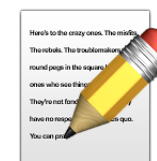


Three Things You Should Keep in Mind While Collecting Your Own Data

1. Structure your data



A few tips...

- ▶ Keep the same format across your spreadsheet (ex: number vs text)
- ▶ Don't use comments or colors in cells; instead use flags!
- ▶ Structure numbers the same (ex: thousand separator)

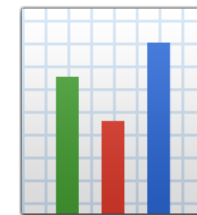
Instead of this....

A	B	C	D	E	F	G
ID	Player	Date_of_birth	Day	Month	Year	Age
1	Sergio Romero	2/22/87	22	February	1987	29
2	Nahuel Guzmán	2/10/86	10	February	1986	30
3	Mariano Andújar	7/30/83	30	July	1983	33
4	Pablo Zabaleta	1/16/85	16	January	1985	31
5	Martín Demichelis	12/20/80	20	December	1980	35
6	Marcos Rojo	3/20/90	20	March	1990	26
7	Nicolás Otamendi	2/12/88	12	February	1988	28
8	Ramiro Funes Mori	3/5/91	5	March	1991	25
9	Facundo Roncaglia	2/10/87	10	February	1987	29
10	Gabriel Mercado	3/18/87	18	March	1987	29
11	Emmanuel Más	1/15/89	15	January	1989	27
12	Mateo Musacchio	8/26/90	26	August	1990	25
13	Javier Mascherano	6/8/84	8	June	1984	32
14	Ángel Di María	2/14/88	14	February	1988	28

Format correctly

A	B	C	D	E	F	G	H	I	J	K
ID	Player	Date_of_birth	Day	Month	Year	Age	Caps	Goals	Club	Checked
1	Sergio Romero	2/22/87	22	February	1987	29	79	0	Manchester	1
2	Nahuel Guzmán	2/10/86	10	February	1986	30	6	0	UANL	1
3	Mariano Andújar	7/30/83	30	July	1983	33	11	0	Estudiantes	0
4	Pablo Zabaleta	1/16/85	16	January	1985	31	54	0	Manchester	1
5	Martín Demichelis	12/20/80	20	December	1980	35	50	2	Espanyol	1
6	Marcos Rojo	3/20/90	20	March	1990	26	49	2	Manchester	1
7	Nicolás Otamendi	2/12/88	12	February	1988	28	38	2	Manchester	0
8	Ramiro Funes Mori	3/5/91	5	March	1991	25	13	0	Everton	0
9	Facundo Roncaglia	2/10/87	10	February	1987	29	12	0	Celta	0
10	Gabriel Mercado	3/18/87	18	March	1987	29	10	2	Sevilla	0
11	Emmanuel Más	1/15/89	15	January	1989	27	4	0	San Lorenzo	0
12	Mateo Musacchio	8/26/90	26	August	1990	25	3	0	Villarreal	0

2. Data dictionary



A dictionary for ^{YOUR} data

A	B	C	D
Type	Key	Meaning	Format
personal	total_personnel	Personal insgesamt	INT
personal	fully_employed	Vollzeitbeschäftigt	INT
personal	part-time	Teilzeitbeschäftigt	INT
personal	nursing_and_care	Pflege und Betreuung	INT
personal	social_care	soziale Betreuung	INT
costs	food	Verpflegung	INT
beds	1-bed	1-Bett-Zimmern	INT
beds	2-bed	2-Bett-Zimmern	INT
beds	3-bed	3-Bett-Zimmern	INT
beds	4-bed	4 und mehr-Bett-Zimmern	INT
recipients	recipients_nursing_class_1	Pflegestufe I	FLOAT
recipients	recipients_nursing_class_2	Pflegestufe II	FLOAT
recipients	recipients_nursing_class_3	Pflegestufe III	FLOAT

Understandable headers

	A	B	C	D	E	F
1	unique id	total3	total2	namef	namel	other1
2	1	64328	392902	Sam	Smith	738393
3	2	42910	2048349	Bob	Miller	28102
4	3	27293	2302938	Sandhya	Kambhampa	1281912
5	4	91213	2392302	Sam	Jones	12182913
6	5	42323	23826	Nikil	Patel	4289328943
7	6	492303	3422	Samantha	Smith	32849
8	7	23900	3423422	John	Doe	29401

WRONG

Understandable headers

	A	B	C	D	E	F
1	unique_id	total_debt	total_revenue	first_name	last_name	taxes
2	1	64328	392902	Sam	Smith	738393
3	2	42910	2048349	Bob	Miller	28102
4	3	27293	2302938	Sandhya	Kambhampa	1281912
5	4	91213	2392302	Sam	Jones	12182913
6	5	42323	23826	Nikil	Patel	4289328943
7	6	492303	3422	Samantha	Smith	32849
8	7	23900	3423422	John	Doe	29401

3. Data diary



Keep track of your work

Data diary for Police-simple.csv created from ez-police-combined.csv

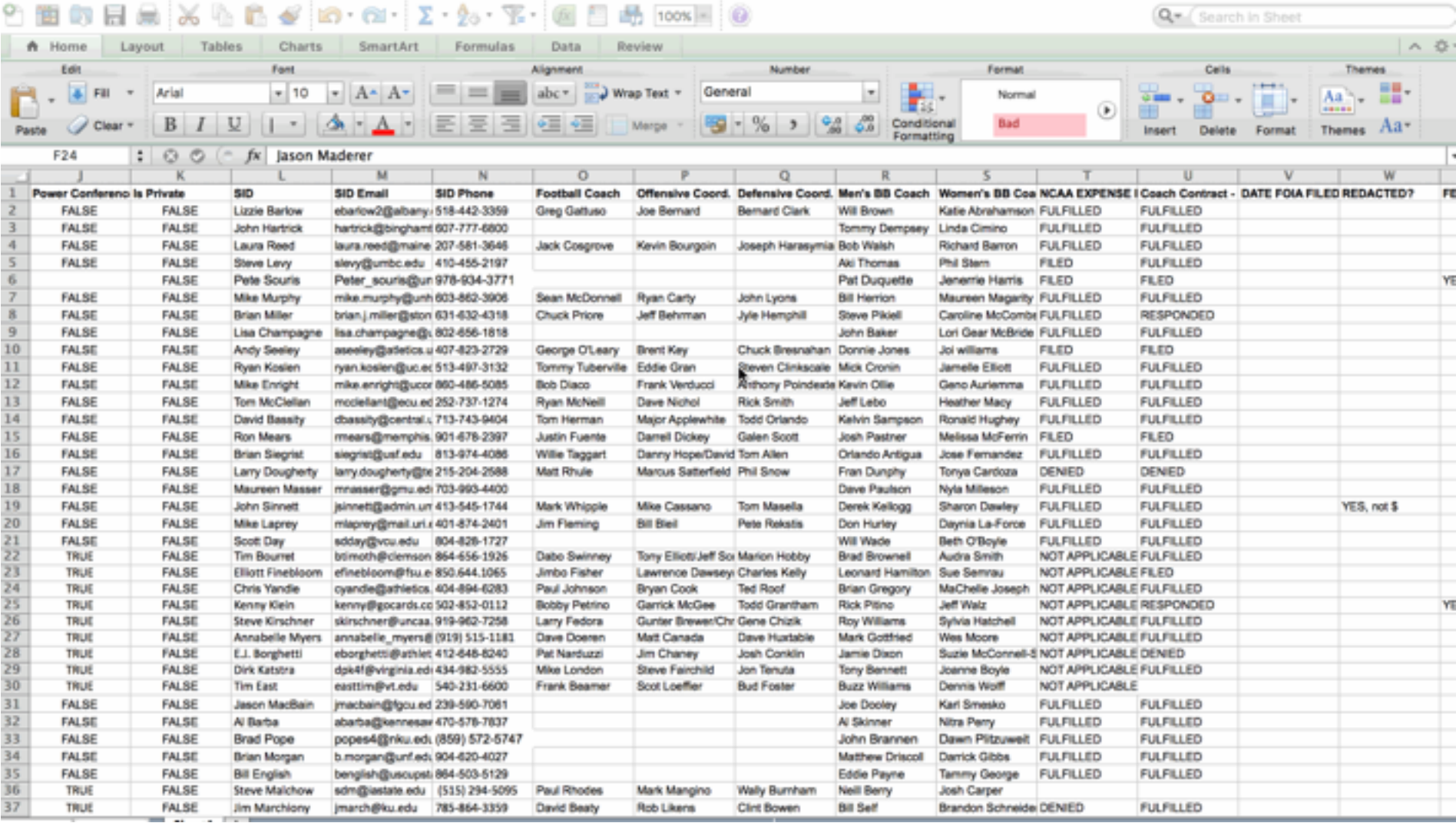
#9/8/16:

- cleaning the data in Open Refine: created new columns for each description and the title field
- formulas used:
 - + value.trim() -> shortdescriptionclean, projecttitleclean, longdescriptionclean
 - + value.toLowerCase()-> shortdescriptionclean, projecttitleclean, longdescriptionclean

Make your work reproducible

```
{,
{
  "op": "core/column-addition",
  "description": "Create column projecttitle-clean at index 7 based on
    column projecttitle using expression grel:value",
  "engineConfig": {
    "mode": "row-based",
    "facets": []
  },
  "newColumnName": "projecttitle-clean",
  "columnInsertIndex": 7,
  "baseColumnName": "projecttitle",
  "expression": "grel:value",
  "onError": "keep-original"
},
```

Track your work



The screenshot shows an Excel spreadsheet with the following columns (from left to right):

- Power Conference
- Private
- SID
- SID Email
- SID Phone
- Football Coach
- Offensive Coord.
- Defensive Coord.
- Men's BB Coach
- Women's BB Coach
- NCAA EXPENSE
- Coach Contract -
- DATE FOIA FILED
- REDACTED?
- FE

The spreadsheet contains 37 rows of data, each representing an individual. The data is organized into columns for various roles and contact details. The first two columns, 'Power Conference' and 'Private', contain 'FALSE' or 'TRUE' values. The 'SID' column contains names. The 'SID Email' and 'SID Phone' columns contain email addresses and phone numbers. The 'Football Coach' column contains names. The 'Offensive Coord.' and 'Defensive Coord.' columns contain names. The 'Men's BB Coach' and 'Women's BB Coach' columns contain names. The 'NCAA EXPENSE' column contains 'FULFILLED', 'FILED', 'DENIED', 'NOT APPLICABLE', or 'RESPONDED'. The 'Coach Contract -' column contains 'FULFILLED' or 'FILED'. The 'DATE FOIA FILED' column contains 'YES, not \$' or 'YES, not \$'. The 'REDACTED?' column contains 'YES, not \$' or 'YES, not \$'. The 'FE' column contains 'YES, not \$' or 'YES, not \$'.

@sandhya_k
#IJAsia16

Thank you! Questions?

Tweet: @sandhya__k
sjkambhampati@gmail.com