Level 5 – Continuous Improvement with Industry 4.0

Analyse 4 – ANOVA, Correlation and Regression

MANUFACTURE ADVANCED DESIGN ENGINEERING
Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David
Cronfa Gymdeithasol Ewrop
European Social Fund

Analyse – T Test / Analysis of Variance (ANOVA)

All the data looks normal but how do we test to see if they are different?

2 sample T test – tests the means of two populations to see if they are the same

ANOVA – analysis of variance tests the means of two or more populations to see if they are the same

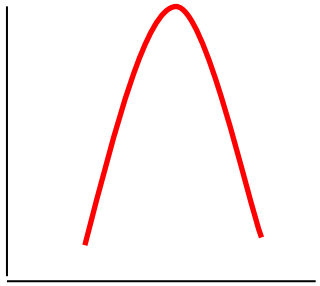
Analyse – ANOVA

Analysis of variance - ANOVA

analysis of variance tests the means of two or more populations to see if they are the same

Our Hypothesis is

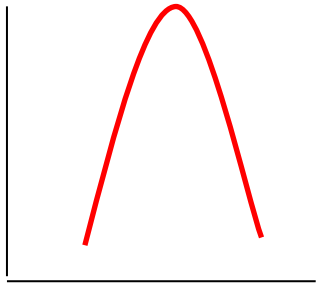
Ho : Mu1 = Mu2 = Mu3 = Mu4



Analyse – ANOVA

One way ANOVA tests involves one factor it tests the means of sets of results to see if they are the same - it does not tell us which one is different

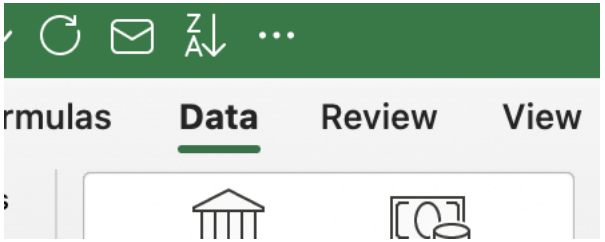
Two way ANOVA involves many factors



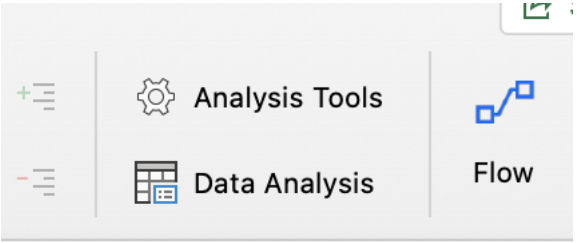
Analyse – ANOVA

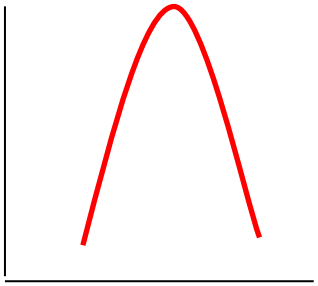
70	66	72	70
73	70	70	65
71	73	68	66
72	66	74	66
72	68	74	65
69	67	76	67
68	68	75	68
68	70	73	69
71	69	71	69
71	68	74	67
68	69	73	70
74	69	71	66
72	68	75	65
71	72	68	70
72	72	68	69
71	72	71	67
68	70	75	68
71	70	73	68
71	68	72	71
71	72	71	68

- Open new excel spreadsheet and key in the following numbers into a column and click on 'Data' tab



- Click on the 'Data Analysis' icon

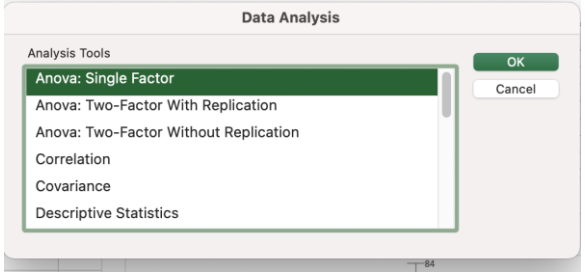




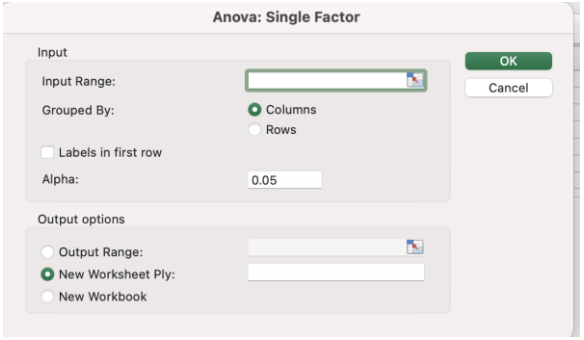
Analyse – ANOVA

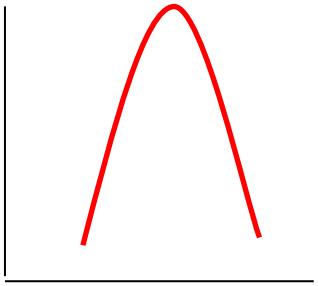
70	66	72	70
73	70	70	65
71	73	68	66
72	66	74	66
72	68	74	65
69	67	76	67
68	68	75	68
68	70	73	69
71	69	71	69
71	68	74	67
68	69	73	70
74	69	71	66
72	68	75	65
71	72	68	70
72	72	68	69
71	72	71	67
68	70	75	68
71	70	73	68
71	68	72	71
71	72	71	68

- Click on ANOVA Single Factor



- Paint in the 4 ranges of data and ensure Alpha is 0.05 (confidence of 95%) and press 'OK'

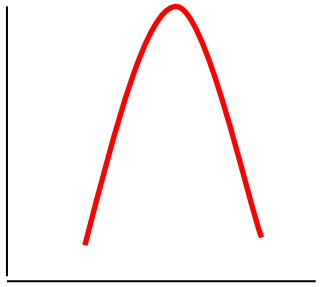




Analyse – ANOVA

70	66	72	70
73	70	70	65
71	73	68	66
72	66	74	66
72	68	74	65
69	67	76	67
68	68	75	68
68	70	73	69
71	69	71	69
71	68	74	67
68	69	73	70
74	69	71	66
72	68	75	65
71	72	68	70
72	72	68	69
71	72	71	67
68	70	75	68
71	70	73	68
71	68	72	71
71	72	71	68

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Column 1	83	5889	70.95	4.00		
Column 2	83	5816	70.07	5.46		
Column 3	83	6193	74.61	13.07		
Column 4	83	5832	70.27	10.08		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1125.602410	3.000000	375.200803	46.036825	0.000000	2.632142
Within Groups	2673.204819	328.000000	8.150015			
Total	3798.807229	331.000000				



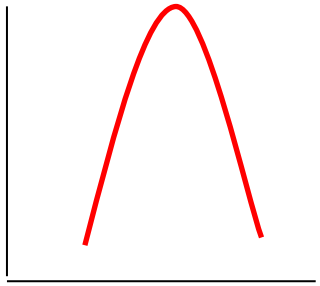
Analyse – ANOVA

Our Hypothesis is

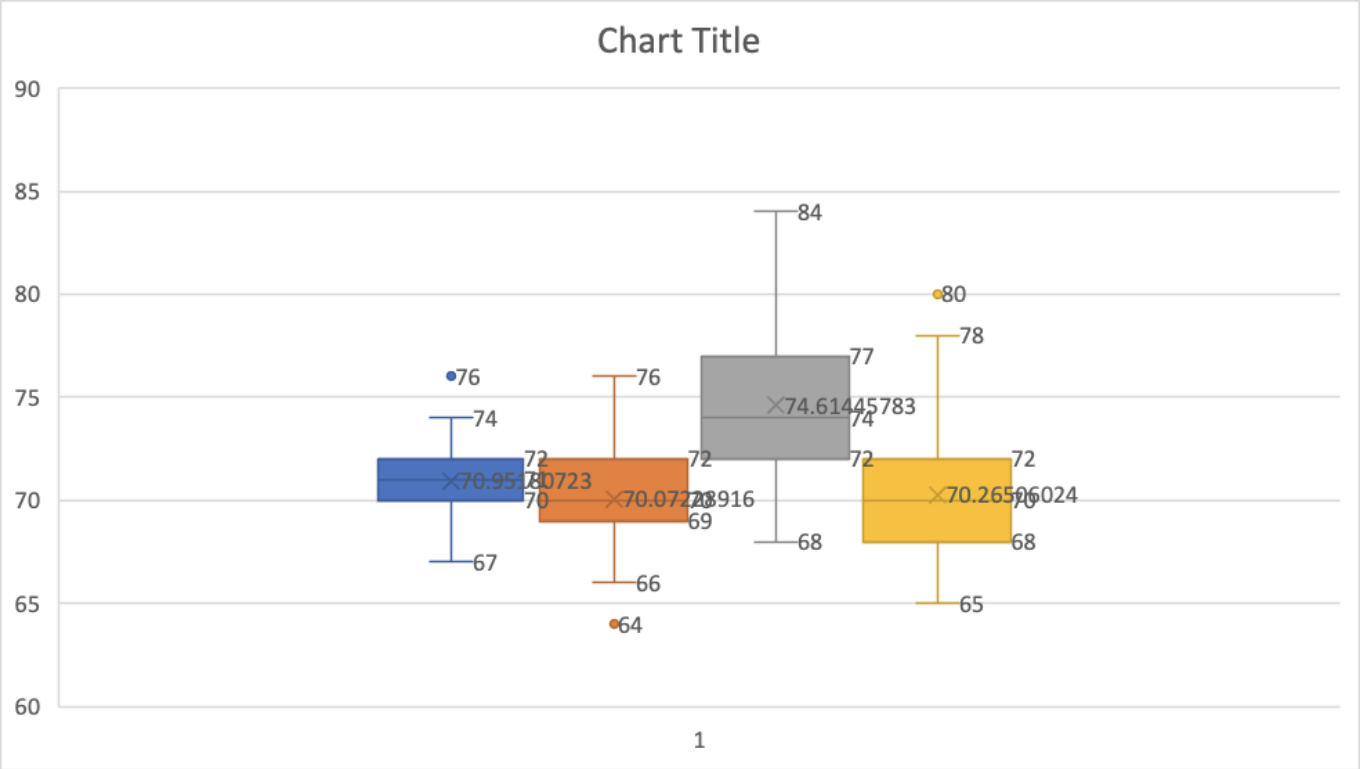
Ho : Mu1 = Mu2 = Mu3 = Mu4

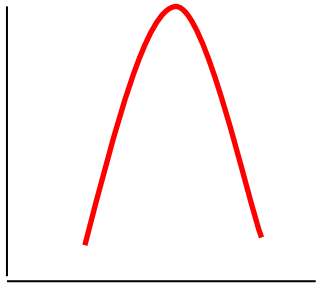
And P = 0.000 therefore we reject the null hypothesis

Ha : At Least One Inequality



Analyse – ANOVA

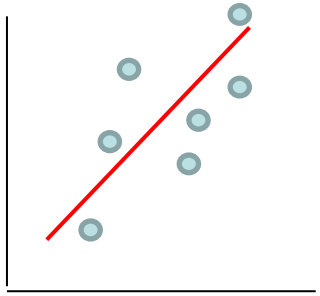




Analyse – Correlation and Regression

Correlation is used to see if there is a relationship between different sets (x, y) of data

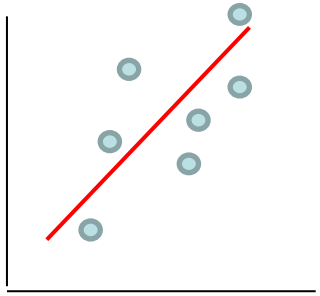
- Machining speed and surface finish
- Hours revised and exam results
- Car milage and second-hand value



Analyse – Correlation and Regression

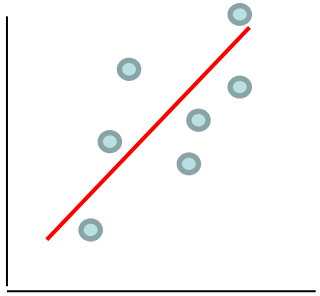
These may be classed in three ways:-

- Positive correlation where one value rises, so does the other
- Negative correlation where one value rises, the other descreases
- No correlation between either



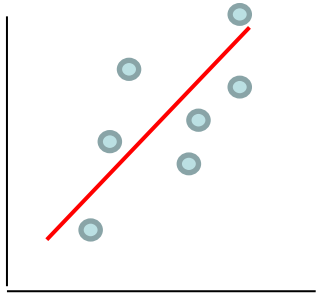
Analyse – Correlation and Regression

The relationship is expressed as ‘r’ and is known as the ‘Pearson correlation coefficient’

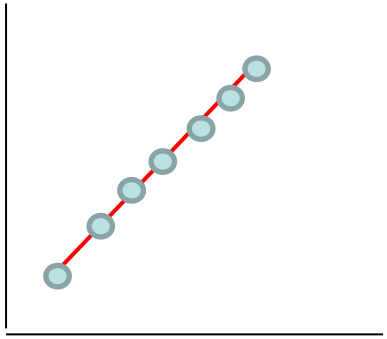


Analyse – Correlation and Regression

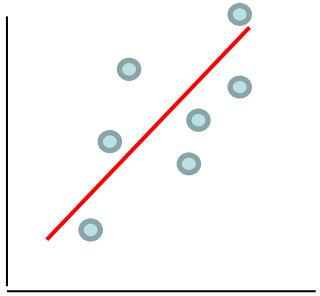
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



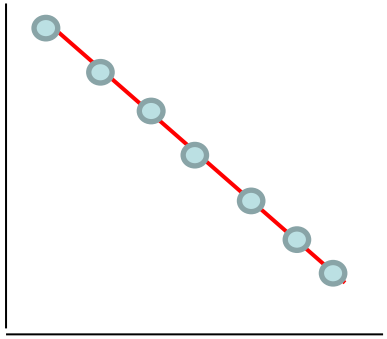
Analyse – Correlation and Regression



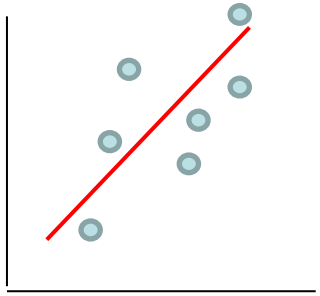
Perfect positive correlation r = 1



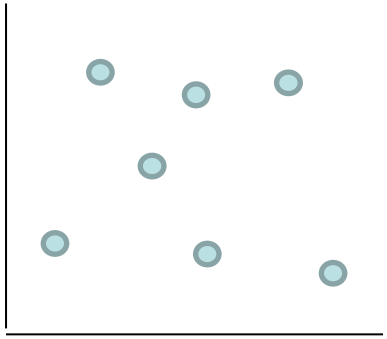
Analyse – Correlation and Regression



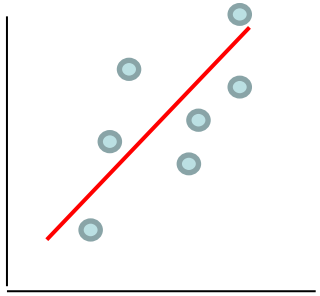
Perfect negative correlation r = -1



Analyse – Correlation and Regression



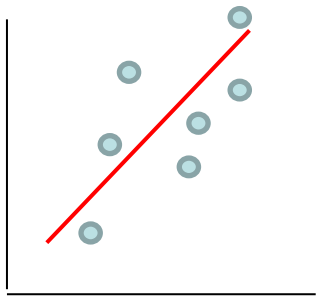
No correlation r = 0



Analyse – Correlation and Regression

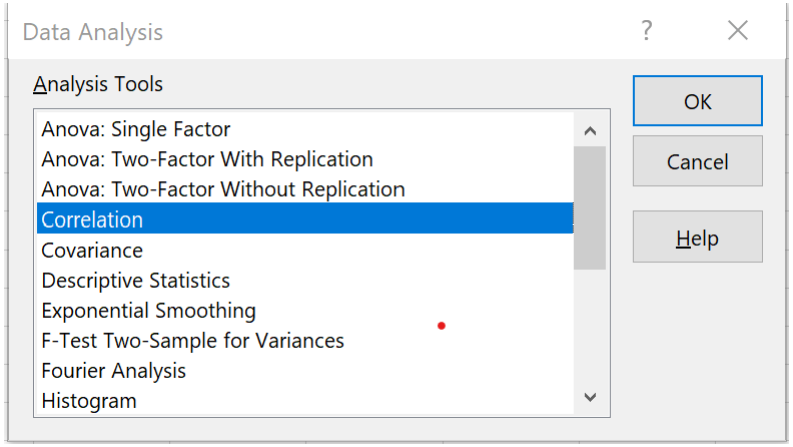
Here is some data I have collected about my car (the one I use on my journey to work!)

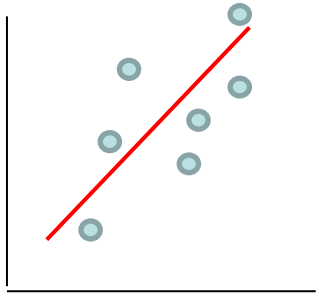
value	milage	servicing
32000	5000	300
29000	10000	400
27000	15000	450
26500	20000	610
26000	25000	400
25000	30000	700
24000	35000	700
22000	40000	610
19000	45000	750



Analyse – Correlation and Regression

value	milage	servicing
32000	5000	300
29000	10000	400
27000	15000	450
26500	20000	610
26000	25000	400
25000	30000	700
24000	35000	700
22000	40000	610
19000	45000	750

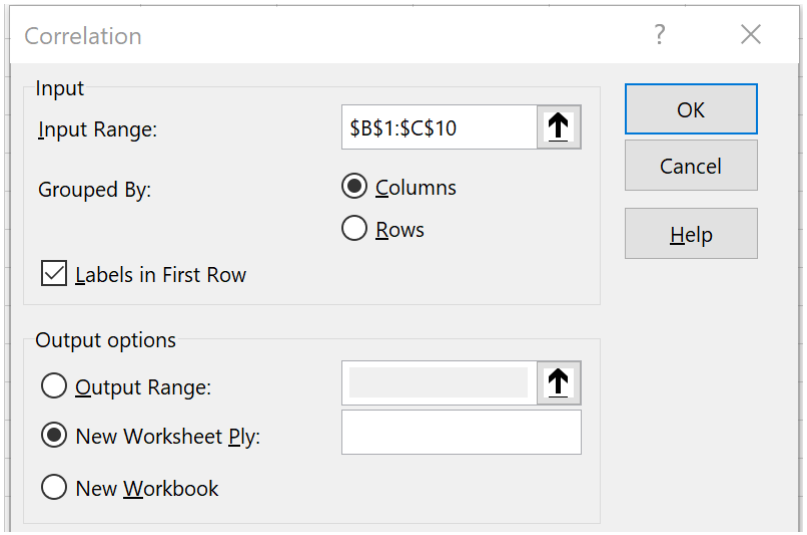


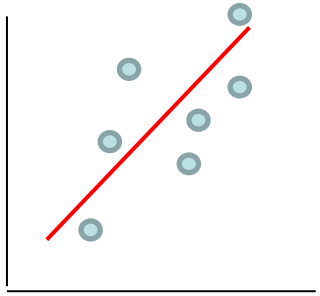


Analyse – Correlation and Regression

value	milage	servicing
32000	5000	300
29000	10000	400
27000	15000	450
26500	20000	610
26000	25000	400
25000	30000	700
24000	35000	700
22000	40000	610
19000	45000	750

Paint in first two columns



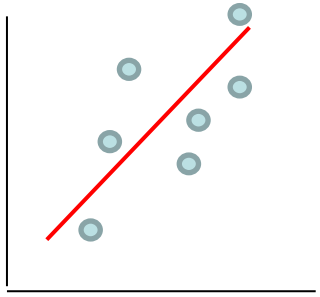


Analyse – Correlation and Regression

value	milage	servicing
32000	5000	300
29000	10000	400
27000	15000	450
26500	20000	610
26000	25000	400
25000	30000	700
24000	35000	700
22000	40000	610
19000	45000	750

	value	milage
value	1	
milage	-0.96957	1

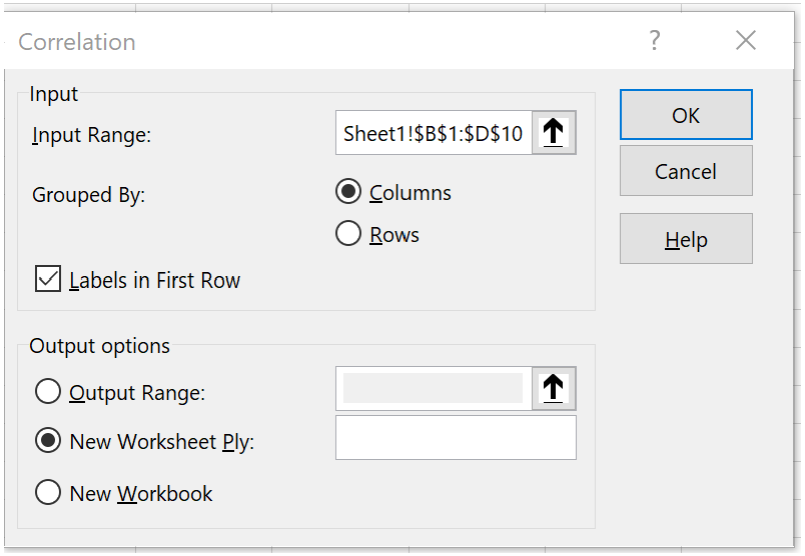
There is a strong (negative) correlation between value and milage

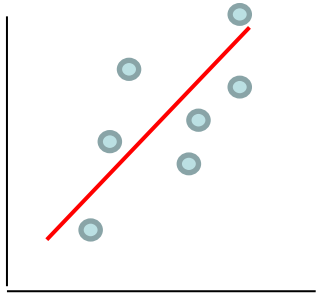


Analyse – Correlation and Regression

value	milage	servicing
32000	5000	300
29000	10000	400
27000	15000	450
26500	20000	610
26000	25000	400
25000	30000	700
24000	35000	700
22000	40000	610
19000	45000	750

Paint in all columns



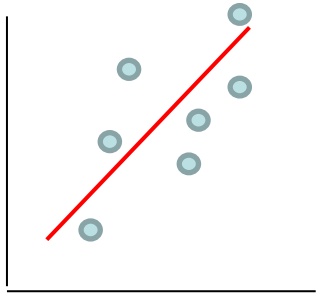


Analyse – Correlation and Regression

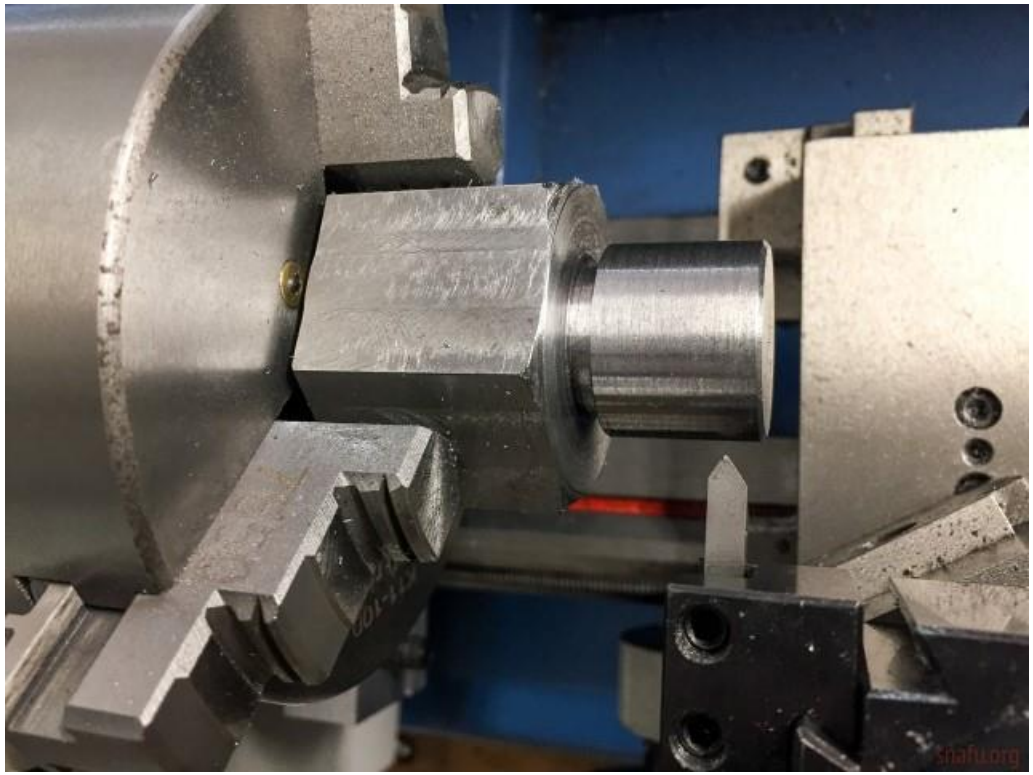
value	milage	servicing
32000	5000	300
29000	10000	400
27000	15000	450
26500	20000	610
26000	25000	400
25000	30000	700
24000	35000	700
22000	40000	610
19000	45000	750

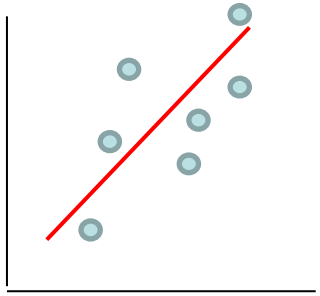
	value	milage	servicing
value	1		
milage	-0.96957	1	
servicing	-0.83992	0.8516	1

There is a strong (negative) correlation between value and servicing.
Strong (positive) correlation between milage and servicing



Analyse – Correlation and Regression

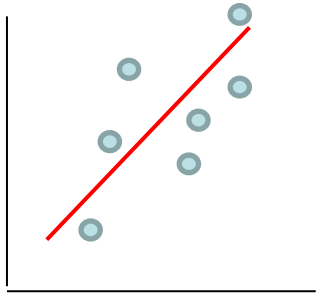




Analyse – Correlation and Regression

Machine shop collected data to see if there was any correlation between size and finish (outputs and other parameters)

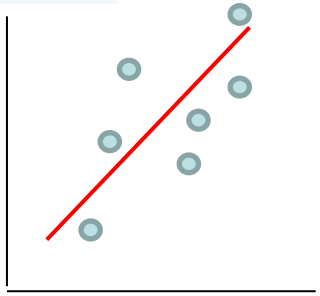
size off machine	surface finish	chuck speed	temperature	motor amps
10.94	2	448	18	18
10.82	1	447	21	18
10.78	1	446	21	17
10.62	1	441	20	17
10.58	1	440	20	17
10.58	1	435	21	17
10.57	1	434	22	17
10.56	1	432	21	16
10.54	2	428	20	16
10.52	2	428	22	16
10.52	2	426	22	16
10.48	3	423	21	16
10.47	3	421	22	16
10.45	3	417	20	15
10.42	3	416	21	15
10.41	4	415	22	15
10.40	4	415	22	15
10.39	5	408	21	15
10.39	5	405	21	15
10.38	5	401	18	15



Analyse – Correlation and Regression

Machine shop collected data to see if there was any correlation between size and finish (outputs and other parameters

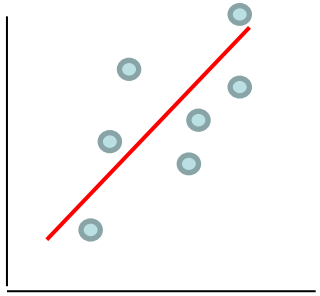
	size off machine	surface finish	chuck speed	temperature	motor amps
size off machine	1				
surface finish	-0.701728373	1			
chuck speed	0.908005172	-0.911895623	1		
temperature	-0.314750053	-0.059889904	-0.089815367	1	
motor amps	0.918230007	-0.807241187	0.937190844	-0.198216948	1



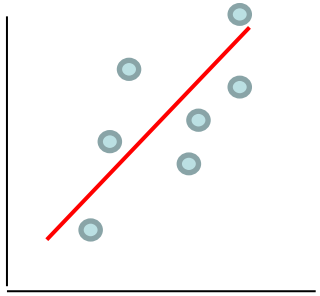
Analyse – Correlation and Regression

Linear Regression is used with correlation to see if we can predict future performance

It defines the relationship between x and y more accurately than correlation alone



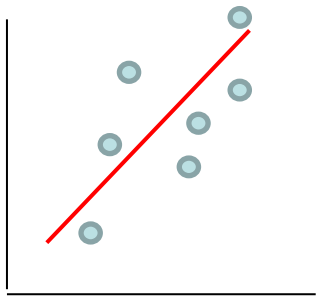
Analyse – Correlation and Regression



Analyse – Correlation and Regression

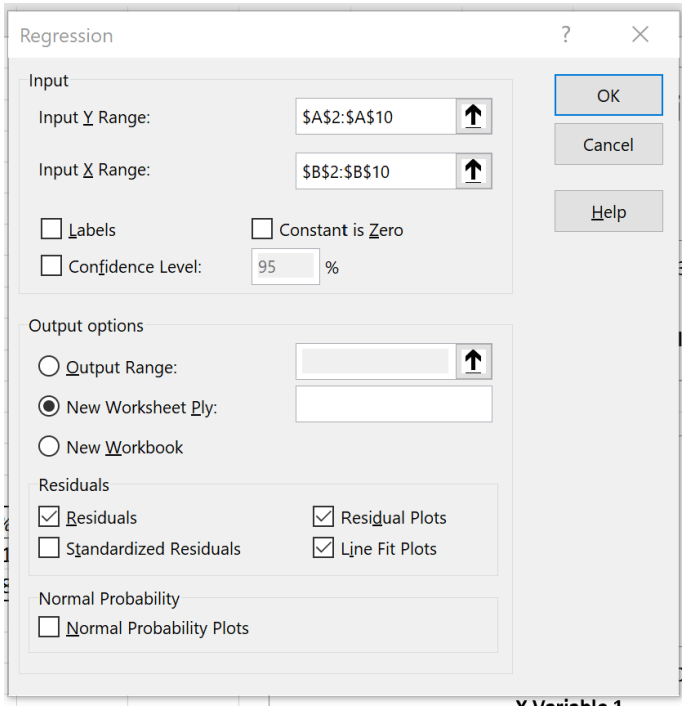
Here is some data I have collected about my car (the one I use on my journey to work!)

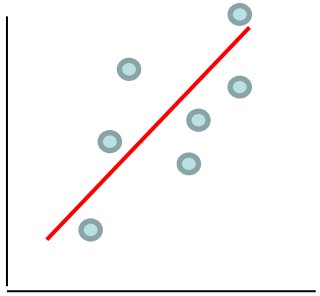
value	milage	servicing
32000	5000	300
29000	10000	400
27000	15000	450
26500	20000	610
26000	25000	400
25000	30000	700
24000	35000	700
22000	40000	610
19000	45000	750



Analyse – Correlation and Regression

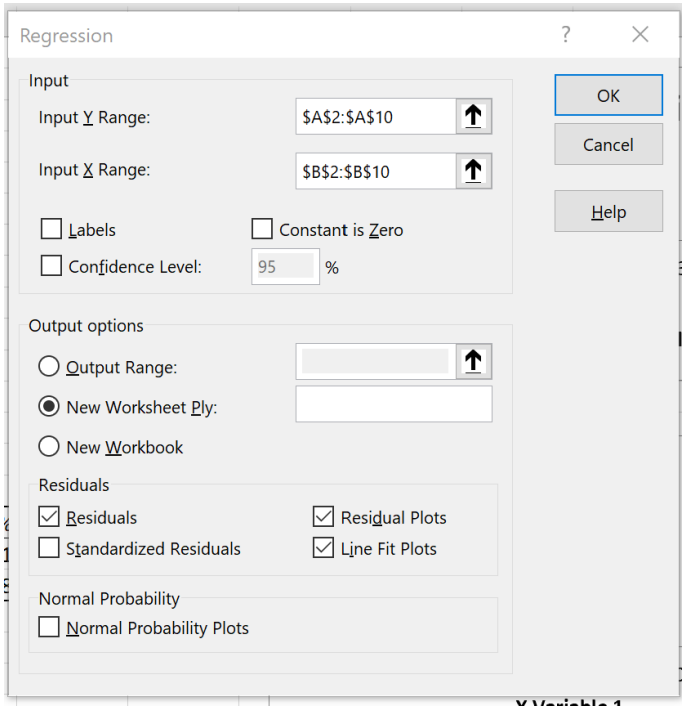
value	milage	servicing
32000	5000	300
29000	10000	400
27000	15000	450
26500	20000	610
26000	25000	400
25000	30000	700
24000	35000	700
22000	40000	610
19000	45000	750

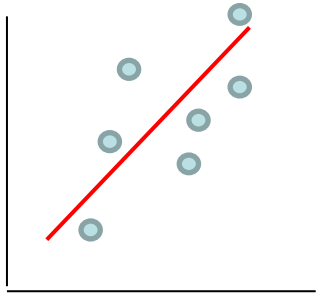




Analyse – Correlation and Regression

value	milage	servicing
32000	5000	300
29000	10000	400
27000	15000	450
26500	20000	610
26000	25000	400
25000	30000	700
24000	35000	700
22000	40000	610
19000	45000	750





Analyse – Correlation and Regression

SUMMARY OUTPUT

This is the correlation same as found earlier

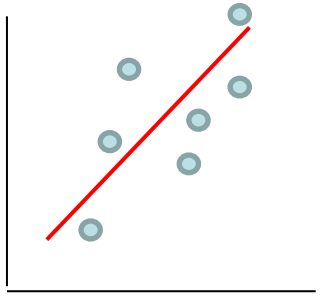
This is the correlation squared

Regression Statistics	
Multiple R	0.969575
R Square	0.940075
Adjusted R Square	0.931514
Standard Error	991.7317
Observations	9

ANOVA

	df	SS	MS	F	Significance F
Regression	1	1.08E+08	1.08E+08	109.8126	1.57E-05
Residual	7	6884722	983531.7		
Total	8	1.15E+08			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	32319.44	720.4764	44.85844	7.15E-10	30615.79	34023.1	30615.79	34023.10034
X Variable 1	-0.26833	0.02506	-10.4792	1.57E-05	-0.32888	-0.20778	-0.32888	-0.207783814



From straight line graph *y=mx+c*

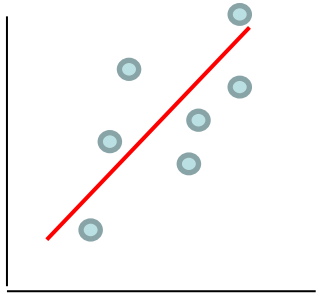
From straight line graph

Y = predicted value

M = slope

X = independent variable

C = intercept



Analyse – Correlation and Regression

Whats my car value at 100,000 miles?

Independent variable

$$y = mx+c$$

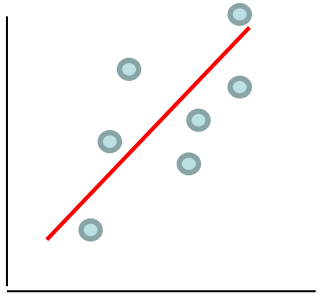
$$value = 100{,}000*-0.2683+32319$$

Intercept

$$value = £5{,}489$$

$$value = 45{,}000*-0.2683+32319$$

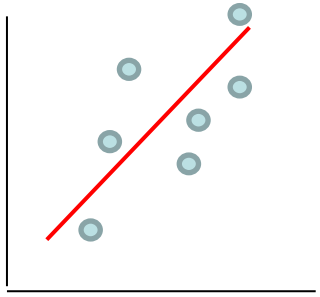
$$value = £20{,}245$$



Analyse – Correlation and Regression

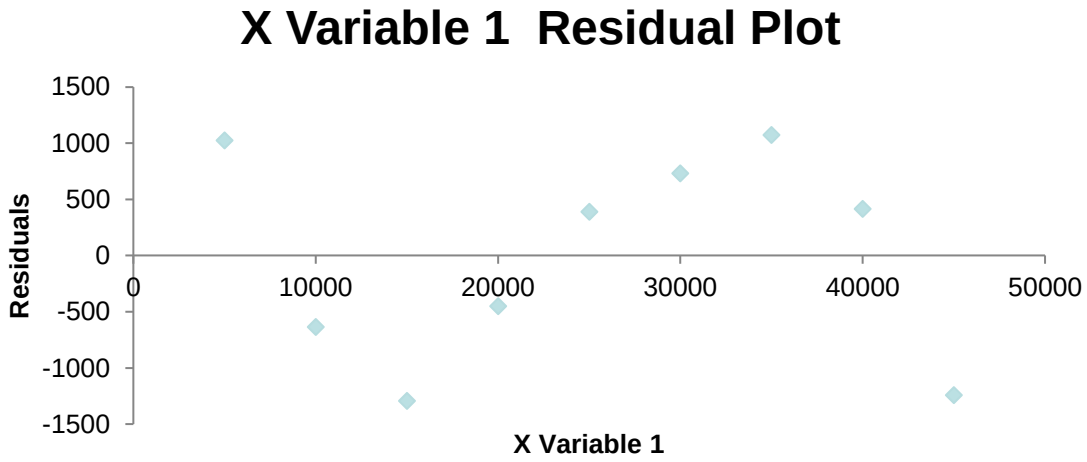
Residuals - how much do results vary from the predicted values

RESIDUAL OUTPUT			
Observation	Predicted Y	Residuals	
1	30977.78	1022.222	
2	29636.11	-636.111	
3	28294.44	-1294.44	
4	26952.78	-452.778	
5	25611.11	388.8889	
6	24269.44	730.5556	
7	22927.78	1072.222	
8	21586.11	413.8889	
9	20244.44	-1244.44	

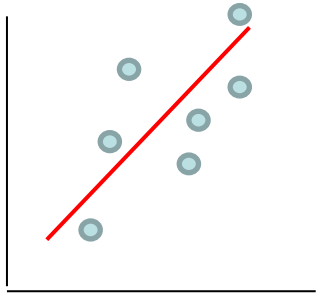


Analyse – Correlation and Regression

Residuals - how much do results vary from the predicted values



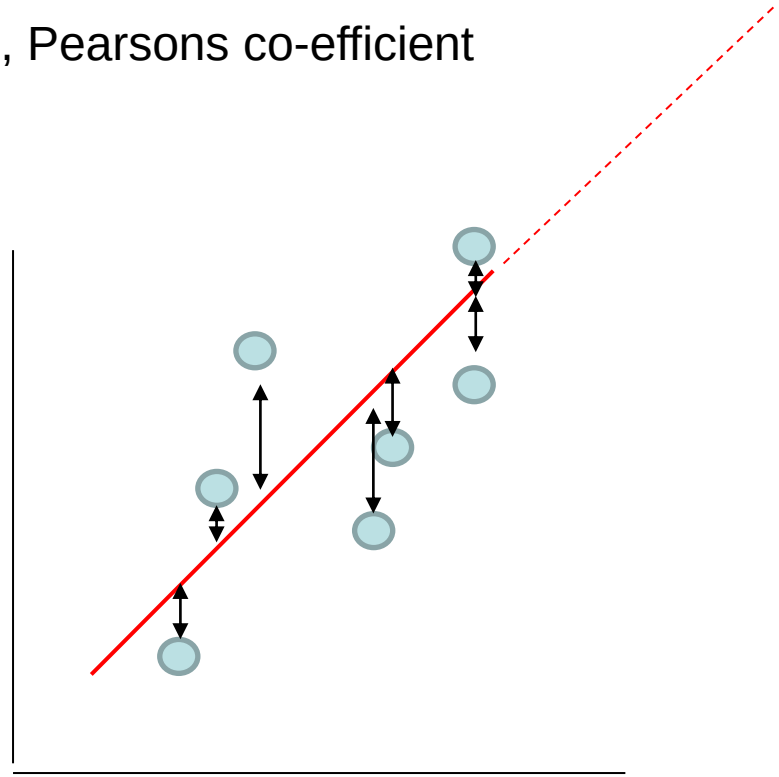
We want the residuals to be evenly spaced around zero, some positive, some negative



Scatter plots

R, Pearsons co-efficient

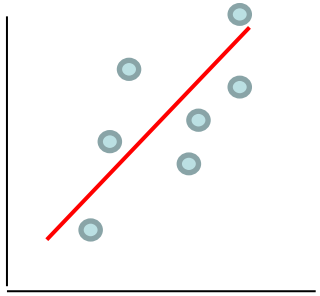
Y=mx+c



residuals

Analyse – Correlation and Regression

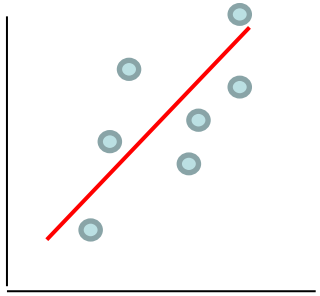
I have been collecting some data on my farm!



Analyse – Correlation and Regression

I have been collecting some data on my farm!

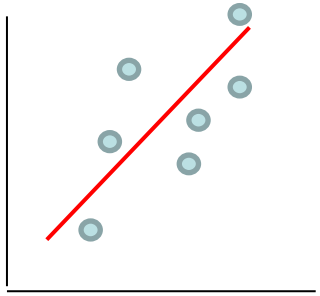
field size (Ha)	pH value	N applied Kg/Ha)	Diesel (L/Ha)	Barley Yield (t)	water content (%)	Distance from Shed (m)	cost of harvesting (Ha)
4.2	5.3	1025	13.02	14.03	14.2	285	£92.50
3.7	5.3	875	11.53	13.07	13.9	470	£83.00
5.6	5.7	1367	18.05	23.04	14.3	560	£125.00
5.2	6.1	1282	15.46	24.5	12.7	325	£115.00
3.8	5.9	1100	10.07	16.89	13.5	678	£88.00
4.7	5.8	1203	16.02	22.07	14.1	528	£103.00
6.2	5.6	1485	18.51	24.5	17.5	457	£130.00
3.4	5.5	970	9.07	13.25	16.2	365	£75.00
5.6	5.9	1501	17.53	25.97	16.8	210	£125.00
7.2	5.8	1846	26.07	31.8	16.5	25	£160.00



Analyse – Correlation and Regression

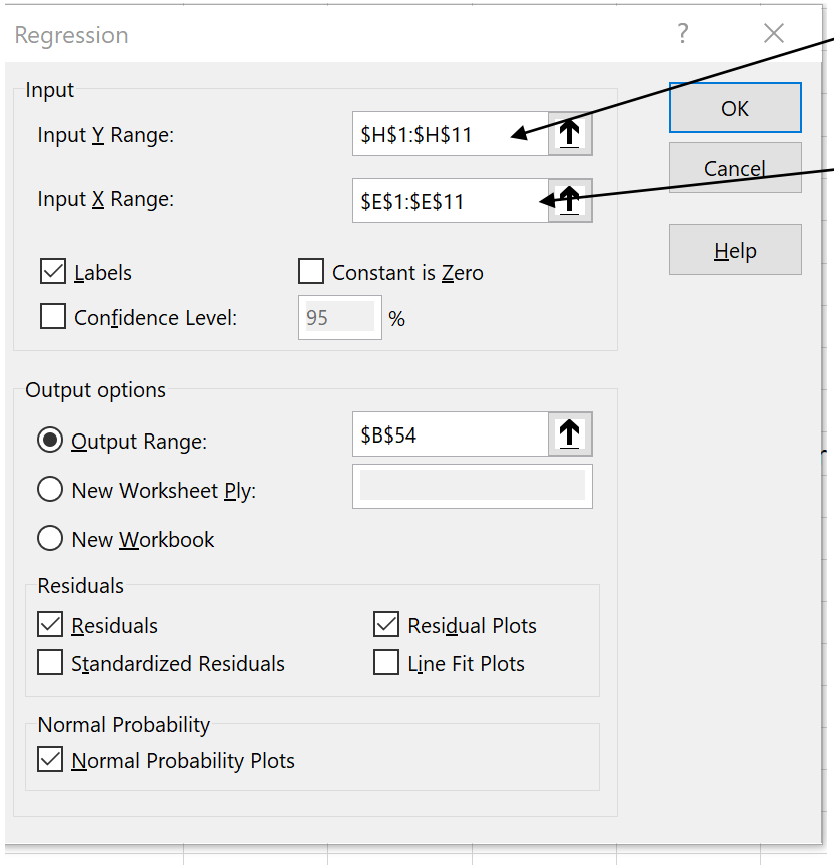
Here are the results:-

	field size (Ha)	pH value	N applied Kg/Ha)	Diesel (L/Ha)	Barley Yield (t)	water content (%)	Distance from Shed (m)	cost of harvesting (Ha)
field size (Ha)	1							
pH value	0.407844974	1						
N applied Kg/Ha)	0.968382691	0.516683173	1					
Diesel (L/Ha)	0.972421431	0.33510267	0.940619794	1				
Barley Yield (t)	0.945441392	0.659042576	0.964319103	0.920129478	1			
water content (%)	0.486194283	-0.137580716	0.534526785	0.432697789	0.373504003	1		
Distance from Shed (m)	-0.531712427	-0.041624799	-0.554110119	-0.584015357	-0.484899409	-0.454969135	1	
cost of harvesting (Ha)	0.994859797	0.444188691	0.97639778	0.976507576	0.955440095	0.44549946	-0.540498829	1



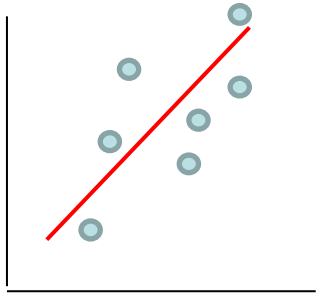
Analyse – Correlation and Regression

I now want to find relationship between cost of harvesting and barley yield



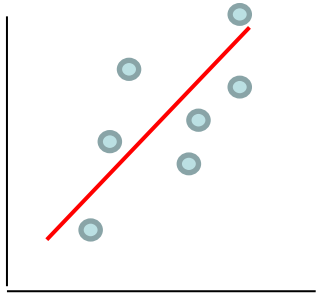
Cost of harvesting

Barley yield



Analyse – Correlation and Regression

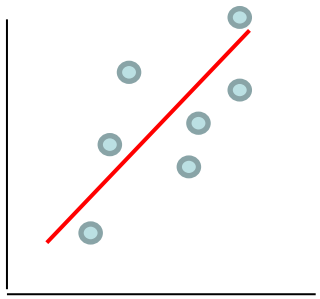
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.96							
R Square	0.91							
Adjusted R Square	0.90							
Standard Error	8.20							
Observations	10.00							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1.00	5629.67	5629.67	83.81	0.00			
Residual	8.00	537.36	67.17					
Total	9.00	6167.03						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	26.90	9.40	2.86	0.02	5.21	48.58	5.21	48.58
X Variable 1	3.96	0.43	9.15	0.00	2.96	4.95	2.96	4.95



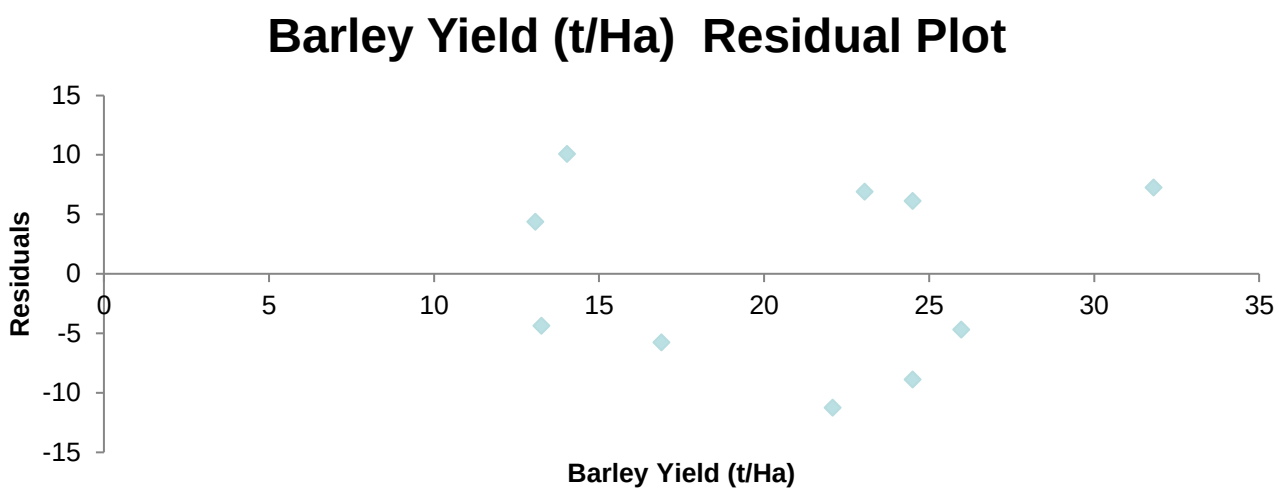
Analyse – Correlation and Regression

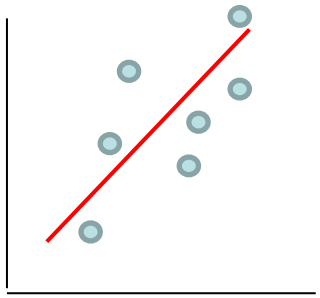
t	x variable	intercept	predicted	actual	residual
14.03	3.96	26.90	£82.42	£92.50	10.084

RESIDUAL OUTPUT			
Observation	Predicted Y	Residuals	Standard Residuals
1	82.41594183	10.08405817	1.305041598
2	78.61694506	4.383054945	0.567238798
3	118.0711095	6.928890471	0.896711438
4	123.8487505	-8.848750464	-1.145172634
5	93.73378641	-5.733786408	-0.74204548
6	114.2325399	-11.23253987	-1.453673862
7	123.8487505	6.151249536	0.796072017
8	79.32925695	-4.329256951	-0.56027646
9	129.6659643	-4.665964283	-0.60385188
10	152.7369551	7.263044856	0.939956465

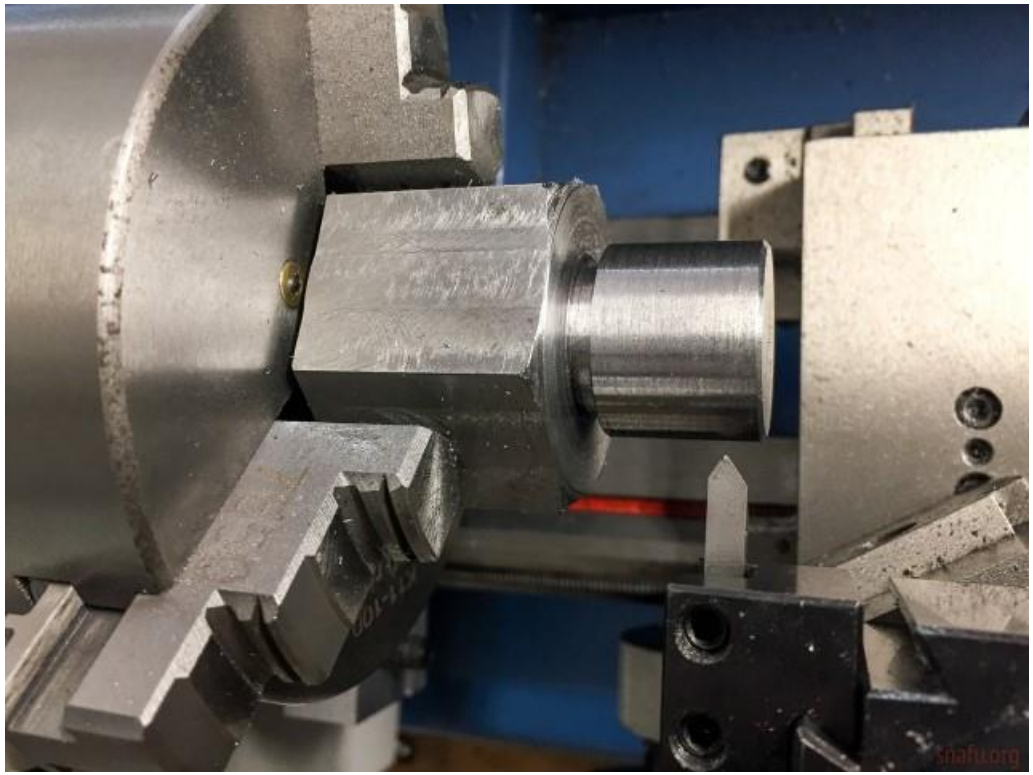


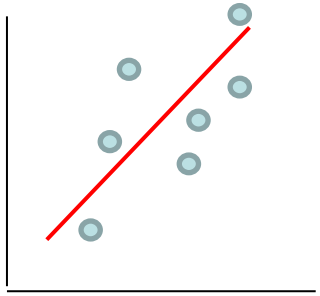
Analyse – Correlation and Regression





Analyse – Correlation and Regression

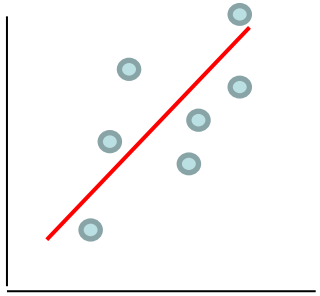




Analyse – Correlation and Regression

Machine shop collected data to see if there was any correlation between size and chuck speed (outputs and other parameters)

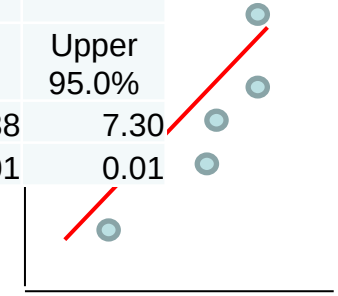
size off machine	surface finish	chuck speed	temperature	motor amps
10.94	2	448	18	18
10.82	1	447	21	18
10.78	1	446	21	17
10.62	1	441	20	17
10.58	1	440	20	17
10.58	1	435	21	17
10.57	1	434	22	17
10.56	1	432	21	16
10.54	2	428	20	16
10.52	2	428	22	16
10.52	2	426	22	16
10.48	3	423	21	16
10.47	3	421	22	16
10.45	3	417	20	15
10.42	3	416	21	15
10.41	4	415	22	15
10.40	4	415	22	15
10.39	5	408	21	15
10.39	5	405	21	15
10.38	5	401	18	15



Analyse – Correlation and Regression

See if we can predict size finish by chuck speed

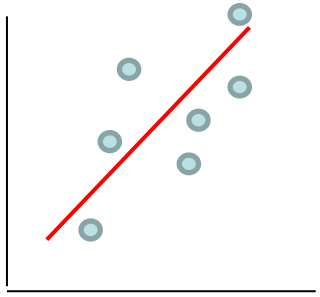
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.91							
R Square	0.82							
Adjusted R Square	0.81							
Standard Error	0.07							
Observations	20.00							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1.00	0.37	0.37	84.55	0.00			
Residual	18.00	0.08	0.00					
Total	19.00	0.45						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	6.34	0.46	13.86	0.00	5.38	7.30	5.38	7.30
chuck speed	0.01	0.00	9.20	0.00	0.01	0.01	0.01	0.01



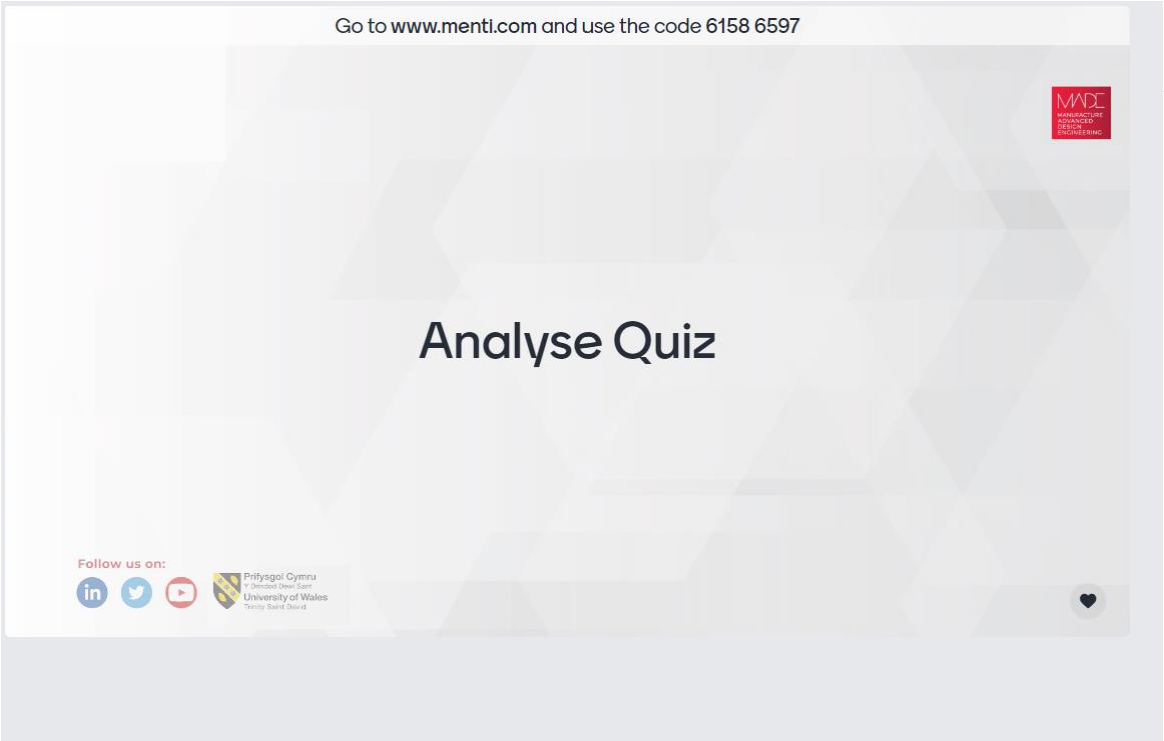
Analyse – Correlation and Regression

chuck speed	x variable	intercept		predicted		actual		residual
401	0.01	6.34		10.29		10.38		0.09

RESIDUAL OUTPUT			
Observation	Predicted size off machine	Residuals	Standard Residuals
1.00	10.75	0.19	2.88
2.00	10.75	0.07	1.17
3.00	10.74	0.04	0.70
4.00	10.69	-0.07	-1.03
5.00	10.68	-0.10	-1.50
6.00	10.63	-0.05	-0.73
7.00	10.62	-0.05	-0.73
8.00	10.60	-0.04	-0.58
9.00	10.56	-0.02	-0.28
10.00	10.56	-0.04	-0.59
11.00	10.54	-0.02	-0.28
12.00	10.51	-0.03	-0.44
13.00	10.49	-0.02	-0.29
14.00	10.45	0.00	0.01
15.00	10.44	-0.02	-0.30
16.00	10.43	-0.02	-0.31
17.00	10.43	-0.03	-0.46
18.00	10.36	0.03	0.46
19.00	10.33	0.06	0.92
20.00	10.29	0.09	1.38



Analyse – the tools

- No data or experience – all we can offer is an opinion
- No data – we use our experience to get through
- Some data - data collected we just look at the numbers
- Grouped data - basic charts and diagrams
- Descriptive statistics - mean, median, mode, Std Dev etc
- Characterising statistics - Cp, Cpk,
- Control statistics - SPC charts (Mr Shewhart!)
- Inferential statistics - normality, correlation, regression, etc

Project Work