Exploratory Data Analysis

Project4: Estimation of average claim cost using Gamma and Tweedie distribution.

Name: Jitesh Rawat

Reg.no - 20237

Introduction.

The data provided is Crash Analysis System(CAS) data.This data comes from the Waka Kotahi Crash Analysis System (CAS), which records all traffic crashes reported to us by the NZ Police. CAS covers crashes on all New Zealand roadways or places where the public have legal access with a motor vehicle.

The data has **758757 rows and 72 columns.**

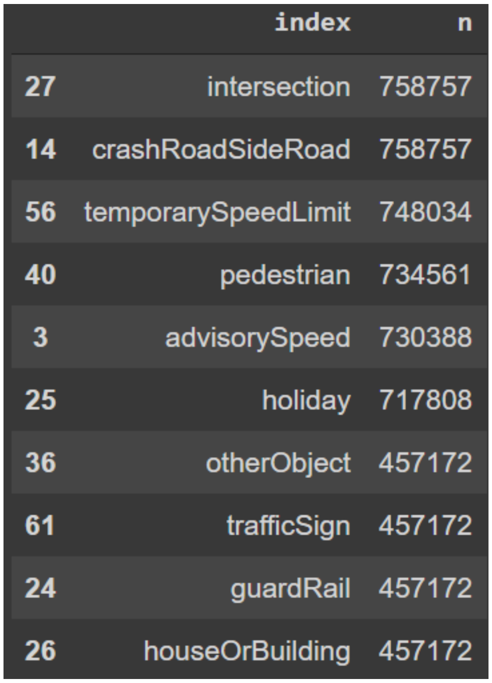
Columns provided in dataset are :

'OBJECTID', 'advisorySpeed', 'areaUnitID', 'bicycle', 'bridge', 'bus', 'carStationWagon', 'cliffBank', 'crashDirectionDescription', 'crashFinancialYear', 'crashLocation1', 'crashLocation2', 'crashRoadSideRoad', 'crashSeverity', 'crashSHDescription', 'crashYear', 'debris', 'directionRoleDescription', 'ditch', 'fatalCount', 'fence', 'flatHill', 'guardRail', 'holiday', 'houseOrBuilding', 'intersection', 'kerb', 'light', 'meshblockId', 'minorInjuryCount', 'moped', 'motorcycle', 'NumberOfLanes', 'objectThrownOrDropped', 'otherObject', 'otherVehicleType', 'overBank', 'parkedVehicle', 'pedestrian', 'phoneBoxEtc', 'postOrPole', 'region', 'roadCharacter', 'roadLane', 'roadSurface', 'roadworks', 'schoolBus', 'seriousInjuryCount', 'slipOrFlood', 'speedLimit', 'strayAnimal', 'streetLight', 'suv', 'taxi', 'temporarySpeedLimit', 'tlaId', 'tlaName', 'trafficControl', 'trafficIsland', 'trafficSign', 'train', 'tree', 'truck', 'unknownVehicleType', 'urban', 'vanOrUtility', 'vehicle', 'waterRiver', 'weatherA', 'weatherB'.

Exploratory Data Analysis

Since there are 72 columns we will check some of them individually and some of them in a combine plot. But first we will check the information of the data whether all entries are available or there are any null values.

For that we will create a table that holds the sum number of null values in each of the column then we will sort it and find the columns that has max number of null values and how many such columns exist.



	index	n
27	intersection	758757
14	crashRoadSideRoad	758757
56	temporarySpeedLimit	748034
40	pedestrian	734561
3	advisorySpeed	730388
25	holiday	717808
36	otherObject	457172
61	trafficSign	457172
24	guardRail	457172
26	houseOrBuilding	457172

By looking at the table we can see the first 2 columns have null values in all the entries as the number of rows matches the number of null values in 'intersection' and 'crashRoadSideRoad'. Other than that we have next 4 columns which are having null values atleast more than 7 lakhs.

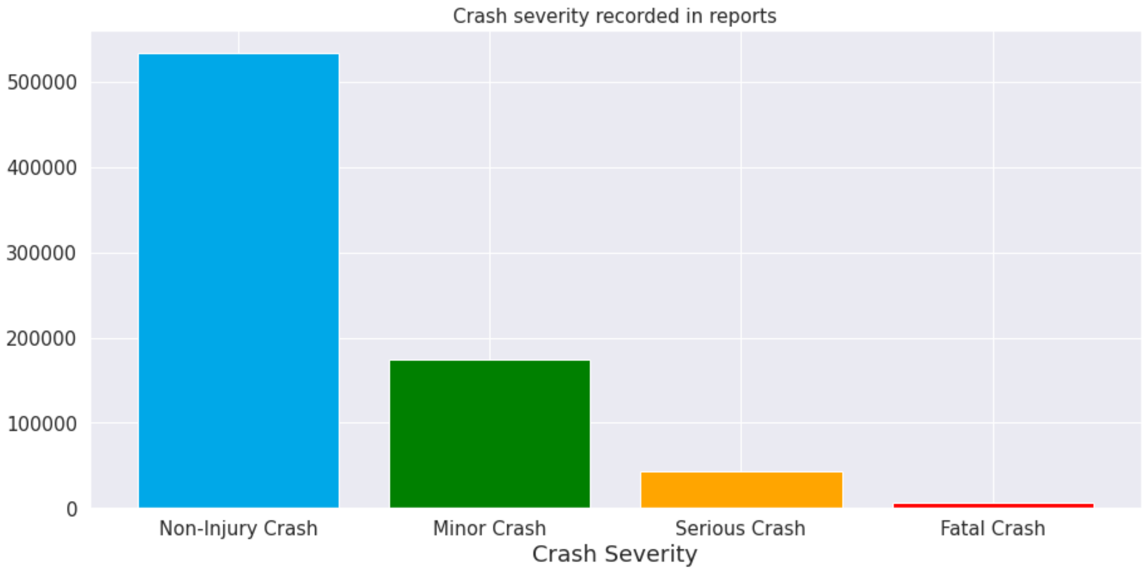
Plotting similar categories of variables.

Since there are 72 columns it would be difficult to understand what each one is trying to depict if we plot all of them together. So we would try to find similar types of columns which provide similar informations and plot them.

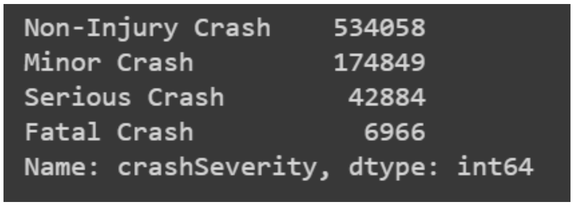
1. Crash Severity, Injury and fatal counts.

In this section we would see histograms of columns that are related to crash severity or intensity of injury.

A) Crash severity



From above plot we observe that most of the times crash type is non-injury crash following with minor and severe ones. We see very few fatal crashes were found in crash report lets look at their numbers to get more clarity.

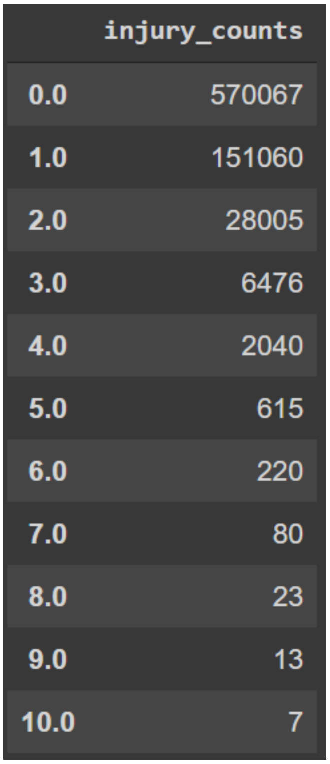


```
Non-Injury Crash     534058
Minor Crash          174849
Serious Crash         42884
Fatal Crash            6966
Name: crashSeverity, dtype: int64
```

B) Minor injury count

Other than crash severity we also have variables like 'minorInjuryCount' and 'fatalCount' with which we can see and verify the injury counts.

