Entering / cleaning data 1

Reading data into R

Basics of getting data into R

Basic approach:

- Download data to your computer
- Save the data in your R Project directory for the project you're using it for or in a subdirectory within that directory ("data" is a good name for this subdirectory)
- Read data into R (functions in `readr`: `read_csv`, `read_table`, `read_delim`, `read_fwf`, etc.)
- Check to make sure the data came in correctly (`dim`, `ncol`, `nrow`, `head`, `tail`, `str`, `colnames`)

What kind of data can you get into R?

The sky is the limit…

- **Flat files**
- Files from other statistical packages (SAS, Excel, Stata, SPSS)
- Tables on webpages
- Data in a database (e.g., SQL)
- Data stored in XML and JSON
- Complex data formats (e.g., netCDF files from climate folks, MRI data stored in Analyze, NIfTI, and DICOM formats)
- Data through APIs (e.g., GoogleMaps, Twitter, many government agencies)
- Incredibly messy data using `scan` and `readLines`

Flat files

R can read in data from *a lot* of different formats. The only catch: you need to tell R how to do it.

To start, we'll look at **flat files**, which are plain text files (i.e., you can read them when you open them in a text editor, unlike a file in a binary format, like an Excel or Word file) with a two-dimenional structure (a row for each observation and a column for each variable).

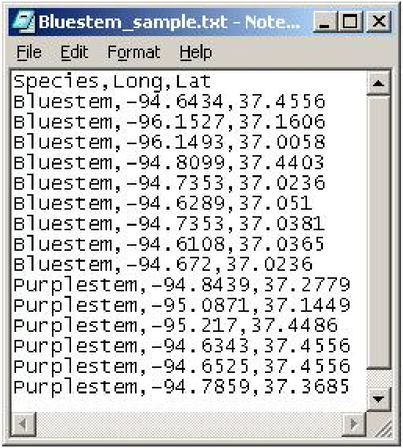
Types of flat files

There are two main types of flat files:

1. **Fixed width files**: Each column is a certain number of characters wide. (If you printed it out, you could draw vertical lines that separate the columns.)
2. **Delimited files**: In each row, a certain symbol (**delimiter**) separates the data into columns values for that observation.
    - ".csv": Comma-separated values
    - ".tab", ".tsv": Tab-separated values
    - Other possible delimiters: colon, semicolon, pipe ("|")

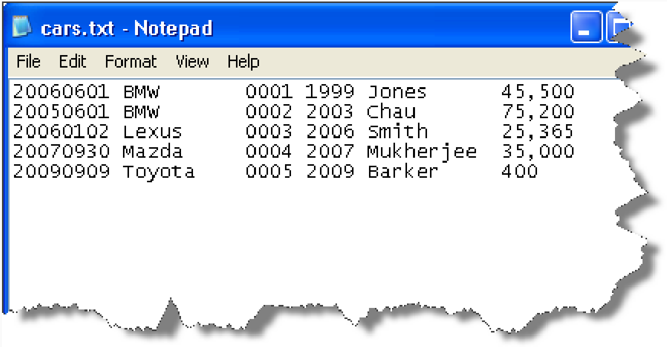
See if you can identify what types of files the following files are. . .

What type of file?



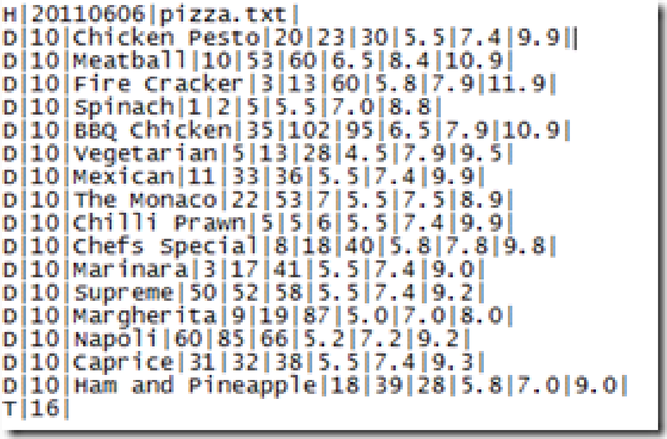
```
Species,Long,Lat
Bluestem,-94.6434,37.4556
Bluestem,-96.1527,37.1606
Bluestem,-96.1493,37.0058
Bluestem,-94.8099,37.4403
Bluestem,-94.7353,37.0236
Bluestem,-94.6289,37.051
Bluestem,-94.7353,37.0381
Bluestem,-94.6108,37.0365
Bluestem,-94.672,37.0236
Purplestem,-94.8439,37.2779
Purplestem,-95.0871,37.1449
Purplestem,-95.217,37.4486
Purplestem,-94.6343,37.4556
Purplestem,-94.6525,37.4556
Purplestem,-94.7859,37.3685
```

What type of file?



