Challenge description

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

F1	F2	F3	F4	Rep	MODEL_OF	MODEL_OFU	MODEL_TIM	CBC_OF	CBC_TIME	CBC_FC	CBC_PC	NEWCBC_O	EWCBC_TIN	NEWCBC_FC
3	400	9	A	1	197	197	0.29	197	0.02	0	1	197	0.056	0
3	400	9	A	2	162	162	0.24	162	0.02	0	0	162	0.016	0
3	400	9	A	3	165	165	0.85	165	0.01	0	0	165	0.023	0
3	400	6	A	4	178	178	0.311	178	0.02	0	1	178	0.012	0
3	400	9	A	5	220	220	0.72	220	0.03	0	1	220	0.038	0
3	400	7	A	6	162	162	1.085	162	0.03	0	2	162	0.014	0
3	400	8	A	7	108	108	1.58	108	0.05	0	1	108	0.092	0
3	400	7	A	8	201	201	0.11	201	0.02	0	1	201	0.016	0
3	400	10	A	9	201	201	0.58	201	0.05	0	2	201	0.031	0
3	400	8	A	10	156	156	0.19	156	0.05	1	1	156	0.074	0
3	500	8	A	1	159	159	0.95	159	0.04	0	2	159	0.018	0

This data set shows the results of three different algorithms (Model, CBC and NEWCBC) applied to instances characterized by 4 features (F1-F4). Each experiment was repeated 10 times (Rep, but notice that in 2 cases we only made one repetition) 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. **MODEL_OF** represents the value of the objective function obtained by the algorithm M
2. **MODEL_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. **CBC_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
7. **NEWCBC_OF** represents the value of the objective function obtained by the NEWCBC algorithm
8. **NEWCBC _Time** represents the time spent by the NEWCBC algorithm
9. **NEWCBC _FC** represents the number of FC-type actions executed by the NEWCBC algorithm 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 (first verify that the conditions for the selected test holds) and appropriate visualizations.
3. Evaluate the efficiency of the algorithms using the TIME value to compare the algorithms. Use a statistical test (first verify that the conditions for the selected test holds) 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? Is it possible to predict the TIME or OF variable? (you can make them categorical).

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 (e.g. for results visualization).