

Week 23 Correlation

The techniques of correlation and regression are used to determine linear relationships between two variables.

Correlation is concerned with measuring how well a straight line describes the relationship between the two variables.

Regression is concerned with finding the equation of the 'best' straight line to describe the relationship.

Case Study

The Society of London Theatres (SOLT) collates data each year relating to the attendance figures in the London theatres.

The data published in January 2012 is shown below (the excel file is available in the "Data" folder

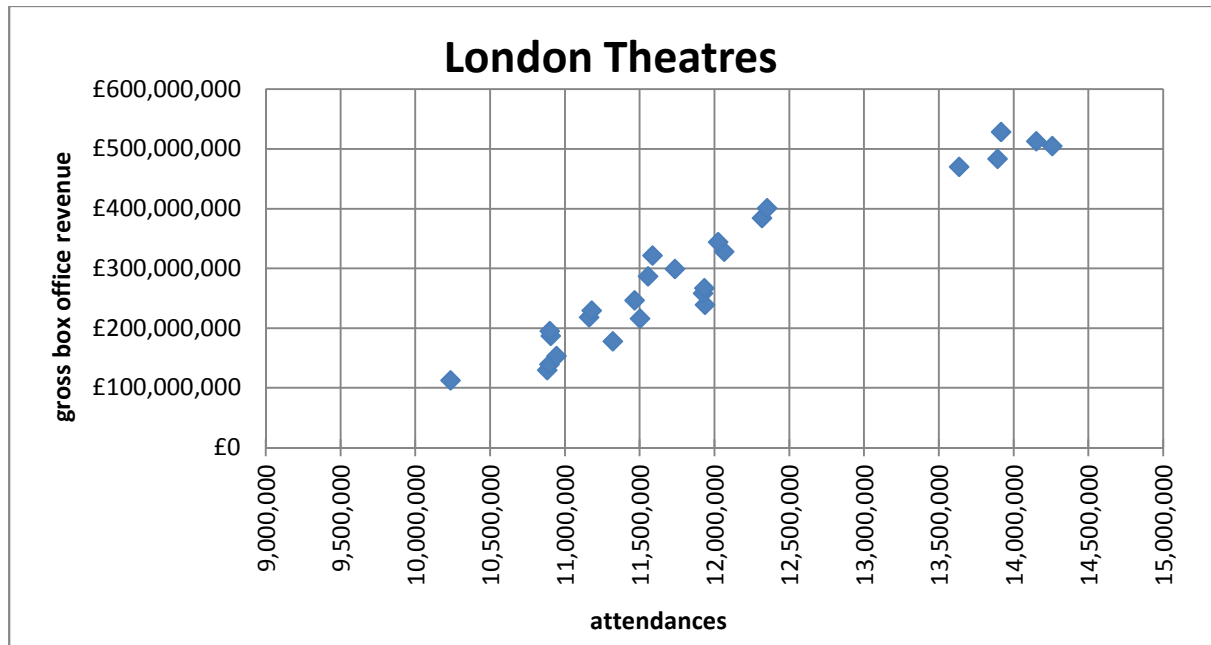
year	attendances	Revenue	number of theatres open	number of performances	number of new productions
1986	10,236,362	£112,067,737	42	16543	213
1987	10,880,791	£129,586,817	42	16603	212
1988	10,897,384	£139,338,056	43	16970	228
1989	10,944,760	£153,250,633	42	16436	237
1990	11,321,288	£177,903,841	40	15887	187
1991	10,905,385	£186,789,856	39	15508	192
1992	10,899,928	£194,772,331	41	15916	193
1993	11,503,358	£215,619,208	41	15922	198
1994	11,163,023	£217,762,521	41	16063	208
1995	11,937,671	£238,740,636	43	17163	208
1996	11,179,326	£229,016,639	41	16084	186
1997	11,466,248	£246,081,850	39	15568	195
1998	11,924,518	£257,920,278	41	16018	207
1999	11,931,954	£266,565,215	44	17089	265
2000	11,555,178	£286,557,500	43	16633	252
2001	11,734,767	£298,989,461	44	17035	264
2002	12,064,100	£327,971,671	44	17090	221
2003	11,585,446	£321,485,161	42	16664	225
2004	12,025,073	£343,674,090	43	17235	225
2005	12,318,625	£383,941,704	45	17406	221
2006	12,351,495	£400,852,925	43	16912	268
2007	13,636,212	£469,939,749	44	17455	243
2008	13,892,460	£483,349,423	45	18275	241
2009	14,257,922	£504,765,690	45	17899	260
2010	14,152,230	£512,331,808	46	18615	264
2011	13,915,185	£528,375,874	n/a	18061	256

http://solt.co.uk/downloads/pdfs/pressroom/london_box_office_figures_2011.pdf

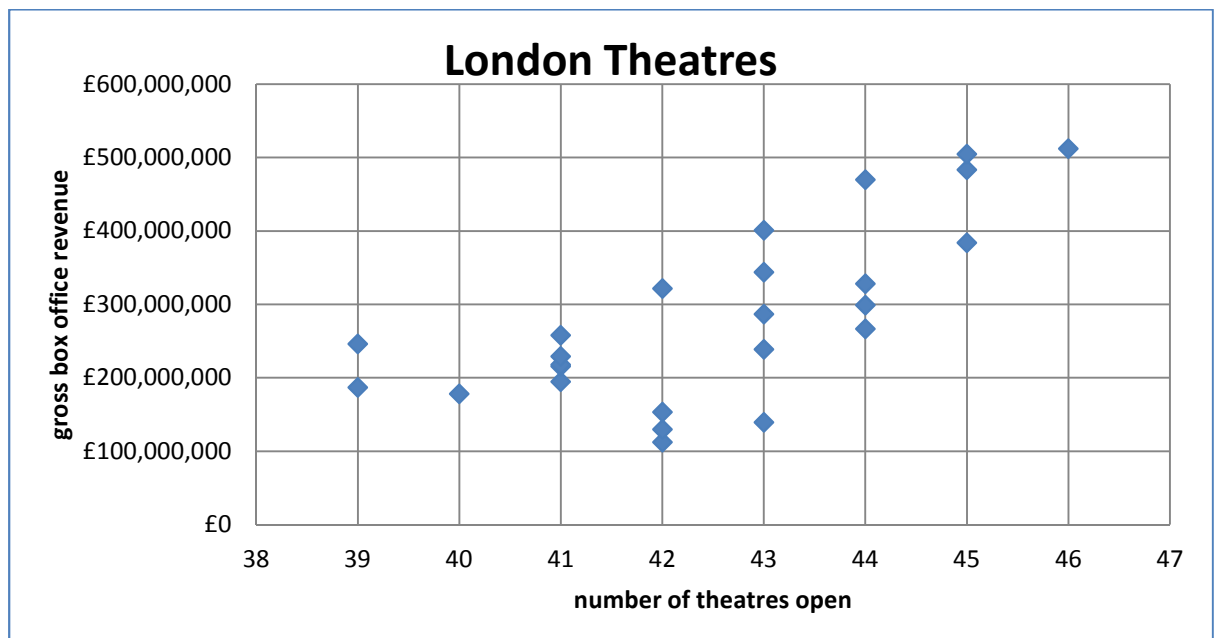
We are going to decide which other variable is the best predictor for “ gross box office revenue”

In order to judge which other variable might be the best predictor for revenue, we can use scatter graphs.

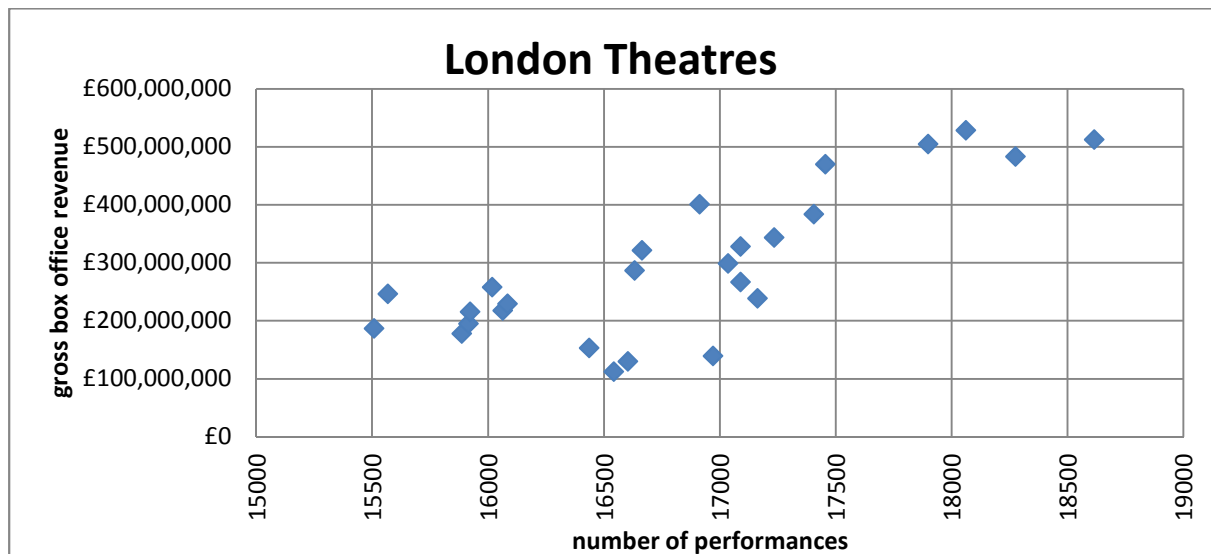
1. Gross box office Revenue and Attendance



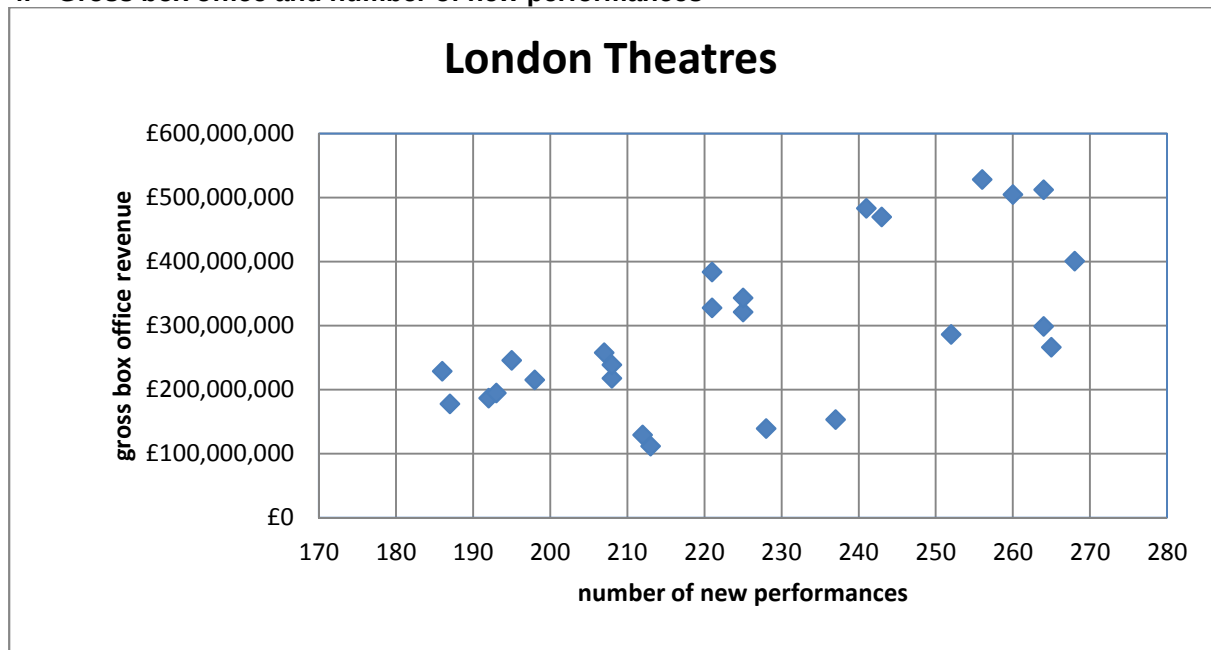
2. Gross box office revenue and number of theatres open



3. Gross box office and number of performances



4. Gross box office and number of new performances



From these scatter graphs we can see that:

- The higher the number of attendances at London theatres, the higher the box office revenue
- The higher the number of theatres open, the higher the box office revenue
- The higher the number of performances at London theatres, the higher the box office revenue
- The higher the number of new performances at London theatres, the higher the box office revenue

Which one looks best for prediction purposes?

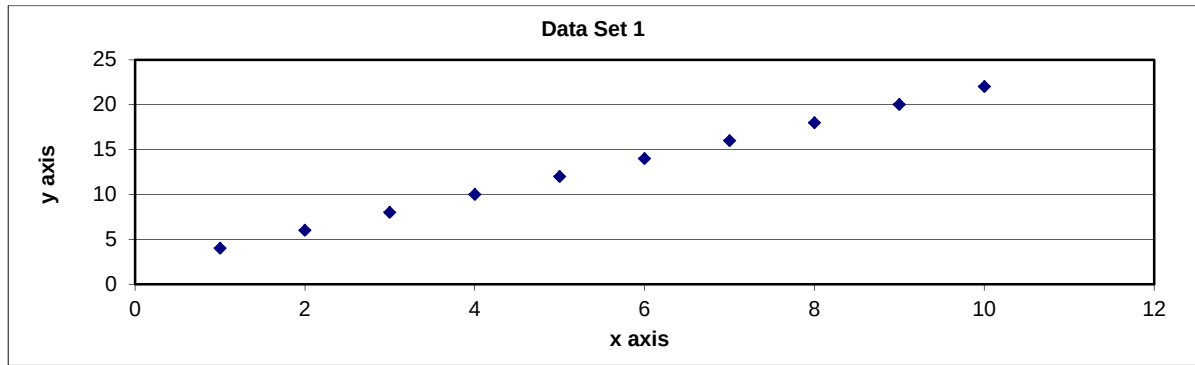
Correlation

Correlation measures the degree to which the relationship between the two variables is linear. The following examples show 4 sets of data and the corresponding scatter diagrams.

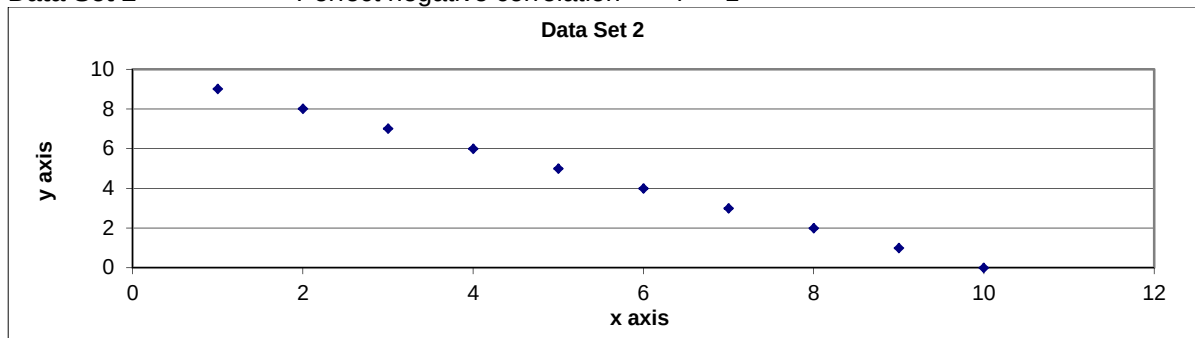
Data Set 1

Perfect positive correlation

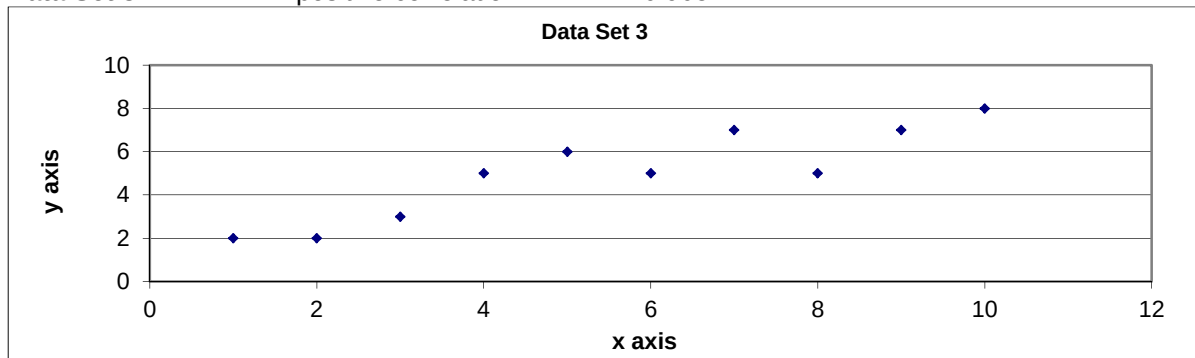
$r = +1$



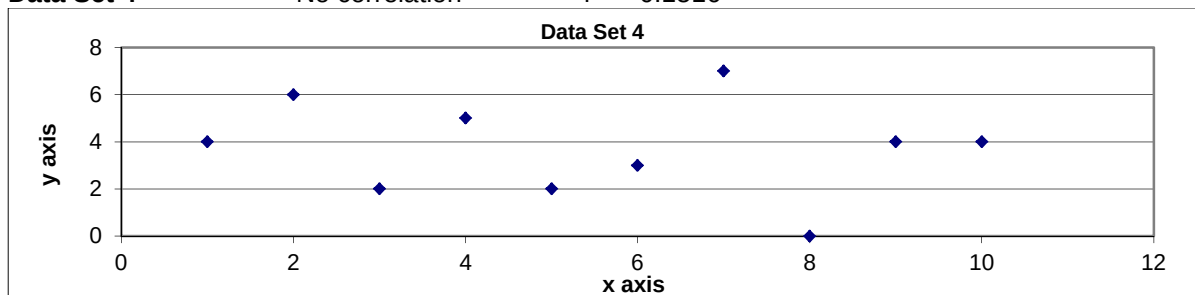
Data Set 2 Perfect negative correlation $r = -1$



Data Set 3 positive correlation $r = +0.9052$



Data Set 4 No correlation $r = -0.1516$



The term '**perfect** correlation' applies only when the points on the scatter diagram lie **exactly** on a straight line.

The term '**positive** correlation' applies when the 'best' straight line through the points will have a **positive gradient**.

'**Negative** correlation' indicates a **negative** gradient.

Measuring Correlation

The correlation coefficient, symbol r , measures the degree of linear correlation between two variables.

This coefficient can take any value between -1 and +1.

+1	indicates	perfect positive correlation
0	indicates	no linear correlation
-1	indicates	perfect negative correlation

The formula for the **Pearson's Product Moment correlation coefficient, r** is:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}}$$

For the data sets shown before the values of r are:

Data Set	1	2	3	4
Value of r	+1	-1	0.9052	-0.1516

In general we consider positive correlation coefficient of above 0.8 to be fairly high, i.e. there is a linear relationship between the two variables.

If the correlation coefficient is between -1 and -0.8 this also indicates that there is a fairly high level of linear relationship between the two variables but in a negative fashion.

How to use the Formula:

In order to demonstrate how the formula can be used we will use data set 3 as this set shows a high level (of non-perfect) correlation.

Example: **Data Set 3**

x	y	xy	x ²	y ²
1	2	2	1	4
2	2	4	4	4
3	3	9	9	9
4	5	20	16	25
5	6	30	25	36
6	5	30	36	25
7	7	49	49	49
8	5	40	64	25
9	7	63	81	49
10	8	80	100	64
$\sum x = 55$	$\sum y = 50$	$\sum xy = 327$	$\sum x^2 = 385$	$\sum y^2 = 290$

n = 10

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}}$$

$$r = \frac{10 \times 327 - 55 \times 50}{\sqrt{(10 \times 385 - (55)^2)(10 \times 290 - (50)^2)}}$$

$$r = \frac{3270 - 2750}{\sqrt{(3850 - 3025)(2900 - 2500)}}$$

$$r = \frac{520}{\sqrt{(825 \times 400)}} = \frac{520}{\sqrt{330000}} = \frac{520}{574.46} = 0.9052$$

Thus the correlation coefficient for Data Set 3 is 0.9052

This indicates a fairly high level of positive correlation.

Using Excel Commands

The value of r can be found from Excel by the command: = CORREL(x range, y range)

Notes:

1. You must type the comma between the x range and the y range, i.e. after highlighting the x range, type a comma before highlighting the y range.
2. The command =CORREL(y range, x range) gives the same answer, i.e. the order of the data columns does not matter.

$$\text{Coefficient of Determination} = (\text{Correlation Coefficient})^2 = r^2$$

The value of r^2 is known as the coefficient of determination. It is usually expressed as a percentage. It gives the level of variation in 'y' that can be explained by the variation in 'x'.

In data set 3 $r = 0.9052$ thus $r^2 = 0.8194$

Thus 81.94% (82 %) of the variation in y can be explained by the variation in x. The other 18% is unaccounted for but would be due to other factors. These factors will depend upon the nature of the data.

In data set 4 $r = -0.1516$ thus $r^2 = 0.02298$

Thus 2.298% (2.3%) of the variation in y can be explained by the variation in x.
As the other 97.7% is unaccounted for, the actual nature of the relationship is not known.

A low value of r^2 indicates that the relationship is not linear.

Returning to the case study on London Theatres:

We can produce a table of correlations:

	<i>attendances</i>	<i>gross box office revenue</i>	<i>average number of theatres open</i>	<i>number of performances</i>	<i>number of new productions</i>
attendances	1				
gross box office revenue	0.9595	1			
average number of theatres open	0.7543	0.7521	1		
number of performances	0.8224	0.7943	0.9593	1	
number of new productions	0.6158	0.6423	0.7964	0.7492	1

From this table we can see:

Revenue and attendances	$r = 0.9595$
Revenue and theatres open	$r = 0.7521$
Revenue and performances	$r = 0.7943$
Revenue and new productions	$r = 0.6423$

If we want to be able to predict the gross box office revenue using one other variable, we should choose "attendances" as this relationship has the highest correlation coefficient.

The value of the correlation coefficient, r , between gross box office revenue" and "attendances" is 0.9595 , thus $r^2 = 0.9206$

We could then report on our findings:

There is a strong linear relationship between the number of attendances at London Theatres and the Gross box office revenue.

The value of the correlation coefficient between the number of attendances and revenue is 0.9595 indicating a strong, positive correlation. Furthermore 92% of the variation in the revenue can be explained by the variation in the number of attendances. The remaining 8% will be due to other factors; such as the prices of the theatre tickets, etc.

Warning:

A strong correlation between two variables does NOT imply that one variable causes the other to occur.

IT worksheet on correlation and regression

Load up the data file on londontheatres.xls from weblearn.

To produce a Correlation Matrix

Firstly, you need to add in the analysis toolpak

File Options Add-Ins Choose Analysis Toolpak then GO

From the menu choose Analysis Toolpak and then Ok

This will produce the analysis toolpak under the data tab

Choose the **Data** tab

Choose **Data Analysis**

Choose **Correlation**, click **OK**

Highlight the columns Attendances to new productions (including the headings but not the "years")

Tick the "labels in first row" box

Select any blank cell for the output range (the output must not touch the data table)

	A	B	C	D	E	F	G	H	I	J	K
1	year	attendances	gross box office revenue	average number of theatres open	number of performances	number of new productions					
2	1986	10,236,362	£112,067,737	42	16543	213					
3	1987	10,880,791	£129,586,817	42	16603	212					
4	1988	10,897,384	£139,338,056	43	16970	228					
5	1989	10,944,760	£153,250,633	42	16436						
6	1990	11,321,288	£177,903,841	40	15887						
7	1991	10,905,385	£186,789,856	39	15508						
8	1992	10,899,928	£194,772,331	41	15916						
9	1993	11,503,358	£215,619,208	41	15922						
10	1994	11,163,023	£217,762,521	41	16063						
11	1995	11,937,671	£238,740,636	43	17163						
12	1996	11,179,326	£229,016,639	41	16084						
13	1997	11,466,248	£246,081,850	39	15568						
14	1998	11,924,518	£257,920,278	41	16018						
15	1999	11,931,954	£266,565,215	44	17089						
16	2000	11,555,178	£286,557,500	43	16633						

Correlation

Input

Input Range:

Grouped By: ☒ Columns ☐ Rows

☒ Labels in first row

Output options: ☒ Output Range:

☐ New Worksheet Ply:
☐ New Workbook

OK

Cancel

Help

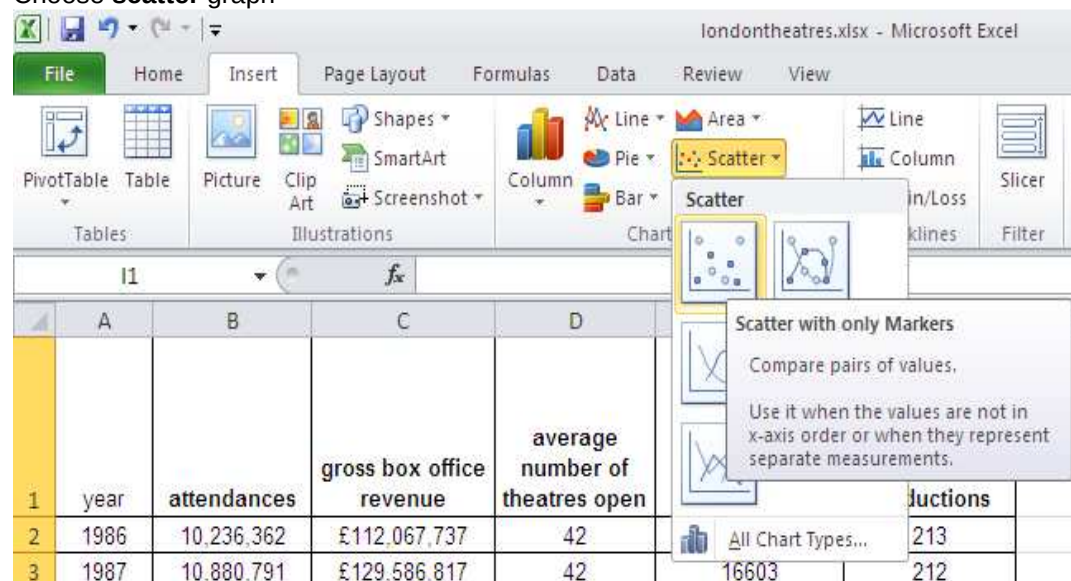
Click OK

The resulting output is :

		gross box office revenue	average number of theatres open	number of performances	number of new productions
attendances	1				
gross box office revenue	0.95952971	1			
average number of theatres open	0.75429333	0.752133	1		
number of performances	0.82244044	0.794324	0.959342	1	
number of new productions	0.61578115	0.642332	0.796411	0.749190768	1

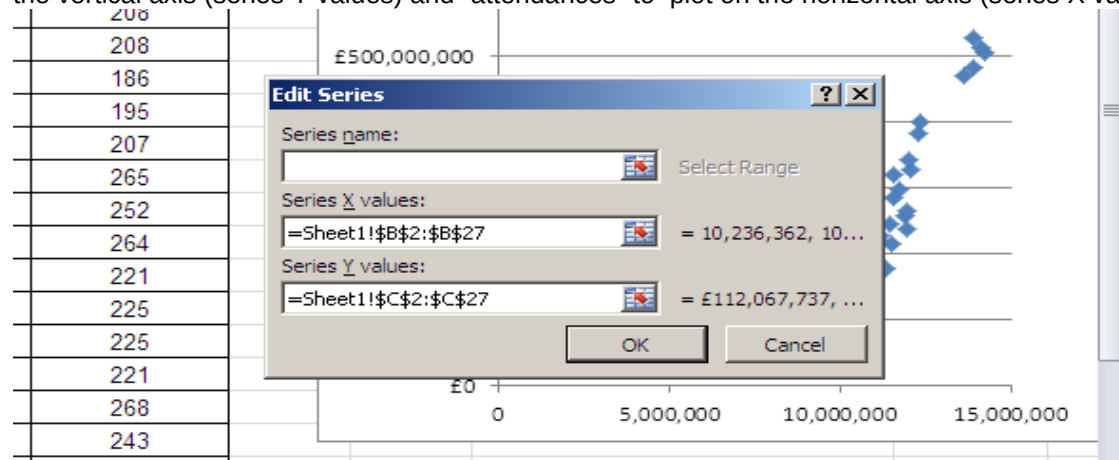
To draw a scatter graph

Choose the **Insert** tab
Choose **scatter** graph

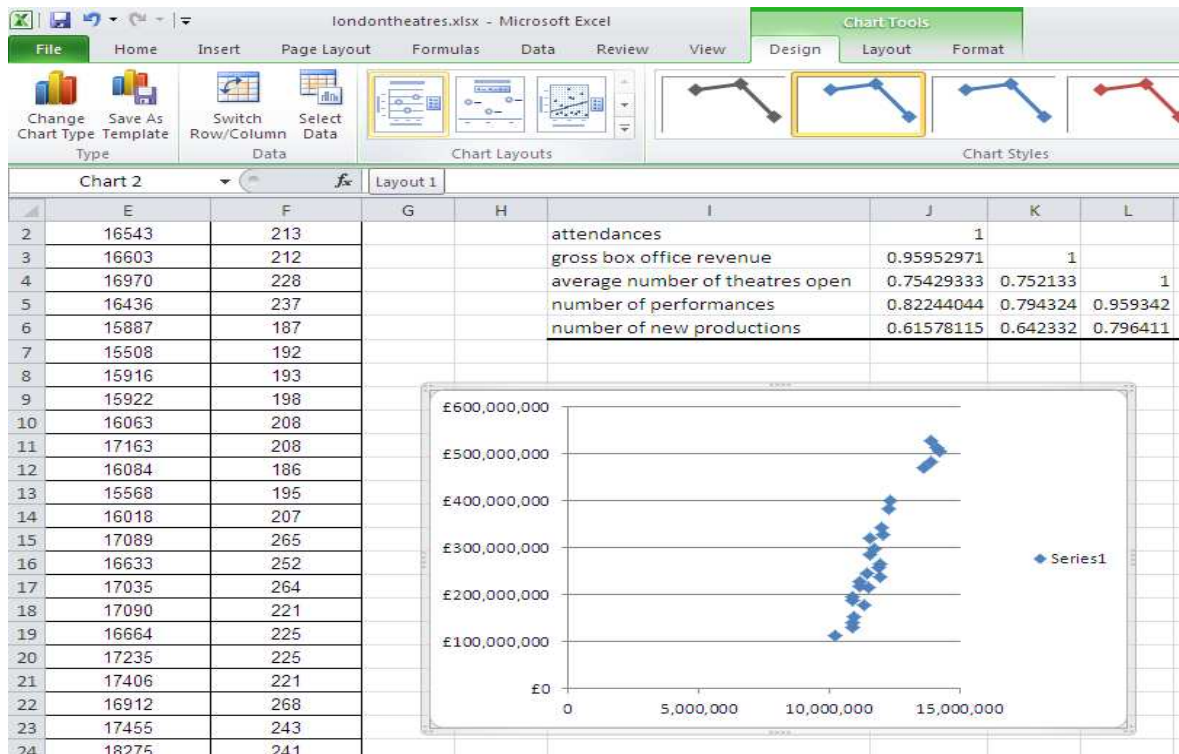


Choose the scatter graph with “only markers” (that is the one without any lines)
A blank graph will then appear.
Right click on the graph and choose “**select data**”
In the next dialogue box choose “**add**”

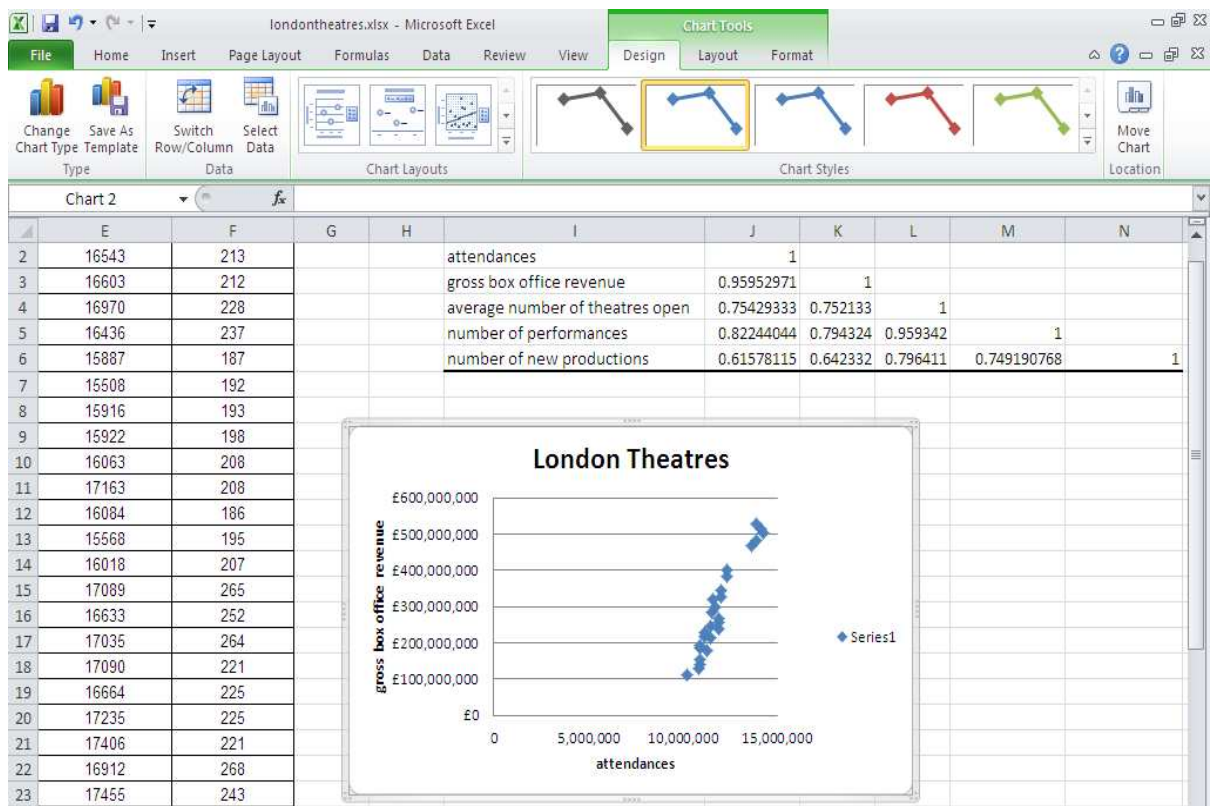
As we know that the best predictor for “revenue” is “attendances”, we will select “revenue” to plot on the vertical axis (series Y values) and “attendances” to plot on the horizontal axis (series X values).



Click OK
Click OK

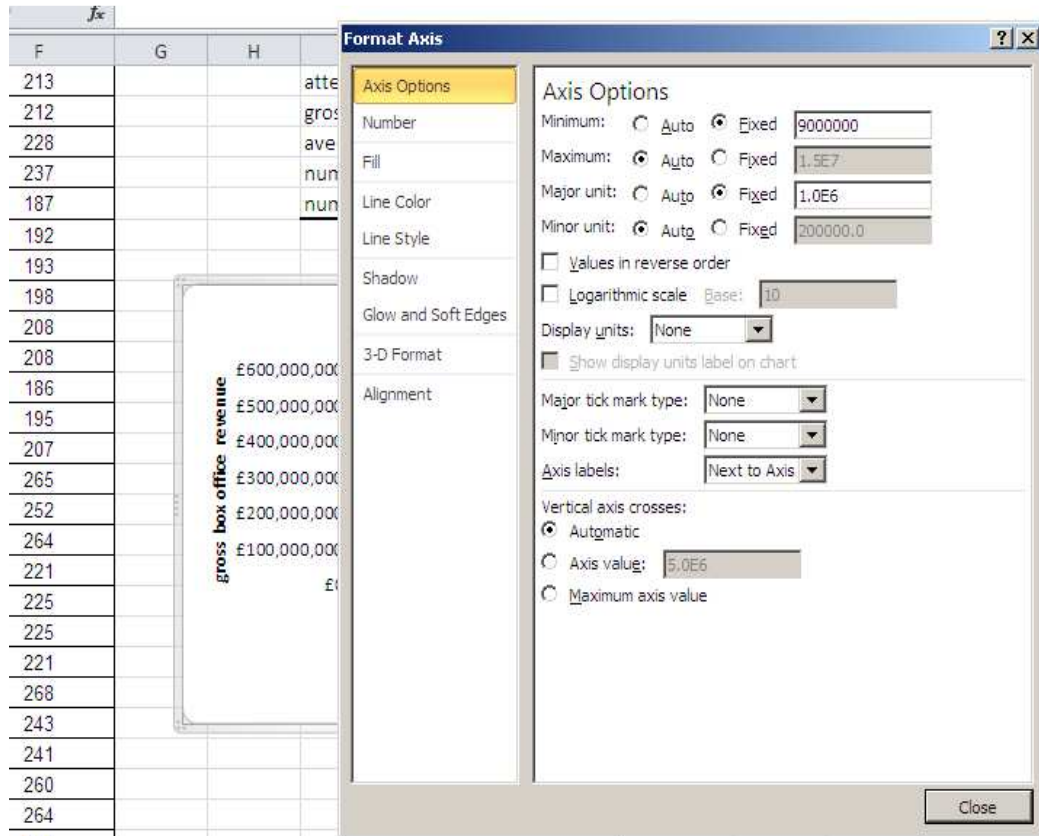


To tidy up the graph and add labels, we need to select the best design. Using the design tab , choose the layout that enables us to title the graph and label the axes

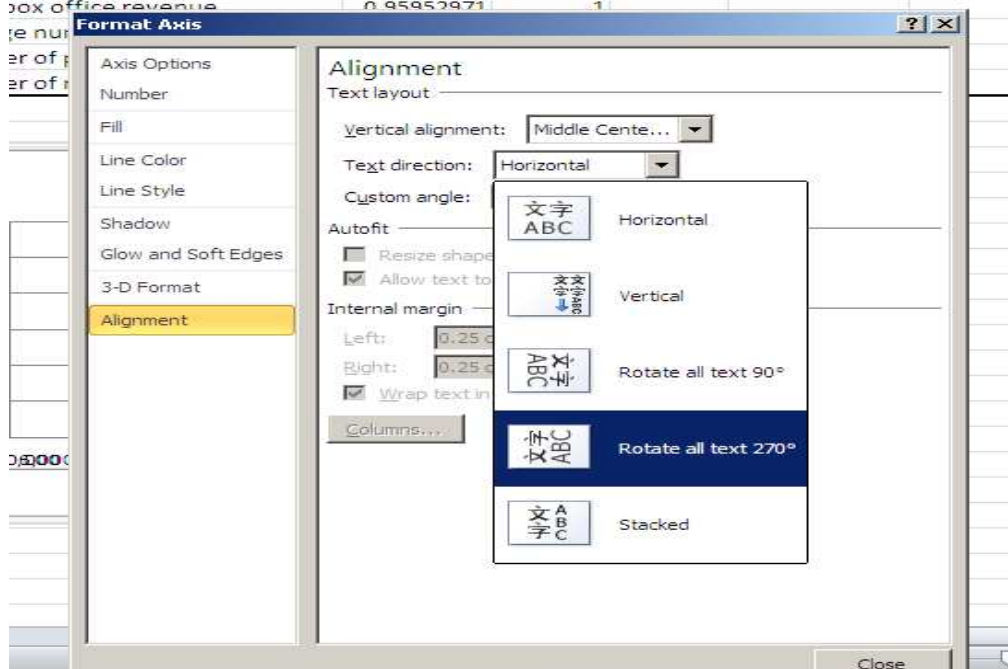


We can now add the title and label the axes by over-typing.

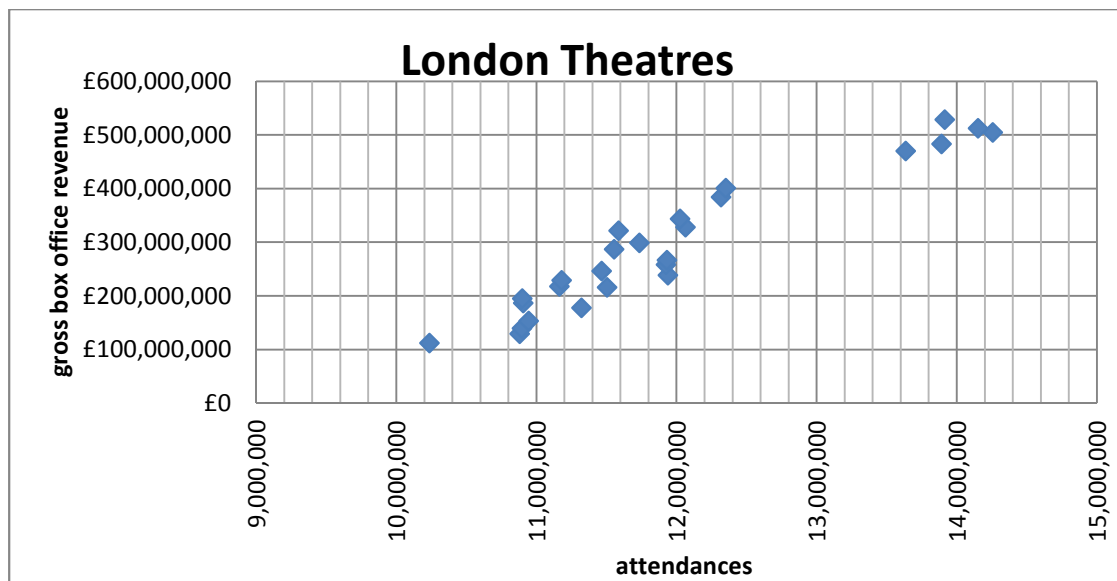
By right clicking on the horizontal axis , choosing "format axis "we can change the labelling so that it starts at 9 million and goes up in steps of 1 million



We can then use the “alignment “ on the left had side to present the numbers vertically

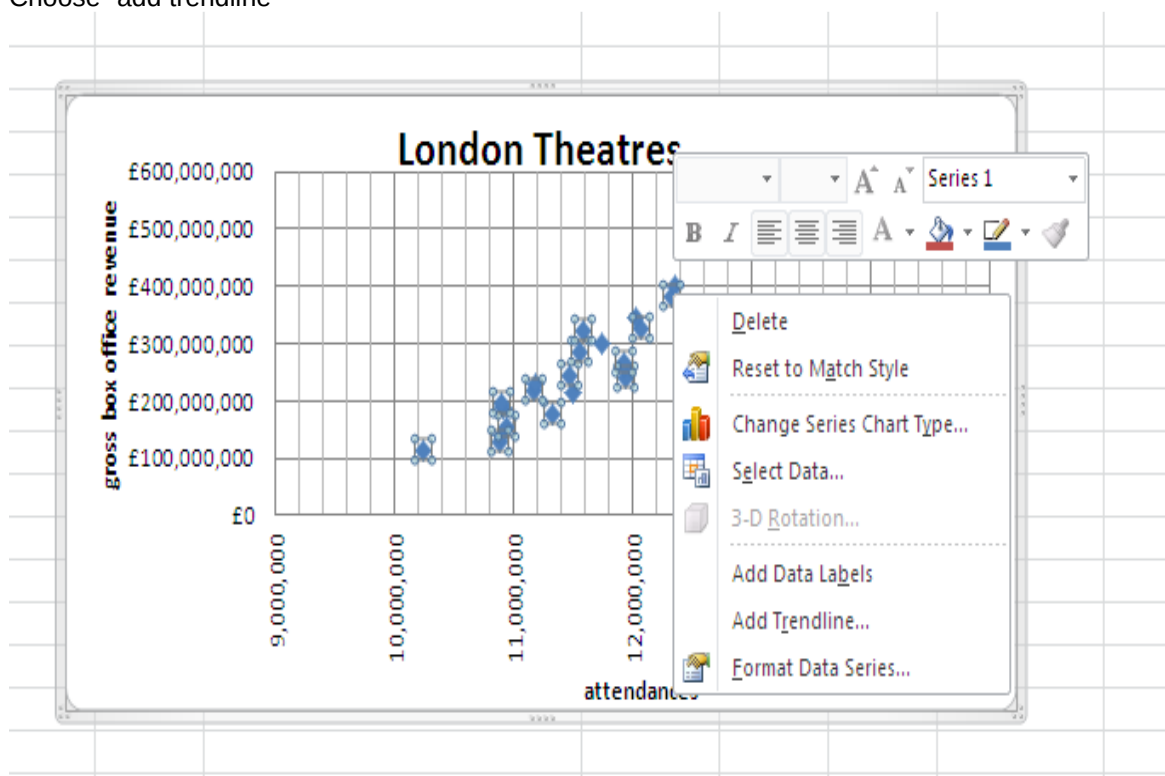


The resulting graph is :

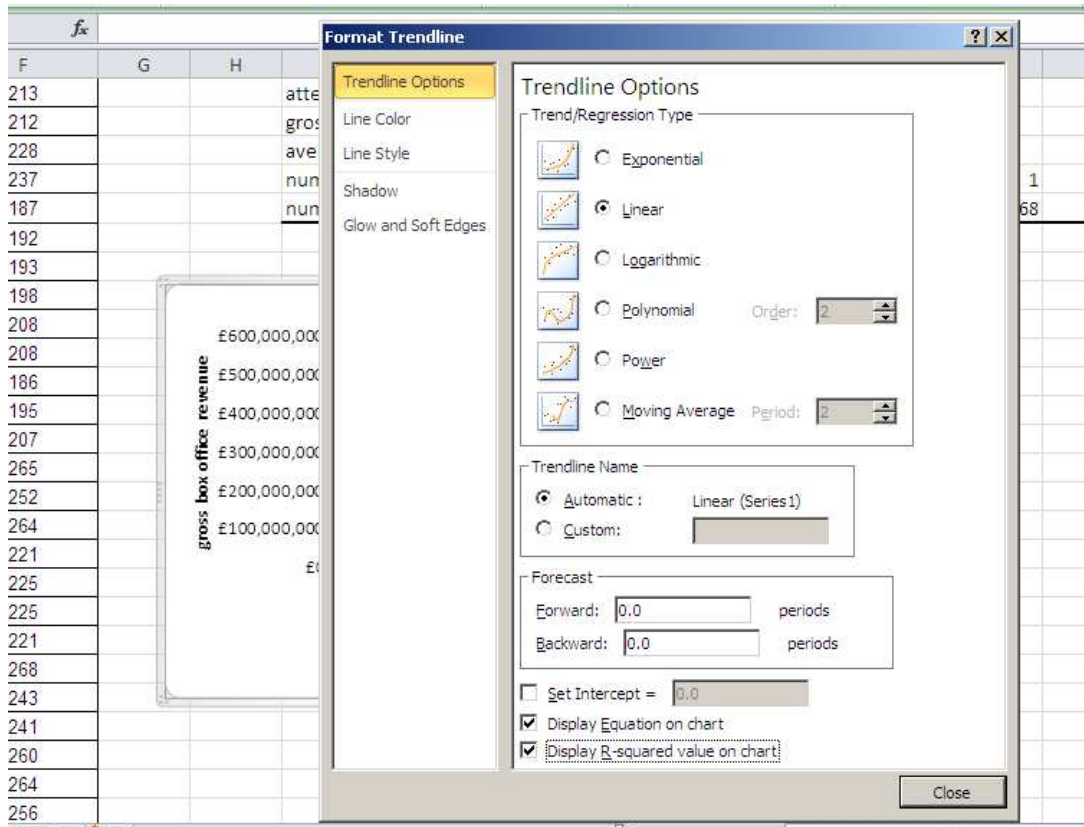


To add a regression line and equation

Right Click on any one of the plotted points
Choose “add trendline”



Choose “linear” and tick the bottom two boxes (display equation and display r squared)



Close the dialogue box to get the graph

