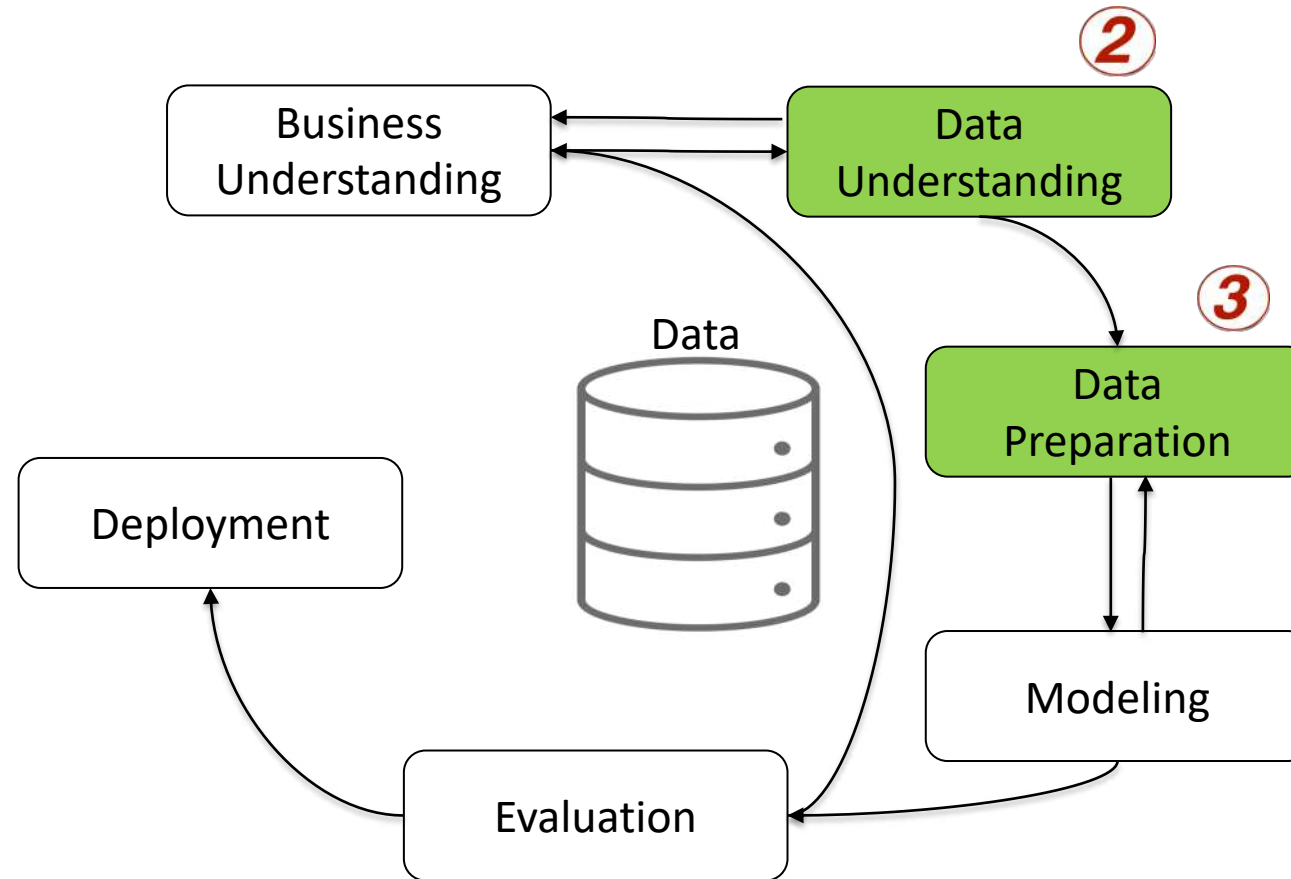# Text Data Mining
For Business Decisions

Module 3

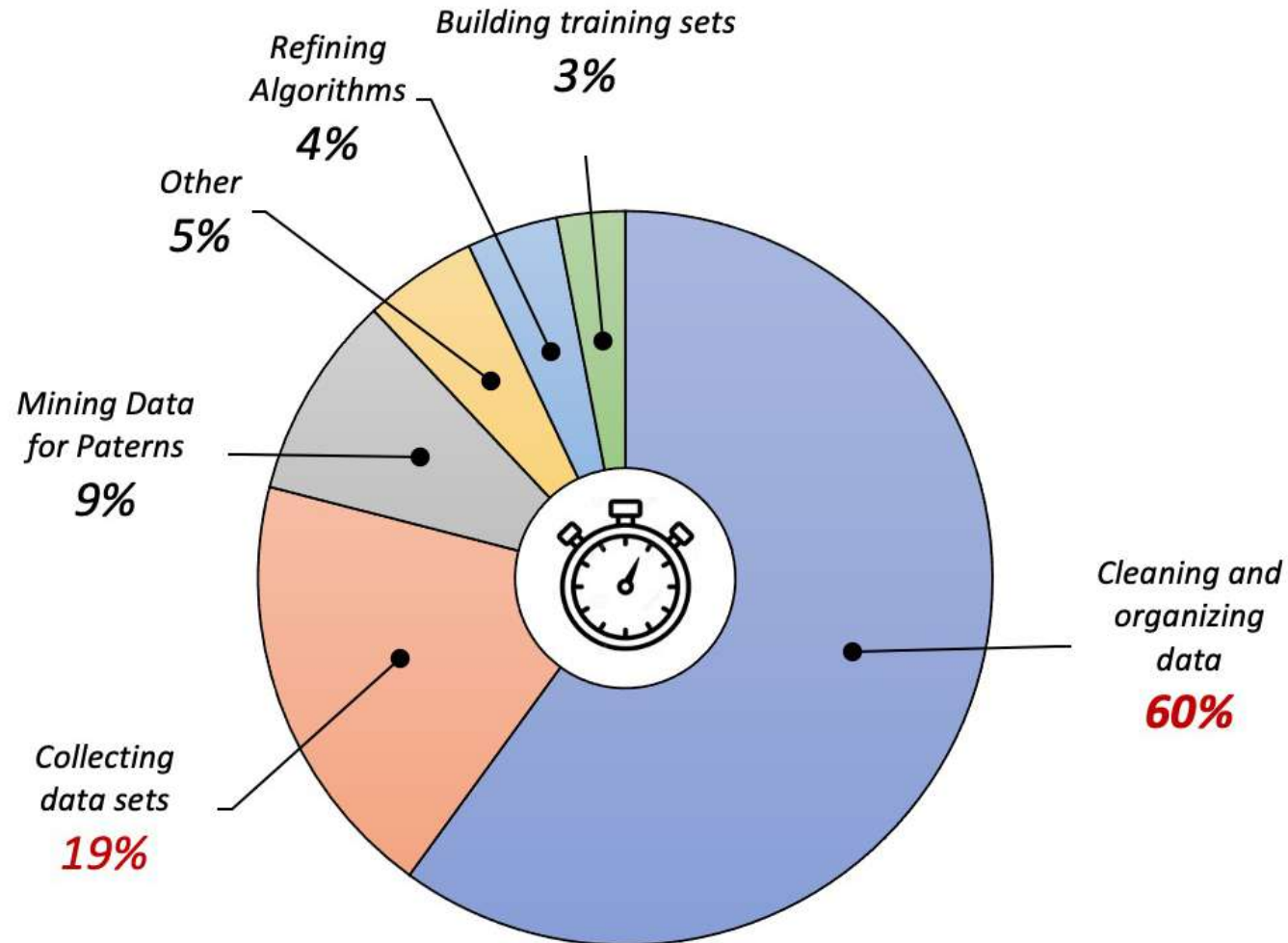Data Preparation

Data Scraping, Cleaning and Annotating

Data Preparation and the CRISP-DM Process

Acquiring, cleaning and organizing data takes up 80% of time of an analyst's time
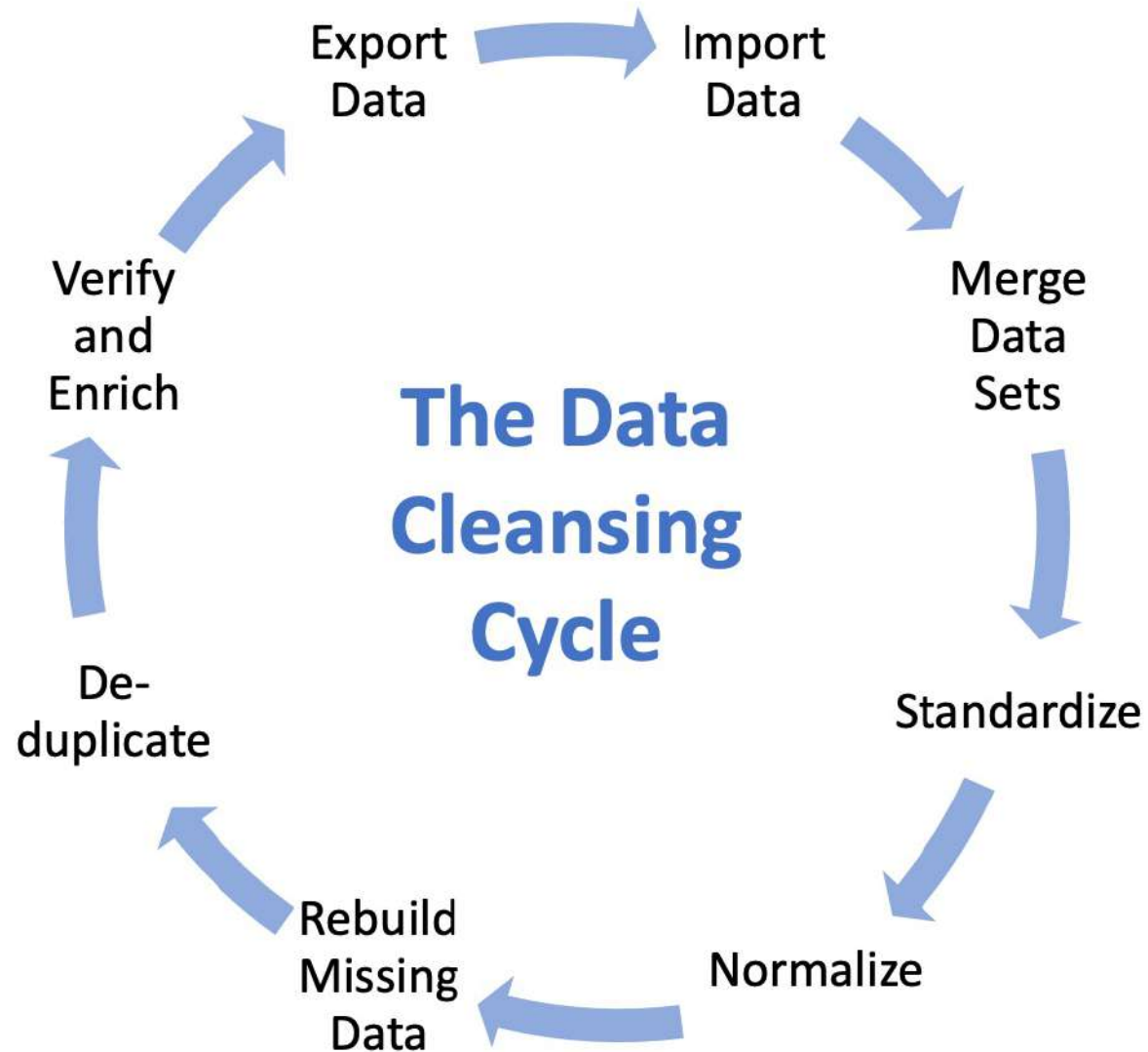
What do Data Scientists Spend Most of their Time Doing?

Building training sets
3%

Refining Algorithms
4%

Other
5%

Mining Data for Paterns
9%

Cleaning and organizing data
60%

Collecting data sets
19%

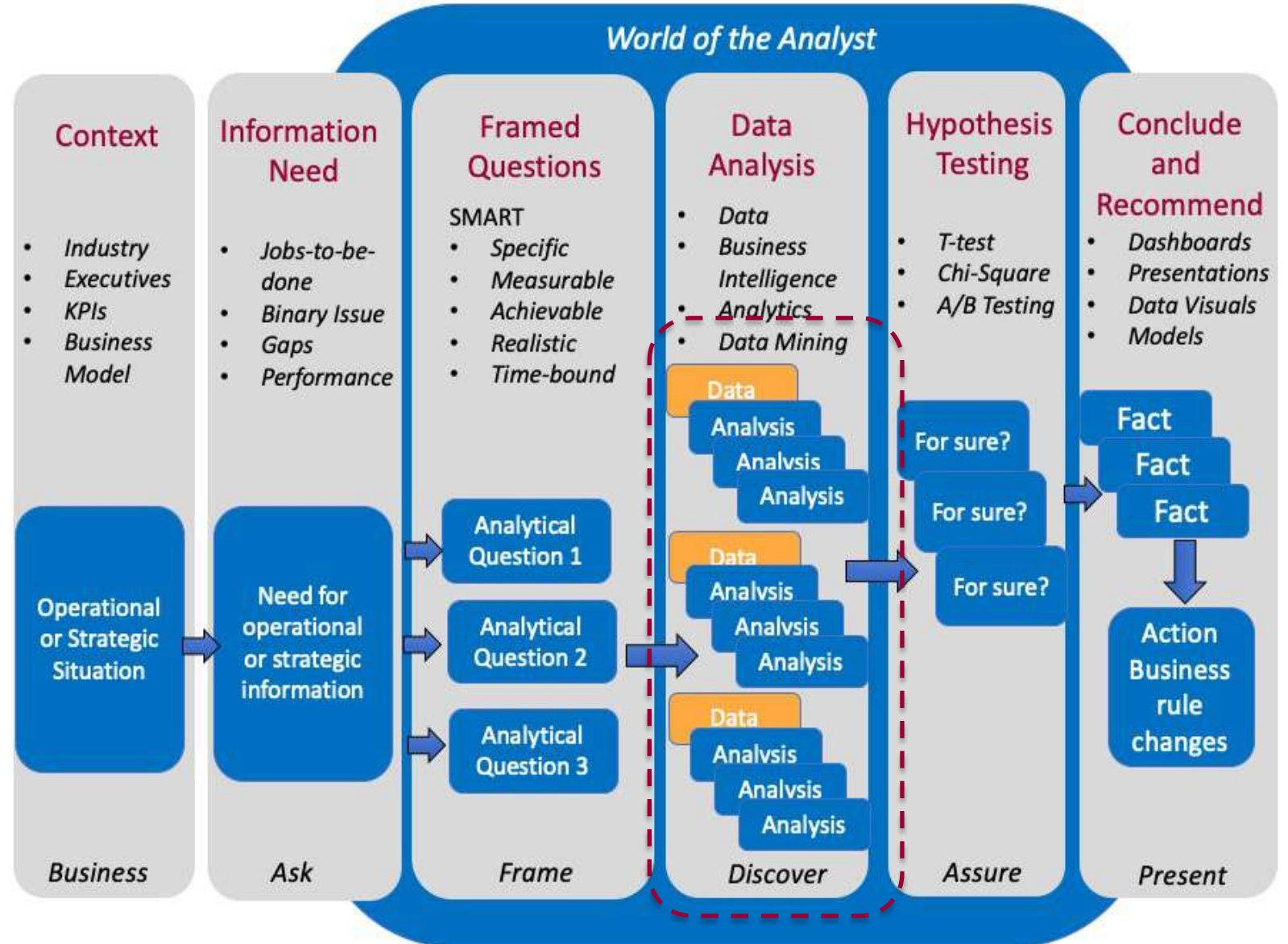***What does this mean?***
The more tools you know and the cleverer you are in shaping a data set the sooner to get to analysis.

FORTINO
Global Education
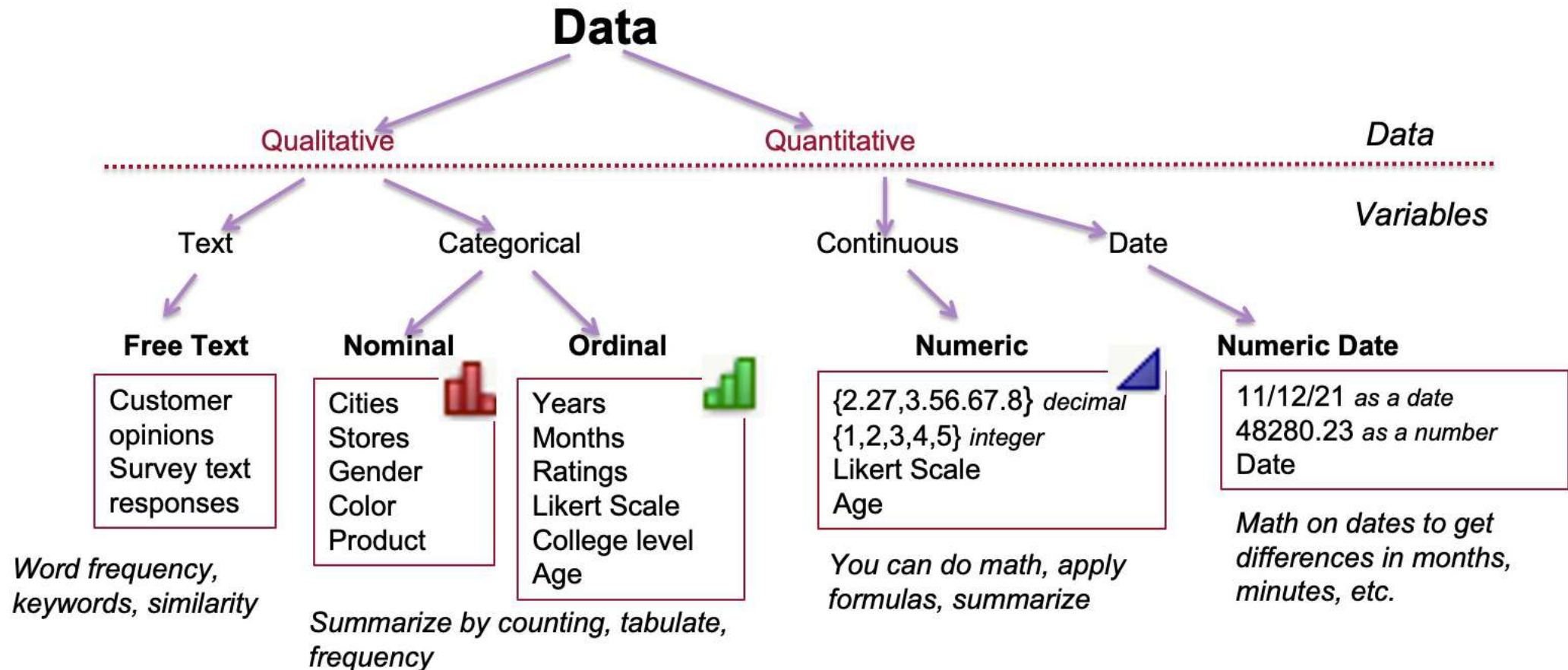
The Data Cleaning Cycle

Where does data preparation fit in the world of the analyst?

World of the Analyst

Context
- Industry
- Executives
- KPIs
- Business Model

Operational or Strategic Situation

Business

Information Need
- Jobs-to-be-done
- Binary Issue
- Gaps
- Performance

Need for operational or strategic information

Ask

Framed Questions
SMART
- Specific
- Measurable
- Achievable
- Realistic
- Time-bound

Analytical Question 1

Analytical Question 2

Analytical Question 3

Frame

Data Analysis
- Data
- Business Intelligence
- Analytics
- Data Mining

Data
Analysis
Analysis
Analysis

Data
Analysis
Analysis
Analysis

Data
Analysis
Analysis
Analysis

Discover

Hypothesis Testing
- T-test
- Chi-Square
- A/B Testing

For sure?

For sure?

For sure?

Assure

Conclude and Recommend
- Dashboards
- Presentations
- Data Visuals
- Models

Fact
Fact
Fact

Action Business rule changes

Present

FORTINO
Global Education

The World of Data and Variables

Two kinds of data, *four* kinds of variables

Data

Qualitative   Quantitative   *Data*

Variables

Text   Categorical   Continuous   Date

Free Text

Customer
opinions
Survey text
responses

*Word frequency,
keywords, similarity*

Nominal

Cities
Stores
Gender
Color
Product

Ordinal

Years
Months
Ratings
Likert Scale
College level
Age

*Summarize by counting, tabulate,
frequency*

Numeric

{2.27,3.56.67.8} *decimal*
{1,2,3,4,5} *integer*
Likert Scale
Age

*You can do math, apply
formulas, summarize*

Numeric Date

11/12/21 *as a date*
48280.23 *as a number*
Date

*Math on dates to get
differences in months,
minutes, etc.*

FORTINO
Global Education

Where are Variables Defined?

- Data tables should come with information about the variables and the nature of the rows of data – it is called *metadata* (data about the data)

- The definitions are in a table or separate document called the *data dictionary*

- It should be consulted when doing data cleaning- it gives the ranges and possible values

Example Data Dictionary

- The data file with survey responses for the San Francisco passenger survey has a typical associated data dictionary
  - https://www.flysfo.com/media/customer-survey-data

- The data dictionary can be found in one of the tabs in the Excel data set

	A	B	C	D	E	F	G
1	Field	Question/Description					
2	RESPNUM	Software added ID number					
3	CCGID	CCG Assigned ID number					
4	RUNID	Run on which survey was conducted					
5	INTDATE	Date of interview					
6	DAY	Day of Interview					
7		1	Sunday				
8		2	Monday				
9		3	Tuesday				
10		4	Wednesday				
11		5	Thursday				
12		6	Friday				
13		7	Saturday				
14	Gate	Gate Number					
15	BAREA	Boarding Area – based on gate as follows:					
16		A	Gates 1-12				
17		B	Gates 20-39				
18		C	Gates 40-48				
19		D	Gates 50-59				
20		E	Gates 60-69				
21		F	Gates 70-90				
22		G	Gates 91-102				
23	STRATA	Strata					
24		1	AM (Flights departing before 11 am)				
25		2	MID (Flights departing 11 am to 5 pm)				
26		3	PM (Flights departing after 5 pm)				
27	PEAK	Peak					
28		1	Domestic peak – domestic flights departing 8 am to 1 pm				
29		2	Domestic offpeak – domestic flights departing before 8 am or after 1 pm				
30		3	International flights				

Digital Encoding of Data

- Words and number storage in a digital computer
- Digital computer stores and manipulate data as 0's and 1's
  - 0011010010001110011110010100010000100
- Must encode *language* and expressions into sequence of zeros and ones
- Digital Codes
  - What are some early examples of encoding of *language* for storage and transmission?
  - https://www.w3.org/International/questions/qa-what-is-encoding

FORTINO
Global Education

Character Encoding

- If you use anything other than the most basic English text, people may not be able to read the content you create unless you declare what character encoding you used.

- For example, you may intend the text to look like this:

Author: Guðrún Guðmundsdóttir. Title: Introduction to character encoding (文字符号化入門). Copyright © 2004-2007 W3C® (MIT, ERCIM, Keio).

- but it may actually display like this:

Author: GuÃ°rÃºn GuÃ°mundsdÃ³ttir. Title: Introduction to character encoding (æ–‡å—ç¬¦å·åŒ–å…¥é–€). Copyright Â© 2004-2007 W3CÂ® (MIT, ERCIM, Keio).

Character Encoding Schemes

- Words and sentences in text are created from *characters*. Examples of characters include the Latin letter á or the Chinese ideograph 請 or the Devanagari character ह.

- Characters that are needed for a specific purpose are grouped into a *character set* .

- To refer to characters in an unambiguous way, each character is associated with a number, called a *code point*.
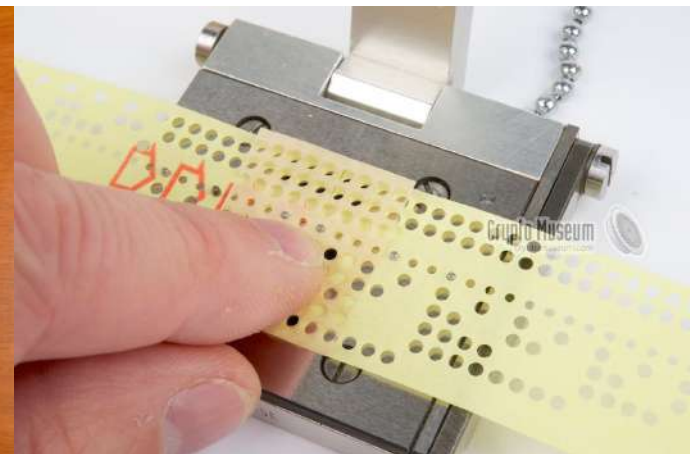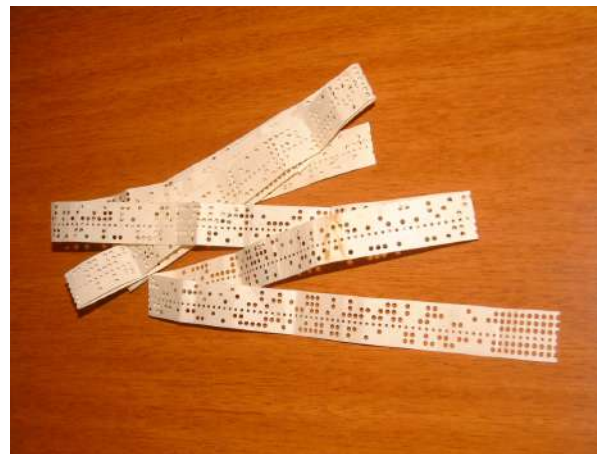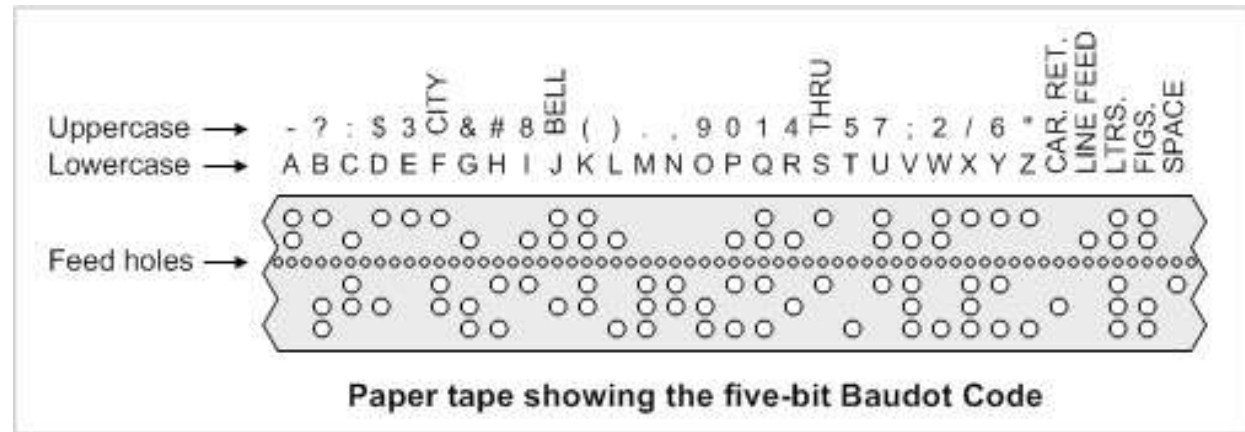
- The characters are stored in the computer as one or more *byte*

FORTINO
Global Education

Early Form of Encoding: Morse Code

- **Morse code** is a method of transmitting text information as a series of on-off tones, lights, or clicks that can be directly understood by a skilled listener or observer without special equipment.

International Morse Code

1. The length of a dot is one unit.
2. A dash is three units.
3. The space between parts of the same letter is one unit.
4. The space between letters is three units.
5. The space between words is seven units.

A more efficient code for transmission: The Baudot Code

- 5-character code for encoding of the characters in the English language

Paper tape showing the five-bit Baudot Code

FORTINO
Global Education

A modern Code for the Computer Era: ASCII-7

- 7-bit code for encoding of English characters and control codes

- What's the 7-bit sequence for they letter A?

- 01000001

b7 b6 b5				Column	0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1
b4	b3	b2	b1	Row	0	1	2	3	4	5	6	7
0	0	0	0	0	NUL	DLE	SP	0	@	P	`	p
0	0	0	1	1	SOH	DC1	!	1	A	Q	a	q
0	0	1	0	2	STX	DC2	"	2	B	R	b	r
0	0	1	1	3	ETX	DC3	#	3	C	S	c	s
0	1	0	0	4	EOT	DC4	$	4	D	T	d	t
0	1	0	1	5	ENQ	NAK	%	5	E	U	e	u
0	1	1	0	6	ACK	SYN	&	6	F	V	f	v
0	1	1	1	7	BEL	ETB	'	7	G	W	g	w
1	0	0	0	8	BS	CAN	(	8	H	X	h	x
1	0	0	1	9	HT	EM	)	9	I	Y	i	y
1	0	1	0	10	LF	SUB	*	:	J	Z	j	z
1	0	1	1	11	VT	ESC	+	;	K	[	k	{
1	1	0	0	12	FF	FS	,	<	L	\	l		
1	1	0	1	13	CR	GS	–	=	M	]	m	}
1	1	1	0	14	SO	RS	.	>	N	^	n	~

Codes Compared

Message	S	T	A	R	T	3	#
Morse	● ● ●	▬	● ▬	● ▬ ●	▬	● ● ● ▬	
Baudot							
ASCII-7	01010011	01010100	01000001	01010010	01010100	00110011	00100011
UFT-8	01010011	01010100	01000001	01010010	01010100	00110011	00100011

FORTINO
Global Education

Unicode UTF-8

- **UTF-8** is a [variable-width](#) [character encoding](#) used for electronic communication.

- Defined by the Unicode Standard, the name is derived from *Unicode* (or *Universal Coded Character Set*) *Transformation Format – 8-bit*.

- It can encode over 1 million unique characters

- The most common set uses 4 bytes per character

- As an analyst, you should nowadays always [choose the UTF-8 character encoding](#) for your content or data. This [Unicode](#) encoding is a good choice because you can use a single character encoding to handle any character you are likely to need. This greatly simplifies things.

UTF-8 Encoding

- UTF-8 uses 1 byte to represent characters in the ASCII set, two bytes for characters in several more alphabetic blocks, and three bytes for the rest of the BMP. Supplementary characters use 4 bytes.

	A	ℵ	好	不
Code point	U+0041	U+05D0	U+597D	U+233B4
UTF-8	41	D7 90	E5 A5 BD	F0 A3 8E B4
UTF-16	00 41	05 D0	59 7D	D8 4C DF B4
UTF-32	00 00 00 41	00 00 05 D0	00 00 59 7D	00 02 33 B4

Data Types

- Characters or character data
  - "A", "B", "C"…"a', "b", "c"… "%", "#", "@"…"+", "/", "="…
- Numeric
  - Numeric data is defined as data which consists of digits as opposed to letters of the alphabet of special characters.
  - Integer 1,2,3,4 or 34590 or 12863456223
  - Decimal 4.56 or 3.1417 or 17032.45002
  - Floating point 3.78E12

Data Storage
Organization Protocols

- A way to organize what the stored data means so computer programs can interpret them properly
- Meta data about the stored characters – the character set
- What does each character mean? Context!
- When is a "2" a 2?
  - a *character* as in "21 Jump Street"
  - An *integer* as in 2 items
  - Or part of a *decimal* as in 8.2 ounces
  - Or part of a *floating-point number* as in the circumference of the earth is 2.49E3 miles
  - A *code* (a category, a label) as in 1=male, 2= female

FORTINO
Global Education

Some Data Storage/Organization Protocols

- File structures
  - <mark>CSV</mark>
  - JSON
  - XML and HTML
  - <mark>XLS</mark>, XLXS
- Usually identified by the file extension:
  - <mark>.csv</mark>, .json, .xml, .xsl, <mark>xlxs</mark>

FORTINO
Global Education

Comma Separated Values

- A comma-separated values (CSV) file stores tabular data (numbers and text) in plain-text form

- Plain text means that the file is a sequence of characters, with no data that has to be interpreted instead, as binary numbers.

- A CSV file consists of any number of records, separated by line breaks of some kind; each record consists of fields, separated by some other character or string, most commonly a literal comma or tab.

- Usually, all records have an identical sequence of fields.

Delimiter Separated Values

- Formats that use delimiter-separated values (also DSV) store two-dimensional arrays of data by separating the values in each row with specific delimiter characters.

- Most database and spreadsheet programs are able to read or save data in a delimited format.

- Any character may be used to separate the values, but the most common delimiters are
  - comma (CSV)
  - tab
  - colon
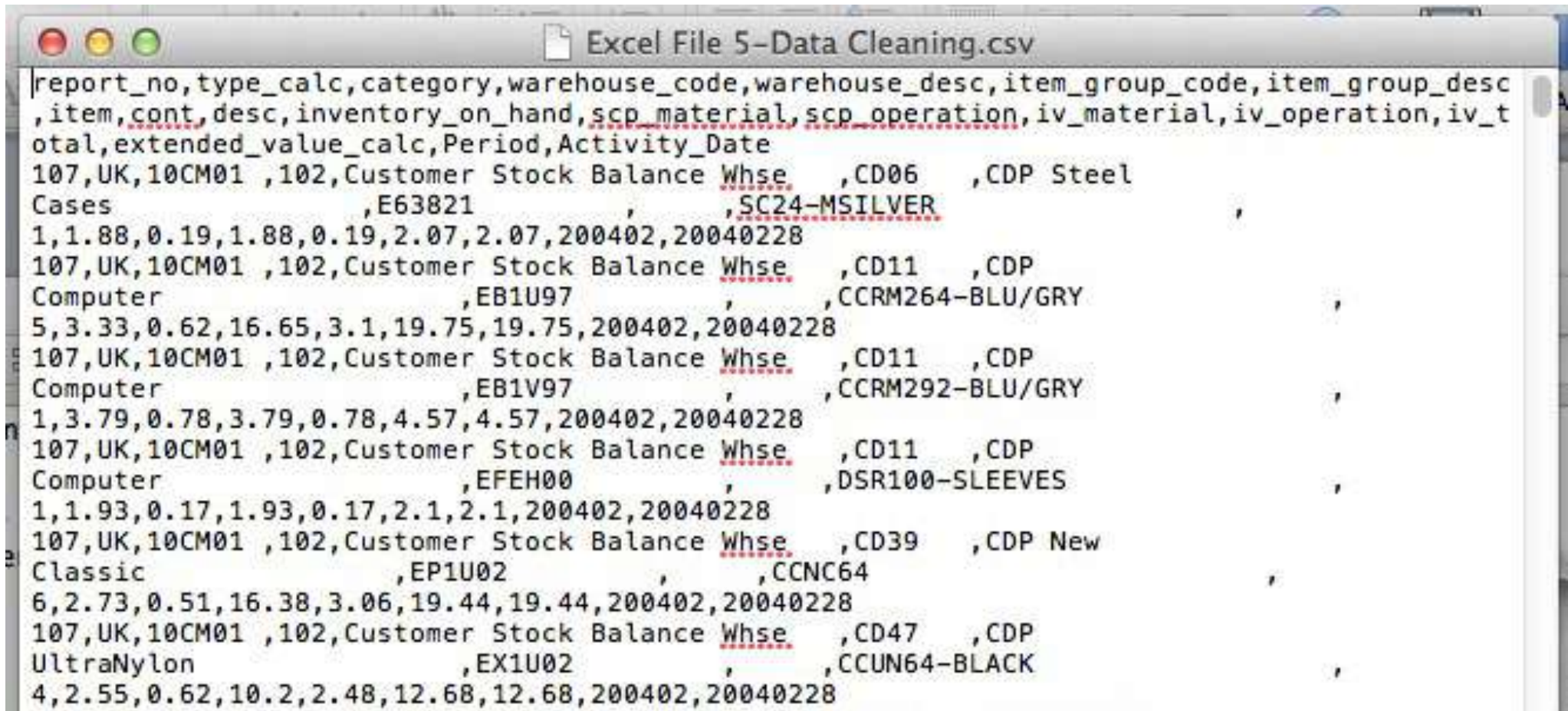
Comma Separated Values

- In a comma-separated values (CSV) file the data items are separated using commas as a delimiter

- Column headers are sometimes included as the first line, and each subsequent line is a row of data.

- The lines are separated by newlines.

- Example

```
"Date","Pupil","Grade"
"25 May","Bloggs, Fred","C"
"25 May","Doe, Jane","B"
"15 July","Bloggs, Fred","A"
"15 April","Muniz, Alvin ""Hank""","A"
```

Wrappers

- Note the use of the double quote to enclose each field.
- This prevents the comma in the actual field value from being interpreted as a field separator.
- This necessitates a way to "escape" the field wrapper itself, in this case the double quote; it is customary to double the double quotes actually contained in a field as with those surrounding "Hank".
- In this way, any ASCII text including newlines can be contained in a field.

CSV Example


```
Excel File 5-Data Cleaning.csv

report_no,type_calc,category,warehouse_code,warehouse_desc,item_group_code,item_group_desc
,item,cont,desc,inventory_on_hand,scp_material,scp_operation,iv_material,iv_operation,iv_t
otal,extended_value_calc,Period,Activity_Date
107,UK,10CM01 ,102,Customer Stock Balance Whse    ,CD06    ,CDP Steel
Cases                    ,E63821          ,        ,SC24-MSILVER                      ,
1,1.88,0.19,1.88,0.19,2.07,2.07,200402,20040228
107,UK,10CM01 ,102,Customer Stock Balance Whse    ,CD11    ,CDP
Computer                 ,EB1U97          ,        ,CCRM264-BLU/GRY              ,
5,3.33,0.62,16.65,3.1,19.75,19.75,200402,20040228
107,UK,10CM01 ,102,Customer Stock Balance Whse    ,CD11    ,CDP
Computer                 ,EB1V97          ,        ,CCRM292-BLU/GRY              ,
1,3.79,0.78,3.79,0.78,4.57,4.57,200402,20040228
107,UK,10CM01 ,102,Customer Stock Balance Whse    ,CD11    ,CDP
Computer                 ,EFEH00          ,        ,DSR100-SLEEVES               ,
1,1.93,0.17,1.93,0.17,2.1,2.1,200402,20040228
107,UK,10CM01 ,102,Customer Stock Balance Whse    ,CD39    ,CDP New
Classic                  ,EP1U02          ,        ,CCNC64                        ,
6,2.73,0.51,16.38,3.06,19.44,19.44,200402,20040228
107,UK,10CM01 ,102,Customer Stock Balance Whse    ,CD47    ,CDP
UltraNylon               ,EX1U02          ,        ,CCUN64-BLACK                 ,
4,2.55,0.62,10.2,2.48,12.68,12.68,200402,20040228
```

Format of Files for Easy Analysis

- Most analysis programs require that the data file be in a simple tabular form
- The columns are variables with the name of the variable at the top of the column
- The rows are the values of the variables for each instance of the data set
- The rows in the table are all about the same thing
  - We don't mix information about the characteristics of cars for sale in the lot with detailed customer information – one is an inventory file, the other a customer information file.
- We shape tables of information to have this structure – called a flat file or "flat file format"

Case Study: Flat File Format

Each column is a variable, a "feature", of the data set, or attributes of the population

Each row is an example, or "instantiation" of the population

	A	B	C	D	E
1	name	age	sex	height	weight
2	KATIE	12	F	59	95
3	LOUISE	12	F	61	123
4	JANE	12	F	55	74
5	JACLYN	12	F	66	145
6	LILLIE	12	F	52	64
7	TIM	12	M	60	84
8	JAMES	12	M	61	128
9	ROBERT	12	M	51	79
10	BARBARA	13	F	60	112
11	ALICE	13	F	61	107
12	SUSAN	13	F	56	67
13	JOHN	13	M	65	98

Flat File Format

No empty columns

It is good form to have first cell in A1 position

First row is always the variable names and each column **must** have a name

Do not mix population rows, all rows are from the same population

Nothing to the right of the last column

No completely empty rows

Not necessary but it is good form to remove all unnecessary formatting of the cells (coloring, bolding, borders, italics)

Nothing below the last row

	A	B	C	D	E	F
1	name	age	sex	height	weight	
2	KATIE	12	F	59	95	
3	LOUISE	12	F	61	123	
4	JANE	12	F	55	74	
5	JACLYN	12	F	66	145	
6	LILLIE	12	F	52	64	
7	TIM	12	M	60	84	
8	JAMES	12	M	61	128	
9	ROBERT	12	M	51	79	
10	BARBARA	13	F	60	112	
11	ALICE	13	F	61	107	
12	SUSAN	13	F	56	67	
13	JOHN	13	M	65	98	
14	JOE	13	M	63	105	
15	MICHAEL	13	M	58	95	
16	DAVID	13	M	59	79	
17	JUDY	14	F	61	81	
18	ELIZABETH	14	F	62	91	
19	LESLIE	14	F	65	142	
20	CAROL	14	F	63	84	
21	PATTY	14	F	62	85	
22	FREDRICK	14	M	63	93	
23	ALFRED	14	M	64	99	
24	HENRY	14	M	65	119	
25	LEWIS	14	M	64	92	
26	EDWARD	14	M	68	112	
27	CHRIS	14	M	64	99	
28	JEFFERY	14	M	69	113	
29	MARY	15	F	62	92	
30	AMY	15	F	64	112	
31	ROBERT	15	M	67	128	
32	WILLIAM	15	M	65	111	
33	CLAY	15	M	66	105	
34	MARK	15	M	62	104	
35	DANNY	15	M	66	106	
36	MARTHA	16	F	65	112	
37	MARIAN	16	F	60	115	
38	PHILLIP	16	M	68	128	
39	LINDA	17	F	62	116	
40	KIRK	17	M	68	134	
41	LAWRENCE	17	M	70	172	
42						
43						

Case Study: Is this a Flat File?

- *Why not?*
- *How would you fix it?*

	A	B	C	D	E	F
1	PIZZA	BAKERY	SHOES	GIFTS	PETS	
2	80	150	48	100	25	
3	125	40	35	96	80	
4	35	120	95	35	30	
5	58	75	45	99	35	
6	110	160	75	75	30	
7	140	60	115	150	28	
8	97	45	42	45	20	
9	50	100	78	100	75	
10	65	86	65	120	48	
11	79	87	125	50	20	
12	35	90			50	
13	85				75	
14	120				55	
15					60	
16					85	
17					110	
18						
19	**Business Startup Costs**					
20						
21	The following data represent business startup costs (thousands of dollars) for shops.					
22	PIZZA = startup costs for pizza					
23	BAKERY = startup costs for baker/donuts					
24	SHOES = startup costs for shoe stores					
25	GIFTS = startup costs for gift shops					
26	PETS = startup costs for pet stores					
27	Reference: *Business Opportunities Handbook*					
28						

FORTINO
Global Education

Case Study: Is this a Flat File?

- *Why not?*
- *How would you fix it?*

	A	B	C	D	E	F	G	H	I	J	K	L
1	PURCHASED	BIRTHDATE	SEX	ISSUEDATE	ID	Phone	NAME	Start Dat	Gender	BIRTHDATE	Sales	
2	52750	08.24.05	F	01.01.05	10	3978	CHRISTINE I HAAS	01.01.05	F	08.24.05	52750	
3	41250	02.02.88	M	02.12.05	20	3476	MICHAEL L THOMPSON	02.12.05	M	02.02.88	41250	
4	19180	11.12.79	M	02.23.05	30	961	DANIEL S SMITH	02.23.05	M	11.12.79	19180	
5	38250	05.11.81	F	03.03.05	50	4738	SALLY A KWAN	03.03.05	F	05.11.81	38250	
6	40175	09.15.65	M	03.24.05	60	6789	JOHN B GEYER	03.24.05	M	09.15.65	40175	
7	32250	07.07.85	M	04.05.05	70	6423	EVA D PULASKI	04.05.05	M	07.07.85	32250	
8	36170	05.26.53	F	04.11.05	90	7831	EILEEN W HENDERSON	04.11.05	F	05.26.53	36170	
9	29750	05.15.81	F	05.05.05	100	5498	THEODORE Q SPENSER	05.05.05	F	05.15.81	29750	
10	26150	12.18.56	M	05.16.05	110	972	VINCENZO G LUCCHESSI	05.16.05	M	12.18.56	26150	
11	46500	11.05.69	M	05.30.05	120	3490	SEAN O'CONNELL	05.30.05	M	11.05.69	46500	
12	29250	10.18.82	M	06.19.05	130	4501	DAVID BROWN	06.19.05	M	10.18.82	29250	
13	27740	05.29.81	M	06.19.05	140	4578	DOLORES M QUINTANA	06.19.05	M	05.29.81	27740	
14	23800	09.15.65	F	07.07.05	150	1793	HEATHER A NICHOLLS	07.07.05	F	09.15.65	23800	
15	28420	01.19.86	F	07.07.05	160	4510	BRUCE ADAMSON	07.07.05	F	01.19.86	28420	
16	25280	05.17.87	M	07.26.05	170	3782	ELIZABETH R PIANKA	07.26.05	M	05.17.87	25280	
17	22250	14/24/55	F	07.28.05	170	2890	MASATOSHI J YOSHIMURA	07.28.05	F	14/24/55	22250	
18	24680	01.05.51	M	08.07.05	190	9001	MARIA L PERES	08.07.05	M	01.05.51	24680	
19	27380	05.26.53	F	08.15.05	200	1682	MARILYN S SCOUTTEN	08.15.05	F	05.26.53	27380	
20	21340	02.21.89	F	08.17.05	210	2986	JAMES H WALKER	08.17.05	F	02.21.89	21340	
21	20450	06.25.52	M	08.29.05	220	4501	DAVID BROWN	08.29.05	M	06.25.52	20450	
22	27740	05.29.81	M	09.11.05	230	942	WILLIAM T JONES	09.11.05	M	05.29.81	27740	
23	18270	02.23.53	M	09.12.05	240	672	JENNIFER K LUTZ	09.12.05	M	02.23.53	18270	
24	29840	03.19.00	F	09.14.05	250	2094	JAMES J JEFFERSON	09.14.05	F	03.19.00	29840	
25	22180	05.30.75	M	09.15.05	260	3780	SALVATORE M MARINO	09.15.05	M	05.30.75	22180	
26	28760	03.31.54	M	09.15.05	270	961	DANIEL S SMITH	09.15.05	M	03.31.54	28760	
27	19180	11.12.79	M	09.30.05	280	8953	SYBIL P JOHNSON	09.30.05	M	11.12.79	19180	
28	17250	10.05.76	F	09.30.05	290	9001	MARIA L PEREZ	09.30.05	F	10.05.76	17250	
29	27380	05.26.53	F	09.30.05	300	8997	ETHEL R SCHNEIDER	09.30.05	F	05.26.53	27380	
30	26250	03.28.76	F	09.30.05	310	4502	JOHN R PARKER	09.30.05	F	03.28.76	26250	
31	15340	07.09.86	M	10.10.05	320	2095	PHILIP X SMITH	10.10.05	M	07.09.86	15340	
32	17750	10.27.76	H	10.11.05	330	3332	MAUDE F SETRIGHT	10.11.05	H	10.27.76	17750	
33	15900	04.21.71	F	10.30.05	340	9990	RAMLAL V MEHTA	10.30.05	F	04.21.71	15900	
34	19950	08.11.72	M	11.21.05	360	2103	WING LEE	11.21.05	M	08.11.72	19950	
35	25370000	07.18.81	M	12.05.05	380	5698	JASON R GOUNOT	12.05.05	M	07.18.81	25370000	
36	23840	05.17.66	M	12.05.05	400	5698		12.05.05	M	05.17.66	23840	
37							CUSTNO					
38												
39												

Rubbermaid Flat File With Text Data

	A	B	C	D	E	F
1	brand	product	review.date	rating	reviews.text	review.title
2	Rubbermaid	Rubbermaid174 Reveal Spray Mop	11/19/16	5	I really love the mop. Only concern is the nozzl	Used Many Mops This Is Quit
3	Rubbermaid	Rubbermaid174 Reveal Spray Mop	7/2/10	5	I am a handicapped man, with limited use of m	The Perfect Mop
4	Rubbermaid	Rubbermaid174 Reveal Spray Mop	3/20/15	1	I would give it 0 star if I could. This is my secon	Rubbermaid reveal mop
5	Rubbermaid	Rubbermaid174 Reveal Spray Mop	12/30/13	5	This mop is perfect! It's thin so it gets under cr	You Must Buy This Mop!
6	Rubbermaid	Rubbermaid174 Reveal Spray Mop	7/19/17	4	I really like this mop. It does a great job in my l	I Like It But...
7	Rubbermaid	Rubbermaid174 Reveal Spray Mop	1/16/17	2	I purchased the Reveal Spray Mop about year a	Very disappointed!
8	Rubbermaid	Rubbermaid174 Reveal Spray Mop	2/1/15	2	I came across the warning that this mop had a	Defective Spray Trigger
9	Rubbermaid	Rubbermaid174 Reveal Spray Mop	5/12/11	1	I hate this mop! I bought it to replace a Swiffer	does not get floors clean
10	Rubbermaid	Rubbermaid174 Reveal Spray Mop	9/26/11	3	A great idea and worked well for about 6 mont	reveal spray mop
11	Rubbermaid	Rubbermaid174 Reveal Spray Mop	8/20/16	1	This is my second Rubbermaid spray mop and	the bottle leaks
12	Rubbermaid	Rubbermaid174 Reveal Spray Mop	10/22/13	3	I saw this at someone's house and loved it, so	Reveal Spray Mop
13	Rubbermaid	Rubbermaid174 Reveal Spray Mop	3/4/14	1	Oddly enough I wrote Rubbermaid on line toda	Reveal Spray Mop
14	Rubbermaid	Rubbermaid174 Reveal Spray Mop	11/23/14	4	I've had my Reveal mop for a couple of years a	Years Of Good Service
15	Rubbermaid	Rubbermaid174 Reveal Spray Mop	2/19/14	4	I bought this mop about a month ago and I use	Very green
16	Rubbermaid	Rubbermaid174 Reveal Spray Mop	1/22/15	1	I have purchased 3 of these mops and they hav	Poor Quality
17	Rubbermaid	Rubbermaid174 Reveal Spray Mop	7/16/16	3	I loved my reveal but the trigger kept getting s	spray trigger
18	Rubbermaid	Rubbermaid174 Reveal Spray Mop	3/12/15	5	I bought this mop about 8 - 9 months ago. I wa	Loved It When It Worked.
19	Rubbermaid	Rubbermaid174 Reveal Spray Mop	9/22/10	5	I love this mop. I teach Home Economics at RM	Reveal Spray Mop
20	Rubbermaid	Rubbermaid174 Reveal Spray Mop	8/7/12	1	I have purched 3 of the reveal spray mops (one	Ro
21	Rubbermaid	Rubbermaid174 Reveal Spray Mop	7/31/16	3	I bought this mop after my sister raved about I	Great Idea But Problems With
22	Rubbermaid	Rubbermaid174 Reveal Spray Mop	9/5/14	2	I have had a reveal mop for a while now and al	Frustrating And Disappointing
23	Rubbermaid	Rubbermaid174 Reveal Spray Mop	1/1/14	1	After about 6 months the sprayer only worked	Hate this product
24	Rubbermaid	Rubbermaid174 Reveal Spray Mop	3/8/13	2	Rubbermaid had a great idea when they came	Not impressed
25	Rubbermaid	Rubbermaid174 Reveal Spray Mop	11/1/14	1	I have owned two of these mops now and both	Two Mops Purchased In The L
26	Rubbermaid	Rubbermaid174 Reveal Spray Mop	12/27/10	5	Bought this mop to replace a rag mop. Absolut	Love this mop
27	Rubbermaid	Rubbermaid174 Reveal Spray Mop	4/15/14	1	When I first saw this mop i love the idea of the	Great Concept, Bad design

Scraping and pasting and saving UTF-8 text file

- Exercise 1 – saving a Word document as a text file
  - Resume
- Exercise 2 – scraping text from a web page and creating a text file
  - LinkedIn profile
- Exercise 3 – Saving PDF as a text file
  - LinkedIn PDF profile

Additional Processing

- The text files created in Word have an additional element that must be removed – the carriage return
- Open the Text file in Word and using the Replace function replace all carriage returns (^p) with a space
- Resave it as a text file
- It can be scraped into the compute buffer ready to be pasted into a data cell in Excel
- Example of not removing ^p and pasting into Excel- error!

1 Text file produced by Adobe from original PDF file

2 Text file processed by Word to strip CRs

3 Resulting text file without CRs

FORTINO Global Education