Advanced Analytics & Dashboard Design

Author Angela North

Exercise 6.1: Sourcing Open Data

Import Libraries

In [1]:

```
import pandas as pd
import numpy as np
import numbers
import chart_studio
import plotly
from plotly.offline import init_notebook_mode, iplot
import chart_studio.plotly as py
import plotly.graph_objs as go
from plotly import tools
import folium
# from folium import plugins

init_notebook_mode(connected=True)
```

Load Data set and Set data frame

In [2]:

```
gun_violence_df = pd.read_csv('gun-violence-data_01-2013_03-2018.csv')
```

Glimpse of Data

In [3]:

```
# head of data set
gun_violence_df.head()
```

Out[3]:

	incident_id	date	state	city_or_county	address	n_killed	n_injured	
0	461105	2013-01-01	Pennsylvania	Mckeesport	1506 Versailles Avenue and Coursin Street	0	4	http://www.gunvic
1	460726	2013-01-01	California	Hawthorne	13500 block of Cerise Avenue	1	3	http://www.gunvio
2	478855	2013-01-01	Ohio	Lorain	1776 East 28th Street	1	3	http://www.gunvio
3	478925	2013-01-05	Colorado	Aurora	16000 block of East Ithaca Place	4	0	http://www.gunvio
4	478959	2013-01-07	North Carolina	Greensboro	307 Mourning Dove Terrace	2	2	http://www.gunvio

5 rows × 29 columns

In [4]:
```
# Last values of the data set
gun_violence_df.tail()
```

Out[4]:

	incident_id	date	state	city_or_county	address	n_killed	n_injured	
239672	1083142	2018-03-31	Louisiana	Rayne	North Riceland Road and Highway 90	0	0	http://www.gu
239673	1083139	2018-03-31	Louisiana	Natchitoches	247 Keyser Ave	1	0	http://www.gu
239674	1083151	2018-03-31	Louisiana	Gretna	1300 block of Cook Street	0	1	http://www.gu
239675	1082514	2018-03-31	Texas	Houston	12630 Ashford Point Dr	1	0	http://www.gu
239676	1081940	2018-03-31	Maine	Norridgewock	434 Skowhegan Rd	2	0	http://www.gu

5 rows × 29 columns

Statistical Overview of the Data

In [5]:
```
gun_violence_df.describe() ##describes only numeric data
```

Out[5]:

	incident_id	n_killed	n_injured	congressional_district	latitude	lo
count	2.396770e+05	239677.000000	239677.000000	227733.000000	231754.000000	231754
mean	5.593343e+05	0.252290	0.494007	8.001265	37.546598	-89
std	2.931287e+05	0.521779	0.729952	8.480835	5.130763	14
min	9.211400e+04	0.000000	0.000000	0.000000	19.111400	-171
25%	3.085450e+05	0.000000	0.000000	2.000000	33.903400	-94
50%	5.435870e+05	0.000000	0.000000	5.000000	38.570600	-86
75%	8.172280e+05	0.000000	1.000000	10.000000	41.437375	-80
max	1.083472e+06	50.000000	53.000000	53.000000	71.336800	97

The information regarding the numerical columns of the statistics on gun violence is described in the table that was presented earlier. Because the information is only provided for the numeric columns, and there is no information provided about the data that is missing, we have developed a more in-depth tool that will describe the information for all of the attributes below.

Check for Missing Data

In [6]:

```
# Function to describe more information for all the attributes
def brief(data):

    df = data.copy()

    print("This dataset has {} Rows {} Attributes".format(df.shape[0],df.shape
    print('\n')

    real_valued = {}
    symbolics = {}


    for i,col in enumerate(df.columns, 1):
        Missing = len(df[col]) - df[col].count()

        counter = 0
        for val in df[col].dropna():
            if isinstance(val, numbers.Number):
                    counter += 1

        if counter != len(df[col].dropna()):
            arity = len(df[col].dropna().unique())
            symbolics[i] = [i, col, Missing, arity]
        else:
            Mean, Median, Sdev, Min, Max = df[col].mean(), df[col].median(), d
            real_valued[i] =  [i, col, Missing, Mean, Median, Sdev, Min, Max]


    #Create array containing list of real valued
    real_valued_array = [real_valued[keys] for keys in real_valued.keys()]
    real_valued_transformed = np.array(real_valued_array).T

    symbolic_array = [symbolics[keys] for keys in symbolics.keys()]
    symbolic_transformed = np.array(symbolic_array).T

    # return symbolic_transformed
    real_cols = ['Attribute_ID', 'Attribute_Name', 'Missing', 'Mean', 'Median'
    sym_cols = ['Attribute_ID', 'Attribute_Name', 'Missing','arity']



    index = range(1, len(real_valued.keys())+1)
    real_val_df = pd.DataFrame(data={unit[0]:unit[1] for unit in zip(real_cols


    index_sym = range(1, len(symbolics.keys())+1)
    sym_val_df = pd.DataFrame(data={unit[0]:unit[1] for unit in zip(sym_cols,

    text = ("real valued attributes" + "\n" + "---------------------"
            + "\n" + str(real_val_df) + "\n"  + "non-real valued attributes"
            + "\n" + "-------------------" + "\n" + str(sym_val_df))

    return text
```

