# Data & IT Handover Document

Henry - 23929804@student.uwa.edu.au
Vinayak - 24066272@student.uwa.edu.au

## Raw Data Collection and Cleaning

It all starts with the raw data which is given to us by the Law Team. They research the countries and years in which the legislation was passed. The dataset primarily consists of **Countries**, and **4 different legislations** and a **Nationwide** column which tells when a country has covered all the four legislations. More details are given in the HOPE report. Please read the report to understand more about the dataset. Our job is to analyze the data based on the years in which the laws were passed. As data science/IT interns, we only deal with the numbers. So the original dataset was taken and reference numbers, gatekeeper's body of policy and notes were removed. Some extra columns such as levels, difference between each legislation were added to better understand the data. You can remove any of these columns(**NOT THE PRIMARY ONES**) based on the requirement of analysis you are doing.

| | Time Period | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Countries | Nationwide Legislation on | Reference Nationwide | In Penal Institutions/Or | Reference in Penal | In Schools | Reference in Schools | In Out of Home care | Reference out of Home | In Homes | Reference in home | Gatekeeper bodies of policy | Notes |
| Sweden | 1979 | [1] | 1979 | [4] | 1958 | [4] | 1958 | [93] | 1979 | [1] | Ministry: Ministry of Health and Social Affairs (119) | In Schools: "Elite secondary schools: 1918 |
| Finland | 1983 | [2] | 1983 | [2] | 1914 | [2] | 1983 | [2] | 1983 | [2] | Ministry: Ministry of Social Affairs and Health (91) | In penal Institutions/orphanages: No details of prohibiting legi |
| Austria | 1989 | [19] | 2011 | [19] | 1974 | [19] | 1989 | [19] | 1989 | [19] | Ministry: The Federal Chancellery (126) | In 1992 a Supreme Court case confirmed the prohibition of v |
| Denmark | 1997 | [27] | 1933 | [27] | 1967 | [27] | 2007 | [27] | 1997 | [27] | Ministry: Ministry for Family and Consumer Affairs (28) | In penal institutions: corporal punishment has been unlawful |
| Latvia | 1998 | [29] | 1998 | [29] | 1998 | [29] | 1998 | [29] | 1998 | [29] | Ministry: Ministry of Welfare (30) | One third or 32 percent of parents in Latvia use corporal pu |
| Germany | 2000 | [34] | 2000 | [34] | 1994 | [34] | 2000 | [34] | 2000 | [34] | Ministry: Federal Ministry for Family, Senior Citizens, Women and Youth (131) | In the GDR, Corporal punishment was abolished in 1949. (18 |
| Israel | 2000 | [37] | 2000 | [37] | 2000 | [37] | 2000 | [37] | 2000 | [37] | Ministry: Ministry of Labor, Social Affairs and Social Services (134) | In penal institutions: corporal punishment is unlawful as a di |
| Bulgaria | 2000 | [96] | 2000 | [96] | 2000 | [96] | 2000 | [96] | 2000 | [96] | Ministry: Ministry of Labor and Social Policy (40) | Corporal punishment was prohibited in 2000 in the Child Prot |
| Turkmenistan | 2002 | [44] | 2002 | [44] | 2013 | [44] | 2002 | [44] | 2002 | [44] | The Ministry of Labour and Social Protection of the Population of Turkmenistan | |
| Iceland | 2003 | [194] | 2003 | [194] | 2003 | [194] | 2003 | [194] | 2003 | [194] | Ministry: Ministry of Education and Children (142) | |
| Ukraine | 2003 | [52] | 1996 | [52] | 1991 | [52] | 2003 | [52] | 2003 | [52] | Ministry: Ministry of Reintegration of the Temporarily Occupied Territories of | In penal institutions: CP is unlawful as a disciplinary measure |
| Romania | 2004 | [50] | 1969 | [50] | 1948 | [50] | 2004 | [50] | 2004 | [50] | Ministry: Ministry of Labour and Social Justice (144) | |

**ORIGINAL DATASET (LAW TEAM)**

| Serial No. | Countries | Nationwide | Penal Institutions | Schools | Out of Home Care | In Home care | Levels | Nation - Penal | Nation - Schools |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Afghanistan | | 2019 | 2008 | | | 2 | | |
| 2 | Albania | 2017 | 2017 | 1995 | 2017 | 2008 | 4 | 0 | 22 |
| 3 | Algeria | | | 2008 | | | 1 | | |
| 4 | Andorra | 2014 | 2007 | 2014 | 2014 | 2014 | 4 | 7 | 0 |
| 5 | Angola | | | | | | 0 | | |
| 6 | Antigua and Barbuda | | | | | | 0 | | |
| 7 | Argentina | 2016 | 2010 | 2016 | 2016 | 2016 | 4 | 6 | 0 |
| 8 | Armenia | | | | | | 0 | | |
| 9 | Aruba | 2016 | 2016 | 2016 | 2016 | 2012 | 4 | 0 | 0 |
| 10 | Australia | | | 2015 | 2012 | | 2 | | |
| 11 | Austria | 2011 | 2011 | 1974 | 1989 | 1989 | 4 | 0 | 37 |
| 12 | Azerbaijan | | 2012 | 2009 | | | 2 | | |
| 13 | Bahamas | | 1984 | | 2015 | | 2 | | |
| 14 | Bahrain | | | 1992 | | | 1 | | |
| 15 | Bangladesh | | | 2011 | | | 1 | | |

**Cleaned Dataset ( DATA SCIENCE/IT TEAM)**

Also you can put validation checks / constraints if you like. For Example - I have a small constraint in Legislations Columns for 4 digits. If you try to put a year value in 3 digits it gives an error to prevent typos. Similarly, you can put constraints to improve data integrity and quality.

