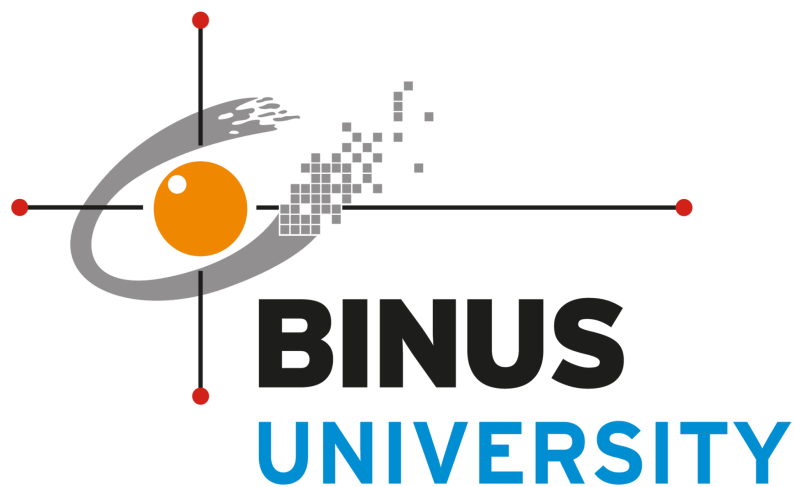


Mass Shootings in United States (2018-2022) Analytics

A Project Report of
Big Data Processing Course



Submitted by:

2440047173 - Alexander Ryu Fenando
2440004395 - Even Owen Thamrin
2440003770 - Kent Samuel Claudio
2440030456 - Vetri Marvel Budiman
2440018141 - Yulius Obi

COMPUTER SCIENCE MAJOR
BINA NUSANTARA UNIVERSITY

2022

1. Latar belakang masalah

Penembakan massal adalah insiden di mana empat orang atau lebih (tidak termasuk si penembak) tertembak dan terbunuh. Jumlah penembakan massal yang melanda negara Amerika Serikat terlalu tinggi. Dari tahun 2021 saja sudah terdapat 692 jumlah kasus penembakan, yang menimbulkan 3519 korban yang terluka maupun korban jiwa.

1 dari 4 korban penembakan massal adalah anak-anak atau remaja. Adegan mengerikan penembakan massal telah menghantui masyarakat Amerika Serikat. Banyaknya regulasi mengenai kepemilikan senjata api di setiap negara bagian belum bisa menghapus insiden penembakan massal. Oleh karena itu, kami membuat analisis penembakan massal di setiap negara bagian Amerika Serikat untuk memberikan gambaran mengenai jumlah kasus penembakan massal yang ada di setiap negara bagian Amerika Serikat, serta meningkatkan kesadaran akan bahayanya kepemilikan senjata api dan dampak buruknya.

2. Tujuan

- Menampilkan jumlah kasus dan jumlah korban pada insiden penembakan massal di setiap negara bagian di Amerika Serikat dalam rentang 2018 - 2022.
- Mencari tahu dan menampilkan hari, bulan, tahun, musim dan public holiday yang paling sering terjadi insiden penembakan massal dalam rentang 2018 - 2022.
- Menampilkan dan menganalisa apa yang bisa didapat dari deskripsi terjadinya penembakan massal.
- Menampilkan informasi terkait keketatan regulasi mengenai penggunaan senjata api di Amerika Serikat berdasarkan banyaknya regulasi di setiap negara bagian pada tahun 2017.

3. Manfaat

- Meningkatkan kesadaran akan bahayanya kepemilikan senjata api dan dampak buruknya.
- Mengetahui ketat dan longgarnya aturan mengenai kepemilikan dan penggunaan senjata api di setiap negara bagian.

4. Metodologi

4.1. Dataset 1

Kami menggunakan dataset yang berasal dari website Kaggle yang berjudul “Mass Shooting in United States (2018 - 2022).” Dataset ini mengandung dokumentasi kasus penembakan massal di setiap negara bagian Amerika Serikat dari tahun 2018 - 2022.

1. Contain

Dataset ini memiliki 5 file dengan format csv yang dibedakan dengan tahun dari daftar penembakan massal yaitu dari tahun 2018 sampai tahun 2022

2. Columns

Setiap file memiliki jumlah kolom yang sama yaitu 6 kolom. Kolom tersebut antara lain:

- a. Date: berisi tanggal terjadinya penembakan massal
 - i. Tipe data: date
- b. State: berisi dimana penembakkan massal terjadi
 - i. Tipe data: string
- c. Dead: jumlah korban meninggal dari penembakkan massal
 - i. Tipe data: integer
- d. Injured: jumlah korban terluka dari penembakkan massal
 - i. Tipe data: integer
- e. Total: total dari jumlah korban terluka dan meninggal
 - i. Tipe data: integer
- f. Description: deskripsi atau laporan singkat mengenai detail bagaimana terjadinya penembakan tersebut
 - i. Tipe data: string

Sumber Dataset:

https://www.kaggle.com/datasets/hemil26/mass-shootings-in-united-states-20182022?select=shootings_2022.csv

4.2. Dataset 2

Kami juga menggunakan dataset yang berasal dari website Kaggle yang berjudul “Firearms Provisions in US States”. Dataset ini mengandung jumlah regulasi tentang kepemilikan senjata api di setiap negara bagian Amerika Serikat.

1. Contain

Dataset ini memiliki 2 file dengan format csv dan xls yang dibagi menjadi negara yang memiliki regulasi bersenjata dari daftar regulasi yang berlaku dari tahun 1991 sampai tahun 2017. Kami menggunakan file raw_data.csv yang tersedia.

2. Columns

Dalam file 'raw_data.csv' memiliki jumlah kolom sebanyak 136 kolom. Namun, kami hanya memakai kolom yang kami perlukan yaitu 'state', 'year', dan 'lawtotal'. Deskripsi dari kolom-kolom tersebut adalah sebagai berikut;

- a. state: berisi dimana penembakkan massal terjadi
 - i. Tipe data: String (nama negara bagian)
- b. year: berisi tahun regulasi tersebut dikemukakan.
 - i. Tipe data: integer
- c. lawtotal: berisi jumlah regulasi yang berlaku mengenai senjata api
 - i. Tipe data: number

Sumber dataset:

<https://www.kaggle.com/datasets/jboysen/state-firearms>

4.3. Tools

Kami menggunakan Google Colab dan bahasa pemrograman Python untuk melakukan analisis

5. Processing data

5.1. Import library yang diperlukan

- A. numpy
- B. pandas
- C. matplotlib.pyplot
- D. plotly.express
- E. wordcloud
- F. collections

```
import library

library yang digunakan

[4] !pip install wordcloud

Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: wordcloud in /usr/local/lib/python3.7/dist-packages (1.5.0)
Requirement already satisfied: numpy>=1.6.1 in /usr/local/lib/python3.7/dist-packages (from wordcloud) (1.21.6)
Requirement already satisfied: pillow in /usr/local/lib/python3.7/dist-packages (from wordcloud) (7.1.2)

[5] import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt # making plots from data
import plotly.express as px # graphing interactive map from data

from wordcloud import WordCloud, STOPWORDS
import collections
```

