Data Warehousing and Business Intelligence Project

on

Demography

Dinesh Gauda
17169836

MSc/PGDip Data Analytics – 2018/9

Submitted to: Dr. Simon Caton

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

Student Name:	Dinesh Gauda
Student ID:	17169836
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:	Demography

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

Demography

Dinesh Gauda
17169836

November 26, 2018

Abstract

Demography lets one understand the relationship between economic, social, cultural, and biological processes influencing a population Wikipedia (2018) as a result helping in better social planning, insurance forecasting, labor market analysis, economic development and identifying potential solution for social and economic problems [1]. Objective of this project is to build a data warehouse using different demographics for 146 countries around the world over a period of eight years ranging from 2006 to 2013 which would help in answering different business requirements.

This project uses Kimball's bottom-up approach to build a data warehouse which would help in finding list of countries with low Life Expectancy Rate and finding effect of Infant Mortality Rate and Life Expectancy Rate on Population as well as finding relationship between Literacy Rate and GDP.

1 Introduction

A demographic transition leads to a change in the supply and demand of labor, thus affecting the labor market which indeed have an impact on economic growth. Previous work has decomposed the population growth into its fertility and mortality components, and examined their independent effects on economic growth which helps in solving demographic problems. Bloom & Williamson (1997)

Objective of this project is to build a data warehouse which would help answer below business requirements:

(Req-1) Does Infant Mortality Rate and Fertility Rate have effect on Population ?

The project aims to compare data of above three measures for 146 countries over eight years of span and get better understanding of the above three measures will enable to tackle problems and find potential solutions related to them.

(Req-2) Locations with low Life Expectancy Rate.

The project aims to find list of locations with low life expectancy rate that will help in identifying locations with scarce expectancy rate so that more research can be done on finding reasons behind the lack in expectancy rate and resolve them.

[1] https://www.suda.su.se/education/what-is-demography

(Req-3) Is GDP dependent on Literacy Rate ?

The project aims to find relationship between gdp and literacy rate that would help in deciding whether to invest on improving literacy rate to achieve higher gdp or not.

2 Data Sources

This project uses five data sources and nine datasets to build a dataware house on Demography.

Source	Type	Brief Summary
Wikipedia	Structured	This data source provided literacy rate country wise which was the core requirement for my data warehousing model to have measures country wise.
Indexmundi	Unstructured	This data source provided the literacy rate of countries which where missing in Wikipedia data source .
Statista	Structured	This data source provides the worldwide life expectancy rate year wise, but for limited years only.
Data.World	Structured	This data source provided four datasets, first having population country wise across years, second one with fertility rate country wise across years, third one having the list of countries with there respective region, and the final one having life expectancy rate country wise across years.
Kaggle	Structured	This data source provides two datasets, first having gdp country wise across years and second one having infant mortality country wise.

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

2.1 Source 1: Statista

Dataset of global life expectancy at birth from 2006 to 2016 from Statista: `https://www.statista.com/statistics/805060/life-expectancy-at-birth-worldwide/` provides worldwide life expectancy as a measure and year as dimension in my data model.Data from this data source is downloaded in xlsx format. 6

This data source addresses the business requirements listed in Section 1 (Req 2).

2.2 Source 2: Wikipedia

List of countries by literacy rate from Wikipedia: `https://en.wikipedia.org/wiki/List_of_countries_by_literacy_rate` provides literacy rate as a measure and country name and gender type as dimensions in my data model. Data from this data source is scraped using R.

This data source addresses the business requirements listed in Section 1 (Req 3).

2.3 Source 3: Indexmundi

Literacy rate of countries from Indexmundi: `https://www.indexmundi.com/belgium/literacy.html` provides literacy rate of countries whose data is missing in Wikipedia data source as a measure and country name and gender type as dimensions in my data model. Data from this data source is scraped from multiple web pages using R. 9

This data source addresses the business requirements listed in Section 1 (Req 3).

2.4 Source 4: Data.World

Four datasets were accessed from Data.world: `https://data.world/bhavnachawla/population-fertility-rate-life-expectancy` which provides population, fertility rate and life expectancy rate as measures and year and location details(country and region) as dimensions in my data model. Data from this data source is downloaded in csv format using R. 5

This data source is used to address multiple business requirements listed in Section 1 (Req 1) and (Req 2).

2.5 Source 5: Kaggle

Two datasets were accessed from Kaggle: `https://www.kaggle.com/resulcaliskan/countries-gdp` and `https://www.kaggle.com/fernandol/countries-of-the-world` which provides gdp and infant mortality rate as measures and country and year as dimensions in my data model. Data from this data source is downloaded in csv and xls format using R. 7, 8

This data source is used to address multiple business requirements listed in Section 1 (Req 3).

3 Related Work

Demography is very vast and lot of research and work has been done before in this field. Few of the previous work/research has been studied in this project with the help of papers written on them.

3.1 Usage of similar data in previous works:

Following are the list of previous works that have used similar datasets:
1) In Cameron & Cameron (2006) the authors used data for developing countries from 1990 to 2002 to find the economic benefit of having higher literacy rate.
2) In Li (2015) the author have used data for 120 developing countries from 1970 to 2014 to analyze relationship between fertility rate and economic growth.
3) In Lee & Tuljapurkar (1994) researchers used data for G7 countries (Canada, France, Germany, Italy, Japan, the United Kingdom, and the United States) from 1950 to 1994 to analyze the trend in Child Mortality Rate.

4) In Oeppen & Vaupel (2002) they used data for Australia Iceland Japan The Netherlands New Zealand Norway Sweden and Switzerland from 1840 to 2000 to analysis the trend in Life Expectancy and predict Life Expectancy for the years to come.

3.2 Domain knowledge from previous findings:

Below are the findings from previous works:
1) In Cameron & Cameron (2006) the researchers found that countries with higher literacy rate has higher economic growth.
2) In Sonnega (2006) author has said that because of improved nutrition and the control of infectious diseases there was drastic reduction in child and infant mortality in first half of 20th century and thus resulting in increased life expectancy rate.
3) In Haines (1998) the author concluded that for United States of America from 1850s to 1940s reduction in mortality rate resulted in reduction in fertility rate.
4) In Oeppen & Vaupel (2002) the author said that for 160 years, life expectancy has steadily increased by a quarter of a year per year, an extraordinary constancy of human achievement.

3.3 Value added to previous work from this project:

In above papers researchers have taken few countries or a single country's demographics over years to analyze relationship between different demographics. This project aims to analyze demographics of 146 countries for a span of eight years from 2006 to 2013 to answer business requirements listed in Section 1.

4 Data Model

Data model of this project consists of three dimensions and seven measures.

4.1 Dimensions:

4.1.1 Dimension Location:

First dimension in the data model is location dimension which consists of Location_ID, Continent, Region and Country. Location dimension is created using data related to location from Data.World data source and function countrycode to get continent name from country name of R. Each and every measures in this data model can be drilled down or rolled up using the hierarchies created in location dimension.
Hierarchies created in location dimension which facilitates drill down and roll up in this dimension are:
.Continent
..Region
...Country

4.1.2 Dimension Date:

Second dimension in the data model is date dimension which consists of Date_ID, Decade and Year. Date dimension in this data model can be created from multiple data sources

but, It is using data from Data.World and R. Measures like population, fertility rate, life expectancy and gdp can be drilled down or rolled up using the hierarchies created in date dimension.
Hierarchies created in date dimension which facilitates drill down and roll up in this dimensions are:
.Decade
..Year

4.1.3 Dimension Gender:

Third and final dimension in the data model is gender dimension which consists of Gender_ID and Gender. This dimension is created using data generated after combining data from Wikipedia and Indexmundi. In this dimension drill down or roll up is not possible just filtering on the basis of gender is possible.

4.2 Fact:

Fact table in the data model consist if Literacy Rate, GDP, Population, Infant Mortality, Fertility Rate, Life Expectancy and Worldwide Life Expectancy as measures along with Fact_ID(Primary key), Location_ID, Date_ID and Gender_ID as foreign key to establish relationship between fact table and dimension tables of the data model.

Literacy rate and gdp can be drilled down or rolled up in location dimension and filtered using gender dimension which can used to analyze dependency of gdp on literacy rate in various continent/region/countries as mentioned in business requirements listed in Section 1 (Req 3).

Population, infant mortality and fertility rate are used to find effect of infant mortality and fertility rate on population in various continent/region/countries over the decade/years as mentioned in business requirements listed in Section 1 (Req 1).

Last two measures in fact table life expectancy and worldwide life expectancy are used to find list of locations with low life expectancy rate as mentioned in business requirements listed in Section 1 (Req 2).

