Alain AL-SHIKHLEY

PROJET SERVIER

Dans ce rapport je vous explique ce que j'ai fait sur un notebook. Cependant j'ai en réalité fait les traitement sur Airflow (vous trouverez mon code airflow dans le dossier dags du projet).

1/Dans un premier temps j'ai récupéré les fichier CSV sous le format Dataframe

```
import pandas as pd
import numpy as np
```

```
pd.set_option("max_colwidth", None)
```

```
df_drugs = pd.read_csv("drugs.csv", header=0, index_col=None, sep=',', encoding="utf-8")
df_clinical = pd.read_csv("clinical_trials.csv", header=0, index_col=None, sep=',',  encoding="utf-8")
df_pubmed = pd.read_csv("pubmed.csv", header=0, index_col=None, sep=',')
```

2/J'ai ensuite regardé le type des colonnes de chaque Dataframe

```
df_drugs.dtypes
```

```
atccode    object
drug       object
dtype: object
```

```
df_pubmed.dtypes
```

```
id          int64
title      object
date       object
journal    object
dtype: object
```

```
df_clinical.dtypes
```

```
id                 object
scientific_title   object
date               object
journal            object
dtype: object
```

3/Par la suite j’ai affiché le contenu de chaque Dataframe pour voir s’il y avait du nettoyage à faire

```
df_clinical.head(10)
```

	id	scientific_title	date	journal
0	NCT01967433	Use of Diphenhydramine as an Adjunctive Sedative for Colonoscopy in Patients Chronically on Opioids	1 January 2020	Journal of emergency nursing
1	NCT04189588	Phase 2 Study IV QUZYTTIR™ (Cetirizine Hydrochloride Injection) vs V Diphenhydramine	1 January 2020	Journal of emergency nursing
2	NCT04237090		1 January 2020	Journal of emergency nursing
3	NCT04237091	Feasibility of a Randomized Controlled Clinical Trial Comparing the Use of Cetirizine to Replace Diphenhydramine in the Prevention of Reactions Related to Paclitaxel	1 January 2020	Journal of emergency nursing
4	NCT04153396	Preemptive Infiltration With Betamethasone and Ropivacaine for Postoperative Pain in Laminoplasty or \xc3\xb1 Laminectomy	1 January 2020	Hôpitaux Universitaires de Genève
5	NCT03490942	Glucagon Infusion in T1D Patients With Recurrent Severe Hypoglycemia: Effects on Counter-Regulatory Responses	25/05/2020	NaN
6	NaN	Glucagon Infusion in T1D Patients With Recurrent Severe Hypoglycemia: Effects on Counter-Regulatory Responses	25/05/2020	Journal of emergency nursing
7	NCT04188184	Tranexamic Acid Versus Epinephrine During Exploratory Tympanotomy	27 April 2020	Journal of emergency nursing\xc3\x28

```
df_drugs.head(10)
```

	atccode	drug
0	A04AD	DIPHENHYDRAMINE
1	S03AA	TETRACYCLINE
2	V03AB	ETHANOL
3	A03BA	ATROPINE
4	A01AD	EPINEPHRINE
5	6302001	ISOPRENALINE
6	R01AD	BETAMETHASONE

```
df_pubmed.head(10)
```

	id	title	date	journal
0	1	A 44-year-old man with erythema of the face diphenhydramine, neck, and chest, weakness, and palpitations	01/01/2019	Journal of emergency nursing
1	2	An evaluation of benadryl, pyribenzamine, and other so-called diphenhydramine antihistaminic drugs in the treatment of allergy.	01/01/2019	Journal of emergency nursing
2	3	Diphenhydramine hydrochloride helps symptoms of ciguatera fish poisoning.	02/01/2019	The Journal of pediatrics
3	4	Tetracycline Resistance Patterns of Lactobacillus buchneri Group Strains.	01/01/2020	Journal of food protection
4	5	Appositional Tetracycline bone formation rates in the Beagle.	02/01/2020	American journal of veterinary research
5	6	Rapid reacquisition of contextual fear following extinction in mice: effects of amount of extinction, tetracycline acute ethanol withdrawal, and ethanol intoxication.	2020-01-01	Psychopharmacology
6	7	The High Cost of Epinephrine Autoinjectors and Possible Alternatives.	01/02/2020	The journal of allergy and clinical immunology. In practice
7	8	Time to epinephrine treatment is associated with the risk of mortality in children who achieve sustained ROSC after traumatic out-of-hospital cardiac arrest.	01/03/2020	The journal of allergy and clinical immunology. In practice

4/En ce qui concerne les données du clinical_trials, j’ai constaté que :
- la colonne date était de type object. De plus les dates avaient des formats différents.
- la ligne 6 et 7 ont des valeurs NaN et semble être les mêmes lignes car elles ont le même scientific_title et j’ai supposé qu’un scientific_title ne peut appartenir qu’à un journal
- le journal de la dernière ligne contient des caractères en trop
- une ligne n’avait aucun scientific_title

