Queen Mary
University of London

EMS702UP Statistical Thinking and Applied Machine Learning

W2.2 Exploratory Data Analysis & Visualising data

Jun Chen

Exploratory Data Analysis & Visualisation

Exploratory Data Analysis (EDA) is the activity by which a data set is explored and organised ***in order*** so that information it contains is made clear.

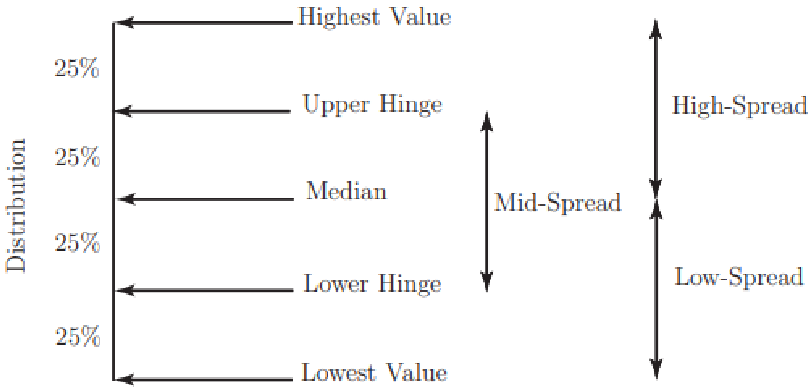
The basic principles followed in EDA are:

- To measure the location and spread of a distribution we use statistics which are **resistant** to departures from normality;
- To summarise shape location and spread we use several statistics rather than just two;
- **Visual displays** as well as **numerical displays** are used to summarise information obtained about shape, location and spread.

Data visualisation is an important skill in applied statistics and machine learning.

- Data visualisation can be helpful when exploring and getting to know a dataset and can help with identifying **patterns**, **corrupt data**, **outliers**, and much more.
- With a little domain knowledge, data visualisations can be used to express and demonstrate key relationships in plots and charts that are more visceral to yourself and stakeholders than measures of association or significance.

Exploratory Data Analysis & Visulisation – Five Number Summary Statistics



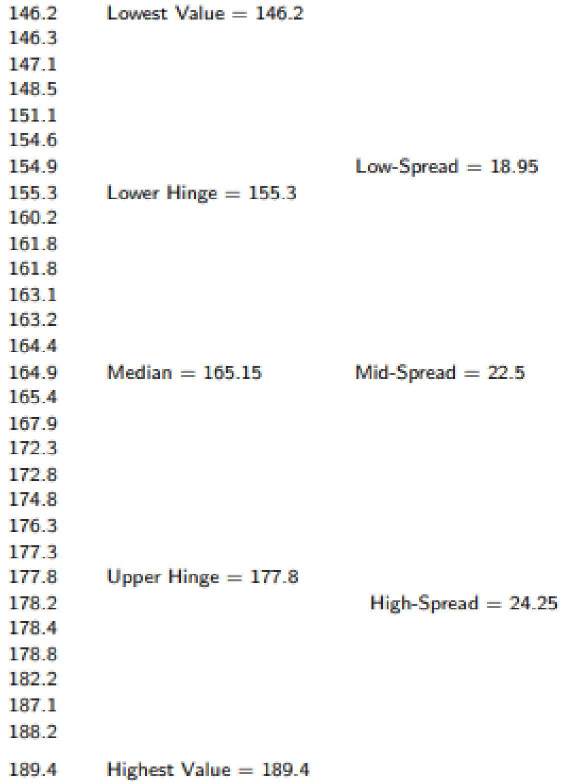
- The lower and upper hinges are the lower and upper sample quartiles.
- The mid-spread is the inter-quartile range (IQR).
- The five-number summary, especially when used in conjunction with the three spreads shown in above gives an adequate representation of a non-symmetrical distribution.
- The median and the hinges are unaffected by changes in extreme values.

Exploratory Data Analysis & Visulisation – Five Number Summary Statistics

Example 9

Find the five number summary and the mid-spread, high-spread and low-spread for the sample from Example 3.

Solution



146.2	Lowest Value = 146.2	
146.3		
147.1		
148.5		
151.1		
154.6		
154.9		Low-Spread = 18.95
155.3	Lower Hinge = 155.3	
160.2		
161.8		
161.8		
163.1		
163.2		
164.4		
164.9	Median = 165.15	Mid-Spread = 22.5
165.4		
167.9		
172.3		
172.8		
174.8		
176.3		
177.3		
177.8	Upper Hinge = 177.8	
178.2		High-Spread = 24.25
178.4		
178.8		
182.2		
187.1		
188.2		
189.4	Highest Value = 189.4	

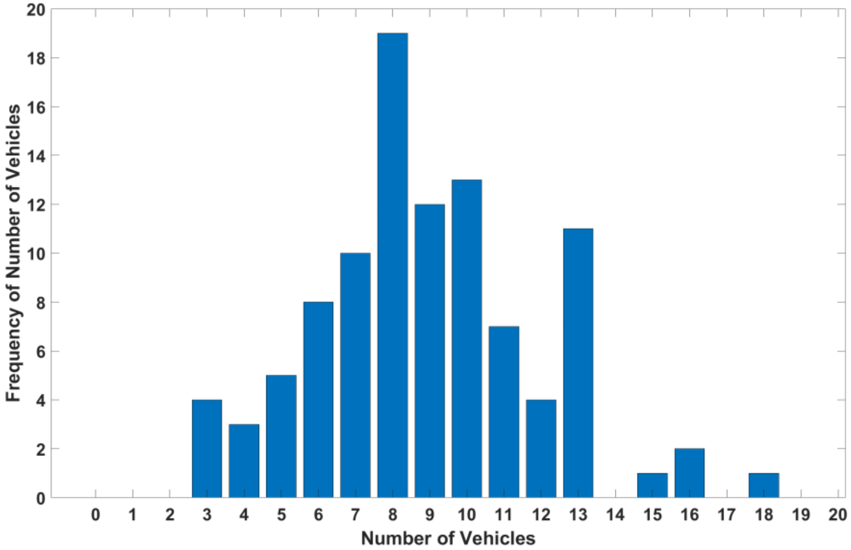
Exploratory Data Analysis & Visulisation – Bar Charts

Bar charts can be used to represent the frequencies for **categorical** or **discrete** data. In a bar chart, we draw a bar for each value of the variable. The length of the bar is **proportional** to the frequency for that value. Bars can be drawn vertically or horizontally.

Example 10

Construct a bar chart for the sample from Example 2.

Solution



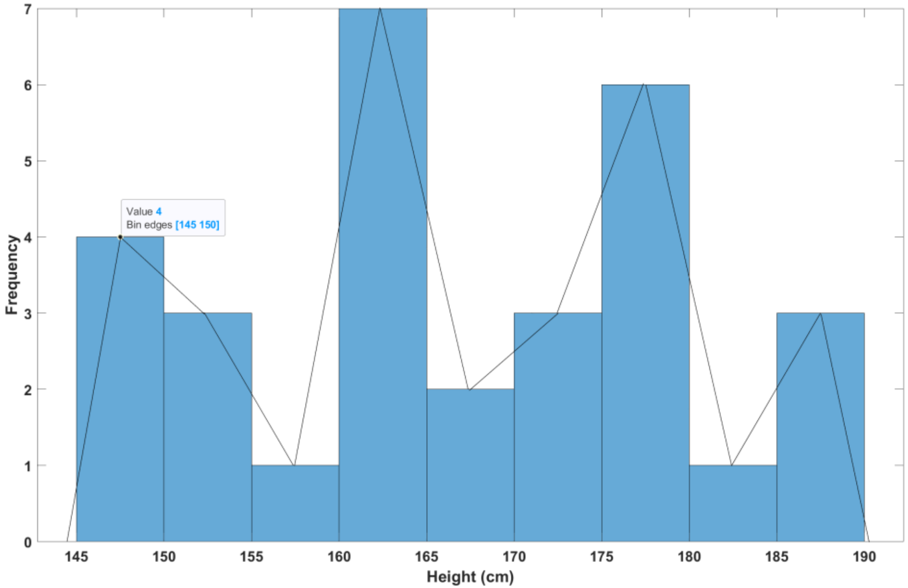
Exploratory Data Analysis & Visulisation – Histograms

Histogram is used to represent the frequency distribution of a sample of **continuous** data. In a histogram, the base of each block or column is the class interval on the x-axis. If the class intervals are of *equal width*, the height of the columns are *proportional* to the frequencies.

Example 11

Construct a histogram for the sample from Example 3.

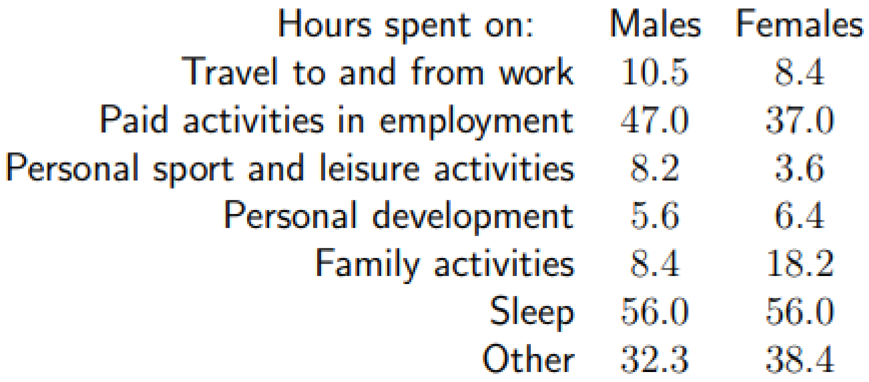
Solution



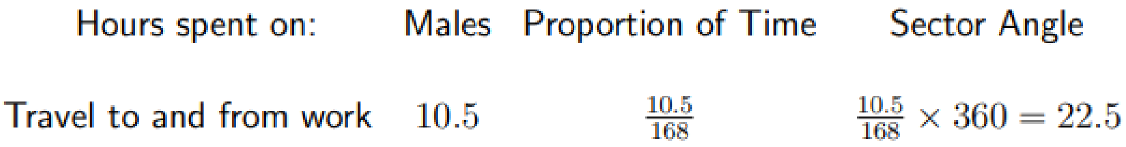
The **approximate shape** of the **distribution** of data is indicated by a **frequency polygon** which is formed by joining the mid-points of the tops of the blocks forming the histogram with straight lines.

Exploratory Data Analysis & Visulisation – Pie Charts

A **pie chart** is simply a circular diagram where the circle is divided into sectors and the angles of the sectors are proportional to the quantity represented. Since the total area of the circle is fixed, pie charts are considered to be useful for representing **proportions of a total**.



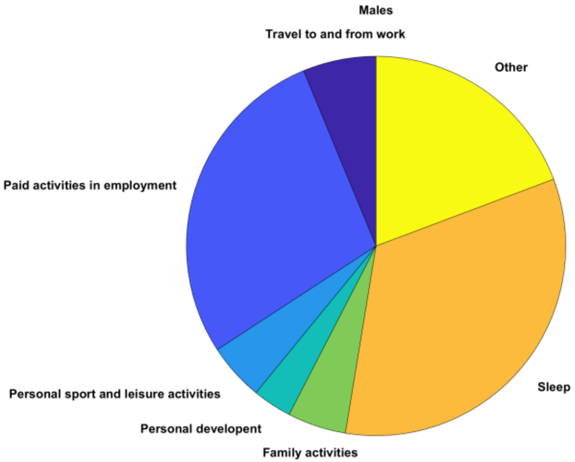
Hours spent on:	Males	Females
Travel to and from work	10.5	8.4
Paid activities in employment	47.0	37.0
Personal sport and leisure activities	8.2	3.6
Personal development	5.6	6.4
Family activities	8.4	18.2
Sleep	56.0	56.0
Other	32.3	38.4

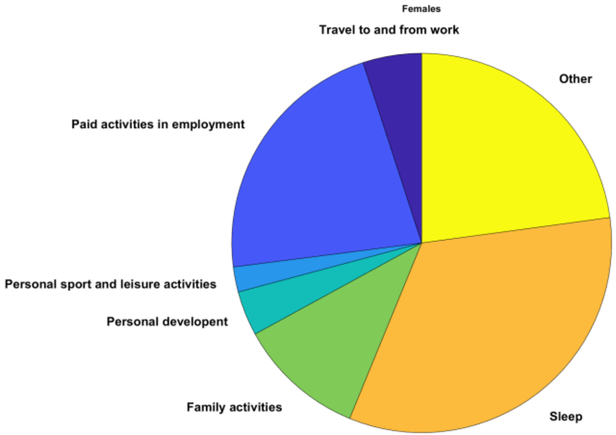


Hours spent on:	Males	Proportion of Time	Sector Angle
Travel to and from work	10.5	$\frac{10.5}{168}$	$\frac{10.5}{168} \times 360 = 22.5$

Exploratory Data Analysis & Visulisation – Pie Charts

A **pie chart** is simply a circular diagram where the circle is divided into sectors and the angles of the sectors are proportional to the quantity represented. Since the total area of the circle is fixed, pie charts are considered to be useful for representing **proportions of a total**.





Exploratory Data Analysis & Visulisation – The Box-and-Whisker Diagram

Boxplots are useful to summarise the distribution of a data sample as an alternative to the histogram. The diagram is constructed as follows.

The Box

- The left-hand vertical is placed at the lower hinge;
- The right-hand vertical is placed at the upper hinge;
- The vertical in the box is placed at the median.

The Whiskers

- Find the greatest value which is within 1.5 mid-spread of the upper hinge;
- Find the least value which is within 1.5 mid-spread of the lower hinge;
- Connect the greatest and least values to the box by means of dashed lines (sometimes solid lines with a bar).

The Outlying Values

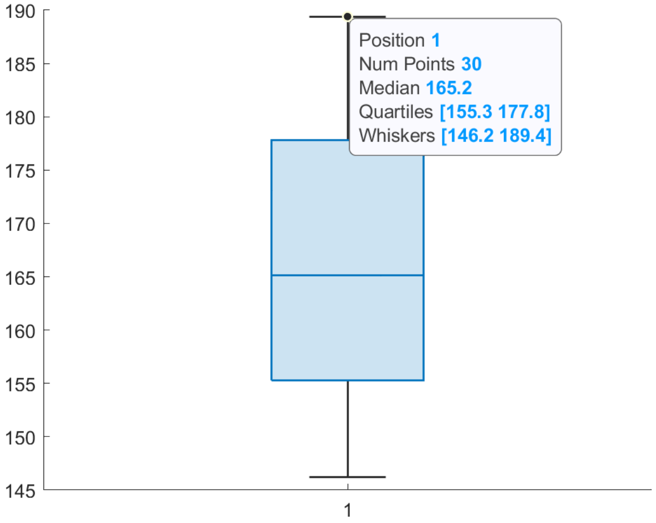
- Mark as large dots any values which are more than 1.5 mid-spreads from the hinges.

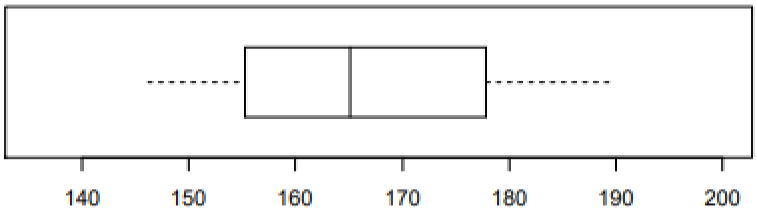
Exploratory Data Analysis & Visulisation – The Box-and-Whisker Diagram

Example 12

Construct a box-and-whisker diagram representing the sample from Example 3. Does the box-and-whisker diagram tell you that the data set that you are working with is symmetrical?

Solution





Exploratory Data Analysis & Irregularities – Outliers

Irregularities refers to something in a sample distribution that departures from normality, e.g. extreme values lie well outside the range of most of a sample, asymmetric distribution, and multiple peaks.

Outliers can be extremely important for the following reasons:

- They can have misleading effect on statistics such as the mean and standard deviation;
- Their occurrence may be due to incorrect observation, measurement or recording. In this case it is often possible to correct the data;
- Their presence can induce a false skewness (see **Section** 3.2) in a data set;
- They may actually be members of a population not under consideration. E.g, data on road traffic speeds collected at a point in a highway may be intended to provide information on the speeds of motor vehicles but the data are likely to include some observations for bicycles and other slower-moving types of traffic.

There is no standard precise definition, but some simple criteria do exist which may be used to detect outliers and accept or reject outliers.

Exploratory Data Analysis & Irregularities – Outliers

Two criteria for the detection of outliers are given below.

Criterion 1

For variables where the distribution has a "**normal**" shape, we can expect only about 1 in 1000 observations to lie more than **3.3 standard deviations** away from the mean. So we could treat any value further than 3.3 standard deviations from the mean as an **outlier**.

An observation x would be regarded as an outlier if

$$\left|\frac{x-\bar{x}}{s}\right| > 3.3$$

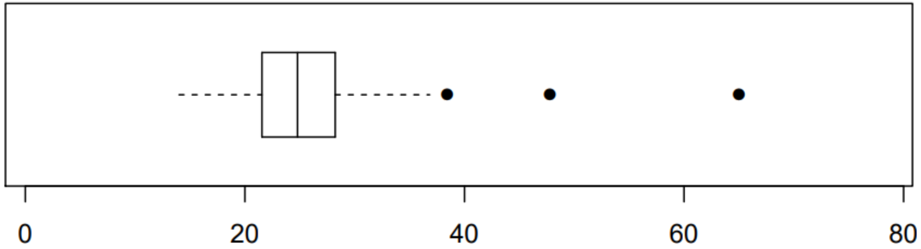
Criterion 1 essentially implies that a value has less than 1 in a 1000 of chance of occurring naturally outside the range defined by 3.3 standard deviations from the mean. Note that the property that 0.1% of observations are more than 3.3 standard deviations from the mean really refers to the population mean and standard deviation. Here, we have used $\bar{x}$ and s the sample mean and standard deviation. However, this will be a reasonable approximation in reasonably **large samples**, say $\boldsymbol{n > 30}$.

