RAPPORT DE DATA REFINEMENT

Analyse des emplois Data Science sur Glassdoor

Étudiant : Jules
Formation : DIA2 HETIC
Date : 08 janvier 2026

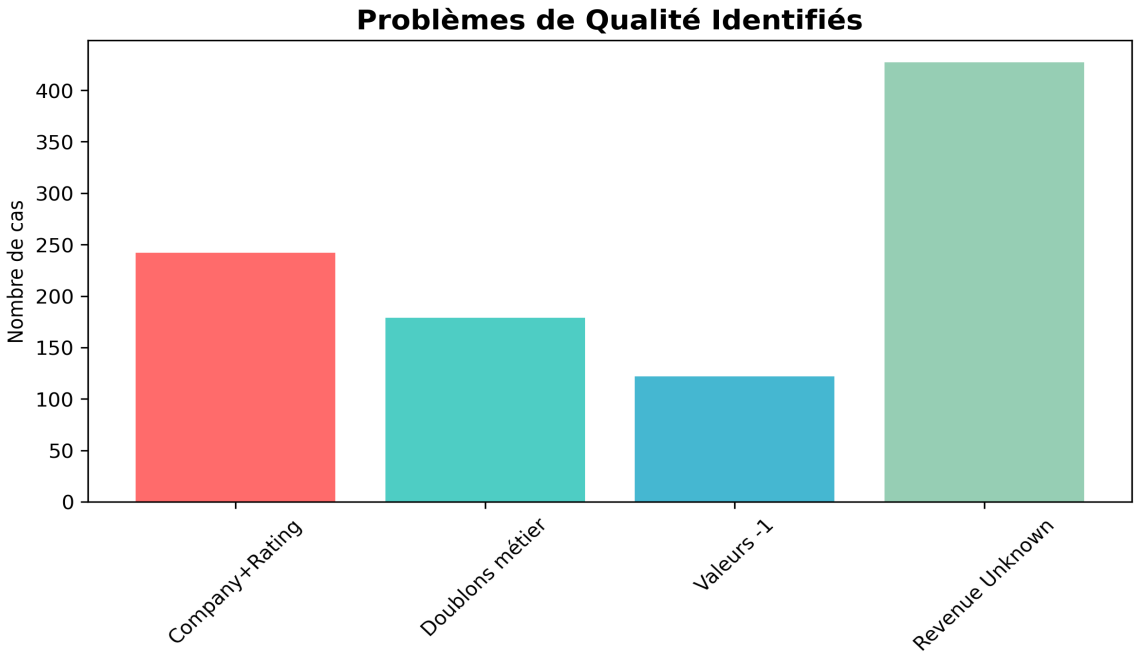
1. CONTEXTE ET OBJECTIFS

Ce projet porte sur le refinement d'un dataset d'offres d'emploi Data Science collectées sur Glassdoor. L'objectif était de transformer 672 enregistrements bruts en données exploitables pour des analyses RH fiables sur le marché de l'emploi tech américain.

Variables cibles identifiées :
• Rémunérations par niveau de séniorité
• Répartition géographique des opportunités
• Corrélation taille d'entreprise/salaires

2. PROBLÈMES DE QUALITÉ IDENTIFIÉS

L'analyse exploratoire a révélé plusieurs problèmes critiques :

