

Dataset source: Higher Education Student Data

(Chart 7 - HE students by subject area and sex 2014/15 to 2018/19)

URL: <https://data.europa.eu/data/datasets/higher-education-student-data?locale=en>

Columns: AcademicYear, SubjectArea, Sex, Number

Analyze Gender Education by Subject

Input Data (2) - Configuration

Connect a File or Database

W Project-chart-7 HE students by subject area and sex 201415 to 201819.csv

Set Up a Connection

Use Data Connection Manager (DCM)

Options

Name	No
5 Delimiters	-
6 First Row Contains Field Names	<input checked="" type="checkbox"/>
7 Field Length	254
8 Start Data Import on Line	14
9 Ignore Delimiters in	Quotes

Preview (first 100 records)

Academic Year	Subject area	Sex
2014/15	Medicine & dentistry	Female
2014/15	Medicine & dentistry	Male
2014/15	Subjects allied to medicine	Female
2014/15	Subjects allied to medicine	Male
2014/15	Biological sciences	Female
2014/15	Biological sciences	Male
2014/15	Veterinary science	Female
2014/15	Veterinary science	Male
2014/15	Agriculture & related subjects	Female
2014/15	Agriculture & related subjects	Male
2014/15	Physical sciences	Female
2014/15	Physical sciences	Male
2014/15	Mathematical sciences	Female
2014/15	Mathematical sciences	Male
2014/15	Computer science	Female

(Input Data)

Start data import on line 14 as the previous lines contain the data source and some metadata.

Filter (29) - Configuration

Select Basic or Custom Filter

Basic filter

Academic Year Is not empty

Custom filter

fx [IsEmpty([Academic Year])]

Workflow Diagram:

```

graph LR
    A[chart-7 HE students by subject area and sex 201415 to 201819.csv] --> B[IsEmpty([Academic Year])]
    B --> C[Education Dataset.xlsx Query=Sheet1]
  
```

Results - Filter (29) - Out - True

Record	Academic Year	Subject area	Sex	Number
184	2018/19	Historical & philosophical studies	Male	36785
185	2018/19	Creative arts & design	Female	117860
186	2018/19	Creative arts & design	Male	63530
187	2018/19	Education	Female	111050
188	2018/19	Education	Male	31675
189	2018/19	Combined	Female	19750
190	2018/19	Combined	Male	11995

We found some records doesn't contain any data except for the sex and this was preventing us from

correctly splitting the data, so we filtered the data on academic year when it is not empty true then output it in Education Dataset file to work on it. **(Filter)**

The screenshot shows the Alteryx Designer interface with the 'Tile' tool configured. The 'Tile Method' is set to 'Equal Records' and the 'Number of Tiles' is set to 2. The 'Results - Tile (13) - Input' pane displays a table with 4 fields: Academic Year, Subject area, Sex, and Number. The data is split into two tiles based on the configuration.

Record	Academic Year	Subject area	Sex	Number
1	2014/15	Medicine & dentistry	Female	37335
2	2014/15	Medicine & dentistry	Male	28660
3	2014/15	Subjects allied to medicine	Female	218510
4	2014/15	Subjects allied to medicine	Male	56815
5	2014/15	Biological sciences	Female	128775
6	2014/15	Biological sciences	Male	62570
7	2014/15	Veterinary science	Female	4495
8	2014/15	Veterinary science	Male	1405

Using **Tile** method we choose equal records and number of tiles 2 to split the data into 2 sets.

The screenshot shows the Alteryx Designer interface with the 'Filter' tool configured. The 'Basic filter' is set to 'Tile_Num = 1'. The 'Results - Filter (16) - Input' pane displays a table with 6 fields: Academic Year, Subject area, Sex, Number, Tile_Num, and Tile_SequenceNum. The data is filtered based on the configuration.

Record	Academic Year	Subject area	Sex	Number	Tile_Num	Tile_SequenceNum
92	2016/17	Computer science	Male	83710	1	92
93	2016/17	Engineering & technology	Female	29025	1	93
94	2016/17	Engineering & technology	Male	136085	1	94
95	2016/17	Architecture, building & planning	Female	19350	1	95
96	2016/17	Architecture, building & planning	Male	31905	2	1
97	2016/17	Social studies	Female	139915	2	2
98	2016/17	Social studies	Male	81685	2	3
99	2016/17	Law	Female	55985	2	4

Using filter on Tile_Num we will export the data into 2 excel files the first half is in Destination_1 if filter is True and the second half is in Destination_2 if it is false.

Workflow - Configuration

Canvas Options

- Layout Direction: Horizontal
- Annotations: Show
- Connection Progress: Show Only When Running

Workflow

Education Dataset.xlsx
Query=Sheet15

[Tile_Num] = 1

Destination_1.xlsx
Query=Destination_1

Destination_2.xlsx
Query=Destination_2

Results - Workflow - Messages

All 0 Errors 0 Conv Errors 0 Warnings 6 Info 3 Files

Designer x64 The Designer x64 reported: Allocating requested memory would be more than available physical memory. Reverting to 2354.9 MB of memory.

Designer x64 The Designer x64 reported: This is AMP Engine; running 12 worker threads; memory limit 2354.9 MB.

Tile (11) Number of Tiles was set to 2

Input Data (11) 180 records were read from "E:\COMPUTER SCIENCE\LEVEL 4\2 Second Term\Data Warehouse\DW Project\Education Dataset.xlsx" ("Sheet15")

Tile (11) 2 tiles were generated.

Filter (16) 95 records were True and 95 were False

Output Data (17) 95 records were written to "E:\COMPUTER SCIENCE\LEVEL 4\2 Second Term\Data Warehouse\DW Project\Destination_1.xlsx" (Destination_1)

Output Data (18) 95 records were written to "E:\COMPUTER SCIENCE\LEVEL 4\2 Second Term\Data Warehouse\DW Project\Destination_2.xlsx" (Destination_2)

Designer x64 Finished running DW Project part 2.yxmd in 0.6 seconds using AMP engine.

Transformations

1: filter records on specific condition (custom condition: **Filter** out subjects that contain either studies or combined)

Workflow - Configuration

Select Basic or Custom Filter

Basic filter

Select column... =

Custom filter

[Contains([Subject area], "studies")] &&
[Contains([Subject area], "combined")]

Workflow

Destination_1.xlsx
Query=Destination_1

[Contains([Subject area], "studies")] &&
[Contains([Subject area], "combined")]

Destination_2.xlsx
Query=Destination_2

Results - Filter (2) - Out - True

6 of 6 Fields Cell Viewer 79 records displayed Search Data Metadata Actions

Record	Academic Year	Subject area	Sex	Number	Title_Num	Title_SequenceNum
73	2016/17	Mathematical sciences	Female	16265	1	89
74	2016/17	Mathematical sciences	Male	27580	1	90
75	2016/17	Computer science	Female	17390	1	91
76	2016/17	Computer science	Male	83710	1	92
77	2016/17	Engineering & technology	Female	29025	1	93
78	2016/17	Engineering & technology	Male	136085	1	94
79	2016/17	Architecture, building & planning	Female	19350	1	95

2: choose any string column and uppercase the first character only (Using Formula)

The screenshot shows the Alteryx Designer interface with the Formula tool configured. The output column is named "Subject area" and the formula is: `Uppercase(Left([Subject area],1)) + LowerCase(Substring([Subject area],1,100))`. The data type is set to V_String and the size is 255. The results pane shows 6 fields: Academic Year, Subject area, Sex, Number, Title_Num, and Title_SequenceNum. The data is displayed in a table with 79 records.