Also after cleaning please audit your data multiple times to make this data error proof.

# Visualization Insights

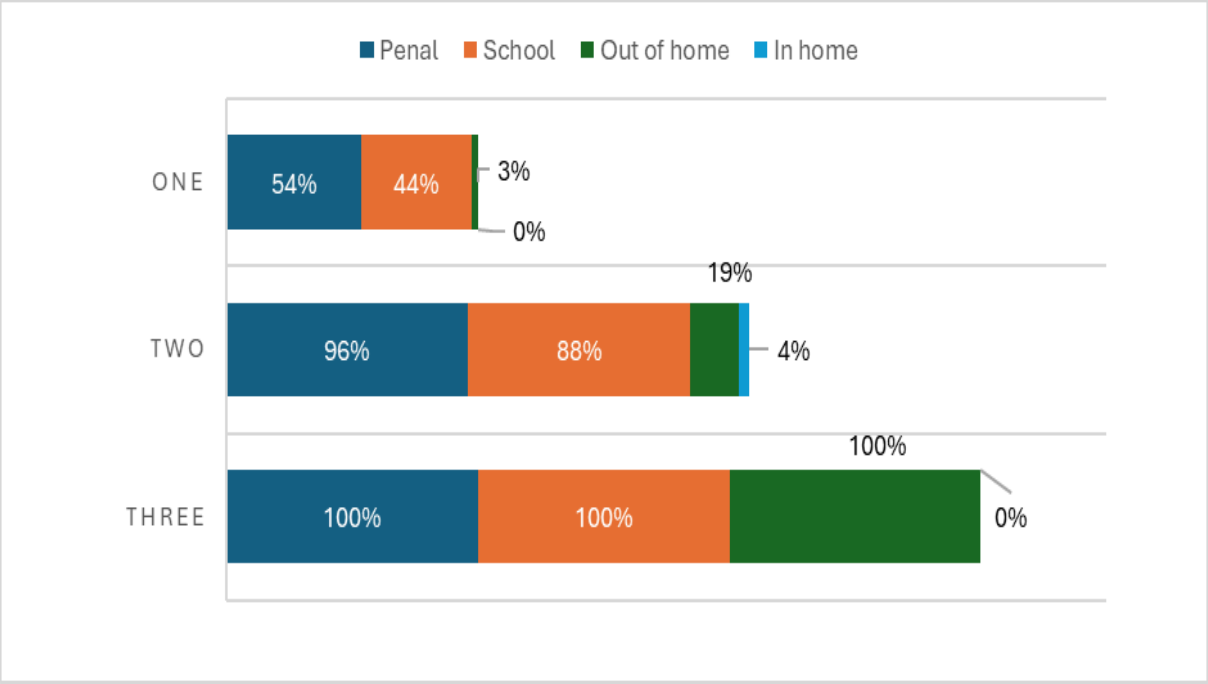
Our next step is to analyze the data which was given to us.

## Tools & Programming Language Used

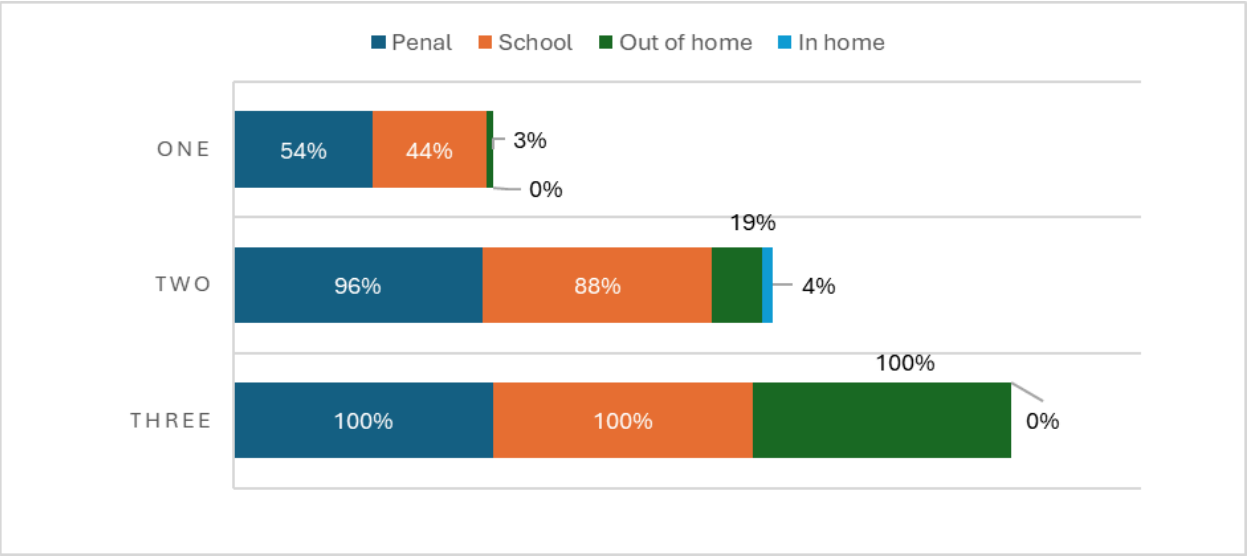
We used primarily Tableau , Python , R and Excel for our analysis. Feel free to use any tool that is convenient for you. I will take you through each tool one by one.

