National College of Ireland

Data Warehousing and Business Intelligence Project

on

Tourism

Arth Vyas
x17170516

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:	Arth Vyas
Student ID:	x17170516
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:	Tourism

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

Tourism

Arth Vyas
x17170516

November 26, 2018

Abstract

This project is based on Tourism, major three dependent domains were incorporated in this project as mentioned. They are Hotel industry, GDP of country and factor influencing tourist to visit specific country. Tourism is growing and successful industry hence, if proper analysis in this sector will be carried out this industry can do wonders. to being with the analysis has been done with respect to country and time frame. Required data sets were gather through different sources. A data warehouse is build to store Raw data and dumped into data marts after appropriate changes done to it. The flow of collecting Raw data, finding relationship in data, cleaning of data, loading of data, transforming of data, process of cube creation, automation of whole process through data warehouse and analysis services. The business requirement overall process in the project is addressed in this report.

1 Introduction

Tourism is the emerging sector in todays era. We have often seen that various cities in world are solely dependent on tourism for earing and people leaving have engrossed them self in turning the visit of tourist in that city memorable by providing following amenities. Now questions come which are those amenities which visitors look for while visiting particular place? In this project amenities and factors affecting Tourism and country will be address. The data is captured from different source to analysis the tourism for following requirement.

(Req-1) Project first requirement is to analysis World Ranking of Country influencing tourism rate.

(Req-2) Second requirement is to get an insight on country GDP with respect to hotel revenue.

(Req-3) Third requirement is to have a glance on hotel occupancy rate effected by overseas visitors.

2 Data Sources

Building a Data Warehouse and Business Intelligence project three different sources are used to carry out the project on Tourism. They will be described briefly in following section with the source details, data type and summary.

Source	Type	Brief Summary
Statista	Structured	Used to find the effect of Revenue, occupancy and visitors on Tourism.
OECD	Structured	Used to find the impact of hotel revenue on GDP of country.
Media.beam.usnews	Unstructured	Used to find does country ranking in different area inspire visitors to visit particular country.

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

2.1 Source 1: Statista

The Data from Statista caters the part of Hotel Occupancy rate, Hotel revenue and Overseas Visitors to the countries.Hotel Occupancy US: :https://www.statista.com/statistics/200161/us-annual-accomodation-and-lodging-occupancy-rate/ Publish Date: Jan 2018. Data Contains of 4 Columns and 22 Rows. Hotel Occupancy UK: :https://www.statista.com/statistics/323728/hotel-occupancy-ate-change-in-provinces-i
Publish Date: Sept 2018. Data contains of 4 Columns and 17 Rows. Both the countries data were merged and made in 5 columns and 21 Rows as well as appropriate clean is done. Link Hotel Revenue US: :https://www.statista.com/statistics/245841/total-revenue-of-the-us-hotel-industry/. Link Hotel Revenue UK :https://www.statista.com/statistics/323729/hotel-revpar-change-in-provinces-in-the-united-kingdom
Publish Date: Jul 2018. Data Contains of 3 Columns and 22 Rows. Both the countries data were merged and made in 5 columns and 35 Rows as well as appropriate clean is done.

2.2 Source 2: OCED

From this source GDP for US and UK has been fetched which will allow to compare that is there any impact of hotel revenue on GDP for respective countries. Link: :https://data.oecd.org/gdp/real-gdp-forecast.htm. Data Contains of 8 Columns and 529 Rows. Clean data consist of 5 Columns and 23 Rows.

2.3 Source 3: Media.beam.usnews.com

This Data is about the overall world ranking of the countries in different field i.e. Cultural influence, Heritage, adventure and overall rank. Analysis to be made in such a way to know that is the ranking of countries in several field is inspiring the visitors to visit. World Ranking of country for 2017: :https://media.beam.usnews.com/6a/c8/bccd653643b983c3e1ff671dcf13/171110-best-countries-overall-rankings-2017.pdf. (Copyright 2017 U.S. News World Report LP. All rights reserved.) This PDF Consist of 13 pages with 4 tables and description with 4 tables having 11 columns and 80 Rows, clean table have 8 Columns and 2 Rows. World Ranking of country for 2018: :https://media.beam.usnews.com/ce/e7/fdca61cb496da027ab53bef37a24/171110-best-countries-overall-rankings-2018.pdf. (Copyright 2018 U.S.News World Report LP All rights reserved.)This PDF Consist of 13 pages with 4 tables and description with 4 tables having 11 columns and 80 Rows, clean table have 8 Columns

and 2 Rows. Final table is created by merging the above cleaned table for both countries having 8 Columns and 4 Rows.