Record	Academic Year	Subject area	Sex	Number	Title_Num	Title_SequenceNum
73	2016/17	Mathematical sciences	Female	16265	1	89
74	2016/17	Mathematical sciences	Male	27580	1	90
75	2016/17	Computer science	Female	17390	1	91
76	2016/17	Computer science	Male	83710	1	92
77	2016/17	Engineering & technology	Female	29025	1	93
78	2016/17	Engineering & technology	Male	136085	1	94
79	2016/17	Architecture, building & planning	Female	19350	1	95

3: split any column into many columns (Using Text To Columns)

The screenshot shows the Alteryx Designer interface with the Text To Columns tool configured. The column to split is "Academic Year" and the delimiter is "/". The number of columns is set to 2. The output root name is "End Academic Year". The results pane shows 8 fields: Academic Year, Subject area, Sex, Number, Title_Num, Title_SequenceNum, End Academic Year, and End Academ. The data is displayed in a table with 79 records.

Record	Academic Year	Subject area	Sex	Number	Title_Num	Title_SequenceNum	End Academic Year	End Academ
1	2014/15	Medicine & dentistry	Female	37335	1	1	2014	15
2	2014/15	Medicine & dentistry	Male	28660	1	2	2014	15
3	2014/15	Subjects allied to medicine	Female	218510	1	3	2014	15
4	2014/15	Subjects allied to medicine	Male	56815	1	4	2014	15
5	2014/15	Biological sciences	Female	128775	1	5	2014	15
6	2014/15	Biological sciences	Male	62570	1	6	2014	15
7	2014/15	Veterinary science	Female	4495	1	7	2014	15
8	2014/15	Veterinary science	Male	1405	1	8	2014	15

We wanted to add '20' before the End Academic Year2, So we used **Formula**:

The screenshot shows the Alteryx Designer interface. On the left, the 'Formula (9) - Configuration' pane is open, showing the 'Output Column' 'End Academic Year2' with a value of 2015. The formula entered is '20 * [End Academic Year2]'. The data type is 'V_String' and the size is '255'. The main workspace shows a workflow diagram with the following tools: 'Destination, Table Query = Destination', 'Contains (Subject area, "studies")', 'Subject area = Uppercase(Left (Subject area), 1)) + LowerCase (Substring (Subject area, 1))', and 'End Academic Year2 = 20 * [End Academic Year2]'.

Using **SELECT**, we removed the unused columns, renamed the new columns and adjusted the type of students number.

The screenshot shows the Alteryx Designer interface. On the left, the 'Select (10) - Configuration' pane is open, showing the 'Options' tab. The columns are listed with their types and sizes. The 'Results - Select (10) - Input' pane shows a table with 8 fields and 79 records displayed.

Field	Type	Size	Rename	Description
Academic Year	V_String	255		
Subject area	V_String	255		
Sex	V_String	255		
Number	Int64	8	#Students	
Title_Num	Double	8		
Title_SequenceNum	Double	8		
End Academic Year1	V_String	255	Start Academic Year	TextToCol
End Academic Year2	V_String	255	End Academic Year	TextToCol
Unknown	Unknown	0		Dynamic

Record	Academic Year	Subject area	Sex	Number	Title_Num	Title_SequenceNum	End Academic Year1	End Academic Year2
1	2014/15	Medicine & dentistry	Female	37335	1	1	2014	2015
2	2014/15	Medicine & dentistry	Male	28660	1	2	2014	2015
3	2014/15	Subjects allied to medicine	Female	218510	1	3	2014	2015
4	2014/15	Subjects allied to medicine	Male	56815	1	4	2014	2015
5	2014/15	Biological sciences	Female	128775	1	5	2014	2015
6	2014/15	Biological sciences	Male	62570	1	6	2014	2015
7	2014/15	Veterinary science	Female	4495	1	7	2014	2015
8	2014/15	Veterinary science	Male	1405	1	8	2014	2015

4: replace any white spaces with underscore for any string column (Using **Formula REGEX Replace**)

The screenshot shows the Alteryx Designer interface with the Formula tool configured to replace spaces with underscores in the 'Subject area' field. The formula used is `REGEX_Replace([Subject area], " ", "_")`. The data preview shows 79 records with columns: Subject area, Sex, #Students, Start Academic Year, and End Academic Year.

Record	Subject area	Sex	#Students	Start Academic Year	End Academic Year
1	Medicine & dentistry	Female	37335	2014	2015
2	Medicine & dentistry	Male	28660	2014	2015
3	Subjects allied to medicine	Female	218510	2014	2015
4	Subjects allied to medicine	Male	56815	2014	2015
5	Biological sciences	Female	128775	2014	2015
6	Biological sciences	Male	82570	2014	2015
7	Veterinary science	Female	4495	2014	2015
8	Veterinary science	Male	1405	2014	2015

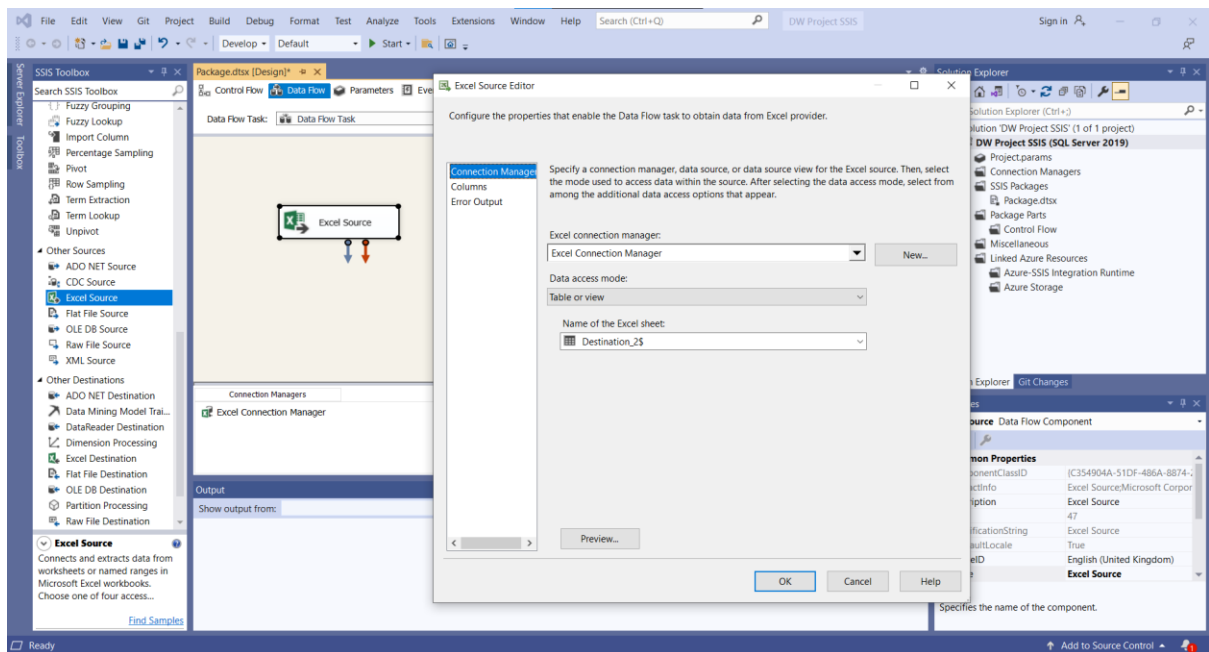
Divide the current Destination1 aka. First Half again into two parts. (using **Tile**, **Filter**, **Select**, **Output Data**)

The screenshot shows the Alteryx Designer interface with the workflow extended to split the data into two parts. The workflow includes a Tile tool, a Filter tool, and an Output Data tool. The data preview shows 39 records with columns: Subject area, Sex, #Students, Start Academic Year, and End Academic Year.

