- What file formats will you be gathering or creating for this project?

The main file format are the csv files that the datasets come in. Other files will be .ipynb jupyter notebook where the code is and a keynote file where I will have my presentation. I will also have a readme text file inside my jupyter notebook where other users can read my code documentation for replication and analysis.

- How are you storing your data?

I have my local files stored under a Data folder and a Code folder. I do not have screenshots of my data source since they are from a UC website. I also saved my code and csv files in the iSchool's Box. For further backup, I have my code and datasets in a public repository on GitHub: https://github.com/slieu2/Open-Mashup. I've already experience the surprise of not being to open my files in the iSchool's box storage. So far, I have three places where I'm storing my data and code. As a fourth measure, I'm going to email to myself these files.

- How are you preserving and protecting all data values from changes, either accidental or purposeful?

Since these are csv files, I'm reading them in as data frames in pandas. I have not made changes to the original files and I don't anticipate that I will. In case I have to make changes to the original csv file, I'll rename them with "_altered" and keep both the original and the altered files I'll be working off of.

- What will your final data file look like? Describe it in terms of file format, structure, size, and other details. This won't be a perfectly certain answer, but you should provide as many details as you can think of. Provide ranges or options for areas where you are uncertain.

My final data csv file will contain the following possible columns:

county_name

fall_term or year

high_school

application_GPA

Total Population per county

% Total Population: White Alone

% Total Population: Hispanic or Latino
% Total Population: Black or African American Alone

% Total Population: American Indian and Alaska Native Alone

% Total Population: Asian Alone

% Total Population: Native Hawaiian and Other Pacific Islander Alone

% Total Population: Some Other Race Alone

% Total Population: Two or More Races

% Population 25 Years and Over: Less than High School

% Population 25 Years and Over: High School Graduate or More (Includes Equivalency)

% Population 25 Years and Over: Some College or More

% Population 25 Years and Over: Bachelor's Degree or More

% Population 25 Years and Over: Master's Degree or More

% Population 25 Years and Over: Professional School Degree or More

% Population 25 Years and Over: Doctorate Degree

% Civilian Population 16 to 19 Years: Not High School Graduate, Not Enrolled (Dropped Out)

% Civilian Population 16 to 19 Years: High School Graduate, or Enrolled (in School)

Median Household Income (In 2017 Inflation Adjusted Dollars)

% Population Under 18 Years of Age for Whom Poverty Status Is Determined: Living in Poverty

% Population Under 18 Years of Age for Whom Poverty Status Is Determined: At or Above Poverty Level

Households

% Households with Housing Costs more than 30% of Income.

The difficulty lies in the breakdown of multiple high schools per county and their corresponding GPAs. I want to capture the overall GPA per county to compare them with the demographics information to see if highly educated and/or wealthy counties have higher GPAs compared to poorer counties. Given there are 58 counties in California, and many high school in each county, I need to figure out what kind of join I can do on the data. I'm thinking a left outer join or a right outer join, but I'm not there yet, code-wise.

I also have the percentage of each ethnicities of incoming freshmen by year, but it's not correlated to county nor ethnicity per UC campus. I would like to see how ethnically representative each university based on the county in which they are situated.

- How do you plan to disseminate the data once the project is complete?

Since all of my datasets are public, I have them on a GitHub repository.

- Describe any hand edits, curation, cleaning, or other alterations that are needed for the data. Explain your systems in place for ensuring that the original data values are recoverable and your changes reviewable.

All original datasets are kept in their original forms as they were downloaded: as csv, excel or text files. I will convert them to csv files when I start to read them in jupyter notebook.

The original demographics file contains a lot of unwanted columns and an unnecessary first row. I'll also have to remove the "County, California" suffix for each row of the the county_name column as I will use the county name to join to the other GPA dataset.

I have a pell grant amounts by university that I will need to integrate by that data set is by state and city, not by county. Fortunately, I'm only looking at nine campuses so I'm thinking the integration won't be too painful. Perhaps, once I extract the nine UC campuses from this dataset, I can then "add" a county column for the join.

- Describe the stages of the project (that you determined from your workflows exercise from class) and the associated due dates you anticipate for them.

Week 8 - 1) Data and project management documentation due.

2) Extract pieces of data from csv files for remaining data sets.

3) Create GitHub repo.

Week 9 - Filter out columns and start to compile the big csv file that I will use for data analysis.

Week 10 - Write reproducible notebook.

Week 11 - Data documentation assessment

Week 12 - Turn in final dataset design.

Week 13 - Midpoint check-in. GitHub due.

Week 14- Reproducible notebook draft due as integrated in Jupyter Notebook.

Week 15 - Conference talk proposal

Week 16 - Submission of completed project.