3 Related Work

The following five paper has been reviewed before deciding the topic for this project. Here are the links given below: Reference Paper1: https://files.eric.ed.gov/fulltext/ED155110.pdf Reference Paper 2: https://files.eric.ed.gov/fulltext/EJ1134236.pdf Reference Paper 3: https://www.emeraldinsight.com/doi/full/10.1108/IJCHM-10-2014-0507 Reference Paper 4: https://www.emeraldinsight.com/doi/full/10.1108/09596119710191056 Reference Paper 5: https://www.emeraldinsight.com/doi/pdfplus/10.1108/TR-11-2014-0056 The major insight was gained by reading the reference paper were, it is mentioned that tourism is the growing industry in the world. According to the world tourist organization international tourist grew by 4 percent in 2011 by 980 million. It is also found that travel and tourism have account for 9.2 percent of global GDP. Hence due to tourism hotel industry is in huge demand for facilitate the accommodation requirement for tourist. Based on triple bottom line guidelines, companies mission in such a context is achieving competitive advantage by means of a business model considering

three key aspects: people, profit, and planet. There is a study based on daily occupancy of hotels by tourist was predicted by analyzing forecast.

There are plethora amenities which attracts the tourists location like landscape, adventure, heritage and cultural importance. By referring all the paper, the topic tourism was selected and tried to give different dimension of analyzing the tourism factors by comparing revenue with GDP, Occupancy with Visitors and the non-trivial approach was experiment to analyze the attraction of tourist towards the country by ranking in different sectors.

4 Data Model

The data has been selected for this project is based on continent, country and specific duration of time span. All the measure in Fact table share the same dimensions of dimCountry and dimyear. Measure are taken for following items Adventure, Cultural Influence, Heritage, overall Rank, Hotel Revenue, Overseas Visitors, Hotel occupancy Rate and GDP. The respective measure has been chosen to analysis the aim of project to check does these measure are really affecting the Tourism of country. For example, one of the requirement is to know if there is any effect on country GDP as the numbers of overseas visitors are visiting the country and halting at hotel due to which hotel revenue get increase.

Figure 5. Project targets to answer all the business intelligence queries, In which DimContry allows to compare the stats and trend between Continents and country with respect to all measures present in Fact table dimyear will leverage all the queries by adding a specific period of time to it. The dimensions are choose to give fair idea for all requirement which is going to be derived in this project. At present there is only one country included for each continent and time duration from 2002 to 2018 is taken and will be used appropriately for each queries.

