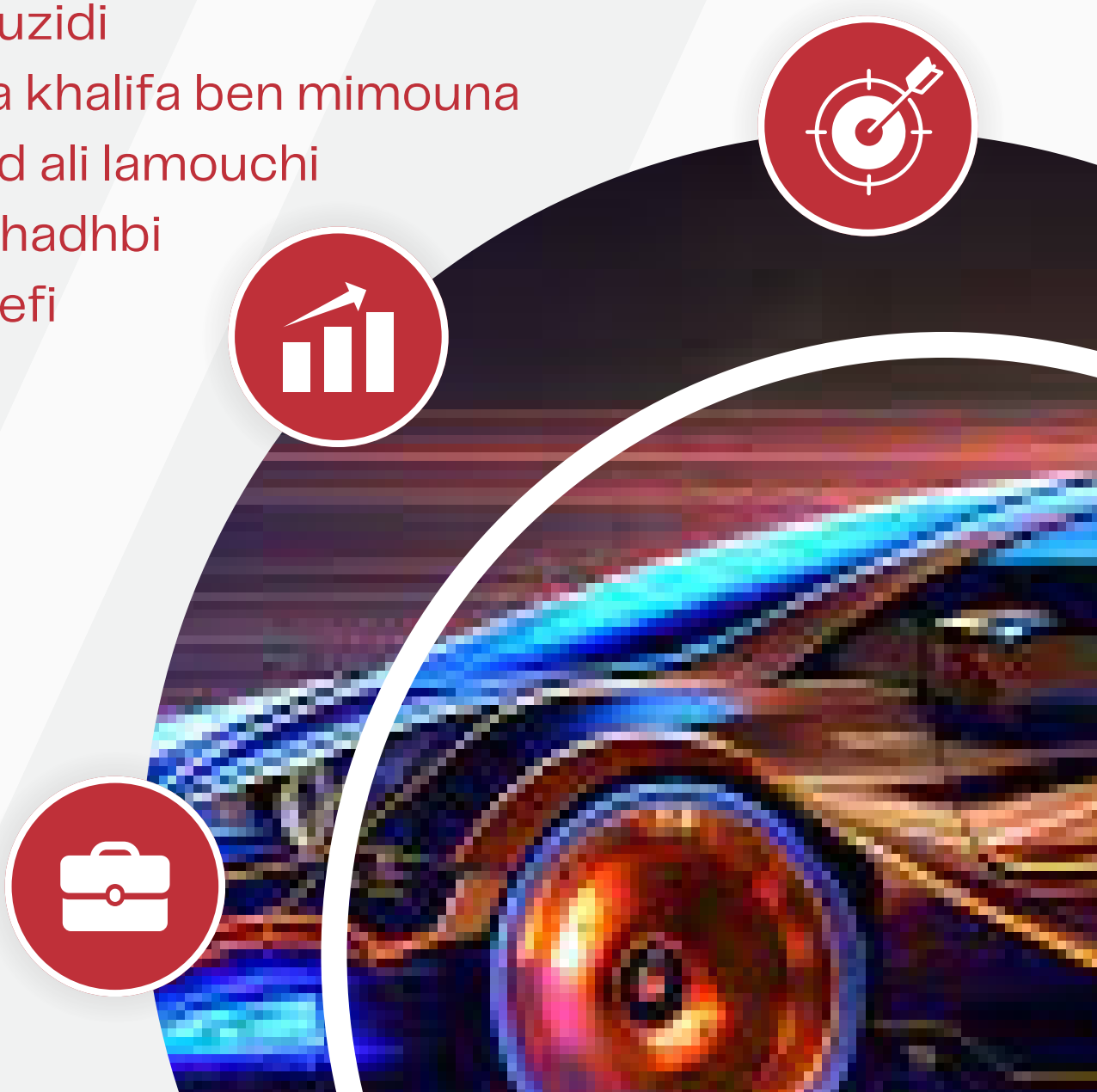
Data science project

Exploration et analyse de données

Fatma mrabti
Dorra Bouzidi
Oussema khalifa ben mimouna
Mohamed ali lamouchi
Fatma mhadhbi
Aymen kefi



Sommaire

1-Objectif

L'objectif principal du code est de combiner les données dispersées dans plusieurs fichiers CSV en une seule entité structurée, représentée sous la forme d'un DataFrame pandas.

2- Technologies utilisées

Environnement de developpement

Bibliothéque

Langage de programmation

Python

3- Extraction et nettoyage des données

3.1 Extraction des donnés

Dans cette partie on a extrait les données qu'on a selectionner selon les categories

Script

```
import pandas as pd
import glob

#usagers
# Get a list of all CSV files in the directory
files = glob.glob('DataScienceProject/Data/usagers*.csv')

# Initialize an empty DataFrame to store the combined data
combined_data = pd.DataFrame()
print(files)
# Loop through each file and append its data to the combined DataFrame
for file in files:
    # Read the first row of the file to infer data types
    dtypes = pd.read_csv(file, nrows=1).dtypes.to_dict()

    # Read the entire CSV file using inferred data types
    df = pd.read_csv(file, dtype=dtypes, encoding='UTF-8', sep=';', quotechar='"')
    combined_data = pd.concat([combined_data, df], ignore_index=True)

# Now 'combined_data' contains data from all CSV files
combined_data.to_excel('DataScienceProject/combinedData/usagers.xlsx', index=False)
combined_data.head
```

Details du script

1- Identification des fichiers

- Utilise le module glob pour obtenir une liste de chemins de fichiers pour les fichiers CSV correspondant à un motif spécifié.

