National College of Ireland

Data Warehousing and Business Intelligence Project

on

Analysis of Top European Airlines based on Revenue, Number of Employees, Number of Passengers and Service Quality.

OLUWASUREFUNMI MARIA IDOWU

18158188

MSc Data Analytics – 2019/20

Submitted to: Sean Heeney

National College of Ireland
Project Submission Sheet – 2019/2020
School of Computing

Student Name:	Oluwasurefunmi Maria Idowu
Student ID:	18158188
Programme:	MSc Data Analytics
Year:	2019/20
Module:	Data Warehousing and Business Intelligence
Lecturer:	Sean Heeney
Submission Due Date:	12/04/2019
Project Title:	Analysis of Top European Airlines based on Revenue, Number of Employees, Number of Passengers and Service Quality

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 01/21/2018
- ☐ 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 Top European Airlines based on Revenue, Number of Employees, Number of Passengers and Service Quality

OLUWASUREFUNMI MARIA IDOWU
18158188

April 12, 2020

Abstract

The aim of this project is to analyze the relationship between airlines based on their total number of passengers, number of employees, passenger handling rights, service quality and revenue. It places strong focus on analyzing the customer service quality of the largest airlines in the world, based on their number of employees. Furthermore it hopes to help airline companies use their data in a more resourceful way. It also further emphasizes the need for customer service quality in airline companies, proving that quality is better than quantity. This means that an airline with more resources might not necessarily render better services but if it does not use its resources in the right way.

1 Introduction

'Three stone cutters were asked about their jobs. The first said he was paid to cut stones. The second replied that he used special techniques to shape stones in an exceptional way, and proceeded to demonstrate his skills. The third stone cutter just smiled and said: 'I build cathedrals.' Unknown' (Sturdy and Graham, 2012, p. 144) According to Wixom, B. et al. (2008), Data Warehousing is not a destination, but rather regarded as a journey. It is an evolution that never stops. Maintaining airline customers, as well as prevailing over rivals in the industry is a very important in developing markets as well as dying markets. Sales technologies including the simplicity of their online experience, revenue management and Salesforce automation, Frequent Flyer Program, are key instruments in any aircrafts competitive technique. However, there are better alternatives that are existent in recent times. To have a deeper understanding of your customers (and frequent passengers), the practice of data warehousing to deliver factual business intelligence cannot be overemphasized. (Williams and OConnell, 2011)

The main objectives of this project are the following: a. To show that the airlines with more passengers may not necessarily have a higher revenue. b. To analyze the service quality of airlines based on their number of employees and to show that less is more, sometimes. c. To analyze the handling of airline passenger rights, based on number of passengers and number of employees.

The core business requirements for this project are as follows:

Source	Type	Brief Summary
Statista	Structured	I used Statista because it is an authentic and widely used source of data which can be manipulated to create useful business models. It contains the very key data which helped me produce reasonable information. More so, it was a necessary requirement as the licence was paid for by the school. One csv file (Top 100 Companies: Airlines) was downloaded from this source.
Wikipedia	Unstructured	This is one of the most widely used data sources on the web. Generally speaking, it is very easy to use and it usually contains very precise information which are easy to understand. This is the major reason I chose it. It was scraped and converted into a csv file. The process is explained later in the project.
Data.world	Structured	This is my major data source. I especially picked out this dataset because it contains the major data I was looking for. Also, the publisher, Airhelp, is a well renowned company with a huge passion for airlines and air service quality

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

(Req-1) Accessibility: This means that the data and information produced are simple and easy to understand and will serve as a foundation for simple decision making.

(Req-2) Adaptive and Resilient: This means that the data is tolerable to unavoidable business changes and the data warehouse changes will not invalidate the data.

(Req-3) Improved decision making is aided with the right data and the correct use of analytical tools like described in this project.

(Req-4) Security: The data warehouse is secure and cannot be accessed by unauthorized users.

(Req-5) Acceptability: The data is reliable and trustworthy.

This project aims to build a data warehouse that can be used by airlines in order to facilitate better service quality to their customers.

2 Data Sources

I used three sources of data for this project and all of them do not have ethical issues.

2.1 Source 1: Statista

The statista airlines dataset downloaded as an Excel file from: `https://www.statista.com/study/44541/top-100-airlines/` It provides 41 columns of information on the top 100 airline companies worldwide.

This includes the rank, company name, listing ID, exchange, primary ISIC code, country code, street, city, region, zip code, phone, website, IR website, number of employees from 2013 to 2017 and revenue within the same period.

However, only 3 columns were relevant to the project. This will be explained later in the transformation process.

The dataset meets all business requirements listed in Section 1 in the following ways. Due to its historical nature, its authenticity can be proven. It is a structured data, and was obtained within the last year (March 2019). It will be used in collaboration with other data to achieve the first objective of this project.

