National College of Ireland

Data Warehousing and Business Intelligence Project

on

Analysis of air traffic and opinion of customers for the airline

Lalit Pathak
x1810088

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:	Lalit Pathak
Student ID:	x1810088
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:	Analysis of air traffic and opinion of customers for the airline

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:	December 3, 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

Analysis of air traffic and opinion of customers for the airline

Lalit Pathak
x1810088

December 3, 2018

Abstract

Everyone loves to go on a long holiday to get rid of, of the hectic routine work. These routine breaks give us new enthusiasm and paved a way in shifting our energy to the next level. While planning to go on holiday, we need to consider some factors which will help us in making our holidays more desirable. The factors that need to be considered could be the right destination, the right time to visit the destination, the climatic conditions at the destination, the transport facilities and their services and many more. Here, I am considering few of these factors and creating a database to generate BI queries that will bring some insights to the customers or a traveller. Along with those factors I am considering the ratings and reviews of various airlines which surely bring a decisive move to the travellers.

1 Introduction

I am taking the data of the airlines which are flying between U.S and Spain during the year 2015 to 2017 on a quarterly basis. I have taken the parameters like Origin Airport, Destination Airport, City and Country of Origin, City and Country of Destination. I have sorted out this data on yearly, quarterly and on monthly basis. Along with that, I am taking number of passengers flying by various airlines for the respective time period. One of my data is focusing on the monthly temperature of Spain for the year 2015 to 2017. It will help me to do the analysis that how significantly the temperature impacts the passengers visit to Spain. The study done by (Domonkos P, Farre X, Duro J A, (2010)), shows that the summer climate especially in the months of July and August attract more tourists.

I have done the web scrapping of Trip Advisor website to get the ratings for the specific airlines on various parameters like Passengers volume, Value for money. I have also done the sentimental analysis on the basis of passengers reviews about their experiences with the airlines. The sentimental analysis gives the positive, negative and the overall sentiment scores with respect to 10 airlines serving passengers.

I will clean and transform all these data in R as per requirement of my facts and dimensions. I will then create a data warehouse where all the cleaned data will be stored and will be used make sensible BI queries which in return generate some qualitative conclusions.

(Req-1) Analysis of passangers traffic from USA to Spain

Source	Type	Brief Summary
STATISTA	STRUCTURED	AVERAGE MEAN TEMPERATURE OF SPAIN FOR YEAR 2015-17
Bureau of Transportation Statistic	STRUCTURED	PASSENGER,AIRLINE,AIRPORT INFORMATION FROM THIS DATA SOURCE
TRIPADVISOR	UNSTRUCTURED	REVIEWS AND RATINGS FROM THIS DATA SOURCE

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

(Req-2) Req 2 Analysis of Airline according to ratings

(Req-3) Req 3 openion of the customers about the airline

2 Data Sources

Project uses 3 different data sources. First source of data is Statista from Where average mean temperature of the Spain for the year 2015-2017 is collected. and this is quarterly data source which is then clean and converted as per project requirement. The second source of data is from Bureau of Transportation Statistic of the unites states department of transportation. from this website contain information about Carrier, origin, destination, time which includes carrier name, airline id, carrier code, origin airport id ,destination airport id , origin country ,destination country and duration is selected. Mainly this data set provides passengers count for year 2015-2017.Data from this website is stored in local system and processed through R code. Final source of the data is from trip advisor from where rating and reviews of the airline data is extracted. This data is used to to identify to do the sentimental analysis Of the airline and to check whether it is worth to spend money to the airline

2.1 Source 1: Statista

Dataset Monthly average mean temperature in Spain from 2015 to 2017 (in degrees Celsius) `https://www.statista.com/statistics/802774/monthly-mean-temperature-in-spain/` was selected from the statista. This Dataset contains average monthly temperature which used as the fact in the data model of this project. temperature is correlated with the air passenger traffic travelling towards Spain from the US. Dataset contain 2 columns which includes month and year in single column and average monthly temp in c. This dataset is clean using R Code which includes removing 1st three rows which is unnecessary and then converting abbreviations of month like (jan =January) and for the year(15=2015). To create new column for the year and month . Generated quarter column from month and year column of the dataset. Value of March 2016 of avg mean temp column contains 100c which is not possible so corrected the value by taking mean of March 2015 and 17 before loading into fact table. One unique ID generated by concatenating month, year, and quarter from the dataset This value is unique for each row of the dataset which is used as primary key in the Data model. Release date is January 2018.

2.2 Source 2: Bureau of Transportation Statistic

Aviation data source is accumulate from the Bureau of Transportation Statistics (BTS) is an independent statistical agency under the Department of Transportation (DOT) of the united states of the America which use as 2nd source of the data in this Project. This source contain 3 dataset for the year 2015,2016,2017.These dataset contain the information about Unique carried Code, Airline id, Carrier name, Passenger information, Origin airport id ,origin, origin country, Origin city, destination airport id, origin airport id ,destination, destination country, destination city, year, quarter, month all these column are used to create dimension table of the data model as these column contains the information of passenger travelling to Spain through which airline and destination airport and all this information is quarter wise information for the 2015 to year 2017.This data set is locally stored and cleaning is used to merge all the data set for the year 2015,16 and 17 to create single dataset. This processed dataset is used to create dimension table such as DIM_AIRLINE and DIM_AIRPORT. These dimension table contains unique information of airline and airport. Dimension Airport has column Airport_ID which is used as primary key for the source airport id and destination airport id in Fact table. Similarly Airline ID from the Dimension Airline is used as the primary key which is unique value. Release date is May 2018.

2.3 Source 3: Trip advisor

Trip advisor has reviews and rating information for all the airlines. In This Project Trip advisor is used as 3rd source of the data and web scrapping technique is used to accumulate the data of required airline. Iteration of the url in R code is used to gather the required information from the trip advisor refer R code . R code output contains information about the airline name, total number of reviews, overall rating, legroom rating, customer service rating, value for the money, positive sentiment and negative sentiment which is used as the measures in the fact table for the 2nd and 3rd BI query.

