ROYAUME DU MAROC

--*-*-*

HAUT COMMISSARIAT AU PLAN

--*-*-*-*-*-*-*

INSTITUT NATIONAL

DE STATISTIQUE ET D'ECONOMIE APPLIQUEE

INSEA

Le Mini-Projet de module :

Bases de données décisionnelles

*****

Création et analyse d'un DataWarehouse à partir d'un jeu de données des accidents de la ville New York en 2020

- Préparé par : ***BOUCHFAR ANASS***
- Sous la direction de : **Pr. *HILAL IMANE***
- Année universitaire : 2020/2021

Master de recherche en Systèmes d'Information et Systèmes Intelligents

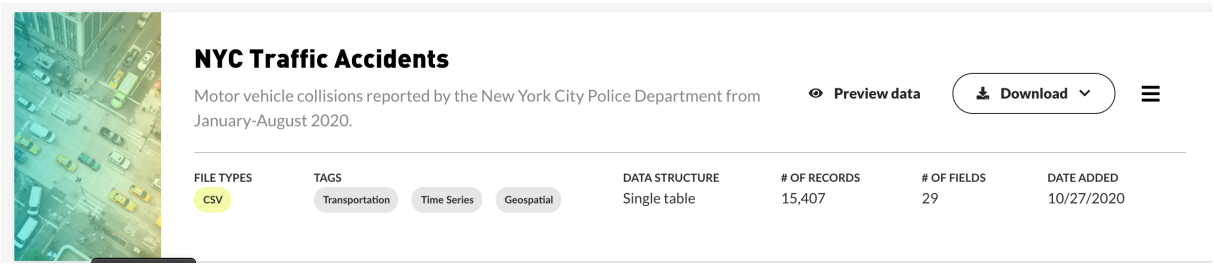
(M2SI – 1A – S2 – P1)

1. Introduction

Dans ce travail je vais vous présenter ensemble du processus décisionnel depuis la phase d'extraction de données jusqu'au la phase du rapport, et pour cela je vais utiliser comme solutions **Talend** pour la phase d'**ETL, MYSQL** comme SGBD et pour la partie reporting **POWER BI** qui est une solution d'analyse de données de Microsoft. Il permet de créer des visualisations de données personnalisées et interactives avec une interface suffisamment simple pour que les utilisateurs finaux créent leurs propres rapports et tableaux de bord

Et comme données on va étudier les accidents dans la ville New York City durant l'année 2020 qu'on a pris à partir de la source suivante :

https://www.mavenanalytics.io



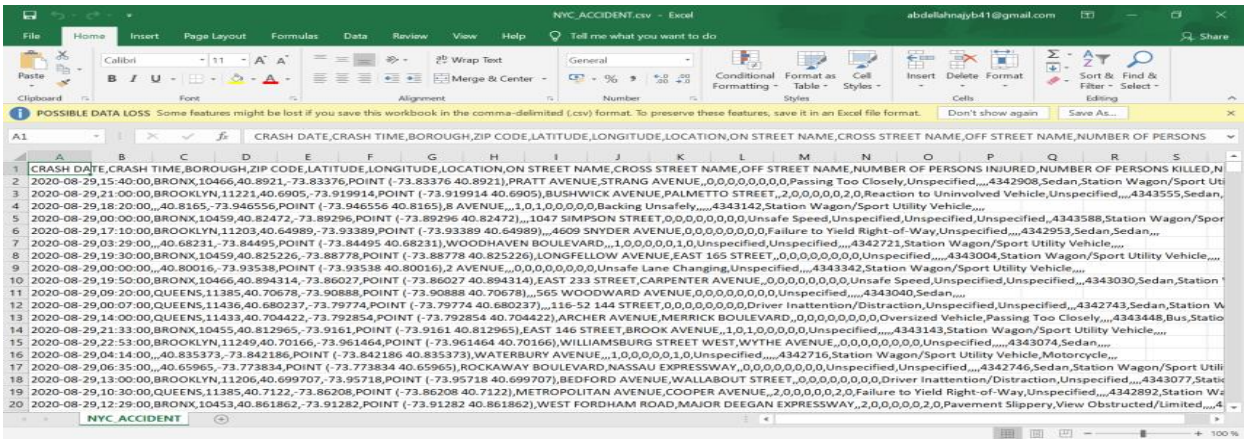
2. Implémentation

je vais essayer de démontrer les différentes phases que j'ai effectué afin de créer mon système décisionnel, pour cela je vais diviser mon processus en 3 parties

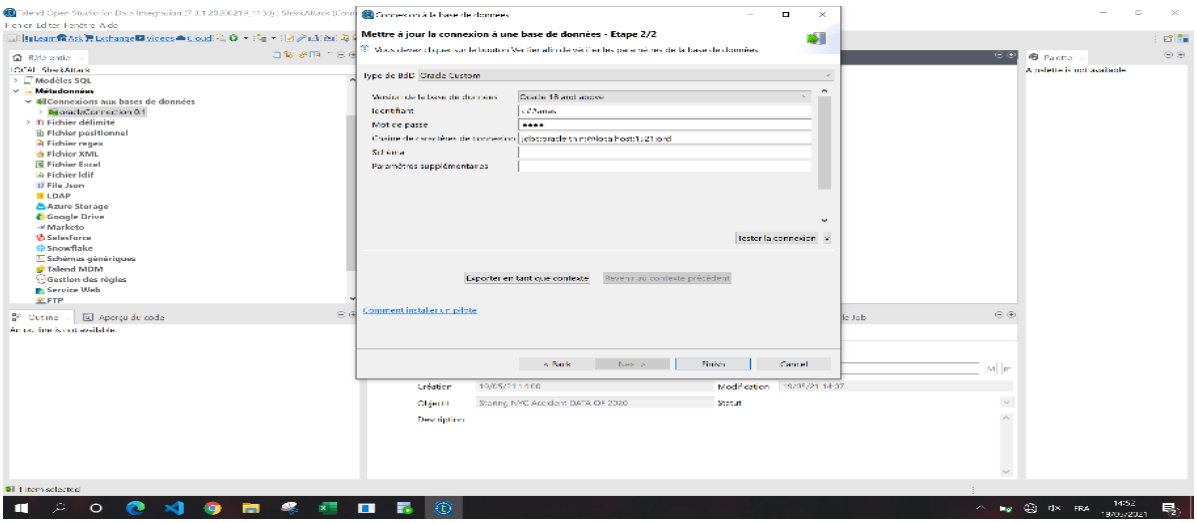
2.1. ETL

Dans cette phase j'ai essayé d'extrait les informations nécessaire à l'analyse, les transforme en un format qui peut répondre aux besoins opérationnels et la charge dans un Data Warehouse. Pour cela j'ai utilisé **Talend.** Et voila quelques étapes que j'ai suivi au long de cette phase.

