

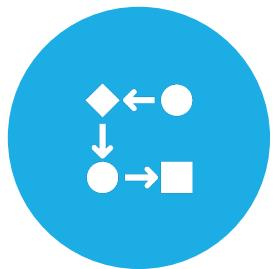
LABORATORY OF DATA SCIENCE

Group 24 Project Presentation – Davide Chen & Andrea Ribellino

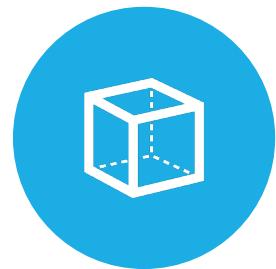
PROJECT PIPELINE



DATA WAREHOUSE
BUILDING



SSIS SOLUTION



MULTIDIMENSIONAL
DATA ANALYSIS



REPORTING



DATA WAREHOUSE BUILDING

Group 24 Project Presentation –
Davide Chen & Andrea Ribellino

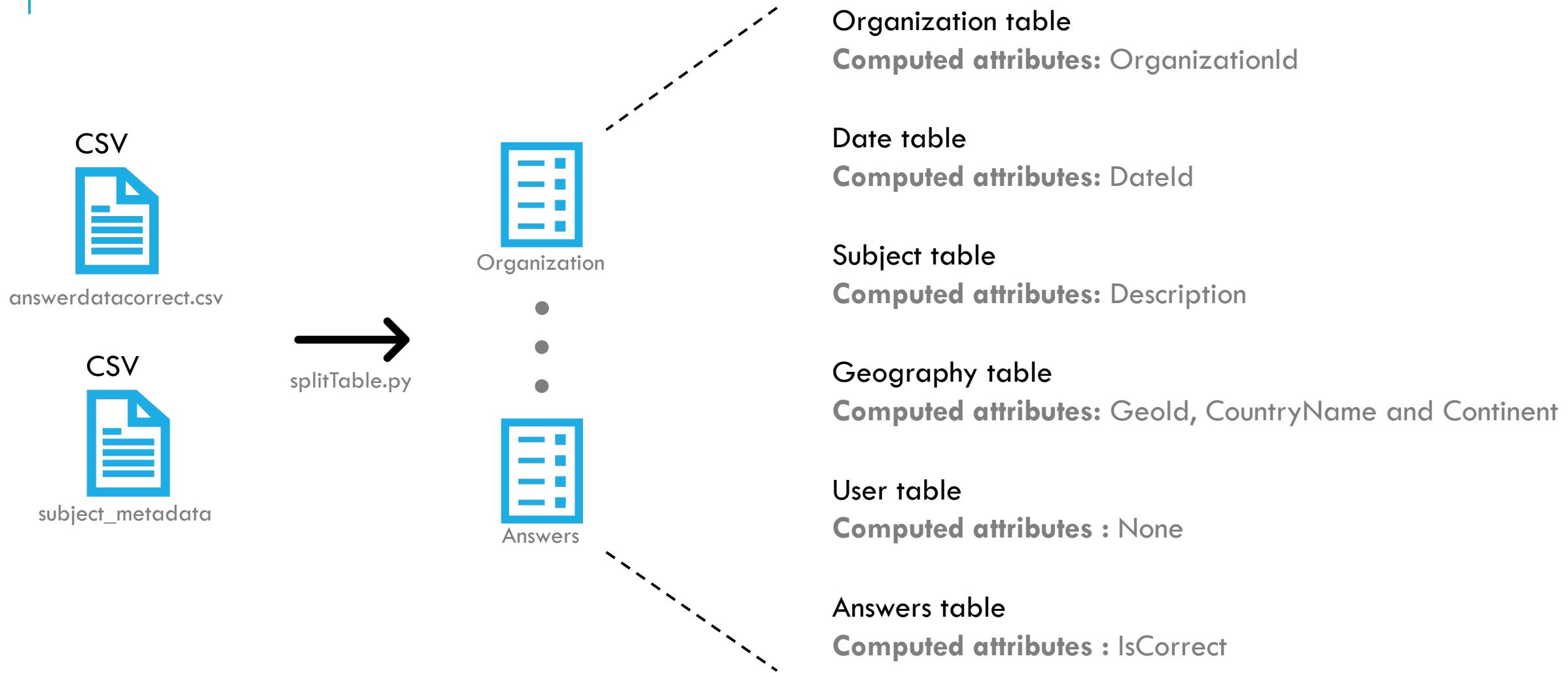
DATA UNDERSTANDING & PREPROCESSING

- The `answerdatacorrect.csv` data frame has 17 columns and 538835 rows with 0 missing value:
 - AnswerId has 538835 unique values and 0 duplicates
 - UsrId has 13630 unique values and 525205 duplicates
 - SubjectId has 412 unique values and 538423 duplicates
 - Organization table will have a total 24640 unique rows (given by [GroupId, QuizId, SchemeOfWorkId] unique rows)
 - Geography table will have a total 76 unique rows (given by Region unique rows)
- The CountyCode column has the following unique values:

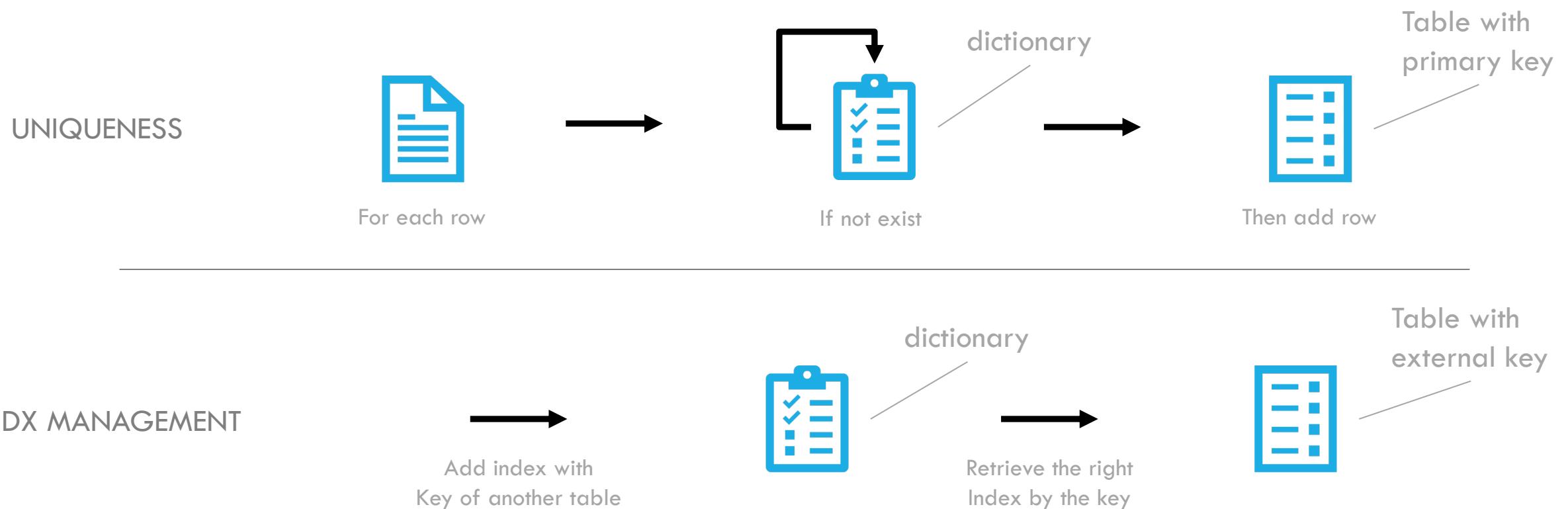
[de, us, ie, it, nz, es, uk, fr, ca, be, au]

In order to match [ISO 3166-1 alpha-2] standard of country code, a transformation 'uk' to 'gb' was applied.

