PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013)

UE19CS203 – STATISTICS FOR DATA SCIENCE

Unit-1 - Introduction to Data Science

QUESTION BANK – SOLVED

Data Visualization Techniques – Histogram

Exercises for Section 1.3

1. **The weather in Los Angeles is dry most of the time, but it can be quite rainy in the winter. The rainiest month of the year is February. The following table presents the annual rainfall in Los Angeles, in inches, for each February from 1965 to 2006.**

0.2	**3.7**	**1.2**	**13.7**	**1.5**	**0.2**	**1.7**
0.6	**0.1**	**8.9**	**1.9**	**5.5**	**0.5**	**3.1**
3.1	**8.9**	**8.0**	**12.7**	**4.1**	**0.3**	**2.6**
1.5	**8.0**	**4.6**	**0.7**	**0.7**	**6.6**	**4.9**
0.1	**4.4**	**3.2**	**11.0**	**7.9**	**0.0**	**1.3**
2.4	**0.1**	**2.8**	**4.9**	**3.5**	**6.1**	**0.1**

a. Construct a stem-and-leaf plot for these data. (Exclude)
b. Construct a histogram for these data.
c. Construct a dotplot for these data. (Exclude)
d. Construct a boxplot for these data. Does the boxplot show any outliers?

[Text Book Exercise – Section 1.3 – Q. No.1 – Pg. No. 39]

Solution:

b. Construct a histogram for these data.

Step: 1 – Prepare the Data

Arrange the values in ascending order (number of data points (n) = 42)

0.0	0.1	0.1	0.1	0.1	0.2	0.2
0.3	0.5	0.6	0.7	0.7	1.2	1.3
1.5	1.5	1.7	1.9	2.4	2.6	2.8

3.1	3.1	3.2	3.5	3.7	4.1	4.4
4.6	4.9	4.9	5.5	6.1	6.6	7.9
8.0	8.0	8.9	8.9	11.0	12.7	13.7

Step: 2 – Identify the Bin Widths

By using the Freedman – Diaconis , the bin width / class intervals can be found.

$$\text{Bin size} = 2\,\frac{\text{IQR}(x)}{\sqrt[3]{n}}$$

Find the IQR (InterQuartile Range)

$$IQR = Q_3 - Q_1$$

Quartile 1, $Q_1 = 0.25\ (n+1) = 0.25\ (43) = 10.75$

0.0	0.1	0.1	0.1	0.1	0.2	0.2
0.3	0.5	0.6	0.7	0.7	1.2	1.3

$$\frac{0.6 + 0.7}{2} = 0.65$$

Quartile 3, $Q_3 = 0.75\ (n+1) = 0.75\ (43) = 32.25$

3.1	3.1	3.2	3.5	3.7	4.1	4.4
4.6	4.9	4.9	5.5	6.1	6.6	7.9

$$\frac{5.5 + 6.1}{2} = 5.8$$

IQR = 5.8 – 0.65 = 5.15

Substitute in the formula, let's find the Bin width

$$\frac{2 * 5.15}{\sqrt[3]{42}} = 2.9 \sim 3$$

Step: 3 – Build the frequency distribution table

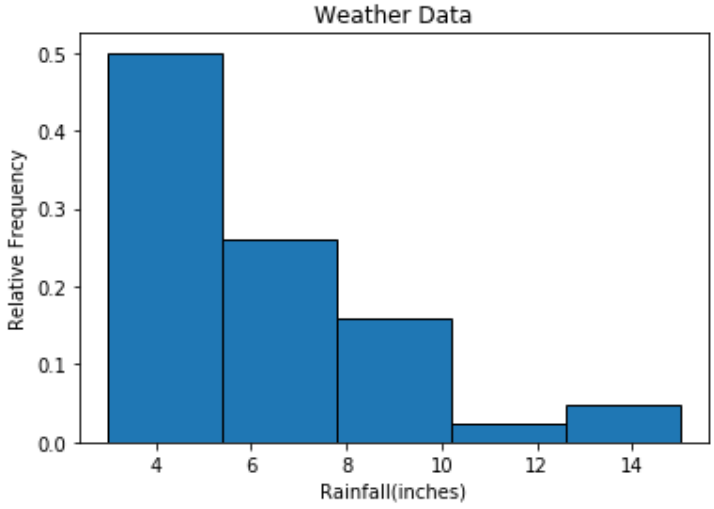
Class	Frequency	Relative Frequency	Density
0 – 3	21	0.5	0.1667
3 – 6	11	0.2619	0.0873
6 – 9	7	0.1667	0.0555
9 – 12	1	0.0238	0.0073
12 - 15	2	0.0476	0.0159
	Sum = 42	**Sum = 1**	

Step: 4 – Find the number of bins / buckets

Number of bins/buckets = $\frac{\text{Max} - \text{Min}}{\text{Bin} - \text{Size}} = \frac{15-0}{3} = 5$

Step: 5 – Plot the Histogram

Here is one histogram. Other choices for the endpoints are possible.



2. **Following are measurements of soil concentrations (in mg/kg) of chromium (Cr) and nickel (Ni) at 20 sites in the area of Cleveland, Ohio. These data are taken from the article "Variation in North American Regulatory Guidance for Heavy Metal Surface Soil Contamination at Commercial and Industrial Sites" (A. Jennings and J. Ma, *J Environment Eng*, 2007:587–609).**

Cr	**34**	**1**	**511**	**2**	**574**	**496**	**322**	**424**
	269	**140**	**244**	**252**	**76**	**108**	**24**	**38**
	18	**34**	**30**	**191**				

Ni	**23**	**22**	**55**	**39**	**283**	**34**	**159**	**37**

	61	**34**	**163**	**140**	**32**	**23**	**54**	**837**
	64	**354**	**376**	**471**				

a. Construct a histogram for each set of concentrations.
b. Construct comparative boxplots for the two sets of concentrations.
c. Using the boxplots, what differences can be seen between the two sets of concentrations?