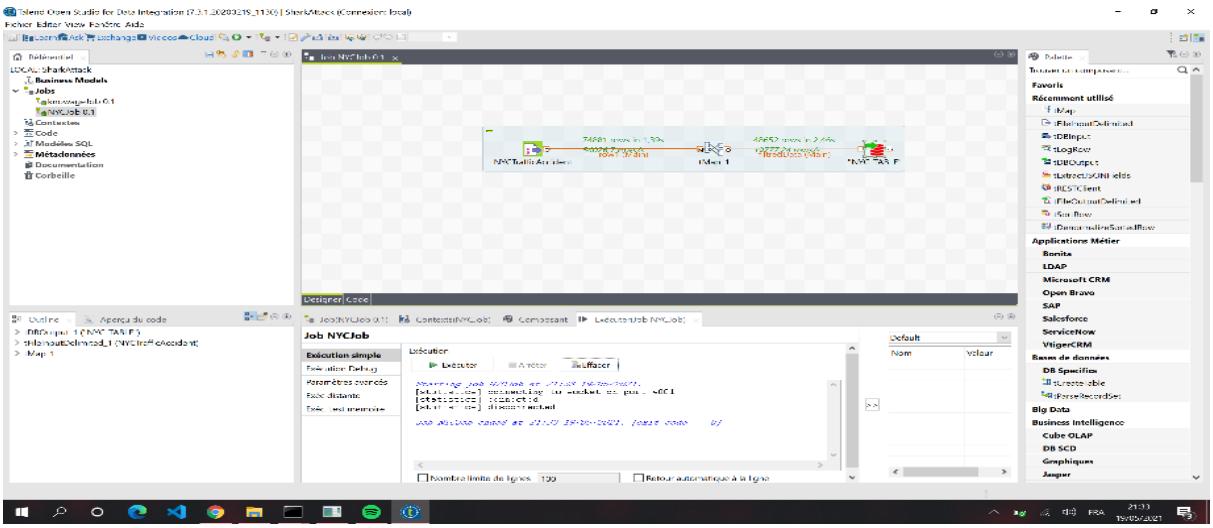
Premièrement j'ai commencé par l'importation de données à partir d'un fichier csv nommée NYC_accident.csv



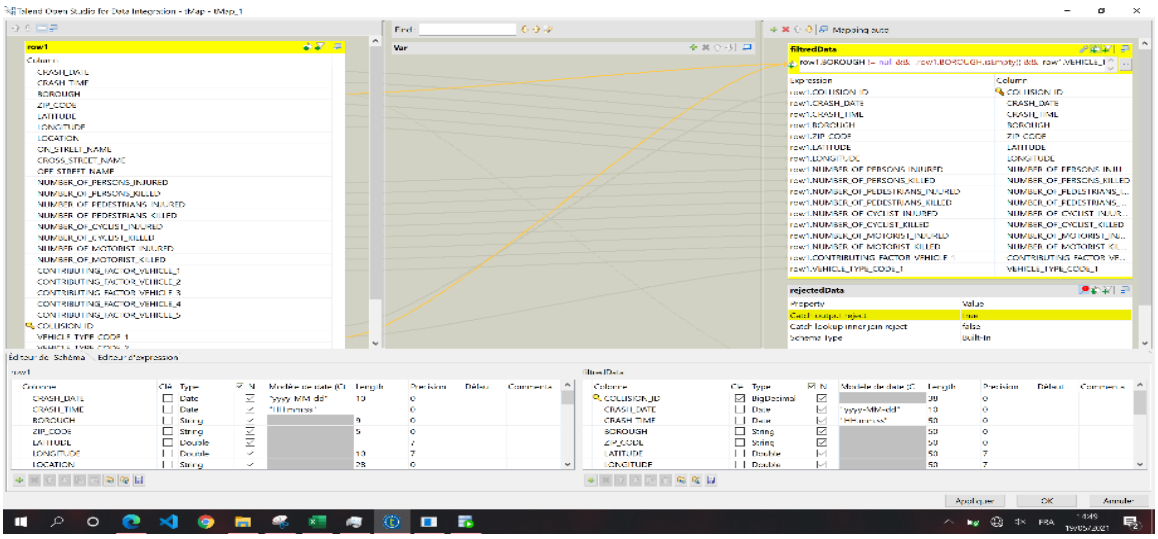
Après j'ai créer une connexion vers notre base de données mysql dans laquelle on va stocker les données traitées dans un table :



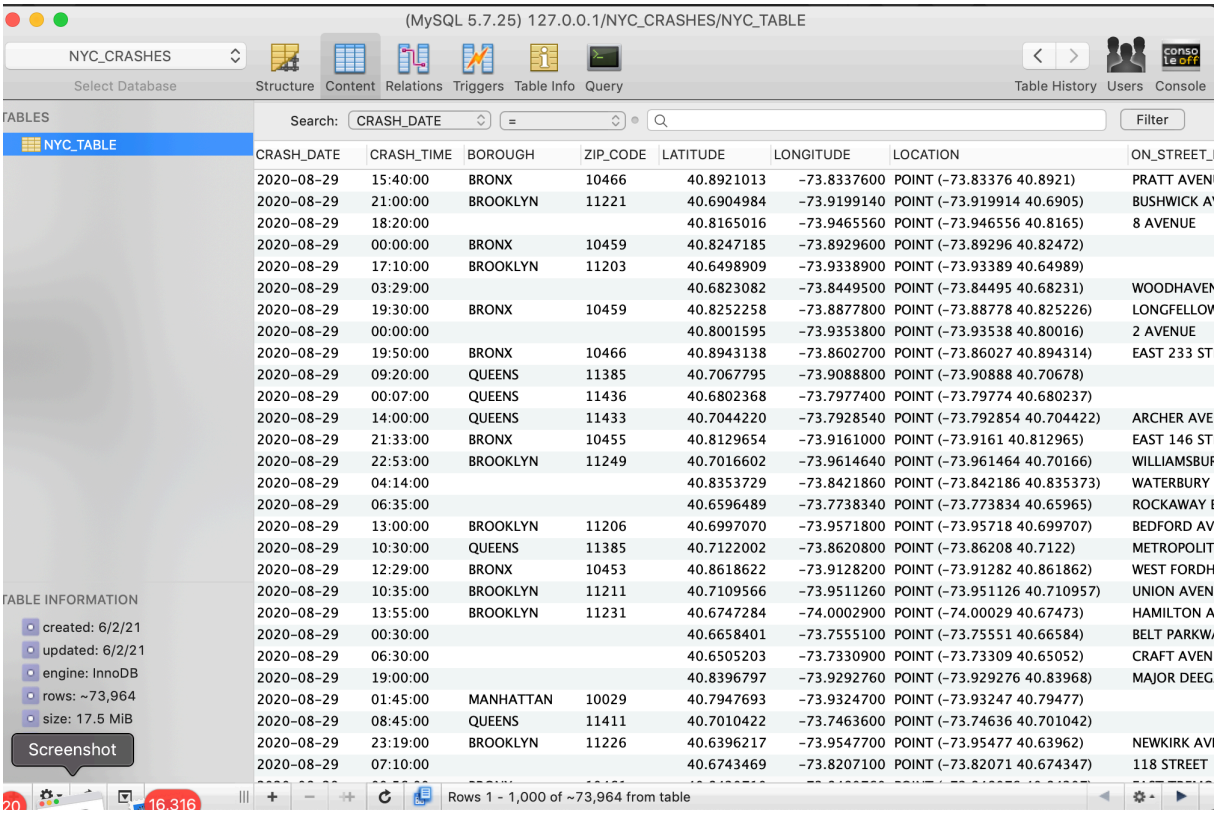
Après j'ai créé mon job dans lequel j'ai traité et filtrer mes données et les charger dans une base de données mysql, et voila la structure de mon Job est comme suivant :



