Tien Duong
Jul-8-2020

Project: Investigation the relation of two datasets (Population Census and FBI Gun Record)

Table of contents

- Introduction
- Data Wrangling
- Exploratory Data Analysis
- Conclusions

Introduction

Welcome to the exploratory data analysis of the relation between population census dataset and gun record information. The gun record information comes from the FBI's national instant Criminal Background check system or NICS. Whenever there is a firearm purchase, gun shop's owner will run a check through the NICS system to ensure that the buyer meet all of the qualification before their purchase. Accompanying the NICS dataset is the U.S. census dataset of which contain serveral variables at the state level. Most variables have only one data point per state (2016), but a few have data for more than one year (poverty).

Census link (https://www.census.gov/)

Census and FBI gun Data (https://d17h27t6h515a5.cloudfront.net/topher/2017/November/5a0a5623_ncis-and-census-data/ncis-and-census-data.zip)

Libary using in this data investigation report

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

Data Wrangling

In [2]:
```
# Importing the FBI gun data record
gun_info = pd.read_csv('gun_data.csv')
gun_info.head()
```

Out[2]:

	month	state	permit	permit_recheck	handgun	long_gun	other	multiple	admin	prepawı
0	2017-09	Alabama	16717.0	0.0	5734.0	6320.0	221.0	317	0.0	
1	2017-09	Alaska	209.0	2.0	2320.0	2930.0	219.0	160	0.0	
2	2017-09	Arizona	5069.0	382.0	11063.0	7946.0	920.0	631	0.0	
3	2017-09	Arkansas	2935.0	632.0	4347.0	6063.0	165.0	366	51.0	
4	2017-09	California	57839.0	0.0	37165.0	24581.0	2984.0	0	0.0	

5 rows × 27 columns

In [3]:
```
# Importing the census data record
census = pd.read_csv('U.S. Census Data.csv')
census.head()
```

Out[3]:

	Fact	Fact Note	Alabama	Alaska	Arizona	Arkansas	California	Colorado	Connecticut	D
0	Population estimates, July 1, 2016, (V2016)	NaN	4,863,300	741,894	6,931,071	2,988,248	39,250,017	5,540,545	3,576,452	
1	Population estimates base, April 1, 2010, (V2...	NaN	4,780,131	710,249	6,392,301	2,916,025	37,254,522	5,029,324	3,574,114	
2	Population, percent change - April 1, 2010 (es...	NaN	1.70%	4.50%	8.40%	2.50%	5.40%	10.20%	0.10%	
3	Population, Census, April 1, 2010	NaN	4,779,736	710,231	6,392,017	2,915,918	37,253,956	5,029,196	3,574,097	
4	Persons under 5 years, percent, July 1, 2016, ...	NaN	6.00%	7.30%	6.30%	6.40%	6.30%	6.10%	5.20%	

5 rows × 52 columns

In [4]:
```
# checking for missing data
gun_info.isna().sum(), census.isna().sum()
```

Out[4]:
```
(month                          0
state                          0
permit                        24
permit_recheck             11385
handgun                       20
long_gun                      19
other                       6985
multiple                       0
admin                         23
prepawn_handgun             1943
prepawn_long_gun            1945
prepawn_other               7370
redemption_handgun          1940
redemption_long_gun         1941
redemption_other            7370
returned_handgun           10285
returned_long_gun          10340
returned_other             10670
rentals_handgun            11495
rentals_long_gun           11660
private_sale_handgun        9735
private_sale_long_gun       9735
private_sale_other          9735
return_to_seller_handgun   10010
return_to_seller_long_gun   9735
return_to_seller_other     10230
totals                         0
dtype: int64,
Fact                5
Fact Note          57
Alabama            20
Alaska             20
Arizona            20
Arkansas           20
California         20
Colorado           20
Connecticut        20
Delaware           20
Florida            20
Georgia            20
Hawaii             20
Idaho              20
Illinois           20
Indiana            20
Iowa               20
Kansas             20
Kentucky           20
Louisiana          20
Maine              20
Maryland           20
Massachusetts      20
Michigan           20
Minnesota          20
Mississippi        20
```

```
Missouri          20
Montana           20
Nebraska          20
Nevada            20
New Hampshire     20
New Jersey        20
New Mexico        20
New York          20
North Carolina    20
North Dakota      20
Ohio              20
Oklahoma          20
Oregon            20
Pennsylvania      20
Rhode Island      20
South Carolina    20
South Dakota      20
Tennessee         20
Texas             20
Utah              20
Vermont           20
Virginia          20
Washington        20
West Virginia     20
Wisconsin         20
Wyoming           20
dtype: int64)
```

In [5]:
```
# Fill in the missing data with 0
gun_info.fillna(0,inplace=True), census.fillna(0,inplace=True)
```

Out[5]: (None, None)

Checking both dataframe for missing data or abnormalities

In [6]:
```
1 gun_info.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12485 entries, 0 to 12484
Data columns (total 27 columns):
 #   Column                     Non-Null Count  Dtype  
---  ------                     --------------  -----  
 0   month                      12485 non-null  object 
 1   state                      12485 non-null  object 
 2   permit                     12485 non-null  float64
 3   permit_recheck             12485 non-null  float64
 4   handgun                    12485 non-null  float64
 5   long_gun                   12485 non-null  float64
 6   other                      12485 non-null  float64
 7   multiple                   12485 non-null  int64  
 8   admin                      12485 non-null  float64
 9   prepawn_handgun            12485 non-null  float64
 10  prepawn_long_gun           12485 non-null  float64
 11  prepawn_other              12485 non-null  float64
 12  redemption_handgun         12485 non-null  float64
 13  redemption_long_gun        12485 non-null  float64
 14  redemption_other           12485 non-null  float64
 15  returned_handgun           12485 non-null  float64
 16  returned_long_gun          12485 non-null  float64
 17  returned_other             12485 non-null  float64
 18  rentals_handgun            12485 non-null  float64
 19  rentals_long_gun           12485 non-null  float64
 20  private_sale_handgun       12485 non-null  float64
 21  private_sale_long_gun      12485 non-null  float64
 22  private_sale_other         12485 non-null  float64
 23  return_to_seller_handgun   12485 non-null  float64
 24  return_to_seller_long_gun  12485 non-null  float64
 25  return_to_seller_other     12485 non-null  float64
 26  totals                     12485 non-null  int64  
dtypes: float64(23), int64(2), object(2)
memory usage: 2.6+ MB
```

In [7]:
```
1 census.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 85 entries, 0 to 84
Data columns (total 52 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Fact            85 non-null     object
 1   Fact Note       85 non-null     object
 2   Alabama         85 non-null     object
 3   Alaska          85 non-null     object
 4   Arizona         85 non-null     object
 5   Arkansas        85 non-null     object
 6   California      85 non-null     object
 7   Colorado        85 non-null     object
 8   Connecticut     85 non-null     object
 9   Delaware        85 non-null     object
 10  Florida         85 non-null     object
 11  Georgia         85 non-null     object
 12  Hawaii          85 non-null     object
 13  Idaho           85 non-null     object
 14  Illinois        85 non-null     object
 15  Indiana         85 non-null     object
 16  Iowa            85 non-null     object
 17  Kansas          85 non-null     object
 18  Kentucky        85 non-null     object
 19  Louisiana       85 non-null     object
 20  Maine           85 non-null     object
 21  Maryland        85 non-null     object
 22  Massachusetts   85 non-null     object
 23  Michigan        85 non-null     object
 24  Minnesota       85 non-null     object
 25  Mississippi     85 non-null     object
 26  Missouri        85 non-null     object
 27  Montana         85 non-null     object
 28  Nebraska        85 non-null     object
 29  Nevada          85 non-null     object
 30  New Hampshire   85 non-null     object
 31  New Jersey      85 non-null     object
 32  New Mexico      85 non-null     object
 33  New York        85 non-null     object
 34  North Carolina  85 non-null     object
 35  North Dakota    85 non-null     object
 36  Ohio            85 non-null     object
 37  Oklahoma        85 non-null     object
 38  Oregon          85 non-null     object
 39  Pennsylvania    85 non-null     object
 40  Rhode Island    85 non-null     object
 41  South Carolina  85 non-null     object
 42  South Dakota    85 non-null     object
 43  Tennessee       85 non-null     object
 44  Texas           85 non-null     object
 45  Utah            85 non-null     object
 46  Vermont         85 non-null     object
 47  Virginia        85 non-null     object
 48  Washington      85 non-null     object
 49  West Virginia   85 non-null     object
```

```
 50  Wisconsin       85 non-null     object
 51  Wyoming         85 non-null     object
dtypes: object(52)
memory usage: 34.7+ KB
```

Data cleaning

First we will be observing the census dataset properties, and then we will be moving onto the cleaning process of the gun dataset.

Cleaning the census dataset

In [8]:
```
# census dataset containt footnote inside the data,
# therefore we will initiate slicing method to drop out those footnote
census.drop(census.index[65:],inplace=True)
census.tail()
# Census dataset no longer contain any of the footnotes
```

Out[8]:

	Fact	Fact Note	Alabama	Alaska	Arizona	Arkansas	California	Colorado	Connect
60	Veteran-owned firms, 2012	0	41,943	7,953	46,780	25,915	252,377	51,722	31
61	Nonveteran-owned firms, 2012	0	316,984	56,091	427,582	192,988	3,176,341	469,524	281
62	Population per square mile, 2010	0	94.4	1.2	56.3	56	239.1	48.5	7
63	Land area in square miles, 2010	0	50,645.33	570,640.95	113,594.08	52,035.48	155,779.22	103,641.89	4,84
64	FIPS Code	0	"01"	"02"	"04"	"05"	"06"	"08"	

5 rows × 52 columns

Below is what we dropped from the census's footnote

NOTE: FIPS Code values are enclosed in quotes to ensure leading zeros remain intact.

Value Notes
1 Includes data not distributed by county.

Fact Notes
(a) Includes persons reporting only one race
(b) Hispanics may be of any race, so also are included in applicable race categories

(c) Economic Census - Puerto Rico data are not comparable to U.S. Economic Census data

Value Flags

- Either no or too few sample observations were available to compute an estimate, or a ratio of medians cannot be calculated because one or both of the median estimates falls in the lowest or upper interval of an open ended distribution.
  D Suppressed to avoid disclosure of confidential information
  F Fewer than 25 firms
  FN Footnote on this item in place of data
  NA Not available
  S Suppressed; does not meet publication standards
  X Not applicable
  Z Value greater than zero but less than half unit of measure shown

NOTE: FIPS Code values are enclosed in quotes to ensure leading zeros remain intact.	
Value Notes	
1	Includes data not distributed by county.
Fact Notes	
(a)	Includes persons reporting only one race
(b)	Hispanics may be of any race, so also are included in applicable race categories
(c)	Economic Census - Puerto Rico data are not comparable to U.S. Economic Census data
Value Flags	
-	Either no or too few sample observations were available to compute an estimate, or a ratio of medians cannot be calculated because one or both of the median estimates falls in the lowest or upper interval of an open ended distribution.
D	Suppressed to avoid disclosure of confidential information
F	Fewer than 25 firms
FN	Footnote on this item in place of data
NA	Not available
S	Suppressed; does not meet publication standards
X	Not applicable
Z	Value greater than zero but less than half unit of measure shown

Cleaning the FBI Gun dataset

Extracting the columns from gun_info dataset that we will be workin
g on.

In [9]:
```
# We will only be using the state, permit and totals column
gun_info = gun_info[['month','state','permit','totals']]
gun_info.head()
```

Out[9]:

	month	state	permit	totals
0	2017-09	Alabama	16717.0	32019
1	2017-09	Alaska	209.0	6303
2	2017-09	Arizona	5069.0	28394
3	2017-09	Arkansas	2935.0	17747
4	2017-09	California	57839.0	123506

In [10]:
```
# Checking for the data type of each column
gun_info.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12485 entries, 0 to 12484
Data columns (total 4 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   month   12485 non-null  object 
 1   state   12485 non-null  object 
 2   permit  12485 non-null  float64
 3   totals  12485 non-null  int64  
dtypes: float64(1), int64(1), object(2)
memory usage: 390.3+ KB
```

Exploratory Data Analysis

Question?

There are 3 exploratory questions for the analysis

- What census data is most associated with high gun per capita?
- Which states have had the highest growth in gun registration?
- What is the overall trend of gun purchases?

- What census data is most associated with high gun per capita?

The census data that associate with gun per capita would be the population and veteran per state. We will be comparing the total population with the totals firearms information that we have. Also, we will be observing the permit and total guns information on record.

Census dataset manipulation

In [11]:
```
census.iloc[[0,20]]
```

Out[11]:

	Fact	Fact Note	Alabama	Alaska	Arizona	Arkansas	California	Colorado	Connecticut	D
0	Population estimates, July 1, 2016, (V2016)	0	4,863,300	741,894	6,931,071	2,988,248	39,250,017	5,540,545	3,576,452	
20	Veterans, 2011-2015	0	363,170	69,323	505,794	220,953	1,777,410	391,725	199,331	

2 rows × 52 columns

In [12]:
```
ating the total population row inside the census dataset using iloc
census.iloc[0]
vert the extracted data into it own dataframe + reverse the index and column
s_q1 = pd.DataFrame({'Jul-1-2016 population':df1})
s_q1.drop(['Fact Note','Fact'],inplace=True)
s_q1.head()
```

Out[12]:

	Jul-1-2016 population
Alabama	4,863,300
Alaska	741,894
Arizona	6,931,071
Arkansas	2,988,248
California	39,250,017

