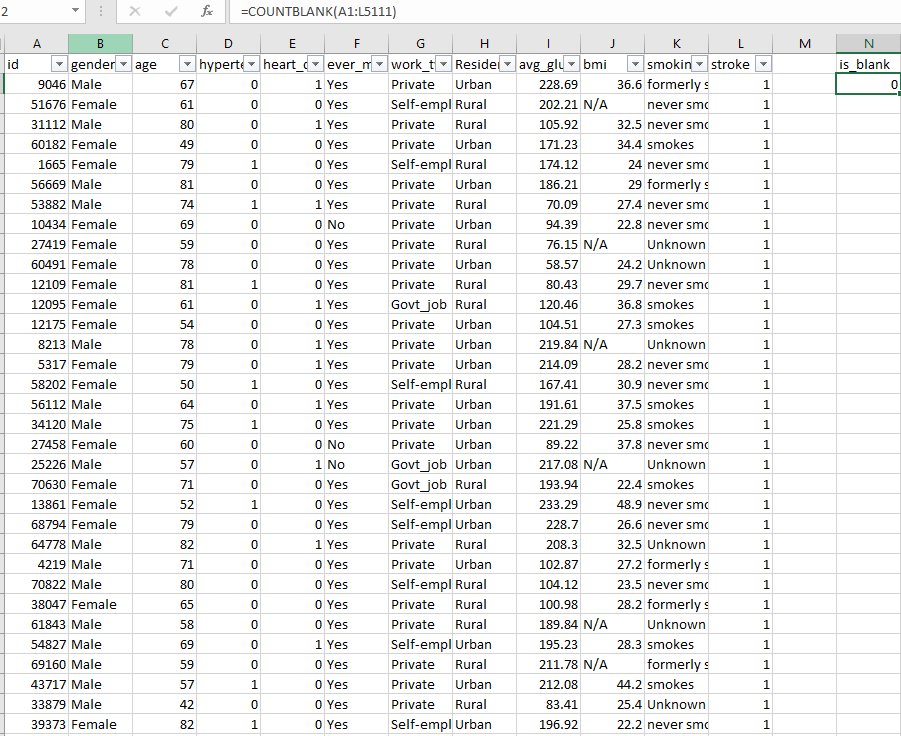
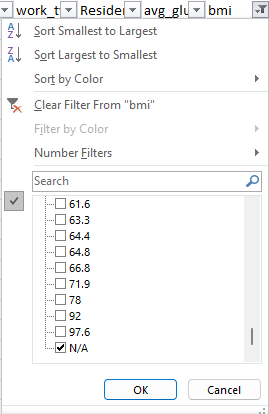
**Handling Missing Values**

For this assignment, I have selected heatstroke dataset

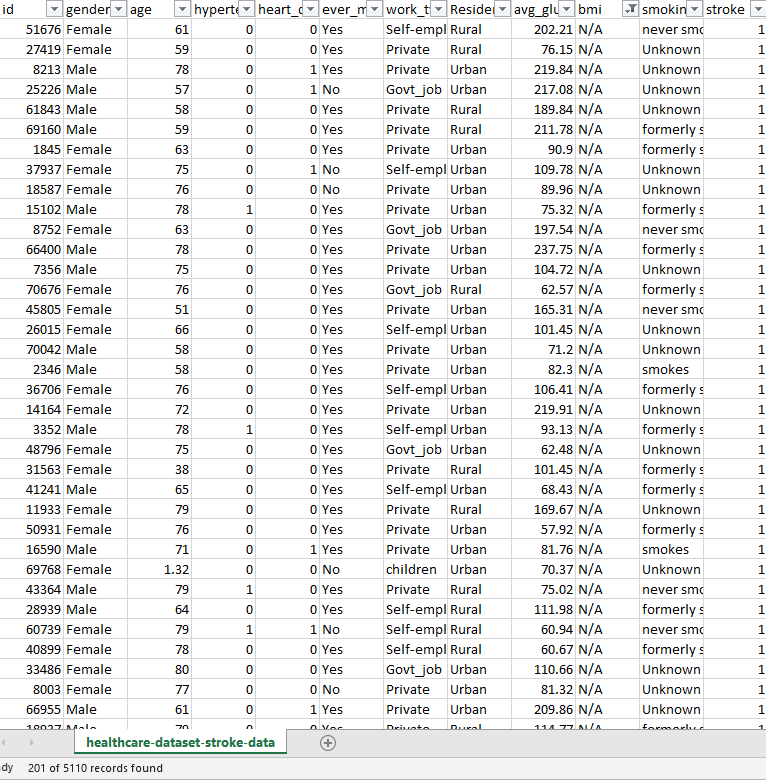
In the dataset, there is no blank values or empty cell found initially. To count the missing values countblank is used