Record	Subject area	Sex	#Students	Start Academic Year	End Academic Year
1	Physical_sciences	Female	38285	2015	2016
2	Physical_sciences	Male	56555	2015	2016
3	Mathematical_sciences	Female	16025	2015	2016
4	Mathematical_sciences	Male	27060	2015	2016
5	Computer_science	Female	16480	2015	2016
6	Computer_science	Male	79700	2015	2016
7	Engineering_&_technology	Female	27745	2015	2016
8	Engineering_&_technology	Male	135300	2015	2016

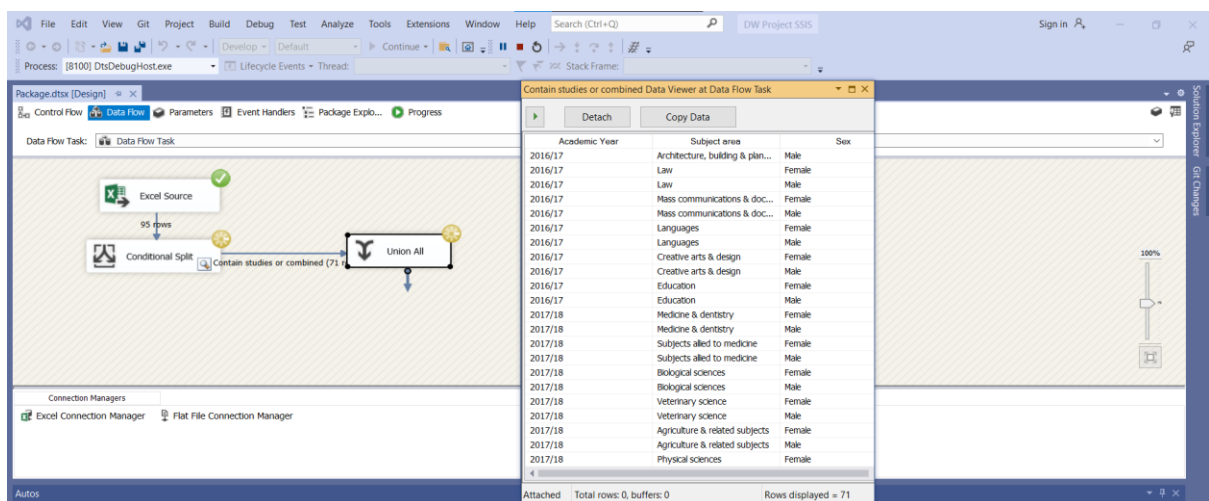
NOW Let's move to SSIS

Import Desination_2 using Excel Source

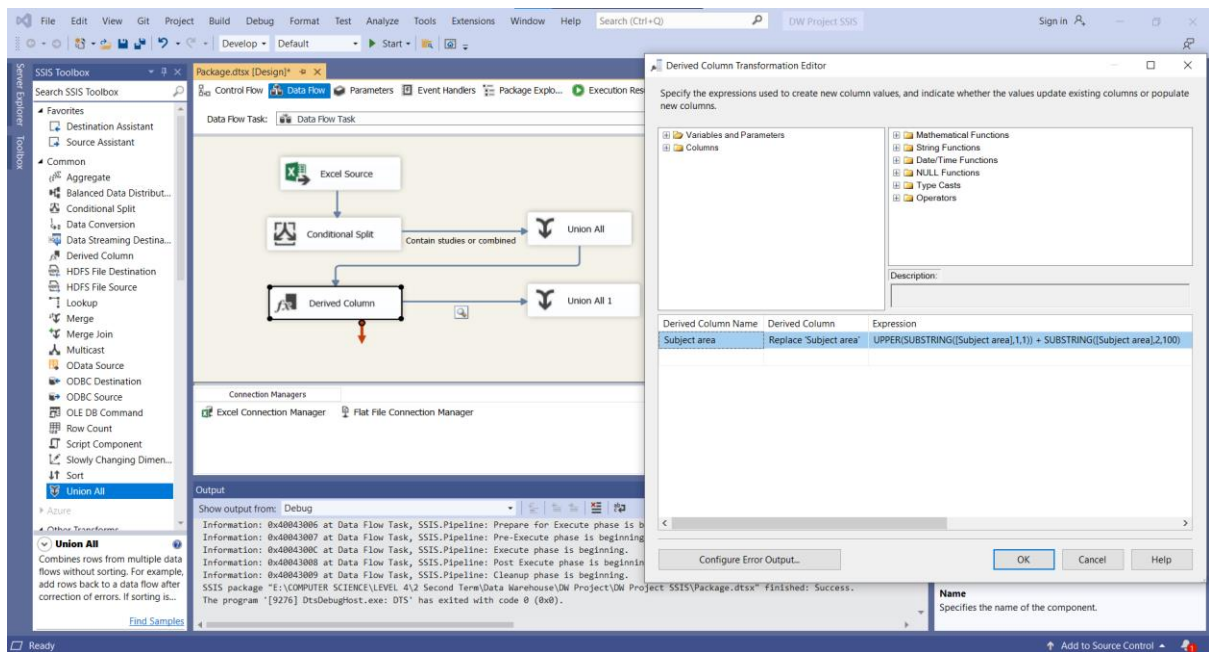


Transformations

1: filter records on specific condition (custom condition: Filter out subjects that contain either studies or combined using **Conditional Split**)



2: choose any string column and uppercase the first character only (Using Derived Column)

