Data description

After searching the internet for automobile crashes datasets I discovered that New Zealand Transport Agency has an open data repository, that includes *"Crash Analysis System (CAS) data"* (on a personal note I live in New Zealand, so this project could give me some interesting insights that I can relate to). This dataset is the automobile crash information I need. The dataset is available in several formats, I will use the CSV version, found here as disaggregated crash data.

The crash data provides information about each traffic crash reported by the New Zealand Police since 1 January 2000. The dataset is updated quarterly and the current set was updated in July 2020 and includes crash data up to the 29th of February. The data covers automobile crashes on all New Zealand roadways or places where the public have legal access with a motor vehicle. It is important to note that not all crashes are reported to the Police and the level of reporting increases with the severity of the crash. Due to the nature of non-fatal crashes, it is believed that these are under-reported.

The particular dataset I will be using for my modelling has a total of 725548 samples, with 72 features on each of them. The New Zealand Transport Agency provides a metafile or a webpage with a brief description of the features. A table presenting my compilation of the descriptions is included in Appendix A. The feature with the information regarding the crash severity is named "*crashSeverity*", it is a categorical feature that can take the values: 'Fatal crash', 'Serious crash', 'Minor crash', 'Non-injury crash'. This is determined by the worst injury sustained in the crash at the time of entry.

The features in the dataset are both categorical and numerical, and a few spatial geodata ones used to indicate the location of the crash site.

Several categorical features provide information about different attributes and conditions of the crash event. All these features have a small number of possible values. For example, the feature "*roadLane*" informs about the road configuration at the crash site. Its possible values are: '1-way', '2-way', 'Median' and 'Off road' which are self-explanatory. Another example, is the feature "*light*" which informs about the natural light conditions at the time and place of the crash. Its possible values are: 'Bright Sun', 'Overcast', 'Twilight, 'Dark' or 'Unknown', which again are self-explanatory.

The dataset has a large number of numerical features, which are in most cases counters providing details about a variety of objects that could be involved in the crash. For example, the feature *"parkedVehicle"* indicates how many times a parked or unattended vehicle was struck in the crash. The feature "debris" indicates how many times debris, boulders or items dropped or thrown from a vehicle(s) were struck in the crash. Additionally, there are other counters such as *"seriousInjuryCount"* which informs about the number of people seriously injured in the crash. A similar counter *"fatalCount"* provides the number of fatalities for a crash of severity '*Fatal crash*'. So, for example, a crash of severity *'Serious crash'* would have a *"fatalCount"* of zero and most probably a value greater than zero for *"seriousInjuryCount"*.

Another group of numerical features are values describing various speeds, such as "speedLimit" which provides the speed limit in force at the crash site at the time of the crash or "temporarySpeedLimit" which provides the temporary speed limit at the crash site if one exists (e.g. for road works). Besides these, some other numerical features provide information about the year of the crash, the number of lanes on the road of the crash and identifiers.

The next step is to explore the data and obtain insights that will determine if it is possible to reduce the number of samples and/or features and the placement of missing values and the possibility of replacing them.

