Rubric

As of July 12, 2021 we're still finishing the fine details of this rubric. The exact wording and levels might change a some before we finalize (within the next few days). We're not going to make substantial changes, so you should feel empowered as a team to start working on your project.

Introduction (5 levels)

Is the introduction clear? Is the research question specific and well defined? Does the introduction motivate a specific concept to be measured and explain how it will be operationalized. Does it do a good job of preparing the reader to understand the model specifications?

5. An introduction that is scored in the top level has very successfully made the case for the reader. It will have placed the specific question in context and clearly defined the concepts that are under investigation. The introduction will have connected the problem with the statistics and data that will be used, and have generally created a compelling case for continuing to read the report.

6. An introduction that scores in the fourth level has done a significant amount of successful work. The reader can place the topic area and the research question and can understand and appreciate why this is a question that demands a data based answer. The concepts that are under investigation should be free of ambiguity. Keeping this introduction from a full, four-point score might be: small issues with writing, some lack of continuity in the argument, lack of clarity in concept definition or some other such small issue that keeps this from being a wholly convincing introduction.

7. An introduction that scores in the third level has done some successful work. The reader can place the topic area and question. Often times, introductions in this level identify the broad issue at hand, but it might be too sweeping and fail to connect the broad issue to the specific research question. Or, specific aspects of the research question might not be fully developed. The concepts under investigation might not be clearly defined, or might have multiple facets that are not explored or not defined.

8. An introduction that scores in the second level has made some attempt to describe the problem and its importance, but it is missing significant parts of this argument. The language might be poorly chosen, the argument loosely formed, or the context or concepts left unexplained; there is considerable room for improvement.

9. An introduction that is scoring at the first level has given a very cursory treatment for the setting of the question and why an answer based on statistics is necessary. To the extent that concepts are identified for investigation, they are unclear, unspecific, unscientific, or contentious.

Operationalization (4 levels)

Operationalization is the process whereby the team argues for a specific measurement of their concept.

4. A report that is scored in the top level on operationalization will have done a wholly adequate job connecting the concepts from the introduction to a set of clearly, and precisely defined variables and measures. From concepts, the report will have identified data and

variables that clearly do a good job of measuring these concepts. The argument will be clear, concise, and complete, potentially including a statement of the data or variables that the team chose not to use, and the justifications for these choices.

5. A report that is scored in the third level on operationalization will have done good work connecting the concepts from the introduction to a set of well-defined variables. Keeping this report from full marks might be some imprecision in argument moving from concept to measurement. A report at this level may have made reasonable choices, but has not provided clear justification for those choices.

6. A report that scored in the second level of operationalization will have attempted to connect the concepts from the introduction to measures in the data. But, a report at this level will have run into significant challenges. The choice of variables used by a team might not match the concept: there may be better variables available, or choices made might be given only a cursory justification or explanation. To be clear, finding a perfect measure might not be possible, but a report should make a clear argument that justifies whatever measure it chooses.

7. A report that is scored in the first level operationalization will have failed to connect the concepts that have been identified in the introduction to the data in any meaningful way. As a consequence, it is not possible for the reader to know that any analysis that comes from this report has any bearing on the question at hand.

Data Wrangling (3 levels)

There are several steps to working with data for this report: downloading and procuring data; reshaping and joining disparate data sources; engineering or building variables; and, finally, the statistical analysis itself. The code for each of these steps should be well organized. This maximizes readability of code that you are writing, and it also maximizes your team's ability to contribute and debug problems when they arise. We would probably separate each step into its own file (for example `download_data.R`, `clean_cdc_data.R`, `clean_airbnb_data.R`, and `merge_data.R`) but there are other models that you could choose provided the organization is clear.

3. A report that is scored in the top level on data wrangling will have succeeded to produce a modern, legible data pipeline from raw data to data for analysis. Because there are many pieces of data that are being marshaled, the reports that score in the top level will have refactored different steps in the data handling into separate files, functions, or other units. It should be clear what, and how, any additional features are derived from this data. The analysis avoid defining additional data frames when a single source of truth data would suffice.

4. A report that scores in the second level data wrangling will have tried, but not fully achieved the aims for a modern, legible data pipeline from data to analysis. None of the problems cast doubt on the results, but might mean that it would be difficult to contribute to this project in the future, difficult to read this analysis for the present reader, or some other such flaw. At this level, data wrangling can still have issues that keep it from being professional-grade, but these are the types of issues that might be expected for early-programmers: variable names might be communicative, but clumsy; pipelines might work, but be inefficient, etc. This level of data wrangling might have several version of the data that do not maintain referential integrity (in the case of this data, an example of a problem is writing code based on column position rather than column names), or it might have several versions of derived data from the source data (i.e. `anes_data_republicans`, `anes_data_democrats`) that could generate issues with a future pipeline.

