**INTRODUCTION**

**Purpose:**

The Aim is to reduce the original file into a desired file (i.e. without N/A and Duplicates).

**Scope:**

This program aims to achieve following Goals:

* Separate the single N/A data whose Aircraft Registration Number=” N/A”, Flight Number, Flight registration Number, Departure Airport Code and Arrival Airport Code does not match with other data.
* Separate the N/A data with Duplicates data whose Aircraft Registration Number=” N/A”, Flight Number, Flight registration Number, Departure Airport Code and Arrival Airport Code does match with other data.
* Separate the Duplicates data whose Flight Number, Flight registration Number, Departure Airport Code and Arrival Airport Code does match with other data.
* A merge data file is produce using N/A data with Duplicate data file.
* The final file is produce using merge file and drop\_duplicates() using pandas.

**Acronyms**

* Name\_of\_file: the input file name the user will give.
* Duplicates: defined for the rows whose Flight Number, Flight registration Number, Departure Airport Code, Arrival Airport Code does match with other data.

**Overview:**

The program is to filter the original file into a clean data file consisting of data without Aircraft Registration Number = ’N/A’ and merged duplicate rows. The role of user is to input the file name they want to clean.

The program will create 6 file naming:

* Name\_of\_file \_na\_data\_single.csv
* Name\_of\_file \_duplicate\_data\_with\_na.csv
* Name\_of\_file \_duplicate\_data\_without\_na.csv
* Name\_of\_file \_merge.csv
* Name\_of\_file \_notnull\_data.csv
* Name\_of\_file \_finalfile.csv

\_na\_data\_single file: consist of only Aircraft Registration Number = ’N/A’ without Duplicates.

\_duplicate\_data\_with\_na.csv file: consist of Aircraft Registration Number = ’N/A’ with their Duplicates.

\_duplicate\_data\_without\_na.csv file: consists of the duplicate row only from \_duplicate\_data\_with\_na files.

\_merge.csv file: consist of the merged rows using duplicate \_data\_with\_na.csv file. The merging of duplicate rows done w.r.t Flight registration Number, Departure Airport Code, Arrival Airport Code columns.

\_notnull\_data.csv file: consist of the data records in original file whose rows does not consist of Aicraft Registration Number =’N/a’ using notnull() method of pandas.

\_finalfile.csv file: it is the desired output user wants.

**Programming language**:

* Python 3.7.3

**Module:**

* Csv
* Pandas

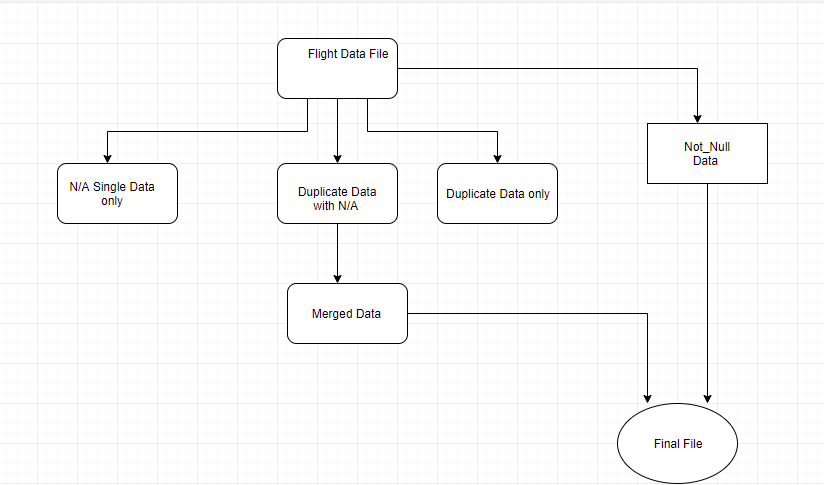
**Built-in functions used in source code:**

1. Input()
2. Open()
3. List()
4. Range()
5. Len()
6. Close()
7. Print()
8. csv.reader()
9. csv.writer()
10. read\_csv()
11. to\_csv()
12. notnull()
13. drop\_duplicates()

**Dependency in source code:**

* The \_duplicate\_data\_with\_na.csv is generated using original file using for loop and if and else condition on original file. This file is later used to generate merged data file.
* The merge file is generated using \_duplicate\_data\_with\_na.csv file using for loop and if and else condition on file data. This file is further used for generating final file.
* The final file can be generated using merge.csv file data and \_notnull\_data.csv file, where merge file is generated in above part of code

**Flow Chart:**



**Structure of Source Code:**

Step 1: importing modules i.e pandas and csv

Step 2: user input i.e name of file(without .csv extension)

Step 3: Filter Data with Single N/A in Na\_data\_single file.