2.2 Source 2: Wikipedia

The second dataset was web scraped from Wikipedia at https://en.wikipedia.org/wiki/List-of-largest-airlines-in-Europe via https://wikitable2csv.ggor.de/ which converted it into a csv file. It provides 22 rows of information on the largest airlines in Europe.

These rows include the rank, country, airline, number of passengers from 2006 to 2019, fleet size and destinations, source, alliance and association. The transformation process will be explained later in this project.

However, relevant to the project are 3 columns: airline, country and most recent number of passengers.

This dataset meet the business requirement explained in section 1 in the following ways: It is an unstructured and historical data, from a good source. It will be used with other data to achieve the objectives of this project.

2.3 Source 3: Data.world

My third and final dataset was sourced from https://data.world/dataremixed/2018-airline-rankings-by-airhelp and published by AirHelp in June 6 2018 at https://www.airhelp.com/da/airhelp-score/rangliste-for-flyselskabe.

The data consists of 8 coulmns including the rank, airline, country, service quality, handling of passenger rights, airhelp score and year.

However, the relevant columns used for this project were airline, country, punctuality, service quality and handling of passenger rights.

This dataset meets the business requirement as discussed in the first section such that it is structured and from an authentic source. Being the major dataset, it will be used with other data to achieve the second and third objective of this project.

3 Related Work

Kimball (2008) has clarified that asides the use of text and numbers, comparisons can be better interpreted graphically. In his words, The true goal of the data warehouse and business intelligence system is to assist the end user make decisions. I completely agree because, if airlines could visualize their important data related to their field, it will

aid good decision making and thereby contribute to the overall progress of the business. Similarly, Revels and Nussbaumer (2013) addressed data warehousing advantages in airline industry as well as its challenges. They also encouraged airlines to adapt predictive analysis to develop their decision making techniques in order to gain strong competitive advantage.

In the aviation industry, the optimization of the profits acquired from each flight, is solely identified with the optimization of revenues, due to the fact that the expenses brought about are basically fixed at any rate, temporarily. D'Alfonso et al., (2011)

A similar research has been performed on the effect of service quality (pre-flight, in-flight and post-flight services). Findings indicated that these services had a statistically significant impact on passenger satisfaction, and customer satisfaction as a mediating variable also had a significant relationship with customer loyalty. This contrast here is the methodology used (SPSS) and the sources of data (primary and secondary). Using primary source of data (survey) she also deduced that passengers perceived satisfaction in different ways based on the several services of airline, for instance, food quality or onboard and offboard facilities. (Namukasa, 2013)

Andronie M (2015) investigated the diverse business intelligence accessible in the market, stating that airlines with expansive volumes of data can extract data either by using a general or specific industry Business Intelligence tool. He therefore concluded that using a general BI tool is not as effective as an airline industry based BI system because the latter will be able to deal with specific issues thereby solving specific problems related to similar companies.

James J. H. and Liou (2008) conducted a study combining the Factor Analysis and Variable Consistency Dominance-based Rough Set Approach (VC-DRSA) as an operational apparatus for anticipating the acquiring customer choices in the airline industry. The determined principles can assist the aircraft with creating legitimate systems for various classes of customers and improve the customer relationship management within airlines.

In my research, I will explain how to build a data warehouse for an airline to show the correlations between airline employees and service quality, passenger handling rights and number of passengers, as well as passenger count and revenue. To carry out this analysis, I will be using the following tools: R Studio for data cleaning and transformation, Microsoft SQL Server Management Studio (SSMS) for management analysis services, and finally Microsoft Visual Studio Community 2017 and Tableau for data visualization.

4 Data Model

In this section, I provide details of my dimensions.

As seen in Figure 1, I have 2 dimensions and one Fact table which were carefully thought out in order to meet the business requirements as discussed earlier.

4.1 Dimension: Airline

Airline ID: This is the primary key of the Airline dimension. It was generated to create a relationship between the Fact Table and the dimension table. It has numeric values. It was mapped from the ID column in the Employees table. It was added to the dimension to aid cube deployment process. Airline: This column shows the full names of the airlines. It was sourced from the Employees table. It offers a form of drill drown, as users can

