DATA MANAGEMENT AND DATABASE DESIGN

(ASSIGNMENT 1)

BY:

TeamASquare

ANINDITA BAISHYA (NUID: 001387422)

ABHI PATODI (NUID: 001404833)

Abstract:

In this assignment, we planned to create a conceptual relation between three different data sources via Web Scraping, Web API and CSV Dataset. After collecting the data, we made it accurate and complete by cleaning the irrelevant parts of the data. Post which the data is split into three tables with the relevant attributes.

Data Theme:

The assignment’s theme is based on the “Best Airlines” worldwide.

Data Sources:

We created a database of multiple attributes which has its information captured from different data sources using Web scraping, Web API and a CSV Dataset.

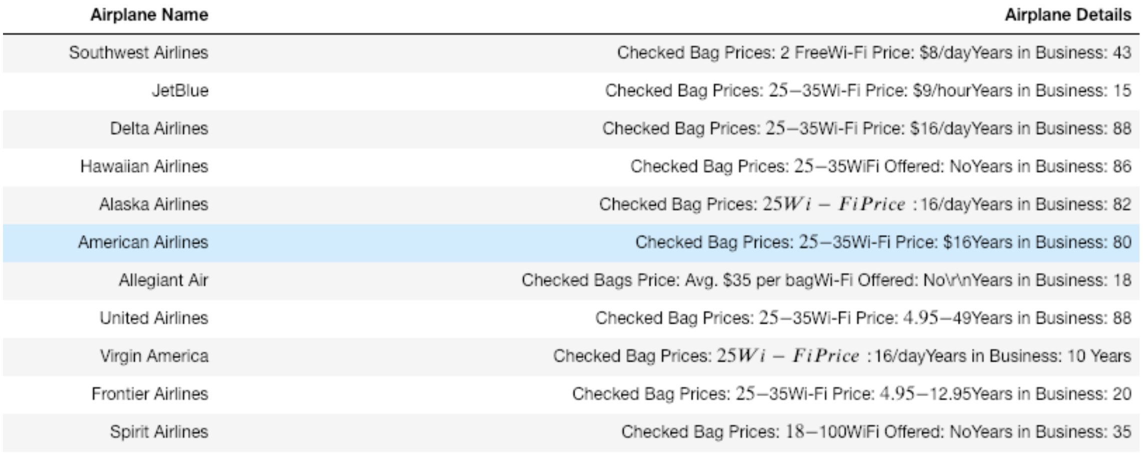
The various attributes collected from the three sources are listed below:

#Using Web Scraper:

We scrapped the attributes from the website:

URL: https://bestcompany.com/airlines and created a dataframe from the website using BeautifulSoup

1. Airplane Name – gives the name of the airplanes
2. Airplane Details – gives the baggage information, Wi-fi details, Years in the business

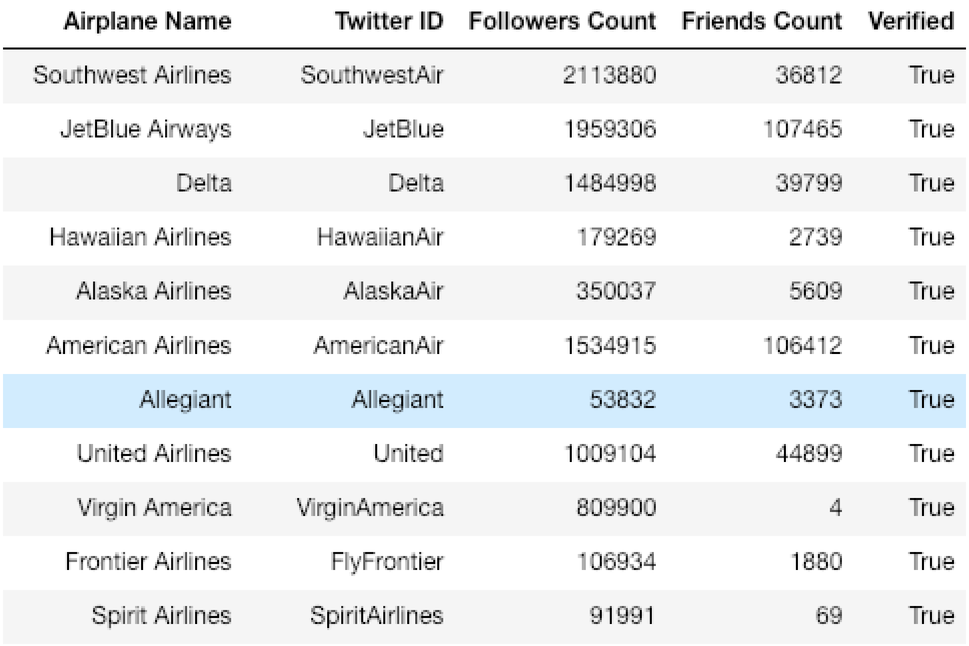


Airplane Name	Airplane Details
Southwest Airlines	Checked Bag Prices: 2 FreeWi-Fi Price: $8/dayYears in Business: 43
JetBlue	Checked Bag Prices: 25–35Wi-Fi Price: $9/hourYears in Business: 15
Delta Airlines	Checked Bag Prices: 25–35Wi-Fi Price: $16/dayYears in Business: 88
Hawaiian Airlines	Checked Bag Prices: 25–35WiFi Offered: NoYears in Business: 86
Alaska Airlines	Checked Bag Prices: $25Wi - FiPrice$:16/dayYears in Business: 82
American Airlines	Checked Bag Prices: 25–35Wi-Fi Price: $16Years in Business: 80
Allegiant Air	Checked Bags Price: Avg. $35 per bagWi-Fi Offered: No\r\nYears in Business: 18
United Airlines	Checked Bag Prices: 25–35Wi-Fi Price: 4.95–49Years in Business: 88
Virgin America	Checked Bag Prices: $25Wi - FiPrice$:16/dayYears in Business: 10 Years
Frontier Airlines	Checked Bag Prices: 25–35Wi-Fi Price: 4.95–12.95Years in Business: 20
Spirit Airlines	Checked Bag Prices: 18–100WiFi Offered: NoYears in Business: 35

#Using Web API:

We collected the below listed attributes and created a data frame using the Twitter APIs :

1. Airplane Name – gives the name of the airplanes
2. Twitter ID – gives the user name of their social account in twitter
3. Followers Count – gives the list of follower’s count
4. Friends Count – gives the list of account’s count they follow
5. Verified – states if their social accounts are verified or not

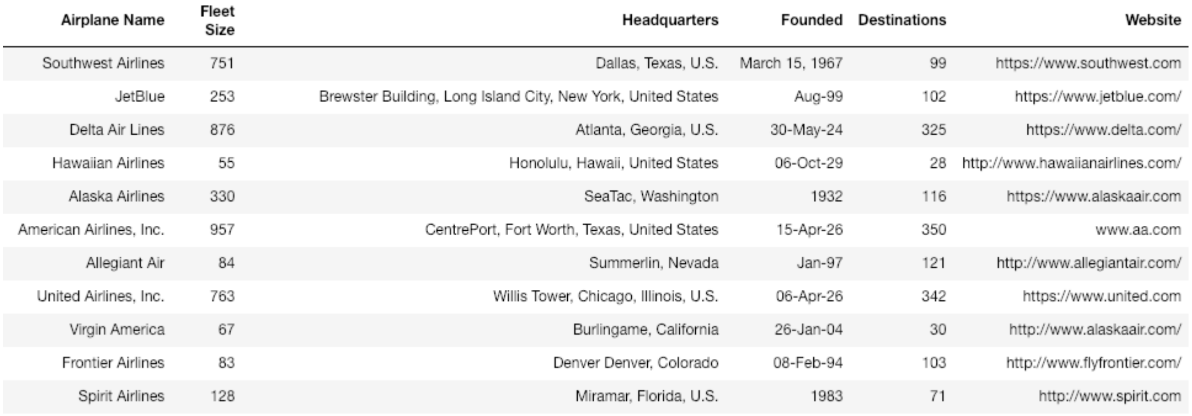


Airplane Name	Twitter ID	Followers Count	Friends Count	Verified
Southwest Airlines	SouthwestAir	2113880	36812	True
JetBlue Airways	JetBlue	1959306	107465	True
Delta	Delta	1484998	39799	True
Hawaiian Airlines	HawaiianAir	179269	2739	True
Alaska Airlines	AlaskaAir	350037	5609	True
American Airlines	AmericanAir	1534915	106412	True
Allegiant	Allegiant	53832	3373	True
United Airlines	United	1009104	44899	True
Virgin America	VirginAmerica	809900	4	True
Frontier Airlines	FlyFrontier	106934	1880	True
Spirit Airlines	SpiritAirlines	91991	69	True

#CSV Dataset:

We created the CSV dataset by scraping Wikipedia pages of relevant best airlines.

1. Airplane Name – gives the names of the airplanes
2. Fleet Size – refers to the number of planes or aircraft of similar model operated by an airline
3. Headquarters – their main office
4. Founded – the day they started off
5. Destinations – to how many destinations it flies to
6. Website – their websites to refer



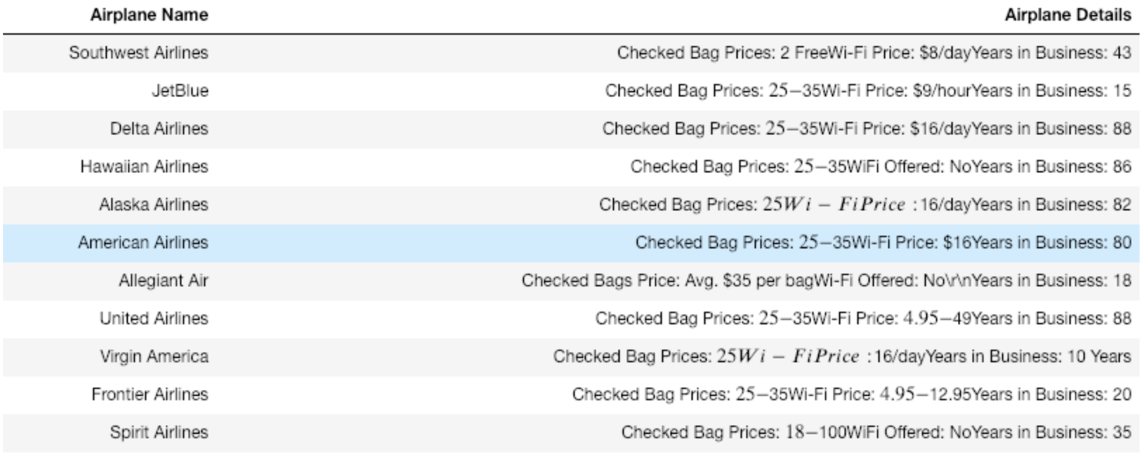
Airplane Name	Fleet Size	Headquarters	Founded	Destinations	Website
Southwest Airlines	751	Dallas, Texas, U.S.	March 15, 1967	99	https://www.southwest.com
JetBlue	253	Brewster Building, Long Island City, New York, United States	Aug-99	102	https://www.jetblue.com/
Delta Air Lines	876	Atlanta, Georgia, U.S.	30-May-24	325	https://www.delta.com/
Hawaiian Airlines	55	Honolulu, Hawaii, United States	06-Oct-29	28	http://www.hawaiianairlines.com/
Alaska Airlines	330	SeaTac, Washington	1932	116	https://www.alaskaair.com
American Airlines, Inc.	957	CentrePort, Fort Worth, Texas, United States	15-Apr-26	350	www.aa.com
Allegiant Air	84	Summerlin, Nevada	Jan-97	121	http://www.allegiantair.com/
United Airlines, Inc.	763	Willis Tower, Chicago, Illinois, U.S.	06-Apr-26	342	https://www.united.com
Virgin America	67	Burlingame, California	26-Jan-04	30	http://www.alaskaair.com/
Frontier Airlines	83	Denver Denver, Colorado	08-Feb-94	103	http://www.flyfrontier.com/
Spirit Airlines	128	Miramar, Florida, U.S.	1983	71	http://www.spirit.com

Data Reformat and Conceptual Schema:

Collected the data from three sources and separated the datasets into the following three tables.

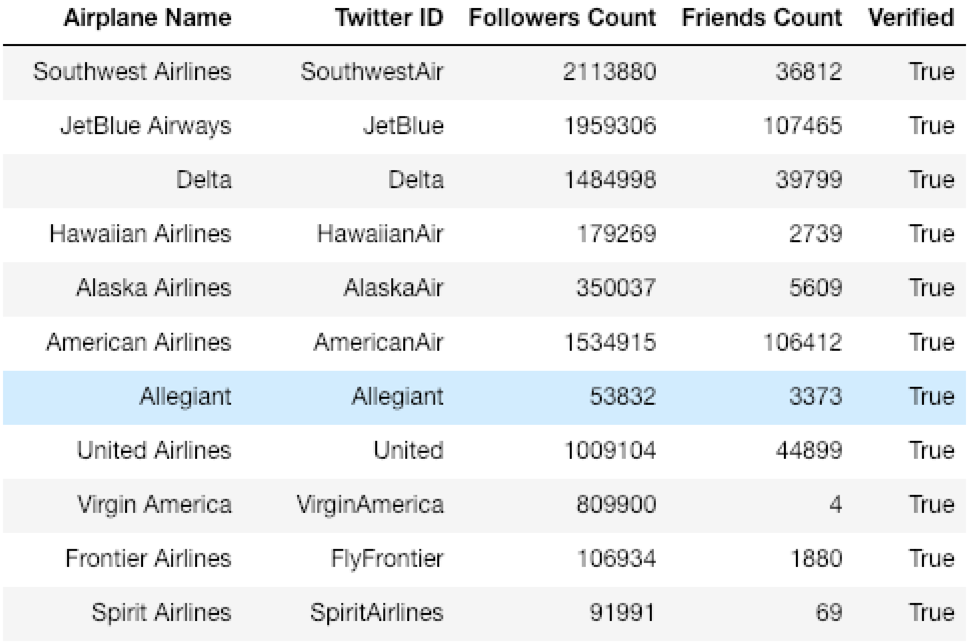
1. Baggage_Details
   a. Airplane Name

b. Airplane Details



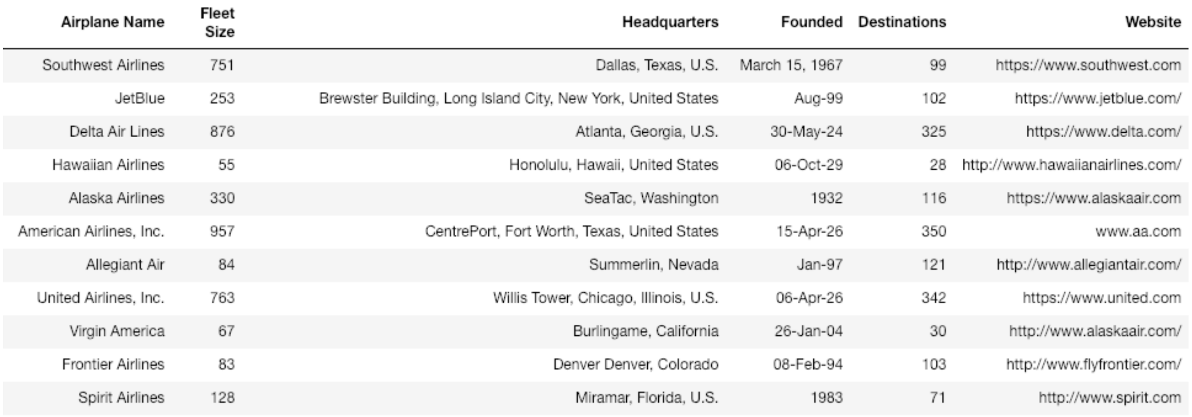
Airplane Name	Airplane Details
Southwest Airlines	Checked Bag Prices: 2 FreeWi-Fi Price: $8/dayYears in Business: 43
JetBlue	Checked Bag Prices: $25-35$Wi-Fi Price: $9/hourYears in Business: 15
Delta Airlines	Checked Bag Prices: $25-35$Wi-Fi Price: $16/dayYears in Business: 88
Hawaiian Airlines	Checked Bag Prices: $25-35$WiFi Offered: NoYears in Business: 86
Alaska Airlines	Checked Bag Prices: $25Wi-FiPrice:$16/dayYears in Business: 82
American Airlines	Checked Bag Prices: $25-35$Wi-Fi Price: $16Years in Business: 80
Allegiant Air	Checked Bags Price: Avg. $35 per bagWi-Fi Offered: No\r\nYears in Business: 18
United Airlines	Checked Bag Prices: $25-35$Wi-Fi Price: $4.95-49$Years in Business: 88
Virgin America	Checked Bag Prices: $25Wi-FiPrice:$16/dayYears in Business: 10 Years
Frontier Airlines	Checked Bag Prices: $25-35$Wi-Fi Price: $4.95-12.95$Years in Business: 20
Spirit Airlines	Checked Bag Prices: $18-100$WiFi Offered: NoYears in Business: 35

2. Twitter_Specifications
   a. Airplane Name
   b. Twitter ID
   c. Followers Count
   d. Friends Count
   e. Verified



Airplane Name	Twitter ID	Followers Count	Friends Count	Verified
Southwest Airlines	SouthwestAir	2113880	36812	True
JetBlue Airways	JetBlue	1959306	107465	True
Delta	Delta	1484998	39799	True
Hawaiian Airlines	HawaiianAir	179269	2739	True
Alaska Airlines	AlaskaAir	350037	5609	True
American Airlines	AmericanAir	1534915	106412	True
Allegiant	Allegiant	53832	3373	True
United Airlines	United	1009104	44899	True
Virgin America	VirginAmerica	809900	4	True
Frontier Airlines	FlyFrontier	106934	1880	True
Spirit Airlines	SpiritAirlines	91991	69	True

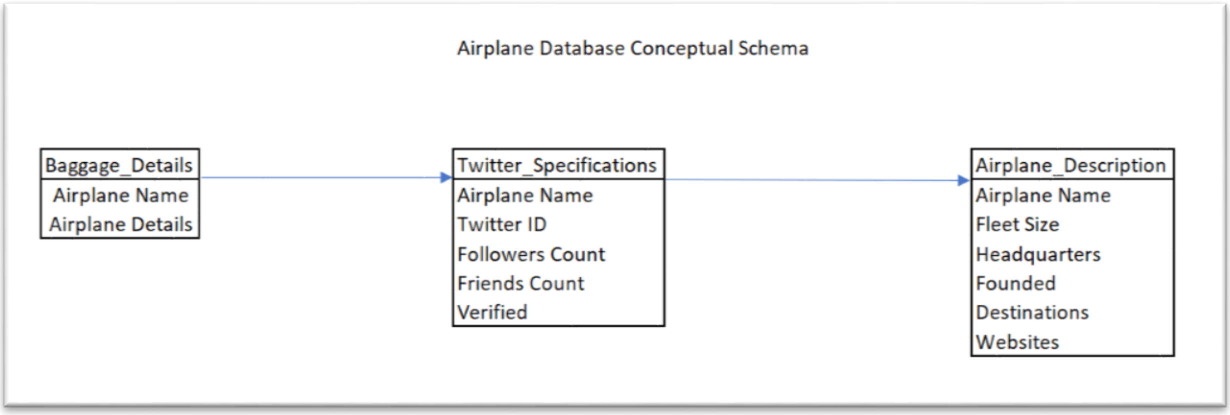
3. Airplane_Description
   a. Airplane Name
   b. Fleet Size
   c. Headquarters
   d. Founded
   e. Destinations
   f. Website



Airplane Name	Fleet Size	Headquarters	Founded	Destinations	Website
Southwest Airlines	751	Dallas, Texas, U.S.	March 15, 1967	99	https://www.southwest.com
JetBlue	253	Brewster Building, Long Island City, New York, United States	Aug-99	102	https://www.jetblue.com/
Delta Air Lines	876	Atlanta, Georgia, U.S.	30-May-24	325	https://www.delta.com/
Hawaiian Airlines	55	Honolulu, Hawaii, United States	06-Oct-29	28	http://www.hawaiianairlines.com/
Alaska Airlines	330	SeaTac, Washington	1932	116	https://www.alaskaair.com
American Airlines, Inc.	957	CentrePort, Fort Worth, Texas, United States	15-Apr-26	350	www.aa.com
Allegiant Air	84	Summerlin, Nevada	Jan-97	121	http://www.allegiantair.com/
United Airlines, Inc.	763	Willis Tower, Chicago, Illinois, U.S.	06-Apr-26	342	https://www.united.com
Virgin America	67	Burlingame, California	26-Jan-04	30	http://www.alaskaair.com/
Frontier Airlines	83	Denver Denver, Colorado	08-Feb-94	103	http://www.flyfrontier.com/
Spirit Airlines	128	Miramar, Florida, U.S.	1983	71	http://www.spirit.com

Conceptual Schema Diagram:

We created an object model diagram showing the relation between three tables having **"Airplane Name"** as the primary key.



Airplane Database Conceptual Schema

Baggage_Details
- Airplane Name
- Airplane Details

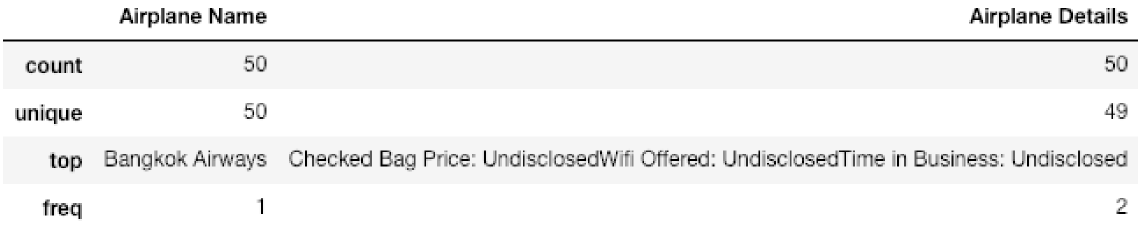
Twitter_Specifications
- Airplane Name
- Twitter ID
- Followers Count
- Friends Count
- Verified

Airplane_Description
- Airplane Name
- Fleet Size
- Headquarters
- Founded
- Destinations
- Websites

Audit Accuracy:

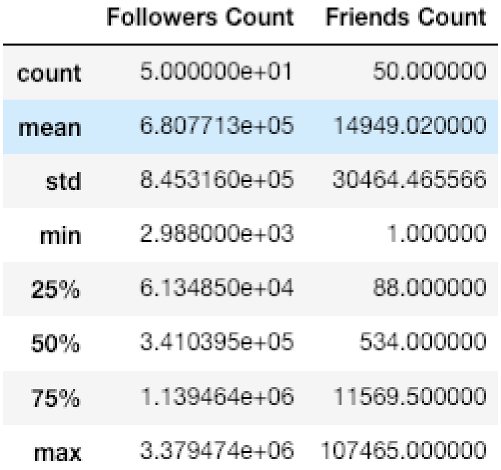
We used Python to check if there are any duplicate values and delete the rows which has values.

Baggage_Details Table:



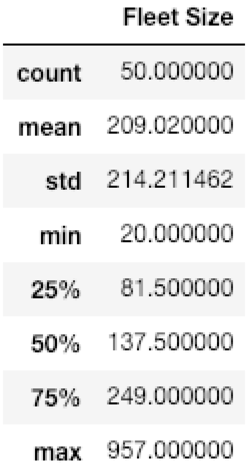
	Airplane Name	Airplane Details
count	50	50
unique	50	49
top	Bangkok Airways	Checked Bag Price: UndisclosedWifi Offered: UndisclosedTime in Business: Undisclosed
freq	1	2

Twitter_Specifications Table:



	Followers Count	Friends Count
count	5.000000e+01	50.000000
mean	6.807713e+05	14949.020000
std	8.453160e+05	30464.465566
min	2.988000e+03	1.000000
25%	6.134850e+04	88.000000
50%	3.410395e+05	534.000000
75%	1.139464e+06	11569.500000
max	3.379474e+06	107465.000000

Airplane_Description Table:



	Fleet Size
count	50.000000
mean	209.020000
std	214.211462
min	20.000000
25%	81.500000
50%	137.500000
75%	249.000000
max	957.000000

Audit Completeness:

We collected the data of the best rated airlines from the real world website: https://bestcompany.com/airlines that gives us the list of the best airlines with respect to the below key points :

1. User Review Index Score (60% of overall score)
   - Star Rating of Reviews (30%)
   - No of Reviews (30%)

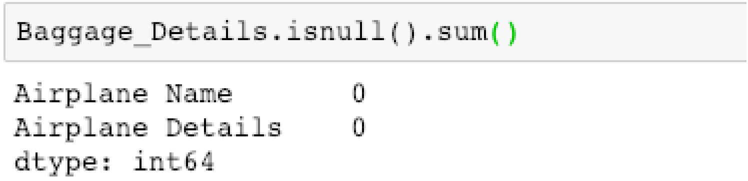
2. Market Index Score (45% of overall score)
   - Recurring Fees (14%)
   - One-Time Fees (12%)
   - Contract of Warranty Length (10%)
   - Brand Search Volume (2%)
   - Time in Business (2%)

And from that list we picked the top 50 airlines which defines the completeness of the data.

Audit Consistency/Uniformity:

We checked and verified that the data does not contain any missing values for any column. All the data throughout is consistent and has no null values.

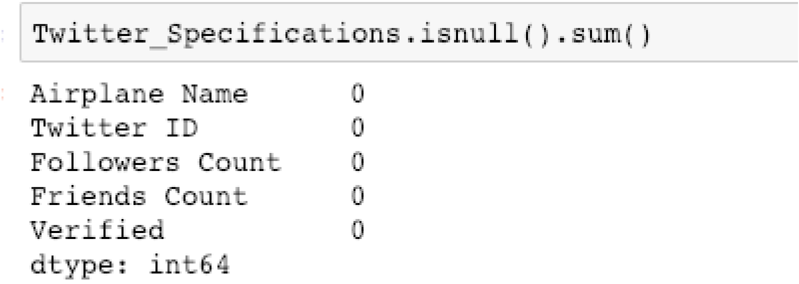
Baggage_Details Table



```
Baggage_Details.isnull().sum()
```

```
Airplane Name       0
Airplane Details    0
dtype: int64
```

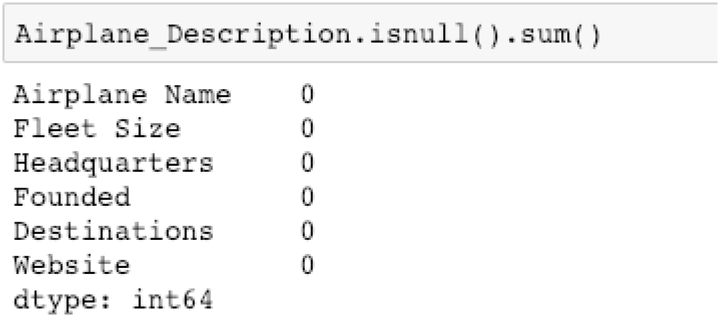
Twitter_Specifications Table:



```
Twitter_Specifications.isnull().sum()
```

```
Airplane Name      0
Twitter ID         0
Followers Count    0
Friends Count      0
Verified           0
dtype: int64
```

Airplane_Description Table:



```
Airplane_Description.isnull().sum()
```

```
Airplane Name    0
Fleet Size       0
Headquarters     0
Founded          0
Destinations     0
Website          0
dtype: int64
```

Citations and References:

Each code in this assignment is self-developed and is not copied from any website. Please refer the Jupytor notebook TeamASquare_Assignment1.ipynb attached along with this document for the code.

References were taken from the below websites:

https://www.pythonforbeginners.com
https://www.py4e.com/book.php
https://www.w3schools.com/python/

Text License:

This is a human-readable summary of (and not a substitute for) the license. Disclaimer.

You are free to:

Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.