3 Related Work

I have gone through some papers that help me to build some relations in my project work. I have accessed papers which gave me a reference to correlate temperature with tourists attraction for Spain. The research work done by (Domonkos P, Farre X, Duro J A, (2010)), The two most important explanations of the great heat tolerance is the air conditioning, which is general in Spanish hotels, and that the vast majority of foreign tourists are beach users. There are further evidences that the Mediterranean is a great tourist attraction and most tourists prefer to visit it in summer when the weather is continuously warm and sunny and the temperature of sea water is pleasant for bathing. It fortify that climatic potential of Spain flourishes its tourism and make many travellers as its their favourite destination.

I have done search on sentimental analysis, where the authors (Gao, B., Hu, N. and Bose, I. (2017)) suggests that, This study investigates if reviewers' pattern of rating is consistent over time and predictable. Two interesting results emerge from the econometric analyses using publicly available data from TripAdvisor.com. First, reviewers' rating behavior is consistent over time and across products. Furthermore, most of the variation in their future rating behavior can be explained by their rating behavior in the past rather

than by the observed average rating. Second, reviews by reviewers with higher absolute bias in rating in the past receive more helpful votes in future. We further divide the bias in rating into intrinsic bias (driven by intrinsic reviewer characteristics) and extrinsic bias (driven by influences beyond intrinsic bias) and document that intrinsic bias plays a more significant role in influencing helpful votes for reviews than extrinsic bias. Our results are robust to different product categories and different definition of bias.

4 Data Model

Star Schema is used to create the data model of the project .Data model contains three Dimension table and one fact table. Dimension tables are DIM_AIRLINE, DIM_AIRPORT, DIM_TIME .Each Dimension table hold unique Key called as primary key. In Dimension Airport airport id is used as the primary key which is referenced by foreign key present in the Fact table. Dimension airport contain airport ID, country of the airport and city in which airport is located . This information is used to calculate the number trips airline had to the airport. Data type of column AIRPORT_ID,COUNTRY_CODE,CITY_CODE,CITY_NAME are INTEGER,VARCHAR,VARCHAR,VARCHAR,VARCHAR Second dimension table is DIM_AIRLINE which holds information about the name of the airline, ID of the airline and code of the airline used by Department of Transportation of United states of America. Airline id is used as the primary key in DIM_AIRPORT table. Dimension tables are created from the 2nd source of data ref .Data type of column AIRLINE_ID,CARRIER_CODE, CARRIER_NAME are INTEGER,VARCHAR,VARHCAR Third Dimension is Time Dimension which holds information of Year, Quarter, Month this table contains TID as the primary key which is referenced by foreign key in the fact table. Fact table contain all the measure and output of BI quires are calculated from the fact table. Airline id, Airport_id and T_ID are used as the foreign key in the fact table these keys are used to fetch information from the Three Dimension table. Apart from foreign keys fact table contains other information like Passenger travelled from US to Spain, AVG_TEMP, TOTALREVIEWS, OVERALL_RATING ,LEGROOM_RATING , CUSTOMER_SERVICE_RATING, CLEANLINESS_RATING, FOOD_AND_BEVERAGE_RATING , SEAT COMFORT RATING, VALUE_FOR_MONEY_RATING, CHECK-IN AND BOARDING, IN-FLIGHT ENTERTAINMENT(WIFI, TV, FILMS, POSITIVE_SENTIMENT, NEGATIVE_SENTIMENT, OVERALL_SENTIMENT.These all attributes of the column Fact are used to build SAAS cube.

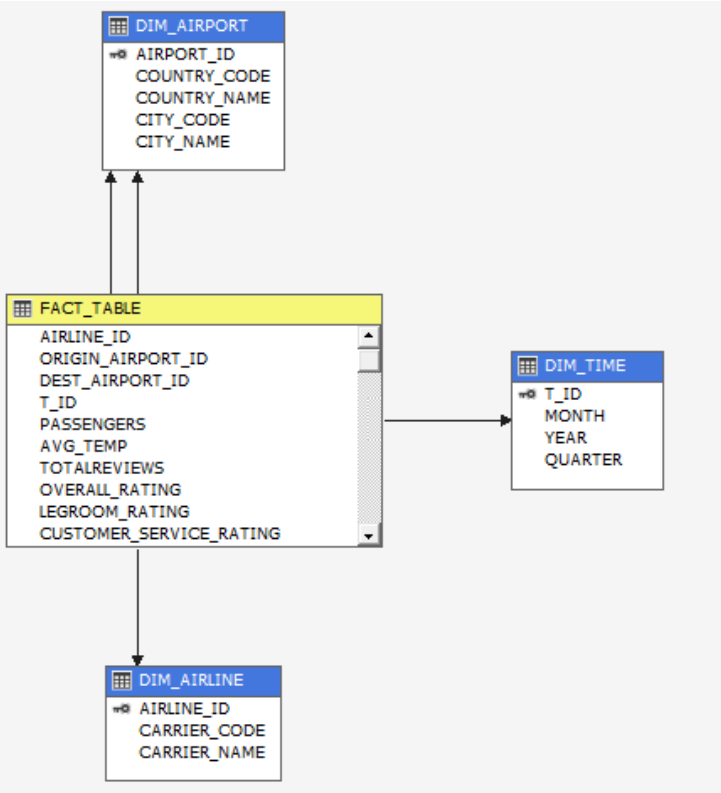

Figure 1: Star Schema

5 Logical Data Map

In this section, describe your logical data map, i.e. how every row of every data source is handled such that it is a part of your star schema.

