Unit 3
Bivariate Distribution

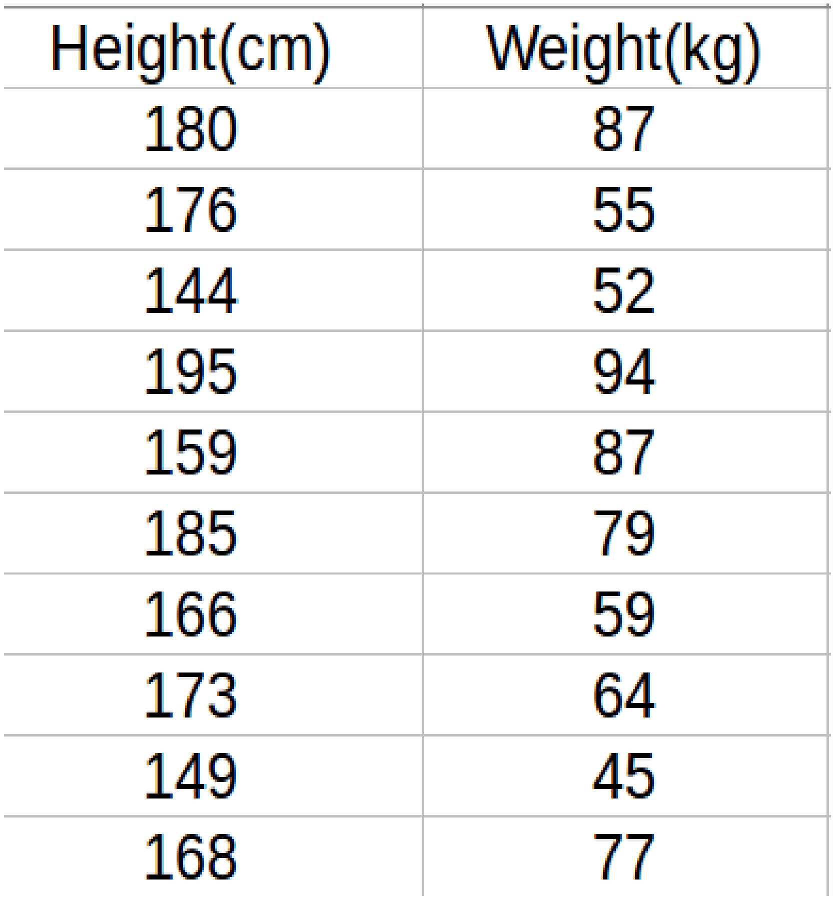
Bivariate Analysis

- Data consisting of
  - Two variables – bivariate data
  - More than two variables – multivariate data
- Most real time data is multivariate
- How can one visualize multivariate data ?
  - using sequence of two dimensions
  - Or using 3 dimensions
- Bivariate analysis can be done on two numerical variables or two categorical variables

Bivariate analysis on Numerical Data

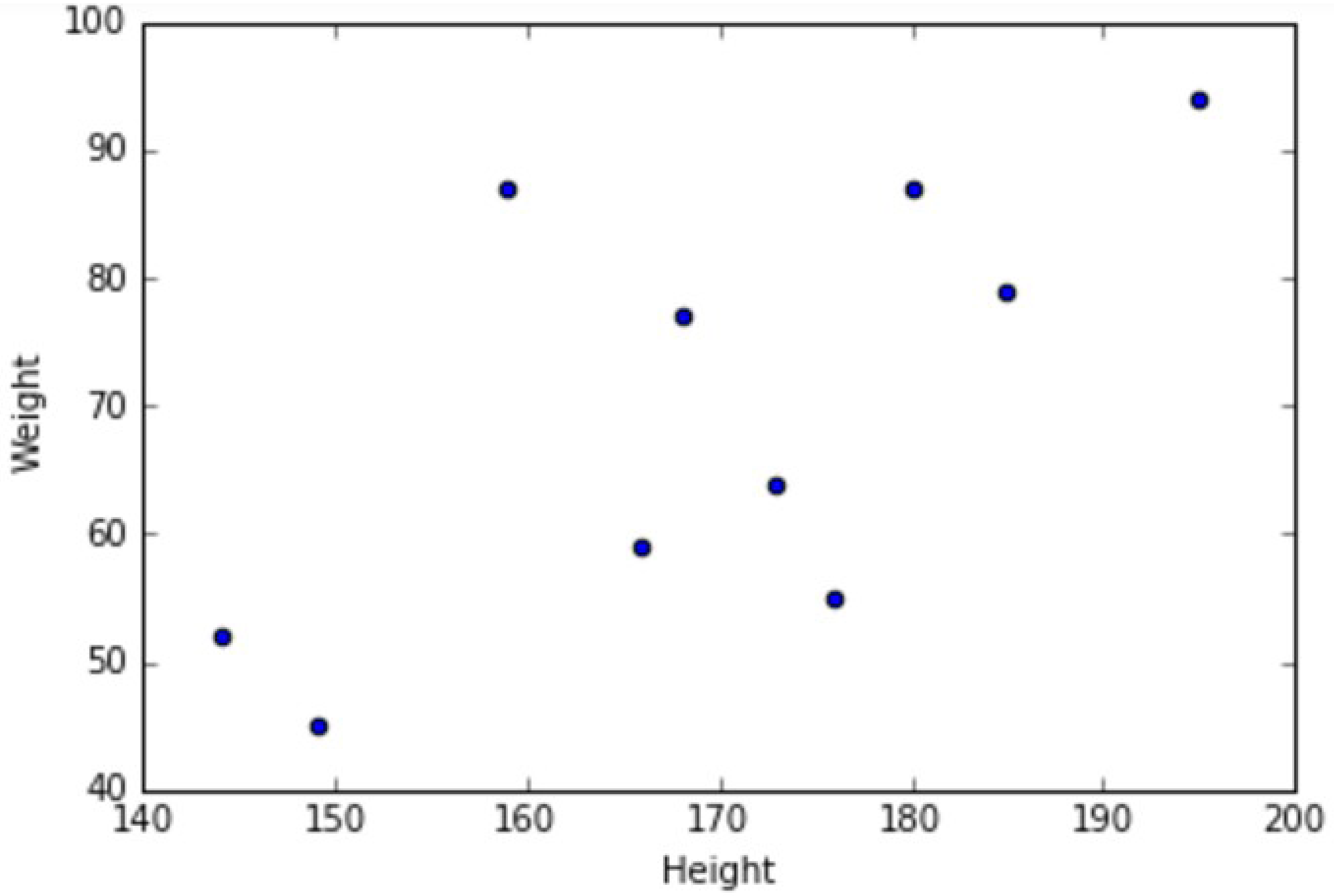
- Does the study of hours increases the % of the students?
- Does the regular attendance increases the % of students?
- To answer the above question, we have to go for bivariate analysis
- The idea behind this is to find any linear relationship or association between two variables.
- Correlation coefficient - The statistical tool that measures the strength and direction of the linear relationship between two numerical variables

Problem 1 – Construct a Scatter Plot



Height(cm)	Weight(kg)
180	87
176	55
144	52
195	94
159	87
185	79
166	59
173	64
149	45
168	77

Problem 1 - Solution

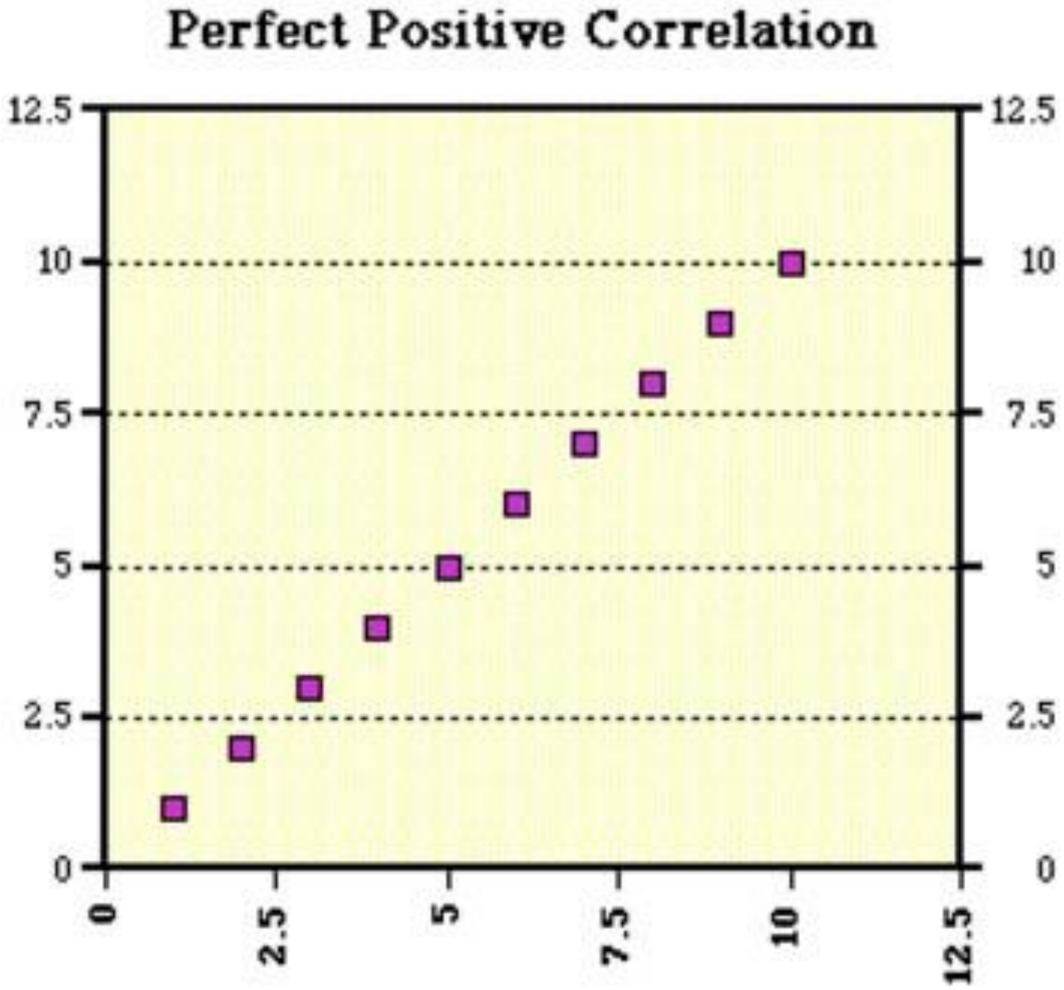


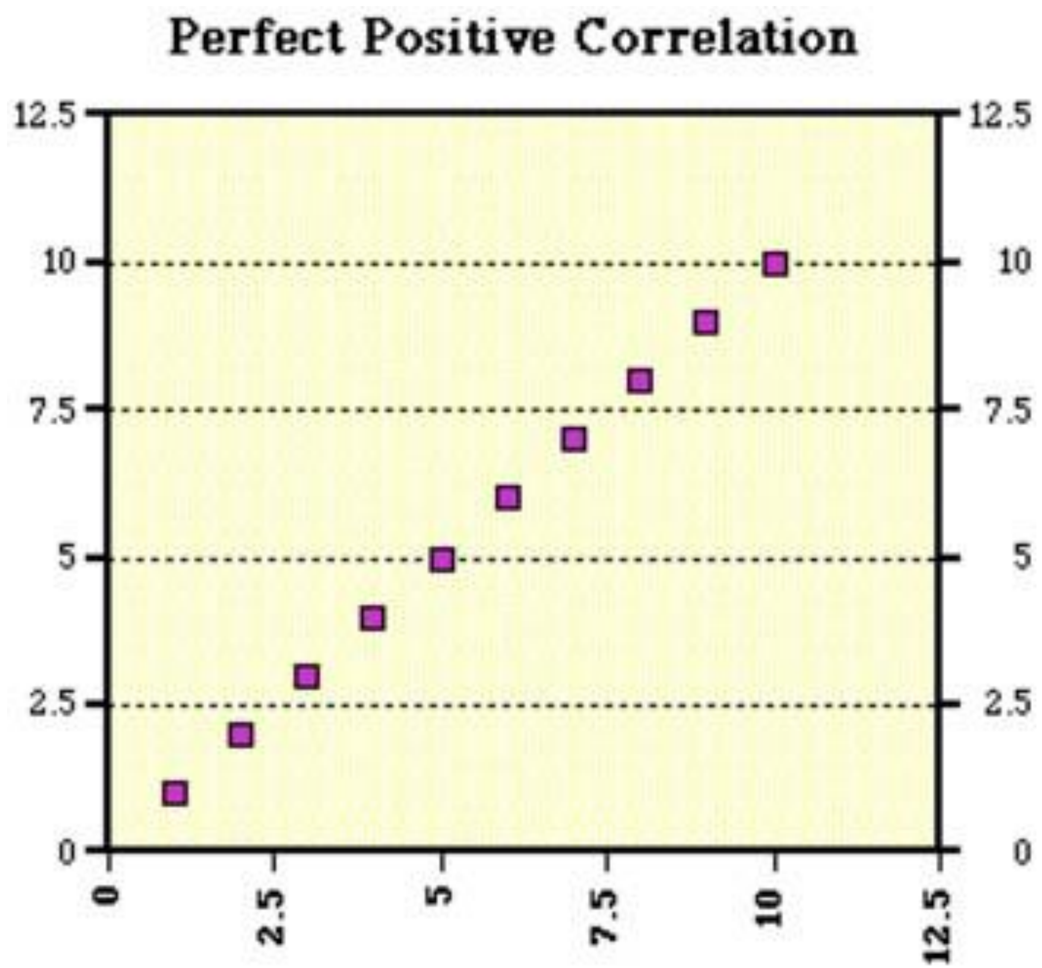
Analysing a Scatter Plot

- The relationship between two variables is called their **Correlation.**
- Correlation is a statistical measure that indicates the **extent to which two or more variables fluctuate together.**
- The closer the data points come when plotted to **making a straight line**, the higher the correlation between the two variables, or **stronger the relationship**.

Analysing a Scatter Plot

- **A positive correlation** indicates the extent to which those variables increase or decrease in parallel. A **perfect positive correlation** is given the **value of 1.**
- **A negative correlation** indicates the extent to which one variable increases as the other decreases. **A perfect negative correlation** is given the **value of -1.**
- If there is **absolutely no correlation** present the **value given is 0.**


Perfect Positive Correlation
12.5
10
7.5
5
2.5
0
12.5
10
7.5
5
2.5
0
0
2.5
5
7.5
10
12.5


Perfect Positive Correlation
12.5
10
7.5
5
2.5
0
12.5
10
7.5
5
2.5
0
0
2.5
5
7.5
10
12.5

- The closer the number is to 1 or -1, the stronger the correlation, or the stronger the relationship between the variables.
- The closer the number is to 0, the weaker the correlation.
- So something that seems to kind of correlate in a positive direction might have a value of 0.67, whereas something with an extremely weak negative correlation might have the value -0.21.

Examples

- **Perfect Positive correlation :** The total amount of money spent on tickets at the movie theatre with the number of people who go.
- **Negative correlation :** The number of days required to do a complete a project with programmers
- **Strong but not perfect positive correlation:** The number of hours students spent studying for an exam versus the grade received.

This won't be a perfect correlation because:

- Two people could spend the same amount of time studying and get different grades.

But in general the rule will hold true that as the amount of time studying increases so does the grade received.

Correlation is Not Causation

- When the fluctuation of one variable reliably predicts a similar fluctuation in another variable, there's often a tendency to think that means that the change in one causes the change in the other.
- However, correlation does not imply causation, which says that a correlation does not mean that one thing causes the other.
- There may be an unknown factor that influences both variables similarly.

Problem

- Which graph would have a correlation of:

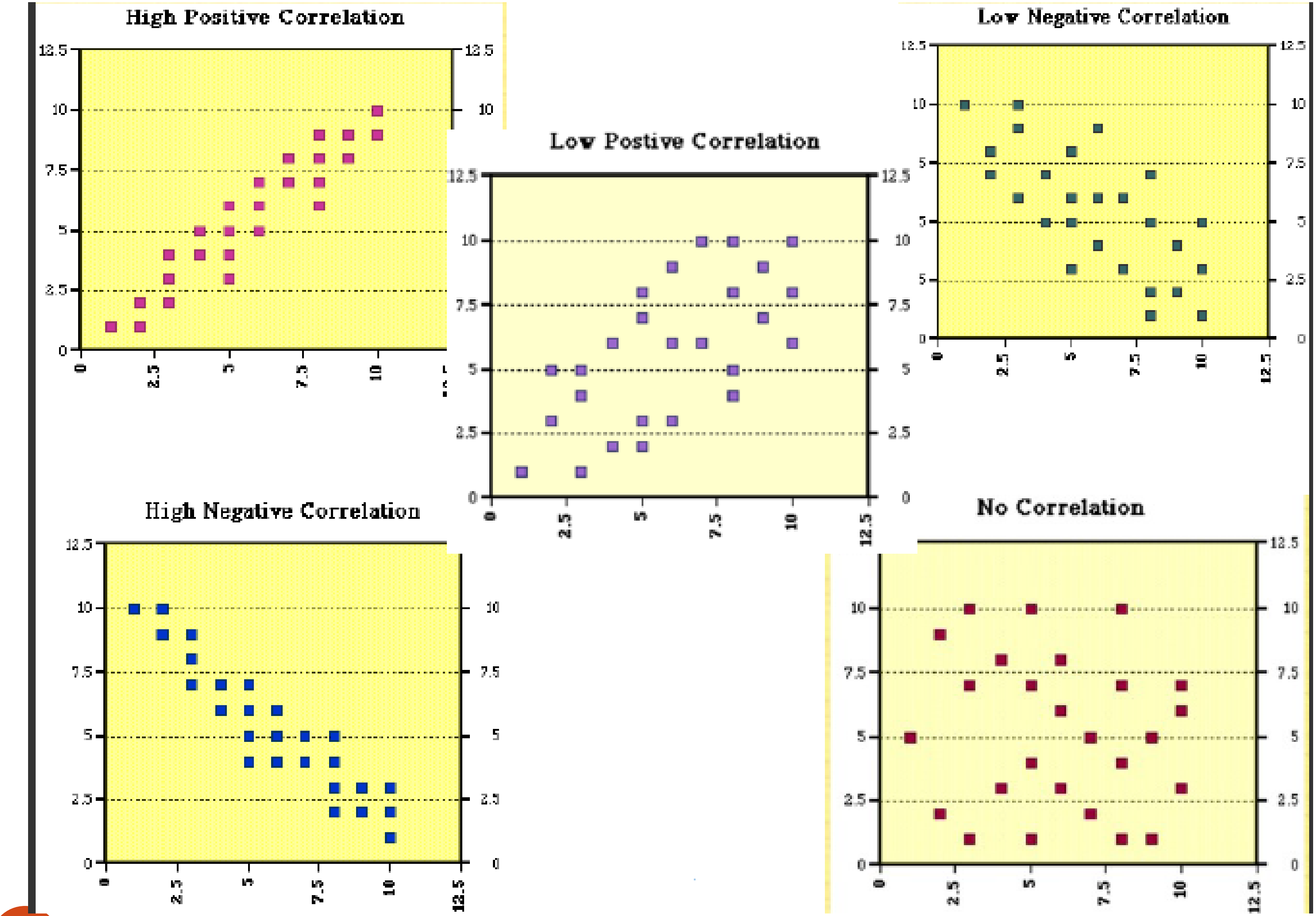
  1) 0 ?

  2) 0.7?

  3) -0.7?

  4) 0.3?

  5) -0.3?



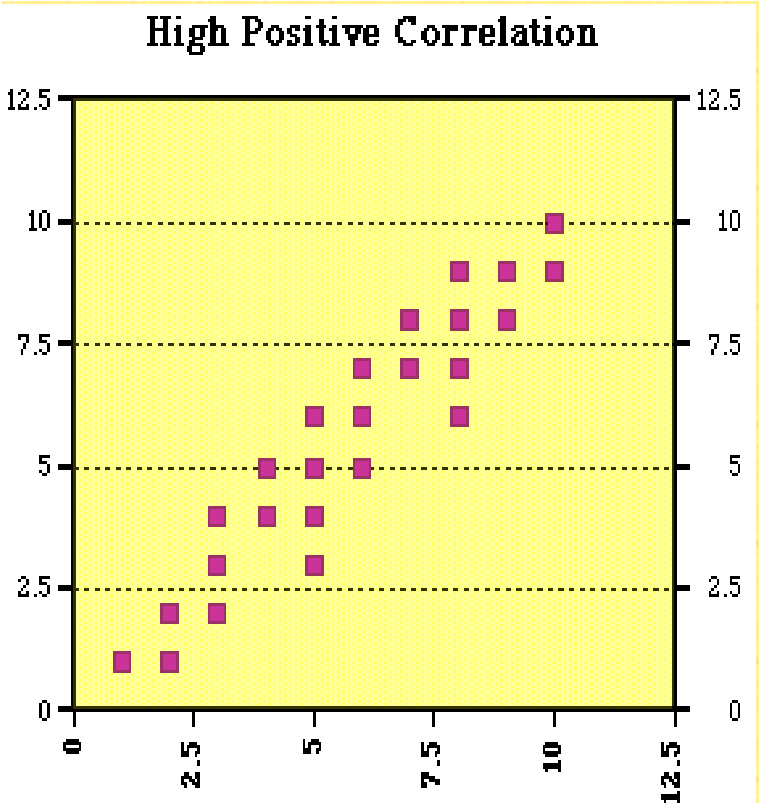
High Positive Correlation

Low Postive Correlation

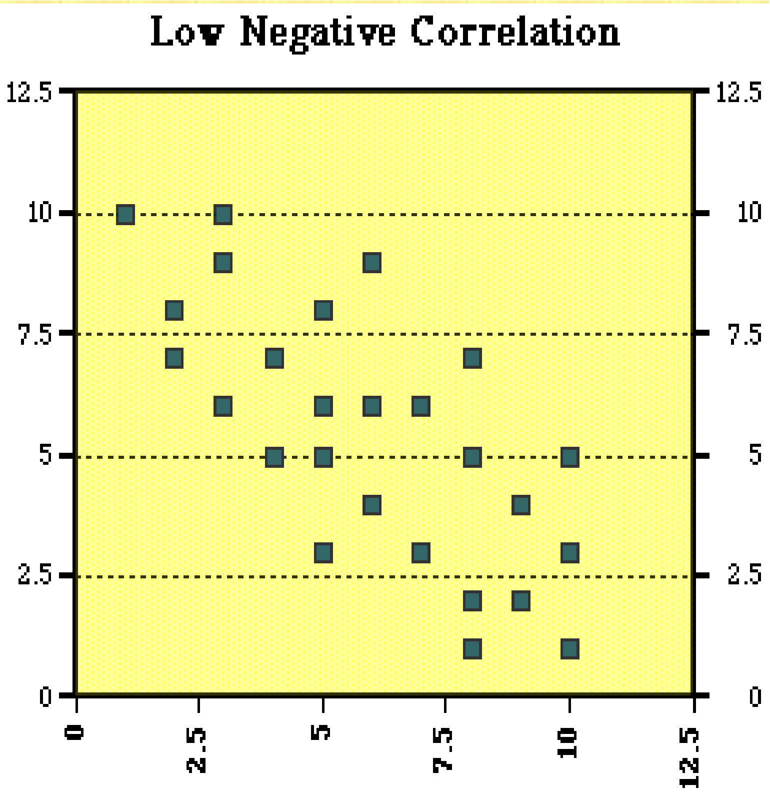
Low Negative Correlation

High Negative Correlation

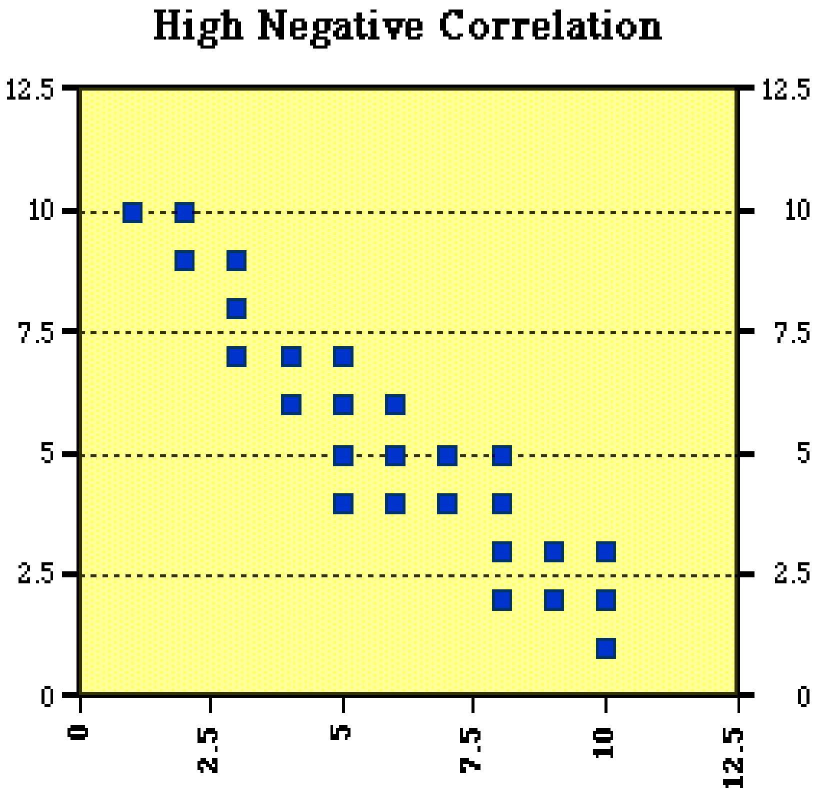
No Correlation



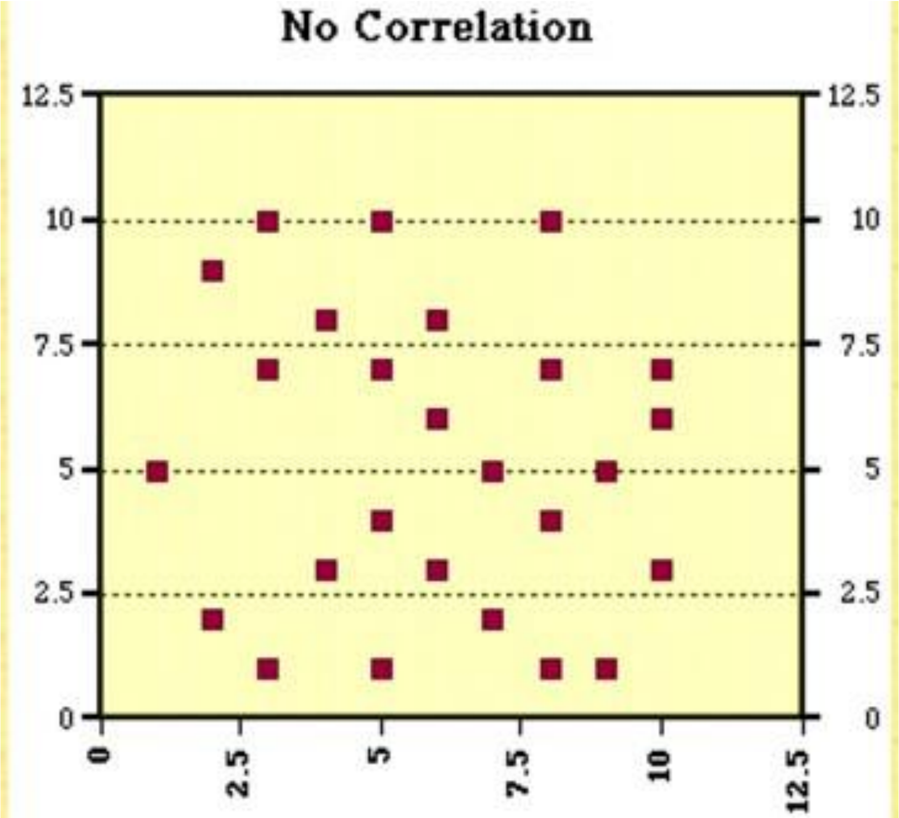
- The first graph seems to have a pretty strong positive correlation, so it would have a value of about 0.7. You can see that the band of data points that is angled upward is relatively thin so there is not a whole lot of variation in the results when one variable is entered.



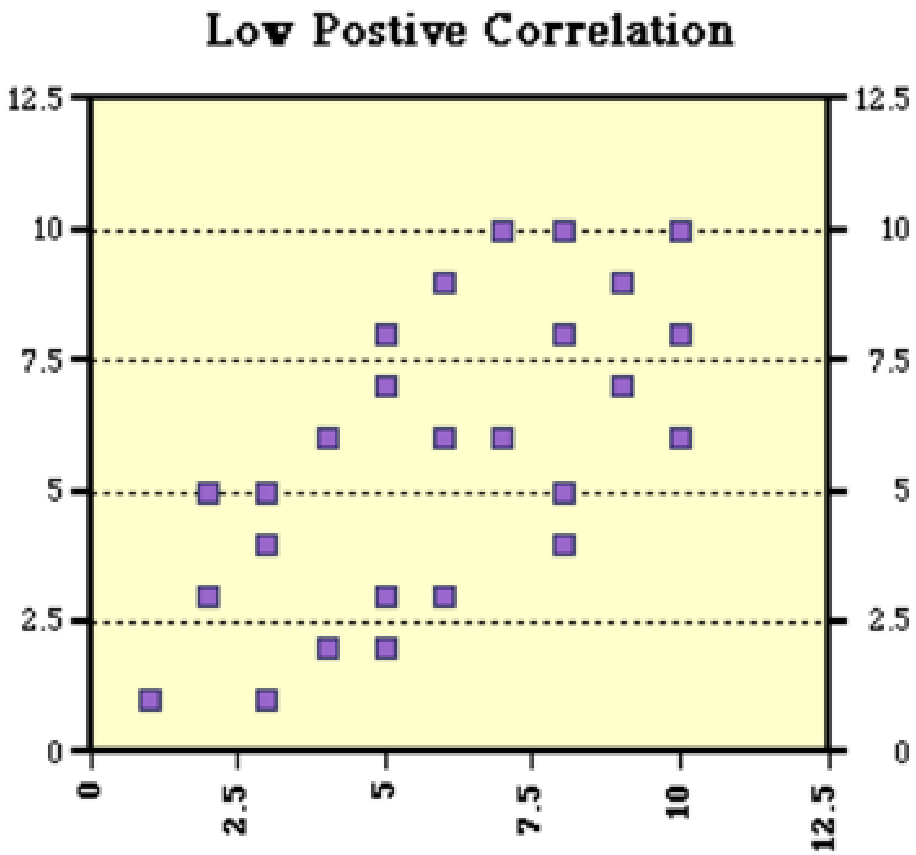
- The data points of the second graph are much more spread out, although they definitely follow a downward pattern. Therefore, it would be a good guess to say that this is roughly a -0.3 correlation.



- The third graph also has a negative correlation, but the data points are much tighter indicating a higher correlation. Therefore, this would probably have a value of about -0.7.



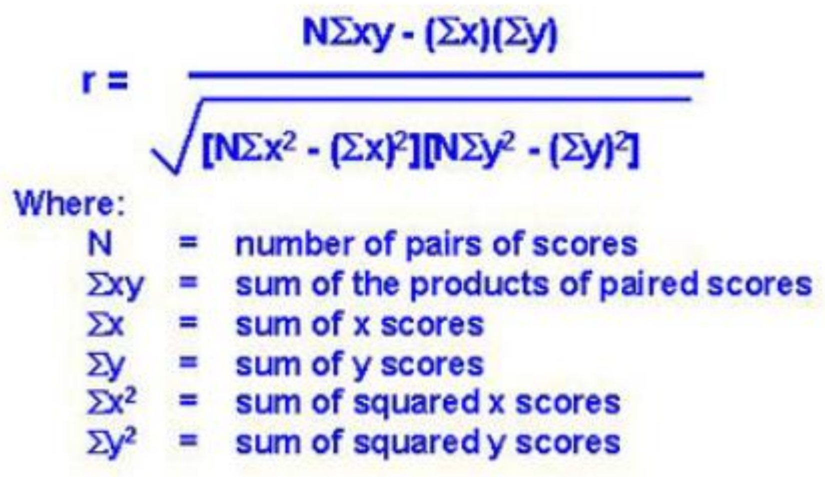
- The fourth graph does not seem to have a correlation at all. There is no pattern to where the data points lie. They do not seem to go in any particular direction. Therefore this data has a correlation value of 0.



- The last graph appears to have a positive correlation, although the data points are not very close together. This graph would probably have a value of 0.3.

How does correlation coefficient calculated?

- Is it the right way to scale two variables where their measurement is in different scale? For eg: study time is measured in hours and marks in units.
- Earlier Galton correlation was done using absolute values.
- Later Pearson correlation was proposed, which is done based on z scores. i.e, it is done on the mean and sd. Hence the ideal correlation coefficient method is pearson correlation coefficient. It is represented by r. It is also represented as Pearson R test.
- Correlation coefficient methods – pearson, kendall, spearman



$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

N	=	number of pairs of scores
Σxy	=	sum of the products of paired scores
Σx	=	sum of x scores
Σy	=	sum of y scores
Σx^2	=	sum of squared x scores
Σy^2	=	sum of squared y scores

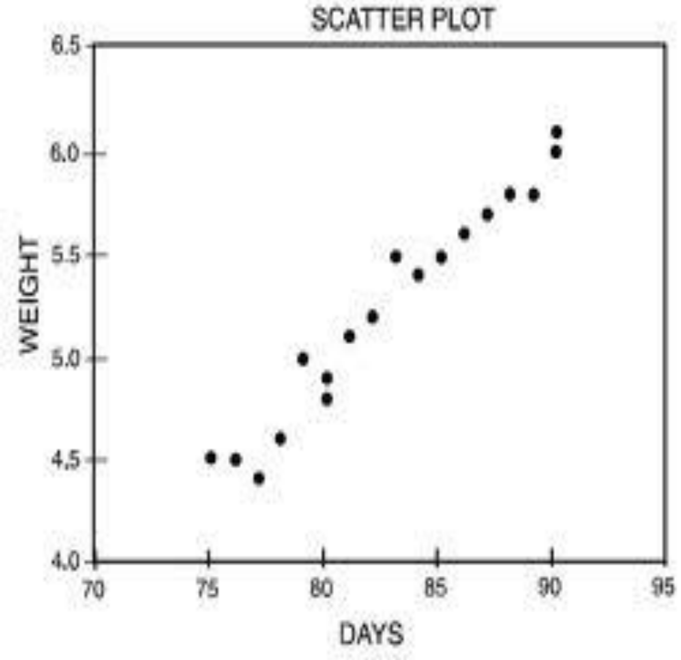
Person	Age(x)	Score(y)	xy	x^2	y^2
1	20	30	600	400	900
2	24	20	480	576	400
3	17	27	459	289	729
Total	61	77	1539	1265	2029

Quiz

- A correlation tell us
  - how much one variable causes another to vary.
  - the direction and strength of a linear relationship between two quantitative variables.
  - the frequency of scores for a quantitative variable.
  - the number of data points in a scatterplot that are outliers.

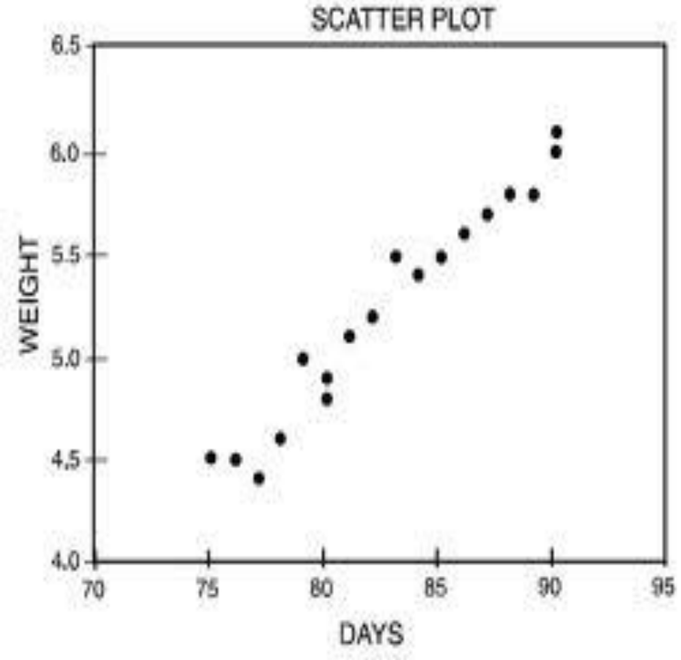
Quiz

- **A researcher plotted the weights of each of his lab mice against their age (in days). This produced the following scatterplot**
- How would you describe the direction of the relationship shown in this scatterplot?
  - Positive correct
  - Negative
  - There is no relationship



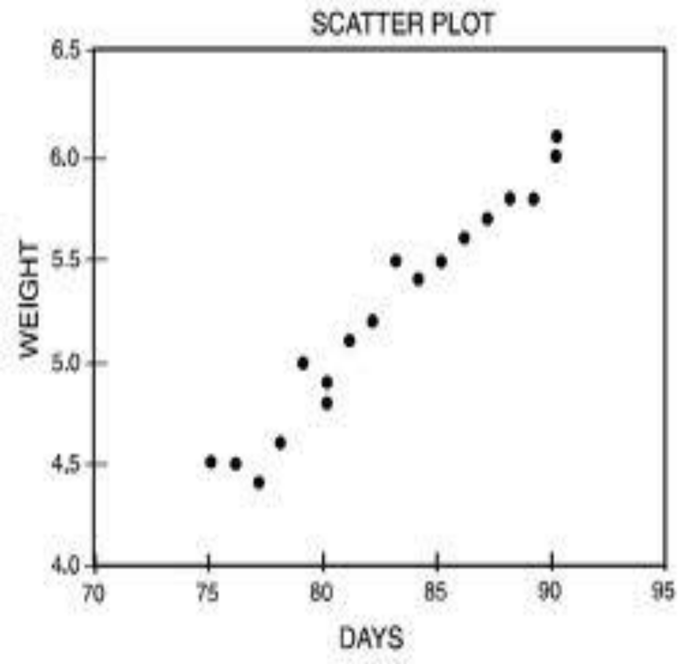
Quiz

- How would you describe the strength of the relationship in this scatterplot?
  - Weak (the data points are widely scattered)
  - Moderately strong (there is a clear pattern, but there are a few significant outliers)
  - Very strong (nearly all the data points sit on or near the linear trend line)
  - None (there is no visible relationship)



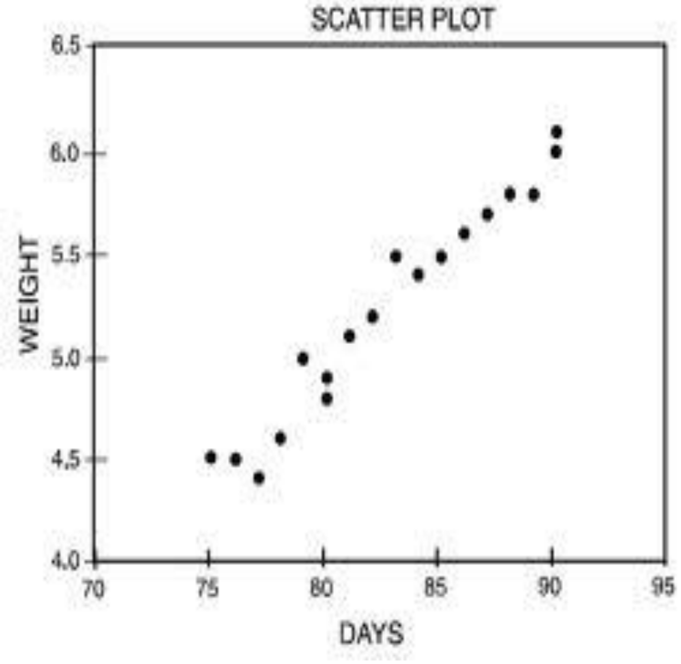
Quiz

- Which of the following best describes the relationship between these two variables?
  - As mice age, their growth rate slows down.
  - The mice get heavier as they age.



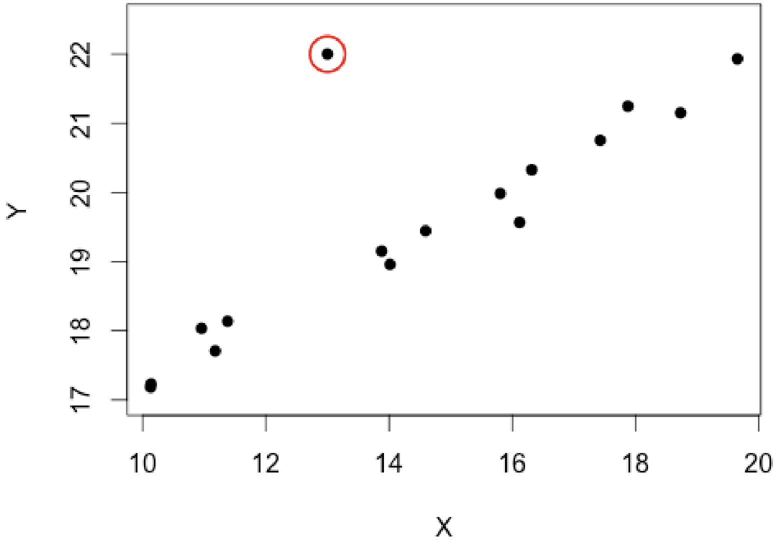
Quiz

- Which would be the most likely value of the Pearson correlation coefficient, r, for this relationship between age and weight?
  - -0.87
  - 0.56
  - 0.93 correct
  - 1.00



Quiz

- Look at the scatterplot below. Select the answer that best describes what would happen to the value of the correlation coefficient r_{xy} if the circled point were removed from the analysis.
  - The value of r_{xy} would increase. correct
  - Removing the outlier would have no effect on the correlation coefficient.
  - The value of r_{xy} would decrease.
  - The circled point is not an outlier. It fits with the trend of the data.



How to do it R?

- To find correlation between two vectors
  - cor(x,y)
- To find correlation between multiple vectors (correlation matrix)
  - Make the subset of continuous variables
  - cor(subset)
- Correlation graph for two variable
  - plot(x,y) #scatter plot
  - abline(lm(y~x)) #trendline
  - text(x~y, labels=col_name,cex, pos, offset) #adding labels to the plot
  - 1, 2, 3 or 4 down, left, up or right

How to do it R?

- Matrix plot for correlation matrix
  - Make the subset of continuous variables
  - Result = Corrplot(subset)
  - Convert the result into matrix
  - Library(corrplot)
  - corrplot(result_matrix, method,type)
    - method – circle,number, pie,color
    - type – upper, lower

Removal of the outlier

- **Do scatter plot**
- **Draw the regression line**
- **Look for any outlier value in the plot**
- **If it is above the line find max, if it is below the line find min**
  - **Find the observation which has that value**
    - **which(y == max(y))**
  - Remove that observation from the dataset
    - Dataset=dataset[-which_result,]
- Calculate again the correlation coefficient

How to do for categorical data?

- Suppose, you want find whether there is a relationship between Gender and grade?
- Is there a possible way ??? Contingency table
- Contingency table - Intersection of characteristics of two categorical variables

		Grade					**Total**
Gender		S	A	B	C	D	
	M						
	F						
Total							

Marginal distribution - Gender

Marginal Distribution

- Table percentage – usually ignored. Generally don't answer out question
- Row percentages
- Column Percentages
- Find out which type of percentage tells more impactful story on the data
- Examine the conditional distribution with marginal distribution
- Among the two variables, determine which is the outcome variable

Conditional Probability/distribution

- P(A|B)=P(A) => No Relationship exists .
- If it fails then there is a relationship between two variables
- If they do not match with one another or come close to match one another => relationship between two variables

How to determine relationship?

- Determine the variable of interest
- Determine the marginal distribution
- Determine the conditional distribution
- Relationship
  - Compare conditional distribution and the marginal distribution of each class, if they donot match with one another or come close to match one another, then we have a relationship between i.e, they are dependent with each other
  - If $P(A)=P(A|B)$, then they are independent i.e, they do not have any relationship

Quiz

Given is a contingency table showing data from a University of Texas Southwestern Medical Center study on Hepatitis C.

- How many simple events were possible for participants in this study?
- What was the total number of participants in this study?
- What was the marginal distribution for Hepatitis status in this study?

	Tattoo in Commercial Parlor	**Tattoo Done Elsewhere**	**No Tattoo**	**Total**
Has Hep C	17	8	18	43
Does Not	35	53	495	583
Total	52	61	513	626

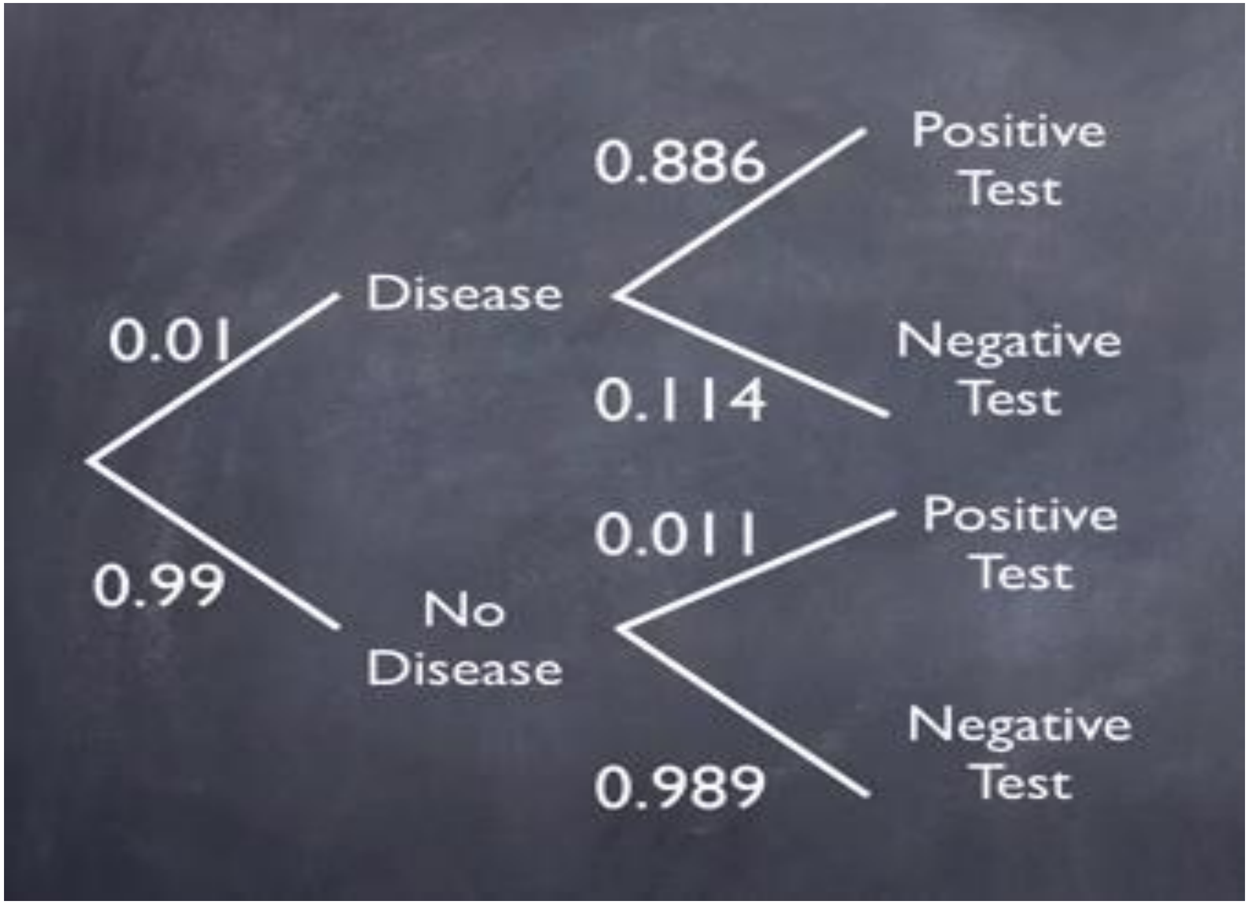
Quiz

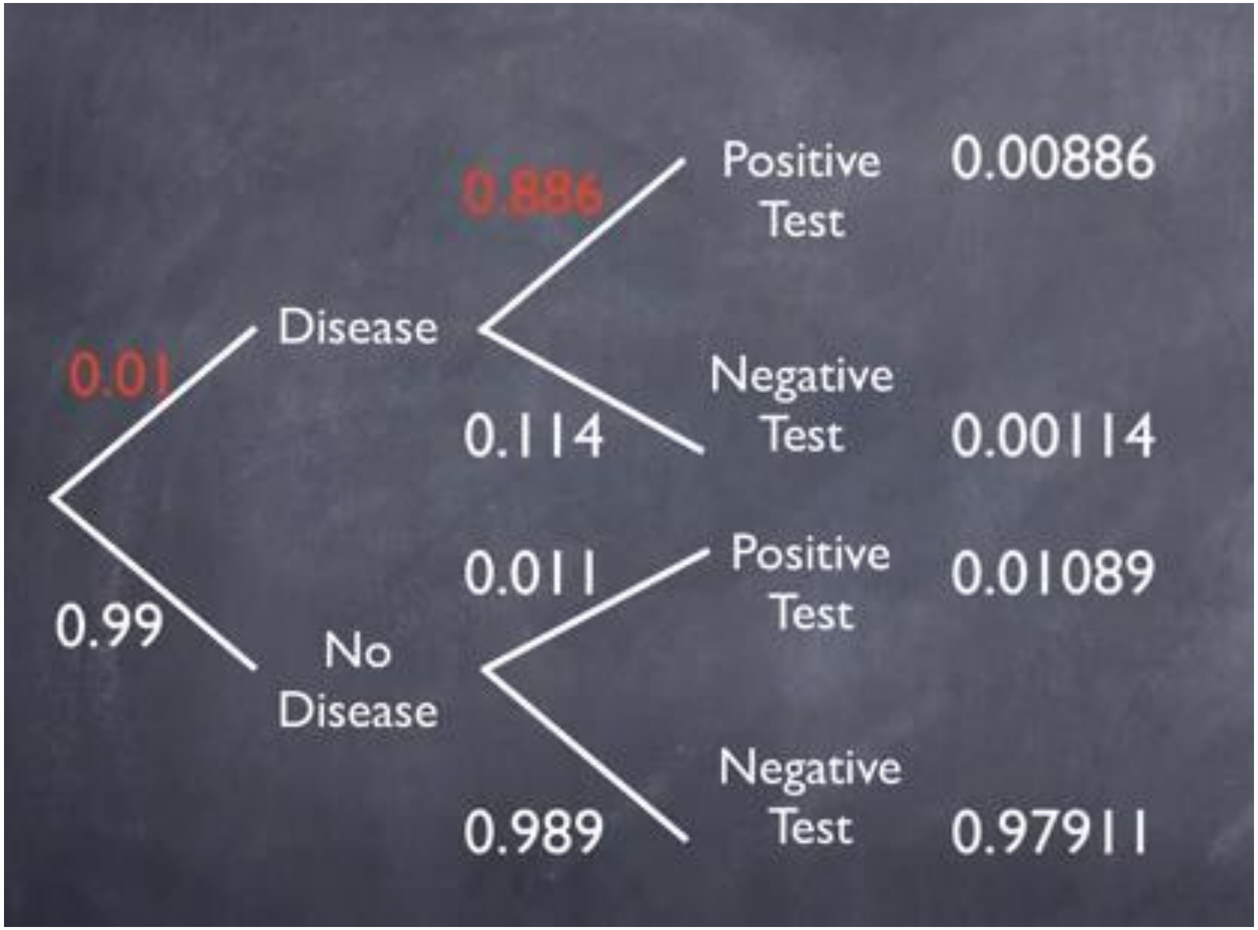
- Overall, what percentage of participants had a tattoo?
- What percentage of those participants with Hepatitis C had a tattoo done in a commercial parlor
- What percentage of those who had a tattoo done in a commercial parlor have Hepatitis C?

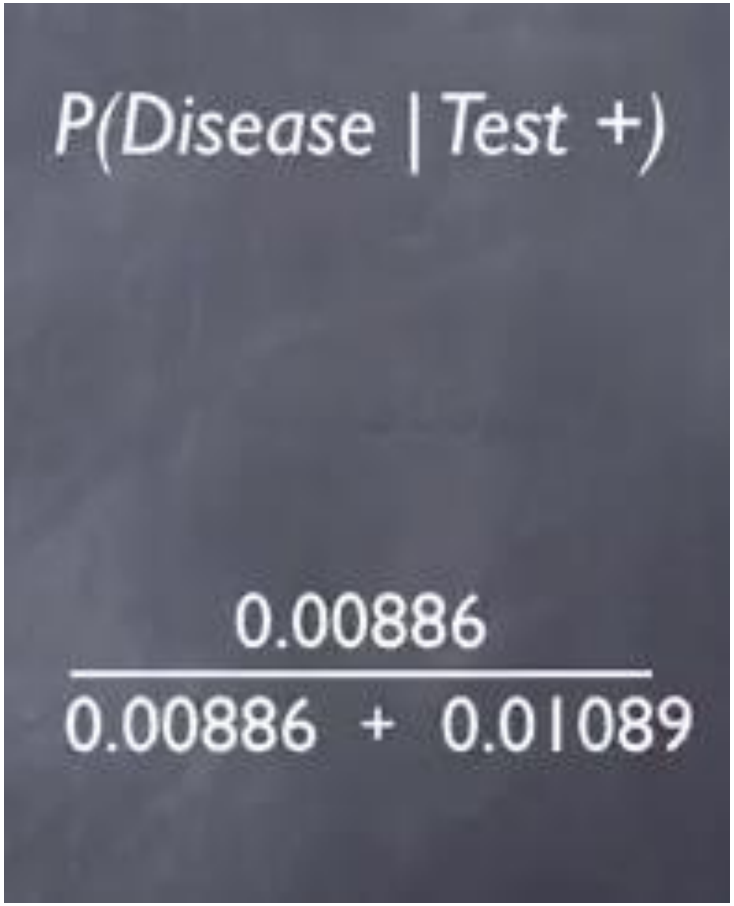
	Tattoo in Commercial Parlor	**Tattoo Done Elsewhere**	**No Tattoo**	**Total**
Has Hep C	17	8	18	43
Does Not	35	53	495	583
Total	52	61	513	626

Reverse Conditional Probability

- What is the probability of having a disease given that you have tested positive?
- Sensitivity- probability of testing positive in the presence of the disease
- Specificity – probability of testing negative when the disease is not present
- P(Disease | test+)


0.01
0.99
Disease
No Disease
0.886
0.114
0.011
0.989
Positive Test
Negative Test
Positive Test
Negative Test


0.01
Disease
0.99
No
Disease
0.886
0.114
0.011
0.989
Positive
Test
Negative
Test
Positive
Test
Negative
Test
0.00886
0.00114
0.01089
0.97911



P(Disease | Test +)

$$\frac{0.00886}{0.00886 + 0.01089}$$

Bivariate analysis in R

- Analysis
  - Contingency table - table(x,y)
  - Proportion table – prop.table(table(x,y))
  - Row conditional probability – prop.table(table(x,y),1)
  - Column conditional probability – prop.table(table(x,y),2)
- Plotting
  - Frequency - barplot(table(x,y),legend=T)
  - Relative frequency – barplot(prop.table(table(x,y),2))