POLYTECH SORBONNE

Projet : Analyse de données

Equipe : Yannick ZHANG Léon CHUANG

Table des matières

1 Jeu de données

1.1 Introduction

Le jeu de données "LinkedIn Job Postings (2023 - 2024)" est prélevé à partir Kaggle. Et notre se trouve sur ce lien github.

1.1.1 Contenu

Le jeu de données se compose de 11 fichiers CSV, recueillis à partir de Linkedin, ils contiennent des détails sur des entreprises, des offres d'emploi, des secteurs d'activité, des compétences nécessaires, ainsi que des échelles de rémunération.

```
linkedin
├── company_details
│   ├── companies.csv
│   ├── company_industries.csv
│   ├── company_specialities.csv
│   └── employee_counts.csv
├── job_details
│   ├── benefits.csv
│   ├── job_industries.csv
│   ├── job_skills.csv
│   └── salaries.csv
├── job_postings.csv
└── maps
    ├── industries.csv
    └── skills.csv
```

1.1.2 Problèmatique

Ce jeu de données contient diverses dimensions d'une poste de travail et peut être utilisé pour effectuer des analyses de données sans supervision afin d'en lire des significations plus intuitives. Depuis la pandémie, de nombreux emplois commencent à permettre le télétravail. Au début, la plupart de ces emplois concernaient l'industrie technologique et les développeurs. Cependant, de plus en plus de postes rejoignent également la ligne du travail à distance. Par exemple, les restaurants de New York, disposent d'un personnel de comptoir de chat vidéo à distance pour réduire les coûts de main-d'œuvre. Nous utilisons cette information (la possibilité de télétravail) comme label pour mieux comprendre quels emplois sont les plus susceptibles d'être effectués à distance.

1.2 Prétraitement

Afin de réaliser une analyse, nous avons modifié les données en suivant plusieurs étapes de prétraitement.

Nous avons commencé par filtrer le type de travail pour ne garder que les emplois à temps plein, éliminant ainsi les autres types d'emplois qui ne intéressaient pas pour notre analyse. Ensuite, nous avons aussi filtré les données par période de paie pour ne conserver que les offres avec une période de paie mensuelle ou annuelle. Les autres périodes de paie, comme les paiements hebdomadaires ou journaliers, ont été exclus pour assurer la cohérence des données et faciliter la comparaison des salaires. Puis nous avons choisi de conservé uniquement les lignes où au moins une information sur le salaire était présente. En effet, cela nous permettait de nous concentrer sur les offres d'emploi avec des données de rémunération

disponibles Pour simplifier ces informations, nous avons utilisé la médiane lorsqu'elle était disponible, car elle est moins sensible aux valeurs extrêmes que la moyenne. Et lorsque la médiane n'était pas disponible, nous l'avons calculé manuellement avec avec le salaire maximum et le salaire minimum. Mais si seulement l'un des deux était disponible nous avons utilisé la valeur directement. Enfin, nous avons converti les salaires mensuels en salaires annuels afin d'uniformiser les données et de permettre une comparaison plus facile entre les offres d'emploi.

Pour la suite des traitement nous avons remplacé les valeurs manquantes pour certaines variables numériques par 0, telles que le nombre de candidatures (*applies*), le nombre de vues (*views*) et l'autorisation du télétravail (*remote_allowed*). En effet, ce choix est basé sur des hypothèses raisonnables : par exemple, si le nombre de candidatures ou de vues n'est pas spécifié, il est probable qu'aucune candidature n'a été reçue ou que l'offre n'a pas été vue. De même, si l'autorisation du télétravail n'est pas mentionnée, cela peut être interprété par défaut comme non-autorisé

Quant aux variables temporelles, qui étaient toutes en timestamps (un format de données représentant le nombre de secondes écoulées depuis le 1er janvier 1970), les timestamps ont été convertis en dates lisibles afin de faciliter l'analyse. Nous avons également calculé la durée entre la date de publication et la date d'expiration des offres pour avoir une idée de la durée pendant laquelle les offres restent ouvertes.

Nous avons aussi supprimé des colonnes non nécessaires pour notre analyse, telles que les descriptions des postes, les URLs, les informations sur la devise (*currency*), le type de compensation (*compensation_type*), les types de travail formatés (*formatted_work_type*) etc. Car ces colonnes ont été jugées non pertinentes pour notre analyse ou redondantes avec d'autres informations.

Enfin, en raison du grand nombre de lieux différents dans la colonne (*location*), nous les avons regroupés d'abord regroupés districts puis aussi en régions. Cette catégorisation a permis d'analyser les tendances géographiques de manière plus structurée.

Et la répartition des districts a été effectuée selon l'image suivante :

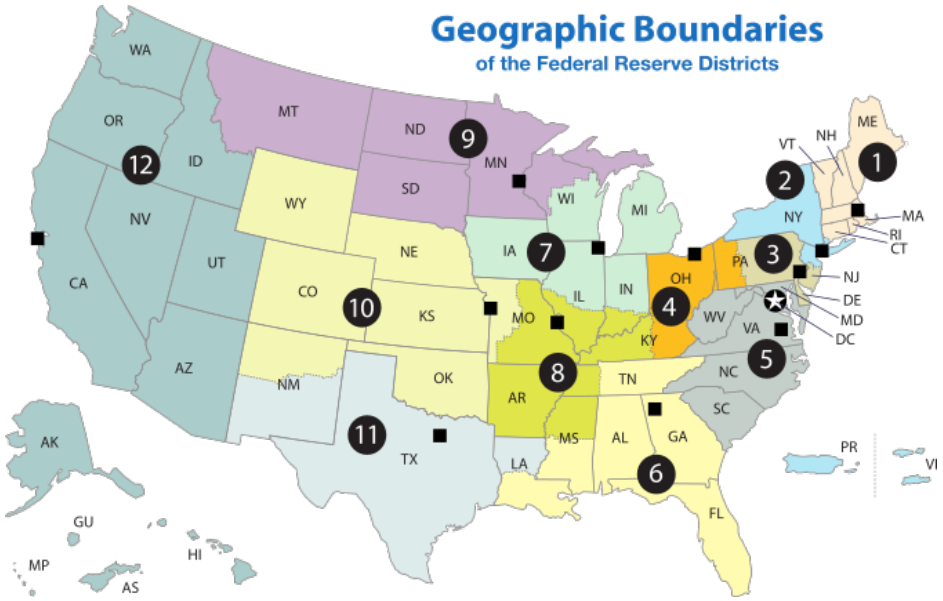


FIGURE 1 – Répartition des district

Quant à la répartition des régions selon cette image :

