2018 Data Analysis Report

This report is a summary of the pandas profiling done on the 2018 traffic, including all four datasets, acronymed as 2018 loc, 2018 occ, 2018 tbd, and 2018 unt.

Dataset 1: 2018 loc:

- Overall statistics:

Overview | Alerts 40 | Reproduction

Dataset statistics

Number of variables	57
Number of observations	142406
Missing cells	1837557
Missing cells (%)	22.6%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	61.9 MiB
Average record size in memory	456.0 B

Variable types

Numeric	36
Categorical	14
Text	6
Boolean	1

- Dataset Observations:

The dataset 2018 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. However, there were some missing full forms for the dataset and its data types, as seen below through loc.info():

```
 #   Column                                 Non-Null Count   Dtype
---  ------                                 --------------   -----
 0   Collision Number                       142406 non-null  int64
 1   Trafficway                             142406 non-null  int64
 2   Type of Record (L-Location data)       142406 non-null  object
 3   Day of Week                            142406 non-null  int64
 4   On Route Auxiliary                     141569 non-null  float64
 5   Lane Number of Collision               120403 non-null  float64
 6   Base Distance Direction                142406 non-null  object
 7   Light Condition                        142406 non-null  int64
 8   Weather Condition                      142406 non-null  int64
 9   Road Condition                         142406 non-null  int64
 10  Road Surface Conition                  142406 non-null  int64
 11  First Harmful Event Location           142406 non-null  int64
 12  Crosswalk Indicator                    142406 non-null  int64
 13  Date of Collision                      142406 non-null  float64
 14  On Route Street Name                   125730 non-null  object
 15  Investigating Jurisdiction Code        142406 non-null  object
 16  Number of Fatalities                   142406 non-null  float64
 17  Number of Non-Fatal Injuries           142406 non-null  float64
 18  Amended or Corrected Indicator         13144 non-null   object
 19  Base Intersection Route Category       142406 non-null  int64
 20  Base Intersection Route Auxiliary      140690 non-null  float64
 21  Second Intersection Route Category     142406 non-null  int64
 22  Second Intersection Route Auxiliary    139794 non-null  float64
 23  Base Intersection Street Name          133443 non-null  object
 24  alss                                   134946 non-null  object
 25  hzd                                    115 non-null     float64
 26  Number of Buses                        235 non-null     float64
 27  Number of Persons Transported Immediately 1648 non-null  float64
 28  Number of Towed Units                  3979 non-null    float64
 29  Latitude of Collision (special format) 142406 non-null  int64
 30  Longitude of Collision (special format) 142406 non-null int64
 31  Junction Type                          142406 non-null  int64
 32  Other Contributing Factor 1            28955 non-null   float64
```

```
 33  Other Contributing Factor 2                                   5455 non-null    float64
 34  Other Contributing Factor 3                                   680 non-null     float64
 35  Other Contributing Factor 4                                   77 non-null      float64
 36  School Bus Involved                                           142406 non-null  int64
 37  Work Zone Indicator                                           142406 non-null  int64
 38  Work Zone Type                                                3027 non-null    float64
 39  Work Zone Location                                            3031 non-null    float64
 40  Workers Present Indicator                                     3078 non-null    float64
 41  Currently Junk field, was Badge Number of Investigating Officer  142393 non-null  object
 42  Traffic Control Type                                          142406 non-null  int64
 43  Number of Units (Vehicle and Non-Motorists)                   142406 non-null  int64
 44  County of Collision                                           142406 non-null  int64
 45  rtn                                                           130551 non-null  float64
 46  brn                                                           98242 non-null   float64
 47  srn                                                           83901 non-null   float64
 48  First Harmful Event                                           142406 non-null  int64
 49  Primary Contributing Factor                                   142406 non-null  int64
 50  tim                                                           142406 non-null  int64
 51  bdo                                                           142406 non-null  object
 52  pnt                                                           142406 non-null  int64
 53  pat                                                           142406 non-null  int64
 54  On Route Category                                             142406 non-null  int64
 55  Alcohol/Drug Involved Driver in Collision                     142406 non-null  object
 56  Direction of Lane                                             125101 non-null  object
dtypes: float64(21), int64(25), object(11)
memory usage: 61.9+ MB
```

Here, out of 57 variables, 9 variables are not in the 2018 loc layout. However, the layout does have 57 variable acronyms and names. This can lead to the conclusion that some of the variable names and acronyms are not used in the actual 2018 loc dataset that are present in the loc layout.

- Variable Observations:

Collision Number:

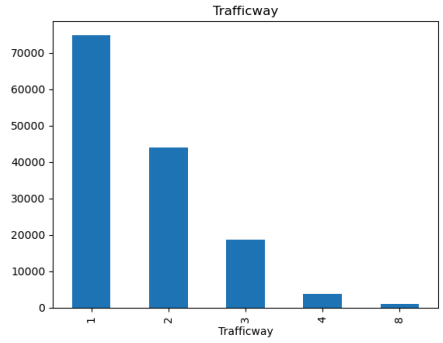
Disregarded because it is used to give an ID to every collision, and therefore cannot be analyzed (or show anything that can support any conclusions)

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

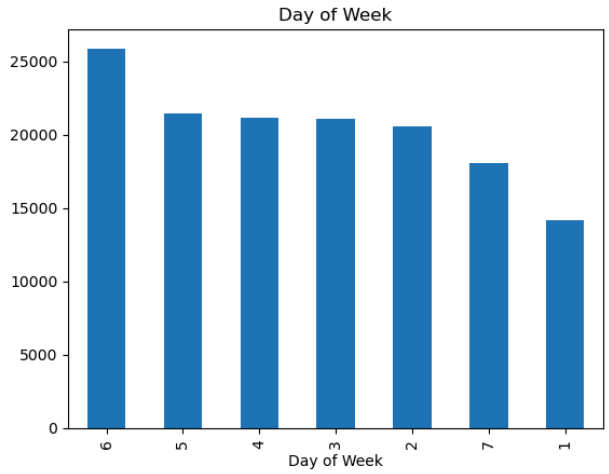
1 Two-Way, Not Divided
2 Two-Way,Divided,Unprotected Median
3 Two-Way,Divided,Barrier
4 One-Way
8 Other

, 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

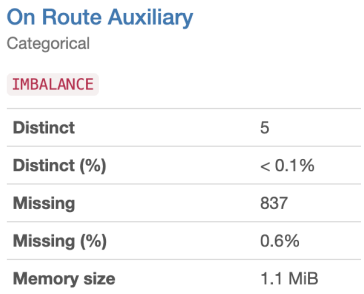
Day of the Week:

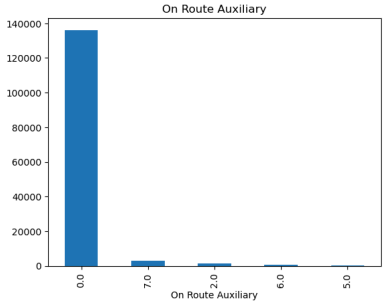


There are no missing values, and the numbers 1-7 stand for the days of the week (shown below)

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):





There are some missing values (0.6%) from the data, which limit accurate data analysis. The numerical values have translational value, as indicated below:

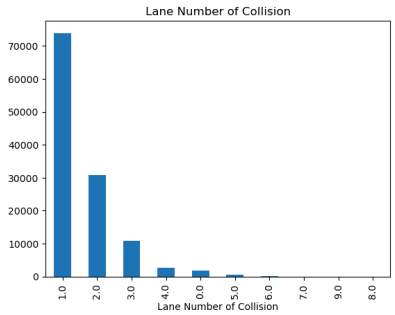
0 Main
2 Alternate
5 Spur
6 Connection
7 Business
9 Other . Here, a predominant amount of collisions occurred on main routes, which makes sense, because they would be the busiest.

