CITS5504

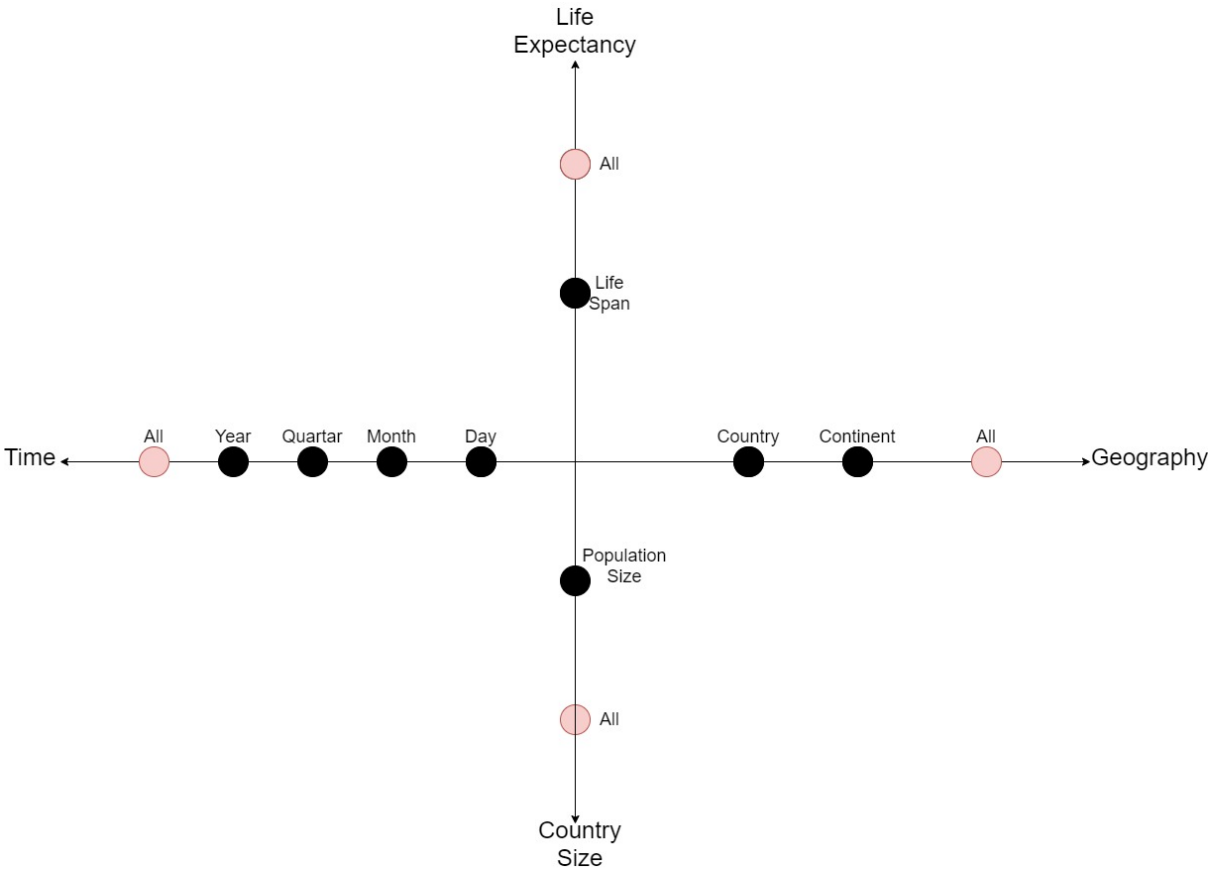
PROJECT 1

ISAAC HUANG

(23019722)

Table of Contents

1. STARNET



2. ETL

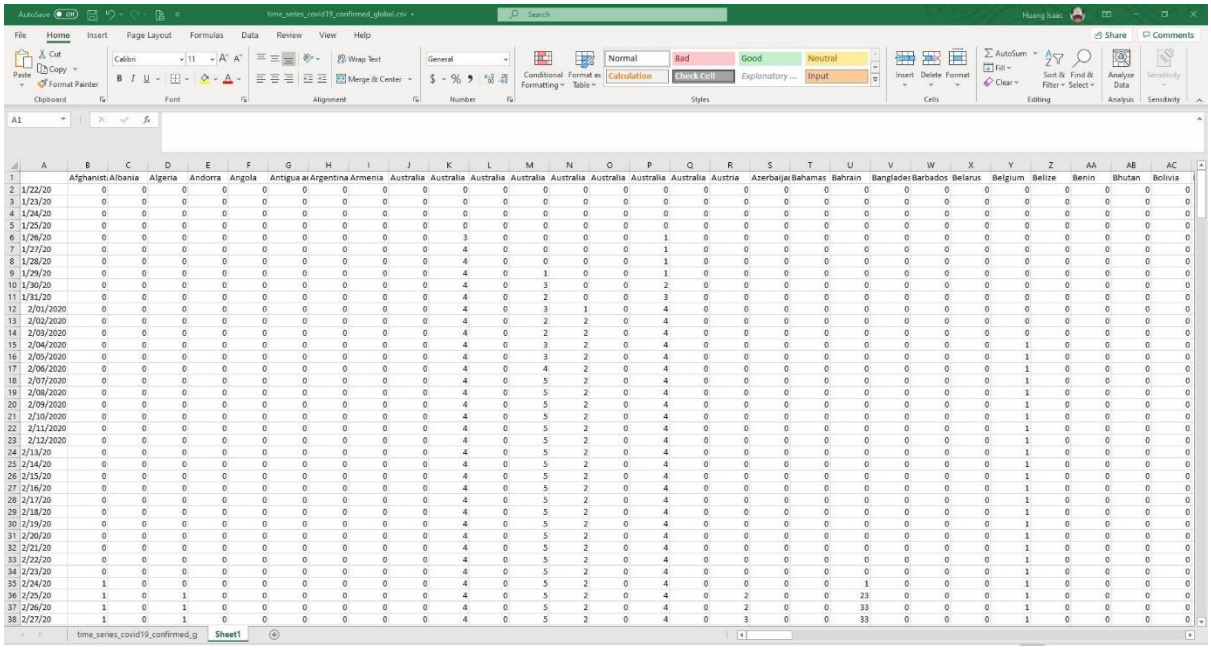
2.1 Extraction

Valid data cross all data sources are **from 22 Jan 2020 to 1 Mar 2021**, so here we only extract data between those dates. Latitude and longitude data are not relevant, so we discard them here.

2.2 Transformation

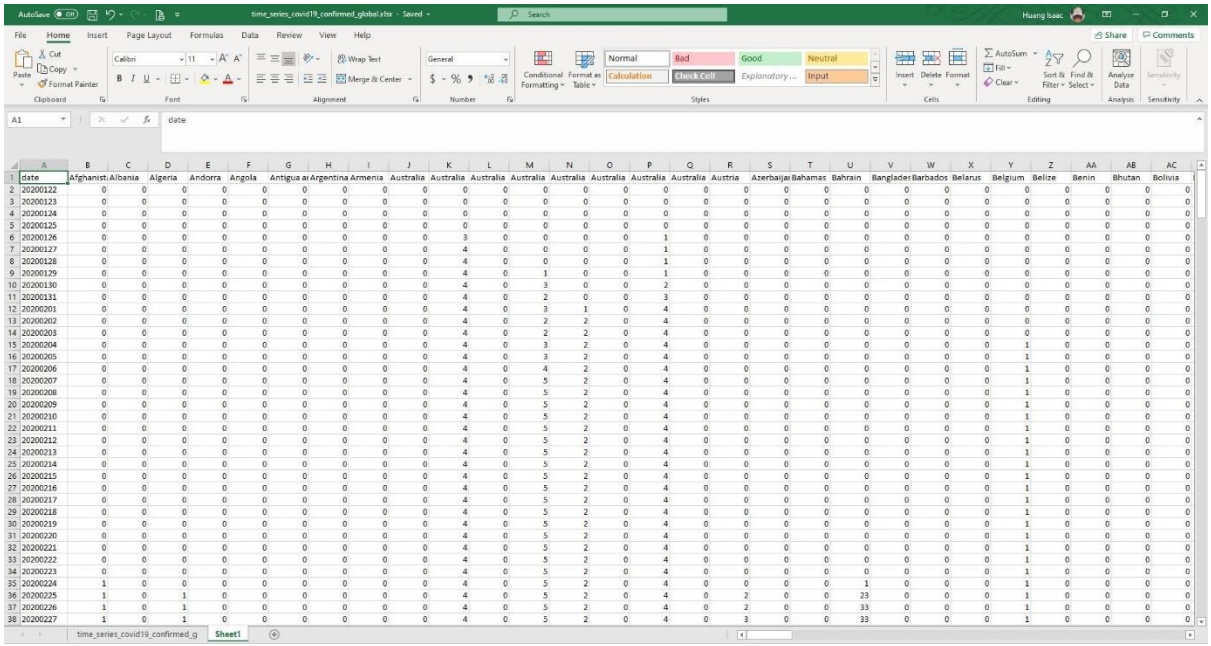
2.2.1 Transposing

Transpose rows and columns for later transformation:



2.2.2 Reconciling Data

We have two different date formats here, so we need to reconcile them into one format, I chose pure number format to display date here, in case some data file type we use in the future does not support special characters:



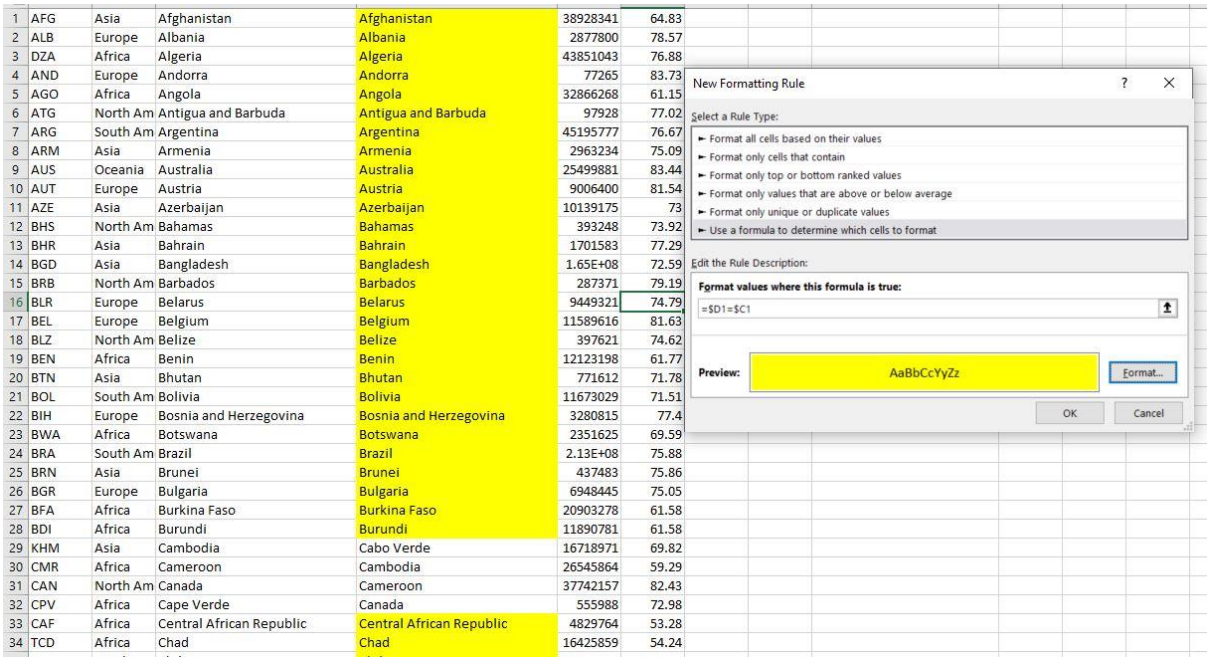
2.2.3 Cleaning and Integration

A. Mapping province / state data into the same country, as they are fragmented.

B. Handling NULL values: I fill up Kosovo's life expectancy data, as it is the only null cell, and we can easily find certified information online.

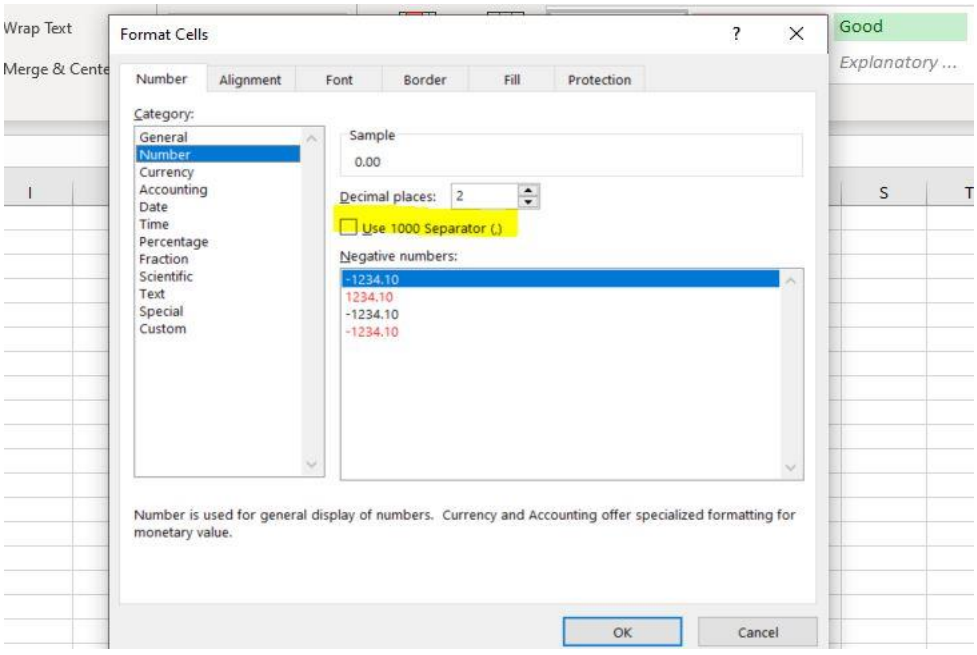
C. Removing "Dimond Princess", "MS Zaandam" and "Micronesia", as they are not a country.