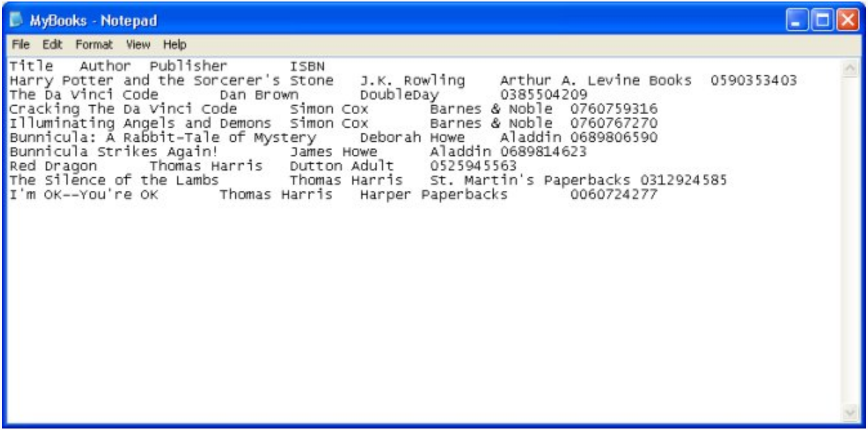
```
20060601 BMW         0001 1999 Jones      45,500
20050601 BMW         0002 2003 Chau       75,200
20060102 Lexus       0003 2006 Smith      25,365
20070930 Mazda       0004 2007 Mukherjee  35,000
20090909 Toyota      0005 2009 Barker     400
```

What type of file?



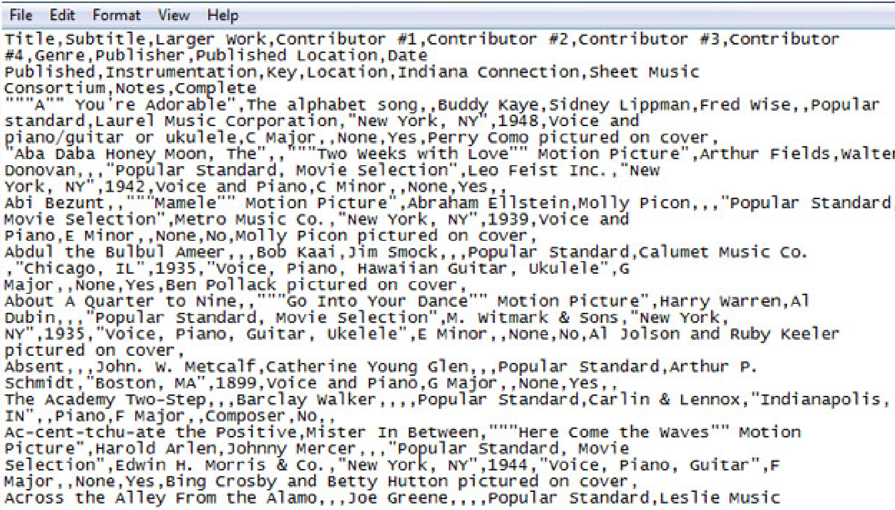
```
H|20110606|pizza.txt|
D|10|Chicken Pesto|20|23|30|5.5|7.4|9.9||
D|10|Meatball|10|53|60|6.5|8.4|10.9|
D|10|Fire Cracker|3|13|60|5.8|7.9|11.9|
D|10|Spinach|1|2|5|5.5|7.0|8.8|
D|10|BBQ Chicken|35|102|95|6.5|7.9|10.9|
D|10|Vegetarian|5|13|28|4.5|7.9|9.5|
D|10|Mexican|11|33|36|5.5|7.4|9.9|
D|10|The Monaco|22|53|7|5.5|7.5|8.9|
D|10|Chilli Prawn|5|5|6|5.5|7.4|9.9|
D|10|Chefs Special|8|18|40|5.8|7.8|9.8|
D|10|Marinara|3|17|41|5.5|7.4|9.0|
D|10|Supreme|50|52|58|5.5|7.4|9.2|
D|10|Margherita|9|19|87|5.0|7.0|8.0|
D|10|Napoli|60|85|66|5.2|7.2|9.2|
D|10|Caprice|31|32|38|5.5|7.4|9.3|
D|10|Ham and Pineapple|18|39|28|5.8|7.0|9.0|
T|16|
```

What type of file?