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
RAW_DATA_AIRLINE	AIRLINE_ID	DIM_AIRLINE	AIRLINE_ID	Dimension	One to One mapping as all the cleaning and transformation is done in R code
RAW_DATA_AIRLINE	CARRIEAR_NAME	DIM_AIRLINE	CARRIEAR_NAME	Dimension	One to One mapping
RAW_DATA_AIRLINE	CARRIER_CODE	DIM_AIRLINE	CARRIER_CODE	Dimension	One to One mapping
RAW_DATA_AIRPORT	AIRPORT_COUNTRY	DIM_AIRPORT	AIRPORT_COUNTRY	Dimension	One to One mapping
RAW_DATA_AIRPORT	AIRPORT_CITY	DIM_AIRPORT	AIRPORT_CITY	Dimension	One to One mapping
RAW_DATA_AIRPORT	AIRPORT_ID	DIM_AIRPORT	AIRPORT_ID	Dimension	One to One mapping
RAW_DATA_TEMPERATURE	YEAR	DIM_TIME	YEAR	Dimension	One to One mapping as all the cleaning and transformation is done in R code
RAW_DATA_TEMPERATURE	QUARTER	DIM_TIME	QUARTER	Dimension	One to One mapping as all the cleaning and transformation is done in R code
RAW_DATA_TEMPERATURE	MONTH	DIM_TIME	MONTH	Dimension	One to One mapping as all the cleaning and transformation is done in R code
RAW_DATA_TEMPERATURE	T_ID	DIM_TIME	T_ID	Dimension	One to One mapping as all the cleaning and transformation is done in R code
RAW_DATA_AVIATION	AIRLINE_ID	AIRLINE	AIRLINE_ID	STG table of FACT	one to one mapping
RAW_DATA_AVIATION	ORIGIN_AIRPORT_ID	AIRLINE	ORIGIN_AIRPORT_ID	STG table of FACT	one to one mapping

Continued on next page

Table 3 – *Continued from previous page*

Source	Column	Destination	Column	Type	Transformation
RAW_DATA_AVIATION	DEST_AIRPORT_ID	AIRLINE	DEST_AIRPORT_ID	STG table of FACT	one to one mapping
RAW_DATA_TEMPERATURE	T_ID	AIRLINE	T_ID	STG table of FACT	one to one mapping
RAW_DATA_AVIATION	PASSENGERS	AIRLINE	PASSENGERS	STG table of FACT	one to one mapping
RAW_DATA_AVIATION	CARRIER_NAME	AIRLINE	CARRIER_NAME	STG table of FACT	one to one mapping
RAW_DATA_AVIATION	MONTH	AIRLINE	MONTH	STG table of FACT	one to one mapping
RAW_DATA_AVIATION	YEAR	AIRLINE	YEAR	STG table of FACT	one to one mapping
RAW_DATA_TEMPERATURE	AVG_TEMP	AIRLINE	AVG_TEMP	STG table of FACT	one to one mapping
RAW_DATA_AIRLINES AND REVIEWS	TOTALREVIEWS	AIRLINE	TOTALREVIEWS	STG table of FACT	cast function is use to convert value of varhcar into decimal
RAW_DATA_AIRLINES AND REVIEWS	OVERALL_RATING	AIRLINE	OVERALL_RATING	STG table of FACT	cast function is use to convert value of varhcar into decimal
RAW_DATA_AIRLINES AND REVIEWS	LEGROOM_RATING	AIRLINE	LEGROOM_RATING	STG table of FACT	cast function is use to convert value of varhcar into decimal
RAW_DATA_AIRLINES AND REVIEWS	CUSTOMER_SERVICE_RATING	AIRLINE	CUSTOMER_SERVICE_RATING	STG table of FACT	cast function is use to convert value of varhcar into decimal

Continued on next page

Table 3 – *Continued from previous page*

Source	Column	Destination	Column	Type	Transformation
RAW_DATA_AIRLINE_AND_REVIEWS	CLEANLINESS_RATING	STG_AIRLINE	CLEANLINESS_RATING	STG table of FACT	cast function is use to convert value of varhcar into decimal
RAW_DATA_AIRLINE_AND_REVIEWS	FOOD_AND_BEVERAGE_RATING	STG_AIRLINE	FOOD_AND_BEVERAGE_RATING	STG table of FACT	cast function is use to convert value of varhcar into decimal
RAW_DATA_AIRLINE_AND_REVIEWS	SEAT_RATING	STG_AIRLINE	SEAT	STG table of FACT	cast function is use to convert value of varhcar into decimal
RAW_DATA_AIRLINE_AND_REVIEWS	VALUE_FOR_MONEY_RATING	STG_AIRLINE	VALUE_FOR_MONEY_RATING	STG table of FACT	cast function is use to convert value of varhcar into decimal
RAW_DATA_AIRLINE_AND_REVIEWS	CHECK-IN_RATING	STG_AIRLINE	CHECK-IN	STG table of FACT	cast function is use to convert value of varhcar into decimal
RAW_DATA_AIRLINE_AND_REVIEWS	IN-FLIGHT_RATING	STG_AIRLINE	IN-FLIGHT	STG table of FACT	cast function is use to convert value of varhcar into decimal
RAW_DATA_AIRLINE_AND_REVIEWS	POSITIVE_SENTIMENT	STG_AIRLINE	POSITIVE_SENTIMENT	STG table of FACT	cast function is use to convert value of varhcar into decimal
RAW_DATA_AIRLINE_AND_REVIEWS	NEGATIVE_SENTIMENT	STG_AIRLINE	NEGATIVE_SENTIMENT	STG table of FACT	cast function is use to convert value of varhcar into decimal
RAW_DATA_AIRLINE_AND_REVIEWS	OVERALL_SENTIMENT	STG_AIRLINE	OVERALL_SENTIMENT	STG table of FACT	cast function is use to convert value of varhcar into decimal
AIRLINE	ORIGIN_AIRPORT_ID	FACT_TABLE	ORIGIN_AIRPORT_ID	FACT	one to one mapping

