**Projet: Accidents routiers en France**

**Le contexte et objectif du projet:**

Chaque année, les forces de l'ordre en France recueillent des informations détaillées sur les accidents survenus sur les voies ouvertes à la circulation publique, impliquant au moins un véhicule et entraînant au moins une victime nécessitant des soins. Ces informations sont sauvegardées dans des fiches BAAC (bulletin d’analyse des accidents corporels) annuels.

Les informations extraites des fiches BAAC sont disponibles dans une base de données annuelle, couvrant les années de 2005 à 2022. Ces bases de données contiennent des informations détaillées sur les accidents sous format brute et sont structurées en quatre rubriques principales : CARACTÉRISTIQUES, LIEUX, VÉHICULES et USAGERS.

Chaque rubrique contient des informations spécifiques sur les accidents, les lieux, les véhicules impliqués et les usagers impliqués respectivement. Par exemple, la rubrique CARACTÉRISTIQUES contient des détails sur l'heure, la lumière, les conditions atmosphériques, etc. tandis que la rubrique VEHICULES contient des informations sur les véhicules impliqués tels que le type de véhicule, le sens de circulation, etc.

Dans ce contexte, l’objectif de notre projet est d’exploiter au mieux les données des accidents routiers des dernières dix années pour fournir un outil permettant d'analyser les accidents de la route dans le pays afin de réduire le nombre de décès et d'accidents en général.

Le projet comporte quatres étapes essentielles:

Exploration des données:

* identification du périmètre du projet.
* Analyse approfondie des données afin d'identifier les informations pertinentes à partir de diverses sources.
* Mise en œuvre de méthodes de nettoyage et d'intégrité pour garantir la qualité des données.

Modélisation des données

* Identification des KPIs à présenter
* Création des users stories de ces KPI
* Choisir un schéma de données convenable à nos données
* Implémentation des bases de données

Construction du pipeline ETL/ELT

Création du Dashboard analytique

**Livrable 1: Exploration de données non structurées**

Trello: https://trello.com/invite/b/grMqUwjw/ATTI917fbd94545948210a75cb33e79135e4A9E45EAF/projet-fil-rouge-accidents

L'objectif initial de ce rapport est d'entreprendre l'exploration, la visualisation et le prétraitement des bases de données en question.

**Périmètre du projet:**

**périmètre géographique:** France métropolitaine et / ou non départements d’Outre-mer (Guadeloupe, Guyane, Martinique, La Réunion et Mayotte ).

( données exclues: les autres territoires d’outre-mer (Saint-Pierre-et-Miquelon, Saint-Barthélemy, Saint-Martin, Wallis-et-Futuna, Polynésie française et Nouvelle-Calédonie) car disponible qu’à partir de 2019 dans l’open data.)

**Sources de données:** https://www.onisr.securite-routiere.gouv.fr/sites/default/files/2019-02/Guide%20BAAC%202017.pdf

Chaque BAAC annuel comprend plusieurs variables qui décrivent de façon précise les différents facteurs reliés aux accidents.

Compte tenu de la nature différentes de ces variables, elles sont classées selon 4 rubriques sous la forme pour chacune d'elles d'un fichier au format csv.

* rubrique CARACTÉRISTIQUES : décrit les circonstances générales de l’accident
* rubrique LIEUX : lieux des accidents qui peuvent être multiple en cas d’intersection
* rubrique VEHICLES : détails des véhicules impliqués
* rubrique USAGERS : personnes impliqués

**rubrique CARACTÉRISTIQUES**

**les variables clés:**

Num\_Acc

jour ( on peut avoir la date)

mois

an

lum (1 – Plein jour 2 – Crépuscule ou aube 3 – Nuit sans éclairage public 4 – Nuit avec éclairage public non allumé 5 – Nuit avec éclairage public allumé)

dep ( on peut extraire de département en utilisant le code INSEE)

com( code donné par l‘INSEE. Le code est composé du code INSEE)

agg (Localisation : 1 – Hors agglomération 2 – En agglomération)

Intersection : 1 – Hors intersection 2 – Intersection en X 3 – Intersection en T 4 – Intersection en Y 5 – Intersection à plus de 4 branches 6 – Giratoire 7 – Place 8 – Passage à niveau 9 – Autre intersection )??