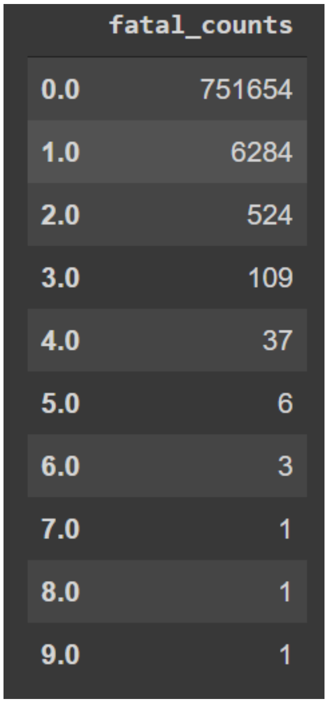
The table below shows one important fact that around 5,70,000 crash reports had 0 minor injuries. And there were atleast 1.5 lakh crash reports were 1 person had minor injury due to a crash.



	injury_counts
0.0	570067
1.0	151060
2.0	28005
3.0	6476
4.0	2040
5.0	615
6.0	220
7.0	80
8.0	23
9.0	13
10.0	7

C) Fatal counts

This table verifies the number of fatat counts and we can infer that positively most of the crashes(around 7.5 lakh) had 0 number of fatal casualties related to crash. And the number drops drastically where 6284 crash reports having 1 person with fatal casualty in a crash.

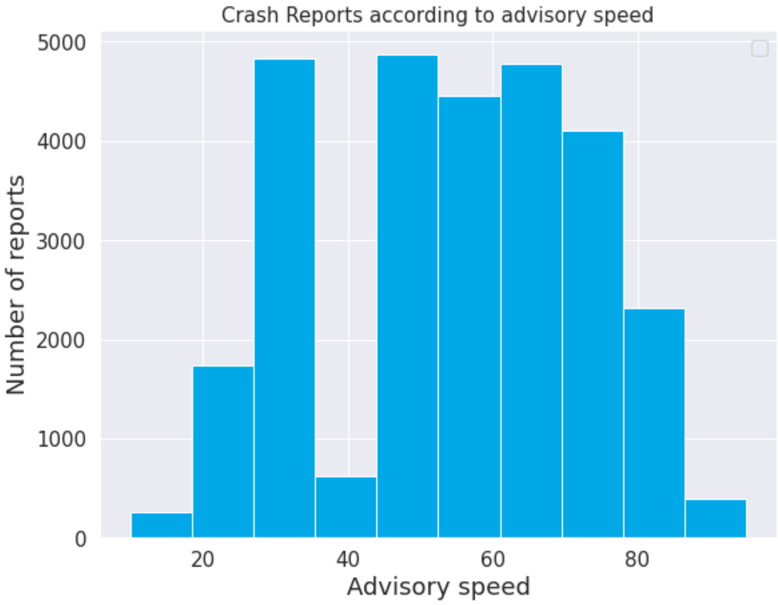


	fatal_counts
0.0	751654
1.0	6284
2.0	524
3.0	109
4.0	37
5.0	6
6.0	3
7.0	1
8.0	1
9.0	1

2. Speedlimit and advisory speed.

A) Advisory Speed

The advisory speed has most of null values and only around 5000 legit entries.

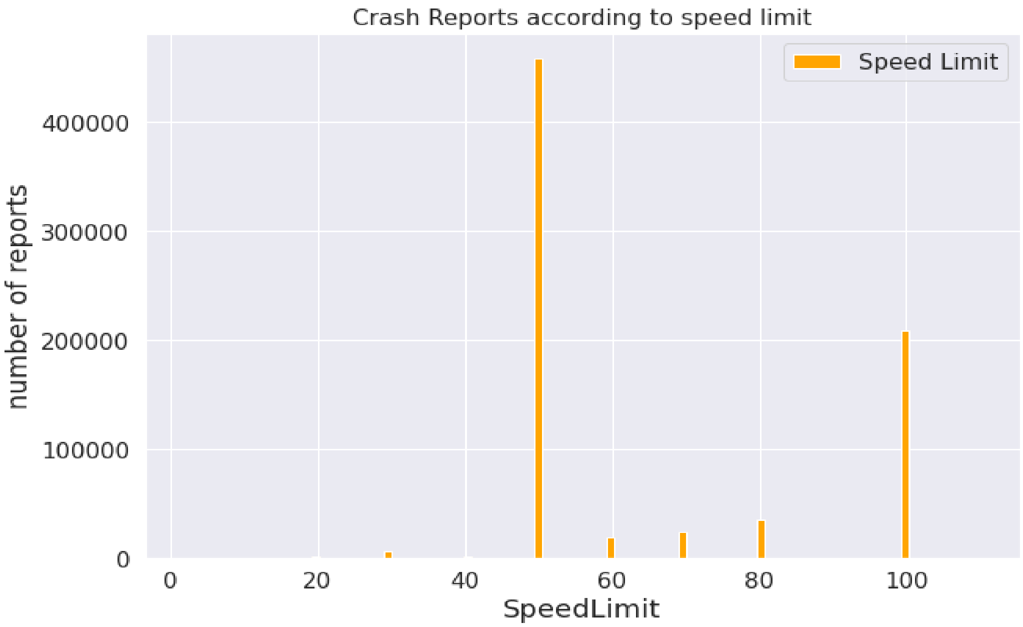


Crash Reports according to advisory speed

It is wierd to see that the bar is high where at 30 and 50, 60 ,70 and then it reduces gradually in 80, 90. But if we observe the data is about advisory speeds and crashes are less prone where advisory speed is 40. But a hike at the bar at advisory speed 30 is an anomaly. And about advisory speed 70,80, 90 we can understand that these are not advisory speed generally unless the road is highway or its empty.

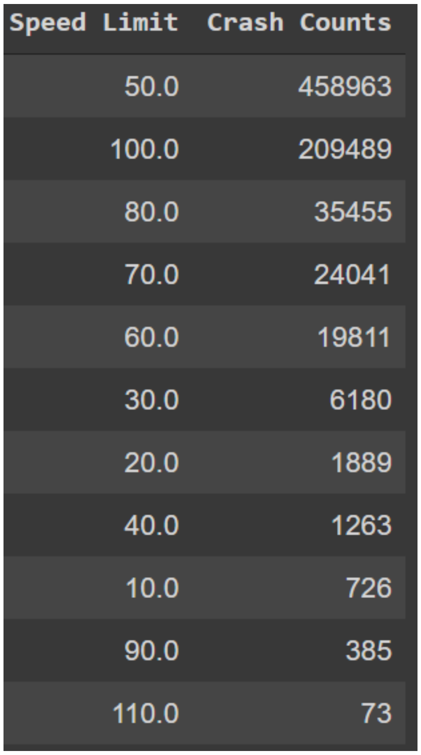
B) Speed limit.

Speed limit is the speed limit recorded at time of crash. It can be an important factor if we were to determine crash.



From the plot we can infer very high number of crash reports when speed limit is 50 i.e around 4.5 lakhs. And surprisingly crashes at speed 60, 70, 80 are very less as compared to crashes where speed limit was 50. But crash reports which recorded 100 as speed limit at crash site is also around 2 lakhs.

We can also look at exact numbers to get better idea.



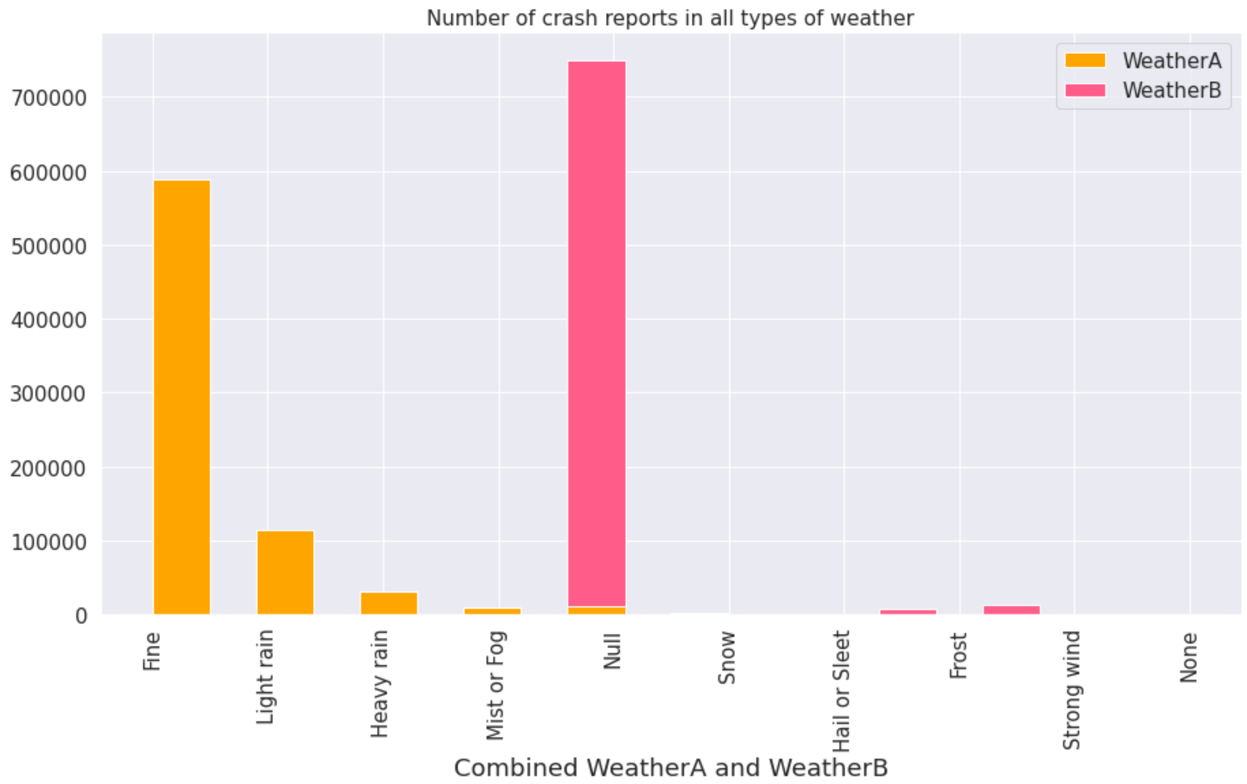
Speed Limit	Crash Counts
50.0	458963
100.0	209489
80.0	35455
70.0	24041
60.0	19811
30.0	6180
20.0	1889
40.0	1263
10.0	726
90.0	385
110.0	73

