STAT 440 Final Project

Data Analysis for OpenFlights

University of Illinois

Zhuoyun Wang

Qianxiao Zhang

INTRODUCTION

This report will introduce our data analysis based on the data provided by the OpenFlights website. We will focus on processing and analyzing three datasets: “airports”, “airlines”, and “routes” from the website. The analysis will base on the relationship between the three datasets, from which we can get information about the operations of the airlines and airports. Beginning with introducing our original data, the report will proceed by discussing the issues within the data, our data cleaning and validating methods, and the methods and strategies for doing our tasks; Then we will show the results of our analysis by presenting the output tables and discuss any issues that are noticeable.

METHODS

Data Description

The three datasets from OpenFlights were downloaded as .txt files originally. As of January 2017, the dataset “Airports” contains over 10,000 airports, train stations, and ferry terminals, each of which with its unique Airport ID. The information provided by this dataset includes the airports’ IDs, names, main cities served, countries/territories, etc.; The dataset “Airlines” contains 5,888 airlines as of January 2012. The variables of this dataset include a unique ID for each airline, the airline’s name, the country where the airline is incorporated, and whether the airline was still active at that time, etc.; The dataset “routes” contains 6,7663 routes between 3,321 airports on 548 airlines globally as of June 2014. The information it provides includes the airlines’ codes and IDs, the source airports and their IDs, the destination airports and their IDs, the stops of the flight, etc.

Data Processing, Cleaning, and Validating

Since all three downloaded datasets are delimited standard data, we used the list input without specifying informats. For the dataset “airlines”, we used DSD and MISSOVER options because the original dataset has missing values. Input variables are both character variables, including “name”, “alias”, “two-letter code”, “three-letter code”, “callsign”, “country”, and “active”, and numeric variables, which is “ID”. The dataset “airports” was read by using the FLOWOVER option and the RETAIN statement. Input variables included both character and numeric information of airports. We read the dataset “routes” by specifying DSD and MISSOVER options. The variable “ID” was read as characters since missing values were assigned “\N”, therefore numeric reading did not work. This variable was cleaned later and explained in the data validation part. Other important variables included “destID”, “sourceID”, and “stops”.

Proceeding to data validating and cleaning, we decided to focus only on the airlines that were still active at that time, and the ID should be numeric and not missing. This would require us to remove the observations with value “N” for the variable “Active” or missing values for the variable “ID”. We should also convert character “ID” values to numeric ones by using the “input” function in SAS if applicable; As for the dataset “airports”, we believe that we should also convert the character “Airport_ID” values to numeric ones if necessary; Besides, for our analysis, the variables “ID”, “sourceID”, and “destID” of the dataset “routes” should be non-missing numeric values, which required us to remove observations with missing values and convert character values to numeric ones if necessary. The goal of removing undesirable observations would be achieved by using the WHERE statement under the data step in SAS, and converting data types would be achieved by using the INPUT function in SAS.

Task Description and Strategy Explanation

We addressed several different tasks in this project. Our methods and strategies are listed below:

1. We found out which active airlines were used frequently based on the number of route records they had. We would solve this by selecting the IDs with top frequencies from the dataset "routes", using the GROUP BY and ORDER BY statements in SQL, and then use SQL's inner joins to match the IDs selected with the IDs of the dataset "airlines" so that we could find the names and countries of these airline IDs.
2. We wondered the most popular source airports based on route records. We tackled the problem by selecting the source IDs with top frequencies from the dataset "routes", using the GROUP BY and ORDER BY statements. Then we use inner joins to find the names, cities, and countries of the owners of these IDs.
3. After noticing that most flights were non-stop, we wanted to know the percentages of non-stop and one-stop flights, given that these were the only two conditions. We used the COUNT function and GROUP BY statements to get the numbers of non-stop and one-stop flights and then calculated their percentages.
4. Based on the outputs of the prior tasks, it was surprising to see that only a tiny part of the flights had one stop, so we wondered the active airlines that operated these flights. We used a subquery in SQL to select the airline IDs, which had one stop based on the dataset "routes", names, as well as countries of the airlines from the dataset "airlines".
5. Chicago has two international airports: O'Hare and Midway. We wondered what the most popular destinations and the most frequently used airlines are regarding these two airports. We used SQL subqueries to match "source_ID" in "routes_cleaned" datasets with the "Aiport_ID" in the "Airport" dataset to pick out Midway and O'Hare Airport. And we used the count() function to get the total numbers of airlines and used the ORDER BY statement to sort the counts.