Gambar 1

- 5.2. Melakukan download dataset dari Kaggle ke local komputer
- 5.3. Memuat dan membaca dataset ke Google colab

```
Load dataset

Read dan import dataset in csv files

[6] df2018 = pd.read_csv('shootings_2018.csv')
df2019 = pd.read_csv('shootings_2019.csv')
df2020 = pd.read_csv('shootings_2020.csv')
df2021 = pd.read_csv('shootings_2021.csv')
df2022 = pd.read_csv('shootings_2022.csv')
merge_dfyear = pd.concat([df2018, df2019, df2020, df2021, df2022])

merge_dfyear.reset_index(drop=True, inplace=True)

merge_dfyear.tail()
```

	Date	State	Dead	Injured	Total	Description
2351	01/01/2022	Georgia	1	3	4	After officers were dispatched to respond to a...
2352	01/01/2022	Wisconsin	1	3	4	A man was killed, and three others wounded, in...
2353	01/01/2022	Indiana	0	4	4	Four people were wounded at a shooting at a Ne...
2354	01/01/2022	Colorado	2	2	4	Two adults were killed, and two wounded, in an...
2355	01/01/2022	Missouri	0	4	4	Four adults were wounded in the early morning ...

Gambar 2

- 5.4. Melakukan data cleaning
 - A. Melakukan reset index pada dataframe
 - B. Mengecek dan mencari apakah terdapat data yang memiliki value null

- C. Mencari keakuratan dari kolom Total yang didapat dari jumlah kolom Dead dengan kolom Injured. Setelah menemukan data yang tidak sesuai atau tidak akurat, kami menghapus data tersebut.

```
▼ Data Cleaning

Cek dan Find null value in dataframe

[7] merge_dfyear.isnull().sum()

Date          0
State          0
Dead           0
Injured        0
Total          0
Description    0
dtype: int64
```

Gambar 3

```
mencari total dari jumlah injured dan jumlah kill yang tidak sesuai dengan jumlah di kolom total

[8] merge_dfyear.shape

error_kolom = merge_dfyear[(merge_dfyear['Dead'] + merge_dfyear['Injured'] ) != merge_dfyear['Total']]
print(error_kolom)

merge_dfyear = merge_dfyear[~((merge_dfyear['Dead'] + merge_dfyear['Injured'] ) != merge_dfyear['Total'])]

[9] merge_dfyear.shape

error_kolom2 = merge_dfyear[(merge_dfyear['Dead'] + merge_dfyear['Injured'] ) != merge_dfyear['Total']]
# print(error_kolom2)
```

Gambar 4

5.5. Menampilkan tipe data dan informasi kolom dataset

▼ Detail dan Basic Statistics

Mengetahui nama dan tipe data kolom, describe detail

```
[10] print(merge_dfyear.info())
      print('\nTipe data kolomn')
      print(merge_dfyear.dtypes)
      print('\n=====\\n')
      print(merge_dfyear.columns)
      print('\nBasic Statistics')
      merge_dfyear.describe()
```

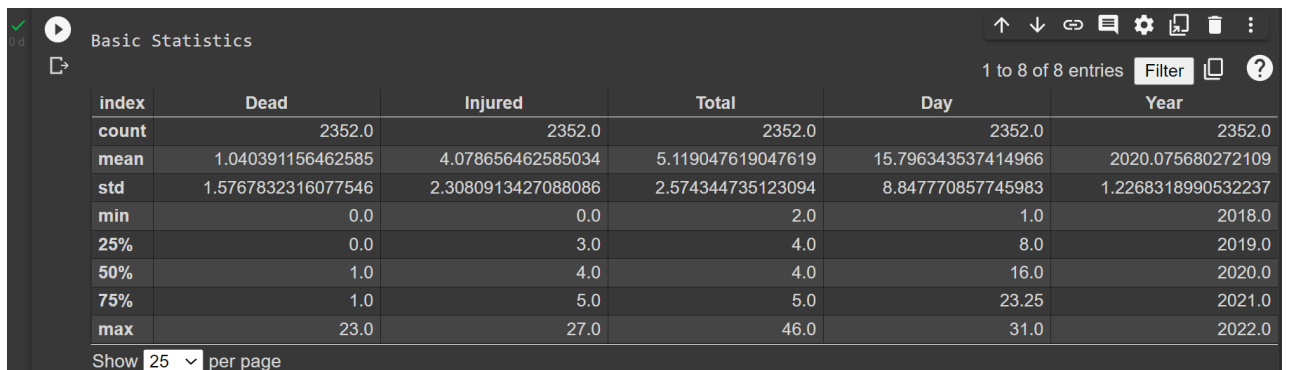
Gambar 5

```
> <class 'pandas.core.frame.DataFrame'>
Int64Index: 2352 entries, 0 to 2355
Data columns (total 6 columns):
#   Column          Non-Null Count  Dtype
---  -
0   Date             2352 non-null   object
1   State            2352 non-null   object
2   Dead             2352 non-null   int64
3   Injured          2352 non-null   int64
4   Total            2352 non-null   int64
5   Description      2352 non-null   object
dtypes: int64(3), object(3)
memory usage: 128.6+ KB
None

Tipe data kolomn
Date            object
State           object
Dead            int64
Injured         int64
Total           int64
Description     object
dtype: object
```

Gambar 6

5.6. Menampilkan data basic statistics seperti count, mean, dan lainnya



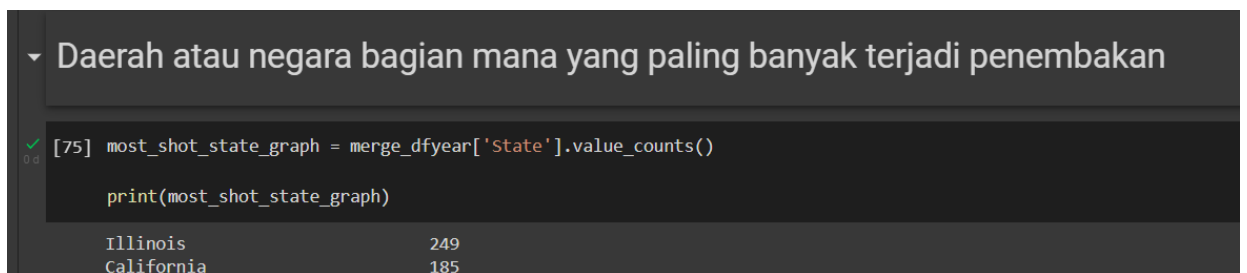
index	Dead	Injured	Total	Day	Year
count	2352.0	2352.0	2352.0	2352.0	2352.0
mean	1.040391156462585	4.078656462585034	5.119047619047619	15.796343537414966	2020.075680272109
std	1.5767832316077546	2.3080913427088086	2.574344735123094	8.847770857745983	1.2268318990532237
min	0.0	0.0	2.0	1.0	2018.0
25%	0.0	3.0	4.0	8.0	2019.0
50%	1.0	4.0	4.0	16.0	2020.0
75%	1.0	5.0	5.0	23.25	2021.0
max	23.0	27.0	46.0	31.0	2022.0

Show 25 per page

Gambar 7

5.7. Selanjutnya, kami melakukan analisis dengan 4 poin utama yaitu

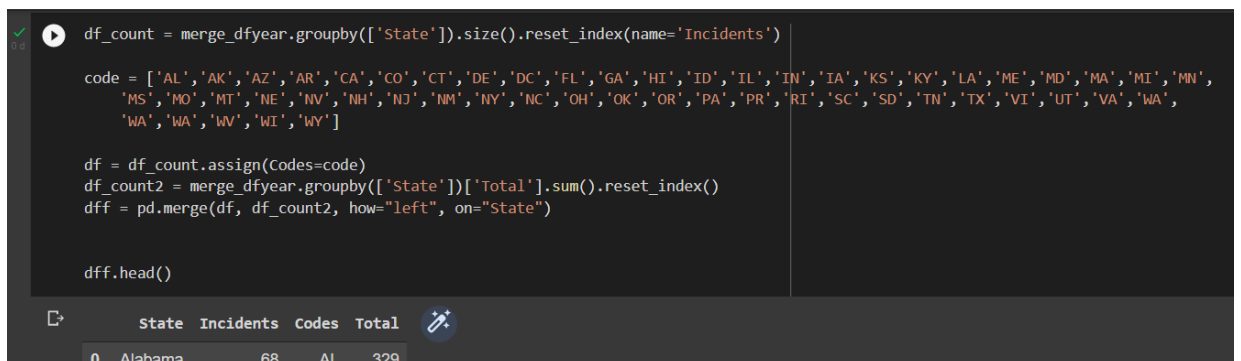
A. Menganalisis jumlah kasus penembakan dan jumlah korban penembakan di setiap negara bagian



```
[75] most_shot_state_graph = merge_dfyear['State'].value_counts()
      print(most_shot_state_graph)
```

Illinois	249
California	185

Gambar 8



```
df_count = merge_dfyear.groupby(['State']).size().reset_index(name='Incidents')
code = ['AL','AK','AZ','AR','CA','CO','CT','DE','DC','FL','GA','HI','ID','IL','IN','IA','KS','KY','LA','ME','MD','MA','MI','MN',
        'MS','MO','MT','NE','NV','NH','NJ','NM','NY','NC','OH','OK','OR','PA','PR','RI','SC','SD','TN','TX','VI','UT','VA','WA',
        'WA','WA','WV','WI','WY']
df = df_count.assign(Codes=code)
df_count2 = merge_dfyear.groupby(['State'])['Total'].sum().reset_index()
dff = pd.merge(df, df_count2, how="left", on="State")
dff.head()
```

	State	Incidents	Codes	Total
0	Alabama	68	AL	329

Gambar 9

B. Menganalisis kolom 'Date'

- Mengetahui hari yang paling sering terjadi penembakan

➤ Sering terjadinya penembakan di tanggal apa

Berdasarkan Hari

```
[82] merge_dfyar['DateConvert'] = pd.to_datetime(merge_dfyar['Date'])
      merge_dfyar['DayOfWeek'] = merge_dfyar['DateConvert'].dt.day_name()

merge_dfyar['Date'] = pd.to_datetime(merge_dfyar['Date'])
dftime = merge_dfyar['Date'].iloc[0]
merge_dfyar['Day'] = merge_dfyar['Date'].apply(lambda time: time.day)
merge_dfyar['Month'] = merge_dfyar['Date'].apply(lambda time: time.month)
merge_dfyar['Year'] = merge_dfyar['Date'].apply(lambda time: time.year)

print(merge_dfyar['DayOfWeek'].value_counts().sort_index(axis=0))

# print("\n DAY DAY DAY")
print(merge_dfyar['Month'])
```

Gambar 10

- Mengetahui bulan yang paling sering terjadi penembakan

Berdasarkan Bulan

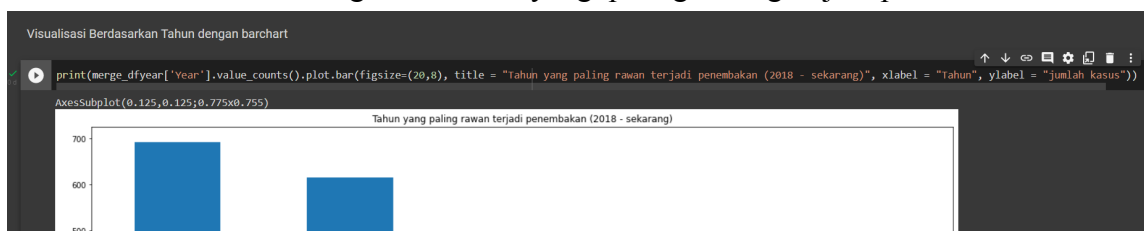
```
[84] dmap={1:'Jan',2:'Feb',3:'Mar',4:'Apr',5:'Mei',6:'Jun',7:'Jul', 8:'Aug', 9:'Sep', 10:'Oct', 11:'Nov', 12:'Dec'}
      merge_dfyar['Month']= merge_dfyar['Month'].map(dmap)

print(merge_dfyar)
```

	Date	State	Dead	Injured	Total	\
0	2018-12-31	Ohio	3	2	5	

Gambar 11

- Mengetahui tahun yang paling sering terjadi penembakan



Gambar 12

- Mengetahui musim yang paling sering terjadi penembakan


```
Dari kolom gender, kami mencari gender pelaku dengan mencari kata yang menandakan hal yang dilakukan atau merujuk pada pelaku.

# mencari gender pelaku menggunakan pattern
merge_dfyar['pelaku_gender'] = merge_dfyar['Description'].str.contains("man killed | men killed | man shot | man wounded | men wounded | a man | a men | a gunman | gunmen ", case=False)

# pelaku = merge_dfyar.loc[merge_dfyar['pelaku_gender'] == True]

# error_gen = merge_dfyar[merge_dfyar['gender'] != merge_dfyar['pelaku_gender']]

# error_gen

merge_dfyar
# pelaku
```

Gambar 16

- Menganalisis kombinasi kata yang ada dan menampilkan kata yang paling banyak muncul pada kolom 'Description'.

```
Menganalisis kombinasi kata yang ada dan menampilkan kata yang paling banyak muncul pada kolom 'Description'.

[94]
text = " ".join(merge_dfyar['Description'])
print ("There are {} words in the combination of all descriptions.".format(len(text)))

There are 272452 words in the combination of all descriptions.

[95] # Create stopwords:
stopwords = STOPWORDS

# Generate a word cloud image
wordcloud_desc = WordCloud(stopwords=stopwords, max_font_size=50, background_color="white").generate(text)

filtered_words = [word for word in text.split() if word not in stopwords]
counted_words = collections.Counter(filtered_words)

word_count = {}

for letter, count in counted_words.most_common(50):
    word_count[letter] = count

for i,j in word_count.items():
    print('Word: {0}, count: {1}'.format(i,j))

Word: people, count: 1114
```

Gambar 17

D. Menganalisis aturan atau regulasi penggunaan senjata api di setiap negara bagian

Dengan menggunakan dataset yang baru yaitu mengenai jumlah aturan atau regulasi yang ada di setiap negara bagian Amerika Serikat, kami melakukan analisis mengenai jumlah aturan dan mengelompokkan menjadi 3 kelas yaitu Ketat, Sedang dan Longgar

▼ Feature Extraction

Membuat tabel baru mengenai data tiap negara bagian, berkaitan dengan aturan penggunaan senjata api di tahun 2017
Kemudian, mengelompokan jumlah aturan menjadi 3 kelas yaitu Ketat, Sedang dan Longgar

```
[80] df_law = pd.read_csv('raw_data.csv')

df_law2 = df_law.loc[df_law.year == 2017]

total_data = pd.DataFrame(df_law2)
plt.figure(figsize=(10,15))
plt.xticks(rotation=90)
plt.barh(df_law2.state,df_law2.lawtotal)
```

Gambar 18

```
df_law3 = df_law2[['state', 'lawtotal']].reset_index(drop=True)
# df.loc[df["gender"] == "male", "gender"] = 1

# df_law3['lawtotal'] = np.where((df_law3['lawtotal']<))

def f(row):
    if row['lawtotal'] < 30:
        val = 'Rendah'
    elif row['lawtotal'] < 65:
        val = 'Sedang'
    else:
        val = 'Ketat'
    return val

df_law3['Keketatan Regulasi'] = df_law3.apply(f, axis=1)

print(df_law3)
```

Gambar 19

6. Visualisasi

6.1. Basic statistics

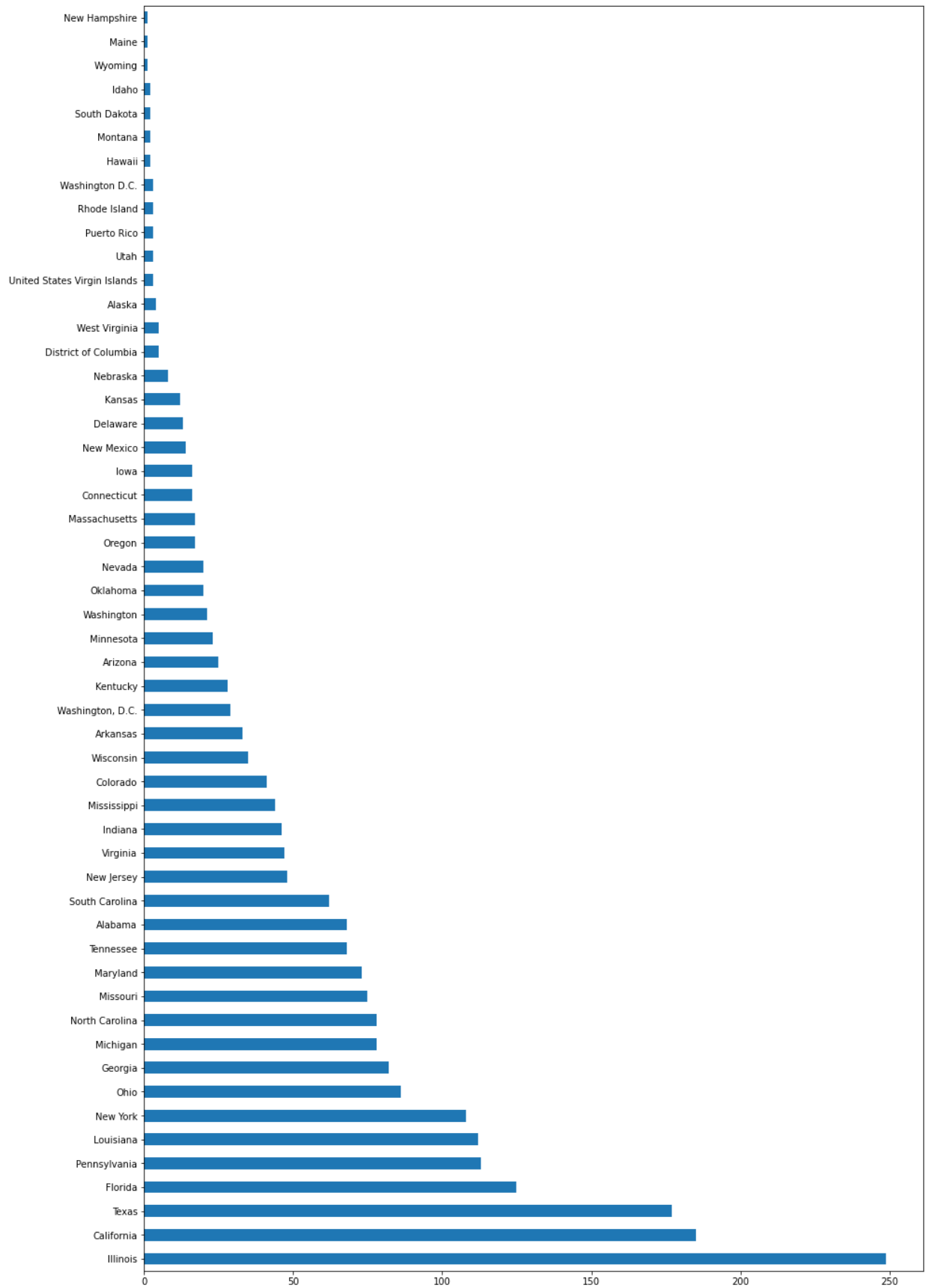
Menggunakan tabel yang dihasilkan oleh functions pandas

Basic Statistics						
	Dead	Injured	Total	Day	Year	
count	2352.000000	2352.000000	2352.000000	2352.000000	2352.000000	
mean	1.040391	4.078656	5.119048	15.796344	2020.075680	
std	1.576783	2.308091	2.574345	8.847771	1.226832	
min	0.000000	0.000000	2.000000	1.000000	2018.000000	
25%	0.000000	3.000000	4.000000	8.000000	2019.000000	
50%	1.000000	4.000000	4.000000	16.000000	2020.000000	
75%	1.000000	5.000000	5.000000	23.250000	2021.000000	
max	23.000000	27.000000	46.000000	31.000000	2022.000000	

Gambar 20

6.2. Total jumlah kasus penembakan yang terjadi di setiap negara bagian

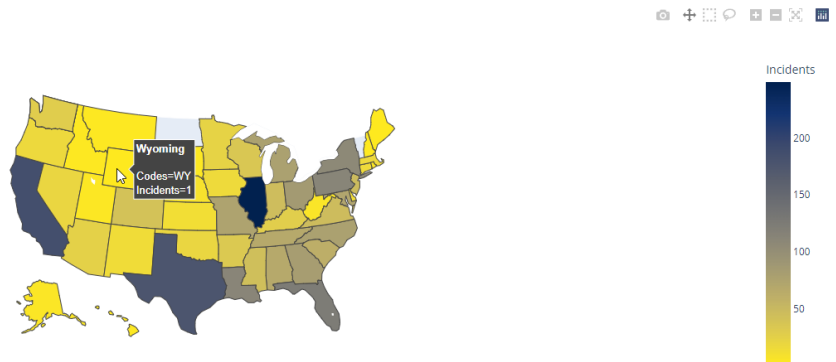
menggunakan bar chart



Gambar 21

Selain itu, kami memvisualisasikan juga menggunakan Spatial Map

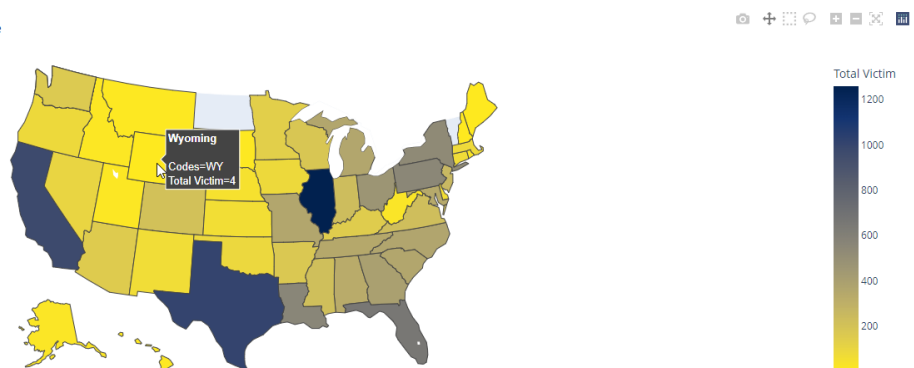
Incidents Map since 2018 by State



Gambar 22

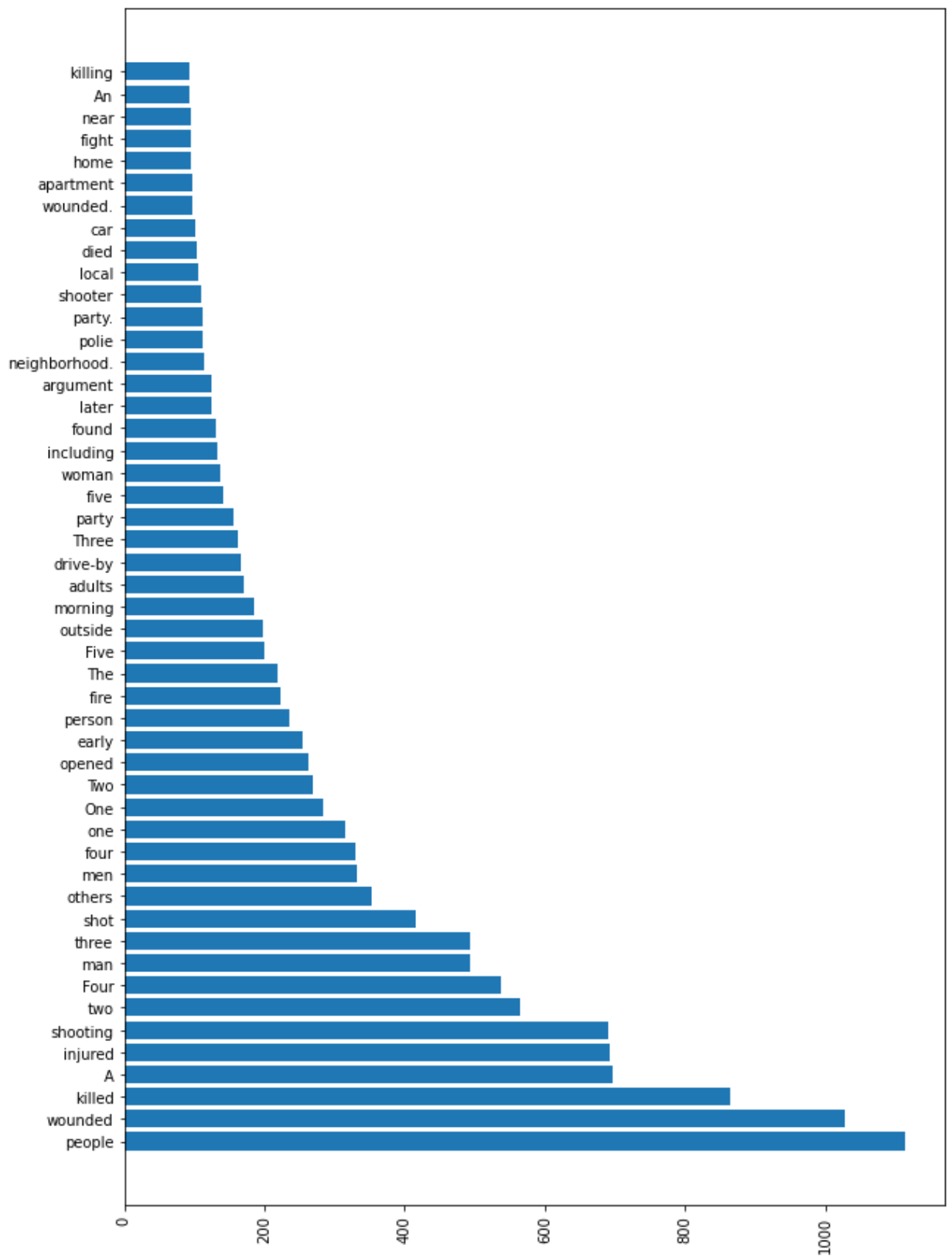
6.3. Total jumlah korban dalam setiap negara bagian Menggunakan spatial map

Total Victim Map since 2018 by State



Gambar 23

6.4. Attribute baru yaitu 'gender' dan 'pelaku_gender' Menambahkan kolom atau attribute baru ke dataframe

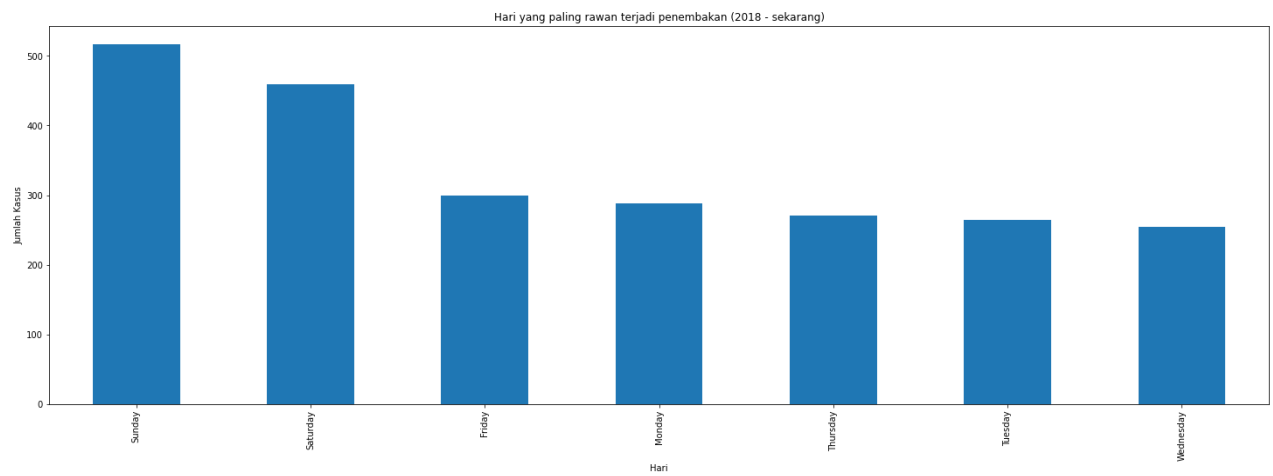


Gambar 26

6.6. Penembakan sering terjadi dikelompokkan berdasarkan

- Hari

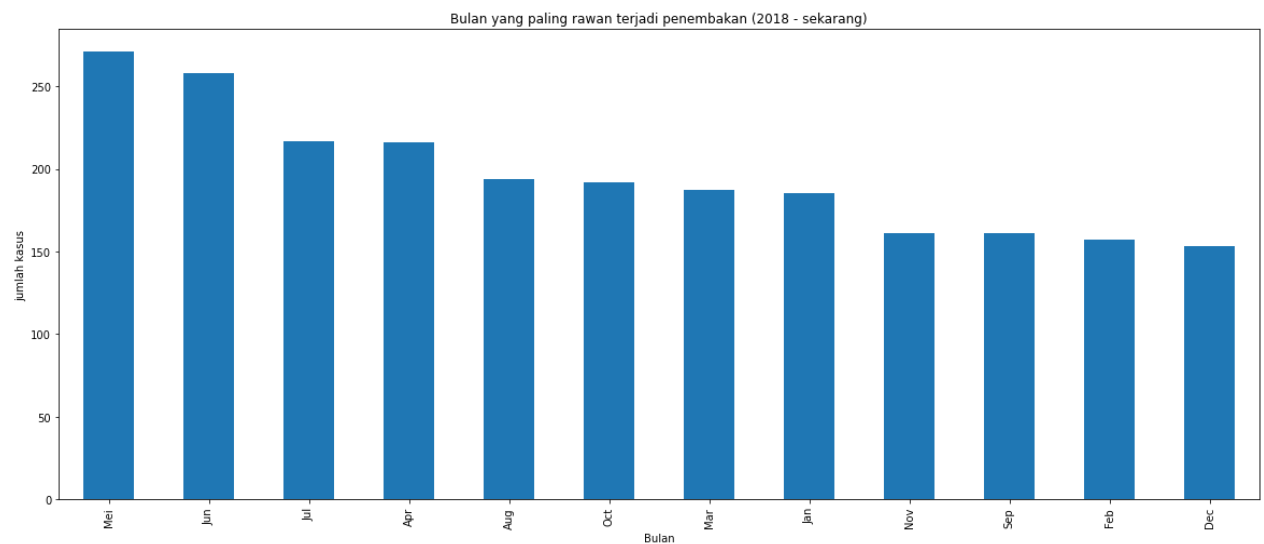
Menggunakan barchart



Gambar 27

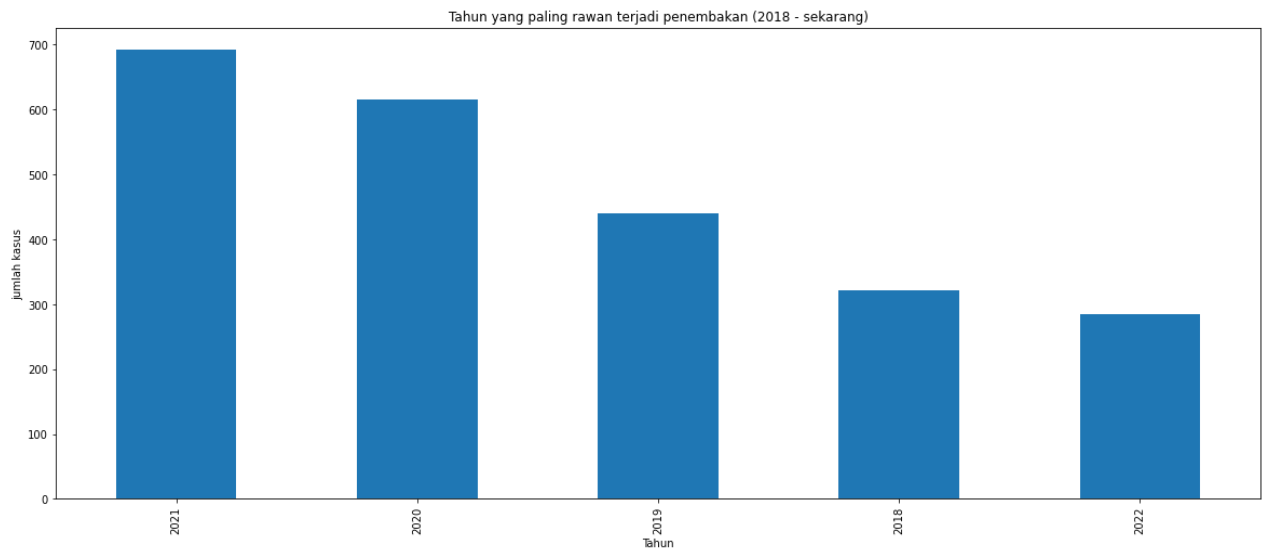
- Bulan

Menggunakan barchart



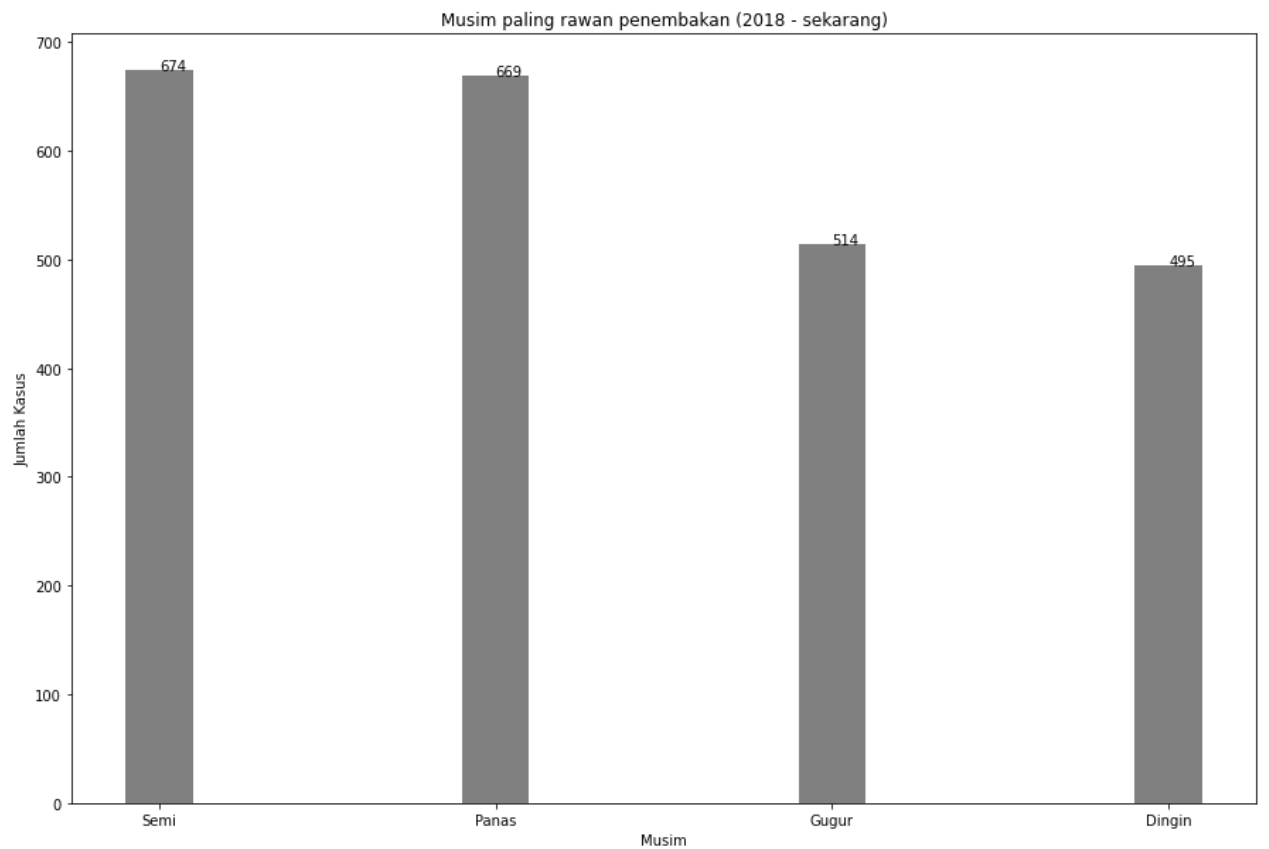
Gambar 28

- Tahun
Menggunakan barchart



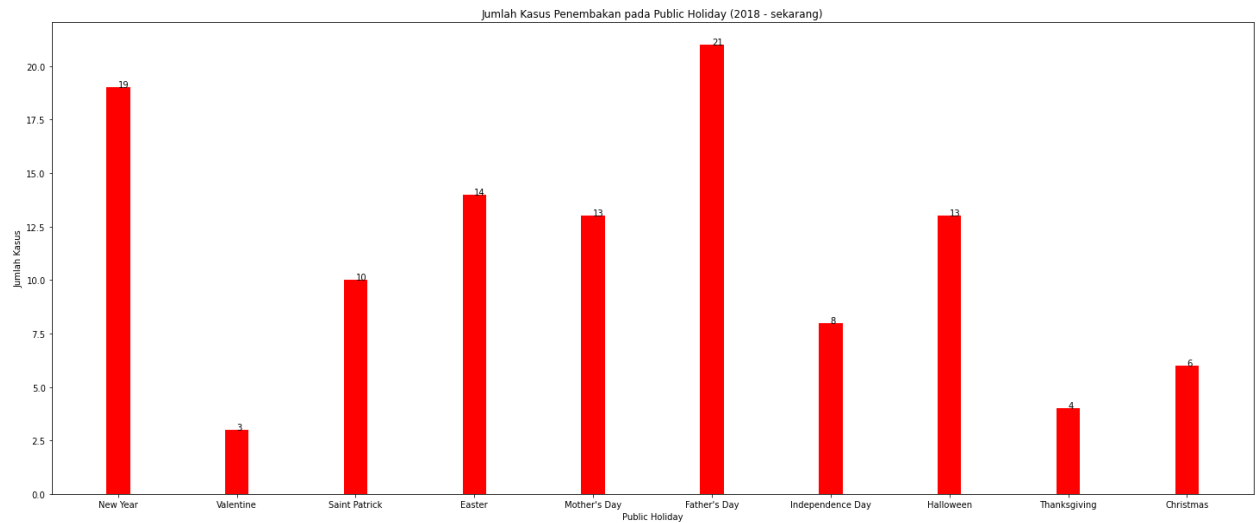
Gambar 29

- Musim
Menggunakan barchart



Gambar 30

- Public Holiday
Menggunakan barchart



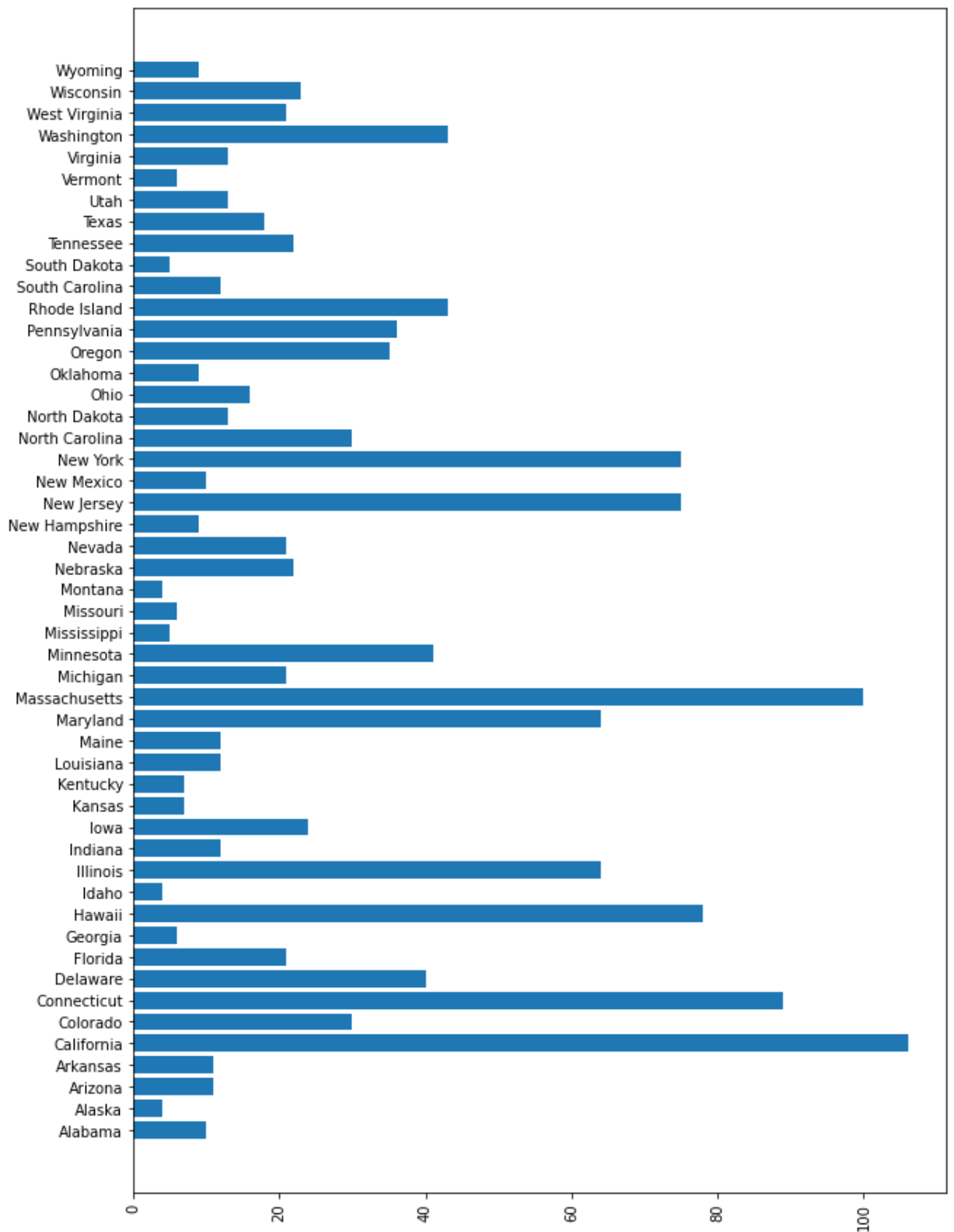
Gambar 31

- 6.7. Aturan atau regulasi penggunaan senjata api di setiap negara bagian dan dikelompokkan menjadi 3 kelas yaitu Ketat, Sedang dan Longgar
Menggunakan tabel baru

	state	lawtotal	Keketatan	Regulasi
0	Alabama	10		Rendah
1	Alaska	4		Rendah
2	Arizona	11		Rendah
3	Arkansas	11		Rendah
4	California	106		Ketat
5	Colorado	30		Sedang
6	Connecticut	89		Ketat
7	Delaware	40		Sedang
8	Florida	21		Rendah
9	Georgia	6		Rendah
10	Hawaii	78		Ketat
11	Idaho	4		Rendah
12	Illinois	64		Sedang
13	Indiana	12		Rendah
14	Iowa	24		Rendah
15	Kansas	7		Rendah
16	Kentucky	7		Rendah
17	Louisiana	12		Rendah
18	Maine	12		Rendah
19	Maryland	64		Sedang
20	Massachusetts	100		Ketat
21	Michigan	21		Rendah
22	Minnesota	41		Sedang
23	Mississippi	5		Rendah
24	Missouri	6		Rendah
25	Montana	4		Rendah
26	Nebraska	22		Rendah
27	Nevada	21		Rendah
28	New Hampshire	9		Rendah
29	New Jersey	75		Ketat
30	New Mexico	10		Rendah
31	New York	75		Ketat
32	North Carolina	30		Sedang
33	North Dakota	13		Rendah
34	Ohio	16		Rendah
35	Oklahoma	9		Rendah
36	Oregon	35		Sedang
37	Pennsylvania	36		Sedang
38	Rhode Island	43		Sedang
39	South Carolina	12		Rendah
40	South Dakota	5		Rendah

Gambar 32

menggunakan bar chart



Gambar 33

7. Penutup

Dalam setiap poin analisis yang dilakukan kami mendapatkan, pertama adalah negara bagian Illinois memiliki jumlah kasus penembakan terbanyak dengan jumlah 249 kasus dan jumlah korban terbanyak juga dipegang oleh Illinois sebanyak 1260 korban. Kedua, kami mendapatkan hasil bahwa hari paling banyak terjadi penembakan adalah hari Minggu dan diikuti oleh hari Sabtu. Tetapi dalam data tersebut, didapatkan bahwa hari Jumat memiliki selisih yang cukup banyak dengan hari Sabtu yaitu sebanyak 160 kasus penembakan. Oleh karena itu, kami berpikir bahwa akhir pekan memungkinkan meningkatnya peluang terjadinya penembakan karena banyak penduduk yang tidak menjalani kesibukan apapun dan lebih banyak melakukan aktivitas sosial dengan orang lain.

Selanjutnya, berdasarkan data jumlah penembakan menurut bulan dan tahunnya, bulan Mei memiliki jumlah kasus terbanyak dengan jumlah 271 kasus penembakan dan diikuti oleh bulan Juni. Lalu, untuk kasus jumlah penembakan pertahunnya, tahun dengan jumlah kasus terbanyak adalah tahun 2021 dengan jumlah kasus penembakan sebanyak 692 dan diikuti oleh tahun 2020. Kami berhipotesis bahwa tahun 2020 dan 2021 merupakan tahun dengan jumlah kasus terbanyak karena dipengaruhi oleh kondisi masyarakat saat pandemi COVID-19. Selain bulan dan tahun, musim juga mempengaruhi banyaknya kasus penembakan dalam negara-negara bagian Amerika Serikat. Analisis kami mendapatkan bahwa kasus penembakan terjadi paling banyak pada musim semi dengan jumlah kasus penembakan sebanyak 674 dan diikuti oleh musim panas. Perbandingan drastis antara musim gugur dan panas juga memberikan hipotesis bahwa libur musim panas berkemungkinan meningkatkan kasus penembakan karena pada musim panas penduduk menjalani aktivitas dengan orang lain.

Ketiga, Selama tahun 2017, kami dapat mengetahui bahwa California memiliki 106 regulasi penggunaan senjata api yang merupakan regulasi yang sangat ketat. Namun cukup ironis jika dilihat dari insiden dan jumlah korban yang terjadi. Berbanding terbalik dengan negara yang memiliki regulasi yang rendah, seperti Alaska dan Idaho yang memiliki regulasi senjata api yang rendah namun jumlah korban dan insiden yang terjadi jauh lebih rendah dibandingkan California. Kesimpulan yang kami dapatkan, tingkat kasus penembakan massal menggunakan senjata api di Amerika Serikat sangat tinggi meskipun sudah ada banyak regulasi mengenai kepemilikan senjata api. Terakhir, dengan menggunakan kolom 'Description', kami dapat mengetahui jumlah kombinasi kata yang ada dan dapat diketahuinya gender pelaku dari kolom tersebut. Untuk kedepannya, yang harus dikembangkan dari analisis kami adalah data mengenai regulasi tentang kepemilikan senjata api di setiap negara bagian Amerika Serikat, karena data yang kami gunakan pada saat ini adalah data dari tahun 2017, dan ada kemungkinan bahwa regulasi-regulasi tersebut sudah diperbaharui oleh pemerintah Amerika Serikat.