2021

CAB430 Assessment 1

PROBLEM SOLVING TASK 1

N10415483 JIYAN ZHU, N10505024 SHU DU

Tasks

Task 1:

a) Data profile

North America

Element	Data type	Format	Domain of Values	Minimum	Maximum	Average	Frequency of Distinct	Data issues
survey_Id	String	Xxx-xxxx	4012 distinct, 4010 unqiue	100	5119	Not int	Most frequent: 5111: 2, 5110: 2.	Survey_id should be all distinct value, because it is a primary key
survey_date	String	dd*mm*yyyy	93 distinct, 8unique	Not int	Not int	Not int	Most frequent: 06*07*2020	None
region	String	Xx	NA, SA, OC	Not int	Not int	Not int	NA: 3559, SA: 385, OC: 70.	The file is called North America, but there are data coming from another region, like SA and OC.
country	String	Xx	22 distinct, 4 unique	Not int	Not int	Not int	Most frequent: US: 3250	Same with Region, some country is not from NA
platitude	String	platitude	3958 distinct, 3902 unique	Not int	Not int	Not int	Many values appeared twice	None
Plenitude	String	platitude	3958 distinct, 3902 unique	Not int	Not int	Not int	Most frequent: -77.0807: 3	None

Participant	String	Participant	4004 distinct, 3994 unique	Not int	Not int	Not int	10 ids with appear twice.	Participant id should be unique, because it is a primary key.
Gender	String	XX	Female, male, other.	Not int	Not int	Not int	Female: 2053, Male: 1949, Other: 12.	None
Age	String	XX	0_10-100_110	Not int	Not int	Not int	30_40: 986, 40_50: 763, 20_30: 598, 50_60: 592, 60_70: 530, 70_80: 338, 80_90: 89, 10_20: 81. 90_100: 24, 0_10: 10, 100_110:3.	None
Height	String	Xxx	47 distinct, 8 unqiue	110	238	171.75	Most frequent: 178: 297	None
Weight	String	xx-xxx	68 distinct, 2 unqiue	44	180	84.61	Most frequent: 74: 202	None

Bmi	String	xx.x	371 distinct, 78 unqiue	11.9	125	28.6	Most frequent: 27.7: 96, 29: 92	None
Blood type	String	Xx/xxx	Unknown, op, ap, on, an, bp, adp, bn, ?, abn	Not int	Not int	Not int	Unknown: 1162, Op: 953, Ap: 923, On: 303, Bp: 251, Abp: 131, Bn: 59, Abn: 31, ?: 9.	Yes, “?” is missing value.
Insurance	String	xx-xxxx	Yes, no, blank	Not int	Not int	Not int	Yes: 3321, No: 461, Blank: 232.	None
Income	String	xx-xxxx	Med, high, low, blank, gov, ?	Not int	Not int	Not int	Med: 1860, High: 1631, Low: 343, Blank: 93, Gov: 81,	Yes, “?” is missing value

							?: 6.	
Race	String	XX	White, Hispanic, Asian, mixed, black, other, other, blank.	Not int	Not int	Not int	White: 3158, Hispanic: 350, Asian: 229, Mixed: 130, Black: 82, Other: 40, Blank: 25.	None
Immigrant	String	XX	Native, immigrant, blank	Not int	Not int	Not int	Native: 3526, Immigrant: 449, Blank: 39.	None
Response_id	String	Xxxxx	4004 distinct, 3994 unique.	Not int	Not int	Not int	10 values appear twice.	None
Smoking	String	XX	Never, quit10, quit5, vape, yeslight, yesmedium, yesheavy, ?.	Not int	Not int	Not int	Never: 2663, Quit10: 367, Quit5: 340, Quit0: 190, Vape: 136, Yeslight: 132, Yesmedium: 125, Yesheavy: 44,	Yes, there is missing values

							?: 17.	
Contact_count	String	Xx	22 distinct.	0	21	7.48	Most frequent: 1: 472	None
House_count	String	Xx	11 distinct	1	11	3.14	Most frequent: 2: 1568.	None
Public_transport_count	String	Xx	14 distinct, 1 uunique	Not int	Not int	Not int	Most frequent: 0: 3810	None
Working	String	XX	Stopped, travel critical, never, home, travel non critical.	Not int	Not int	Not int	Stopped: 1309, Never: 1201, Travel critical: 821, Home: 349, Travel non critical: 312, ?: 9	Yes, there are missing values.
Covid19_positive	String	X	1, 0	Not int	Not int	Not int	0: 2575, 1: 1439.	None
Covid19_symptoms	String	X	1, 0	Not int	Not int	Not int	0: 3544, 1: 470.	None
Covid_contact	String	X	1, 0	Not int	Not int	Not int	0: 3527, 1: 487.	None

Asthma	String	X	1, 0	Not int	Not int	Not int	0: 3511, 1: 503.	None
Kidney_isease	String	X	1, 0	Not int	Not int	Not int	0: 3965, 1: 49.	None
Liver_disease	String	X	1, 0	Not int	Not int	Not int	0:3990, 1: 24.	None
Compromised_immune	String	X	1, 0	Not int	Not int	Not int	0: 3773, 1: 241	None
Heart_disease	String	X	1, 0	Not int	Not int	Not int	0: 3875, 1: 139	None
Lung_disease	String	X	1, 0	Not int	Not int	Not int	0: 3924, 1: 90.	None
Diabetes	String	X	1, 0	Not int	Not int	Not int	0: 3674, 1: 340.	None
Hiv_postive	String	X	1, 0	Not int	Not int	Not int	0:4003, 1: 11.	None
Hypertension	String	X	1, 0	Not int	Not int	Not int	0: 3291, 1: 723.	None
Other_chronic	String	X	1, 0	Not int	Not int	Not int	0: 3738, 1: 276	None

Nursing_home	String	X	1, 0	Not int	Not int	Not int	0:3985, 1: 29	None
Health_worker	String	X	1, 0	Not int	Not int	Not int	0: 3772, 1: 242	None
Risk_infection	String	X	31 distinct, 7 unique	Not int	Not int	Not int	Most frequent: 5: 1617.	None
Risk_mortality	String	x.xx	278 distinct, 176 unique	Not int	Not int	Not int	Most frequent: 0.05.	None

