Quelles entreprises correspondent aux métiers de la data aujourd'hui?

assas sabah

Février 2022

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

0.1 Introduction

Le traitement des données nécessite diverses compétences distinctes qui forment des métiers bien spécifiques et complémentaires. En effet le travail des données comporte des débouchées variées telles que data analyst, data scientist et data engineer.
Le métier de Data Scientist est l'un des plus recherchés. Il consiste à analyser les données, et à exploiter le Machine Learning pour établir des modèles prédictifs. Ainsi, les entreprises sont en mesure de prendre de meilleures décisions.
Le Data Engineer ou ingénieur des données a pour rôle de créer des pipelines permettant d'automatiser la collecte de données en provenance de sources variées, de les nettoyer, et de les transférer aux Data Analysts et aux Data Scientists.
Le Data Analyst se charge quant à lui de collecter et d'analyser les données, et partage les résultats de ses travaux sous forme de visualisations ou de rapports.
Les entreprises traitant leurs données peuvent employer l'une ou l'autre de ces fonctions selon des spécificités propres à chacune d'entre elles. Analysons quelles sont les caractéristiques d'entreprise correspondant le mieux au métier de la data qui sont le plus demandés.
Tout d'abord nous allons étudier la situation du marché du travail des métiers de la data récemment en France.
Ensuite, nous nous intéresserons à la relation entre le type de profil recherché et les caractéristiques des entreprises. Enfin, à partir de ces données, nous allons créer des groupes en utilisant une méthode de classification et un algorithme non supervisé donc deux algorithmes seront testés pour voir celui qui permet d'obtenir la classification la plus fine possible. Une fois cette étape réalisée, nous réaliserons une description des groupes obtenus pour faciliter le processus de la recherche d'emplois liés à la data en se rendant directement auprès des entreprises concernées.
Ce projet s'adresse à toutes les entreprises recherchant quel profil correspond à leurs besoins ainsi qu'à chaque personne spécialisée dans le traitement des données qui recherche le type d'entreprise qui irait le mieux avec ses qualités professionnelles.

objectifs :

Dresser un portrait de la situation actuelle du marché du travail concernant les métiers de la data en France ;

Segmenter les entreprises sur la base de données portant sur leur statut ;

Décrire les caractéristiques de nos groupes pour savoir quel type d'entreprise recrute quel type de profil.

0.2 Acquisition des données

Les données utilisées dans cette étude ont été téléchargées à partir de trois sites Web, nous avons d'abord collecté des informations sur les opportunités d'emploi disponibles pour travailler en tant que data analyst, data scientist et data engineer avec le nom de la compagnie, le nombre d'offres, le niveau requis, l'emplacement et le type de travail à partir du site linkedin. Ensuite, nous avons eu recours au site société.fr pour récupérer le nombre d'établissements, le secteur, le nombre d'employés, le lieu de siège et la période depuis la création des entreprises concernées. A la fin, nous avons utilisé le site BFM Verif pour ajouter certaines informations liées à la situation financière de ces compagnies comme le chiffre d'affaires, les charges d'exploitation et les salaires.

0.3 Situation du marché du travail des métiers de la data en France

La situation actuelle du marché du travail des métiers de la data en France a été étudié sur 102 entreprises, tel que il y a plus de 432 offres disponibles sur le site linkedin. Les variables qui ont servi à décrire cette situation sont : le nom de la compagnie, le nombre d'offres, le titre de poste, la région, le secteur, l'emplacement et le type de travail.

0.3.1 Nombre d'offres pour chaque type de profil et de travail

Les résultats obtenus pour chaque type de profil sont présentés dans le tableau ci-dessous.

Table 1 – Nombre d'offres pour chaque type de profil et de travail

Type de profil	Nbre d'offres	Nbre d'entreprises	ratio off/ent	Type de travail	Nbre d'offres
Data Analyst	161	72	2.23	CDD	6
				CDI	121
				Stage	34
Data Scientist	146	52	2.8	CDD	2
				CDI	93
				Stage	51
Data Engineer	125	39	3.2	CDD	1
				CDI	110
				Stage	14
Total	432	102	4.23	//	//

Nous avons constaté que les data analysts sont plus recherchés que les data scientists et les data engineers et les contrats de travail sont souvent de type CDI.
Pour savoir si une entreprise a besoin de beaucoup data-worker ou si un seul suffit, nous avons calculé le rapport entre le nombre d'offres et le nombre entreprises pour chaque type de profil où nous n'avons trouvé qu'une seule entreprise demandant plus de trois data engineers et nous avons également remarqué que ceux-ci n'ont pas beaucoup d'opportunités pour effectuer un stage.

0.3.2 Nombre d'offres par secteur, type et lieu de travail

Le tableau ci-dessous présente les dix premières lignes du nombre d'offres pour chaque secteur, le type et l'emplacement du travail après tri en ordre décroissant.

