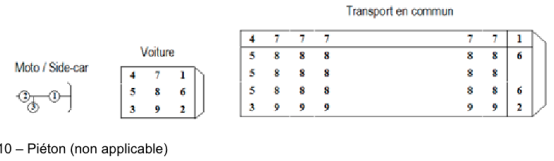
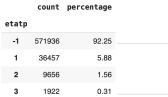
	Number of lines in the table : 273226	CHARACTERISTICS						
# Column	**Name of the Column**	**Description**	**Variable's type**	**Percentage of missing values**	**Categorical / Quantitative**	**Distribution**	**Comments**	**Keep?**
		What does this variable represent (from a business perspective ?)	*int64, float etc...* *If "object", develop.*	*in %*		*For categorical variables with less than 10 categories, list all categories.* *For quantitative variables, detail the distribution (basic descriptive statistics)*	*Free text*	*Should we keep or drop the column?*
1	*Num_Acc*	*Unique accident identifier number*	*integer*	0,00%	Unique Value			Keep
2	*mois (rename to month)*	*Month of the accident*	*integer*	0,00%	Categorical	*Integer from 1 to 12*		Keep
3	*jour (rename to day)*	*Day of the accident*	*integer*	0,00%	Categorical	*Integer from 1 to 31*		Keep
4	*hrmn*	*Time (hour and minute) of the accident*	*object*	0,00%	Datetime	*Datetime given in military time*		Keep
5	*lum (rename to light_cond)*	*Light conditions*	*integer*	0,00%	Categorical - 3 to 5 categories	*1 – Daylight* *2 – Dawn or dusk* *3 – Night without street lighting* *4 – Night with unlit street lighting* *5 – Night with lit street lighting*		Keep
6	*agg (rename to loc_type)*	*Location type*	*integer*	0,00%	Categorical - Binary	*1 – Outside built-up area* *2 – In built-up area*		Keep
7	*int (rename to intersc_type)*	*Intersection type*	*integer*	0,00%	Categorical - 6 to 10 categories	*1 – Not at intersection* *2 – Crossroads (X)* *3 – T-junction* *4 – Y-junction* *5 – Intersection with more than 4 branches* *6 – Roundabout* *7 – Square* *8 – Level crossing* *9 – Other intersection*		Keep
8	*atm (rename to weather_cond)*	*Weather conditions*	*integer*	0,01%	Categorical - 6 to 10 categories	*-1 – Not recorded* *1 – Normal* *2 – Light rain* *3 – Heavy rain* *4 – Snow or hail* *5 – Fog or smoke* *6 – Strong wind / storm* *7 – Dazzling conditions (sun glare, etc.)* *8 – Overcast* *9 – Other*		Keep
9	*col (rename to collision_type)*	*Type of collision*	*integer*	0,00%	Categorical - 6 to 10 categories	*-1 – Not recorded* *1 – Head-on (two vehicles)* *2 – Rear-end (two vehicles)* *3 – Side impact (two vehicles)* *4 – Chain collision (three or more vehicles)* *5 – Multiple collisions (three or more vehicles)* *6 – Other collision* *7 – No collision*		Keep
10	*com*	*Commune – INSEE commune code*	*object*	0,00%	Categorical			Keep
11	*adr*	*Postal address (for accidents in built-up areas)*	*object*	1,46%	Categorical			Drop
12	*lat*	*Latitude*	*object*	0,00%	Quantitative			Keep
13	*long*	*Longitude*	*object*	0,00%	Quantitative			Keep
14	*dep*	*Department – INSEE department code*	*object*	0,00%	Categorical		Could be a way to differenciate between mainland and oversees departments	Keep
15	*year*	*Year of the accident*	*integer*	0,00%	Categorical	*Integers from 2019 to 2023*		Keep

	Number of lines in the table : 467169	VEHICLES						
# Column	Name of the Column	Description	Variable's type	Percentage of missing values	Categorical / Quantitative	Distribution	Comments	Keep?
