DS\_630 Final Project Report

Data Analysis on Gun Violence

**1. Introduction**

**1.1 Problem Statements :**

The objective of this data analysis project is to represent the comprehensive collection of information related to incidents involving firearms in various contexts. This dataset is a valuable resource for understanding the magnitude, patterns, and dynamics of gun-related incidents, including their locations, impact, circumstances, and other related factors. It is instrumental in the domains of public safety, criminal justice, policy formulation, and advocacy, providing crucial insights into the complex issue of gun violence. The problem statements guiding our analysis are as follows:

* Total number of cases registered per each state?
* Compare and analyze the gun violence incidents in urban (city) and rural (county) areas?
* Determine if there's any correlation between the presence of specific representatives and the frequency or severity of incidents.
* Investigate whether certain demographic groups are more vulnerable or prone to involvement in such incidents.
* Determine if incidents involving stolen guns have distinct characteristics compared to incidents with legally owned guns.
* Analyze whether certain geographical areas are more susceptible to gun violence.
* Determine which age groups are more involved in these cases.
* Determine the total number of individuals killed in each state.
* Determine which gender is primarily involved in this violence.
* Which year had the highest number of cases?

**Methodology**

**Data Collection:**

We obtained a dataset containing relevant information on incidents, gender, state, urban, rural areas, events, age types, density, range, Year.

**Data Preprocessing:**

Data preprocessing involved handling null values, removing repeated values and normalizing numerical features. Categorical variables were encoded, and outliers were addressed to ensure the quality of the dataset for analysis.

**Exploratory Data Analysis:**

Exploratory Data Analysis was performed to gain initial insights into the dataset. Descriptive statistics, visualizations, and correlation analyses were utilized to understand the distribution of variables and relationships within the data.

**Detailed Findings :**

**Gun Violence Incidents in Urban and Rural Areas:**

The first code block analyzes and compares gun violence incidents in urban (city) and rural (county) areas. It creates a bar chart showing the number of incidents in the top 5 cities or counties.

**Correlation with Specific Representatives:**

The second code block attempts to determine if there's any correlation between the presence of specific representatives and the frequency or severity of incidents. It counts the occurrences of different participant statuses and visualizes the results in a bar chart.

**Demographic Groups and Vulnerability:**

The third code block investigates whether certain demographic groups are more vulnerable or prone to involvement in gun violence incidents. It analyzes the characteristics of incidents involving different participant demographics, presenting the results in a pie chart and bar chart.

**Stolen Guns vs. Legally Owned Guns:**

The fourth code block explores whether incidents involving stolen guns have distinct characteristics compared to incidents with legally owned guns. It plots a bar chart showing the total number of incidents involving stolen guns over the years.

**Geographical Areas Susceptibility:**

The fifth code block focuses on analyzing whether certain geographical areas are more susceptible to gun violence. It visualizes the count of participant relationship types in violent events using a bar plot.

**Age Groups Involvement:**

The sixth code block determines which age groups are more involved in gun violence incidents. It creates a bar chart showing the count of incidents for different age groups.

**Total Individuals Killed per State:**

The seventh code block calculates and visualizes the total number of individuals killed in each state. It includes a bar chart displaying the count for each state.

**Gender Involvement:**

The eighth code block analyzes the gender involvement in gun violence incidents. It uses kernel density estimation (KDE) plots to show the distribution of male and female participants over the years.

**Yearly Trends:**

The ninth code block determines which year had the highest number of gun violence cases. It uses a count plot to visualize the yearly trends.

**Comparison with Initial Expectations**

Initial expectations would depend on the context and hypotheses. If you had specific assumptions or hypotheses before the analysis, you should compare the findings with those expectations to validate or revise them.

Code:

# Compare and analyze the gun violence incidents in urban (city) and rural (county) areas?

urban\_vs\_rural = df['city\_or\_county'].value\_counts().head(5)

# Plotting the bar chart

plt.figure(figsize=(8, 5))

urban\_vs\_rural.plot(kind='bar', color=['skyblue', 'lightgreen'])

plt.xlabel('Location Type')

plt.ylabel('Number of Incidents')

plt.title('Gun Violence Incidents in Urban and Rural Areas')

plt.xticks(rotation=0)

plt.tight\_layout()

# Show the plot

plt.show()''

#Determine if there's any correlation between the presence of specific representatives and the frequency or severity of incidents.

