Report:

**Question #1:** How do countries relate to one another in terms of a set of shared attributes - how can we compare these countries in an appealing way and at the same time simplify and summarize the data? I would like to answer this question using **Agglomerate Clustering** because we will be able to see countries in relation to one another and will be a visual summarization of the data.

**Question #2:** Using the following:

v= Total known number of vaccinations given in the month of February and March for country x.

p= v/population of country x

M= Mean value of p for all countries.

Threshold: If p is less than M for country x then class value of N. If p is more than or equal to M for country x then class value of Y.

Using a training set of random countries (amounting to 60% of the total countries) what decision trees and sets of rules can we formulate using this class value and a variety of different attributes such as total fertility, health, gdpp, etc.? Using the testing ser, we can test this model and determine the accuracy and coverage of the rules.

**Summary of original datasets:**

Dataset #1

Vaccination Progress - <https://www.kaggle.com/gpreda/covid-world-vaccination-progress>

Summary: Vaccination data for countries on a daily basis (time series). This includes data such as number of shots administered on a certain day, total people vaccinated count, etc.

Dataset #2

Country Data: <https://www.kaggle.com/rohan0301/unsupervised-learning-on-country-data?select=Country-data.csv>

Summary: Country data. Includes attributes such as child mortality, exports, imports, income, gdpp, etc.

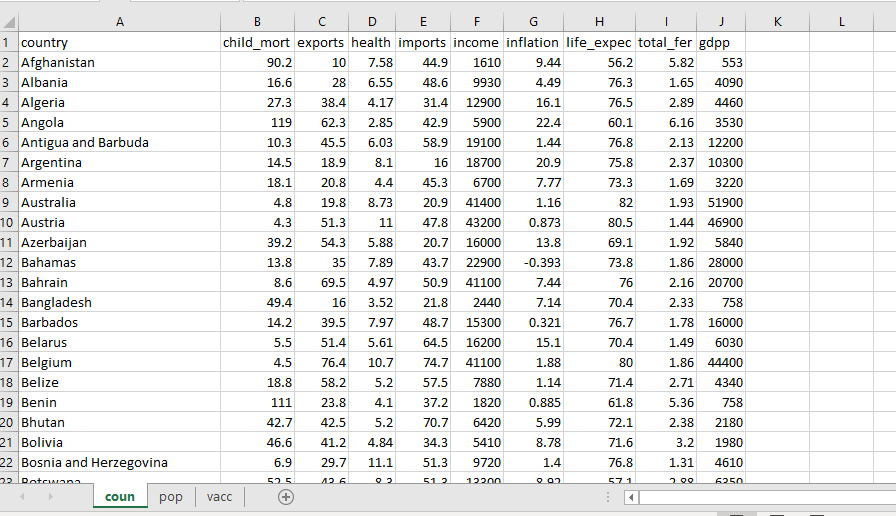
Dataset #3

Population Data: <https://www.kaggle.com/tanuprabhu/population-by-country-2020>

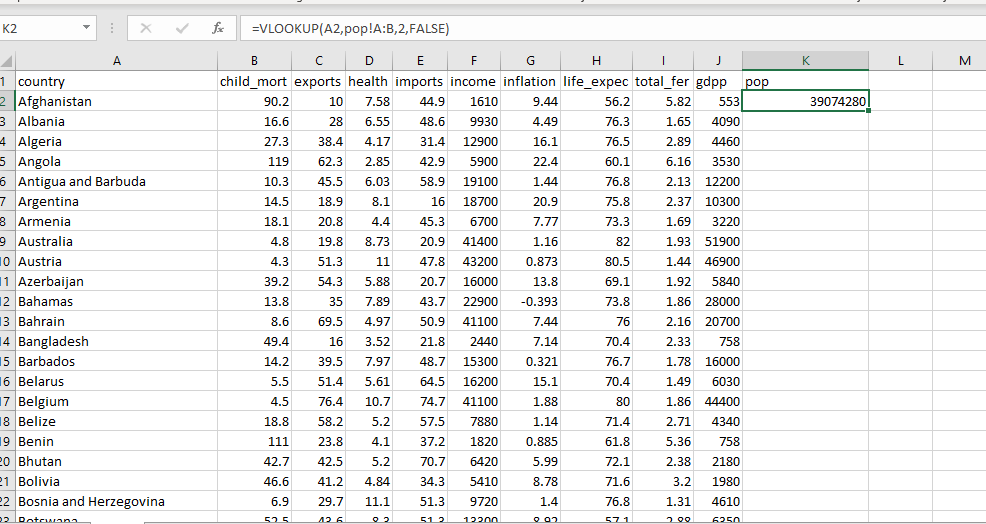
Summary: Includes data on the 2020 population of countries.