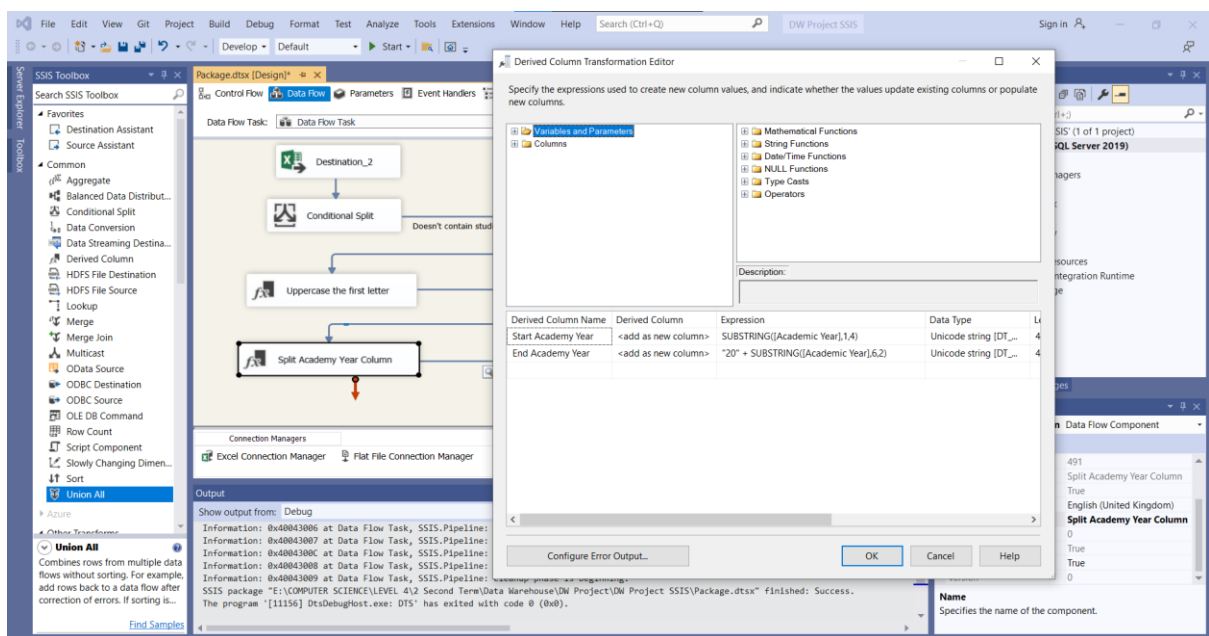


The screenshot shows the SSIS Package Designer with a Data Flow Task. The Data Flow Task contains an Excel Source, a Conditional Split, a Derived Column, and a Union All. The Derived Column transformation is selected, and the Derived Column Transformation Editor is open. The editor shows the following table:

Derived Column Name	Derived Column	Expression
Subject area	Replace 'Subject area'	UPPER(SUBSTRING([Subject area],1,1)) + SUBSTRING([Subject area],2,100)

The editor also shows a list of functions and operators on the right, including Mathematical Functions, String Functions, Date/Time Functions, NULL Functions, Type Casts, and Operators.

3: split any column into many columns (Using Derived Column)

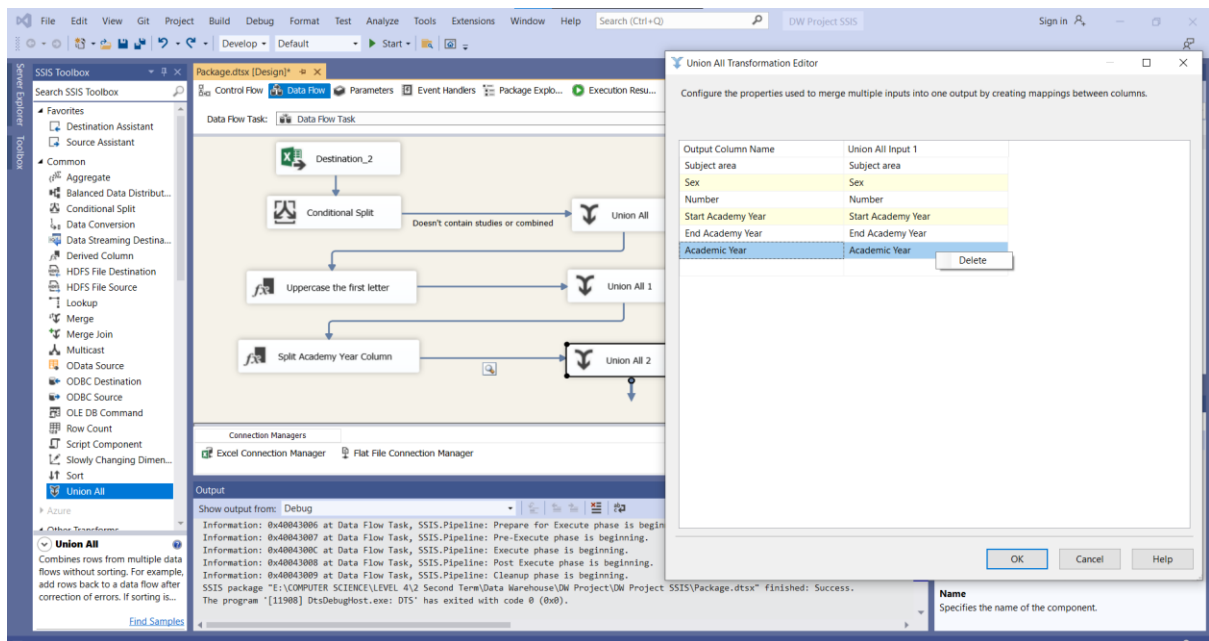


The screenshot shows the SSIS Package Designer with a Data Flow Task. The Data Flow Task contains a Destination_2, a Conditional Split, an Uppercase the first letter, and a Split Academy Year Column. The Split Academy Year Column transformation is selected, and the Derived Column Transformation Editor is open. The editor shows the following table:

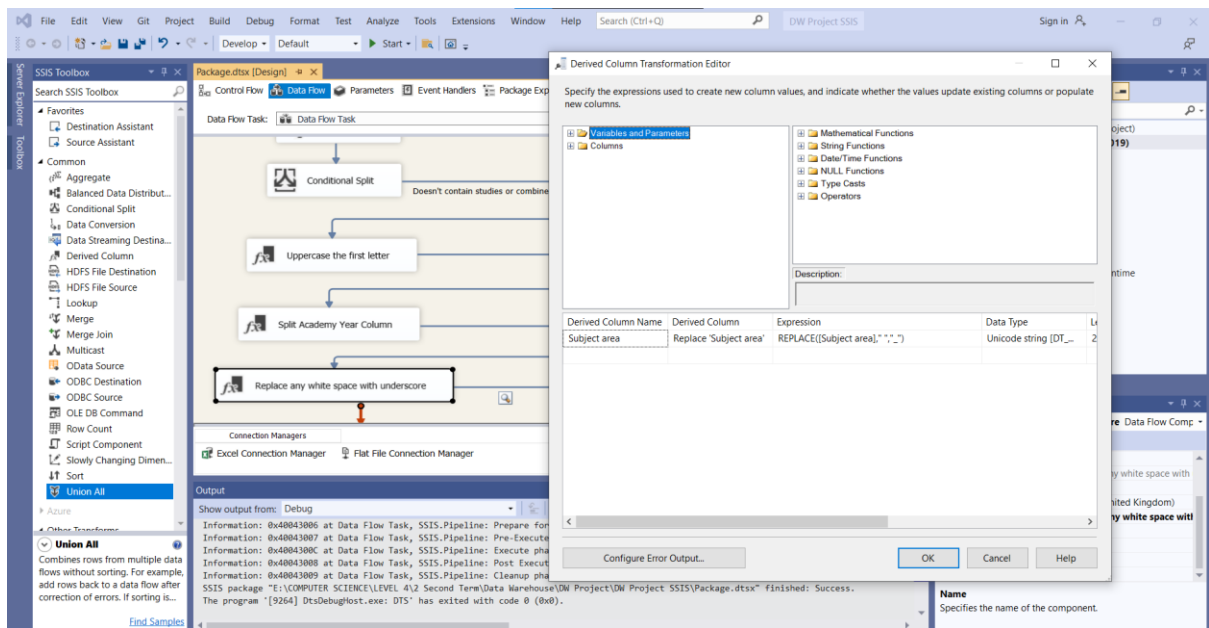
Derived Column Name	Derived Column	Expression	Data Type
Start Academy Year	<add as new column>	SUBSTRING([Academic Year],1,4)	Unicode string [DT_...]
End Academy Year	<add as new column>	"20" + SUBSTRING([Academic Year],6,2)	Unicode string [DT_...]

