**A. Data Collection**

* **A.1 Informed consent**: If there are human subjects, have they given informed consent, where subjects affirmatively opt-in and have a clear understanding of the data uses to which they consent?
  + We can assume that if people took this survey, they gave informed consent. (The onus goes onto the original publisher.)
* **A.2 Collection bias**: Have we considered sources of bias that could be introduced during data collection and survey design and taken steps to mitigate those?
  + Since this is a LGBTQ+ survey, cis straight people have been intentionally excluded from this survey. This bias is intentional.
  + Identity and country wise, the data is very balanced minimizing bias within minority groups.
  + Each question has language that introduces bias, but that is the main goal of this research so bias here does not need to be mitigated, but instead expressed intentionally.
* **A.3 Limit PII exposure**: Have we considered ways to minimize exposure of personally identifiable information (PII) for example through anonymization or not collecting information that isn't relevant for analysis?
  + No personally identifiable information was present in the public dataset.

**B. Data Storage**

* **B.1 Data security**: Do we have a plan to protect and secure data (e.g., encryption at rest and in transit, access controls on internal users and third parties, access logs, and up-to-date software)?
  + The data was posted publicly on Kaggle, so there is no need for data security.
* **B.2 Right to be forgotten**: Do we have a mechanism through which an individual can request their personal information be removed?
  + We will follow the original publishers’ actions. If this dataset is taken down, we will also delete the dataset from our devices when we notice the dataset is not present.
* **B.3 Data retention plan**: Is there a schedule or plan to delete the data after it is no longer needed?
  + We will delete the data after this project to respect people’s right to be forgotten.

**C. Analysis**

* **C.1 Missing perspectives**: Have we sought to address blindspots in the analysis through engagement with relevant stakeholders (e.g., checking assumptions and discussing implications with affected communities and subject matter experts)?
  + It is out of the scope of this project to find further research or metrics to support our results.
  + We will be including all the surveys as to not exclude any group unless specified by a research question.
* **C.2 Dataset bias**: Have we examined the data for possible sources of bias and taken steps to mitigate or address these biases (e.g., stereotype perpetuation, confirmation bias, imbalanced classes, or omitted confounding variables)?
  + Although the dataset has balanced features, this is a European-centric dataset. Most of the world is excluded from this dataset.
  + Additionally, language used in this dataset is going to be European-centric.
* **C.3 Honest representation**: Are our visualizations, summary statistics, and reports designed to honestly represent the underlying data?
  + Visualizations for before and after preprocessing will be presenting to be transparent.
  + Peer review will help clarify any discrepancies within the visualizations.
* **C.4 Privacy in analysis**: Have we ensured that data with PII are not used or displayed unless necessary for the analysis?
  + We do not have PII to publish.
* **C.5 Auditability**: Is the process of generating the analysis well documented and reproducible if we discover issues in the future?
  + We will be posting the code to GitHub in case anyone else wants to investigate our analysis.

**D. Modeling**

* **D.1 Proxy discrimination**: Have we ensured that the model does not rely on variables or proxies for variables that are unfairly discriminatory?
  + Yes, we are intentionally limiting the scope of this project to the LGBTQ+ community.
* **D.2 Fairness across groups**: Have we tested model results for fairness with respect to different affected groups (e.g., tested for disparate error rates)?
  + The dataset is unusually balanced for a real world dataset, so the fairness for different groups should be minimized.
  + We are limited to looking at a European-centric scope of groups.
* **D.3 Metric selection**: Have we considered the effects of optimizing for our defined metrics and considered additional metrics?
  + Optimizations made in this model may affect groups of people differently, but overall it should not discriminate in any way that is harmful to any specific groups.
* **D.4 Explainability**: Can we explain in understandable terms a decision the model made in cases where a justification is needed?
  + Using NMF and PCA, it will be possible to explain the decisions of the model when looking at cluster personas.
* **D.5 Communicate bias**: Have we communicated the shortcomings, limitations, and biases of the model to relevant stakeholders in ways that can be generally understood?
  + Since this project will be presented in class, all stakeholders will have an advanced understanding of the dataset which should minimize assumptions.
  + Additional peer review will help avoid leading peers to inaccurate conclusions.