**Preprocessing Agglomerative:**

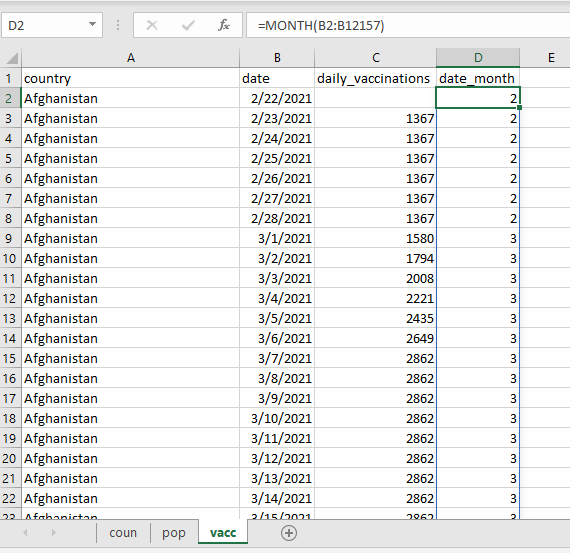
1. We will start out with making the country sheet are base sheet.



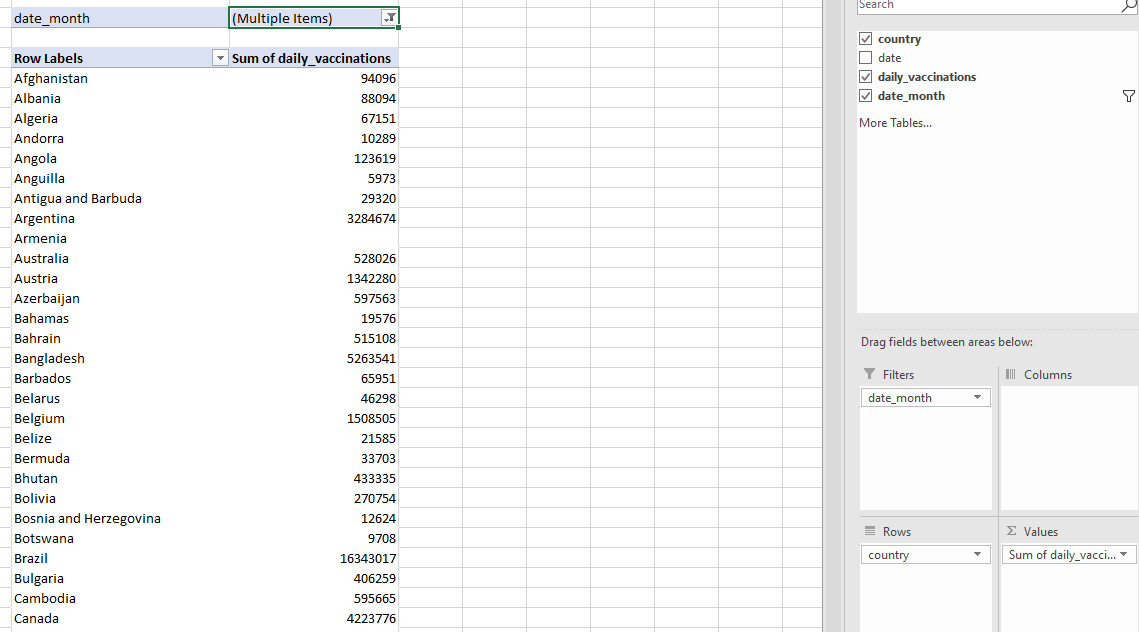
1. We then will use a VLOOKUP to add the population from the pop sheet.