Ensuite, j'ai configuré l'outil tMap que j'ai utilisé afin de filtrer mes données (éliminer les champs vides et nul) et structurer mes données de sortie :

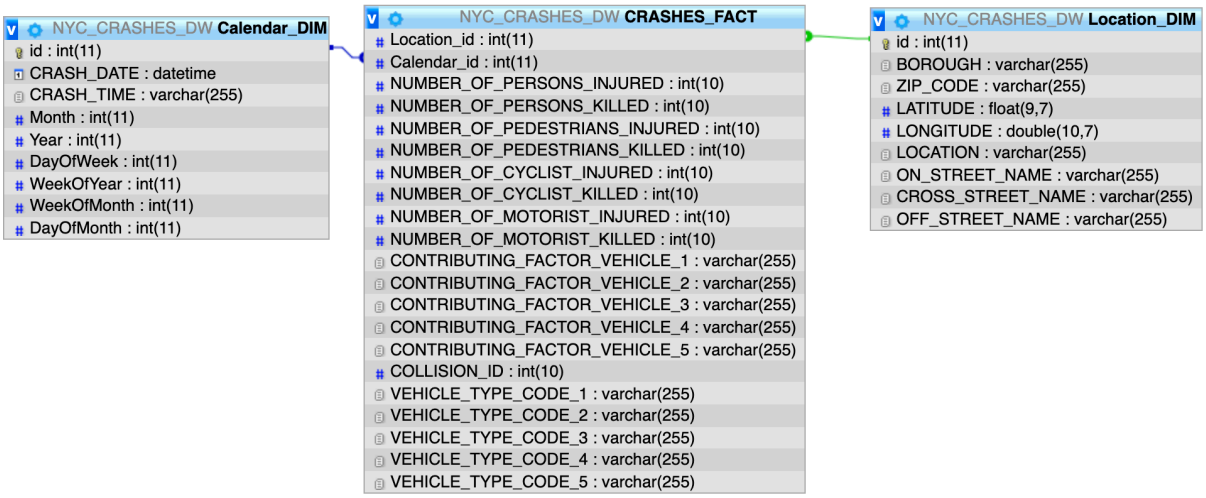


Et après l'exécution de mon job j'ai eu mes données transformées au niveau de notre base de données mysql :



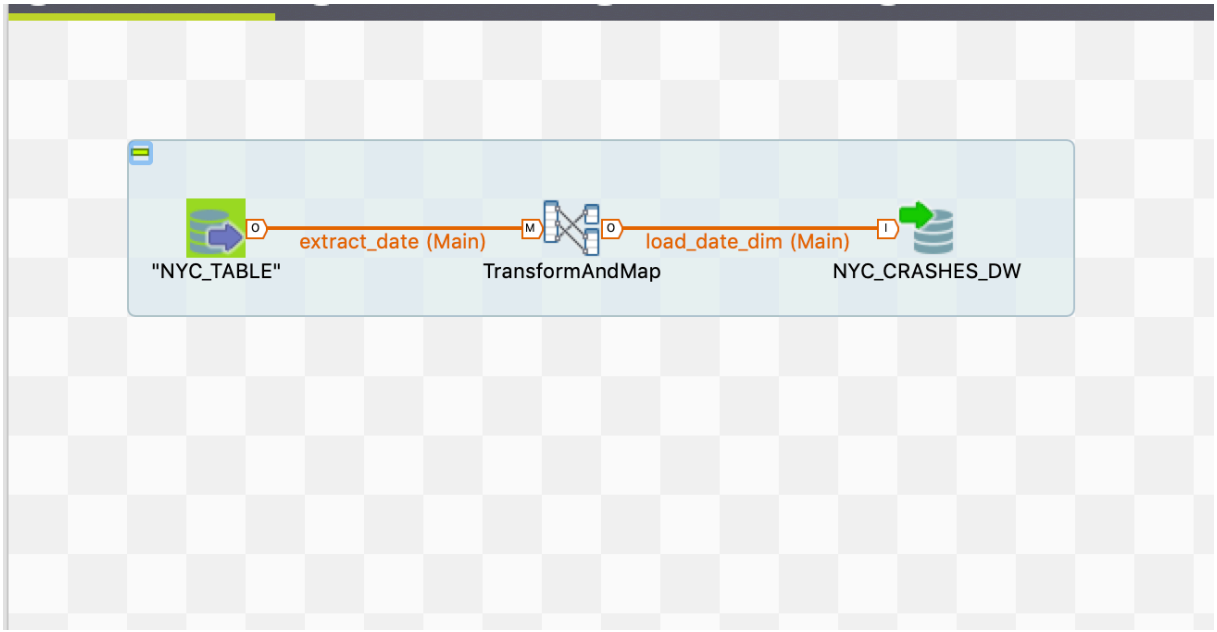
Et voila la première phase est terminé et avant de passer à la deuxième phase je dois faire une modélisation de mon data Warehouse, cette modélisation qui est une modélisation multidimensionnelle contenant la table de fait qui est la mesure et les tables de dimensions qui sont liées à ce fait en formant une modélisation en étoile.

