Hello,

As an initial disclaimer, I tried to preserve observations in the dataset by manipulating those observations that were initially outside the scope of the mapped dataset requirements, but some observations did have to be dropped as I could not find a way to retain them, but still meet the mapped dataset requirements.

The following transformations were made to this dataset in order to map the existing fields and contained values to the required fields and values in the database:

1. Take out titles, suffixes, quoted middle names
   1. Compress cutoff first names with periods
   2. Hyphenate middle names with first names or spaced last names
2. Split the name column into a first\_name and last\_name column
3. Undo hyphenation to split the name column
4. Drop the name column as it does not exist in the mapped dataset
5. Drop the applied column as it does not exist in the mapped dataset
6. Reposition the first\_name and last\_name columns from the end of the dataset to its position in the mapped dataset
7. Rename columns to match the mapped dataset
8. There was 1 observation in the last\_name that did not have a value, so I manually set that value to “NoLastName” to preserve the observation
9. I filled in the missing inquiry\_date observations with the epoch date since snaplogic uses JavaScript to manipulate the data and epoch time is what JavaScript bases it datetime calculations off.
10. Converted date time formats to match mapped data set format requirements
11. Replaced the values in the application\_status\_override column to match the mapped dataset required values
12. Used the application\_status\_override values to override the values in the current\_status column
13. Dropped the application\_status\_override column as it did not exist in the mapped dataset
14. Dropped the missing observations from the current\_status column as there was not a “missing data” or “no data” acceptable value in the mapped dataset
15. Replaced values in the institution\_name column to match the mapped dataset requirements
16. Replaced values in the gender column to match the mapped dataset requirements
17. Replaced values in the ok\_to\_text column to match the mapped dataset requirements
18. Replaced values in the pell\_eligible column to match the mapped dataset requirements
19. Filled in missing observations in the pell\_eligible column with “Not Specified” in order to preserve those observations
20. Created a function to cutoff characters in the program column to match a 15-character length limit and then applied that limit to each row in the column
21. There was 1 duplicate observation in the student\_id / external\_id column so I dropped the duplicate observation and kept the first observed value
22. Exported the file to a csv for loading into the DB, but you can connect to a SQL DB instance and load the dataframe via a DB connector.

I do have a few questions about the dataset:

1. In the ok\_to\_text column, should we replace the missing observations with False and assume that it’s not ok\_to\_text unless we receive affirmation or just keep the missing values?
2. I noticed some values with blank application\_date values but would have current\_status values of admitted, application withdrawn, started application, submitted application, etc. so there is clearly application activity. Should I replace those missing values with the inquiry dates so there is a relative value on file in the DB or how do we want to handle those?
3. On the inquiry\_date where I filled in the epoch time, I thought about replacing those values with the application date where the application date exists, but I wasn’t sure if having inquiry\_date values that have the same value as the application\_date would have caused any issues.

I have included a fully documented Python script that matches the data to the mapped dataset. The Python script walks through each transformation step above and provides additional information on the logic behind each step. Please let me know if you have any questions about the output file and if I need to make any additional changes based on the answers to my questions above.

Thank you,