Data Fundamentals Lab 3: Data Cleaning

In this lab, you will clean data scraped from a web page to make it easier to analyse. You will fix data labels, spellings, formatting, and arrangement.

Crime Statistics in Tanzania

In this lab, you will clean the ***Criminal Offences and Road Accidents, Tanzania, 2013 - 2016*** dataset from the Tanzania in Figures 2016 report. Cleaning data is an important first step to make the data ready for analysis.

1. Open Tanzania in Figures 2016 and save a copy of the PDF locally on your computer.

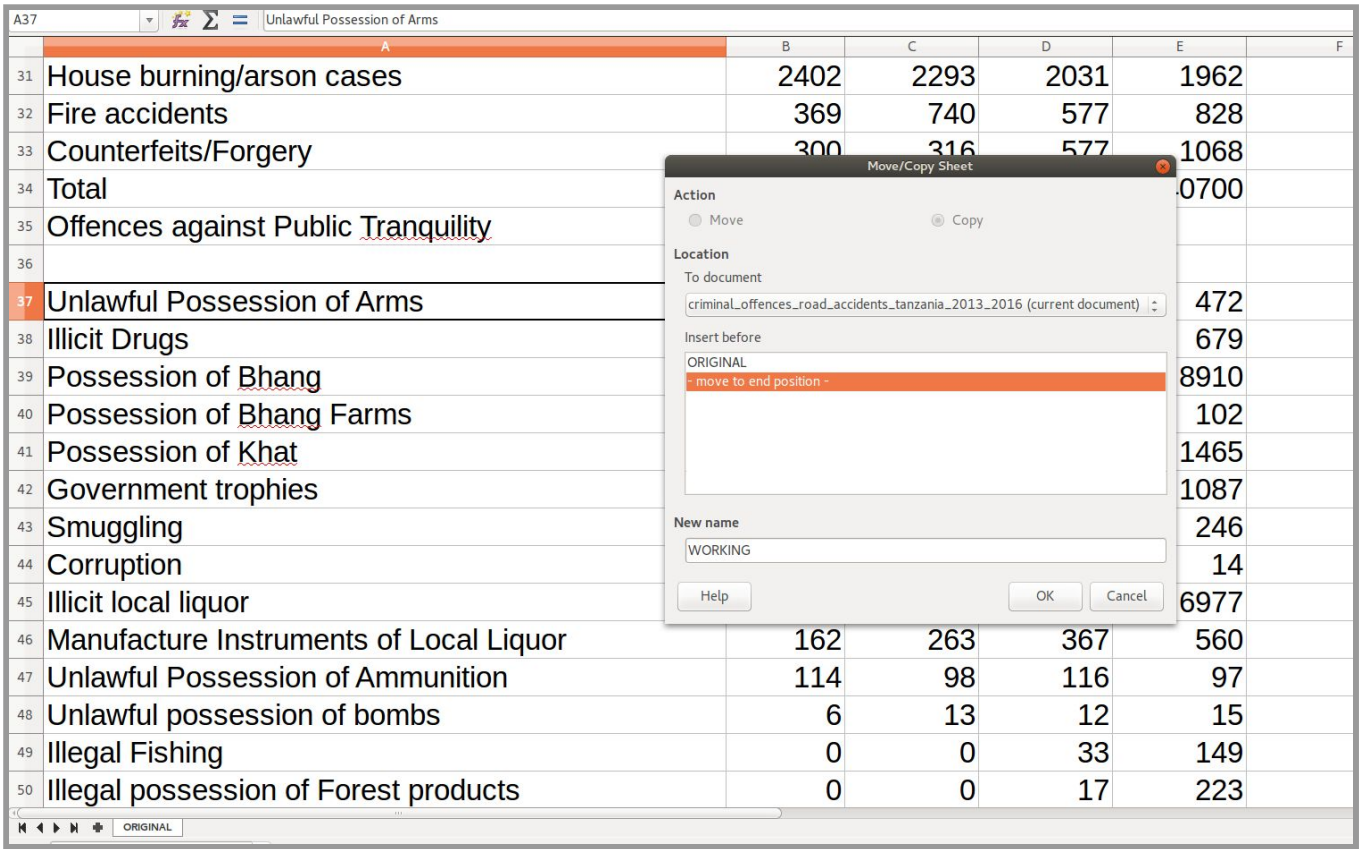
2. Using Tabula, scrape the Criminal Offences and Road Accidents, Tanzania 2013 - 2016 table on **pages 79 - 81**.

