UNIVERSITÀ DI PISA

LDS Project – part 2

Group 8:

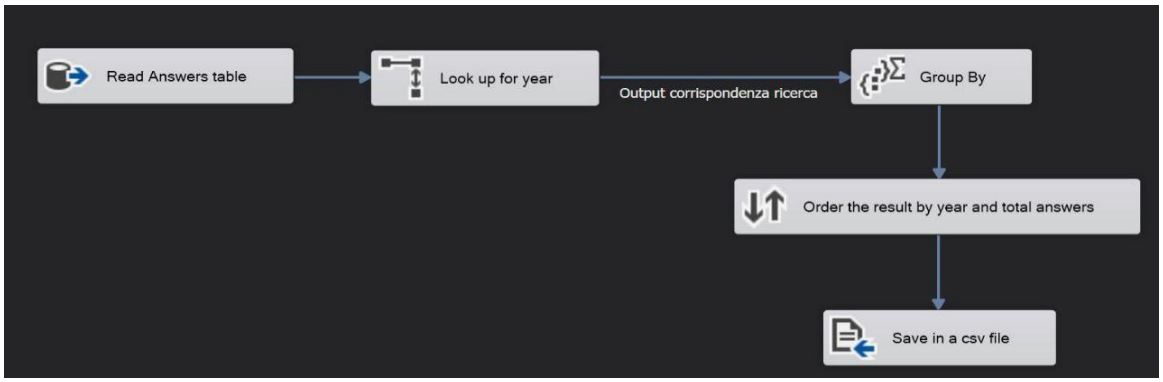
Sabatini Emanuele

Volpi Federico

Assignment 0

For every year, the users ordered by total number of answers

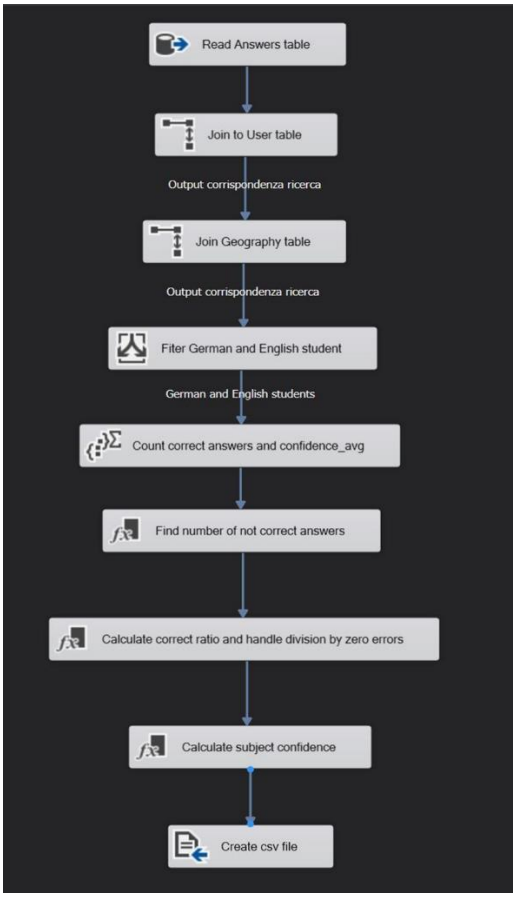
We read Answers table by taking answerid, userid, dateid columns. A Lookup is done for the year (search) in the table. Then, we grouped by year, user id and count answers given. Then, we have done a sort where first we have year ascending and total answer descending for each user id, then we export to csv.



Assignment 1

> For every subject, the *correctness confidence index* is defined as the ratio between correct answers and wrong answers multiplied by the average confidence. Provide such index for every subject, considering only German and English students.

We read Answers table, do a Lookup as a join on user table and then one on Geography. Then we select only German and English students with a filter, through the appropriate expression. We do a groupby subjectid, sum iscorrect (which has value 0 and 1) to get the number of correct answers. We continue by counting iscorrect on tot_answers to get the number of wrong answers and confidence avg for the average. At this point, we introduce a derived column finds the wrong answers, total answers minus correct answers. Then we do an operation that is used to create a correct ratio where we say if the value of not_correct is zero, we put the value of iscorrect, otherwise the ratio of correct to not_correct. In this way we also handle divisions by zero to avoid errors, then we do another derived column where we have correct ratio by confidence, as requested in the assignment, with the round. We decided to use three derived columns, to better handle any problems and make each step clearer.



Assignment 2

> For each region, the percentage of correct answers with respect to the country of origin.

We have imported an OLE DB Origin from the Answers table, then created two Lookups. Firstly, one for the User table and the other for the Geography table to join them together and get the iscorrect attribute. Then, we have used a multicast tool for two different group by, the former by country_name and region with sum function on iscorrect, the latter by just country_name and the count function on iscorrect. This particular set-up for the second group by was chosen based on the interpretation of the assignment as to find the percentage of correct answers with respect to the total answers of the country of origin, not the total correct answers. Then, with the merge join we have merged together the different results of the aggregation and with a derived column computed the percentage.