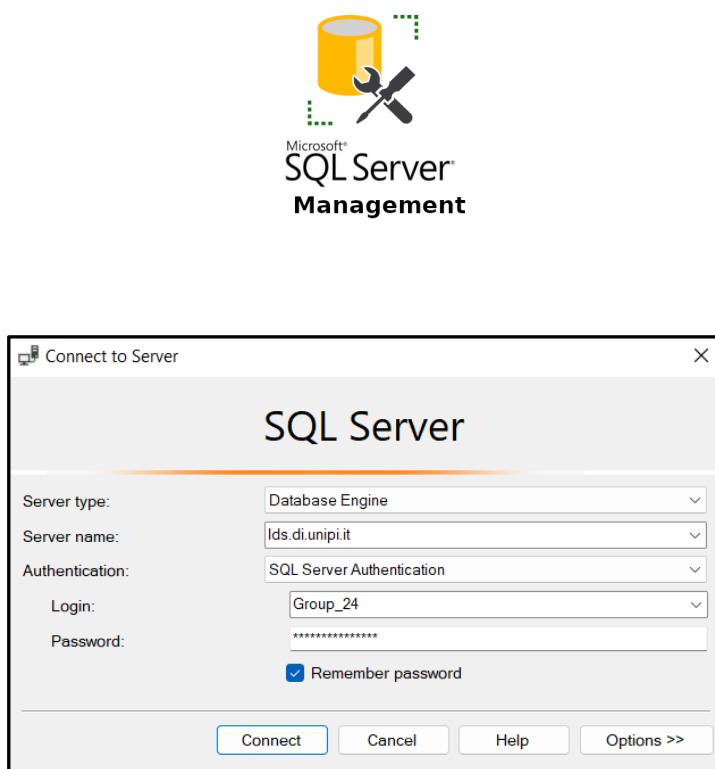
SPLITTING TABLES



UNIQUENESS & INDEX KEY MANAGEMENT



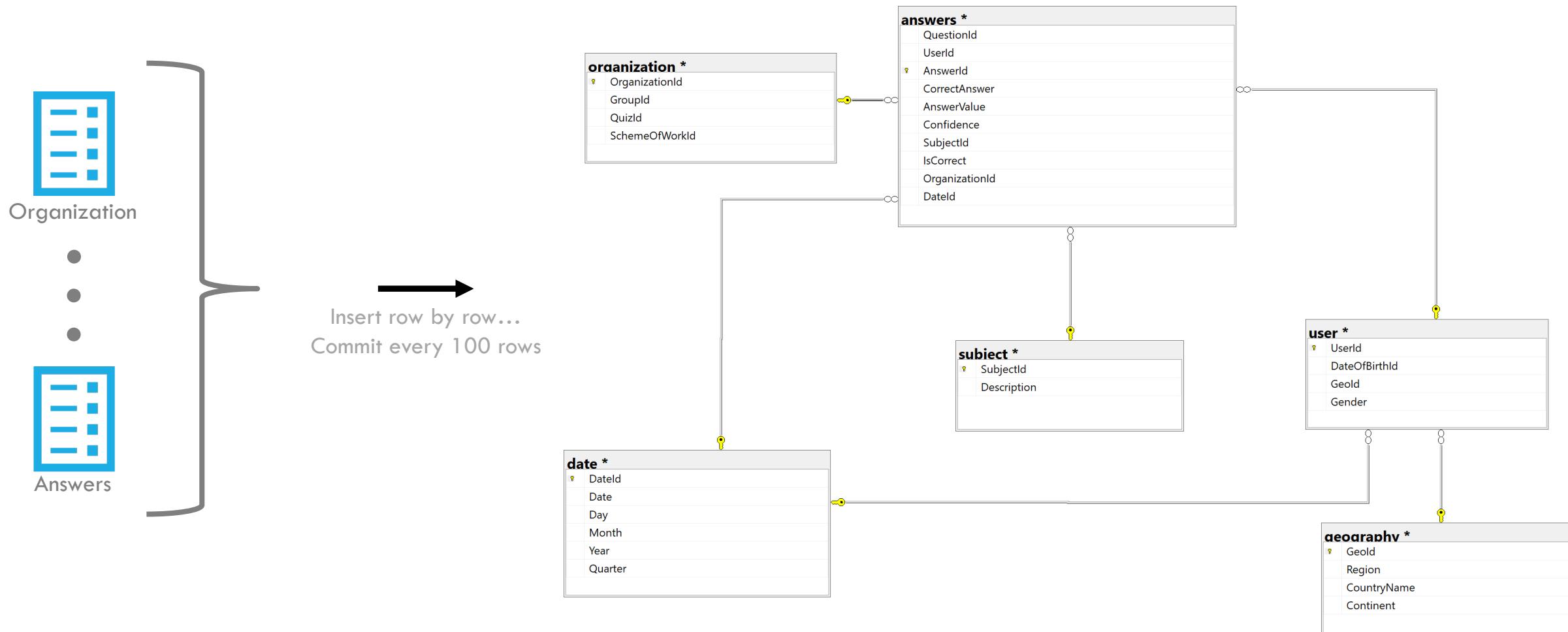
CREATION OF GROUP_24_DB



The screenshot shows the Microsoft SQL Server Management Studio interface. The title bar reads 'MVSalvatore.Group_24_DB - Group_24.answers - Microsoft SQL Server Management Studio'. The 'Object Explorer' pane on the left lists various databases and objects, including 'Group_15_DB' through 'Group_24_DB', 'Database Diagrams', and 'Tables'. The 'Tables' section shows three tables: 'System Tables', 'FileTables', 'External Tables', 'Graph Tables', 'Group_24.answers', 'Group_24.date', and 'Group_24.geography'. The 'Group_24.answers' table is selected in the 'Table Designer' pane on the right, which displays its columns:

Column Name	Data Type	Allow Nulls
QuestionId	int	<input type="checkbox"/>
UserId	int	<input type="checkbox"/>
AnswerId	int	<input type="checkbox"/>
CorrectAnswer	int	<input type="checkbox"/>
AnswerValue	int	<input type="checkbox"/>
Confidence	float	<input type="checkbox"/>
SubjectId	varchar(50)	<input type="checkbox"/>
IsCorrect	int	<input type="checkbox"/>
OrganizationId	int	<input type="checkbox"/>
Dated	int	<input type="checkbox"/>

LOADING DATA





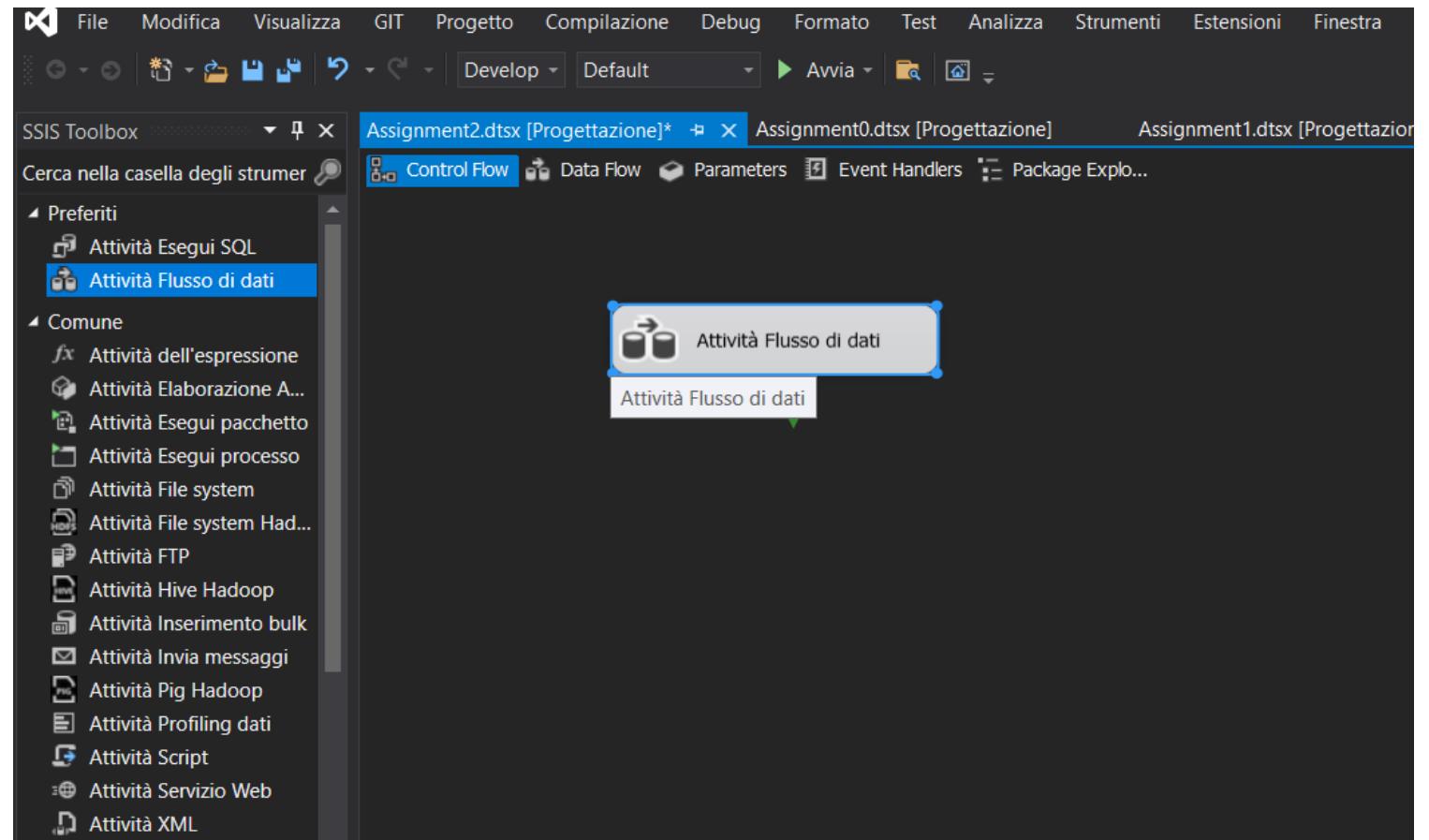
SISS – INTEGRATION SERVICE

Group 24 Project Presentation –
Davide Chen & Andrea Ribellino