3. Export the scraped data as a CSV and open in Microsoft Excel or LibreOffice Calc.

4. Save As and label the data as **criminal_offences_road_accidents_tanzania_2013_2016.xlsx .**

5. Rename the current (and only) sheet as **ORIGINAL**.

6. Create a clean copy of the tab and label it **WORKING**. We will be working on this sheet for the cleaning exercises.



	A	B	C	D	E
31	House burning/arson cases	2402	2293	2031	1962
32	Fire accidents	369	740	577	828
33	Counterfeits/Forgery	300	316	577	1068
34	Total				0700
35	Offences against Public Tranquility				
36					
37	Unlawful Possession of Arms				472
38	Illicit Drugs				679
39	Possession of Bhang				8910
40	Possession of Bhang Farms				102
41	Possession of Khat				1465
42	Government trophies				1087
43	Smuggling				246
44	Corruption				14
45	Illicit local liquor				6977
46	Manufacture Instruments of Local Liquor	162	263	367	560
47	Unlawful Possession of Ammunition	114	98	116	97
48	Unlawful possession of bombs	6	13	12	15
49	Illegal Fishing	0	0	33	149
50	Illegal possession of Forest products	0	0	17	223

7. Open the original Tanzania in Figures 2016 PDF you saved to compare the result of the scraping and what to clean.

8. The first thing to notice is the empty rows in the scraped data set. These do not exist in the original table. Let us get rid of them.

9. Delete rows 3, 13, 36 and 56. The last row now should be row 59.

51	Crime Type	2013	2014	2015	2016
52	Road Traffic Accidents, Death and Injured Persons				
53	Major	24480	15420	8777	10297
54	Minor	663722	1110252	1381705	2200442
55	Total	688202	1125672	1390482	2210739
56	Road accidents	24480	15420	8777	10297
57	Fatal accidents	3545	3106	2909	2840
58	Deaths	4091	3857	3574	3381
59	Injured Persons	21536	15230	9993	9549
60					

10. Next, let's focus on the header row (row 1). The header labels are not so clear. We can improve upon them:
    a. Recorded incidents in 2013
    b. Recorded incidents in 2014
    c. Recorded incidents in 2015
    d. Recorded incidents in 2016

11. Rows 10, 32, 50 and 55 represent some kind of totals. In a clean dataset, values that can be regenerated from other values in the dataset are considered redundant and should be removed. Total or sum values, averages, minimums and maximums are common examples of these. For this dataset, we will delete these **rows (10, 32, 50 and 55)**. The last row after these deletions should be **row 55**.

49	Road Traffic Accidents, Death and Injured Persons				
50	Major	24480	15420	8777	10297
51	Minor	663722	1110252	1381705	2200442
52	Road accidents	24480	15420	8777	10297
53	Fatal accidents	3545	3106	2909	2840
54	Deaths	4091	3857	3574	3381
55	Injured Persons	21536	15230	9993	9549
56					

12. We can delete **Row 48** since it is a repetition of the original header and is not necessary anymore (since the dataset is now combined). The last row after this action should be **Row 54**.

13. Next, you may notice that rows 2, 10, 31 and 48 are originally sub-headings for the different categories of offences. We can think of these as indicators of who or what the offences were committed against. We should create a new column for this indicator. Label this column as ***"Offence Against"*** in **Column F**.

