

Preprocessing

Why is it needed?

Know our data

Get the overall picture of data

Learn some statistical measures

Descriptive Data Summarization

Distributive Measure: sum, count

Algebraic Measure: mean, weighted mean

Holistic Measure (expensive): median

It is the middle value.

Mode – value occurring most frequently

Midrange – average of largest and smallest values

Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 36, 40, 45, 46, 52, 70

What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).

What is the midrange of the data?

Descriptive Data Summarization

The data characteristics we discussed are called Central Tendencies

Another important characteristic is called Dispersion or variance of the data

- Range
- Five-number summary (based on quartiles)
- Interquartile range

Descriptive Data Summarization

Percentile:

The k th **percentile** of a set of data in numerical order is the value $x(i)$ having the property that k percent of the data entries lie at or below $x(i)$. The median is the 50th percentile.

25 percentile == 20.25

That means values from 0 to 20.25 make for 25% of the data.

50 percentile == 25

That means values from 0 to 25 make for 50% of the data.

Values from 25 and above constitute for the other 50% of the data.

Descriptive Data Summarization

Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

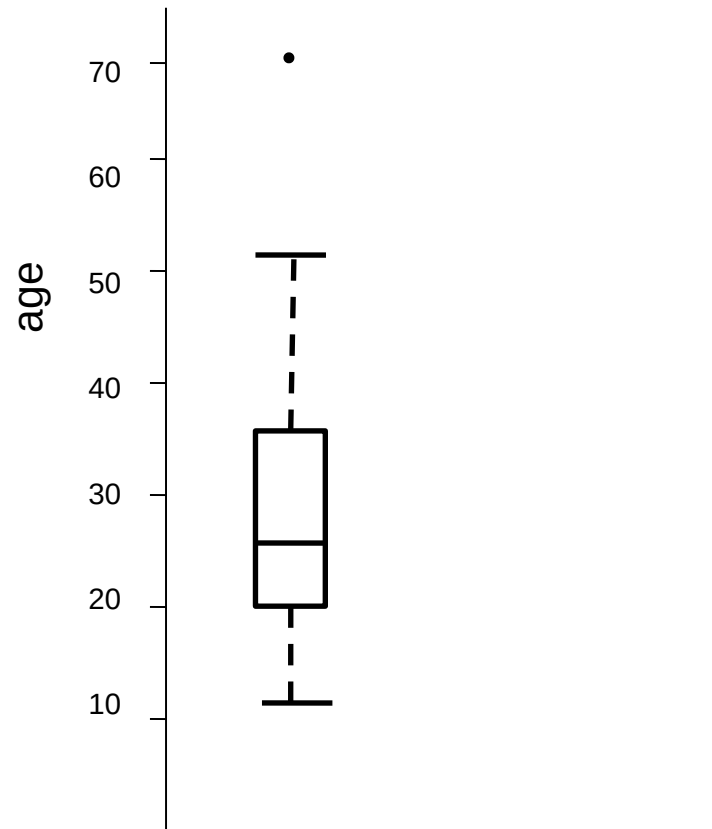
Can you find (roughly) the first quartile ($Q1$) and the third quartile ($Q3$) of the data?

The first quartile (corresponding to the 25th percentile) of the data is: 20. The third quartile (corresponding to the 75th percentile) of the data is: 35.

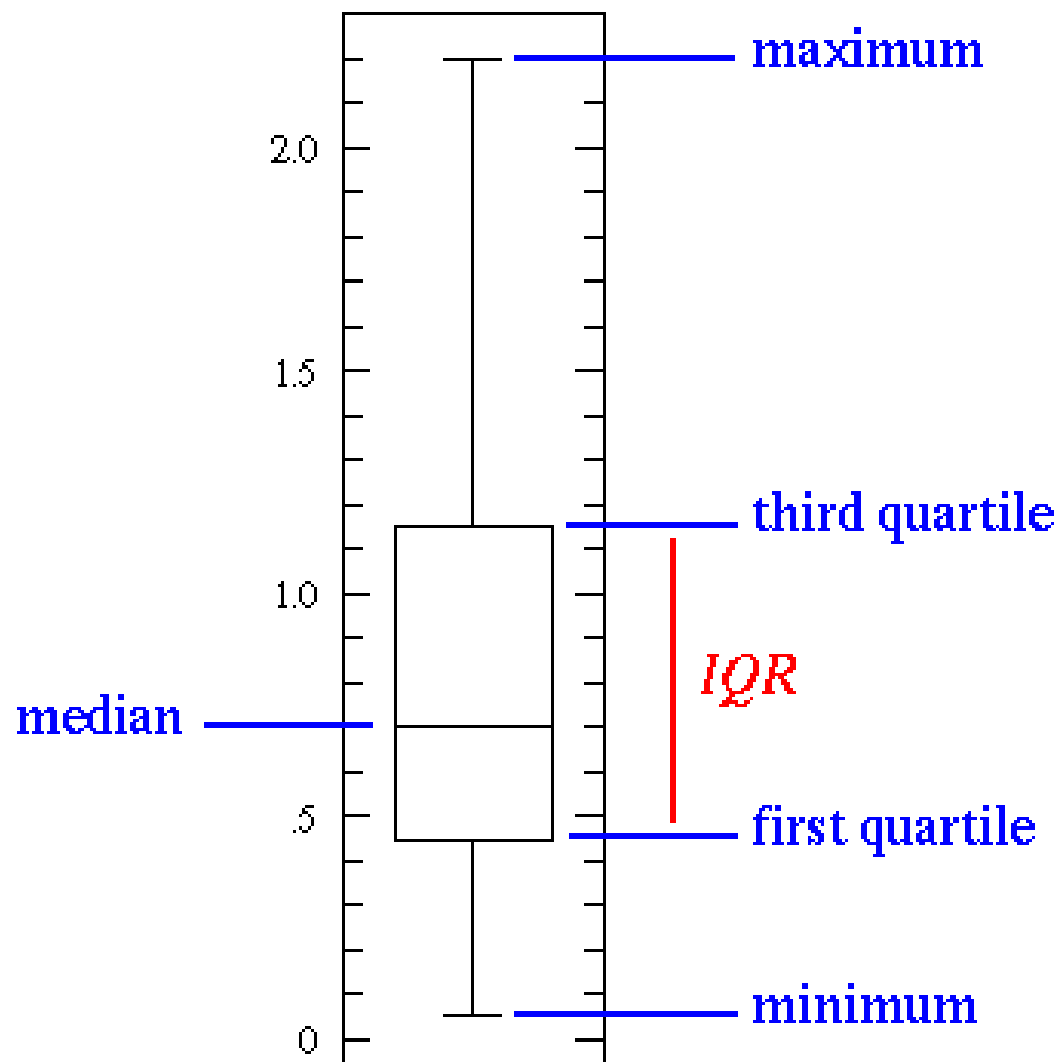
Descriptive Data Summarization

Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

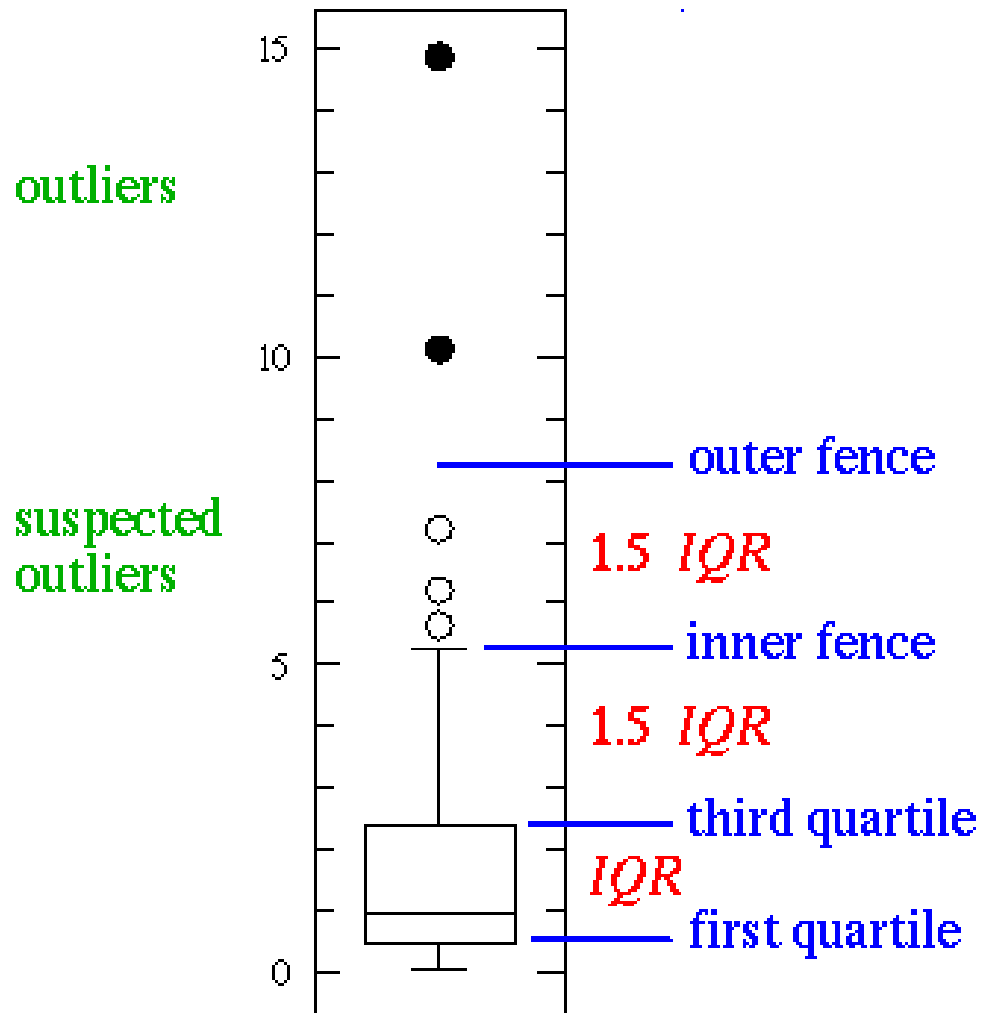
Show a *boxplot* of the data



Box Plot



Box Plot (cont.)



Variance and Standard Deviation

The variance of N observations is:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 = \frac{1}{N} \left[\sum x_i^2 - \frac{1}{N} (\sum x_i)^2 \right]$$

Square root of the variance is called **standard deviation**.

Variance and Standard Deviation

The variance of N observations is:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 = \frac{1}{N} \left[\sum x_i^2 - \frac{1}{N} (\sum x_i)^2 \right]$$

Square root of the variance is called **standard deviation**.

Ex.

Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result.

<i>age</i>	23	23	27	27	39	41	47	49	50
<i>%fat</i>	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
<i>age</i>	52	54	54	56	57	58	58	60	61
<i>%fat</i>	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

Calculate the mean, median and standard deviation of *age* and *%fat*.

Variance and Standard Deviation

The variance of N observations is:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 = \frac{1}{N} \left[\sum x_i^2 - \frac{1}{N} (\sum x_i)^2 \right]$$

Square root of the variance is called **standard deviation**.

Ex.

Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result.

<i>age</i>	23	23	27	27	39	41	47	49	50
<i>%fat</i>	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
<i>age</i>	52	54	54	56	57	58	58	60	61
<i>%fat</i>	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

Calculate the mean, median and standard deviation of *age* and *%fat*.

The mean is 46.44, the median is 51, and the standard deviation is 12.85. For the variable *%fat* the mean is 28.78, the median is (???), and the standard deviation is 8.99

Variance and Standard Deviation

The variance of N observations is:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 = \frac{1}{N} \left[\sum x_i^2 - \frac{1}{N} (\sum x_i)^2 \right]$$

Square root of the variance is called **standard deviation**.

Ex.

Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result.

<i>age</i>	23	23	27	27	39	41	47	49	50
<i>%fat</i>	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
<i>age</i>	52	54	54	56	57	58	58	60	61
<i>%fat</i>	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

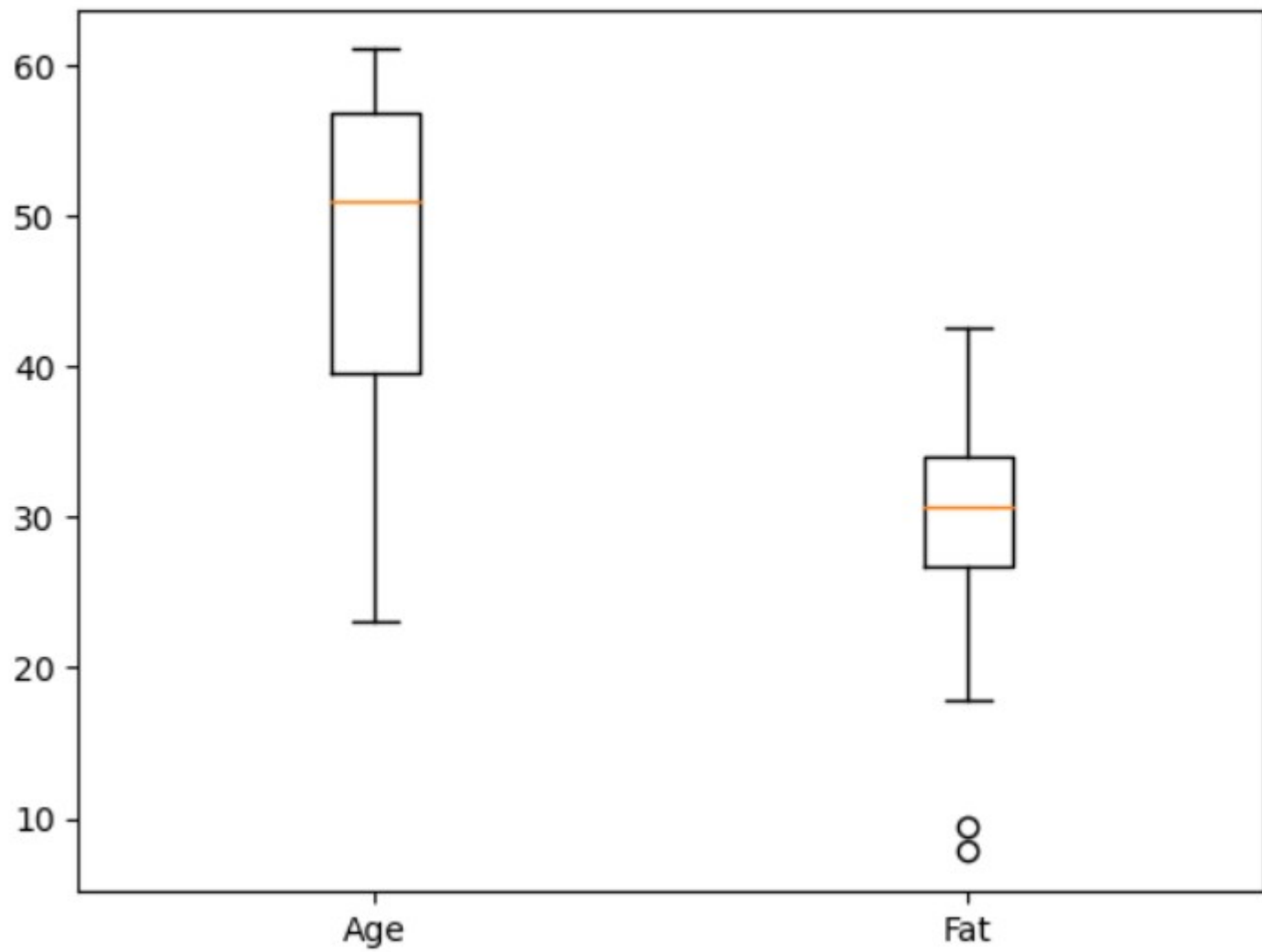
Calculate the mean, median and standard deviation of *age* and *%fat*.

The mean is 46.44, the median is 51, and the standard deviation is 12.85. For the variable *%fat* the mean is 28.78, the median is 30.7, and the standard deviation is 8.99

Boxplot

Draw the boxplots for *age* and *%fat*.

<i>age</i>	23	23	27	27	39	41	47	49	50
<i>%fat</i>	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
<i>age</i>	52	54	54	56	57	58	58	60	61
<i>%fat</i>	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7



Scatter Plot

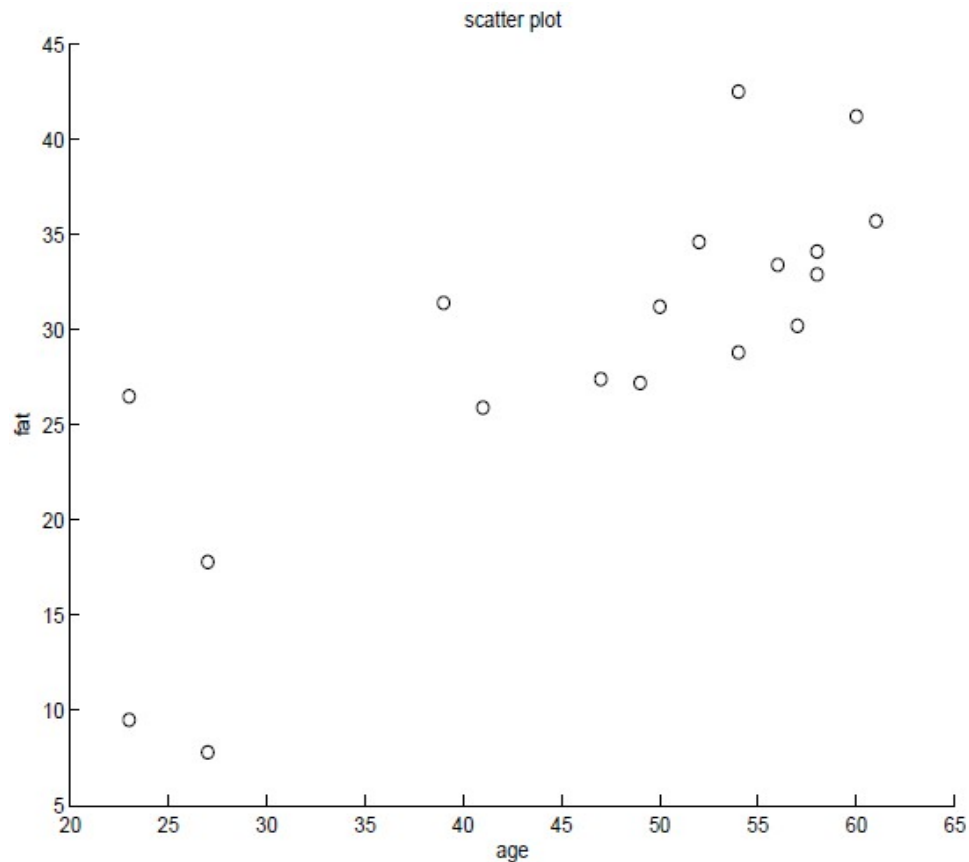
Draw a *scatter plot* based on these two variables.

<i>age</i>	23	23	27	27	39	41	47	49	50
<i>%fat</i>	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
<i>age</i>	52	54	54	56	57	58	58	60	61
<i>%fat</i>	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

Scatter Plot

Draw a *scatter plot* based on these two variables.

<i>age</i>	23	23	27	27	39	41	47	49	50
<i>%fat</i>	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
<i>age</i>	52	54	54	56	57	58	58	60	61
<i>%fat</i>	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7



Correlation

Calculate the *correlation coefficient* (Pearson's product moment coefficient). Are these two variables positively or negatively correlated?

<i>age</i>	23	23	27	27	39	41	47	49	50
<i>%fat</i>	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
<i>age</i>	52	54	54	56	57	58	58	60	61
<i>%fat</i>	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

Correlation

Calculate the *correlation coefficient* (Pearson's product moment coefficient). Are these two variables positively or negatively correlated?

<i>age</i>	23	23	27	27	39	41	47	49	50
<i>%fat</i>	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
<i>age</i>	52	54	54	56	57	58	58	60	61
<i>%fat</i>	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

The formula for Pearson's product moment coefficient is:

$$r_{A,B} = \frac{\sum_{i=1}^N (a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^N (a_i b_i) - N\bar{A}\bar{B}}{N\sigma_A\sigma_B}$$

Correlation

Calculate the *correlation coefficient* (Pearson's product moment coefficient). Are these two variables positively or negatively correlated?

<i>age</i>	23	23	27	27	39	41	47	49	50
<i>%fat</i>	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
<i>age</i>	52	54	54	56	57	58	58	60	61
<i>%fat</i>	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

The formula for Pearson's product moment coefficient is:

$$r_{A,B} = \frac{\sum_{i=1}^N (a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^N (a_i b_i) - N\bar{A}\bar{B}}{N\sigma_A\sigma_B}$$

0.82

Correlation

Calculate the *correlation coefficient* (Pearson's product moment coefficient). Are these two variables positively or negatively correlated?

<i>age</i>	23	23	27	27	39	41	47	49	50
<i>%fat</i>	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
<i>age</i>	52	54	54	56	57	58	58	60	61
<i>%fat</i>	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

The formula for Pearson's product moment coefficient is:

$$r_{A,B} = \frac{\sum_{i=1}^N (a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^N (a_i b_i) - N\bar{A}\bar{B}}{N\sigma_A\sigma_B}$$

0.82; since it's > 0 , they are positively correlated
Scatter plot also showed same thing. Refer to earlier slide...

Thank You!