Continued on next page

Table 3 – *Continued from previous page*

Source	Column	Destination	Column	Type	Transformation
AIRLINE	DEST_AIRPORT_ID	FACT_TABLE	DEST_AIRPORT_ID	FACT	one to one mapping
AIRLINE	PASSENGERS	FACT_TABLE	T_ID	FACT	one to one mapping
AIRLINE	CARRIER_NAME	FACT_TABLE	PASSENGERS	FACT	one to one mapping
AIRLINE	AVG_TEMP	FACT_TABLE	AVG_TEMP	FACT	one to one mapping
AIRLINE	TOTALREVIEWS	FACT_TABLE	TOTALREVIEWS	FACT	one to one mapping
AIRLINE	OVERALL_RATING	FACT_TABLE	OVERALL_RATING	FACT	one to one mapping
AIRLINE	LEGROOM_RATING	FACT_TABLE	LEGROOM_RATING	FACT	one to one mapping
AIRLINE	CUSTOMER_SERVICE_RATING	FACT_TABLE	CUSTOMER_SERVICE_RATING	FACT	one to one mapping
AIRLINE	CLEANLINESS_RATING	FACT_TABLE	CLEANLINESS_RATING	FACT	one to one mapping
AIRLINE	FOOD_AND_BEVERAGE_RATING	FACT_TABLE	FOOD_AND_BEVERAGE_RATING	FACT	one to one mapping
AIRLINE	SEAT	FACT_TABLE	SEAT	FACT	one to one mapping
AIRLINE	VALUE_FOR_MONEY_RATING	FACT_TABLE	VALUE_FOR_MONEY_RATING	FACT	one to one mapping
AIRLINE	CHECK-IN	FACT_TABLE	CHECK-IN	FACT	one to one mapping
AIRLINE	IN-FLIGHT	FACT_TABLE	IN-FLIGHT	FACT	one to one mapping
AIRLINE	POSITIVE_SENTIMENT	FACT_TABLE	POSITIVE_SENTIMENT	FACT	one to one mapping
AIRLINE	NEGATIVE_SENTIMENT	FACT_TABLE	NEGATIVE_SENTIMENT	FACT	one to one mapping
AIRLINE	OVERALL_SENTIMENT	FACT_TABLE	OVERALL_SENTIMENT	FACT	one to one mapping

6 ETL Process

The ETL process of the project in SSIS divided into 8 parts in:

Step 1 First part check whether the tables are present into the staging area. Data is truncated from all the staging tables if the tables are available.

Step 2 In the Second Step R code is used to Extract the records from the trip advisor. This code is divide into two parts 1st part extract rating and total number of review information for each airline. Second part of the code is actually extract all the reviews from the trip advisor and generate .csv file for each airline which contains user id, quote and reviews for the airline. These files are stored on the local machine. And in next step these files are picked to do the sentimental analysis of the airline to generate positive, negative and overall sentimental score of the code. Finally this part is merge with the output of 1st part of the code and stored as file on local machine which contains data for the table RAW_DATA_AIRLINE_RATING_REVIEWS. This is completely automated process

Step 3 In the Third step of the ETL raw dataset which is manually stored into the local machine from the Statista is used and performed cleaning operation and created T_ID .This process creates file which contain data for the table [RAW_DATA_TEMPERATURE].

Step 4 Three Raw dataset from Bureau of Transportation Statistic are used. Data cleaning operation is performed to check null value in the dataset. All the dataset are merge into single dataset and records where origin country is US is filter. And single file stored on local machine which is used as data for the table [RAW_DATA_AVIATION]. This step used file created in step 4 to create data file for Raw_Data_airport table Raw_data_Airline.

Step 5 This step 1st check whether dimension tables are present create Dimension table if tables are not created initially and load the data into dimension table from the Staging table Generated from step 1 to 4.

Step 6 This step is used to create table AIRLINE which contain all the measure value and this table must be populated before loading data into the fact table.

Step 7 In this step data in the table AIRLINE is loaded from table RAW_DATA_AVIATION, RAW_DATA_AIRLINE_RATING AND REVIEWS, RAW_DATA_TEMPERATURE. Fact table is loaded from the table airline.

Step 8 In this process cube is automatically deployed to the server. Before applying automation of cube in SSIS workflow, Deployed cube manually through SSAS .

7 Application

1) Relationship between the airline traffic to from USA to Spain and effect of temperature on Passenger traffic.

2) Second Requirement is to identify the rating of the airline and total number passengers travel from this Airlines.

3) Third Requirement is customer opinions about the airline and trips of the airline from USA to Spain for the year 2015-2017.

7.1 BI Query 1: Best quarter of the year to visit Spain considering temperature

Two sources of data used 1st is Bureau of Transportation Statistic and second is Statista to build this query. As we can clearly see from the chart quarter 3 of each Year is the best period to visit to the Spain because of warm temperature, as illustrated in Figure 2.

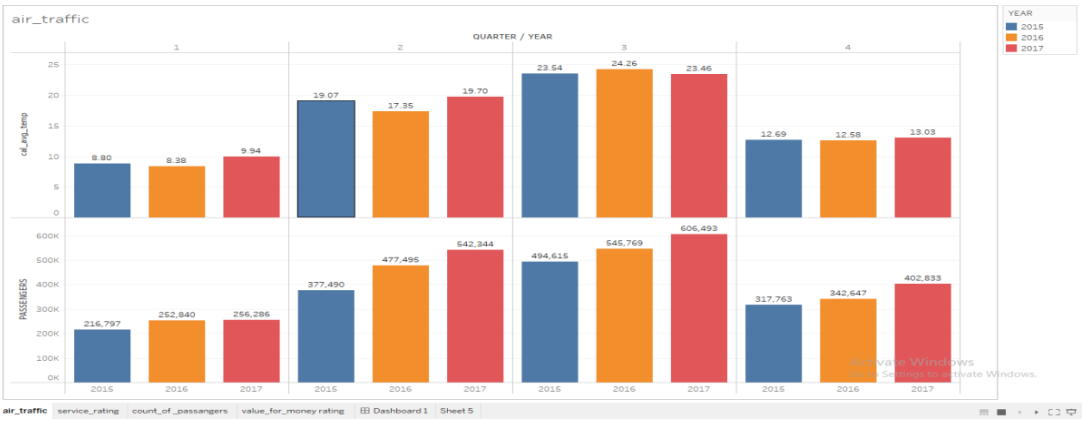

Figure 2: Results for BI Query 1

7.2 BI Query 2: Which among the top three airlines by passengers volume is the best as far as value for money and customer rating is concerned

Data sources used here are Bureau of Transportation Statistic and Trip advisor. Dashboard contain 3 sheets:

1. Count_of_Passenger
2. Service rating
3. value for money rating

From the graph we can easily gather below information and say that Delta Airline is the best among the top three airlines by passenger volume and value for money rating and customer service rating. Illustrated in Figure 3.

7.3 BI Query 3: Which is the worst airline as per the sentimental analysis and what number of trip covered by this airline from the united states to Spain

Data sources used are Bureau of Transportation Statistic and Trip advisor. The worst airline is the airline which is having maximum number of negative overall score And from the bar chart Norwegian Air Shuttle ASA has maximum negative overall score and total number of trips by this airline to Spain from USA are 26. Illustrated in Figure 4.

7.4 Discussion

With the help of BI queries result we can see that almost all the airline among the present airline from US to Spain have negative opinions of the Customers and this shows

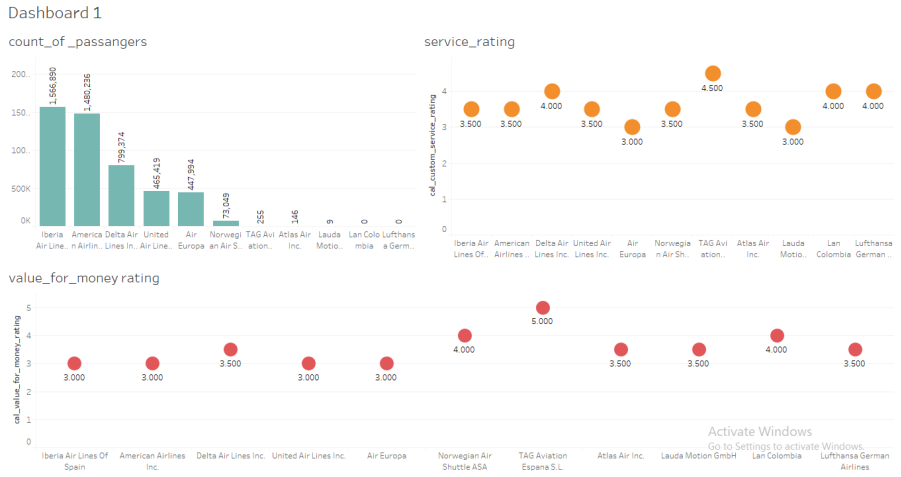

Figure 3: Results for BI Query 2

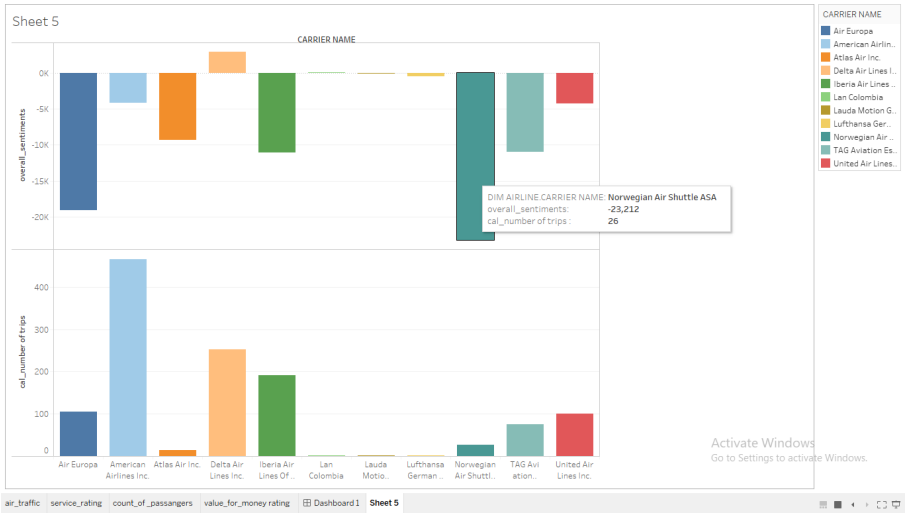

Figure 4: Results for BI Query 3

the frustration of the customer for the different aspect such value for money, legroom rating, Customer service rating .Only delta airline has positive opinion among the top 3 airlines considering volume of Passenger. This analysis also shows that number of passengers increases in the 3rd quarter of the every year in Spain and this useful for the airline companies to increase the flight frequency in 3rd quarter between the USA and Spain.

8 Conclusion and Future Work

As we know air traffic is increasing day by day and lot of people prefer airline as first choice of travelling so this project is analysing the customer traffic from spin to USA for the 2015-2017 And the overall opinion of the customer about the Airline. And as per the analysis on available dataset it is Important for the airline companies need lot of improvement in different aspects to provide good satisfactory services to the customer. I will implement SCD-2 in this project and also use DTExec.exe to trigger SSIS workflow from the Command line.

