2021 Data Analysis Report

This report is a summary of the pandas profiling done on the 2021 traffic, including all four datasets, acronymed as 2021 loc, 2021 occ, 2021 tbd, and 2021 unt.

Dataset 1: 2021 loc:

- Overall statistics:

Dataset statistics	
Number of variables	57
Number of observations	147724
Missing cells	1537371
Missing cells (%)	18.3%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	64.2 MiB
Average record size in memory	456.0 B

Variable types	
Numeric	36
Categorical	13
Text	7
Boolean	1

- Dataset Observations:

The dataset 2021 loc was cross-referenced with the loc data, which contained all the full form of the acronyms (referred to as layout) for better data analysis, as seen below through loc.info():

```
0   Collision Number                                    147724 non-null  int64
1   Trafficway                                          147724 non-null  int64
2   Type of Record (L-Location data)                    147724 non-null  object
3   Day of Week                                         147724 non-null  int64
4   On Route Auxiliary                                  146449 non-null  float64
5   Lane Number of Collision                            145298 non-null  object
6   Base Distance Direction                             147724 non-null  object
7   Light Condition                                     147724 non-null  int64
8   Weather Condition                                   147724 non-null  int64
9   Road Condition                                      147724 non-null  int64
10  Road Surface Conition                               147724 non-null  int64
11  First Harmful Event Location                        147724 non-null  int64
12  Crosswalk Indicator                                 147724 non-null  int64
13  Date of Collision                                   147724 non-null  float64
14  On Route Street Name                                137237 non-null  object
15  Investigating Jurisdiction Code                     147724 non-null  object
16  Number of Fatalities                                147724 non-null  float64
17  Number of Non-Fatal Injuries                        147724 non-null  float64
18  Amended or Corrected Indicator                      20980 non-null   object
19  Base Intersection Route Category                    147724 non-null  int64
20  Base Intersection Route Auxiliary                   145844 non-null  float64
21  Second Intersection Route Category                  147724 non-null  int64
22  Second Intersection Route Auxiliary                 145221 non-null  float64
23  Base Intersection Street Name                       142033 non-null  object
24  Second Intersection Street Name                     143447 non-null  object
25  Currently Junk field                                84769 non-null   float64
26  Number of Buses                                     84814 non-null   float64
27  Number of Persons Transported Immediately           85163 non-null   float64
28  Number of Towed Units                               85786 non-null   float64
29  Latitude of Collision (special format)              147724 non-null  object
30  Longitude of Collision (special format)             147724 non-null  int64
31  Junction Type                                       147724 non-null  int64
32  Other Contributing Factor 1                         25096 non-null   float64
33  Other Contributing Factor 2                         4660 non-null    float64
34  Other Contributing Factor 3                         557 non-null     float64
35  Other Contributing Factor 4                         72 non-null      float64
36  School Bus Involved                                 147724 non-null  int64
37  Work Zone Indicator                                 147724 non-null  int64
38  Work Zone Type                                      2608 non-null    float64
39  Work Zone Location                                  2608 non-null    float64
40  Workers Present Indicator                           2664 non-null    float64
41  Currently Junk field, was Badge Number of Investigating Officer  147703 non-null  object
42  Traffic Control Type                                147724 non-null  int64
43  Number of Units (Vehicle and Non-Motorists)         147724 non-null  int64
44  County of Collision                                 147724 non-null  int64
45  On Route Street Number                              135987 non-null  float64
46  Base Intersetion Street Number                      102776 non-null  float64
47  Second Intersection Street Number                   86256 non-null   float64
48  First Harmful Event                                 147724 non-null  int64
49  Primary Contributing Factor                         147724 non-null  int64
50  Military Time of Collision                          147724 non-null  int64
51  Base Distance Offset (from collision location to the base intersection in miles)  147724 non-null  object
52  Military Time of Police Notification                147724 non-null  int64
53  Military Time of Police Arrival at Scene            147724 non-null  int64
54  On Route Category                                   147724 non-null  int64
55  Alcohol/Drug Involved Driver in Collision           147724 non-null  object
56  Direction of Lane                                   129977 non-null  object
dtypes: float64(20), int64(24), object(13)
memory usage: 64.2+ MB
```

- Variable Observations:

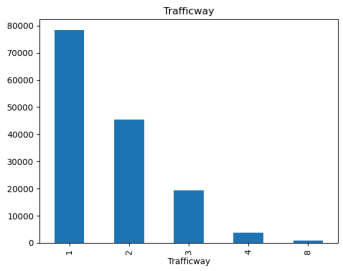
Collision Number:

Disregarded because it is used to give an ID to every collision, and therefore cannot be analyzed (or show anything that can support any conclusions)

Trafficway:

Trafficway
Categorical

Distinct	5
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	1.1 MiB



There are no missing values, and the numbers stand for different values

1 Two-Way, Not Divided
2 Two-Way,Divided,Unprotected Median
3 Two-Way,Divided,Barrier
4 One-Way
8 Other

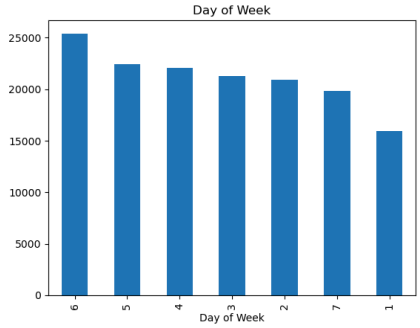
, so the frequency graph above indicates that the most number of collisions occurred on a two-way that was not divided, which makes sense, because it is the least protected by physical barriers.

Type of Record:

Disregarded because the constant L, indicates all the data in the dataset is locator data

Day of the Week:

Distinct	7	**Minimum**	1
Distinct (%)	< 0.1%	**Maximum**	7
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	4.1489264	**Memory size**	1.1 MiB



There are no missing values, and the numbers 1-7 stand for the days of the week (shown below)

1 Sunday
2 Monday
3 Tuesday
4 Wednesday
5 Thursday
6 Friday
7 Saturday

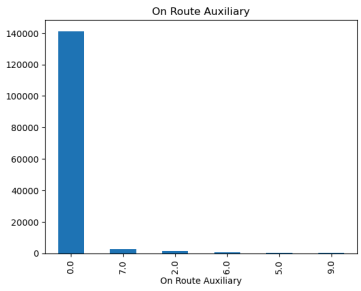
. This demonstrates that the most frequent day of collisions is Friday, probably due to the end of the week, and a rush to get home (or people going out)

On Route Auxiliary (was this the main road/common road taken):

On Route Auxiliary
Real number (ℝ)

ZEROS

Distinct	6	**Minimum**	0
Distinct (%)	< 0.1%	**Maximum**	9
Missing	1275	**Zeros**	141233
Missing (%)	0.9%	**Zeros (%)**	95.6%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	0.19126112	**Memory size**	1.1 MiB



There are some missing values (0.9%) from the data, which limit accurate data analysis. The numerical values have translational value, as indicated below:

0 Main
2 Alternate
5 Spur
6 Connection
7 Business
9 Other

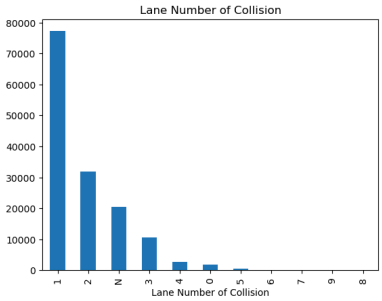
. Here, a predominant amount of collisions occurred on main routes, which makes sense, because they would be the busiest.

Lane Number of Collision:

Lane Number of Collision
Categorical

MISSING

Distinct	11
Distinct (%)	< 0.1%
Missing	2426
Missing (%)	1.6%

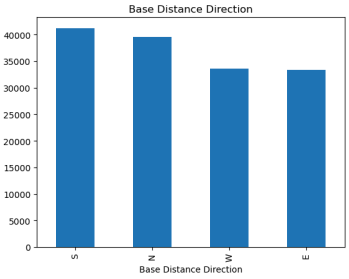


Here, there are a lot of missing values (1.6%). There is no key in the loc layout, so the lane values are up to interpretation and must be standardized

Base Distance Direction:

Base Distance Direction
Categorical

Distinct	4
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%



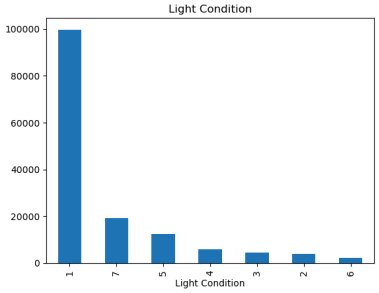
Here, there are no missing values. The greatest collisions occur in the south direction, per the frequency graph, as indicated by the key (seen below)

N North
S South
E East
W West
U Unknown

Light Conditions:

Light Condition
Real number (ℝ)

Distinct	7	**Minimum**	1
Distinct (%)	< 0.1%	**Maximum**	7
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	2.3956161	**Memory size**	1.1 MiB



There are no missing values. The values of 1-7 are associated with weather conditions, as shown below:

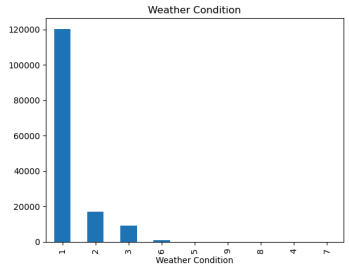
1 Daylight
2 Dawn
3 Dusk
4 Dark (Lighting Unspecified)
5 Dark (Street Lamp Lit)
6 Dark (Street Lamp Not Lit)
7 Dark (No Lights)

. The most often light conditions when collisions occurred was daylight, which makes sense, because that's when most people will be on the road

Weather Conditions:

Weather Condition
Real number (ℝ)

Distinct	9	**Minimum**	1
Distinct (%)	< 0.1%	**Maximum**	9
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1.2786751	**Memory size**	1.1 MiB



There are no missing values. The values 1-9 signify weather conditions as below:

1 Clear, No Adverse Conditions
2 Rain
3 Cloudy
4 Sleet Or Hail
5 Snow
6 Fog,Smog,Smoke
7 Blowing Sand, Soil, Dirt Or Snow
8 Severe Cross Winds, High Wind
9 Unknown

. The most collisions occur in clear conditions, which makes sense, because that's the most common weather condition over time

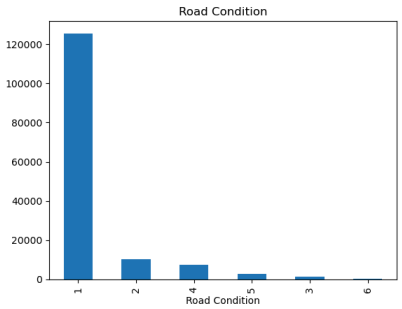
Road Condition + Road Surface Condition:

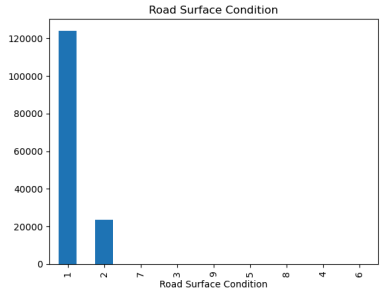
Road Condition
Real number (ℝ)

Distinct	6	**Minimum**	1
Distinct (%)	< 0.1%	**Maximum**	6
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1.3251672	**Memory size**	1.1 MiB