The editor also shows a list of functions and operators on the right, including Variables and Parameters, Columns, Mathematical Functions, String Functions, Date/Time Functions, NULL Functions, Type Casts, and Operators.

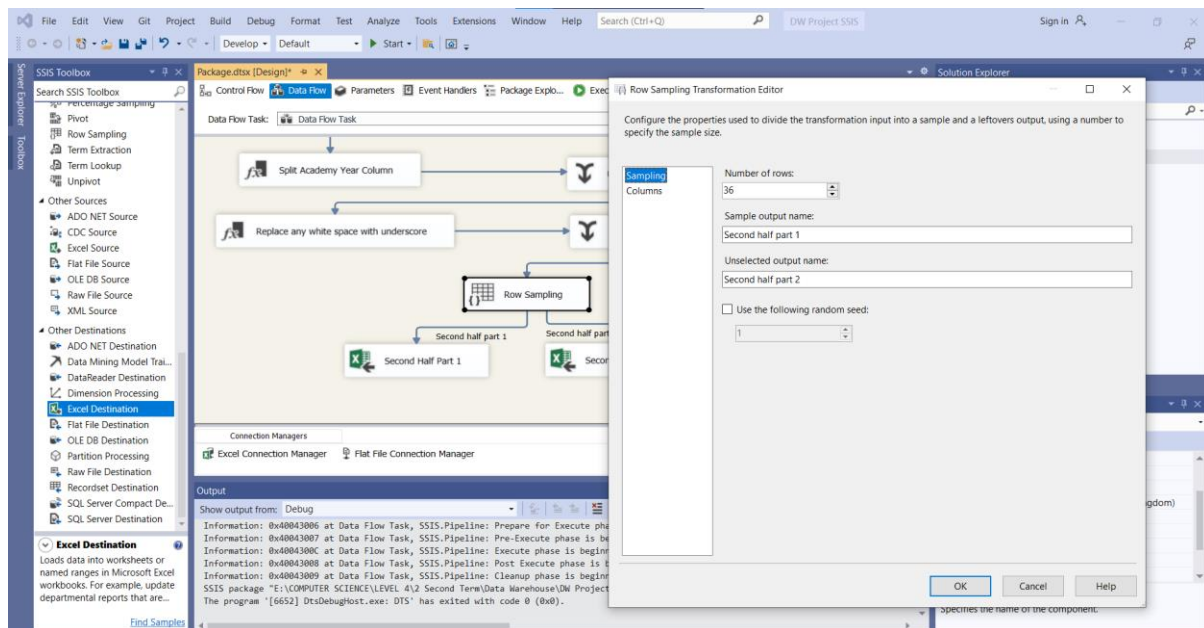
we deleted the unused columns:



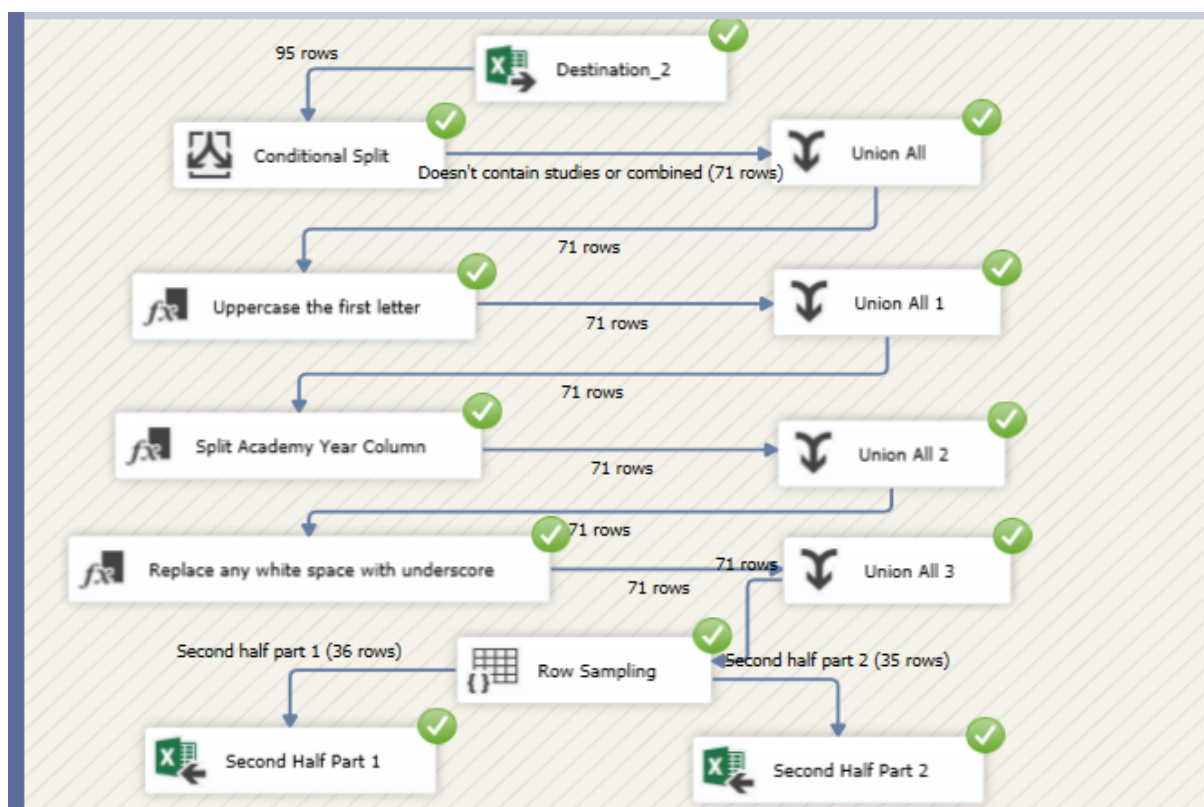
4: replace any white spaces with underscore for any string column (Using Derived Column)



Divide the current Destination2 aka. Second Half again into two parts. (using **Row Sampling Transformation, Excel Destination**)