atm(Conditions atmosphériques : -1 – Non renseigné 1 – Normale 2 – Pluie légère 3 – Pluie forte 4 – Neige - grêle 5 – Brouillard - fumée 6 – Vent fort - tempête 7 – Temps éblouissant 8 – Temps couvert 9 – Autre )

col (Type de collision : -1 – Non renseigné 1 – Deux véhicules - frontale 2 – Deux véhicules – par l’arrière 3 – Deux véhicules – par le côté 4 – Trois véhicules et plus – en chaîne 5 – Trois véhicules et plus - collisions multiples 6 – Autre collision 7 – Sans collision)

lat Latitude

Long Longitude => à modifier pour qu’on puisse les dessiner sur un maps

**rubrique LIEUX**

**Variables clés:**

Num\_Acc

catr(Catégorie de route)

circ (Régime de circulation)

prof (Profil en long décrit la déclivité de la route à l'endroit de l'accident)

infra( Aménagement - Infrastructure)

vma (Vitesse maximale autorisée sur le lieu et au moment de l’accident)?? maybe

**rubrique VÉHICULES**

**Variable clés:**

Num\_Acc

id\_vehicule ( permet de relier chaque véhicule au usagers impliqués dans l’accident)

catv (Catégorie du véhicule)

obsm (Obstacle mobile heurté)

motor (Type de motorisation du véhicule)

occutc Nombre d’occupants dans le transport en commun ?? nb personnes dans la voiture

**rubrique USAGERS**

Num\_Acc

id\_vehicule ( permet de relier usagers à véhicule)

grav (Gravité de blessure de l'usager)

sexe

An\_nais

locp (Localisation du piéton)

**clés pour join:**

Num\_Acc : relie usagers, lieux aux caractéristiques

Num\_Acc, num\_Veh: relie usagers à véhicule

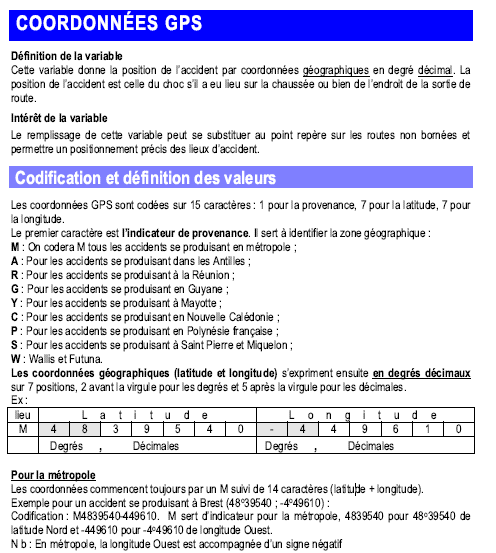
idée modification:

* on modifie le dep et commune avec les codes INSEE.
* on modifie les variables catégorielles par leur désignation pour les utiliser convenablement après dans les analyses.
* on a 3 variables: jour/mois/année:on peut créer une nouvelle variable date et supprimer jour
* on travaille sur la france métropolitaine uniquement
* identifier les KPI qu’on veut visualiser pour se concentrer sur leurs données:
  + nb accidents / commun / departement
  + type d’accident le plus commun
  + cause d'accident — (infra / temps/ déclivité route ..)

Différence entre les fichiers:

rubrique CARAC après / avant 2019

* lat et long : exprimées en unités différentes pour -2019 (Coordonnées géographiques en degrés décimaux)



* variable gps n'existe pas dans la version +2019
* variable an dans la version -2019 est présenté sur 2 chiffre (17) et +2019 (2017)
* variable heure est différente ( : n’est pas présente pour les versions -2019: exemple: 2017)

rubrique véhicules

* id\_vehicule n’est pas présente dans -2019
* les colonnes ne sont pas dans le même ordre dans attention pour la concaténation après

rubrique usagers

* secu1, secu2 et secu3 pour +2019 vs secu pour -2019 ( colonnes à supprimer)
* id\_vehicule n’est pas présente dans -2019
* id\_suager ( à partir de 2021)

rubrique lieux

* Avant 2019 la vitesse maximale dans un lieu n’existait pas.

suggestions next dans le sprint:

* chacun prend en charge une base de données: lui applique les transformations nécessaire, concatenate les différentes bases annuels par sujet.
* chacun rédige la partie expliquant les différentes sources de données accompagné des exemples de données collectées.

