Rapport d'analyse et de nettoyage des données

Projet : Étude des offres d'emploi en lien avec les métiers Tech

---

1. Description du jeu de données

1.1. Données collectées

- **Source :** CSV France_Travail initial
- **Taille du jeu de données :**
  - **Nombre de lignes :** 458 113
  - **Nombre de colonnes :** 60

1.2. Données après nettoyage

- **Taille du jeu de données :**
  - **Nombre de lignes :** 665
  - **Nombre de colonnes :** 17
- **Colonnes conservées :**
  - id, intitule, description, dateCreation, dateActualisation, lieuTravail_latitude, lieuTravail_longitude, lieuTravail_libelle, romeCode, typeContrat, experienceExige, alternance, origineOffre_urlOrigine, dureeTravailLibelleConverti, competences, qualitesProfessionnelles, formations.

---

2. Processus de nettoyage et transformation

2.1. Tableau des modifications

Type d'erreur	Problème identifié	Actions réalisées	Résolution
Tri des données	Trop de métiers et code Rome différents	Sélection de deux code Rome : M1403 (data analyst et scientist) et M1805 (développeur)	1435 lignes restantes.
Valeurs manquantes	Latitudes et longitudes manquantes dans la colonne lieuTravail_latitude et lieuTravail_longitude.	Suppression des lignes avec valeurs manquantes (dropna).	665 lignes restantes.
Colonnes inutiles	Présence de colonnes non pertinentes pour l'analyse.	Sélection uniquement des colonnes nécessaires.	Conservation de 17 colonnes.
Données imbriquées	Dans les datas de l'API : certaines informations (lieu de travail, url d'origine) sont dans des colonnes composites.	Extraction des sous-informations dans des colonnes dédiées (lieuTravail_latitude, lieuTravail_longitude, lieuTravail_libelle, origineOffre_urlOrigine).	Préparation pour la base de données (V2).
Format des dates	Format non uniforme dans dateCreation et dateActualisation.	Conversion des dates au format standard ISO8601.	Dates uniformisées dans l'API.
Doublons	Présence de doublons dans les données.	Suppression des doublons avec drop_duplicates().	Unicité des données assurée.
Traitement texte	Récupération des compétences spécifiques depuis la colonne description.	Utilisation d'un LLM pour l'extraction des compétences.	Identification des compétences techniques. Implémentation dans la V2

---

3. Analyse et visualisation des données à partir du CSV

3.1. Analyse par code ROME

- **Proportion des offres**
  - **M1403 :** 49.2% soit 327 offres
  - **M1805 :** 50.8% soit 338 offres
- **Code ROME M1403 (Data Analyst/Scientist) :**

    - **Île-de-France :** 81 offres (24,7%).
    - **Grand-Est :** 18 offres (5.5%).
- **Code ROME M1805 (Développeurs) :**
    - **Île-de-France :** 40 offres (11.83%).
    - **Grand-Est :** 14 offres (4.14%).

3.2. Répartition géographique

- **Carte interactive créée :**
    - **Visualisation :** Localisation des offres (latitudes et longitudes).
    - **Fonctionnalité :** Accès direct aux offres via l'URL origineOffre_urlOrigine.

3.3. Compétences identifiées

Extraction des compétences depuis la colonne **description** à l'aide d'un LLM :

- **Compétences techniques (hard skills) :** Analyse de données, SQL, Python, visualisation de données (Power BI, Tableau), IA.
- **Compétences comportementales (soft skills) :** Esprit d'analyse, rigueur, communication, autonomie.
- **Techniques (savoir-faire) :** Langages de programmation, analyse de données, outils de développement.

3.4. Analyse par type de contrat et expérience

- Répartition des offres :
    - **Type de contrat :** Alternance, CDI, CDD, Freelance.
    - **Niveau d'expérience :** Débutant, intermédiaire, confirmé.

---

4. Étude des données issues de l'API pour M1805

4.1. Caractéristiques des données collectées

- **Période :** Du 25 novembre au 8 décembre.
- **Nombre de lignes :** 474
- **Nombre de colonnes :** 39

4.2. Transformations nécessaires pour insertion en base de données (BDD)

- **Colonnes composites à traiter :**
    - lieuTravail : Extraction des informations **latitude**, **longitude**, et **libellé**.
    - origineOffre : Extraction de urlOrigine.

- **Améliorations prévues pour la V2 :**
    - Configurer les codes Rome pour M1403 et d'autres.
    - Implémentation du bouton de mise à jour des offres dans l'interface.
    - Intégration du LLM pour trier les compétences
    - Automatisation des extractions.

---

5. Conclusion et axes d'amélioration

5.1. Résumé des actions principales

- Nettoyage des données pour éliminer les valeurs manquantes et les doublons.
- Conservation des colonnes nécessaires pour l'analyse.
- Utilisation d'un LLM pour identifier les compétences clés (à intégrer dans un des micro-services).
- Visualisation géographique et analyse par région et type de métier.

5.2. Limites rencontrées

- Manque de temps pour transformer entièrement les données API.
- Certaines données composites nécessitent des modifications pour la version 2 de l'application.
- Problèmes de timeout lors de la demande manuelle de mise à jour.

5.3. Recommandations

- Automatiser le traitement des données de l'API.
- Configurer l'API pour d'autres codes Rome.
- Intégration du LLM afin de récupérer les compétences dans description.
- Amélioration des filtres dynamiques pour l'analyse par contrat et expérience.
- Ajout d'une visualisation par rapport aux top 5 des compétences