

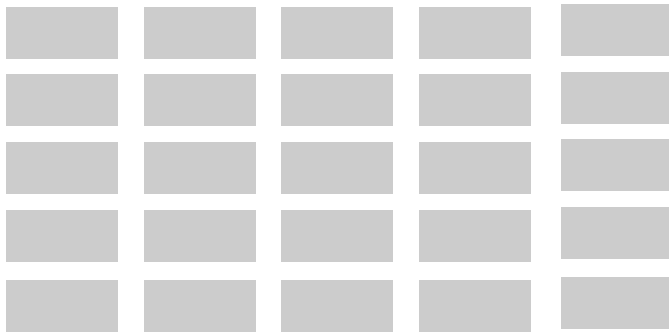
Data Tables (intro)

Stat 133 with Gaston Sanchez

Creative Commons Attribution Share-Alike 4.0 International CC BY-SA

Tabular vs Non-tabular

tabular



First part of the course

non-tabular



By the way

Data is ready to
be analyzed

Textbook / Classroom
monumental assumption

VS

Data requires
some processing

Rarely taught and/or
discussed

Tables

A Data Table

Name	Gender	Homeland	Birthyr	Mass	Height	Jedi
Luke	male	Tatooine	-99999	77kg	1.72m	yes
Leia	female	Alderaan	19BBY	49kg	1.50m	no
Obi-Wan	male	Stewjon	57BBY	77kg	1.82m	yes
Han	MALE	Corellia	29BBY	80kg	1.80m	no
Anakin	male	Tatooine	41.9BBY	84kg	1.88m	NA
Amidala	female	Naboo	46BBY	45kg	1.65m	no



Name	Gender	Homeland	Birthyr	Mass	Height	Jedi
Luke	male	Tatooine	-99999	77kg	1.72m	yes
Leia	female	Alderaan	19BBY	49kg	1.50m	no
Obi-Wan	male	Stewjon	57BBY	77kg	1.82m	yes
Han	MALE	Corellia	29BBY	80kg	1.80m	no
Anakin	male	Tatooine	41.9BBY	84kg	1.88m	NA
Amidala	female	Naboo	46BBY	45kg	1.65m	no

BBY: Before the Battle of Yavin

Tabular Data

Most data analysis methods (e.g. statistical, predictive models, machine learning, data mining, AI algorithms) require data in **tabular format**



Not necessarily the way data is collected, stored, organized, formatted

And even when data is in
a table format ...

Name	Gender	Homeland	Birthyr	Mass	Height	Jedi
Luke	male	Tatooine	-99999	77kg	1.72m	yes
Leia	female	Alderaan	19BBY	49kg	1.50m	no
Obi-Wan	male	Stewjon	57BBY	77kg	1.82m	yes
Han	MALE	Coreellia	29BBY	80kg	1.80m	no
Anakin	male	Tatooine	41.9BBY	84kg	1.88m	NA
Amidala	female	Naboo	46BBY	45kg	1.65m	no

there's still a lot of work to do
to have data ready to be analyzed

Some types of Tables

About “tables”

The term “Data Table” is a generic one

There are several flavors of tables depending on what the rows and columns represent

Let’s briefly review some of the most common types of tables that are found in statistical studies



Leia



Luke



Han

Heterogeneous data (most general type of table)

Name	Sex	Height	Weight
Luke	male	1.72	77
Leia	female	1.50	49
Han	male	1.80	80

Binary (e.g. presence-absence)

Which products do you buy?

	Beer	Wine	Juice	Coffee	Tea
Luke	1	1	1	0	0
Leia	0	1	1	0	1
Han	1	0	1	1	0

(1 = yes, 0 = no)

Modalities

How often do you drink the following products?

	Beer	Wine	Juice	Coffee	Tea
Luke	3	2	3	1	1
Leia	1	2	3	2	3
Han	3	2	2	3	1

(1=never, 2=sometimes, 3=always)

Preference

How much do you like the following juice flavors?