Road Surface Conition
Real number (ℝ)

Distinct	9	**Minimum**	1
Distinct (%)	< 0.1%	**Maximum**	9
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1.1696407	**Memory size**	1.1 MiB





Here, there is no missing data. The numerical values of 1-6 for road condition stand for

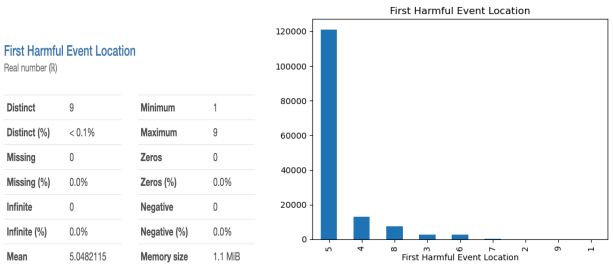
1 Straight-Level
2 Straight-On Grade
3 Straight-Hillcrest
4 Curve-Level
5 Curve-On Grade
6 Curve-Hillcrest

, whereas the values of 1-9 for road surface condition is

1 Dry
2 Wet
3 Snow
4 Slush
5 Ice
6 Contaminant (Sand, mud, Dirt, oil, Etc.)
7 Water(Standing)
8 Other
9 Unknown

. The most common are straight-level, dry roads, which is how most roads are

First Harmful Event Location:

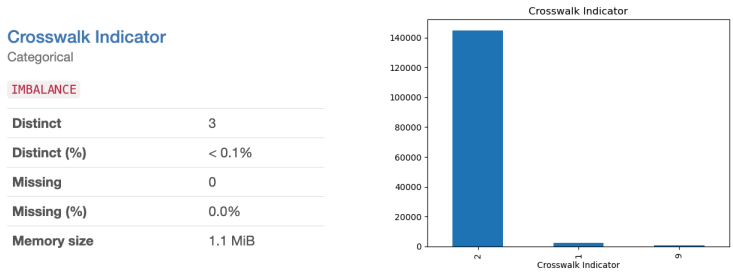


There are no missing values. The values 1-9 signify the different locations where the collisions have occurred

1 Gore
2 Island
3 Median
4 Roadside
5 Roadway
6 Shoulder
7 Sidewalk
8 Outside Trafficway
9 Unknown

. The most often ones have happened on the roadway itself, where most cars are

Crosswalk Indicator:



There is no missing data. This one is different because the layout has a Y for yes, N for no, and U and a blank for unknown, but this profile shows numerical values, creating issues in data analysis.

Date of Collision:

Disregarded because it just provides a timestamp for the collision, nothing else

On Route Street Name:

Disregarded because it does not specify county name or location, so many streets of the same name are repeated, causing problems in data analysis

Investigating Jurisdiction Code:

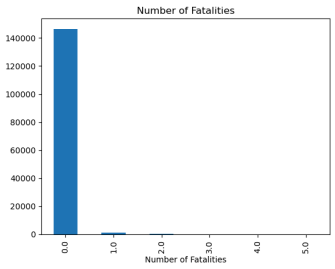
Disregarded because it provides an authoritative code rather than a zip code, which hinders data analysis

Number of Fatalities:

Number of Fatalities
Real number (ℝ)

ZEROS

Distinct	6	Minimum	0
Distinct (%)	< 0.1%	Maximum	5
Missing	0	Zeros	146612
Missing (%)	0.0%	Zeros (%)	99.2%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	0.0081097181	Memory size	1.1 MiB



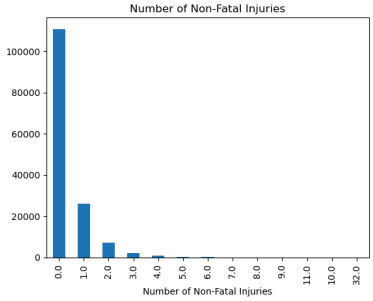
There is no missing data. Here, most collisions occurred have had no fatalities, a good indication

Number of Non-Fatal Injuries:

Number of Non-Fatal Injuries
Real number (ℝ)

ZEROS

Distinct	13	Minimum	0
Distinct (%)	< 0.1%	Maximum	32
Missing	0	Zeros	110865
Missing (%)	0.0%	Zeros (%)	75.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	0.36281173	Memory size	1.1 MiB



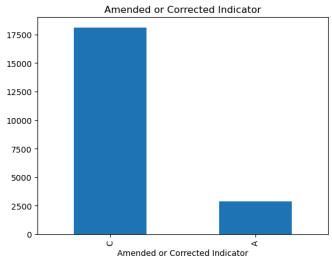
There is no missing data. Here, the highest frequency of non-fatal injuries is from 0-3, indicating very little injuries in these collisions.

Amended or Corrected Indicator:

Amended or Corrected Indicator
Categorical

MISSING

Distinct	2
Distinct (%)	< 0.1%
Missing	126744
Missing (%)	85.8%
Memory size	1.1 MiB

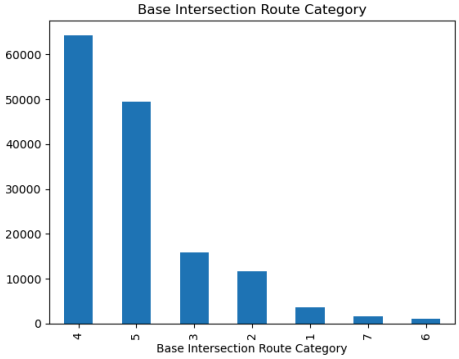


Most of the data is missing (85.8%). However, the data that is available, most of the indicators are C (corrected) while very little are A (amended)

Base Intersection Route Category:

Base Intersection Route Category
Real number (ℝ)

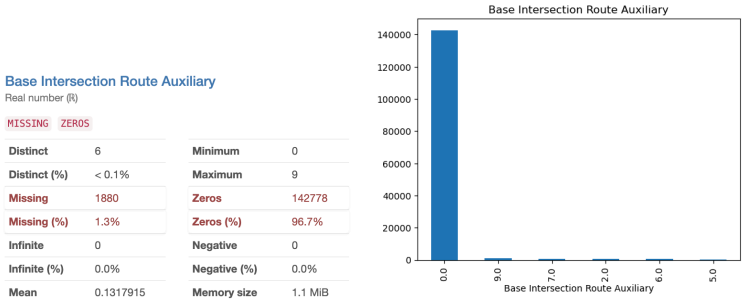
Distinct	7	Minimum	1
Distinct (%)	< 0.1%	Maximum	7
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	4.0446576	Memory size	1.1 MiB



There is no missing data, and the values are corresponded to:

1 Interstate
2 US Primary
3 SC Primary
4 Secondary
5 County
6 Other , so most common were collisions on secondary and county roads, which is unexpected

Base Intersection Route Auxiliary:



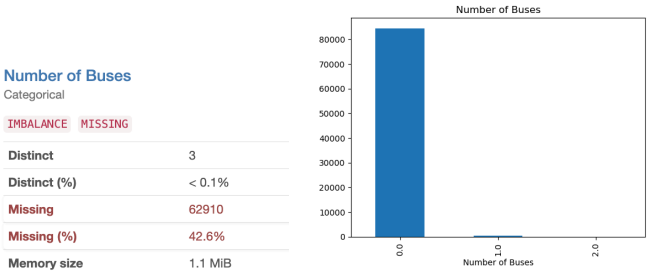
There are some missing values (1.3%) from the data, which limit accurate data analysis. The numerical values have translational value, as indicated below:

0 Main
2 Alternate
5 Spur
6 Connection
7 Business
9 Other. Here, a predominant amount of collisions occurred on main routes, which makes sense, because they would be the busiest.

Second Intersection Route Category/Auxiliary:
Disregarded because they did not add much to the data analysis

Base/Second Intersection Street Name:
Disregarded because it does not specify county name or location, so many streets of the same name are repeated, causing problems in data analysis

Number of Buses:



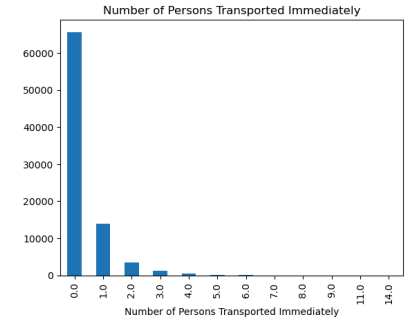
Almost half of the dataset (42.6%) is missing, but the data that is available shows that no buses is involved in a bus-related collision

Number of Persons Transported Immediately:

Number of Persons Transported Immediately
Real number (ℝ)

MISSING ZEROS

Distinct	12	Minimum	0
Distinct (%)	< 0.1%	Maximum	14
Missing	62561	Zeros	65657
Missing (%)	42.3%	Zeros (%)	44.4%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	0.33305543	Memory size	1.1 MiB



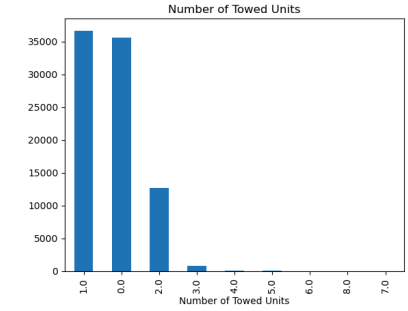
Almost half of the dataset here is missing (42.3%), but when the data is available, anywhere from 0-3 people are transported immediately, which supports the non-fatality of the collesions

Number of Towed Units:

Number of Towed Units
Real number (ℝ)

MISSING ZEROS

Distinct	9	Minimum	0
Distinct (%)	< 0.1%	Maximum	8
Missing	61938	Zeros	35577
Missing (%)	41.9%	Zeros (%)	24.1%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	0.75521647	Memory size	1.1 MiB



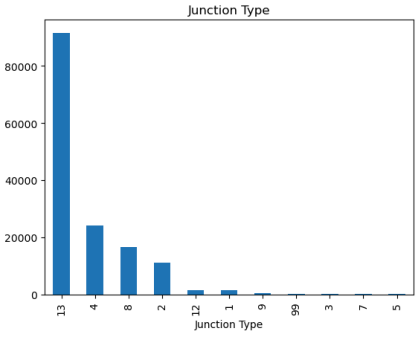
Almost half of the dataset (41.9%) is missing, but when the data is available, mostly 1 or 2 vehicles is towed

Latitude + Longitude of Collision:
Disregarded because it provides no analytical data, just exact location, which can be used, but not for exploratory purposes.