D. As country names appear differently across dataset, we use Excel to search and reconcile different names:



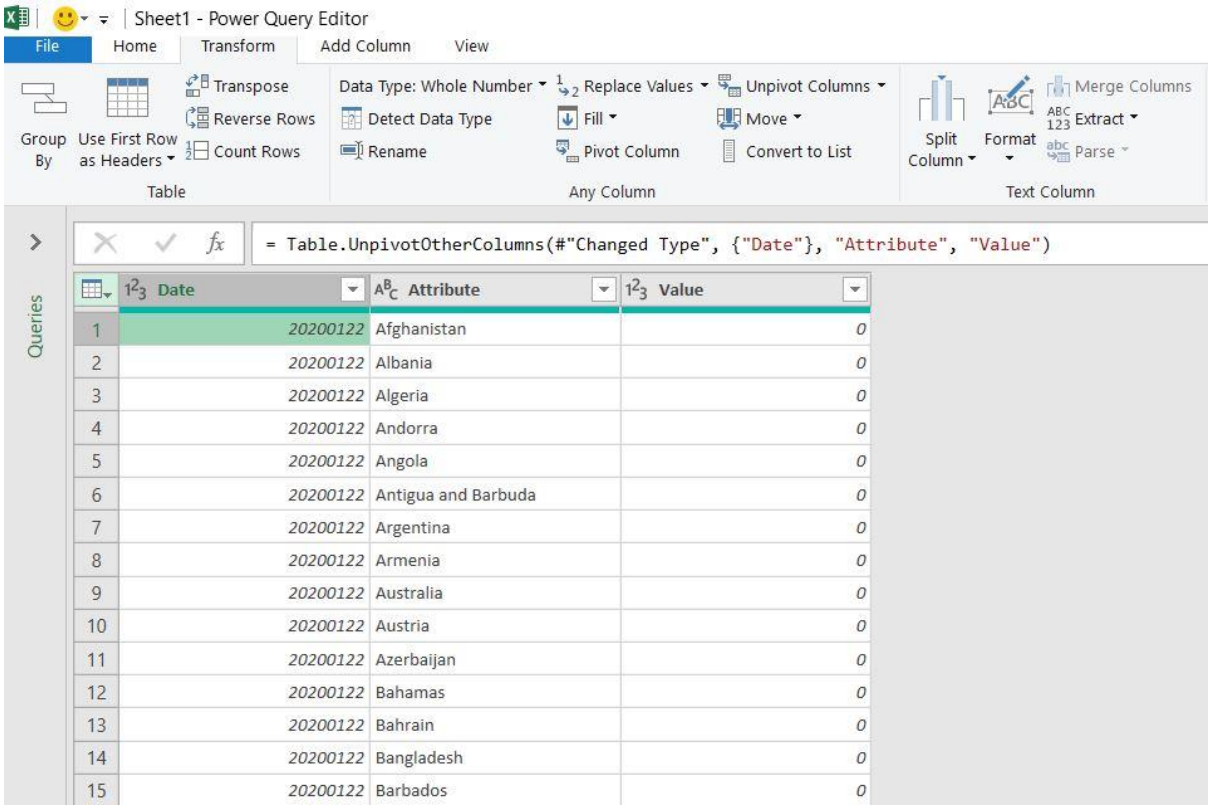
1	AFG	Asia	Afghanistan	Afghanistan	38928341	64.83
2	ALB	Europe	Albania	Albania	2877800	78.57
3	DZA	Africa	Algeria	Algeria	43851043	76.88
4	AND	Europe	Andorra	Andorra	77265	83.73
5	AGO	Africa	Angola	Angola	32866268	61.15
6	ATG	North Am	Antigua and Barbuda	Antigua and Barbuda	97928	77.02
7	ARG	South Am	Argentina	Argentina	45195777	76.67
8	ARM	Asia	Armenia	Armenia	2963234	75.09
9	AUS	Oceania	Australia	Australia	25499881	83.44
10	AUT	Europe	Austria	Austria	9006400	81.54
11	AZE	Asia	Azerbaijan	Azerbaijan	10139175	73
12	BHS	North Am	Bahamas	Bahamas	393248	73.92
13	BHR	Asia	Bahrain	Bahrain	1701583	77.29
14	BGD	Asia	Bangladesh	Bangladesh	1.65E+08	72.59
15	BRB	North Am	Barbados	Barbados	287371	79.19
16	BLR	Europe	Belarus	Belarus	9449321	74.79
17	BEL	Europe	Belgium	Belgium	11589616	81.63
18	BLZ	North Am	Belize	Belize	397621	74.62
19	BEN	Africa	Benin	Benin	12123198	61.77
20	BTN	Asia	Bhutan	Bhutan	771612	71.78
21	BOL	South Am	Bolivia	Bolivia	11673029	71.51
22	BIH	Europe	Bosnia and Herzegovina	Bosnia and Herzegovina	3280815	77.4
23	BWA	Africa	Botswana	Botswana	2351625	69.59
24	BRA	South Am	Brazil	Brazil	2.13E+08	75.88
25	BRN	Asia	Brunei	Brunei	437483	75.86
26	BGR	Europe	Bulgaria	Bulgaria	6948445	75.05
27	BFA	Africa	Burkina Faso	Burkina Faso	20903278	61.58
28	BDI	Africa	Burundi	Burundi	11890781	61.58
29	KHM	Asia	Cambodia	Cabo Verde	16718971	69.82
30	CMR	Africa	Cameroon	Cambodia	26545864	59.29
31	CAN	North Am	Canada	Cameroon	37742157	82.43
32	CPV	Africa	Cape Verde	Canada	555988	72.98
33	CAF	Africa	Central African Republic	Central African Republic	4829764	53.28
34	TCD	Africa	Chad	Chad	16425859	54.24

E. Removing special characters like "Korea, South" or number "1,000" separators, because we use "," as our Field Terminator in SQL script:

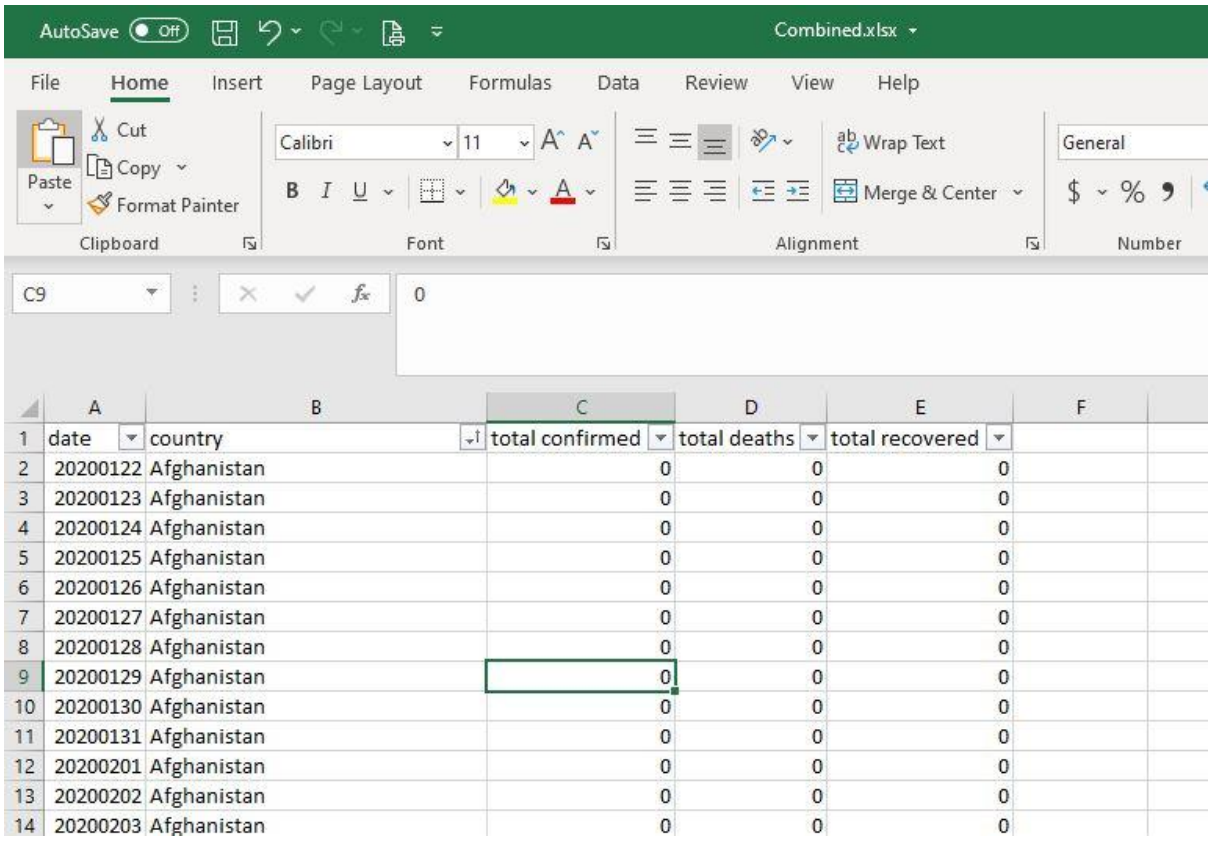


2.2.4 Joining

Use Excel Power Query to unpivot columns and join multiple attributes into one. Apply this method three times for confirmed, recovered cases and deaths, then join into one file.



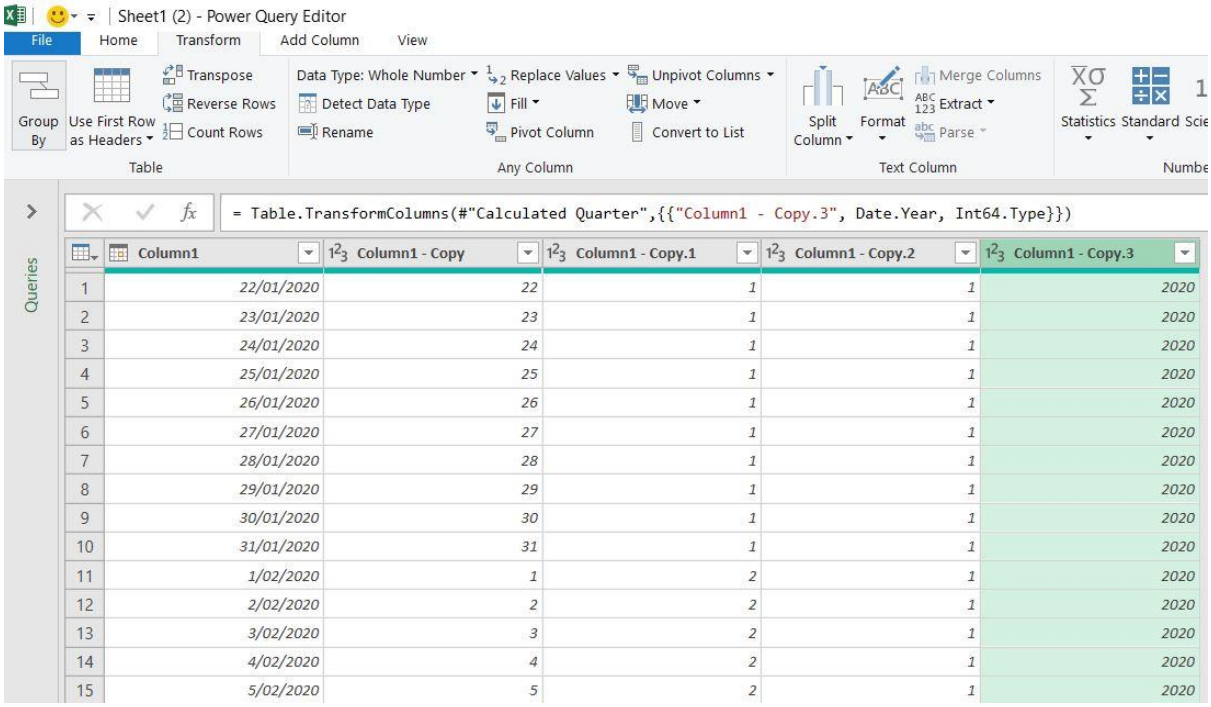
	Date	Attribute	Value
1	20200122	Afghanistan	0
2	20200122	Albania	0
3	20200122	Algeria	0
4	20200122	Andorra	0
5	20200122	Angola	0
6	20200122	Antigua and Barbuda	0
7	20200122	Argentina	0
8	20200122	Armenia	0
9	20200122	Australia	0
10	20200122	Austria	0
11	20200122	Azerbaijan	0
12	20200122	Bahamas	0
13	20200122	Bahrain	0
14	20200122	Bangladesh	0
15	20200122	Barbados	0



	A	B	C	D	E	F
1	date	country	total confirmed	total deaths	total recovered	
2	20200122	Afghanistan	0	0	0	
3	20200123	Afghanistan	0	0	0	
4	20200124	Afghanistan	0	0	0	
5	20200125	Afghanistan	0	0	0	
6	20200126	Afghanistan	0	0	0	
7	20200127	Afghanistan	0	0	0	
8	20200128	Afghanistan	0	0	0	
9	20200129	Afghanistan	0	0	0	
10	20200130	Afghanistan	0	0	0	
11	20200131	Afghanistan	0	0	0	
12	20200201	Afghanistan	0	0	0	
13	20200202	Afghanistan	0	0	0	
14	20200203	Afghanistan	0	0	0	

2.2.5 Splitting

Very convenient to use Excel Power Query to split “day”, “month”, “quarter” and “year” out of dates for time dimension csv file as well:



	Column1	Column1 - Copy	Column1 - Copy.1	Column1 - Copy.2	Column1 - Copy.3
1	22/01/2020	22	1	1	2020
2	23/01/2020	23	1	1	2020
3	24/01/2020	24	1	1	2020
4	25/01/2020	25	1	1	2020
5	26/01/2020	26	1	1	2020
6	27/01/2020	27	1	1	2020
7	28/01/2020	28	1	1	2020
8	29/01/2020	29	1	1	2020
9	30/01/2020	30	1	1	2020
10	31/01/2020	31	1	1	2020
11	1/02/2020	1	2	1	2020
12	2/02/2020	2	2	1	2020
13	3/02/2020	3	2	1	2020
14	4/02/2020	4	2	1	2020
15	5/02/2020	5	2	1	2020

