3. Data Cleaning

Data cleaning is an essential step that should be conducted before analyzing and visualizing.

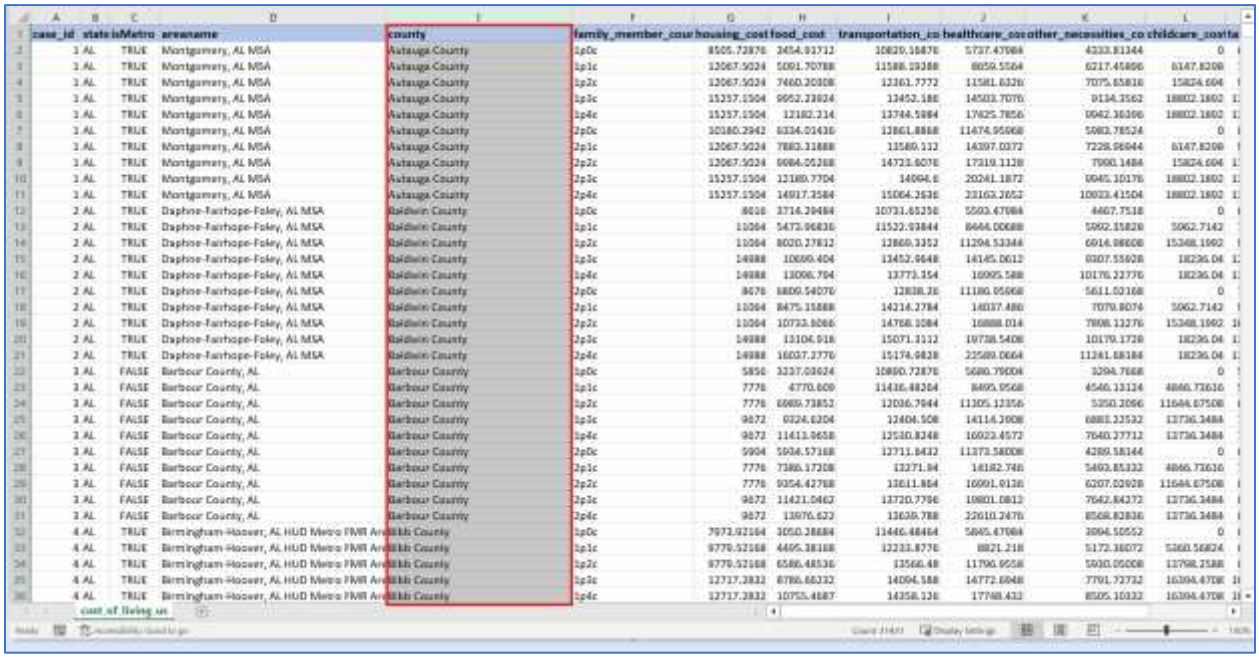
Data cleaning involves fixing or removing incorrect, incorrectly formatted, duplicate, or incomplete data within a dataset.

3.1. Removing Repeated Word

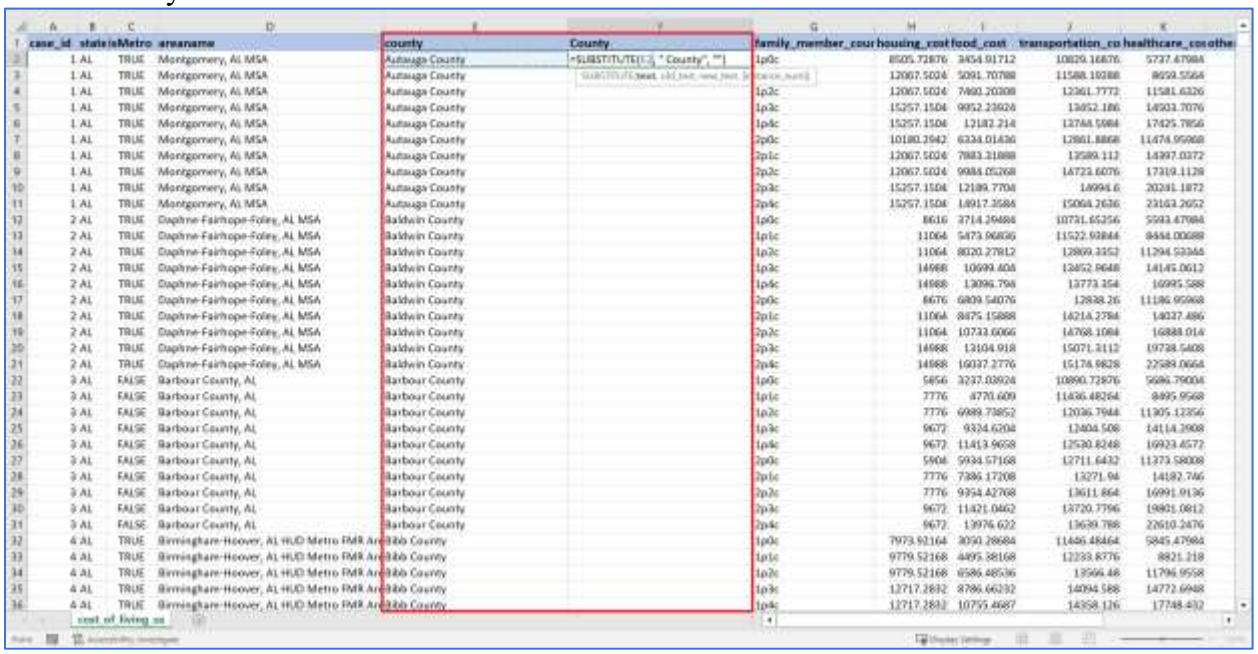
Column Name: County

Before:

The "County" column includes value like "Los Angeles/Autauga County," where the word "County" was repeated. To fix this and display only the county name, we used a function called SUBSTITUTE() to take out the extra word.

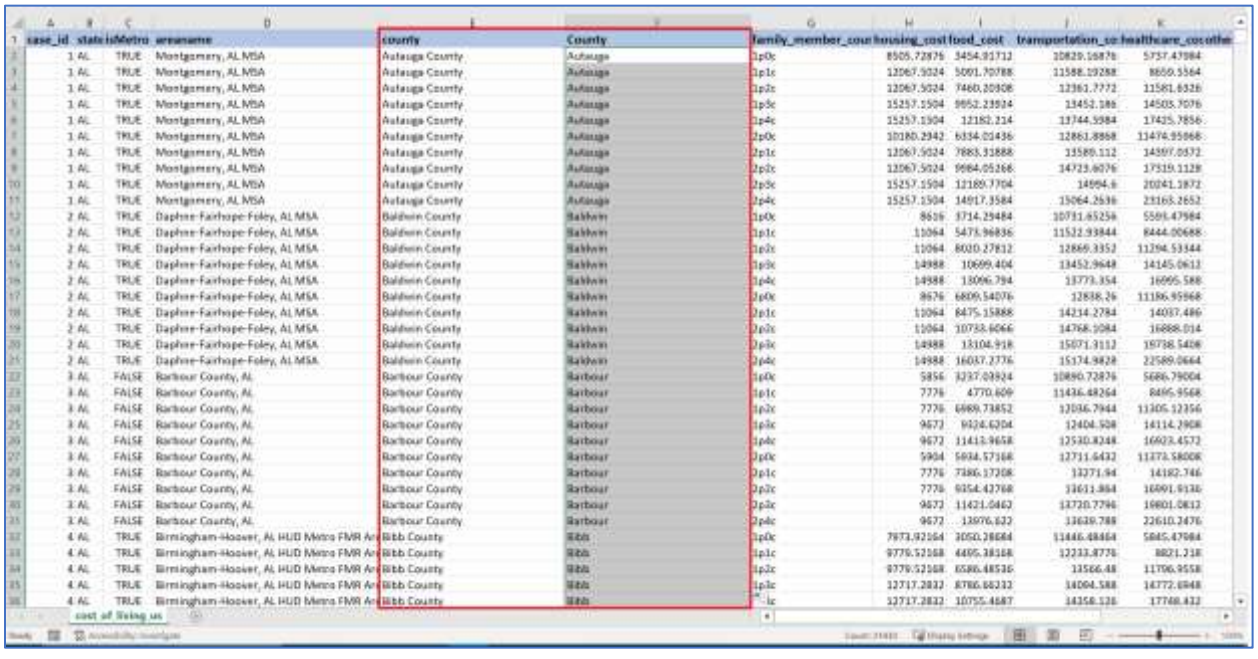


SUBSTITUTE () is used to replace the word “County” with “ “(blank spaces). Column E refers to the "county" column.



After:

Once the repeated word has been removed, it now displays only the county name.



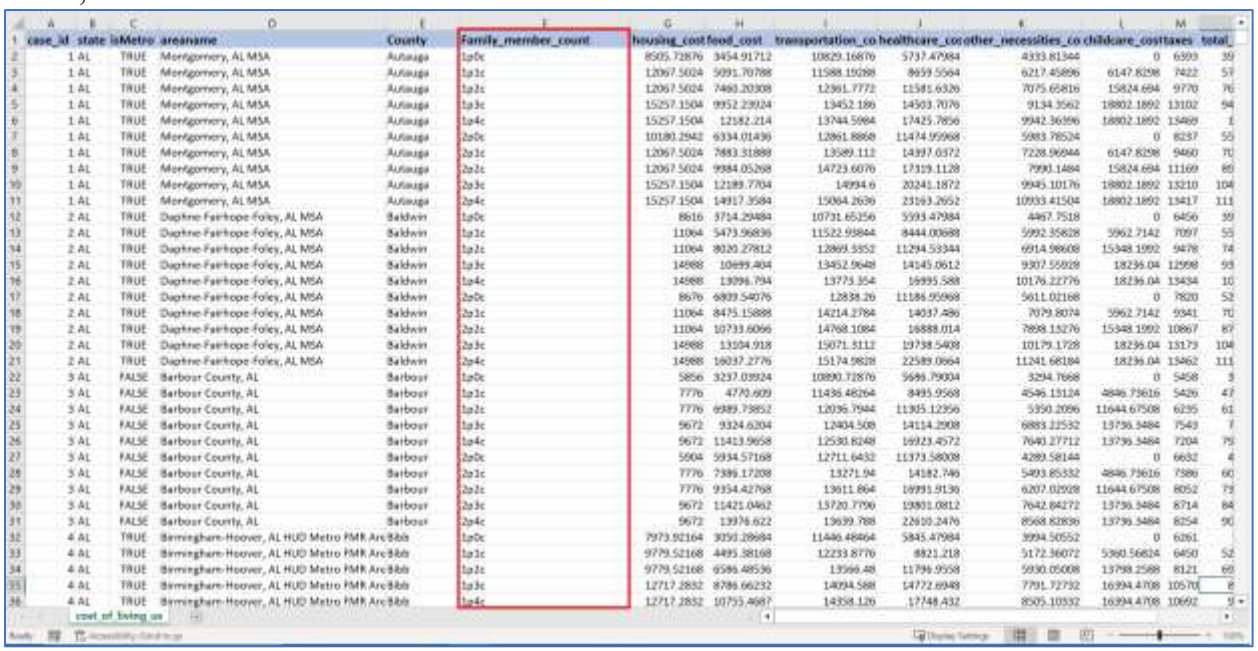
3.2. Separating Value from one column to two column

Before Column Name: Family Member Count

After Cleaning Column Name: Adult_count

Before:

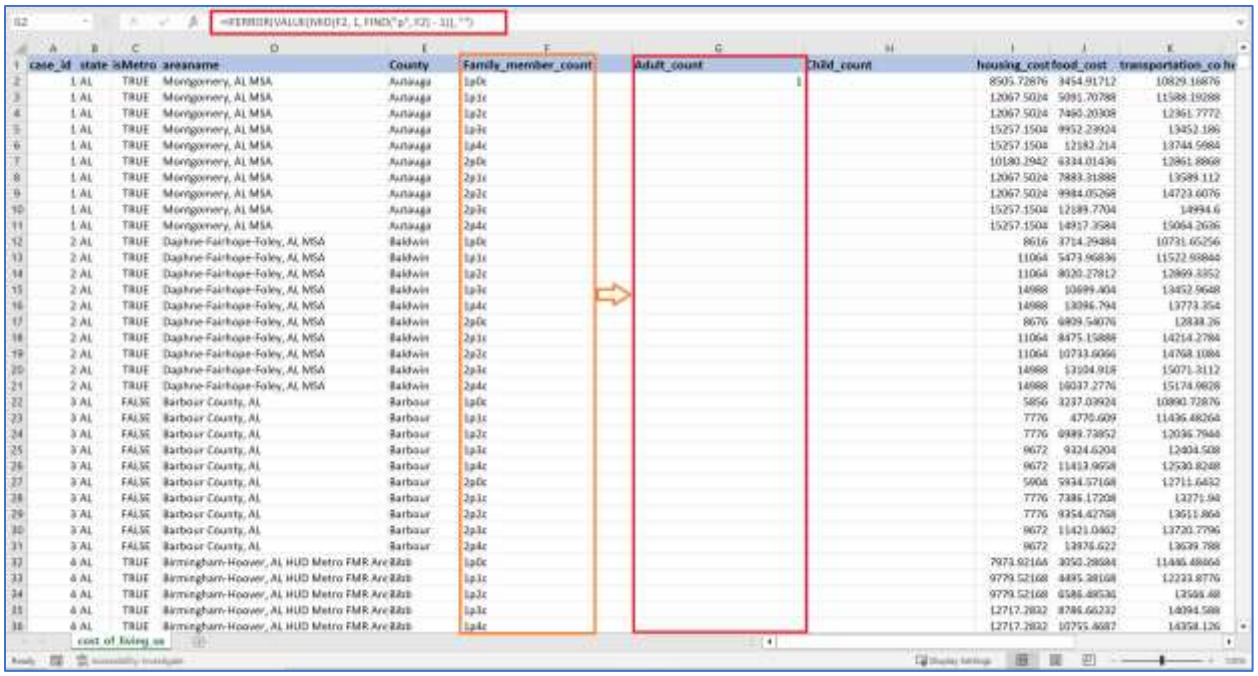
The “Family Member Count” column contains two types of values: "p" indicates the number of adults, and "c" indicates the number of children.