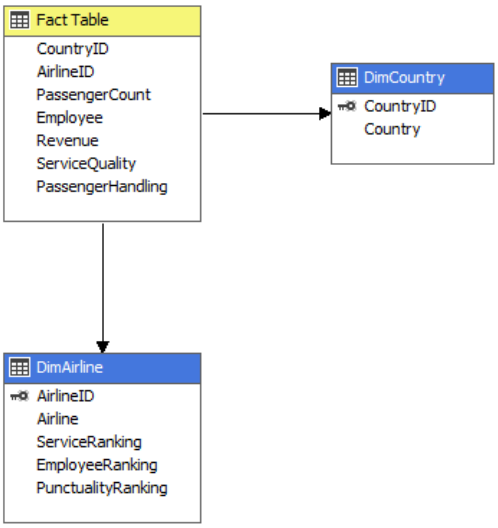

Figure 1: Star Schema showing the attributes of my dimensions, including the Fact Table

view data based on the name of the airline and it is necessary to distinguish information between different airlines. It is also used to describe all the facts in the fact table. Airline (which represents the airline name) is located in all 3 tables but was particularly sourced from the Employee table because it has the most number of rows. ServiceRanking, EmployeeRanking and PunctualityRanking: These columns were generated to define the quality of the dimension.

4.2 Dimension: Country

Country ID: This is the primary key of the Country dimension. It was generated to enable a relationship between the Fact Table and the dimension table. It has numeric values ranging from 1 to 71 and was sourced from the Ranking table. It was added to the dimension to aid cube deployment process. Country: This shows the country name where the airlines were established. It was sourced from the Ranking table. Like the airline name, it offers a form of drill down as users can view the location of these airlines. The column country is located in the 2 tables. The Ranking table and the Passenger table.

4.3 Fact Table

PassengerCount: This is the most recent number of each airline passengers in millions. It is was sourced from https://en.wikipedia.org/wiki/List-of-largest-airlines-in-Europe. (Wikipedia). It is mapped from the column Passengers in the Passenger table and coerced into the fact tale. It is necessary to show this column as it meets the basic objective of this project.

Service Quality: This shows the overall service quality of each airline. It was sourced from https://data.world/dataremixed/2018-airline-rankings-by-airhelp (Data.world). It is important to show this column in order to meets the main objective of this project.

Employee: This is the latest number of employees present in each airline. It was

sourced from It is was sourced from https://www.statista.com/study/44541/top-100-airlines/ (Statista). It is necessary to show this column as it meets the first objective of this project.

Revenue: This is the latest value each airline's revenue in US dollars. It was sourced from https://www.statista.com/study/44541/top-100-airlines/ (Statista). It is mapped from the column Revenue in the Employee table and coerced into the fact tale. It is necessary to show this column as it meets one of the objectives of this project.

PassengerHandling: This is the rating (out of 10) of the way airlines handle their rights and claims of their passengers as regards delays, rescheduling and cancellation of flights. It was sourced from https://data.world/dataremixed/2018-airline-rankings-by-airhelp. It is necessary to show this column as it meets the third objective of this project.

5 Logical Data Map

Below is the description of my logical data map.

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

Source	Column	Destination	Column	Type	Transformation
Statista	Airline	DimAirline	Airline	Dimension	Spaces were replaced with underscores for uniformity purposes. Airlines with inconsistent and duplicate values were removed
Statista	AirlineID	DimAirline	AirlineID	Fact and Dimension	Generated to create relationship with the fact table $
Statista	Employee Ranking	DimAirline	Employee Ranking	Dimension	Created for quality definition of the dimension
Data.world	Service Quality	FactTable	Service Ranking	Dimension	Created for quality definition of the dimension $
Data.world	Service Quality	FactTable	Service Quality	Fact	No transformation needed $
Data.world	Punctuality Ranking	DimAirline	Punctuality Ranking	Dimension	Created for quality definition of the dimension
Wikipedia	Passengers	FactTable	Passenger Count	Fact	The missing values for passengers from 2015 to 2018 were replaced with the median of all 4 years. The Column name was changed from 2018 to Passengers as I was dealing with only the recent number of passengers.$
Wikipedia	Country	DimCountry	Country	Dimension	The correct country was selected for rows with more than one country

Continued on next page

Table 3 – *Continued from previous page*

Source	**Column**	**Destination**	**Column**	**Type**	**Transformation**
Statista	Employee	FactTable	Employee	Fact	The column showing the latest number of employees was changed from 'latest' to 'Employees' and all commas in the values were removed
Statista	Revenue	FactTable	Revenue	Fact	The column showing the latest Revenue was changed from 'latest' to 'Revenue' and all commas in the values were removed
Data.world	Handling of passenger rights	FactTable	Passenger Handling	Fact	No transformation needed
Data.world	CountryCode	DimCountry	Country	Dimension	Each country codes were changed manually to the full names of the countries
Statista	Country	DimCountry	Country	Dimension	No transformation needed but was used in the SQL joining process
hline					

6 ETL Process

This section describes the high-level strategy of the ETL process, very specific details will be highlighted in the video.