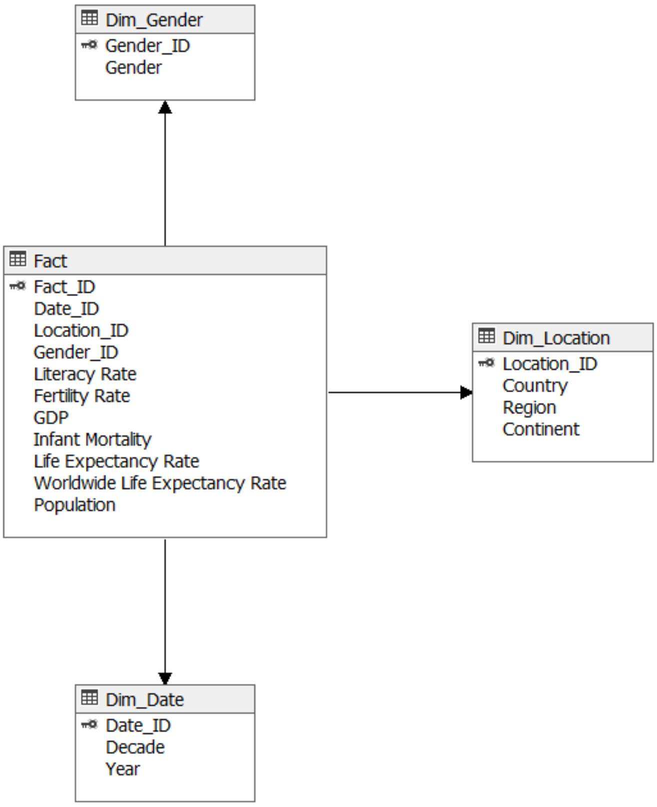


Figure 1: Star schema

5 Logical Data Map

Table 3: Logical Data Map describing all transformations, sources and destinations for all components of the data model illustrated in Figure 1

Source	Column	Destination	Column	Type	Transformation
Data.World	Year values from 1960 to 2013	Dim_Date	Year	Dimension	Remodelled the dataset by transforming the columns(Years 1960-2013) into rows with header as Years besides their respective Country column. Removed rows having years values other than 2006 to 2013. Changed the data type of Year to numeric. Removed columns which are not required.
Data.World	Year values from 1960 to 2013	Fact	Population	Fact	Remodelled the dataset by transforming each unique row cells with population values into columns cells of new column Population with their respective country and year. Removed rows with empty or NA values in population column.
Data.World	Year values from 1960 to 2013	Dim_Date	Decade	Dimension	Created new column header Decade using remodelled column Year and inserted values as per year value.
Data.World	Country Name	Dim_Location	Country	Dimension	Kept countries which are common in all datasets without empty or NA values in a row. Changed letter case of countries to lower case.
Data.World	Country Name	Dim_Location	Continent	Dimension	Created new column header Continent and inserted values as per country value using R function countrycode. Removed columns which are not required.

Continued on next page

Table 3 – *Continued from previous page*

Source	**Column**	**Destination**	**Column**	**Type**	**Transformation**
Data.World	Region	Dim_Location	Region	Dimension	Replaced symbol '&' with string 'and'.
Data.World	Country Name	Dim_Location	Country	Dimension	Kept countries which are common in all datasets without empty or NA values in a row. Changed letter case of Country values to lower case.
Data.World	Year values from 1960 to 2013	Dim_Date	Year	Dimension	Did Same transformation as in first row but now on fertility rate dataset.
Data.World	Year values from 1960 to 2013	Fact	Fertility Rate	Fact	Did same transformation as in second row but now on fertility rate dataset.
Data.World	Year values from 1960 to 2013	Dim_Date	Decade	Dimension	Did Same transformation as in third row but now on fertility rate dataset.
Data.World	Country Name	Dim_Location	Country	Dimension	Did Same transformation as in fourth row but now on fertility rate dataset.
Data.World	Year values from 1960 to 2013	Dim_Date	Year	Dimension	Did same transformation as in first row but now on life expectancy rate dataset.
Data.World	Year values from 1960 to 2013	Fact	Life Expectancy Rate	Fact	Did same transformation as in second row but now on life expectancy rate dataset.
Data.World	Year values from 1960 to 2013	Dim_Date	Decade	Dimension	Did same transformation as in third row but now on life expectancy rate dataset.
Data.World	Country Name	Dim_Location	Country	Dimension	Did same transformation as in fourth but now on life expectancy rate dataset.

Continued on next page

Table 3 – *Continued from previous page*

Source	**Column**	**Destination**	**Column**	**Type**	**Transformation**
Kaggle	Year values from 1960 to 2017	Dim_Date	Year	Dimension	Did same transformation as in third row but now for different year span(1960 to 2017)
Kaggle	1960,...,2017	Fact	GDP	Fact	Remodelled the dataset by transforming each unique row cells with gdp values into columns cells of new column GDP with their respective country and year. Removed rows with empty or NA values in GDP column. Converted gdp values from US dollars to million US dollars.
Kaggle	1960,...,2013	Dim_Date	Decade	Dimension	Did same transformation as in third but now on gdp dataset.
Kaggle	Country Name	Dim_Location	Country	Dimension	Did same transformation as in fourth but now on gdp dataset.
Kaggle	Country	Dim_Location	Country	Dimension	Removed leading and trailing white spaces. Kept countries which are common in all datasets without empty or NA values in a row. Changed letter case to lower case.
Kaggle	Infant mortality (per 1000 births)	Fact	Infant Mortality	Fact	Replaced symbol ',' with '.'.
Wikipedia	Country	Dim_Location	Country	Dimension	Merged all rows in this dataset and all rows from dataset indexmundi. Kept countries which are common in all datasets without empty or NA values in a row. Changed letter case of countries to lower case. Removed columns which are not required.

Continued on next page

Table 3 – *Continued from previous page*

Source	**Column**	**Destination**	**Column**	**Type**	**Transformation**
Wikipedia	Literacy rate(all)	Fact	Literacy Rate	Fact	Inserted values in the column Literacy Rate and inserted Overall in column Gender foe each new row inserted. Removed values having value as 'not reported by UNESCO 2015' or NA or empty in column Literacy Rate. Removed percentage symbol from Literacy Rate values.
Wikipedia	Male literacy rate	Fact	Literacy Rate	Fact	Did same transformation as for Literacy rate(all) but now for Male literacy rate.
Wikipedia	Female literacy rate	Fact	Literacy Rate	Fact	Did same transformation as for Literacy rate(all) but now for Female literacy rate.
Wikipedia	Literacy rate(all), Male literacy rate, Female literacy rate	Dim_Gender	Gender	Dimension	Created new column header Gender and inserted values into the column based on the values of literacy rate.
Indexmundi	Country	Dim_Location	Country	Dimension	Replaced symbol '_' with empty space. Kept countries which are common in all datasets without empty or NA values in a row. Changed letter case of countries to lower case.
Indexmundi	Total literacy rate	Fact	Literacy Rate	Fact	Added values in the column Literacy Rate of Fact for Gender Overall. Removed values having empty or NA value in column Literacy Rate. Removed percentage symbol from Literacy Rate values.
Indexmundi	Male literacy rate	Fact	Literacy Rate	Fact	Did same transformation as for Total literacy rate but now for Male literacy rate.
Indexmundi	Female literacy rate	Fact	Literacy Rate	Fact	Did same transformation as for Literacy rate(all) but now for Female literacy rate.

Continued on next page

Table 3 – *Continued from previous page*

Source	**Column**	**Destination**	**Column**	**Type**	**Transformation**
Indexmundi	Total literacy rate, Female literacy rate, Male literacy rate	Dim_Gender	Gender	Dimension	Created new column header Gender and inserted values into the column based on the values of literacy rate.
Statista	Year	Dim_Date	Year	Dimension	Kept rows with year values that are common in all datasets(2006-2013) without empty or NA values in a row.
Statista	Year	Dim_Date	Decade	Dimension	Created new column header Decade and inserted values as per year value.
Statista	Life expectancy at birth in years	Fact	Worldwide Life Expectancy Rate	Fact	Inserted the values of Worldwide Life Expectancy Rate for the years common in all datasets.

6 ETL Process

6.1 Extraction:

This project access data from five different data sources to build a data warehouse on demography.

6.1.1 Wikipedia

Data about literacy rate of different countries gender wise is scraped from this data source using functions read_html and read_table of rvest package in R. Automatic scraping of data from this data source is done by executing R scripts in SSIS with the help of Execute Process Task component.

6.1.2 Indexmundi

Gender wise literacy rate of countries whose data is missing in Wikipedia data source is scraped from this data source. To get data of all missing countries multiple webpages are accessed dynamically one by one and then data from all the web pages are parsed to get the desired information. Automatic scraping of data from this data source is done by executing R scripts in SSIS with the help of Execute Process Task component.

6.1.3 Statista

Worldwide life expectancy for a span of eight years from 2003 to 2013 is accessed from statista. Data from this data source is downloaded in xls format. Extraction of data from this data source is not automated as there is no api provided by Statista to download files and requires login credentials to download files from the website.

6.1.4 Kaggle

From this data source two datasets has been accessed which contains data related to gdp and infant mortality rate. Data related to gdp is downloaded in csv format with the help of api provided by kaggle and data related to infant mortality is downloaded first in the form of zip and then extracted to get the data in csv format. Extraction of data from from this data source is automated by done by executing R scripts in SSIS with the help of Execute Process Task component.

6.1.5 Data.World

From this data source four datasets has been accessed which contains data related to fertility rate, life expectancy rate, location and population. Data related to life expectancy rate is downloaded using library httr of R in csv format and for fertility rate, location and population in xls format. Extraction of data from this data source is automated by executing R scripts in SSIS with the help of Execute Process Task component.

6.2 Cleaning:

Once the data from all sources were scraped and downloaded, they were cleaned, merged and filtered accordingly so that they can used to create fact and dimensions. Cleaning, merging and filtering of data sources is automated by executing R scripts in SSIS with the help of Execute Process Task component.

