Task 2

For this task you continue to work with SparkSQL. 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.

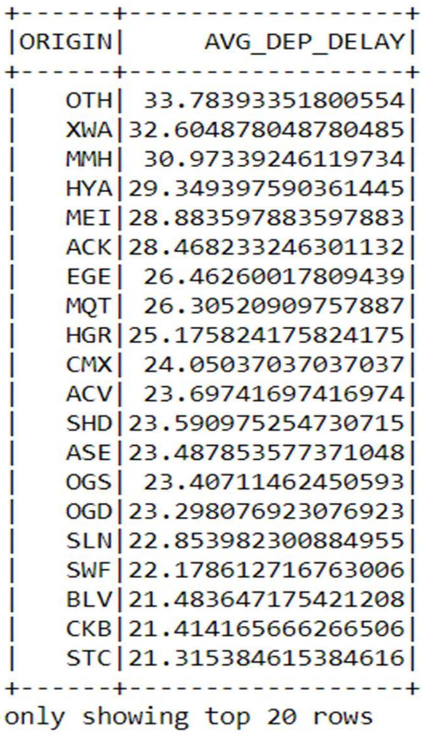
Regarding the second task, first I create a subset name df2 of the original cleaned flights_data. We shall check also for duplicate rows in this dataset (duplicates are 17427) and drop them.

After these data cleaning operations of the subset, we have to examine for outliers regarding the number of flights and drop the observations that below to the lowest 1% percentile. We group by the column Origin and aggregate the number of flights, after we sort this output with ascending order because we want to find which airport had the lowest number of flights. The value of the 1% percentile is 82, this means that airports with total number of flights below to 82 are belonging in the lowest 1% of the sample. Specifically, the airports, labeled as AKN, PGV, GST, DLG belong to the lowest percentile of 1% thus we exclude these observations from our subset.

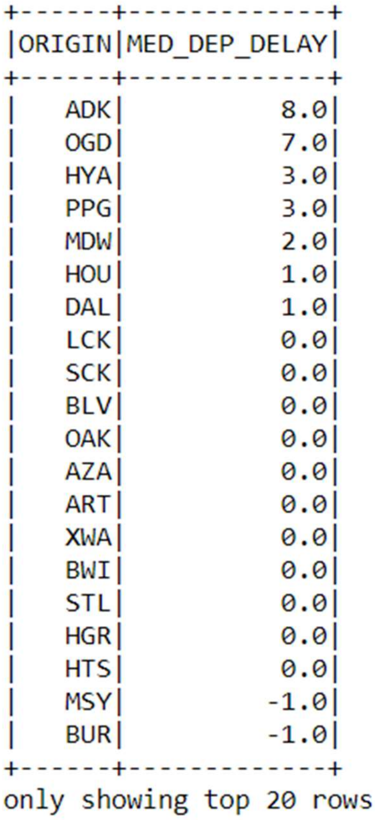
Now, we are going to answer in the question and find the average and median departure delay by airport (ORIGIN) and by airways (CARRIER). We implement basic group and order by operations for columns ORIGIN and CARRIER separately regarding the median we want to calculate the percentile of 0.5 which indicates the value which the 50% is below it.

Finally, after the calculations we can say that the highest average departure delay by airport is 33 minutes in the airport labeled OTH while the median departure delay by airport is 8 minutes in the airport labeled ADK. On the other hand the highest average departure delay by airways is 17.97 minutes for the airways labeled as B6 while the highest meadian value for departure delay by airways is 0 . For the final implementation of the question, we have to export the first 100 observations of the results in csv format files. Output of the code is shown below.