		What does this variable represent (from a business perspective ?)	*int64, float etc...* *If "object", develop.*	*in %*		*For categorical variables with less than 10 categories, list all categories.* *For quantitative variables, detail the distribution (basic descriptive statistics)*	*Free text*	*Should we keep or drop the column?*
1	*Num_Acc*	*Foreign key for Characteristics table*	*integer*	0,00%	Categorical			Keep
2	*senc (rename to direc_travel)*	*Direction of travel*	*integer*	0,00%	Categorical - 3 to 5 categories	*-1 – Not recorded* *0 – Unknown* *1 – Increasing PK/PR/address number* *2 – Decreasing PK/PR/address number* *3 – No reference point*		Drop
3	*catv (rename to vehicle_cat)*	*Vehicle category*	*integer*	0,00%	Categorical - more than 10 categories	*00 – Undetermined* *01 – Bicycle* *02 – Moped <50 cm3* *03 – Microcar (body quadricycle, formerly motor tricycle)* *07 – Passenger car (VL)* *10 – Light utility vehicle (1.5T <= GVWR <= 3.5T)* *13 – HGV (3.5T < GVWR <= 7.5T)* *14 – HGV (> 7.5T)* *15 – HGV + trailer* *16 – Road tractor* *17 – Road tractor + semi-trailer* *20 – Special vehicle* *21 – Agricultural tractor* *30 – Scooter <50 cm3* *31 – Motorcycle >50 cm3 and <=125 cm3* *32 – Scooter >50 cm3 and <=125 cm3* *33 – Motorcycle >125 cm3* *34 – Scooter >125 cm3* *35 – Light quad <=50 cm3 (non-body)* *36 – Heavy quad >50 cm3 (non-body)* *37 – Bus* *38 – Coach* *39 – Train* *40 – Tramway* *41 – 3RM <=50 cm3* *42 – 3RM >50 cm3 and <=125 cm3* *43 – 3RM >125 cm3* *50 – Motorized personal mobility device* *60 – Non-motorized personal mobility device* *80 – Electrically assisted bicycle* *99 - Other vehicle*	Group categories.	Keep
4	*occutc (rename to num_passenger_pub)*	*Number of passengers in public transport*	*float*	99,16%	Quantitative	*mean = 0,08; sd = 2,2; min = 0; max = 9; Q1, Q2 and Q3 = 0*		Drop
5	*obs (rename to fixed_obs_hit)*	*Fixed obstacle hit*	*integer*	0,00%	Categorical - more than 10 categories	*-1 – Not recorded* *0 – None* *1 – Parked vehicle* *2 – Tree* *3 – Metal guardrail* *4 – Concrete barrier* *5 – Other guardrail* *6 – Building/wall/bridge pier* *7 – Sign support/emergency phone post* *8 – Pole* *9 – Street furniture* *10 – Parapet* *11 – Island/refuge/high bollard* *12 – Kerb* *13 – Ditch/embankment/rock face* *14 – Other fixed obstacle on carriageway* *15 – Other fixed obstacle on pavement/shoulder* *16 – Off-road departure without obstacle* *17 – Culvert/aqueduct*		Keep
6	*obsm (rename to mobile_obs_hit)*	*Mobile obstacle hit*	*integer*	0,00%	Categorical - 6 to 10 categories	*-1 – Not recorded* *0 – None* *1 – Pedestrian* *2 – Vehicle* *4 – Rail vehicle* *5 – Domestic animal* *6 – Wild animal* *9 – Other*		Keep

	Number of lines in the table : 467169	VEHICLES						
# Column	Name of the Column	Description	Variable's type	Percentage of missing values	Categorical / Quantitative	Distribution	Comments	Keep?