1. Next we have to clean the vaccination data in the vacc sheet to look as the follows. The MONTH formula is used.

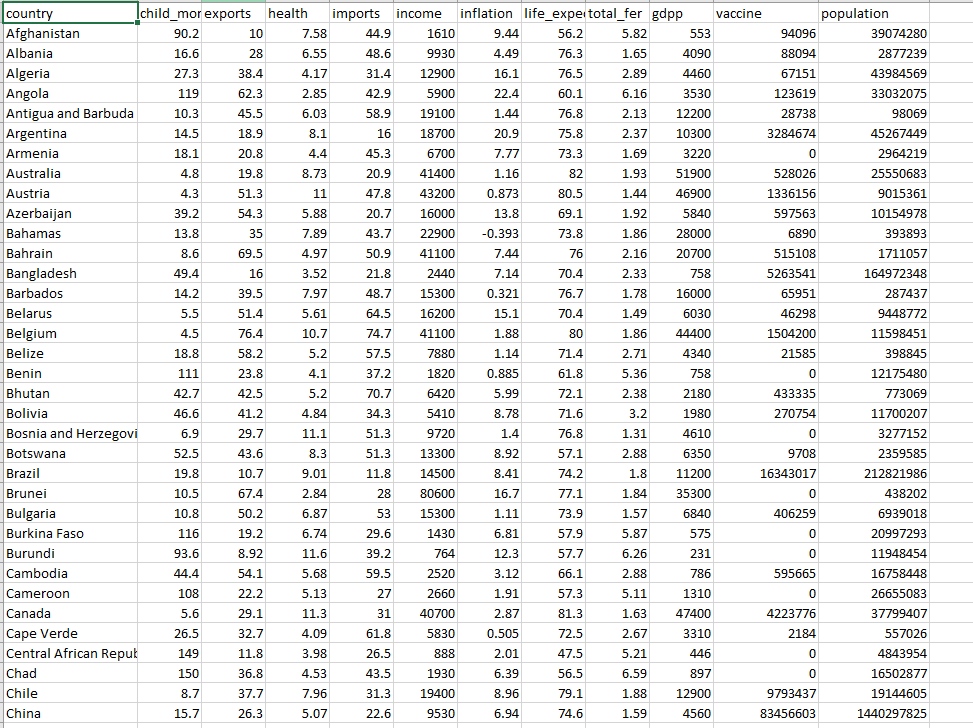


1. After this we will add a filter to the sheet and create a Pivot Table in order to have the summation of the daily vaccines for the months of February and March for every country. I have chosen the months of February and March the table calculated the sums.

