|  |
| --- |
| Institute of Technology Tallaght |
| Interactive Media Design |
| Data Visualization CA1 |

|  |
| --- |
| Michael Smith x00107586  10-31-2016 |

Contents

[CA Summary 2](#_Toc466409757)

[Background 2](#_Toc466409758)

[Seven Stages 3](#_Toc466409759)

[Stage 1 – Acquire 4](#_Toc466409760)

[Stage 2 – Parse 5](#_Toc466409761)

[Stage 3 – Filter 6](#_Toc466409762)

[Stage 4 – Mine 7](#_Toc466409763)

[Stage 5 – Represent 8](#_Toc466409764)

[Stage 6 – Refine 11](#_Toc466409765)

[Stage 7 – Interact 11](#_Toc466409766)

[Problems and Solutions 12](#_Toc466409767)

[Conclusion 14](#_Toc466409768)

# CA Summary

The idea of this CA was to find a dataset and, applying the 7 stages methodology, create a dashboard consisting of several visualizations that would display informative data to a person in an easy to understand way.

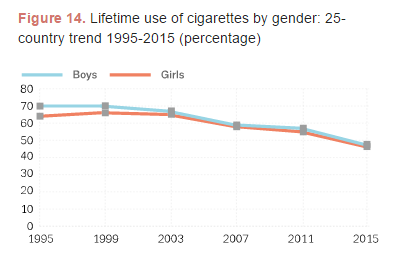
There are several websites containing datasets but in my opinion the challenge of finding a dataset was trying to find one or more that could display something meaningful. After changing datasets numerous times, I finally decided to create a dashboard on youth and more specifically the percentage of 15-16-year-old school children that have tried alcohol, cigarettes and illicit drugs before in their lives. In my opinion it appears that a lot less young people these days are doing these things and I wanted to see if there was any data that could back up my opinion. So, I went to the European School Survey Project on Alcohol and Other Drugs (ESPAD) website and found some datasets on the idea I had chosen. I managed to find the data I was looking for, not just regarding Ireland but also for a number of European Countries, and using these datasets created several visualizations to see if there was any truth in what I thought.

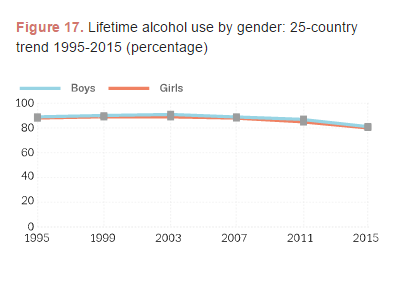
# Background

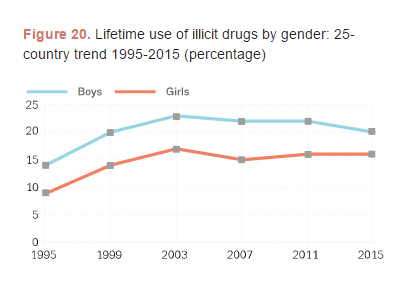
When given the CA, I was struggling to think of an idea that I would like to use. I toyed around with a few ideas but could not settle on one that I would like to proceed with. I wanted to find some data that I could relate to and that would display some meaningful results. I am a smoker myself, and whilst having a cigarette during college I realised that not many of my younger classmate’s smoke. This made me think back to when I began smoking at the age of 15 or 16 and made me wonder whether smoking was as “popular” now for the youth as it was when I was younger and began myself. I thought that I could also include information on how prevalent the use of alcohol and drugs, along with smoking, are nowadays among secondary school children throughout Europe.

drugsandalcohol.ie have previously done similar research into the subject and have created tables displaying results from all of Ireland. The ESPAD have also released past reports that contain data and visualizations regarding the topic.

