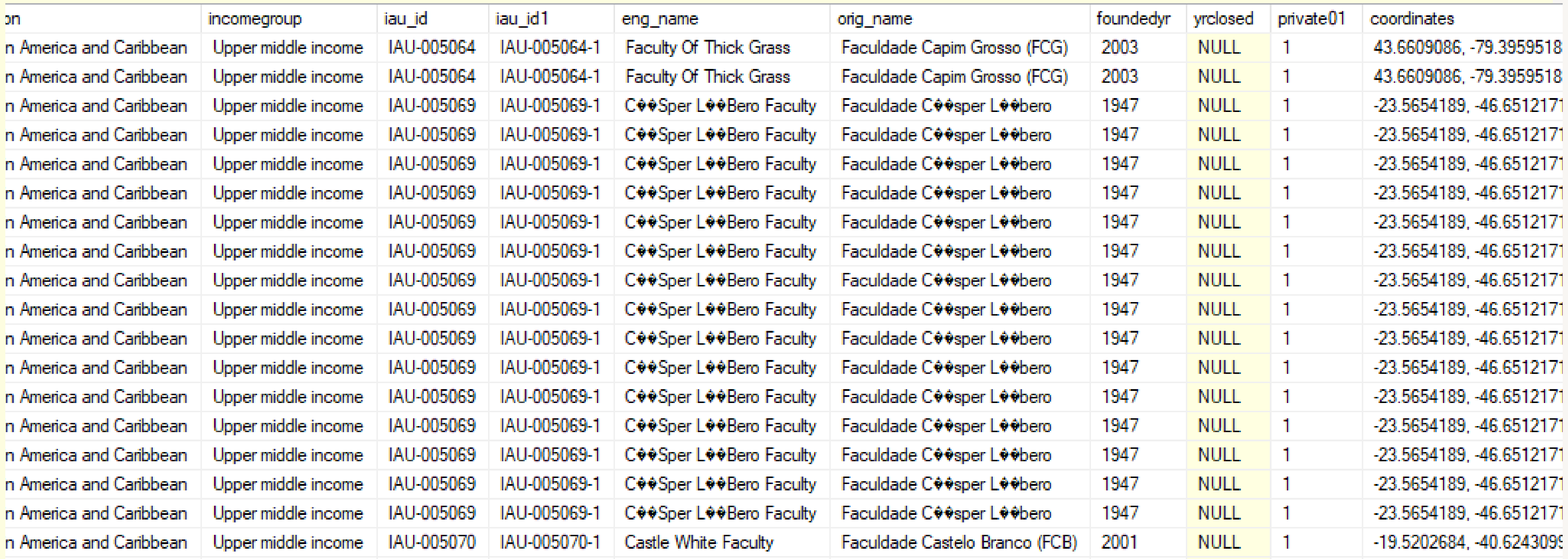
University Enrollments

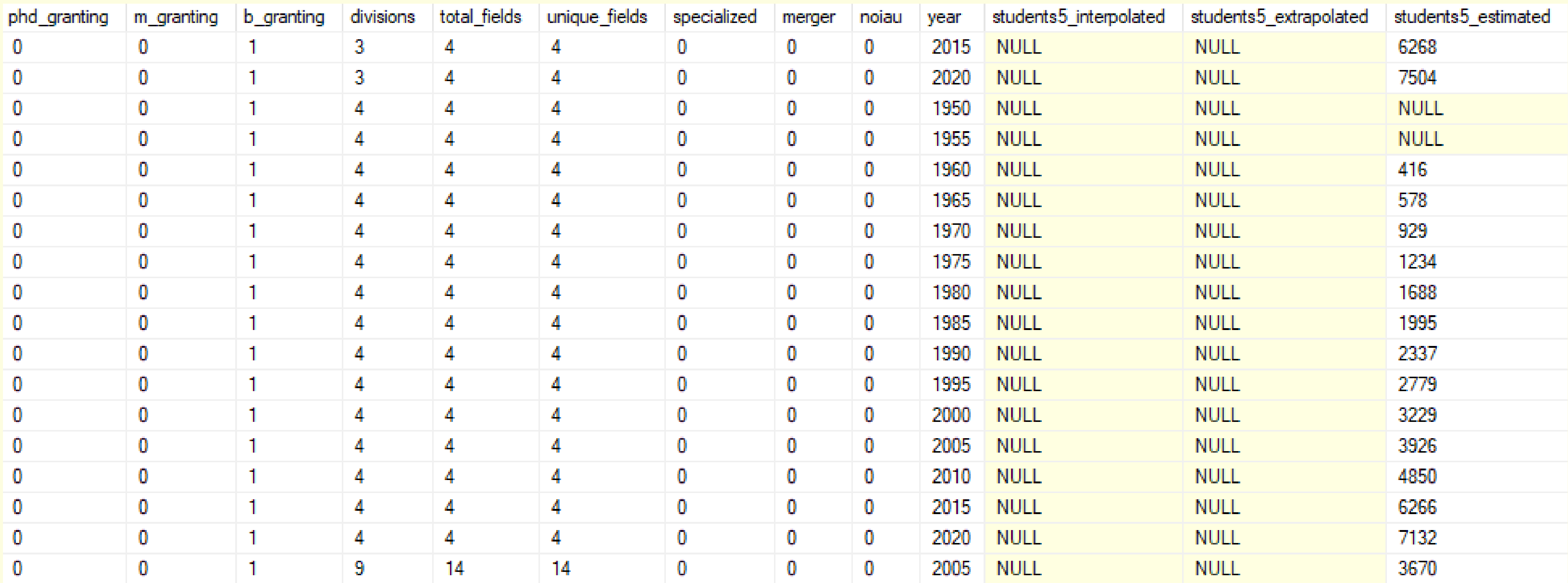
Βασίλειος Δημόπουλος
t8190038
Αλέξιος Φιλιππακόπουλος
p3190212

JANUARY 12, 2023

Enrollments Dataset



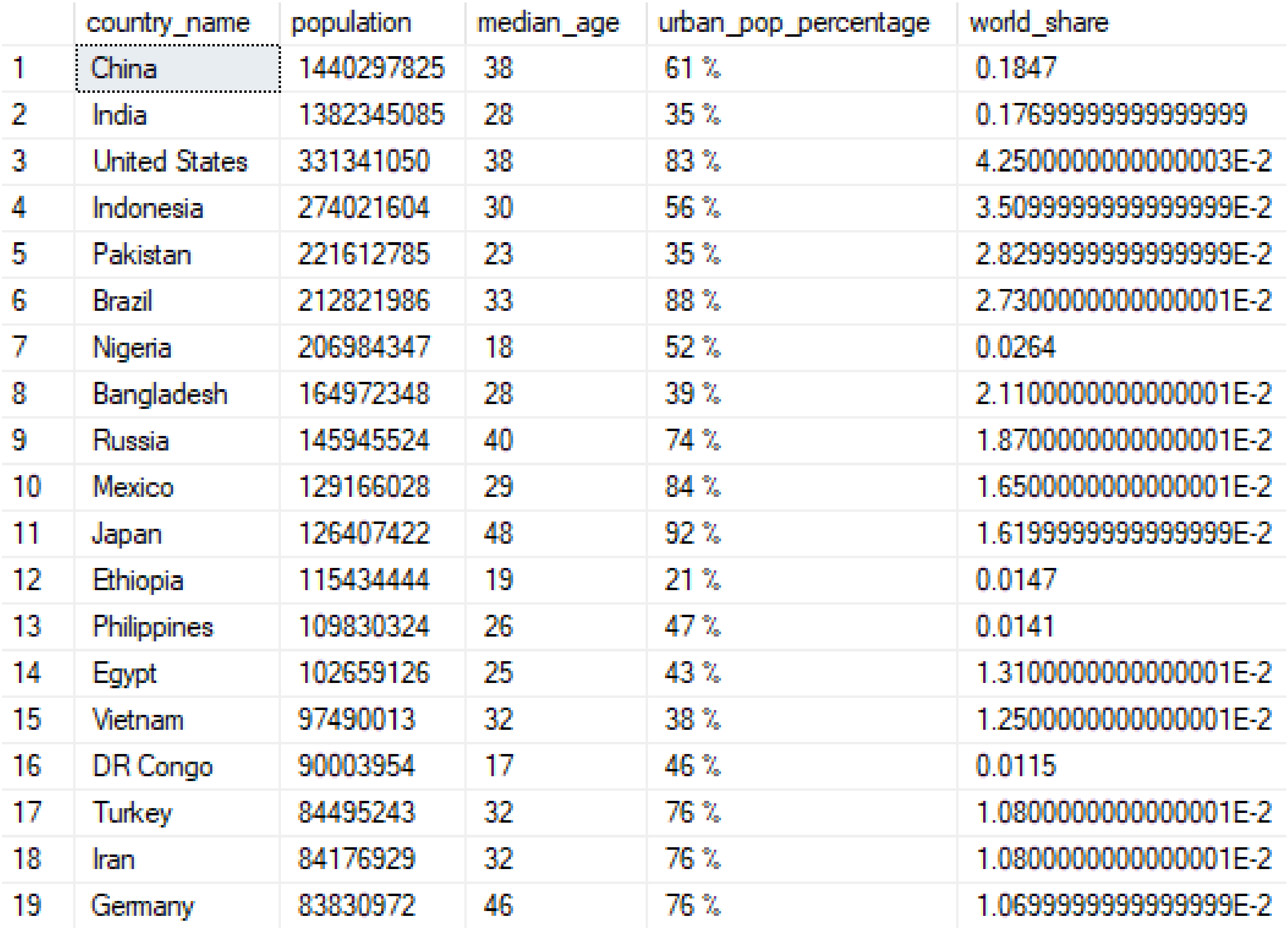
on	incomegroup	iau_id	iau_id1	eng_name	orig_name	foundedyr	yrclosed	private01	coordinates
n America and Caribbean	Upper middle income	IAU-005064	IAU-005064-1	Faculty Of Thick Grass	Faculdade Capim Grosso (FCG)	2003	NULL	1	43.6609086, -79.3959518
n America and Caribbean	Upper middle income	IAU-005064	IAU-005064-1	Faculty Of Thick Grass	Faculdade Capim Grosso (FCG)	2003	NULL	1	43.6609086, -79.3959518
n America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947	NULL	1	-23.5654189, -46.6512171
n America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947	NULL	1	-23.5654189, -46.6512171
n America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947	NULL	1	-23.5654189, -46.6512171
n America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947	NULL	1	-23.5654189, -46.6512171
n America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947	NULL	1	-23.5654189, -46.6512171
n America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947	NULL	1	-23.5654189, -46.6512171
n America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947	NULL	1	-23.5654189, -46.6512171
n America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947	NULL	1	-23.5654189, -46.6512171
n America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947	NULL	1	-23.5654189, -46.6512171
n America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947	NULL	1	-23.5654189, -46.6512171
n America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947	NULL	1	-23.5654189, -46.6512171
n America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947	NULL	1	-23.5654189, -46.6512171
n America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947	NULL	1	-23.5654189, -46.6512171
n America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947	NULL	1	-23.5654189, -46.6512171
n America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947	NULL	1	-23.5654189, -46.6512171
n America and Caribbean	Upper middle income	IAU-005070	IAU-005070-1	Castle White Faculty	Faculdade Castelo Branco (FCB)	2001	NULL	1	-19.5202684, -40.6243099



phd_granting	m_granting	b_granting	divisions	total_fields	unique_fields	specialized	merger	noiau	year	students5_interpolated	students5_extrapolated	students5_estimated
0	0	1	3	4	4	0	0	0	2015	NULL	NULL	6268
0	0	1	3	4	4	0	0	0	2020	NULL	NULL	7504
0	0	1	4	4	4	0	0	0	1950	NULL	NULL	NULL
0	0	1	4	4	4	0	0	0	1955	NULL	NULL	NULL
0	0	1	4	4	4	0	0	0	1960	NULL	NULL	416
0	0	1	4	4	4	0	0	0	1965	NULL	NULL	578
0	0	1	4	4	4	0	0	0	1970	NULL	NULL	929
0	0	1	4	4	4	0	0	0	1975	NULL	NULL	1234
0	0	1	4	4	4	0	0	0	1980	NULL	NULL	1688
0	0	1	4	4	4	0	0	0	1985	NULL	NULL	1995
0	0	1	4	4	4	0	0	0	1990	NULL	NULL	2337
0	0	1	4	4	4	0	0	0	1995	NULL	NULL	2779
0	0	1	4	4	4	0	0	0	2000	NULL	NULL	3229
0	0	1	4	4	4	0	0	0	2005	NULL	NULL	3926
0	0	1	4	4	4	0	0	0	2010	NULL	NULL	4850
0	0	1	4	4	4	0	0	0	2015	NULL	NULL	6266
0	0	1	4	4	4	0	0	0	2020	NULL	NULL	7132
0	0	1	9	14	14	0	0	0	2005	NULL	NULL	3670

- Data from 17000+ Universities from 180+ countries
- Data for each university every 5 yers
- Over 27 columns with data on degree grantings, location, fields etc and enrollments for every 5 years since the 50s
- Source: https://www.kaggle.com/datasets/michaelbryantds/university-enrollments-dataset

Population Dataset

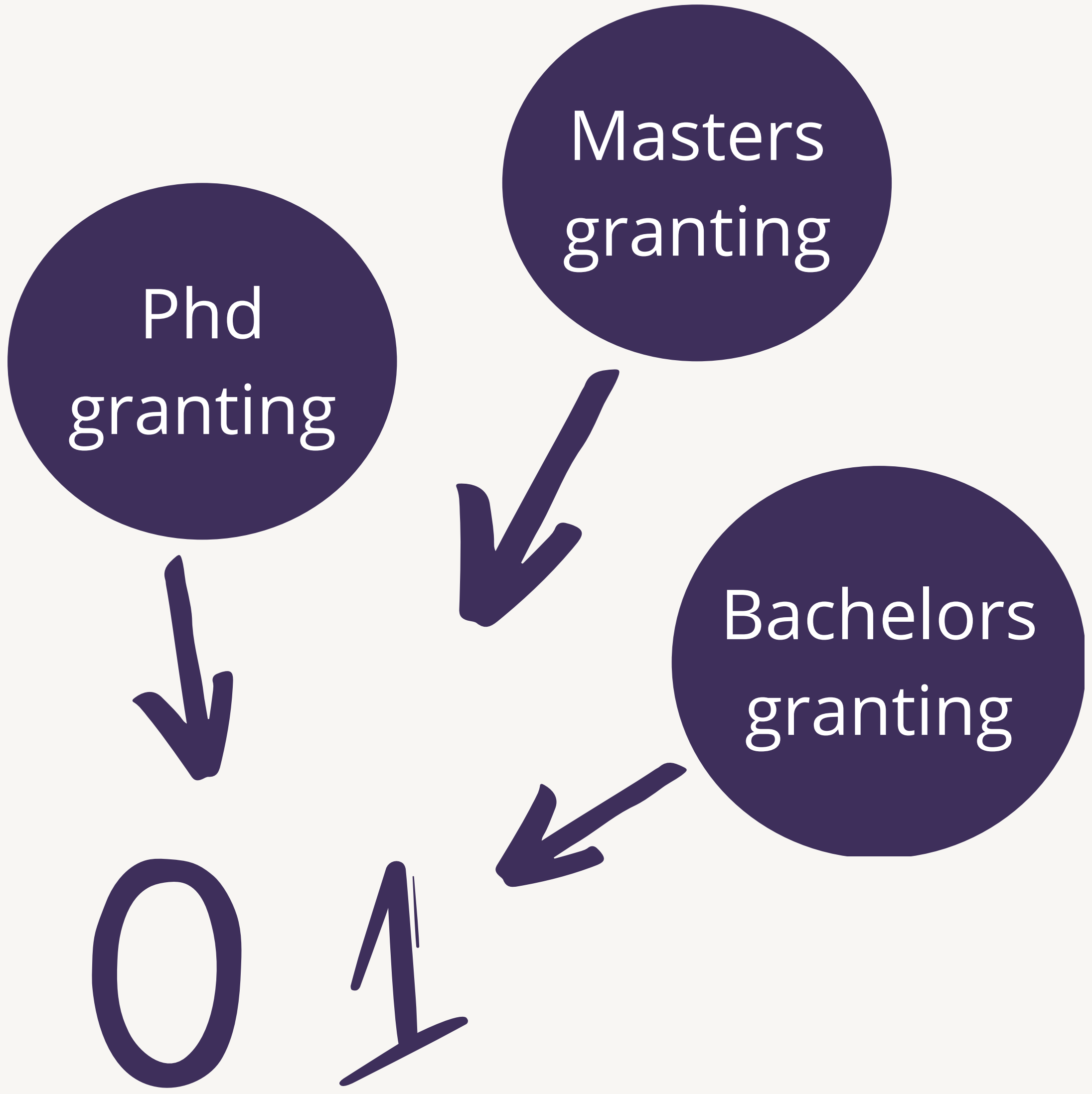


	country_name	population	median_age	urban_pop_percentage	world_share
1	China	1440297825	38	61 %	0.1847
2	India	1382345085	28	35 %	0.17699999999999999
3	United States	331341050	38	83 %	4.2500000000000003E-2
4	Indonesia	274021604	30	56 %	3.5099999999999999E-2
5	Pakistan	221612785	23	35 %	2.8299999999999999E-2
6	Brazil	212821986	33	88 %	2.7300000000000001E-2
7	Nigeria	206984347	18	52 %	0.0264
8	Bangladesh	164972348	28	39 %	2.1100000000000001E-2
9	Russia	145945524	40	74 %	1.8700000000000001E-2
10	Mexico	129166028	29	84 %	1.6500000000000001E-2
11	Japan	126407422	48	92 %	1.6199999999999999E-2
12	Ethiopia	115434444	19	21 %	0.0147
13	Philippines	109830324	26	47 %	0.0141
14	Egypt	102659126	25	43 %	1.3100000000000001E-2
15	Vietnam	97490013	32	38 %	1.2500000000000001E-2
16	DR Congo	90003954	17	46 %	0.0115
17	Turkey	84495243	32	76 %	1.0800000000000001E-2
18	Iran	84176929	32	76 %	1.0800000000000001E-2
19	Germany	83830972	46	76 %	1.0699999999999999E-2

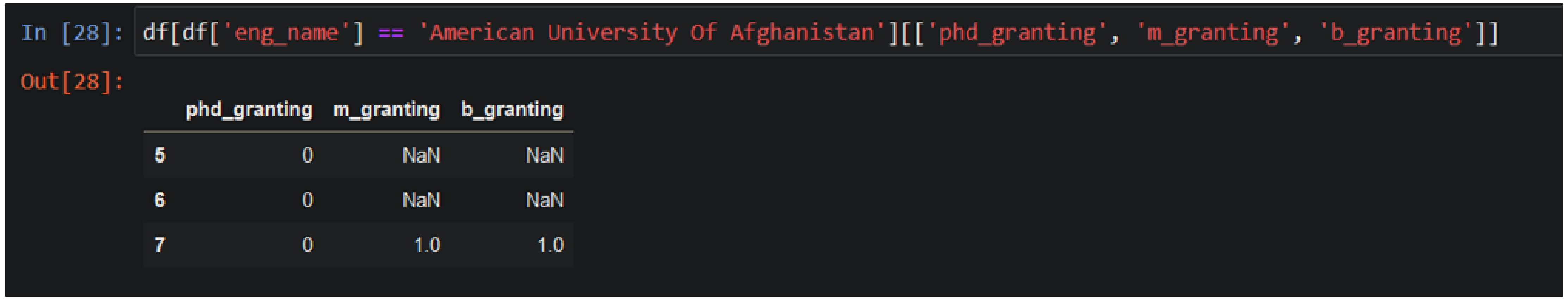
- Contains data for 200+ countries
- Consists of 11 columns such as country, population, median age, world share…
- Source: Kaggle

Data Manipulation (1)

Information about the offered degrees of each institution was found only on the last index.

