COMP5721M: Programming for Data Science

Coursework 3: Data Analysis Project

US Gun Death Data Analysis (1999 - 2019)

- *Achyuth Jayakumar Maneesha*, *mm22ajm*@leeds.ac.uk
- *Nikhil Shivasankar*, *mm22ns*@leeds.ac.uk
- *Rohini Krishna Preetha*, *mm22rkp*@leeds.ac.uk
- *Unnikrishnan Radhakrishnan*, *mm22ur*@leeds.ac.uk

Project Plan

The Data (10 marks)

Recent headlines made it very evident how much the US has been harmed by gun fatalities over the past few decades. This project aims to examine the US county-level gun fatalities from 1999 to 2019. The link to the original dataset is https://www.kaggle.com/datasets/ahmedeltom/us-gun-deaths-by-county-19992019 (https://www.kaggle.com/datasets/ahmedeltom/us-gun-deaths-by-county-19992019).

The dataset is sourced and edited from data.world. The data is accurate as it is taken from trusted and reliable sources as the dataset was created from the U.S.-Mexico border regions 1999–2019 multiple causes of death data from the 2020–released CDC WONDER Online Database. The Multiple Cause of Death Files, 1999–2019, was created using the information provided by the 57 jurisdictions that handle vital statistics under the Vital Statistics Cooperative Program.

The parameters are:

- **Year:** Data type: integer- Year of fire arm discharge
- **County:** Data type: string- Name of the county where the fire arm discharge occurred
- **State Name:** Data type: string- Name of the state where the firearm discharge occurred
- **Deaths:** Data type: integer- Count of deaths caused by firearm discharge in a given year
- **Population:** Data type: integer- The county's population for the given year
- **Crude Rate:** Data type: float- firearms deaths per 100,000 population for the given county and year

- **Crude Rate Lower 95% Confidence Interval:** Data type: float- The lower limit of the interval in which out 95% of the crude rate are true
- **Crude Rate Upper 95% Confidence Interval:** Data type: float- The upper limit of the interval in which out 95% of the crude rate are true
- **Age Adjusted Rate:** Data type: float- Age adjusted rates are weighted averages of age-specific death rates, with the weights corresponding to a given population by age. They are used to compare the relative risk of death between groups and across time. An age-adjusted rate is the rate that would have existed if the year's age-specific rates had prevailed in a population with the same age distribution as the fixed population. Age-adjusted mortality rates should be regarded as relative indices rather than direct or actual indicators of mortality risk.
- **Age Adjusted Rate Lower 95% Confidence Interval:** Data type: float- The lower limit of the interval in which out 95% of the age adjusted rate are true
- **Age Adjusted Rate Upper 95% Confidence Interval:** Data type: float- The upper limit of the interval in which out 95% of the age adjusted rate are true

Project Aim and Objectives (5 marks)

The aim of this project is to analyze about the gun deaths occured in United States between 1999 and 2019. This project's objective can be categorised as:

- Cleaning the data
- Analyze the shape and size of the data
- Observe the Year of Death v/s Number of Death
- Review the correlation between Age adjusted rate and Population
- Analyze the states which have the most number of deaths based on states
- Search for the Safest state to reside in US
- Analyze the general trend using visualization
- Determine the states with rates of gun violence are constantly increasing
- Compare the results of the analysis using the crude rate and the total number of deaths.
- Analyse the correlation between 'Age Adjusted Rate', 'Crude Rate', 'Year' and 'Population'

Specific Objective(s)

- **Objective 1:** *Identify the top 10 states where gun deaths are most prevalent.*
- **Objective 2:** *Determine which counties and states have the highest crude rates for the five years with the most fatalities.*
- **Objective 3:** *Determine which state are the safest and riskiest to live in.*
- **Objective 4:** *Find the average age of the population in the top 10 states with the highest crude rate. Also analyse the correlation between population and age adjusted rate.*

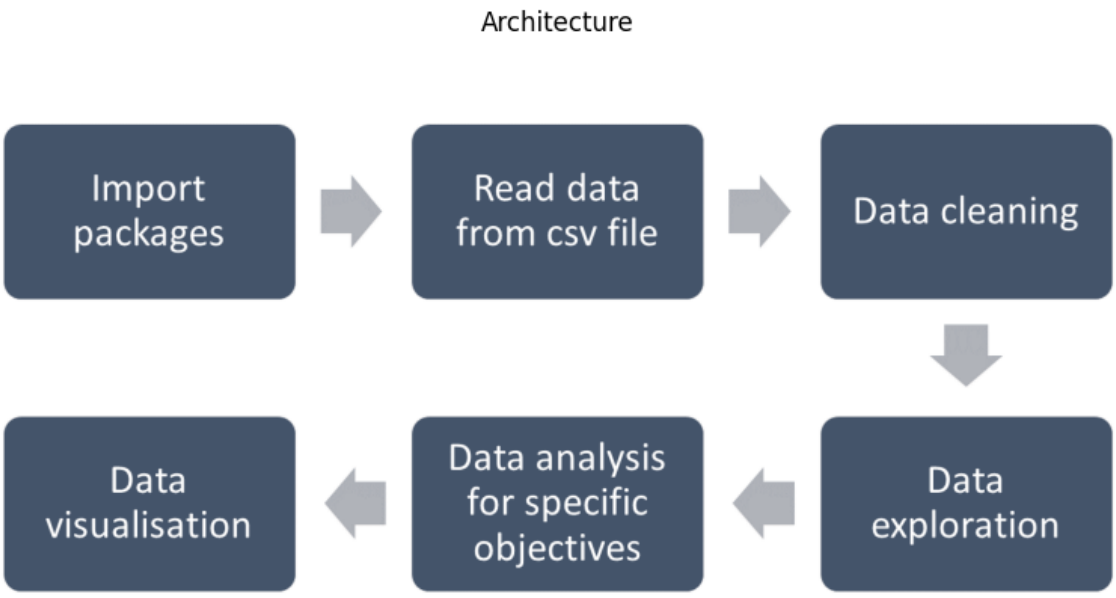
System Design (5 marks)

Architecture

In [56]:
```
#Importing all the necessary packages
import pandas as pd #for handling dataframe
import matplotlib.pyplot as plt #to make changes in visualization
import numpy as np # for working with arrays
import matplotlib.colors as mcolors #for colors in visualization
import matplotlib.image as image #for plotting images
import math #for mathematical functions
import seaborn as sns #for high-level interfaces
from scipy.stats import spearmanr #for probability distributions an
import warnings #to exclude warnings
warnings.filterwarnings('ignore')
```

The architecture for the proposed project can be be summarised as shown in Fig a.

In [57]:
```
img=image.imread('flowchart.png')
plt.figure(figsize=(10,10))
plt.axis('off')
plt.title('Architecture')
architecture=plt.imshow(img)
```



Import packages: Import the necessary packages for the project.

Read data from csv file: Use pandas dataframe to read and load the data from the selected 'gun deaths us 1999 2019.csv' file.

Data cleaning: Remove records that have null values in columns that are needed for calculations and analysis.

Data exploration: Find the outliers and check if they fall within their respective intervals.

Data analysis for specific objectives: Analyse the data for specified objectives.

Data visualisation: Generate visual representations using barcharts, line graphs, boxplots, heatmaps, etc. to analyse the results.

Processing Modules and Algorithms

The main processing modules for this project are:

- **Importing packages:** Firstly import the packages such as pandas, matplotlib, numpy, math, seaborn, and scipy to do calculations and plot visuals necessary for the project.
- **Loading the data into a dataframe:** Load the data from 'gun deaths us 1999 2019.csv' file into the dataframe 'GUN DEATH DF.'
- **Data cleaning:** Remove 'GUN DEATH DF' records with null values in the column 'Crude Rate'. Eliminate unwanted columns such as 'County Code,' 'State,' and 'State Code' from 'GUN DEATH DF'.
- **Findining Outliers:** Use boxplots for 'Crude Rate' and 'Age Adjusted Rate' to look for outliers. Examine 'Age Adjusted Rate' and 'Crude Rate' to see if they fall within the bounds of their respective upper and lower confidence intervals.
- **Implementing functions to achieve specific objectives:** Perform qualitative and quantitative analysis for each specific objective using pandas dataframe manipulation techniques.
- **Data Visualization:** Finally visualise the data using boxplot, line graph, bar graph, heatmap, etc. to analyse the results of each specific objective.

Program Code (15 marks)

Importing the chosen csv file into the 'GUN_DEATH_DF' dataframe

In [58]:
```
GUN_DEATH_DF=pd.read_csv("gun_deaths_us_1999_2019.csv")

GUN_DEATH_DF.columns
```

Out[58]:
```
Index(['Unnamed: 0', 'Year', 'County', 'County Code', 'State', 'St
ate_Name',
       'State Code', 'Deaths', 'Population', 'Crude Rate',
       'Crude Rate Lower 95% Confidence Interval',
       'Crude Rate Upper 95% Confidence Interval', 'Age Adjusted R
ate',
       'Age Adjusted Rate Lower 95% Confidence Interval',
       'Age Adjusted Rate Upper 95% Confidence Interval'],
      dtype='object')
```

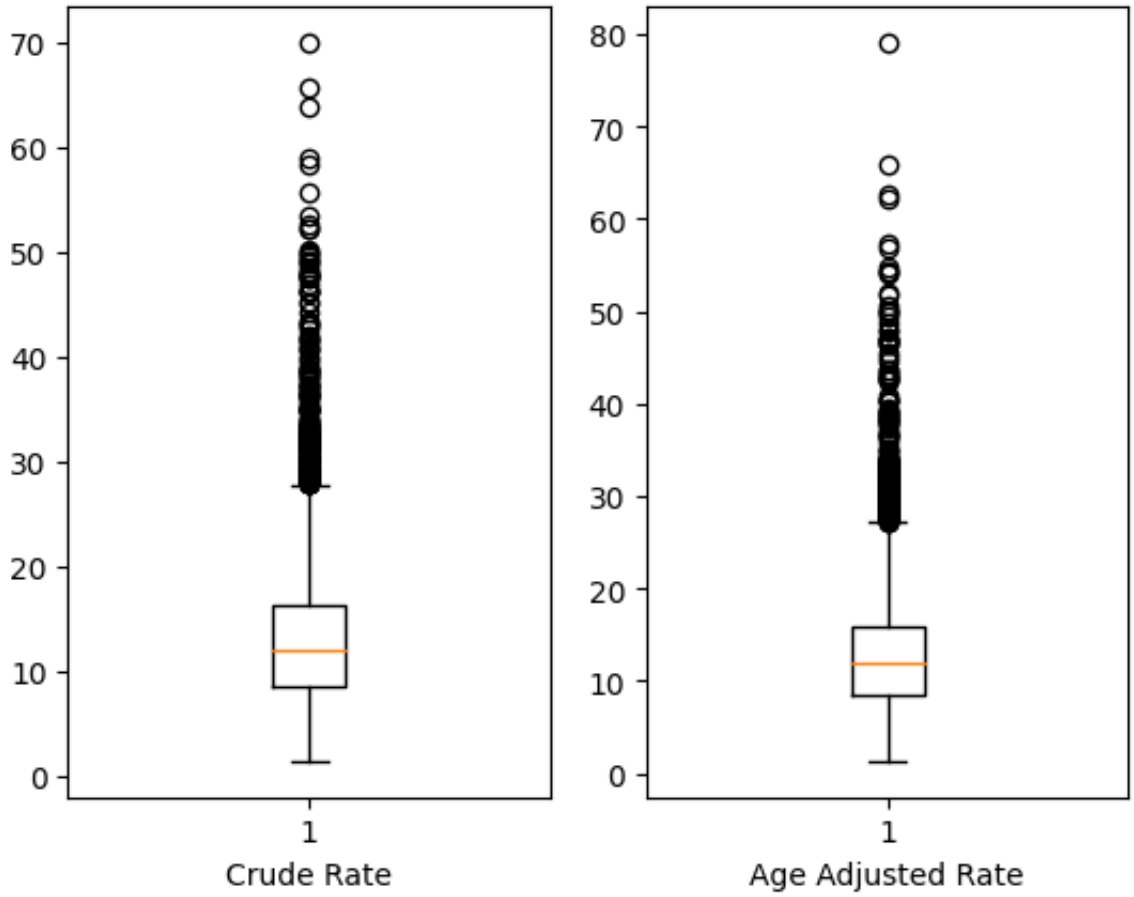
Data Cleaning

Eliminating null values from the Crude Rate column, removing unwanted columns such as 'County Code', 'State' and' State Code' from the dataframe and the copying the cleansed data to a data frame 'GUN_DEATH_DF'