Below are some visualizations from the ESPAD report







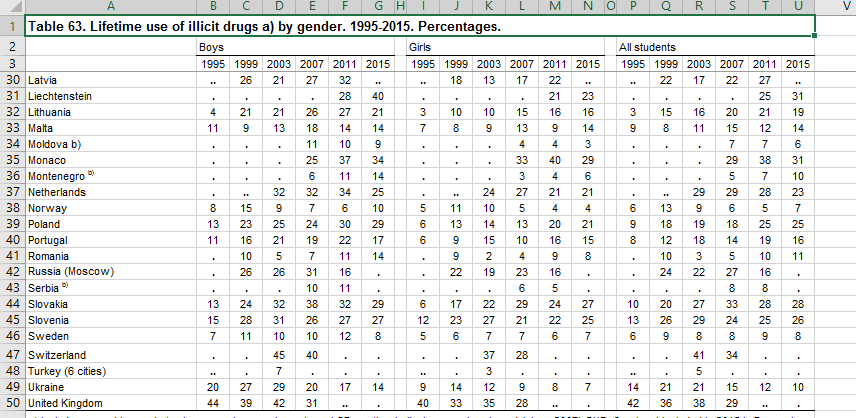
The above pictures are visualizations from the ESPAD 2015 report. They display similar information to what I am trying to create with my visualizations. What I noticed with this report though is that they stick to the one visualization which is the dual lines display. Also in my CA, I have added the average per year and the average of both girls and boys so that it can be easier to compare a country as a whole to the average for all countries.

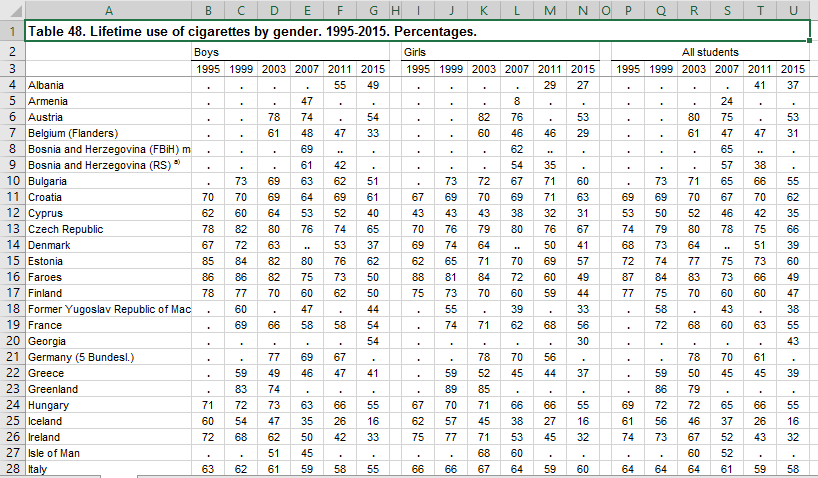
# Seven Stages

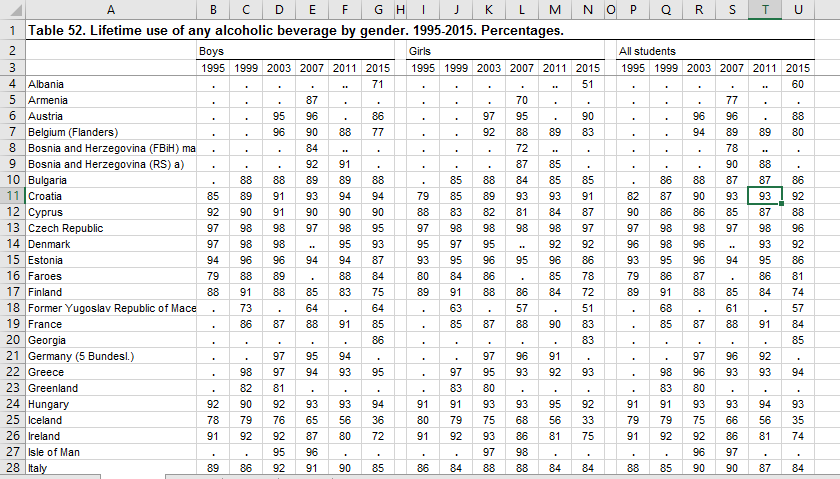
To transform the original data that I began with I used the seven stages as a guideline to help convert my data to a point where I could create the visualizations I wanted. I will now explain stage by stage what I did to the dataset at each of the seven stages involved.

