In [1]:
```
import pandas as pd
import numpy as np
import csv

from pltfunctions import hist_kde_plots
from math import sqrt

import matplotlib.pyplot as plt
import seaborn as sns

from edafunctions import df_remove_columns_threshold as rmcol
from edafunctions import df_merge_dataframes_left as merle
```

Basic Data Import and Cleaning

In [2]:
```
dfvehicles = pd.read_csv(r"data/TrafficCrashes-Vehicle.csv", low_memory=False
```

In [12]:
```
dfvehicles = rmcol(dfvehicles)
```

In [13]:
```
dfvehicles.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 925738 entries, 0 to 925737
Data columns (total 16 columns):
 #   Column               Non-Null Count   Dtype  
---  ------               --------------   -----  
 0   CRASH_UNIT_ID        925738 non-null  int64  
 1   CRASH_RECORD_ID      925738 non-null  object 
 2   RD_NO                918507 non-null  object 
 3   CRASH_DATE           925738 non-null  object 
 4   UNIT_NO              925738 non-null  int64  
 5   UNIT_TYPE            924349 non-null  object 
 6   VEHICLE_ID           904074 non-null  float64
 7   MAKE                 904069 non-null  object 
 8   MODEL                903927 non-null  object 
 9   VEHICLE_DEFECT       904074 non-null  object 
 10  VEHICLE_TYPE         904074 non-null  object 
 11  VEHICLE_USE          904074 non-null  object 
 12  TRAVEL_DIRECTION     904074 non-null  object 
 13  MANEUVER             904074 non-null  object 
 14  OCCUPANT_CNT         904074 non-null  float64
 15  FIRST_CONTACT_POINT  898669 non-null  object 
dtypes: float64(2), int64(2), object(12)
memory usage: 113.0+ MB
```

In [14]:
```
dfcrash = pd.read_csv(r"data/TrafficCrashes-Crashes.csv", low_memory=False)
```

In [15]:
```
dfcrash = rmcol(dfcrash)
```

In [16]:
```
dfcrash.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 453873 entries, 0 to 453872
Data columns (total 38 columns):
 #   Column                          Non-Null Count   Dtype  
---  ------                          --------------   -----  
 0   CRASH_RECORD_ID                 453873 non-null  object 
 1   RD_NO                           450376 non-null  object 
 2   CRASH_DATE                      453873 non-null  object 
 3   POSTED_SPEED_LIMIT              453873 non-null  int64  
 4   TRAFFIC_CONTROL_DEVICE          453873 non-null  object 
 5   DEVICE_CONDITION                453873 non-null  object 
 6   WEATHER_CONDITION               453873 non-null  object 
 7   LIGHTING_CONDITION              453873 non-null  object 
 8   FIRST_CRASH_TYPE                453873 non-null  object 
 9   TRAFFICWAY_TYPE                 453873 non-null  object 
 10  ALIGNMENT                       453873 non-null  object 
 11  ROADWAY_SURFACE_COND            453873 non-null  object 
 12  ROAD_DEFECT                     453873 non-null  object 
 13  REPORT_TYPE                     443012 non-null  object 
 14  CRASH_TYPE                      453873 non-null  object 
 15  DAMAGE                          453873 non-null  object 
 16  DATE_POLICE_NOTIFIED            453873 non-null  object 
 17  PRIM_CONTRIBUTORY_CAUSE         453873 non-null  object 
 18  SEC_CONTRIBUTORY_CAUSE          453873 non-null  object 
 19  STREET_NO                       453873 non-null  int64  
 20  STREET_DIRECTION                453870 non-null  object 
 21  STREET_NAME                     453872 non-null  object 
 22  BEAT_OF_OCCURRENCE              453868 non-null  float64
 23  NUM_UNITS                       453873 non-null  int64  
 24  MOST_SEVERE_INJURY              452971 non-null  object 
 25  INJURIES_TOTAL                  452981 non-null  float64
 26  INJURIES_FATAL                  452981 non-null  float64
 27  INJURIES_INCAPACITATING         452981 non-null  float64
 28  INJURIES_NON_INCAPACITATING     452981 non-null  float64
 29  INJURIES_REPORTED_NOT_EVIDENT   452981 non-null  float64
 30  INJURIES_NO_INDICATION          452981 non-null  float64
 31  INJURIES_UNKNOWN                452981 non-null  float64
 32  CRASH_HOUR                      453873 non-null  int64  
 33  CRASH_DAY_OF_WEEK               453873 non-null  int64  
 34  CRASH_MONTH                     453873 non-null  int64  
 35  LATITUDE                        451411 non-null  float64
 36  LONGITUDE                       451411 non-null  float64
 37  LOCATION                        451411 non-null  object 
dtypes: float64(10), int64(6), object(22)
memory usage: 131.6+ MB
```

