Plan for the week

- M: Introduction to Spreadsheets
- W: Descriptive statistics
  - Measures of central tendency
- F: Section
  - More advanced spreadsheet functionality (sort, filter, pivot tables, etc.)

Databases vs. Spreadsheets

Raw data can be unmanageable

Data Source

- What company appears most often in the data set?
- What nationality is most common?
- Who is the youngest person in the dataset?
- What is the average net worth of all the people in the dataset?

name	networth	nationality	company	age
Carlos Slim Helu	$54.5 Billion	Mexico	América Móvil	77
David Koch	$48.3 Billion	U.S.	Koch Industries	77
Charles Koch	$48.3 Billion	U.S.	Koch Industries	81
Bill Gates	$86 Billion	U.S.	Microsoft	60
Sheldon Adelson	$30.4 Billion	U.S.	Las Vegas Sands	83
Warren Buffett	$75.6 Billion	U.S.	Berkshire Hathaway	85
Sergey Brin	$39.8 Billion	U.S.	Google	43
Larry Ellison	$52.5 Billion	U.S.	Oracle	72
Wang Jianlin	$31.3 Billion	China	Dallan Wanda Group	62
Li Ka-shing	$31.2 Billion	Hong Kong	CK Hutchison Holding	89
Liliane Bettencourt	$39.5 Billion	French	L'Oreal	94
Larry Page	$39.8 Billion	U.S.	Google	43
Amancio Ortega	$71.3 Billion	Spain	Zara	81
Jim Walton	$34 Billion	U.S.	Walmart	60
Rob Walton	$34.1 Billion	U.S.	Walmart	72
Bernard Arnault	$41.5 Billion	French	Louis Vuitton	68
Alice Walton	$33.8 Billion	U.S.	Walmart	67
Jeff Bezos	$72.8 Billion	U.S.	Amazon	53
Mark Zuckerberg	$56 Billion	U.S.	Facebook	33
Michael Bloomberg	$47.5 Billion	U.S.	Bloomberg	75

Raw data can be unmanageable

Data Source

- What company appears most often in the data set?
  - Walmart
- What nationality is most common?
  - U.S.
- Who is the youngest person?
  - Mark Zuckerberg
- What is the average networth of the people in the dataset?
  - $48.41 Billion

name	networth	nationality	company	age
Carlos Slim Helu	$54.5 Billion	Mexico	América Móvil	77
David Koch	$48.3 Billion	U.S.	Koch Industries	77
Charles Koch	$48.3 Billion	U.S.	Koch Industries	81
Bill Gates	$86 Billion	U.S.	Microsoft	60
Sheldon Adelson	$30.4 Billion	U.S.	Las Vegas Sands	83
Warren Buffett	$75.6 Billion	U.S.	Berkshire Hathaway	85
Sergey Brin	$39.8 Billion	U.S.	Google	43
Larry Ellison	$52.5 Billion	U.S.	Oracle	72
Wang Jianlin	$31.3 Billion	China	Dallan Wanda Group	62
Li Ka-shing	$31.2 Billion	Hong Kong	CK Hutchison Holding	89
Liliane Bettencourt	$39.5 Billion	French	L'Oreal	94
Larry Page	$39.8 Billion	U.S.	Google	43
Amancio Ortega	$71.3 Billion	Spain	Zara	81
Jim Walton	$34 Billion	U.S.	Walmart	60
Rob Walton	$34.1 Billion	U.S.	Walmart	72
Bernard Arnault	$41.5 Billion	French	Louis Vuitton	68
Alice Walton	$33.8 Billion	U.S.	Walmart	67
Jeff Bezos	$72.8 Billion	U.S.	Amazon	53
Mark Zuckerberg	$56 Billion	U.S.	Facebook	33
Michael Bloomberg	$47.5 Billion	U.S.	Bloomberg	75

Raw data can be unmanageable

Data Source

- It is often impossible to answer these or similar queries, or to establish any trends, by eyeballing the data.
- Computational processing is necessary for data beyond a trivial size.

name	networth	nationality	company	age
Carlos Slim Helu	$54.5 Billion	Mexico	América Móvil	77
David Koch	$48.3 Billion	U.S.	Koch Industries	77
Charles Koch	$48.3 Billion	U.S.	Koch Industries	81
Bill Gates	$86 Billion	U.S.	Microsoft	60
Sheldon Adelson	$30.4 Billion	U.S.	Las Vegas Sands	83
Warren Buffett	$75.6 Billion	U.S.	Berkshire Hathaway	85
Sergey Brin	$39.8 Billion	U.S.	Google	43
Larry Ellison	$52.5 Billion	U.S.	Oracle	72
Wang Jianlin	$31.3 Billion	China	Dallan Wanda Group	62
Li Ka-shing	$31.2 Billion	Hong Kong	CK Hutchison Holding	89
Liliane Bettencourt	$39.5 Billion	French	L'Oreal	94
Larry Page	$39.8 Billion	U.S.	Google	43
Amancio Ortega	$71.3 Billion	Spain	Zara	81
Jim Walton	$34 Billion	U.S.	Walmart	60
Rob Walton	$34.1 Billion	U.S.	Walmart	72
Bernard Arnault	$41.5 Billion	French	Louis Vuitton	68
Alice Walton	$33.8 Billion	U.S.	Walmart	67
Jeff Bezos	$72.8 Billion	U.S.	Amazon	53
Mark Zuckerberg	$56 Billion	U.S.	Facebook	33
Michael Bloomberg	$47.5 Billion	U.S.	Bloomberg	75

Databases

- A database is a structured set of data that are easily accessible in various ways
- Database software tools facilitate the automated management of data
  - Storing, searching, modifying, and extracting information in a database
- Database systems can manage a very large quantity of data

Database Software

- Database Management Systems: software that handles storage, retrieval, and updating of databases
  - Examples: Oracle, MySQL, Access
- Statistics Packages
  - Examples: SAS, SPSS, Stata, R
- Spreadsheets: they do a little of both, at human scale, and for human comprehension, with minimal human effort!

Spreadsheets

- Suitable for managing relatively small databases
- Designed to facilitate human comprehension of data
- Useful for visualizing both raw data, in tabular form, and summaries of data in the form of charts, etc.
- Data management is usually accomplished manually, not via automated tools

History of Databases

Public School Database then

- Before computers were commonplace, people used other systems to track student data
- Schools often used paper-based books (can you imagine?!) for grades, attendance, student contact info, etc.
- These books would be divided into many small squares where teachers would record relevant information

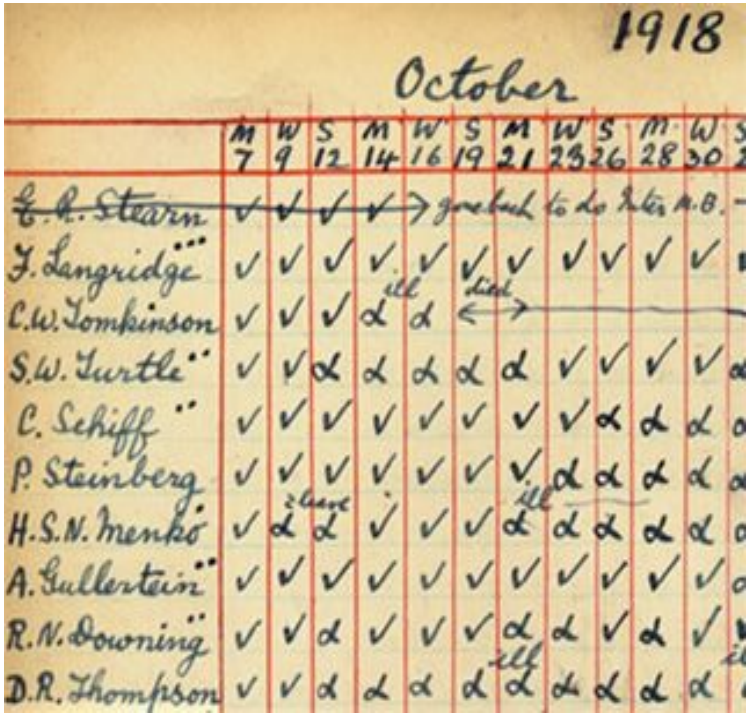


Image Source

Difficulties with paper-based records

- Mistakes are hard to correct and quickly become quite messy
- It is difficult to search through the whole book (slowly or quickly)
- Calculating summary statistics is difficult (e.g., how many days did R.N. Downing miss school?)
- Records can be lost forever or damaged due to water, fire, etc., because it is expensive to store backups

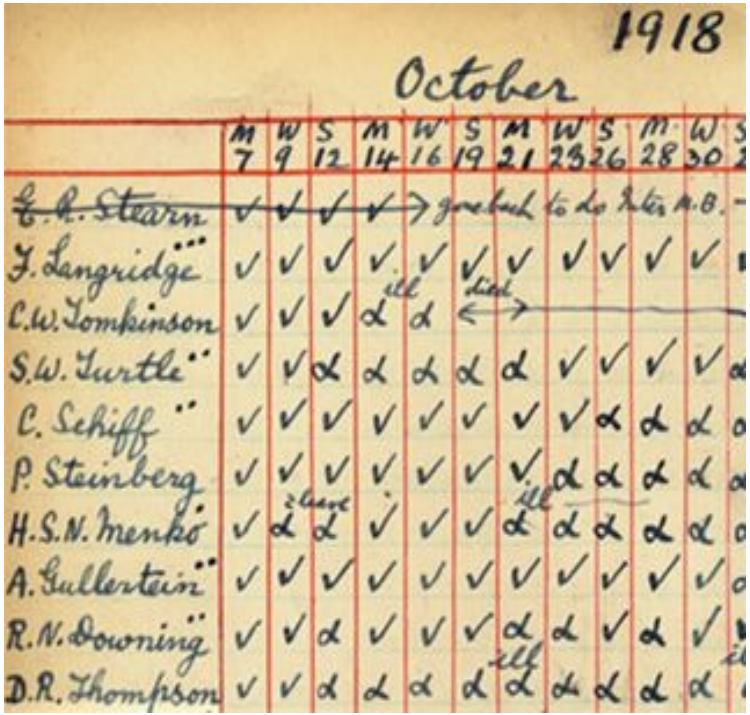


Image Source

Public School Database now

- Make heavy use of (hopefully, secure) spreadsheets
- Mistakes are easy to fix (and make!) in spreadsheets
  - One can simply overwrite an entry
- Can easily search for specific entries
  - E.g., did R.N. Downing attend school on 10/14?
- Spreadsheets can also be used to easily calculate summary statistics, like grade-point averages and total number of days absent
- Records are backed up on some online system, so they are less likely to be damaged like paper books

	A	B	C	D	E	F	G
7	Student Name		Monday	Tuesday	Wednesday	Thursday	Friday
8			15 Nov-15	16 Nov-15	17 Nov-15	18 Nov-15	19 Nov-15
9	01	Lee, John	PR	PR	PR	PR	PR
10	02	Williams, Adam	PR	PR	PR	PR	PR
11	03	Williams, Sarah	PR	PR	PR	PR	PR
12	04	Doe, Jane	PR	**AB**	**AB**	PR	**AB**
13	05	Liu, Kat	PR	**AB**	PR	PR	PR
14	06	Smith, Elizabeth	PR	PR	PR	PR	PR
15	07	Sanders, Shane	**AB**	PR	PR	PR	PR
16	08	White, Andres	PR	PR	PR	PR	PR
17	09	Aaronson, Allison	PR	PR	PR	PR	PR
18	10	Mapple, Noah	PR	PR	PR	PR	PR
19	11	Grande, John	PR	PR	PR	PR	PR
20	12	McDonell, Josh	PR	PR	PR	PR	PR
21	13	Spear, Terry	PR	PR	PR	PR	PR
22	14	Greene, Tom	PR	PR	PR	PR	PR
23	15	Snow, Daniel	PR	PR	PR	PR	PR
24	16	Daniels, Jack	PR	PR	**AB**	PR	PR
25	17	Daniels, Bob	PR	PR	PR	PR	PR
26	18	West, Williams	PR	PR	PR	PR	PR
27	19						
28	20						
29	21						
30	22						
31	23						
32	24						
33	25						
34	26						
35	27						
36	28						
37	29						
38	30						
39		**Nº PResent per Day:**	**17**	**16**	**16**	**18**	**17**
40		**Nº ABsent per Day:**	**1**	**2**	**2**	**0**	**1**
41	**18**	**Total:**	**18**	**18**	**18**	**18**	**18**
42		**Daily Attendance %:**	**94**	**89**	**89**	**100**	**94**

ATTENDANCE:
Select..
PR if Present
AB if Absent

Weekly Attendance Record

Teacher:		Ms. Smith				Group Name:	4th Grade			
WEEK BEGINNING: (dd/mm/yyyy)		11/15/2015				Level:				
		Sunday								
Student Name		Monday	Tuesday	Wednesday	Thursday	Friday	Notes	ATTENDANCE		
		15 Nov-15	16 Nov-15	17 Nov-15	18 Nov-15	19 Nov-15		Present	Absent	
01	Lee, John	PR	PR	PR	PR	PR	Was an hour late on Wednesday (Dentist)	5	0	
02	Williams, Adam	PR	PR	PR	PR	PR		5	0	
03	Williams, Sarah	PR	PR	PR	PR	PR		5	0	
04	Doe, Jane	PR	AB	AB	PR	AB		2	3	
05	Liu, Kat	PR	AB	PR	PR	PR		4	1	
06	Smith, Elizabeth	PR	PR	PR	PR	PR		5	0	
07	Sanders, Shane	AB	PR	PR	PR	PR		4	1	
08	White, Andres	PR	PR	PR	PR	PR		5	0	
09	Aaronson, Allison	PR	PR	PR	PR	PR		5	0	
10	Mapple, Noah	PR	PR	PR	PR	PR		5	0	
11	Grande, John	PR	PR	PR	PR	PR		5	0	
12	McDonell, Josh	PR	PR	PR	PR	PR		5	0	
13	Spear, Terry	PR	PR	PR	PR	PR		5	0	
14	Greene, Tom	PR	PR	PR	PR	PR		5	0	
15	Snow, Daniel	PR	PR	PR	PR	PR		5	0	
16	Daniels, Jack	PR	PR	AB				4	1	
17	Daniels, Bob	PR	PR	PR				5	0	
18	West, Williams	PR	PR	PR				5	0	
19								0	0	
20								0	0	
21								0	0	
22								0	0	
23								0	0	
24								0	0	
25								0	0	
26								0	0	
27								0	0	
28								0	0	
29								0	0	
30								0	0	
18	N° PResent per Day:	17	16	16	18	17	Week Attendance Totals:	84	6	
	N° ABsent per Day:	1	2	2	0	1				
	Total:	18	18	18	18	18	Days Recorded:	6		
	Daily Attendance %:	94	89	89	100	94	Week Attendance %:	93	7	

ATTENDANCE:
Select..
PR if Present
AB if Absent

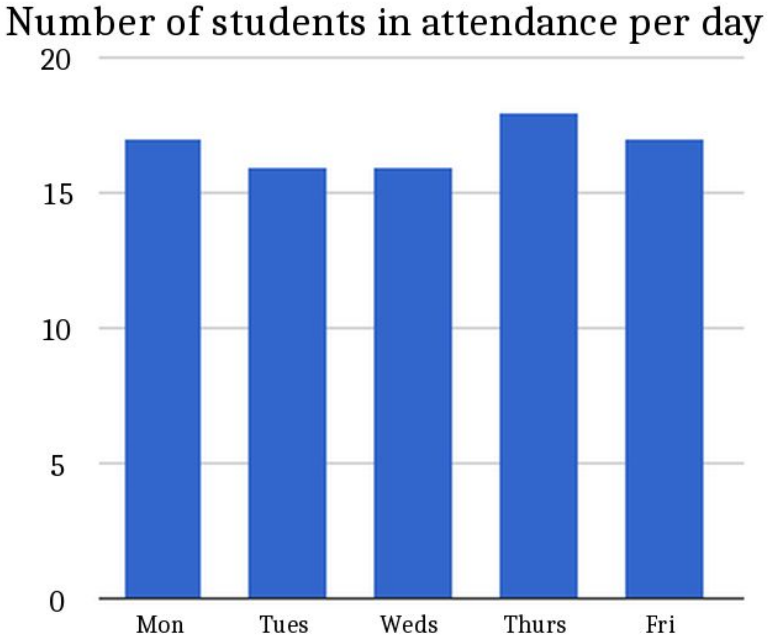
Summarizing the Data

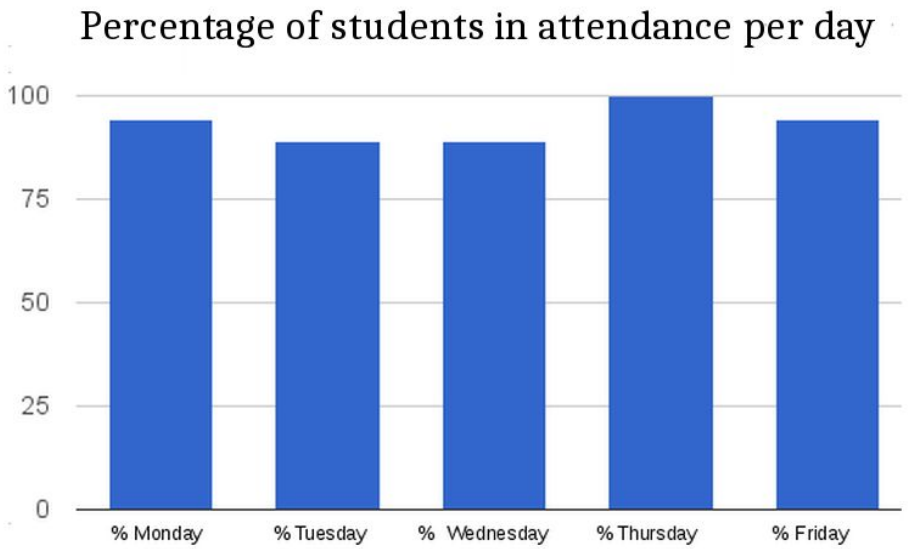
Number of students in attendance per day

	A	B	C	D	E
1	Data				
2	Sum - Monda	Sum - Tuesda	Sum - Wedne	Sum - Thursd	Sum - Friday
3	17	18	16	16	17

- Student with the most absences: Jane Doe
- Student with the most Monday absences: Shane Sanders
- Etc.

(Coarse) Visualizations





History of Spreadsheets

History of Spreadsheets

- The word spreadsheet comes from the word “spread,” as in a newspaper or magazine
- “Spreadsheet” was used to refer to bookkeeping ledgers with all revenue, costs, taxes, etc. spread on a single sheet of paper (or across two pages of a bound ledger)
- In 1961, Professor Richard Mattessich pioneered computerized spreadsheets for business accounting on main frames: "Spreadsheet: Its First Computerization (1961-1964)"

VisiCalc: Dan Bricklin & Bob Frankston

Subject: RE: Spreadsheets history
Date: Wed, 12 May 1999 17:25:22 -0400
From: Dan Bricklin
Organization: Trellix Corporation
To: 'Dan Power'
CC: "'Bob_Frankston@frankston.com'"

Dan,

As I said I'd do in my email a few weeks ago (I've included a copy below) I've posted on my web site a first attempt at addressing what's so special about VisiCalc that makes people call it the "first" electronic spreadsheet as we know them today. As to the story of coming up with the idea and refining it into a product, I have the beginnings on my site (which you've read already, I assume) and Bob and I are working on more information about design decisions, exactly what happened, etc. Since we were often the only ones there, I hope you'll put some trust in our story (which we'll try to back up with pictures, check with friends from that time, etc.). I'll be seeing some of my classmates from Harvard in a few weeks and will try to get more first-hand accounts of the history.

Here's the URL for my posting:

http://www.bricklin.com/firstspreadsheetquestion.htm

If you want some nit picks on your piece (if not, skip this paragraph): I programmed the first "working prototypes" (not "working version") in the fall of 1978 (Oct? Nov? I'll have to check -- I think it was over a long weekend), not the summer. I recruited Bob to do a real, assembler version, instead of the prototype in Integer Basic. The prototypes actually had a better interface in some cases (context sensitive help, for example). Bob was to build production code (faster speed, better arithmetic, scrolling, etc.). Since one of the things that made VisiCalc special was the implementation and the details only worked out in that implementation (some of which we will cover in that material Bob and I are working on) Bob is referred to as the "co-creator" of VisiCalc. If you want to include his many contributions in the category of invention, then he certainly is a "co-inventor", but there is no question he was a co-creator. Fylstra, Bob, and I first talked about VisiCalc in the fall (after summer vacation I call fall) of 1978. I'll get the actual dates of some of these meetings later (for the record, not that many people care...). I don't think the number of sold copies ever made it over several hundred thousand, but I don't have the records ("about 1 million" sounds nice and if you round 500K+ up, then it's probably correct). [Enough nits... Tell me if you want things at this detail. :)]

Let me know what you think.

Thanks again for your interest in the history of this area.

-DanB

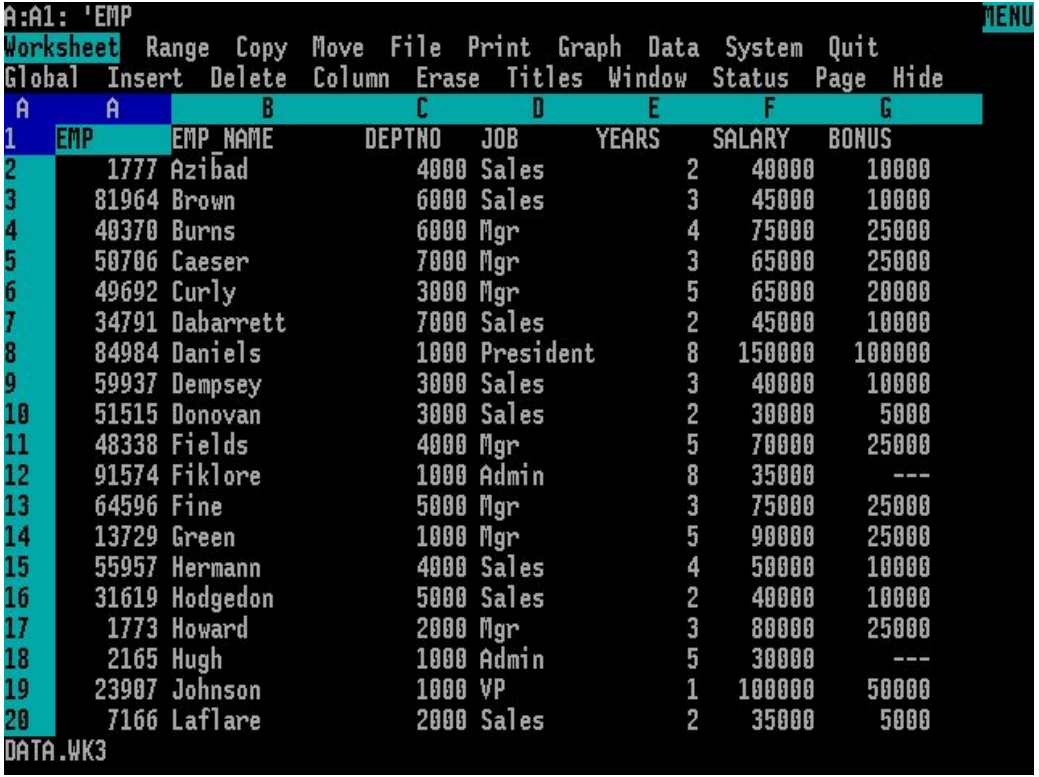
VisiCalc

Big idea: Instant automatic recalculation based on formulas stored in the cells referencing other cells

Image Source

Lotus 1-2-3

- January 1983
- Mitch Kapor and Jonathan Sachs
- More powerful than VisiCalc
- Very popular
- Lotus 1-2-3 is thought to be one of the reasons the IBM PC was so successful. Likewise, for VisiCalc and the Apple II.

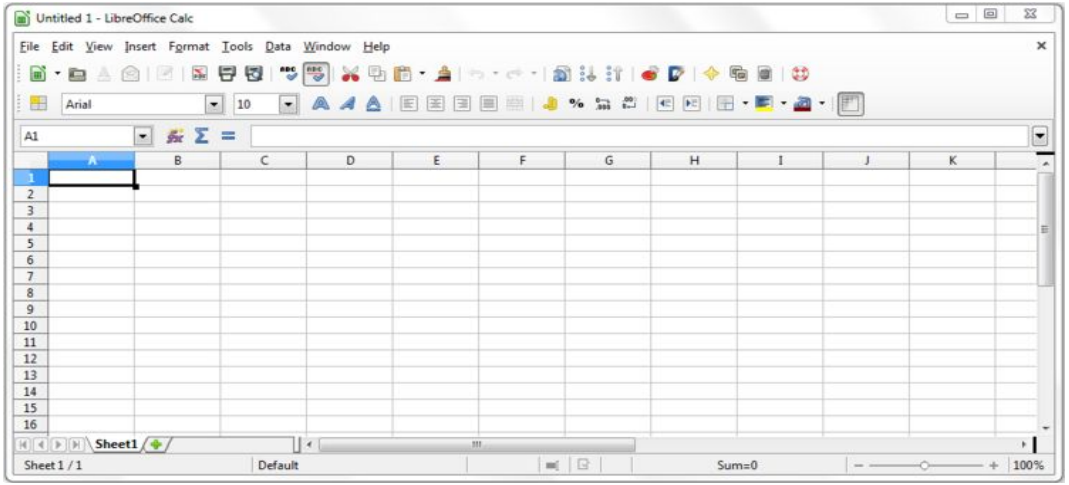


Microsoft Excel

- Mid 1980’s
- First spreadsheet to use a GUI, or graphical user interface, meaning not an entirely text-based interface
- Was a big selling point for Apple’s Macintosh; people bought it because it came with Excel
- In 1987, Microsoft released Windows with Excel

Today there are many options

- Microsoft Excel
- Google Spreadsheets
- LibreOffice Calc
  - The screenshots in these slides are of LibreOffice Calc, because it used to be used in CS 100



Qualitative vs. Quantitative Data

Data can be either qualitative or quantitative

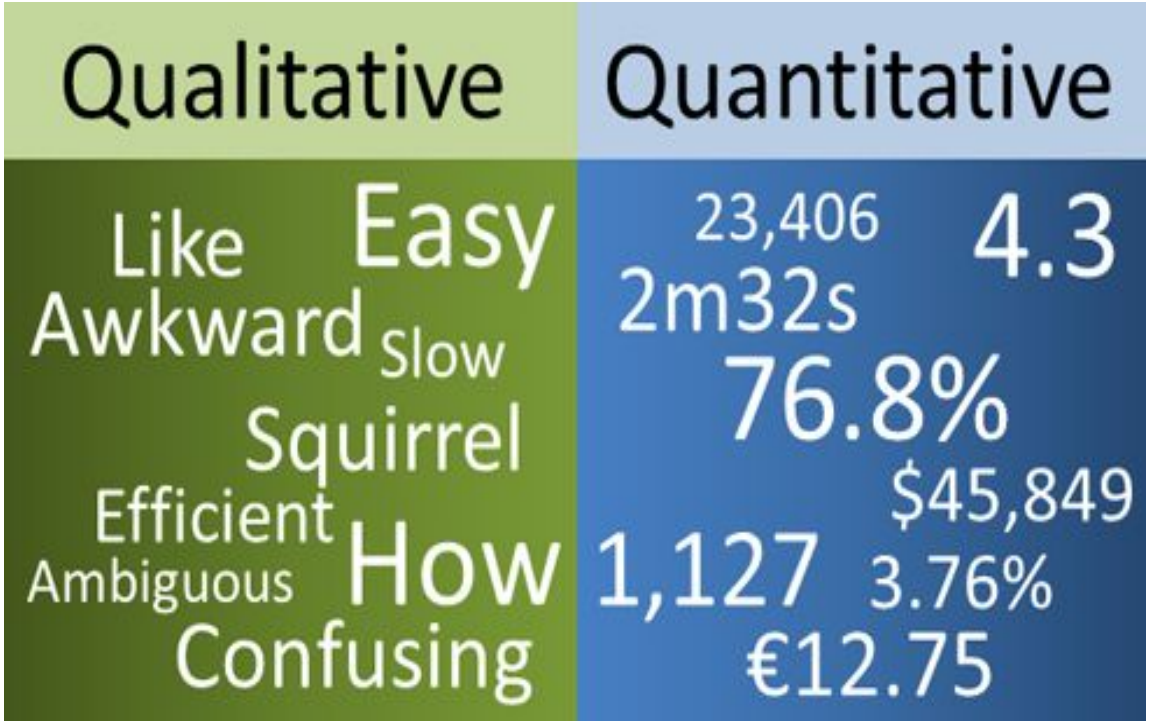
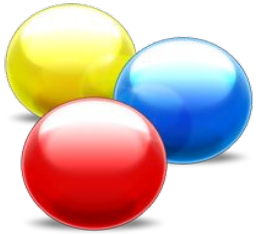


Image Source

Image Source

Qualitative data



Qualitative data describe qualities, like color, texture, smell, taste, appearance, etc.

Many qualitative data are categorical: e.g.,

- the color of a ball (yellow, blue, or red)
- the brand of a product purchased (brand A, B, or C)
- whether a person is employed (yes or no)

Qualitative data can be nominal or ordinal

- Nominal means that there is no natural order among the values

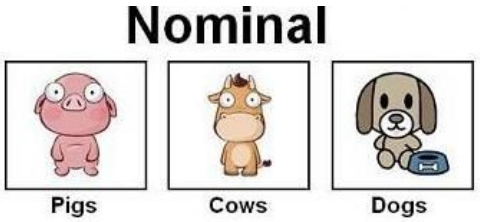


Image source

- Ordinal means that there is a natural ordering

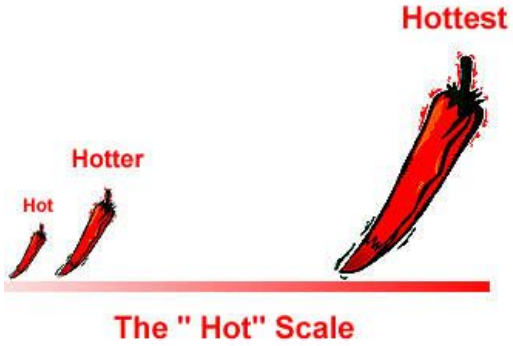


Image Source

Quantitative data

Quantitative data take on numerical values, so are typically ordinal

Examples:

- age, weight, height, income, etc.
- the value of a country's exports
- a batter's number of home runs

Image Source

Quantitative data can be either discrete or continuous

- Data are discrete if the measurements are necessarily integral (i.e., integers)

- Data are continuous if the measurements can take on any value, usually within some range