6.3 Truncate and Staging:

To avoid insertion of duplicate entries into database truncation of tables are done by executing a sql query. Automatic clearing of data from table is done by executing sql in SSIS with the help of Execute SQl task in SSIS. As data in database cannot be truncated if there exists relationships in database such as primary key-foreign key relationship, all such relationships were dropped before clearing data from database. Datasets which are now in csv formats were loaded into database with the help of Flat File Source and OLE DB Destination components in SSIS.

6.4 Transformation:

Once all raw files were loaded into database, Raw tables are accessed in SSIS with the help of OLE DB Source component and then Dimension table is created using raw tables after removing duplicates entries ins raw tables with help of Sort component in SSIS. To create fact table all raw files and dimension tables are joined using Lookup and Merge Join component in SSIS.

6.5 Loading:

Loading of dimensions and fact into databases has been done with the help of OLE DB Destination component after which relationship with dimension and fact is established by executing SQL query in SSIS with the help of Execute SQL task component.

6.6 OLAP Cube:

The process for creating OLAP cube starts by creating a connection with Database Server to fetch Dimension and Fact tables in SSAS and then establish a connection with Analysis Service Server where the OLAP cube is to be deployed. Once connections are established populated Dimensions and Fact tables from database server are loaded into Analysis Service Server and then using Dimension and Fact tables OLAP cube is created after which hierarchies for dimensions are created and finally cube is deployed in SSAS. To automate redeployment of OLAP cube first dimensions are deployed followed by fact with the help of Analysis Services Processing Task in SSIS. The whole ETL process right from extraction of data source to deployment of cube is automated in SSIS.

7 Application

Following BI Queries are used to address and demonstrate the attainment of the business requirements noted in Section 1.

7.1 BI Query 1: Relationship between Infant Mortality, Fertility Rate and Population.

For this query, the contributing sources of data are: Data.World and Kaggle.
The general findings are that 1) Population is not only dependent on fertility rate but also on infant mortality rate and 2) Locations with low infant mortality rate has low fertility rate.Figure 2.

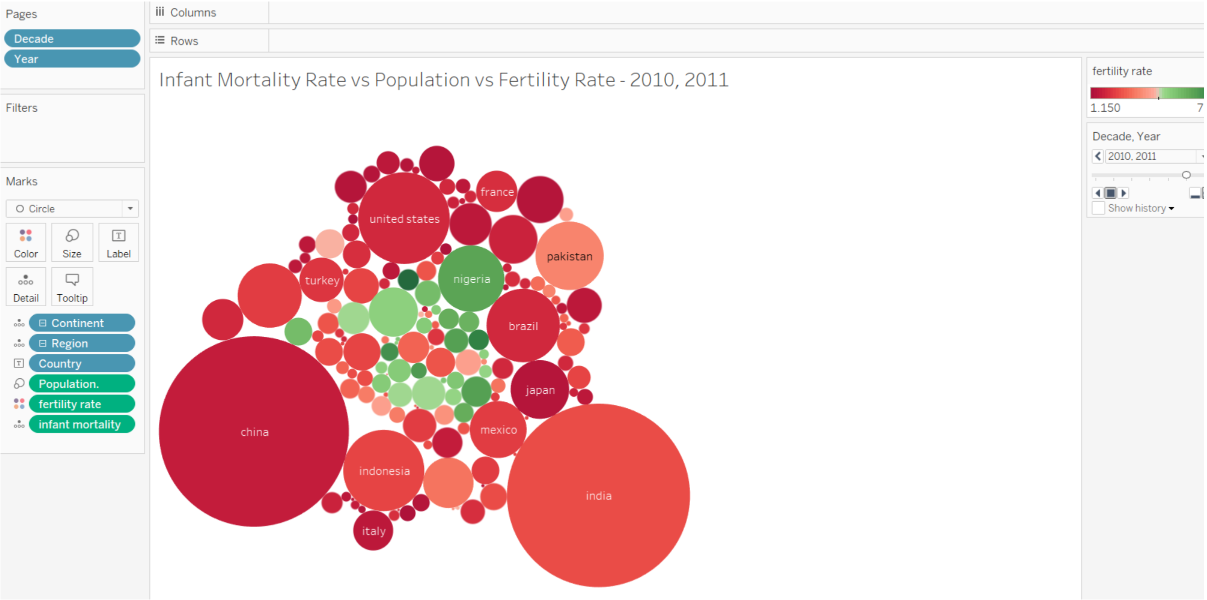


Figure 2: Results for BI Query 1

7.2 BI Query 2: Comparison of Life Expectancy Rate of individual country with Worldwide Life Expectancy Rate.

For this query, the contributing sources of data are: Statista and Data.World.
The general findings are that African countries have lower life expectancy rate compared to other countries in the world as illustrated in Figure 3.

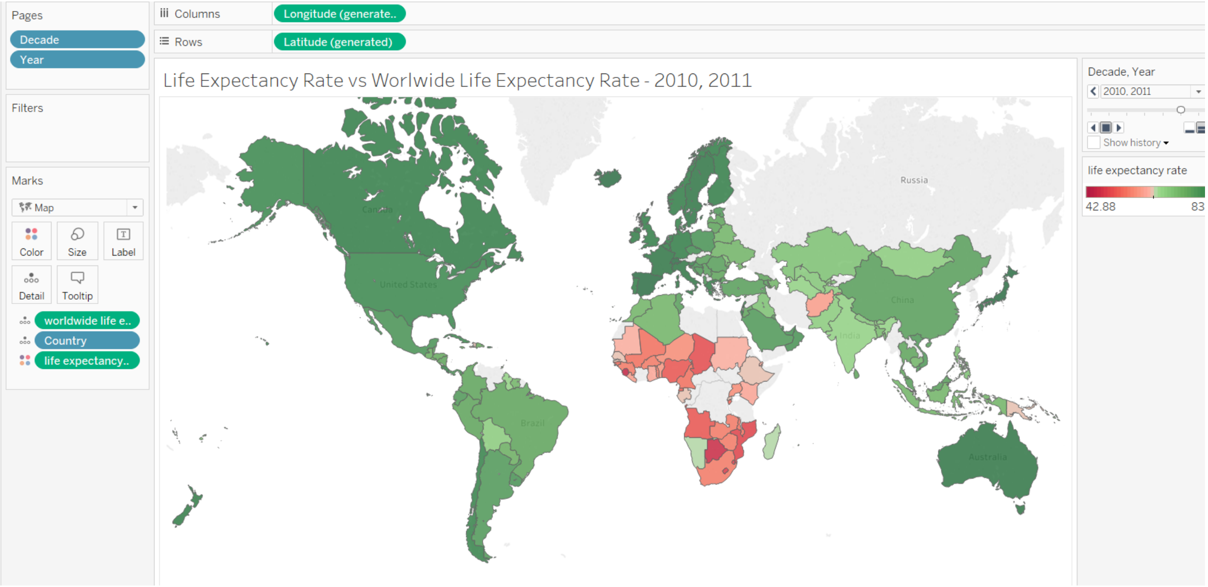
7.3 BI Query 3: Relationship between Literacy Rate and GDP.

For this query, the contributing sources of data are: Wikipedia, Indexmundi and Kaggle.
The general findings are that GDP is directly dependent on literacy rate as illustrated in Figure 4.

7.4 Discussion

7.4.1 First BI query:

In first business query comparison of Infant Mortality Rate, Fertility Rate and Population across continents, regions & countries over the decades & years is done and following things has been observed:
1. Countries with higher fertility rate such as Niger, Mali, Chad have lower population and higher infant mortality rate.


Columns: Longitude (generate..

Rows: Latitude (generated)

Life Expectancy Rate vs Worlwide Life Expectancy Rate - 2010, 2011

Decade, Year: 2010, 2011

Figure 3: Results for BI Query 2

2. Countries with higher population such as China, India have low fertility rate and low infant mortality rate.
Similar trends were found in continents and regions.
With the help pf above observations it can be said that population not only depends on fertility rate but also on infant mortality rate.

In relation to the findings of previous work which stated that reduction in infant mortality rate resulted in reduction in fertility rate. Haines (1998) Findings from this business query also shows that countries with lower infant mortality rate has lower fertility rate.

7.4.2 Second BI query:

In second business query Life Expectancy Rate of individual countries are compared with Worldwide Life Expectancy Rate across decades & years and following observations have been made :
1. African countries such as Botswana, Mozambique, Chad have lower life expectancy rate compare to other counties of the world
2. Worldwide life expectancy rate as well as life expectancy rate of individual countries over the years is constantly increasing.
With the help of above observations it can be said African countries have lower life expectancy rate compared to rest of the countries in the world.

In relation to the findings of previous work which stated that for 160 years life expectancy has steadily increased by a quarter of a year per year.Oeppen & Vaupel (2002) Findings from this business query also shows that there is increase in life expectancy over the years.

7.4.3 Third BI query:

In third business query, comparison of GDP with Literacy Rate across continents, regions & countries over the decades & years is done and following observations are made:

1. Continent such as Europe with higher literacy rate higher better GDP.
2. Continent such as Africa with lower literacy rate has lower GDP.
3. Similar trend can be observed in regions and countries.

From above observations it can be concluded that GDP of a location is dependent on literacy rate of a location

In relation to the findings of previous work which stated that countries with higher literacy rate has higher economic growth. Cameron & Cameron (2006) Findings from this business query also shows that locations with higher literacy rate has higher GDP.

