Data Visualisation

Chapter 3

Data Sources & Outliers

DR. TONY LYONS

AUTUMN SEMESTER 2019

B Sc (Hons) Multimedia Applications Development

B Sc (Hons) Creative Computing

WATERFORD INSTITUTE OF TECHNOLOGY

Contents

1 Data Sourcing

1.1 Primary Data

Definition 1.1: Primary Data Source

Primary Data are data collected by a researcher directly.

Features of Primary Data

- The data has never been collected before, either by a particular method, or at a specific period of time.
- Primary data is often collected when no other such data set exists from other sources.
- The method of sourcing the data e.g. questions asked, experiments run etc, may be tailored to meet the objectives of the research.
- However, collection of primary data may be very expensive, and if related to institutes such as schools, hospitals, banks etc. this may require institutional approval before being collected.
- Issues of consent and confidentiality are of extreme importance.
- In many cases, primary data should be the second option, after secondary data, since it is better to use current information relating to a given field before collecting new data.

Sources of Primary Data

Primary data are data collected directly by the researcher themselves. Some sources include

- Interviews
- Observations
- Surveys
- Questionnaires
- Experiments
- Case studies
- Longitudinal studies
- Eyewitnesses
- Legal documents

1.2 Secondary Data

Definition 1.2: Secondary Data

Secondary Data is data which has not been sourced directly by the researcher, but rather has been obtained from outside sources.

Features of Secondary Data

- The data has already been collected by outside researchers, which saves large financial and time resources for the research project.

- The data already exists on record somewhere, and with wide access to the internet for most people, this can mean a huge amount of data readily available.

- The use of secondary data can often help clarify the aims of a research project, and is often used prior to primary data to help focus the research aims.

- However, in contrast the primary data, the quality of secondary data may in some cases be questionable, sue to bias or other factors in the data collection.

- In some cases the data may not be appropriate for the research being carried out.

- It is also possible that the data is incomplete, in that certain data may have been omitted.

- It may be the case that the secondary data is no longer valid for current environments, for instance census data must be routinely updated.

Sources of Secondary Data

- Central Statistics Office
- Government Departments
- National Archives
- Media, e.g. stock market listings
- International bodies for example, World bank, WTO, UN
- Universities
- Libraries

2 Quartiles and IQR

The detection of outliers in data sets relies on finding the first and third quartiles Q_1 and Q_3 of a data-set, along with the inter-quartile range IQR. We give the definitions of these three terms here.

First Quartile

Definition 2.1: First or Lower Quartile (Q_1)

- Given a data set with N data points, the first quartile Q_1 is the data point **below** which one quarter of the data lies.
- The value of Q_1 corresponds to the data value at the $\left(\frac{N+1}{4}\right)^{\text{th}}$ data point.
- If $\frac{N+1}{4}$ is not a whole number, then the average of the data values either side of the $\left(\frac{N+1}{4}\right)^{\text{th}}$ should be averaged to give Q_1

Third Quartile

Definition 2.2: Third or Upper Quartile (Q_3)

- Given a data set with N data points, the first quartile Q_3 is the data point **above** which one quarter of the data lies.
- The value of Q_3 corresponds to the data value at the $\left(\frac{3(N+1)}{4}\right)^{\text{th}}$ data point.
- If $\frac{3(N+1)}{4}$ is not a whole number, then the average of the data values either side of the $\left(\frac{3(N+1)}{4}\right)^{\text{th}}$ should be averaged to give Q_3

Median

Like the quartiles, the median of a data set also divides the set, in this case, it is the number M which splits the data set into two halves.

Definition 2.3: Median (M)

- The median M of a data set corresponds to the data value at the $\left(\frac{N+1}{2}\right)^{\text{th}}$ data point.
- If $\frac{N+1}{2}$ is not a whole number, then the average of the data values either side of the $\left(\frac{N+1}{2}\right)^{\text{th}}$ should be averaged to give M.