## Stage 1 – Acquire

The first stage of the CA was to actually acquire the dataset that I would use. This stage was difficult for me as I could not settle on a dataset that I was happy to use. I originally began with a film dataset and then switched to a number of datasets regarding population, and then finally settled on the datasets I used regarding alcohol, cigarette and drug usage in 15-16-year-old European school children. To get the information I needed I looked for numerous sources as I originally just wanted to do the dataset about Ireland so I looked for websites that would have the information that I needed but could not find anything close to what I felt I needed. I then found the European School Survey Project on Alcohol and Other Drugs (ESPAD) website and managed to get the information I needed from that. I also saw that it contained information on the same topics about several European countries so decided then to include them all instead of just concentrating on Ireland. Below are the datasets I included in my CA.



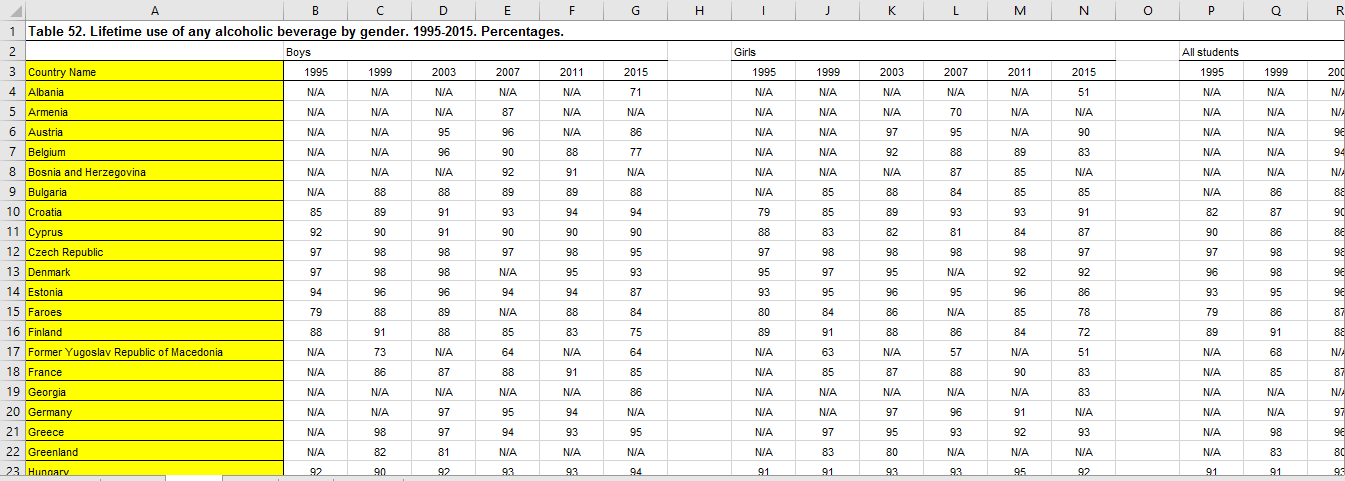




These images are just some of the datasets that I used for the CA. All the links to the excel sheets containing the data are available at <http://www.espad.org/report/additional-tables>

