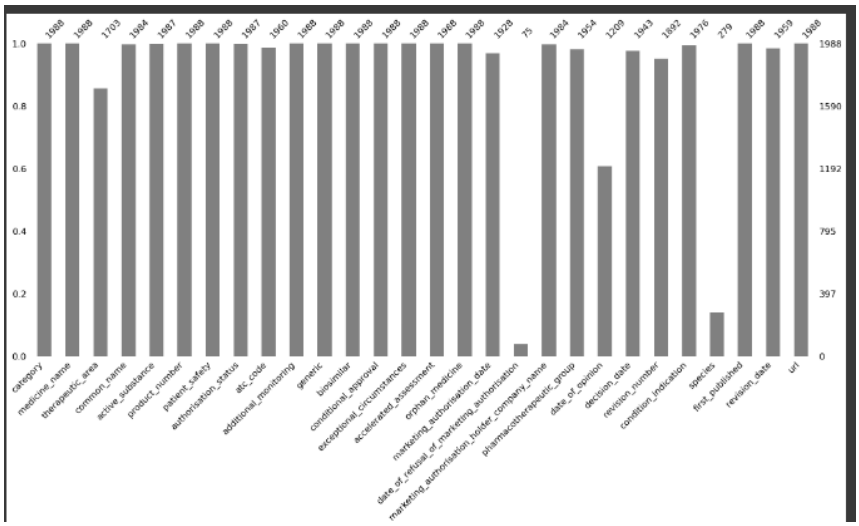
Machine Learning

Projet: Exploration et analyse des données

Aissatou BALDE & Aïda SOW

Typologie des variables: Après analyse l'ensemble des 28 colonnes de notre dataset sont des variables catégorielles

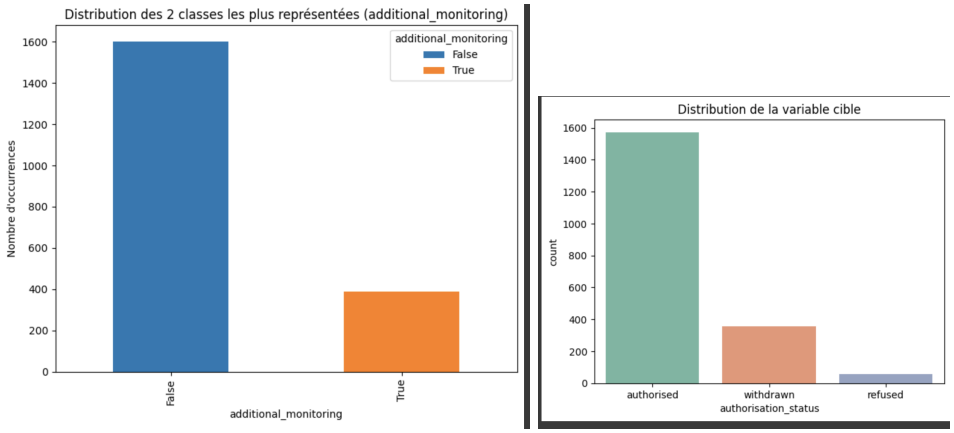
Proportion de données manquantes de notre dataset



- **Valeurs manquantes dans les colonnes: P**our les valeurs manquantes, on voit qu'il ya que plusieurs colonnes en ont. Cependant il y a quatre (04) colonnes qui ont plus de valeurs manquantes que les autres. Ce sont les colonnes:
  - **therapeutic_area:** avec **1703 valeurs renseignées sur 1988,**
  - **date_of_refusal_of_marketing_authorisation:** avec **75 valeurs renseignées sur 1988**
  - **date_of_opinion: avec 1209 renseignées sur les 1988**
  - **species: avec seulement 279 valeurs renseignées sur les 1988**

Analyse univariée:

- **Sélection des colonnes:** Suppression des colonnes avec plus de 1000 modalités pour réduire la taille du dataset
- **Etude des tableaux des fréquences pour chaque variable pour déterminer la distribution**
- **Visualisation des deux fréquences les plus importantes pour chaque variable**



- **Analyse bivariée: Matrice de corrélation avec le coefficient Phi**