**Livrable 2: Modélisation des données**

**TRANSFORMATIONs:**

| **TABLE** | **PÉRIODE** | **TRANSFORMATION** |
| --- | --- | --- |
| **carac** | 2022 - 2019 | * suppression des NA |
| 2018 - 2016  2015 - 2012 | * supprimer la colonne “**gps**” parce qu’elle n'existe pas dans les versions plus récentes * modifier les col “**lat**” et “**long**” vers le système décimal * enlever le 0 de la col “**dep**” |
| 2015 - 2012 | * les col “**lat**” et “**long**” présentent des “0” ( tester dans une carte si c’est normal et c’est ‘ pas renseigné’ ) * les col “**lat**” et “**long**” présentent beaucoup de valeurs nulles |
| TOUT | * merge avec code département * créer une col “**date**” * split col “**hrmn**” pour avoir seulement les heures * supprimer la col “ **adr**”, elle va provoquer des problèmes d’encodage. * suppression des NA * NOTE: suppression des cols “**lat**” et “**long**” vu qu’elles présentent bcp de valeurs nulles à partir de 2016 |
|  | | |
| **usagers** | 2022 - 2019 | * A partir des colonnes “**sec1**” et “**sec2**” et “**sec3** ”, on peut créer une col “**utilisation\_secu**” où:   -1 : non renseigné  0: Non  1: Oui   * suppression col “**id\_usager**” et “**id\_vehicule**” vu qu’elle ne sont pas compatibles avec l’ancienne version. |
| 2018-2012 | * split col “**secu**” et garder le second concerne l’utilisation de l’Équipement de sécurité et créer une col “**utilisation\_secu**” où:   -1 : non renseigné  0: Non  1: Oui |
| TOUT | * on peut ajouter une col “age” a partir de la colonne “**an\_nais**” * drop les colonnes secu * supprimer la colonne “**place**” parce qu’elle n’est pas clé. ( sauf si on veut analyser les accidents par cat : transport en commun ou véhicule privé) * la col “**catu**” présente le num 4 (catégorie déplacée, à partir de l’année 2018, vers le fichier "Véhicules" Catégorie du véhicule : 99) * supprimer les “**locp**” et “**actp**”, "**etatp**” parce qu’elles ne sont pas représentatives, la valeur 0 ou -1( non renseigné) représente plus que 90% des données ( à vérifier) * idée pour “**trajet"? #dans quel cas de figure les accidents sont réalisés** * **actp: pour voir si une personne toute seule est plus exposé à un accident** * **locp: localisation de l’usager** |

| **TABLE** | **PÉRIODE** | **TRANSFORMATION** |
| --- | --- | --- |
| **Véhicules** | 2019-2022 | * drop col “**id\_vehicule**” * col “**motor**” n'existe pas dans les versions précédentes   (mais on peut la laisser en ajoutant plutôt des nulls et filtrer après parce que je trouve que c’est un indice intéressant ) |
| 2012-2018 | * col “**obsm**”: (objet mobile heurté) ne contient pas le 0 (aucun) et -1 ( non renseigné) donc présente bcp plus de NaN ou “ “:   => On peut remplacer ces valeurs par -1 ou 0 pour ne pas perdre de la donnée. |
| TOUT | * col “**catv**” : on a 48 catégories spécifiques: regrouper sous des cat plus génériques l:   1: VL (véhicule léger)  2: motor  3: trottinette, etc   * regrouper col “**manv** “ pour une granularité plus macro:   Tournant  Dépassant  Sans changement de direction  Dans le couloir bus  Changeant de file  En faisant demi-tour  => au lieu de 24 options, on peut avoir 8 significatives   * drop col “**obs**”: parce qu’elle ‘ est pas représentative,, la valeur 0 représente plus que 80% des données ( à vérifier) * drop col “ **occutc**”: présente le nombre d’occupants dans le transport public, je ne pense pas qu’elle soit utile pour nos analyses. * col ‘ **choc’** avec col **‘intersection**”: peut nous dire les accidents les plus courants dans les intersections. |

L'analyse des catégories de manœuvres aide à:

* Comprendre les causes des accidents : comprendre les circonstances qui ont mené à l'incident => savoir si l’accident était dû à une erreur humaine, à une condition routière dangereuse, ou à d'autres facteurs.
* Élaboration de stratégies préventives : les autorités peuvent développer des stratégies ciblées pour les prévenir => modification de la conception des routes, l'amélioration de la signalisation, ou la mise en œuvre de campagnes de sensibilisation spécifiques.
* Amélioration des lois et réglementations : L'analyse des accidents peut révéler des lacunes dans les normes de sécurité routière => adopter de nouvelles lois ou règlements visant à améliorer la sécurité.
* Planification urbaine et gestion du trafic : les manœuvres d'accident peuvent influencer la planification urbaine => Par exemple, si les accidents se produisent souvent dans les zones où les véhicules font demi-tour, cela pourrait indiquer la nécessité de construire des carrefours giratoires ou des passages supplémentaires.

