Task 2

The objective is to create reports on the average and median departure delays of (a) all the airports, and (b) all the airways in the dataset. You should give four reports, two for the airports (average/median delays) and two for the airways (average/median delays). Each report is a CSV file containing one line for each airport/airway and the lines of each file should be ordered (in descending order) based on the corresponding criterion (average/median delay). No header files are required for these files. An extra instruction you have from your supervisor is that you should take care of some data outliers: you should not consider in your analysis any airports/airways that have extremely low number of flights; the criterion is that any airport/airway belonging in the lowest 1% percentile, regarding the number of flights, should be omitted.

First, we had to create the dataset upon which we worked with. We read the whole international flights file and then we created a dataset for each one of the reports we had to deliver. The "flights" table contained those airports (from the "ORIGIN" column only, because the "DEST" column contained the same airports) that conducted flights having as total more than the 1% of all the flights (1% percentile), while the "airways" table consisted of the airways (from CARRIER column) belonging to the 1% percentile as well. In order to have more accurate results we used both the negative and positive values of the column "DEP_DELAY" (Departure delays). Also, the records contained null values were replaced with the 0 number (meaning no delay for that flight).

After that, we created a table containing the airports and for each one of them the average delay of the departures. Then we exported that data in a csv file (named "task2-ap-avg.csv") without having a header and being ordered in descending order by the average. We also, rounded the numbers in 3 decimals so as to have more readable results. The same procedure was done and for finding the average departure delays of all the airways in the dataset. (exported in file named "task2-aw-avg.csv")

Later, we had to process the data in order to find the median of both the airports and the airways. To begin with, the median is the value of the records that placing them in ascending order, it separates equally the upper half of the dataset with the lower one. So, if the size of the dataset is an odd number then the median is exactly the value explained previously, but if the length of the dataset is an even number then the median is defined by the average of

the middle and the next from it number. So, for finding the average departure delays, we sorted them in ascending order, and we grouped them by airport. Then, we created an index iterating the number of delays for each airport. After that, we found for each airport the length of its records and we assigned the middle to a new column. If the size of the airport's flights were an even number, then we assigned to a column "P" the middle and to "P1" the next number after the middle. Later, it was necessary to match each number, to the line referring to the departure delay of this specific airport and after that, we had to join the two tables. So, now we have a table with an airport and two other columns referring to the departure delays of the middle rows. So, we took the average of those two delays and found the median of each airport. Finally, we exported the above table to a csv file named "task2-ap-med.csv" ordering it in descending order by median value. It is worth mentioning that in this table we observed that for all the airports the middle row and the one next to it had the same departure delay.

For the fourth report we followed the above steps, and we exported the specific table to a file named "task2-aw-med.csv". This report contains the median values of all the airways of our dataset.