In [17]:
```
dfpeople = pd.read_csv(r"data/TrafficCrashes-People.csv", low_memory=False)
```

In [18]:
```
dfpeople = rmcol(dfpeople)
```

In [19]:
```
dfpeople.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1006093 entries, 0 to 1006092
Data columns (total 11 columns):
 #   Column                Non-Null Count    Dtype  
---  ------                --------------    -----  
 0   PERSON_ID             1006093 non-null  object 
 1   PERSON_TYPE           1006093 non-null  object 
 2   CRASH_RECORD_ID       1006093 non-null  object 
 3   RD_NO                 998607 non-null   object 
 4   VEHICLE_ID            985919 non-null   float64
 5   CRASH_DATE            1006093 non-null  object 
 6   SEX                   991169 non-null   object 
 7   SAFETY_EQUIPMENT      1003090 non-null  object 
 8   AIRBAG_DEPLOYED       986732 non-null   object 
 9   EJECTION              993588 non-null   object 
 10  INJURY_CLASSIFICATION  1005547 non-null  object 
dtypes: float64(1), object(10)
memory usage: 84.4+ MB
```

Create a merged data table on CRASH_RECOR

In [20]:
```
merge = 'CRASH_RECORD_ID'
```

In [21]:
```
dfmerge = pd.merge(dfvehicles, dfcrash, how='left', on=merge)
```

In [22]:
```
dfmerge.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 925738 entries, 0 to 925737
Data columns (total 53 columns):
 #   Column                          Non-Null Count   Dtype  
---  ------                          --------------   -----  
 0   CRASH_UNIT_ID                   925738 non-null  int64  
 1   CRASH_RECORD_ID                 925738 non-null  object 
 2   RD_NO_x                         918507 non-null  object 
 3   CRASH_DATE_x                    925738 non-null  object 
 4   UNIT_NO                         925738 non-null  int64  
 5   UNIT_TYPE                       924349 non-null  object 
 6   VEHICLE_ID                      904074 non-null  float64
 7   MAKE                            904069 non-null  object 
 8   MODEL                           903927 non-null  object 
 9   VEHICLE_DEFECT                  904074 non-null  object 
 10  VEHICLE_TYPE                    904074 non-null  object 
 11  VEHICLE_USE                     904074 non-null  object 
 12  TRAVEL_DIRECTION                904074 non-null  object 
 13  MANEUVER                        904074 non-null  object 
 14  OCCUPANT_CNT                    904074 non-null  float64
 15  FIRST_CONTACT_POINT             898669 non-null  object 
 16  RD_NO_y                         918507 non-null  object 
 17  CRASH_DATE_y                    925738 non-null  object 
 18  POSTED_SPEED_LIMIT              925738 non-null  int64  
 19  TRAFFIC_CONTROL_DEVICE          925738 non-null  object 
 20  DEVICE_CONDITION                925738 non-null  object 
 21  WEATHER_CONDITION               925738 non-null  object 
 22  LIGHTING_CONDITION              925738 non-null  object 
 23  FIRST_CRASH_TYPE                925738 non-null  object 
 24  TRAFFICWAY_TYPE                 925738 non-null  object 
 25  ALIGNMENT                       925738 non-null  object 
 26  ROADWAY_SURFACE_COND            925738 non-null  object 
 27  ROAD_DEFECT                     925738 non-null  object 
 28  REPORT_TYPE                     898568 non-null  object 
 29  CRASH_TYPE                      925738 non-null  object 
 30  DAMAGE                          925738 non-null  object 
 31  DATE_POLICE_NOTIFIED            925738 non-null  object 
 32  PRIM_CONTRIBUTORY_CAUSE         925738 non-null  object 
 33  SEC_CONTRIBUTORY_CAUSE          925738 non-null  object 
 34  STREET_NO                       925738 non-null  int64  
 35  STREET_DIRECTION                925732 non-null  object 
 36  STREET_NAME                     925736 non-null  object 
 37  BEAT_OF_OCCURRENCE              925728 non-null  float64
 38  NUM_UNITS                       925738 non-null  int64  
 39  MOST_SEVERE_INJURY              924182 non-null  object 
 40  INJURIES_TOTAL                  924202 non-null  float64
 41  INJURIES_FATAL                  924202 non-null  float64
 42  INJURIES_INCAPACITATING         924202 non-null  float64
 43  INJURIES_NON_INCAPACITATING     924202 non-null  float64
 44  INJURIES_REPORTED_NOT_EVIDENT   924202 non-null  float64
 45  INJURIES_NO_INDICATION          924202 non-null  float64
 46  INJURIES_UNKNOWN                924202 non-null  float64
 47  CRASH_HOUR                      925738 non-null  int64  
```

```
 48  CRASH_DAY_OF_WEEK             925738 non-null  int64
 49  CRASH_MONTH                   925738 non-null  int64
 50  LATITUDE                      920858 non-null  float64
 51  LONGITUDE                     920858 non-null  float64
 52  LOCATION                      920858 non-null  object
dtypes: float64(12), int64(8), object(33)
memory usage: 381.4+ MB
```

