INFO101: Tabular Data

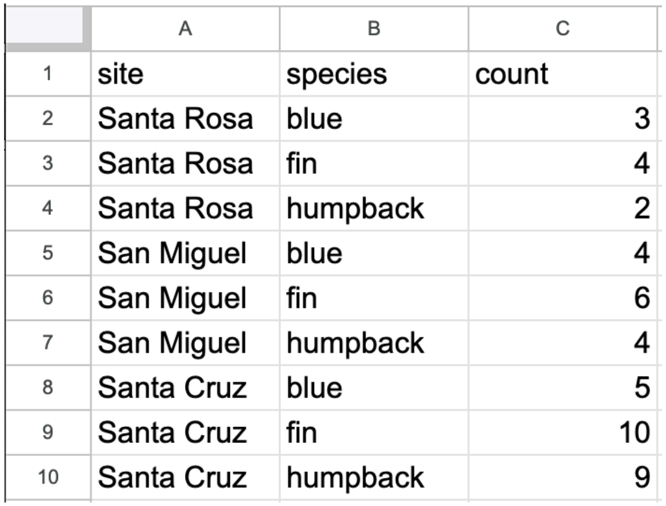
What makes data tidy?

MARINCS 100B | Intro to Marine Data Science | Winter 2025

Key concepts

1) Make it a rectangle
2) Dont confuse the computer
3) Consistant names and forma

Make it a rectangle

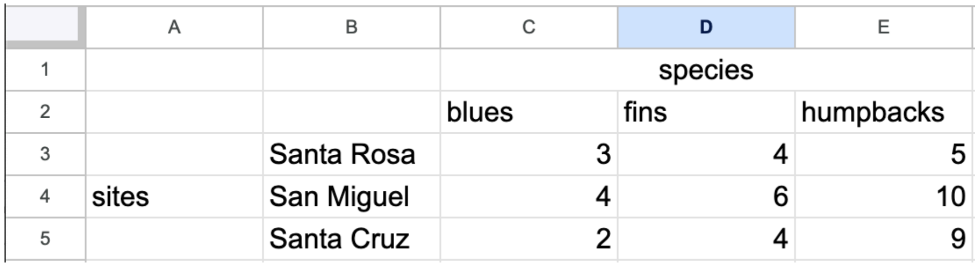


	A	B	C
1	site	species	count
2	Santa Rosa	blue	3
3	Santa Rosa	fin	4
4	Santa Rosa	humpback	2
5	San Miguel	blue	4
6	San Miguel	fin	6
7	San Miguel	humpback	4
8	Santa Cruz	blue	5
9	Santa Cruz	fin	10
10	Santa Cruz	humpback	9

One row per
observation
(left-right)

One column per
imformation type
(up and down)

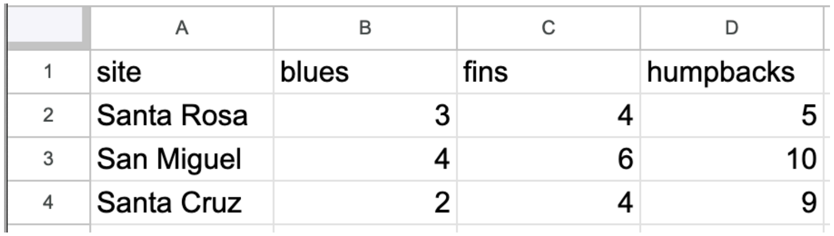
Non-rectangular examples



	A	B	C	D	E
1			species		
2			blues	fins	humpbacks
3		Santa Rosa	3	4	5
4	sites	San Miguel	4	6	10
5		Santa Cruz	2	4	9

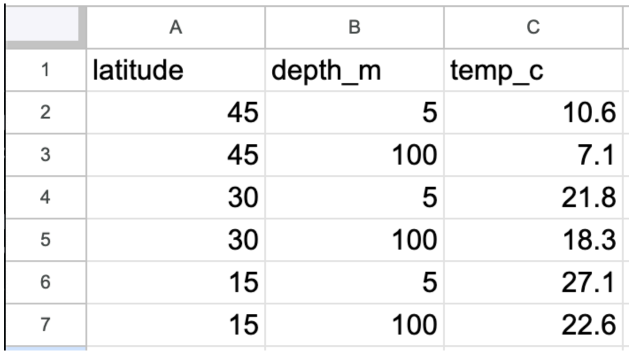
Multiple lines
of headers

Is a rectangle but does not follow one column per variable. In this case the whale species is one variable and “species” should be the header of the column.



