**Cleansing Your Data with Alteryx**

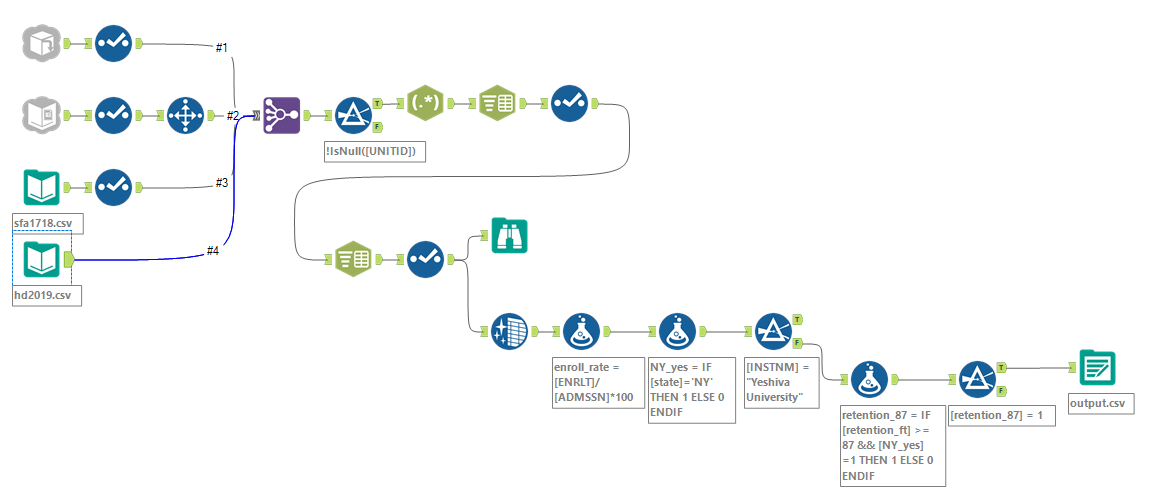
**by Qi Sun**

The datasets were downloaded from Integrated Postsecondary Education Data System, <https://nces.ed.gov/ipeds/datacenter/DataFiles.aspx?goToReportId=7> .

For the assignment of last week, three datasets were used, they are institution application data, retention data, and financial aid data. For this week, I added one more dataset to the last one. The new dataset contains institution Characteristics data, i.e. institutional name and address.

Three distinct data sources are 1. Amazon S3, 2. Google Sheets, and 3. csv files from the local drive.

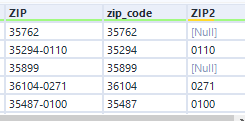
**Here is a screenshot of the workflow:**

****

**Tools used for this workflow are:**

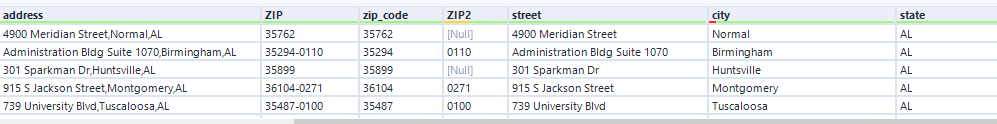
1. ‘Input Data’: input data from 1) Amazon S3 about application; 2) Google Sheets about student retention; 3) two csv files from the local drive about student financial aid.
2. ‘Select’: for the Amazon S3 and local csv files, I made changes to the data types from string to int. For the Google Sheets file, I made changes to the column names by using ‘Select’.
3. ‘Select Records’: for the Google Sheets file, I made selection to delete the first row that contains the column names.
4. ‘Join Multiple’: join 4 data files together by field ‘UNITID’.
5. ‘Filter’: select rows without missing UNITID.
6. ‘**Parse – RegEx**’: for ‘ZIP’ column, the default data type is string. I replaced the leading and trailing whitespace with ‘0’ by using regular expression ‘^\s+|\s+$.
7. ‘**Parse - Text To Columns**’: for the ‘ZIP’ column, the data are displayed as ZIP+4 code. I’ll get zip codes by splitting this column into two using delimiter ‘-‘ . Then, I used ‘Select’ to change ’ZIP1’ column name to ‘zip\_code’.

The screenshot below shows the first 5 records of the results.



‘**Parse -Text To Columns**’: for the ‘address’ column, I split it into street name, city, and state using delimiter ‘,‘. Then, I used ‘Select’ to change column names to street, city, state.