In [59]:
```
GUN_DEATH_DF=GUN_DEATH_DF[GUN_DEATH_DF["Crude Rate"].notnull()]

GUN_DEATH_DF=GUN_DEATH_DF[['Year','County','State_Name','Deaths','P

GUN_DEATH_DF.describe()
GUN_DEATH_DF.dtypes
```

Out[59]:
```
Year                                                 int64
County                                              object
State_Name                                          object
Deaths                                               int64
Population                                           int64
Crude Rate                                         float64
Crude Rate Lower 95% Confidence Interval           float64
Crude Rate Upper 95% Confidence Interval           float64
Age Adjusted Rate                                  float64
Age Adjusted Rate Lower 95% Confidence Interval    float64
Age Adjusted Rate Upper 95% Confidence Interval    float64
dtype: object
```

In [60]:
```
#boxplot to find outliers
fig,(ax1,ax2)= plt.subplots(1,2)

ax1.boxplot(GUN_DEATH_DF["Crude Rate"])
ax1.set(xlabel='Crude Rate')
ax2.boxplot(GUN_DEATH_DF["Age Adjusted Rate"])
ax2.set(xlabel='Age Adjusted Rate')

plt.show()
```



In [61]:
```
#checking if age adjusted and crude rates lie in 95% Confidence Int
(GUN_DEATH_DF["Crude Rate"]>= GUN_DEATH_DF["Crude Rate Lower 95% Co
```

Out[61]: 6871

In [62]:
```
(GUN_DEATH_DF["Crude Rate"]<= GUN_DEATH_DF["Crude Rate Upper 95% Co
```

Out[62]: 6871

In [63]:
```
(GUN_DEATH_DF["Age Adjusted Rate"]>= GUN_DEATH_DF["Age Adjusted Rat
```

Out[63]: 6871

In [64]:
```
(GUN_DEATH_DF["Age Adjusted Rate"]<= GUN_DEATH_DF["Age Adjusted Rat
```

Out[64]: 6871

General Data Analysis

Before moving on to our specific objectives, we will do several studies on the data (as indicated below) to better understand our data, such as information about fatalities in various states and so on.

Prior to viewing the dataframe as a whole, the dataframe's size and shape are assessed.

In [65]:
```
GUN_DEATH_DF.shape
```

Out[65]: (6871, 11)

In [66]:
```
GUN_DEATH_DF.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 6871 entries, 0 to 14355
Data columns (total 11 columns):
 #   Column                                             Non-Null Cou
nt  Dtype
---  ------                                             ------------
--  -----
 0   Year                                               6871 non-nul
l   int64
 1   County                                             6871 non-nul
l   object
 2   State_Name                                         6871 non-nul
l   object
 3   Deaths                                             6871 non-nul
l   int64
 4   Population                                         6871 non-nul
l   int64
 5   Crude Rate                                         6871 non-nul
l   float64
 6   Crude Rate Lower 95% Confidence Interval           6871 non-nul
l   float64
 7   Crude Rate Upper 95% Confidence Interval           6871 non-nul
l   float64
 8   Age Adjusted Rate                                  6871 non-nul
l   float64
 9   Age Adjusted Rate Lower 95% Confidence Interval    6871 non-nul
l   float64
 10  Age Adjusted Rate Upper 95% Confidence Interval    6871 non-nul
l   float64
dtypes: float64(6), int64(3), object(2)
memory usage: 644.2+ KB
```

In [67]:
```
GUN_DEATH_DF
```

Out[67]:

	Year	County	State_Name	Deaths	Population	Crude Rate	Crude Rate Lower 95% Confidence Interval	Crude Rate Upper 95% Confidence Interval	A
0	1999	Baldwin County	Alabama	22	137555	15.99	10.02	24.21	
1	1999	Calhoun County	Alabama	29	114910	25.24	16.90	36.24	
6	1999	Etowah County	Alabama	23	104002	22.11	14.02	33.18	
8	1999	Jefferson County	Alabama	149	662845	22.48	18.87	26.09	
12	1999	Madison County	Alabama	27	274692	9.83	6.48	14.30	
...	...	...	...	...	...	...	...	...	
14335	2018	Laramie County	Wyoming	27	98976	27.28	17.98	39.69	
14336	2018	Natrona County	Wyoming	25	79115	31.60	20.45	46.65	
14339	2019	Laramie County	Wyoming	28	99500	28.14	18.70	40.67	
14340	2019	Natrona County	Wyoming	24	79858	30.05	19.26	44.72	
14355	2018	Cass County	North Dakota	20	181516	11.02	6.73	17.02	

6871 rows × 11 columns

The dataframe "GUN_DEATH_DF" is then grouped into a new dataframe based on the year and state name.

In [68]:
```
#Grouping the dataframe with respect to year and statename to a new

DF_grpd_by_state_n_yr = GUN_DEATH_DF.groupby(['Year', 'State_Name'])
DF_grpd_by_state_n_yr
```

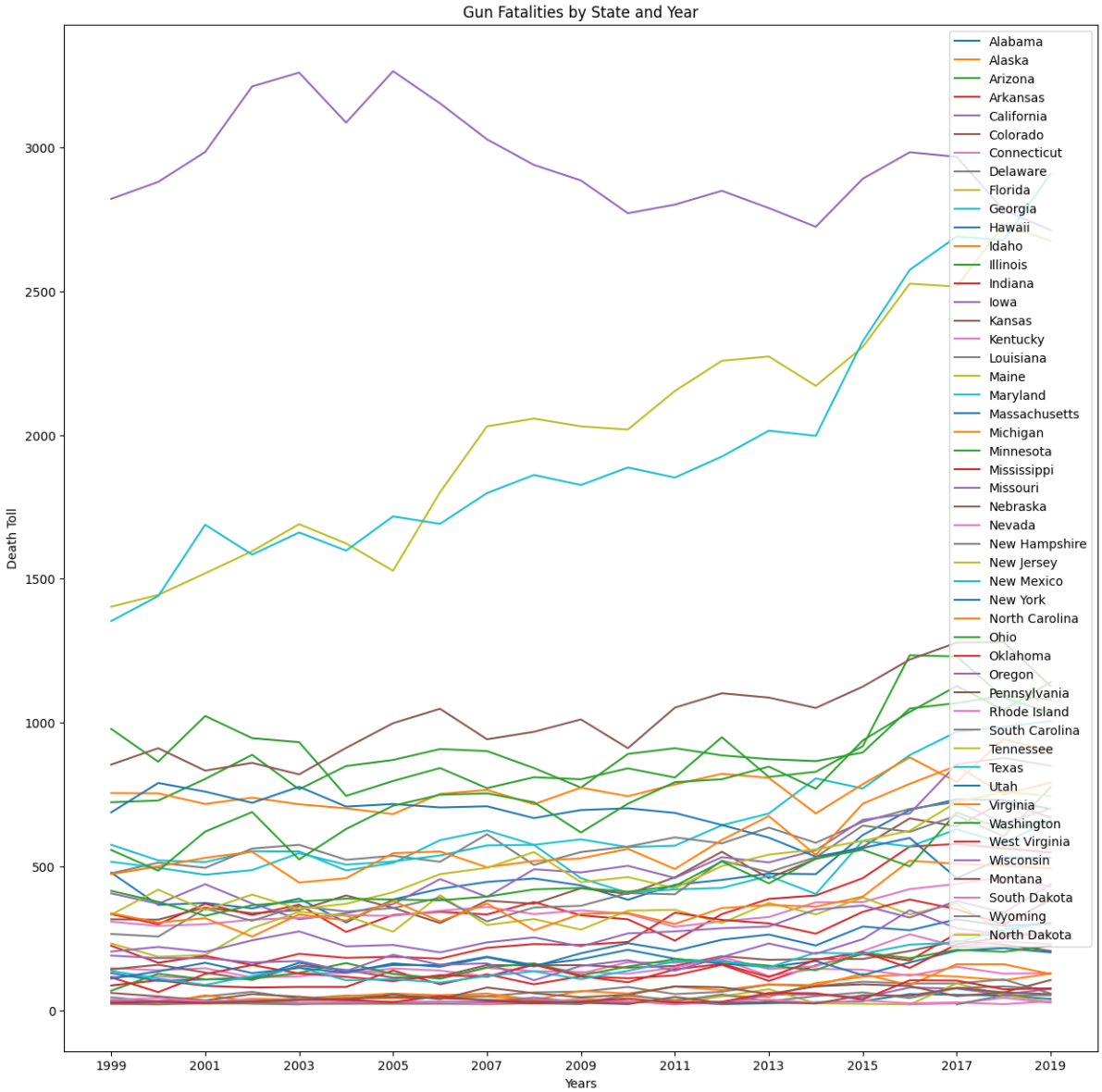
Out[68]:

	Year	State_Name	Deaths	Population	Crude Rate	Crude Rate Lower 95% Confidence Interval	Crude Rate Upper 95% Confidence Interval	Age Adjusted Rate	Ra Co
0	1999	Alabama	478	2374564	229.29	156.86	325.04	227.11	
1	1999	Alaska	27	259348	10.41	6.86	15.15	10.91	
2	1999	Arizona	723	4320548	97.48	73.79	126.85	92.22	
3	1999	Arkansas	86	508948	31.78	22.39	44.08	31.66	
4	1999	California	2821	30500740	207.99	158.77	268.28	213.07	
...	...	...	...	...	...	...	...	...	
939	2019	Virginia	493	4322491	235.73	159.53	337.05	245.10	
940	2019	Washington	655	6237403	118.48	87.00	157.77	117.86	
941	2019	West Virginia	76	370656	65.62	42.04	97.73	64.49	
942	2019	Wisconsin	269	2527033	63.09	43.76	88.33	60.64	
943	2019	Wyoming	52	179358	58.19	37.96	85.39	60.08	

944 rows × 10 columns

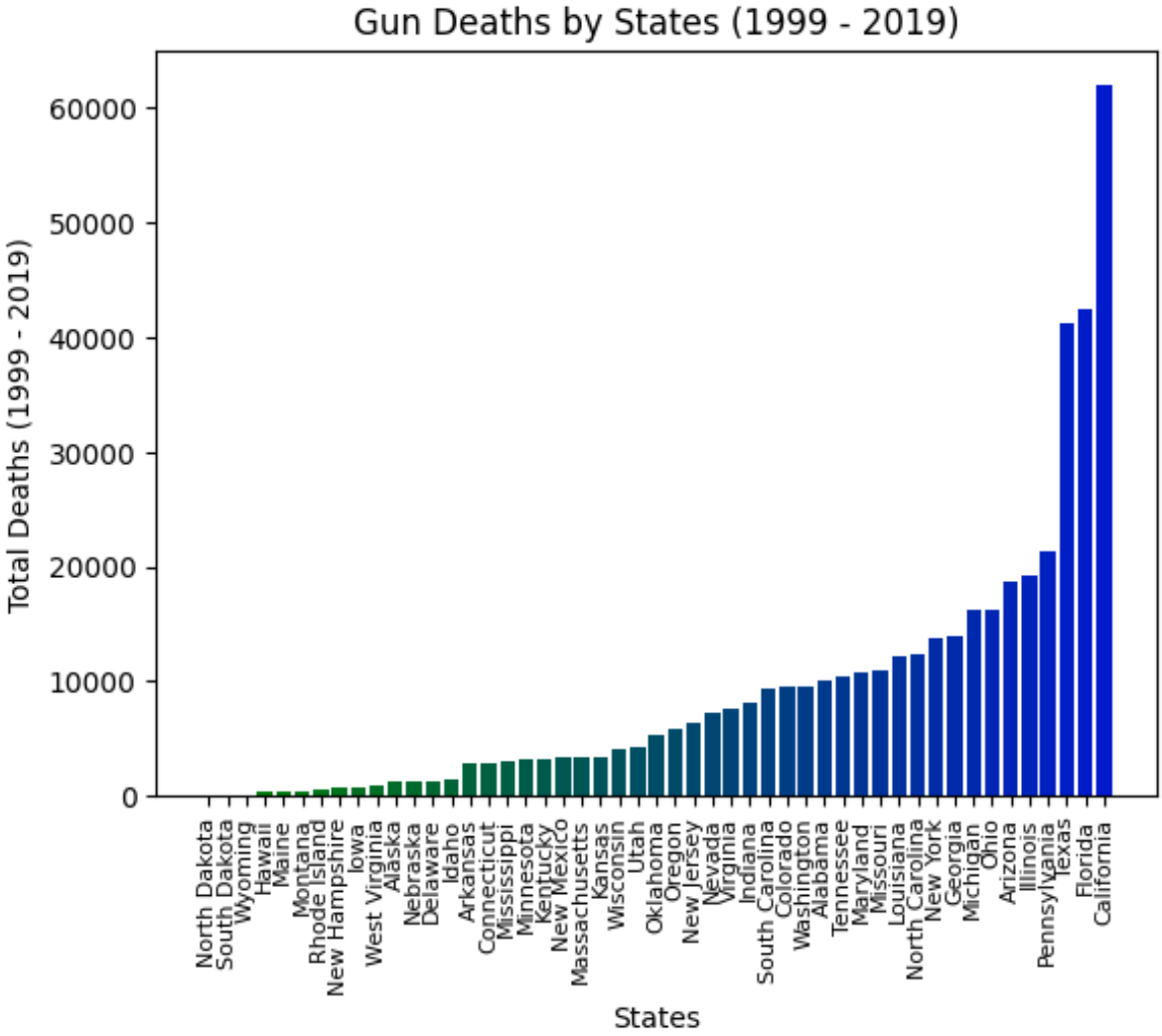
Using the dataframe mentioned above, we can utilise visualisation (in this case, a line graph) to show how the number of fatalities varies across all states for each year.

