

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>



NYC Traffic Accidents

Motor vehicle collisions reported by the New York City Police Department from January-August 2020.

[Preview data](#) [Download](#) [☰](#)

FILE TYPES	TAGS	DATA STRUCTURE	# OF RECORDS	# OF FIELDS	DATE ADDED
CSV	Transportation Time Series Geospatial	Single table	15,407	29	10/27/2020

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

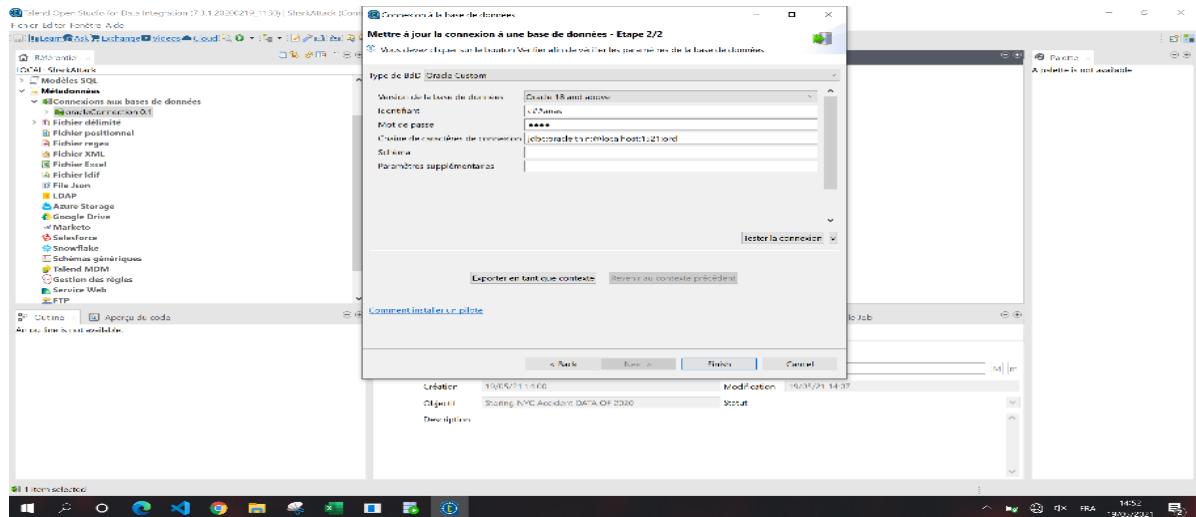
2.1. ETL

Dans cette phase j'ai essayé d'extraire 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.

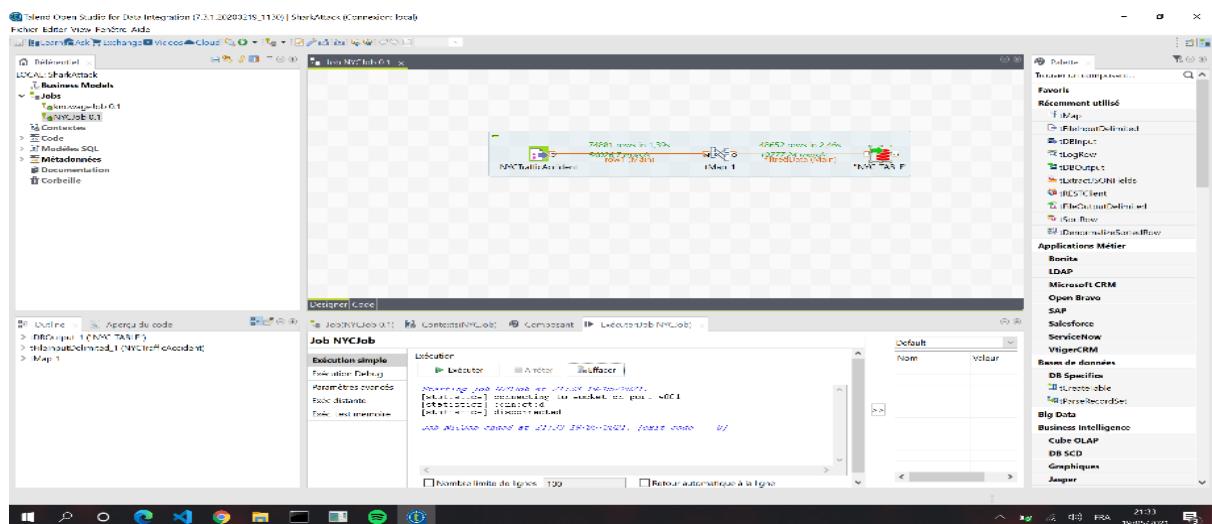
Premièrement j'ai commencé par l'importation de données à partir d'un fichier csv nommée NYC_accident.csv

