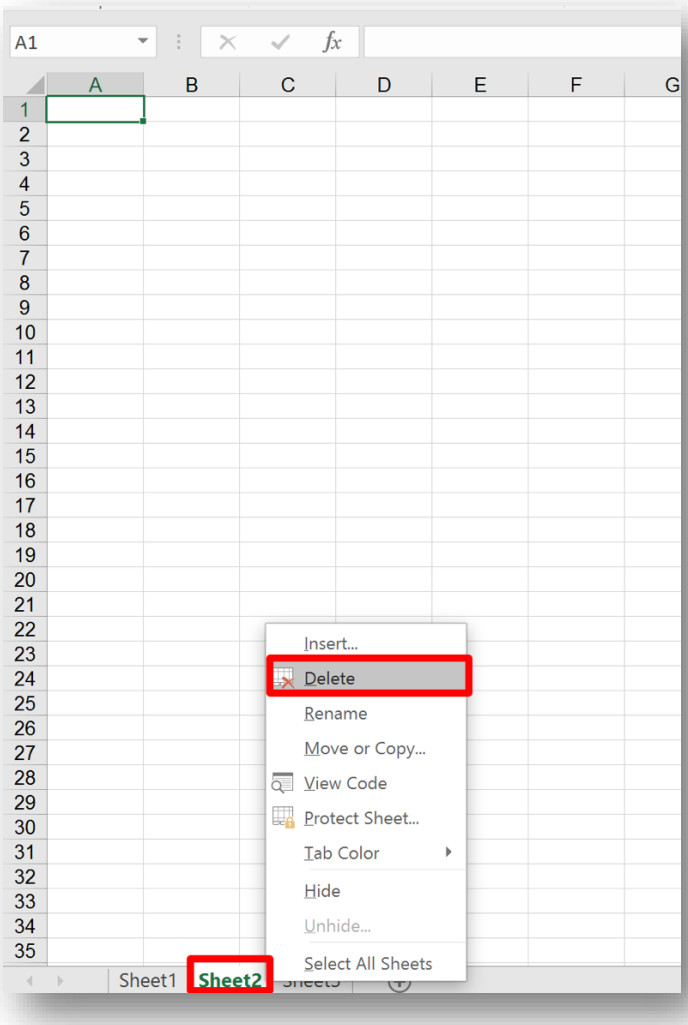
Data Cleaning

In this lab, you will clean data scraped from a PDF or web page in preparation for analysis. You will fix data labels, spelling, formatting, and standardize data entry. For a review of data standardization, review Module 3.

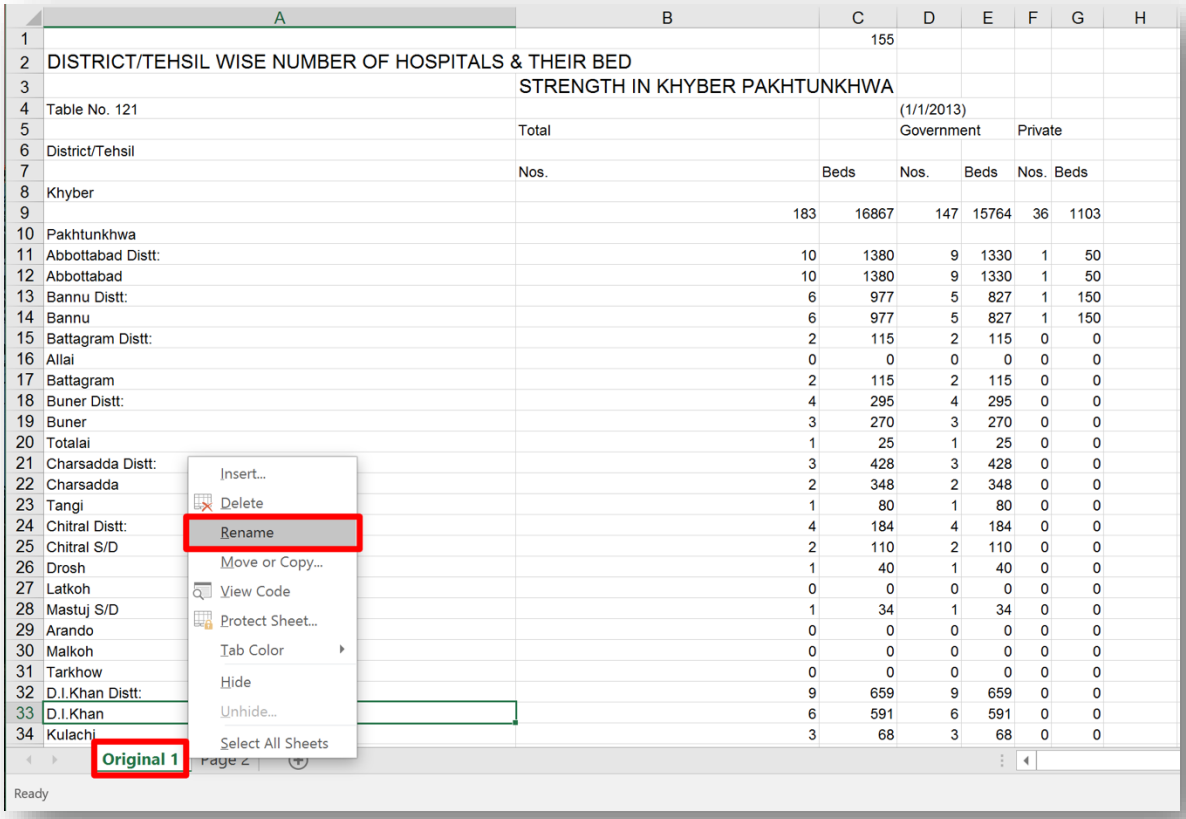
Working in Excel First Steps

There are some basic best practices for working with data in Excel: steps data journalists always take before modifying data. These steps are to ensure the integrity of the original data. Saving an untouched copy of your data is important in case of legal challenges and in case you make mistakes and need to revert to the original dataset. We will be working with a workbook, which is an Excel data file. We will also refer to tabs and sheets, which are the separate pages of each workbook.

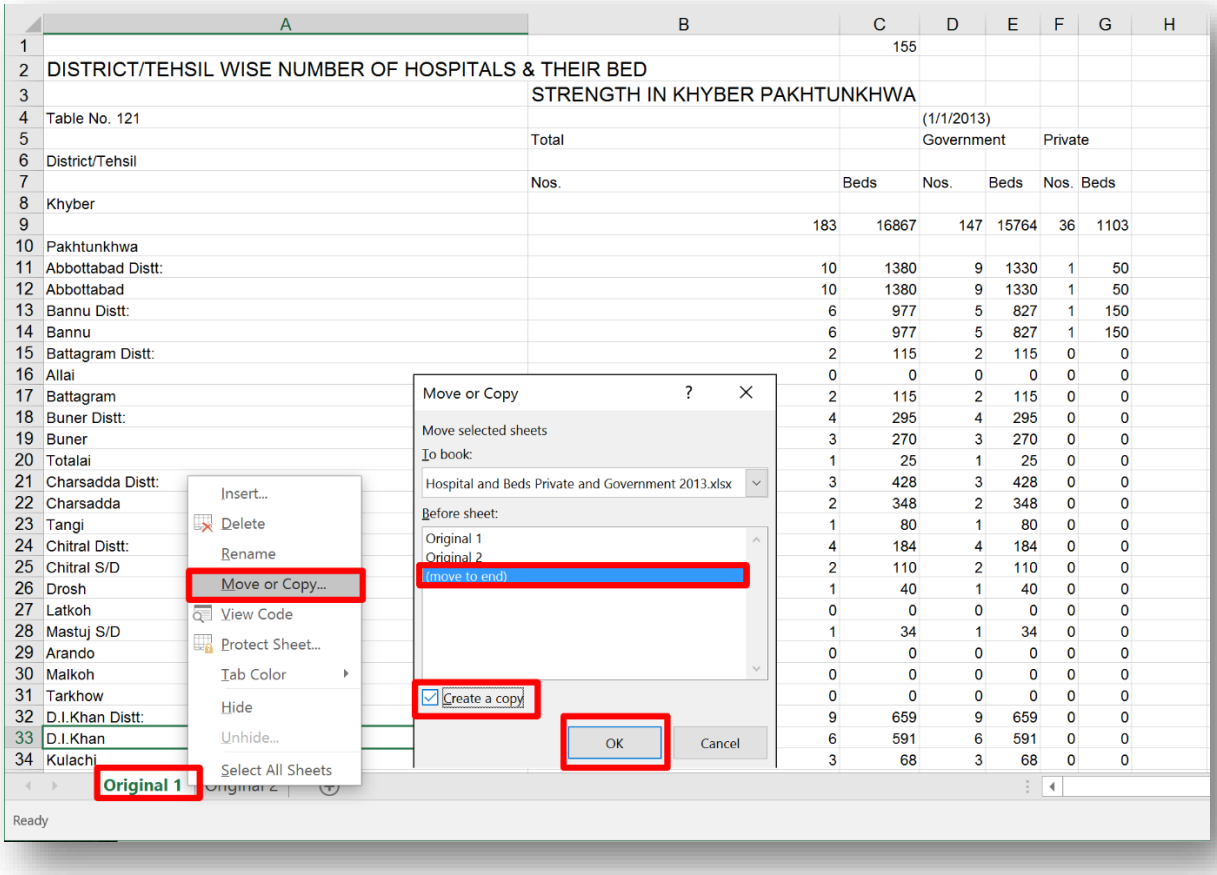
1. Open **Hospital and Beds Private and Government 2013.xlsx** The data scrapped was originally on two pages, reflected in the two Excel tabs.
2. **Delete empty sheets:** in case is necessary to delete any not data sheet by using right click on tab and use Delete function.