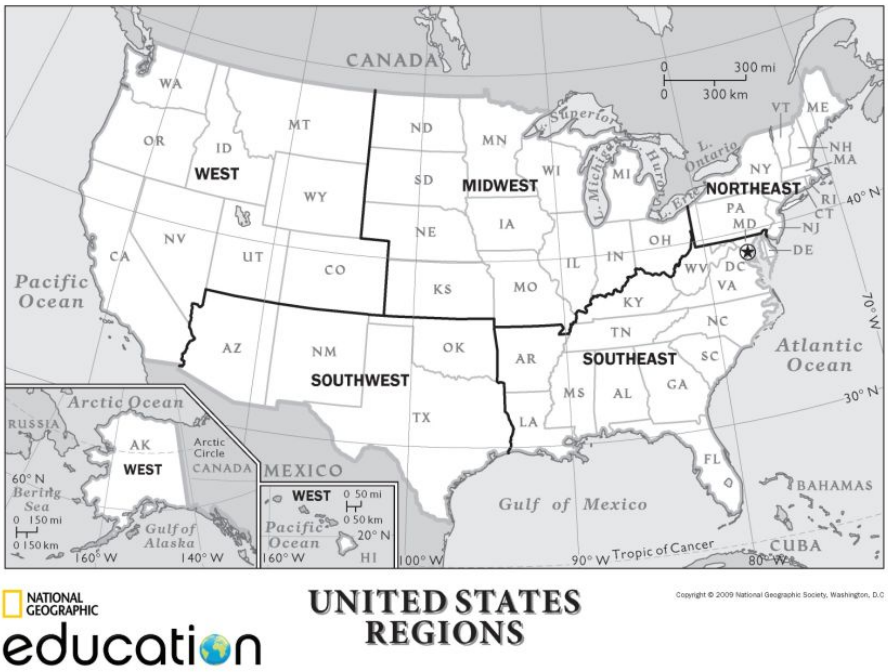


FIGURE 2 – Répartition des region

En conclusion, toutes ces étapes de prétraitement ont permis d'obtenir un jeu de données propre, cohérent et prêt pour des analyses plus approfondies.

2 Analyse

2.1 ACP

L'objectif ici était de comprendre les relations entre les différentes variables quantitatives et qualitatives, notamment le nombre de candidatures, le nombre de vues, le salaire, le niveau d'expérience requis, et la durée des annonces.

Avant de lancer l'ACP, les variables quantitatives ont été sélectionnées et normalisées (centrées et réduites) pour assurer une échelle comparable. Puis les variables qualitatives ont été intégrées à l'analyse pour évaluer leur influence sur les composantes principales.

Les variables quantitatives utilisées sont les suivantes :

— *applies* : Nombre de candidatures

— *views* : Nombre de vues

— *salary* : Salaire

— *formatted_experience_level* : Niveau d'expérience requis (converti en échelle de 0 à 5)

— *duration_days* : Durée des annonces en jours

Les variables qualitatives utilisées sont les suivantes :

— *district* : District géographique de l'offre

— *region* : Région géographique de l'offre

— *remote_allowed* : Indication si le télétravail est permis

— *sponsored* : Indication si l'offre est sponsorisée

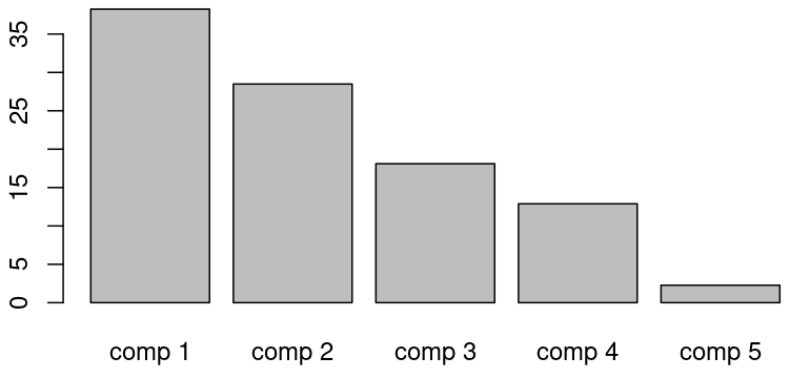


FIGURE 3 – Pourcentage de la variance expliqué par axe

Pour l'ACP, nous avons d'abord étudié les valeurs propres afin de décider combien de dimensions garder. Et il semble approprié de considérer les trois premières dimensions (comp.1, comp.2, comp.3) pour une analyse détaillée car elles cumulent plus de 80% de la variance expliquée.

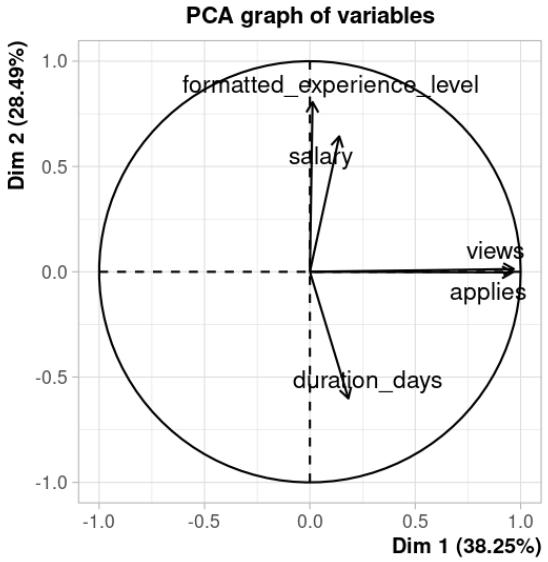


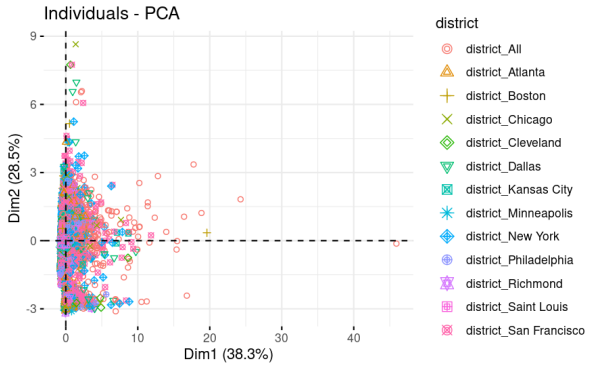
FIGURE 4 – Cercle des corrélations de l'ACP

On peut observer sur le cercle de corrélation que la première composante principale (Dim 1) est influencée par les variables *views* et *applies*. D'ailleurs *views* et *applies* sont pratiquement alignés le long de l'axe 1. Ce qui indique aussi une forte corrélation positive entre ces deux variables et suggère que Dim 1 peut être interprétée comme un indicateur de la popularité des offres d'emploi.

