

1 MÉTHODOLOGIE {#methodologie}

Approche générale

1. **Collecte et consolidation** : Import du dataset brut
2. **Audit qualité** : Analyse des données manquantes et aberrantes
3. **Nettoyage intelligent** : Traitement différencié selon la nature des manquements
4. **Enrichissement** : Création de variables dérivées (tranches d'âge, périodes)
5. **Exploration multi-axes** : Analyses temporelle, spatiale, démographique
6. **Synthèse** : Extraction d'insights et formulation de recommandations

Outils utilisés

- **Python 3.x** (Pandas, NumPy, Matplotlib, Seaborn)
- **Jupyter Notebook** (documentation et reproductibilité)
- **Missingno** (visualisation des données manquantes)




2 PRÉPARATION DES DONNÉES {#preparation}

2.1 Audit initial

Dimensions du dataset

- 1 000 000 lignes (incidents)
- 28 colonnes (variables)

Qualité globale

- Données dupliquées :  Vérifiées (aucune duplication)
- Données manquantes :  Présentes dans 12 colonnes
- Format des dates :  Nécessite normalisation

2.2 Traitement des données manquantes

Approche méthodologique

Nous avons classifié chaque variable selon le mécanisme de disparition (MCAR, MAR, MNAR) pour appliquer le traitement le plus approprié.

Variable	% Manquant	Type	Traitement	Justification
Crime Code 2/3/4	>93%	MNAR	Suppression	Crimes secondaires peu pertinents
Cross Street	Variable	MNAR	"PRECISE ADDRESS"	Absence = adresse précise disponible
Mocodes	15.09%	MNAR	"NO RECORD"	Absence = pas d'antécédents

Variable	% Manquant	Type	Traitement	Justification
Weapon Used Cd/Desc	67.44%	MNAR	"NO WEAPON"	Absence = pas d'arme utilisée
Vict Descent/Sex	~14%	MAR	"NO VICTIM" ou "X"	Corrélés (pas de victime humaine)
Premis Cd/Desc	Variable	MAR	0 / "NO DESCRIPTION"	Lieu non identifiable
Crm Cd 1	Faible	Simple	Copie Crm Cd	Redondance des codes
Status	Minimal	Simple	"IC"	Valeur par défaut système

Résultat

✔ 100% des données manquantes traitées sans perte d'information pertinente

2.3 Normalisation des formats

Dates

- **Avant** : MM/JJ/AAAA HH:MM:SS (format mixte)
- **Après** : JJ/MM/AAAA (format standard FR, heures séparées)

Heures

- **Problème détecté** : Valeurs >2400 (minutes au-delà de minuit)
- **Solution** : Conversion en minutes depuis minuit, puis en heures standard
- **Format final** : 0-23 (24h)

2.4 Création de variables dérivées

Nouvelle variable	Description	Usage
Year	Année extraite	Analyse tendances annuelles
Month	Mois extrait	Saisonnalité
Day of Week	Jour de la semaine	Patterns hebdomadaires
Is Weekend	Booléen weekend	Comparaison semaine/weekend
Is Night	Booléen nuit (22h-04h)	Criminalité nocturne
Age Bracket	Tranches d'âge (0-17, 18-25, etc.)	Profils victimes