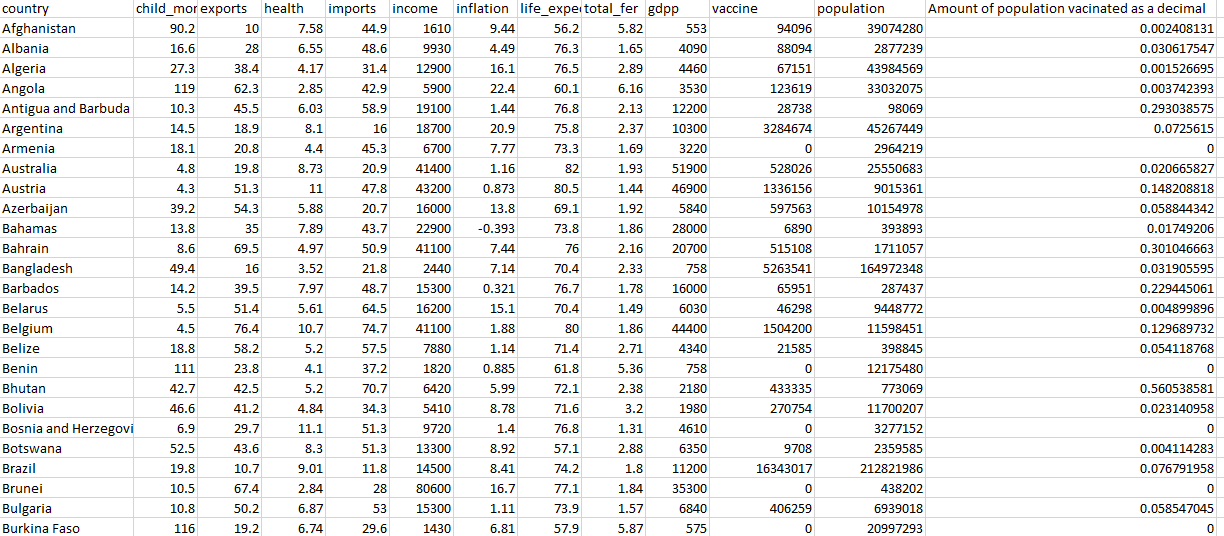


1. The sheet should look like the following with the VLOOKUP for the vaccination data working.

In addition, the gaps for the missing vaccine & population data have been filled in. This was done by finding any NULL attribute that have resulted from the VLOOKUPS and comparing the way the country was spelled in both sheets and making sure it is spelled the same way. This was done manually as there are not an overwhelming number of NULL attributes.



1. Next, we want to get rid of the vaccine and the population columns and add a percentage of the population vaccinated column. We will do this by simply dividing (for country x) the corresponding vaccine value by the population. This we will keep as a decimal.
2. There we have it, we are done preprocessing this dataset to work for Agglomerate clustering. The final dataset should look like the following:



**Data Mining Agglomerate:**

As a reminder, we are using Agglomerate clustering to cluster the above tuples. We will be able to visually see likeness among countries and which countries are nearest to which. The algorithm that will be used takes in the data as a text file in the following format:

Afghanistan Albania Algeria

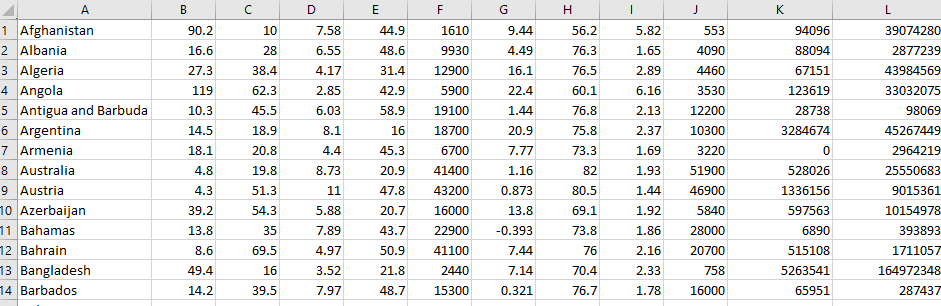
90.2 10 7.58

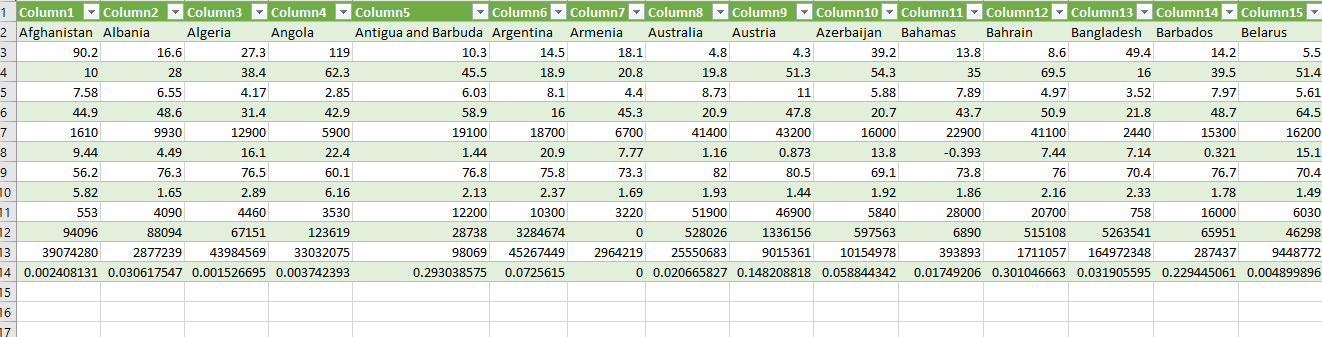
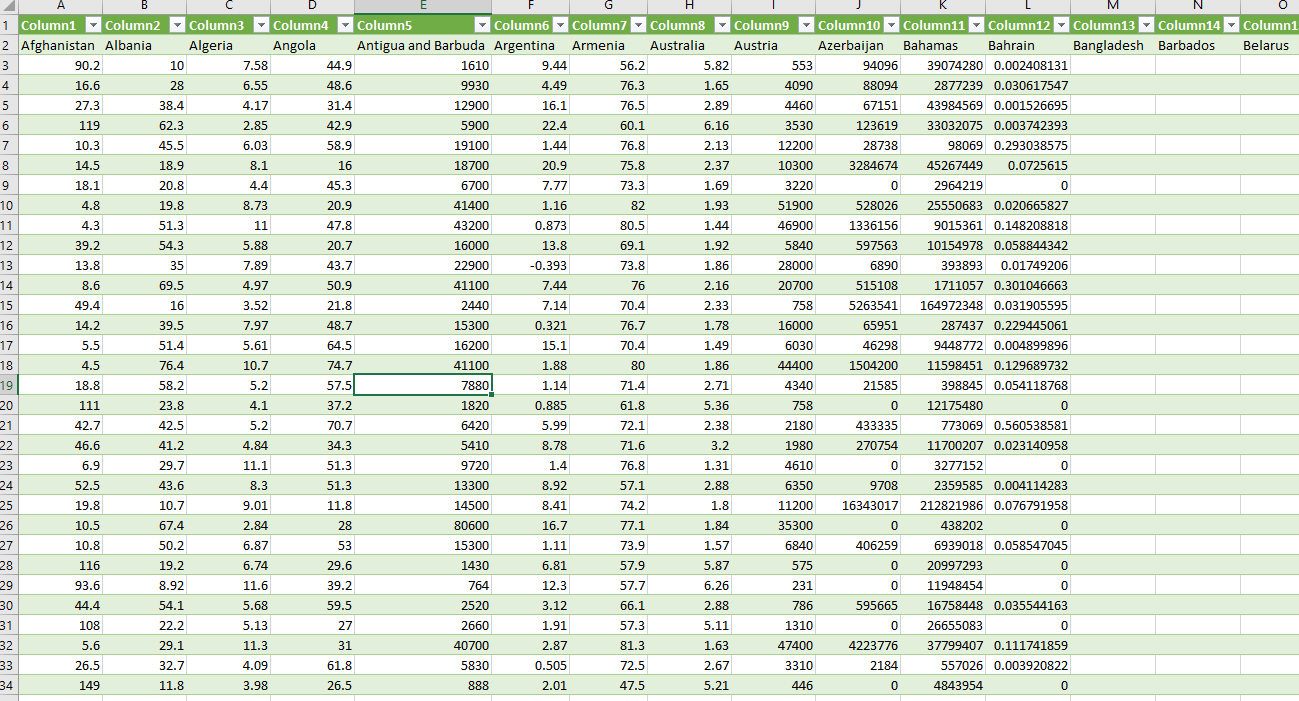
16.6 28 6.55

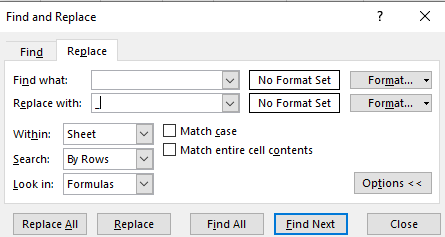
27.3 38.4 4.17

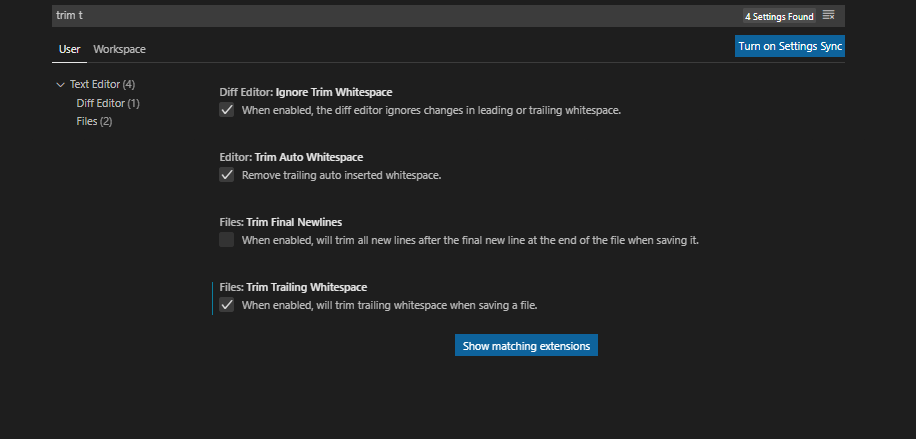
Thus, we must re-preprocess the data in order to fit the needs of this algorithm.