Other Region

Element	Data type	Format	Domain of Values	Minimum	Maximum	Average	Frequency of Distinct	Data issues
survey_Id	Float	Xxx-xxxx	Unique	103	5124	not int	Unique	None
survey_date	Date	Dd/mm/yyyy	86 distinct, 18 unique.	Not int	Not int	Not int	Most frequent date: 07/06/2020	None
region	varchar	XX	EU, AS, AF	Not int	Not int	Not int	EU: 743, AS: 194, AF: 76.	None
country	varchar	XX	69 distinct, 15 unique	Not int	Not int	Not int	Most frequent country: GB: 343.	None
platitude	float	platitude	1000 distinct, 987 unique	Not int	Not int	Not int	There are 13 values appear twice.	None?
Plenitude	float	platitude	1000 distinct, 1003 unique	Not int	Not int	Not int	There are 12 values appear twice	None?
Participant	float	Participant	1001 distinct, 989 unique	Not int	Not int	Not int	5 id appeared twice	Id should be a primary key, so it must be unique
Gender	varchar	XX	Male, female	Not int	Not int	Not int	Male: 648, Female: 365.	None

Age	varchar	XX	0_10-100_110	Not int	Not int	Not int	30_40: 223, 40_50: 195, 50_60: 175, 20_30: 141, 60_70: 124, 70_80: 56, 10_20: 45, 80_90; 30, 90_100: 13, 0_10: 2, 100_110: 2.	None
Height	float	Xxx	35 distinct, 5 unique	110	198	172.28	Most Frequent: 170: 100.	None
Weight	float	xx-xxx	56 distinct, 8 unique	44	180	79.76	Most frequent: 70: 75	None
Bmi	float	xx.x	228 distinct, 72 unique	15	114	26..83	Most frequent: 27.1: 25	None
Blood type	varchar	XX/xxx	Unknown, ap, op, bo, on, adp, an, bn, abn, ?	Not int	Not int	Not int	Unknown: 272, Ap: 254, Op: 208,	Yes, “?” is missing value.

							Bp: 90, On: 73, Adp: 43, An: 31, Bn: 17, Abn: 7, ?: 5	
Insurance	varchar	xx-xxxx	Yes, no, blank	Not int	Not int	Not int	Yes: 580, No: 313, Blank: 120.	None
Income	varchar	xx-xxxx	Med, high, low, blank, gov.	Not int	Not int	Not int	Med: 459, High: 406, Low: 92, Blank: 37, Gov: 19	None
Race	varchar	XX	White, Asian, mixed, Hispanic, other, black, blank	Not int	Not int	Not int	White: 741, Asian: 160, Mixed: 38, Hispanic: 27, Other: 24,	None

							Black: 14, Blank: 9	
Immigrant	varchar	XX	Native, immigrant, blank	Not int	Not int	Not int	Native: 800, Immigrant: 198, Blank: 15.	None
Response_id	float	Xxxxx	1008 distinct, 1003 unique.	Not int	Not int	Not int	There are 5 values appeared twice.	This is a primary key, which means they should all be unique.
Smoking	varchar	XX	Never, quit10, quit5, quit0, yesmedium, yeslight, vape, yesheavy.	Not int	Not int	Not int	Never: 573, Quit10: 105, Quit5: 89, Quit0: 76, Yesmedium: 65, Yeslight: 43, Vape: 37, Yesheavy: 25.	None
Contact_count	float	XX	22 distinct, 2 unique	0	21	7.77	Most frequent: 10: 115	None
House_count	float	XX	2, 3, 4, 1, 5, 6, 11, 7, 8, 10, 9	1	11	3.14	Most frequent: 2: 368.	None

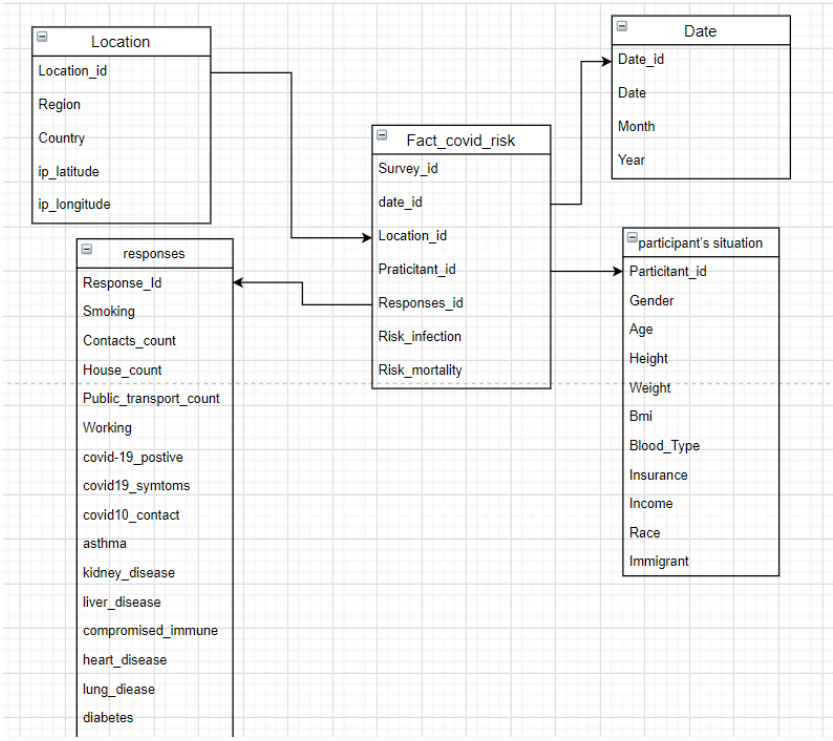
Public_transport_count	float	XX	0, 1, 2, 3, 4, 5, 3, 7, 8, 10	Not int	Not int	Not int	Most frequent: 0: 946	None
Working	varchar	XX	Never, stopped, travel critical, travel non critical, home, ?	Not int	Not int	Not int	Never: 318, Stopped: 285, Travel critical: 161, Travel non critical: 139, Home: 104, ?: 6	There are missing values in this column.
Covid19_positive	float	X	1, 0	Not int	Not int	Not int	0: 680, 1: 333	None
Covid19_symptoms	float	X	1, 0	Not int	Not int	Not int	0: 907, 1: 106.	None
Covid_contact	float	X	1, 0	Not int	Not int	Not int	0: 922, 1: 91	None
Asthma	float	X	1, 0	Not int	Not int	Not int	0: 896, 1: 117.	None
Kidney_disease	float	X	1, 0	Not int	Not int	Not int	0: 986, 1: 27.	None
Liver_disease	float	X	1, 0	Not int	Not int	Not int	0: 989, 24.	None

Compromised_immune	float	X	1, 0	Not int	Not int	Not int	0: 939, 74.	None
Heart_disease	float	X	1, 0	Not int	Not int	Not int	0: 965, 1: 48	None
Lung_disease	float	X	1, 0	Not int	Not int	Not int	0: 981, 32.	None
Diabetes	float	X	1, 0	Not int	Not int	Not int	0: 939, 1: 74.	None
Hiv_postive	float	X	1, 0	Not int	Not int	Not int	0: 1006, 1: 7	None
Hypertension	float	X	1, 0	Not int	Not int	Not int	0: 835, 1: 178	None
Other_chronic	float	X	1, 0	Not int	Not int	Not int	0: 935, 1: 78.	None
Nursing_home	float	X	1, 0	Not int	Not int	Not int	0: 996, 1: 17.	None
Health_worker	float	X	1, 0	Not int	Not int	Not int	0: 940, 1: 73.	None
Risk_infection	float	X	23 distinct, 7 unqiue	Not int	Not int	Not int	Most frequent: 5: 375	None

Risk_mortality	float	x.xx	313 distinct, 7 unique	Not int	Not int	Not int	Most frequent: 0.05: 243	None

Above two tables contain all the columns from both Other regions and North America. The default type for all columns in North America are strings, while for Other Regions are mainly float and varchar. The format for each column is mostly the same between two tables except ‘date’. Domain values shows how many distinct and unique values. Minimum, maximum, and average gives us a view on the data range of each column. Frequency of Distinct shows what is the majority class in our data. Lastly, the data requires to be cleaned as it contains nullable and invalid values, which are found to be empty values and “?”.

b) & c)



The start schema design is used as the structure of the database. In the centre is the fact table, which contains Survey_id, date_id, location_id, praticitant_id, responses_id, Risk_infection and risk mortality. The other four are the dimension tables, Date, Location, Responses and Participant's Situation. Date dimension table contains Date, Month, Year, and primary key Date_id. Location dimension table contains Location_id, region, country, platitude, and ip_longitude. The Participant's situation dimension table contains Participant_Id, gender, height, weight, bmi, blood_type, insurance, income, race and immigrant. The last dimension table is Reponses. The responses dimension table contains Response_Id, smoking, contacts_count, House_count, Public_transport_count, working, Covid-19_postive, covid19_symtoms, covid10_contact, asthma, kidney_disease, liver_disease, compromised_immune, heart_disease, lung_diease, diabetes, hiv_postive, hypertensioin, other_chroinc, nursing_home and hearth_worker.

Task 2:

a) SQL scripts for database creation

```
CREATE DATABASE Assignment
GO

USE Assignment

CREATE TABLE Dates
( Date_id date NOT NULL,
  Date int,
  Month int,
  Year int,
  PRIMARY KEY ( Date_id )
);

CREATE TABLE Locations
( Location_id nvarchar (255) NOT NULL ,
Country nvarchar (255),
Region nvarchar (255),
Ip_latitude nvarchar (255),
Ip_longitude nvarchar (255),
PRIMARY KEY ( Location_id )
);

CREATE TABLE Participant
(
Paticipant_id int NOT NULL,
Gender nvarchar (255),
Age nvarchar (255),
Height int,
Weight int,
Bmi float,
Blood_type nvarchar (255),
Insurance nvarchar (255),
Income nvarchar (255),
Race nvarchar (255),
Immigrant nvarchar (255),
PRIMARY KEY ( Paticipant_id )
);
```

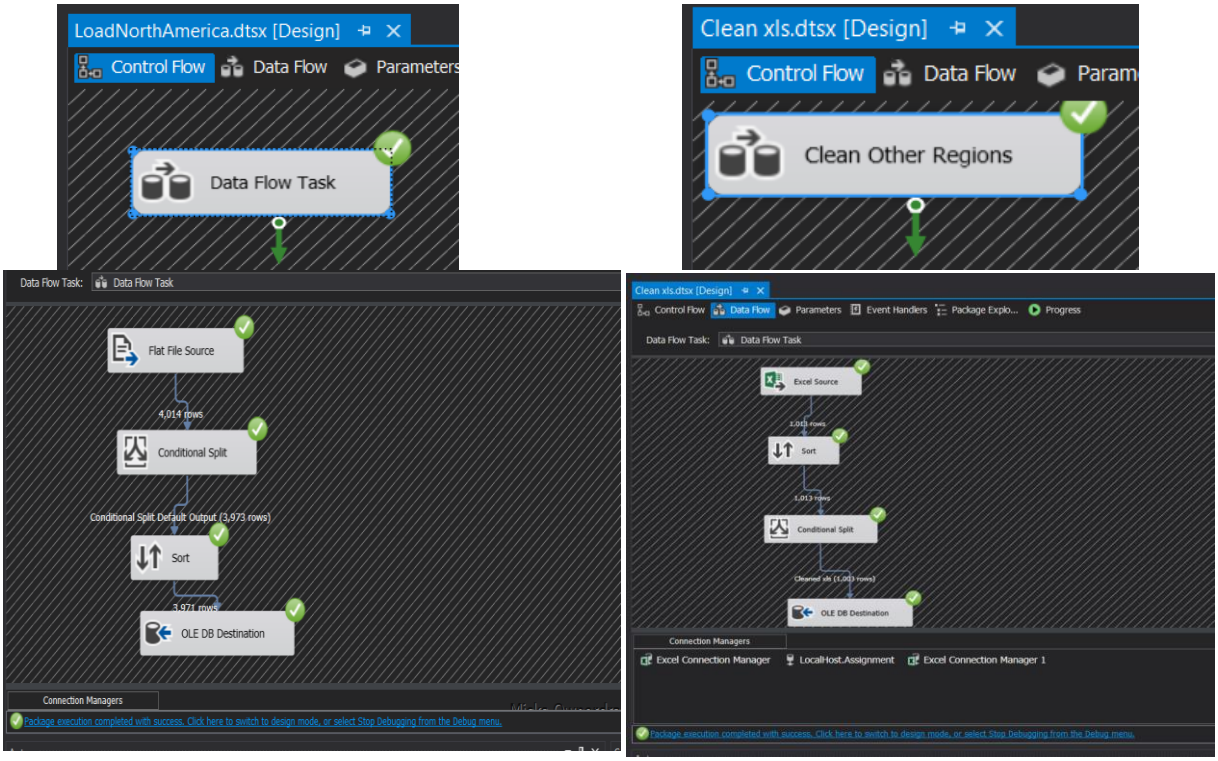
```
CREATE TABLE Response
(
Response_id int NOT NULL,
Smoking nvarchar (255),
Contact_count int,
House_count int,
Public_transport_count int,
Working nvarchar (255),
covid_19_positive int,
covid_19_symptoms int,
covid19_contact int,
asthma int,
kidney_disease int,
liver_disease int,
compromised_immune int,
heart_disease  int,
lung_disease int,
diabetes int,
hiv_positive int,
hypertension int,
other_chronic int,
nursing_home int,
health_worker int,
PRIMARY KEY ( Response_id )
);


CREATE TABLE Fact_covid_risk
(
Survey_id int NOT NULL,
Date_id date,
Location_id nvarchar (255),
Paticipant_id int,
Response_id int,
risk_infection int,
risk_mortality float,
PRIMARY KEY (Survey_id),
FOREIGN KEY ( Date_id ) REFERENCES Dates ( Date_id ),
FOREIGN KEY ( Location_id) REFERENCES Locations ( Location_id ),
FOREIGN KEY ( Paticipant_id ) REFERENCES Participant ( Paticipant_id ),
FOREIGN KEY ( Response_id ) REFERENCES Response ( Response_id )
)
```

b) A screenshot of the ER diagram of your database.

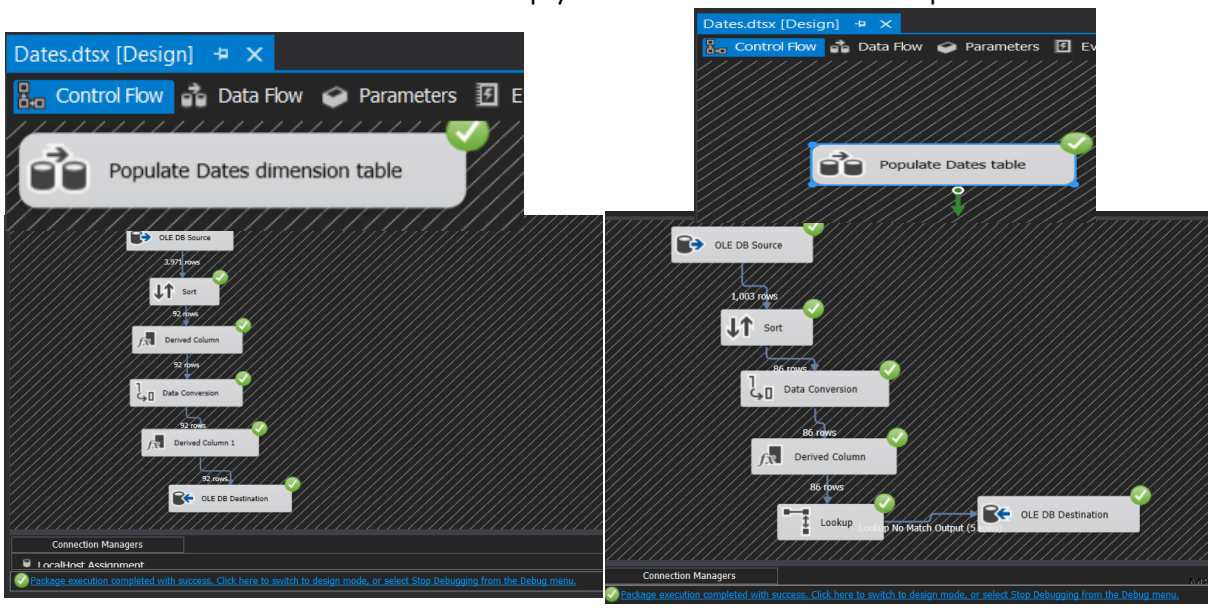
Task 3

a) An overview of your ETL application

There will be two ETL processes, one for North America, and one for Other Regions.



The purpose of these two control flows is to remove nullable and invalid values from the original datasets and load them into SQL server. By using conditional spilt, it allows the data set spilt out all the rows contain "?" and values that are empty. Sort were used to remove duplicate records.