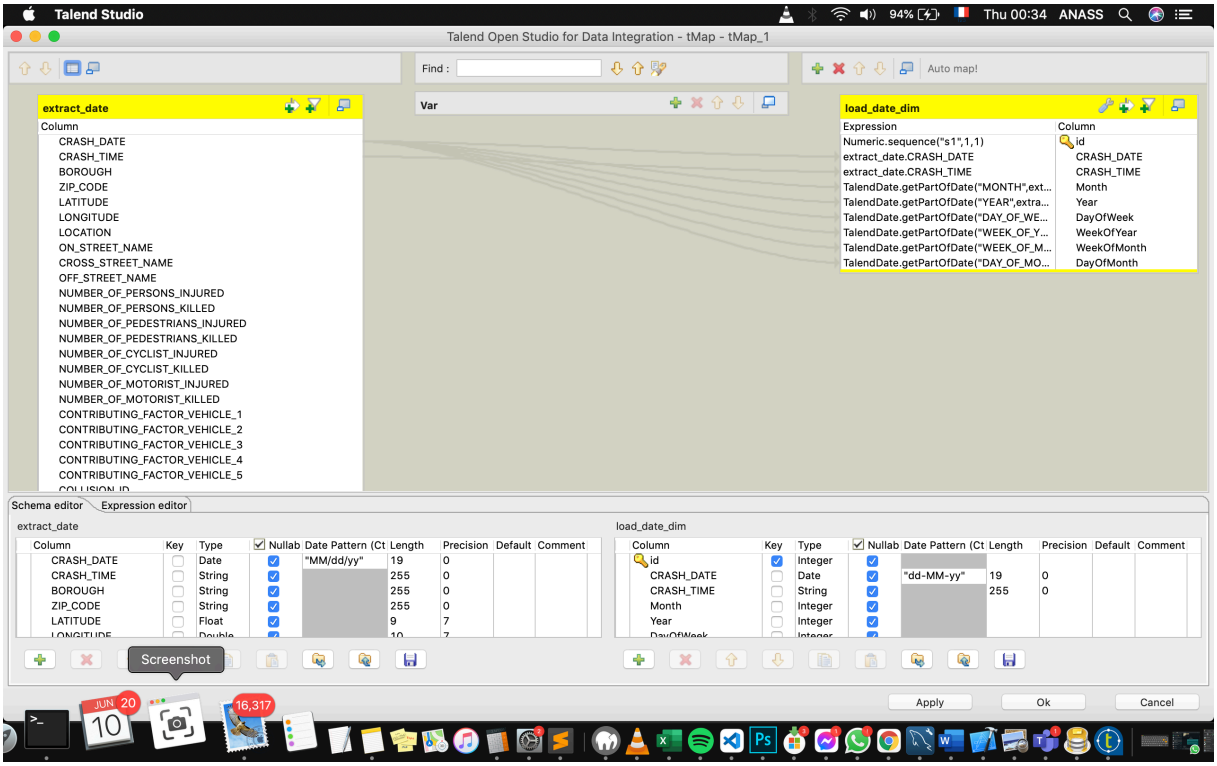
Dans mon cas la table de fait est les accidents à New York City lié à la dimension temps présenté par la table **Calendar_DIM** et la dimension du lieu présenté par la table Location_DIM et chaque table contenant des attributs



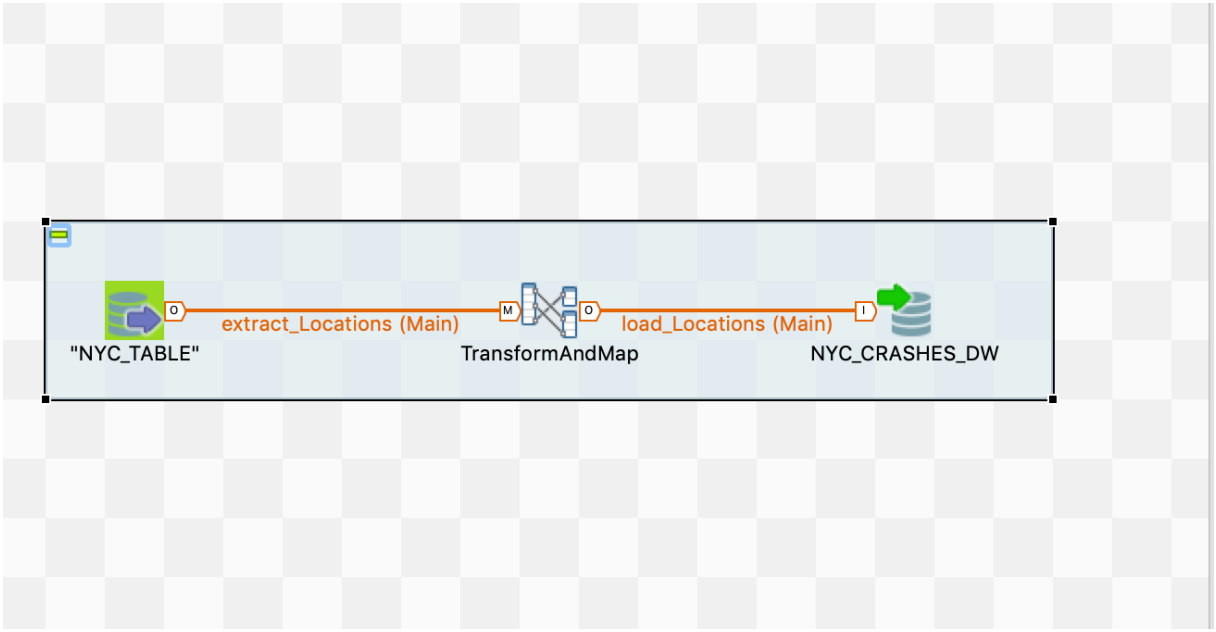
Donc après qu'on fait notre modélisation on va commencer ETL(extraction-Transform-LOAD)