	A	B	C	D
1	site	blues	fins	humpbacks
2	Santa Rosa	3	4	5
3	San Miguel	4	6	10
4	Santa Cruz	2	4	9

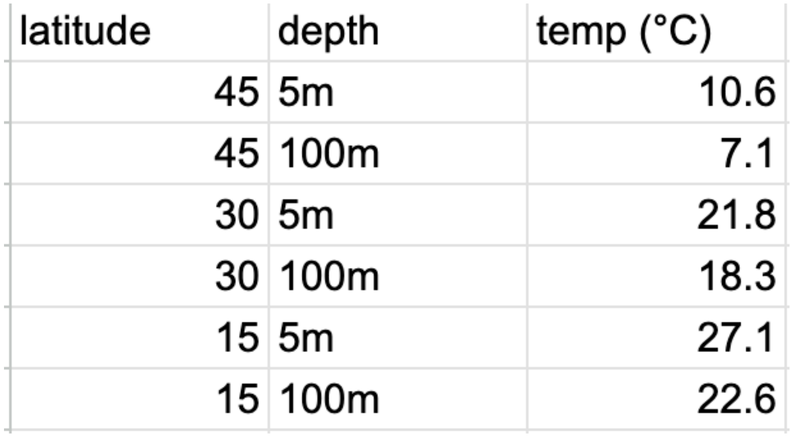
Don’t confuse the computer



	A	B	C
1	latitude	depth_m	temp_c
2	45	5	10.6
3	45	100	7.1
4	30	5	21.8
5	30	100	18.3
6	15	5	27.1
7	15	100	22.6

1) column names look like variable names
2) cells contain one value of one type of data

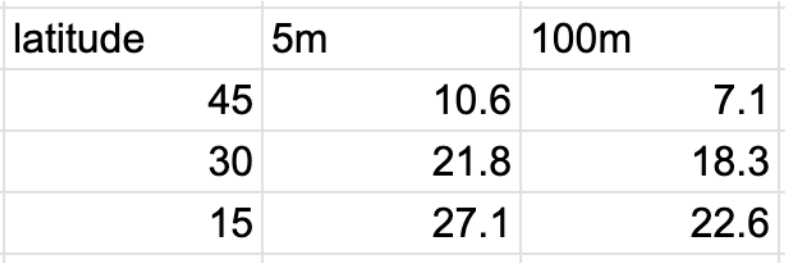
Confusing examples



latitude	depth	temp (°C)
45	5m	10.6
45	100m	7.1
30	5m	21.8
30	100m	18.3
15	5m	27.1
15	100m	22.6

Temp column: computer wont understand header. There is a space, parenthesis, and a degree symbol. Remeber, headers should be in the same format as if you were naming a function in R

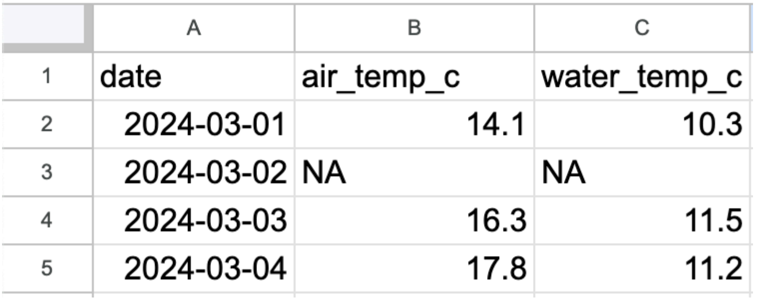
Depth column: has a number and a letter (Ex: 100m) the computer does not know what this means. Should only contain number



latitude	5m	100m
45	10.6	7.1
30	21.8	18.3
15	27.1	22.6

Computers dont like seeing the header with a number in it as the number is seen as a variable and not as a name. wide-format data

Consistent names and formats



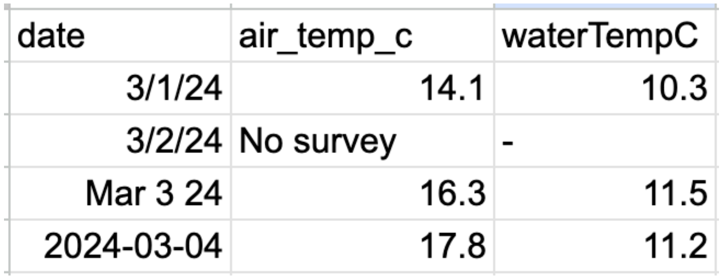
	A	B	C
1	date	air_temp_c	water_temp_c
2	2024-03-01	14.1	10.3
3	2024-03-02	NA	NA
4	2024-03-03	16.3	11.5
5	2024-03-04	17.8	11.2

1) Want column names to be readable with consistant formating

2)Dates, etc. should follow universal conventions

3) missing values are clearly indicated(NA)

Inconsistent examples



date	air_temp_c	waterTempC
3/1/24	14.1	10.3
3/2/24	No survey	-
Mar 3 24	16.3	11.5
2024-03-04	17.8	11.2

Formating is not consistant with the headers.