Exploratory Data Analysis & Irregularities – Outliers

Two criteria for the detection of outliers are given below.

Criterion 2

Outlier observations can be defined to be more than 1.5 mid-spreads (or IQRs) from the hinges (or quartiles). **Extreme outliers** can be defined to be more than 3 mid-spreads from the hinges.

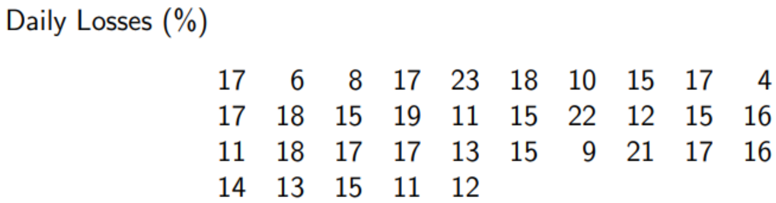


While all values classified as outliers should be investigated, this is particularly true of those classified as extreme outliers.

Exploratory Data Analysis & Irregularities – Outliers

Example 13

Manufacturing processes generally result in a certain amount of wasted material. For reasons of cost, companies need to keep such wastage to a minimum. The following data were gathered over a five-week period by a manufacturing company whose production lines run seven days per week. The numbers given represent the percentage wastage of the amount of material used in the manufacturing process.



Daily Losses (%)

17	6	8	17	23	18	10	15	17	4
17	18	15	19	11	15	22	12	15	16
11	18	17	17	13	15	9	21	17	16
14	13	15	11	12					

1. Find the mean and standard deviation of the percentage losses of material over the five-week period.

2. Assuming that the losses are roughly normally distributed, apply an appropriate criterion to decide whether any of the losses are smaller or larger than might be expected by chance.

Solution

Exploratory Data Analysis & Irregularities – Outliers

Example 13 Continued…

Solution

Using Criterion 1, we will treat any value further than 3.3 standard deviation from the mean as an outlier.

$\bar{x} = 14.69$ $s = 4.22$

Apply $\left|\frac{x_0 - \bar{x}}{s}\right|$ to the values of the table.

0.55	-2.06	-1.58	0.55	1.97	0.79	-1.11	0.07	0.55	-2.53
0.55	0.79	0.07	1.02	-0.87	0.07	1.73	-0.64	0.07	0.31
-0.87	0.79	0.55	0.55	-0.40	0.07	-1.35	1.50	0.55	0.31
-0.61	-0.40	0.07	-0.87	-0.64					

All above values are smaller than 3.3, so the daily losses are all within the range indicated by chance variation.

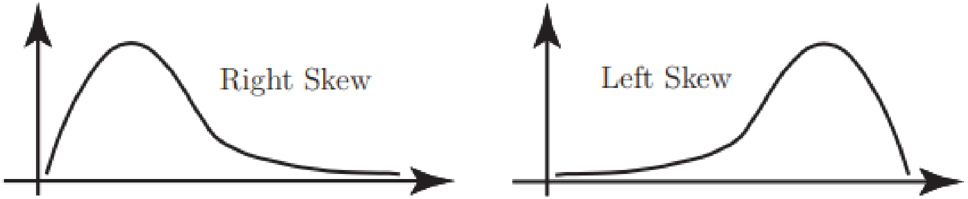
Exploratory Data Analysis & Irregularities – Skewness

- The regions on either side of the distribution where the frequencies die out are called the ***left*** and ***right*** **tails** of the distribution.

- In a symmetric (or Normal) distribution, the left and right tails are like mirror images of each other. Symmetric unimodal distributions are common but there are exceptions.

- The distributions of some variables tend to be asymmetric (i.e. **skewed**), with one tail longer than the other.

- If the longer tail is on the right/left, this is called *right* (*positive)/left (negative) skew*.



A skewed distribution cannot be represented purely by two numbers, say the mean and standard deviation. **Five-number summary statistics** can be used in this case to describe such a distribution.

Exploratory Data Analysis & Irregularities – Multimodal Distribution

A distribution with more than one peak is called a **multimodal distribution**. A distribution with exactly two peaks is called a **bimodal distribution**. Such distributions can be very difficult to summarise. In this case, the stem-and-leaf plot (not covered in this module and please refer to [1] in Further Readings) is more informative than the box-and-whisker plot.

Further Readings

[1] HELM Workbook 36 Descriptive Statistics

https://nucinkis-lab.cc.ic.ac.uk/HELM/HELM_Workbooks_36-40/WB36-all.pdf

Queen Mary
University of London