participant\_statuses = {}

def participant\_type\_count(age\_type):

if str(age\_type) != 'nan':

ages = age\_type.split("||")

if len(ages) == 1:

ages = ages[0].split("|")

for age in ages:

try:

if age.split('::')[1] in participant\_statuses.keys():

participant\_statuses[age.split('::')[1]] += 1

else:

participant\_statuses[age.split('::')[1]] = 0

except:

if age.split(':')[1] in participant\_statuses.keys():

participant\_statuses[age.split(':')[1]] += 1

else:

participant\_statuses[age.split(':')[1]] = 0

df['participant\_status'].apply(participant\_type\_count);

participant\_statuses = {k: v for k, v in sorted(participant\_statuses.items(), key=lambda item: -item[1])}

age\_df = pd.DataFrame(participant\_statuses.items(), columns = ['participant statuses', 'Count'])

fig = px.bar(age\_df, x='participant statuses', y='Count')

# fig.update\_layout(xaxis\_tickangle=45)

fig.show()

#Investigate whether certain demographic groups are more vulnerable or prone to involvement in such incidents.

from collections import Counter

import re

from plotly.offline import download\_plotlyjs, init\_notebook\_mode, plot, iplot

total\_incidents = []

for i, each\_inc in enumerate(df['incident\_characteristics'].fillna('Not Available')):

split\_vals = [x for x in re.split('\|', each\_inc) if len(x)>0]

total\_incidents.append(split\_vals)

if i == 0:

unique\_incidents = Counter(split\_vals)

else:

for x in split\_vals:

unique\_incidents[x] +=1

unique\_incidents = pd.DataFrame.from\_dict(unique\_incidents, orient='index')

colvals = unique\_incidents[0].sort\_values(ascending=False).index.values

find\_val = lambda searchList, elem: [[i for i, x in enumerate(searchList) if (x == e)][0] for e in elem]

a = np.zeros((df.shape[0], len(colvals)))

for i, incident in enumerate(total\_incidents):

aval = find\_val(colvals, incident)

a[i, np.array(aval)] = 1

incident = pd.DataFrame(a, index=df.index, columns=colvals)

prominent\_incidents = incident.sum()[[4, 5, 6, 9, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20,

21, 23,22,24,45,51]]

fig = {

'data': [

{

'labels': prominent\_incidents.index,

'values': prominent\_incidents,

'type': 'pie',

'hoverinfo':'label+percent+name',

"domain": {"x": [0, .45]},

}

],

'layout': {'title': 'Prominent Incidents of Gun Violence',

'showlegend': False}

}

iplot(fig)

#Determine if incidents involving stolen guns have distinct characteristics compared to incidents with legally owned guns.

df['gun\_stolen'] = df['gun\_stolen'].fillna('Null')

df['gun\_stolen'] = df['gun\_stolen'].str.replace('::',',')

df['gun\_stolen'] = df['gun\_stolen'].str.replace('|',' ')

df['gun\_stolen'] = df['gun\_stolen'].str.replace(',',' ')

df['gun\_stolen']= df['gun\_stolen'].str.replace('\d+', '')

df['Stolenguns']=df['gun\_stolen'].apply(lambda x: x.count('Stolen'))

df['stolenguns']=df['gun\_stolen'].apply(lambda x: x.count('Stolen'))

df['Stolengunstotal'] = df['Stolenguns'] + df['stolenguns']

df\_year\_stolenguns = df[['year','Stolengunstotal']].groupby(['year'], as\_index = False).sum()

df\_year\_stolenguns[['year','Stolengunstotal']].set\_index('year').plot(kind='bar')

# Analyze whether certain geographical areas are more susceptible to gun violence.

relation = df['participant\_relationship']

relation = relation[relation.notnull()]

relation = relation.str.replace("[:|0-9]"," ").str.upper()

relation1 = pd.DataFrame({"count":[len(relation[relation.str.contains("FAMILY")]),

len(relation[relation.str.contains("ROBBERY")]),

len(relation[relation.str.contains("FRIENDS")]),

len(relation[relation.str.contains("AQUAINTANCE")]),

len(relation[relation.str.contains("NEIGHBOR")]),

len(relation[relation.str.contains("INVASION")]),

len(relation[relation.str.contains("CO-WORKER")]),

len(relation[relation.str.contains("GANG")]),

len(relation[relation.str.contains("RANDOM")]),

len(relation[relation.str.contains("MASS SHOOTING")])],

"category":["FAMILY","ROBBERY","FRIENDS","AQUAINTANCE","NEIGHBOR","INVASION","CO-WORKER","GANG","RANDOM","MASS SHOOTING"]})

relation1

plt.figure(figsize=(14,5))

sns.barplot("category","count",data=relation1,palette="prism")

plt.title("COUNT PLOT FOR PARTICPANT RELATION TYPE IN VIOLENT EVENTS")

age\_types = {}

def age\_count(age\_type):

if str(age\_type) != 'nan':

ages = age\_type.split("||")

if len(ages) == 1:

ages = ages[0].split("|")

for age in ages:

try:

if age.split('::')[1] in age\_types.keys():

age\_types[age.split('::')[1]] += 1

else:

age\_types[age.split('::')[1]] = 0

except:

if age.split(':')[1] in age\_types.keys():

age\_types[age.split(':')[1]] += 1

else:

age\_types[age.split(':')[1]] = 0

df['participant\_age\_group'].apply(age\_count)

