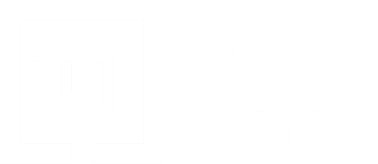
A black and blue sign with text

Description automatically generated

|  |  |
| --- | --- |
| Compte Rendu de Projet ADC  Violence Armée aux États-Unis (2013-2018) | |
| Encadré par : Élaborer par : | |
| Pr. CHEMLAL Yman | ERFOUDI Anass  MOUSDIK Ismail  DAIF Abdellah  EDDAOUDI Mohamed |



Contents

[I. Contexte du Projet 3](#_Toc183950670)

[II. Objectifs du Projet 4](#_Toc183950671)

[III. Description des Données 5](#_Toc183950672)

[IV. Étapes Réalisées 6](#_Toc183950673)

[1. Préparation des Données 6](#_Toc183950674)

[2. Analyse Exploratoire 6](#_Toc183950675)

[3. Visualisations Dynamiques dans Power BI 7](#_Toc183950676)

[V. Résultats Principaux 8](#_Toc183950677)

[VI. Applications Pratiques 9](#_Toc183950678)

[VII. Limitations 10](#_Toc183950679)

[VIII. Recommandations 11](#_Toc183950680)

[IX. Conclusion 12](#_Toc183950681)

# Contexte du Projet

La violence armée est un problème majeur aux États-Unis, avec des impacts sociaux, économiques, et humains importants. Ce projet vise à analyser un jeu de données contenant des informations détaillées sur les incidents de violence armée, couvrant la période 2013-2018, afin d'identifier des tendances et d'apporter des éclairages utiles pour les politiques publiques.

# Objectifs du Projet

* Comprendre les tendances temporelles des incidents de violence armée.
* Identifier les zones géographiques les plus affectées.
* Explorer les caractéristiques des incidents (nombre de victimes, type d’armes impliquées, etc.).
* Créer un tableau de bord interactif pour visualiser les résultats et faciliter l'interprétation des données.

# Description des Données

## Introduction et Choix du Jeu de Données

Question 1 : **Quel jeu de données avez-vous choisi, et pourquoi ce sujet est-il pertinent ou intéressant pour l’analyse ?**  
 Le jeu de données choisi concerne la violence armée aux États-Unis entre 2013 et 2018. Ce sujet est pertinent car la violence armée représente un problème critique aux États-Unis, affectant de nombreux aspects de la société, notamment la sécurité publique, la santé et les politiques sociales. Analyser ces données peut aider à identifier des tendances, des zones sensibles, et les facteurs associés à ces incidents. Cela permet également de fournir des insights pour orienter les politiques publiques et les stratégies de prévention.

Question 2 : **Quelles sources de données avez-vous explorées pour identifier un jeu de données pertinent et répondant aux critères (au moins 500 lignes) ?**  
Pour identifier un jeu de données pertinent, plusieurs bases de données publiques ont été explorées, notamment **Kaggle**, Data.gov et Gun Violence Archive. Le jeu de données final provient de Gun Violence Archive, une source fiable qui compile les incidents liés aux armes à feu à partir de rapports des forces de l'ordre, des médias et des bases de données publiques. Ce jeu de données a été choisi car il contient plus de **260 000** lignes, satisfaisant largement les critères de taille minimale, et offre une couverture détaillée des incidents.

## Importation et Exploration Préliminaire des Données

Question 1 : **Quelles sont les principales caractéristiques de votre jeu de données (nombre de colonnes, types de variables) ?**  
Le jeu de données comporte **29 colonnes** couvrant une variété d'informations, notamment :

* **Variables temporelles** : date (date de l'incident).
* **Variables géographiques** : state, city\_or\_county, latitude, longitude.
* **Variables sur les victimes** : n\_killed (nombre de morts), n\_injured (nombre de blessés).
* **Caractéristiques des incidents** : incident\_characteristics (ex. fusillades de masse, possession illégale d'armes).
* **Détails des participants** : participant\_age, participant\_gender, participant\_type.

Le jeu de données est hétérogène, incluant des variables qualitatives, quantitatives, et des données textuelles structurées à l’aide de séparateurs (||).

Question 2 : **Avez-vous identifié des données manquantes ou des valeurs aberrantes lors de votre première exploration ?**

**Si oui, quelles sont-elles ?**  
Oui, plusieurs problèmes ont été identifiés lors de l'exploration préliminaire :

* **Données manquantes** : Certaines colonnes, incident\_url, source\_url, incident\_url\_fields\_missing et location\_description, contiennent des valeurs nulles ou absentes pour de nombreux enregistrements.
* **Valeurs aberrantes** : Dans certaines colonnes, comme n\_guns\_involved, des valeurs sont soit absentes, soit incohérentes (par exemple, des incidents enregistrés avec zéro armes impliquées).
* **Colonnes avec des séparateurs** : Des colonnes comme participant\_age et incident\_characteristics utilisent le séparateur || pour combiner plusieurs valeurs, ce qui complique leur traitement direct.

# Étapes Réalisées

## Netoyage et Préparation des Données

Question 1 : **Quelles étapes avez-vous effectuées pour nettoyer les données (traitement des valeurs manquantes, suppression des doublons, transformation des formats) ?**

Lors de la phase de nettoyage des données, les étapes suivantes ont été réalisées :

1. **Suppression des colonnes inutiles** :
   * Les colonnes suivantes ont été supprimées car elles ne contribuent pas à l'analyse :
     + incident\_url, source\_url, incident\_url\_fields\_missing (liens vers des sources externes).
   * Les colonnes participant\_relationship, gun\_stolen, et gun\_type ont également été supprimées, car elles contiennent des valeurs vides dans presque tous les enregistrements.
2. **Vérification des doublons** :
   * Une analyse des doublons a été effectuée sur la base des colonnes principales (incident\_id, date, state). Aucune duplication n’a été identifiée, ce qui garantit l’unicité des enregistrements.
3. **Traitement des formats** :
   * Les formats de la colonne date ont été standardisés au format AAAA-MM-JJ.
   * Les valeurs textuelles dans des colonnes comme incident\_characteristics ont été uniformisées pour éviter les variations dues à des erreurs de typographie.
4. **Gestion des valeurs manquantes** :
   * Les colonnes ayant des valeurs manquantes importantes et peu exploitables (comme latitude et longitude) ont été conservées pour une analyse partielle géographique.

Question 2 : **Quelles transformations avez-vous appliquées pour rendre les données exploitables ?**

Les transformations suivantes ont été appliquées :

1. **Catégorisation et normalisation des données sur les participants** :
   * Les colonnes relatives aux participants (participant\_age\_group, participant\_gender, participant\_status, et participant\_type) ont été normalisées et catégorisées :
     + **participant\_age\_group** : Groupes d’âge comme Enfant (<18 ans) et Adulte (18+ ans).
     + **participant\_gender** : Normalisation des valeurs (Male, Female).
     + **participant\_status** : Statut des participants (e.g., Killed, Injured, Arrested).
     + **participant\_type** : Types comme Victim, Subject-Suspect.

A screenshot of a computer

Description automatically generated

1. **Création d'une nouvelle variable** :
   * Une colonne nommée **n\_participant\_involved** a été ajoutée pour calculer automatiquement le nombre total de participants impliqués dans chaque incident.
   * Ce calcul a été réalisé en comptant les occurrences dans les colonnes relatives aux participants (participant\_age\_group, etc.).
2. **Création d’une table calendrier :**
   * Une table calendrier a été générée pour enrichir l’analyse temporelle. Cette table contient des colonnes spécifiques telles que :
     + **Année** : Pour faciliter l’analyse des tendances annuelles.
     + **Mois** : Pour identifier les variations saisonnières et mensuelles.
   * La table calendrier a été liée à la colonne date du dataset principal, permettant d'effectuer des analyses temporelles plus précises et d’améliorer la cohérence des données.

