1. L’élève a-t-il utilisé une fonction du module cluster de scikit-learn pour laquelle le nombre de clusters est un paramètre à fournir (typiquement, k-Means) ? Noter que DBSCAN ne s’applique pas ici, on sait combien de clusters on veut.

2. Il est recommandé de scaler les données avant d’appliquer une méthode de réduction de dimension… voire avant le clustering.

3. Le pourcentage de variance expliqué par les deux premières composantes d’une ACP sur ces données est faible. On préfèrera une méthode non-linéaire telle que tSNE.

4. Pour la visualisation : est-elle claire ? Tous les points apparaissent-ils bien sur le graphe ?

5. Pour l’évaluation intrinsèque, le plus simple avec scikit-learn est d’utiliser le coefficient de silhouette. Il est positif, ce qui veut dire que chaque point est bien dans le meilleur cluster possible. On peut aussi implémenter soi-même une des autres méthodes vues dans le chapitre 2, bien sûr.

6. Pour la comparaison aux étiquettes, le plus simple est d’utiliser l’indice de Rand ajusté car nous l’avons déjà rencontré. Scikit-learn propose néanmoins de nombreuses autres méthodes utilisant labels\_true dans la liste à l’URL :

<http://scikit-learn.org/stable/modules/clustering.html#clustering-evaluation>