POSSIBLE DATA LOSS		Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.																	
		Don't show again																	
A1	B1	C1	D1	E1	F1	G1	H1	I1	J1	K1	L1	M1	N1	O1	P1	Q1	R1	S1	
1	CRASH DATE	CRASH TIME	BOROUGH	ZIP CODE	LATITUDE	LONGITUDE	LOCATION	ON STREET NAME	CROSS STREET NAME	OFF STREET NAME	NUMBER OF PERSONS INJURED	NUMBER OF PERSONS KILLED	N						
2	2020-08-29 15:40:00	.00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00	BRONX	10466	40.8921	-73.8337	POINT (-73.8337640.8921)	PRATT AVENUE	STRANG AVENUE	,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	Passing Too Closely,Unspecified,,	,4342908,Sedan,Station Wagon/Sport Utility Vehicle,,							
3	2020-08-29 21:00:00	.00,BROOKLYN	12112	40.6095	-73.99194	POINT (-73.99194040.6095)	BUSHWICK AVENUE	PALMETTO STREET	,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	Reaction to Uninvolved Vehicle,Unspecified,,	,4343555,Sedan,,								
4	2020-08-29 18:20:00	.00,8165,-73.946556	POINT (-73.94655640.8165)	B8 AVENUE	,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	Backing Unsafely,,	,4343142,Station Wagon/Sport Utility Vehicle,,												
5	2020-08-29 00:00:00	.00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00	BRONX	10459	40.82472	-73.892996	POINT (-73.89299640.82472)	1047 SIMPSON STREET	,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	Unsafe Speed,Unspecified,Unspecified,Unspecified,,	,4343588,Station Wagon/Sport Utility Vehicle,,								
6	2020-08-29 00:00:00	.00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00	BRONX	10459	40.825226	-73.88778	POINT (-73.88778040.825226)	WOOHOO HAVEN BOULEVARD	,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	Failure to Stop at a Stop Sign,Unspecified,,	,4342723,Sedan,Station Wagon/Sport Utility Vehicle,,								
7	2020-08-29 19:30:00	.00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00	BRONX	10459	40.825226	-73.88778	POINT (-73.88778040.825226)	LONGFELLOW AVENUE	165 EAST 165 STREET	,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	Unsafe Speed,Unspecified,,	,4343004,Station Wagon/Sport Utility Vehicle,,							
8	2020-08-29 19:30:00	.00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00	BRONX	10459	40.825226	-73.88778	POINT (-73.88778040.825226)	WOOHOO HAVEN BOULEVARD	,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	Failure to Stop at a Stop Sign,Unspecified,,	,4342723,Sedan,Station Wagon/Sport Utility Vehicle,,								
9	2020-08-29 00:00:00	.00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00	BRONX	10406	-73.93538	POINT (-73.9353840.80016)	2 AVENUE	,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	Unsafe Lane Changing,Unspecified,,	,4343342,Station Wagon/Sport Utility Vehicle,,									
10	2020-08-29 19:50:00	.00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00	BRONX	10466	40.894314	-73.86027	POINT (-73.86027040.894314)	EAST 233 STREET	CARPENTER AVENUE	,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	Unsafe Speed,Unspecified,Unspecified,,	,4343030,Sedan,Station Wagon,,							
11	2020-08-29 20:00:00	.00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00	QUEENS	11385	40.70678	-70.90888	POINT (-70.90888040.70678)	555 WOODWARD AVENUE	,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	Unspecified,Unspecified,,	,4343040,Sedan,,,								
12	2020-08-29 07:00:00	.00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00	QUEENS	11436	40.680237	-73.77974	POINT (-73.77974040.680237)	116-52 141 STREET	,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	Drive Inattention/Distraction,Unspecified,Unspecified,,	,4342743,Sedan,Station Wagon,,								
13	2020-08-29 07:00:00	.00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00	QUEENS	11436	40.680237	-73.77974	POINT (-73.77974040.680237)	116-52 141 STREET	,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	Drive Inattention/Distraction,Unspecified,Unspecified,,	,4342743,Sedan,Station Wagon,,								
14	2020-08-29 21:33:00	.00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00	BRONX	10459	40.82965	-73.95178	POINT (-73.95178040.82965)	71 AVENUE	,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	Passing Too Closely,Unspecified,,	,4343448,Bus,Station Wagon,,								
15	2020-08-29 23:53:00	.00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00	BROOKLYN	11240	40.70166	-73.061464	POINT (-73.061464040.70166)	1 WILLIAMSBURG STREET	WEST 1ST AVENUE	,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	Failure to Yield Right-of-Way,Unspecified,,	,4342892,Bus,Station Wagon,,							
16	2020-08-29 04:14:00	.00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00	QUEENS	11385	73.84186	-73.84186	POINT (-73.84186040.85373)	WATERBURY AVENUE	,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	Unspecified,Unspecified,,	,4342716,Station Wagon/Sport Utility Vehicle,Motorcycle,,								
17	2020-08-29 03:35:00	.00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00	QUEENS	11385	40.65965	-73.773834	POINT (-73.77383404.65965)	ROCKAWAY BOULEVARD	NASSAU EXPRESSWAY	,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	Unspecified,Unspecified,,	,4342746,Sedan,Station Wagon/Sport Utility Vehicle,,							
18	2020-08-29 13:00:00	.00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00	BROOKLYN	11206	40.699707	-73.95178	POINT (-73.95178040.699707)	BEDFORD AVENUE	WALLABOUT STREET	,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	Driver Inattention/Distraction,Unspecified,,	,4343077,Station Wagon/Sport Utility Vehicle,,							
19	2020-08-29 10:30:00	.00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00	QUEENS	11385	40.7122	-73.86208	POINT (-73.86208040.7122)	METROPOLITAN AVENUE	COOPER AVENUE	,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	Failure to Yield Right-of-Way,Unspecified,,	,4342892,Station Wagon,,							
20	2020-08-29 12:29:00	.00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00,00	BRONX	10453	40.861862	-73.91282	POINT (-73.91282040.861862)	WEST FORDHAM ROAD	MAJOR DEEGAN EXPRESSWAY	,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	Pavement Slippery,View Obstructed/Limited,,								

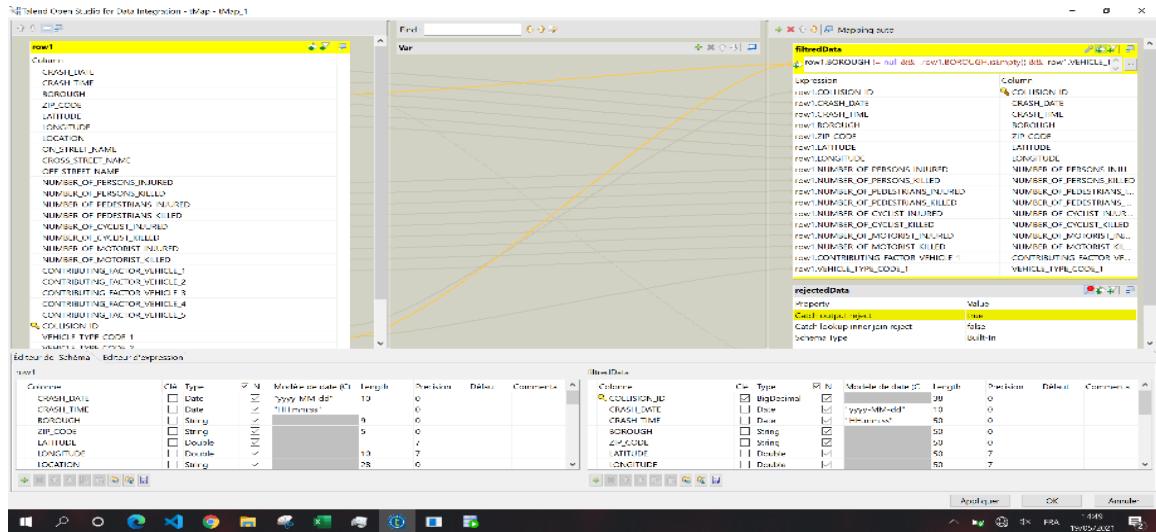
Après j'ai créer une connexion vers notre base de données mysql dans laquelle on va stocker les données traitées dans un table :



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 :



Ensute, 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 :

