The data are from the paper:

Exploring Relationships in Body Dimensions

Grete Heinz and Louis J. Peterson
San José State University

Roger W. Johnson and Carter J. Kerk
South Dakota School of Mines and Technology

Journal of Statistics Education Volume 11, Number 2 (2003), https://goo.gl/WMXNbj

Topics from GCE AS and A Level Mathematics covered in Sections 1:

- Interpret diagrams for single-variable data.
- Interpret measures of central tendency and variation, extending to standard deviation.
- Recognise and interpret possible outliers in data sets and statistical diagrams.
- Select or critique data presentation techniques in the context of a statistical problem.
- Be able to clean data (including dealing with missing data, errors and outliers).

Before carrying out an investigation it is important to become familiar with the data and how they have been collected.

Data collection

Below is an extract from the paper:

> 'Body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender, are given for 507 physically active individuals, 247 men and 260 women.
>
> The first two authors and technicians trained by the authors took the measurements of the 247 men and 260 women in the dataset associated with this article. These were primarily individuals in their twenties and early thirties, with a scattering of older men and women, all physically active (several hours of exercise a week). The initial measurements were taken at San José State University and at the U.S. Naval Postgraduate School in Monterey, California, primarily by the first two authors. Additional measurements were performed by technicians in dozens of California health and fitness clubs. Every effort was made to assure consistency in these different settings by having one of the authors (G.H.) monitor the technicians' measurement techniques.
>
> We caution the reader that the dataset does not constitute a random sample from a well-defined population. While we feel that the dataset can be used to illustrate a number of inferential techniques in statistics, such inference is technically invalid here as the sample is not a representative sample from a well-defined population.'

Refer to the extract from the paper to discuss the following questions.

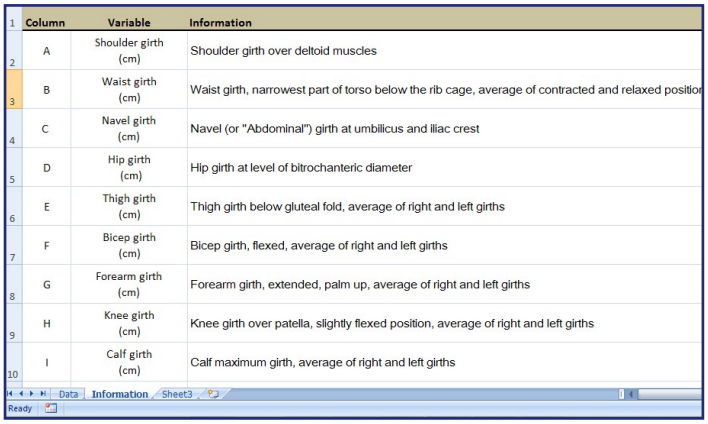
a). **Why is it important to train the researchers to take the measurements?**
To be sure of consistency in all measurements.

b). **Why is the sample of adults in this data set not random?**
It seems that the universities and health clubs from where the samples were selected were not chosen randomly and no information is given about how the members were selected.

c). **What type of sampling method has been used?**
Probably opportunity sampling, i.e. the members of the sample were chosen as they were available (or volunteered). They were not chosen randomly.

d). **What does the author mean by not 'a well-defined' population?**
The target population has not been stated. Is the target population all adults, active adults, American adults or those who attend or work at an American university or attend health and fitness clubs? The paper does not make this clear.

e). **What are inferential techniques?**
Inferential statistics are used to try to infer (predict) from the sample data to make judgements about the population.



Column	Variable	Information
A	Shoulder girth (cm)	Shoulder girth over deltoid muscles
B	Waist girth (cm)	Waist girth, narrowest part of torso below the rib cage, average of contracted and relaxed positio
C	Navel girth (cm)	Navel (or "Abdominal") girth at umbilicus and iliac crest
D	Hip girth (cm)	Hip girth at level of bitrochanteric diameter
E	Thigh girth (cm)	Thigh girth below gluteal fold, average of right and left girths
F	Bicep girth (cm)	Bicep girth, flexed, average of right and left girths
G	Forearm girth (cm)	Forearm girth, extended, palm up, average of right and left girths
H	Knee girth (cm)	Knee girth over patella, slightly flexed position, average of right and left girths
I	Calf girth (cm)	Calf maximum girth, average of right and left girths

Data | Information | Sheet3
Ready

The data are in the Excel spreadsheet '**Adult Measurements.xlsx**'.

