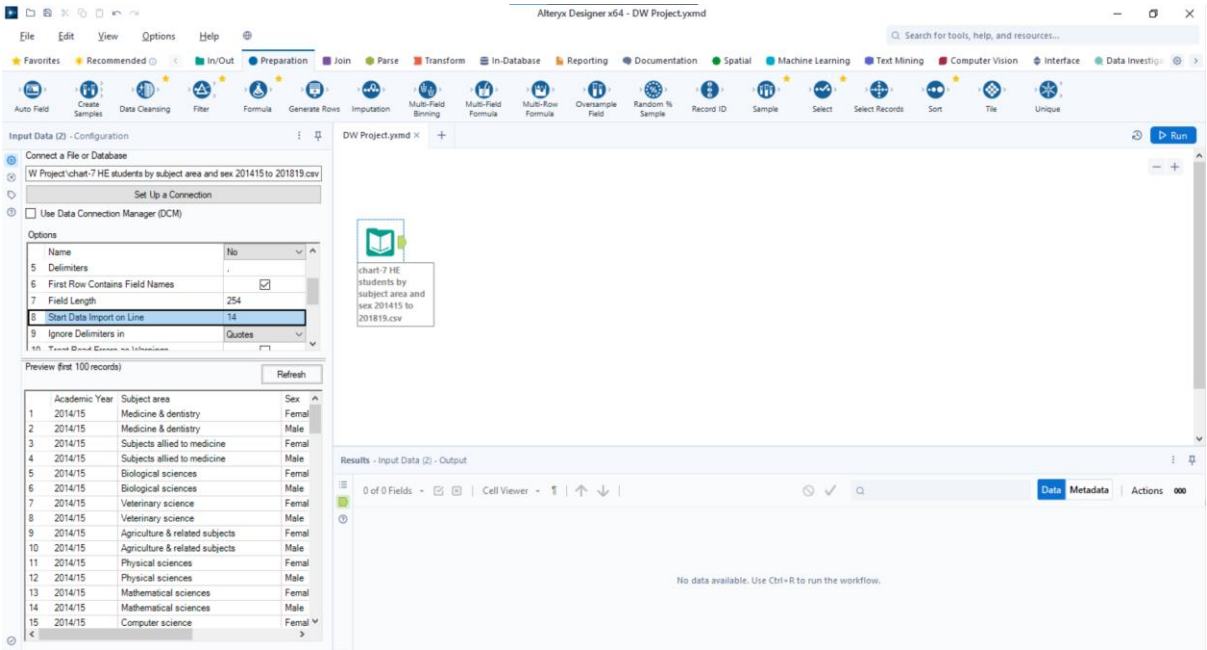
Dataset source: Higher Education Student Data

(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

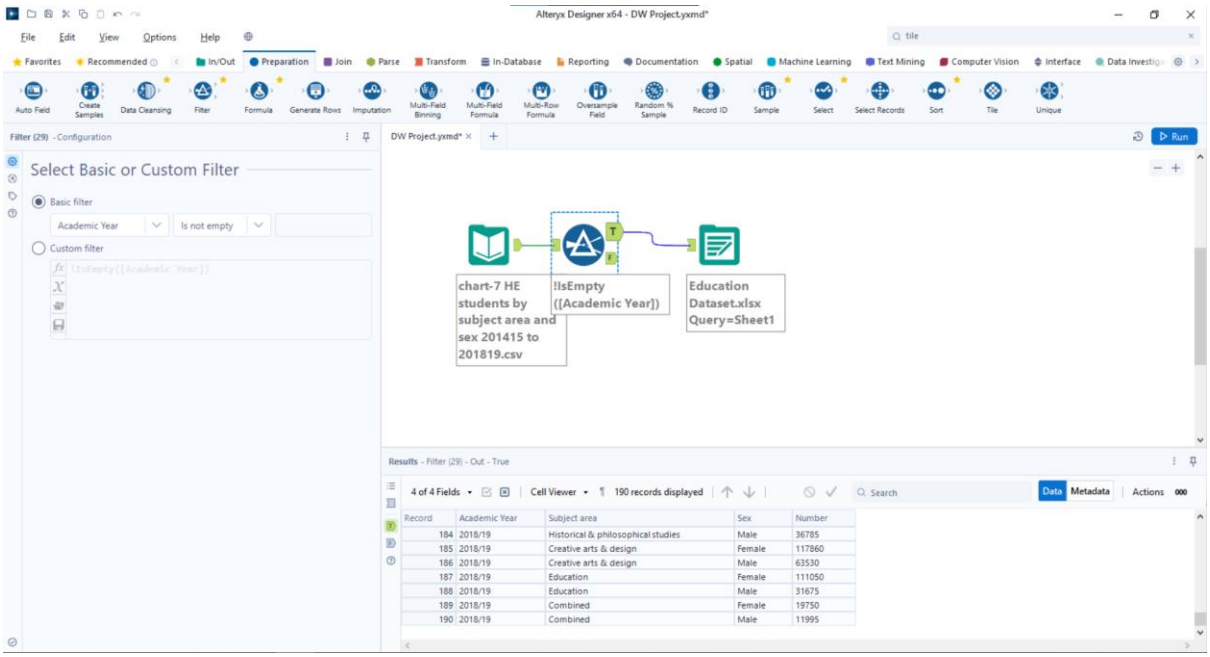
Columns: AcademicYear, SubjectArea, Sex, Number

Analyze Gender Education by Subject



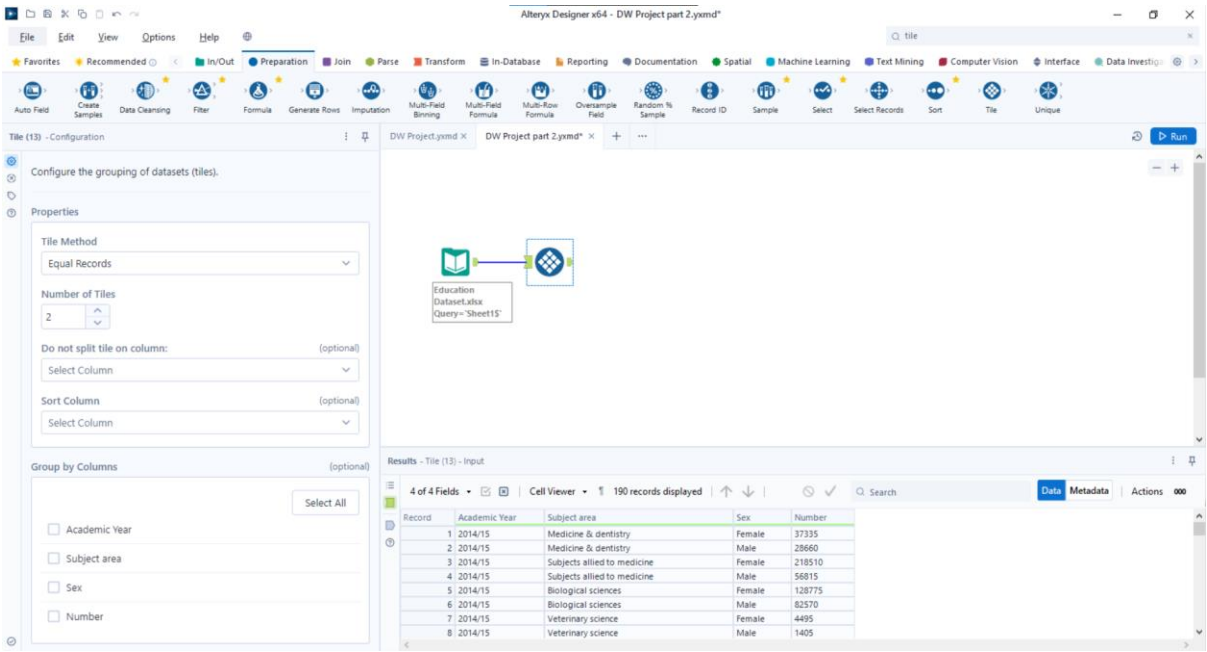
(Input Data)

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

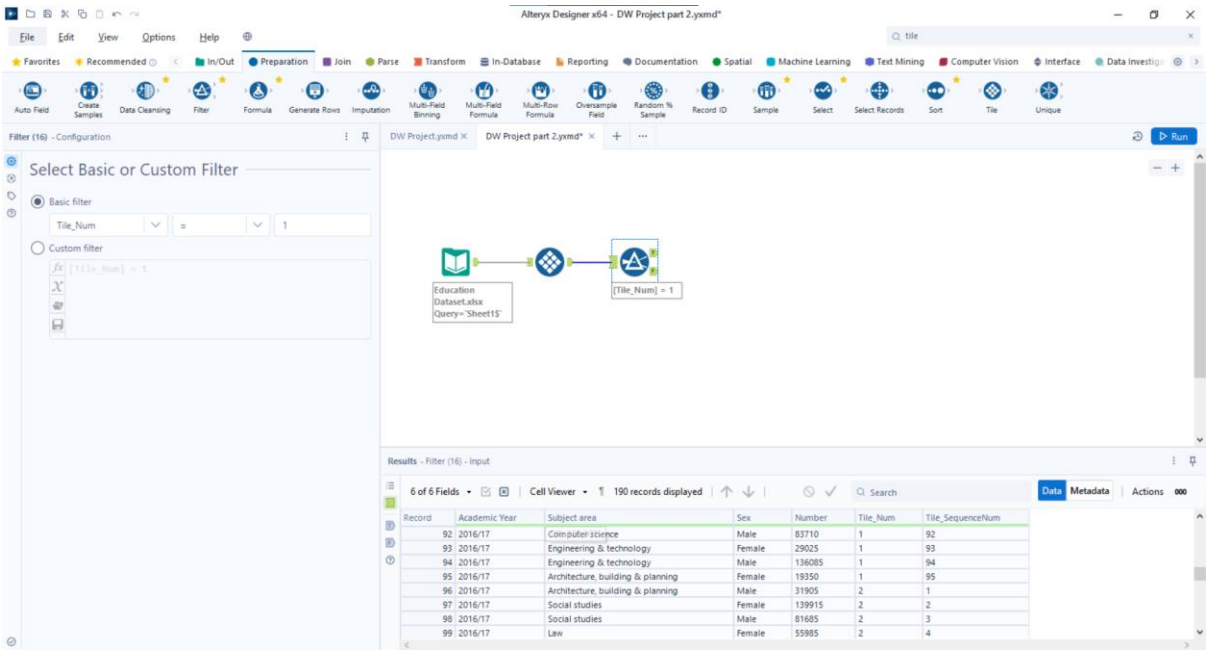


We found some records doesn't contain any data except for the sex and this was preventing us from