8 Conclusion and Future Work

Findings from this project are 1) Population not only depends on fertility rate but also on infant mortality rate. 2) Countries with lower infant mortality rate has lower fertility rate. 3) Life Expectancy rate has increased over the years. 4) GDP is dependent on literacy rate.

In this project a data warehouse is implemented which analyze different demographics for 146 countries over a span of eight years which gives insights about the topic thus helping in answering business requirements, but the span of years for which analysis is done is very small to generalize the findings.

So a follow up project can be done which would cover more years so as to be sure that the findings are correct and not just happened by chance. Also, In this project the scope is limited to countries so to get more insights analysis for states/county of the country can be done which will give better understanding of how results of each states/county aggregates to give result for the country.

References

Bloom, D. & Williamson, J. (1997), Demographic transitions and economic miracles in emerging asia, Technical report.

Cameron, D. J. & Cameron, S. (2006), 'The economic benefits of increased literacy', *Education for All Global Monitoring Report UNESCO* .

Haines, M. R. (1998), 'The relationshp between infant and child mortality and fertility: Some historical and contemporary evidence for the united states'.

Lee, R. D. & Tuljapurkar, S. (1994), 'Stochastic population forecasts for the united states: Beyond high, medium, and low', *Journal of the American Statistical Association* **89**(428), 1175–1189.

Li, Y. (2015), 'The relationship between fertility rate and economic growth in developing countries'.

Oeppen, J. & Vaupel, J. W. (2002), 'Broken limits to life expectancy', *American Association for the Advancement of Science (AAAS)* **296**(5570), 1029–1031.

Sonnega, A. (2006), 'The future of human life expectancy: Have we reached the ceiling or is the sky the limit?'.

Wikipedia (2018), ‘Demography — Wikipedia, the free encyclopedia’, http://en.wikipedia.org/w/index.php?title=Demography&oldid=870578838. [Online; accessed 26-November-2018].

