

RAPPORT DE PROJET

**Thème** : **Mise en place d’un système Big Data pour une étude épidémiologique**

Membres du groupe :

* - Guillon Xavier
* - KENFACK Ivana
* - LONTSIE Dilane
* -MADJOU Flora

[**I. Contexte 3**](#_sqh7gw61ms4a)

[**Étape 1. Etude de besoin 3**](#_io95q9a9t8ob)

[1. Le choix de l'épidémie 3](#_5vyju4kwul2b)

[2. Objectifs du projet 3](#_6efgbo2z3qs3)

[3. Les sources 4](#_d4ce0p7ga0u5)

[**Étape 2 . Mise en place de l’architecture de stockage 4**](#_1afoqpb65c60)

[1. Installation de l’environnement de travail 4](#_rhr75s578cr7)

[**Étape 3 : Traitement des données 6**](#_7gszjvrz2ipa)

[2. Traitement dans Rstudio 7](#_f65qjf9enzzq)

[1. Mise en place du conteneur de rstudio 7](#_9aczp5p57xgb)

[2. Connection à RStudio 7](#_euxa5pai9a5q)

[3. Obtenir les données à partir de l’ordinateur vers docker 8](#_s1uf6kq1s2p8)

[4. Construction du data lake 9](#_rio7cip6wq2p)

[5. Charger les résultats des analyses dans l’ordinateur 9](#_zgsk5lwi79xb)

[6. Nettoyage des données avec rstudio 9](#_n4nnuzvbfixs)

[7. Analyses avec rstudio 11](#_kxk9joq1t7l3)

[Étape 4 : Visualisation des résultats 11](#_lmoxc2rx7pcv)

# Contexte

Dans le cadre de notre cours de Big Data, il nous a été demandé de mener un projet d’analyse avec une approche Big Data. De ce fait, nous avons choisi de nous porter vers une étude épidémiologique, qui sera le cœur de notre projet que nous explicitons au fil des pages suivantes. Le travail effectué sera structuré en 5 grandes étapes : l’étude du besoin, la mise en place de l’environnement de stockage et de traitement, l’ingestion/extraction des données, le traitement de données et la visualisation des données.

## Étape 1. Etude de besoin

### Le choix de l'épidémie

Le thème du projet étant assigné, il fallait trouver une problématique en lien avec l’évolution de l'épidémie, son impact selon les critères que nous avons dû déterminer. La première décision qu’il nous a fallu prendre a été le choix de l’épidémie à étudier et nous avons décidé de nous intéresser à l’évolution du SIDA dans le monde.

### Objectifs du projet

La tâche la plus importante lors de cette première étape a été de décider quels allaient être les objectifs de notre projet. Après de longues réflexions nous avons décidé de choisir comme objectif :

* Identifier les populations les plus touchées, et les raisons pour lesquelles elles le sont. Il sera question pour nous d’effectuer une analyse descriptive sur l’évolution du VIH.
* Réaliser une analyse prédictive des populations qui seront les plus touchées par le VIH.
* Évaluer l'efficacité des interventions de prévention et de traitement: Les données sur le SIDA nous aideront à aider à évaluer l'efficacité des interventions de prévention, telles que les programmes de dépistage, l’accès aux traitements antirétroviraux et la promotion de comportements sexuels à faible risque.