In [69]:
```
# Plot deaths by year for all states

lgnd = []
plt.figure(figsize=(15,15))
for state in DF_grpd_by_state_n_yr['State_Name'].unique():
    plt.plot(DF_grpd_by_state_n_yr[DF_grpd_by_state_n_yr.State_Name
    lgnd.append(state)
plt.xticks(np.arange(DF_grpd_by_state_n_yr['Year'].min(), DF_grpd_b
plt.legend(lgnd, loc = "upper right")
plt.title("Gun Fatalities by State and Year")
plt.xlabel("Years")
plt.ylabel("Death Toll")
plt.show()
```



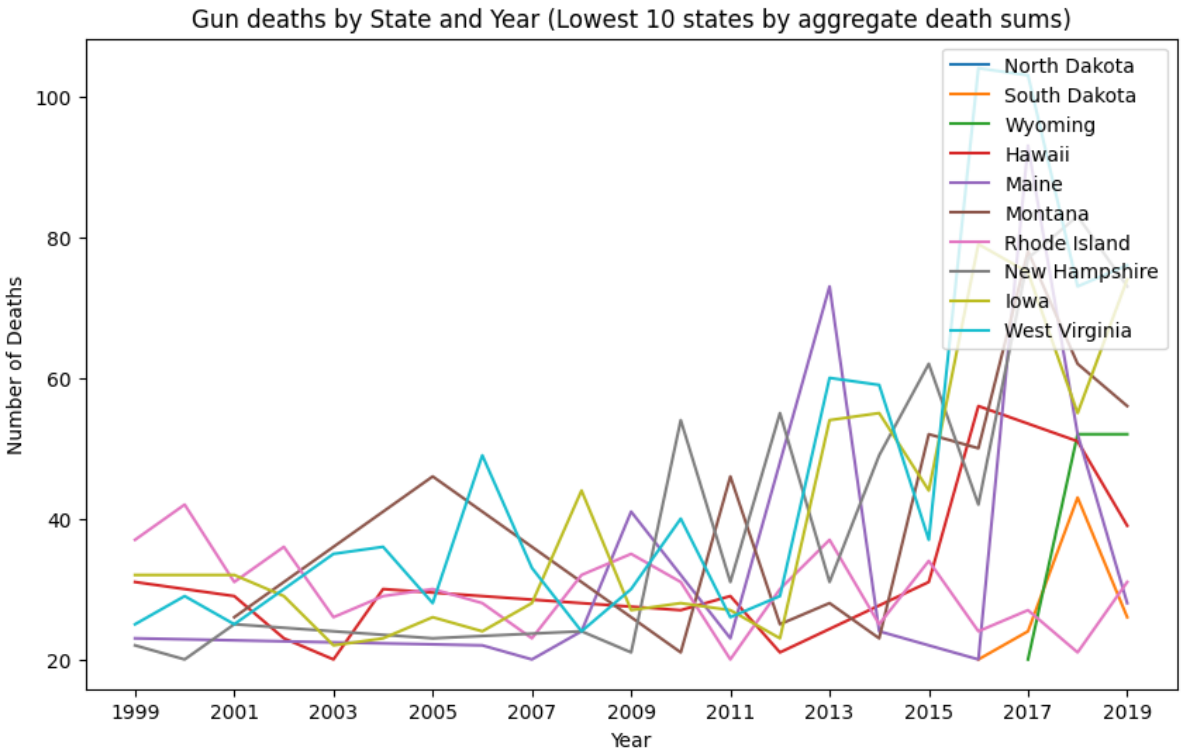
We then create a dataframe that contains the aggregate deaths per state in order to see the gun fatalities in various US states from the years 1999 to 2019. Through visualisation, we can see how the number of gun deaths varies between the different states.

In [70]:
```
#The aggregate deaths by the different states throughout the years


#Creating the required dataframe
DF_agg_dths_by_state = GUN_DEATH_DF.groupby("State_Name").agg("sum"
DF_agg_dths_by_state = DF_agg_dths_by_state.sort_values(by="Deaths"

#Visualisation of deaths in differeent states via bar graph
color_list = [(0, 'green'), (1, 'blue')]
N = 60
x = np.arange(N).astype(float)
rvb = mcolors.LinearSegmentedColormap.from_list("", color_list)
plt.bar(DF_agg_dths_by_state['State_Name'], DF_agg_dths_by_state['D
plt.xticks(fontsize=8, rotation=90)
plt.xlabel("States")
plt.ylabel("Total Deaths (1999 - 2019)")
plt.title("Gun Deaths by States (1999 - 2019)")
plt.show()
```



A new dataframe is created to store the data of 10 states with the lowest aggregate gun deaths. Number of gun deaths in each of these states from 1999 to 2019 can be observed in the plot

In [71]:
```
# To find the 10 states with the lowest aggregate gun deaths
states_lwst10_deaths = GUN_DEATH_DF.groupby('State_Name').agg('sum'
plt.figure(figsize=(10,6))
legend = []
for state in states_lwst10_deaths:
    legend.append(state)
    plt.plot(DF_grpd_by_state_n_yr[DF_grpd_by_state_n_yr.State_Name
plt.xticks(np.arange(DF_grpd_by_state_n_yr['Year'].min(), DF_grpd_b
plt.legend(legend, loc = "upper right")
plt.title("Gun deaths by State and Year (Lowest 10 states by aggreg
plt.xlabel("Year")
plt.ylabel("Number of Deaths")
plt.show()
```



Objective 1

The aim of this study is to identify the top 10 states where gun fatalities have occurred most frequently throughout time. As a result, we have categorised the dataframe "GUN_DEATH_DF" according to the year and state name into a new dataframe, which has been called.

In [72]:
```
#Objective 1
#Identifying the top 10 states where gun deaths are most prevalent

DF_grpd_by_state_n_yr
```

Out[72]:

	Year	State_Name	Deaths	Population	Crude Rate	Crude Rate Lower 95% Confidence Interval	Crude Rate Upper 95% Confidence Interval	Age Adjusted Rate	Ra Cc
0	1999	Alabama	478	2374564	229.29	156.86	325.04	227.11	
1	1999	Alaska	27	259348	10.41	6.86	15.15	10.91	
2	1999	Arizona	723	4320548	97.48	73.79	126.85	92.22	
3	1999	Arkansas	86	508948	31.78	22.39	44.08	31.66	
4	1999	California	2821	30500740	207.99	158.77	268.28	213.07	
...	...	...	...	...	...	...	...	...	
939	2019	Virginia	493	4322491	235.73	159.53	337.05	245.10	
940	2019	Washington	655	6237403	118.48	87.00	157.77	117.86	

The dataframe "top_states_deaths" is then created by categorising the dataframe "GUN_DEATH_DF" according to state name, that recorded the top 10 most fatalities.

In [73]:
```
#Creating a new dataframe with top 10 deaths per state

top_states_deaths = GUN_DEATH_DF.groupby('State_Name').agg('sum').r
top_states_deaths
```

Out[73]:
```
4       California
8          Florida
42           Texas
37    Pennsylvania
12        Illinois
2          Arizona
34            Ohio
21        Michigan
9          Georgia
31        New York
Name: State_Name, dtype: object
```

To observe how the deaths in these ten states compare to one another, we choose to visualise the aforementioned dataframe using a line graph.

In [74]:
```
#A Line graph is used to visualise the top 10 fatalities per states

def top10states_line():
    plt.figure(figsize=(10,6))
    legend = []
    for state in top_states_deaths:
        legend.append(state)
        plt.plot(DF_grpd_by_state_n_yr[DF_grpd_by_state_n_yr.State_
    plt.xticks(np.arange(DF_grpd_by_state_n_yr['Year'].min(), DF_gr
    plt.legend(legend, loc = "upper right")
    plt.title("Fig 1.1: Gun deaths in State and Year (Top 10 States
    plt.xlabel("Year")
    plt.ylabel("Number of Deaths")
    plt.show()
```

Similarly, we may verify with the aid of a bar graph.

In [75]:
```
#bar plot
def top10states_bar():

    DF_grpd_by_state_n_yr = GUN_DEATH_DF.groupby(['Year', 'State_Na
    legend=[]
    for state in top_states_deaths:
        legend.append(state)
        plt.bar(DF_grpd_by_state_n_yr[DF_grpd_by_state_n_yr.State_N
    plt.xticks(np.arange(DF_grpd_by_state_n_yr['Year'].min(), DF_gr
    plt.legend(legend, loc = "upper right")
    plt.title("Fig 1.2: Gun deaths in State and Year (Top 10 States
    plt.xlabel("Year")
    plt.ylabel("Number of Deaths")
    plt.show()
```

It is obvious from the analysis above that California is the state where most deaths occur. The function below reveals which county in California had the highest death rate.

In [76]:
```
# Shootings by county in California

#We know most deaths occured in this state
def cali_county_death():
    State_name = "California"

    DF_state_shooting = GUN_DEATH_DF.groupby(['State_Name', 'County
    DF_state_shooting = DF_state_shooting[DF_state_shooting['State_
    N = len(DF_state_shooting['County'])
    plt.figure(figsize = (10, 4))
    plt.bar(DF_state_shooting['County'], DF_state_shooting['Deaths'
    plt.xticks(fontsize=12, rotation=90)
    plt.title("Fig 1.3: {} State Counties and Gun Fatalities (1999
    plt.xlabel("County Name ({})".format(State_name))
    plt.ylabel("Aggregate Gun Fatalities (1999 -2019)")
    plt.show()
```

Objective 2

The aim of this objective is to find the years having the highest total number of gun deaths and analyse the death patterns statewise and countywise

Dictionary created to list the aggregate death in each year

In [77]:
```
death_agg=GUN_DEATH_DF.groupby("Year")["Deaths"].sum().sort_values(
death_agg #display the dictionary to represent the count of aggrega
```

Out[77]:
```
{2017: 27522,
 2019: 27491,
 2018: 27202,
 2016: 26278,
 2015: 24130,
 2012: 21699,
 2013: 21684,
 2014: 21534,
 2011: 20395,
 2007: 20039,
 2010: 19999,
 2008: 19945,
 2009: 19631,
 2006: 19606,
 2005: 19106,
 2003: 18901,
 2002: 18626,
 2001: 18302,
 2004: 18184,
 1999: 17588,
 2000: 17385}
```

Function defined to show the aggregate death in each year

In [78]:
```
def year_vs_death_agg():
    DF_aggr_dths_by_yr = GUN_DEATH_DF.groupby("Year").agg("sum").re
    plt.figure(figsize = (7, 7))
    plt.pie(DF_aggr_dths_by_yr['Deaths'], labels=DF_aggr_dths_by_yr
    plt.title("Fig 2.1: Total Gun Deaths in USA (1999 - 2019)")
    plt.show()
```

Dataframe used to obtain information of gun deaths from top 5 death years

In [79]:
```
DB_COPY=GUN_DEATH_DF[(GUN_DEATH_DF["Year"]== 2015) | (GUN_DEATH_DF[
```

Function to plot the relation between state vs aggregate death for each year

In [80]:
```
def state_vs_death_agg():
    DB2= DB_COPY.groupby("State_Name").mean("Deaths").sort_values("
    plt.figure(figsize = (12, 6)) #setting figure size
    plt.bar(DB2["State_Name"],DB2["Deaths"]) #barplot of state name
    plt.xlabel("State")
    plt.ylabel("Number of deaths")
    plt.title("Fig 2.2: State vs Number of deaths")
    plt.xticks(fontsize=12,rotation=90) #rotating the label for bet
    plt.show()
```

