Suvo_Ganguli_Assignment_1.1

January 13, 2024

1 Assignment 1.1

Name: Subhabrata (Suvo) Ganguli

Date: Jan 13, 2024

For this assignment, you will refer to the textbook to solve the practice exercises. **Use Python to answer any coding problems (not R, even if indicated in your textbook).** Use Jupyter Notebook, Google Colab, or a similar software program to complete your assignment. Submit your answers as a **PDF or HTML** file. As a best practice, always label your axes and provide titles for any graphs generated on this assignment. Round all quantitative answers to 2 decimal places.

1.1 Problem # 1.1.

In the 2018 election for Senate in California, a CNN exit poll of 1882 voters stated that 52.5% voted for the Democratic candidate, Diane Feinstein. Of all 11.1 million voters, 54.2% voted for Feinstein.

(a) What was the (i) subject, (ii) sample, (iii) population?

(i) Subject: Voting results of Democratic Candidate, Diana Feinstein
(ii) Sample: 1882 voters from the CNN exit poll
(iii) Population: 11.1 million voters

1.2 Problem # 1.2.

The `Students` data file at http://stat4ds.rwth-aachen.de/data/Students.dat responses of a class of 60 social science graduate students at the University of Florida to a questionnaire that asked about *gender* (1 = female, 0 = male), *age*, *hsgpa* = high school GPA (on a four-point scale), cogpa = college GPA, *dhome* = distance (in miles) of the campus from your home town, *dres* = distance (in miles) of the classroom from your current residence, *tv* = average number of hours per week that you watch TV, *sport* = average number of hours per week that you participate in sports or have other physical exercise, *news* = number of times a week you read a newspaper, *aids* = number of people you know who have died from AIDS or who are HIV+, *veg* = whether you are a vegetarian (1 = yes, 0 = no), *affil* = political affiliation (1 = Democrat, 2 = Republican, 3 = independent), *ideol* = political ideology (1 = very liberal, 2 = liberal, 3 = slightly liberal, 4 = moderate, 5 = slightly conservative, 6 = conservative, 7 = very conservative), *relig* = how often you attend religious services (0 = never, 1 = occasionally, 2 = most weeks, 3 = every week), *abor* = opinion about whether abortion should be legal in the first three months of pregnancy (1 = yes, 0 = no), *affirm* = support affirmative action (1 = yes, 0 = no), and *life* = belief in life after death (1 = yes, 2 = no, 3 = undecided). You will use this data file for some exercises in this book.

(a) Practice accessing a data file for statistical analysis with your software by going to the book's website and copying and then displaying this data file.

(b) Using responses on *abor*, state a question that could be addressed with (i) descriptive statistics, (ii) inferential statistics.

'(a) Code and outputs are shown below:

[1]:
```
# Import necessary libraries
import numpy as np
import pandas as pd

# Read in the Students data file
students = pd.read_csv('Students.dat', sep='\s+')

# Show first five rows of the data
students.head()
```

[1]:

	subject	gender	age	hsgpa	cogpa	dhome	dres	tv	sport	news	aids	\
0	1	0	32	2.2	3.5	0	5.0	3.0	5	0	0	
1	2	1	23	2.1	3.5	1200	0.3	15.0	7	5	6	
2	3	1	27	3.3	3.0	1300	1.5	0.0	4	3	0	
3	4	1	35	3.5	3.2	1500	8.0	5.0	5	6	3	
4	5	0	23	3.1	3.5	1600	10.0	6.0	6	3	0	

	veg	affil	ideol	relig	abor	affirm	life
0	0	2	6	2	0	0	1
1	1	1	2	1	1	1	3
2	1	1	2	2	1	1	3
3	0	3	4	1	1	1	2
4	0	3	1	0	1	0	2

(b)

(c) Descriptive Statistics: Find the number of students who are for or against whether abortion should be legal in the first three months of pregnancy. Use a bargraph for the data.

(ii) Inferential Statistics: Estimate the number of students in the overall population who are for or against whether abortion should be legal in the first three months of pregnancy (from the sample statistics).

1.3 Problem # 1.3.

Identify each of the following variables as categorical or quantitative: (a) Number of smartphones that you own; (b) County of residence; (c) Choice of diet (vegetarian, nonvegetarian); (d) Distance, in kilometers, commute to work

(a) Number of smartphones that you own: Quantitative

(b) County of residence: Categorical

(c) Choice of diet (vegetarian, nonvegetarian): Categorical

(d) Distance, in kilometers, commute to work: Quantitative

1.4 Problem # 1.4.

Give an example of a variable that is (a) categorical; (b) quantitative; (c) discrete; (d) continuous

(a) Categorical: County of residence

(b) Quantative: BMI of a sample of population

(c) Discrete: Number of cell phones owned by individuals

(d) Continuous: Distance, in kilometers, commute to work

1.5 Problem # 1.10.

Analyze the `Carbon_West` (http://stat4ds.rwth-aachen.de/data/Carbon_West.dat) data file at the book's website by **(a)** constructing a frequency distribution and a histogram, **(b)** finding the mean, median, and standard deviation. Interpret each.

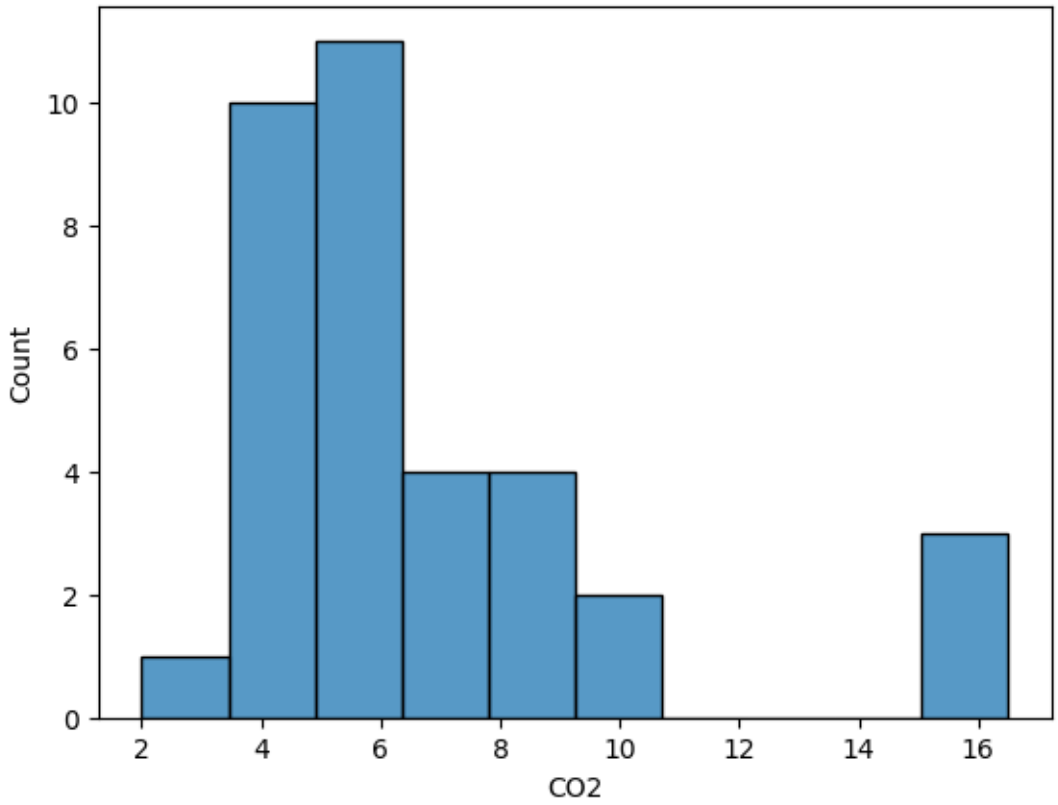
(a) Code and outputs are shown below:

[2]:
```
# Import seaborn library
import seaborn as sns

# Read in the data file
data = pd.read_csv('Carbon_West.dat', sep='\s+')

# Show data
print(data.head())
print(' ')

# Frequency distribution
print('Frequency Distribution:')
df = data['CO2'].value_counts()
print(df)

# Histogram
sns.histplot(x = data.CO2, bins = 10)
```

```
      Nation   CO2
0    Albania   2.0
1  Australia  15.4
2    Austria   6.9
3    Belgium   8.3
4     Bosnia   6.2

Frequency Distribution:
CO2
4.3     3
6.2     3
```

```
5.3     3
3.5     2
5.9     2
6.5     1
4.5     1
5.0     1
5.7     1
9.3     1
7.7     1
9.9     1
3.6     1
5.4     1
4.4     1
2.0     1
7.3     1
15.4    1
8.9     1
4.6     1
8.7     1
9.2     1
4.0     1
15.1    1
8.3     1
6.9     1
16.5    1
Name: count, dtype: int64
```

[2]: <Axes: xlabel='CO2', ylabel='Count'>



[3]:
```
# Data description

data.describe()
```

[3]:
```
              CO2
count  35.000000
mean    6.717143
std     3.356949
min     2.000000
25%     4.450000
50%     5.900000
75%     8.000000
max    16.500000
```

For the C02 data of 35 countries, the mean CO2 emmision is 6.72, the median is 5.9 and the standard deviation is 3.36. Since, the mean is greater than the median, the distribution can be right-skewed. This can be confirmed by the histogram plot above.

1.6 Problem # 1.11.

According to Statistics Canada, for the Canadian population having income in 2019, annual income had a median of $35,000 and mean of $46,700. What would you predict about the shape of the

distribution? Why?

Since the mean is greater the median, the histogram will be right-skewed. This is because a larger portion of the data is concentrated towards the right tail based on the mean data.

1.7 Problem # 1.13.

A report indicates that public school teacher's annual salaries in New York city have an approximate mean of $69,000 and standard deviation of $6,000. If the distribution has approximately a bell shape, report intervals that contain about (a) 68%, (b) 95%, (c) all or nearly all salaries. Would a salary of $100,000 be unusual? Why?

Since the mean (mu) is $69,000 and standard deviation is $6,000: * Approximately 68% of the data will be within mu +/- sigma = [$ 63,000, $ 75,000] * Approximately 95% of the data will be within mu+/- 2*sigma = [$ 57,000, $ 81,000]* Approximately all of the data will be within mu +/- 3*sigma = [$ 51,000, $ 87,000]

A salary of $ 100,000 will be unusual since it does not fall within the 3-sigma bounds.

1.8 Problem # 1.17.

From the `Murder` data file (http://stat4ds.rwth-aachen.de/data/Murder.dat) at the book's website, use the variable murder, which is the murder rate (per 100,000 population) for each state in the U.S. in 2017 according to the FBI Uniform Crime Reports. At first, do not use the observation for D.C. (DC). Using software:
(a) Find the mean and standard deviation and interpret their values.
(b) Find the five-number summary, and construct the corresponding box plot. Interpret.
(c) Now include the observation for D.C. What is affected more by this outlier: The mean or the median? The range or the inter-quartile range?

Answer:

[4]:
```
# Read in the data file
data = pd.read_csv('Murder.dat', sep='\s+')
data_noDC = data.drop(data[data['state'] == 'DC'].index)

# Show data
print('Data without DC row:')
print(data_noDC)
print(' ')

# Mean without DC data
mean_noDC = data_noDC['murder'].mean().round(2)
print('Mean (without DC) = ' + str(mean_noDC))

# Standard Deviation without DC data
stddev_noDC = data_noDC['murder'].std().round(2)
print('Std. Dev. (without DC) = ' + str(stddev_noDC))
```

```
Data without DC row:
   state  murder
```

0	AK	8.4
1	AL	8.3
2	AR	8.6
3	AZ	5.9
4	CA	4.6
5	CO	3.9
6	CT	2.8
7	DE	5.6
8	FL	5.0
9	GA	6.7
10	HI	2.7
11	ID	1.9
12	IL	7.8
13	IN	6.0
14	IO	3.3
15	KS	5.5
16	KY	5.9
17	LA	12.4
18	MA	2.5
19	MD	9.0
20	ME	1.7
21	MI	5.7
22	MN	2.0
23	MO	9.8
24	MS	8.2
25	MT	3.9
26	NC	5.8
27	ND	1.3
28	NE	2.2
29	NH	1.0
30	NJ	3.6
31	NM	7.1
32	NV	2.3
33	NY	2.8
34	OH	6.1
35	OK	6.2
36	OR	2.5
37	PA	5.8
38	RI	1.9
39	SC	7.8
40	SD	2.9
41	TN	7.8
42	TX	5.0
43	UT	2.4
44	VA	5.3
45	VT	2.2
46	WA	3.1
47	WI	3.2