		What does this variable represent (from a business perspective ?)	*int64, float etc...* *If "object", develop.*	*in %*		*For categorical variables with less than 10 categories, list all categories.* *For quantitative variables, detail the distribution (basic descriptive statistics)*	*Free text*	*Should we keep or drop the column?*
7	*choc (rename to int_impact_point)*	*Initial impact point*	*integer*	0,00%	Categorical - 6 to 10 categories	*-1 – Not recorded* *0 – None* *1 – Front* *2 – Front right* *3 – Front left* *4 – Rear* *5 – Rear right* *6 – Rear left* *7 – Right side* *8 – Left side* *9 – Multiple impacts (rollover)*		Keep
8	*manv (rename to maneuvre)*	*Main maneuver before accident*	*integer*	0,00%	Categorical - more than 10 categories	*-1 – Not recorded* *0 – Unknown* *1 – No change of direction* *2 – Same lane, same direction* *3 – Between two lanes* *4 – Reversing* *5 – Wrong way* *6 – Crossing central reservation* *7 – Bus lane, same direction* *8 – Bus lane, opposite direction* *9 – Merging* *10 – U-turn* *11 – Lane change left* *12 – Lane change right* *13 – Drift left* *14 – Drift right* *15 – Left turn* *16 – Right turn* *17 – Overtaking left* *18 – Overtaking right* *19 – Crossing carriageway20 – Parking maneuver* *21 – Avoidance maneuver* *22 – Opening door* *23 – Stopped (not parked)* *24 – Parked (with occupants)* *25 – Driving on pavement* *26 – Other maneuvers*		Keep
9	*num_veh*	*Vehicle identifier (alphanumeric)*	*object*	0,00%	Categorical		*Identification of the vehicle taken back for each user occupying this vehicle (including pedestrians who are attached to vehicles that hit them) - alphanumeric code - change to string*	Keep
10	*year*	*Year of accident*	*integer*	0,00%	Categorical	*Integers from 2019 to 2023*		Keep
11	*id_vehicle*	*Unique vehicle identifier (numeric)*	*object (integer)*	0,00%	Unique Value		*Together with num_veh used to merge to the characteristics file - change to integer*	Keep
12	*motor*	*Vehicle engine type*	*integer*	0,00%	Categorical - 6 to 10 categories	*-1 – Not recorded* *0 – Unknown* *1 – Fossil fuels* *2 – Hybrid electric* *3 – Electric* *4 – Hydrogen* *5 – Human powered* *6 – Other*		Keep

	Number of lines in the table : 619807	USERS							
# Column	Name of the Column	Description	Variable's type	Percentage of missing values	Categorical / Quantitative	Distribution	Comments	Keep?	
		What does this variable represent (from a business perspective ?)	*int64, float etc...* *If "object", develop.*	*in %*		*For categorical variables with less than 10 categories, list all categories.* *For quantitative variables, detail the distribution (basic descriptive statistics)*	*Free text*	*Should we keep or drop the column?*	
0	Num_Acc (rename num_accident)	Accident identifier	int64	0,00%	Categorical - more than 10 categories	Foreign key - Characteristics table - used to merge		Keep	
1	id_vehicule (rename id_vehicle)	Unique numeric identifier for each vehicle involved in the accident	object	0,00%	Categorical - more than 10 categories	Foreign key - Vehicles table - used to merge	Change type to integer	Keep	
2	num_veh (rename num_vehicle)	Alphanumeric vehicle identifier used to link each user (driver, passenger, pedestrian) to the corresponding vehicle	object	0,00%	Categorical - more than 10 categories	Foreign key - Vehicle table -used to merge	Change type to string due to alphanumeric values	Keep	