5. A report that scores in the first level on data wrangling has significant issues in how the data has been prepared for analysis.

Data Understanding (3 levels)

In order for a reader to understand or ascribe meaning to your results, they need to understand enough about the data that they can place your results in context. Remember: at the time that you present an analysis, you and the team know more about your data than anybody else. You will have to communicate what the data is, how it was generated, and also communicate how specific variables that you are using in your analysis are distributed. This can be done with by referencing tables, figures, and summary statistics in the narrative. You might ask yourselves, "Overall, does the report demonstrate a thorough understanding of the data? Does the report convey this understand to its reader – can the reader, through reading this report, come to the same understanding that the team has come to?"

3. A report that is scored in the top level will describe the provenance of the data; the audience will know the source of the data, the method used to collect the data, the units of observation of the data, and the distributions of important features of the data that will be used in the analysis. The report will identify anomalies, censored scales, artifacts of binning, and prominent clusters.
4. A report that is scored in the second level will leave the reader with a good understanding of the data distribution. Keeping the report from a perfect score might be a failure to comment on some feature of the data that would help the reader to contextualize the results.
5. A report that is scored in the first level will leave the reader with an insufficient understanding of the distribution to fully contextualize the results. A report that includes an output dump (a plot or R output that is not discussed in the text) will also score in this level.

Presenting Evidence (2 levels)

In a data science report, every argument that you make should be supported by evidence from your data. This argument should be compelling – you should choose the most effective data to make your point – and it should be parsimonious – you should not provide data as evidence that does not clearly move your argument forward.

2. Every single plot or output included in the report was discussed as a narrative point in the prose of the report. Every claim based on data that was made in the report was supported by evidence in the form of a table, figure, or supporting statistic.
3. There are claims made in the report that could have been supported with data, but were not. Or, there are plots, tables, or statistics presented in the report that are superfluous or do not push the argument forward.

Visual Design (5 levels)

5. A report that is scored in the top level will include plots that effectively transmit information, engage the reader's interest, maximize usability, and follow best practices of data visualization. Titles and labels will be informative and written in plain english, avoiding variable names or other artifacts of R code. Plots will have a good ratio of information to space or information to ink; a large or complicated plot will not be used when simple plot or table would show the same information more directly. Axis limits will be chosen to minimize visual distortion and avoid misleading the viewer. Plots will be free of visual artifacts created by binning. Colors and line types will be chosen to reinforce the meanings of variable levels and with thought given to accessibility for the visually-impaired.

6. A report that is scored in the fourth level will include plots that effectively transmit information, do not present major obstacles to usability, and follow best practices for data visualization. Keeping the report from top score might be some distracting visual artifact, axis labels that do not line up properly with histogram bars, poorly chosen bin widths, a legend that conceals important information, or some other aspect in which the plot might be improved.

7. A report that is scored in the third level will include plots that contain important information and can be effectively used by most readers. Keeping the report from a top score might be an instance in which a complicated or large plot is used when a compact plot or table would display the same information more effectively, redundant plots that substantially overlap in the information they show, or a moderate problem with axes, binning, labels, or colors.

8. A report that scores in the second level will have chosen a plot that communicates a point to the reader, but that could be more convincing in its effect. The plot may not have multiple issues that interfere with usability; it may be poorly titled and labeled; or the choice of presentation may not have fully communicated the pattern that the team wants to make.

9. A report that is scored in the first level will have serious issues with its visual design.

Model Specification (5 levels)

5. A report that is scored in the top level will have chosen a set of regression models that strongly support the goal of the study. The concept to be measured will be represented in a key coefficient(s) in every model. Variables will have appropriate transformations applied. Ordinal variables will not be treated as metric. If the goal is explanation, variables will be chosen considering where they appear in the causal graph: proxies will be chosen for important omitted variables if possible, outcome variables / mediators will not be included on the right hand side. Each specification will be a sound, defensible model, and the specifications will be different from each other, outlining the space of reasonable modeling choices.

6. A report that is scored in this level will have chosen a set of regression models that are aligned with the goal of the study. The concept to be measured will be represented in a key coefficient(s) in every model. Keeping these models from full points might be a questionable choice of a variable transformation, a predictor that can be regarded as an outcome variable, or too much similarity between two specifications.

7. A report that is scored in the third level will have chosen a set of regression models that are mostly aligned with the goal of the study. The models will be largely defensible, even if some functional forms are not perfectly chosen, or if one model has a few too many variables. For an explanatory model, variables will be chosen considering where they appear in the causal graph, and the purpose of each variable will be clear, even if there is some question about whether the variable is an outcome or not.

8. A report that is scored in the second level will have a set of regression models that fall short of supporting the goal of the study in some minor way,

9. A report that is scored in the first level will have a set of regression models that fall short of supporting the goal of the study in some major way.