Interquartile Range

The interquartile range (IQR) of a data set is the distance between to first and third quartiles. Specifically

Definition 2.4: Interquartile Range(IQR)

- Given a data set with first quartile Q_1 and third quartile Q_3, the interquartile range of the set is

$$IQR = Q_3 - Q_1.$$

Example 2.1

Find the first and third quartiles along with the median of the following data set

4	1	6	9	3
8	12	1	5	20
9	11	4	4	3
2	3	5	3	7
3	5	9	10	12

Solution. Since the data set is not ordered we must do so before we calculate the quartiles or the median. The resulting data set is given by

1	1	2	3	3
3	3	3	4	4
4	5	5	5	6
7	8	9	9	9
10	11	12	12	20

which we now use instead. Since there are N=25 data points in the data set, the first quartile is at the $\frac{26}{4} = 6.5^{\text{th}}$ position. The third quartile is at the $\frac{78}{4} = 19.5^{\text{th}}$ position. Since neither of these are whole numbers we mus take averages, and so Q_1 is given by the average of the 6^{th} and 7^{th} data values, giving

$$Q_1 = \frac{3+3}{2} = 3.$$

Likewise, the third quartile is given by the average of the 19^{th} and 20^{th} data values, giving

$$Q_3 = \frac{9+9}{2} = 9.$$

This means that one quarter of the data values have value less than 3, while one quarter have value greater than 9.

The median is found at the $\frac{25+1}{2} = 13^{\text{th}}$ position, meaning

$$M = 5,$$

in which case half the data have value less than 5 and half have value greater than 5.

3 Detection of Data Outliers

Definition 3.1: Outliers

An outlier is an observation that lies an abnormal distance from the other values in a data sample.

3.1 Stem & Leaf Plots

Often the overall shape of a graph representing a data set can indicate the presence of outliers. For instance, the appearance of tails in a histogram can indicate the presence of outliers in the tail region of the histogram (though the presence of a tail does not necessarily mean there are outliers present). Stem & Leaf plots are an effective means of indicating the presence of outliers in a data set, since the data is represented directly. These plots often display the shape of the data set directly in that we can see if the data is skewed left, skewed right, centered, just from the shape of the stem & leaf plot.

Example 3.1: Stem and leaf plot

The daily distances traveled by a sales representative during a working month are tabulated daily, with the data given in the table below

53	104	94	85	152	121	97
84	63	68	63	62	144	122
101	97	92	95	57	85	59
122	106	78	84	93	109	111

Using this data, answer the following:

(i) Create an appropriately labeled stem & leaf plot for this data.

(ii) From this stem & leaf plot, determine if the data is skewed-left, skewed-right or centered.

(iii) Determine the best measure of centrality, based on the skewedness of the data.

Solution.

(i) The stem and leaf plot is given by

Stem	**Leaf**					
5	7	9				
6	2	3	3	3	8	
7	8					
8	7	8				
9	2	3	4	5	7	7
10	1	4	6	9		
11	1					
12	1	2	2			
13						
14	4					
15	2					

Table 1: Stem & leaf plot of daily distances traveled by a sales representative.
Key: $5|7 = 57$

(ii) Turning the stem & leaf plot counter-clockwise by 90° we find

Stem	**Leaf**
5	7 9
6	2 3 3 3 8
7	8
8	7 8
9	2 3 4 5 7 7
10	1 4 6 9
11	1
12	1 2 2
13	
14	4
15	2

The data appears to be **skewed-right**, based on the shape of the rotated stem & leaf plot.

(iii) Since the data is **skewed-right**, the best measure of centrality is the **median**.

3.2 Tukey's Criteria

While a stem & leaf plot is a good first indication of whether or not a data set contains outliers, it cannot be used to definitively say there are or are not outliers present. To detect the presence of outliers numerically we use **Tukey's Criteria**, which classifies outliers as **mild** or **extreme**.