14. Type "Person" into **cell F3** to indicate that this offence was committed against a person.

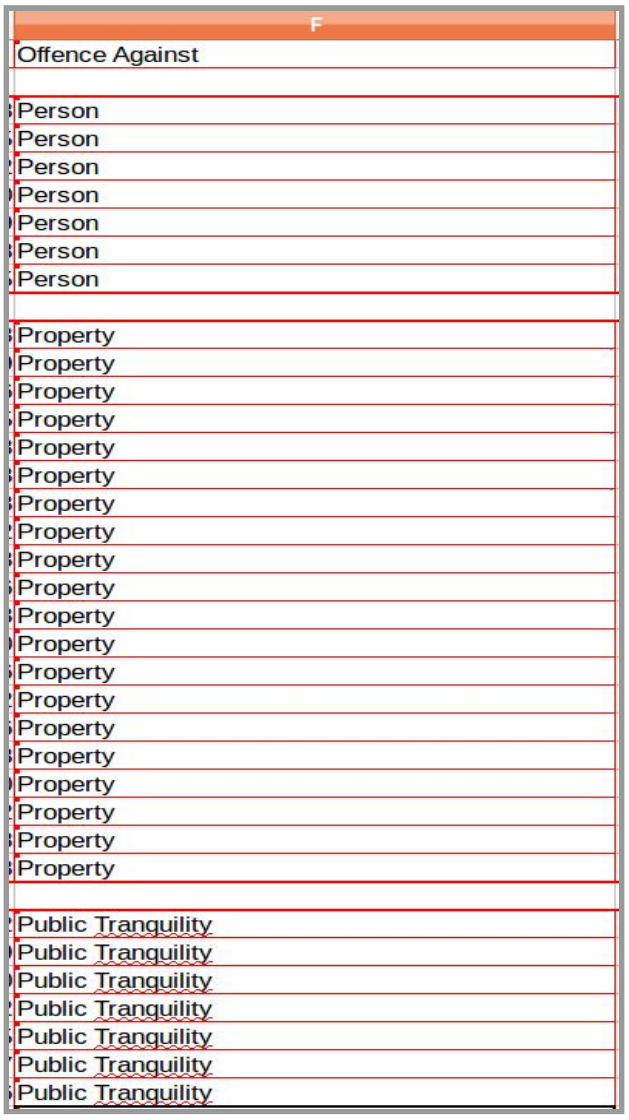
15. While cell F3 is still selected, move the cursor to the bottom-right corner of the cell until you see a **plus sign.** Click, hold and drag the box all the way down to cell F9 (this is the last row under the ***Offences Against Person*** section).

	A	B	C	D	E	F
1	Crime Type	Recorded incidents in 2013	Recorded incidents in 2014	Recorded incidents in 2015	Recorded incidents in 2016	Offence Against
2	Offences Against Person					
3	Killings/Murder	3929	3775	3560	3318	Person
4	Rape	6105	6028	5802	7645	Person
5	Un-natural Offence	820	944	928	1202	Person
6	Child Stealing	192	146	146	170	Person
7	Child Desertion	243	237	205	159	Person
8	Defilement	10	15	12	18	Person
9	Human Trafficking	36	21	45	55	Person
10	Offences Related to Property					

16. Repeat similar steps for the other offences as follows:
    a. Property: F11 to F30

    b. Public Tranquility: F32 to F47

    c. Road Traffic Accidents, Death and Injured Persons: F49 to F54



F
Offence Against
Person
Person
Person
Person
Person
Person
Person
Property
Property
Property
Property
Property
Property
Property
Property
Property
Property
Property
Property
Property
Property
Property
Property
Property
Property
Public Tranquility
Public Tranquility
Public Tranquility
Public Tranquility
Public Tranquility
Public Tranquility
Public Tranquility

17. Now we can delete rows 2, 10, 31 and 48 since this data is captured in column F.

18. Our next goal is to ensure that each cells contains an appropriate value. Let us start with filling any cell that's empty. Cells **B14** and **C14** are the only cells containing no value. This could be as a result of no data being available or being recorded at the time of reporting. To correct this, type **NA** into cells B14 and C14.

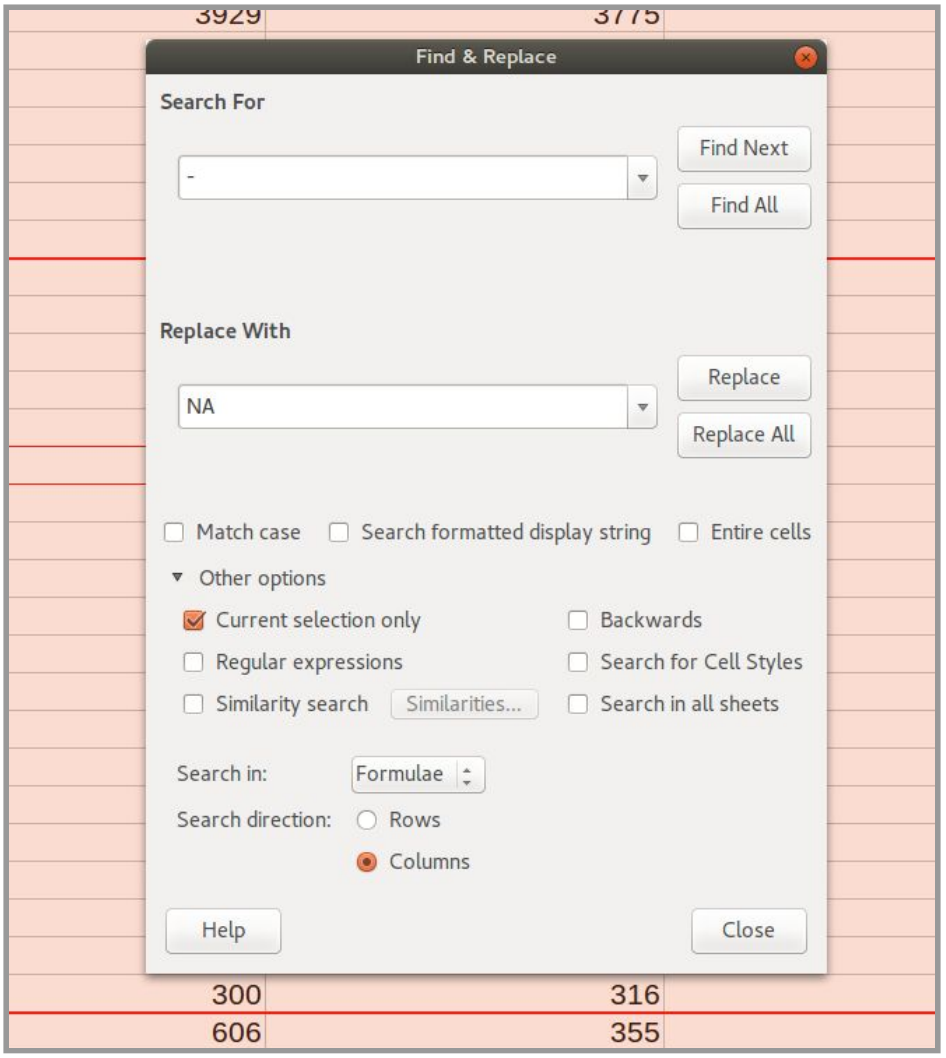
12	Robbery with Violence	5710	5294	4507
13	Robbery through fire arms -Breaking in	23017	21479	20337
14	Property theft	NA	NA	179
15	Theft of Motorcycle	4695	5232	5317
16	Theft of motor vehicles	464	427	488

19. Lastly, we need to replace the **"-"** character in cells **B25** and **C10.** One possible option will be to use the Find & Replace feature where we can replace **"-"** with **"NA"**. When we do this however, we also end up replace the hyphenated words in cells A4, A13, A17 and A20. This is not what we want. We need to be more specific.

	A	B	C
1	Crime Type	Recorded incidents in 2013	Recorded incidents in 2014
2	Killings/Murder	3929	3775
3	Rape	6105	6028
4	UnNAnatural Offence	820	944
5	Child Stealing	192	146
6	Child Desertion	243	237
7	Defilement	10	15
8	Human Trafficking	36	21
9	Guns theft / theft of Arms	76	59
10	Robbery in Highway	3	NA
11	Armed Robbery	1266	1127
12	Robbery with Violence	5710	5294
13	Robbery through fire arms NABreaking in	23017	21479
14	Property theft	NA	NA
15	Theft of Motorcycle	4695	5232
16	Theft of motor vehicles	464	427
17	Fake money NACounterfeiting of Bank Notes	664	517
18	Livestock theft	5307	5119
19	Bank robbery	116	55
20	Theft in the public sectors NAParastatal Organization	158	59
21	Theft in the trade unions/Cooperative union	13	32
22	Theft in the Local Government	11	6
23	Crime Type	2013	2014
24	Theft in the Central Government	14	8
25	Theft in political parties	NA	2
26	House burning/arson cases	2402	2293
27	Fire accidents	369	740

20. To fix this, do the following:
    a. Select only columns B and C.
    b. Type ***Ctrl + H***

c. Type **"-"** in the Find section and **"NA"** in the Replace section. Note: the quotes should not be included when typing.
d. Click on "***Other options***".
e. Select the ***"Current selection only"*** option
f. Click on **Replace all**.