However, there are N/A values in BMI index column. To find them applied a filter on the column and select all the entries with NA Values as seen in screenshot



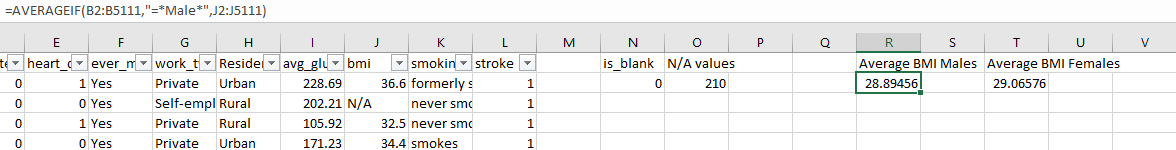
Filtering out the bmi column where the value is N/A, 201 records were found.



To fill these missing values, I have calculated average BMI of both genders male and females by using average if formula.

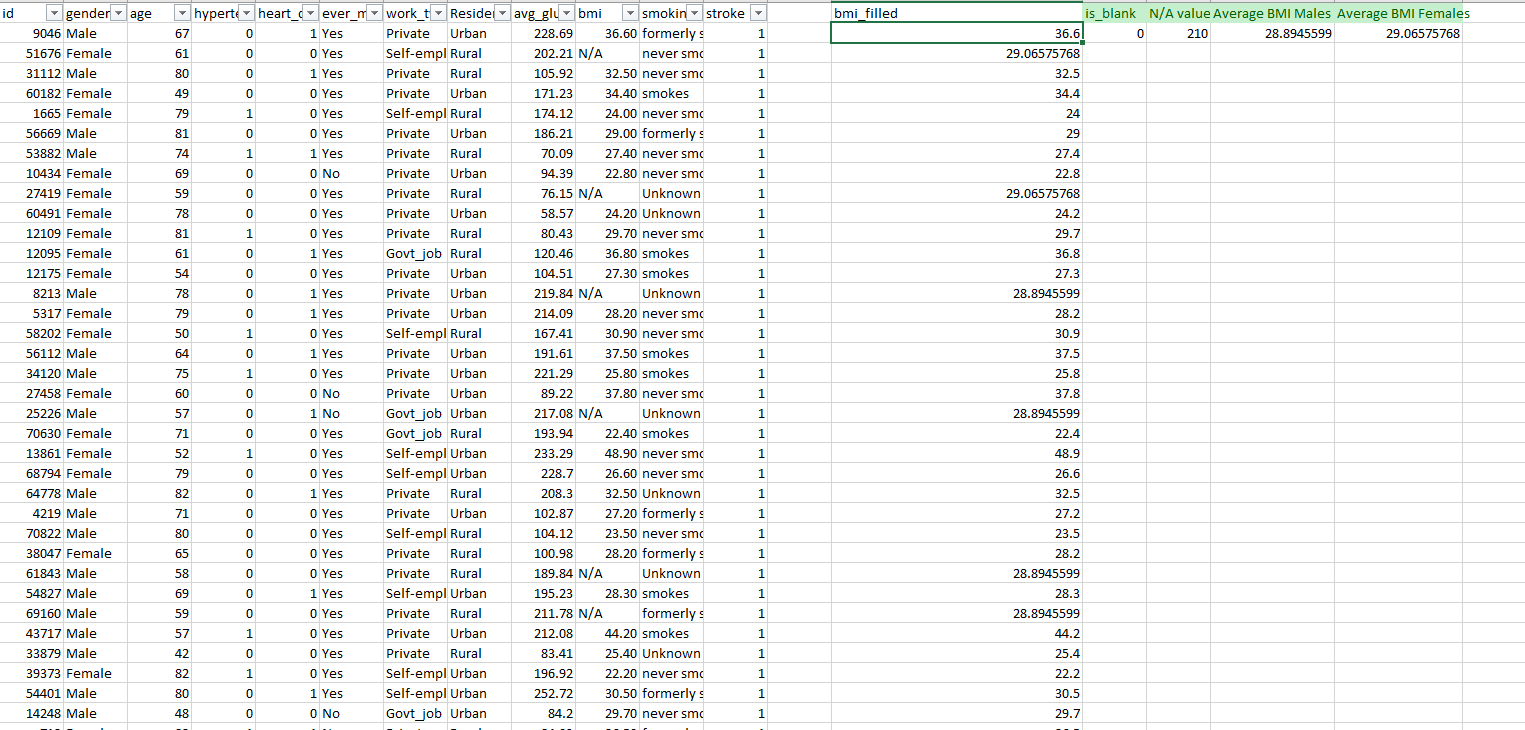


Applying this on dataset, we have found following average values for both males and females



For handling this missing data, I have filled N/A values with average of their gender. If original value is N/A and gender is Male, fill it with its mean value.



After applying this if else statement, all of the N/A values are filled with their respective mean.  


**Outliers Detection**

As a part of this dataset, outliers can be found in numerical columns like age, BMI and glucose level.

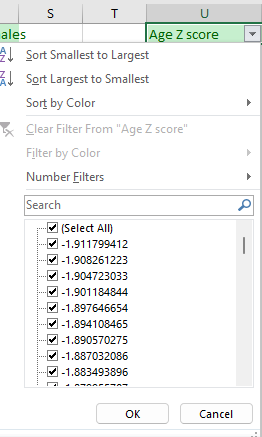
To detect outliers, I computed the z score of that column which is calculated using the formula:

In case z-score lies above than 3 or less than -3, the particular record is considered as outlier. I computed z-score of columns age, BMI and glucose level and found all the values in expected ranges so no outliers found.   
For calculating z-score, I use the following formula in excel:

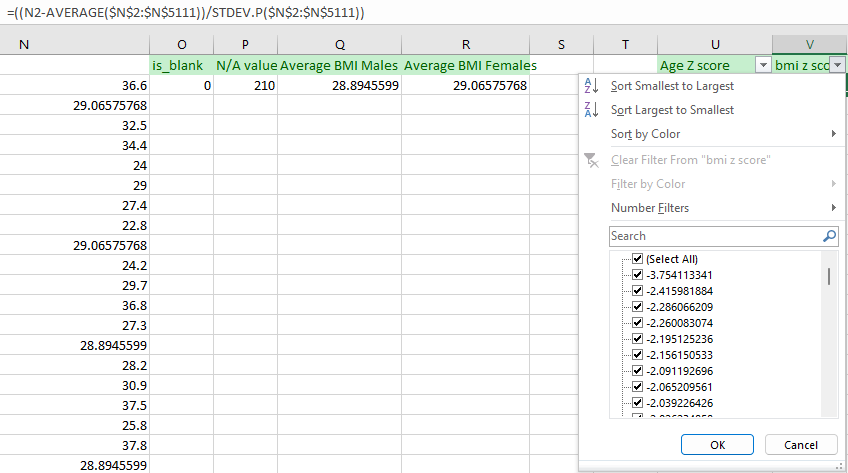


where c is the age column

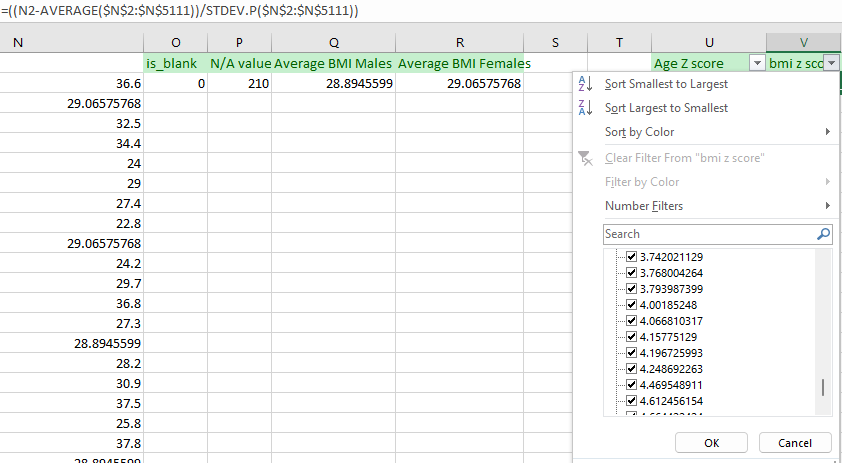
The by filtering z-score values, all values lies between -1.9 to 1.7, no outliers detected.



