**TrueCue Women + Data Hackathon**
**Team 17: Data Femmes**
**Team Members: Katy Oliver, Bharti Kaila, Liz Trumble, Shivali Malhotra**
**Facilitator: Jawwad Adel**

## Introduction

TrueCue Consulting, in conjunction with Alteryx and Concentra Consulting, organized a data hackathon geared towards women Data Analysts and hackers. The aim of the hackathon was to formulate interesting problem statements around Sustainability (coming off the heels of the COP26 Summit).

Our team of 4 professionals and 1 facilitator from TrueCue successfully participated in Round 1 of the Competition and submitted our findings and presentation in the form of a video on 25/11/2021.

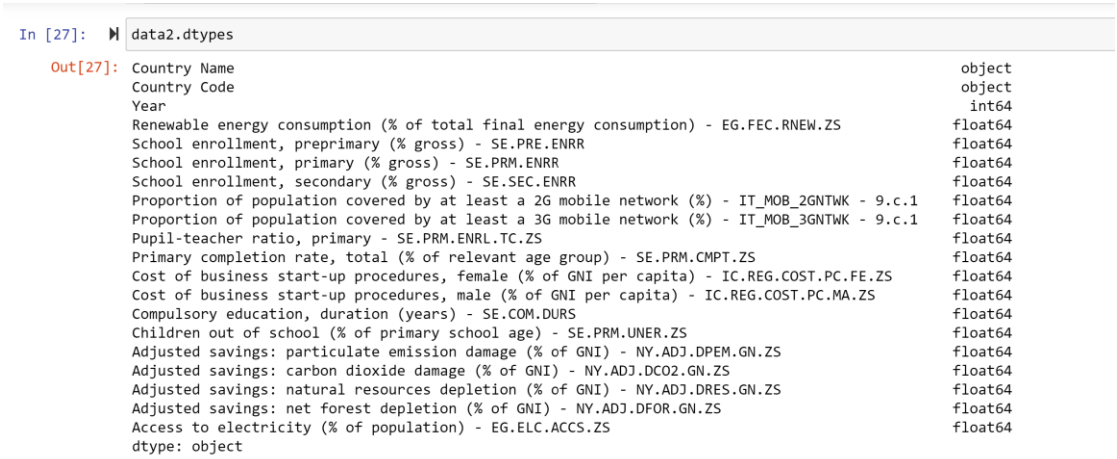
## Our chosen topic for the Hackathon was:

"Does Better and Increased Access To Education, and Education for Women specifically, imply that a country is closer to attaining Sustainable Development Goals?"

## Our Approach to the Hackathon

## Methods and Approach to Data:

1. Using the Sustainability Hackathon Dataset, with variables and demographic data sourced from the UNSD and World Bank Data, we cleaned, processed, and manipulated the data into a usable format with tools such as Python and Alteryx. This dataset contained demographic information for 173 countries over the years 2000-2018. The data original had 51 odd variables, out of which we dropped 32 variables and were left with 19 relevant demographic factors that we wanted to study closely.



```
In [27]: data2.dtypes

Out[27]: Country Name                                                                    object
         Country Code                                                                    object
         Year                                                                             int64
         Renewable energy consumption (% of total final energy consumption) - EG.FEC.RNEW.ZS    float64
         School enrollment, preprimary (% gross) - SE.PRE.ENRR                           float64
         School enrollment, primary (% gross) - SE.PRM.ENRR                              float64
         School enrollment, secondary (% gross) - SE.SEC.ENRR                            float64
         Proportion of population covered by at least a 2G mobile network (%) - IT_MOB_2GNTWK - 9.c.1    float64
         Proportion of population covered by at least a 3G mobile network (%) - IT_MOB_3GNTWK - 9.c.1    float64
         Pupil-teacher ratio, primary - SE.PRM.ENRL.TC.ZS                                float64
         Primary completion rate, total (% of relevant age group) - SE.PRM.CMPT.ZS       float64
         Cost of business start-up procedures, female (% of GNI per capita) - IC.REG.COST.PC.FE.ZS    float64
         Cost of business start-up procedures, male (% of GNI per capita) - IC.REG.COST.PC.MA.ZS      float64
         Compulsory education, duration (years) - SE.COM.DURS                            float64
         Children out of school (% of primary school age) - SE.PRM.UNER.ZS               float64
         Adjusted savings: particulate emission damage (% of GNI) - NY.ADJ.DPEM.GN.ZS    float64
         Adjusted savings: carbon dioxide damage (% of GNI) - NY.ADJ.DCO2.GN.ZS          float64
         Adjusted savings: natural resources depletion (% of GNI) - NY.ADJ.DRES.GN.ZS    float64
         Adjusted savings: net forest depletion (% of GNI) - NY.ADJ.DFOR.GN.ZS           float64
         Access to electricity (% of population) - EG.ELC.ACCS.ZS                        float64
         dtype: object
```

2. We joined this data with EPI or Environmental Performance Index data (the EPI is a score that evaluates a country's environmental performance after taking into account various quantitative and qualitative metrics of sustainability).

3. We imputed missing values in the data through filling missing values with the mean of a column, and also used the MICE package in R for certain columns.