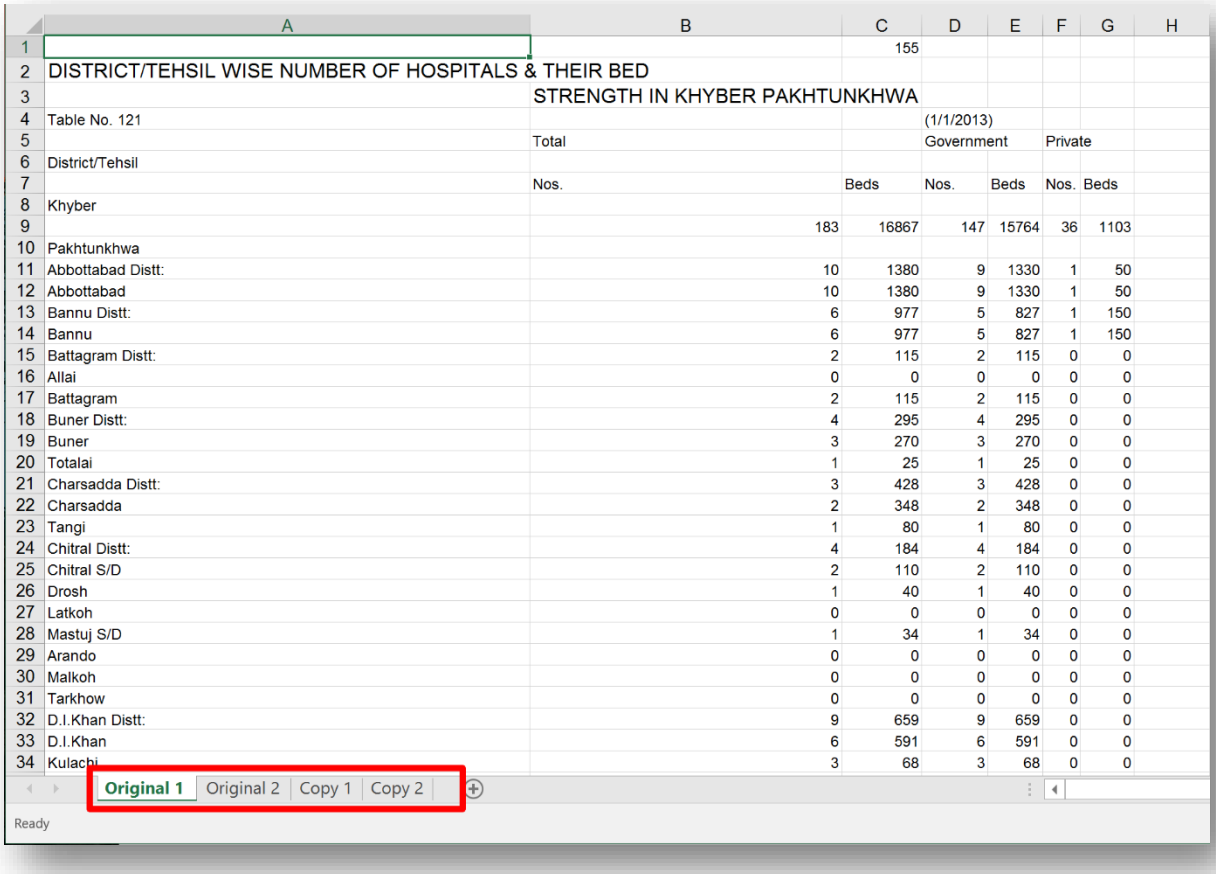
3. **Rename Sheets**: Rename Sheet 1 by right click on Page 1 and use the Rename function. Rename as Original 1 do the same with Sheet 2 and rename it Original 2



4. **Make a copy:** Right click on Original 1 sheet use Move or Copy function, click on move to the end and create a copy.



5. **Rename the copy:** Rename Original 1 (2) Sheet as Copy 1. Do the same with Original 2 and Save as Copy 2.



	A	B	C	D	E	F	G	H
1			155					
2	DISTRICT/TEHSIL WISE NUMBER OF HOSPITALS & THEIR BED							
3		STRENGTH IN KHYBER PAKHTUNKHWA						
4	Table No. 121			(1/1/2013)				
5		Total		Government		Private		
6	District/Tehsil							
7		Nos.	Beds	Nos.	Beds	Nos.	Beds	
8	Khyber							
9		183	16867	147	15764	36	1103	
10	Pakhtunkhwa							
11	Abbottabad Distt:	10	1380	9	1330	1	50	
12	Abbottabad	10	1380	9	1330	1	50	
13	Bannu Distt:	6	977	5	827	1	150	
14	Bannu	6	977	5	827	1	150	
15	Battagram Distt:	2	115	2	115	0	0	
16	Allai	0	0	0	0	0	0	
17	Battagram	2	115	2	115	0	0	
18	Buner Distt:	4	295	4	295	0	0	
19	Buner	3	270	3	270	0	0	
20	Totalai	1	25	1	25	0	0	
21	Charsadda Distt:	3	428	3	428	0	0	
22	Charsadda	2	348	2	348	0	0	
23	Tangi	1	80	1	80	0	0	
24	Chitral Distt:	4	184	4	184	0	0	
25	Chitral S/D	2	110	2	110	0	0	
26	Drosh	1	40	1	40	0	0	
27	Latkoh	0	0	0	0	0	0	
28	Mastuj S/D	1	34	1	34	0	0	
29	Arando	0	0	0	0	0	0	
30	Malkoh	0	0	0	0	0	0	
31	Tarkhow	0	0	0	0	0	0	
32	D.I.Khan Distt:	9	659	9	659	0	0	
33	D.I.Khan	6	591	6	591	0	0	
34	Kulachi	3	68	3	68	0	0	

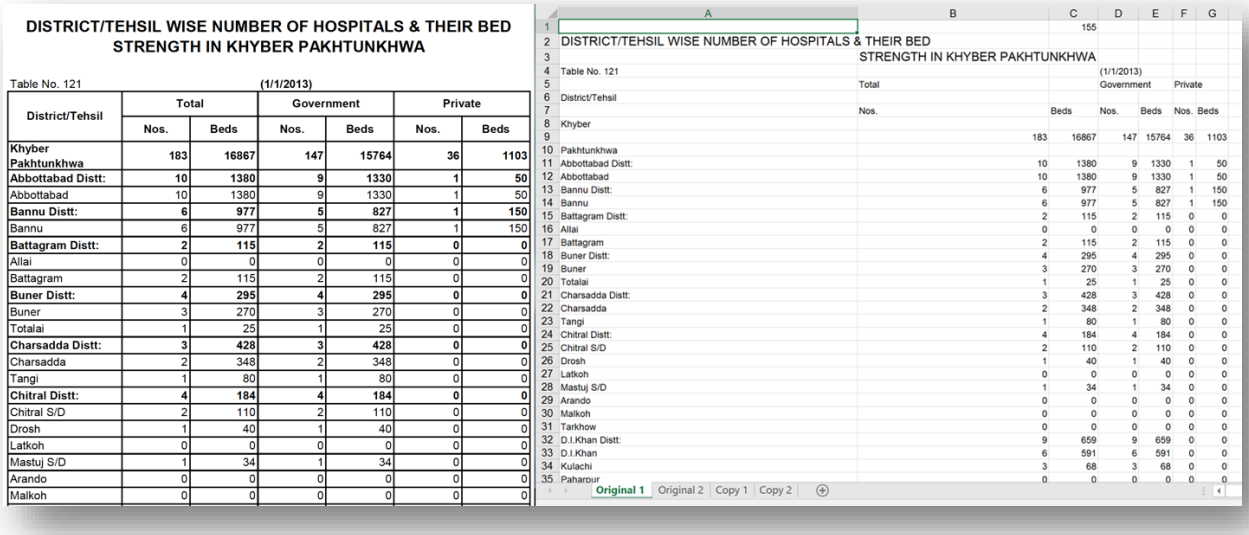
Original 1 | Original 2 | Copy 1 | Copy 2

Ready

6. Save changes

Hospital and Beds Private and Government 2013

In Lab 1 Scraping Data from PDFs and Images, you scraped a PDF document using Zamzar.com called Hospital and Beds Private and Government 2013. The scraped version is saved in the folder for the Cleaning Data Lab. Compare them.



DISTRICT/TEHSIL WISE NUMBER OF HOSPITALS & THEIR BED STRENGTH IN KHYBER PAKHTUNKHWA

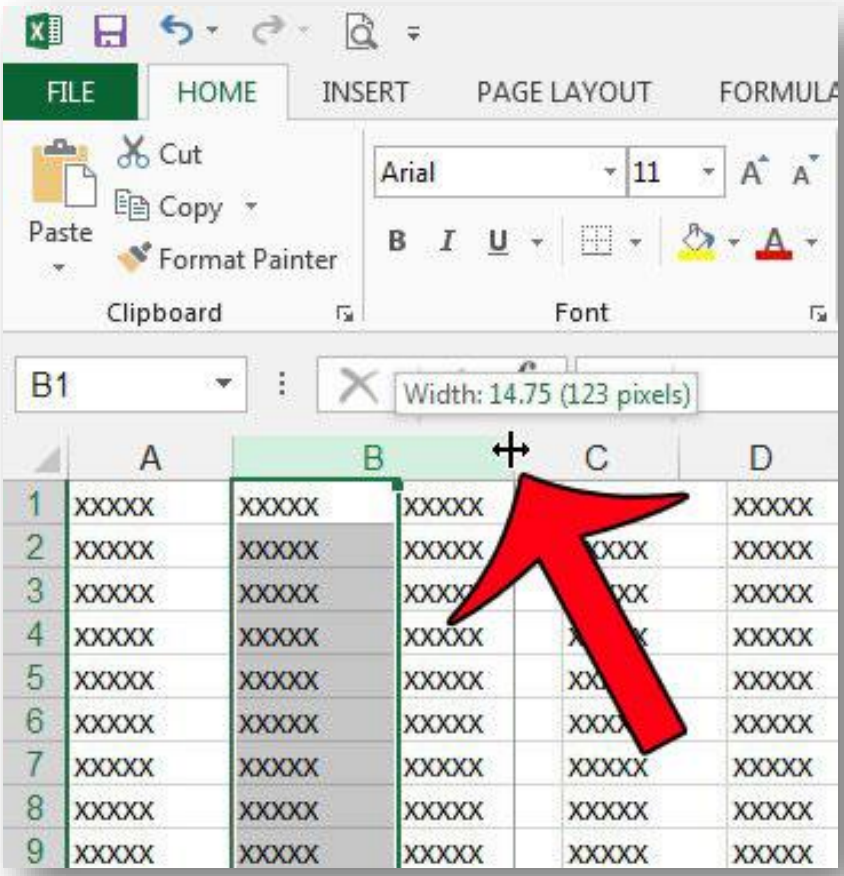
Table No. 121 (1/1/2013)

District/Tehsil	Total		Government		Private	
	Nos.	Beds	Nos.	Beds	Nos.	Beds
Khyber Pakhtunkhwa	**183**	**16867**	**147**	**15764**	**36**	**1103**
Abbottabad Distt:	**10**	**1380**	**9**	**1330**	**1**	**50**
Abbottabad	10	1380	9	1330	1	50
Bannu Distt:	**6**	**977**	**5**	**827**	**1**	**150**
Bannu	6	977	5	827	1	150
Battagram Distt:	**2**	**115**	**2**	**115**	**0**	**0**
Allai	0	0	0	0	0	0
Battagram	2	115	2	115	0	0
Buner Distt:	**4**	**295**	**4**	**295**	**0**	**0**
Buner	3	270	3	270	0	0
Totalai	1	25	1	25	0	0
Charsadda Distt:	**3**	**428**	**3**	**428**	**0**	**0**
Charsadda	2	348	2	348	0	0
Tangi	1	80	1	80	0	0
Chitral Distt:	**4**	**184**	**4**	**184**	**0**	**0**
Chitral S/D	2	110	2	110	0	0
Drosh	1	40	1	40	0	0
Latkoh	0	0	0	0	0	0
Mastuj S/D	1	34	1	34	0	0
Arando	0	0	0	0	0	0
Malkoh	0	0	0	0	0	0