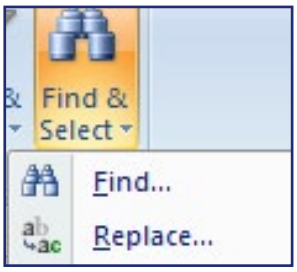
The worksheet named **Data** has15 variables in the columns with column headers and data for 507 respondents in the rows.

The worksheet named **Information** has the definitions for the variables and coding for gender.

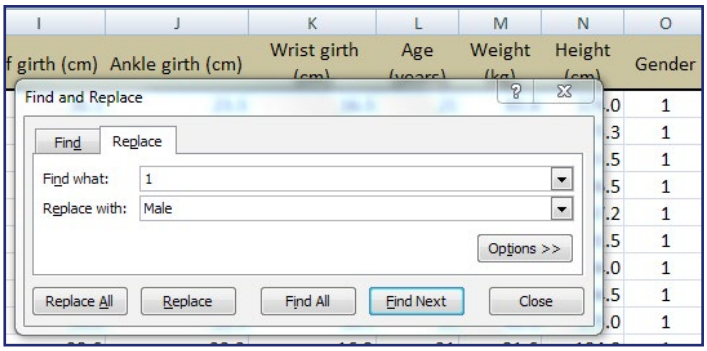
At this stage it is a good idea to replace the coding for gender with text.
For example replace **1** with **Male** and **0** with **Female**.

To use the Replace function in Excel

- Select the **Home** tab
- Select the column for '**Gender**' (column **O**)



- Select **Find & Select**
- Then **Replace**



- Enter **1** in **Find what**
- Enter **Male** in **Replace with**
- **Replace All**
- **Close**
- Repeat for the Female code.
- **Save your work**

Note: You must **select column O before Replace all** otherwise every 0 will be replaced by Male in the whole spreadsheet.

Process

Below is an extract from the worksheet **Data** in the workbook '**Adult Measurements.xls**'.

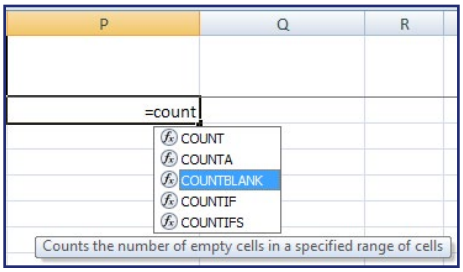
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Shoulder girth (cm)	Waist girth (cm)	Navel girth (cm)	Hip girth (cm)	Thigh girth (cm)	Bicep girth (cm)	Forearm girth (cm)	Knee girth (cm)	Calf girth (cm)	Ankle girth (cm)	Wrist girth (cm)	Age (years)	Weight (kg)	Height (cm)	Gender
2	106.2	71.5	74.5	93.5	51.5	32.5	26.0	34.5	36.5	23.5	16.5	21	65.6	174.0	Male
3	110.5	79.0	86.5	94.8	51.5	34.4	28.0	36.5	37.5	24.5	17.0	23	71.8	175.3	Male
4	115.1	83.2	82.9	95.0	57.3	33.4	28.8	37.0	37.3	21.9	16.9	28	80.7	193.5	Male
5	104.5	77.8	78.8	94.0	53.0	31.0	26.2	37.0	34.8	23.0	16.6	23	72.6	186.5	Male
6	107.5	80.0	82.5	98.5	55.4	32.0	28.4	37.7	38.6	24.4	18.0	22	78.8	187.2	Male
7	119.8	82.5	80.1	95.3	57.5	33.0	28.0	36.6	36.1	23.5	16.9	21	74.8	181.5	Male
8	123.5	82.0	84.0	101.0	60.9	42.4	32.3	40.1	40.3	23.6	18.8	26	86.4	184.0	Male
9	120.4	76.8	80.5	98.0	56.0	34.1	28.0	39.2	36.7	22.5	18.0	27	78.4	184.5	Male
10	111.0	68.5	69.0	89.5	50.0	33.0	26.0	35.5	35.0	22.0	16.5	23	62.0	175.0	Male
11	119.5	77.5	81.5	99.8	59.8	36.5	29.2	38.3	38.6	22.2	16.9	21	81.6	184.0	Male
12	117.1	81.9	81.0	98.4	60.5	34.6	27.9	38.9	40.1	23.2	16.2	23	76.6	180.0	Male