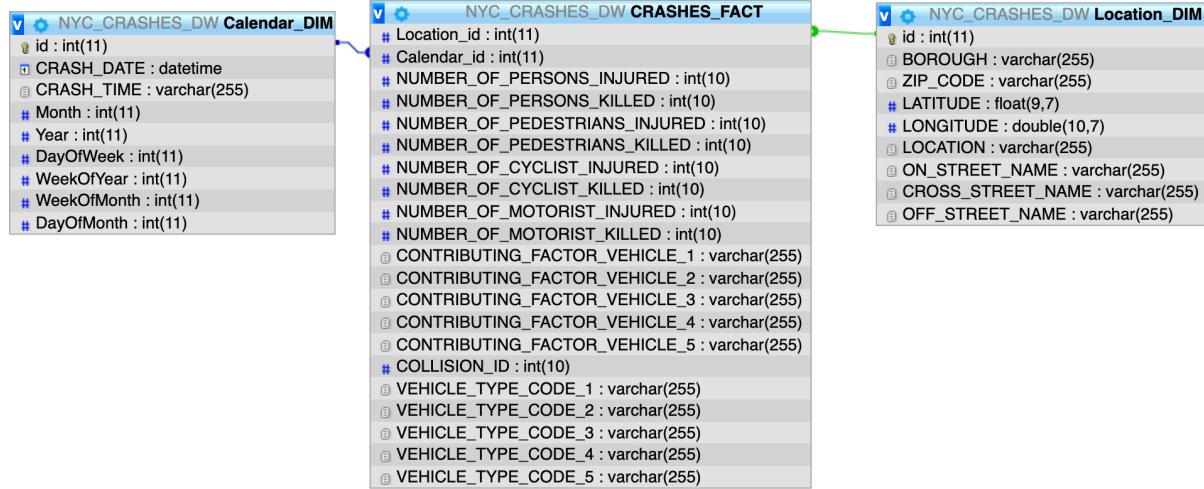


Et après l'exécution de mon job j'ai eu mes données transformées au niveau de notre base de données mysql :

(MySQL 5.7.25) 127.0.0.1/NYC_CRASHES/NYC_TABLE								
NYC_CRASHES		Structure	Content	Relations	Triggers	Table Info	Query	
Select Database								
TABLES	Search: CRASH_DATE							
NYC_TABLE	CRASH_DATE	CRASH_TIME	BOROUGH	ZIP_CODE	LATITUDE	LONGITUDE	LOCATION	ON_STREET_U
	2020-08-29	15:40:00	BRONX	10466	40.892103	-73.8337600	POINT (-73.83376 40.8921)	PRATT AVEN
	2020-08-29	21:00:00	BROOKLYN	11221	40.6904983	-73.9199140	POINT (-73.919914 40.6905)	BUSHWICK A'
	2020-08-29	18:20:00			40.8165016	-73.9465560	POINT (-73.946556 40.8165)	8 AVENUE
	2020-08-29	00:00:00	BRONX	10459	40.8247185	-73.8929600	POINT (-73.89296 40.82472)	
	2020-08-29	17:10:00	BROOKLYN	11203	40.6498909	-73.9338900	POINT (-73.93389 40.64989)	
	2020-08-29	03:29:00			40.6823082	-73.8449500	POINT (-73.84495 40.68231)	WOODHAVEN
	2020-08-29	19:30:00	BRONX	10459	40.8252258	-73.8877800	POINT (-73.88778 40.825226)	LONGFELLOW
	2020-08-29	00:00:00			40.8001595	-73.9353800	POINT (-73.93538 40.80016)	2 AVENUE
	2020-08-29	19:50:00	BRONX	10466	40.8943138	-73.8602700	POINT (-73.86027 40.894314)	EAST 233 ST
	2020-08-29	09:20:00	QUEENS	11385	40.7067795	-73.9088800	POINT (-73.90888 40.70678)	
	2020-08-29	00:07:00	QUEENS	11436	40.6802368	-73.7977400	POINT (-73.79774 40.680237)	
	2020-08-29	14:00:00	QUEENS	11433	40.7044220	-73.7928540	POINT (-73.792854 40.704422)	ARCHER AVE
	2020-08-29	21:33:00	BRONX	10455	40.8129654	-73.9161000	POINT (-73.9161 40.812965)	EAST 146 ST
	2020-08-29	22:53:00	BROOKLYN	11249	40.7016602	-73.9614640	POINT (-73.961464 40.70166)	WILLIAMSBUF
	2020-08-29	04:14:00			40.8537293	-73.8421860	POINT (-73.842186 40.8537293)	WATERBURY
	2020-08-29	06:35:00			40.6596489	-73.7738340	POINT (-73.773834 40.65965)	ROCKAWAY I
	2020-08-29	13:00:00	BROOKLYN	11206	40.6997070	-73.9571800	POINT (-73.95718 40.699707)	BEDFORD AV
	2020-08-29	10:30:00	QUEENS	11385	40.7122002	-73.8620800	POINT (-73.86208 40.7122)	METROPOLIT
	2020-08-29	12:29:00	BRONX	10453	40.8618622	-73.9128200	POINT (-73.91282 40.861862)	WEST FORDH
	2020-08-29	10:35:00	BROOKLYN	11211	40.7109566	-73.9511260	POINT (-73.951126 40.710957)	UNION AVEN
	2020-08-29	13:55:00	BROOKLYN	11231	40.6747284	-74.0002900	POINT (-74.00029 40.67473)	HAMILTON A
	2020-08-29	00:30:00			40.6658401	-73.7555100	POINT (-73.75551 40.66584)	BELT PARKW
	2020-08-29	06:30:00			40.6505203	-73.7309000	POINT (-73.73090 40.65052)	CRAFT AVEN
	2020-08-29	19:00:00			40.8396797	-73.9292760	POINT (-73.929276 40.83968)	MAJOR DEEG
	2020-08-29	01:45:00	MANHATTAN	10029	40.7947693	-73.9324700	POINT (-73.93247 40.79477)	
	2020-08-29	08:45:00	QUEENS	11411	40.7010422	-73.7463600	POINT (-73.74636 40.701042)	
	2020-08-29	23:19:00	BROOKLYN	11226	40.6396217	-73.9547700	POINT (-73.95477 40.63962)	NEWKIRK AVI
	2020-08-29	07:10:00			40.6743469	-73.8207100	POINT (-73.82071 40.674347)	118 STREET