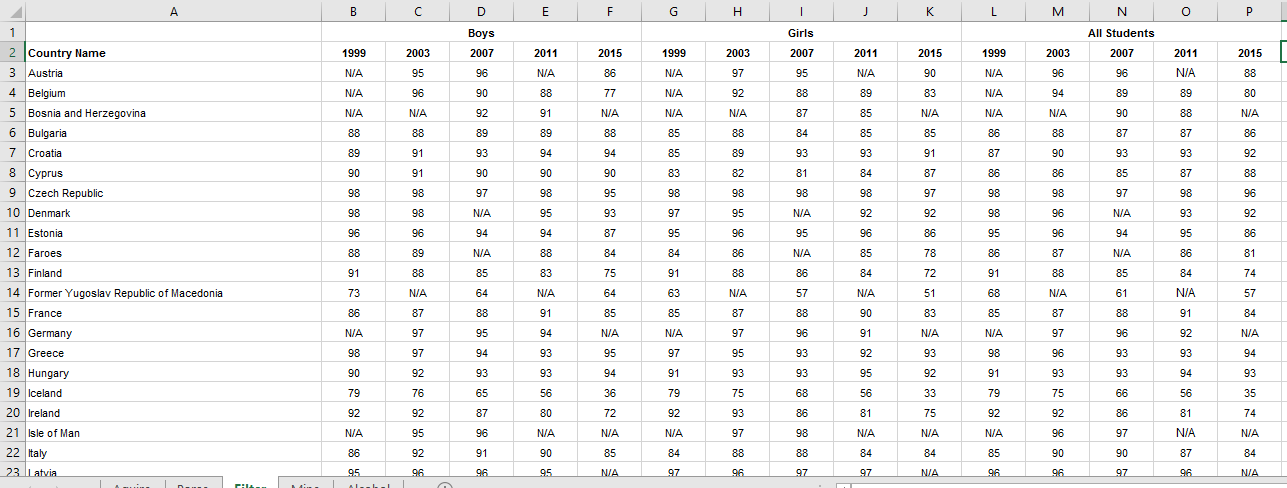
## Stage 2 – Parse

For the second stage the idea was to structure the data into a more understandable form. I began this by going through each country and year and replacing the cells were there was no information available. I simply put N/A into the cells were there was no information or just a dot. I then ordered the data alphabetically by country to make it easier to read and navigate. I also cleaned up the country names themselves. I had previously used OpenRefine to parse the previous datasets I was going to use and thought about using it again for this dataset but in the end decided to just change the values manually as there was not too much to change. I also tried to use Azure Machine Learning to parse the data. I had used this technology as part of my 4th year project and realised that I could change the data by putting my dataset through this. Unfortunately, this did not work out how I had planned and I could not get it to work correctly so again decided that changing the values in excel itself would be the best option. Below is an image of the data after I had completed the Parse stage.



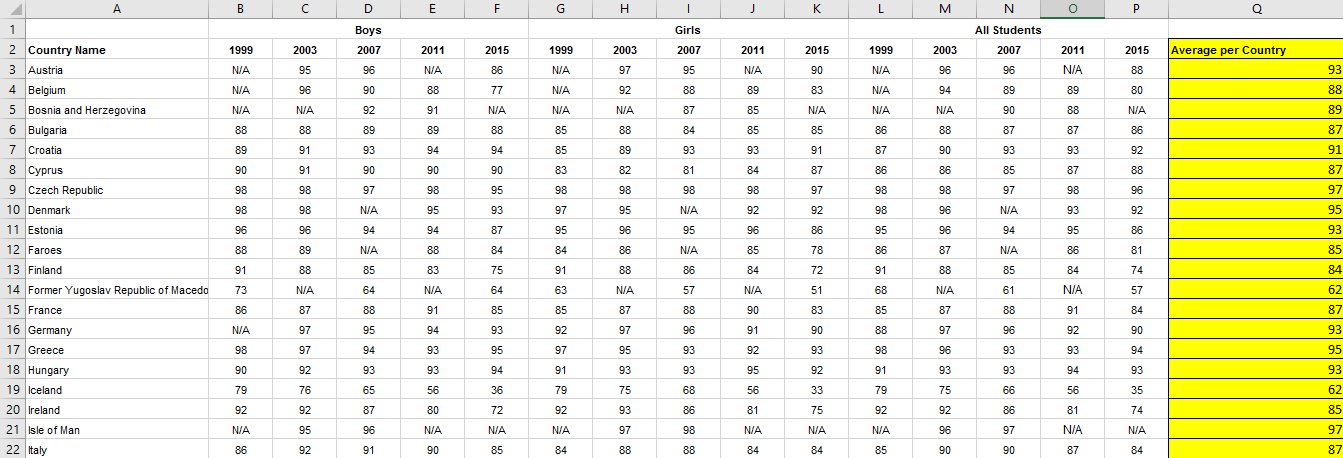
## Stage 3 – Filter

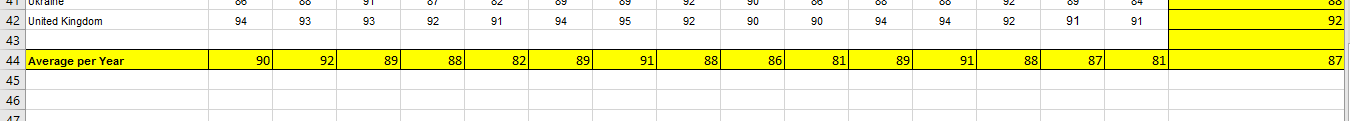
When I reached the Filter stage of the process I already knew that there were certain things that I wanted to remove from the data. Firstly, I noticed that the column for the year 1995 had a lot of missing values. I thought about this and figured that I would be better off excluding the whole column as I would still have information on each topic from 1999 – 2015 which would be enough for the visualizations I wanted to create. So, I removed that column for boys, girls and all students. I also removed the blank columns between the boys, girls and all students. In addition, I removed any text that was in the dataset that was not critical to the information I wanted. Finally, I removed a small number of countries that I felt there was not enough data that I could display something meaningful through a visual representation. Below is an image of a dataset after going through the Filter stage.



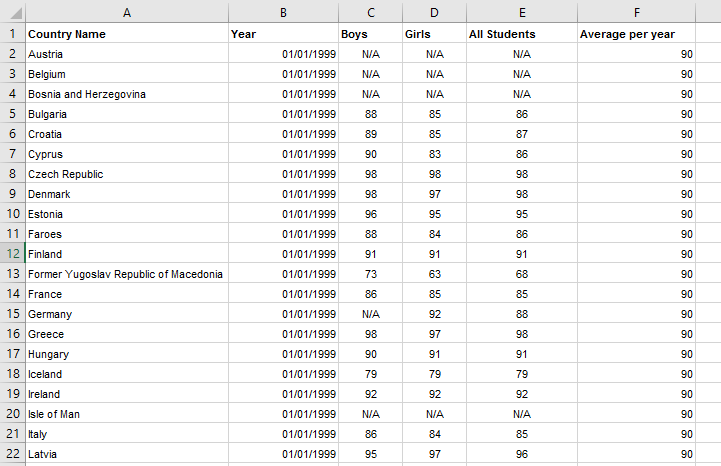
## Stage 4 – Mine

The Mine stage of the seven-stage process involved me creating an additional column and row within the dataset that would display an average figure depending on what dataset I was working on. For example, in the Alcohol dataset I took the data for each year for each country and created a column at the end that would display the average. I also created an average for every country for each year which I would go on to use in the data visualizations. Below is an image of the dataset with both the added fields after the Mine stage.





In addition to the Mine stage I also added another step myself. When connecting to a dataset I realised that I was having trouble ordering by year the way I had my dataset laid out. To counteract this I decided to rearrange the dataset so that now I had a column that would display the year. This meant that I would have to list each country 5 times as that was the amount of years I was working with (1999, 2003, 2007, 2011, and 2015) This also meant that I could now just have one column for each of Boys, Girls and All Students. Below is an image of the dataset after I rearranged it.