Model Results and Interpretation (5 levels)

5. A report that scores in the top level will have interpreted the results of the model appropriately, drawn a conclusion about statistical significance; placed these results in context with some statement about practical significance that is compelling to the reader; and will have done so in a way that is clear, concise, and correct.
6. A report that scores in the fourth level will have interpreted the results of the model appropriately, drawn a conclusion about statistical significance; placed these results in context with some statement about practical significance that is compelling to the reader. Keeping this from full points might be some lack of clarity or concision; or some very slight error in the interpretation of the model.
7. A report that scores in the third level will have done good work to interpret the results, but the explanation of practical significance will not fully communicate the importance of the results to the reader. The argument might use a complicated effect size measure when a simpler or more understandable one is available. While the statistics might not be incorrect, they are not making a wholly compelling argument.
8. A report that scores in the second level will have some considerable errors in the model interpretation. Either the results will be misunderstood – statistically significant results might be interpreted as non-significant for example – or will have failed to successfully connect the results of the model with an interpretation of what these results mean.
9. A report that scores in the first level will have very serious issues with how the model results are interpreted. They may be wrong; non-existent; mis-characterized; or some other such very serious issue.

Assumptions of the Model (4 levels)

Has the team presented a sober assessment of the CLM assumption that might be problematic for their model? Have they presented their analysis about the consequences of these problems (including random sampling) for the models they estimate? Did they use visual tools or statistical tests, as appropriate? Did they respond appropriately to any violations?

4. A report that scores in the top-level has provided an unassailable assessment of the assumptions of the model. This does not mean that the model necessarily satisfies all assumptions, but that where it does not the reader is promptly notified of the issue, its consequences for the analysis, and potential strategies to remedy these consequences.
5. A report that scores in the third-level has provided a fair assessment of the assumptions of the model. The team may have failed to satisfy a skeptical reader who: a skeptic might raise concerns that have not been addressed, or might reasonably disagree with the team's assessment of an assumption violation.
6. A report that scores in the second-level has provided some work to assess the modeling assumptions, but a skeptical reading might quickly point out flaws in the statistical reasoning, or raise issues that the team has not addressed.
7. A report that scores in the bottom-level has provided a cursory treatment of modeling assumptions. Some assessments might be incorrect or non-existent. Reasonable questions from a skeptical reader might be left unaddressed.

Omitted Variables Bias (3 levels)

Did the report miss any important sources of omitted variable bias? Are the estimated directions of bias correct? Was their explanation clear? Is the discussion connected to whether the key effects are real or whether they may be solely an artifact of omitted variable bias?

3. A report that scores in the top-level has thought carefully about omitted variables in a causal model; their thinking is done in a way that would satisfy a skeptical reader. It need not propose to measure all omitted featured, but after the discussion of omitted variables even a skeptical reader should acknowledge that the proposed cause-and-effect relationship is plausible.
4. A report that scores in the second-level has engaged with the concept of omitted variables, but has not fully addressed the topic. A skeptical reader might propose reasonable features that the team has not considered that would confound the proposed causal relationship.
5. A report that scores in the first-level has engaged with the concept of omitted variables, but with some gaps in the argument. There may be some problem in an argument for the direction of bias. A skeptical reader might pose a fundamental problem for the cause-and-effect relationship that the report has failed to address.

Conclusion and Impacts (3 points)

3. A top-level conclusion will summarize the team's work, provide the reader with a reminder of the initial question, and will provide recommendations *based on the analysis* that are reasonable generalizations of the narrow statistical modeling work that the team has undertaken. Recommendations made in this section will identify the evidence from the reports' analysis, and will also identify limitations or context that shapes in interpretation of these recommendations.
4. A second-level conclusion might too simply restate the introduction of the report. It might not be as closely-related to the analysis as it should be. A report with a conclusion that scores in the second-level is better than a report that does not have a conclusion, but it could be strengthed to contribute more.
5. A first-level conclusion does not make much, if any, contribution to the report overall. This conclusion likely is a rote restatement of the introduction; or, it does not relate to the analysis conducted in the report. A report with an introduction ranked in the first level would most likely be just as strong without this conclusion.

Overall Effect (3 levels)

3. A report that scores in the top level will have met expectations for professionalism in data-based writing, reasoning and argument for this point in the course. It can be presented, as is, to another student in the course, and that student could read, interpret and take away the aims, intents, and conclusions of the report.
4. A report that scores in the second level will be largely professional; largely clearly written and structured; but will have some problem that inhibits the reader from being able to clearly reason from the report.
5. A report that scores in the first level will have significant issues in its presentation. These could be spelling, language, argument, formatting, or other issues that cause problems for the reader in their ability to read, evaluate, and take action on what is reported.