Write up your work in a document of of approximately 8 pages. This should answer all the prompts and specific questions below, but also read as a single connected report. This report is for the client, not for a statistics professor. You may assume that your audience has a basic understanding of introductory statistics and knows what a regression is. Everything else must be explained.

The document should include all text and figures. Code should be integrated with R Markdown, and **hidden**. You must push your finished `.Rmd` to Github. This assignment is like HW 3: analyze the data and write a report which answers all the questions.

While there are wrong answers, there are many possible right answers. Any data analysis decisions or conclusions that you make should be justified and explained. Your job is to correctly ananlyze the data, not force the analysis to match a pre-conceived idea. Discuss whether or not your results match your hypothesis.

Data and research problem

Many people assume that violence, while perhaps dangerous or evil, is more *effective* politically than non-violence. In this homework, we will examine whether, in fact, non-violent political movements are more or less likely to achieve their goals than violent ones. Moreover, we will look at the conditions which make non-violence more or less likely to succeed.

Our data set, gathered by political scientists who have studied exactly these questions, is `navc.csv` in your repo. The units of analysis here are political movements or campaigns. For each movement, the data records:

- The name of the movement `campaign`;
- The country the movement was in `country`;
- The peak year of the movement's activity `year`;
- Whether the movement fully achieved its aims (1.0), achieved partial success (0.5), or failed (0) `outcome`;
- An indicator variable `nonviol`, 1 for non-violent movements and 0 for others;
- A quantitative measure of how democratic the government of the country was, from -10 for very un-democratic governments to a possible maximum of +10 `democracy`;
- An indicator for the government being under international sanctions `sanctions`;
- An indicator for whether the government received aid from other governments to help deal with the movement `aid`;
- An indicator for the movement's receiving aid from foreign governments `support`;
- An indicator for the government's using violence to repress the movement `viol.repress`;
- An indicator for whether substantial portions of the security (military and police) forces of the government sided with the movement `defect`;
- The duration of the movement, in days `duration`.

Specific analytic issues you must address In general, are non-violent movements more likely to be successful than violent ones? Does violent repression by the government make movements more or less likely to be successful, and is there a difference in this effect between movements which are themselves violent and non-violent? Similarly, what is the effect of foreign aid to the government and to the movement? Do non-violent movements become more likely to succeed as the government becomes more democratic? Does the difference in probability of success between violent and non-violent movements vary with how democratic the government is? (Hint: this question means that your model should probably interact a smoother of `democracy` with a factor for `nonviol`. To do this, use `s(democracy, by=nonviol)`.) All of these should be answered with reference to the results in your model (or models).

Models Use a generalized additive model with a logistic link function; smooth all continuous predictor variables, and include all categorical variables, except `campaign` and `country`, as your default. (Departures from this should be carefully justified.) Be sure to include the year as a predictor variable, and explain the interpretation of your estimated effects for the year. Some of the analytic issues above may be most easily addressed through including interaction terms, or through fitting different models on subsets of the data; describe any such variations, and the reasons for your choices.

Note 1: Before fitting a model with a logistic link function, you will need to re-code partial successes as either successes or failures. Explain which one you chose, and briefly justify your decision. Either is fine, but you must justify your choice.

Inferential statistics and model assessment You **CANNOT** assume that `R`'s default standard errors or p-values on estimated regression coefficients can be trusted. Uncertainty should be assessed using either the bootstrap or simulation procedures. (Be sure to explain why you used the procedure you did.) If you need to compare two models in terms of predictive accuracy, this should **NOT** be done through `R`'s default significance tests or R^2's or MSE, but through cross-validation. Exceptions will be made if you can successfully argue that the default calculations are reliable *for this problem.*

Model checking The answers you give to the substantive analytical questions rest on your estimated model, so you need to include some assessment of the model's goodness of fit. The exact way in which you do this is left up to your initiative; it may help to remember that the model is predicting probabilities of success. Be sure to describe your procedure and explain why you chose it, that is, why it is appropriate to answer the questions at hand. Possible things you can examine are prediction accuracy relative to an appropriate baseline, calibration, and residual plots.

Format

Your main report should have the following sections:

- **Introduction** describing the scientific problem and the data set, possibly including *relevant* summary statistics or exploratory graphs. Do NOT simply copy the text above as this will result in no credit.
- **Models** with subsections
  - Describing the specification of the model (or models) you estimated, and explaining why you decided to use those specifications rather than others;
  - Giving the relevant estimated coefficients and/or functions (possibly in visual form), along with suitable measures of uncertainty;
  - Checking the goodness of fit of the model, including a description of the test procedures you used, why you chose those ways of checking the model, what the results were, and what they told you about the ability of the model to describe the data set.
- **Results and conclusions** answering the analytical questions quantitatively, and with suitable measures of uncertainty, with reference to your estimated model or models.

You may assume that the reader has a general familiarity with the contents of 431, and with the models and methods we have covered so far in the course, but will need to be reminded of any details. The reader should not be assumed to have any prior familiarity with the data set.

Numerical results Numerical quantities should be written out to appropriate precision, i.e., neither more nor fewer significant digits than appropriate.

Grading rubric

Words (4 / 4) The text is laid out cleanly, with clear divisions and transitions between sections and subsections. The writing itself is well-organized, free of grammatical and other mechanical errors, divided into complete sentences logically grouped into paragraphs and sections, and easy to follow from the presumed level of knowledge.

Numbers (2 / 2) All numerical results or summaries are reported to suitable precision, and with appropriate measures of uncertainty attached when applicable.

Pictures (4 / 4) Figures and tables are easy to read, with informative captions, axis labels and legends, and are placed near the relevant pieces of text or referred to with convenient labels.

Code (5 / 5) The code is formatted and organized so that it is easy for others to read and understand. It is indented, commented, and uses meaningful names. It only includes computations which are actually needed to answer the analytical questions, and avoids redundancy. Code borrowed from the notes, from books, or from resources found online is explicitly acknowledged and sourced in the comments. Functions or procedures not directly taken from the notes have accompanying tests which check whether the code does what it is supposed to. The text of the report is free of intrusive blocks of code. With regards to R Markdown, all calculations are actually done in the file as it knits, and only relevant results are shown.

Analysis (10 / 10) Variables are examined individually and bivariately. Features/observations are discussed with appropriate figure or tables. The relevance of the EDA to the modeling is clearly explained. The model's formulation is clearly related to the substantive questions of interest. The model's assumptions are checked by means of appropriate diagnostic plots or formal tests; if the model is re-formulated, the changes are both well-motivated by the diagnostics, and still allow the model to answer the original substantive question. Limitations from un-fixable problems are clearly noted. The substantive questions about real estate pricing are answered as precisely as the data and the model allow. The chain of reasoning from estimation results about models, or derived quantities, to substantive con- clusions is both clear and convincing. Contingent answers ("if X, then Y, but if Z, then W") are likewise described as warranted by the model and data. If uncertainties in the data and model mean the answers to some questions must be imprecise, this too is reflected in the discussion.

Extra credit (0 / 0) Up to five points may be awarded for reports which are unusually well-written, where the code is unusually elegant, where the analytical methods are unusually insightful, or where the analysis goes beyond the required set of analytical questions.