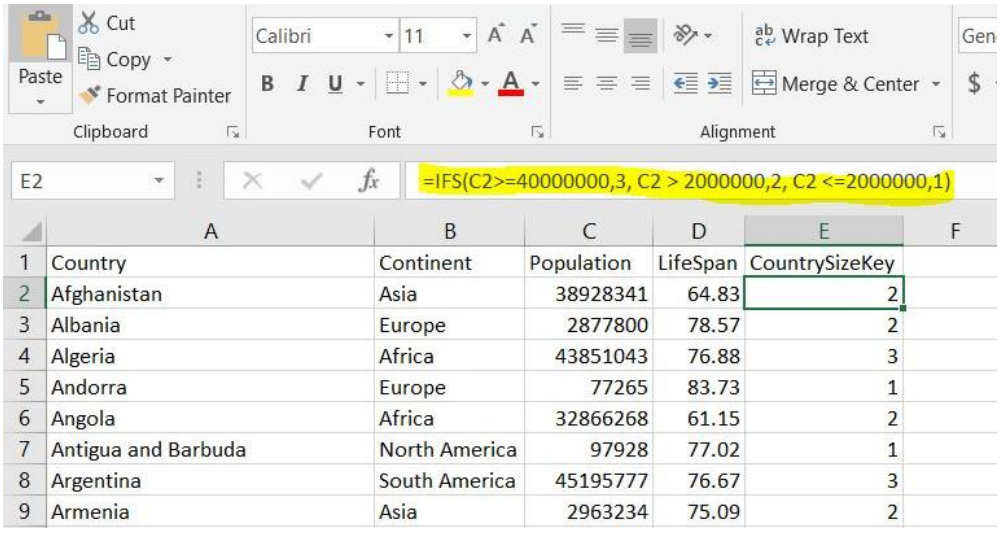
2.2.6 Sorting

A. Use some easy Excel functions to get daily case numbers from total case numbers.

B. Use Excel Sort & Filter to make sure the attribute orders are the same in both fact and dimension tables.

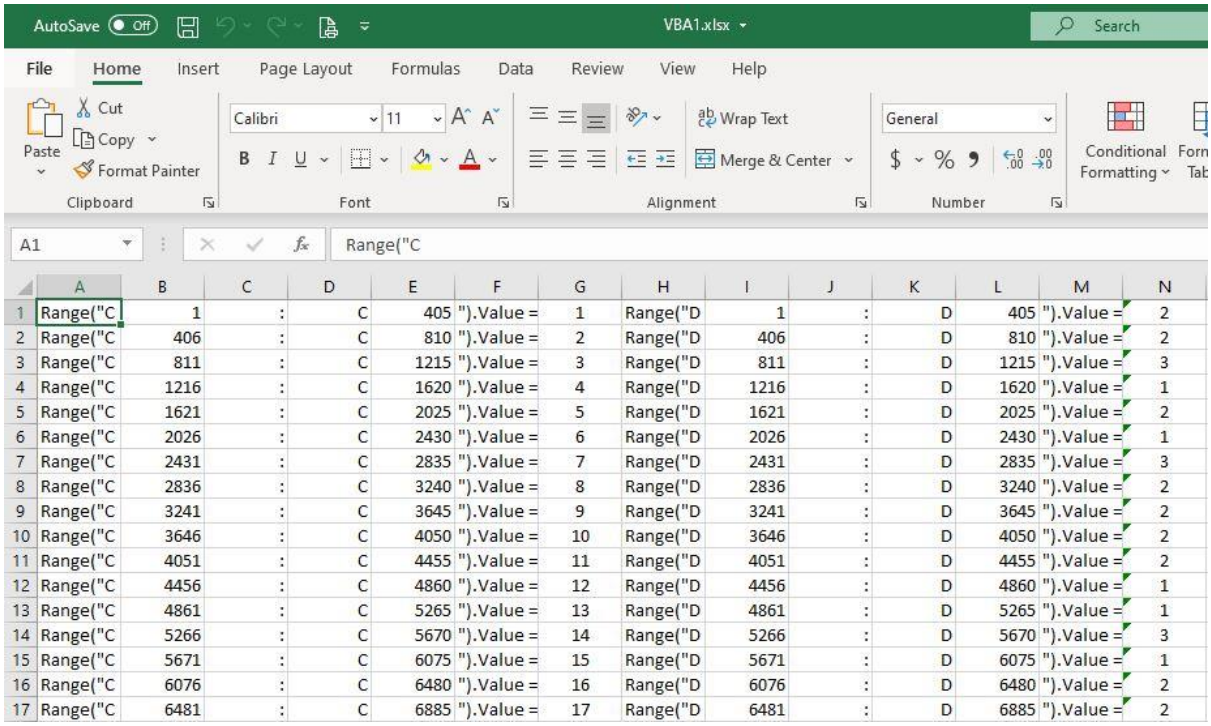
2.2.7 Conversion

A. Use Excel “IFS” function to convert Population and Life Span into CountrySizeKey and LifeExpectancyKey:

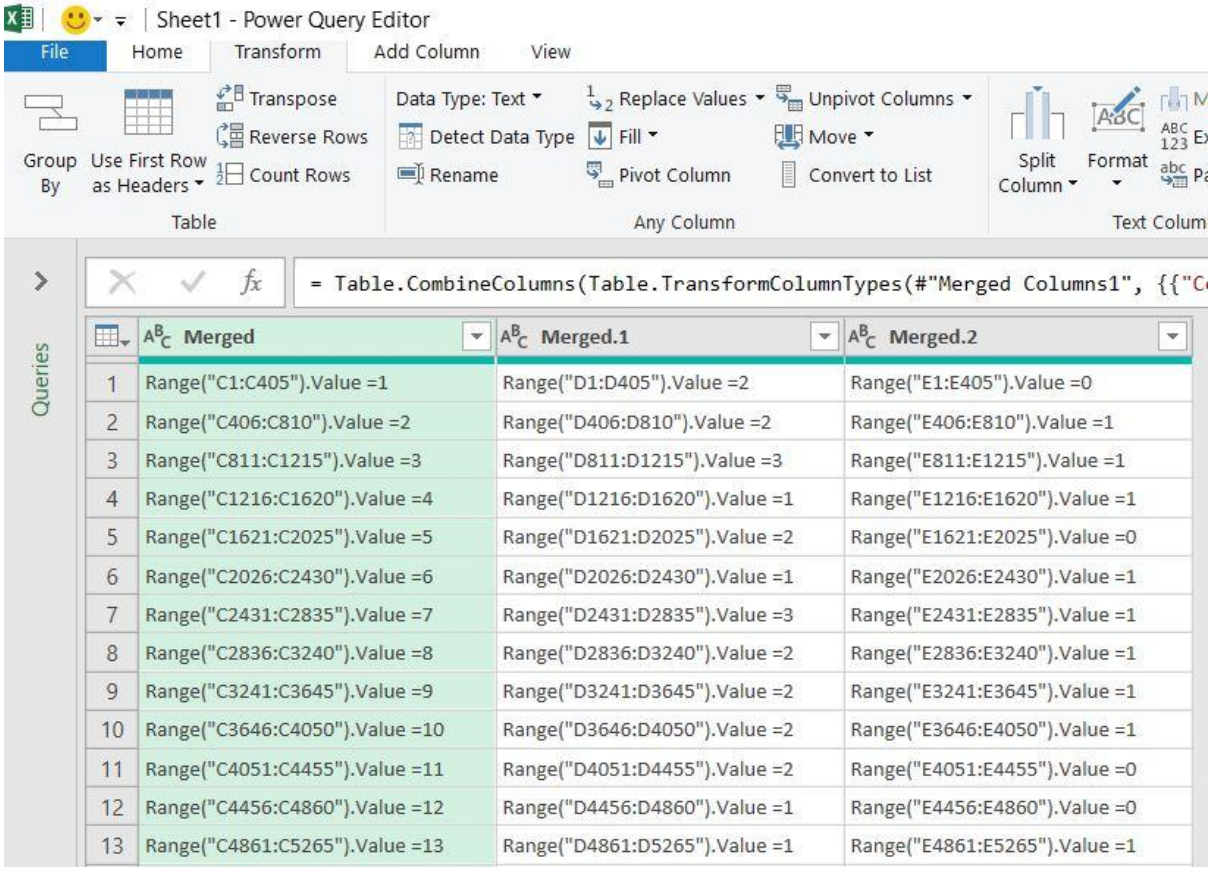


	A	B	C	D	E	F
1	Country	Continent	Population	LifeSpan	CountrySizeKey	
2	Afghanistan	Asia	38928341	64.83	2	
3	Albania	Europe	2877800	78.57	2	
4	Algeria	Africa	43851043	76.88	3	
5	Andorra	Europe	77265	83.73	1	
6	Angola	Africa	32866268	61.15	2	
7	Antigua and Barbuda	North America	97928	77.02	1	
8	Argentina	South America	45195777	76.67	3	
9	Armenia	Asia	2963234	75.09	2	

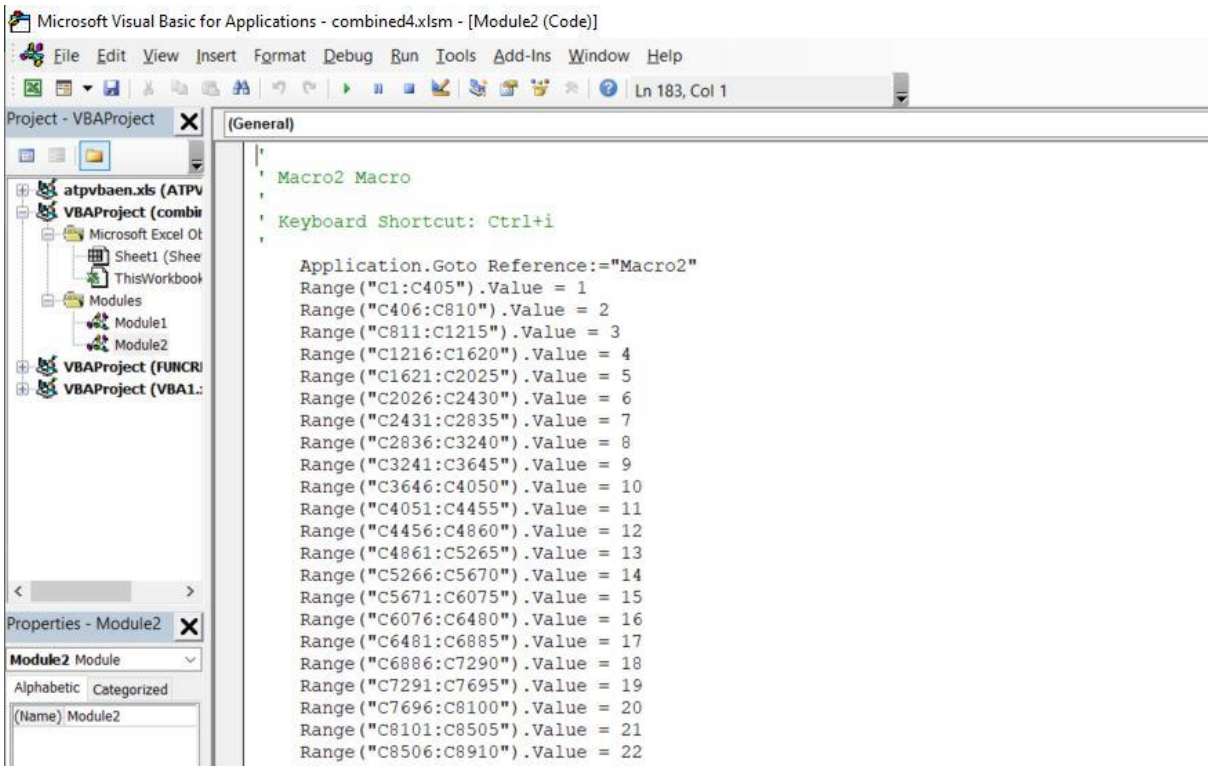
B. Use Excel and Power Query to generate functions, then paste on Excel VBA to convert and insert Country.key, CountrySize.key and LifeExpectancy.key into fact table spreadsheet.

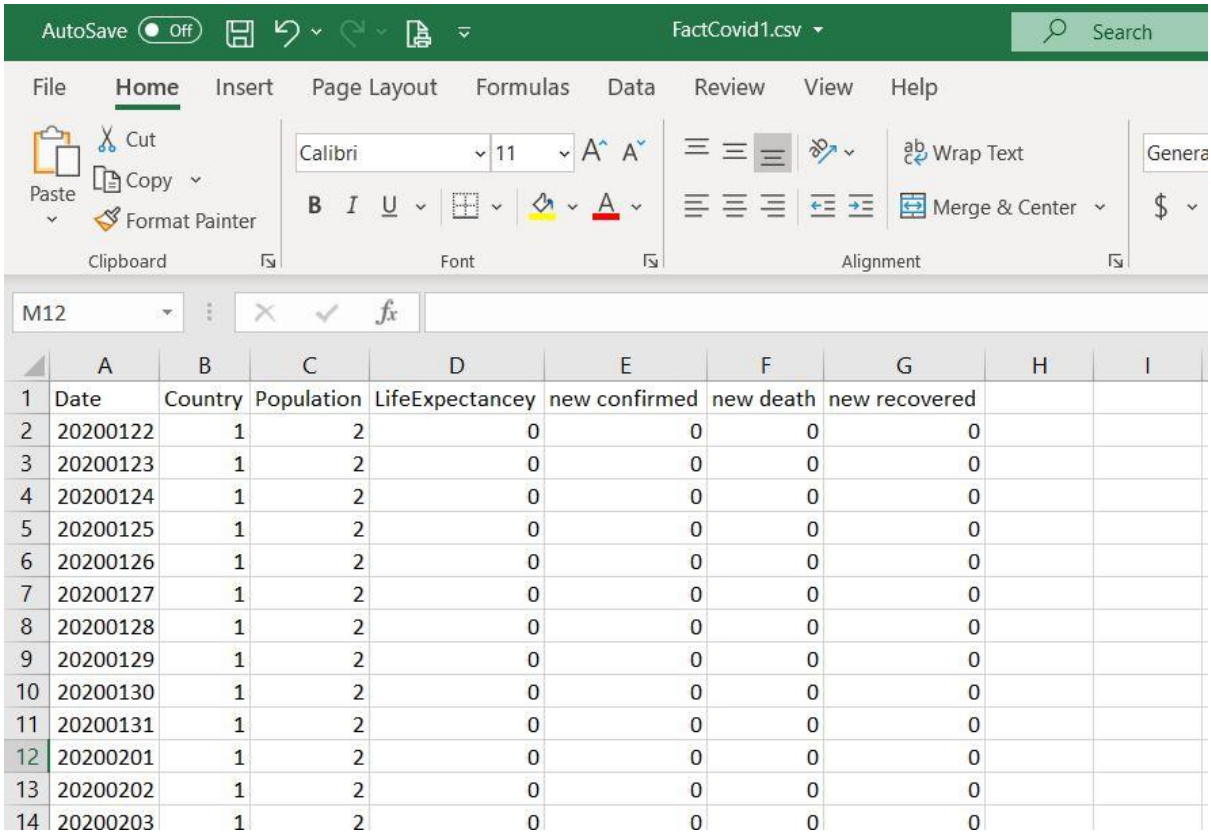


	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Range("C	1	:	C	405	").Value =	1	Range("D	1	:	D	405	").Value =	2
2	Range("C	406	:	C	810	").Value =	2	Range("D	406	:	D	810	").Value =	2
3	Range("C	811	:	C	1215	").Value =	3	Range("D	811	:	D	1215	").Value =	3
4	Range("C	1216	:	C	1620	").Value =	4	Range("D	1216	:	D	1620	").Value =	1
5	Range("C	1621	:	C	2025	").Value =	5	Range("D	1621	:	D	2025	").Value =	2
6	Range("C	2026	:	C	2430	").Value =	6	Range("D	2026	:	D	2430	").Value =	1
7	Range("C	2431	:	C	2835	").Value =	7	Range("D	2431	:	D	2835	").Value =	3
8	Range("C	2836	:	C	3240	").Value =	8	Range("D	2836	:	D	3240	").Value =	2
9	Range("C	3241	:	C	3645	").Value =	9	Range("D	3241	:	D	3645	").Value =	2
10	Range("C	3646	:	C	4050	").Value =	10	Range("D	3646	:	D	4050	").Value =	2
11	Range("C	4051	:	C	4455	").Value =	11	Range("D	4051	:	D	4455	").Value =	2
12	Range("C	4456	:	C	4860	").Value =	12	Range("D	4456	:	D	4860	").Value =	1
13	Range("C	4861	:	C	5265	").Value =	13	Range("D	4861	:	D	5265	").Value =	1
14	Range("C	5266	:	C	5670	").Value =	14	Range("D	5266	:	D	5670	").Value =	3
15	Range("C	5671	:	C	6075	").Value =	15	Range("D	5671	:	D	6075	").Value =	1
16	Range("C	6076	:	C	6480	").Value =	16	Range("D	6076	:	D	6480	").Value =	2
17	Range("C	6481	:	C	6885	").Value =	17	Range("D	6481	:	D	6885	").Value =	2



