pdfsubject=DWBI Project
pdfsubject=DWBI Project

National College of Ireland

Data Warehousing and Business Intelligence Project

on

Global Incidence and Mortality Analysis on Tuberculosis

Atif Feroz Jeelani
x17169992

MSc Data Analytics – 2018/9

Submitted to: Dr. Simon Caton

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Atif Feroz Jeelani
Student ID:	x17169992
Programme:	MSc Data Analytics
Year:	2018/9
Module:	Data Warehousing and Business Intelligence
Lecturer:	Dr. Simon Caton
Submission Due Date:	26/11/2018
Project Title:	Global Incidence and Mortality Analysis on Tuberculosis

I hereby certify that the information contained in this (my submission) is information pertaining to my own individual work that I conducted for this project. All information other than my own contribution is fully and appropriately referenced and listed in the relevant bibliography section. I assert that I have not referred to any work(s) other than those listed. I also include my TurnItIn report with this submission.

ALL materials used must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is an act of plagiarism and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	November 26, 2018

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Table 1: Mark sheet – do not edit

Criteria	Mark Awarded	Comment(s)
Objectives	of 5	
Related Work	of 10	
Data	of 25	
ETL	of 20	
Application	of 30	
Video	of 10	
Presentation	of 10	
Total	of 100	

Project Check List

This section capture the core requirements that the project entails represented as a check list for convenience.

- ☒ Used LaTeX template
- ☐ Three Business Requirements listed in introduction
- ☐ At least one structured data source
- ☐ At least one unstructured data source
- ☐ At least three sources of data
- ☐ Described all sources of data
- ☐ All sources of data are less than one year old, i.e. released after 17/09/2017
- ☐ Inserted and discussed star schema
- ☐ Completed logical data map
- ☐ Discussed the high level ETL strategy
- ☐ Provided 3 BI queries
- ☐ Detailed the sources of data used in each query
- ☐ Discussed the implications of results in each query
- ☐ Reviewed at least 5-10 appropriate papers on topic of your DWBI project

Global Incidence and Mortality Analysis on Tuberculosis

Atif Feroz Jeelani
x17169992

November 26, 2018

Abstract

Tuberculosis remains a major global health problem. While much progress has been made with tuberculosis control, the World Health Organization (WHO) estimates that 9 million people developed tuberculosis in 2013 and that 1.5 million died, including 360 000 people who were infected with human immunodeficiency virus (HIV) . People suffering from human immunodeficiency virus have high potential to develop active tuberculosis disease than people who do not suffer with HIV .In many countries tuberculosis diagnosis still relies on older out dated tools and hence resulting in a rapid rate of tuberculosis incidence giving rise to large number of mortality rates across the globe.

This project report explains the specic problem. Here I have taken data from different sources which gives me a global incidence and mortality rates of the people across more than 190 countries over a time span of sixteen years (2000 to 2016), suffering from tuberculosis and also people who had HIV who were than prone to TB. This report will elaborate and also provide an analysis of Incidence and Mortality score at a Global level and compare them with patients countries with HIV patients . The resulting queries and logical data mapping and stratifying of ETL is also explained in this report.

1 Introduction

Tuberculosis is the leading infectious cause of death worldwide, claiming 1.6 million lives annually Bloom (2018) .It has now exceeded infection with the human immunodeficiency virus and acquired immunodeficiency syndrome (HIV/AIDS) and malaria as the worlds largest cause of death from an infectious disease. The World Health Organization (WHO) estimates that there are 10.4 million new cases and 1.7 million deaths annually, including 0.4 million deaths in people with HIV infection, of which 95 percent occur in low-and-middle-income countries. Omara F Dogar & Qadeer3 (2017*a*)
In this study, we would be analyzing and comparing the incidences, mortality rates, gender disparity, Tuberculosis HIV Co-infection. We will be using horizontal bars and maps of the tableau to compare the various dimensions of the leading infectious cause of the death worldwide, i.e., Tuberculosis. By the end of the study, we would be clear with the results as to which country leads in the incidence cases and which country witnesses the highest mortality rates. We will also be clear with the impact of HIV on the incidences and mortality due to tuberculosis. Gender ranking would help understand, if there is any

discrepancy in the mortality rates between the genders infected with the Tuberculosis. This study would help in understanding the need for the specific countries to take serious steps to cure this disease and move towards prosperity
Three analysis i will be carrying out in this Data warehousing project are mentioned as below:

(Req-1) Is there significant regional disparity of the global incidence and mortality rates for tuberculosis at a Global scale ?

(Req-2) What world regions records the highest incidences of tuberculosis inclusive of HIV ? And what are the death/mortality rates of tuberculosis exclusive of HIV across global regions?

(Req-3) A comparison of gender differentiation based on Mortality rates ? Compared with the cases of incidence of TB based on Ranking .

2 Data Sources

Source	Type	Brief Summary
World Health Organization	Structured (csv)	This data-set provides information related to incident rates of TB across more than 190 countries from year 2000 to 2017 and also displays the people who were HIV positive and had TB which has been analyzed to form a BI query.
World Health Organization	Structured (csv)	The total cases of death rate has been measured in the following data-set who suffered from tuberculosis from across 190 countries from year 2000 to 2017 .
Statista	Structured (csv)	Statistics for total case of tuberculosis recorded in European countries for the year 2016
UNDP	Unstructured (pdf)	Here measures of incidence and mortality cases on gender differentiation is recorded for the year 2011 to 2013
World Health Organization	Structured (csv)	The people who recovered from TB and were relapsed are comprised in this data set for across the globe from year 2000 to 2017
Git-Hub	Structured	This data-set is extracted and used to provide a multidimensional drill down to the model based on country ,region ,sub region wise.

Table 2: Summary of sources of data used in the project

2.1 Source 1: World Health Organization

The first data-set is taken from World Health Organization (WHO) . The data set is from Global Health observatory data repository and is structured , http://apps.who.int/gho/data/node.main.1320?lang=en. This data set comprises of six columns of information related to the number of incidences of tuberculosis across more than 190 countries recorded from year 2000 to 2017 and also records for patients who had TB and were HIV positive which would be analyzed . It also contains other information like incidences of TB per 100,000 population and more which wont be used . The data was loaded in R , cleaned and transformed accordingly.

2.2 Source 2: World Health Organization

The second data set was also sourced from World Health Organization (WHO). The data is sourced from Global health observatory data repository and is structured data, http://apps.who.int/gho/data/node.main.1317?lang=en. The data that was extracted from this source is based on Mortality rates of tuberculosis country wise ranging from year 2000 to 2017 for more than 190 countries . There are four columns which represent country , year , number of deaths due to tuberculosis and death rate for 100,000 population having TB who are tested as HIV negative. However, the information about mortality across different countries are being analyzed in this project.

2.3 Source 3: Statista

The third data-set is taken from Statista . The data is of type structured , https://www.statista.com/statistics/630987/tuberculosis-cases-reported-in-europe/. The data-set comprises of 2 columns only. The statistic displays the number of cases of tuberculosis reported in European countries for the year 2016 which is analyzed to generate a BI query.

2.4 Source 4: UNDP (United Nations Development Programme)

The fourth data-set is taken from UNDP . The data is of type unstructured as it has been extracted from a Discussion paper under title Gender and tuberculosis which was available in .pdf format , .The data-set comprises of six columns which explains the rate of TB incidence and death rate on a gender differentiation from year 2011 to 2013 . The data was extracted , transformed and cleaned in a table format which was done using R. http://www.undp.org/content/dam/undp/library/HIV-AIDS/Gender%20HIV%20and%20Health/Gender%20and%20TB%20UNDP%20Discussion%20Paper%20(1).pdf

2.5 Source 5: World Health Organization

The fifth data-set was also taken from World Health Organization (WHO). The data-set is structured, http://apps.who.int/gho/data/node.main.1323?lang=en .The data set comprises of five columns , and depicts the relapse cases i.e. of different people across more than 190 countries from a range of year 2000 to 2017 which will be used in analysis .

2.6 Source 6: Git-Hub

The fifth data-set is sourced from Git-hub ,
`https://github.com/lukes/ISO-3166-Countries-with-Regional-Codes/blob/master/all/all.csv` .This data gives us the drill down on country to their respective codes and regions The only reason to add this data in my work was for the purpose of multidimensional drill down and hence maintaining a hierarchy to my country dimension

3 Related Work

In order to conduct my study, I have used the data sets that include statistics on the incidences of tuberculosis, HIV, relapse rates and gender disparity worldwide.
The results I arrived at are in sync with the prior work being done in this domain. Therefore, the following literature depicts the results which were arrived by using the data sets that I have used for my study. I have therefore tried to cover the various themes that cover my study topic.

3.1 Reducing Global Tuberculosis DeathsTime For India To Step Up.

Authors tried to explain through statistical reasons as to why India is the ground for tuberculosis. They outlined various factors that lead to the mortality in India due to tuberculosis, these factors include, malnutrition, smoking, tobacco consumption, implementation failure, under investment, a poor health system, and definitely pathetic tuberculosis care in the private sector. Incidence of tuberculosis was witnessed 20 times higher in homeless people even in industrialized cities. If we see, for India, homelessness, crowded environments and poor hygienic practices are the bane for such infections.
Authors came up with the required solutions as well; they suggested that there should be increase in the public funding, improvement in the quality and testing in approaches, viz, decentralized molecular diagnostics and information communication technology.