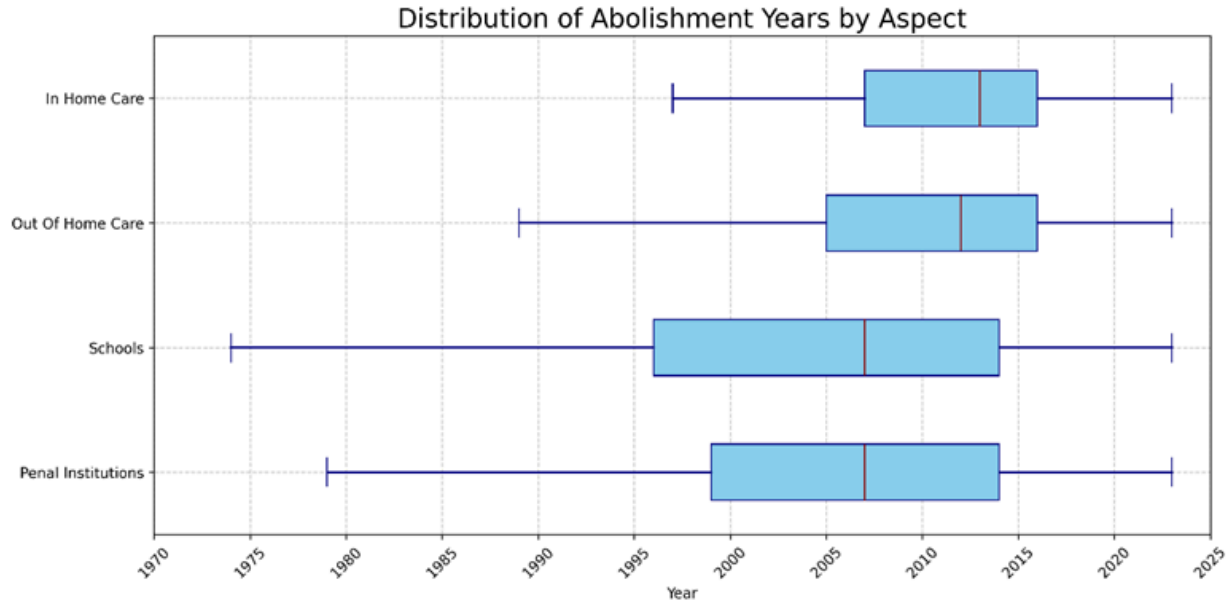
### Excel

This is the most widely used tool in the industry and practically this tool can be used for anyything from cleaning , modeling and making visualizations. I used tableau and R because I had more experience with those tools and graphs were also dynamic. Henry used python to make graphs which are also being used in our HOPE report. After cleaning , these graphs were made in Excel.




These are some of the graphs which were made but not included in the report. More graphs can be found in the HOPE report.