Appendix

R code example

```
scrapping_literacy_rate_wiki.R:
setwd("C:\\Users\\Dell\\Desktop\\Countries\\R")
library(rvest)

#scraping html table using rvest from wikipedia
literacy_rate <- read_html("https://en.wikipedia.org/wiki/List_of_countries_
literacy_rate_table = html_table(html_nodes(literacy_rate, "table")[[4]], fi

literacy_rate_table = tail(literacy_rate_table, -1)

names(literacy_rate_table) <- c(’Country’, ’Literacy␣rate(all)’, ’Male␣liter

write.csv(literacy_rate_table, ’raw_literacy_rate_wiki.csv’)

scraping_literacy_rate_missing.R (From Indexmundi.com):
setwd("C:\\Users\\Dell\\Desktop\\Countries\\R")
library(rvest)
library(stringr)
source("util.R")


literacyRateDf = data.frame(matrix(ncol = 4))
colnames(literacyRateDf) = c(’Country’, ’Total␣literacy␣rate’, ’Male␣literac
countries = c(’andorra’, ’australia’, ’belgium’, ’barbados’, ’canada’, ’czec


for(i in 1:length(countries)){
urlContent = read_html(paste("https://www.indexmundi.com/",countries[i],"/li

allData = urlContent %>% html_nodes("div.c") %>% html_text()

location_total = instr(allData, ’total␣population:’) + 18
literacyRate_total = substr(allData, location_total, location_total+5)
literacyRate_total = numextract(literacyRate_total)

location_male = instr(allData, ’male:’) + 5
literacyRate_male = substr(allData, location_total, location_total+5)
literacyRate_male = numextract(literacyRate_male)

location_female = instr(allData, ’female:’) + 7
literacyRate_female = substr(allData, location_total, location_total+5)
literacyRate_female = numextract(literacyRate_female)

rowContent = c(countries[i], literacyRate_total, literacyRate_male, literacy
literacyRateDf =rbind(literacyRateDf, rowContent)
}
literacyRateDf = literacyRateDf[-1, ]
write.csv(literacyRateDf, 'raw_literacy_rate_missing.csv')

extraction_from_data_world:
setwd("C:\\Users\\Dell\\Desktop\\Countries\\R")
library("httr")
library("readxl")

GET("https://query.data.world/s/enywjzjweteawkuzd3llgwsnmkrwmq", write_disk(
df <- read_excel(tf)
write.csv(df,"raw_population.csv", row.names = F)

GET("https://query.data.world/s/ad4vnhjplgn6tmogwlusdcwzpjzdvu", write_disk(
df <- read_excel(tf)
write.csv(df,"raw_location.csv", row.names = F)

GET("https://query.data.world/s/d54pheg25sygfg5ntsfx226zcon7ji", write_disk(
df <- read_excel(tf)
write.csv(df,"raw_fertility_rate.csv", row.names = F)

GET("https://query.data.world/s/sg4z5wzztp4d5hjqgzcwxj44gvyh7u", write_disk(
df <- read.csv(tf)
write.csv(df,"raw_life-epectancy-rate_data_world.csv", row.names = F)


extraction_from_Kaggle.R:

setwd("C:\\Users\\Dell\\Desktop\\Countries\\R")

#delete file if already exist
fileName <- "raw_gdp_kaggle.xls"
if (file.exists(fileName)) file.remove(fileName)

fileName <- "raw_infant_mortality.csv"
if (file.exists(fileName)) file.remove(fileName)

system("kaggle␣datasets␣download␣-f␣w_gdp.xls␣resulcaliskan/countries-gdp")
file.rename("w_gdp.xls", "raw_gdp_kaggle.xls")


system("kaggle␣datasets␣download␣-d␣fernandol/countries-of-the-world")
zipF<- "countries-of-the-world.zip"
outDir<-"C:\\Users\\Dell\\Desktop\\Countries\\R"
unzip(zipF,exdir=outDir)
file.rename("countries␣of␣the␣world.csv", "raw_infant_mortality.csv")
```

```
#delete zip
fileName <- "countries-of-the-world.zip"
if (file.exists(fileName)) file.remove(fileName)


cleaning.R:
setwd("C:\\Users\\Dell\\Desktop\\Countries\\R")
library(xlsx)
library(stringr)
source("util.R")

#Infant mortality
filename = ’raw_infant_mortality.csv’
infant_mortality = read.csv(filename, header = T, stringsAsFactors = F)

infant_mortality = infant_mortality[, c(1,8)]
column_names <- c(’Country’, ’Infant␣Mortality’)
colnames(infant_mortality) <- column_names
infant_mortality$‘Infant Mortality‘ <- gsub("[/,]",".",infant_mortality$‘Inf
infant_mortality$Country <- gsub("[/&]","and",infant_mortality$Country)
infant_mortality$Country = trim(infant_mortality$Country)
infant_mortality$Country = tolower(infant_mortality$Country)

write.csv(infant_mortality,’infant_mortality_intermediate.csv’, row.names=FA

#Population

filename = ’raw_population.csv’
df = read.csv(filename, header=T,stringsAsFactors = F)
colLength= ncol(df);
df = df[, c(1,colLength, colLength-1, colLength-2, colLength-3, colLength-4,
colnames(df) = c(’Country’, ’2013’, ’2012’, ’2011’, ’2010’, ’2009’, ’2008’,
df$Country= tolower(df$Country)

write.csv(df,’population_intermediate.csv’, row.names=FALSE)


#Fertility Rate
filename = ’raw_fertility_rate.csv’
df = read.csv(filename, header=T,stringsAsFactors = F)
colLength= ncol(df);
df = df[, c(1,colLength, colLength-1, colLength-2, colLength-3, colLength-4,
colnames(df) = c(’Country’, ’2013’, ’2012’, ’2011’, ’2010’, ’2009’, ’2008’,
df$Country= tolower(df$Country)

write.csv(df,’fertility_rate_intermediate.csv’, row.names=FALSE)


#GDP
filename = ’raw_gdp_kaggle.xls’
df <- read.xlsx(filename, sheetIndex = 1, startRow=4)
```

```
colLength= ncol(df);
df = df[, c(1,colLength-4, colLength-5, colLength-6, colLength-7, colLength-
colnames(df) = c(’Country’, ’2013’, ’2012’, ’2011’, ’2010’, ’2009’, ’2008’,
df$Country= tolower(df$Country)

write.csv(df,’gdp_intermediate.csv’, row.names=FALSE)


#Life Expectancy

filename = ’raw_life-epectancy-rate_data_world.csv’
life_expectancy_rate = read.csv(filename, header=T,stringsAsFactors = F)

colLength= ncol(life_expectancy_rate);
life_expectancy_rate = life_expectancy_rate[, c(1,colLength, colLength-1, co
colnames(life_expectancy_rate) = c(’Country’, ’2013’, ’2012’, ’2011’, ’2010’
life_expectancy_rate$Country= tolower(life_expectancy_rate$Country)

countries = life_expectancy_rate$Country

lifeExpectancy <- data.frame(matrix(ncol = 3, nrow= 0))
column_names <- c("Country", "Year", "Life.expectancy")
colnames(lifeExpectancy) <- column_names

for (i in 2: ncol(life_expectancy_rate)) {
  lengthOfDataFrame = nrow(lifeExpectancy)
  for (j in 1 : length(countries)) {
    Year = as.numeric(gsub("[^0-9.]", "",  colnames( life_expectancy_rate )[
    lifeExpectancy[ lengthOfDataFrame + j, ] = c(life_expectancy_rate$Countr
  }
}


filename = ’raw_life_expectancy_rate_statista.xlsx’
df <- read.xlsx(filename, sheetIndex = 2, startRow=5)
df = df[(c(1:8)), ]
colnames(df) =  c(’Year’, ’Life.expectancy’)
write.csv(df,’life_expectancy_rate_statista.csv’, row.names=FALSE)

lifeExpectancy_Statista = read.csv("life_expectancy_rate_statista.csv", head

life_expectancy_rate<- data.frame(matrix(ncol = 8, nrow= 0))
column_names <- c("2006", "2007", "2008", "2009", "2010", "2011", "2012", "2
colnames(life_expectancy_rate) <- column_names

for (i in 1 : ncol(life_expectancy_rate)) {
  life_expectancy_rate[1,i] = lifeExpectancy_Statista$Life.expectancy[i]
}

life_expectancy_rate_data_frame <- data.frame(matrix(ncol = 4, nrow= 0))
column_names <- c("Country", "Year", "Life␣Expectancy␣Rate", "Worldwide␣Life
```

```r
colnames(life_expectancy_rate_data_frame) <- column_names

for (j in 1 : length(lifeExpectancy$Country)) {
  year = lifeExpectancy$Year[j]
  location = which( colnames(life_expectancy_rate)== as.character(year) )
  life_expectancy_rate_data_frame[j, ] = c(lifeExpectancy$Country[j], lifeEx
}

life_expectancy_rate_data_frame$Country= tolower(life_expectancy_rate_data_f

fileName <- "life_expectancy_rate_statista.csv"
if (file.exists(fileName)) file.remove(fileName)

write.csv(life_expectancy_rate_data_frame, "life_expectancy_rate_intermediat


#Literacy Rate

datawiki = read.csv("raw_literacy_rate_wiki.csv", header=T,stringsAsFactors
datawiki$X <- NULL
datawiki$Gender.difference <- NULL
datawiki$non.unesco <- NULL

datawiki = datawiki[datawiki$Male.literacy.rate != ’not␣reported␣by␣UNESCO␣2

datawiki$Literacy.rate.all. = numextract(datawiki$Literacy.rate.all.)
datawiki$Male.literacy.rate = numextract(datawiki$Male.literacy.rate)
datawiki$Female.literacy.rate = numextract(datawiki$Female.literacy.rate)
datawiki$Country = str_to_lower(datawiki$Country)

colnames(datawiki) = c(’Country’, ’Literacy␣rate␣all’,’Male␣literacy␣rate’,
write.csv(datawiki, ’literacy_rate_wiki.csv’, row.names=FALSE)

dataMissing = read.csv("raw_literacy_rate_missing.csv", header=T,stringsAsFa
dataMissing$X <- NULL
dataMissing$Country = str_to_lower(dataMissing$Country)
dataMissing$Country <- gsub(’_’, ’␣’, dataMissing$Country)

colnames(dataMissing) = c(’Country’, ’Literacy␣rate␣all’,’Male␣literacy␣rate

write.csv(dataMissing, ’literacy_rate_missing.csv’, row.names=FALSE)

dataMissing = read.csv("literacy_rate_missing.csv", header=T,stringsAsFactor
datawiki = read.csv("literacy_rate_wiki.csv", header=T,stringsAsFactors = F)

dataMissing$X <- NULL
datawiki$X <- NULL

literacyData <- merge(dataMissing, datawiki,all.x = TRUE, all.y = TRUE)
literacyData$X <- NULL
```

```
fileName <- "literacy_rate_missing.csv"
if (file.exists(fileName)) file.remove(fileName)

fileName <- "literacy_rate_wiki.csv"
if (file.exists(fileName)) file.remove(fileName)

write.csv(literacyData, ’literacy_rate_intermediate.csv’, row.names = F)


#Location

filename = ’raw_location.csv’
df = read.csv(filename, header=T,stringsAsFactors = F)

df$Country.Code = NULL
df$SpecialNotes = NULL
df$IncomeGroup = NULL
colnames(df)[colnames(df)=="Country.Name"] <- "Country"
df$Country= tolower(df$Country)

write.csv(df, ’location_intermediate.csv’ ,row.names = F)

filtering.R:

setwd("C:\\Users\\Dell\\Desktop\\Countries\\R")
library(stringr)
library(countrycode)

#read csv files
literacy_rate = read.csv("literacy_rate_intermediate.csv", header = T, strin
gdp = read.csv("gdp_intermediate.csv", header = T, stringsAsFactors = F)
fertility_rate = read.csv("fertility_rate_intermediate.csv", header = T, str
life_expectancy_rate = read.csv("life_expectancy_rate_intermediate.csv", hea
meta_data = read.csv(’location_intermediate.csv’, header=T,stringsAsFactors
infant_mortality = read.csv(’infant_mortality_intermediate.csv’, header = T,
population = read.csv("population_intermediate.csv", header = T, stringsAsFa


#Delete Intermediate files
fileName <- "literacy_rate_intermediate.csv"
if (file.exists(fileName)) file.remove(fileName)

fileName <- "gdp_intermediate.csv"
if (file.exists(fileName)) file.remove(fileName)

# fileName <- "population_intermediate.csv"
# if (file.exists(fileName)) file.remove(fileName)

fileName <- "fertility_rate_intermediate.csv"
if (file.exists(fileName)) file.remove(fileName)
```

```
fileName <- "life_expectancy_rate_intermediate.csv"
if (file.exists(fileName)) file.remove(fileName)

fileName <- "location_intermediate.csv"
if (file.exists(fileName)) file.remove(fileName)

fileName <- "infant_mortality_intermediate.csv"
if (file.exists(fileName)) file.remove(fileName)

fileName <- "population_intermediate.csv"
if (file.exists(fileName)) file.remove(fileName)


#merge by country
all_data_by_country = Reduce(function(...) merge(..., by='Country'), mget(c(
#remove rows with NA
all_data_by_country =  na.omit(all_data_by_country)


#filter on the basis of common countries
literacy_rate = literacy_rate[literacy_rate$Country %in% all_data_by_country
gdp = gdp[gdp$Country %in% all_data_by_country$Country,]
fertility_rate = fertility_rate[fertility_rate$Country %in% all_data_by_coun
life_expectancy_rate = life_expectancy_rate[life_expectancy_rate$Country %in
meta_data = meta_data[meta_data$Country %in% all_data_by_country$Country, ]
infant_mortality = infant_mortality[infant_mortality$Country %in% all_data_b

#infant mortality
write.csv(infant_mortality, 'infant_mortality.csv', row.names = F)


#population
countries = population$Country

population_data_frame <- data.frame(matrix(ncol = 4, nrow= 0))
column_names <- c("Country", "Decade", "Year", "Population")
colnames(population_data_frame) <- column_names

for (i in 2: ncol(population)) {
  lengthOfDataFrame = nrow(population_data_frame)
  for (j in 1 : length(countries)) {
    Year = as.numeric(gsub("[^0-9.]", "",  colnames( population )[i]))
    Decade = Year - (Year %% 10)
    population_data_frame[ lengthOfDataFrame + j, ] = c(population$Country[j
  }
}
write.csv(population_data_frame, 'population.csv', row.names=FALSE)
```

```
#location
column_names <- c("Country", "Region")
colnames(meta_data) <- column_names
meta_data$Continent <- factor(countrycode(sourcevar = meta_data[, "Country"]
                                          origin = "country.name",
                                          destination = "continent"))
meta_data$Region = gsub("&", "and", meta_data$Region)
write.csv(meta_data, ’location.csv’, row.names=FALSE)


#life_expectancy_rate
column_names <- c("Country", "Year", "Life␣Expectancy␣Rate", "Worldwide␣Life
colnames(life_expectancy_rate) <- column_names
life_expectancy_rate$Decade = 2010
for (j in 1 : length(life_expectancy_rate$Country)) {
  Decade = life_expectancy_rate$Year[j] - (life_expectancy_rate$Year[j] %% 1
  life_expectancy_rate$Decade[j]= Decade
}
write.csv(life_expectancy_rate, ’life_expectancy_rate.csv’, row.names=FALSE)

#literacy_rate
countries = literacy_rate$Country

literacy_rate_data_frame <- data.frame(matrix(ncol = 3, nrow= 0))
column_names <- c("Country", "Gender", "Literacy␣Rate")
colnames(literacy_rate_data_frame) <- column_names
Gender = c(’’, ’Overall’, ’Male’, ’Female’)
for (i in 2: ncol(literacy_rate)) {
  lengthOfDataFrame = nrow(literacy_rate_data_frame)
  for (j in 1 : length(countries)) {
    literacy_rate_data_frame[ lengthOfDataFrame + j, ] = c(literacy_rate$Cou
  }
}
write.csv(literacy_rate_data_frame, ’literacy_rate.csv’, row.names=FALSE)


#GDP
countries = gdp$Country

gdp_data_frame <- data.frame(matrix(ncol = 4, nrow= 0))
column_names <- c("Country","Decade", "Year", "GDP")
colnames(gdp_data_frame) <- column_names

for (i in 2: ncol(gdp)) {
  lengthOfDataFrame = nrow(gdp_data_frame)
  for (j in 1 : length(countries)) {
    Year = as.numeric(gsub("[^0-9.]", "",  colnames( gdp )[i]))
    Decade = Year - (Year %% 10)
    gdp_data_frame[ lengthOfDataFrame + j, ] = c(gdp$Country[j], Decade, Yea
  }
```

```
}
gdp_data_frame$GDP = as.numeric(gdp_data_frame$GDP)
gdp_data_frame$GDP = round(gdp_data_frame$GDP / 1e6, 1)  #Converting it to m
write.csv(gdp_data_frame, ’gdp.csv’, row.names=FALSE)


#fertility rate
countries = fertility_rate$Country

fertility_rate_data_frame <- data.frame(matrix(ncol = 4, nrow= 0))
column_names <- c("Country", "Decade", "Year", "Fertility␣Rate")
colnames(fertility_rate_data_frame) <- column_names

for (i in 2: ncol(fertility_rate)) {
  lengthOfDataFrame = nrow(fertility_rate_data_frame)
  for (j in 1 : length(countries)) {
    Year = as.numeric(gsub("[^0-9.]", "",  colnames( fertility_rate )[i]))
    Decade = Year - (Year %% 10)
    fertility_rate_data_frame[ lengthOfDataFrame + j, ] = c(fertility_rate$C
  }
}
write.csv(fertility_rate_data_frame, ’fertility_rate.csv’, row.names=FALSE)

util.R:
instr <- function(str1,str2,startpos=1,n=1){
  aa=unlist(strsplit(substring(str1,startpos),str2))
  if(length(aa) < n+1 ) return(0);
  return(sum(nchar(aa[1:n])) + startpos+(n-1)*nchar(str2) )
}

numextract <- function(string){
  str_extract(string, "\\-*\\d+\\.*\\d*")
}

trim <- function (x) gsub("^\\s+|\\s+$", "", x)

#SQL queries

#Drop primary key foreign key relationships
BEGIN
ALTER TABLE Fact
DROP CONSTRAINT Gender_ID, Location_ID, Date_ID
END;

#Clear data from table

BEGIN
TRUNCATE TABLE Raw_Population
TRUNCATE TABLE Raw_Fertility_Rate
TRUNCATE TABLE Raw_GDP
TRUNCATE TABLE Raw_Life_Expectancy_Rate
TRUNCATE TABLE Raw_Literacy_Rate
TRUNCATE TABLE Raw_Infant_Mortality
TRUNCATE TABLE Raw_Location
TRUNCATE TABLE Dim_Gender
TRUNCATE TABLE Dim_Date
TRUNCATE TABLE Dim_Location
TRUNCATE TABLE Fact
 END;

#Add primary key foreign key relationships
BEGIN
ALTER TABLE Fact
ADD CONSTRAINT Date_ID
FOREIGN KEY (Date_ID) REFERENCES Dim_Date(Date_ID)

ALTER TABLE Fact
ADD CONSTRAINT Location_ID
FOREIGN KEY (Location_ID) REFERENCES Dim_Location(Location_ID)

ALTER TABLE Fact
ADD CONSTRAINT Gender_ID
FOREIGN KEY (Gender_ID) REFERENCES Dim_Gender(Gender_ID)
END;

#Create table
CREATE TABLE [dbo].[Dim_Date](
        [Date_ID] [int] IDENTITY(1,1) NOT NULL,
        [Decade] [varchar](50) NULL,
        [Year] [varchar](50) NULL
        )

CREATE TABLE [dbo].[Dim_Gender](
        [Gender_ID] [int] IDENTITY(1,1) NOT NULL,
        [Gender] [varchar](50) NULL
        )

CREATE TABLE [dbo].[Dim_Location](
        [Location_ID] [int] NOT NULL,
        [Country] [varchar](50) NULL,
        [Region] [varchar](50) NULL,
        [Continent] [varchar](50) NULL
        )

CREATE TABLE [dbo].[Fact](
        [Fact_ID] [int] IDENTITY(1,1) NOT NULL,
        [Date_ID] [int] NULL,
        [Location_ID] [int] NULL,
        [Gender_ID] [int] NULL,
        [Literacy Rate] [numeric](28, 0) NULL,
        [Fertility Rate] [numeric](18, 3) NULL,
        [GDP] [numeric](28, 3) NULL,
        [Infant Mortality] [numeric](18, 2) NULL,
        [Life Expectancy Rate] [numeric](28, 2) NULL,
        [Worldwide Life Expectancy Rate] [numeric](28, 2) NULL,
        [Population] [numeric](18, 2) NULL)

CREATE TABLE [dbo].[Raw_Fertility_Rate](
        [Country] [varchar](50) NULL,
        [Decade] [varchar](50) NULL,
        [Year] [varchar](50) NULL,
        [Fertility Rate] [numeric](18, 3) NULL
)

CREATE TABLE [dbo].[Raw_GDP](
        [Country] [varchar](50) NULL,
        [Decade] [varchar](50) NULL,
        [Year] [varchar](50) NULL,
        [GDP] [decimal](28, 3) NULL
)

CREATE TABLE [dbo].[Raw_Infant_Mortality](
        [Country] [varchar](50) NULL,
        [Infant Mortality] [numeric](18, 2) NULL
)

CREATE TABLE [dbo].[Raw_Life_Expectancy_Rate](
        [Country] [varchar](50) NULL,
        [Decade] [varchar](50) NULL,
        [Year] [varchar](50) NULL,
        [Life Expectancy Rate] [decimal](28, 2) NULL,
        [Worldwide Life Expectancy Rate] [decimal](28, 2) NULL
)

CREATE TABLE [dbo].[Raw_Literacy_Rate](
        [Country] [varchar](50) NULL,
        [Gender] [varchar](50) NULL,
        [Literacy Rate] [decimal](28, 2) NULL
)

CREATE TABLE [dbo].[raw_Location](
        [Location_ID] [int] IDENTITY(1,1) NOT NULL,
        [Country] [varchar](50) NULL,
        [Region] [varchar](50) NULL,
        [Continent] [varchar](50) NULL)

CREATE TABLE [dbo].[Raw_Population](
        [Country] [varchar](50) NULL,
        [Decade] [varchar](50) NULL,
        [Year] [varchar](50) NULL,
        [Population] [numeric](18, 2) NULL
)
```

```
#References for the code :
#https://rdrr.io/cran/rvest/man/html_table.html
#http://stla.github.io/stlapblog/posts/Numextract.html
#https://stackoverflow.com/questions/14249562/find-location-of-character-in-
#https://stackoverflow.com/questions/29478584/intersection-of-multiple-dataf
#https://stackoverflow.com/questions/11254524/omit-rows-containing-specific-
#https://stackoverflow.com/questions/29478584/intersection-of-multiple-dataf
#http://rprogramming.net/rename-columns-in-r/
#https://github.com/Kaggle/kaggle-api
#https://stackoverflow.com/questions/33203800/unzip-a-zip-file
#https://stackoverflow.com/questions/10758965/how-do-i-rename-files-using-r
#https://stackoverflow.com/questions/14219887/how-to-delete-a-file-with-r
#https://stackoverflow.com/questions/11936339/replace-specific-characters-wi
#https://stackoverflow.com/questions/14249562/find-location-of-character-in-
#https://stackoverflow.com/questions/19252663/extracting-decimal-numbers-fro
#https://stackoverflow.com/questions/2261079/how-to-trim-leading-and-trailin
#https://www.google.com/url?q=https://stackoverflow.com/questions/47510141/g
#https://stackoverflow.com/questions/14096814/merging-a-lot-of-data-frames
#https://stackoverflow.com/questions/35352914/floor-a-year-to-the-decade-in-
```