So we get the idea that there are more violation of speedlimit at 50 which leads to max number of crashes.

3. Weather , lights and directions.

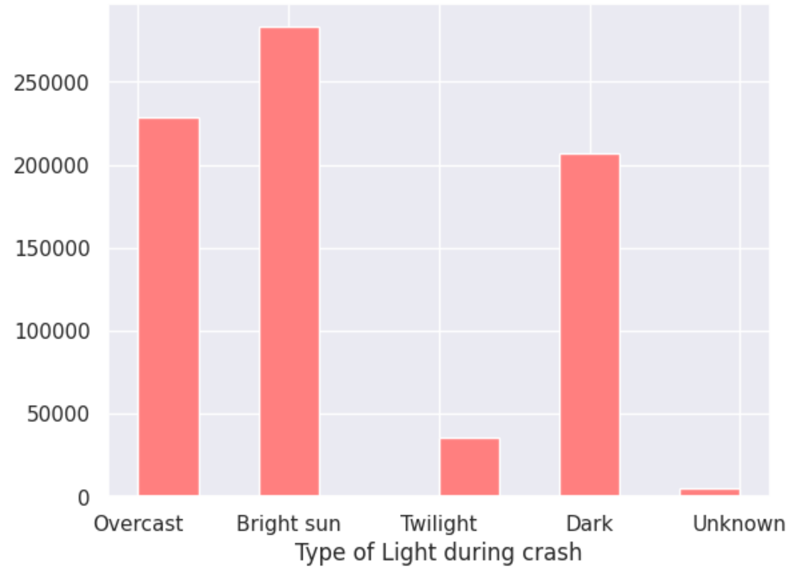
A) Weather

There are 2 variables WeatherA and WeatherB that indicate the weather at the crash time. The below plot shows combine values of both variables.



The list of weather contains Fine, Light rain, Heavy rain, Mist or Fog, Snow, Hail or Sleet, Frost and Strong wind. Most of the information is provided by WeatherA variable where we can see that almost 6 lakh crashes had Fine weather during crash. Also around 1.2 lakh had light rain. For WeatherB variable we can see that most of them are null values and few thousand entries for strong wind and frost.

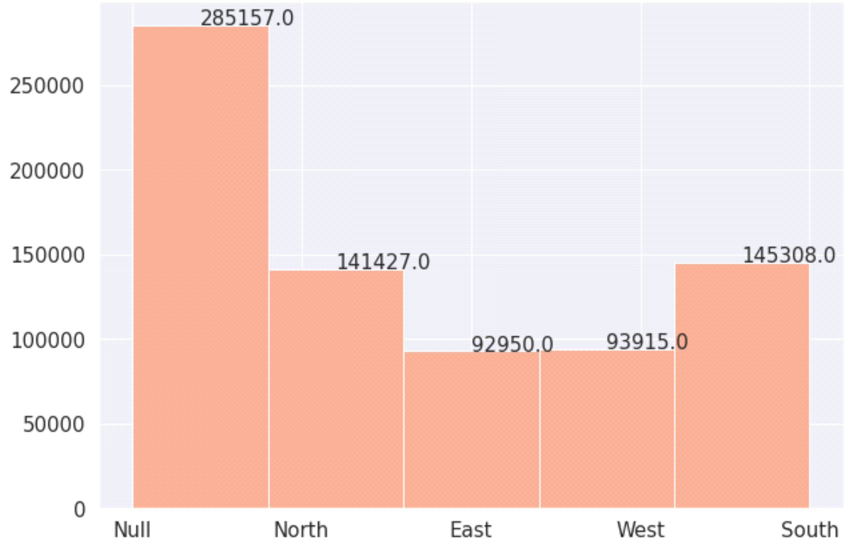
B) Light



The light column shows the type of light present on the crash site. The possible values provided are Overcast, Bright sun, Twilight, Dark and Unknown. The Bright sun light seems to have most number of crashes with more than 2.5 lakh crashes which is slightly more than Overcast type. Very less number of crashes had Twilight type of light.

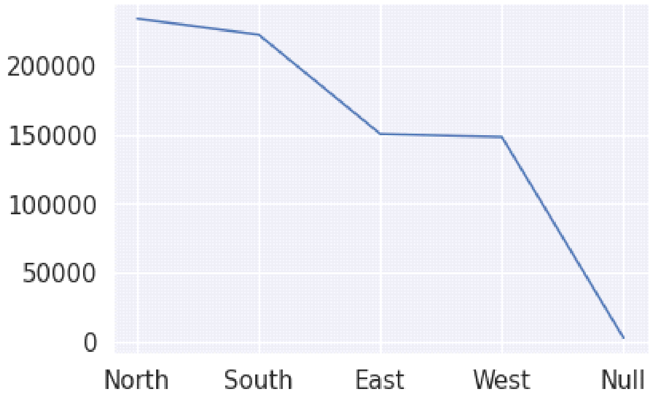
C) Crash direction description

This variable has the direction (dirn) of the crash from the reference point. Values possible are 'North', 'East', 'South' or 'West'.



North and south have almot equal number of around 1.4 lakh crashes, and its same case with east and west with below 95 thousand crash reports.

D) Direction role description



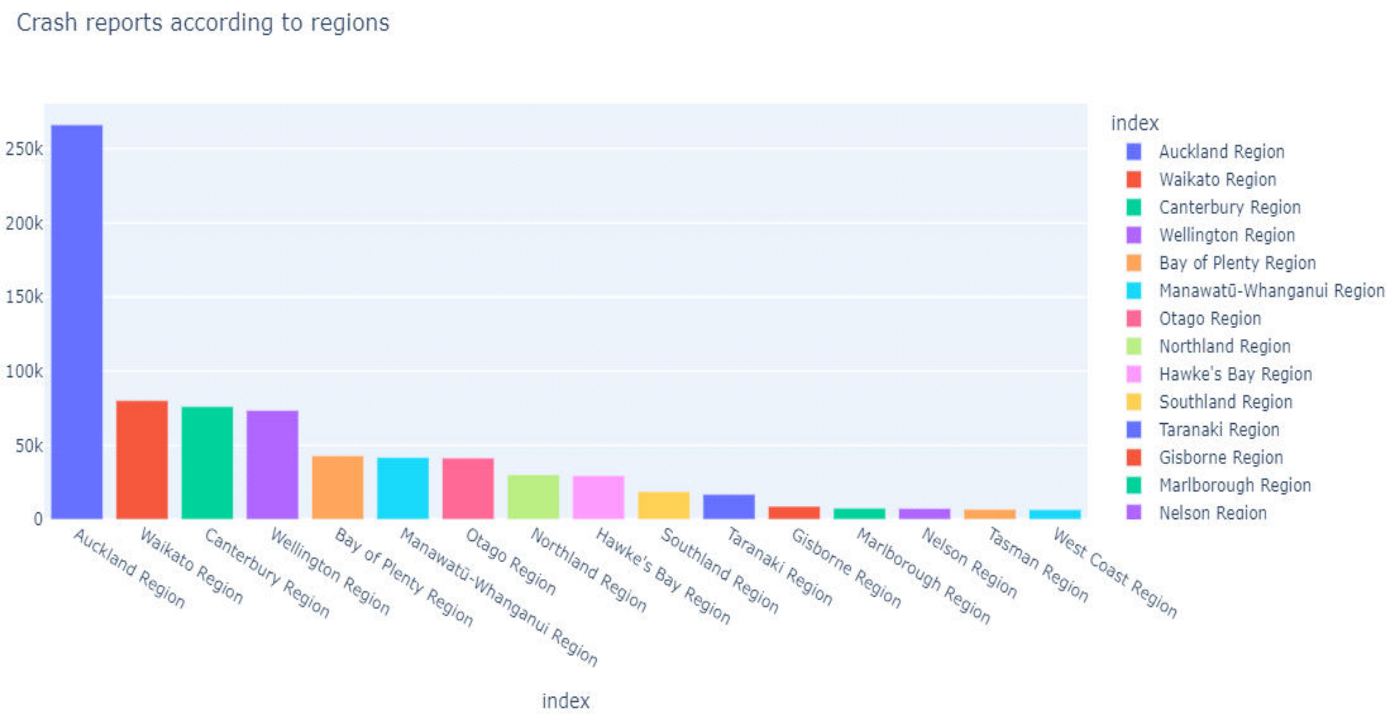
Principle vehicles directed towards north directions seem to have most number of crashes followed by South. East and West directed vehicles have least and almost equal number of crashes.