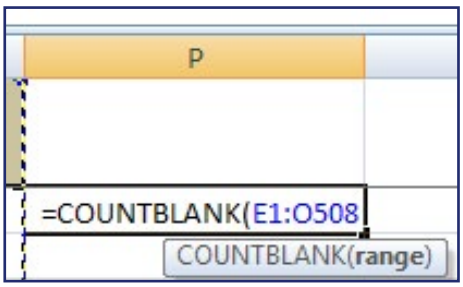
1. Are there any missing values?

There are no missing values.

There are several ways to check for missing values.
One way is to use the **COUNTBLANK** function in Excel.

To use the COUNTBLANK function in Excel





- Select **P1**, type **=count.**

A list of COUNT functions will appear.

- Double click on **COUNTBLANK**.

This function counts the number of empty cells in a specified range.

- Select **E1** to **O508** and **Enter**

Do not select the whole columns A to O as there are thousands of empty cells in these columns.

The result is 0.

- **Save your work**

A problem with this COUNTBLANK method is missing values could be labelled N/A, * or have some other coding.

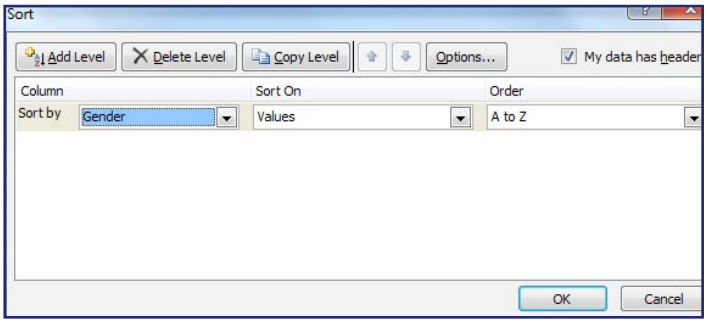
Another way is to sort the data according to the variable you are interested in, either alphabetically or numerically depending on the type of variable, and check the column for coded missing values.

If data are missing, do not delete the whole row of data. Either accept the value is missing or, if possible, try to find the missing value.

To sort in Excel

- Select the **Data** tab
- Select columns **A** to **O**

- Select **Sort**

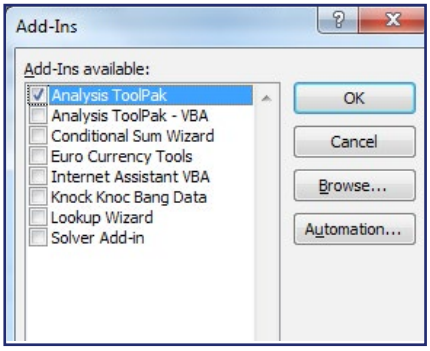


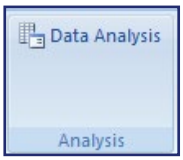
- Tick **My data has headers**
- Select **Gender** in the **Sort by** box
- **OK.**
- Inspect the data for coded missing values.

2. Comment on the shape of the distribution for the variable 'Age'.

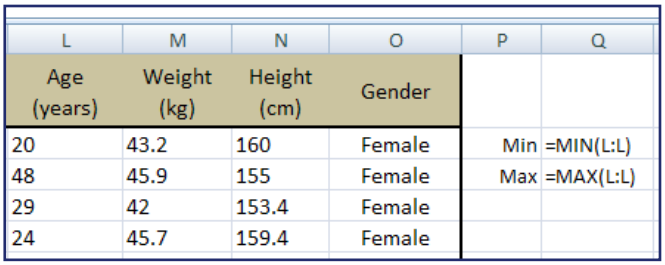
One way to investigate the shape of a distribution is to plot a histogram.

To plot a histogram in Excel



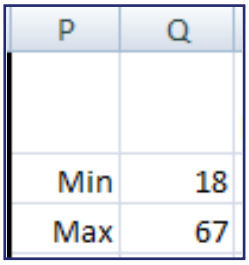


- First make sure the **Data Analysis Toolpak** is in Excel
- Press at the same time **Alt T I** (i.e. Alt and capital T and capital I)
- The **Add-Ins** menu will appear.
- Tick **Analysis ToolPak** then **OK**.
- **Data Analysis** will appear in the **Data** tab.
- In Excel, bins are defined as the upper boundaries and lower limits of each class.