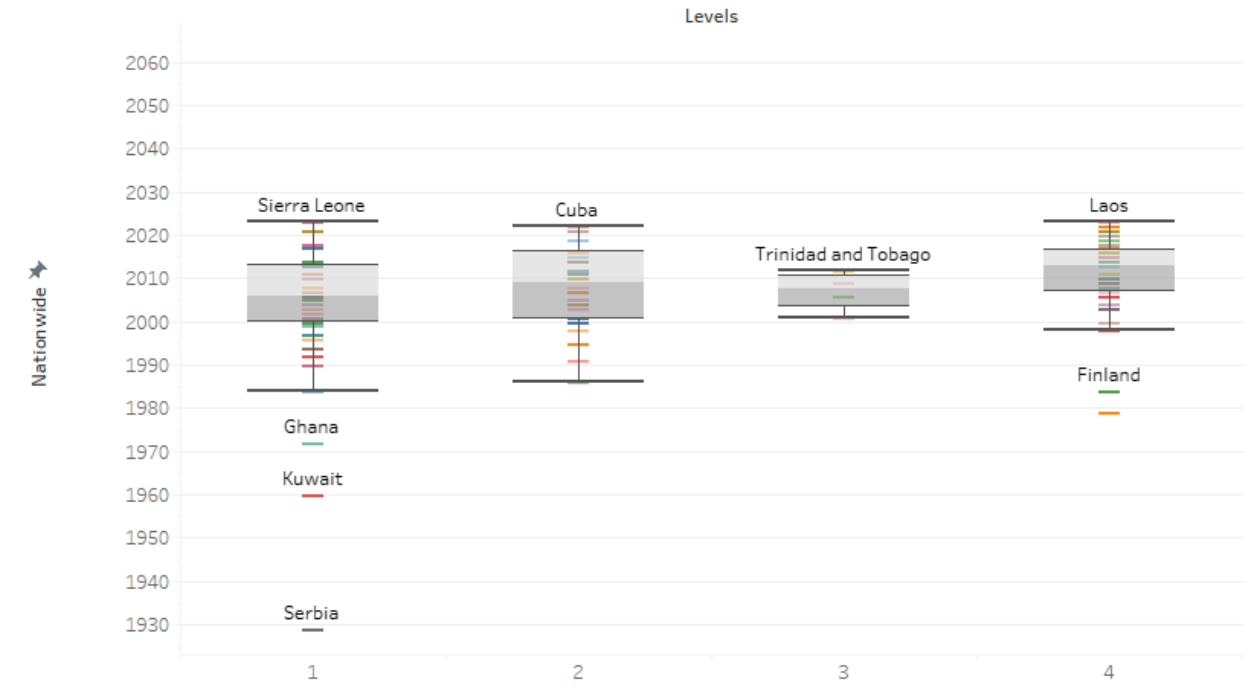

VBA was also used (The codes are given in the Handover Material). It was used when we had to increase reference numbers by some unit or change texts over multiple areas. So instead of doing manually or making big formulas I used that. Please explore that if you are more interested otherwise ignore this. It just makes things easier.

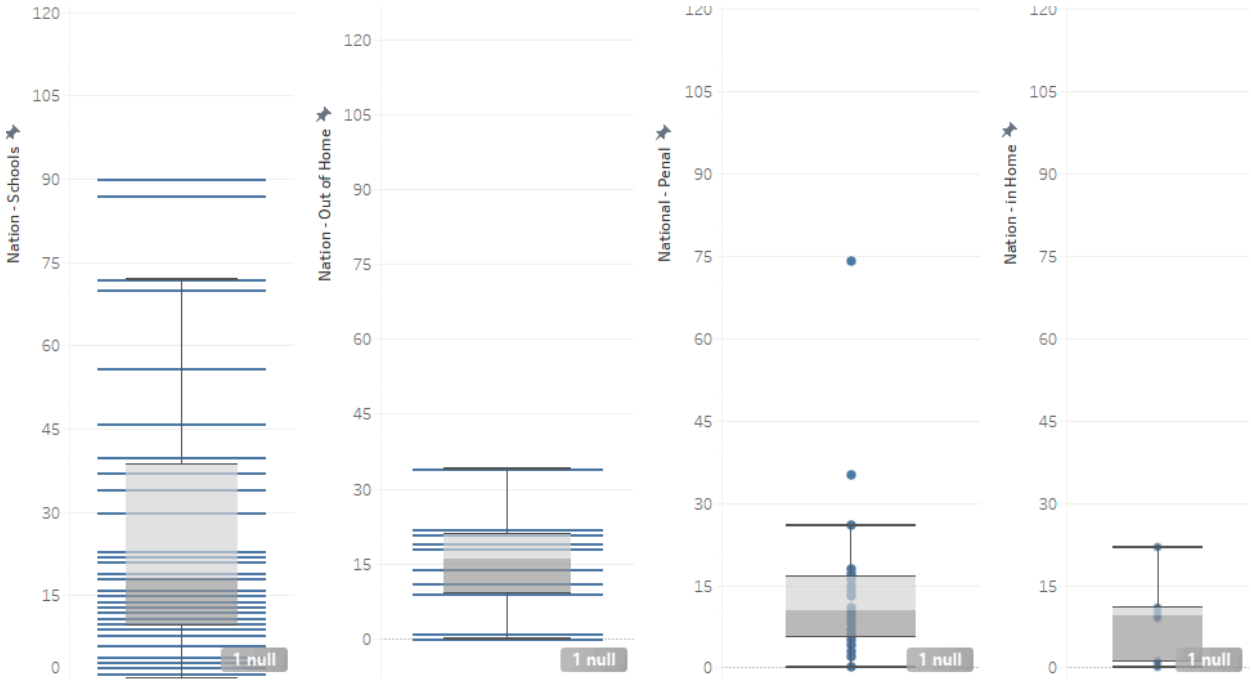
**Tableau**

This tool and power BI is widely used in industries for making dynamic visualizations. The only challenge is you must know how to use these tools otherwise you will waste time understanding it's functions. But honestly with determination anything can be learned, so give it time and you