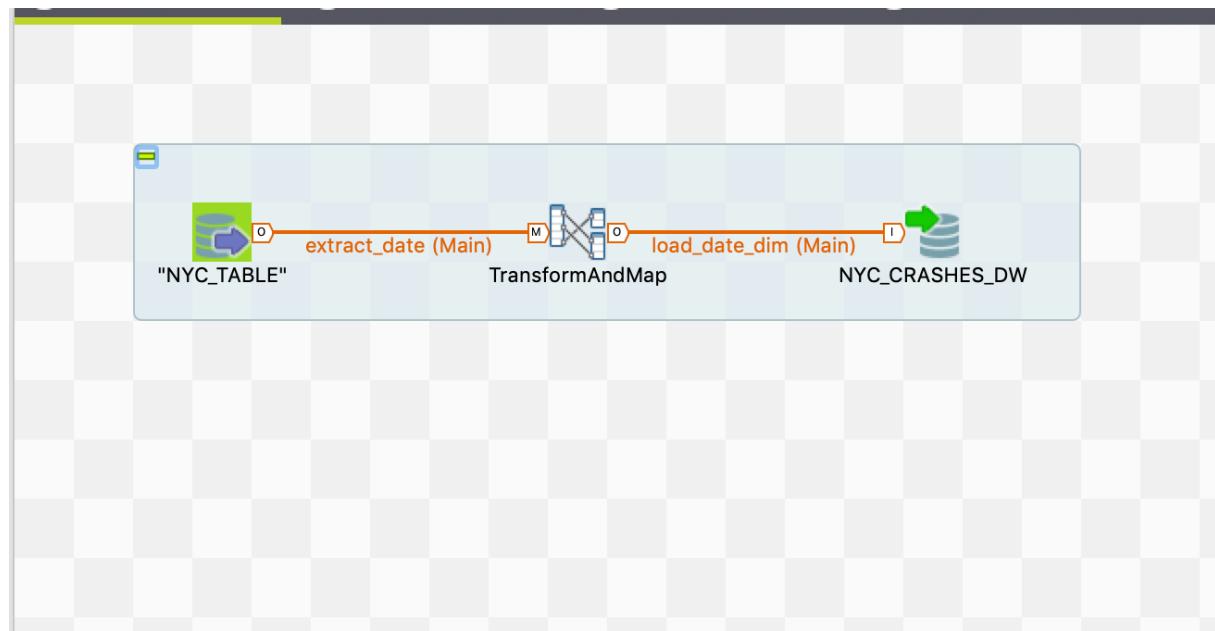
Et voilà la première phase est terminé et avant de passer à la deuxième phase je dois faire une modélisation de mon data Warehouse, cette modélisation qui est une modélisation multidimensionnelle contenant la table de fait qui est la mesure et les tables de dimensions qui sont liées à ce fait en formant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

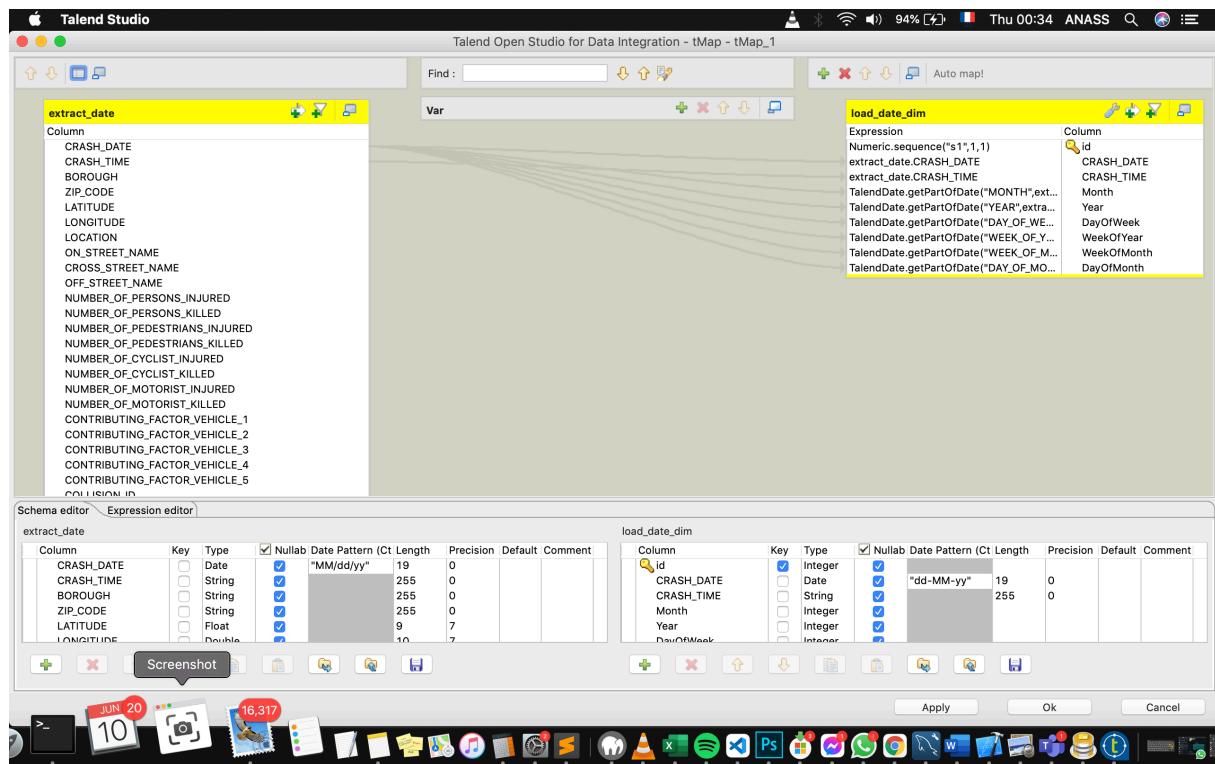
Dans mon cas la table de fait est les accidents à New York City lié à la dimension temps présenté par la table **Calendar_DIM** et la dimension du lieu présenté par la table **Location_DIM** et chaque table contenant des attributs



