Investigate_a_Dataset

June 27, 2018

1 Project: Investigate FBI GUN DATASET

1.1 Table of Contents

Introduction

2 The provided data set has two different files, one with census data for each state and other with the firearms details for each state. Census data has over 60 parameter for each state for a particular year

3 Therefore the common fields between two sheets is the state and the year. We have to analyze this data set using this fields with more interest to give the findings about the firearm data of a particular state for a month or year or so..

3.1 Posing Questions

4 1.What census variable or fact value is most associated with high gun per capita per state? Ceusus data includes state as variable, and there are 65 differnt census measurement as value of Fact.

5 2.Which states have had the highest growth and the lowest growth in gun registrations from Apr 2010 to Jul 2016?

6 3.What is the overall trend of gun purchases by year or by year and month?

In [55]:
```
# Use this cell to set up import statements for all of the packages that you
#   plan to use.
```

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
# Remember to include a 'magic word' so that your visualizations are plotted
#   inline with the notebook. See this page for more:
#   http://ipython.readthedocs.io/en/stable/interactive/magics.html
% matplotlib inline
census = pd.read_csv("US_Census_Data.csv")
gun = pd.read_excel("gun_data.xlsx")
```

Data Wrangling

Tip: In this section of the report, you will load in the data, check for cleanliness, and then trim and clean your dataset for analysis. Make sure that you document your steps carefully and justify your cleaning decisions.

6.0.1 General Properties

```
In [56]: # Load your data and print out a few lines. Perform operations to inspect data

         #   types and look for instances of missing or possibly errant data.

         # Step 1 Will check the data types of columns in gun data

         gun.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12485 entries, 0 to 12484
Data columns (total 27 columns):
month                        12485 non-null object
state                        12485 non-null object
permit                       12461 non-null float64
permit_recheck               1100 non-null float64
handgun                      12465 non-null float64
long_gun                     12466 non-null float64
other                        5500 non-null float64
multiple                     12485 non-null int64
admin                        12462 non-null float64
prepawn_handgun              10542 non-null float64
prepawn_long_gun             10540 non-null float64
prepawn_other                5115 non-null float64
redemption_handgun           10545 non-null float64
redemption_long_gun          10544 non-null float64
redemption_other             5115 non-null float64
returned_handgun             2200 non-null float64
returned_long_gun            2145 non-null float64
returned_other               1815 non-null float64
rentals_handgun              990 non-null float64
rentals_long_gun             825 non-null float64
```

```
private_sale_handgun         2750 non-null float64
private_sale_long_gun        2750 non-null float64
private_sale_other           2750 non-null float64
return_to_seller_handgun     2475 non-null float64
return_to_seller_long_gun    2750 non-null float64
return_to_seller_other       2255 non-null float64
totals                       12485 non-null int64
dtypes: float64(23), int64(2), object(2)
memory usage: 2.6+ MB
```

6.0.2 Based on the above info() function we could note that there are 12485 rows. There are no rows with all NAN values since there are several columns with 12485 entries.

6.0.3 There are lot of missing values in many columns for example, 'permit_recheck', 'permit', 'other', 'rentals_handgun', 'rentals_long_gun', 'private_sale_handgun', 'private_sale_long_gun', 'private_sale_other'. The missing values are to be replaced with mean per column before analyzing the data set.

6.1 One more important point to be noted is that the number of guns should be a integer not a float data_type since gun count be in decimal values like 3.5 or 4.5 etc. The data type should be converted to integer from float.

In [57]:
```
# Lets Check for any duplicated rows in census and gun data in below code;

dup_gun = sum(gun.duplicated())
dup_census = sum(census.duplicated())
print("there are "+ dup_gun.astype(str) +' duplicate rows in gun data')
print("there are "+ dup_census.astype(str) +' duplicate rows in census data')
```

```
there are 0 duplicate rows in gun data
there are 3 duplicate rows in census data
```

6.1.1 From the above result we could note that there are no duplicate rows in gun_data whereas there are 3 duplicated rows in census data. So we will drop these rows during data cleaning steps.

In [58]:
```
# Lets check with data types in census dataset also:

census.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 85 entries, 0 to 84
Data columns (total 52 columns):
Fact               80 non-null object
Fact Note          28 non-null object
Alabama            65 non-null object
Alaska             65 non-null object
```

```
Arizona           65 non-null object
Arkansas          65 non-null object
California        65 non-null object
Colorado          65 non-null object
Connecticut       65 non-null object
Delaware          65 non-null object
Florida           65 non-null object
Georgia           65 non-null object
Hawaii            65 non-null object
Idaho             65 non-null object
Illinois          65 non-null object
Indiana           65 non-null object
Iowa              65 non-null object
Kansas            65 non-null object
Kentucky          65 non-null object
Louisiana         65 non-null object
Maine             65 non-null object
Maryland          65 non-null object
Massachusetts     65 non-null object
Michigan          65 non-null object
Minnesota         65 non-null object
Mississippi       65 non-null object
Missouri          65 non-null object
Montana           65 non-null object
Nebraska          65 non-null object
Nevada            65 non-null object
New Hampshire     65 non-null object
New Jersey        65 non-null object
New Mexico        65 non-null object
New York          65 non-null object
North Carolina    65 non-null object
North Dakota      65 non-null object
Ohio              65 non-null object
Oklahoma          65 non-null object
Oregon            65 non-null object
Pennsylvania      65 non-null object
Rhode Island      65 non-null object
South Carolina    65 non-null object
South Dakota      65 non-null object
Tennessee         65 non-null object
Texas             65 non-null object
Utah              65 non-null object
Vermont           65 non-null object
Virginia          65 non-null object
Washington        65 non-null object
West Virginia     65 non-null object
Wisconsin         65 non-null object
Wyoming           65 non-null object
```

```
dtypes: object(52)
memory usage: 34.6+ KB
```

```
In [59]: # lets just have a look at what kind of data is there in census data for better underst

         census.head(3)
```

```
Out[59]:                                                  Fact Fact Note    Alabama  \
         0       Population estimates, July 1, 2016,  (V2016)       NaN  4,863,300
         1  Population estimates base, April 1, 2010,  (V2...       NaN  4,780,131
         2  Population, percent change - April 1, 2010 (es...       NaN      1.70%

             Alaska    Arizona   Arkansas  California   Colorado Connecticut Delaware  \
         0  741,894  6,931,071  2,988,248  39,250,017  5,540,545   3,576,452  952,065
         1  710,249  6,392,301  2,916,025  37,254,522  5,029,324   3,574,114  897,936
         2    4.50%      8.40%      2.50%       5.40%     10.20%       0.10%    6.00%

             ...    South Dakota Tennessee       Texas       Utah  Vermont   Virginia  \
         0   ...          865454   6651194  27,862,596  3,051,217  624,594  8,411,808
         1   ...          814195   6346298  25,146,100  2,763,888  625,741  8,001,041
         2   ...           0.063     0.048      10.80%     10.40%   -0.20%      5.10%

           Washington West Virginia  Wisconsin  Wyoming
         0  7,288,000     1,831,102  5,778,708  585,501
         1  6,724,545     1,853,011  5,687,289  563,767
         2      8.40%        -1.20%      1.60%    3.90%

         [3 rows x 52 columns]
```

```
In [60]: #WHILE WE OBSERVED WE COULD NOTE THAT FACT NOTE VALUES WERE NAN FOR 3 ROWS ALSO, SO JUS

         census['Fact Note'].isnull().sum()
```

```
Out[60]: 57
```

6.1.2 Based on the above results of info(), head() and isnull(), we can see the data type of these state columns are string, but they're acutally numbers and some columns with percentage.

6.1.3 The string data is to be converted into numeric data type for grouping and calculating for data exploration.

6.1.4 Out of 65 row data values there are 57 nan values for fact note columns, which should be considered for removing to reduce the missing values for census data.

6.1.5 Data Cleaning!!! —- Lets proceed with data cleaning based on basic observation done until now.

```
In [61]: # Lets replace NAN field with mean of each column data for gun data
```

```
gun = gun.fillna(gun.mean(), axis = 0, inplace = True)
```

```
In [62]: # now testing for above step whether we have done replacement correctle

         gun.isnull().sum().sum()

Out[62]: 0
```

6.1.6 The above cell has returned us 0, which means the that gun data has been replaced with mean values for NAN values for respective columns

```
In [63]: ### lets drop duplicate rows in census data

         census.drop_duplicates(inplace=True)

In [64]: ## now checking for duplicated rows in census data

         sum(census.duplicated())

Out[64]: 0
```

6.1.7 We have now removed the three duplicate rows in census data set, evidence is that we have got 0 output for the above code

```
In [65]: # We could note that fact note column had too many missing values , its better to drop
         census.drop('Fact Note', axis = 1, inplace = True)

In [66]: #after removing the above column, census data should not be consisting any further nan
         #But found there were 17 rows with no data and I missed that initially. NOw using tail
         census.tail(17)

         #SO i would like to drop rows from 65 to end for clear data exploaration

Out[66]:                                                   Fact Alabama Alaska Arizona  \
         65                                                 NaN     NaN    NaN     NaN
         66  NOTE: FIPS Code values are enclosed in quotes ...     NaN    NaN     NaN
         68                                         Value Notes     NaN    NaN     NaN
         69                                                   1     NaN    NaN     NaN
         71                                          Fact Notes     NaN    NaN     NaN
         72                                                 (a)     NaN    NaN     NaN
         73                                                 (b)     NaN    NaN     NaN
         74                                                 (c)     NaN    NaN     NaN
         76                                         Value Flags     NaN    NaN     NaN
         77                                                   -     NaN    NaN     NaN
         78                                                   D     NaN    NaN     NaN
         79                                                   F     NaN    NaN     NaN
         80                                                  FN     NaN    NaN     NaN
         81                                                 NaN     NaN    NaN     NaN
         82                                                   S     NaN    NaN     NaN
```

