Scrapping des données

Elaboré par :

- Rekik Wiem
- Dhaouadi Mohamed Amine
- Dghaies Yasser
- Ferjani Mohamed Iheb
- Battikh Anis
- Achour Nouha

I. Introduction

Ce script Python utilise les bibliothèques Selenium et BeautifulSoup pour extraire des données à partir d'une page Web contenant un tableau de données. L'objectif principal est de récupérer les informations de la page sur les collisions de véhicules à New York depuis le portail de données ouvert de la ville.

II. Données Source

Le tableau de données présente les accidents liés aux collisions de véhicules automobiles à New York, fournissant des détails sur chaque incident spécifique. Chaque ligne de ce tableau représente un événement de collision distinct, présentant des informations telles que la date, l'heure, l'emplacement, les véhicules impliqués, et d'autres détails pertinents. Ces données sont compilées à partir des rapports officiels de la police, offrant une vue d'ensemble complète et détaillée des collisions de véhicules à moteur dans la région

III. Outils

IV. Script

```
: from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.chrome.options import Options
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.common.exceptions import StaleElementReferenceException
from bs4 import BeautifulSoup
import pandas as pd
import time

def initialize_driver():
    options = Options()
    options.add_argument('--headless')
    return webdriver.Chrome(options=options)

def wait_for_element(driver, locator, timeout=10):
    return WebDriverWait(driver, timeout).until(EC.presence_of_element_located(locator))

def click_and_retry(driver, element, locator, retry_attempts=3):
    for _ in range(retry_attempts):
        try:
            element.click()
            return True
        except StaleElementReferenceException:
            print("Stale Element Reference Exception. Retrying...")
            element = wait_for_element(driver, locator)
        except Exception as e:
            print(f"An error occurred: {str(e)}")
            return False
    return False

def scrape_table_data(driver, table_locator):
    table = wait_for_element(driver, table_locator)
    table_html = table.get_attribute('outerHTML')
    soup = BeautifulSoup(table_html, 'html.parser')
    html_table = soup.find('table')
    return pd.read_html(str(html_table))[0]
```

```python
def main():
    url = "https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95"
    driver = initialize_driver()

    try:
        driver.get(url)
        driver.implicitly_wait(10)

        table_locator = (By.CLASS_NAME, 'socrata-table.frozen-columns')

        if not (table := wait_for_element(driver, table_locator)):
            print("No table  found on the page with class 'socrata-table frozen-columns'.")
            return

        dfs = []

        total_pages = 500

        for page in range(1, total_pages + 1):
            next_button_locator = (By.XPATH, "//button[@class='pager-button-next']")
            next_button = wait_for_element(driver, next_button_locator)

            if not click_and_retry(driver, next_button, next_button_locator):
                print("Failed to click next button. Exiting.")
                break

            time.sleep(3)

            df = scrape_table_data(driver, table_locator)

            print(df.head())

            dfs.append(df)

        result_df = pd.concat(dfs, ignore_index=True)

        result_df = result_df.dropna(axis=1, how='all')
```

ChatGPT - Google Chrome

```python
        result_df = result_df.dropna(axis=1, how='all')

        result_df.to_excel("DataCollisions.xlsx", index=False)
        print("Cleaned data successfully exported DataCollisions.xlsx")

    except Exception as e:
        print(f"An error occurred: {str(e)}")

    finally:
        driver.quit()

if __name__ == "__main__":
    main()
```

```
2                    NaN  HARLEM RIVER DRIVE RAMP              NaN
```

Étape 1 : Initialisation du WebDriver

Le script débute en initialisant un WebDriver Chrome en mode headless à l'aide de la bibliothèque Selenium. Cette étape permet une navigation automatisée sur la page Web sans interface graphique.

Étape 2 : Navigation vers la Page Cible

Le WebDriver accède à la page de données des collisions de véhicules à New York en utilisant l'URL spécifié. Une attente implicite garantit que la page est entièrement chargée avant de procéder à la collecte des données.

Étape 3 : Scraping de Données

Le script identifie le tableau de données sur la page (repéré par la classe CSS 'socrata-table.frozen-columns') et commence le processus d'extraction des informations. La pagination automatique est gérée pour collecter les données de plusieurs pages.

