DATA EXCERCISES

ACTIVITY 1

- The activity list of IQ score are : 118, 123, 124, 125, 127, 128, 129, 130, 133, 136, 138, 141, 142, 149, 150, 154. Do Frequency distribution table with classes.
- Answer

Interval	Frequency
118-125	4
126-133	6
134-141	3
142-149	2
150-157	2

In the above table we can see the cluster between 134-157

ACTIVITY 4

10-15	4
16-20	1
21-25	3
26-30	7
31-35	8
36-40	9
41-45	5

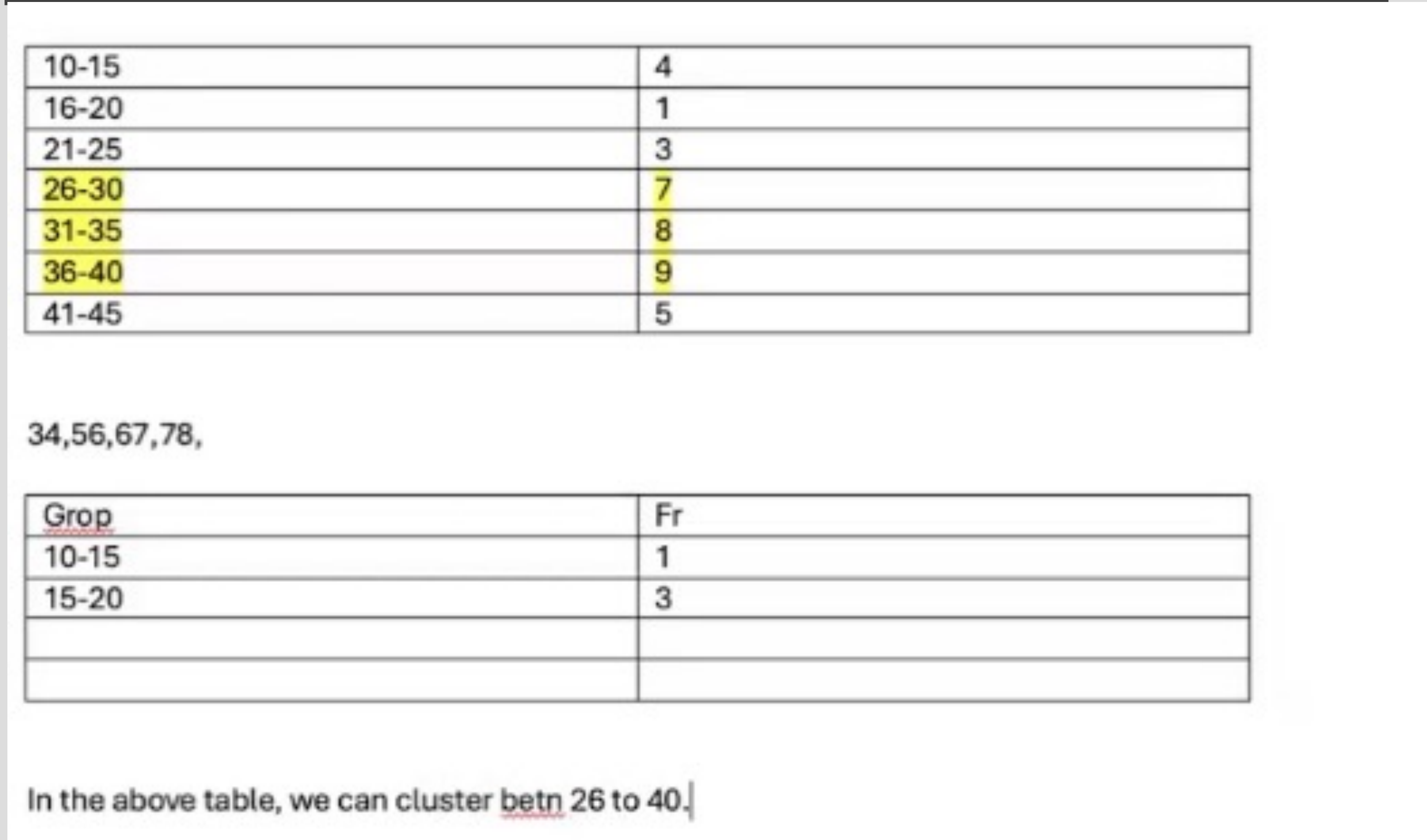
In the above table we can see the cluster between 26-40

FREQUENCY DISTRIBUTION

Holidays	Frequency (f)	Percentage
0	7	7/36
1	18	18/36 = 50%
2	6	6/36
3	4	4/36
4	1	¼=25%
Total	36	