L'attribut relatif au type de motorisation aide à:

* Comprendre les implications de sécurité spécifiques : Différents types de motorisation peuvent avoir des caractéristiques de performance uniques qui influencent le comportement du véhicule en cas d'accident. Par exemple, les véhicules électriques ont tendance à avoir un centre de gravité plus bas, ce qui peut affecter leur stabilité en cas de collision.
* Planification de l'infrastructure de recharge : En identifiant les zones avec une concentration élevée d'accidents impliquant des véhicules électriques ou hybrides, les urbanistes peuvent mieux planifier l'emplacement des stations de recharge pour améliorer l'accessibilité et la sécurité.
* Adaptation des services d'urgence : La connaissance du type de motorisation peut être cruciale pour les interventions d'urgence en cas d'accident. Par exemple, les véhicules électriques nécessitent des procédures d'urgence spécifiques en raison du risque d'électrocution ou d'incendie de la batterie.

| **TABLE** | **PÉRIODE** | **TRANSFORMATION** |
| --- | --- | --- |
| **Lieux** | 2019-2022 | * col “ **vitesse\_max**” : soit à droper soit à faire des analyses à part. |
| 2012-2018 | * Dans certaines colonnes ( **circ, vosp, prof, plan, surf**) les champs vides sont à remplacer par -1 ( non renseigné) pour s'aligner avec les nouvelles versions. * drop col “**env1** “ : n’existe pas et n’est pas interessante. |
| TOUT | * drop' **voie'**, **'v2'**, **'lartpc'**, **'v1'**, **'larrout':** valeurs manquantes ou non pertinentes * col “**pr**” et “**pr1**” ?? * On peut regrouper les options de col “**surf**” (Etat de la surface): en   -1: non renseigné/autre  0:normal  1:mouillé/anormal/glissante   * Pour la col”**infra**”, 90% des données des années est 0 (aucune infra) : ce qui logique mais je ne sais pas si elle est représentative ou non ( à droper ou non ) |

| **TABLE** | **PÉRIODE** | **TRANSFORMATION** |
| --- | --- | --- |
| **Immatricule.** | TOUT | * on peut modifer le CNIT par le type de voiture pour faire des analyses selon le type de voiture |

**KPIs:**

|  |  |  |
| --- | --- | --- |
| **Axe d’analyse** | **KPIs** | **User Story** |
| **Vu d’ensemble** | Nombre total d’accidents | *En tant qu'analyste de la sécurité routière, je veux connaître le nombre total d'accidents sur la période étudiée* |
| Nombre d’accidents annuels | *En tant qu'analyste de la sécurité routière, je veux connaître le nombre d'accidents par année.* |
| Variation annuelle des accidents | *En tant qu'analyste de la sécurité routière, je veux connaître l'évolution annuelle des accidents.* |
| Nombre d’accidents par conditions atmosphériques | *En tant qu'analyste de la sécurité routière, je veux connaître l’impact des conditions atmosphériques sur les accidents pour étudier l’infrastructure.* |
| Nombre d’accidents par type de collision ( type d’accidents) | *En tant qu'analyste de la sécurité routière, je veux connaître la répartition des types de collision accentuer le développement sur les mesures de protection les plus pertinentes.* |
| Répartition des victimes selon l'âge et le sexe | *En tant qu'analyste de la sécurité routière, je veux connaître le nombre des victimes par age et par sexe.* |
| Nombre de victimes par type de gravité d’accidents par catégorie d'usagers. ( même par âge). | *En tant qu'analyste de la sécurité routière, je veux connaître le nombre de victimes par type de gravité d'accidents , catégorie d'usager la plus impliquée dans chaque type et l'âge des chaque catégorie.* |
| Occurrence d'accidents par nombre de voitures impliqués | *En tant qu'analyste de la sécurité routière, je veux connaître le nombre d'accidents par nombre de voitures par accidents.* |
|  | | |
| **Geo spatial** | Nombre d’accidents par département. | *En tant qu'analyste de la sécurité routière, je veux visualiser le nombre d’accidents par département sur une carte ( on peut ajouter par gravité).* |
| Endroits avec forte concentration d'accidents | *En tant qu'analyste de la sécurité routière, je veux identifier sur le map les zones où le nombre d’accidents est > à la moyenne.* |
| Nombre d’accidents impliquants des piétons(Obstacle mobile heurté) | *En tant qu'analyste de la sécurité routière, je souhaite identifier les zones à risque élevé pour les piétons afin de proposer des améliorations de l'infrastructure piétonne.* |
| Top départements selon le nombre d’accidents | *En tant qu'analyste de la sécurité routière, je veux connaître les départements avec le grand nombre d’accidents.* |
| Nombre d'accidents répartis par type d’agglomération et type de route (autoroutes, nationales..) | *En tant qu'analyste de la sécurité routière, je veux connaître quels types de routes sont les plus accidentogènes.* |
| Nombre d’accidents par condition d’éclairage | *En tant qu'analyste de la sécurité routière, je veux connaître l’impact des conditions d’éclairage sur les accidents.* |
| Nombre d'accidents par type d’intersection(X,T,Y,..) | *En tant qu'analyste de la sécurité routière, je veux analyser le nombre d'accidents par type d'intersection pour identifier celles les plus dangereuses et proposer des mesures de signalisation pour améliorer la sécurité.* |
| Répartition des accidents par état de la surface (normale, mouillée, enneigée) | *En tant qu'analyste de la sécurité routière, je veux analyser la répartition des accidents selon le type de la surface pour connaître l’influence et ce dernier sur les accidents.* |