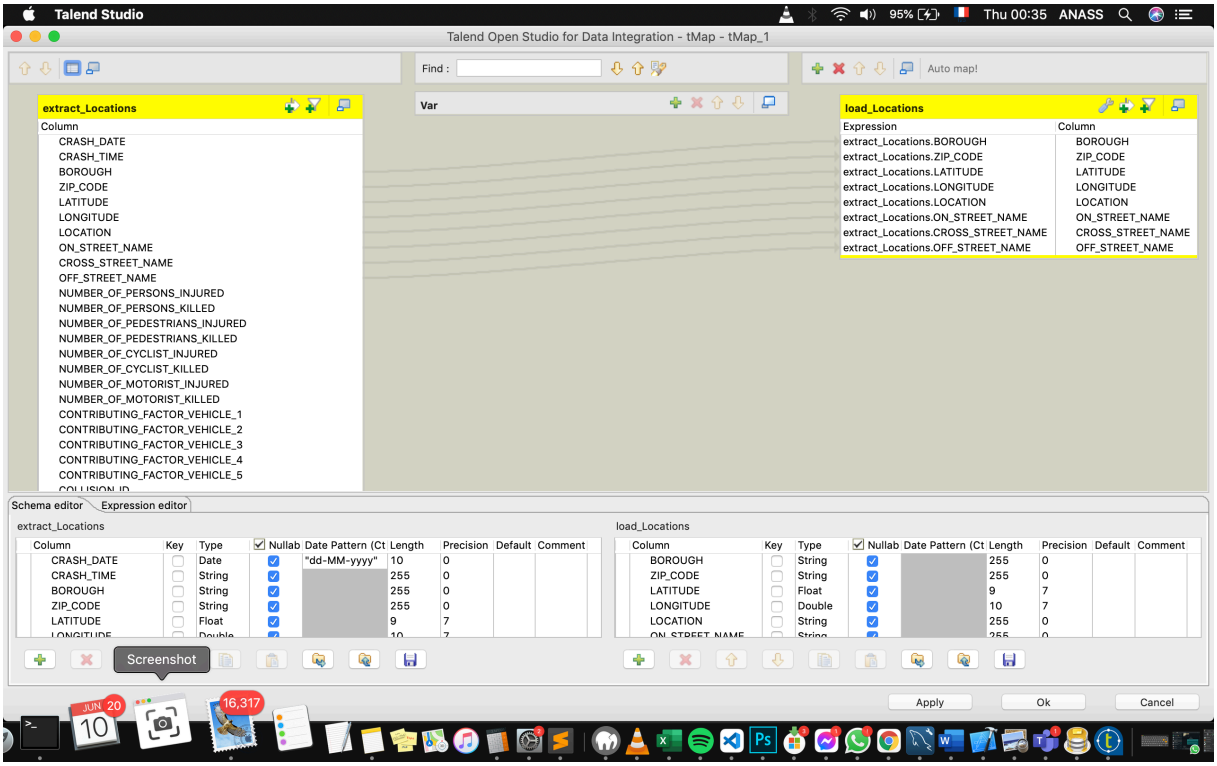
➢ ETL Table dimension Calendar :