From the table we understand 50% have take 18 (1 day) holidays

GROUP FREQUENCY DISTRIBUTION



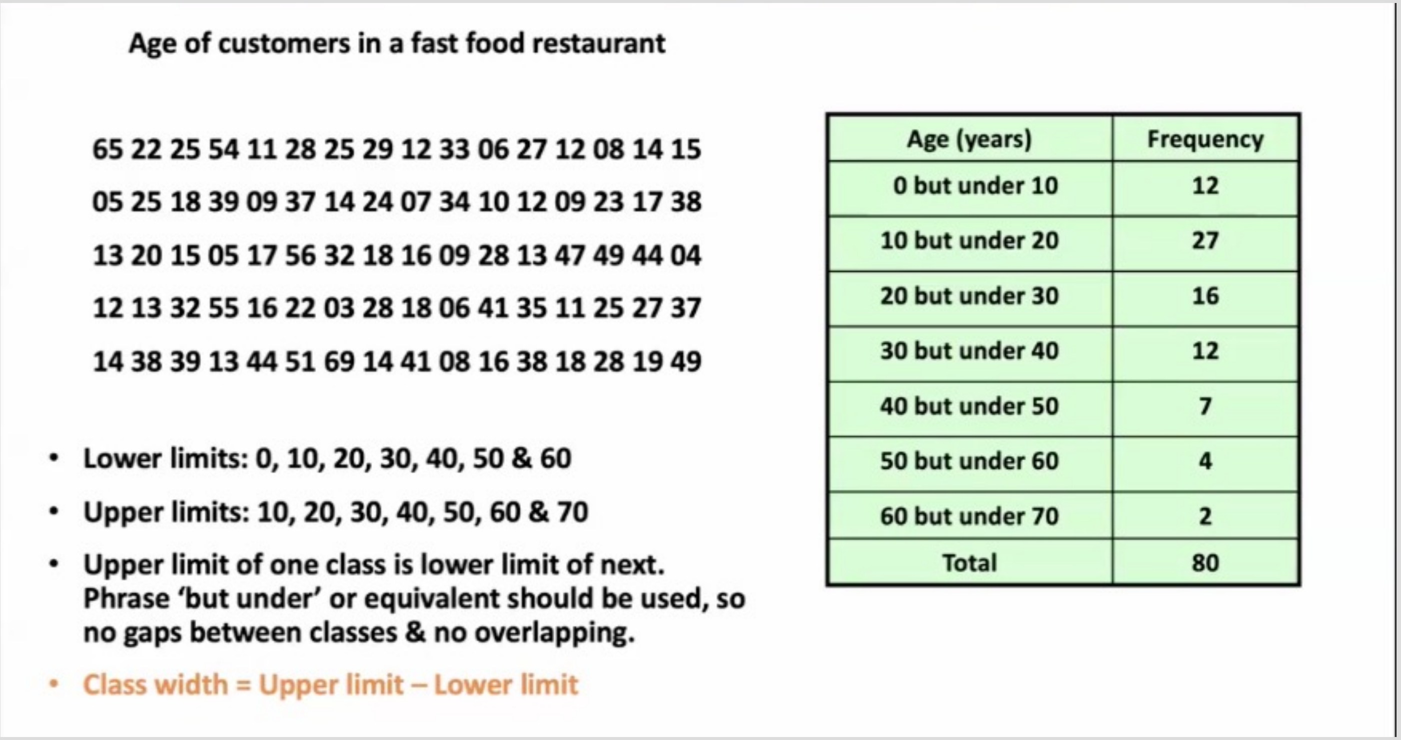
10-15	4
16-20	1
21-25	3
26-30	7
31-35	8
36-40	9
41-45	5

34,56,67,78,

Grop	Fr
10-15	1
15-20	3

In the above table, we can cluster betn 26 to 40.