	Merged	Merged.1	Merged.2
1	Range("C1:C405").Value =1	Range("D1:D405").Value =2	Range("E1:E405").Value =0
2	Range("C406:C810").Value =2	Range("D406:D810").Value =2	Range("E406:E810").Value =1
3	Range("C811:C1215").Value =3	Range("D811:D1215").Value =3	Range("E811:E1215").Value =1
4	Range("C1216:C1620").Value =4	Range("D1216:D1620").Value =1	Range("E1216:E1620").Value =1
5	Range("C1621:C2025").Value =5	Range("D1621:D2025").Value =2	Range("E1621:E2025").Value =0
6	Range("C2026:C2430").Value =6	Range("D2026:D2430").Value =1	Range("E2026:E2430").Value =1
7	Range("C2431:C2835").Value =7	Range("D2431:D2835").Value =3	Range("E2431:E2835").Value =1
8	Range("C2836:C3240").Value =8	Range("D2836:D3240").Value =2	Range("E2836:E3240").Value =1
9	Range("C3241:C3645").Value =9	Range("D3241:D3645").Value =2	Range("E3241:E3645").Value =1
10	Range("C3646:C4050").Value =10	Range("D3646:D4050").Value =2	Range("E3646:E4050").Value =1
11	Range("C4051:C4455").Value =11	Range("D4051:D4455").Value =2	Range("E4051:E4455").Value =0
12	Range("C4456:C4860").Value =12	Range("D4456:D4860").Value =1	Range("E4456:E4860").Value =0
13	Range("C4861:C5265").Value =13	Range("D4861:D5265").Value =1	Range("E4861:E5265").Value =1





	A	B	C	D	E	F	G	H	I
1	Date	Country	Population	LifeExpectancey	new confirmed	new death	new recovered		
2	20200122	1	2	0	0	0	0		
3	20200123	1	2	0	0	0	0		
4	20200124	1	2	0	0	0	0		
5	20200125	1	2	0	0	0	0		
6	20200126	1	2	0	0	0	0		
7	20200127	1	2	0	0	0	0		
8	20200128	1	2	0	0	0	0		
9	20200129	1	2	0	0	0	0		
10	20200130	1	2	0	0	0	0		
11	20200131	1	2	0	0	0	0		
12	20200201	1	2	0	0	0	0		
13	20200202	1	2	0	0	0	0		
14	20200203	1	2	0	0	0	0		

2.3 Loading

2.3.1 Executing SQL Script

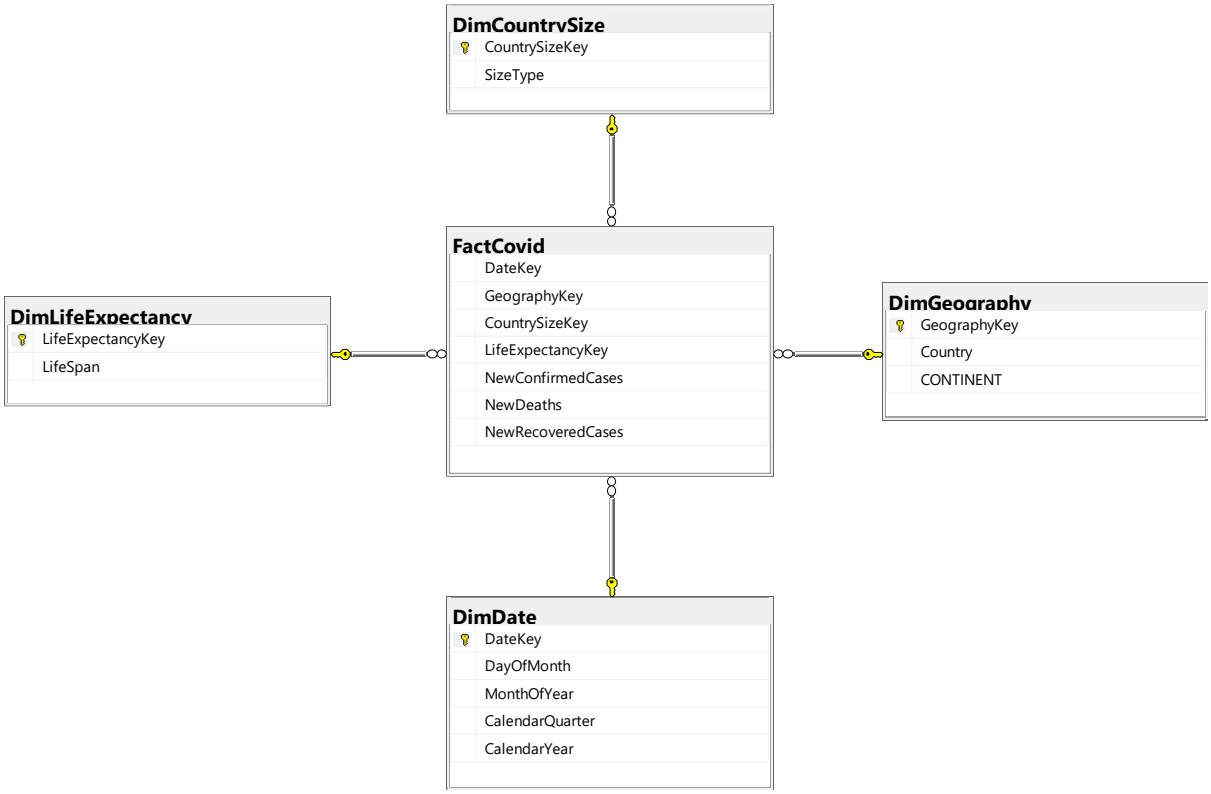
Run and debug SQL script until there is no bugs anymore.

A. Bug1: “,” in “Korea, South” and number “1,000” separators have been removed at step 2.2.3.E

B. Bug2: DATAFILETYPE=’widechar’ does not support UTF8. Hence, change it to ‘char’.

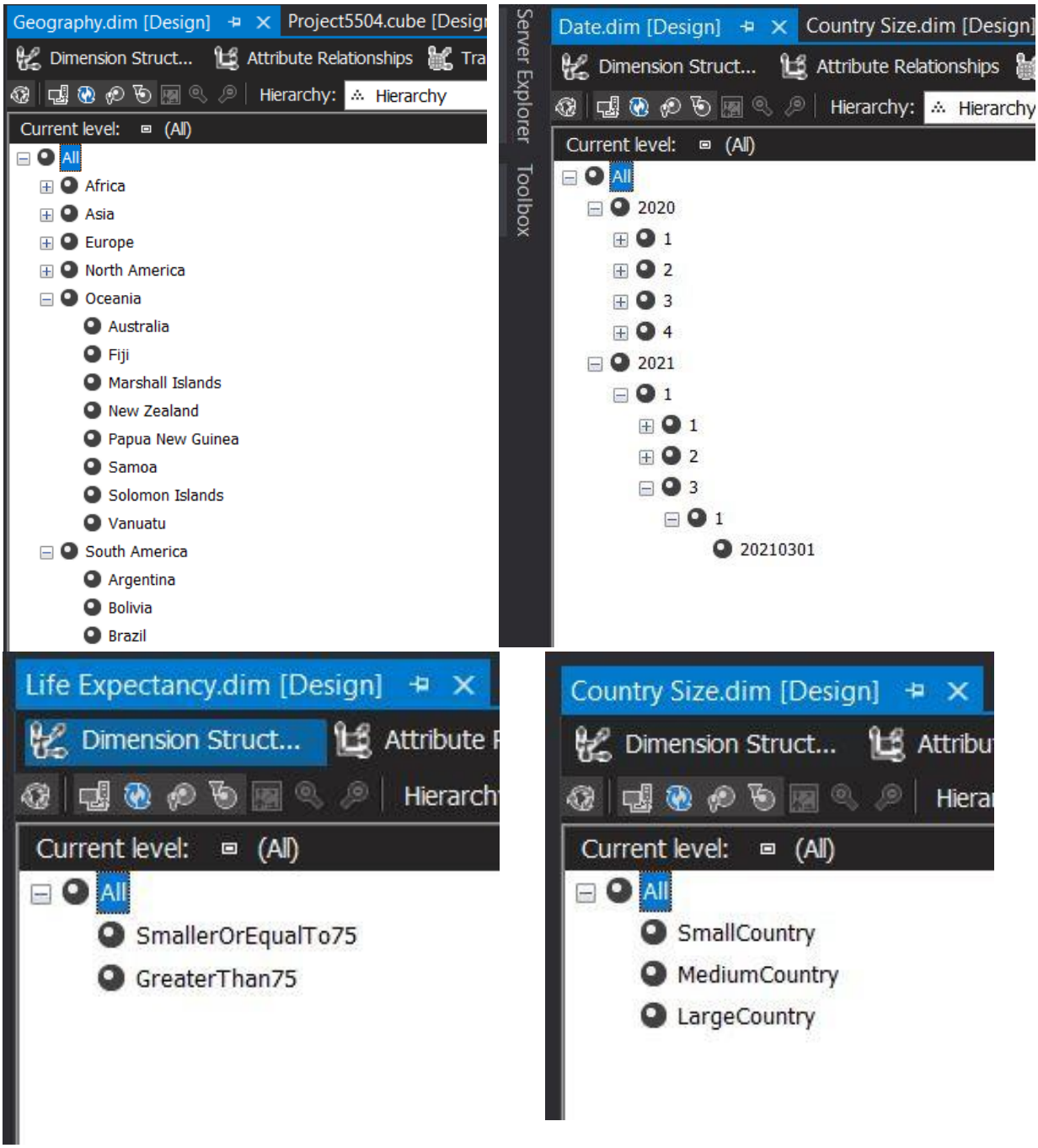
2.3.2 Star Schema / Database Diagram

Check if fact table, dimension tables and the relationships are all correctly implemented.



2.3.3 Hierarchy

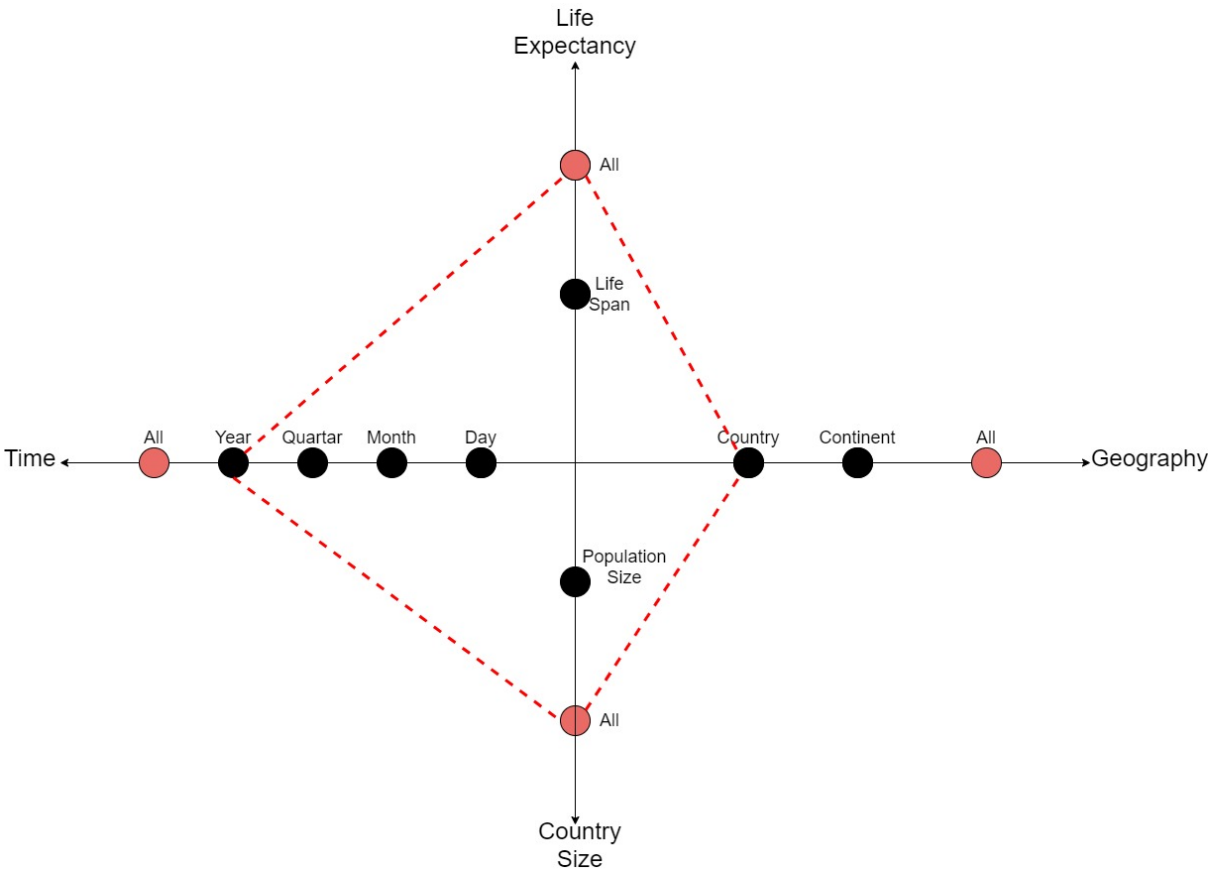
Check if the hierarchy of dimension tables can be correctly displayed.



