Cleansing Your Data with Alteryx

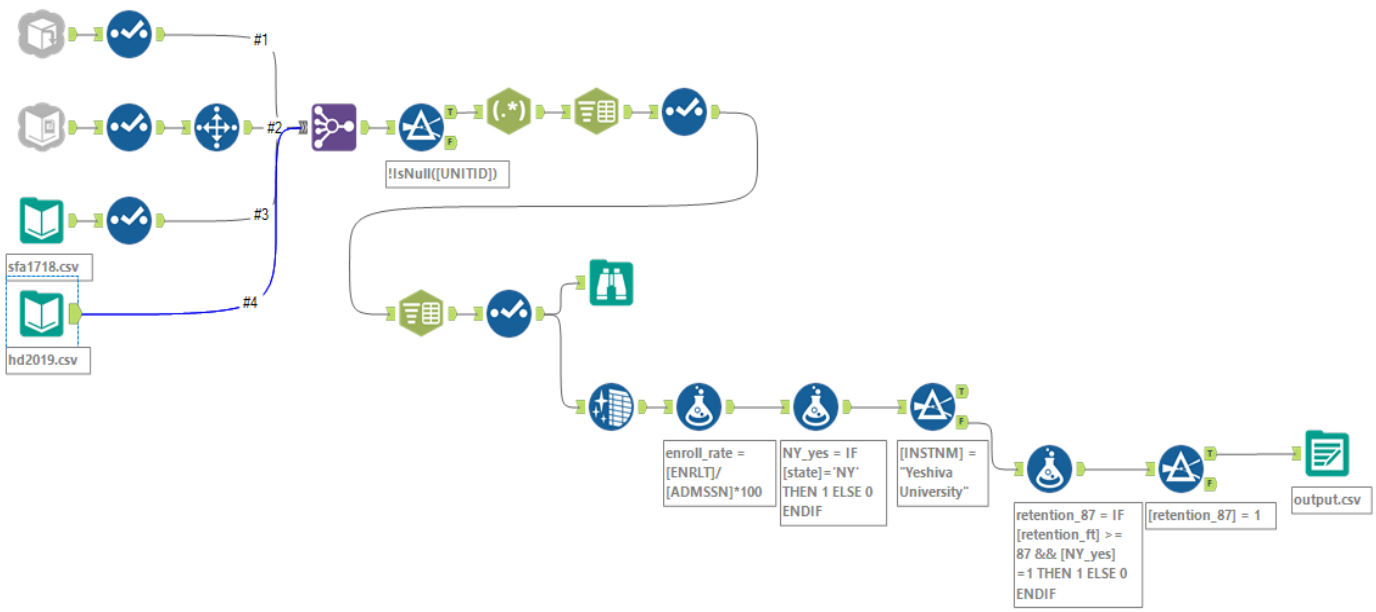
by Qi Sun

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.

ZIP	zip_code	ZIP2
35762	35762	[Null]
35294-0110	35294	0110
35899	35899	[Null]
36104-0271	36104	0271
35487-0100	35487	0100

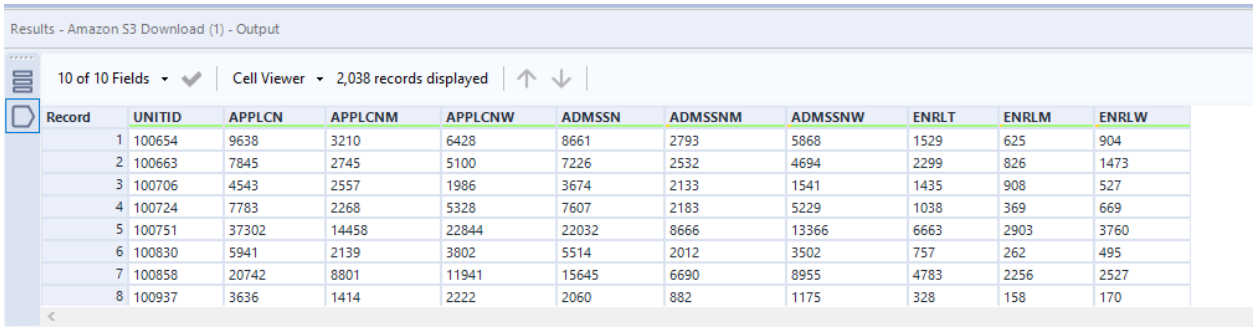
   '**Parse -Text To Columns**': for the 'address' column, I split it into street name, city, and state using delimiter ','. Then, I used 'Select' to change column names to street, city, state.
   The screenshot below shows the first 5 records of the results.