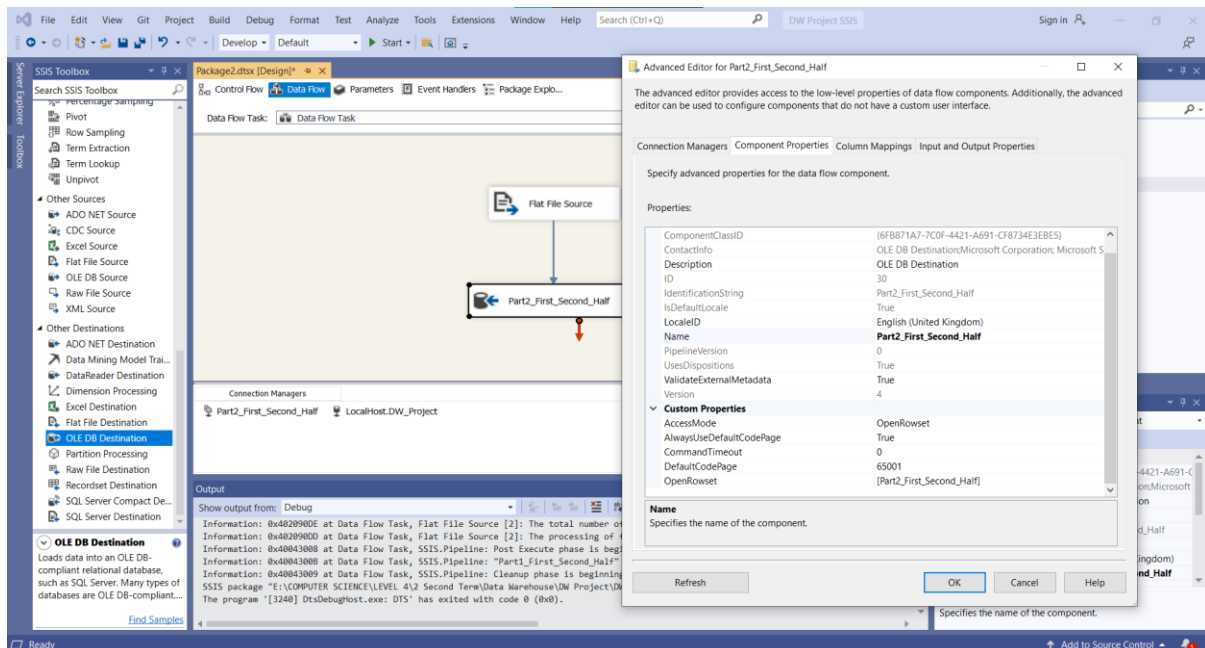
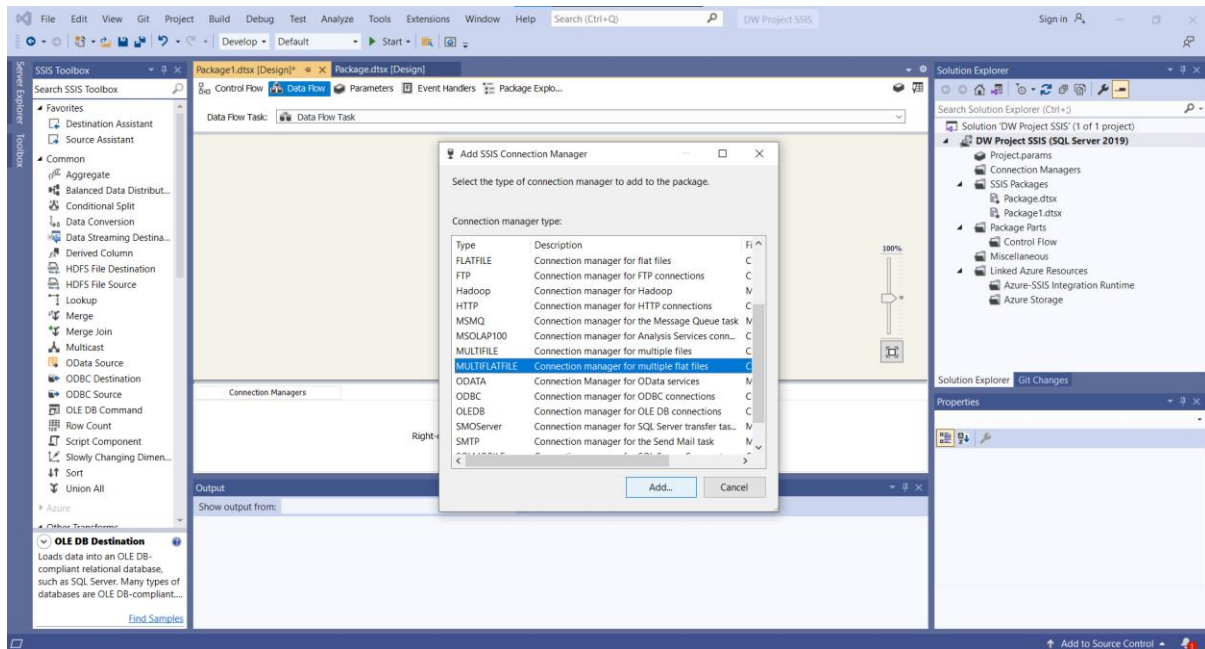


Final Result

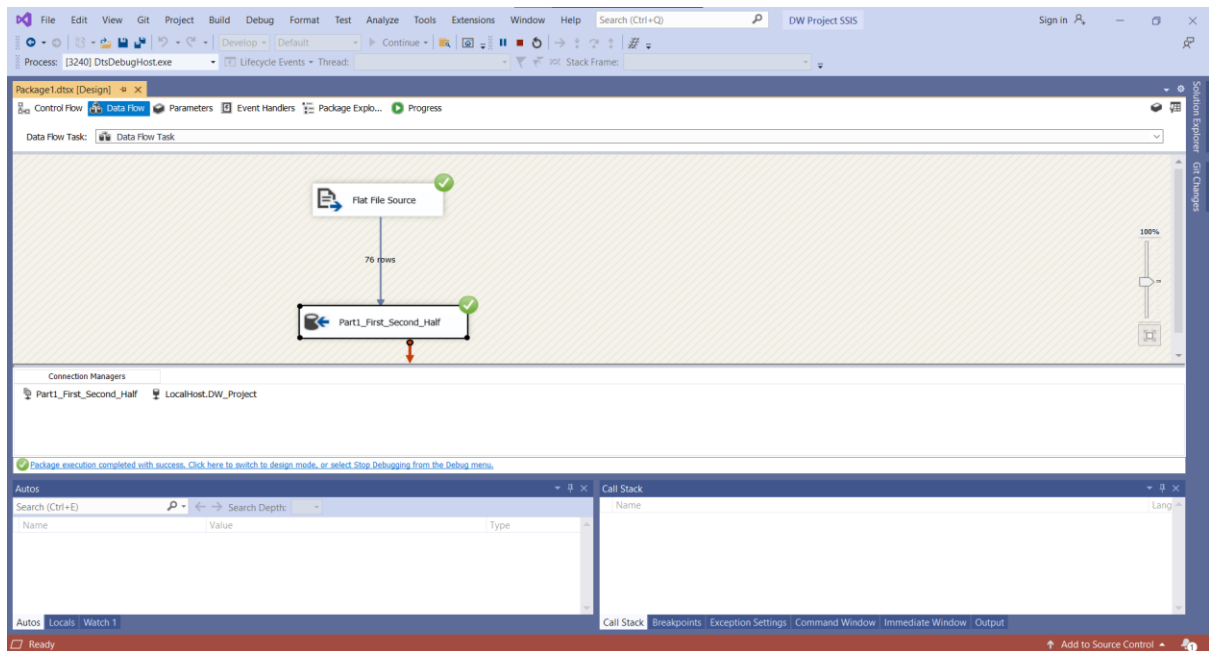


Load part 1 from first and second halves in destination 1 in SQL Server:

Using MULTIPLEFLATFILE to make the connection to folder that contain them.