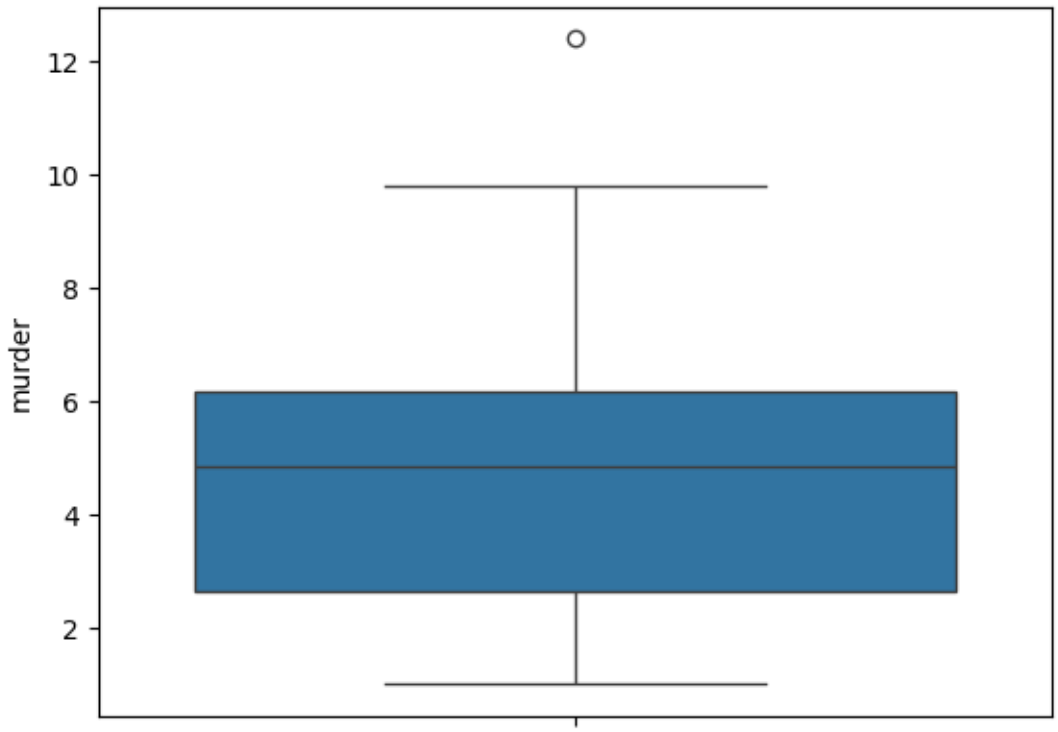
```
48    WV     4.7
49    WY     2.6

Mean (without DC) = 4.87
Std. Dev. (without DC) = 2.59
```

(a) The mean and standard deviation presents a description of the data if the data is normally distributed. However, without looking at the actual data using boxplot or histogram, it cannot be determined if the normal distribution is a good representation of the data or not.

[5]:
```
# Five number summary without DC data
print(data_noDC.describe().loc[['min', '25%', '50%', '75%', 'max']])

# Boxplot
sns.boxplot(data = data_noDC['murder'])
```

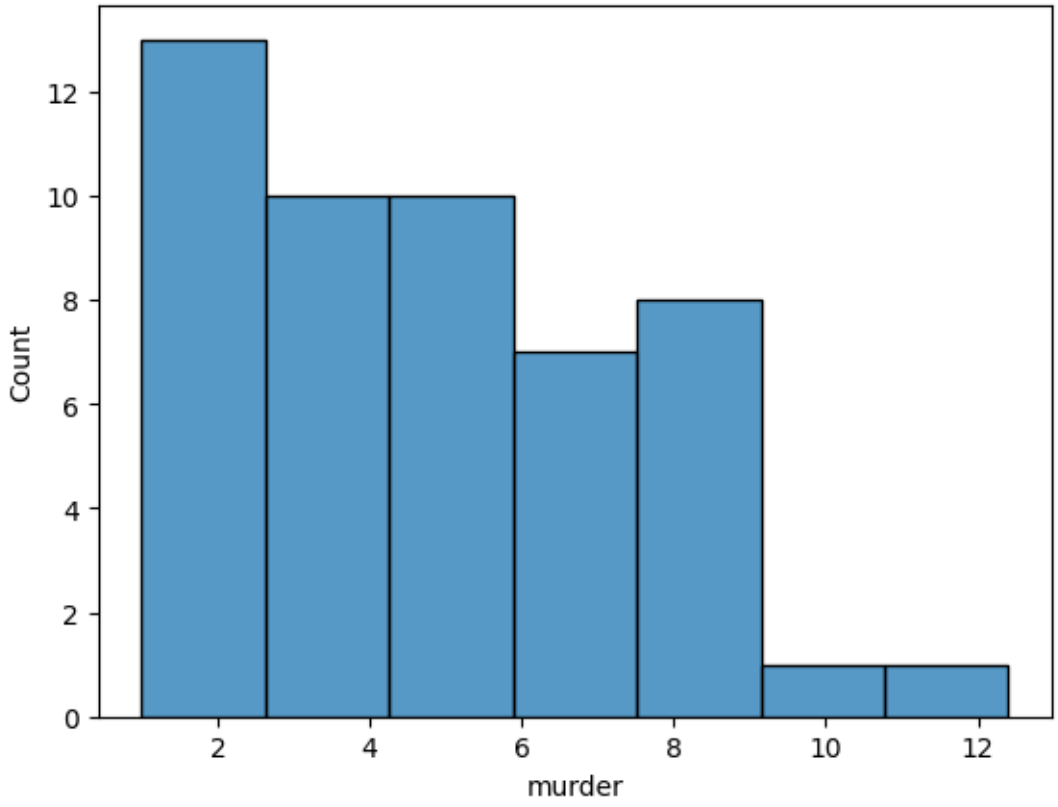
```
     murder
min   1.000
25%   2.625
50%   4.850
75%   6.175
max  12.400
```

[5]: <Axes: ylabel='murder'>



[6]:
```
# Histogram
sns.histplot(data = data_noDC, x = 'murder')
```

[6]:
```
<Axes: xlabel='murder', ylabel='Count'>
```



(b)The boxplot can be used to determine if the data is normally distributed, left-skewed or right-skewed.

Since the boxplot is not symmetric, we can conclude that the data is not normally distributed. In addition, since the median is towards the upper limit of the box, it can lead to a conclusion that the data is left-skewed. But if we look at the whiskers, we see that the maximum (in the top) is further out than the minimum value (one in the bottom). That can lead to the conclusion that the data is right-skewed. These are contradicting interpretations.