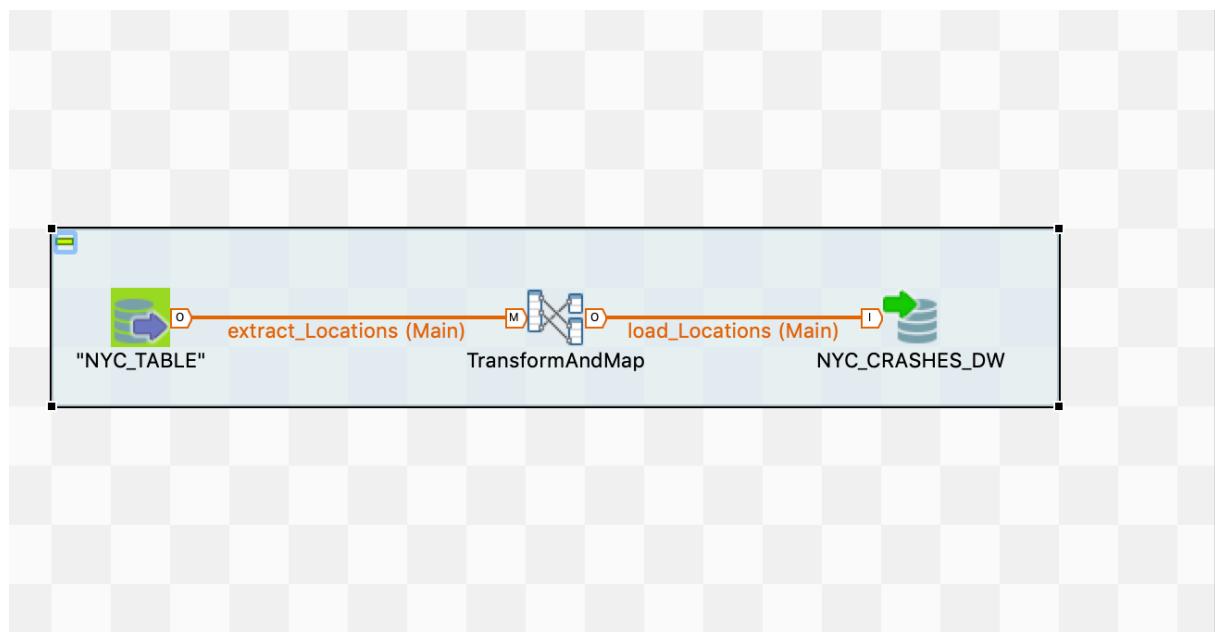
Donc après qu'on fait notre modélisation on va commencer ETL(extraction-Transform-LOAD)

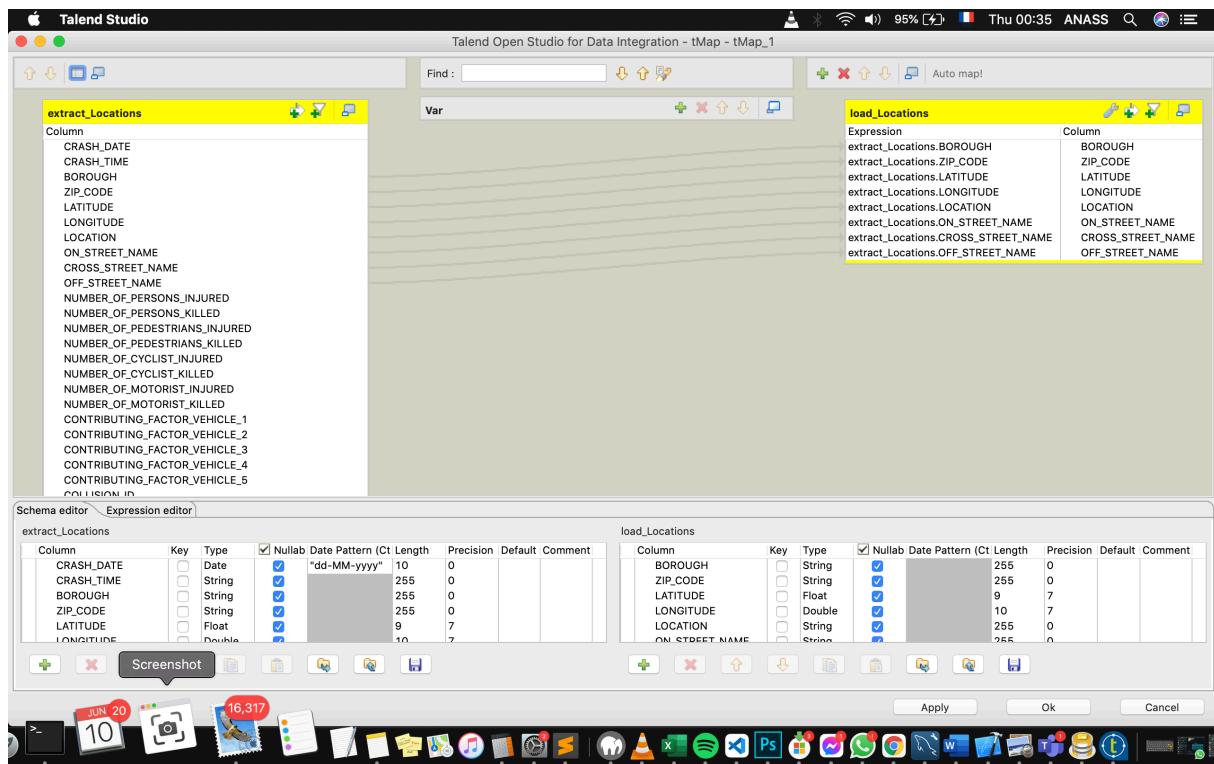
➤ ETL Table dimension Calendar :