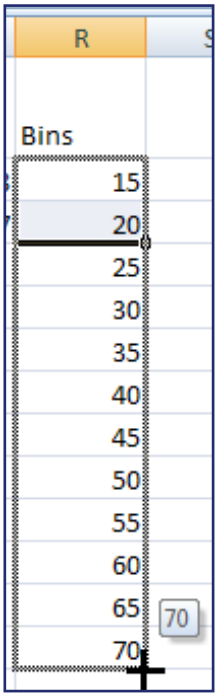


L	M	N	O	P	Q
Age (years)	Weight (kg)	Height (cm)	Gender		
20	43.2	160	Female	Min	=MIN(L:L)
48	45.9	155	Female	Max	=MAX(L:L)
29	42	153.4	Female		
24	45.7	159.4	Female		

- Find the minimum and maximum values of the data to decide on the values of the bins.
- Select **P2** type **Min**.
- Select **Q2** type **=MIN** (and either select column L or type **L:L**) then **Enter**.
- Select **P3** type **Max**.
- Select **Q3** type **=MAX** (and either select column L or type **L:L**) then **Enter**.

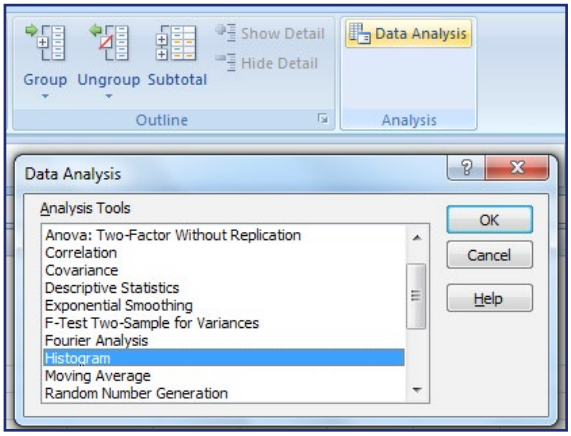


P	Q
Min	18
Max	67

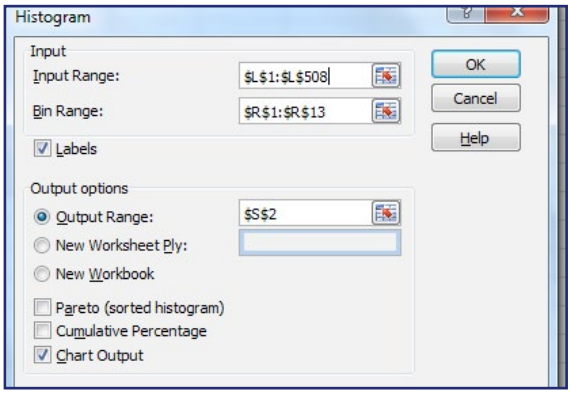


- Since the minimum value is 18 and the maximum is 67 use bins 15, 20, 25, ..., 70.
- Select **R1** and type **Bins**.
- Select **R2** and enter **15**. Select **R3** and enter **20**.
- Copy the pattern 15, 20, 25, ..., 70.
- Select **R2:R3,** take the cursor to the bottom right hand corner of **R3** until you see a small cross then click and drag. Stop at 70 (**R13**).

Plot the histogram

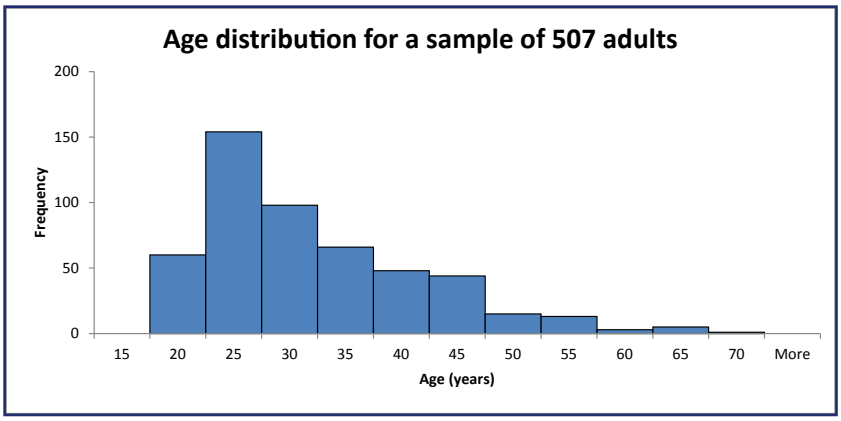


- Select the **Data** tab then **Data Analysis** then **Histogram** and **OK**
- Select **Input Range** (by clicking into the box) then select the data – column **L** or **L1:L508**
- Select **Bin Range** then **R1:R13**
- Tick **Labels** so **L1** and **R1** will be recognised as headers



- Select **Output Range** then select **S2**
- Tick **Chart Output** then **OK**
- The output graph is not a histogram as it has gaps and it is not labelled properly.
- Remove the legend.
- Click on **Frequency** then **Delete**
- Close the gaps between the columns.
- Right click on a bar.
- Select **Format Data Series** then **Series Options.**
- Close **Gap width** to **0%**.
- Add a border to the columns.

- Select **Border Color** then **Solid line**, open **Color** and choose **black** and **Close**.
- Add a title and label the axes.
- Click on **horizontal axis** then type in **Age (years)** then **Enter**.
- Vertical axis is fine.
- Click on **title** then type in '**Age distribution for a sample of 507 adults**', then **Enter.**



When Excel processes the histogram function it will count how many adults are greater than 0 and less than or equal to 15 and enters the count into **S2**. This is 0.

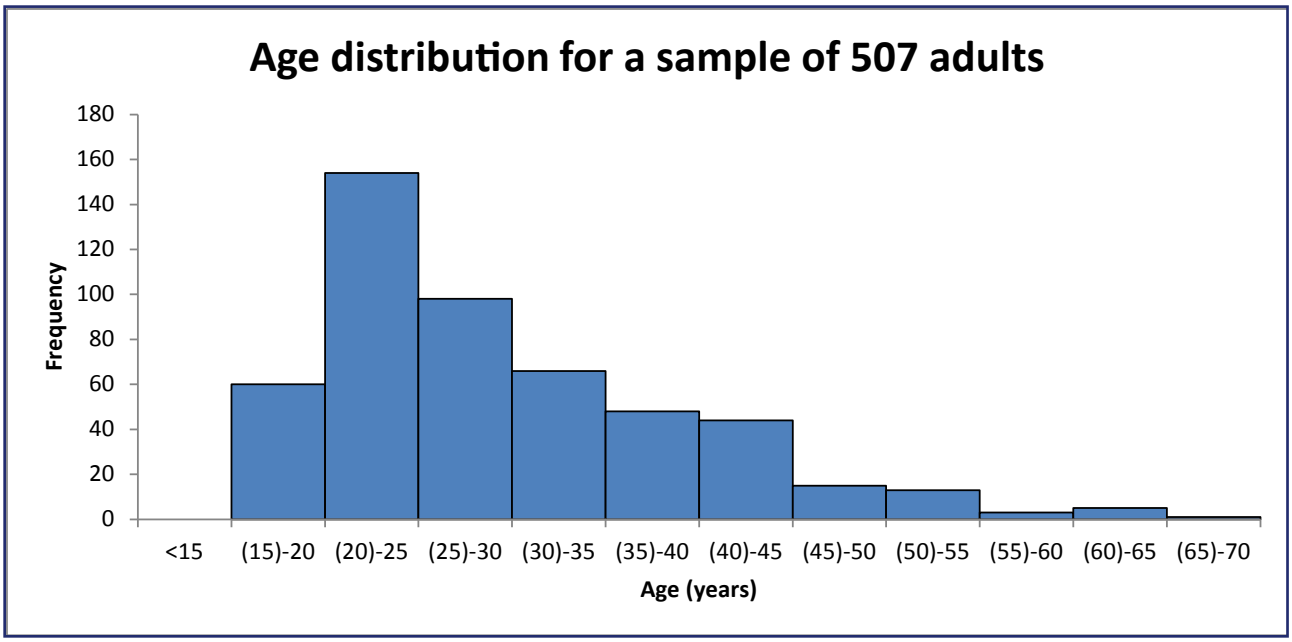
The frequency in **T3** is the number of adults who are greater than 15 and less than or equal to 20. The frequency in **T4** is the number of adults who are greater than 20 and less than or equal to 25, and so on.