```
83                                              X     NaN    NaN     NaN
84                                              Z     NaN    NaN     NaN

   Arkansas California Colorado Connecticut Delaware Florida   ...    \
65      NaN        NaN      NaN         NaN      NaN     NaN   ...
66      NaN        NaN      NaN         NaN      NaN     NaN   ...
68      NaN        NaN      NaN         NaN      NaN     NaN   ...
69      NaN        NaN      NaN         NaN      NaN     NaN   ...
71      NaN        NaN      NaN         NaN      NaN     NaN   ...
72      NaN        NaN      NaN         NaN      NaN     NaN   ...
73      NaN        NaN      NaN         NaN      NaN     NaN   ...
74      NaN        NaN      NaN         NaN      NaN     NaN   ...
76      NaN        NaN      NaN         NaN      NaN     NaN   ...
77      NaN        NaN      NaN         NaN      NaN     NaN   ...
78      NaN        NaN      NaN         NaN      NaN     NaN   ...
79      NaN        NaN      NaN         NaN      NaN     NaN   ...
80      NaN        NaN      NaN         NaN      NaN     NaN   ...
81      NaN        NaN      NaN         NaN      NaN     NaN   ...
82      NaN        NaN      NaN         NaN      NaN     NaN   ...
83      NaN        NaN      NaN         NaN      NaN     NaN   ...
84      NaN        NaN      NaN         NaN      NaN     NaN   ...

   South Dakota Tennessee Texas Utah Vermont Virginia Washington  \
65          NaN       NaN   NaN  NaN     NaN      NaN        NaN
66          NaN       NaN   NaN  NaN     NaN      NaN        NaN
68          NaN       NaN   NaN  NaN     NaN      NaN        NaN
69          NaN       NaN   NaN  NaN     NaN      NaN        NaN
71          NaN       NaN   NaN  NaN     NaN      NaN        NaN
72          NaN       NaN   NaN  NaN     NaN      NaN        NaN
73          NaN       NaN   NaN  NaN     NaN      NaN        NaN
74          NaN       NaN   NaN  NaN     NaN      NaN        NaN
76          NaN       NaN   NaN  NaN     NaN      NaN        NaN
77          NaN       NaN   NaN  NaN     NaN      NaN        NaN
78          NaN       NaN   NaN  NaN     NaN      NaN        NaN
79          NaN       NaN   NaN  NaN     NaN      NaN        NaN
80          NaN       NaN   NaN  NaN     NaN      NaN        NaN
81          NaN       NaN   NaN  NaN     NaN      NaN        NaN
82          NaN       NaN   NaN  NaN     NaN      NaN        NaN
83          NaN       NaN   NaN  NaN     NaN      NaN        NaN
84          NaN       NaN   NaN  NaN     NaN      NaN        NaN

   West Virginia Wisconsin Wyoming
65           NaN       NaN     NaN
66           NaN       NaN     NaN
68           NaN       NaN     NaN
69           NaN       NaN     NaN
71           NaN       NaN     NaN
72           NaN       NaN     NaN
```

```
73            NaN        NaN      NaN
74            NaN        NaN      NaN
76            NaN        NaN      NaN
77            NaN        NaN      NaN
78            NaN        NaN      NaN
79            NaN        NaN      NaN
80            NaN        NaN      NaN
81            NaN        NaN      NaN
82            NaN        NaN      NaN
83            NaN        NaN      NaN
84            NaN        NaN      NaN

[17 rows x 51 columns]
```

In [67]:
```
#I will select data from 0:64 which gives me required data only... Or i could have gone

census = census[0:65]
```

In [68]:
```
#Now lets check for NAN values in processed census dataset
census.isnull().sum().sum()
```

Out[68]:
```
0
```

6.1.8 Finally I was able to retain the data required for our analysis. As above result indicated that there are no more nan values in census data. Lets go ahead with conversion of data columns into required data type in gun data and census data. Here we go

In [69]:
```
states = []
for state in census.columns:
    states.append(state)
states.remove('Fact')
states
```

Out[69]:
```
['Alabama',
 'Alaska',
 'Arizona',
 'Arkansas',
 'California',
 'Colorado',
 'Connecticut',
 'Delaware',
 'Florida',
 'Georgia',
 'Hawaii',
 'Idaho',
 'Illinois',
 'Indiana',
 'Iowa',
 'Kansas',
```

```
 'Kentucky',
 'Louisiana',
 'Maine',
 'Maryland',
 'Massachusetts',
 'Michigan',
 'Minnesota',
 'Mississippi',
 'Missouri',
 'Montana',
 'Nebraska',
 'Nevada',
 'New Hampshire',
 'New Jersey',
 'New Mexico',
 'New York',
 'North Carolina',
 'North Dakota',
 'Ohio',
 'Oklahoma',
 'Oregon',
 'Pennsylvania',
 'Rhode Island',
 'South Carolina',
 'South Dakota',
 'Tennessee',
 'Texas',
 'Utah',
 'Vermont',
 'Virginia',
 'Washington',
 'West Virginia',
 'Wisconsin',
 'Wyoming']
```

In [70]:
```
# I have extracted list of states from census data frame and excluded the Fact as its n
#Now we will remove all non digit characters and convert it to float as they're to be n

for state in states:
    #below expression to remove all non digit characters
    census[state].replace(regex=True,inplace=True,to_replace=r'\D',value=r'')
    #we will convert data type to float and lets ignore the nan values
    census[state] = pd.to_numeric(census[state], downcast = 'float', errors = 'ignore')
#lets check the data types in census data using dtypes()

census.dtypes
```

Out[70]:
```
Fact                 object
Alabama             float32
```

```
Alaska            float32
Arizona           float32
Arkansas          float32
California        float32
Colorado          float32
Connecticut       float32
Delaware          float32
Florida           float32
Georgia           float32
Hawaii            float32
Idaho             float32
Illinois          float32
Indiana           float32
Iowa              float32
Kansas            float32
Kentucky          float32
Louisiana         float32
Maine             float32
Maryland          float32
Massachusetts     float32
Michigan          float32
Minnesota         float32
Mississippi       float32
Missouri          float32
Montana           float32
Nebraska          float32
Nevada            float32
New Hampshire     float32
New Jersey        float32
New Mexico        float32
New York          float32
North Carolina    float32
North Dakota      float32
Ohio              float32
Oklahoma          float32
Oregon            float32
Pennsylvania      float32
Rhode Island      float32
South Carolina    float32
South Dakota      float32
Tennessee         float32
Texas             float32
Utah              float32
Vermont           float32
Virginia          float32
Washington        float32
West Virginia     float32
Wisconsin         float32
```

```
Wyoming            float32
dtype: object
```

6.2 Hurray, we've converted all the columns into float except the column with label Fact..

In [71]:
```
#let's separate the year and month values in the gun dataset with column label month

gun['year'] = gun['month'].apply(lambda x: x.split('-')[0]).astype(int)
gun['months'] = gun['month'].apply(lambda x: x.split('-')[1]).astype(int)
```

In [72]:
```
# a small check to see if we have added the necessary year and months columns

gun.head(3)
```

Out[72]:
```
     month    state   permit  permit_recheck  handgun  long_gun  other  \
0  2017-09  Alabama  16717.0             0.0   5734.0    6320.0  221.0
1  2017-09   Alaska    209.0             2.0   2320.0    2930.0  219.0
2  2017-09  Arizona   5069.0           382.0  11063.0    7946.0  920.0

   multiple  admin  prepawn_handgun   ...    rentals_long_gun  \
0       317    0.0             15.0   ...                 0.0
1       160    0.0              5.0   ...                 0.0
2       631    0.0             13.0   ...                 0.0

   private_sale_handgun  private_sale_long_gun  private_sale_other  \
0                   9.0                   16.0                 3.0
1                  17.0                   24.0                 1.0
2                  38.0                   12.0                 2.0

   return_to_seller_handgun  return_to_seller_long_gun  \
0                       0.0                        0.0
1                       0.0                        0.0
2                       0.0                        0.0

   return_to_seller_other  totals  year  months
0                     3.0   32019  2017       9
1                     0.0    6303  2017       9
2                     0.0   28394  2017       9

[3 rows x 29 columns]
```

In [73]:
```
# we have to convert all data type of all columns except for month, state... so below i
cols = []
for column in gun.columns:
    cols.append(column)
del cols[:2]
cols
#since we need to remove first two columns from the cols list, let do that
```

Out[73]:
```
['permit',
 'permit_recheck',
 'handgun',
 'long_gun',
 'other',
 'multiple',
 'admin',
 'prepawn_handgun',
 'prepawn_long_gun',
 'prepawn_other',
 'redemption_handgun',
 'redemption_long_gun',
 'redemption_other',
 'returned_handgun',
 'returned_long_gun',
 'returned_other',
 'rentals_handgun',
 'rentals_long_gun',
 'private_sale_handgun',
 'private_sale_long_gun',
 'private_sale_other',
 'return_to_seller_handgun',
 'return_to_seller_long_gun',
 'return_to_seller_other',
 'totals',
 'year',
 'months']
```

In [74]:
```
# now lets convert data type of these columns into integers and using dtypes() lets ver
gun[cols] = gun[cols].applymap(np.int64)

gun.dtypes
#after this step we could note that all data types are integers except month and state
```

Out[74]:
```
month                      object
state                      object
permit                      int64
permit_recheck              int64
handgun                     int64
long_gun                    int64
other                       int64
multiple                    int64
admin                       int64
prepawn_handgun             int64
prepawn_long_gun            int64
prepawn_other               int64
redemption_handgun          int64
redemption_long_gun         int64
```