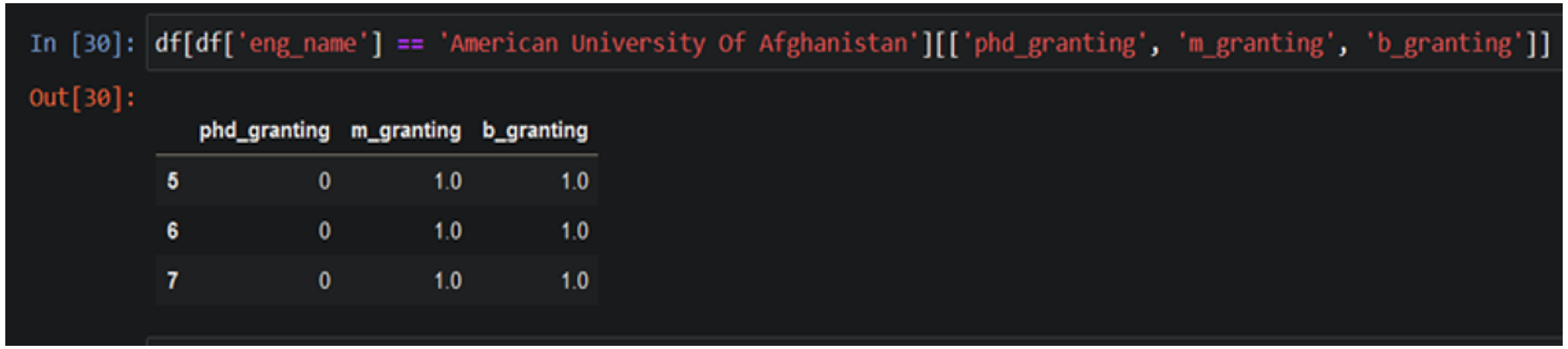


Before:



```
In [28]: df[df['eng_name'] == 'American University Of Afghanistan'][['phd_granting', 'm_granting', 'b_granting']]
Out[28]:
     phd_granting  m_granting  b_granting
  5             0         NaN         NaN
  6             0         NaN         NaN
  7             0         1.0         1.0
```

After:

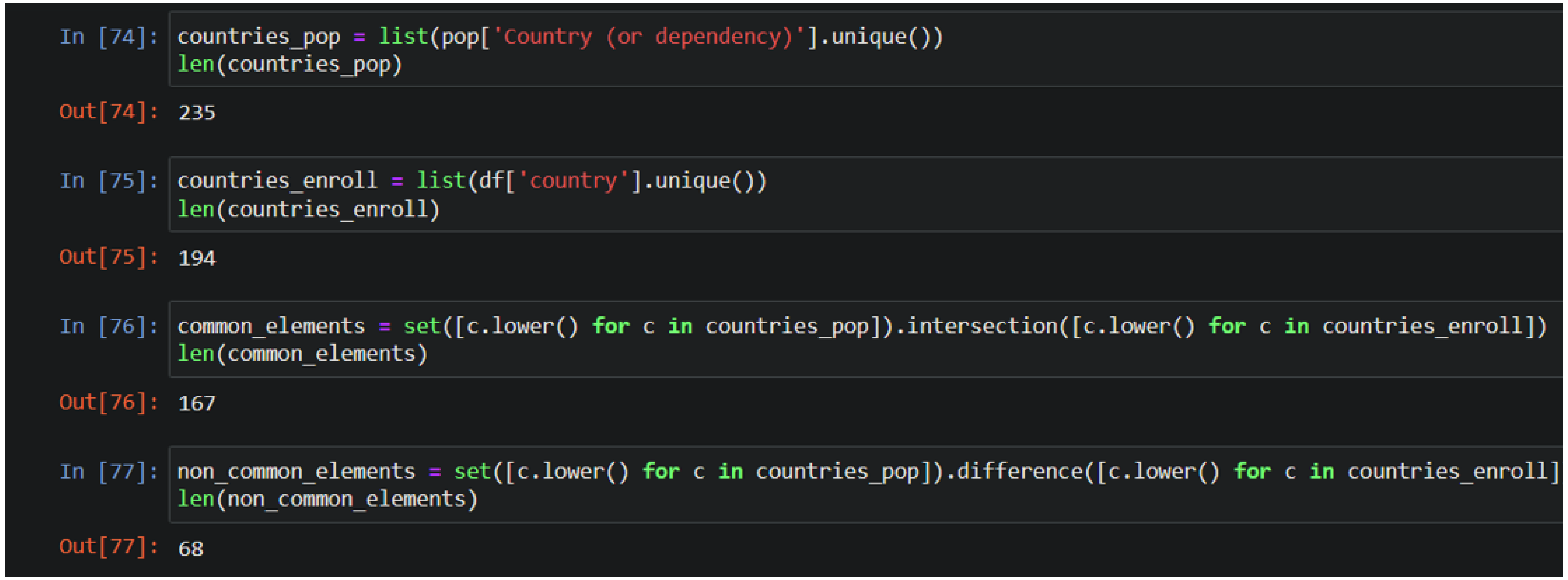


```
In [30]: df[df['eng_name'] == 'American University Of Afghanistan'][['phd_granting', 'm_granting', 'b_granting']]
Out[30]:
     phd_granting  m_granting  b_granting
  5             0         1.0         1.0
  6             0         1.0         1.0
  7             0         1.0         1.0
```

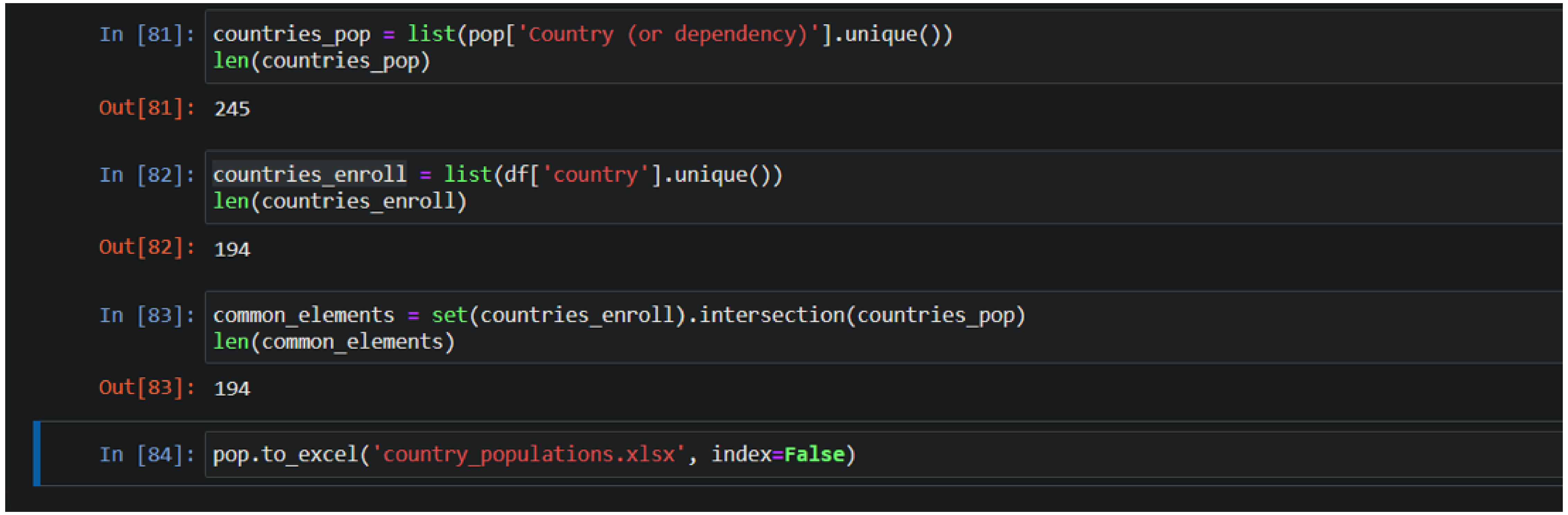
Data Manipulation (2)

Between the two datasets we had 68 countries that did not match.

Through analyzing the data, we found that most of these countries were referenced in a different manner across each dataset. For the ones that did not match we searched the web and filled the columns ourselves resulting in:



```
In [74]: countries_pop = list(pop['Country (or dependency)'].unique())
         len(countries_pop)

Out[74]: 235

In [75]: countries_enroll = list(df['country'].unique())
         len(countries_enroll)

Out[75]: 194

In [76]: common_elements = set([c.lower() for c in countries_pop]).intersection([c.lower() for c in countries_enroll])
         len(common_elements)

Out[76]: 167

In [77]: non_common_elements = set([c.lower() for c in countries_pop]).difference([c.lower() for c in countries_enroll]
         len(non_common_elements)

Out[77]: 68
```



```
In [81]: countries_pop = list(pop['Country (or dependency)'].unique())
         len(countries_pop)

Out[81]: 245

In [82]: countries_enroll = list(df['country'].unique())
         len(countries_enroll)

Out[82]: 194

In [83]: common_elements = set(countries_enroll).intersection(countries_pop)
         len(common_elements)

Out[83]: 194

In [84]: pop.to_excel('country_populations.xlsx', index=False)
```

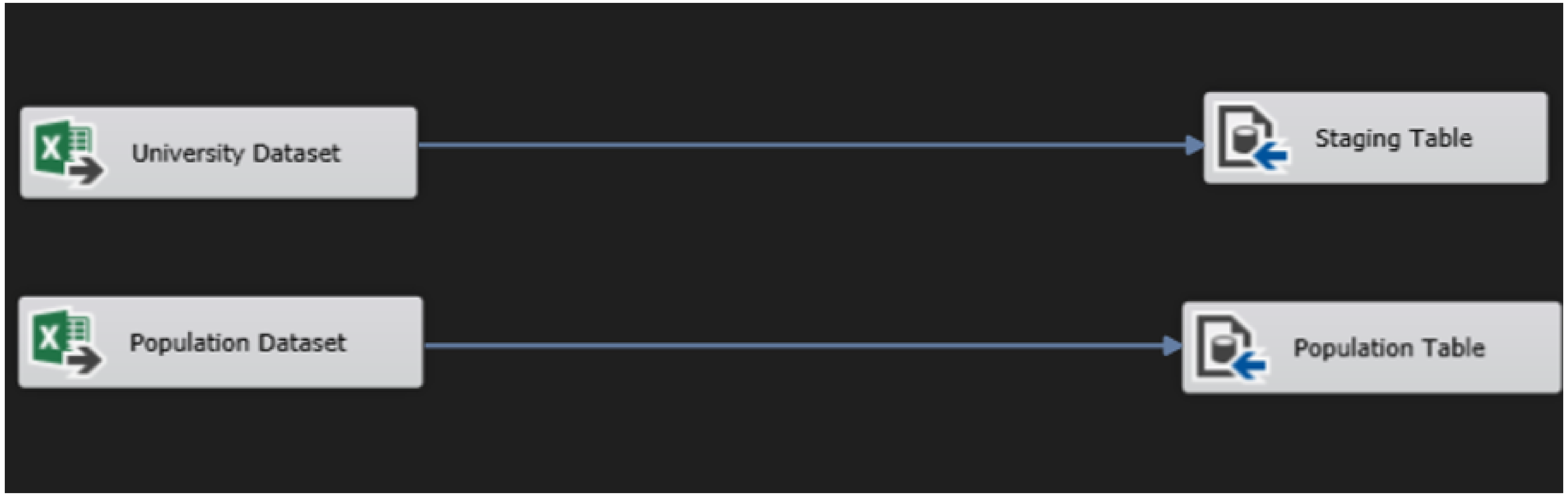
Data Warehousing

We used Microsoft Visual Studios Integration Services Project to load our data to the SQL Server, create staging, dimension and fact tables

Microsoft
SQL Server

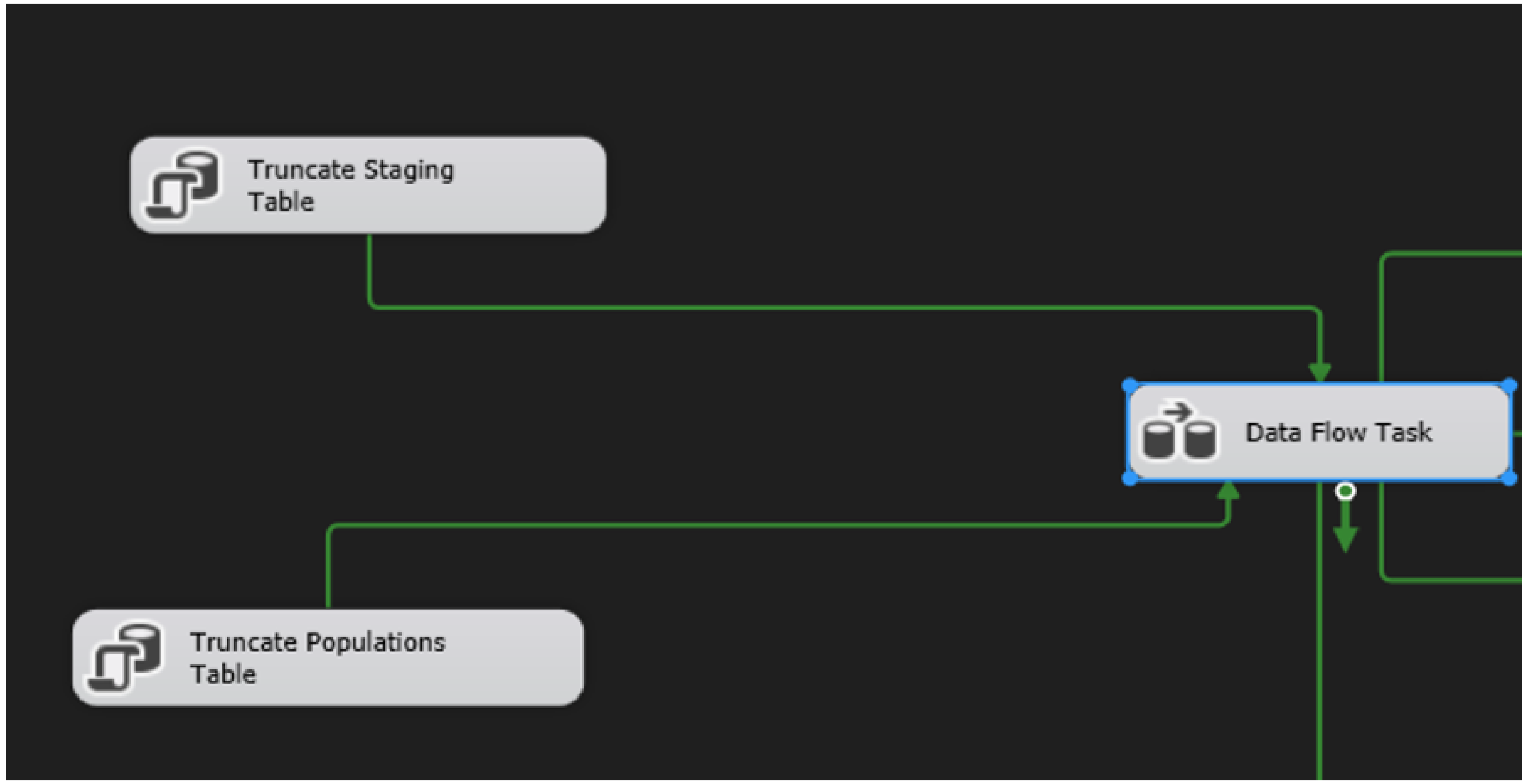
Data Warehousing (1)

We started by creating our relational database in SSMS and introducing a Data Flow Task whose purpose is to create two tables based on the two datasets that we used.



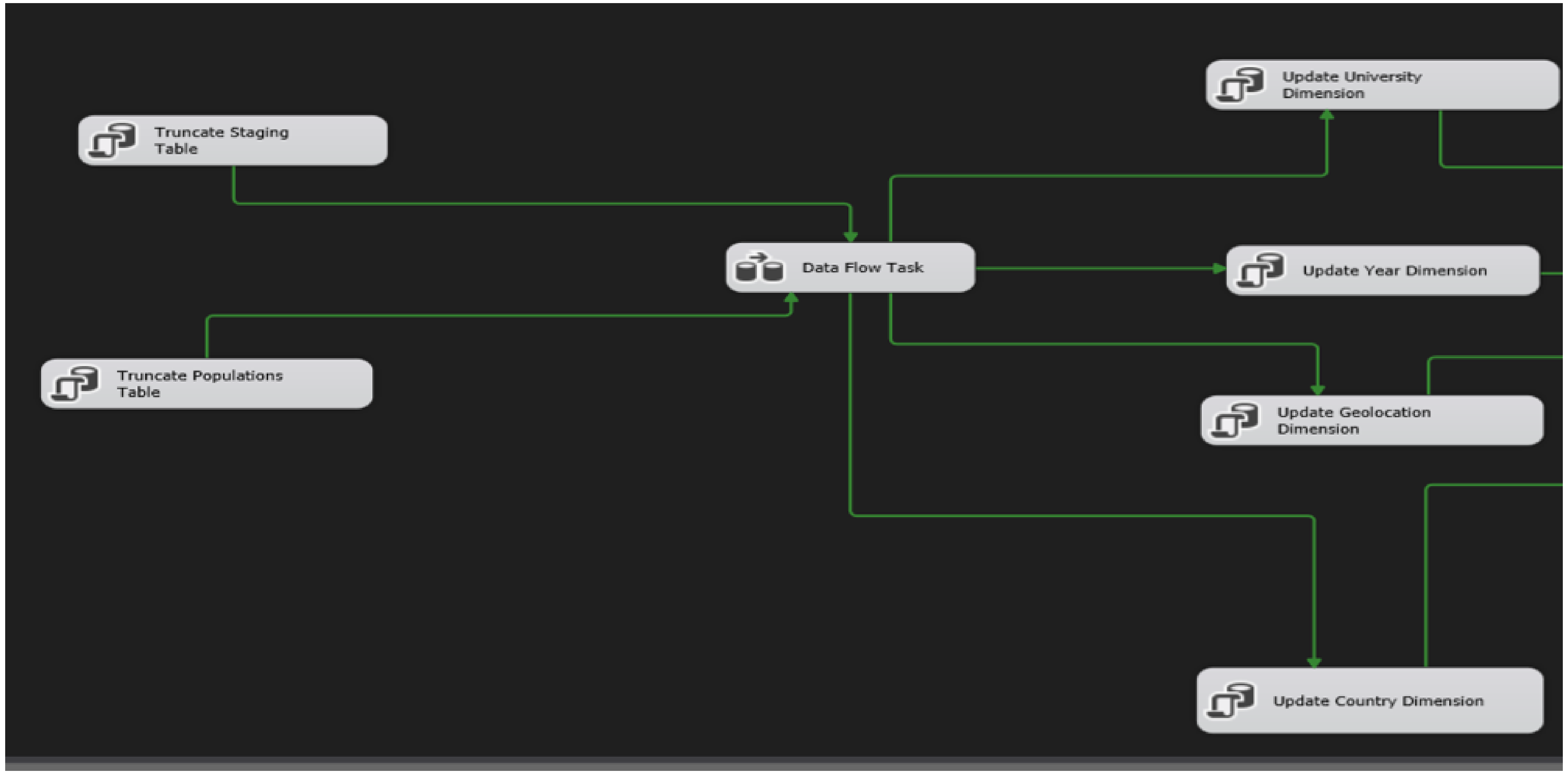
Data Warehousing (2)

The next step was to create two Execute SQL Tasks that are connected to the Data Flow Task and truncate the tables before the data insertion in order to avoid stacking the data on top of itself each time.



