Challenge description

The objective of this challenge is to analyze using Apache Spark a data set containing the results of an experiment involving two algorithms. The data set is constituted by a file in csv format (or xls) with the following structure:

F1	F2	F3	F4	Rep	M_OF	M_TIME(s)	CBC_OF	CBC_TIME	CBC_FC	CBC_PC
3	400	9	A	1	361	0,266	361	0,462	0	0
3	400	9	A	2	306	0,511	306	0,106	0	0
3	400	9	A	3	271	1,052	271	1,646	0	4
3	400	6	A	4	263	0,556	263	0,179	0	1
3	400	9	A	5	307	4,706	307	4,633	0	4
3	400	7	A	6	270	1,139	270	0,057	0	0
3	400	8	A	7	251	0,726	251	0,188	0	1
3	400	7	A	8	293	1,6	293	0,023	0	0

This data set shows the results of two different algorithms (M and CBC) applied to instances characterized by 4 features (F1-F4). Each experiment was repeated 10 times (Rep) generating randomly the instances. It is important to note that the F3 feature is actually a range of values and at each repetition a new value is randomly generated within the range. For this reason, instances belonging to the same group can have distinct F3 values.

In addition:

1. **M_OF** represents the value of the objective function obtained by the algorithm M
2. **M_Time** represents the time spent by the algorithm M
3. **CBC_OF** represents the value of the objective function obtained by the CBC algorithm
4. **M_Time** represents the time spent by the CBC algorithm
5. **CBC_FC** represents the number of FC-type actions executed by the CBC algorithm
6. **CBC_PC** represents the number of PC-type actions executed by the CBC algorithm

You are asked to:

1. Carry out a descriptive statistical analysis of the table data
2. Evaluate the effectiveness of the algorithms using the OF value to compare the algorithms. Use a statistical test and appropriate visualizations.
3. Evaluate the efficiency of the algorithms using the TIME value to compare the algorithms. Use a statistical test and appropriate visualizations.
4. Study the correlation between features and values of OF/Time, FC and PC. Which features have the greatest impact on the execution time of the two algorithms?

Finally, we ask you to use primarily the features made available by Spark but if necessary you can also use those provided by libraries of the Python language (eg. for results visualization).