```
redemption_other              int64
returned_handgun              int64
returned_long_gun             int64
returned_other                int64
rentals_handgun               int64
rentals_long_gun              int64
private_sale_handgun          int64
private_sale_long_gun         int64
private_sale_other            int64
return_to_seller_handgun      int64
return_to_seller_long_gun     int64
return_to_seller_other        int64
totals                        int64
year                          int64
months                        int64
dtype: object
```

Exploratory Data Analysis ### Finally we have done with data cleaning based on our view, now let's move on to data analysis :)

6.2.1 Research Question 1 :What census data is most associated with high gun per capita?

6.2.2 Scenario: In order to calculate the gun per capita, the gun totals and population for each state is to be fetched firstly(which means we have to combine to both gun and census data). On my view of table I could note that state in census data is divided by 50 columns, however, in gun data state is only one column which has 46 different state value.

6.2.3 Further more, in census data-fact column, there are all kind of census measurement for state, for example: population estimate,July 1st,2016(v2016) and popluation estimate base, april1,2016(v2016). These two variables can be used for comparing data on 2010 and 2016, and fact column will be used to analyze the association between gun per capital and these measurements.

6.3 WORK AROUND:

6.3.1 As the data present in two different ways on these two datasets., firstly we will need to transpose the census data so that state values are in single columns to that of gun data. Then we can summarize the data based on gun totals by year 2010 and 2016 as we have population data for the same. SO we will merge gun data with transposed data based on state column. We will calculate the gun per capita and list the highest 5 per capital states in 2010 and 2016.

6.3.2 We will plot a scatter plot against fact value to find the associaton between gun per capital and fact value.

In [75]:
```
#Transposing census data and removing the index on fact

census.set_index('Fact', inplace=True)
census_T = census.T.reset_index()
```

```
#lets rename the column to state from index to have a better picture of data.
census_T.rename(columns={'index':'state'}, inplace = True)
census_T.head()
```

```
Out[75]: Fact       state  Population estimates, July 1, 2016,  (V2016)  \
0       Alabama                                         4863300.0
1        Alaska                                          741894.0
2       Arizona                                         6931071.0
3      Arkansas                                         2988248.0
4    California                                        39250016.0

Fact  Population estimates base, April 1, 2010,  (V2016)  \
0                                              4780131.0
1                                               710249.0
2                                              6392301.0
3                                              2916025.0
4                                             37254520.0

Fact  Population, percent change - April 1, 2010 (estimates base) to July 1, 2016,  (V2
0                                                 170.0
1                                                 450.0
2                                                 840.0
3                                                 250.0
4                                                 540.0

Fact  Population, Census, April 1, 2010  \
0                             4779736.0
1                              710231.0
2                             6392017.0
3                             2915918.0
4                            37253956.0

Fact  Persons under 5 years, percent, July 1, 2016,  (V2016)  \
0                                                 600.0
1                                                 730.0
2                                                 630.0
3                                                 640.0
4                                                 630.0

Fact  Persons under 5 years, percent, April 1, 2010  \
0                                             640.0
1                                             760.0
2                                             710.0
3                                             680.0
4                                             680.0

Fact  Persons under 18 years, percent, July 1, 2016,  (V2016)  \
0                                                2260.0
```

```
1                                                2520.0
2                                                2350.0
3                                                2360.0
4                                                2320.0

Fact  Persons under 18 years, percent, April 1, 2010  \
0                                           2370.0
1                                           2640.0
2                                           2550.0
3                                           2440.0
4                                           2500.0

Fact  Persons 65 years and over, percent,  July 1, 2016,  (V2016)    ...      \
0                                                1610.0              ...
1                                                1040.0              ...
2                                                1690.0              ...
3                                                1630.0              ...
4                                                1360.0              ...

Fact  All firms, 2012  Men-owned firms, 2012  Women-owned firms, 2012  \
0            374153.0               203604.0                 137630.0
1             68032.0                35402.0                  22141.0
2            499926.0               245243.0                 182425.0
3            231959.0               123158.0                  75962.0
4           3548449.0              1852580.0                1320085.0

Fact  Minority-owned firms, 2012  Nonminority-owned firms, 2012  \
0                        92219.0                       272651.0
1                        13688.0                        51147.0
2                       135313.0                       344981.0
3                        35982.0                       189029.0
4                      1619857.0                      1819107.0

Fact  Veteran-owned firms, 2012  Nonveteran-owned firms, 2012  \
0                       41943.0                      316984.0
1                        7953.0                       56091.0
2                       46780.0                      427582.0
3                       25915.0                      192988.0
4                      252377.0                     3176341.0

Fact  Population per square mile, 2010  Land area in square miles, 2010  \
0                                944.0                        5064533.0
1                                 12.0                       57064096.0
2                                563.0                       11359408.0
3                                 56.0                        5203548.0
4                               2391.0                       15577922.0

Fact  FIPS Code
```

```
0           1.0
1           2.0
2           4.0
3           5.0
4           6.0

[5 rows x 66 columns]
```

6.3.3 we have transposed data in census data so that we can combine with gun data.

In [76]:
```
#lets get data for only 2010 and 2016
gun_16 = gun[gun['year'] == 2016]
gun_10 = gun[gun['year'] == 2010]
gun_10.head()
```

Out[76]:
```
        month       state  permit  permit_recheck  handgun  long_gun  other  \
4455  2010-12     Alabama     413            1165    13978     24298    152
4456  2010-12      Alaska       0            1165     2553      3950     93
4457  2010-12     Arizona    2082            1165     9943      9814    219
4458  2010-12    Arkansas    2582            1165     5816     12455     62
4459  2010-12  California   24901            1165    24519     32100      0

      multiple  admin  prepawn_handgun   ...    rentals_long_gun  \
4455       569      0                6   ...                   0
4456       146      0                3   ...                   0
4457       431      0                6   ...                   0
4458       257      1                8   ...                   0
4459         0      2                0   ...                   0

      private_sale_handgun  private_sale_long_gun  private_sale_other  \
4455                    14                     11                   1
4456                    14                     11                   1
4457                    14                     11                   1
4458                    14                     11                   1
4459                    14                     11                   1

      return_to_seller_handgun  return_to_seller_long_gun  \
4455                         0                          0
4456                         0                          0
4457                         0                          0
4458                         0                          0
4459                         0                          0

      return_to_seller_other  totals  year  months
4455                       0   43266  2010      12
4456                       0    7036  2010      12
4457                       0   23942  2010      12
4458                       0   23821  2010      12
```

```
4459                       0   81522  2010      12

[5 rows x 29 columns]
```

In [77]:
```
#lets group data by state and gun total for 2010 and 2016

guntotal_16 = gun_16.groupby(['state'])['totals'].sum().reset_index()
guntotal_10 = gun_10.groupby(['state'])['totals'].sum().reset_index()
guntotal_16.head()
```

Out[77]:
```
        state   totals
0     Alabama   616947
1      Alaska    87647
2     Arizona   416279
3    Arkansas   266014
4  California  2377167
```

6.3.4 Now we have go gun totals for state wise...

In [78]:
```
#lets merge now 2010 and 2016 data ..

guntotal = guntotal_16.merge(guntotal_10, on = 'state', how = 'inner', suffixes=('_16',
guntotal.head()
```

Out[78]:
```
        state  totals_16  totals_10
0     Alabama     616947     308607
1      Alaska      87647      65909
2     Arizona     416279     206050
3    Arkansas     266014     191448
4  California    2377167     816399
```

In [79]:
```
#now we should to merge our totals data with census data...

result = guntotal.merge(census_T, on = 'state', how = 'inner')
result.head()
```

Out[79]:
```
        state  totals_16  totals_10  \
0     Alabama     616947     308607
1      Alaska      87647      65909
2     Arizona     416279     206050
3    Arkansas     266014     191448
4  California    2377167     816399

   Population estimates, July 1, 2016,  (V2016)  \
0                                    4863300.0
1                                     741894.0
2                                    6931071.0
3                                    2988248.0
4                                   39250016.0
```

```
   Population estimates base, April 1, 2010,  (V2016)  \
0                                          4780131.0
1                                           710249.0
2                                          6392301.0
3                                          2916025.0
4                                         37254520.0

   Population, percent change - April 1, 2010 (estimates base) to July 1, 2016,  (V2016
0                                              170.0
1                                              450.0
2                                              840.0
3                                              250.0
4                                              540.0

   Population, Census, April 1, 2010  \
0                          4779736.0
1                           710231.0
2                          6392017.0
3                          2915918.0
4                         37253956.0

   Persons under 5 years, percent, July 1, 2016,  (V2016)  \
0                                              600.0
1                                              730.0
2                                              630.0
3                                              640.0
4                                              630.0

   Persons under 5 years, percent, April 1, 2010  \
0                                          640.0
1                                          760.0
2                                          710.0
3                                          680.0
4                                          680.0

   Persons under 18 years, percent, July 1, 2016,  (V2016)    ...      \
0                                             2260.0         ...
1                                             2520.0         ...
2                                             2350.0         ...
3                                             2360.0         ...
4                                             2320.0         ...