Junction Type:

Junction Type
Real number (ℝ)

Distinct	11	Minimum	1
Distinct (%)	< 0.1%	Maximum	99
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	10.161368	Memory size	1.1 MiB



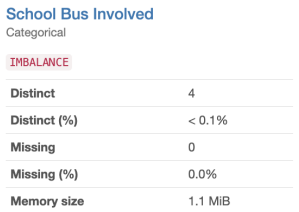
There is no missing data. The values of 1-99 have meaning indicated below:

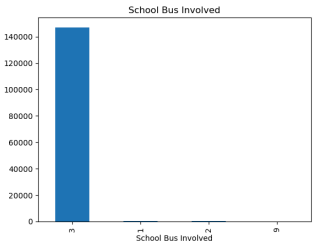
01 Cross-Over
02 Driveway
03 Five/More Points
04 Four-Way Intersection
05 Railway Grade Crossing
07 Shared Use Paths Or Trail
08 T-Intersection
09 Traffic Circle
12 Y-Intersection
13 Non-Junction
99 Unknown , therefore most collisions occur in non-junctions

Other Contributing Factors 1, 2, 3, 4:

Disregarded because their numerical value doesn't have any meaning in the layout that is useful for data analysis, and most of the data is missing

School Bus Involved:

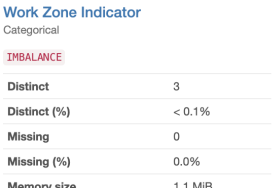




There is no missing data. However, the values 1-9 correspond with the values below:

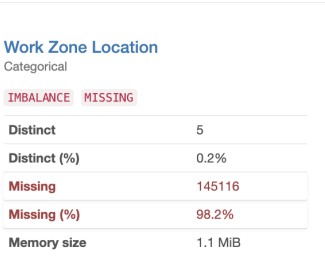
1 Yes, Directly
2 Yes, Indirectly
3 No
9 Unknown , and the data shows that most collisions occur with 3 school buses

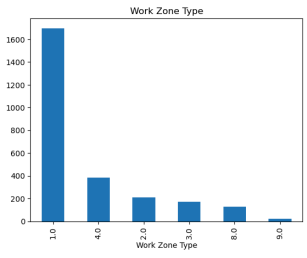
Work Zone Indicator:

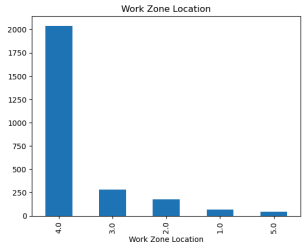


There is no missing data, and the layout for this variable does not match the categorical labels of 1, 2, or 9 (it has Y for yes, N for No, and U and blank for unknown)

Work Zone Type and Location:







Most of both datasets (98.2% each) have missing values. But the ones they do have, have the values below:

Work Zone Type	1 Shoulder/Median Work 2 Lane Shift/Cross-Over 3 Intermittent/Moving Work 4 Lane Closure 8 Other 9 Unknown
Work Zone Location	1 Before First Signal 2 Advanced Warning Area 3 Transition Area 4 Activity Area 5 Termination Area

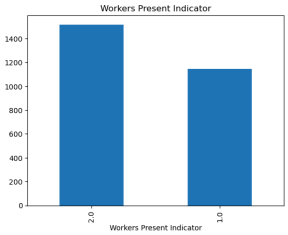
Most collisions occur in active shoulder work areas

Workers Present Indicator:

Workers Present Indicator
Categorical

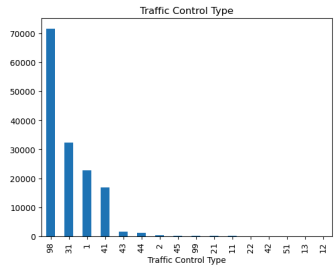
MISSING

Distinct	2
Distinct (%)	0.1%
Missing	145060
Missing (%)	98.2%
Memory size	1.1 MiB



Most of the data is missing (98.2%), but where it is present, it cannot be determined because the layout does not match

Traffic Control Type:



There is no missing data, and the values, according to the layout mean:

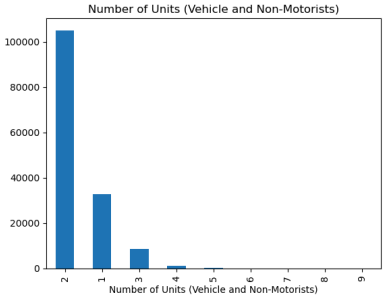
01 Stop And Go Light
02 Flashing Traffic Signal
11 RR (X-Bucks, Lights And Gates)
12 RR (X-Bucks And Lights)
13 RR (X-Bucks Only)
21 Officer Or Flagman
22 Oncoming Emergency Vehicle
31 Pavement Markings(Only)
41 Stop Sign
42 School Zone Sign
43 Yield Sign
44 Work Zone
45 Other Warning Signs
51 Flashing Beacon
98 None
99 Unknown

, which means most collisions have no traffic control type, contributing to the fact that most collisions occurred in areas without any significant traffic control

Number of Units (Vehicle and Non-Motorists):

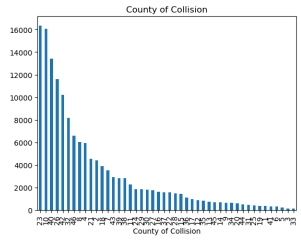
Number of Units (Vehicle and Non-Motorists)
Real number (ℝ)

Distinct	9	**Minimum**	1
Distinct (%)	< 0.1%	**Maximum**	9
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1.8575248	**Memory size**	1.1 MiB



There are no missing values, but mostly there are 2-3 vehicle involved collisions, as shown in the graph

County of Collision:



There are no missing values, and the values from 1-46 can be explained by the key in the layout:

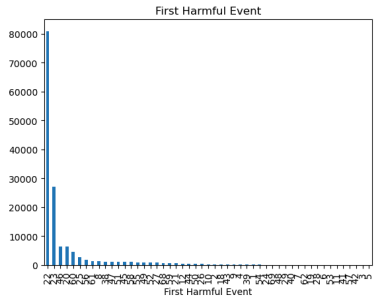
01 Abbeville
02 Aiken
03 Allendale
04 Anderson
05 Bamberg
06 Barnwell
07 Beaufort
08 Berkeley
09 Calhoun
10 Charleston
11 Cherokee
12 Chester
13 Chesterfield
14 Clarendon
15 Colleton
16 Darlington
17 Dillon
18 Dorchester
19 Edgefield
20 Fairfield
21 Florence
22 Georgetown
23 Greenville
24 Greenwood
25 Hampton
26 Horry
27 Jasper
28 Kershaw
29 Lancaster
30 Laurens
31 Lee
32 Lexington
33 McCormick
34 Marion
35 Marlboro
36 Newberry
37 Oconee
38 Orangeburg
39 Pickens
40 Richland
41 Saluda
42 Spartanburg
43 Sumter
44 Union
45 Williamsburg
46 York

, and the frequency graph shows the highest in Greenville, which is also a populous county

First Harmful Event:

First Harmful Event
Real number (ℝ)

Distinct	51	**Minimum**	1
Distinct (%)	< 0.1%	**Maximum**	69
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	27.080007	**Memory size**	1.1 MiB

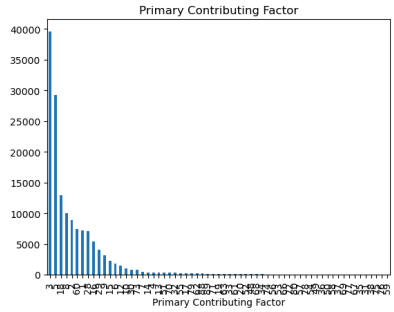


No missing data, and the values 1-99 correspond with the following:

00 None Listed	None Listed
01 Non-Collision	Cargo/Equip Loss Or Shift
02 Non-Collision	Cross Median/Center Line
03 Non-Collision	Downhill Runaway
04 Non-Collision	Equipment Failure
05 Non-Collision	Fire/Explosion
06 Non-Collision	Immersion
07 Non-Collision	Jackknife
08 Non-Collision	Overturn/Rollover
09 Non-Collision	Run Off Road Left
10 Non-Collision	Run Off Road Right
11 Non-Collision	Separation Of Units
12 Non-Collision	Spill (Two Wheel Vehicle)
18 Non-Collision	Other Non-Collision
19 Non-Collision	Unknown Non-Collision
20 Collision: Object Not Fixed	Animal (Deer Only)
21 Collision: Object Not Fixed	Animal (Not Deer)
22 Collision: Object Not Fixed	Motor Vehicle (In Transport)
23 Collision: Object Not Fixed	Motor Vehicle (Stopped)
24 Collision: Object Not Fixed	Motor Vehicle (Other Roadway)
25 Collision: Object Not Fixed	Motor Vehicle (Parked)
26 Collision: Object Not Fixed	Pedalcycle
27 Collision: Object Not Fixed	Pedestrian
28 Collision: Object Not Fixed	Railway Vehicle
29 Collision: Object Not Fixed	Work Zone Maint. Equip.
38 Collision: Object Not Fixed	Other Movable Object
39 Collision: Object Not Fixed	Unknown Movable Object
40 Collision: Object Fixed	Bridge Overhead Structure
41 Collision: Object Fixed	Bridge Parapet End
42 Collision: Object Fixed	Bridge Pier Or Abutment
43 Collision: Object Fixed	Bridge Rail
44 Collision: Object Fixed	Culvert
45 Collision: Object Fixed	Curb
46 Collision: Object Fixed	Ditch
47 Collision: Object Fixed	Embankment
48 Collision: Object Fixed	Equipment
49 Collision: Object Fixed	Fence
50 Collision: Object Fixed	Guardrail End
51 Collision: Object Fixed	Guardrail Face
52 Collision: Object Fixed	HWY Traffic Sign Post
53 Collision: Object Fixed	Impact Attenuator/Crash Cushion
54 Collision: Object Fixed	Light Luminaire Support
55 Collision: Object Fixed	Mailbox
56 Collision: Object Fixed	Median Barrier
57 Collision: Object Fixed	Overhead Sign Support
58 Collision: Object Fixed	Other(Post,Pole,Support,Etc.)
59 Collision: Object Fixed	Other(Wall,Bldg,Tunnel,Etc.)
60 Collision: Object Fixed	Tree
61 Collision: Object Fixed	Utility Pole
62 Collision: Object Fixed	Workzone Maint. Equip.
68 Collision: Object Fixed	Other
69 Collision: Object Fixed	Unknown Fixed Object
99 Collision: Object Fixed	Unknown

, thus most collisions have been with another transport vehicle in motion

Primary Contributing Factor:



No missing data, and the values 1-89 correspond with:

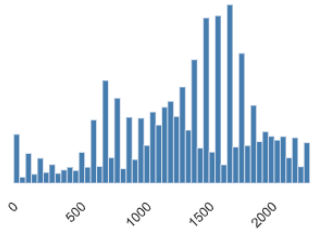
01 1 Disregarded Signs/Signals/Etc.	
02 1 Distracted/Inattention	50 3 Non-Motorist Inattentive
03 1 Driving Too Fast for Conditions	51 3 Lying &/Or Illegally In Roadway
04 1 Exceeded Authorized Speed Limit	52 3 Non-Motorist Failed To Yield ROW
05 1 Failed To Yield Right of Way	53 3 Not Visible(Dark Clothing)
06 1 Ran Off Road	54 3 Non-Motorist Disregarded Signs/Signals/Etc
07 1 Fatigued/Asleep	55 3 Improper Crossing
08 1 Followed Too Closely	56 3 Darting
09 1 Made an Improper Turn	57 3 Wrong Side Of Road
10 1 Medical Related	58 3 Other Non-Motorist Factor
12 1 Aggressive Operation of Vehicle	59 3 Non-Motorist Unknown
13 1 Over-Correcting/Over-Steering	60 4 Animal In Road
14 1 Swerving To Avoid Object	61 4 Glare
15 1 Wrong Side/Wrong Way	62 4 Obstruction
16 1 Driver Under Influence	63 4 Weather Condition
17 1 Vision Obscured (Within Unit)	66 3 Non-Motorist Under Influence
18 1 Improper Lane Usage/Change	67 3 Other Person Under Influence
19 1 On Cell Phone	68 4 Other Environmental Factor
20 1 Texting	69 4 Unknown Environmental Factor
28 1 Other Improper Action	70 5 Brakes
29 1 Unknown	71 5 Steering
30 2 Debris	72 5 Power Plant
31 2 Non-Highway Work	73 5 Tires/Wheels
32 2 Obstruction In Roadway	74 5 Lights
33 2 Road Surface Condition (ie. Wet)	75 5 Signals
34 2 Rut, Holes, Bumps	76 5 Windows/Windshield
35 2 Shoulders(None,Low,Soft,High)	77 5 Restraint Systems
36 2 Traffic Control Device(ie. Missing)	78 5 Truck Coupling
37 2 Work Zone(Constr/Maintenance/Util)	79 5 Cargo
38 2 Worn, Travel-Polished Surface	80 5 Fuel System
48 2 Other Roadway Factor	88 5 Other Vehicle Defect
49 2 Unknown	89 5 Unknown Vehicle Defect