|  |  |  |
| --- | --- | --- |
| **Axe d’analyse** | **KPIs** | **User Story** |
| **Temporel** | Variation annuelle des accidents | *En tant qu'analyste de la sécurité routière, je veux connaître l'évolution annuelle des accidents pour identifier des années avec un nombre d’accidents plus élevé.* |
| Evolution d'accidents dans le temps par mois | *En tant qu'analyste de la sécurité routière, je veux analyser la répartition des accidents par mois pour identifier les périodes de l'année où les risques sont plus élevés* |
| Nombre d'accidents par jour de la semaine (weekday/weekend) | *En tant qu'analyste de la sécurité routière, je veux connaître la répartition des accidents par jour de la semaine pour optimiser les préventions.* |
| Nombre d'accidents par heure du jour. | *En tant qu'analyste de la sécurité routière, je veux identifier les heures de pointe en termes d'accidents pour améliorer la gestion du trafic et la signalisation.* |
| Evolution de la gravité de blessure dans le temps | *En tant qu'analyste de la sécurité routière, je veux connaître l'évolution de la gravité des accidents* |
|  | | |
| **Véhicules** | Nombre accidents par type de voiture | *En tant qu'analyste de la sécurité routière, je veux connaître le nombre d'accidents par type de voiture pour identifier les catégories de risque* |
| Répartition de la gravité d’accidents par type de voiture | *En tant qu'analyste de la sécurité routière, je veux analyser la gravité des accidents en fonction du type de voiture (robustesse voiture)* |
| Nombre d’accidents par modèle de voiture (CNIT) | *En tant qu'analyste de la sécurité routière, je veux connaitre les modèles de voiture les plus souvent impliqués dans des accidents pour des fins préventives* |
| Répartition de gravité d’accidents par modèle de voiture (CNIT) | *En tant qu'analyste de la sécurité routière, je veux connaître la gravité des accidents pour différents modèles de voiture afin de détecter si certains modèles sont moins robustes.* |
| Nombre d’accidents par âge de voiture | *En tant qu'analyste de la sécurité routière, je veux connaitre comment l'âge des véhicules influe sur le nombre d'accidents.* |
| Nombre d'accidents par manœuvre | *En tant qu'analyste de la sécurité routière, je veux identifier les manœuvres de véhicule les plus souvent associées aux accidents et proposer des améliorations éliminant ce genre de manœuvre risqués.* |
| Effet utilisation de sécurité sur la gravité de l’accident | *En tant qu'analyste de la sécurité routière, je veux voir si l'utilisation des sécurités influence la gravité ou non.* |

**MODÉLISATION:**

* faire un benchmark avec les différentes possibilités
* créer le diagramme
* créer le fichier d'implémentation
* tester si possible avec certaines valeurs

Pour la modélisation de données, on opte pour un modèle en **étoile**. Ce modèle est idéal pour les data warehouses car il facilite les **requêtes analytiques complexes** tout en maintenant une **structure simple et compréhensible**.

Exemple UML:

