

How to load data

Dealing with PDFs

- PDFs are terrible files for storing data
- Some PDFs are machine-readable, others aren't
 - Machine-readable means the underlying data in the PDF is still intact
 - How can you tell? Try to highlight the text.

Dealing with PDFs

- If you can highlight the data in the PDF, it **IS** machine-readable
 - Try an online tool like Tabula to get the data into to a spreadsheet
- If you can't highlight it, i.e. if it looks like a scanned document or a photo, it's **NOT** machine-readable
 - You'll need OCR (optical character recognition software to convert into machine-readable data. Try Acrobat or Document Cloud or [new Excel hotness](#)
 - OR go back to your data source and ask for a machine-readable file

File Types

- Data can be stored in lots of file types
- Text files - the simplest. Basically a text document of data
- File type is usually .txt
 - Obviously a 10,000-page document of data isn't useful. We have to separate it somehow.
 - Generally two ways to do this

File Types

- Check the file name extension (.csv, .txt, etc.)
- Open the file up in a text editor and browse
- Do you have a header row?
- What's separating each field?

1	1749	01	1749.042	96.7	-1.0	-1
2	1749	02	1749.123	104.3	-1.0	-1
3	1749	03	1749.204	116.7	-1.0	-1
4	1749	04	1749.288	92.8	-1.0	-1
5	1749	05	1749.371	141.7	-1.0	-1
6	1749	06	1749.455	139.2	-1.0	-1
7	1749	07	1749.538	158.0	-1.0	-1
8	1749	08	1749.623	110.5	-1.0	-1
9	1749	09	1749.707	126.5	-1.0	-1
10	1749	10	1749.790	125.8	-1.0	-1
11	1749	11	1749.874	264.3	-1.0	-1
12	1749	12	1749.958	142.0	-1.0	-1
13	1750	01	1750.042	122.2	-1.0	-1
14	1750	02	1750.123	126.5	-1.0	-1
15	1750	03	1750.204	148.7	-1.0	-1
16	1750	04	1750.288	147.2	-1.0	-1
17	1750	05	1750.371	150.0	-1.0	-1
18	1750	06	1750.455	166.7	-1.0	-1
19	1750	07	1750.538	142.3	-1.0	-1
20	1750	08	1750.623	171.7	-1.0	-1
21	1750	09	1750.707	152.0	-1.0	-1
22	1750	10	1750.790	109.5	-1.0	-1
23	1750	11	1750.874	105.5	-1.0	-1
24	1750	12	1750.958	125.7	-1.0	-1
25	1751	01	1751.042	116.7	-1.0	-1
26	1751	02	1751.123	72.5	-1.0	-1
27	1751	03	1751.204	75.5	-1.0	-1
28	1751	04	1751.288	94.0	-1.0	-1
29	1751	05	1751.371	101.2	-1.0	-1
30	1751	06	1751.455	84.5	-1.0	-1
31	1751	07	1751.538	110.5	-1.0	-1
32	1751	08	1751.623	99.7	-1.0	-1
33	1751	09	1751.707	39.2	-1.0	-1

File Type: Fixed Width

- You must manually put breaks in the data to tell your computer where a new field ends and another begins
- If you have a fixed width file, make sure your source gives you a guide or schema that tells you where to put the breaks. Don't guess!
 - Example

File Type: Delimited

- The data fields are separated by a delimited (often punctuation) that tells our computer where a new field ends and another begins
 - This is more common than fixed width
 - Common delimiters
 - Commas
 - Tabs
 - Pipes (|)

File Type: CSV

- Commas are the most common delimiters
- delimited files that use commas are also known as CSVs, aka comma-separated values
- File type can be .csv instead of .txt
- But what if you have commas in a field?
 - A text qualifier (usually quote marks) tells the computer to ignore commas within fields
 - Example: AT&T, Inc. will look like “AT&T, Inc.” in the text file

1 HCPCS Code,HCPCS Description,HCPCS Drug Indicator,Place of Service,Number of Providers,Number of Services,Number of Unique Beneficiary/Provider Interactions,Number of Distinct Medicare Beneficiary/Per Day Services,Average Submitted Charge Amount,Minimum Submitted Charge Amount,Maximum Submitted Charge Amount,Standard Deviation of Submitted Charge Amount,Average Medicare Allowed Amount,Minimum Medicare Allowed Amount,Maximum Medicare Allowed Amount,Standard Deviation of Medicare Allowed Amount,Average Medicare Payment Amount,Minimum Medicare Payment Amount,Maximum Medicare Payment Amount,Standard Deviation of Medicare Payment Amount

2 A0425,"Ground mileage, per statute mile",N,F,10343,140801494.5,6755692,12671918,15.0939463,0.076552661,408.2278481,8.08359992,7.801376767,0.048680084,10.74028436,0.925850875,6.133360523,0.8.593333333,0.726987114

3 A0425,"Ground mileage, per statute mile",N,O,16,3458.5,824,1304,12.51162064,6.87,120,2.463704996,7.093404655,6.87,10.74,0.064904357,5.577900824,5.385714286,8.42,0.05160961

4 A0426,"Ambulance service, advanced life support, non-emergency transport, level 1 (als 1)",N,F,3000,325543.3,282932,317966,791.3980132,7,28437,458.3650808,262.1513796,7,347.09,19.76073411,203.955559,5.556,277.67,15.76392806