FREQUENCY DISTRIBUTION



Age of customers in a fast food restaurant

65 22 25 54 11 28 25 29 12 33 06 27 12 08 14 15

05 25 18 39 09 37 14 24 07 34 10 12 09 23 17 38

13 20 15 05 17 56 32 18 16 09 28 13 47 49 44 04

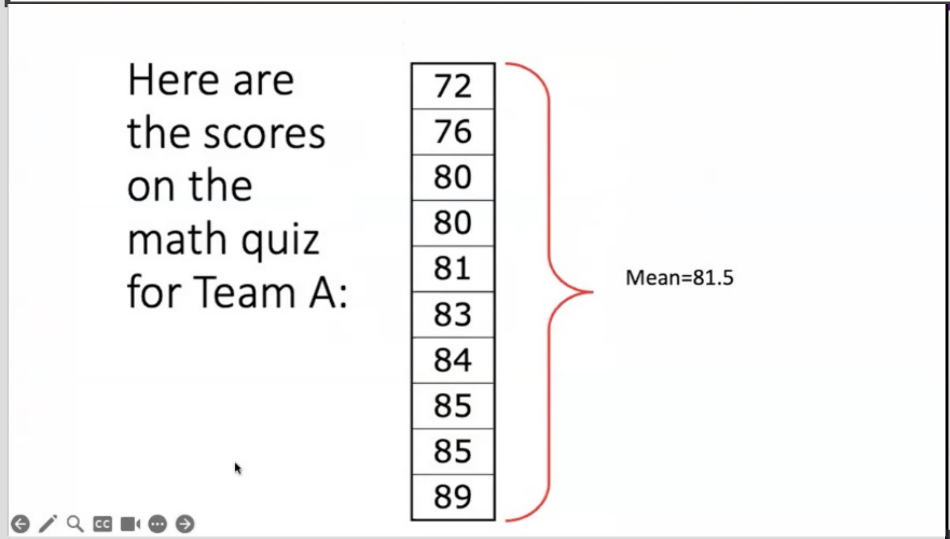
12 13 32 55 16 22 03 28 18 06 41 35 11 25 27 37

14 38 39 13 44 51 69 14 41 08 16 38 18 28 19 49

- **Lower limits: 0, 10, 20, 30, 40, 50 & 60**
- **Upper limits: 10, 20, 30, 40, 50, 60 & 70**
- **Upper limit of one class is lower limit of next. Phrase 'but under' or equivalent should be used, so no gaps between classes & no overlapping.**
- **Class width = Upper limit – Lower limit**

Age (years)	Frequency
0 but under 10	12
10 but under 20	27
20 but under 30	16
30 but under 40	12
40 but under 50	7
50 but under 60	4
60 but under 70	2
Total	80

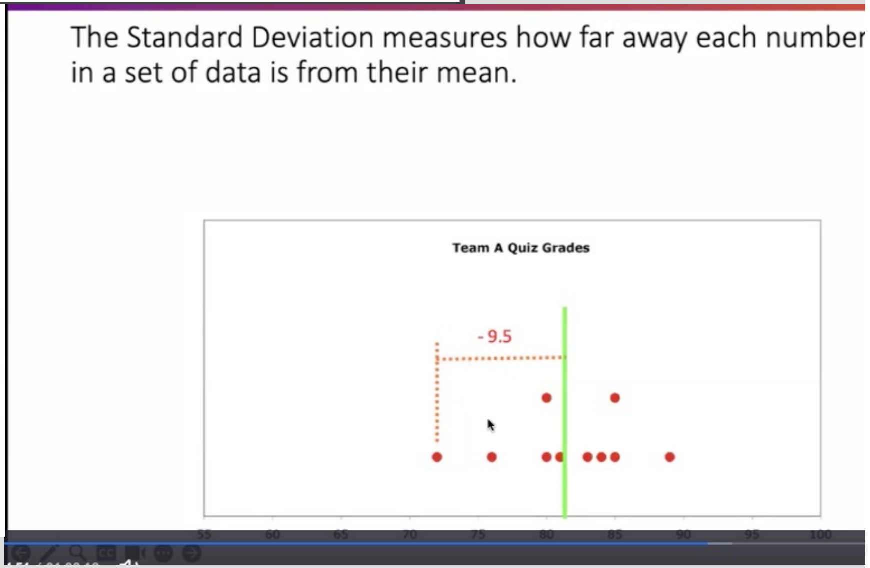
STANDARD DEVIATION



Here are the scores on the math quiz for Team A:

72
76
80
80
81
83
84
85
85
89

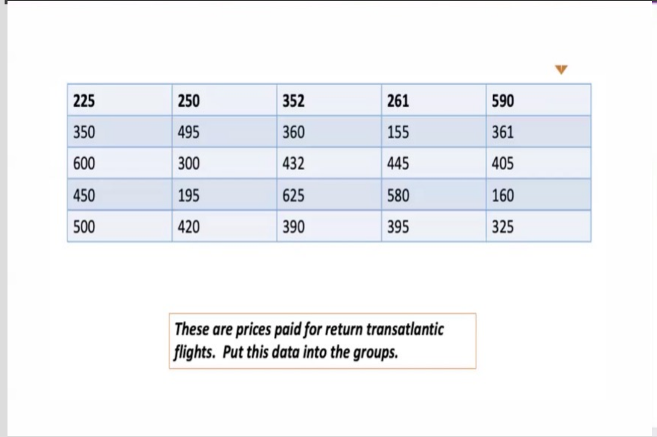
Mean=81.5



The Standard Deviation measures how far away each number in a set of data is from their mean.

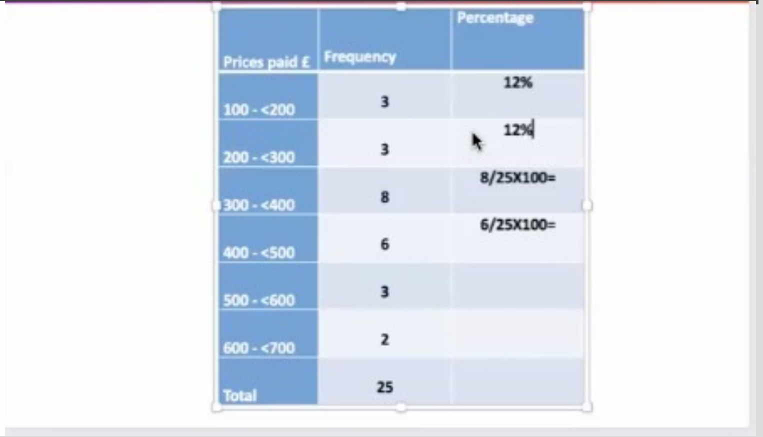
The difference between Mean and lowest value is -9.5
The difference between Mean and highest value is 7

ACTIVITY 1



225	250	352	261	590
350	495	360	155	361
600	300	432	445	405
450	195	625	580	160
500	420	390	395	325

These are prices paid for return transatlantic flights. Put this data into the groups.



Prices paid £	Frequency	Percentage
100 - <200	3	12%
200 - <300	3	12%
300 - <400	8	8/25X100=
400 - <500	6	6/25X100=
500 - <600	3	
600 - <700	2	
Total	25	

ACTIVITY 2

- Listed below are maximum daily temperatures (in degrees Celsius) in Iqaluit from June 2 to June 16: 2.8, 7.3, 9.6, 8.9, 11.4, 6.7, 5.8, 5.5, 6.7, 6.2, 9.0, 8.2, 7.6, 8.5, 6.7
- Find the range, Interquartile range, Median

Answer

- Ordered the data - 2.8, 5.5, 5.8, 6.2, 6.7, 6.7, 6.7, 7.3, 7.6, 8.2, 8.5, 8.9, 9.0, 9.6, 11.4
- Range : Maximum Value – Minimum Value = 11.4-2.8= 8.6
- Median : 7.3
- Interquartile Range (IQR) = Q3- Q1 = 8.9 – 6.2 = 2.7
- Q1 = {2.8, 5.5, 5.8, 6.2, 6.7, 6.7, 6.7} = 6.2
- Q3 = {7.6, 8.2, 8.5, 8.9, 9.0, 9.6, 11.4} = 8.9

MODE

2,3,3,4,5 - Mode = 3

2,3,3,4,5,2 – Mode – 2,3 [data is bi modal]

2,3,3,4,5,2,4 – Mode – 2,3,4 [Tri modal]