case_id	state	isMetro	areaname	County	Family_member_count	housing_cost	food_cost	transportation_co	healthcare_co	other_necessities_co	childcare_cost	taxes	total_
1	AL	TRUE	Montgomery, AL MSA	Autauga	1p0c	8505.72876	3454.91712	10829.16876	5737.47984	4333.81344	0	6393	39
1	AL	TRUE	Montgomery, AL MSA	Autauga	1p1c	12067.5024	5091.70788	11588.19288	8659.5564	6217.45896	6147.8298	7422	57
1	AL	TRUE	Montgomery, AL MSA	Autauga	1p2c	12067.5024	7480.20308	12361.7772	11581.6326	7075.65816	15824.694	9770	76
1	AL	TRUE	Montgomery, AL MSA	Autauga	1p3c	15257.1504	9952.23924	13452.186	14503.7076	9134.3562	18802.1892	13102	94
1	AL	TRUE	Montgomery, AL MSA	Autauga	1p4c	15257.1504	12182.214	13744.5984	17425.7856	9942.36396	18802.1892	13469	1
1	AL	TRUE	Montgomery, AL MSA	Autauga	2p0c	10180.2942	6334.01436	12861.8868	11474.99568	5983.78524	0	8237	55
1	AL	TRUE	Montgomery, AL MSA	Autauga	2p1c	12067.5024	7883.31888	13589.112	14397.0372	7228.96944	6147.8298	9460	70
1	AL	TRUE	Montgomery, AL MSA	Autauga	2p2c	12067.5024	9984.05268	14723.6076	17319.1128	7990.1484	15824.694	11169	85
1	AL	TRUE	Montgomery, AL MSA	Autauga	2p3c	15257.1504	12189.7704	14994.6	20241.1872	9945.10176	18802.1892	13210	104
1	AL	TRUE	Montgomery, AL MSA	Autauga	2p4c	15257.1504	14917.3584	15064.2636	23163.2652	10933.41504	18802.1892	13417	111
2	AL	TRUE	Daphne-Fairhope-Foley, AL MSA	Baldwin	1p0c	8616	3714.29484	10731.65256	5593.47984	4467.7518	0	6456	39
2	AL	TRUE	Daphne-Fairhope-Foley, AL MSA	Baldwin	1p1c	11064	5473.96836	11522.95844	8444.00688	5992.35828	5962.7142	7097	55
2	AL	TRUE	Daphne-Fairhope-Foley, AL MSA	Baldwin	1p2c	11064	8020.27812	12869.5352	11294.53344	6914.98608	15348.1992	9478	74
2	AL	TRUE	Daphne-Fairhope-Foley, AL MSA	Baldwin	1p3c	14988	10899.404	13452.9648	14145.0612	9307.55028	18236.04	12998	99
2	AL	TRUE	Daphne-Fairhope-Foley, AL MSA	Baldwin	1p4c	14988	13096.794	13773.354	16995.588	10176.22776	18236.04	13434	10
2	AL	TRUE	Daphne-Fairhope-Foley, AL MSA	Baldwin	2p0c	8676	6809.54076	12838.26	11186.95968	5611.02168	0	7820	52
2	AL	TRUE	Daphne-Fairhope-Foley, AL MSA	Baldwin	2p1c	11064	8475.15888	14214.2784	14037.486	7079.8074	5962.7142	9341	70
2	AL	TRUE	Daphne-Fairhope-Foley, AL MSA	Baldwin	2p2c	11064	10733.6066	14768.1084	16888.014	7898.15276	15348.1992	10867	87
2	AL	TRUE	Daphne-Fairhope-Foley, AL MSA	Baldwin	2p3c	14988	13104.918	15071.3112	19738.5408	10179.1728	18236.04	13179	104
2	AL	TRUE	Daphne-Fairhope-Foley, AL MSA	Baldwin	2p4c	14988	16037.2776	15174.9828	22589.0664	11241.68184	18236.04	13462	111
3	AL	FALSE	Barbour County, AL	Barbour	1p0c	5856	3237.03924	10890.72876	5686.79004	3294.7668	0	5458	3
3	AL	FALSE	Barbour County, AL	Barbour	1p1c	7776	4770.609	11436.48264	8493.9568	4546.13124	4846.73616	5426	47
3	AL	FALSE	Barbour County, AL	Barbour	1p2c	7776	6989.73852	12036.7944	11305.12356	5350.2096	11644.67508	6235	61
3	AL	FALSE	Barbour County, AL	Barbour	1p3c	9672	9324.6204	12404.508	14114.2908	6883.22532	13736.3484	7543	7
3	AL	FALSE	Barbour County, AL	Barbour	1p4c	9672	11413.9656	12530.8248	16923.4572	7640.27712	13736.3484	7204	75
3	AL	FALSE	Barbour County, AL	Barbour	2p0c	5904	5934.57168	12711.6432	11373.58008	4289.58144	0	6632	4
3	AL	FALSE	Barbour County, AL	Barbour	2p1c	7776	7386.17208	13271.94	14182.746	5403.85332	4846.73616	7586	60
3	AL	FALSE	Barbour County, AL	Barbour	2p2c	7776	9354.42768	13611.864	16991.9136	6207.02928	11644.67508	8052	73
3	AL	FALSE	Barbour County, AL	Barbour	2p3c	9672	11421.0462	13720.7796	19801.0812	7642.84272	13736.3484	8714	84
3	AL	FALSE	Barbour County, AL	Barbour	2p4c	9672	13976.622	13639.788	22610.2476	8568.82836	13736.3484	8254	90
4	AL	TRUE	Birmingham-Hoover, AL HUD Metro FMR Are	Bibb	1p0c	7973.92164	3050.28684	11446.48464	5845.47984	3994.50552	0	6261	
4	AL	TRUE	Birmingham-Hoover, AL HUD Metro FMR Are	Bibb	1p1c	9779.52168	4495.38168	12233.8776	8821.218	5172.36072	5360.56824	6450	52
4	AL	TRUE	Birmingham-Hoover, AL HUD Metro FMR Are	Bibb	1p2c	9779.52168	6586.48536	13566.48	11796.9558	5930.05008	13798.2588	8121	69
4	AL	TRUE	Birmingham-Hoover, AL HUD Metro FMR Are	Bibb	1p3c	12717.2832	8786.66232	14094.588	14772.6948	7791.72732	16394.4708	10570	8
4	AL	TRUE	Birmingham-Hoover, AL HUD Metro FMR Are	Bibb	1p4c	12717.2832	10755.4687	14358.126	17748.432	8505.10332	16394.4708	10692	9

For analysis purposes, we need to separate this value into Adult_count and Child_count.

- <u>Number of adult(Adult_count):</u>



case_id	state	isMetro	areaname	County	Family_member_count	Adult_count	Child_count	housing_cost	food_cost	transportation_co
1	AL	TRUE	Montgomery, AL MSA	Autauga	1p0c	1		8505.72876	3454.91712	10829.16876
1	AL	TRUE	Montgomery, AL MSA	Autauga	1p1c			12067.5024	5091.70788	11588.19288
1	AL	TRUE	Montgomery, AL MSA	Autauga	1p2c			12067.5024	7480.20308	12361.7772
1	AL	TRUE	Montgomery, AL MSA	Autauga	1p3c			15257.1504	9952.23924	13452.186
1	AL	TRUE	Montgomery, AL MSA	Autauga	1p4c			15257.1504	12182.214	13744.5984
1	AL	TRUE	Montgomery, AL MSA	Autauga	2p0c			10180.2942	6334.01436	12861.8868
1	AL	TRUE	Montgomery, AL MSA	Autauga	2p1c			12067.5024	7883.31888	13589.112
1	AL	TRUE	Montgomery, AL MSA	Autauga	2p2c			12067.5024	9984.05268	14723.6076
1	AL	TRUE	Montgomery, AL MSA	Autauga	2p3c			15257.1504	12189.7704	14994.6
1	AL	TRUE	Montgomery, AL MSA	Autauga	2p4c			15257.1504	14917.3584	15064.2636
2	AL	TRUE	Daphne-Fairhope-Foley, AL MSA	Baldwin	1p0c			8616	3714.29484	10731.65256
2	AL	TRUE	Daphne-Fairhope-Foley, AL MSA	Baldwin	1p1c			11064	5473.96836	11522.95844
2	AL	TRUE	Daphne-Fairhope-Foley, AL MSA	Baldwin	1p2c			11064	8020.27812	12869.5352
2	AL	TRUE	Daphne-Fairhope-Foley, AL MSA	Baldwin	1p3c			14988	10899.404	13452.9648
2	AL	TRUE	Daphne-Fairhope-Foley, AL MSA	Baldwin	1p4c			14988	13096.794	13773.354
2	AL	TRUE	Daphne-Fairhope-Foley, AL MSA	Baldwin	2p0c			8676	6809.54076	12838.26
2	AL	TRUE	Daphne-Fairhope-Foley, AL MSA	Baldwin	2p1c			11064	8475.15888	14214.2784
2	AL	TRUE	Daphne-Fairhope-Foley, AL MSA	Baldwin	2p2c			11064	10733.6066	14768.1084
2	AL	TRUE	Daphne-Fairhope-Foley, AL MSA	Baldwin	2p3c			14988	13104.918	15071.3112
2	AL	TRUE	Daphne-Fairhope-Foley, AL MSA	Baldwin	2p4c			14988	16037.2776	15174.9828
3	AL	FALSE	Barbour County, AL	Barbour	1p0c			5856	3237.03924	10890.72876
3	AL	FALSE	Barbour County, AL	Barbour	1p1c			7776	4770.609	11436.48264
3	AL	FALSE	Barbour County, AL	Barbour	1p2c			7776	6989.73852	12036.7944
3	AL	FALSE	Barbour County, AL	Barbour	1p3c			9672	9324.6204	12404.508
3	AL	FALSE	Barbour County, AL	Barbour	1p4c			9672	11413.9656	12530.8248
3	AL	FALSE	Barbour County, AL	Barbour	2p0c			5904	5934.57168	12711.6432
3	AL	FALSE	Barbour County, AL	Barbour	2p1c			7776	7386.17208	13271.94
3	AL	FALSE	Barbour County, AL	Barbour	2p2c			7776	9354.42768	13611.864
3	AL	FALSE	Barbour County, AL	Barbour	2p3c			9672	11421.0462	13720.7796
3	AL	FALSE	Barbour County, AL	Barbour	2p4c			9672	13976.622	13639.788
4	AL	TRUE	Birmingham-Hoover, AL HUD Metro FMR Are	Bibb	1p0c			7973.92164	3050.28684	11446.48464
4	AL	TRUE	Birmingham-Hoover, AL HUD Metro FMR Are	Bibb	1p1c			9779.52168	4495.38168	12233.8776
4	AL	TRUE	Birmingham-Hoover, AL HUD Metro FMR Are	Bibb	1p2c			9779.52168	6586.48536	13566.48
4	AL	TRUE	Birmingham-Hoover, AL HUD Metro FMR Are	Bibb	1p3c			12717.2832	8786.66232	14094.588
4	AL	TRUE	Birmingham-Hoover, AL HUD Metro FMR Are	Bibb	1p4c			12717.2832	10755.4687	14358.126

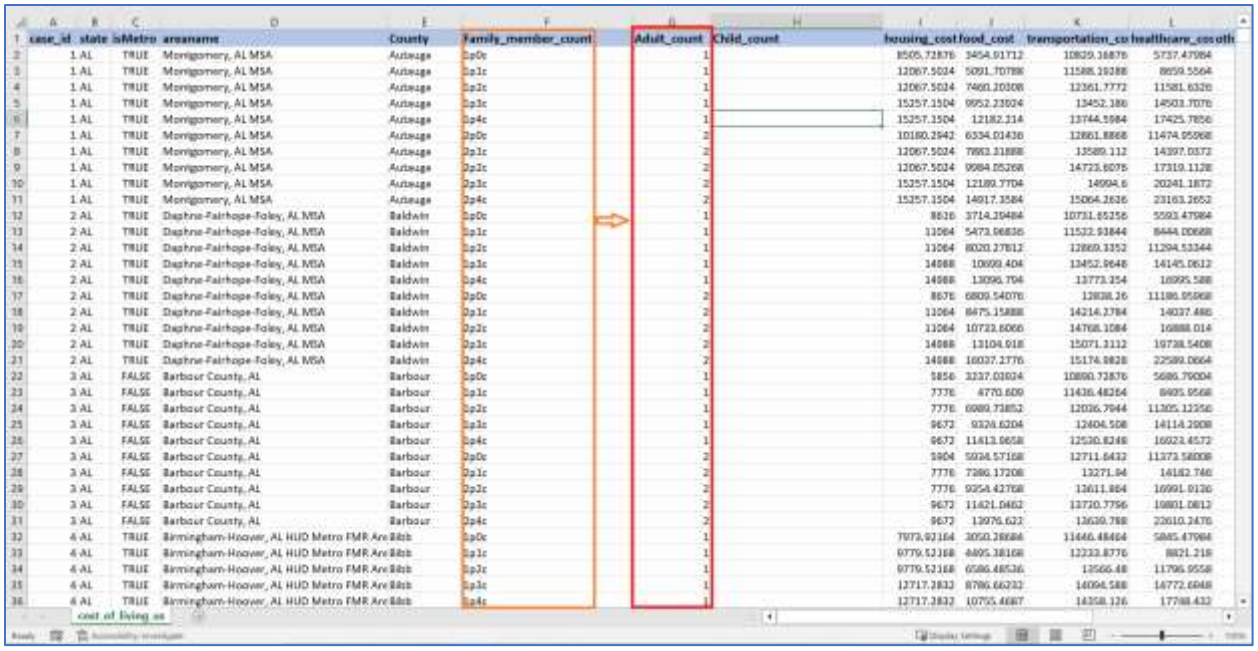
- FIND("p", F2):This part finds where the letter "p" is in the text of cell F2.
- MID(F2, 1, FIND("p", F2) - 1):** This extracts the part of the text in F2 before the "p."
- VALUE(...): Tries to turn the extracted text into a number. If the text has numbers, it turns them into a numeric format.

- IFERROR(..., ""):Checks for errors. If there is an error (like if there are no numbers before the "p"), it shows an empty space instead of an error.

So, basically, this whole thing takes the numbers from the beginning of the text in F2 until it hits the "p." If there's an issue (no numbers before "p"), it just gives an empty space.

After:

Now, the "Adult_count" column contains the numeric value before the letter "p" from the original text.

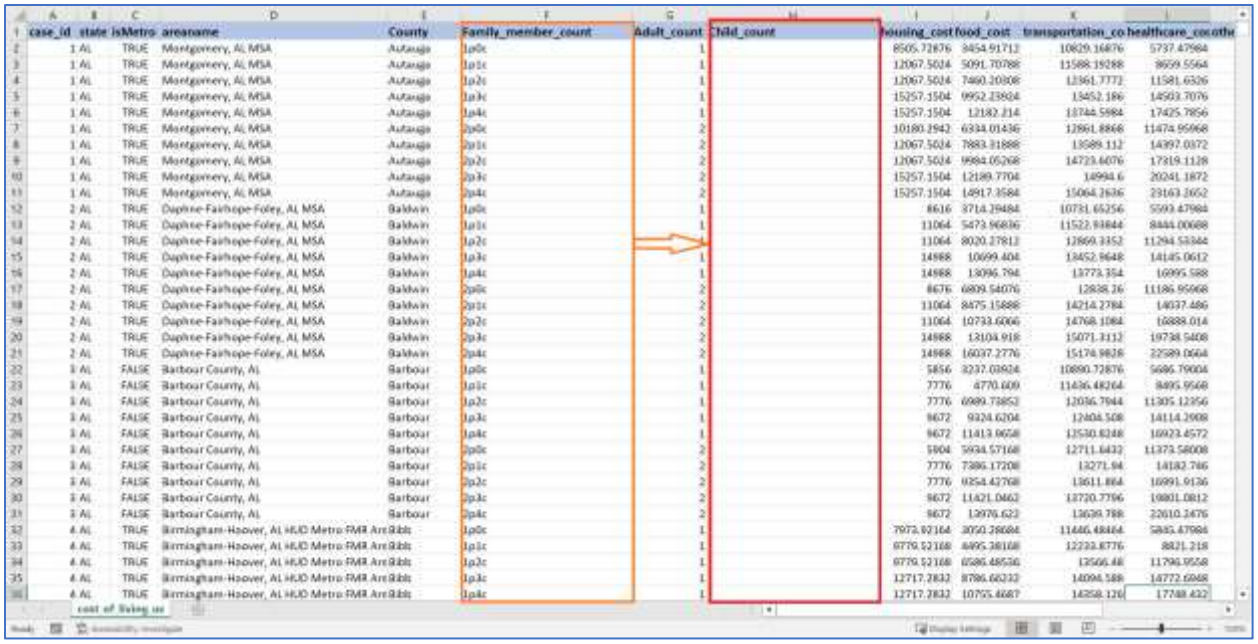


- Number of Childs(Child_count):

  Before Column Name: Family Member Count

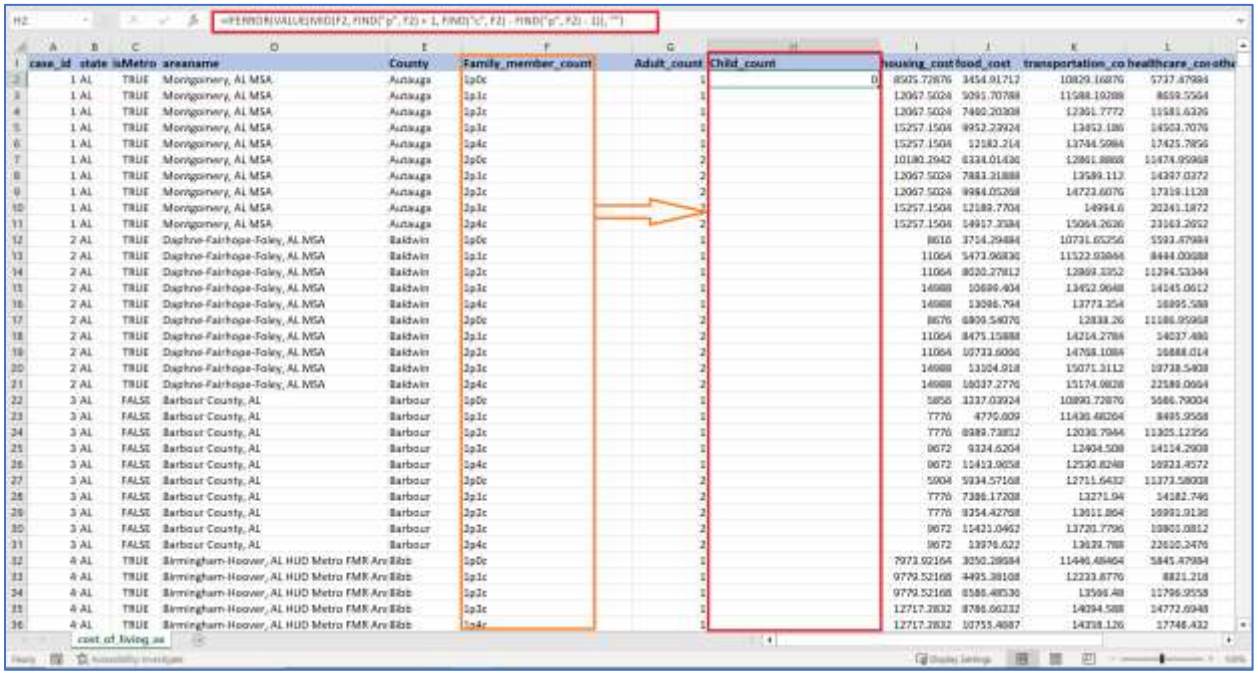
  After Cleaning Column Name: Child_count

Before:



After:

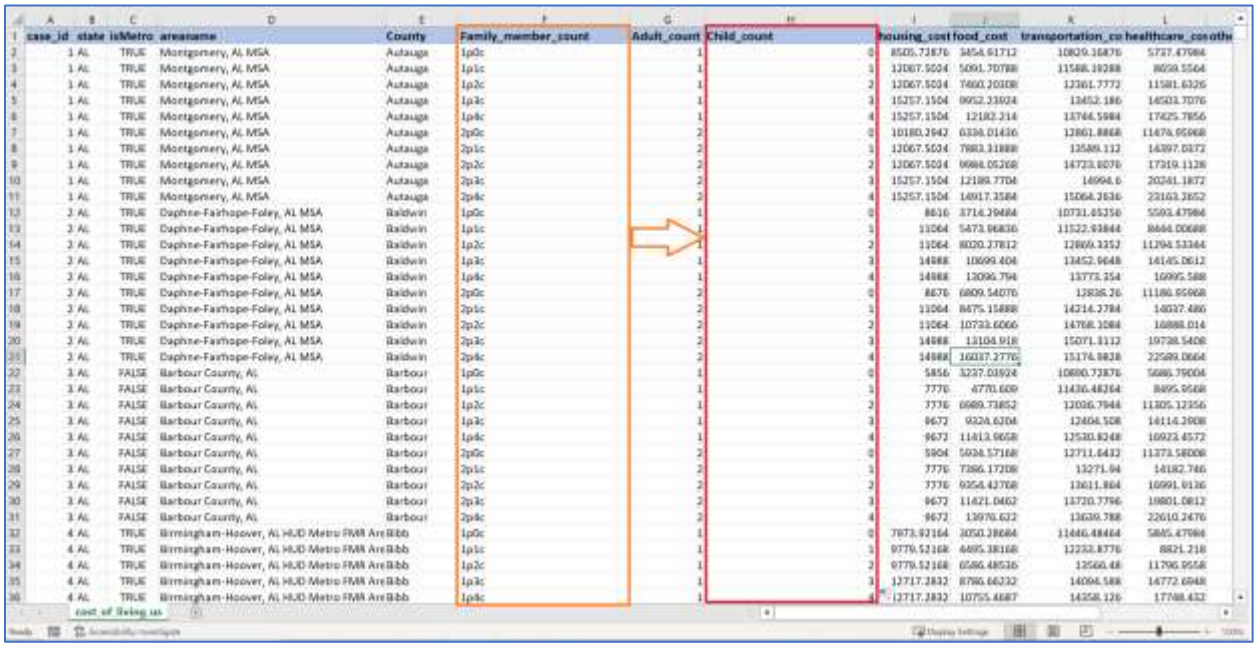
The number of children has been separated from the "Family_member_count" column and placed in a new column called "Child_count."



- FIND("p", F2): Finds where the letter "p" is in the text of cell F2.
- FIND("c", F2): Finds where the letter "c" is in the text of cell F2.
- FIND("p", F2) + 1: Adjusts the starting position for the MID function to begin one character after "p.

- FIND("c", F2) - FIND("p", F2) - 1: Calculates the length of the substring to be extracted by measuring the distance between "p" and "c" and subtracting 1 to exclude "p."
- MID(F2, FIND("p", F2) + 1, FIND("c", F2) - FIND("p", F2) - 1): Extracts the characters between "p" and "c" in the text in cell F2.
- VALUE(...): Tries to turn the extracted text into a number. If there are numbers, it converts them into a numeric format.
- IFERROR(..., ""): Checks for errors. If there's an error (like if there are no numbers between "p" and "c"), it shows an empty space instead of an error.

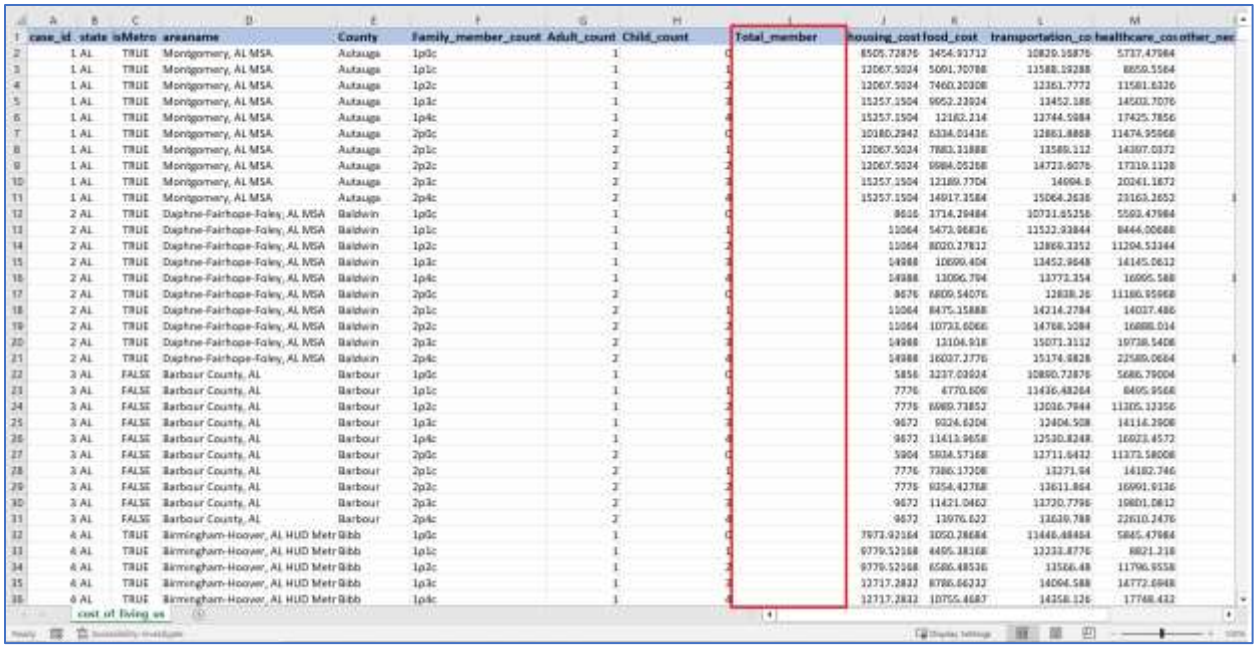
This formula gets the numbers between "p" and "c" in the text of cell F2. If there's an issue (no numbers between "p" and "c"), it just gives an empty space.

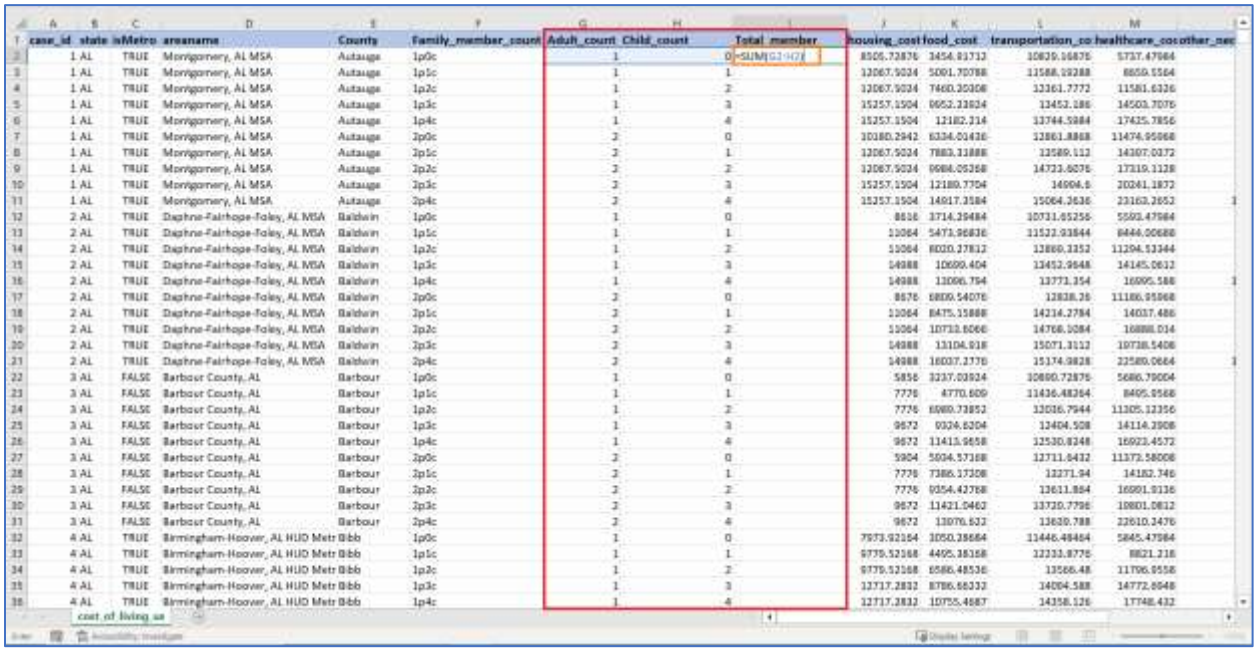


3.3. Creating Additional Columns Based on Existing Data

Column Name: Total_member

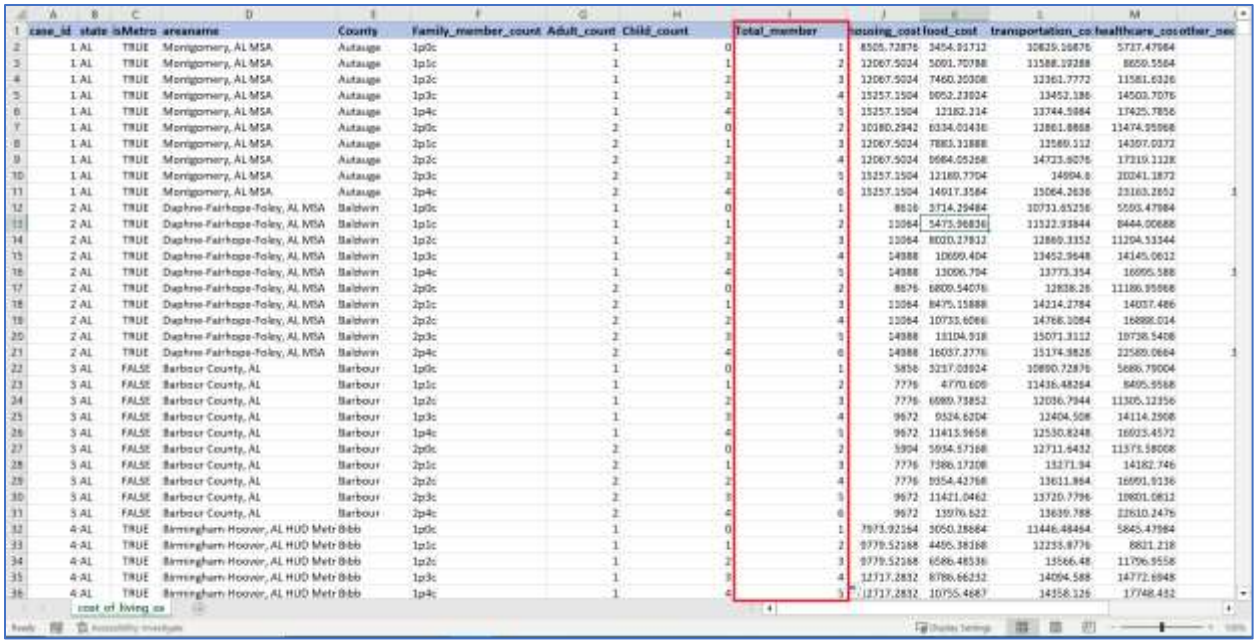
Before:



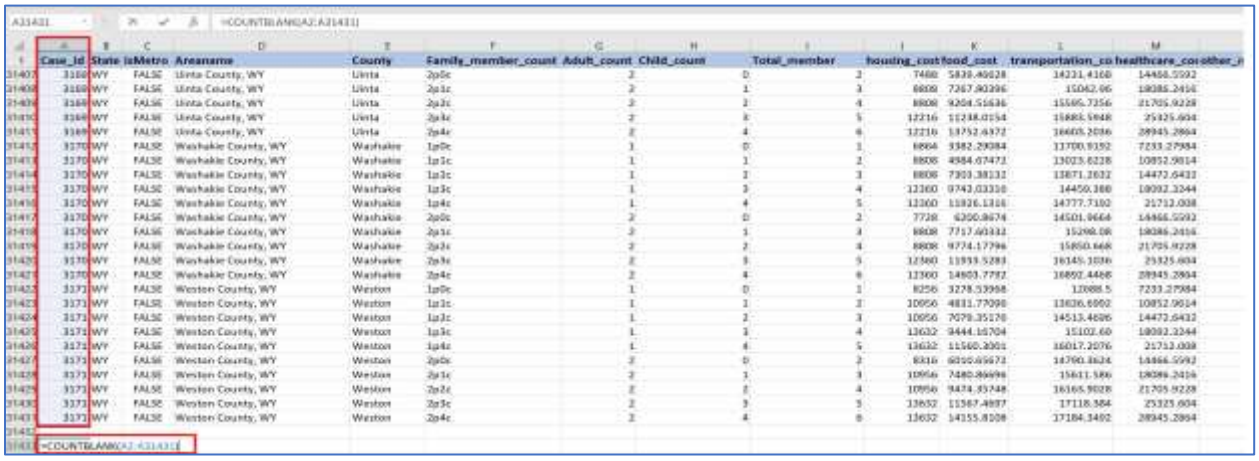


Creating a new column titled "Total_member" by summing the counts from "Adult_count" and "Child_count" to represent the total number of members in a family.

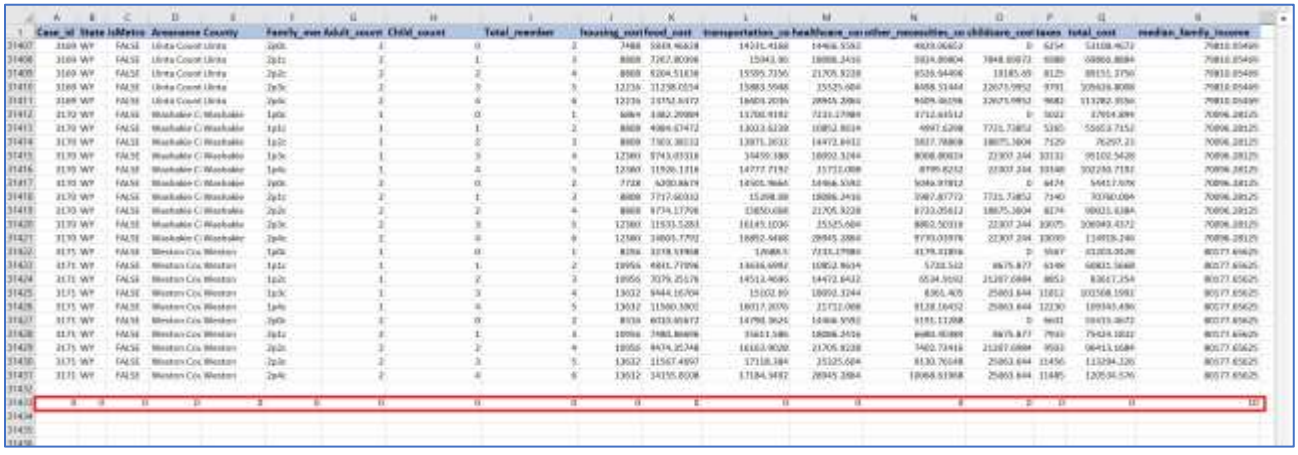
- Using the SUM() function to calculate the total number of family members by adding the counts from both "Adult_count" and "Child_count."



3.4. Check for a Blank Values:



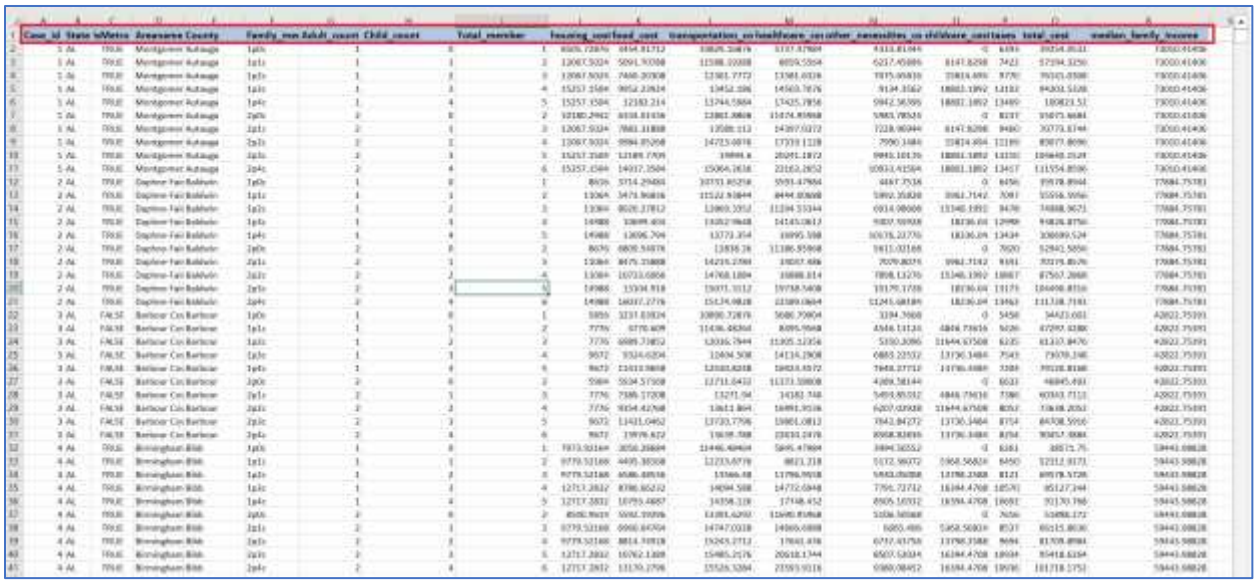
After:



3.5. Renaming the column header:

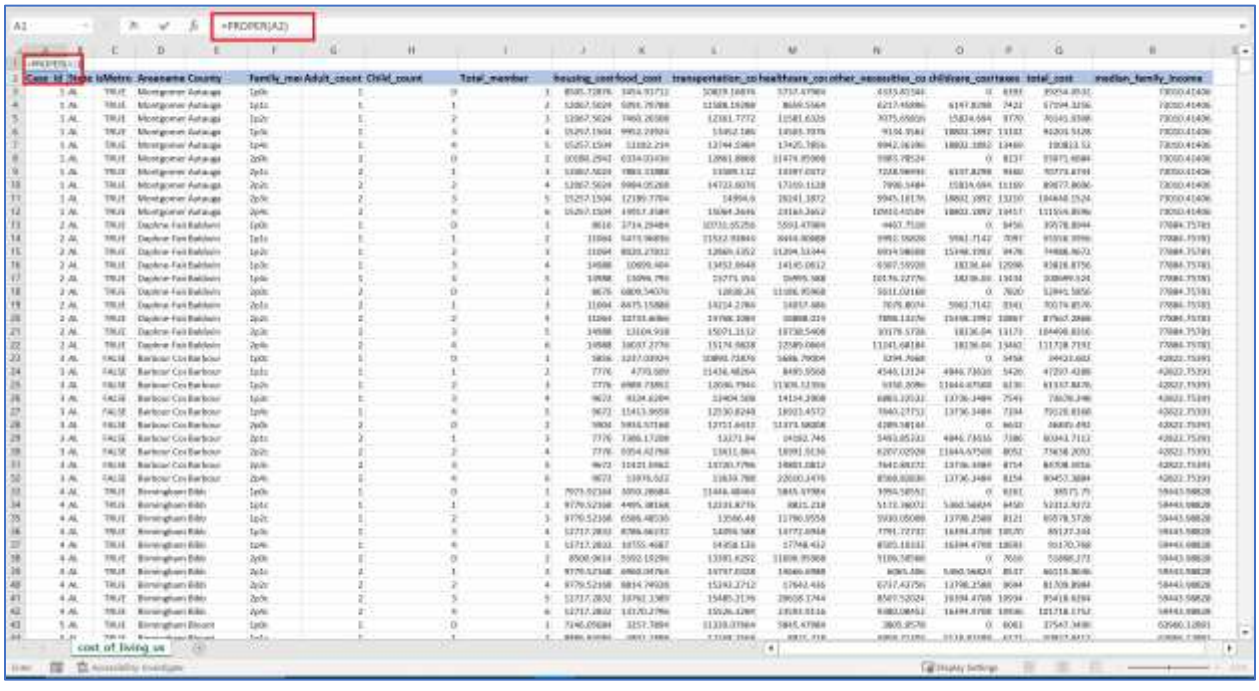
During the data cleaning process, inconsistencies in column names were identified, with some names appearing in uppercase and others in lowercase. To standardize naming conventions and facilitate data manipulation and error prevention, all column names were renamed.

