**Phase IV and V Rubric**

| **Category** | **Less developed projects** | **Typical projects** | **More developed projects** |
| --- | --- | --- | --- |
| **Introduction** | Less focused and organized. They may jump to technical details without explaining why results are important.  Research questions are not clearly stated and/or results are not clearly summarized at the end of the introduction. | Provides background information and context.  Introduces key terms and data sources.  Outlines research question(s).  Ends with a brief summary of findings. | All expectations of typical projects +  clearly describes why the setting is important and what is at stake in the results of the analysis. Even if the reader doesn’t know much about the subject, they know why they care about the results of your analysis. |
| **Data description** | Simple description of some aspects of the dataset, little consideration for sources.  The description is missing answers to applicable questions detailed in the "Datasheets for Datasets" paper. | Answers all relevant questions (see below) in the "Datasheets for Datasets" paper. | All expectations of typical projects + credits and values data sources. |
| **Preregistration statement** | The preregistered analyses could be performed using the data collected, but it is not clear how they fit in the context of the real-world application from which the data originated. | The preregistered analyses are contextualized by the real-world application to a certain degree. The analyses are not described in a way that persuades the reader their results would be interesting, whether or not they turn out to be statistically significant. | The preregistered analyses reflect deep and critical thinking about the real-world application from which the data originates. The analyses are described in a way that persuades the reader that their results will be interesting, whether or not they turn out to be statistically significant. |
| **Data analysis** | Code closely matches examples from class, and does not go much further.  Analyses selected are not clearly purposeful.  Preregistered analyses are not presented. | Code goes further than the examples presented in class.  Analyses selected are purposeful and further the data narrative, but questions raised are not adequately addressed.  Preregistered analyses are presented. | All expectations of typical projects + analyses are carefully selected to answer all reasonable questions.  Questions raised by one analysis are addressed in subsequent analyses. |
| **Evaluation of significance** | Metrics of statistical significance are present, but not interpreted for the reader and/or relevant to the analysis performed. | Metrics of statistical significance appropriate to the analysis performed are presented and are interpreted to some degree for the reader. | Metrics of statistical significance appropriate to the analysis performed are presented and clearly interpreted for the reader. Limitations of significance metrics are acknowledged. |
| **Interpretation and conclusions** | Results are presented as numeric values and plots, with little to no written discussion.  Values are printed out of context, with no/few labels. | Values are interpreted in a way that is clear and addresses what the values mean and explain to some extent why they are important.  Values are printed with clear labels. | Interprets numeric values in a way that supports a clear story and conclusion creatively ties analysis together to present the results of the analysis through a well-written discussion.  Values are presented in context and with clear labels. |
| **Limitations** | The limitations are not explained in depth.  There is no mention of how these limitations may affect the meaning of results. | Identifies potential harms and data gaps, and describes how these could affect the meaning of results. | Creatively identifies potential harms and data gaps, and describes how these could affect the meaning of results, as well as the impact of results on people. |
| **Writing** | May have spelling and grammatical errors, or awkward or incomplete sentences, indicating that they were written in haste without editing. | Language will be polished and free from errors *(Note: if your group does not include a native English speaker, make a note of that in your submission)*. | Writing is clear and complicated ideas are presented such that they are immediately understandable. |
| **Organization and focus** | Work appears to have been done independently by team members and then merged at the last moment.  Analyses may be exhaustive but carry little meaning or interpretation.  There is not a very clear story throughout the entire report. | Most elements of the project are clear and provide a connected conclusion. Some parts could have been removed to make the report more focused.  There is a clear story that flows throughout most of the report. | All elements of the project support a clear and connected conclusion. Every part is essential and cohesive.  There is a clear story to the entire report that flows throughout. |

Priority questions from Datasheets for Datasets:

* What are the observations (rows) and the attributes (columns)?
* Why was this dataset created?
* Who funded the creation of the dataset?
* What processes might have influenced what data was observed and recorded and what was not?
* What preprocessing was done, and how did the data come to be in the form that you are using?
* If people are involved, were they aware of the data collection and if so, what purpose did they expect the data to be used for?
* Where can your raw source data be found, if applicable? Provide a link to the raw data (hosted on Github, in a [Cornell Google Drive](https://support.google.com/drive/answer/2494822?co=GENIE.Platform%3DDesktop&hl=en) or [Cornell Box](https://it.cornell.edu/box)).