Lane Number of Collision:

Lane Number of Collision
Real number (ℝ)

MISSING ZEROS

Distinct	10	Minimum	0
Distinct (%)	< 0.1%	Maximum	9
Missing	22003	Zeros	1749
Missing (%)	15.5%	Zeros (%)	1.2%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	1.5043147	Memory size	1.1 MiB

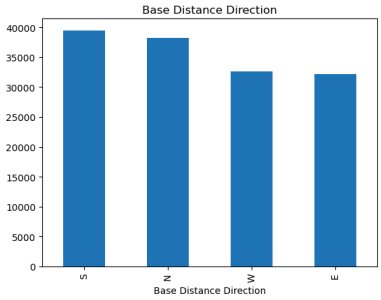


Here, there are a lot of missing values (15.5%). There is no key in the loc layout, so the lane values are up to interpretation and must be standardized

Base Distance Direction:

Base Distance Direction
Categorical

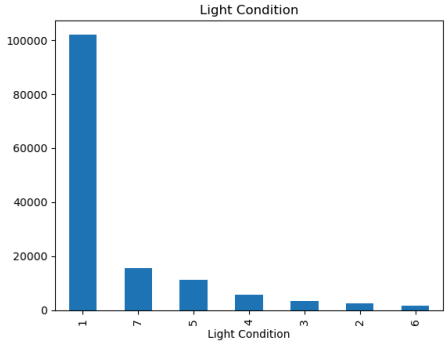
Distinct	4
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	1.1 MiB



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:

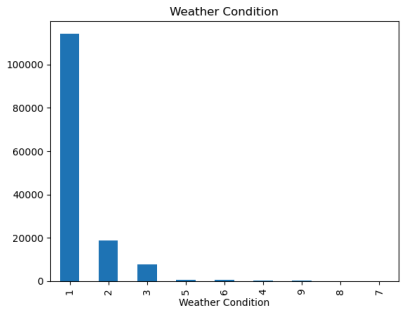


There are no missing values. The values of 1-7 are associated with weather conditions, as shown below:

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 Condition:

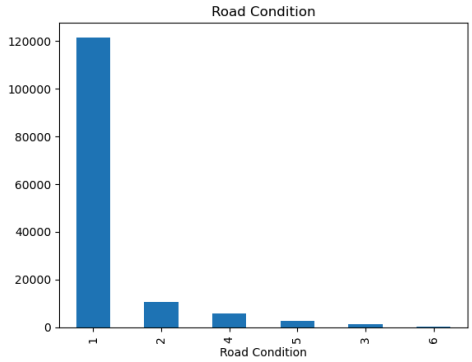


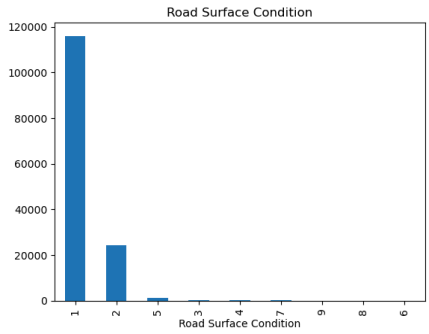
There are no missing values. The values 1-9 signify weather conditions as below:

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

. Most collisions occur in clear conditions

Road Condition + Road Surface Condition:





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

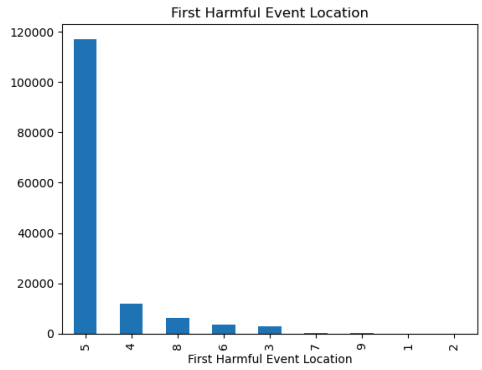
1 Straight-Level
2 Straight-On Grade
3 Straight-Hillcrest
4 Curve-Level
5 Curve-On Grade
6 Curve-Hillcrest

, 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

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

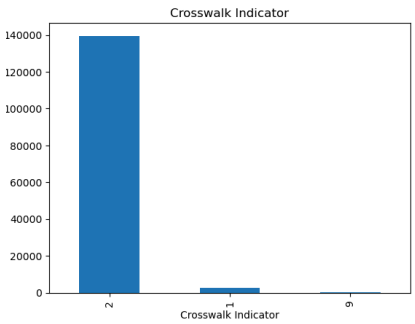
. The most often ones have happened on the roadway itself

Crosswalk Indicator:

Crosswalk Indicator
Categorical

IMBALANCE

Distinct	3
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	1.1 MiB



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:

On Route Street Name
Text

MISSING

Distinct	17213
Distinct (%)	13.7%
Missing	16676
Missing (%)	11.7%
Memory size	1.1 MiB

There is a good portion of the data missing (11.7%). Also, the word tag cloud cannot be used for analysis because it does not specify the location (city-wise), so street names can have repeats all throughout the state.

Investigating Jurisdiction Code:

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

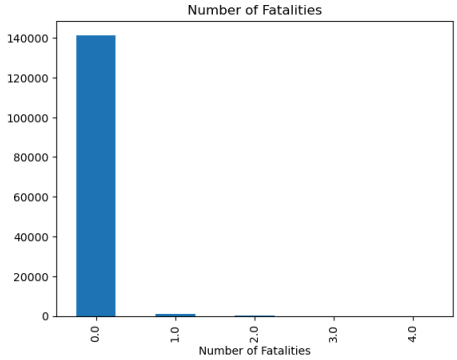
Number of Fatalities:

Number of Fatalities

Categorical

`IMBALANCE`

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



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

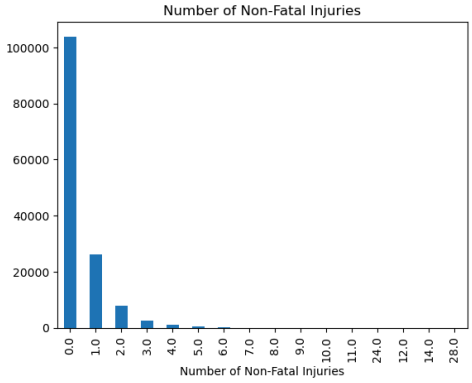
Number of Non-Fatal Injuries:

Number of Non-Fatal Injuries

Real number (ℝ)

`ZEROS`

Distinct	16	**Minimum**	0
Distinct (%)	< 0.1%	**Maximum**	28
Missing	0	**Zeros**	103900
Missing (%)	0.0%	**Zeros (%)**	73.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	0.40765839	**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.

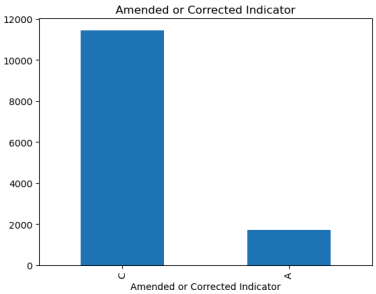
Amended or Corrected Indicator:

Amended or Corrected Indicator

Categorical

`MISSING`

Distinct	2
Distinct (%)	< 0.1%
Missing	129262
Missing (%)	90.8%
Memory size	1.1 MiB



Most of the data is missing (90.8%). However, the data that is available, most of the indicators are C (corrected) while very little are A (amended)

Base/Secondary Route Category/Auxiliary:

These four variables are disregarded because they don't add too much to the dataset or its understanding because they are just indicative of route selection (two of the datasets are missing values as well)

Base Intersection Street Name:

Base Intersection Street Name

Text

MISSING

Distinct	34762
Distinct (%)	26.1%
Missing	8963
Missing (%)	6.3%
Memory size	1.1 MiB

6.3% of the data is missing, as well as the word tag cloud does not help identifying the specific streets/areas where most collisions occur as there are many repeats in street names throughout the state

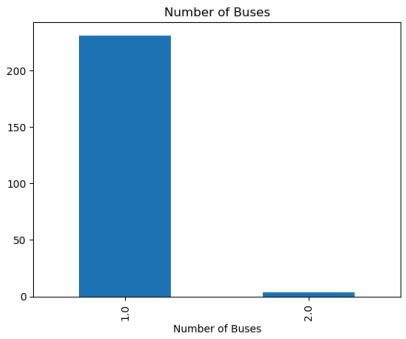
Number of Buses:

Number of Buses

Categorical

IMBALANCE MISSING

Distinct	2
Distinct (%)	0.9%
Missing	142171
Missing (%)	99.8%
Memory size	1.1 MiB



Most of the dataset (99.8%) is missing, but the data that is available indicates that mostly one bus is involved in a bus-related collision

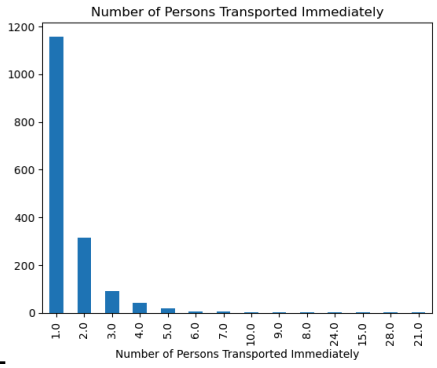
Number of Persons Transported Immediately:

Number of Persons Transported Immediately

Real number (ℝ)

MISSING

Distinct	14	Minimum	1
Distinct (%)	0.8%	Maximum	28
Missing	140758	Zeros	0
Missing (%)	98.8%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	1.5442961	Memory size	1.1 MiB



Most of the dataset here is missing (98.8%), but when the data is available, anywhere from 0-5 people are transported immediately

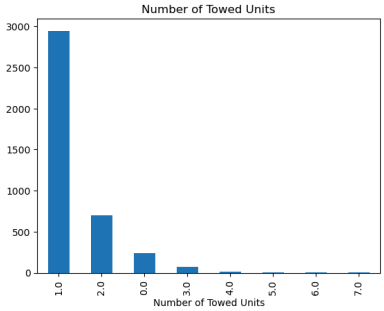
Number of Towed Units:

Number of Towed Units

Real number (ℝ)

MISSING

Distinct	8	**Minimum**	0
Distinct (%)	0.2%	**Maximum**	7
Missing	138427	**Zeros**	239
Missing (%)	97.2%	**Zeros (%)**	0.2%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	1.1711485	**Memory size**	1.1 MiB



Most of the dataset (97.2%) is missing, but when the data is available, mostly 1 vehicle is towed

Latitude + Longitude of Collision:

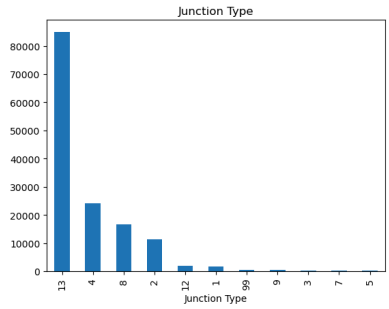
Disregarded because it provides no analytical data, just exact location, which can be used, but not for exploratory purposes.

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	10.112228	**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 railway crossings, shared paths, intersections, traffic circles, y-intersections, and non-junctions

Other Contributing Factors 1, 2, 3, 4:

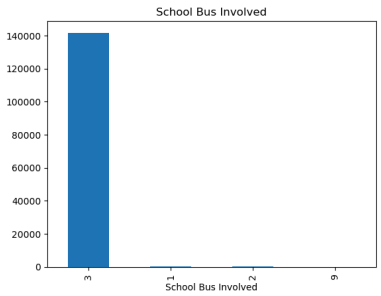
Disregarded because their numerical value doesn't have any meaning in the layout that is useful for data analysis

School Bus Involved:

School Bus Involved
Categorical

IMBALANCE

Distinct	4
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	1.1 MiB



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

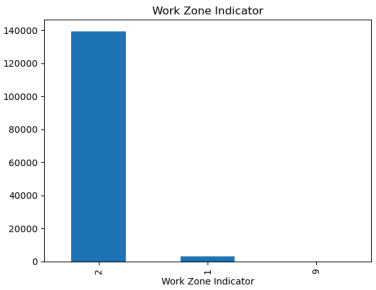
, and the data shows that most collisions occur without a school bus

Work Zone Indicator:

Work Zone Indicator
Categorical

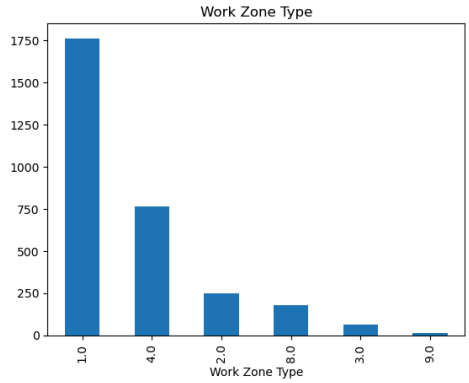
IMBALANCE

Distinct	3
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	1.1 MiB



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)

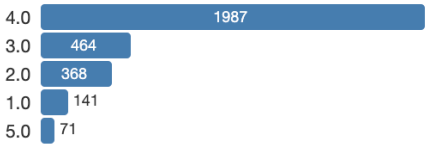
Work Zone Type and Location:



Work Zone Location
Categorical

MISSING

Distinct	5
Distinct (%)	0.2%
Missing	139375
Missing (%)	97.9%
Memory size	1.1 MiB



Both datasets have mostly 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

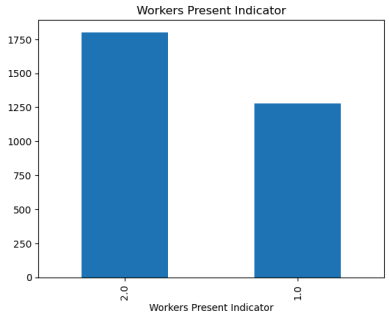
Most collisions occur in shoulder active worker areas

Workers Present Indicator:

Workers Present Indicator
Categorical

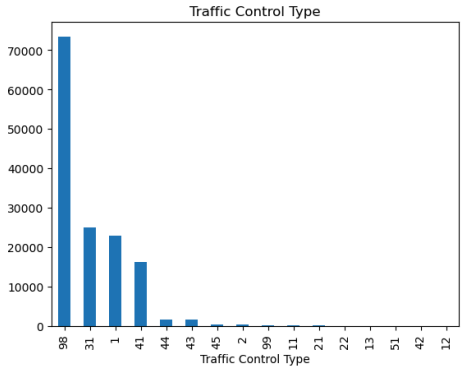
MISSING

Distinct	2
Distinct (%)	0.1%
Missing	139328
Missing (%)	97.8%
Memory size	1.1 MiB



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

Traffic Control Type:



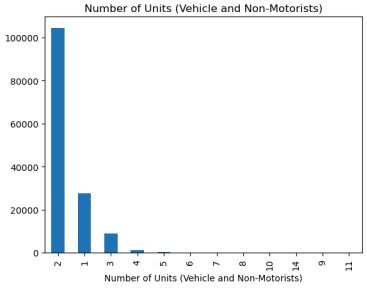
There is no missing data, and the values, according to the layout mean:

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 an unknown traffic control type

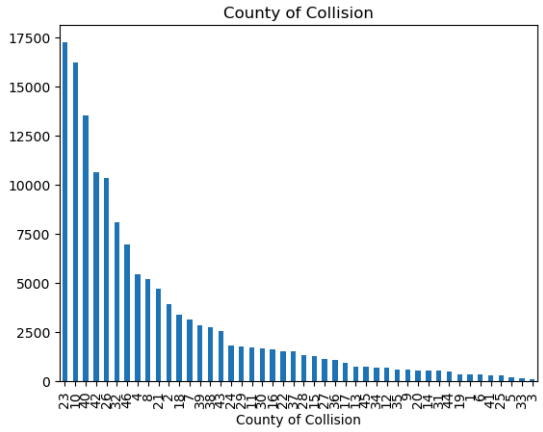
Number of Units (Vehicle and Non-Motorists):

Distinct	12	**Minimum**	1
Distinct (%)	< 0.1%	**Maximum**	14
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
hide	1.894246	**Memory size**	1.1 MiB