### Les sources

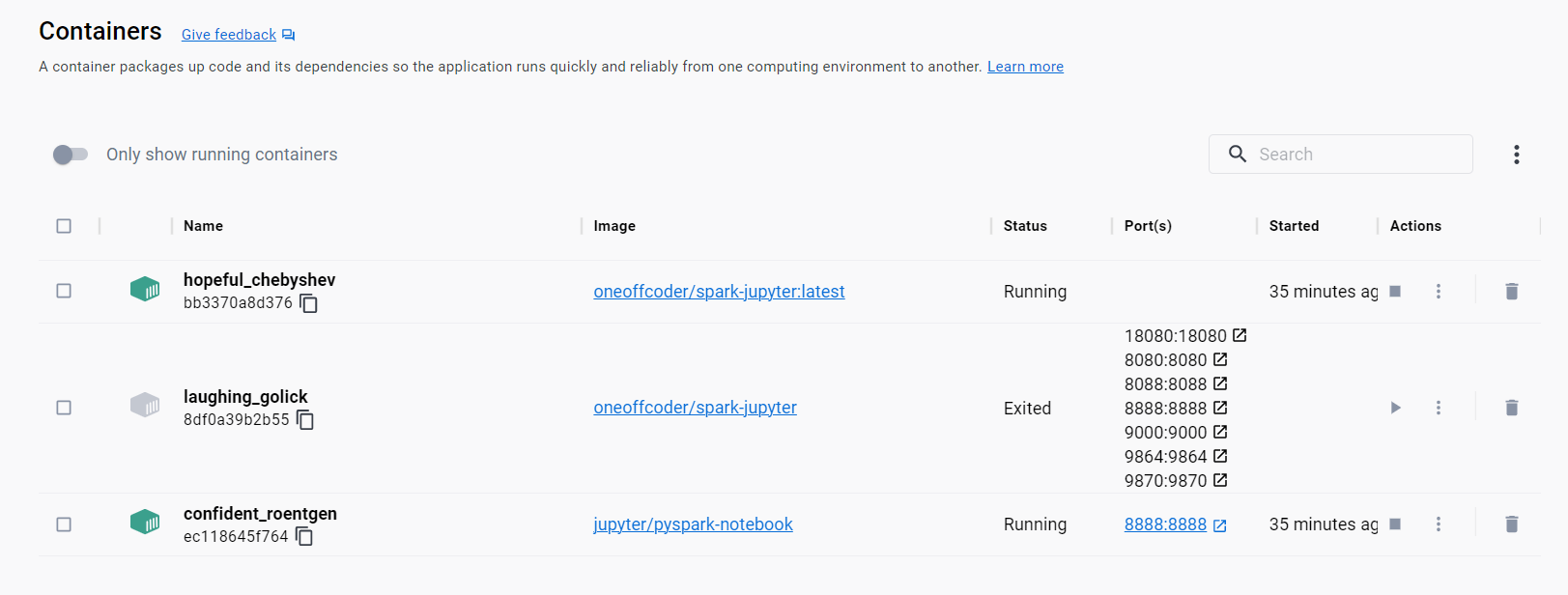
Cette étape a consisté à identifier les sources de données que nous allons utiliser pour pouvoir réaliser notre projet Big Data. Après de longues recherches, notre choix s’est porté sur le dépôt de données disponibles sur Kaggle et sur World Health Organisation (who). Les principaux avantages de ces dépôts sont qu’il est mis à jour régulièrement et qu’il est très complet (regroupe toutes les informations dont nous avions besoin pour le projet).

## Étape 2 . Mise en place de l’architecture de stockage

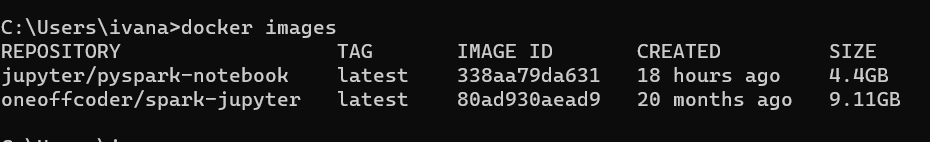
Cette étape avait pour principal but la mise en place de notre environnement de travail. La technologie que nous avons choisie pour notre projet est Hadoop qui est un framework libre et open source qui permet le stockage et le traitement de larges volumes de données dans un environnement distribué (sur plusieurs machines distinctes).