VISUAL STUDIO 2019

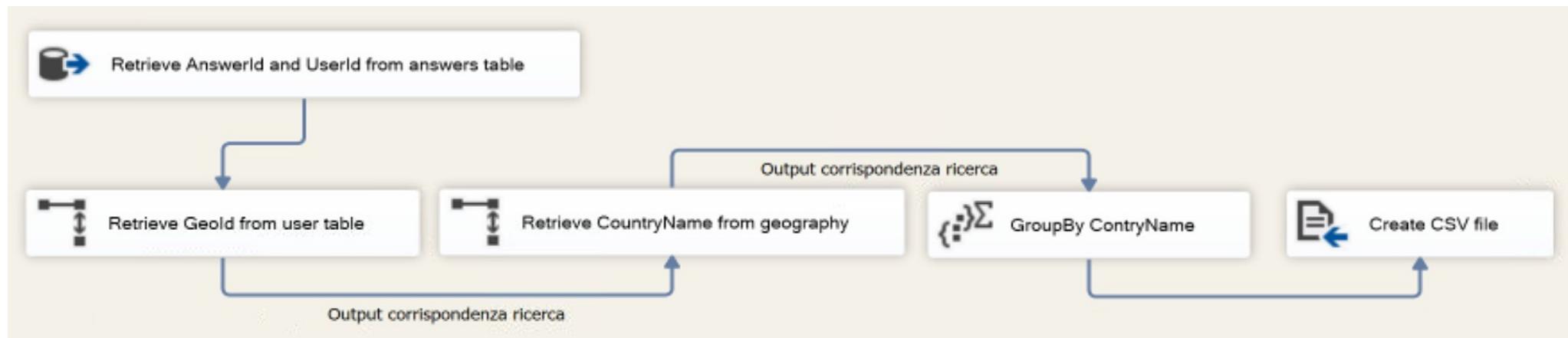


Visual
Studio



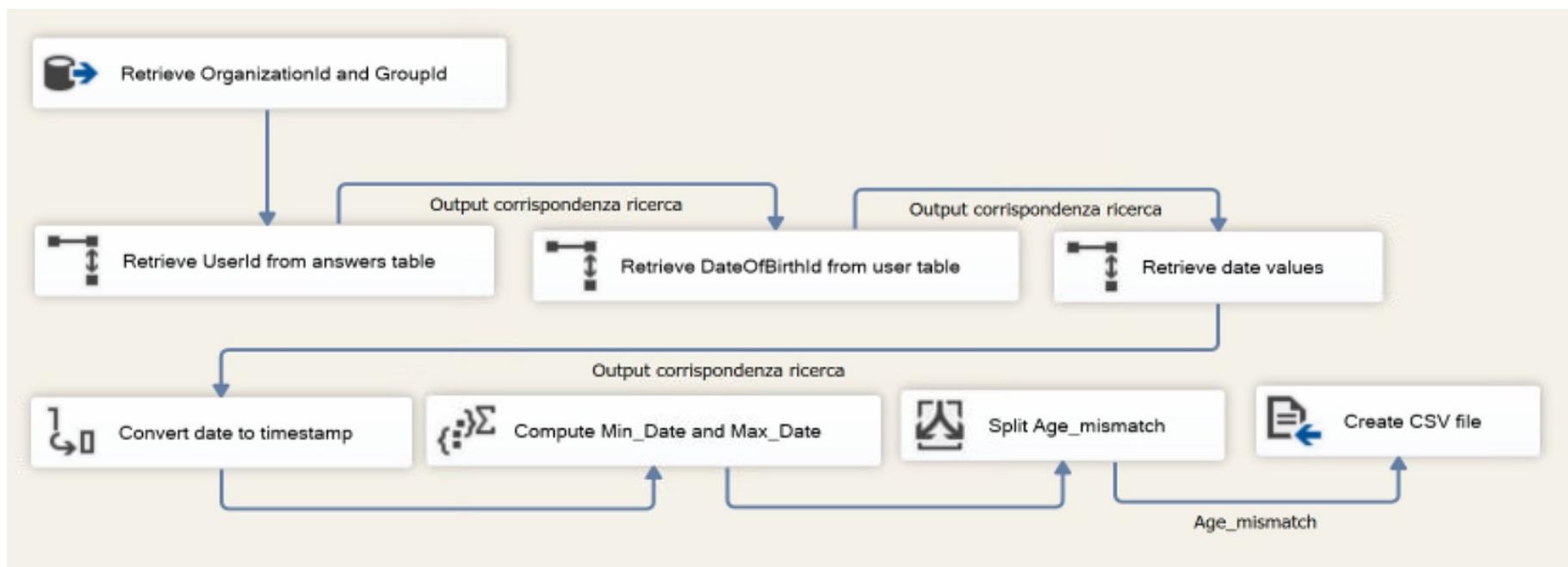
ASSIGNMENT 0

- For every country, the number of total answers.



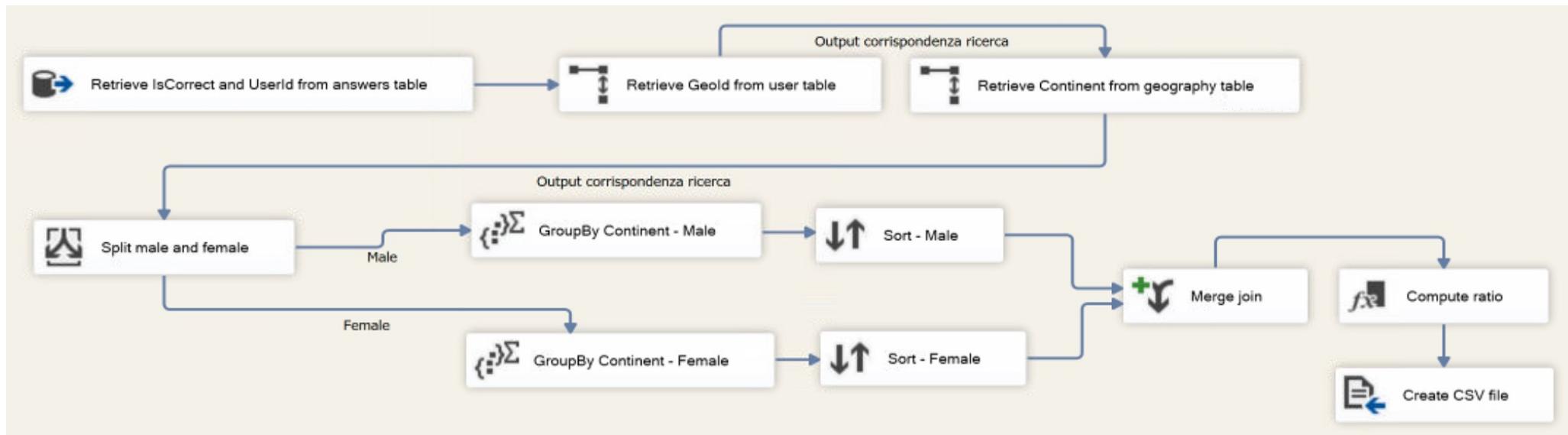
ASSIGNMENT 1

- A group (identified by GroupId) is said to have an age mismatch if the difference between the date of birth of the youngest participating student and the oldest is greater than 365 days. List all the groups with an age mismatch.



ASSIGNMENT 2

- For each continent the ratio between correct answers of males and correct answers of females





