Data Antipatterns

An antipattern is an example of something that you should not do. In the notes you saw several examples of how to collect and organize data. Here, we will see several examples of things you should strongly avoid doing in your own work ...

Antipattern #1

What is the issue here and how could we fix it?

	A	B	C	D	E	F
1		**country**	**party**	**birth_year**	**year_started**	
2	Sanna Marin	Finland	Social Democrats	1985	2019	
3	Élisabeth Borne	France	LREM	1961	2022	
4	Κυριάκος Μητσοτάκης	Greece	New Democracy	1968	2019	
5	Katrín Jakobsdóttir	Iceland	Left-Green	1976	2017	
6	Kaja Kallas	Estonia	Reform Party	1977	2021	
7						
8						
9						

Antipattern #1 : Solution

Problem: The first column does not have a name.

Solution: Add column name.

	A	B	C	D	E	F
1	**name**	**country**	**party**	**birth_year**	**year_started**	
2	Sanna Marin	Finland	Social Democrats	1985	2019	
3	Élisabeth Borne	France	LREM	1961	2022	
4	Κυριάκος Μητσοτάκης	Greece	New Democracy	1968	2019	
5	Katrín Jakobsdóttir	Iceland	Left-Green	1976	2017	
6	Kaja Kallas	Estonia	Reform Party	1977	2021	
7						
8						
9						

Antipattern #2

What is the issue here and how could we fix it?

	A	B	C	D	E	F
1	**name**	**country**	**party**	**birth_year**	**year_started**	
2	Sanna Marin	Finland	Social Democrats	1985	2019	
3	Élisabeth Borne	France	LREM	1961	2022	
4	Κυριάκος Μητσοτάκης	Greece	New Democracy	1968	2019	
5	Katrín Jakobsdóttir	Iceland	Left-Green	1976*	2017	
6	Kaja Kallas	Estonia	Reform Party	1977	2021	
7						
8						
9						

Antipattern #2 : Solution

Problem: A numeric variable has a non-numeric element (here a *).
Solution: Include a notes column to contain the special information.

	A	B	C	D	E	F
1	**name**	**country**	**party**	**birth_year**	**year_started**	**notes**
2	Sanna Marin	Finland	Social Democrats	1985	2019	
3	Élisabeth Borne	France	LREM	1961	2022	
4	Κυριάκος Μητσοτάκης	Greece	New Democracy	1968	2019	
5	Katrín Jakobsdóttir	Iceland	Left-Green	1976	2017	*
6	Kaja Kallas	Estonia	Reform Party	1977	2021	
7						
8						
9						

Antipattern #3

What is the issue here and how could we fix it?

	A	B	C	D	E	F	G	H
1								
2		Trial 1						
3		ALPHA	BETA	GAMMA	DELTA	EPSILON	ZETA	
4	A	37	78	62	50	87	97	
5	B	88	39	11	95	5	36	
6	C	67	16	30	93	5	13	
7	D	91	10	90	59	93	67	
8	E	6	63	38	83	22	97	
9	F	10	51	48	32	14	51	
10	G	8	94	73	86	16	27	
11								
12		Trial 2						
13		ALPHA	BETA	GAMMA	DELTA	EPSILON	ZETA	
14	A	92	52	11	71	100	55	
15	B	65	16	61	79	89	17	
16	C	14	46	67	93	78	90	
17	D	81	6	8	21	45	74	
18	E	43	12	27	54	15	81	
19	F	6	84	59	66	11	86	
20	G	54	9	67	24	65	17	
21								
22								

Antipattern #3 : Solution

Problem: Two tables on one sheet.

Solution: Combine into a single table, adding a column for the trial.

	A	B	C	D	E	F	G	H	I
1	latin	trial	alpha	beta	gamma	delta	epsilon	zeta	
2	A	1	37	78	62	50	87	97	
3	B	1	88	39	11	95	5	36	
4	C	1	67	16	30	93	5	13	
5	D	1	91	10	90	59	93	67	
6	E	1	6	63	38	83	22	97	
7	F	1	10	51	48	32	14	51	
8	G	1	8	94	73	86	16	27	
9	A	2	92	52	11	71	100	55	
10	B	2	65	16	61	79	89	17	
11	C	2	14	46	67	93	78	90	
12	D	2	81	6	8	21	45	74	
13	E	2	43	12	27	54	15	81	
14	F	2	6	84	59	66	11	86	
15	G	2	54	9	67	24	65	17	
16									
17									
18									
19									

Antipattern #3 : Solution

Another Solution: Create a longer format for the data with one measurment on each row.

Note: This is format is preferred for reasons we will see when learning about data pivots in a future set of notes.