In [23]:
```
dfmerged = pd.merge(dfmerge, dfpeople, how='left', on=merge)
```

In [24]:
```
dfmerged.describe()
```

	CRASH_UNIT_ID	UNIT_NO	VEHICLE_ID_x	OCCUPANT_CNT	POSTED_SPEED_
count	2.115954e+06	2.115954e+06	2.065023e+06	2.065023e+06	2.115954e+06
mean	4.978058e+05	3.374134e+00	4.741220e+05	1.388109e+00	2.883798e+01
std	2.860458e+05	2.597244e+03	2.699553e+05	1.404195e+00	5.998828e+00
min	2.000000e+00	0.000000e+00	2.000000e+00	0.000000e+00	0.000000e+00
25%	2.496780e+05	1.000000e+00	2.419340e+05	1.000000e+00	3.000000e+01
50%	4.984790e+05	2.000000e+00	4.755130e+05	1.000000e+00	3.000000e+01
75%	7.466410e+05	2.000000e+00	7.083555e+05	2.000000e+00	3.000000e+01
max	9.906910e+05	3.778035e+06	9.388350e+05	9.900000e+01	9.900000e+01

8 rows × 21 columns

In [25]:
```
dfmerged = dfmerged.dropna() # because of amount of data, am going to remove
```

In [26]:
```
dfmerged.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1897113 entries, 0 to 2115933
Data columns (total 63 columns):
 #   Column                          Dtype
---  ------                          -----
 0   CRASH_UNIT_ID                   int64
 1   CRASH_RECORD_ID                 object
 2   RD_NO_x                         object
 3   CRASH_DATE_x                    object
 4   UNIT_NO                         int64
 5   UNIT_TYPE                       object
 6   VEHICLE_ID_x                    float64
 7   MAKE                            object
 8   MODEL                           object
 9   VEHICLE_DEFECT                  object
 10  VEHICLE_TYPE                    object
 11  VEHICLE_USE                     object
 12  TRAVEL_DIRECTION                object
 13  MANEUVER                        object
 14  OCCUPANT_CNT                    float64
 15  FIRST_CONTACT_POINT             object
 16  RD_NO_y                         object
 17  CRASH_DATE_y                    object
 18  POSTED_SPEED_LIMIT              int64
```

In [27]:
```
# what other columns can be dropped right away?
# drop_columns = ['RD_NO_x, TRAVEL_DIRECTION, RD_NO_y, DATE_POLICE_NOTIFIED,
dfmerged = dfmerged.drop(columns=['CRASH_RECORD_ID', 'RD_NO_x', 'TRAVEL_DIREC
```

In [28]:
```
dfmerged = dfmerged.drop(columns=['RD_NO', 'VEHICLE_ID_y'])
```

In [29]:
```
dfmerged.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1897113 entries, 0 to 2115933
Data columns (total 48 columns):
 #   Column                          Dtype  
---  ------                          -----  
 0   CRASH_DATE_x                    object 
 1   UNIT_TYPE                       object 
 2   MAKE                            object 
 3   MODEL                           object 
 4   VEHICLE_DEFECT                  object 
 5   VEHICLE_TYPE                    object 
 6   VEHICLE_USE                     object 
 7   MANEUVER                        object 
 8   OCCUPANT_CNT                    float64
 9   CRASH_DATE_y                    object 
 10  POSTED_SPEED_LIMIT              int64  
 11  TRAFFIC_CONTROL_DEVICE          object 
 12  DEVICE_CONDITION                object 
 13  WEATHER_CONDITION               object 
 14  LIGHTING_CONDITION              object 
 15  FIRST_CRASH_TYPE                object 
 16  TRAFFICWAY_TYPE                 object 
 17  ALIGNMENT                       object 
 18  ROADWAY_SURFACE_COND            object 
 19  ROAD_DEFECT                     object 
 20  REPORT_TYPE                     object 
 21  CRASH_TYPE                      object 
 22  DAMAGE                          object 
 23  PRIM_CONTRIBUTORY_CAUSE         object 
 24  SEC_CONTRIBUTORY_CAUSE          object 
 25  BEAT_OF_OCCURRENCE              float64
 26  NUM_UNITS                       int64  
 27  MOST_SEVERE_INJURY              object 
 28  INJURIES_TOTAL                  float64
 29  INJURIES_FATAL                  float64
 30  INJURIES_INCAPACITATING         float64
 31  INJURIES_NON_INCAPACITATING     float64
 32  INJURIES_REPORTED_NOT_EVIDENT   float64
 33  INJURIES_NO_INDICATION          float64
 34  INJURIES_UNKNOWN                float64
 35  CRASH_HOUR                      int64  
 36  CRASH_DAY_OF_WEEK               int64  
 37  CRASH_MONTH                     int64  
 38  LATITUDE                        float64
 39  LONGITUDE                       float64
 40  PERSON_ID                       object 
 41  PERSON_TYPE                     object 
 42  CRASH_DATE                      object 
 43  SEX                             object 
 44  SAFETY_EQUIPMENT                object 
 45  AIRBAG_DEPLOYED                 object 
 46  EJECTION                        object 
 47  INJURY_CLASSIFICATION           object 
```

```
dtypes: float64(11), int64(5), object(32)
memory usage: 709.2+ MB
```