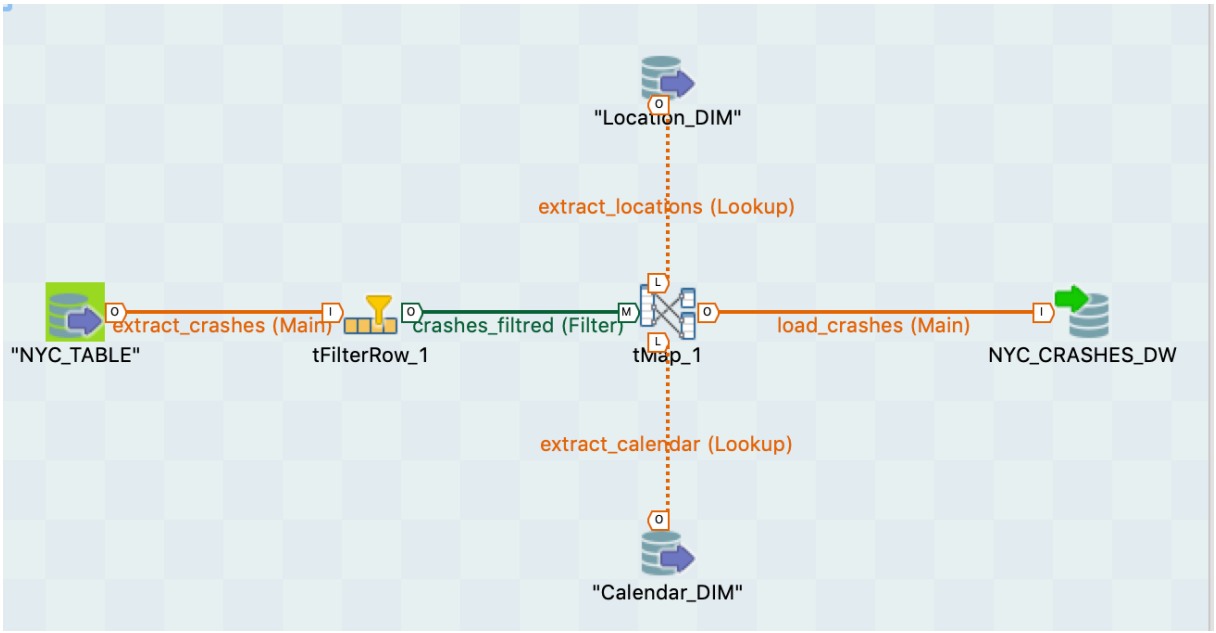


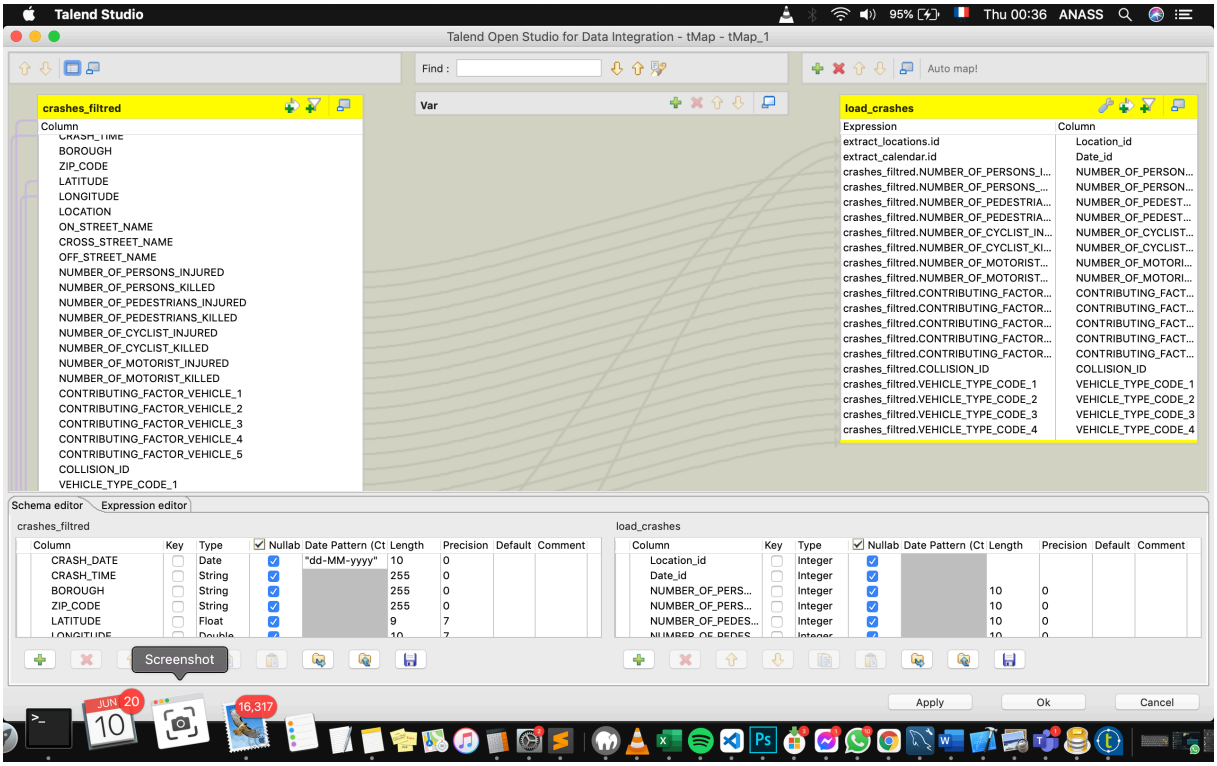
➢ ETL Table dimension Locations :

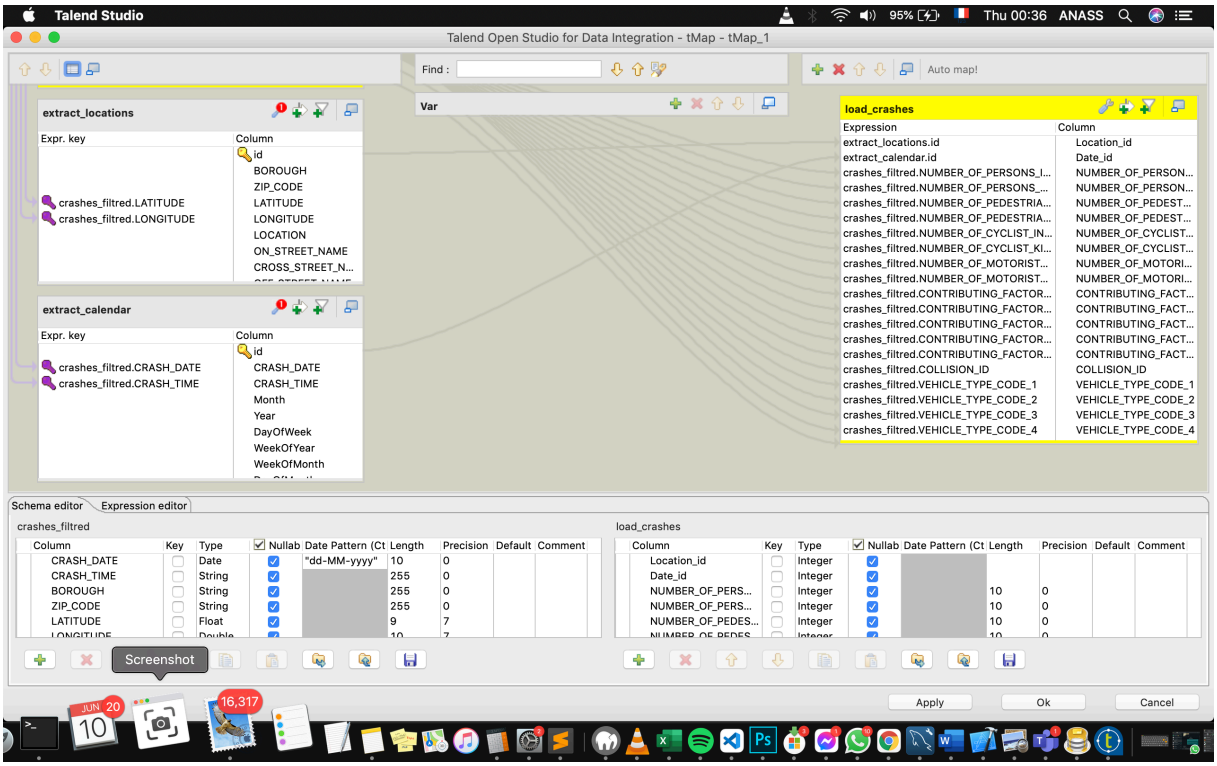




➢ ETL Table fait Crashes :







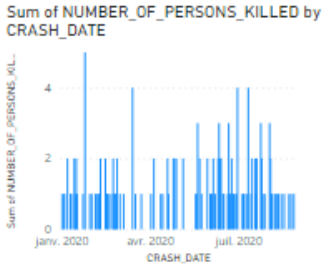
2.2 analyse & Reporting avec Power BI :

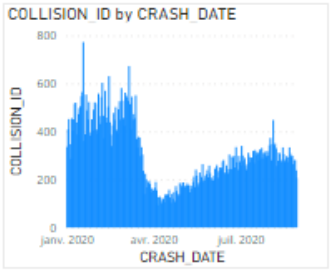
- **L’ensemble des graphes**

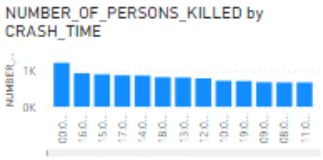
34,38 %
taux manquants zipcode in Locations