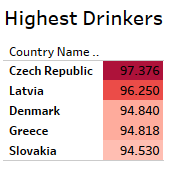


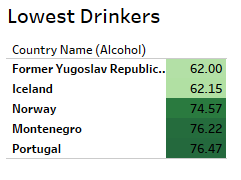
## Stage 5 – Represent

During this stage I chose the graphical representations for the datasets I had chosen. I began by creating a bar chart that would show the data I had found on the percentage of 15-16 year olds that had previously tried alcohol in their lives. I first took the average percentage of all students per year for each individual country and used this figure for each of the years in the bar chart. I then took the average per year through every country in the dataset and displayed this as a line over the bars. This enabled me to compare each country against the average and from this I could tell which countries were above or below the average and also I could gather from the bar chart whether the percentage of students that had tried alcohol was increasing or decreasing over time which in itself I felt gave some very interesting results. Below is an image of the graph.

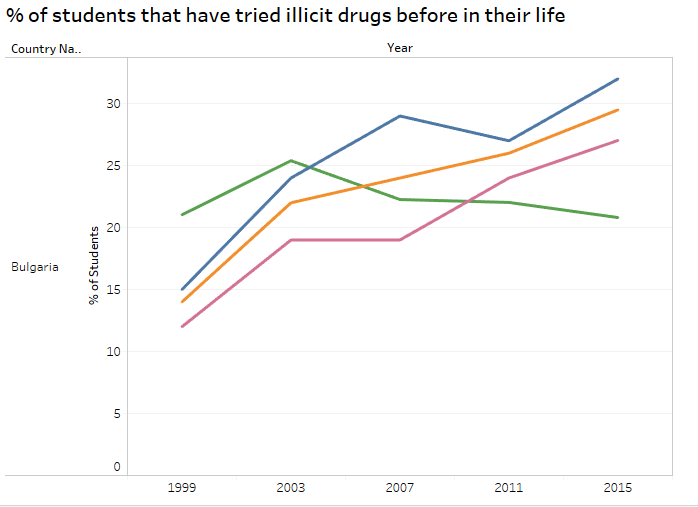


I also decided to include a table that would display the 5 countries with the highest and lowest average percentage of children for Alcohol. I decided to use a highlight table as I felt it would be the easiest way of displaying this information. Both tables are displayed below.

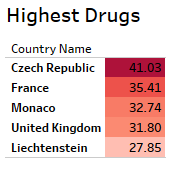


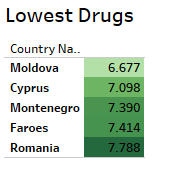


I next wanted to display the percentage of 15-16 year olds that had previously tried any illicit drug in their life. When I say illicit in this context this means any drug including Cannabis, amphetamines, cocaine, crack, ecstasy, LSD or any other hallucinogenic, heroin and GHB. For this graph I wanted to do something different than the previous graph I had made so I decided to create a graph composed of lines. In this graph I display the percentage of Boys, the percentage of Girls, the percentage of All Students for a specified country and also the average percentage based on all countries in the survey. Again, this graph displayed some very interesting results. Below is an image of the graph.



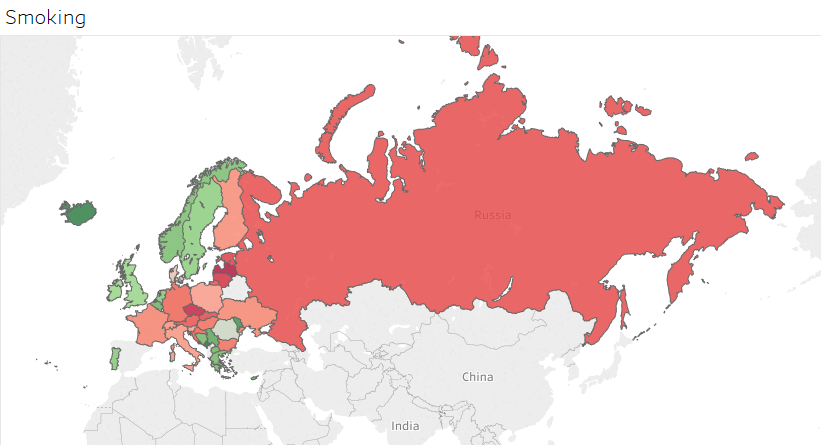
As before with the results of alcohol I wanted to display the 5 countries with the highest average percentage for drugs as well so again I used the highlight table format to keep things simple and so that somebody viewing the dashboard would know what these tables represent after reading the tables about alcohol. Below are images of the tables.