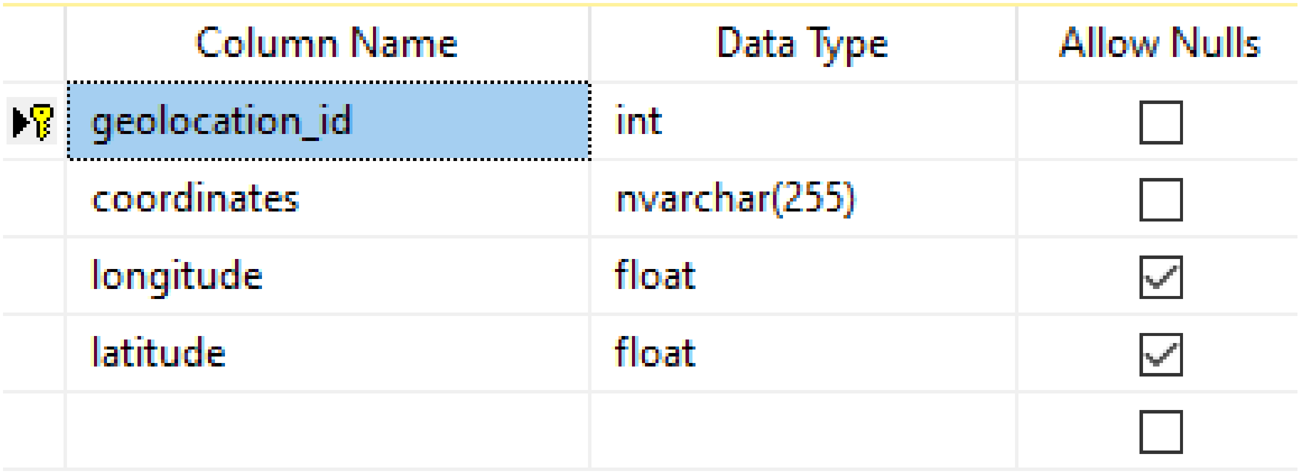
Data Warehousing (3)

Afterwards we created our four dimension tables (Year, University, Country, Geolocation) in SSMS.
Then we created four Execute SQL Tasks that update said tables by inserting the appropriate values.
These tasks are connected to the Data Flow Task.


Truncate Staging Table
Truncate Populations Table
Data Flow Task
Update University Dimension
Update Year Dimension
Update Geolocation Dimension
Update Country Dimension

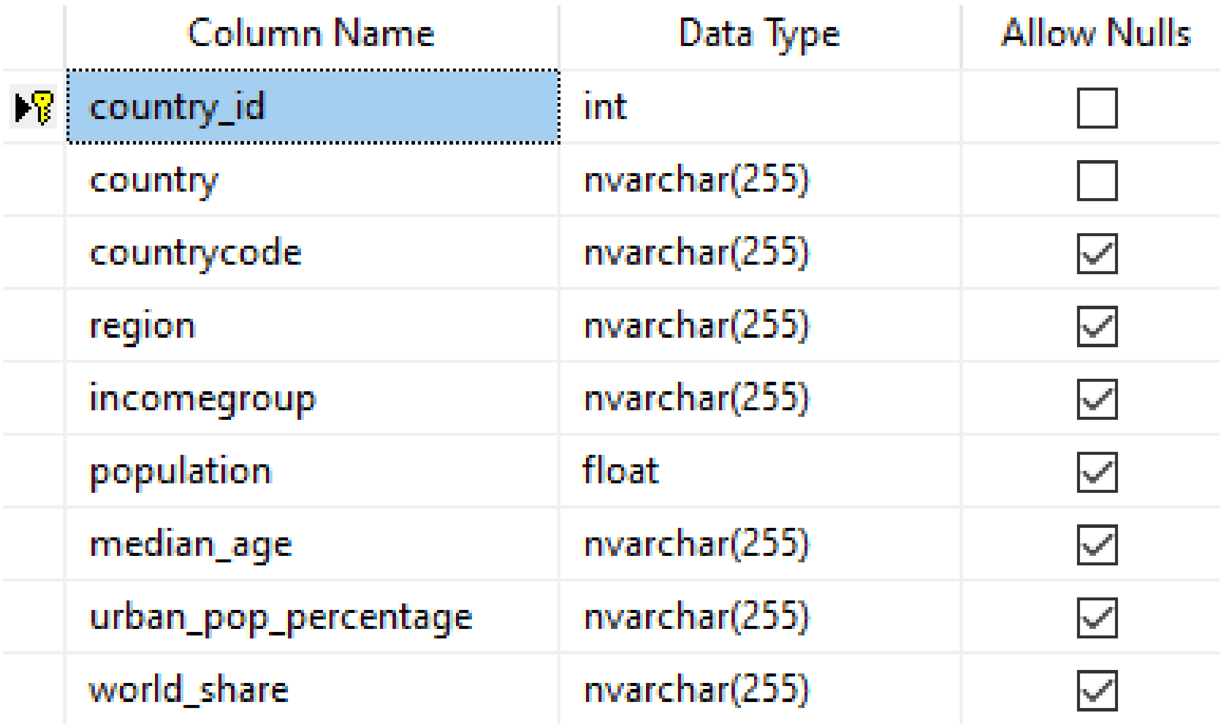
Dimensions

Geolocation Dimension



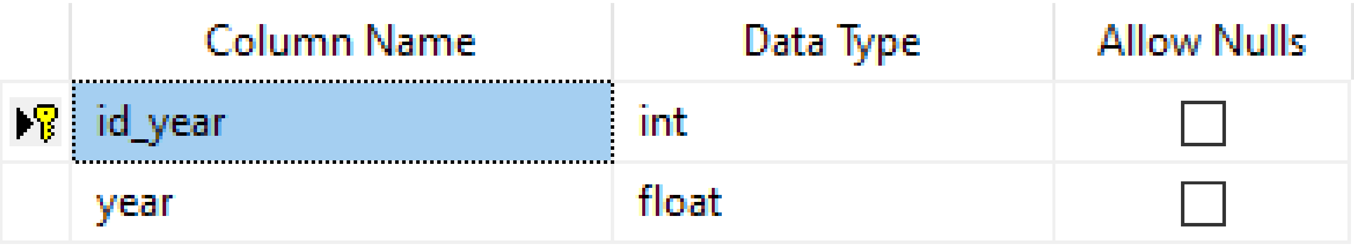
	Column Name	Data Type	Allow Nulls
▶🔑	geolocation_id	int	□
	coordinates	nvarchar(255)	□
	longitude	float	☑
	latitude	float	☑
			□

Country Dimension



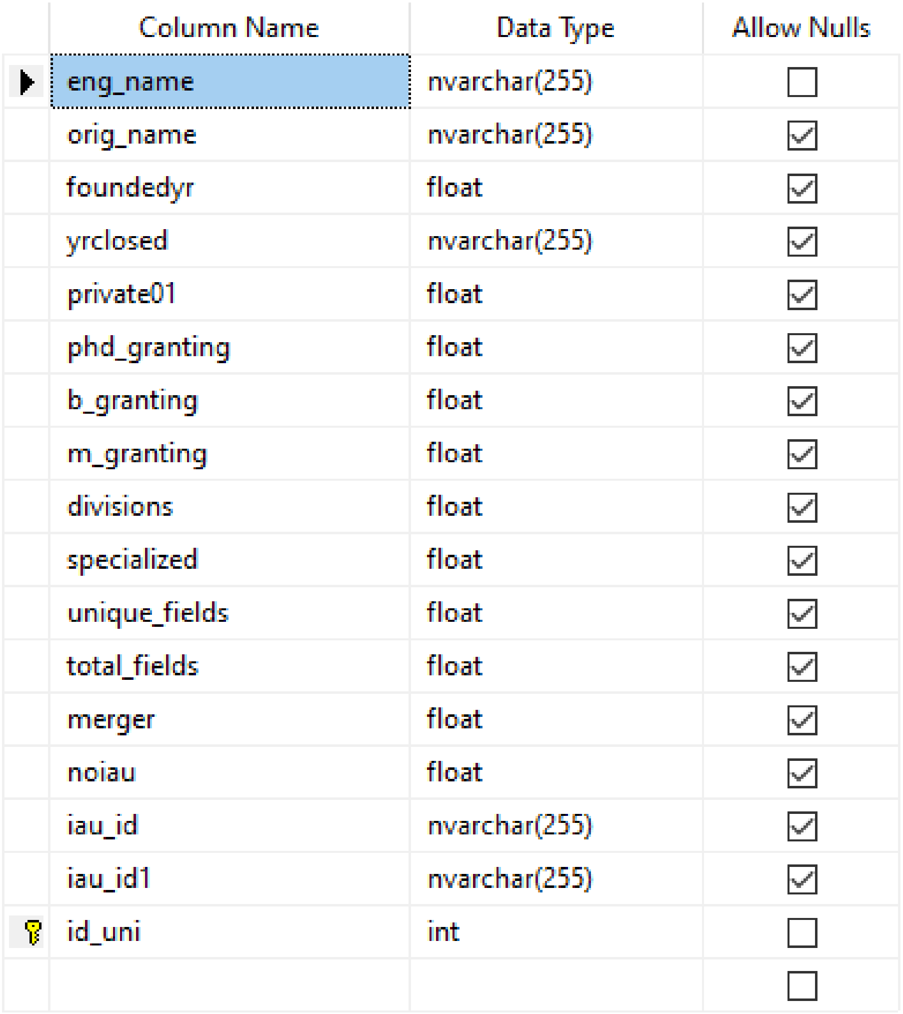
	Column Name	Data Type	Allow Nulls
▶🔑	country_id	int	□
	country	nvarchar(255)	□
	countrycode	nvarchar(255)	☑
	region	nvarchar(255)	☑
	incomegroup	nvarchar(255)	☑
	population	float	☑
	median_age	nvarchar(255)	☑
	urban_pop_percentage	nvarchar(255)	☑
	world_share	nvarchar(255)	☑

Year Dimension



	Column Name	Data Type	Allow Nulls
▶🔑	id_year	int	□
	year	float	□

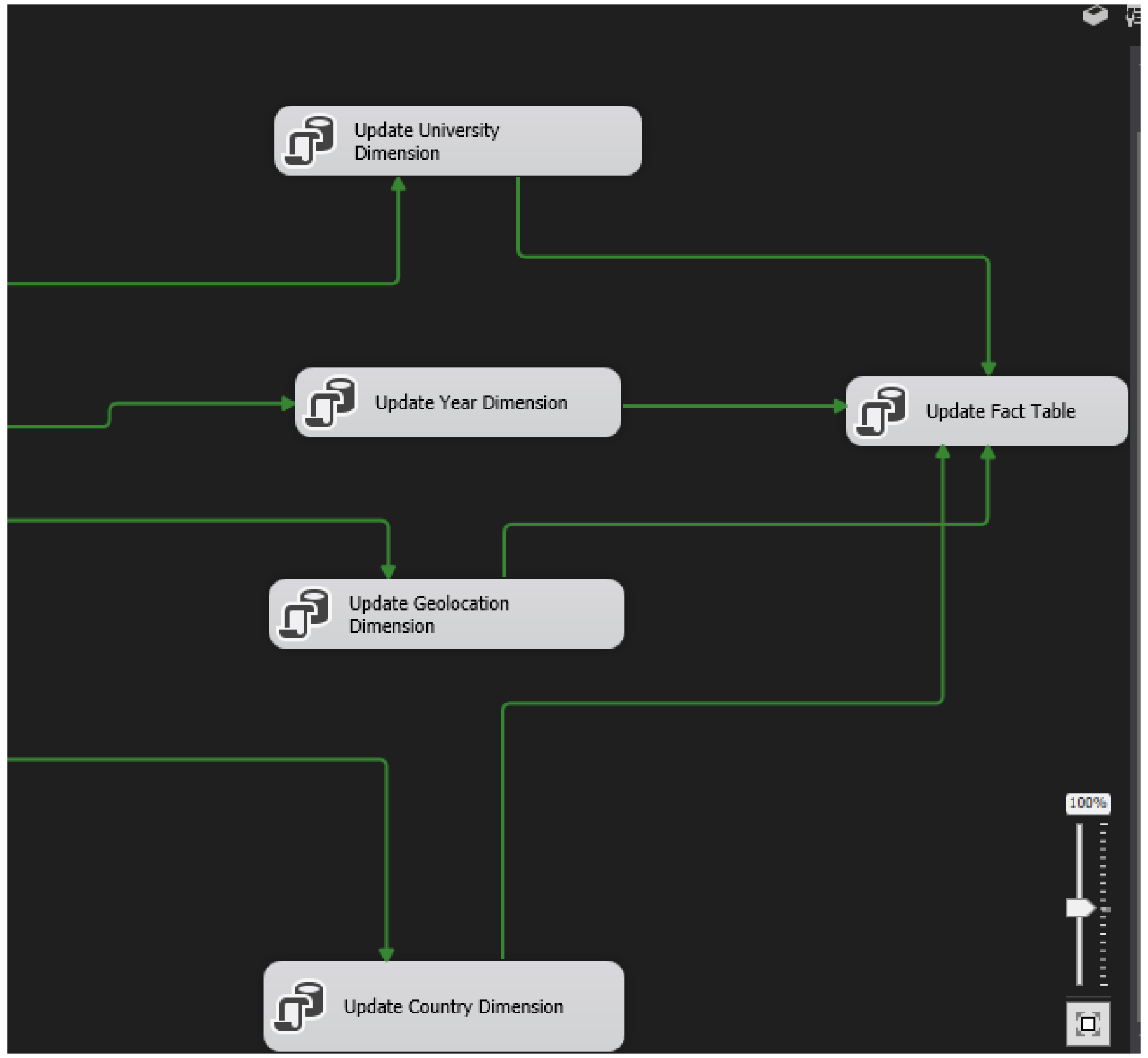
University Dimension



	Column Name	Data Type	Allow Nulls
▶	eng_name	nvarchar(255)	□
	orig_name	nvarchar(255)	☑
	foundedyr	float	☑
	yrclosed	nvarchar(255)	☑
	private01	float	☑
	phd_granting	float	☑
	b_granting	float	☑
	m_granting	float	☑
	divisions	float	☑
	specialized	float	☑
	unique_fields	float	☑
	total_fields	float	☑
	merger	float	☑
	noiau	float	☑
	iau_id	nvarchar(255)	☑
	iau_id1	nvarchar(255)	☑
🔑	id_uni	int	□
			□

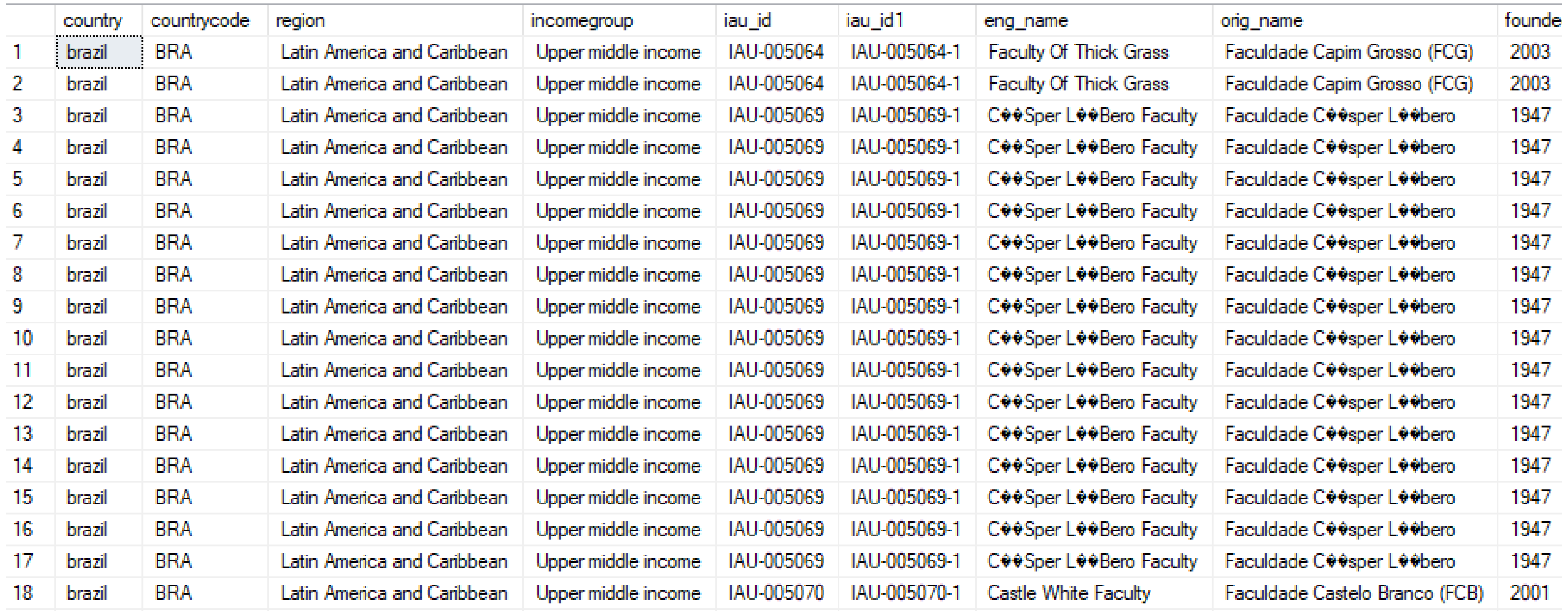
Data Warehousing (4)

Afterwards we created our Fact table in SSMS.
Then we created an Execute SQL Task that joins the dimension ids to the staging table and selects them into the fact table
This task is connected to the Dimension Tasks.