3. FOOTPRINT / DATA CUBE / POWER BI

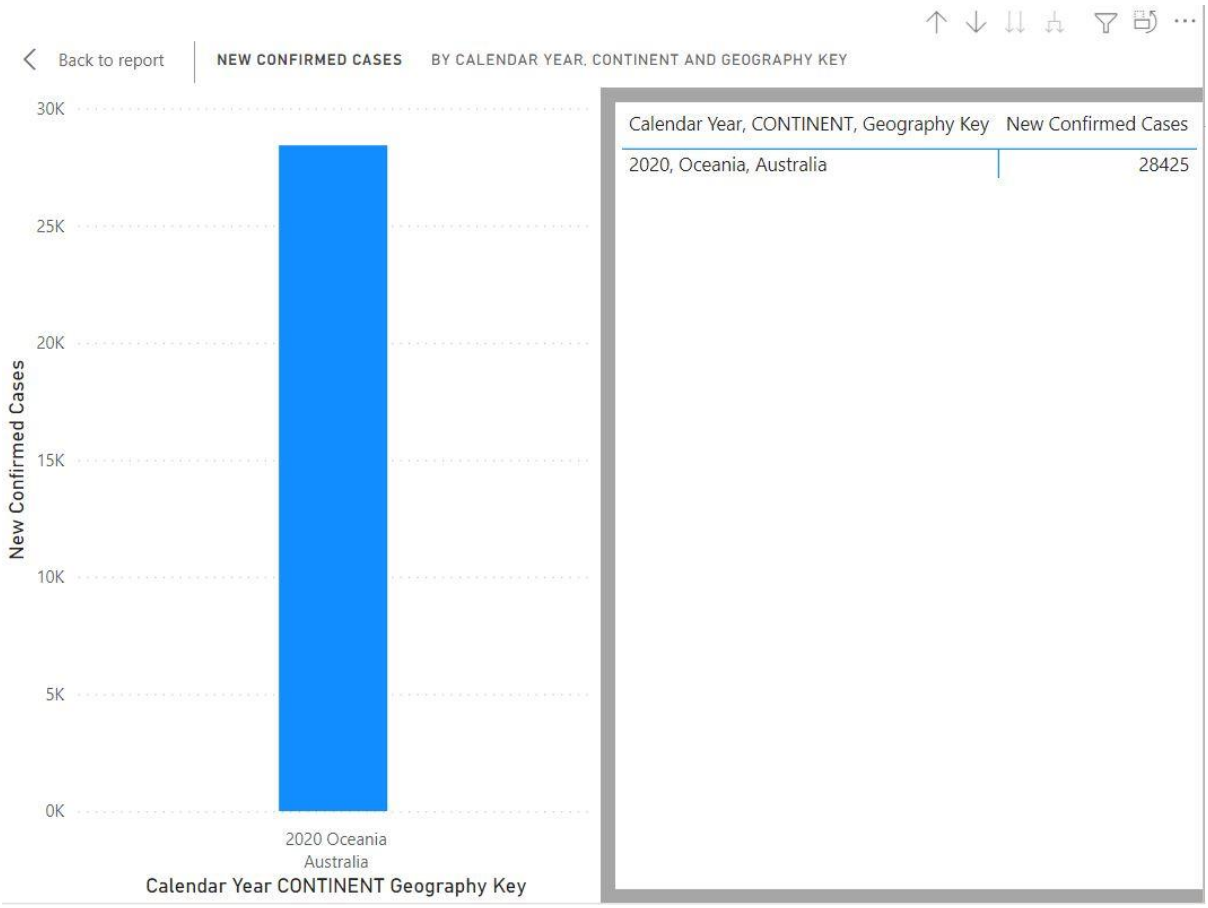
3.1 Business Queries

3.1.1 What is the total number of confirmed cases in Australia in 2020?

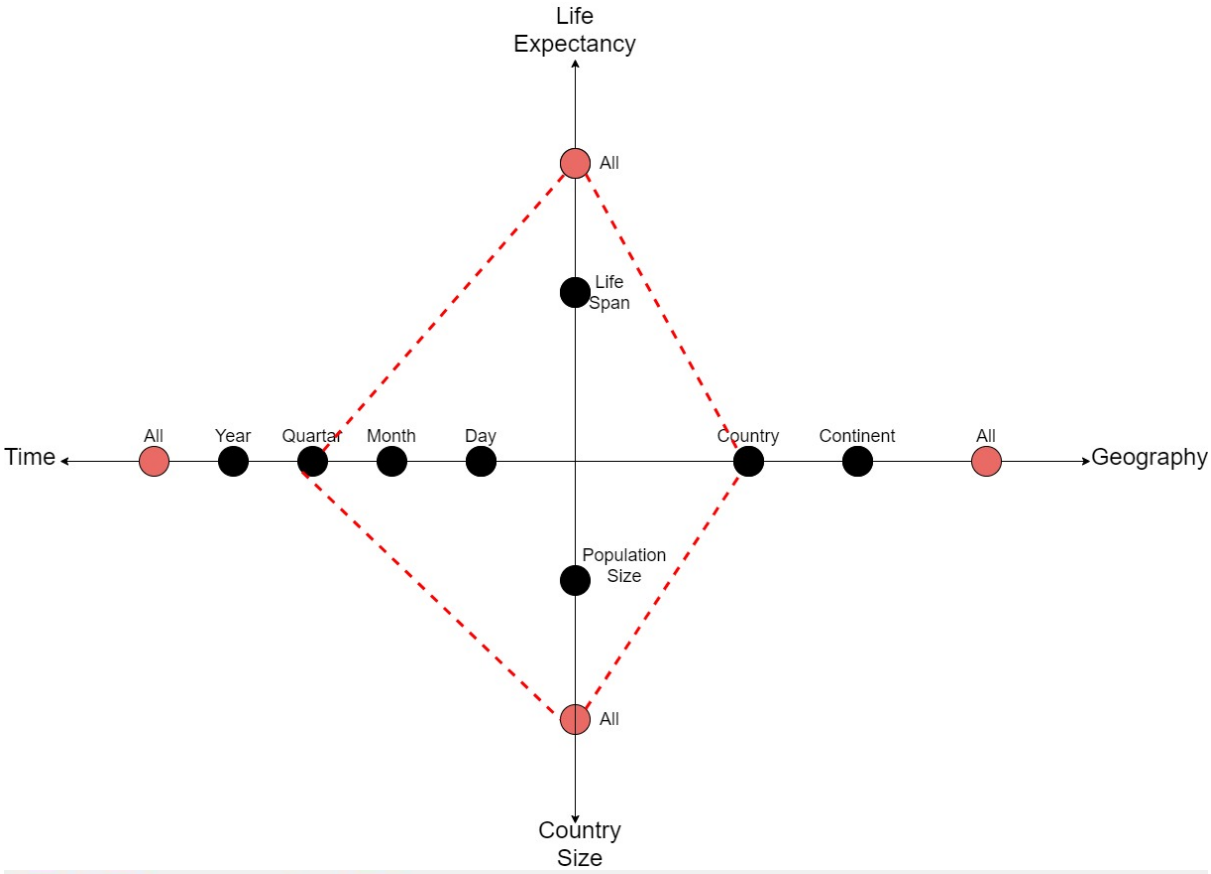


Dimension	Hierarchy	Operator	Filter Expression
Date	Calendar Year	Contains	2020
Geography	Country	Contains	Australia
<Select dimension>			

New Confirmed Cases
28425

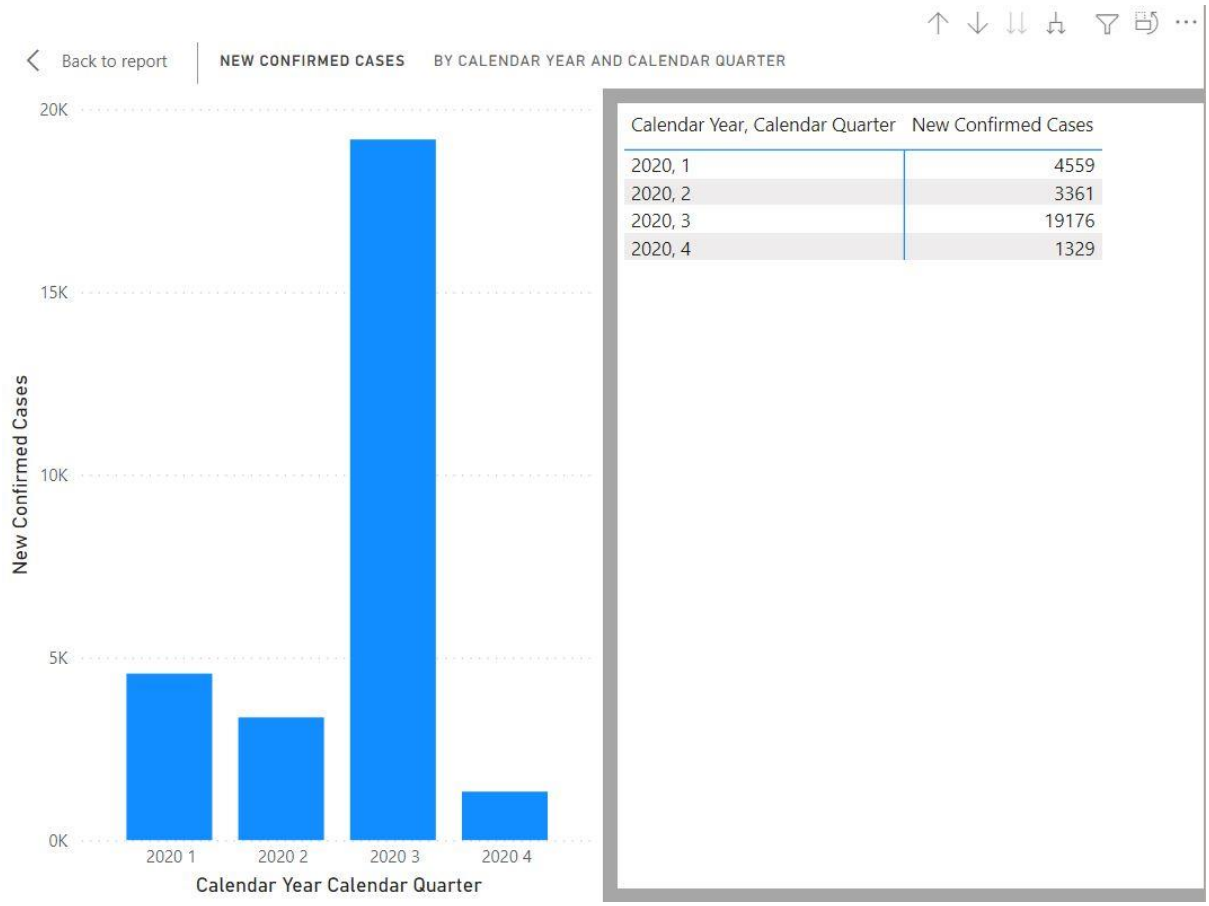

Back to report
NEW CONFIRMED CASES
BY CALENDAR YEAR, CONTINENT AND GEOGRAPHY KEY
30K
25K
20K
15K
10K
5K
0K
New Confirmed Cases
2020 Oceania
Australia
Calendar Year CONTINENT Geography Key
Calendar Year, CONTINENT, Geography Key
New Confirmed Cases
2020, Oceania, Australia
28425

3.1.2 What is the number of confirmed cases in each quarter of 2020 in Australia?

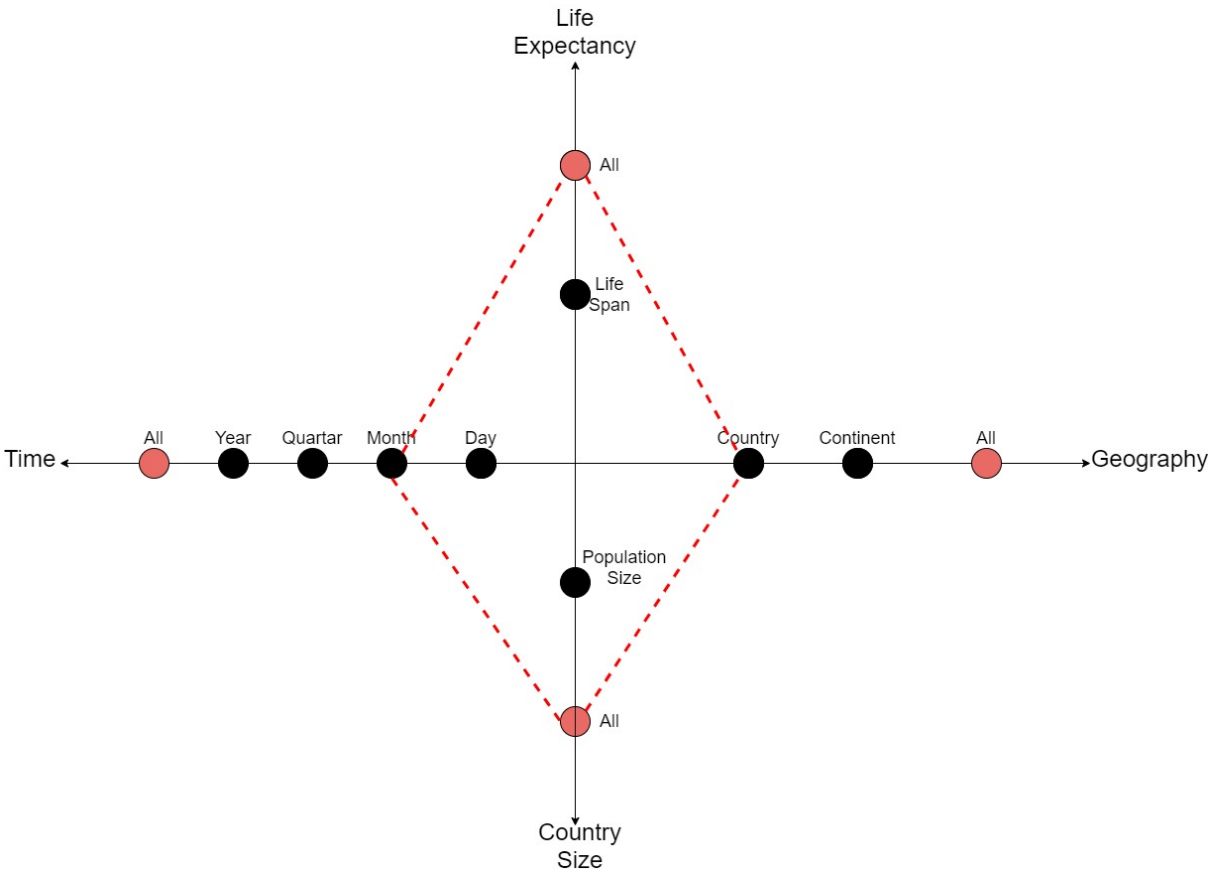


Dimension	Hierarchy	Operator	Filter Expression
Date	Calendar Year	Contains	2020
Geography	Country	Contains	Australia
<Select dimension>			