6.1 Cleaning, Transformation and Challenges

Statista: This data was downloaded as a csv file and loaded into R studio for cleaning. a. Before it was loaded into R studio for cleaning, the first sheet including the meta data was removed as it was irrelevant to the project. b. All unnecessary columns were removed, leaving only the relevant columns to the project. c. The column names were renamed to suit the other datasets and present them in an orderly fashion. d. All spaces in the airline column were replaced with underscores so that they could be joined swiftly in SQL. e. The file was renamed 'Employee' and exported back to the folder.

Wikipedia: a. This data was sourced from https://en.wikipedia.org/wiki/List-of-largest-airlines-in-Europe and extracted via https://wikitable2csv.ggor.de then loaded into R studio b. In this dataset, I had to deal with missing values of passengers in several years by calculating the median of each airline's passengers from 2015 to 2018. I did not consider the other years (2006-2014 and 2019) for cleaning because the missing data was too much, so I simply deleted those columns . Since I was dealing with only latest values, I used only values in 2018 and changed the column name to 'Passengers.' c. I changed 'Airline/Holdings to 'Airline' d. Countries with more than one country value were corrected

Data.world: This dataset was sourced from https://data.world/dataremixed/2018-airline-rankings-by-airhelp and was downloaded as a csv file. It did not need much cleaning as it contained no missing values and all column names were correct and suitable for the project to go on successfully.

6.2 Staging

After the cleaning and transformation, I exported the files into Visual Studio in order to integrate it into the database 1. A connection was made between the database Airline (created in SSMS via automation process in R) 2. The connection for the 3 datasets (raw csv files) was made (Flat file Source to OLEDB Destination) 3. I brought the R scripts into SSIS by using the excute process task component in the SSIS Toolbox 4. SQL statements were used to create the Dimensions (DimAirline and DimCountry) and excute SQL task companent was used for this purpose. 5. In order to created the fact table, all tables (Employee, Passenger and Ranking) were joined in SQL based on airline and countries. 6. The cube was processed using the SSAS environment. 7. The data source and view were created and the dimension tables were created from the existing dimension tables in SSMS and SSIS. 8. By choosing the correct measure (Fact Table) and Dimensions (DimCountry and DimAirline), the cube was successfully deployed. The tables had to be truncated for the process to be executable numerous times, and also to avoid data duplication. The cube was processed and deployed and the automation was complete.

6.3 Degree of Automation

The following automation took place in order to make the ETL process run multiple times.

I automated the cleaning and transformation process in R studio using the Execute Processing task component of SSIS. The cube deployment was included in the automation process as well. This was done through the Analysis Processing Task Component of SSIS.

7 Application

This section will brief you through my case studies and how meets the business requirements noted in Section 1.

Case Study 1: This case study meets the following requirements Accessibility: This means that the data and information produced are simple and easy to understand and will serve as a foundation for simple decision making. Adaptive and Resilient: This means that the data is tolerable to unavoidable business changes and the data warehouse changes will not invalidate the data. Improved decision making is aided with the right data and the correct use of analytical tools like described in this project. Security: The data warehouse is secure and cannot be accessed by unauthorized users. Acceptability: The data is reliable and trustworthy.

It is also non trivial as it comes from different data sources.

Case Study 2: This case study meets the following requirements Accessibility: This means that the data and information produced are simple and easy to understand and will serve as a foundation for simple decision making. Adaptive and Resilient: This means that the data is tolerable to unavoidable business changes and the data warehouse changes will not invalidate the data. Improved decision making is aided with the right data and the correct use of analytical tools like described in this project. Security: The data warehouse is secure and cannot be accessed by unauthorized users. Acceptability: The data is reliable and trustworthy.

Case Study 3:This case study meets the following requirements Accessibility: This means that the data and information produced are simple and easy to understand and will serve as a foundation for simple decision making. Adaptive and Resilient: This means that the data is tolerable to unavoidable business changes and the data warehouse changes will not invalidate the data. Improved decision making is aided with the right data and the correct use of analytical tools like described in this project. Security: The data warehouse is secure and cannot be accessed by unauthorized users. Acceptability: The data is reliable and trustworthy.

7.1 BI Query 1: Which Airline has the most number of passengers and how does it affect its Revenue?

For this query, the contributing sources of data are data.world and Statista and the contributing columns are PassengerCount (Fact Table), Revenue(FactTable), Country (DimCountry) and Airline (DimAirline). The general findings are that airlines with more passengers have a higher revenue, but there was an exception in the case of. In specific cases, it was found that the International Group Airlines -IAG (which has subsidiaries such as British Airways and Iberia) takes the lead with about almost 2 million passengers,

having the highest revenue as well. However some companies like Easyjet has lower passenger numbers but higher revenue.