Étape 4 : Gestion des Éléments Dynamiques

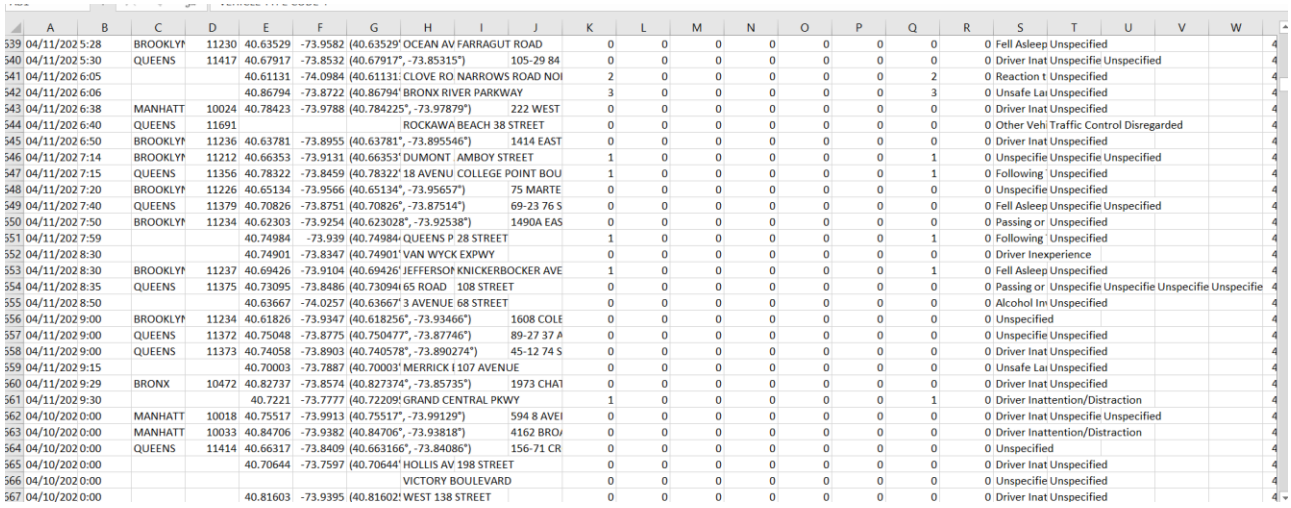
Les mécanismes de gestion des éléments dynamiques, tels que l'utilisation de l'attente explicite et la gestion des exceptions Stale Element Reference, sont mis en œuvre pour assurer la fiabilité du scraping, même sur des pages dynamiques.

Étape 5 : Nettoyage des Données

Une fois les données collectées, le script les nettoie en supprimant les colonnes qui ne contiennent que des valeurs nulles, optimisant ainsi la qualité et la lisibilité du jeu de données.