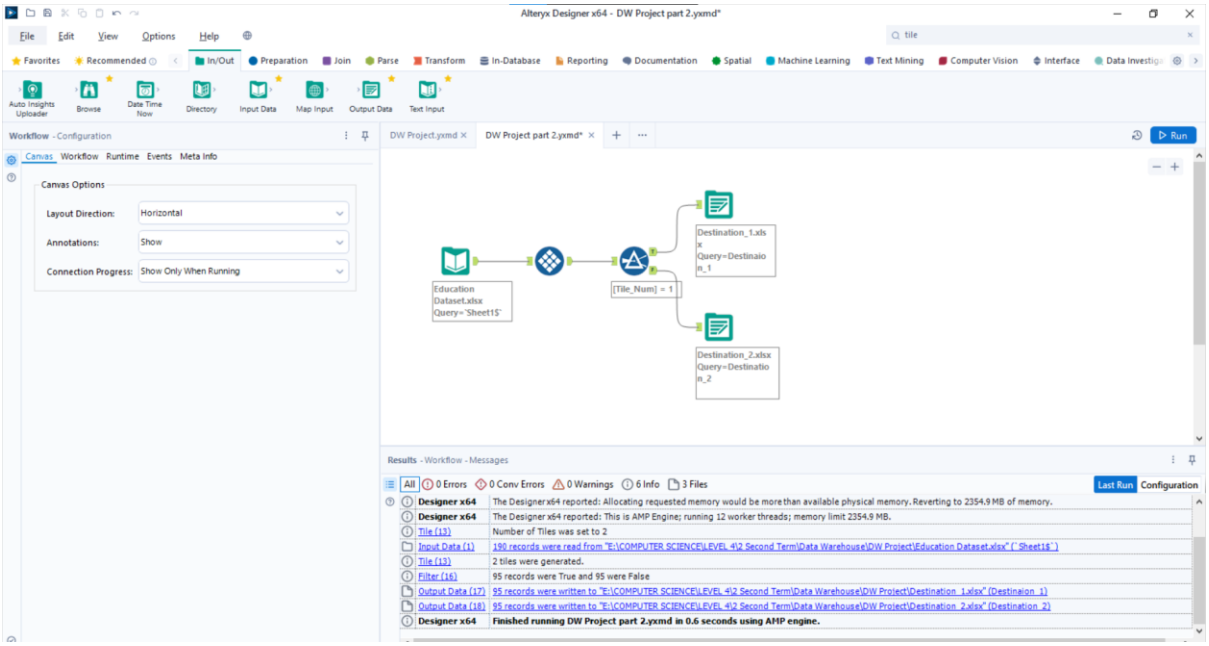
correctly splitting the data, so we filtered the data on academic year when it is not empty true then output it in Education Dataset file to work on it. **(Filter)**



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

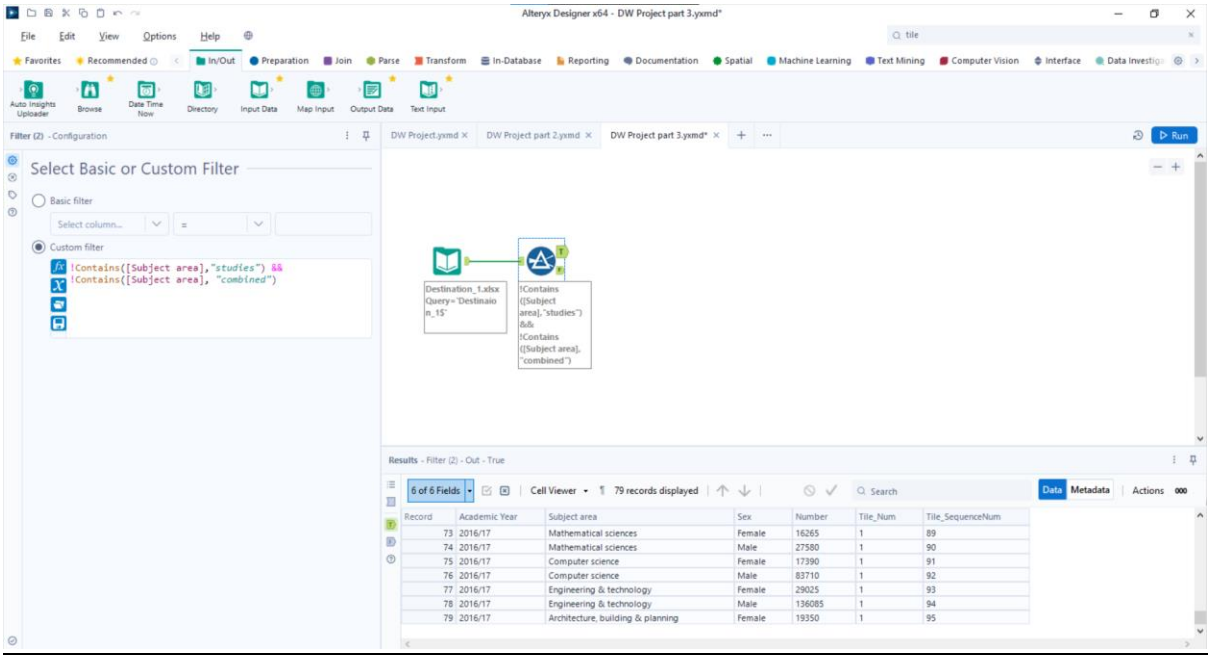


Using filter on Tile_Num we will export the data into 2 excel files the first half is in Destination_1 if filter is True and the second half is in Destination_2 if it is false.