The way mortality due to tuberculosis in Wales and England eventually declined and the significant factors foe this decrease could be attributed to improved housing and sanitation, and decreased crowding, same is expected for India a s well with the programmes like Swachh Bharat mission , housing for all by 2022 and development of smart cities along with addition to the health budget for fighting tuberculosis would definitely mitigate the burden of tuberculosis in India Pai M (2017)

3.2 Tuberculosis/Hiv Co-Infection

Author in this article highlights HIV to be the most significant risk factor that triggers the Tuberculosis or activates it in those people who are already with mycobacterium tuberculosis infection. Asia and Africa has seen the more burden of this co infection comparatively.
Author analysed India to be the worlds top ranked country that experiences the highest burden of Tuberculosis as well as HIV infection. Increasing incidence of tuberculosis has been linked to poverty, other significant factors could be under nutrition, smoking, HIV, diabetes, etc. Infants could be at risk for HIV infection, if mother is co infected with tuberculosis and HIV, which would eventually lead to premature birth of an infant or low

body weight or infant mortality and higher maternal mortality.
Author in this article also highlights the 4 Is policy in order to address the co infection of tuberculosis and HIV that includes infection control, ionized preventive chemotherapy, case finding, integration of tuberculosis and HIV services, family planning, and immunization services.
Author concluded by laying the stress on strengthening of health systems so that long term care is considered. Swaminathan (2017)

3.3 A Study On The Relapse Rate Of Tuberculosis And Related Factors In Korea Using Nationwide Tuberculosis Notification Data

In this study researcher attempted to estimate the relapse rate of Tuberculosis and showcase his interest to outline the relevant factors for the same. In order to carry out this study, he used nationwide Tuberculosis notification data in Korea. Data of 2005 was used and it was cross checked as well. Associated factors were analyzed using multivariate logistic regression with the variables, such as age, sex, and registration type, medication, sputum smear test, and outcome of treatment.
The relapse of tuberculosis was found to be more in men, based on chi square and multivariate logistic regression analysis. It was also analyzed that relapse occurred more in patients who were in their 40s, 50s, and 60s, this could possibly be attributed to latent tuberculosis infection. Whereas, there was seen no such relation in the patients who were registered with public health centres, with available treatment results.
Other relevant studies indicate the factors that are responsible for relapse of tuberculosis, which include alcoholism, irregular intake of medicine, smoking, and diseases associated with decline in immune competence and prior history of tuberculosis infection.
This study had a serious limitation as clinical data could be attained, due to which reinfection and recurrence rates, that are the two main categories of relapse, thus were not able to get identified. Hyungmin Lee (2015)

3.4 Tuberculosis Incidence In Prisons: A Systematic Review

In this study, researcher concluded the evidence that was published already, i.e., the incidence inclusive of LTBI as well as Tuberculosis in Prisons. It was reported that globally speaking, transmission of Tuberculosis is much higher in prisons as compared to general population.
For this study, researcher identified 23 Studies that depicted the incidence of TB/LTBI among prisoners as well as the staff in the prison. Incidence of TB in relevant population was estimated using WHO data, whereas, incidence of LTBI was estimated from the studies themselves. After this, researcher calculated the ratio between both the incidences for general population in the prison. IRR varied for both the studies, for LTBI, it was 26.4, that means for LTBI in prisons, average incidence was 26.4 times higher in the general population. Whereas, IRR for Tuberculosis was 23.
Thus, this study review confirms using peer reviewed data belonging to both Low/High income countries, that risk to suffer from tuberculosis is higher in prison then in general population.
There is a scope for future study that could include studying the impact of conditions in prisons on TB transmission and the population attributable risk of prison to community spread. Iacopo Baussano 1 (2011)

3.5 Gender Differences In Tuberculosis Notification In Pakistan

Author in this study attempted to show that in Pakistan, rates of new smear positive TB notification are higher significantly in young women who are less than 45 years of age than males in comparison to India and also it was reported that these rates do not skew significantly to males in older age, the way they do in India.

In order to calculate the age and sex specific rates of TB notification for the year 2008 for India Pakistan, data from WHO global TB control 2009 report over 29 year span(1980-2008) for over 200 countries territories was used, along with population estimates from US Bureau International database were used. In order to calculate the age specific differences in proportion of female cases, chi square and odds ratio was being used. Whereas, Indian population was used as a reference group with 95 percent confidence interval.

It was estimated that the notification rates are approximately equal in males and females, which is not true for age specific notification rates, wherein it is higher in females as compared to males who were aged less than 45 years. In Pakistan, female notification outnumbers the females in India. Excess of notification in India was seen in the females less than 15 years of age. Thus the proportion of notified female cases in India is significantly lower than Pakistan for every age group.

Researcher concluded that improvement in living standards and public health could effectively control TB in the long run. While as in the short run, it is important to understand the key factors that contribute to the infection and the disease in the woman that would eventually be of great importance to TB control cOmara F Dogar & Qadeer3 (2017*b*)

Research Gap :
Lot of work has already been done in this domain, but this study provides a birds view to the researcher for future study, as it is a comparative analysis of the continents, and it also attempted to study the other possible dimensions. Since it is based on the trend analysis, it is convenient to understand the issue, interpret the results, and come up with the desired recommendations.
Therefore, this study provides a better understanding of the topic and has a future scope of further research. My analysis mainly puts forward the incidence and mortality comparison on global level .So it could be used in future to deal with better analysis on tuberculosis incidence case studies related to HIV patients and on gender differentiation analysis.

4 Data Model

In my project I made use of Ralph Kimball bottom up approach, according to Kimballs bottom up approach the data model can either consist of one fact table or more than on fact table with its respective dimension tables attached to it . In the dimensional table their exists information and primary key which is inserted and used by fact table as foreign keys . Also the fact table contains all the measures.

A star schema was made in this project out of six dimensional tables . All dimensional tables consists of primary key used by the fact table as a foreign key. First of all , the raw data was imported directly to SSMS directly using R after the transformation and

cleaning . Then attribute from the raw table were distinguished and converted to dimensional table. After this primary key was created for each dimensional table to uniquely identify each row. Incidence_id, Mortality_id , Country_id , Relapse_id, statistaa_id are all primary keys . Followed by creating all facts I joined all the dimensional tables and created a fact table containing all the measures and foreign keys. Below are all the dimensional tables which follow a star schema,

a) Dim_Country : This dimension will be a role playing dimension as it is used in nearly all the queries formed for analysis of the data. Table contains four columns namely country name, region ,sub region and country id which here is a primary key. This dimensional table also provides a multidimensional drill down and hence give a hierarchy to analysis of data .

	Country_ID	Country_Name	Region	Sub_Region
1	4	Afghanistan	Asia	Southern Asia
2	8	Albania	Europe	Southern Europe
3	12	Algeria	Africa	Northern Africa
4	16	American Samoa	Oceania	Polynesia

Figure 1: Dim Country

b) Dim_Incidence : This dimension consists of seven columns which comprise information related to incidence of tuberculosis and people who suffer TB having HIV . I have defined incidence_id as an identity int which is here primary key for this dimension. It consists of data on global level for about more than 190 countries. It was joined with Dim_country as country in common for country_id to uniquely identify each country for different years .

	Country_ID	Country_Name	Region	Sub_Region
1	4	Afghanistan	Asia	Southern Asia
2	8	Albania	Europe	Southern Europe
3	12	Algeria	Africa	Northern Africa
4	16	American Samoa	Oceania	Polynesia

Figure 2: Dim Incidence

c) Dim_Mortality : This dimension consists of five columns .In this dimension mortality rates are recorded and also for people who underwent HIV test but were diagnosed negative . Here mortality_id is the primary key which is an identity int to uniquely identify the entire table . The raw table was also joined with Dim_country as country in common for country_id to form this dimension.

	mortality_id	country_id	year	nod_exhiv	nod_hivnegpop
1	10001	4	2016	11000	33
2	10002	4	2015	13000	38
3	10003	4	2014	14000	42
4	10004	4	2013	13000	43

Figure 3: Dim Mortality

d) Dim_Relapse : The composition of this dimension is also the same I have joined its raw table with Dim_country to uniquely identify for each country and also identity int is formed as relapse_id which is the primary key here . The table comprises of medical cover and cases of number of patients that relapsed.

	relapse_id	country_id	year	tb_relapse	tbinsurance
1	10001	999	9999	0	0
2	10002	4	2016	41954	64
3	10003	4	2015	35878	56
4	10004	4	2014	31746	51
5	10005	4	2013	30507	51

Figure 4: Dim Relapse

e) Dim_statistaa :This dimension contains 4 columns after composition which comprises of TB incidence in Europe in year 2016 . Here, statista_id was set as primary key in this case which uniquely defines each row. Also it was joined with Dim_country to obtain country_id column.

	statistaa_id	country_id	Country	Incidence
1	10	999	Dummy	0
2	11	40	Austria	634
3	12	56	Belgium	1047
4	13	100	Bulgaria	1603

Figure 5: Dim Statista

.

f)Dim_Gender_cases : This dimension is a gender differentiation table which comprises of six columns with men , women and children incidence on tuberculosis and their death . Here I have set gender_case_id as primary key which is defined as an identity.

	gender_case_ID	year	cases	women	men	children
1	101	9999	Dummy	0	0	0
2	102	2011	Incident Cases 2011	290000	530000	500000
3	103	2012	Incident Cases 2012	290000	517000	530000
4	104	2013	Incident Cases 2013	330000	515000	550000

Figure 6: Dim gender cases

g) Fact table : All the dimensional table are being related to each other using the fact table where all the primary key from every dimensional table are present which are referenced as foreign keys in this table. Foreign Key is used here as a reference to create a Join between Fact and Dimension Table. This table stored the most granulated information of my data analysis . Measures from all the dimension tables are fetched in this table which will be used in the analysis of data .

	incidence_id	mortality_id	relapse_id	statista_id	gender_case_ID	country_id	tbcases	tbper_pop	tbcases_hiv	tbper_pophiv	tb_relapse	tbinsurance	nod_exhiv	nod_hivnegpop	incidenc
1	10001	10001	10002	10	101	4	65000	189	280	1	41954	64	11000	33	0
2	10002	10002	10003	10	101	4	64000	189	270	1	35878	56	13000	38	0
3	10003	10003	10004	10	101	4	62000	189	260	1	31746	51	14000	42	0
4	10004	10004	10005	10	104	4	60000	189	240	1	30507	51	13000	43	0

Figure 7: Fact table

The measures present in the fact table in combination with the attributes present in the dimension is used for the following analysis :

Is there significant regional disparity of the global incidence and mortality rates for tuberculosis at a Global scale ?