### Installation de l’environnement de travail

* Nous avons utilisé Docker pour mettre en place notre environnement de travail



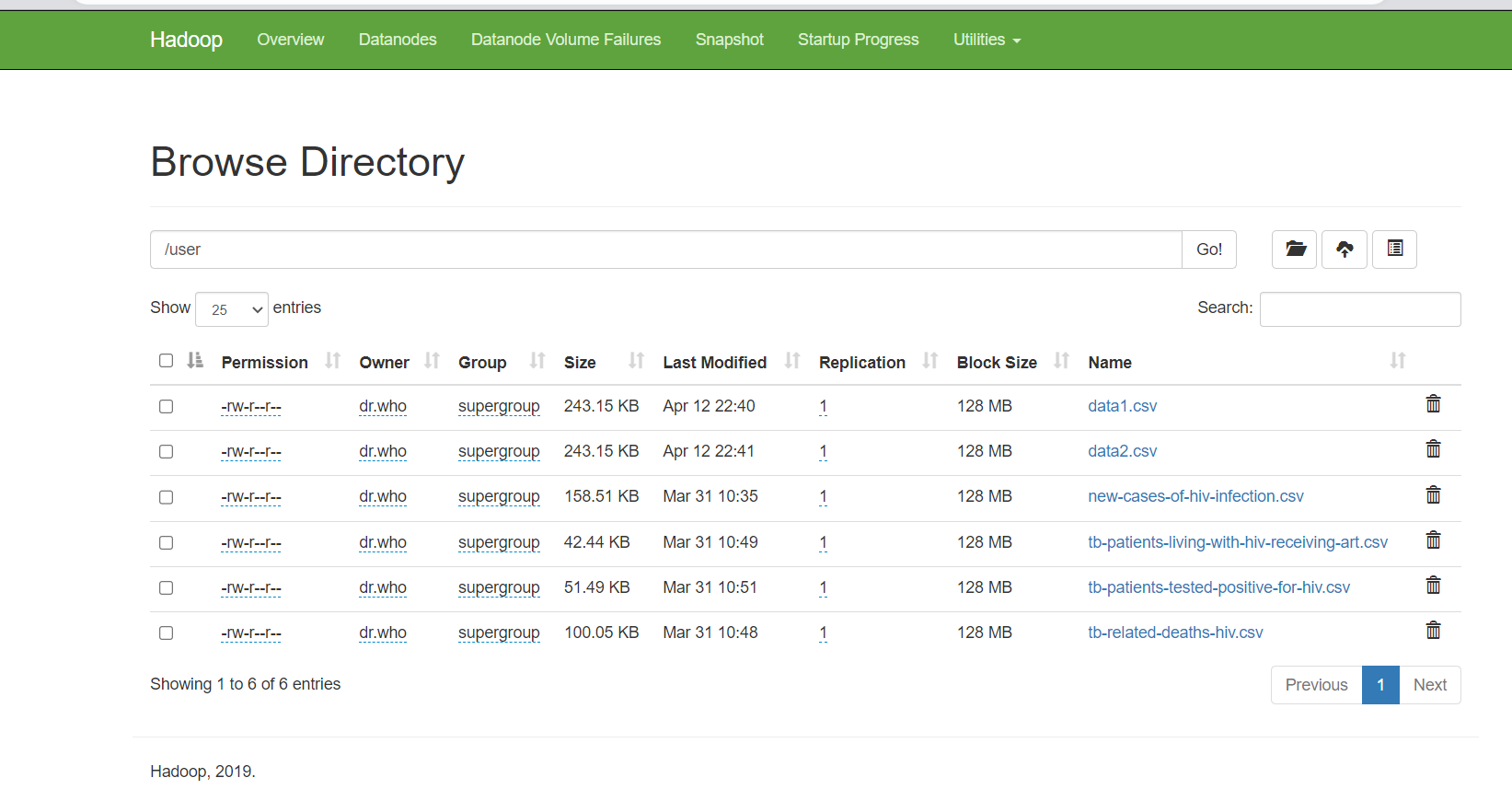
* Nous avons opté pour un cluster Hadoop constitué d’un maître et de deux esclaves, pour se faire nous avons importé une image contenant les éléments dont nous aurons besoin pour nos travaux, nous avons utilisé l’image offerte oneoffcoder à travers ce lien <https://github.com/oneoffcoder/docker-containers/tree/master/spark-jupyter>
* Télécharger l’image docker à l’aide de la commande **docker pull oneoffcoder/spark-jupyter**

On remarque bien la présence de notre image parmi les différentes images déjà présentes sur docker.