2- Combinaison des données

- Utilise une compréhension de liste pour lire et concaténer les données de chaque fichier CSV dans un seul DataFrame pandas (combined_data). Spécifie l'encodage UTF-8, le point-virgule (comme séparateur et les guillemets doubles (") comme caractère de citation lors de la lecture des fichiers.

3- Sauvegarde des données

- Enregistre les données combinées dans un fichier Excel nommé 'usagers.xlsx' dans un répertoire désigné ('DataScienceProject/combinedData/'). Le paramètre index=False garantit que l'index du DataFrame est exclu du fichier Excel.

3- Résultat et conclusion

3.1 Résultat

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Num_Acc	id_usager	d_vehicule	num_veh	place	catu	grav	sexe	an_nais	trajet	secu1	secu2	secu3	locp	actp	etatp
2	2,02E+11	267 638	201 764	B01	1	1	3	1	2000	1	0	9	-1	0	0	-1
3	2,02E+11	267 639	201 765	A01	1	1	1	1	1978	1	1	-1	-1	0	0	-1
4	2,02E+11	267 636	201 762	A01	1	1	4	1	1983	0	1	-1	-1	0	0	-1
5	2,02E+11	267 637	201 763	B01	1	1	3	1	1993	0	1	-1	-1	0	0	-1
6	2,02E+11	267 634	201 761	A01	1	1	1	1	1995	1	1	0	-1	0	0	-1
7	2,02E+11	267 635	201 761	A01	10	3	3	2	1959	4	0	-1	-1	3	3	1
8	2,02E+11	267 631	201 758	A01	1	1	1	1	2000	-1	-1	0	-1	-1	-1	-1
9	2,02E+11	267 632	201 759	D01	1	1	2	2	2014	5	0	-1	-1	-1	-1	-1
10	2,02E+11	267 627	201 754	A01	1	1	4	2	1997	1	1	-1	-1	-1	-1	-1
11	2,02E+11	267 628	201 755	Z01	1	1	-1	-1		-1	8	-1	-1	-1	-1	-1
12	2,02E+11	267 625	201 752	B01	1	1	4	1	2009	2	0	-1	-1	0	0	-1
13	2,02E+11	267 626	201 753	A01	1	1	1	1	1976	5	1	-1	-1	0	0	-1
14	2,02E+11	267 622	201 750	B01	4	2	4	2	2002	0	1	0	-1	-1	-1	-1
15	2,02E+11	267 623	201 750	B01	1	1	1	1	2001	5	1	5	-1	-1	-1	-1
16	2,02E+11	267 624	201 751	A01	1	1	4	1	1991	0	1	5	-1	-1	-1	-1
17	2,02E+11	267 619	201 748	B01	1	1	3	1	1972	5	2	-1	-1	0	0	-1
18	2,02E+11	267 620	201 748	B01	2	2	4	2	1984	5	8	-1	-1	0	0	-1
19	2,02E+11	267 621	201 749	A01	1	1	1	1	1971	5	1	-1	-1	0	0	-1
20	2,02E+11	267 616	201 746	A01	1	1	1	1	1981	1	1	0	-1	-1	-1	-1
21	2,02E+11	267 617	201 747	B01	2	2	4	2	1936	5	1	0	-1	-1	-1	-1
22	2,02E+11	267 618	201 747	B01	1	1	1	1	1935	5	1	0	-1	-1	-1	-1
23	2,02E+11	267 611	201 744	A01	1	1	1	1	1976	5	8	8	-1	-1	-1	-1
24	2,02E+11	267 612	201 745	B01	4	2	4	2	1982	0	8	8	-1	-1	-1	-1
25	2,02E+11	267 613	201 745	B01	8	2	4	2	1960	5	8	8	-1	-1	-1	-1
26	2,02E+11	267 614	201 745	B01	8	2	4	2	1959	5	8	8	-1	-1	-1	-1
27	2,02E+11	267 615	201 745	B01	1	1	4	2	1994	4	1	8	-1	-1	-1	-1
28	2,02E+11	267 609	201 742	A01	1	1	1	1	1951	5	1	8	-1	-1	-1	-1
29	2,02E+11	267 610	201 743	B01	1	1	4	1	1989	0	8	0	-1	-1	-1	-1

3.2 Conclusion

En résumé, le code est une solution efficace et concise pour agréger des données provenant de plusieurs fichiers CSV, ce qui le rend adapté aux tâches de prétraitement des données dans notre projet.