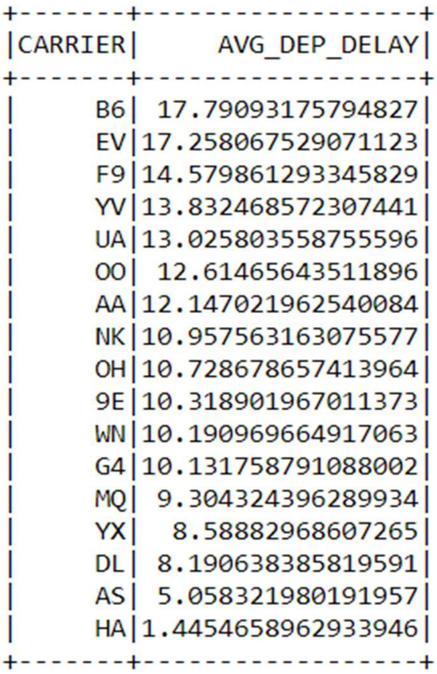
Figure 1 - By Origin avg



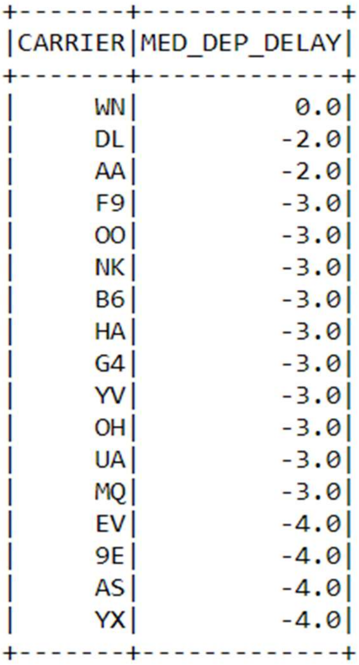
```
+------+------------------+
|ORIGIN|     AVG_DEP_DELAY|
+------+------------------+
|   OTH| 33.78393351800554|
|   XWA|32.604878048780485|
|   MMH| 30.97339246119734|
|   HYA|29.349397590361445|
|   MEI|28.883597883597883|
|   ACK|28.468233246301132|
|   EGE| 26.46260017809439|
|   MQT| 26.30520909757887|
|   HGR|25.175824175824175|
|   CMX| 24.05037037037037|
|   ACV| 23.69741697416974|
|   SHD|23.590975254730715|
|   ASE|23.487853577371048|
|   OGS| 23.40711462450593|
|   OGD|23.298076923076923|
|   SLN|22.853982300884955|
|   SWF|22.178612716763006|
|   BLV|21.483647175421208|
|   CKB|21.414165666266506|
|   STC|21.315384615384616|
+------+------------------+
only showing top 20 rows
```



```
+------+-------------+
|ORIGIN|MED_DEP_DELAY|
+------+-------------+
|   ADK|          8.0|
|   OGD|          7.0|
|   HYA|          3.0|
|   PPG|          3.0|
|   MDW|          2.0|
|   HOU|          1.0|
|   DAL|          1.0|
|   LCK|          0.0|
|   SCK|          0.0|
|   BLV|          0.0|
|   OAK|          0.0|
|   AZA|          0.0|
|   ART|          0.0|
|   XWA|          0.0|
|   BWI|          0.0|
|   STL|          0.0|
|   HGR|          0.0|
|   HTS|          0.0|
|   MSY|         -1.0|
|   BUR|         -1.0|
+------+-------------+
only showing top 20 rows
```



```
+-------+------------------+
|CARRIER|     AVG_DEP_DELAY|
+-------+------------------+
|     B6| 17.79093175794827|
|     EV|17.258067529071123|
|     F9|14.579861293345829|
|     YV|13.832468572307441|
|     UA|13.025803558755596|
|     OO| 12.61465643511896|
|     AA|12.147021962540084|
|     NK|10.957563163075577|
|     OH|10.728678657413964|
|     9E|10.318901967011373|
|     WN|10.190969664917063|
|     G4|10.131758791088002|
|     MQ| 9.304324396289934|
|     YX|  8.58882968607265|
|     DL| 8.190638385819591|
|     AS| 5.058321980191957|
|     HA|1.4454658962933946|
+-------+------------------+
```



```
+-------+-------------+
|CARRIER|MED_DEP_DELAY|
+-------+-------------+
|     WN|          0.0|
|     DL|         -2.0|
|     AA|         -2.0|
|     F9|         -3.0|
|     OO|         -3.0|
|     NK|         -3.0|
|     B6|         -3.0|
|     HA|         -3.0|
|     G4|         -3.0|
|     YV|         -3.0|
|     OH|         -3.0|
|     UA|         -3.0|
|     MQ|         -3.0|
|     EV|         -4.0|
|     9E|         -4.0|
|     AS|         -4.0|
|     YX|         -4.0|
+-------+-------------+
```