We will present and discuss our results in the following section.

RESULTS

The first two tables were generated by the PROC CONTENTS procedure in SAS. We can see that there is a total of 6,161 observations in the SAS dataset "airlines", and the variable "ID" is numeric as we expected. Also, we use the PROC MEANS procedure to show the total values and missing values for the variable "ID". The result is shown in the third table, and we know that there are no missing levels for this variable. Since we were only interested in the airlines that were still active at that time, we trimmed the observations down by creating a new SAS dataset called "airlines_cleaned" and removing the observations of which the statuses were non-active. The result is presented at the fourth table below, from which we know that there were 1,254 active airlines left.

Data Set Name	WORK.AIRLINES	**Observations**	6161

Member Type	DATA	**Variables**	8
Engine	V9	**Indexes**	0
Created	05/05/2019 12:49:50	**Observation Length**	136
Last Modified	05/05/2019 12:49:50	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Alphabetic List of Variables and Attributes			
#	**Variable**	**Type**	**Len**
8	Active	Char	8
7	Country	Char	20
1	ID	Num	8
2	Name	Char	50
3	alias	Char	8
6	callsign	Char	20
4	code2	Char	8
5	code3	Char	8

Analysis Variable : ID	
N	**N Miss**
6161	0

Data Set Name	WORK.AIRLINES_CLEANED	**Observations**	1254
Member Type	DATA	**Variables**	8
Engine	V9	**Indexes**	0
Created	05/05/2019 12:49:50	**Observation Length**	136
Last Modified	05/05/2019 12:49:50	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Moving on to the SAS dataset “airports”, the first two tables below show the outputs of the PROC CONTENTS procedure in SAS, and we know from them that there are 7,543 observations and 14 variables in total in this dataset. The variable “Airport_ID” is numeric as we expected. The third table, which is generated by the PROC MEANS procedure, informs us that there are no missing values for the variable “Airport_ID”. As the properties of this dataset seem to be aligned with our data validation guidelines, we will leave the dataset as it is for future analysis.

Data Set Name	WORK.AIRPORTS	**Observations**	7543
Member Type	DATA	**Variables**	14
Engine	V9	**Indexes**	0
Created	05/05/2019 12:49:50	**Observation Length**	232
Last Modified	05/05/2019 12:49:50	**Deleted Observations**	0
Protection		**Compressed**	NO

Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Alphabetic List of Variables and Attributes			
#	**Variable**	**Type**	**Len**
1	Airport_ID	Num	8
3	City	Char	30
4	Country	Char	30
2	Name	Char	50
9	al	Num	8
5	code3	Char	8
6	code4	Char	8
11	dst	Char	2
7	lat	Num	8
8	long	Num	8
14	source	Char	15
10	timezone	Char	6
13	type	Char	15
12	tz	Char	30

Analysis Variable: Airport_ID	
N	**N Miss**
7543	0

Then we examined the properties of the SAS dataset “routes”. By performing the PROC CONTENTS procedure, we know that there is a total of 67,663 observations and 9 variables in this dataset. However, as shown in the second table, all the variables “ID”, “sourceID”, and “destID” are characters, so we have to convert them to numeric values. The third table shows the number of observations which have missing values for “ID”, “sourceID”, or “destID”, and we will remove these observations for our analysis, using the WHERE statement under the data step in SAS. After validating and cleaning the dataset, the SAS dataset “routes_cleaned” has 9 variables with 66,765 observations, and all the variables mentioned above become numeric, as known in the fourth and fifth tables generated by the PROC CONTENTS procedure. The last table was the output of the PROC MEANS procedure, showing that there are no missing values in the dataset “routes_cleaned” for the three variables mentioned above.