Enfin on lance l'exécution des différents conteneurs grâce à la commande:

docker run -it -p 9870:9870 -p 8088:8088 -p 8080:8080 -p 18080:18080 -p 9000:9000 -p 8888:8888 -p 9864:9864 -v $HOME/git/docker-containers/spark-jupyter/ubuntu/root/ipynb:/root/ipynb -e PYSPARK\_MASTER=spark://localhost:7077 oneoffcoder/spark-jupyter

Ainsi notre cluster Hadoop est lancé et ainsi on a procédé au stockage des données qui seront par la suite répliquées



## Étape 3 : Traitement des données

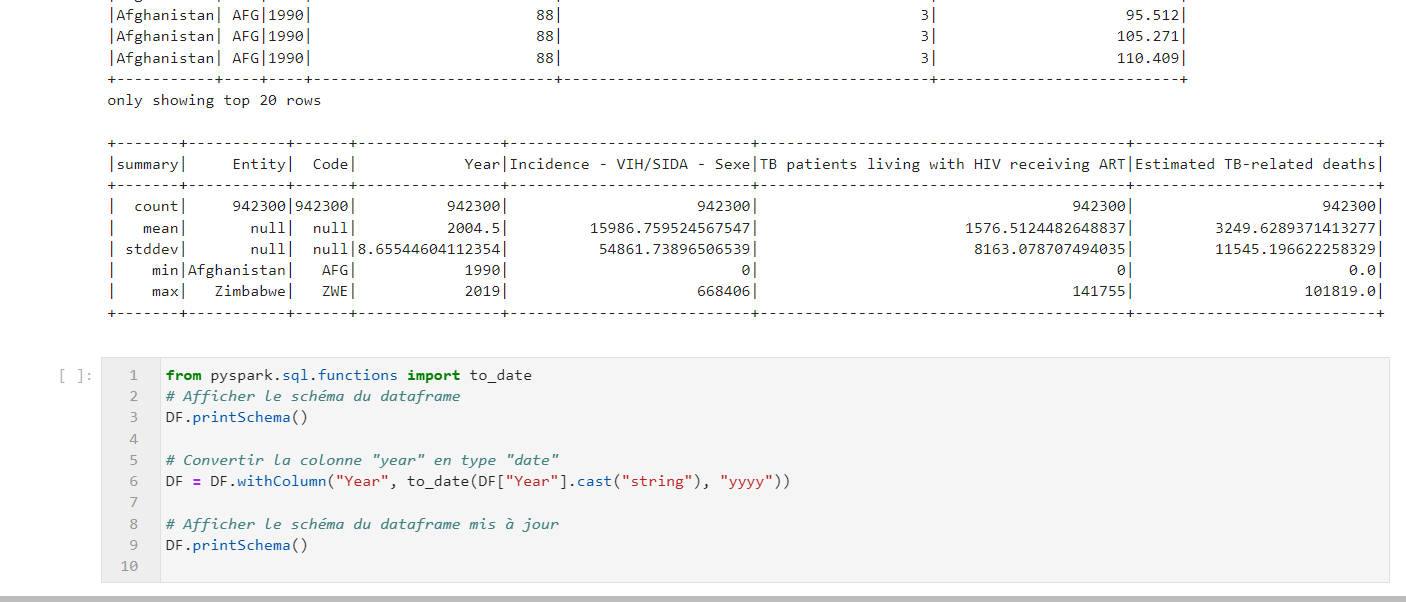
1. Traitement dans SPARK

Le traitement des données avec Pyspark est un processus qui permet d’analyser et de manipuler de grandes quantités de données en utilisant la puissance de calcul distribuée. C’est ainsi que nous avons opté pour le nettoyage et le traitement de données afin de sortir des analyses pertinentes.

Tout d’abord, il était important pour nous de charger les données dans une session spark en utilisant des fonctions de lecture de fichiers telles que read(). Ensuite, nous avons procédé au nettoyage des données en utilisant des transformations PySpark pour supprimer les données manquantes, filtrer les valeurs aberrantes et formater les données en vue d’une analyse ultérieure.



Une fois les données nettoyées, il est temps d’effectuer des opérations de transformation pour identifier les populations les plus touchées par le VIH. Par exemple des opérations de jointure et de regroupement pour combiner les données de différents ensembles de données.



### Traitement dans Rstudio

#### Mise en place du conteneur de rstudio

Dans cette partie du projet, nous allons accéder aux services de rstudio grâce à un mappage de ports. Nous allons mapper un port dans docker avec un port de la machine locale, c’est-à-dire que nous allons relier un port du conteneur vers un port de l’ordinateur, afin d'exécuter un conteneur docker à partir de l’image ‘rocker/rstudio’. Le port 8787 du conteneur est mappé avec le port 8787 d'un ordinateur en local . La commande **docker container** “ls” permet de voir sur quels ports de la machine sont exposés un conteneur.