Definition 3.2: Mild outliers

A data set has mild outliers if there are data values x such that

$$x < Q_1 - 1.5IQR \quad \text{or} \quad x > Q_3 + 1.5IQR$$

Definition 3.3: Extreme outliers

A data set has extreme outliers if there are data values x such that

$$x < Q_1 - 3IQR \quad \text{or} \quad x > Q_3 + 3IQR$$

Example 3.2: Tukey's criteria for outliers

Use Tukey's criteria to detect the possible presence of outliers in the data set given in **Example** 2.1.

Solution. To remind ourselves, we tabulate the ordered form of the data set again

1	1	2	3	3
3	3	3	4	4
4	5	5	5	6
7	8	9	9	9
10	11	12	12	20

We already know that the first and third quartiles of this data set are

$$Q_1 = 3 \quad Q_3 = 9.$$

Hence the interquartile range is

$$IQR = 9 - 3 = 6.$$

According to **Tukey's criteria**, there are extreme outliers in the data if there are data values x such that

$$\begin{cases} x < 3 - 3(6) \text{ or } x < -15 \\ x > 9 + 3(6) \text{ or } x > 27. \end{cases}$$

Since there are no data values less than -15 or greater than 27, it follows that there are **no extreme outliers**. Likewise, there are mild outliers if there are data values x such that

$$\begin{cases} x < 3 - 1.5(6) \text{ or } x < -6 \\ x > 9 + 1.5(6) \text{ or } x > 18. \end{cases}$$

While there are no data values less than -6, there is one data value greater than 18 (i.e. the data value 20), and so there is **one mild outlier**.

3.3 Box plots

Box plots combine a quantitative analysis and a visual representation of the data. They are an effective way of representing the data while highlighting the presence of outliers in the data.

3.3.1 The 5 number summary

When creating a box plot to represent a data set, we use the **5 number summary** of the data set, these numbers being

1. Minimum
2. First Quartile
3. Median
4. Third Quartile
5. Maximum.

Remark 3.1: Maximum and minimum of data sets

When drawing a box plot, the maximum and minimum values used should not be outliers. If this is the case then the smallest and largest values which are not outliers are used as the new minimum and maximum respectively.

3.3.2 Creating the box plot

When creating a box plot of a data set we always follow the following procedure

(I) Obtain the 5 number summary of the data set.

(II) Test the data set for outliers.

(III) Draw a rectangular box, marking the end of the box as Q_1 and Q_3.

(IV) Draw the fences of the plot. If the data has no outliers then the fences are the maximum and minimum. If there are outliers, the smallest and largest values which are not outliers.

(V) Mark any outliers outside these fences with a dot or a circle.

(VI) Draw *whiskers* connecting the ends of the box to the fences.

Example 3.3

Create a box plot to represent the data sets

1	1	2	3	3
3	3	3	4	4
4	5	5	5	6
7	8	9	9	9
10	11	12	12	20

Solution. The **5 number summary** of this data set is given by

Minimum 1

First Quartile 3

Median 5

Third Quartile 9

Maximum 20

which was obtained using **R**.

Next, upon testing for outliers it is found that the data value 20 is an outlier. Hence, we used

$$Q_3 + 1.5\text{IQR} = 9 + 1.5(9 - 3) = 18$$

as the **maximum for the box plot.**

Following the rest of the steps outlined in (I)–(IV), we create a box plot to represent this data given by