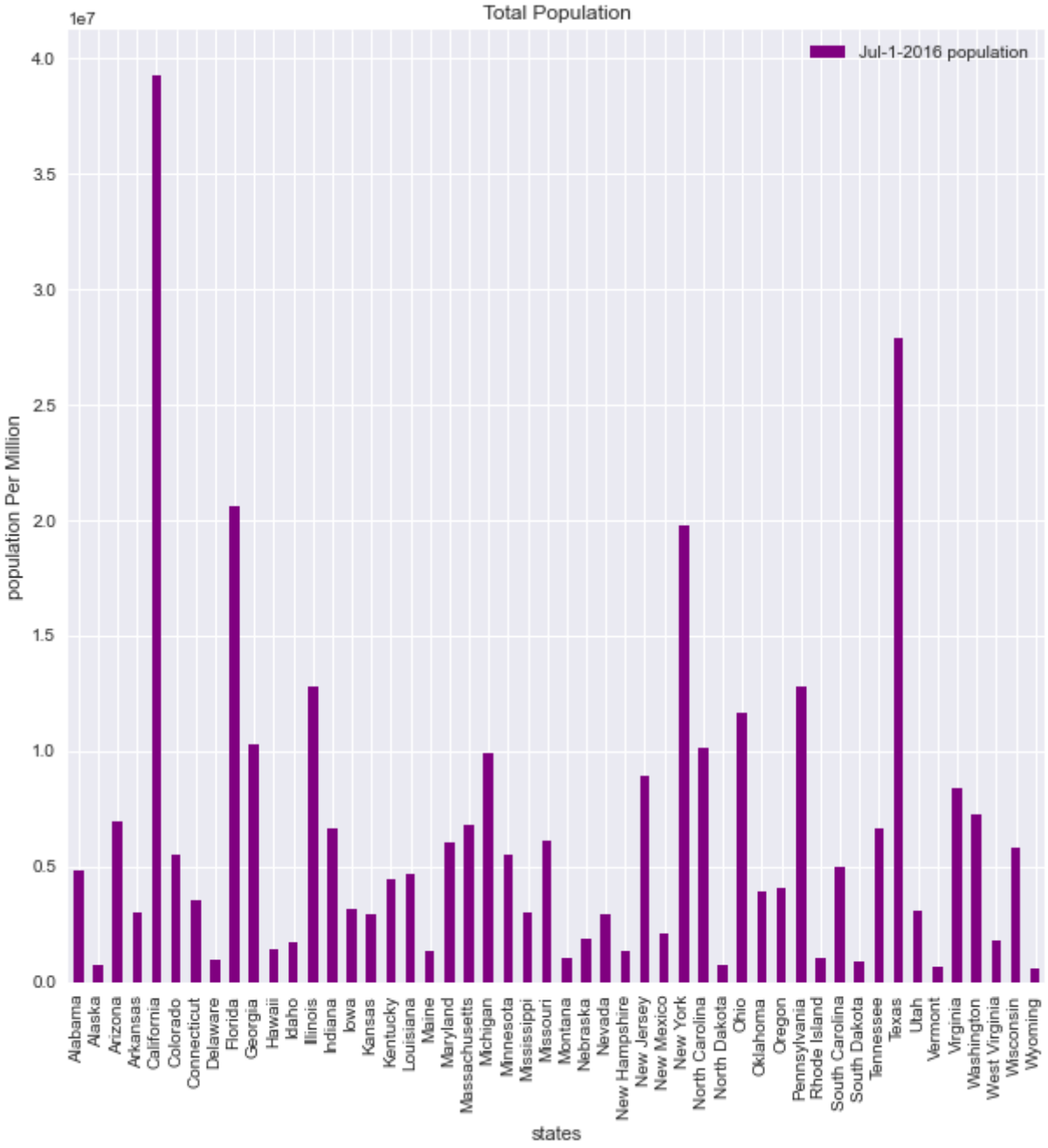
In [13]:
```
# Manipulating the data into the proper type by using to_numeric
# Using replace method to removing the comma inside our columns
census_q1.replace(',','', regex=True,inplace=True)
c = census_q1.select_dtypes(object).columns
census_q1[c] = census_q1[c].apply(pd.to_numeric, errors='coerce')
census_q1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 50 entries, Alabama to Wyoming
Data columns (total 1 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Jul-1-2016 population  50 non-null     int64
dtypes: int64(1)
memory usage: 800.0+ bytes
```