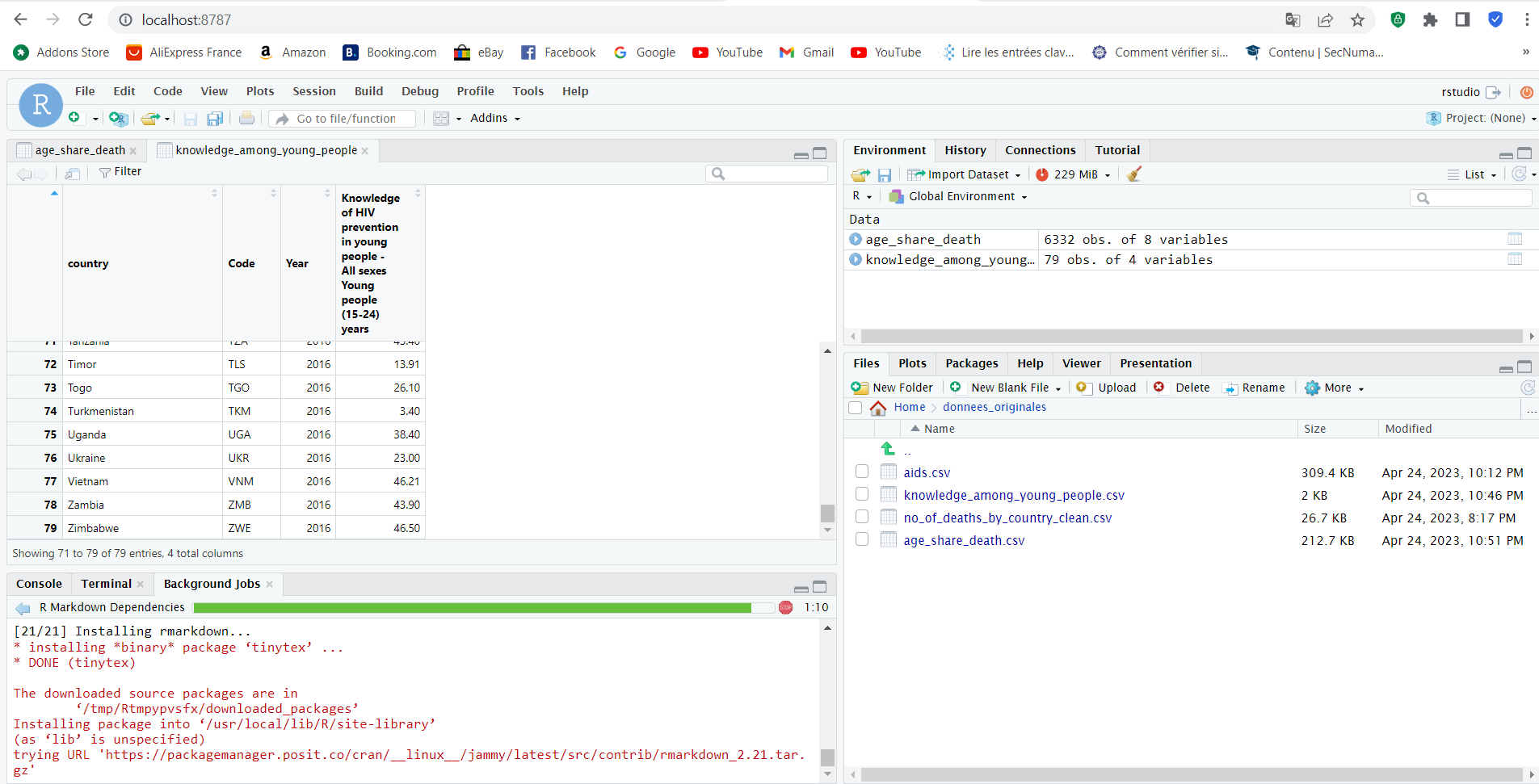
Une image contenant texte, capture d’écran, Police

Description générée automatiquement

#### Connection à RStudio

Nom d'utilisateur: **rstudio**; Mot de passe= **monmotdepasse**

Obtenir l’environnement de travail utilisé à l'intérieur du container docker :



#### Obtenir les données à partir de l’ordinateur vers docker

Nous allons créer un dossier nommé « mes\_données » et charger les data sets à l’intérieur :

* Ouvrir l’invite de commande
* Exécuter la commande suivante : docker cp chemin\_vers \_le\_fichier\_source nom\_du\_container:/home/rstudio . Dans le cadre de nos analyses : **docker cp C:\I2\_Flora\SEMESTRE\_2\Big\_Data\Projet\_Big\_data\data\_set\originaux\new-cases-of-hiv-infection.csv mon\_image\_rstudio:/home/rstudio**

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

Résultat :

Une image contenant texte, capture d’écran, Police, logiciel

Description générée automatiquement

Nous allons répéter cette commande pour tous nos data set.