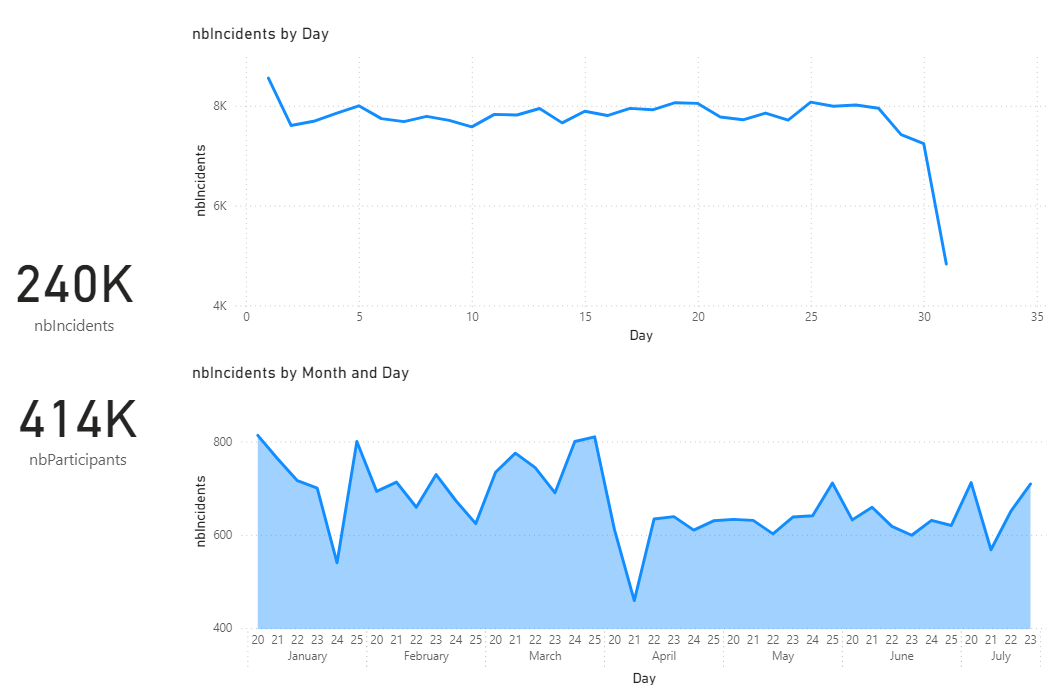
Cette approche a permis de structurer les données temporelles de manière efficace, tout en rendant l’analyse plus flexible et intuitive.

## Analyse Exploratoire

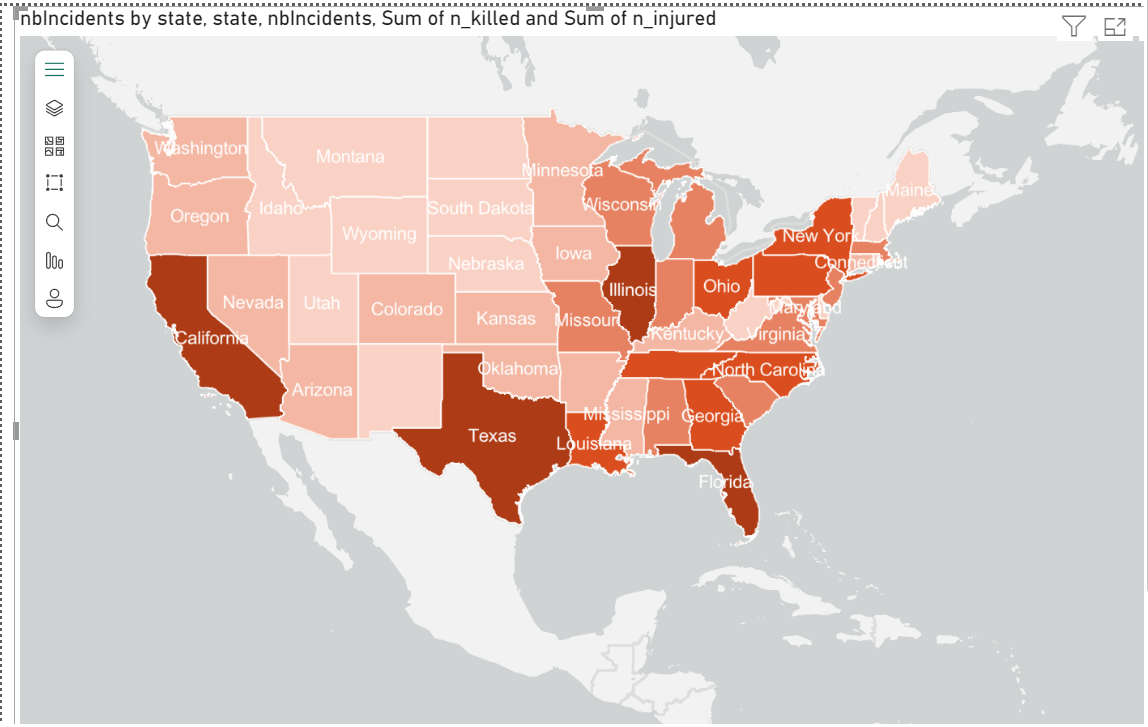
* **Analyse temporelle** :
  + Calcul du nombre total d’incidents par année et par mois.
  + Identification des tendances (augmentation ou diminution des incidents).

