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 – Box Plot

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:

d. Construct a boxplot for these data. Does the boxplot show any outliers?

Step: 1 – Prepare the data

Arrange the data in order (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 – Construct Five Number Summary

1) Median:

$$\frac{2.8 + 3.1}{2} = 2.95$$

2) 1^{st} Quartile = (0.25) (n + 1) = 0.25 * 43 = 10.75

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

3) 3^{rd} Quartile = (0.75) (n + 1) = 0.75 * 43 = 32.25

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

4) Minimum Value = 10

5) Maximum Value= 59

Step: 3 – Compute Inter-Quartile Range

Next compute the Inter-Quartile Range

IQR = 5.8 – 0.65 = 5.15

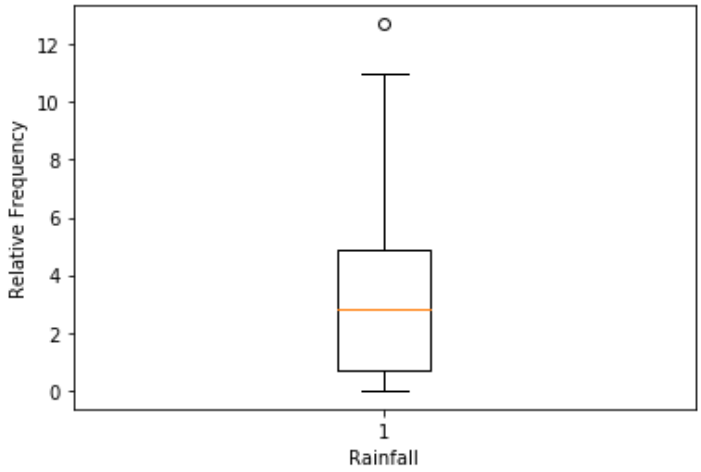
Step: 4 – Find the Outliers if present.

Lower = Q_1 – 1.5 * IQR = 0.65 – (1.5 * 5.15) = -7.075

Upper = Q_3 + 1.5 * IQR = 5.8 + (1.5 * 5.15) = 13.525

Step: 5 – Construct a Box Plot

There is one outlier.



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:

b. Construct comparative boxplots for the two sets 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 Five Number Summary

Five Number Summary	Chromium Concentration	Nickel Concentration
Median	$\frac{108 + 140}{2} = 124$	$\frac{55 + 61}{2} = 58$
st Quartile	(0.25)(n+1) = 0.25 * 21 = 5.25 $\frac{30 + 34}{2} = 32$	(0.25)(n+1) = 0.25 * 21 = 5.25 $\frac{34 + 34}{2} = 34$
3rd Quartile	(0.75)(n+1) = 0.75*21 = 15.75 $\frac{269 + 322}{2} = 295.5$	(0.75)(n+1) = 0.75*21 = 15.75 $\frac{163 + 283}{2} = 223$
Minimum	1	22
Maximum	574	837

Step: 3 – Compute Inter-Quartile Range

Next compute the Inter-Quartile Range

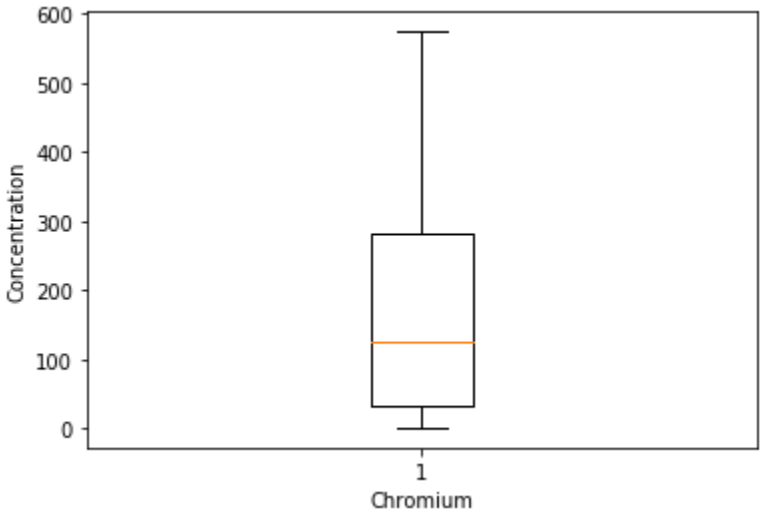
	Chromium Concentration	Nickel Concentration
IQR	295.5 – 32 = 263.5	223 – 34 = 189