Step 4: filter Data with N/A with and without duplicate in Duplicate\_with\_na and Duplicate \_without\_na file respectively.

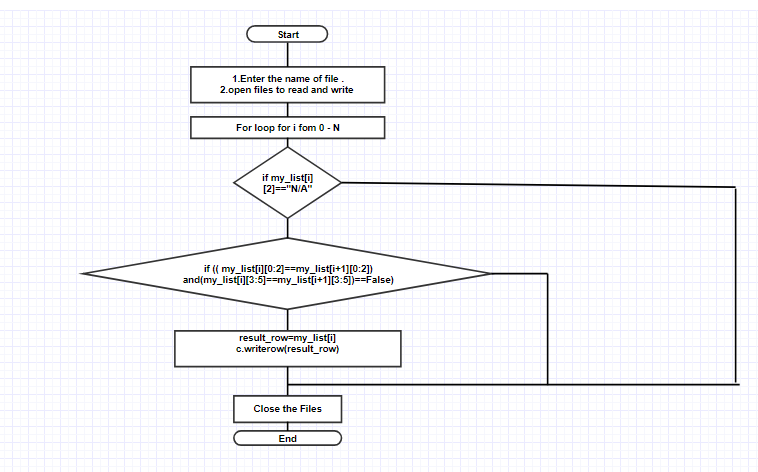
Step 5: now using duplicate\_with\_na file a newmerge file is created which consist of the merged data of N/A row and duplicate rows in one row w.r.t the first five columns.

Step 6: A not\_null file is created using panda’s method which consist of all the rows whose second column is not null, i.e where Aircraft Registration Number is not null.

Step7: The Not\_null and merge file is used to create a final file.

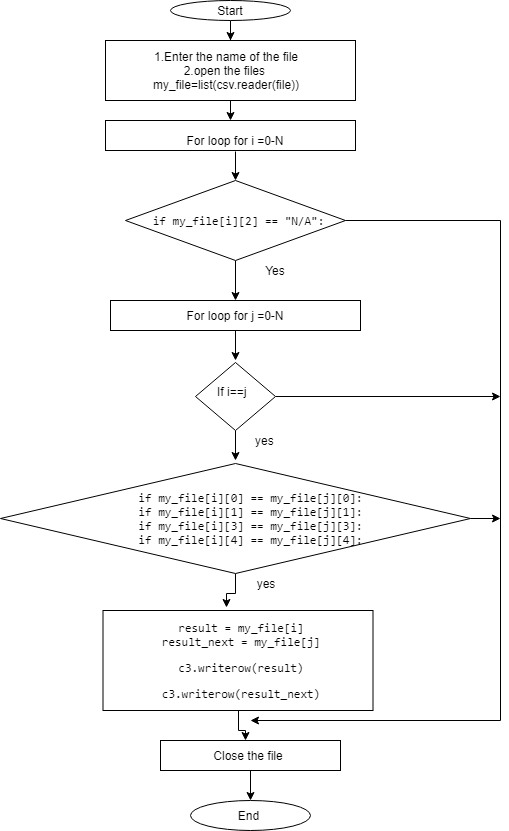
\*\* the files created does not consist of headers, the header list is later included in file using pandas method.