Function to visualise the crude rate with respect to each state for the last 5 years

In [81]:
```
def cruderate_vs_state():
    DB1= DB_COPY.groupby("State_Name").mean("Crude Rate").sort_valu
    plt.figure(figsize = (12, 6))
    plt.bar(DB1["State_Name"],DB1["Crude Rate"]) #barplot of state
    plt.title("Fig 2.3: State vs Crude Rate ")
    plt.xlabel("States")
    plt.ylabel("Crude Rate")
    plt.xticks(fontsize=12,rotation=90)
    plt.show()
```

Function to plot County vs Crude Rate with highest rates

In [82]:
```
def county_vs_cruderate_largest():
    DB3=DB_COPY.groupby("County").mean("Crude Rate").sort_values(by
    plt.figure(figsize = (12, 6))
    plt.plot(DB3["County"],DB3["Crude Rate"])
    plt.title("Fig 2.4: County vs Crude Rate Highest")
    plt.xlabel("County")
    plt.ylabel("Crude Rate")
    plt.xticks(fontsize=12,rotation=90)
    plt.show()
```

Function to plot County vs Crude Rate with smallest rates

In [83]:
```
def county_vs_cruderate_smallest():
    DB4=DB_COPY.groupby("County").mean("Crude Rate").sort_values(by
    plt.figure(figsize = (12, 6))
    plt.plot(DB4["County"],DB4["Crude Rate"])
    plt.title("Fig 2.5: County vs Crude Rate Smallest")
    plt.xticks(fontsize=12,rotation=90)
    plt.xlabel("County")
    plt.ylabel("Crude Rate")
    plt.show()
```

Objective 3

The dataset is arranged by county. Hence the dataset is grouped by state names to determine the safest and riskiest states to live in the United States.

In [84]:
```
dff=GUN_DEATH_DF.groupby(['State_Name']).sum('Deaths').reset_index(
```

Since utilising the mean of the crude rate yields unreliable results, the crude rate for each state is calculated independently using the population and death toll for each state.

In [85]:
```
dff=GUN_DEATH_DF.groupby(['State_Name']).sum('Deaths').reset_index(
dff["Crude Rate"]=[dff.loc[i]["Deaths"]/dff.loc[i]["Population"]*10
```

Since the only variables required to analyse the safest and riskiest states to reside in the United States are crude rate and state name, all columns other than those two are excluded.

In [86]:
```
dff=dff.filter(['State_Name','Crude Rate'], axis=1)
```

The new list is ordered in ascending order based on Crude rate, and the safest state is printed to determine which state is the safest to live in the USA.

In [87]:
```
 ##safest
dff=dff.sort_values(by=['Crude Rate'])
dff[:1]
```

Out[87]:

	State_Name	Crude Rate
10	Hawaii	3.317639

Similar to this, another list is ordered in descending order according to the crude rate, and the riskiest state is printed to determine which state in the United States is the riskiest to live in.

In [88]:
```
##Riskiest
dff=dff.sort_values(by=['Crude Rate'],ascending=False)
dff[:1]
```

Out[88]:

	State_Name	Crude Rate
48	Wyoming	27.206347

To plot a horizontal bar graph between Crude rate and State name.

In [89]:
```
dff=dff.sort_values(by=['Crude Rate'])
#Graph
def gr():
    X = dff['Crude Rate']
    Y = dff['State_Name']

    plt.figure(figsize=(10, 10))
    plt.barh(Y, X, color='g')
    plt.title("Fig 3.1: Crude rate in different states")
    plt.xlabel("Crude Rate")
    plt.ylabel("State Name")

    plt.show()
```

To plot a bar graph between County wise crude rate of the state Wyoming.

In [90]:
```
DF_county_high = GUN_DEATH_DF.groupby(['State_Name', 'County']).sum
DF_county_high = DF_county_high[DF_county_high['State_Name'] == "Wy
DF_county_high["Crude Rate"]=[DF_county_high.loc[i]["Deaths"]/DF_co

def gr2():
    N = len(DF_county_high['County'])
    #plt.figure(figsize = (2, 2))
    plt.bar(DF_county_high['County'], DF_county_high['Crude Rate'],
    plt.xticks(fontsize=12, rotation=90)
    plt.title("Fig 3.2: Wyoming Counties v/s Crude Rate")
    plt.xlabel("County Name")
    plt.ylabel("Crude Rate")
    plt.show()
```

To plot a bar graph between County wise crude rate of the state Hawaii.

In [91]:
```
DF_county_low = GUN_DEATH_DF.groupby(['State_Name', 'County']).sum(
DF_county_low = DF_county_low[DF_county_low['State_Name'] == "Hawai
DF_county_low["Crude Rate"]=[DF_county_low.loc[i]["Deaths"]/DF_coun

def gr3():
    N = len(DF_county_low['County'])
    plt.figure(figsize = (10, 4))
    plt.bar(DF_county_low['County'], DF_county_low['Crude Rate'], w
    plt.xticks(fontsize=12, rotation=90)
    plt.title("Fig 3.3: Hawaii Counties v/s Crude Rate")
    plt.xlabel("County Name")
    plt.ylabel("Crude Rate")
    plt.show()
```

Objective 4

To find the average age of the population in top 10 states with highest crude rate, GUN_DEATH_DF dataframe has been grouped by 'State_Name' and the mean of the 'Age Adjusted Rate' and sorted in the descending order of 'Crude Rate'. This result is stored in df_n dataframe.

In [92]:
```
df_n=GUN_DEATH_DF.groupby('State_Name').mean('Age Adjusted Rate').s
#Grouping the dataframe values by State_name using mean of Age Adju
```

The results can be visualised using the following bar charts.

Plotting State_Name v/s Crude Rate in a bar chart for top ten states with highest crude rate

In [93]:
```
def bar_crude():
    dfc=zip(df_n['State_Name'],df_n['Crude Rate'])# creating an ite
    plt.figure(figsize=(8,6)) #defining the figure size
    x,y=zip(*dfc)#mapping the items in tuple dfc to x and y
    plt.bar(x,y,color='green') #plotting the bar chart
    plt.xlabel('State Name')# labelling x axis
    plt.ylabel('Crude Rate')# labelling y axis
    plt.title('Fig 4.1: Top ten states with highest crude rate') #

    dfc=dict(zip(df_n['State_Name'],df_n['Crude Rate']))#dictionary

    DFA=sorted(dfc.items(),key=lambda x:(x[1],x[0])) #sorting in th

    for l,m in DFA:
        plt.text(l,m,format(m,'0.2f'),horizontalalignment='center',
    plt.xticks(rotation = 90) #aligning the x-axis label

    plt.show()
```

Plotting State_Name v/s Age Adjusted Rate in a bar chart for top ten states with highest crude rate

In [94]:
```
def bar_age_adjusted_rate():
    df=zip(df_n['State_Name'],df_n['Age Adjusted Rate'])# creating
    df_a=dict(zip(df_n['State_Name'],df_n['Age Adjusted Rate']))#di
    plt.figure(figsize=(8,6))#defining the figure size
    l,m=zip(*df)#mapping the items in tuple dfc to x and y
    plt.bar(l,m, color='midnightblue' ) #plotting the bar chart
    plt.xlabel('State Name')# labelling x axis
    plt.ylabel('Age Adjusted Rate')# labelling y axis
    plt.title('Fig 4.2: Average age of population in top 10 states

    DFA=sorted(df_a.items(),key=lambda x:(x[1],x[0])) #sorting in t

    for l,m in DFA:
        plt.text(l,m,format(m,'0.2f'),horizontalalignment='center',
    plt.xticks(rotation = 90)#aligning the x-axis label
    plt.show()
```

Analysing the correlation between Age Adjusted Rate and Population

In [95]:
```
coef, p = spearmanr(GUN_DEATH_DF['Age Adjusted Rate'], GUN_DEATH_DF[
print('%.3f'%coef)
```

```
-0.531
```

Plotting heatmap to analyse the correlation between 'Age Adjusted Rate', 'Crude Rate', 'Year' and 'Population'

In [96]:
```
def correlation_map():
    fig, ax = plt.subplots(figsize=(10,6))
    GUN_DEATH_DF_sub=GUN_DEATH_DF[['Age Adjusted Rate','Crude Rate'
    sns.heatmap(GUN_DEATH_DF_sub.corr(method='spearman'),annot=True
    plt.title('Fig 4.3: Correlation between Age Adjusted Rate, Crud
```

Project Outcome (10 + 10 marks)

Overview of Results

Firstly, the states with the highest number of gun fatalities over time were identified using number of deaths. These figures indicate that California had the highest number of gun fatalities.

Secondly, the analysis of gun deaths across states was then done using crude rate. Since the number of fatalities did not take geographical area or population into consideration, it was clear that the crude rate provided a more realistic analysis of the deaths.

According to an analysis based on states and crude rate, Wyoming has the highest crude rate. As a result, living there was the riskiest option. However, Hawaii has the lowest crude rate, making it the safest state to live in.

Moreover, based on counties and the crude rate, observations were made. They demonstrate that New York County has the lowest crude rate and Petersburg City has the highest. Another unsettling finding of this investigation is that in Wyoming, people in their late twenties are most likely to die from a gunshot. The investigation made it quite evident that gunfire has an impact on people of all ages, but particularly on young adults who are the society's future.