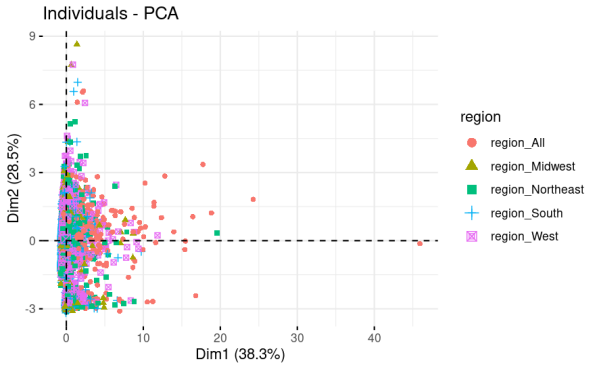
Quant à l'axe 2 (Dim 2), il est principalement influencé par les variables *salary* et *formatted_experience_level*, suggérant une relation entre la rémunération et le niveau de responsabilité des postes. De plus, la présence de *duration_days* sur cette dimension pourrait indiquer que la durée pendant laquelle un poste reste ouvert peut être liée à des niveaux de salaire ou d'expérience plus élevés, possiblement en raison de la difficulté

à trouver des candidats qualifiés pour des rôles plus spécialisés ou exigeants.

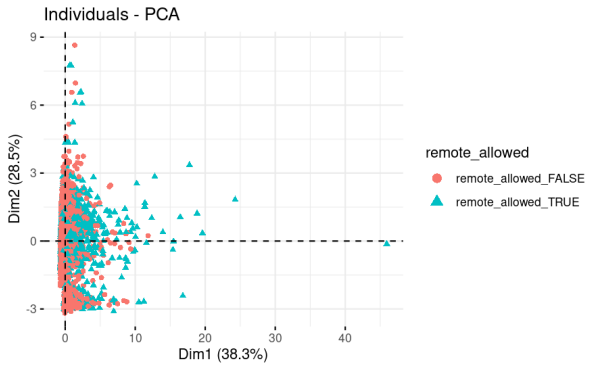
Ensuite, pour comprendre l'impact des variables qualitatives sur les composantes principales, nous avons habillé les individus (offres d'emploi) selon ces variables. Les graphiques suivants montrent la distribution des individus selon les dimensions principales, en fonction des variables qualitatives : *district*, *region*, *remote_allowed*, et *sponsored*.



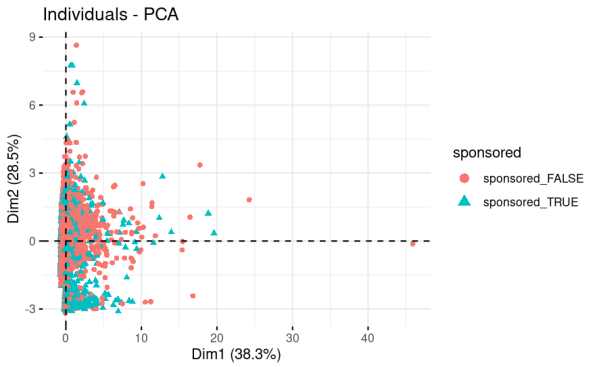
(a) Habillage par District



(b) Habillage par Région



(c) Habillage par Télétravail



(d) Habillage par Sponsoring

FIGURE 5 – Habillage des individus selon les variables qualitatives

Pour les quatre graphes, les points sont plutôt bien dispersés le long de l'axe de Dim 2, ce qui signifie que les variables qualitatives présentent une variabilité similaire sur cette dimension.

Interprétation des Graphes

- **Graphe 1 (District)** : La dispersion le long de l'axe 2 peut indiquer une diversité dans les caractéristiques des offres d'emploi selon les districts. Cependant, on n'observe pas de regroupement net. Cela signifie que les différences entre les districts, en termes de visibilité des offres d'emploi et de leurs caractéristiques comme le salaire ou le niveau d'expérience, ne sont pas suffisamment marquées. Les offres sont relativement uniformes à travers les différents districts.

- **Graphe 2 (Région)** : C'est un peu la même chose que l'analyse avec les districts, mais on peut observer que les offres d'emploi classées sous la catégorie "All" semblent se distinguer légèrement des autres en étant positionnées plus à droite sur la première dimension (Dim 1). Cela pourrait signifier que les offres ayant moins de contraintes (ici de localisation) pourraient potentiellement attirer plus de vues ou de candidatures.

- **Graphe 3 (Télétravail)** :

```
FALSE  TRUE
 6300  1439
```

On peut déjà remarquer qu'il y a très peu d'offres qui autorisent le travail à distance (remote_allowed_TRUE). De plus, on voit que les offres d'emploi permettant le travail à distance semblent se regrouper sur l'axe central, tandis que celles qui ne le permettent pas (remote_allowed_FALSE) sont plus dispersées le long de la Dim 1. Cette disposition suggère que les offres autorisant le travail à distance pourraient être plus populaires ou plus visibles (plus de vues et de candidatures), possiblement dues à une demande croissante pour la flexibilité dans les conditions de travail.

- **Graphe 4 (Sponsoring)** :

```
FALSE  TRUE
 5571  2168
```

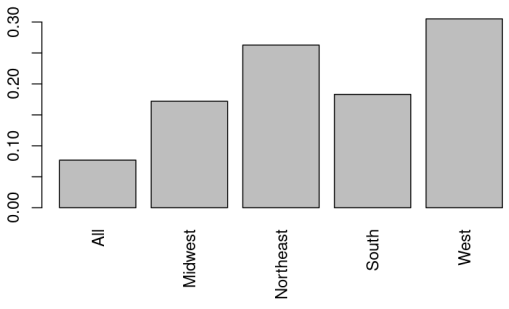
Même s'il y a un peu plus d'annonces qui sont sponsorisées comparé au travail à distance, cela reste très peu. Il semble aussi que le sponsoring des offres d'emploi n'a pas d'impact significatif sur le nombre de vues ou de candidatures reçues. En effet, nous pouvons observer que les points pour les offres sponsorisées (sponsored_TRUE) et non sponsorisées (sponsored_FALSE) sont dispersés de manière quasi similaire le long de la Dim 1.

En conclusion, l'ACP a révélé des relations intéressantes sur la popularité et les caractéristiques des offres d'emploi sur LinkedIn. La popularité des offres est fortement corrélée au nombre de vues et de candidatures, tandis que la qualité des postes, en termes de salaire et de niveau d'expérience, influence également la durée pendant laquelle les postes restent ouverts. Les analyses ont aussi montré que les variables qualitatives telles que le district, la région, le télétravail et le sponsoring présentent des variabilités intéressantes mais parfois limitées.

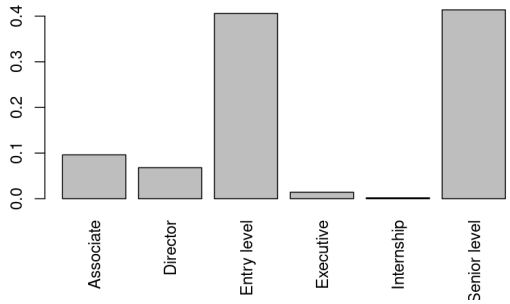
2.2 AFC

Analyse avant AFC

Afin de comprendre les relations entre le niveau d'expérience requis pour les postes et les régions géographiques. Nous avons d'abord examiné la relation entre le niveau d'expérience et la région. Pour ce faire, nous avons calculé une table de contingence entre le niveau d'expérience et la région, puis visualisé les proportions d'offres d'emploi par région et par niveau d'expérience.



(a) Barplot par region



(b) Barplot par level

Le premier barplot montre que la région Ouest et Nord-Est regroupent plus de 50% des offres (Figure **??**). Cela peut s'expliquer par la forte concentration d'industries de haute technologie dans des zones

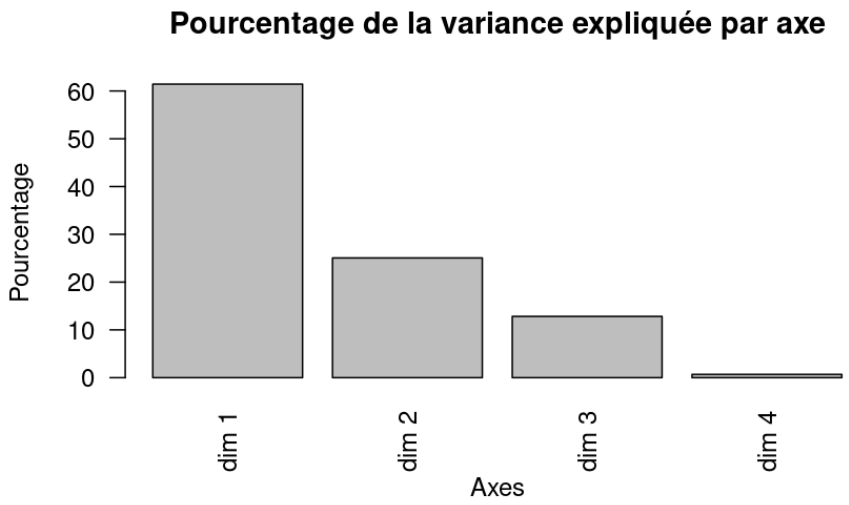
telles que la Silicon Valley et Seattle (pour l'Ouest), ainsi que par la présence de grands centres financiers et éducatifs comme New York et Boston (Nord-Est). En effet, cette région est connu pour ses industries financières, ses institutions académiques et de ses secteurs de la santé et de la biotechnologie, qui attirent un grand nombre de professionnels.

Puis sur le second barplot on observe que les postes de niveau "Entry level" et "Senior level" sont les plus fréquents, ce qui peut indiquer une demande élevée pour des positions de début de carrière et des rôles plus avancés.

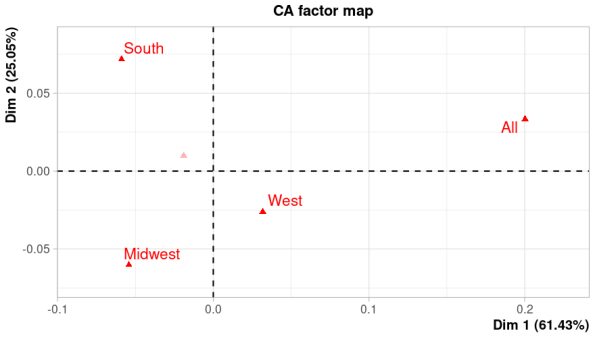
Nous avons ensuite effectué un test d'indépendance (test Chi-2) pour évaluer la relation entre ces deux variables qualitatives. L'hypothèse nulle (H0) de ce test est que le niveau d'expérience requis pour les postes et les régions géographiques sont indépendants. Cependant, les résultats du test (p-value = 1.119e-05) nous ont conduits à rejeter cette hypothèse nulle. En effet, avec un p-value aussi faible, nous avons une preuve suffisante pour conclure que le niveau d'expérience requis et la région géographique des offres d'emploi ne sont pas indépendants, avec une haute intervalle de confiance. Il est donc pertinent de réaliser une AFC pour explorer plus en détail le lien entre ces 2 variables.