**Flow chart for Na\_data\_single file:**



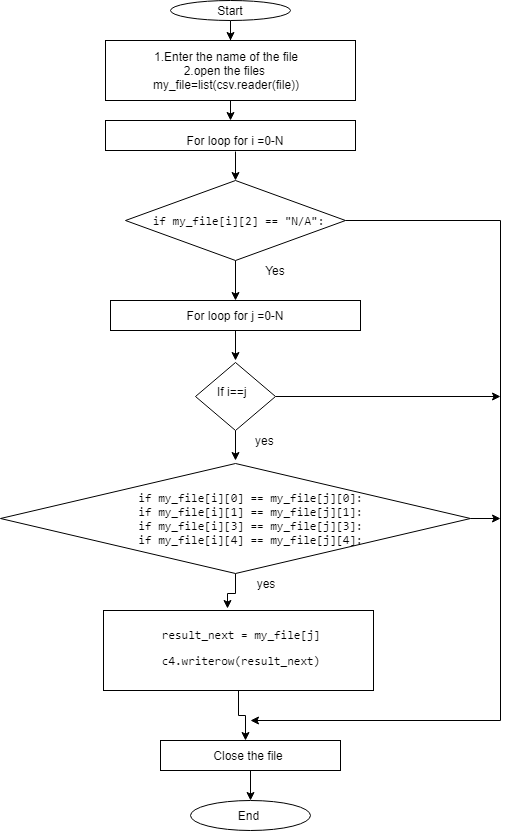
* the na data single file does not consist of headers , the headers are included using pandas method as shown below.
* **na\_data\_only=pd.read\_csv(Name\_of\_file + '\_na\_data\_single.csv', header=None)  
  na\_data\_only.columns=['Flight Number','Flight Departure Date','Aircraft Registration Number','Departure Airport Code','Arrival Airport Code','Aircraft Payload','Available Payload','Maximum Zero Fuel Weight (Kgs)','Actual Zero Fuel Weight (Kgs)','Take Off Fuel','Maximum Take Off Weight (Kgs)','Actual Take Off Weight (Kgs)','Trip Fuel Leg','Maximum Landing Weight (Kgs)','Actual Landing Weight (Kgs)','First Class Flown Passenger','Business Class Flown Passenger','Economy Class Flown Passenger','Total Male','Total Female','Total Children','Total Infants','Total Passenger Weight (Kgs)','Total Compartment Weight (Kgs)','Total Baggage Weight Leg (Kgs)','Total Cargo Weight Leg** (Kgs)','Total Mail Weight Leg (Kgs)','Total Engg Weight Leg (Kgs)','Dry Operating Weight (Kgs)','Cockpit Crew Count','Cabin Crew Count','First Class Non Revenue Passenger','Business Class Non Revenue Passenger','Economy Class Non Revenue Passenger']  
  na\_data\_only.to\_csv(Name\_of\_file + '\_na\_data\_single.csv',index=False)

**Flow chart for duplicate with na file:**



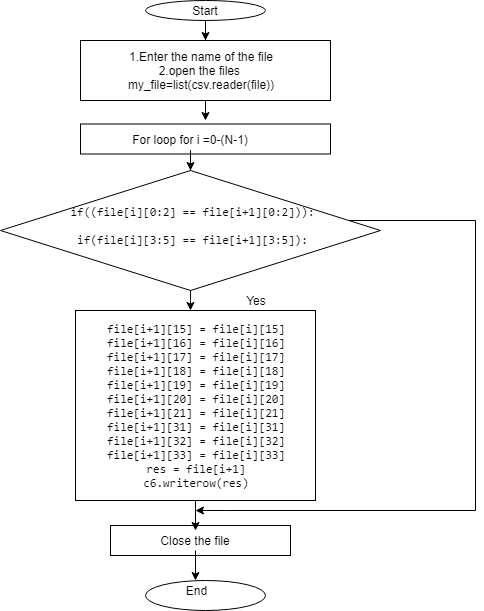
* the duplicate data with na file does not consist of headers which are included using pandas as shown below.
* **dup\_with\_na=pd.read\_csv(Name\_of\_file + '\_duplicate\_data\_with\_na.csv',header=None)  
  dup\_with\_na.columns=['Flight Number','Flight Departure Date','Aircraft Registration Number','Departure Airport Code','Arrival Airport Code','Aircraft Payload','Available Payload','Maximum Zero Fuel Weight (Kgs)','Actual Zero Fuel Weight (Kgs)','Take Off Fuel','Maximum Take Off Weight (Kgs)','Actual Take Off Weight (Kgs)','Trip Fuel Leg','Maximum Landing Weight (Kgs)','Actual Landing Weight (Kgs)','First Class Flown Passenger','Business Class Flown Passenger','Economy Class Flown Passenger','Total Male','Total Female','Total Children','Total Infants','Total Passenger Weight (Kgs)','Total Compartment Weight (Kgs)','Total Baggage Weight Leg (Kgs)','Total Cargo Weight Leg (Kgs)','Total Mail Weight Leg (Kgs)','Total Engg Weight Leg (Kgs)','Dry Operating Weight (Kgs)','Cockpit Crew Count','Cabin Crew Count','First Class Non Revenue Passenger','Business Class Non Revenue Passenger','Economy Class Non Revenue Passenger']  
  dup\_with\_na.to\_csv(Name\_of\_file + '\_duplicate\_data\_with\_na.csv',index=False)**