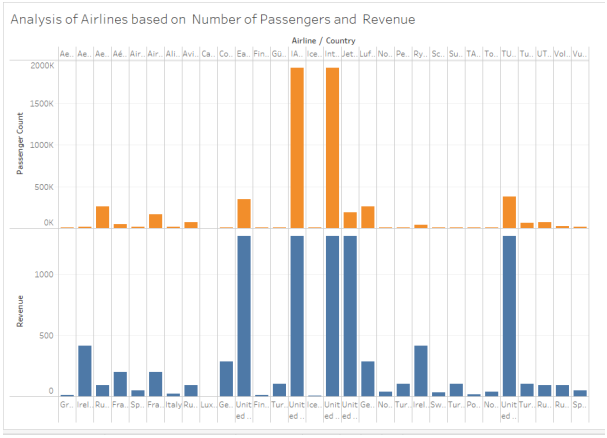

Figure 2: Results for BI Query 1

7.2 BI Query 2: Which airline company has the most number of employees and what is its Service Quality?

For this query, the contributing sources of data data.world and are and the contributing columns are Service Quality(Fact Table), Employee (FactTable) Country (DimCountry) and Airline (DimAirline). The general findings as illustrated in Figure **??** are that Easy Jet takes the lead with the highest number of employees and highest service quality while TAP- Tap Air Portugal Flight has the lowest service quality. In comparison to other airlines, it is obvious that not all airlines with high employee number have a higher service quality.

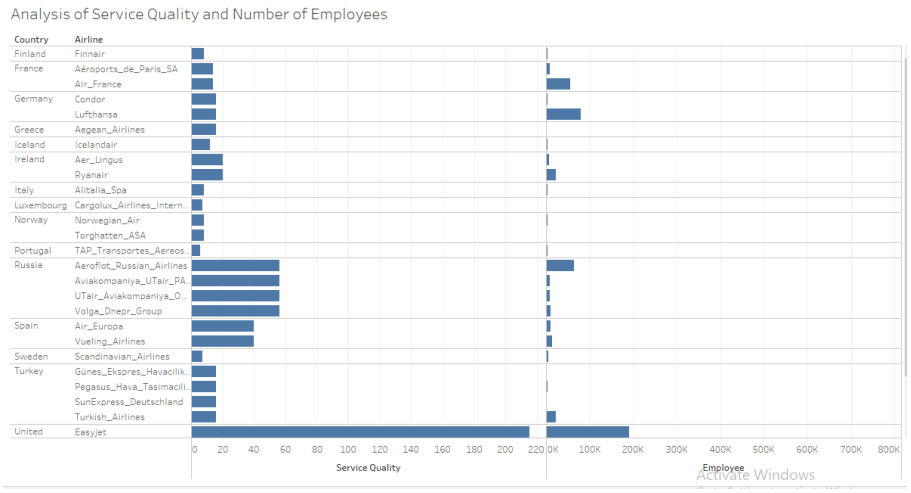

Figure 3: Results for BI Query 2

7.3 BI Query 3:Which country has the best at Handling Passenger rights and what is its total employee and passenger number?

For this query, the contributing sources of data are: Data.world, Statista and Wikipedia and the contributing columns are PassengerCount (FactTable), PassengerHandling (FactTable) Employee (FactTable) Country (DimCountry) and Airline(DimAirline). The general findings are that the United Kingdom tops the charts again as the same company the International Group Airlines -IAG takes the lead as illustrated in Figure **??**. IAG has the highest number of employees and passengers, as well as the best passenger rights.

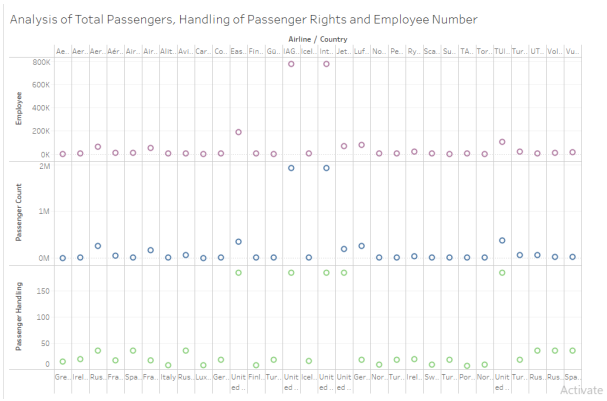

Figure 4: Results for BI Query 3

7.4 Discussion

From the BI queries seen above, it is a noteable fact that some the United Kingdom has the highest service quality overall, but it could be due to the fact that the company listed includes subsidiaries. Looking at some other airlines, we can see that not all the airlines with high number of passengers have high quality service when it comes to passenger handling rights. Some of them have less human resources (employees) but better way of handling passenger rights.