# Determine which age groups are more involved in these cases.

age\_types = {k: v for k, v in sorted(age\_types.items(), key=lambda item: -item[1])}

age\_df = pd.DataFrame(age\_types.items(), columns = ['Age Type', 'Count'])

fig = px.bar(age\_df, x='Age Type', y='Count')

fig.update\_layout(xaxis\_tickangle=45)

fig.show()

# Determine the total number of individuals killed in each state.

states = {'Alabama': 'AL','Alaska': 'AK','American Samoa': 'AS','Arizona': 'AZ','Arkansas': 'AR','California': 'CA','Colorado': 'CO','Connecticut': 'CT','Delaware': 'DE','District of Columbia': 'DC','Florida': 'FL','Georgia': 'GA','Guam': 'GU','Hawaii': 'HI','Idaho': 'ID','Illinois': 'IL','Indiana': 'IN','Iowa': 'IA','Kansas': 'KS','Kentucky': 'KY','Louisiana': 'LA','Maine': 'ME','Maryland': 'MD','Massachusetts': 'MA','Michigan': 'MI','Minnesota': 'MN','Mississippi': 'MS','Missouri': 'MO','Montana': 'MT','Nebraska': 'NE','Nevada': 'NV','New Hampshire': 'NH','New Jersey': 'NJ','New Mexico': 'NM','New York': 'NY','North Carolina': 'NC','North Dakota': 'ND','Northern Mariana Islands':'MP','Ohio': 'OH','Oklahoma': 'OK','Oregon': 'OR','Pennsylvania': 'PA','Puerto Rico': 'PR','Rhode Island': 'RI','South Carolina': 'SC','South Dakota': 'SD','Tennessee': 'TN','Texas': 'TX','Utah': 'UT','Vermont': 'VT','Virgin Islands': 'VI','Virginia': 'VA','Washington': 'WA','West Virginia': 'WV','Wisconsin': 'WI','Wyoming': 'WY'}

df['state code'] = df['state'].apply(lambda x : states[x])

df.groupby('state code')['n\_killed'].sum().sort\_values(ascending = False).reset\_index()

plt.figure(figsize=(15,10))

sns.countplot(x=df['state code'],data=df)

# Determine which gender is primarily involved in this violence.

df["participant\_gender"] = df["participant\_gender"].fillna("0::Unknown")

def gen(n) :

gen\_rows = []

gen\_row = str(n).split("||")

for i in gen\_row :

g\_row = str(i).split("::")

if len(g\_row) > 1 :

gen\_rows.append(g\_row[1])

return gen\_rows

gen\_series = df.participant\_gender.apply(gen)

df["total\_participant"] = gen\_series.apply(lambda x: len(x))

df["male\_participant"] = gen\_series.apply(lambda i: i.count("Male"))

df["female\_participant"] = gen\_series.apply(lambda i: i.count("Female"))

df["unknown\_participant"] = gen\_series.apply(lambda i: i.count("Unknown"))

genderwise\_total = df[["total\_participant", "male\_participant", "female\_participant", "unknown\_participant"]].groupby(df["year"]).sum()

dp\_gen\_plot=sns.kdeplot(genderwise\_total['male\_participant'], shade=True, color="r",label='male\_participant')

dp\_gen\_plot=sns.kdeplot(genderwise\_total['female\_participant'], shade=True, color="b",label='female\_participant')

plt.legend()

plt.xlabel('Range')

del(genderwise\_total)

# Which year had the highest number of cases?

df['date']=pd.to\_datetime(df['date'])

df['year'] = df['date'].dt.year

df['month']=df['date'].dt.month

plt.figure(figsize=(9,6))

sns.countplot(x='year', data=df)

plt.xlabel('Year')

plt.ylabel('Count')

plt.title('Highest number of cases')

# Show the plot

plt.show()

**Future Work**

Further analysis could explore correlations between different variables (e.g., participant demographics and incident characteristics).

Time series analysis could provide insights into temporal patterns and potential factors influencing changes in gun violence over the years.

**Conclusion**

Gun violence is a complex issue influenced by various factors, including location, demographics, and incident characteristics.

The analysis provides a foundation for understanding patterns and trends, offering valuable insights for policymakers, researchers, and the general public.

Thank you,

Yaswanth Akkisetty

Rajashekar Reddy Dinasarapu