TABLE 2 – Nombre d'offres par secteur, type et lieu de travail

Secteur	Type de travail	Emplacement	Nbre d'offres
Activité du conseil pour les affaires	CDI	Paris	22
Activité du conseil en systèmes et logiciels	CDI	Paris	21
Activité du conseil en systèmes et logiciels	CDI	Levallois-perret	16
Activité de la production de film	CDI	Paris	12
Activité de la programmation informatique	CDI	Paris	9
Activité de l'édition de logiciels applicatifs	CDI	Paris	9
Activité des autres intermédiations monétaires	CDI	Paris	8
Activité du conseil en systèmes et logiciels	CDI	Lille	7
Activité de l'entretien et réparation	CDI	Lyon	6
Activité du conseil pour les affaires	Stage	Paris	6

D'après les informations ci-dessus, il existe différents secteurs qui ont besoin des métiers de la data en particulier le secteur d'activité du conseil en systèmes et logiciels informatiques qui embauchent le plus en CDI. On trouve aussi que la plupart des offres d'emploi sont disponibles sur Paris.

0.3.3 Conclusion

Grâce à nos données, nous avons constaté qu'il existe des disparités entre les trois métiers de la data en termes d'emploi par entreprise. Pour comprendre cette différence, il peut être intéressant de se pencher sur les états financiers de ces entreprises qui peuvent permettre d'établir des liens avec nos données.

0.4 Lien entre le type de profil recherché et les caractéristiques des entreprises

Les données financières des entreprises ont été collectées sur le site BFM Verif.

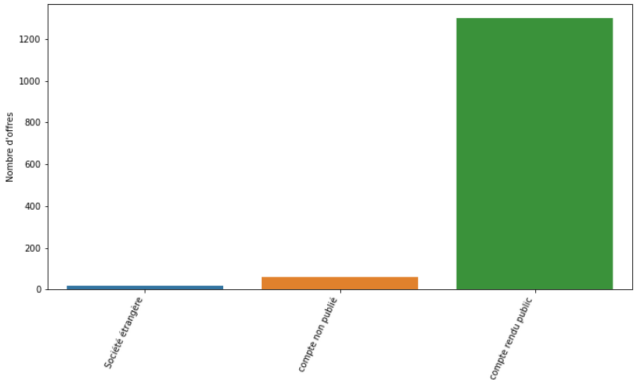


FIGURE 1 – Publication des résultats

Nous avons constaté que certaines entreprises n'affichent pas publiquement leurs comptes, y compris les sociétés étrangères, nous avons donc dû supprimer les offres d'emploi disponibles auprès de ces entreprises et continuer notre étude sur 79 entreprises. Les entreprises ne publient pas leurs comptes chaque année, il a été fait le choix de prendre la moyenne des chiffres déclarés pour chaque entreprise et chaque variable. Ainsi, cela nous permet d'obtenir un nombre de données assez important.

0.4.1 Corrélation entre les variables

Nous avons cherché à voir s'il y avait des corrélations entre les variables.

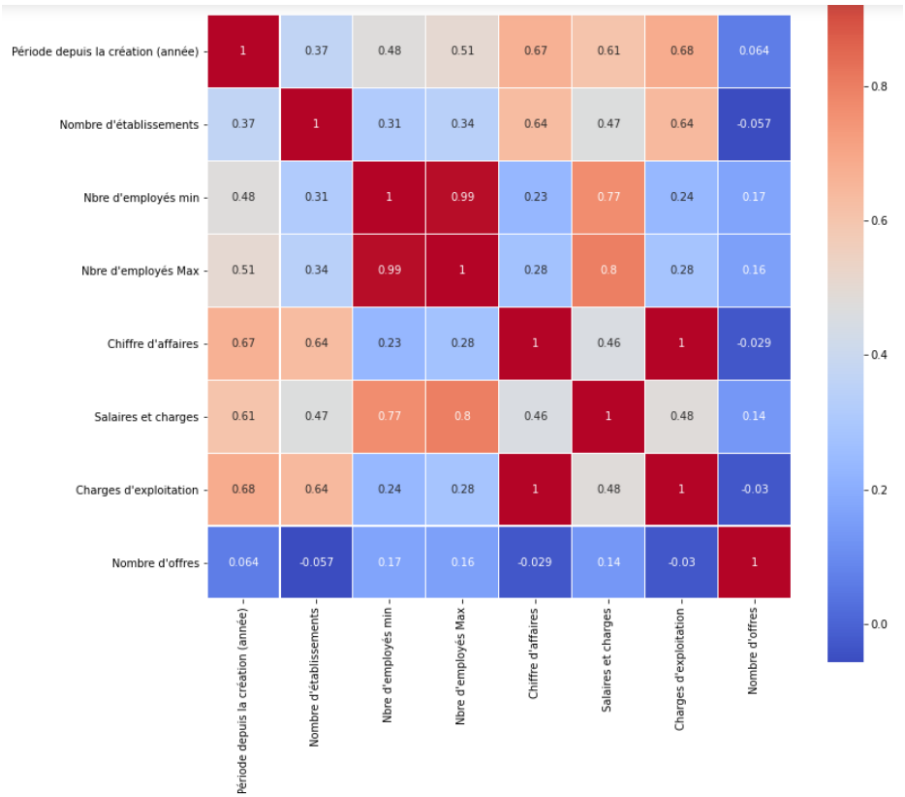


FIGURE 2 – Matrice de corrélations des variables

On remarque qu'il y a une forte corrélation entre le chiffre d'affaires et les charges d'exploitation ainsi qu'entre le nombre d'employés max et le nombre d'employés min, ainsi pour éviter la redondance on a supprimé les variables de charges d'exploitation et le nombre d'employés max.Cette dernière variable a été supprimé du fait l'imprécision du chiffre trop important donné par les entreprises.