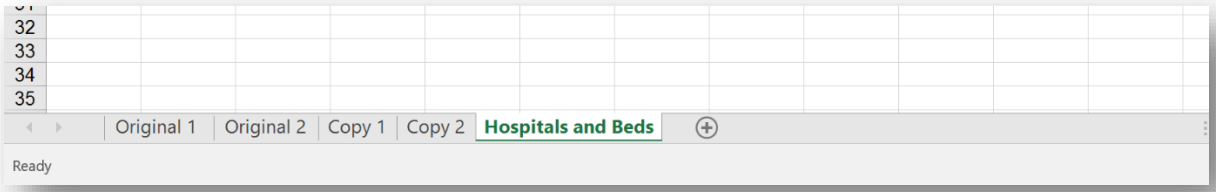
	A	B	C	D	E	F	G
1			155				
2	DISTRICT/TEHSIL WISE NUMBER OF HOSPITALS & THEIR BED						
3		STRENGTH IN KHYBER PAKHTUNKHWA					
4	Table No. 121			(1/1/2013)			
5		Total		Government		Private	
6	District/Tehsil						
7		Nos.	Beds	Nos.	Beds	Nos.	Beds
8	Khyber						
9		183	16867	147	15764	36	1103
10	Pakhtunkhwa						
11	Abbottabad Distt:	10	1380	9	1330	1	50
12	Abbottabad	10	1380	9	1330	1	50
13	Bannu Distt:	6	977	5	827	1	150
14	Bannu	6	977	5	827	1	150
15	Battagram Distt:	2	115	2	115	0	0
16	Allai	0	0	0	0	0	0
17	Battagram	2	115	2	115	0	0
18	Buner Distt:	4	295	4	295	0	0
19	Buner	3	270	3	270	0	0
20	Totalai	1	25	1	25	0	0
21	Charsadda Distt:	3	428	3	428	0	0
22	Charsadda	2	348	2	348	0	0
23	Tangi	1	80	1	80	0	0
24	Chitral Distt:	4	184	4	184	0	0
25	Chitral S/D	2	110	2	110	0	0
26	Drosh	1	40	1	40	0	0
27	Latkoh	0	0	0	0	0	0
28	Mastuj S/D	1	34	1	34	0	0
29	Arando	0	0	0	0	0	0
30	Malkoh	0	0	0	0	0	0
31	Tarkhow	0	0	0	0	0	0
32	D.I.Khan Distt:	9	659	9	659	0	0
33	D.I.Khan	6	591	6	591	0	0
34	Kulachi	3	68	3	68	0	0
35	Paharpur	0	0	0	0	0	0

Original 1 | Original 2 | Copy 1 | Copy 2

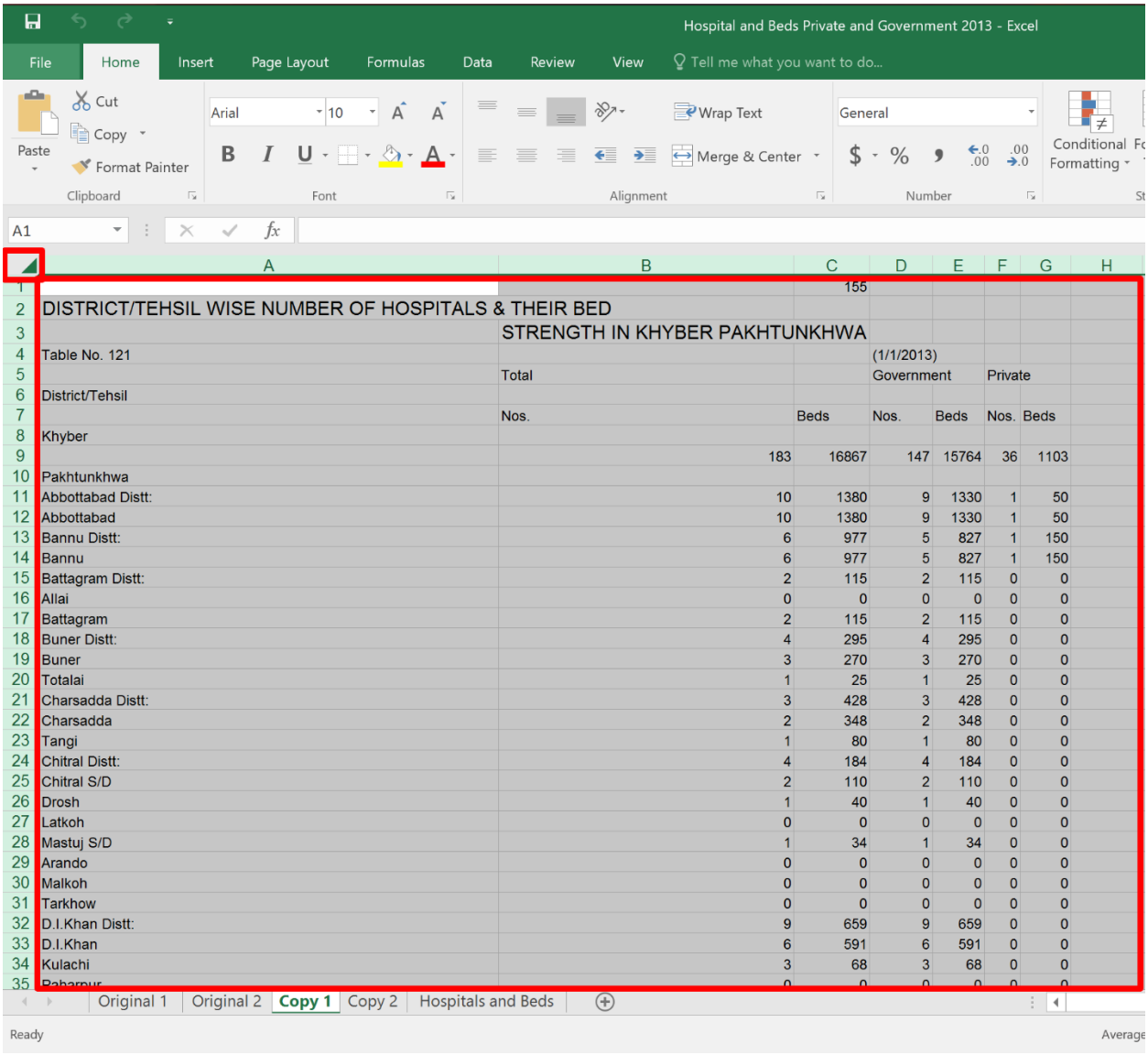
7. To make it easier to see the information in the spreadsheet, resize the columns. Move your cursor to between two columns until you see an icon with two arrows pointing in opposite directions, move the cursor and resize columns. You can also double click to resize automatically.