Visualizing the total 2016 population

In [14]:
```
# Changing the graph style
plt.style.use('seaborn')
```

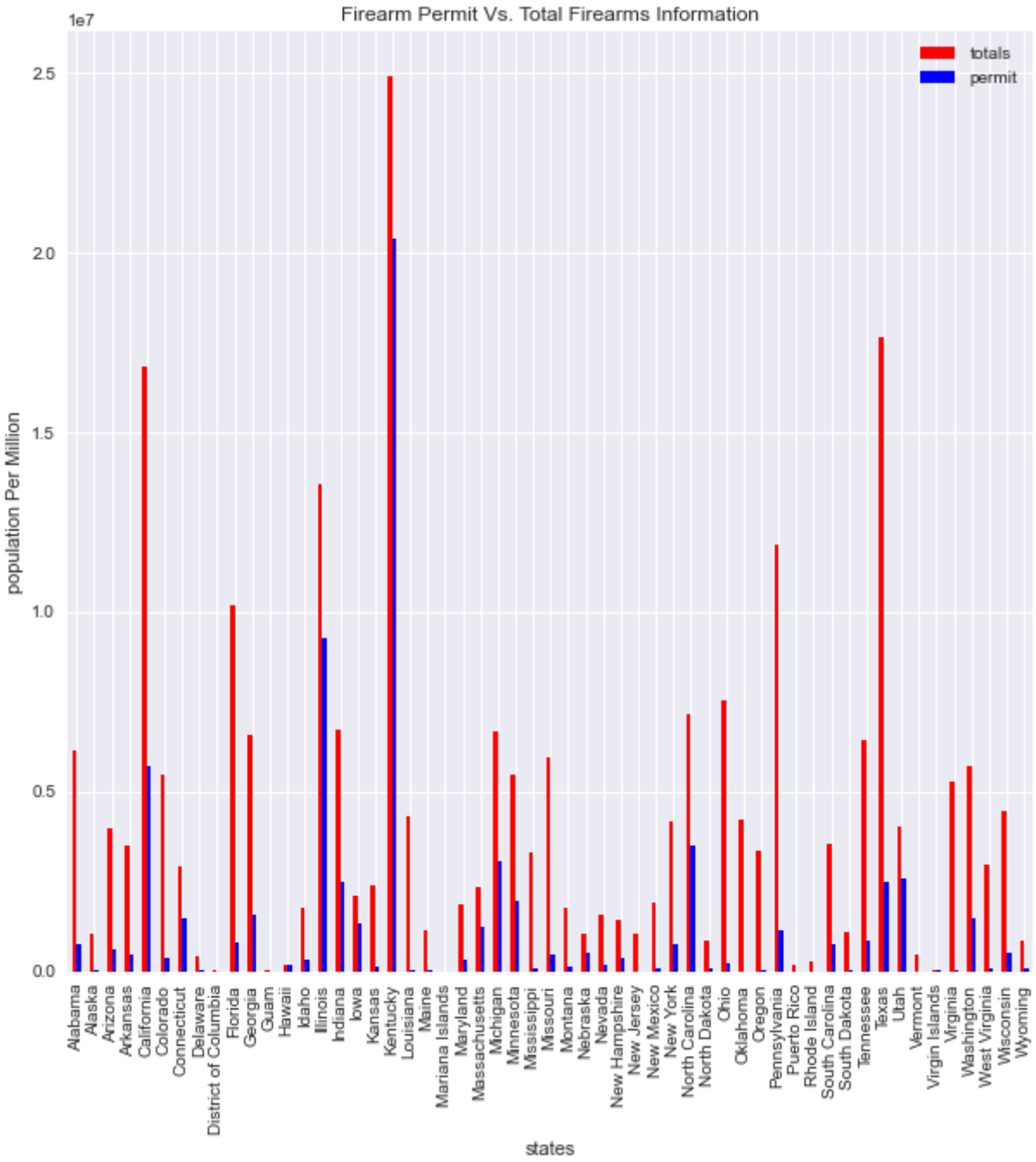
In [15]:
```
ax= census_q1.plot(kind='bar',figsize=(10,10),color='purple')
ax.set(title='Total Population',
xlabel='states',
ylabel='population Per Million')
ax.legend().set_visible(True)
```



Gun dataset manipulation

In [16]:
```
# Narrow down the dates of gun record on file,
# matching it with the census date
gun_q1 = gun_info.query('month <= "2016-07"')
gun_q1 = gun_q1.groupby('state').agg({'totals':'sum','permit':'sum'})
```

In [17]:
```
ax= gun_q1.plot(kind='bar',figsize=(10,10),color=['red','blue'])
ax.set(title='Firearm Permit Vs. Total Firearms Information',
xlabel='states',
ylabel='population Per Million')
ax.legend().set_visible(True)
```



In [18]:
```
# joining census dataset into firearms dataset
mbined = census_q1.merge(gun_q1, left_on=census_q1.index, right_on=gun_q1.ir
mbined.info()
```