Analyse de l'AFC

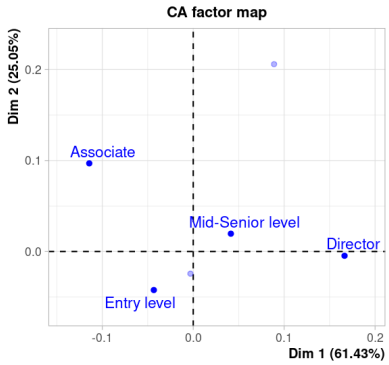
Tout comme ce qu'on a fait pour l'AFC, pour l'AFC nous avons aussi éudiée les valeurs propres et grâce aux barplot ci-dessous on a vu que les deux premiers axes représentaient plus de 80% de la variance totale. Nous avons donc décidé de retenir ces deux axes.



Ensuite, nous avons étudié séparément les contributions des deux variables (régions et niveaux d'expérience) sur les deux axes retenus



(a) Graphique des régions bien représentées



(b) Graphique des niveaux d'expérience bien représentés

Pour les régions, les contributions les plus élevées à l'axe 1 proviennent des régions "All" et "South". All" se trouve tout à droite, représentant des offres d'emploi généralistes avec une portée nationale, tandis que "South" est tout à gauche, indiquant alors des offres spécifiques à cette région, qui sont peut-être liées à des industries régionales telles que l'agriculture ou la fabrication. Cette distinction entre généralistes et spécifiques suggère une différenciation dans les compétences recherchées et les besoins du marché du travail régional.

Sur le second axes, les 2 contributions les plus élevés sont ceux des régions "South" et "Midwest". Le "South", situé plus haut sur le graphe, se distingue par son implication dans des secteurs innovants comme l'énergie et l'aérospatiale, surtout dans des États comme le Texas et la Floride. En revanche, le "Midwest", plus bas sur l'axe, se concentre sur des industries traditionnelles comme la fabrication et l'agriculture, notamment dans des États comme l'Ohio et le Michigan. On peut donc supposer que cet axe indique le degré d'innovation ou de technologie des offres d'emploi dans ces régions, avec le "South" privilégiant des secteurs plus novateurs et le "Midwest" se concentrant davantage sur des industries traditionnelles.

Pour les niveaux d'expérience, les contributions les plus élevées à l'axe 1 proviennent des niveaux "Director" et "Associate". "Director" est positionné à tout à droite tandis que "Associate" tout à gauche. On peux donc supposer que l'axe 1 capte probablement les différences en termes de responsabilité ou de complexité des postes. Les postes de direction impliquent des responsabilités plus élevées, tandis que les postes d'associé sont généralement de niveau inférieur.

Puis pour l'axe 2, c'est "Associate" qui a la contribution la plus significative et est positionné légèrement au-dessus de l'axe horizontal. "Entry level" apparaît aussi avec une contribution notable sur cet axe, mais est positionné un peu plus en bas. Ces deux types de niveaux peuvent partager une similarité dans la compréhension générale du travail, car ils sont tous deux souvent associés à des rôles d'initiation où les employés apprennent les bases de leurs fonctions et de l'organisation. Cependant, ils peuvent différer en termes d'autonomie, niveau de responsabilité ou de la portée des tâches. Les employés en début de carrière peuvent avoir un niveau d'autonomie plus faible et nécessiter une supervision dans leurs tâches, tandis que les associés peuvent être plus autonomes. L'axe 2 pourrait donc refléter les variations dans l'autonomie et la responsabilité au sein des postes de début de carrière.

Interprétation globale

Les résultats montrent des relations intéressantes entre les niveaux d'expérience et les régions. Par exemple, "Entry level" est très proche avec la région "Midwest" ce qui suggère que cette zone a une demande plus importante pour des postes de niveau débutant, ce qui est semble assez cohérent. En effet, ce dernier se concentre plus dans les industries traditionnelles, et ces secteurs offrent souvent de nombreuses opportunités d'emploi ne nécessitant pas une expérience préalable.

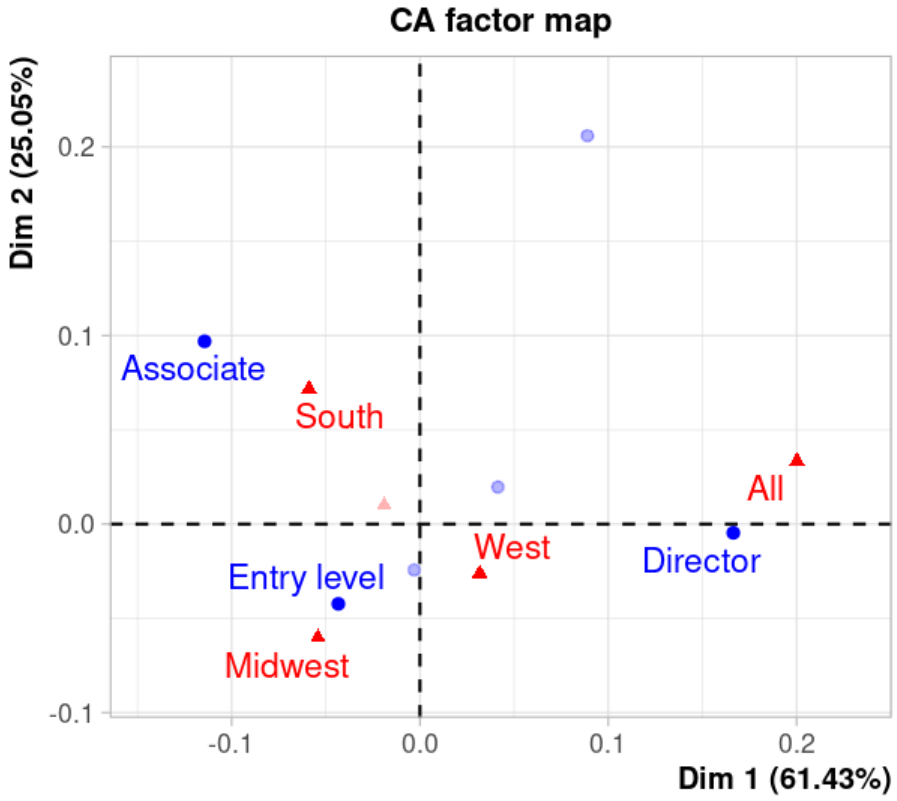


FIGURE 8 – Graphique des niveaux d'expérience et des régions bien représentés

De même, Associate" est très proche de "South". Cela suggère que la région "South" pourrait avoir une demande significative pour des postes d'associé, ce qui est due à ses activités plus orientés dans les secteurs innovants tels que l'énergie et l'aérospatiale. Dans ces domaines, les postes d'associé exigent généralement un peu plus d'expérience et de compétences que les postes de niveau "Entry level", car ils sont souvent impliqués dans des activités de recherche et développement plus avancées.