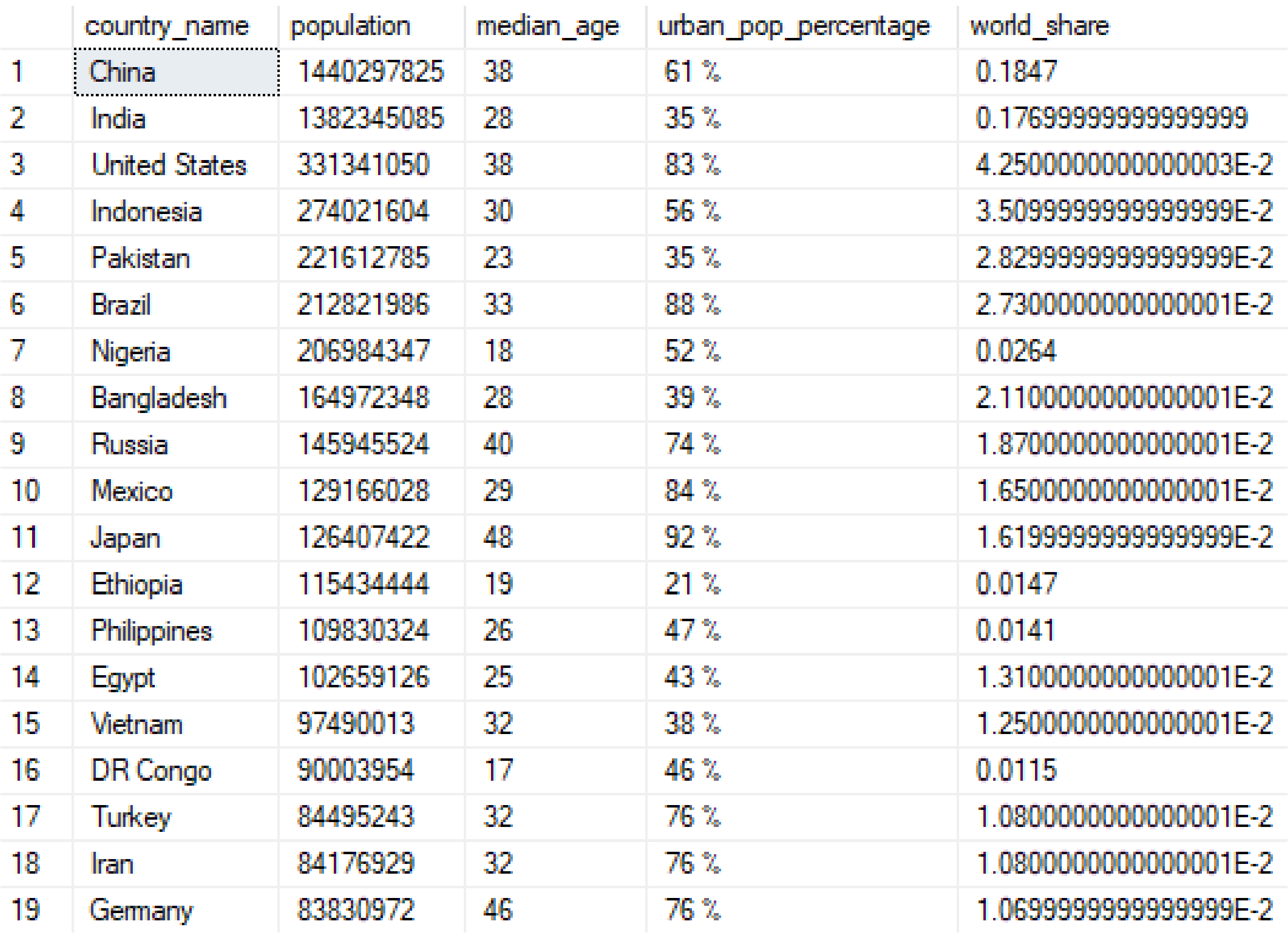


Result tables

We run the processes and filled the SQL Server Tables. Below we can see the enrollment and population tables.

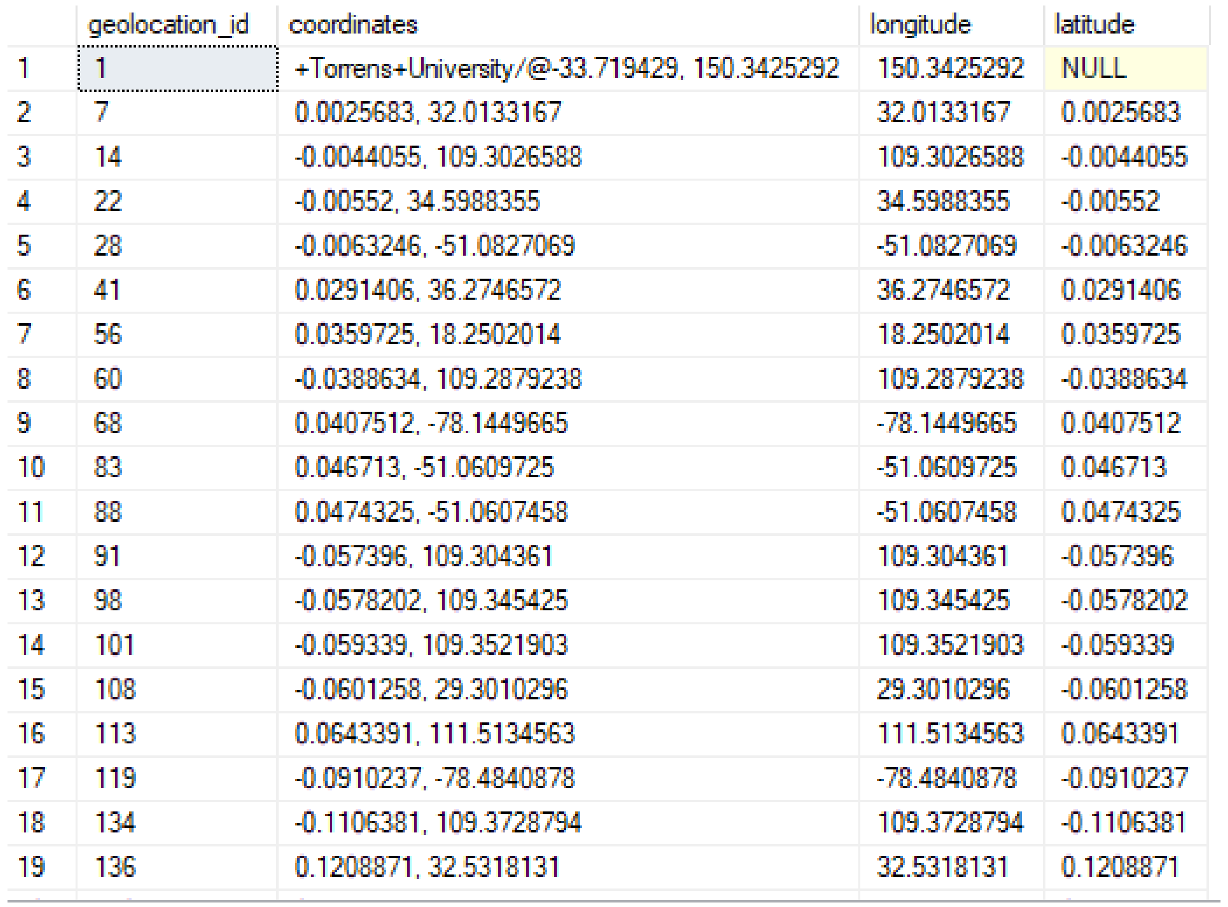


	country	countrycode	region	incomegroup	iau_id	iau_id1	eng_name	orig_name	founde
1	brazil	BRA	Latin America and Caribbean	Upper middle income	IAU-005064	IAU-005064-1	Faculty Of Thick Grass	Faculdade Capim Grosso (FCG)	2003
2	brazil	BRA	Latin America and Caribbean	Upper middle income	IAU-005064	IAU-005064-1	Faculty Of Thick Grass	Faculdade Capim Grosso (FCG)	2003
3	brazil	BRA	Latin America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947
4	brazil	BRA	Latin America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947
5	brazil	BRA	Latin America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947
6	brazil	BRA	Latin America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947
7	brazil	BRA	Latin America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947
8	brazil	BRA	Latin America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947
9	brazil	BRA	Latin America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947
10	brazil	BRA	Latin America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947
11	brazil	BRA	Latin America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947
12	brazil	BRA	Latin America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947
13	brazil	BRA	Latin America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947
14	brazil	BRA	Latin America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947
15	brazil	BRA	Latin America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947
16	brazil	BRA	Latin America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947
17	brazil	BRA	Latin America and Caribbean	Upper middle income	IAU-005069	IAU-005069-1	C��Sper L��Bero Faculty	Faculdade C��sper L��bero	1947
18	brazil	BRA	Latin America and Caribbean	Upper middle income	IAU-005070	IAU-005070-1	Castle White Faculty	Faculdade Castelo Branco (FCB)	2001

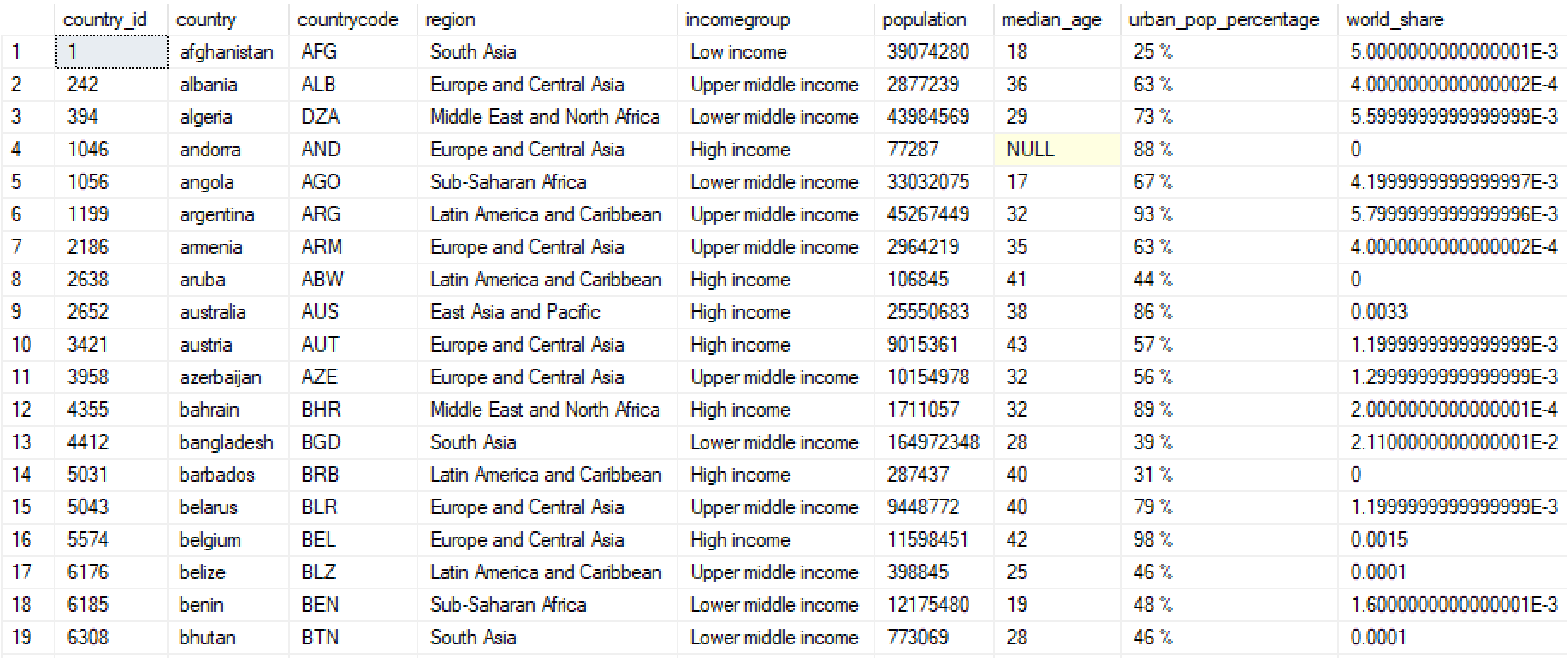


	country_name	population	median_age	urban_pop_percentage	world_share
1	China	1440297825	38	61 %	0.1847
2	India	1382345085	28	35 %	0.17699999999999999
3	United States	331341050	38	83 %	4.2500000000000003E-2
4	Indonesia	274021604	30	56 %	3.5099999999999999E-2
5	Pakistan	221612785	23	35 %	2.8299999999999999E-2
6	Brazil	212821986	33	88 %	2.7300000000000001E-2
7	Nigeria	206984347	18	52 %	0.0264
8	Bangladesh	164972348	28	39 %	2.1100000000000001E-2
9	Russia	145945524	40	74 %	1.8700000000000001E-2
10	Mexico	129166028	29	84 %	1.6500000000000001E-2
11	Japan	126407422	48	92 %	1.6199999999999999E-2
12	Ethiopia	115434444	19	21 %	0.0147
13	Philippines	109830324	26	47 %	0.0141
14	Egypt	102659126	25	43 %	1.3100000000000001E-2
15	Vietnam	97490013	32	38 %	1.2500000000000001E-2
16	DR Congo	90003954	17	46 %	0.0115
17	Turkey	84495243	32	76 %	1.0800000000000001E-2
18	Iran	84176929	32	76 %	1.0800000000000001E-2
19	Germany	83830972	46	76 %	1.0699999999999999E-2

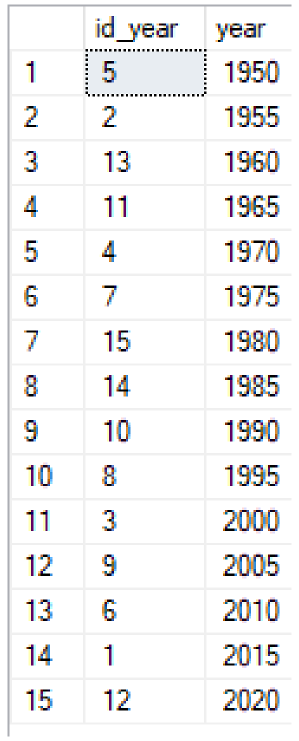
Dimension Tables



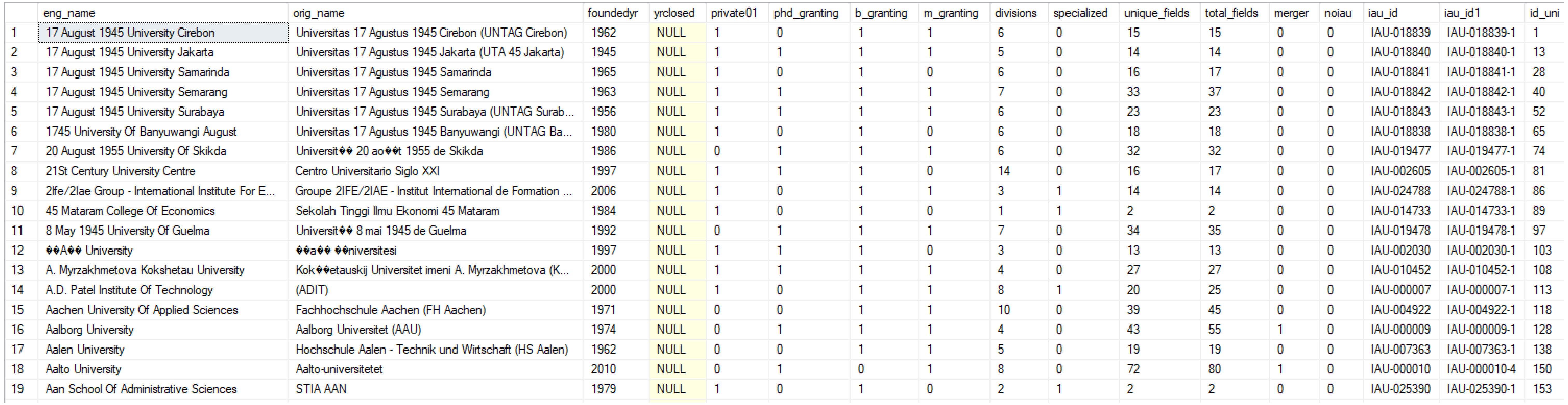
	geolocation_id	coordinates	longitude	latitude
1	1	+Torrens+University/@-33.719429, 150.3425292	150.3425292	NULL
2	7	0.0025683, 32.0133167	32.0133167	0.0025683
3	14	-0.0044055, 109.3026588	109.3026588	-0.0044055
4	22	-0.00552, 34.5988355	34.5988355	-0.00552
5	28	-0.0063246, -51.0827069	-51.0827069	-0.0063246
6	41	0.0291406, 36.2746572	36.2746572	0.0291406
7	56	0.0359725, 18.2502014	18.2502014	0.0359725
8	60	-0.0388634, 109.2879238	109.2879238	-0.0388634
9	68	0.0407512, -78.1449665	-78.1449665	0.0407512
10	83	0.046713, -51.0609725	-51.0609725	0.046713
11	88	0.0474325, -51.0607458	-51.0607458	0.0474325
12	91	-0.057396, 109.304361	109.304361	-0.057396
13	98	-0.0578202, 109.345425	109.345425	-0.0578202
14	101	-0.059339, 109.3521903	109.3521903	-0.059339
15	108	-0.0601258, 29.3010296	29.3010296	-0.0601258
16	113	0.0643391, 111.5134563	111.5134563	0.0643391
17	119	-0.0910237, -78.4840878	-78.4840878	-0.0910237
18	134	-0.1106381, 109.3728794	109.3728794	-0.1106381
19	136	0.1208871, 32.5318131	32.5318131	0.1208871



	country_id	country	countrycode	region	incomegroup	population	median_age	urban_pop_percentage	world_share
1	1	afghanistan	AFG	South Asia	Low income	39074280	18	25 %	5.0000000000000001E-3
2	242	albania	ALB	Europe and Central Asia	Upper middle income	2877239	36	63 %	4.0000000000000002E-4
3	394	algeria	DZA	Middle East and North Africa	Lower middle income	43984569	29	73 %	5.5999999999999999E-3
4	1046	andorra	AND	Europe and Central Asia	High income	77287	NULL	88 %	0
5	1056	angola	AGO	Sub-Saharan Africa	Lower middle income	33032075	17	67 %	4.1999999999999997E-3
6	1199	argentina	ARG	Latin America and Caribbean	Upper middle income	45267449	32	93 %	5.7999999999999996E-3
7	2186	armenia	ARM	Europe and Central Asia	Upper middle income	2964219	35	63 %	4.0000000000000002E-4
8	2638	aruba	ABW	Latin America and Caribbean	High income	106845	41	44 %	0
9	2652	australia	AUS	East Asia and Pacific	High income	25550683	38	86 %	0.0033
10	3421	austria	AUT	Europe and Central Asia	High income	9015361	43	57 %	1.1999999999999999E-3
11	3958	azerbaijan	AZE	Europe and Central Asia	Upper middle income	10154978	32	56 %	1.2999999999999999E-3
12	4355	bahrain	BHR	Middle East and North Africa	High income	1711057	32	89 %	2.0000000000000001E-4
13	4412	bangladesh	BGD	South Asia	Lower middle income	164972348	28	39 %	2.1100000000000001E-2
14	5031	barbados	BRB	Latin America and Caribbean	High income	287437	40	31 %	0
15	5043	belarus	BLR	Europe and Central Asia	Upper middle income	9448772	40	79 %	1.1999999999999999E-3
16	5574	belgium	BEL	Europe and Central Asia	High income	11598451	42	98 %	0.0015
17	6176	belize	BLZ	Latin America and Caribbean	Upper middle income	398845	25	46 %	0.0001
18	6185	benin	BEN	Sub-Saharan Africa	Lower middle income	12175480	19	48 %	1.6000000000000001E-3
19	6308	bhutan	BTN	South Asia	Lower middle income	773069	28	46 %	0.0001