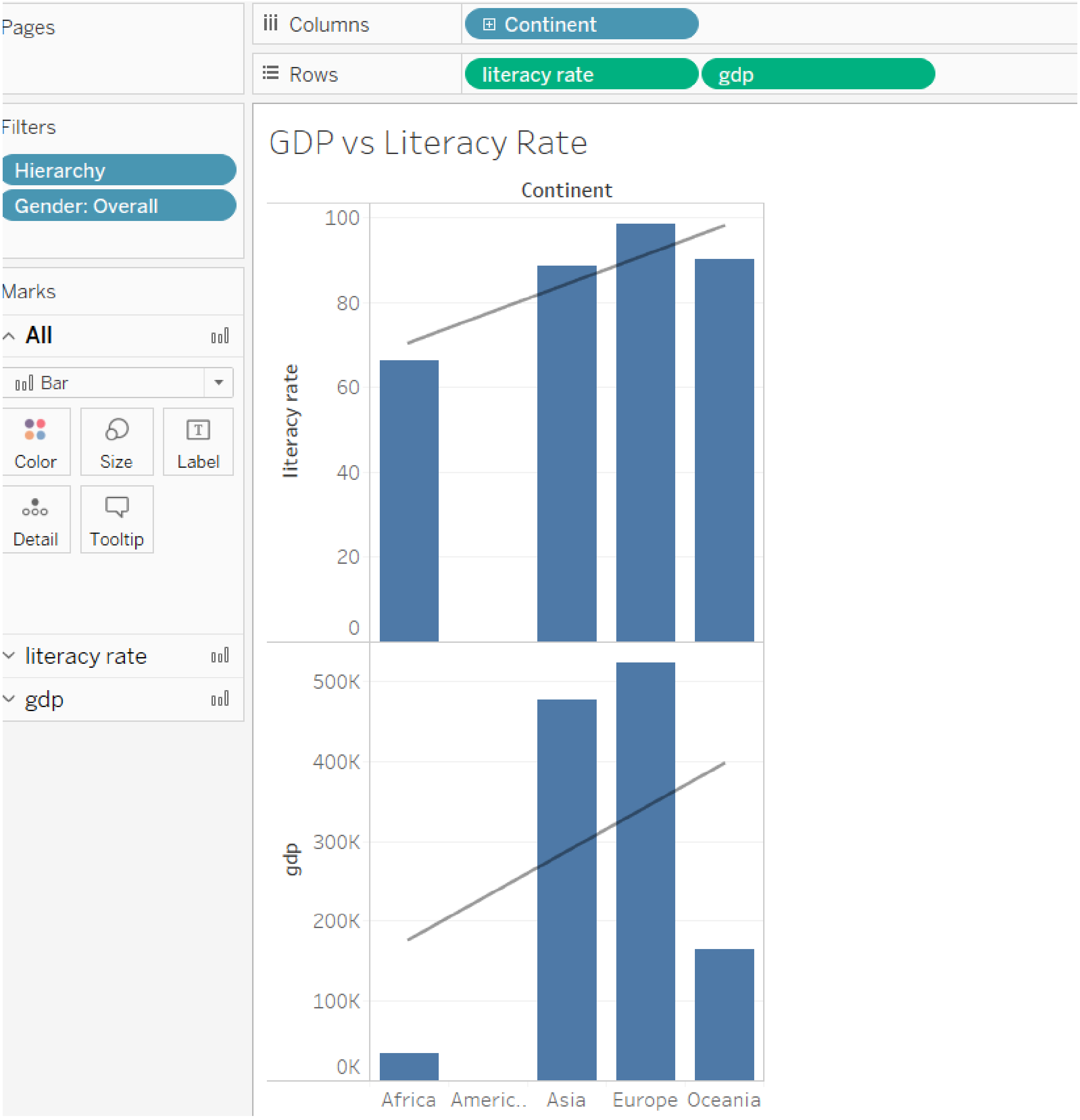
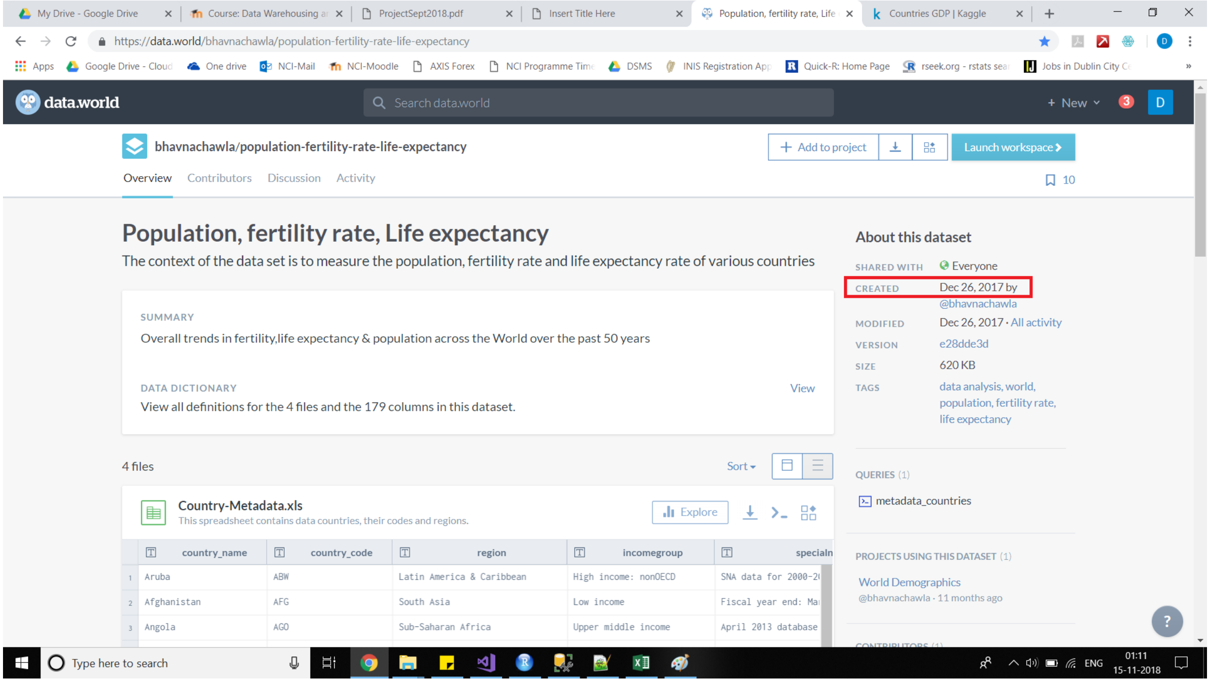