0.4.2 Relation entre les variables liées au type de profil

D'après la matrice de corrélations le nombre d'offres n'a aucun lien avec les autres variables donc nous avons décidé d'étudier la corrélation du nombre d'offres de chaque type de profil avec les autres variables.

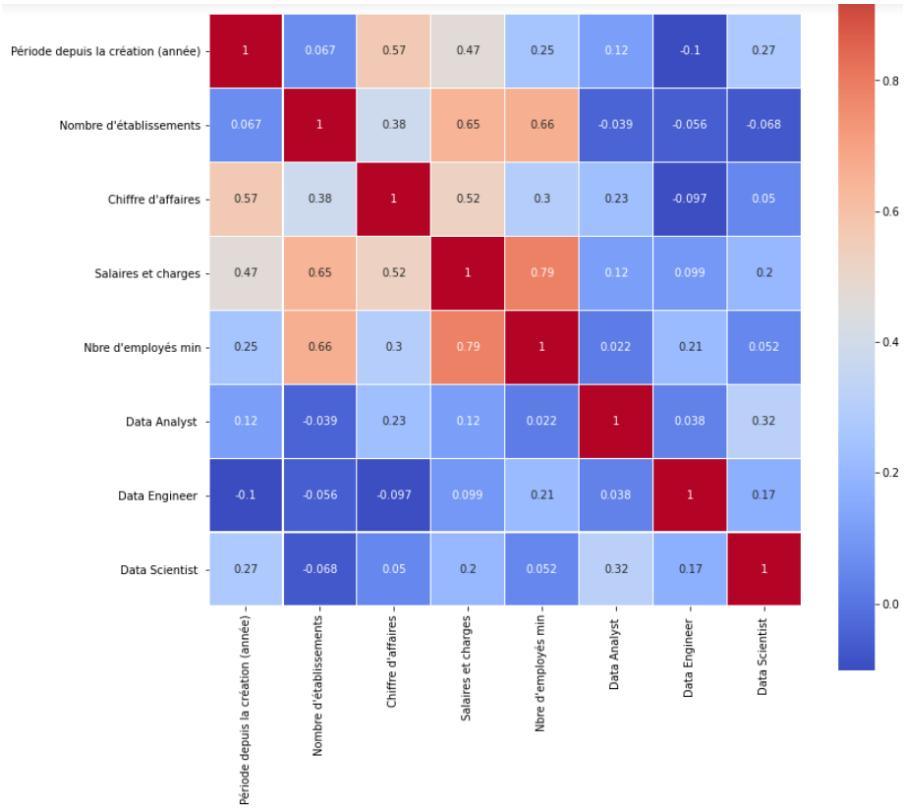


FIGURE 3 – Matrice de corrélations des variables en lien avec le type de profil

A l'aide de la matrice des corrélations, nous observons que finalement les caractéristiques des entreprises ont un impact sur le nombre d'offres pour chaque type de profil, comme le nombre d'offres aux data analysts a un lien avec le chiffre d'affaires.
Pour clarifier davantage ce que nous avons vu précédemment, nous avons tracé les graphiques par paire selon le type de profil.