Calendar Quarter	New Confirmed Cases
1	4559
2	3361
3	19176
4	1329

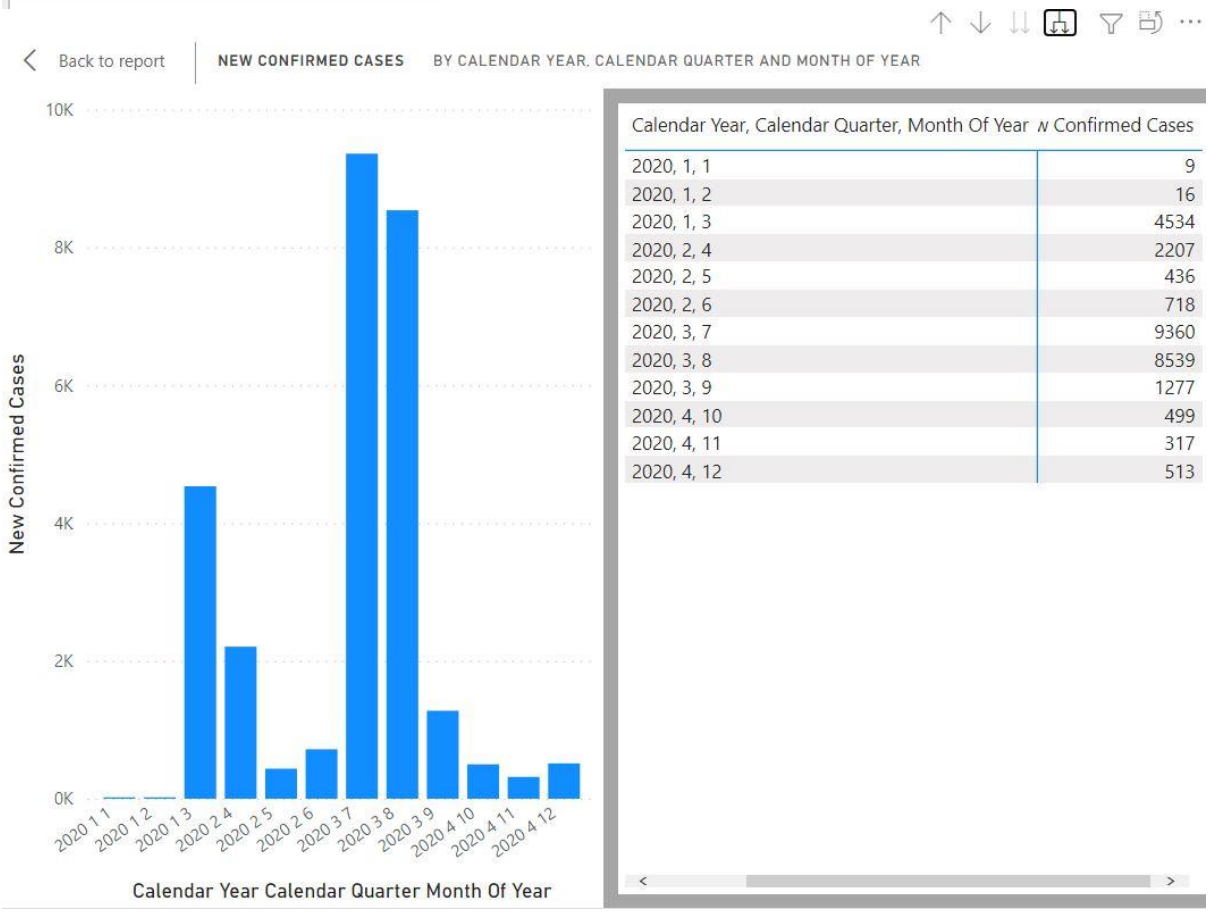


3.1.3 What is the number of confirmed cases in each month of 2020 in Australia?



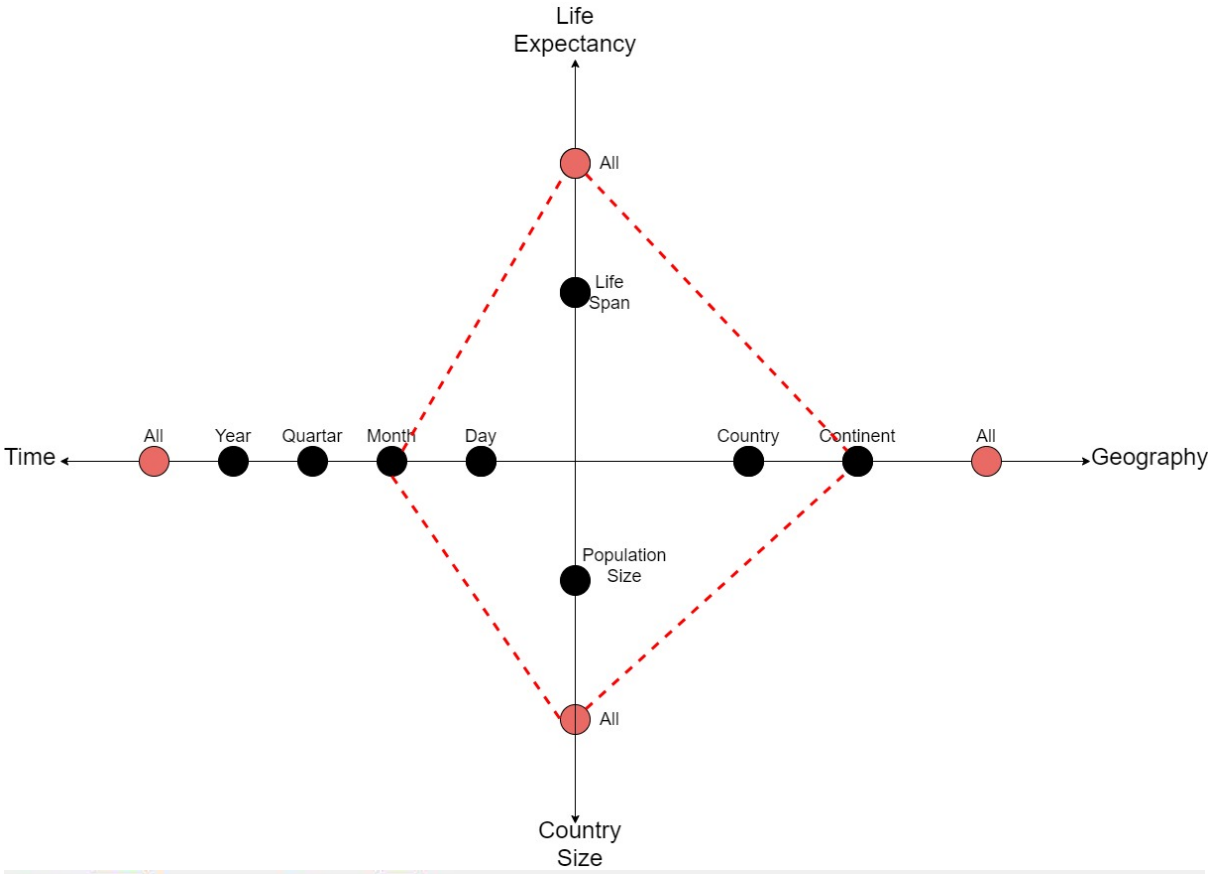
Dimension	Hierarchy	Operator	Filter Expression
Date	Calendar Year	Contains	2020
Geography	Country	Contains	Australia
<Select dimension>			

Month Of Year	New Confirmed Cases
1	9
10	499
11	317
12	513
2	16
3	4534
4	2207
5	436
6	718
7	9360
8	8539
9	1277



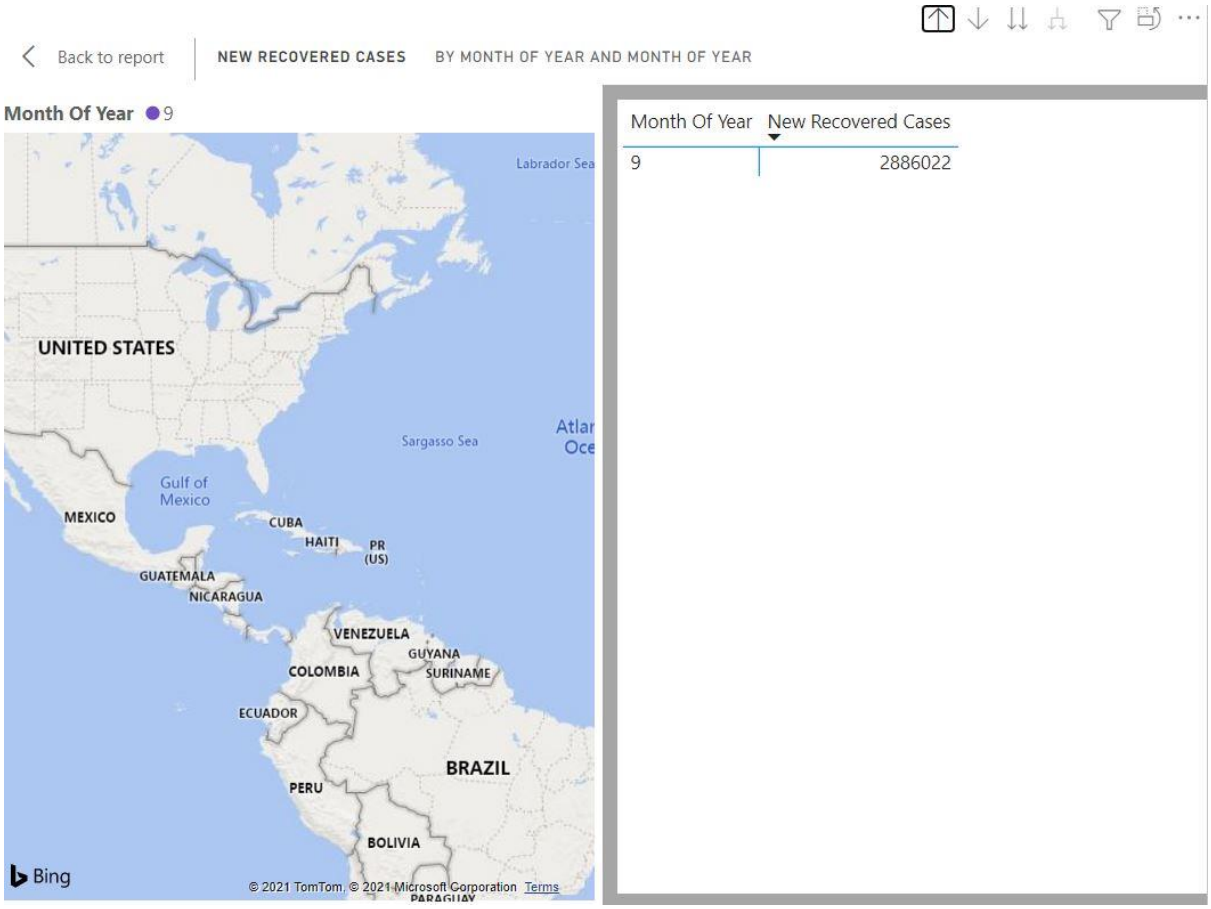
Calendar Year, Calendar Quarter, Month Of Year	Confirmed Cases
2020, 1, 1	9
2020, 1, 2	16
2020, 1, 3	4534
2020, 2, 4	2207
2020, 2, 5	436
2020, 2, 6	718
2020, 3, 7	9360
2020, 3, 8	8539
2020, 3, 9	1277
2020, 4, 10	499
2020, 4, 11	317
2020, 4, 12	513

3.1.4 In Sept 2020, how many recovered cases are there in the region of the Americas?

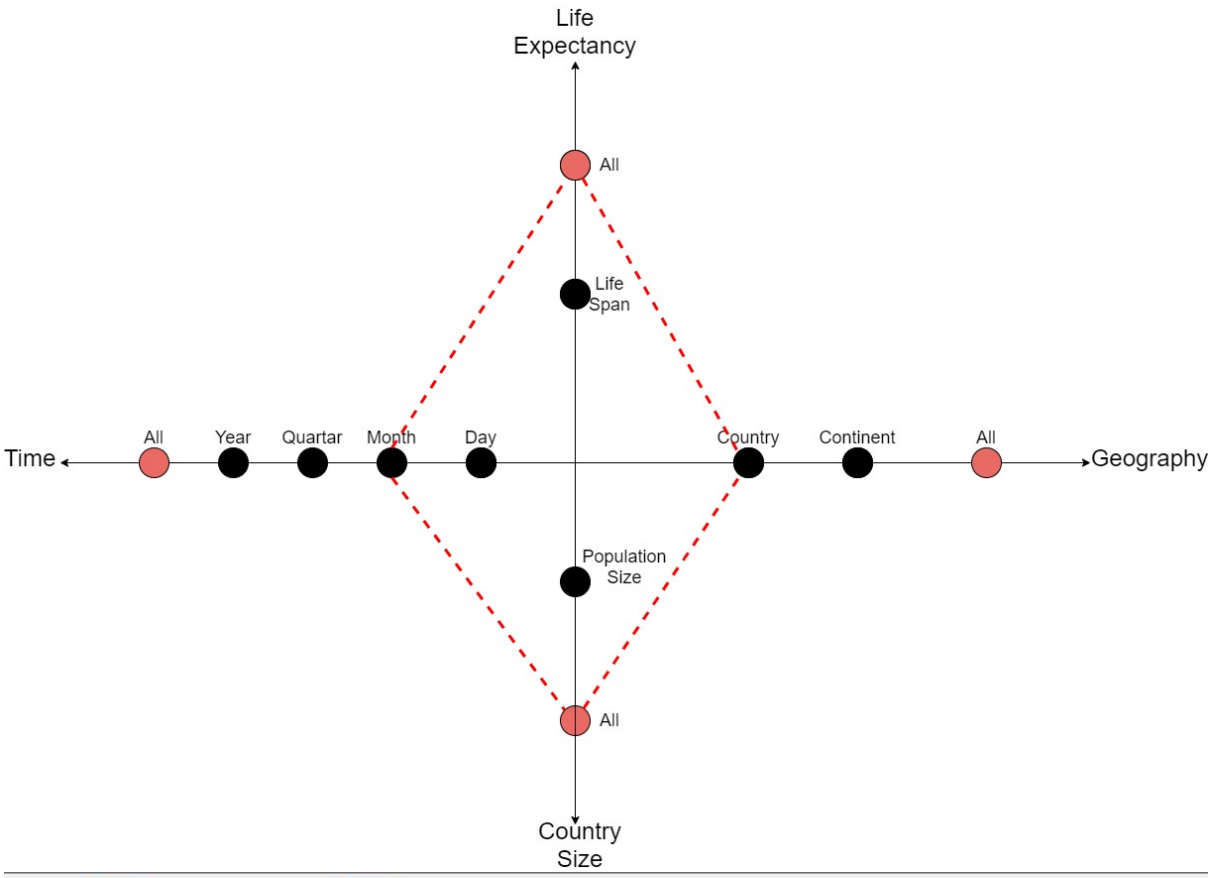


Dimension	Hierarchy	Operator	Filter Expression
Date	Calendar Year	Contains	2020
Date	Month Of Year	Equal	{ 9 }
Geography	CONTINENT	Contains	America
<Select dimension>			