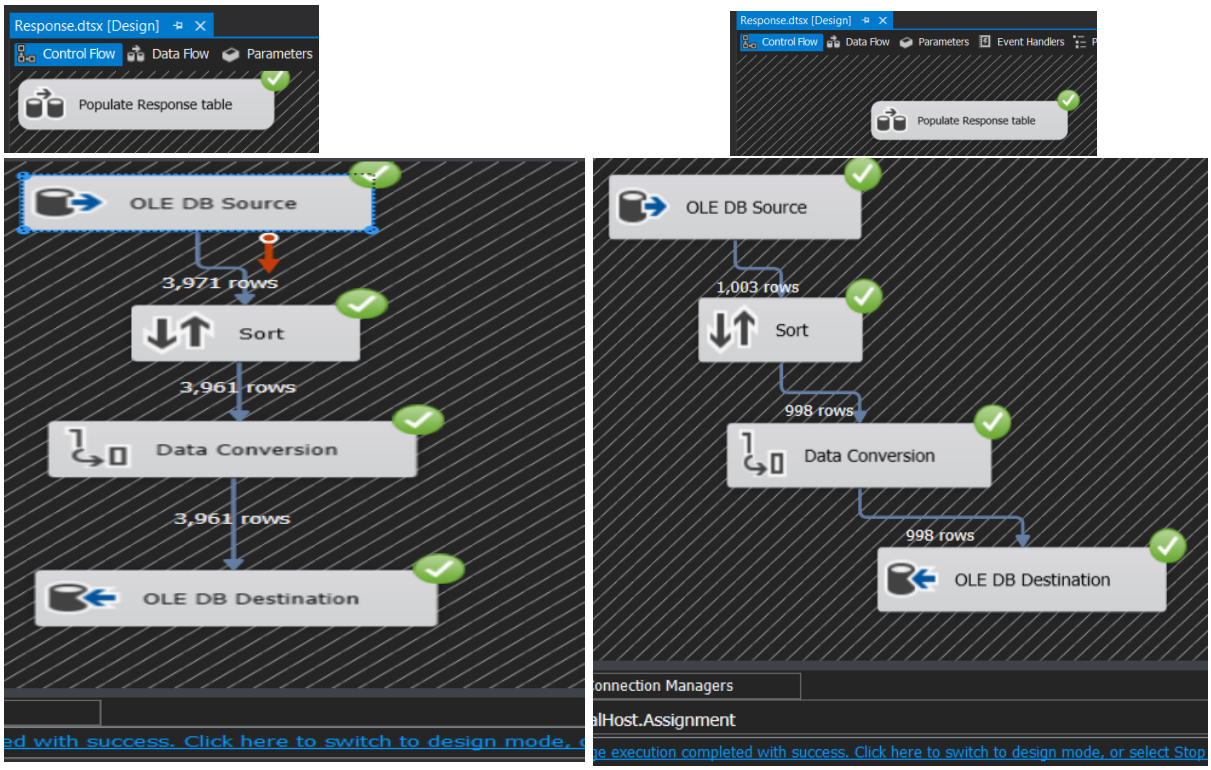


After cleaned both data sets, the next step will be to import all the date's value in to the data base. First, sort the dates by ascending order and remove duplicate dates records. In North America's cleaned data, Date_id was created by replace "*" with "/" from Survey_date, it will allow those value

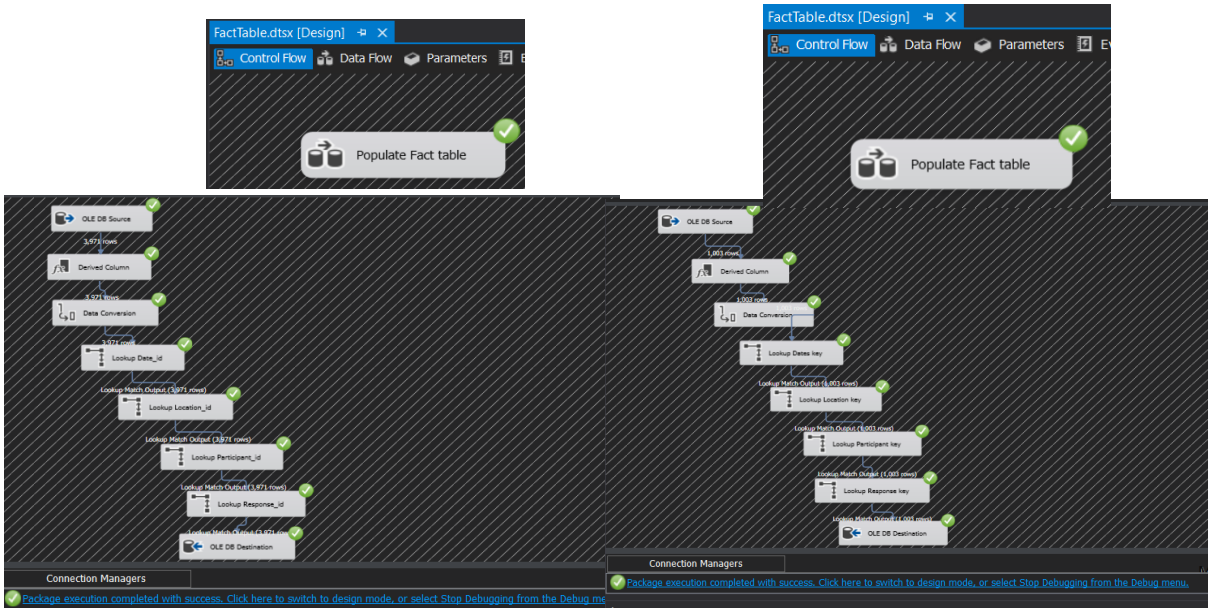
to convert to DATE type.While the date in Other Regions is correctly formated and nothing needs to be replaced. Then, the Date, Month, Year attributes are created accordingly. Because the date data in Other Regions were populated after North America's, so a look up transformation was used to filter exsisting records.

For these control flows, the data from two sources were transformed and loaded into the Location table. Location id was generated by concating the longitude and latitude values.

To populate Participant table, duplicate Participant id were removed first, and the rest were converted to the correct data type before loading into destination table.

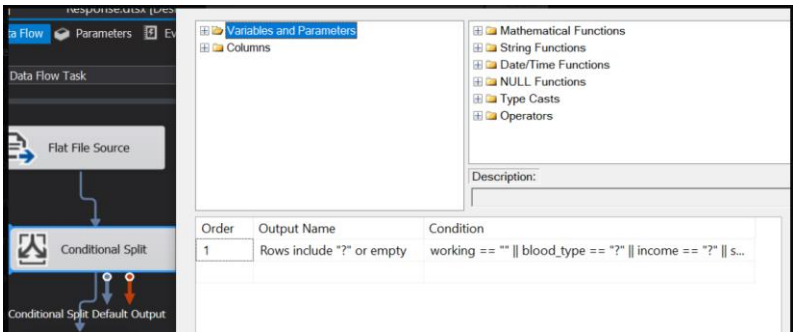


Similar to the previous control flow, the relative data were loaded into Response table .

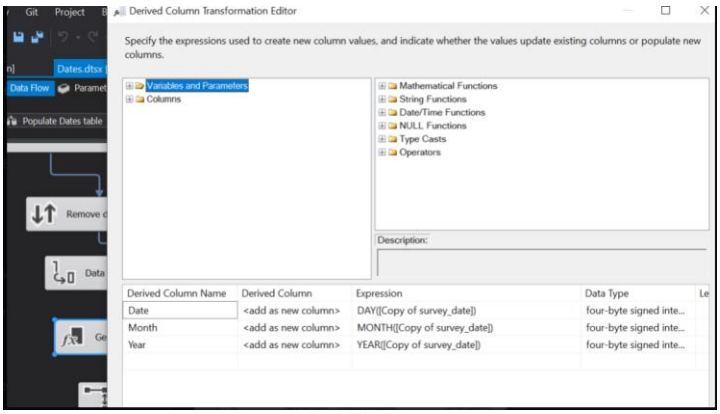


Having populated all dimension tables, the relative data is loaded into the fact table. By using several look up transformations, the matched attribute from dimension tables can be inserted into the fact table.