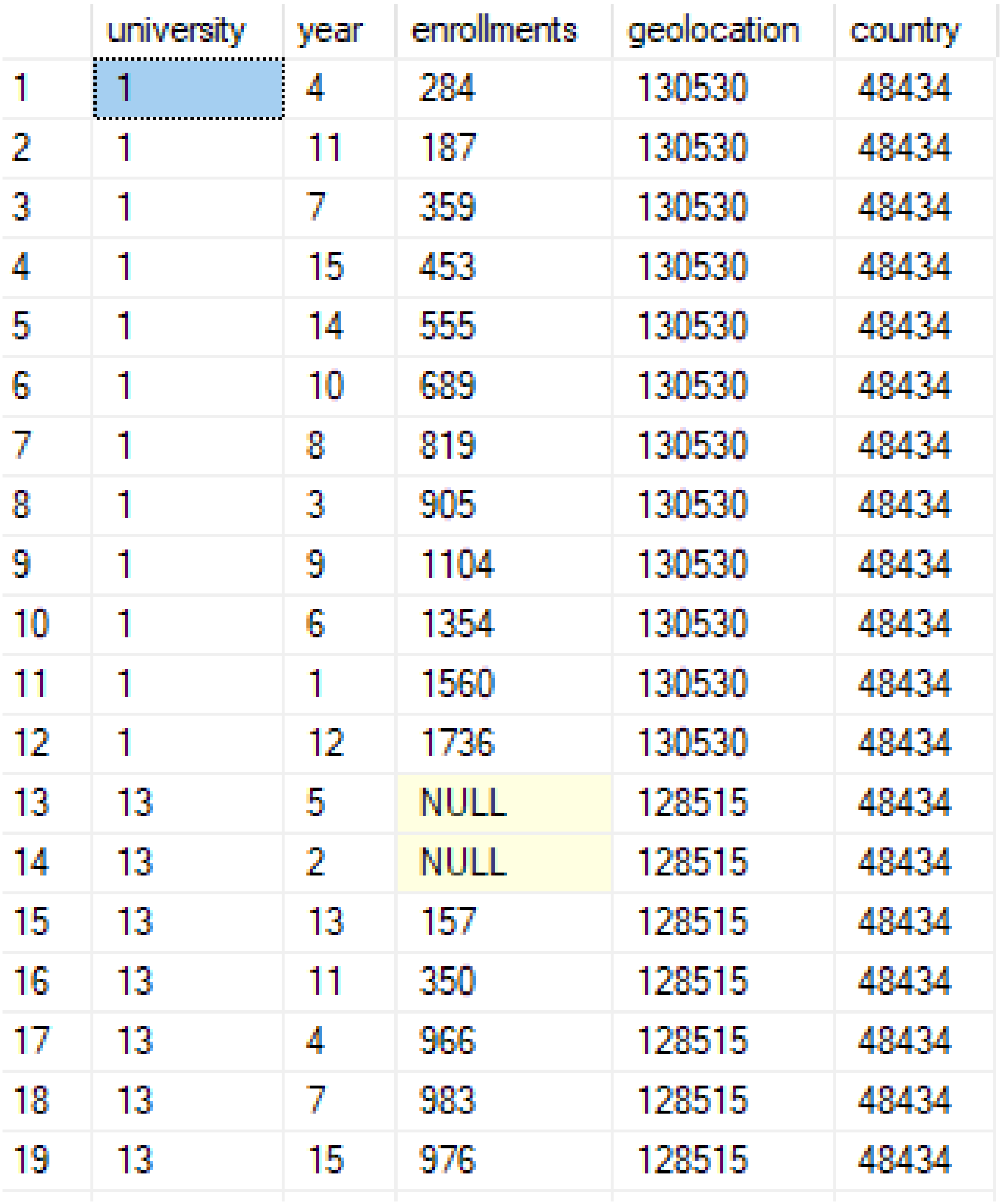


	id_year	year
1	5	1950
2	2	1955
3	13	1960
4	11	1965
5	4	1970
6	7	1975
7	15	1980
8	14	1985
9	10	1990
10	8	1995
11	3	2000
12	9	2005
13	6	2010
14	1	2015
15	12	2020



	eng_name	orig_name	foundedyr	yrclosed	private01	phd_granting	b_granting	m_granting	divisions	specialized	unique_fields	total_fields	merger	noiau	iau_id	iau_id1	id_uni
1	17 August 1945 University Cirebon	Universitas 17 Agustus 1945 Cirebon (UNTAG Cirebon)	1962	NULL	1	0	1	1	6	0	15	15	0	0	IAU-018839	IAU-018839-1	1
2	17 August 1945 University Jakarta	Universitas 17 Agustus 1945 Jakarta (UTA 45 Jakarta)	1945	NULL	1	1	1	1	5	0	14	14	0	0	IAU-018840	IAU-018840-1	13
3	17 August 1945 University Samarinda	Universitas 17 Agustus 1945 Samarinda	1965	NULL	1	0	1	0	6	0	16	17	0	0	IAU-018841	IAU-018841-1	28
4	17 August 1945 University Semarang	Universitas 17 Agustus 1945 Semarang	1963	NULL	1	1	1	1	7	0	33	37	0	0	IAU-018842	IAU-018842-1	40
5	17 August 1945 University Surabaya	Universitas 17 Agustus 1945 Surabaya (UNTAG Surab...	1956	NULL	1	1	1	1	6	0	23	23	0	0	IAU-018843	IAU-018843-1	52
6	1745 University Of Banyuwangi August	Universitas 17 Agustus 1945 Banyuwangi (UNTAG Ba...	1980	NULL	1	0	1	0	6	0	18	18	0	0	IAU-018838	IAU-018838-1	65
7	20 August 1955 University Of Skikda	Universit�� 20 ao��t 1955 de Skikda	1986	NULL	0	1	1	1	6	0	32	32	0	0	IAU-019477	IAU-019477-1	74
8	21St Century University Centre	Centro Universitario Siglo XXI	1997	NULL	1	1	1	0	14	0	16	17	0	0	IAU-002605	IAU-002605-1	81
9	2Ife/2Iae Group - International Institute For E...	Groupe 2IFE/2IAE - Institut International de Formation ...	2006	NULL	1	0	1	1	3	1	14	14	0	0	IAU-024788	IAU-024788-1	86
10	45 Mataram College Of Economics	Sekolah Tinggi Ilmu Ekonomi 45 Mataram	1984	NULL	1	0	1	0	1	1	2	2	0	0	IAU-014733	IAU-014733-1	89
11	8 May 1945 University Of Guelma	Universit�� 8 mai 1945 de Guelma	1992	NULL	0	1	1	1	7	0	34	35	0	0	IAU-019478	IAU-019478-1	97
12	��A�� University	��a�� ��niversitesi	1997	NULL	1	1	1	0	3	0	13	13	0	0	IAU-002030	IAU-002030-1	103
13	A. Myrzakhmetova Kokshetau University	Kok��etauskij Universitet imeni A. Myrzakhmetova (K...	2000	NULL	1	1	1	1	4	0	27	27	0	0	IAU-010452	IAU-010452-1	108
14	A.D. Patel Institute Of Technology	(ADIT)	2000	NULL	1	0	1	1	8	1	20	25	0	0	IAU-000007	IAU-000007-1	113
15	Aachen University Of Applied Sciences	Fachhochschule Aachen (FH Aachen)	1971	NULL	0	0	1	1	10	0	39	45	0	0	IAU-004922	IAU-004922-1	118
16	Aalborg University	Aalborg Universitet (AAU)	1974	NULL	0	1	1	1	4	0	43	55	1	0	IAU-000009	IAU-000009-1	128
17	Aalen University	Hochschule Aalen - Technik und Wirtschaft (HS Aalen)	1962	NULL	0	0	1	1	5	0	19	19	0	0	IAU-007363	IAU-007363-1	138
18	Aalto University	Aalto-universitetet	2010	NULL	0	1	0	1	8	0	72	80	1	0	IAU-000010	IAU-000010-4	150
19	Aan School Of Administrative Sciences	STIA AAN	1979	NULL	1	0	1	0	2	1	2	2	0	0	IAU-025390	IAU-025390-1	153

Fact Table



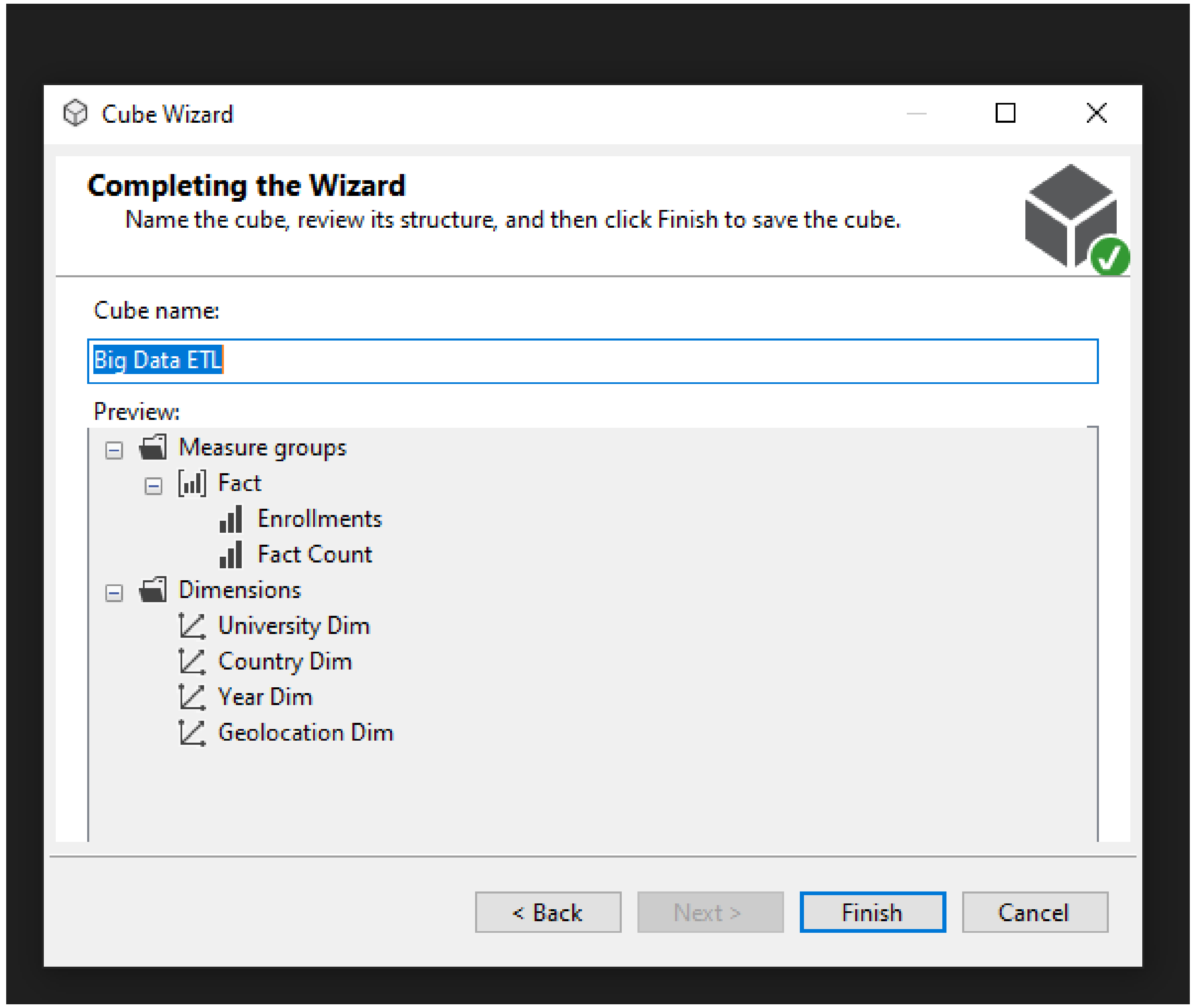
	university	year	enrollments	geolocation	country
1	1	4	284	130530	48434
2	1	11	187	130530	48434
3	1	7	359	130530	48434
4	1	15	453	130530	48434
5	1	14	555	130530	48434
6	1	10	689	130530	48434
7	1	8	819	130530	48434
8	1	3	905	130530	48434
9	1	9	1104	130530	48434
10	1	6	1354	130530	48434
11	1	1	1560	130530	48434
12	1	12	1736	130530	48434
13	13	5	NULL	128515	48434
14	13	2	NULL	128515	48434
15	13	13	157	128515	48434
16	13	11	350	128515	48434
17	13	4	966	128515	48434
18	13	7	983	128515	48434
19	13	15	976	128515	48434

Cubes

We used Microsoft Visual Studios Analysis Services Multidimensional Project to load our views from SQL Server, create process and deploy the cube of our data

Cube (1)

We added our dimensions and fact table to create the view from our SQL Server Data Source. We set as measures the enrollments and fact count



Cube Wizard

Completing the Wizard

Name the cube, review its structure, and then click Finish to save the cube.

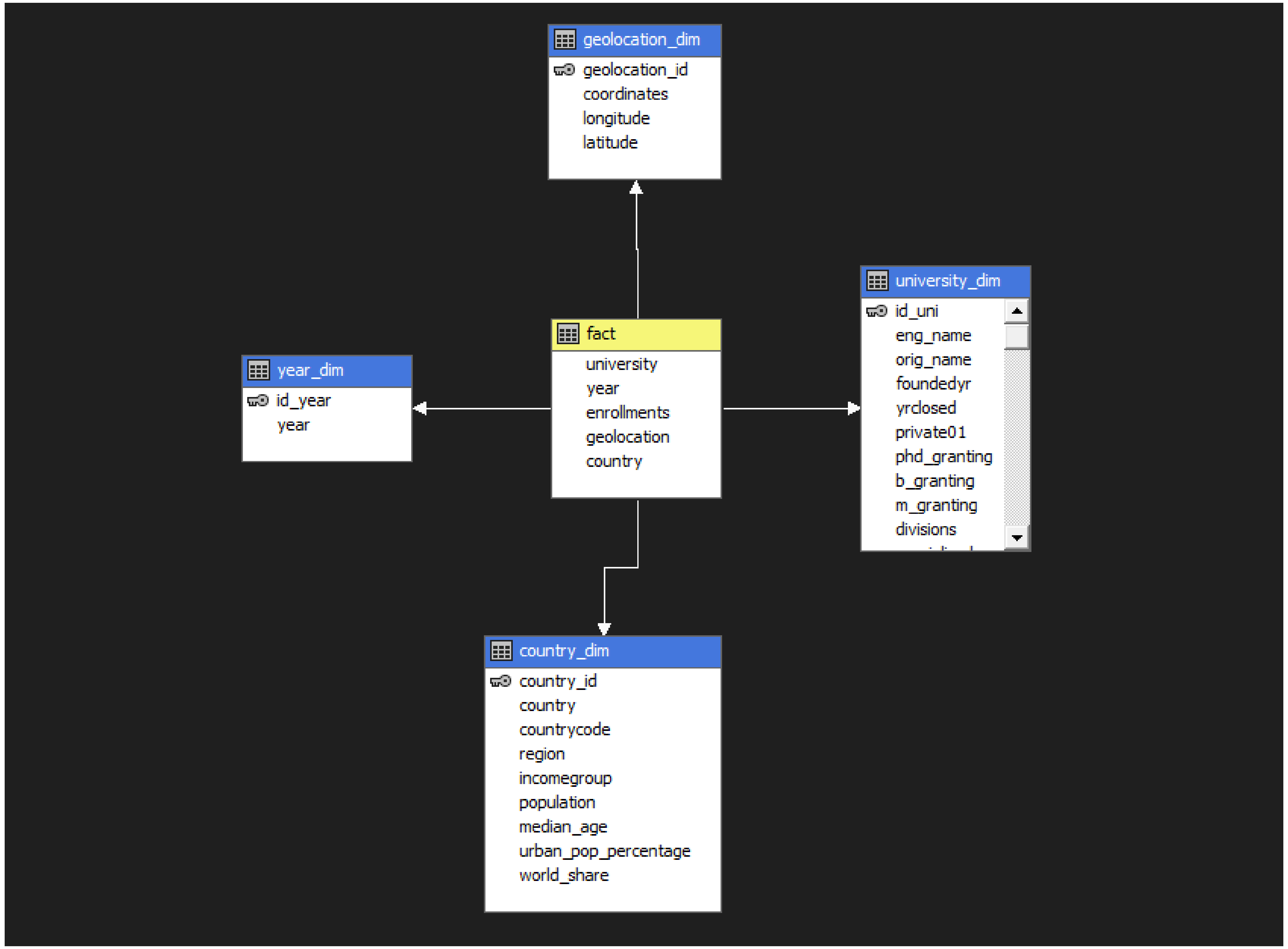
Cube name: Big Data ETL

Preview:

- Measure groups
  - Fact
    - Enrollments
    - Fact Count
- Dimensions
  - University Dim
  - Country Dim
  - Year Dim
  - Geolocation Dim

