Zewail City of Science, Technology, and Innovation
University of Science and Technology

Communications and Information

Engineering

Business Report

Statistical Inference and Data Analysis - FALL2022

Prepared By

Abdullah Kamal Muhammad 201801271
Ibrahim Salah Abdelaal 201800224
Mohamed Abdelwahed 201801978
Mohamed Nasr 201801675

Supervised By
"Dr. Mahmoud Abdelaziz"
Eng Asmaa
Eng Anhar

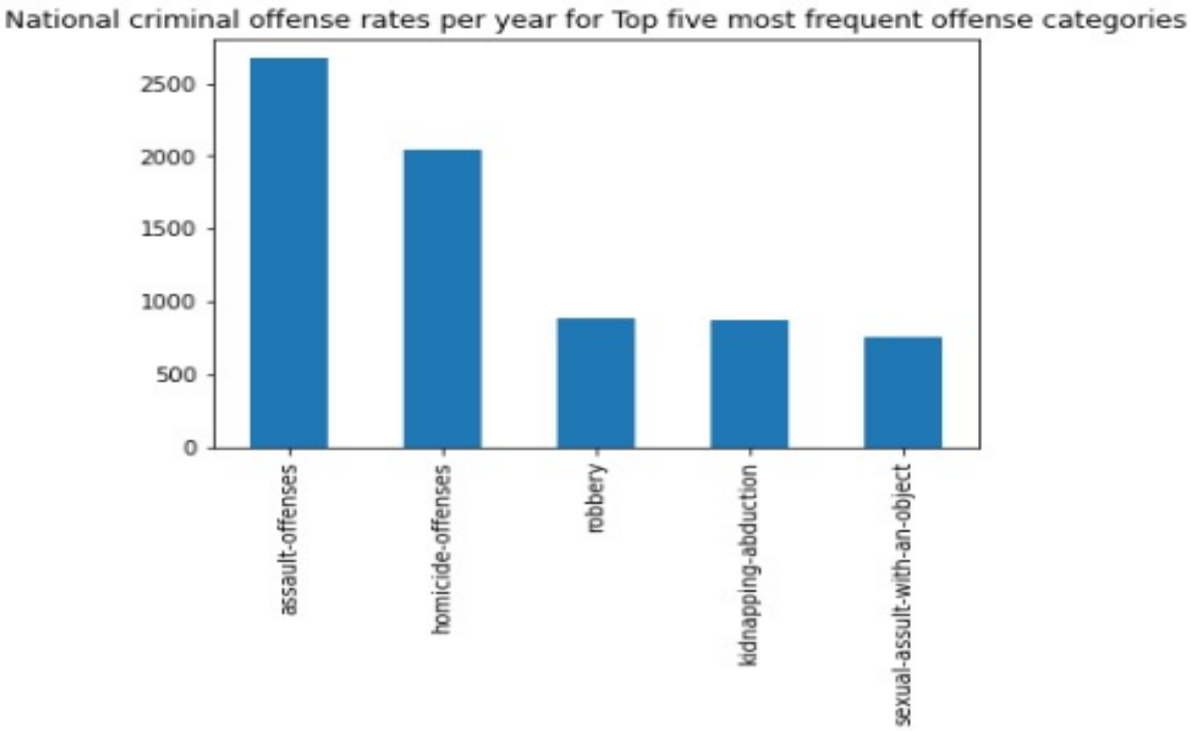
2022/2023

Introduction:

The main source of data on criminal victimization in the country is the BJS National Crime Victimization Survey (NCVS). Each year, statistics are collected from a sample of approximately 240,000 people in approximately 150,000 households, which is nationally representative. People are questioned on the frequency, traits, and effects of criminal victimization in the US. The NCVS gathers data on household property crimes (such as burglary/trespassing, motor vehicle theft, and other types of theft) as well as nonfatal personal crimes (such as rape or sexual assault, robbery, aggravated and simple assault, and personal larceny). Respondents to the survey disclose personal data about themselves, including their age, sex, race, and ethnicity (e.g., Hispanic origin), marital status, level of education, and income, as well as if they have ever been victims of victimization. The NCVS records each victimization incidence.

In this project we will use the publicly available crime data offered by the Federal Bureau of Investigation (FBI) and the Bureau of Justice Statistics (BJS) to analyze the patterns of crime in the U.S across time, regions, and demographics.
We find the relationships among different features and the maximum category in each feature

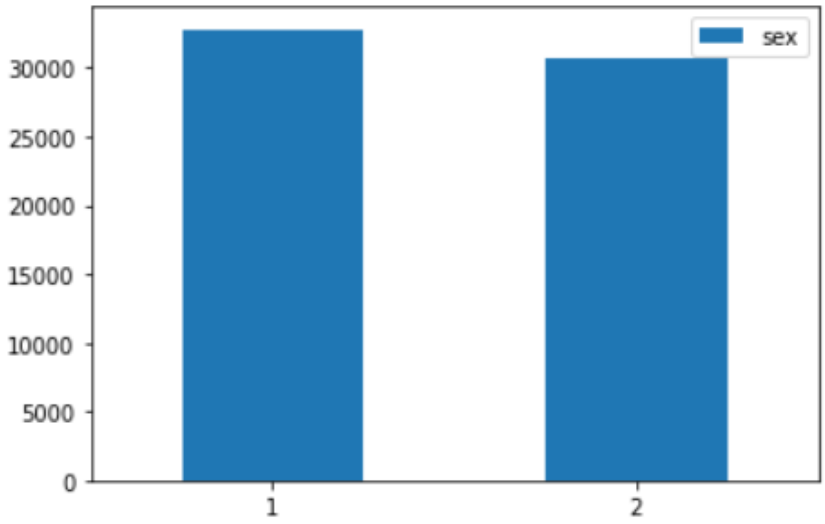
Plots:



- Visualizing the counts and plotting the frequency of the each age class show that class of assault offensive are the highest and the greater class among different crimes groups

1-

```
[1]  ##Plotting the frequency
     import seaborn as sns
     import matplotlib.pyplot as plt
     ax =pd.DataFrame(frequenc_sex).plot.bar(rot=0)
```



1	Male
2	Female

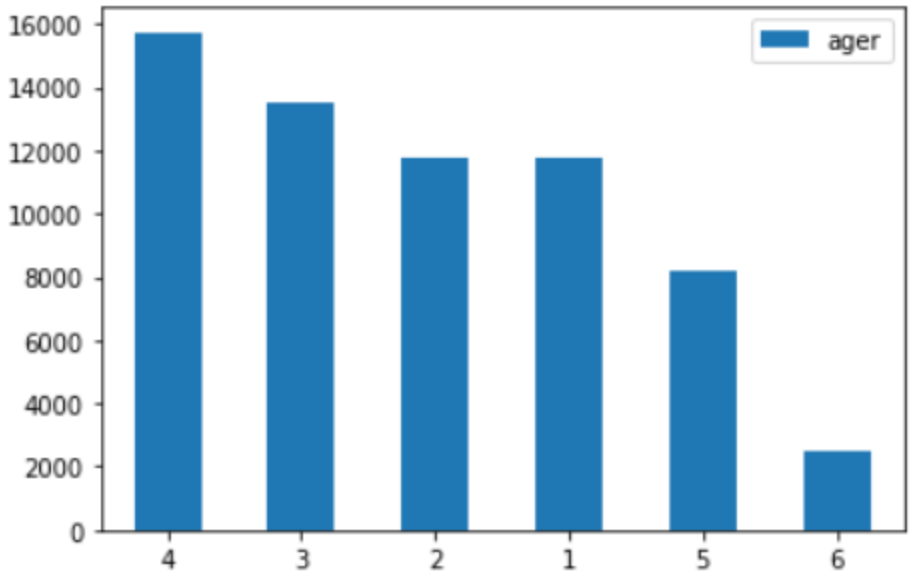
- Visualizing the counts and plotting the frequency of the each gender
  1 is male and 2 is for female
  We note that the number of males is greater than the number of females