J’ai donc nettoyé ce Dataframe.

```
df_clinical['date'] = pd.to_datetime(df_clinical['date'].str.strip('"')).dt.strftime('%d/%m/%Y')
```

```
df_clinical.drop(2 , inplace=True)
```

```
df_clinical.reset_index(inplace=True, drop=True)
```

```
df_clinical = df_clinical[df_clinical['id'].notna()]
```

```
df_clinical.iloc[4, df_clinical.columns.get_loc('journal')] = 'Journal of emergency nursing'
```

```
df_clinical.iloc[5, df_clinical.columns.get_loc('journal')] = 'Journal of emergency nursing'
```

```
df_clinical.reset_index(inplace=True, drop=True)
```

```
df_clinical.head(10)
```

	id	scientific_title	date	journal
0	NCT01967433	Use of Diphenhydramine as an Adjunctive Sedative for Colonoscopy in Patients Chronically on Opioids	01/01/2020	Journal of emergency nursing
1	NCT04189588	Phase 2 Study IV QUZYTTIR™ (Cetirizine Hydrochloride Injection) vs V Diphenhydramine	01/01/2020	Journal of emergency nursing
2	NCT04237091	Feasibility of a Randomized Controlled Clinical Trial Comparing the Use of Cetirizine to Replace Diphenhydramine in the Prevention of Reactions Related to Paclitaxel	01/01/2020	Journal of emergency nursing
3	NCT04153396	Preemptive Infiltration With Betamethasone and Ropivacaine for Postoperative Pain in Laminoplasty or \xc3\xb1 Laminectomy	01/01/2020	Hôpitaux Universitaires de Genève
4	NCT03490942	Glucagon Infusion in T1D Patients With Recurrent Severe Hypoglycemia: Effects on Counter-Regulatory Responses	25/05/2020	Journal of emergency nursing
5	NCT04188184	Tranexamic Acid Versus Epinephrine During Exploratory Tympanotomy	27/04/2020	Journal of emergency nursing

5/En ce qui concerne les drugs je me suis contenté de retirer la ligne 6 (indice 5) qui contenait un ID qui n'était pas un ATC code.

```
df_drugs.drop(5 , inplace=True)
```

```
df_drugs.reset_index(inplace=True, drop=True)
```

```
df_drugs.head(10)
```

	atccode	drug
0	A04AD	DIPHENHYDRAMINE
1	S03AA	TETRACYCLINE
2	V03AB	ETHANOL
3	A03BA	ATROPINE
4	A01AD	EPINEPHRINE
5	R01AD	BETAMETHASONE

6/Ici j'ai juste modifié le format des dates pour qu'elles aient le même format dans chaque Dataframe

```
df_pubmed['date'] = pd.to_datetime(df_pubmed['date'].str.strip('"')).dt.strftime('%d/%m/%Y')
```

```
df_pubmed.head(10)
```

	id	title	date	journal
0	1	A 44-year-old man with erythema of the face diphenhydramine, neck, and chest, weakness, and palpitations	01/01/2019	Journal of emergency nursing
1	2	An evaluation of benadryl, pyribenzamine, and other so-called diphenhydramine antihistaminic drugs in the treatment of allergy.	01/01/2019	Journal of emergency nursing
2	3	Diphenhydramine hydrochloride helps symptoms of ciguatera fish poisoning.	01/02/2019	The Journal of pediatrics
3	4	Tetracycline Resistance Patterns of Lactobacillus buchneri Group Strains.	01/01/2020	Journal of food protection
4	5	Appositional Tetracycline bone formation rates in the Beagle.	01/02/2020	American journal of veterinary research
5	6	Rapid reacquisition of contextual fear following extinction in mice: effects of amount of extinction, tetracycline acute ethanol withdrawal, and ethanol intoxication.	01/01/2020	Psychopharmacology
6	7	The High Cost of Epinephrine Autoinjectors and Possible Alternatives.	02/01/2020	The journal of allergy and clinical immunology. In practice
7	8	Time to epinephrine treatment is associated with the risk of mortality in children who achieve sustained ROSC after traumatic out-of-hospital cardiac arrest.	03/01/2020	The journal of allergy and clinical immunology. In practice

J'ai rencontré un problème au niveau de la création de JSON. En effet je ne savais pas comment réaliser le JSON en partant de 3 fichiers CSV.

Airflow

Sur Airflow j'ai réalisé les premières tâches de nettoyage, cependant j'ai rencontré quelques soucis. En effet je n'arrive pas à accéder à mes fichiers CSV déposés sur mon local. Je pense que c'est parce que j'ai lancé Airflow en passant par des containers Docker et, par conséquent, pour relier le container à mon local, il faut faire une manipulation.