New Recovered Cases
2886022

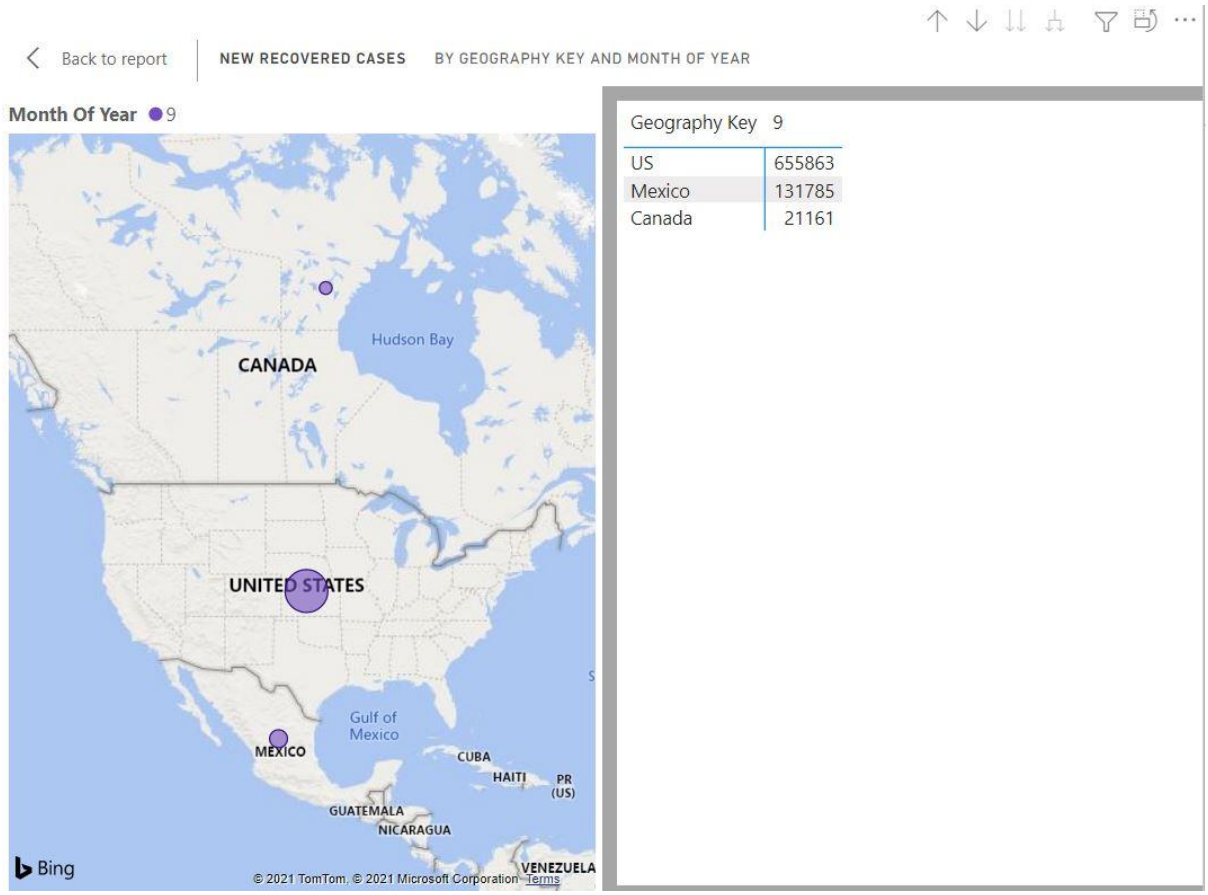

Back to report
NEW RECOVERED CASES
BY MONTH OF YEAR AND MONTH OF YEAR
Month Of Year 9
Labrador Sea
UNITED STATES
Sargasso Sea
Gulf of Mexico
MEXICO
CUBA
HAITI
PR (US)
GUATEMALA
NICARAGUA
VENEZUELA
GUYANA
SURINAME
COLOMBIA
ECUADOR
BRAZIL
PERU
BOLIVIA
PARAGUAY
Bing
© 2021 TomTom, © 2021 Microsoft Corporation Terms
Month Of Year
New Recovered Cases
9
2886022

3.1.5 How many recovered cases in the United States, Canada and Mexico, respectively, in Sep 2020?

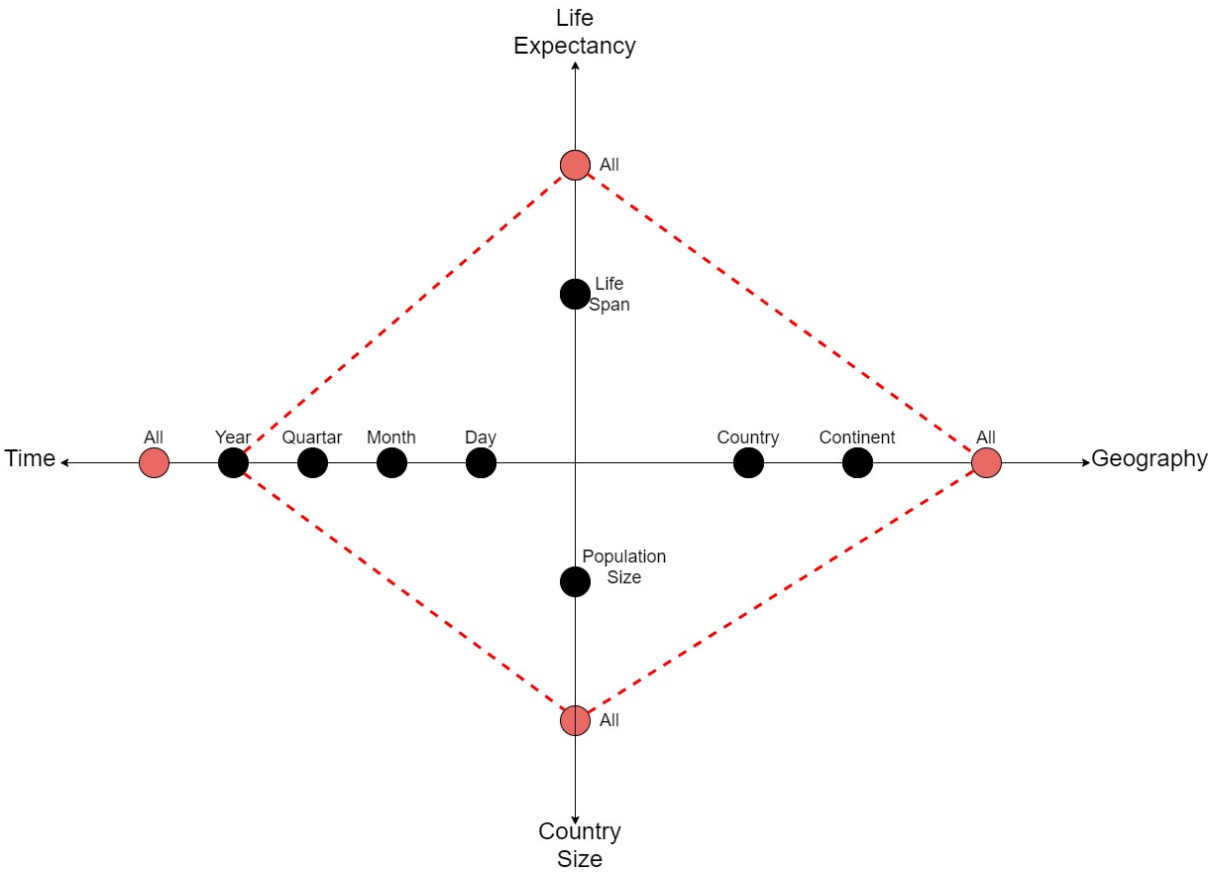


Dimension	Hierarchy	Operator	Filter Expression
Date	Calendar Year	Contains	2020
Date	Month Of Year	Equal	{ 9 }
Geography	Country	Equal	{ Canada, Mexico, US }
<Select dimension>			

Country	New Recovered Cases
Canada	21161
Mexico	131785
US	655863



3.1.6 What is the total number of covid deaths worldwide in 2020?



Dimension	Hierarchy	Operator	Filter Expression
Date	Calendar Year	Contains	2020
<Select dimension>			

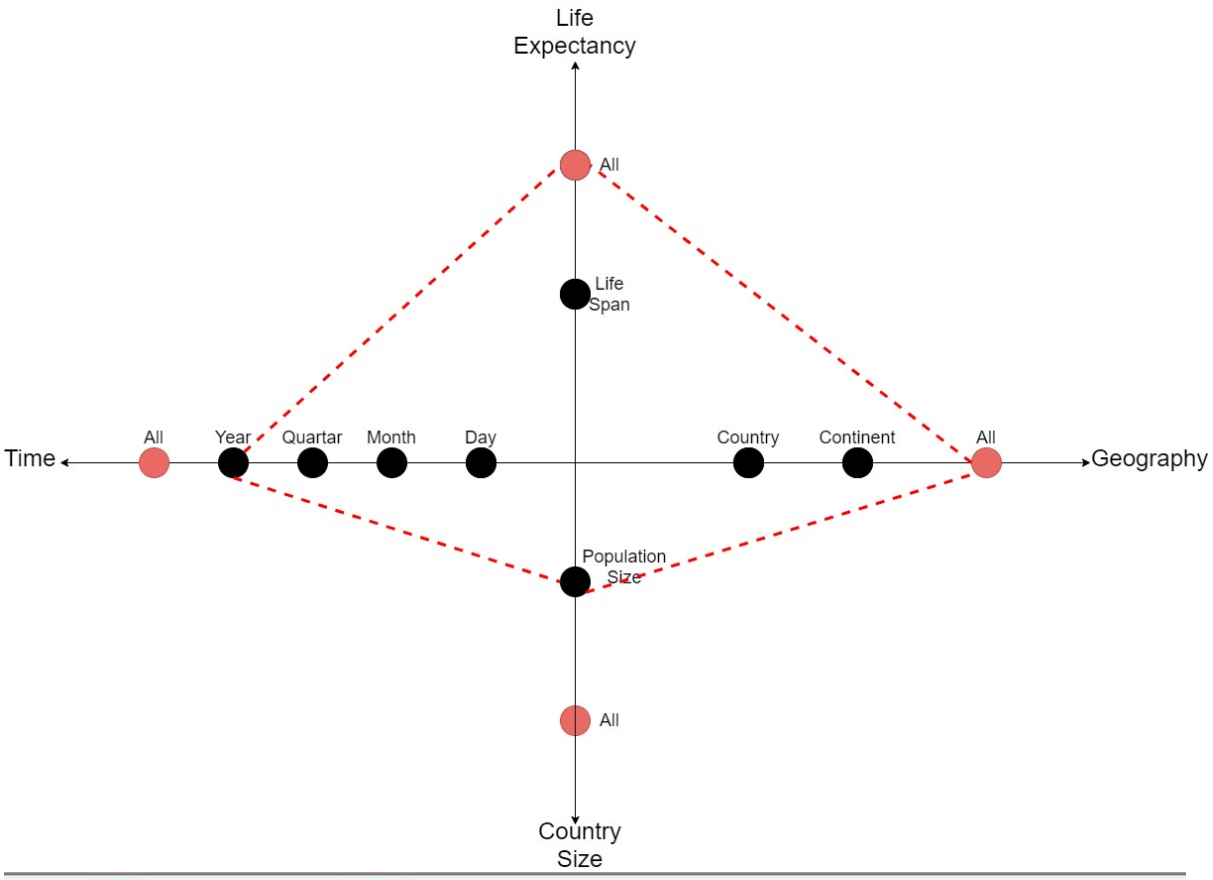
New Deaths
1829167

Back to report | NEW DEATHS BY CALENDAR YEAR

Calendar Year ● 2020

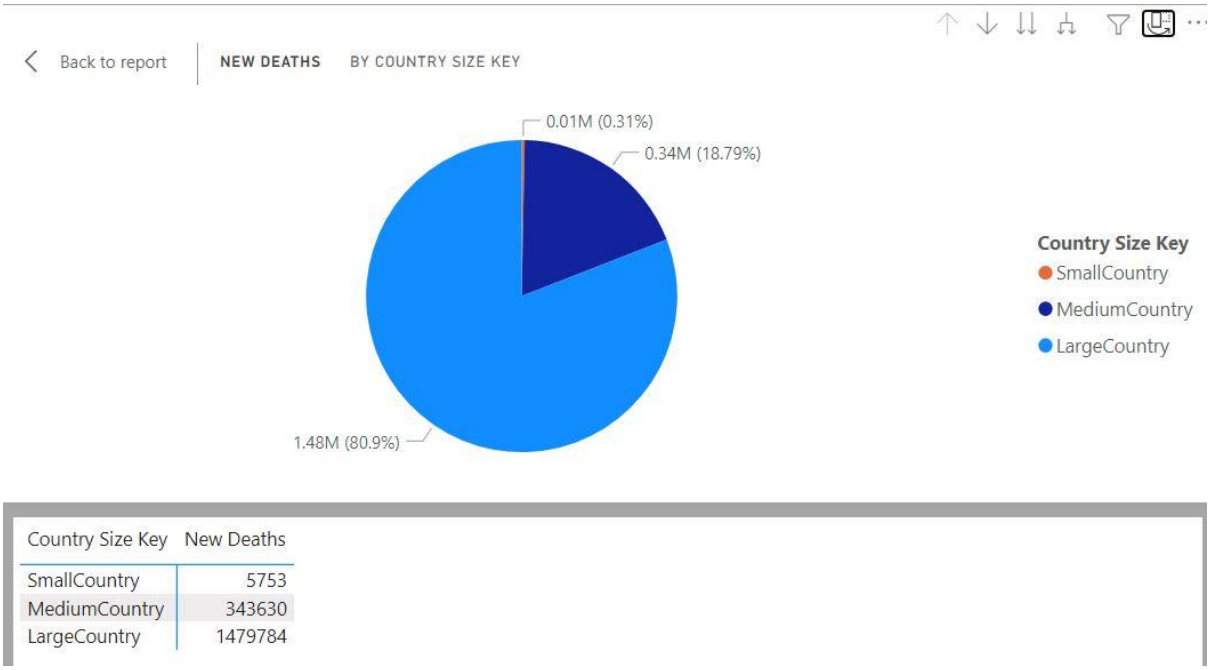
Calendar Year	New Deaths
2020	1829167

3.1.7 What is the total number of covid deaths in large countries, medium countries and small countries, respectively, in 2020?



