Contexte

On a besoin de nettoyer les données et de transformer les colonnes nécessaires afin d'informer les parties prenantes de manière efficace en se basant sur un tableau de bord permettant de comparer les emplois dans le domaine de la science des données.

Description du jeu de données

Le jeu de données fourni contient des informations sur différents emplois dans le domaine de la science des données. Voici une description des colonnes qui pourraient nous intéresser :

- Job Title (Titre d'emploi) : Le titre du poste proposé.
- Job Description (Description de l'emploi) : Une description détaillée des responsabilités et des exigences du poste.
- Company Name (Nom de l'entreprise) : Le nom de l'entreprise proposant le poste.
- Location (Emplacement) : L'endroit où se situe le poste.
- Industry (Industrie) : Le secteur d'activité de l'entreprise qui propose le poste.
- Estimate Min Salary : Le salaire minimum estimé pour le poste.
- Estimate Max Salary : Le salaire maximum estimé pour le poste.

Objectifs

1. Nettoyer les données et les préparer pour un tableau de bord.
2. Comparer et informer les parties prenantes sur l'emploi dans le domaine de la science des données.

Étapes

1. Importation des données
    a. Télécharger et charger les données à partir de : https://www.kaggle.com/datasets/andrewmvd/data-scientist-jobs
    b. Afficher les 5 premières lignes pour avoir un aperçu des données.

2. Nettoyage des données
    a. Suppression de colonnes non pertinentes : supprimer les colonnes : `index`, `Rating`, `Headquarters`, `Size`, `Founded`, `Type of ownership`, `Revenue`, `Sector`, `Easy Apply`, `Competitors`, `Unnamed: 0`.
    b. Nettoyage de la colonne 'Salary Estimate' :
        i. Identifier et afficher le nombre de lignes où la colonne 'Salary Estimate' ne respecte pas le format : `$chiffre_minK-$chiffre_maxK (texte)`.
        ii. Supprimer ces lignes.
    c. Suppression de lignes non pertinentes : supprimer toutes les lignes où la colonne 'Industry' contient la valeur `-1`.
    d. Gestion des valeurs nulles et manquantes : identifier et traiter les valeurs nulles ou manquantes selon la stratégie appropriée (imputation, suppression, etc.).

3. Transformation des données
    a. Transformation des titres de postes :
        i. Transformer la valeur de la colonne `Job Title` selon les sous-chaînes spécifiques :
            1. En `Data Analyst` si contient `analyst`.
            2. En `Data Scientist` si contient `scientist` ou `science`.
            3. En `Data Engineer` si contient `engineer`.
    b. Filtrage des catégories de métiers : garder uniquement les lignes correspondant aux catégories `Data Analyst`, `Data Scientist` et `Data Engineer`.
    c. Transformation de la colonne 'Salary Estimate' :
        i. Remplacer `K` par `000`.
        ii. Extraire les valeurs minimales et maximales des salaires et les diviser en deux colonnes : `Estimate Min Salary` et `Estimate Max Salary`.
        iii. Supprimer la colonne `Salary Estimate`.
    d. Extraction de la colonne 'Company Name' : extraire la partie avant le caractère `\n`.
    e. Extraction de la colonne 'Location' : extraire la partie avant le caractère `,`.

4. Analyse des données
    a. Analyse des compétences requises : compter le nombre de postes contenant les sous-chaînes suivantes : `sql`, `python`, `power bi`, `tableau`, `excel`, `r`. Supposons que ce soient les compétences dont nous souhaitons trouver la fréquence d'apparition dans chacune des trois catégories d'emploi.
    b. Convertir le résultat en dataframe :

	sql	python	power bi	tableau	excel	r
Data Analyst	507	243	88	247	462	60
Data Scientist	627	933	35	175	769	155
Data Engineer	473	515	42	112	245	40

    c. Proposer une visualiser du résultat
    d. Calculer la moyenne arrondie des salaires estimés (Min et Max) par titre d'emploi et visualiser le résultat
    e. Vérifier si les salaires estimés des Data Analyst contiennent des valeurs aberrantes
    f. Analyse exploratoire des données (EDA) : utiliser la bibliothèque `ydata_profiling` pour créer une analyse exploratoire des données.

Bonus

Proposer deux axes d'analyse de votre choix

Livrables

- Fichier jupyter notebook sous forme d'un rapport bien lisible et compréhensible