4.Region

In this section we will plot variables that are related to region or locations.

A) Region

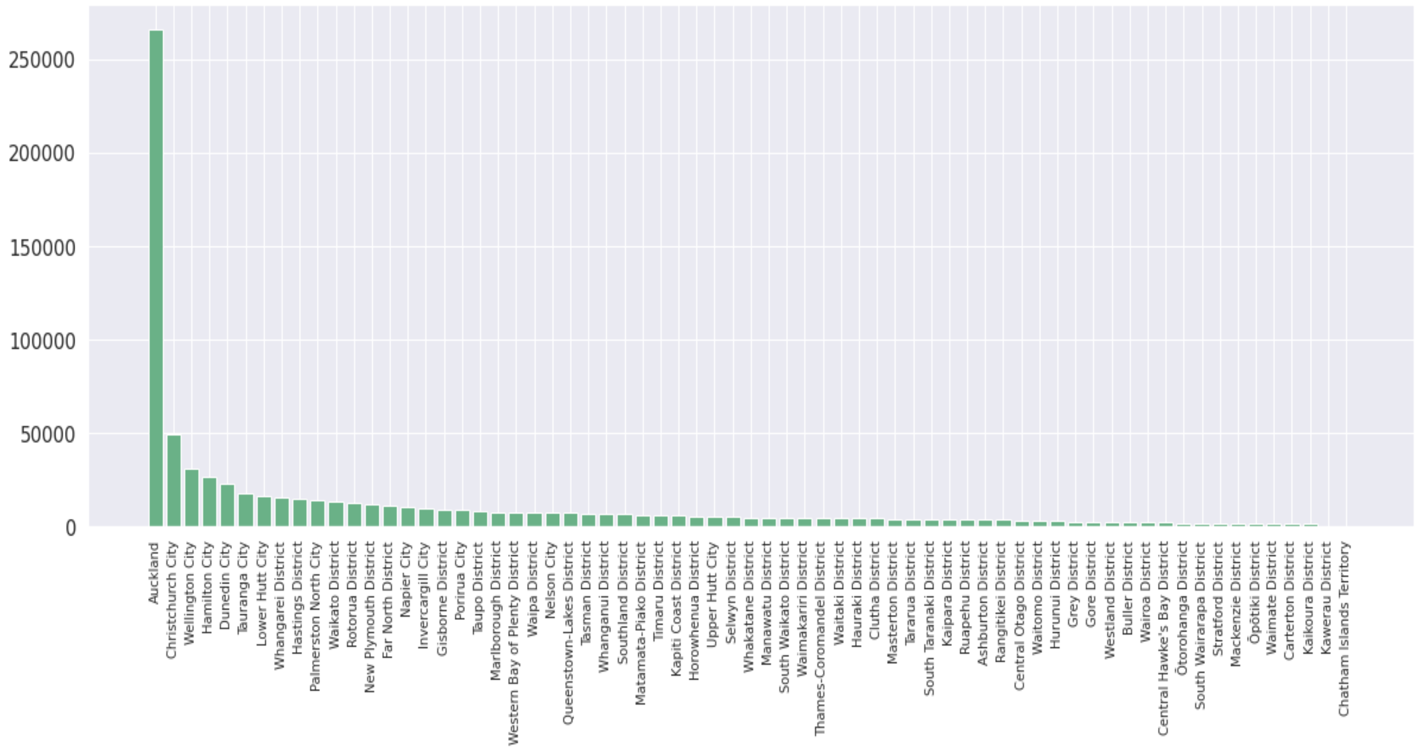
Region variable gives info about number of crashes over the local government regions.



From above plot we see that Auckland region has max number with totally over 250k-above crashes. And then Walkato, Canterbury and Wellington regions with above 50k whereas rest all regions have below 50k crash records.

B) TLA names

The 'tlaName' variable has territorial local authority (TLA) names recorded for those crash reports.

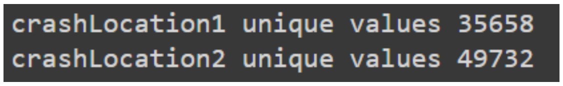


Seems like TLA names variable is not much different from the Region variable as the name of places seem to match other than some extra regions and here also we observe same inference that Auckland has drastically high number of crashes than rest of the regions.

C) Crash Locations.

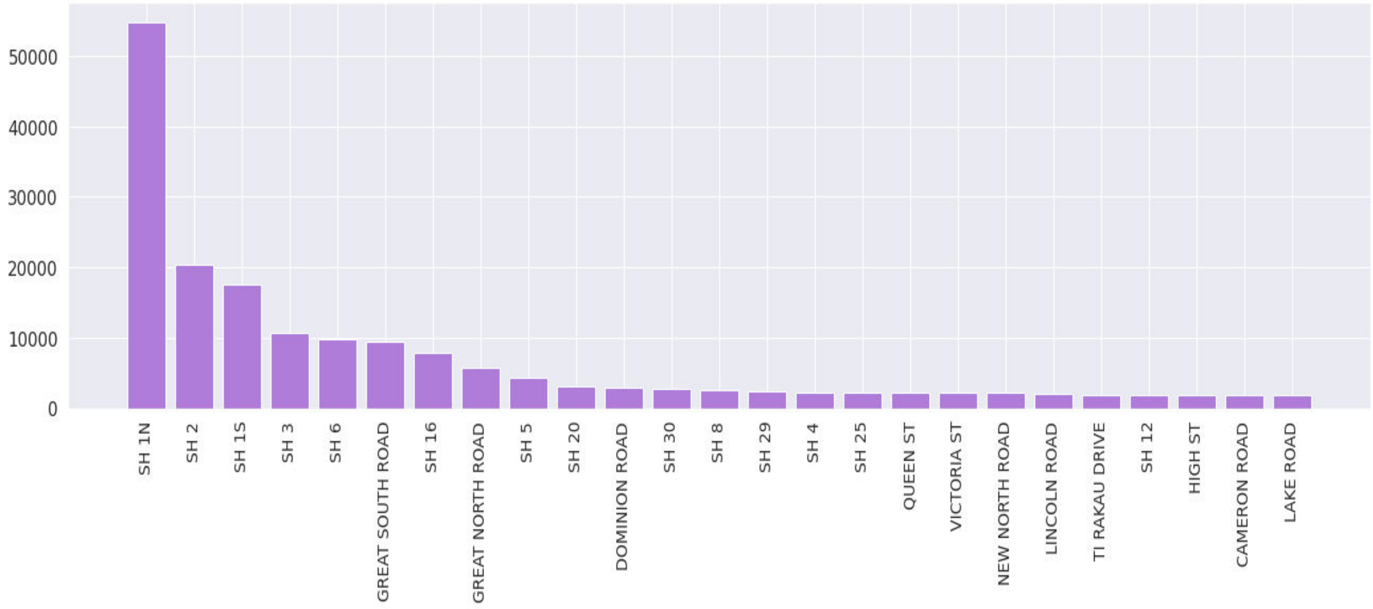
For crash locations we have 2 variables 'crashLocation1' and 'crashLocation2'. Crash locations consists of either road name, route position (RP), landmark, or other used for location descriptions in reports.