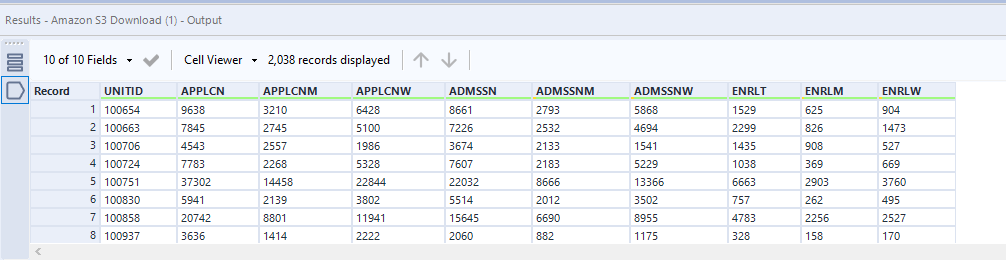
The screenshot below shows the first 5 records of the results.



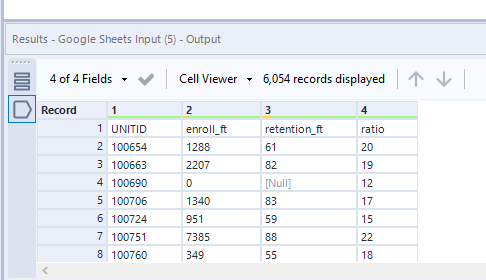
1. ‘Browse’: view a snapshot of the data. There are 2,038 records with 24 fields in this dataset.
2. ‘Data Cleansing’: replace null numerical values in this dataset with 0, replace missing string values with blanks.
3. ‘**Formula’**:
4. create a column to show the enrollment rate by using formula: total enroll/total admission.
5. **conditional if statement**: create a numerical variable ‘NY\_yes’ to show if the institution is in New York State by using the statement ‘IF [state]='NY' THEN 1 ELSE 0 ENDIF’.
6. create a column to show institutions that are in New York State and their retention rate is greater than that of Yeshiva University. We can get a lot of information here, including these institutions’ enrollment rate and financial aid info. Before this step, I used ‘filter’ tool to check YU’s retention rate to get the number (87).
7. ‘Output Data’: export data to a csv file saved at the local drive.

**Screenshots of the inputted data:**

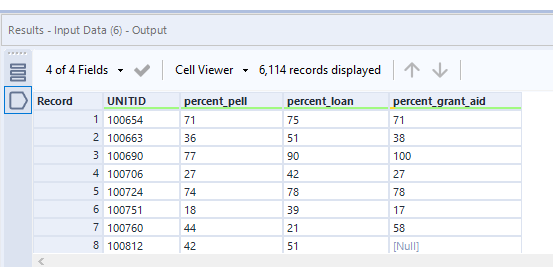
1) Amazon S3 about application:

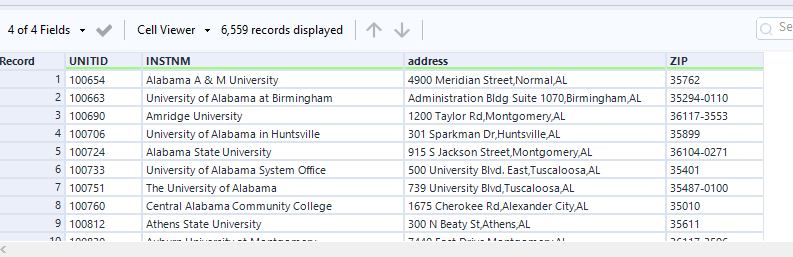


2) Google Sheets about student retention:

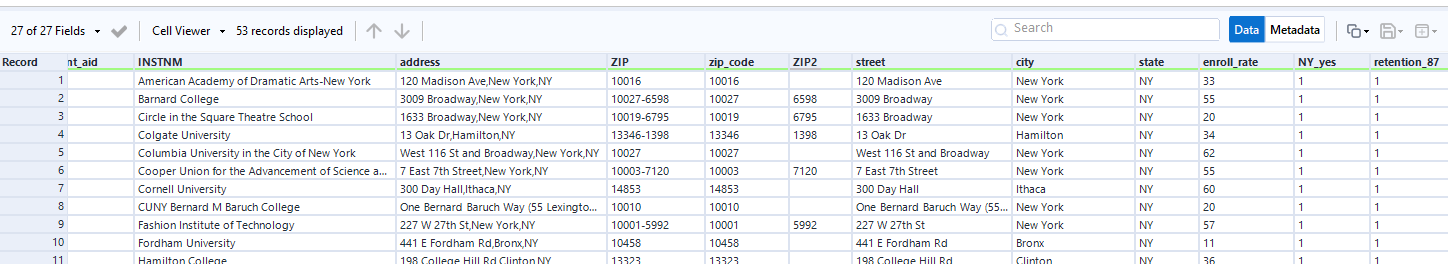


3) csv file from the local drive about student financial aid:



csv file from the local drive about institution Characteristics:

**Screenshot of the outputted data:**

****