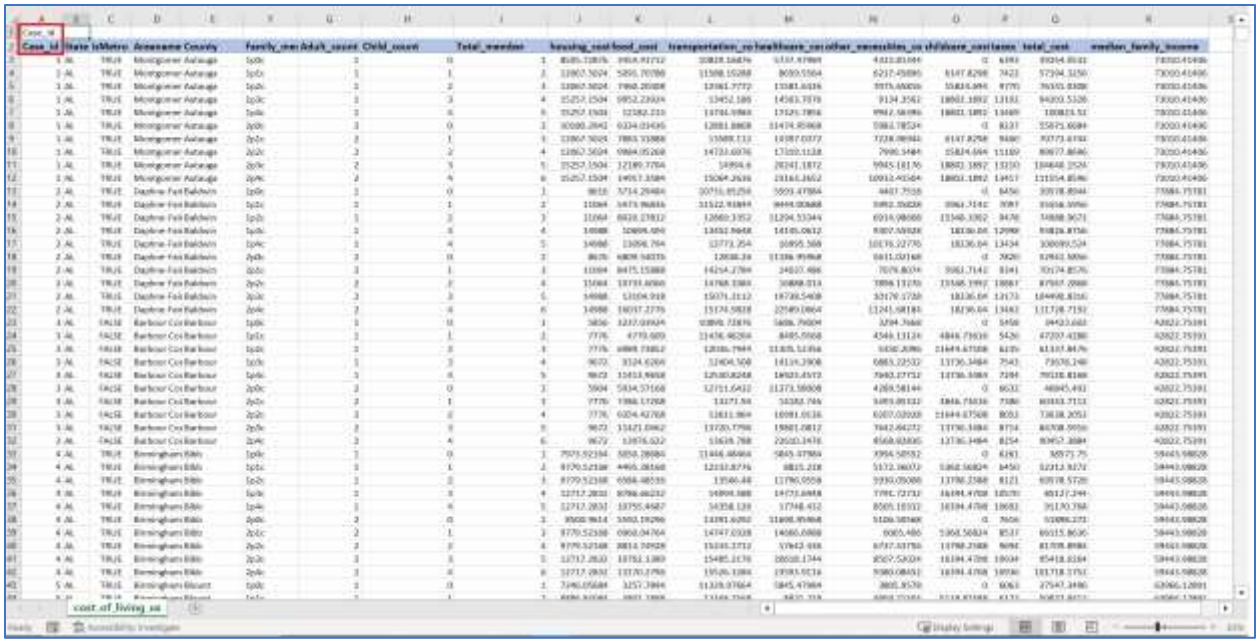
Before:



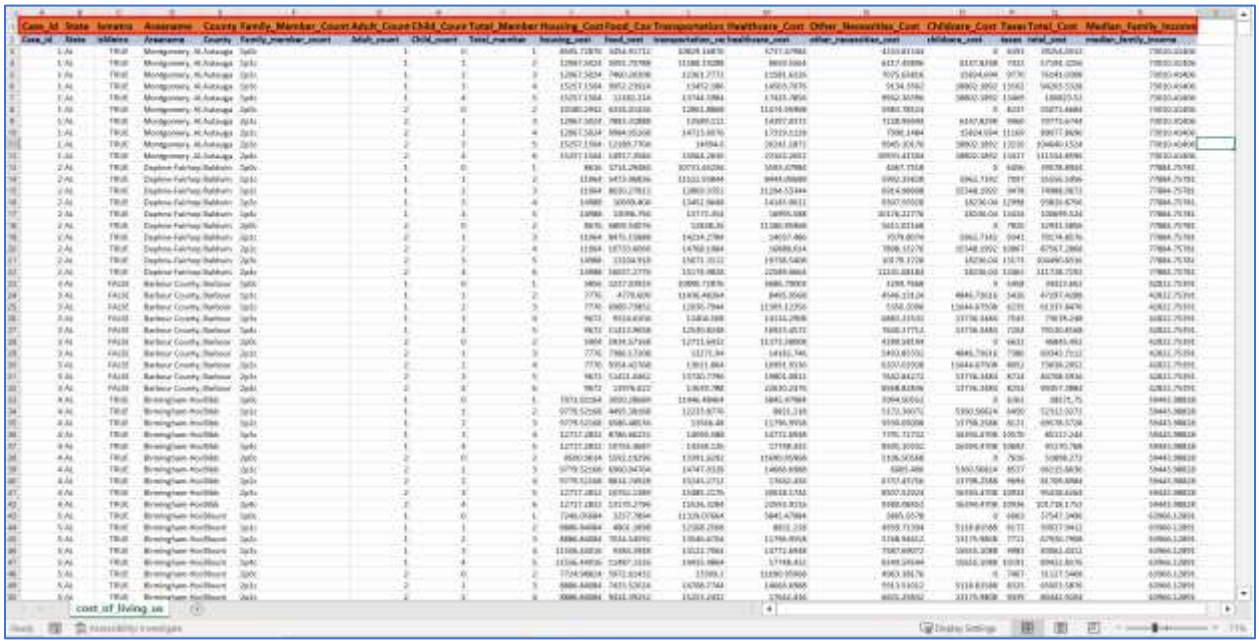
After:

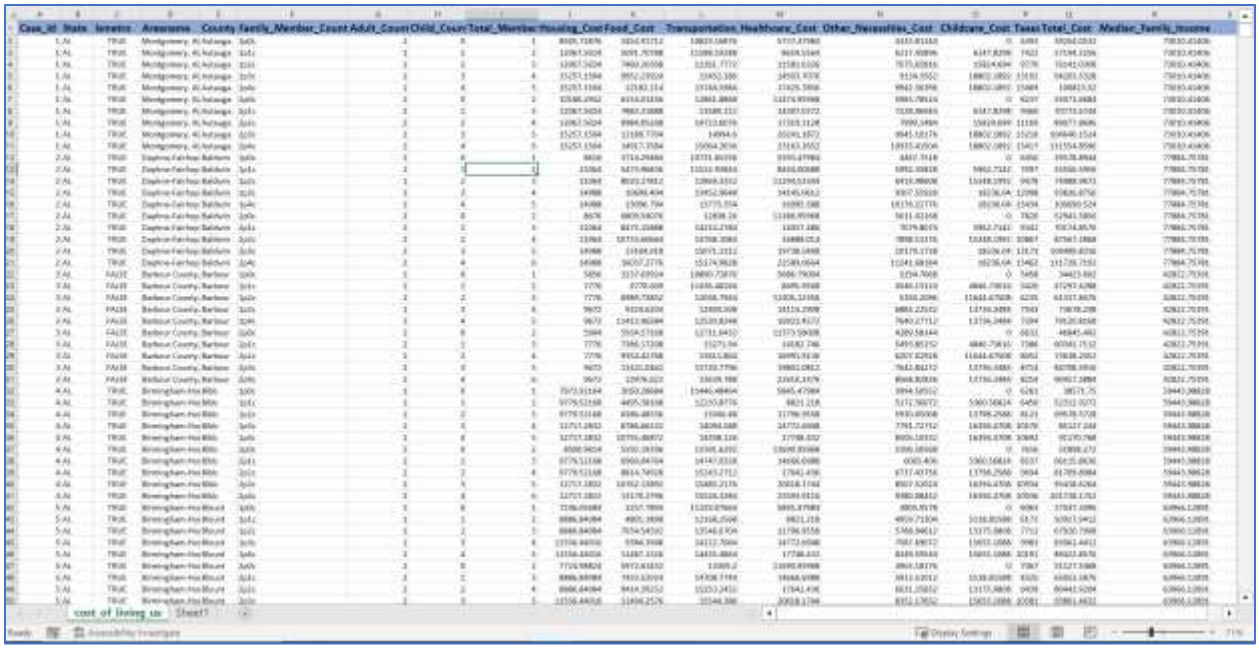
The PROPER() function in Excel capitalizes the first letter of each word in a text string. Here the header name in column A, “Case_id”, has been transformed to “Case_Id” using the PROPER() function.





Performed the same step for rest of the columns.

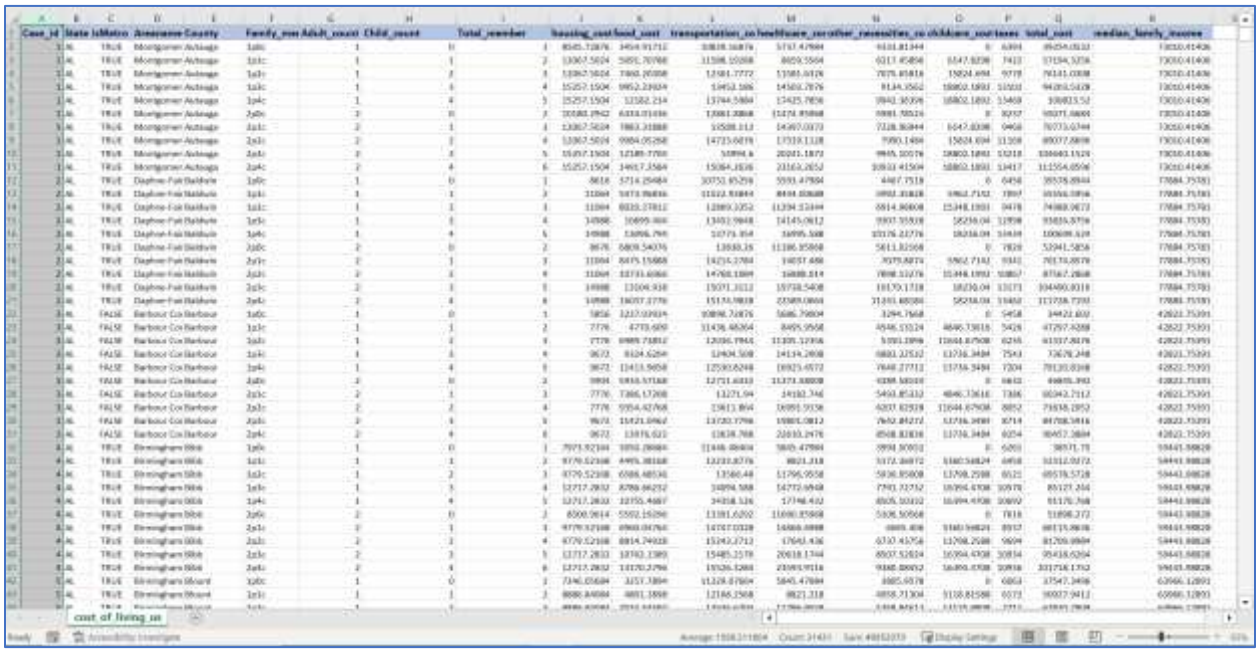




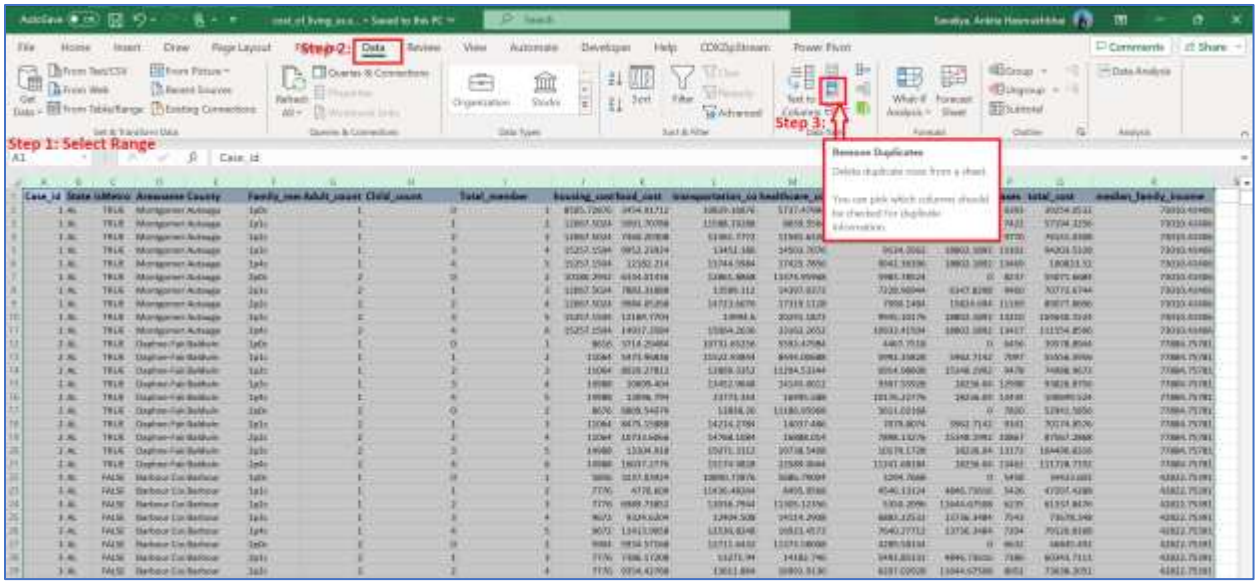
3.6. Removing Duplicate values:

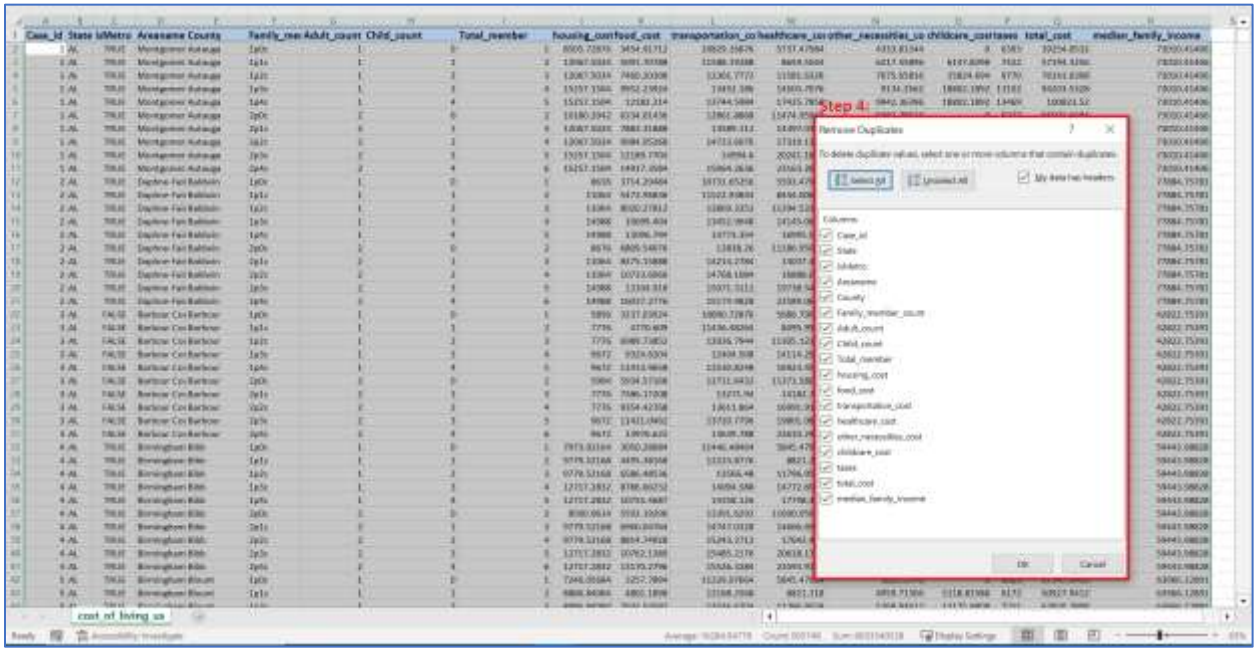
In our data cleaning process, the initial step involved removing duplicate rows. It is crucial to remove duplicate values for the true analysis. And another reason was that this action ensures data consistency and enables us to obtain precise insights, which in turn supports informed decision-making based on the data.

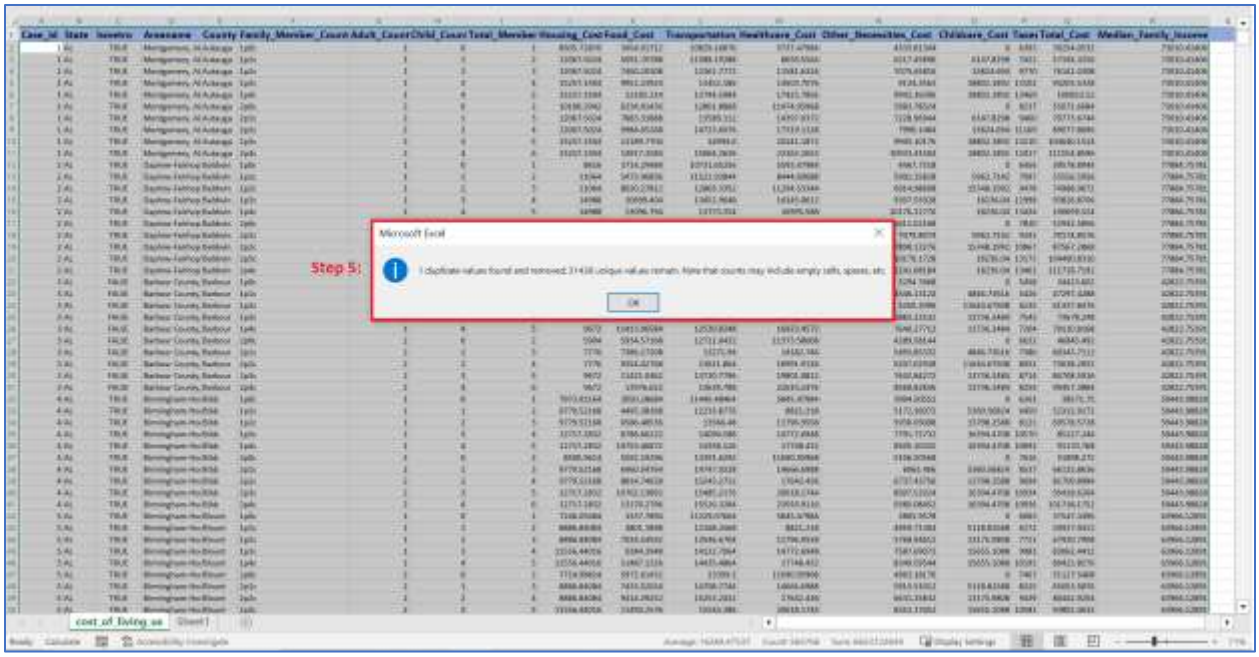
Before:



To check how many duplicates rows in our dataset, we performed the following steps.

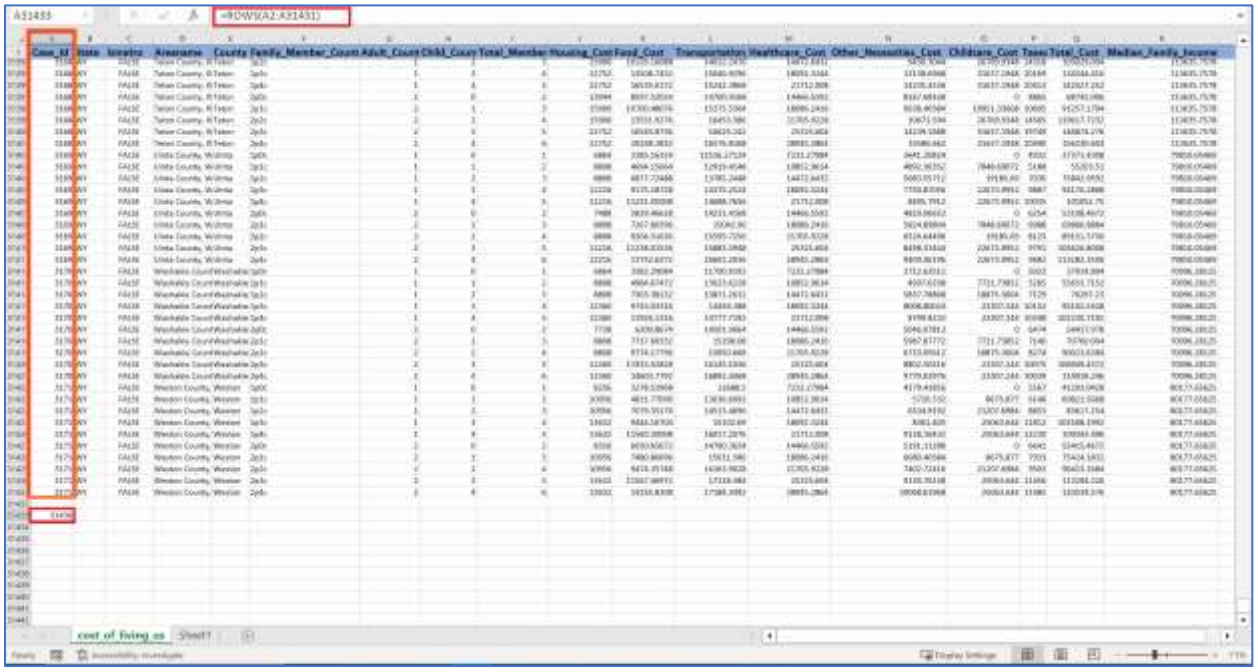

Step 1: Select Range
Step 2:
Step 3:
Remove Duplicates


Step 4:
Remove Duplicates



After:

There was only one duplicate row in the dataset. After removing the duplicate entry, the 'rows()' function was utilized to determine the number of remaining rows in the dataset.

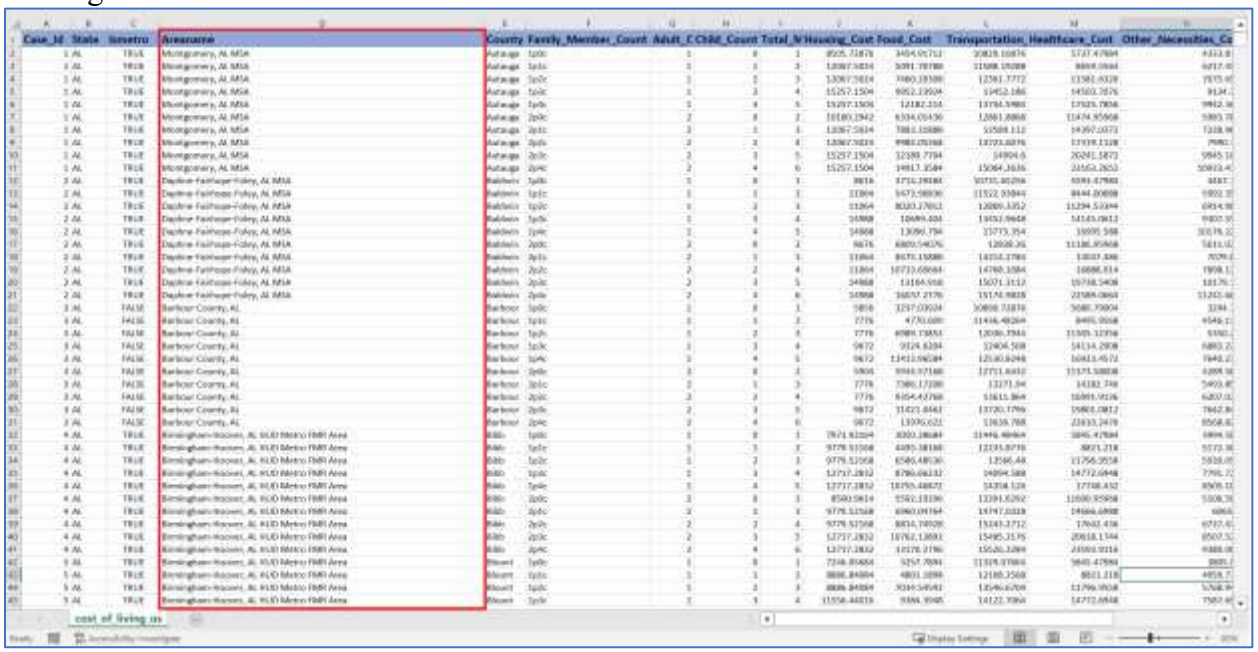


3.7. Extracting Area Name

Before:

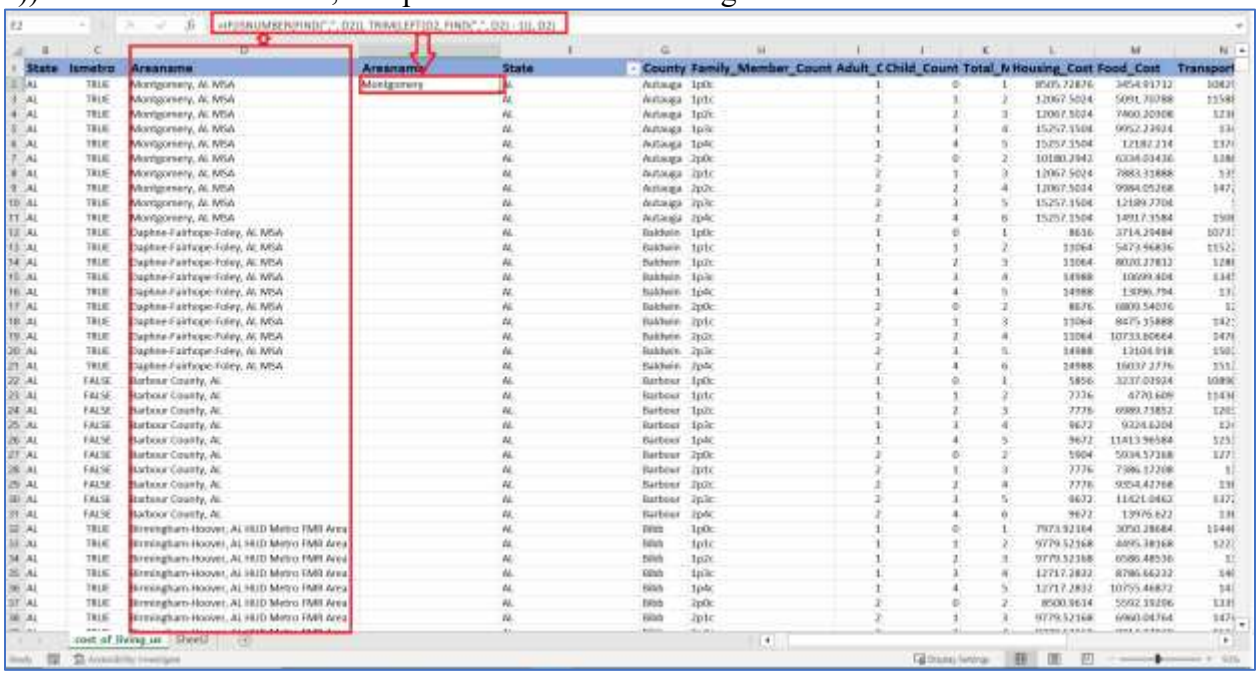
Column Name: Areaname

We focused on extracting relevant information for our analysis, specifically the area from the existing values.