5 A0427,"Ambulance service, advanced life support, emergency transport, level 1 (als1-emergency)",N,F,8378,4972850.9,3490463,4936591,923.3939244,90.67,28655.35714,460.1374707,417.5188687,90.67,549.57,32.27543877,324.7140335,0,436.73,25.20193278

6 A0427,"Ambulance service, advanced life support, emergency transport, level 1 (als1-emergency)",N,O,3,480,382,480,589.71875,445,600,6.62807024,421.2877292,397.57,425.45,1.099862878,330.9389375,311.7,340.36,0.977981327

7 A0428,"Ambulance service, basic life support, non-emergency transport, (bls)",N,F,4973,6812508.1,2053295,4844933,537.1170175,95,28454.625,275.6785558,220.8932109,95,289.24,17.11772737,172.9915218,0,228.65,13.46258727

8 A0429,"Ambulance service, basic life support, emergency transport (bls-emergency)",N,F,9722,2731110.7,2014926,2716724,669.2708485,76.87969925,31869.9,288.9939803,358.1562734,75.70225564,462.79,29.92258515,278.1289712,59.70285714,370.23,23.76236904

9 A0429,"Ambulance service, basic life support, emergency transport (bls-emergency)",N,O,3,822,572,822,499.6046594,252.03,500,9.047700509,354.757944,252.03,423,4.301793185,278.7129805,197.59,331.63,3.378427184

10 A0430,"Ambulance service, conventional air services, transport, one way (fixed wing)",N,F,89,10826,10072,10793,14438.15216,700,29903.92308,5062.165345,4190.477053,700,4731.355452,412.6544457,3275.365067,547.9038462,3652.114057,324.2669331

11 A0431,"Ambulance service, conventional air services, transport, one way (rotary wing)",N,F,275,56410,54843,56302,17309.74183,360,27709.38536,4177.623604,4599.301837,360,5510.41,296.4455132,3585.952542,282.24,4356.8975,233.0686692

12 A0432,"Paramedic intercept (pi), rural area, transport furnished by a volunteer ambulance company which is prohibited by state law from billing third party payers",N,F,59,3146,2641,3143,638.4824698,125,1492.44,161.4826762,375.3057629,125,407.41,29.98904694,288.0348601,98,320.7589655,24.78549587

13 A0433,"Advanced life support, level 2 (als 2)",N,F,5459,111967.8,108047,111419,1101.45624,2.06,5141.55,549.7118639,602.5650325,2.06,795.42,53.82315441,468.8264066,1.62,636.34,42.87535231

14 A0434,"Specialty care transport (sct)",N,F,1300,104281.7,79913,97874,1865.055575,40,32209.9,1148.997883,744.0198998,22.12932166,940.05,63.95298324,583.2122558,17.58468271,742.0133333,50.91906057

15 A0435,"Fixed wing air mileage, per statute mile",N,F,89,1980144.2,10020,10729,104.9972547,6.93,278.7912166,50.98005411,11.75196535,6.93,12.48003565,1.019015093,9.202622582,5.433166667,9.984006024,0.805605927

16 A0436,"Rotary wing air mileage, per statute mile",N,F,272,3322854,54437,55885,186.5635308,33.32,281.8598527,56.57891524,31.5708137,22.20998926,33.32000599,1.813376021,24.65698949,16.025,26.656,1.410686318

17 A0999,Unlisted ambulance service,N,F,486,3678.2,2767,3396,698.2032162,12,20283.27,495.3732378,357.4373144,7.089705882,3286.46,127.2913899,280.9220189,5.555,2576.59,100.0887829

18 A4212,Non-coring needle or stylet with or without catheter,N,O,4,19,18,19,6.236842105,5,17,2.966339322,0.405263158,0.01,6.5,1.454431951,0.324736842,0.01,5.2,1.162530672

19 A4215,"Needle, sterile, any size, each",N,O,22,182,121,137,6.50967033,0.18,80,8.595171502,0.037417582,0.01,5,0.368865846,0.031153846,0.008194444,4,0.205003222

data FIELD types

Just like there are different types of data files, there are different types of data fields (aka columns aka variables).

How you upload them can make a huge difference.

data FIELD types

- Text/character
 - Alphanumeric
 - Short or long
 - You don't need to do calculations with these things
 - Ex: names, addresses, descriptions, **zip codes**
- Numeric
 - Things you want to calculate with
 - Ex: counts, sums, dollars amounts
 - Can be integers or decimal numbers
- Dates/times
 - Ex: 7/22/2019, July-19, 1:32PM
 - Storing these as dates or times instead of text will make it a lot easier to do calculations, for example, calculating how many days between admission and discharge

Why should you care about types of data fields?

- Different field types sort differently!!
- If you're not careful when you load your data, Excel/Google sheets will try and be helpful and guess your data types.
 - It's usually wrong.
 - **THUS -> DO NOT JUST DOUBLE CLICK YOUR DATA FILE TO OPEN IT!!!!**

Computers are dumb

How does Excel/Google sheets get things wrong when guessing field types?

- Drops leading zeros
 - New England zip codes almost always start with zero, e.g. 02901
 - Excel will assume it's a number. 02901 becomes 2901. BAD BAD BAD
 - Avoid this by explicitly loading as a text field
- Reformat numbers to dates
 - Maybe there's a code in your data, like 11-53
 - Excel assumes this is November 1953 and loads it as 11/1/1953.
 - Avoid this by explicitly loading as a text field