**Metadata Information**

**Information on new data created**

**Note** These csv’s are not present in the folder due to space restriction but can be found and generated using the ipynb’s code. The name is same as mentioned

1. **Valid\_airports.csv** – Generated in part1.ipynb file. This data has following columns

* TYPE
* NAME
* IATA\_CODE

It only contains information for medium/large sized airports.

**USE**: It helps in filtering those flights information which fly between medium/large airports.

1. **Busiest\_flight\_routes.csv** – Generated in part1.ipynb. This data has following columns
   * ORIGIN
   * DESTINATION
   * Number of flights

**USE**: Used to find out top 10 busiest round-trip routes.

1. **all\_flights\_profit\_data.csv** – Generated in part2.ipynb. This data has following columns
   * FL\_DATE
   * ORIGIN
   * DESTINATION
   * DEP\_DELAY
   * ARR\_DELAY
   * AIR\_TIME
   * DISTANCE
   * OCCUPANCY
   * AVERAGE\_ITIN\_FARE
   * Fare\_confidence
   * TYPE\_ORIGIN
   * TYPE\_DESTINATION
   * Revenue
   * Cost
   * Profit

**USE**: Used to calculate top 10 most profitable routes. Also used in the calculation of top 10 recommended flights as average profit, fare confidence and average delay was generated from this data.

1. **Fare\_confidence feature**- new metric which was devised by me which helped in deciding how much confidence/trust we should put in the average fare between a route that we calculated.

It plays a very crucial role while deciding the recommended flights. More information on this can be found in the section 2.2 in the report.

1. **Profit confidence feature**- new metric which is the product of average profit and fare confidence for that route.
2. **Delay feature** - simply an average value of arrival and departure delay for a route.