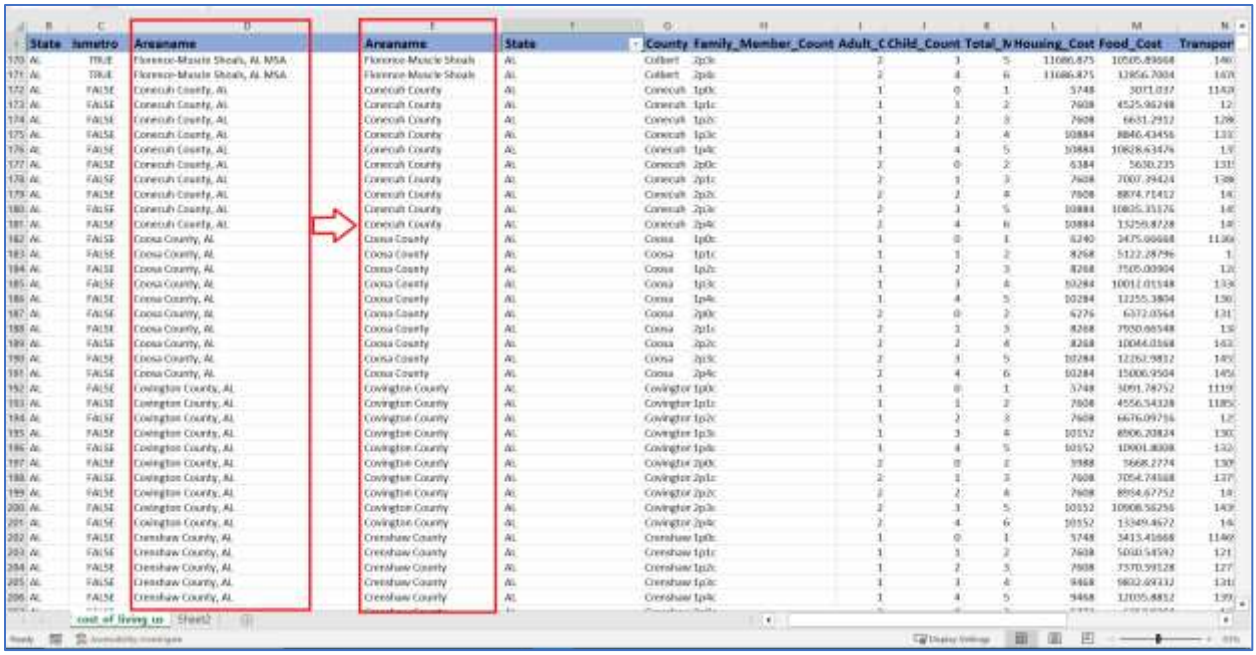


After:

This formula checks if there is a comma in the string using ISNUMBER(FIND(",", D2)). If a comma is found, it extracts the value before the comma using TRIM(LEFT(D2, FIND(",", D2) - 1)). If no comma is found, it copies the entire value using D2.



Simply drag it down for the entire column. It will extract the value before the comma or copy the entire value if there is no comma.

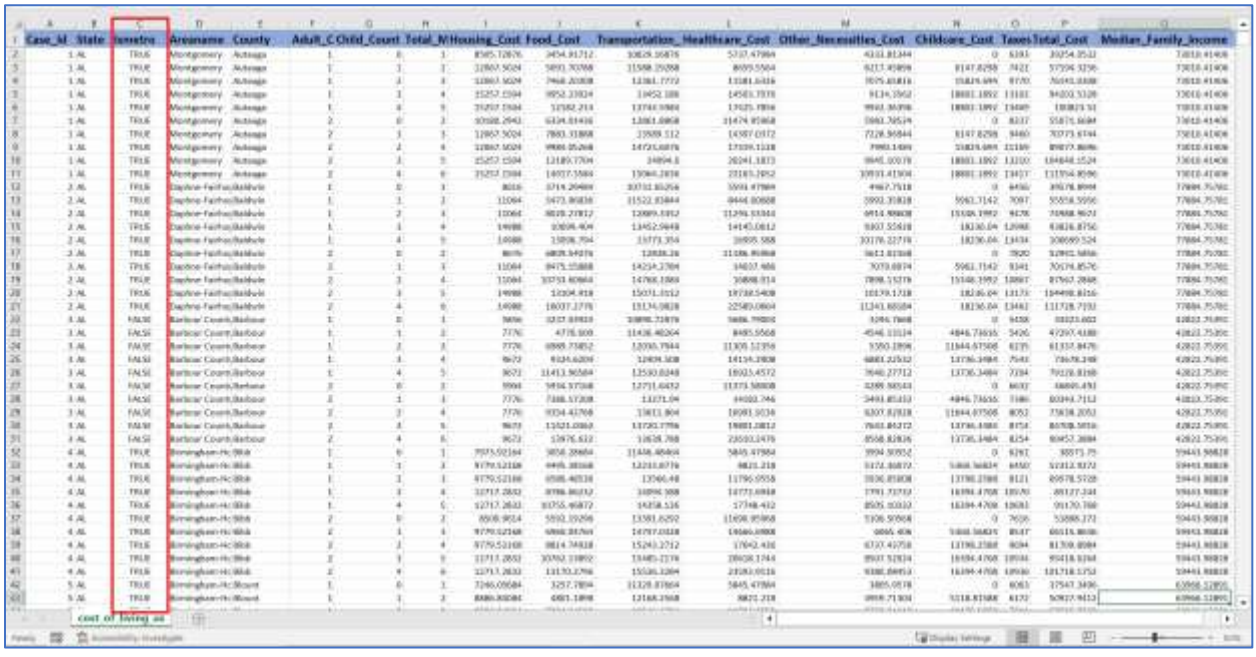


3.8. Removing extra spaces

Column Name: Ismetro

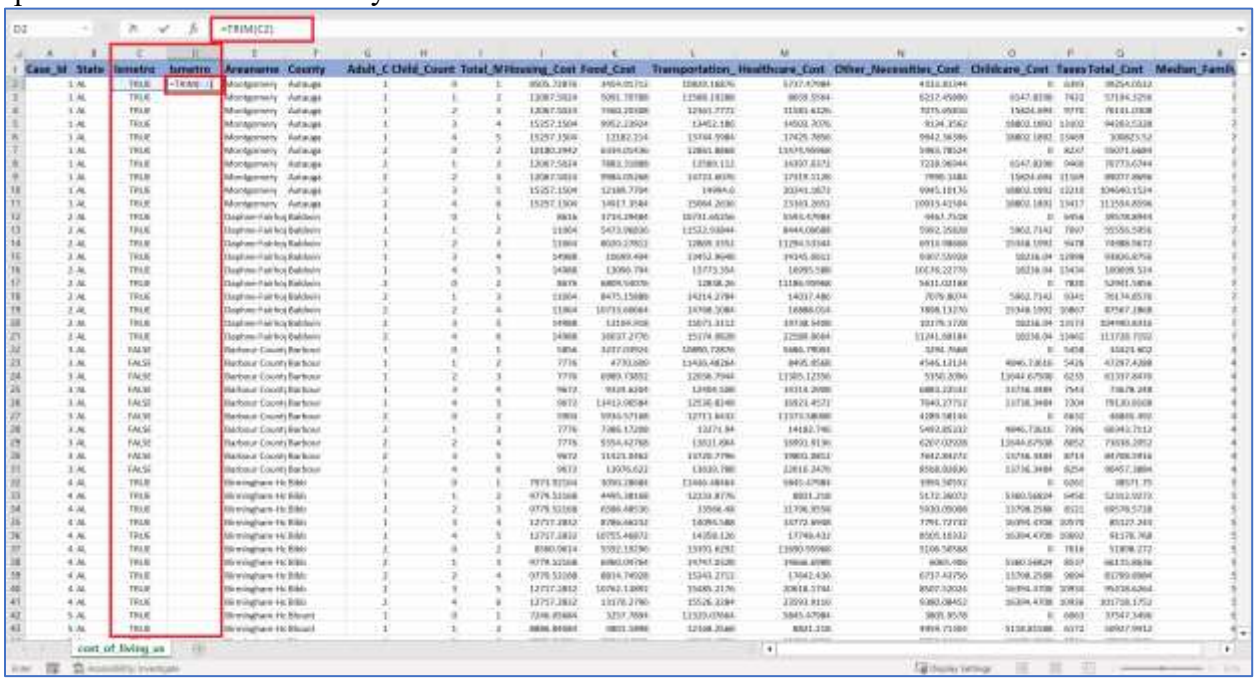
Before:

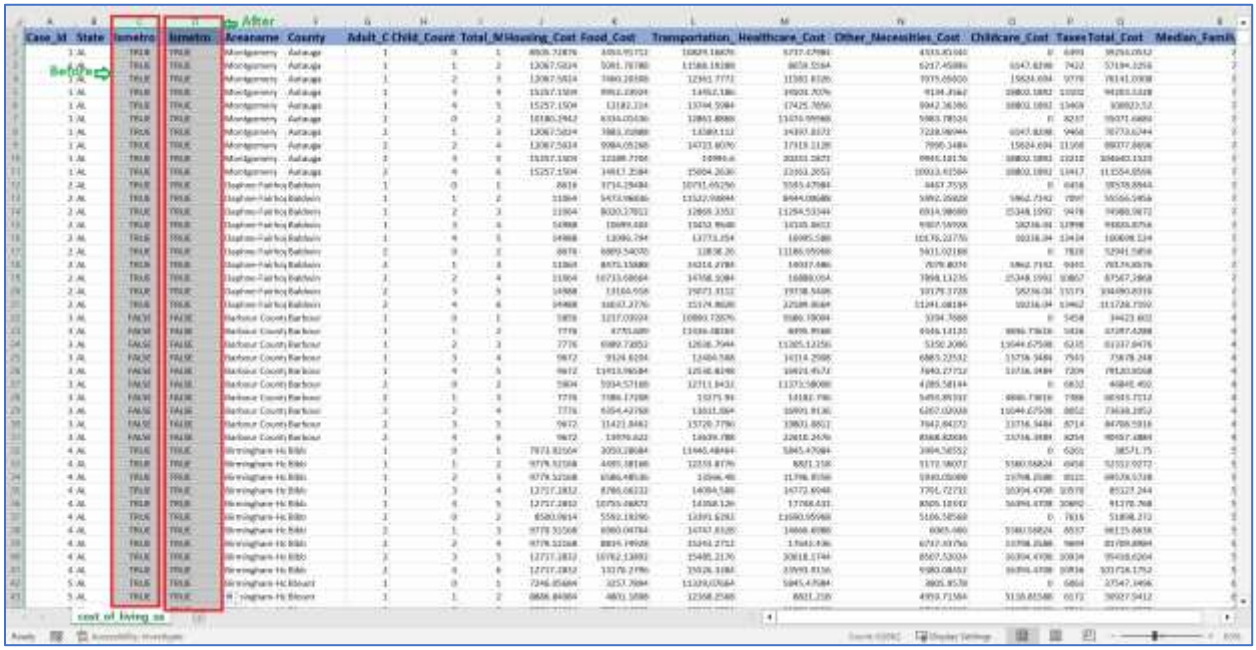
The "Ismetro" column initially contained leading spaces, and those spaces have been removed for clarity.