8 Conclusion and Future Work

The aim of this project was to analyze airline companies based on their number of passengers, employees and other service qualities. If an airline invests in Talent Management by giving their employees necessary training based on service quality, they can use more resources better. Finally, we can conclude that less is indeed more. Future work can be done on the analysis of airline punctuality rate using historical data.

...

9 References

Andronie, M. (2015) Airline Applications of Business Intelligence Systems, INCAS Bulletin, 7(3), pp. 153160. Academic Search Complete. Available at: http://eds.b.ebscohost.com/eds/pdfv

d9af-410e-a0d5-25c2e02b69dc

D'Alfonso, Tiziana Malighetti, Paolo Redondi, Renato. (2011) The pricing strategy of Ryanair Airline Industry: Strategies, Operations and Safety. 119-141. Research Gate. Available at https://www.researchgate.net/publication/289029343$_{T}he_{p}ricing_{s}trategy_{o}f_{R}yanair[Access$

James J.H., Liou (2009). A Novel Decision Rules Approach for Customer Relationship Management of the Airline Market Expert Systems with Applications. 36(3) Part 1:4374-4381. ScienceDirect. DOI: 10.1016/j.eswa.2008.05.002 [Assessed 10 April 2019]

Kimball, R. (2008) Drill Down to Ask Why, Part 2, DM Review, 18(8), p. 31. Business Source Complete. Available at: http://search.ebscohost.com/login.aspx?direct=trueAuthType=ip,cook livescope=sitecustid=ncirlib [Accessed: 3 April 2019].

Namukasa, J. (2013) The influence of airline service quality on passenger satisfaction and loyalty : The case of Uganda airline industry The TQM Journal, 25(5), pp.520-532. Emerald Insight. Available at: https://doi.org/10.1108/TQM-11-2012-0092 [Accessed 10 April 2019].

Sturdy and Graham R. (2012) Customer Relationship Management Using Business Intelligence. Business Source Complete. Available at https://www.sciencedirect.com/science/article/pii/B97 [Accessed 2 April 2019].

Williams, G., and OConnell, J. F. (2011) Air Transport in the 21st Century: Key Strategic Developments. EBSCO Host. Available at: http://search.ebscohost.com/login.aspx?direct=trueAut livescope=site [Accessed 11 April 2019].

Wixom, B., Watson, H., Reynolds, AM Hoffer, J. (2008) Continental Airlines Continues to Soar with Business Intelligence, Information Systems Management, 25(2), pp. 102112. doi: 10.1080/10580530801941496. [Accessed 10 April 2019]