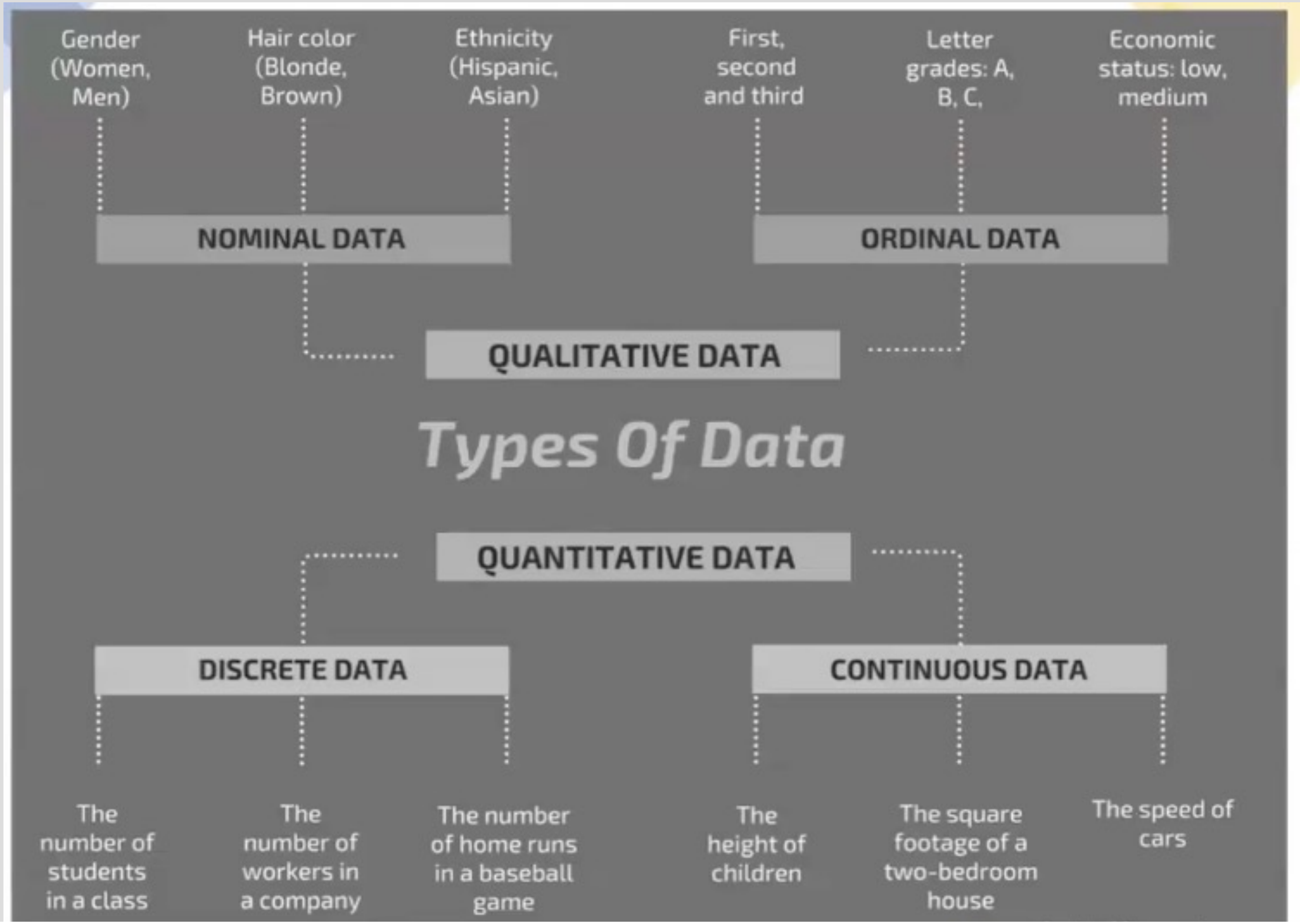
ACTIVITY 3

- Are the following ratio, interval, ordinal or normal data ? And why do they meet each classification ?

Number of males and females in a primary school	Nominal Data
A depression rating scale	Interval Data
A pain scale	Interval Data
Number of people from each region of the UK who voted for a labour government	Nominal Data
Money in Pence	Ratio Data
Intelligent Rating scale	Interval Data
Number of Children in swimming pool who received gold, silver and bronze	Ordinal data
Weight measurement of cohort of ladies in swimming club	Ratio Data
Patient Satisfaction Survey	Interval Data

TYPES OF DATA

Eye color	Nominal Data
Weight of a person	Continuous data
Flavor of ice-cream	Nominal Data
Educational level	Ordinal Data
Market share price	Continuous data
Total number of students present in class	Discrete Data
Wifi frequency	Continuous Data
Cost of a cell phone	Continuous Data
Gender	Nominal Data
Ranking in army	Ordinal Data



Types Of Data

QUALITATIVE DATA

NOMINAL DATA

Gender (Women, Men)

Hair color (Blonde, Brown)

Ethnicity (Hispanic, Asian)