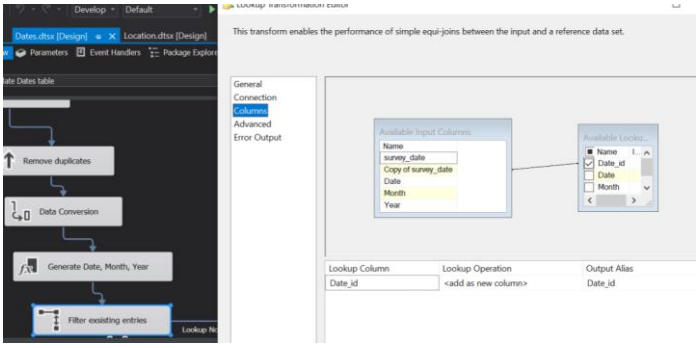
b) Explanation of the Transformations used in your ETL application



This is an example where a Conditional Split transformation is used. The purpse of this condlitional split is to clean the dataset by removing nullable and invalid records. Inside the condition is rule on what to use, which are any rows contain "?" and any rows that are empty. Conditional Split was not only used in the Response dimension table, it is also used in all other diemsion tables to help clean the data.



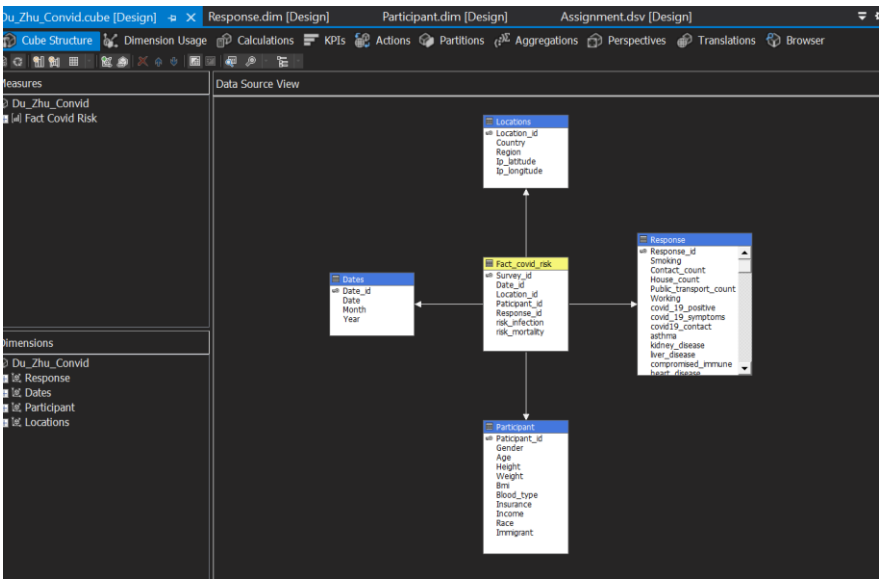
Above screen shot is a devided column transformation. The purpse of this transformation is to create 3 new attributes for Date table. The given data sources do not have the column Date, Month and year. So, this transformation allows us to generate these 3 columns. Similarly, this transformation was used to derive location id for location table.

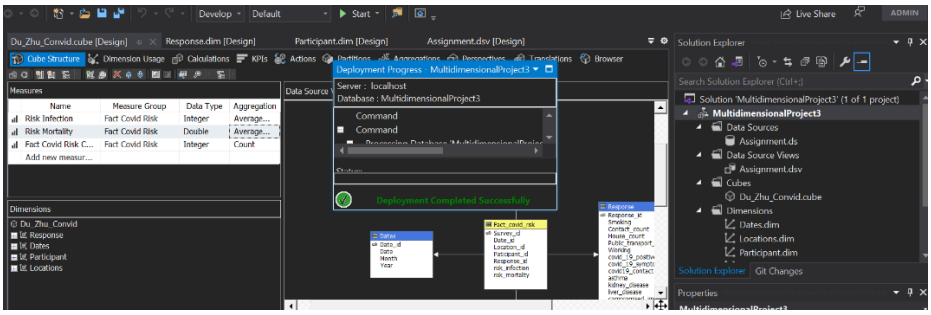


This shows a look up transformation. The purpse of this transformation is to filter exsisting records from the destination Date table. Because the date records from North America were loaded into the Date table before Other Regions, so it requires a look up transformation to pass only the non exsisting records, or it would violate the primary key constraint.

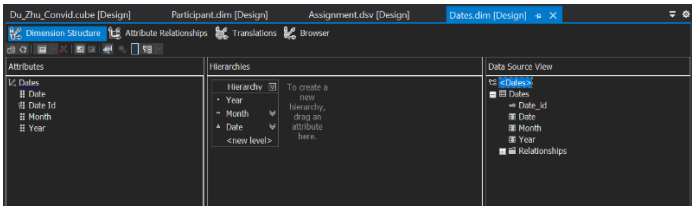
Task 4:

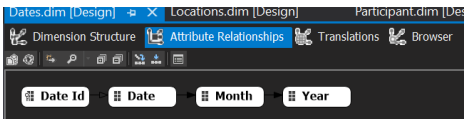
a) Provide a description to each data cube



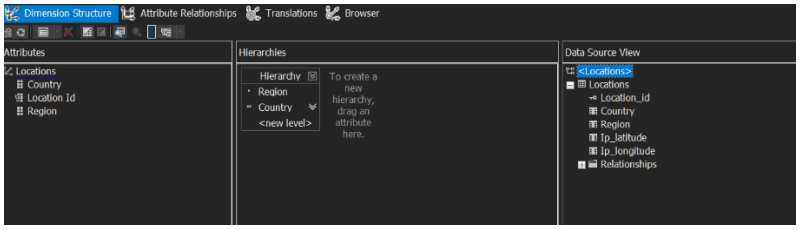


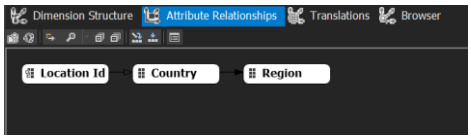
There is only one cube created, however, the cube contains all four dimension table and one fact table. This cube provides three measure, which are averge for Risk infection, Risk mortaility and Count of Records. The cube name contains both students' surname.