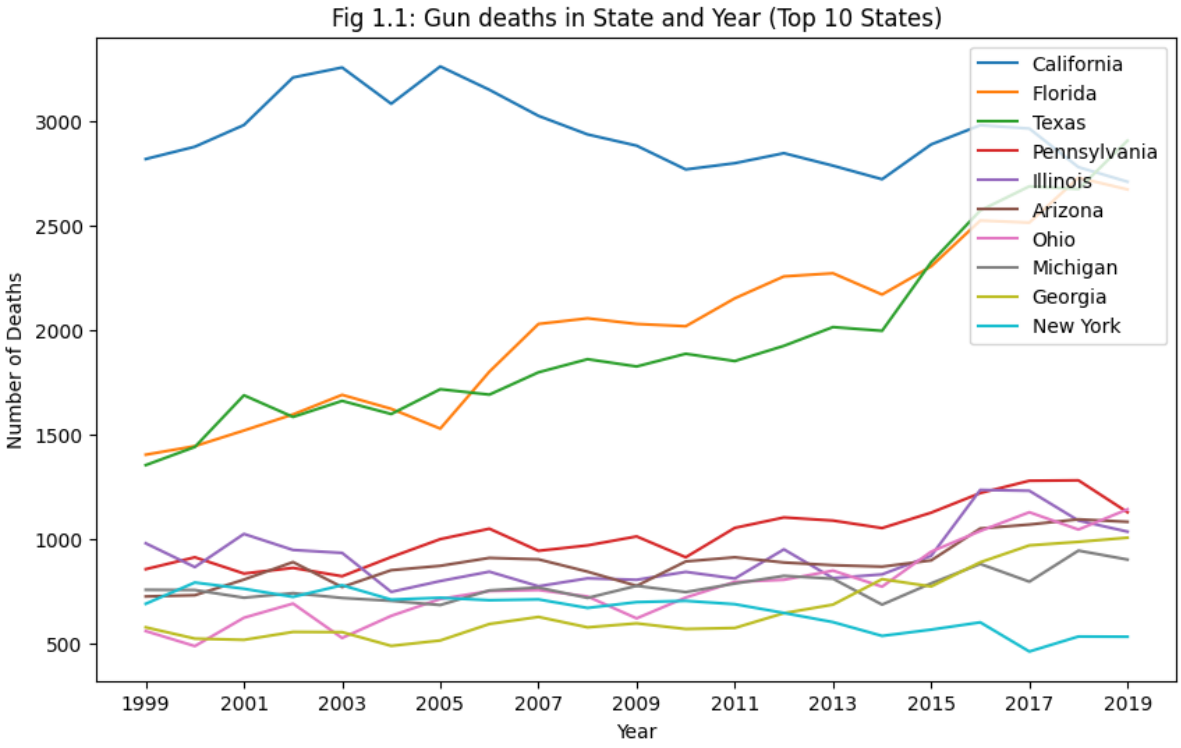
Objective 1

Explanation of Results

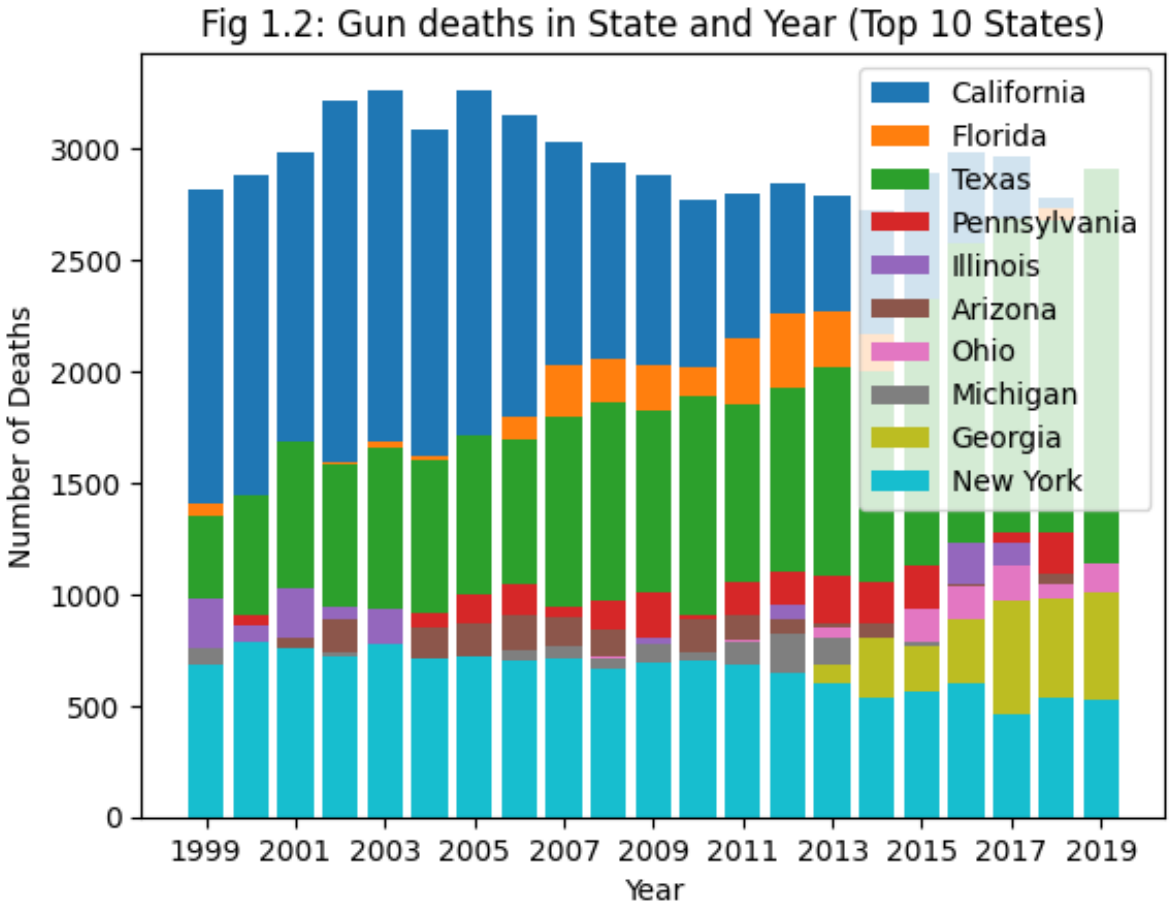
The dataset consists of the gun fatalities that have occurred in the USA throughout the years (1999 to 2019). The basis of this objective was to find out the top ten states that were most affected by gun attacks in the country throughout the time frame 1999-2019. In the general data analysis, a line graph lets us visualise and understand how the overall impact of gun deaths throughout the years affected the country, with which we can assess how the fatalities in those 50 states varied from one another. We examined the fatalities in these ten states using these figures to see how they differ from one another. Both the bar and line graphs provide us with a unique perspective on how the fatalities fluctuate in each state in each year. The bar graph displays all the data for each state in bars corresponding to each specific year, while the line graph shows the fatalities that occurred in each state individually. From these figures, it's very evident that California has suffered the most deaths compared to any other state throughout the given timeframe. These figures make it very clear that, during the course of the specified period, California has had the highest number of fatalities relative to all other states. Given that the dataset breaks down gun deaths in the USA by county, we can also see which county in California experienced the most impact over the years, due to it having the highest number of fatalities. Figure 1.3 makes it clear that Los Angeles county had the greatest number of fatalities and that it was far larger than the other counties, with LA having about 20,000 death cases compared to San Diego (which came in second) having only about 5,000 instances. These figures make it very clear that, during the course of the specified period, California has had the highest number of fatalities relative to all other states. By altering the state name as input in the cali_county_death() function, we can analyse the county-level death rate of other states in a manner similar to this.