References
Appendix

R code example

Airline Rating:

```
# code to fetch data from the website #Tripadvisor

# install.packages("rvest")
# install.packages("purrr")
# install.packages("XML")
# install.packages("plyr")
# install.packages("tidytext")
# install.packages("reshape")
# install.packages("fetch")
# install.packages("textcat")
# install.packages("cld2")
# install.packages("cld3")
# install.packages("tidyverse")


library(rvest)
library(XML)
library(plyr)
library(dplyr)
library(purrr)
library(cld2)
library(cld3)
library(tidytext)
library(tidyr)
library(tidyverse)




# scrap airline name and Number of Totalreviews.
# values in x vector is combine with url_base and gather the airline name,to
# map df applay  function(i) to the vector  x and create dataframe with colu
# sprintf  returns a  vector which is combination of url_base text and varia
# Read_html function will read the content from a .html file assign this val
# html_nodes function select value of the sepecific class from the html_docu
# html_text extract text from the html nodes output.
# Gsub is use to remove the string "reviews" assosiated with each observatio

setwd("C:\\Users\\MOLAP\\Documents\\R\\TRIP_ADVISOR_REVIEWS")
x<-c("d8729113","d8729177","d8729108","d8729026","d15052991","d8729060","d87
url_base <- "https://www.tripadvisor.ie/Airline_Review-%s"

map_df(x, function(a) {
```

```
  webpage <- read_html(sprintf(url_base, a))
  data.frame(AIRLINE_NAME=html_text(html_node(webpage, ".heading_height")),
             Totalreviews=gsub("reviews","",html_text(html_nodes(webpage, ".
             stringsAsFactors=FALSE)
})-> AIRLINE_DETAILS

AIRLINE_DETAILS$Totalreviews<- gsub(",","",AIRLINE_DETAILS$Totalreviews)

#code is used to scrap Rating Names for the airlines.Consider only one airli
#all the airlines.

url1 <- 'https://www.tripadvisor.ie/Airline_Review-d8729160-Reviews-Swiss-In
webpage1<-read_html(url1)
RATING_NAME<-rbind(html_text(html_nodes(webpage1,"#AIRLINE_DETAIL_MAIN_WRAPP
RATING_NAME<-c("overall_rating",RATING_NAME)

# scrap overall_rating,Customer Servic,Legroom,Seat Comfort,Cleanliness,Valu
# Check-in and Boarding,Food and Beverag,In-flight entertainment (WiFi, TV,
# past funuction combine tripadvisor url with element of x vector for each p
# select nodes with class .ui_bubble_rating and create list for all the rati
# Rating contain list of 11 list one list for each  airline page.

Rating <- lapply(paste0('https://www.tripadvisor.ie/Airline_Review-', x),
                 function(url){
                   url %>% read_html() %>%
                     html_nodes(".ui_bubble_rating")
                 })

# Rating output is list of list. Each list contain xml_nodeset.
# Need to extraact xml attributes from xml_nodeset for each list to get the
# lapply is used to apply function to the xml_attr of each rating list to cr
# extracting  1st 9 rows from alt column of each dataframe as contains ratin



df.RATING=c()
for(i in 1:11)
{
  df.RATING[[i]]<-bind_rows(lapply(xml_attrs(Rating[[i]]),function(x) data.f
  df.RATING[[i]]<-gsub("of 5 bubbles","",head(df.RATING[[i]]$alt,9))
}

#combine all char lists in df.rating into one variable of the data frame.
df.RATING2<-data.frame(cbind((df.RATING)))

# craeting dataframe all_rating
ALL_RATINGS<-data.frame(t(sapply(df.RATING,c)))

#asssign column name to the data frame
colnames(ALL_RATINGS)<-c(RATING_NAME)
```

```r
# combine data frame Airline details which contains airlinename and total nu
# all_ratings which contains 9 type of ratings for each airline.
AVIATION_DF<-cbind(AIRLINE_DETAILS,ALL_RATINGS)

# remove the unwanted right spaces from the column airline name and "\n"
AVIATION_DF$AIRLINE_NAME<-trimws(AVIATION_DF$AIRLINE_NAME, which = c( "right
AVIATION_DF$AIRLINE_NAME<-gsub("\n","",AVIATION_DF$AIRLINE_NAME)

# changing the airline name as per the names in DIM_airline table.
# Due to this it is easy to load the rating values and total reviews number


AVIATION_DF$AIRLINE_NAME<-revalue(AVIATION_DF$AIRLINE_NAME,c("Lufthansa" = "
                                                             "United␣Airline
                                                             "LAN␣Airlines␣(
                                                             "Atlas␣Air" ="A
                                                             "TAG" = "TAG␣Av
                                                             "Delta␣Air␣Line
                                                             "American␣Airli
                                                             "Iberia"= "Iber
                                                             "Air␣Europa" =
                                                             "Laudamotion"=
                                                             "Norwegian" ="N


#SENTIMENTAL ANALYSIS CODE #