Enfin, les postes de "Director" sont proches de "All", suggérant une demande dispersée à travers tout le pays pour ces positions à haute responsabilité. Ce qui signifie que les compétences de leadership et de gestion sont valorisées de manière uniforme à travers différentes régions, reflétant la nécessité pour les entreprises, quelle que soit leur localisation, de recruter des cadres supérieurs capables de diriger des équipes et de piloter des stratégies d'affaires efficacement.

2.3 Clustering

Pour cette analyse, nous avons utilisé deux méthodes principales : la classification ascendante hiérarchique (CAH) et le k-means. Ces méthodes nous ont permis de regrouper les offres d'emploi en fonction de leurs caractéristiques quantitatives.

Les données ont été préparées en sélectionnant et en normalisant (centrant et réduisant) les variables quantitatives suivantes :

— *applies* : Nombre de candidatures

— *views* : Nombre de vues

— *salary* : Salaire

— *formatted_experience_level* : Niveau d'expérience requis (converti en échelle de 0 à 5)

— *duration_days* : Durée des annonces en jours

Les distances entre les observations ont été calculées à partir des données normalisées.

Nous avons d'abord déterminé le nombre optimal de clusters en examinant l'inertie intra-classe pour différents nombres de clusters. Le graphique suivant montre la proportion de l'inertie intra-classe en fonction du nombre de clusters.

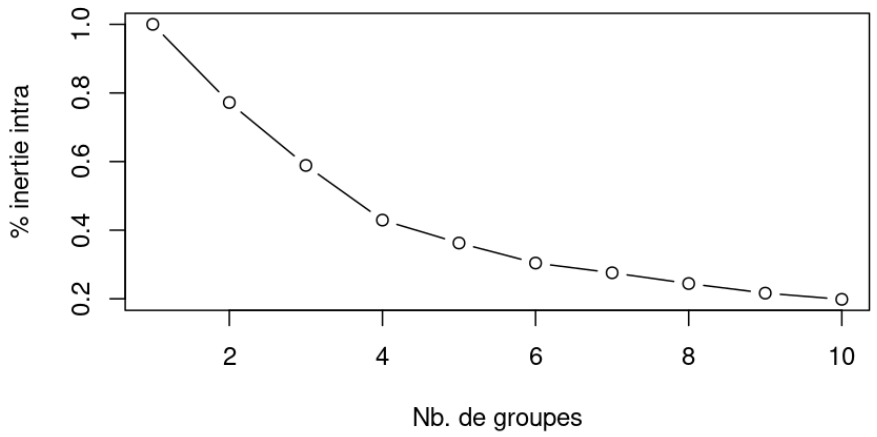
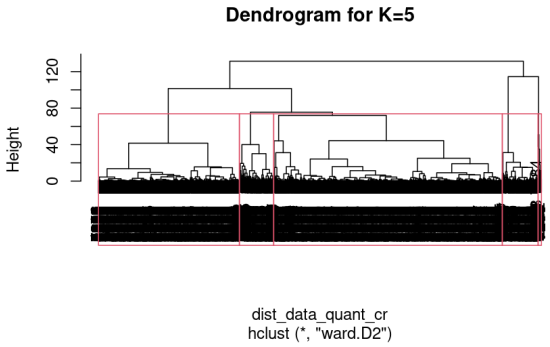


FIGURE 9 – Inertie intra-classe en fonction du nombre de clusters

À partir de K = 5 ou 6 clusters, l'ajout d'un groupe supplémentaire ne diminue pas significativement la part d'inertie intra-classe. Nous avons donc décidé de procéder avec K = 5. En effet, en observant les dendrogramme, la différence entre K = 5 et K = 6 est quasi négligeable. Ainsi nous avons appliqué les méthodes de clustering CAH et k-means avec K = 5.



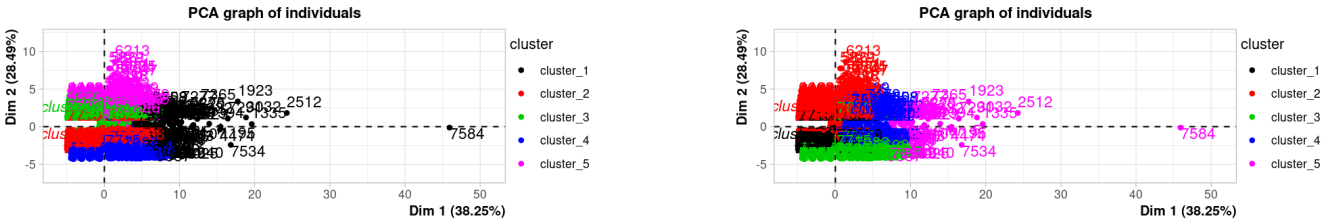
(a) Dendrogramme CAH pour K=5

Group <int>	applies <dbl>	views <dbl>	salary <dbl>	formatted_experience_level <dbl>	duration_days <dbl>
1	3.389273	24.62129	97145.59	1.065421	31.82706
2	5.369554	33.30145	127291.18	2.997747	31.77805
3	21.046326	108.50799	97411.37	1.217252	180.13992
4	70.088333	317.58833	121329.53	2.383333	31.19502
5	381.655172	1246.39655	124139.83	1.948276	63.83501

(b) Clusters K-means pour K=5

FIGURE 10 – Clustering pour K=5

Nous avons visualisé les clusters à l'aide de l'ACP pour les données des clusters obtenus avec les deux méthodes (CAH et k-means).



(a) ACP K-means pour K=5

(b) ACP CAH pour K=5

FIGURE 11 – Visualisation des clusters par ACP

Et la table de comparaison ci-dessous montre la correspondance entre les clusters obtenus par CAH et ceux obtenus par K-means. On observe une bonne concordance entre les deux méthodes, bien que quelques différences subsistent dans les affectations des observations aux clusters.