****

2-

Age:

```
##Plotting the frequency
import seaborn as sns
import matplotlib.pyplot as plt
ax =pd.DataFrame(frequenc_age).plot.bar(rot=0)
```



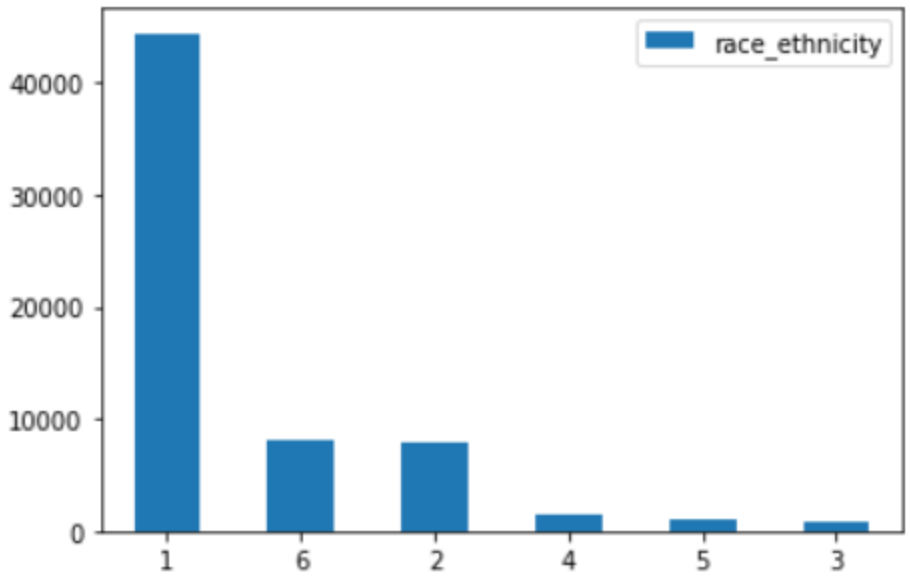
1	12-17
2	18-24
3	25-34
4	35-49
5	50-46
6	65:

- Visualizing the counts and plotting the frequency of the each age class show that class 4 which is 35-49 are the highest and the greater class among different age groups

****

3-

```
##Plotting the frequency
import seaborn as sns
import matplotlib.pyplot as plt
ax =pd.DataFrame(frequenc_race).plot.bar(rot=0)
```



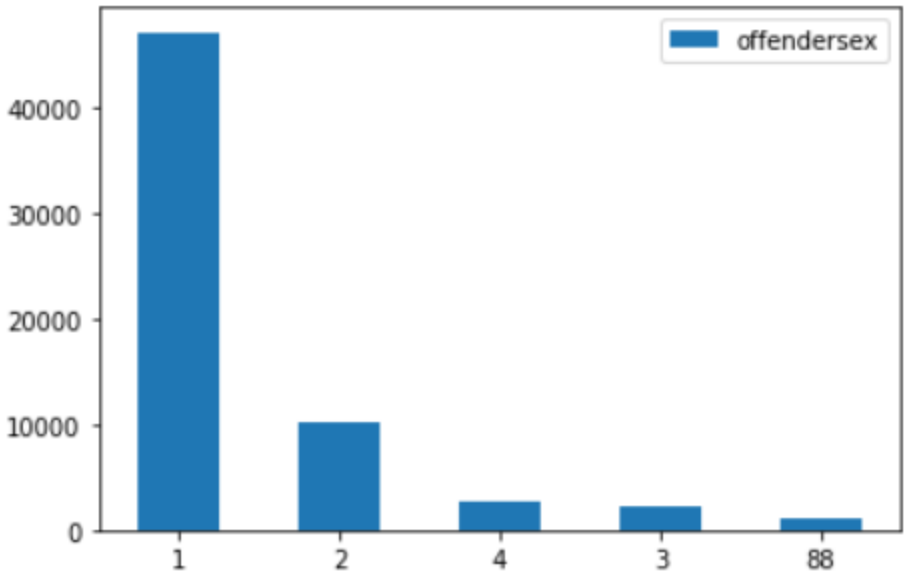
1	White
2	Black
3	American
4	Indian/ Alaska
5	Native Asian/Native Hawaiian
6	/Other Pacific Islander More than one race

- Visualizing the counts and plotting the frequency of the each age class show that class 1 which is white are the highest and the greater class among different race ethnicity groups

****

4-

```
##Plotting the frequency
import seaborn as sns
import matplotlib.pyplot as plt
ax =pd.DataFrame(frequenc_sex).plot.bar(rot=0)
```



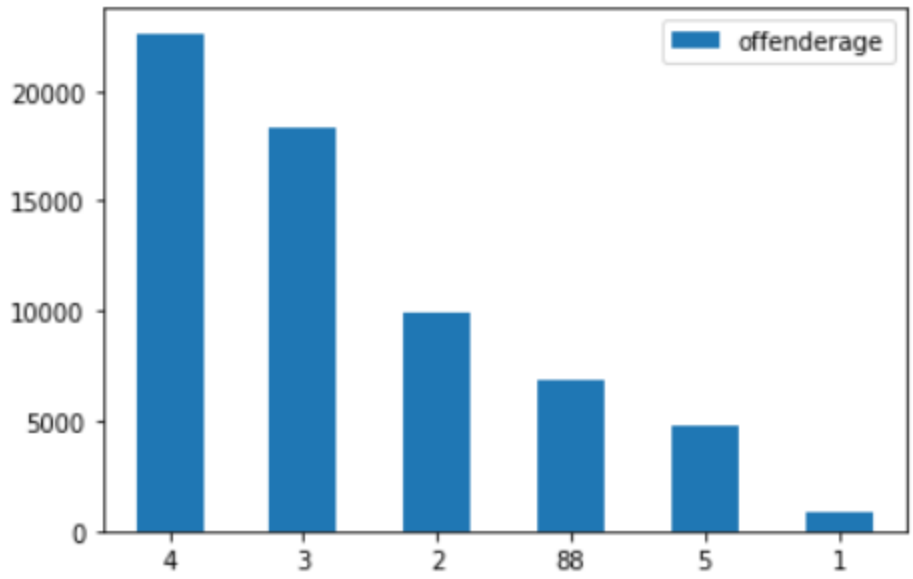
1	Male
2	Female
3	Both male and female
4	offenders
88	Unknown Residue

- Visualizing the counts and plotting the frequency of the each age class show that class 1 which is male are the highest and the greater class among different offendersex groups

****

5-

```
##Plotting the frequency
import seaborn as sns
import matplotlib.pyplot as plt
ax =pd.DataFrame(frequenc_age).plot.bar(rot=0)
```