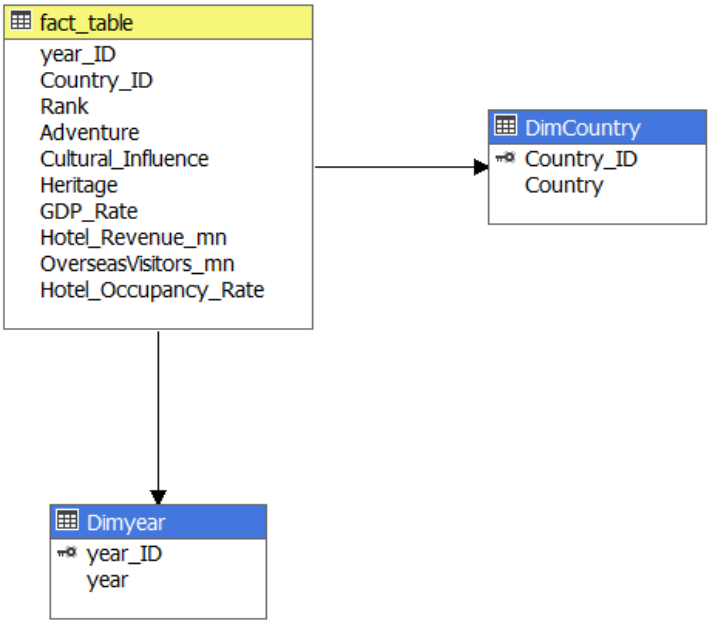

Figure 1: star schema

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 5

Source	Column	Destination	Column	Type	Transformation
1	Overseas Visitors_mn	Fact_Table	Overseas Visitors_mn	Fact	Column is created after merging two tables for two different countries, Reworked Column heading, rounded figure with no decimals in million, deletion of unwanted rows.
1	Hotel _Occupancy _Rate	Fact_Table	Hotel _Occupancy _Rate	Fact	Column is created after merging two tables for two different countries, Reworked Column heading, rounded figure with no decimals in percent, deletion of unwanted rows and Columns $
1	Hotel _Revenue _mn	Fact_Table	Hotel _Revenue _mn	Fact	Column is created after merging two tables for two different countries, Reworked Column heading, rounded figure with no decimals in million, deletion of unwanted rows and Columns
1	GDP_Rate	Fact_Table	GDP_Rate	Fact	Reworked column heading and deleted unwanted rows.
1	Rank	Fact_Table	Rank	Fact	Column is created after merging two scrape PDF tables for two different countries, Reworked Column heading, deletion of unwanted rows and Columns
1	Revenue	FactTable	Revenue	Fact	Rounded to nearest million $
1	Adventure	Fact_Table	Adventure	Fact	Column is created after merging two scrape PDF tables for two different countries, Reworked Column heading, deletion of unwanted rows and Columns

Continued on next page

Table 3 – *Continued from previous page*

Source	**Column**	**Destination**	**Column**	**Type**	**Transformation**
1	Cultural _Influence	Fact_Table	Cultural _Influence	Fact	Column is created after merging two scrape PDF tables for two different countries, Reworked Column heading, deletion of unwanted rows and Columns
1	Heritage	Fact_Table	Heritage	Fact	Column is created after merging two scrape PDF tables for two different countries, Reworked Column heading, deletion of unwanted rows and Columns
1,2,3	Country	DimCountry	Country	Dimension	Reworked Country name for similarity, Column is merged with 10 tables. Unwanted rows and columns are deleted.
1,2,3	year	Dimyear	year	Dimension	Reworked year format for similarity, Column is merged with 10 tables. Unwanted rows and columns are deleted.
1,2,3	Country_ID	Dimyear, Fact_Table	Country_ID	Dimension, Fact	A separate CSV was generated and merged with all 10 tables for creating primary key in SSMS
1,2,3	Year_ID	Dimyear, Fact_Table	Year_ID	Dimension, Fact	A separate CSV was generated and merged with all 10 tables for creating primary key in SSMS

6 ETL Process

The required data is fetch from the different data score and data types after reshaping the data as explained in logical data map, finally created 5 CSV which is ready to load in different database tables. Now following are the steps followed for loading of the data in Data warehouse:

6.1 First New Database is created in SQL server management studio for storage of captured data for the project.

6.2 Inserted Execute Process Task in SQL server integration services. In this task load Rscript.exe file with the executable location and working directory of R in the system. Which will load the 5 CSV data in empty table created in next step.

6.3 In SQL server integration services 5 Data flow has been inserted according to the requirement of the project Data. Next flat files and OLE DB Destination were included in each Data flow. While loading the files location in flat file a database connection is made between SSIS and SSMS. Here the columns are mapped with each other from flat files to OLE DB Destination to ensure the correct data is inserted in each column as desired when the execution will take place.

6.4 Execute SQL task is inserted in the staging area to create a Empty table by merging all the Data coming from Data flow. Here the left join is used for merging table with common column with each other and hence a Raw Table is created by merging all the 5 CSV.

6.5 Now the Dimension and Fact table is created in SSMS, the values are inserted from Raw Data table called Emptytable through SSIS. The primary key is assigned to Year _ID and Country _ID for Dimyear and DimCountry respectively.

6.6 Now the staging area is ready to load the data in the tables and check if the command has executed properly and all the table is filled with the mapped files in the data flow. Another two Execute SQL task will be created for Dimension and Fact table in which insert command will allow to populate the table.