Étape 6 : Exportation des Données Nettoyées

Les données nettoyées sont ensuite exportées vers un fichier Excel, facilitant leur utilisation dans des analyses ultérieures ou d'autres applications.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	
539	04/11/202	5:28	BROOKLYN	11230	40.63529	-73.9582	(40.63529	OCEAN AV	FARRAGUT	ROAD	0	0	0	0	0	0	0	0	Fell Asleep	Unspecified				4
540	04/11/202	5:30	QUEENS	11417	40.67917	-73.8532	(40.67917,	-73.85315")	105-29 84		0	0	0	0	0	0	0	0	Driver Inat	Unspecifie	Unspecified			4
541	04/11/202	6:05			40.61131	-74.0984	(40.61131	CLOVE RO	NARROWS	ROAD NOI	2	0	0	0	0	0	2	0	Reaction t	Unspecified				4
542	04/11/202	6:06			40.86794	-73.8722	(40.86794	BRONX RIVER PARKWAY			3	0	0	0	0	0	3	0	Unsafe La	Unspecified				4
543	04/11/202	6:38	MANHATT	10024	40.78423	-73.9788	(40.784225",	-73.97879")	222 WEST		0	0	0	0	0	0	0	0	Driver Inat	Unspecified				4
544	04/11/202	6:40	QUEENS	11691				ROCKAWA	BEACH 38	STREET	0	0	0	0	0	0	0	0	Other Vehi	Traffic Control Disregarded				4
545	04/11/202	6:50	BROOKLYN	11236	40.63781	-73.8955	(40.63781°,	-73.895546")	1414 EAST		0	0	0	0	0	0	0	0	Driver Inat	Unspecified				4
546	04/11/202	7:14	BROOKLYN	11212	40.66353	-73.9131	(40.66353	DUMONT	AMBOY	STREET	1	0	0	0	0	0	1	0	Unspecifie	Unspecifie	Unspecified			4
547	04/11/202	7:15	QUEENS	11356	40.78322	-73.8459	(40.78322	18 AVENU	COLLEGE	POINT BOU	1	0	0	0	0	0	1	0	Following	Unspecified				4
548	04/11/202	7:20	BROOKLYN	11226	40.65134	-73.9566	(40.65134",	-73.95657")	75 MARTE		0	0	0	0	0	0	0	0	Unspecifie	Unspecified				4
549	04/11/202	7:40	QUEENS	11379	40.70826	-73.8751	(40.70826°,	-73.87514")	69-23 76 S		0	0	0	0	0	0	0	0	Fell Asleep	Unspecifie	Unspecified			4
550	04/11/202	7:50	BROOKLYN	11234	40.62303	-73.9254	(40.623028",	-73.92538")	1490A EAS		0	0	0	0	0	0	0	0	Passing or	Unspecified				4
551	04/11/202	7:59			40.74984	-73.939	(40.74984	QUEENS P	28 STREET		1	0	0	0	0	0	1	0	Following	Unspecified				4
552	04/11/202	8:30			40.74901	-73.8347	(40.74901	VAN WYCK	EXPWY		0	0	0	0	0	0	0	0	Driver Inexperience					4
553	04/11/202	8:30	BROOKLYN	11237	40.69426	-73.9104	(40.69426	JEFFERSO	KNICKERBOCKER	AVE	1	0	0	0	0	0	1	0	Fell Asleep	Unspecified				4
554	04/11/202	8:35	QUEENS	11375	40.73095	-73.8486	(40.73094	65 ROAD	108 STREET		0	0	0	0	0	0	0	0	Passing or	Unspecifie	Unspecifie	Unspecifie	Unspecifie	4
555	04/11/202	8:50			40.63667	-74.0257	(40.63667	3 AVENUE	68 STREET		0	0	0	0	0	0	0	0	Alcohol Im	Unspecified				4
556	04/11/202	9:00	BROOKLYN	11234	40.61826	-73.9347	(40.618256",	-73.93466")	1608 COLE		0	0	0	0	0	0	0	0	Unspecified					4
557	04/11/202	9:00	QUEENS	11372	40.75048	-73.8775	(40.750477",	-73.87746")	89-27 37 A		0	0	0	0	0	0	0	0	Unspecifie	Unspecified				4
558	04/11/202	9:00	QUEENS	11373	40.74058	-73.8903	(40.740578",	-73.890274")	45-12 74 S		0	0	0	0	0	0	0	0	Driver Inat	Unspecified				4
559	04/11/202	9:15			40.70003	-73.7887	(40.70003	MERRICK I	107 AVENUE		0	0	0	0	0	0	0	0	Unsafe La	Unspecified				4
560	04/11/202	9:29	BRONX	10472	40.82737	-73.8574	(40.827374",	-73.85735")	1973 CHAT		0	0	0	0	0	0	0	0	Driver Inat	Unspecified				4
561	04/11/202	9:30			40.7221	-73.7777	(40.72209	GRAND CENTRAL	PKWY		1	0	0	0	0	0	1	0	Driver Inattention/Distraction					4
562	04/10/202	0:00	MANHATT	10018	40.75517	-73.9913	(40.75517",	-73.99129")	594 8 AVE		0	0	0	0	0	0	0	0	Driver Inat	Unspecifie	Unspecified			4
563	04/10/202	0:00	MANHATT	10033	40.84706	-73.9382	(40.84706°,	-73.93818")	4162 BRO		0	0	0	0	0	0	0	0	Driver Inattention/Distraction					4
564	04/10/202	0:00	QUEENS	11414	40.66317	-73.8409	(40.663166",	-73.84086")	156-71 CR		0	0	0	0	0	0	0	0	Unspecified					4
565	04/10/202	0:00			40.70644	-73.7597	(40.70644	HOLLIS AV	198 STREET		0	0	0	0	0	0	0	0	Driver Inat	Unspecified				4
566	04/10/202	0:00						VICTORY	BOULEVARD		0	0	0	0	0	0	0	0	Unspecifie	Unspecified				4
567	04/10/202	0:00			40.81603	-73.9395	(40.81602	WEST 138	STREET		0	0	0	0	0	0	0	0	Driver Inat	Unspecified				4

