# Data Warehousing and Business Intelligence Project

on

IMPACT OF TOURISM ON GROWTH OF A COUNTRY

VARUN GOWDA

18129129

MSc/PGDip Data Analytics – 2019/20

Submitted to: Sean Heeney

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

| | |
|---|---|
| **Student Name:** | Varun Gowda |
| **Student ID:** | 18129129 |
| **Programme:** | MSc Data Analytics |
| **Year:** | 2019/20 |
| **Module:** | Data Warehousing and Business Intelligence |
| **Lecturer:** | Sean Heeney |
| **Submission Due Date:** | 12/04/2019 |
| **Project Title:** | IMPACT OF TOURISM ON GROWTH OF A COUNTRY |

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:** | April 12, 2019 |

**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

# IMPACT OF TOURISM ON GROWTH OF A COUNTRY

VARUN GOWDA
18129129

12/04/2019

## 1 Introduction

Tourism has a positive impact on the economic growth of a country. It plays a major role in contribution to employment, social stability and improvement in a country. In 2014, it was seen that global tourism made up to 3.7 percent of the worlds GDP, that is roughly around US 2.5 billion dollars.

There is a five percent annual increase in number of international tourists worldwide. When a country is frequently visited by tourists, it is seen that there is an increase in jobs, and there are overall better living conditions for the locals in that country. Tourism encourages governments and concerned organizations to maintain the historical places, tourist attractions in good condition. Studies have shown that directly or indirectly, travel and tourism create nearly eleven percent of the total jobs in the world.

Data Warehouse is a relational database which holds structured data that is designed for query and analysis. It maintains records of data, analyzing the data to gain a better understanding of the business and to improve the business. The repository can be queried for multiple business questions and we can come up a summary or a forecast from the data warehouses data.

Business Intelligence is a technology driven process for analyzing data, and presenting actionable information to help corporate executives, governments, business owners and other end users make more informed business decisions. Data Warehousing involves getting data from multiple sources, does the ETL workflow and inputs the data in data warehouse in a standard format. Business Intelligence starts from here, where we report the data, analyze the data and create visualizations. Business intelligence usually refers to the information that is available to the enterprise to make decisions on different aspects of the enterprise like profit scale, data comparison, visualization of data, analysis etc.

In my project, I have tried to analyze the economic conditions of countries which are most visited by international tourists. An attempt is been made to study the impact of tourism on growth of a country

## 2 Data Sources

Data has to be extracted from different sources, and then it must be all loaded into a single staging process using ETL workflow. Data can be sourced from many sources, it can be structured or unstructured. All these data files have to be made into one standard format and brought into the data warehouse. In this project I have used three structured datasets and one unstructured dataset.
The datasets are:

1. Structured dataset:
https://tcdata360.worldbank.org/indicators/govt.tat.spend?indicator=24661viz=line$_{c}hartyears = 1995, 2028$

The above dataset gives us data of countries, their tourism expenditures, growth percentage, percentage share of GDP and similar other information.

2. Unstructured dataset:
https://en.wikipedia.org/wiki/World$_{T}ourism_{r}ankings$

The above dataset gives us information on world tourism rankings. It is compiled by United Nations World Tourism Organization. It contains information of countries in order of number of international visitor arrivals, income generated by inbound tourism, and expenditure of outbound tourists.

3. Structured dataset (Statista):
https://www.statista.com/statistics/261726/countries-ranked-by-number-of-international-tourist-arrivals/ The above dataset consists of information of the countries with the largest number of tourist arrivals in the year 2017. These countries are shortlisted in the first dataset and all the relevant information with respect to these countries are extracted from the first dataset.

4. Structured dataset:
https://datacatalog.worldbank.org/dataset/gdp-ranking The above dataset consists of information of GDP of the countries. It is a time series type of data and has year to year changes in the level of income of an economy. It has information related to economic growth of a country.

TECHNOLOGIES USED:
1. R STUDIO is used for extracting information from unstructured data and also data cleaning.
2. SSMS and SSIS for creation fact and dim tables.
3. Tableau is used as a visualization tool.

# 3 Related Work

There has been a lot of work and research conducted to show the direct relation between contribution of tourism and a countrys economic growth. The main reason why governments and associations support or engage in promotion of tourism is because of its impact on the countrys development. Tourism creates employment, generates revenue, and the overall economic transactions are increased. GDP is always almost taken as a measure to check how much tourism contributes to a growth of a country economically. There are many articles which propose various methods to measure the consequences of tourism on the economy.

One example of a research to study the tourism impact on economic development was conducted by Proenca and Soukiazis (2005). They find in their study that if one percent increases in accommodation capacity in tourism field in Portuguese regions increases 0.01 percent in per capita income.

The impact of tourism on economy must be studied regionally, nationally and internationally. Local communities need to understand the impact it has on their region. There is a range of methods, which vary from just guess work to complicated math models which are used to estimate tourism impact. Studies differ in terms of quality and accuracy rate. Also, it depends on what features of tourism are involved. Tourism is a relatively new field in global economic trades. In the past decade, it is a major contributor to the foreign income sources of many nations. It also plays a vital role in the economic, cultural and social development of many nations.

What is found in common referring to researches undertaken to study the impact of tourism on a country is that the impact is generally positive. It triggers the economic growth. I found out that tourism is a vey vital part of an economic activity of many countries which are developed or developing. Many countries focus a lot of money and energy towards the development of tourism. It is also seen that tourism is used by international agencies and other related institutions as a key factor to ease poverty. A study conducted in Kenya shows that if there is a steady 5 percent increase in number of arrivals in Kenya, it will empower 1.83 percent of the local population to come out of the poverty line. The total average household income had also increased.

I have tried to gauge the impact the tourism creates on growth of countries in terms of GDP value, GDP ranks, growth ranks of countries, etc.

# 4 Data Model

RALPH KIMBALL APRROACH: This type of approach suggests we start constructing several data marts that perform the analytical needs of the departments followed by merging the data marts for consistency through an information bus. Therefore, this type of approach received the bottom up tittle. Kimball believes that various data marts that store information in dimensional models to address quickly the requirements of various departments and various areas of the enterprise data. Kimball trusts that creating in multi-dimensional model which is star schema and snowflakes is the best way to approach the data model. The users can understand the data, analyze it, aggregate and traverse the data inconsistencies in an easy way. To add to this, data marts are defined as atomic and aggregated that both use dimensional model. This type of approach stresses on integration for consistent results. For my project, I am using Kimballs approach.