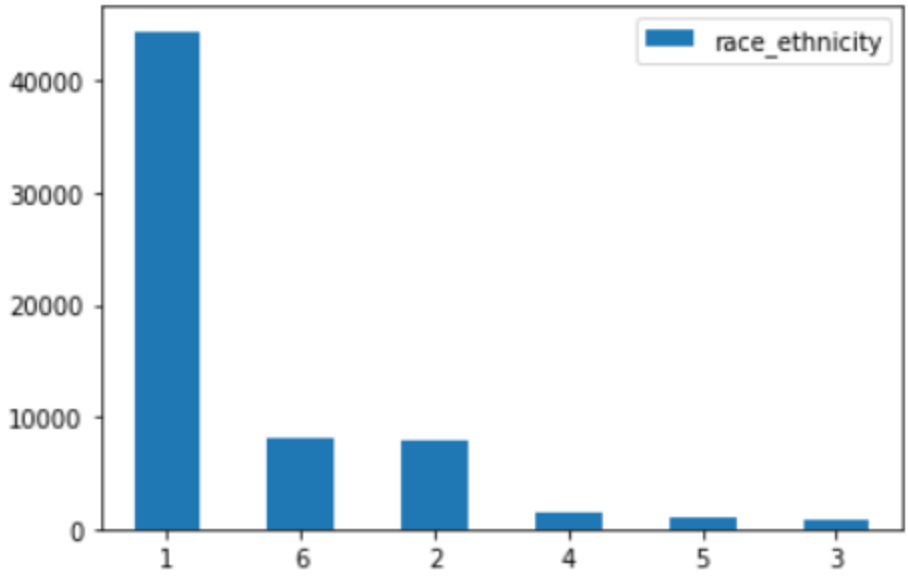
1	11 or younger
2	12-17
3	18-29
4	30 or older
5	Multiple offenders of various ages
88	Residue

- Visualizing the counts and plotting the frequency of the each age class show that class 4 which is 30 or older are the highest and the greater class among different offenderage groups

****

6-

```
##Plotting the frequency
import seaborn as sns
import matplotlib.pyplot as plt
ax =pd.DataFrame(frequenc_race).plot.bar(rot=0)
```



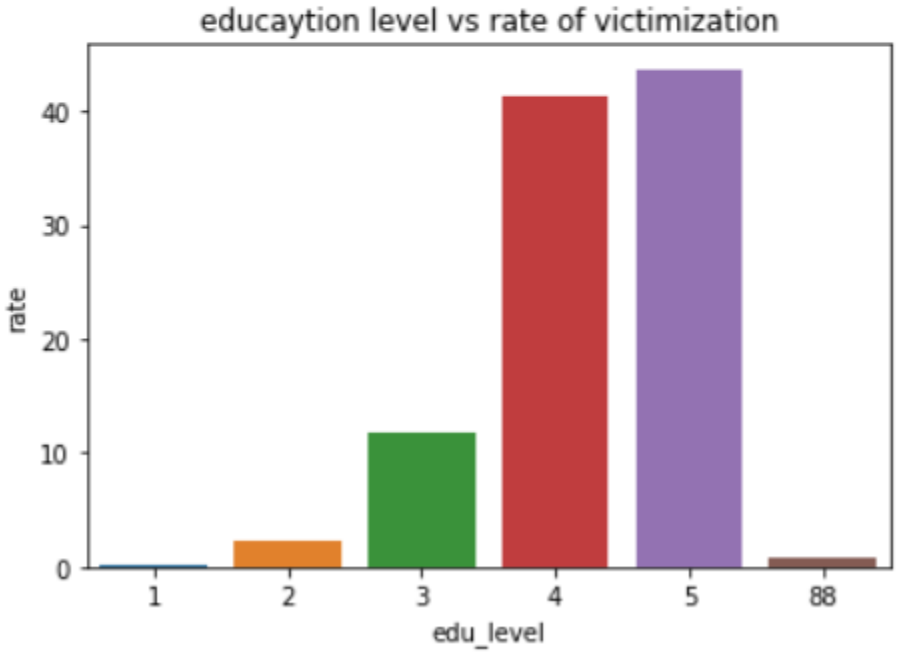
1	Non-Hispanic white
2	Non-Hispanic black
3	Non-Hispanic American Indian/ Alaska Native
4	Non-Hispanic Asian/ Native Hawaiian/ Other Pacific Islander
5	Non-Hispanic more than one race
6	Hispanic

- Visualizing the counts and plotting the frequency of the each age class show that class 1 which is Non- Hispanic white are the highest and the greater class among different race ethnicity groups

****

7-

```
level=list(edu_vicitm.to_frame().index)
rate=[(count/Total_count*100) for count in edu_vicitm]
zipped = list(zip(level, rate))
df = pd.DataFrame(zipped, columns=['edu_level', 'rate'])
sns.barplot(x ='edu_level', y ='rate', data = df)
# Show the plot
plt.title("educaytion level vs rate of victimization")
plt.show()
```



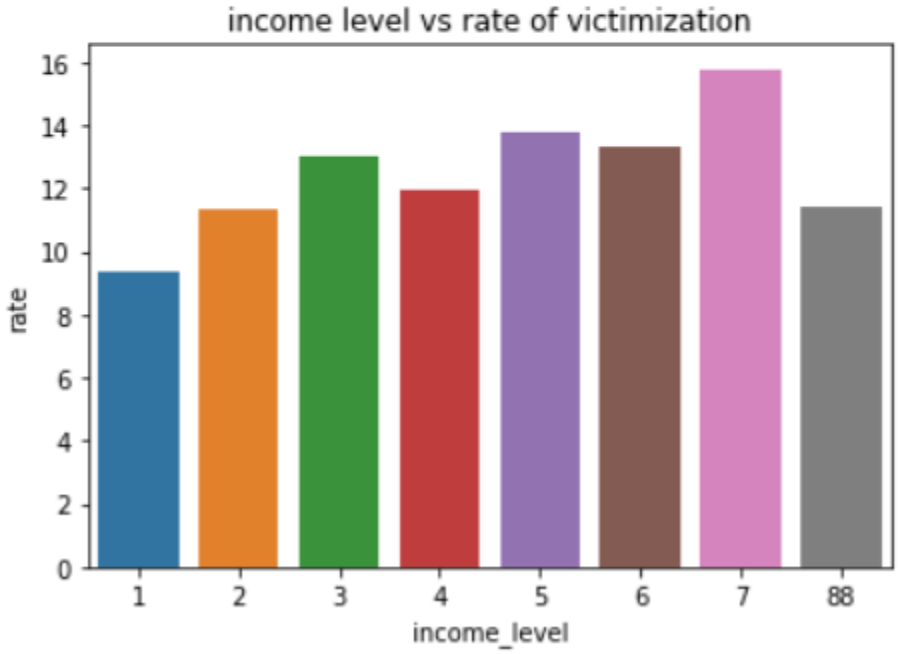
1	No schooling
2	Grade school
3	Middle school
4	High school
5	College
88	Residue

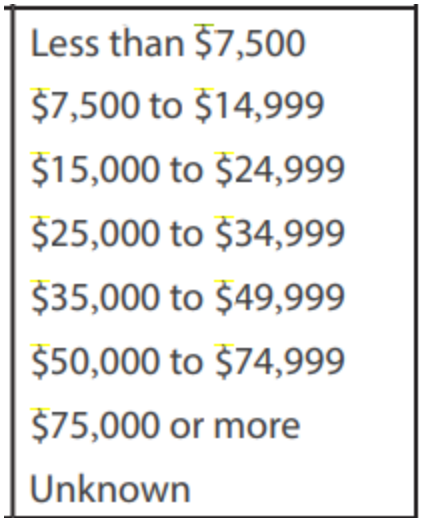
- Visualizing the counts and plotting the frequency of the each age class show that class 5 which is College are the highest and the greater class among different education level groups

****

8-

```
level=list(income_vicitm.to_frame().index)
rate=[(count/Total_count*100) for count in income_vicitm]
zipped = list(zip(level, rate))
df = pd.DataFrame(zipped, columns=['income_level', 'rate'])
sns.barplot(x ='income_level', y ='rate', data = df)
# Show the plot
plt.title("income level vs rate of victimization")
plt.show()
```



