UNIVERSITÀ DI PISA

LUCA PALLA - 533605

GIULIA CALVO - 544434

LDS- GROUP 21_DB: PART 3

ASSIGNMENT 0

In order to create a new cube, the first step after connecting with our data and create views, was to construct dimension which refers to answer's fact table. They are:

- Date: DateId, Date, Day, Month, Quarter, Year.
- Organization: OrganizationId, GroupId, QuizId, SchemeOfWorkId.
- Subject: SubjectId, Description.
- User1: UserId, DateId, GeoId, Gender, Region, CountryCode, Continent.
- Geography: GeoId, Region, CountryCode, Continent.

As it is possible to observe, User1 also contains geography features. This was necessary to overlap a problem which didn't permit to correct visualize the geography dimension inside the cube. When dimension was created, the hierarchies between attributes were considered. Date and Geography (and so also User1) presented some attributes which were hierarchies. In order to exploit this fact, DayMonthQuarterYear and RegionCountryContinent hierarchies were created.

The next step was the creation of the cube. Using the dimension created, the answer's cube was created. As default, two group of measures were created in order to answer the query requested in subsequent assignment. They are:

- Answer group which contains "Conteggio di Answer" and "Is Correct" (the sum of correct answer).
- User1 group which contains "Conteggio di User".

Before start describing next assignments, is preferred to discuss all the changes and data manipulation made on the cube in order to answer queries with MDX language or create useful dashboard with PowerBI.
As first, new columns were added by "Nuovo calcolo denominato" function in tables' views; they are:

- IsIncorrect: it contains boolean values; 1 when answer is incorrect and 0 when is correct.
- Gender_BI: It contains 3 distinct values; F when Gender was 1, M when Gender was 2 and "Not specified" when Gender was 0.
- Continent_BI: It contains 3 different values which identify the 3 different continents in our data. This column was created in order to overlap a visualization problem in PowerBI (before NA was considered as Nambia instead of North America).
- Region_BI: It was created due to problems in localization of some regions such as "Atlantic provinces" renamed "Atlantic Canada" and "south island central" renamed "Te Waipounamu" both suggested by Wikipedia.

All the columns which end with "_BI" there will be useful in the last two assignments.
As second a new hierarchy were generated including the new columns for geography and it was called "Geo_BI". As for features ending in "_BI" it will be useful in the last assignments.
As last step a new measure was created; it was named "Is Incorrect" and it is the sum of "IsIncorrect" column.

ASSIGNMENT 1

Show the percentage increase or decrease in correct answers with respect to the previous year for each student.

- For this query 3 solutions were implemented to practice with MDX and to give different point of views. Although they are different, they shared the same case statement which highlights 4 different situations:

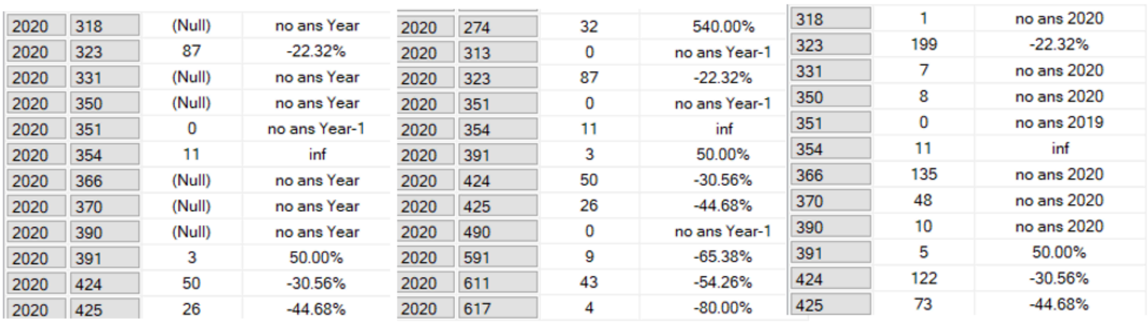
- The first is when the currentmember (or 2020 in third solution) is empty. In this case will be returned 'no ans Year' (no ans 2020).

- The second is when the prevmember (or 2019 in the third solution) is empty. In this case will be returned "no ans Year-1" (no ans 2019).

- The third case is when the absolute difference in correct answer between one year and the previous one is 0. In this case 0% will be returned. This case was necessary to be specificized because there were situations where students had 0 answer correct in both years and this returned an indefinite form of the type 0/0.

- The last one is when data are present for both years and so we just performed the formula for compute the ratio between correct answer of the current year and the previous one.