DATA MODELLING PROCESS: We can see below the star schema of my project. There is a basic assumption made that facts present in data have very high volume, so we attempt to make them narrow. Here in the star schema, each dimension is de-normalized into one table. By den-normalization we get simpler queries and very fast aggregations.

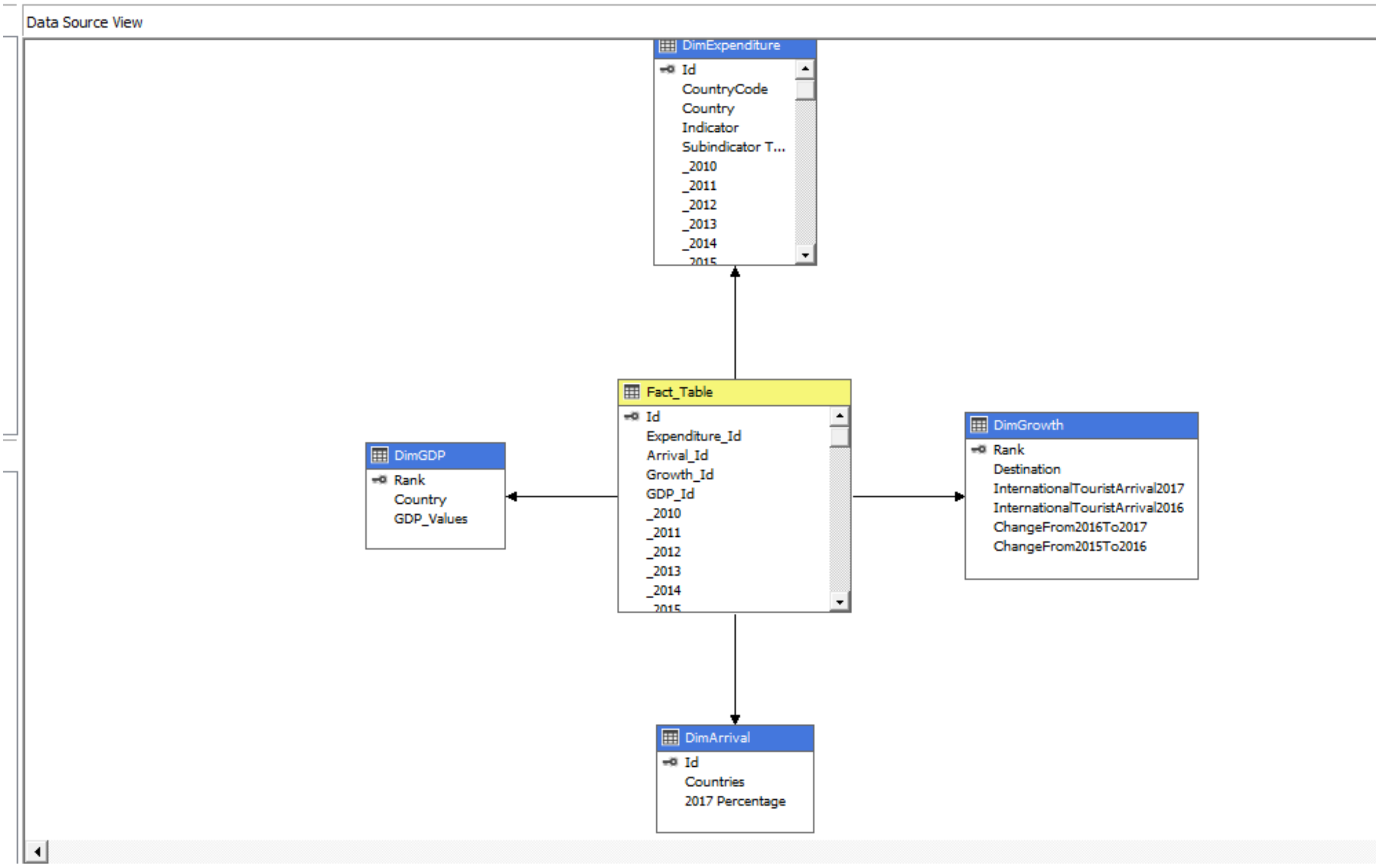


Figure 1: STAR SCHEMA

The dimension tables are de-normalized heavily so that joins can be removed from the database because of the reason that they are not effective or efficient when we try to run a query against it. The basic methodology of the star schema is that we try to take out as much joins as we can by de-normalizing the dimension and make the largest table (fact table) as normalized as we can. The star schema type of approach is widely used to construct data warehouse and dimensional data marts. It is most commonly used because it is very easily understandable and end users can navigate and understand the subject before making a query.

The star schema consists of fact table and dimension tables. The dimensions are only connected to the central fact table. Facts may change with time, but dimensions are constant or even if they change, they do so at a very slow rate. The fact table is in the center and has entities like Arrival_Id , Growth_Id, GDP_Id, and expenditure values of years.

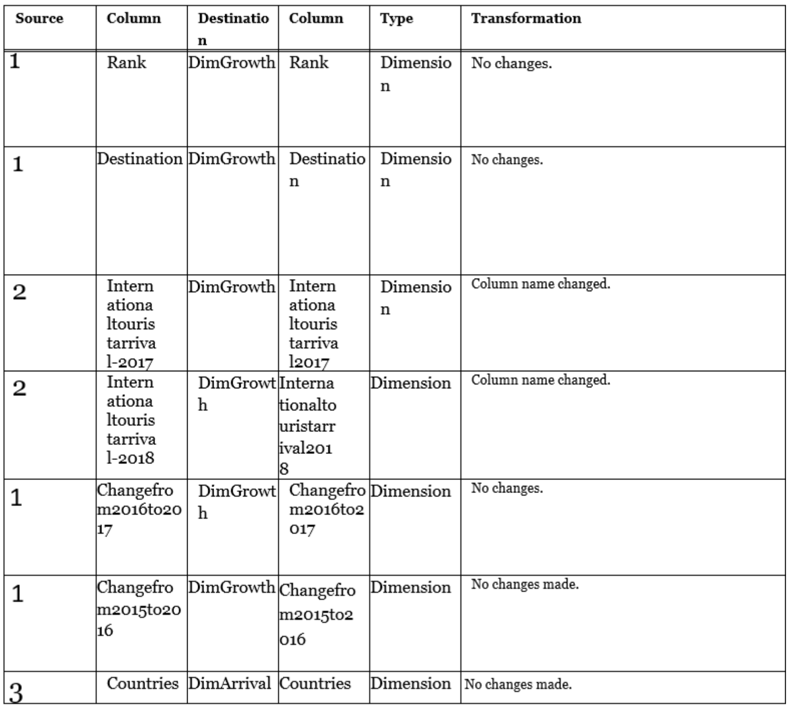
I have four dimension tables which are DimExpenditure, DimArrival, DimGDP, DimGrowth. All these dimension tables provide factors to the fact table. In DimExpenditure the factors present are country, and the expenditure value of each year starting from 2010. In DimArrival the factors are arrival countries and international tourist arrival percentage of 2017. In DimGDP, as the name suggests there is country wise GDP values. In DimGrowth the factors are growth ranks of countries, international tourist arrivals of 2016 and 2017. Percentage change in tourists from 2015 and 2016 and also from 2016 and 2017. All these are linked to the fact table in the center.