, so most common contributing factor was driving too fast

Military Time of Collision:

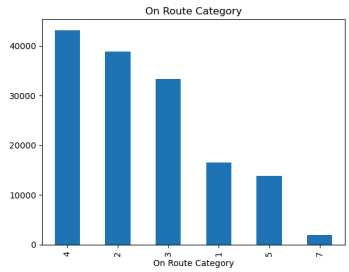
Military Time of Collision
Real number (ℝ)

Distinct	1440	**Minimum**	0
Distinct (%)	1.0%	**Maximum**	2359
Missing	0	**Zeros**	548
Missing (%)	0.0%	**Zeros (%)**	0.4%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1357.4809	**Memory size**	1.1 MiB



No missing data, but most of the times are in the afternoon, which are peak driving hours when there are a lot of people on the roads (due to getting off work/school)

On Route Category:



There is no missing data, and the values are corresponded to:

1 Interstate
2 US Primary
3 SC Primary
4 Secondary
5 County
6 Other

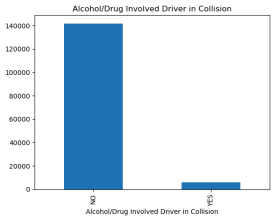
, so most common was on secondary roads

Alcohol/Drug Involved Driver in Collision:

Alcohol/Drug Involved Driver in Collision
Boolean

IMBALANCE

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	144.4 KiB



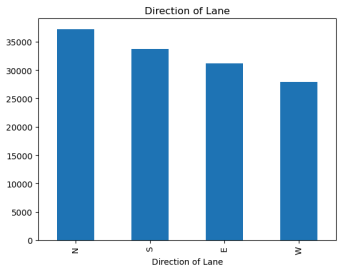
There is no missing data, and most collisions didn't include an intoxicated driver

Direction of Lane:

Direction of Lane
Categorical

MISSING

Distinct	4
Distinct (%)	< 0.1%
Missing	17747
Missing (%)	12.0%
Memory size	1.1 MiB



12.0% of the data is missing, but most of the collision lanes were facing north.

Dataset 2: 2021 occ:

- Overall statistics:

Dataset statistics	
Number of variables	17
Number of observations	362070
Missing cells	1276396
Missing cells (%)	20.7%
Duplicate rows	151
Duplicate rows (%)	< 0.1%
Total size in memory	47.0 MiB
Average record size in memory	136.0 B

Variable types	
Numeric	6
Categorical	9
Unsupported	1
Text	1

- Dataset Observations:

The dataset 2021 occ was cross-referenced with the occ data, which contained all the full form of the acronyms (referred to as layout) for better data analysis, as seen below through occ.info():

```
0   Collision Number                  362070 non-null  int64
1   Unit Number                       362070 non-null  int64
2   Person Seating Location           362070 non-null  int64
3   Record Type                       360170 non-null  object
4   Currently Junk Variable           0 non-null       float64
5   Person Gender                     362070 non-null  object
6   Person Race                       362070 non-null  object
7   Person Age                        341319 non-null  float64
8   Restraint/Safety Device           362070 non-null  int64
9   Location After Impact             362070 non-null  int64
10  Injury Status                     362070 non-null  int64
11  Motorcycle Head Injury            6396 non-null    object
12  Ejection Status                   362070 non-null  int64
13  Transported to Medical Facility?  362070 non-null  int64
14  Transport by whom                 46717 non-null   float64
15  Air Bag Deployment                362070 non-null  int64
16  Person Zip Code                   141422 non-null  object
```

- Variable Observations:

<u>Collision Number:</u>

Disregarded because it is used to give an ID to every collision, and therefore cannot be analyzed (or show anything that can support any conclusions)

<u>Unit Number:</u>

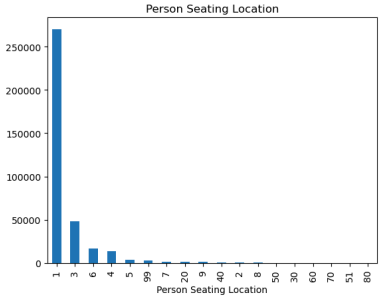
Disregarded because it is used to give an ID to the collision

<u>Person Seating Location:</u>

Person Seating Location

Real number (ℝ)

Distinct	18	Minimum	1
Distinct (%)	< 0.1%	Maximum	99
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	2.8251277	Memory size	2.8 MiB



No missing data, and the values 1-99 correspond to the values below:

01 Driver
02 1st Row Middle
03 1st Row Right
04 2nd Row Left
05 2nd Row Middle
06 2nd Row Right
07 3rd Row Left
08 3rd Row Middle
09 3rd Row Right
20 Pedestrian
30 Trailing Unit
40 Bus Or Van (4th Row Or Higher)
50 Other Enclosed Area (Nontrailing)
51 Other Unenclosed Area (Nontrailing)
60 Sleeper Of Cab
70 Riding On Unit Exterior
80 Lap
99 Unknown/NA, thus most collisions occurred with only the driver in the car

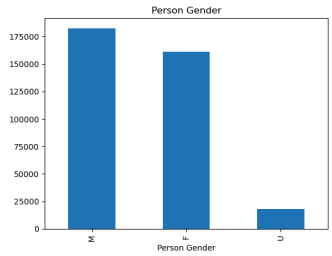
Record Type:

Disregarded because it has a constant value of O, which, according to the 2021 occ raw data layout, stands for a person. Therefore, its constant value indicates accident data only where a person is involved

Person Gender:

Person Gender
Categorical

Distinct	3
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	2.8 MiB

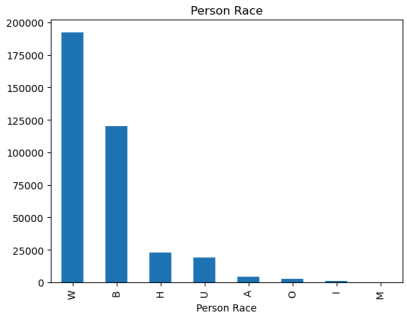


No missing data, and most people involved in collisions are male (M is male, F is female, and U is unknown)

Person Race:

Person Race
Categorical

Distinct	8
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	2.8 MiB



No missing data, and the values are as such:

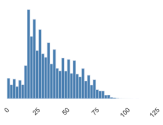
W Caucasian
B African American
O Other
I Alaskan Native/American Indian
A Asian/Pacific Islander
H Hispanic
U Unknown, thus most are caucasian in collisions

Person Age:

Person Age
Real number (ℝ)

MISSING

Distinct	109	**Minimum**	0
Distinct (%)	< 0.1%	**Maximum**	121
Missing	20751	**Zeros**	2017
Missing (%)	5.7%	**Zeros (%)**	0.6%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	36.701775	**Memory size**	2.8 MiB



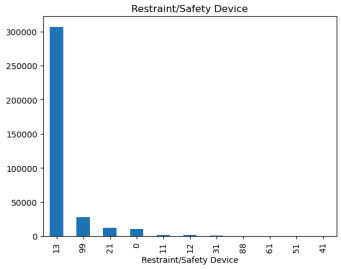
5.7% of the data is missing, but most collisions occur in ages below 25 and above 20

Restraint/Safety Device:

Restraint/Safety Device
Real number (ℝ)

ZEROS

Distinct	11	**Minimum**	0
Distinct (%)	< 0.1%	**Maximum**	99
Missing	0	**Zeros**	10278
Missing (%)	0.0%	**Zeros (%)**	2.8%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	19.693443	**Memory size**	2.8 MiB



There is no missing data, and the values are as such:

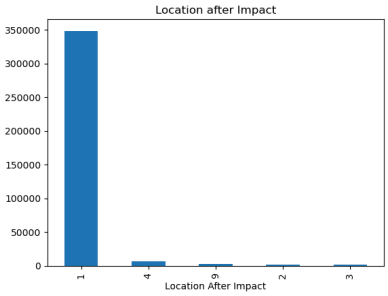
00 None Used
11 Shoulder Belt Only
12 Lap Belt Only
13 Shoulder And Lap Belt
21 Child Safety Seat
31 Helmet
41 Protective Pads
51 Reflective Clothing
61 Lighting
88 Other
99 Unknown, thus the most collisions only had something similar to a shoulder/lap belt

Location after Impact:

Location After Impact
Categorical

IMBALANCE

Distinct	5
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%



No missing data, and the values are as such:

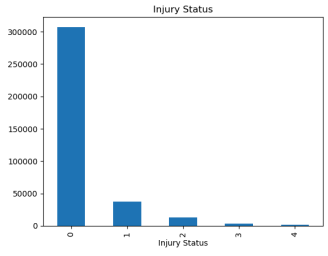
1 Not Trapped
2 Extricated(Mech Means)
3 Freed(Non-Mech Means)
4 Not Applicable
9 Unknown, thus most collisions resulted in no trapping of people

Injury Status:

Injury Status
Categorical

IMBALANCE

Distinct	5
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%



No missing data, and values are as such:

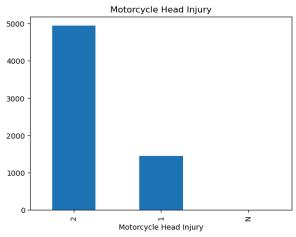
0 No Apparent Injury
1 Possible Injury
2 Suspected Minor Injury
3 Suspected Serious Injury
4 Fatal Injury, thus most collisions had no apparent injury

Motorcycle Head Injury:

Motorcycle Head Injury
Categorical

IMBALANCE MISSING

Distinct	3
Distinct (%)	< 0.1%
Missing	355674
Missing (%)	98.2%



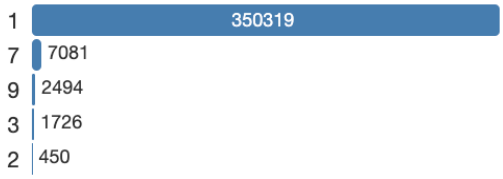
97.8% of the data is missing, but the one that is available suggests that there was no head injury (1 is yes, 2 is no, 9 is unknown)

Ejection Status:

Ejection Status
Categorical

IMBALANCE

Distinct	5
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%



No missing data, and values are as such:

1 Not Ejected
2 Partially Ejected
3 Totally Ejected
7 Not Applicable
9 Unknown

, thus most weren't ejected

Transported To a Medical Facility?