Appendix

All Codes Used for this Project

```
#Cleaning Process for EMPLOYEES TABLE
setwd("C:/Users/MOLAP/OneDrive - National College of Ireland/DWBIProject/Airline")

Employees = read.csv("StatistaData.csv")

Employees <- Employees[,-c(3:5, 7:22, 34:50)]
Employees <- Employees[,-c(5:12)]
Employees <- Employees[,-c(5)]

colnames(Employees) <- c("Rank","Airline","Country", "Employees","Revenue")

Employees$Airline = sub(" ","_",Employees$Airline)
Employees$Airline = sub(" ","_",Employees$Airline)
Employees$Airline = sub(" ","_",Employees$Airline)
Employees$Airline = sub(" ","_",Employees$Airline)
Employees$Airline = sub(" ","_",Employees$Airline)
```

```
Employees$Airline = sub("Delta_Air_Lines__","Delta_Air_Lines",Employees$Airline)
Employees$Airline = sub("Jet_Airways__","Jet_Airways",Employees$Airline)
Employees$Airline = sub("WestJet_Airlines_","WestJet_Airlines",Employees$Airline)
Employees$Airline = sub("El_Al_Israel_Airlines_","El_Al_Israel_Airlines",Employees$Ai

Employees$Revenue = sub(",","",Employees$Revenue)
Employees$Employees = sub(",","",Employees$Employees)

Employees$Revenue = as.numeric(Employees$Revenue)
Employees$Employees = as.numeric(Employees$Employees)

write.csv(Employees, "RawData_Employees.csv")

Employees <- read.csv("RawData_Employees.csv")
df <- Employees
df <- data.frame(Employees)
df
library(RODBC)
connection=odbcDriverConnect("driver={SQL Server};server=localhost;database=Airline;T
sqlSave(connection,Employees,tablename="RawData_Employees",rownames="Id",addPK=TRUE)
close(connection)

#Cleaning Process for RANKING TABLE
setwd("C:/Users/MOLAP/OneDrive - National College of Ireland/DWBIProject/Airline")
Ranking = read.csv("Ranking.csv")

colnames(Ranking) <- c("Rank","Airline","Country","Punctuality","Service_Quality","Ha
                       "AirHelp_Score","Year")

#removed duplicate rows
Ranking <- Ranking[-c(72),]
#removed irrelvant column
Ranking <- Ranking[,-c(8)]

Ranking$Airline = sub(" ","_",Ranking$Airline)
Ranking$Airline = sub(" ","_",Ranking$Airline)
Ranking$Airline = sub(" ","_",Ranking$Airline)

Ranking$Airline = sub("WestJet_Airlines__","WestJet_Airlines",Ranking$Airline)
Ranking$Airline = sub("Swiss_International_Air_Lines_","Swiss_International_Air_Lines

write.csv(Ranking,"RawData_Ranking.csv")

setwd("C:/Users/MOLAP/OneDrive - National College of Ireland/DWBIProject/Airline")
Ranking <- read.csv("RawData_Ranking.csv")
df <- Ranking
df <- data.frame(Ranking)
```

```
df
library(RODBC)
connection=odbcDriverConnect("driver={SQL Server};server=localhost;database=Airline;T
sqlSave(connection,Ranking,tablename="RawData_Ranking",rownames="Id",addPK=TRUE)
close(connection)

#Cleaning Process for PASSENGERS TABLE
setwd("C:/Users/MOLAP/OneDrive - National College of Ireland/DWBIProject/Airline")
Passengers = read.csv("WikipediaData.csv")


Passengers <- Passengers[,-c(20:22)]

colnames(Passengers) <- c("Rank","Country","Airline","2019","PassengerCount","2017","
                          ,"2010","2009","2008","2007","2006","Fleet","Destinations")

#Deleted the following columns because they had too many na values
Passengers$`2019`= NULL
Passengers$`2014`= NULL
Passengers$`2013`= NULL
Passengers$`2012`= NULL
Passengers$`2011`= NULL
Passengers$`2010`= NULL
Passengers$`2009`= NULL
Passengers$`2008`= NULL
Passengers$`2007`= NULL
Passengers$`2006`= NULL

#Deleted Fleet and Destination coloumns because they were irrelevant to me
Passengers$Fleet = NULL
Passengers$Destinations = NULL

Passengers$Airline = sub(" ","_",Passengers$Airline)
Passengers$Airline = sub(" ","_",Passengers$Airline)

Passengers$Airline = sub("SmartWings_Group_","SmartWings_Group",Passengers$Airline)
Passengers$Airline = sub("Icelandair_","Icelandair",Passengers$Airline)
Passengers$Airline = sub("Norwegian_Air_","Norwegian_Air",Passengers$Airline)

# I deleted the rows with more than two na values.
Passengers <- Passengers[-c(43),]
Passengers <- Passengers[-c(39),]

sum(is.na(Passengers))

#replace na values with the median
AE <- c(Passengers[17,4],Passengers[17,5],Passengers[17,6],Passengers[17,7])
MedAE <- median(AE, na.rm = T)
```

```
Passengers[17,4][is.na(Passengers[17,4])] <- MedAE

Jet2 <- c(Passengers[18,4],Passengers[18,5],Passengers[18,6],Passengers[18,7])
MedJet2 <- median(Jet2, na.rm = T)
Passengers[18,4][is.na(Passengers[18,4])] <- MedJet2

flybe <- c(Passengers[19,4],Passengers[19,5],Passengers[19,6],Passengers[19,7])
MedFlybe <- median(flybe, na.rm = T)
Passengers[19,(4:7)][is.na(Passengers[19,(4:7)])] <- MedFlybe