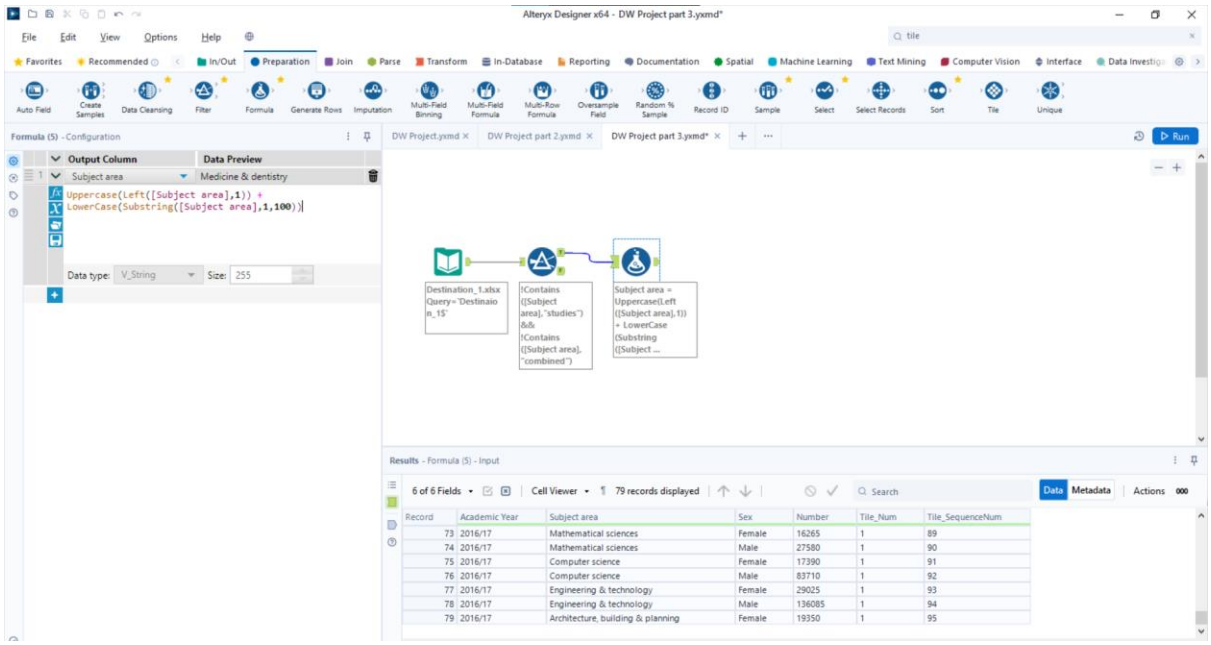


Transformations

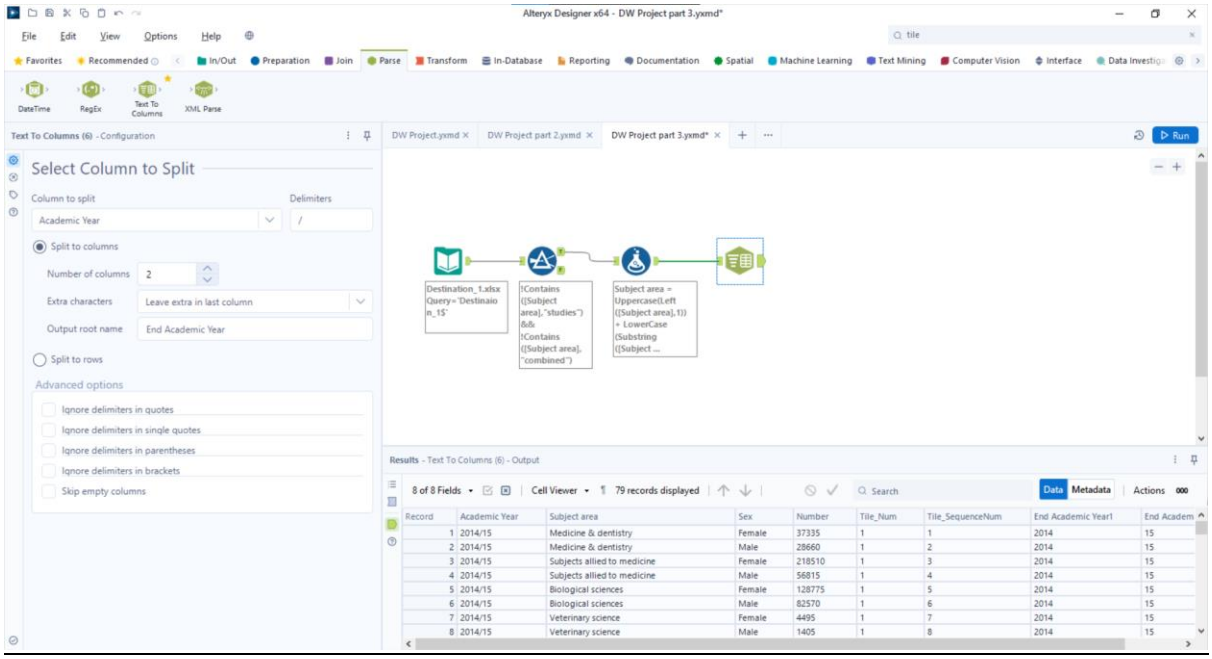
1: filter records on specific condition (custom condition: **Filter** out subjects that contain either studies or combined)



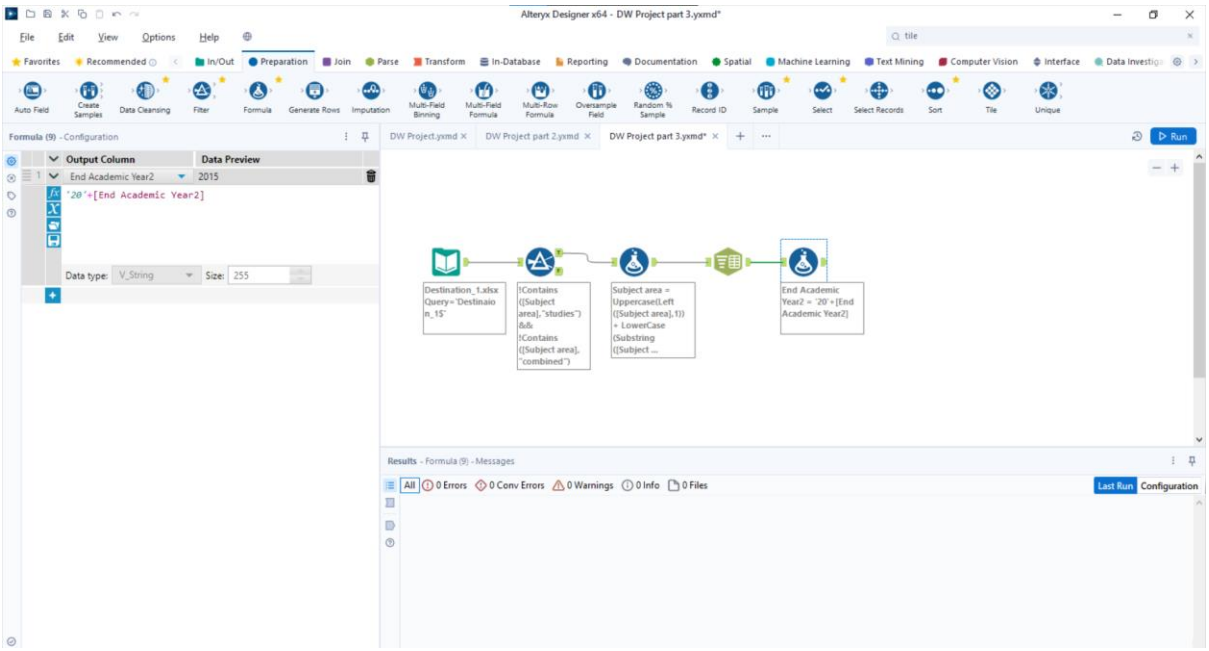
2: choose any string column and uppercase the first character only (Using **Formula**)



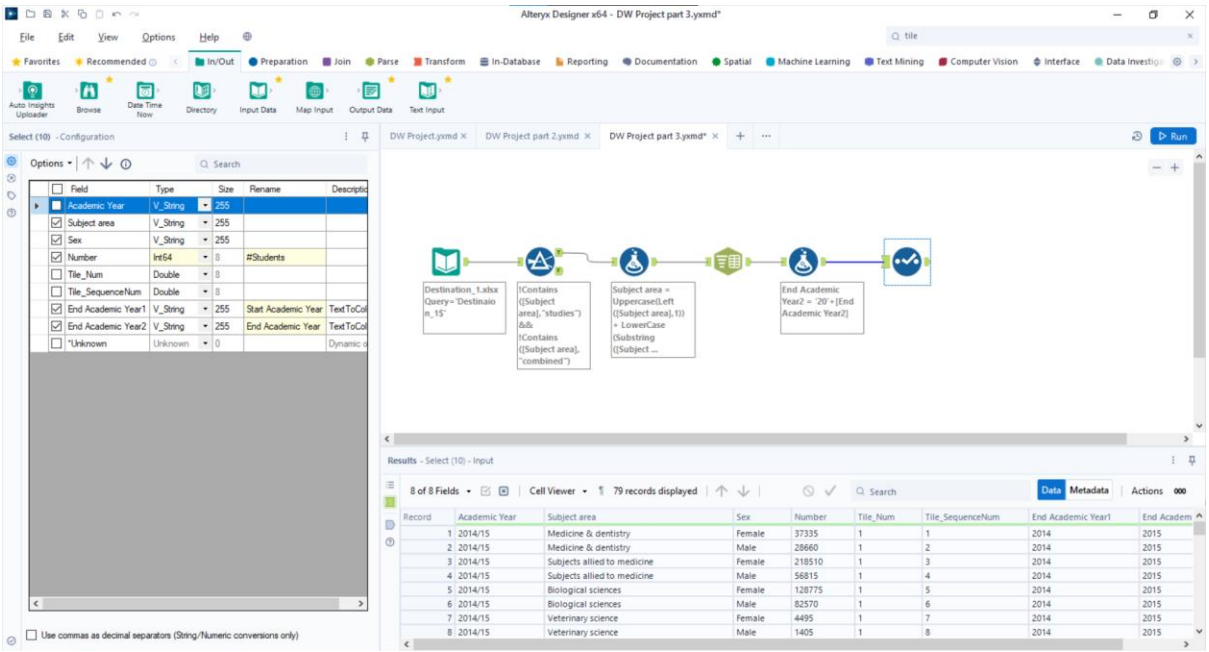
3: split any column into many columns (Using **Text To Columns**)



