Data Analysis Theories

Data Analysis Theories

- **Exploratory Data Analysis**

  exploring the information at hand for clues that hint at a bigger meaning

- **Confirmatory Data Analysis**

  CDA focuses on utilizing traditional statistical tools such as *confidence*, *inference, and significance* to evaluate the data and challenge any assumptions we made during EDA.

- **Grounded Theory**

  involves the collection and analysis of information at the same time (less popular)

EDA vs CDA

	Exploratory Data Analysis (EDA)	Confirmatory Data Analysis (CDA)
Reasoning Type	Inductive	Deductive
Goal	Pattern Recognition and Hypothesis generation	Estimation, Modeling, Hypothesis testing
Applied Data	Observation Data (data collected without well-defined hypothesis)	Experimental data (data collected through formally designed experiments)
Techniques	Descriptive Statistics, Data Visualization, Clustering Analysis, Process Mining…	Traditional statistical techniques of inference, significance, and confidence
Advantages	• No assumptions required • Promotes deeper understanding of the data	• Precise • Well-established theory and methods
Disadvantages	• No conclusive answers • Difficult to avoid bias produced by overfitting	• Required unrealistic assumptions • Difficult to notice unexpected results

Exploratory Data Analysis (EDA)

- What is Exploratory Data Analysis?
- Why do we do EDA?
- What are the steps in EDA?
- What are the tools used for EDA?
- What happens if we don’t do EDA?

EDA: Introduction

- Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

- By completing the EDA you will have many plots, heat-maps, frequency distribution, graphs, correlation matrix along with the hypothesis by which any individual can understand what your data is all about and what insights you got from exploring your data set.

- By doing this, one can get to know whether the selected features are good enough to model, are all the features required, are there any correlations based on which we can either go back to the Data Preprocessing step or move on to modeling.

EDA: Highlights

a) Description of data
b) Handling missing data/Outliers
c) Understanding relationships and new insights through plots

EDA (a): Description of data

- We need to know the different kinds of data and other statistics of our data before we can move on to the other steps.

- For numeric data, the result's index will include count, mean, std, min, max;
  as well as lower, 50 and upper percentiles.
  - SUMMARY STATISTICS
  - DESCRIPTIVE STATISTICS

Describing a numeric `Series`.

```
>>> s = pd.Series([1, 2, 3])
>>> s.describe()
count    3.0
mean     2.0
std      1.0
min      1.0
25%      1.5
50%      2.0
75%      2.5
max      3.0
dtype: float64
```

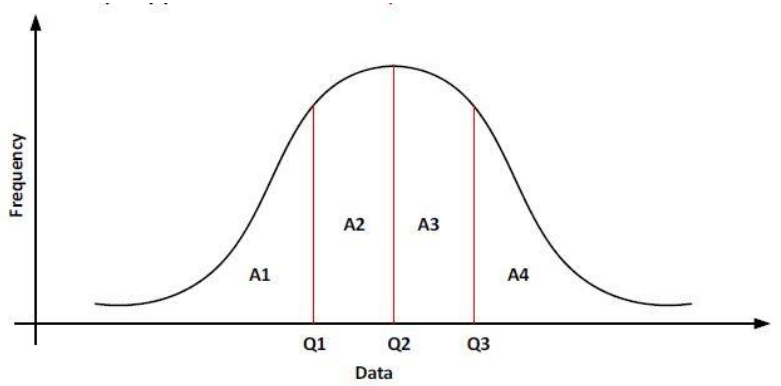
- For object/Categorical data (e.g. strings or timestamps), the result's index will include count, unique, top, and freq.

Describing a categorical `Series`.

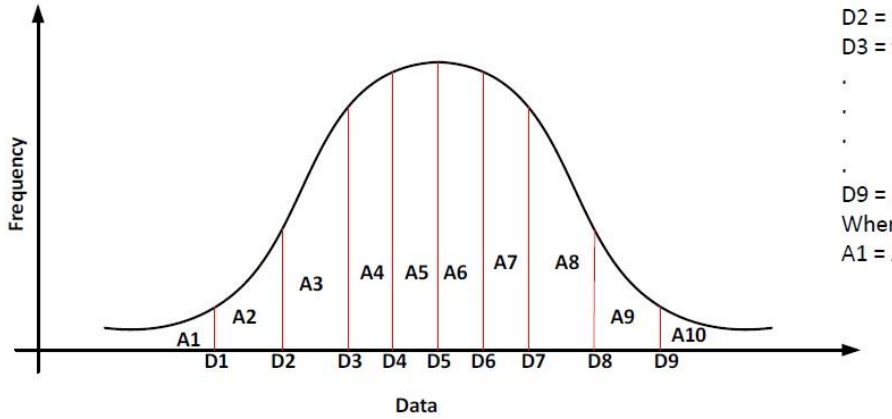
```
>>> s = pd.Series(['a', 'a', 'b', 'c'])
>>> s.describe()
count     4
unique    3
top       a
freq      2
dtype: object
```

Quartiles, Deciles, Percentiles

- **Median** is the middle point in the axis frequency distribution curve, and it divides the area under the curve **into two equal parts**, having the same area in the left, and in the right.
- The area under the curve may be divided **into four equal parts**, called as **quartiles**.
- In the same procedure divide the area for **ten equally pieces** and each area is called **deciles**.
- Finally where divided the area for **hundred equally pieces** and each area is called **percentiles**.



Q1 = first quartile
Q2 = second quartile
Q3 = third quartile
Where:
A1 = A2 = A3 = A4



D1 = first decile
D2 = second decile
D3 = third decile
.
.
.
.
D9 = ninthe decile
Where:
A1 = A2 = A3 = A4= . . . .=A10

Quartiles, Deciles, Percentiles (contd.)

The same procedure for division is done for finding percentiles for any frequency distributed curve

P1 = first percentile

P2 = second percentile

P3 = third percentile

.

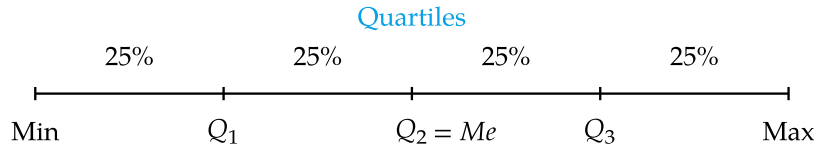
.

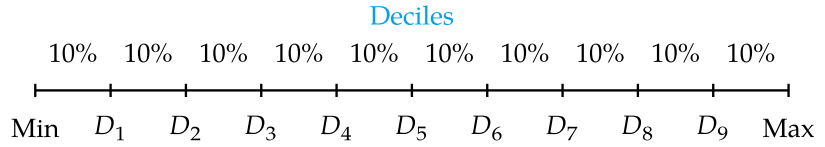
.

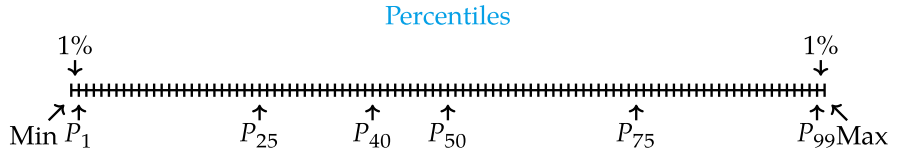
P99 = ninety ninth percentile

Where:

A1 = A2 = A3 = A4 = ...= A100







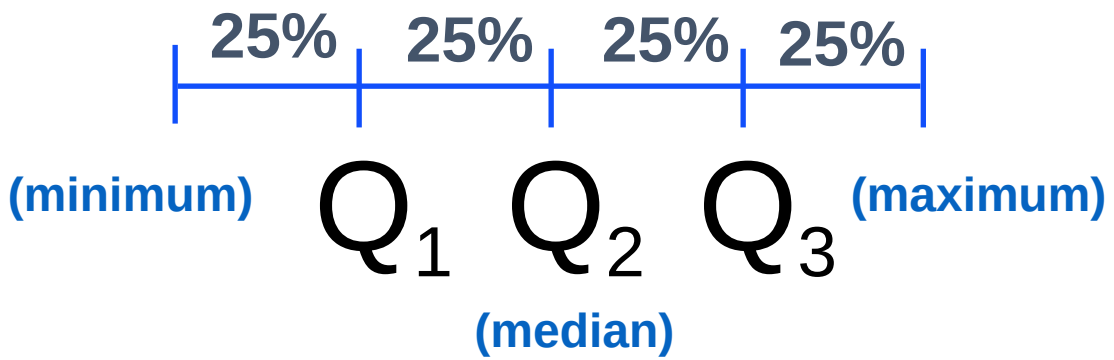
To find the quartiles, or deciles, or percentiles we follow the same procedure to find the median. Arrangement the data in ascending form only.

If numbering arrangement of quartiles, deciles, and percentiles is fraction then its value is for the number greater than it, if true number the value is the mean of its and the greater numbers.

Quartiles

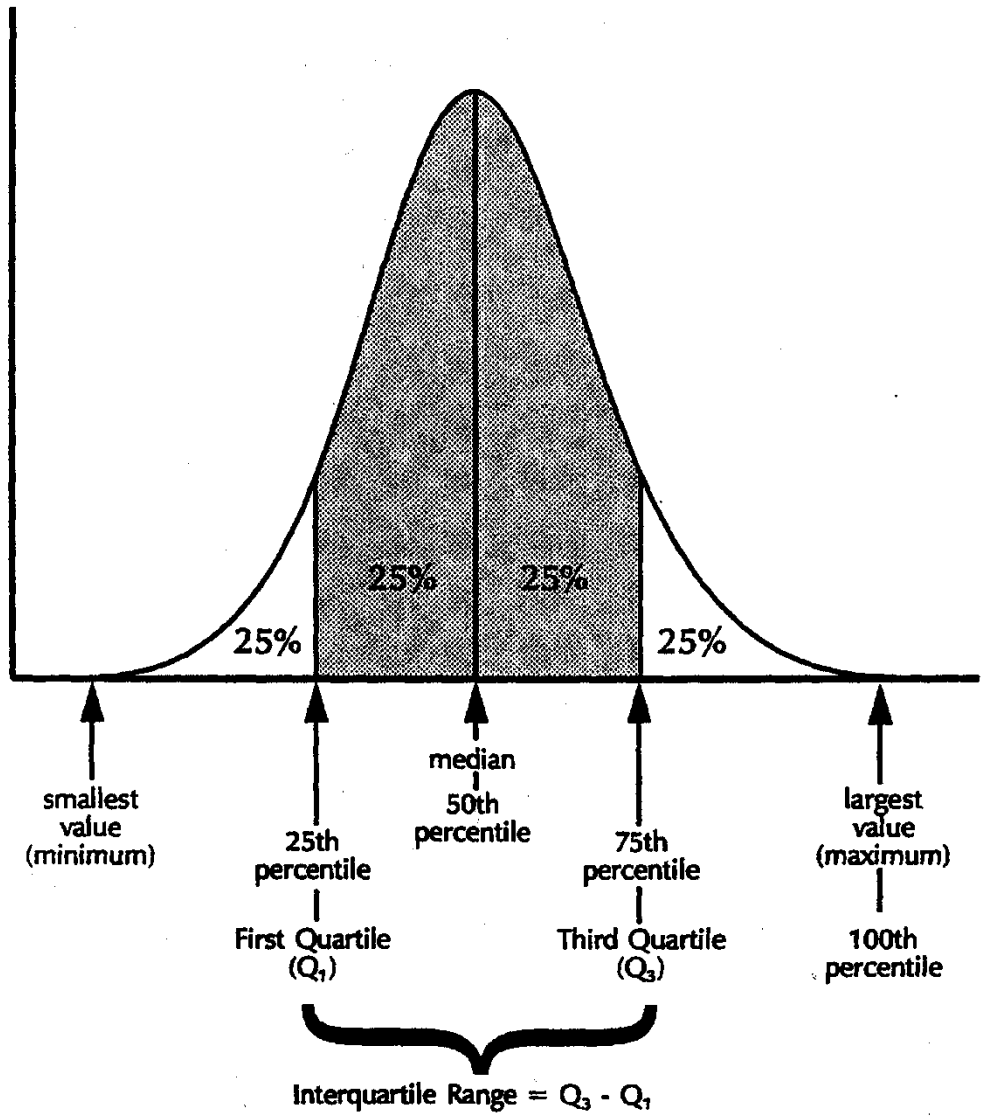
Q_1, Q_2, Q_3

divides ranked scores into four equal parts



Inter quartile Range

- The inter quartile range is Q_3-Q_1
- 50% of the observations in the distribution are in the inter quartile range.
- The right figure shows the interaction between the quartiles, the median and the inter quartile range.



An Example

Sample Number	Unsorted Values
1	25
2	27
3	20
4	23
5	26
6	24
7	19
8	16
9	25
10	18
11	30
12	29
13	32
14	26
15	24
16	21
17	28
18	27
19	20
20	16
21	14

An Example

Sample Number	Unsorted Values	Ranked Values
1	25	14
2	27	16
3	20	16
4	23	18
5	26	19
6	24	20
7	19	20
8	16	21
9	25	23
10	18	24
11	30	24
12	29	25
13	32	25
14	26	26
15	24	26
16	21	27
17	28	27
18	27	28
19	20	29
20	16	30
21	14	32

An Example

Sample Number	Unsorted Values	Ranked Values	
1	25	14	Minimum
2	27	16	
3	20	16	
4	23	18	
5	26	19	LQ or Q_1
6	24	20	
7	19	20	
8	16	21	
9	25	23	
10	18	24	Md or Q_2
11	30	24	
12	29	25	
13	32	25	
14	26	26	
15	24	26	UQ or Q_3
16	21	27	
17	28	27	
18	27	28	
19	20	29	
20	16	30	Maximum

Q:
How to find
those Q1, Q2, Q3?

Questions

1. Can we find the position and values of quartiles (Q1, Q2, Q3), deciles and percentiles?
2. Given a value from a dataset, can we find which quartile it would fall?

How to find Quartiles?

To find the quartiles of a dataset or sample, follow the step-by-step guide below.

1.Count the number of observations in the dataset (n).

2.Sort the observations from smallest to largest.

3.Find the (value of) first quartile ($Q_{t=1}$):

1. Calculate its position, $q_t = n * (t / 4)$.
2. If q_t is an integer,
   then the first quartile is the mean of the numbers at positions q_t and $q_t + 1$.
3. If q_t is **not** an integer,
   then round it up. The value at this position is the first quartile, $Q_{t=1}$.

4. Find the second quartile ($Q_{t=2}$) following the process described at 3 [Remember, here: $q_t = n * (2 / 4)$.]
5. Find the second quartile ($Q_{t=3}$) following the process described at 3 [Remember, here: $q_t = n * (3 / 4)$.]

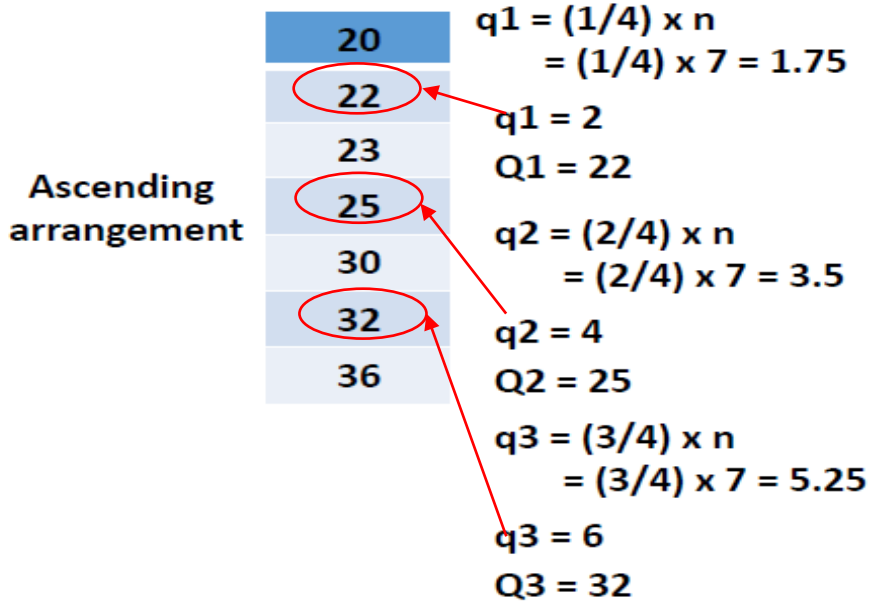
QUARTILES finding example

Example:

find the quartiles Q1, Q2, and Q3 of the following data 20, 30, 25, 23, 22, 32, 36

Solution:

Arrange data in ascending form, and n = 7 odd number



At a glance:

To find Q_t, we first find its position,

$q_t = \frac{t}{4}\, n$, where

n = total data sample

$t = t^{th}$-quartile to find

If q_t is integer, $Q_t = [val(q_t) + val(q_t + 1)]/2$

Else, $Q_t = val\,(\text{round}(q_t))$

QUARTILES finding example

Example:

find the quartiles Q1, Q2, and Q3 of the following data 20, 30, 25, 23, 22, 32, 36, 18

Solution:

Arrange data in ascending form, and n = 8 even number

Ascending arrangement
18
20
22
23
25
30
32
36

q1 = (1/4) x n
= (1/4) x 8 = 2

q1 = mean of (2), and (3)

Q1 = (20+22)x(1/2) = 21

---

q2 = (2/4) x n
= (2/4) x 8 = 4

q2 = mean of (4), and (5)

Q2 = (23+25)x(1/2) = 24

---

q3 = (3/4) x n
= (3/4) x 8 = 6

q3 = mean of (6), and (7)

Q3 = (30+32)x(1/2) = 31

DECILES

Example:

find the desiles D1, D5, and D8 of the following data 20, 30, 25, 23, 22, 32, 36

Solution:

Arrange data in ascending form, and n = 7 odd number

Ascending arrangement

20
22
23
25
30
32
36

d1 = (1/10) x n
= (1/10) x 7 = 0.7

d1 = 1

D1 = 20

d5 = (5/10) x n
= (5/10) x 7 = 3.5

d5 = 4

D5 = 25

d8 = (8/10) x n
= (8/10) x 7 = 5.6

d8 = 6

D8 = 32

At a glance:

To find D_t, we first find its position,

$$d_t = \frac{t}{10}\, n \text{ , where}$$

n = total data sample

$t = t^{th}$-decile to find

If d_t is integer, $D_t = [val(d_t) + val(d_t + 1)]/2$

Else, $D_t = val\ (round(d_t))$

Example:

find the desiles D1, D5, and D8 of the following data 20, 30, 25, 23, 22, 32, 36, 18

Solution:

Arrange data in ascending form, and n = 8 even number

Ascending arrangement

18
20
22
23
25
30
32
36

$d1 = (1/10) \times n$

$= (1/10) \times 8 = 0.8$

$d1 = 1$

$D1 = 18$

$d5 = (5/10) \times n$

$= (5/10) \times 8 = 4$

$d5 =$ mean of (4), and (5)

$D5 = (23+25) \times (1/2) = 24$

$d8 = (8/10) \times n$

$= (8/10) \times 8 = 6.4$

$d8 = 7$

$D8 = 32$

PERCENTILES

Example:

find the percentiles P8, P50, and P85 of the following data 20, 30, 25, 23, 22, 32, 36

Solution:

Arrange data in ascending form, and n = 7 odd number

Ascending arrangement

20
22
23
25
30
32
36

p8 = (8/100) x n
= (8/100) x 7 = 0.56

p8 = 1

P8 = 20

---

p50 = (50/100) x n
= (50/100) x 7 = 3.5

p50 = 4

P50 = 25

---

p85 = (85/100) x n
= (85/100) x 7 = 5.95

p85 = 6

P85 = 32

At a glance:

To find P_t, we first find its position,

$$p_t = \frac{t}{100}\, n \text{ , where}$$

n = total data sample

$t = t^{th}$-percentile to find

If p_t is integer, $P_t = [val(p_t) + val(p_t + 1)]/2$

Else, $P_t = val\ (\text{round}(p_t))$

Example:

find the percentiles P8, P50, and P85 of the following data 20, 30, 25, 23, 22, 32, 36, 18

Solution:

Arrange data in ascending form, and n = 8 even number

Ascending arrangement

18
20
22
23
25
30
32
36

p8 = (8/100) x n
= (8/100) x 8 = 0.64

p8 = 1

P8 = 18

---

p50 = (50/100) x n
= (50/100) x 8 = 4

p50 = mean of (4), and (5)

P50 = (23+25)x(1/2) = 24

---

p85 = (85/100) x n
= (85/100) x 8 = 6.8

p85 = 7

P85 = 32

Questions

1. Can we find the position and values of quartiles (Q1, Q2, Q3), deciles and percentiles?
2. Given a value from a dataset, can we find which quartile/decile/percentile it would fall?

Note: We call the r^{th} *percentile* the value such that r percent of the data fall at or below that value.

Example

If you score in the 75^{th} percentile, then 75% of the population scored lower than you.

Question

Suppose the test scores were

22, 34, 68, 75, 79, 79, 81, 83, 84, 87, 90, 92, 96, and 99

If your score was the 75, in what percentile did you score?

Solution

There were 14 scores reported and there were 4 scores at or below yours. We divide

$$\frac{4}{14} \; 100\% = 29$$

So you scored in the 29^{th} percentile.

Remember?

$$p_t = \frac{t}{100}\, n$$

$$t \times n = p_t \times 100$$

$$t = \frac{p_t}{n} \times 100$$

Here,
$t = t^{th}$*-percentile*
Position (75), $p_t = 4$
Total Sample, $n = 14$

Some More Range

Interquartile Range: $Q_3 - Q_1$

Semi-interquartile Range: $\frac{Q_3 - Q_1}{2}$

Midquartile: $\frac{Q_1 + Q_3}{2}$

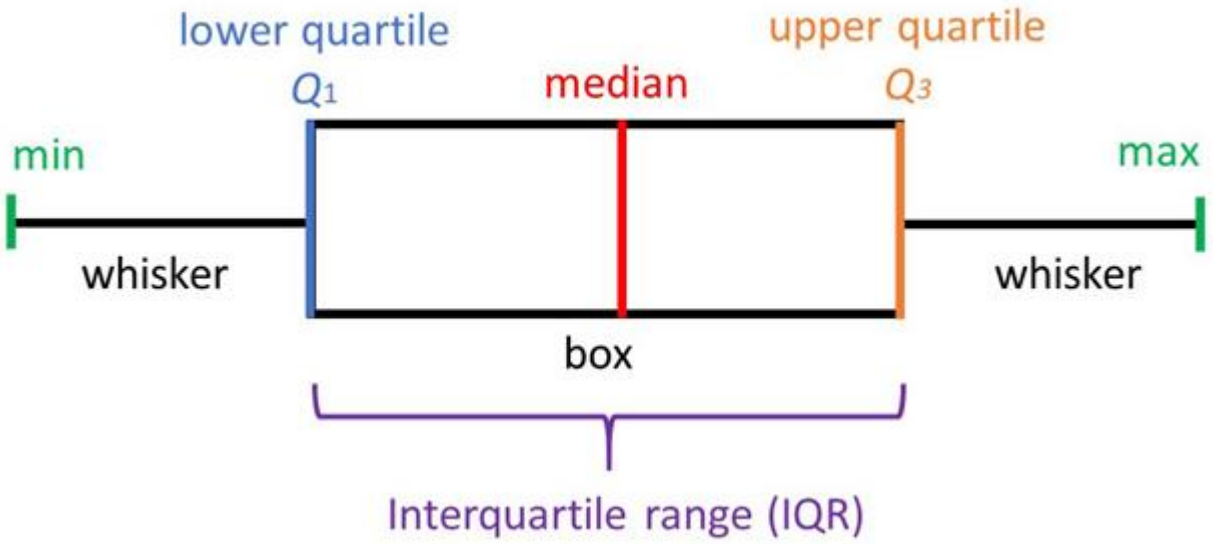
10–90 Percentile Range: $P_{90} - P_{10}$

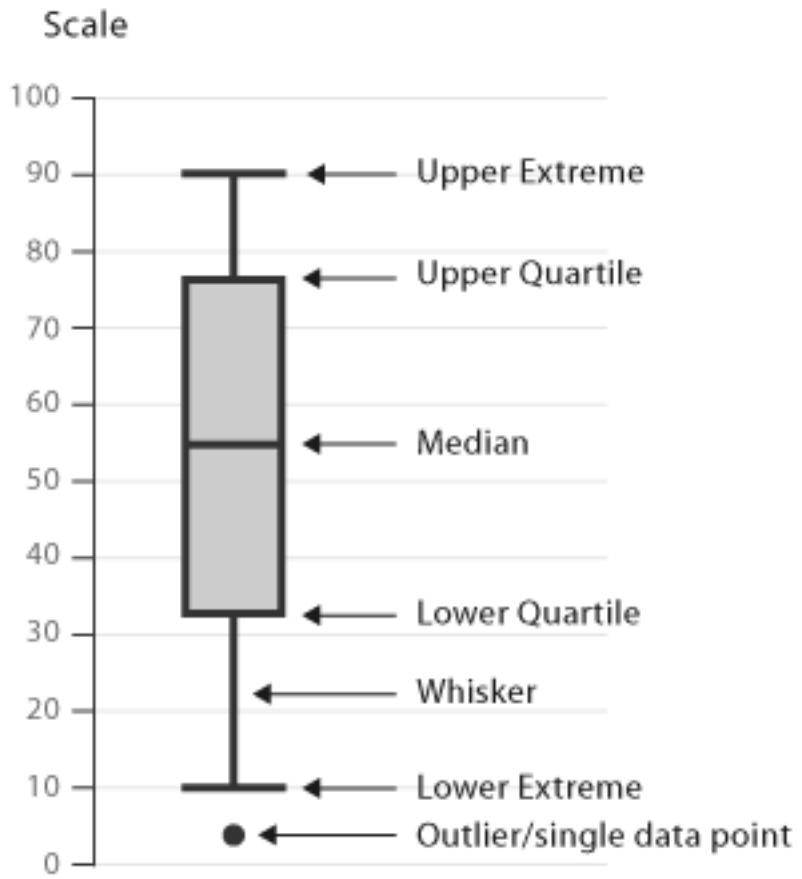
EDA (b): Handling missing data/Outliers

- Some of the methods for detecting and handling outliers:
    - Box Plot
    - Scatter plot
    - Z-score
    - IQR(Inter-Quartile Range)

Boxplots

Box-and-Whisker Diagram

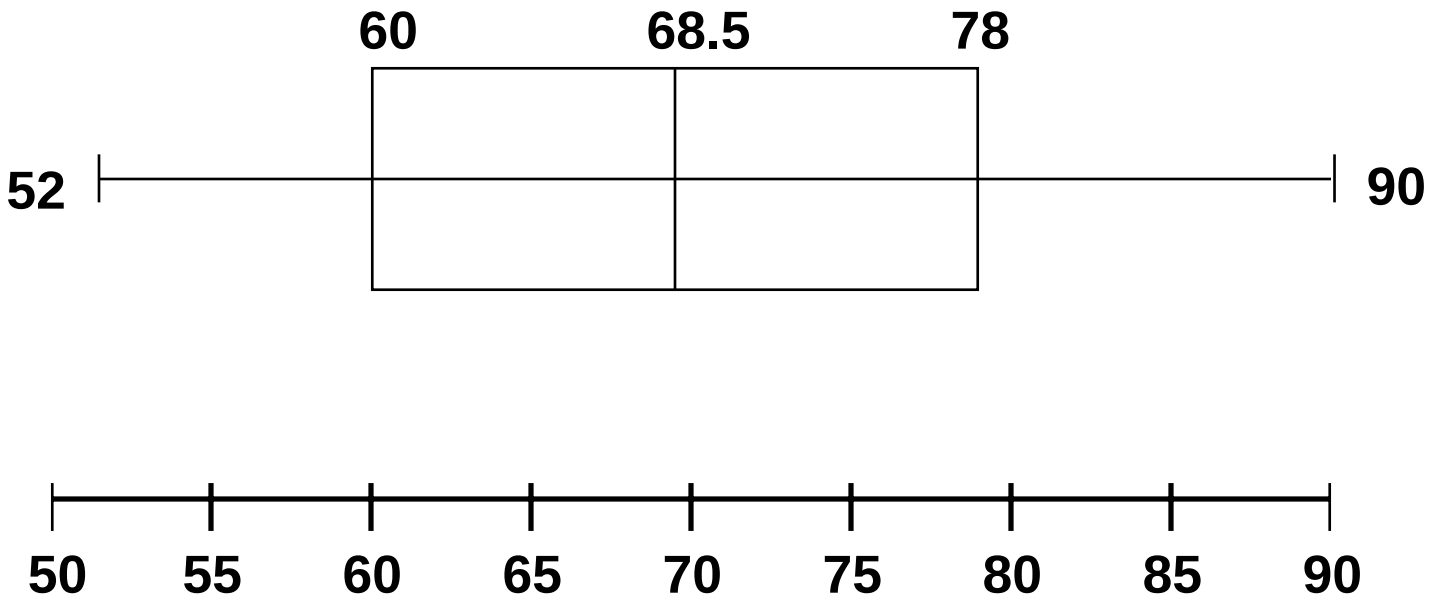




The whiskers extend from either side of the box.
The whiskers represent the ranges for the bottom 25% and the top 25% of the data values, excluding outliers.

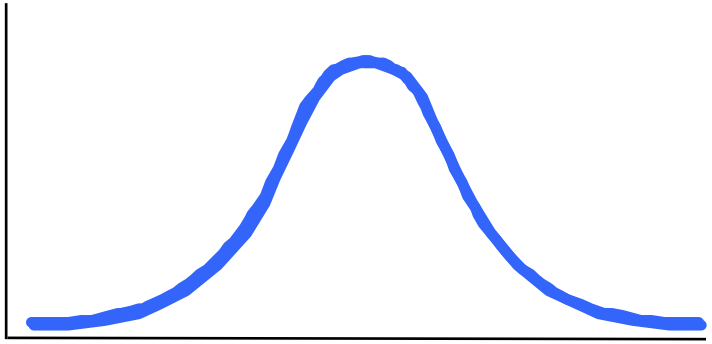
Boxplots: Numeric Example

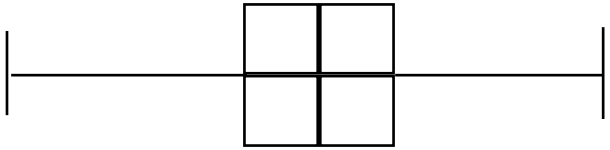
Box-and-Whisker Diagram



Boxplot of Pulse Rates (Beats per minute) of Smokers

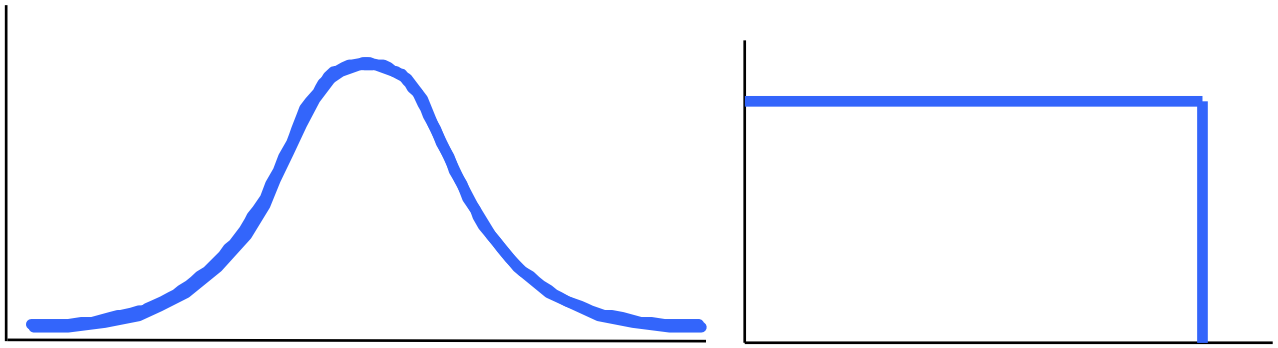
Boxplots

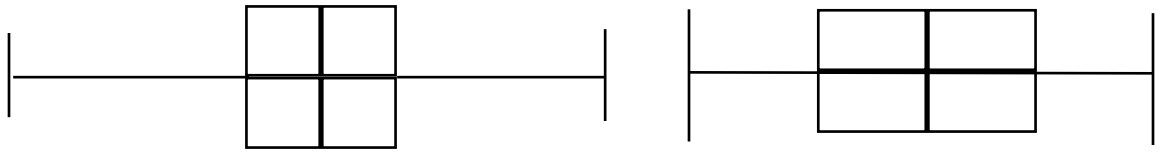




Normal

Boxplots

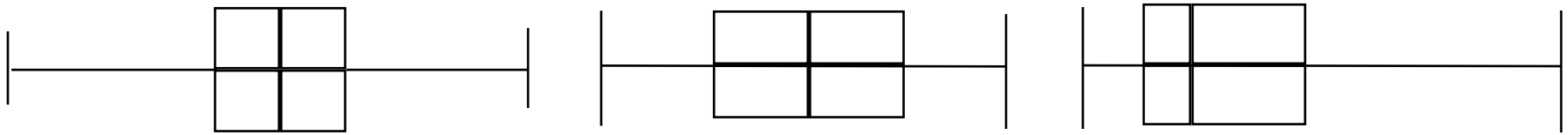




Normal

Uniform

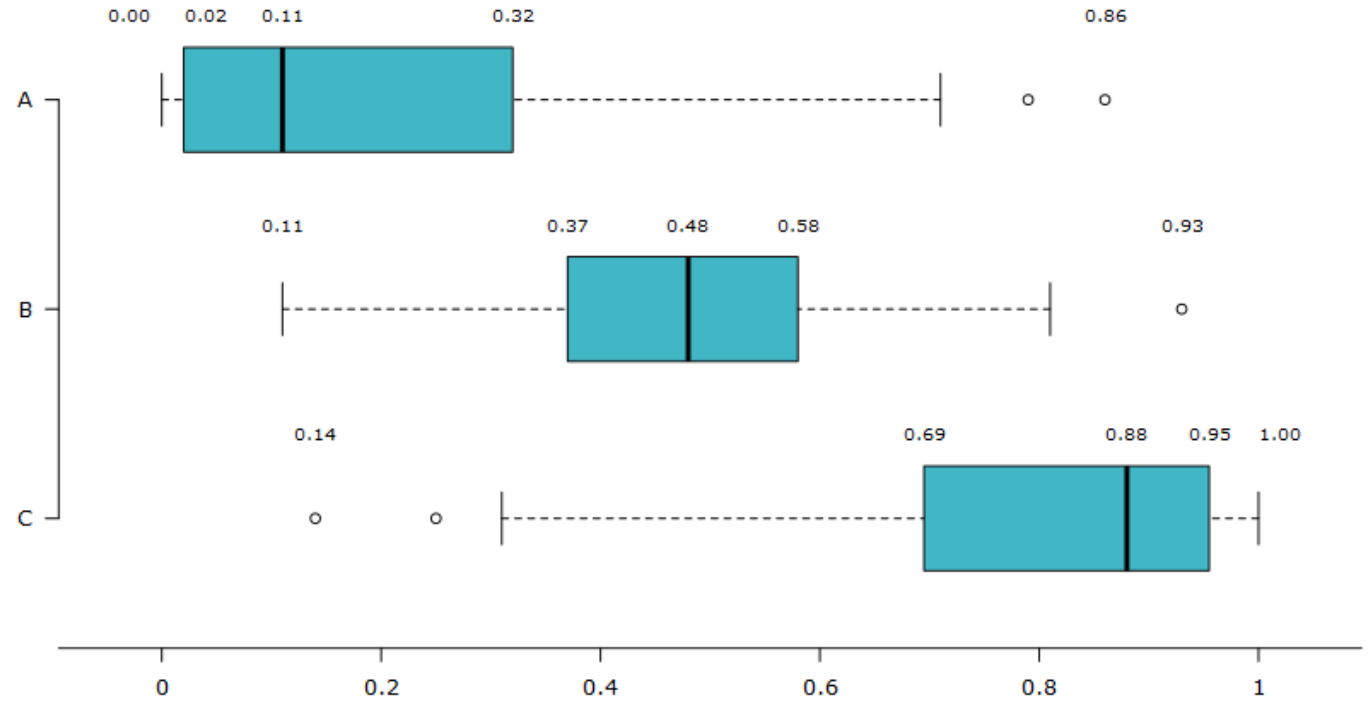
Boxplots



Normal **Uniform** **Skewed**

Boxplot: Detecting Outlier

Chart 4.5.2.1
Box and whisker plots and five-number summaries of distributions A, B and C



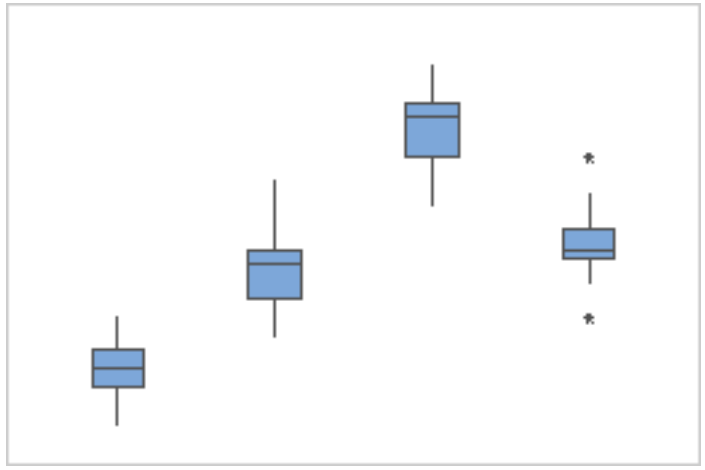
Interpreting Boxplot

Skewed Data

Right-skewed

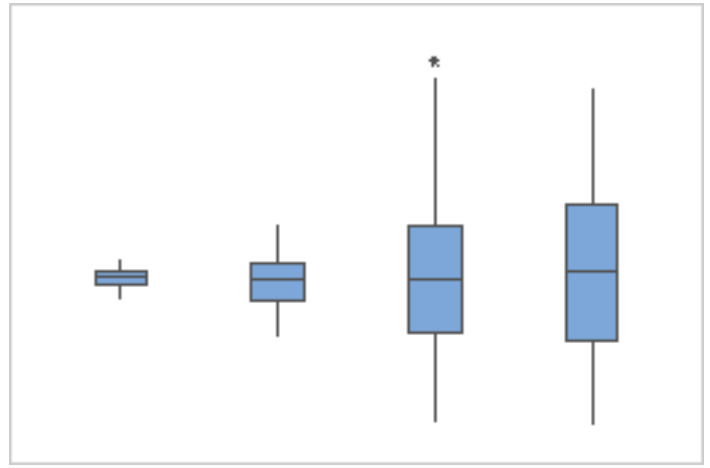
Left-skewed

Centers



Look for differences between the centers of the groups. For example, the following boxplot shows the thickness of wire from four suppliers. The median thicknesses for some groups seem to be different.

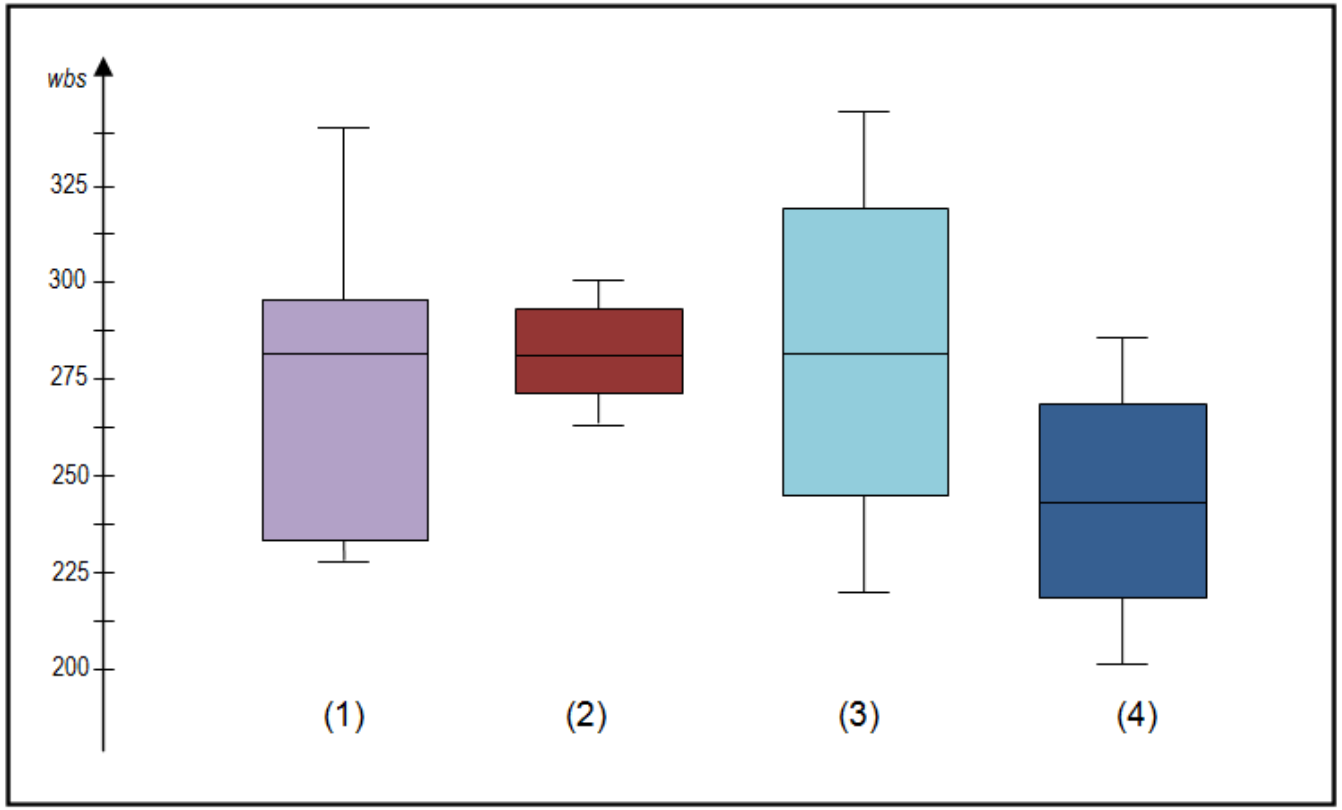
Spread



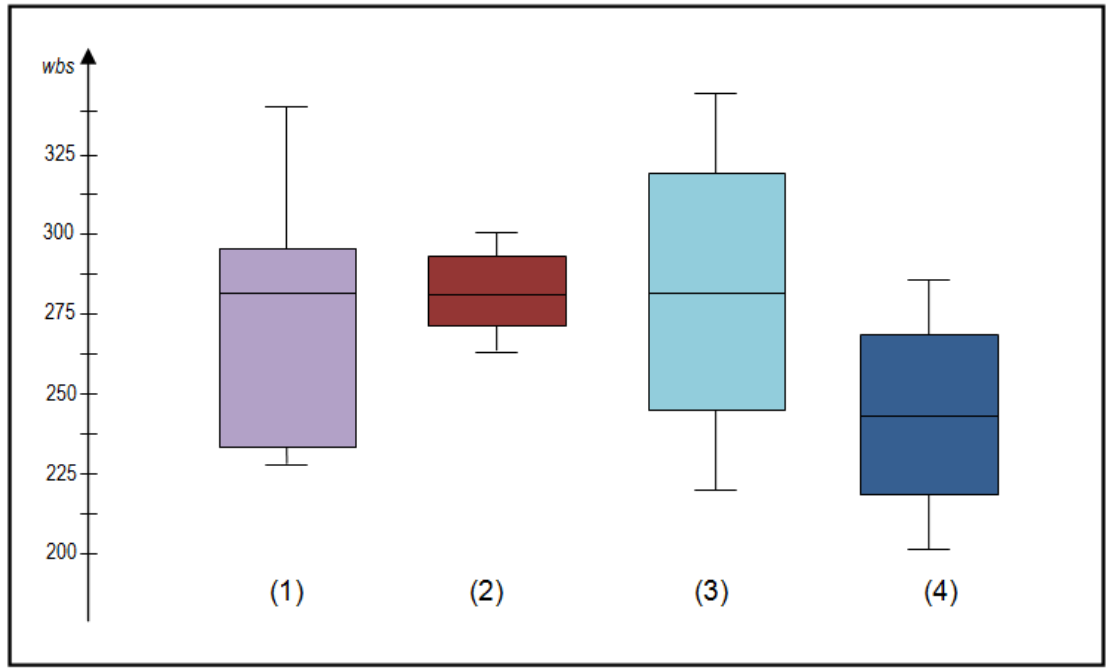
Look for differences between the spreads of the groups. For example, the following boxplot shows the fill weights of cereal boxes from four production lines. The median weights of the groups of cereal boxes are similar, but the weights of some groups are more variable than others.

Task: Interpreting Boxplots (contd.)

Ratings over a subject is taken from 4-student groups, and the boxplot of the responses are given here.



What are your observations after watching these boxplots?



Some general observations about box plots

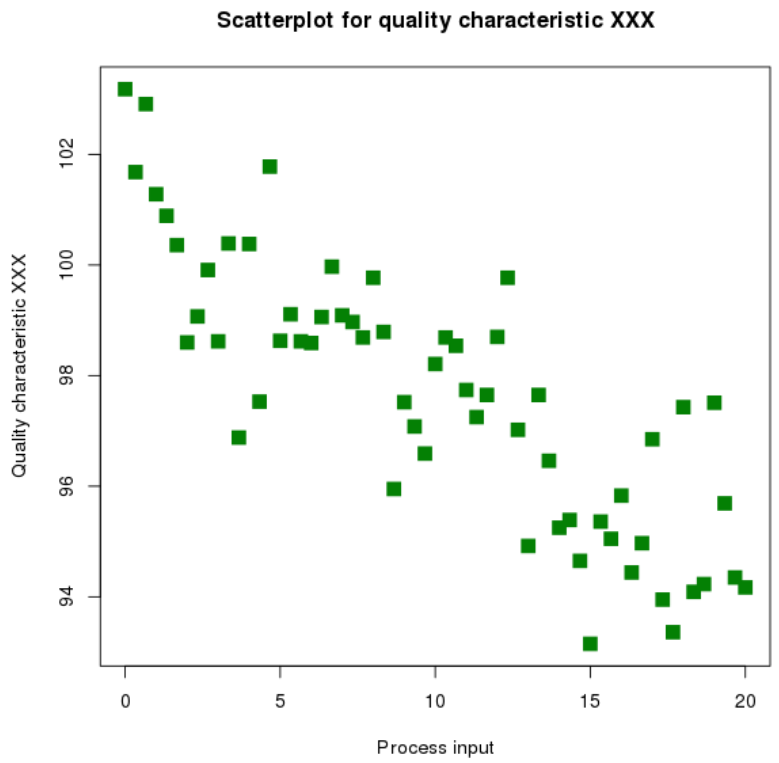
- **The box plot is comparatively short** – see example (2). This suggests that overall students have a high level of agreement with each other.
- **The box plot is comparatively tall** – see examples (1) and (3). This suggests students hold quite different opinions about this aspect or sub-aspect.
- **One box plot is much higher or lower than another** – compare (3) and (4) – This could suggest a difference between groups. For example, the box plot for boys may be lower or higher than the equivalent plot for girls. Follow this up by looking at the *Items at a Glance* reports.
- **Obvious differences between box plots** – see examples (1) and (2), (1) and (3), or (2) and (4). Any obvious difference between box plots for comparative groups is worthy of further investigation in the *Items at a Glance* reports.
- Your school box plot is much higher or lower than the national reference group box plot. This also suggests an area of difference that could be explored further in the *Items in Detail* reports and through consultation.
- **The 4 sections of the box plot are uneven in size** – See example (1). This shows that many students have similar views at certain parts of the scale, but in other parts of the scale students are more variable in their views. The long upper whisker in the example means that students views are varied amongst the most positive quartile group, and very similar for the least positive quartile group. The *Items in Detail* reports can be used to explore this further.
- **Same median, different distribution** – See examples (1), (2), and (3). The medians (which generally will be close to the average) are all at the same level. However the box plots in these examples show very different distributions of views.
  It always important to consider the pattern of the whole distribution of responses in a box plot.