LOT <- c(Passengers[21,4],Passengers[21,5],Passengers[21,6],Passengers[21,7])
MedLOT <- median(LOT, na.rm = T)
Passengers[21,4][is.na(Passengers[21,4])] <- MedLOT

SWGroup <- c(Passengers[22,4],Passengers[22,5],Passengers[22,6],Passengers[22,7])
MedSW <- median(SWGroup, na.rm = T)
Passengers[22,7][is.na(Passengers[22,7])] <- MedSW

Virgin <- c(Passengers[25,4],Passengers[25,5],Passengers[25,6],Passengers[25,7])
MedV <- median(Virgin, na.rm = T)
Passengers[25,4][is.na(Passengers[25,4])] <- MedV

blueair <- c(Passengers[26,4],Passengers[26,5],Passengers[26,6],Passengers[26,7])
MedBlu <- median(blueair, na.rm = T)
Passengers[26,4][is.na(Passengers[26,4])] <- MedBlu

Volotea <- c(Passengers[27,4],Passengers[27,5],Passengers[27,6],Passengers[27,7])
MedVol <- median(Volotea, na.rm = T)
Passengers[27,7][is.na(Passengers[27,7])] <- MedVol

Belevia <- c(Passengers[32,4],Passengers[32,5],Passengers[32,6],Passengers[32,7])
MedBel <- median(Belevia, na.rm = T)
Passengers[32,(4:7)][is.na(Passengers[32,(4:7)])] <- MedBel

AS <- c(Passengers[34,4],Passengers[34,5],Passengers[34,6],Passengers[34,7])
MedAS <- median(AS, na.rm = T)
Passengers[34,4][is.na(Passengers[34,4])] <- MedAS

Tarom <- c(Passengers[36,4],Passengers[36,5],Passengers[36,6],Passengers[36,7])
MedTarom <- median(Tarom, na.rm = T)
Passengers[36,4][is.na(Passengers[36,4])] <- MedTarom

AA <- c(Passengers[37,4],Passengers[37,5],Passengers[37,6],Passengers[37,7])
medianAA <- median(AA, na.rm = T)
Passengers[37,(6:7)][is.na(Passengers[37,(6:7)])] <- medianAA

Luxair <- c(Passengers[40,4],Passengers[40,5],Passengers[40,6],Passengers[40,7])
MedLux <- median(Luxair, na.rm = T)
```

```
Passengers[40,4][is.na(Passengers[40,4])] <- MedLux

Adria <- c(Passengers[45,4],Passengers[45,5],Passengers[45,6],Passengers[45,7])
MedAdria <- median(Adria, na.rm = T)
Passengers[45,4][is.na(Passengers[45,4])] <- MedAdria

Passengers = subset(Passengers,select = -c(5,6,7))

write.csv(Passengers, file = "RawData_Passengers.csv", row.names = F)

setwd("C:/Users/MOLAP/OneDrive - National College of Ireland/DWBIProject/Airline")
Passengers <- read.csv("RawData_Passengers.csv")
df <- Passengers
df <- data.frame(Passengers)
df
library(RODBC)
connection=odbcDriverConnect("driver={SQL Server};server=localhost;database=Airline;T
sqlSave(connection,Passengers,tablename="RawData_Passengers",rownames="Id",addPK=TRUE
close(connection)

#Query for Dimensions

Create Table DimCountry(
CountryID int identity(1,1) primary key,
Country varchar(50))

Drop table DimAirline
Create Table DimAirline(
AirlineID int identity(1,1) primary key,
Airline varchar(50),
ServiceRanking int,
EmployeeRanking int,
PunctualityRanking int)



Insert into DimCountry(Country)
Select Distinct Country from DB_Staging_Ranking


Insert into DimAirline (Airline,
ServiceRanking,
EmployeeRanking,
PunctualityRanking)
Select Distinct [e].[Airline],
   [e].[Rank],
  [p].[Punctuality],
  [pa].[Rank]
```

```
from DB_Staging_Employee as e
Inner join [dbo].[DB_Staging_Ranking] as p ON [p].[airline] = [e].[Airline]
Inner Join [dbo].[DB_Staging_Passengers] as pa ON [pa].[Airline] =[e].[Airline]

truncate table DimCountry

truncate table DimAirline


#Query for Fact Table
Create table [Fact Table] (

CountryID int ,
AirlineID int,
PassengerCount numeric,
Employee numeric,
Revenue numeric,
ServiceQuality numeric,
PassengerHandling numeric
)
insert into [Fact Table](
CountryID ,
AirlineID ,
PassengerCount,
Employee ,
Revenue ,
ServiceQuality ,
PassengerHandling
)

Select  Distinct [dc].[CountryID],
[da].[AirlineID],
  [e].[Employees],
  [e].[Revenue],
      [pa].[PassengerCount],
  [p].[Service_Quality],
  [p].[Handling_of_passenger_rights]
from DB_Staging_Employee as e
Inner Join [dbo].[DB_Staging_Passengers] as pa ON [pa].[Country] =[e].[Country]
Inner join [dbo].[DB_Staging_Ranking] as p ON [p].[Country] = [e].[Country]
Inner Join [dbo].[DimAirline] as da ON [da].[Airline] = [e].[Airline]
Inner Join [dbo].[DimCountry] as dc ON [dc].[Country] = [e].[Country];

Select * from [dbo].[DB_Staging_Ranking]
Select * from [dbo].[DB_Staging_Employee]
Select * from [dbo].[DB_Staging_Passengers]
```

```
Alter Table [dbo].[Fact Table]
Drop Constraint FK_Country
Go

Alter Table [dbo].[Fact Table]
Drop Constraint FK_Airline
Go

Truncate Table [dbo].[Fact Table]
Alter Table [dbo].[Fact Table]
Add Constraint FK_Country
FOREIGN KEY([CountryID]) References
DimCountry(CountryID)

go
Alter Table [dbo].[Fact Table]
Add Constraint FK_Airline
FOREIGN KEY([AirlineID]) References
DimAirline(AirlineID)
```