Missing values: one says “no survey” while one says “-“

Dates are not consistant. Should be formated with year, month, day. Ex: 2024-03-04

Recap

1) make it a rectangle

2) dont confuse the
computer

3) consistant names and formats

New vocabulary and lingering questions

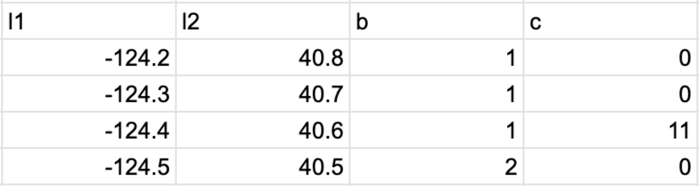
New vocabulary

Snake-case
Camel-case
Wide-format

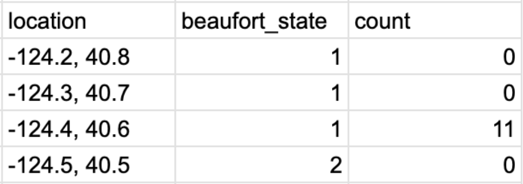
Lingering questions

Exercises

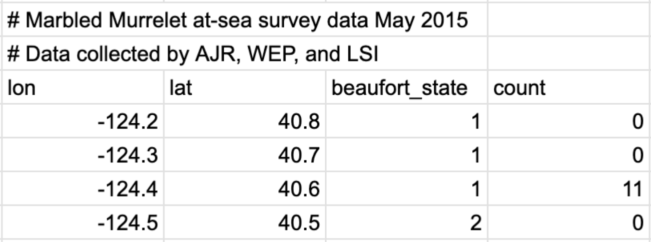
Match the tables to the tidy rule they violate



l1	l2	b	c
-124.2	40.8	1	0
-124.3	40.7	1	0
-124.4	40.6	1	11
-124.5	40.5	2	0



location	beaufort_state	count
-124.2, 40.8	1	0
-124.3, 40.7	1	0
-124.4, 40.6	1	11
-124.5, 40.5	2	0



# Marbled Murrelet at-sea survey data May 2015			
# Data collected by AJR, WEP, and LSI			
lon	lat	beaufort_state	count
-124.2	40.8	1	0
-124.3	40.7	1	0
-124.4	40.6	1	11
-124.5	40.5	2	0

Rule 1 - make it a rectangle

Table 3

Rule 2 - don’t confuse the computer

Table 2

Rule 3 - use consistent names and formats

Table 1

INFO101: Tabular Data

Creating and importing data frames in R

MARINCS 100B | Intro to Marine Data Science | Winter 2025

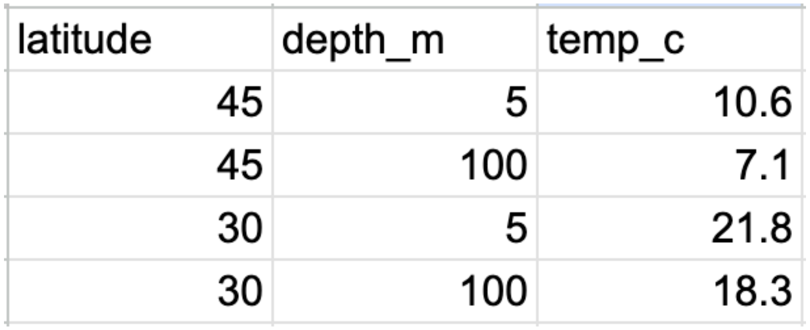
Key concepts

1) “Data frames” workhorses of data science

2) DFs are 2-D with rows and columns

3) create data frames manually, more often we’ll import from file

Two views, same data



latitude	depth_m	temp_c
45	5	10.6
45	100	7.1
30	5	21.8
30	100	18.3

Spreadsheet
software view

Creating a data frame

```
# How to create a data frame manually
noaa_survey <- data. frame(latitude = c(45, 45, 30, 30),
depth_m = c(5, 100, 5, 100),
temp_c = c(10.6, 7.1, 21.8,))
```

This is how you format a table in R.
Similar how we wrote fuctions, We have
the functuion(data. frame) the parameter
name (Latitude) and the colunm values.

