My motivation for this project stems from my interest in ethical considerations in mental health research using data science. This project also incorporates my passion for art in a way that encourages community engagement, self-advocacy, and data literacy. Lastly, I believe greater awareness among administrators, faculty, and students about the current climate for the disabled students' community is much needed. This project combines my interests by analyzing if disabled students feel as if they are effectively represented in datasets generated from university surveys. This project will also explore if students resonate with the topics of mental health and wellbeing presented through university survey questions. Students' responses and perspectives on these topics will be portrayed in a community created data mural, which is a mural that visualizes data. Ultimately, my project aims to highlight how we should bring the subjects of campus surveys to the frontlines of data collection and analysis. I believe that this approach may help create datasets that better capture the individual's identity and mental self, which in turn allows for research that better serves the community.

The two datasets I will be using in my project were generated from surveys conducted by the Student Experience in the Research University (SERU) Consortium and the University of California Undergraduate Experience Survey (UCUES). The SERU Consortium survey was conducted from May to June 2020 across nine research universities. It gauged the climate of students' lives and health during the pandemic. In particular, I would like to focus on the SERU report, "The Experiences of Undergraduate Students with Physical, Learning, Neurodevelopmental, and Cognitive Disabilities During the Pandemic" in order to evaluate what questions these surveys should ask to accurately represent the disabled students community and to inform the wider UC Berkeley community of the disabled students' circumstances. I will also be relying on 2020 data from the University of California Undergraduate Experience Survey

(UCUES). This is a biennial survey conducted across the nine UC campuses. I will hone in on questions under the following sections of the survey: Satisfaction, Academic Experience and Globalization, and Campus Climate for Diversity. The survey questions from both the SERU Consortium survey and the UCUES survey that I will be focusing on revolve around topics of student mental health and wellbeing on campus.

I will also be relying on qualitative data gathered from focus groups that consist of members from the disabled students community on campus. Students will engage with the data visualizations from the SERU survey and with visualizations I create from UCUES datasets. They will be asked questions that highlight how they may feel about their representation in the datasets, how effective the data visualizations are in describing their experiences, and what they think is missing from the data presented. A transcript of their responses and demographics will be used in the analysis portion of the project.

Since I will be using the visualizations provided by the SERU Consortium survey, there will be no data processing with that survey data. However, I will be processing the UCUES dataset so that I could create effective and compelling visualizations from the given information. The data processing techniques that I will be using include changing numerical data values to their categorical counterparts (e.g. changing the value 1 to the categorical value it represents, “Strongly Disagree”), creating an additional column that specifies different demographic variables from the survey (first generation status, race/ethnicity, gender, discipline, and student level), and merging the datasets based on questions from the survey in order to look for potential relationships between different questions or demographic variables.

In addition, I will be using natural language processing (NLP) techniques to analyze students' thoughts and feelings in response to the UCUES and SERU data visualizations. The NLP methods I will focus on are topic modeling, Word2vec, and sentiment analysis. I will be using the most common topic modeling algorithm, Latent Dirichlet Allocation (LDA), which will return a series of word-probability pairs that quantify how well a word captures the topic of a student's response. The model's input will be the collection of students' transcribed responses and the estimated number of probable topics--I will try different amounts of topics to determine how many topics are representative of students' responses. I will also be using the Word2vec algorithm which will return a list of words for each interview question ordered by which words are predicted to best capture the overall sentiment of the question topic. Finally, I will be performing a sentiment analysis using the VADER (Valence Aware Dictionary for Sentiment Reasoning) model in order to gauge students' emotions and sentiment regarding the data visualizations presented to them. These three techniques will help me understand how students feel about their representation in these campus surveys. They will also provide some insight into how much students resonate with the data on student mental health and wellbeing.

Based on these methods, there is not an obvious outcome or model output since the dataset created from the student focus groups is dependent on the students' opinions. However, I anticipate that there will be relationships between different demographic variables and feelings of belonging and stress in regards to the UCUES dataset. In particular, there may be a relationship between disciplines and levels of stress due to the competitive nature of STEM majors. There may also be a relationship between first-generation students, levels of stress, and feelings of belonging since students may not know what to expect when starting college. These

proposed outcomes are influenced by my personal experiences, therefore, the actual results may vary.

It is also important to consider the potential biases involved in my research project. Possible blind spots may likely relate to the datasets' particular focus on students' experience during the 2020 school year, which was during the height of the pandemic. The degree to which students may resonate with data presented to them would depend on how students experience life on campus over a year into the pandemic. Additional bias may be introduced due to the size of the focus groups; the participants' responses may not be reflective of the wider disabled students community. Another ethical concern to consider is participants' feelings towards sharing their responses to the datasets; their responses may expose potentially sensitive information, such as the state of their mental health and wellbeing and their particular disability. Before participating in the focus group, students will be told what questions they will be asked and how their data would be used if given their approval.

Ultimately, this project is relevant to research and student life since it addresses areas for improvement in current research methods and highlights a way that students can bring attention to their needs and concerns. Students' responses to campus surveys' visualizations may inform campus survey creators on what questions may be more representative of the disabled student population. Participants' feedback may also give more insight into what visualizations are most effective, which researchers and campus administrators could take into account when creating visualizations for the wider community. The results of this project may help pave the way for wider use of community-based participatory research when conducting campus climate surveys/research. This project may also inspire greater use of art-based methods and data literacy techniques to communicate research findings and encourage community engagement.