Step: 4 – Find the Outliers if present.

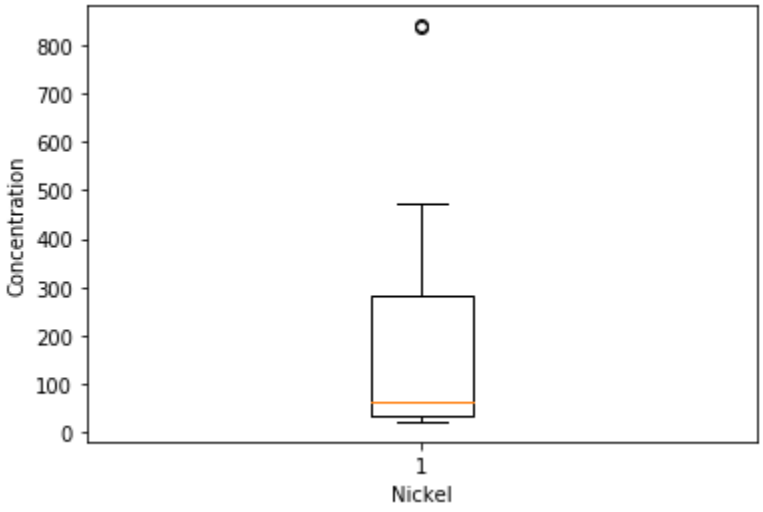
	Chromium Concentration	Nickel Concentration
Lower	32 – (1.5) (263.5) = -363.25	34 – (1.5) (189) = -249.5
Upper	295.5 + (1.5) (263.5) = 690.75	223 + (1.5) (189) = 506.5

Step: 5 – Construct a Box Plot

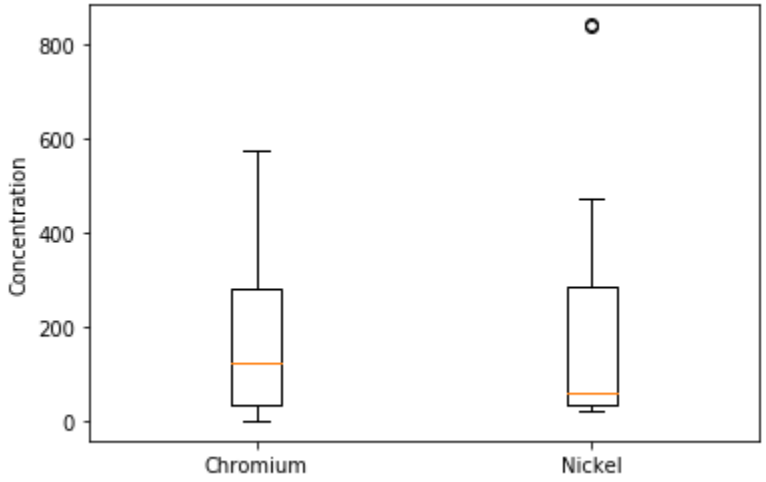
Chromium Concentration



Nickel Concentration



Comparative BoxPlots



c. Using the boxplots, what differences can be seen between the two sets of concentrations?

The concentrations of nickel are on the whole lower than the concentrations of chromium. The nickel concentrations are highly skewed to the right, which can be seen from the median being much closer to the first quartile than the third. The chromium concentrations are somewhat less skewed. Finally, the nickel concentrations include an outlier.