#	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
6476	03/20/202	14:00	BROOKLYN	11226	40.65279	-73.9469	(40.65279	NEW YORK	LINDEN BOULEVARD		1	0	0	0	1	0	0	0	Unspecified		
6477	03/20/202	14:00	BROOKLYN	11236	40.64141	-73.9123	(40.64140	FARRAGUT	EAST 86 STREET		1	0	0	0	0	0	1	0	Unspecific	Unspecified	
6478	03/20/202	14:05	BRONX	10467				EAST FORC	PELHAM PARKWAY		1	0	0	0	1	0	0	0	Failure to	Unspecified	
6479	03/20/202	14:08	QUEENS	11435	40.69759	-73.8115	(40.69759	94 AVENU	138 PLACE		1	0	0	0	0	0	1	0	Passing or	Unspecified	
6480	03/20/202	14:30	QUEENS	11434	40.68257	-73.7927	(40.68256	116 AVENI	SUTPHIN BOULEVARD		0	0	0	0	0	0	0	0	Unspecific	Unspecified	
6481	03/20/202	14:30			40.70042	-73.9518	(40.70041	MADISON	AVENUE		0	0	0	0	0	0	0	0	Lost Comr	Unspecific	Unspecified
6482	03/20/202	14:34	QUEENS	11354	40.75586	-73.8362	(40.75586°, -73.83615°)			131-04 40	0	0	0	0	0	0	0	0	Turning im	Unspecified	
6483	03/20/202	14:36	QUEENS	11373	40.7418	-73.8788	(40.7418°, -73.8788°)			86-12 WHI	1	0	0	0	0	0	1	0	Turning im	Unspecified	
6484	03/20/202	14:38	MANHATT	10028	0	0	(0.0°, 0.0°)			1536 3 AVI	0	0	0	0	0	0	0	0	Backing Ur	Unspecified	
6485	03/20/202	14:45			40.68316	-73.9381	(40.68316	HALSEY	STREET		1	0	1	0	0	0	0	0	Driver Inattention/Distraction		
6486	03/20/202	14:48	BROOKLYN	11229	0	0	(0.0°, 0.0°)			2069 EAST	0	0	0	0	0	0	0	0	Backing Ur	Unspecified	
6487	03/20/202	14:48	MANHATT	10017	40.75374	-73.9678	(40.755742°, -73.96783°)			348 EAST 4	0	0	0	0	0	0	0	0	Driver Inat	Unspecified	
6488	03/20/202	14:49	BRONX	10467	0	0	(0.0°, 0.0°)	BRONX PA	WAINING AVENUE		0	0	0	0	0	0	0	0	Driver Inat	Unspecified	
6489	03/20/202	14:50	BRONX	10472	40.82975	-73.8737	(40.82975	HARROD A	WESTCHESTER AVENU		0	0	0	0	0	0	0	0	Unspecified		
6490	03/20/202	14:56	QUEENS	11432	0	0	(0.0°, 0.0°)	JAMAICA A	168 PLACE		0	0	0	0	0	0	0	0	Driver Inat	Unspecified	
6491	03/20/202	15:00	BRONX	10460	40.83824	-73.8767	(40.83823	EAST 177	BRONX PARK AVENUE		0	0	0	0	0	0	0	0	Driver Inat	Driver Inattention/Dist	
6492	03/20/202	15:00	BROOKLYN	11221	0	0	(0.0°, 0.0°)			860 MONR	0	0	0	0	0	0	0	0	Driver Inat	Unspecified	
6493	03/20/202	15:00			40.88042	-73.8443	(40.88042	EAST 223	STREET		0	0	0	0	0	0	0	0	Unsafe Sp	Unspecified	
6494	03/20/202	15:04			40.70245	-73.8594	(40.70245	JACKIE ROBINSON	PKWY		0	0	0	0	0	0	0	0	Reaction to	Unknown Vehicle	
6495	03/20/202	15:05	BRONX	10466	40.88721	-73.8608	(40.887207°, -73.86082°)			3949 WHIT	2	0	2	0	0	0	0	0	Driver Inexperience		
6496	03/20/202	15:12			40.67269	-73.7354	(40.67268	LAURELTON	PARKWAY		0	0	0	0	0	0	0	0	Unspecific	Unspecified	
6497	03/20/202	15:25	BRONX	10455	40.80988	-73.9091	(40.80988	EAST 149	BRUCKNER BOULEVA		1	0	0	0	0	0	1	0	Unspecific	Unspecified	
6498	03/20/202	15:25	QUEENS	11354	40.76582	-73.8237	(40.76582	35 AVENU	PARSONS BOULEVARD		1	0	1	0	0	0	0	0	Failure to	Yield Right-of-Way	
6499	03/20/202	15:30	QUEENS	11372	40.7487	-73.874	(40.7487°, -73.87395°)			92-17 ROC	1	0	0	0	1	0	0	0	Unspecified		
6500	03/20/202	15:31	BROOKLYN	11217	0	0	(0.0°, 0.0°)			188 FLATB	0	0	0	0	0	0	0	0	Passenger	View Obstructed/Limit	
6501	03/20/202	15:40			0	0	(0.0°, 0.0°)	CHRYSTIE	STREET		1	0	0	0	0	0	1	0	Physical Di	Unspecified	