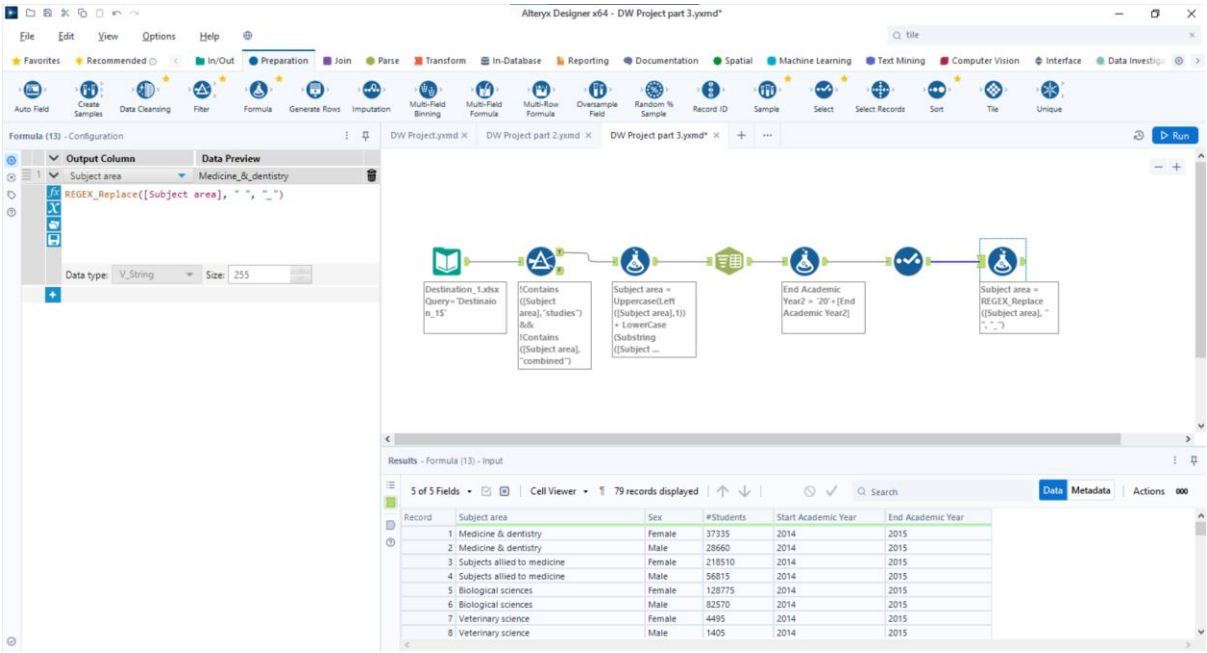
We wanted to add ‘20’ before the End Academic Year2, So we used **Formula:**



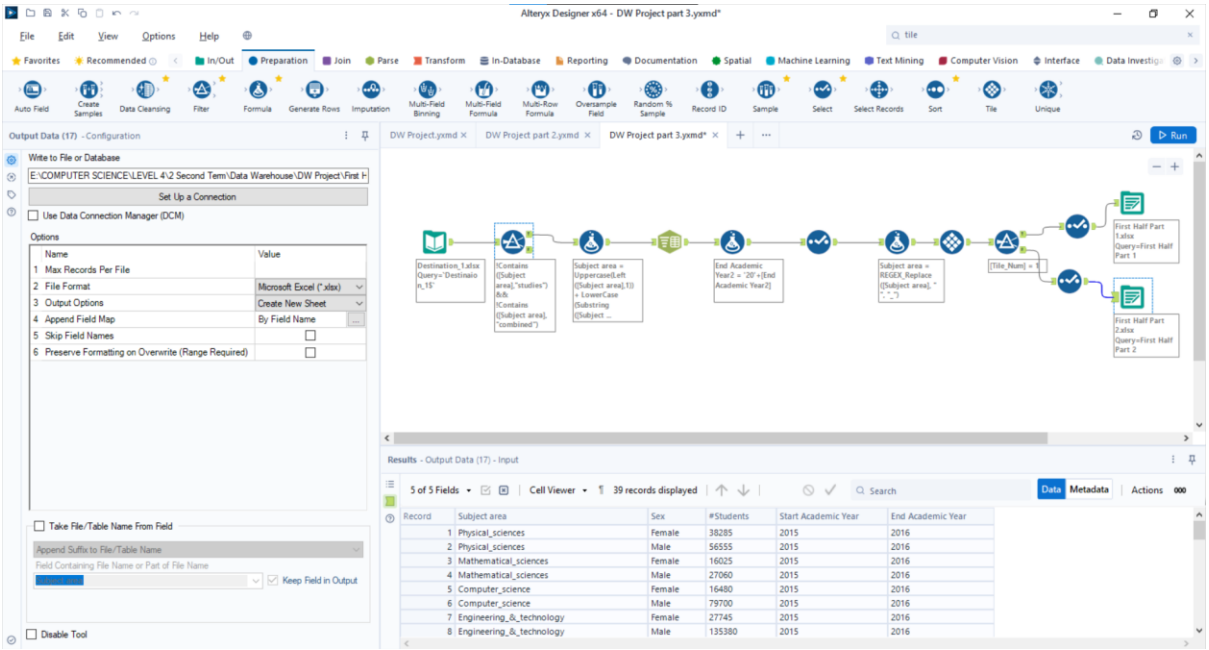
Using **SELECT**, we removed the unused columns, renamed the new columns and adjusted the type of students number.



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

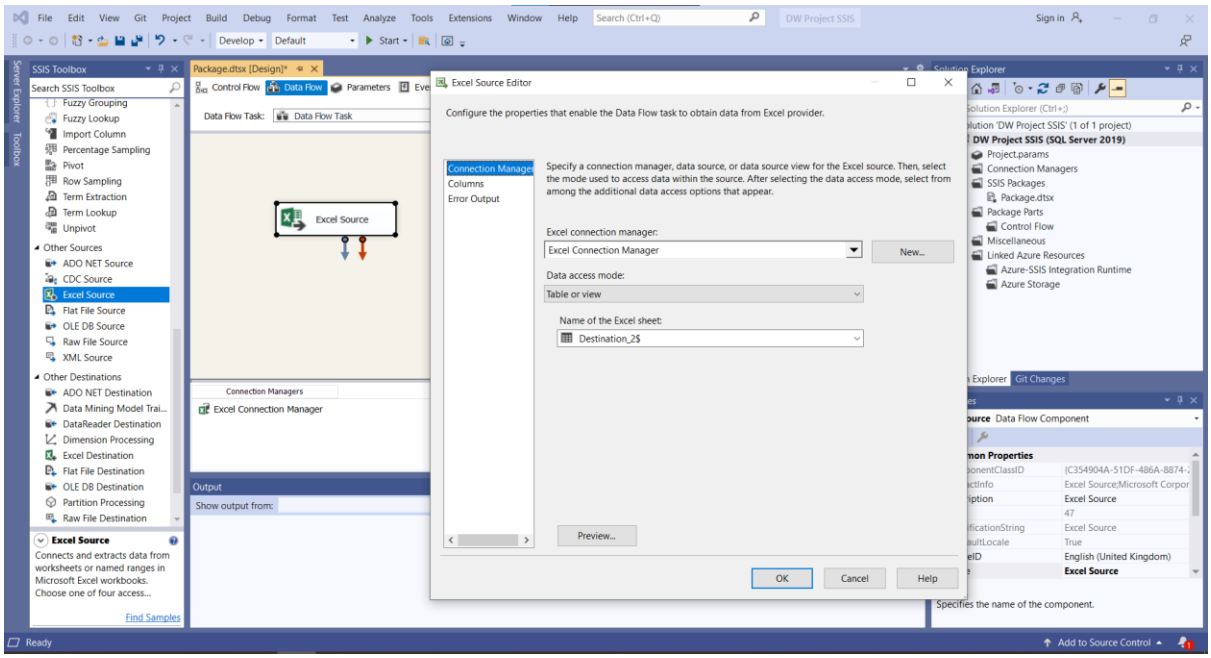


Divide the current Destination1 aka. First Half again into two parts. (using **Tile, Filter, Select, Output Data**)