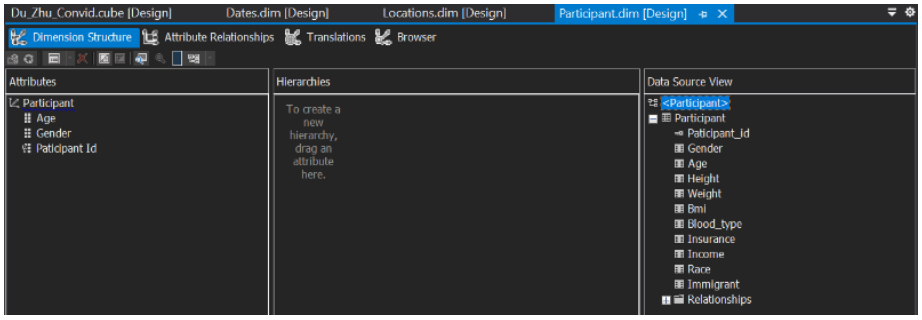


This is the Date dimension, which has 3 attributes Date, Month and Year. The hierarchy relationship is Date> Month > Year.

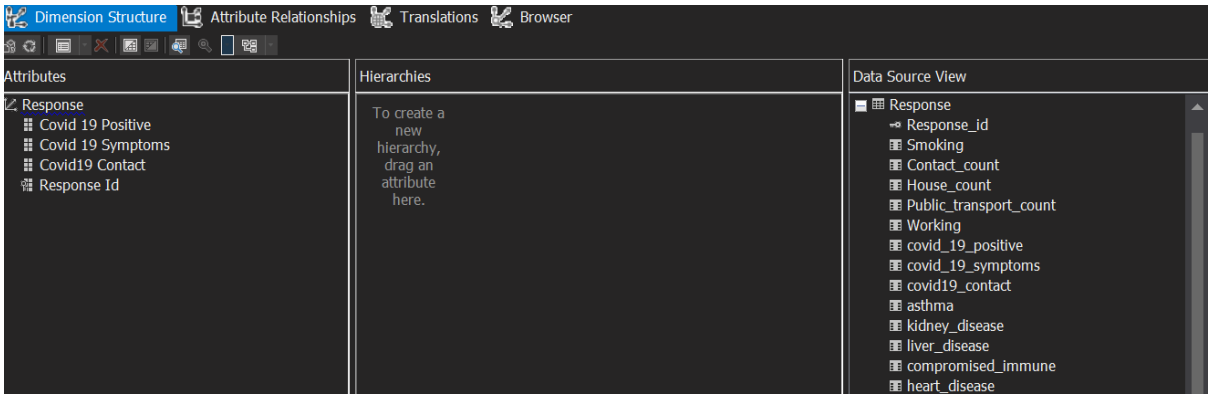




This is the Location dimension, which has 2 attributes Country and Region. The hierarchy relationship is Location ID > Country > Region.

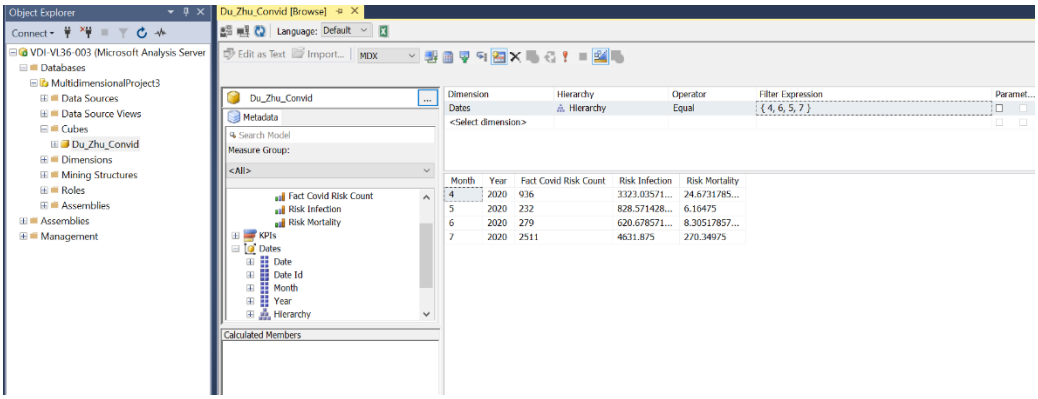


For the participant dimension, Age and Gender were selected among a list of 10 from Participant table.

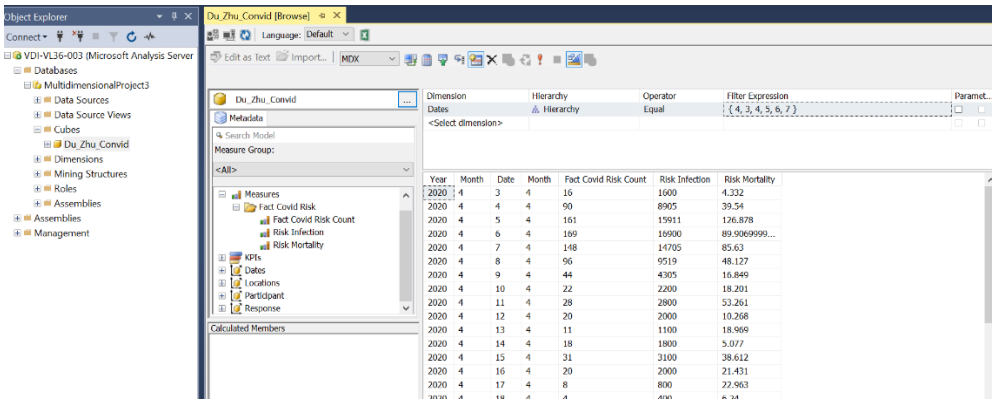


Lastly, the 3 attributes related to Convid 19 results were chosen to be in the Response dimension. There is no require for a hierarchy relationship for attributes in the Participant and Response dimension.