   All firms, 2012  Men-owned firms, 2012  Women-owned firms, 2012  \
0         374153.0               203604.0                 137630.0
1          68032.0                35402.0                  22141.0
2         499926.0               245243.0                 182425.0
3         231959.0               123158.0                  75962.0
```

```
4       3548449.0             1852580.0              1320085.0

   Minority-owned firms, 2012  Nonminority-owned firms, 2012  \
0                     92219.0                       272651.0
1                     13688.0                        51147.0
2                    135313.0                       344981.0
3                     35982.0                       189029.0
4                   1619857.0                      1819107.0

   Veteran-owned firms, 2012  Nonveteran-owned firms, 2012  \
0                    41943.0                      316984.0
1                     7953.0                       56091.0
2                    46780.0                      427582.0
3                    25915.0                      192988.0
4                   252377.0                     3176341.0

   Population per square mile, 2010  Land area in square miles, 2010  \
0                             944.0                        5064533.0
1                              12.0                       57064096.0
2                             563.0                       11359408.0
3                              56.0                        5203548.0
4                            2391.0                       15577922.0

   FIPS Code
0        1.0
1        2.0
2        4.0
3        5.0
4        6.0

[5 rows x 68 columns]
```

6.3.5 We have got the data set result in most best way to analyse based on a state.. Lets now start calculating the descriptive statistics for this dataset.

In [80]:
```
#we are including a new column per capita for 2016 and 2010 :

result['Gun_Per_Capital_2016'] = result['totals_16']/result['Population estimates, July
result['Gun_Per_Capital_2010'] = result['totals_10']/result['Population estimates base,
```

In [81]:
```
#Let's display the to 5 highest per capital for 2010
result.nlargest(5,'Gun_Per_Capital_2010')
```

Out[81]:
```
            state  totals_16  totals_10  \
16       Kentucky    3676847    2385579
43           Utah     294907     553134
25        Montana     136337     101095
1          Alaska      87647      65909
```

```
47  West Virginia     242350     159550

    Population estimates, July 1, 2016,  (V2016)  \
16                                   4436974.0
43                                   3051217.0
25                                   1042520.0
1                                     741894.0
47                                   1831102.0

    Population estimates base, April 1, 2010,  (V2016)  \
16                                        4339344.0
43                                        2763888.0
25                                         989414.0
1                                          710249.0
47                                        1853011.0

    Population, percent change - April 1, 2010 (estimates base) to July 1, 2016,  (V201
16                                              220.0
43                                             1040.0
25                                              540.0
1                                               450.0
47                                              120.0

    Population, Census, April 1, 2010  \
16                          4339367.0
43                          2763885.0
25                           989415.0
1                            710231.0
47                          1852994.0

    Persons under 5 years, percent, July 1, 2016,  (V2016)  \
16                                              620.0
43                                              830.0
25                                              600.0
1                                               730.0
47                                              550.0

    Persons under 5 years, percent, April 1, 2010  \
16                                          650.0
43                                          950.0
25                                          630.0
1                                           760.0
47                                          560.0

    Persons under 18 years, percent, July 1, 2016,  (V2016)  \
16                                             2280.0
43                                             3020.0
25                                             2180.0
```

```
1                                              2520.0
47                                             2050.0

           ...           Women-owned firms, 2012  Minority-owned firms, 2012  \
16         ...                          106011.0                     27258.0
43         ...                           76269.0                     24423.0
25         ...                           35449.0                      5578.0
1          ...                           22141.0                     13688.0
47         ...                           39065.0                      5777.0

    Nonminority-owned firms, 2012  Veteran-owned firms, 2012  \
16                       296155.0                    33208.0
43                       218826.0                    18754.0
25                       102746.0                    11486.0
1                         51147.0                     7953.0
47                       104785.0                    12912.0

    Nonveteran-owned firms, 2012  Population per square mile, 2010  \
16                      282704.0                            1099.0
43                      219807.0                             336.0
25                       93393.0                              68.0
1                        56091.0                              12.0
47                       94960.0                             771.0

    Land area in square miles, 2010  FIPS Code  Gun_Per_Capital_2016  \
16                        3948634.0       21.0              0.828683
43                        8216962.0       49.0              0.096652
25                       14554580.0       30.0              0.130776
1                        57064096.0        2.0              0.118140
47                        2403821.0       54.0              0.132352

    Gun_Per_Capital_2010
16              0.549756
43              0.200129
25              0.102177
1               0.092797
47              0.086103

[5 rows x 70 columns]
```

In [82]:
```
#Let's display the to 5 highest per capital for 2016
result.nlargest(5,'Gun_Per_Capital_2016')
```

Out[82]:
```
            state  totals_16  totals_10  \
16       Kentucky    3676847    2385579
13        Indiana    1436725     345650
12       Illinois    1924070     695300
47  West Virginia     242350     159550
```

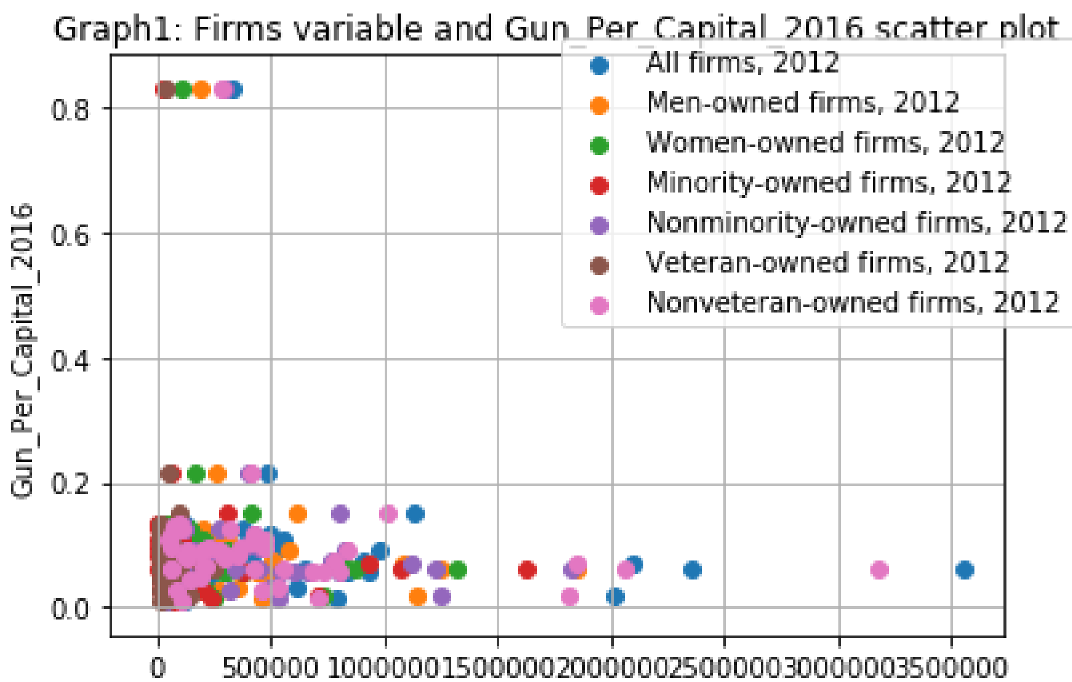
```
25       Montana     136337     101095

   Population estimates, July 1, 2016,  (V2016)  \
16                                    4436974.0
13                                    6633053.0
12                                   12801539.0
47                                    1831102.0
25                                    1042520.0

   Population estimates base, April 1, 2010,  (V2016)  \
16                                          4339344.0
13                                          6484136.0
12                                         12831574.0
47                                          1853011.0
25                                           989414.0

   Population, percent change - April 1, 2010 (estimates base) to July 1, 2016,  (V201
16                                              220.0
13                                              230.0
12                                               20.0
47                                              120.0
25                                              540.0

   Population, Census, April 1, 2010  \
16                         4339367.0
13                         6483802.0
12                        12830632.0
47                         1852994.0
25                          989415.0

   Persons under 5 years, percent, July 1, 2016,  (V2016)  \
16                                              620.0
13                                              640.0
12                                              600.0
47                                              550.0
25                                              600.0

   Persons under 5 years, percent, April 1, 2010  \
16                                         650.0
13                                         670.0
12                                         650.0
47                                         560.0
25                                         630.0

   Persons under 18 years, percent, July 1, 2016,  (V2016)  \
16                                             2280.0
13                                             2380.0
12                                             2290.0
```

```
47                                             2050.0
25                                             2180.0

            ...           Women-owned firms, 2012  Minority-owned firms, 2012  \
16          ...                          106011.0                     27258.0
13          ...                          162798.0                     61252.0
12          ...                          417500.0                    311684.0
47          ...                           39065.0                      5777.0
25          ...                           35449.0                      5578.0

    Nonminority-owned firms, 2012  Veteran-owned firms, 2012  \
16                       296155.0                    33208.0
13                       405090.0                    45174.0
12                       795129.0                    89110.0
47                       104785.0                    12912.0
25                       102746.0                    11486.0

    Nonveteran-owned firms, 2012  Population per square mile, 2010  \
16                      282704.0                             1099.0
13                      412543.0                              181.0
12                     1006885.0                             2311.0
47                       94960.0                              771.0
25                       93393.0                               68.0

    Land area in square miles, 2010  FIPS Code  Gun_Per_Capital_2016  \
16                        3948634.0       21.0              0.828683
13                        3582611.0       18.0              0.216601
12                        5551893.0       17.0              0.150300
47                        2403821.0       54.0              0.132352
25                       14554580.0       30.0              0.130776