What world regions records the highest incidences of tuberculosis inclusive of HIV ?

And what are the death/mortality rates of tuberculosis exclusive of HIV across global regions?

A comparison of gender differentiation based on Mortality rates ? Compared with the cases of incidence of TB based on Ranking .

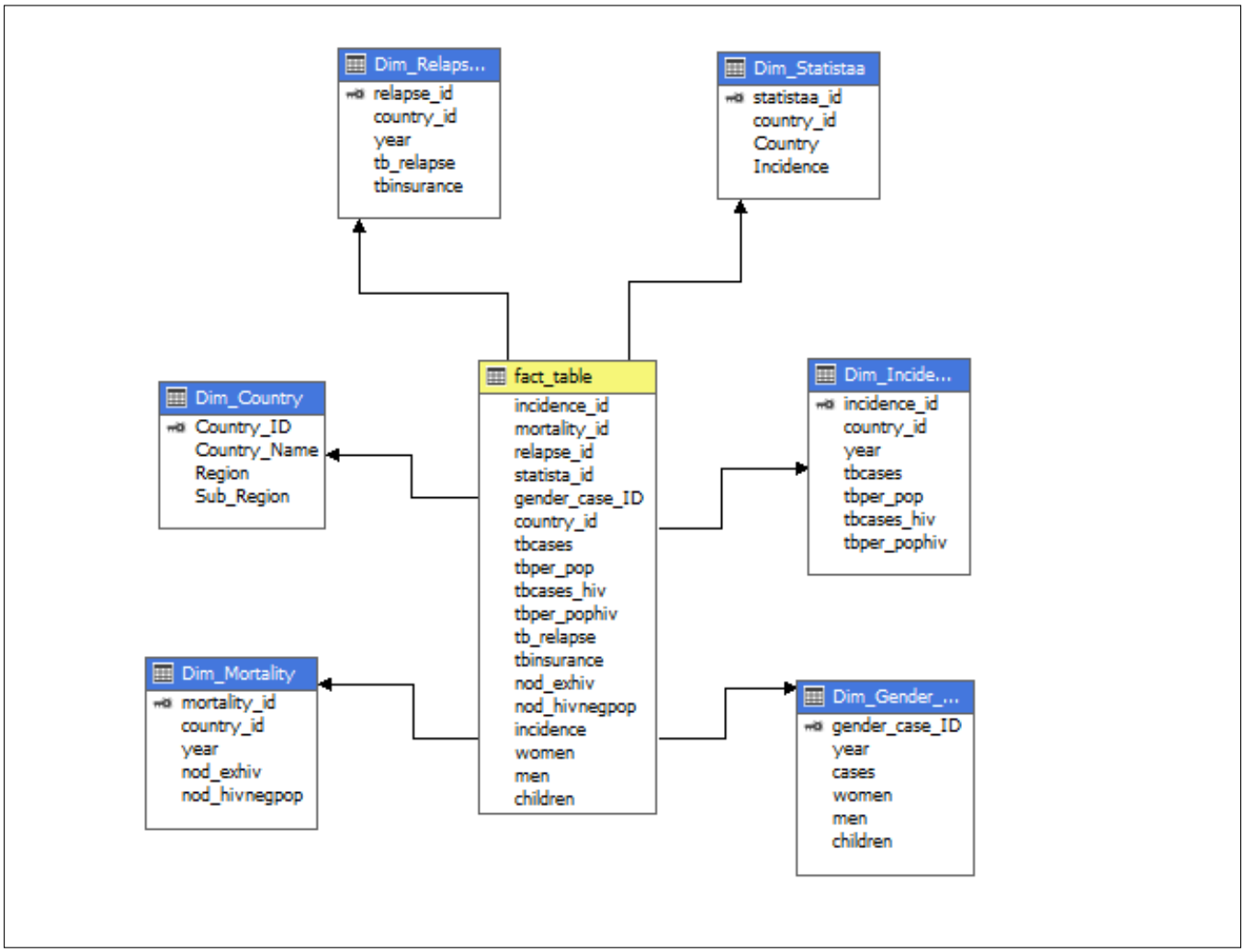


Figure 8: Star Schema

5 Logical Data Map

This section, represents logical data map, i.e. how every row of every data source is handled such that it is a part of my star schema.

Source	Column	Destination	Column	Type	Transformation
1	tbcases	Fact_Table	tbcases	Fact	In this column unwanted text inside braces were removed using G-sub function in R .
1	tbcases_hiv	Fact_Table	tbcases_hiv	Fact	First of all unwanted text inside braces were removed using G-sub function in R. Column had 'no-values' which had to be changed as type numeric to generate Null values which were replaced to 0 $
1	tbper_pophiv	Fact_Table	tbper_pophiv	Fact	First of all unwanted text inside braces were removed using G-sub function in R. Column had 'no-values' which had to be changed as type numeric to generate Null values which were replaced to 0
1	Year	Dim_Incidence	year	Dimention	the coloumn was transformed to type date$
2	nod_exhiv	Fact table	nod_exhiv	Fact	Unwanted text inside braces were removed using G-sub function in R. Column had 'no-values' which had to be changed as type numeric to generate Null values which were replaced to 0
2	year	Dim_Mortality	year	Dimention	Entire columnn was changed to type date$
2	nod_hivnegpop	Fact Table	nod_hivnegpop	Fact	Here the unwanted number from braces were removed using G-sub in R and than 'no-values' space in columns were replaced by 0 by changing the type to numeric to generate Null
5	tb_relapse	Fact_Table	tb_relapse	Fact	Unwanted text inside braces were removed using G-sub function in R. Column had 'no-values' which had to be changed as type numeric to generate Null values which were replaced to 0

Continued on next page

Table 3 – *Continued from previous page*

Source	Column	Destination	Column	Type	Transformation
5	tbinsurance	Fact_Table	tbinsurance	Fact	Here the unwanted number from braces were removed using G-sub in R and than 'no-values' space in columns were replaced by 0 by changing the type to numeric to generate Null
6	Country_name	Dim_Country	Country_Name	Dimension	In this coloumn null values were checked and simply removed by adding NULL in that row(row deleted) $
6	Region	Dim_Country	Region	Dimension	Here null values were present inside the column which were deleted in R
6	Sub_Region	Dim_Country	Sub_Region	Dimension	The row containing null values were simply deleted in R$
4	Cases	Dim_Gender	Cases	Dimension	This column was scrapped off a pdf using R which went under different operations like rbind , cbind and string splitter. Once the column was attained in a tabular after that the column name were changed .
4	women	Fact_table	women	Fact	This column was scrapped off a pdf using R which went under different operations like rbind , cbind and string splitter. Once the column was attained in a tabular form it had unwanted operators which were removed using g-sub and also the values were changed from 1m to 1,000,000 .
4	men	Fact_table	men	Fact	This column was scrapped off a pdf using R which went under different operations like rbind , cbind and string splitter. Once the column was attained in a tabular form it had unwanted operators which were removed using g-sub and also the values were changed from 1m to 1,000,000 .

Continued on next page

Table 3 – *Continued from previous page*

Source	**Column**	**Destination**	**Column**	**Type**	**Transformation**
4	children	Fact_Table	children	Fact	This column was scrapped off a pdf using R which went under different operations like rbind , cbind and string splitter. Once the column was attained in a tabular form it had unwanted operators which were removed using g-sub and also the values were changed from 1m to 1,000,000
4	year	Dim_Gender	year	Fact	This column was made using string split function on the column 'cases' from the same table to obtain a year column
3	Incidence	Fact_table	Incidence	Fact	Here no such transformation was performed except the first two columns were removed in R and column name was given
3	Country	Dim_Statista	country	Dimension	In this column the top two columns were removed and the column name was given

6 ETL Process

ETL stands for extraction transform and load . In this process how the data has been extracted from its source , cleaned and then the process of loading it into the Data Warehouse is explained.

Extraction

The source of data that can used to load into the Data Warehouse can be of any type . It can be in a structured , semi-structured or unstructured form . For structured data might be in a .csv extension , JSON , a XML file or it might generated from an API . For unstructured the data might be web crapped or a tabular data might be scrapped off a .pdf .

In this project the Structured data was extracted in a .csv extension from 2 sources . And for unstructured data a tabular data was scrapped off which was embedded in text from a report which was in a PDF format.

The first structured dataset is the Incidence data which was downloaded as a CSV file from WHO ,Incidence by country section of Tuberculosis . It consisted of six columns from which I removed the unwanted columns in MS-Excel . After that the table was loaded in R where operation was performed on columns . For all the numeric columns which I could use for measures I removed all the unwanted numeric numbers covered in braces using G-sub function . At some places no-value was present to take care of this, columns were transformed to type numeric to generate null values and then the no-values value was replaced with 0 using R . Finally , the cleaned data set was directly transferred to SSMS using RODBC library to the staging area as Stg_incidence.