8. Create another sheet by clicking on the plus icon on the Sheets menu; name it Hospitals and Beds

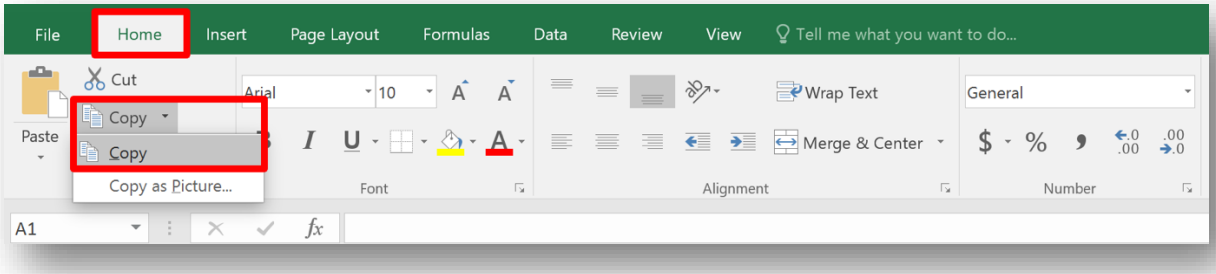


9. Go to Copy 1 Select all content with Control + A or click on the arrow in the top corner between A and 1.

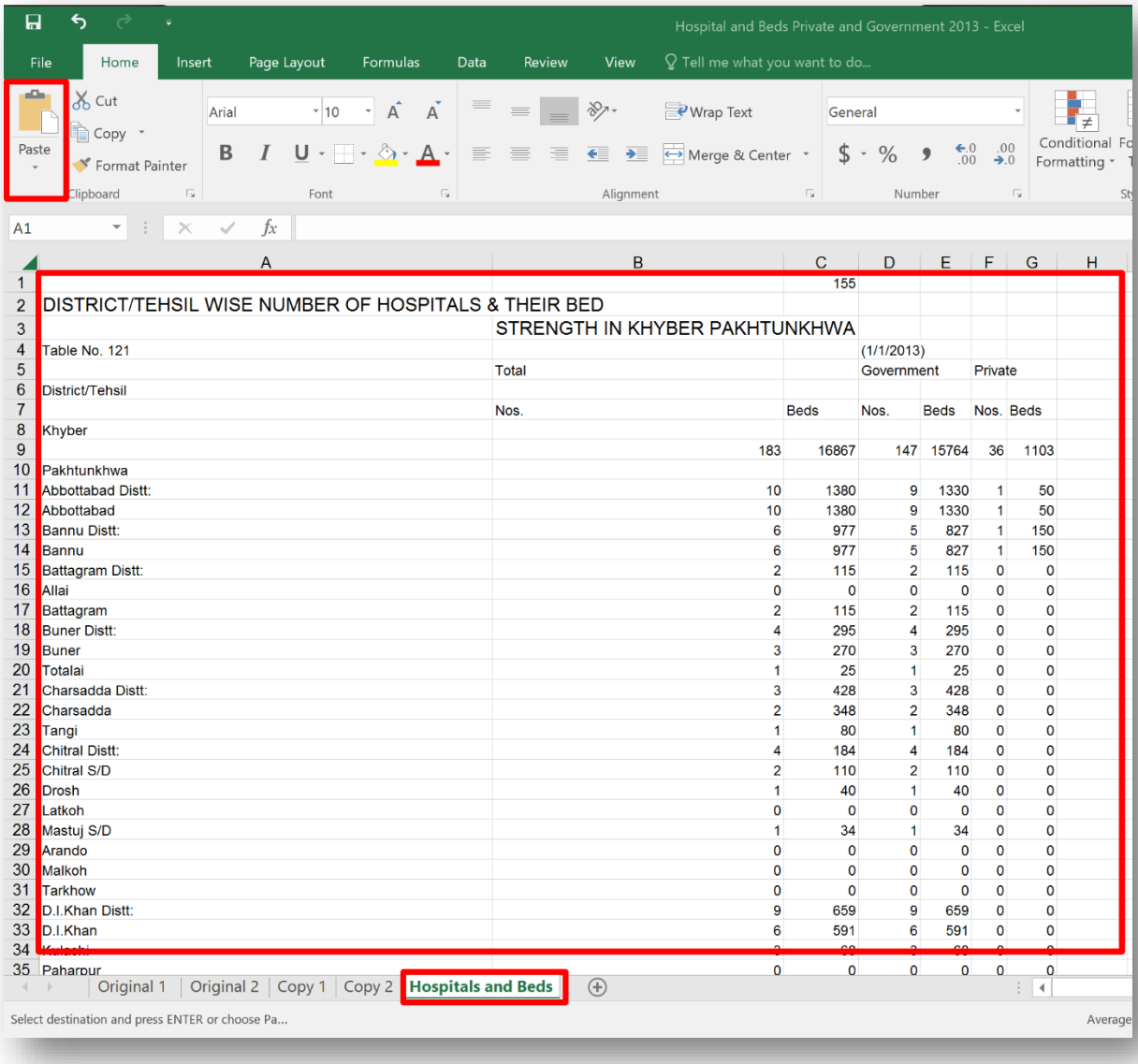


	A	B	C	D	E	F	G	H
1			155					
2	DISTRICT/TEHSIL WISE NUMBER OF HOSPITALS & THEIR BED							
3		STRENGTH IN KHYBER PAKHTUNKHWA						
4	Table No. 121			(1/1/2013)				
5		Total		Government		Private		
6	District/Tehsil							
7		Nos.	Beds	Nos.	Beds	Nos.	Beds	
8	Khyber							
9		183	16867	147	15764	36	1103	
10	Pakhtunkhwa							
11	Abbottabad Distt:	10	1380	9	1330	1	50	
12	Abbottabad	10	1380	9	1330	1	50	
13	Bannu Distt:	6	977	5	827	1	150	
14	Bannu	6	977	5	827	1	150	
15	Battagram Distt:	2	115	2	115	0	0	
16	Allai	0	0	0	0	0	0	
17	Battagram	2	115	2	115	0	0	
18	Buner Distt:	4	295	4	295	0	0	
19	Buner	3	270	3	270	0	0	
20	Totalai	1	25	1	25	0	0	
21	Charsadda Distt:	3	428	3	428	0	0	
22	Charsadda	2	348	2	348	0	0	
23	Tangi	1	80	1	80	0	0	
24	Chitral Distt:	4	184	4	184	0	0	
25	Chitral S/D	2	110	2	110	0	0	
26	Drosh	1	40	1	40	0	0	
27	Latkoh	0	0	0	0	0	0	
28	Mastuj S/D	1	34	1	34	0	0	
29	Arando	0	0	0	0	0	0	
30	Malkoh	0	0	0	0	0	0	
31	Tarkhow	0	0	0	0	0	0	
32	D.I.Khan Distt:	9	659	9	659	0	0	
33	D.I.Khan	6	591	6	591	0	0	
34	Kulachi	3	68	3	68	0	0	
35	Paharpur	0	0	0	0	0	0	

10. Go to Home – Clipboard – Copy or use Control C.

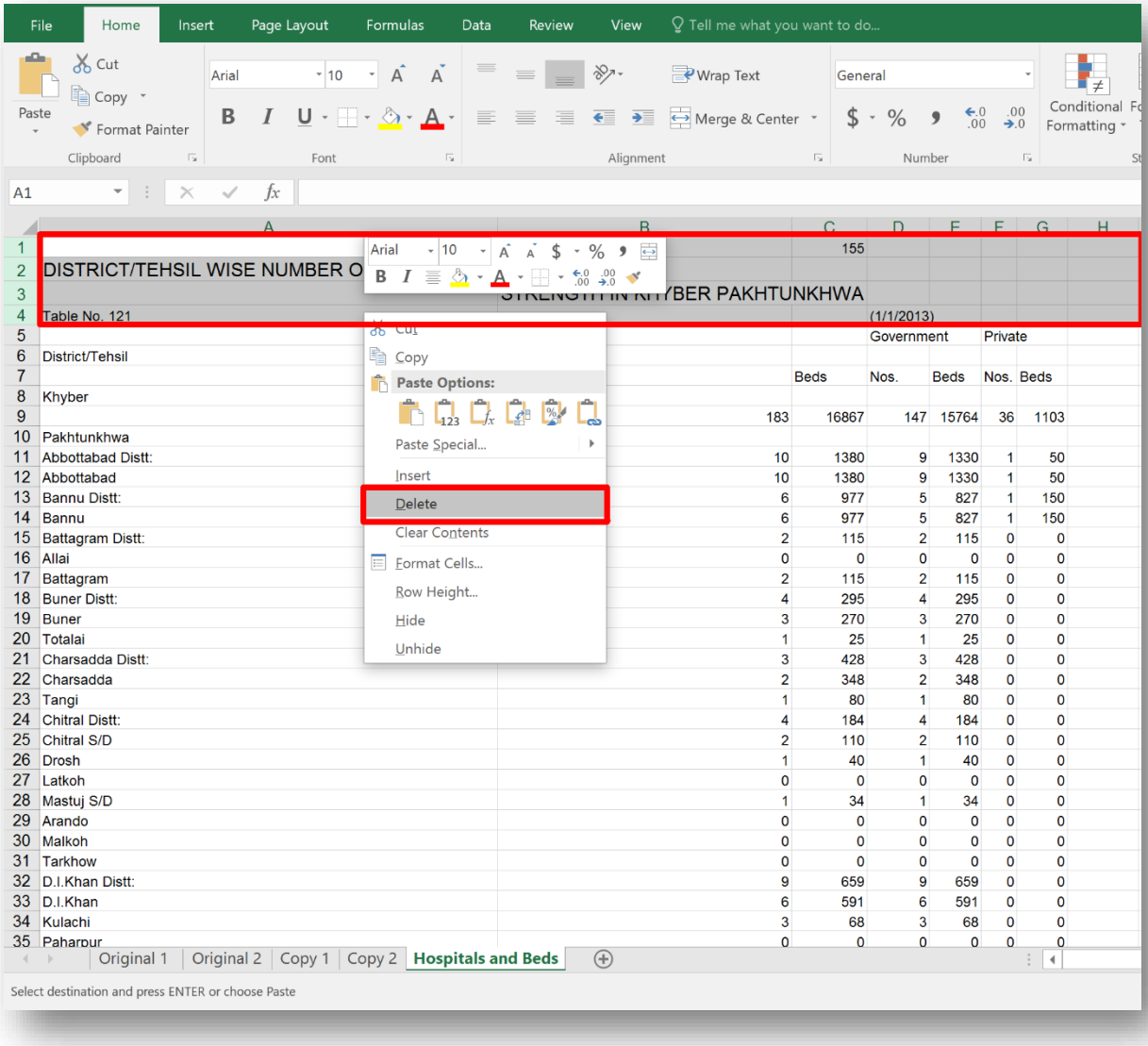


11. Go to Hospital and Beds sheet, go to A1 cell and go Paste.

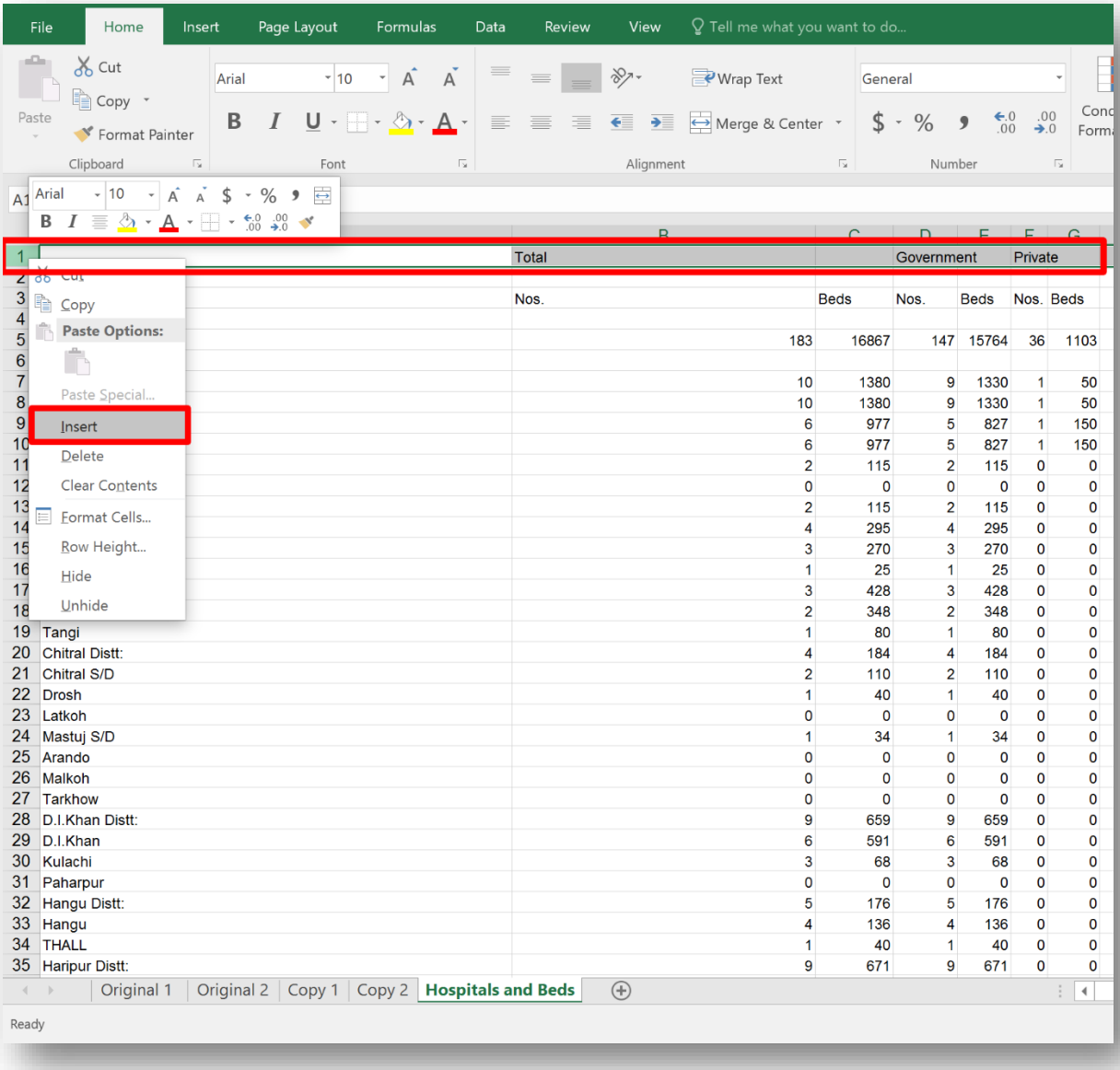


	A	B	C	D	E	F	G	H
1			155					
2	DISTRICT/TEHSIL WISE NUMBER OF HOSPITALS & THEIR BED							
3		STRENGTH IN KHYBER PAKHTUNKHWA						
4	Table No. 121			(1/1/2013)				
5		Total		Government		Private		
6	District/Tehsil							
7		Nos.	Beds	Nos.	Beds	Nos.	Beds	
8	Khyber							
9		183	16867	147	15764	36	1103	
10	Pakhtunkhwa							
11	Abbottabad Distt:	10	1380	9	1330	1	50	
12	Abbottabad	10	1380	9	1330	1	50	
13	Bannu Distt:	6	977	5	827	1	150	
14	Bannu	6	977	5	827	1	150	
15	Battagram Distt:	2	115	2	115	0	0	
16	Allai	0	0	0	0	0	0	
17	Battagram	2	115	2	115	0	0	
18	Buner Distt:	4	295	4	295	0	0	
19	Buner	3	270	3	270	0	0	
20	Totalai	1	25	1	25	0	0	
21	Charsadda Distt:	3	428	3	428	0	0	
22	Charsadda	2	348	2	348	0	0	
23	Tangi	1	80	1	80	0	0	
24	Chitral Distt:	4	184	4	184	0	0	
25	Chitral S/D	2	110	2	110	0	0	
26	Drosh	1	40	1	40	0	0	
27	Latkoh	0	0	0	0	0	0	
28	Mastuj S/D	1	34	1	34	0	0	
29	Arando	0	0	0	0	0	0	
30	Malkoh	0	0	0	0	0	0	
31	Tarkhow	0	0	0	0	0	0	
32	D.I.Khan Distt:	9	659	9	659	0	0	
33	D.I.Khan	6	591	6	591	0	0	
34	Kulachi	3	68	3	68	0	0	
35	Paharpur	0	0	0	0	0	0	

12. First, we need to standardize our column headers my removing extraneous text above the column names. Select rows 1, 2, 3 and 4. Right click on the selected data and click on Delete.