SASS – ANALYSIS SERVICE

Group 24 Project Presentation –
Davide Chen & Andrea Ribellino

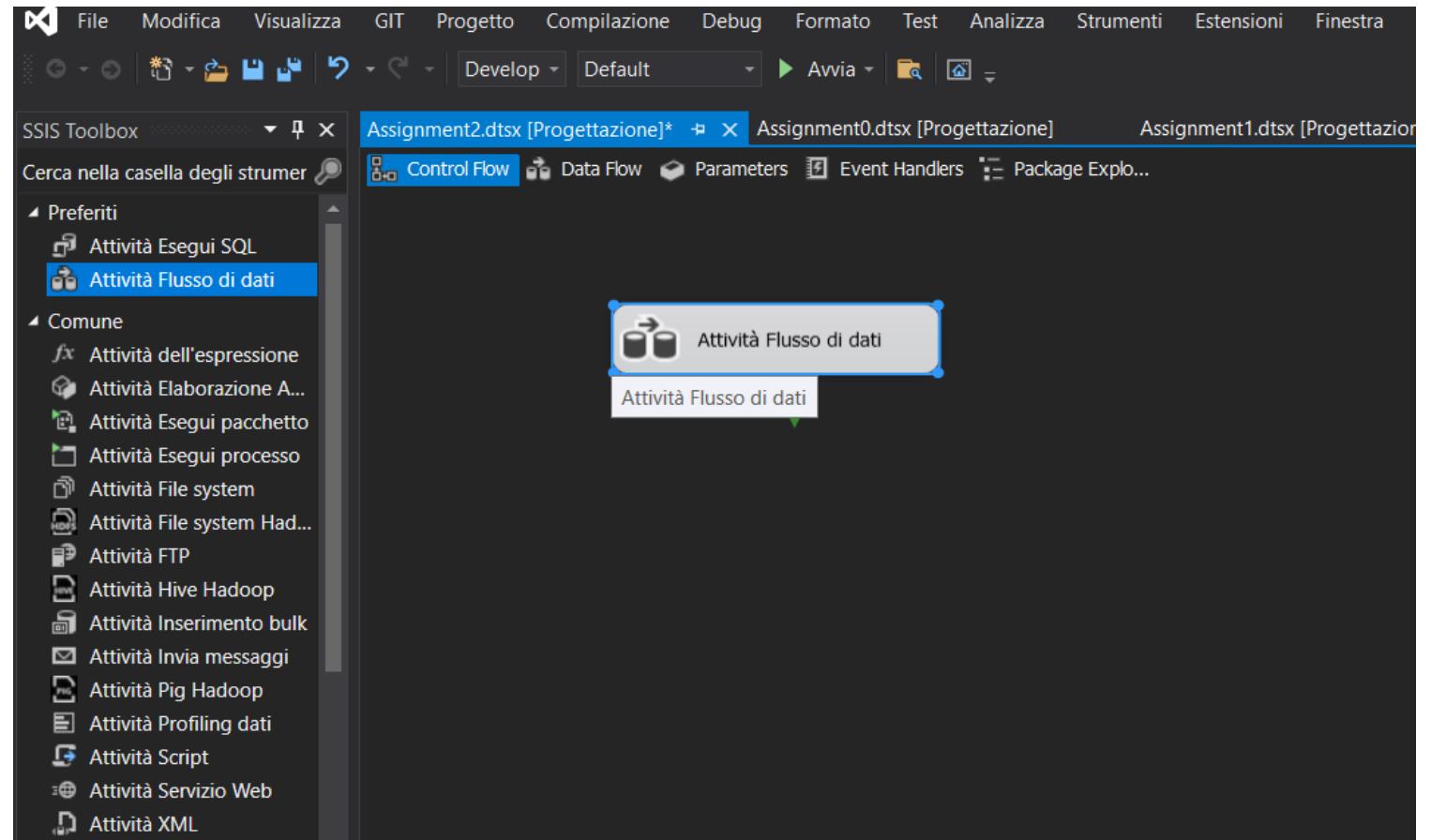
VISUAL STUDIO 2019 – CREATION OF OLAP CUBE



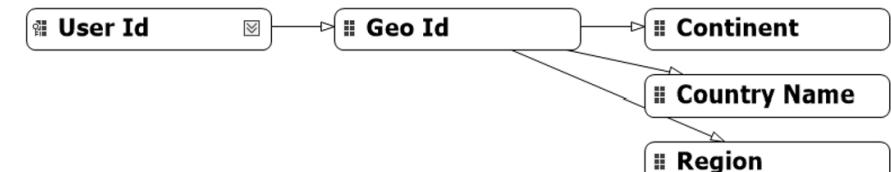
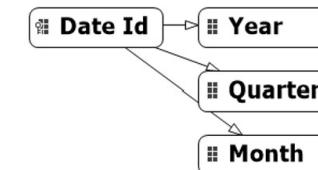
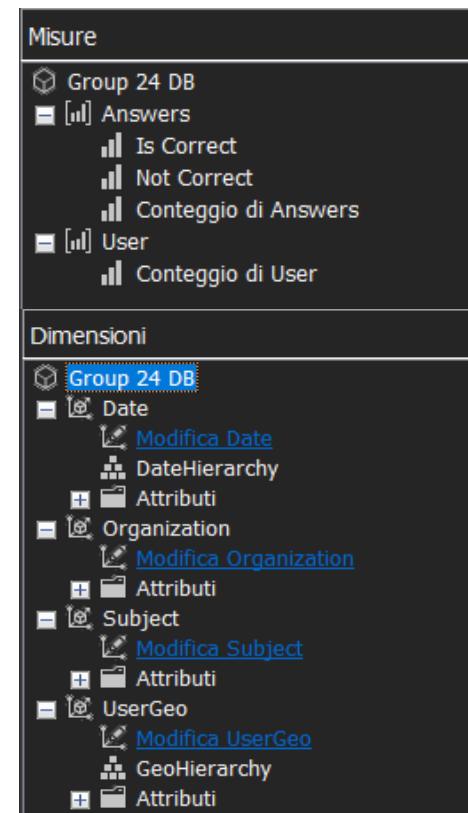
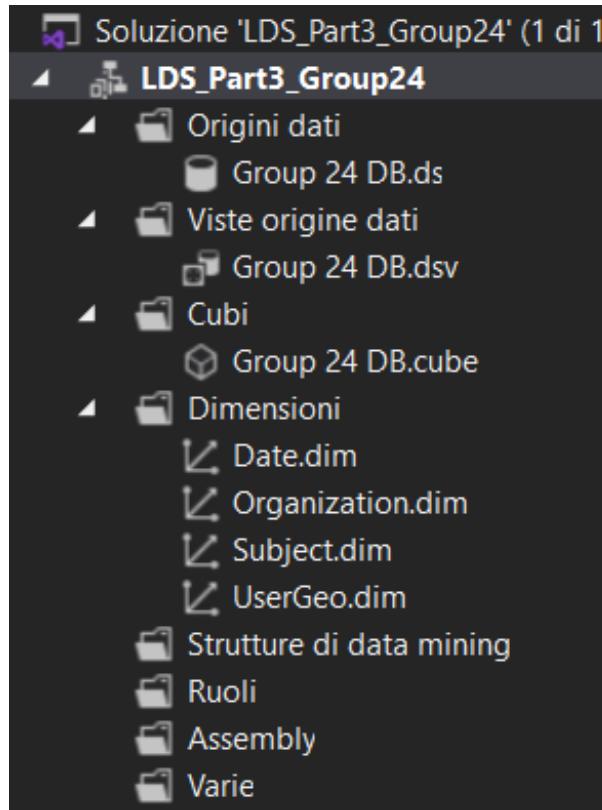
Visual
Studio



Microsoft®
SQL Server®
Analysis Services



CREATION OF OLAP CUBE



SQL MANAGEMENT SERVICE - ANALYSIS SERVICE



Connect to Server

SQL Server

Server type: Analysis Services

Server name:

Authentication: Windows Authentication

User name: DAVIDECHEN61A9\davide

Password:

Remember password

Connect **Cancel** **Help** **Options >>**

The screenshot shows the Microsoft SQL Server Management Studio interface. The title bar indicates the session is for MDX_Group24.mdx - http://lds.di.unipi.it/olap/msmdpump.dll.LDS_Part3_Group24 (DAVIDECHEN61A9\davide) - Microsoft SQL Server Management Studio.