# SCRAP REVIEWS FOR AIRLINE "AIR_EUROPA" FROM TRIP ADVISOR.AND THIS MODULE I
# THERE ARE MORE THAN  120000 REVIEWS ON TRIP ADVISOR FOR ALL 10 AIRLINES AN
# Total number of pages of "AIR_EUROPA" on tripadvisor are 108 each page con
# x is vector which contain values from 10 to 1080 and seperated by 10
# map df applay  function(i) to the vector  x and create dataframe.
# sprintf  returns a  vector which is combination of url_base text and varia
# Read_html function will read the content from a .html file assign this val
# html_nodes function select value of the sepecific class from the html_docu
# html_text extract text from the html nodes output.


x<-seq(10,1080,by=10)
url_base <- "https://www.tripadvisor.ie/Airline_Review-d8729002-Reviews-or%d
map_df(x, function(i) {

  webpage <- read_html(sprintf(url_base, i))
  reviews<-html_nodes(webpage,"#REVIEWS␣.innerBubble")
  id<-html_node(reviews,".quote␣a")%>%
    html_attr("id")
  quote<-html_text(html_node(reviews ,".quote␣span"))
  review <-html_text(html_node(reviews,".entry␣.partial_entry"))
  df<-data.frame(id,quote,review,Airline_name="Air␣Europa", stringsAsFactors
}) -> Air_Europa
```

```
write.csv(Air_Europa,"Air_Europa.csv",row.names = F )

# scrapped data of Airline review from trip advisor and is use as the input
# This function will pick all the files from the working directory with file

temp = list.files(pattern="*.csv")
for (i in 1:length(temp))
  assign(temp[i], read.csv(temp[i],stringsAsFactors = F))


# creating single data frame from all the dataframe

DATA<-data.frame(rbind(Atlas_Air.csv,Air_Europa.csv,American_Airlines.csv,De
                       Iberia.csv,LAN_Airlines.csv,Laudmotion.csv,
                       Lufthans.csv,TAG.csv,united_airlines.csv,Norwegian.cs


DATA$review<-gsub(pattern="\\W",replace="␣",DATA$review)
DATA$review<-gsub(pattern="\\d",replace="␣",DATA$review)
DATA$review<-tolower(DATA$review)
DATA$review<-gsub(pattern = "\\b[a-z]\\b{1}",replace="␣",DATA$review)



# language detection mechanisam is used cld2 and cld3 are most reliable comp
DATA<-DATA %>% mutate(
airlin_name=DATA$Airline_name,
cld2=cld2::detect_language(text = review ,plain_text = FALSE),
cld3=cld3::detect_language(text = review)) %>%
select(review,cld2,cld3,Airline_name) %>%
filter(cld2=="en"& cld3=="en")



#data<-DATA rm(data1)

text_df<-data_frame(text=DATA$review,airline_name=DATA$Airline_name)
afin <-get_sentiments("afinn")
bing <-get_sentiments("bing")
nrc<-get_sentiments("nrc")

text_df<-unique(text_df)
text_df<-text_df %>% unnest_tokens(word,text)
text_df<-unique(text_df)
text_df$ID <- seq.int(nrow(text_df))


# afinsent<-text_df %>%
# inner_join(afin) %>%
# summarise(sentiment= sum(score))
#bing
```

```
bingSent <-text_df %>%
inner_join(bing)%>%
spread(sentiment, ID, fill = 0) %>%
mutate(sentiment = positive - negative)
final_sentiment<-aggregate(bingSent[, 3:5], list(bingSent$airline_name), mea
colnames(final_sentiment)<-c("AIRLINE_NAME","POSITIVE","NEGATIVE","SENTIMENT
final_sentiment<-cbind(AVIATION_DF,final_sentiment[,2:4])
setwd("C:\\Users\\MOLAP\\Documents\\R")
write.csv(final_sentiment,"RAW_DATA_AIRLINE_RATING.csv",row.names = F )

# all_positive<- aggregate(bingSent$positive, by=list(Category=bingSent$airl
# all_negative<- aggregate(bingSent$negative, by=list(Category=bingSent$airl
# all_sentiments<-aggregate(bingSent$sentiment, by=list(Category=bingSent$ai
# final_sentiment<- data.frame(cbind(all_positive$Category,all_positive$x,al
# colnames(final_sentiment)<-c("AIRLINE_NAME","POSITIVE","NEGATIVE","SENTIME
# as.data.frame((table(bingSent$airline_name)))
# aggregate(bingSent[, 3:5], list(bingSent$airline_name), mean)
# sapply(split(bingSent[,3:5],bingSent$airline_name)),mean
# by(bingSent[3:5],bingSent$airline_name,mean)

Raw Airline Data:

# This Code will pick Raw data Aviation file and create RAW_DIM_AIRLINE

library(dplyr)
setwd("C:\\Users\\MOLAP\\Documents\\R")
RAW_DATA_AIRLINES<-read.csv("Raw_DATA_AVIATION.csv") # read file Raw_DATA_AV
RAW_DATA_AIRLINES<-distinct(RAW_DATA_AIRLINES[,c(2:4)]) # select distinct va
                                                        # column 2:4 of the
colnames(RAW_DATA_AIRLINES)<-c("AIRLINE_ID","CARRIER_CODE","CARRIER_NAME") #
write.csv(RAW_DATA_AIRLINES,"RAW_DATA_AIRLINE.CSV",row.names=FALSE) # wrire


Raw Data Airport:

# code to ceate RAW_DIM_AIRPORT
# This will pick Raw data Aviation file and create RAW_DIM_AIRPORT

library(dplyr)

setwd("C:\\Users\\MOLAP\\Documents\\R")   # set working directory as C:\\Use
RAW_DATA_AIRPORT  <-read.csv("Raw_DATA_AVIATION.csv") # pick RAW_DATA_AVIATI
sapply(RAW_DATA_AIRPORT, function(x) sum(is.na(x)))  # TO check the NA value
RAW_DATA_AIRPORT1 <-distinct(RAW_DATA_AIRPORT[,c(6:10)]) # select distinct v
RAW_DATA_AIRPORT2 <-distinct(RAW_DATA_AIRPORT[,c(11:15)] ) # select distinct
colnames(RAW_DATA_AIRPORT1)<-c("COUNTRY_CODE","COUNTRY_NAME","CITY_CODE","CI
colnames(RAW_DATA_AIRPORT2)<-c("COUNTRY_CODE","COUNTRY_NAME","CITY_CODE","CI
RAW_DATA_AIRPORT<-rbind(RAW_DATA_AIRPORT1,RAW_DATA_AIRPORT2) # binding two d
                                                             # unique record
write.csv(RAW_DATA_AIRPORT, file = "Raw_DATA_AIRPORT.csv",row.names=FALSE)
```

