KAUNAS UNIVERSITY OF TECHNOLOGY

INFORMATICS FACULTY

INTRODUCTION TO ARTIFICIAL INTELLIGENCE DATA ANALYSIS LAB WORK REPORT

Report author: Uysal Demirci

Lecturer: doc.dr. Germanas Budnikas

Kaunas, 2024

1. Select (create) a dataset to perform this and other laboratory works. Your choice must be approved by the tutor.

Selected Dataset: Guns

Link: https://vincentarelbundock.github.io/Rdatasets/csv/AER/Guns.csv

Description: Guns is a balanced panel of data on 50 US states, plus the District of Columbia (for a total of 51 states), by year for 1977–1999.

Format: A data frame containing 1,173 observations on 13 variables.

state factor indicating state.

year factor indicating year.

violent violent crime rate (incidents per 100,000 members of the population).

murder murder rate (incidents per 100,000).

robbery robbery rate (incidents per 100,000).

prisoners incarceration rate in the state in the previous year (sentenced prisoners per 100,000 residents; value for the previous year).

afam percent of state population that is African-American, ages 10 to 64.

cauc percent of state population that is Caucasian, ages 10 to 64.

male percent of state population that is male, ages 10 to 29.

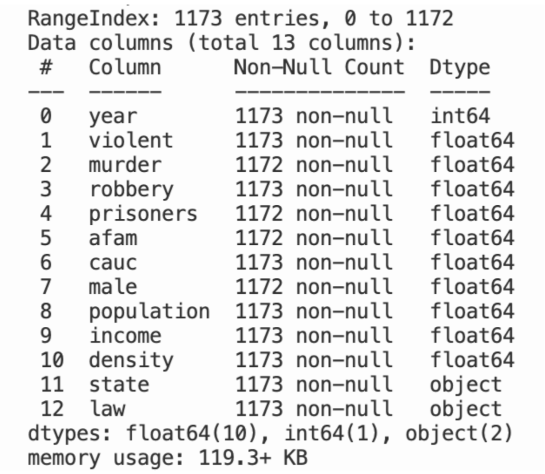
population state population, in millions of people.

income real per capita personal income in the state (US dollars).

density population per square mile of land area, divided by 1,000.

law factor. Does the state have a shall carry law in effect in that year?

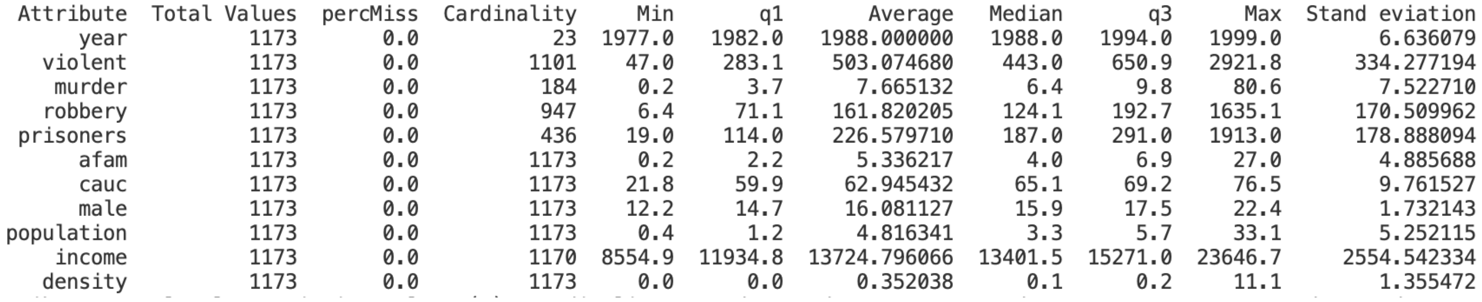
The columns have the following datatypes:



```
RangeIndex: 1173 entries, 0 to 1172
Data columns (total 13 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   year        1173 non-null   int64  
 1   violent     1173 non-null   float64
 2   murder      1172 non-null   float64
 3   robbery     1173 non-null   float64
 4   prisoners   1172 non-null   float64
 5   afam        1172 non-null   float64
 6   cauc        1173 non-null   float64
 7   male        1172 non-null   float64
 8   population  1173 non-null   float64
 9   income      1173 non-null   float64
 10  density     1173 non-null   float64
 11  state       1173 non-null   object 
 12  law         1173 non-null   object 
dtypes: float64(10), int64(1), object(2)
memory usage: 119.3+ KB
```

2. For each numeric type attribute calculate:

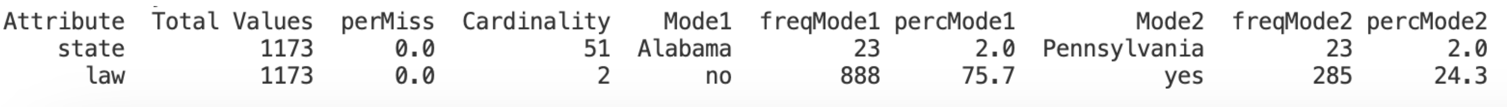
- total number of values,
- percentage of missing values,
- cardinality,
- minimum (min) and maximum (max) values,
- 1st and 3rd quartiles,
- average,
- median,
- Standard deviation.



```
  Attribute  Total Values  percMiss  Cardinality     Min       q1      Average   Median       q3      Max  Stand eviation
       year          1173       0.0           23  1977.0   1982.0  1988.000000   1988.0   1994.0   1999.0        6.636079
    violent          1173       0.0         1101    47.0    283.1   503.074680    443.0    650.9   2921.8      334.277194
     murder          1173       0.0          184     0.2      3.7     7.665132      6.4      9.8     80.6        7.522710
    robbery          1173       0.0          947     6.4     71.1   161.820205    124.1    192.7   1635.1      170.509962
  prisoners          1173       0.0          436    19.0    114.0   226.579710    187.0    291.0   1913.0      178.888094
       afam          1173       0.0         1173     0.2      2.2     5.336217      4.0      6.9     27.0        4.885688
       cauc          1173       0.0         1173    21.8     59.9    62.945432     65.1     69.2     76.5        9.761527
       male          1173       0.0         1173    12.2     14.7    16.081127     15.9     17.5     22.4        1.732143
 population          1173       0.0         1173     0.4      1.2     4.816341      3.3      5.7     33.1        5.252115
     income          1173       0.0         1170  8554.9  11934.8 13724.796066  13401.5  15271.0  23646.7     2554.542334
    density          1173       0.0         1173     0.0      0.0     0.352038      0.1      0.2     11.1        1.355472
```

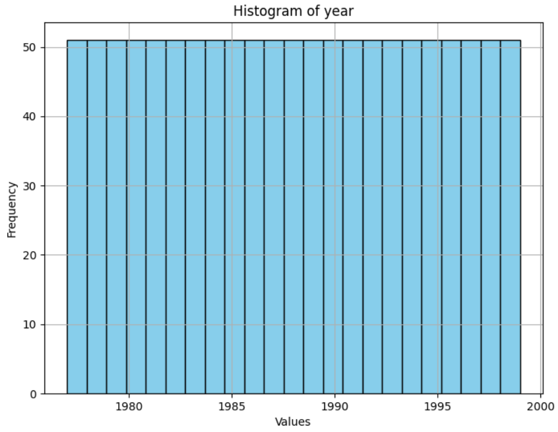
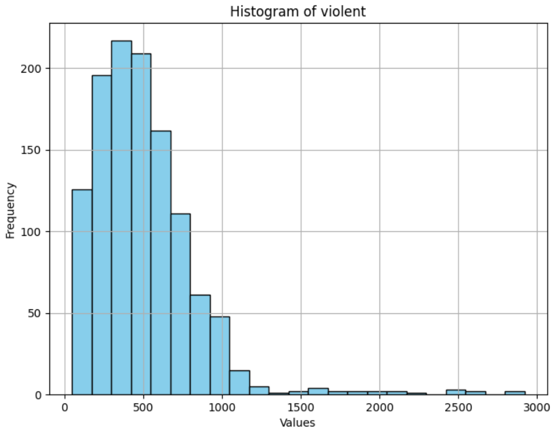
3. For each *category* type attribute calculate:

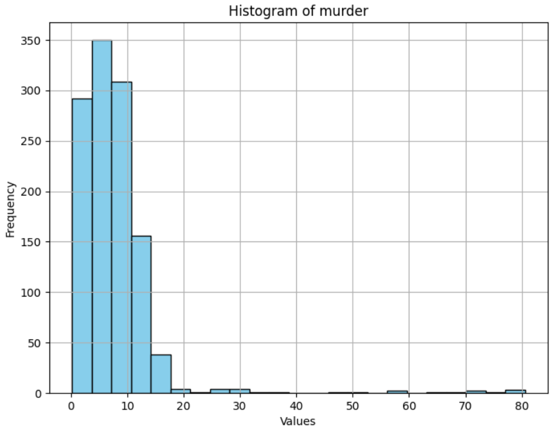
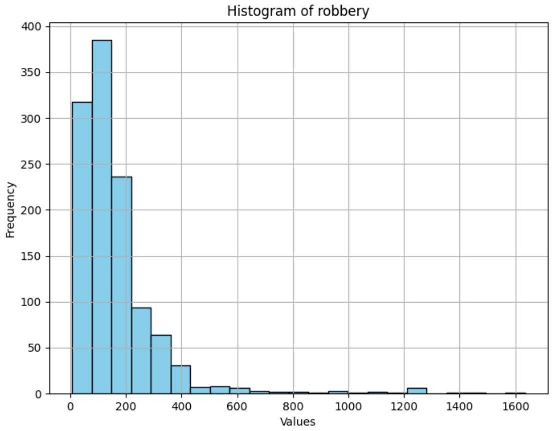
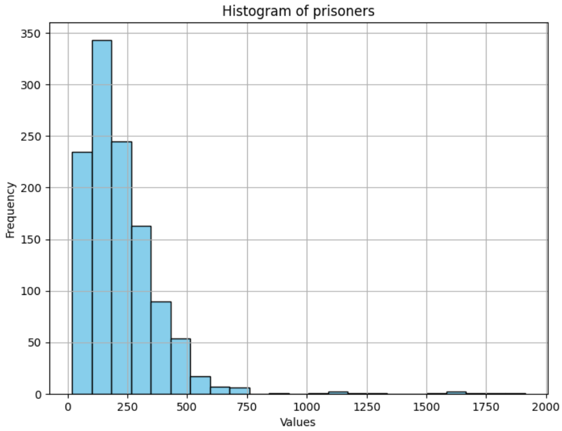
- total number of values,
- percentage of missing values,
- cardinality,
- mode,
- The frequency of the mode
- Percentage value of the mode
- Second mode value (mode 2),
- Frequency value for Mode 2,
- Percentage of Mode 2.

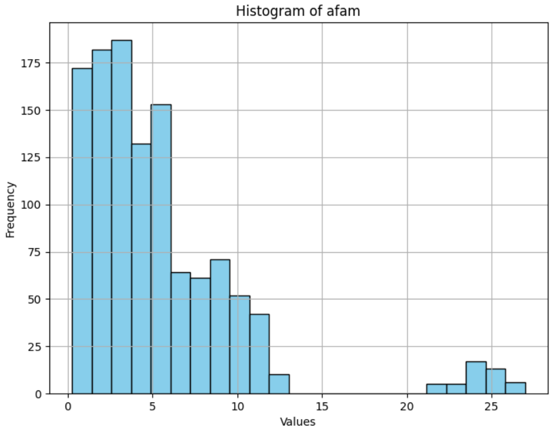
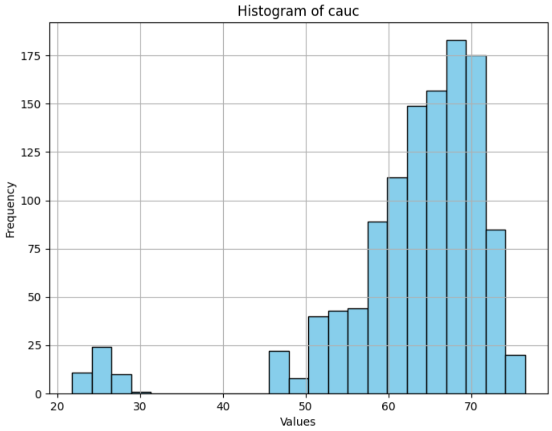
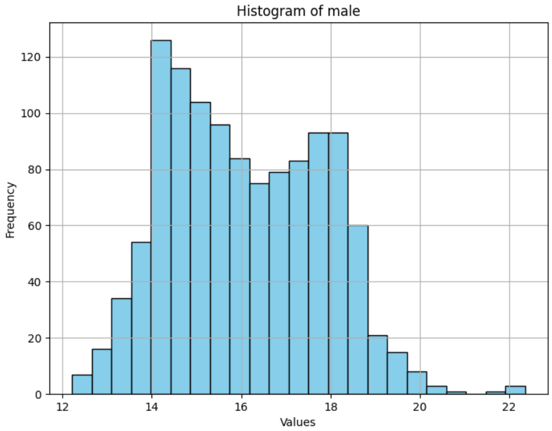


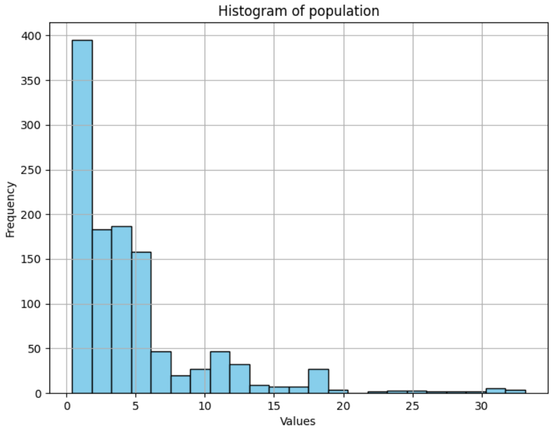
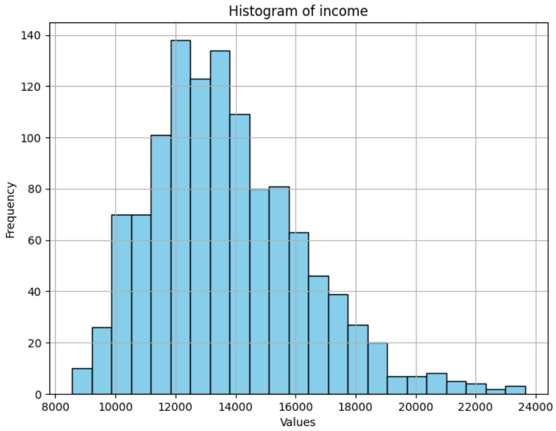
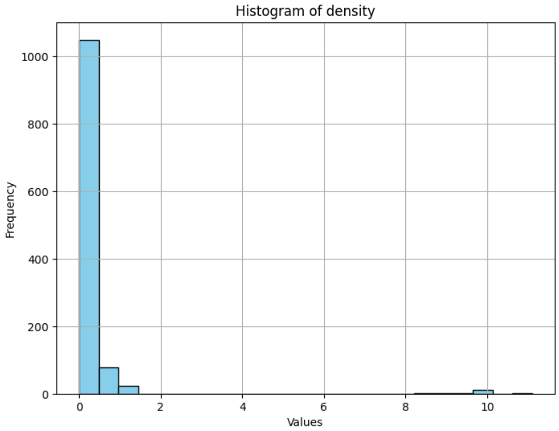
```
Attribute  Total Values  perMiss  Cardinality    Mode1  freqMode1 percMode1        Mode2  freqMode2 percMode2
    state          1173      0.0           51  Alabama         23       2.0  Pennsylvania        23       2.0
      law          1173      0.0            2       no        888      75.7          yes        285      24.3
```

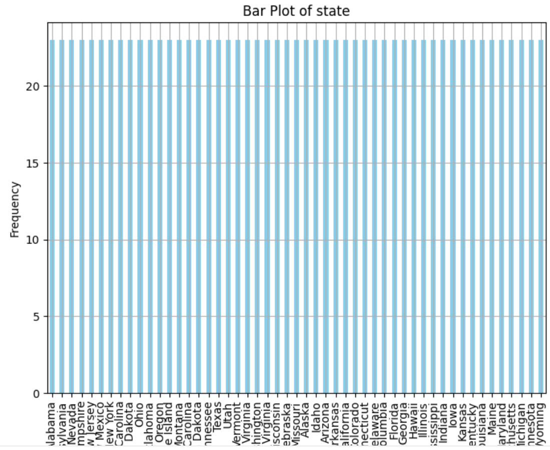
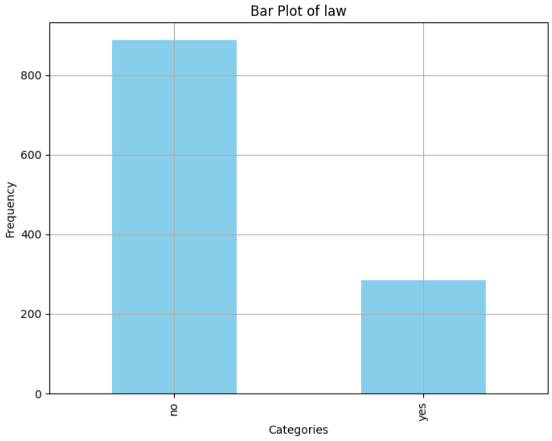
4. Draw histograms of attributes. Provide descriptions of the distribution (eg, normal, exponential, etc.) and what conclusions can be drawn from it.

Histogram	Description
 	Uniform Distribution
 	Skewed right

Histogram	Description
	Normal Distribution
	Normal Distribution. Outliers 600 and 1600 should be removed
	Normal Distribution. Outliers between 800 and 2000 should be removed.

Histogram	**Description**
	Skewed right
	Normal Distribution
	Normal Distribution

Histogram	Description
	Skewed right
	Normal Distribution
	Skewed right. Outliers between 1 and 12 should be removed.

Histogram	Description
	Uniform Distribution
	Only two categories, the rate of states without a carry law is higher than the rate of states with a carry law.

5. Identify data quality problems: missing values, cardinality problems, outliers. Provide a plan for resolving these issues, which will be implemented programmatically (e.g., missing values for a categorical attribute based on the attribute estimate of the mode, extreme values being removed or corrected).

There are no values missing. So there is to be done.

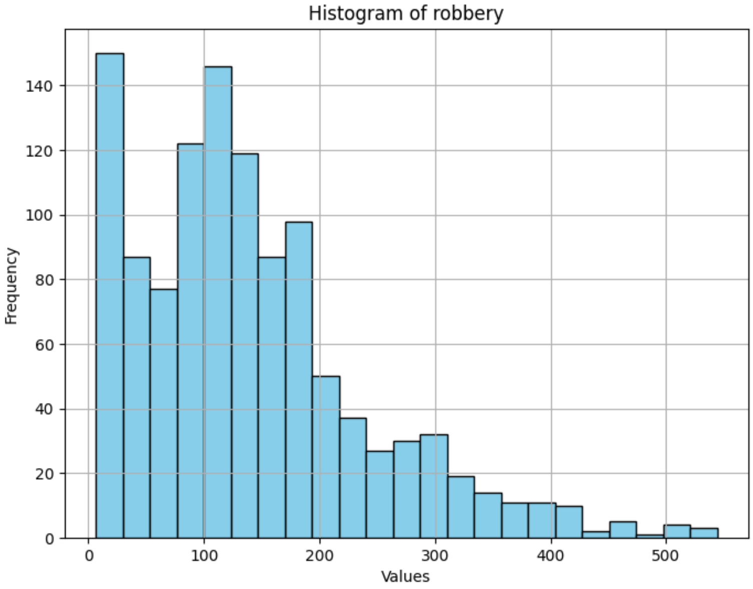
There are some outliers that should be removed, like I marked in the histograms above.

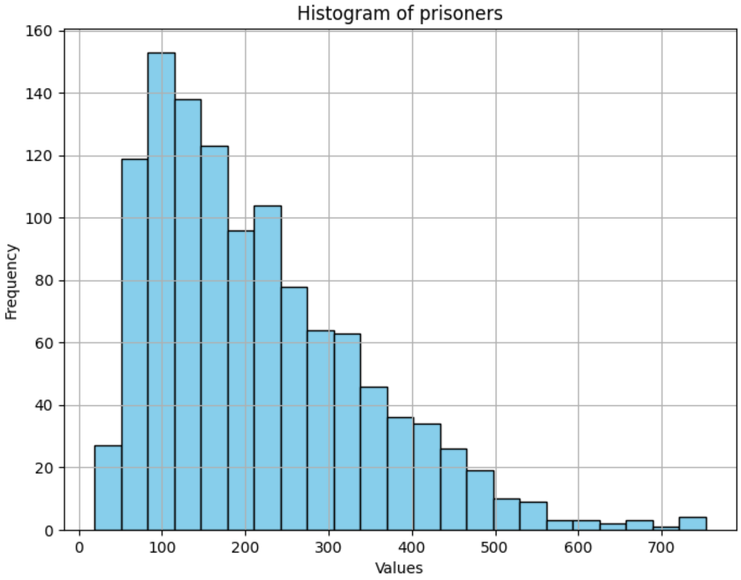
For histogram of robbery, I have removed the values higher than 600.

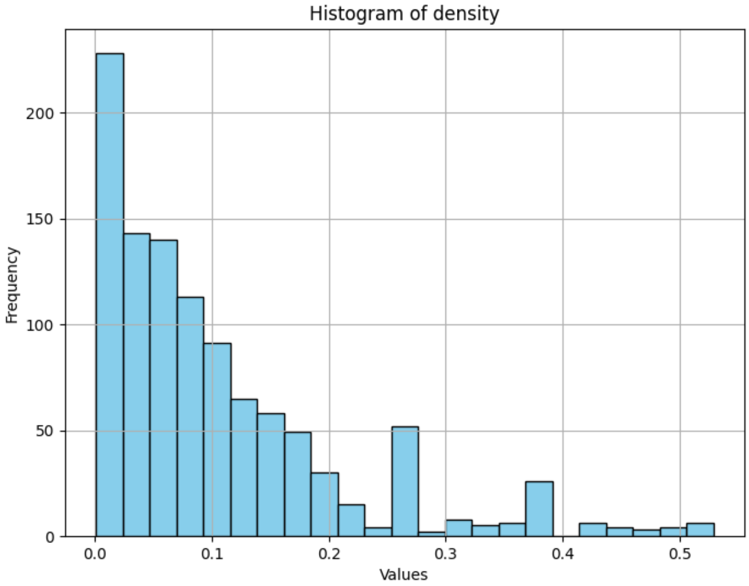
For histogram of prisoners, I have removed the values higher than 800.

In histogram of density, I made sure that only values up to 1 are shown on the graph.

After the all problems solved, the three histogram look like this:


Histogram of robbery
Frequency
0
20
40
60
80
100
120
140
0
100
200
300
400
500
Values

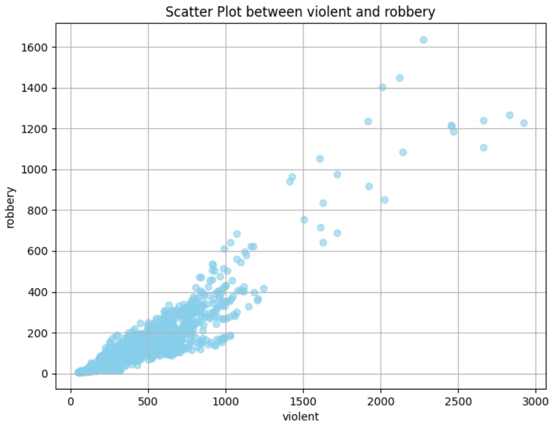
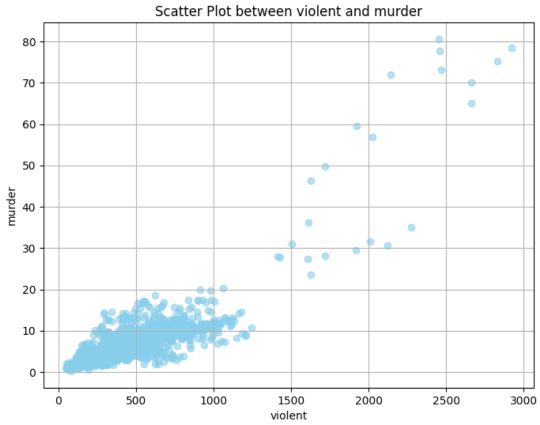

Histogram of prisoners
Frequency
0
20
40
60
80
100
120
140
160
0
100
200
300
400
500
600
700
Values

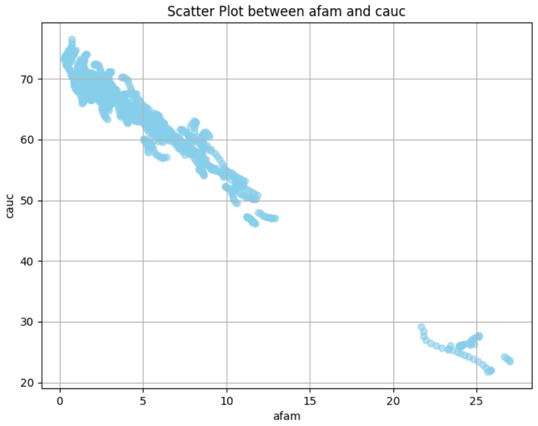

Histogram of density
Frequency
0
50
100
150
200
0.0
0.1
0.2
0.3
0.4
0.5
Values

6. Establish relationships between attributes using visualization techniques

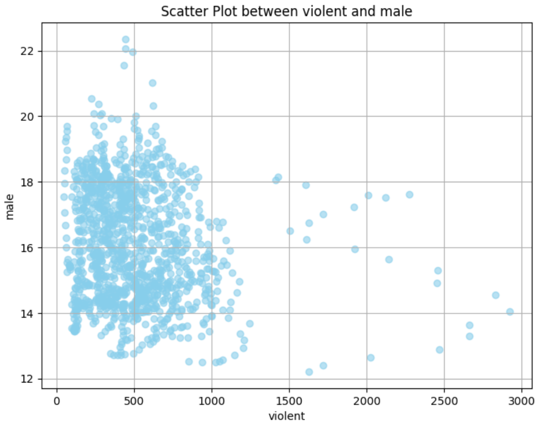
- **For numeric type attributes: Using a scatter plot type graph, provide multiple (2-3) examples with strong linear attribute dependency (direct or inverse correlation) and multiple examples with non-correlated (weakly correlated) attributes. Comment on results.**

- **Provide an SPLOM diagram (Scatter Plot Matrix).**

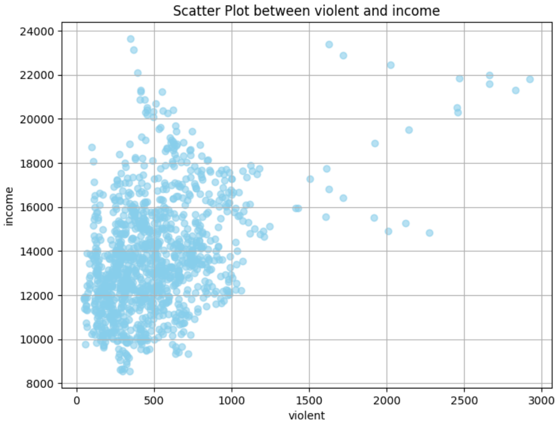
Scatter Plot	**Description**
 Scatter Plot between violent and robbery 	Strong Linear Correlation: As the number of violent incidents increases, the number of robberies has also increased.
 Scatter Plot between violent and murder 	Strong Linear Correlation: As the number of violent incidents increases, the number of murders has also increased.



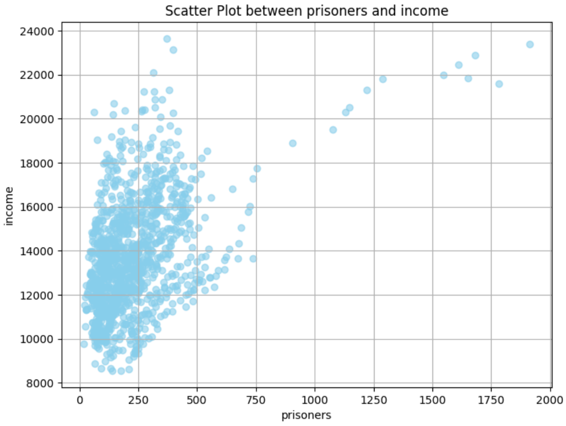
Strong Linear Correlation:
"As the percentage of the state population that is African-American, ages 10 to 64, increases, the percentage of the state population that is Caucasian, ages 10 to 64, has decreased."



No correlation between percent of state population that is male, ages 10 to 29 and real per capita personal income in the state (US dollars).

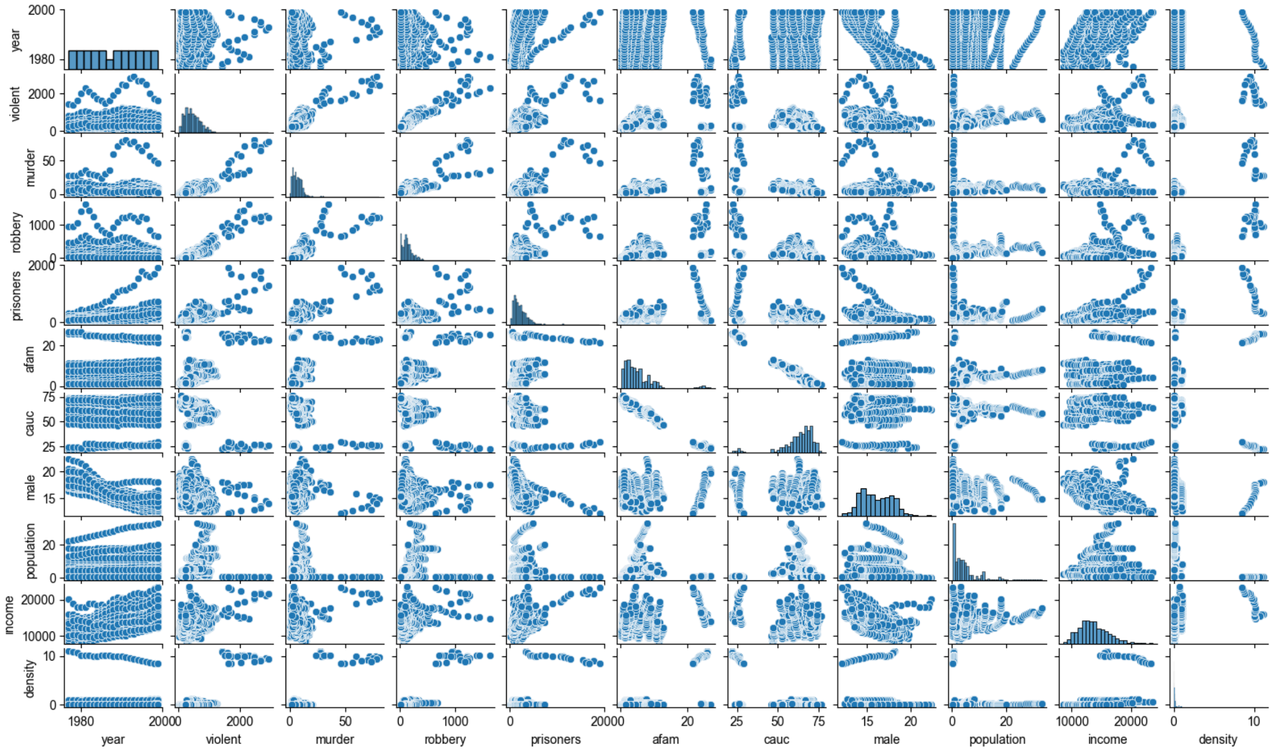


No correlation between violent crime rate (incidents per 100,000 members of the population) and real per capita personal income in the state (US dollars).

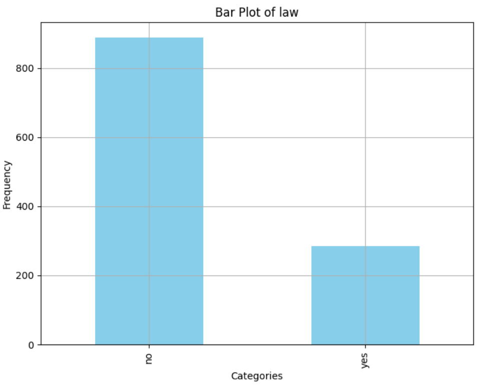


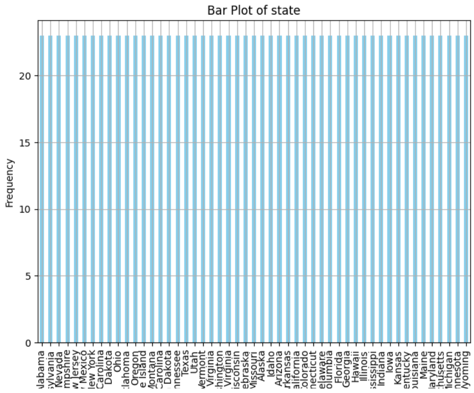
Linear correlation:
As the incarceration rate in the state in the previous year (sentenced prisoners per 100,000 residents; value for the previous year) increases, the real per capita personal income in the state (US dollars) has increased.

SPLOM-Diagram:

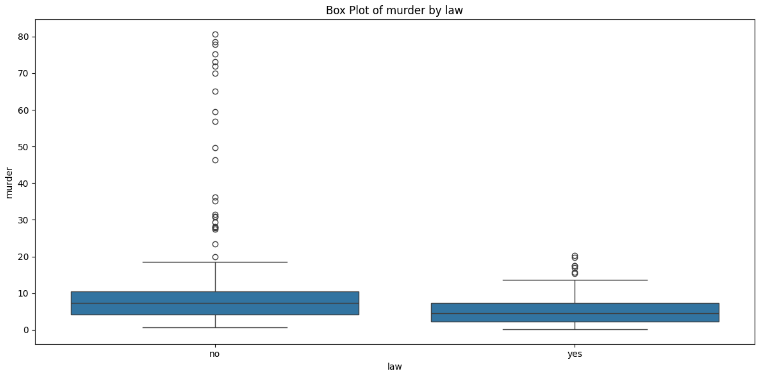
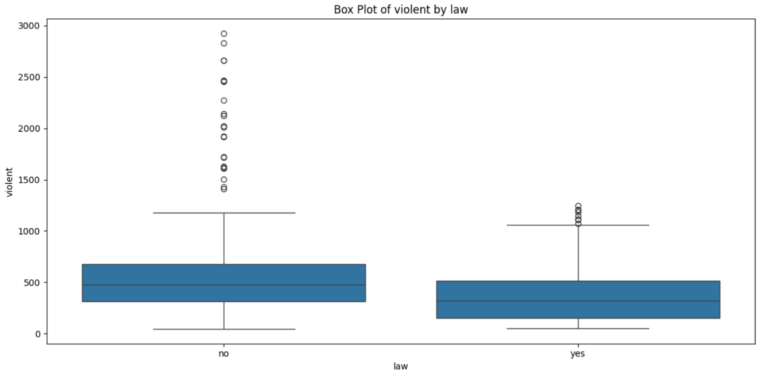


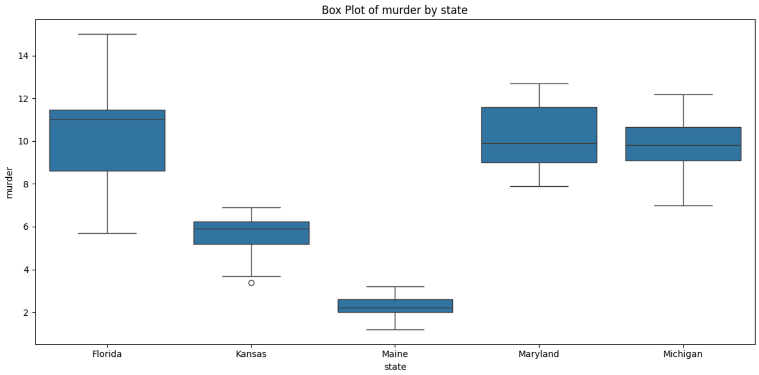
***For categorical attributes:* Using the bar plot type diagram, give some (2-3) examples of attribute frequency and comment on the results.**

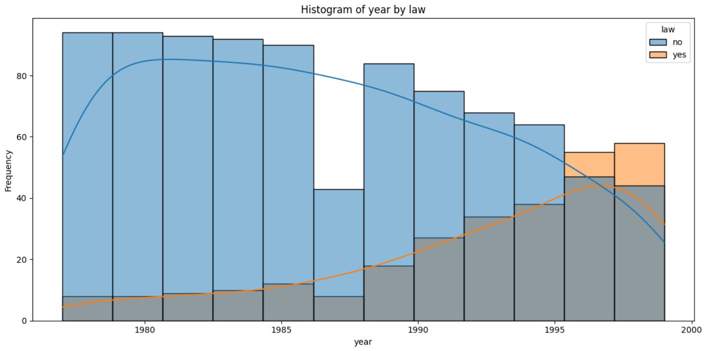
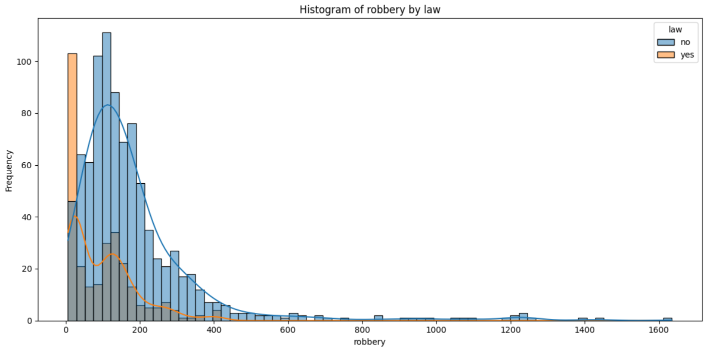
Bar Plot	**Legend**	**Comment**
	Yes The state have a shall carry law in effect in that year No The state doesn't have a shall carry law in effect in that year	Most of states don't have a shall carry law in effect in that year

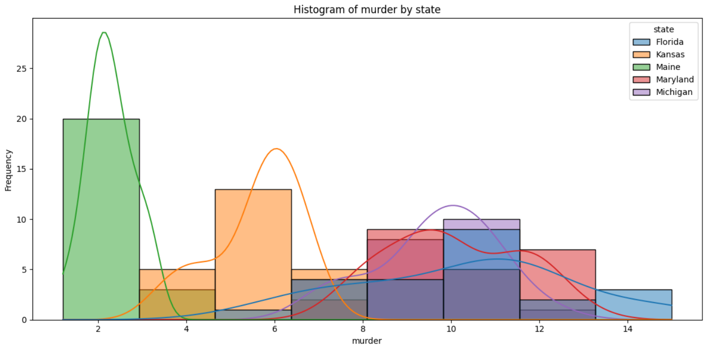
	Each state name indicates the state from which the study was conducted.	Equal amounts of data have been collected from all states.

Provide some (2-3) examples of histograms and box plot diagrams depicting relationships between categorical and numeric type variables.

Boxplot	**Description**
	In states don't have a shall carry law in effect in that year, the murder rate is higher compared to states that have a shall carry law in effect in that year
	In states don't have a shall carry law in effect in that year, the violent rate is higher compared to states that have a shall carry law in effect in that year

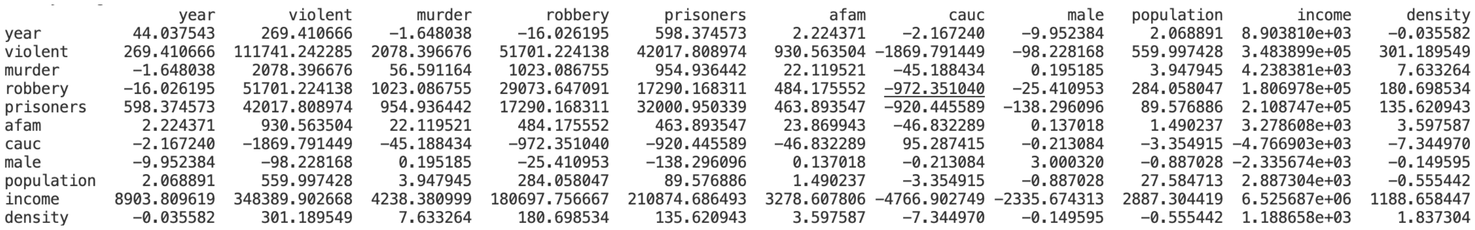
	In randomly selected 5 states, we do not observe the same murder rates. In some states, the murder rate is low.

Histogram	Description
	Over the years, states have started to implement shall carry laws.
	It has been observed that the rate of robbery is higher when there is no shall carry law in place.

	Among the randomly selected 5 states, it has been observed that Florida has the highest murder rate.

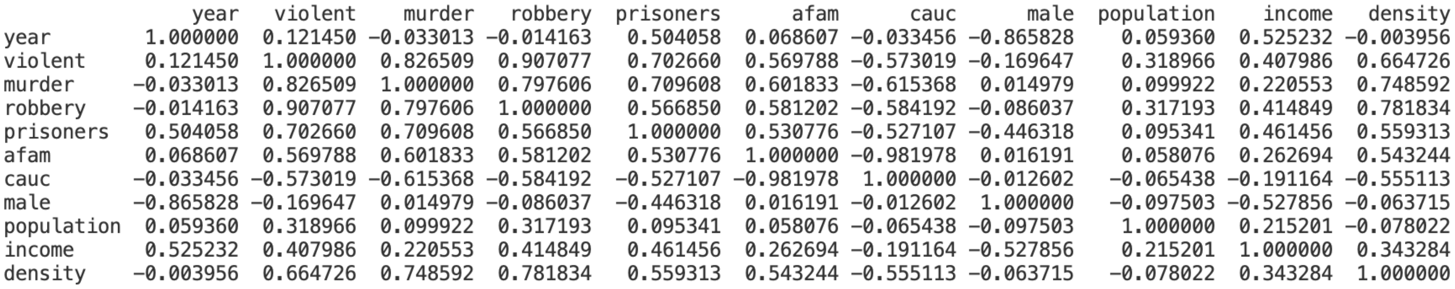
7. Calculate the covariance and correlation values between continuous attributes and graphically represent the correlation matrix. Comments on the results.

Covariance:

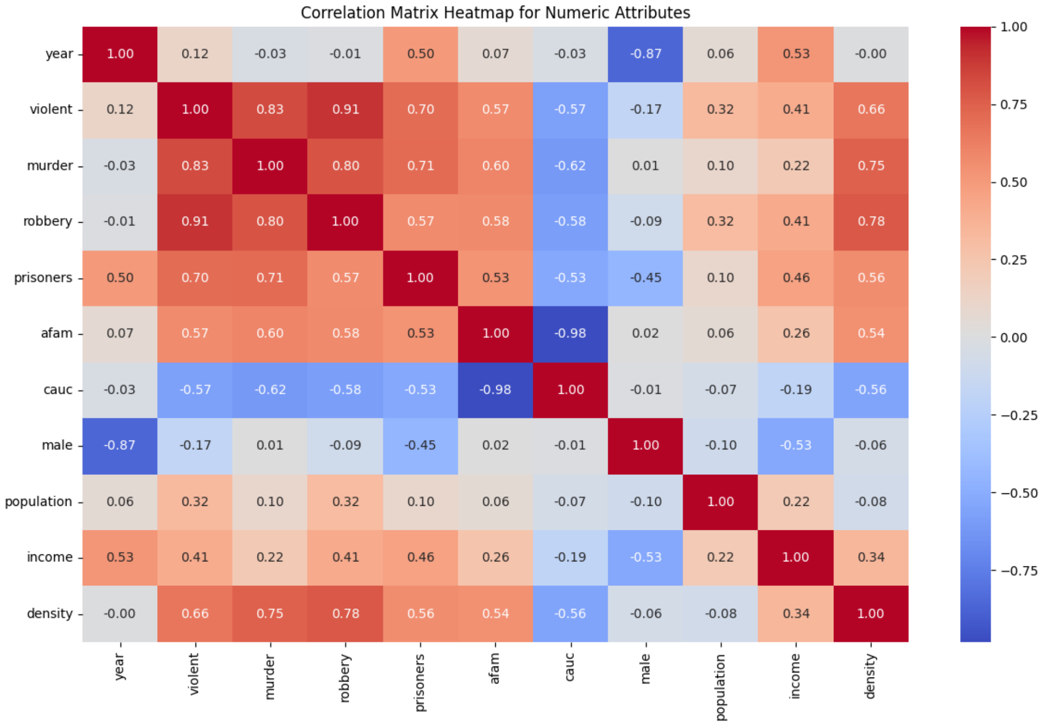


```
                  year         violent       murder        robbery      prisoners         afam         cauc         male  population        income     density
year         44.037543      269.410666    -1.648038     -16.026195     598.374573     2.224371    -2.167240    -9.952384    2.068891  8.903810e+03   -0.035582
violent     269.410666  111741.242285  2078.396676   51701.224138   42017.808974   930.563504 -1869.791449   -98.228168  559.997428  3.483899e+05  301.189549
murder       -1.648038     2078.396676    56.591164    1023.086755     954.936442    22.119521   -45.188434     0.195185    3.947945  4.238381e+03    7.633264
robbery     -16.026195    51701.224138  1023.086755   29073.647091   17290.168311   484.175552  -972.351040   -25.410953  284.058047  1.806978e+05  180.698534
prisoners   598.374573    42017.808974   954.936442   17290.168311   32000.950339   463.893547  -920.445589  -138.296096   89.576886  2.108747e+05  135.620943
afam          2.224371      930.563504    22.119521     484.175552     463.893547    23.869943   -46.832289     0.137018    1.490237  3.278608e+03    3.597587
cauc         -2.167240    -1869.791449   -45.188434    -972.351040    -920.445589   -46.832289    95.287415    -0.213084   -3.354915 -4.766903e+03   -7.344970
male         -9.952384      -98.228168     0.195185     -25.410953    -138.296096     0.137018    -0.213084     3.000320   -0.887028 -2.335674e+03   -0.149595
population    2.068891      559.997428     3.947945     284.058047      89.576886     1.490237    -3.354915    -0.887028   27.584713  2.887304e+03   -0.555442
income     8903.809619   348389.902668  4238.380999  180697.756667  210874.686493  3278.607806 -4766.902749 -2335.674313 2887.304419  6.525687e+06 1188.658447
density      -0.035582      301.189549     7.633264     180.698534     135.620943     3.597587    -7.344970    -0.149595   -0.555442  1.188658e+03    1.837304
```

Correlation:

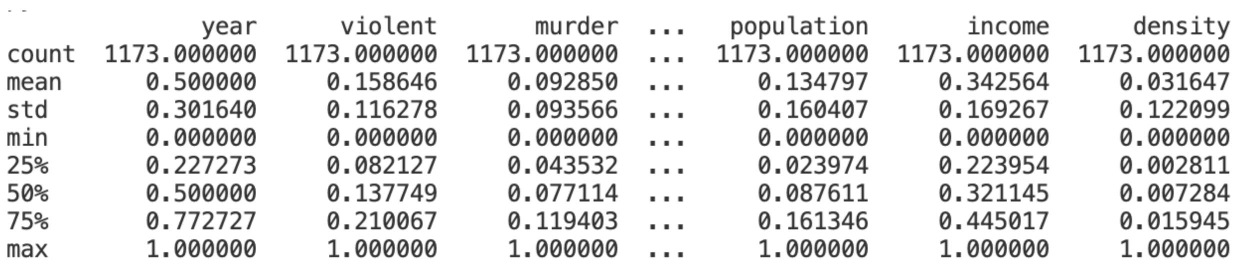


```
                year   violent    murder   robbery  prisoners      afam      cauc      male  population    income   density
year        1.000000  0.121450 -0.033013 -0.014163   0.504058  0.068607 -0.033456 -0.865828    0.059360  0.525232 -0.003956
violent     0.121450  1.000000  0.826509  0.907077   0.702660  0.569788 -0.573019 -0.169647    0.318966  0.407986  0.664726
murder     -0.033013  0.826509  1.000000  0.797606   0.709608  0.601833 -0.615368  0.014979    0.099922  0.220553  0.748592
robbery    -0.014163  0.907077  0.797606  1.000000   0.566850  0.581202 -0.584192 -0.086037    0.317193  0.414849  0.781834
prisoners   0.504058  0.702660  0.709608  0.566850   1.000000  0.530776 -0.527107 -0.446318    0.095341  0.461456  0.559313
afam        0.068607  0.569788  0.601833  0.581202   0.530776  1.000000 -0.981978  0.016191    0.058076  0.262694  0.543244
cauc       -0.033456 -0.573019 -0.615368 -0.584192  -0.527107 -0.981978  1.000000 -0.012602   -0.065438 -0.191164 -0.555113
male       -0.865828 -0.169647  0.014979 -0.086037  -0.446318  0.016191 -0.012602  1.000000   -0.097503 -0.527856 -0.063715
population  0.059360  0.318966  0.099922  0.317193   0.095341  0.058076 -0.065438 -0.097503    1.000000  0.215201 -0.078022
income      0.525232  0.407986  0.220553  0.414849   0.461456  0.262694 -0.191164 -0.527856    0.215201  1.000000  0.343284
density    -0.003956  0.664726  0.748592  0.781834   0.559313  0.543244 -0.555113 -0.063715   -0.078022  0.343284  1.000000
```



8. Perform data normalization.

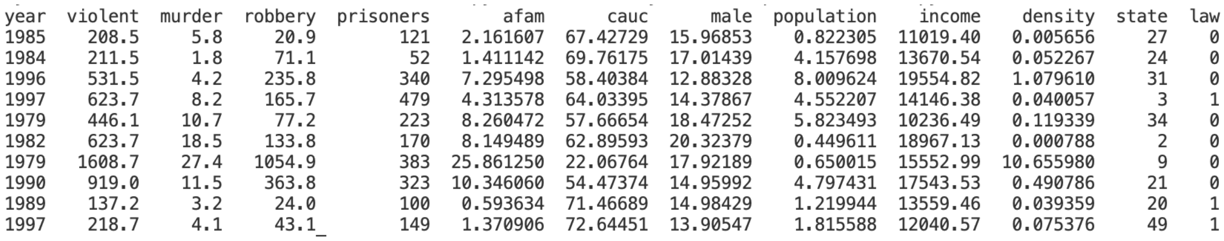
I converted all values in each column to values between 0 and 1.



```
              year      violent       murder  ...   population       income      density
count  1173.000000  1173.000000  1173.000000  ...  1173.000000  1173.000000  1173.000000
mean      0.500000     0.158646     0.092850  ...     0.134797     0.342564     0.031647
std       0.301640     0.116278     0.093566  ...     0.160407     0.169267     0.122099
min       0.000000     0.000000     0.000000  ...     0.000000     0.000000     0.000000
25%       0.227273     0.082127     0.043532  ...     0.023974     0.223954     0.002811
50%       0.500000     0.137749     0.077114  ...     0.087611     0.321145     0.007284
75%       0.772727     0.210067     0.119403  ...     0.161346     0.445017     0.015945
max       1.000000     1.000000     1.000000  ...     1.000000     1.000000     1.000000
```

9. Convert categorical variables to numeric type variables.

I converted to categorical 'law' variable to numeric using 1 for yes and 0 for no. For to convert categorical 'state' attributes to numeric, I gave a number to each state from starting 1.



```
year  violent  murder  robbery  prisoners       afam      cauc      male  population    income    density  state  law
1985    208.5     5.8     20.9        121   2.161607  67.42729  15.96853    0.822305  11019.40   0.005656     27    0
1984    211.5     1.8     71.1         52   1.411142  69.76175  17.01439    4.157698  13670.54   0.052267     24    0
1996    531.5     4.2    235.8        340   7.295498  58.40384  12.88328    8.009624  19554.82   1.079610     31    0
1997    623.7     8.2    165.7        479   4.313578  64.03395  14.37867    4.552207  14146.38   0.040057      3    1
1979    446.1    10.7     77.2        223   8.260472  57.66654  18.47252    5.823493  10236.49   0.119339     34    0
1982    623.7    18.5    133.8        170   8.149489  62.89593  20.32379    0.449611  18967.13   0.000788      2    0
1979   1608.7    27.4   1054.9        383  25.861250  22.06764  17.92189    0.650015  15552.99  10.655980      9    0
1990    919.0    11.5    363.8        323  10.346060  54.47374  14.95992    4.797431  17543.53   0.490786     21    0
1989    137.2     3.2     24.0        100   0.593634  71.46689  14.98429    1.219944  13559.46   0.039359     20    1
1997    218.7     4.1     43.1_       149   1.370906  72.64451  13.90547    1.815588  12040.57   0.075376     49    1
```