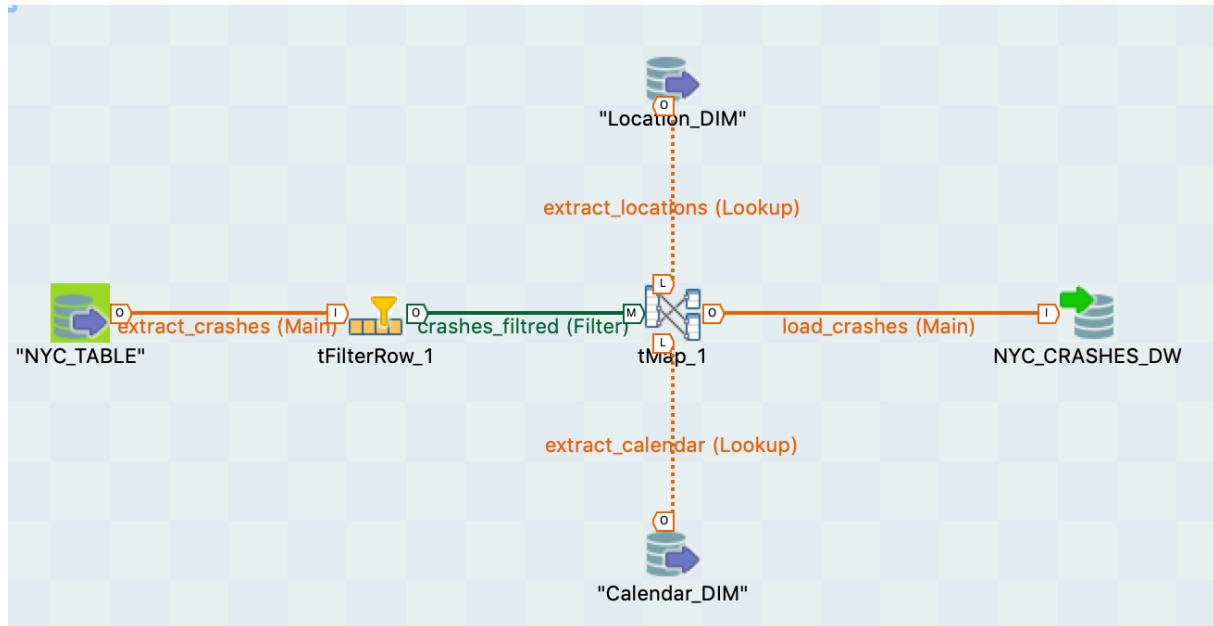


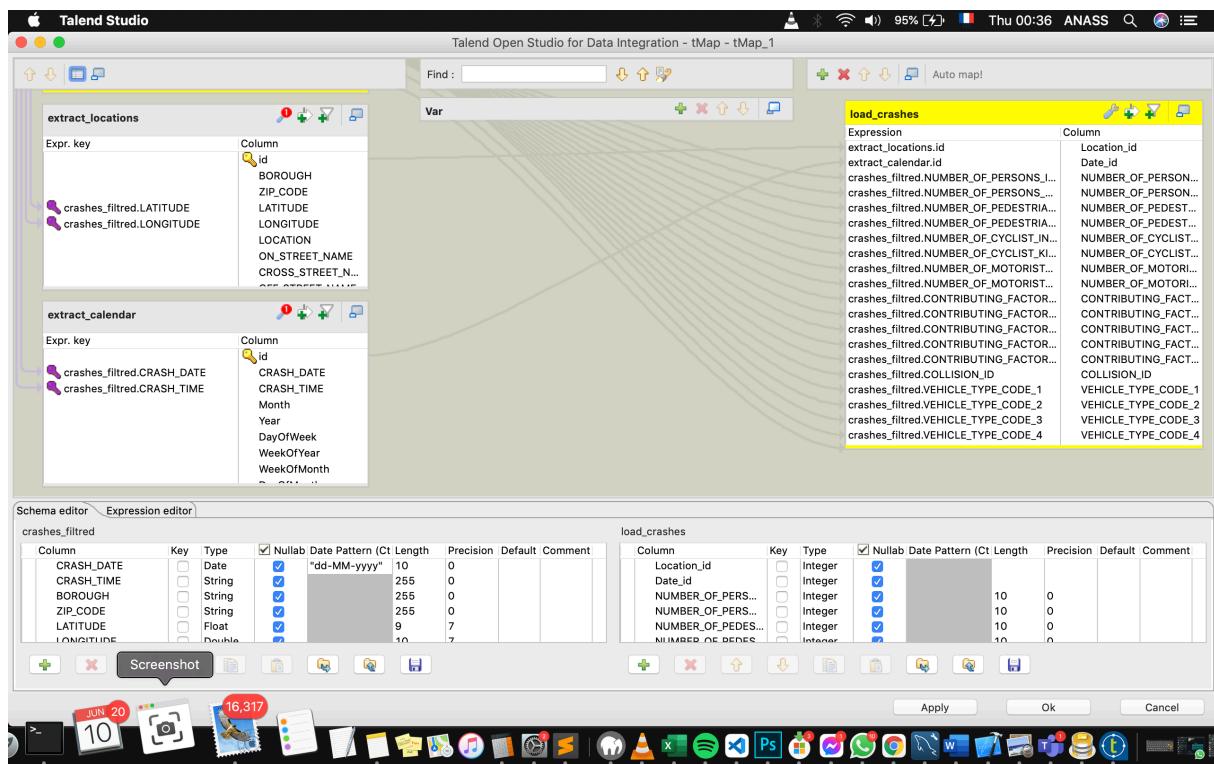
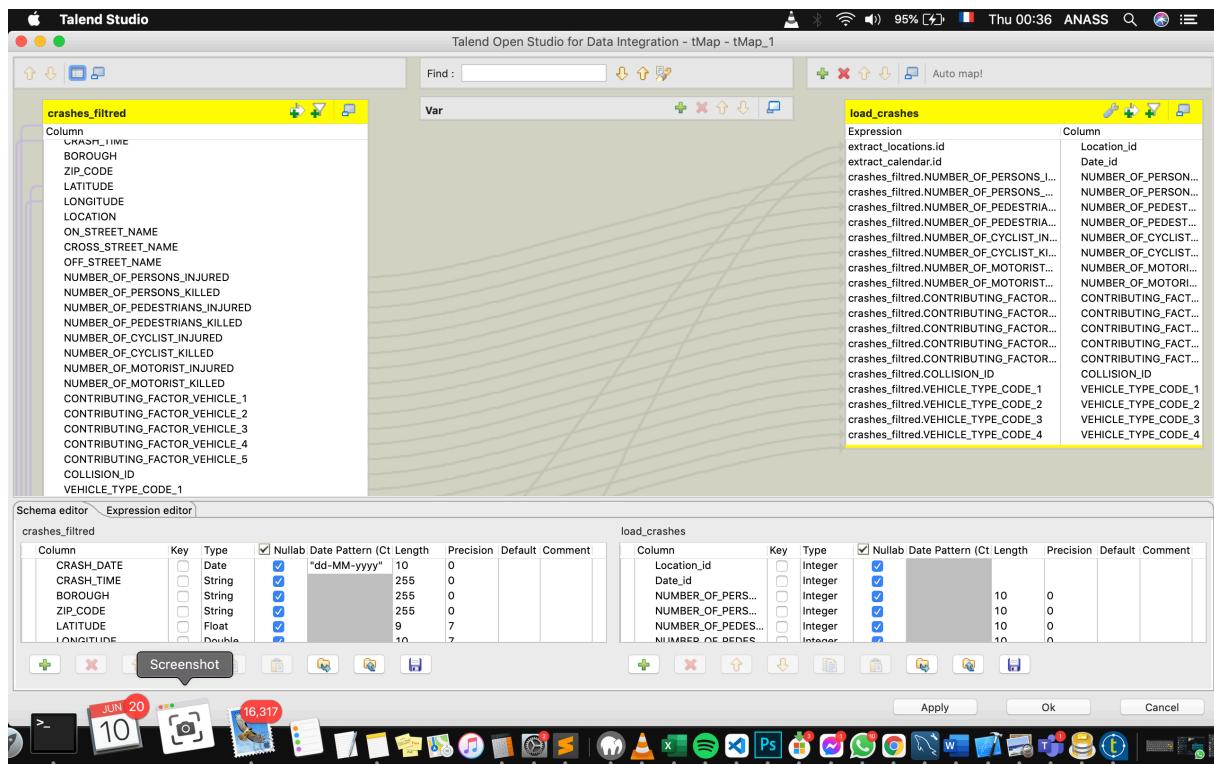
➤ ETL Table dimension Locations :





➤ ETL Table fait Crashes :