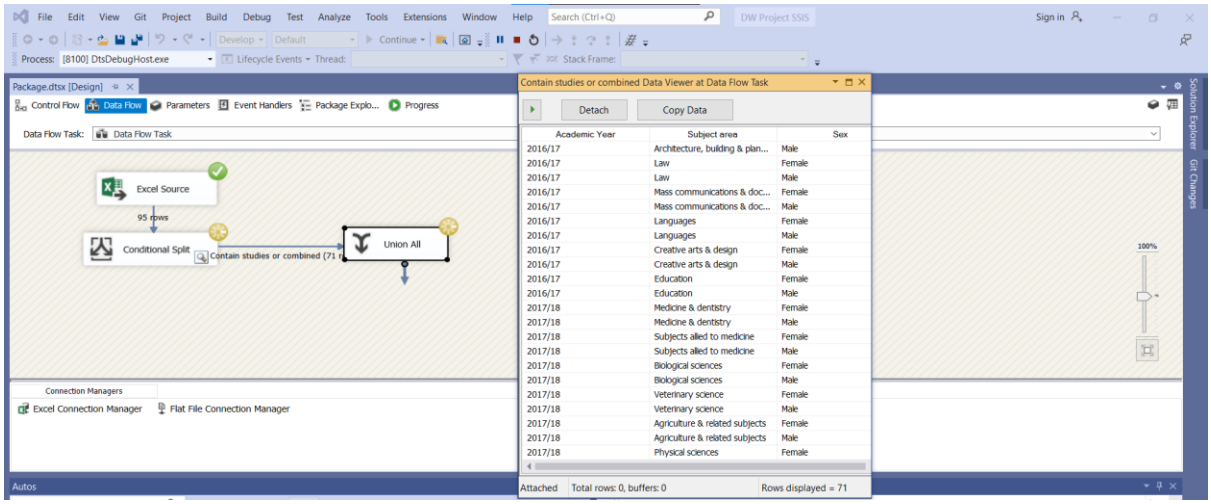
NOW Let’s move to SSIS

Import Desination_2 using **Excel Source**

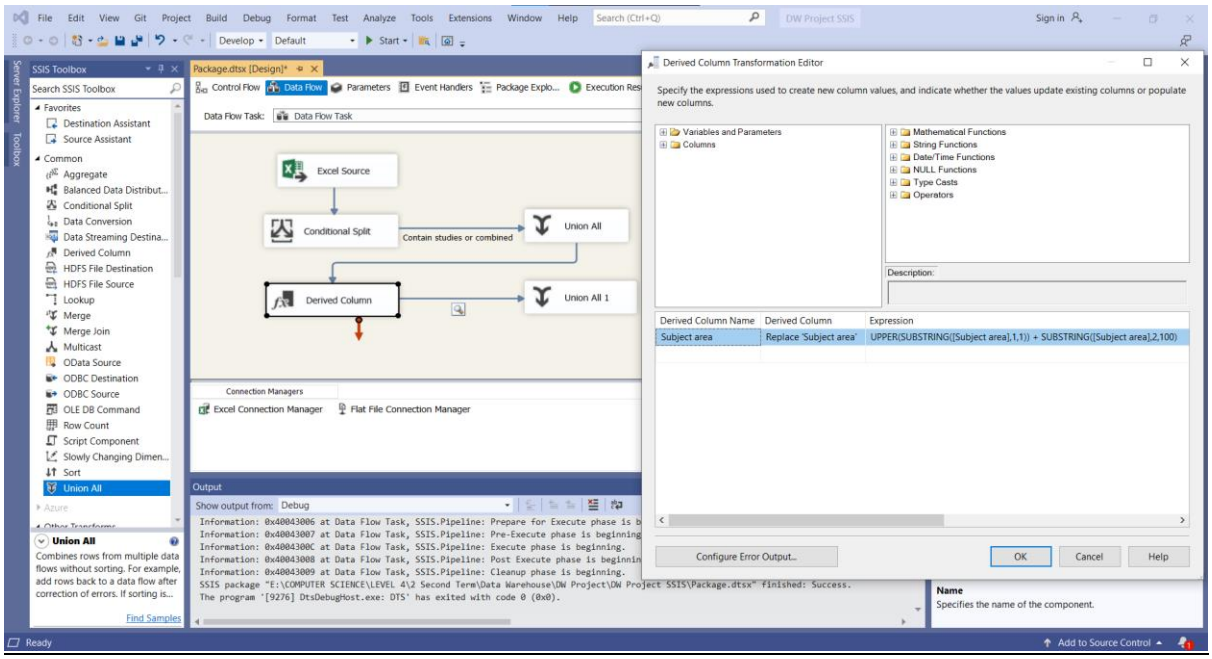


Transformations

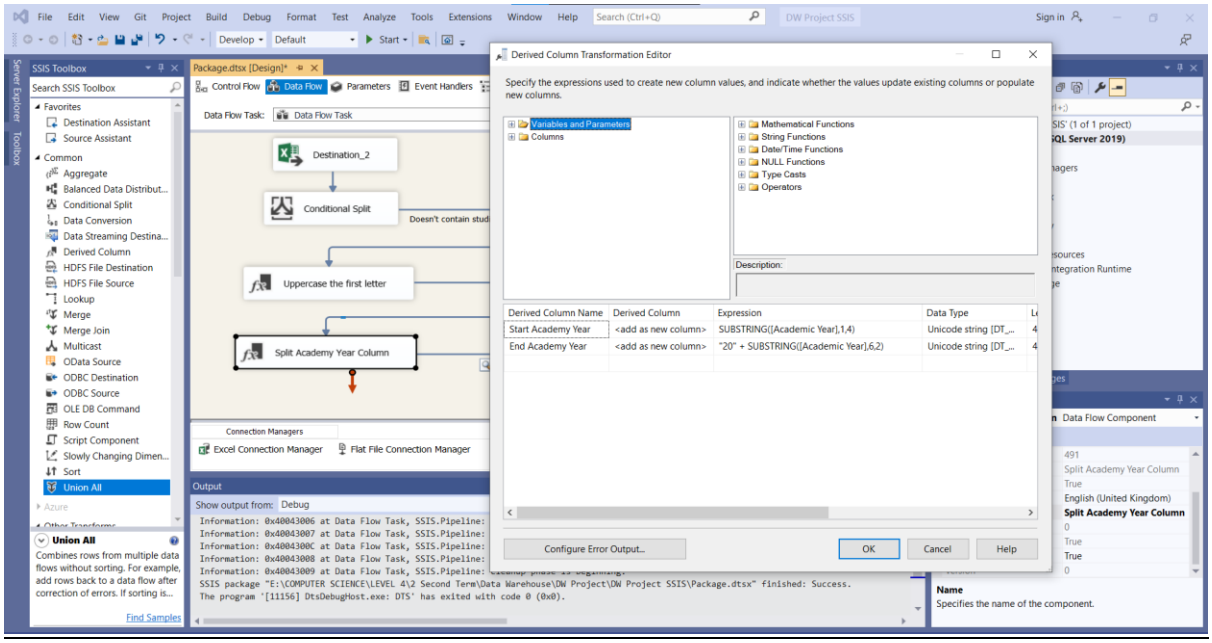
1: filter records on specific condition (custom condition: Filter out subjects that contain either studies or combined using **Conditional Split**)



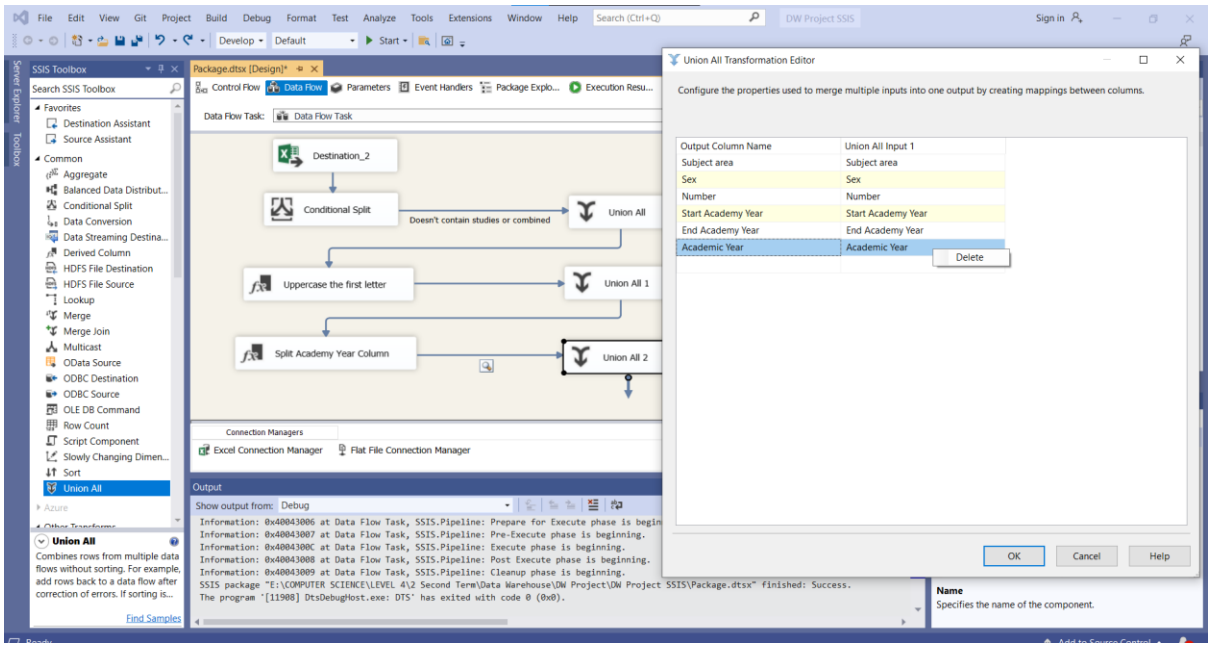
2: choose any string column and uppercase the first character only (Using **Derived Column**)



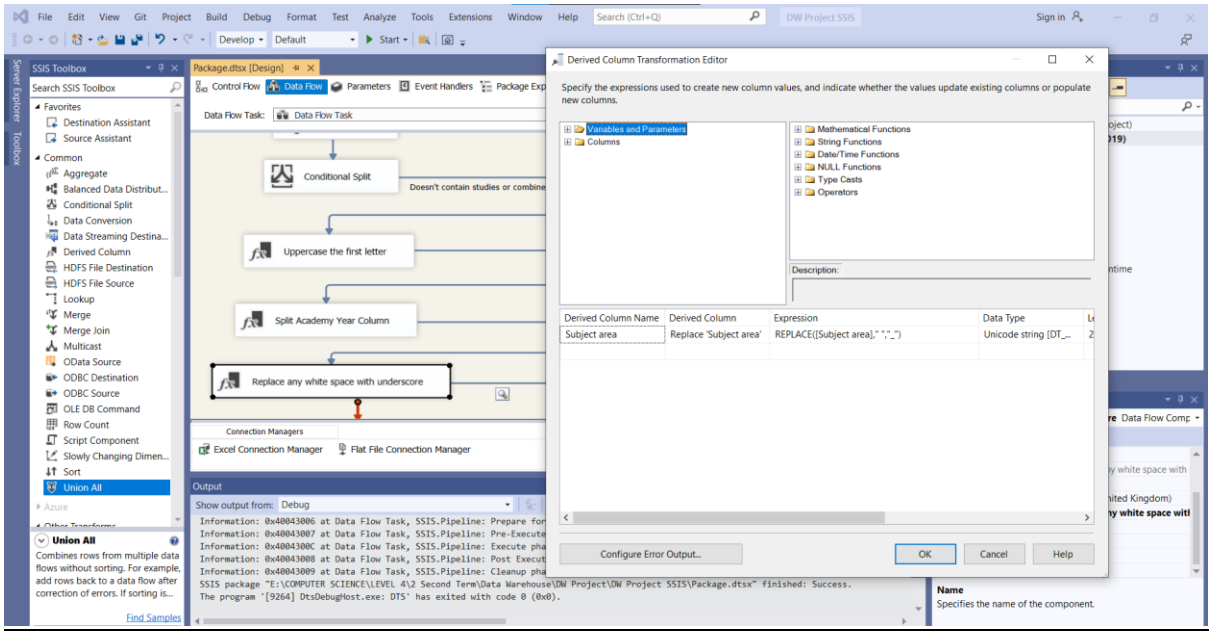
3: split any column into many columns (Using **Derived Column**)