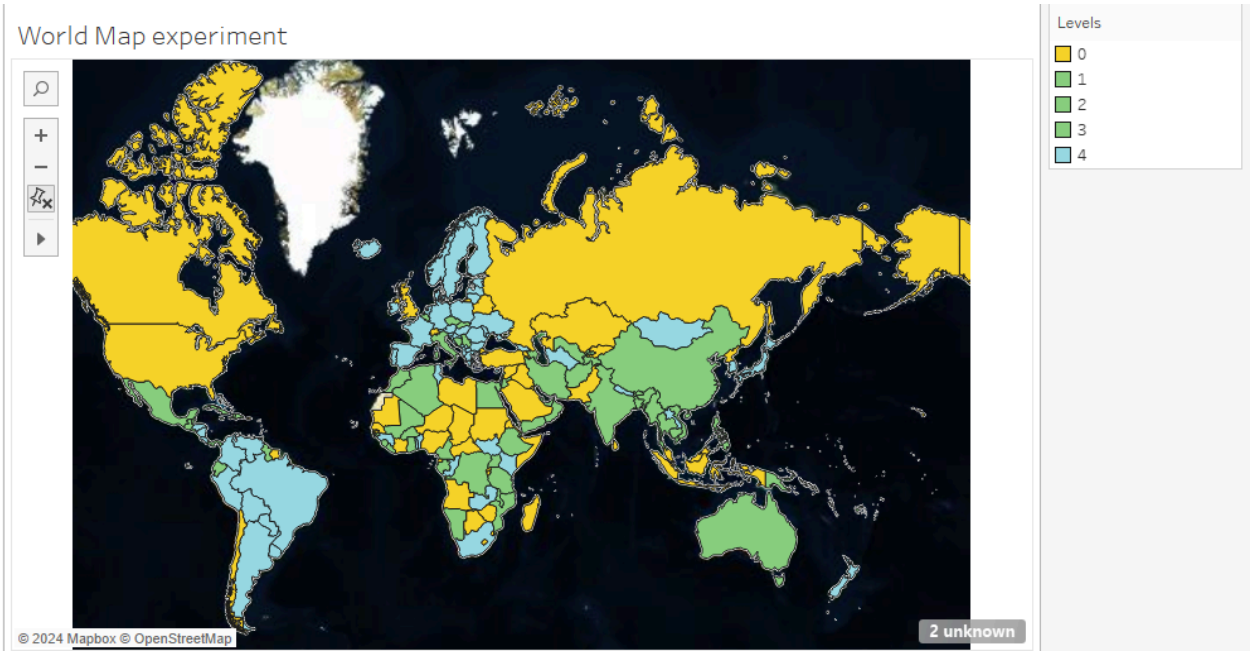
may learn something new. Some of the graphs which initially were made but not used in the report are given below.

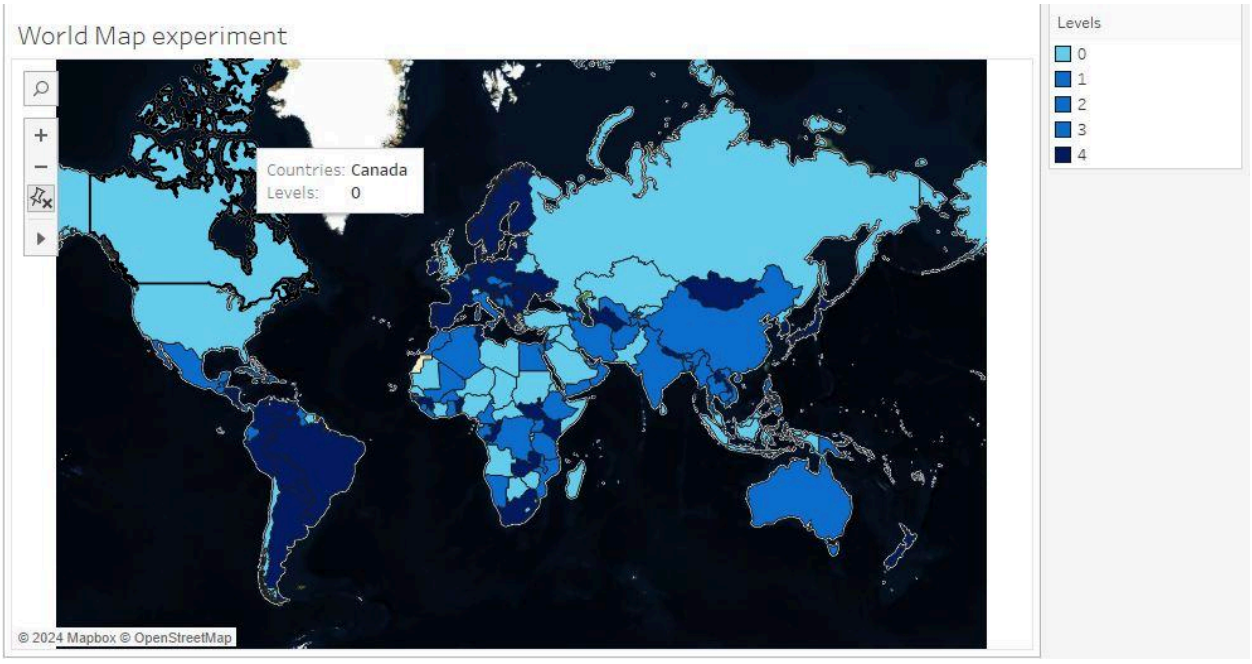
These graphs were made on wrong data because of testing purposes hence results make no sense. These are just to give you ideas on what we can do tableau.



