3th of June, 2020,

Dear colleagues,

You will find hereafter the business statement of the case I have selected for the IBM Data Science Capstone project, as well as a description of the data needed and the concrete objectives associated with the use case.

The business statement is the following: in the previous exercice, a caracterisation of the cities of New York and Toronto has been made separately on the basis of data gathered from the Four Square APIs. It would be of practical interest for commuters to be able to identify the main differences between the two and to benchmark them with a European city (in this case, Brussels).

As a reminder, this clusterisation has put in relief the similarities and differences between the neighboroughs of the two cities, on a certain number of dimensions and features of the most appreciated venues of the cities.

However, the exercice has been made separately for each of the two cities which a priori have quite comparable positionings in their respective countries, as being very diverse and financial capitals in their own country. A first objective of this work will be to compare the two cities on the basis of these same dimensions.

A second step will be to extend the comparison with a European representative city. We have chosen Brussels for its multicutural dimension and its central positioning with respect to European institutions and some international companies.

The added value of this exercice is to bring light on significant similarities and differences between the three cities, and this can be useful as showcases of elements to be taken into account for mobilities between these cities (and in particular between (New York, Toronto on one side, Brussels on the other side).

In order to perform the exercice, the considered data perimeter will be based on the following sources (and related assumptions):

- Four Square related data of most appreciated venues irrespective of their nature (with zoom on specific categories if needed)

- List of neighborhoods of the 3 cities with their respective coordinates, obtained from specific local sources; we will keep the number of neighborhoods in each city in the same order of magnitude

We intend to analyse these data with a KNN algorithm which will be tuned so as to perform:

- an analysis with 3 clusters, so as to qualify the global level of specificity of each town with respect to the 2 others

- an analysis with a higher amount of clusters so as to produce a more fine-grained analysis.

Each cluster will be analyzed and characterised with respect to their more salient points. Finally, a PCA analysis will be performed so as to identify the more significant variables of the analysis, so as to sharpen the diagnosis.

The methodology of analysis and the data analysis done will be exposed in a synthesis document.