For crashLocation1 and crashLocation2 the location values are :


```
crashLocation1 unique values 35658
crashLocation2 unique values 49732
```

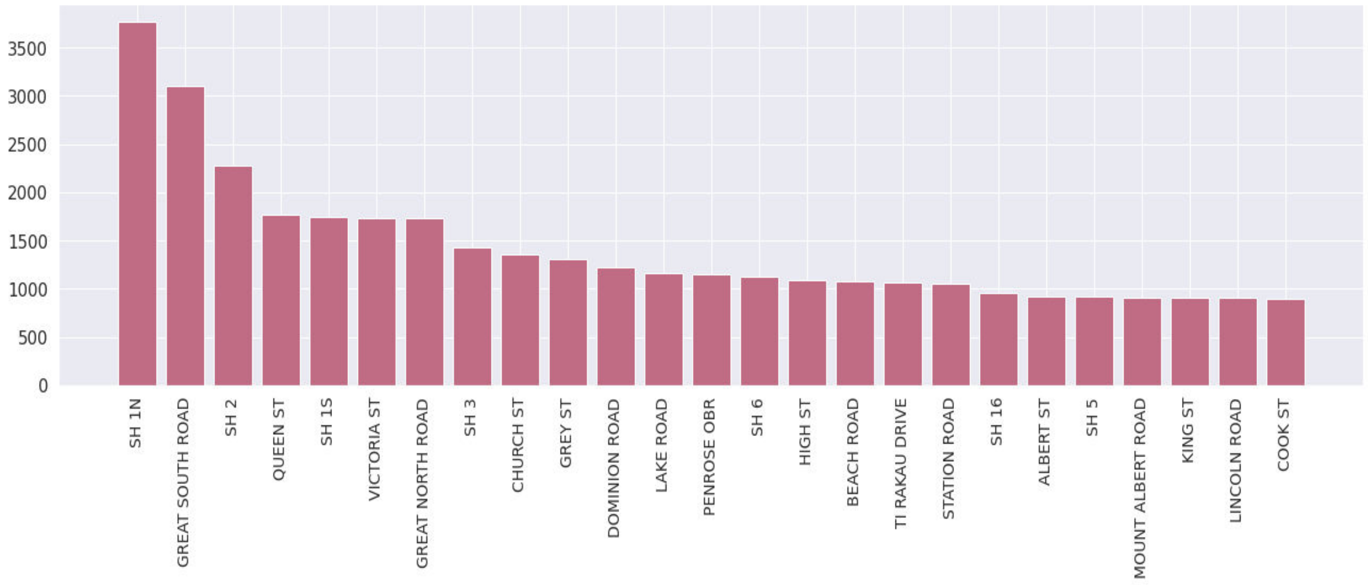
It would be difficult to plot all values and infer something from it so we will sort the values and plot initial few values for both variables

-crashLocation1



In crash location1 SH1N have some 55k number of crashes followed by SH2, SH1S. SH here means State Highway.Remaining locations seem to have lower than 10k crashes.

-crashLocation2



In crashLocation2 as well we can see that SH1 has max number of accidents. But a point to note here is that max number of crashes is below 4k in this variable whereas in crashLocation1 it was above 55k. From this we can infer that most of the crashes are almost evenly distributed around the locations in this variable.

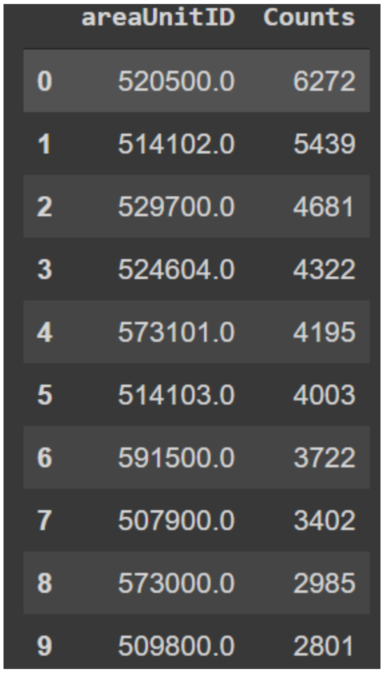
D) **areaUnitID**

This columns have crashes according to area unit Ids which is unique identifier of an area unit.

This column has 1874 unique values and there is no point in plotting crash counts for 1874 values and get inference.

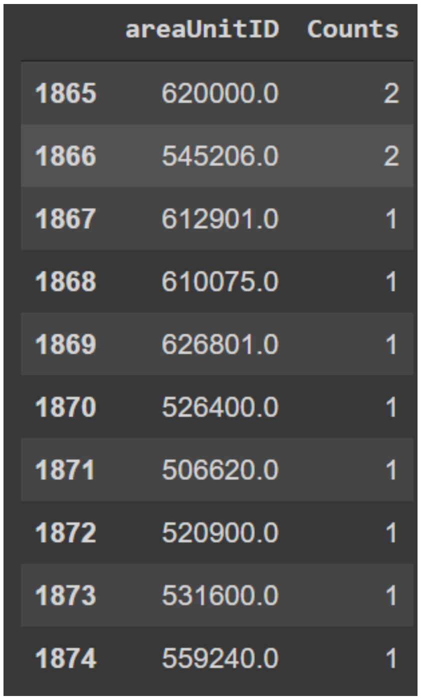
Instead we will sort these values with most crash counts and show first and last 10 areaUnitIDs with crash counts.

-First 10 areaUnitIDs show that an area unit with most crashes had 6272 crashes.



	areaUnitID	Counts
0	520500.0	6272
1	514102.0	5439
2	529700.0	4681
3	524604.0	4322
4	573101.0	4195
5	514103.0	4003
6	591500.0	3722
7	507900.0	3402
8	573000.0	2985
9	509800.0	2801

-last 10 areaUnitIDs. We can see that 8 area units had only 1 crash in them!

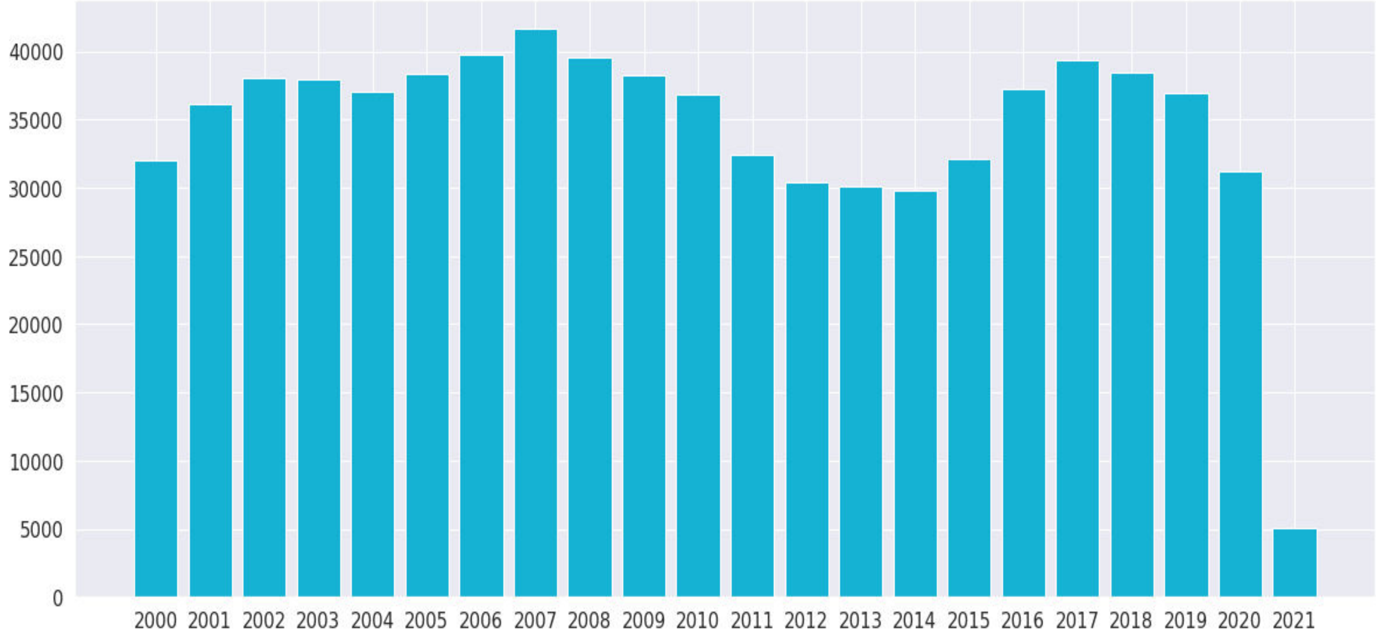


	areaUnitID	Counts
1865	620000.0	2
1866	545206.0	2
1867	612901.0	1
1868	610075.0	1
1869	626801.0	1
1870	526400.0	1
1871	506620.0	1
1872	520900.0	1
1873	531600.0	1
1874	559240.0	1