2.2 analyse & Reporting avec Power BI :

- L'ensemble des graphes

34,38 %

taux manquants zipcode in Locations

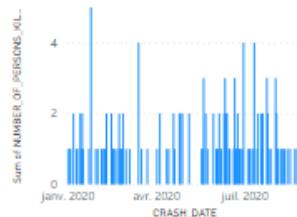
7,94 %

taux manquants latitude in Locations

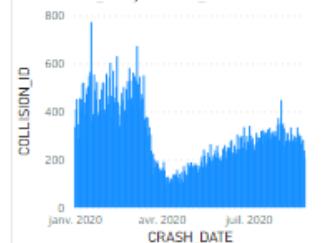
0,00 %

taux manquants time in calendar

Sum of NUMBER_OF_PERSONS_KILLED by CRASH_DATE



COLLISION_ID by CRASH_DATE



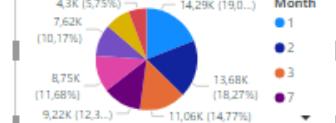
NUMBER_OF_PERSONS_KILLED by CRASH_TIME



4

max of persons killed

NUMBER_OF_PERSONS_KILLED by Month



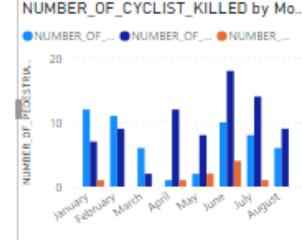
NUMBER_OF_PERSONS_KILLED by CONTRIBUTING_FACTOR_VEHICLE_1



COLLISION_ID by LATITUDE and LONGITUDE



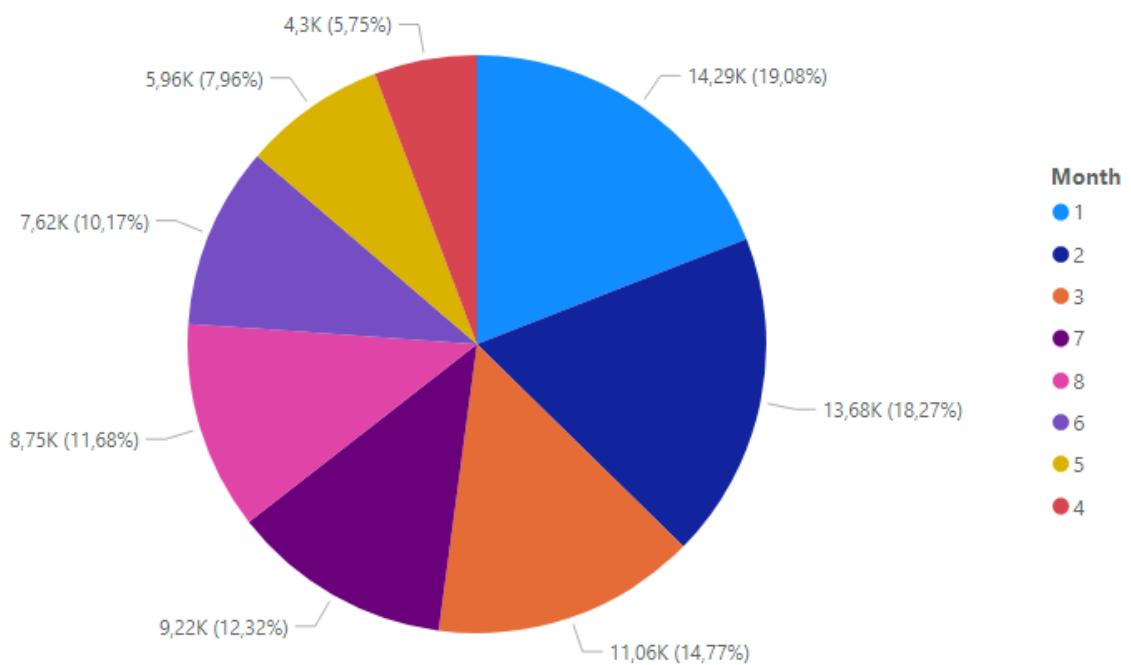
NUMBER_OF_PEDESTRIANS_KILLED, NUMBER_OF_MOTORIST_KILLED and NUMBER_OF_CYCLIST_KILLED by Mo...



- Nombres des personnes morts par mois

[Back to report](#)