CAH / K-means	1	2	3	4	5
1	0	2402	1	1	57
2	0	441	2752	0	801
3	8	0	0	614	4
4	76	234	180	0	110
5	58	0	0	0	0

TABLE 1 – Comparaison des clusters CAH et K-means pour K=5

Il semble que ces regroupements ne soient pas conformes aux variables qualitatives existantes. D'après les résultats du tableau (b), ce regroupement semble concerner des emplois de popularité différentes sur LinkedIn. Le résultat de ces groupes peut nécessiter une analyse plus approfondie pour les interpréter

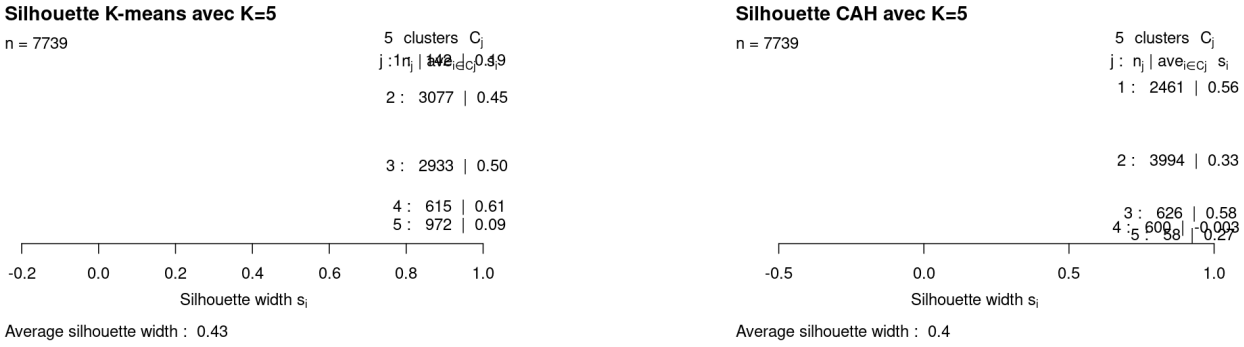


FIGURE 12 – Silhouette pour K-means et CAH avec k = 5

Et l'analyse des silhouettes pour les méthodes K-means et CAH avec K=5 révèle des différences significatives dans la qualité des clusters. Pour K-means, la largeur moyenne de silhouette est de 0.43, indiquant une cohésion et une séparation modérées des clusters. Notamment, le Cluster 4 montre une

bonne cohésion avec une silhouette moyenne de 0.61, tandis que le Cluster 5, avec une silhouette moyenne de 0.09, est mal défini.

Pour CAH, la largeur moyenne de silhouette est légèrement inférieure à 0.4, avec le Cluster 1 (0.56) et le Cluster 3 (0.58) montrant une bonne définition. Cependant, le Cluster 4 (-0.03) indique une mauvaise assignation des observations.

En conclusion, K-means montre une meilleure performance globale par rapport à CAH, bien que les deux méthodes rencontrent des difficultés similaires pour certains clusters, en particulier le Cluster 5 pour K-means et le Cluster 4 pour CAH.

2.4 Analyse supervisée

Pour cette analyse, nous avons utilisé plusieurs modèles d'apprentissage supervisé afin de prédire la variable binaire *remote_allowed* (télétravail autorisé ou non) à partir des autres variables disponibles dans le jeu de données prétraité. Les méthodes utilisées sont :

- Analyse Discriminante Linéaire (LDA)
- Analyse Discriminante Quadratique (QDA)
- Arbre de Décision (CART)
- Forêt Aléatoire (Random Forest)
- AdaBoost
- Régression Logistique Lasso

Nous avons d'abord importé les données et converti les variables nécessaires en facteurs. Les variables pertinentes pour cette analyse incluent le nombre de candidatures (*applies*), le nombre de vues (*views*), le salaire (*salary*), la durée des annonces (*duration_days*), le niveau d'expérience (*formatted_experience_level*), la région (*region*), le type d'entretien (*application_type*) et si l'annonce est sponsorisée (*sponsored*).

Ensuite, nous avons divisé les données en deux ensembles : un ensemble d'entraînement (80% des données) et un ensemble de test (20% des données). Cependant il y'avait un déséquilibre dans les données et pour y remédier nous avons utilisé la méthode SMOTE pour équilibrer l'ensemble d'entraînement.

Performances des Modèles

Nous avons comparé les modèles en termes de précision et d'AUC (Area Under the Curve). Les résultats sont résumés dans le tableau suivant :

Modèle	Précision	AUC
LDA	87.5%	0.7578
QDA	87.4%	0.7640
CART	84.7%	0.7485
Random Forest	86.5%	0.7814
AdaBoost	89.2%	0.8435
Logistic Lasso	79.5%	0.7090

Table 2 – Comparaison des performances des modèles

Le modèle AdaBoost a montré la meilleure performance en termes d'AUC et de précision, suivi par le modèle Random Forest. Ces deux modèles sont bien adaptés pour capturer les complexités des données.

Interprétation des Résultats

Les résultats montrent que le modèle AdaBoost offre les meilleures performances en termes de précision et d'AUC, ce qui en fait le modèle de choix pour cette analyse. Le modèle Random Forest suit de