address	ZIP	zip_code	ZIP2	street	city	state
4900 Meridian Street,Normal,AL	35762	35762	[Null]	4900 Meridian Street	Normal	AL
Administration Bldg Suite 1070,Birmingham,AL	35294-0110	35294	0110	Administration Bldg Suite 1070	Birmingham	AL
301 Sparkman Dr,Huntsville,AL	35899	35899	[Null]	301 Sparkman Dr	Huntsville	AL
915 S Jackson Street,Montgomery,AL	36104-0271	36104	0271	915 S Jackson Street	Montgomery	AL
739 University Blvd,Tuscaloosa,AL	35487-0100	35487	0100	739 University Blvd	Tuscaloosa	AL

8. 'Browse': view a snapshot of the data. There are 2,038 records with 24 fields in this dataset.
9. 'Data Cleansing': replace null numerical values in this dataset with 0, replace missing string values with blanks.
10. '**Formula'**:
    1) create a column to show the enrollment rate by using formula: total enroll/total admission.
    2) **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'.
    3) 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).
11. 'Output Data': export data to a csv file saved at the local drive.

Screenshots of the inputted data:

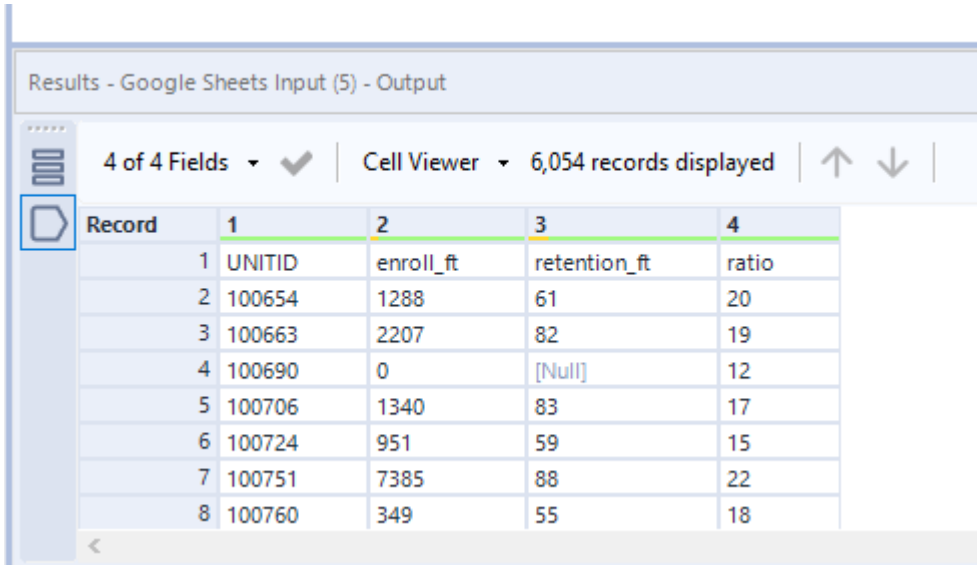
1) Amazon S3 about application:


Results - Amazon S3 Download (1) - Output

10 of 10 Fields | Cell Viewer | 2,038 records displayed

Record	UNITID	APPLCN	APPLCNM	APPLCNW	ADMSSN	ADMSSNM	ADMSSNW	ENRLT	ENRLM	ENRLW
1	100654	9638	3210	6428	8661	2793	5868	1529	625	904
2	100663	7845	2745	5100	7226	2532	4694	2299	826	1473
3	100706	4543	2557	1986	3674	2133	1541	1435	908	527
4	100724	7783	2268	5328	7607	2183	5229	1038	369	669
5	100751	37302	14458	22844	22032	8666	13366	6663	2903	3760
6	100830	5941	2139	3802	5514	2012	3502	757	262	495
7	100858	20742	8801	11941	15645	6690	8955	4783	2256	2527
8	100937	3636	1414	2222	2060	882	1175	328	158	170