The Object Explorer pane on the left lists various database objects, including cubes like group_30_tennis, Group_4_DB, Group_4_part_3, Group_5_tennis, Group_8, group12, group12_cube_project, Group13_2022, Group13_cube, Group15_Cube, Group16_2022, Group17_LDS, Group17_part3, Group18_LDS, Group19_Part3, group2_2022, Group24_tennis, group4_i_foodmart, Group6_Cube, Grup 31 Cube, Grup16pj, Grupo14_Task3_real, LDS_Cube_Group.11, LDS_Group10_Part3, LDS_Part3_Group24, Data Sources, Group 24 DB, Data Source Views, Cubes, Dimensions, Mining Structures, Roles, and Assemblies.

The MDXQuery_Group2...CHEN61A9\davide) pane shows the MDX script being developed:

```
-- Assignment 1
WITH MEMBER difference AS
CASE
WHEN ([Measures].[Is Correct],[Date].[Year].&[2020]) THEN 1
WHEN ([Measures].[Is Correct],[Date].[Year].&[2020]) THEN 1
WHEN ([Measures].[Is Correct],[Date].[Year].&[2019]) THEN 1
WHEN ([Measures].[Is Correct],[Date].[Year].&[2019]) THEN 1
ELSE (([Measures].[Is correct],[Date].[Year].&[2020]) - ([Measures].[Is correct],[Date].[Year].&[2019])) / ([Measures].[Is correct],[Date].[Year].&[2019])
END,
format_string="percent"

MEMBER is_correct_2019 AS
([Date].[Year].&[2019],[Measures].[Is Correct])

MEMBER is_correct_2020 AS
[Measures].[Is Correct] - is_correct_2019

SELECT filter([UserGeo].[User Id].[User Id], difference)
{is_correct_2019, is_correct_2020, difference} ON COLUMNS
FROM[Group 24 DB]

-- Assignment 2
WITH MEMBER correct_percentage AS
([Measures].[Is Correct])/([Measures].[Conteggio di Answers])
format_string = "Percent"
```

MDX QUERY - 1

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

```
1 -- Assignment 1
2 WITH MEMBER difference AS
3 CASE
4 WHEN ([Measures].[Is Correct],[Date].[Year].&[2020]) = 0 then null
5 WHEN ([Measures].[Is Correct],[Date].[Year].&[2020]) = null then null
6 WHEN ([Measures].[Is Correct],[Date].[Year].&[2019]) = 0 then null
7 WHEN ([Measures].[Is Correct],[Date].[Year].&[2019]) = null then null
8 ELSE (([Measures].[Is correct],[Date].[Year].&[2020])-([Measures].[Is correct],[Date].[Year].&[2019]))
9             /([Measures].[Is correct],[Date].[Year].&[2019])
10 END,
11 format_string="percent"
12
13 MEMBER is_correct_2019 AS
14 ([Date].[Year].&[2019],[Measures].[Is Correct])
15
16 MEMBER is_correct_2020 AS
17 [Measures].[Is Correct] - is_correct_2019
18
19 SELECT filter([UserGeo].[User Id].[User Id], difference <> null) on rows,
20 {is_correct_2019, is_correct_2020, difference} on columns
21 FROM[Group 24 DB]
```

Idea di output

	is_correct_2019	is_correct_2020	variation
userID			
userID			
userID			

$$\text{variation} = \frac{\text{correct_answer_of_2020} - \text{correct_answer_of_2019}}{\text{correct_answer_of_2019}}$$

	is_correct_2019	is_correct_2020	difference
38	75.00	45.00	-40.00%
99	4.00	63.00	1475.00%
132	21.00	7.00	-66.67%
164	4.00	1.00	-75.00%
182	27.00	1.00	-96.30%
199	6.00	11.00	83.33%
218	185.00	56.00	-69.73%
274	5.00	32.00	540.00%
323	112.00	87.00	-22.32%

MDX QUERY - 2

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

```
1 -- Assignment 2
2 WITH MEMBER correct_percentage AS
3 ([Measures].[Is Correct]/[Measures].[Conteggio di Answers]),
4 format_string = "Percent"
5
6 SELECT correct_percentage ON COLUMNS,
7 [Subject].[Subject Id].[Subject Id] ON ROWS
8 FROM [Group 24 DB]
```

OUTPUT

	correct_percentage
Maths, Advanced Pure, Functions, Composite Functions	37.64%
Maths, Advanced Pure, Functions, Function Notation	42.21%
Maths, Advanced Pure, Functions, Inverse Functions	42.36%
Maths, Algebra, Algebraic Fractions, Adding and Subtracting Algebra...	57.86%
Maths, Algebra, Algebraic Fractions, Multiplying and Dividing Algebra...	39.59%
Maths, Algebra, Algebraic Fractions, Simplifying Algebraic Fractions	55.85%
Maths, Algebra, Algebraic Fractions, Simplifying Algebraic Fractions	55.13%

$$\text{correct_percentage} = \frac{\text{Number_of_correct_answer_of_given_subject}}{\text{Number_of_answer_of_that_subject}}$$

MDX QUERY - 3

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

```
1 -- Assignment 3
2 WITH MEMBER avg_incorrect AS
3 ([[UserGeo].[Continent], [UserGeo].[User Id].[ALL],
4 [Measures].[Not Correct])/([UserGeo].[Continent], [UserGeo].[User Id].[ALL], [
5     Measures].[Conteggio di User])),
6 format_string = "standard"
7
8 SELECT {[Measures].[Not Correct], avg_incorrect} ON COLUMNS,
9 filter ([[UserGeo].[Continent].[Continent], [UserGeo].[User Id].[User Id]],
10 [Measures].[Not Correct] >= avg_incorrect) ON ROWS
11 FROM [Group 24 DB];
```

OUTPUT

		Not Correct	avg_incorrect
EU	35	26.00	14.31
EU	38	69.00	14.31
EU	99	91.00	14.31
EU	218	24.00	14.31
EU	257	61.00	14.31