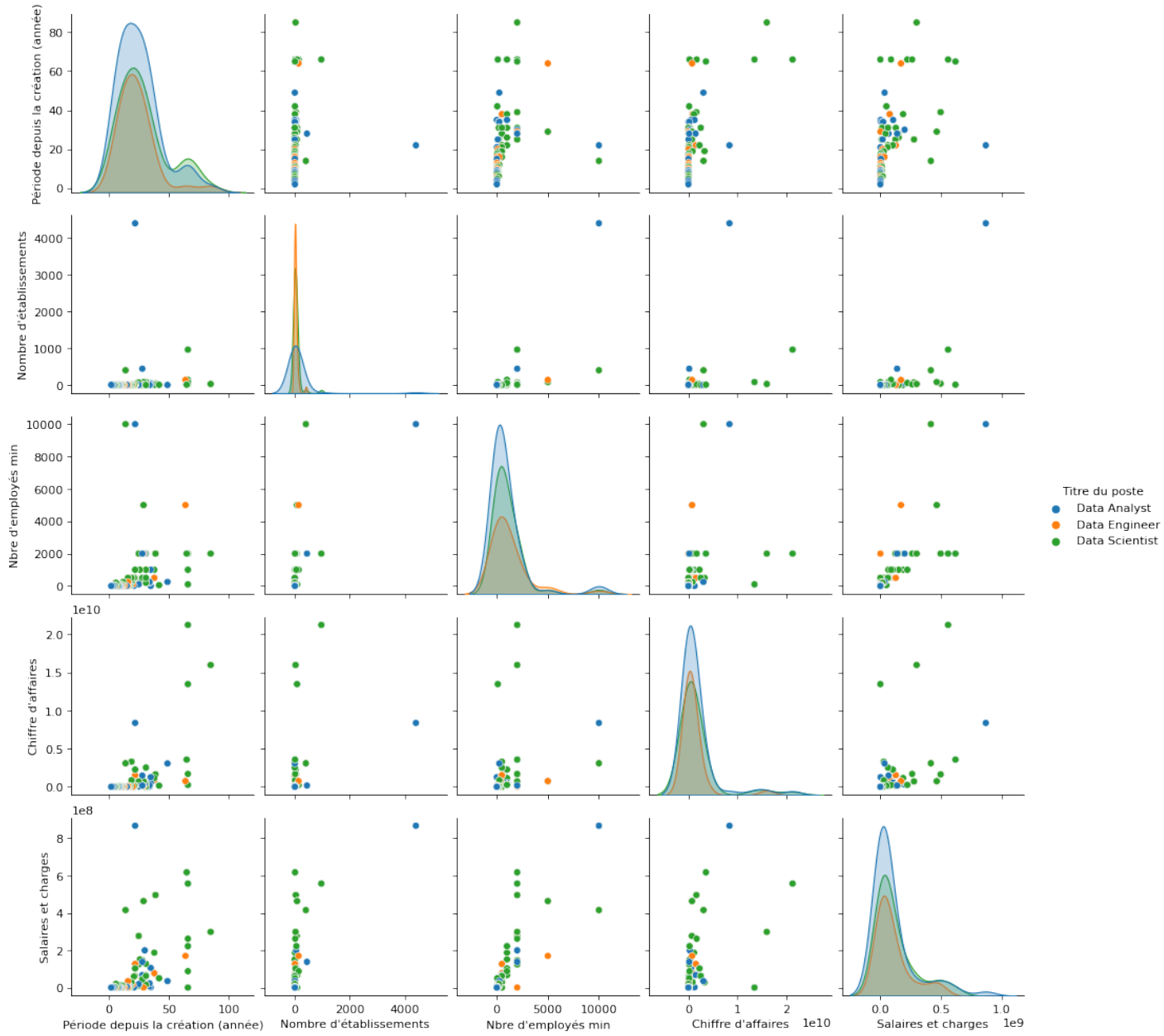


FIGURE 4 – Graphiques par paire selon le type de profil

On voit qu'il n'y a pas beaucoup de différence entre les entreprises qui emploient les trois métiers de la data.

0.4.3 Conclusion

L'analyse des données sur la santé financière des entreprises ne montre pas beaucoup de distinction entre les grandes et les petites entreprises quant à la recherche de data-worker, c'est pourquoi nous avons utilisé des algorithmes de classification qui permettent de créer des groupes d'entreprises avec des particularités proches.

0.5 Modélisation : Utilisation d'algorithmes de clustering

Les techniques de machine learning peuvent permettre de créer des groupes d'entreprises afin de faciliter le processus de la recherche d'emplois liés à la data. Pour répondre à notre problématique, les algorithmes de classification non supervisée vont nous être utiles. Dans la suite de cette étude, nous allons faire appel à l'algorithme du K-means et la classification ascendante hiérarchique. A l'issue du développement de ces algorithmes, on pourra comparer les groupes obtenus pour avoir une idée de celui qui semble plus performant.

0.5.1 L'algorithme du K-means

L'algorithme du K-means permet de créer des clusters en spécifiant le nombre de clusters que l'on souhaite obtenir. Pour déterminer le nombre de clusters, on utilise la méthode du coude qui se base sur les distorsions des variables de notre étude.

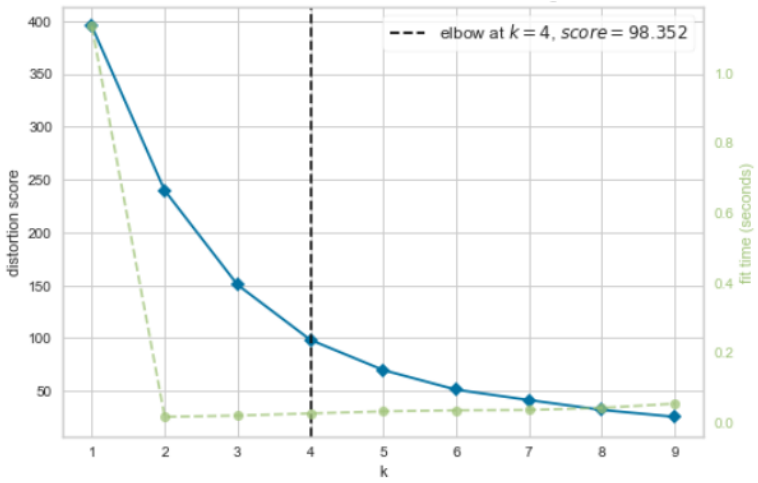


FIGURE 5 – Méthode du coude

Cette méthode nous montre que le nombre idéal de clusters serait 4.