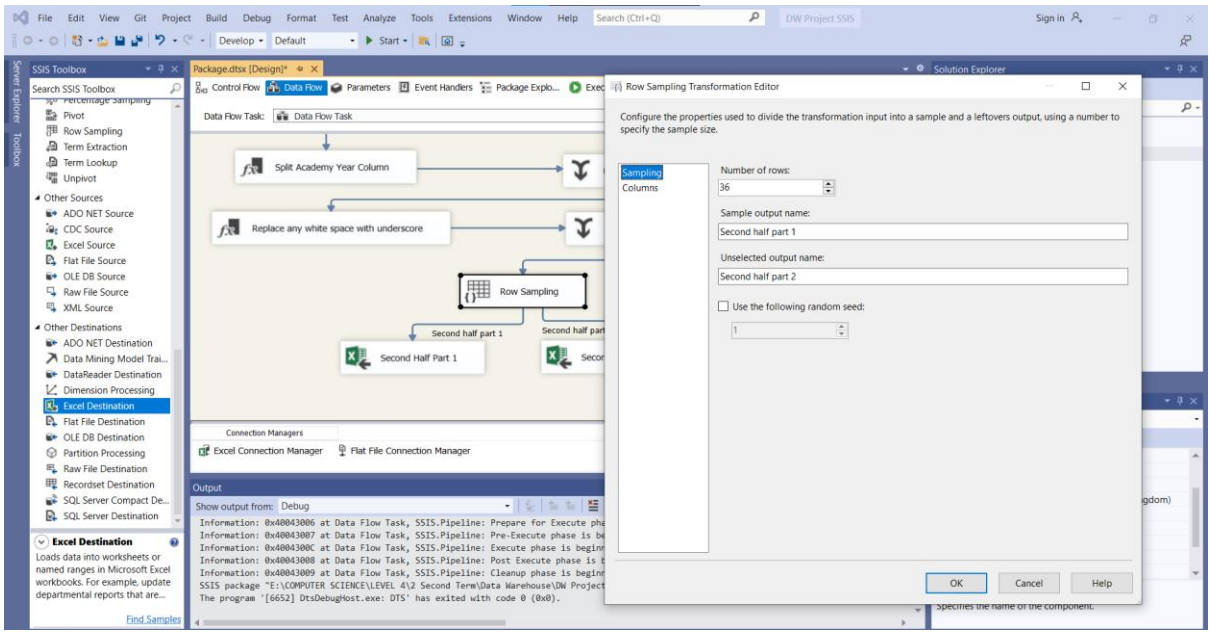
we deleted the unused columns:



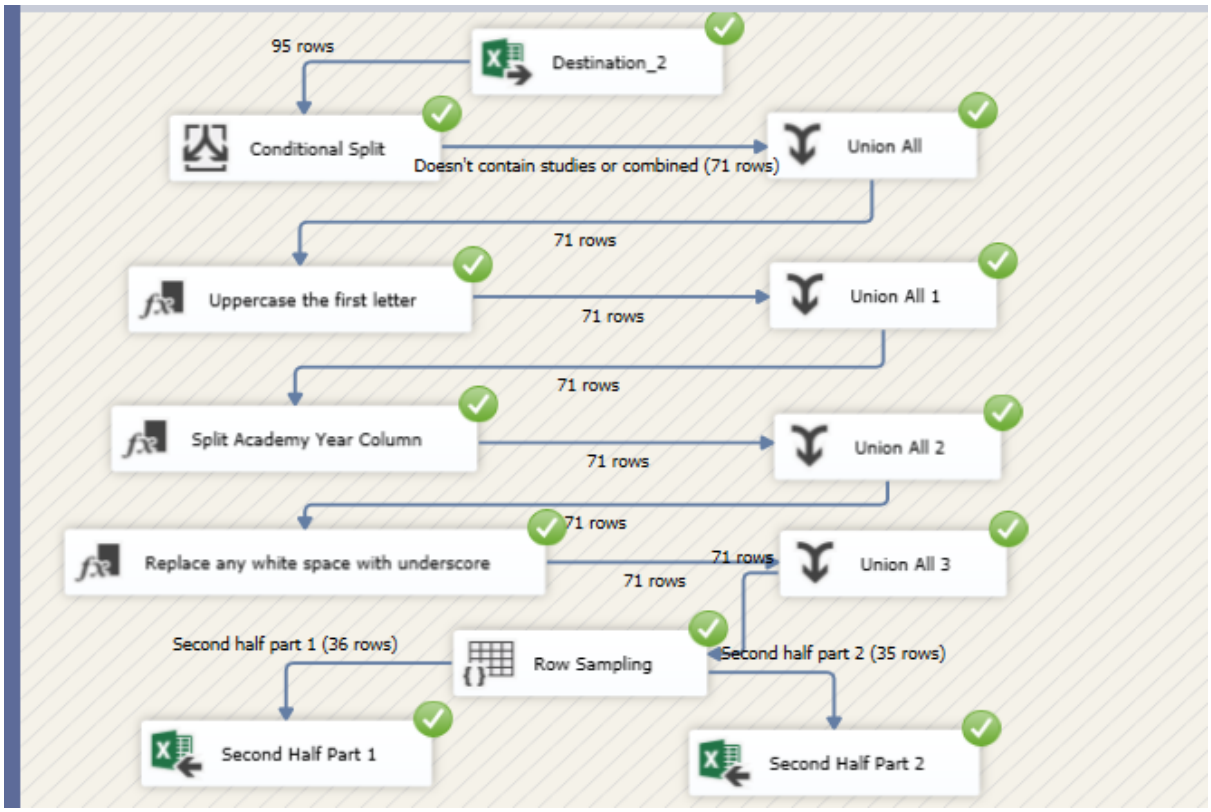
4: replace any white spaces with underscore for any string column (Using **Derived Column**)