Figure 4: Results for BI Query 3



bhavnachawla/population-fertility-rate-life-expectancy

Population, fertility rate, Life expectancy

The context of the data set is to measure the population, fertility rate and life expectancy rate of various countries

SUMMARY

Overall trends in fertility,life expectancy & population across the World over the past 50 years

DATA DICTIONARY

View all definitions for the 4 files and the 179 columns in this dataset.

4 files

Country-Metadata.xls

This spreadsheet contains data countries, their codes and regions.

	country_name	country_code	region	incomegroup	specialn
1	Aruba	ABW	Latin America & Caribbean	High income: nonOECD	SNA data for 2000-2(
2	Afghanistan	AFG	South Asia	Low income	Fiscal year end: Ma
3	Angola	AGO	Sub-Saharan Africa	Upper middle income	April 2013 database

About this dataset

SHARED WITH Everyone

CREATED Dec 26, 2017 by @bhavnachawla

MODIFIED Dec 26, 2017 · All activity

VERSION e28dde3d

SIZE 620 KB

TAGS data analysis, world, population, fertility rate, life expectancy

QUERIES (1)

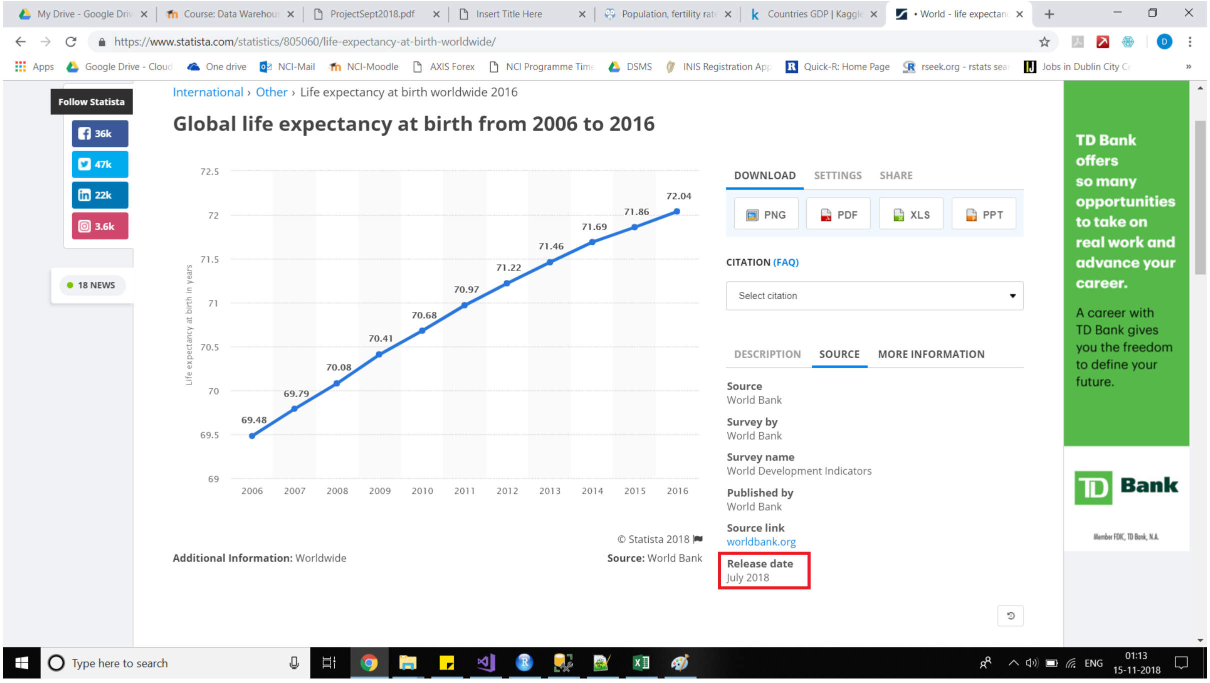
metadata_countries

PROJECTS USING THIS DATASET (1)

World Demographics
@bhavnachawla · 11 months ago

CONTRIBUTORS (1)

Figure 5: Data.World Released Date



International › Other › Life expectancy at birth worldwide 2016

Global life expectancy at birth from 2006 to 2016

Additional Information: Worldwide

Source: World Bank

CITATION (FAQ)