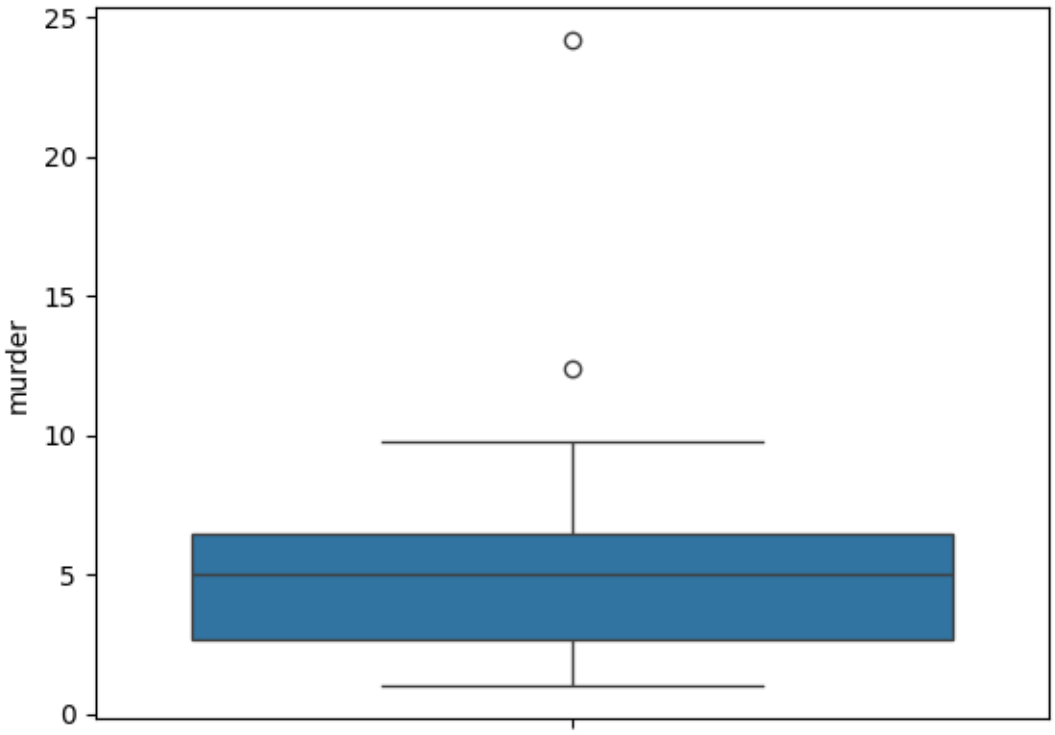
So, without looking at the actual histogram of the data, it cannot be interpreted from the boxplot if the data is left-skewed or right-skewed. The histogram verifies that the data is right-skewed.

[7]:
```
# Mean
mean = data['murder'].mean().round(2)
print('Mean (with DC) = ' + str(mean))

# Standard Deviation
```

```
stddev = data['murder'].std().round(2)
print('Std. Dev. (with DC) = ' + str(stddev))

# Five number summary with DC data
print(data.describe().loc[['min', '25%', '50%', '75%', 'max']])
sns.boxplot(data = data['murder'])
```

```
Mean (with DC) = 5.25
Std. Dev. (with DC) = 3.73
     murder
min    1.00
25%    2.65
50%    5.00
75%    6.45
max   24.20
```

[7]: <Axes: ylabel='murder'>



(c) With addition of the DC row:

- The mean changes from 4.87 to 5.25 (increase of 0.38)
- The median changes from 4.85 to 5.00 (increase of 0.15)

Thus the mean changes more than the median towards the right. This is expected as the DC data is an outlier on the right side of the distribution, causing it to be further right-skewed.

1.9 Problem # 1.18.

The `Income` data file (http://stat4ds.rwth-aachen.de/data/Income.dat) at the book's website reports annual income values in the U.S., in thousands of dollars.

(a) Using software, construct a histogram. Describe its shape.

(b) Find descriptive statistics to summarize the data. Interpret them.

(c) The kernel density estimation method finds a smooth-curve approximation for a histogram. At each value, it takes into account how many observations are nearby and their distance, with more weight given those closer. Increasing the bandwidth increases the influence of observations further away. Plot a smooth-curve approximation for the histogram of income values. Summarize the impact of increasing and of decreasing the bandwidth substantially from the default value.

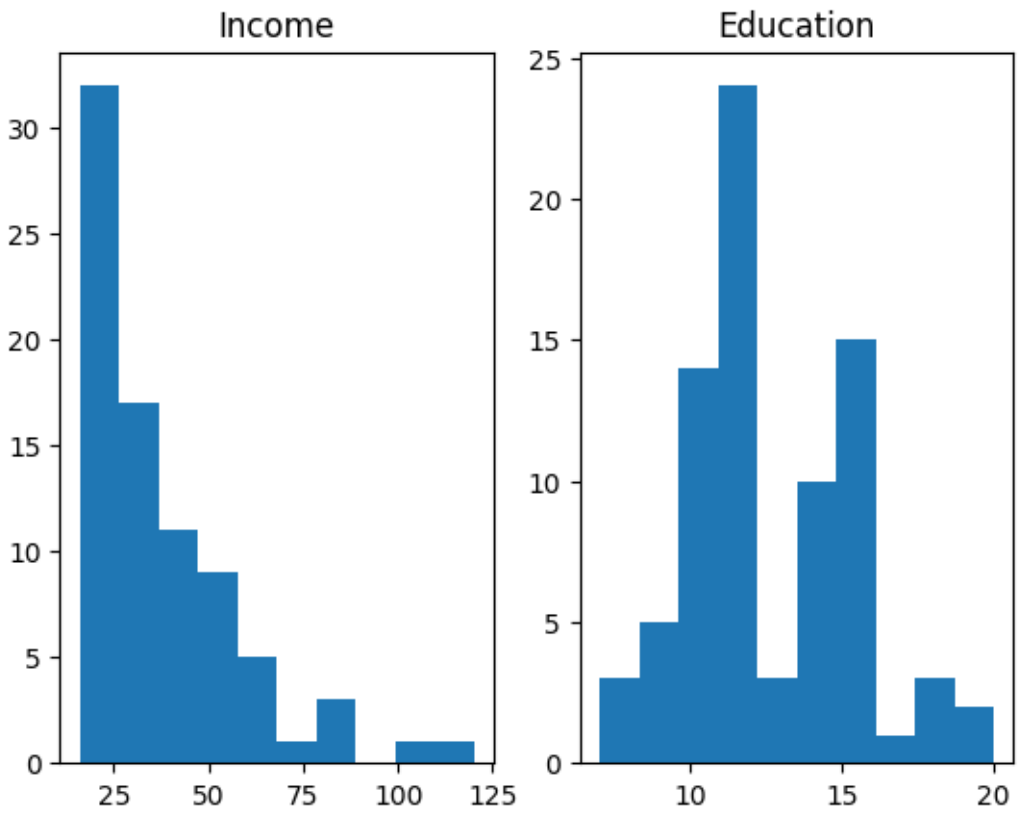
(d) Construct and interpret side-by-side box plots of income by race (B = Black, H = Hispanic, W = White). Compare the incomes using numerical descriptive statistics

Answer:

[8]:
```
# Include library
import matplotlib.pyplot as plt

# Read in the data file
data = pd.read_csv('Income.dat', sep='\s+')

# See the data
print(data.head())

# Plots
fig, ax = plt.subplots(1,2)
ax[0].hist(data = data, x = 'income')
ax[0].set_title('Income')
ax[1].hist(data = data, x = 'education')
ax[1].set_title('Education')
```

```
   income  education race
0      16         10    B
1      18          7    B
2      26          9    B
3      16         11    B
4      34         14    B
```

[8]: `Text(0.5, 1.0, 'Education')`



(a) The histograms of the income and the education levels are shown in the above plots.

The histogram of the income is right-skewed. The histogram of the education level is primarily bi-modal.