In [30]:
```
dfmerged['CRASH_TYPE'].unique() # multicollinearity with most severe injury c
```

```
array(['NO INJURY / DRIVE AWAY', 'INJURY AND / OR TOW DUE TO CRASH'],
      dtype=object)
```

In [31]:
```
dfmerged['MOST_SEVERE_INJURY'].unique() # target classification column, INJUI
```

```
array(['NO INDICATION OF INJURY', 'NONINCAPACITATING INJURY',
       'REPORTED, NOT EVIDENT', 'INCAPACITATING INJURY', 'FATAL'],
      dtype=object)
```

In [32]:
```
dfmerged['OCCUPANT_CNT'].unique() # occupant count
```

```
array([ 1.,  0.,  2.,  3.,  5.,  4., 37.,  6.,  8.,  9., 13.,  7., 35.,
       26., 20., 16., 15., 14., 12., 44., 18., 22., 36., 11., 10., 19.,
       30., 33., 24., 43., 60., 34., 17., 39., 25., 27., 21., 29., 41.,
       28., 47., 38., 99.])
```

In [33]:
```
dfmerged['PERSON_TYPE'].unique()
```

```
array(['DRIVER', 'PASSENGER', 'NON-CONTACT VEHICLE'], dtype=object)
```

In [34]:
```
dfmerged['INJURIES_FATAL'].unique()
```

```
array([0., 1., 2., 3.])
```

In [35]:
```
dfmerged['INJURY_CLASSIFICATION'].unique() # gives injury on a per individual
```

```
array(['NO INDICATION OF INJURY', 'NONINCAPACITATING INJURY',
       'REPORTED, NOT EVIDENT', 'INCAPACITATING INJURY', 'FATAL'],
      dtype=object)
```

In [36]:
```
dfmerged['MANEUVER'].unique()
```

```
array(['TURNING LEFT', 'STRAIGHT AHEAD', 'SLOW/STOP IN TRAFFIC',
       'UNKNOWN/NA', 'CHANGING LANES', 'PARKED', 'PASSING/OVERTAKING',
       'MERGING', 'BACKING', 'STARTING IN TRAFFIC', 'OTHER',
       'AVOIDING VEHICLES/OBJECTS', 'SLOW/STOP - LOAD/UNLOAD',
       'SKIDDING/CONTROL LOSS', 'NEGOTIATING A CURVE', 'TURNING RIGHT',
       'ENTER FROM DRIVE/ALLEY', 'U-TURN', 'PARKED IN TRAFFIC LANE',
       'LEAVING TRAFFIC LANE TO PARK', 'SLOW/STOP - LEFT TURN',
       'ENTERING TRAFFIC LANE FROM PARKING', 'DRIVERLESS',
       'SLOW/STOP - RIGHT TURN', 'DIVERGING', 'TURNING ON RED',
       'DRIVING WRONG WAY', 'DISABLED'], dtype=object)
```

