How to Organize Data

Learning how to organize a dataset is one of the most important skills you should leave this class with. The notes you read for today focused on the mechanics of doing data entry.

Here we will look at some of the higher-level concepts related to collecting data.

The central concept for understanding how to organize data is to know the **unit of observation**, the thing that one particular observation represents.

An Example

Consider the following dataset showing four female tennis players. Notice that the last column is trying to put multiple pieces of information in one value.

	A	B	C
1	**player**	**nationality**	**majors_won**
2	Ashleigh Barty	AUS	Wimbledon|French Open
3	Emma Raducanu	GBR	US Open
4	Barbora Krejčíková	CZE	French Open
5	Naomi Osaka	JPN	Australian Open|US Open

unit of observation: **Player**

First Normal Form (1NF)

Here is an alternative that only has one piece of information in each cell. This table is said to be in the First Normal Form (1NF).

1NF requires that the data be in a tabular format with only one individual pieces of information in each cell.

	A	B	C	D	E	F
1	**player**	**nationality**	**won_wimbledon**	**won_us_open**	**won_french_open**	**won_australian_open**
2	Ashleigh Barty	AUS	1	0	1	0
3	Emma Raducanu	GBR	0	1	0	0
4	Barbora Krejčíková	CZE	0	0	1	0
5	Naomi Osaka	JPN	1	0	0	1

unit of observation: **Player**

Another Example

Here is another example of a similar dataset. This one gives information about each Grand Slam itself. It is in 1NF but notice that it duplicates information.

	A	B	C	D	E
1	**tournament**	**year**	**tournament_country**	**winner**	**winner_nationality**
2	Australian Open	2020	AUS	Sofia Kenin	USA
3	French Open	2020	FRA	Iga Świątek	POL
4	Wimbledon	2020	GBR		
5	US Open	2020	USA	Naomi Osaka	JPN
6	Australian Open	2021	AUS	Naomi Osaka	JPN
7	French Open	2021	FRA	Barbora Krejčíková	CZE
8	Wimbledon	2021	GBR	Ashleigh Barty	AUS
9	US Open	2021	USA	Emma Raducanu	GBR

unit of observation: **tournament x year**

Third Normal Form (3NF)

We can fix the problem of duplication by creating three different tables. Now we only have information that is directly about the unit of observation on each table with no duplication.

These tables are in Third Normal Form (3NF).

	A	B	C
1	**tournament**	**year**	**winner**
2	Australian Open	2020	Sofia Kenin
3	French Open	2020	Iga Świątek
4	Wimbledon	2020	
5	US Open	2020	Naomi Osaka
6	Australian Open	2021	Naomi Osaka
7	French Open	2021	Barbora Krejčíková
8	Wimbledon	2021	Ashleigh Barty
9	US Open	2021	Emma Raducanu

unit of observation: **tournament x year**

	A	B
1	**tournament**	**tournament_country**
2	Australian Open	AUS
3	French Open	FRA
4	Wimbledon	GBR
5	US Open	USA

unit of observation: **tournament**

	A	B
1	**player**	**nationality**
2	Sofia Kenin	USA
3	Iga Świątek	POL
4	Naomi Osaka	JPN
5	Barbora Krejčíková	CZE
6	Ashleigh Barty	AUS
7	Emma Raducanu	GBR

unit of observation: **player**

Second Normal Form?

You may wonder why we skipped 2NF. It does exist as an intermediate form. The type of duplication in the **tournament_country** is called a functional dependency whereas the duplication in **winner_nationality** is a transitive dependency. 2NF only requires that we remove the first type.

	A	B	C	D	E
1	**tournament**	**year**	**tournament_country**	**winner**	**winner_nationality**
2	Australian Open	2020	AUS	Sofia Kenin	USA
3	French Open	2020	FRA	Iga Świątek	POL
4	Wimbledon	2020	GBR		
5	US Open	2020	USA	Naomi Osaka	JPN
6	Australian Open	2021	AUS	Naomi Osaka	JPN
7	French Open	2021	FRA	Barbora Krejčíková	CZE
8	Wimbledon	2021	GBR	Ashleigh Barty	AUS
9	US Open	2021	USA	Emma Raducanu	GBR

unit of observation: **tournament x year**

Wide vs. Long

Consider the two following datasets, which contain the same information. Both are in 3NF, but have different units of observation. The second format is longer than the first; the first is wider than the second. We will come back to this concept soon.

	A	B	C	D	E	F
1	**player**	**nationality**	**won_wimbledon**	**won_us_open**	**won_french_open**	**won_australian_open**
2	Ashleigh Barty	AUS	1	0	1	0
3	Emma Raducanu	GBR	0	1	0	0
4	Barbora Krejčíková	CZE	0	0	1	0
5	Naomi Osaka	JPN	1	0	0	1

unit of observation: **player**

	A	B	C
1	**player**	**nationality**	**grand_slam**
2	Ashleigh Barty	AUS	US Open
3	Ashleigh Barty	AUS	French Open
4	Emma Raducanu	GBR	US Open
5	Barbora Krejčíková	CZE	French Open
6	Naomi Osaka	JPN	Australian Open
7	Naomi Osaka	JPN	US Open

unit of observation: **tournament x player**

How to Organize Data

Here are my general rules for how to collect data in terms of these concepts:

Always respect 1NF when collecting data.

Following 3NF is nice, but you can break this in the interest of simplicity, particularly when doing the initial data collection.

Try to think of how you can make the dataset longer rather than wider.

In the upcoming classes we will see how two-table and pivot data verbs help us deal with these concepts.