5. Crash reports according to years.

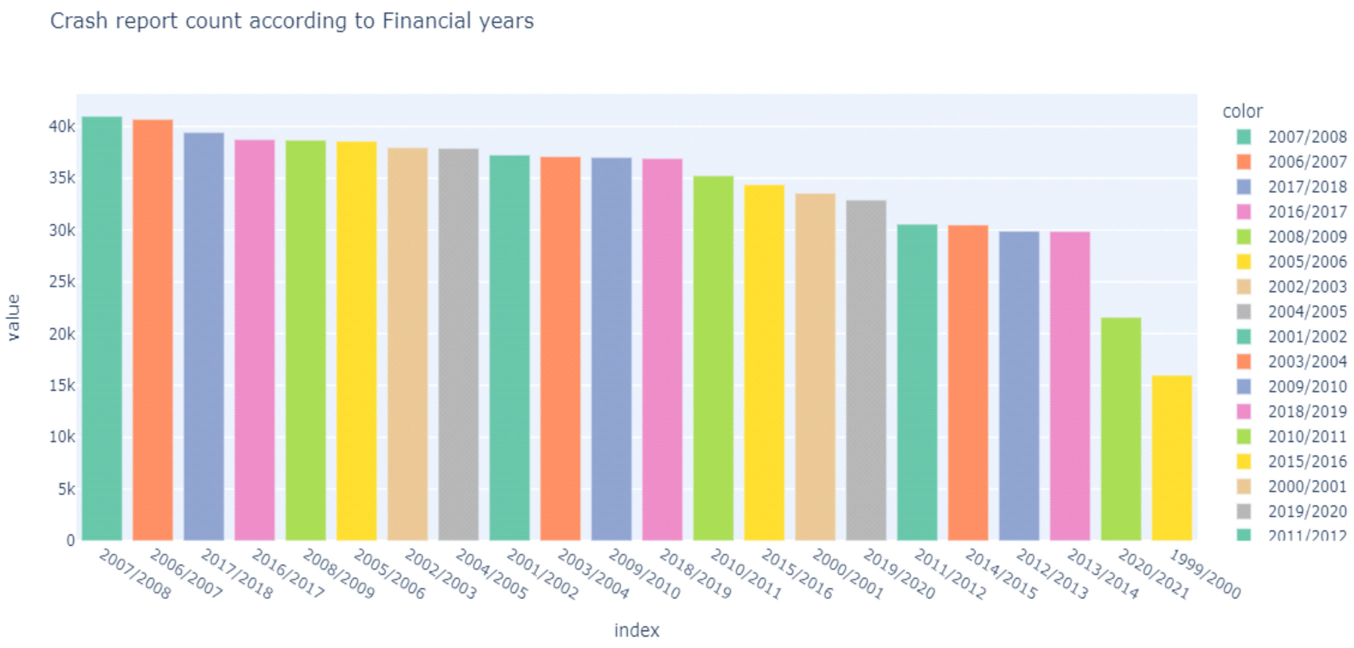
2 variables that give information yearwise crashYear and crashFinancialYear.

-crashYear



In crashYear we observe that there are atleast > 25k motor vehicle accidents per year except 2021 which is ongoing. And 2007 seem to have most number of crashes in that year followed by 2017 with most number of crashes. The legit least amount of crash reports can be seen for year 2014 with slightly around 25k crashes.

-crashFinancialYear



crashFinancialYear has recorded 2007/2008 to be the year with most number of crashes. The least amount we see here from 1999/2000 can be because the financial year starts from April 1st and ends on 31st March.

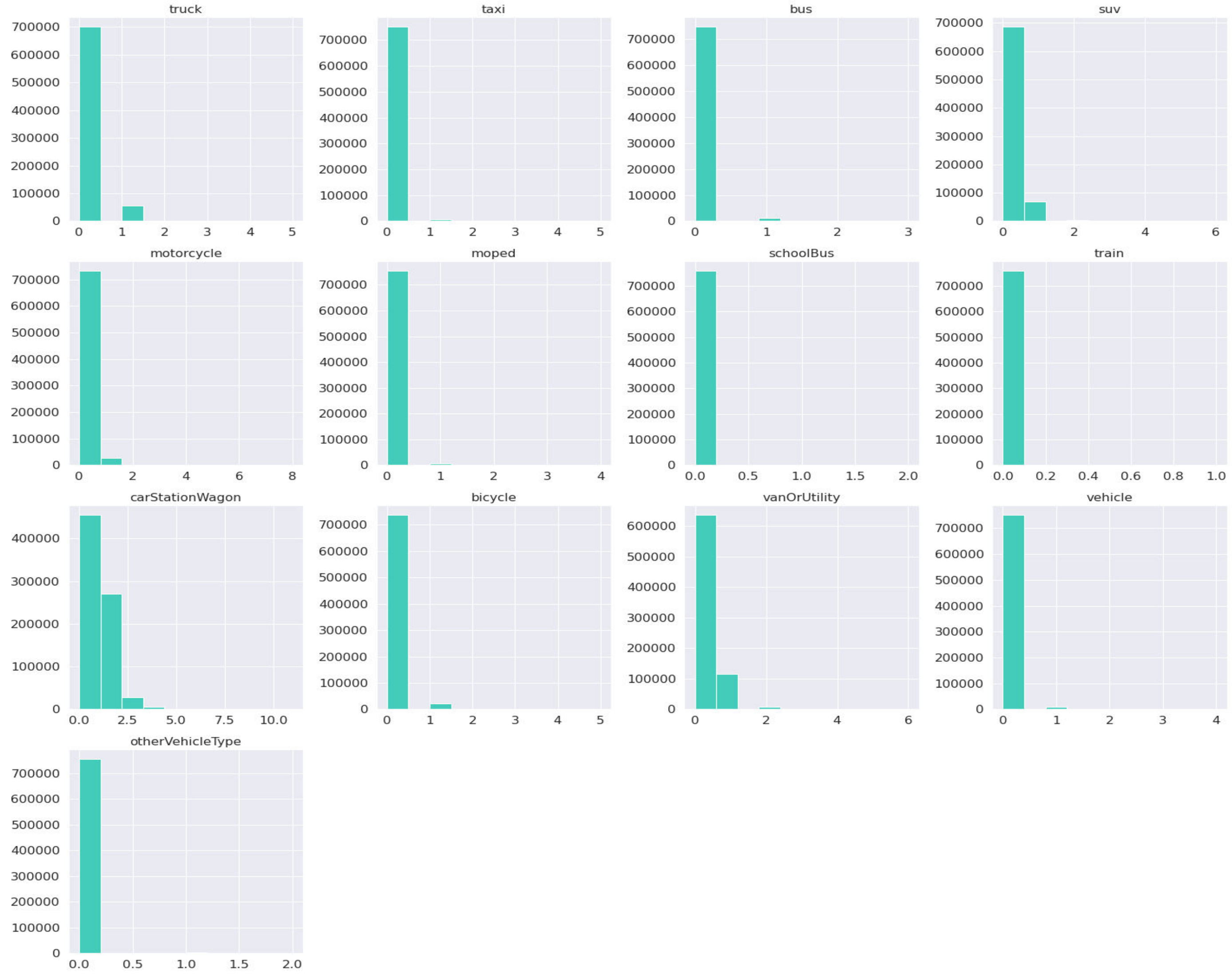
6. Vehicle types.

This section includes all the vehicle type columns in which either 0 , 1 or so units of that vehicle types were present in a crash report.

The plots include columns -

truck, taxi, bus, suv, motorcycles, moped, schoolbus, train, car stationwagon, bicycle, van or utility, vehicle(stationary), other vehicletypes.

These columns have number of each type invovled in a crash.



Most of the times we see that highest value is 0 in all plots.If we were to take an inference from these plots, we can observe that the most vehicletypes that have atleast 1 of them involved in crashes are carStationWagons with above 2.5 lakh crashes followed by vanOrUtility types with around 1lakh crashes and then suvs and trucks.

7. Objects that were struck during crashes.

These plots depict variables that have certain objects and their numbers that were struck during a crash.

variables are :

'**bridge**' - Derived variable to indicate how many times a bridge, tunnel, the abutments, handrails were struck in the crash.

'cliffBank' - Derived variable to indicate how many times a 'cliff' or 'bank' was struck in the crash. This includes retaining walls

'**debris**' - Derived variable to indicate how many times debris, boulders or items dropped or thrown from a vehicle(s) were struck in the crash

'**ditch**' - Derived variable to indicate how many times a 'ditch' or 'waterable drainage channel' was struck in a crash.

'**fence**' - Derived variable to indicate how many times a 'fence' was struck in the crash. This includes letterbox(es), hoardings, private roadside furniture, hedges, sight rails, etc.

'**guardRail**' - Derived variable 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.

'**houseOrBuilding**' - Derived variable to indicate how many times a houses, garages, sheds or other buildings(Bldg) were struck in the crash

'**kerb**' - Derived variable to indicate how many times a kerb was struck in the crash, that contributed directly to the crash.