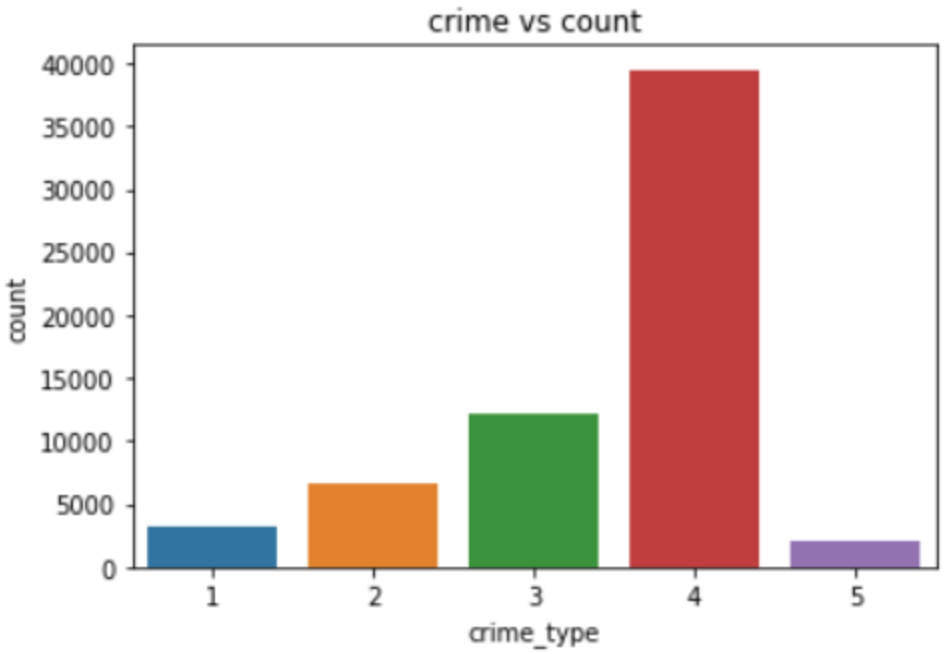


Form the plotting, we note that the highest level is class 7 which is $75,000 or more

****

9-

```
[ ]  ###most crime nonfatal
     crime_types_count=victim_data['newoff'].value_counts()
     level=list(crime_types_count.to_frame().index)
     rate=[(count) for count in crime_types_count]
     zipped = list(zip(level, rate))
     df = pd.DataFrame(zipped, columns=['crime_type', 'count'])
     sns.barplot(x ='crime_type', y ='count', data = df)
     # Show the plot
     plt.title("crime vs count")
     plt.show()
```



*Most crime reported is 4 *

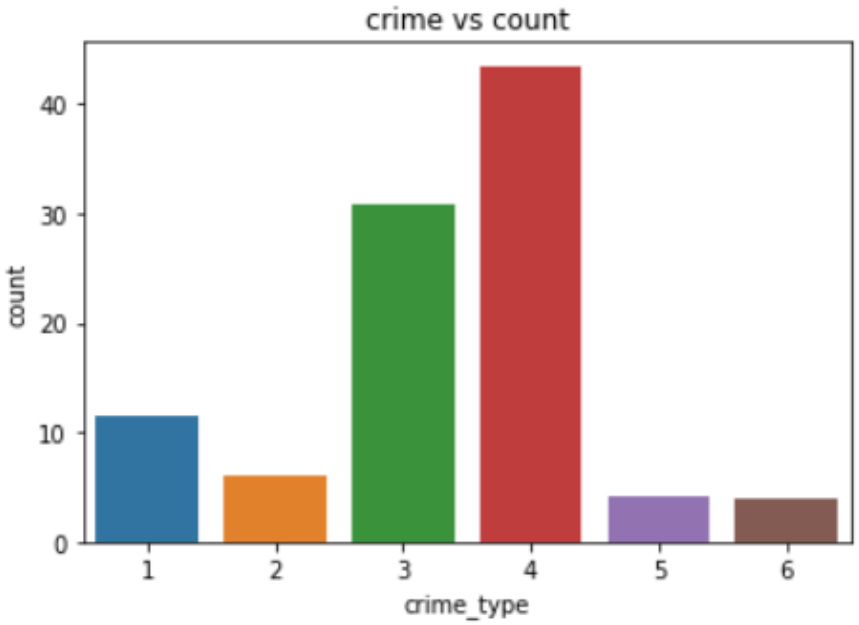
Form the plotting, we note that the highest level is class 4 which is Simple assault

1	Rape/sexual assault
2	Robbery
3	Aggravated assault
4	Simple assault
5	Personal theft/larceny

****

10-

```
###realtion ship crime type
crime_relation_count=victim_data['direl'].value_counts()
level=list(crime_relation_count.to_frame().index)
rate=[(count)/Total_count*100 for count in crime_relation_count]
zipped = list(zip(level, rate))
df = pd.DataFrame(zipped, columns=['crime_type', 'count'])
sns.barplot(x ='crime_type', y ='count', data = df)
# Show the plot
plt.title("crime vs count")
plt.show()
```



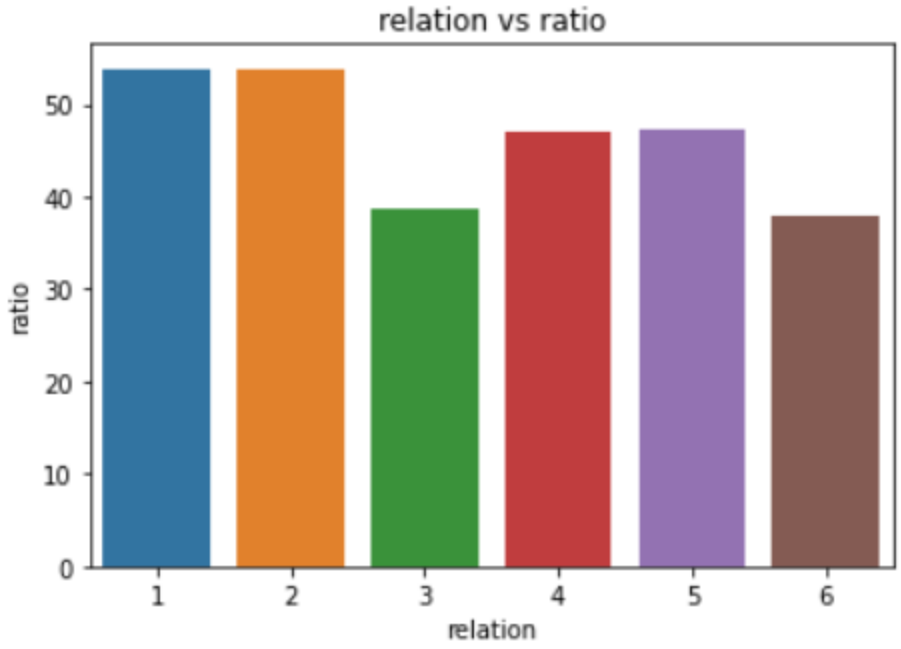
1	Intimates
2	Other relatives
3	Well known/casual acquaintance
4	Strangers
5	Do not know relationship
6	Do not know number of offenders

- Visualizing the counts and plotting the frequency of the each age class show that class 4 which is Strangers are the highest and the greater class among different crime relation groups

****

11-

```
###realtion ship crime type
level=list(ratio.to_frame().index)
rate=[(count) for count in ratio]
zipped = list(zip(level, rate))
df = pd.DataFrame(zipped, columns=['relation', 'ratio'])
sns.barplot(x ='relation', y ='ratio', data = df)
# Show the plot
plt.title("relation vs ratio")
plt.show()
```