There are no missing values, but mostly there are 1-2 vehicle collisions, as shown in the graph

County of Collision:

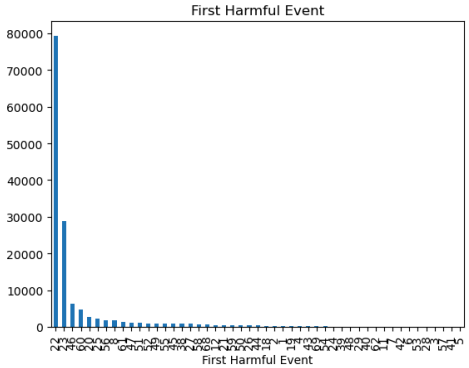


There are no missing values, and the values from 1-46 can be explained by the key in the layout:

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:

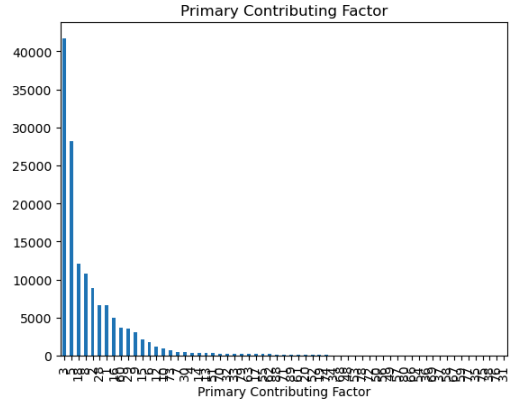


No missing data, and the values 1-99 correspond with the following:

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 movement

Primary Contributing Factor:

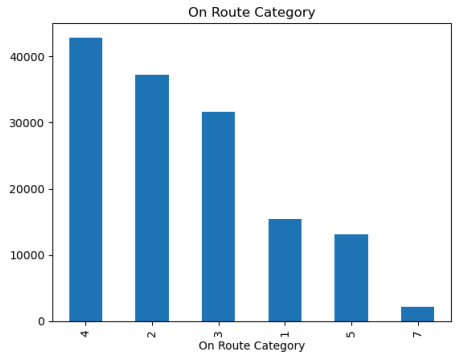


No missing data, and the values 1-89 correspond with:

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 speeding

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 routes

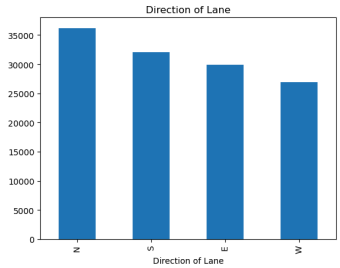
Alcohol/Drug Involved Driver in Collision:

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	139.2 KiB

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

Direction of Lane:

Distinct	4
Distinct (%)	< 0.1%
Missing	17305
Missing (%)	12.2%
Memory size	1.1 MiB



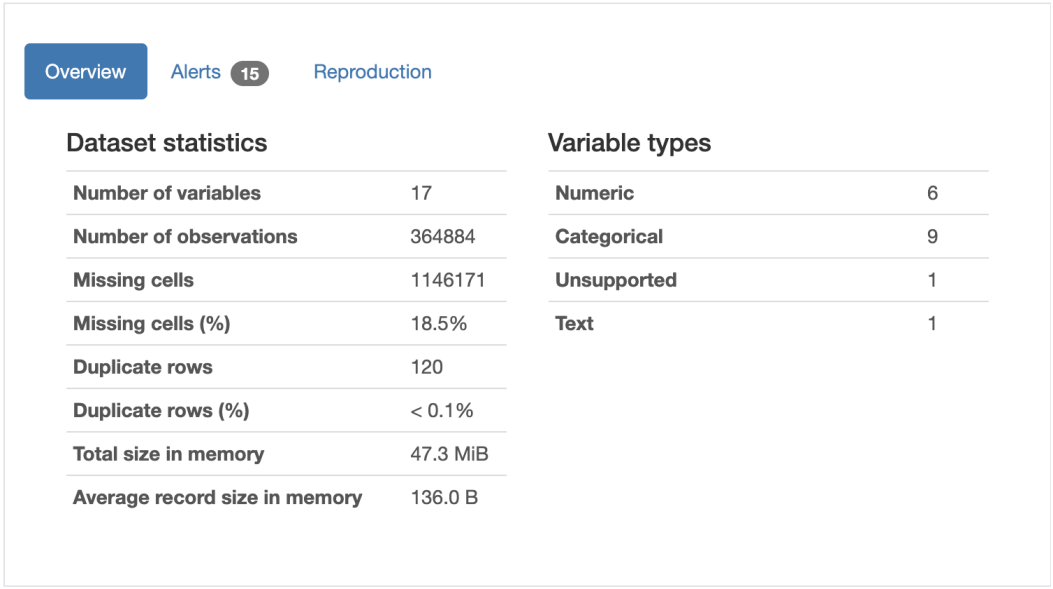
12.2% of the data is missing, but most of the collision lanes were facing north.

Variables not explored:

Hzd, alss, rtn, brn, and srn because there is no knowledge of what they mean in the loc layout, and thus cannot be analyzed.

Dataset 2: 2018 occ:

- Overall statistics:



Overview | Alerts 15 | Reproduction

Dataset statistics

Number of variables	17
Number of observations	364884
Missing cells	1146171
Missing cells (%)	18.5%
Duplicate rows	120
Duplicate rows (%)	< 0.1%
Total size in memory	47.3 MiB
Average record size in memory	136.0 B

Variable types

Numeric	6
Categorical	9
Unsupported	1
Text	1

- Dataset Observations:

The dataset 2018 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():

```
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   Collision Number                364884 non-null  int64
 1   Unit Number                     364884 non-null  int64
 2   Person Seating Location         364884 non-null  int64
 3   Record Type                     364884 non-null  object
 4   Currently Junk Variable         0 non-null       float64
 5   Person Gender                   364884 non-null  object
k to scroll output; double click to hide
 6   Person Ejection                 364884 non-null  object
 7   Person Age                      346529 non-null  float64
 8   Restraint/Safety Device         364884 non-null  int64
 9   Location After Impact           364884 non-null  int64
 10  Injury Status                   364864 non-null  float64
 11  Motorcycle Head Injury          7890 non-null    float64
 12  Ejection Status                 364884 non-null  int64
 13  Transported to Medical Facility? 364884 non-null  int64
 14  Transport by whom               49126 non-null   float64
 15  Air Bag Deployment              364884 non-null  int64
 16  Person Zip Code                 274724 non-null  object
dtypes: float64(5), int64(8), object(4)
memory usage: 47.3+ MB
```

- Variable Observations:

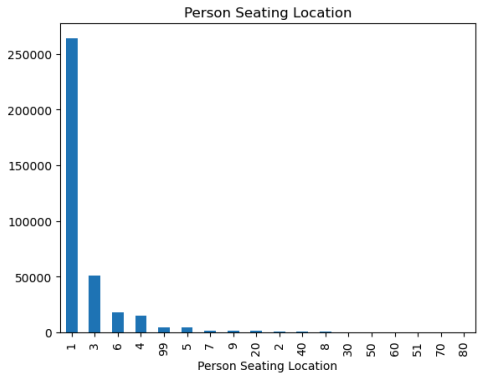
<u>Collision Number:</u>

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

Person Seating Location:



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:

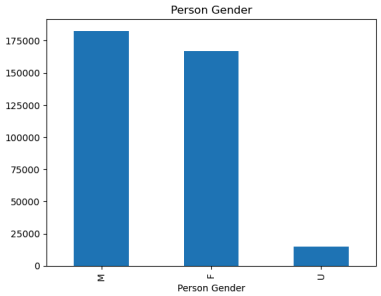
Disregarded because it has a constant value of O, which, according to the 2018 occ raw data layout, stands for a person. Therefore, its constant value indicates accident data only where a person is involved

Current Junk Variable:

Disregarded because it was unsupported by pandas profiling, thus was rejected + was also the column that had no data (a Nan type)

