



Group 17

Nimal Kumarr Arunkumar (801317922)

Samhitha Mudiam (801329504)

Hrushika Pappuri (801060321)

Tanuj Darla (801316530)

Shravan Naidu Gollu (801318790)

Big Data Analytics for Competitive Advantage

ITCS 6100

Deliverable 2

Gun_violence_data

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

Importing the dataset

```
In [2]: gun_data = pd.read_csv('Gun_violence_data.csv', dtype=str)
```

```
In [3]: gun_data
```

Out[3]:

	incident_id	date	day	state	city_or_county	address	n_killed	n_inj
0	461105	01-Jan-13	Tuesday	Pennsylvania	Mckeesport	1506 Versailles Avenue and Coursin Street	0	
1	460726	01-Jan-13	Tuesday	California	Hawthorne	13500 block of Cerise Avenue	1	
2	478855	01-Jan-13	Tuesday	Ohio	Lorain	1776 East 28th Street	1	
3	478925	05-Jan-13	Saturday	Colorado	Aurora	16000 block of East Ithaca Place	4	
4	478959	07-Jan-13	Monday	North Carolina	Greensboro	307 Mourning Dove Terrace	2	
...
241392	1293054	11-Sep-20	Friday	California	Torrance	22501 Hawthorne Blvd	3	
241393	1292941	12-Sep-20	Saturday	Arizona	Yuma	3800 block of E County 18 關 St	1	
241394	1291779	13-Sep-20	Sunday	Arkansas	Jonesboro	3516 Galaxy St	1	
241395	1289877	14-Sep-20	Monday	South Carolina	Columbia	1709 Decker Blvd	0	
241396	1289618	15-Sep-20	Tuesday	Florida	Tallahassee	2020 W Pensacola St	0	

241397 rows × 31 columns

Displaying the shape of dataset

In [4]: `gun_data.shape`

Out[4]: (241397, 31)

Displaying info of dataset

In [5]: `gun_data.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 241397 entries, 0 to 241396
Data columns (total 31 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   incident_id                          241397 non-null object
1   date                                241397 non-null object
2   day                                  241397 non-null object
3   state                                241397 non-null object
4   city_or_county                       241397 non-null object
5   address                              224899 non-null object
6   n_killed                             241397 non-null object
7   n_injured                            241395 non-null object
8   incident_url                         239677 non-null object
9   source_url                           239209 non-null object
10  incident_url_fields_missing          239677 non-null object
11  congressional_district               227733 non-null object
12  gun_stolen                           140179 non-null object
13  gun_type                             140226 non-null object
14  incident_characteristics             239351 non-null object
15  latitude                             231754 non-null object
16  location_description                 42089 non-null  object
17  longitude                             231754 non-null object
18  n_guns_involved                      140226 non-null object
19  notes                                158660 non-null object
20  participant_age                      147379 non-null object
21  participant_age_group                197558 non-null object
22  participant_gender                   203315 non-null object
23  participant_name                     117424 non-null object
24  participant_relationship              15774 non-null  object
25  participant_status                   212051 non-null object
26  participant_type                     214814 non-null object
27  sources                              239068 non-null object
28  state_house_district                 200905 non-null object
29  state_senate_district                207342 non-null object
30  critical                             241397 non-null object
dtypes: object(31)
memory usage: 57.1+ MB
```

Displaying missing values from the dataset

In [6]: `gun_data.isnull().sum()`

```
Out[6]: incident_id      0
        date            0
        day            0
        state          0
        city_or_county  0
        address        16498
        n_killed       0
        n_injured      2
        incident_url    1720
        source_url      2188
        incident_url_fields_missing 1720
        congressional_district 13664
        gun_stolen     101218
        gun_type       101171
        incident_characteristics 2046
        latitude       9643
        location_description 199308
        longitude      9643
        n_guns_involved 101171
        notes          82737
        participant_age 94018
        participant_age_group 43839
        participant_gender 38082
        participant_name 123973
        participant_relationship 225623
        participant_status 29346
        participant_type 26583
        sources         2329
        state_house_district 40492
        state_senate_district 34055
        critical        0
        dtype: int64
```

Display the number and name of columns

```
In [7]: print('Columns :', list(gun_data))
        print('Number of columns :', len(list(gun_data)))
```

```
Columns : ['incident_id', 'date', 'day', 'state', 'city_or_county', 'address',
'n_killed', 'n_injured', 'incident_url', 'source_url', 'incident_url_fields_missing', 'congressional_district', 'gun_stolen', 'gun_type', 'incident_characteristics', 'latitude', 'location_description', 'longitude', 'n_guns_involved', 'notes', 'participant_age', 'participant_age_group', 'participant_gender', 'participant_name', 'participant_relationship', 'participant_status', 'participant_type', 'sources', 'state_house_district', 'state_senate_district', 'critical']
Number of columns : 31
```

Drop the less featured columns

```
In [8]: gun_data = gun_data.drop(['incident_url', 'source_url', 'incident_url_fields_missing', 'gun_type', 'incident_characteristics', 'latitude', 'n_guns_involved', 'notes', 'participant_age', 'participant_name', 'participant_relationship', 'participant_status', 'sources', 'state_house_district', 'state_senate_district'])
```

```
In [9]: gun_data.head()
```

```
Out[9]:
```

	incident_id	date	day	state	city_or_county	address	n_killed	n_injured	city
0	461105	01-Jan-13	Tuesday	Pennsylvania	Mckeesport	1506 Versailles Avenue and Coursin Street	0	4	F
1	460726	01-Jan-13	Tuesday	California	Hawthorne	13500 block of Cerise Avenue	1	3	F
2	478855	01-Jan-13	Tuesday	Ohio	Lorain	1776 East 28th Street	1	3	F
3	478925	05-Jan-13	Saturday	Colorado	Aurora	16000 block of East Ithaca Place	4	0	
4	478959	07-Jan-13	Monday	North Carolina	Greensboro	307 Mourning Dove Terrace	2	2	F

```
In [10]: gun_data.shape
```

```
Out[10]: (241397, 9)
```

Renaming the columns

```
In [11]: gun_data.rename(columns={"n_killed": "killed", "n_injured": "injured"}, inplace=True)
gun_data.head()
```

Out[11]:	incident_id	date	day	state	city_or_county	address	killed	injured	critical
0	461105	01-Jan-13	Tuesday	Pennsylvania	Mckeesport	1506 Versailles Avenue and Coursin Street	0	4	FALSE
1	460726	01-Jan-13	Tuesday	California	Hawthorne	13500 block of Cerise Avenue	1	3	FALSE
2	478855	01-Jan-13	Tuesday	Ohio	Lorain	1776 East 28th Street	1	3	FALSE
3	478925	05-Jan-13	Saturday	Colorado	Aurora	16000 block of East Ithaca Place	4	0	TRUE
4	478959	07-Jan-13	Monday	North Carolina	Greensboro	307 Mourning Dove Terrace	2	2	FALSE

In [12]: `gun_data.isnull().sum()`

```
Out[12]: incident_id      0
         date            0
         day            0
         state          0
         city_or_county  0
         address      16498
         killed         0
         injured        2
         critical       0
         dtype: int64
```

In [13]: `gun_data.dtypes`

```
Out[13]: incident_id      object
         date            object
         day            object
         state          object
         city_or_county  object
         address        object
         killed         object
         injured        object
         critical       object
         dtype: object
```

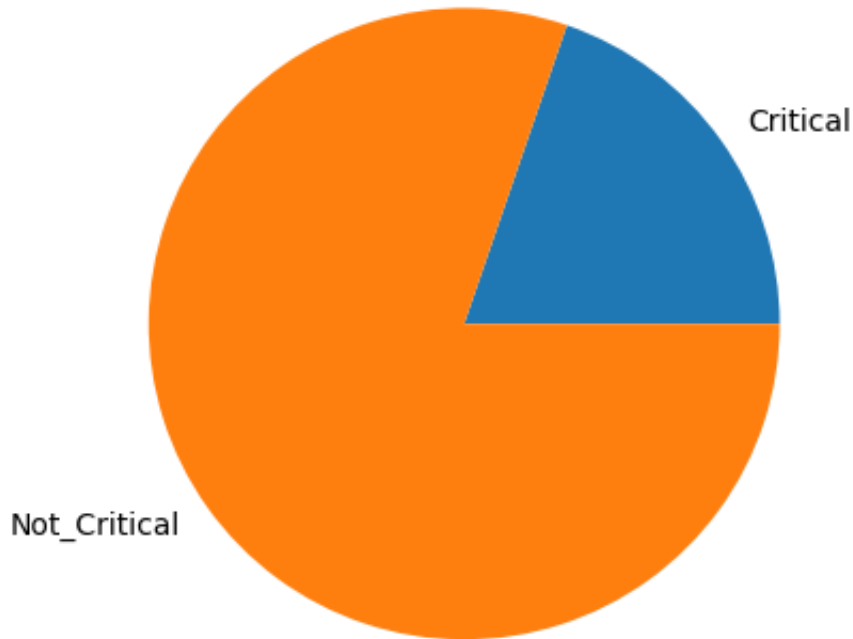
Displaying the critical and non-critical cases

```
In [21]: non_critical = len(gun_data[gun_data['critical'] == 'FALSE'])
critical = len(gun_data[gun_data['critical'] == 'TRUE'])
```

```
In [15]: print('Number of Critical cases : ', critical)
print('Number of Non-critical cases : ', non_critical)

plt.pie([critical,non_critical], labels=['Critical', 'Not_Critical'], radius
plt.show()
```

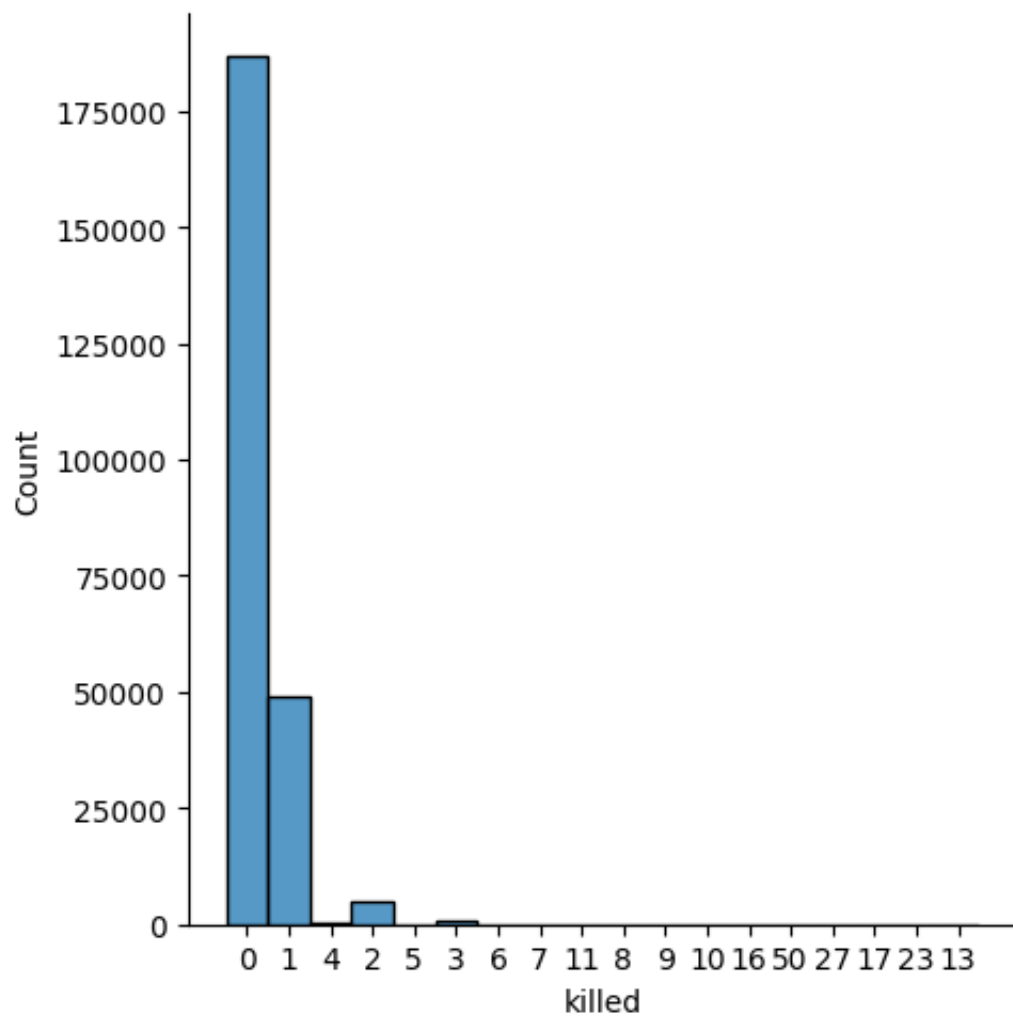
Number of Critical cases : 47607
Number of Non-critical cases : 193790



Displaying the killed count

```
In [16]: sns.displot(gun_data,x="killed",binwidth=10)
```

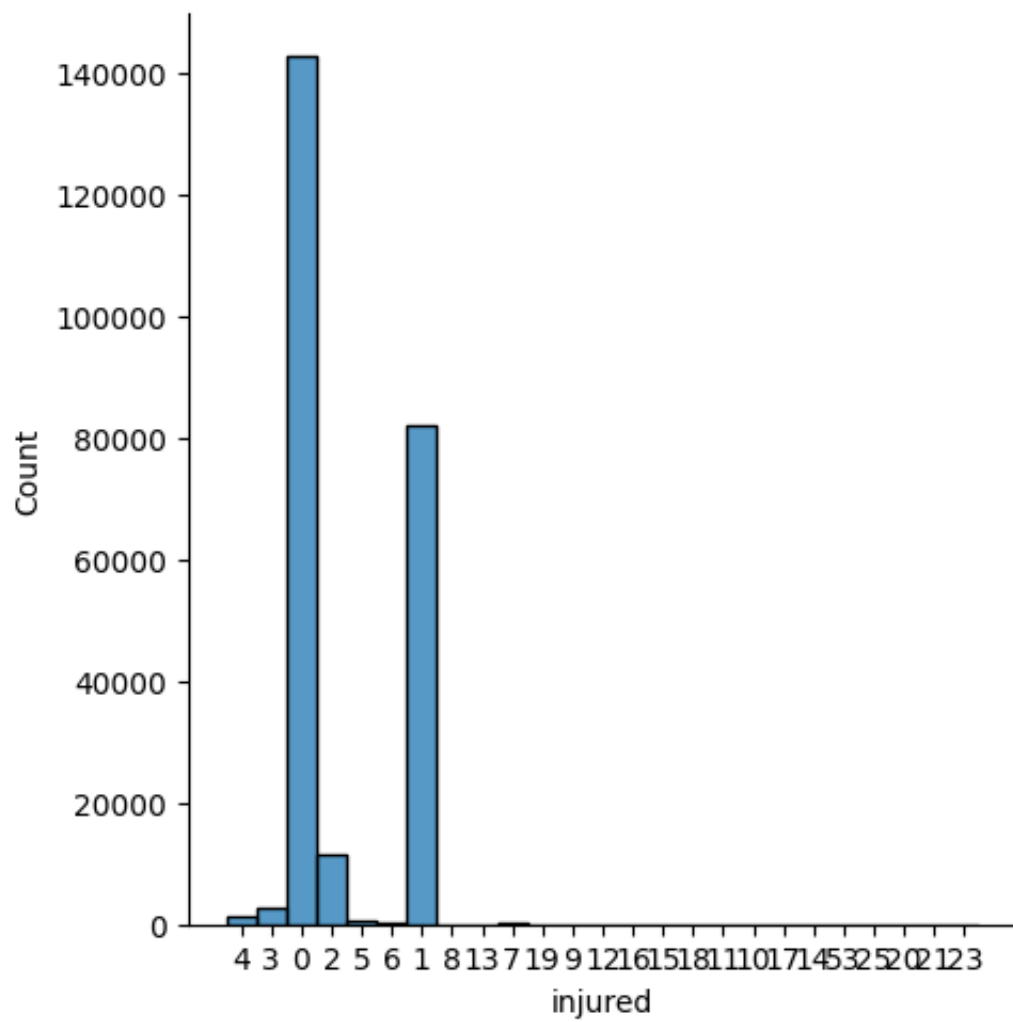
```
Out[16]: <seaborn.axisgrid.FacetGrid at 0x7fc2bcd25a60>
```

Displaying the injured count

```
In [17]: sns.displot(gun_data,x="injured",binwidth=10)
```

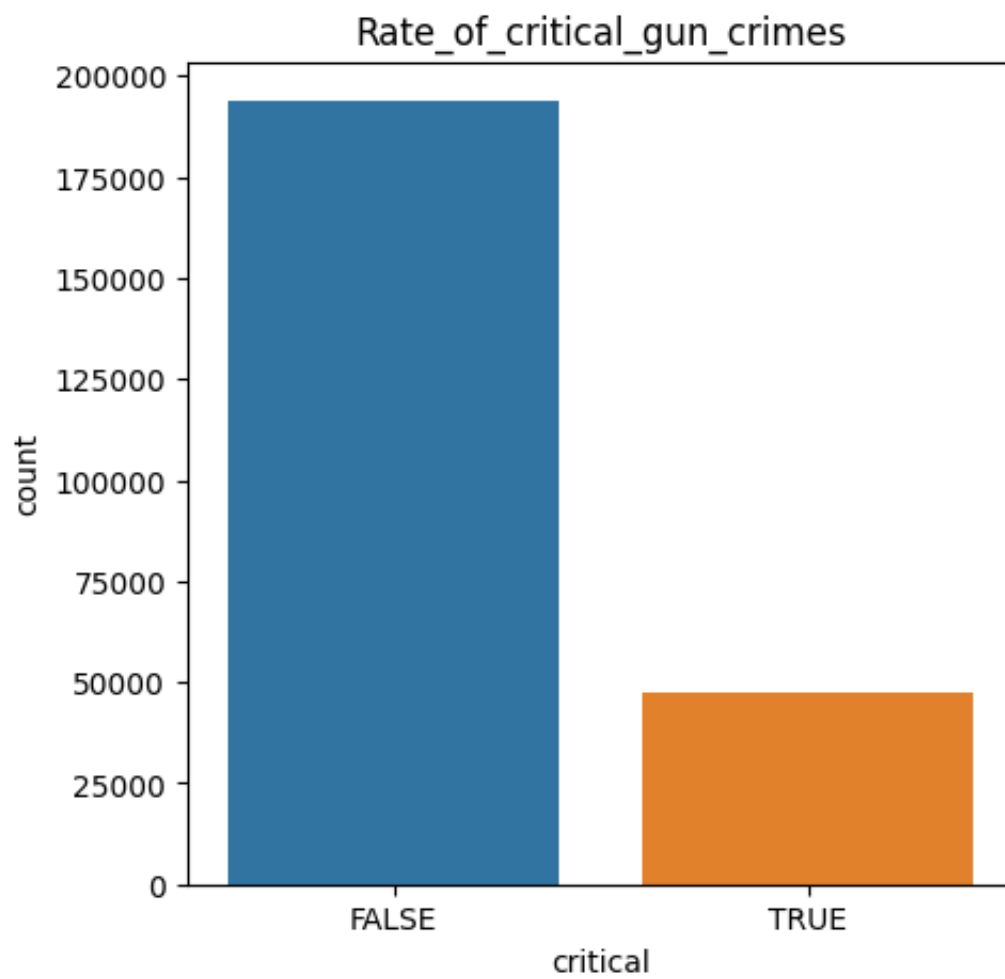
```
Out[17]: <seaborn.axisgrid.FacetGrid at 0x7fc2bcd15b80>
```



Checking the rate of critical gun crimes

```
In [18]: sns.countplot(x = gun_data['critical'])  
fig = plt.gcf()  
fig.set_size_inches(5,5)  
plt.title('Rate_of_critical_gun_crimes')
```

```
Out[18]: Text(0.5, 1.0, 'Rate_of_critical_gun_crimes')
```



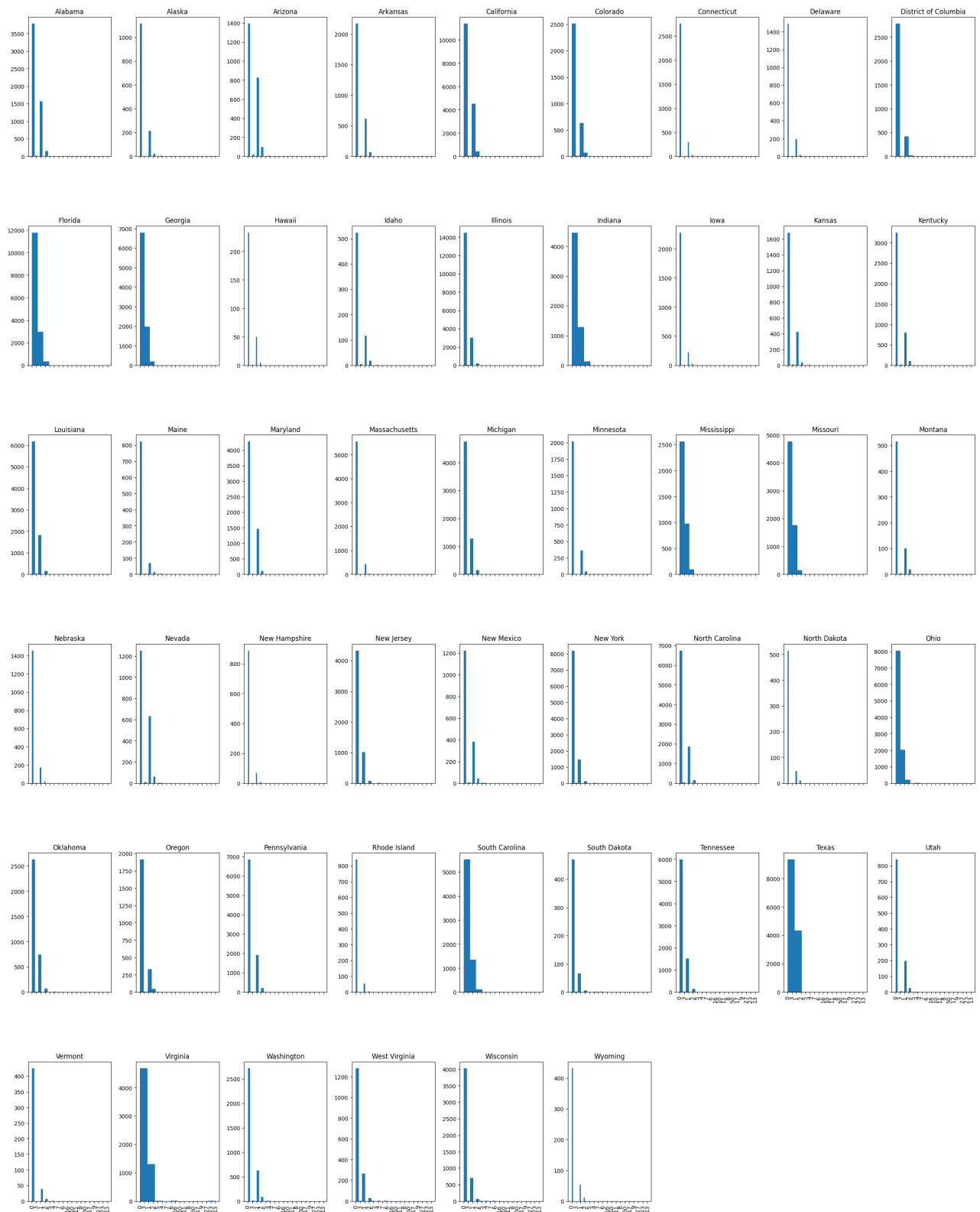
Displaying the killed count by the states

```
In [19]: gun_data.hist(column='killed',by='state',layout=(6,9),figsize=(30,40),sharex
```

```

Out[19]: array([[<Axes: title={ 'center': 'Alabama' }>,
<Axes: title={ 'center': 'Alaska' }>,
<Axes: title={ 'center': 'Arizona' }>,
<Axes: title={ 'center': 'Arkansas' }>,
<Axes: title={ 'center': 'California' }>,
<Axes: title={ 'center': 'Colorado' }>,
<Axes: title={ 'center': 'Connecticut' }>,
<Axes: title={ 'center': 'Delaware' }>,
<Axes: title={ 'center': 'District of Columbia' }>],
[<Axes: title={ 'center': 'Florida' }>,
<Axes: title={ 'center': 'Georgia' }>,
<Axes: title={ 'center': 'Hawaii' }>,
<Axes: title={ 'center': 'Idaho' }>,
<Axes: title={ 'center': 'Illinois' }>,
<Axes: title={ 'center': 'Indiana' }>,
<Axes: title={ 'center': 'Iowa' }>,
<Axes: title={ 'center': 'Kansas' }>,
<Axes: title={ 'center': 'Kentucky' }>],
[<Axes: title={ 'center': 'Louisiana' }>,
<Axes: title={ 'center': 'Maine' }>,
<Axes: title={ 'center': 'Maryland' }>,
<Axes: title={ 'center': 'Massachusetts' }>,
<Axes: title={ 'center': 'Michigan' }>,
<Axes: title={ 'center': 'Minnesota' }>,
<Axes: title={ 'center': 'Mississippi' }>,
<Axes: title={ 'center': 'Missouri' }>,
<Axes: title={ 'center': 'Montana' }>],
[<Axes: title={ 'center': 'Nebraska' }>,
<Axes: title={ 'center': 'Nevada' }>,
<Axes: title={ 'center': 'New Hampshire' }>,
<Axes: title={ 'center': 'New Jersey' }>,
<Axes: title={ 'center': 'New Mexico' }>,
<Axes: title={ 'center': 'New York' }>,
<Axes: title={ 'center': 'North Carolina' }>,
<Axes: title={ 'center': 'North Dakota' }>,
<Axes: title={ 'center': 'Ohio' }>],
[<Axes: title={ 'center': 'Oklahoma' }>,
<Axes: title={ 'center': 'Oregon' }>,
<Axes: title={ 'center': 'Pennsylvania' }>,
<Axes: title={ 'center': 'Rhode Island' }>,
<Axes: title={ 'center': 'South Carolina' }>,
<Axes: title={ 'center': 'South Dakota' }>,
<Axes: title={ 'center': 'Tennessee' }>,
<Axes: title={ 'center': 'Texas' }>,
<Axes: title={ 'center': 'Utah' }>],
[<Axes: title={ 'center': 'Vermont' }>,
<Axes: title={ 'center': 'Virginia' }>,
<Axes: title={ 'center': 'Washington' }>,
<Axes: title={ 'center': 'West Virginia' }>,
<Axes: title={ 'center': 'Wisconsin' }>,
<Axes: title={ 'center': 'Wyoming' }>, <Axes: >, <Axes: >,
<Axes: >]], dtype=object)

```



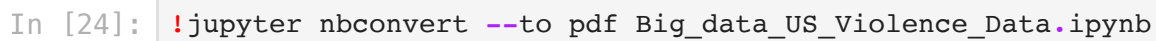
Displaying the injured count by the states

```
In [20]: gun_data.hist(column='injured',by='state',layout=(6,9),figsize=(30,40),share
```

```

Out[20]: array([[<Axes: title={ 'center': 'Alabama' }>,
<Axes: title={ 'center': 'Alaska' }>,
<Axes: title={ 'center': 'Arizona' }>,
<Axes: title={ 'center': 'Arkansas' }>,
<Axes: title={ 'center': 'California' }>,
<Axes: title={ 'center': 'Colorado' }>,
<Axes: title={ 'center': 'Connecticut' }>,
<Axes: title={ 'center': 'Delaware' }>,
<Axes: title={ 'center': 'District of Columbia' }>],
[<Axes: title={ 'center': 'Florida' }>,
<Axes: title={ 'center': 'Georgia' }>,
<Axes: title={ 'center': 'Hawaii' }>,
<Axes: title={ 'center': 'Idaho' }>,
<Axes: title={ 'center': 'Illinois' }>,
<Axes: title={ 'center': 'Indiana' }>,
<Axes: title={ 'center': 'Iowa' }>,
<Axes: title={ 'center': 'Kansas' }>,
<Axes: title={ 'center': 'Kentucky' }>],
[<Axes: title={ 'center': 'Louisiana' }>,
<Axes: title={ 'center': 'Maine' }>,
<Axes: title={ 'center': 'Maryland' }>,
<Axes: title={ 'center': 'Massachusetts' }>,
<Axes: title={ 'center': 'Michigan' }>,
<Axes: title={ 'center': 'Minnesota' }>,
<Axes: title={ 'center': 'Mississippi' }>,
<Axes: title={ 'center': 'Missouri' }>,
<Axes: title={ 'center': 'Montana' }>],
[<Axes: title={ 'center': 'Nebraska' }>,
<Axes: title={ 'center': 'Nevada' }>,
<Axes: title={ 'center': 'New Hampshire' }>,
<Axes: title={ 'center': 'New Jersey' }>,
<Axes: title={ 'center': 'New Mexico' }>,
<Axes: title={ 'center': 'New York' }>,
<Axes: title={ 'center': 'North Carolina' }>,
<Axes: title={ 'center': 'North Dakota' }>,
<Axes: title={ 'center': 'Ohio' }>],
[<Axes: title={ 'center': 'Oklahoma' }>,
<Axes: title={ 'center': 'Oregon' }>,
<Axes: title={ 'center': 'Pennsylvania' }>,
<Axes: title={ 'center': 'Rhode Island' }>,
<Axes: title={ 'center': 'South Carolina' }>,
<Axes: title={ 'center': 'South Dakota' }>,
<Axes: title={ 'center': 'Tennessee' }>,
<Axes: title={ 'center': 'Texas' }>,
<Axes: title={ 'center': 'Utah' }>],
[<Axes: title={ 'center': 'Vermont' }>,
<Axes: title={ 'center': 'Virginia' }>,
<Axes: title={ 'center': 'Washington' }>,
<Axes: title={ 'center': 'West Virginia' }>,
<Axes: title={ 'center': 'Wisconsin' }>,
<Axes: title={ 'center': 'Wyoming' }>, <Axes: >, <Axes: >,
<Axes: >]], dtype=object)

```



```
!jupyter nbconvert --to pdf Big_data_US_Violence_Data.ipynb
```

```

[NbConvertApp] Converting notebook Big_data_US_Violence_Data.ipynb to pdf
[NbConvertApp] Support files will be in Big_data_US_Violence_Data_files/
[NbConvertApp] Making directory ./Big_data_US_Violence_Data_files
[NbConvertApp] Making directory ./Big_data_US_Violence_Data_files
[NbConvertApp] Making directory ./Big_data_US_Violence_Data_files
[NbConvertApp] Making directory ./Big_data_US_Violence_Data_files
[NbConvertApp] Making directory ./Big_data_US_Violence_Data_files
[NbConvertApp] Making directory ./Big_data_US_Violence_Data_files
[NbConvertApp] Writing 57068 bytes to notebook.tex
[NbConvertApp] Building PDF
Traceback (most recent call last):
  File "/usr/local/bin/jupyter-nbconvert", line 8, in <module>
    sys.exit(main())
  File "/usr/local/lib/python3.9/dist-packages/jupyter_core/application.py",
line 277, in launch_instance
    return super().launch_instance(argv=argv, **kwargs)
  File "/usr/local/lib/python3.9/dist-packages/traitlets/config/application.
py", line 992, in launch_instance
    app.start()
  File "/usr/local/lib/python3.9/dist-packages/nbconvert/nbconvertapp.py", 1
ine 423, in start
    self.convert_notebooks()
  File "/usr/local/lib/python3.9/dist-packages/nbconvert/nbconvertapp.py", 1
ine 597, in convert_notebooks
    self.convert_single_notebook(notebook_filename)
  File "/usr/local/lib/python3.9/dist-packages/nbconvert/nbconvertapp.py", 1
ine 560, in convert_single_notebook
    output, resources = self.export_single_notebook(
  File "/usr/local/lib/python3.9/dist-packages/nbconvert/nbconvertapp.py", 1
ine 488, in export_single_notebook
    output, resources = self.exporter.from_filename(
  File "/usr/local/lib/python3.9/dist-packages/nbconvert/exporters/exporter.
py", line 189, in from_filename
    return self.from_file(f, resources=resources, **kw)
  File "/usr/local/lib/python3.9/dist-packages/nbconvert/exporters/exporter.
py", line 206, in from_file
    return self.from_notebook_node(
  File "/usr/local/lib/python3.9/dist-packages/nbconvert/exporters/pdf.py",
line 194, in from_notebook_node
    self.run_latex(tex_file)
  File "/usr/local/lib/python3.9/dist-packages/nbconvert/exporters/pdf.py",
line 164, in run_latex
    return self.run_command(
  File "/usr/local/lib/python3.9/dist-packages/nbconvert/exporters/pdf.py",
line 111, in run_command
    raise OSError(
OSError: xelatex not found on PATH, if you have not installed xelatex you ma
y need to do so. Find further instructions at https://nbconvert.readthedocs.io/en/latest/install.html#installing-tex.

```


Visualizations