Star Schema

We can see below the star schema that was created by Visual Studio showcasing our dimensions and measurements



geolocation_dim
- geolocation_id
- coordinates
- longitude
- latitude

year_dim
- id_year
- year

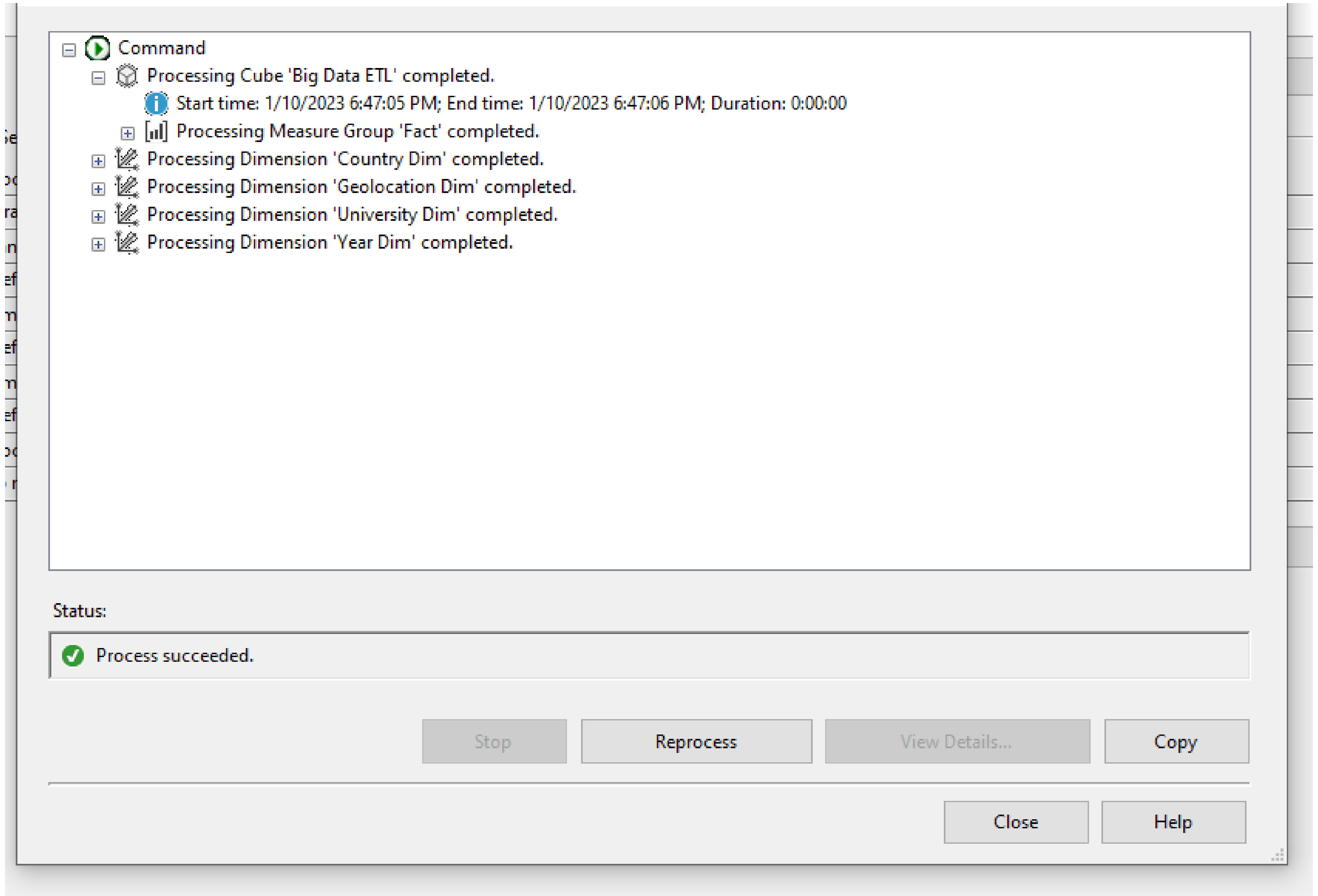
fact
- university
- year
- enrollments
- geolocation
- country

university_dim
- id_uni
- eng_name
- orig_name
- foundedyr
- yrclosed
- private01
- phd_granting
- b_granting
- m_granting
- divisions

country_dim
- country_id
- country
- countrycode
- region
- incomegroup
- population
- median_age
- urban_pop_percentage
- world_share

Cube (2)

We processed the cube and ran it succesfully



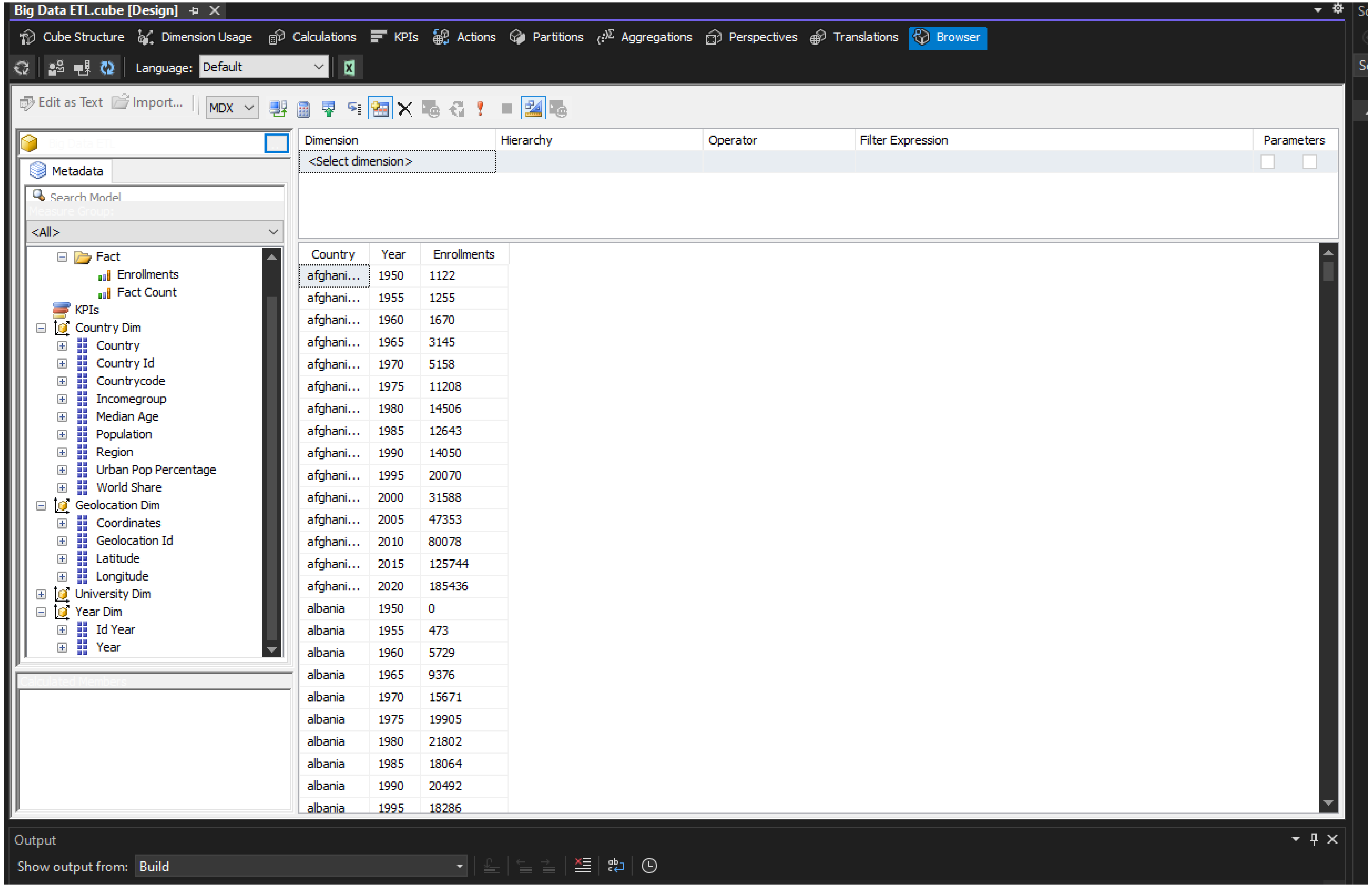
```
⊟ Command
  ⊟ Processing Cube 'Big Data ETL' completed.
        Start time: 1/10/2023 6:47:05 PM; End time: 1/10/2023 6:47:06 PM; Duration: 0:00:00
    ⊞ Processing Measure Group 'Fact' completed.
  ⊞ Processing Dimension 'Country Dim' completed.
  ⊞ Processing Dimension 'Geolocation Dim' completed.
  ⊞ Processing Dimension 'University Dim' completed.
  ⊞ Processing Dimension 'Year Dim' completed.
```

Status:

✓ Process succeeded.

Cube (3)

We can browse it to see it's working correctly



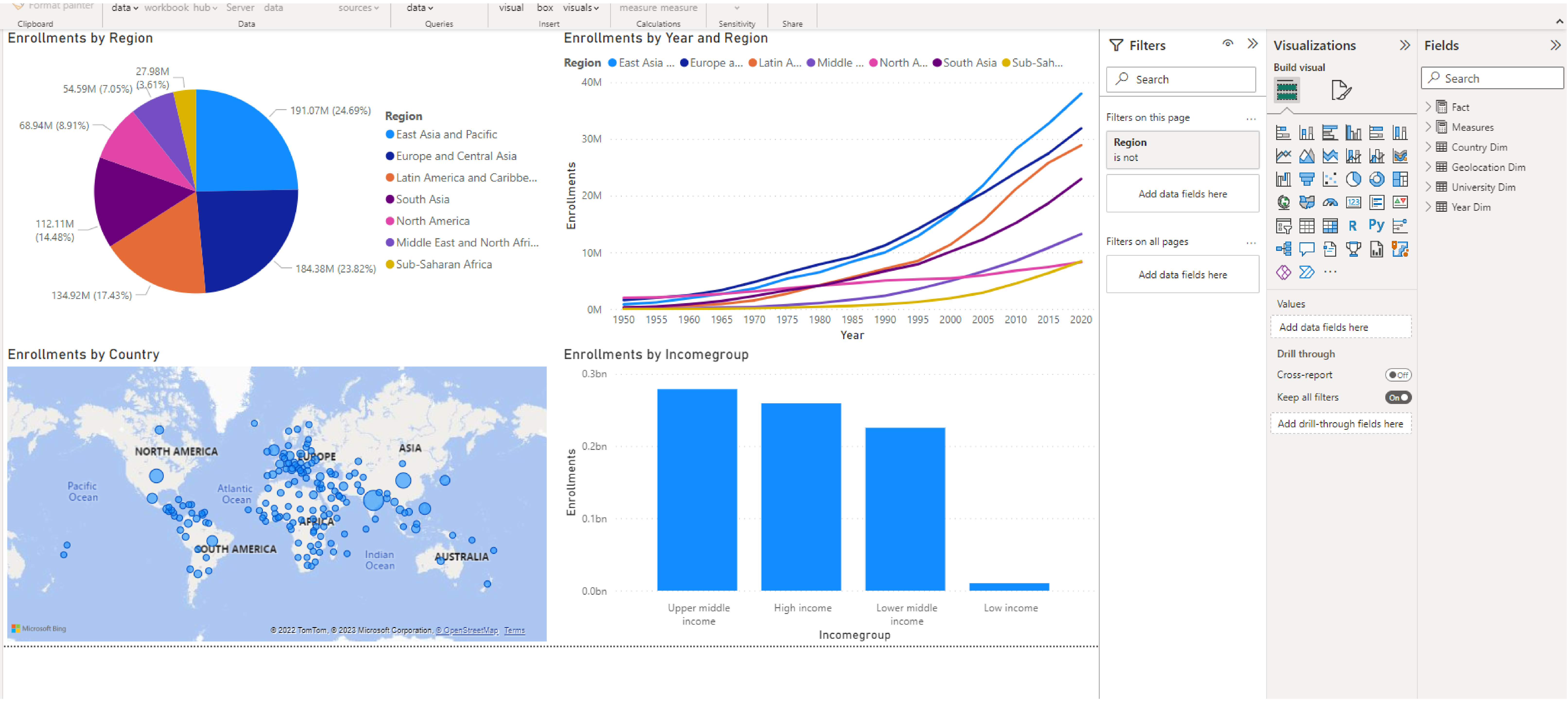
Country	Year	Enrollments
afghani...	1950	1122
afghani...	1955	1255
afghani...	1960	1670
afghani...	1965	3145
afghani...	1970	5158
afghani...	1975	11208
afghani...	1980	14506
afghani...	1985	12643
afghani...	1990	14050
afghani...	1995	20070
afghani...	2000	31588
afghani...	2005	47353
afghani...	2010	80078
afghani...	2015	125744
afghani...	2020	185436
albania	1950	0
albania	1955	473
albania	1960	5729
albania	1965	9376
albania	1970	15671
albania	1975	19905
albania	1980	21802
albania	1985	18064
albania	1990	20492
albania	1995	18286

Data Visualization

We connected our Analysis Services Project, our cube specifically to PowerBi Desktop to create dashboards and Visualizations

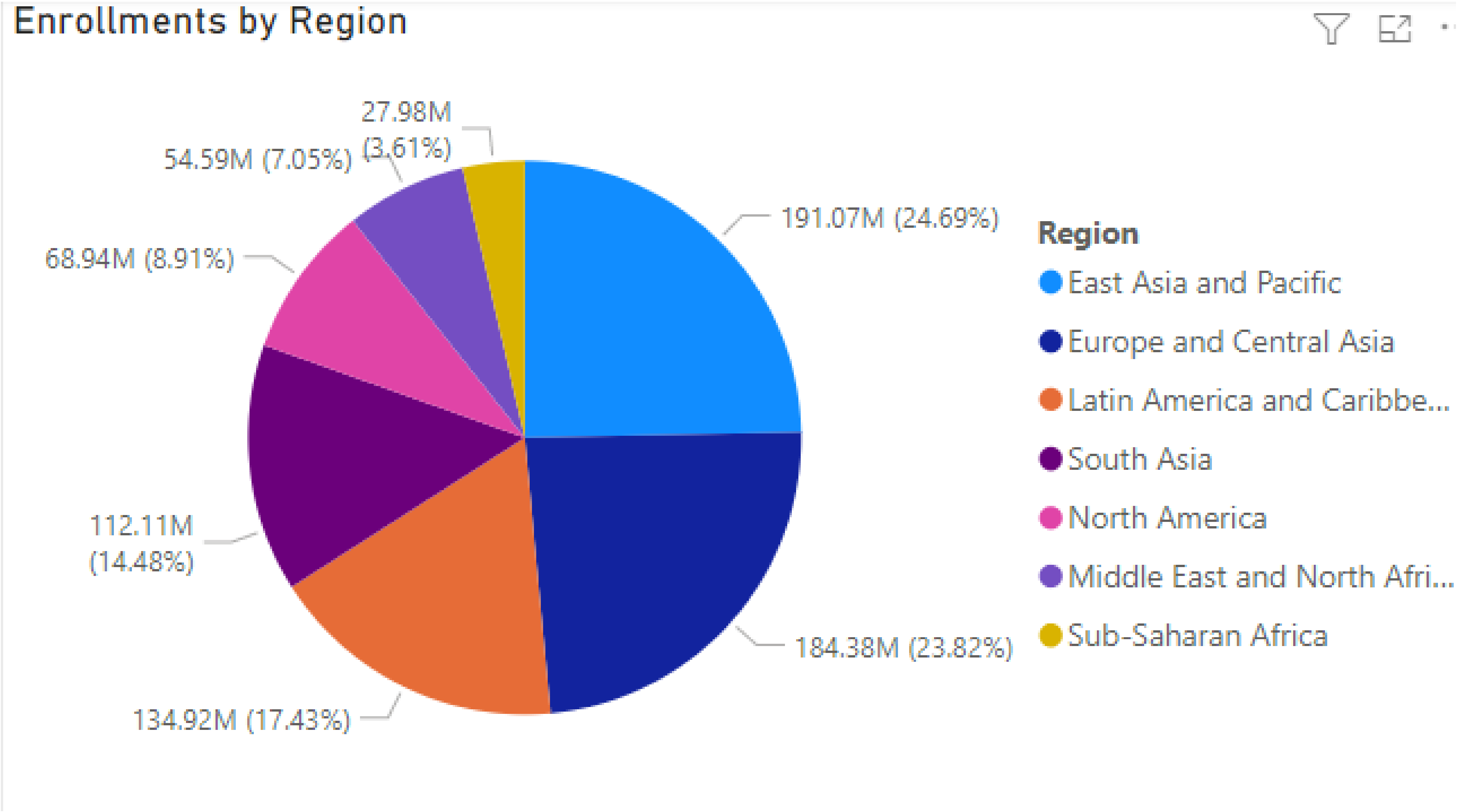
Visualizations (1)

We firstly created the dashboard below



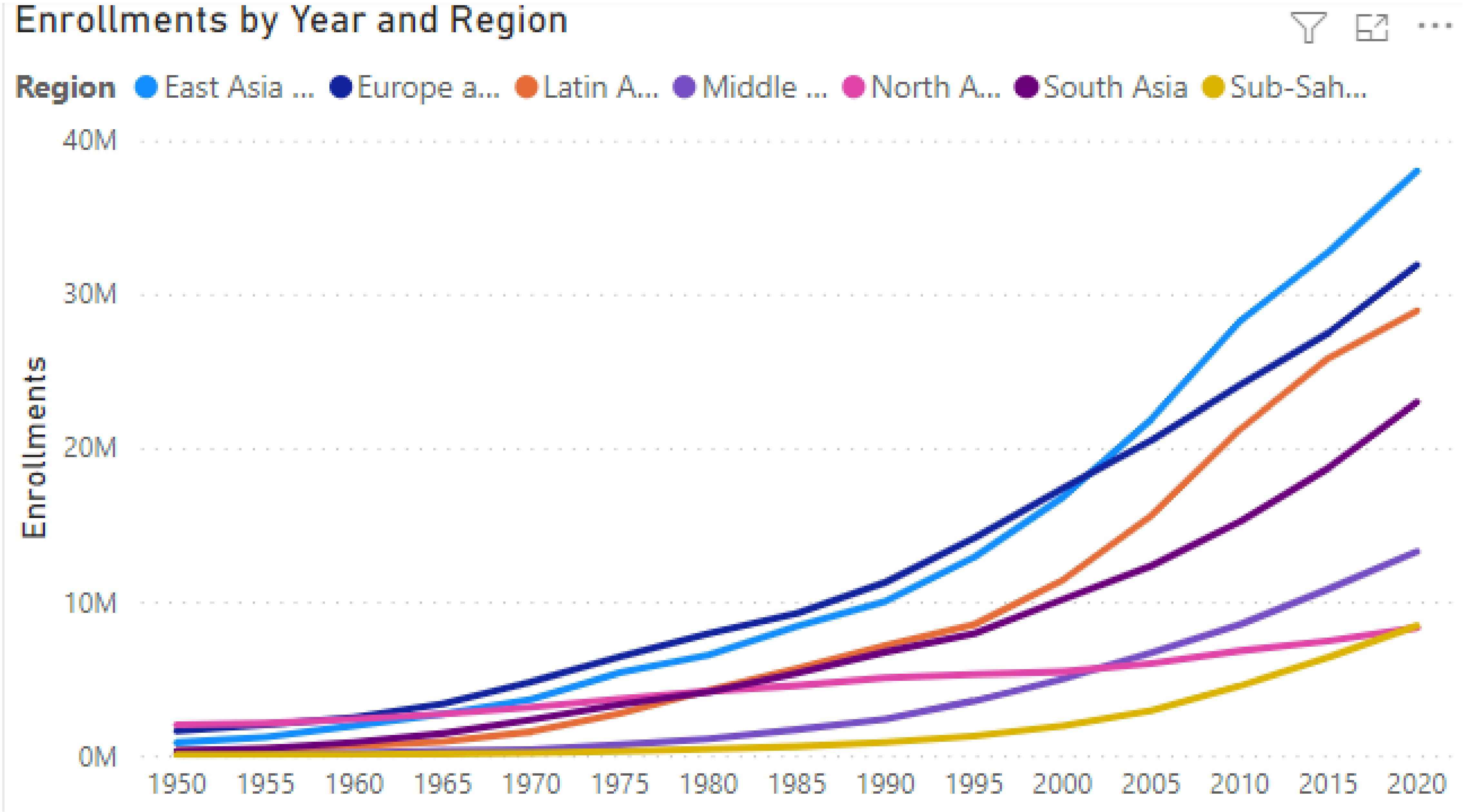
Visualizations (2)