Transported to Medical Facility?
Categorical

IMBALANCE

Distinct	3
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%

No missing data, and the values are as such:

1 Transported To Medical Facility
2 Not Transported To Medical Facility
3 Unknown

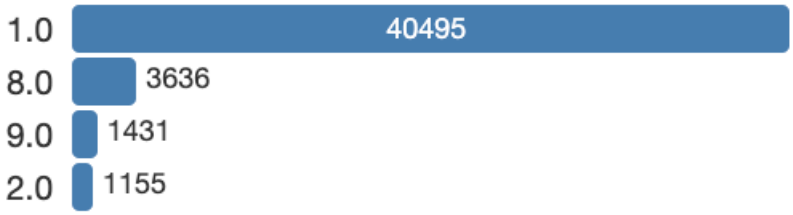
, most weren't transported to a medical facility

Transported by Whom?

Transport by whom
Categorical

IMBALANCE MISSING

Distinct	4
Distinct (%)	< 0.1%
Missing	315353
Missing (%)	87.1%



Most of the data is missing (87.1%), but values are as such:

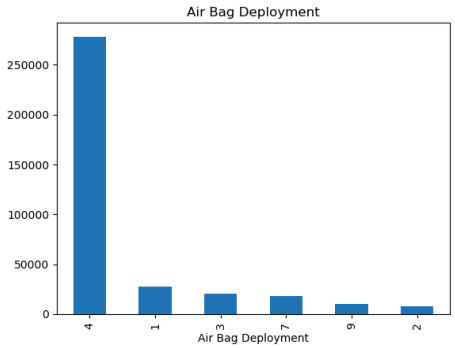
1 EMS
2 Police
8 Other
9 Unknown

, so most were transported by EMS

Air Bag Deployment:

Air Bag Deployment
Real number (ℝ)

Distinct	6	**Minimum**	1
Distinct (%)	< 0.1%	**Maximum**	9
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	3.9594802	**Memory size**	2.8 MiB



No missing data, and the values are as such:

1 Deployed Front
2 Deployed Side
3 Deployed Both
4 Not Deployed
7 Not Applicable
9 Deployment Unknown , most were not deployed during a collision

Person Zip Code:

Person Zip Code
Text

MISSING

Distinct	6440
Distinct (%)	4.6%
Missing	220648
Missing (%)	60.9%
Memory size	2.8 MiB

60.9% of the data is missing, but the two most common zip codes are 29483 and 29445

Dataset 3: 2021 tbd:

- Overall statistics:

Dataset statistics	
Number of variables	22
Number of observations	3368
Missing cells	20327
Missing cells (%)	27.4%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	579.0 KiB
Average record size in memory	176.0 B

Variable types	
Numeric	5
Text	6
Categorical	11

- Dataset Observations:

The dataset 2021 tbd was cross-referenced with the tbd data, which contained all the full form of the acronyms (referred to as layout) for better data analysis, as seen below through tbd.info():

```
0   Collision Number                    3368 non-null   int64
1   Unit Number                         3368 non-null   int64
2   Carrier Name                        2987 non-null   object
3   Carrier Street                      2971 non-null   object
4   Carrier City                        2965 non-null   object
5   Carrier State                       2826 non-null   object
6   Carrier Zip                         2910 non-null   object
7   Carrier DOT Number                  2173 non-null   object
8   Access Control                      2900 non-null   float64
9   Carrying Hazardous Materials?       3085 non-null   float64
10  Hazardous Materials Placard?        2925 non-null   float64
11  Hazardous Materials Class           178 non-null    float64
12  Hazardous Materials ID              72 non-null     object
13  Hazardous Materials Released?       2960 non-null   float64
14  Gross Veh Weight Rating/Combo Rating 3090 non-null  float64
15  Vehicle Configuration               3110 non-null   float64
16  Trailer Length1 Code                2312 non-null   float64
17  Trailer Length2 Code                444 non-null    float64
18  Trailer Width1 Code                 2289 non-null   float64
19  Trailer Width2 Code                 442 non-null    float64
20  Citation Issued                     3368 non-null   int64
21  Carrier Type                        3026 non-null   float64
```

- Variable Observations:

<u>Collision Number:</u>

Disregarded because it is used to give an ID to every collision, and therefore cannot be analyzed

<u>Unit Number:</u>

Disregarded because it is used to give an ID to the collision

<u>Carrier Name + Street:</u>

Disregarded because it is used to give a name to the people involved in the collision

<u>Carrier City</u>:

Disregarded because it provides no insight into the actual collisions themselves

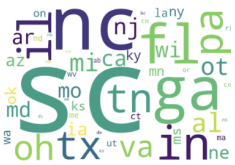
<u>Carrier State:</u>

Carrier State

Text

MISSING

Distinct	52
Distinct (%)	1.8%
Missing	542
Missing (%)	16.1%
Memory size	26.4 KiB



16.1% of the data is missing, but the most common carrier state is South Carolina, which makes sense, considering that is where the data is from

Carrier Zip:

Carrier Zip
Text

MISSING

Distinct	1266
Distinct (%)	43.5%
Missing	458
Missing (%)	13.6%
Memory size	26.4 KiB

13.6% of the data is missing, but most common zip code is 29201

Carrier Dot Number:

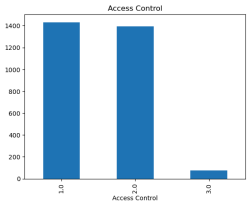
Disregarded because it does not add much to the data analysis

Access Control:

Access Control
Categorical

MISSING

Distinct	3
Distinct (%)	0.1%
Missing	468
Missing (%)	13.9%



13.9% of the data is missing, but the values are as such:

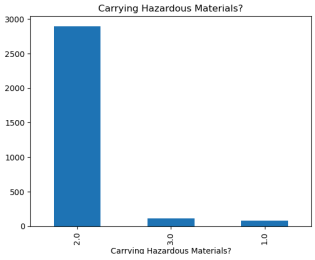
1 No Access Control
2 Full Access Control
3 Partial Access Control
0 Unknown Access Control , thus most had no access control or full access control

Carrying Hazardous Materials?

Carrying Hazardous Materials?
Categorical

IMBALANCE MISSING

Distinct	3
Distinct (%)	0.1%
Missing	283
Missing (%)	8.4%



8.4% of the data is missing, but the values is as follows:

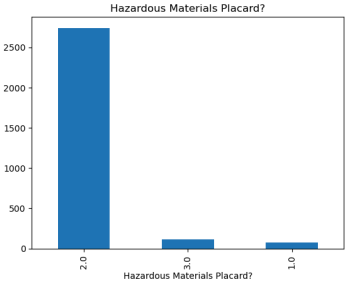
0 Blank
1 Yes
2 No
3 Unknown/Hit&Run , so most collisions didn't have any hazardous materials

Hazardous Materials Placard?

Hazardous Materials Placard?
Categorical

IMBALANCE MISSING

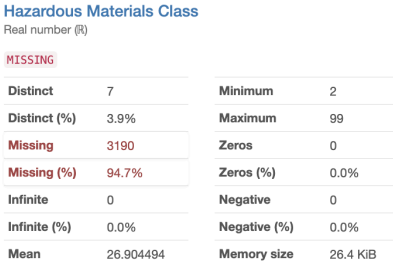
Distinct	3
Distinct (%)	0.1%
Missing	443
Missing (%)	13.2%



13.2% of the data is missing, but the value is as follows:

0 Blank
1 Yes
2 No
3 Unknown/Hit&Run , so most placards were not there

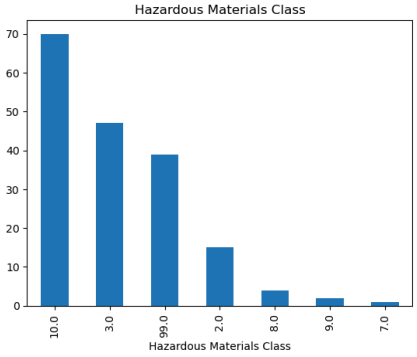
Hazardous Materials Class:



Hazardous Materials Class
Real number (ℝ)

MISSING

Distinct	7	Minimum	2
Distinct (%)	3.9%	Maximum	99
Missing	3190	Zeros	0
Missing (%)	94.7%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	26.904494	Memory size	26.4 KiB



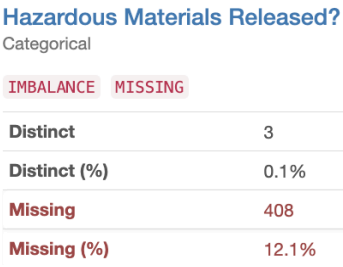
94.7% of the dataset is missing, but the values present are as follows:

00 Blank
01 Explosives
02 Gases
03 Flammable Liquids
04 Flammable Solids
05 Oxidizing Substance
06 Poison/Infectious Substance
07 Radioactive
08 Corrosives
09 Miscellaneous Goods
10 No Placard
99 Unknown/Hit&Run , so most fell under no placArd

Hazardous Materials ID:

Disregarded because they're just numbers that have no meaning in the layout

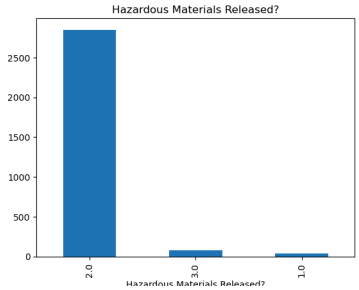
Hazardous Materials Released?



Hazardous Materials Released?
Categorical

IMBALANCE MISSING

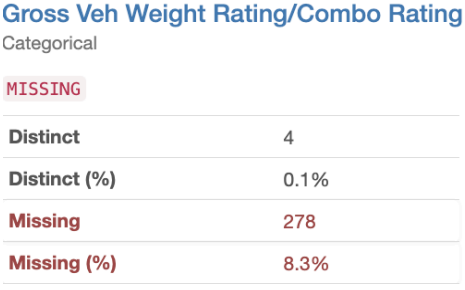
Distinct	3
Distinct (%)	0.1%
Missing	408
Missing (%)	12.1%



12.1% of the data is missing, and the values are as follows:

0 Blank
1 Yes
2 No
3 Unknown/Hit&Run , so most materials weren't released

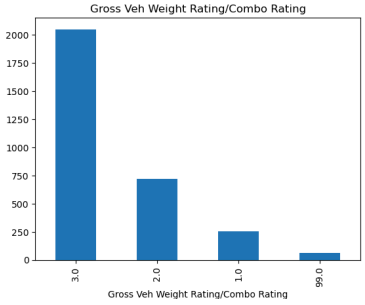
Gross Veh Weight Rating/Combo Rating:



Gross Veh Weight Rating/Combo Rating
Categorical

MISSING

Distinct	4
Distinct (%)	0.1%
Missing	278
Missing (%)	8.3%



8.3% of the dataset is missing, but the values is as follows:

01 <= 10,000 lbs
02 10,000 - 26,000 lbs
03 > 26,000 lbs
99 Unknown/Hit&Run

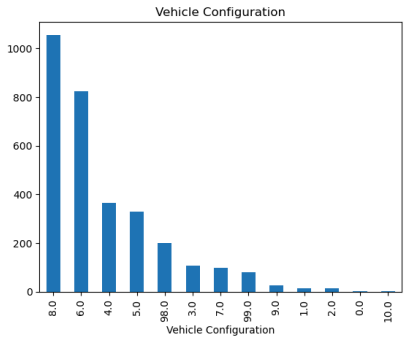
, so the most vehicles weighed greater than 26,000 lbs