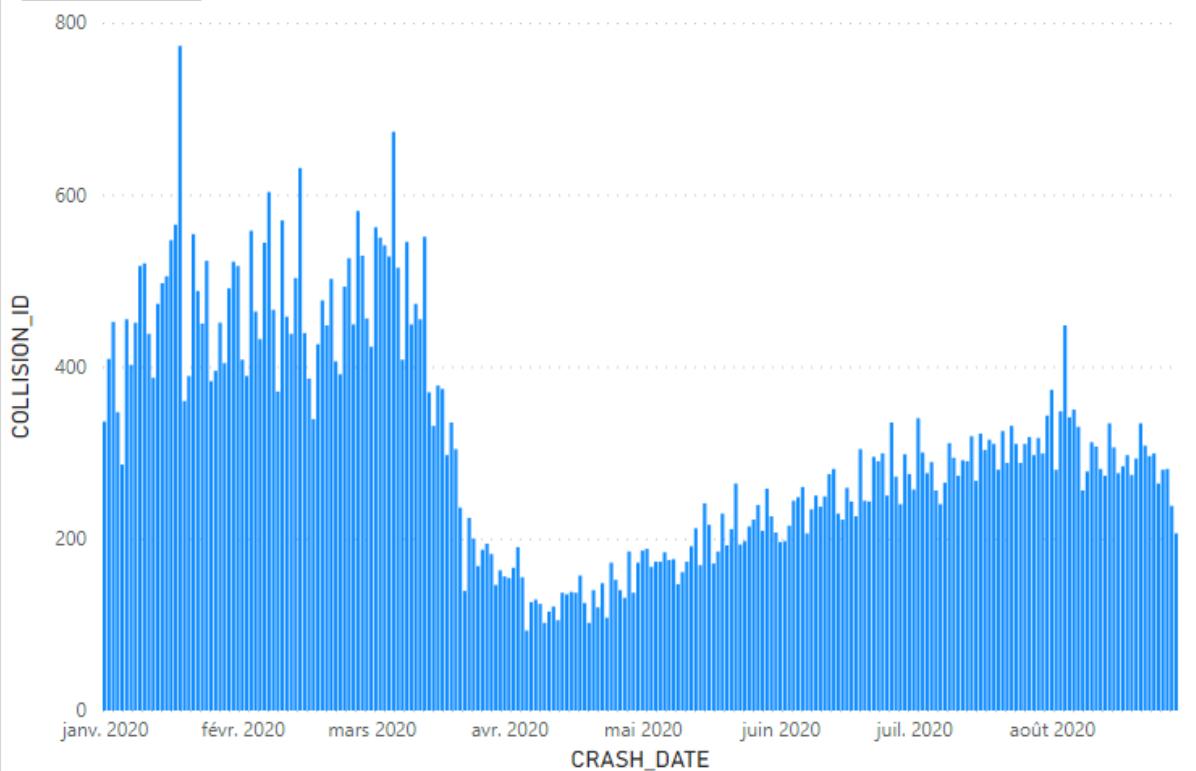
NUMBER_OF_PERSONS_KILLED BY MONTH



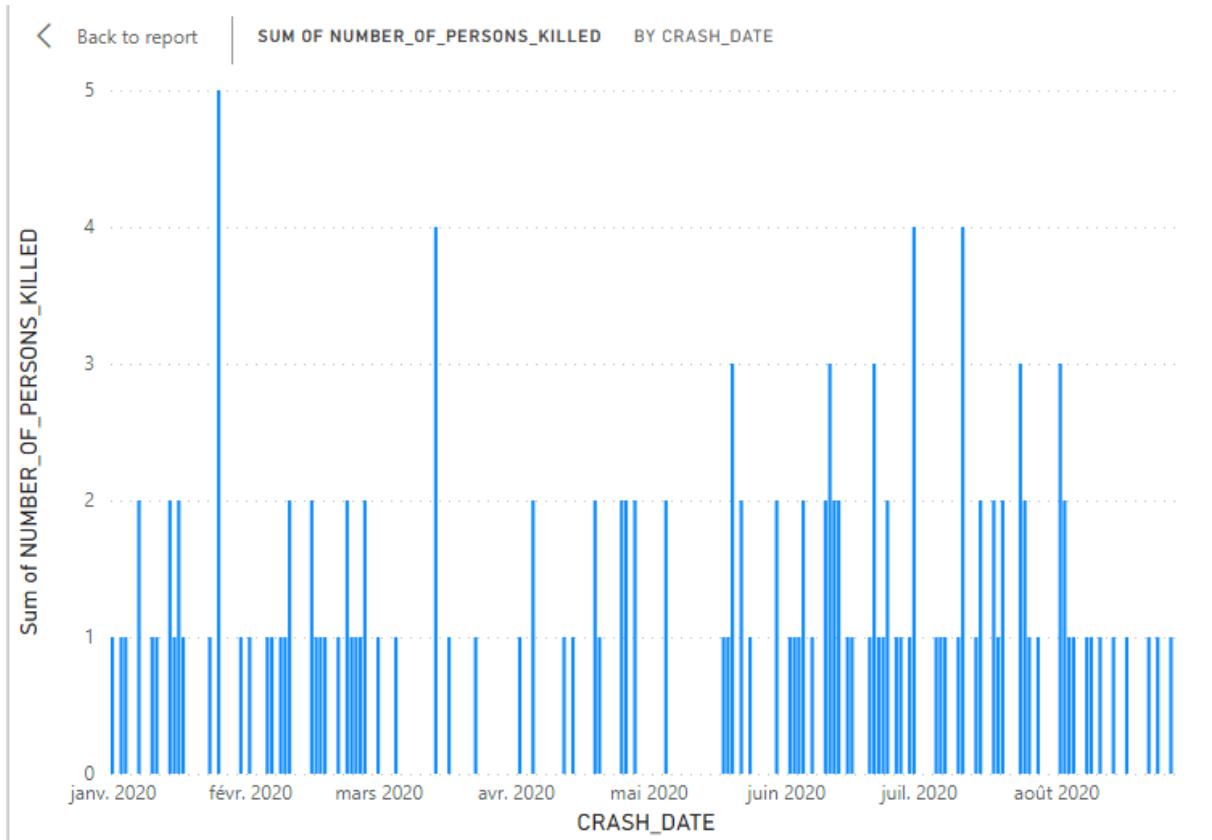
- Nombres des accidents par date

[Back to report](#)

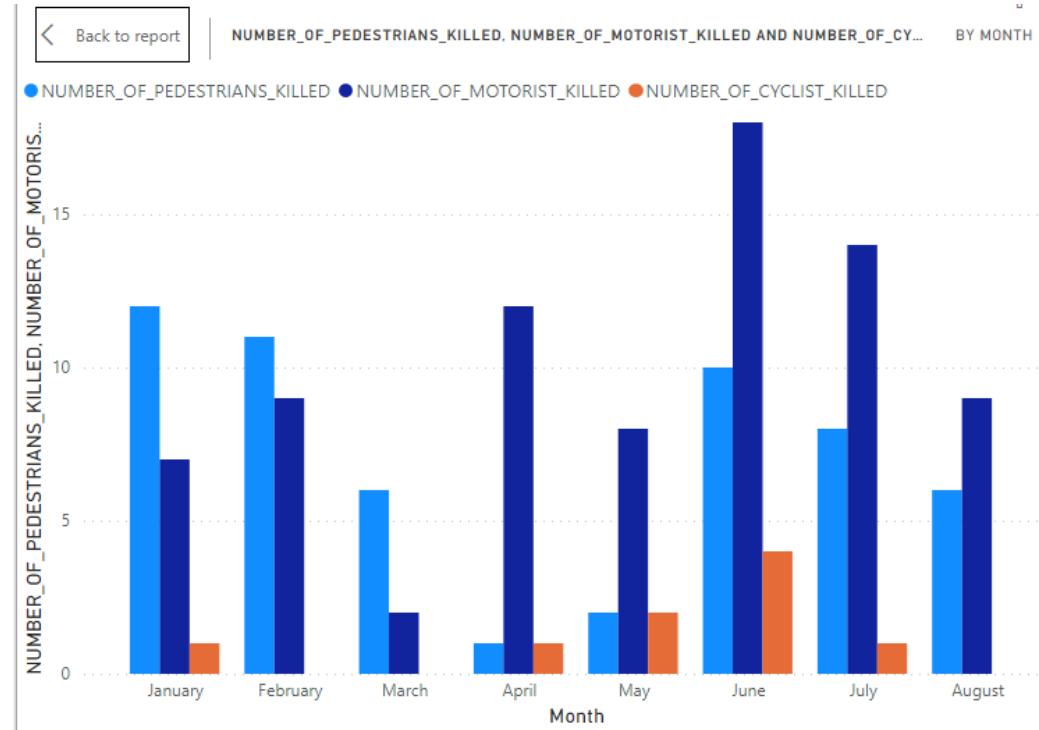
COLLISION_ID BY CRASH_DATE



- **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)

