Tidy data

The bread and butter of every data scientist

Paola Suárez

PyLadies Hamburg | 08.March.2021

About me

- i like to bake bread
- trained as a scientist
- data consultant - *Charité Universitätsmedizin Berlin*
- developer & data scientist - *ThoughtWorks*
- @psrmx

Science dog, the internet.

My experience with data

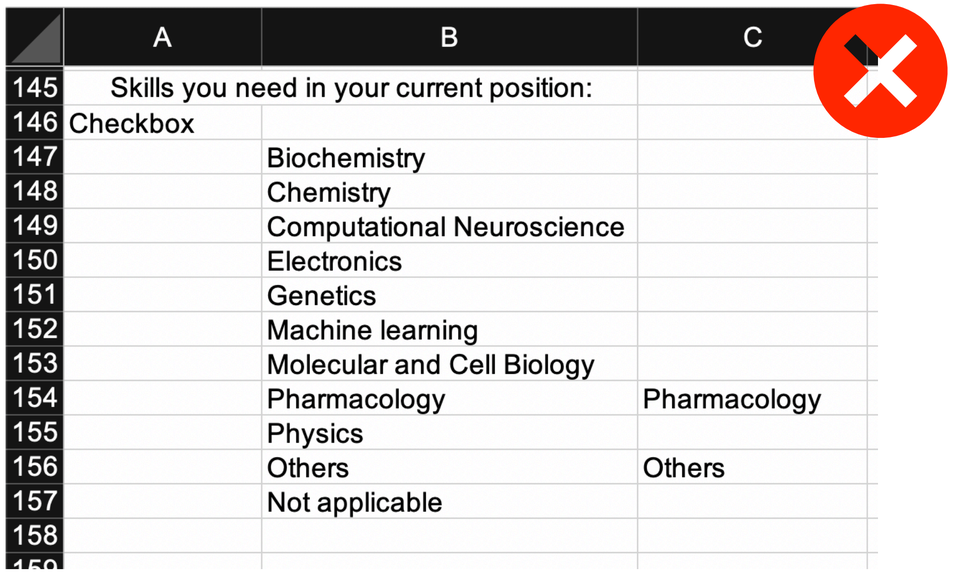
- analysis of satisfaction surveys by the university's alumni
  - goal - to prepare the students for their work life

Please indicate your level of agreement with the following statements:

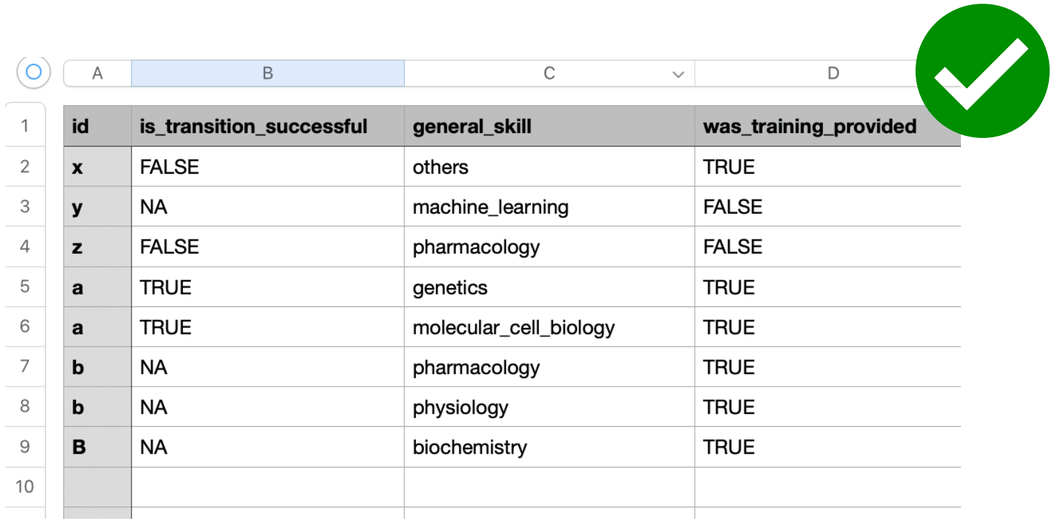
	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Store hours are appropriate for my shopping needs	○	○	○	○	○
Store atmosphere and decor are appealing	○	○	○	○	○
Merchandise sold is of the highest quality	○	○	○	○	○
The merchandise sold is a good value for the money	○	○	○	○	○
Advertised merchandise in stock	○	○	○	○	○
I am very satisfied with the merchandise I purchased	○	○	○	○	○

	A	B	C
145	Skills you need in your current position:		
146	Checkbox		
147		Biochemistry	
148		Chemistry	
149		Computational Neuroscience	
150		Electronics	
151		Genetics	
152		Machine learning	
153		Molecular and Cell Biology	
154		Pharmacology	Pharmacology
155		Physics	
156		Others	Others
157		Not applicable	
158			
159			

Time to wrangle with data



	A	B	C
145	Skills you need in your current position:		
146	Checkbox		
147		Biochemistry	
148		Chemistry	
149		Computational Neuroscience	
150		Electronics	
151		Genetics	
152		Machine learning	
153		Molecular and Cell Biology	
154		Pharmacology	Pharmacology
155		Physics	
156		Others	Others
157		Not applicable	
158			



	A	B	C	D
1	**id**	**is_transition_successful**	**general_skill**	**was_training_provided**
2	**x**	FALSE	others	TRUE
3	**y**	NA	machine_learning	FALSE
4	**z**	FALSE	pharmacology	FALSE
5	**a**	TRUE	genetics	TRUE
6	**a**	TRUE	molecular_cell_biology	TRUE
7	**b**	NA	pharmacology	TRUE
8	**b**	NA	physiology	TRUE
9	**B**	NA	biochemistry	TRUE
10				

- identify your variables
- record one observation per row
- assign an identifier key to each observation

Then, I realised that…

> “Tidy datasets are all alike, but every messy dataset is messy in its own way.”
>
> - Hadley Wickham

What is tidy data?