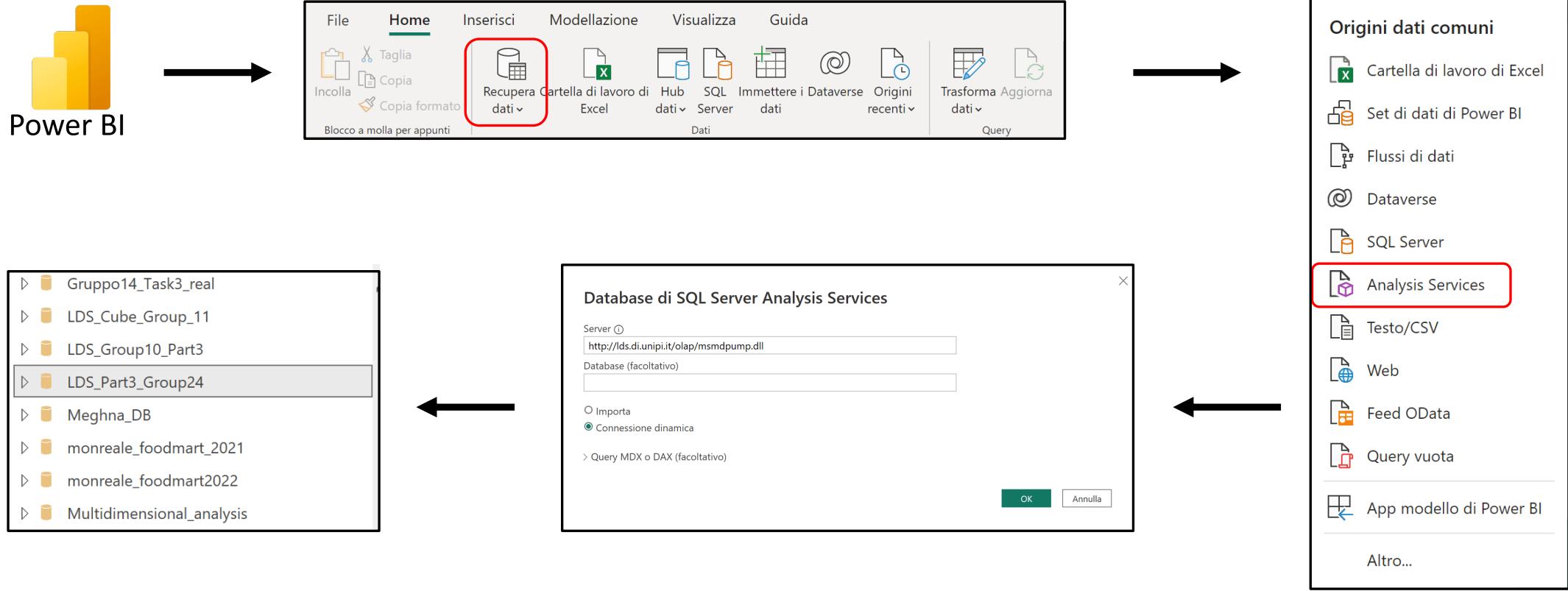
$$\text{avg_incorrect} = \frac{\text{sum_of_incorrect_answer_of_students_by_continent}}{\text{Number_of_student_in_that_continent}}$$