Each dimension table has an Id (primary key) which is linked to one of the columns in the fact table. The fact tables generally have some numeric value or records, and all the descriptive information is stored in our dimensional tables. Fact tables are generally large because they contain granular level of data, whereas dimension tables are usually small compared to fact tables as they consist of attributes or a small record of data. But each attribute can relate or explain to a large number of data present in fact tables.

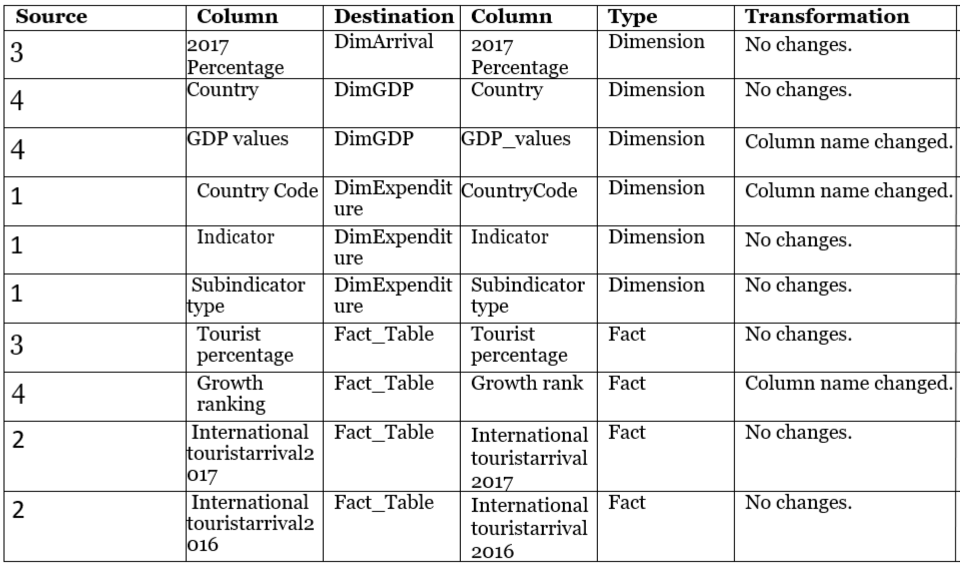
# 5 Logical Data Map

In this section, we describe our logical data map, i.e. how every row of every data source is handled and information about any changes that took place.

Logical Data Map describing all transformations, sources and destinations for all components of the data model illustrated here