0.5.1.1 K=4

L'algorithme du K-means nous a permis de déterminer quatre clusters que l'on a projeté dans les deux premières de projections d'une analyse en composantes principales.

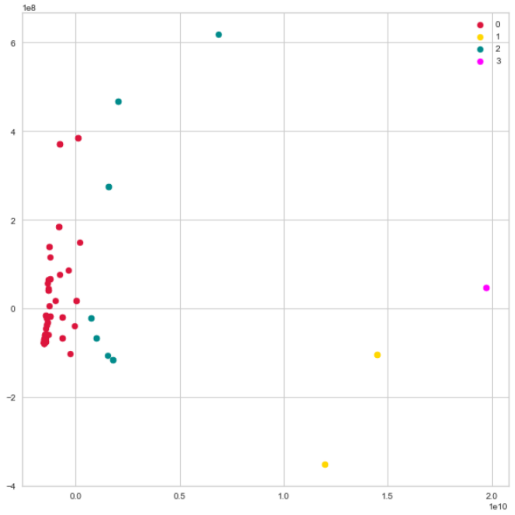


FIGURE 6 – Représentation des clusters du K-Means (avec K = 4)

Nous avons remarqué que le dernier cluster ne contenait qu'un seul individu, nous avons donc voulu approfondir pour connaître le nom et le type de cette entreprise, ce qui incité l'algorithme k-means à la classer seule dans un groupe.

Période depuis la création (année)	Nombre d'établissements	Chiffre d'affaires	Salaires et charges	Nbre d'employés min	clusters	N° Siren	Nom de la compagnie	Région	Secteur	Titre du poste	Nombre d'offres
22	4407	8.376600e+09	867175000.0	10000.0	2	428268023	Groupe Casino	SAINT-ETIENNE	Activité des hypermarchés	Data Analyst	1

FIGURE 7 – Cluster 02

Le groupe casino est né de la cession d'action de l'entreprise casino qui a été créée le 1898 et d'après le site statista cette entreprise possède 407 établissements. Ce qui justifie le résultat trouvé on a donc décidé de compter la période depuis la création de l'entreprise Casino et de corriger le nombre d'établissements.
Ensuite, nous avons effectué encore une fois l'algorithme de K-means.

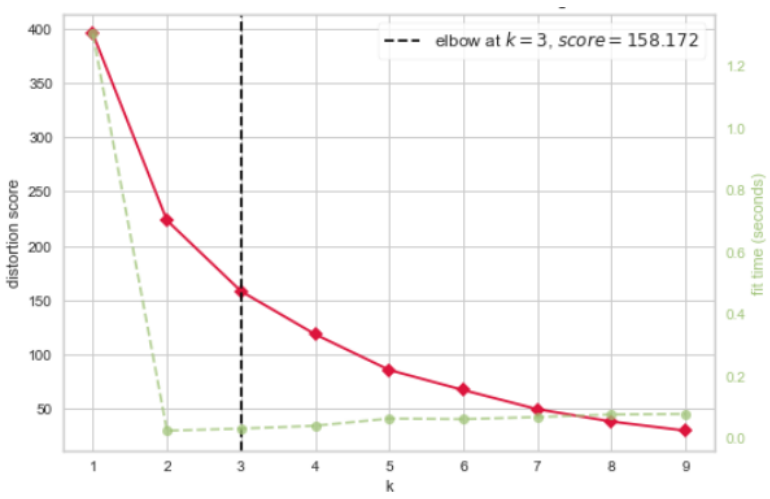


FIGURE 8 – Méthode du coude 2

La méthode du coude nous a permis de voir que le nombre optimal de clusters était de 3.

0.5.1.2 K=3

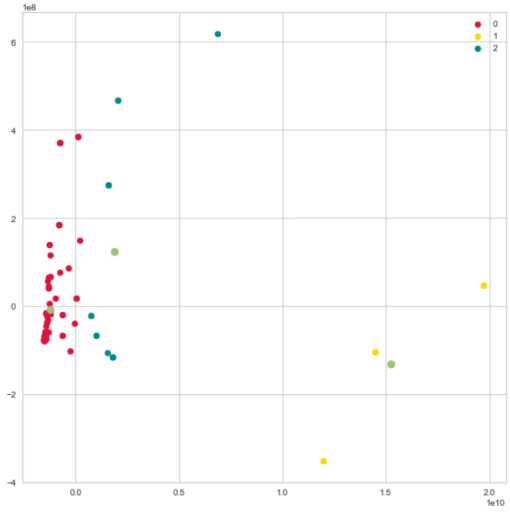


FIGURE 9 – Représentation des clusters du K-Means (avec K = 3)

l'algorithme du K-means nous permet d'obtenir trois groupes plutôt bien équilibrés et bien répartis sur les deux premières composantes de l'ACP.
Pour plus de lisibilité, nous nous sommes penchés sur les scores moyens de nos variables d'intérêt dans ces trois clusters.
Les données en logarithme sont présentées dans le graphique ci-dessous.