ORDINAL DATA

First, second and third

Letter grades: A, B, C,

Economic status: low, medium

QUANTITATIVE DATA

DISCRETE DATA

The number of students in a class

The number of workers in a company

The number of home runs in a baseball game

CONTINUOUS DATA

The height of children

The square footage of a two-bedroom house

The speed of cars

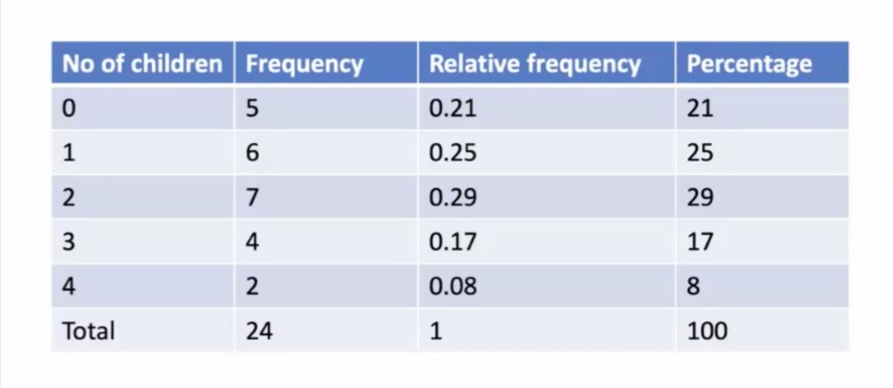
HYPOTHESIS

- There is an effect of weight on the body's physical movement
  - Null Hypothesis :There is no effect of weight on the body's physical movement.
  - Alternative hypothesis :There is an effect of weight on the body's physical movement.
- Girls are performing better than boys in Maths test
  - Null Hypothesis : Girls are not performing netter than boys in the Maths test or Girls and boys performance are same in maths tests.
  - Alternative Hypothesis : Girls are performing better than boys in Maths tests.
- A and B are highly related
  - Null Hypothesis :A and B are Not related
  - Alternative Hypothesis :A and B are related.

CONFIDENCE INTERVAL

- 95% CI (lower linmit and upper limit)
- 95% CI is presented Odd Risk, Relative Risk, Hazard Rist, 1 or not
- 95% CI is also presented mean differences, 0 or not
- If a 95% CI contains 0, it is non sig (0.23 to 0.45) does ot contain 0? Yes, Non
- If a 95% CI contains 1, it is no sig (0.78 to 1.34) does it contain 1 ? Yes, Non