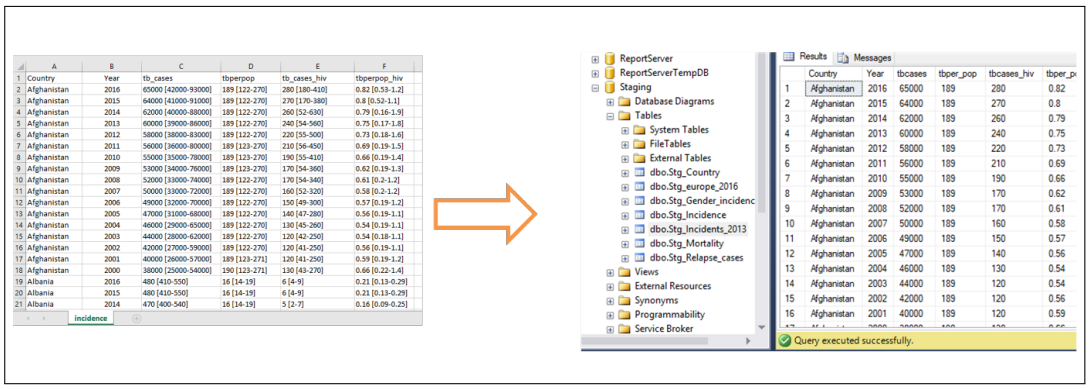
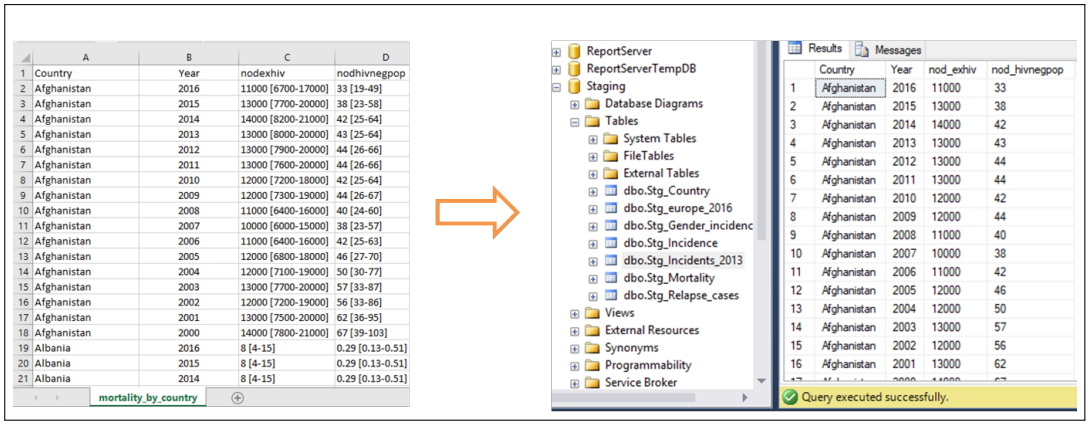


Figure 9: Stg Incidence

The second structured data is the mortality data which was downloaded as a CSV file from the source WHO . It consisted of 4 columns . The table was opened in R where operations were performed to clean the data. First of all the unwanted numeric numbers covered in braces were removed using G-sub function . Some places no-values character were present. To remove it I changed the column to numeric and null values were generated which were replaced by 0 . Finally , the cleaned dataset was directly transferred to SSMS using RODBC library to the staging area as Stg_mortality.



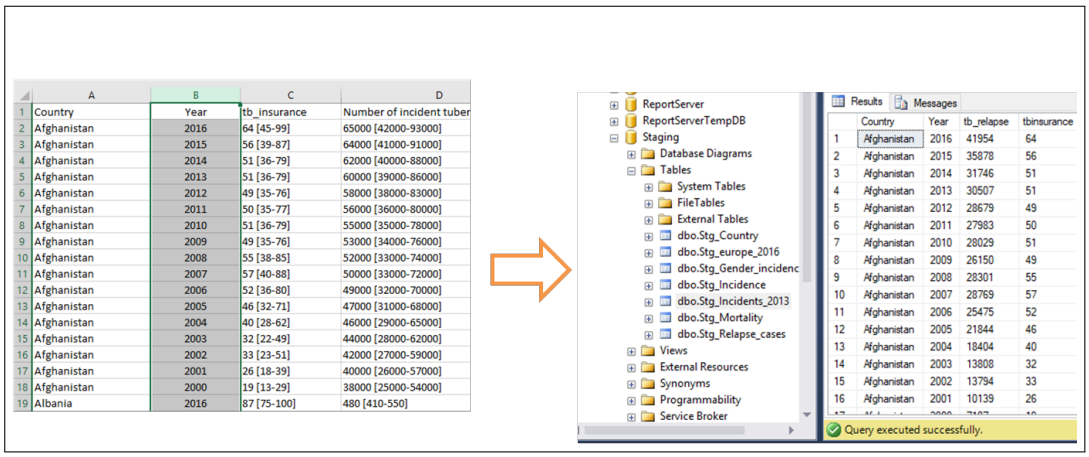
	A	B	C	D
1	Country	Year	nodexhiv	nodhivnegpop
2	Afghanistan	2016	11000 [6700-17000]	33 [19-49]
3	Afghanistan	2015	13000 [7700-20000]	38 [23-58]
4	Afghanistan	2014	14000 [8200-21000]	42 [25-64]
5	Afghanistan	2013	13000 [8000-20000]	43 [25-64]
6	Afghanistan	2012	13000 [7900-20000]	44 [26-66]
7	Afghanistan	2011	13000 [7600-20000]	44 [26-66]
8	Afghanistan	2010	12000 [7200-18000]	42 [25-64]
9	Afghanistan	2009	12000 [7300-19000]	44 [26-67]
10	Afghanistan	2008	11000 [6400-16000]	40 [24-60]
11	Afghanistan	2007	10000 [6000-15000]	38 [23-57]
12	Afghanistan	2006	11000 [6400-16000]	42 [25-63]
13	Afghanistan	2005	12000 [6800-18000]	46 [27-70]
14	Afghanistan	2004	12000 [7100-19000]	50 [30-77]
15	Afghanistan	2003	13000 [7700-20000]	57 [33-87]
16	Afghanistan	2002	12000 [7200-19000]	56 [33-86]
17	Afghanistan	2001	13000 [7500-20000]	62 [36-95]
18	Afghanistan	2000	14000 [7800-21000]	67 [39-103]
19	Albania	2016	8 [4-15]	0.29 [0.13-0.51]
20	Albania	2015	8 [4-15]	0.29 [0.13-0.51]
21	Albania	2014	8 [4-15]	0.29 [0.13-0.51]

mortality_by_country

	Country	Year	nod_exhiv	nod_hivnegpop
1	Afghanistan	2016	11000	33
2	Afghanistan	2015	13000	38
3	Afghanistan	2014	14000	42
4	Afghanistan	2013	13000	43
5	Afghanistan	2012	13000	44
6	Afghanistan	2011	13000	44
7	Afghanistan	2010	12000	42
8	Afghanistan	2009	12000	44
9	Afghanistan	2008	11000	40
10	Afghanistan	2007	10000	38
11	Afghanistan	2006	11000	42
12	Afghanistan	2005	12000	46
13	Afghanistan	2004	12000	50
14	Afghanistan	2003	13000	57
15	Afghanistan	2002	12000	56
16	Afghanistan	2001	13000	62

Figure 10: Stg Mortality

The third structured data is the relapse data downloaded in CSV format from the source WHO. It consists of five columns , the unwanted columns were removed using MS-Excel .One column had unwanted numeric values present in braces which were removed using R . Also it had the same no-vale character which was changed to 0 by first changing the column as numeric and then replacing the generated null vales by zero. Finally, the cleaned dataset was directly transferred to SSMS in the staging area as Stg_Relapse_cases.



	A	B	C	D
1	Country	Year	tb_insurance	Number of incident tuber
2	Afghanistan	2016	64 [45-99]	65000 [42000-93000]
3	Afghanistan	2015	56 [39-87]	64000 [41000-91000]
4	Afghanistan	2014	51 [36-79]	62000 [40000-88000]
5	Afghanistan	2013	51 [36-79]	60000 [39000-86000]
6	Afghanistan	2012	49 [35-76]	58000 [38000-83000]
7	Afghanistan	2011	50 [35-77]	56000 [36000-80000]
8	Afghanistan	2010	51 [36-79]	55000 [35000-78000]
9	Afghanistan	2009	49 [35-76]	53000 [34000-76000]
10	Afghanistan	2008	55 [38-85]	52000 [33000-74000]
11	Afghanistan	2007	57 [40-88]	50000 [33000-72000]
12	Afghanistan	2006	52 [36-80]	49000 [32000-70000]
13	Afghanistan	2005	46 [32-71]	47000 [31000-68000]
14	Afghanistan	2004	40 [28-62]	46000 [29000-65000]
15	Afghanistan	2003	32 [22-49]	44000 [28000-62000]
16	Afghanistan	2002	33 [23-51]	42000 [27000-59000]
17	Afghanistan	2001	26 [18-39]	40000 [26000-57000]
18	Afghanistan	2000	19 [13-29]	38000 [25000-54000]
19	Albania	2016	87 [75-100]	480 [410-550]

	Country	Year	tb_relapse	tbinsurance
1	Afghanistan	2016	41954	64
2	Afghanistan	2015	35878	56
3	Afghanistan	2014	31746	51
4	Afghanistan	2013	30507	51
5	Afghanistan	2012	28679	49
6	Afghanistan	2011	27983	50
7	Afghanistan	2010	28029	51
8	Afghanistan	2009	26150	49
9	Afghanistan	2008	28301	55
10	Afghanistan	2007	28769	57
11	Afghanistan	2006	25475	52
12	Afghanistan	2005	21844	46
13	Afghanistan	2004	18404	40
14	Afghanistan	2003	13808	32
15	Afghanistan	2002	13794	33
16	Afghanistan	2001	10139	26

Figure 11: Dim Relapse

The fourth structured data is the European incidence In year 2016 from the source Statista. It consisted of 2 rows . The file was downloaded in the xml format and then opened R where the unwanted first two rows were deleted and also the correct column names were given using R . The cleaned data was then transferred to SSMS directly in the staging area under the name Stg_Europe_2016.