Person Gender:

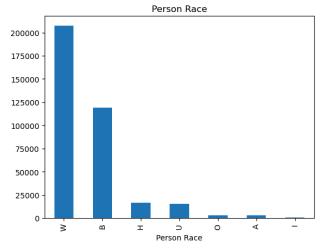
Distinct	3
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	2.8 MiB



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

Person Race:

Distinct	7
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	2.8 MiB



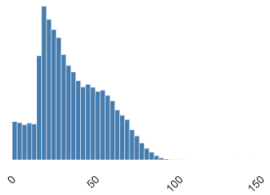
No missing data, and the values are as such:

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:

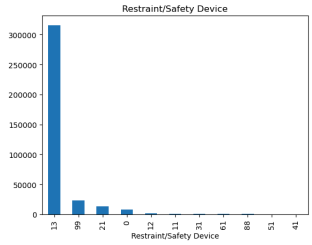
Distinct	110	**Minimum**	0
Distinct (%)	< 0.1%	**Maximum**	150
Missing	18355	**Zeros**	2205
Missing (%)	5.0%	**Zeros (%)**	0.6%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	36.553674	**Memory size**	2.8 MiB



5% of the data is missing, but most collisions occur in ages below the age of 50

Restraint/Safety Device:

Distinct	11	**Minimum**	0
Distinct (%)	< 0.1%	**Maximum**	99
Missing	0	**Zeros**	8373
Missing (%)	0.0%	**Zeros (%)**	2.3%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	18.636594	**Memory size**	2.8 MiB



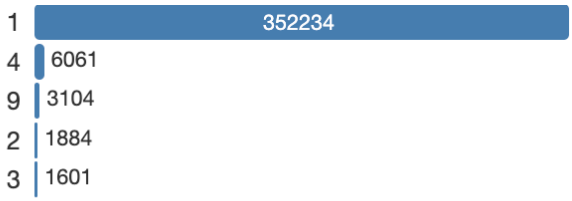
There is no missing data, and the values are as such:

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:

istinct	5
istinct (%)	< 0.1%
lissing	0
lissing (%)	0.0%
lemory size	2.8 MiB



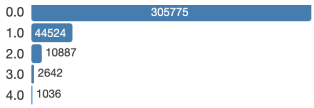
No missing data, and the values are as such:

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:

Distinct	5
Distinct (%)	< 0.1%
Missing	20
Missing (%)	< 0.1%
Memory size	2.8 MiB



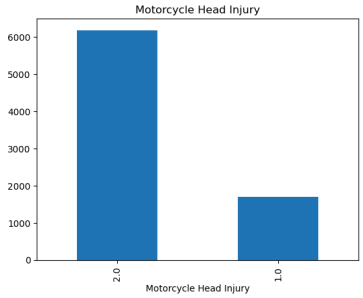
No missing data, and values are as such:

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:

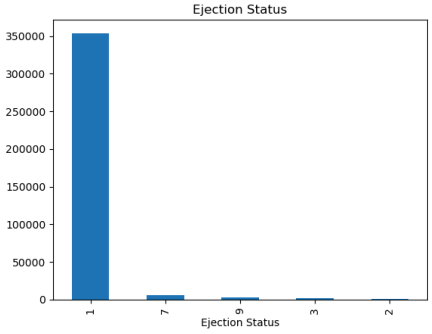
Distinct	2
Distinct (%)	< 0.1%
Missing	356994
Missing (%)	97.8%
Memory size	2.8 MiB



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:

Distinct	5
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	2.8 MiB

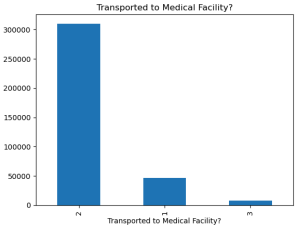


No missing data, and values are as such:

1 Not Ejected
2 Partially Ejected
3 Totally Ejected
7 Not Applicable
9 Unknown, thus most weren't ejected

Transported To a Medical Facility?

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

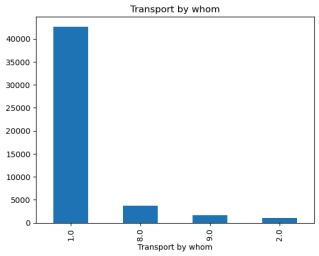


No missing data, and the values are as such:

1 Transported To Medical Facility
2 Not Transported To Medical Facility
3 Unknown, most weren't transported to a medical facility

Transported by Whom?

Distinct	4
Distinct (%)	< 0.1%
Missing	315758
Missing (%)	86.5%

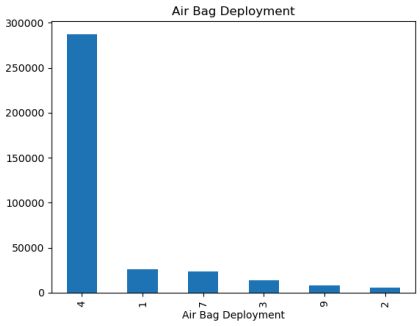


Most of the data is missing (88.5%), but values are as such:

1 EMS
2 Police
8 Other
9 Unknown, so most were transported by EMS

Air Bag Deployment:

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	4.0128918	**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:

Distinct	9400
Distinct (%)	3.4%
Missing	90160
Missing (%)	24.7%
Memory size	2.8 MiB

24.7% of the data is missing, but the two most common zip codes are 29483 and 29445

Dataset 3: 2018 tbd:

- Overall statistics:

Dataset statistics	
Number of variables	22
Number of observations	4299
Missing cells	20670
Missing cells (%)	21.9%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	739.0 KiB
Average record size in memory	176.0 B

Variable types	
Numeric	4
Text	6
Categorical	12

- Dataset Observations:

The dataset 2018 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():

```
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Collision Number                    4299 non-null   int64
 1   Unit Number                         4299 non-null   int64
 2   Carrier Name                        4196 non-null   object
 3   Carrier Street                      4196 non-null   object
 4   Carrier City                        4197 non-null   object
 5   Carrier State                       4051 non-null   object
 6   Carrier Zip                         4191 non-null   object
 7   Carrier DOT Number                  2692 non-null   object
 8   Access Control                      4199 non-null   float64
 9   Carrying Hazardous Materials?       4200 non-null   float64
 10  Hazardous Materials Placard?        4198 non-null   float64
 11  Hazardous Materials Class           119 non-null    float64
 12  Hazardous Materials ID              115 non-null    object
 13  Hazardous Materials Released?       4198 non-null   float64
 14  Gross Veh Weight Rating/Combo Rating  4200 non-null   float64
 15  Vehicle Configuration               4235 non-null   float64
 16  Trailer Length1 Code                2940 non-null   float64
 17  Trailer Length2 Code                980 non-null    float64
 18  Trailer Width1 Code                 2928 non-null   float64
 19  Trailer Width2 Code                 979 non-null    float64
 20  Citation Issued                     4299 non-null   int64
 21  Carrier Type                        4197 non-null   float64
dtypes: float64(12), int64(3), object(7)
memory usage: 739.0+ KB
```

- 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>

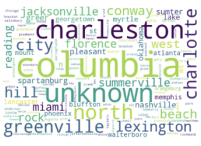
Disregarded because it is used to give a name to the people involved in the collision

<u>Carrier City</u>:

Carrier City
Text

MISSING

Distinct	1164
Distinct (%)	27.7%
Missing	102
Missing (%)	2.4%
Memory size	33.7 KiB



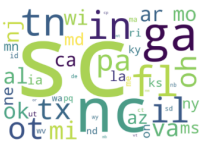
,

2.4% of the data is missing, but the most common city these trucking agency collisions occur is Columbia

<u>Carrier State:</u>

MISSING

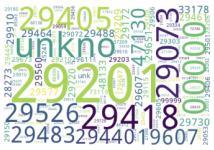
Distinct	51
Distinct (%)	1.3%
Missing	248
Missing (%)	5.8%
Memory size	33.7 KiB



5.8% 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:

Distinct	1550
Distinct (%)	37.0%
Missing	108
Missing (%)	2.5%
Memory size	33.7 KiB



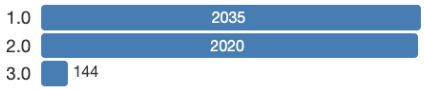
2.5 % of the data is missing, but most common zip code is 29201

Carrier Dot Number:

Disregarded because it does not add much to the data analysis

Access Control:

Distinct	3
Distinct (%)	0.1%
Missing	100
Missing (%)	2.3%
Memory size	33.7 KiB



2.3% 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?