	A	B	C	D	E
1	**trial**	**letter**	**greek**	**number**	
2	1	A	ALPHA	37	
3	1	A	BETA	78	
4	1	A	GAMMA	62	
5	1	A	DELTA	50	
6	1	A	EPSILON	87	
7	1	A	ZETA	97	
8	1	B	ALPHA	88	
9	1	B	BETA	39	
10	1	B	GAMMA	11	
11	1	B	DELTA	95	
12	1	B	EPSILON	5	
13	1	B	ZETA	36	
14	1	C	ALPHA	67	
15	1	C	BETA	16	
16	1	C	GAMMA	30	
17	1	C	DELTA	93	
18	1	C	EPSILON	5	
19	1	C	ZETA	13	

Antipattern #4

What is the issue here and how could we fix it?

	A	B	C	D	E	F	G	H
1		fall_semester			spring_semester			
2	**student_name**	**dsst289**	**math211**	**rhcs103**	**dsst389**	**math212**	**rhcs104**	
3	Sally	79	75	93	95	100	77	
4	Bob	84	77	80	97	82	88	
5	Jill	77	72	87	78	71	98	
6	Jack	91	98	91	78	99	98	
7	Mary	86	85	88	84	92	80	
8								
9								
10								

Antipattern #4 : Solution

Problem: Hierarchical column names.
Solution: Spread data into more rows and fewer columns so that the metadata about the original features become full features in the new data.

	A	B	C	D
1	**student_name**	**course**	**semester**	**grade**
2	Sally	dsst289	fall	79
3	Bob	dsst289	fall	84
4	Jill	dsst289	fall	77
5	Jack	dsst289	fall	91
6	Mary	dsst289	fall	86
7	Sally	math211	fall	75
8	Bob	math211	fall	77
9	Jill	math211	fall	72
10	Jack	math211	fall	98
11	Mary	math211	fall	85
12	Sally	rhcs103	fall	93
13	Bob	rhcs103	fall	80
14	Jill	rhcs103	fall	87
15	Jack	rhcs103	fall	91
16	Mary	rhcs103	fall	88
17	Sally	dsst389	spring	95
18	Bob	dsst389	spring	97
19	Jill	dsst389	spring	78
20	Jack	dsst389	spring	78
21	Mary	dsst389	spring	84

Antipattern #5

What is the issue here and how could we fix it?

	A	B	C	D
1	**student_name**	**year**	**fav_foods**	
2	Sally	sophmore	hotdogs; pizza; nachos	
3	Bob	sophmore	bagels; sandwiches	
4	Jill	junior	ice cream	
5	Jack	junior	milkshakes; cookies	
6	Mary	senior	ribeye; lobster; caviar	
7				
8				

Antipattern #5

Note: Why is this is not a good solution? Think about using data formated like this to determine the most frequently listed favorite food from a large dataset. It turns out it is quite complicated with the tools we have available to us!

Problem: Multiple values in one cell.

Bad Solution: Create new numbered columns.

	A	B	C	D	E
1	**student_name**	**year**	**fav_food1**	**fav_food2**	**fav_food3**
2	Sally	sophmore	hotdogs	pizza	nachos
3	Bob	sophmore	bagels	sandwiches	
4	Jill	junior	ice cream		
5	Jack	junior	milkshakes	cookies	
6	Mary	senior	ribeye	lobster	cavier
7					
8					
9					

Antipattern #5 : Solution

Problem: Multiple values in one cell.

Solution: Expand data to include one row for each item.

	A	B	C	D
1	**student_name**	**year**	**fav_foods**	
2	Sally	sophmore	hotdogs	
3	Sally	sophmore	pizza	
4	Sally	sophmore	nachos	
5	Bob	sophmore	bagels	
6	Bob	sophmore	sandwiches	
7	Jill	junior	ice cream	
8	Jack	junior	milkshakes	
9	Jack	junior	cookies	
10	Mary	senior	ribeye	
11	Mary	senior	lobster	
12	Mary	senior	cavier	
13				

Antipattern #5

Note: This is format is preferred for reasons we will see when learning about data normalization in a future set of notes.

Another Solution: Same as before, but create a second table to avoid repeating information.

	A	B
1	**student_name**	**year**
2	Sally	sophmore
3	Bob	sophmore
4	Jill	junior
5	Jack	junior
6	Mary	senior
7		

	A	B	C
1	**student_name**	**fav_foods**	
2	Sally	hotdogs	
3	Sally	pizza	
4	Sally	nachos	
5	Bob	bagels	
6	Bob	sandwiches	
7	Jill	ice cream	
8	Jack	milkshakes	
9	Jack	cookies	
10	Mary	ribeye	
11	Mary	lobster	
12	Mary	cavier	
13			