[9]:
```
print('Income statistics:')
print(data['income'].describe())
print(' ')
print('Education statistics:')
print(data['education'].describe())
```

```
Income statistics:
count     80.000000
mean      37.525000
std       20.672843
min       16.000000
25%       22.000000
50%       30.000000
75%       46.500000
max      120.000000
Name: income, dtype: float64
```

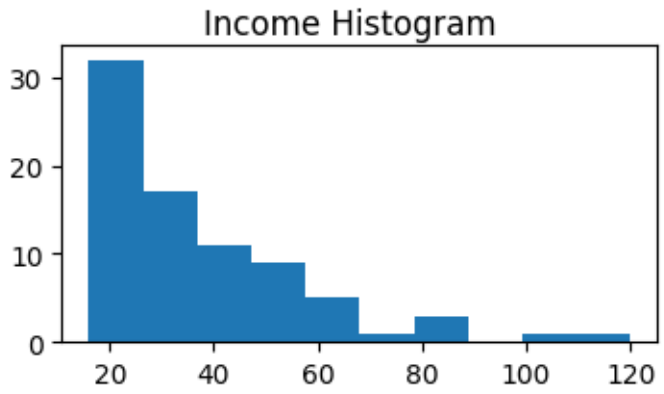
```
Education statistics:
count    80.000000
mean     12.687500
std       2.871042
min       7.000000
25%      10.000000
50%      12.000000
75%      15.000000
max      20.000000
Name: education, dtype: float64
```

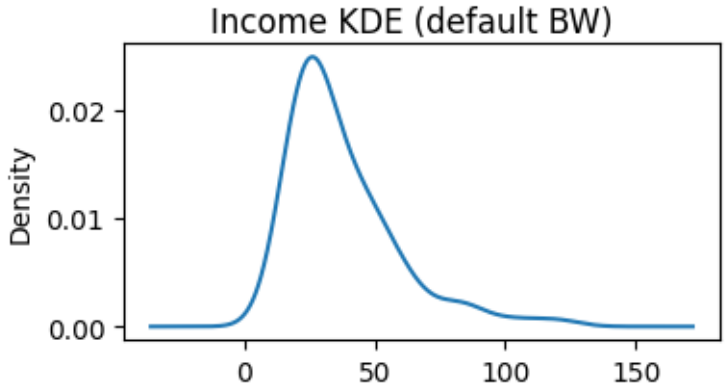
(b) For the Income distribution, the mean is more than the median. This means the distribution is probably right-skewed - which is confirmed when looking at the histogram plot.

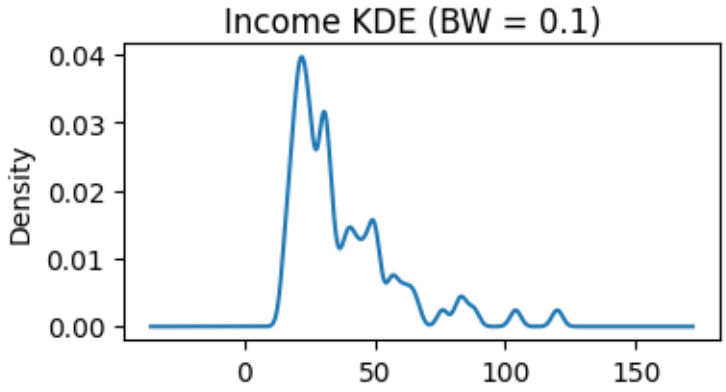
For the Education distribution, the mean and median are very close. This can mean that the distribution is normally distributed. However, this is not the case. The histogram shows that the distribution is bi-modal, roughly symetrically distributed around the mean value.

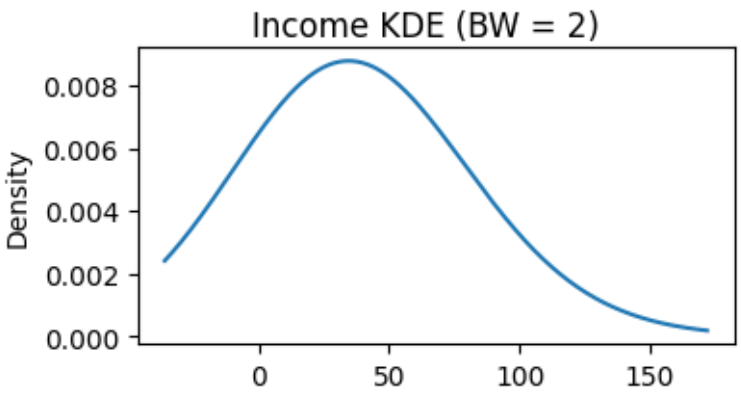
[10]:
```
# Get income data
income = data['income']

# Create histogram and kde plots
plt.figure(figsize=(4,2))
plt.hist(income)
plt.suptitle('Income Histogram')

plt.figure(figsize=(4,2))
income.plot.kde()
plt.suptitle('Income KDE (default BW)')

plt.figure(figsize=(4,2))
income.plot.kde(bw_method = 0.1)
plt.suptitle('Income KDE (BW = 0.1)')

plt.figure(figsize=(4,2))
income.plot.kde(bw_method = 2)
plt.suptitle('Income KDE (BW = 2)')
```