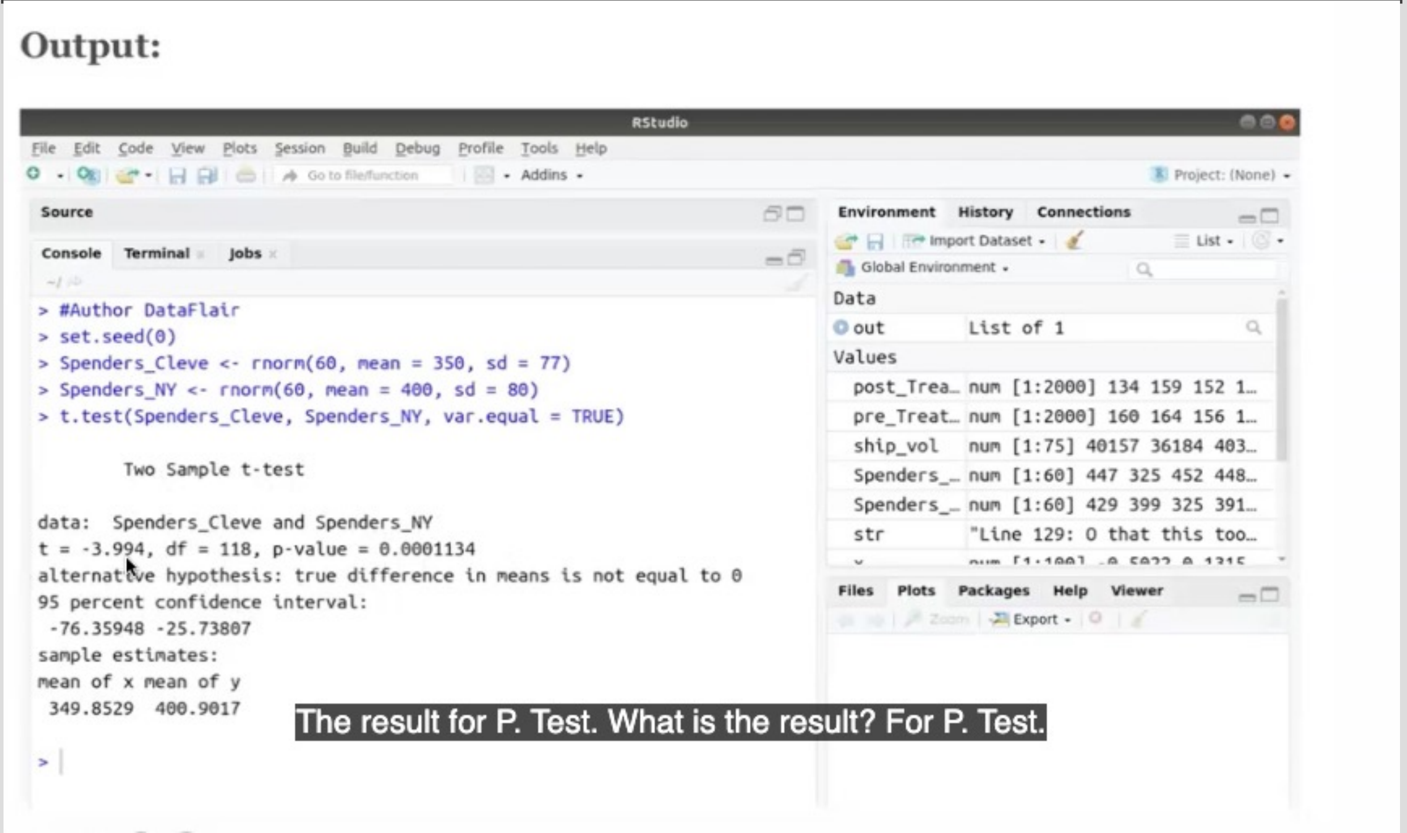
EXERCISE



No of children	Frequency	Relative frequency	Percentage
0	5	0.21	21
1	6	0.25	25
2	7	0.29	29
3	4	0.17	17
4	2	0.08	8
Total	24	1	100

You need to mention in in 29% households have got 2 children, not this sentence nicely. Don't just write 29%, right? 29% households have got 2 children. You will get 20 out of 20.

TEST QUESTION



Output:

```
> #Author DataFlair
> set.seed(0)
> Spenders_Cleve <- rnorm(60, mean = 350, sd = 77)
> Spenders_NY <- rnorm(60, mean = 400, sd = 80)
> t.test(Spenders_Cleve, Spenders_NY, var.equal = TRUE)

        Two Sample t-test

data:  Spenders_Cleve and Spenders_NY
t = -3.994, df = 118, p-value = 0.0001134
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -76.35948 -25.73807
sample estimates:
mean of x mean of y
 349.8529  400.9017

> 
```

```
Data
out          List of 1
Values
post_Trea…   num [1:2000] 134 159 152 1…
pre_Treat…   num [1:2000] 160 164 156 1…
ship_vol     num [1:75] 40157 36184 403…
Spenders_…   num [1:60] 447 325 452 448…
Spenders_…   num [1:60] 429 399 325 391…
str          "Line 129: 0 that this too…
```

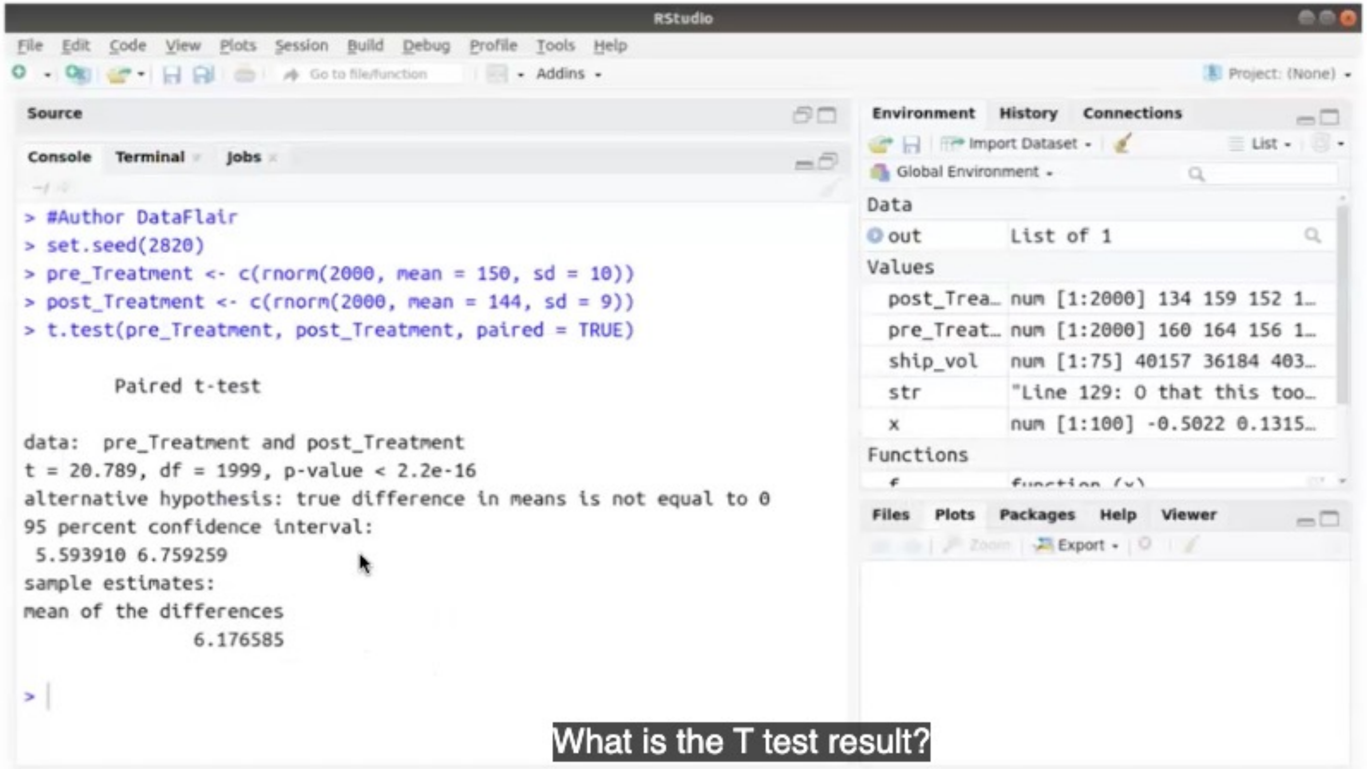
The result for P. Test. What is the result? For P. Test.