	Orange	Apple	Grapefruit	Carrot
Luke	3	5	2	1
Leia	1	4	5	2
Han	2	1	3	5

(1=don't like at all, 5=like it very much)

Cross-table (e.g. frequency tables)

Character	Episode IV	Episode V	Episode VI
Luke	254	128	112
Leia	57	114	56
Han	153	182	124
Vader	41	56	43
Emperor	0	5	39

Number of dialogues per character

Proximities, Dissimilarities, and Distances

	Luke	Leia	Han	Vader	Emperor
Luke	1	0.8	0.7	0.3	0.2
Leia	0.8	1	0.8	0.2	0.2
Han	0.7	0.8	1	0.1	0.1
Vader	0.3	0.2	0.1	1	0.9
Emperor	0.2	0.2	0.1	0.9	1

1 = perfect similarity, 0 = totally different

Tabular (rectangular) Data



Leia Organa
Female
1.50m tall



Luke Skywalker
Male
1.72m tall



Han Solo
Male
1.80m tall

How to store tables in a file?

name	gender	height
Leia Organa	female	1.50
Luke Skywalker	male	1.72
Han Solo	male	1.80



Analyst /Scientist

How to store data cells?

What type of format?

columns

rows

How do we “encode” cells in a file?


Plain text formats

Text files

A common way to store data in tabular form is via text files

Confusingly, people may refer to text files as:

- Plain text files
- Formatted text files
- Enriched text files



“Let’s take the term **text files** to mean a file that consists mainly of ASCII characters... and that uses newline characters to give humans the perception of lines”

Norman Matloff (2011)

The Art of R Programming

Plain Text Files

By “text files” we mean plain text files

Plain text as an umbrella term for any file that is in a human-readable form (.txt, .csv, .xml, .html, .md)

Text files stored as a sequence of characters

Each character stored as a single byte of data

Text files can be read and edited with a text editor (not to confuse with *word processor*)

Reading

Introduction to Data Technologies (ItDT) by Paul Murrell

<https://www.stat.auckland.ac.nz/~paul/ItDT/itdt-2010-11-01.pdf>

- Preface
- Chapter 1
- Chapter 5

So ...

How do you store data tables?

What type of formats can you use?

Plain Text Formats

There are 2 main subtypes of plain text formats, depending on how the separated values are identified in a row

1) Delimited formats

2) Fixed-width formats

Character Delimited Text

A common way to store data in tabular form is via text files

To store the data we need a way to separate data values

Each line represents a “row”

The idea of columns is conveyed with delimiters

Common Delimited Formats

Delimiter	Description
" "	White space
" , "	Comma
" \t "	Tab
" ; "	Semicolon

Consider the following data table

Name	Gender	Homeland	Born	Jedi
Anakin	male	Tatooine	41.9BBY	yes
Amidala	female	Naboo	46BBY	no
Luke	male	Tatooine	19BBY	yes
Leia	female	Alderaan	19BBY	no
Obi-Wan	male	Stewjon	57BBY	yes
Han	male	Corellia	29BBY	no
Palpatine	male	Naboo	82BBY	no
R2-D2	unknown	Naboo	33BBY	no

Tab delimited format

Each value separated by a tab `"\t"`

Typical file extensions: `txt`

Name	Gender	Homeworld	Born	Jedi
Anakin	male	Tatooine	41.9BBY	yes
Amidala	female	Naboo	46BBY	no
Luke	male	Tatooine	19BBY	yes
Leia	female	Alderaan	19BBY	no
Obi-Wan	male	Stewjon	57BBY	yes
Han	male	Corellia	29BBY	no
Palpatine	male	Naboo	82BBY	no
R2-D2	unknown	Naboo	33BBY	no

Space delimited format

Each value separated by a space " "

Typical file extensions: txt

```
Name Gender Homeworld Born Jedi
Anakin male Tatooine 41.9BBY yes
Amidala female Naboo 46BBY no
Luke male Tatooine 19BBY yes
Leia female Alderaan 19BBY no
Obi-Wan male Stewjon 57BBY yes
Han male Corellia 29BBY no
Palpatine male Naboo 82BBY no
R2-D2 unknown Naboo 33BBY no
```

Comma delimited format

Each value separated by a comma “,”

Typical file extensions: csv

```
Name,Gender,Homeworld,Born,Jedi
Anakin,male,Tatooine,41.9BBY,yes
Amidala,female,Naboo,46BBY,no
Luke,male,Tatooine,19BBY,yes
Leia,female,Alderaan,19BBY,no
Obi-Wan,male,Stewjon,57BBY,yes
Han,male,Corellia,29BBY,no
Palpatine,male,Naboo,82BBY,no
R2-D2,unknown,Naboo,33BBY,no
```

Semicolon (european) delimited format

Each value separated by a semicolon ";"

Typical file extensions: csv

```
Name;Gender;Homeworld;Born;Jedi
Anakin;male;Tatooine;41.9BBY;yes
Amidala;female;Naboo;46BBY;no
Luke;male;Tatooine;19BBY;yes
Leia;female;Alderaan;19BBY;no
Obi-Wan;male;Stewjon;57BBY;yes
Han;male;Corellia;29BBY;no
Palpatine;male;Naboo;82BBY;no
R2-D2;unknown;Naboo;33BBY;no
```

Other characters as delimiters

Each value separated by a bar `"|"`

You can make up your own extension: *bar*

```
Name | Gender | Homeworld | Born | Jedi
Anakin | male | Tatooine | 41.9BBY | yes
Amidala | female | Naboo | 46BBY | no
Luke | male | Tatooine | 19BBY | yes
Leia | female | Alderaan | 19BBY | no
Obi-Wan | male | Stewjon | 57BBY | yes
Han | male | Corellia | 29BBY | no
Palpatine | male | Naboo | 82BBY | no
R2-D2 | unknown | Naboo | 33BBY | no
```

CSV files

The Comma-Separated Value (CSV) format is a special case of a plain text format.

Although not a formal standard, CSV files are very common and are a quite reliable plain text delimited format that at least solves the problem of where the fields are in each row of the file.

CSV files are a common way to transfer data from a spreadsheet to other software

What if a comma is part
of a field?

Commas in fields

Name	Gender	Homeland	Born
Skywalker, Anakin	male	Tatooine	41.9BBY
Amidala, Padme	female	Naboo	46BBY
Skywalker, Luke	male	Tatooine	19BBY
Organa, Leia	female	Alderaan	19BBY
Kenobi, Obi-Wan	male	Stewjon	57BBY
Solo, Han	male	Corellia	29BBY

Commas within a comma delimited format

Commas that are part of a field must be surrounded by quotes ", "

```
Name , Gender , Homeworld , Born , Jedi  
"Skywalker, Anakin", male , Tatooine , 41.9BBY  
"Amidala, Padme", female , Naboo , 46BBY  
"Skywalker, Luke", male , Tatooine , 19BBY  
"Organa, Leia", female , Alderaan , 19BBY  
"Kenobi, Obi-Wan", male , Stewjon , 57BBY  
"Solo, Han", male , Corellia , 29BBY
```

There's also fixed-width

Fixed-width format

Each value separated with a pre-specified width (in characters)

Name	Gender	Homeworld	Born	Jedi
Anakin	male	Tatooine	41.9BBY	yes
Amidala	female	Naboo	46BBY	no
Luke	male	Tatooine	19BBY	yes
Leia	female	Alderaan	19BBY	no
Obi-Wan	male	Stewjon	57BBY	yes
Han	male	Corellia	29BBY	no
Palpatine	male	Naboo	82BBY	no
R2-D2	unknown	Naboo	33BBY	no