Visualisation

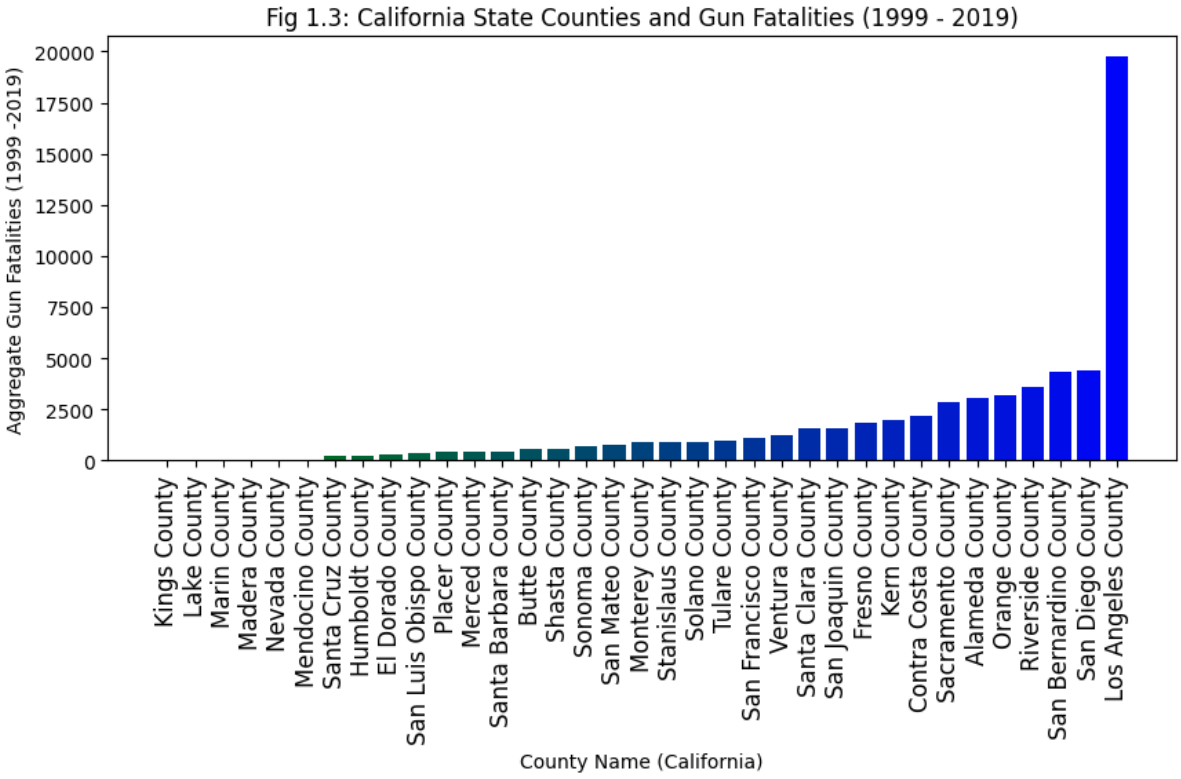
In [97]:
```
top10states_line()
```



In [98]:
```
top10states_bar()
```



In [99]:
```
cali_county_death()
```



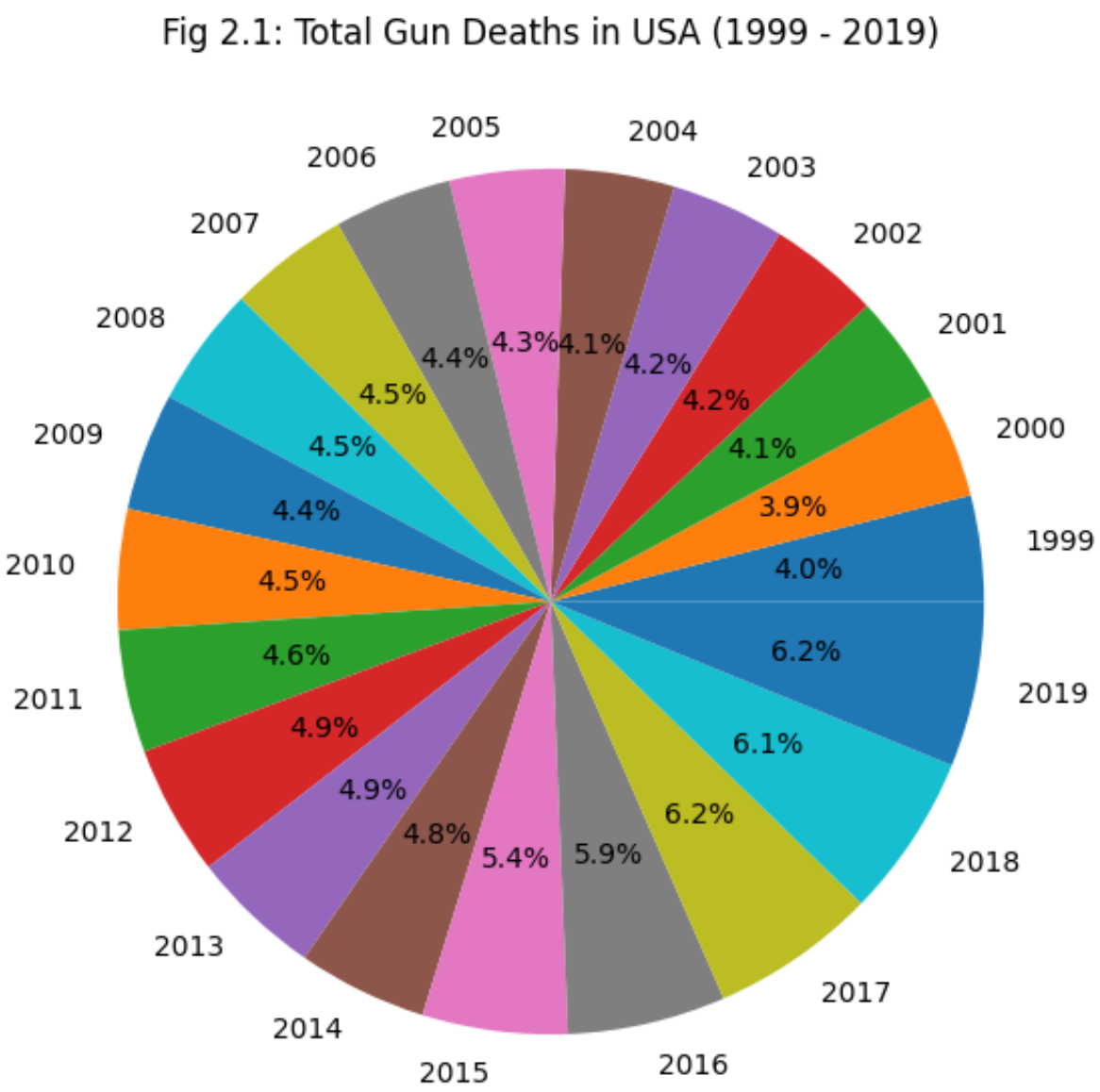
Objective 2

Explanation of Results

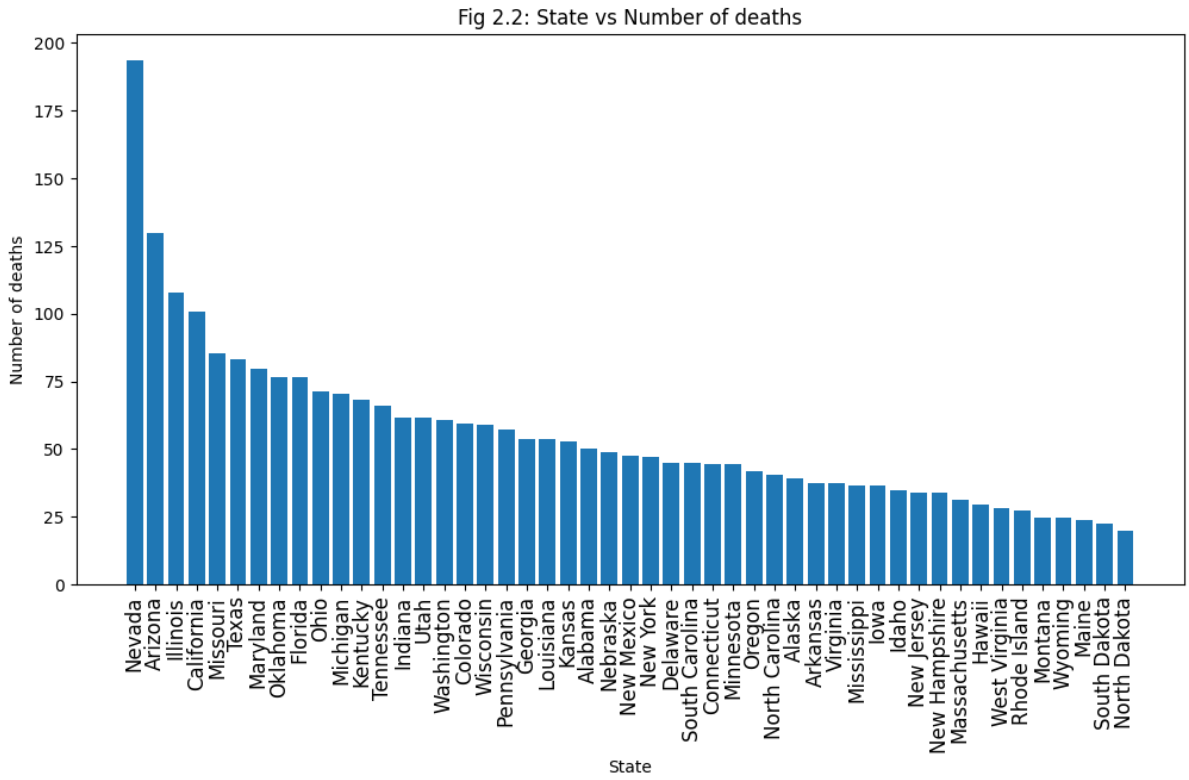
The total number of deaths according to year is displayed in a dictionary. To confirm the same distribution as that depicted in Fig. 2.1, a pie chart was created. The pie chart makes it clear that the years 2019, 2018, 2017, 2016, and 2015 had the highest death rates, with a percentage distribution of 6.2, 6.1, 6.2, 5.9 and 5.4 percent respectively. We determined the lowest and highest death states, as shown in Fig. 2.2, by comparing the trend of State vs. Mean of Deaths for each State. The graph shows that Vermont has the fewest deaths while Arizona has the most. Another plot based on crude rate was created to confirm the outcome for each state as shown in Fig 2.3. From the bar plot, we can see that Wyoming has the highest crude rate which is contrary to the number of deaths plotted in Fig 2.2. This is because the crude rate is a measure of deaths per 100000 population and hence the number of deaths in a state alone cannot determine whether the region is dangerous or not hence crude rate is the best measure to analyse the deaths in several regions. Vermont has the lowest crude rate (Fig 2.3) and has one of the lowest number of deaths (Fig 2.2). In addition to that Arizona had the highest number of deaths as shown in Fig 2.2 but according to the crude rate plot (Fig 2.3), it is only among one of the highest. Another analysis was done on Counties taking the same five-year time frame. Two plots were done, one plotting the counties having the highest crude rates (Fig 2.4) and another having the lowest crude rates (Fig 2.5). Both the line plots show that Petersburg city has the highest crude rate (Fig 2.4) and New York County has the lowest crude rate based on the crude rate analysis for each county. Hence to conclude the objective prioritising crude rate as the measure of gun deaths in each state and county, Wyoming is the most notorious state for gun death violence while Vermont is the least notorious. Analysing the county crude rate across the US we got Petersburg city county having the highest crude rate and New York county having the lowest.

Visualisation

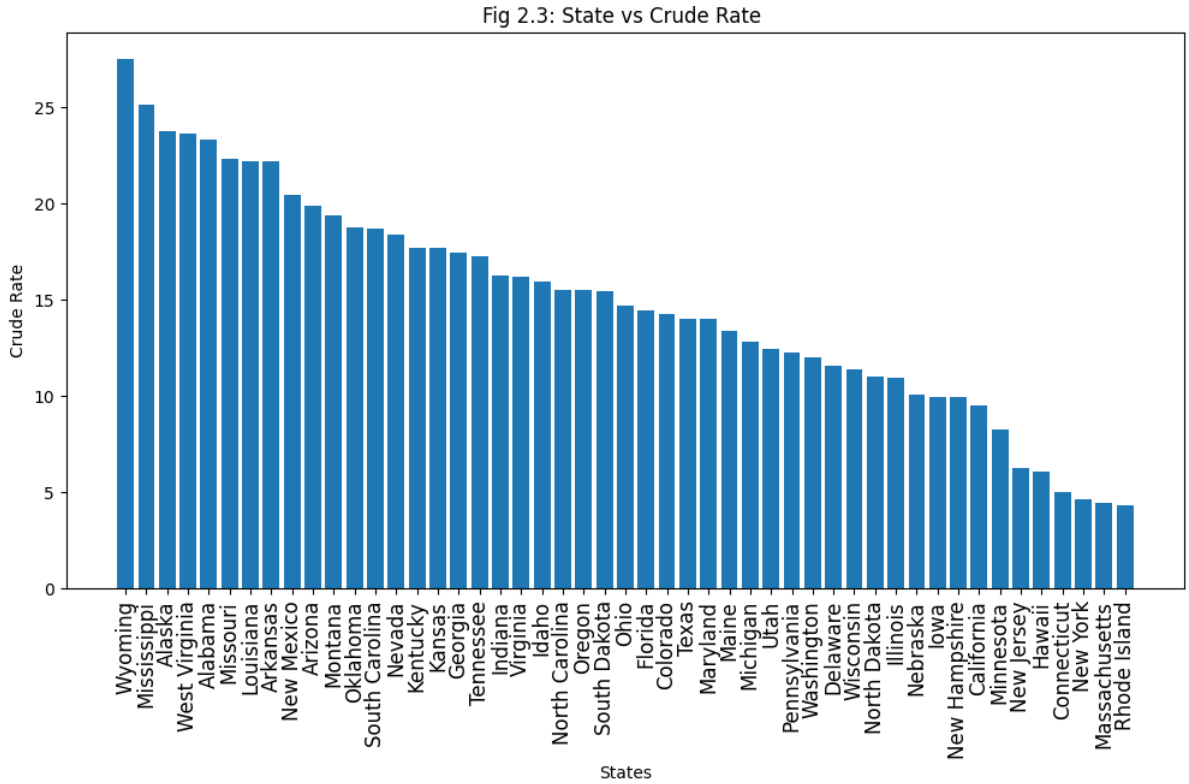
In [100]:
```
year_vs_death_agg()
```



In [101]:
```
state_vs_death_agg()
```

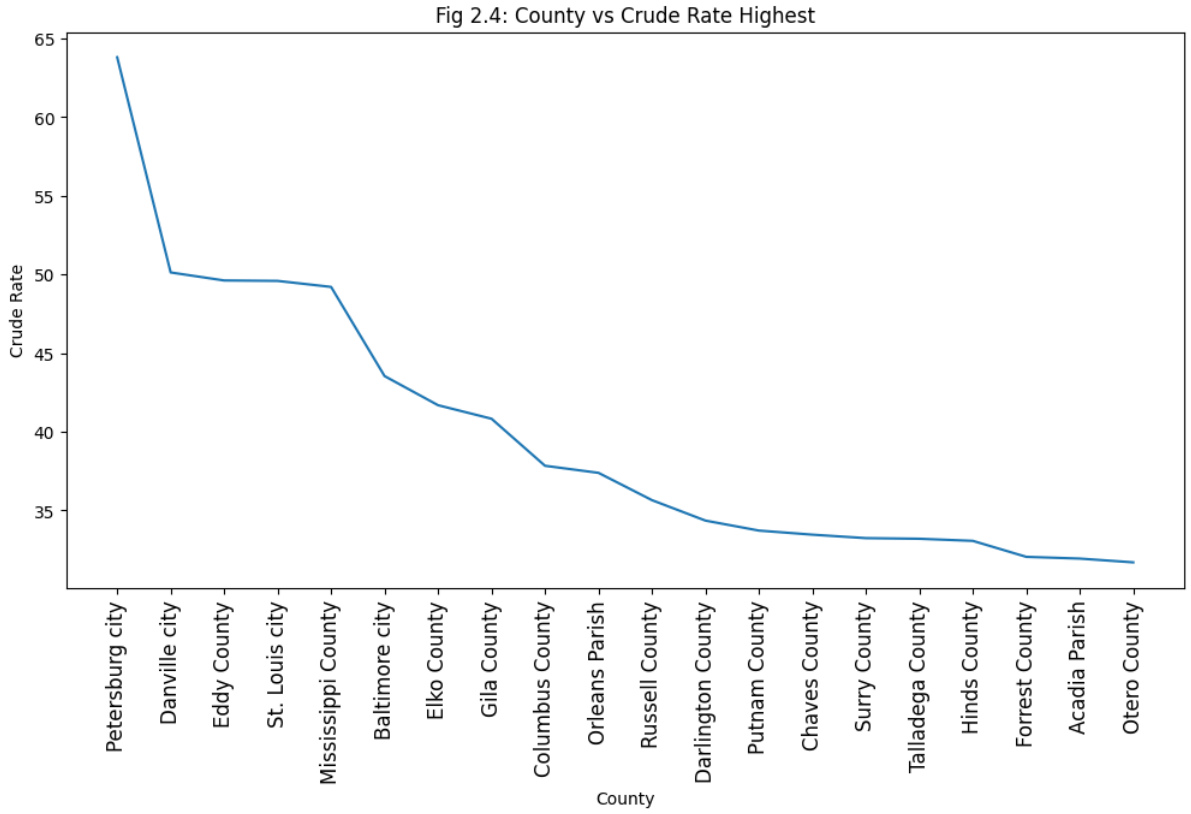


In [102]:
```
cruderate_vs_state()
```



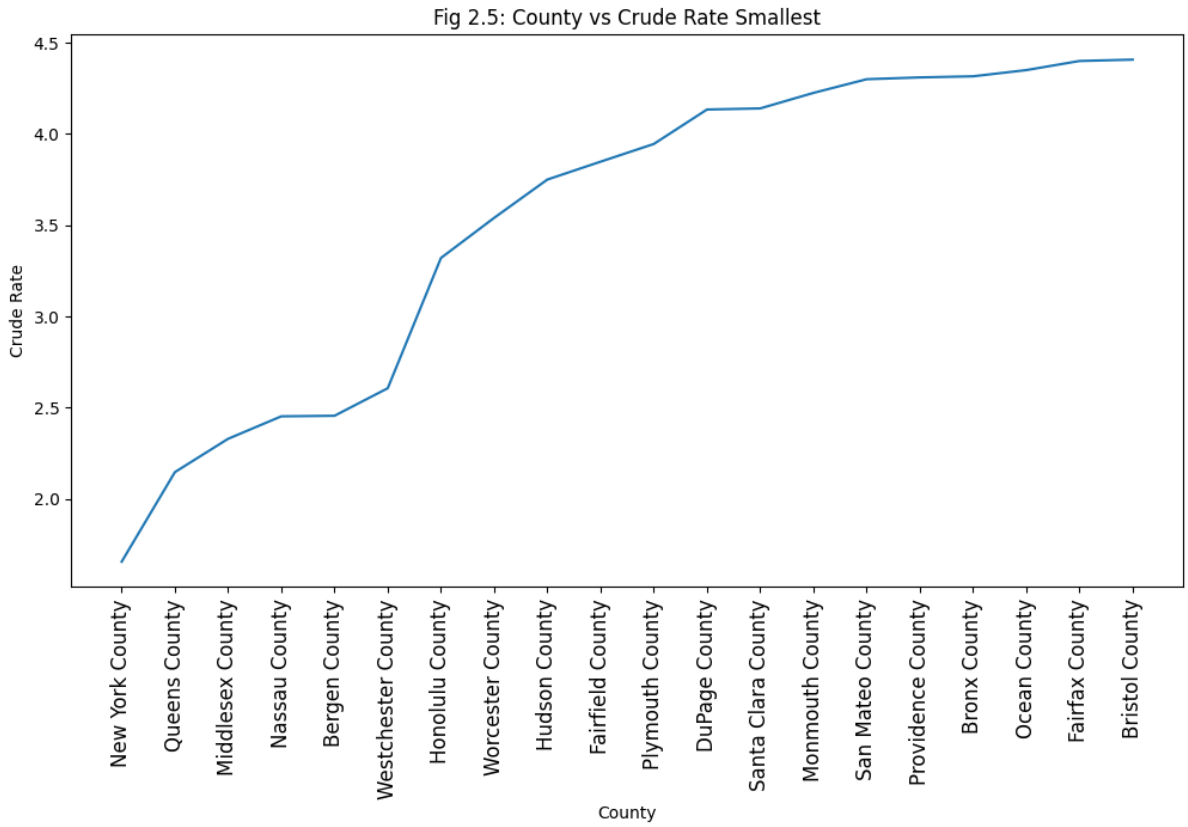
- The following is a line plot to show the counties having highest crude rate for the year 2015 to 2019

In [103]:
```
county_vs_cruderate_largest()
```



- The following is a line plot to show the counties having lowest crude rate for the year 2015 to 2019

In [104]:
```
county_vs_cruderate_smallest()
```



Objective 3

Explanation of Results

The dataset is arranged by county. To identify the safest and riskiest states to live in the United States, the dataset is grouped by state names. Additionally, if we analyse based on the number of deaths, the results are completely different.

Fig 3.1 depicts the crude rate in several US states. The graph makes it clear that Wyoming is the most dangerous state to live in. It has the highest crude rate of 27.2%. Hawaii, on the other hand, is the safest state to live in. Its crude rate is the lowest at 3.31%. As seen in fig 3.2, in the list of counties from the riskiest state, Wyoming, Natrona County has the highest crude rate. It has a crude rate of 30.94%. Laramie county on the other hand has the least crude rate in this list with a crude rate of 20.34%.

As observed in fig 3.3, in the list of counties from the safest state, Hawaii, Hawaii County has the highest crude rate. It has a crude rate of 11.51%. Honolulu county, on the other hand, has the least crude rate in this list with a crude rate of 3.04%.

Moreover, the county with the highest crude rate in Hawaii is safer than the county with the least crude rate in Wyoming.

Visualisation

In [105]:
```
gr()
```

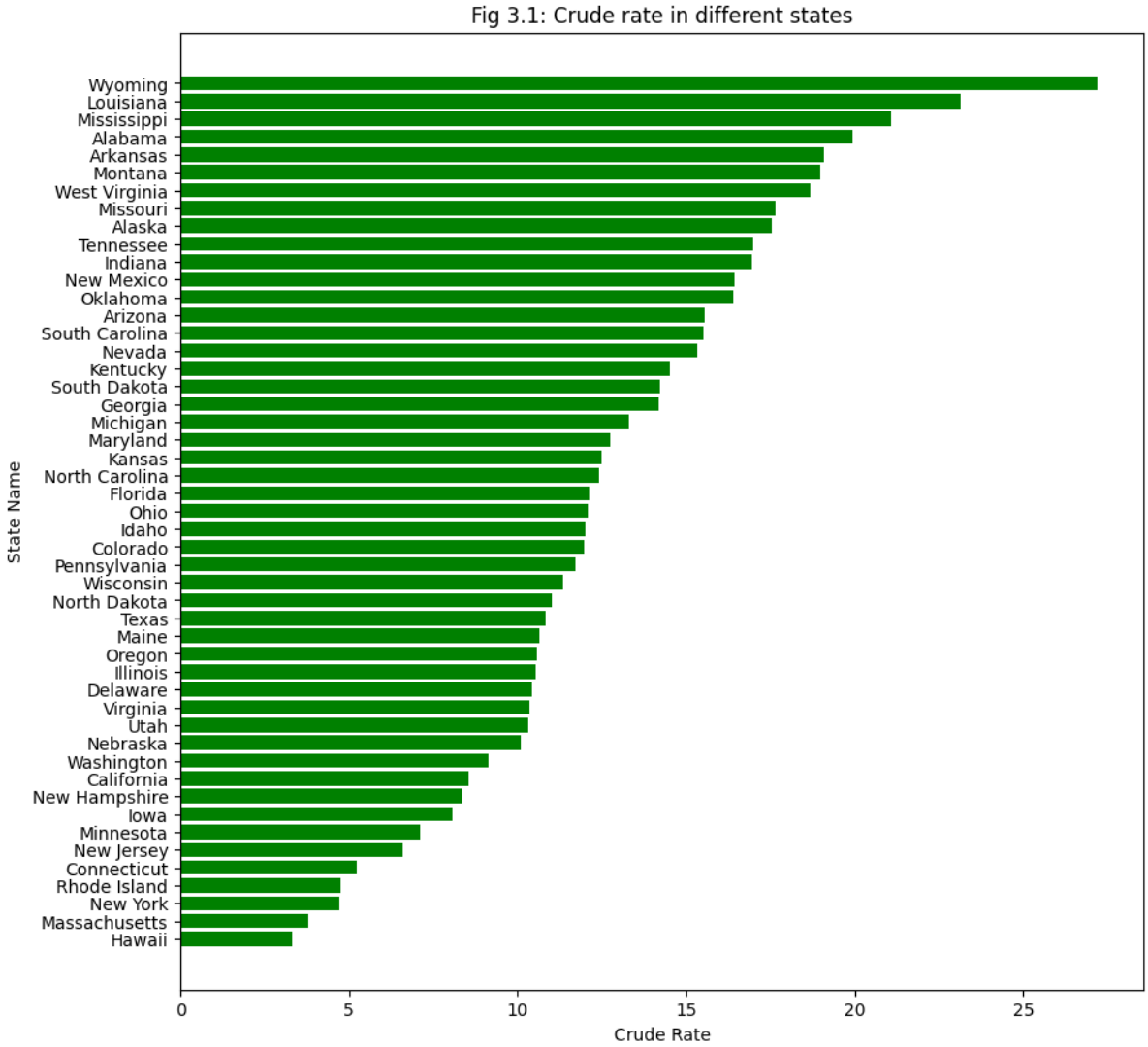


Fig 3.1: Crude rate in different states
State Name
Wyoming
Louisiana
Mississippi
Alabama
Arkansas
Montana
West Virginia
Missouri
Alaska
Tennessee
Indiana
New Mexico
Oklahoma
Arizona
South Carolina
Nevada
Kentucky
South Dakota
Georgia
Michigan
Maryland
Kansas
North Carolina
Florida
Ohio
Idaho
Colorado
Pennsylvania
Wisconsin
North Dakota
Texas
Maine
Oregon
Illinois
Delaware
Virginia
Utah
Nebraska
Washington
California
New Hampshire
Iowa
Minnesota
New Jersey
Connecticut
Rhode Island
New York
Massachusetts
Hawaii
0
5
10
15
20
25
Crude Rate

In [106]:
```
gr2()
```

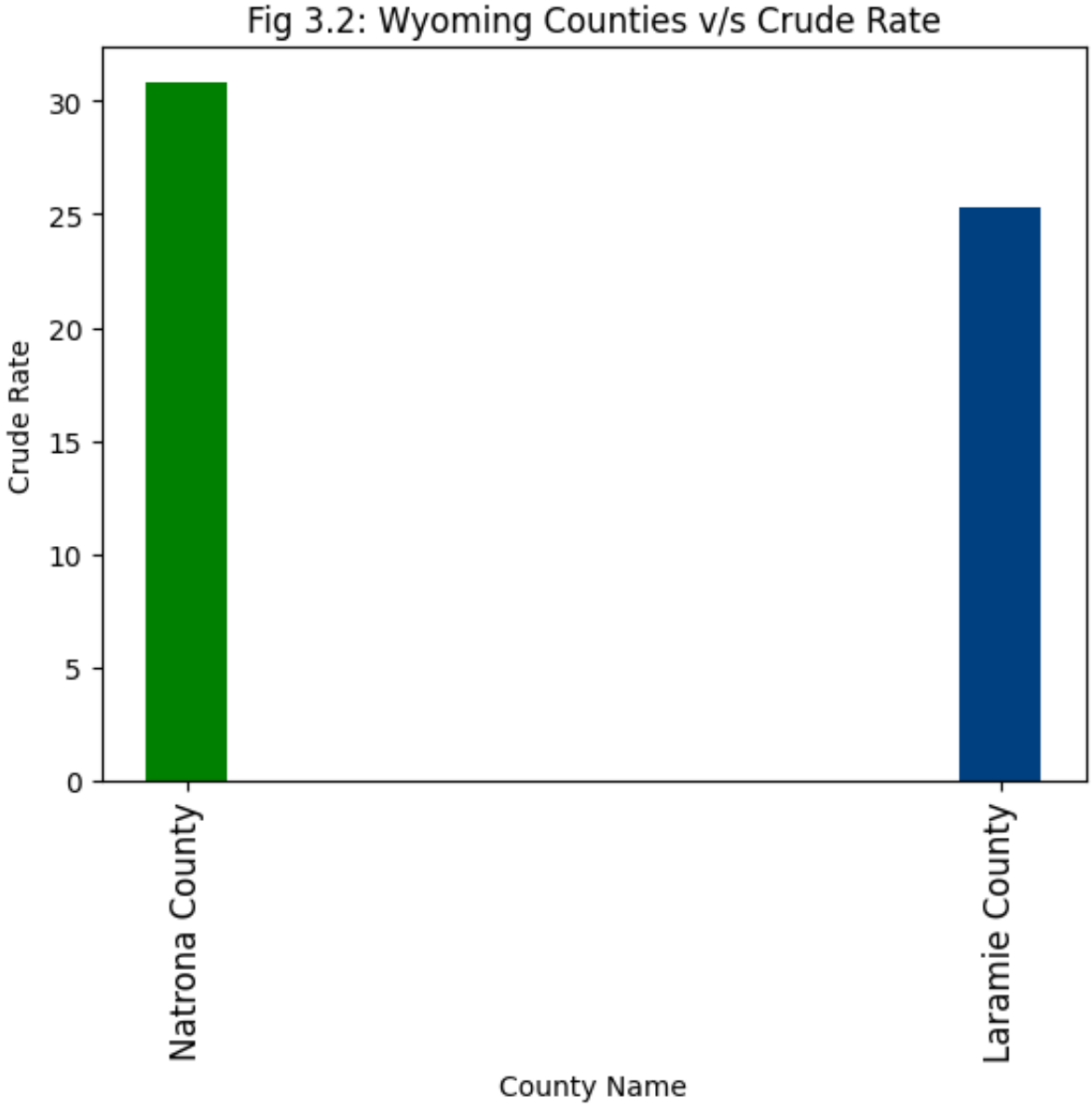


Fig 3.2: Wyoming Counties v/s Crude Rate
Crude Rate
0
5
10
15
20
25
30
Natrona County
Laramie County
County Name

In [107]:
```
gr3()
```

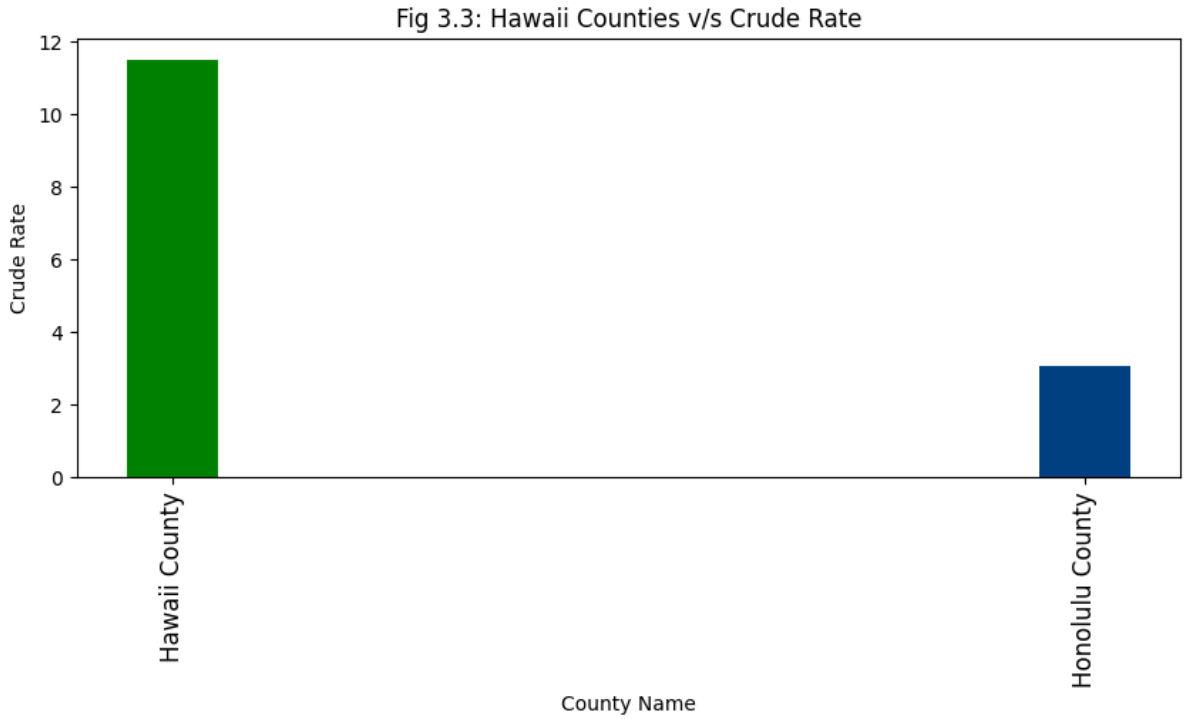


Fig 3.3: Hawaii Counties v/s Crude Rate
Crude Rate
0
2
4
6
8
10
12
Hawaii County
Honolulu County
County Name