This needs to be shown on the x-axis scale. A number in brackets shows this is a limit and that the exact value is not included in the column.

Correct the scale on the horizontal axis:

Bins	*Frequency*
<15	0
(15)-20	60
(20)-25	154
(25)-30	98
(30)-35	66
(35)-40	48
(40)-45	44
(45)-50	15
(50)-55	13
(55)-60	3
(60)-65	5
(65)-70	1

- Select **S2** and type **< 15** then **Enter**
- Select **S3** and type **(15)–20** then **Enter** and so on
- Select **S15:T15** and **Delete**.



Comment on the shape of the distribution for the variable 'Age'.

The distribution of the variable 'Age' is positively skewed i.e. there are a few very large values.

3. What is the average age of the sample?

Based on the shape of the distribution of the variable 'Age', what measure of location would be the most appropriate to use?

As the distribution of the variable 'Age' is positively skewed the mean will be greater than the median. This is because when calculating a mean, all the values are included. Therefore the better measure of location to use is the median as this is the middle value and not affected by very large or very small values.

To investigate this, use the functions in Excel to complete the table below::

	Age (years)
Mean	
Standard deviation	
Minimum	
Lower quartile, Q1	
Median, Q2	
Upper quartile, Q3	
Maximum	
Interquartile range, IQR	

Copy the table above into **P5** to **Q13**.

To calculate summary statistics in Excel

To enter a function in Excel you must start with **=.**

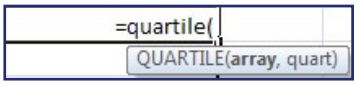
As you enter the function a message appears telling you what the function will do. As you enter an open bracket a message will appear to let you know in what format to enter the information.

Calculate a mean:

- Select **Q6**.
- Type **=AVERAGE(L:L)** then **Enter**.

The function AVERAGE calculates the mean for an array of data.

Calculate quartiles



The function **QUARTILE** calculates the three quartiles for an array of data.
The function asks for an 'array' and a 'quart'.

The 'array' is the column with the data.
The 'quart' is to let Excel know which quartile to calculate.

- Enter 1 for the lower quartile (Q1)
  2 for the median (Q2)
  3 for the upper quartile (Q3).

The interquartile range (IQR) is given by: $Q_3 - Q_1$

Continue by entering the functions as shown in the table below.

Statistic	Excel function
Mean	=AVERAGE(L:L)
Population standard deviation	=STDEVP(L:L)
Minimum	=MIN(L:L)
Lower quartile, Q_1	=QUARTILE(L:L,1)
Median, Q_2	=QUARTILE(L:L,2)
Upper quartile, Q_3	=QUARTILE(L:L,3)
Maximum	=MAX(L:L)
Interquartile range, IQR	=Q11-Q9

- Reduce the number of decimal place values for all the statistics in the table to 2.
- Select the statistics in the table and click on the  icon the appropriate number of times.
- **Save your work**

	Age (years)
Mean	30.18
Standard deviation	9.60
Minimum	18.00
Lower quartile, Q1	23.00
Median, Q2	27.00
Upper quartile, Q3	36.00
Maximum	67.00
Interquartile range, IQR	13.00

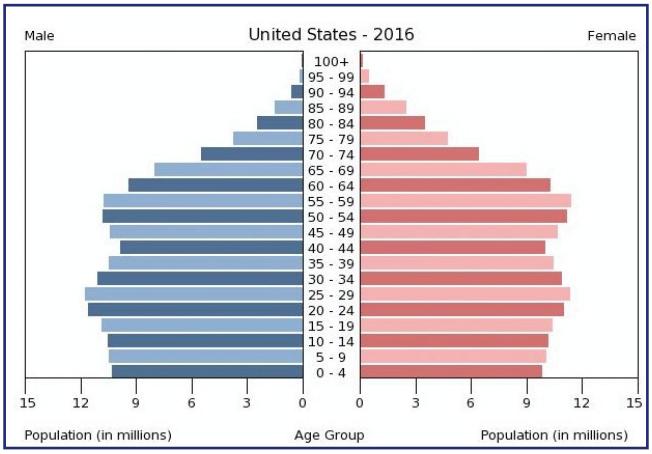
Report
The median age of the adults in this sample is 27 years.