The fifth structured dataset is the country sub-region dataset sourced from UNO . The dataset comprised of some unwanted columns which were removed using R . The region

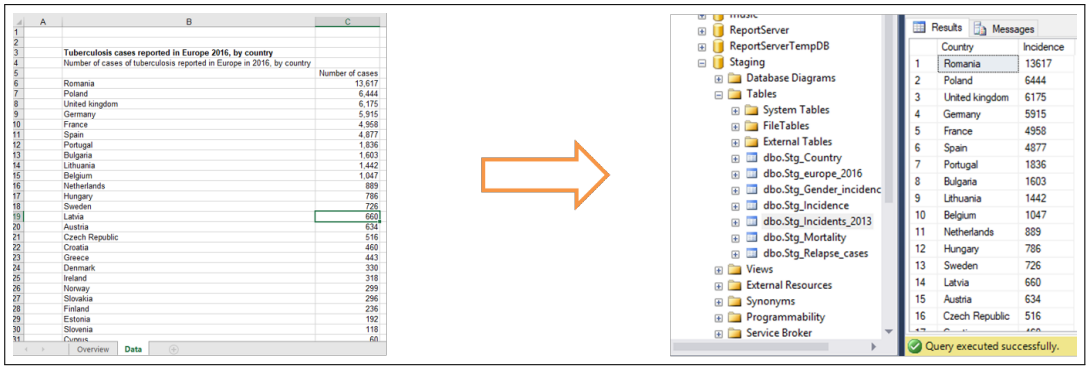


Figure 12: Stg Europe

and sub region columns consisted of NA values which were removed and alternation in the column names were made. This cleaned data was then transferred to SSMS directly from R in the staging area named as Stg_country.

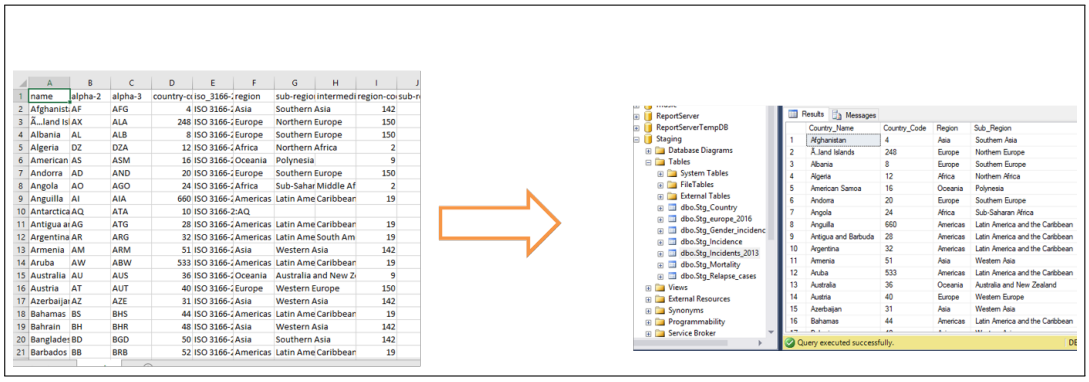


Figure 13: Stg Country

The last data was my unstructured dataset which is taken from a UNDP report which was scrapped from a PDF using R . The tabular data was inscribed in text . While scrapping , all the contents of data had to undergo different R operations to attain a tabular form using R functions like cbind , splitstackshape , tabulizer and tidyverse . After that desired column names were given to table . After attain a tabular structure the data cleaning like removing of star mark . In column, women, men and children I changed the values like 1m was rounded to 1,000,000 by removing m from the column and multiplying the column by 1,000,000. Also unnecessary sign operators were removed using G-sub. Also the unwanted columns were removed .The cleaned data was than transferred to SSMS in the staging area under name Stg_gender_incidence_mortality.

Transformation and loading

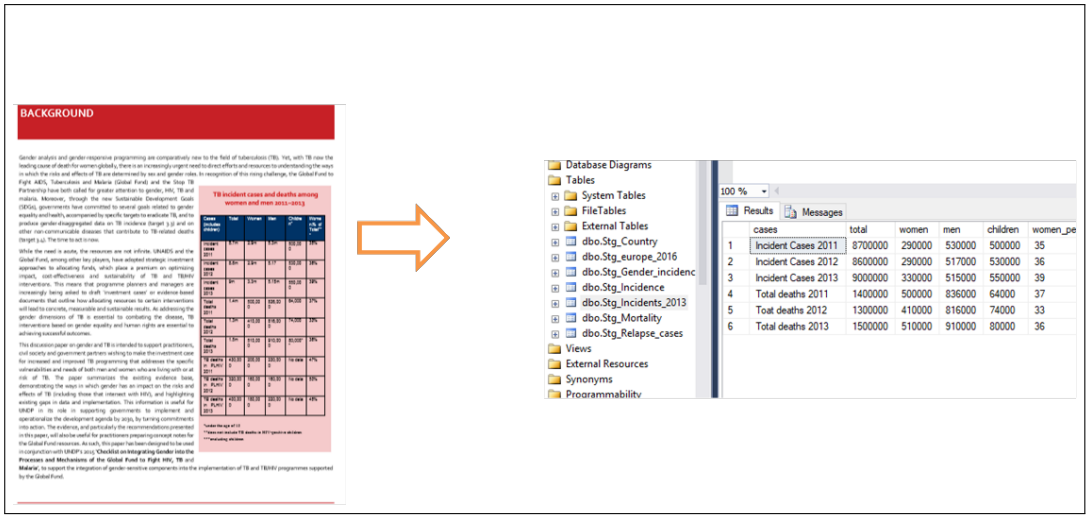


Figure 14: Stg Gender

In my project once all the data was extracted from its source and cleaned it was directly loaded inside the SQL server 2016 database through R under the database name Stage . This is used for the staging area and the database for Dimensional table for dimensions and facts.

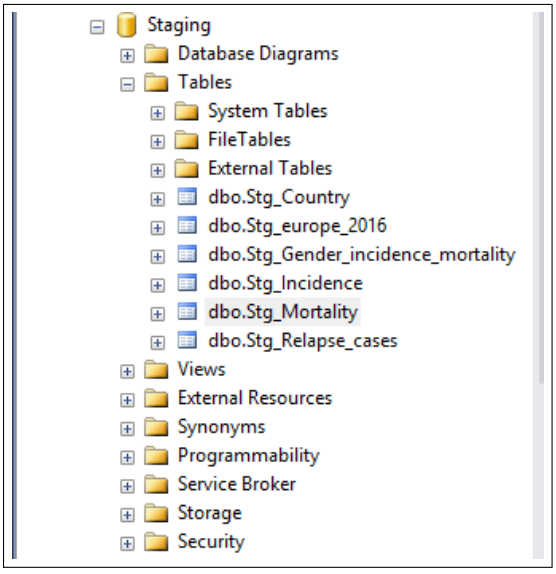


Figure 15: Staging Database

Each of the above mention structured data behaved as raw data shown in Fig above. After all my raw data is loaded inside SSMS , dimensions which were required for this data model such as DimIncidence , DimMortality ,DimRelapse cases , DimStatistaa were created using SSMS . It was done by CREATE TABLE query ,creating new table and inserting values for it and giving different joins to it manually . For each of these dimensional table primary key was defined. After creating these dimensions in SSMS the insert queries were then inserted into SQL task feature in SSIS

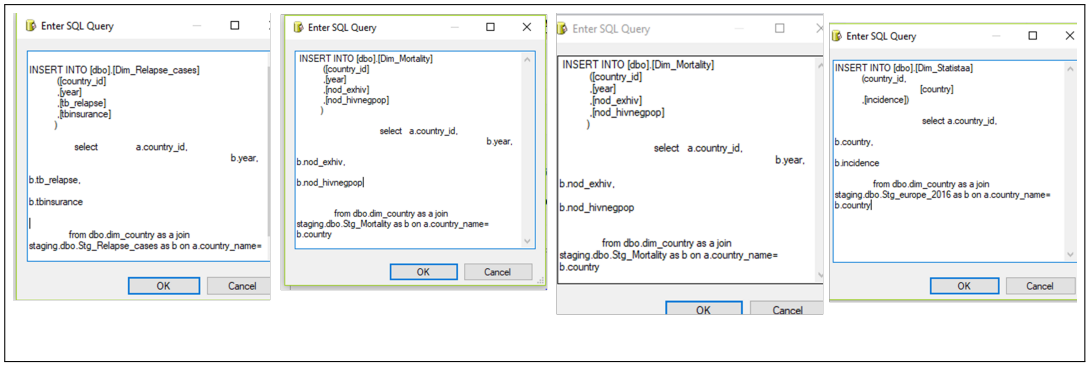


Figure 16: Insert query for dimension in execute task

For Dim gender and Dim Country , I have created their dimensions in SSIS using the execute SQL task where they were automatically loaded inside database Dimension table . Here sort transformation followed by slowly changing dimension was used to create these two dimensions