1. Delete the 1st row to make the sheet look the following.



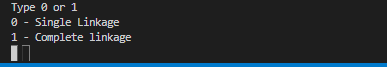
1. Select the whole sheet and click Data -> From Table/Range
2. When in the pop-out box click Transform -> Transpose
3. Then click File->Close & Load
4. It your data should look like this:
5. Next select rows 3-14 copy and paste(transpose) it right below the existing data. Then delete the original copied data and replace it with the pasted data (which is now transposed).
6. We can post the data only as text in a new sheet, getting rid of the colors.
7. We also need to get rid of the spacing within country names and change them with an underscore. For example, United States will become United\_States – this is so that the program does not consider United on country and States another. This can be done by finding all spaces within the sheet and changing them with an underscore.



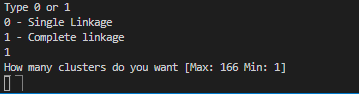
1. Export this sheet as Text (Tab delimited)
2. Within the text file find all tabs and replace with a space.
3. In addition, make sure to get rid of extra whitespaces after each line, this can be done easily in Visual Studio Code by going to File -> Preferences and searching “trim t”

Now that we have re-preprocessed the data, we can proceed to run the algorithm and decipher the results.

1. The algorithm works as the following:



One chooses whether they want to use single or complete linkage.

1. Next we choose how many clusters we want: 
2. For our analysis we will analyze the results of complete linkage & single linkage using 8 clusters. I am choosing this because we can average near 20 countries in each cluster if we use 8 clusters.

Single Linkage:

Complete Linkage:



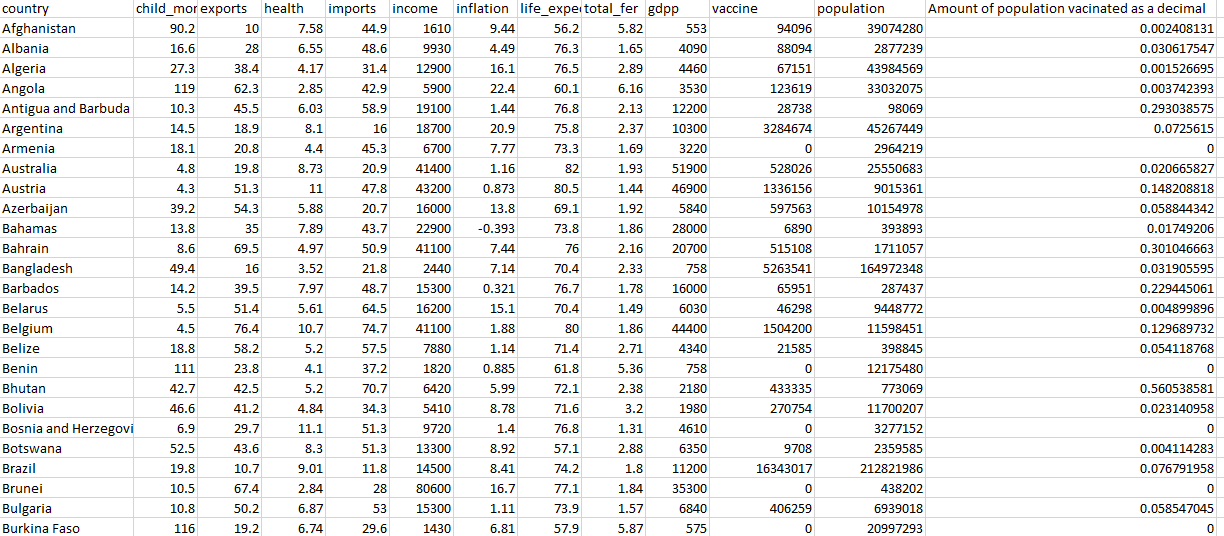
1. These results are difficult to analyze without being biased towards certain countries, especially countries that I have no knowledge of. It would be awkward and improper for me to assume certain results and question why I did not get them. However, there seems, from glancing, some results I did not expect. The issue for me is, are these results unexpected because the algorithm did not work as it should have or I have an unconscious bias of how the results “should be”.