6.7 Inserted the Execute SQL Task for applying the Truncate command to clear the existing data in the table so that when the execution will process again there will be no duplicate values will be inserted.

6.8 Created SQL server analysis services package, where the Data source, Data source view, dimension and cube creation was taking place. After the deployment of cube. The aim of this process is to gather raw data, extract transform and load data as soon as the cube process will complete the successful, creation of Data warehouse done and data is ready for analysis.

6.9 Now created a connection of analysis services of SSMS with SSIS for automation of Cube deployment, the Database source data file and initial identity catalog is developed. Inserted Sequence container in which two Analysis services processing task for assigning Dimension and fact.

6.10 Now SSIS is ready to execute in fully automation mode, the loading of data in database and deployment of cube will be executed with the execution command.

The major challenges faced to build a Raw Data for this project were to gather appropriate Data. Raw Data consist two countries data for United States and United Kingdom. The time span is considered year from 2002 to 2018. All the nine table were available for each country separately so, the next step was to merge both countries together with the available time span for each gathered data for Tourism. Total six table extracted from statista. Two table were scrape from two different PDF and merge

to make a table. One another structured data was taken and clean appropriately as mentioned, i.e, data related Hotel Occupancy Rate, Overseas Visitors, Hotel Revenue and data extracted table from PDF. Basically, general cleaning for all the table were merging table, reordering columns, deletion of rows/columns, addition of column where the data was missing, formatting the values. In the end all the code for each table were club together and 5 different CSV were generated for creating the automation of loading the data into the empty table.

7 Application

For the first requirement, the data of world Ranking of United States and United Kingdom in the respecting factors like Cultural Influence, heritage, adventure and overall ranking with respect to overseas visitors in the duration of 2017 2018 is to be compared. So that we will come to know is there an effect of Country ranking influencing visitors to visit specific country.

For the second requirement, we can analysis the trend of GDP over the period of span for United states and United Kingdom. GDP and revenue goes hand in hand so, by comparing the hotel revenue trend will explain whether GDP is increasing or decreasing with respect to hotel revenue.

For the third Requirement, here the relationship of Overseas visitors with respect to hotel occupancy rate is going to be checked. It is evident that if Overseas Visitors plan visit to country they will need a place for accommodation. The trend will explain is that the case or other way around.

7.1 BI Query 1: Which factor is influencing visitors to visit specific country

For this query, the contributing sources of data are Overseas visitors included form source 1 and ranking of country in different factors are fetched from source 3.

It is seen from the graph that the ranking of United States expect Adventure is high with comparison to United Kingdom and the overseas visitors are more visiting United states. which is an indicating that visitors are influenced by the Ranking of different sectors.

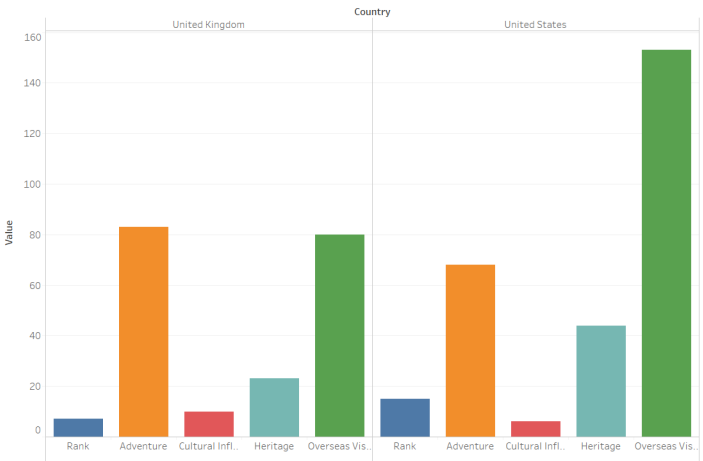

Figure 2: Results for BI Query 1

7.2 BI Query 2:Which country GDP is affected by the Hotel Revenue

For this query, the contributing sources of data is GDP table from Source 2 and Hotel Revenue table from Source 1.