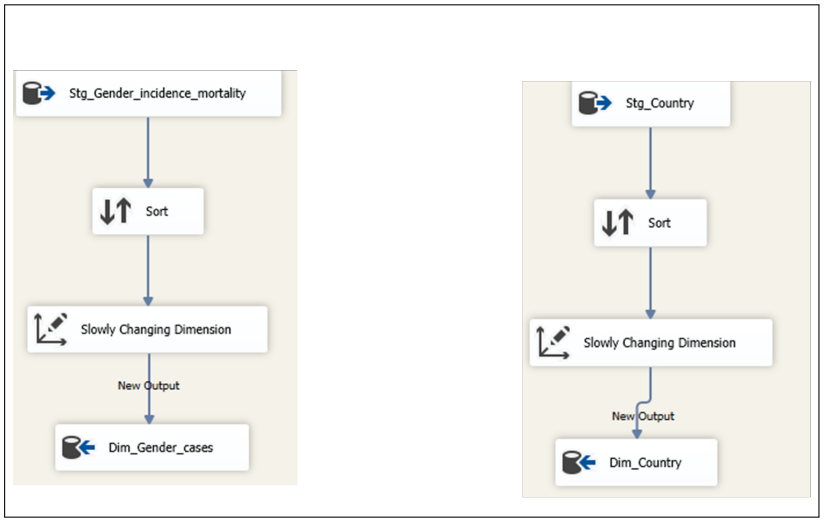
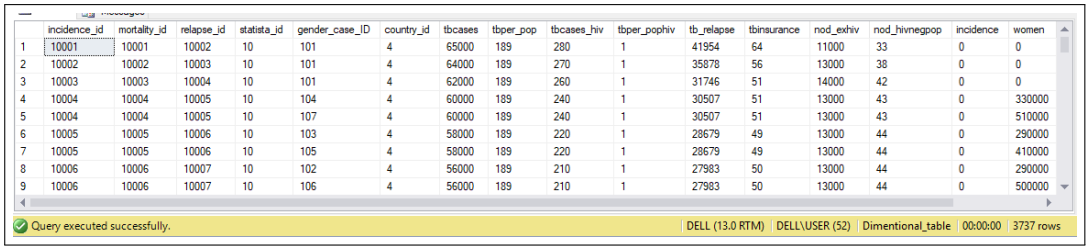


Figure 17: Slowly changing dimension

Finally after the creation and loading of all the dimensions inside the Dimension table , a fact table was created in SSMS using CREATE TABLE query (query attached in appendix). After this to load the table INSERT query was written where all the tables from dimensional table and raw data table have been merged and linked by using different SQL joins . Also foreign key references were specified for each primary key to represent all the dimensions. For reloading the data into the dimensional table , a SQL task feature was pulled in SSIS under name Truncate table where insert query was written .Truncate is used to empty the contents present in fact table.

After the fact table is successfully loaded with all the measures and foreign keys, the cube for dimensional table was successfully deployed in SSAS. This all process of pulling



	incidence_id	mortality_id	relapse_id	statista_id	gender_case_ID	country_id	tbcases	tbper_pop	tbcases_hiv	tbper_pophiv	tb_relapse	tbinsurance	nod_exhiv	nod_hivnegpop	incidence	women
1	10001	10001	10002	10	101	4	65000	189	280	1	41954	64	11000	33	0	0
2	10002	10002	10003	10	101	4	64000	189	270	1	35878	56	13000	38	0	0
3	10003	10003	10004	10	101	4	62000	189	260	1	31746	51	14000	42	0	0
4	10004	10004	10005	10	104	4	60000	189	240	1	30507	51	13000	43	0	330000
5	10004	10004	10005	10	107	4	60000	189	240	1	30507	51	13000	43	0	510000
6	10005	10005	10006	10	103	4	58000	189	220	1	28679	49	13000	44	0	290000
7	10005	10005	10006	10	105	4	58000	189	220	1	28679	49	13000	44	0	410000
8	10006	10006	10007	10	102	4	56000	189	210	1	27983	50	13000	44	0	290000
9	10006	10006	10007	10	106	4	56000	189	210	1	27983	50	13000	44	0	500000

Figure 18: Fact Table

data from staging database and loading them in dimension database and fact table and finally the deployment of the cube was all automated .

OLAP cube :

Overview of the cube in SQL server Analysis service.

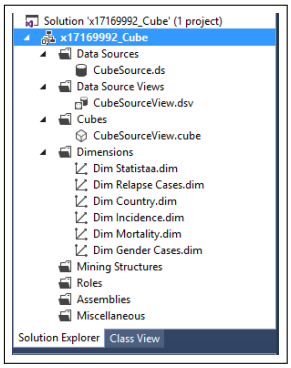


Figure 19: OLAP Cube

Deployment of the cube

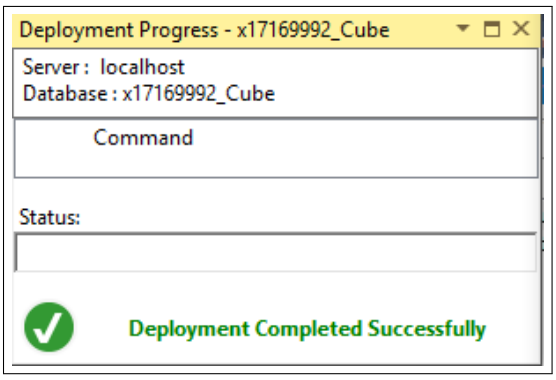


Figure 20: Cube Deployed Successfully

Overview of the star schema of OLAP cube

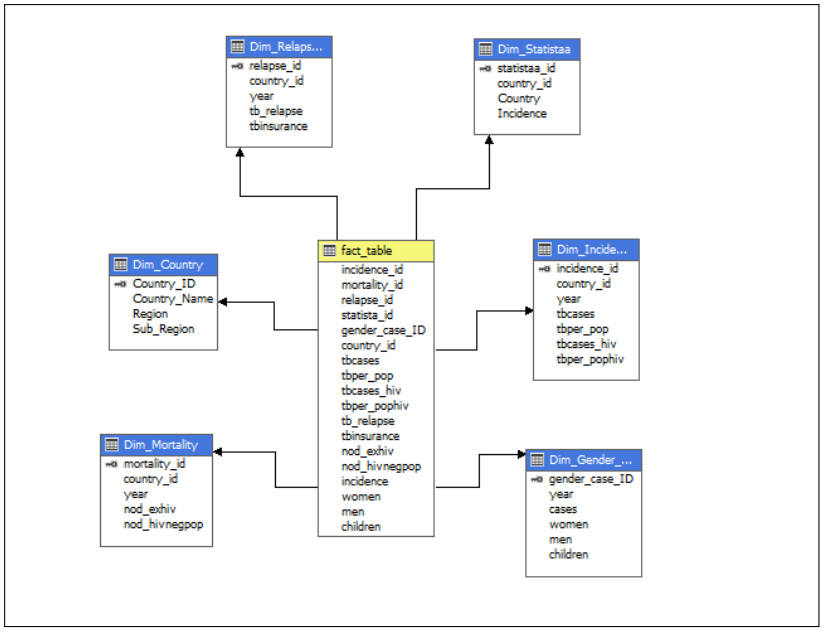


Figure 21: Start Schema

7 Application

7.1 BI Query 1: Is there significant regional disparity of the global incidence and mortality rates for tuberculosis at a Global scale ?

In order to identify and compare which country has the highest incidence for tuberculosis and compare it with the mortality rates at a Global level data has been taken from two structured source of data . From which the tuberculosis cases and mortality cases of more than 190 countries have been taken across a time period from 2000 to 2017 from World health Organization .
From the given map visualization tool using tableau it was demonstrated that India has been recorded as the country with the highest rate of Incidence as well as for Mortality with incidence recorded as 60,670,000 and 10,153,000 for mortality . Where as China stands second with incidence of .22,223,000. and Indonesia with mortality of 24,910,000.

7.2 BI Query 2: What world regions records the highest incidences of tuberculosis inclusive of HIV ? And what are the death/mortality rates of tuberculosis exclusive of HIV across global regions?

-A REGIONAL , SUB-REGIONAL AND COUNTRY WISE ANALYSIS
For my Second query to identify and compare which global region has the highest number of death due to tuberculosis excluding HIV compared to incidences of tuberculosis with HIV , I have taken data from two structured sources .One data gives my query a

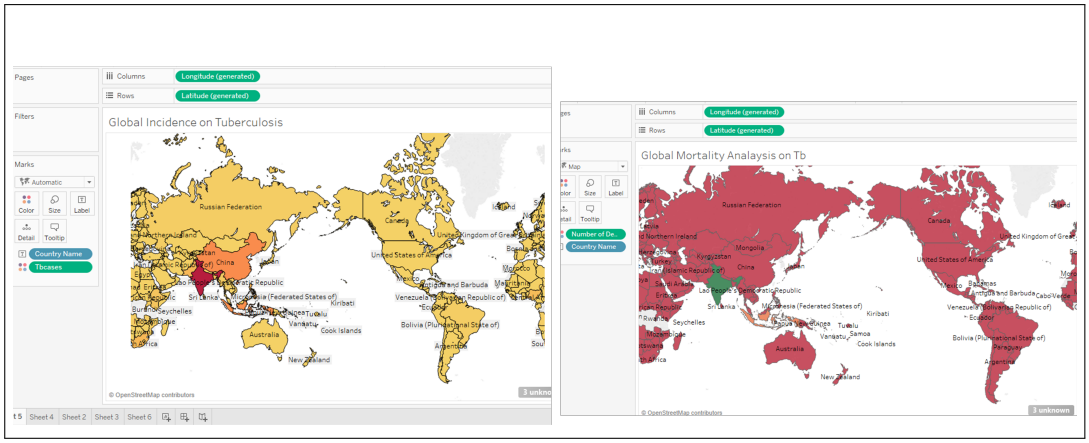


Figure 22: Case Study 1

hierarchical outlook of global region ,sub region and country wise analysis and the second source represents incidences and mortality cases from 2000 to 2017 at a global scale.
From the given horizontal bars visualized with the help of tableau , it is clearly evident that Asia records the highest number of death for people excluding HIV which is 19,284,910 followed by Africa with 6,860,463 where as when it comes to incidences of TB for HIV positive Africa records the highest incidences of 16,876,219 followed by Asia .