    Gun_Per_Capital_2010
16              0.549756
13              0.053307
12              0.054187
47              0.086103
25              0.102177

[5 rows x 70 columns]
```

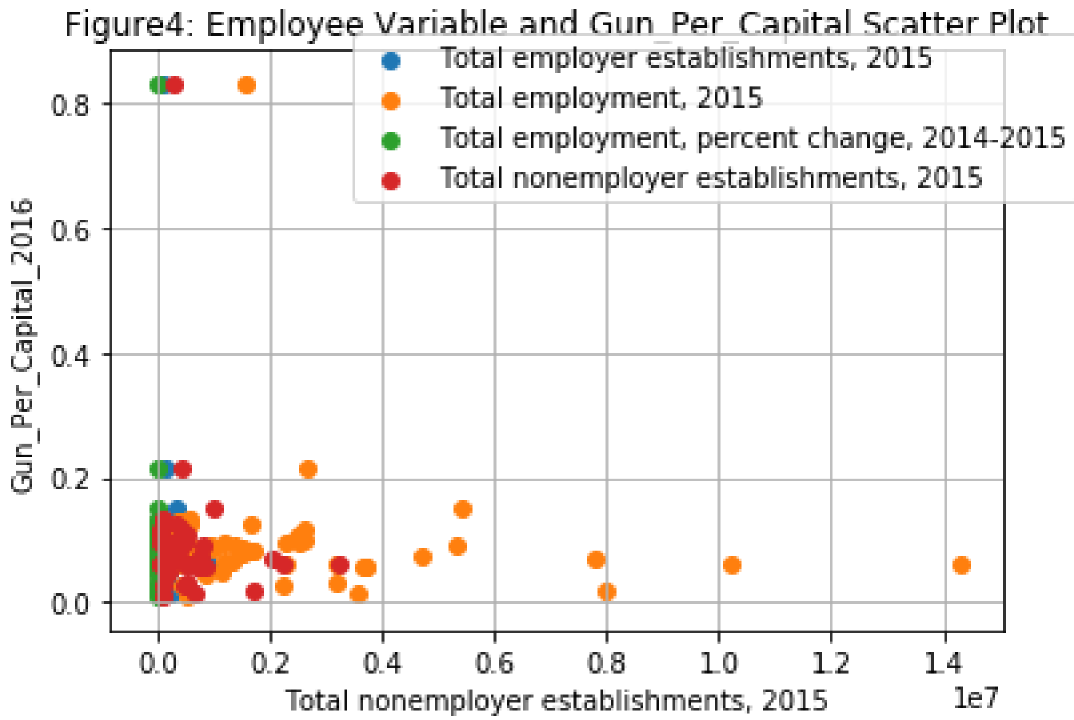
In [89]:
```
#lets drop non fact value from result to data frame fact
fact = result.drop(['Gun_Per_Capital_2010', 'state', 'FIPS Code', 'totals_16', 'totals_
```

In [91]:
```
#now lets visualize the insights by plotting a scatter plot for all fact variable
#Graph1: Firms related variable and Gun_Per_Capital
#wil use for loop to create a scatter plot for firms variable in single graph

for col in list(fact):
```

```
    if 'firms' in col:
        plt.scatter(fact[col], fact['Gun_Per_Capital_2016'], label = col)
plt.ylabel("Gun_Per_Capital_2016")
plt.title("Graph1: Firms variable and Gun_Per_Capital_2016 scatter plot")
plt.grid(True)
plt.legend(bbox_to_anchor = (1.1, 1.05))
plt.show()
```



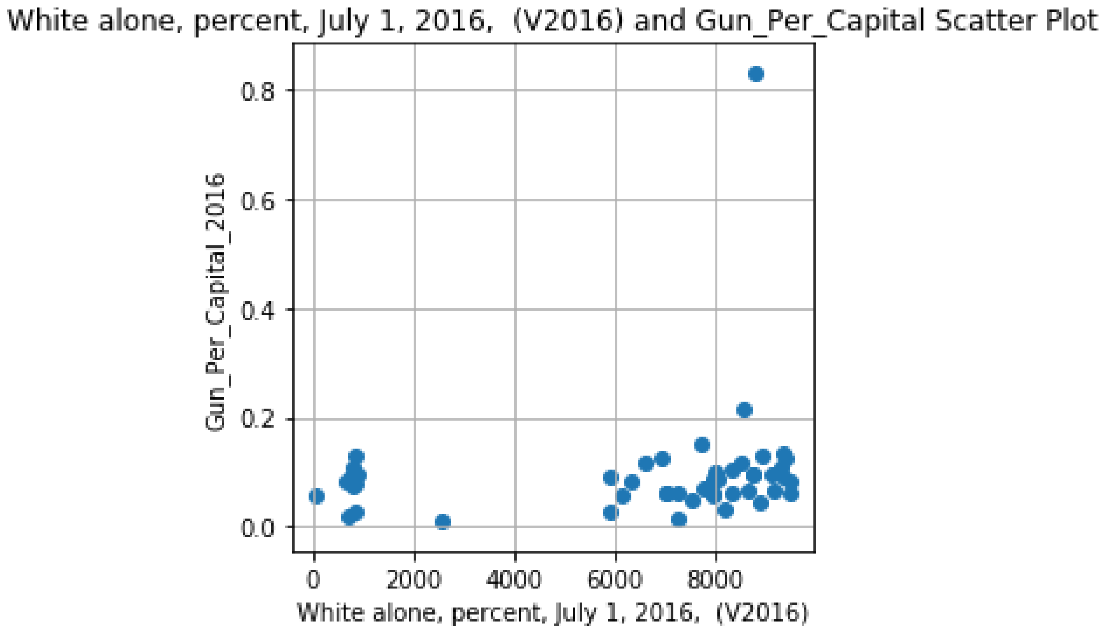

