Update for Admit Decline Survey:

Every step on the flow is described. The numbers of the steps are associated with the boxes on the flow, for example, 1.) is the Union step at the start of the flow. 🡪 2.) is the clean step right after it.

A screenshot of a computer

Description automatically generated

1. Add all the Admit Decline Survey Sheets to Tableau Prep, then Union them by dragging each of the Admit Decline sheets to each other. While in the Union, most fields containing the same values have the same field names, however some years have different names for similar fields such as “What were the deciding factors you used when choosing your future Graduate Business Program? (Please Select all that Apply):” in the Admit Decline 5 (2022) and “What were the deciding factors upon choosing your future MBA/MS Program? (Please Select all that Apply): - Selected Choice” in Admit Decline (2018). To fix this, while in the union dropbox drag and drop the fields that hold the same values but different column names on another.
2. Removed irrelevant rows such as ‘Recipient email’, ‘user language’, etc., that do hold any value in analysis afterwards.
3. Grouped University Names that were misspelled, had different capitalizations, abbreviations, etc. Did the same with program names as well.
4. Ran a python script (cleaning-script.py) which used the longitude and latitude fields of the survey to find the location of the admits and for American applicants their states as well.

**SPLIT into two branches:**

**Bottom Branch (Whole dashboard except for Sankey):**

1. Removed the Longitude and Latitude columns, cleaned up other fields by removing numbers and extra spaces.
2. Ran a python script (replace.py) that cleaned field 'To which other schools did you apply? (If applicable)’ so that it would be ready for splitting using commas. Also sorted fields "What were the deciding factors upon choosing your future MBA/MS Program? (Please Select all that Apply): -" and "Program at Krannert to which you applied: -  Selected Choice" alphabetically so that any duplicates can be grouped up (it could not be grouped up if it was not sorted because you could have had (DS, MBA) and (MBA, DS) for programs applied at and they would be classed differently as two unique values when it is the same thing.
3. Must always have a buffer step or any step before going from one script step to another.
4. Script (stopwords.py) that removed stop words from open-ended responses, and conducted sentiment analysis, adding features to the dataset.
5. Buffer
6. Script (uni\_names.py) that replaces common abbreviations, misspelling of university names and more with the full name of the university.
7. Buffer with a small group fix
8. Script (unisplit.py) split and expanded dataset with single observations of the multi-valued features of universities applied to, programs applied to, and deciding factors.
9. Trimmed spaced after the split to make sure that values that might have been split with an extra space value were the same as the actual value (ex. ‘ USA’ is the same as ‘USA’)

**Top Branch (For Sankey Use Only):**

1. Programs at Krannert applied to were multivalued, with at most two values in each response, so this step splits them into two columns and cleans them.
2. Pivot the columns that were just split to include all the values from both columns independently.
3. Aggregate the values based on the programs at Krannert applied to, then pick the features wanted to be aggregated as well
4. Add a ‘Link’ column of just ones to then join with the link.csv spreadsheet which has data values that are required to build the Sankey.
5. Join link.csv and step 4 on the link column with an inner join