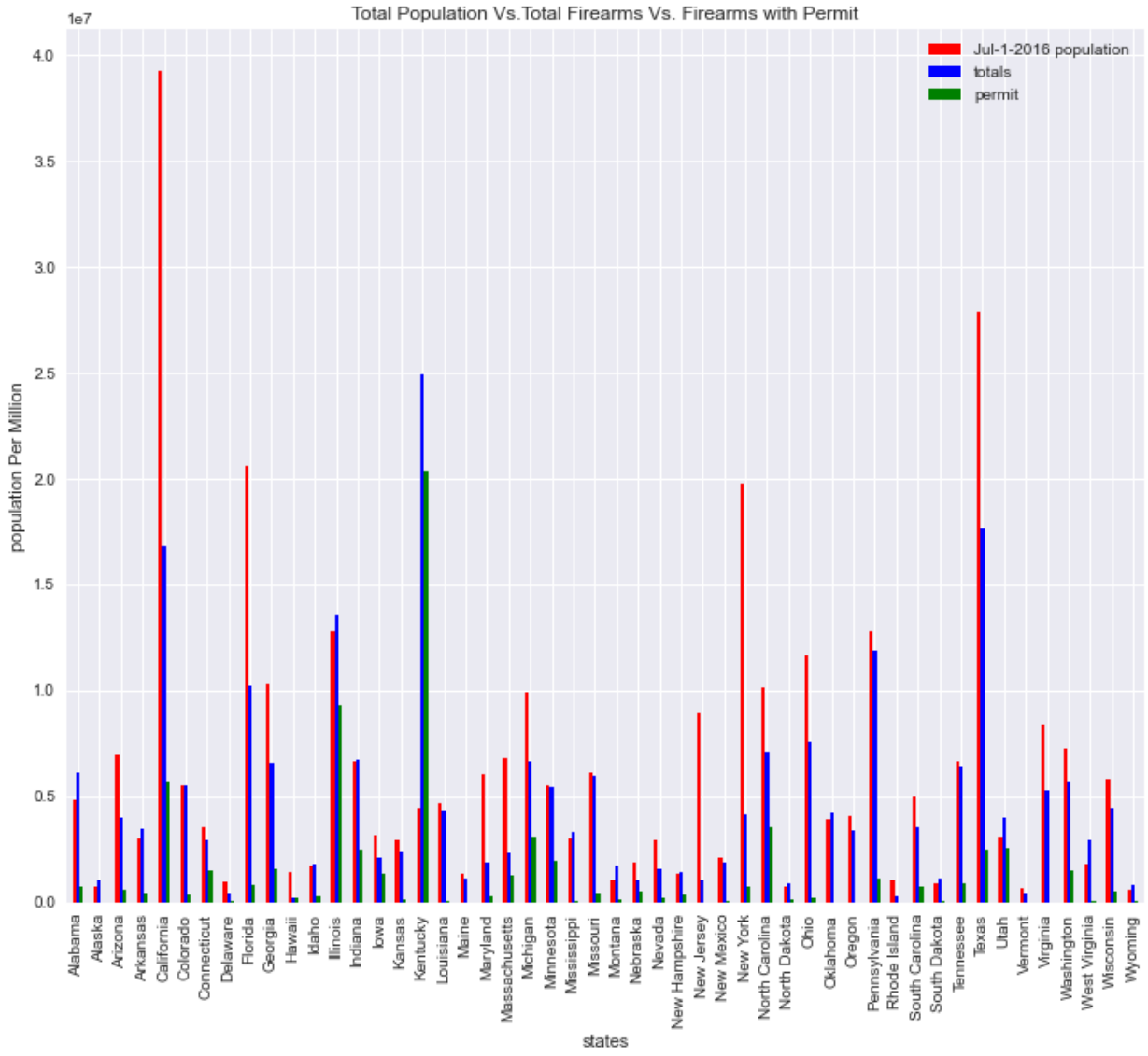
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 50 entries, 0 to 49
Data columns (total 4 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   key_0                  50 non-null     object
 1   Jul-1-2016 population  50 non-null     int64
 2   totals                 50 non-null     int64
 3   permit                 50 non-null     float64
dtypes: float64(1), int64(2), object(1)
memory usage: 2.0+ KB
```

In [19]:
```
df_combined.head()
```

Out[19]:

	key_0	Jul-1-2016 population	totals	permit
0	Alabama	4863300	6129783	742682.0
1	Alaska	741894	1039997	9879.0
2	Arizona	6931071	3960606	604209.0
3	Arkansas	2988248	3470885	458960.0
4	California	39250017	16807520	5685338.0

In [20]:
```
f_combined.plot(x='key_0',kind='bar',figsize=(12,10),color=['red','blue','gr
t(title='Total Population Vs.Total Firearms Vs. Firearms with Permit',
l='states',
l='population Per Million')
grid().set_visible(True)
```

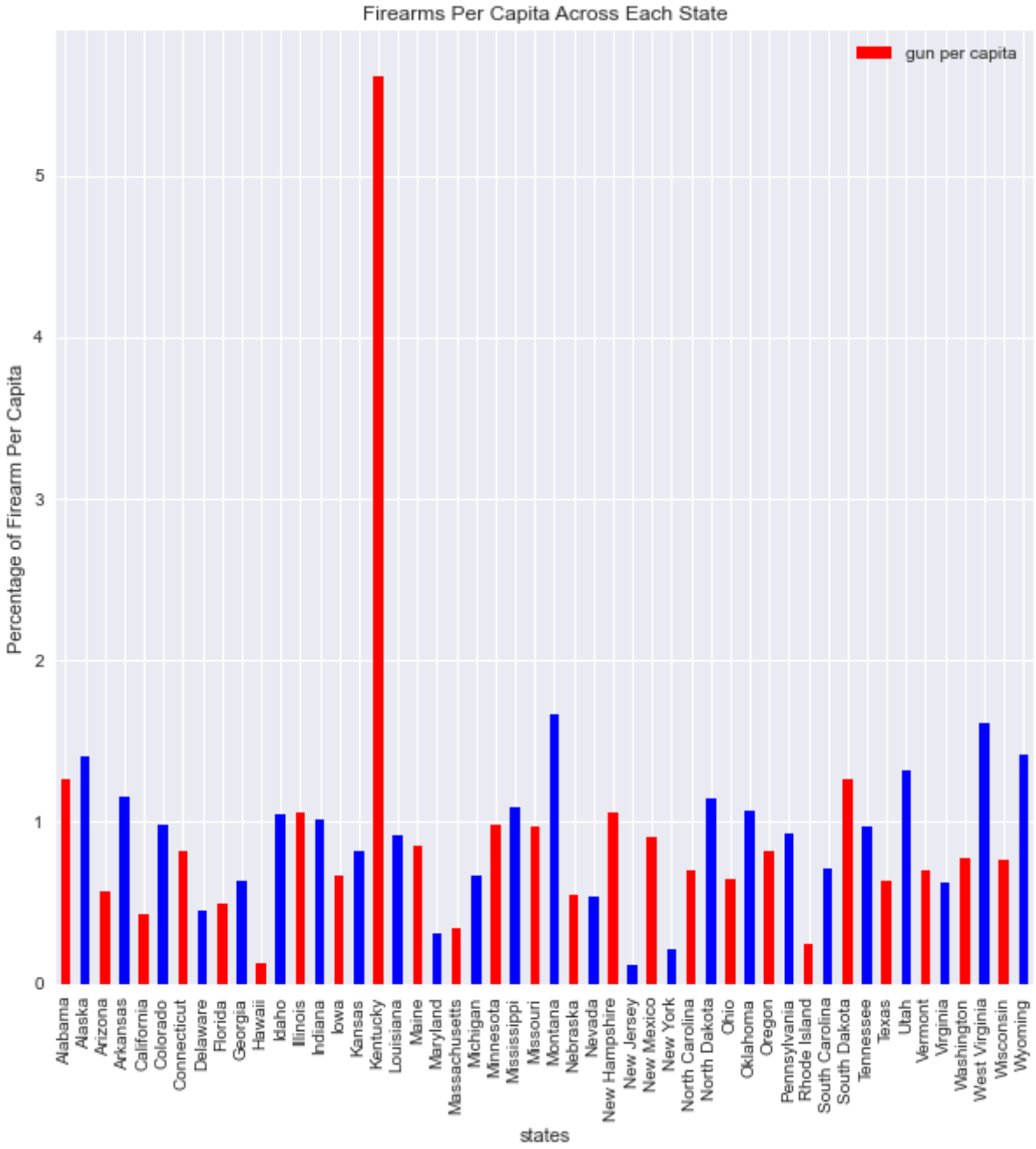


In [21]:
```
# Total percengtage of gun owner per capita
df_combined['gun per capita'] = df_combined['totals']/df_combined['Jul-
```

In [22]:
```
df_combined['gun per capita'].describe()
```

Out[22]:
```
count    50.000000
mean      0.922652
std       0.767042
min       0.113772
25%       0.627696
50%       0.823736
75%       1.060950
max       5.614070
Name: gun per capita, dtype: float64
```

In [23]:
```