Vehicle Configuration:

Vehicle Configuration
Real number (ℝ)

MISSING

Distinct	13	Minimum	0
Distinct (%)	0.4%	Maximum	99
Missing	258	Zeros	2
Missing (%)	7.7%	Zeros (%)	0.1%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	14.557556	Memory size	26.4 KiB



1.5% of the dataset is missing, but values are as follows:

00 Passenger Car w/ Hazmat
01 Light Truck w/ Hazmat
02 Bus 9-15 people
03 Bus 16+ people
04 Single Unit Truck 2 axles 6+ tires
05 Single Unit Truck 3 or more axles
06 Truck w/ Trailer
07 Truck-Tractor Only Bobtail
08 Truck w/ Semi-Trailer
09 Tractor w/ Double Trailers
10 Tractor w/ Triple Trailers
98 Other/Unable to Classify
99 Unknown/ Hit&Run

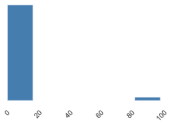
, thus most vehicles were trucks with semi trailers

Trailer Length Code 1 and 2:

Trailer Length1 Code
Real number (ℝ)

MISSING ZEROS

Distinct	6	Minimum	0
Distinct (%)	0.3%	Maximum	99
Missing	1056	Zeros	652
Missing (%)	31.4%	Zeros (%)	19.4%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	4.8512111	Memory size	26.4 KiB



More details

Trailer Length2 Code
Categorical

IMBALANCE MISSING

Distinct	5
Distinct (%)	1.1%
Missing	2924
Missing (%)	86.8%

31.4% and 86.8%, respectively of the dataset is missing, but the values are as follows:

00 No Trailer
01 Less than 480 inches
02 481 - 576 inches
03 577 inches or more
99 Unknown/Hit&Run

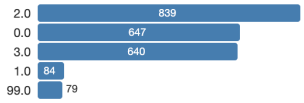
, so most didn't have a trailer

Trailer Width Code 1 and 2:

Trailer Width1 Code
Categorical

MISSING

Distinct	5
Distinct (%)	0.2%
Missing	1079
Missing (%)	32.0%
Memory size	26.4 KiB



More details

Trailer Width2 Code
Categorical

IMBALANCE MISSING

Distinct	5
Distinct (%)	1.1%
Missing	2926
Missing (%)	86.9%

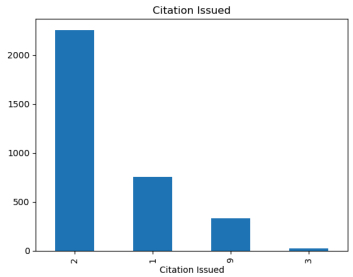
32% and 88.9% of the dataset is missing, but the values are as follows:

00 No Trailer
01 Less than 60 inches
02 61 - 84 inches
03 84 inches or more
99 Unknown/Hit&Run, so most vehicles had either a 61-84 inch wide trailer or did not have one at all

Citation Issued:

Citation Issued
Categorical

Distinct	4
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%



No missing data, but values are as follows:

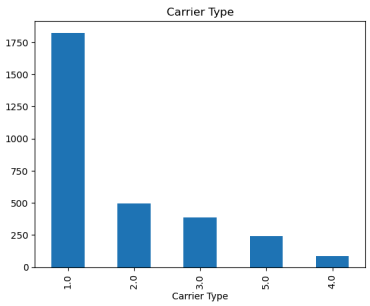
1 Yes
2 No
3 Pending
9 Unknown, so for most cases, citations weren't issued

Carrier Type:

Carrier Type
Categorical

MISSING

Distinct	5
Distinct (%)	0.2%
Missing	342
Missing (%)	10.2%



10.2% of the data is missing, but the value is as follows:

1 Interstate
2 Intrastate
3 Not in Commerce - Other Truck/Bus
4 Not in Commerce - Government
5 Other Operation/Not Specified

, so most of the carriers are found on the interstate

Dataset 4: 2021 unt:

- Overall statistics:

Dataset statistics	
Number of variables	47
Number of observations	274401
Missing cells	4184623
Missing cells (%)	32.4%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	98.4 MiB
Average record size in memory	376.0 B

Variable types	
Numeric	23
Categorical	16
Text	8

- Dataset Observations:

The dataset 2021 unt was cross-referenced with the tbd data, which contained all the full form of the acronyms (referred to as layout) for better data analysis, as seen below through unt.info():

```
---  ------                                 --------------   -----
0   Collision Number                        274401 non-null  int64
1   Unit Number                             274401 non-null  int64
2   Driver Sex                              274401 non-null  object
3   Driver Race                             274401 non-null  object
4   Drivers License State                   256583 non-null  object
5   Drivers License Class                   248753 non-null  object
6   Vehicle Make                            264078 non-null  object
7   Vehicle Registration Plate State        273140 non-null  object
8   Vehicle Registration Plate Year         260741 non-null  object
9   Contributed to Collision                274401 non-null  object
10  Speed Limit                             269419 non-null  float64
11  Citation Violation Code 1               67724 non-null   object
12  Currently Junk                          67513 non-null   object
13  Currently Junk                          8135 non-null    object
14  Direction of Travel                     274401 non-null  object
15  Unit Type                               274401 non-null  int64
16  Vehicle Use                             274401 non-null  int64
17  Vehicle Attachments                     274401 non-null  object
18  Action Prior to Impact                  274401 non-null  int64
19  Property Damage                         152876 non-null  float64
20  Towed                                   144650 non-null  float64
21  Extent of Deformity                     274401 non-null  int64
22  Most Harmful Event                      274401 non-null  int64
23  Property Damage 2                       149997 non-null  float64
24  Alcohol/Drug Information                274401 non-null  object
25  Citation Violation Code 2               8172 non-null    object
26  Truck/Bus Supplemental Form Required    274401 non-null  object
27  Manner of Collision                     148263 non-null  float64
28  Underride / Override                    274401 non-null  int64
29  Alcohol Test Given                      7493 non-null    float64
30  Drug Test Given                         5868 non-null    float64
31  Alcohol Test Type                       4605 non-null    float64
32  Drug Test Type                          1716 non-null    float64
33  Drug Test Results                       788 non-null     float64
34  Vehicle Body Type                       1016 non-null    float64
35  Sequence of Events1                     274396 non-null  float64
36  Sequence of Events2                     48898 non-null   float64
37  Sequence of Events3                     19828 non-null   float64
38  Sequence of Events4                     7049 non-null    float64
39  Estimated Collision Speed               272076 non-null  object
40  Unit Damage(in dollars)                 273717 non-null  float64
41  First Deformed Area                     274392 non-null  float64
42  Most Deformed Area                      274401 non-null  int64
43  Alcohol Test Results                    3072 non-null    float64
44  Vehicle Identification Number           258048 non-null  object
45  Number of Occupants                     274401 non-null  int64
46  CDL licensed required                   274401 non-null  object
```

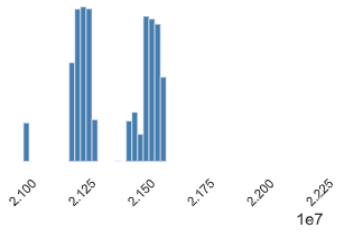
- Variable Observations:

Collision Number:

Collision Number

Real number (ℝ)

Distinct	147724	Minimum	21000003
Distinct (%)	53.8%	Maximum	22201210
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	21381302	Memory size	2.1 MiB



No missing data, but the value is as follows:

01 Vehicle	Backing
02 Vehicle	Changing Lanes
03 Vehicle	Entering Traffic Lane
04 Vehicle	Leaving Traffic Lane
05 Vehicle	Making U-Turn
06 Vehicle	Movement Essentially Straight Ahead
07 Vehicle	Overtaking/Passing
08 Vehicle	Parked
09 Vehicle	Slowing Or Stopped In Traffic
10 Vehicle	Turning Left
11 Vehicle	Turning Right
21 Non-Motorist	Approaching/Leaving Vehicle
22 Non-Motorist	Entering/Crossing Location
23 Non-Motorist	Playing/Working On Vehicle
24 Non-Motorist	Pushing Vehicle
25 Non-Motorist	Standing
26 Non-Motorist	Walking/Playing/Cycling
27 Non-Motorist	Working
88 Other/Unknown	Other
99 Other/Unknown	Unknown

Unit Number:

Disregarded because it is used to give an ID to the collision

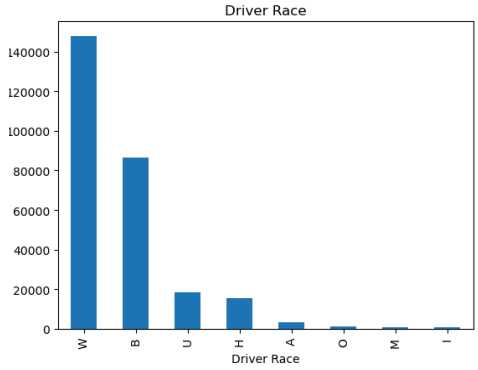
Driver Race + Sex:

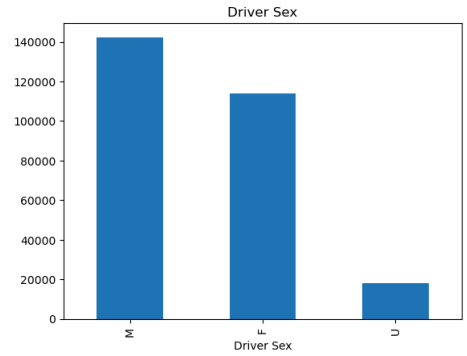
Driver Sex
Categorical

Distinct	3
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	2.1 MiB

Driver Race
Categorical

Distinct	8
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	2.1 MiB





No missing values, but values are as follows:

W	Caucasian
B	African American
O	Other
I	Alaskan Native/American Indian
A	Asian/Pacific Islander
H	Hispanic
U	Unknown

, so Most drivers are male and caucasian (M is male, F is female, and U is unknown)

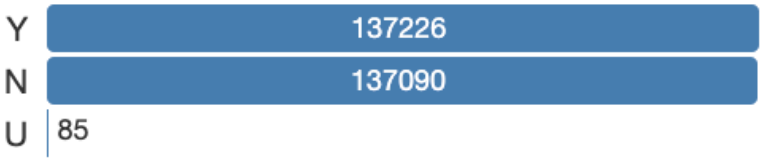
Driver License State, Class, Vehicle Make, Vehicle Registration Plate State, and Vehicle Registration Plate Year:

All are disregarded because it is not helpful in road safety as most are from South Carolina (where the dataset is from)

Contributed to Collision:

Contributed to Collision
Categorical

Distinct	3
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	2.1 MiB



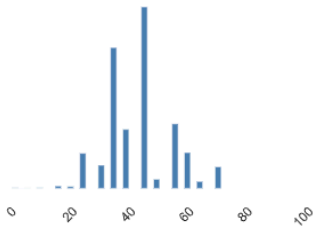
No missing data, and there is a greater number of yes than no, although they are close (Y for yes, N for no, and U and a blank for unknown)

Speed Limit:

Speed Limit
Real number (ℝ)

MISSING

Distinct	47	Minimum	0
Distinct (%)	< 0.1%	Maximum	99
Missing	4982	Zeros	221
Missing (%)	1.8%	Zeros (%)	0.1%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	43.258939	Memory size	2.1 MiB