- A special mention is also needed for 0 division results. We decided to not custom the retuned value by MDX and so to keep "inf" as answer. This decision was made in order to better distinguish this situation from the others above. In fact, "Inf" tell us that a student which gave no correct answer in the previous year had this year gave some correct answer. If we decided to returned a given percentage, we will not be able to highlights these students which inevitably will be confused with others.



2020	318	(Null)	no ans Year
2020	323	87	-22.32%
2020	331	(Null)	no ans Year
2020	350	(Null)	no ans Year
2020	351	0	no ans Year-1
2020	354	11	inf
2020	366	(Null)	no ans Year
2020	370	(Null)	no ans Year
2020	390	(Null)	no ans Year
2020	391	3	50.00%
2020	424	50	-30.56%
2020	425	26	-44.68%

2020	274	32	540.00%
2020	313	0	no ans Year-1
2020	323	87	-22.32%
2020	351	0	no ans Year-1
2020	354	11	inf
2020	391	3	50.00%
2020	424	50	-30.56%
2020	425	26	-44.68%
2020	490	0	no ans Year-1
2020	591	9	-65.38%
2020	611	43	-54.26%
2020	617	4	-80.00%

318	1	no ans 2020
323	199	-22.32%
331	7	no ans 2020
350	8	no ans 2020
351	0	no ans 2019
354	11	inf
366	135	no ans 2020
370	48	no ans 2020
390	10	no ans 2020
391	5	50.00%
424	122	-30.56%
425	73	-44.68%

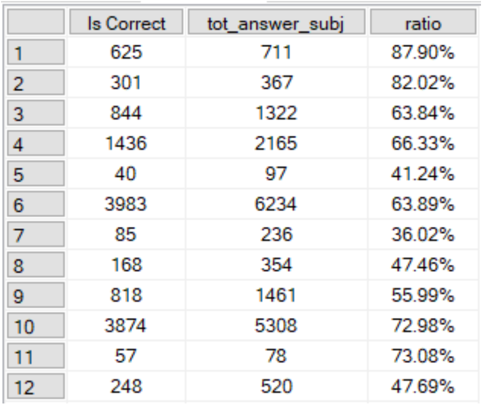
From left to right a small visualization of result for answer "assignment 1", "assignment 1.1", "assignment 1.2"

In the first solution "set" was used in order to restrict the years to visualize. As showed in the first image all the 4 kinds of results are present. In this case "inf" say to us that the student 354 the previous year gave 0 answer correct while this year he/she gives 11 answers correct.
Student 351 instead gives 0 correct answers this year but due to missing answers in the previous year is not possible to compute the ratio.
In the second solution instead of use set, all years were passed but the "nonempty" function was applied on rows. In this case the most significant difference is that we "lose" students which don't give any answers in this year although they gave some answers the previous year. In fact, looking at the image in the middle no result are showed for them (no "no ans Year").
In the third solution we consider the fact that answer's data are only available for year 2019 and 2020.

Considering these aspects, we decided to construct a query which specifically work only on these years (differently to the first one which only restrict the visualization of results). In this way only the results for 2020 are showed.

ASSIGNMENT 2

For each subject show the total correct answers in percentage with respect to the total answers of that subject.



	Is Correct	tot_answer_subj	ratio
1	625	711	87.90%
2	301	367	82.02%
3	844	1322	63.84%
4	1436	2165	66.33%
5	40	97	41.24%
6	3983	6234	63.89%
7	85	236	36.02%
8	168	354	47.46%
9	818	1461	55.99%
10	3874	5308	72.98%
11	57	78	73.08%
12	248	520	47.69%

In this case we provided a unique query. We decided to distinguish two possible cases:

- The first is when we have the subject, but no answer are registered for it. In this case the ratio will return an indefinite form. For this reason, we decided to return "no answer" (although no case are present in our data).

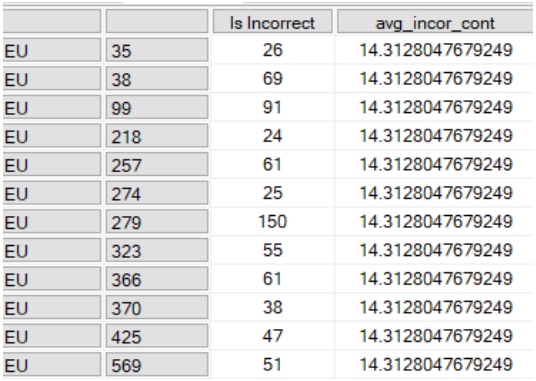
- The second is when we have all data so we can perform the ratio.

- In the image at the left is given a view of the solution where also correct answer and total answer are reported.