SQL:

```
fact stage table load
INSERT INTO [dbo].[AIRLINE]
(
[AIRLINE\_ID],
[ORIGIN\_AIRPORT\_ID],
[DEST\_AIRPORT\_ID],
[T\_ID],
[PASSENGERS]  ,
[CARRIER\_NAME],
[YEAR],
[MONTH],
[TOTALREVIEWS],
[AVG\_TEMP],
[OVERALL\_RATING],
[LEGROOM\_RATING],
[CUSTOMER\_SERVICE\_RATING],
[CLEANLINESS\_RATING],
[FOOD\_AND\_BEVERAGE\_RATING],
[SEAT COMFORT RATING],
[VALUE\_FOR\_MONEY\_RATING],
[CHECK-IN AND BOARDING] ,
[IN-FLIGHT ENTERTAINMENT(WIFI, TV, FILMS)],
[POSITIVE\_SENTIMENT],
[NEGATIVE\_SENTIMENT],
[OVERALL\_SENTIMENT]
)
SELECT
A.[AIRLINE\_ID],
A.[ORIGIN\_AIRPORT\_ID],
A.[DEST\_AIRPORT\_ID],
B.[T\_ID],
A.[PASSENGERS],
A.[CARRIER\_NAME],
A.[YEAR],
A.[MONTH],
E.Totalreviews,
CAST(B.AVERAGE\_TEMP as [dec](16,2)),
CAST(E.OVERALL\_RATING as [dec](16,2)),
CAST(E.[Legroom] as [dec](16,2)),
CAST(E.[Customer Service] as [dec](16,2)),
CAST(E.[Cleanliness] as [dec](16,2)),
CAST(E.[Food and Beverage] as [dec](16,2)),
CAST(E.[Seat Comfort] as [dec](16,2)),
CAST(E.[Value for Money] as [dec](16,2)),
CAST(E.[Check-in and Boarding] as [dec](16,2)),
CAST(E.[In-flight entertainment (WiFi, TV, films)] as [dec](16,2)),
CAST(E.[POSITIVE] as [dec](16,2)),
CAST(E.[NEGATIVE] as [dec](16,2)),
```

```
CAST(E.[SENTIMENTS] as [dec](16,2))
FROM [dbo].[RAW\_DATA\_AVIATION] A
INNER JOIN  [dbo].[RAW\_DATA\_AIRLINE\_RATING AND REVIEWS] E
ON A.CARRIER\_NAME=E.AIRLINE\_NAME
INNER JOIN [dbo].[RAW\_DATA\_TEMPERATURE] B
ON A.MONTH=B.MONTH AND A.YEAR =B.YEAR
fact table load
USE [AIRLINE]
GO
TRUNCATE TABLE [dbo].[FACT\_TABLE]
INSERT INTO [dbo].[FACT\_TABLE]
(
[AIRLINE\_ID],
[ORIGIN\_AIRPORT\_ID] ,
[DEST\_AIRPORT\_ID],
[T\_ID],
[PASSENGERS],
[TOTALREVIEWS],
[AVG\_TEMP],
[OVERALL\_RATING] ,
[LEGROOM\_RATING],
[CUSTOMER\_SERVICE\_RATING],
[CLEANLINESS\_RATING],
[FOOD\_AND\_BEVERAGE\_RATING],
[SEAT COMFORT RATING],
[VALUE\_FOR\_MONEY\_RATING],
[CHECK-IN AND BOARDING],
[IN-FLIGHT ENTERTAINMENT(WIFI, TV, FILMS)],
[POSITIVE\_SENTIMENT],
[NEGATIVE\_SENTIMENT],
[OVERALL\_SENTIMENT]
)
SELECT
A.[AIRLINE\_ID],
A.[ORIGIN\_AIRPORT\_ID],
A.[DEST\_AIRPORT\_ID],
A.[T\_ID],
[PASSENGERS],
[TOTALREVIEWS],
[AVG\_TEMP],
[OVERALL\_RATING] ,
[LEGROOM\_RATING],
[CUSTOMER\_SERVICE\_RATING],
[CLEANLINESS\_RATING],
[FOOD\_AND\_BEVERAGE\_RATING],
[SEAT COMFORT RATING],
[VALUE\_FOR\_MONEY\_RATING],
[CHECK-IN AND BOARDING],
[IN-FLIGHT ENTERTAINMENT(WIFI, TV, FILMS)],
A.[POSITIVE\_SENTIMENT],
A.[NEGATIVE\_SENTIMENT],
```

```
A.[OVERALL\_SENTIMENT]
FROM [dbo].[AIRLINE] A
INNER JOIN [dbo].[DIM\_AIRLINE] B
ON A.AIRLINE\_ID=B.AIRLINE\_ID
INNER JOIN [dbo].[DIM\_AIRPORT] C
ON A.ORIGIN\_AIRPORT\_ID = C.AIRPORT\_ID
INNER JOIN [dbo].[DIM\_AIRPORT] D
ON A.DEST\_AIRPORT\_ID = D.AIRPORT\_ID
INNER JOIN [dbo].[DIM\_TIME] E
ON A.T\_ID =E.T\_ID
```

References: Peter DOMONKOS1, Xavier FARR2, Juan Antonio DURO2 1Centro de Cambio Climtico, Univ. Rovira i Virgili, Campus Terres de lEbre, Tortosa 2 Departamento de Economa y CREIP, Univ. Rovira i Virgili, Tarragona

http://fundacion.usal.es/conaec/pendrive/ficheros/ponencias/ponencias3/06-Impactos.pdf

Gao, B., Hu, N. and Bose, I. (2017) Follow the herd or be myself? An analysis of consistency in behavior of reviewers and helpfulness of their reviews, Decision Support Systems, 95, pp. 111. doi: 10.1016/j.dss.2016.11.005.