In [37]:
```
dfmerged['PRIM_CONTRIBUTORY_CAUSE'].unique()
```

```
array(['UNABLE TO DETERMINE', 'FOLLOWING TOO CLOSELY',
       'FAILING TO YIELD RIGHT-OF-WAY', 'IMPROPER LANE USAGE',
       'IMPROPER OVERTAKING/PASSING', 'NOT APPLICABLE',
       'IMPROPER BACKING', 'FAILING TO REDUCE SPEED TO AVOID CRASH',
       'DISTRACTION - FROM INSIDE VEHICLE', 'WEATHER',
       'DISREGARDING STOP SIGN', 'PHYSICAL CONDITION OF DRIVER',
       'VISION OBSCURED (SIGNS, TREE LIMBS, BUILDINGS, ETC.)',
       'DRIVING SKILLS/KNOWLEDGE/EXPERIENCE',
       'IMPROPER TURNING/NO SIGNAL',
       'EXCEEDING SAFE SPEED FOR CONDITIONS',
       'EQUIPMENT - VEHICLE CONDITION', 'DRIVING ON WRONG SIDE/WRONG WAY',
       'OPERATING VEHICLE IN ERRATIC, RECKLESS, CARELESS, NEGLIGENT OR AGGRESSIVE MANNER',
       'EXCEEDING AUTHORIZED SPEED LIMIT', 'DISREGARDING TRAFFIC SIGNALS',
       'DISREGARDING ROAD MARKINGS',
       'ROAD ENGINEERING/SURFACE/MARKING DEFECTS',
       'EVASIVE ACTION DUE TO ANIMAL, OBJECT, NONMOTORIST', 'TEXTING',
       'UNDER THE INFLUENCE OF ALCOHOL/DRUGS (USE WHEN ARREST IS EFFECTED)',
       'DISTRACTION - FROM OUTSIDE VEHICLE', 'ANIMAL',
       'ROAD CONSTRUCTION/MAINTENANCE',
       'CELL PHONE USE OTHER THAN TEXTING',
       'DISREGARDING OTHER TRAFFIC SIGNS',
       'HAD BEEN DRINKING (USE WHEN ARREST IS NOT MADE)',
       'TURNING RIGHT ON RED', 'PASSING STOPPED SCHOOL BUS',
       'DISTRACTION - OTHER ELECTRONIC DEVICE (NAVIGATION DEVICE, DVD PLAYER, ETC.)',
       'DISREGARDING YIELD SIGN',
       'MOTORCYCLE ADVANCING LEGALLY ON RED LIGHT',
       'BICYCLE ADVANCING LEGALLY ON RED LIGHT', 'RELATED TO BUS STOP',
       'OBSTRUCTED CROSSWALKS'], dtype=object)
```

In [38]:
```
dfmerged['SEC_CONTRIBUTORY_CAUSE'].unique()
```

```
array(['UNABLE TO DETERMINE', 'NOT APPLICABLE',
       'FAILING TO REDUCE SPEED TO AVOID CRASH',
       'DRIVING SKILLS/KNOWLEDGE/EXPERIENCE', 'IMPROPER LANE USAGE',
       'FOLLOWING TOO CLOSELY',
       'VISION OBSCURED (SIGNS, TREE LIMBS, BUILDINGS, ETC.)',
       'IMPROPER OVERTAKING/PASSING', 'FAILING TO YIELD RIGHT-OF-WAY',
       'OPERATING VEHICLE IN ERRATIC, RECKLESS, CARELESS, NEGLIGENT OR AGGRESSIVE MANNER',
       'DRIVING ON WRONG SIDE/WRONG WAY', 'WEATHER',
       'EXCEEDING SAFE SPEED FOR CONDITIONS',
       'MOTORCYCLE ADVANCING LEGALLY ON RED LIGHT',
       'IMPROPER TURNING/NO SIGNAL', 'EQUIPMENT - VEHICLE CONDITION',
       'DISREGARDING OTHER TRAFFIC SIGNS',
       'HAD BEEN DRINKING (USE WHEN ARREST IS NOT MADE)',
       'ROAD ENGINEERING/SURFACE/MARKING DEFECTS',
       'DISREGARDING TRAFFIC SIGNALS', 'EXCEEDING AUTHORIZED SPEED LIMIT',
       'CELL PHONE USE OTHER THAN TEXTING', 'IMPROPER BACKING',
       'PHYSICAL CONDITION OF DRIVER', 'TEXTING',
       'DISTRACTION - FROM INSIDE VEHICLE',
       'UNDER THE INFLUENCE OF ALCOHOL/DRUGS (USE WHEN ARREST IS EFFECTED)',
       'ROAD CONSTRUCTION/MAINTENANCE',
       'BICYCLE ADVANCING LEGALLY ON RED LIGHT', 'DISREGARDING STOP SIGN',
       'DISTRACTION - FROM OUTSIDE VEHICLE', 'ANIMAL',
       'PASSING STOPPED SCHOOL BUS', 'DISREGARDING ROAD MARKINGS',
       'EVASIVE ACTION DUE TO ANIMAL, OBJECT, NONMOTORIST',
       'TURNING RIGHT ON RED', 'DISREGARDING YIELD SIGN',
       'DISTRACTION - OTHER ELECTRONIC DEVICE (NAVIGATION DEVICE, DVD PLAYER, ETC.)',
       'RELATED TO BUS STOP', 'OBSTRUCTED CROSSWALKS'], dtype=object)
```