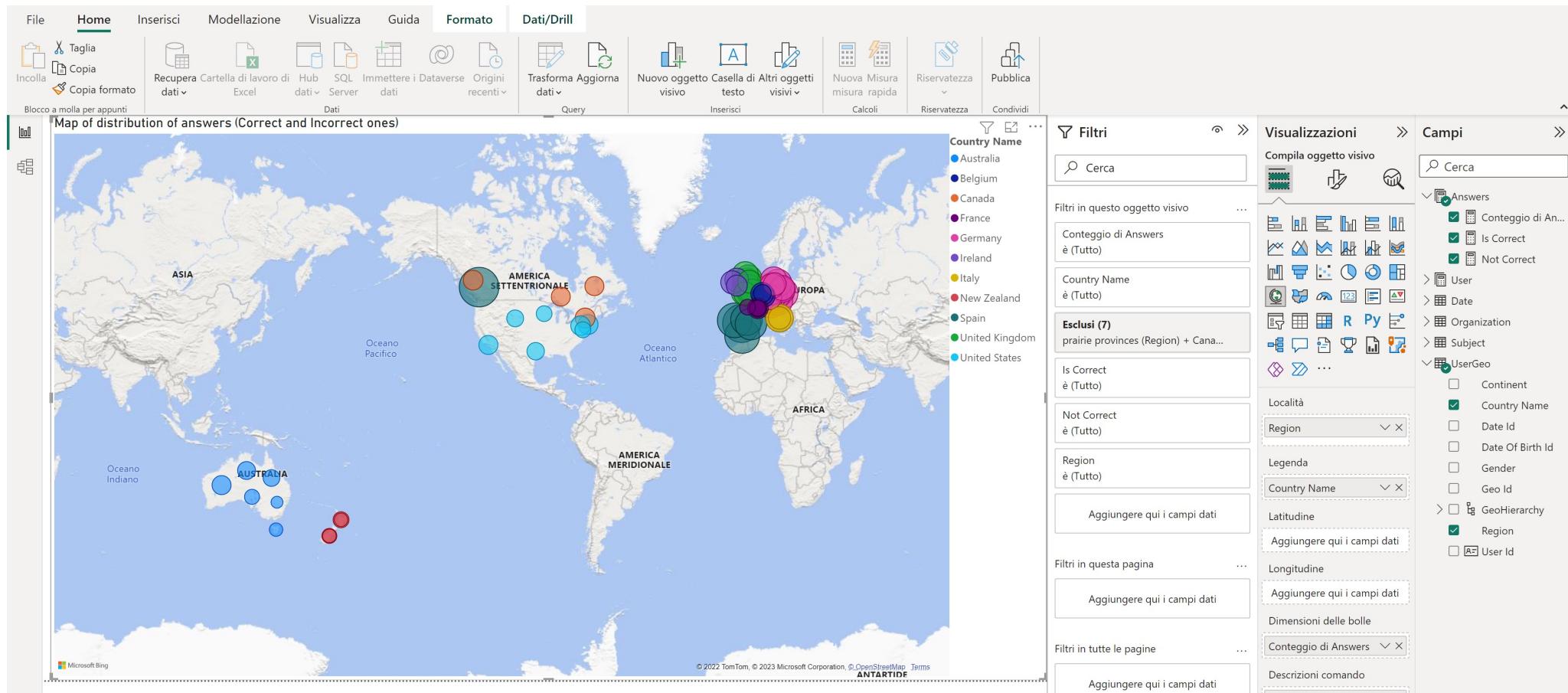
REPORTING

Group 24 Project Presentation –
Davide Chen & Andrea Ribellino

MICROSOFT POWER BI - CONNECTION



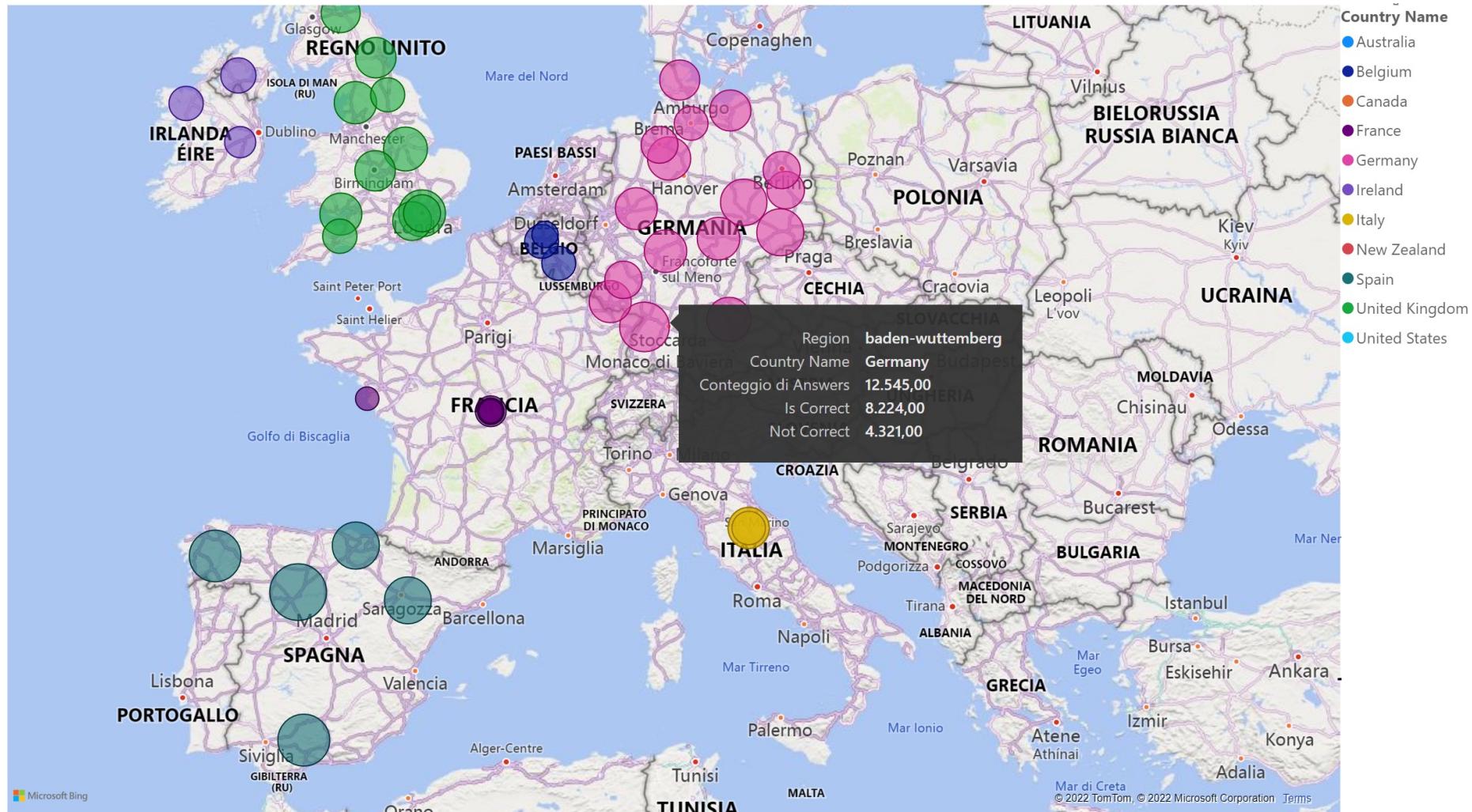
MICROSOFT POWER BI - USAGE



MAP OF ANSWERS DISTRIBUTIONS

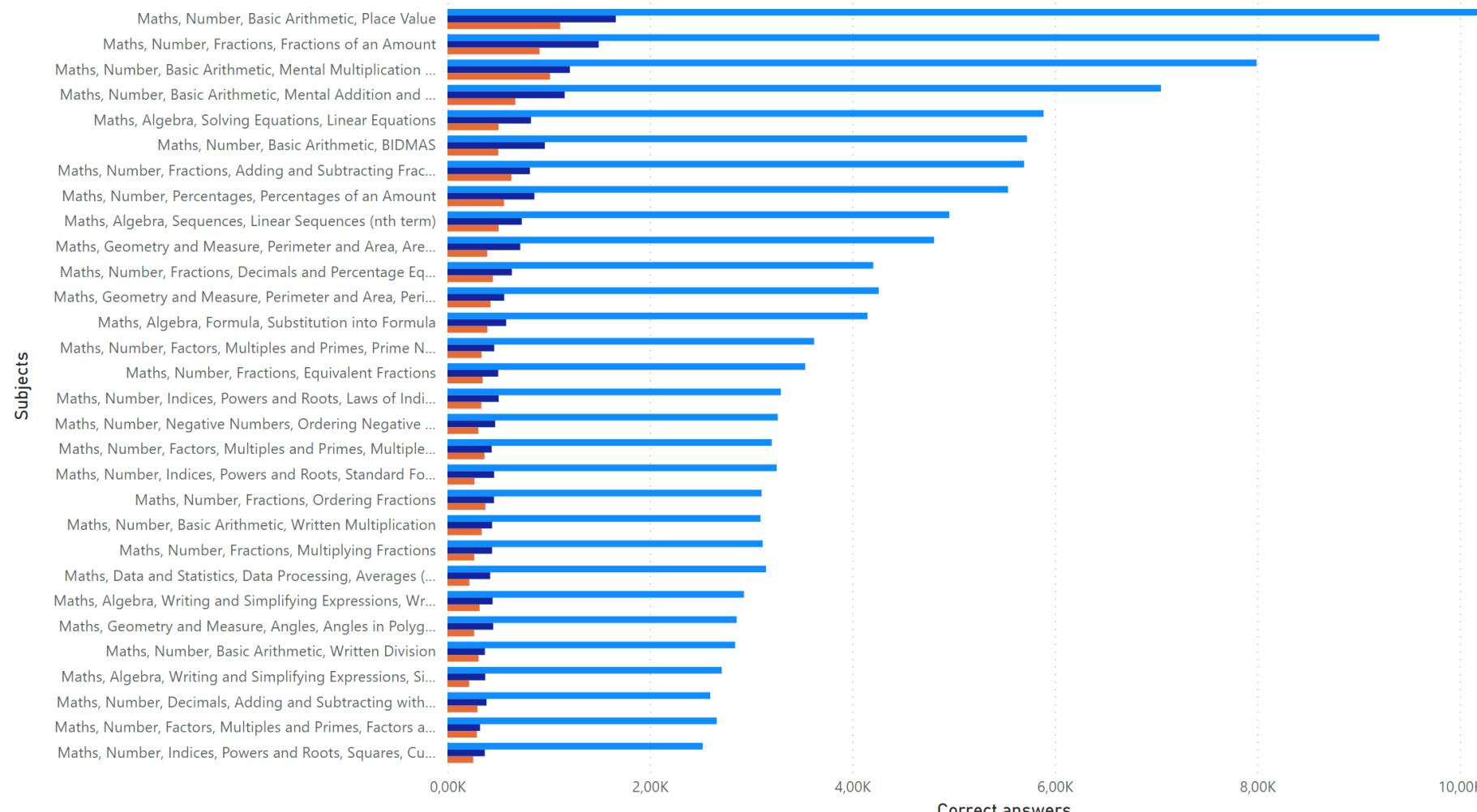


ZOOM-IN TO EUROPE



CORRECT ANSWERS BY SUBJECT AND CONTINENT

Continent ● EU ● NA ● OC



THANKS FOR THE ATTENTION

Group 24

Davide Chen

Email: d.chen@studenti.unipi.it

Andrea Ribellino

Email: a.ribellino@studenti.unipi.it