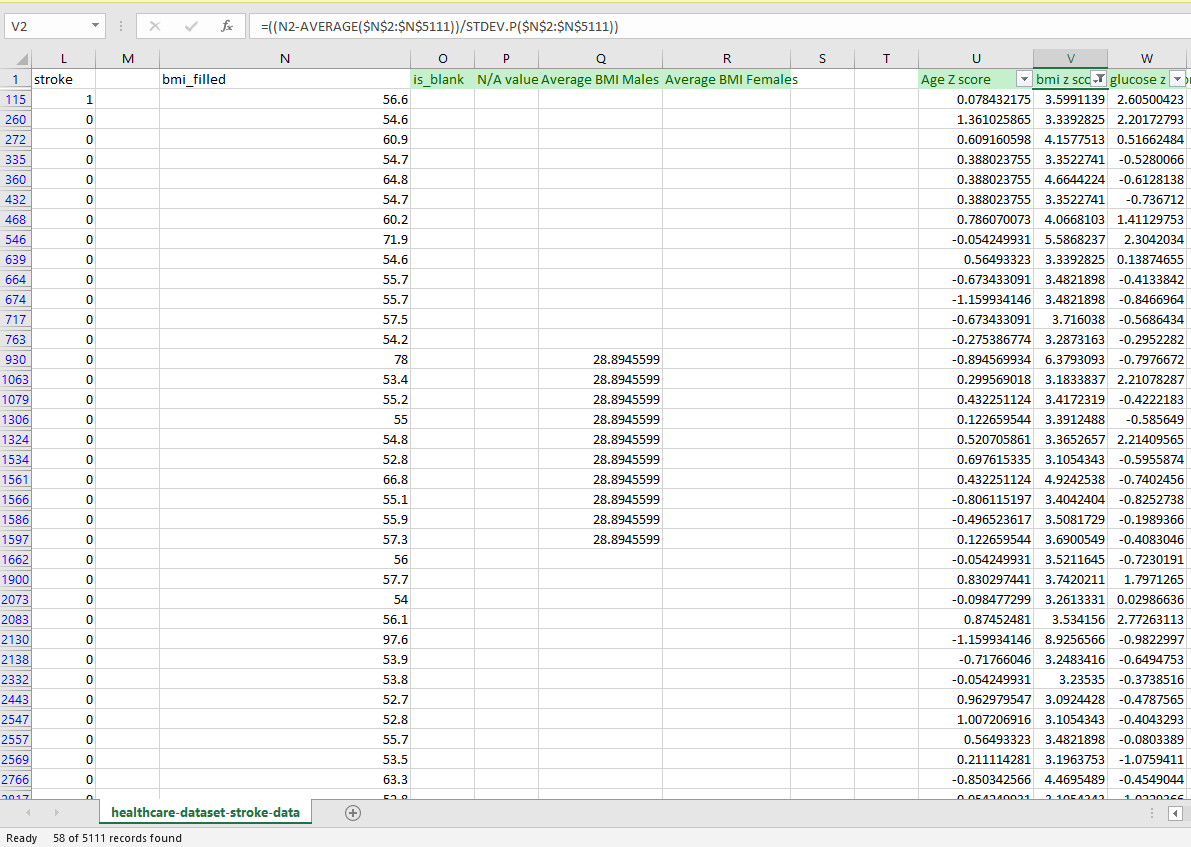
For BMI values:



In case of BMI values, there are several records that are outliers and ranging outside the normal z score values so filtering them out.