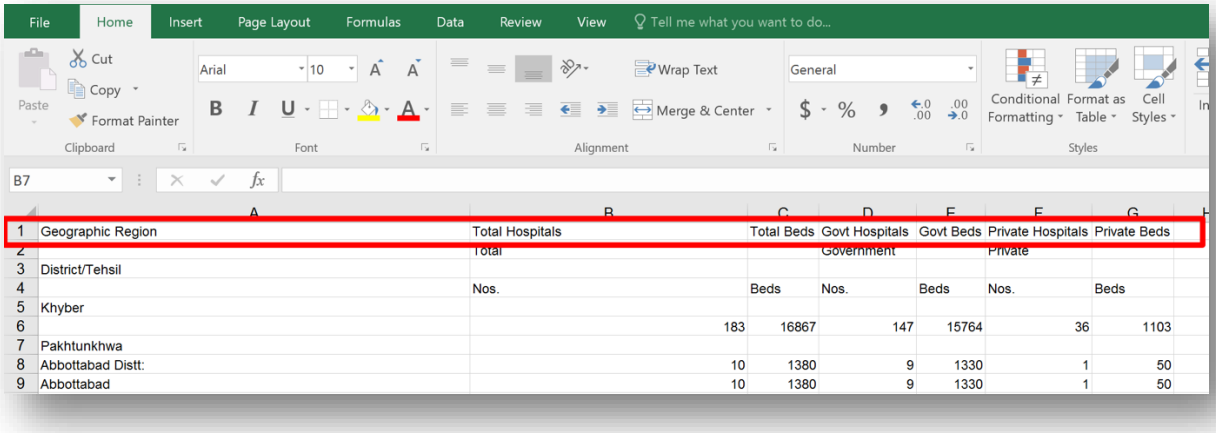
13. Select row 1, right click and click Insert to add a blank row



		Total		Government		Private	
		Nos.	Beds	Nos.	Beds	Nos.	Beds
		183	16867	147	15764	36	1103
		10	1380	9	1330	1	50
		10	1380	9	1330	1	50
		6	977	5	827	1	150
		6	977	5	827	1	150
		2	115	2	115	0	0
		0	0	0	0	0	0
		2	115	2	115	0	0
		4	295	4	295	0	0
		3	270	3	270	0	0
		1	25	1	25	0	0
		3	428	3	428	0	0
		2	348	2	348	0	0
19	Tangi	1	80	1	80	0	0
20	Chitral Distt:	4	184	4	184	0	0
21	Chitral S/D	2	110	2	110	0	0
22	Drosh	1	40	1	40	0	0
23	Latkoh	0	0	0	0	0	0
24	Mastuj S/D	1	34	1	34	0	0
25	Arando	0	0	0	0	0	0
26	Malkoh	0	0	0	0	0	0
27	Tarkhow	0	0	0	0	0	0
28	D.I.Khan Distt:	9	659	9	659	0	0
29	D.I.Khan	6	591	6	591	0	0
30	Kulachi	3	68	3	68	0	0
31	Paharpur	0	0	0	0	0	0
32	Hangu Distt:	5	176	5	176	0	0
33	Hangu	4	136	4	136	0	0
34	THALL	1	40	1	40	0	0
35	Haripur Distt:	9	671	9	671	0	0

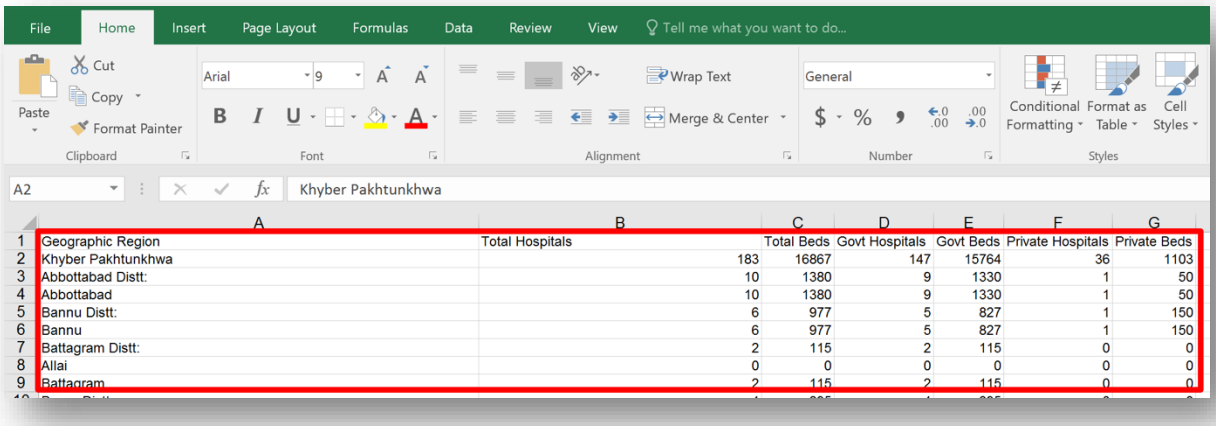
14. Label columns

- A1: Geographic Region
- B1: Total Hospitals
- C1 : Total Beds
- D1: Govt Hospitals
- E1: Govt Beds
- F1: Private Hospitals
- G1: Private Beds



	A	B	C	D	E	F	G
1	Geographic Region	Total Hospitals	Total Beds	Govt Hospitals	Govt Beds	Private Hospitals	Private Beds
2		Total		Government		Private	
3	District/Tehsil						
4		Nos.	Beds	Nos.	Beds	Nos.	Beds
5	Khyber						
6		183	16867	147	15764	36	1103
7	Pakhtunkhwa						
8	Abbottabad Distt:	10	1380	9	1330	1	50
9	Abbottabad	10	1380	9	1330	1	50

15. In cell A6 retype Khyber Pakhtunkwa and Delete row 2, 3, 4, 5 and 7.

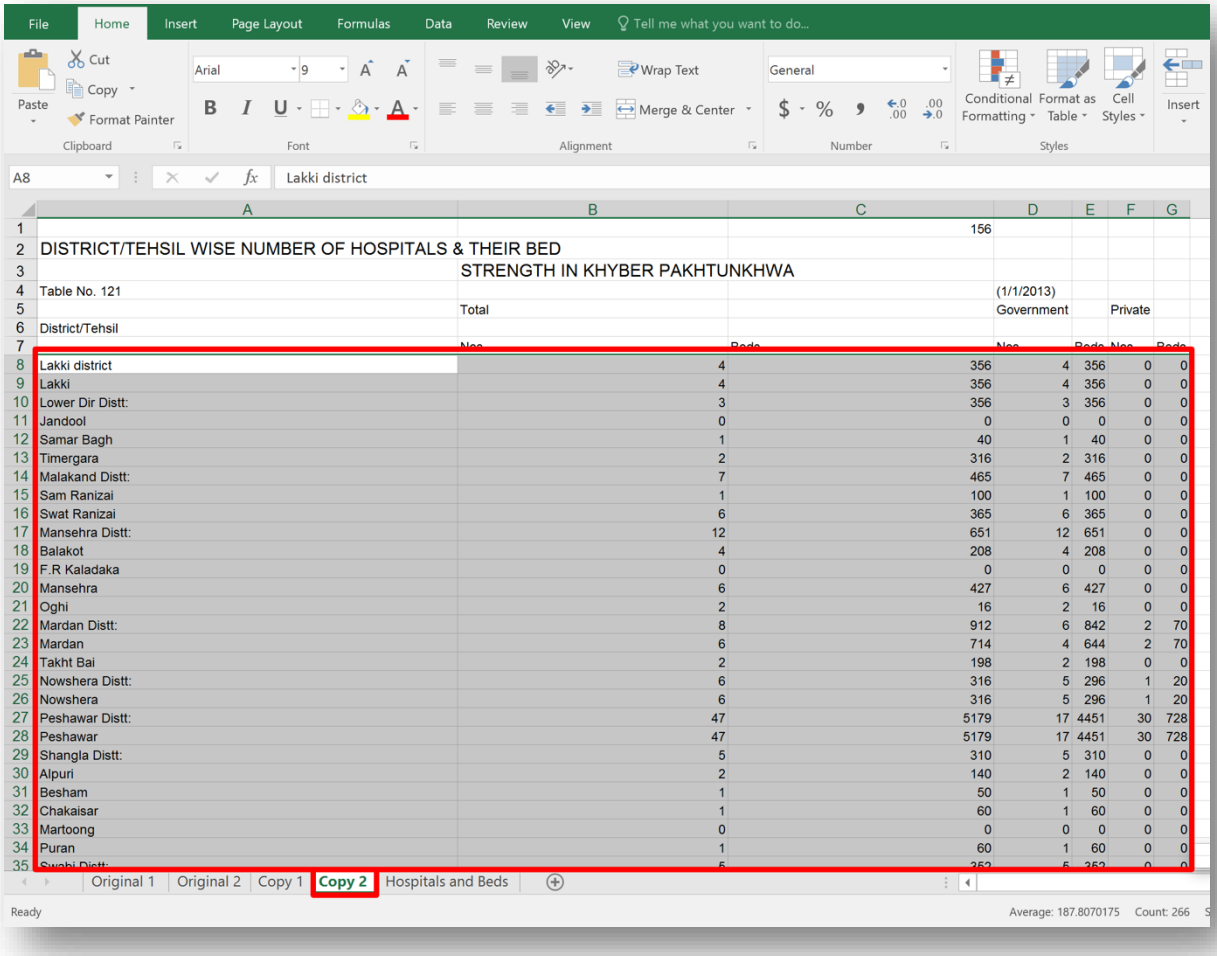


	A	B	C	D	E	F	G
1	Geographic Region	Total Hospitals	Total Beds	Govt Hospitals	Govt Beds	Private Hospitals	Private Beds
2	Khyber Pakhtunkhwa	183	16867	147	15764	36	1103
3	Abbottabad Distt:	10	1380	9	1330	1	50
4	Abbottabad	10	1380	9	1330	1	50
5	Bannu Distt:	6	977	5	827	1	150
6	Bannu	6	977	5	827	1	150
7	Battagram Distt:	2	115	2	115	0	0
8	Allai	0	0	0	0	0	0
9	Battagram	2	115	2	115	0	0