Fixed-width format

Each value separated with a pre-specified width (in characters)

Name	Gender	Homeworld	Born	Jedi
Anakin	male	Tatooine	41.9BBY	yes
Amidala	female	Naboo	46BBY	no
Luke	male	Tatooine	19BBY	yes
Leia	female	Alderaan	19BBY	no
Obi-Wan	male	Stewjon	57BBY	yes
Han	male	Corellia	29BBY	no
Palpatine	male	Naboo	82BBY	no
R2-D2	unknown	Naboo	33BBY	no

10

8

10

8

4

Pros and Cons of data tables

Advantages

Simplicity

Common formats (csv, tsv, txt, dat, *etc*)

Can be opened and modified with a text editor

Can also be opened in spreadsheet software

Easy to understand for most users

Can be read in data analysis software

Disadvantages

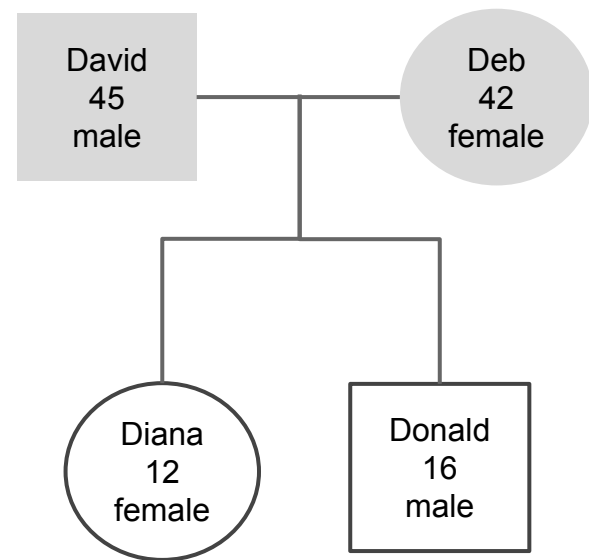
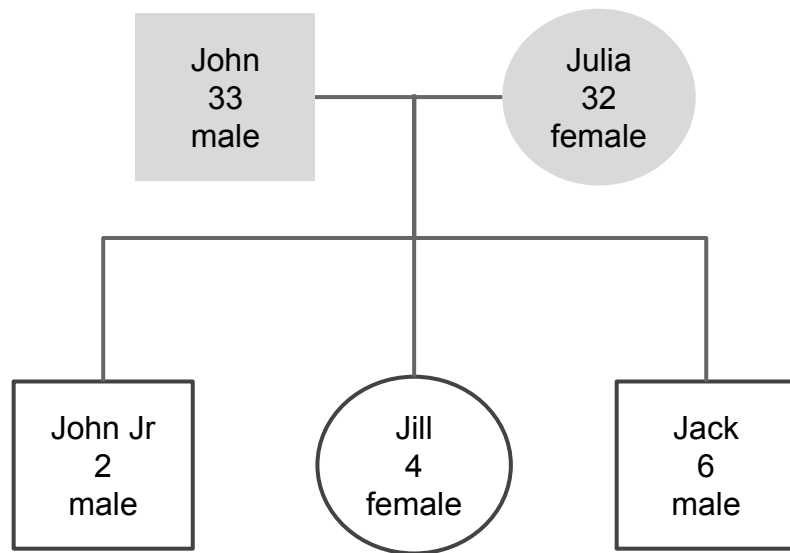
Not good for hierarchical or nested structures

Values may not be well self-described

Difficult to include metadata

Long column names are cumbersome

Hierarchical Data



Father	Mother	Name	Age	Sex
		John	33	male
		Julia	32	female
John	Julia	Jack	6	male
John	Julia	Jill	4	female
John	Julia	John Jr	2	male
		David	45	male
		Debbie	42	female
David	Debbie	Donald	16	male
David	Debbie	Dianne	12	female