In [39]:
```
# There are a few more things to do to clean data
# 1) I want to reduce amount of data - I only want crash data from year 2018
# 2) Remove rows that have 'Unkown' or 'Other' in the VEHICLE_DEFECT column
# 3) Remove Unknown/NA from VEHICLE_TYPE, VEHICLE_USE, MANEUVER,
# 4) Remove any rows with POSTED_SPEED_LIMIT less than 15 mph
# 5) Remove Unknown from TRAFFIC_CONTROL_DEVICE
# 6) Remove Unknown from DEVICE_CONDITION
# 7) Remove longitude/latitude coordinates outside of Chicago area
```

In [40]:
```
dfmerged['CRASH_DATE_x'] = pd.to_datetime(dfmerged.CRASH_DATE_x)
dfmerged['CRASH_DATE_x'] = pd.DatetimeIndex(dfmerged['CRASH_DATE_x']).year
```

In [41]:
```
dfmerged.head()
```

	CRASH_DATE_x	UNIT_TYPE	MAKE	MODEL	VEHICLE_DEFECT	VEHICLE_TYPE	V
0	2015	DRIVER	FORD	Focus	NONE	PASSENGER	F
1	2015	DRIVER	FORD	Focus	NONE	PASSENGER	F
2	2015	DRIVER	NISSAN	Pathfinder	NONE	SPORT UTILITY VEHICLE (SUV)	F
3	2015	DRIVER	NISSAN	Pathfinder	NONE	SPORT UTILITY VEHICLE (SUV)	F
4	2015	DRIVER	FORD	F150	UNKNOWN	VAN/MINI-VAN	U

5 rows × 48 columns

In [42]:
```
df_recent = dfmerged[dfmerged.CRASH_DATE_x >= 2015]
```

In [43]:
```
df_recent.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1897085 entries, 0 to 2115933
Data columns (total 48 columns):
 #   Column                          Dtype  
---  ------                          -----  
 0   CRASH_DATE_x                    int64  
 1   UNIT_TYPE                       object 
 2   MAKE                            object 
 3   MODEL                           object 
 4   VEHICLE_DEFECT                  object 
 5   VEHICLE_TYPE                    object 
 6   VEHICLE_USE                     object 
 7   MANEUVER                        object 
 8   OCCUPANT_CNT                    float64
 9   CRASH_DATE_y                    object 
 10  POSTED_SPEED_LIMIT              int64  
 11  TRAFFIC_CONTROL_DEVICE          object 
 12  DEVICE_CONDITION                object 
 13  WEATHER_CONDITION               object 
 14  LIGHTING_CONDITION              object 
 15  FIRST_CRASH_TYPE                object 
 16  TRAFFICWAY_TYPE                 object 
 17  ALIGNMENT                       object 
 18  ROADWAY_SURFACE_COND            object 
 19  ROAD_DEFECT                     object 
 20  REPORT_TYPE                     object 
 21  CRASH_TYPE                      object 
 22  DAMAGE                          object 
 23  PRIM_CONTRIBUTORY_CAUSE         object 
 24  SEC_CONTRIBUTORY_CAUSE          object 
 25  BEAT_OF_OCCURRENCE              float64
 26  NUM_UNITS                       int64  
 27  MOST_SEVERE_INJURY              object 
 28  INJURIES_TOTAL                  float64
 29  INJURIES_FATAL                  float64
 30  INJURIES_INCAPACITATING         float64
 31  INJURIES_NON_INCAPACITATING     float64
 32  INJURIES_REPORTED_NOT_EVIDENT   float64
 33  INJURIES_NO_INDICATION          float64
 34  INJURIES_UNKNOWN                float64
 35  CRASH_HOUR                      int64  
 36  CRASH_DAY_OF_WEEK               int64  
 37  CRASH_MONTH                     int64  
 38  LATITUDE                        float64
 39  LONGITUDE                       float64
 40  PERSON_ID                       object 
 41  PERSON_TYPE                     object 
 42  CRASH_DATE                      object 
 43  SEX                             object 
 44  SAFETY_EQUIPMENT                object 
 45  AIRBAG_DEPLOYED                 object 
 46  EJECTION                        object 
 47  INJURY_CLASSIFICATION           object 
```

```
dtypes: float64(11), int64(6), object(31)
memory usage: 709.2+ MB
```