Demo in R

How to create
and import
data frames

New vocabulary and lingering questions

New vocabulary

csv.file
dir()

Lingering questions

Exercises

Complete the exercises in exercises/exercises101b.R

INFO101: Tabular Data

Indexing data frames

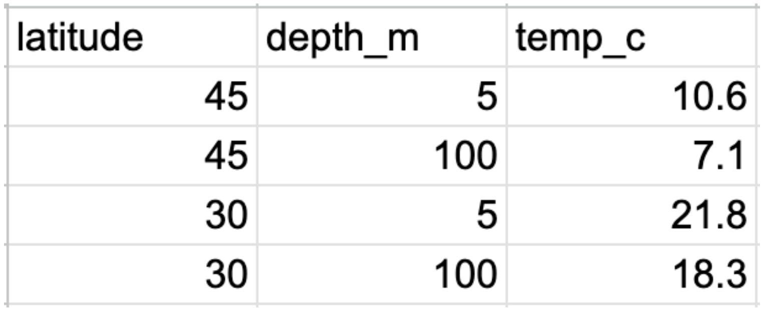
MARINCS 100B | Intro to Marine Data Science | Winter 2025

Key concepts

1) Index with [ ]
2) But 2-D -> [r, c]

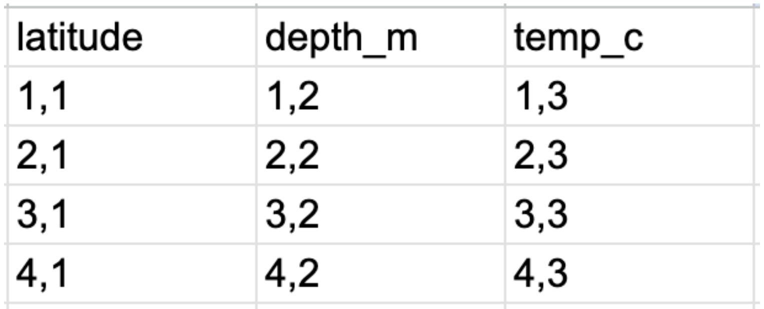
How to index into data frames

noaa_survey



latitude	depth_m	temp_c
45	5	10.6
45	100	7.1
30	5	21.8
30	100	18.3

index-> cell



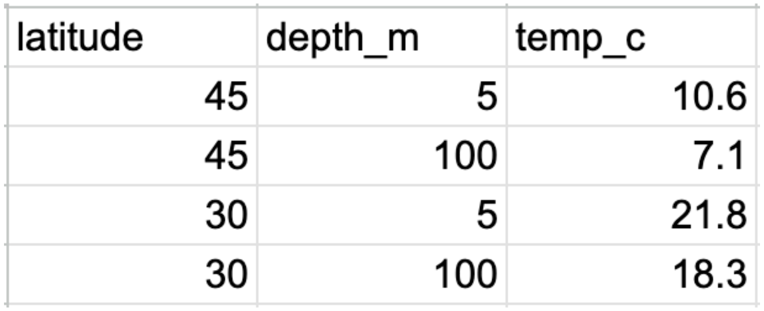
latitude	depth_m	temp_c
1,1	1,2	1,3
2,1	2,2	2,3
3,1	3,2	3,3
4,1	4,2	4,3

noaa_survey [1,1] = first row and first column
[2, 2:3] = second row and second and third columns
[3:4, 2:3] = third row, fourth column and second row, third column

noaa_survey[4,1] <- 50 = chaneges the values of fourth row, column one to the value of 50

Pull rows and columns from data frames

noaa_survey



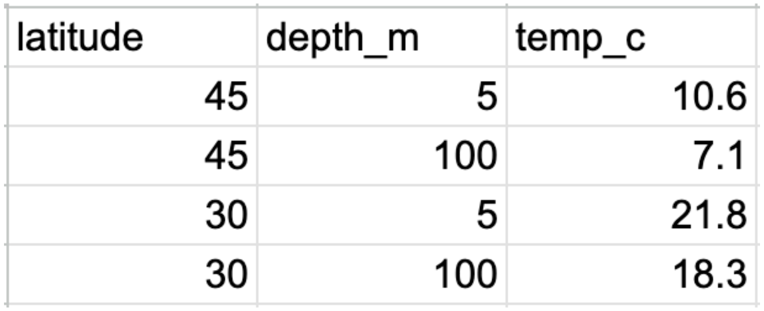
latitude	depth_m	temp_c
45	5	10.6
45	100	7.1
30	5	21.8
30	100	18.3

if you want entire row = noaa_survey[1,]
If you want entire column = noaa_survey[,1]

noaa_survey$latitude

Filtering rows

noaa_survey



latitude	depth_m	temp_c
45	5	10.6
45	100	7.1
30	5	21.8
30	100	18.3

```
noaa_survey[noaa_survey$latitude==45,]
```

this line of code gives us the rows where
latitude = 45

MAKE SURE TO ADD COMA

New vocabulary and lingering questions

New vocabulary

```
noaa_survey[noaa_survey$latitude==45,]
```

index with [ ]

Lingering questions

Exercises

Complete the exercises in exercises/exercises101c.R