Distinct	3
Distinct (%)	0.1%
Missing	99
Missing (%)	2.3%

2.3% 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

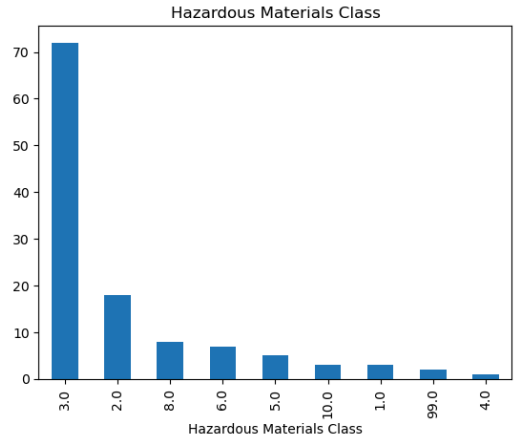
Hazardous Materials Placard?

Distinct	3
Distinct (%)	0.1%
Missing	101
Missing (%)	2.3%

2.3% 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:



97.2% 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 flammable liquids

Hazardous Materials ID:

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

Hazardous Materials Released?

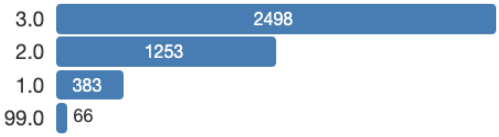
Distinct	3
Distinct (%)	0.1%
Missing	101
Missing (%)	2.3%

2.3% 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:

Distinct	4
Distinct (%)	0.1%
Missing	99
Missing (%)	2.3%

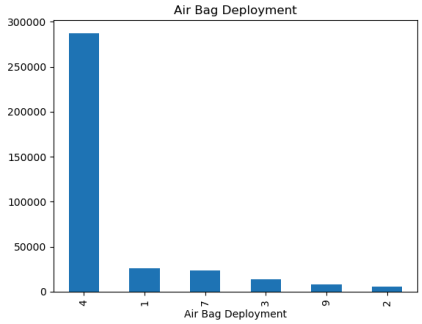


2.3% of the dataset is missing, but the values is as follows:

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:

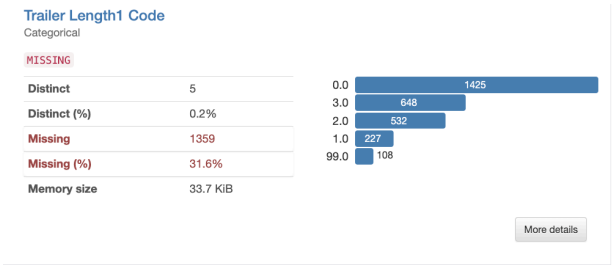
Distinct	13	Minimum	0
Distinct (%)	0.3%	Maximum	99
Missing	64	Zeros	6
Missing (%)	1.5%	Zeros (%)	0.1%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	16.401653	Memory size	33.7 KiB

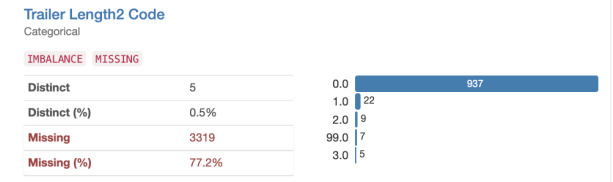


1.5% of the dataset is missing, but values are as follows:

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 cars or trucks

Trailer Length Code 1 and 2:





A majority of the dataset is missing, but the values are as follows:

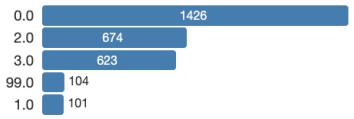
00 No Trailer
01 Less than 480 inches
02 481 - 576 inches
03 577 inches or more
99 Unknown/Hit&Run , so most didn't have a trailer

Trailer Width Code 1 and 2:

Trailer Width1 Code
Categorical

MISSING

Distinct	5
Distinct (%)	0.2%
Missing	1371
Missing (%)	31.9%
Memory size	33.7 KiB

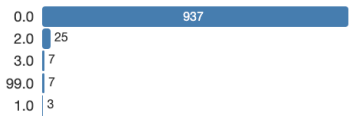


More details

Trailer Width2 Code
Categorical

IMBALANCE MISSING

Distinct	5
Distinct (%)	0.5%
Missing	3320
Missing (%)	77.2%

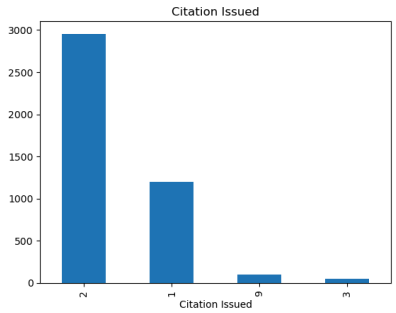


A majority of the dataset is missing, but the values are as follows:

00 No Trailer
01 Less than 60 inches
02 61 - 84 inches
03 84 inches or more
99 Unknown/Hit&Run, so most didn't have a trailer

Citation Issued:

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

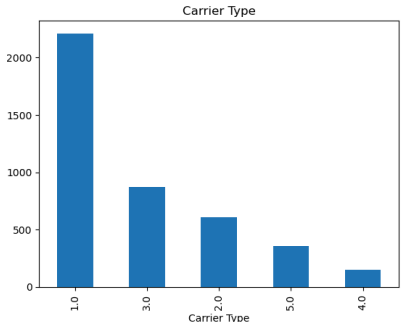


No missing data, but values are as follows:

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

Carrier Type:

Distinct	5
Distinct (%)	0.1%
Missing	102
Missing (%)	2.4%



2.4% 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: 2018 unt:

- Overall statistics:

Dataset statistics	
Number of variables	47
Number of observations	269752
Missing cells	4526288
Missing cells (%)	35.7%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	96.7 MiB
Average record size in memory	376.0 B

Variable types	
Numeric	23
Categorical	15
Text	8
Unsupported	1

- Dataset Observations:

The dataset 2018 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                        269752 non-null  int64
  1   Unit Number                             269752 non-null  int64
  2   Driver Sex                              269752 non-null  object
  3   Driver Race                             269752 non-null  object
  4   Drivers License State                   260522 non-null  object
  5   dlc                                     249143 non-null  object
  6   Vehicle Make                            256691 non-null  object
  7   Vehicle Registration Plate State        268590 non-null  object
  8   Vehicle Registration Plate Year         252535 non-null  object
  9   Contributed to Collision                269752 non-null  object
  10  Speed Limit                             264347 non-null  float64
  11  Citation Violation Code 1               73401 non-null   object
  12  vlc1                                    73007 non-null   object
  13  vlc2                                    6708 non-null    object
  14  Direction of Travel                     269752 non-null  object
  15  Unit Type                               269752 non-null  int64
  16  Vehicle Use                             269752 non-null  int64
  17  Vehicle Attachments                     269752 non-null  object
  18  Action Prior to Impact                  269752 non-null  int64
  19  Property Damage                         12525 non-null   float64
  20  Towed                                   0 non-null       float64
  21  Extent of Deformity                     269752 non-null  int64
  22  Most Harmful Event                      269752 non-null  int64
  23  Property Damage 2                       2453 non-null    float64
  24  Alcohol/Drug Information                269752 non-null  object
  25  Citation Violation Code 2               6761 non-null    object
  26  Truck/Bus Supplemental Form Required    269752 non-null  object
  27  Manner of Collision                     142433 non-null  float64
  28  Underride / Override                    269752 non-null  int64
  29  Alcohol Test Given                      12230 non-null   float64
  30  Drug Test Given                         12122 non-null   float64
  31  Alcohol Test Type                       2761 non-null    float64
  32  Drug Test Type                          933 non-null     float64
  33  Drug Test Results                       630 non-null     float64
  34  Vehicle Body Type                       4201 non-null    float64
  35  Sequence of Events1                     269752 non-null  int64
  36  Sequence of Events2                     47719 non-null   float64
  37  Sequence of Events3                     18260 non-null   float64
  38  Sequence of Events4                     6263 non-null    float64
  39  Estimated Collision Speed               266551 non-null  object
  40  Unit Damage(in dollars)                 269609 non-null  float64
  41  First Deformed Area                     269736 non-null  float64
  42  Most Deformed Area                      269752 non-null  int64
  43  Alcohol Test Results                    2350 non-null    float64
  44  Vehicle Identification Number           244287 non-null  object
  45  Number of Occupants                     269752 non-null  int64
  46  CDL licensed required                   269752 non-null  object
dtypes: float64(17), int64(11), object(19)
```

Here, it is evident that out of 46 variables, three don't have a value in the layout. This is strange because the layout has 47 variables mentioned.

- Variable Observations:

<u>Collision Number:</u>

Distinct	142406	Minimum	1803823
Distinct (%)	52.8%	Maximum	18693237
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	18558334	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

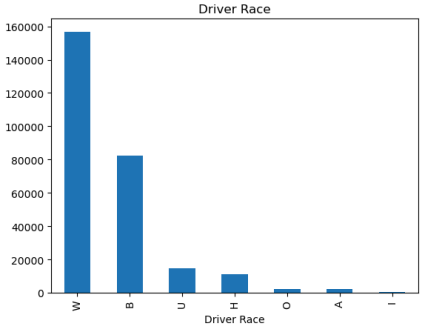
Driver Race + Sex:

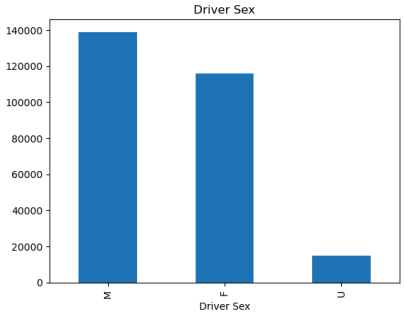
Driver Sex
Categorical

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

Driver Race
Categorical

Distinct	7
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

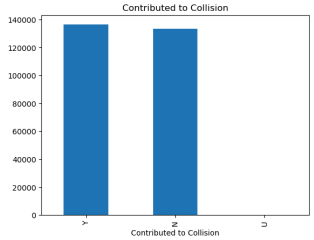
, so Most drivers are male and caucasian (M is male, F is female, and U is unknown)

Driver License State, 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:

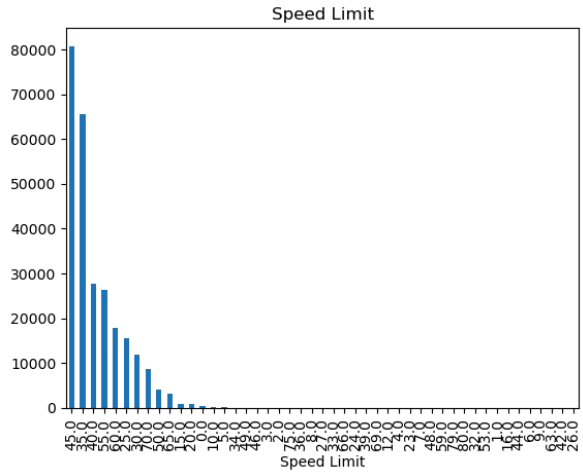
Distinct	3
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%



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:

Distinct	47	Minimum	0
Distinct (%)	< 0.1%	Maximum	80
Missing	5405	Zeros	356
Missing (%)	2.0%	Zeros (%)	0.1%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	43.004679	Memory size	2.1 MiB



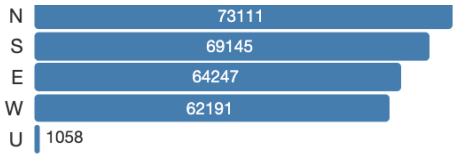
2% of the dataset is missing, but most collisions occurred when the speed limit is around 45 mph

Citation Violation Code 1 and 2:

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

Direction of Travel:

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

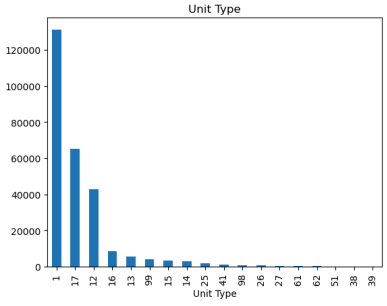


There is no missing data, and a majority of these vehicles were traveling north

Unit Type:

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	10.028945	Memory size	2.1 MiB

