



DATA SCIENCE: PHASE 1: WEB SCRAPPING DES DONNEES

Réalisé par :

Naoures Hidri
Mohamed Aziz Ghalleb
Alaa Galai
Malek Kouissi
Houcem Ghiloufi

I. Problématique

La croissance continue du nombre de véhicules et de conducteurs suscite une préoccupation croissante dans chaque pays. Dans ce contexte, il devient crucial d'évaluer le taux d'accidents routiers et d'identifier les principales causes sous-jacentes. Cette problématique souligne l'importance de comprendre les facteurs contribuant aux accidents, ouvrant ainsi la voie à des analyses approfondies et à la mise en place de mesures visant à améliorer la sécurité routière.

II. Objectif

L'objectif de cette phase consiste à recueillir des données au Los Angeles à partir du site Web dédié aux statistiques sur les collisions de la circulation. À l'aide d'un script écrit en Python, nous utilisons la bibliothèque "Selenium" pour extraire ces données et les sauvegarder sous forme de fichiers CSV.

III. Solution

Notre approche utilise un script Python qui automatisera le processus de collecte de données à partir du site spécifié. En utilisant la bibliothèque "Selenium", le script parcourt le code du HTML et JavaScript du site pour extraire les informations pertinentes liées aux collisions de la circulation. Ces données sont ensuite organisées et stockées dans des fichiers CSV, offrant ainsi une méthode efficace et automatisée pour analyser et traiter les statistiques sur les accidents routiers au Los Angeles.

IV. Outils utilisés

Les outils utilisés au cours de cette phases sont les suivants :

☛ Langage de programmation :



☛ Les bibliothèques :



☛ Navigateur Web :



☛ Environnement de développement et IDE :



V. Script Python

```
In [ ]: from selenium import webdriver
        from selenium.webdriver.common.by import By
        from selenium.webdriver.support.ui import WebDriverWait
        from selenium.webdriver.support import expected_conditions as EC
        import os
        import time

        # Set up the WebDriver (in this example, we'll use Chrome)
        driver = webdriver.Chrome()

        # URL of the website
        url = "https://data.lacity.org/Public-Safety/Traffic-Accidents-by-date/2mzm-av8t"

        # Open the website
        driver.get(url)

        try:
            # Find and click on the "Export" button
            export_button = WebDriverWait(driver, 10).until(EC.element_to_be_clickable((By.CSS_SELECTOR, "button.btn.btn-simple.btn-sm.d
            export_button.click()

            # Find and click on the "CSV" button within the pop-up
            csv_button = WebDriverWait(driver, 10).until(EC.element_to_be_clickable((By.CSS_SELECTOR, "ul.featured-download-links li.dow
            csv_button.click()

            # Give some time for the download to start
            time.sleep(5) # Adjust the time as needed

            # Specify the default download directory
            default_download_directory = r'C:\DataScience'
```

```

# Give some time for the download to start
time.sleep(5) # Adjust the time as needed

# Specify the default download directory
default_download_directory = r'C:\DataScience'

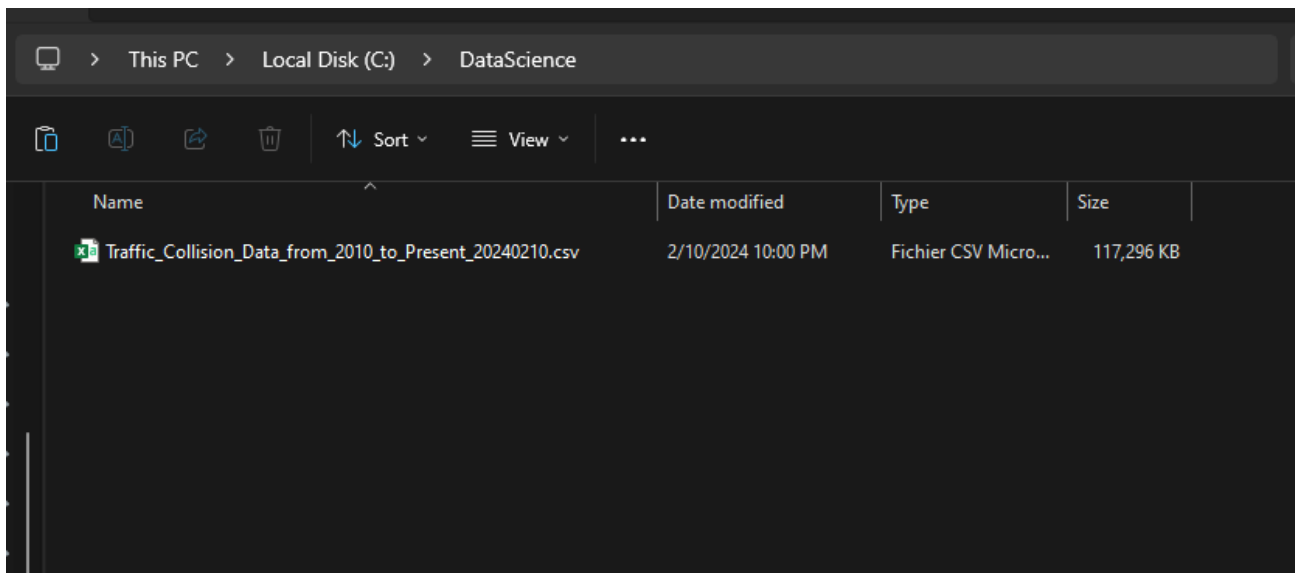
# Check if the file has been downloaded
while True:
    # Check if the file exists in the download directory
    files = os.listdir(default_download_directory)
    if any(file.endswith('.csv') for file in files):
        time.sleep(2) # Wait for a while before checking again
        print("CSV file has been downloaded successfully.")
    else:
        print("CSV file has been downloaded successfully.")
        break


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

# Close the browser
driver.quit()

