Contexte de l'analyse

Elu métier le plus sexy par la Harvard Business Review en octobre 2012, le data scientist représente un profil rare qui exige de nombreuses compétences.

A partir d'un dataset, vous réaliserez :

- un clustering non supervisé afin d'identifier le nombre optimal de groupes de profils techniques distinctes
- une prédiction des profils manquants

Données

data.csv contient 6 variables :

- 'Entreprise' correspond à une liste d'entreprises fictive
- 'Metier' correspond au métier parmi data scientist, lead data scientist, data engineer et data architecte
- 'Technologies' correspond aux compétences maîtrisées par le profil
- 'Diplome' correspond à son niveau scolaire (Bac, Master, PhD,...)
- 'Experience' correspond au nombre d'années d'expériences
- 'Ville' correspond au lieu de travail

Questions :

1) Combien y a t-il d'observations dans ce dataset? Y a t-il des valeurs manquantes?
2) Réaliser l'imputation des valeurs manquantes pour la variable "Experience" avec :
    a. la valeur médiane pour les data scientists
    b. la valeur moyenne pour les data engineers
3) Combien d'années d'expériences ont, en moyenne, chacun des profils : le data scientist, le lead data scientist et le data engineer en moyenne ?¶

4) Faire la représentation graphique de votre choix afin de comparer le nombre moyen d'années d'expériences pour chaque métier
5) Transformer la variable continue 'Experience' en une nouvelle variable catégorielle 'Exp_label' à 4 modalités : débutant, confirmé, avancé et expert. Veuillez expliquer votre choix de la règle de transformation.
6) Quelles sont les 5 technologies les plus utilisées ? Faites un graphique
7) Réaliser une méthode de clustering non supervisée de votre choix pour faire apparaître le nombre de clusters que vous jugerez pertinents. Donnez les caractéristiques de chacun des clusters.
    a. Justifier le nombre de clusters
    b. Justifier la performance de votre algorithme grâce à une métrique.
    c. Interpréter votre résultat.
8) Réaliser la prédiction des métiers manquants dans la base de données par l'algorithme de votre choix
    a. Justifier la performance de votre algorithme grâce à une métrique.
    b. Interpréter votre résultat.