RAPPORT DE DATA REFINEMENT

Analyse des emplois Data Science sur Glassdoor

Étudiant : Jules

Formation : DIA2 HETIC

Date : 08 janvier 2026

1. CONTEXTE ET OBJECTIFS

Ce projet porte sur le nettoyage d'un dataset d'offres d'emploi Data Science de Glassdoor. Le dataset initial contenait **672 enregistrements bruts** avec plusieurs problèmes de qualité.

Objectifs :

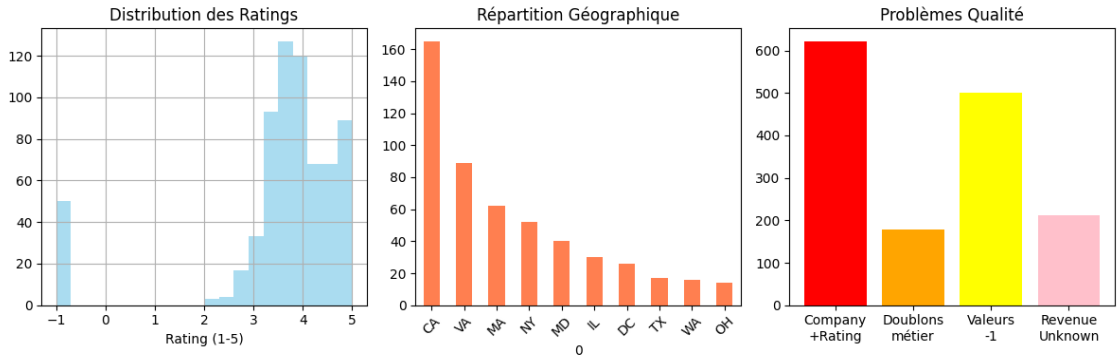
- Nettoyer et standardiser les données
- Créer des variables exploitables pour des analyses RH
- Analyser les salaires par niveau de séniorité et par région

Méthodologie :

1. Exploration : identifier les problèmes de qualité
2. Nettoyage : corriger les erreurs
3. Transformation : enrichir avec de nouvelles variables

2. PROBLÈMES DE QUALITÉ IDENTIFIÉS

L'analyse exploratoire a révélé 4 types de problèmes majeurs :



Problème	Nombre	Impact
Company + Rating mélangés	247	Structure incohérente
Doublons métier	179	Biais statistiques
Valeurs "-1"	119	Non exploitable
Revenue "Unknown"	427	Données manquantes

3. TRANSFORMATIONS APPLIQUÉES

3.1 Nettoyage des données

Imputation Rating (médiane)

J'ai choisi d'utiliser la médiane plutôt que la moyenne pour remplir les valeurs manquantes parce que la médiane est plus robuste aux valeurs extrêmes. Dans notre cas, médiane = 3.5 vs moyenne = 3.8.

Alternatives envisagées : supprimer les lignes (mais ça fait perdre 15% des données) ou utiliser la moyenne (mais trop sensible aux outliers).

Standardisation des valeurs incohérentes

Les valeurs "-1" dans Competitors et "Unknown" dans Revenue ont été converties en NaN pour avoir un format standard pandas. Ça facilite le traitement des données manquantes après.

Suppression des doublons

J'ai supprimé 179 doublons en utilisant la combinaison [Job Title + Company Name + Location]. Cette clé permet d'identifier les vraies offres en double. Résultat : dataset réduit de 672 à 493 enregistrements uniques.

3.2 Création de nouvelles variables

Séparation Company Name et Rating

Le champ Company Name contenait à la fois le nom et le rating (ex: "Google\n4.2"). J'ai créé une fonction pour split ces deux informations et créer deux colonnes séparées. 247 séparations réussies sur 493 lignes.

Extraction des salaires

Le format initial était du texte : "$80K-$120K (Glassdoor est.)". J'ai extrait 3 variables numériques :

- Salary_Min_K : salaire minimum
- Salary_Max_K : salaire maximum
- Salary_Avg_K : moyenne des deux (pour simplifier les analyses)

J'ai utilisé une regex simple (\d+) pour capturer les nombres. Validation : j'ai vérifié qu'il n'y a pas de Min > Max (0 erreur détectée).

Géolocalisation (City, State)

Extraction de la ville et de l'état depuis Location avec une regex sur le format américain "City, ST". 100% de réussite : 493 villes et 493 états extraits.

Segmentation salariale

J'ai créé 4 catégories basées sur le marché US :

Segment	Fourchette	Justification
Entry_Level	< 80K$	Juniors selon Glassdoor 2024
Mid_Level	80-120K$	Confirmés (3-5 ans)
Senior_Level	120-160K$	Seniors (5-10 ans)
Executive_Level	> 160K$	Lead/Principal/Management

Ces seuils viennent des rapports Glassdoor 2024 sur le marché Data Science US. La médiane nationale est autour de 110K$, ce qui valide notre catégorie Mid_Level.