3	place	Allows to locate the place occupied in the vehicle by the user at the time of the accident	int64	0,00%	Categorical - 6 to 10 categories	General codes (apply to all vehicles): -1 – Not specified NaN – Missing (no value recorded in dataset) Motorcycle / Sidecar 1 – Driver (rider) 2 – Passenger behind the rider (pillion) 3 – Sidecar passenger Car (Passenger car) 1 – Driver seat 2 – Front passenger seat 3 – Rear side seat (right) 4 – Rear side seat (left) 5 – Rear middle seat 6 – Front middle seat (bench, if present) 7 – Other seat in the vehicle (e.g., 3rd row) 8 – Other/unspecified passenger position 9 – Unknown seat Public transport (bus/coach, etc.) 1 – Driver 6 – Door/platform area 7 – Seated passenger 8 – Standing passenger (aisle/interior) 9 – Other/unspecified area Pedestrian 10 – Pedestrian (not applicable) - not found in dataset	1. Value 0 is in dataset, but no describtion what it means 2.Change type to int64 instead of float64	Drop	
4	catu (rename category_user)	User category	int64	0,00%	Categorical - 3 to 5 categories	1 - Driver 2 - Passenger 3 - Pedestrian		Keep	
5	grav (rename injury_severity)	Severity of the accident: The injured users are classified into three categories of victims plus the uninjured	int64	0,00%	Categorical - 3 to 5 categories	-1 - Not specified 1 - Unscathed 2 - Killed 3 - Hospitalized wounded 4 - Light injury	Main indicator the measure for the project; -1 probably means not specified? There are 419 users without grav value (0,07 % - so no major effect when excluded)	Keep	
6	sexe (rename to gender)	Sex of the user	int64	0,00%	Categorical - Binary	-1 - Not specified 1 - Male 2 - Female	1. Decide if we want to leave the missing values out or replace it with mode or the distrubution between male/female	Keep	
7	an_nais (rename year_of_birth)	Year of birth of the user	float64	1,38%	Categorical - more than 10 categories	Ranging from 1900.0 to 2023.0	1. Change the type to int64 instead of float64 2. Change or add Age (year of birth minus year) 3. Decide if we want to leave the missing values out or replace it with mean	Keep	
8	trajet (rename to trip_purpose)	Reason for traveling at the time of the accident	int64	0,00%	Categorical - 6 to 10 categories	-1 - Not specified 0 - ?? 1 - Home - work 2 - Home - school 3 - Shopping - Shopping 4 - Professional use 5 - Promenade - leisure 9 - Other	No check what 0 means due to dropping the column	Keep	

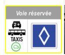
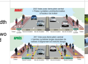
	Number of lines in the table : 619807	USERS							
# Column	Name of the Column	Description	Variable's type	Percentage of missing values	Categorical / Quantitative	Distribution	Comments	Keep?	
		What does this variable represent (from a business perspective ?)	*int64, float etc...* *If "object", develop.*	*in %*		*For categorical variables with less than 10 categories, list all categories.* *For quantitative variables, detail the distribution (basic descriptive statistics)*	*Free text*	*Should we keep or drop the column?*	