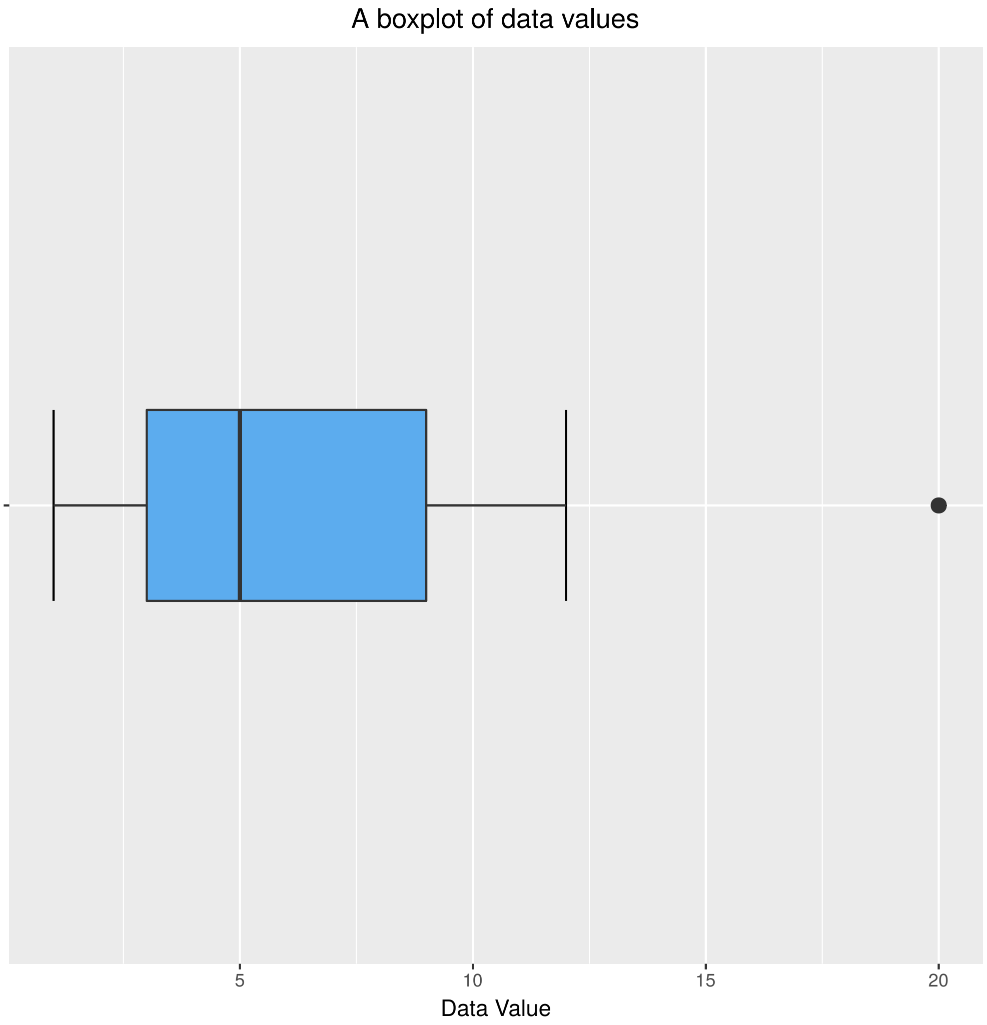

Figure 1: A box plot with an outlier.

3.3.3 Comparing data sets with box plots

It is also possible to compare various data sets using different box plots on the same chart. This is particularly useful when we want to compare the centre and spread of comparable data sets.

Example 3.4: Comparing data sets

A lecturer divides a class into two groups of ten, and assigns each group an exam covering the same material. The grades of each group are listed in the table below:

Student	1	2	3	4	5	6	7	8	9	10
Group 1	23	56	43	45	78	61	41	36	51	4
Group 2	45	47	39	51	52	43	28	40	32	38

Create two box plots on the same graph to represent these data. Comment on the grades of each group, with reference to the box plot.

Solution. The 5 number summaries of the two groups are given in the table below

Number	**Group 1**	**Group 1**
Minimum	4	28
1st Quartile	29.5	35
Median	44	41.5
3rd Quartile	53.5	46
Maximum	28	52

Applying Tukey's criteria to each group, only Group 1 has an outlier, which is the grade 4. Using all of this we now draw a pair of box plots on the same chart for these data sets, as shown in Figure 2.