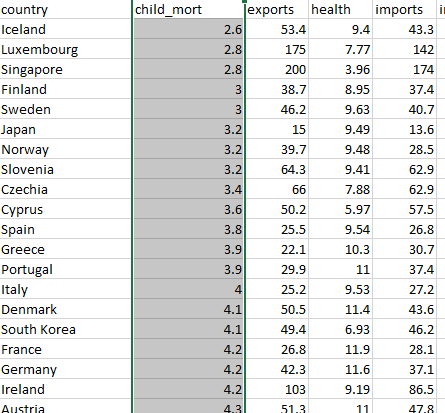
**Summary Agglomerative**

1. I am happy with how my question was answered. My question being**:** How do countries relate to one another in terms of a set of shared attributes - how can we compare these countries in an appealing way and at the same time simplify and summarize the data?
   1. My algorithm was able to relate countries to one another.
   2. We can compare countries based on the preferred number of clusters and preferred linkage method.
   3. The output is shown in a simple way and summarizes the data in an alluring fashion.

**Preprocessing ID3**

1. Let us start preprocessing by using what we had done for Agglomerative and build upon it. This is out sheet:

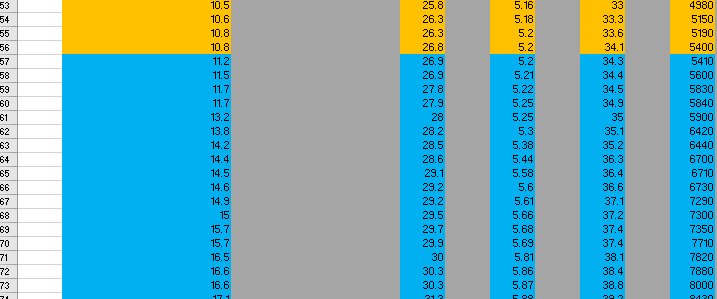
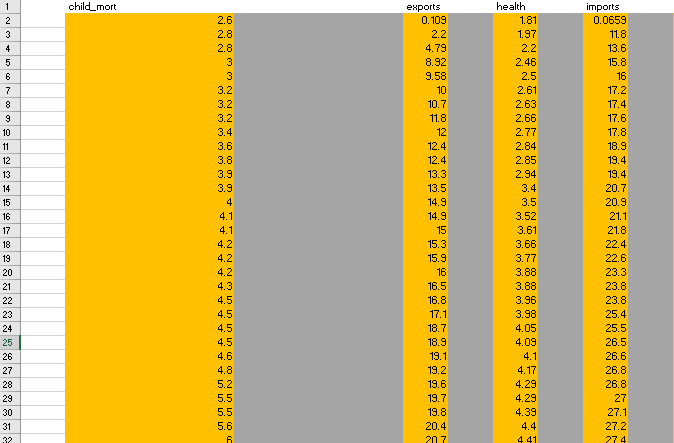


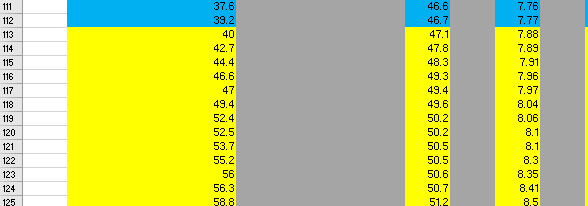
1. Second, we need to turn each of these attributes into categorical attributes. We will achieve this by binning. Lets start off by binning the child\_mort attribute into three different bins.
2. Select the child\_mort column and filter from smallest to largest: 
3. We have 167 countries so out bins will be:

Low – lowest 55 countries

Med – middle 56 countries

High – highest 56 countries.

1. Copy and paste the child\_mort column into a new sheet s. Select the next column, filter by smallest to largest and copy into sheet s. Continue this process for all the attributes. Format it so it looks like this(The colors are changing based on the aforementioned intervals)



1. Add the corresponding value in the grey column and transfer that back onto the main sheet. Make sure you filter by that attribute from lowest to highest before pasting.