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

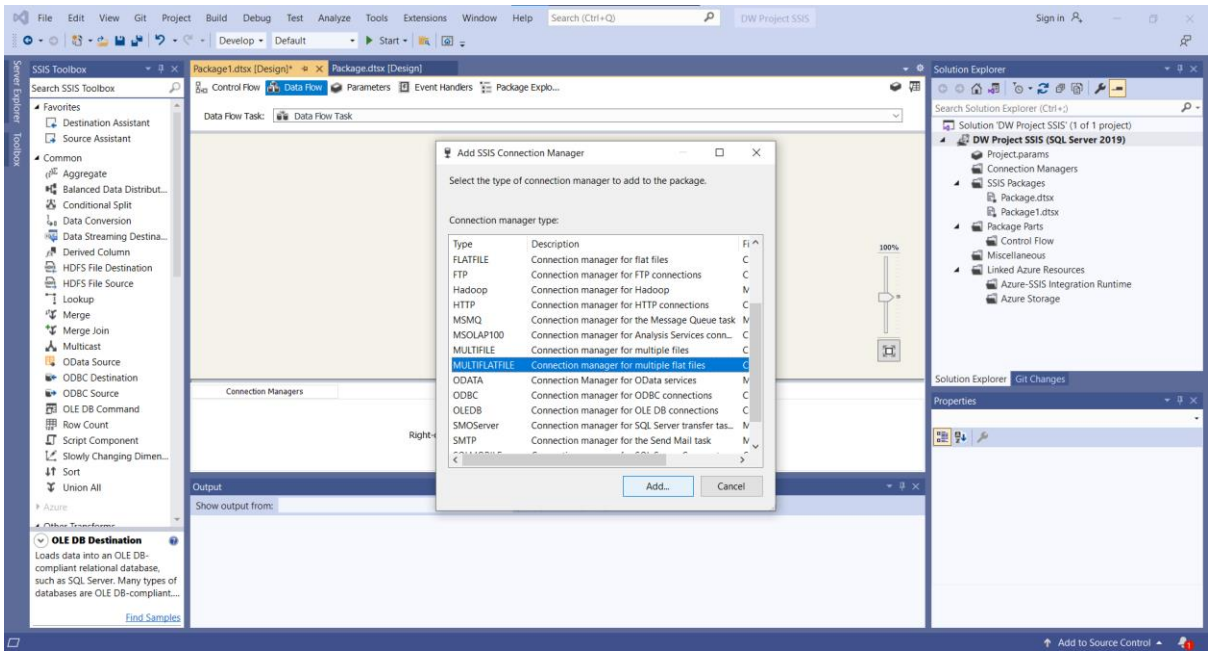


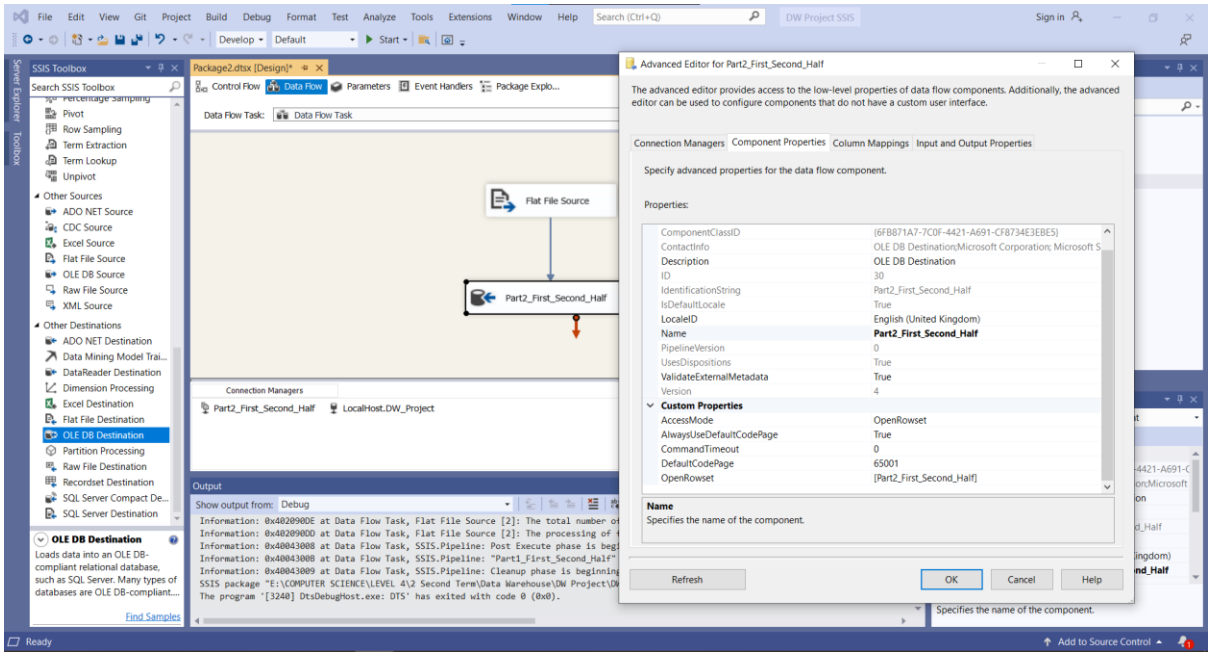
Final Result



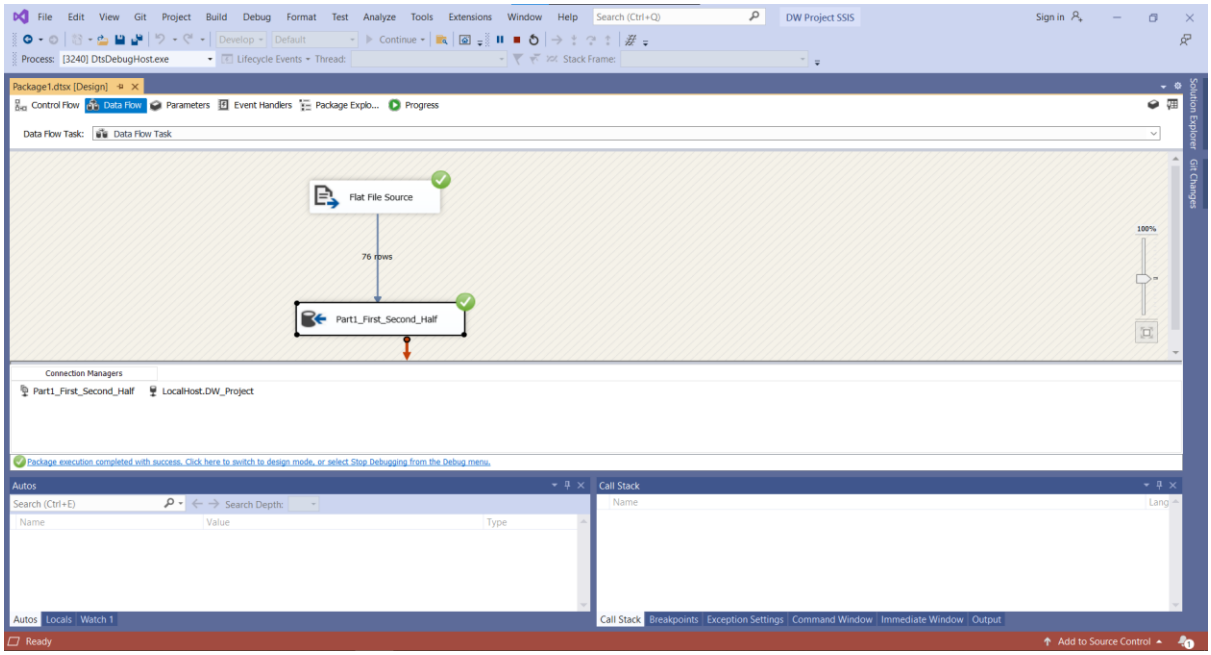
Load part 1 from first and second halves in destination 1 in SQL Server:

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



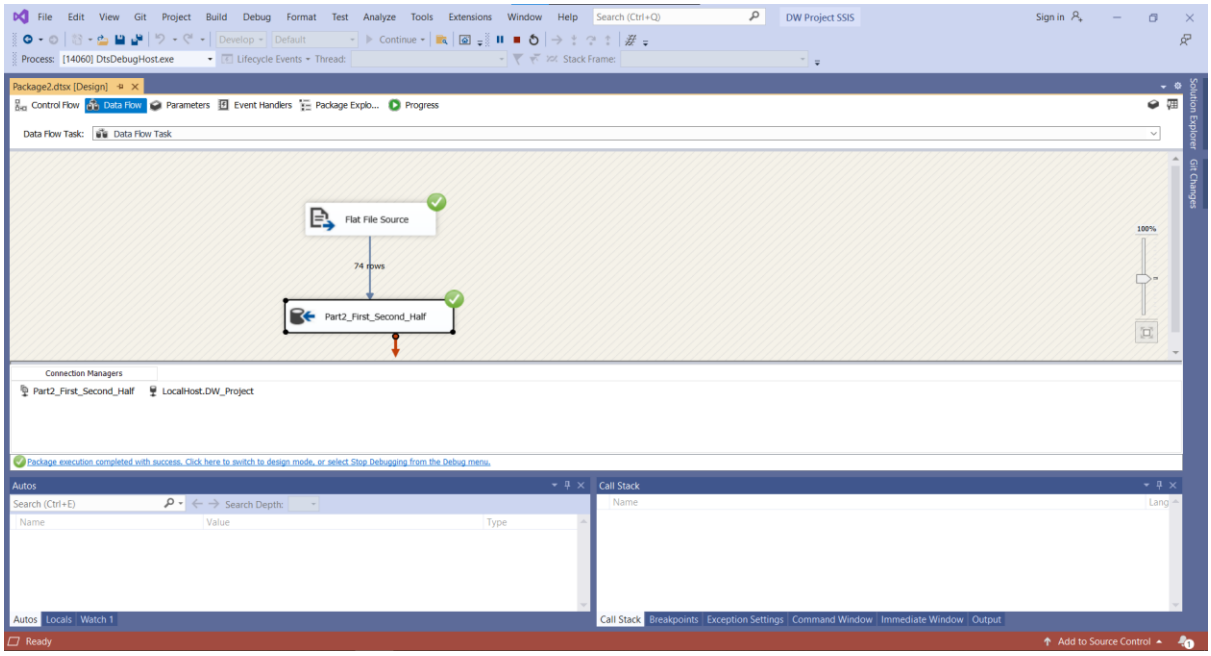


We change the OLE DB Destination advanced editor setting DefaultCodePage to 65001 and make AlwaysUseDefaultCodePage True to solve the column cannot be processed because more than one code page. (From: Stakeoverflow)

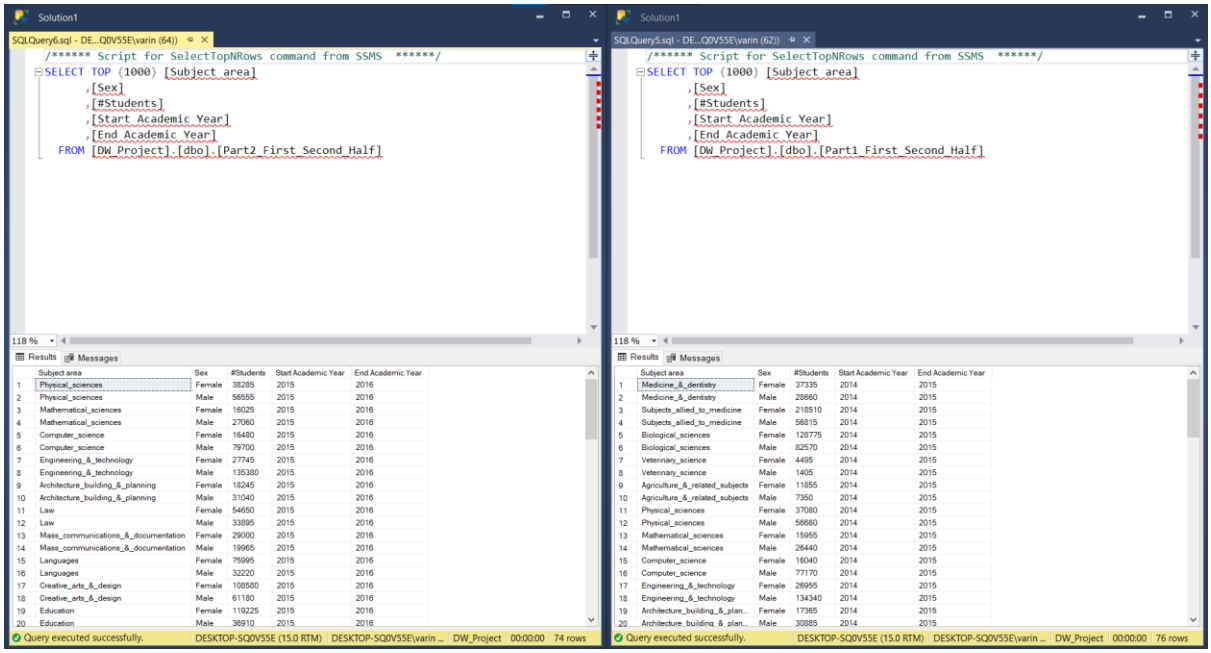


and here it is 76 rows (40 rows from First Half Part1 & 36 rows from Second Half Part1)