16. Delete (Continued) cell G44

17. Go to Copy 2 rename cell A8 as Lakki District

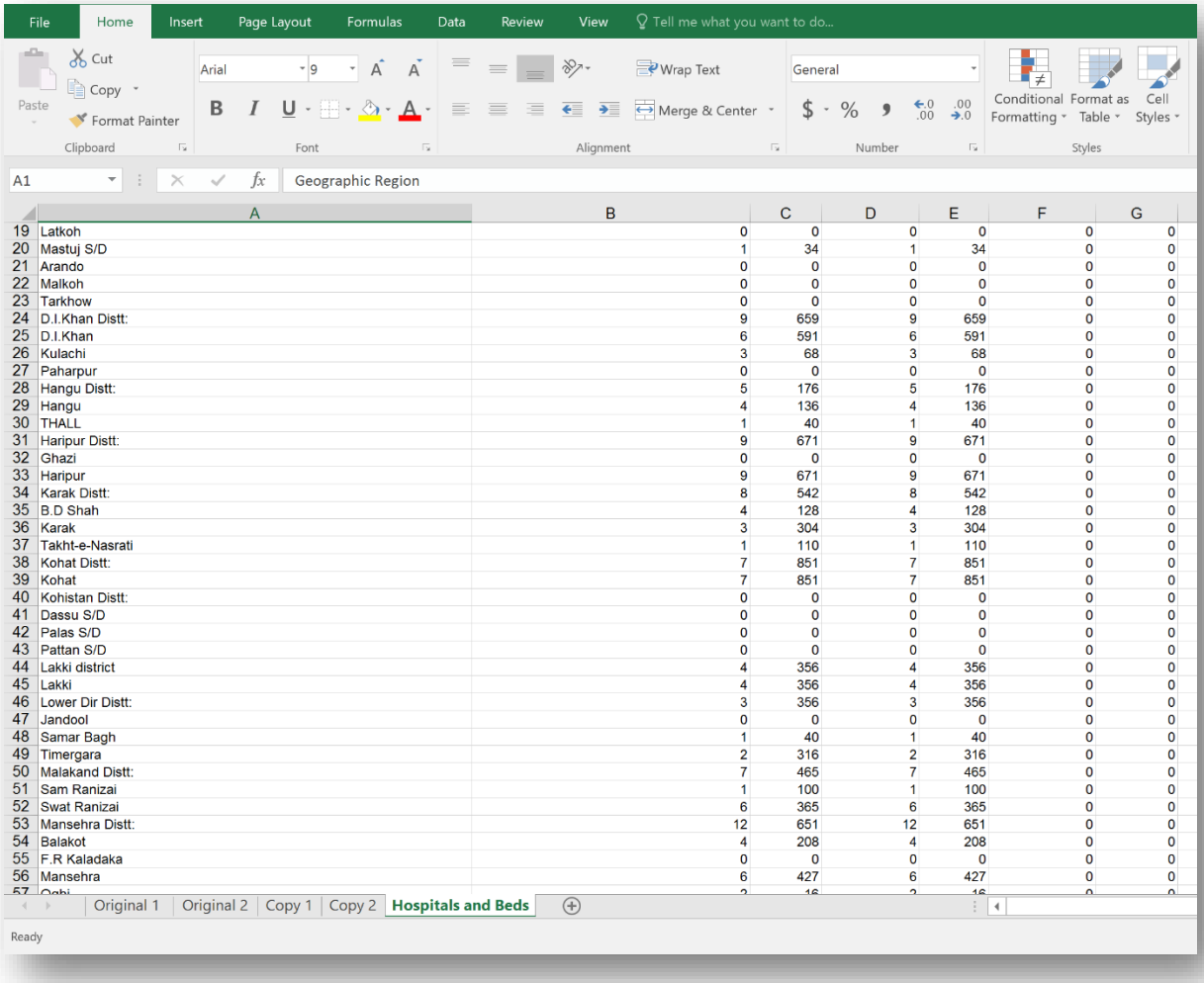
18. In Copy 2 sheet, select cells A8 to G8 and from there select the rows with data (until row 45)



	A	B	C	D	E	F	G
1			156				
2	DISTRICT/TEHSIL WISE NUMBER OF HOSPITALS & THEIR BED						
3		STRENGTH IN KHYBER PAKHTUNKHWA					
4	Table No. 121			(1/1/2013)			
5		Total		Government		Private	
6	District/Tehsil						
7		Nos	Beds	Nos	Beds	Nos	Beds
8	Lakki district	4	356	4	356	0	0
9	Lakki	4	356	4	356	0	0
10	Lower Dir Distt:	3	356	3	356	0	0
11	Jandool	0	0	0	0	0	0
12	Samar Bagh	1	40	1	40	0	0
13	Timergara	2	316	2	316	0	0
14	Malakand Distt:	7	465	7	465	0	0
15	Sam Ranizai	1	100	1	100	0	0
16	Swat Ranizai	6	365	6	365	0	0
17	Mansehra Distt:	12	651	12	651	0	0
18	Balakot	4	208	4	208	0	0
19	F.R Kaladaka	0	0	0	0	0	0
20	Mansehra	6	427	6	427	0	0
21	Oghi	2	16	2	16	0	0
22	Mardan Distt:	8	912	6	842	2	70
23	Mardan	6	714	4	644	2	70
24	Takht Bai	2	198	2	198	0	0
25	Nowshera Distt:	6	316	5	296	1	20
26	Nowshera	6	316	5	296	1	20
27	Peshawar Distt:	47	5179	17	4451	30	728
28	Peshawar	47	5179	17	4451	30	728
29	Shangla Distt:	5	310	5	310	0	0
30	Alpuri	2	140	2	140	0	0
31	Besham	1	50	1	50	0	0
32	Chakaisar	1	60	1	60	0	0
33	Martoong	0	0	0	0	0	0
34	Puran	1	60	1	60	0	0
35	Swabi Distt:	5		5		0	0

19. Copy the data selected.

20. Go to Hospital and Beds cell A44 and paste the data selected.



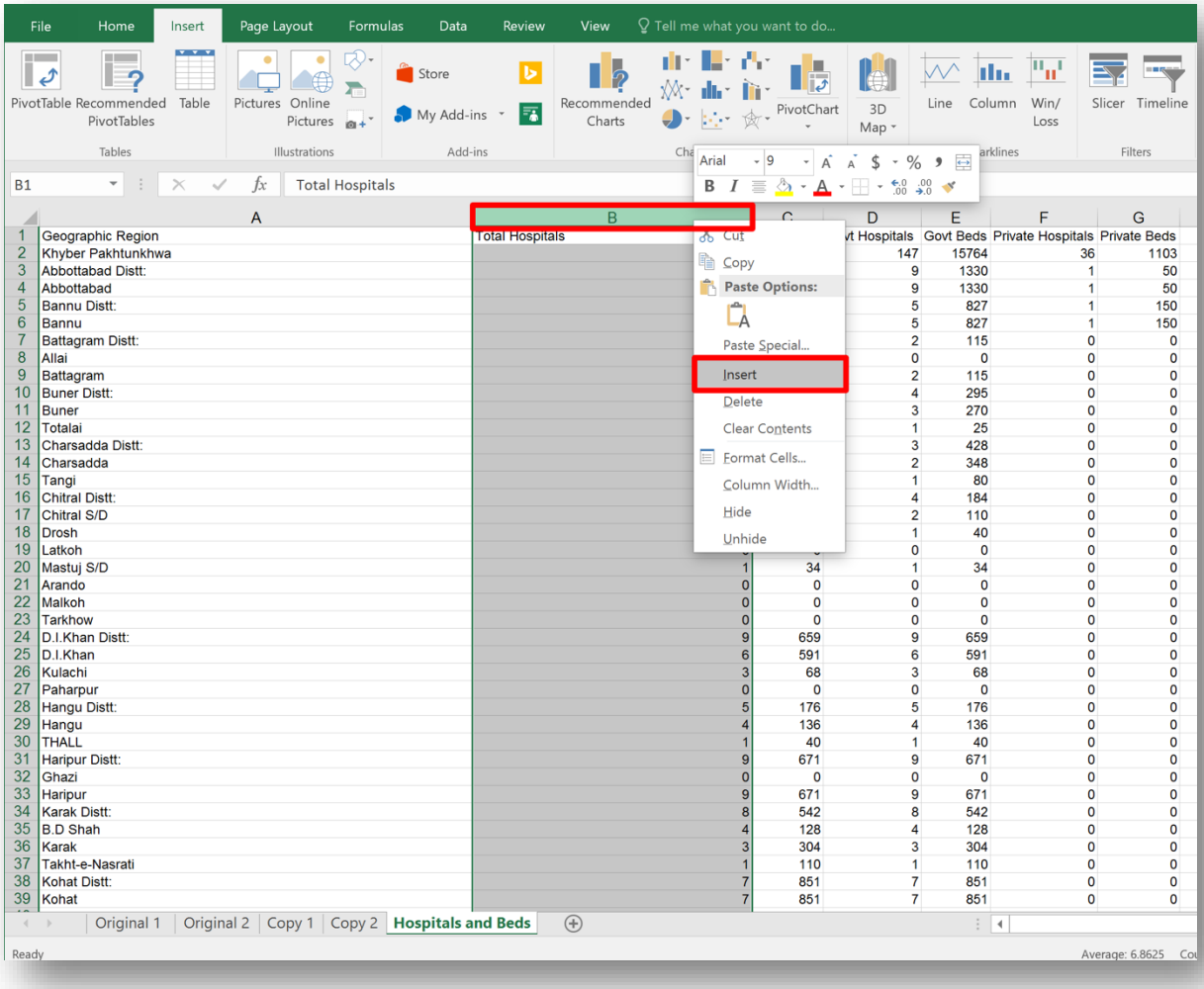
	A	B	C	D	E	F	G
19	Latkoh	0	0	0	0	0	0
20	Mastuj S/D	1	34	1	34	0	0
21	Arando	0	0	0	0	0	0
22	Malkoh	0	0	0	0	0	0
23	Tarkhow	0	0	0	0	0	0
24	D.I.Khan Distt:	9	659	9	659	0	0
25	D.I.Khan	6	591	6	591	0	0
26	Kulachi	3	68	3	68	0	0
27	Paharpur	0	0	0	0	0	0
28	Hangu Distt:	5	176	5	176	0	0
29	Hangu	4	136	4	136	0	0
30	THALL	1	40	1	40	0	0
31	Haripur Distt:	9	671	9	671	0	0
32	Ghazi	0	0	0	0	0	0
33	Haripur	9	671	9	671	0	0
34	Karak Distt:	8	542	8	542	0	0
35	B.D Shah	4	128	4	128	0	0
36	Karak	3	304	3	304	0	0
37	Takht-e-Nasrati	1	110	1	110	0	0
38	Kohat Distt:	7	851	7	851	0	0
39	Kohat	7	851	7	851	0	0
40	Kohistan Distt:	0	0	0	0	0	0
41	Dassu S/D	0	0	0	0	0	0
42	Palas S/D	0	0	0	0	0	0
43	Pattan S/D	0	0	0	0	0	0
44	Lakki district	4	356	4	356	0	0
45	Lakki	4	356	4	356	0	0
46	Lower Dir Distt:	3	356	3	356	0	0
47	Jandool	0	0	0	0	0	0
48	Samar Bagh	1	40	1	40	0	0
49	Timergara	2	316	2	316	0	0
50	Malakand Distt:	7	465	7	465	0	0
51	Sam Ranizai	1	100	1	100	0	0
52	Swat Ranizai	6	365	6	365	0	0
53	Mansehra Distt:	12	651	12	651	0	0
54	Balakot	4	208	4	208	0	0
55	F.R Kaladaka	0	0	0	0	0	0
56	Mansehra	6	427	6	427	0	0

Hospital and Beds Private and Government 2013 (2)

Now we have the Hospital and Bed in Khyber Pakhtunkhwa data in a format that matches the original PDF. But it doesn't yet meet all of our requirements for data standardization in order to work with the data in the next analysis lab.