For Example this shows how many times it took for countries to become nationwide. From here we see that countries who passed laws INHomes took almost 9 years to become nationwide (Results are wrong).



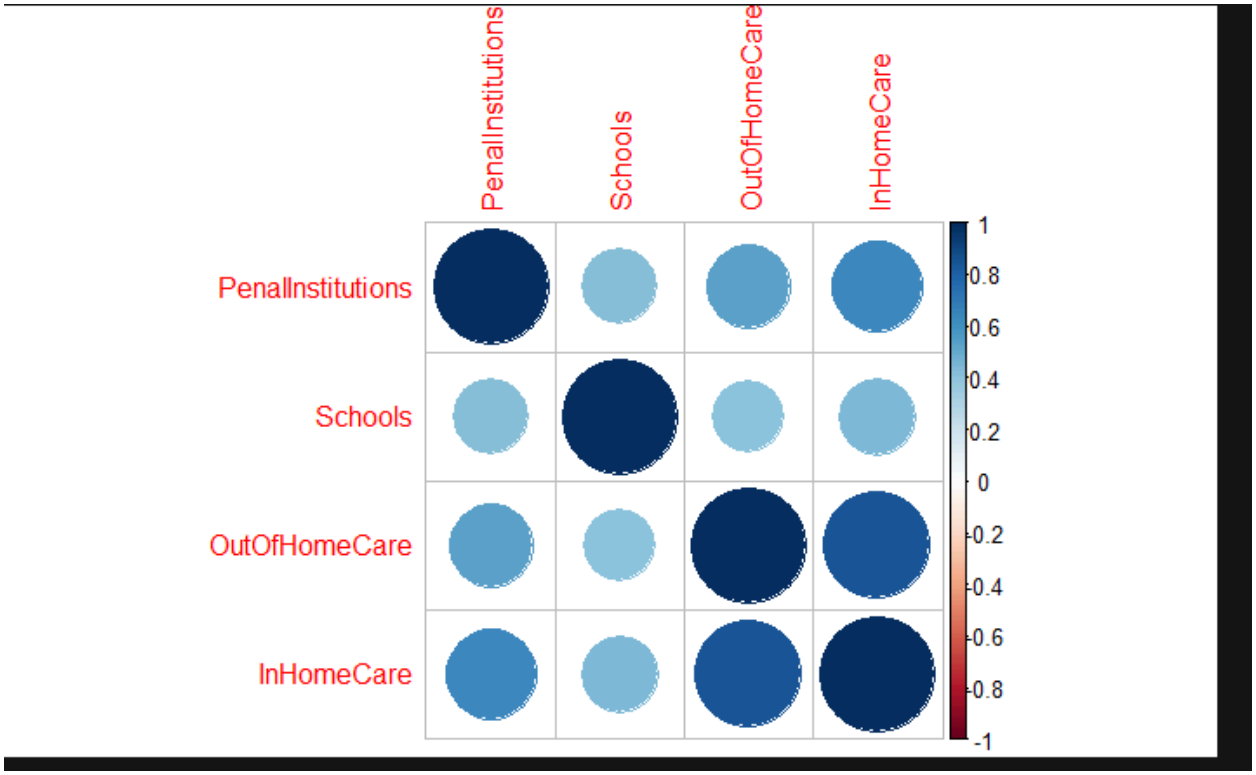




My personal favorite. It was not used in the report.

**R**

This tool is widely for for statistical modeling in industries and research. I will show you how. There are many things which we can do like taking out correlations, Exploratory Data Analysis , regression models and even machine learning.

```
 PenalInstitutions    Schools     OutOfHomeCare    InHomeCare
 Min.   :1933      Min.   :1845   Min.   :1958   Min.   :1979
 1st Qu.:1999      1st Qu.:1996   1st Qu.:2005   1st Qu.:2007
 Median :2007      Median :2007   Median :2012   Median :2013
 Mean   :2006      Mean   :1999   Mean   :2009   Mean   :2011
 3rd Qu.:2014      3rd Qu.:2014   3rd Qu.:2016   3rd Qu.:2016
 Max.   :2023      Max.   :2023   Max.   :2023   Max.   :2023
 NA's   :85        NA's   :91     NA's   :125    NA's   :137
```