[10]:
```
Text(0.5, 0.98, 'Income KDE (BW = 2)')
```


Income Histogram
30
20
10
0
20
40
60
80
100
120


Income KDE (default BW)
Density
0.02
0.01
0.00
0
50
100
150


Income KDE (BW = 0.1)
Density
0.04
0.03
0.02
0.01
0.00
0
50
100
150

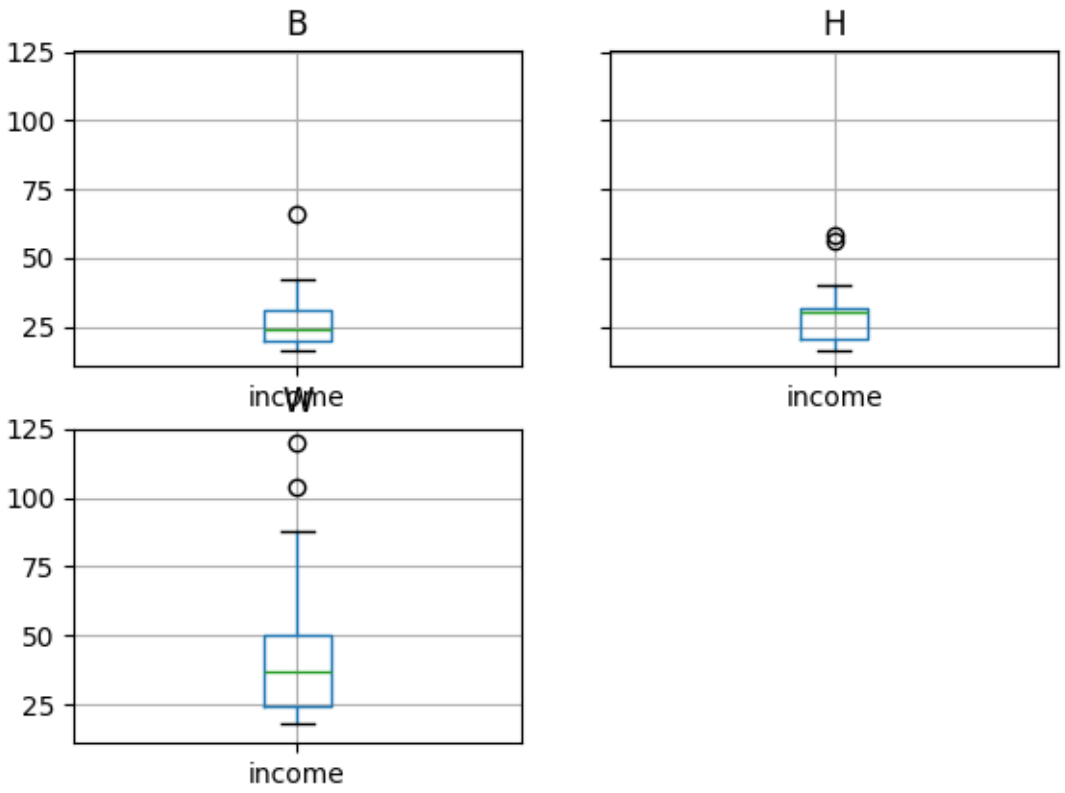


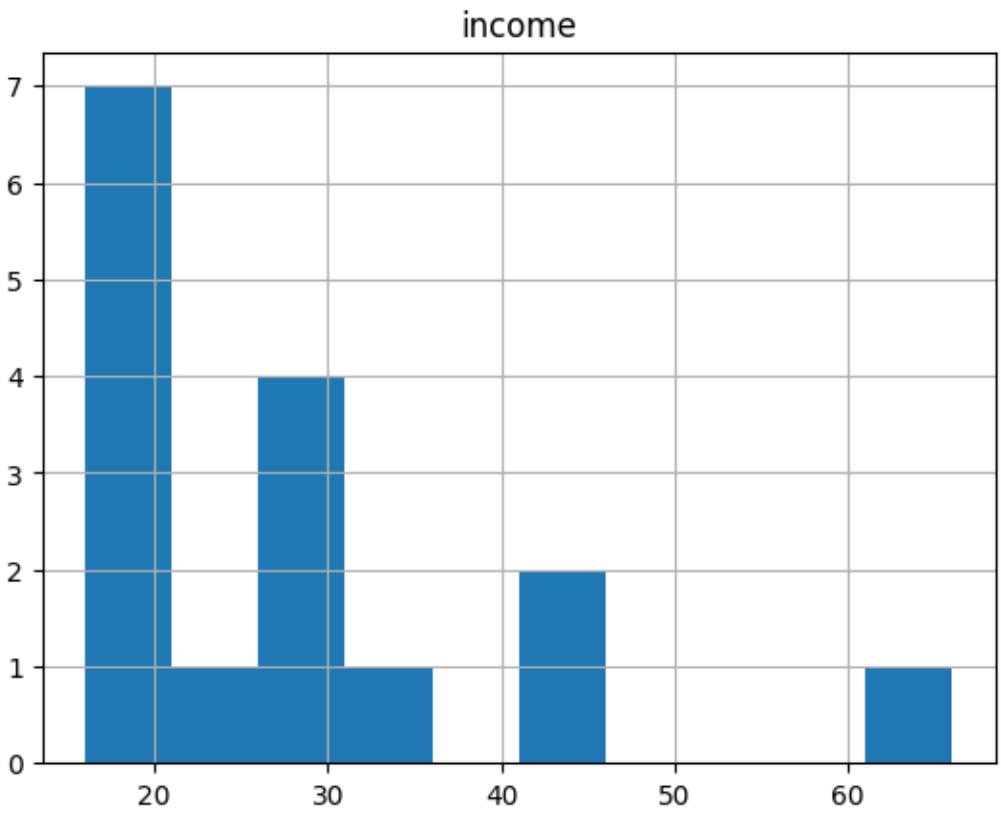
(c) The code and the outputs for the actual histogram and KDE's for different bandwidths are shown above.

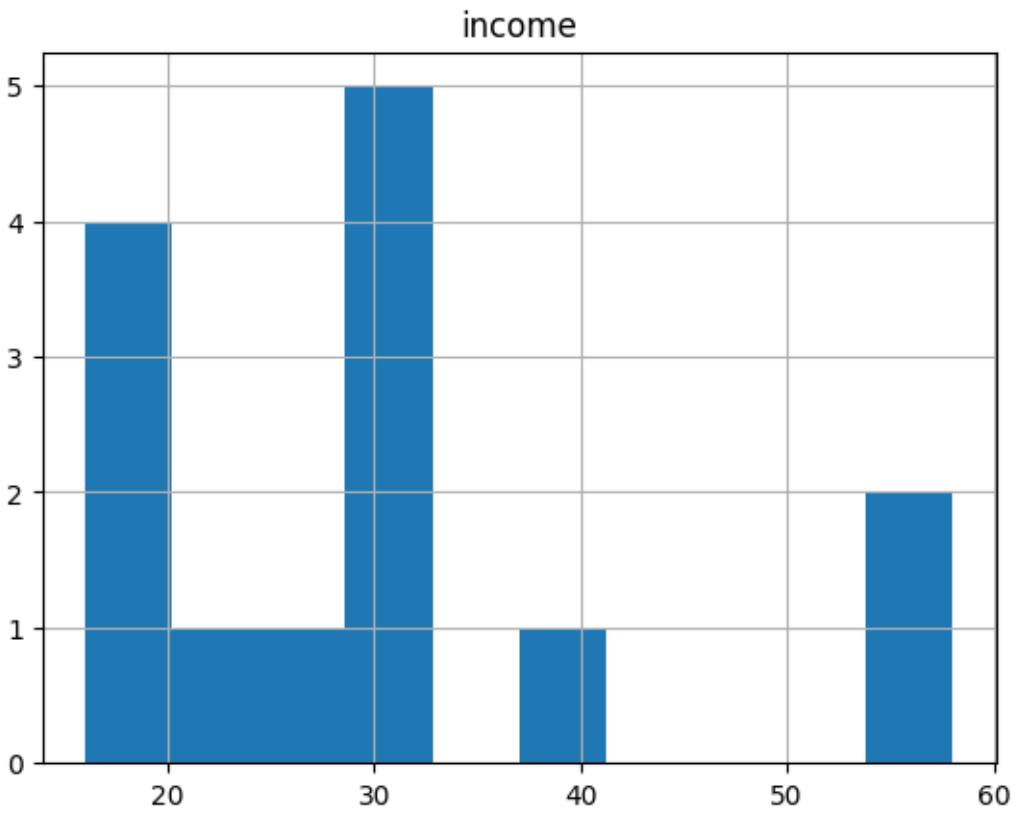
Using a small bandwidth value leads to over-fitting, while using a large bandwidth value results in under-fitting. In addition, with larger bandwidth, the data which are further away influences the curve more.