'**objectThrownOrDropped**' - Derived variable to indicate how many times objects were thrown at or dropped on vehicles in the crash.

'**otherObject**' - Derived variable to indicate how many times an object was struck in a crash and the object struck was not pre-defined. This variable includes stockpiled materials, rubbish bins, fallen poles, fallen trees, etc.

'**overBank' -** Derived variable to indicate how many times an embankment was struck or driven over during a crash. This variable includes other vertical drops driven over during a crash.

'parkedVehicle' - Derived variable to indicate how many times a parked or unattended vehicle was struck in the crash. This variable can include trailers.

'pedestrian' - Derived variable to indicate how many pedestrians were involved in the crash. This includes pedestrians on skateboards, scooters and wheelchairs.

'phoneBoxEtc' - Derived variable to indicate how many times a telephone kiosk traffic signal controllers, bus shelters or other public furniture was struck in the crash

'postOrPole' - Derived variable 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)

'roadworks' - Derived variable to indicate how many times an object associated with 'roadworks' (including signs, cones, drums, barriers, but not roadwork vehicles) was struck during the crash

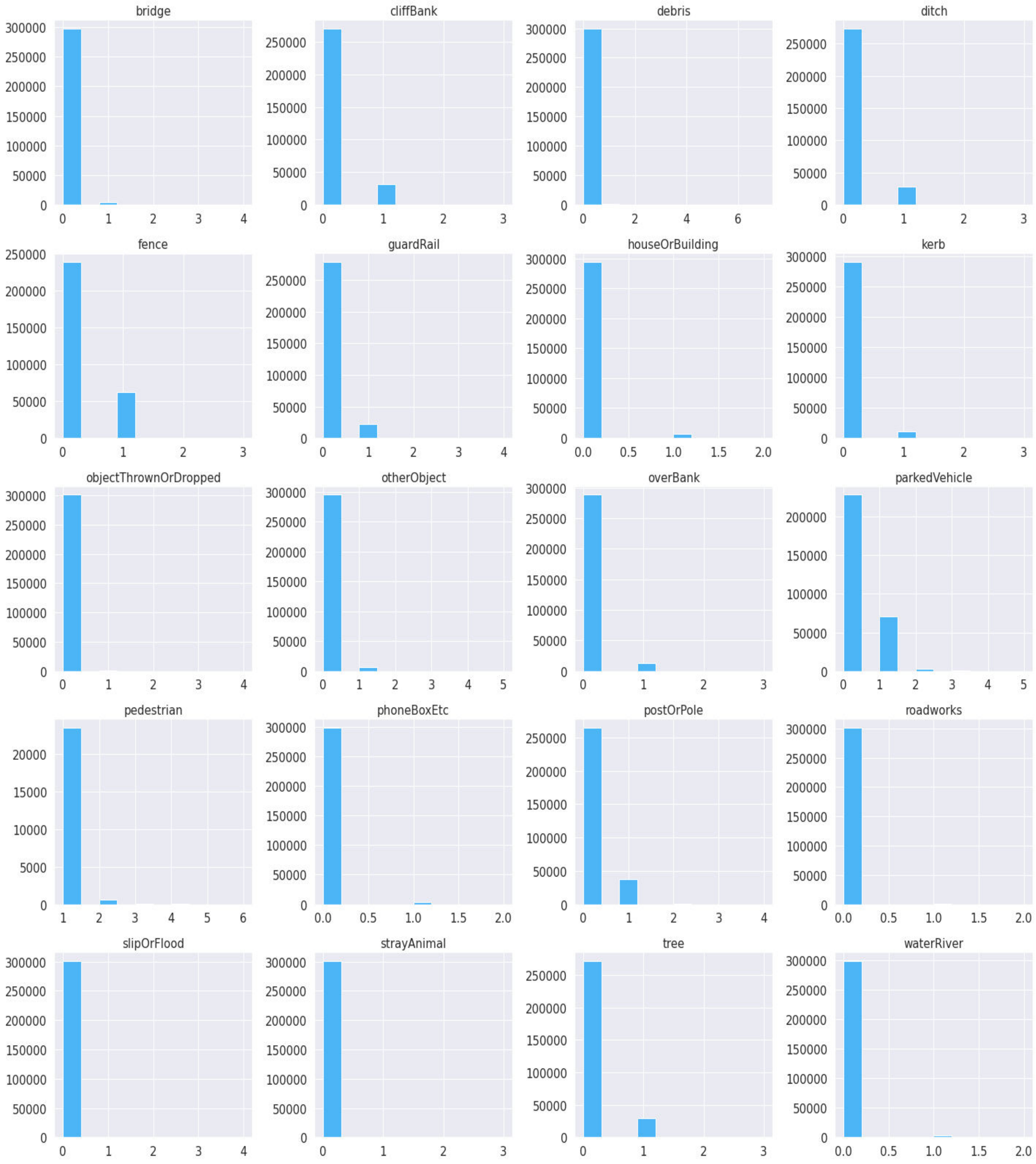
'slipOrFlood' - Derived variable to indicate how many times landslips, washouts or floods (excluding rivers) were objects struck in the crash

'strayAnimal' - Derived variable to indicate how many times a stray animal(s) was struck in the crash. This variable includes wild animals such as pigs, goats, deer, straying farm animals, house pets and birds.

'streetLight' - The street lighting at the time of the crash. Possible values 'On', 'Off', 'None' or ' Unknown'.

'tree' - Derived variable to indicate how many times trees or other growing items were struck during the crash.

'waterRiver' - Derived variable to indicate how many times a body of water (including rivers, streams, lakes, the sea, tidal flates, canals, watercourses or swanps) was struck in the crash.

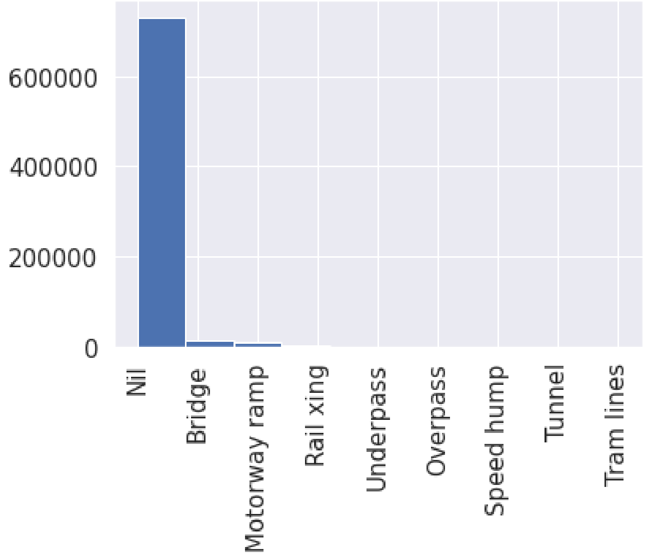


From above plots we can see that atleast parked vehicle among the objects were struck most of the times. Then fences, ditches, posts or poles and trees. We see that most of the time one objects are struck during the crash.

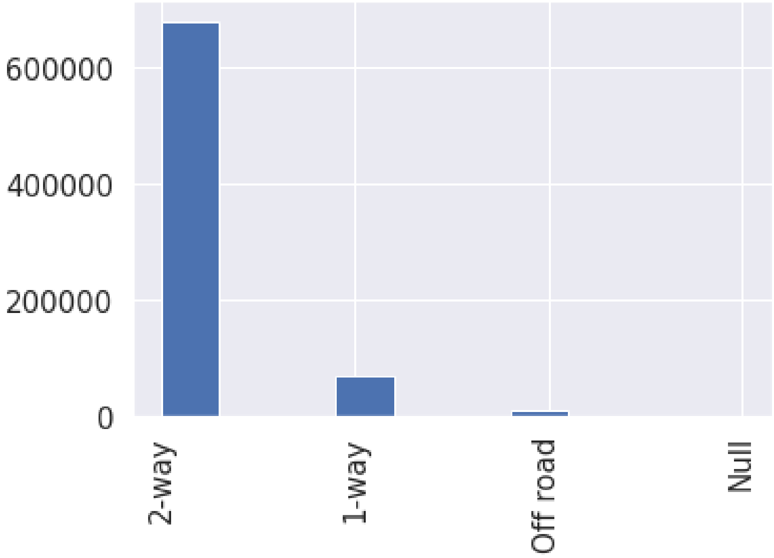
8. Road

In this section we have road related variables -

'**roadCharacter**' - The general nature of the road. Possible values include 'Bridge', 'Motorway Ramp', ‘Rail crossing’ or 'Nil'.