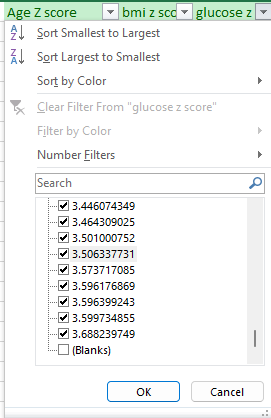
58 records were found outside the range of -3 to 3

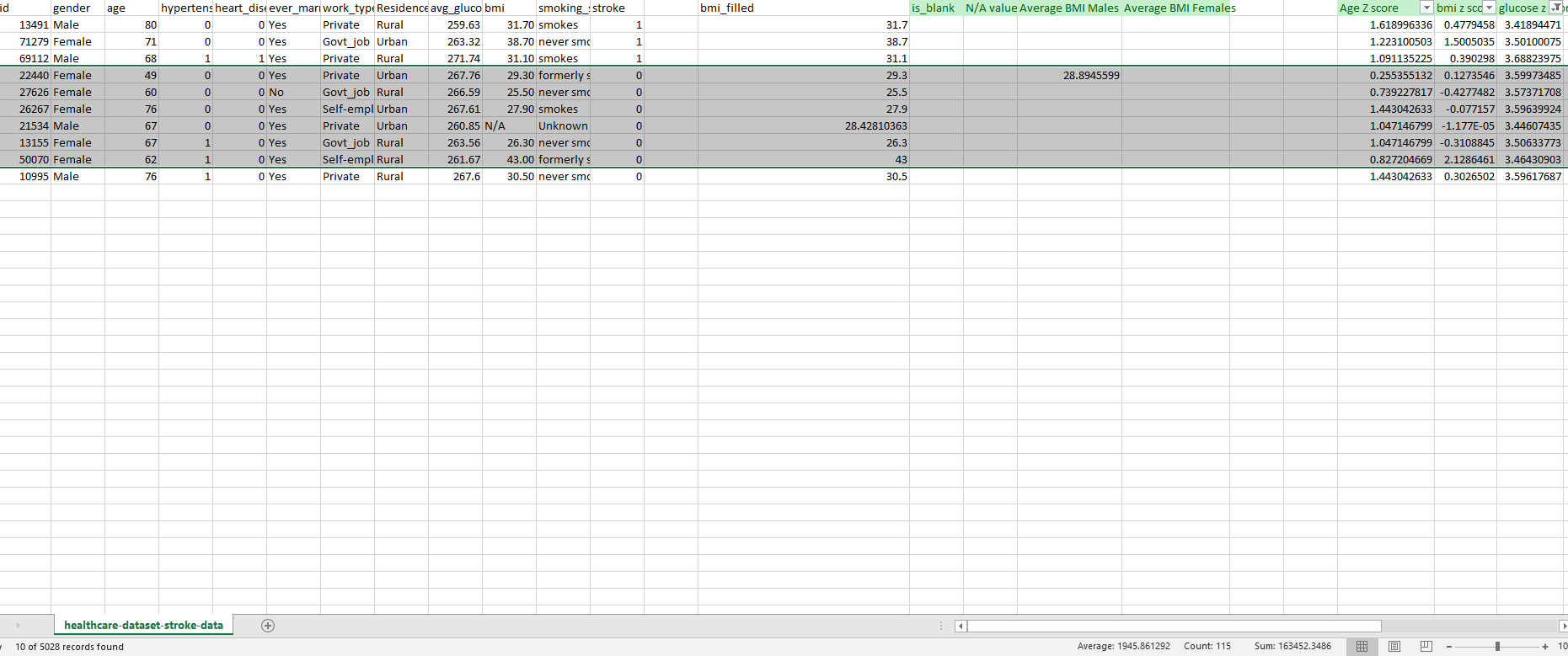


And then as mean and standard deviation updates 26 more records were declared as outliers, so removed them for data as well