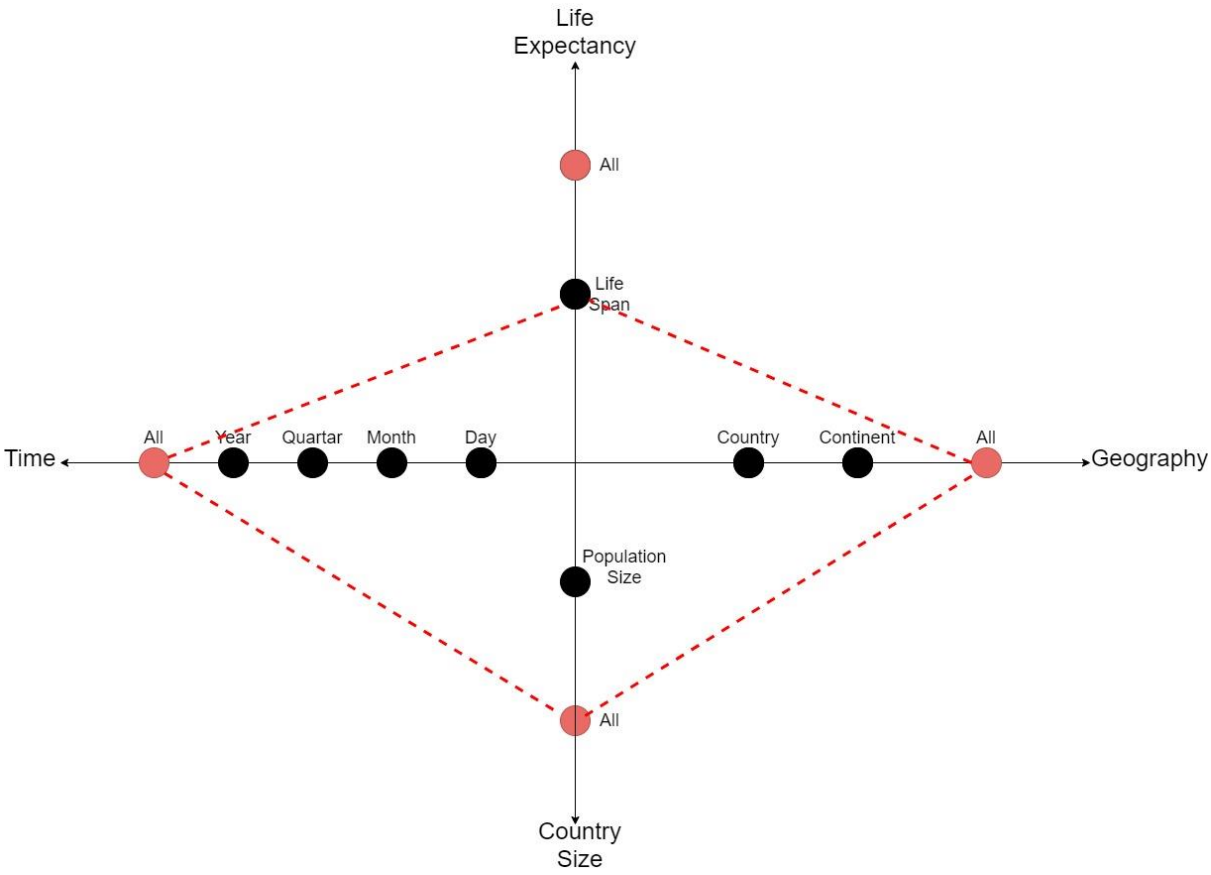
Dimension	Hierarchy	Operator	Filter Expression
Date	Calendar Year	Contains	2020
<Select dimension>			

Country Size Key	New Deaths
SmallCountry	5753
MediumCountry	343630
LargeCountry	1479784

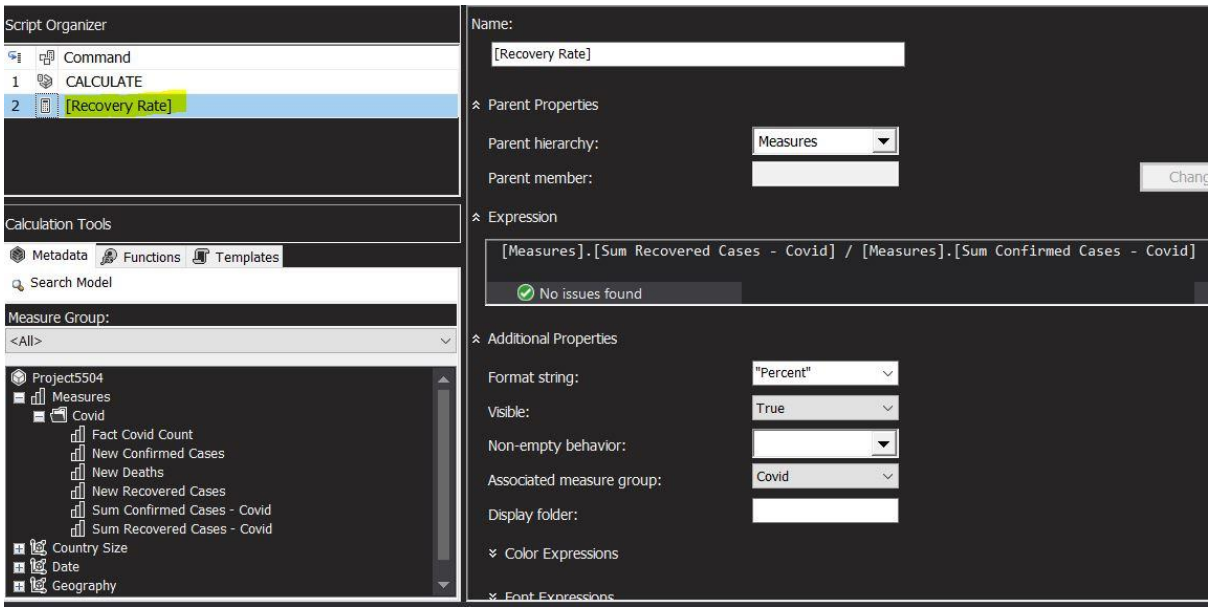


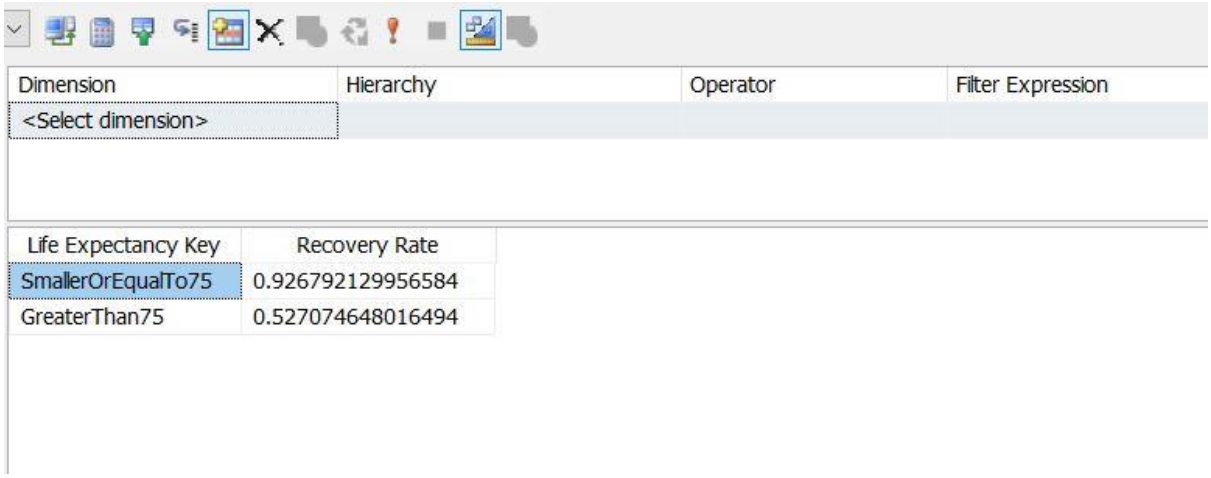
Country Size Key	New Deaths
SmallCountry	5753
MediumCountry	343630
LargeCountry	1479784

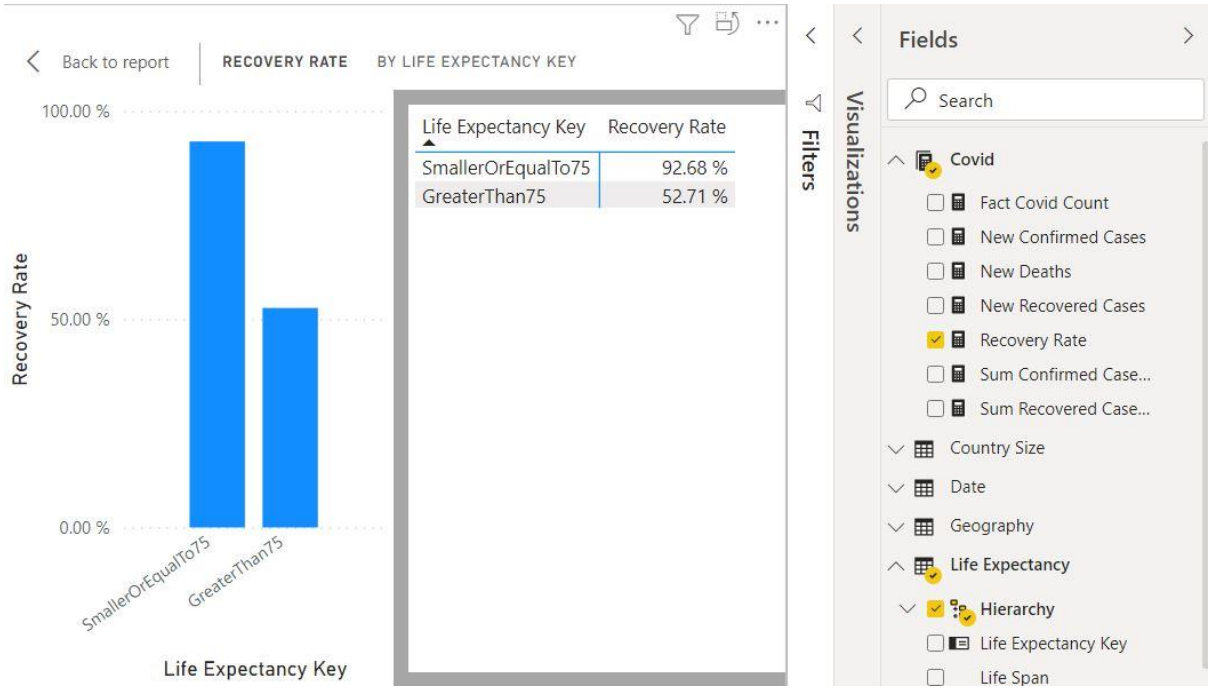
3.1.8 One conditional comparison query: Do countries with a life expectancy greater than 75 have a higher recovery rate?



Here, we need to define a new measure before running queries:

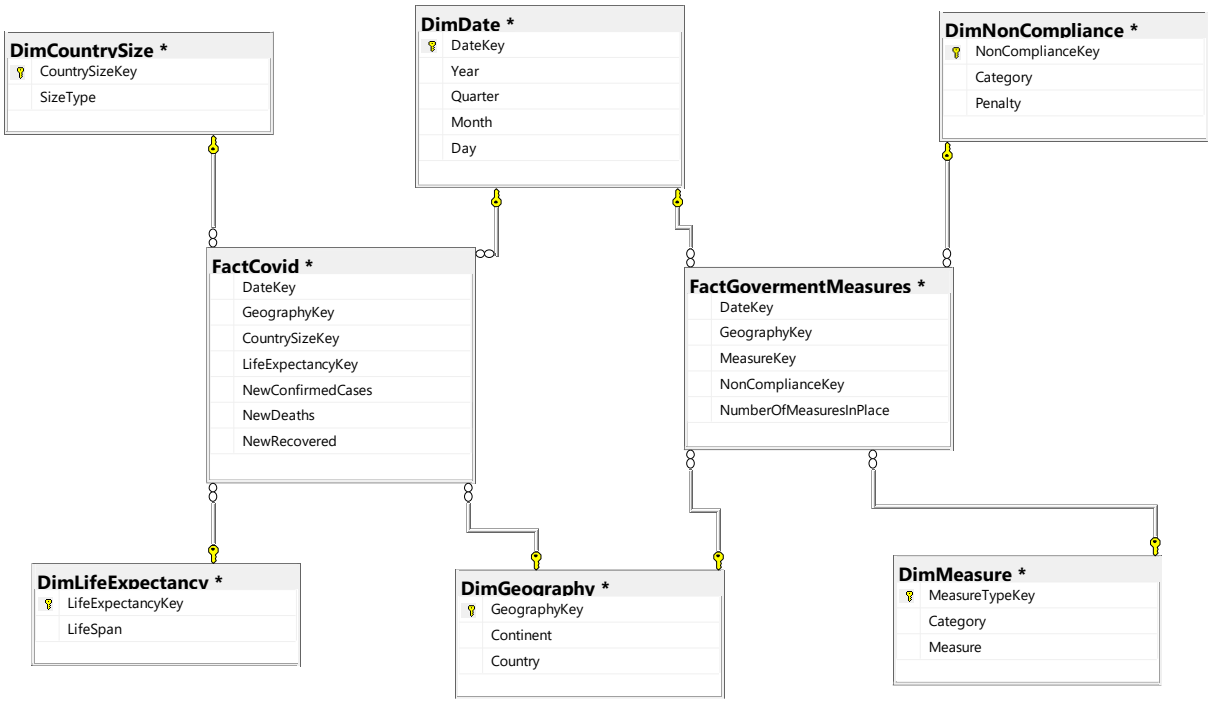

Script Organizer
Command
1 CALCULATE
2 [Recovery Rate]
Calculation Tools
Metadata
Functions
Templates
Search Model
Measure Group:
<All>
Project5504
Measures
Covid
Fact Covid Count
New Confirmed Cases
New Deaths
New Recovered Cases
Sum Confirmed Cases - Covid
Sum Recovered Cases - Covid
Country Size
Date
Geography
Name:
[Recovery Rate]
Parent Properties
Parent hierarchy:
Measures
Parent member:
Expression
[Measures].[Sum Recovered Cases - Covid] / [Measures].[Sum Confirmed Cases - Covid]
No issues found
Additional Properties
Format string:
"Percent"
Visible:
True
Non-empty behavior:
Associated measure group:
Covid
Display folder:
Color Expressions


Dimension
Hierarchy
Operator
Filter Expression
<Select dimension>
Life Expectancy Key
Recovery Rate
SmallerOrEqualTo75
0.926792129956584
GreaterThan75
0.527074648016494


Back to report
RECOVERY RATE
BY LIFE EXPECTANCY KEY
100.00 %
50.00 %
0.00 %
Recovery Rate
SmallerOrEqualTo75
GreaterThan75
Life Expectancy Key
Life Expectancy Key
Recovery Rate
SmallerOrEqualTo75
92.68 %
GreaterThan75
52.71 %
Filters
Visualizations
Fields
Search
Covid
Fact Covid Count
New Confirmed Cases
New Deaths
New Recovered Cases
Recovery Rate
Sum Confirmed Case...
Sum Recovered Case...
Country Size
Date
Geography
Life Expectancy
Hierarchy
Life Expectancy Key
Life Span

4. GALAXY SCHEMA

4.1 Database Diagram



4.2 Business Query Example:

In 2020, did The US have lower average daily confirmed case number when there was one or more social distancing measure in place, compared to the other days when there was no any government measures in place?