Data Analysis Boot Camp

Student Lab Guide

29400LG_7.0_2017

Data Analysis Boot Camp
Lab Guide

Sentiment Analysis

Unstructured data is a class, or type, of data that is not organized in the sense that it conforms to a well-defined data model. Typically, unstructured data is composed of pure text but is often mixed with other types of data in the larger collection. For example, numbers indicating quantities can be embedded in a long string of text that comprises cooking instructions.

In contrast to unstructured data, structured data conforms to a data model and is significantly easier to parse when the data model is understood. A data model can be as simple defining when a new piece of data begins (such as comma delimited files) to defining and enforcing types and sizes for pieces of data as well as the interrelationships between them (relational databases).

Many forms of data in the real world are a mixture of both. Consider an email, or a tweet: the complete description of the thing we call an email conforms to a data model with fields such as headers, destination and source addresses, and content body. However, inside the content body, the details of an email (or a tweet) are plain text with meaning to the reader, but not conforming to any particular data model which is easily parsed by a computer algorithm.

While computers can efficiently parse text in email messages, figuring out what the text means is very difficult due to the wide array of languages and colloquialisms in everyday communication. Add to that the shortcuts of the common tweet and the computer's job is really hard. For instance, consider the following:

"Thanks a lot, local airline, I got home on time with the delay"
"Thanks a lot, local airline, no way I am getting home now"

What does the phrase "Thanks a lot…" tell you about the feeling of the passenger sending this message?

The exercise of parsing the messages to determine how the customer feels about performance of the company is called sentiment analysis. The messages are parsed by sophisticated modern software called natural language processors, isolating and tagging parts of speech, comparing structural patterns with known rules and perhaps databases of examples of structures in different contexts and assigning weights to the sentiment expressed. When the natural language processors are finished with chewing through all of the messages that are of interest to the company, we, as analysts, have a bunch of numbers assigned to the messages that indicate the sentiment of the messenger.

It is important to remember that, in such natural language processing as sentiment analysis, the analytics are performed on the numbers, not on the text. The natural language processors dissect the text and a model maps characteristics of the text as a whole to numerical values which define the mood. Then, we can crunch the numbers and attempt to identify trends in the messages.

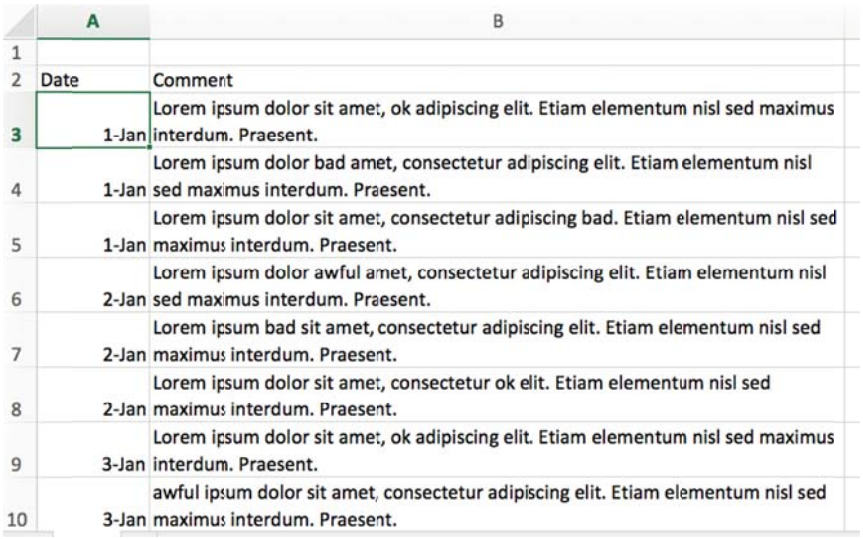
In any analysis of a complex problem, the quality of the analysis depends on the quality of the models and the tools as well as the ability of the analyst. A poor language processing algorithm will many important aspects of common conversation. When combined with a trivial weighting mechanism, sentiment analysis is not a very enlightening technique from which to base any sound business decisions. In this exercise, we will use the worst of both the language parse and the weighting scheme.

The company is new to engaging on social media. When reviewing tweets from 2014, the company notices many bad tweets about its customer service. Beginning in January of 2015, the company begins a yearlong quality initiative. The company monitors tweets to see how the initiative is working. We will use Excel to search the tweets for single instances of several keywords that we decided will make a good model for analysis. Each of the keywords, if found in the text at all, will assure an assignment of a weight based upon that keyword. Once the weights have been assigned, we will use Excel to look for trends in the data by simply probing data with different views.

Note: It is easy to fixate on the quality of the search/replace function as natural language processing. Don't let it distract from the task at hand. This is a basic example to illustrate transforming unstructured text to something that can be analyzed.

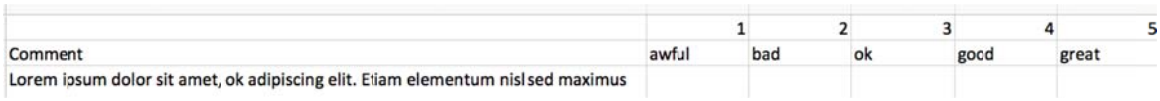
Step by Step

1) Open Sentiment Analysis.xlsx and observe data



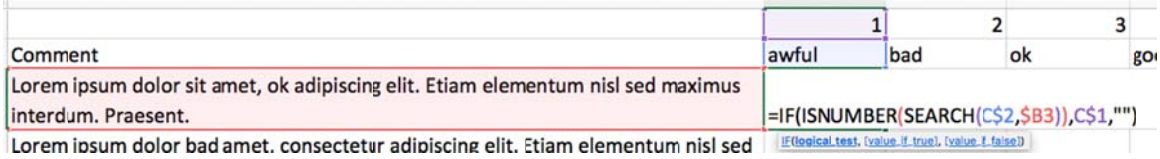
	A	B
1		
2	Date	Comment
3	1-Jan	Lorem ipsum dolor sit amet, ok adipiscing elit. Etiam elementum nisl sed maximus interdum. Praesent.
4	1-Jan	Lorem ipsum dolor bad amet, consectetur adipiscing elit. Etiam elementum nisl sed maximus interdum. Praesent.
5	1-Jan	Lorem ipsum dolor sit amet, consectetur adipiscing bad. Etiam elementum nisl sed maximus interdum. Praesent.
6	2-Jan	Lorem ipsum dolor awful amet, consectetur adipiscing elit. Etiam elementum nisl sed maximus interdum. Praesent.
7	2-Jan	Lorem ipsum bad sit amet, consectetur adipiscing elit. Etiam elementum nisl sed maximus interdum. Praesent.
8	2-Jan	Lorem ipsum dolor sit amet, consectetur ok elit. Etiam elementum nisl sed maximus interdum. Praesent.
9	3-Jan	Lorem ipsum dolor sit amet, ok adipiscing elit. Etiam elementum nisl sed maximus interdum. Praesent.
10	3-Jan	awful ipsum dolor sit amet, consectetur adipiscing elit. Etiam elementum nisl sed maximus interdum. Praesent.

2) Add words for analysis – awful, bad, ok, good, great
3) Add a weight assignment – 1,2,3,4,5



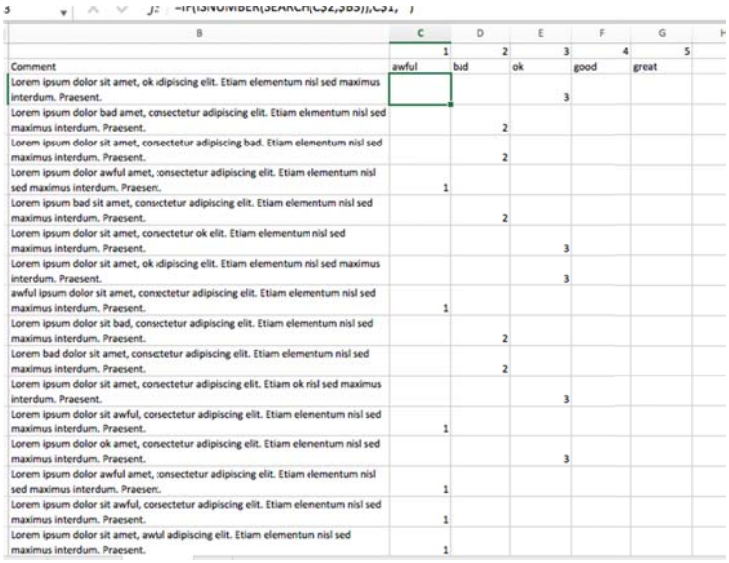
	1	2	3	4	5
Comment	awful	bad	ok	good	great
Lorem ipsum dolor sit amet, ok adipiscing elit. Etiam elementum nisl sed maximus					

4) Add formula as shown



	1	2	3
Comment	awful	bad	ok
Lorem ipsum dolor sit amet, ok adipiscing elit. Etiam elementum nisl sed maximus interdum. Praesent.	=IF(ISNUMBER(SEARCH(C$2,$B3)),C$1,"")		
Lorem ipsum dolor bad amet, consectetur adipiscing elit. Etiam elementum nisl sed	IF(logical_test, [value_if_true], [value_if_false])		

5) View final "sentiment" weight



	B	C	D	E	F	G
		1	2	3	4	5
	Comment	awful	bad	ok	good	great
	Lorem ipsum dolor sit amet, ok adipiscing elit. Etiam elementum nisl sed maximus interdum. Praesent.			3		
	Lorem ipsum dolor bad amet, consectetur adipiscing elit. Etiam elmentum nisl sed maximus interdum. Praesent.		2			
	Lorem ipsum dolor sit amet, consectetur adipiscing bad. Etiam elementum nisl sed maximus interdum. Praesent.		2			
	Lorem ipsum dolor awful amet, consectetur adipiscing elit. Etiam elementum nisl sed maximus interdum. Praesen.	1				
	Lorem ipsum bad sit amet, consctetur adipiscing elit. Etiam elementum nisl sed maximus interdum. Praesent.		2			
	Lorem ipsum dolor sit amet, consectetur ok elit. Etiam elementum nisl sed maximus interdum. Praesent.			3		
	Lorem ipsum dolor sit amet, ok adipiscing elit. Etiam elementum nisl sed maximus interdum. Praesent.			3		
	awful ipsum dolor sit amet, consectetur adipiscing elit. Etiam elementum nisl sed maximus interdum. Praesent.	1				
	Lorem ipsum dolor sit bad, consctetur adipiscing elit. Etiam elementum nisl sed maximus interdum. Praesent.		2			
	Lorem bad dolor sit amet, consctetur adipiscing elit. Etiam elementum nisl sed maximus interdum. Praesent.		2			
	Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam ok risl sed maximus interdum. Praesent.			3		
	Lorem ipsum dolor sit awful, consectetur adipiscing elit. Etiam elementum nisl sed maximus interdum. Praesent.	1				
	Lorem ipsum dolor ok amet, consectetur adipiscing elit. Etiam elementum nisl sed maximus interdum. Praesent.			3		
	Lorem ipsum dolor awful amet, consectetur adipiscing elit. Etiam elementum nisl sed maximus interdum. Praesen.	1				
	Lorem ipsum dolor sit awful, consectetur adipiscing elit. Etiam elementum nisl sed maximus interdum. Praesent.	1				
	Lorem ipsum dolor sit amet, awful adipiscing elit. Etiam elementun nisl sed maximus interdum. Praesent.	1				

6) Perform any analysis suitable to the problem (covered in more detail later in course) – below we see a Pivot Table and a frequency plot of the Pivot Table data)

	A	B	C	D	E	F
1						
2						
3	Row Labels	Count of awful	Count of bad	Count of ok	Count of good	Count of great
4	Jan	32	34	24	3	0
5	Feb	27	19	29	12	0
6	Mar	14	27	38	14	0
7	Apr	5	23	30	32	0
8	May	0	29	28	33	3
9	Jun	0	16	28	34	12
10	Jul	0	7	37	31	18
11	Aug	0	4	28	30	31
12	Sep	0	0	20	36	34
13	Oct	0	0	21	34	38
14	Nov	0	0	6	35	49
15	Dec	0	0	3	33	57
16	**Grand Total**	**78**	**159**	**292**	**327**	**242**

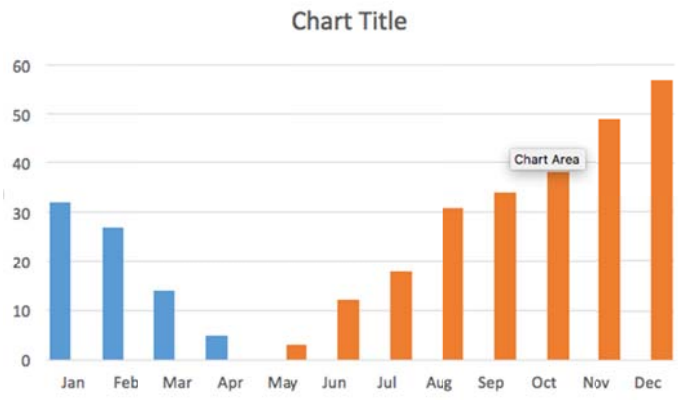


Chart Title
60
50
40
30
20
10
0
Chart Area
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

Data Quality

As we have discussed in the previous section, data in the real world is varied in its form and format. It is also very messy within those forms and formats. That phrase that describes this is data quality. It refers to the state of the data as we bring it into our organization.

The phrase "data quality", in this instance, refers to the state of the data as it is put through an Extract-Transform-Load (ETL) process. In the case of external data coming from a CSV file, there are concerns of completeness of the set, data types in the set being consistent, and consistency of identifying information (this is not an exhaustive list of possibilities).

Due to security requirements at a client site, a survey for the services provided to the client must be given with paper surveys. The paper surveys are then taken to a company which optically scans them and returns a CSV file of the answers to the questions in the survey (bubbling in one of answers 0-9). For a variety of reasons this can be prone to error and the incoming data must be profiled and cleaned before imported to the company database.

The file survey.csv contains the following columns

- Respondent Identifier
- Last Name
- First Name
- Q1 – Q10 (the answer to each of 10 survey questions)

There are multiple responses by each of several respondents as they answer the survey repeatedly over the course of several months. The student should notice the following things:

- Errors in spelling of the last name
- The respondent identifiers are correct regardless of spelling
- A period in various places where an answer (0-9) should be

These can cause the following problems

- Misspelled names can cause entries of the same person to look like another person
- A '.' where there should be a number means that any analysis may yield an incorrect summary
- A '.' can also lead to a type problem when importing into a database because it is not an integer when the database table will be expecting an integer

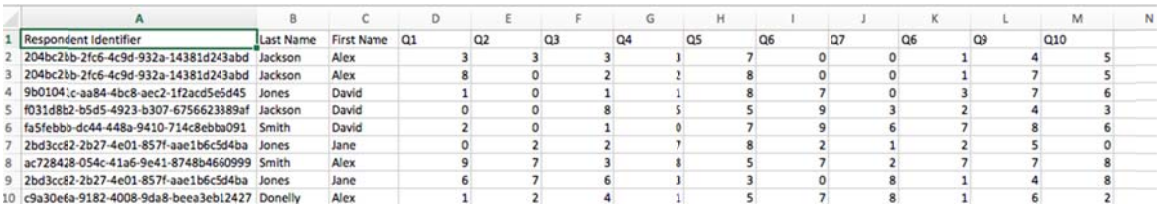
The problems can be found by turning the data into a table in Excel and filtering for a single respondent identifier. This will limit the table to show what should be all the same last name and first name, but the misspellings will become apparent. This is an example of why there should be an SSOT.

If the tools are available, the student can also try to import this into SQL Server and illustrate the type problem on problem.

Lastly, the most important data problem is found by using Excel to find the average of column Q4 in two ways. The first is use the Excel function AVERAGE() which will calculate about 0.5. The second is to use the SUM()/1000 which will calculate about 0.25. The reason is that about 40% of the questions in Q4 are not answered. This leads to poor sampling relative to the expected number of responses. AVERAGE() correctly throws away the unanswered entries, however, it is important to know the correct number of answers to draw correct conclusions from the numbers.

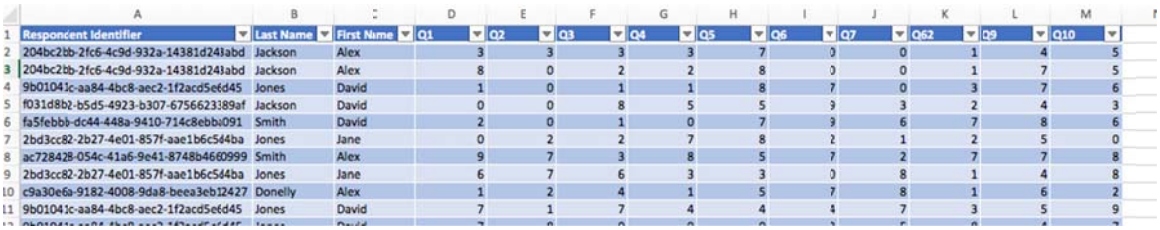
Step by Step

1) Open survey.csv and observe the data



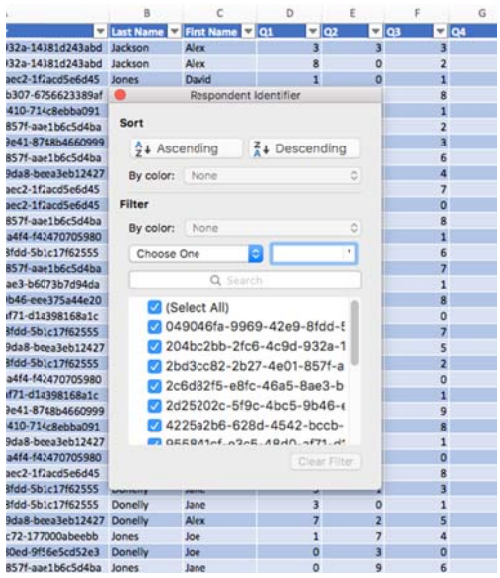
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Respondent Identifier	Last Name	First Name	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q6	Q9	Q10	
2	204bc2bb-2fc6-4c9d-932a-14381d243abd	Jackson	Alex	3	3	3	3	7	0	0	1	4	5	
3	204bc2bb-2fc6-4c9d-932a-14381d243abd	Jackson	Alex	8	0	2	2	8	0	0	1	7	5	
4	9b01041c-aa84-4bc8-aec2-1f2acd5e6d45	Jones	David	1	0	1	1	8	7	0	3	7	6	
5	f031d8b2-b5d5-4923-b307-6756623389af	Jackson	David	0	0	8	5	5	9	3	2	4	3	
6	fa5febbb-dc44-448a-9410-714c8ebba091	Smith	David	2	0	1	0	7	9	6	7	8	6	
7	2bd3cc82-2b27-4e01-857f-aae1b6c5d4ba	Jones	Jane	0	2	2	7	8	2	1	2	5	0	
8	ac728428-054c-41a6-9e41-8748b4660999	Smith	Alex	9	7	3	8	5	7	2	7	7	8	
9	2bd3cc82-2b27-4e01-857f-aae1b6c5d4ba	Jones	Jane	6	7	6	3	3	0	8	1	4	8	
10	c9a30e6a-9182-4008-9da8-beea3eb12427	Donelly	Alex	1	2	4	1	5	7	8	1	6	2	

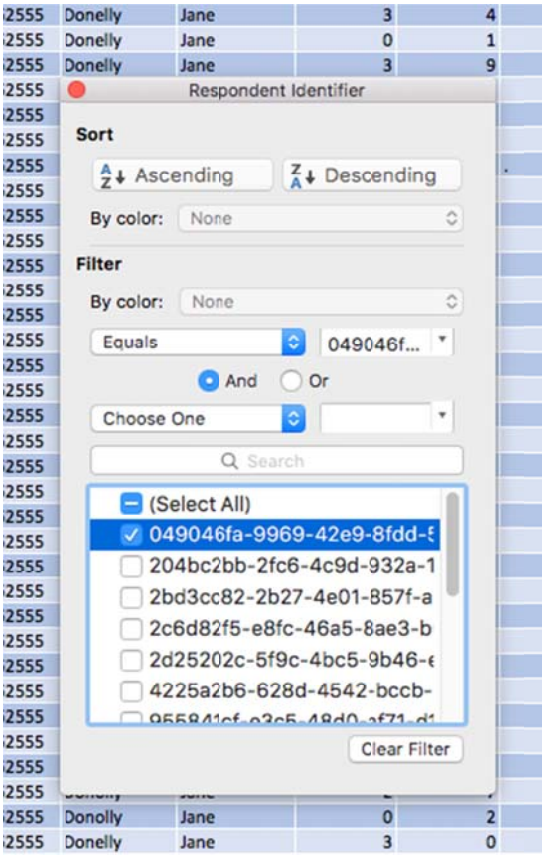
2) Create a table from the data (use the Excel Table option)



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Respondent Identifier	Last Name	First Name	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q62	Q9	Q10
2	204bc2bb-2fc6-4c9d-932a-14381d243abd	Jackson	Alex	3	3	3	3	7	0	0	1	4	5
3	204bc2bb-2fc6-4c9d-932a-14381d243abd	Jackson	Alex	8	0	2	2	8	0	0	1	7	5
4	9b01041c-aa84-4bc8-aec2-1f2acd5e6d45	Jones	David	1	0	1	1	8	7	0	3	7	6
5	f031d8b2-b5d5-4923-b307-6756623389af	Jackson	David	0	0	8	5	5	9	3	2	4	3
6	fa5febbb-dc44-448a-9410-714c8ebba091	Smith	David	2	0	1	0	7	9	6	7	8	6
7	2bd3cc82-2b27-4e01-857f-aae1b6c544ba	Jones	Jane	0	2	2	7	8	2	1	2	5	0
8	ac728428-054c-41a6-9e41-8748b4660999	Smith	Alex	9	7	3	8	5	7	2	7	7	8
9	2bd3cc82-2b27-4e01-857f-aae1b6c544ba	Jones	Jane	6	7	6	3	3	0	8	1	4	8
10	c9a30e6a-9182-4008-9da8-beea3eb12427	Donelly	Alex	1	2	4	1	5	7	8	1	6	2
11	9b01041c-aa84-4bc8-aec2-1f2acd5e6d45	Jones	David	7	1	7	4	4	4	7	3	5	9

3) Filter the data in order to characterize it



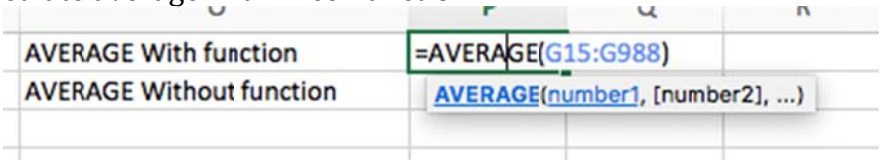


4) Note duplication of identifiers implies misspellings in names

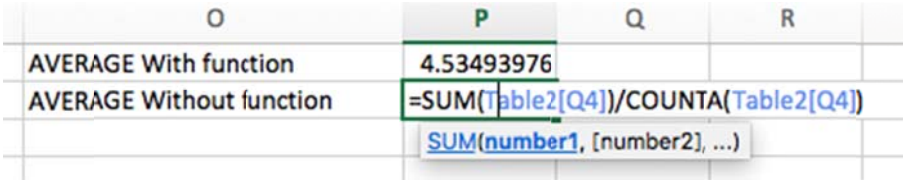
5) Note periods where numbers are expected to be found and its implications on importing data

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Respondent Identifier	Last Name	First Name	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q62	Q9	Q10
15	049046fa-9969-42e9-8fdd-5b1c17f62555	Donelly	Jane	3	4	6	4	0	2	2	4	1	2
20	049046fa-9969-42e9-8fdd-5b1c17f62555	Donelly	Jane	0	1	7	1	5	8	3	9	2	3
12	049046fa-9969-42e9-8fdd-5b1c17f62555	Donelly	Jane	3	9	2	8	2	4	4	0	0	2
10	049046fa-9969-42e9-8fdd-5b1c17f62555	Donelly	Jane	3	1	3	1	6	2	0	1	3	2
11	049046fa-9969-42e9-8fdd-5b1c17f62555	Donelly	Jane	3	0	1	4	8	2	2	6	4	8
19	049046fa-9969-42e9-8fdd-5b1c17f62555	Donelly	Jane	5	2	3	5	7	2	7	7	2	3
16	049046fa-9969-42e9-8fdd-5b1c17f62555	Donelly	Jane	4	7			9	5	0	9	3	1
01	049046fa-9969-42e9-8fdd-5b1c17f62555	Donelly	Jane	7	9	3		6	7	6	6	9	0
45	049046fa-9969-42e9-8fdd-5b1c17f62555	Donelly	Jane			2		3	0	7	2	9	9
54	049046fa-9969-42e9-8fdd-5b1c17f62555	Donelly	Jane	5	4	4	1	1	8	9	4	8	5
14	049046fa-9969-42e9-8fdd-5b1c17f62555	Donelly	Jane		3	3		9	5	0	8	9	8
31	049046fa-9969-42e9-8fdd-5b1c17f62555	Donelly	Jane	1	2	2	8	0	5	3	0	9	3
54	049046fa-9969-42e9-8fdd-5b1c17f62555	Donelly	Jane	0	2	0		4	5	6	7	5	6
64	049046fa-9969-42e9-8fdd-5b1c17f62555	Donelly	Jane	5	3	6	5	8	0	8	9		5
66	049046fa-9969-42e9-8fdd-5b1c17f62555	Donelly	Jane	5	6	3	8	7	9	2	3		7
71	049046fa-9969-42e9-8fdd-5b1c17f62555	Donelly	Jane	2	5	3	5	9	3	9	2	7	4

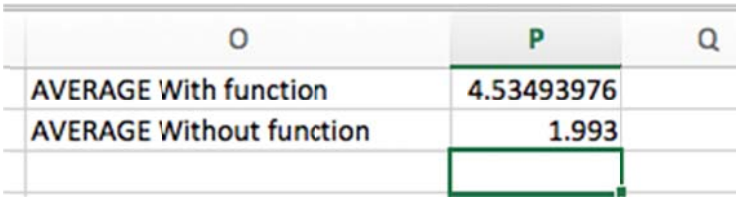
6) Calculate average with Excel function



7) Calculate average using SUM() in Excel



8) Note the difference between the two in Q4 and its implications for data interpretation



Coin Toss with Two Coins

To determine the probability of getting any combination of heads and tails when flipping two coins at one time, we can set up a table with each coin along one dimension of the table (columns and rows). We can combine the column and the row to label the outcomes.

To determine the probability of getting each combination, we first count all of the possible outcomes. Counting each of these possibilities, and recognizing that HT and TH are really the same outcome, we get the numbers that are labeled as Frequency in the figure. Further, we calculate the relative frequency by dividing the frequency for the combination by the total possible outcomes We see the final calculation gives a 50% chance of getting an HT combination and only a 25% chance of getting the HH or TT combination when flipping two coins

Step by Step

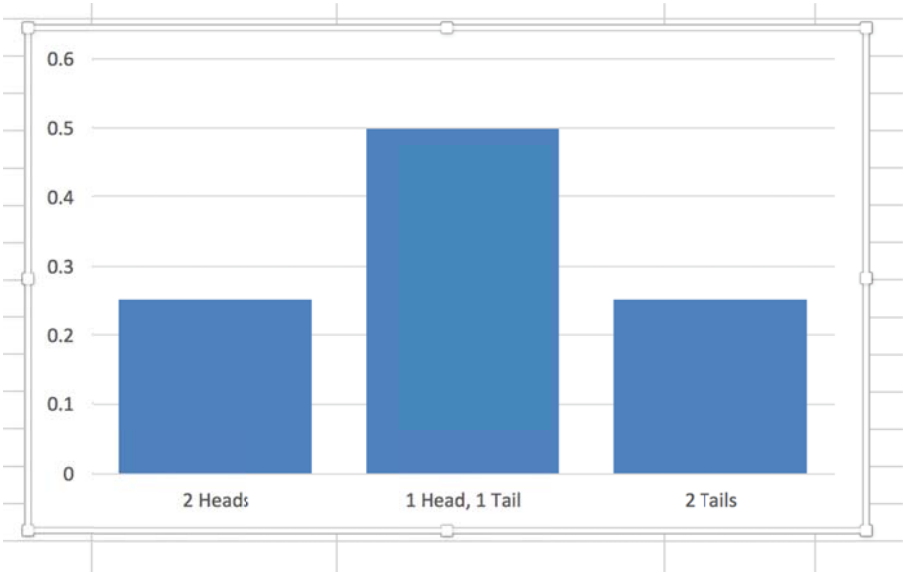
1) Create table with single coin options as columns and other coin options as rows
2) Fill in table with options by combining row and column at the cell in question

A	B	C
Event Space for flipping two coins at the same time		
	H	T
H	HH	HT
T	TH	TT

3) Calculate the total possible outcomes by counting the number of cells in the table
4) Calculate the frequency by counting the number of ways the selected option appears in the table
5) Calculate the relative frequency by dividing the frequency by the total possible outcomes

E	F	G
Possible Outcomes	4	
Outcome	Frequency	Relative Frequency
2 Heads	1	=F6/F3
1 Head, 1 Tail	2	0.5
2 Tails	1	0.25

6) Plot the results in a frequency plot



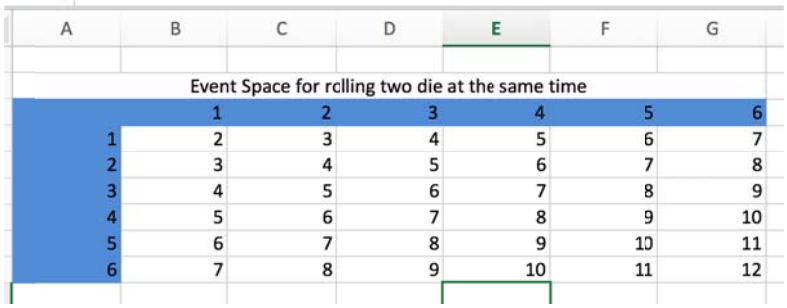
Dice Roll with Two Die

We can extend this technique by looking at the more complicated probability of throwing dice. We consider the two six sided die by labeling a column header and a row header with each possible outcome of single sided die. The matrix lists the possible combinations of the dice roll, ranging from 2 to 12. Again, we enumerate all the outcomes, count the total possible outcomes, 36, and count the possible number of each outcome

Step by Step

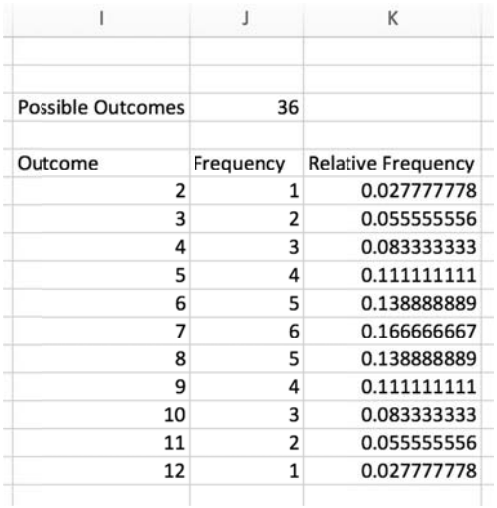
1) Create table with single die options in column and other dies options in rows

2) Fill in the table with combinations of the row and column at the cell in question



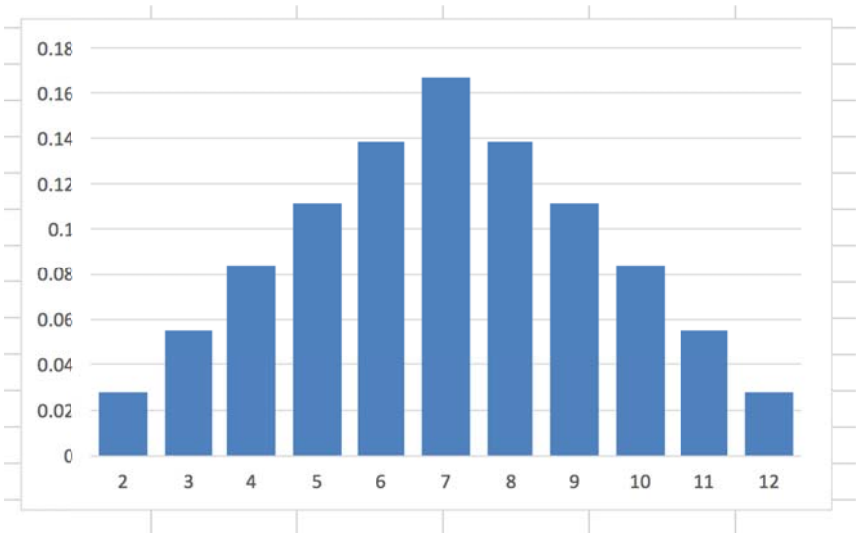
	A	B	C	D	E	F	G
	Event Space for rolling two die at the same time						
		1	2	3	4	5	6
	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

3) Calculate the total possible outcomes by counting the number of cells in the table
4) Calculate the frequency by counting the number of ways the selected option appears in the table
5) Calculate the relative frequency by dividing the frequency by the total possible outcomes



I	J	K
Possible Outcomes	36	
Outcome	Frequency	Relative Frequency
2	1	0.027777778
3	2	0.055555556
4	3	0.083333333
5	4	0.111111111
6	5	0.138888889
7	6	0.166666667
8	5	0.138888889
9	4	0.111111111
10	3	0.083333333
11	2	0.055555556
12	1	0.027777778

6) Plot results in a frequency plot



Single Set of Test Grades

The primary central tendency of a distribution of data is the average. There are 3 types of averages covered in this class:

Arithmetic Mean
The arithmetic mean is the sum of the elements of a set divided by the total number of elements in that set
Median
The value that separates the upper half of the set from the lower half of the set
Mode
The most frequent value of the set

These along with the standard deviation describe pretty well how the data set is laid out. The visualization of how it is laid out is done with the frequency distribution chart. This is simply a vertical bar chart of the count of the grades when they are divided into buckets, or bins. The frequency distribution chart makes it clear what kind of distribution the analyst is working with in the data set.

A professor has given a test and they have been graded. The data set is the list of grades without reference to a student because the analysis should be performed independently of who the students were. The professor wants to know if the test was a fair test given that an average of 75 is expected and the spread should be fairly "tight" around that median value. Too wide of a spread may mean the class has too varied a skill set. A low average with a small spread probably means that the test was too hard, as the whole group did below expected values.

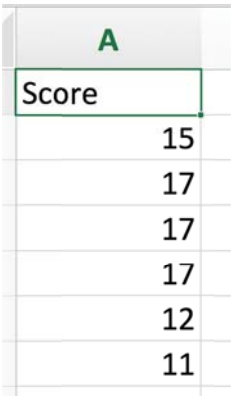
The first thing that the analyst should notice is that the numbers are not out of 100%, or are awfully low if they are. Therefore, the first question should be "out of how many points?". The answer is 20 points. The scores should be scaled in another column and the scaled scores out of 100% should be used.

This exercise can be used in two different places. The first is an introduction to Excel and the FREQUENCY() function where only the notion of an average is understood by the class. In this case, the instructor cannot use the standard deviation or the normal distribution because they will not have been introduced yet. The second is after these concepts have been introduced and to illustrate them.

The instructor can use the same data set in both places so that the students are familiar with the data set when it revisited.

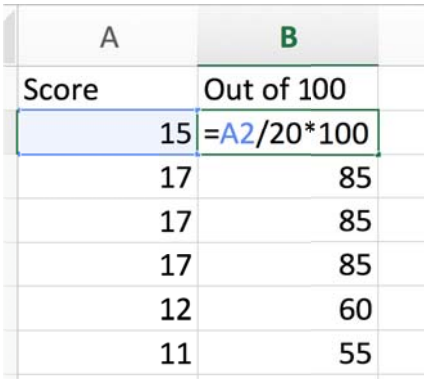
Step by Step

1) Open Gradebook Single.csv and observe a list of grades on a scale of 20 points



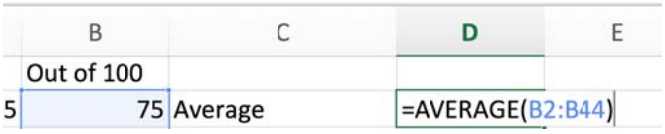
A
Score
15
17
17
17
12
11

2) Rescale the grades by dividing the grades in the list by the max number of points and multiplying by 100



A	B
Score	Out of 100
15	=A2/20*100
17	85
17	85
17	85
12	60
11	55

3) Calculate the average of the grades



	B	C	D	E
	Out of 100			
5	75	Average	=AVERAGE(B2:B44)	

4) Calculate the standard deviation of the grades (more detailed discussion later in the course)

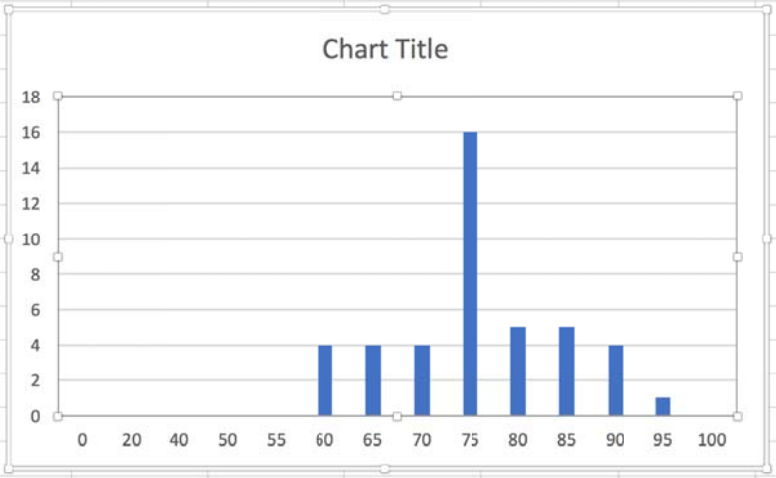
B	C	D	E
Out of 100			
75	Average	75.581395	
85	Standard Deviation	=STDEV.P(B2:B44)	

5) Separate the grades into bins by using the FREQUENCY() function in Excel

	B	C	D	E	F
5	75		Bins	Count in Bin	
7	85		0	=FREQUENCY(B2:B44,D3:D16)	
7	85		20		

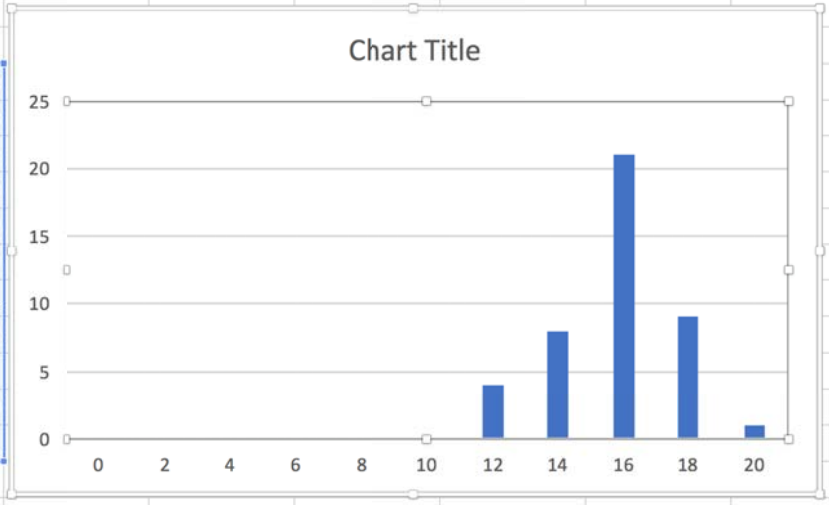
6) Plot the frequency diagram

Bins	Count in Bin
0	0
20	0
40	0
50	0
55	0
60	4
65	4
70	4
75	16
80	5
85	5
90	4
95	1
100	0



7) For comparison, try it with the original, unscaled data

Bins	Count in Bin
0	0
2	0
4	0
6	0
8	0
10	0
12	4
14	8
16	21
18	9
20	1



Full Semester Test Grades

For the same scenario described above, there are three exams. Each of the exam distributions illustrates a different problem with the exam.

A good exam – the average and the standard deviation are where they should be for an exam (around middle C and tightly distributed)
A tough exam – the average is low and the deviation is a little too high indicating a hard exam and a broad level of capabilities
A bimodal exam – the exam has two normal curves on top of each other, illustrating, possibly, a semester in which a really good group of students are trying to get ahead, and a poor group of students are retaking the class
Each of the distributions says something about the class, the instructor, and the exam. Experience is usually the only interpreter

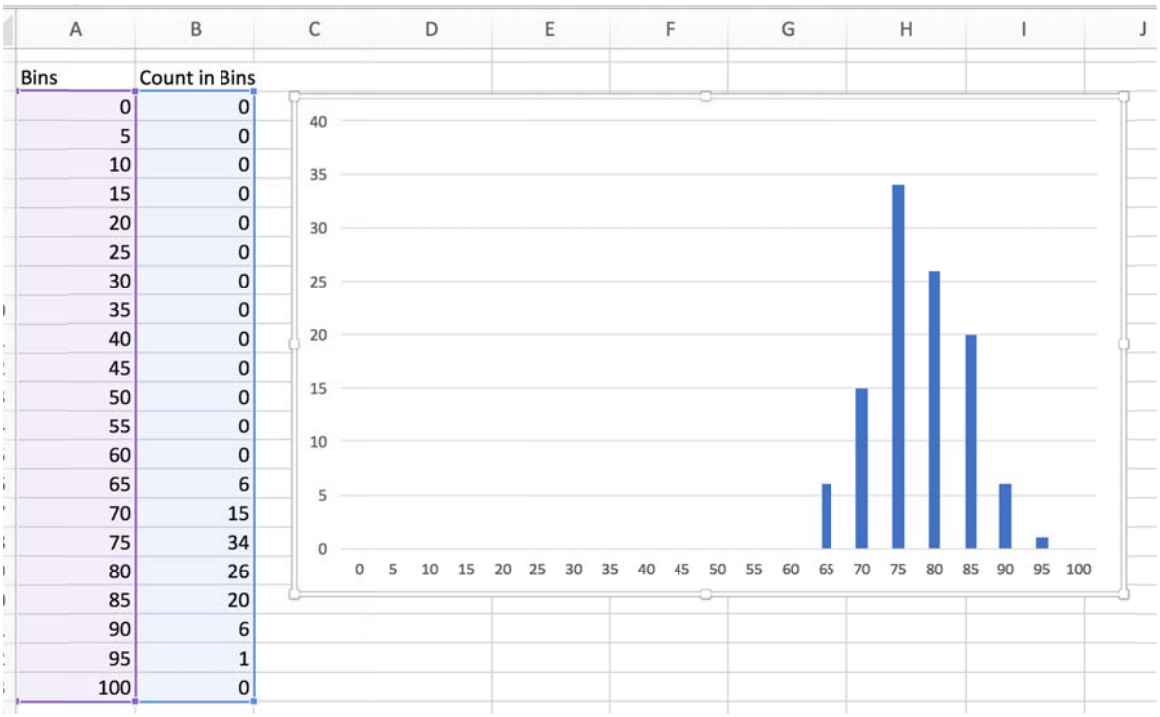
Note: Using simple bins like A,B,C,D and F will not show the normal nature of the distribution. It is better to build bins from 0 to 100 in steps of 5.

Step by Step

1) Open Gradebook.csv and note the data – student id and three test grades for each

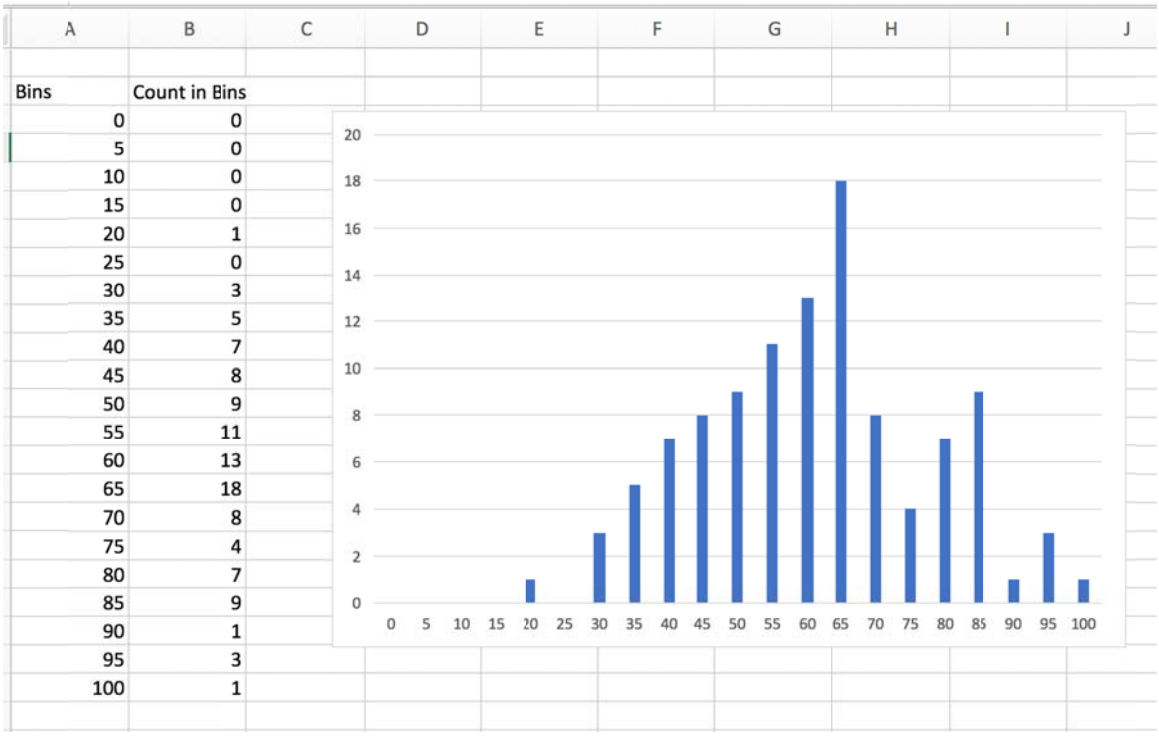
A	B	C	D
Student ID	Test 1	Test 2	Test 3
12836	83	74	28
98550	72	42	26
70656	72	92	38
92645	82	85	25
15558	81	52	26
94452	63	29	24
89538	82	44	27
37490	70	38	22
68678	81	70	30

2) Perform the steps in Single Set of Test Grades for each of the test grades in this file
3) Try different bin arrangements
4) Plot the results of Test 1



Bins	Count in Bins
0	0
5	0
10	0
15	0
20	0
25	0
30	0
35	0
40	0
45	0
50	0
55	0
60	0
65	6
70	15
75	34
80	26
85	20
90	6
95	1
100	0

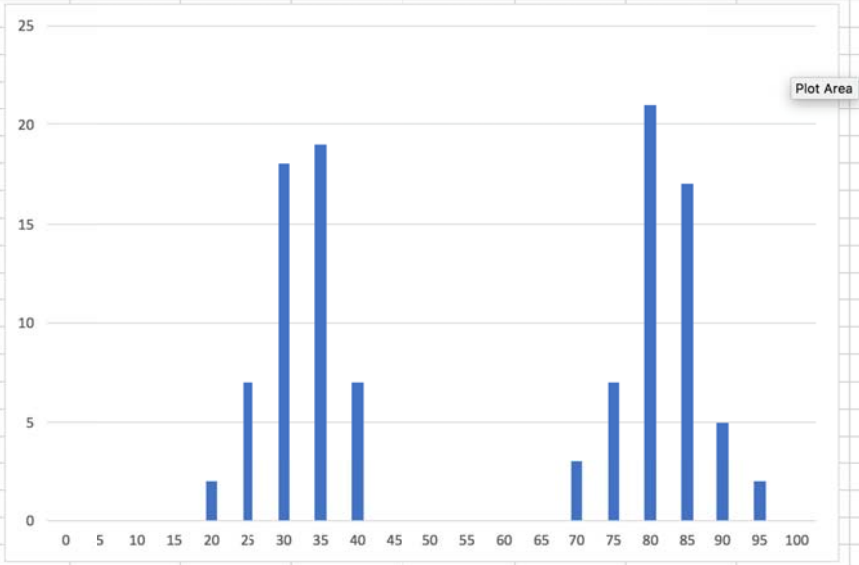
5) Plot the results of Test 2



Bins	Count in Bins
0	0
5	0
10	0
15	0
20	1
25	0
30	3
35	5
40	7
45	8
50	9
55	11
60	13
65	18
70	8
75	4
80	7
85	9
90	1
95	3
100	1

6) Plot the results of Test 3

Bins	Count in Bins
0	0
5	0
10	0
15	0
20	2
25	7
30	18
35	19
40	7
45	0
50	0
55	0
60	0
65	0
70	3
75	7
80	21
85	17
90	5
95	2
100	0



7) Draw conclusions and make recommendations

Heights of Men

Independent events are when the probability of one outcome does not affect the probability of another event occurring. Take, for instance, the measurement of the height of individuals in a group. Assuming we control for genetics by not measuring people in the same family (or other such variables), the heights of people in a group are such that result of measuring one person does not affect the result of measuring another person. The measurements are independent events. In this exercise, we characterize the results of these measurements and observe the result of the measurement of many independent events.

In the Heights Of Men file, there is data of a large group of men whose height is measured and reported in inches. A sample of the data is shown in figure X. We can use Excel to calculate the average and the standard deviation

Step by Step

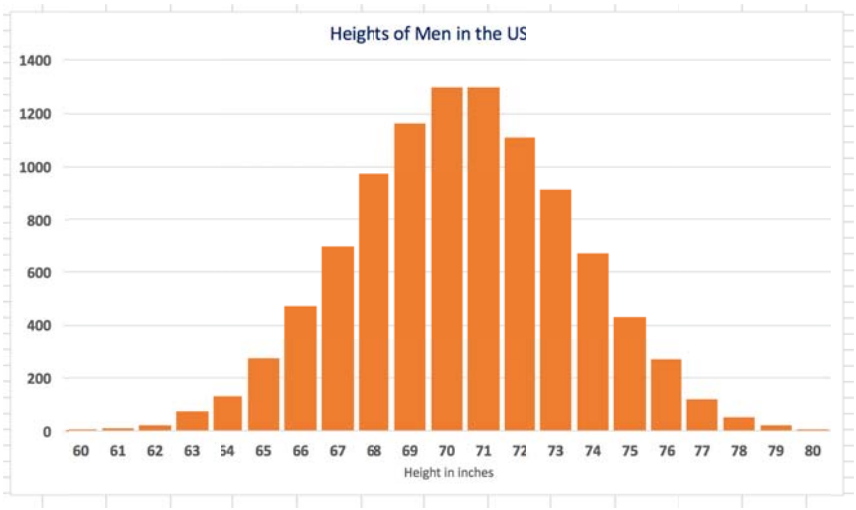
1) Open Heights of Men.csv and define the data to look at

Person Measured	Height (inches)
1	72.2
2	70.2
3	72.0
4	66.9
5	67.2
6	73.2
7	69.6
8	70.3
9	72.0
10	74.2
11	67.2
12	70.8

2) Calculate the average
3) Calculate the standard deviation

AVERAGE	70.0
Std Dev	3.0120648

4) Create the bins
5) Parse with FREQUENCY()
6) Plot the data


Heights of Men in the US
1400
1200
1000
800
600
400
200
0
60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80
Height in inches

Call Center Analysis

Analysis of data often begins with plotting the distribution and calculating the average and standard deviation. This gives us an overall view of the data we are investigating. Once we have identified the distribution, we can use Excel to ask questions about the distribution itself to gather information, make plans, and draw conclusions.

A call center manager has decided to motivate the employees by offering a cash incentive for handling a higher call-per-hour at the center. In order to determine what the incentive number (calls per hour) should be, the analyst needs to profile the current performance of the floor and the manager needs to evaluate the likelihood of achieving certain milestones in order to set a high bar which is still accessible by people on the floor.

The data is list of calls handled per hour by a random selected call center employee. Each log entry is a measure of someone different on the floor so that the entire call center is represented in the data. By calculating the average and the standard deviation, then plotting the data and considering the data meets the following criteria:

- Few occurrences (calls per hour)
- Independent events (measurements are not related to each other)

The Poisson distribution is selected as the proper model for the data and the problem. Using Excel's functionality, we can ask questions like P(calls/hour > 7) or P(calls/hour > 8) and so forth. By evaluating these options, the manager can begin to determine what reasonable incentives may be.

Step by Step

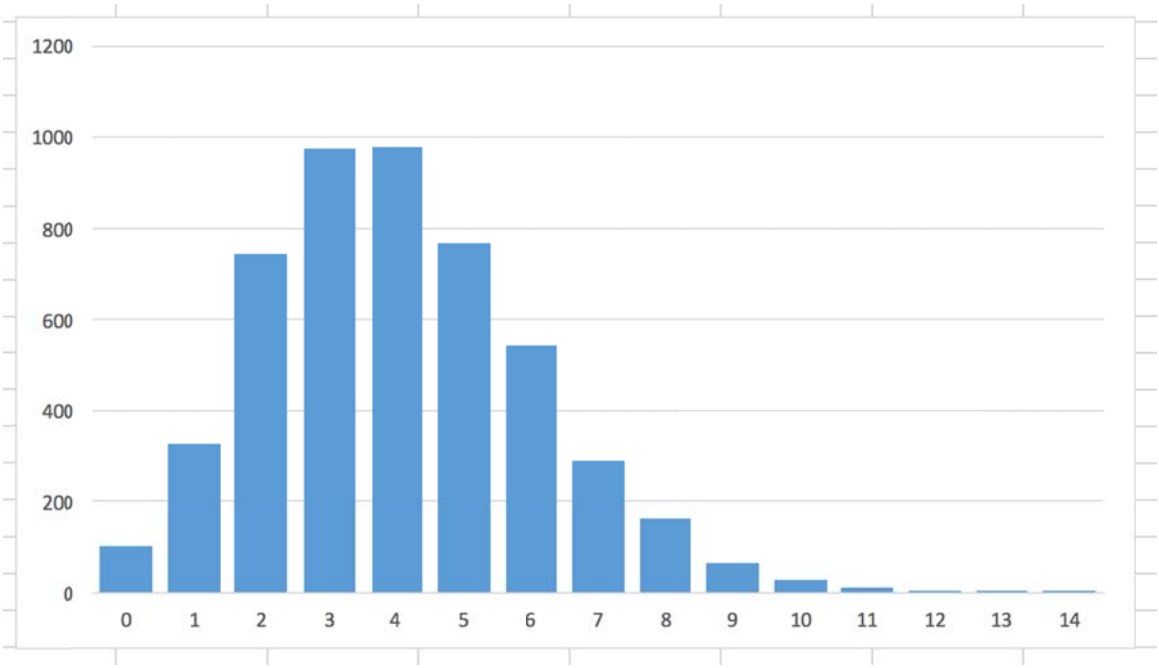
1) Open the file and read the data

A	B
Measurement Log Number	Calls per Hour
1	7
2	5
3	3
4	3
5	5
6	1
7	7
8	5
9	8
10	2
11	6
12	3
13	3
14	7
15	5

2) Define the bins
3) Use the FREQUENCY() function to parse the data

E	F	G
Buckets		
0	=FREQUENCY(B2:B5001,E2:E16)	
1	326	
2	743	
3	977	
4	980	
5	766	
6	542	
7	291	
8	161	
9	64	
10	29	
11	11	
12	5	
13	1	
14	2	

4) Plot the results



5) Calculate, for example, the probability that there will be 6 or less calls covered in an hour (average of the data is 4.03 calls/hr)

P(X>6)	=1-POISSON.DIST(6,4.03, TRUE)

Software Project Plan

A software project manager has asked all of the teams who worked on a completed project to report on the number of hours spent on the tasks specified in the worksheet. The analyst's job is to report on the distribution of the hours as pieces of the total number of hours.

In order to make the most effective report on the data, the we need to perform the following the steps:

Order the data according to the hours spent (descending)
Build a column which calculates the hours as a fraction of the total number of hours
Build a column which accumulates the fraction in the column above
Plot the last two columns with two y axes on the same plot

You will observe the 80/20 rule

Step by Step

1) Open the file and read the data

2) Order the data where the Actual Hours are listed in descending order

A	B
Task Designation	**Actual Hours**
Debug	550
Test Design	433
Mock implementation	76
Test Implementation	45
Project Design	31
Requirements Gathering	21
Test Environment Setup	19
Architecture Assessment	10
Production Enviroment Creation	8
Update meetings	7
Actual Project Hours	**1200**

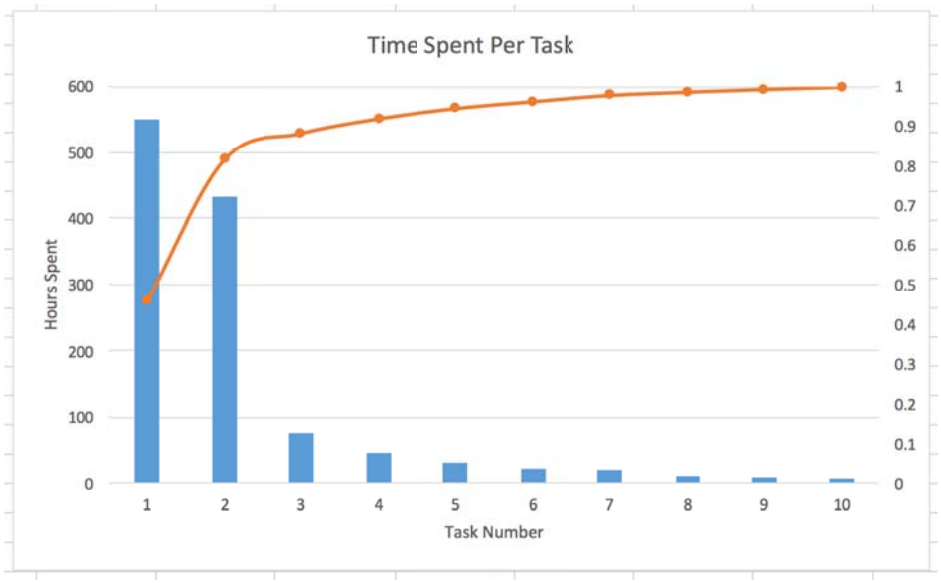
3) Calculate the percentage of the total that each task occupies in Hours as Percent

A	B	C
Task Designation	**Actual Hours**	**Hours as Percent**
Debug	550	0.46
Test Design	433	0.36
Mock implementation	76	0.06
Test Implementation	45	0.04
Project Design	31	0.03
Requirements Gathering	21	0.02
Test Environment Setup	19	0.02
Architecture Assessment	10	0.01
Production Enviroment Creation	8	0.01
Update meetings	7	0.01

4) Add a column called Cumulative and keep a running sum of the Hours as Percent

A	B	C	D
Task Designation	**Actual Hours**	**Hours as Percent**	**Cumulative**
Debug	550	0.46	0.46
Test Design	433	0.36	0.82
Mock implementation	76	0.06	0.88
Test Implementation	45	0.04	0.92
Project Design	31	0.03	0.95
Requirements Gathering	21	0.02	0.96
Test Environment Setup	19	0.02	0.98
Architecture Assessment	10	0.01	0.99
Production Enviroment Creation	8	0.01	0.99
Update meetings	7	0.01	1.00

5) Plot a frequency chart of the Actuals Hours
6) Plot a connected scatterplot of the Cumulative Hours as Percent



Candy Factory

When there is a series of data points available, it is often useful to fit a curve to the data in order to determine if there is a correlation between the variables that are being plotted from the data set. If there is a correlation between the variables that is useful and convincing, then the curve fit can support certain conclusions and have useful, if limited, predictive powers. In this class, we are working with only two variables at a time.

A candy factory has a problem with defects on the factory line. As such, the factory begins a quality improvement initiative. Every two weeks after an initial first week baseline sample, the product is checked for defects and the defects counted and logged. In the interim weeks, the engineering team works on the line to improve the process. The data represents the first 30 weeks of the initiative. The goal is to determine if it is working and to predict the date when zero defects will occur.

The spreadsheet has a simple set of data representing the amount of defects found on the week listed. Plotting the data as a scatter plot shows a possible linear correlation. Fit a curve and display the equation and the square correlation and discuss how the fit show a clear success in the initiative. Perhaps discuss how the improvement may be taking too long depending upon what the unknown constraints may be. The discussion should center around the clear improvement trend, but how other factors may have to be known to determine if the initiative is truly successful from a business perspective.

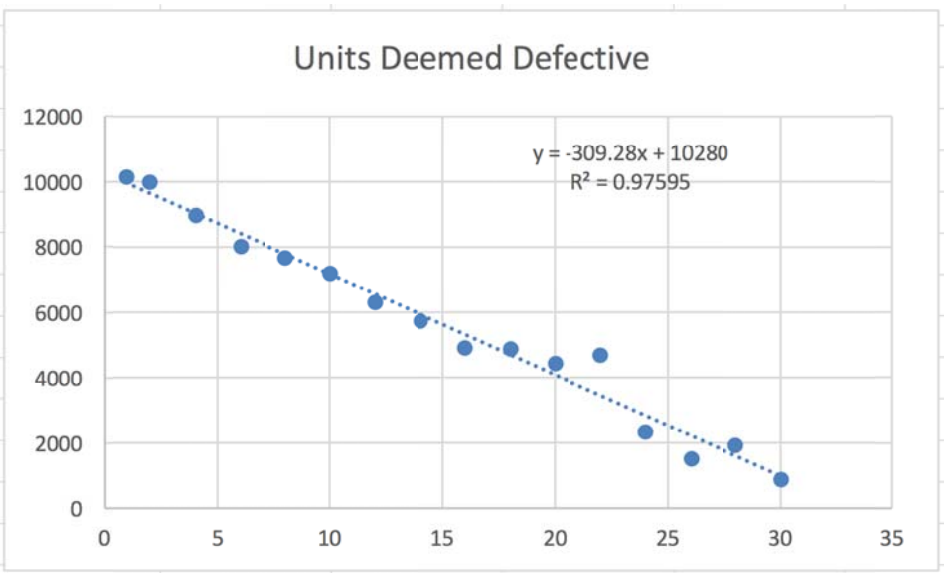
Finally, use the Excel forecasting tool to predict the day at which there will be zero defects. Discuss how this is really an unrealistic goal in the real world of a real factory (not that zero is unrealistic on that day, but how continued zeros is).

Step by Step

1) Open Candy Factory.xlsx and read the data

Week of Sampling	Units Deemed Defective
1	10159
2	10006
4	8990
6	8028
8	7698
10	7209
12	6335
14	5747
16	4932
18	4913
20	4441
22	4720
24	2361
26	1544
28	1952
30	906

2) Plot the data as a scatterplot
3) Fit a trendline to the scatterplot and print the equation of the line as well as the R^2 value



Nurf Foam Factory

This is the same as the candy factory problem but with Nurf Rockets (not be confused with their competitor, Nerf™).

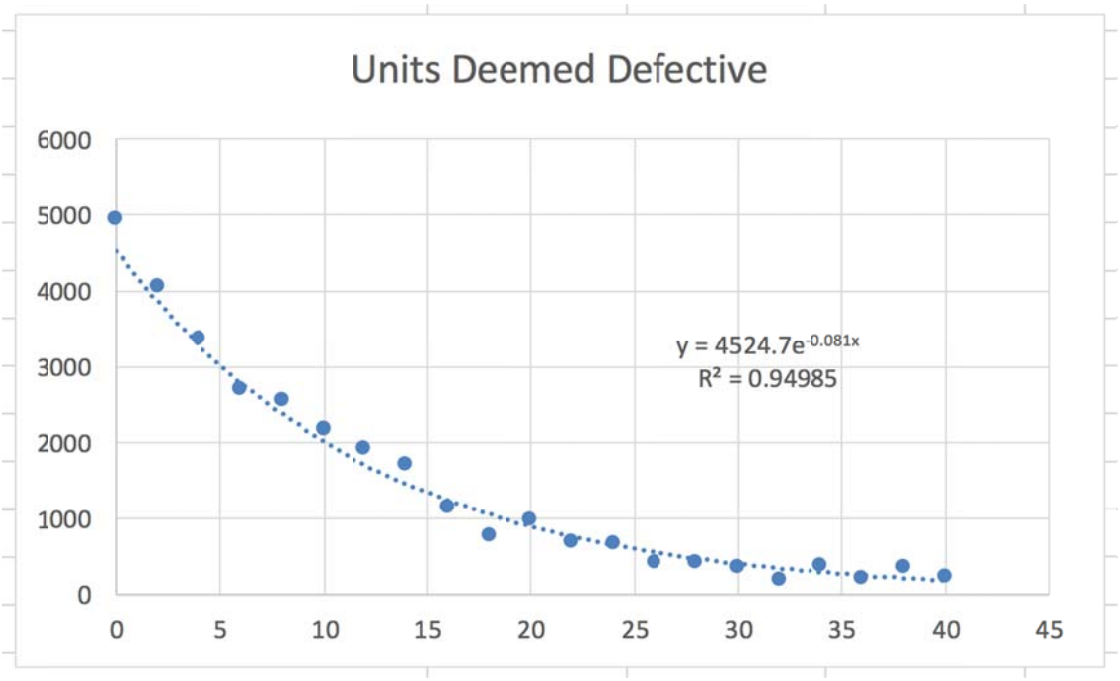
The data in this example is not a linear fit. The fit is an exponential curve. This is more realistic in a factory setting. Initial efforts yield large improvements, then there is smaller and smaller improvement going forward (think Pareto-like). The key point about this fit is that it will never be zero. It will always get closer and closer, but will not be zero. In a real factory, the team must designate a "maximum number" of defects to act as threshold to consider the program a success. Aiming for zero is nice, but after a certain point, the cost to continue the improvement will outweigh the losses it prevents. Those values are of course dependent on the factory etc. so the instructor cannot place true numbers on the thresholds in the Nurf factory.

Step by Step

1) Open the file and read the data

Week of Sampling	Units Deemed Defective
0	4931
2	4067
4	3375
6	2706
8	2556
10	2172
12	1931
14	1705
16	1154

2) Plot the data as a scatterplot
3) Fit a trendline to the scatterplot and print the equation of the line as well as the R^2 value


Units Deemed Defective
6000
5000
4000
3000
2000
1000
0
0
5
10
15
20
25
30
35
40
45
y = 4524.7e-0.081x
R² = 0.94985

Cans of Beer at the Ballpark

The owner of a small local baseball stadium has gathered data over the last few years about the cans of beer sold at the stadium. The data collected is the date, the year, the temperature on game day, and the number of cans that were sold on that date. The owner has requested an analysis that will provide some insight into inventory ordering so that a more "just in time" approach can be taken for ordering. The inventory takes up space when not sold and there are unhappy customers when there are is not enough for game day.

The data should be examined in order of the fields so that data profiling continues to be an activity performed in the course. A pivot table can immediately be built and the Date field vs cans of beer examined. The user should observe that the data as is cannot support year over year analysis because the date field implies 2016 and the actual year is in a separate column. However, logically, it should be pointed out that, excepting holidays, the dates fall on different days each year and is unlikely to have a pattern that can inform us about inventory needs.

The next field should be temperature. When examining temperature and cans of beer sold, the aggregates can be plotted and a linear fit is observed over the range of temperatures available.

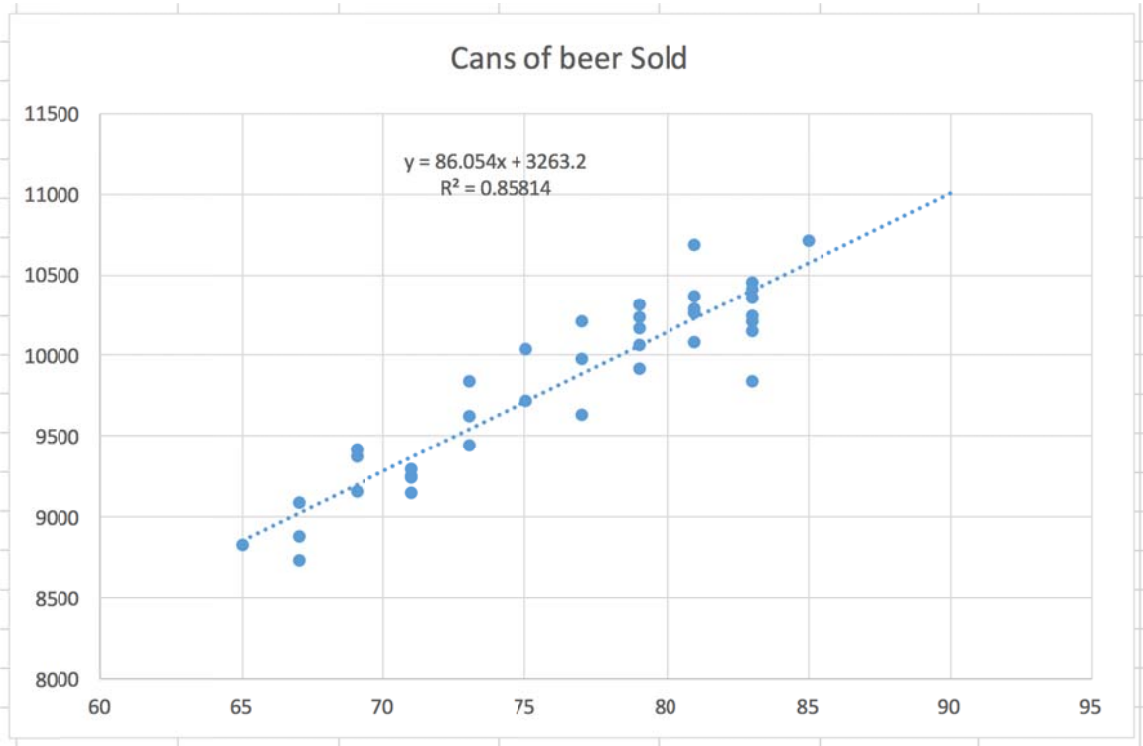
Some things to note: While the model can predict inventory requirements over time (the correlation is Cans of Beer vs Temperature) over the range of temperatures available, what happens when the temperature is 100 degrees or 40 degrees? The temperature can exceed the tolerance of many game attendees which will break the model fit.

Step by Step

1) Open the file and read the data

A	B	C	D
Date	Year	Temperature	Cans of beer Sold
1-Jun	2010	71	9150
20-Jun	2010	81	10084
12-Jul	2010	71	9242
28-Jul	2010	83	10361
3-Aug	2010	65	8829
16-Aug	2010	71	9253
29-Aug	2010	85	10713
2-Sep	2010	81	10689
19-Sep	2010	67	8884
5-Oct	2010	69	9155
1-Jun	2011	81	10369
20-Jun	2011	83	10149

2) Plot the data as a scatterplot
3) Fit a trendline to the scatterplot and print the equation of the line as well as the R^2 value



Customer Retention

The data fit for this data set has several options. The best fit is actually not attainable in Excel. The data is best fit by an exponential of the form 1-exp(-x), which is only approximated by a function of the form of ln(x). Since there a multiple high quality fits, the analyst has to decide how to present the various options.

A new company has launched a software product which has created a new market. The software is a service (SaaS) product which has a rapid adoption early in the life of the product, but the product is experiencing a slowdown in customer subscriptions. The team needs to decide if the next 5 – 10 time periods (the forecast) warrants investment in new server hardware for the influx in customers. The analyst should examine the data and make recommendations for the short term investment.

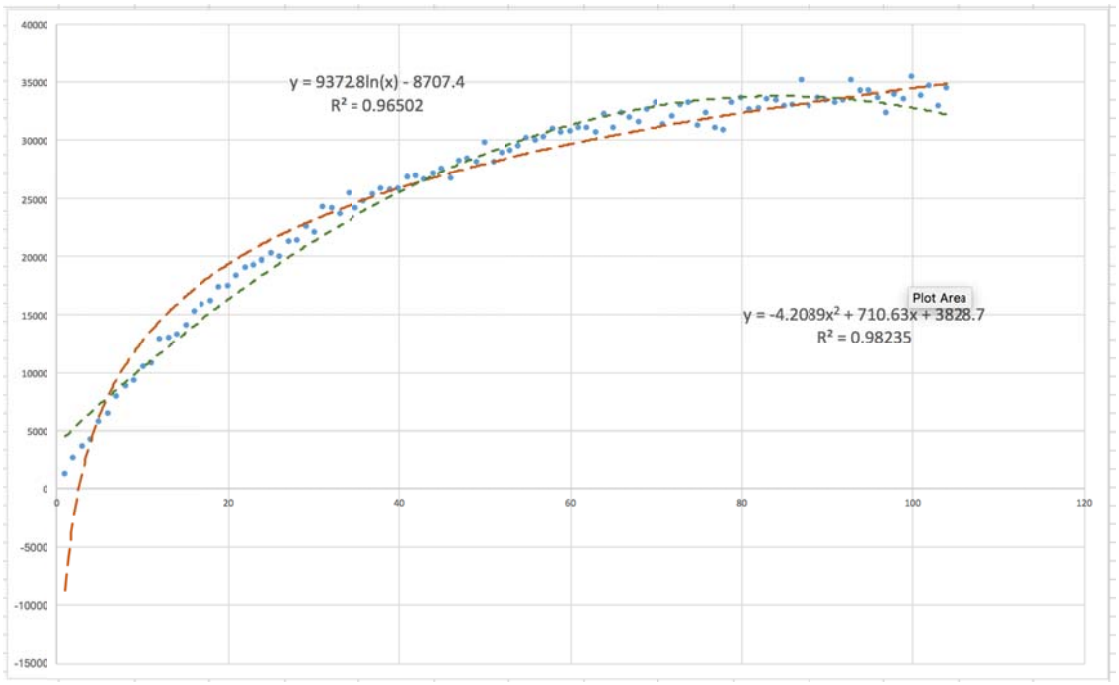
Notice that there are two fits Excel can provide: one in which there is continued growth and one in which there is a downturn. Neither fit models the data particularly well in earlier time periods. What can be seen from the fit in later time periods? Can one say with absolute certainly which of the models is an accurate fit ? These should be points of discussion with the class.

Step by Step

1) Open the file and read the data

Week	Customer Count
1	1266
2	2599
3	3658
4	4225
5	5748
6	6364
7	7846
8	8787
9	9273
10	10448

2) Plot the data as a scatterplot
3) Fit a trendline to the scatterplot and print the equation of the line as well as the R^2 value


y = 9372.8ln(x) - 8707.4
R² = 0.96502
y = -4.2089x² + 710.63x + 3828.7
R² = 0.98235
Plot Area
40000
35000
30000
25000
20000
15000
10000
5000
0
-5000
-10000
-15000
0
20
40
60
80
100
120

Monte Carlo

Monte Carlo is a name given to a class of computational methods designed to extract useful information from complex models (or systems) with probabilistic techniques. Monte Carlo has its roots in the nuclear research program at Los Alamos National Laboratory where early computer simulations were performed in nuclear research.

The key idea is to apply elements of randomness to a complex problem and infer results from the model in some statistical way. To use Monte Carlo, one needs a model, a random number generator, and a way run the simulation (model) with the random numbers as inputs many times in such a way that the results may be analyzed statistically. We will refer to the simulation as "playing the game".

The Monte Carlo examples used in this class consist of three parts. First, we use historical data and perform descriptive analytics to determine the shape of the distribution with which the random numbers must conform. In actuality, this is provided at the beginning of any exercise, but each lab scenario is founded on this idea. Second, the model is given. In some instances, the data distribution alone is the model, in others it is separate. Third, some software package is used to play the game. In the case of Excel, "What-if?" tables are used.

You have a business laying concrete foundations for small offices, homes, patios, etc. You are considering bidding for a new job. The delivery dates are negotiable, but once agreed to, there is a contractual obligation to pay a penalty for every day late on the delivery. You have 10 years of job data in a database and you analyze that data to determine how your job completions have been distributed. The result of the data analysis is that the delivery time for a job this size is 20 days with a standard deviation of 2 days. The median and mode are also 20 days and a frequency plot suggests that the distribution is normal.

You will use this information to create a model which you will use in a Monte Carlo analysis of the likelihood of exceeding a chosen day as our late period. The purpose is to determine what our tolerance may be for being late knowing that competitors' bids will be aggressive and not allow us to choose an arbitrary date which we know we can make.

A second pass at the exercise is to suppose that information has come to you there is a possible 2-day strike coming in the concrete manufacturers' union. Your source estimates an 80% chance of the strike occurring. How do you incorporate this into the model? What is the result of the analysis?

Step by Step

1) Analyze data to define parameters of a model

Average Delivery Time (days)	20
St Dev (sigma)	2

2) Implement parameters in a model

	A	B	C	D	E	F
				Model	=NORM.INV(RAND(), B2, B3)	
	Average Delivery Time (days)	20				
	St Dev (sigma)	2				

3) Create a column enumerating trials

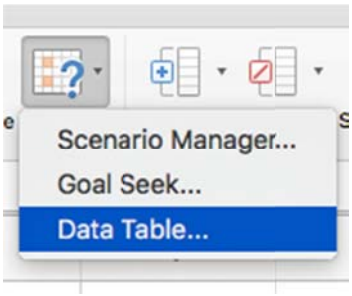
4) Fill one corresponding cell with result of model

# of Game played	
1	=E1
2	
3	
4	
5	
6	
7	
8	
9	
10	

5) Highlight table of trials

# of Game played	
1	22.120263
2	
3	
4	
5	
6	
7	
8	
9	
10	

6) Select Data Table from What If tool



7) Query the model with questions

Number of times greater than 22	=COUNTIF(B7:B16,">22")
% greater than 22	10