Same steps to load the Part 2:



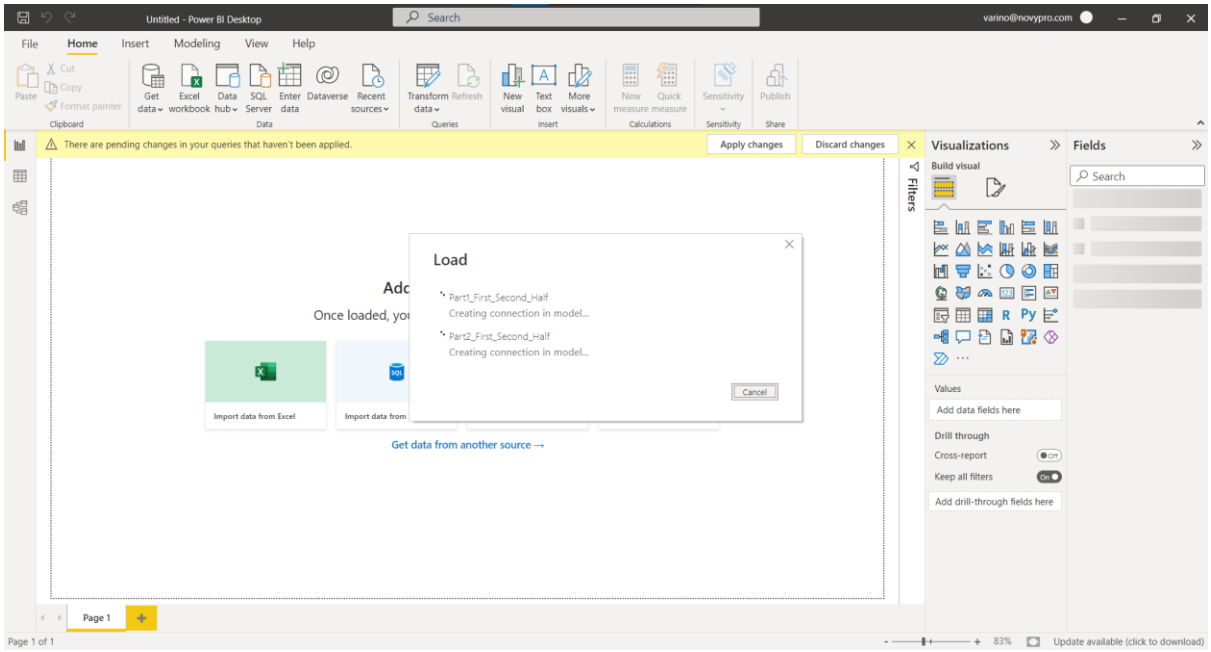
and here it is 74 rows (39 rows from First Half Part1 & 35 rows from Second Half Part1)



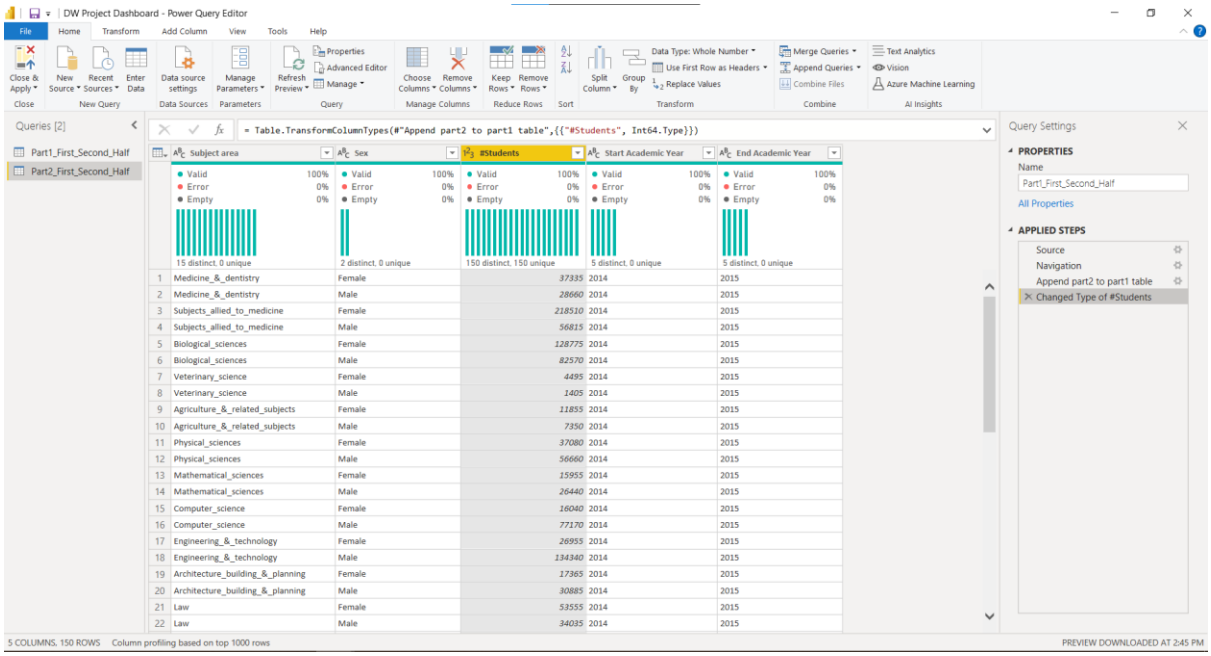
Both of them are now in SQL Server.

NOW Let’s move to Power BI

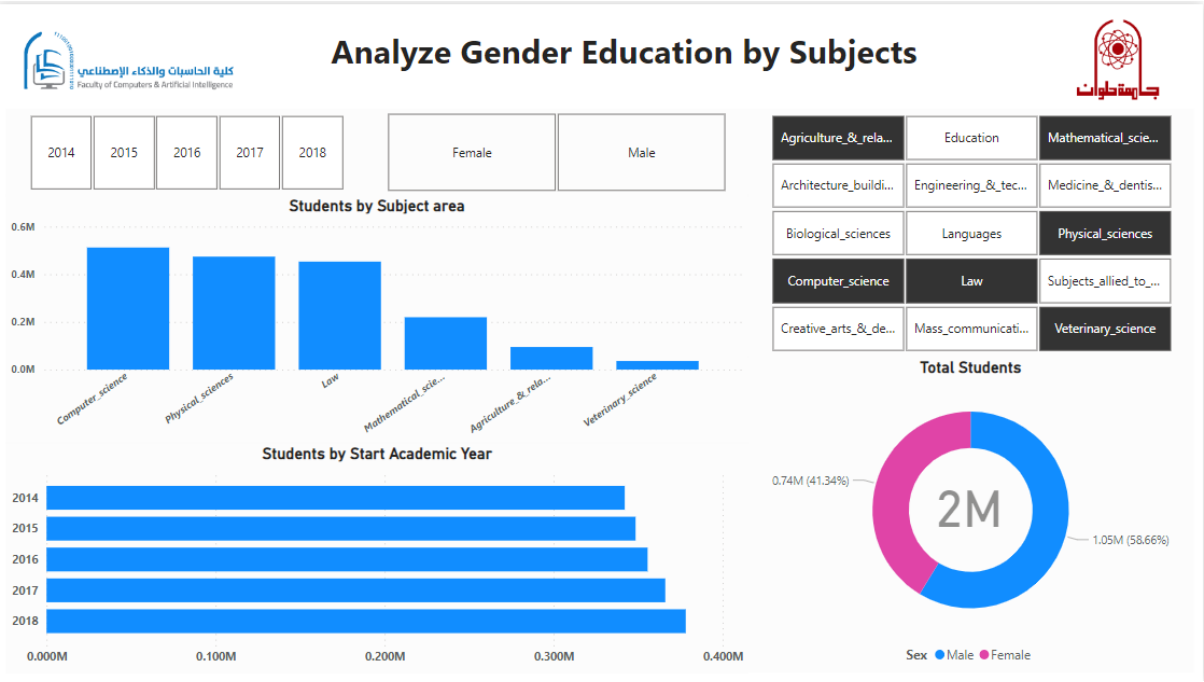
Load our data from SQL Server



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 فهمى

201900878 مينا طارق نجيب

201900232 بيشوي سمير لمعى

201900882 مينا مفيد موريس

201900190 اندرو سعيد وهيب

201900598 ماركو ماجد فؤاد