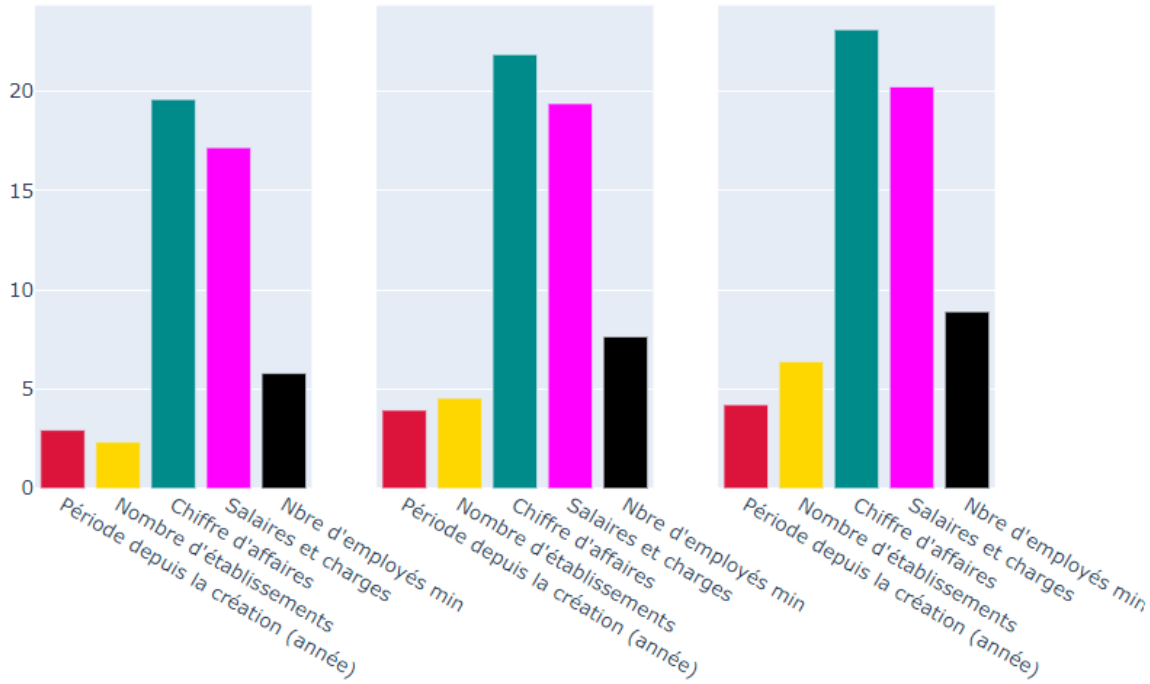
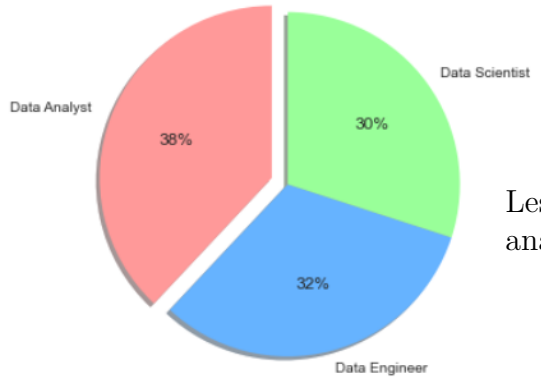


FIGURE 10 – Scores moyens des différentes variables dans nos trois clusters (K-means)

Nous voyons clairement que l'algorithme du K-means a différencié de la manière ou les petites entreprises sont dans le premier groupe, les entreprises de taille moyenne dans le second groupe, tandis que les grandes entreprises dans le dernier groupe.

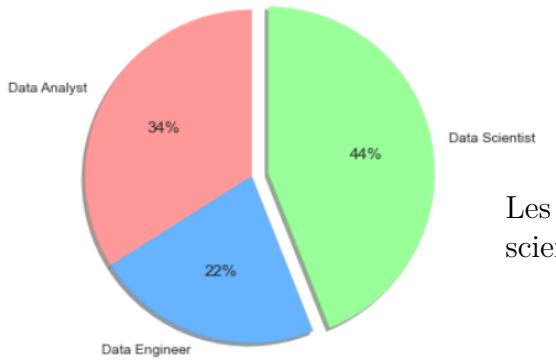
0.5.1.3 Nombre d'offres pour chaque type de profil dans les trois clusters

Après avoir étudié les scores moyens de nos variables dans les différents clusters, nous nous sommes intéressés au nombre d'offres pour chaque type de profil.



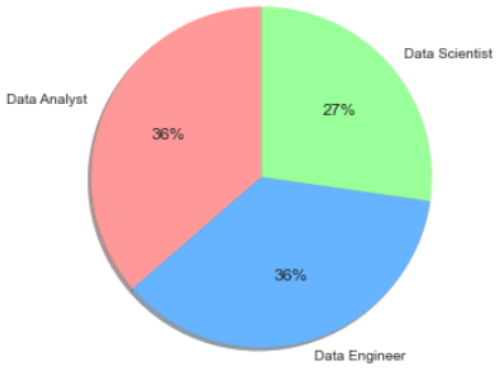
Les entreprises de premier cluster recrutent plus les Data analysts.