Data Set Name	WORK.ROUTES	**Observations**	67663
Member Type	DATA	**Variables**	9
Engine	V9	**Indexes**	0
Created	05/05/2019 12:49:51	**Observation Length**	56
Last Modified	05/05/2019 12:49:51	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Alphabetic List of Variables and Attributes

#	Variable	Type	Len
2	ID	Char	8
1	code	Char	3
5	dest	Char	5
6	destID	Char	8
9	equipment	Char	8
7	share	Char	1
3	source	Char	5
4	sourceID	Char	8
8	stops	Num	8

Data Set Name	WORK.A	**Observations**	898
Member Type	DATA	**Variables**	9
Engine	V9	**Indexes**	0
Created	05/05/2019 12:49:51	**Observation Length**	56
Last Modified	05/05/2019 12:49:51	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Data Set Name	WORK.ROUTES_CLEANED	**Observations**	66765
Member Type	DATA	**Variables**	9
Engine	V9	**Indexes**	0
Created	05/05/2019 12:49:51	**Observation Length**	56
Last Modified	05/05/2019 12:49:51	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Alphabetic List of Variables and Attributes			
#	**Variable**	**Type**	**Len**
7	ID_airline	Num	8
1	code	Char	3
3	dest	Char	5
9	dest_ID	Num	8
6	equipment	Char	8
4	share	Char	1
2	source	Char	5
8	source_ID	Num	8

5	stops	Num	8

The table below shows the top 3 airlines with the most records in the dataset “routes”. The variables “ID”, “Name”, and “Country” provide relative information about these airlines, and the variable “ID_frequency” represents the times that they show up in the “routes” dataset. These airlines are listed in a descending order based on their frequencies.

0.38	Name	Country	ID_frequency
4296	Ryanair	Ireland	2484
24	American Airlines	United States	2352
5209	United Airlines	United States	2180

The table below shows the top 3 most popular source airports based on the records in the “routes” dataset. The variables “source_ID”, “Name”, “City”, “Country”, “source”, and “Source_Freq” represents the airports’ IDs, the cities they are located in, the countries they are located in, their abbreviated names, and the frequencies that they show up as source airports in the “routes” dataset, respectively. They are listed in a descending order based on their frequencies.

source_ID	Name	City	Country	source	Source_Freq
3682	Hartsfield Jackson Atlanta International Airport	Atlanta	United States	ATL	915
3830	Chicago O'Hare International Airport	Chicago	United States	ORD	558
3364	Beijing Capital International Airport	Beijing	China	PEK	535

The table below shows the percentages of fights without stops and with one stop to the total number of flights recorded in the “routes” dataset. It’s somewhat counterintuitive to see that only 0.0165% of the flights recorded had one stop, and there were no flights with more than one stop recorded.

Percentage	stops
99.9835%	0

0.0165%	1

After noticing the shocking percentage of the flights with one stop, we were curious about the airlines that operated these flights. The table below lists all the active airlines which had flights with one stop, including the airlines' IDs, names, and the countries that they belonged to.

ID	Name	Country
330	Air Canada	Canada
1316	AirTran Airways	United States
1623	Canadian North	Canada
1936	Cubana de Aviación	Cuba
4319	Scandinavian Airlines System	Sweden
4547	Southwest Airlines	United States

As the two tables show, regarding Chicago O'Hare International Airport, the top three airlines are United Airlines (161), United Airlines (124), and US Airways (118). The top three destinations are ATL (Atlanta, 20), MSY (Paris, 13), and CDG (New Orleans, 10). While the top three airlines at Midway Airport are Southwest Airlines (62), AirTran Airways (61), and Volaris (4), the top three destinations are ATL (Atlanta, 5), MSP (Minneapolis, 4), and DTW (Detroit, 3).

Name	Destination	Count	air_name	Airline Count
Chicago O'Hare International Airport	MSY	13	American Airlines	124
Chicago O'Hare International Airport	ATL	20	United Airlines	161
Chicago O'Hare International Airport	CDG	10	US Airways	118

Name	Destination	Count	Airline Name	Airline Count
Chicago Midway International Airport	MSP	4	AirTran Airways	61

Chicago Midway International Airport	ATL	5	Southwest Airlines	62
Chicago Midway International Airport	DTW	3	Volaris	4

CONCLUSION

As we can see from the tables above, the United States seems to be a hub for the airline industry. The United States had two airlines ranking among the top three airlines with the highest frequencies and two airports ranking among the top three most popular source airports globally. This information could be used to analyze passengers' preferences for airlines and flights. The information we got from this project can be used by data analysts for future analysis about airlines' profitability, airports' utilization, etc. The marketing departments of the airlines should also take relative statistics information into account when they formulate marking strategies.