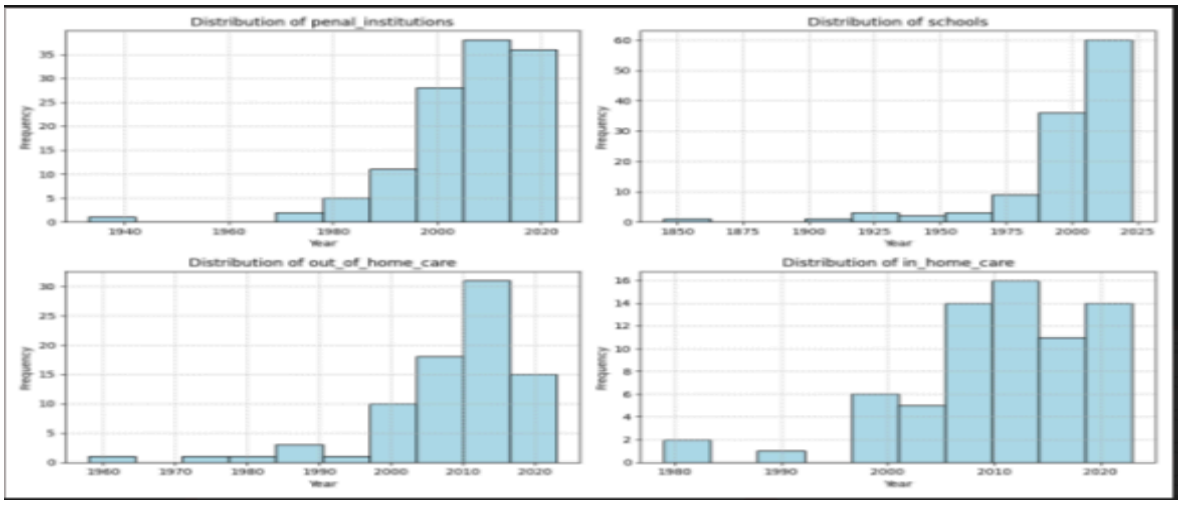

```
                  PenalInstitutions   Schools OutOfHomeCare InHomeCare
PenalInstitutions         1.0000000 0.4282113     0.5334596  0.6329747
Schools                   0.4282113 1.0000000     0.4011317  0.4477254
OutOfHomeCare             0.5334596 0.4011317     1.0000000  0.8486892
InHomeCare                0.6329747 0.4477254     0.8486892  1.0000000
Warning: package 'corrplot' was built under R version 4.3.3
corrplot 0.92 loaded
```

```
Call:
lm(formula = Nationwide ~ `Penal Institutions` + Schools + `Out of Home Care` +
    `In Home care`, data = clean_data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5956 -1.5487 -0.3416  0.4466 16.0442

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         366.12863   88.56172   4.134 0.000107 ***
`Penal Institutions`  0.03559    0.03460   1.029 0.307607
Schools               0.01233    0.01410   0.874 0.385368
`Out of Home Care`    0.17799    0.06354   2.801 0.006753 **
`In Home care`        0.59272    0.08958   6.616 9.27e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.107 on 63 degrees of freedom
Multiple R-squared:  0.8621,    Adjusted R-squared:  0.8534
F-statistic: 98.48 on 4 and 63 DF,  p-value: < 2.2e-16
```

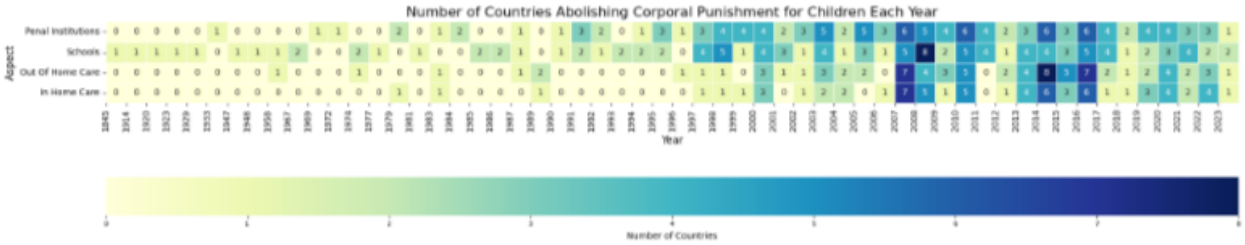
The results are just estimates and please don;t use these as reference for anything. Please do your own EDA and try to make models. Keep in mind the assumptions for making linear,multi-regression models. ( residual analysis, normalization etc.) which you may have studied or are studying in lectures.