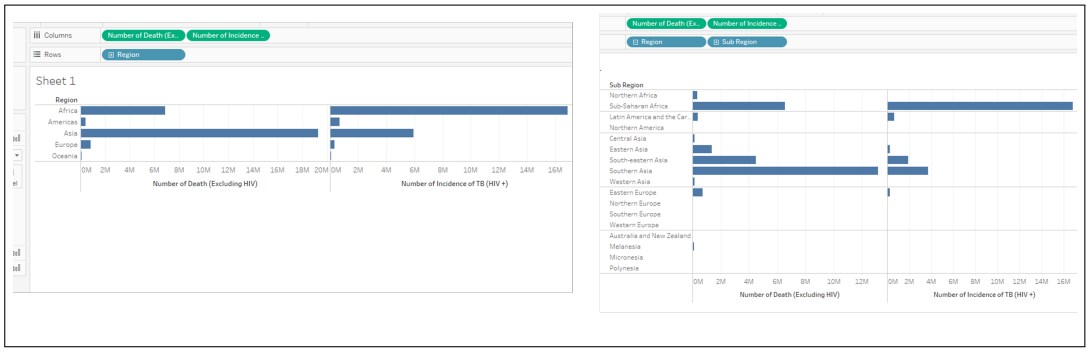


Figure 23: Case Study 2

7.3 BI Query 3: A comparison of gender differentiation based on Mortality rates ? Compared with the cases of incidence of TB based on Ranking .

For my final query I have used two sources of data one is unstructured and compared it with the structured data. All the gender data was sourced from UNDP that illustrates the data over the year 2011 to 2013 for male , female and children.
To study the discrepancies between the gender on mortality rates and rank them accordingly, we used the horizontal bar representation available in tableau. Data was taken for the year 2011, 2012, 2013. We compared the number of incidences of tuberculosis and the gender wise death rates in this time period. The Graph illustrates an unpredictable

analysis of death rate in gender compared to their incidence of TB in the year 2013. It depicts that when death rate was recorded as rank 1 (i.e , its highest value) the incidence of Tb was recorded as lowest with rank 3.

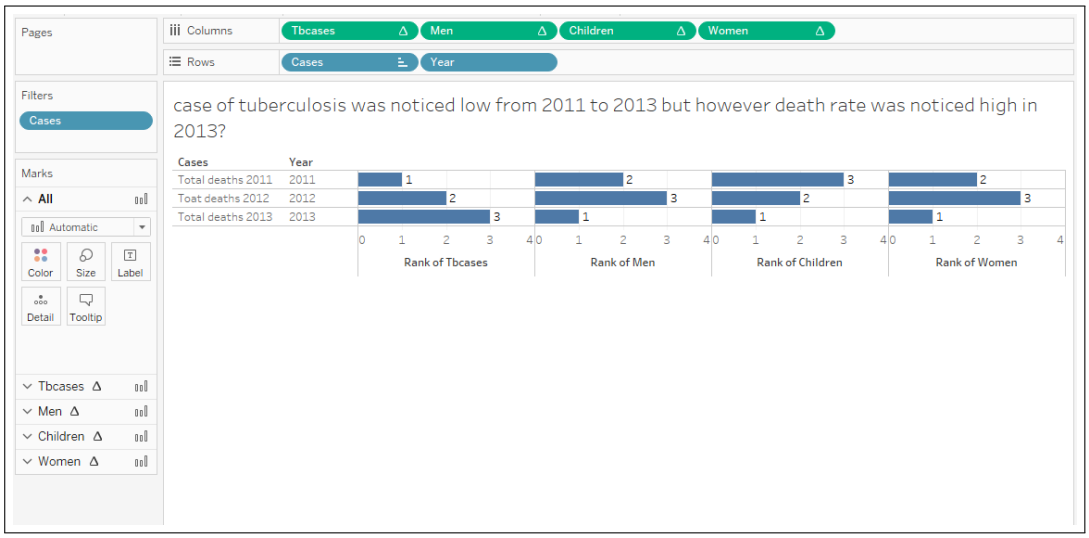


Figure 24: Case Study 3

7.4 Discussion

The first bi query attempts to identify and compare which country has the highest incidence for tuberculosis and compare it with the mortality rates at a Global level, As already mentioned above, the given map demonstrated that India has been recorded as the country with the highest rate of Incidence as well as for Mortality. Since India is a developing nation, it is very much prone to all such menaces. Pai M, Correa N, Mistry N, Jha P in their paper Reducing Global Tuberculosis DeathsTime For India To Step Up has also explained as to why India is the ground for tuberculosis. They outlined various factors that lead to the mortality in India due to tuberculosis, these factors include, malnutrition, smoking, tobacco consumption, implementation failure, underinvestment, a poor health system, and definitely pathetic tuberculosis care in the private sector. S. Swaminathan in his paper Tuberculosis/Hiv Co-Infection also analyzed India to be the worlds top ranked country that experiences the highest burden of Tuberculosis. According to him, in India, Increasing incidence of tuberculosis has been linked to poverty, other significant factors could be under nutrition, smoking, HIV, diabetes, etc.

The Second query attempts to identify and compare as which global region has the highest number of deaths due to tuberculosis excluding HIV as compared to incidences of tuberculosis with HIV. It was clearly evident from the given horizontal bars of tableau, that Asia records the highest number of deaths for the people excluding HIV, followed by Africa. Whereas when it comes to incidences of TB for HIV positive, Africa records the highest incidences followed by Asia. S. Swaminathan in his paper Tuberculosis/Hiv Co-Infection also highlights HIV to be the most significant risk factor that triggers the Tuberculosis or activates it in those people who are already with mycobacterium tuberculosis

infection. Asia and Africa has seen the more burden of this co infection comparatively. He analyzed India to be the worlds top ranked country that experiences the highest burden of Tuberculosis as well as HIV infection.

The Third query compares the gender differentiation based on Mortality rates which are compared with the cases of incidence of TB based on Ranking. To study the discrepancies between the gender on mortality rates and rank them accordingly, we used the horizontal bar representation available in tableau. The Graph above illustrates an unpredictable analysis of death rate in gender compared to their incidence of TB in the year 2013. Andrew J. Codlin ,* Saira Khowaja , Zhongxue Chen , Mohammad H. Rahbar , Ejaz Qadeer , Ismat Ara , Joseph B. McCormick ,Susan P. Fisher-Hoch , and Aamir J. Khan in their article Gender Differences In Tuberculosis Notification In Pakistan, estimated that the notification rates are approximately equal in males and females, which is not true for age specific notification rates, wherein it is higher in females as compared to males who were aged less than 45 years. In Pakistan, female notification outnumbers the females in India. Excess of notification in India was seen in the females less than 15 years of age. Thus the proportion of notified female cases in India is significantly lower than Pakistan for every age group. Whereas, Omara F Dogar, Sarwat K Shah, Abrar A Chughtai, and Ejaz Qadeer, in ther study titled Gender Disparity In Tuberculosis Cases In Eastern And Western Provinces Of Pakistan, concluded that sex specific differences were seen higher in females in western provinces of Pakistan as compared to eastern provinces. Countries like Norway, Denmark, England, and Wales depicted the similar trends. There was a common thread in all the mentioned countries along with Pakistan; all were subjected to disaster either natural or manmade those lead to mass displacement of population. There are couple of things that western province has faced, it has firstly accommodated externally displaced people due to entry of refugees from Afghanistan, also it has witnessed mass movement of internally displaced people due to earthquakes, conflicts or floods. These inferences depict that females are specifically affected by TB in west part of Pakistan. This could be either due to female exposure to those factors that predispose them to higher TB notification.

8 Conclusion and Future Work

This study outlines that India has the highest incidence of tuberculosis followed by china, whereas, death rates due to tuberculosis are also higher in India followed by Indonesia. This study could lay stress on the relevant measures that the concerned countries need to undertake. Suppose, India topped the list for both number of incidences as well as death rates due to tuberculosis. This indicates that the concerned system of that country should consider their health system on a very serious note, and take measures accordingly, that would improve their health system in the future. It should also check the poverty levels of the country; they could consider imposing a strict ban on the use of alcohol and smoking as they are the major cause for this life taking disease. The government could impose the duties on the consumption of alcohol and tobacco so that the consumption diminishes eventually.
Medical Awareness and proper preventive measures could help in safeguarding from HIV, which was found to be a serious trigger in progressing the tuberculosis. There was also found a gender disparity in the incidences the mortality rates, it could thus be

concluded that improvement in living standards without gender bias and public health could effectively control TB in the long run.
There is a further scope in this study as a researcher could formulate the policies that would be effective enough for TB control, considering the gender disparity, hygiene, health system, medical conditions, etc, and the study could include studying the impact of conditions of the people who are infected with tuberculosis as well.

References

Bloom, B. R. (2018), 'New promise for vaccines against tuberculosis', pp. 1672–1674.

Hyungmin Lee, J. K. (2015), 'Division of epidemic intelligence service, korea centers for disease control and prevention, cheongju, korea. b department of internal medicine, college of medicine, catholic university of korea, seoul, korea'.

Iacopo Baussano 1, Brian G. Williams3, P. N. M. B. U. F. F. S. (2011), 'Cancer epidemiology unit, upo a. avogadro and cpo-piemonte, novara, italy, 2division of epidemiology, public health and primary care, faculty of medicine, imperial college of science, technology and medicine, london, united kingdom, 3 stop tb department, world health organization, geneva, switzerland, 4 ser, epidemiological department, veneto region, castelfranco veneto, italy'.

Omara F Dogar, Sarwat K Shah2, A. A. C. & Qadeer3, E. (2017*a*), 'The coming of age of drug-susceptibility testing for tuberculosis', pp. 1474–1475.

Omara F Dogar, Sarwat K Shah2, A. A. C. & Qadeer3, E. (2017*b*), 'gender disparity in tuberculosis cases in easter'.

Pai M, Correa N, M. N. J. P. (2017), 'Reducing global tuberculosis deathstime for india to step up', **7**(1), 1174–1176.

Swaminathan, S. (2017), ', national institute for research in tuberculosis (icmr), chennai, india'.