FIGURE 11 – Nombre d'offres pour chaque type de profil de cluster 1



Les entreprises de seconde cluster recrutent plus les Data scientists.

FIGURE 12 – Nombre d'offres pour chaque type de profil de cluster 2



Les entreprises de troisième cluster recrutent moins les Data scientists.

FIGURE 13 – Nombre d'offres pour chaque type de profil de cluster 3

0.5.2 La classification ascendante hiérarchique : le dendrogramme

Une autre méthode utilisée pour créer des groupes d' entreprises : la classification ascendante hiérarchique. Pour cette classification, on a divisé notre échantillon en trois clusters. Les résultats sont présentés ci-dessous.

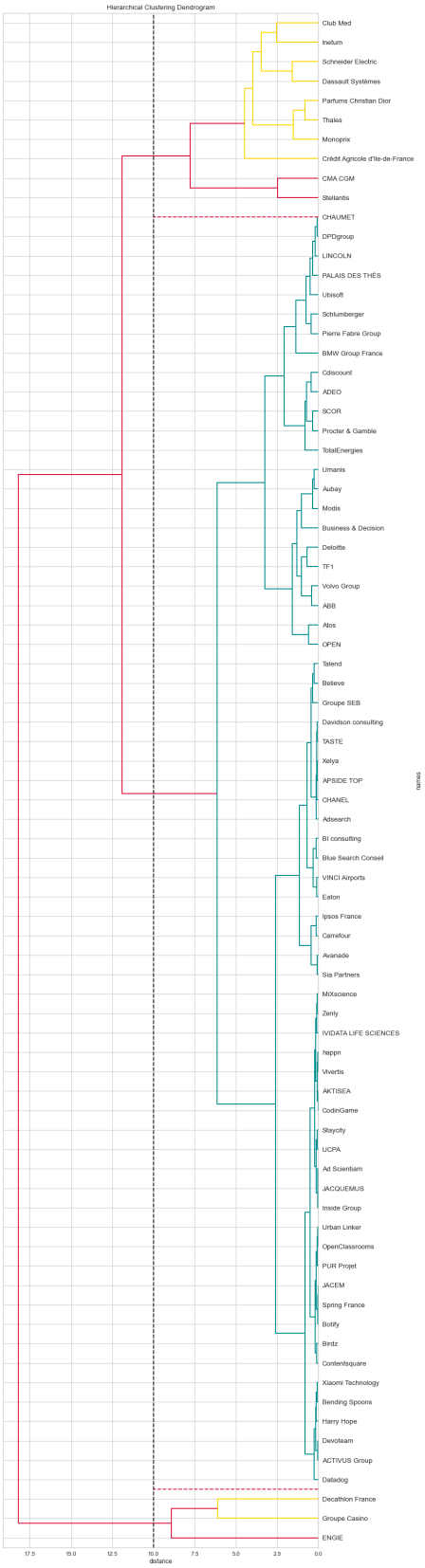


FIGURE 14 – Dendrogramme de la classification ascendante hiérarchique des entreprises avec 3 clusters

Pour voir si cette segmentation est optimale, on va s'intéresser à la distribution des clusters selon les caractéristiques des entreprises.

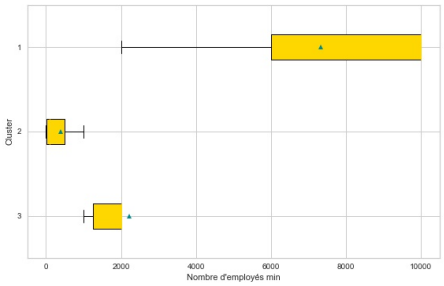


FIGURE 15 – Comparaison des distributions des clusters selon le Nbre d'employés des entreprises

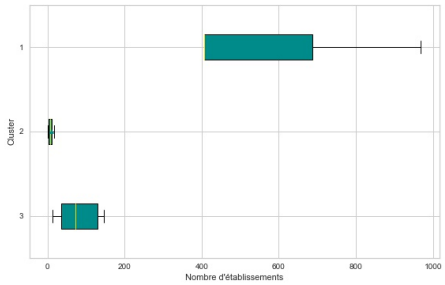


FIGURE 16 – Comparaison des distributions des clusters selon le nombre d'établissements des entreprises

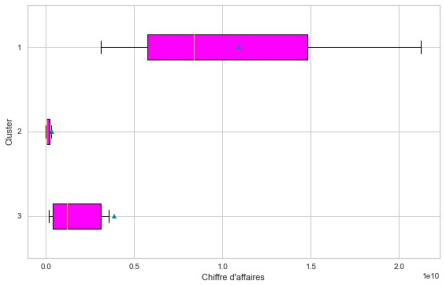


FIGURE 17 – Comparaison des distributions des clusters selon le chiffre d'affaires des entreprises

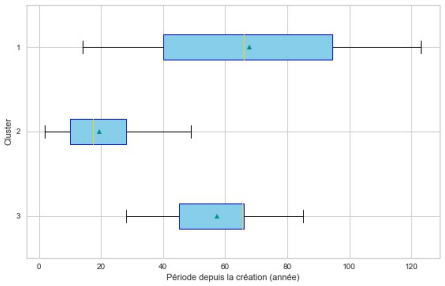


FIGURE 18 – Comparaison des distributions des clusters selon la période depuis la création des entreprises

Ce modèle nous a permis d'obtenir trois clusters différents.