#### Construction du data lake

Pour construire le data Lake dans rstudio, nous avons créé deux répertoires dans l’environnement de travail de rstudio partitionné de la manière suivante :

* Un répertoire donnees\_originales pour tous les différents dataset qui seront analysé
* Un répertoire donnees\_analysees qui va stocker les résultats issus des analyses

Il est nécessaire d’importer nos fichiers dans l’environnement R avec la fonction read\_csv qui permet de lire un fichier CSV et de stocker les données qu'il contient dans un objet.

Une image contenant texte, logiciel, Icône d’ordinateur, nombre

Description générée automatiquement

#### Charger les résultats des analyses dans l’ordinateur

Utiliser la commande suivante : Docker cp source destination ; en remplaçant « source » et « destinations » par les chemins d'accès du fichiers sources vers sa destination dans le pc

Exemple : **docker cp mon\_image\_rstudio:/home/rstudio/ new\_cases\_of\_hiv\_infection C:\I2\_Flora\SEMESTRE\_2\Big\_Data\Projet\_Big\_data\data\_set\donnes\_analysees**

#### Nettoyage des données avec rstudio

* Examiner le type de colonnes

Une image contenant texte, Police, capture d’écran

Description générée automatiquement

* Vérifier s’il y a des valeurs manquantes

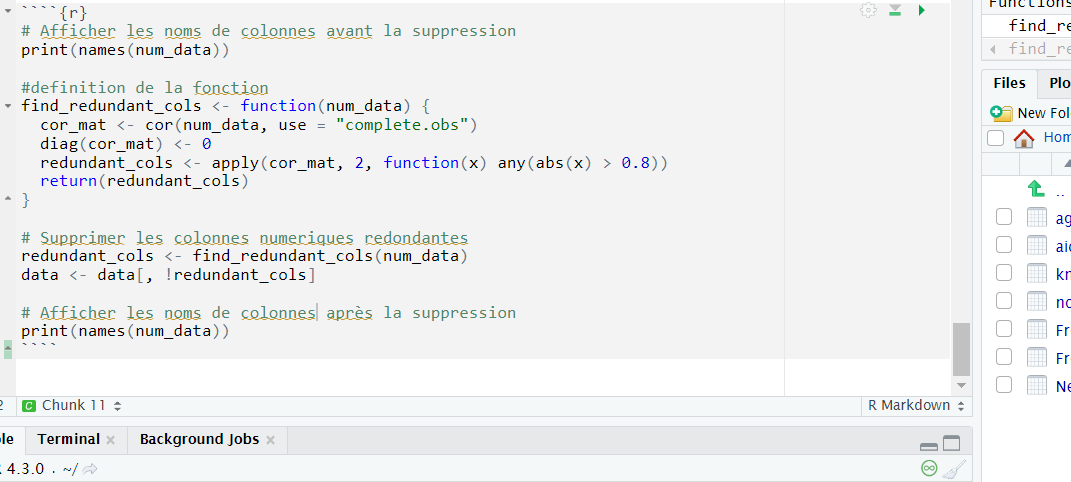
Une image contenant texte, capture d’écran, Police, ligne

Description générée automatiquement

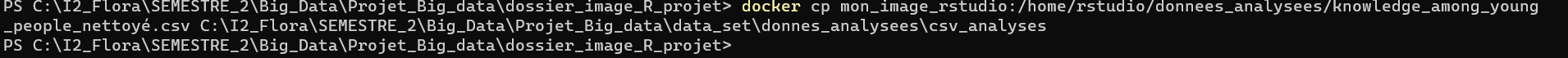
* Supprimer les valeurs manquantes ou colonnes vides
* Identifier les colonnes redondantes ou inutiles et les supprimer