4. Explain why it better to use the IQR than the standard deviation as a measure of spread.
Since the calculation of the standard deviation includes all the values this is distorted by the few high values. The IQR is the middle 50% of the data and is not affected by a few very high or very low values.

5. Are all age groups of the United States well represented in this sample?

A population pyramid illustrates the age and gender structure of a country's population. The population is distributed along the horizontal axis, with males shown on the left and females on the right. The male and female populations are broken down into age groups represented as horizontal bars along the vertical axis, with the youngest age groups at the bottom and the oldest at the top.

The population pyramid for the United States in 2016 shows that the population does start to fall away slightly between 60 and 70 years but not to the extent of this sample of adults. Therefore not all age groups are well represented in this sample



Source: *index mundi* https://goo.gl/1L2tqG

6. Are there any errors or outliers?

Plotting a box plot in Excel is tricky and outliers are not shown.
Use the formula to calculate the limits for outliers.

Outlier limit	Age (years)
Lower limit = Q1 – 1.5 × IQR	*23 – 1.5 × 13 = 3.5*
Upper limit = Q3 + 1.5 × IQR	*36 + 1.5 × 13 = 55.5*

Since the minimum value is 18 there are no lower outliers. However, the maximum value is over 55.5 so there is at least one upper outlier.

To inspect the adults who have ages over the upper limit for outliers:
Sort the data according to 'Age' and investigate the data for adults over the upper outlier limit. *(55.5)*

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Shoulder girth (cm)	Waist girth (cm)	Navel girth (cm)	Hip girth (cm)	Thigh girth (cm)	Bicep girth (cm)	Forearm girth (cm)	Knee girth (cm)	Calf girth (cm)	Ankle girth (cm)	Wrist girth (cm)	Age (year	Weight (kg)	Height (cm)	Gender
495	107.6	84.9	83.7	97.9	51.8	28.0	25.2	37.5	36.0	21.3	15.1	53	64.5	167.6	Male
496	117.0	98.5	99.9	103.6	57.5	33.7	29.0	38.7	36.7	24.3	17.7	54	85.0	176.5	Male
497	113.2	90.1	90.8	96.6	55.0	32.9	26.5	35.6	37.9	22.5	17.8	55	74.1	167.6	Male
498	110.6	109.2	104.4	101.7	56.4	34.0	29.2	39.3	38.5	24.3	17.4	55	94.1	185.4	Male
499	122.4	98.0	98.0	99.6	56.7	36.4	29.2	40.9	42.1	26.1	19.5	55	87.3	179.1	Male
500	93.0	72.3	91.7	97.8	59.1	31.0	26.3	39.9	39.7	21.6	15.4	56.0	63.6	162.6	Female
501	107.8	84.7	92.9	96.7	54.6	32.8	25.3	36.6	35.0	23.5	17.1	56.0	70.9	175.3	Male
502	120.2	90.5	92.3	95.2	52.4	35.8	28.7	32.5	36.5	23.7	17.5	60.0	73.6	175.3	Male
503	112.8	94.8	98.2	98.6	48.3	31.1	27.0	37.7	36.8	24.6	18.4	62	76.4	185.4	Male
504	112.0	81.5	80.5	89.5	52.0	30.3	25.0	36.0	35.0	22.4	16.7	62	66.8	167.6	Male
505	104.0	76.0	83.0	93.0	54.5	29.5	26.0	37.0	34.5	22.8	17.4	62	64.6	167.0	Male
506	99.7	73.7	92.7	101.2	57.4	28.1	23.3	33.9	33.3	21.2	14.9	64	58.6	156.2	Female
507	123.5	98.6	99.2	103.3	55.3	35.0	29.3	35.9	36.0	23.5	18.1	65	80.0	174.0	Male
508	108.2	75.3	96.3	103.6	59.9	31.5	25.4	35.7	35.8	21.1	16.3	67	67.7	170.2	Female

Are there any outliers?
Looking at the measurements for the other variables for the over 55s, these seem to be similar to other adults. Also, there is a mix of males and females. Therefore these are not errors and will be left in the dataset.

(Note: there are outliers for most of the continuous variables in these data. However, none of these outliers are over 55.)

7. What is considered to be the minimum age for an adult in this research?
18 years old.

8. What is the target population?
The target population has not been well defined in the paper. However, the sample has been selected from a population of American adults who are physically active. This will be considered the target population.