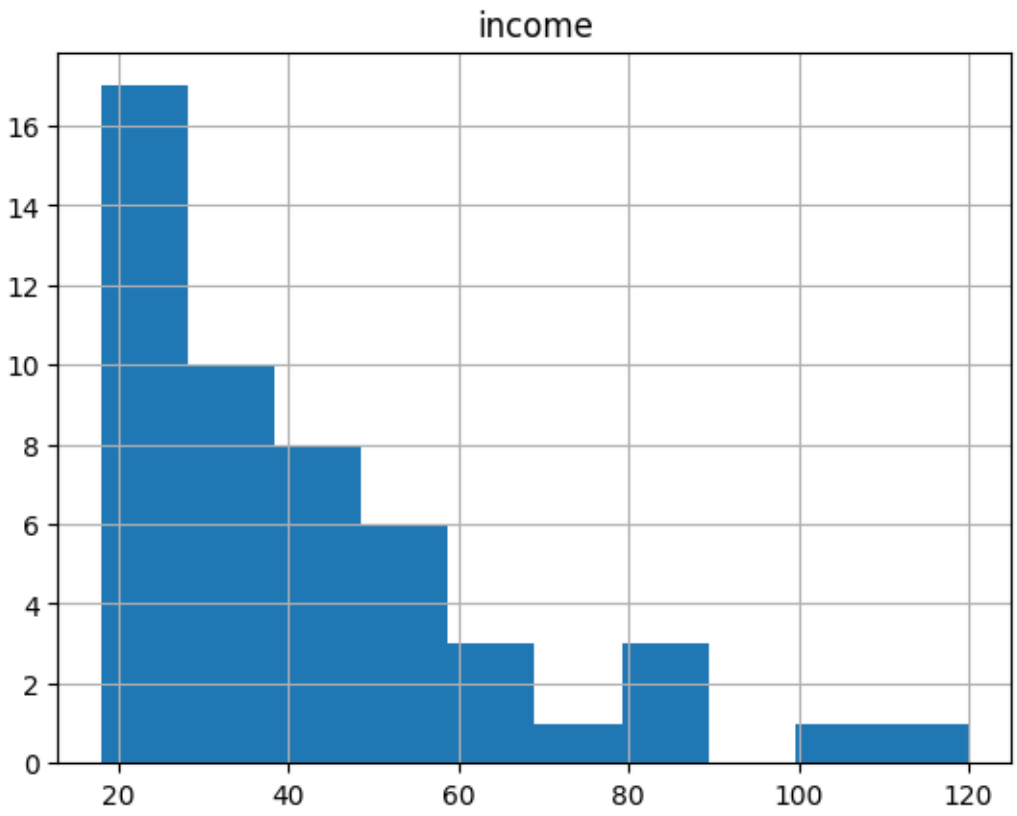
[11]:
```
# Boxplots grouped by Race
data.groupby('race').boxplot(column='income')

# Histogram grouped by Race
data.groupby('race').hist(column='income')
```

[11]:
```
race
B    [[Axes(0.125,0.11;0.775x0.77)]]
H    [[Axes(0.125,0.11;0.775x0.77)]]
W    [[Axes(0.125,0.11;0.775x0.77)]]
dtype: object
```


B
125
100
75
50
25
income
H
income
W
125
100
75
50
25
income


income
0
1
2
3
4
5
6
7
20
30
40
50
60


income
0
1
2
3
4
5
20
30
40
50
60



[12]:
```
data2 = data[['income', 'race']]
data2.groupby('race').describe()
```

[12]:

	income							
	count	mean	std	min	25%	50%	75%	max
race								
B	16.0	27.75	13.284076	16.0	19.5	24.0	31.0	66.0
H	14.0	31.00	12.812254	16.0	20.5	30.0	32.0	58.0
W	50.0	42.48	22.869854	18.0	24.0	37.0	50.0	120.0

(d) Based on the boxplots and the description statistics, we can interpret the following;

- Hispanic: The mean is slighty more than the median, the upper quartile is slightly more than the lower quartile, and the top whisker is more than the bottom whisker. This means the distribution can be right-skewed.
- Black: This is difficult to interpret since the 50% and 75% quartile are very closeby hinting the distribution is left skewed. However, mean and the median values are closeby which hints that the distribution is near normal. These are contradictory interpretations.
- White: The mean is more than the median, the upper quartile is more than the lower quartile, and the top whisker is more than the bottom whisker. This means the distribution can be right-skewed.
- Further we can say, that the White distribution is more right-skewed than the Hispanic

population.

These observations can be verified by the histogram plots shown above.

1.10 Problem # 1.19.

The `Houses` data file (http://stat4ds.rwth-aachen.de/data/Houses.dat) at the book's website lists the selling price (thousands of dollars), size (square feet), tax bill (dollars), number of bathrooms, number of bedrooms, and whether the house is new (1 = yes, 0 = no) for 100 home sales in Gainesville, Florida. Let's analyze the selling prices.

(a) Construct a frequency distribution and a histogram. Describe the shape.

(b) Find the percentage of observations that fall within one standard deviation of the mean. Why is this not close to 68%?

(c) Construct a box plot, and interpret.

(d) Use descriptive statistics to compare selling prices according to whether the house is new.

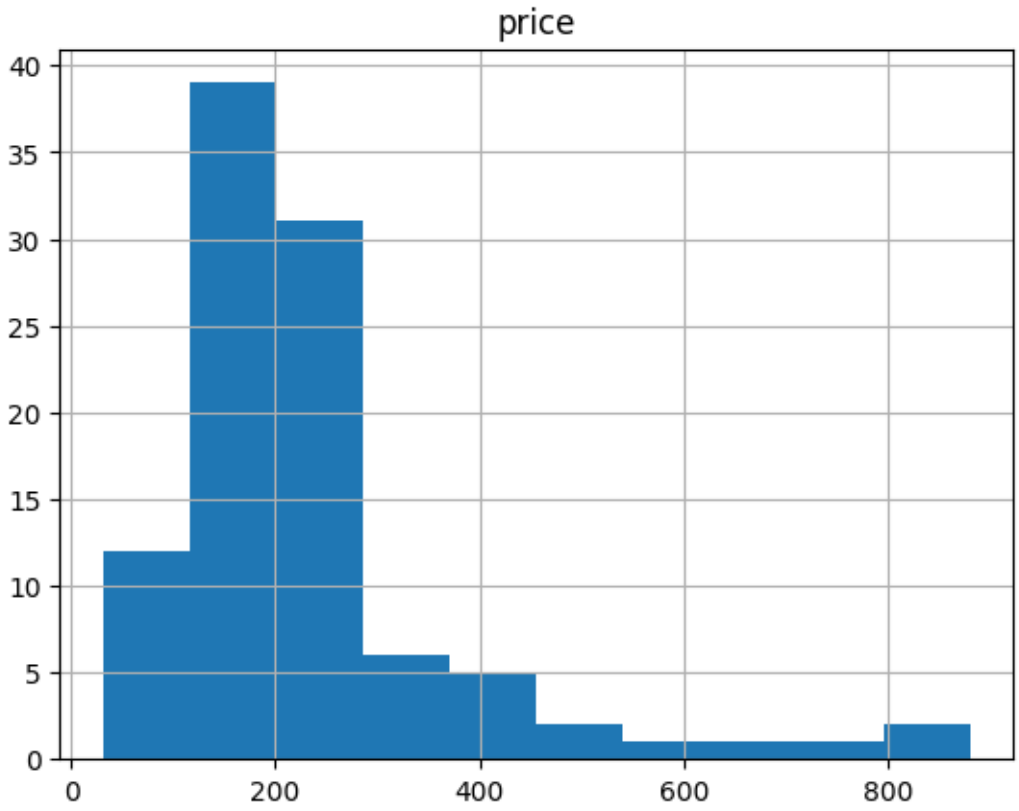
Answer:

[13]:
```
# Read in the data file
data = pd.read_csv('Houses.dat', sep='\s+')

# Data head
print(data.head())

# Frequency distribution
print('\nFrequency Distribution:')
price = data['price'].value_counts()
print(price)

# Histogram
data.hist(column='price')
```

```
   case   price  size  new  taxes  bedrooms  baths
0     1  419.85  2048    0   3104         4      2
1     2  219.75   912    0   1173         2      1
2     3  356.55  1654    0   3076         4      2
3     4  300.00  2068    0   1608         3      2
4     5  239.85  1477    0   1454         3      3

Frequency Distribution:
price
150.00    3
205.50    3
127.50    3
210.00    3
104.85    2
```

```
          ..
90.00      1
190.50     1
129.00     1
75.00      1
190.80     1
Name: count, Length: 81, dtype: int64
```

[13]:
```
array([[<Axes: title={'center': 'price'}>]], dtype=object)
```



(a) The histogram shows that the data is right-skewed.

[14]:
```
# Calculate price statistics
price_mean = data['price'].mean().round(2)
price_std = data['price'].std().round(2)
print('Mean price = ' + str(price_mean))
print('Stddev price = ' + str(price_std))

# Convert data to list
price_np = np.array(data['price'].values.tolist())
```

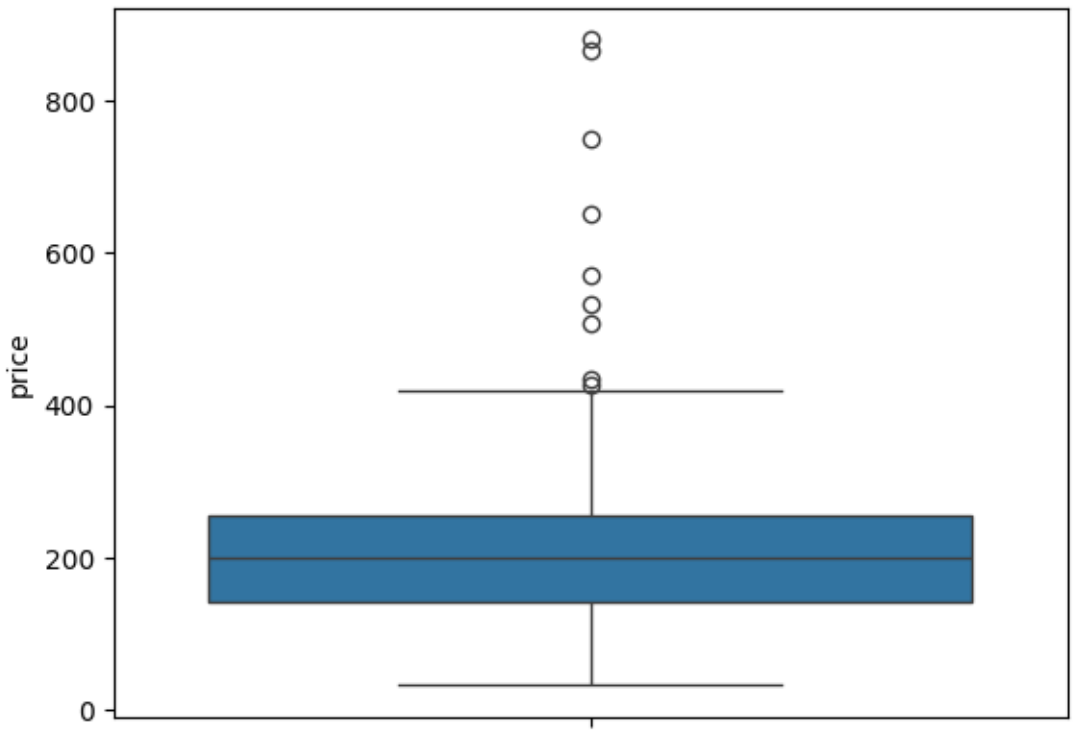
```
# 1 stddev range
val1 = price_mean - price_std
val2 = price_mean + price_std

# Find percentage of data within 1 stddev
count = 0
for i in range(len(price_np)):
    if (price_np[i] >= val1) and (price_np[i] <= val2):
        count = count + 1

val3 = count/len(price_np)*100
print('Percentage data within 1 stddev = ' + str(val3))
```

```
Mean price = 233.0
Stddev price = 151.89
Percentage data within 1 stddev = 85.0
```

(b) The percentage of data within 1 standard deviation is 85%. This is not close to 68% since the data distribution is not normal.

[15]:
```
sns.boxplot(data = data['price'])
```

[15]:
```
<Axes: ylabel='price'>
```



(c) The boxplot is shown above.

The top whisker is slightly more than the bottom whisker. Further, there are several outliers on the top. This can be interpreted as the distribution is right-skewed.

[16]:
```
data2 = data[['price','new']]
data2.groupby('new').describe()
```

[16]:

	price							
	count	mean	std	min	25%	50%	75%	max
new								
0	89.0	207.851124	121.039149	31.50	135.00	190.8	240.000	880.50
1	11.0	436.445455	219.832789	158.85	256.95	427.5	519.675	866.25

(d) The above descriptive statistics shows that the mean price of new houses is approximately \$229K more than the mean price of old houses. Further, the standard deviation of new house price is much higher (approximately \$99K) than old houses.

However, it is interesting to observe than the maximum price of one of the old houses is more than that of the new ones.