ASSIGNMENT 3

Show the students having a total incorrect answer greater or equal than the average incorrect answers in each continent.

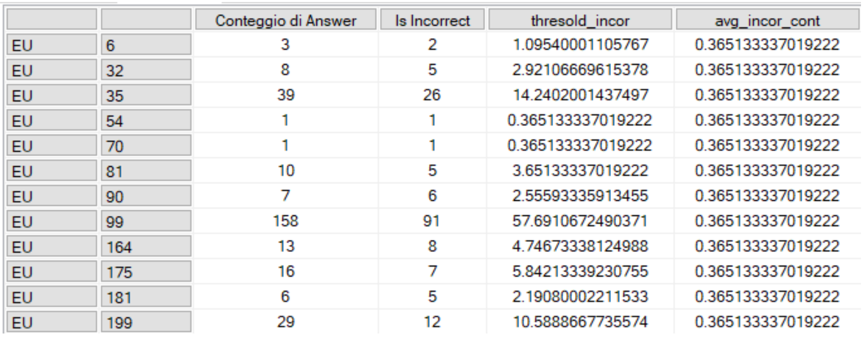
Here two queries were implemented.



		Is Incorrect	avg_incor_cont
EU	35	26	14.3128047679249
EU	38	69	14.3128047679249
EU	99	91	14.3128047679249
EU	218	24	14.3128047679249
EU	257	61	14.3128047679249
EU	274	25	14.3128047679249
EU	279	150	14.3128047679249
EU	323	55	14.3128047679249
EU	366	61	14.3128047679249
EU	370	38	14.3128047679249
EU	425	47	14.3128047679249
EU	569	51	14.3128047679249

A. ***The first one*** represents what we literally thought was asked by the assignment. We computed the average incorrect answer for each continent using the new measure "IsIncorrect" and "Conteggio di User1".

In fact, for each continent we computed the total incorrect answer given and we divided it by the total number of students. It returned that in Europe each student gave 14.31 incorrect answer while in North America and Oceania this value was respectively 14.82 and 15.04. After average was computed, we reported all the students which had a total incorrect answer greater than the average value of the continent. All worked fine but, looking at the data, we saw that a lot of students had a total answer lower than this threshold and so they didn't appear although the relative ratio of incorrect answer over the total was very high. For this reason, a second query was proposed.



		Conteggio di Answer	Is Incorrect	thresold_incor	avg_incor_cont
EU	6	3	2	1.09540001105767	0.365133337019222
EU	32	8	5	2.92106669615378	0.365133337019222
EU	35	39	26	14.2402001437497	0.365133337019222
EU	54	1	1	0.365133337019222	0.365133337019222
EU	70	1	1	0.365133337019222	0.365133337019222
EU	81	10	5	3.65133337019222	0.365133337019222
EU	90	7	6	2.55593335913455	0.365133337019222
EU	99	158	91	57.6910672490371	0.365133337019222
EU	164	13	8	4.74673338124988	0.365133337019222
EU	175	16	7	5.84213339230755	0.365133337019222
EU	181	6	5	2.19080002211533	0.365133337019222
EU	199	29	12	10.5888667735574	0.365133337019222

B. ***The second query*** was constructed computing the average of incorrect answer as the ratio between incorrect answer and total answer for each continent. In this way we had that in Europe the 36.51% of answer where incorrect while for North America and Oceania we obtained a value of 37.42% and 34.82%. After had this value we computed for each student his/her personal threshold multiplying the average value of incorrect answer for continent with the number of answers given by the student. In this way if the student gave several incorrect answers which is greater than the percentage allowed by his/her continent he/she was displayed in solution.

ASSIGNMENT 4

Create a dashboard that shows the geographical distribution of correct answers and incorrect answers

At this point what was created in the previous Assignments has been used: the column Geo_Bi has been used to geolocate the distribution of the data: without these changes the map was unable to point in the right region or Continent.

The dashboard is logically divided in two parts: on the left the “Is Correct” Data and on the right the “Is Incorrect” data. In the middle all the data divided again for Geo_Bi.

The idea here is that, clicking on the part of interest is clear the contribution in the two measures and how it changes during the two years for which we have data.

ASSIGNMENT 5

Create a plot/dashboard of your choosing, that you deem interesting w.r.t. the data available in your cube

Using the column Gender_BI, we decided to go through Gender using the analysis of correct and incorrect answer on the difference between the two years for which we have data and on the performance in different countries. We decided to not include the 10 records labeled as unknown gender from 2020.

The idea here is that, clicking on the part of interest is clear how both gender have performed both in the geographical area (counting how many answers we have from a particular area) but also in years