1.8% of the dataset is missing, but most collisions occurred when the speed limit is around 45 mph

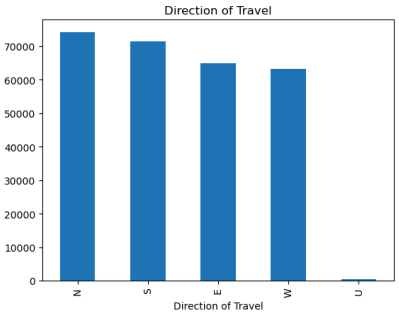
Citation Violation Code 1 and 2:

Disregarded because they don't stand for anything and have a lot of missing data

Direction of Travel:

Direction of Travel
Categorical

Distinct	5
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	2.1 MiB

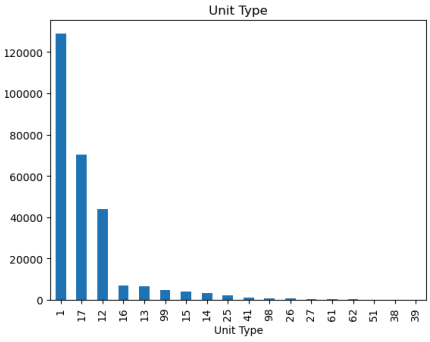


There is no missing data, and a majority of these vehicles were traveling north

Unit Type:

Unit Type
Real number (ℝ)

Distinct	18	Minimum	1
Distinct (%)	< 0.1%	Maximum	99
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	10.514346	Memory size	2.1 MiB



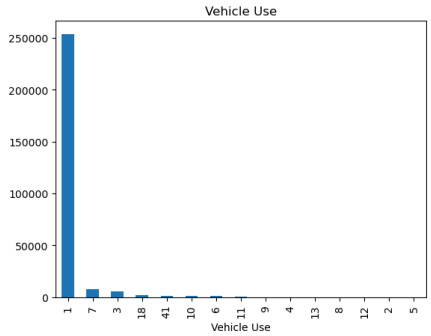
There is no missing data, and values are as such:

01 Automobile
12 Pickup Truck
13 Truck Tractor
14 Other Truck
15 Full Size Van
16 Mini Van
17 SUV
25 Motorcycle
26 Other Motorbike
27 Pedalcycle
38 Animal Drawn Vehicle
39 Animal - Ridden
41 Pedestrian
51 Train
61 School Bus
62 Passenger Bus
98 Other
99 Unknown (Hit & Run Only), so most were automobiles

Vehicle Use:

Vehicle Use
Real number (ℝ)

Distinct	15	**Minimum**	1
Distinct (%)	< 0.1%	**Maximum**	41
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1.6265757	**Memory size**	2.1 MiB



There are no missing data, and the values are such:

01 Personal
02 Driver Training
03 Construction/Maintenance
04 Ambulance
05 Military
06 Transport Passengers
07 Transport Property
08 Farm Use
09 Wrecker Or Tow
10 Police
11 Government
12 Fire Fighting
13 Logging Truck
18 Other
41 Pedestrian

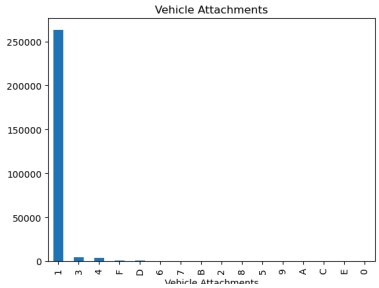
, thus most were personal vehicles

Vehicle Attachments:

Vehicle Attachments
Categorical

IMBALANCE

Distinct	16
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	2.1 MiB



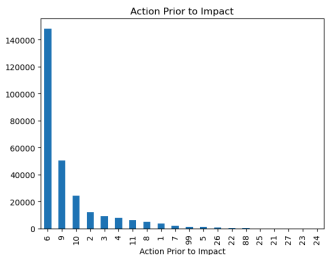
There is no missing data, and the values are as such:

0 Unknown
1 None
2 Mobile Home
3 Semi-Trailer
4 Utility Trailer
5 Farm Trailer
6 Trailer With Boat
7 Camper Trailer
8 Towed Motor Vehicle
9 Petroleum Tanker
A Lowboy Trailer
B Auto Carrier Trailer
C Other Tanker
D Flat Bed
E Twin Trailers
F Other

, so the most common vehicle attachment was nothing

Action Prior to Impact:

Distinct	20	**Minimum**	1
Distinct (%)	< 0.1%	**Maximum**	99
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	7.536693	**Memory size**	2.1 MiB



No missing data, and the values are as such:

1 Given/Known Results
2 Given/Unusable
3 Given/Pending
4 None Given
5 Refused, thus most common was that none were given

Property Damage 1 and 2:

Disregarded because most of the dataset is missing, and there is no value associated in the layout

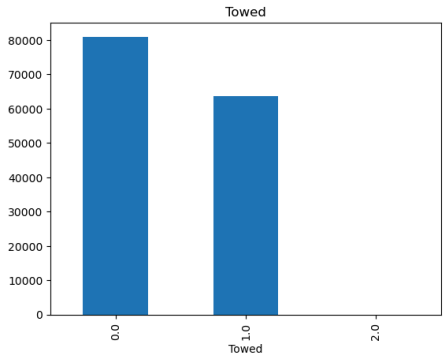
Towed:

Towed

Categorical

MISSING

Distinct	3
Distinct (%)	< 0.1%
Missing	129751
Missing (%)	47.3%



47.3% of the data is missing. The values are 1 for towed and 2 for not towed. This part cannot be analyzed because there is no value for 0

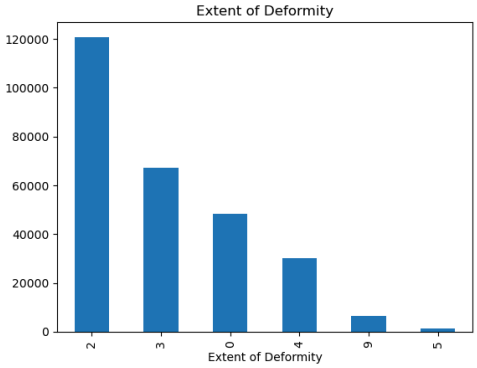
Extent of Deformity:

Extent of Deformity

Real number (ℝ)

ZEROS

Distinct	6	Minimum	0
Distinct (%)	< 0.1%	Maximum	9
Missing	0	Zeros	48429
Missing (%)	0.0%	Zeros (%)	17.6%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	2.2916753	Memory size	2.1 MiB



No missing data, and the values are as follows:

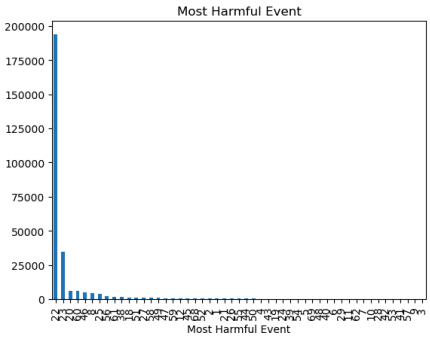
0 None/Minor
2 Functional Damage
3 Disabling Damage
4 Severe/Totaled
5 Not Applicable
9 Unknown, so the most common was functional damage of the car

Most Harmful Event:

Most Harmful Event

Real number (ℝ)

Distinct	51	**Minimum**	1
Distinct (%)	< 0.1%	**Maximum**	69
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	24.607636	**Memory size**	2.1 MiB



No missing data, and values are as follows:

00 None Listed	None Listed
01 Non-Collision	Cargo/Equip Loss Or Shift
02 Non-Collision	Cross Median/Center Line
03 Non-Collision	Downhill Runaway
04 Non-Collision	Equipment Failure
05 Non-Collision	Fire/Explosion
06 Non-Collision	Immersion
07 Non-Collision	Jackknife
08 Non-Collision	Overturn/Rollover
09 Non-Collision	Run Off Road Left
10 Non-Collision	Run Off Road Right
11 Non-Collision	Separation Of Units
12 Non-Collision	Spill (Two Wheel Vehicle)
18 Non-Collision	Other Non-Collision
19 Non-Collision	Unknown Non-Collision
20 Collision: Object Not Fixed	Animal (Deer Only)
21 Collision: Object Not Fixed	Animal (Not Deer)
22 Collision: Object Not Fixed	Motor Vehicle (In Transport)
23 Collision: Object Not Fixed	Motor Vehicle (Stopped)
24 Collision: Object Not Fixed	Motor Vehicle (Other Roadway)
25 Collision: Object Not Fixed	Motor Vehicle (Parked)
26 Collision: Object Not Fixed	Pedalcycle
27 Collision: Object Not Fixed	Pedestrian
28 Collision: Object Not Fixed	Railway Vehicle
29 Collision: Object Not Fixed	Work Zone Maint. Equip.
38 Collision: Object Not Fixed	Other Movable Object
39 Collision: Object Not Fixed	Unknown Movable Object
40 Collision: Object Fixed	Bridge Overhead Structure
41 Collision: Object Fixed	Bridge Parapet End
42 Collision: Object Fixed	Bridge Pier Or Abutment
43 Collision: Object Fixed	Bridge Rail
44 Collision: Object Fixed	Culvert
45 Collision: Object Fixed	Curb
46 Collision: Object Fixed	Ditch
47 Collision: Object Fixed	Embankment
48 Collision: Object Fixed	Equipment
49 Collision: Object Fixed	Fence
50 Collision: Object Fixed	Guardrail End
51 Collision: Object Fixed	Guardrail Face
52 Collision: Object Fixed	HWY Traffic Sign Post
53 Collision: Object Fixed	Impact Attenuator/Crash Cushion
54 Collision: Object Fixed	Light Luminaire Support
55 Collision: Object Fixed	Mailbox
56 Collision: Object Fixed	Median Barrier
57 Collision: Object Fixed	Overhead Sign Support
58 Collision: Object Fixed	Other(Post,Pole,Support,Etc.)
59 Collision: Object Fixed	Other(Wall,Bldg,Tunnel,Etc.)
60 Collision: Object Fixed	Tree
61 Collision: Object Fixed	Utility Pole
62 Collision: Object Fixed	Workzone Maint. Equip.
68 Collision: Object Fixed	Other
69 Collision: Object Fixed	Unknown Fixed Object
99 Collision: Object Fixed	Unknown

, so most were collisions with vehicle in transit

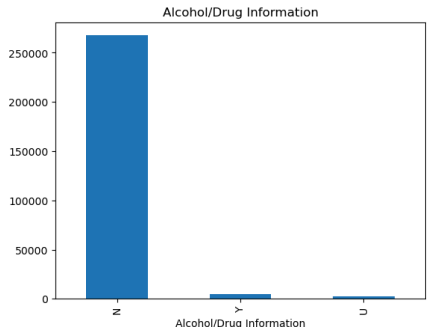
Alcohol/Drug Information:

Alcohol/Drug Information

Categorical

`IMBALANCE`

Distinct	3
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	2.1 MiB



No missing data, so most were not tested for alcohol or drugs (it has Y for yes, N for No, and U and blank for unknown)

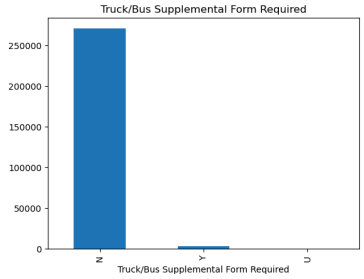
Truck/Bus Supplemental Form Required:

Truck/Bus Supplemental Form Required

Categorical

`IMBALANCE`

Distinct	3
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%



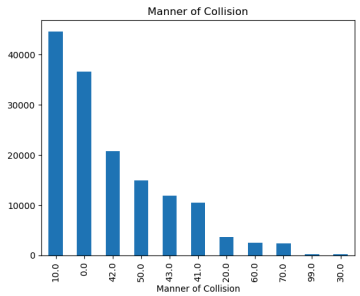
No missing data, so the form was not required for most of them (it has Y for yes, N for No, and U and blank for unknown)

Manner of Collision:

Manner of Collision
Real number (ℝ)

MISSING ZEROS

Distinct	11	Minimum	0
Distinct (%)	< 0.1%	Maximum	99
Missing	126138	Zeros	36633
Missing (%)	46.0%	Zeros (%)	13.4%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	23.110452	Memory size	2.1 MiB



46% of the data is missing, but the values are as such:

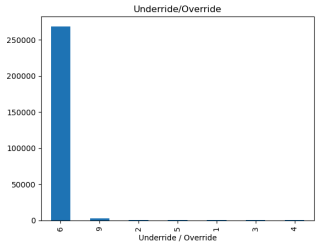
00 Not Collision With Motor Vehicle
10 Rear End
20 Head On
30 Rear To Rear
41 Angle
42 Angle
43 Angle
50 Sideswipe Same Direction
60 Sideswipe Opposite Direction
70 Backed Into
99 Unknown

, so most are rear ends

Underride/Override:

Underride / Override
Real number (ℝ)

Distinct	7	Minimum	1
Distinct (%)	< 0.1%	Maximum	9
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	5.9969752	Memory size	2.1 MiB



No missing data, and the values are as follows:

1 Under-Compartment Intrusion
2 Under-No Intrusion
3 Under-Unknown
4 Over-Motor Vehicle In Transport
5 Over-Other Vehicle
6 None
9 Unknown

, so in most cases, there was no underride or override

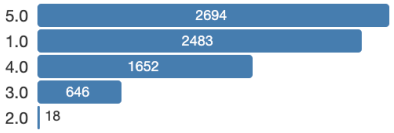
Alcohol Test Given, Results, and Type:

G:

Alcohol Test Given
Categorical

MISSING

Distinct	5
Distinct (%)	0.1%
Missing	266908
Missing (%)	97.3%
Memory size	2.1 MiB

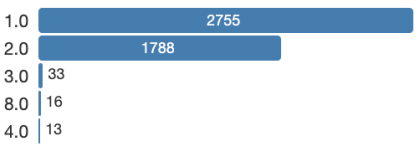


T:

Alcohol Test Type
Categorical

IMBALANCE MISSING

Distinct	5
Distinct (%)	0.1%
Missing	269796
Missing (%)	98.3%



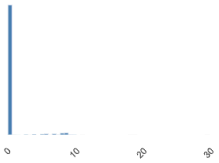
R:

Alcohol Test Results

Real number (ℝ)

MISSING

Distinct	83	Minimum	0
Distinct (%)	2.7%	Maximum	30
Missing	271329	Zeros	1159
Missing (%)	98.9%	Zeros (%)	0.4%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	0.54385091	Memory size	2.1 MiB



A lot of missing data, but the values are as such:

	1 Given/Known Results
	2 Given/Unusable
	3 Given/Pending
	4 None Given
Alcohol Test Given	5 Refused
Alcohol Test Results	
	1 Breath (Alc Only)
	2 Blood
	3 Urine
	4 Serum
Alcohol Test Type	8 Other

, so most times, the person refused, but when they didn't, only breath was checked

<u>Drug Test Given, Results, and Type:</u>

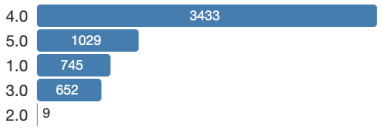
G:

Drug Test Given

Categorical

MISSING

Distinct	5
Distinct (%)	0.1%
Missing	268533
Missing (%)	97.9%
Memory size	2.1 MiB



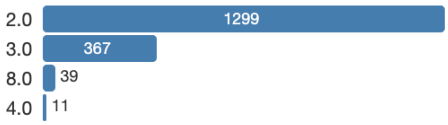
T:

Drug Test Type

Categorical

IMBALANCE MISSING

Distinct	4
Distinct (%)	0.2%
Missing	272685
Missing (%)	99.4%



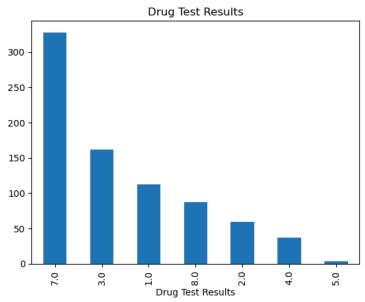
R:

Drug Test Results

Real number (ℝ)

MISSING

Distinct	7	Minimum	1
Distinct (%)	0.9%	Maximum	8
Missing	273613	Zeros	0
Missing (%)	99.7%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	4.9124365	Memory size	2.1 MiB



A lot of missing data, but the values are as such:

Drug Test Given	1 Given/Known Results 2 Given/Unusable 3 Given/Pending 4 None Given 5 Refused
Drug Test Results	1 Amphetamines 2 Cocaine 3 Marijuana 4 Opiates 5 PCP 7 None 8 Other *blank* No Test Given or Negative Result
Drug Test Type	1 Breath (Alcohol Only) 2 Blood 3 Urine 4 Serum 8 Other

, so mostly, no test was given, but when it was, it was mostly a blood test, and found no drugs

Vehicle Body Type:

Disregarded because it cannot be analyzed as there is too much missing data, and doesn't have a key

Sequence of Events 1, 2, 3, and 4:

Disregarded because it does not add to the data analysis

Estimated Collision Speed:

Estimated Collision Speed
Text

Distinct	100
Distinct (%)	< 0.1%
Missing	2325
Missing (%)	0.8%
Memory size	2.1 MiB

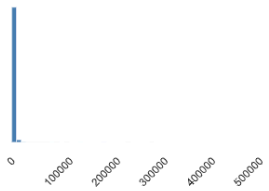
1.2% of the data is missing, but the most common collision speed was around 45 mph

Unit Damage (in dollars):

Unit Damage(in dollars)
Real number (ℝ)

SKEWED ZEROS

Distinct	567	Minimum	0
Distinct (%)	0.2%	Maximum	500100
Missing	684	Zeros	9816
Missing (%)	0.2%	Zeros (%)	3.6%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	2929.0079	Memory size	2.1 MiB



No missing data, and the most were under $1000

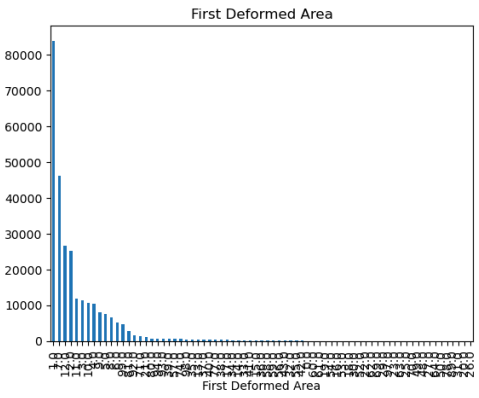
First and Most Deformed Areas:

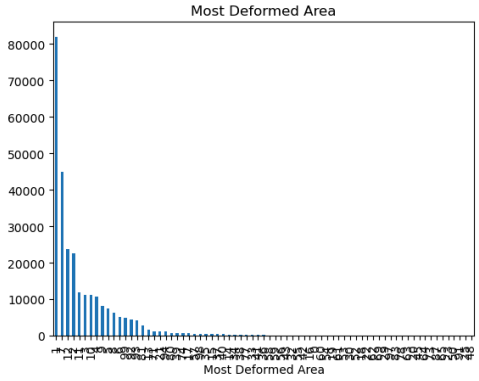
First Deformed Area
Real number (ℝ)

Distinct	73	Minimum	0
Distinct (%)	< 0.1%	Maximum	99
Missing	9	Zeros	16
Missing (%)	< 0.1%	Zeros (%)	< 0.1%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	9.9954736	Memory size	2.1 MiB

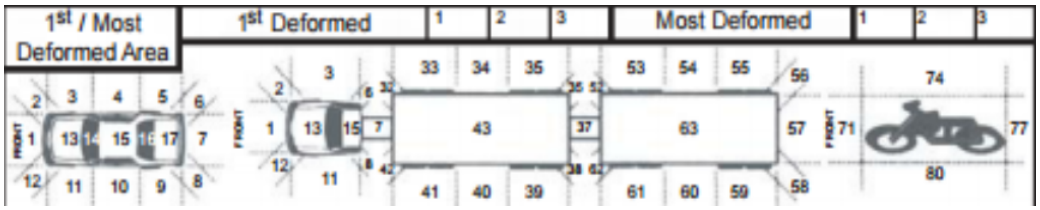
Most Deformed Area
Real number (ℝ)

Distinct	73	Minimum	0
Distinct (%)	< 0.1%	Maximum	99
Missing	0	Zeros	16
Missing (%)	0.0%	Zeros (%)	< 0.1%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	12.17997	Memory size	2.1 MiB





No missing data, and the values are as such:



, so mostly the front

Vehicle Identification Number:

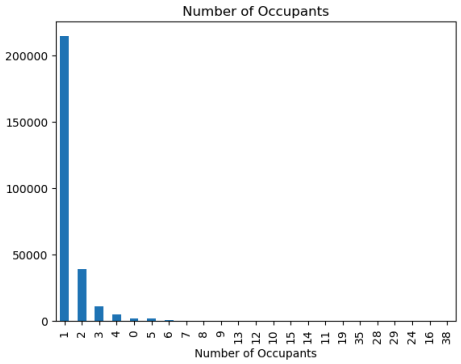
Disregarded because it only provides vehicle identity

Number of Occupants:

Number of Occupants

Real number (ℝ)

Distinct	23	**Minimum**	0
Distinct (%)	< 0.1%	**Maximum**	38
Missing	0	**Zeros**	1942
Missing (%)	0.0%	**Zeros (%)**	0.7%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1.3123822	**Memory size**	2.1 MiB



No missing data, so most collisions were with less than 5 occupants

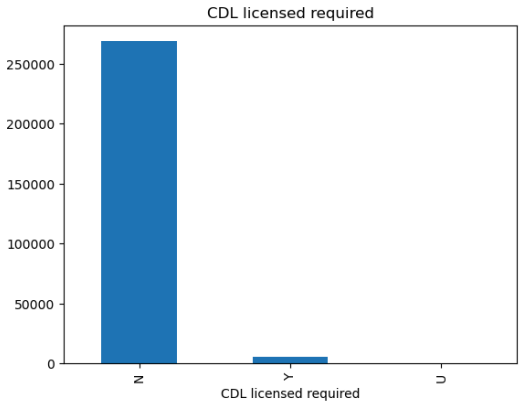
CDL License Required:

CDL licensed required

Categorical

IMBALANCE

Distinct	3
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%



No missing data, so there was no license required (Y for yes, N for no, and U and a blank for unknown)