Université Claude Bernard Lyon 1

Rapport de TP Saadallah Ahmed et Khelili Adrian

I) Réduction de dimensions et Visualisation des données 2
- I.I) Villes 2
- I.II) Crimes 3
- I.III) Startups 3

II) Clustering 4

Université Claude Bernard Lyon 1

I) Réduction de dimensions et Visualisation des données

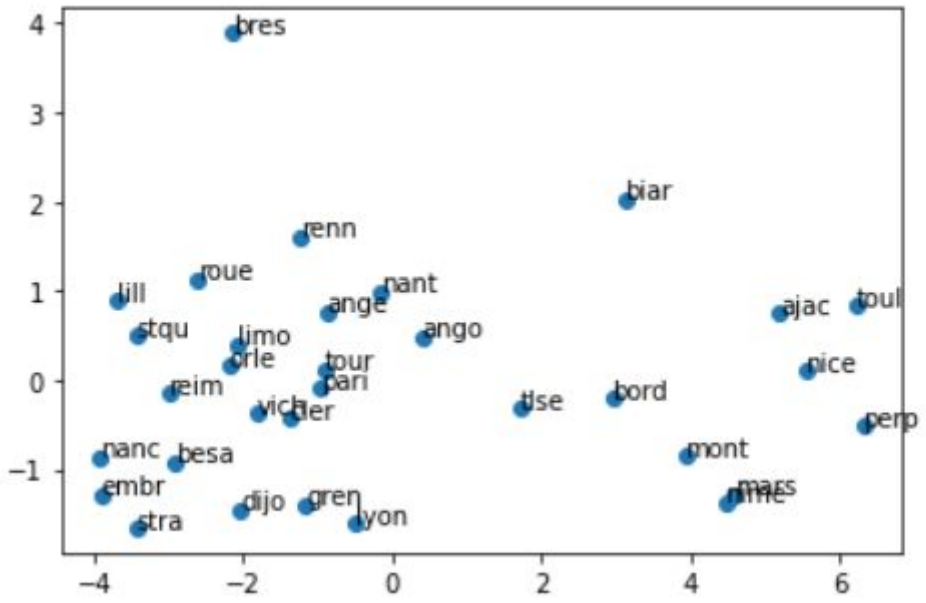
I.I) Villes

1- Le nombre d'axes à retenir pour conserver un minimum de 90% d'information est : 2
Car les deux premiers axes apportent 0.9898 d'information.

2- L'interprétation des deux premiers axes principaux :
Pour le premier axe : si nous avons des valeurs positives, alors cela veut dire que la température moyenne est forte sur toute l'année, à l'inverse les températures moyennes sont très faibles.

Pour le deuxième axe : si nous avons des valeurs élevées pour une ville alors elle a une forte température l'hiver (décembre, janvier, février) et faible l'été (mai, juin, juillet), sinon elle a une forte température l'été et faible en hiver.
(On peut s'apercevoir que Brest a des températures modérées en été et en hiver se trouve isolé tout en haut de l'axe car il ne fait pas très froid en hiver et pas très chaud en été).

3- Représentation graphique des deux axes :



I.II) Crimes

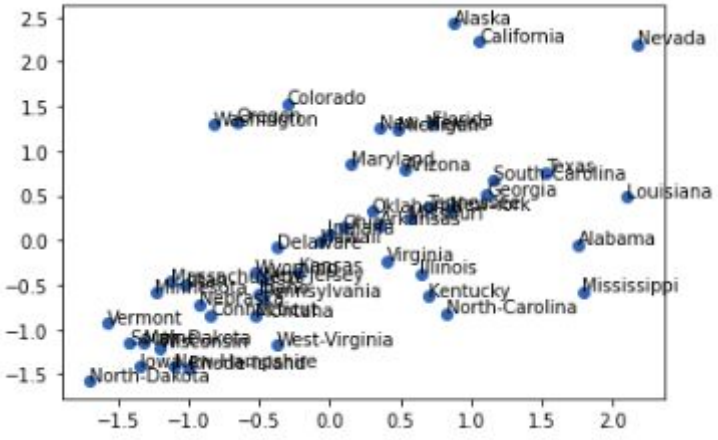
1- Les nombre d'axes à retenir pour conserver un minimum de 90% d'information est : 4
Car les quatre premiers axes apportent 0.9137 d'information.

2- Étant donné que le nombre d'axes minimum à retenir est 4, nous allons utiliser le critère de kaiser. On garde donc les 2 premiers axes malgré la perte d'information.
Si dans le premier axe on a une valeur positive ça veut dire que les délits de tout genre sont élevés dans cette ville.

Pour le deuxième axe, plus les valeurs tendent le bas, plus les délits graves sont élevés et inversement quand les valeurs tendent vers le haut.

(Par exemple : Plus les données sont élevées, plus ces destinations peuvent être préférentielles pour les touristes.)

3- On dessine l'axe, mais il est à noter que ces deux axes nous apportent uniquement 75% d'informations (perte de 25% de l'information), il faut donc rester prudent quant à l'interprétation de cette visualisation :



I.III) Startups

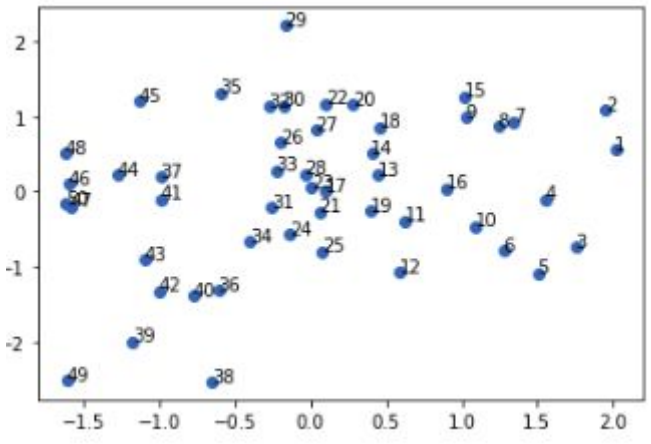
1- Les nombre d'axes à retenir pour conserver un minimum de 90% d'informations est : 2
Car les deux premiers axes apportent 0.9228 d'information.

2-
Premier axe : Plus nous avons des valeurs élevées sur cet axe plus les dépenses R&D, les dépenses Marketing Spend, Bénéfice sont élevés.

Deuxième axe : Plus les valeurs sont faibles, plus les dépenses administratives sont élevées, plus elles sont grandes, moins les dépenses sont élevées.

On peut donc déduire que faire des dépendances administratives n'apportent pas vraiment de bénéfices tandis que les dépenses marketing et dépenses R&D font augmenter les bénéfices. Il vaut donc mieux investir en R&D et en marketing car elles sont corrélées au bénéfice.

3- Représentation graphique des deux axes :



II) Clustering

2- Après avoir appliqué les trois algorithmes de clustering on déduit que le partitionnement en deux clusters est le meilleur car il donne le coefficient silhouette (~ 0.61) le plus proche de 1. Ce qui veut dire qu'en moyenne les points sont plus proches de leurs clusters internes que des clusters externes.

3- Le meilleur algorithme de clustering pour un partitionnement en trois partitions est l'Agglomerative Clustering avec la méthode d'agrégation "average" parce qu'il a donné le plus grand coefficient silhouette (~0.49) comparé aux autres algorithmes et cela veut dire qu'il minimise le mieux les distances à l'intérieur d'un cluster et maximise les distances par rapport aux clusters externes ; cela est dû à (d'après la visualisation graphique) la considération de la ville Brest comme une valeur aberrante et la mettre toute seule dans un cluster.