- meaning of dataset ⟶ structure
- 3 principles:
    1. each variable is a column
    2. each observation is a row
    3. each type of observational unit is a table
- aka Codd’s 3rd normal form

Tidy Data, Hadley Wickham.

Cat tidying up, the internet.

Use case - distance matrix

intercity distance (km)

	berlin	hamburg	cologne	stuttgart	munich
berlin	0	286	573	629	585
hamburg	286	0	425	655	778
cologne	573	425	0	366	575
stuttgart	629	655	366	0	233
munich	585	778	575	233	0

	start	end	distance
0	berlin	hamburg	286
1	berlin	cologne	573
2	berlin	stuttgart	629
3	berlin	munich	585
4	hamburg	cologne	425
5	hamburg	stuttgart	655
6	hamburg	munich	778
7	cologne	stuttgart	366
8	cologne	munich	575
9	stuttgart	munich	233

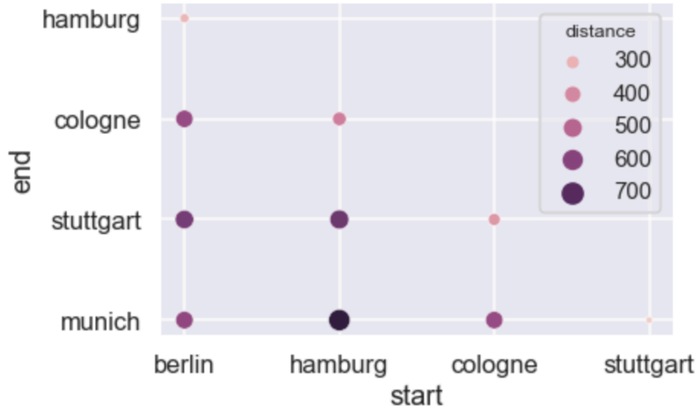
Use case - distance matrix

```
df["distance"].min()
```

```
233
```

```
df.groupby(by="end").agg(sum)
```

end	distance
cologne	998
hamburg	286
munich	2171
stuttgart	1650



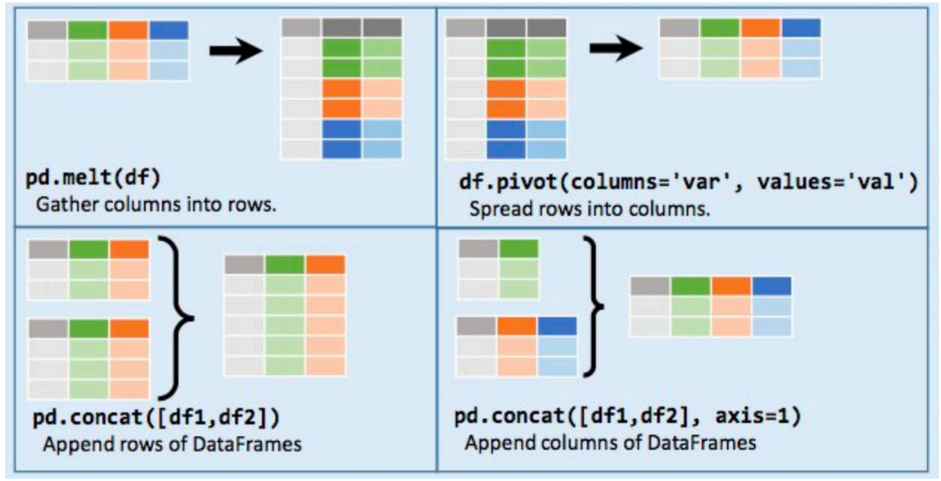
- sort, filter, join, append new observations, new variables, visualise, …

How can we achieve tidy data?

- nested for loops & conditionals
- fancy pandas functions:
  - **melt** (stack columns in rows)
  - **pivot** (unstack rows in columns)
  - **concat**
  - **merge**, **joins**

```
starts = []
ends = []
distances = []


(nrows, ncols) = df.shape
for r in range(nrows):
    for c in range(ncols):
        this_observation = df.iloc[r, c]
        if this_observation not in distances:
            starts.append(df.index[r])
            ends.append(df.columns[c])
            distances.append(this_observation)
```



NumFocus, Data wrangling with pandas cheat sheet.

Why do we like tidy data?

- it makes your life easier!
- saves time & effort
- data cleaning becomes a joy
- unlocks analysis & visualisation powers

- with great power, comes great responsibility

Superhero dogs, the internet.

Call to action ✊🏾

In my country (population of ~126M),
10 women are murdered each **day**.

In the UK (population of ~68M),
13 women are murdered each **month**.

Feminicidios en México, El País.

Femicide Census in UK, End Violence Against Women.

Andrea Murcia (@Usagii_ko), Resistir para existir.

Cute cat saying thank you, the internet.

IWD logo, the internet.

PyLadies Hamburg | 08.March.2021