Appendix

R code for cleaning of structured data sets

```
-----------------------------------------------------------------
for table stg_incidence.
-----------------------------------------------------------------

setwd("/Users/USER/Desktop/dwbi␣data/tb␣datasets")
tb<- read.csv("incidence.csv",header = T,na.strings = c(""))
tb$tbcases<-tb$tb_cases
tb$tbcases<-gsub('\\[.*?\\]', '', tb$tbcases)
tb$tbper_pop<-tb$tbperpop
tb$tbper_pop<-gsub('\\[.*?\\]', '', tb$tbper_pop)
tb$tbcases_hiv<-tb$tb_cases_hiv
```

```
tb$tbcases_hiv<-gsub('\\[.*?\\]', '', tb$tbcases_hiv)
tb$tbper_pophiv<-tb$tbperpop_hiv
tb$tbper_pophiv<-gsub('\\[.*?\\]', '', tb$tbper_pophiv)
tb$tb_cases<-NULL
tb$tbperpop<-NULL
tb$tb_cases_hiv<-NULL
tb$tbperpop_hiv<-NULL
tb$tbper_pophiv<-as.character((tb$tbper_pophiv))
tb$tbper_pophiv<-as.numeric((tb$tbper_pophiv))
tb$tbper_pophiv[is.na(tb$tbper_pophiv)]<-0
tb$tbcases_hiv<-as.character((tb$tbcases_hiv))
tb$tbcases_hiv<-as.integer((tb$tbcases_hiv))
tb$tbcases_hiv[is.na(tb$tbcases_hiv)]<-0
tb$tbper_pop<-as.integer((tb$tbper_pop))
tb$Country<-as.character((tb$Country))
tb$Year<-as.integer((tb$Year))
tb$tbcases<-as.integer((tb$tbcases))
tb$tbcases_hiv<-as.integer((tb$tbcases_hiv))

-----------------------------------------------------------------
For Table stg_mortality
-----------------------------------------------------------------

setwd("/Users/USER/Desktop/dwbi␣data/tb␣datasets")
tb2<- read.csv("mortality_by_country.csv",header = T,na.strings = c(""))
tb2$nod_exhiv<-tb2$nodexhiv
tb2$nod_exhiv<-gsub('\\[.*?\\]', '', tb2$nod_exhiv)
tb2$nod_hivnegpop<-tb2$nodhivnegpop
tb2$nod_hivnegpop<-gsub('\\[.*?\\]', '', tb2$nod_hivnegpop)
tb2$nodexhiv<-NULL
tb2$nodhivnegpop<-NULL
tb2$nod_exhiv<-as.character((tb2$nod_exhiv))
tb2$nod_exhiv<-as.integer((tb2$nod_exhiv))
tb2$nod_hivnegpop<-as.character((tb2$nod_hivnegpop))
tb2$nod_hivnegpop<-as.integer((tb2$nod_hivnegpop))
tb2$Country<-as.character((tb2$Country))

-----------------------------------------------------------------
FOR UNSTRUCTURED DATA stg_Gender_incidence
-----------------------------------------------------------------

install.packages("splitstackshape")
install.packages("tidyverse")
install.packages("tidyr")
library(splitstackshape)
library(rJava)
library(tabulizer)
library(tidyr)
library(tidyverse)

tab2 <- extract_tables("C:\\Users\\USER\\Desktop\\dwbi␣data\\tb␣datasets\\un
```

```
x3<-cbind(tab2[[1]][1:44,2])%>%data.frame()
colnames(x3)<-c("list")
x<-x3[c(13,15,17,19,20,22,23,25,27,29,30,32,33,35,36,38,40,43),]
x<-data.frame(x)
colnames(x)<-c("list")
x$list<-as.character(x$list)
y<-strsplit(x$list,"␣")
z<-do.call(rbind.data.frame, y)
colnames(z)<-c("cases","total","women","men","children","women_percentage","
z$Ca<-NULL
z$Cb<-NULL
z<-z[-c(2,4,6,8,10,12,14,16,18),]
z$cases<-as.character(z$cases)

#############inserting respective column names

z$cases[1]<-"Incident␣Cases␣2011"
z$cases[2]<-"Incident␣Cases␣2012"
z$cases[3]<-"Incident␣Cases␣2013"
z$cases[4]<-"Total␣deaths␣2011"
z$cases[5]<-"Toat␣deaths␣2012"
z$cases[6]<-"Total␣deaths␣2013"
z$cases[7]<-"Tb␣deaths␣in␣PLHIV␣2011"
z$cases[8]<-"Tb␣daths␣in␣PLHIV␣2012"
z$cases[9]<-"Tb␣deaths␣in␣PLHIV␣2013"
c<-z[c(7,8,9),]
c<-c[,c(1,3,4,5,6)]
m<-c("47%","50%","45%")
c<-data.frame(c,m)
colnames(c)<-c("cases","total","women","men","children","women_percentage")
merge.data.frame(z,c)
z$women_percentage<-as.character(z$women_percentage)
c$women_percentage<-as.character(c$women_percentage)
z<-z[-c(7,8,9,10),]
z$children<-as.character(z$children)
z$children[1:3]<-paste(z$children[1:3],"0",  sep = "")
z$women<-as.character(z$women)
z$women[4:6]<-paste(z$women[4:6],"0",  sep = "")
z$men<-as.character(z$men)
z$men[4:6]<-paste(z$men[4:6],"0",  sep = "")

################cleaning

z$total<-z$total<-gsub("m","",z$total)
z$total<-as.numeric(z$total)
z$total<- 1000000 * z$total
z$children<-gsub("\\*","",z$children)
z$children<-gsub("\\,","",z$children)
z$women_percentage<-gsub("\\%","",z$women_percentage)
z$women<-gsub("\\m","",z$women)
z$women<-gsub("\\.","",z$women)
```

```
z$women<-gsub("\\,","",z$women)
z$women<-as.integer(z$women)
z$women[1:3]<-10000*z$women[1:3]
z$men<-gsub("\\m","",z$men)
z$men<-gsub("\\,","",z$men)
z$men<-gsub("\\.","",z$men)
z$men<-as.integer(z$men)
z$men[1]<-10000*z$men[1]
z$men[2:3]<-1000*z$men[2:3]

#####splitting string to year coloumn

z$Year<-substr(z$cases,str_locate(z$cases,"2"),str_locate(z$cases,"2")+4)
####cleaning again
z$women<-as.integer(z$women)
z$men<-as.integer(z$men)
z$children<-as.integer(z$children)
z$Year<-as.integer(z$Year)

-----------------------------------------------------------------
For Statista Data
-----------------------------------------------------------------
setwd("/Users/USER/Desktop/dwbi␣data/tb␣datasets")
install.packages("readxl")
library(readxl)
Exceldata<-read_excel("statista.xlsx",skip=1, sheet = 2)
Exceldata<-Exceldata[-c(1:2),]
colnames(Exceldata)<-c("Country","Incidence")
-----------------------------------------------------------------
for Table tb_relapse_cases
-----------------------------------------------------------------

for table tb_relapse_cases   tb3

setwd("/Users/USER/Desktop/dwbi␣data/tb␣datasets")
tb3<- read.csv("TB_relapse_cases.csv",header = T,na.strings = c(""))
tb3$tbinsurance<-tb3$tb_insurance
tb3$tbinsurance<-gsub('\\[.*?\\]', '', tb3$tbinsurance)
tb3$tb_insurance<-NULL
tb3$Number.of.incident.tuberculosis.cases<-NULL
tb3$tb_relapse<-as.character((tb3$tb_relapse))
tb3$tb_relapse<-as.integer((tb3$tb_relapse))
tb3$tb_relapse[is.na(tb3$tb_relapse)]<-0
tb3$tbinsurance<-as.character((tb3$tbinsurance))
tb3$tbinsurance<-as.integer((tb3$tbinsurance))
tb3$tbinsurance[is.na(tb3$tbinsurance)]<-0
tb3$Country<-as.character((tb3$Country))
tb3$tb_relapse<-as.integer((tb3$tb_relapse))
tb3$tbinsurance<-as.integer((tb3$tbinsurance))

-----------------------------------------------------------------------
```

SQL FACT TABLE QUERY

```
INSERT INTO [dbo].[fact_table]
           ([incidence_id]
           ,[mortality_id]
           ,[relapse_id]
           ,[statista_id]
           ,[gender_case_ID]
           ,[country_id]
           ,[tbcases]
           ,[tbper_pop]
           ,[tbcases_hiv]
           ,[tbper_pophiv]
           ,[tb_relapse]
           ,[tbinsurance]
           ,[nod_exhiv]
           ,[nod_hivnegpop]
           ,[incidence]
           ,[women]
           ,[men]
           ,[children])

                    select
                       a.incidence_id
           ,b.mortality_id
           ,isnull(c.relapse_id,(Select relapse_id from Dim_Relapse_cases wh
           ,isnull(d.statistaa_id,(Select statistaa_id from Dim_Statistaa wh
           ,isnull(e.gender_case_ID,(Select gender_case_ID from Dim_Gender_c
           ,f.country_id
           ,isnull(Cast(a.tbcases as int),0)
           ,isnull(Cast(a.tbper_pop as int),0)
           ,isnull(Cast(a.tbcases_hiv as int),0)
           ,isnull(Cast(a.tbper_pophiv as int),0)
           ,isnull(Cast(c.tb_relapse as int),0)
           ,isnull(Cast(c.tbinsurance as int),0)
           ,isnull(Cast(b.nod_exhiv as int),0)
           ,isnull(Cast(b.nod_hivnegpop as int),0)
           ,isnull(Cast(d.incidence as int),0)
           ,isnull(Cast(e.women as int),0)
           ,isnull(Cast(e.men as int),0)
           ,isnull(Cast(e.children as int),0)


                from dbo.dim_incidence as a
                join Dim_country f on f.country_id=a.country_id
                join Dim_Mortality as b on a.country_id=b.country_id and a.y
                join Dim_Relapse_cases as c on a.country_id=c.country_id and
                left join Dim_Statistaa as d on a.country_id= d.country_id
                left join [dbo].[Dim_Gender_cases] e on e.year=a.year
```