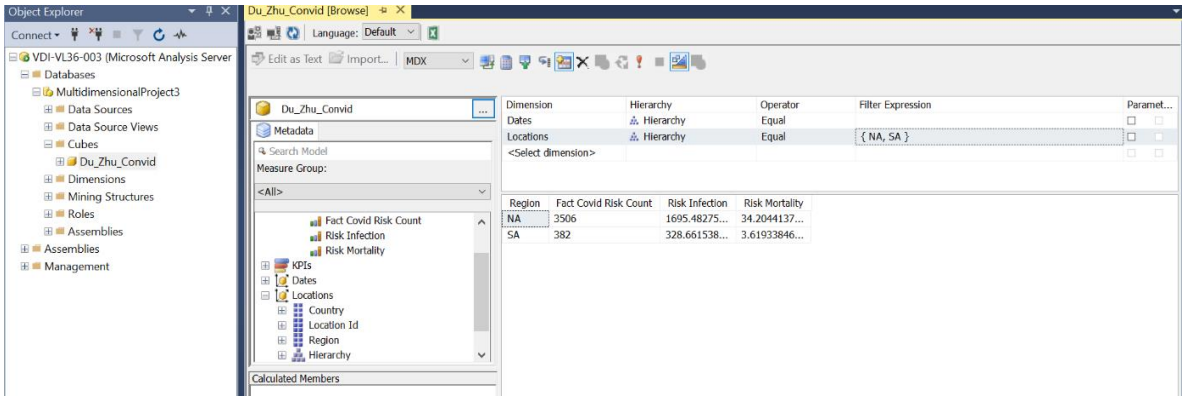
c) example screenshots of a query result



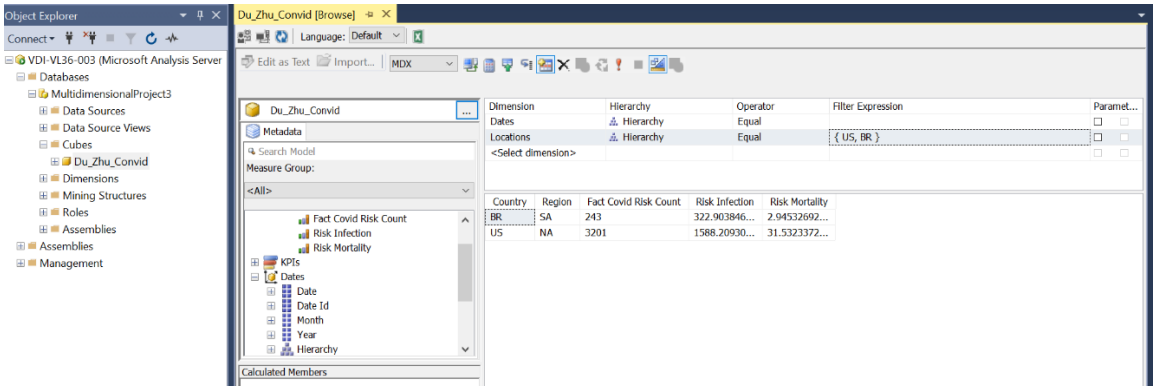
Results for Month between April to July.



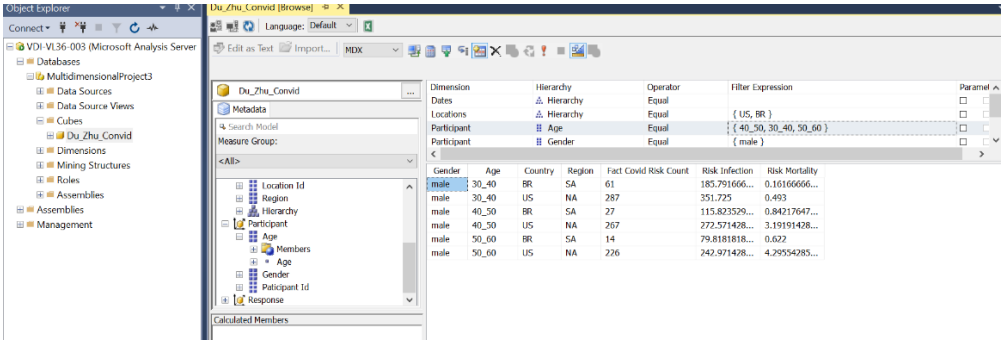
Results on 5 days start from 3rd of Arpil.



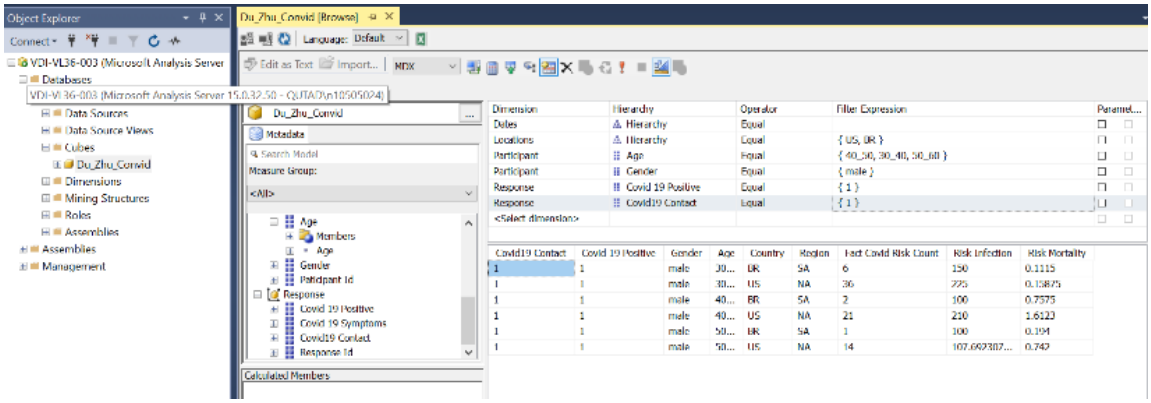
Results for region NA and SA.



Results for country US and BR.



Results for male from age between 30 to age in country NA and BR.



The query result is a extend of the preurse query. This reviews how many of them has contacted covid 19 and tested postive

Marking Sheet

Student(s)

Name	Student ID	Total Marks
JiYan Zhu	N10415483	30 /30
Shu Du	N10505024	

Task 1	**Comments**	**Marks**
Source data analysis	**Completed**	3 /3
The design of a fact table and its dimensions	**Completed**	1 /1
Description of the schema	**Completed**	1 /1
Sub-total mark		5 **/5**

Task 2	**Comments**	**Marks**
SQL statements to create a complete database including all the tables	**Completed**	2 /2
Correct ER diagram of database	**Completed**	1 /1
Sub-total mark		3 **/3**

Task 3	**Comments**	**Marks**
An overview of your ETL application	**Completed**	8 /8
Explain three transformations used in your ETL application.	**Completed**	6 /6

Sub-total mark		14 **/14**

Task 4	**Comments**	**Marks**
Description of the data cube(s)	**Completed**	3 /3
Result screenshots with brief explanations for satisfying the company's expectations.	**Completed**	3 /3
Sub-total mark		6 **/6**

General	**Comments**	**Marks**
Report presentation	**Completed**	2 /2
Sub-total mark		2 **/2**