```
In [92]: #similarly graph2 with employment variables and GUn_per_capital
         for col in list(fact):
             if 'employ' in col:
                 plt.scatter(fact[col],fact['Gun_Per_Capital_2016'], label =col)
                 plt.xlabel(col)
         plt.ylabel("Gun_Per_Capital_2016")
         plt.title("Figure4: Employee Variable and Gun_Per_Capital Scatter Plot")
         plt.grid(True)
         plt.legend(bbox_to_anchor=(1.1, 1.05))
         plt.show()
```



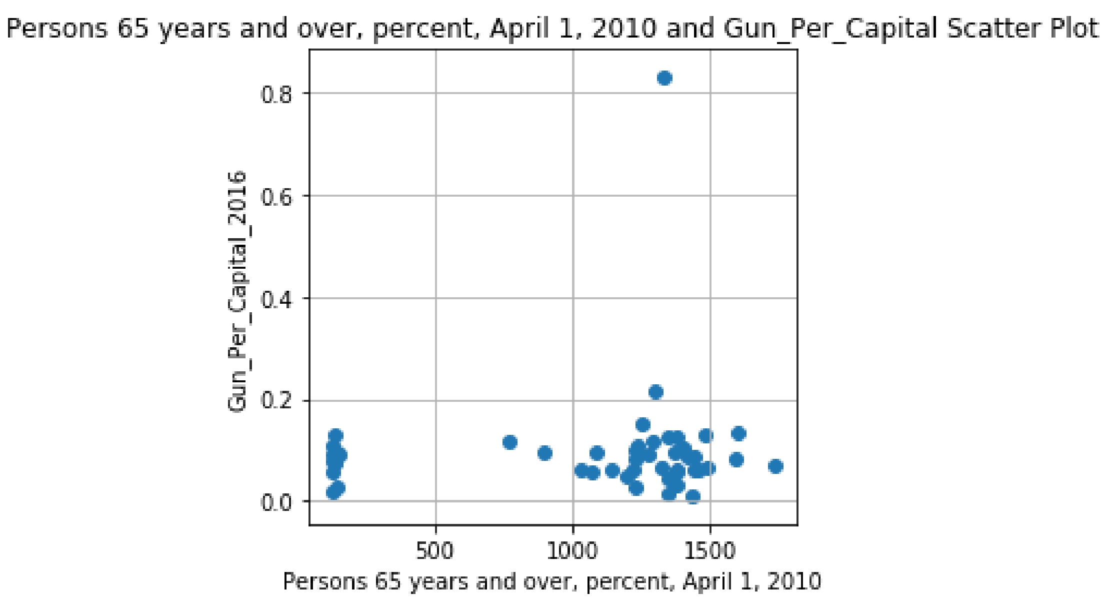
In [93]:
```
#from above two graphs we aren't able to distinguish any significant findings ie., a we
#so lets keep only 6 variables in our cols list and plot a scatter plot
imp =['White alone, percent, July 1, 2016,  (V2016)',
    'Persons 65 years and over, percent, April 1, 2010',
    'Owner-occupied housing unit rate, 2011-2015',
    'Asian alone, percent, July 1, 2016,  (V2016)',
    'Foreign born persons, percent, 2011-2015',
    'Median gross rent, 2011-2015']
```

In [94]:
```
#create scatter plot for all the fact variable in speparate figure, 6 figures
for col in imp:
    plt.figure(figsize=(4,4))
    print(col)
    plt.scatter(fact[col],fact['Gun_Per_Capital_2016'], label =col)
    plt.title(col+" and Gun_Per_Capital Scatter Plot")
    plt.ylabel("Gun_Per_Capital_2016")
    plt.xlabel(col)
    plt.grid(True)
    plt.show()
```

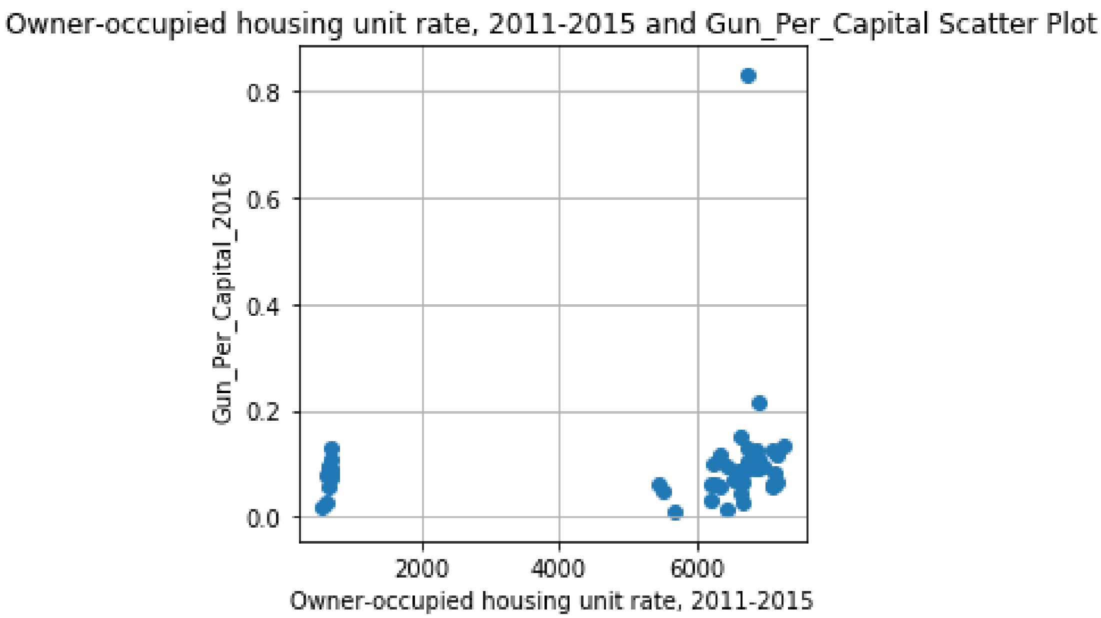
```
White alone, percent, July 1, 2016,  (V2016)
```



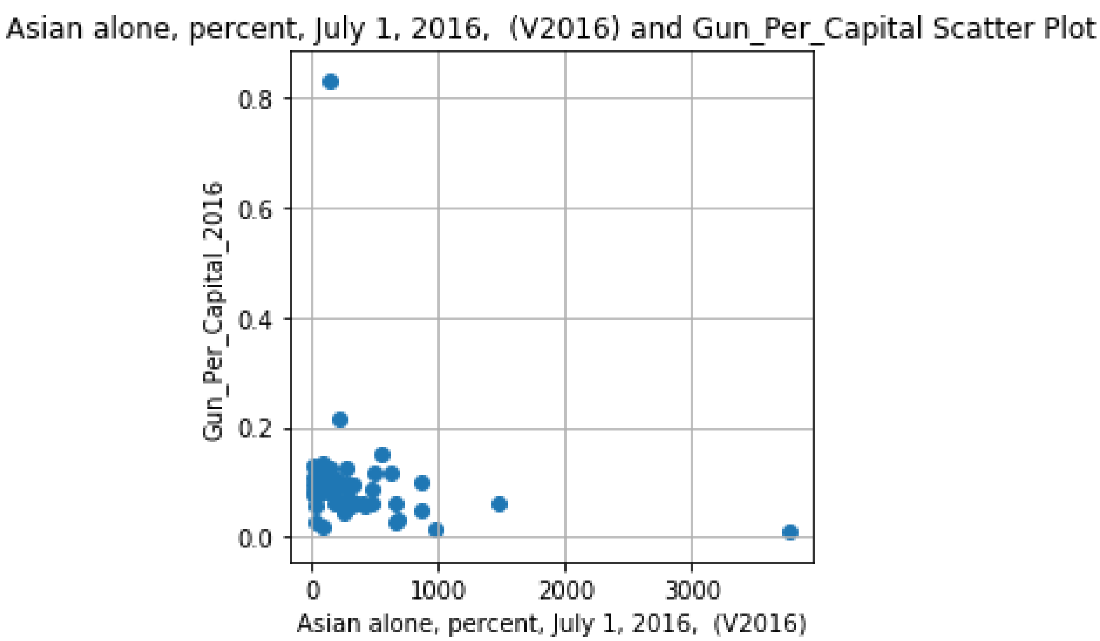
`Persons 65 years and over, percent, April 1, 2010`



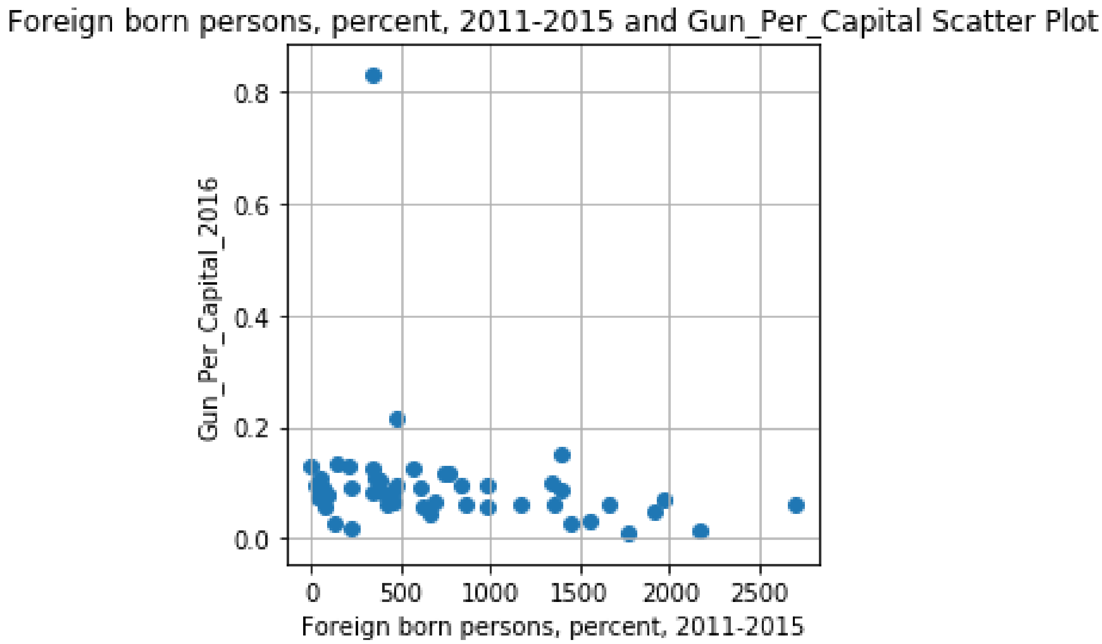
`Owner-occupied housing unit rate, 2011-2015`



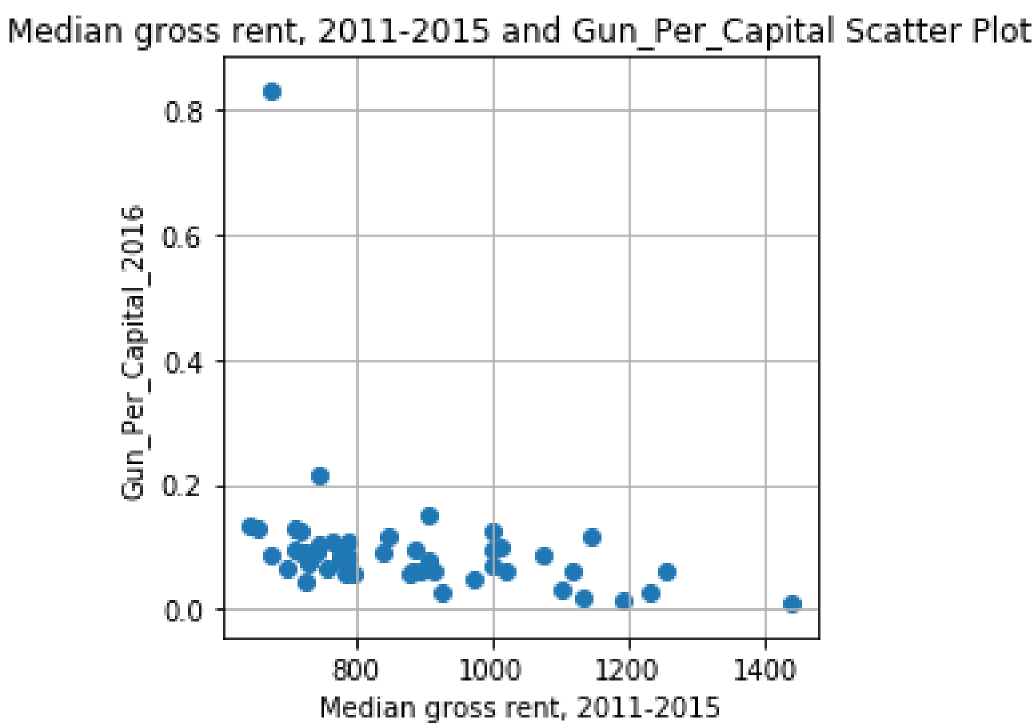
Asian alone, percent, July 1, 2016, (V2016)



Foreign born persons, percent, 2011-2015



Median gross rent, 2011-2015



6.4 Finding for Research question 1: Based on graphs and results

6.5 The Gun and Census data are divided by state with united state. High Gun Per Capita should also be calculated by state.

6.6 Among all the state, Kentucky has the highest Gun per Capita in both 2016 as well in 2010, might be that Kentucky is place where people prefer firearms.

6.7 Based on scatter plot between all fact values and Gun per Capita in one figure and grouping by column name, there is no strong association between any high fact value and high gun per capita.

6.8 However, based on the scatter plot for fact variable and gun per capita, separately, there is some weak association found as following:

6.9 There is positive association between gun per capita and variables which includes:

7 'White alone, percent, July 1, 2016, (v2016)',

8 'Persons 65 years and over, percent, April 1, 2010',

9 'Owner-occupied housing unit rate, 2011-2015',

10

11 the negative association between gun per capita and variables which includes:

12 Asian alone, percent, July 1, 2016, (V2016)

13 Foreign born persons, percent, 2011-2015

14 Median gross rent, 2011-2015

14.0.1 Research Question 2

15 Which states have had the highest growth and the lowest growth in gun registrations?

In [95]:
```
#Calculate the increasing percentage of gun registrations from 2010 to 2016

result['gun_growth'] = result['totals_16']/(result['totals_16']-result['totals_10'])
```

In [97]:
```
# lets find the maximum growth percentage

result['gun_growth'].max()
```

Out[97]:
```
4.031971662526451
```

SO maxixum gun growth is 403.1971%. . .