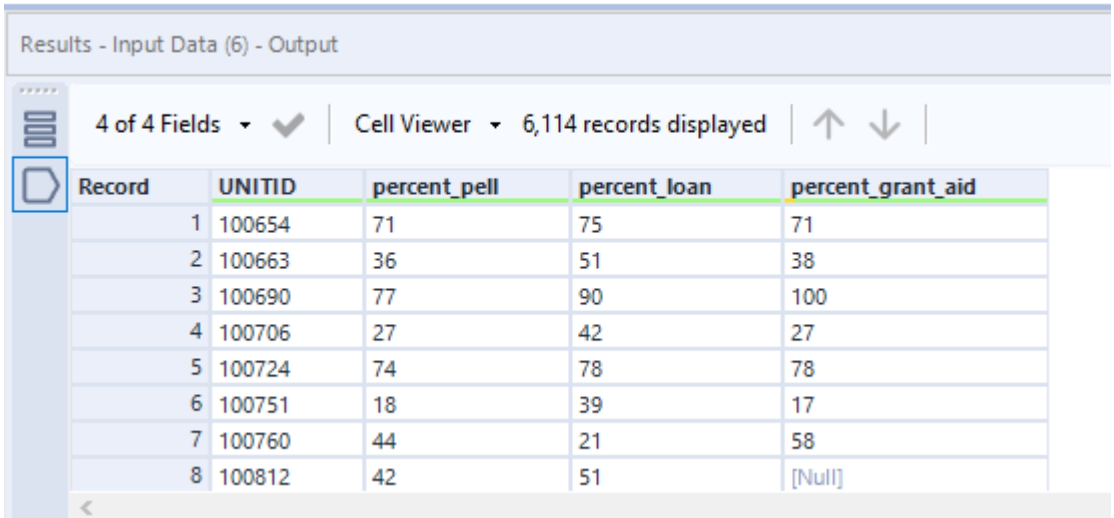
2) Google Sheets about student retention:


Results - Google Sheets Input (5) - Output

4 of 4 Fields | Cell Viewer | 6,054 records displayed

Record	1	2	3	4
1	UNITID	enroll_ft	retention_ft	ratio
2	100654	1288	61	20
3	100663	2207	82	19
4	100690	0	[Null]	12
5	100706	1340	83	17
6	100724	951	59	15
7	100751	7385	88	22
8	100760	349	55	18

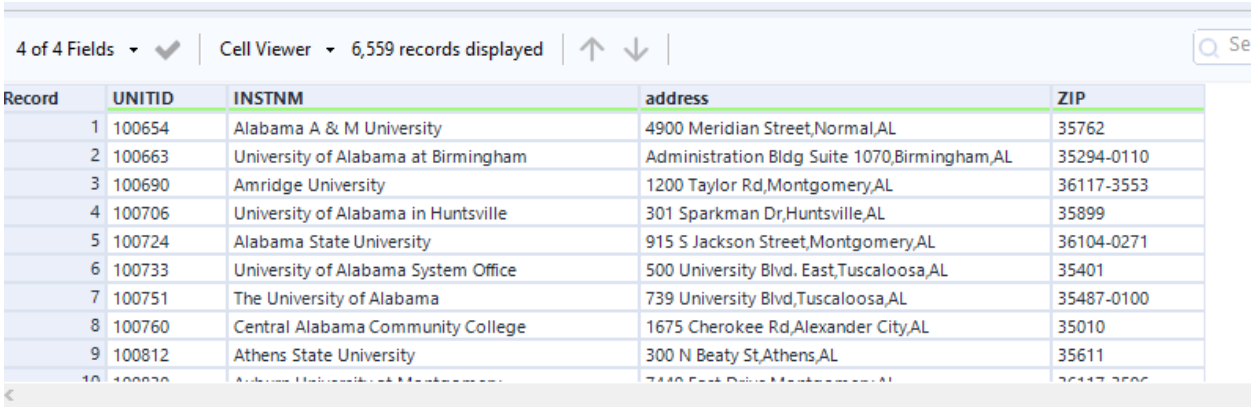
3) csv file from the local drive about student financial aid:


Results - Input Data (6) - Output

4 of 4 Fields | Cell Viewer | 6,114 records displayed

Record	UNITID	percent_pell	percent_loan	percent_grant_aid
1	100654	71	75	71
2	100663	36	51	38
3	100690	77	90	100
4	100706	27	42	27
5	100724	74	78	78
6	100751	18	39	17
7	100760	44	21	58
8	100812	42	51	[Null]