In [7]:

```
%time
print(brief(gun_violence_df))
```

```
CPU times: total: 0 ns
Wall time: 0 ns
This dataset has 239677 Rows 29 Attributes

real valued attributes
---------------------
   Attribute_ID               Attribute_Name Missing                 Mean  \
1             1                  incident_id       0    559334.3464037017   
2             6                     n_killed       0  0.25228953967214207   
3             7                    n_injured       0   0.4940065171042695   
4            10  incident_url_fields_missing       0                  0.0   
5            11       congressional_district   11944    8.001264638853394   
6            15                     latitude    7923    37.54659822311588   
7            17                    longitude    7923   -89.33834822915676   
8            18              n_guns_involved   99451   1.3724416299402393   
9            28         state_house_district   38772    55.44713172892661   
10           29        state_senate_district   32335   20.477110281563792   

      Median                Sdev       Min      Max  
1   543587.0    293128.684285221     92114  1083472  
2        0.0    0.52177887298012         0       50  
3        0.0  0.7299522740842754         0       53  
4        0.0                 0.0     False    False  
5        5.0   8.480834796700318       0.0     53.0  
6    38.5706   5.130763162136701   19.1114  71.3368  
7   -86.2496   14.35954557699743  -171.429  97.4331  
8        1.0   4.678202195031997       1.0    400.0  
9       47.0   42.04811689079994       1.0    901.0  
10      19.0   14.20455963079257       1.0     94.0  
non-real valued attributes
-------------------
   Attribute_ID           Attribute_Name Missing   arity
1             2                     date       0    1725
2             3                    state       0      51
3             4           city_or_county       0   12898
4             5                  address   16497  198037
5             8             incident_url       0  239677
6             9               source_url     468  213989
7            12               gun_stolen   99498     349
8            13                 gun_type   99451    2502
9            14 incident_characteristics     326   18126
10           16     location_description  197588   27595
11           19                    notes   81017  136652
12           20          participant_age   92298   18951
13           21    participant_age_group   42119     898
14           22       participant_gender   36362     873
15           23         participant_name  122253  113488
16           24 participant_relationship  223903     284
17           25       participant_status   27626    2150
18           26         participant_type   24863     259
19           27                  sources     609  217280
```

Based on the analysis presented above, you can deduce that certain properties, such as participant_name and participant_relationship, are missing almost as many values as the total number of records contained in the dataset.

In [8]:
```
gun_violence_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 239677 entries, 0 to 239676
Data columns (total 29 columns):
 #   Column                       Non-Null Count   Dtype  
---  ------                       --------------   -----  
 0   incident_id                  239677 non-null  int64  
 1   date                         239677 non-null  object 
 2   state                        239677 non-null  object 
 3   city_or_county               239677 non-null  object 
 4   address                      223180 non-null  object 
 5   n_killed                     239677 non-null  int64  
 6   n_injured                    239677 non-null  int64  
 7   incident_url                 239677 non-null  object 
 8   source_url                   239209 non-null  object 
 9   incident_url_fields_missing  239677 non-null  bool   
 10  congressional_district       227733 non-null  float64
 11  gun_stolen                   140179 non-null  object 
 12  gun_type                     140226 non-null  object 
 13  incident_characteristics     239351 non-null  object 
 14  latitude                     231754 non-null  float64
 15  location_description         42089 non-null   object 
 16  longitude                    231754 non-null  float64
 17  n_guns_involved              140226 non-null  float64
 18  notes                        158660 non-null  object 
 19  participant_age              147379 non-null  object 
 20  participant_age_group        197558 non-null  object 
 21  participant_gender           203315 non-null  object 
 22  participant_name             117424 non-null  object 
 23  participant_relationship     15774 non-null   object 
 24  participant_status           212051 non-null  object 
 25  participant_type             214814 non-null  object 
 26  sources                      239068 non-null  object 
 27  state_house_district         200905 non-null  float64
 28  state_senate_district        207342 non-null  float64
dtypes: bool(1), float64(6), int64(3), object(19)
memory usage: 51.4+ MB
```

I contribute further to the analysis that was done earlier by providing additional information above. Given the facts shown above, it is very evident that some of the data will require some sort of cleaning.

Data Cleaning

In [9]:

```
# added important missing data point found in the description on Kaggle
missing =  ['sban_1', '2017-10-01', 'Nevada', 'Las Vegas', 'Mandalay Bay 3950 
            '-115.171667', 47, 'Route 91 Harvest Festiva; concert, open fire fr
gun_violence_df.loc[len(gun_violence_df)] = missing

print(gun_violence_df.shape)
drop_columns = gun_violence_df.columns[gun_violence_df.apply(lambda col: col.is
gun_violence_filtered = gun_violence_df.drop(drop_columns, axis=1)
print(gun_violence_filtered.shape)
print('Dropped Columns:', list(drop_columns))
```

```
(239678, 29)
(239678, 26)
Dropped Columns: ['location_description', 'participant_name', 'participant_re
lationship']
```

In [ ]: