



Data Analysis Using Spreadsheets

CS102
Spring 2020

Spreadsheets

Most familiar use is for data presentation

Formulas handy for sales, budgets, and other numeric summaries

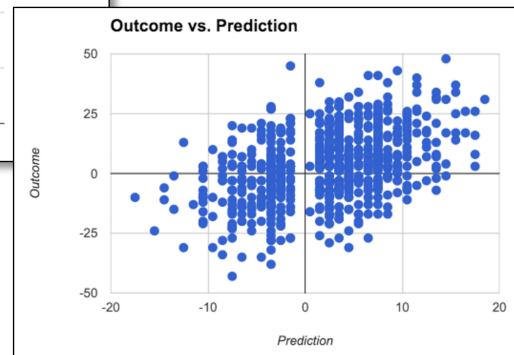
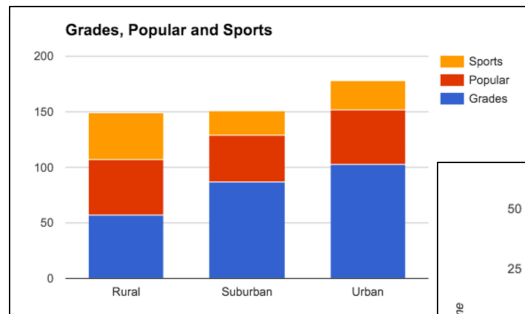
The screenshot shows a spreadsheet for a fitness program. It has columns for exercises (e.g., Core Lift, Squat, Bench Press, Deadlift, Shoulder Press), sets, reps, and weeks. A 'Boring But Big' column is highlighted in yellow. There are also columns for 'Wendler Percentages' (60%, 70%, 75%, 80%, 85%, 90%, 95%, 60%). The spreadsheet is color-coded by week and exercise type.

Formula bar: $=SUM(B4:H4)$

	A	B	C	D	E	F	G	H	I
1									
2									
3		Jan	Feb	Mar	Apr	May	Jun	Jul	Total
4	Person 1	620	768	251	811	664	304	27	3445
5	Person 2	1	928	595	214	317	470	360	2885
6	Person 3	707	481	849	255	548	550	518	3908
7	Person 4	235	110	357	730	739	265	36	2472
8	Person 5	610	508	353	952	643	16	738	3820
9	Person 6	425	648	740	162	865	332	786	3958
10	Person 7	695	751	111	675	736	407	6	3381
11	Person 8	326	449	80	612	779	1000	341	3587
12	Person 9	981	540	509	860	92	631	900	4513
13	Total	4600	5183	3845	5271	5383	3975	3712	31969
14									

Spreadsheets

But also a convenient and powerful tool for analysis of structured data
(And for data visualization)



	A	B	C	D	E	F	G
1	Year	Week	Home	HomeScore	Away	AwayScore	Prediction
2	1998	1	Green_Bay	38	Detroit	19	9.5
3	1998	1	Chicago	23	Jacksonville	24	-8.5
4	1998	1	Minnesota	31	Tampa Bay	7	3.5
5	1998	1	St_Louis	17	New_Orleans	24	3.5
6	1998	1	Cincinnati	14	Tennessee	23	1.5
7	1998	1	Baltimore	13	Pittsburgh	20	-3.5
8	1998	1	Carolina	14	Atlanta	19	4.5
9	1998	1	NY_Giants	31	Washington	24	2.5
10	1998	1	Philadelphia	0	Seattle	38	-3.5
11	1998	1	San_Diego	16	Buffalo	14	1.5
12	1998	1	San_Francisco	36	NY_Jets	30	7.5
13	1998	1	Dallas	38	Arizona	10	5.5
14	1998	1	Indianapolis	15	Miami	24	-3.5
15	1998	1	Kansas_City	28	Oakland	8	7.5
16	1998	1	Denver	27	New_England	21	7.5
17	1998	2	Tennessee	7	San_Diego	13	7.5
18	1998	2	Green_Bay	23	Tampa Bay	15	7.5
19	1998	2	New_Orleans	19	Carolina	14	-3.5
20	1998	2	St_Louis	31	Minnesota	38	-7.5
21	1998	2	Miami	13	Buffalo	7	7.5
22	1998	2	Jacksonville	21	Kansas_City	16	1.5
23	1998	2	NY_Jets	10	Baltimore	24	3.5
24	1998	2	Pittsburgh	17	Chicago	12	11.5
25	1998	2	Atlanta	17	Philadelphia	12	8.5
26	1998	2	Detroit	28	Cincinnati	34	6.5
27	1998	2	Oakland	20	NY_Giants	17	1.5
28	1998	2	Seattle	33	Arizona	14	7.5
29	1998	2	Denver	42	Dallas	23	7.5
30	1998	2	New_England	29	Indianapolis	0	8.5
31	1998	2	Washington	10	San_Francisco	45	-4.5
32	1998	3	Kansas_City	23	San_Diego	7	9.5
33	1998	3	Minnesota	29	Detroit	6	5.5
34	1998	3	Buffalo	33	St_Louis	34	4.5
35	1998	3	Cincinnati	6	Green_Bay	13	-7.5
36	1998	3	Miami	21	Pittsburgh	0	1.5

Spreadsheets

- A surprisingly large fraction of the world's structured data is managed and manipulated in spreadsheets
- Spreadsheets are used by 750 million people — 10% of the world's population

Microsoft Excel is dominant tool

- Many features
- Proprietary and expensive

Google Sheets

- Open and free
- Fewer features, but catching up

What We'll Cover

Spreadsheet basics

- Importing and exporting
- Inserting and deleting
- Formulas

Data operations

- Sorting
- Filtering
- Aggregation
- Joining

Even people with significant spreadsheet experience may learn a few new things

Pivot tables

- Restructuring / aggregation / analysis

In-Class Data Set

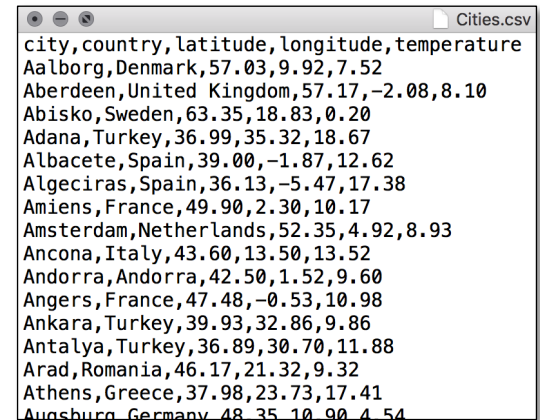
Europe City Temperatures

Cities: city, country, latitude, longitude, temperature

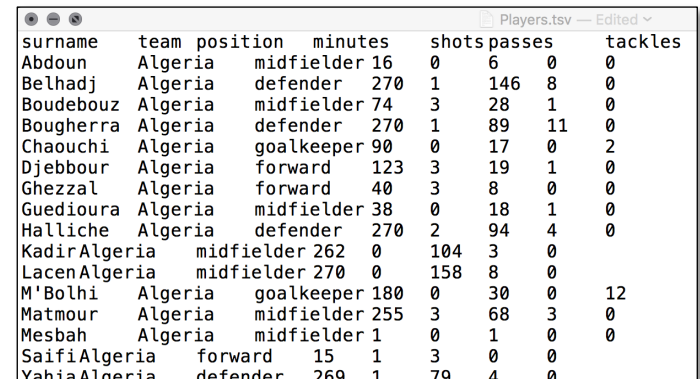
Countries: country, population, EU, coastline

Importing and Exporting

- Structured data in files
 - Comma-separated values (CSV)
 - Tab-separated values (TSV)
- Import into format used by spreadsheet program
- Export from spreadsheet to CSV or TSV (or others)



```
city, country, latitude, longitude, temperature
Aalborg, Denmark, 57.03, 9.92, 7.52
Aberdeen, United Kingdom, 57.17, -2.08, 8.10
Abisko, Sweden, 63.35, 18.83, 0.20
Adana, Turkey, 36.99, 35.32, 18.67
Albacete, Spain, 39.00, -1.87, 12.62
Algeciras, Spain, 36.13, -5.47, 17.38
Amiens, France, 49.90, 2.30, 10.17
Amsterdam, Netherlands, 52.35, 4.92, 8.93
Ancona, Italy, 43.60, 13.50, 13.52
Andorra, Andorra, 42.50, 1.52, 9.60
Angers, France, 47.48, -0.53, 10.98
Ankara, Turkey, 39.93, 32.86, 9.86
Antalya, Turkey, 36.89, 30.70, 11.88
Arad, Romania, 46.17, 21.32, 9.32
Athens, Greece, 37.98, 23.73, 17.41
Augsburg, Germany, 48.35, 10.90, 4.54
```



```
Player Statistics Table
```

surname	team	position	minutes	shots	passes	tackles
Abdoun	Algeria	midfielder	16	0	6	0
Belhadj	Algeria	defender	270	1	146	8
Boudebouz	Algeria	midfielder	74	3	28	1
Bougherra	Algeria	defender	270	1	89	11
Chaouchi	Algeria	goalkeeper	90	0	17	0
Djebbour	Algeria	forward	123	3	19	1
Ghezzal	Algeria	forward	40	3	8	0
Guedioura	Algeria	midfielder	38	0	18	1
Halliche	Algeria	defender	270	2	94	4
Kadir	Algeria	midfielder	262	0	104	3
Lacen	Algeria	midfielder	270	0	158	8
M'Bolhi	Algeria	goalkeeper	180	0	30	0
Matmour	Algeria	midfielder	255	3	68	3
Mesbah	Algeria	midfielder	1	0	1	0
Saifi	Algeria	forward	15	1	3	0
Yahia	Algeria	defender	269	1	79	4

Let's Get Started!

- Inserting and deleting rows
- Inserting and deleting columns
- Formulas

Your Turn

1. Add new column to the left of column F called celsius
2. Use formula to compute values from fahrenheit column E

Note: Celsius = (Fahrenheit - 32) * 5/9

Data Operations

- Sorting
- Filtering
- Aggregation
- Grouped aggregation
- Joining

Your Turn

How many cities in Italy?

Note: There are several ways to solve this one,
some better than others

Data Operations

- Sorting
- Filtering
- Aggregation
- Grouped aggregation
- Joining

Your Turn

What is the average latitude...

1. overall ?
2. for cities with temperature < 10 ?
3. for cities with temperature > 10 ?
4. for cities where both the city name and the country name end in the letter “a” ?

Pivot Tables

For data restructuring, aggregation, general analysis

Convenient and powerful!

Pivot Tables

For data restructuring, aggregation, general analysis

Convenient and powerful!

But pivot tables don't have full spreadsheet functionality – sometimes must copy-paste (special) to new sheet to do further analysis

Your Turn

1. (easy) Which are warmer on average - cities in the EU or cities not in the EU?
2. (harder) What are the western-most and eastern-most countries with no coastline?

For #2: Define the longitude of a country as the *average* longitude of cities in that country, and remember that smaller longitudes are further west.

Explore the features of pivot tables - there are several ways to solve this one!

Data Analysis with Spreadsheets

Convenient and powerful

- Many analyses can be done in “big data” style

No scrolling

A few limitations

- Data size
 - Google sheets: 400,000 cells
- Some analyses are difficult
 - E.g., two cities closest to each other (easy in SQL)

Data Analysis with Spreadsheets

For help while working with spreadsheets:

- Dropdown tips
- Tutorials and help pages (website)
- My favorite: web search