7,94 %
taux manquants latitude in Locations

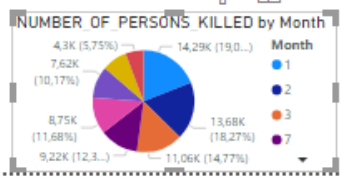
0,00 %
taux manquants time in calendar

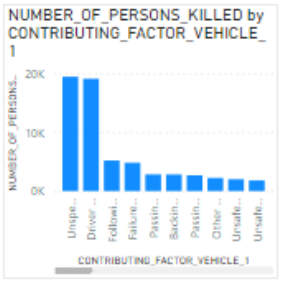


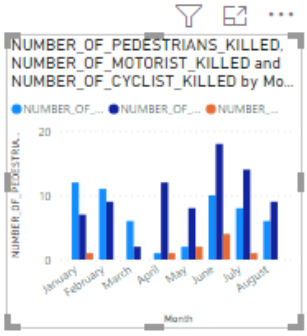




4
max of persons killed

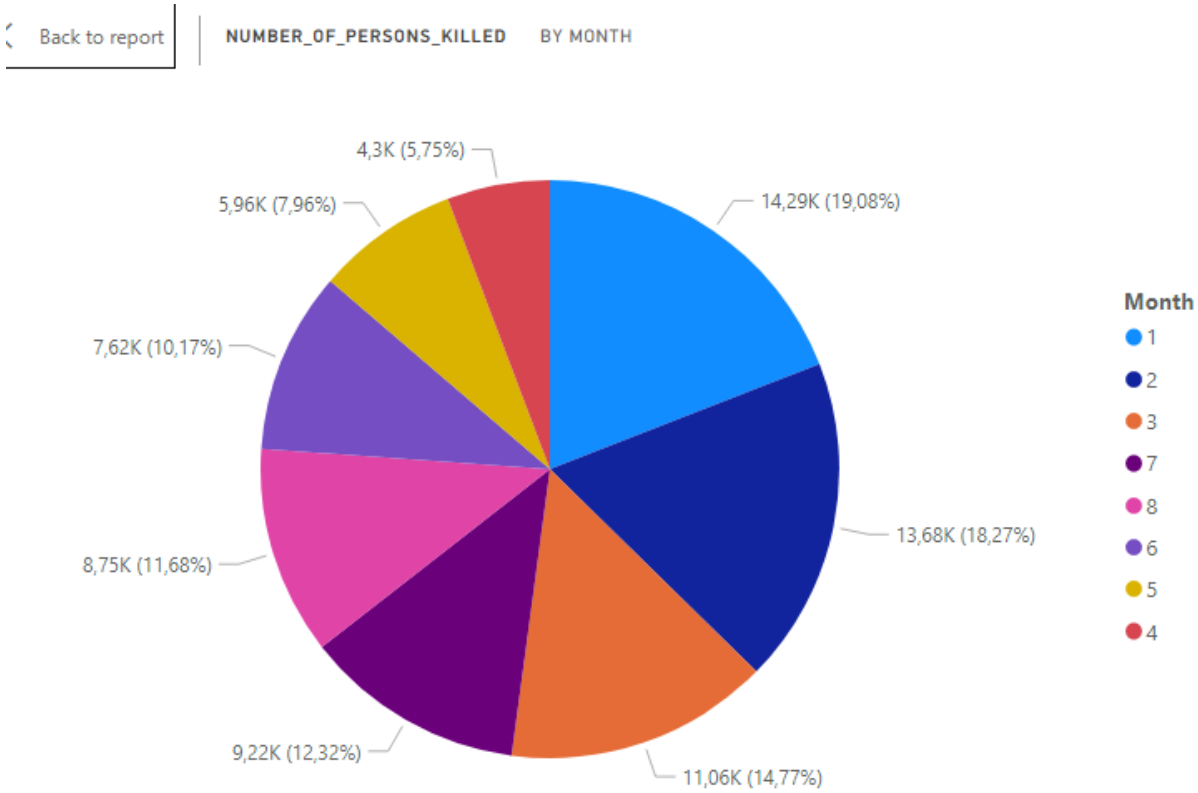




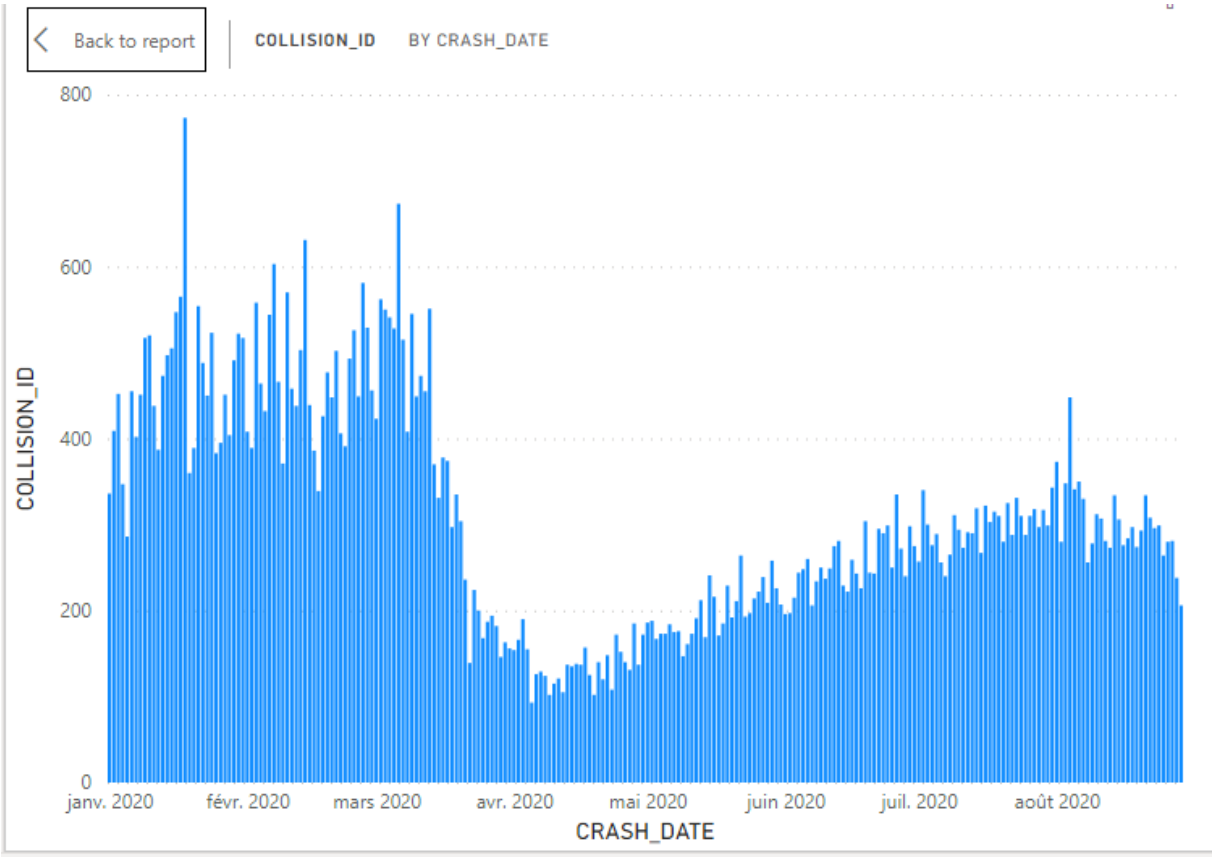


COLLISION_ID by LATITUDE and LONGITUDE