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement



* Renommer des colonnes
* Envoyer les données nettoyées vers le pc



#### Analyses avec rstudio

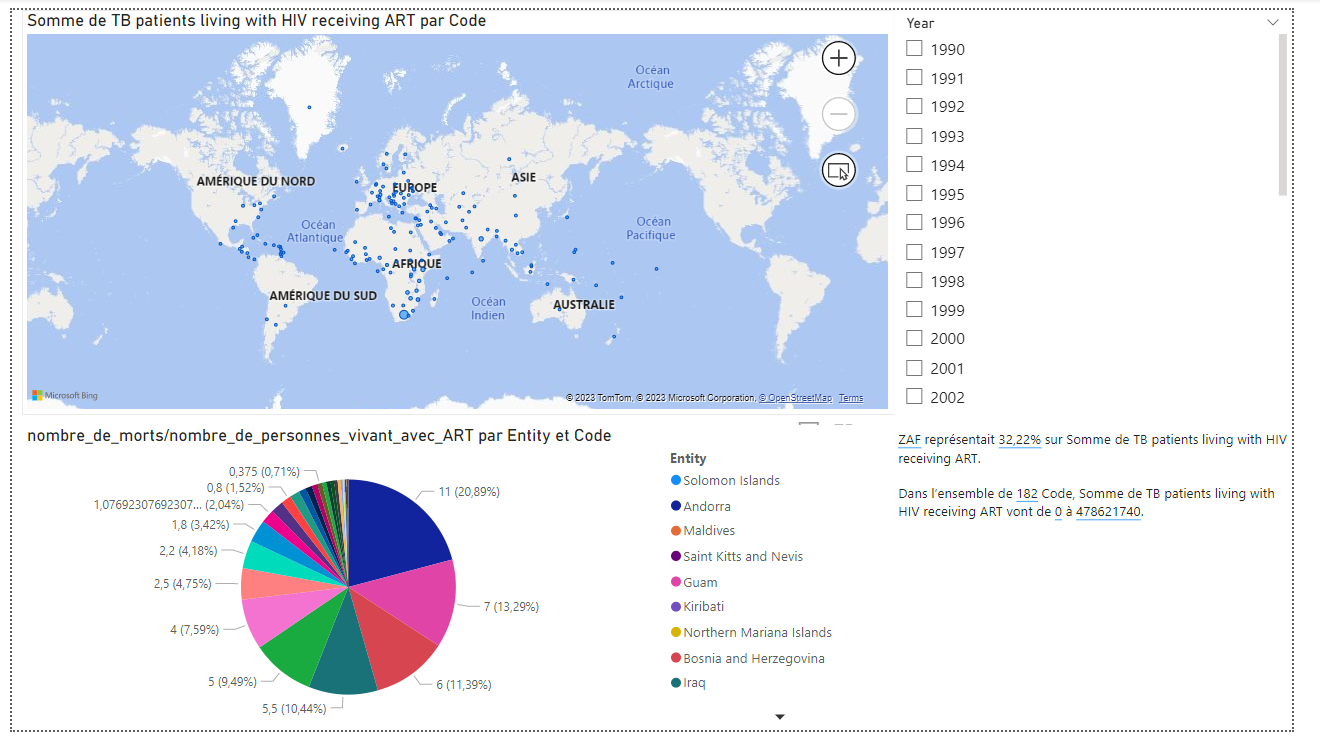
Nous allons effectuer une régression linéaire des data set et répondre aux questions suivantes : quelles populations seront les plus touchées par le sida dans le monde? Qu’est-ce que les pays de ces populations ont en commun?

