Assignment: 00 - Sequential Analysis

Description:

The Bureau of Transport Statistics' On-time Performance (OTP) dataset has information about flights in the USA. The full dataset covers 27 years of air travel and is over 60GB in plain text. For this assignment you should answer the question: Which airlines have the least expensive fares?

Details:

- Individual assignment.
- Data for one month is here (corrected).
- Data definition.
- Write a sequential Java program reading one gzipped file and writing results on the console.
- The only output should be the K and F, one per line, where K is the number of corrupt lines of input (lines not in the same format as the rest and lines with flights that do not pass the sanity test), F is the number of sane flights. Next, output pairs "C p" where C is a carrier two letter code and p is the mean price of tickets. Sort the list by increasing price.
- Th sanity test is:

```
CRSArrTime and CRSDepTime should not be zero

timeZone = CRSArrTime - CRSDepTime - CRSElapsedTime;

timeZone % 60 should be 0

AirportID,  AirportSeqID, CityMarketID, StateFips, Wac should be larger than 0

Origin, Destination,  CityName, State, StateName should not be empty

For flights that not Cancelled:

ArrTime -  DepTime - ActualElapsedTime - timeZone should be zero

if ArrDelay > 0 then ArrDelay should equal to ArrDelayMinutes

if ArrDelay < 0 then ArrDelayMinutes should be zero

if ArrDelayMinutes >= 15 then ArrDel15 should be false
```

- Design the code with care as you will reuse it. Document and test it.
- The reference solution is ~1KLOC and prints 4083 for K, 435940 for F and, for instance, "UA 545.62". Processing time is ~10 seconds.