8	Human Trafficking	36	21
9	Guns theft / theft of Arms	76	59
10	Robbery in Highway	3	NA
11	Armed Robbery	1266	1127
12	Robbery with Violence	5710	5294
13	Robbery through fire arms -Breaking in	23017	21479
14	Property theft	NA	NA
15	Theft of Motorcycle	4695	5232
16	Theft of motor vehicles	464	427
17	Fake money -Counterfeiting of Bank Notes	664	517
18	Livestock theft	5307	5119
19	Bank robbery	116	55
20	Theft in the public sectors -Parastatal Organization	158	59
21	Theft in the trade unions/Cooperative union	13	32
22	Theft in the Local Government	11	6
23	Crime Type	2013	2014
24	Theft in the Central Government	14	8
25	Theft in political parties	NA	2
26	House burning/arson cases	2402	2293
27	Fire accidents	369	740
28	Counterfeits/Forgery	300	316

21. Your data is now ready to use.

	A	B	C	D	E	F
1	Crime Type	Recorded incidents in 2013	Recorded incidents in 2014	Recorded incidents in 2015	Recorded incidents in 2016	Offence Against
2	Killings/Murder	3929	3775	3560	3318	Person
3	Rape	6105	6028	5802	7645	Person
4	Un-natural Offence	820	944	928	1202	Person
5	Child Stealing	192	146	146	170	Person
6	Child Desertion	243	237	205	159	Person
7	Defilement	10	15	12	18	Person
8	Human Trafficking	36	21	45	55	Person
9	Guns theft / theft of Arms	76	59	53	33	Property
10	Robbery in Highway	3	NA	3	29	Property
11	Armed Robbery	1266	1127	913	726	Property
12	Robbery with Violence	5710	5294	4507	3945	Property
13	Robbery through fire arms -Breaking in	23017	21479	20337	19803	Property
14	Property theft	NA	NA	179	193	Property
15	Theft of Motorcycle	4695	5232	5317	5633	Property
16	Theft of motor vehicles	464	427	488	452	Property
17	Fake money -Counterfeiting of Bank Notes	664	517	416	563	Property
18	Livestock theft	5307	5119	4879	5106	Property
19	Bank robbery	116	55	45	23	Property
20	Theft in the public sectors -Parastatal Organization	158	59	72	90	Property
21	Theft in the trade unions/Cooperative union	13	32	2	56	Property
22	Theft in the Local Government	11	6	11	152	Property
23	Crime Type	2013	2014	2015	2016	Property
24	Theft in the Central Government	14	8	11	38	Property
25	Theft in political parties	NA	2	12	0	Property
26	House burning/arson cases	2402	2293	2031	1962	Property
27	Fire accidents	369	740	577	828	Property
28	Counterfeits/Forgery	300	316	577	1068	Property
29	Unlawful Possession of Arms	606	355	444	472	Public Tranquility
30	Illicit Drugs	479	480	481	679	Public Tranquility
31	Possession of Bhang	6821	6747	7550	8910	Public Tranquility
32	Possession of Bhang Farms	0	57	85	102	Public Tranquility
33	Possession of Khat	1331	1206	1272	1465	Public Tranquility
34	Government trophies	884	610	982	1087	Public Tranquility
35	Smuggling	61	79	72	246	Public Tranquility
36	Corruption	15	3	6	14	Public Tranquility
37	Illicit local liquor	5064	4669	5321	6977	Public Tranquility
38	Manufacture Instruments of Local Liquor	162	263	367	560	Public Tranquility
39	Unlawful Possession of Ammunition	114	98	116	97	Public Tranquility
40	Unlawful possession of bombs	6	13	12	15	Public Tranquility
41	Illegal Fishing	0	0	33	149	Public Tranquility
42	Illegal possession of Forest products	0	0	17	223	Public Tranquility
43	Illegal possession of Sea products	0	0	0	3	Public Tranquility
44	Illegal Immigrant	871	599	928	1221	Public Tranquility
45	Major	24480	15420	8777	10297	Road Traffic Accidents, Death and Injured Persons
46	Minor	663722	1110252	1381705	2200442	Road Traffic Accidents, Death and Injured Persons
47	Road accidents	24480	15420	8777	10297	Road Traffic Accidents, Death and Injured Persons
48	Fatal accidents	3545	3106	2909	2840	Road Traffic Accidents, Death and Injured Persons
49	Deaths	4091	3857	3574	3381	Road Traffic Accidents, Death and Injured Persons
50	Injured Persons	21536	15230	9993	9549	Road Traffic Accidents, Death and Injured Persons