APPENDIX
AUTOMOBILE CRASH DATA FEATURES DESCRIPTIONS

Feature name	Feature type	Feature description
X	numerical(geo)	Geographical longitude of the crash.
Y	numerical(geo)	Geographical latitude of the crash.
OBJECTID	numerical	Unique ID for the crash.
advisorySpeed	numerical	The advisory speed at the crash site at the time of the crash.
areaUnitId	numerical(geo)	The unique identifier of an area unit.
bicycle	numerical	Derived feature to indicate how many bicycles were involved in the crash.
bridge	numerical	Derived feature to indicate how many times a bridge, tunnel, the abutments, handrails were struck in the crash.
bus	numerical	Derived feature to indicate how many buses were involved in the crash (excluding school buses which are counted apart).
carStationWagon	numerical	Derived feature to indicate how many cars or station wagons were involved in the crash.
cliffBank	numerical	Derived feature to indicate how many times a cliff or bank was struck in the crash. This includes retaining walls
crashDirectionDescription	categorical(text)	The direction of the crash from the reference point. Values possible are 'North', 'East', 'South' or 'West'.
crashFinancialYear	numerical	The financial year in which a crash occurred, if known.
crashLocation1	text	Part 1 of the 'crash location'. May be a road name, route position (RP), landmark, or other, e.g. 'Ninety Mile Beach'. Used for location descriptions in reports etc.
crashLocation2	text	Part 2 of the 'crash location' (crash_locn). May be a side road name, landmark etc. Used for location descriptions in reports etc.
crashRoadSideRoad	categorical(text)	Indicates whether the principal vehicle in a crash was on the crash road or side road at the time of the crash. Note that 'on side road' can only happen if the crash occurred at an intersection. Possible values are 1: Crash Road, 2: Side Road
crashSeverity	categorical(text)	The severity of a crash. Possible values are 'Fatal crash', 'Serious crash, 'Minor crash', 'Non-injury crash'. This is determined by the worst injury sustained in the crash at time of entry.
crashSHDescription	categorical(text)	Indicates where a crash is reported to have occurred on a State Highway. Possible values include 'Yes' where the crash occurred on a SH, 'No' and 'Unknown'.
crashYear	numerical	Year of the crash (yyyy).
debris	numerical	Derived feature to indicate how many times debris, boulders or items dropped or thrown from a vehicle(s) were struck in the crash.
directionRoleDescription	categorical(text)	The direction (dirn) of the principal vehicle involved in the crash. Possible values are 'North', 'South', 'East' or 'West'.
ditch	numerical	Derived feature to indicate how many times a ditch or water able drainage channel was struck in a crash.
fatalCount	numerical	Number of fatalities for crash of severity 'Fatal crash'.
fence	numerical	Derived feature to indicate how many times a fence was struck in the crash. This includes letterbox(es), hoardings, private roadside furniture, hedges, sight rails, etc.
flatHill	categorical(text)	Whether the road is flat 'Flat' or sloped 'Hill road'.
guardRail	numerical	Derived feature to indicate how many times a guard or guard rail was struck in the crash. This includes 'New Jersey' barriers, 'ARMCO', sand filled barriers, wire catch fences, etc.
holiday	categorical(text)	Indicates whether the crash happened during a holiday and which one. Possible values: ‘None’, ‘Christmas/New Year’, ‘Easter’, ‘Queens Birthday’, ‘Labour Weekend’
houseOrBuilding	numerical	Derived feature to indicate how many times a houses, garages, sheds or other buildings were struck in the crash.
intersection	categorical(text)	Indicate if a crash happened at an 'Intersection', 'At Landmark' or 'Unknown'.
kerb	numerical	Derived feature to indicate how many times a kerb was struck in the crash, that contributed directly to the crash.
light	categorical(text)	The light at the time and place of the crash. Possible values: 'Bright Sun', 'Overcast', 'Twilight, 'Dark' or ' Unknown'.
meshblockId	numerical	UniqueID number for the meshblock.
minorInjuryCount	numerical	Number of minor injured people.
moped	numerical	Derived feature to indicate how many mopeds were involved in the crash.
motorcycle	numerical	Derived feature to indicate how many motorcycles were involved in the crash.

NumberOfLanes	numerical	The number of lanes on the crash road.
objectThrownOrDropped	numerical	Derived feature to indicate how many times objects were thrown at or dropped on vehicles in the crash.
otherObject	numerical	Derived feature to indicate how many times an object was struck in a crash and the object struck was not pre-defined. This feature includes stockpiled materials, rubbish bins, fallen poles, fallen trees, etc.
otherVehicleType	numerical	Derived feature to indicate how many other vehicles (not included in any other category) were involved in the crash.
overBank	numerical	Derived feature to indicate how many times an embankment was struck or driven over during a crash. This feature includes other vertical drops driven over during a crash.
parkedVehicle	numerical	Derived feature to indicate how many times a parked or unattended vehicle was struck in the crash. This feature can include trailers.
pedestrian	numerical	Derived feature to indicate how many pedestrians were involved in the crash. This includes pedestrians on skateboards, scooters and wheelchairs.
phoneBoxEtc	numerical	Derived feature to indicate how many times a telephone kiosk traffic signal controllers, bus shelters or other public furniture was struck in the crash.
postOrPole	numerical	Derived feature to indicate how many times a post or pole was struck in the crash. This includes light, power, phone, utility poles and objects practically forming part of a pole (i.e. 'Transformer Guy' wires).
region	categorical(text)	Identifies the local government (LG) region. The boundaries match territorial local authority (TLA) boundaries.Possible values are: ‘Auckland’, ‘Waikato’, ‘Canterbury’, ‘Wellington’, ‘Bay of Plenty’, ‘Otago’, ‘Manawatu/Wanganui’, ‘Northland’, ‘Hawkes Bay’, ‘Nelson/Marlborough’, ‘Southland’, ‘Taranaki’, ‘Gisborne’, ‘West Coast’.
roadCharacter	categorical(text)	The general nature of the road. Possible values include 'Bridge', 'Motorway Ramp', ‘Rail xing’, Overpass', 'Speed hump', 'Underpass', 'Tunnel', 'Tram lines' or 'Nil'.
roadLane	categorical(text)	The lane configuration of the road. Possible values : '1-way', '2-way, 'Median' and 'Off road'.
roadSurface	categorical(text)	The road surface description applying at the crash site. Possible values: 'Sealed' or 'Unsealed'.
roadworks	numerical	Derived feature to indicate how many times an object associated with roadworks (including signs, cones, drums, barriers, but not roadwork vehicles) was struck during the crash.
schoolBus	numerical	Derived feature to indicate how many school buses were involved in the crash.
seriousInjuryCount	numerical	Number of seriously injured people.
slipOrFlood	numerical	Derived feature to indicate how many times landslips, washouts or floods (excluding rivers) were objects struck in the crash.
speedLimit	numerical	The speed limit in force at the crash site at the time of the crash.
strayAnimal	numerical	Derived feature to indicate how many times a stray animal(s) was struck in the crash. This feature includes wild animals such as pigs, goats, deer, straying farm animals, house pets and birds.
streetLight	categorical(text)	The street lighting at the time of the crash. Possible values 'On', 'Off', 'None' or ' Unknown'.
suv	numerical	Derived feature to indicate how many SUVs were involved in the crash.
taxi	numerical	Derived feature to indicate how many taxis were involved in the crash.
temporarySpeedLimit	numerical	The temporary speed limit at the crash site if one exists (e.g. for road works).
tlaId	numerical	The unique identifier for a territorial local authority (TLA). Each crash is assigned a TLA based on where the crash occurred.
tlaName	categorical(text)	Indicates the local authority who handled the crash. Each region has several TLA.
trafficControl	categorical(text)	The traffic control signals at the crash site. Possible values are 'Traffic Signals', 'Stop Sign', 'Give Way Sign', 'Pointsman', 'School Patrol', 'Nil' or ' Unknown'.
trafficIsland	numerical	Derived feature to indicate how many times a traffic island, medians (excluding barriers) was struck in the crash.
trafficSign	numerical	Derived feature to indicate how many times traffic signage (including traffic signals, their poles, bollards or roadside delineators) was struck in the crash.
train	numerical	Derived feature to indicate how many times a train, rolling stock or jiggers was struck in the crash, whether stationary or moving.
tree	numerical	Derived feature to indicate how many times trees or other growing items were struck during the crash.
truck	numerical	Derived feature to indicate how many trucks were involved in the crash.
unknownVehicleType	numerical	Derived feature to indicate how many vehicles were involved in the crash (where the vehicle type is unknown).

urban	categorical(text)	Derived feature using the 'spd_lim' variable. Possible values are 'Urban' (urban, spd_lim < 80) or 'Open Road' (open road, spd_lim >=80 or 'LSZ')
vanOrUtility	numerical	Derived feature to indicate how many vans or utes were involved in the crash.
vehicle	numerical	Derived feature to indicate how many times a stationary attended vehicle was struck in the crash. This includes broken down vehicles, workmen's vehicles, taxis, buses.
waterRiver	numerical	Derived feature to indicate how many times a body of water (including rivers, streams, lakes, the sea, tidal flats, canals, watercourses or swaps) was struck in the crash.
weatherA	categorical(text)	Indicates weather at the crash time/place. A derived variable using the 'spd_lim' variable. Values that are possible are 'Fine', 'Mist', 'Light Rain', 'Heavy Rain', 'Snow', 'Unknown'.
weatherB	categorical(text)	Second field to detail for the weather at the crash time/place when applicable. Values 'Frost', 'Strong Wind' or 'Unknown'.