Task 1

Your first task is to calculate the average flight delays in the dataset. Your supervisor made it clear that you should choose SparkSQL with Python and DataFrames, so that your code should be compatible with other software products of your company.

Before everything I check for duplicated values in the entire dataset and drop them.

```
number of rows all data: 7422037
number of rows after deleting duplicates: 7418521
duplicate data: 3516
number of rows final: 7418521
```

Regarding the first task, we shall work with a subset of the original dataframe, named df_delay. First, I consider checking for duplicated rows since the total rows of the file are 7,418.521. The findings of these checks are that there are 17610 rows of duplicated data, an extremely low percentage, so I decide to drop them. Afterwards the requested information is the average flight delays, so we have to split the subset in order to calculate the average of the column DEP_DELAY and the average of the column ARR_DELAY separately. The results are that the average time for a departure delay is 10.93 minutes and the average time for an arrival delay is 5.41 minutes.

Tables are shown below.

```
In [107]: # AVG_DEP_DELAY ALL

          AVG_DEP_DELAY = df_delay.select(F.avg("DEP_DELAY")).show()
```

```
+------------------+
|    avg(DEP_DELAY)|
+------------------+
|10.932308482294294|
+------------------+
```

```
In [108]: # AVG_ARR_DELAY ALL

          AVG_ARR_DELAY =  df_delay.select(F.avg("ARR_DELAY")).show()
```

```
+-----------------+
|   avg(ARR_DELAY)|
+-----------------+
|5.424550657897001|
+-----------------+
```