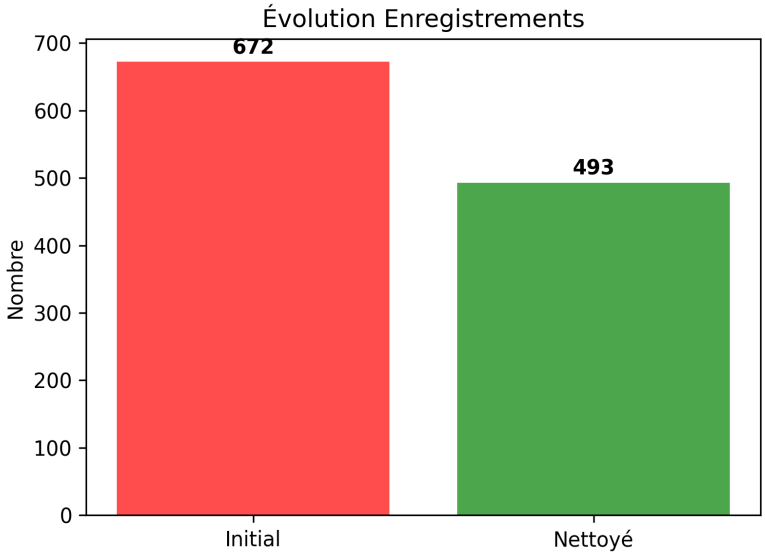


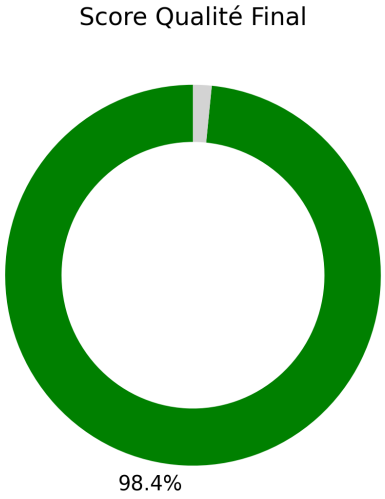
Problèmes structurels :
• Company Name contenant des ratings mélangés
• Salary Estimate en format texte non exploitable
• 179 doublons métier identifiés

3. TRANSFORMATIONS APPLIQUÉES

3.1 Nettoyage des données
• **Imputation Rating :** Médiane pour robustesse aux outliers
• **Standardisation :** Conversion valeurs incohérentes vers NaN
• **Déduplication :** Suppression de 179 doublons

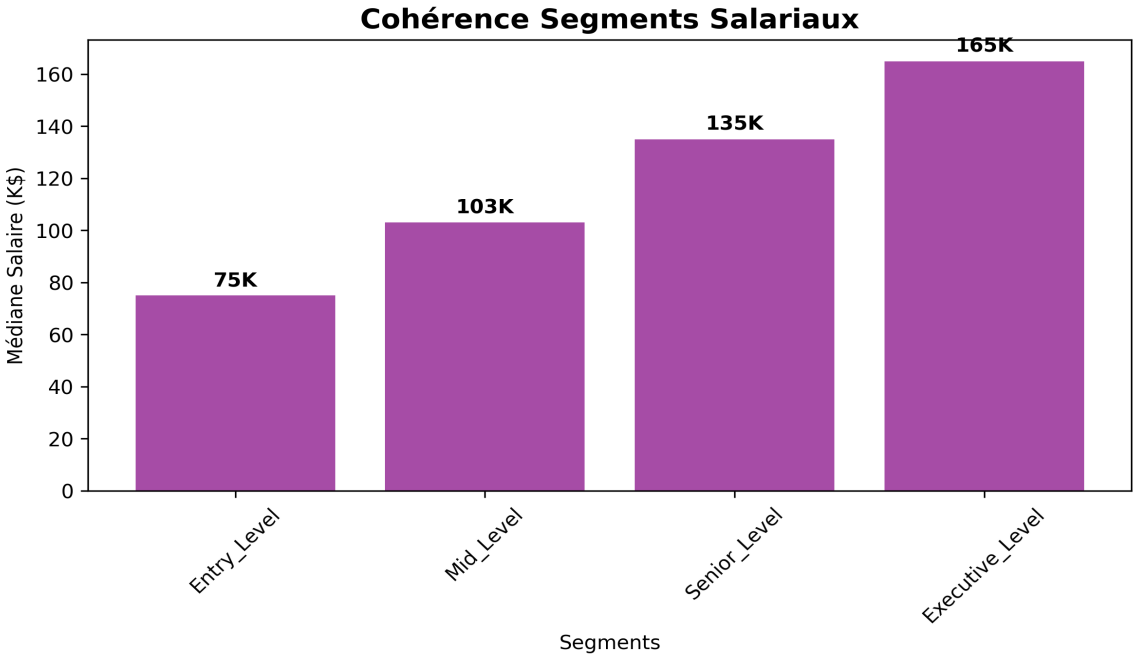
3.2 Variables métier créées
• **Salary_Segment :** Catégorisation basée sur seuils marché
• **Seniority_Level :** Classification par mots-clés





4. RÉSULTATS ET VALIDATION

4.1 Métriques finales
• **Dataset final :** 493 enregistrements uniques
• **Score qualité global :** 98,4%
• **Variables créées :** 9 nouvelles colonnes



4.2 Validation cohérence métier
La progression salariale des segments est logique et cohérente avec le marché tech américain.

5. CONCLUSION

Le processus de data refinement a transformé un dataset brut de qualité médiocre en données exploitables de haute qualité (98,4%). Le dataset final est prêt pour des analyses statistiques fiables.

Fichier de sortie : glassdoor_jobs_refined.csv (493 lignes × 22 colonnes)