**Flow chart for duplicate\_data\_witout\_na file:**



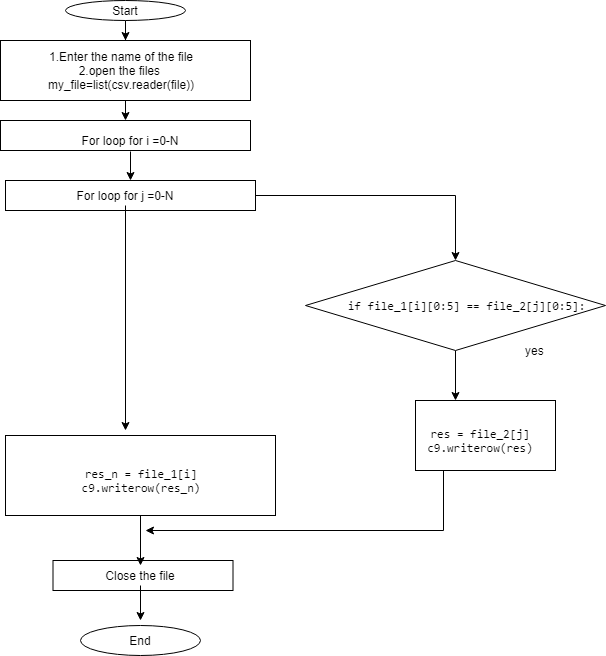
* the duplicate data without na does not consist of headers which are included using pandas method as shown below.
* dup\_without\_na=pd.read\_csv(Name\_of\_file + '\_duplicate\_data\_without\_na.csv',header=None)  
  dup\_without\_na.columns=['Flight Number','Flight Departure Date','Aircraft Registration Number','Departure Airport Code','Arrival Airport Code','Aircraft Payload','Available Payload','Maximum Zero Fuel Weight (Kgs)','Actual Zero Fuel Weight (Kgs)','Take Off Fuel','Maximum Take Off Weight (Kgs)','Actual Take Off Weight (Kgs)','Trip Fuel Leg','Maximum Landing Weight (Kgs)','Actual Landing Weight (Kgs)','First Class Flown Passenger','Business Class Flown Passenger','Economy Class Flown Passenger','Total Male','Total Female','Total Children','Total Infants','Total Passenger Weight (Kgs)','Total Compartment Weight (Kgs)','Total Baggage Weight Leg (Kgs)','Total Cargo Weight Leg (Kgs)','Total Mail Weight Leg (Kgs)','Total Engg Weight Leg (Kgs)','Dry Operating Weight (Kgs)','Cockpit Crew Count','Cabin Crew Count','First Class Non Revenue Passenger','Business Class Non Revenue Passenger','Economy Class Non Revenue Passenger']  
  dup\_without\_na.to\_csv(Name\_of\_file + "\_duplicate\_data\_without\_na.csv", index=False)

**Flow chart for merge file:**



* The merge file does not consist of headers , the headers are included in the file using pandas method i.e
* dup\_merge\_data=pd.read\_csv(Name\_of\_file + '\_merge.csv',header=None)  
  dup\_merge\_data.columns=['Flight Number','Flight Departure Date','Aircraft Registration Number','Departure Airport Code','Arrival Airport Code','Aircraft Payload','Available Payload','Maximum Zero Fuel Weight (Kgs)','Actual Zero Fuel Weight (Kgs)','Take Off Fuel','Maximum Take Off Weight (Kgs)','Actual Take Off Weight (Kgs)','Trip Fuel Leg','Maximum Landing Weight (Kgs)','Actual Landing Weight (Kgs)','First Class Flown Passenger','Business Class Flown Passenger','Economy Class Flown Passenger','Total Male','Total Female','Total Children','Total Infants','Total Passenger Weight (Kgs)','Total Compartment Weight (Kgs)','Total Baggage Weight Leg (Kgs)','Total Cargo Weight Leg (Kgs)','Total Mail Weight Leg (Kgs)','Total Engg Weight Leg (Kgs)','Dry Operating Weight (Kgs)','Cockpit Crew Count','Cabin Crew Count','First Class Non Revenue Passenger','Business Class Non Revenue Passenger','Economy Class Non Revenue Passenger']  
  dup\_merge\_data.to\_csv(Name\_of\_file + "\_merge.csv", index=False)

**Flow chart for Final file:**



* the final file does not conisist of headers which are included suing pandas as shown below.
* output=pd.read\_csv(Name\_of\_file + '\_finalfile.csv')  
  output.drop\_duplicates(subset=['Flight Number','Flight Departure Date','Aircraft Registration Number','Departure Airport Code','Arrival Airport Code'], inplace = True)  
  output.to\_csv(Name\_of\_file + '\_finalfile.csv', index=False)

**Conclusion:**

User only need to enter the Name of file; the rest will be done by the program. The work flow of program is shown in above flow char

t diagrams.