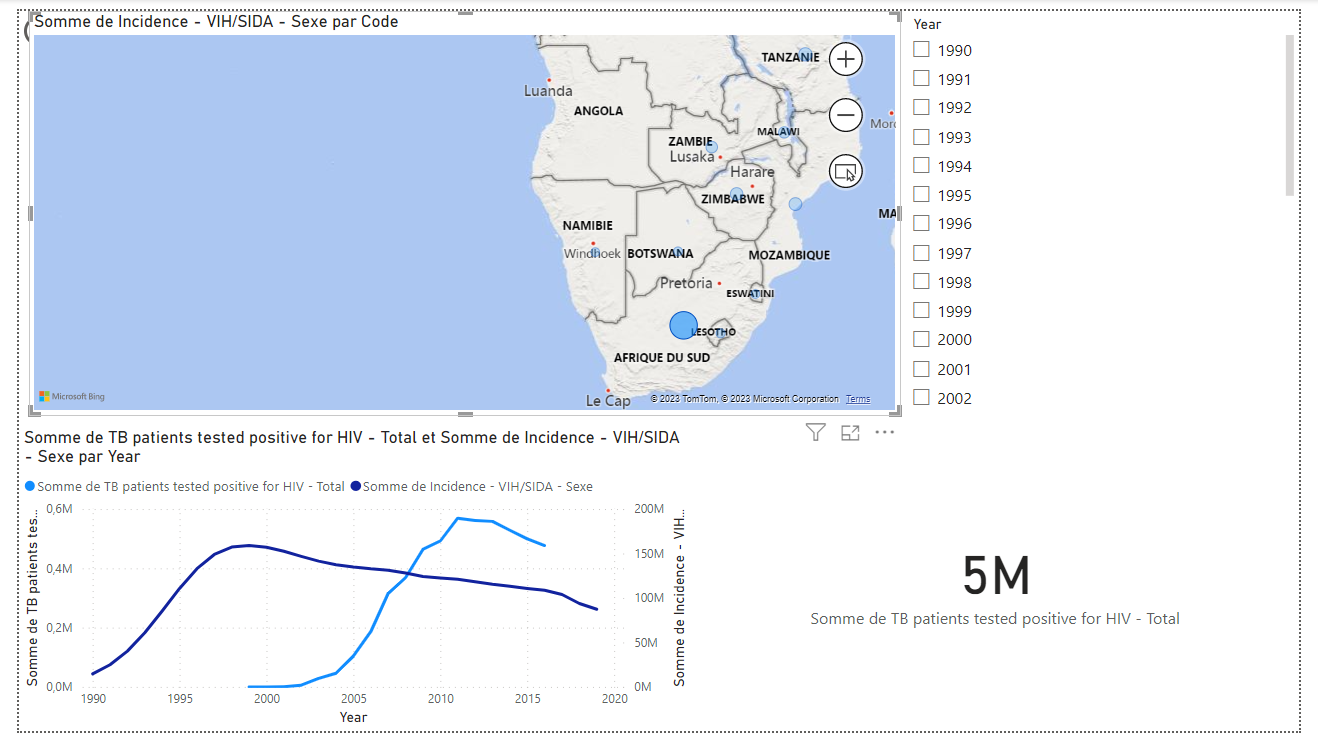
Tout d’abord, nous avons les données qui vont de 1990 à 2022. Nous souhaitons faire une analyse de données entre 2018 et 2022. Nous utilisons la fonction « subset() » de R pour sélectionner les données entre 2017 et 2019.

Nous allons séparer ces données en données d'entraînement et de test car il ne s’agit pas de mémoriser les caractéristiques de l’ensemble des données mais d’apprendre des généralisations qui peuvent être appliqués à d’autres modèles

## Étape 4 : Visualisation des résultats

La visualisation des données est un outil précieux pour l’interprétation des données. Afin de rendre notre analyse plus accessible et compréhensible, nous avons décidé d’élaborer des reportings par l’outil Power BI afin de mieux représenter les résultats d’analyse.





Sources

* https://www.kaggle.com/datasets/imdevskp/hiv-aids-dataset/discussion
* https://www.kaggle.com/datasets/suvojithaldar/aids-dataset
* <https://www.kaggle.com/datasets/programmerrdai/hiv-aids>
* <https://www.kaggle.com/datasets/imdevskp/hiv-aids-dataset?resource=download>
* <https://www.kaggle.com/datasets/neharautela/hivaids?select=antiviral_therapy_among_people.csv>
* <https://www.who.int/data/collections>
* <https://www.who.int/data/gho/data/indicators/indicator-details/GHO/new-hiv-infections-(per-1000-uninfected-population)>
* <https://www.who.int/data/gho/data/indicators/indicator-details/GHO/estimated-antiretroviral-therapy-coverage-among-people-living-with-hiv-(-)>
* <https://www.who.int/data/gho/data/indicators/indicator-details/GHO/reported-number-of-people-receiving-antiretroviral-therapy>