```
In [98]: #List Top 5 rows by gun growth rate descending4

         result.nlargest(5,'gun_growth')
```

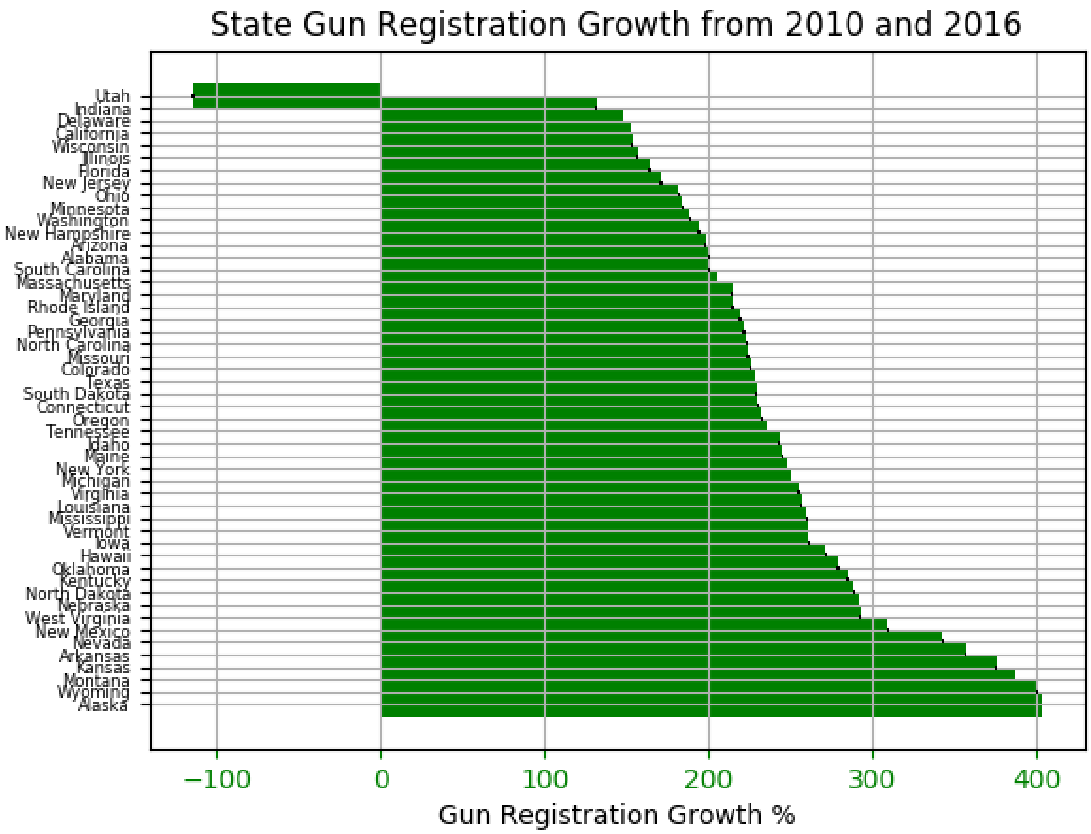
```
Out[98]:        state  totals_16  totals_10  \
         1      Alaska      87647      65909
         49    Wyoming      63594      47709
         25    Montana     136337     101095
         15     Kansas     196548     144156
         3    Arkansas     266014     191448

             Population estimates, July 1, 2016,  (V2016)  \
         1                                         741894.0
         49                                        585501.0
         25                                       1042520.0
         15                                       2907289.0
         3                                        2988248.0

             Population estimates base, April 1, 2010,  (V2016)  \
         1                                                710249.0
         49                                               563767.0
         25                                               989414.0
         15                                              2853129.0
         3                                               2916025.0

             Population, percent change - April 1, 2010 (estimates base) to July 1, 2016,  (V201
         1                                                   450.0
         49                                                  390.0
         25                                                  540.0
         15                                                  190.0
         3                                                   250.0

             Population, Census, April 1, 2010  \
         1                            710231.0
         49                           563626.0
         25                           989415.0
         15                          2853118.0
         3                           2915918.0

             Persons under 5 years, percent, July 1, 2016,  (V2016)  \
         1                                                   730.0
         49                                                  650.0
         25                                                  600.0
         15                                                  670.0
         3                                                   640.0
```

```
    Persons under 5 years, percent, April 1, 2010  \
1                                           760.0
49                                          710.0
25                                          630.0
15                                          720.0
3                                           680.0

    Persons under 18 years, percent, July 1, 2016,  (V2016)     ...      \
1                                                2520.0         ...
49                                               2370.0         ...
25                                               2180.0         ...
15                                               2460.0         ...
3                                                2360.0         ...

    Minority-owned firms, 2012  Nonminority-owned firms, 2012  \
1                      13688.0                        51147.0
49                      4077.0                        55397.0
25                      5578.0                       102746.0
15                     26127.0                       204562.0
3                      35982.0                       189029.0

    Veteran-owned firms, 2012  Nonveteran-owned firms, 2012  \
1                      7953.0                       56091.0
49                     6470.0                       51353.0
25                    11486.0                       93393.0
15                    21610.0                      203401.0
3                     25915.0                      192988.0

    Population per square mile, 2010  Land area in square miles, 2010  \
1                               12.0                       57064096.0
49                              58.0                        9709314.0
25                              68.0                       14554580.0
15                             349.0                        8175872.0
3                               56.0                        5203548.0

    FIPS Code  Gun_Per_Capital_2016  Gun_Per_Capital_2010  gun_growth
1         2.0              0.118140              0.092797    4.031972
49       56.0              0.108615              0.084625    4.003399
25       30.0              0.130776              0.102177    3.868594
15       20.0              0.067605              0.050526    3.751489
3         5.0              0.089020              0.065654    3.567497

[5 rows x 71 columns]
```

In [99]:
```
# lets plot a bar graph for every states gun growth:
#Set the figure size for better visualization
plt.figure(figsize=(30,30))
```

```python
plt.rcdefaults()
fig, ax = plt.subplots()

#Sort result data by gun_growth value
sorted = result.sort_values(by=['gun_growth'])


#create bar chart
y_pos = np.arange(len(sorted['state']))
error = np.random.rand(len(sorted['state']))
ax.barh(y_pos, (sorted['gun_growth']*100), xerr=error, align='center',height=2,linewidt

#set x and y axis lable and make the label readable
ax.set_yticks(y_pos)
ax.set_xlabel("Gun Registration Growth %")
ax.set_yticklabels(sorted['state'],size=6)

#Invert x and y axis
ax.invert_yaxis()  # labels read top-to-bottom

#Set tick colors:
ax.tick_params(axis='x', colors='green')
ax.tick_params(axis='y', colors='black')

#Set the title
plt.title("State Gun Registration Growth from 2010 and 2016")
plt.grid(True)
plt.legend(bbox_to_anchor=(1.1, 1.05))
plt.show()
```