In [44]:

```
df_recent.describe()
```

	CRASH_DATE_x	OCCUPANT_CNT	POSTED_SPEED_LIMIT	BEAT_OF_OCCURREN
count	1.897085e+06	1.897085e+06	1.897085e+06	1.897085e+06
mean	2.018135e+03	1.368861e+00	2.882813e+01	1.235037e+03
std	1.293782e+00	1.296735e+00	6.018050e+00	7.058126e+02
min	2.015000e+03	0.000000e+00	0.000000e+00	1.110000e+02
25%	2.017000e+03	1.000000e+00	3.000000e+01	7.130000e+02
50%	2.018000e+03	1.000000e+00	3.000000e+01	1.211000e+03
75%	2.019000e+03	1.000000e+00	3.000000e+01	1.822000e+03
max	2.020000e+03	9.900000e+01	9.900000e+01	2.535000e+03

In [45]:

```
df1 = df_recent[df_recent['VEHICLE_DEFECT'] != 'UNKNOWN']
df1 = df1[df1['VEHICLE_DEFECT'] != 'OTHER']
df1 = df1[df1['VEHICLE_TYPE'] != 'UNKNOWN/NA']
df1 = df1[df1['TRAFFIC_CONTROL_DEVICE'] != 'UNKNOWN']
df1 = df1[df1['DEVICE_CONDITION'] != 'UNKNOWN']
```

In [46]:

```
df1['VEHICLE_DEFECT'].unique()
```

```
array(['NONE', 'BRAKES', 'TIRES', 'ENGINE/MOTOR', 'FUEL SYSTEM', 'WHEELS',
       'STEERING', 'LIGHTS', 'WINDOWS', 'RESTRAINT SYSTEM', 'CARGO',
       'SUSPENSION', 'SIGNALS', 'EXHAUST', 'TRAILER COUPLING'],
      dtype=object)
```

In [47]:

```
df1 = df1[df1['LONGITUDE'] != 0]
df1 = df1[df1['LATITUDE'] != 0]
```

In [48]:
```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1134909 entries, 0 to 2115933
Data columns (total 48 columns):
 #   Column                          Non-Null Count    Dtype  
---  ------                          --------------    -----  
 0   CRASH_DATE_x                    1134909 non-null  int64  
 1   UNIT_TYPE                       1134909 non-null  object 
 2   MAKE                            1134909 non-null  object 
 3   MODEL                           1134909 non-null  object 
 4   VEHICLE_DEFECT                  1134909 non-null  object 
 5   VEHICLE_TYPE                    1134909 non-null  object 
 6   VEHICLE_USE                     1134909 non-null  object 
 7   MANEUVER                        1134909 non-null  object 
 8   OCCUPANT_CNT                    1134909 non-null  float64
 9   CRASH_DATE_y                    1134909 non-null  object 
 10  POSTED_SPEED_LIMIT              1134909 non-null  int64  
 11  TRAFFIC_CONTROL_DEVICE          1134909 non-null  object 
 12  DEVICE_CONDITION                1134909 non-null  object 
 13  WEATHER_CONDITION               1134909 non-null  object 
 14  LIGHTING_CONDITION              1134909 non-null  object 
 15  FIRST_CRASH_TYPE                1134909 non-null  object 
 16  TRAFFICWAY_TYPE                 1134909 non-null  object 
 17  ALIGNMENT                       1134909 non-null  object 
 18  ROADWAY_SURFACE_COND            1134909 non-null  object 
 19  ROAD_DEFECT                     1134909 non-null  object 
 20  REPORT_TYPE                     1134909 non-null  object 
 21  CRASH_TYPE                      1134909 non-null  object 
 22  DAMAGE                          1134909 non-null  object 
 23  PRIM_CONTRIBUTORY_CAUSE         1134909 non-null  object 
 24  SEC_CONTRIBUTORY_CAUSE          1134909 non-null  object 
 25  BEAT_OF_OCCURRENCE              1134909 non-null  float64
 26  NUM_UNITS                       1134909 non-null  int64  
 27  MOST_SEVERE_INJURY              1134909 non-null  object 
 28  INJURIES_TOTAL                  1134909 non-null  float64
 29  INJURIES_FATAL                  1134909 non-null  float64
 30  INJURIES_INCAPACITATING         1134909 non-null  float64
 31  INJURIES_NON_INCAPACITATING     1134909 non-null  float64
 32  INJURIES_REPORTED_NOT_EVIDENT   1134909 non-null  float64
 33  INJURIES_NO_INDICATION          1134909 non-null  float64
 34  INJURIES_UNKNOWN                1134909 non-null  float64
 35  CRASH_HOUR                      1134909 non-null  int64  
 36  CRASH_DAY_OF_WEEK               1134909 non-null  int64  
 37  CRASH_MONTH                     1134909 non-null  int64  
 38  LATITUDE                        1134909 non-null  float64
 39  LONGITUDE                       1134909 non-null  float64
 40  PERSON_ID                       1134909 non-null  object 
 41  PERSON_TYPE                     1134909 non-null  object 
 42  CRASH_DATE                      1134909 non-null  object 
 43  SEX                             1134909 non-null  object 
 44  SAFETY_EQUIPMENT                1134909 non-null  object 
 45  AIRBAG_DEPLOYED                 1134909 non-null  object 
 46  EJECTION                        1134909 non-null  object 
 47  INJURY_CLASSIFICATION           1134909 non-null  object 
```

```
dtypes: float64(11), int64(6), object(31)
memory usage: 424.3+ MB
```

In [49]:

```
df2 = df1.sample(frac=0.5)
df2.to_csv('ChicagoCrashes.csv')
Sample1 = df1.sample(frac=0.2)
Sample2 = df1.sample(frac=0.2)
Sample3 = df1.sample(frac=0.2)

Sample1.to_csv('Sample1.csv')
Sample2.to_csv('Sample2.csv')
Sample3.to_csv('Sample3.csv')
```

In [50]:
```
Sample1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 226982 entries, 1927557 to 1066360
Data columns (total 48 columns):
 #   Column                          Non-Null Count   Dtype  
---  ------                          --------------   -----  
 0   CRASH_DATE_x                    226982 non-null  int64  
 1   UNIT_TYPE                       226982 non-null  object 
 2   MAKE                            226982 non-null  object 
 3   MODEL                           226982 non-null  object 
 4   VEHICLE_DEFECT                  226982 non-null  object 
 5   VEHICLE_TYPE                    226982 non-null  object 
 6   VEHICLE_USE                     226982 non-null  object 
 7   MANEUVER                        226982 non-null  object 
 8   OCCUPANT_CNT                    226982 non-null  float64
 9   CRASH_DATE_y                    226982 non-null  object 
 10  POSTED_SPEED_LIMIT              226982 non-null  int64  
 11  TRAFFIC_CONTROL_DEVICE          226982 non-null  object 
 12  DEVICE_CONDITION                226982 non-null  object 
 13  WEATHER_CONDITION               226982 non-null  object 
 14  LIGHTING_CONDITION              226982 non-null  object 
 15  FIRST_CRASH_TYPE                226982 non-null  object 
 16  TRAFFICWAY_TYPE                 226982 non-null  object 
 17  ALIGNMENT                       226982 non-null  object 
 18  ROADWAY_SURFACE_COND            226982 non-null  object 
 19  ROAD_DEFECT                     226982 non-null  object 
 20  REPORT_TYPE                     226982 non-null  object 
 21  CRASH_TYPE                      226982 non-null  object 
 22  DAMAGE                          226982 non-null  object 
 23  PRIM_CONTRIBUTORY_CAUSE         226982 non-null  object 
 24  SEC_CONTRIBUTORY_CAUSE          226982 non-null  object 
 25  BEAT_OF_OCCURRENCE              226982 non-null  float64
 26  NUM_UNITS                       226982 non-null  int64  
 27  MOST_SEVERE_INJURY              226982 non-null  object 
 28  INJURIES_TOTAL                  226982 non-null  float64
 29  INJURIES_FATAL                  226982 non-null  float64
 30  INJURIES_INCAPACITATING         226982 non-null  float64
 31  INJURIES_NON_INCAPACITATING     226982 non-null  float64
 32  INJURIES_REPORTED_NOT_EVIDENT   226982 non-null  float64
 33  INJURIES_NO_INDICATION          226982 non-null  float64
 34  INJURIES_UNKNOWN                226982 non-null  float64
 35  CRASH_HOUR                      226982 non-null  int64  
 36  CRASH_DAY_OF_WEEK               226982 non-null  int64  
 37  CRASH_MONTH                     226982 non-null  int64  
 38  LATITUDE                        226982 non-null  float64
 39  LONGITUDE                       226982 non-null  float64
 40  PERSON_ID                       226982 non-null  object 
 41  PERSON_TYPE                     226982 non-null  object 
 42  CRASH_DATE                      226982 non-null  object 
 43  SEX                             226982 non-null  object 
 44  SAFETY_EQUIPMENT                226982 non-null  object 
 45  AIRBAG_DEPLOYED                 226982 non-null  object 
 46  EJECTION                        226982 non-null  object 
 47  INJURY_CLASSIFICATION           226982 non-null  object 
```

```
dtypes: float64(11), int64(6), object(31)
memory usage: 84.9+ MB
```

In [ ]: