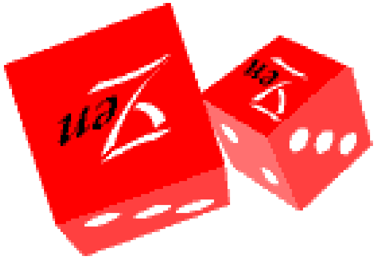
CFG IT, Data and Cyber Security Conference
9 March 2017

“Zest for Enlightenment”

Powerful analytics using Excel and/or statistical software you can use for free

Z/Yen Group Limited
41 Lothbury
London EC2R 7HG
United Kingdom
tel: +44 (20) 7562-9562
www.zyen.com



Z/Yen Overview

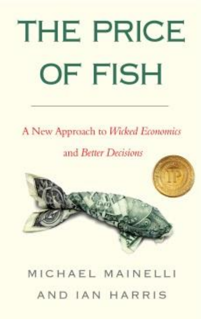
London Accord

LONG FINANCE

The Global Financial Centres Index

WINNER BCS IT PROFESSIONAL AWARDS 2004 ComputerWeekly

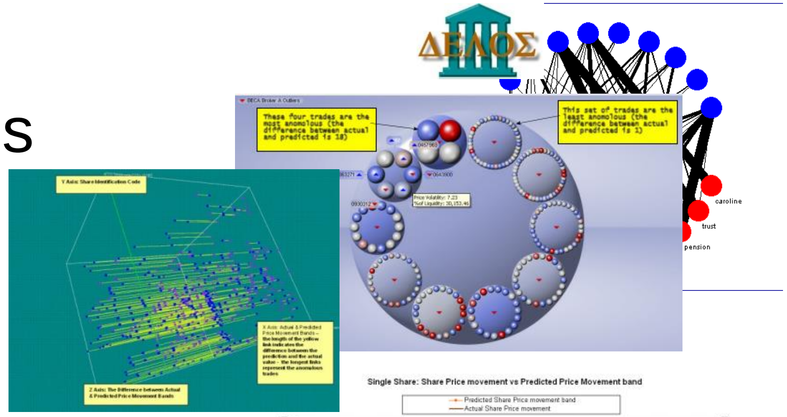
- ♦ Special – City of London's leading commercial think-tank
- ♦ Services – projects, coaching/training, expertise on demand, research
- ♦ Sectors – civil society, technology, finance, professional services

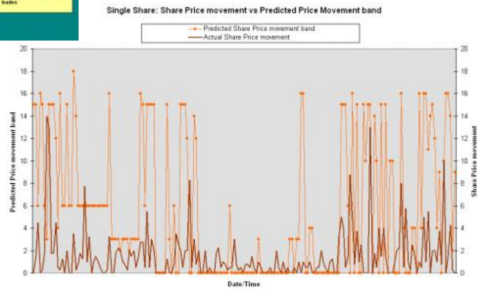

THE PRICE OF FISH
A New Approach to Wicked Economics and Better Decisions
MICHAEL MAINELLI AND IAN HARRIS

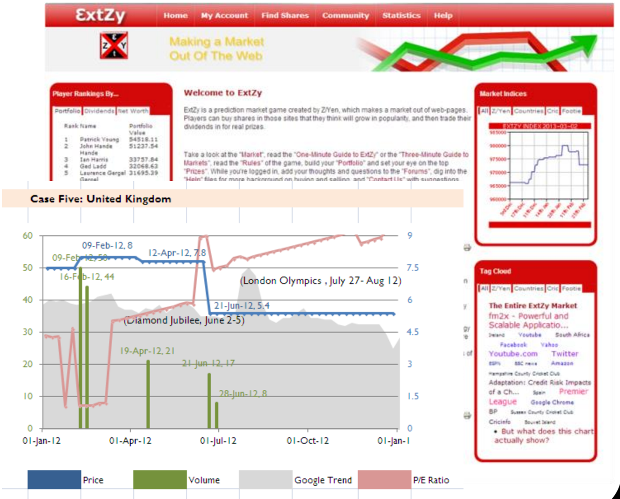
- ➢ **Analytics in Action Award**, UNISON highly commended, 2014
- ➢ Independent Publisher Book Awards Finance, Investment & Economics Gold Prize 2012 for ***The Price of Fish* –** now in paperback
- ➢ British Computer Society **IT Director of the Year** 2004 for PropheZy and VizZy, DTI **Smart Award** 2003 for PropheZy
- ➢ **IT For The Not-For-Profit Sector 2001**
- ➢ *Sunday Times* Book of the Week**, *Clean Business Cuisine, 2000***
- ➢ £1.9M **Foresight Challenge Award** for Financial £aboratory visualising financial risk 1997

Z/Yen – Research & Intelligence

- ♦ Information systems & knowledge management strategies in charities and health (1994-present)
- ♦ PropheZy and VizZy – finance compliance monitoring, charities and health outcomes improvement, (2002-present)
- ♦ Distributed ledgers (1998-present)
- ♦ Prediction markets and bubbles (1998-present) – www.extzy.com
- ♦ Market Intelligence – Charity IT Leaders, GFCI, GIPI & others (1999-present)
- ♦ Avatars For Big Data (2010-2012)


ΔΕΛΟΣ
Y Axis: Share Identification Code
X Axis: Actual & Predicted Price Movement Bands
Z Axis: The Difference between Actual & Predicted Price Movement Bands
These four trades are the most anomolous (the difference between actual and predicted is 18)
This set of trades are the least anomolous (the difference between actual and predicted is 1)
caroline
trust
pension





Rank	Name	Portfolio Value
1	Patrick Young	54518.11
2	John Hande Hande	51237.54
3	Ian Harris	33757.84
4	Ged Ladd	32068.63
5	Laurence Gergel Gergel	31695.39

Welcome to ExtZy

ExtZy is a prediction market game created by Z/Yen, which makes a market out of web-pages. Players can buy shares in those sites that they think will grow in popularity, and then trade their dividends in for real prizes.

Take a look at the "Market", read the "One-Minute Guide to ExtZy" or the "Three-Minute Guide to Markets", read the "Rules" of the game, build your "Portfolio" and set your eye on the top "Prizes". While you're logged in, add your thoughts and questions to the "Forums", dig into the "Help" files for more background on buying and selling, and "Contact Us" with suggestions.

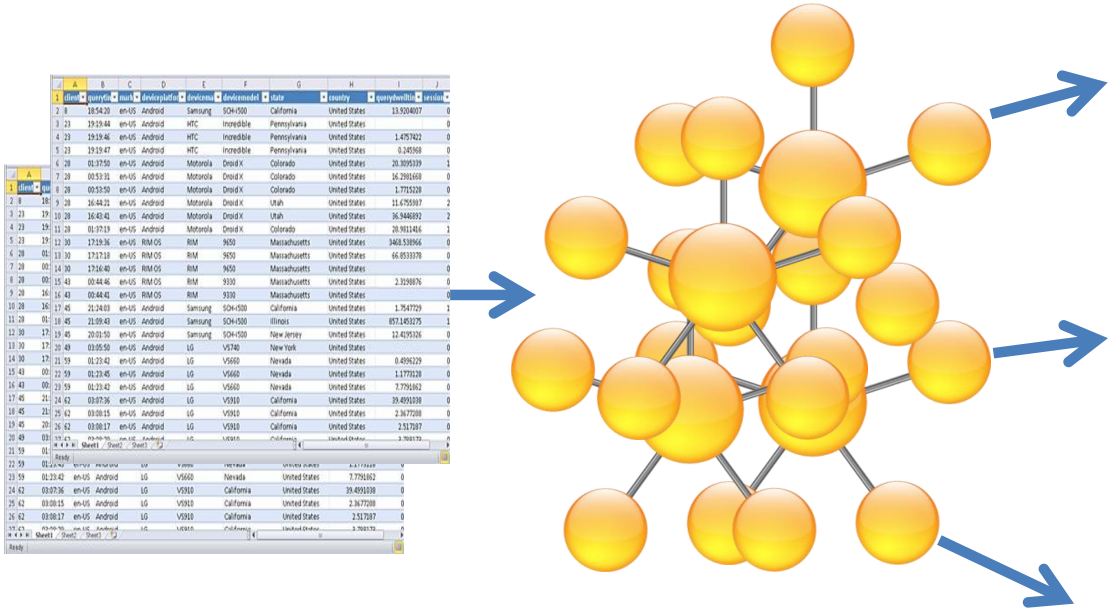
Debunking Myths About Analytics

- ♦ You do not necessarily need big data to deploy powerful machine analytics
- ♦ Does not require expensive software
  - ➢ Open Source software – R – among the best, mathematically, and free **– really, really free**
  - ➢ Excel has many of the statistical functions that used to require specialist software
- ♦ Far more about asking good questions and critical thinking, than maths or stats

© Z/Yen Group
2017

Possible Uses In Charities

Predictive Analytics



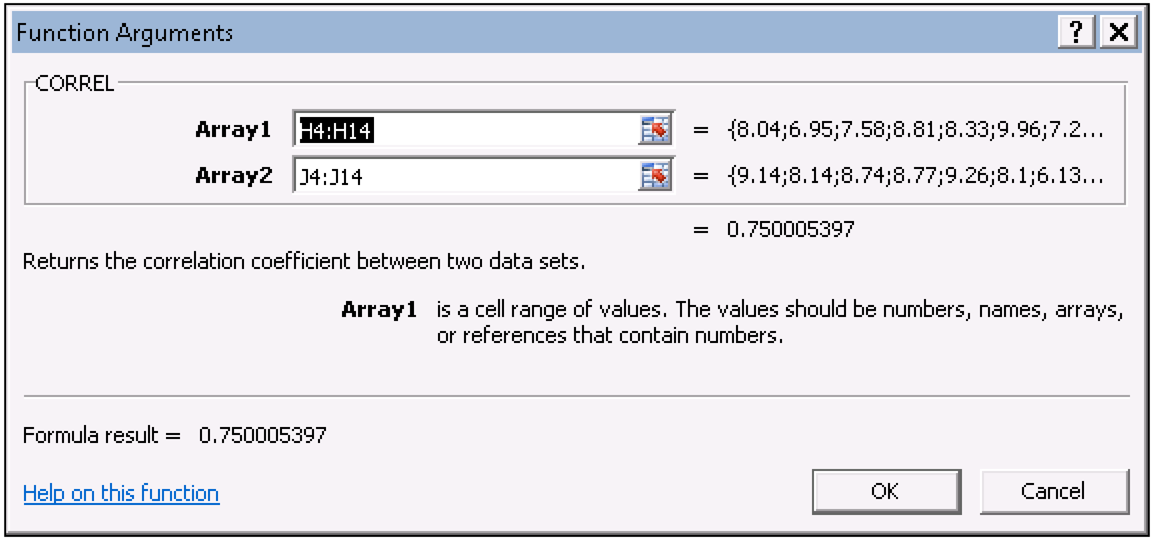
Data

Predictive
Models

Charities Activities	Core Services Delivery
	Service Development
	Marketing
	Fundraising
	Grant making/seeking
	Anomaly Detection
	Risk Management

Excel Functions

- **CORREL** function in Excel - calculates the correlation coefficient between two columns of data.
  - coefficient lies between -1 and 1.



CORREL

Array1 H4:H14 = {8.04;6.95;7.58;8.81;8.33;9.96;7.2...

Array2 J4:J14 = {9.14;8.14;8.74;8.77;9.26;8.1;6.13...

= 0.750005397

Returns the correlation coefficient between two data sets.

Array1 is a cell range of values. The values should be numbers, names, arrays, or references that contain numbers.

Formula result = 0.750005397

Causation?

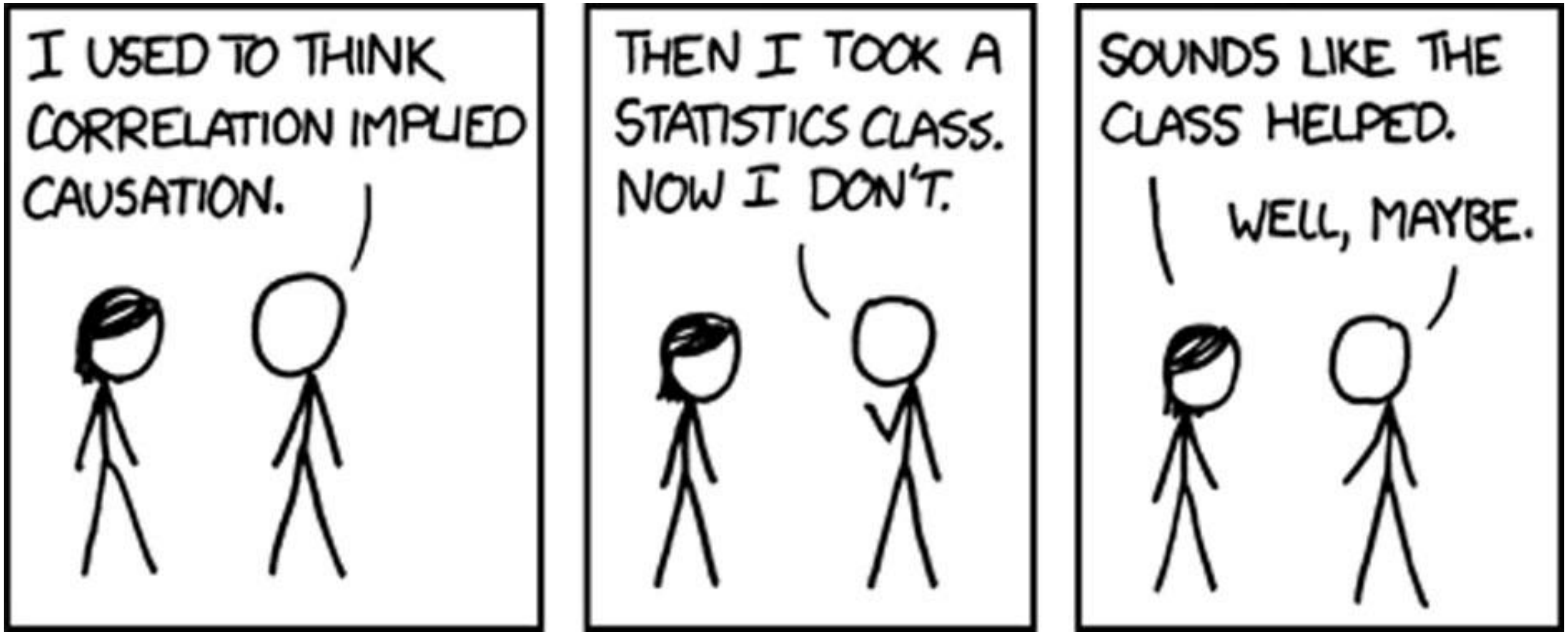

I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

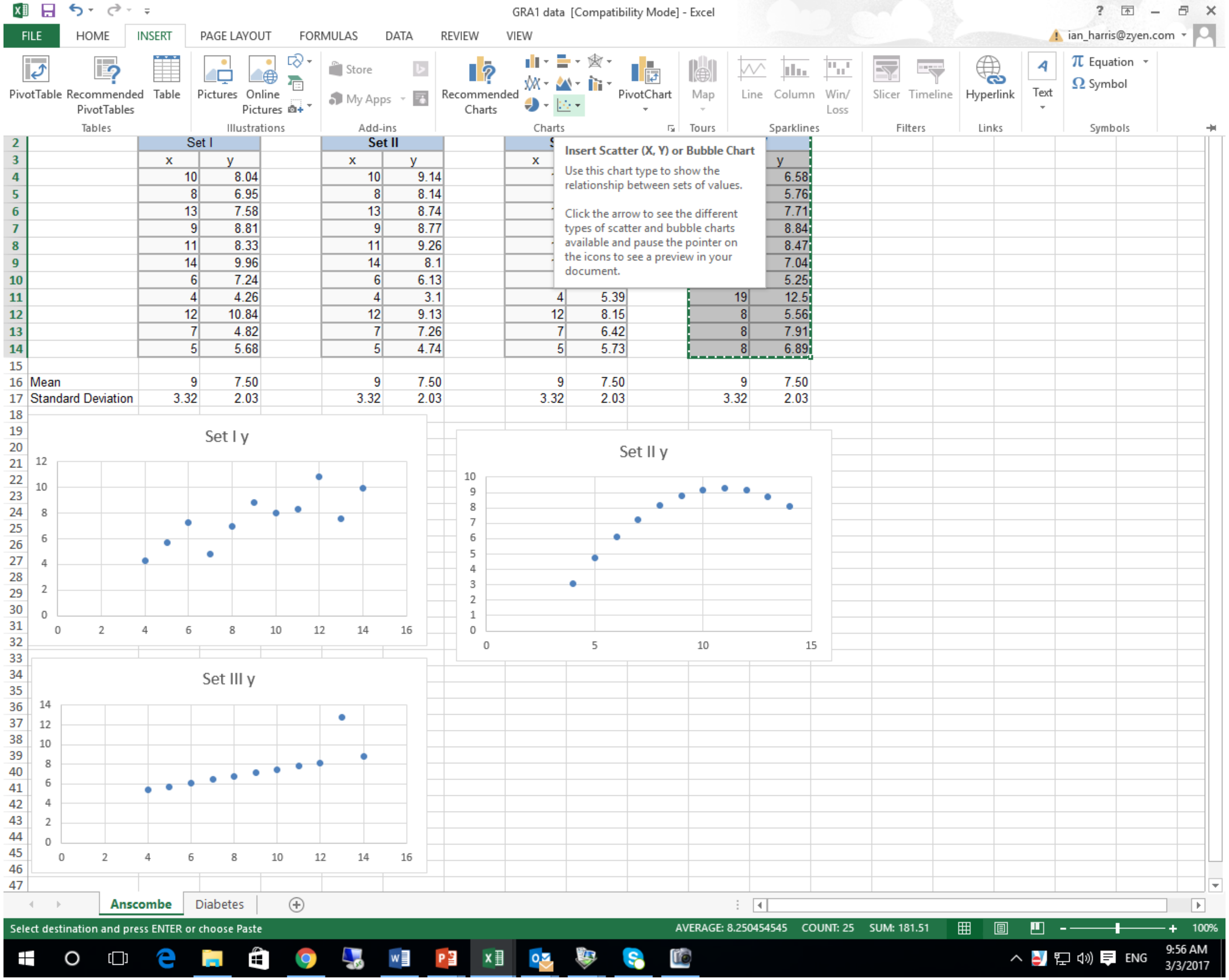
photo source: Correlation from XKCD

The Anscombe Quartet

	Anscombe Quartet										
	Set I			Set II			Set III			Set IV	
	x	y		x	y		x	y		x	y
	10	8.04		10	9.14		10	7.46		8	6.58
	8	6.95		8	8.14		8	6.77		8	5.76
	13	7.58		13	8.74		13	12.74		8	7.71
	9	8.81		9	8.77		9	7.11		8	8.84
	11	8.33		11	9.26		11	7.81		8	8.47
	14	9.96		14	8.1		14	8.84		8	7.04
	6	7.24		6	6.13		6	6.08		8	5.25
	4	4.26		4	3.1		4	5.39		19	12.5
	12	10.84		12	9.13		12	8.15		8	5.56
	7	4.82		7	7.26		7	6.42		8	7.91
	5	5.68		5	4.74		5	5.73		8	6.89
Mean	9	7.50		9	7.50		9	7.50		9	7.50
Standard Deviation	3.32	2.03		3.32	2.03		3.32	2.03		3.32	2.03

© Z/Yen Group
2017

Anscombe Quartet – Make Music



		Set I		Set II		S			
3		x	y	x	y	x			y
4		10	8.04	10	9.14				6.58
5		8	6.95	8	8.14				5.76
6		13	7.58	13	8.74				7.71
7		9	8.81	9	8.77				8.84
8		11	8.33	11	9.26				8.47
9		14	9.96	14	8.1				7.04
10		6	7.24	6	6.13				5.25
11		4	4.26	4	3.1	4	5.39	19	12.5
12		12	10.84	12	9.13	12	8.15	8	5.56
13		7	4.82	7	7.26	7	6.42	8	7.91
14		5	5.68	5	4.74	5	5.73	8	6.89
15									
16	Mean	9	7.50	9	7.50	9	7.50	9	7.50
17	Standard Deviation	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03

Insert Scatter (X, Y) or Bubble Chart

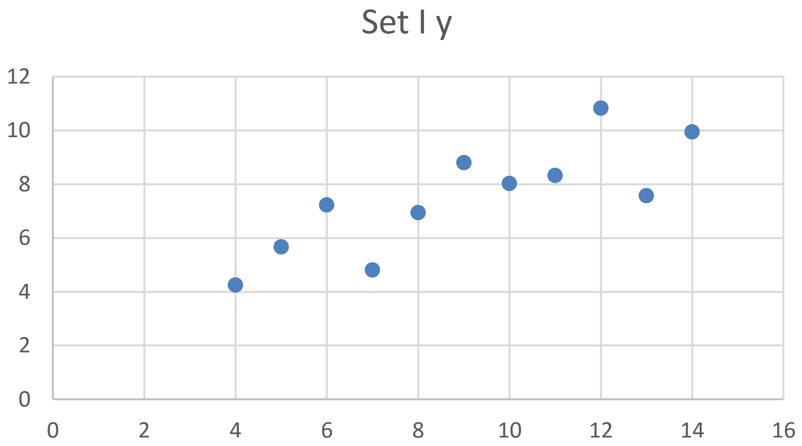
Use this chart type to show the relationship between sets of values.

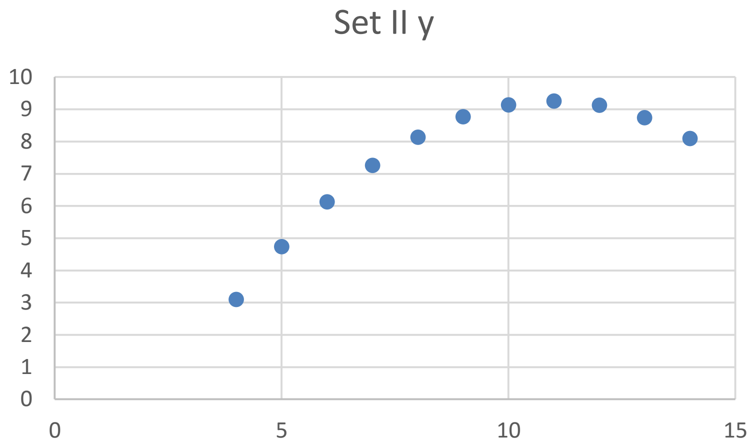
Click the arrow to see the different types of scatter and bubble charts available and pause the pointer on the icons to see a preview in your document.

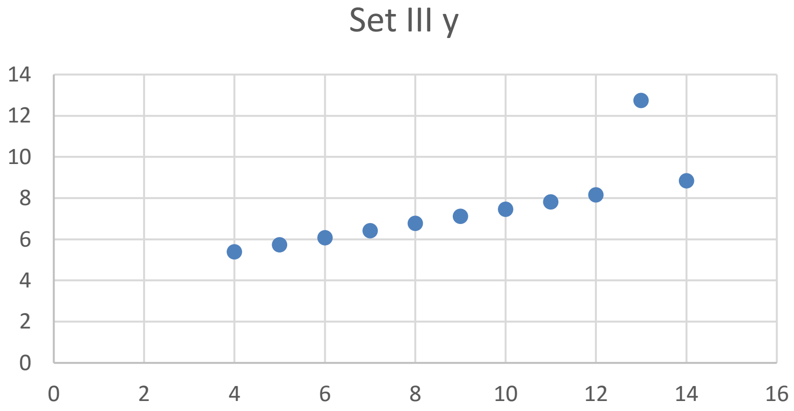
Select destination and press ENTER or choose Paste

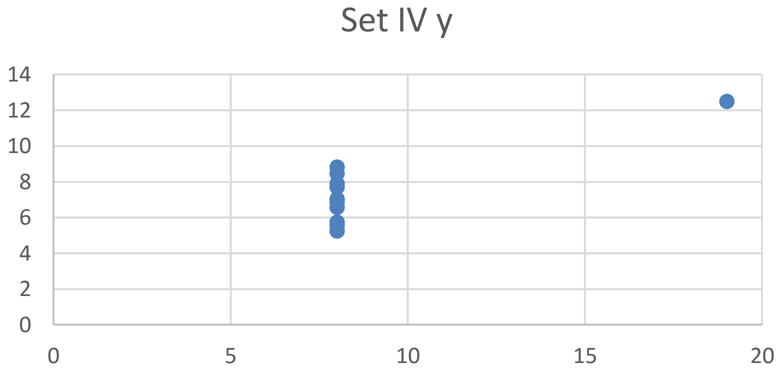
AVERAGE: 8.250454545 COUNT: 25 SUM: 181.51 100%

Anscombe Quartet – Picture This



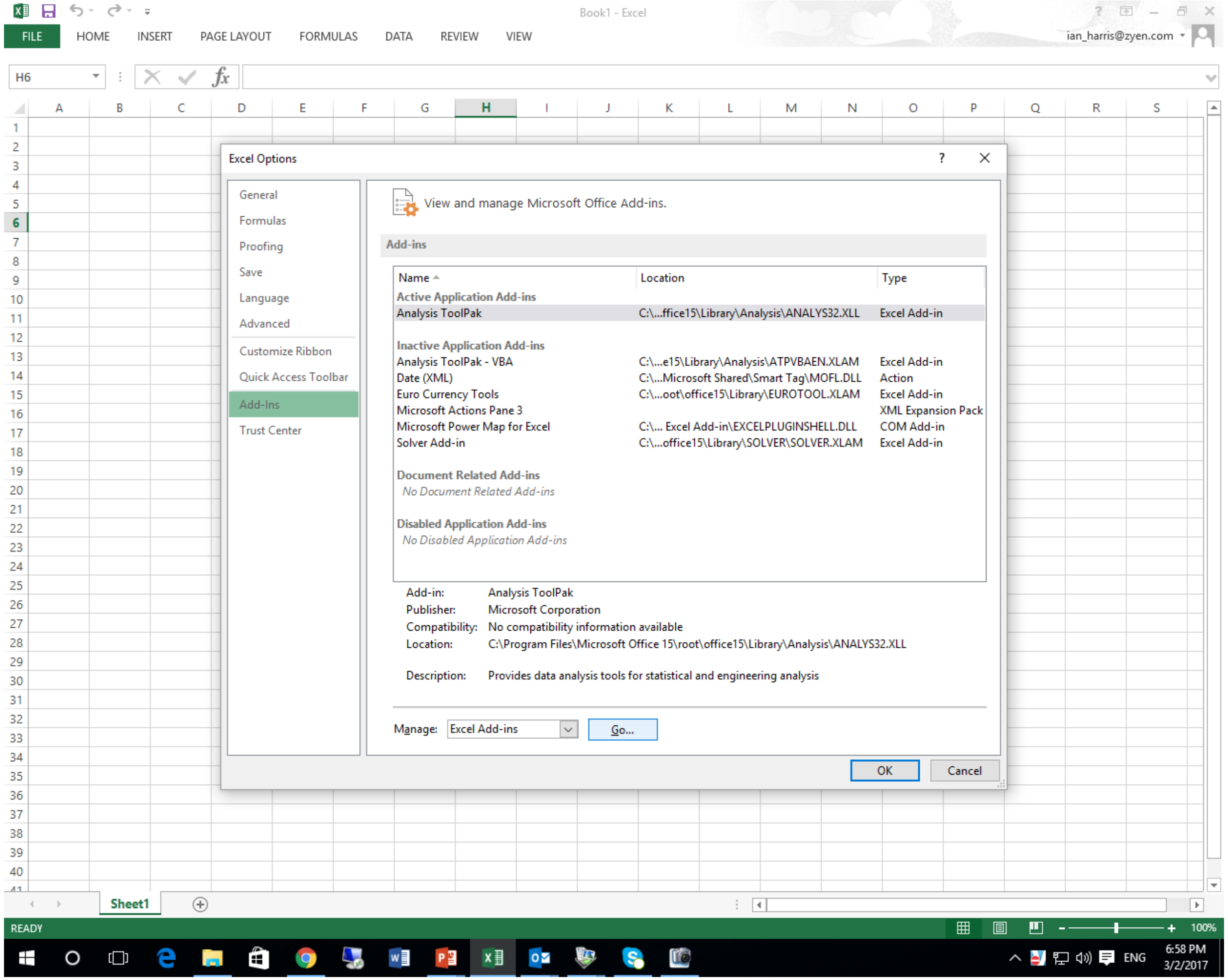






Data Analysis- Excel

File – Options – Add-Ins – Go – Analysis ToolPak



View and manage Microsoft Office Add-ins.

Add-ins

Name	Location	Type
Active Application Add-ins		
Analysis ToolPak	C:\...ffice15\Library\Analysis\ANALYS32.XLL	Excel Add-in
Inactive Application Add-ins		
Analysis ToolPak - VBA	C:\...e15\Library\Analysis\ATPVBAEN.XLAM	Excel Add-in
Date (XML)	C:\...Microsoft Shared\Smart Tag\MOFL.DLL	Action
Euro Currency Tools	C:\...oot\office15\Library\EUROTOOL.XLAM	Excel Add-in
Microsoft Actions Pane 3		XML Expansion Pack
Microsoft Power Map for Excel	C:\... Excel Add-in\EXCELPLUGINSHELL.DLL	COM Add-in
Solver Add-in	C:\...office15\Library\SOLVER\SOLVER.XLAM	Excel Add-in
Document Related Add-ins		
No Document Related Add-ins		
Disabled Application Add-ins		
No Disabled Application Add-ins		

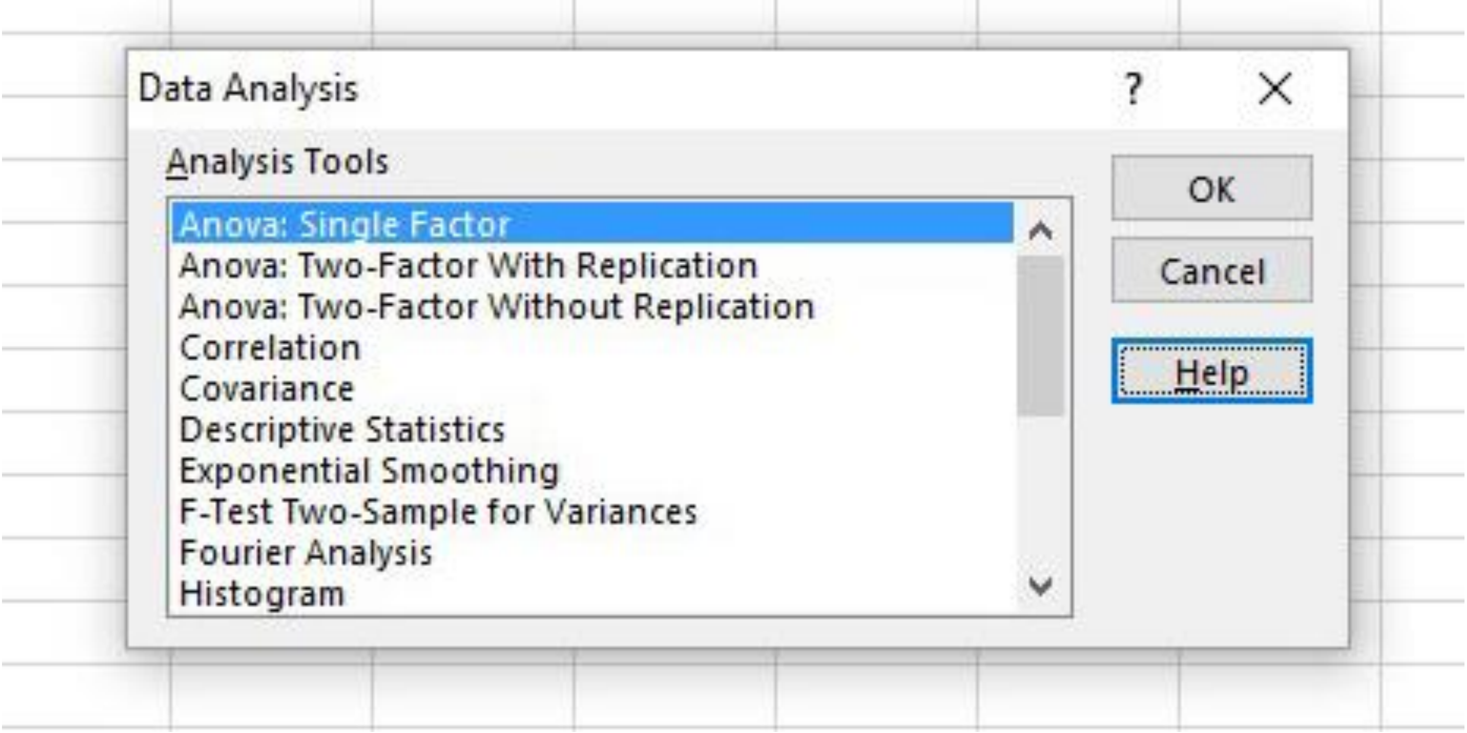
Add-in: Analysis ToolPak
Publisher: Microsoft Corporation
Compatibility: No compatibility information available
Location: C:\Program Files\Microsoft Office 15\root\office15\Library\Analysis\ANALYS32.XLL

Description: Provides data analysis tools for statistical and engineering analysis

Manage: Excel Add-ins [Go...]

[OK] [Cancel]

Analysis Functions A to H



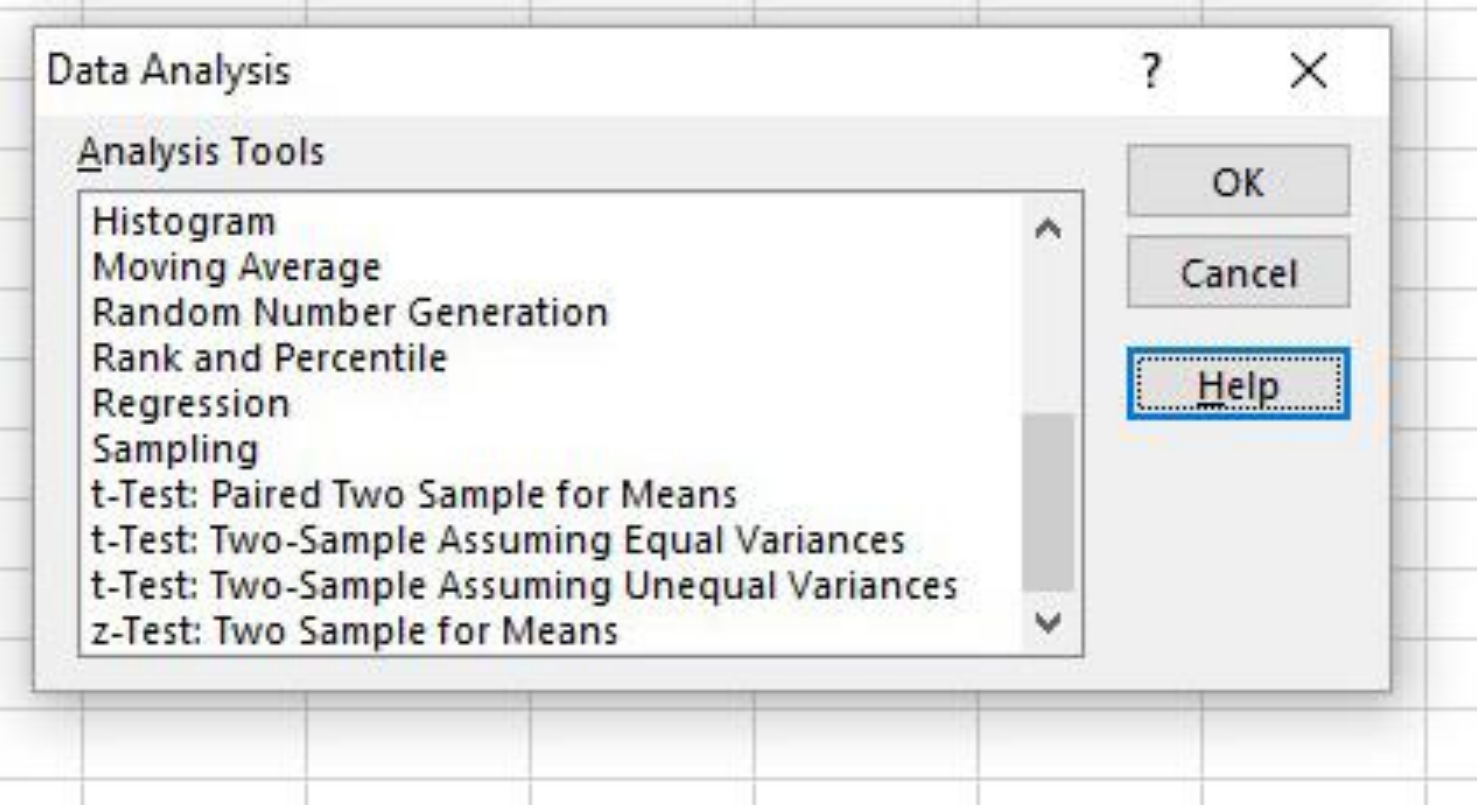
Help actually does help, most of the time…

…plenty of free on-line tutorials if you get stuck

Analysis Functions H to Z

Help actually does help, most of the time…

…plenty of free on-line tutorials if you get stuck



Data Analysis
?
Analysis Tools
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means
OK
Cancel
Help

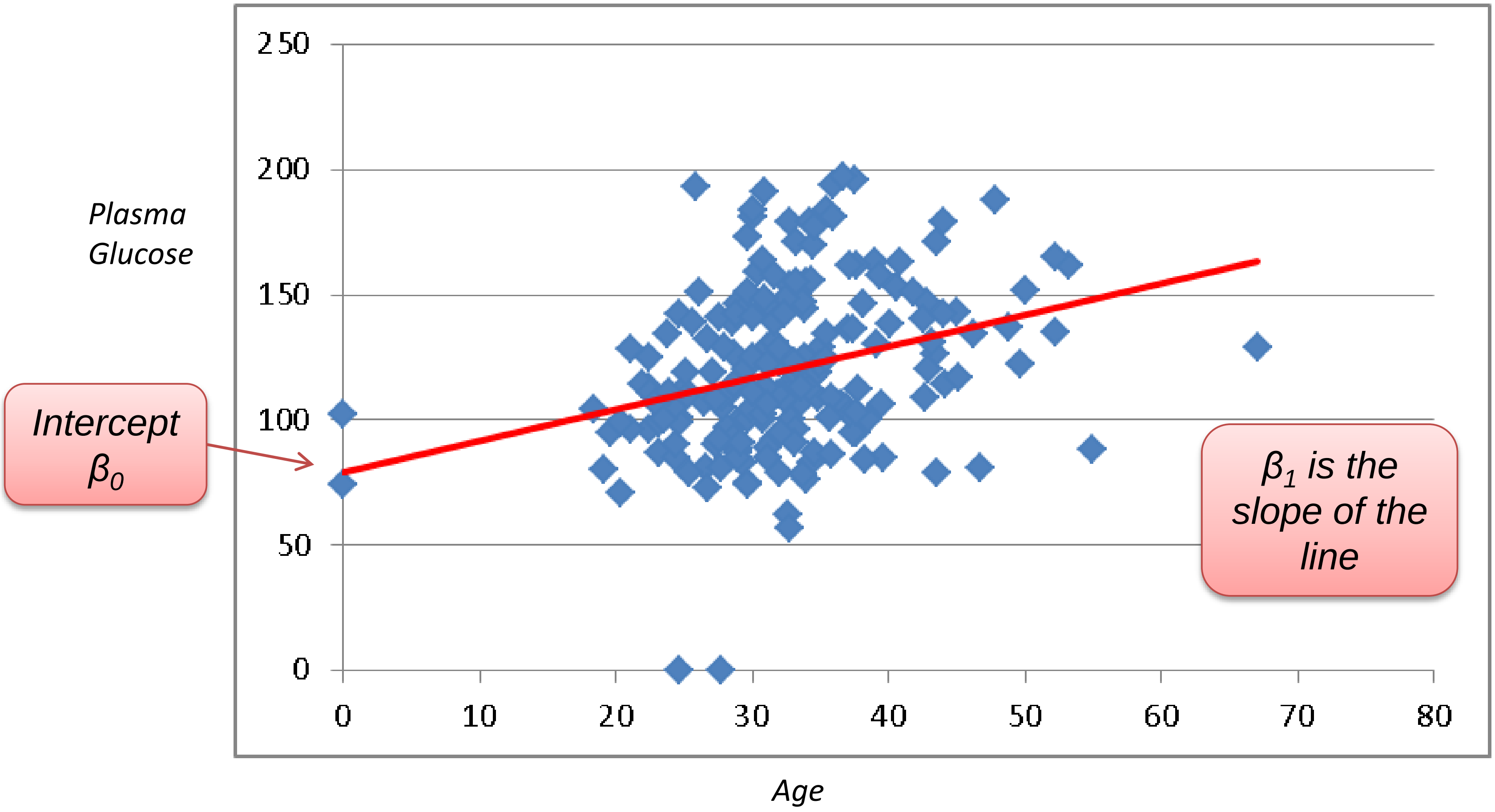
Regression

Definitions:

- *regression* analysis is a statistical process for estimating the relationships among variables.
  - includes many techniques for modeling;
  - the focus is on the relationship between a dependent variable and one or more independent variables;
  - e.g. linear regression, multiple regression.

$$Y = \beta_o + \beta_1 X_1 + \varepsilon$$

Regression Line

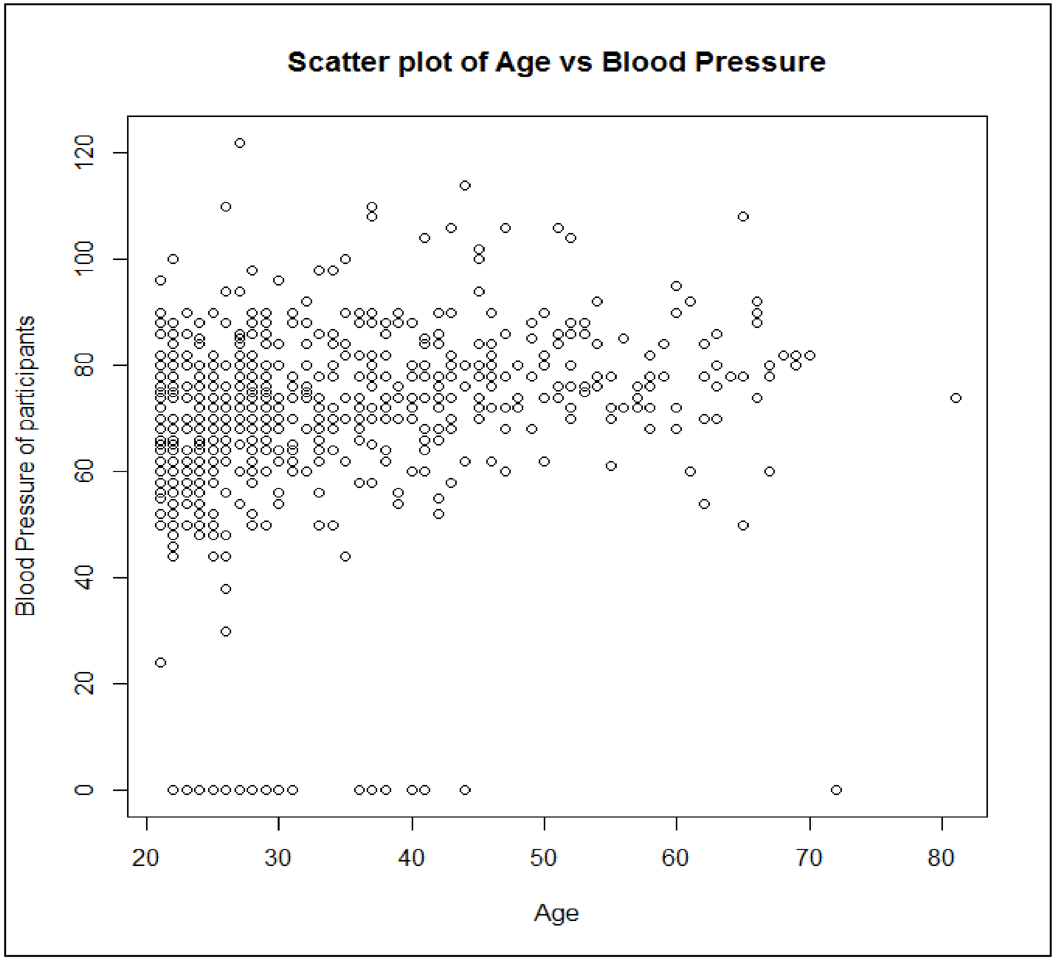


Intercept β_0

β_1 is the slope of the line

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

R example

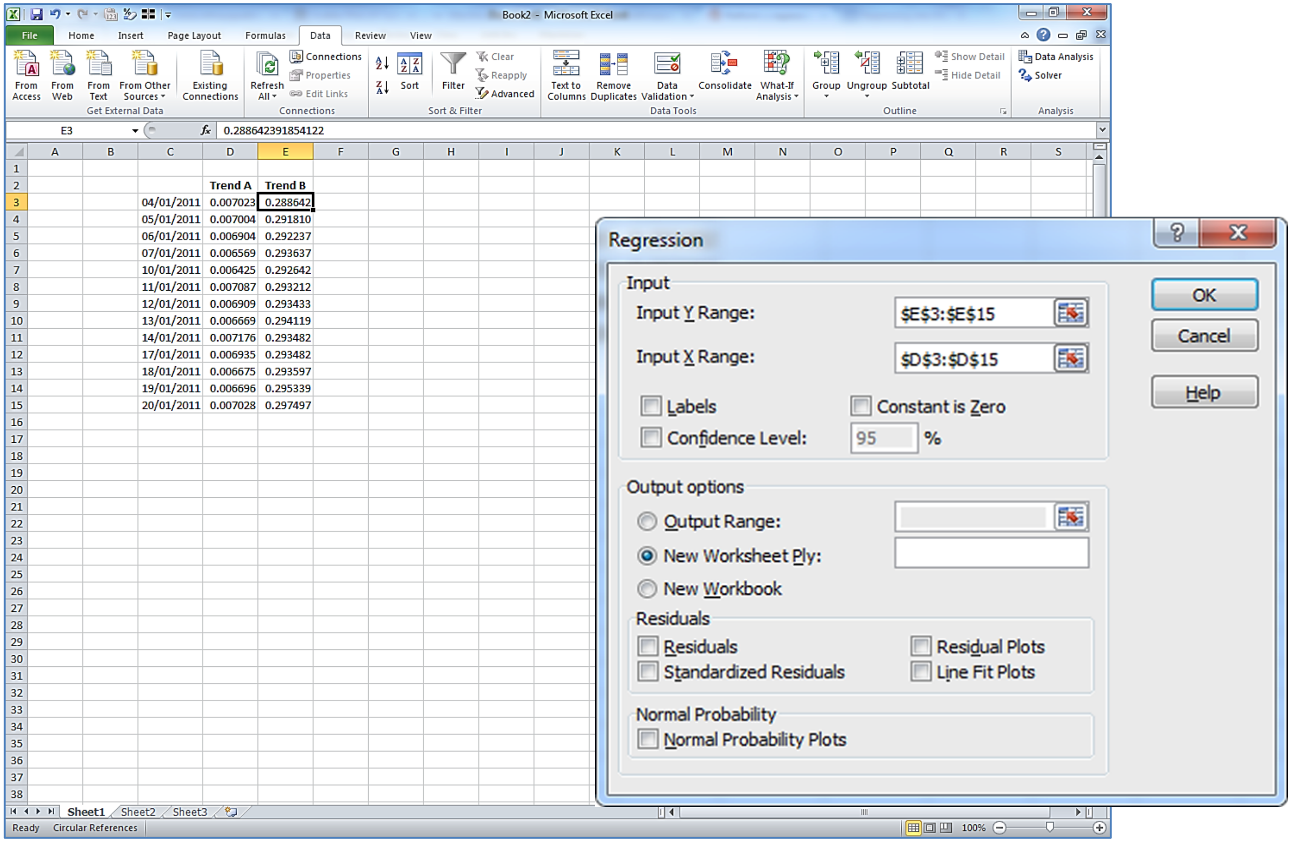


R Code

```
DiabTrain<-read.csv('DiabTrain.csv',head=TRUE)
Age = DiabTrain$Age
BloodPressure= DiabTrain$DiastolicBloodPr
plot(Age,BloodPressure, ylab="Blood Pressure of participants", main = paste("Scatter plot of
Age vs Blood Pressure"))
```

Regression Analysis

The Regression Analysis tool - conducts regression analysis based on the data specified.



E3 0.288642391854122

	Trend A	Trend B
04/01/2011	0.007023	0.288642
05/01/2011	0.007004	0.291810
06/01/2011	0.006904	0.292237
07/01/2011	0.006569	0.293637
10/01/2011	0.006425	0.292642
11/01/2011	0.007087	0.293212
12/01/2011	0.006909	0.293433
13/01/2011	0.006669	0.294119
14/01/2011	0.007176	0.293482
17/01/2011	0.006935	0.293482
18/01/2011	0.006675	0.293597
19/01/2011	0.006696	0.295339
20/01/2011	0.007028	0.297497

Input

Input Y Range: $E$3:$E$15

Input X Range: $D$3:$D$15

□ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

Assumptions & Limitations of Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Key Assumptions:

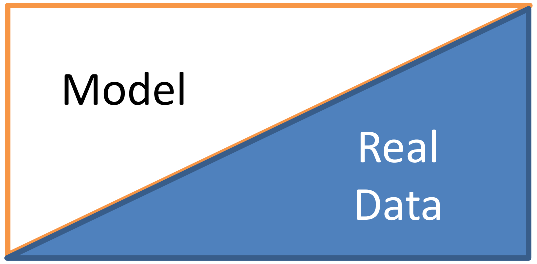
- the relationship between the dependent variable Y and the independent variable X is linear in the slope and intercept parameters β_0 and β_1;
- the independent variable X is not random;
- the expected value of the error term "ε" is 0;
- the variance of the error term is constant for all observations;
- the error term ε is uncorrelated across observations;
- the distribution of the error terms is normal.

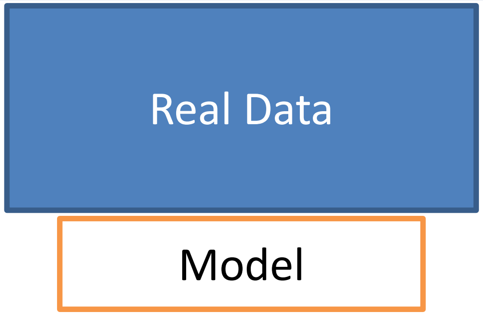
Key Limitations:

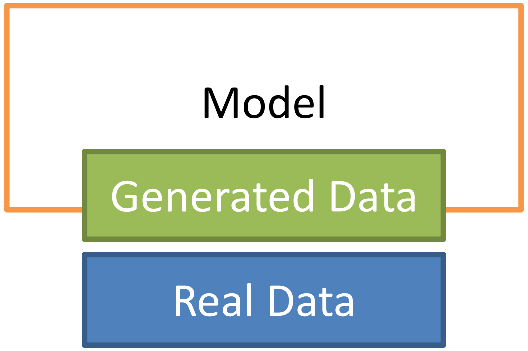
- the estimated parameters and other relationships may change over time;
- in the real world the key assumptions are often unrealistic.

Research Types

Traditional Research	Data-Adaptive Research	Model-dependent Research
Begins with theory/model	Begins with data	Begins with model specification
Uses Classical or Bayesian statistical inference	Searches for useful predictors	Model used to generate data, predictions and make recommendations
Fits models to data	Adapts to the data	Compares generated data with real data
Uses Linear Regression to estimate parameters for linear predictors.	Useful for non-linear relationships & interaction among variables.	Uses simulations and mathematical programming methods.







Monte Carlo simulation

Definition:

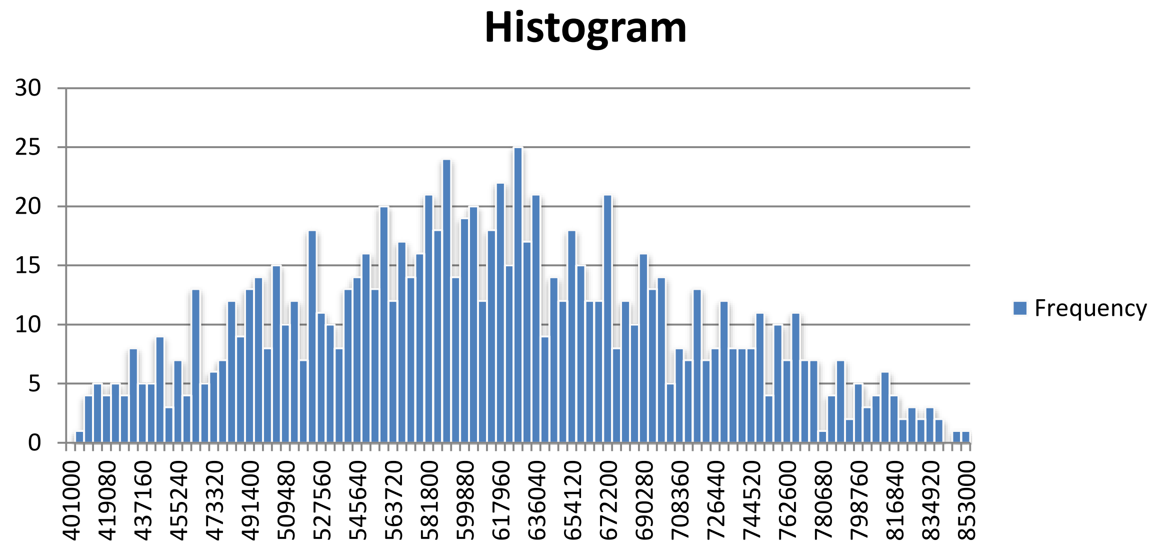
- *Monte Carlo simulation uses repeated sampling to determine the properties of some phenomenon (or behaviour).*
- So called due to methodological and filial link with gambling:
  - play game;
  - record result;
  - (inventor Stanislaw Ulam's uncle reputedly was a regular in that famous casino).

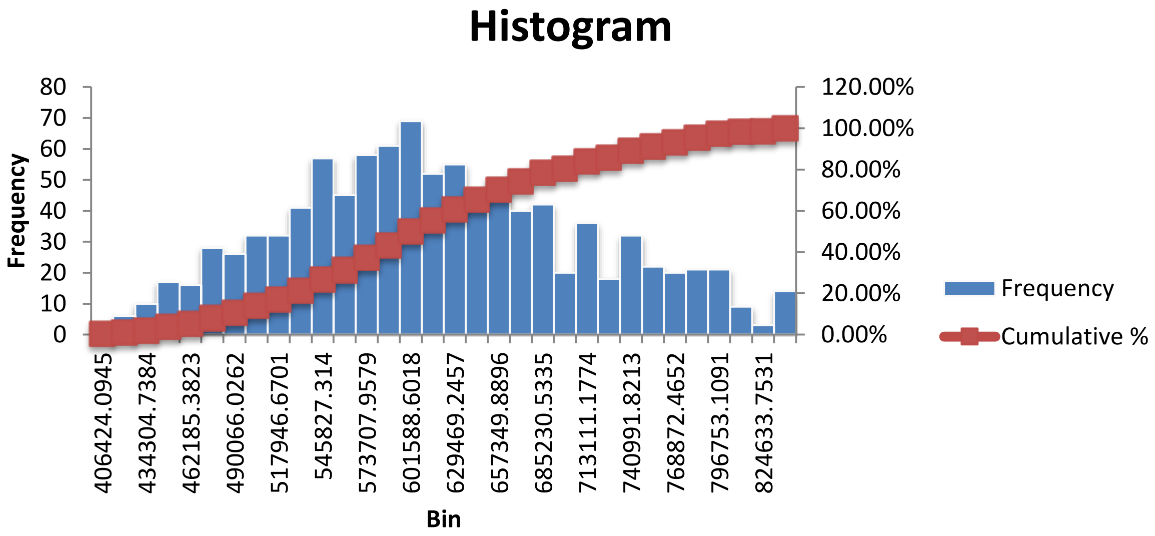
Monte Carlo & Predictive Analytics

- ♦ Example - opportunity tracking in Excel
- ♦ Excel’s random variable can be used to generate uniform and normal distributions for Monte Carlo models
- ♦ Combined with a bit of trigonometry, triangular distributions can also be simulated in Excel
- ♦ Visual expression of results - the use of frequency functions and histograms

© Z/Yen Group
2017

Monte Carlo Visual





Tetlock on Experts and CHAMP

- Philip Tetlock's 18 year study observing 284 experts making 28,000 forecasts. Many/most "experts" hard-pressed to do better than chance, overconfident and reluctant to change their minds in response to new evidence.
- Solution? Use CHAMP:
  - Comparisons are important;
  - Historical trends can help;
  - Average opinions over diverse groups – "the wisdom of crowds";
  - Mathematical models should be taken into account;
  - Predictable biases exist and should be allowed for.

Reference: How To See Into the Future, Tim Harford, Financial Times, 5 September 2014, http://www.ft.com/cms/s/2/3950604a-33bc-11e4-ba62-00144feabdc0.html

Machine Learning

Definition: *machine learning* relates to the construction of algorithmic systems that can learn from data.

- focuses on prediction, based on *known* properties learned from training data;
- includes decision tree learning, neural networks and support vector machines (SVMs);
- can accommodate all five elements of Tetlock's CHAMP – especially good at "P for pesky biases".

Image Credit: mysliderule.com

Markoff (McCarthy/Englebart) Distinction:
Artificial Intelligence? - barely
Intelligence Augmentation? – yes, really!

Some SVM Characteristics

- Copes well with somewhat incomplete and dirty data sets
  - recognises and ignores nulls
  - can be used to clean data
- Enables analysis of many variables at the same time
  - Multi-dimensional
  - Ignores unhelpful variables
  - Curves as well as lines
- Classification, prediction and anomaly detection

Other Advantages

- Machine learning methods are particularly effective in situations where predictive insights need to be uncovered from data sets that are large, diverse and fast changing;
  - outperform traditional methods based on accuracy, scale, and speed.
- Machine learning methods are also useful in analyzing data from multiple sources such as transactional, social media, and other sources
- Stable elements can be embedded in processes yet remain data adaptive (e.g. “Rubies In The Dust” fundraising example and “Rust Never Sleeps” lapsed member recovery process)

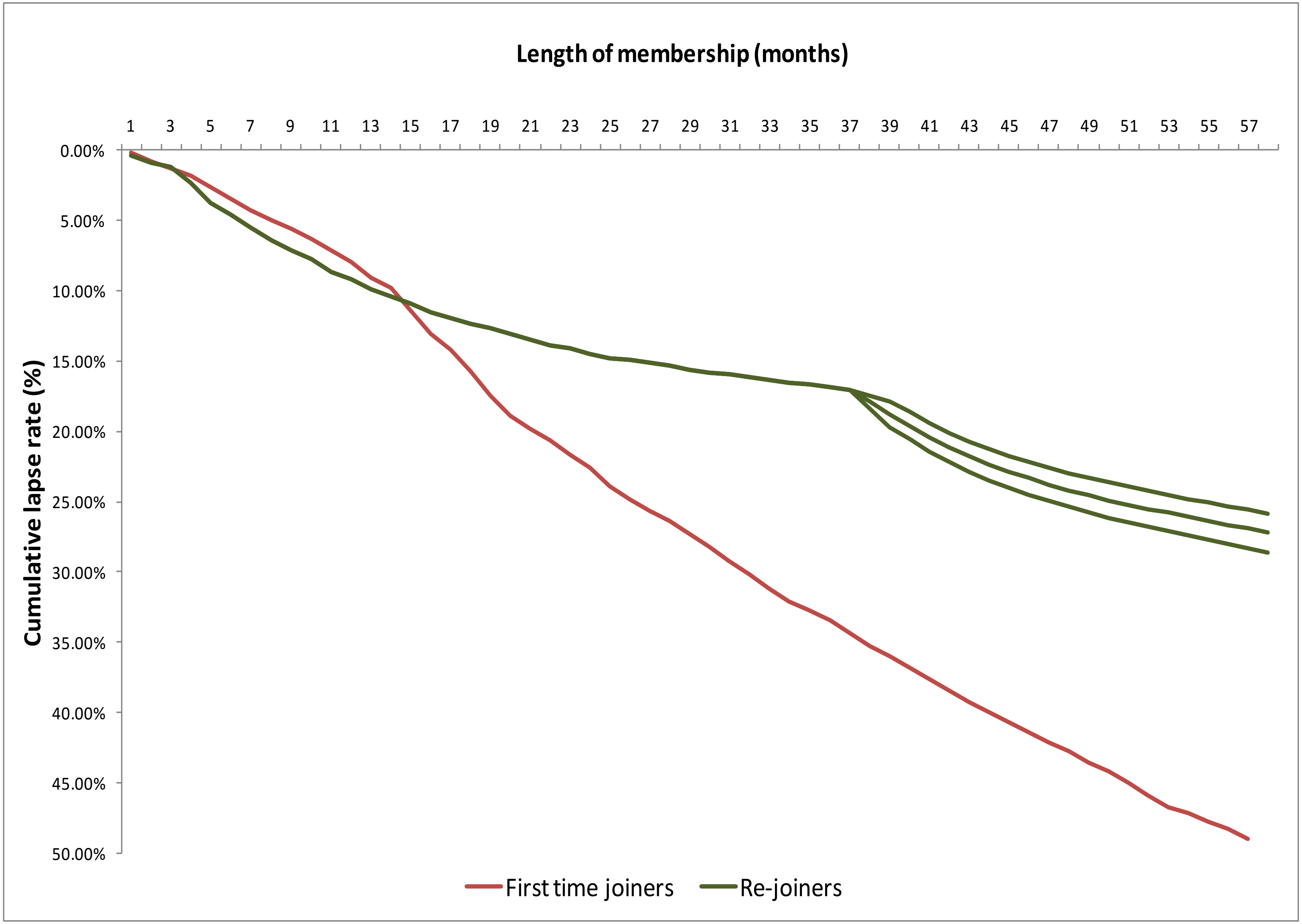
Rekindling Donor Lists - Table

Likelihood block	Potential donors identified by SVM	Actual donors in response to campaign mailshot	PropheZy success rate (%)
Highest Block	3,722	1,645	44.20%
Very High Block	5,837	1,393	23.86%
Quite High Block	6,520	1,239	19.00%
Un-special	103,566	4,828	4.66%
MAILSHOT TOTAL	119,645	9,105	7.61%

Members Rejoining - Table

Propensity of re-joining	Total members	Actual re-joiners*	Actual re-joiner rate
High	192	16	8.33%
Medium	11,742	491	4.18%
Low	16,164	318	1.97%

Members Rejoining - Graph



Further Reading

- Predicting the Effectiveness of Grant-Making, Ian Harris, Michael Mainelli, Peter Grant and Jenny Harrow, 2006, Journal of Strategic Change
- Rubies In the Dust & Rust Never Sleeps, Ian Harris & Mary O'Callaghan, 2012 & 2013, Charity Finance
- Evidence Of Worth In Not-For-Profit Sector Organisations, Ian Harris, Michael Mainelli and Mary O'Callaghan, 2002,Journal of Strategic Change
- How To See Into the Future, Tim Harford, Financial Times, 5 September 2014, http://www.ft.com/cms/s/2/3950604a-33bc-11e4-ba62-00144feabdc0.html
- Machine Learning and Professional Work – A Lookahead To 2040, Ian Harris, SAMi, Autumn 2015