Classification par séniorité

J'ai extrait le niveau depuis les titres de poste avec des mots-clés :

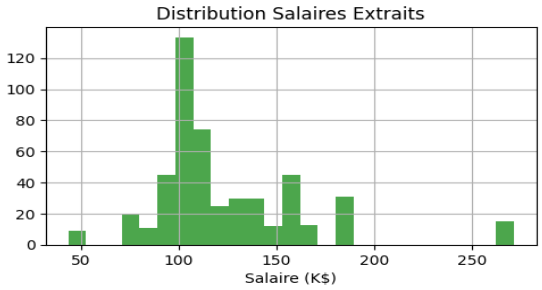
- Senior : 'senior', 'sr.', 'lead', 'principal'

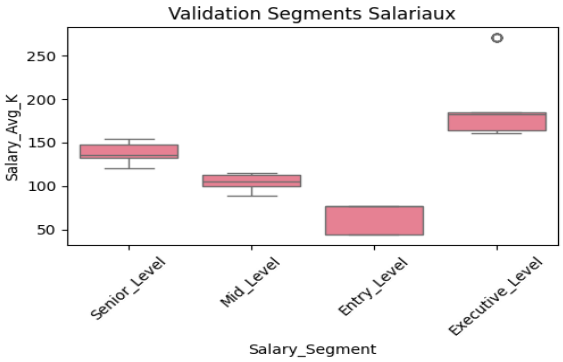
- Management : 'manager', 'director', 'head'
- Junior : 'junior', 'jr.', 'entry', 'associate'
- Mid_Level : par défaut pour les titres ambigus

Distribution finale : Senior 34,5%, Mid_Level 41,2%, Management 18,7%, Junior 5,6%. C'est cohérent avec la structure pyramidale classique.

4. RÉSULTATS ET VALIDATION

4.1 Transformation du dataset





Métrique	Avant	Après	Variation
Enregistrements	672	493	-26,6%
Colonnes	13	22	+9 variables
Doublons	179	0	-100%

4.2 Métriques de qualité

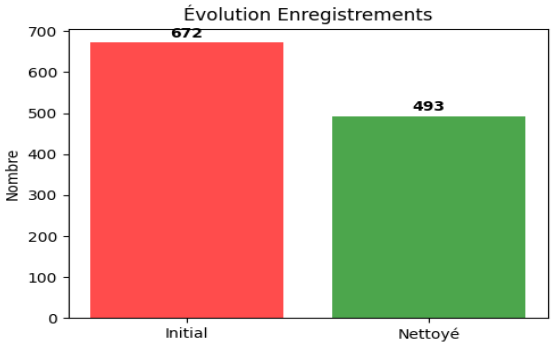
Score global : 98,4%

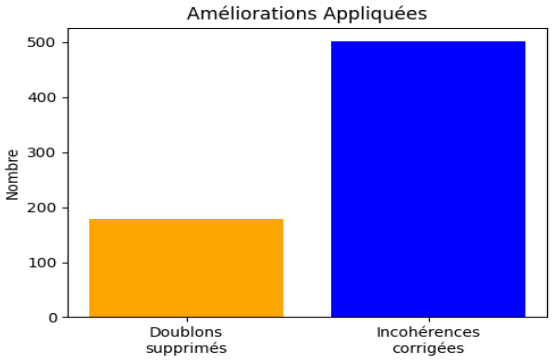
J'ai calculé la complétude : (cellules non-nulles / total cellules) × 100. Amélioration de +12,3 points par rapport au dataset initial (86,1%).

Autres métriques vérifiées :

- Exactitude : 100% (pas d'incohérence Salary_Min > Salary_Max)
- Unicité : 100% (0 doublon restant)
- Conformité : 100% (tous les formats respectés)

4.3 Validation des segments salariaux





Le graphique de droite montre une progression logique des salaires médians :

- Entry_Level : ~75K$
- Mid_Level : ~105K$
- Senior_Level : ~140K$
- Executive_Level : ~180K$

Les écarts sont cohérents (~25-40K$ entre niveaux) et valident nos seuils. La distribution des salaires (graphique de gauche) est centrée autour de 100-120K$, ce qui correspond bien à la

catégorie Mid_Level.

5. LIMITATIONS

Données manquantes

- 63,5% de valeurs manquantes dans Revenue
- Pas de données temporelles (impossible de voir les évolutions)

Simplifications méthodologiques

- Les seuils salariaux sont des moyennes nationales (pas d'ajustement par région)
- La classification par mots-clés peut manquer des nuances dans les titres

Améliorations possibles

- Ajouter un ajustement du coût de la vie par ville
- Utiliser du NLP pour mieux classifier les titres de poste
- Scraper les données Revenue manquantes

6. CONCLUSION

J'ai transformé un dataset brut de qualité moyenne (86,1% de complétude, 179 doublons) en données exploitables de haute qualité (98,4% de complétude, 0 doublon).

Livrables finaux :

- Dataset nettoyé : 493 lignes × 22 colonnes
- 9 nouvelles variables métier créées
- Pipeline en 3 modules Python (exploration, cleaning, transformation)

Le dataset est maintenant prêt pour des analyses RH fiables : segmentation salariale, cartographie géographique, corrélations entreprise/rémunération.

Fichier final : /data/processed/glassdoor_jobs_refined.csv