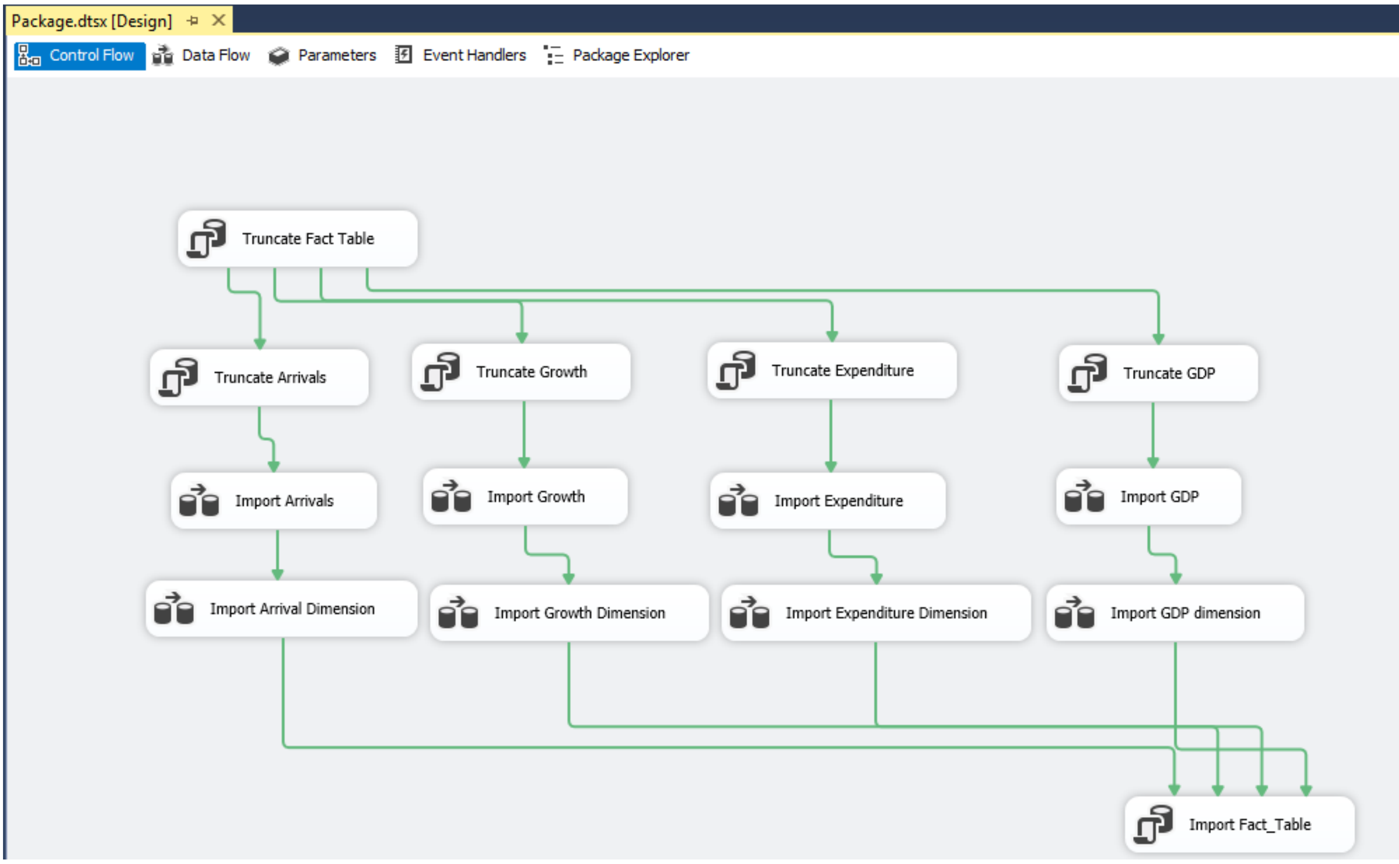


| Source | Column | Destination | Column | Type | Transformation |
|---|---|---|---|---|---|
| 1 | Rank | DimGrowth | Rank | Dimension | No changes. |
| 1 | Destination | DimGrowth | Destination | Dimension | No changes. |
| 2 | Internationaltouristarrival-2017 | DimGrowth | Internationaltouristarrival2017 | Dimension | Column name changed. |
| 2 | Internationaltouristarrival-2018 | DimGrowth | Internationaltouristarrival2018 | Dimension | Column name changed. |
| 1 | Changefrom2016to2017 | DimGrowth | Changefrom2016to2017 | Dimension | No changes. |
| 1 | Changefrom2015to2016 | DimGrowth | Changefrom2015to2016 | Dimension | No changes made. |
| 3 | Countries | DimArrival | Countries | Dimension | No changes made. |

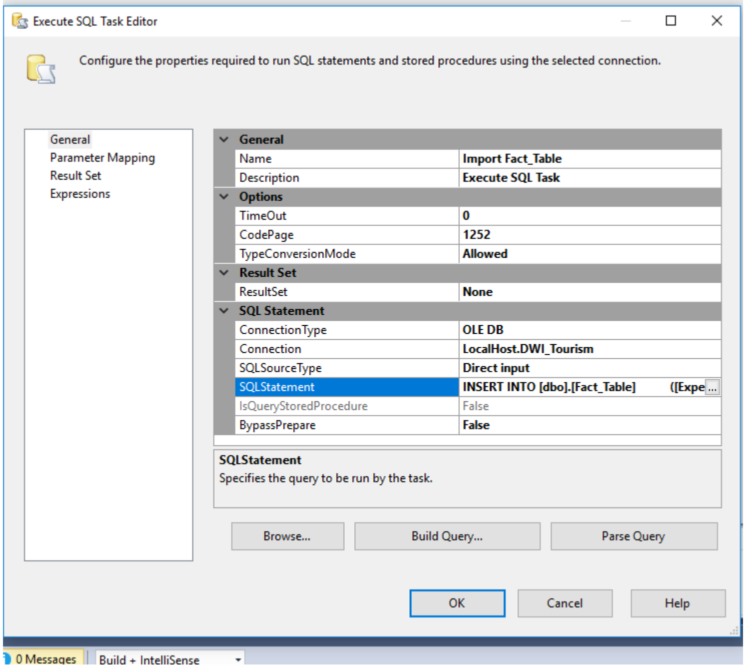


| Source | Column | Destination | Column | Type | Transformation |
|---|---|---|---|---|---|
| 3 | 2017 Percentage | DimArrival | 2017 Percentage | Dimension | No changes. |
| 4 | Country | DimGDP | Country | Dimension | No changes. |
| 4 | GDP values | DimGDP | GDP_values | Dimension | Column name changed. |
| 1 | Country Code | DimExpenditure | CountryCode | Dimension | Column name changed. |
| 1 | Indicator | DimExpenditure | Indicator | Dimension | No changes. |
| 1 | Subindicator type | DimExpenditure | Subindicator type | Dimension | No changes. |
| 3 | Tourist percentage | Fact_Table | Tourist percentage | Fact | No changes. |
| 4 | Growth ranking | Fact_Table | Growth rank | Fact | Column name changed. |
| 2 | International touristarrival2017 | Fact_Table | International touristarrival 2017 | Fact | No changes. |
| 2 | International touristarrival2016 | Fact_Table | International touristarrival 2016 | Fact | No changes. |

# 6 ETL Process



The above picture is the structure of the ETL workflow. As we can see, truncate function is used as the first SQL query. Truncate function deletes all the values present in the database. It is then followed by the staging process where the cleaned datasets are staged, in the sense raw tables are built. After staging, the data is now imported to tables which are created. Below we can see the SQL statement used to import fact table.

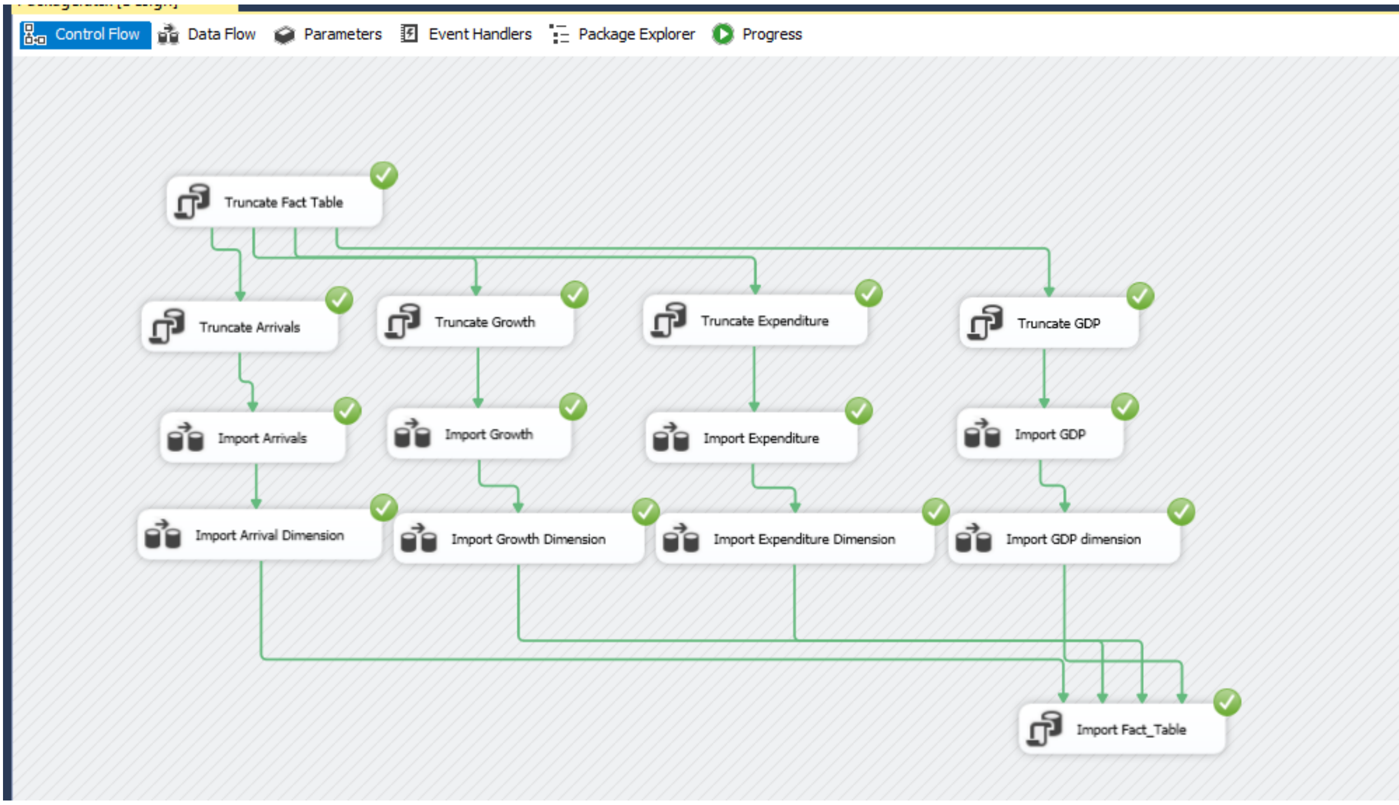


Configure the properties required to run SQL statements and stored procedures using the selected connection.

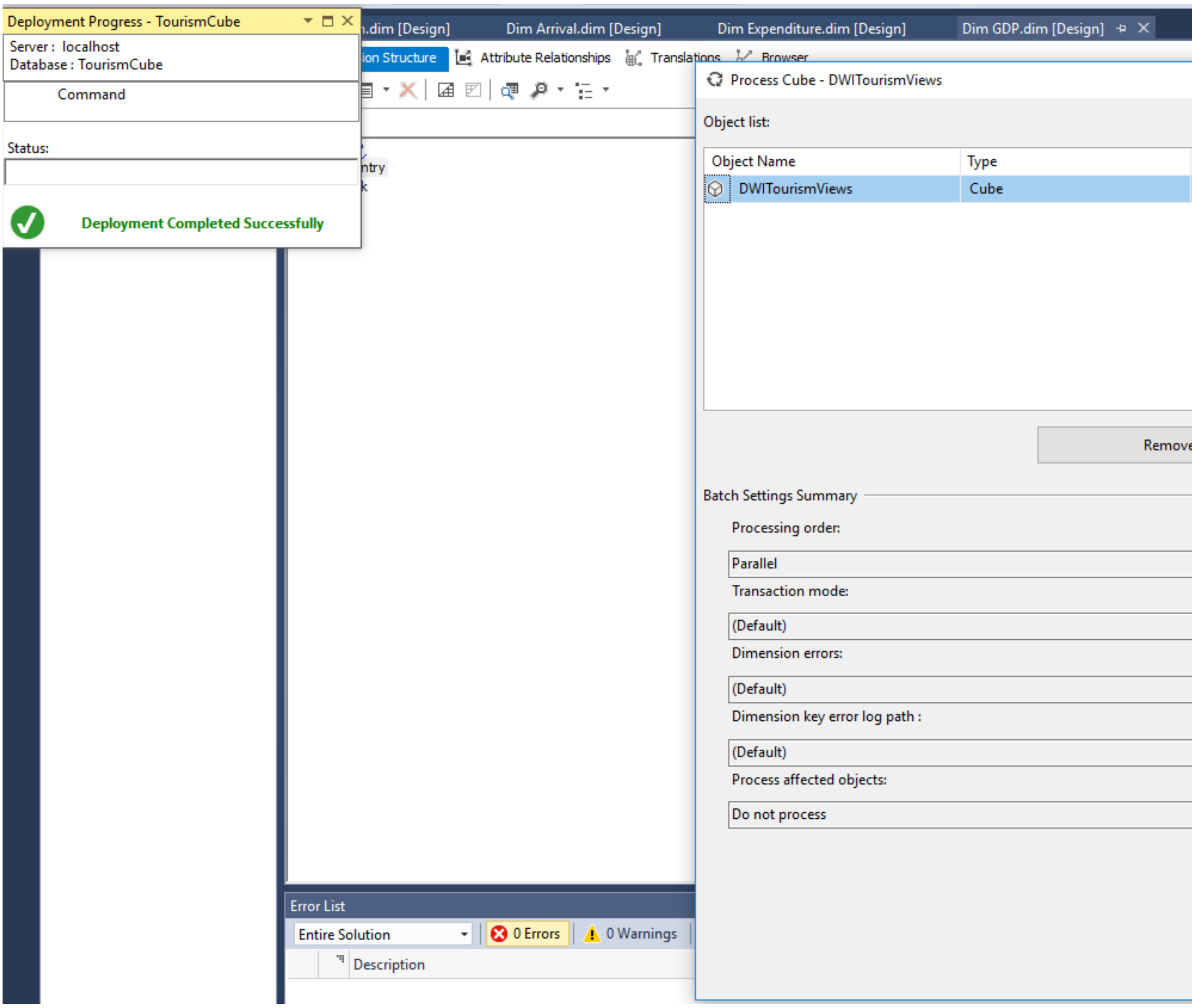
| | |
|---|---|
| **General** | |
| Name | Import Fact_Table |
| Description | Execute SQL Task |
| **Options** | |
| TimeOut | 0 |
| CodePage | 1252 |
| TypeConversionMode | Allowed |
| **Result Set** | |
| ResultSet | None |
| **SQL Statement** | |
| ConnectionType | OLE DB |
| Connection | LocalHost.DWI_Tourism |
| SQLSourceType | Direct input |
| SQLStatement | INSERT INTO [dbo].[Fact_Table] ([Expe |
| IsQueryStoredProcedure | False |
| BypassPrepare | False |

**SQLStatement**
Specifies the query to be run by the task.

Now the dimension tables are created so that only the factors required for the query is extracted. After creating dim tables, data is now loaded into them. Data from source file can be added to database, these are then used to create dimensions in the fact table. The final step of the workflow would be the fact table creation. Truncate function will also be used here to avoid duplication or corruption of data. Every section of the ETL process is automated.



The automated process aids in creation of cube in SSIS which now needs to be deployed. We now move on to deployment of cube using Visual Studio.



Deployment Progress - TourismCube

Server : localhost
Database : TourismCube

Command

Status:

Deployment Completed Successfully

Process Cube - DWITourismViews

Object list:

| Object Name | Type |
|---|---|
| DWITourismViews | Cube |

Batch Settings Summary

Processing order: Parallel

Transaction mode: (Default)

Dimension errors: (Default)

Dimension key error log path : (Default)

Process affected objects: Do not process

Error List

Entire Solution | 0 Errors | 0 Warnings

We process the cube in visual studio and now we can see that cube deployment is completed successfully. Now we can get to the visualization section to visualize the data so that it can answer business queries.

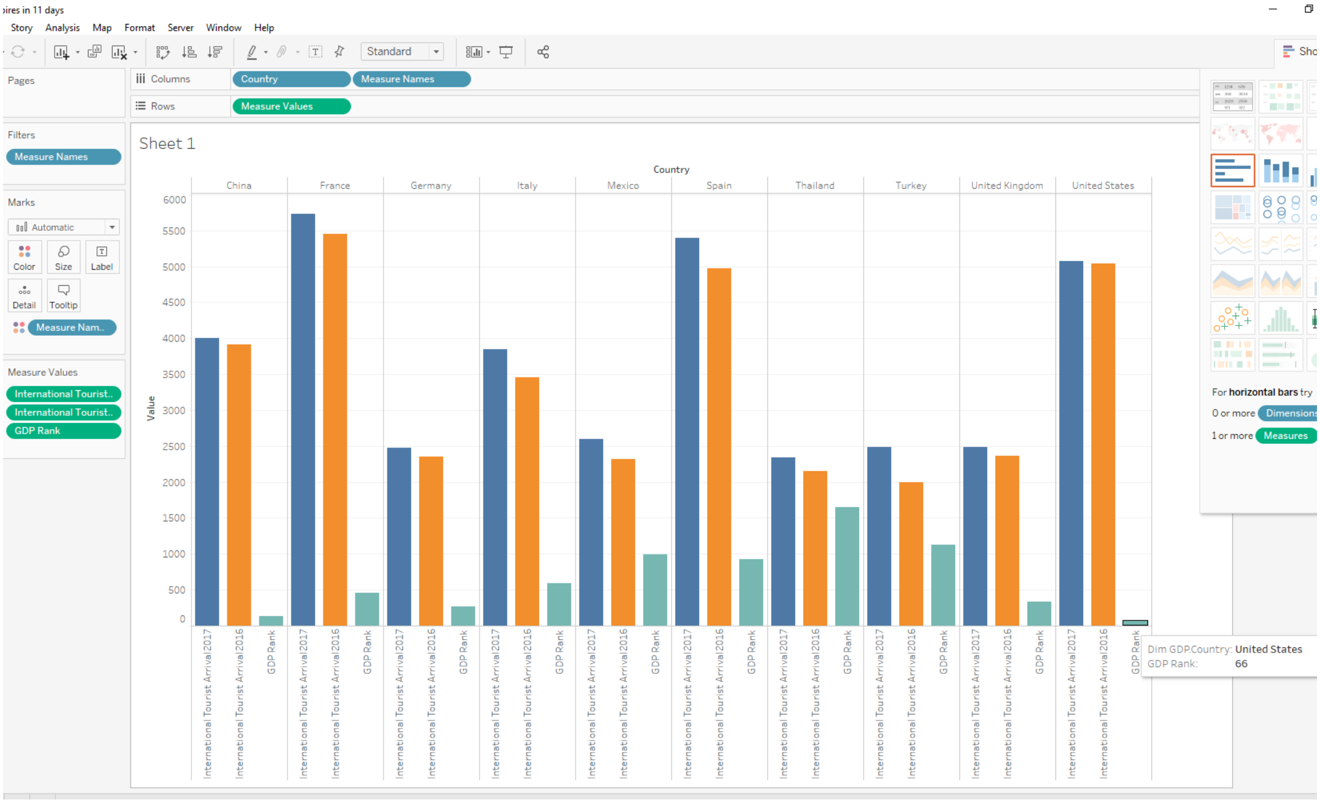
# 7 Application

Tableau is used here for visualization purpose as it is very simple to use, and the data can be easily understood.

## 7.1 BI Query 1: Here we are comparing the international tourist arrivals of 2016 and 2017 of the countries with their GDP ranks.

We can see from the visualization that the countries with a greater number of international tourist arrivals have better GDP ranks. For example, like France, Spain, United States, China, Italy have better GDP ranks than countries like Thailand, Mexico, Turkey.

This shows that number of tourists visiting a country has a direct positive impact on the GDP ranks of a country.



Columns: Country, Measure Names

Rows: Measure Values

Filters: Measure Names

Marks: Automatic; Color: Measure Nam..

Measure Values: International Tourist.., International Tourist.., GDP Rank

Dim GDPCountry: United States
GDP Rank: 66

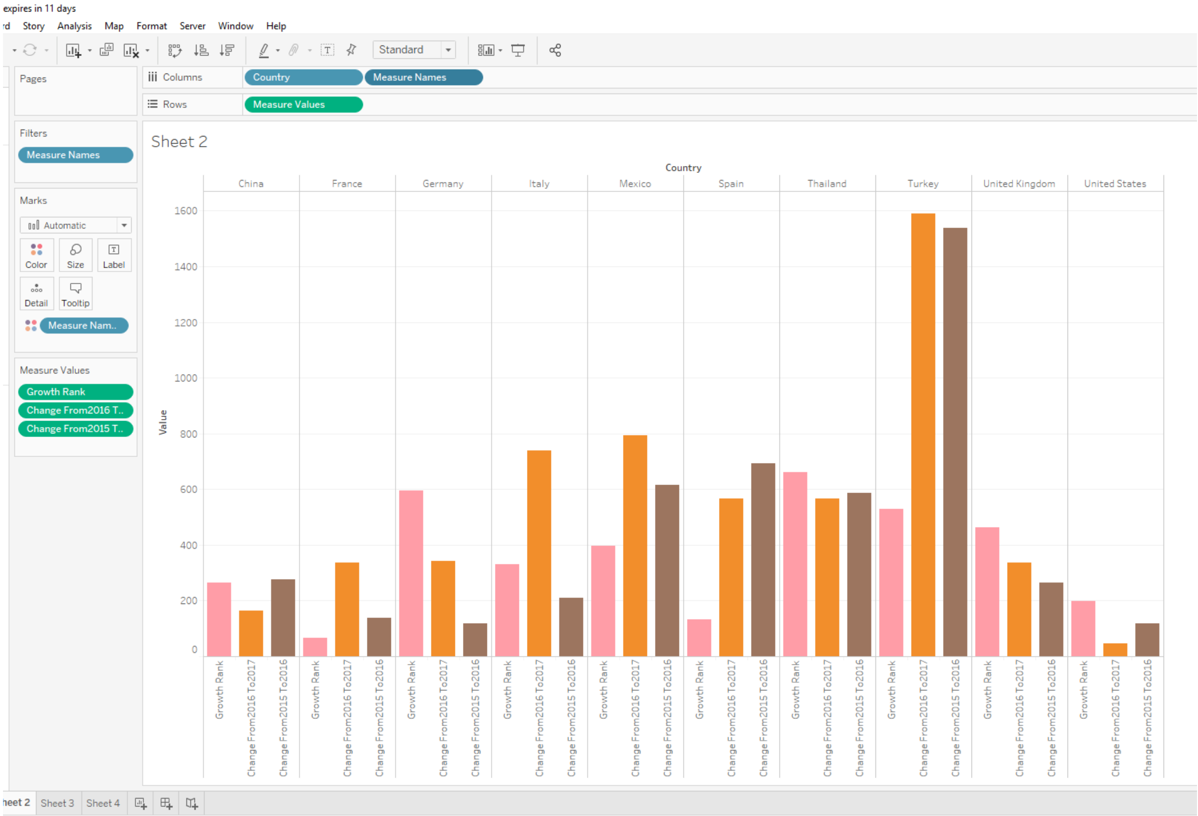
For horizontal bars try
0 or more Dimensions
1 or more Measures

Results for BI Query 1

## 7.2 BI Query 2: . . .

In this visualization, we compare percentage change in tourist arrivals from 2015 to 2016, percentage change in tourist arrivals from 2016 to 2017 with the growth ranks of the countries.

We can see that If we compare the change in percentage of tourists of 2015 and 2016 and that of 2016 and 2017. The growth rank is better for countries which have had an increase in percentage of tourists from the previous year. We can see that in countries such as Turkey, France, Italy, United Kingdom there is a positive change in number of tourists in the year 2017 than in 2016 and the growth ranks are also better. This suggests that a positive influx of tourists means the growth rank of a country will also be better.



expires in 11 days

Columns: Country, Measure Names

Rows: Measure Values

Filters: Measure Names

Marks: Automatic

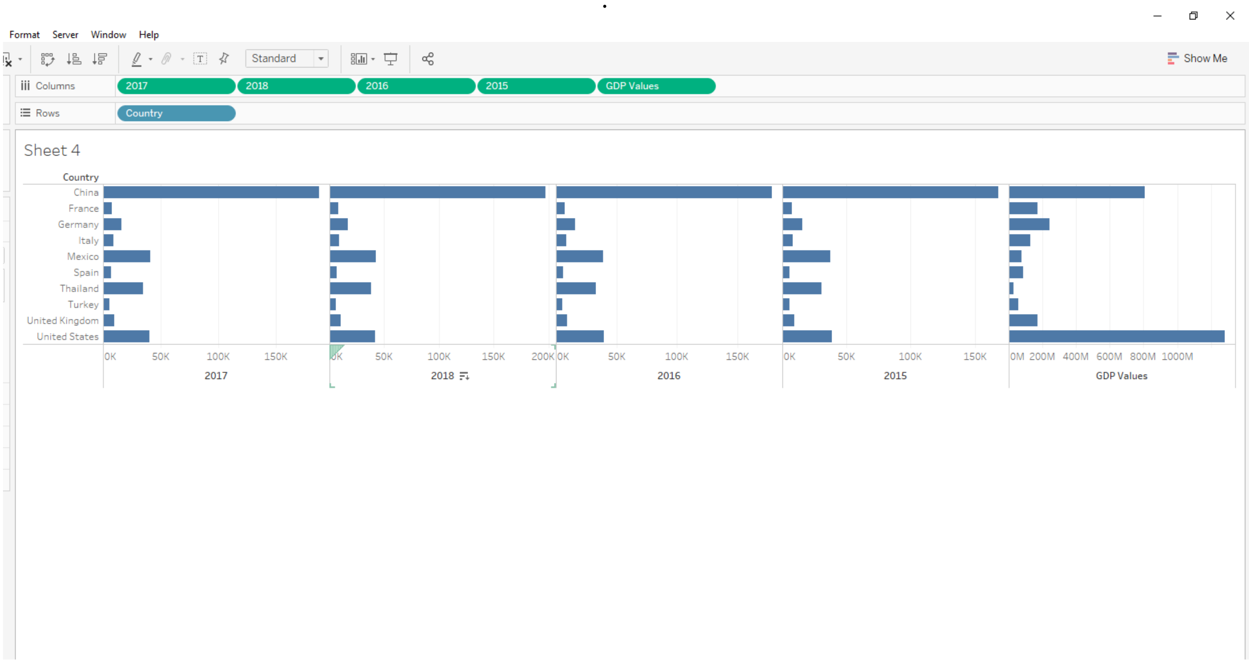
Measure Values: Growth Rank, Change From2016 T.., Change From2015 T..

Results for BI Query 2

## 7.3 BI Query 3: . . .

Here, we compare the expenditure values of countries in the year 2015, 2016, 2017, 2018 and their GDP values in millions.

We find out from the visualizations that countries like China, United States and other countries whose expenditure is more on tourism have better GDP values than other countries whose expenditure is relatively less. This visualization proves that when a country invests more in tourism, it gets more profits and has an economic advantage.



Results for BI Query 3

# 8 Conclusion and Future Work

From our visualizations, it is seen that the number of international tourist arrivals, percentage change in tourists, and the expenditure of countries on tourism all have a direct impact on the growth of a country. It is found that countries which encourage and actively indulge in promoting tourism reap economic benefits such as better GDP values, better growth ranks of countries and GDP values.

Tourism can no longer be neglected, it must be recognized by all the countries as a major contributor to the countrys economic advancement. It has direct and indirect benefits. The governments, related organizations and associations have to step up to the pedestal and make their respective countries capable of handling travel and tourism in a very efficient manner so that it only has a positive impact on the society.

Tourism has to be seen as the major contributor of a countrys GDP which it is today. This has to be more researched by developing countries as it will increase their growth percentage. The importance of travel and tourism has to be made sure that it is known at the core local level to the international level.

Tourism is the key development factor for developing countries. For example, countries like Malaysia has benefited a lot from tourism activities. All these studies and real scenarios have to be analyzed by other countries who are not actively engaging in promoting tourism and they have to use the current technological advancements to match up with the trend and include tourism in their marketing campaign and advertisements etc., and make sure that tourism is used as their major contributor to their economic growth as well.

# References

https://en.wikipedia.org/wiki/Impacts$_of_tourism$

https://www.researchgate.net/publication/228287248

https://msu.edu/course/prr/840/econimpact/pdf/ecimpvol1.pdf

http://dergipark.gov.tr/download/article-file/363129

https://en.wikipedia.org/wiki/World_Tourism_rankings

# Appendix

The codes used in this project are:

data1 ¡- read.csv(file = ”data.csv”,header = TRUE) ¿ View(data1)
¿ FranceData ¡- data1[which(data1$Country.Name = ”France”)]Error : unexpected' =' in”FranceData < -data1[which(data1$Country.Name =”
¿ FranceData ¡- data1[which(data1$Country.Name == ”France”)]Errorin‘[.data.frame‘(data1, which$
== ”France”)) : undefined columns selected
¿ class(data1) [1] ”data.frame
¿ library(sqldf) Error in library(sqldf) : there is no package called sqldf
¿ install.packages(’sqldf’)
Installing package into
Mac/Home/Documents/R/win-library/3.3 (as lib is unspecified) also installing the dependencies gsubfn, proto, RSQLite
package gsubfn successfully unpacked and MD5 sums checked
package proto successfully unpacked and MD5 sums checked

package RSQLite successfully unpacked and MD5 sums checked
package sqldf successfully unpacked and MD5 sums checked

The downloaded binary packages are in:

C:$_packages > library(sqldf)$
$[reachedgetOption(”max.print”) - -omitted11653rows] > specificCountryData <$
$-sqldf(”Select *$
$fromdata1dawhereda.Country.ISO3IN('France','Spain','UnitedStates','China','Italy','Mexico','U$
$ErrorinsqliteSendQuery(con, statement, bind.data) : errorinstatement :$
$nosuchcolumn : da.Country.ISO3$
$> names(data1)$
$[1]”Country.ISO3””Country.Name””Indicator””Subindicator.Type””X1995”[6]”X1996””X1997””X$
$names(data1)[1] < -CountryError : object'Country'notfound$
$> names(data1)[1] < -'Country'$
$> names(data1)[2] < -'CountryName'$
$> specificCountryData < -sqldf(”Select *$
$fromdata1dawhereda.CountryNameIN('France','Spain','UnitedStates','China','Italy','Mexico','U$
$> View(specificCountryData)$

> groupedDataByCountry < −sqldf("Select ∗
fromdata1dawhereda.CountryNameIN('France',' Spain',' UnitedStates',' China',' Italy',' Mexico',' U
ErrorinsqliteSendQuery(con, statement, bind.data) :
errorinstatement : near"GROUPBY" : syntaxerror >
write.csv(specificCountryData, file = "Tourism.csv")
> data1Subindicator.Type ¡-
sub("","",data1Subindicator.Type)
> data1Subindicator.Type ¡-
¿ library(readxl) ¿ statistaData ¡- read$_e$xcel(path =
"statistic$_i$d261726$_c$ountries − with − the − largest − number − of − international −
tourist − arrivals − in − 2017.xlsx", sheet = "Data") > view(statistaData)
Error : couldnotfindfunction"view" > View(statistaData)
> names(statistaData)
[1]"Countrieswiththelargestnumberofinternationaltouristarrivalsin2017"
[2]"" > na.omit(statistaData)
Countrieswiththelargestnumberofinternationaltouristarrivalsin20173France86.9000000000000064S
View(statistaData)
> statistaData < −na.omit(statistaData)
> View(statistaData) >
names(statistaData)[1]"Countrieswiththelargestnumberofinternationaltouristarrivalsin2017"[2]""
¿ names(statistaData)[2] ¡- "Percentage"
¿ View(statistaData)
¿ write.csv(file = "StatistaData.csv")
Error in is.data.frame(x) : argument "x" is missing, with no default
¿ write.csv(statistaData,file = "StatistaData.csv")
¿ pdfData ¡- read.csv(file = "tabula-World Tourism rankings - Wikipedia.csv")
¿ View(pdfData)
¿ wikipadiaData ¡- read.csv(file = "tabula-World Tourism rankings - Wikipedia.csv")
¿ View(wikipadiaData)
¿ names(wikipadiaData)
[1] "Rank" "Destination"
[3] "International.tourist.arrivals..2017..1." "International.tourist.arrivals..2016..1."
[5] "Change..2016.to.2017....." "Change..2015.to.2016....." ¿ names(wikipadiaData)[3] ¡-
"InternationalTouristArrival2017"
¿ names(wikipadiaData)[4] ¡- "InternationalTouristArrival2016"
¿ names(wikipadiaData)[5] ¡- "Change2016To2017"
¿ names(wikipadiaData)[6] ¡- "Change2015To2016"
¿ wikipadiaDataInternationalTouristArrival2017 <
−sub("million","", wikipadiaDataInternationalTouristArrival2017) ¿
wikipadiaDataInternationalTouristArrival2016 <
−sub("million","", wikipadiaDataInternationalTouristArrival2016)
¿ write.csv(wikipadiaData,file = "wikipadia.csv")