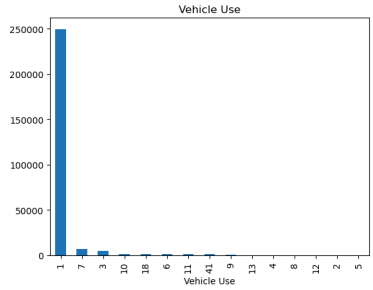
uble click to hide



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, vans, and trucks

Vehicle Use:



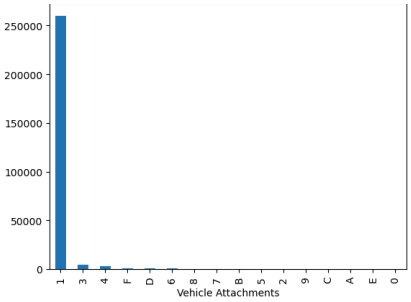
There are no missing data, and the values are such:

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 person vehicles

Vehicle Attachments:

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



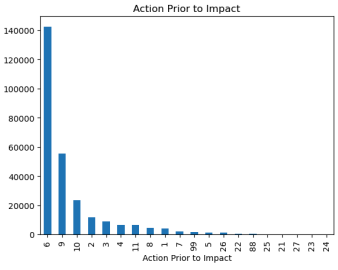
There is no missing data, and the values are as such:

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:

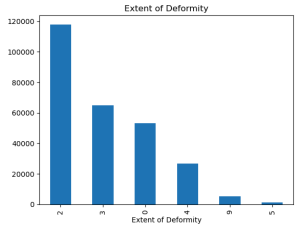
Disregarded because most of the dataset is missing, and there is no value associated in the layout

Towed:

Disregarded because it is not supported by pandas profiling

Extent of Deformity:

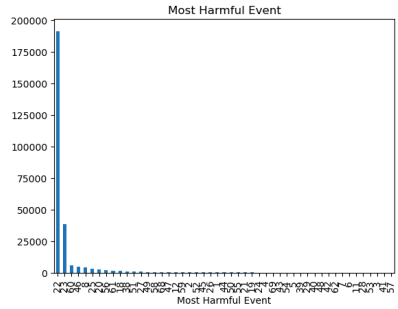
Distinct	6	**Minimum**	0
Distinct (%)	< 0.1%	**Maximum**	9
Missing	0	**Zeros**	53240
Missing (%)	0.0%	**Zeros (%)**	19.7%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	2.1985231	**Memory size**	2.1 MiB



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:



No missing data, and values are as follows:

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

, so most were collisions with a vehicle in transit

Alcohol/Drug Information:

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

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)

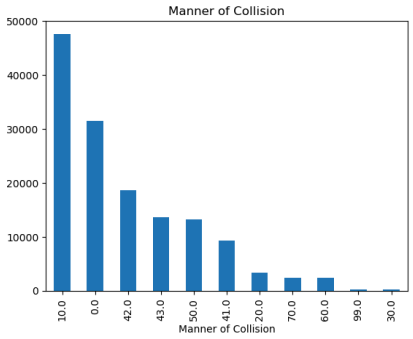
Truck/Bus Supplemental Form Required:

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:

Distinct	11	**Minimum**	0
Distinct (%)	< 0.1%	**Maximum**	99
Missing	127319	**Zeros**	31537
Missing (%)	47.2%	**Zeros (%)**	11.7%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	23.109694	**Memory size**	2.1 MiB

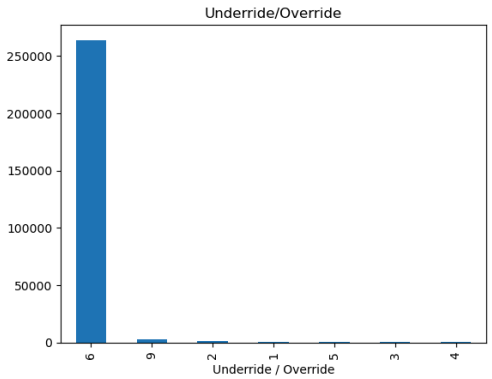


47.2% of the data is missing, but the values are as such:

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:



No missing data, and the values are as follows:

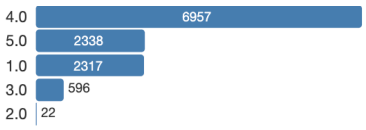
1 Under-Compartment Intrusion
2 Under-No Intrusion
3 Under-Unknown
4 Over-Motor Vehicle In Transport
5 Over-Other Vehicle
6 None
9 Unknown

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

Alcohol Test Given, Results, and Type:

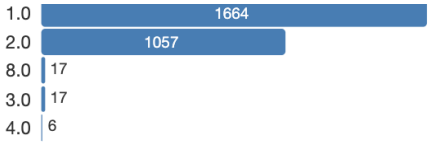
G:

Distinct	5
Distinct (%)	< 0.1%
Missing	257522
Missing (%)	95.5%



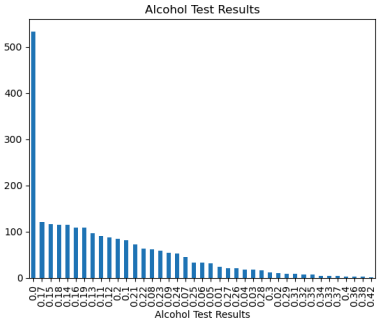
T:

Distinct	5
Distinct (%)	0.2%
Missing	266991
Missing (%)	99.0%



R:

Distinct	41	Minimum	0
Distinct (%)	1.7%	Maximum	0.42
Missing	267402	Zeros	533
Missing (%)	99.1%	Zeros (%)	0.2%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	0.1225617	Memory size	2.1 MiB



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

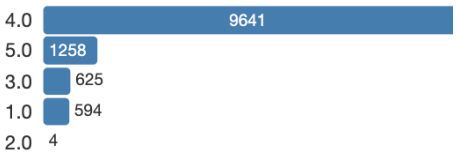
Alcohol Test Given	1 Given/Known Results 2 Given/Unusable 3 Given/Pending 4 None Given 5 Refused
Alcohol Test Results	
Alcohol Test Type	1 Breath (Alc Only) 2 Blood 3 Urine 4 Serum 8 Other

, so most times, none were given, but when they were, only breath was checked

Drug Test Given, Results, and Type:

G:

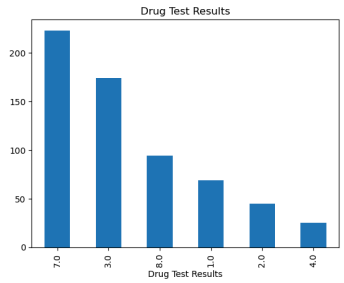
Distinct	5
Distinct (%)	< 0.1%
Missing	257630
Missing (%)	95.5%



T:

Distinct	4
Distinct (%)	0.4%
Missing	268819
Missing (%)	99.7%

R:



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

, no test was given, but when it was, it was mostly a blood test, and found nothing

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

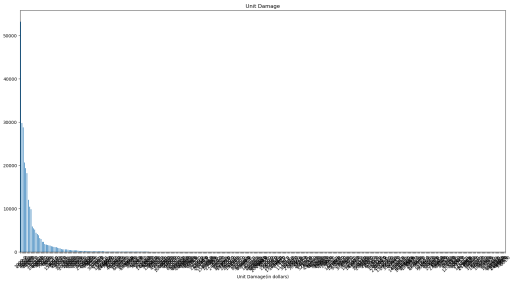
Estimated Collision Speed:

Distinct	98
Distinct (%)	< 0.1%
Missing	3201
Missing (%)	1.2%
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):

Distinct	381	**Minimum**	0
Distinct (%)	0.1%	**Maximum**	220000
Missing	143	**Zeros**	9872
Missing (%)	0.1%	**Zeros (%)**	3.7%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	2502.0595	**Memory size**	2.1 MiB



No missing data, and the most were under $1000

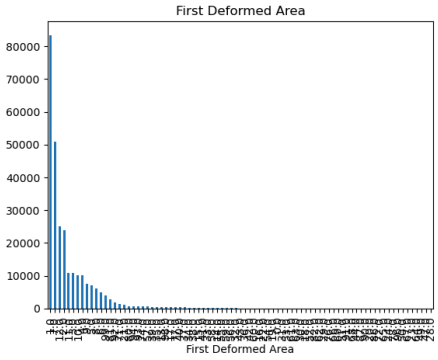
First and Most Deformed Areas:

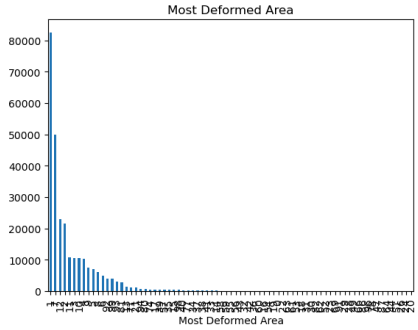
First Deformed Area
Real number (ℝ)

Distinct	83	Minimum	0
Distinct (%)	< 0.1%	Maximum	99
Missing	16	Zeros	9
Missing (%)	< 0.1%	Zeros (%)	< 0.1%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	9.6516038	Memory size	2.1 MiB

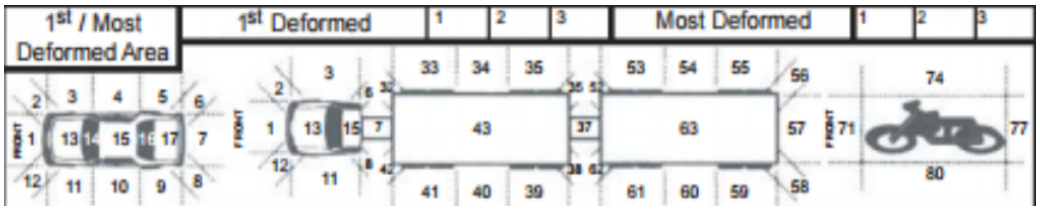
Most Deformed Area
Real number (ℝ)

Distinct	77	Minimum	0
Distinct (%)	< 0.1%	Maximum	99
Missing	0	Zeros	9
Missing (%)	0.0%	Zeros (%)	< 0.1%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	11.265588	Memory size	2.1 MiB





No missing data, and the values are as such:



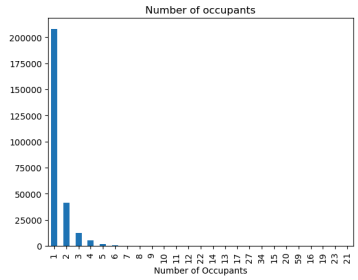
, so mostly the front

Vehicle Identification Number:

Disregarded because it only provides vehicle identity

Number of Occupants:

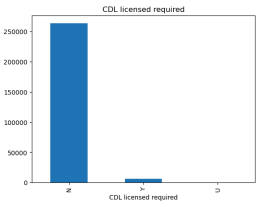
Distinct	25	Minimum	1
Distinct (%)	< 0.1%	Maximum	59
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	1.3526647	Memory size	2.1 MiB



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

CDL License Required:

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



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

Variables Not Explored:

Dlc, vlc1, vlc2 (both are junk)