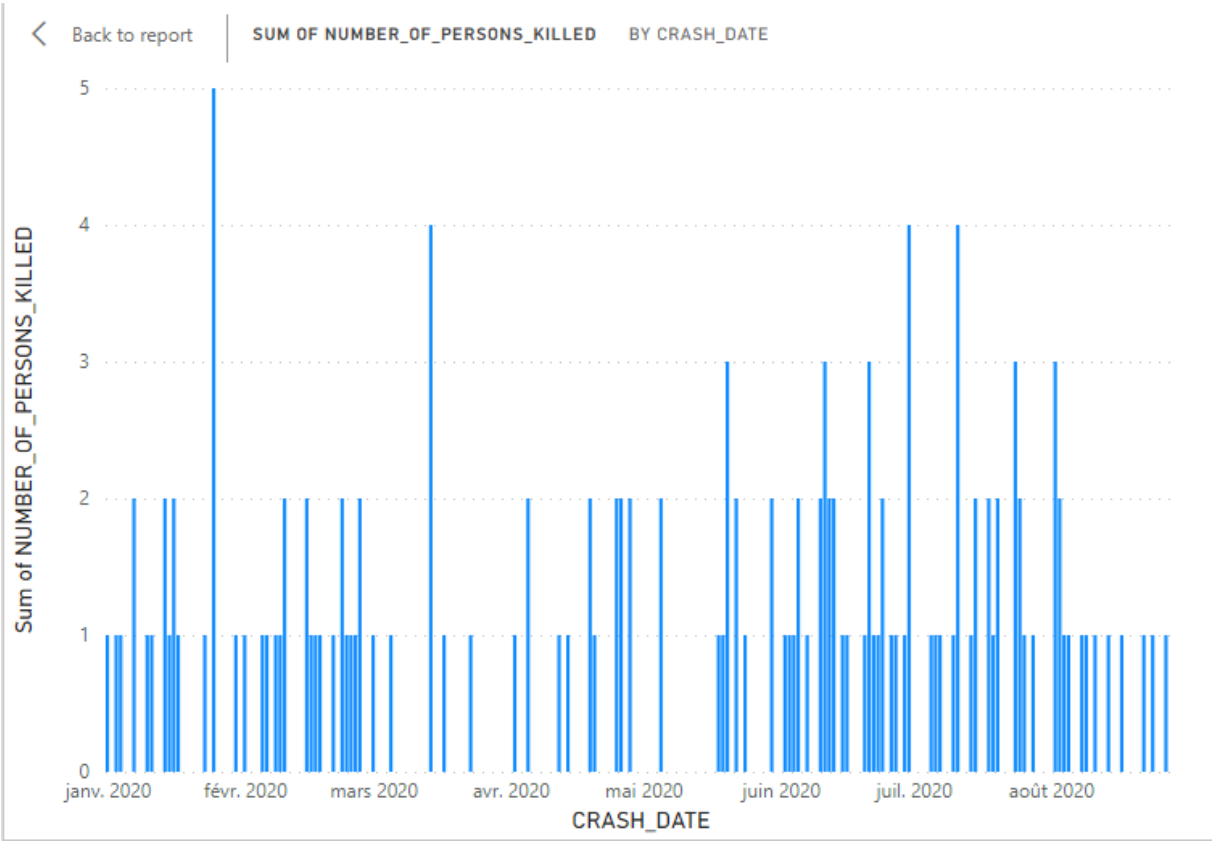
- **Nombres des personnes morts par mois**



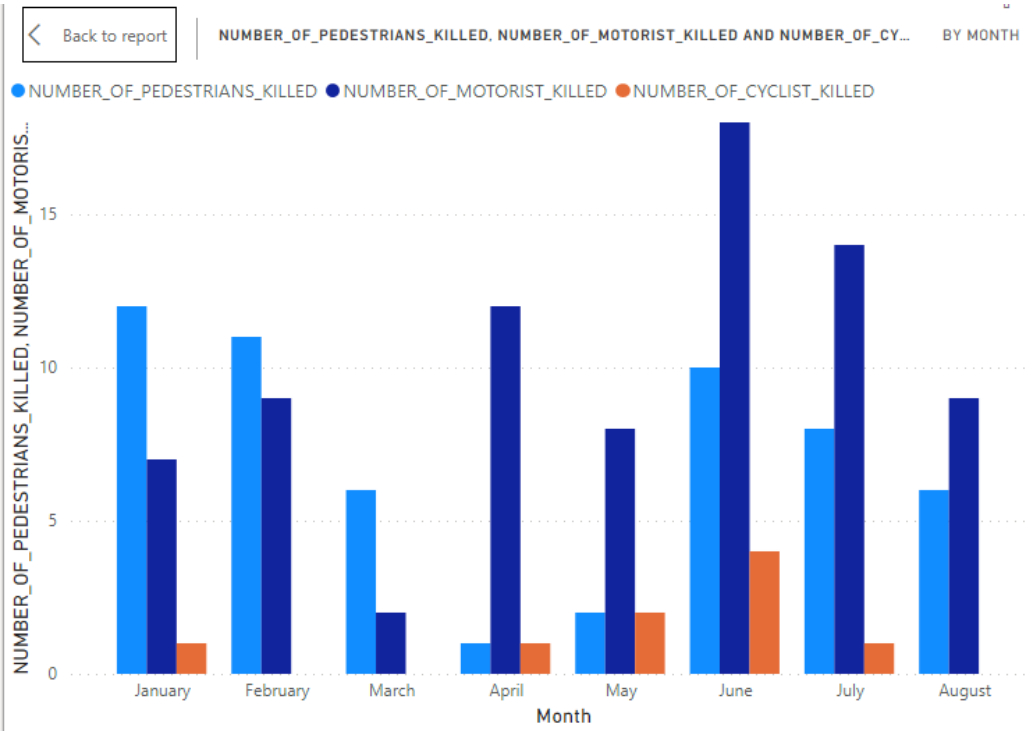
- **Nombres des accidents par date**



- **Nombre des personnes mort par date**



- **Nombre des personnes morts de chaque catégorie par mois**



- **Les emplacements des accidents dans la carte (map avec latitude et longitude)**