
	H3. Economic policies are not effective enough	*econGenRetro* *changeEconomy* *econPersonalRetro*
	H4. Personal beliefs were affected by the crisis	*mii* *bestOnMII* (it seems that the survey does not have beliefs variables, but it is not a problem since we are interested not in beliefs per se, but in the effect of external shock — Brexit is on the list of most salient concerns)
Control variables		*country* *gender* *Age* (wave 4), *age* (wave 12), despite that every other variable has the same spelling capitalization of age is different across the waves

Verifying data quality: after data exploration, I can state that this data ideally fits in the project's research objectives: all necessary variables can be found in the dataset, the number of observations is relatively high, data formats does not prevent me from proceeding with the project. Also, the issue of missing values could be observed for some variables; however, these are very precise questions that are not included in the project's scope. In other words, the project does not face the issue of missing values.

Project plan: In order to answer the research question, the project has several tasks:

1. Due to the academic origin, the initial step within the project is to come up with a coherent theoretical framework, which would provide answers (hypotheses) to the research question — 20 hours
2. Once the theoretical framework is constituted, there is a necessity to conduct an operationalization of dependent and independent variables: which variables from the online surveys do capture the essence of measured, theoretical concepts — 5 hours
3. Recodification of selected dependent and independent variables into binary or categorical — 2 hours
4. Removing irrelevant levels of categorical variables, e.g., "Don't know", "No-none" — 1 hour
5. Modeling is more straightforward because the data is in binary/categorical format; thus, the most appropriate technique is logistic regression for each hypothesis — four models as a result — 15 hours
6. Add complementing visualizations to several steps of modeling and to the output results — 5 hours
7. The last technical task will be results evaluation with the appeal to accuracy, recall, and precision for particular models — 5 hours
8. The final theoretical task is to incorporate obtained results into the ongoing academic discussion — 3 hours

Overall workload: 60 hours

Side comment: Currently, I am not sure what the final output of the project would be. I am aware that only code, poster, and video are assessed, but I am also considering coming up with an article of some sort, which would encapsulate both video and poster-slides and could be circulated across the colleagues.

References:

1. Casal Bértoa, F., Deegan-Krause, K., & Haughton, T. (2017). The volatility of volatility: Measuring change in party vote shares. *Electoral Studies*, *50*, 142–156. https://doi.org/10.1016/j.electstud.2017.09.007
2. Dekalchuk, A. (2014). *Evolution of the Concept of "External Shock" in Various Political Course Traditions*.
3. Franklin, M., Niemi, R., & Whitten, G. (1994). The Two Faces of Tactical Voting. *British Journal of Political Science*, *24*(4), 549–557. https://doi.org/10.1017/S0007123400007006
4. Pedersen, M. N. (1979). THE DYNAMICS OF EUROPEAN PARTY SYSTEMS: CHANGING PATTERNS OF ELECTORAL VOLATILITY. *European Journal of Political Research*, *7*(1), 1–26. https://doi.org/10.1111/j.1475-6765.1979.tb01267.x
5. Sabatier, P. A., & Weible, M. (2007). *The Advocacy Coalition Framework*. 18.
6. Tavits, M. (2005). The Development of Stable Party Support: Electoral Dynamics in Post-Communist Europe. *American Journal of Political Science*, *49*(2), 283–298. https://doi.org/10.1111/j.0092-5853.2005.00123.x