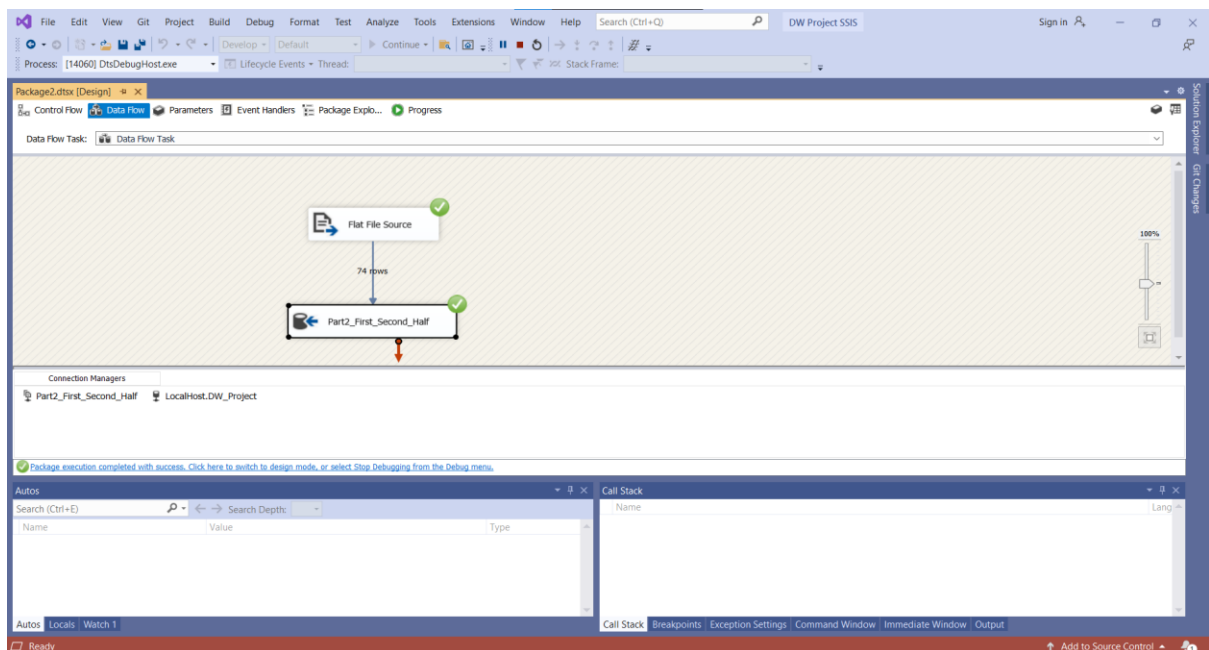


We change the OLE DB Destination advanced editor setting DefaultCodePage to 65001 and make AlwaysUseDefaultCodePage True to solve the column cannot be processed because more than one code page. (From: Stakeoverflow)



and here it is 76 rows (40 rows from First Half Part1 & 36 rows from Second Half Part1)

Same steps to load the Part 2:



and here it is 74 rows (39 rows from First Half Part1 & 35 rows from Second Half Part1)

Left Screenshot (Part2 First Second Half):

```

SELECT TOP (1000) [Subject_area]
, [Sex]
, [#Students]
, [Start_Academic_Year]
, [End_Academic_Year]
FROM [DW_Project].[dbo].[Part2_First_Second_Half]

```

Subject_area	Sex	#Students	Start Academic Year	End Academic Year
1 Physical_sciences	Female	35285	2015	2016
2 Physical_sciences	Male	56585	2015	2016
3 Mathematical_sciences	Female	16025	2015	2016
4 Mathematical_sciences	Male	27060	2015	2016
5 Computer_science	Female	16480	2015	2016
6 Computer_science	Male	79700	2015	2016
7 Engineering_and_technology	Female	27745	2015	2016
8 Engineering_and_technology	Male	135380	2015	2016
9 Architecture_building_and_planning	Female	18245	2015	2016
10 Architecture_building_and_planning	Male	31040	2015	2016
11 Law	Female	54650	2015	2016
12 Law	Male	33895	2015	2016
13 Mass_communications_and_documentation	Female	29000	2015	2016
14 Mass_communications_and_documentation	Male	19965	2015	2016
15 Languages	Female	75995	2015	2016
16 Languages	Male	32220	2015	2016
17 Creative_arts_and_design	Female	108580	2015	2016
18 Creative_arts_and_design	Male	61180	2015	2016
19 Education	Female	119225	2015	2016
20 Education	Male	36910	2015	2016

Right Screenshot (Part1 First Second Half):

```

SELECT TOP (1000) [Subject_area]
, [Sex]
, [#Students]
, [Start_Academic_Year]
, [End_Academic_Year]
FROM [DW_Project].[dbo].[Part1_First_Second_Half]

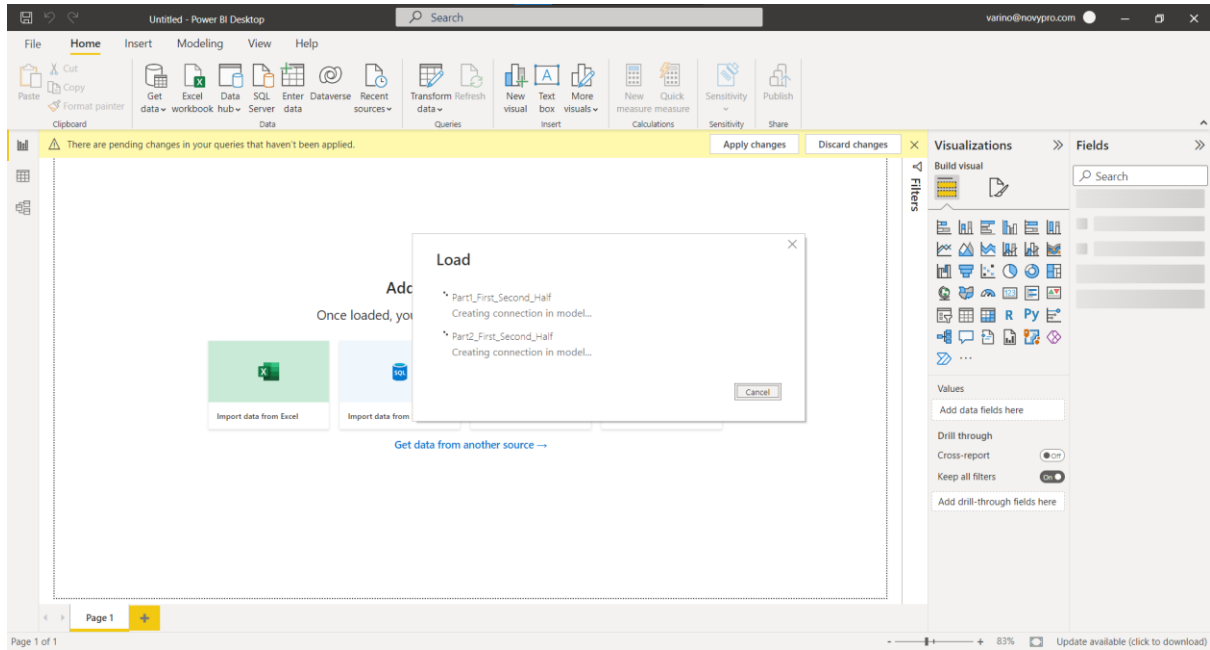
```