9	secu1 (rename safety_equipment1)	If saftey equipment was used and which one	int64	0,00%	Categorical - more than 10 categories	-1 – Not specified 0 – No equipment 1 – Seatbelt 2 – Helmet 3 – Child restraint system 4 – Reflective vest 5 – Airbag (2-wheel/3-wheel vehicles) 6 – Gloves (2-wheel/3-wheel vehicles) 7 – Gloves + Airbag (2-wheel/3-wheel vehicles) 8 – Not determinable 9 – Other	Since 2019, the variables secu1, secu2, and secu3 allow recording up to three types of safety equipment per user (e.g., helmet, gloves, airbag), replacing the previous single-variable (kind of equipment + if used) system. Here we have not only not specified (kinda like missing value) but also not determinable (it was not possible to determine the safety equipment), with 1 % not specified and 12 % not determinable, we have to decide how to work with this variable	Keep	
10	secu2 (rename safety_equipment2)	If second saftey equipment was used and which one	int64	0,00%	Categorical - more than 10 categories	see secu1	see secu1	Keep	
11	secu3 (rename safety_equipment3)	If third saftey equipment was used and which one	int64	0,00%	Categorical - more than 10 categories	see secu1	see secu1	Keep	
12	locp (rename location_pedestrian)	Location of the pedestrian	int64	0,00%	Categorical - 6 to 10 categories	-1 – Not specified 0 – Not applicable (no pedestrian) On the roadway: 1 – More than 50 m from a pedestrian crossing 2 – Within 50 m of a pedestrian crossing On pedestrian crossing: 3 – Without traffic light signal 4 – With traffic light signal Other: 5 – On the sidewalk 6 – On the roadside shoulder 7 – On a refuge island or emergency lane (BAU) 8 – On a service road 9 – Unknown	with 45 % not specified, would probably drop the column	Drop	
13	actp (rename action_pedestrian)	Action of the pedestrian	object	0,00%	Categorical - 6 to 10 categories	-1 – Not specified 0 – Not applicable (no pedestrian) While moving: 1 – In the direction of the oncoming vehicle 2 – In the opposite direction of the vehicle Other: 3 – Crossing 4 – Hidden (obstructed view) 5 – Playing / running 6 – With an animal 9 – Other A – Getting on / off a vehicle B – Unknown	with 40 % not specified, would probably drop the column	Drop	
14	etatp (rename company_pedestrian)	Specify whether the injured pedestrian was alone or not	int64	0,00%	Categorical - 3 to 5 categories	-1 - Not specified 1 - Alone 2 - Accompanied 3 - In a group	with 92 % not specified, would probably drop the column	Drop	
15	id_usager (rename id_user)	Unique user identifier (numeric)	string	38,43%	Categorical - more than 10 categories	First time added in year 2021, therefore missing values for the years, 2019-2020	Drop due to high Percentage of missing values	Drop	
16	year	year of the accident	int64	0,00%	Categorical - 3 to 5 categories	Ranging from 2019 to 2023		Keep	

secu1	count	percentage
1	365320	58.93
2	113314	18.28
8	74595	12.03
0	53120	8.57
-1	7883	1.27
3	3884	0.63
9	658	0.11
6	534	0.09
4	401	0.06
5	246	0.04
7	16	0.00

locp	count	percentage
0	291911	47.08
-1	280083	45.18
3	15048	2.43
2	10413	1.68
4	7207	1.16
1	6968	1.12
5	3770	0.61
9	2375	0.38
6	1445	0.23
8	698	0.11
7	53	0.01

actp	count	percentage
0	322934	52.09
-1	248919	40.15
3	35353	5.70
9	3226	0.52
1	3133	0.51
5	1878	0.30
B	1608	0.26
2	1479	0.24
4	621	0.10
A	540	0.09
6	135	0.02
7	84	0.01
8	61	0.01

etatp	count	percentage
-1	571936	92.25
1	36457	5.88
2	9656	1.56
3	1922	0.31

	Number of lines in the table : 289264	PLACES										
# Column	Name of the Column	Description	Variable's type	Percentage of missing values	Categorical / Quantitative	Distribution	Comments				Keep?	