Something of interest to note is that the Czech Republic has the highest percentage of students that have tried drugs before AND the highest percentage of students that have tried alcohol before!

Finally, for my third dataset which contains information regarding the average percentage of children that had tried cigarettes before I decided to try and display this on a map of Europe which would colour code countries depending on whether they were above or below the average for all countries. Below is an image of this.



These are the graphical representations of the datasets I have chosen. I feel that they display the data in a meaningful way and that a lot of information can be gathered from the images.

## Stage 6 – Refine

For the Refine stage I added headings to each of the graphs that give a name to each graph. I also used colours to represent each specific piece of data and tried to stick to the same colour code throughout. I chose blue and pink to represent boys and girls and then chose orange to represent all students. I chose green to display the average for all countries as well.

When it came to the tables and the world map I chose a green – red scale. Green indicates a better average percent and red represents a worse average percent. I chose these colours for these graphs as I feel people associate green with good and red with bad.

I also added captions to each graph that gives a basic description of what each graph represents.

I feel that the changes made the dashboard a lot easier to understand and also helped identify each graph.

## Stage 7 – Interact

The final stage in the process is to enable a user to interact with the dashboard. In my dashboard I included a single value slider which will display information on both graphs (alcohol and drugs) for whatever country the user has chosen. If the user hovers over any piece of data they will also see what that piece of data represents. Also on the second dashboard, a user can search for a specific country to see what the average percentage of students that had tried cigarettes is. The user can also just search by year and the map will change colour depending whether the average is higher or lower than the European average.

# Problems and Solutions

During the process of finding the dataset and creating the dashboard I came across numerous problems along the way. I will outline some of the problems I encountered.

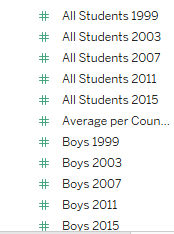
1. Choosing a dataset

When I came to choosing my dataset I spent quite some time searching for one that I was happy to do the CA on. As mentioned before I originally chose to do a film dataset but after talking to you about what visualizations I could create I felt that this would not be a good dataset to continue with.

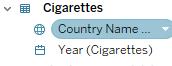
I then changed my dataset to a number of datasets regarding population worldwide and although I found several good datasets, after putting them through the first 4 stages of the seven stage process I realized that there was not much interesting information within the data. I found it hard to try and link the separate datasets together into something cohesive so again I decided to change dataset. Luckily I found the dataset that I eventually used and I am happier with the information that this data gives.

1. Displaying the year the way I wanted to use it.

The next big problem I came across was that in my graphs I wanted to use year as a dimension but found that when I connected my datasets, year was being displayed as a number rather than a date and when I changed the data type from numeric to date it automatically changed each year to 1905. I was perplexed by this for quite some time and eventually decided upon myself that I could create a year column within my data and then read the values in through this. Another reason my year wouldn’t work was because it was just the year alone and no day or month and tableau did not like me using the date this way. To counteract this problem, I changed the date from the year alone to 01/01/YEAR and this in turn fixed the problem I was having when importing the datasets. This edit allowed me to read in the column as a date instead of numeric and after that I had no problem with the year. Below is the way that the year was being displayed when I was having trouble with it, as you can see it takes each column and each year so there was a lot of measures:



When fixed it looked like this:



1. Incorrect graph types

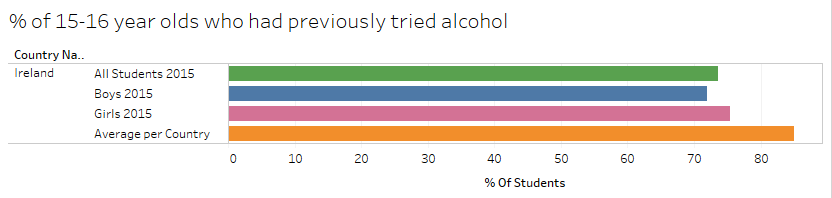
This problem also tied into the year problem I was having. When I first began creating visualizations it was only allowing me to do bar-charts and I wanted to mix my graphs up a bit and have a couple of different ones as I felt all bar-charts and histograms wold not look good. The problem I was having was that for the line graph I wanted to use needed a date dimension and as I previously said I originally could not get the year in as a date. Once I had changed the column in the dataset for the above problem, this also solved this problem.