Source
World Bank

Survey by
World Bank

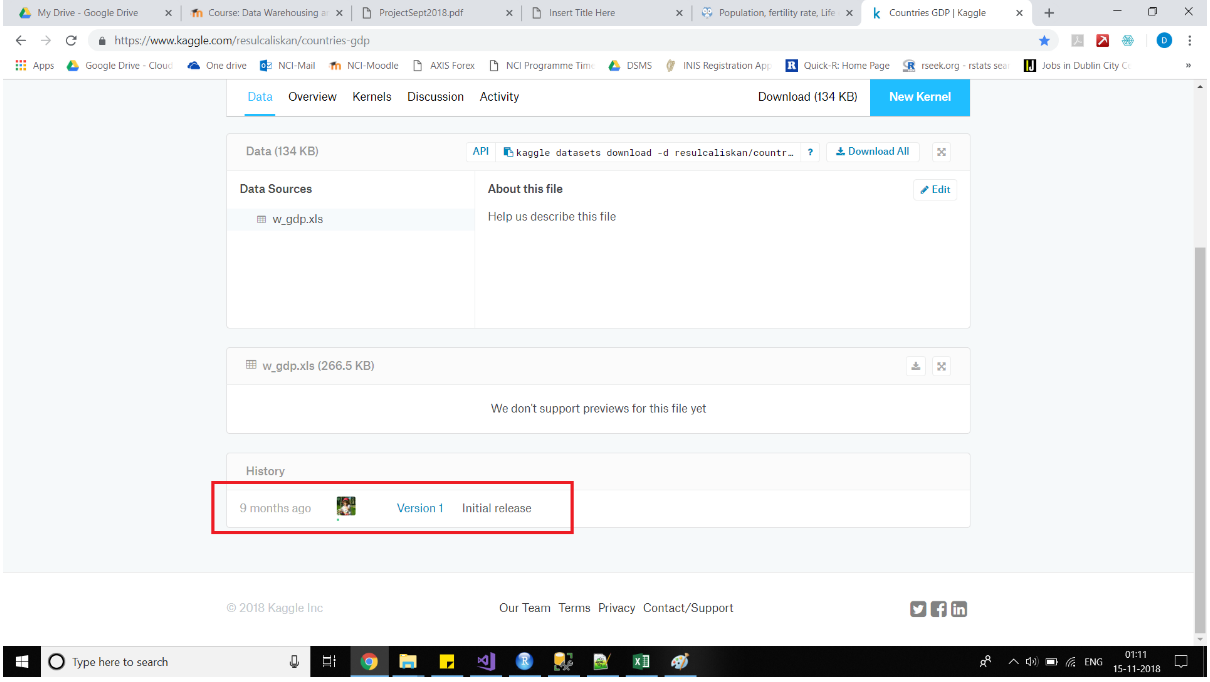
Survey name
World Development Indicators

Published by
World Bank

Source link
worldbank.org

Release date
July 2018

Figure 6: Statista Released Date



https://www.kaggle.com/resulcaliskan/countries-gdp

Data Overview Kernels Discussion Activity Download (134 KB) New Kernel

Data (134 KB) API `kaggle datasets download -d resulcaliskan/countr…` ? Download All

Data Sources

w_gdp.xls

About this file

Help us describe this file

w_gdp.xls (266.5 KB)

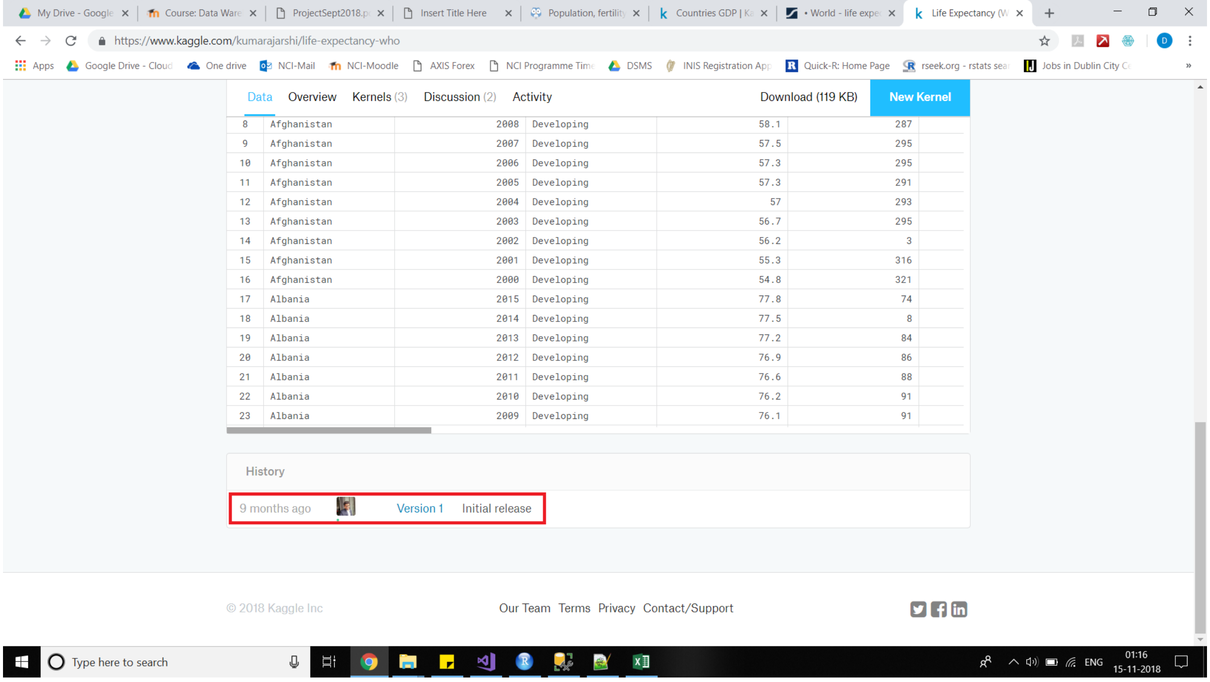
We don't support previews for this file yet

History

9 months ago Version 1 Initial release

© 2018 Kaggle Inc Our Team Terms Privacy Contact/Support

Figure 7: Kaggle GDP dataset Released Date



https://www.kaggle.com/kumarajarshi/life-expectancy-who

Data Overview Kernels (3) Discussion (2) Activity Download (119 KB) New Kernel

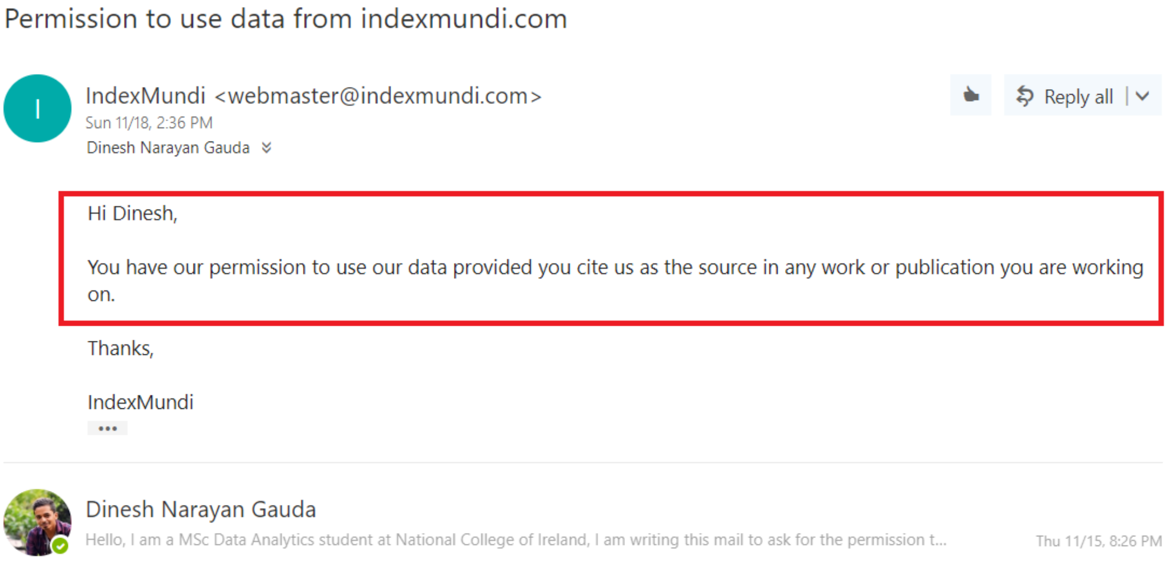
8	Afghanistan	2008	Developing	58.1	287
9	Afghanistan	2007	Developing	57.5	295
10	Afghanistan	2006	Developing	57.3	295
11	Afghanistan	2005	Developing	57.3	291
12	Afghanistan	2004	Developing	57	293
13	Afghanistan	2003	Developing	56.7	295
14	Afghanistan	2002	Developing	56.2	3
15	Afghanistan	2001	Developing	55.3	316
16	Afghanistan	2000	Developing	54.8	321
17	Albania	2015	Developing	77.8	74
18	Albania	2014	Developing	77.5	8
19	Albania	2013	Developing	77.2	84
20	Albania	2012	Developing	76.9	86
21	Albania	2011	Developing	76.6	88
22	Albania	2010	Developing	76.2	91
23	Albania	2009	Developing	76.1	91

History

9 months ago Version 1 Initial release

© 2018 Kaggle Inc Our Team Terms Privacy Contact/Support

Figure 8: Kaggle Life Expectancy Rate dataset Released Date



Permission to use data from indexmundi.com

IndexMundi <webmaster@indexmundi.com>
Sun 11/18, 2:36 PM
Dinesh Narayan Gauda

Hi Dinesh,

You have our permission to use our data provided you cite us as the source in any work or publication you are working on.

Thanks,

IndexMundi

Dinesh Narayan Gauda
Hello, I am a MSc Data Analytics student at National College of Ireland, I am writing this mail to ask for the permission t...
Thu 11/15, 8:26 PM

Figure 9: Permission to scarpe data from Indexmundi.com