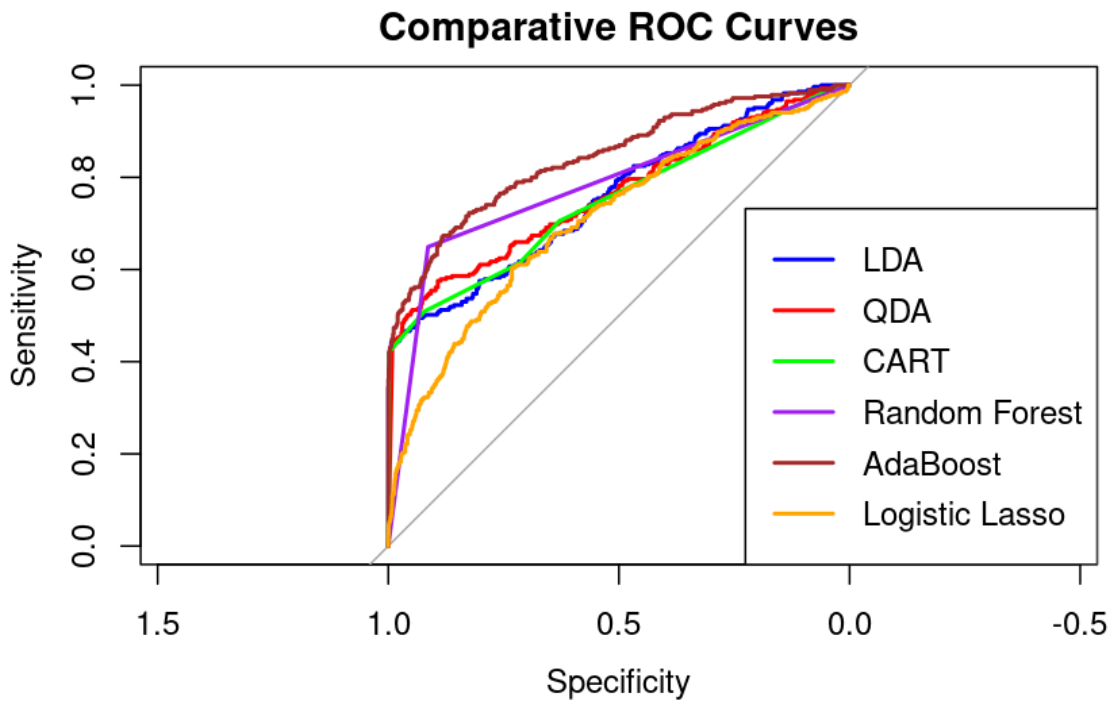


FIGURE 13 – Courbes ROC comparatives des modèles

près en termes de performance, ce qui en fait également un bon choix pour la prédiction de la variable *remote_allowed*.

Les courbes ROC comparatives montrent que AdaBoost et Random Forest offrent une meilleure performance prédictive par rapport aux autres modèles testés. En particulier, AdaBoost, avec une AUC de 0.8435, est le plus efficace pour distinguer entre les offres d'emploi permettant le télétravail et celles qui ne le permettent pas.

La méthode LDA, bien qu'ayant une bonne précision (87.5%), a une AUC inférieure (0.7578), indiquant qu'elle est moins efficace pour les données déséquilibrées. De même, la régression logistique Lasso, avec une précision de 79.5% et une AUC de 0.7090, est moins performante que les autres modèles.

Ensuite, nous avons examiné l'importance des variables pour les modèles AdaBoost et Random Forest afin de mieux comprendre les facteurs influençant le plus la variable *remote_allowed*.

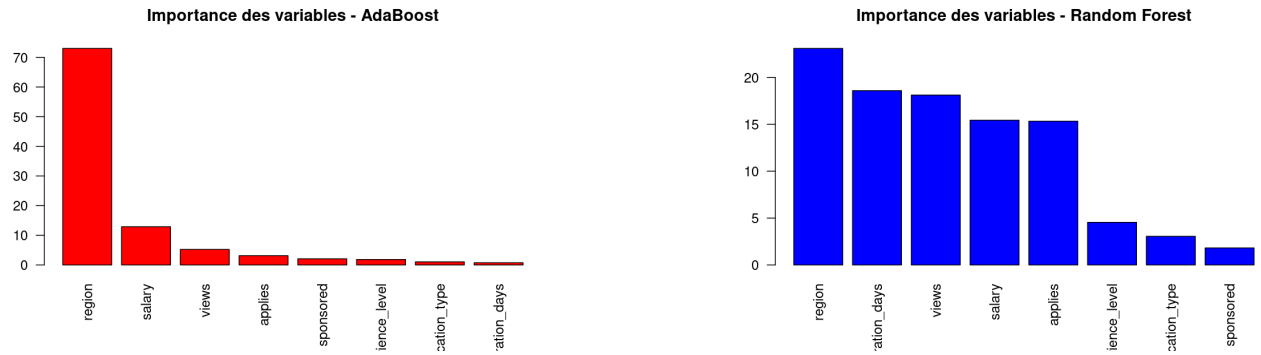


FIGURE 14 – Variable importante des 2 méthodes

L'analyse des graphiques des variables importantes pour les modèles AdaBoost et Random Forest révèle que la région (*region*) est de loin la variable la plus importante pour prédire si un emploi permet le télétravail dans les deux modèles. Cependant, les contributions des autres variables diffèrent légèrement entre les deux modèles. Dans le modèle AdaBoost, le salaire (*salary*) est la deuxième variable la plus importante, mais sa contribution est beaucoup moins significative que celle de la région. Et lees autres variables telles que le nombre de vues (*views*), le nombre de candidatures (*applies*), et le niveau d'expérience (*formatted_experience_level*) ont des importances relativement faibles. En revanche, dans le modèle Random Forest, bien que la région reste toujours la variable la plus importante, la durée des annonces (*duration_days*), le nombre de vues (*views*), et le salaire jouent des rôles plus équilibrés et significatifs. Le nombre de candidatures et le niveau d'expérience sont également notables, mais avec des contributions un peu plus faible. Ces résultats montrent que, bien que la région soit le facteur dominant pour les deux modèles, les autres variables importantes peuvent varier en fonction de l'algorithme utilisé, soulignant la complémentarité des modèles pour capturer différentes facettes des données.

En conclusion, AdaBoost et Random Forest se sont révélés être les meilleurs modèles pour prédire si un emploi permet le télétravail. AdaBoost, en particulier, a obtenu la meilleure précision et AUC. Et la comparaison des courbes ROC (Figure 13) confirme que ces modèles offrent une meilleure performance prédictive par rapport aux autres modèles testés. De plus, les analyses des variables importantes montrent que la région (*region*) est le facteur le plus influent dans les deux modèles, bien que les contributions relatives des autres variables varient légèrement entre les modèles.

Tentatives de Prédiction avec d'autres Variables

Nous avons également tenté de prédire d'autres variables telles que *district* en utilisant les mêmes modèles. Cependant, ces tentatives ont rencontré des problèmes. Par exemple, certaines méthodes ont eu des difficultés avec des données mal équilibrées, ce qui a conduit à des performances médiocres. Ou n'arrivait pas du tout à se lancer. Ces limitations nous ont conduit à nous concentrer principalement sur la prédiction de la variable *remote_allowed*, qui offrait des résultats plus robustes et significatifs.