A blue bar graph with white text

Description automatically generated



* **Analyse géographique** :
  + Distribution des incidents par état et par ville.
  + Création de cartes interactives pour visualiser les zones les plus touchées.



* **Analyse démographique** :
  + Répartition des victimes par âge, sexe et rôle (participant\_type).
  + Comparaison des incidents impliquant différents types d'armes.

A screenshot of a computer

Description automatically generated

## Visualisations Dynamiques dans Power BI

Les visualisations suivantes ont été créées :

1. **Graphique en barres** : Nombre total d'incidents par année.
2. **Carte géographique** : Répartition des incidents dans chaque état.
3. **Diagramme circulaire** : Proportion des incidents selon leurs caractéristiques (fusillades de masse, incidents domestiques, etc.).
4. **Graphique en lignes** : Tendance mensuelle des incidents.
5. **Carte thermique (Heatmap)** : Corrélation entre les incidents, les victimes et les armes impliquées.

# Résultats Principaux

* **Tendances temporelles** :
  + Le nombre d’incidents a augmenté de manière significative entre 2013 et 2017, avant de diminuer légèrement en 2018.
  + Les mois d’été (juin à août) enregistrent généralement plus d’incidents.
* **Tendances géographiques** :
  + Les états comme la Californie, la Floride et le Texas sont les plus touchés.
  + Les grandes villes concentrent une majorité des incidents.
* **Caractéristiques des incidents** :
  + Les fusillades de masse représentent une petite proportion des incidents, mais elles causent un nombre disproportionné de victimes.
  + Les hommes adultes sont les principales victimes et auteurs identifiés.

# Applications Pratiques

* **Politiques publiques** :
  + Renforcement des lois sur les armes dans les états les plus touchés.
  + Ciblage des programmes de prévention dans les zones à risque.
* **Sensibilisation** :
  + Campagnes pour informer le public sur les risques liés aux armes à feu.
* **Éclairages pour la recherche** :
  + Analyse approfondie des facteurs contribuant à la violence armée.

# Limitations

* Données incomplètes : Certaines colonnes manquent de données critiques (par exemple, location\_description).
* Difficulté à standardiser certaines variables (comme incident\_characteristics).
* Nécessité de données additionnelles pour évaluer des facteurs contextuels (population, politique d’armement par état).

# Recommandations

* Étendre l’analyse à d’autres facteurs contextuels (données socio-économiques, densité de population).
* Collaboration avec des experts en politiques publiques pour interpréter les résultats.
* Exploration future des tendances post-2018 pour observer les évolutions.

# Conclusion

Ce projet a permis de mieux comprendre les tendances et caractéristiques des violences armées aux États-Unis. Le tableau de bord interactif créé dans Power BI offre une plateforme puissante pour explorer ces données et les partager avec différents acteurs, qu’ils soient chercheurs, décideurs politiques ou membres du public.