		What does this variable represent (from a business perspective ?)	*int64, float etc...* *If "object", develop.*	*in %*		*For categorical variables with less than 10 categories, list all categories.* *For quantitative variables, detail the distribution (basic descriptive statistics)*	*Free text*				*Should we keep or drop the column?*	
0	Num_Acc	Accident ID	int64	0,00%			IDENTIFIER - LINK TO OTHER TABLES				Keep	
1	catr (rename: road_category)	Category of road	float64	0,00%	Categorical - 6 to 10 categories	1 - Highway 2 - National Road 3 - Departmental Road 4 - Communal Way 5 - Off public network 6 - Parking lot open to public traffic 7 - Urban Metropolitan Road 9 - other	No missing value				Keep	
2	voie (rename: road_nb)	Road number	object	12,38%	Quantitative		Roads in France can contain letters or numbers so makes sense that it is an 'Object'.				Drop	
3	v1 (rename: num_index)	Numeric index of route	float64	3,71%	Quantitative		e.g. Road 2 bis, Road 3 ter				Drop	2019 - 2023 Missing %
4	v2 (rename: lett_index)	Letter index of route	object	91,90%	Quantitative		Only letters so make sense to be an 'Object' e.g. Road 158A would be road_nb = 158 and lett_index = "A"				Drop	
5	circ (rename: traff_reg)	Traffic regime	float64	0,00%	Categorical - 3 to 5 categories	-1 - Not specified 1 - One way 2 - Bidirectional 3 - Separated carriageways 4 - With variable assignment channels					Keep	
6	nbv (rename: traff_lane_nb)	Number of traffic lane	float64	2,22%	Quantitative	X					Keep	
7	vosp (rename: rsrvd_lane)	Reserved lane	float64	0,00%	Categorical - 3 to 5 categories	-1 - Not specified 0 - None 1 - Bike path 2 - Cycle Bank 3 - Reserved channel					Keep	
8	prof (rename: road_slp)	Road slope	float64	0,00%	Categorical - 3 to 5 categories	-1 - Not specified 1 - Dish/Flat 2 - Slope 3 - Hilltop 4- Hill bottom					Keep	
9	pr (rename: ref_point)	Reference point	object	0,00%	Quantitative			Judging from the data it seems to be the 'Point KM' that is noted. We also have the type of road and the road number as catr and road_nb	UNIQUE: array([1.0, 0.0, 24.0, ..., '1800', '1579', '433'], dtype=object) Some are considered numbers (e.g. 24,0, 1,0) some are considered text (e.g. 1800, 433)	CAN CONVERT TO FULL FLOAT	Drop	
10	pr1 (rename: ref_point_dis)	Distance from reference point (meters)	object	0,00%	Quantitative			UNIQUE: array([430.0, 0.0, 630.0, ..., '2012', '191', '2287'], dtype=object) - Some are considered numbers (e.g. 430,0, 630,0) some are considered text (e.g. 2012, 2287)	CAN CONVERT TO FULL FLOAT		Drop	
11	plan (rename: road_algmt)	Road alignement	float64	0,00%	Categorical - 3 to 5 categories	-1 - Not specified 1 - Straight part 2 - Curved on the left 3 - Curved right 4 - In "S"					Keep	
12	lartpc	Width of central median (meters)	float64	99,80%	Quantitative		This describes the width of the central median (the strip separating two directions of a divided road).				Drop	
13	larrout (rename: road_width)	Roadway width (meters)	float64	20,20%	Quantitative		Width of the roadway. DOES NOT TAKE INTO ACCOUNT: TPC, EMERGENCY LANE, PARKING SPACE				Drop	
14	surf	Road surface condition	float64	0,00%	Categorical - 6 to 10 categories	-1 - Not specified 1 - normal 2 - wet 3 - puddles 4 - flooded 5 - snow 6 - mud 7 - icy 8 - fat - oil 9 - other					Keep	
15	infra	Infrastructure	float64	0,00%	Categorical - more than 10 categories	-1 - Not specified 0 - None 1 - Underground - tunnel 2 - Bridge - autopont 3 - Exchanger or connection brace 4 - Railway 5 - Carrefour arranged 6 - Pedestrian area 7 - Toll zone 8 - Construction site 9 - Other					Keep	
16	situ	Accident situation	float64	0,00%	Categorical - 6 to 10 categories	-1 - Not specified 0 - None 1 - On the road 2 - On emergency stop band 3 - On the verge 4 - On the sidewalk 5 - On bike path 6 - On another special track/road 8 - Other					Keep	
17	vma (rename: speed_lim)	Speed limit	float64	0,00%	Quantitative	X	Started to be taken into account from 2019				Keep	
18	annee (rename: year)	Year	int64	0,00%	Quantitative						Keep	