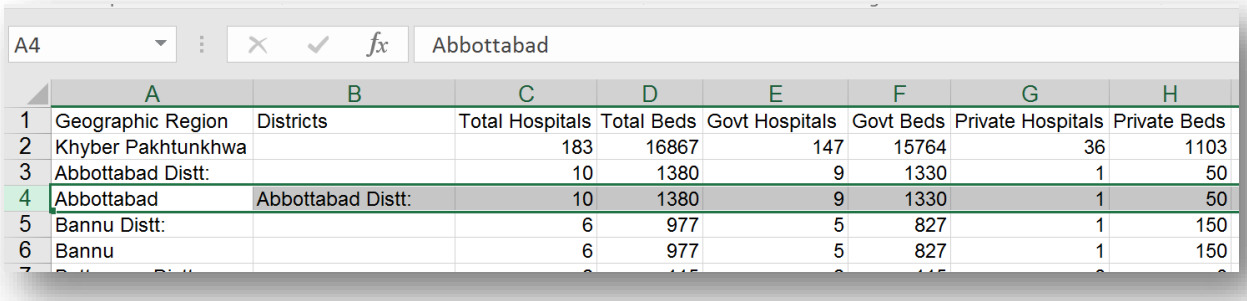
Most of the districts in Khyber Pakhtunkhwa have a subdivision called Tehsil as you can see Abbottabad District has just one tehsil "Abbottabad", but the rows for Abbottabad District and Abbottabad tehsil have the same data because Abbottabad District is the total of hospital and beds in the district. If we try to calculate the total number of hospitals or beds in the district, it will count the same data twice. In the case of Chitral District, there are 7 tehsils and one line for total for the district, preventing us from doing a simple sum. To solve the problem of double counting:

21. First, in Hospital and Beds sheet add another column between A and B by click on Colum B, use right click and select Insert



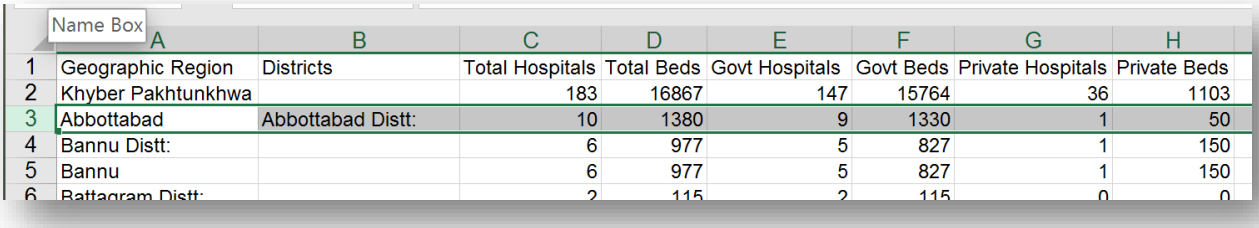
22. Name cell B1: Districts

23. For Abbottabad District copy and paste A3 (Abbottabad Distt:) to B4 cell



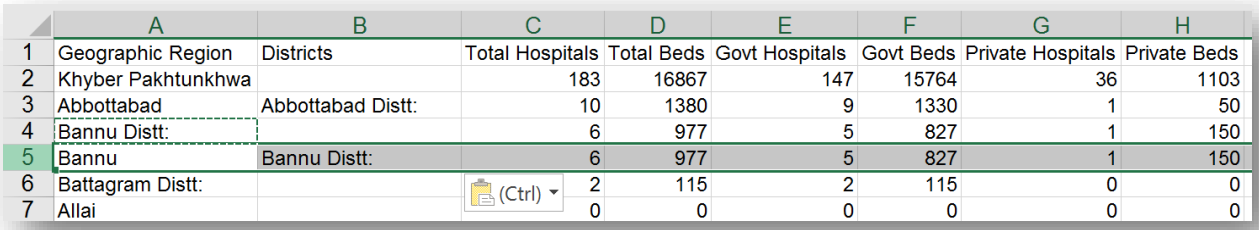
	A	B	C	D	E	F	G	H
1	Geographic Region	Districts	Total Hospitals	Total Beds	Govt Hospitals	Govt Beds	Private Hospitals	Private Beds
2	Khyber Pakhtunkhwa		183	16867	147	15764	36	1103
3	Abbottabad Distt:		10	1380	9	1330	1	50
4	Abbottabad	Abbottabad Distt:	10	1380	9	1330	1	50
5	Bannu Distt:		6	977	5	827	1	150
6	Bannu		6	977	5	827	1	150

24. Then, delete row 3. This leaves us with just one row of data for Abbottabad District.



	A	B	C	D	E	F	G	H
1	Geographic Region	Districts	Total Hospitals	Total Beds	Govt Hospitals	Govt Beds	Private Hospitals	Private Beds
2	Khyber Pakhtunkhwa		183	16867	147	15764	36	1103
3	Abbottabad	Abbottabad Distt:	10	1380	9	1330	1	50
4	Bannu Distt:		6	977	5	827	1	150
5	Bannu		6	977	5	827	1	150
6	Battagram Distt:		2	115	2	115	0	0

25. For Bannu District copy and paste A4 (Bannu Distt:) to B5



	A	B	C	D	E	F	G	H
1	Geographic Region	Districts	Total Hospitals	Total Beds	Govt Hospitals	Govt Beds	Private Hospitals	Private Beds
2	Khyber Pakhtunkhwa		183	16867	147	15764	36	1103
3	Abbottabad	Abbottabad Distt:	10	1380	9	1330	1	50
4	Bannu Distt:		6	977	5	827	1	150
5	Bannu	Bannu Distt:	6	977	5	827	1	150
6	Battagram Distt:		2	115	2	115	0	0
7	Allai		0	0	0	0	0	0

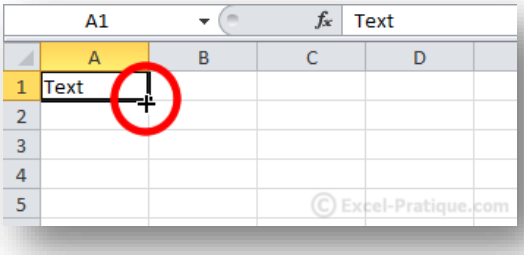
26. Then, delete row 4.

	A	B	C	D	E	F	G	H
1	Geographic Region	Districts	Total Hospitals	Total Beds	Govt Hospitals	Govt Beds	Private Hospitals	Private Beds
2	Khyber Pakhtunkhwa		183	16867	147	15764	36	1103
3	Abbottabad	Abbottabad Distt:	10	1380	9	1330	1	50
4	Bannu	Bannu Distt:	6	977	5	827	1	150
5	Battagram Distt:		2	115	2	115	0	0

27. For Battagram district copy and paste A5 (Battagram Distt:) to B6

	A	B	C	D	E	F	G	H
1	Geographic Region	Districts	Total Hospitals	Total Beds	Govt Hospitals	Govt Beds	Private Hospitals	Private Beds
5	Battagram Distt:		2	115	2	115	0	0
6	Allai	Battagram Distt:	0	0	0	0	0	0
7	Battagram		2	115	2	115	0	0
8	Buner Distt:		4	295	4	295	0	0

28. But, Battagram district has two tehsils, Allai and Battagram. The data also shows you than Allai does not have any hospitals or beds. But **do not delete data even if the columns include only zeros or empty cells**. Go to B6 cell (Battagram Distt:) select it, hover over the lower-right corner of the cell until the cursor icon changes to a black thick cross; drag and drop B6 cell (Battagram Distt:) until B7 to fill in both tehsils



	A	B	C	D	E	F	G	H
1	Geographic Region	Districts	Total Hospitals	Total Beds	Govt Hospitals	Govt Beds	Private Hospitals	Private Beds
5	Battagram Distt:		2	115	2	115	0	0
6	Allai	Battagram Distt:	0	0	0	0	0	0
7	Battagram	Battagram Distt:	2	115	2	115	0	0
8	Buner Distt:		4	295	4	295	0	0

29. And delete row 5 to avoid repeated data as explain previously.

	A	B	C	D	E	F	G	H
1	Geographic Region	Districts	Total Hospitals	Total Beds	Govt Hospitals	Govt Beds	Private Hospitals	Private Beds
2	Khyber Pakhtunkhwa		183	16867	147	15764	36	1103
3	Abbottabad	Abbottabad Distt:	10	1380	9	1330	1	50
4	Bannu	Bannu Distt:	6	977	5	827	1	150
5	Allai	Battagram Distt:	0	0	0	0	0	0
6	Battagram	Battagram Distt:	2	115	2	115	0	0

30. For Buner district copy and paste A7 (Buner Distt:) to B8

	A	B	C	D	E	F	G	H
1	Geographic Region	Districts	Total Hospitals	Total Beds	Govt Hospitals	Govt Beds	Private Hospitals	Private Beds
7	Buner Distt:		4	295	4	295	0	0
8	Buner	Buner Distt:	3	270	3	270	0	0
9	Totalai		1	25	1	25	0	0
10	Charsadda Distt:		3	428	3	428	0	0

31. Buner district has two tehsils, Buner and Totalai, so drag and drop B8 (Buner Distt:) until B9 to fill in the district for both tehsils.

