Tidy and Untidy Data

Vik Gopal
Dept. of Stats and Applied Prob.
Faculty of Science
NUS

2019-10-23

Outline

- What is tidy data?
- Why tidy data?
- Always tidy data?
- Dirty data.

What's in a Dataset?

- Think of a dataset as a collection of **values**.
- Every value is associated with a **variable**, and an **observation**.
- A variable consists of all values that measure the same underlying attribute.
- An observation consists of all values measured on the same unit.

What is Tidy Data?

- The term was coined by Hadley Wickham, in his 2014 paper "Tidy Data".
- It describes a dataset that follows these principles:
  - Each variable forms a column.
  - Each observation forms a row.
  - Each type of observational unit forms a table.

An Example

Consider the following dataset:

	Treatment A	Treatment B
John Smith	-	2
Jane Doe	16	11
Mary Johnson	3	1

An Example

cont'd

There are three variables:

- person
- treatment
- result (value)

Each observation unit corresponds to a person-treatment combination.

An Example

tidy version

person	treatment	result
John Smith	Treatment A	-
Jane Doe	Treatment A	16
Mary Johnson	Treatment A	3
John Smith	Treatment B	2
Jane Doe	Treatment B	11
Mary Johnson	Treatment B	1

Why Tidy Data?

- It works well in R.
- R is a environment for statistical computing.
- There is a set of libraries within R that provide excellent tools for data manipulation.
- They provide a vocabulary of verbs that you can combine to achieve the task that you want.
  - select, filter, arrange, group by, mutate, summarise
- The graphics engine in R requires the data to be in tidy format.
- The modeling functions in R too.

Why Tidy Data?

cont'd

Tidy vs. Untidy Data

- Tidy data is not the holy grail of formats.
- Sometimes, it is not the right choice. Other formats might be more efficient.
- For instance:
  - when working with large corpora of text documents.
  - when working with graph structures.
- Take a look at this page for some more examples: https://simplystatistics.org/2016/02/17/non-tidy-data/

Tidy vs. Dirty Data

Consider the following data on restaurant grades in New York City.

First 5 of 24000 rows:

building	x_coord	y_coord	street	zipcode	borough	cuisine	name
1007	-73.8	40.8	Morris Park Ave	10462	Bronx	Bakery	Morris Park Bake Shop
469	-73.9	40.6	Flatbush Avenue	11225	Brooklyn	Hamburgers	Wendy'S
351	-73.9	40.7	West 57 Street	10019	Manhattan	Irish	Dj Reynolds Pub And Restaurant
2780	-73.9	40.5	Stillwell Avenue	11224	Brooklyn	American	Riviera Caterer
97-22	-73.8	40.7	63 Road	11374	Queens	Jewish/Kosher	Tov Kosher Kitchen

First 5 of 93000 rows:

date	grade	score	id
2014-03-03 08:00:00	A	2	30075445
2013-09-11 08:00:00	A	6	30075445
2013-01-24 08:00:00	A	10	30075445
2011-11-23 08:00:00	A	9	30075445
2011-03-10 08:00:00	B	14	30075445
2014-12-30 08:00:00	A	8	30112340

In how many different ways could it be wrong?

Tidy vs. Dirty Data

cont'd

- Wrong lat/long.
- Missing zipcodes. How to impute?
- NULL grades specified in two different ways.
- Street names spelt with 'First' and '1st'.
- Same restaurant spelt in different ways.

Summary

- Consider converting your data to tidy format.
- It is useful when you intend to analyse/plot/summarise your data in different ways.
- R is not the only tool for data manipulation. You can also use Python.
- In my opinion, cleaning dirty data requires creativity, perseverance, and the ability to check our own biases.