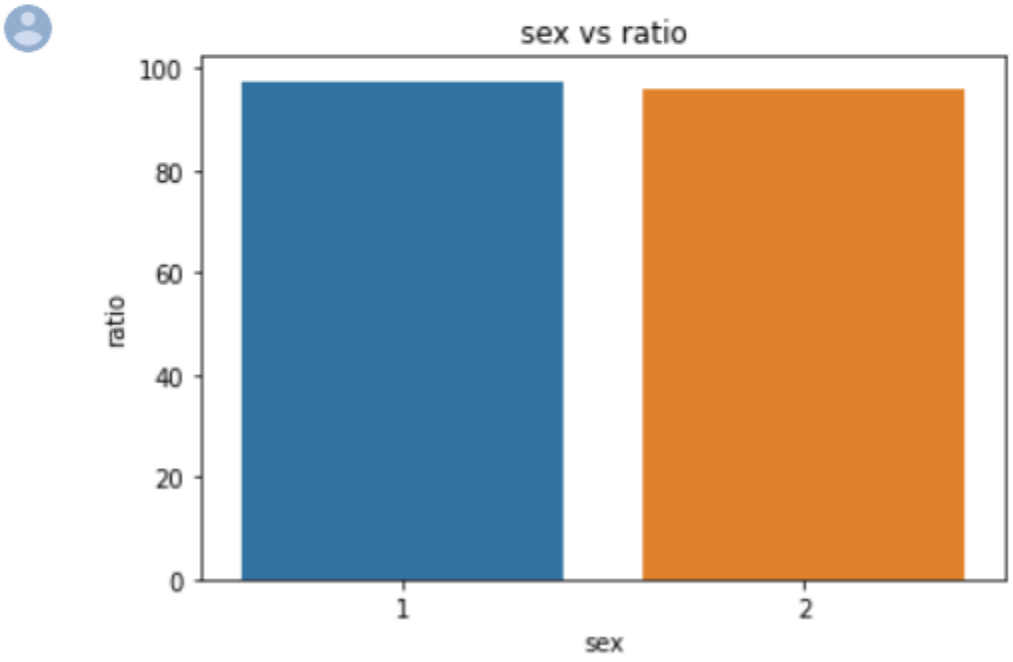
1. Intimates
2. Other relatives
3. Well known/casual acquaintance
4. Strangers
5. Do not know relationship
6. Do not know number of offenders

- Visualizing the counts and plotting the frequency of the each age class show that class 1 and 2 which are Intimates and Other relatives are the highest and the greater class among different relation groups

****

12-

```
###realtion ship crime type
level=list(ratio.to_frame().index)
rate=[(count) for count in ratio]
zipped = list(zip(level, rate))
df = pd.DataFrame(zipped, columns=['sex', 'ratio'])
sns.barplot(x ='sex', y ='ratio', data = df)
# Show the plot
plt.title("sex vs ratio")
plt.show()
```



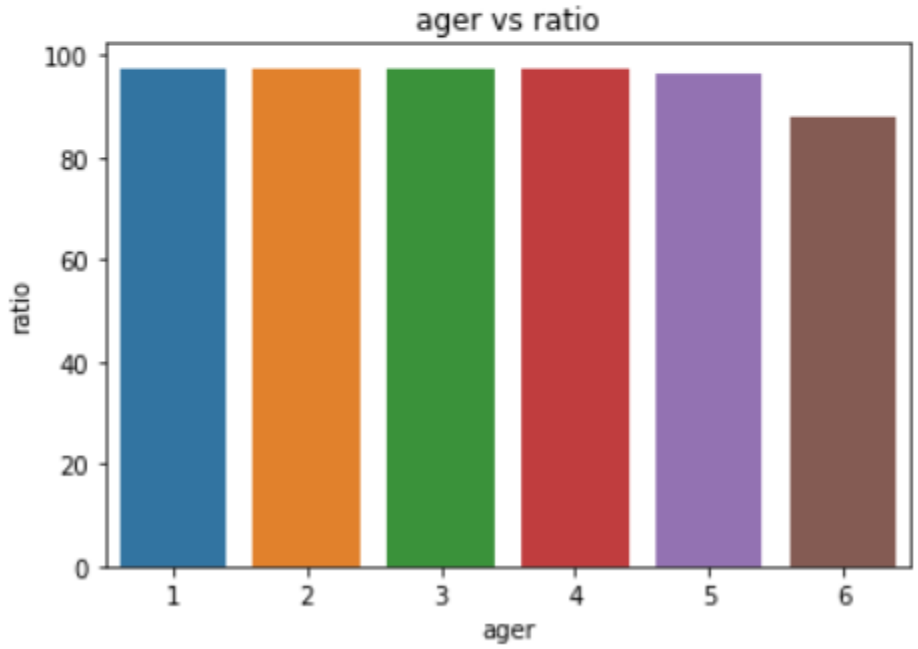
1	Male
2	Female

- Visualizing the counts and plotting the frequency of the each age class show that class 1 which is male are the highest and the greater class among different sex groups

****

13-

```
[ ]  ###realtion ship crime type
     level=list(ratio.to_frame().index)
     rate=[(count) for count in ratio]
     zipped = list(zip(level, rate))
     df = pd.DataFrame(zipped, columns=['ager', 'ratio'])
     sns.barplot(x ='ager', y ='ratio', data = df)
     # Show the plot
     plt.title("ager vs ratio")
     plt.show()
```



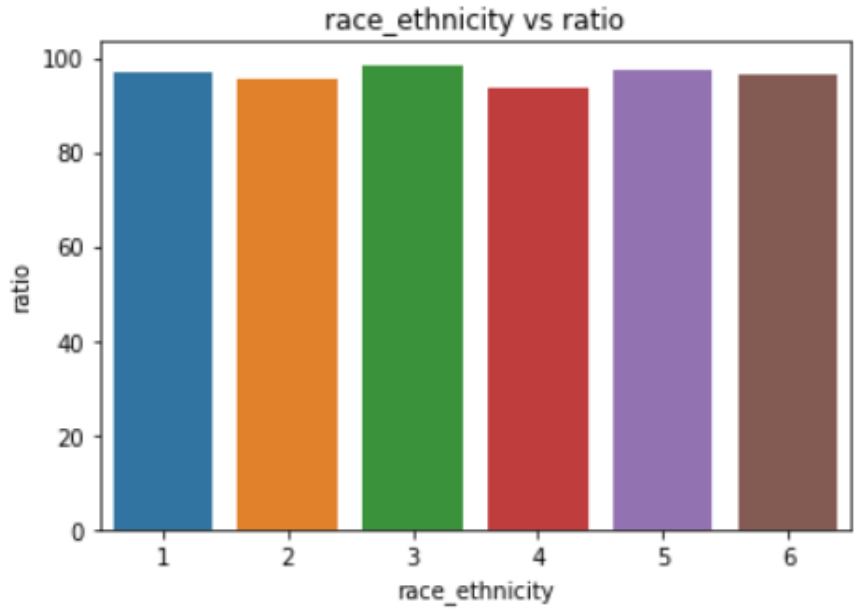
1	12-17
2	18-24
3	25-34
4	35-49
5	50-64
6	65 or older

- Visualizing the counts and plotting the frequency of the each age class show that class 6 which is 65 or older are the lowest and the samplest class among different age groups

****

14-

```
[ ]  ###realtion ship crime type
     level=list(ratio.to_frame().index)
     rate=[(count) for count in ratio]
     zipped = list(zip(level, rate))
     df = pd.DataFrame(zipped, columns=['race_ethnicity', 'ratio'])
     sns.barplot(x ='race_ethnicity', y ='ratio', data = df)
     # Show the plot
     plt.title("race_ethnicity vs ratio")
     plt.show()
```



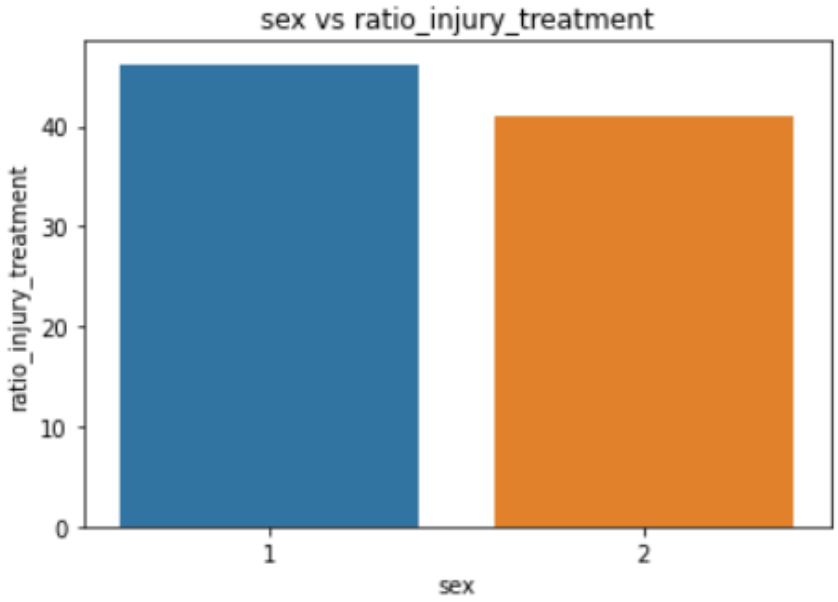
1	Non-Hispanic white
2	Non-Hispanic black
3	Non-Hispanic American Indian/ Alaska Native
4	Non-Hispanic Asian/ Native Hawaiian/ Other Pacific Islander
5	Non-Hispanic more than one race
6	Hispanic

- Visualizing the counts and plotting the frequency of the each age class show that class 3 which is Non Hispanic American Indian and Alaska Native are the highest and the greater class among different race ethnicity groups, although all are similar

****

15-

```
###realtion ship crime type
level=list(ratio.to_frame().index)
rate=[(count) for count in ratio]
zipped = list(zip(level, rate))
df = pd.DataFrame(zipped, columns=['sex', 'ratio_injury_treatment'])
sns.barplot(x ='sex', y ='ratio_injury_treatment', data = df)
# Show the plot
plt.title("sex vs ratio_injury_treatment")
plt.show()
```

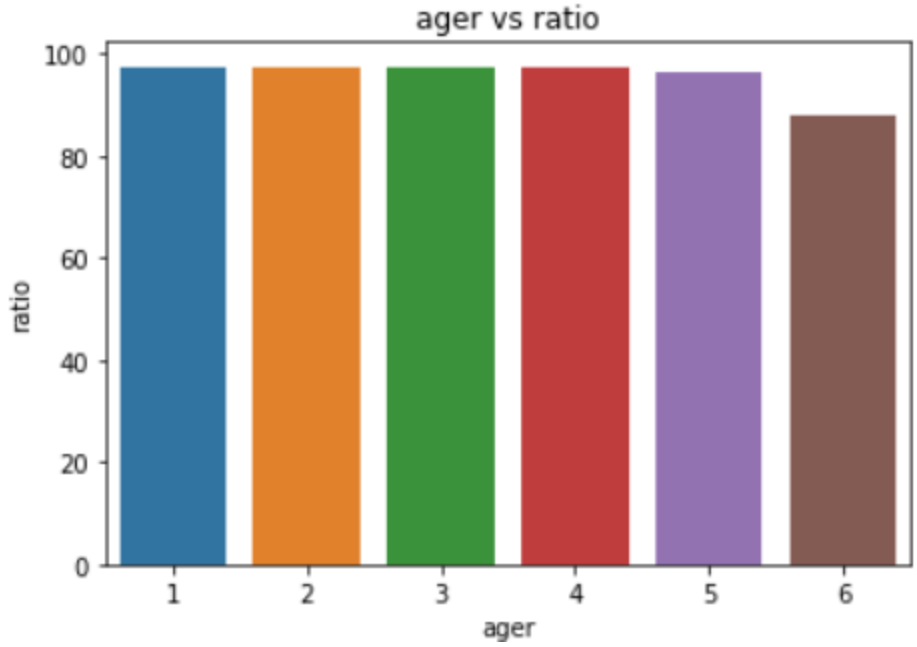


- Visualizing the counts and plotting the frequency of the each age class show that class 1 which is male are the highest and the greater class among different sex groups

****

16-

```
]  ###realtion ship crime type
   level=list(ratio.to_frame().index)
   rate=[(count) for count in ratio]
   zipped = list(zip(level, rate))
   df = pd.DataFrame(zipped, columns=['ager', 'ratio'])
   sns.barplot(x ='ager', y ='ratio', data = df)
   # Show the plot
   plt.title("ager vs ratio")
   plt.show()
```



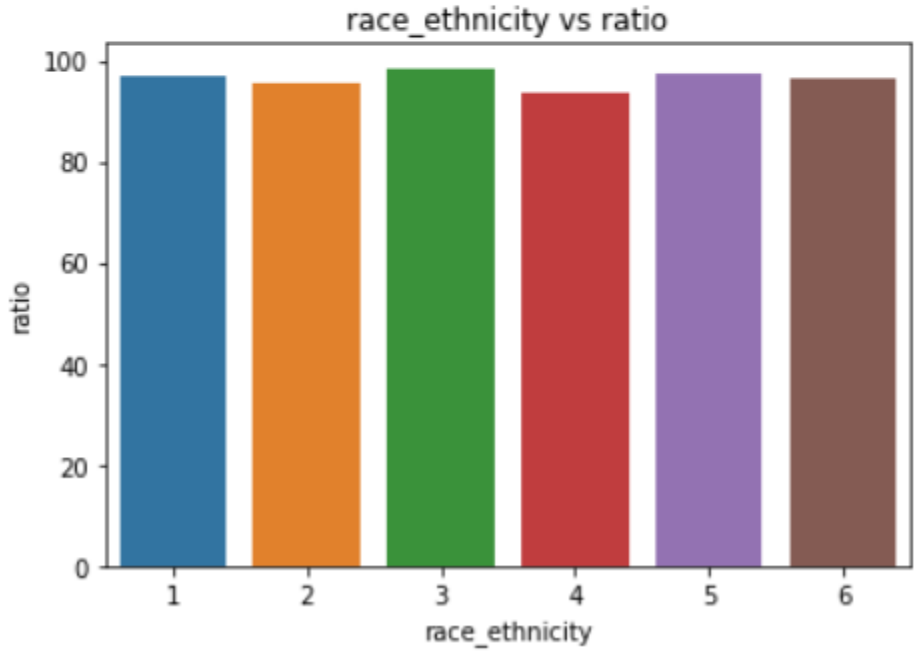
1	12-17
2	18-24
3	25-34
4	35-49
5	50-46
6	65:

- Visualizing the counts and plotting the frequency of the each age class show that class 6 which is white are the lowest and the samplest class among different age groups

****

17-

```
###realtion ship crime type
level=list(ratio.to_frame().index)
rate=[(count) for count in ratio]
zipped = list(zip(level, rate))
df = pd.DataFrame(zipped, columns=['race_ethnicity', 'ratio'])
sns.barplot(x ='race_ethnicity', y ='ratio', data = df)
# Show the plot
plt.title("race_ethnicity vs ratio")
plt.show()
```



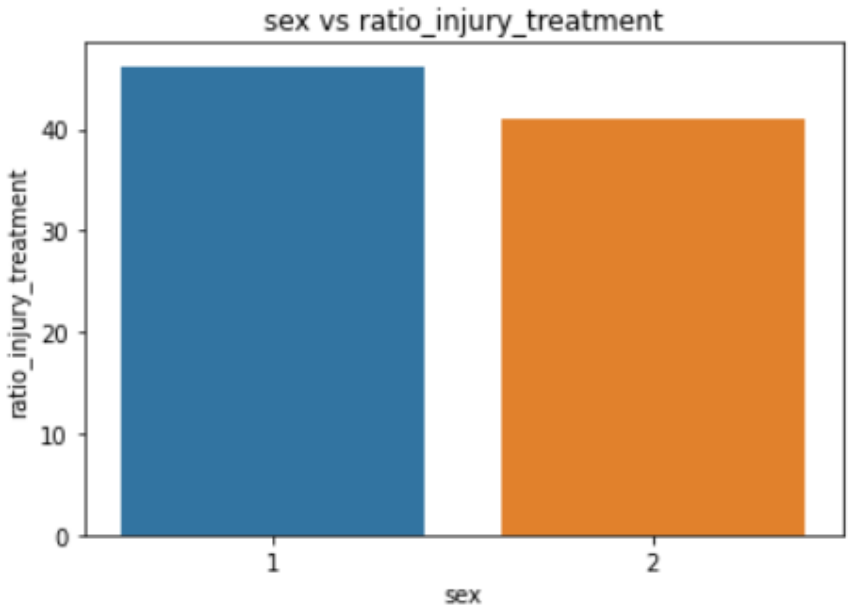
1	Non-Hispanic white
2	Non-Hispanic black
3	Non-Hispanic American Indian/ Alaska Native
4	Non-Hispanic Asian/ Native Hawaiian/ Other Pacific Islander
5	Non-Hispanic more than one race
6	Hispanic

- Visualizing the counts and plotting the frequency of the each age class show that class 3 which is Non Hispanic American Indian and Alaska Native are the highest and the greater class among different race ethnicity groups, although all are similar

****

18-

```
###realtion ship crime type
level=list(ratio.to_frame().index)
rate=[(count) for count in ratio]
zipped = list(zip(level, rate))
df = pd.DataFrame(zipped, columns=['sex', 'ratio_injury_treatment'])
sns.barplot(x ='sex', y ='ratio_injury_treatment', data = df)
# Show the plot
plt.title("sex vs ratio_injury_treatment")
plt.show()
```

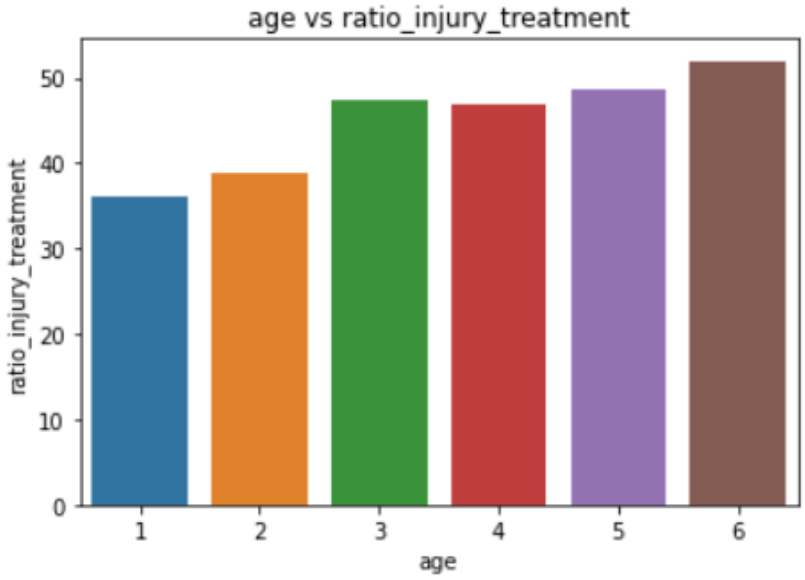


1	Male
2	Female

****

19-

```
###realtion ship crime type
level=list(ratio.to_frame().index)
rate=[(count) for count in ratio]
zipped = list(zip(level, rate))
df = pd.DataFrame(zipped, columns=['age', 'ratio_injury_treatment'])
sns.barplot(x ='age', y ='ratio_injury_treatment', data = df)
# Show the plot
plt.title("age vs ratio_injury_treatment")
plt.show()
```



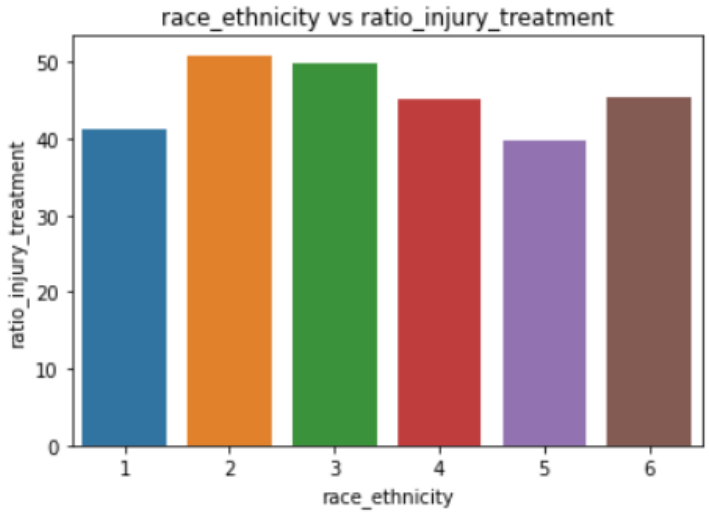
1	12-17
2	18-24
3	25-34
4	35-49
5	50-46
6	65:

- Visualizing the counts and plotting the frequency of the each age class show that class 6 which is 65 or older are the highest and the greater class among different age ratio groups

****

20-

```
###realtion ship crime type
level=list(ratio.to_frame().index)
rate=[(count) for count in ratio]
zipped = list(zip(level, rate))
df = pd.DataFrame(zipped, columns=['race_ethnicity', 'ratio_injury_treatment'])
sns.barplot(x ='race_ethnicity', y ='ratio_injury_treatment', data = df)
# Show the plot
plt.title("race_ethnicity vs ratio_injury_treatment")
plt.show()
```



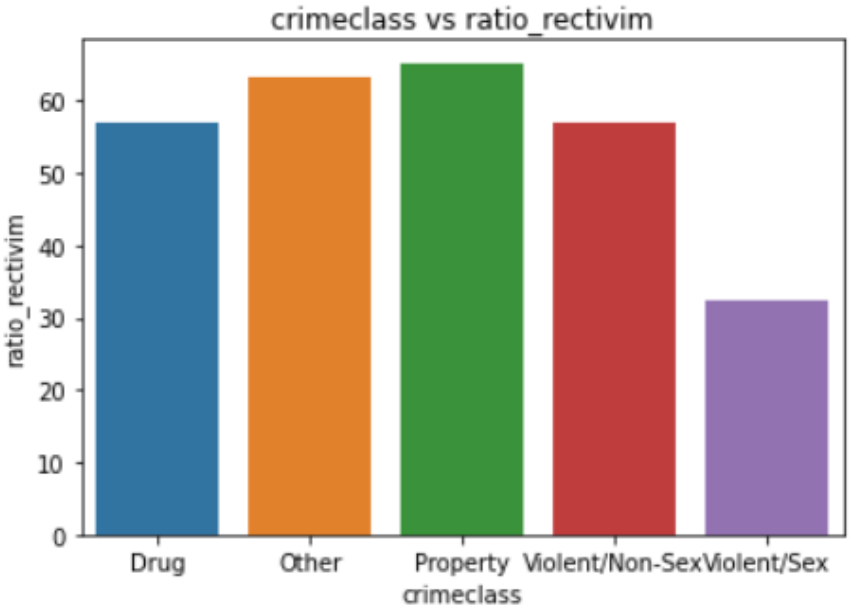
1	Non-Hispanic white
2	Non-Hispanic black
3	Non-Hispanic American Indian/ Alaska Native
4	Non-Hispanic Asian/ Native Hawaiian/ Other Pacific Islander
5	Non-Hispanic more than one race
6	Hispanic

- Visualizing the counts and plotting the frequency of the each age class show that class 2 which is Non Hispanic black are the highest and the greater class among different race ethnicity groups, although all are similar

****

21-

```
###realtion ship crime type
level=list(ratio.to_frame().index)
rate=[(count) for count in ratio]
zipped = list(zip(level, rate))
df = pd.DataFrame(zipped, columns=['crimeclass', 'ratio_rectivim'])
sns.barplot(x ='crimeclass', y ='ratio_rectivim', data = df)
# Show the plot
plt.title("crimeclass vs ratio_rectivim")
plt.show()
```

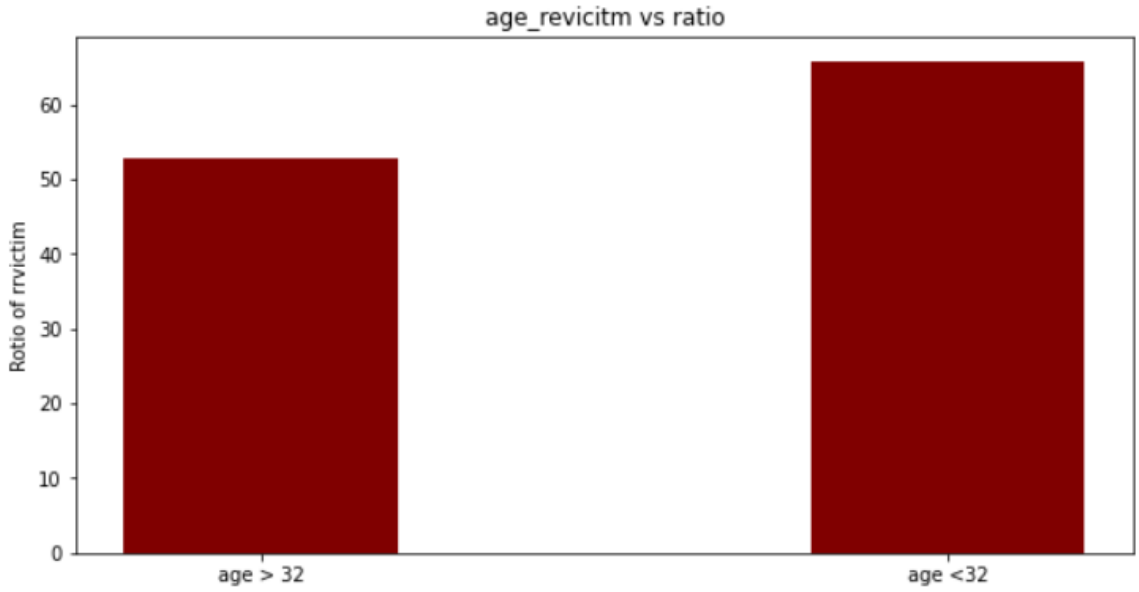


- Visualizing the counts and plotting the frequency of the each age class show that class Property Crime class are the highest and the greater class among different race ethnicity groups, although all are similar

****

22-

```
import numpy as np
import matplotlib.pyplot as plt
ratio=[ratio_more32,ratio_less32]
age=["age > 32","age <32"]
fig = plt.figure(figsize = (10, 5))
plt.bar(age, ratio, color ='maroon',
        width = 0.4)
plt.ylabel("Rotio of rrvictim")
plt.title("age_revicitm vs ratio")
plt.show()
```



- Visualizing the counts and plotting the frequency of the each age class show that class of ages less than is the highest and the greater class among different age groups, although all are similar

****

Bonus Task:

Steps:

- Data cleaning as mentioned
- Using `RandomForestClassifier model`
- Dropping NAN and missing values
- Finding correlation matrix
- Dropping unnecessary columns

- Splitting arrays or matrices into random train and test subsets
- 70 % training dataset and 30 % test datasets
- Creating a RF classifier
- Performing predictions on the test dataset
    - **ACCURACY OF THE MODEL: 0.7141848976711362**

Data:

```
df_firearm=pd.read_csv('/content/drive/MyDrive/stats_proj/Firearm.csv')
df_firearm
```

	state	year	felony	invcommitment	invoutpatient	danger	drugmisdemeanor	alctreatment	alcoholism	relinquishment	...	expartedating	dvrosurrender	dvrosurrendernoconditions
0	Alabama	1991	0	0	0	0	0	0	1	0	...	0	0	0
1	Alabama	1992	0	0	0	0	0	0	1	0	...	0	0	0
2	Alabama	1993	0	0	0	0	0	0	1	0	...	0	0	0
3	Alabama	1994	0	0	0	0	0	0	1	0	...	0	0	0
4	Alabama	1995	0	0	0	0	0	0	1	0	...	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1495	Wyoming	2016	1	0	0	0	0	0	0	0	...	0	0	0
1496	Wyoming	2017	1	0	0	0	0	0	0	0	...	0	0	0
1497	Wyoming	2018	1	0	0	0	0	0	0	0	...	0	0	0
1498	Wyoming	2019	1	0	0	0	0	0	0	0	...	0	0	0
1499	Wyoming	2020	1	0	0	0	0	0	0	0	...	0	0	0

1500 rows × 137 columns

Json file:

```
[ ]  df_NIBRS=pd.read_json('/content/drive/MyDrive/stats_proj/FBI_Crime.json')
     df_NIBRS
```

	state	offense	response
0	HI	aggravated-assault	{'results': [{'count': 1364, 'data_year': 2018...
1	DE	aggravated-assault	{'results': [{'count': 3415, 'data_year': 2001...
2	PR	aggravated-assault	{'results': [], 'pagination': {'count': 0, 'pa...
3	TX	aggravated-assault	{'results': [{'count': 900, 'data_year': 1997}...
4	MA	aggravated-assault	{'results': [{'count': 54, 'data_year': 1994},...
...	...	...	...
4099	PA	all-offenses	{'results': [{'count': 6, 'data_year': 2013}, ...
4100	CT	all-offenses	{'results': [{'count': 77, 'data_year': 1998},...
4101	LA	all-offenses	{'results': [{'count': 537, 'data_year': 2003}...
4102	TN	all-offenses	{'results': [{'count': 394, 'data_year': 1997}...
4103	DC	all-offenses	{'results': [{'count': 91, 'data_year': 2000},...

4104 rows × 3 columns

Correlation among features

Conclusion:

Visualizing the counts and plotting the frequency of the each age class show that:

class of assault offensive are the highest and the greater class among different crimes groups the frequency of the each gender

We note that the number of males is greater than the number of females

35-49 are the highest and the greater class among different age groups

white are the highest and the greater class among different race ethnicity groups
male are the highest and the greater class among different offendersex groups
30 or older are the highest and the greater class among different offenderage groups

Non- Hispanic white are the highest and the greater class among different race ethnicity groups

College are the highest and the greater class among different education level groups

we note that the highest level is class 7 which is $75,000 or more

we note that the highest level is class 4 which is Simple assault
Strangers are the highest and the greater class among different crime relation groups

Intimates and Other relatives are the highest and the greater class among different relation groups

65 or older are the lowest and the samplest class among different age groups

class of ages less than is the highest and the greater class among different age groups, although all are similar

ACCURACY OF THE MODEL: 0.7141848976711362