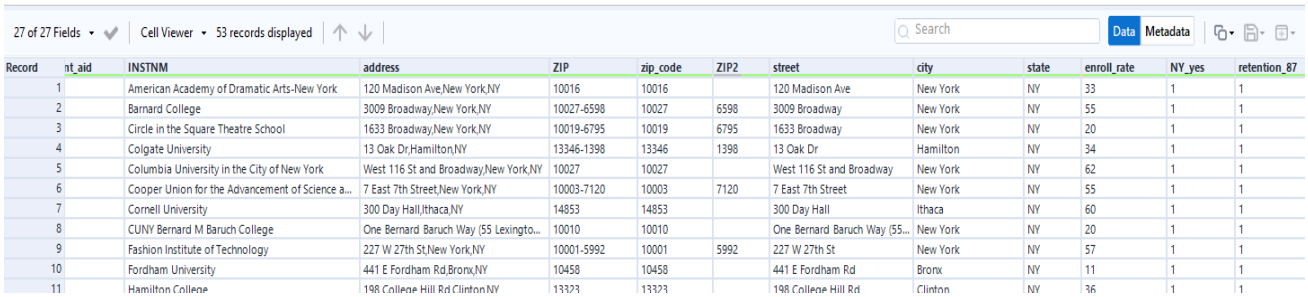
csv file from the local drive about institution Characteristics:



4 of 4 Fields | Cell Viewer | 6,559 records displayed

Record	UNITID	INSTNM	address	ZIP
1	100654	Alabama A & M University	4900 Meridian Street,Normal,AL	35762
2	100663	University of Alabama at Birmingham	Administration Bldg Suite 1070,Birmingham,AL	35294-0110
3	100690	Amridge University	1200 Taylor Rd,Montgomery,AL	36117-3553
4	100706	University of Alabama in Huntsville	301 Sparkman Dr,Huntsville,AL	35899
5	100724	Alabama State University	915 S Jackson Street,Montgomery,AL	36104-0271
6	100733	University of Alabama System Office	500 University Blvd. East,Tuscaloosa,AL	35401
7	100751	The University of Alabama	739 University Blvd,Tuscaloosa,AL	35487-0100
8	100760	Central Alabama Community College	1675 Cherokee Rd,Alexander City,AL	35010
9	100812	Athens State University	300 N Beaty St,Athens,AL	35611

Screenshot of the outputted data:



27 of 27 Fields | Cell Viewer | 53 records displayed | Search | Data | Metadata

Record	nt_aid	INSTNM	address	ZIP	zip_code	ZIP2	street	city	state	enroll_rate	NY_yes	retention_87
1		American Academy of Dramatic Arts-New York	120 Madison Ave,New York,NY	10016	10016		120 Madison Ave	New York	NY	33	1	1
2		Barnard College	3009 Broadway,New York,NY	10027-6598	10027	6598	3009 Broadway	New York	NY	55	1	1
3		Circle in the Square Theatre School	1633 Broadway,New York,NY	10019-6795	10019	6795	1633 Broadway	New York	NY	20	1	1
4		Colgate University	13 Oak Dr,Hamilton,NY	13346-1398	13346	1398	13 Oak Dr	Hamilton	NY	34	1	1
5		Columbia University in the City of New York	West 116 St and Broadway,New York,NY	10027	10027		West 116 St and Broadway	New York	NY	62	1	1
6		Cooper Union for the Advancement of Science a...	7 East 7th Street,New York,NY	10003-7120	10003	7120	7 East 7th Street	New York	NY	55	1	1
7		Cornell University	300 Day Hall,Ithaca,NY	14853	14853		300 Day Hall	Ithaca	NY	60	1	1
8		CUNY Bernard M Baruch College	One Bernard Baruch Way (55 Lexingto...	10010	10010		One Bernard Baruch Way (55...	New York	NY	20	1	1
9		Fashion Institute of Technology	227 W 27th St,New York,NY	10001-5992	10001	5992	227 W 27th St	New York	NY	57	1	1
10		Fordham University	441 E Fordham Rd,Bronx,NY	10458	10458		441 E Fordham Rd	Bronx	NY	11	1	1
11		Hamilton College	198 College Hill Rd,Clinton,NY	13323	13323		198 College Hill Rd	Clinton	NY	36	1	1