Objective 4

Explanation of Results

Find the average age of the population in the top 10 states with the highest crude rates. Also analyse the correlation between population and age adjusted rate

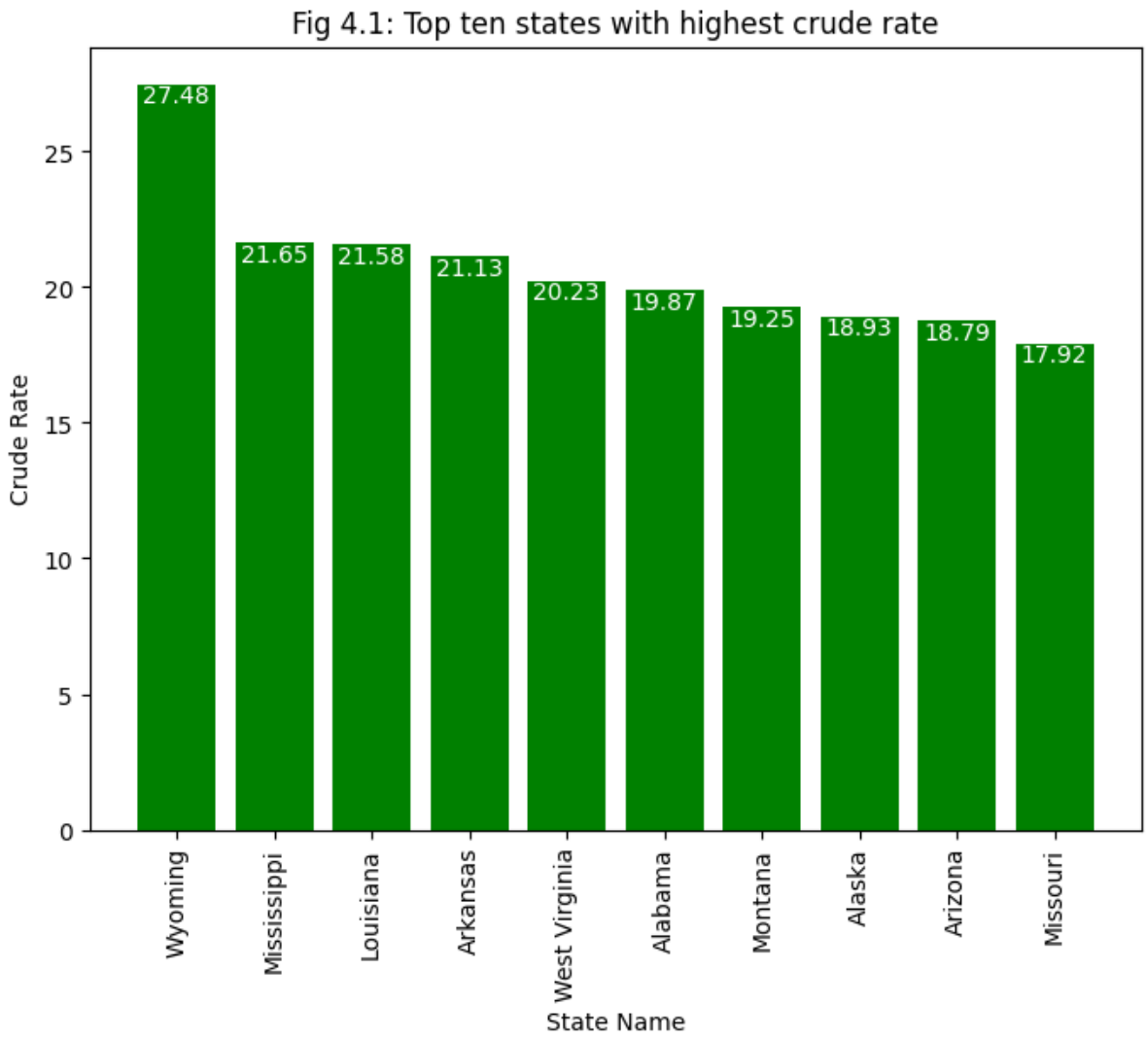
The primary goal was to determine the average age of the population in the top ten states with the highest crude rates. Wyoming has the highest crude rate, followed by Mississippi, Louisiana, Arkansas, West Virginia, Alabama, Montana, Alaska, Arizona, and Missouri, according to our analysis. The crude rate is the number of deaths per 100,000 people. Wyoming's crude rate is 27.48, which means that from 1999 to 2019, nearly 27 people died as a result of firearm discharge in Wyoming alone. The results are depicted in Fig 4.1.

Using the above-mentioned states as a guide, we calculated the average age of the population killed. By analysing Fig 4.2, it has been found that people in their late twenties are the most likely to be killed by a gunshot in Wyoming. People of or under the age of 18 are the most commonly killed in three out of ten states. Three of ten states have an average age of 21. There are six states where people are killed who are under the age of twenty. According to the analysis of these ten states, gunshots disproportionately affect young adults who have the potential to contribute to the overall growth of these states.

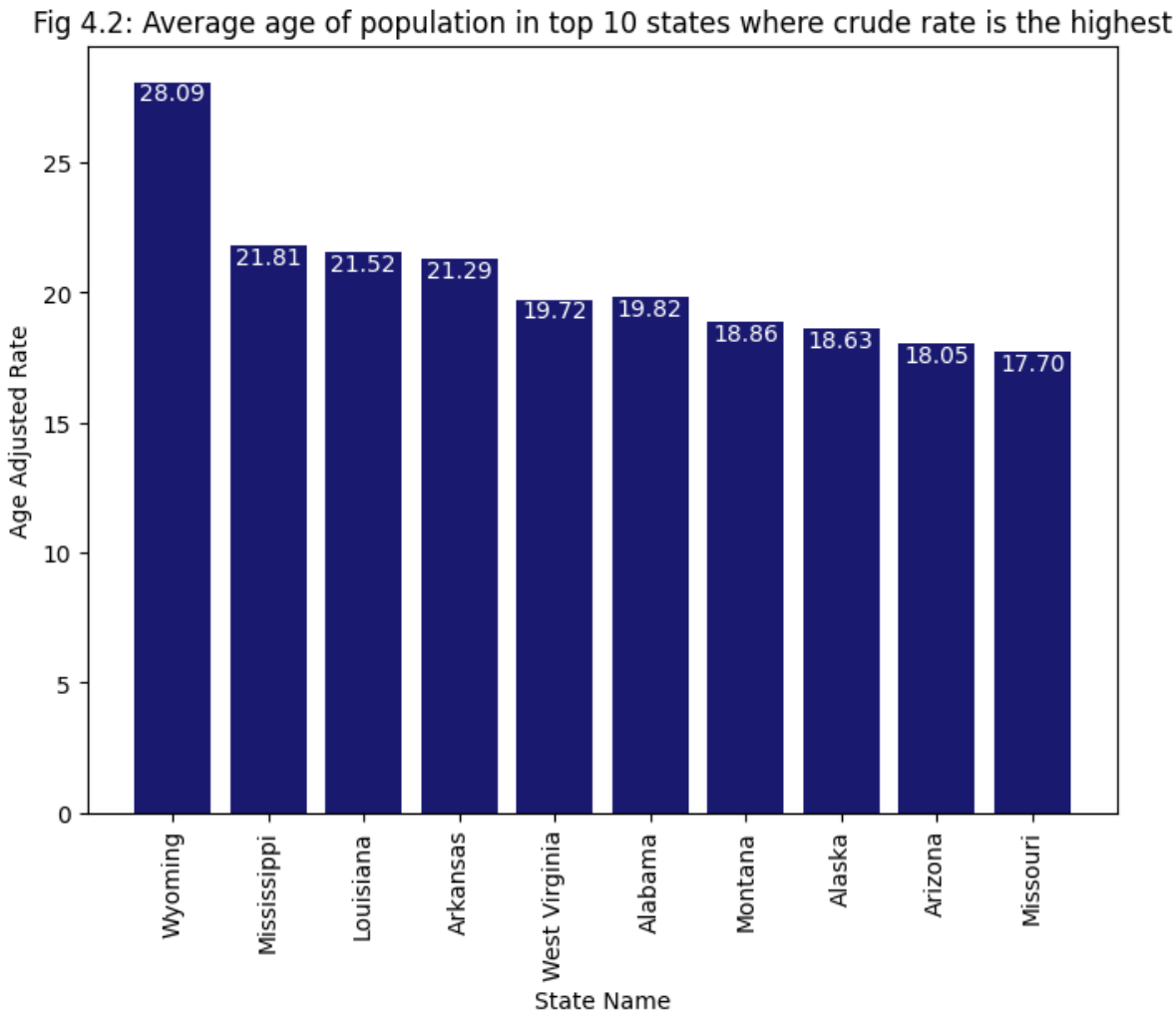
The secondary goal was to determine the relationship between the age adjusted rate and the population. We obtained the spearman's coefficient as -0.53 using the spearman's test. The minus sign indicates a negative correlation, but the coefficient value indicates that there is no correlation between age adjusted rate and population, because the absolute value is less than 0.7. The relationship between population and age adjusted rate, as shown in Fig 4.3, can be discovered using a heatmap, as can the relationship between year, population, crude rate, and age adjusted rate. Furthermore, it has been discovered that the age adjusted rate and the crude rate have a strong correlation.

Visualisation

In [108]:
```
bar_crude()
```



In [109]:
```
bar_age_adjusted_rate()
```



In [110]:
```
correlation_map()
```

Fig 4.3: Correlation between Age Adjusted Rate, Crude Rate, Year and Population

Conclusion (5 marks)

Achievements

The primary goal of this project was to analyse the gun deaths that occurred in United States during a 20-year period. Based on the data, it is evident that from 1999 to 2019, the number of gunshot deaths in Texas and Florida climbed steadily, while California had the highest number of gunshot deaths over nearly the whole time period. This resulted in Texas having the most number of deaths due to gun shots in 2019. From 2015 to 2019, we noticed that Vermont had the fewest fatalities while Arizona had the most. The number of deaths definitely gives an idea of how risky is to reside in a state. However, analysing the crude rate gives a more accurate result for observing the risk factor. We gave our concentration on crude rate rather than number of deaths for further analysis. It was discovered that Wyoming is the riskiest place to live in while Hawaii is the safest one by assessing the crude rate for each state. Moreover, we recognised that there is a significant relationship between the age-adjusted rate and the crude rate. In Wyoming, the state with highest crude rate, it has been learned that individuals in their late 20s had the highest risk of dying from a gunshot. The review of the top ten states with highest crude rate reveals that young adults are adversely impacted by gunshot wounds.

Limitations

Every study was based on crude rate and number of deaths. Lack of information regarding the victim's gender, the proportion of gun deaths brought on by government action, the motive of the murderer, etc. may have aided in a more comprehensive analysis. Accuracy was limited by a few improper columns with null and junk values. Additionally, there were insufficient data to determine if any of these gun deaths were caused by legal intervention.

Future Work

With the right data inputs, it may be possible to determine if these gunshots were deliberate. We could also determine how many of those were gender-specific targets. Moreover, we could determine if any of these involved legal intervention.