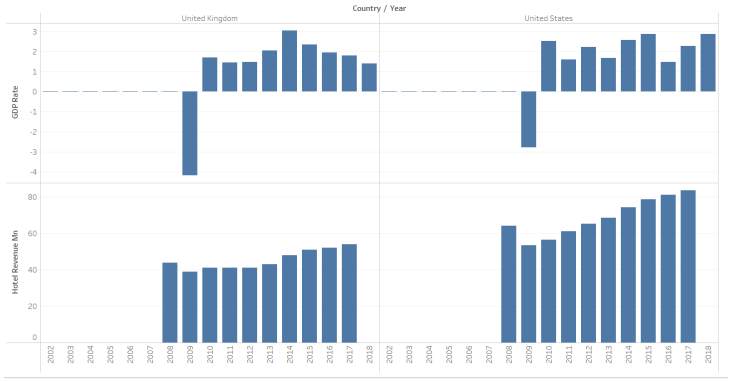

Figure 3: Results for BI Query 2

The graph shows the similarities in the GDP Rate and Hotel Revenue but it can not be straight forwardly inferred it affect directly to GDP.

7.3 BI Query 3:Which Country hotel occupancy is affected the most due to Overseas visitors

For this query, the contributing sources of data are overseas visitors from overseasvisitors table came from source 1 and Hotel occupancy rate is fetch from HotelOccupancy table from source 1. both are fetched from different tables.

Graph represent the trend of overseas visitors and Hotel Occupancy. They both are increasing simultaneously. But it is evident that United States Overseas visitors are really accommodating themselves in hotel as the both graph are identical.

Figure 4: Results for BI Query 3

7.4 Discussion

The literature paper have tired to follow different methods, general discussion will be mentioned here regarding each query carried out in this project and different dimension of analyzing tourism with these three business intelligence query.

BI query 1 the attraction of tourist to visit a country is compared with the ranking of the factors like Adventure, cultural _Influence and heritage which to my mind are the main essence of location which tourist will expect on there visit. on the other hand literature work has gave the insight on the same factors by applying different methods that is attraction factors, is nearest crowd residing nearby famous location, or modes of travel preferred by them and so on. by keeping in this mind result of query one it is evident that irrespective of in factors attraction of country visitors have visited United states more than United Kingdom which can justify that not only attraction factors are important the ranking for specific country in particular factors can all together give new direction to analyze the tourist pattern.

BI query 2 literature work has uplifted the hotel revenue management and stated that Hotel industry is the heart of tourism and Tourism is contributing approximately 9.2 percent to global GDP. So this query targets to analyze in more granular level that is GDP of country is compared with the Hotel revenue as the hotel industry will have significantly contribute to tourism. it is evident from graph that majority of time revenue and GDP trend has shown similarities in BI Query 2 but due to limited data the trend can not straightforwardly address that yes GDP and Hotel revenue goes hand in hand but it pretty evident that yes they have relation. including this dimension of analyzing country GDP and hotel revenue will add more value to the researchers. BI Query 3 Normally Hotel Occupancy have analyzed with respect to forecast. tourist are gonna attract more towards the location if there is pleasant atmosphere. it will be unfair to ignore the people who visits the location irrespective of weather. so this query has incorporated the data of overseas visitors our a period of time lets say from 2008 to 2017. The interesting fact came out from the graph is as the visitors have increase in United States and United Kingdom hotel occupancy rate of the country has increased. on the one side increase visitors capacity over the year and hotel occupancy of united states goes hand in hand. on the other hand in United Kingdom visitors and hotel occupancy has followed the same pattern but i can been seen that all the majority of visitors were not accommodate which will be next the target for future work.

8 Conclusion and Future Work

In conclusion all the data sets and Business intelligence query were made to achieve the aim of pitch in the new dimension of analyzing tourism for the country.Three ideas were implemented and executed in the form of business query. the desired comparison were made between factors influencing visitors, revenue affecting GDP and hotel occupancy relationship with visitors.

The data warehouse created for this project has allowed all the different sources to store together and ETL as well as cube process have leveraged to analysis the data in different dimension as required and successfully help in achieving the required outcome. Connection of analysis software tableau with data warehouse has boost the graphical representation of results. hence, allowed to derived results and finding of queries. This data warehouse can be of great help to hotel industries and tourism industries for analyzing

the tourism rate, hotel revenue, occupancy. Data warehouse is capable of modification and re framed according to the requirements and appropriate analysis could be made.

