Projet I(2023-24)

Technologies Big Data. M1 MIDS/Informatique

Vlady Ravelomanana and Stéphane Boucheron

2024-03-25

Contexte

La *Bases de données annuelles des accidents corporels de la circulation routière - Années de 2005 à 2022* est une donnée ouverte (*open data*) mise à disposition du public par le ministère des transports.

🔗 Bases de données annuelles des accidents corporels de la circulation

https://www.data.gouv.fr/fr/datasets/bases-de-donnees-annuelles-des-accidents-corporels-de-la-circulation-routiere-annees-de-2005-a-2022/

> **i** Note
>
> Les indicateurs labellises se trouvent a l'URL : https://www.onisr.securite-routiere.gouv.fr/node/498

Les données sont présentées sous forme de fichiers annuels au format `csv`. Les données des années 2021 et 20022 sont formées de quatre fichiers portant sur

- Les *caractéristiques* des accidents
- Les *véhicules* impliqués dans les accidents
- Les *usagers* impliqués dans les accidents
- Les *lieux* (voies de circulation) concernées par les accidents

Chaque accident est repéré par un identifiant (colonne `Accident_Id` dans toutes les tables).

> **!** Important
>
> Lire soigneusement le texte `description-des-bases-de-donnees-annuelles-2022.pdf` (13 pages).

Vous travaillerez sur les données des années 2021 et 2022.

Attendus

Un cahier Python (fichier au format `.ipynb`) ou un fichier au format Quarto Markdown (fichier au format `.qmd`) équivalent. Votre fichier ne contiendra pas les sorties (tables et graphiques), mais il devra être *exécutable* sans erreurs dans un environnement où Python 3 et PySpark (>3.0) sont installés.

Si vous le souhaitez, vous pouvez dans une archive (un `.zip`) ajouter une version exécutée de votre cahier/fichier Quarto.

Le ou les fichiers devront être chargés sur Moodle avant l'échéance.

Questions

Dans votre rendu, vous répondrez au moins aux questions qui suivent. Vous pouvez les réinterpréter, les reformuler, et ajouter des questions qui vous semblent intŕessantes.

> **i** Question
>
> **Chargement des fichiers pour les années 2021, 2022**

Vous téléchargez les huit fichiers `.csv` correspondant aux données des années 2021, 2022, dans un sous-répertoire nommé `data` de votre répertoire de travail (celui où se trouve votre cahier/fichier quarto).

Effectuez ce téléchargement de manière programmatique (pas à la main) à l'aide des outils fournis par le module `requests`.

> **i** Question
>
> **Création de dataframes correspondants au fichiers des années 2021, 2022 et nettoyage des données**

Créer un *dataframe* `Spark` ou `Pandas on Spark` pour chaque fichier de données.

Vous veillerez à ce que dans les tables

- les colonnes soient correctement typées (y compris les informations temporelles)
- les valeurs nulles/manquantes soient correctement identifées
- les colonnes catégorielles/qualitatives soient traitées comme des colonnes catégorielles et pas comme des colonnes de type `string`.

Vous veillerez à ce que vos opérations soient reproductibles (qu'il soit possible de traiter les données 2023 si elles n'ont pas subi des changements de forme importants).

Vous éviterez les copiés-collés et suivrez le principe DRY

> **Tip**
>
> Pour le traitement des colonnes catégorielles, si vous n'utilisez pas `Pandas on Spark`, pensez à utiliser `StringIndexer()` dans le module `pyspark.ml.feature`.

> **Question**
>
> **Réunion des années 2021, 20222**

Réalisez la réunion des dataframes correspondant aux années 2021 et 2022.

> **Question**
>
> **Résumés numériques**

Construisez les résumés numériques : moyenne, médiane, quartiles, écart-type, écart interquartile, asymétrie (*skewness*), applatissement (*kurtosis*) pour les colonnes numériques.

Engendrez les objets graphiques correspondant aux boites à moustaches (*boxplot*) de ces colonnes numériques. Affichez ceux que vous jugerez intéressants.

Pour les graphiques, privilégiez `plotly` ou `altair`, plutôt que `matplotlib`/`seaborn`.

> **Question**
>
> **Visualisation**
>
> Construisez des objets graphiques propres à visualiser :
>
> - Répartition des accidents sur la semaine (jours et heures)
> - Répartition des accidents sur les mois de l'année

Question

Profil des usagers

Visualisez le profil des usagers impliqués dans un accident. Distinguez les profils en fonction des circonstances (circulation urbaine, campagne,)

Question

Accidents impliquant des cyclistes et/ou des piétons

Déterminez les lignes qui correspondent à des accidents impliquant au moins un piéton ou un cycliste.

Construisez à nouveau des objets graphiques propres à visualiser :

- Répartition des accidents sur la semaine (jours et heures)
- Répartition des accidents sur les mois de l'année

Construisez des objets graphiques propres à visualiser les caractéristiques des lieux où se sont produits ces accidents.

Question

Usage des types composites

Fabriquer un dataframe avec une ligne par accident, une colonne contenant les véhicules en cause, et les lieux de l'accident.

Tip

Les dataframes de Spark sont susceptibles de représenter des types dits composés, avec les types `ArrayType()`, `StructType()` et `MapType()`.

Question

Sauvegarde au format `parquet`

Sauvegarder vos données nettoyées au format `parquet`. Choisissez un partitionnement qui vous facilite la vie. Motivez vos choix.

Critères de notation

Critère	Points	Détails
Orthographe et grammaire	10%	English/French
Graphiques	20%	choix des types de graphiques, échelles, …
Manipulations de Tables	20%	ETL, SQL
Calcul des Statistiques	10%	Aggrégations, Fenêtres, …
Concision (DRY)	20%	DRY principle at W Wikipedia
Créativité	20%	