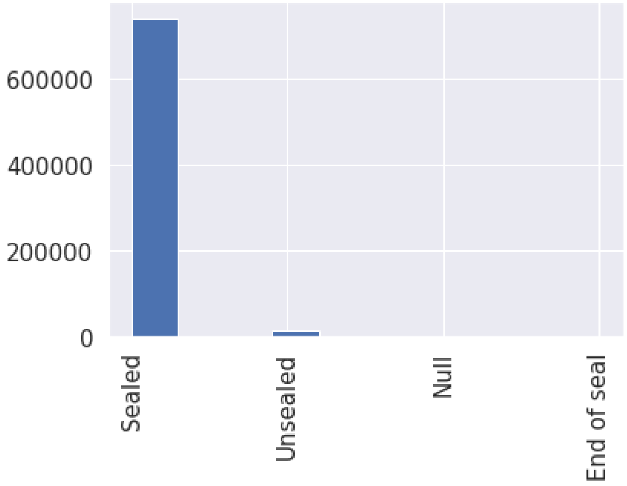


'**roadLane**' - The lane configuration of the road. Possible values : '1' (one way), '2' (two way), 'M' (for where a median exists), 'O' (for off-road lane configuations), ' ' (for unknown or invalid configuarations).

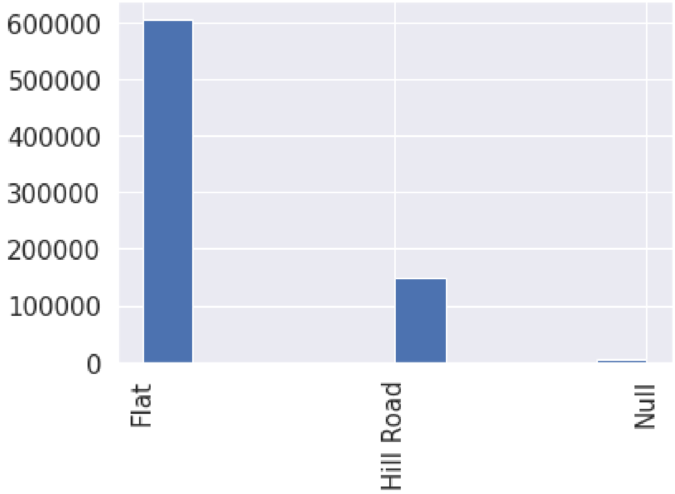


roadLane is a useful plot where we see that most of the crashes occur in 2way lanes than 1-way which is more than around 6.5 lakh.

'**roadSurface**' - The road surface description applying at the crash site. Possible values: 'Sealed' or 'Unsealed'.

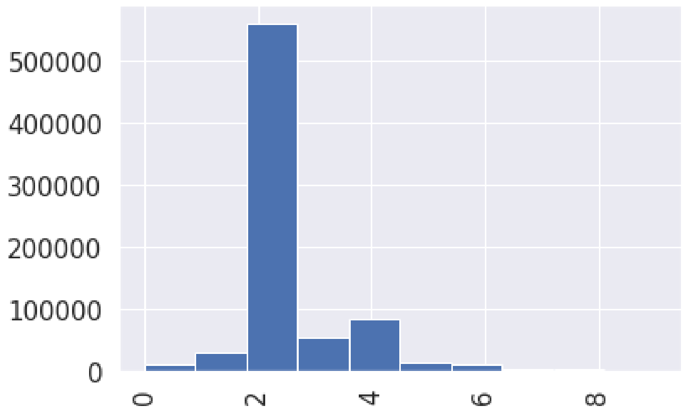


'**flatHill**' - Whether the road is flat or sloped. Possible values include 'Flat or 'Hill'.



In flatHill road type more than 6 lakh crashes occured on flat road.

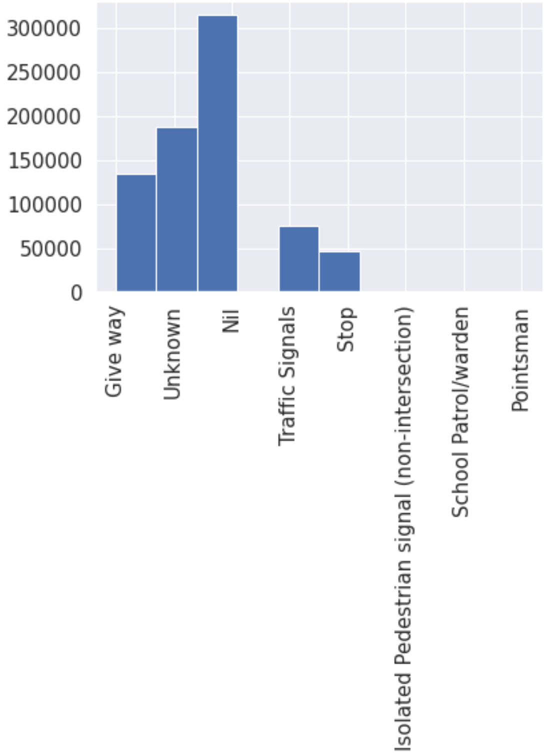
'**NumberOfLanes**' - The number(num) of lanes on the crash road.



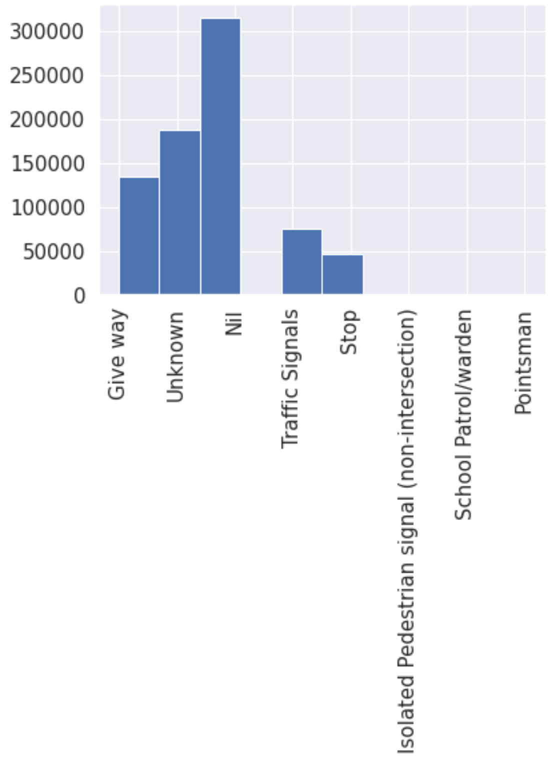
And from number of Lanes variable we see that max number of crashes occured in 2nd lane 5.5 lakh.

9. Other variables

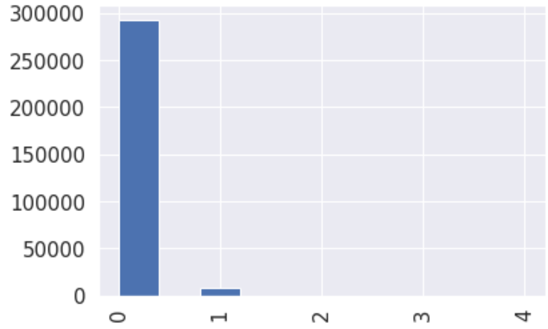
Traffic control - The traffic control (ctrl) signals at the crash site. Possible values are 'Traffic Signals', 'Stop Sign', 'Give Way Sign', 'Pointsman', 'School Patrol', 'Nil' or ' N/A'.



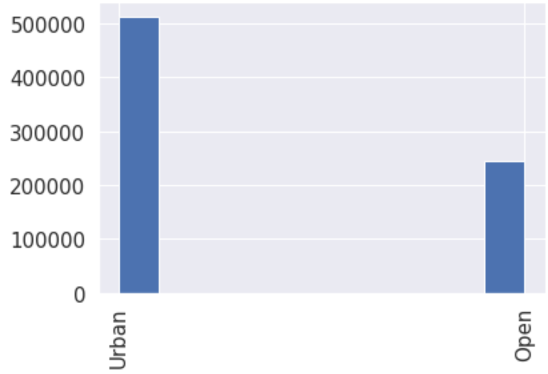
trafficSign - Derived variable to indicate how many times 'traffic signage' (including traffic signals, their poles, bollards or roadside delineators) was struck in the crash.



trafficIsland - Derived variable to indicate how many times a traffic island, medians (excluding barriers)was struck in the crash.



Urban - A derived variable using the 'spd_lim' variable. Possible values are 'Urban' (urban, spd_lim < 80) or 'Open Road' (open road, spd_lim >=80 or 'LSZ').



Conclusion

Ultimately when we have to choose independent variables for our model some variables like speed limit, crash severity, region would be very effective based on what we could see from the data visualizations. Also once we have our response variable correlation of each of independent variables with our response variable would give a clear picture.