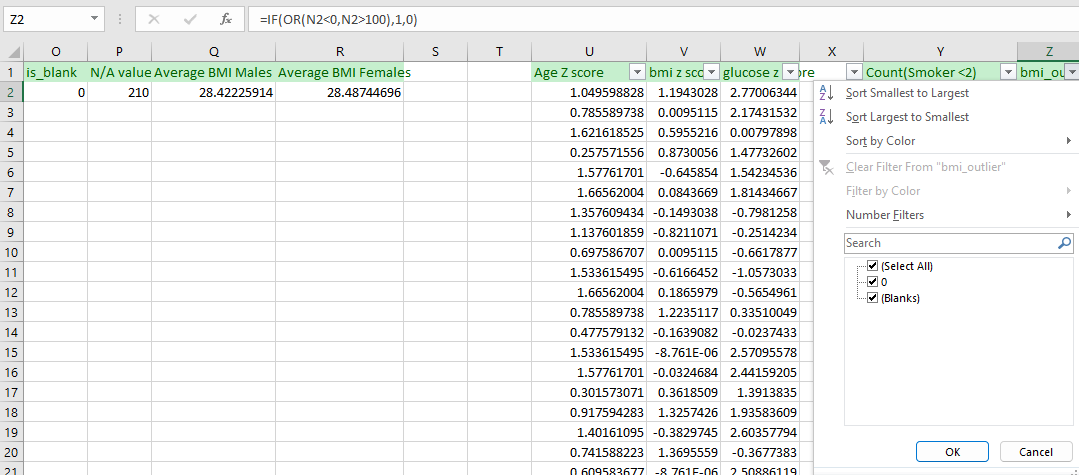


For glucose there are several records lies out of range as well

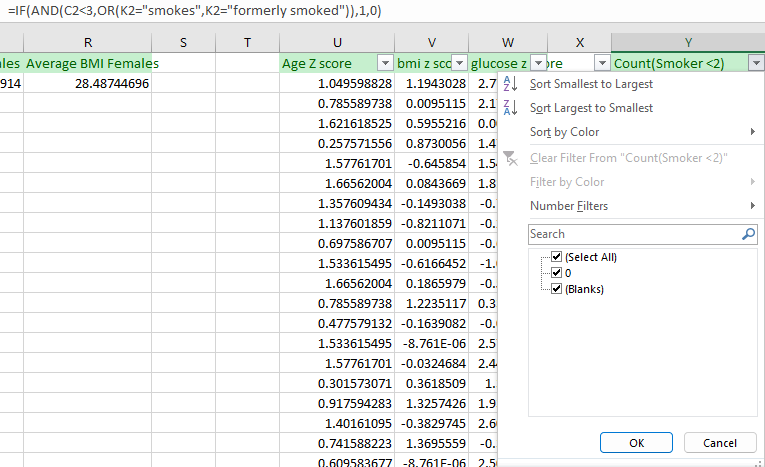


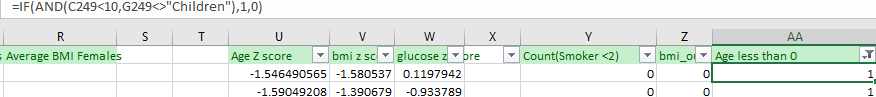
Total 10 records found  


For outlier detection on non-numerical column, create a check if BMI value is less than 1 or greater than 100 which is not possible using if statement that raises a flag to 1 in case of any anomaly. In this particular case, no flags are raised.



Secondly checked if there is a person of age less than 2 and is declared as smoker or formerly smoked than that is not possible as well.

Condition check for smoker less than 2 and no flags raised here as well  


Then another check if a person’s age is less than 10 and marked as employed, two records were found  


Have not removed them as their age is 8 and 9 and marked as private or self-employed so this could be possible in some exceptions so keeping this data.