[Text Book Exercise – Section 1.3 – Q. No. 4 – Pg. No. 39]

Solution:

a. Construct a histogram for each set of concentrations.

Step: 1 – Prepare the Data

Arrange the values in ascending order (number of data points (n) = 20)

Cr	1	2	18	24	30	34	34	38
	76	108	140	191	244	252	269	322
	424	496	511	574				

Arrange the values in ascending order (number of data points (n) = 20)

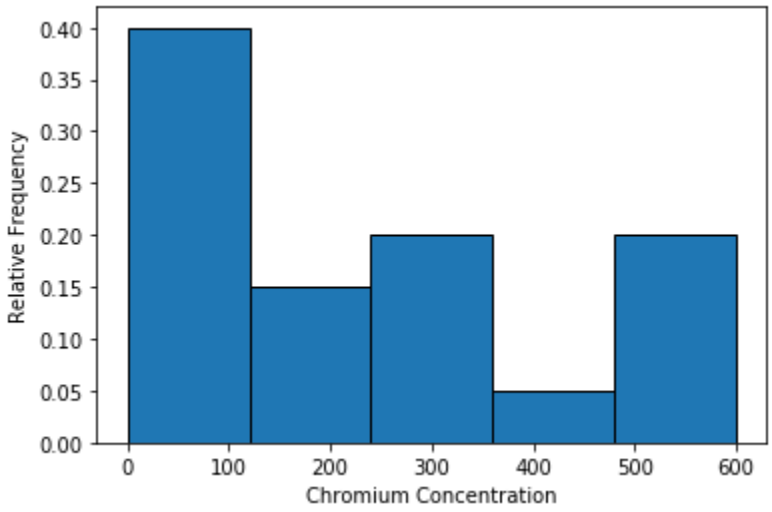
Ni	22	23	23	32	34	34	37	39
	54	55	61	64	140	159	163	283
	354	376	471	837				

Step: 2 – Construct the frequency table

Frequency Table for Chromium Concentration

Class	**Frequency**	**Relative Frequency**
0 – 60	8	0.40
60 – 120	2	0.10
120 - 180	1	0.05
180 – 240	1	0.05
240 – 300	3	0.15
300 – 360	1	0.05
360 – 420	0	0
420 – 480	1	0.05
480 – 540	2	0.10
540 - 600	1	0.05
	Sum = 20	Sum = 1

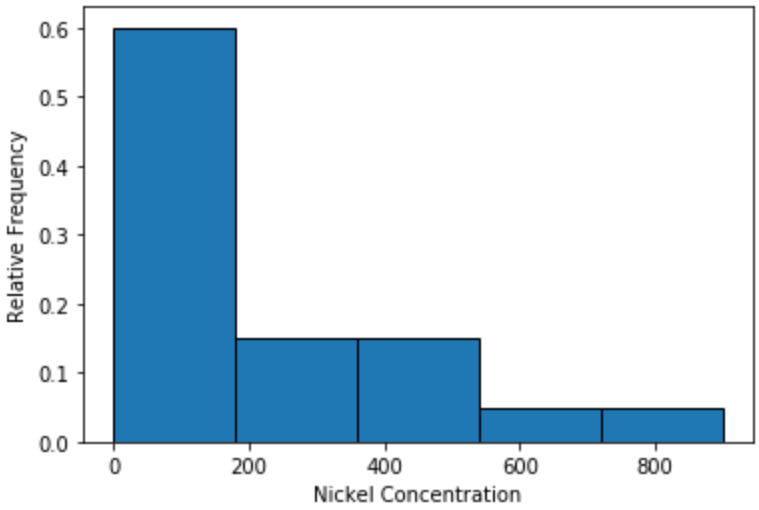
Here is one histogram. Other choices for the endpoints are possible.



Frequency Table for Nickel Concentration

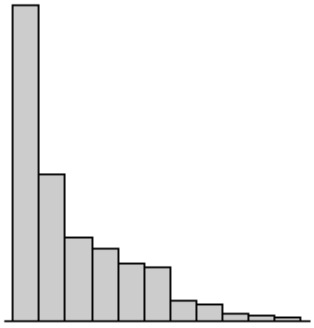
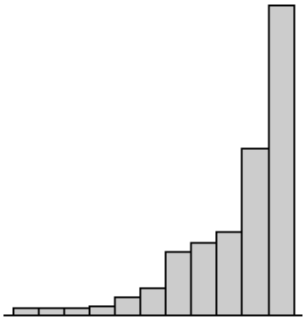
Class	Frequency	Relative Frequency
0 – 90	12	0.60
90 – 180	3	0.15
180 – 270	0	0
270 – 360	2	0.10
360 – 450	1	0.05
450 – 540	1	0.05
540 – 630	0	0
630 – 720	0	0
720 – 810	0	0
810 - 900	1	0.05
	Sum = 20	Sum = 1

Here is one histogram. Other choices for the endpoints are possible.



3. Sketch a histogram for which
a. The mean is greater than the median.
b. The mean is less than the median.
c. The mean is approximately equal to the median.

[Text Book Exercise – Section 1.3 – Q. No. 6 – Pg. No. 39]

a. The histogram should be skewed to the right.	
b. The histogram should be skewed to the left.	

c. The histogram should be approximately symmetric.