On this pie chart we can see the percentage of the total institution enrollments by region (East Asia and Pasific, South Asia etc)



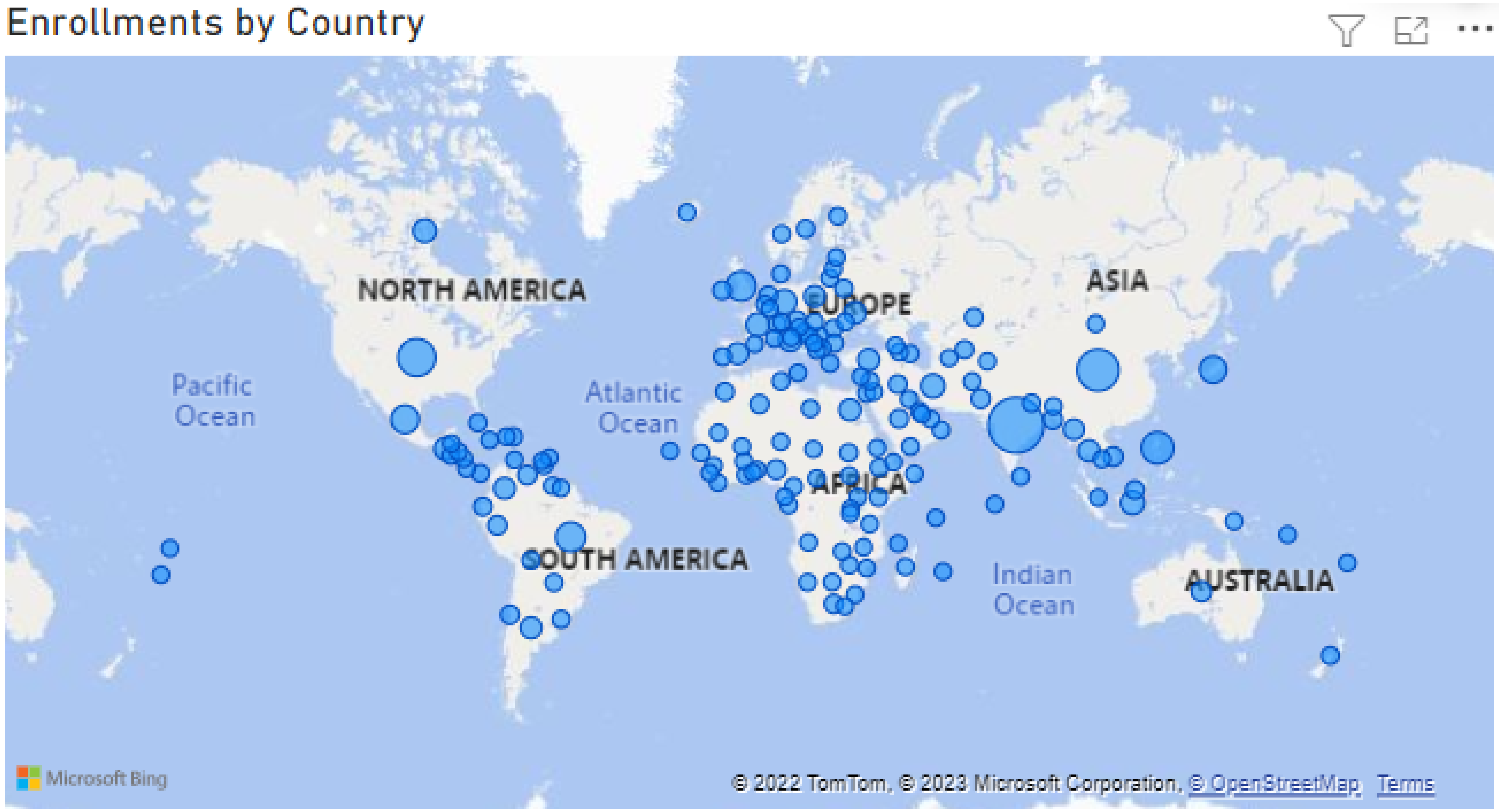
Visualizations (3)

On this line chart we can see the increase of the total institution enrollments by region (East Asia and Pasific, South Asia etc) through the years



Visualizations (4)

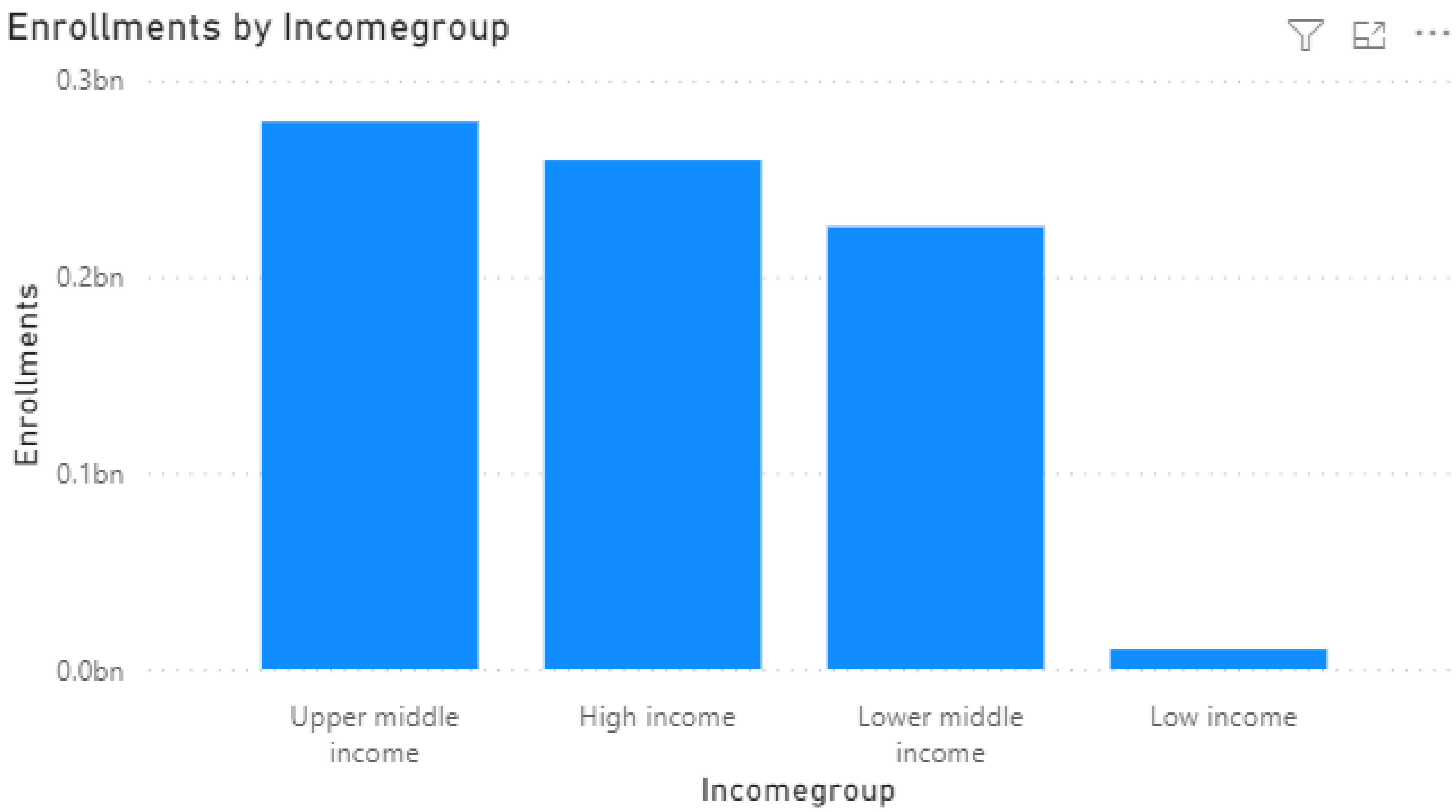
On this map chart we can see the amount of enrollments for every country comparing to all the others, by the size of the dot on the map



Enrollments by Country

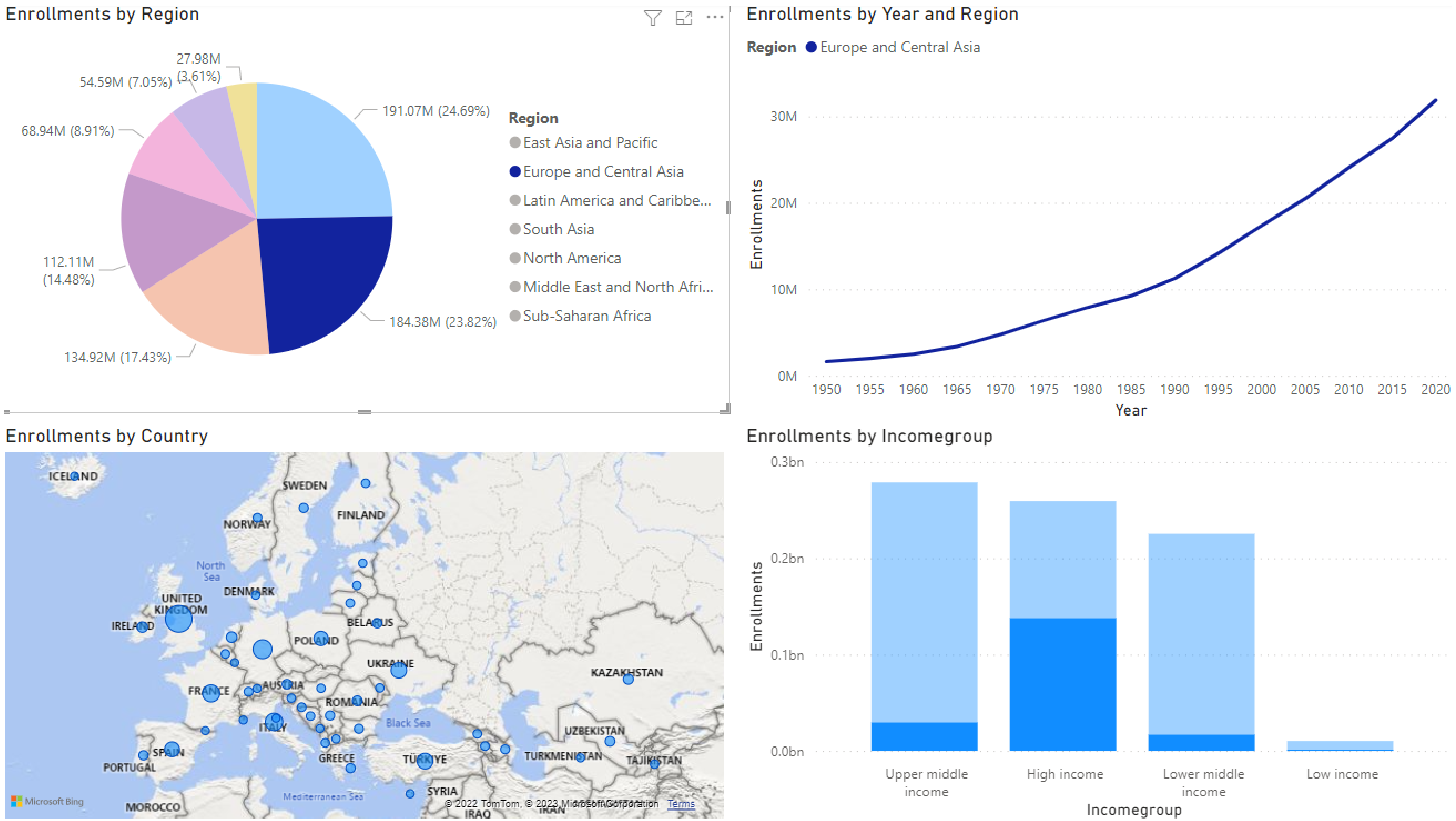
Visualizations (5)

On this column chart we can see the amount of enrollments for every income group worldwide



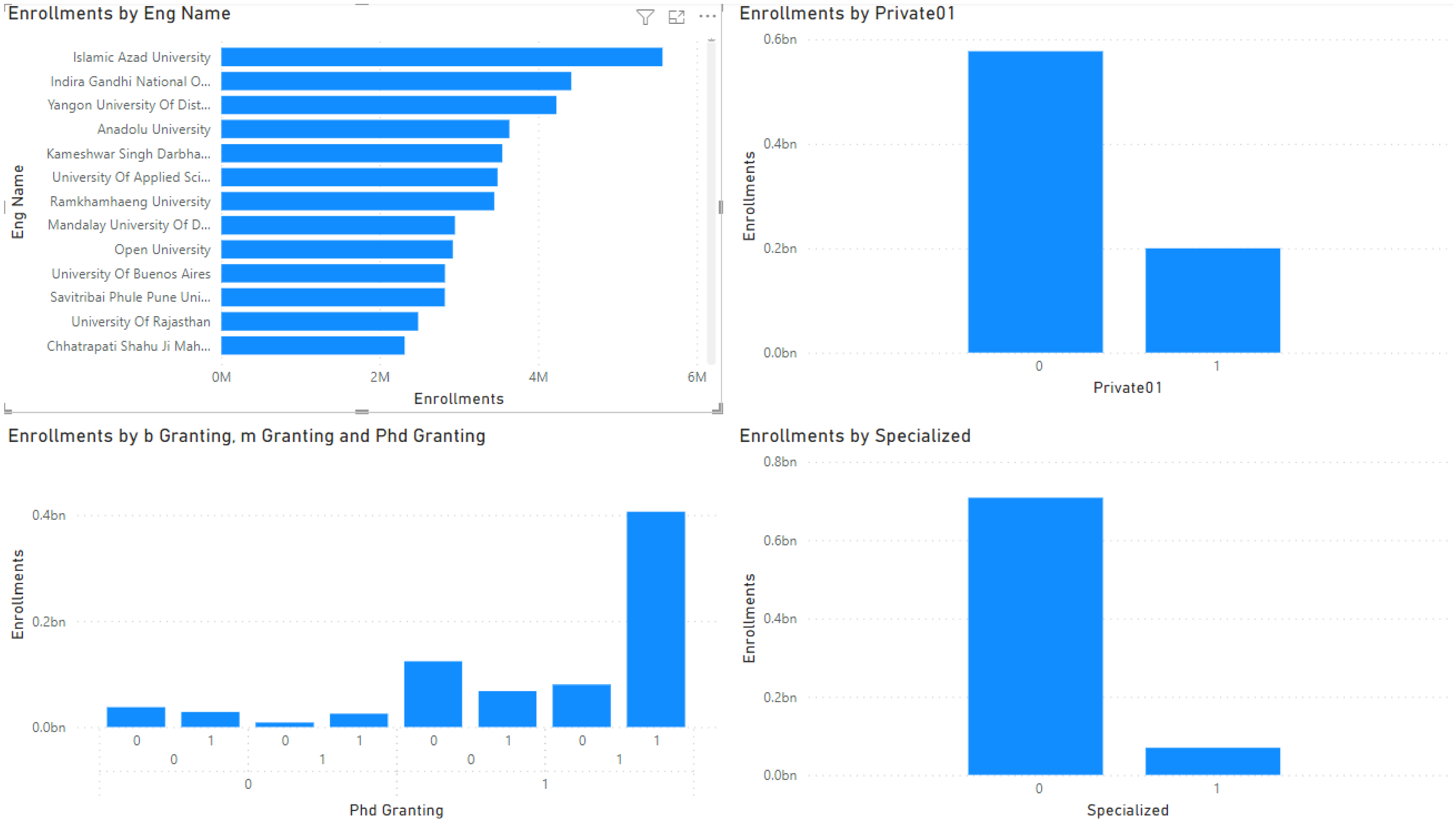
Visualizations (6)

If we click on one of the regions of the pie chart, we can see the information specific to this region and its countries



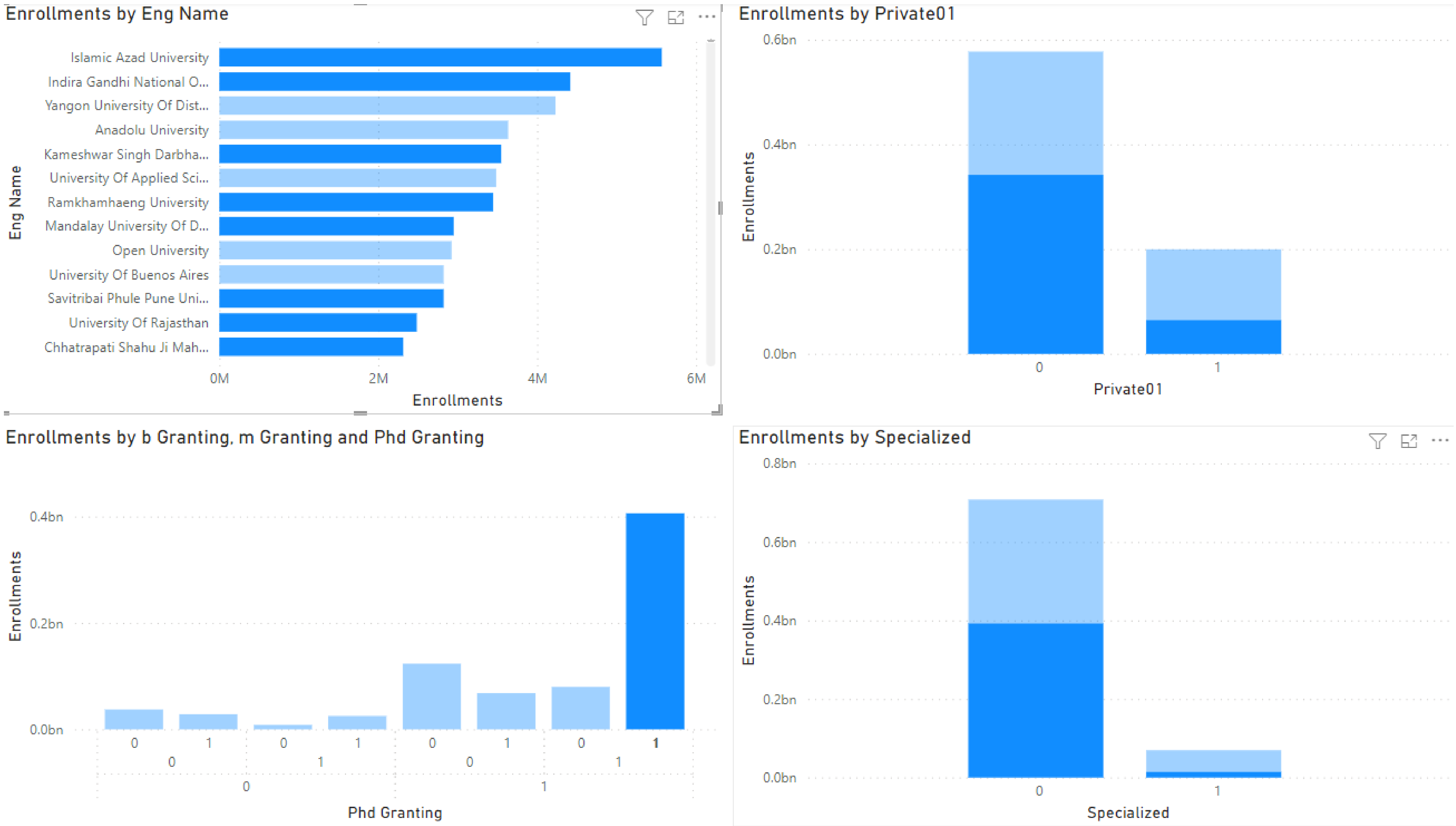
Visualizations (7)

On the dashboard below we can see the top ranking institutions on enrollments (bar chart) and the enrollments depending if the school is private or not, the degrees offered and if it's specialized



Visualizations (8)

Again, if we click on one of the attributes bars, we will see which of the top universities have those characteristics.