```
<matplotlib.figure.Figure at 0x7fe9ebef9d68>
```



16 From the above graph we could see that only Utah and Indian are two states whose gun growth is decreased more 100%

16.1 Further to computations earlier, Alaska has the highest with 403.20% compared to Apr 2010.

16.2 Alaska and Wyoming are the only two states whose growth is more than 400%

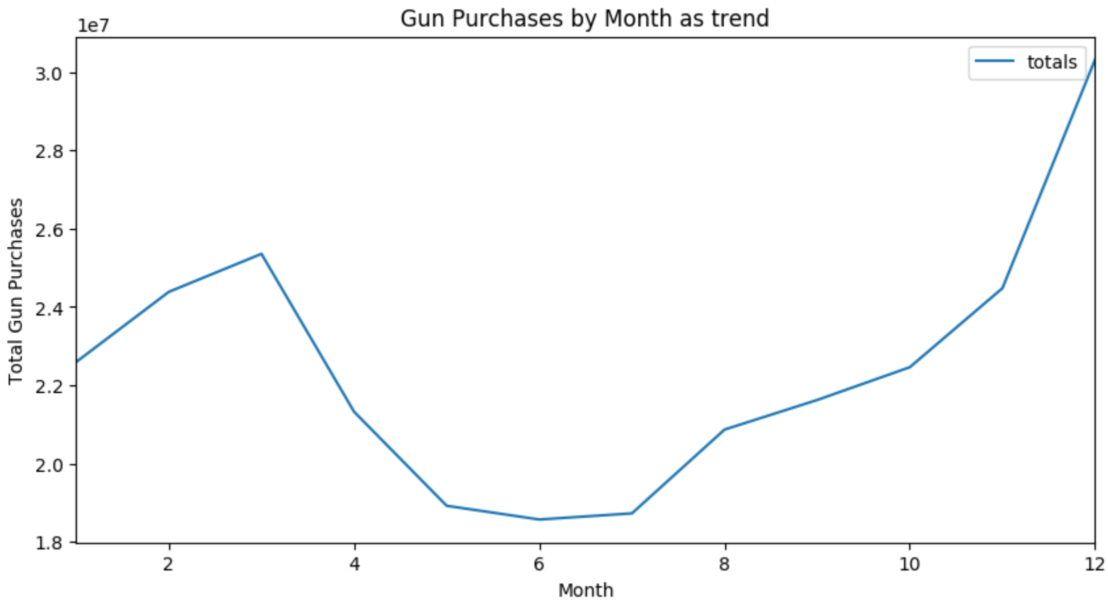
16.2.1 From the graph one more point we could note is that there are 8 countries whose gun growth is more than 300%, we can consider these states are where people trending to possess fire arm(reason might be any security or other reason)

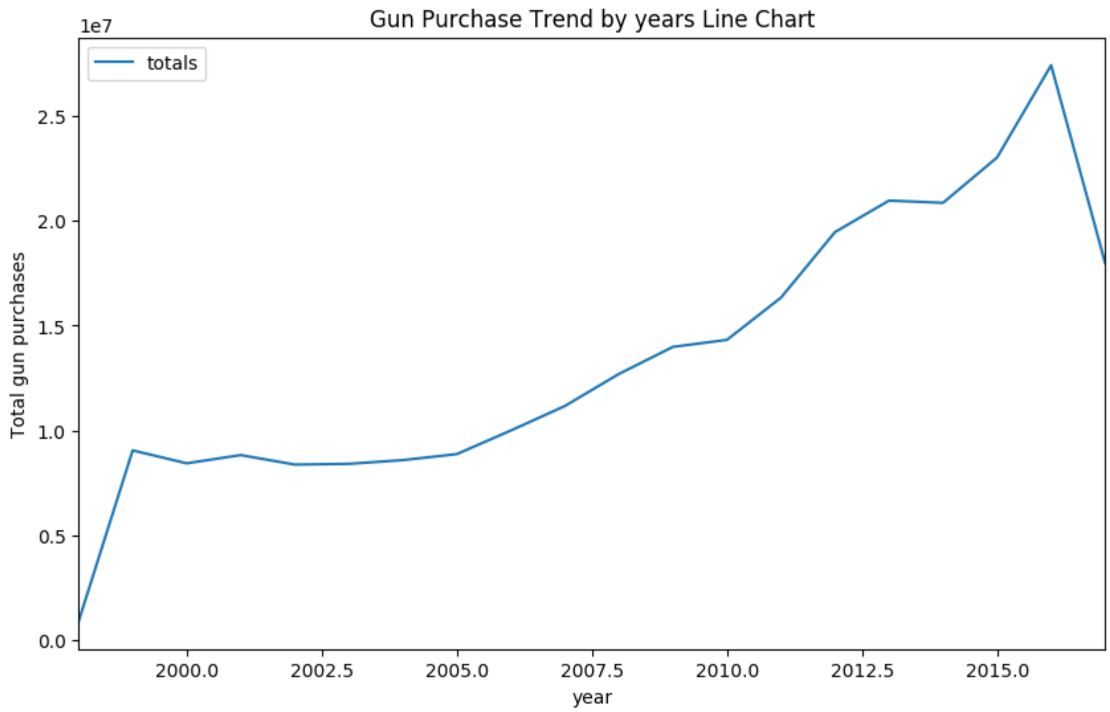
17 Research Question 3:

17.1 We will explore the overall trend for gun purchase from 2010 to 2016 to observe overall trend.

```
In [101]: #Lets create line chart using grouped data by year - months, which can be used to obse

          #let's give a figure sizze
          plt.figure(figsize = (10,5))
```

```python
gun.groupby('months')['totals'].sum().plot(kind='line',sharex=True, sharey=True, layou
#setting x and y axes label name

plt.xlabel('Month')
plt.ylabel('Total Gun Purchases')
plt.legend()
plt.title('Gun Purchases by Month as trend')
plt.show


#Lets Create line chart using grouped data by year, which can be used to observe the c
plt.figure(figsize=(10,6))
gun.groupby('year')['totals'].sum().plot(kind='line')

plt.ylabel('Total gun purchases')
plt.xlabel('year')
plt.title("Gun Purchase Trend by years Line Chart")
plt.legend()
plt.show()
```





18 From Line graph for trends across year, we can tell that:

18.1 From 1998 to 2016 the gun purchases is increasing.

18.2 From 1999 to 2005, the number of guns is almost stable

18.3 The gun purchases of 1 million in 2005 increased to more than 2.5 million by the year 2016

19 From the line trend for gun purchases by month, we can say below:

19.1 The gun purchases is around 2.3 million in january and is increasing till the month of march(where it is almost 2.5 million).

19.2 From march we could see that it has started decreasing and reached near to 1.9 million and this 1.9 million is somewhat constant till the month of July.

19.3 From July to december we could see that slope has started increasing which means that gun purchases are increasing.

19.4 the peak gun purchases is clearly for the month of december with over 3 million purchases and next level peak gun purchases happening between january and march.

Conclusions # FBI Gun and census data are two independent data sets. Their common variables/value include state of United States and year month, which requires data cleaning at first.

We can join these two dataset to see the relationship between gun purchase and census variable. # # Post Question: ### 1.What census variable or fact value is most associated with high gun per capita per state? ### Census data includes state as variable, and there are 65 differnt census measurement as value of Fact. ### 2.Which states have had the highest growth and the lowest growth in gun registrations from Apr 2010 to Jul 2016? ### 3.What is the overall trend of gun purchases by year or by year and month? # # Findings for the Quesitons: # ### 1.The gun and census data are divided by state with United state. High gun per capita should also be calculated by state, except {'District of Columbia','Guam','Mariana Islands','Puerto Rico','Virgin Islands'}. these states' gun total is missing or zero. Among all the state, Kentucky has the highest gun per capita on Jul 2016 and Apr 2010 data. Kentucky,Indiana,Illinois,West Virginia,Montana are the top 5 state who have highes gun per capita on Jul 2016. also, based on the scatter plot for all the fact value in one figure and group by column name scatterplot,there is no strong association between any fact value and high gun per capita. ### However, based on the scatter plot for fact varible and gun per capita, separately, there is some weak association found as following; the positive association between gun per capita and variables which includes: White alone, percent, July 1, 2016, (V2016) Persons 65 years and over, percent, April 1, 2010 owner-occupied housing unit rate, 2011 -2015 ### the negative association between gun per capita and variables which includes: 2011-2015 Asian alone, percent, July 1, 2016, (V2016) Foreign born persons, percent, 2011-2015 Median gross rent, 2011-2015 # ### 2.Alaska had the highest growth in gun registrations in Jul 2017, increasing by 403.20% compare to Apr 2010. ### Additionally,Alaska,Wyoming,Montana,Kansas,Arkansas are the top 5 state with highest growth in gun registrations in Jul 2017.Alaskas and Wyoming are only two state whose growth more than 400% ### also from the gun growth bar chart for all the states, we can see Utah and Indiana are the only two states whose gun growth are descreasing by more than 100%. On the other hand, there are 8 states' gun growth more than 300%, which can be considered outliers. # # 3.From the line chart for gun purchases by years, we can tell that ### from 1998 to 2016, the overall of gun purchases is increasing. ### From 1999 to 2005, the number of gun purchases remains stable, and from 2005 to 2016, the number of gun purchases increase from about 1 million to 2.7 million. From 2016 to 2017, the number of gun purchases goes down, which is partially due to only 9 months in 2017 being calculated. # ### The gun purchases is around 2.3 million in january and is increasing till the month of march(where it is almost 2.5 million). ### From march we could see that it has started decreasing and reached near to 1.9 million and this 1.9 million is somewhat constant till the month of July. ### From July to december we could see that slope has started increasing which means that gun purchases are increasing. ### the peak gun purchases is clearly for the month of december with over 3 million purchases and next level peak gun purchases happening between january and march. # # Limitations: ## I have replaced gun data's missing values with mean of each columns and remove 'Fact Note' column since it has exceeding number of Nan values. Missing data can occur because of nonresponse: no information is provided.for the gun data, missing value can be caused by nonreponse or limitation regulation or lack of gathering data. My solution for replacing missing data with mean and drop null columns are not time consuming. # ## a. For potential improvement, in statistic, probablity distribution graphic can be used to see variable's distribution(normalized/ right-skewed/left skewed) and make prediction of missing value based on mean/ standard deviation. ## b. Also, standardization of datasets before exploring it can help show more clear and strong correlation between variable, for example, gun per capital and Fact metrics. ## c. Additionally, the gun data contains many outliers, scaling using the mean and variance of the data is likely to not work very well. In these cases we would need to search for other methods to give brief about stats for these data.

19.5 Submitting your Project

```
In [103]: from subprocess import call
          call(['python', '-m', 'nbconvert', 'Investigate_a_Dataset.ipynb'])

Out[103]: 0
```