# ETL Project - Final Report

Contents

[**1.** **Data source** 1](#_Toc95231340)

[**2.** **Data extraction** 1](#_Toc95231341)

[**3.** **Data transformation** 1](#_Toc95231342)

[**4.** **Data loading** 7](#_Toc95231343)

[**5.** **Summary** 8](#_Toc95231344)

### **Data source**

Dataset on international Airlines was obtained from Data.gov.au, the central source of Australian open government data. Dataset contains information on scheduled operations of international airlines operating to and from Australia. Monthly data contains information on passengers, freight and mail carried by airline by uplift/discharge country within single flight number services from 1985 to 2021.

Second dataset was obtained from Kaggle and contains information on airline ID, name, alias, IATA and ICAO membership, callsigns, country as well as airlines’ activity. Both datasets were obtained as csv files and chosen to analyse operations for different Airlines.

### **Data extraction**

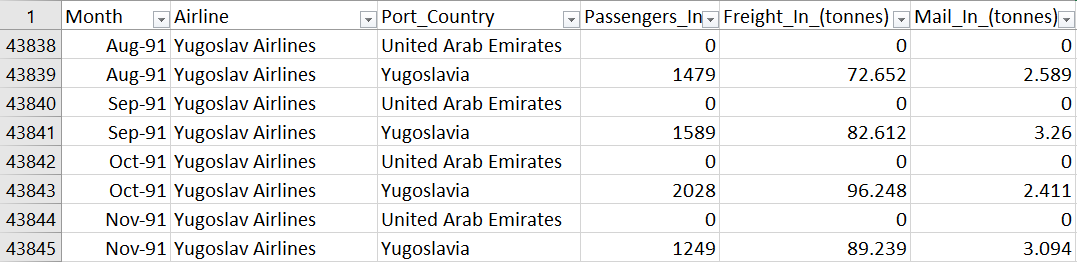
**Data source 1**: International Airlines, Operated Flights and Seats to and from Australia

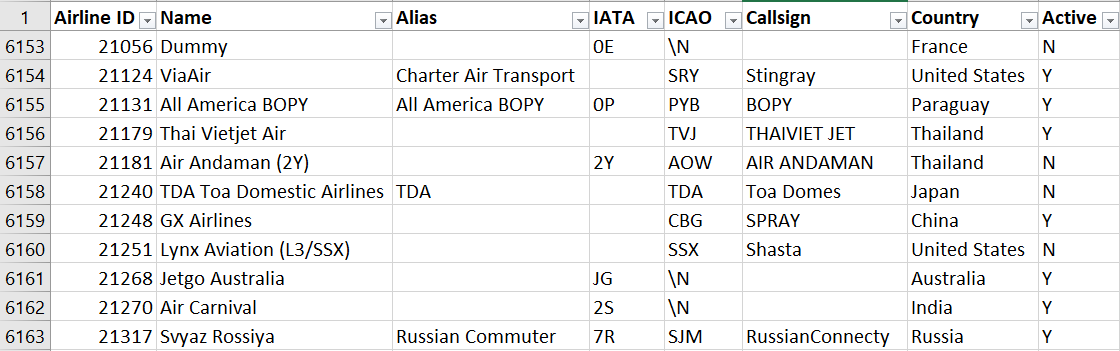
<https://data.gov.au/data/dataset/international-airlines-airline-by-country-of-port-data/resource/809c77d8-fd68-4a2c-806f-c63d64e69842?view_id=c2e9db61-be01-4673-b83e-d8b7f5b9dd8e>

**Data source 2**: Airline database  
<https://www.kaggle.com/open-flights/airline-database?select=airlines.csv>

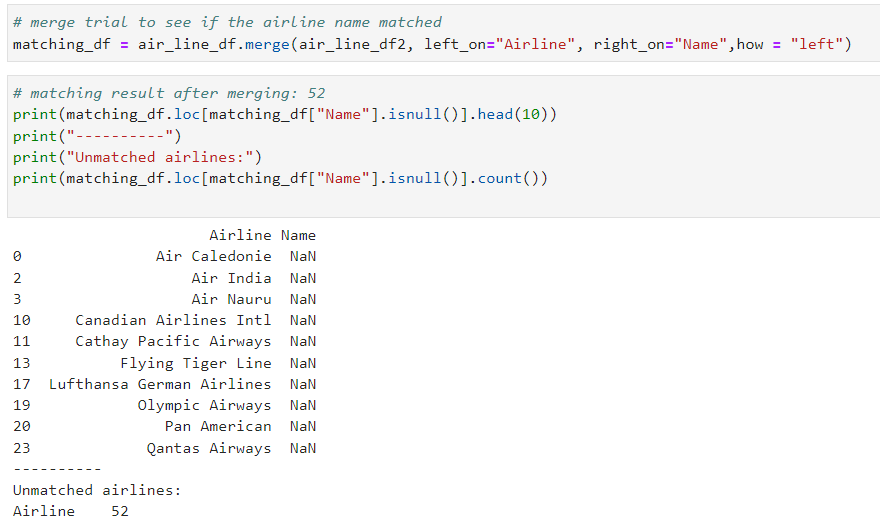
### **Data transformation**

1. Data source 1:

Australian Airports are connected to 74 porting countries. A table with a list of unique porting countries was extracted for data loading. Between data source 1 and 2, there is only one common column, airline name. Excel was used to explore the dataset to identify unmatched names. First step was to retrieve data as a csv file into a Jupyter Notebook. Pandas and Excel were used to preview the data. 