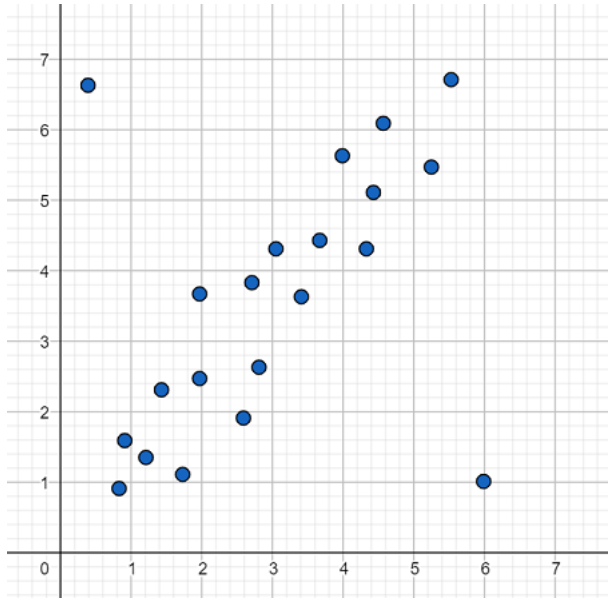
Scatter Plot

- Scatter plots are the **graphs that present the relationship between two variables in a data-set**.

- It represents data points on a two-dimensional plane or on a Cartesian system.

- The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis.





Z-score based Outlier Detection

- While calculating the Z-score we re-scale and center the data and look for data points that are too far from zero.
- These data points which are way too far from zero will be treated as the outliers.
- In most of the cases, a **threshold of 3 or -3 is used** i.e if the Z-score value is greater than or less than 3 or -3 respectively, that data point will be identified as outliers.

$$v' = \frac{v - \mu}{\sigma}$$

IQR based Outlier Detection

IQR = Q3 – Q1.

- Once we have IQR scores below code will give an output with some true and false values (with preset threshold).
- The data point where we have False means values are valid and True indicates the presence of an outlier.

```
print(boston_df_o1 < (Q1-1.5 * IQR)) |(boston_df_o1 > (Q3 + 1.5 *
IQR))
```

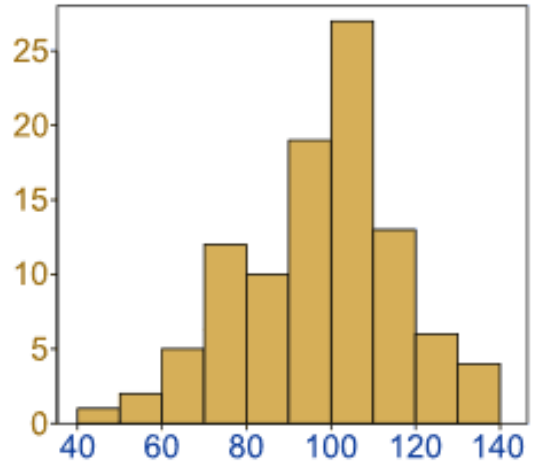
EDA (c): Understanding relationships and new insights through plots

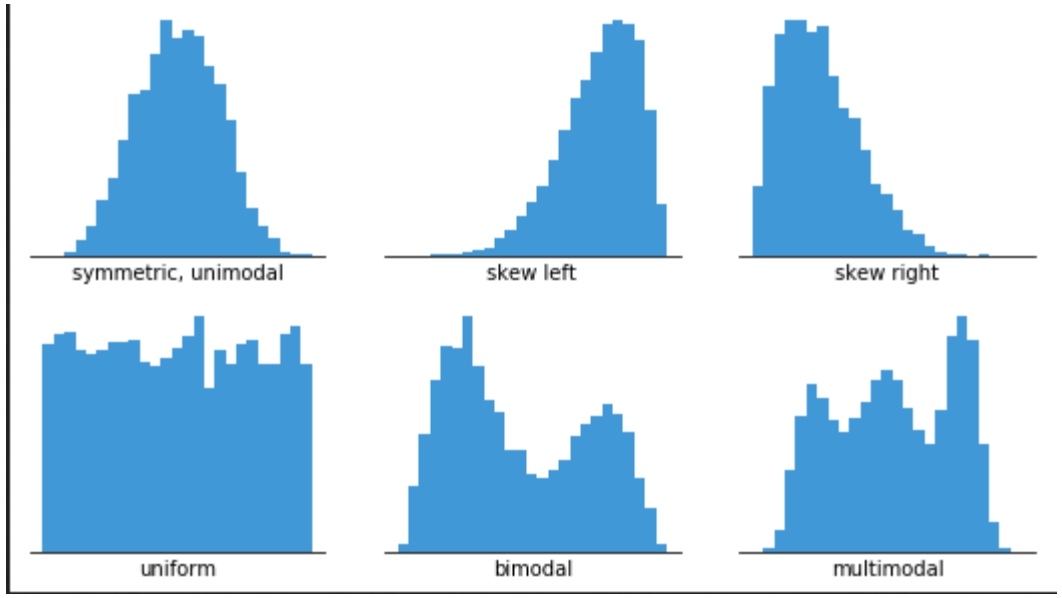
We can get many relations in our data by visualizing our data set. Let’s go through some techniques in order to see the insights.

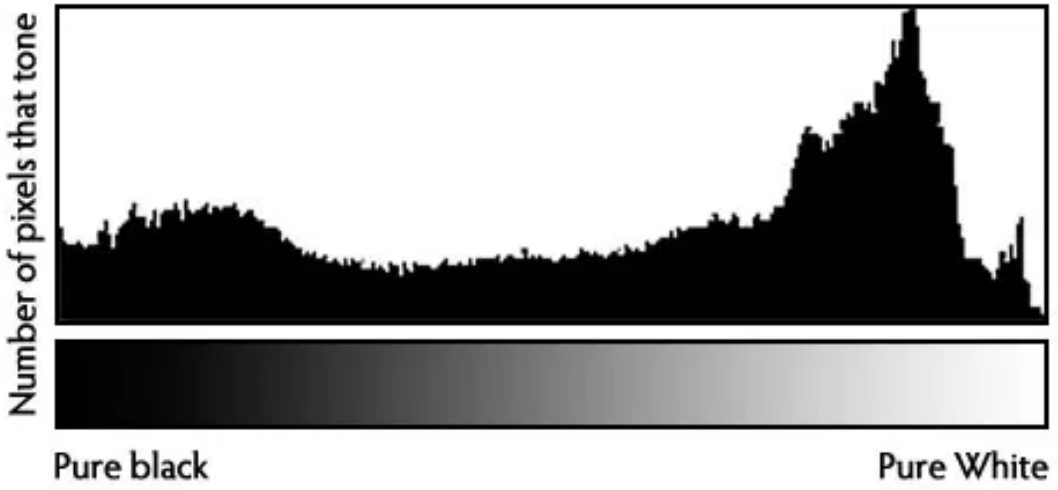
- Histogram
- Heat Maps

Histograms

A histogram is a great tool for quickly assessing a probability distribution that is easily understood by almost any audience.







Heatmaps

- The Heat Map procedure the distribution of a quantitative variable over all combinations of 2 categorical factors.
- If one of the 2 factors represents time, then the evolution of the variable can be easily viewed using the map.
- A gradient color scale is used to represent the values of the quantitative variable.
- The correlation between two random variables is a number that runs from -1 through 0 to +1 and indicates a strong inverse relationship, no relationship, and a strong direct relationship, respectively.

Heatmaps: finding co-related variables