1. Connecting datasets

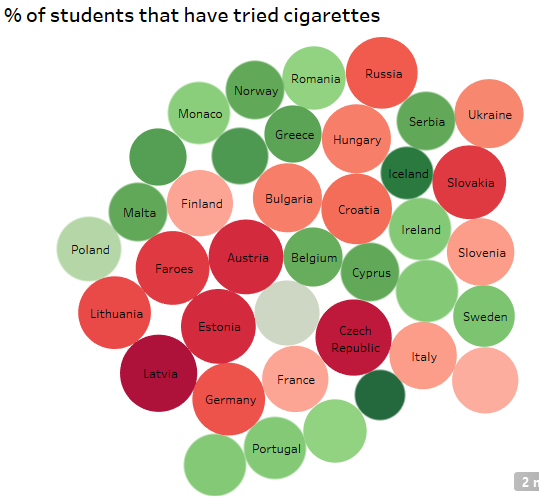
This was a problem I encountered when I tried to connect my first dataset to my second one. I eventually fixed this problem by using an inner join and using Country Name to join the datasets as the country names were the same in each of my datasets. I feel that there may have been a better way to connect my datasets still but I could not find one myself.

These were the biggest problems I encountered during the CA, I also came across other problems with the visualizations but the others were minor enough and weren’t as hard to fix as the above.

In regards to trying graphs that didn’t work this happened a couple of times. I started with a bar chart to display information but I think that the bar chart/line graph displayed the data better so I decided against using the bar chart one. This chart was also created when I was having trouble with the years so you can see that the year is beside the bar and that is not what I wanted. Image is below.



Another visualisation I tried was a bubble chart for the average percent of students that had tried cigarettes. The reason I did not include this graph was because I felt that it did not add anything to the dashboard and that the data it was displaying could be displayed in a much more informative way. Below is what the bubble chart looked like



You can see that even some of the bubbles do not display the countries name so that was also a problem that I felt could be solved by using a different chart. In the end I chose to display this information through a filled map chart and I personally think that this is a much better visualisation as the user can see the country on a map with all the names included.

# Conclusion

I really felt that this CA was going to be a lot easier than it was. There were a lot of tricky parts that I think a lot of people in the class did not really think about until we were doing it. Overall I am happy with the data that I have displayed, there are still some more things I would have liked to try and maybe I could have improved upon the work I did, but in general I like what I managed to create.

I think that the visualizations I created display some very interesting data. For instance, we can see that since 1999 the number of students that have previously tried cigarettes in Ireland has actually been dropping steadily. We can also see that North-Western Europe overall the students are trying cigarettes a lot less than their peers around the rest of Europe. When it comes to alcohol the same can be said with Ireland actually showing a massive drop of almost 20% in the difference of all students that had tried alcohol between 1999 and 2015. Even with drug use the number has been steadily dropping also. On the opposite end of the scale we can see that the Czech Republic seems to have a problem with their school children being exposed to these vices as they scored above average for drugs and alcohol.

In the end, I am happy with the datasets I chose, although not my original idea I feel that the results I got from the data were a lot more interesting than any of the other datasets I wanted to do. I encountered a lot of problems along the way and found that other students also came across similar problems so it was nice to have other people’s input and help on some of the problems. In conclusion, I think this was an interesting CA that I enjoyed doing once I worked out the problems I had and found that tableau is an excellent tool to display visualisations of data that is more interesting and easier to read than from an excel spreadsheet.