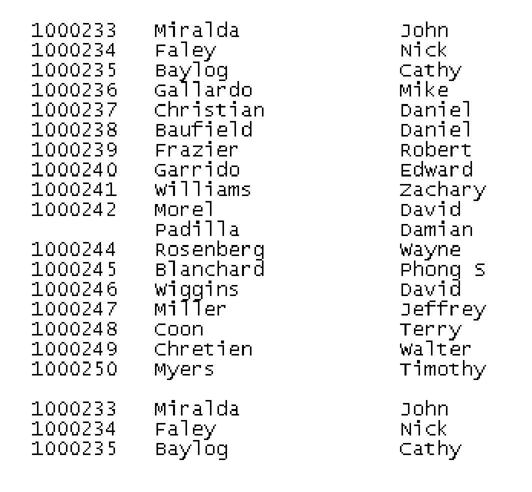
```
Title   Author  Publisher       ISBN
Harry Potter and the Sorcerer's Stone   J.K. Rowling    Arthur A. Levine Books  0590353403
The Da Vinci Code       Dan Brown       DoubleDay       0385504209
Cracking The Da Vinci Code      Simon Cox       Barnes & Noble  0760759316
Illuminating Angels and Demons  Simon Cox       Barnes & Noble  0760767270
Bunnicula: A Rabbit-Tale of Mystery     Deborah Howe    Aladdin 0689806590
Bunnicula Strikes Again!        James Howe      Aladdin 0689814623
Red Dragon      Thomas Harris   Dutton Adult    0525945563
The Silence of the Lambs        Thomas Harris   St. Martin's Paperbacks 0312924585
I'm OK--You're OK       Thomas Harris   Harper Paperbacks       0060724277
```

What type of file?



```
Title,Subtitle,Larger Work,Contributor #1,Contributor #2,Contributor #3,Contributor
#4,Genre,Publisher,Published Location,Date
Published,Instrumentation,Key,Location,Indiana Connection,Sheet Music
Consortium,Notes,Complete
"""A"" You're Adorable",The alphabet song,,Buddy Kaye,Sidney Lippman,Fred Wise,,Popular
standard,Laurel Music Corporation,"New York, NY",1948,Voice and
piano/guitar or ukulele,C Major,,None,Yes,Perry Como pictured on cover,
"Aba Daba Honey Moon, The",,"""Two Weeks with Love"" Motion Picture",Arthur Fields,Walte
Donovan,,,"Popular Standard, Movie Selection",Leo Feist Inc.,"New
York, NY",1942,Voice and Piano,C Minor,,None,Yes,,
Abi Bezunt,,"""Mamele"" Motion Picture",Abraham Ellstein,Molly Picon,,,"Popular Standard
Movie Selection",Metro Music Co.,"New York, NY",1939,Voice and
Piano,E Minor,,None,No,Molly Picon pictured on cover,
Abdul the Bulbul Ameer,,,Bob Kaai,Jim Smock,,,Popular Standard,Calumet Music Co.
,"Chicago, IL",1935,"Voice, Piano, Hawaiian Guitar, Ukulele",G
Major,,None,Yes,Ben Pollack pictured on cover,
About A Quarter to Nine,,"""Go Into Your Dance"" Motion Picture",Harry Warren,Al
Dubin,,,"Popular Standard, Movie Selection",M. Witmark & Sons,"New York,
NY",1935,"Voice, Piano, Guitar, Ukelele",E Minor,,None,No,Al Jolson and Ruby Keeler
pictured on cover,
Absent,,,John. W. Metcalf,Catherine Young Glen,,,Popular Standard,Arthur P.
Schmidt,"Boston, MA",1899,Voice and Piano,G Major,,None,Yes,,
The Academy Two-Step,,,Barclay Walker,,,,Popular Standard,Carlin & Lennox,"Indianapolis,
IN",,Piano,F Major,,Composer,No,,
Ac-cent-tchu-ate the Positive,Mister In Between,"""Here Come the Waves"" Motion
Picture",Harold Arlen,Johnny Mercer,,,"Popular Standard, Movie
Selection",Edwin H. Morris & Co.,"New York, NY",1944,"Voice, Piano, Guitar",F
Major,,None,Yes,Bing Crosby and Betty Hutton pictured on cover,
Across the Alley From the Alamo,,,Joe Greene,,,,Popular Standard,Leslie Music
```

What type of file?



```
1000233   Miralda             John
1000234   Faley               Nick
1000235   Baylog              Cathy
1000236   Gallardo            Mike
1000237   Christian           Daniel
1000238   Baufield            Daniel
1000239   Frazier             Robert
1000240   Garrido             Edward
1000241   Williams            Zachary
1000242   Morel               David
          Padilla             Damian
1000244   Rosenberg           Wayne
1000245   Blanchard           Phong S
1000246   Wiggins             David
1000247   Miller              Jeffrey
1000248   Coon                Terry
1000249   Chretien            Walter
1000250   Myers               Timothy

1000233   Miralda             John
1000234   Faley               Nick
1000235   Baylog              Cathy
```

Types of flat files

Flat files will often end in file extensions like “.txt”, “.csv”, “.fwf”, and “.tsv”.

To figure out the structure of a flat file, start by opening it in a text editor. RStudio can also be used as a text editor to open and explore flat files (right click on the file name and then choose “Open With” and “RStudio”).