Subject_area	Sex	#Students	Start Academic Year	End Academic Year
1 Medicine_and_dentistry	Female	37335	2014	2015
2 Medicine_and_dentistry	Male	28660	2014	2015
3 Subjects_allied_to_medicine	Female	218910	2014	2015
4 Subjects_allied_to_medicine	Male	56815	2014	2015
5 Biological_sciences	Female	128775	2014	2015
6 Biological_sciences	Male	82570	2014	2015
7 Veterinary_science	Female	4495	2014	2015
8 Veterinary_science	Male	1405	2014	2015
9 Agriculture_and_related_subjects	Female	11855	2014	2015
10 Agriculture_and_related_subjects	Male	7350	2014	2015
11 Physical_sciences	Female	37080	2014	2015
12 Physical_sciences	Male	56660	2014	2015
13 Mathematical_sciences	Female	15955	2014	2015
14 Mathematical_sciences	Male	26440	2014	2015
15 Computer_science	Female	16040	2014	2015
16 Computer_science	Male	77170	2014	2015
17 Engineering_and_technology	Female	26955	2014	2015
18 Engineering_and_technology	Male	134240	2014	2015
19 Architecture_building_and_planning	Female	17365	2014	2015
20 Architecture_building_and_planning	Male	30885	2014	2015

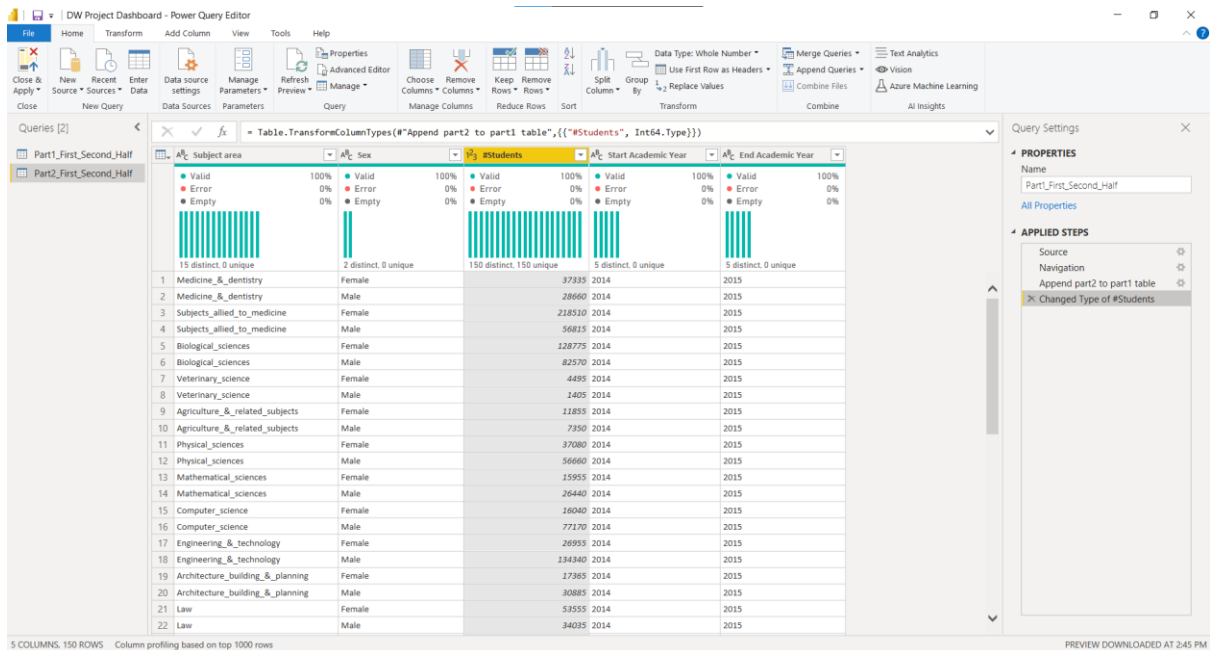
Both of them are now in SQL Server.

NOW Let's move to Power BI

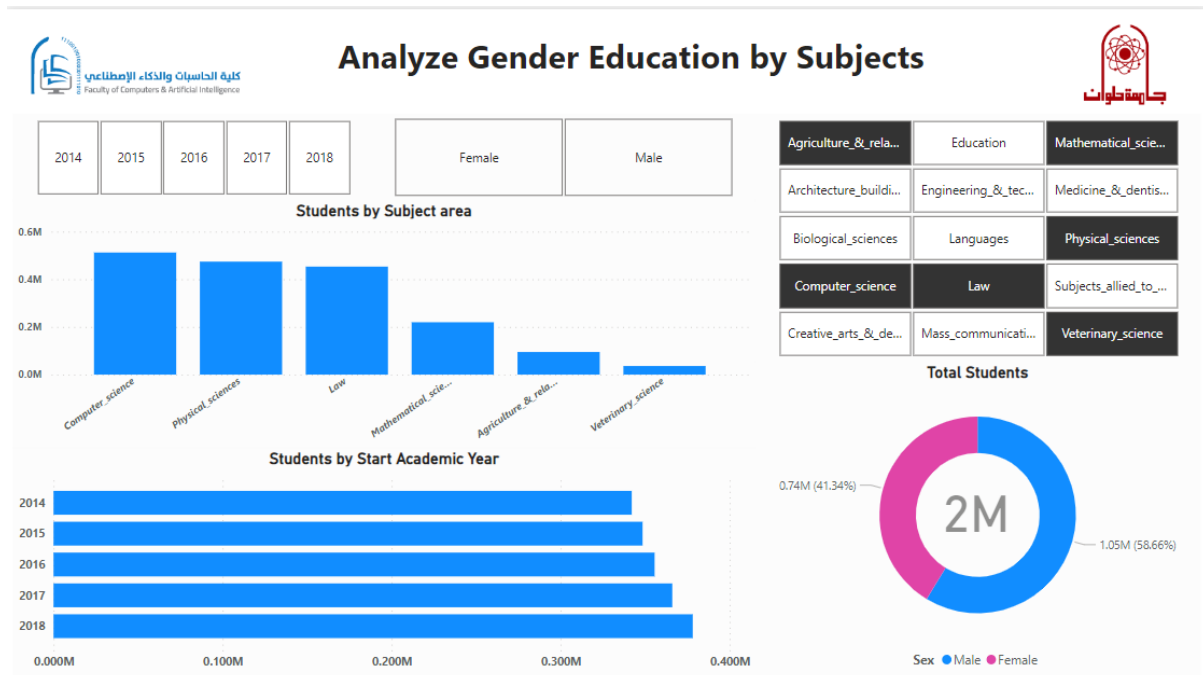
Load our data from SQL Server



Append the second table to the first table and adjust the type of #students



Dashboard



Compare students gender per subject.

Compare number of students over years.

Compare number of students and gender per subject and year.

Thank You!

Team Members:

201900555 فارينو الفريد فهمي

201900878 مينا طارق نجيب

201900232 بيشوي سمير لمعي

201900882 مينا مفيد مورييس

201900190 اندرو سعيد وهيب

201900598 ماركو ماجد فؤاد