0.5.2.1 Nombre d'offres pour chaque type de profil dans les trois clusters

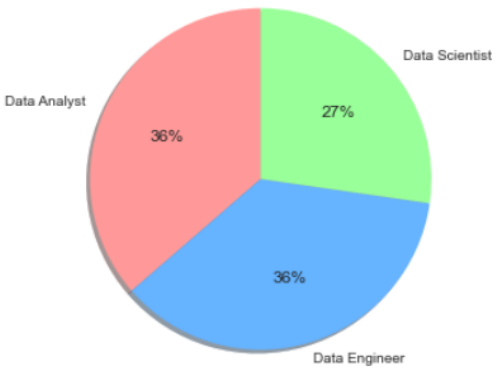


FIGURE 19 – Nombre d'offres pour chaque type de profil de cluster 1

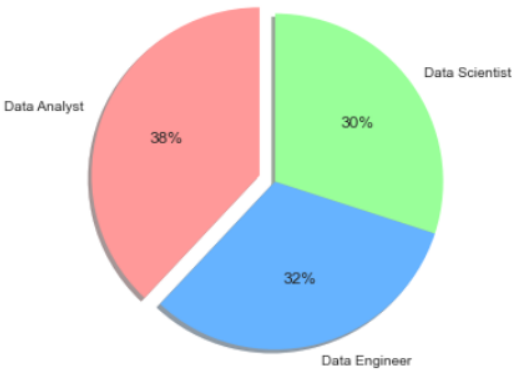


FIGURE 20 – Nombre d'offres pour chaque type de profil de cluster 2

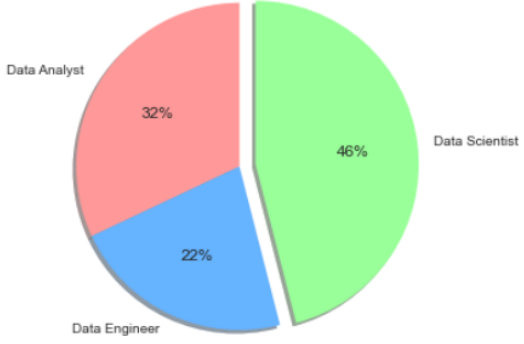


FIGURE 21 – Nombre d'offres pour chaque type de profil de cluster 3

Nous observons que nous avons trouvé presque les mêmes résultats du découpage obtenu avec l'algorithme de K-Means.

0.6 Analyse des résultats

Les trois catégories d'entreprises recherchent des postes dans les trois métiers de la data cependant on note que certaines qualifications sont plus demandées. on suppose que les petites entreprises sont moins regardantes sur la qualification elle veulent surtout une personne capable de gérer l'ensemble du traitement de leurs données, celle ci étant moins quantitatives que pour des entreprise de plus grosses tailles Le data scientist est le plus polyvalent des trois métiers, il maîtrise mieux la collecte des données qu'un analyste et les traite mieux qu'un ingénieur, ce qui nous laisse penser que les entreprises moyennes les préfèrent pour cette capacité d'adaptation. Enfin, les grandes entreprises ayant des besoins plus spécifiques recrutent plus d'analystes et d'ingénieurs, plus compétents dans leur domaines spécifiques, ils entrent mieux dans les grosses structures pouvant répartir le travail plus efficacement.

0.7 Conclusion

Les algorithmes de machine learning ont permis d'établir une segmentation des entreprises pour aider les demandeurs d'emploi dans les métiers de la Data en se focalisant sur des entreprises qui leur correspondent selon leurs particularités.
Il semblerait que la segmentation en trois clusters soit suffisante pour préciser quelles sont les caractéristiques d'entreprise correspondant le mieux au métier de la data.
L'offre des postes entre les Data Analysts, Data Scientists et Data Engineers est relativement équilibrée. La taille de l'entreprise impacte cependant cette répartition. Les petites entreprises embauchent majoritairement des Data Analysts (38%), les moyennes entreprises des Data Scientists (44%), les grandes entreprises des Data Engineers (36%). Nous supposons que l'offre abondante de Data Scientists des moyennes entreprises est lié à leur polyvalence.
Il peut être intéressant de développer ce modèle en y insérant des caractéristiques autres que la santé financière, on pourrait donc continuer ce travail en y ajoutant des variables en lien avec la politique interne de l'entreprise.
.

Appendices

ANNEXE A

TABLEAU DE BORD

Un dashboard résumant toutes les données utilisées a été créée, il est accessible à ce lien :

`https://public.tableau.com/app/profile/assas.sabah/viz/projet8_16432417569340/Tableaudebord1?publish=yes`