Certain limitation which were faced during the project, all the graph and comparison of GDP, Visitors, revenue, occupancy would have been represented strong claim and figures to justify the answers derived from query if more continents and countries were included then better comparison could have been made. solid evidence could be brought to the findings by comparing the data with different continents and country all together.

References

[1] Hyojin Kim, Byung-Gook Kim, (2015) "Economic impacts of the hotel industry: an input-output analysis", Tourism Review, Vol. 70 Issue: 2, pp.132-149, https://doi.org/10.1108/TR-11-2014-0056 Permanent link to this document: https://doi.org/10.1108/TR-11-2014-0056

[2]: Zvi Schwartz, Muzaffer Uysal, Timothy Webb, Mehmet Altin, (2016) "Hotel daily occupancy forecasting with competitive sets: a recursive algorithm", International Journal of Contemporary Hospitality Management, Vol. 28 Issue: 2, pp.267-285, https://doi.org/10.1108/IJC 10-2014-0507 Permanent link to this document: https://doi.org/10.1108/IJCHM-10-2014-0507

[3]: High. Learn. Res. M. Mattera and A. Moreno-Melgarejo Commun. Vol. 2, Num. 4 — December 2012

[4]: matley,Jan. The Geography of Intertational Tourism. ResourcePaper Mc. 76-1.Association of American GeOgraphers,'Washington, National Scien ce FoUndation, Washington, EF-40.83 Plus Postage. BC. Not Availably frog 'EDRS.bibliographies; Cultural Factors; Economic Factors;-Foreign Countries; Geograpkic Distribution;Geographic Location; Geographic Regions; *Geography;*Global Approach; Higher Education-Industry; *International'Relatiofts; Recreatiohal Activities;'Resource Materials; Study Abroad; *Tourism; Travel;Trend Analysis

[5]: Trevor Ward, (1997) "Hotel market trends in the UK", International Journal of Contemporary Hospitality Management, Vol. 9 Issue: 7, pp.270-273, https://doi.org/10.1108/09596119

[6]www.statista.com

[7]www.OCED.com

[8]www.stackoverflow.com

[9]media.beam.usnews.com

Appendix

R code example

```
# Automation Code#--


library("tabulizer")
library(dplyr)
#1st PDF#

pdf <- '/Users/ARTH␣VYAS/Downloads/DWBI␣Dataset/CountryRank17.pdf'
pdf <- extract_tables(pdf)
#View(pdf)
```

```
new <- data.frame(pdf[2])
#View(new)

N6 <- subset(new, select = -c(X2, X4, X7, X8, X9, X10) )
N6<- N6[c(3,7), ]

#View(N6)
N6$year = 2017
N6$Rank = 1
N6$Rank = as.numeric(N6$Rank)
N6[2,6] = 7
N6[1,6] = 3
names(N6) [1]<- paste("Country")
names(N6) [2]<- paste("Adventure")
names(N6) [3]<- paste("Cultural_Influence")
names(N6) [4]<- paste("Heritage")
#View(N6)
N6<- N6[c("year","Country","Rank","Adventure","Cultural_Influence","Heritage

#2st PDF#

pdf1 <- '/Users/ARTH␣VYAS/Downloads/DWBI␣Dataset/CountryRank18.pdf'
pdf1 <- extract_tables(pdf1)
#View(pdf1)
new1 <- data.frame(pdf1[3])
#View(new1)

N7 <- subset(new1, select = -c(X2, X4, X7, X8, X9, X10) )
N7<- N7[c(4,8), ]

#View(N7)
N7$year = 2018
N7$Rank = 1

N7$Rank = as.numeric(N7$Rank)
N7[2,6] = 8
N7[1,6] = 4
names(N7) [1]<- paste("Country")
names(N7) [2]<- paste("Adventure")
names(N7) [3]<- paste("Cultural_Influence")
names(N7) [4]<- paste("Heritage")

#View(N7)
N7<- N7[c("year","Country","Rank","Adventure","Cultural_Influence","Heritage

unstructed <- rbind(N6, N7)                                    #join two data.f
#View(unstructed)                                               #CountryRank#

idy=read.csv('/Users/ARTH␣VYAS/Downloads/DWBI␣Dataset/year_ID.csv')
yearID <- data.frame(idy) #loading yearID CSV in dataframe#
#View(yearID)
```

```
idc=read.csv('/Users/ARTH␣VYAS/Downloads/DWBI␣Dataset/Country_ID.csv')
countryID <- data.frame(idc) #loading countryID CSV in dataframe#
#View(countryID)


hn =merge(countryID,unstructed,by ="Country")
BN <- data.frame(hn)

A1 =merge(yearID,BN,by ="year")


#View(A1)   #Country Rank#
write.csv(A1,'/Users/ARTH␣VYAS/Documents/R/CleanpdfCR.csv', row.names = FALS




#install.packages("readxl")
library("readxl")
#Structured Data source cleaning for Revenue of hotels#

ping <- read_excel("C:/Users/ARTH␣VYAS/Downloads/DWBI␣Dataset/USHR.xlsx", sh
#View(ping)
ping<- ping[-c(1,2,3,4,5,6,7,8,9), ]
names(ping) [1]<- paste("year")
names(ping) [2]<- paste("Hotel_Revenue_mn")
ping$Hotel_Revenue_mn = as.integer(ping$Hotel_Revenue_mn) #round#

ping$Country = "United␣States"  #addition of column for Country"

ping<- ping[c("year","Country","Hotel_Revenue_mn")] #rearranging Columns#


#UK hotels structure data set cleaning#

ting <- read_excel("C:/Users/ARTH␣VYAS/Downloads/DWBI␣Dataset/UKHR.xlsx", sh
#View(ting)
ting<- ting[-c(1,2,13,14), ] #deleting the rows#
names(ting) [1]<- paste("year")
names(ting) [2]<- paste("Hotel_Revenue_mn")
ting$Country = "United␣Kingdom"  #addition of column#

ting<- ting[c("year","Country","Hotel_Revenue_mn")] #rearranging Columns#

source1 <- rbind(ting, ping)                                    #join two data.
#View(source1)                                                 #Hotel_Revenue#


jm =merge(countryID,source1,by ="Country")
BL <- data.frame(jm)
```

```r
A3 =merge(yearID,BL,by ="year")

#View(A3)    #Hotel Revenue#
write.csv(A3,'/Users/ARTH␣VYAS/Documents/R/cleanHR.csv', row.names = FALSE)




# US visitors Cleanning Code#
ying <- read_excel("C:/Users/ARTH␣VYAS/Downloads/DWBI␣Dataset/USVisitors.xls
View(ying)
ying<- ying[-c(1,2,3,4,22,23,24,25), ] #deleting the rows#
names(ying) [1]<- paste("year")
names(ying) [2]<- paste("OverseasVisitors_mn")
ying$Country = "United␣States"  #addition of column#
ying$OverseasVisitors_mn = as.integer(ying$OverseasVisitors_mn)

ying<- ying[c("year","Country","OverseasVisitors_mn")]

# UK visitors Cleanning Code#
ring <- read_excel("C:/Users/ARTH␣VYAS/Downloads/DWBI␣Dataset/UKVisitors.xls
#View(ring)
ring<- ring[-c(1,2), ] #deleting the rows#
names(ring) [1]<- paste("year")
names(ring) [2]<- paste("OverseasVisitors_mn")
ring$Country = "United␣Kingdom"  #addition of column#
ring$year = as.character((ring$year))
ring [17,1] = 2018
ring$OverseasVisitors_mn = as.integer(ring$OverseasVisitors_mn)

ring<- ring[c("year","Country","OverseasVisitors_mn")]
source2 <- rbind(ying, ring)                                    #join two data.
#View(source2)                                                 #Overseasvisitors#

zx =merge(countryID,source2,by ="Country")
BJ <- data.frame(zx)
A5 =merge(yearID,BJ,by ="year")

View(A5)  #Visitors#
write.csv(A5,'/Users/ARTH␣VYAS/Documents/R/cleanvisitors.csv', row.names = F




#GDP Cleaning code#
v5=read.csv('/Users/ARTH␣VYAS/Downloads/DWBI␣Dataset/USGDP.csv')
#View(v5)
v5<- v5[c(309,310,311,312,313,314,315,316,317,318,319,320,321,322,323,324,32
#View(v5)
```

```
v5 <- subset(v5, select = -c(INDICATOR, SUBJECT, MEASURE, FREQUENCY, Flag.Co
#View(v5)
names(v5) [1]<- paste("Country")
names(v5) [2]<- paste("year")
names(v5) [3]<- paste("GDP_Rate")

#View(v5)
v5$Country = as.character(v5$Country)
v5[1,1] = 'United Kingdom'
#View(v5)
v5[2,1] = 'United Kingdom'
v5[3,1] = 'United Kingdom'
v5[4,1] = 'United Kingdom'
v5[5,1] = 'United Kingdom'
v5[6,1] = 'United Kingdom'
v5[7,1] = 'United Kingdom'
v5[8,1] = 'United Kingdom'
v5[9,1] = 'United Kingdom'
v5[10,1] = 'United Kingdom'
v5[11,1] = 'United Kingdom'
v5[12,1] = 'United States'
v5[13,1] = 'United States'
v5[14,1] = 'United States'
v5[15,1] = 'United States'
v5[16,1] = 'United States'
v5[17,1] = 'United States'
v5[18,1] = 'United States'
v5[19,1] = 'United States'
v5[20,1] = 'United States'
v5[21,1] = 'United States'
v5[22,1] = 'United States'

View(v5)                                #GDP#

v5<- v5[c("year","Country","GDP_Rate")]


jk =merge(countryID,v5,by ="Country")
BM <- data.frame(jk)
A2 =merge(yearID,BM,by ="year")


View(A2)      #GDP#
write.csv(A2,'/Users/ARTH VYAS/Documents/R/cleanGDP.csv', row.names = FALSE)




# US Occupancy Cleanning Code#
```

```
cing <- read_excel("C:/Users/ARTH␣VYAS/Downloads/DWBI␣Dataset/USOR.xlsx", sh
cing<- cing[-c(1,2,3,4,5,6,7,8,9), ] #deleting the rows#
#View(cing)
cing_1 = cing[,-3]

names(cing_1) [1]<- paste("year")
names(cing_1) [2]<- paste("Hotel_Occupancy_Rate")
cing_1$Country = 'United␣States'
cing_1<- cing_1[c("year","Country","Hotel_Occupancy_Rate")]
#View(cing_1)
cing_1$Hotel_Occupancy_Rate = as.integer(cing_1$Hotel_Occupancy_Rate)


#UK Occupancy cleanning code#
king <- read_excel("C:/Users/ARTH␣VYAS/Downloads/DWBI␣Dataset/UKOR.xlsx", sh
#View(king)
king<- king[-c(1,2,13,14), ] #deleting the rows#
king$Country = 'United␣Kingdom'
king = king[,-3]
names(king) [1]<- paste("year")
names(king) [2]<- paste("Hotel_Occupancy_Rate")
king<- king[c("year","Country","Hotel_Occupancy_Rate")]
v0 <- rbind(king, cing_1)                                    #join two data.fra
v0$Hotel_Occupancy_Rate = as.integer(v0$Hotel_Occupancy_Rate)
#View(v0)

nm =merge(countryID,v0,by ="Country")
BK <- data.frame(nm)
A4 =merge(yearID,BK,by ="year")

write.csv(A4,'/Users/ARTH␣VYAS/Documents/R/cleanOR.csv', row.names = FALSE)
#View(A4)  #Occupancy#
```

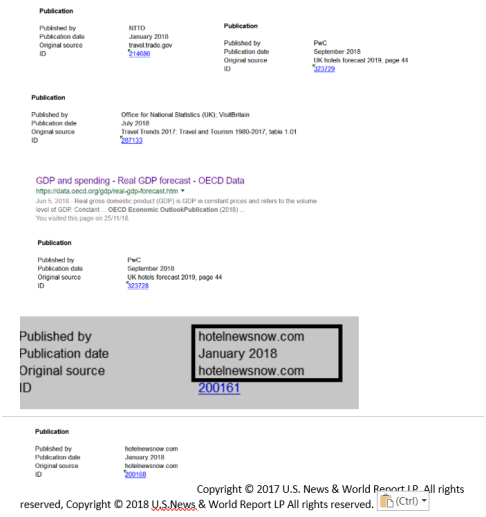

Figure 5: Publication Date