```

VI. Résultat



Name	Date modified	Type	Size
 Traffic_Collision_Data_from_2010_to_Present_20240210.csv	2/10/2024 10:00 PM	Fichier CSV Micro...	117,296 KB

VII. Description des données

Le répertoire des accidents liés aux collisions de véhicules automobiles renferme des informations détaillées sur chaque incident. Chaque ligne représente un accident distinct. Ce tableau de données

recense toutes les collisions de véhicules signalées par les autorités policières de Los Angeles entre 2010 et Janvier 2024.

VIII. Description des variables :

Dans cette section, nous examinons les variables extraites en analysant leurs valeurs pour accéder à des observations pertinentes et prendre des décisions informées en se basant sur les données extraites.

1. **DR Number (Numéro de rapport)** : Ceci est un identifiant unique pour chaque rapport d'accident. Il peut être utilisé pour suivre et référencer spécifiquement chaque incident.
2. **Date Reported (Date de signalement)** : Indique quand l'accident a été signalé. Il peut y avoir un délai entre la date de l'accident et la date à laquelle il a été signalé.
3. **Date Occurred (Date de l'incident)** : La date à laquelle l'accident s'est réellement produit. Cela peut être crucial pour analyser les tendances temporelles des accidents.
4. **Time Occurred (Heure de l'incident)** : L'heure à laquelle l'accident s'est produit. Cela peut aider à identifier les périodes de la journée où les accidents sont les plus fréquents.
5. **Area ID (Identifiant de la zone) / Area Name (Nom de la zone) / Reporting District (District de signalement)** : Ces variables fournissent des informations sur la localisation de l'accident. En les analysant, vous pouvez déterminer quels quartiers ou zones géographiques ont le plus d'accidents.
6. **Crime Code (Code de crime) / Crime Code Description (Description du code de crime)** : Bien que le jeu de données soit axé sur les accidents de la circulation, il semble inclure des codes de crime. Il pourrait être important d'examiner de plus près pourquoi ces informations sont incluses et comment elles pourraient affecter l'analyse des accidents.
7. **MO Codes (Codes MO - Modus Operandi)** : Ces codes peuvent fournir des informations supplémentaires sur la manière dont l'accident s'est produit. Par exemple, s'il s'agit d'une collision de véhicules, il peut y avoir des codes spécifiques indiquant des comportements tels que la vitesse excessive, le non-respect des feux de signalisation, etc.
8. **Victim Age (Âge de la victime) / Victim Sex (Sexe de la victime) / Victim Descent (Origine de la victime)** : Ces variables fournissent des détails sur les personnes impliquées dans l'accident. L'analyse de ces données peut aider à comprendre quel groupe démographique est le plus touché par les accidents de la circulation.
9. **Premise Code (Code du lieu) / Premise Description (Description du lieu)** : Ces variables décrivent le type de lieu où l'accident s'est produit. Cela peut inclure des informations telles que "intersection", "autoroute", "parking", etc. L'analyse de ces données peut révéler les endroits les plus dangereux ou les plus sujets aux accidents.

10. **Address (Adresse de l'incident) / Cross Street (Rue transversale)** : Ces variables fournissent des informations spécifiques sur l'emplacement de l'accident, ce qui peut être crucial pour l'analyse géospatiale.
11. **Location (Coordonnées géographiques)** : Les coordonnées géographiques de l'accident. Cela peut être utilisé pour cartographier précisément les lieux des accidents et analyser les tendances spatiales.

En analysant ces variables, vous pourrez mieux comprendre les facteurs qui contribuent aux accidents de la circulation à Los Angeles, tels que les heures et les endroits les plus dangereux, les groupes démographiques les plus touchés, et les comportements associés aux accidents. Cela peut être