The sample t-test result is showing minus 3.994

The pvalue in this test is 0.00 which is lower than 0.05 so the test is statistically significant

The 95percent confidence interval value does not contain 0 and is significant

TEST QUESTION



```
> #Author DataFlair
> set.seed(2820)
> pre_Treatment <- c(rnorm(2000, mean = 150, sd = 10))
> post_Treatment <- c(rnorm(2000, mean = 144, sd = 9))
> t.test(pre_Treatment, post_Treatment, paired = TRUE)

        Paired t-test

data:  pre_Treatment and post_Treatment
t = 20.789, df = 1999, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 5.593910 6.759259
sample estimates:
mean of the differences
               6.176585

> |
```

```
Data
out          List of 1
Values
post_Trea…   num [1:2000] 134 159 152 1…
pre_Treat…   num [1:2000] 160 164 156 1…
ship_vol     num [1:75] 40157 36184 403…
str          "Line 129: 0 that this too…
x            num [1:100] -0.5022 0.1315…
Functions
f            function (x)
```

What is the T test result?

The t test result is 20.789

The p value is is less than 0.05 so this is significant

Formulate Null and alternative hypothesis

The 95 percent confidence interval does not include 0 and and is statistically significant

NOMINAL DATA (CATEGORICAL, NO ORDER)

- Colors of cars
- Types of pets
- Gender
- Nationality
- Eye Colour
- Car Brands
- Type of Fruit
- Marital status

ORDINAL DATA (CATEGORICAL, WITH ORDER)

- Customer satisfaction ratings: Poor, Fair, Good, Excellent
- Education levels: High school, Bachelor's, Master's, PhD
- Pain levels: Mild, Moderate, Severe
- Military ranks: Private, Corporal, Sergeant, Captain
- Movie Ratings
- Socio Economic Status

DICRETE DATA (QUANTITATIVE, COUNTABLE, WHOLE NUMBERS)

- Number of students in a class: 20, 25, 30
- Number of cars in a parking lot: 10, 15, 22
- Number of books on a shelf: 5, 7, 9
- Number of pets in a household: 2, 3, 4
- Number of Children in family
- Number of Goals in a scored match
- Number of Books in shelf

Typically Integers

CONTINUOUS DATA (QUANTITATIVE, MEASURABLE, ANY VALUE)

- Height of a person: 5.6 feet, 6.2 feet
- Weight of an object: 55.5 kg, 72.3 kg
- Temperature in Celsius: 22.5°C, 36.1°C
- Time to complete a race: 12.45 seconds, 15.67 seconds
- Distance Traveled : 5.2 miles, 10.6 miles

Includes Decimals and Fractions

INTERVAL DATA (NUMERIC DATA WITH EQUAL INTERVALS BETWEEN VALUES BUT NO TRUE ZERO POINT. DIFFERENCES ARE MEANINGFUL, BUT RATIOS ARE NOT.)

- Temperature in Celsius or Fahrenheit: 20°C, 30°C (no absolute zero)
- Years on a calendar: 1990, 2000, 2020 (zero is arbitrary)
- IQ scores: 85, 100, 115
- SAT scores: 400, 600, 800

- Example : Temperature, Calendar dates

RATIO DATA (NUMERIC DATA WITH EQUAL INTERVALS AND A TRUE ZERO POINT, ALLOWING FOR MEANINGFUL COMPARISONS OF BOTH DIFFERENCES AND RATIO)

- Weight in kilograms: 50 kg, 70 kg (0 kg represents no weight)
- Height in meters: 1.5 m, 1.8 m (0 m represents no height)
- Time in seconds: 0 sec, 10 sec, 20 sec
- Distance traveled in kilometers: 0 km, 5 km, 10 km

- Example : Height, Weight, income, distance