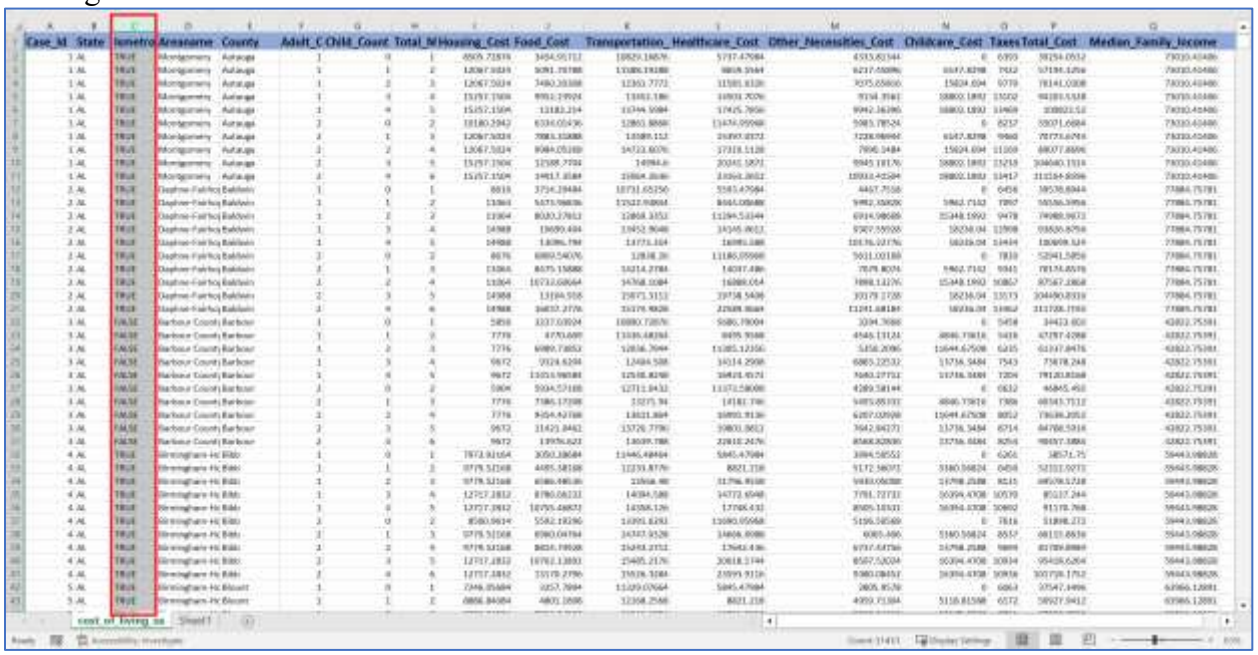
After:

The TRIM() function effectively eliminates unnecessary spaces within the values. Unnecessary spaces have been successfully removed.





Remove the redundant "Ismetro" column (Column C). Replace the final result in Column D with the original values from Column C.

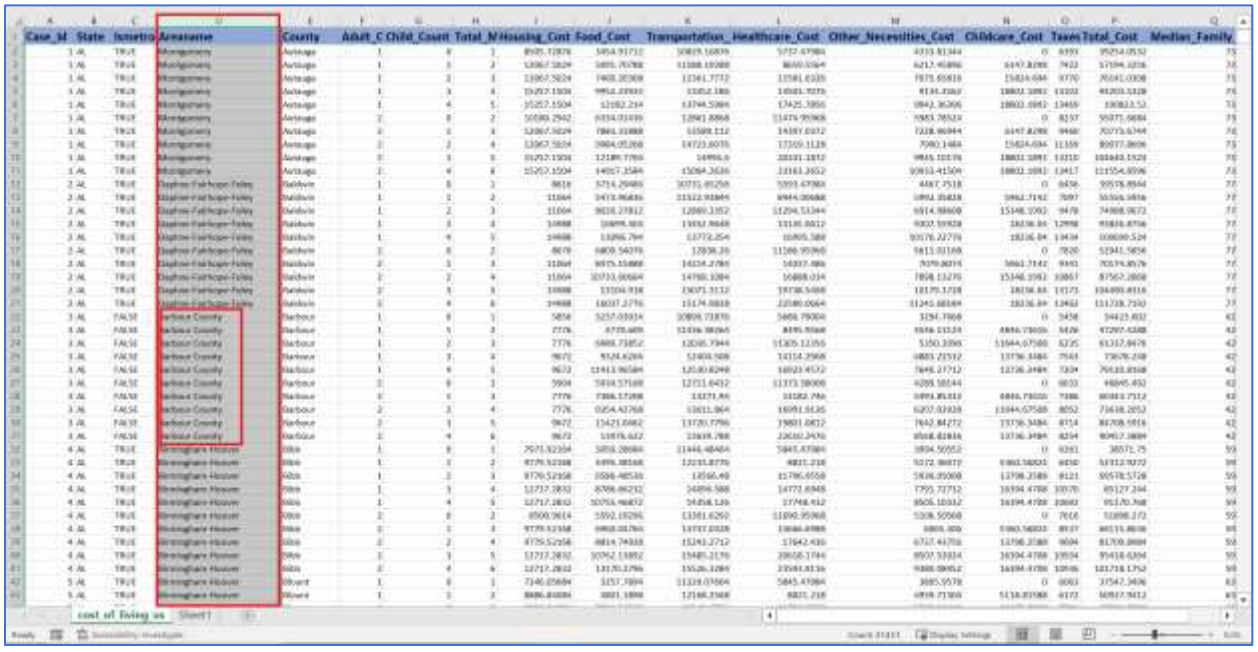


1.9. Removing repeated word

Column Name: Areaname

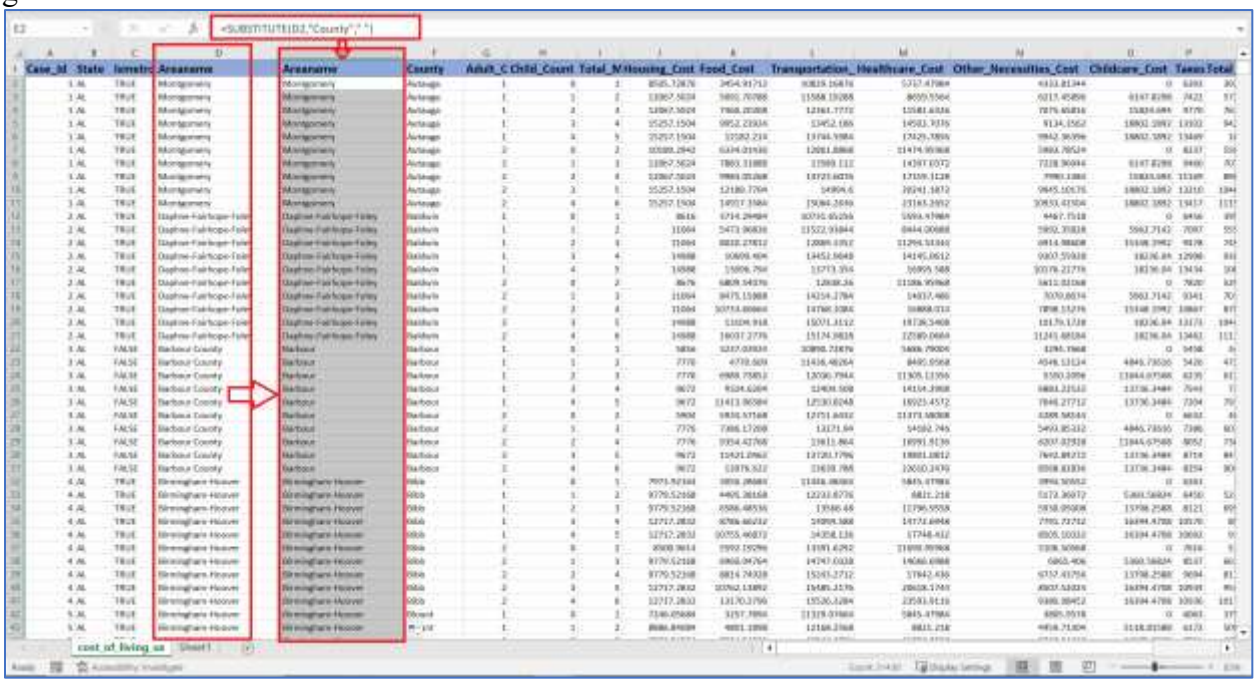
Before:

Removing the word "County" from the 'Areaname' can enhance the overall clarity and simplicity of the name.

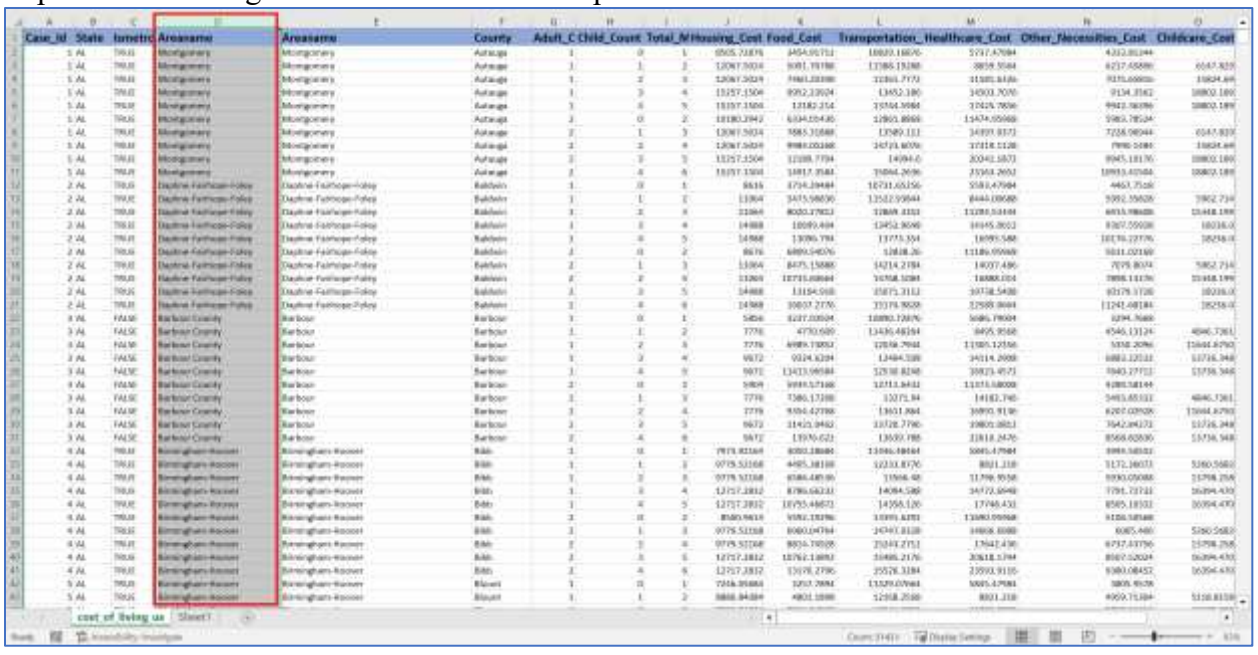


After:

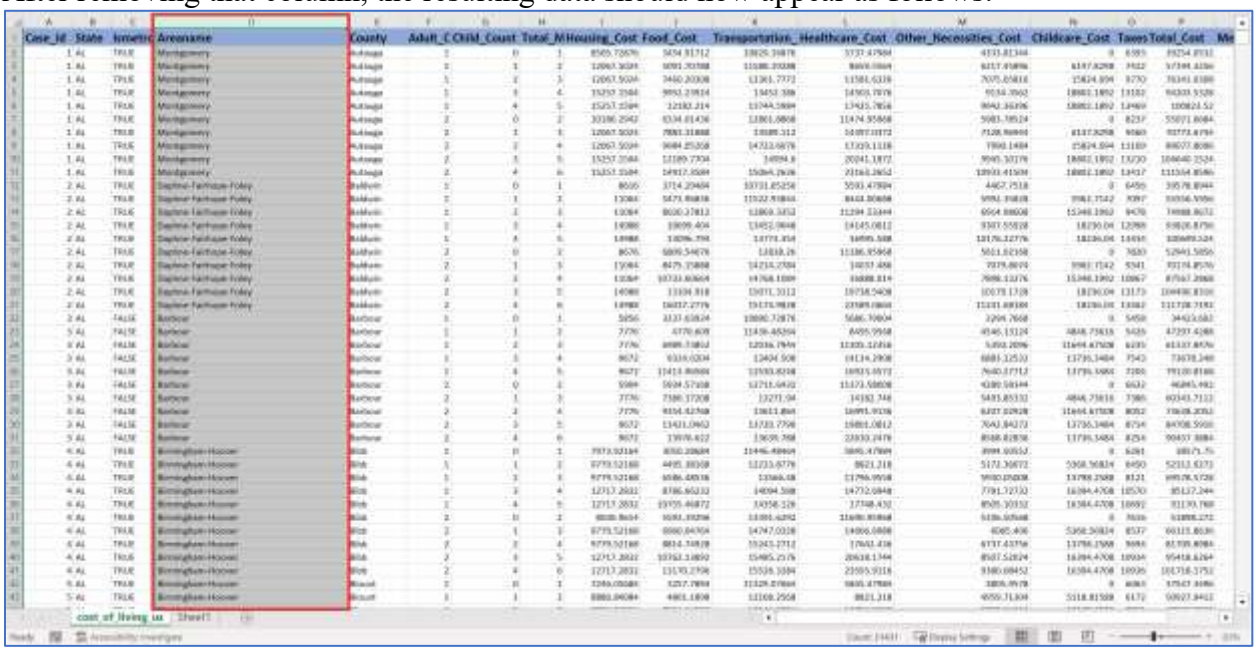
The `SUBSTITUTE()` function is used to substitute the word "County" with a blank space in the given text.



Replace the existing data column D with the updated values.



After removing that column, the resulting data should now appear as follows:



Final Outlook of Dataset:

Our final dataset includes information such as Case_Id, State, Areaname, County,Family_Member_Count, Adult_Count, Child_Count, Total_Member_Count, Housing_Cost, Food_Cost, Transportation_Cost, Healthcare_Cost, Other_Necessities_Cost, Childcare_Cost, Taxes, Total_Cost, and Median_Family_Income.