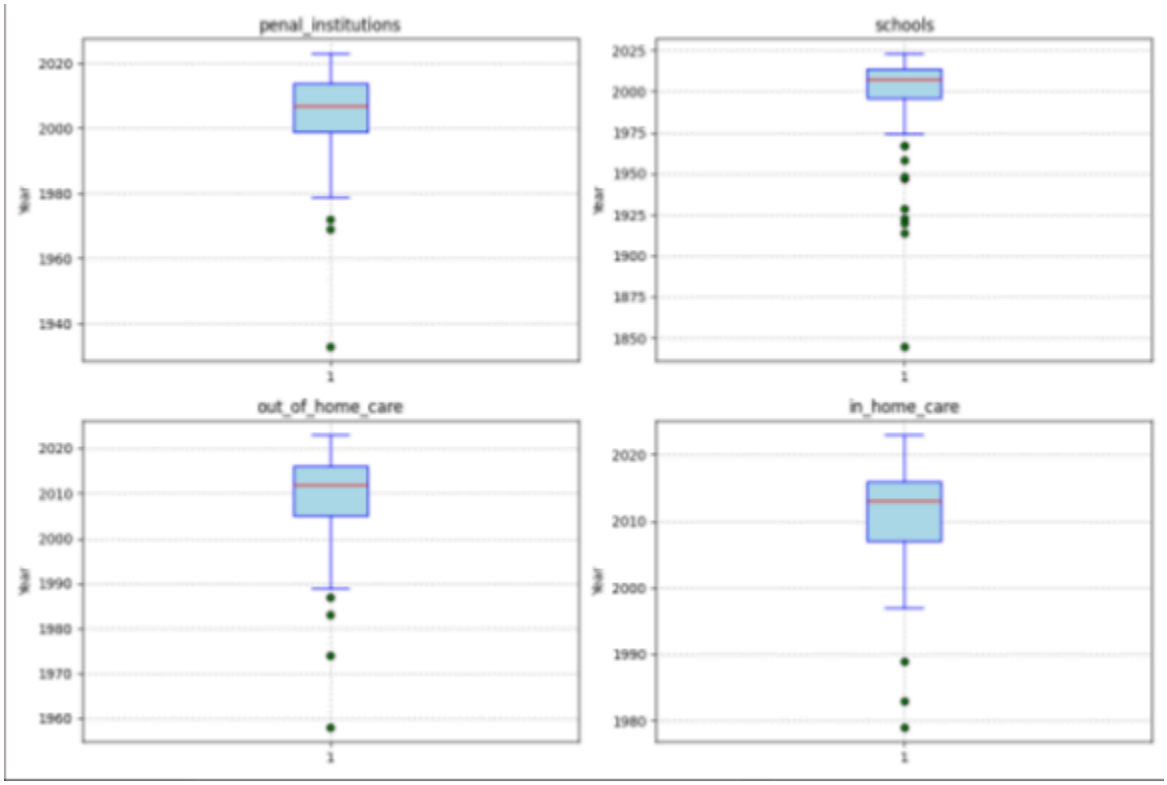
**Python**

Most widely used tool in the industry for IT/Data Science for any work from cybersecurity to data analysis to software development. It is great if you have strong basics in python. For our analysis we only had to make graphs so we did not use it for other purposes. Some of the graphs which were made here are -



| | Min | First Quartile (Q1) | Median | Mean | Third Quartile (Q3) | Max |
|---|---|---|---|---|---|---|
| **penal_institutions** | 1933 | 1999 | 2007 | 2005 | 2014 | 2023 |
| **schools** | 1845 | 1996 | 2007 | 1999 | 2014 | 2023 |
| **out_of_home_care** | 1958 | 2005 | 2012 | 2009 | 2016 | 2023 |
| **in_home_care** | 1979 | 2007 | 2013 | 2010 | 2016 | 2023 |




Some more graphs are used in the report as well.

# Challenges Faced

The major challenge which we faced and took most of the time was data cleaning and interpretation.

Data Science is 70% cleaning and 20% analysis and 10 % sharing. So, make sure you use the correct data for analysis. I audited the old data and with the help of the law team we found it had some errors for legislation years,countries were wrong , missing values. These unexpected things can happen that is why auditing is necessary. It took us almost 30 hours just to clean the raw data form the law team and convert it into working data.

Interpretation was tough. For example nationwide means the country which has covered all the legislation sectors. And our observation is that normally the last legislation to get passed before a country becomes nationwide is In - Homes, so there may be cases where in-home is passed first and rest legislations are passed later. The nationwide year is the latest year in which all legislations are covered and NOT the in-home year necessarily.

Levels denote the number out of 4 which countries have passed the legislations in and not the legislation number. Level 1 means a country has passed a law in any one of the four legislations. You may assign a number to each legislation based on your analysis but then keep it consistent. And let your team-mates and supervisors know so they can understand the data accordingly

Knowledge gaps were common. That's where AI tools come to help. You may use them to learn the process , tools functioning but try to not get too dependent on them. Try to think of your own ideas and ask AI for refinement. Best guide is Professor Garry who will help you understand what you are missing and where you can work upon.

# Future Work

Till now only basic visualizations have been made based on the direct data. Like taking out mean, median etc. I recommend working on working on deeper insights and finding reasons on why the country has / has not passed laws in that field. Use regression , Time-Series Analysis , Machine learning models to understand the behavior of countries and predict when the laws can be passed. Some statistical finding has to be done to get deeper insights. Discuss with your Team-mates, Proff. Gary and supervisor Rebecca. Try to learn what they want to know , give them ideas so you also get a better understanding of what you can work on. Work on different combinations of columns, do research and see if you can come up with something new.

Data Science is all about learning and exploring new things continuously with open mindedness while keeping proper communication with your peers.

All the Best ! : )

Vinny and Henry