# Visualizing gun per capita by each states
ax= df_combined.plot(x='key_0', y='gun per capita',kind='bar',figsize=(
ax.set(title='Firearms Per Capita Across Each State',
xlabel='states',
ylabel='Percentage of Firearm Per Capita')
ax.legend().set_visible(True)
```



Result

Kentucky has the most gun per capita across all U.S territories whi
ch have a number of 5.6 firearm per capita.

Question # 2

Which states have had the highest growth in gun registration?

To answer this question we will looking into the firearm permit vs.
total population. Also, we will be looking into the moving average
of gun permit of the state with the highest number of firearm perm
it.

In [24]:
```
# Retrieving the permit
df_combined['permit'].max()
```

Out[24]: 20400718.0

In [25]:
```
# Finding the state with highest permit
high_ = df_combined.query('permit == 20400718.0')
high_
```

Out[25]:

	key_0	Jul-1-2016 population	totals	permit	gun per capita
16	Kentucky	4436974	24909483	20400718.0	5.61407

Calculating moving average of permit growth

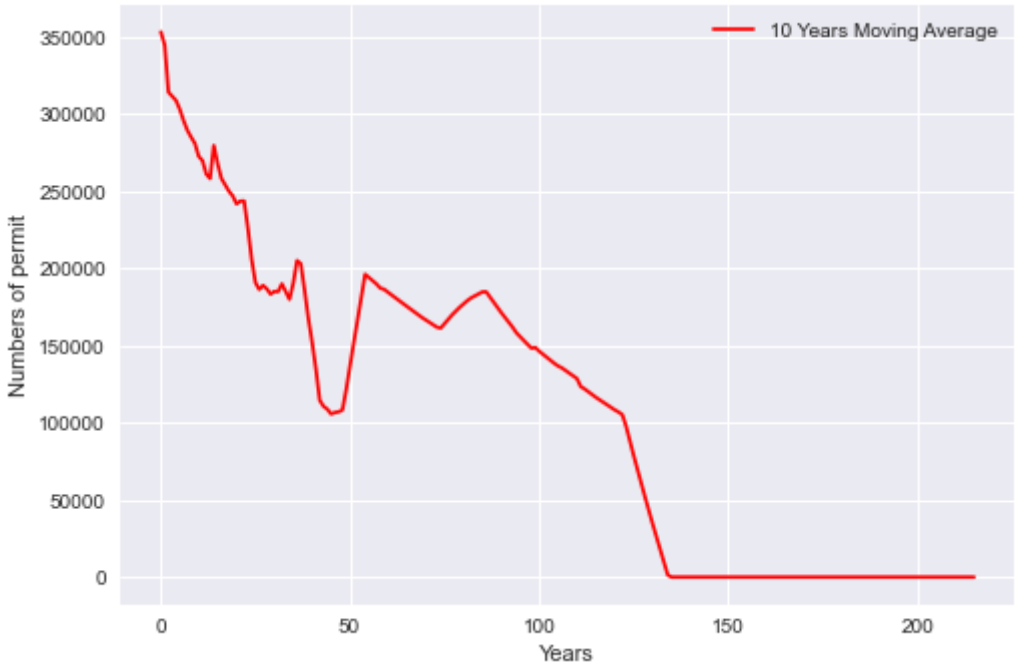
In [26]:
```
# Extracting the original FBI gun dataset
gun_info.head()
```

Out[26]:

	month	state	permit	totals
0	2017-09	Alabama	16717.0	32019
1	2017-09	Alaska	209.0	6303
2	2017-09	Arizona	5069.0	28394
3	2017-09	Arkansas	2935.0	17747
4	2017-09	California	57839.0	123506

In [27]:
```
# Calculate the yearly moving average of permit of Kentucky state
x = gun_info.query('state == "Kentucky"')
x = x.set_index('month')
x1 = x['permit']
```

In [28]:
```
# Using numpy to calculate the moving average
def moving_average(x,w):
    return np.convolve(x,np.ones(w),'valid')/w
y = moving_average(x1,12)
y = pd.DataFrame({'10 Years Moving Average': y})
y.plot(color='red');
plt.ylabel('Numbers of permit')
plt.xlabel('Years');
```



Result

Look like some of the moving average didnt start until later years and eventually there is a near bottom out dips of permits growth in the state of Kentucky.

What is the overall trend of gun purchases?

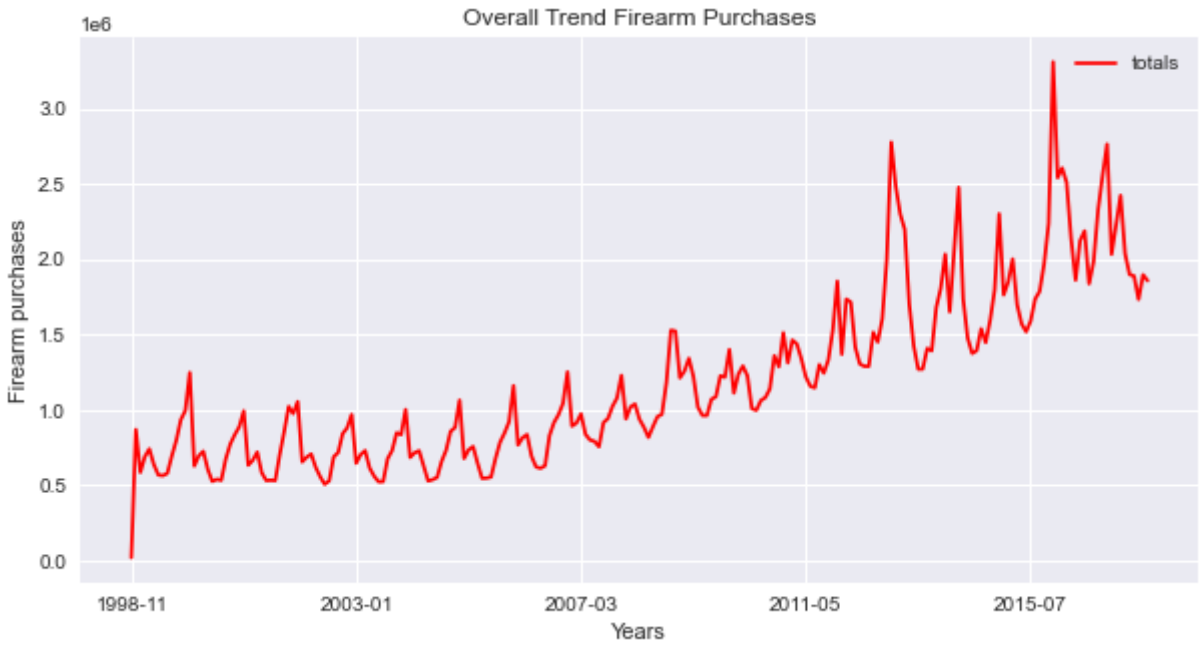
In [29]:
```
# Importing the new FBI gun dataset
over_all = pd.read_csv('gun_data.csv')
over_all.head()
```

Out[29]:

	month	state	permit	permit_recheck	handgun	long_gun	other	multiple	admin	prepawı
0	2017-09	Alabama	16717.0	0.0	5734.0	6320.0	221.0	317	0.0	
1	2017-09	Alaska	209.0	2.0	2320.0	2930.0	219.0	160	0.0	
2	2017-09	Arizona	5069.0	382.0	11063.0	7946.0	920.0	631	0.0	
3	2017-09	Arkansas	2935.0	632.0	4347.0	6063.0	165.0	366	51.0	
4	2017-09	California	57839.0	0.0	37165.0	24581.0	2984.0	0	0.0	

5 rows × 27 columns

In [30]:
```
# Finding the overall trend of gun purchases by using groupby on month co
over_all= over_all.groupby('month').agg({'totals':'sum'})
over_all.plot(color='red',figsize=(10,5))
plt.ylabel('Firearm purchases')
plt.xlabel('Years')
plt.title('Overall Trend Firearm Purchases');
```



Result

The overall trend of firearm purchases increases overtime despite f
or the continuous volatility.

In [ ]:
```
1
```