Comparing the grades of each group using Figure 2, it is clear that the spread of grades in Group 1 is much larger than the spread of grades in Group 2. It is also noticeable that the median grade for Group 1 is slightly higher than the median grade for Group 2. Moreover, Group 1 contains a grade which is classified as an outlier.

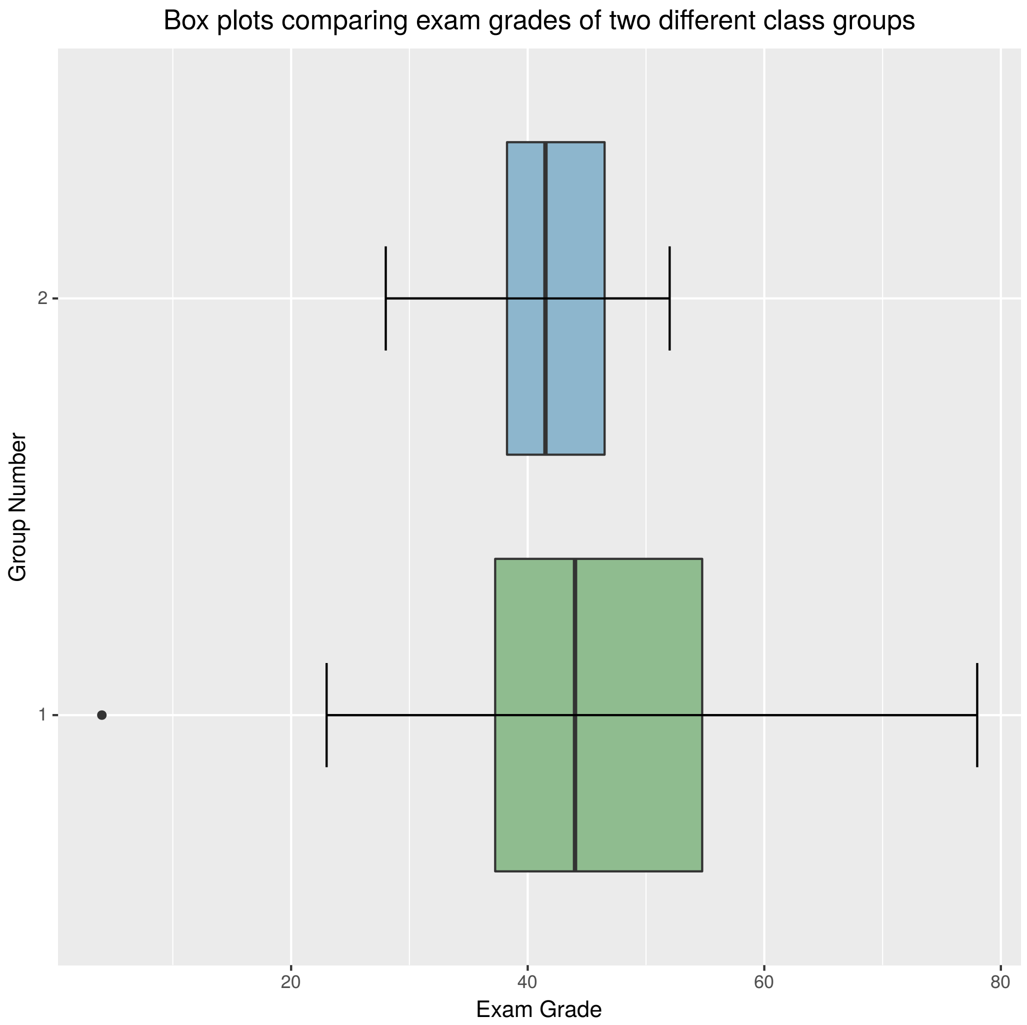

Figure 2: A box plot comparing the exam score of two student groups in a class.