	A	B	C	D	E	F	G	H
1	Geographic Region	Districts	Total Hospitals	Total Beds	Govt Hospitals	Govt Beds	Private Hospitals	Private Beds
7	Buner Distt:		4	295	4	295	0	0
8	Buner	Buner Distt:	3	270	3	270	0	0
9	Totalai	Buner Distt:	1	25	1	25	0	0
10	Charsadda Distt:		3	428	3	428	0	0

32. Delete row 7

	A	B	C	D	E	F	G	H
1	Geographic Region	Districts	Total Hospitals	Total Beds	Govt Hospitals	Govt Beds	Private Hospitals	Private Beds
2	Khyber Pakhtunkhwa		183	16867	147	15764	36	1103
3	Abbottabad	Abbottabad Distt:	10	1380	9	1330	1	50
4	Bannu	Bannu Distt:	6	977	5	827	1	150
5	Allai	Battagram Distt:	0	0	0	0	0	0
6	Battagram	Battagram Distt:	2	115	2	115	0	0
7	Buner	Buner Distt:	3	270	3	270	0	0
8	Totalai	Buner Distt:	1	25	1	25	0	0

33. For Charsadda district copy and paste A9 (Charsadda Distt:) to B10

34. Charsadda has 2 tehsils, Charsadda and Tangi, so drag and drop B10 (Charsadda Distt:) until B11 to fill in the district for both tehsils

35. Delete row 9

	A	B	C	D	E	F	G	H
1	Geographic Region	Districts	Total Hospitals	Total Beds	Govt Hospitals	Govt Beds	Private Hospitals	Private Beds
2	Khyber Pakhtunkhwa		183	16867	147	15764	36	1103
3	Abbottabad	Abbottabad Distt:	10	1380	9	1330	1	50
4	Bannu	Bannu Distt:	6	977	5	827	1	150
5	Allai	Battagram Distt:	0	0	0	0	0	0
6	Battagram	Battagram Distt:	2	115	2	115	0	0
7	Buner	Buner Distt:	3	270	3	270	0	0
8	Totalai	Buner Distt:	1	25	1	25	0	0
9	Charsadda	Charsadda Distt:	2	348	2	348	0	0
10	Tangi	Charsadda Distt:	1	80	1	80	0	0

36. Do the same process with the rest of the districts. Be very careful and remember the use Undo in case of mistakes.

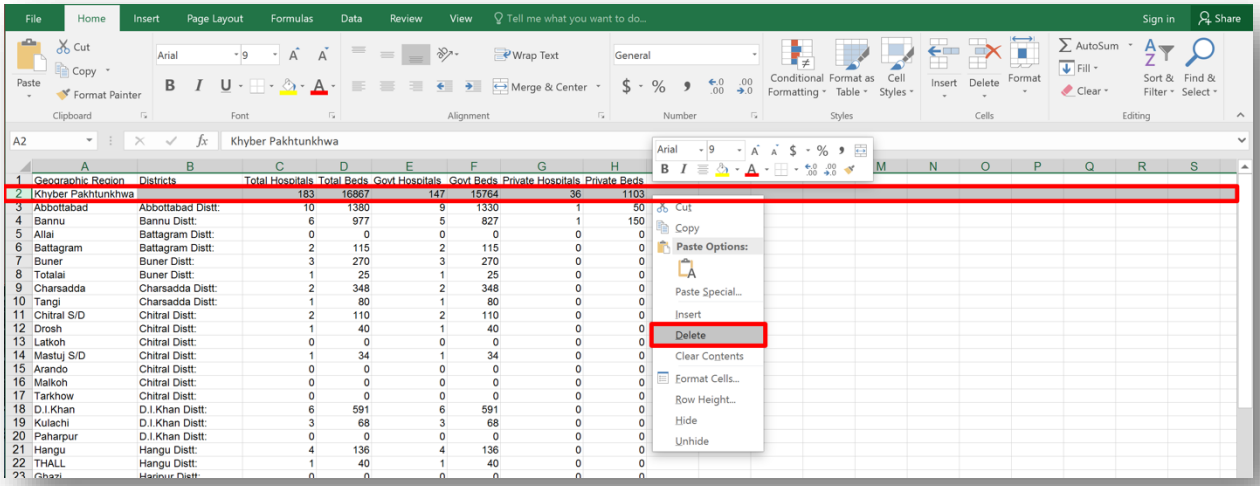
37. There are 24 districts in the data set. And you should finish with a total of 57 rows.

	A	B	C	D	E	F	G	H
1	Geographic Region	Districts	Total Hospitals	Total Beds	Govt Hospitals	Govt Beds	Private Hospitals	Private Beds
2	Khyber Pakhtunkhwa		183	16867	147	15764	36	1103
3	Abbottabad	Abbottabad Distt:	10	1380	9	1330	1	50
4	Bannu	Bannu Distt:	6	977	5	827	1	150
5	Allai	Battagram Distt:	0	0	0	0	0	0
6	Battagram	Battagram Distt:	2	115	2	115	0	0
7	Buner	Buner Distt:	3	270	3	270	0	0
8	Totalai	Buner Distt:	1	25	1	25	0	0
9	Charsadda	Charsadda Distt:	2	348	2	348	0	0
10	Tangi	Charsadda Distt:	1	80	1	80	0	0
11	Chitral S/D	Chitral Distt:	2	110	2	110	0	0
12	Drosh	Chitral Distt:	1	40	1	40	0	0
13	Latkoh	Chitral Distt:	0	0	0	0	0	0
14	Mastuj S/D	Chitral Distt:	1	34	1	34	0	0
15	Arando	Chitral Distt:	0	0	0	0	0	0
16	Malkoh	Chitral Distt:	0	0	0	0	0	0
17	Tarkhow	Chitral Distt:	0	0	0	0	0	0
18	D.I.Khan	D.I.Khan Distt:	6	591	6	591	0	0
19	Kulachi	D.I.Khan Distt:	3	68	3	68	0	0
20	Paharpur	D.I.Khan Distt:	0	0	0	0	0	0
21	Hangu	Hangu Distt:	4	136	4	136	0	0
22	THALL	Hangu Distt:	1	40	1	40	0	0
23	Ghazi	Haripur Distt:	0	0	0	0	0	0
24	Haripur	Haripur Distt:	9	671	9	671	0	0
25	B.D Shah	Karak Distt:	4	128	4	128	0	0
26	Karak	Karak Distt:	3	304	3	304	0	0
27	Takht-e-Nasrati	Karak Distt:	1	110	1	110	0	0
28	Kohat	Kohat Distt:	7	851	7	851	0	0
29	Dassu S/D	Kohistan Distt:	0	0	0	0	0	0
30	Palas S/D	Kohistan Distt:	0	0	0	0	0	0
31	Pattan S/D	Kohistan Distt:	0	0	0	0	0	0
32	Lakki	Lakki Distt:	4	356	4	356	0	0
33	Jandool	Lower Dir Distt:	0	0	0	0	0	0
34	Samar Bagh	Lower Dir Distt:	1	40	1	40	0	0
35	Timergara	Lower Dir Distt:	2	316	2	316	0	0
36	Sam Ranizai	Malakand Distt:	1	100	1	100	0	0
37	Swat Ranizai	Malakand Distt:	6	365	6	365	0	0
38	Balakot	Mansehra Distt:	4	208	4	208	0	0
39	F.R Kaladaka	Mansehra Distt:	0	0	0	0	0	0
40	Mansehra	Mansehra Distt:	6	427	6	427	0	0
41	Oghi	Mansehra Distt:	2	16	2	16	0	0
42	Mardan	Mardan Distt:	6	714	4	644	2	70
43	Takht Bai	Mardan Distt:	2	198	2	198	0	0
44	Nowshera	Nowshera Distt:	6	316	5	296	1	20
45	Peshawar	Peshawar Distt:	47	5179	17	4451	30	728
46	Alpuri	Shangla Distt:	2	140	2	140	0	0
47	Besham	Shangla Distt:	1	50	1	50	0	0
48	Chakaisar	Shangla Distt:	1	60	1	60	0	0
49	Martoong	Shangla Distt:	0	0	0	0	0	0
50	Puran	Shangla Distt:	1	60	1	60	0	0
51	Lahore	Swabi Distt:	0	0	0	0	0	0
52	Swabi	Swabi Distt:	5	352	5	352	0	0
53	Matta	Swat Distt:	1	80	1	80	0	0
54	Swat	Swat Distt:	9	759	9	759	0	0
55	Tank	Tank Distt:	4	165	3	80	1	85
56	Dir S/D	Upper Dir Distt:	4	568	4	568	0	0
57	Wari S/D	Upper Dir Distt:	1	120	1	120	0	0
58								

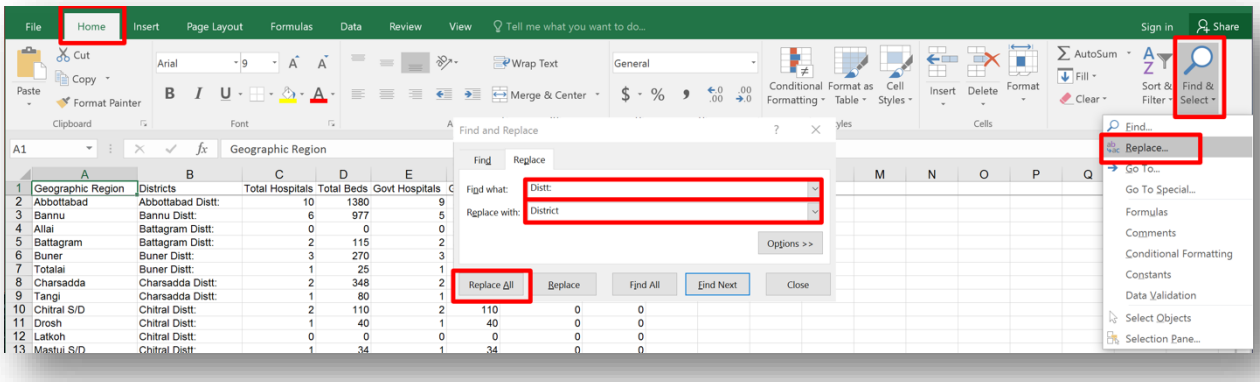
Original 1 | Original 2 | Copy 1 | Copy 2 | **Hospitals and Beds**

38. One way to determine if your cleaning process was successful is to select cells C3 to C57 and click check the total displayed in the bottom right corner. If you get a total of 183, you have found the total number of hospitals in the Khyber Pakhtunkhwa province.

39. Delete row 2 (Khyber Pakhtunkhwa). This is a total of hospitals and beds in the province. This data is easy to calculate through a sum function. More importantly, it does not represent a discreet piece of data so should be deleted to avoid double counting in further calculation.



40. In column B Districts, all the districts are abbreviated as “Distt:” Change the name to District. To do this, go to Home – Editing - Replace. In the new window type in Find what: Distt: and in Replace with: District and click on Replace all.



41. In A21 the Tehsil THALL was type in uppercase. Change it to Thall. The idea is to standardize all formatting.

42. Also in A17, B17, B18, and B19, change D.I. Khan to Dera Ismail Khan with the replace tool. Avoid abbreviations unless there is an established convention for abbreviation.

43. OK, looks like our data set is ready for analysis.

CLEANING EXERCISE:

44. In lab 1 1. Scraping Data from PDFs and Images, you scraped a document called Sanctioned and actual strength of employees.pdf using cometdocs.com and produced an Excel file called Sanctioned and actual strength of employees.xlsx. Use this file for the next cleaning exercise.

45. Open **Sanctioned and actual strength of employees.xlsx**

46. Remember the Excel first steps.

- **a.** Delete Empty sheet
- **b.** Rename sheets
- **c.** Make a copy
- **d.** Rename the copy
- **e.** Save changes.

47. Compare the data in the pdf document with the Excel document

48. Always do cleaning by comparing the source file (in this case the pdf) to the data file you are cleaning.

49. Count how may columns there are in the Excel document and type in the correct names for these columns according to the data.

50. Delete first row to ensure that column headers are in the first row.

51. This data set also mixes data types in Column A between primary categorization and sub-categorization. For example, Cabinet has 17 categories. We do not want to mix data types in a single column.

52. Leave Colum B (Numbers of the Corporation Body)

53. Delete Total at the end of the cleaning

54. The goal it to produce a clean Excel worksheet that is ready for analysis.

55. Be aware that the last page has multiple cleaning problems

56. To verify if your cleaning is complete, check the code in Colum B (Numbers of the Corporation Body). There are 210 distinct entries for Corporation/Body and so you should have 211 rows of data.

57. Also check your totals with the totals in the original PDF.