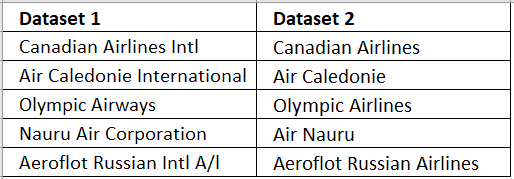


Firstly, it was reviewed whether the Airline name column matched:

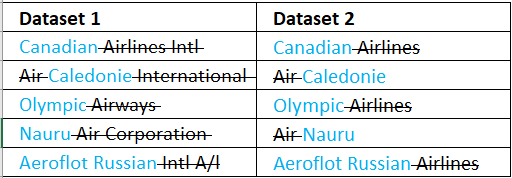


There were 52 unmatched Airline names due to different naming conventions in the two datasets.

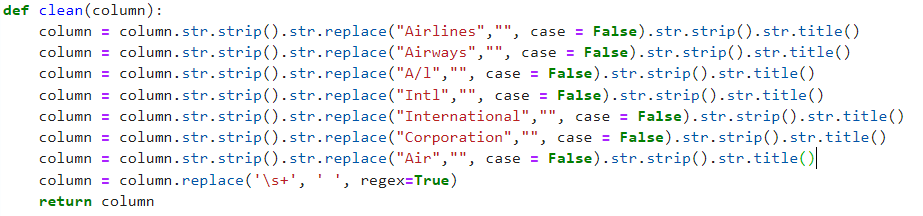
For example:



Due to this issue, common words within airline names were replaced in order to have consistent names.



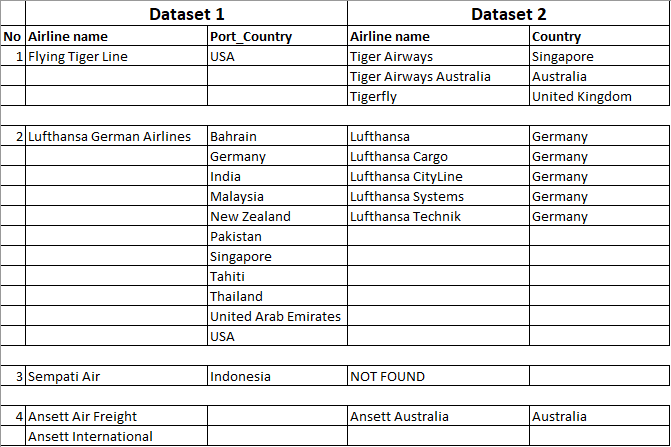
The solution was to develop a predefined function to replace the common words, disregard case sensitive, trim the text, rewrite text in proper format because some airlines have their name all in capital e.g. EVA Air vs Eva Air. Then finally replace multiple white space with a single space:



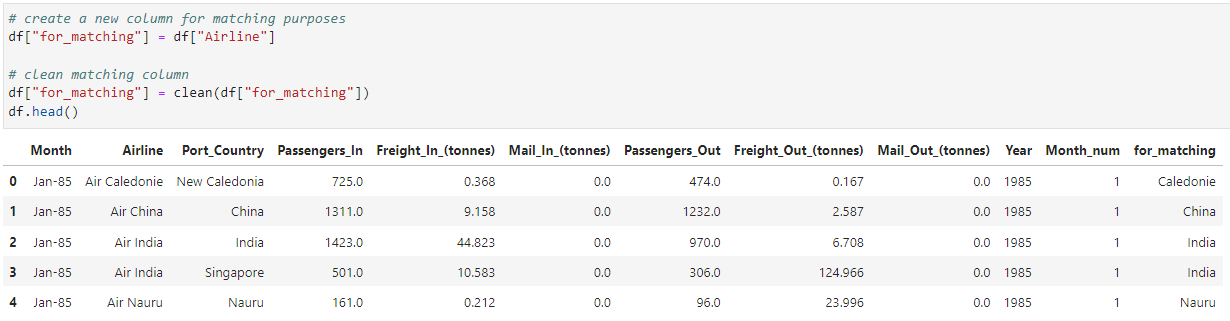
After using the clean() function, 26 unmatched records remained:

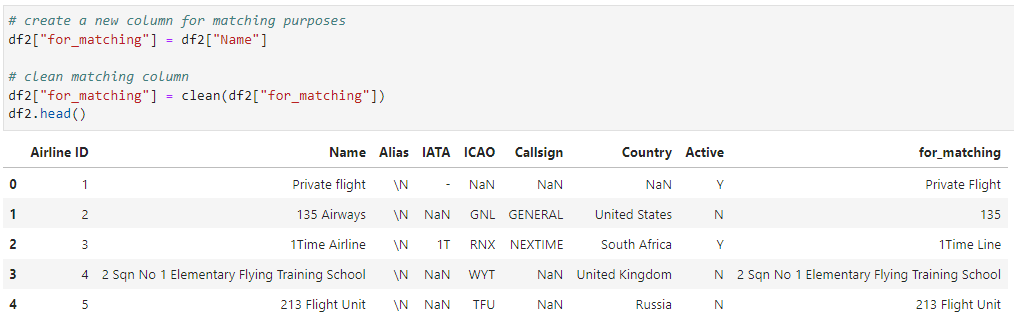


The remaining unmatched airline names are due to ambiguity of these names, e.g. Ansett Air Freight/ Ansett International vs Ansett Australia; or records not found. Therefore, these records were removed from the dataset.

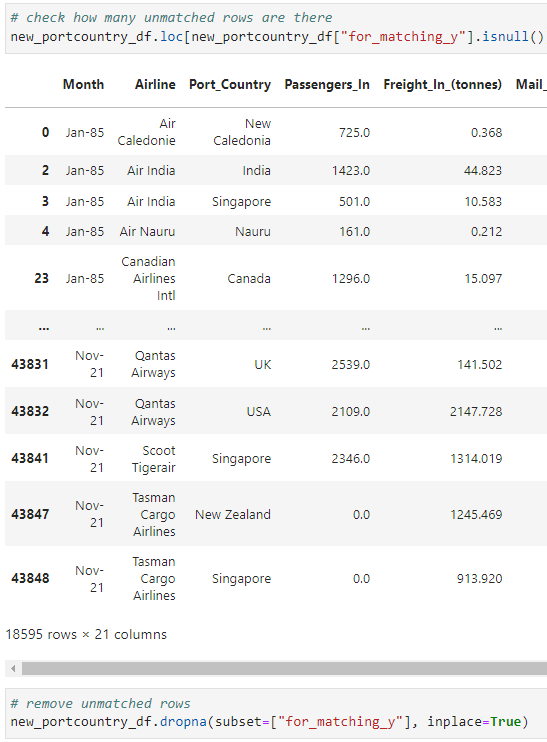


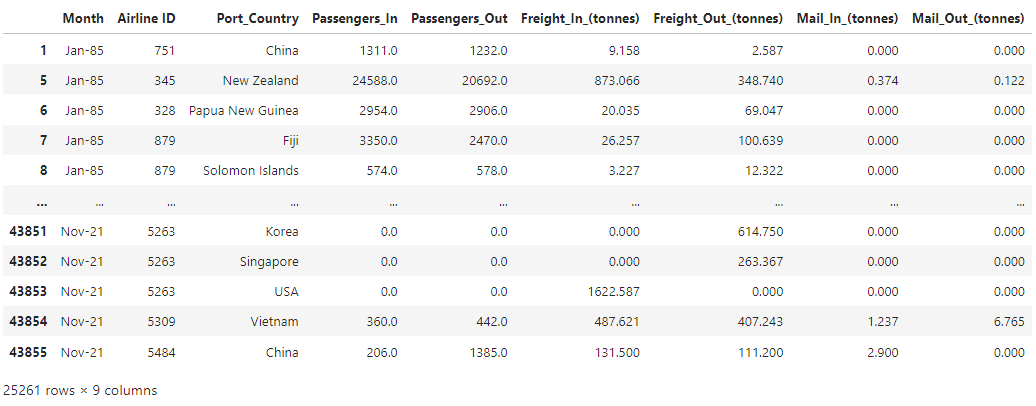
For matching purposes and keeping the original Airline column untouched, a dummy column was created and the name column was cleaned using the below procedure:





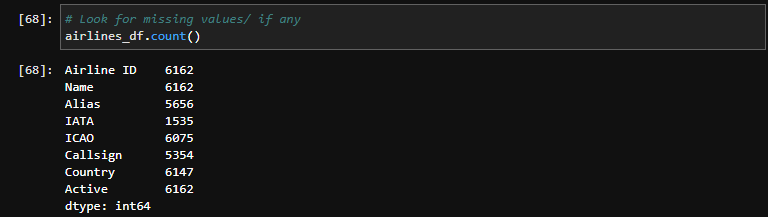
Finally, the two datasets were merged together and remove missing values for further uploading to server





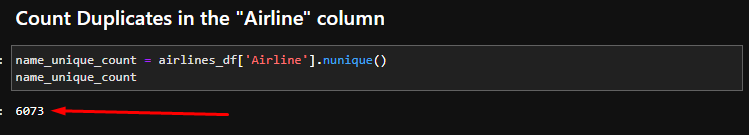
1. Data Source 2:

waschecked for any missing values. Count was performed to check how many fields in the data frame contained information.



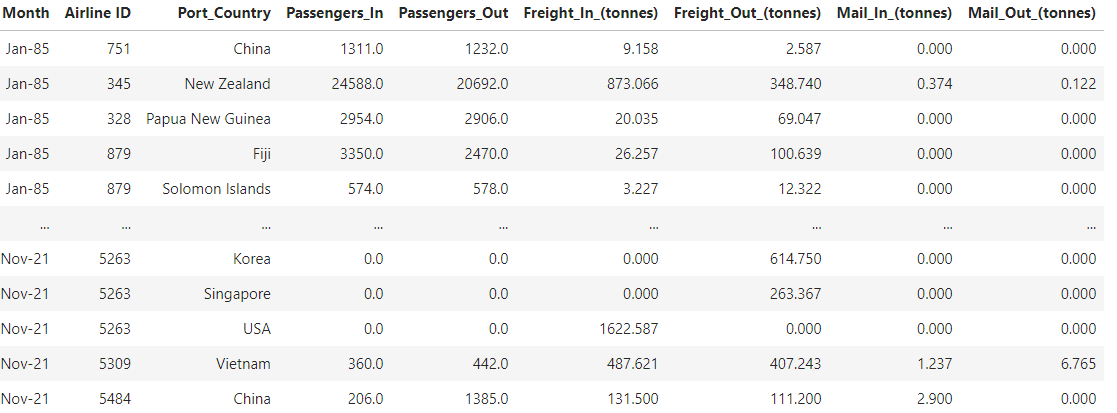
After sketching the ERD, two data sets were merged based on the Airline name. “Name” column was changed to “Airline”.

“Airline” column was checked for duplicate values to reduce the occurrence of an issue for the SQL Schema. As a consequence, 88 duplicate rows were removed.



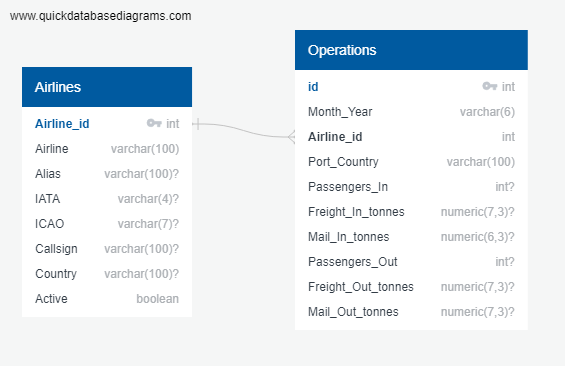
Cleaned data frame was exported to a csv (airlines\_cleaned.csv).

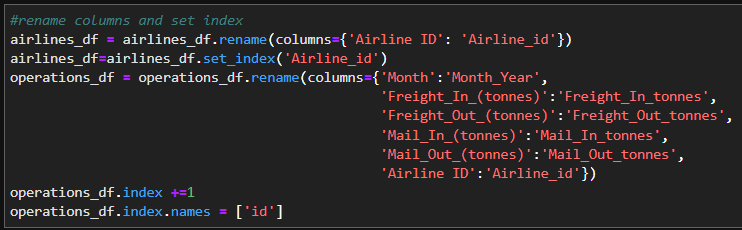
Two datasets were merged. Out of 45404 rows, 18595 rows were unmatched. These unmatched records were dropped and transformed data was exported for further loading.



### 

### **Data loading**

A relational database was created in PostgreSQL and the files were loaded using Python and SQL Alchemy. The Airlines table contains information on Airlines and the Operations table contains information on passengers, freight and mail carried by these airlines. As Airline names were not the same in the two datasets, Airline\_ID was used as the primary key to connect the Airlines and Operations datasets.  
Using the Python column headings were changed and indexes set to fit the data into the SQL format before being loaded into the database.



This database allows a user to analyse the differences in operation of airlines and in Australia, make comparisons between the airlines and find trends over 1985 to 2021.

### **Summary**

The following steps in the ETL process were taken:

1. Choose the data
2. Explore the dataset with excel
3. Read the data and place into a data frame
4. Extract the relevant data from the data frame
5. Transform the data:
   1. Remove duplicates
   2. Drop incomplete records
   3. Amend records
   4. Standardise data types
6. Create new data frames
7. Save final data frames to csv
8. Sketch ERD of the tables
9. Set primary keys
10. Import files to corresponding SQL tables