Data Mining tasks

We used Python Libraries and RapidMiner to transform and clean our data off outliers. Then we used data mining algorithms to get useful insights

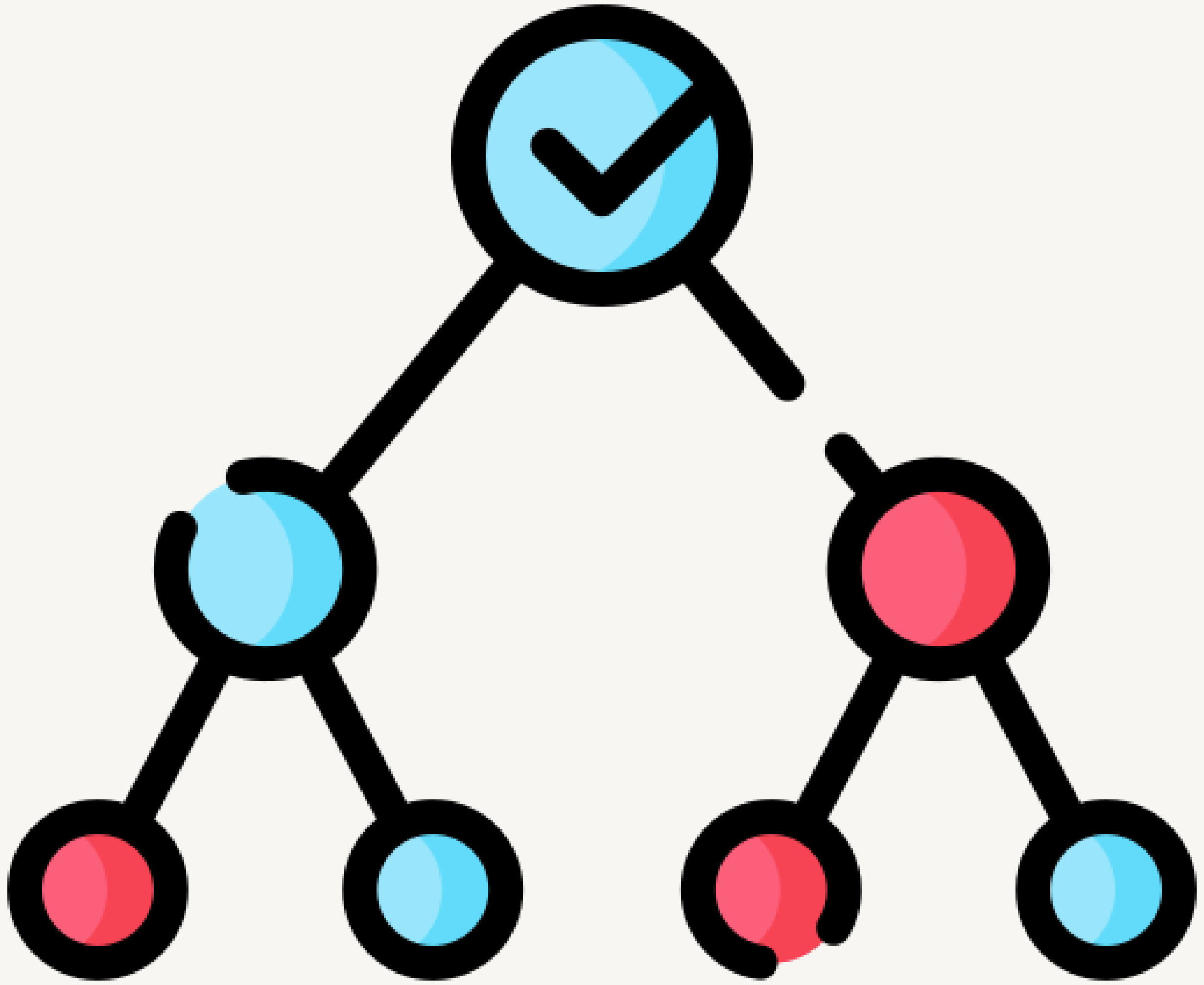
Pandas

Decision Tree

- We wanted to see how the degree grantings, the divisions and the fields of a university would influence the enrollments of each institution

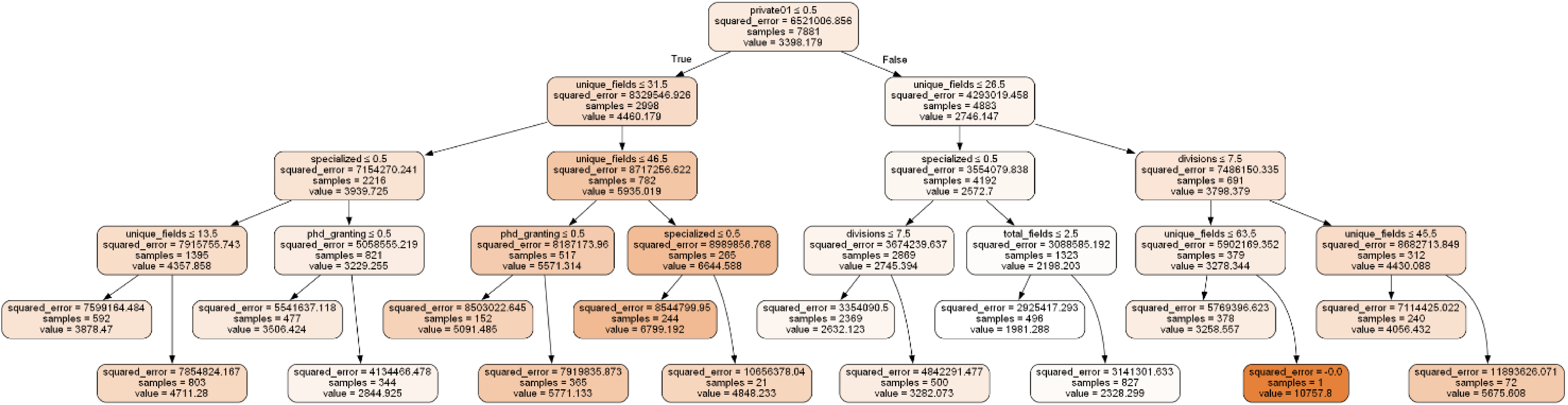
- We grouped all our data by the university id, kept the binary values of the attributes above and got the mean value of the enrollments through the years

- We cleaned our data off outliers and ran a grid search to find the best depth for our tree with python



Decision Tree

We created our tree with python with the max depth set to 4, after the grid search, and the result is the below:



private01 ≤ 0.5
squared_error = 6521006.856
samples = 7881
value = 3398.179

True

False

unique_fields ≤ 31.5
squared_error = 8329546.926
samples = 2998
value = 4460.179

unique_fields ≤ 26.5
squared_error = 4293019.458
samples = 4883
value = 2746.147

specialized ≤ 0.5
squared_error = 7154270.241
samples = 2216
value = 3939.725

unique_fields ≤ 46.5
squared_error = 8717256.622
samples = 782
value = 5935.019

specialized ≤ 0.5
squared_error = 3554079.838
samples = 4192
value = 2572.7

divisions ≤ 7.5
squared_error = 7486150.335
samples = 691
value = 3798.379

unique_fields ≤ 13.5
squared_error = 7915755.743
samples = 1395
value = 4357.858

phd_granting ≤ 0.5
squared_error = 5058555.219
samples = 821
value = 3229.255

phd_granting ≤ 0.5
squared_error = 8187173.96
samples = 517
value = 5571.314

specialized ≤ 0.5
squared_error = 8989856.768
samples = 265
value = 6644.588

divisions ≤ 7.5
squared_error = 3674239.637
samples = 2869
value = 2745.394

total_fields ≤ 2.5
squared_error = 3088585.192
samples = 1323
value = 2198.203

unique_fields ≤ 63.5
squared_error = 5902169.352
samples = 379
value = 3278.344

unique_fields ≤ 45.5
squared_error = 8682713.849
samples = 312
value = 4430.088

squared_error = 7599164.484
samples = 592
value = 3878.47

squared_error = 5541637.118
samples = 477
value = 3506.424

squared_error = 8503022.645
samples = 152
value = 5091.485

squared_error = 8544799.95
samples = 244
value = 6799.192

squared_error = 3354090.5
samples = 2369
value = 2632.123

squared_error = 2925417.293
samples = 496
value = 1981.288

squared_error = 5769396.623
samples = 378
value = 3258.557

squared_error = 7114425.022
samples = 240
value = 4056.432

squared_error = 7854824.167
samples = 803
value = 4711.28

squared_error = 4134466.478
samples = 344
value = 2844.925

squared_error = 7919835.873
samples = 365
value = 5771.133

squared_error = 10656378.04
samples = 21
value = 4848.233

squared_error = 4842291.477
samples = 500
value = 3282.073

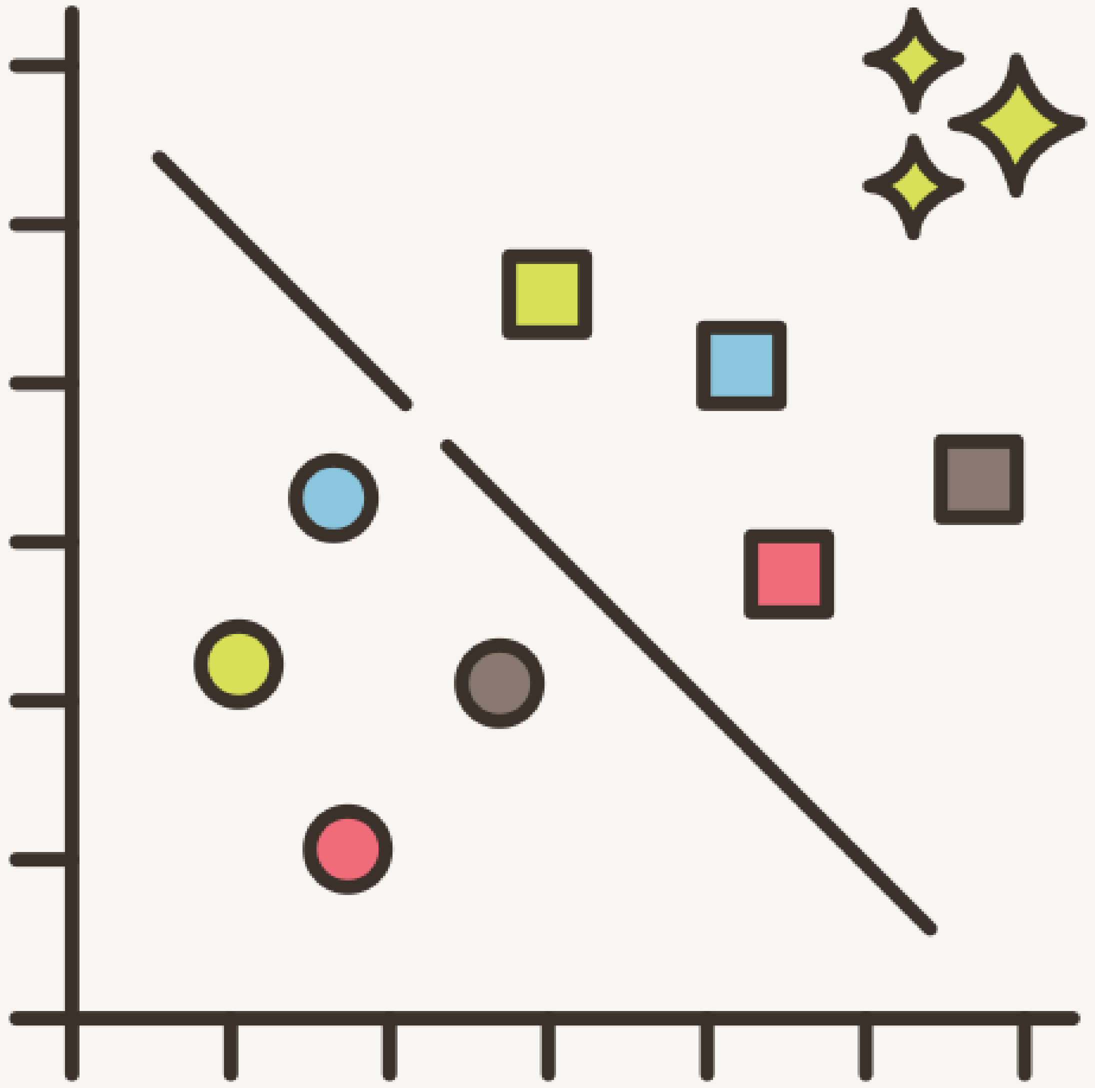
squared_error = 3141301.633
samples = 827
value = 2328.299

squared_error = -0.0
samples = 1
value = 10757.8

squared_error = 11893626.071
samples = 72
value = 5675.608

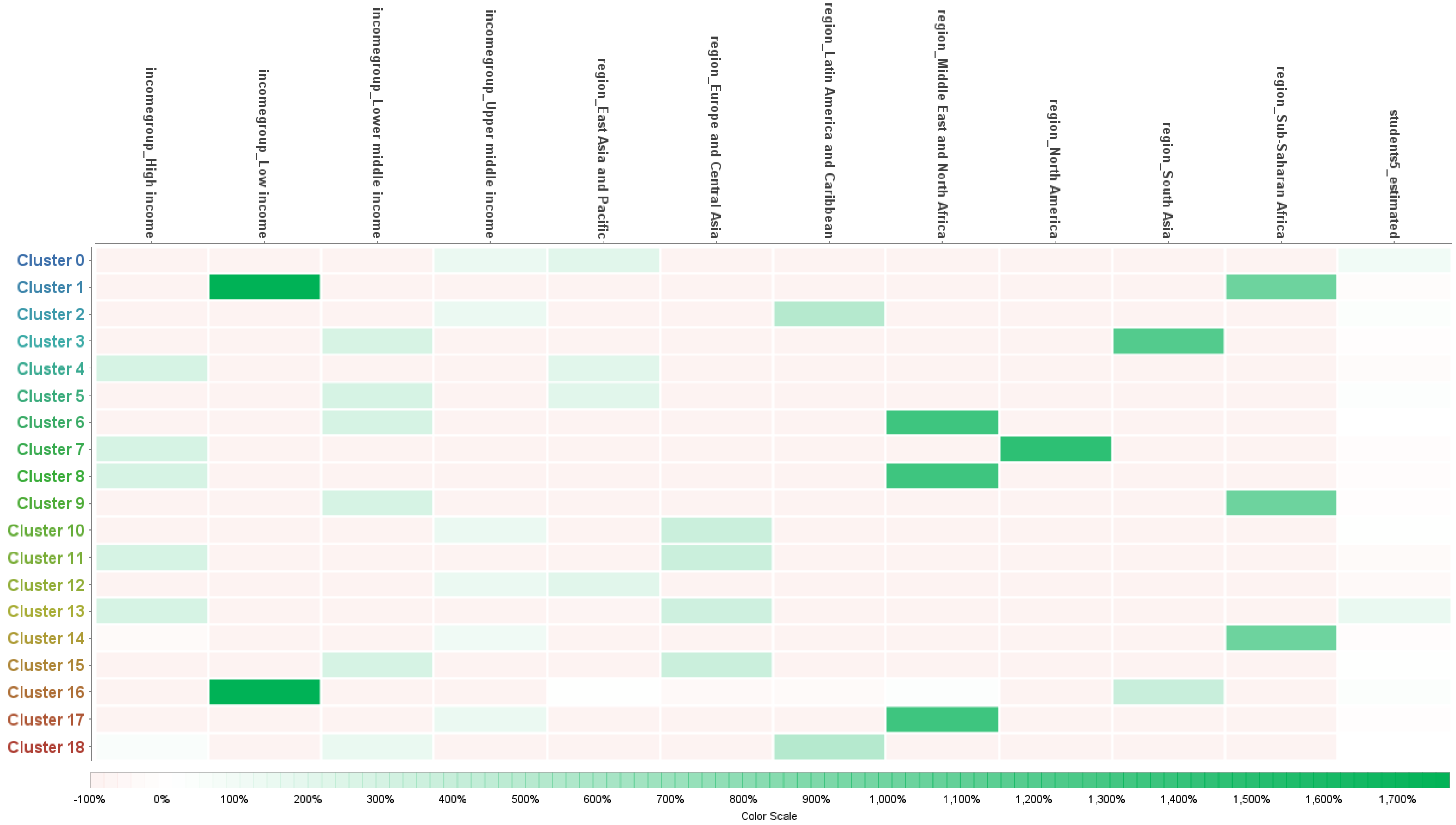
Clustering

- We wanted to see which clusters of region and income group would provide higher enrollment
- We grouped all our data by the university id and kept the text values of income group, region and enrollments. Then we created dummy variables for the regions and income groups having values 0 and 1
- We used standard scaling on our data so the enrollments high value wouldn't influence the results. The we used silhouettes to define the best number for the clusters. Finally, we used the rapidminer k-means algorithm.



Clusters

We used the k-means algorithm with k equal to 19, from our silhouette search, and got the results below:



Clustering Results

Clusters with higher mean enrollments (ordered):

1. High Income - Europe Central Asia
2. Upper/Middle Income - East Asia Pacific
3. Low income - Middle East/North Africa
4. Upper/Middle Income - Latin America
5. Lower/Middle Income - East Asia Pacific
6. Upper/Middle Income - Europe Central Asia