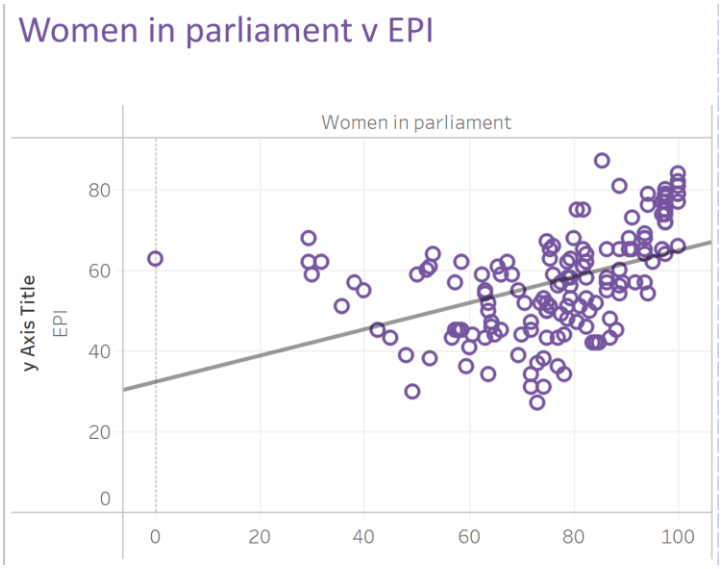
```
In [67]:  for col in missing_cols:
              data2[col] = data2[col].astype(float)
              data2[col] = data2[col].fillna(data2[col].mean())
```

(A snapshot of mean imputation in Python)

4. After our datasets were cleaned, processed, and joined, we imported the new file into Tableau, and built some powerful visuals together
5. We cross-analyzed multiple variables relating to education and employment for women, as well as overall education, to sustainability metrics such as: EPI, adjusted savings from Carbon Depletion, Access to Electricity and Adjusted Savings from Renewable Energy Usage among others.

**Our Findings:**

6. Eventually we decided to visually represent the relationship between the following variables and EPI on a scatterplot:

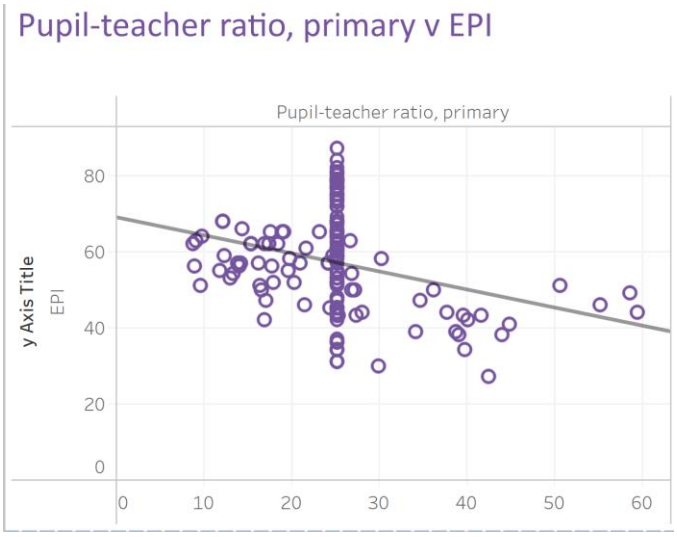



Where we noticed that the higher the number of women in business and law offices (The assumption here is that they are at least Bachelor's Degree Holders or High School Diploma Holders), the better is the environmental performance of that country.

When we analyzed the Duration of Compulsory Education, we found that this had a positive correlation to EPI as well, because the higher the duration of Compulsory Education in a country, the better was its environmental performance.

We similarly noticed that as the ratio of pupils to teachers increased in countries, the environmental performance of those countries also took a hit. (This means that the

number of pupils to every one teacher increased). This could attributed to the fact that there were not enough teachers that could focus on groups of students. In those situations, the quality of education could become compromised.



(Tableau- Screenshot 3)

We ran a regression model for some of the relevant variables:

| | Dependent variable: |
|---|---|
| | EPI2 |
| pupilteacher | -0.033** |
| | (0.014) |
| compulsedu | 0.159*** |
| | (0.059) |
| Electricity | 0.037*** |
| | (0.006) |
| womenlaw | 0.039*** |
| | (0.006) |
| Constant | 52.103*** |
| | (0.887) |
| Observations | 3,287 |
| R2 | 0.080 |
| Adjusted R2 | 0.079 |
| Residual Std. Error | 6.586 (df = 3282) |
| F Statistic | 71.490*** (df = 4; 3282) |
| Note: | *p<0.1; **p<0.05; ***p<0.01 |

**(Stargazer Table in R)**

The Women in law and business variable had a lower weightage compared to the others, however, we notice that for countries like Cuba, where there are 0 women in law, this could also be due to non-availability of data, they still had great environmental performance score of 60. The model is not as well fitted here as it could be, but we could add a quadratic variable to better, fit it, which may not be useful in the long term, because that may not necessarily mean that the variables are statistically significant. All the variables are statistically significant at the 1% significance level, other than the pupils teacher ratio which is significant at the % significant level.

```
data:  model1
BP = 58.996, df = 4, p-value = 4.715e-12

pupilteacher   compulsedu  Electricity     womenlaw
    2.111028     1.178892     2.308466     1.091814
```

On checking for multi-collinearity in the model, we see that the VIF is <10 for all the major variables, which means that the variables are not majorly correlated with each other. However, there could still be an omitted variable bias, as we see some heteroscedasticity in the model. This simply implies the model is not as well fitted, as the estimated values are not close to the true population parameters. We need to explore further. However, at a high level, this visual provides us with good insights and shows us that these variables are statistically significant.
The pupil-teacher ratio variable negatively impacts the model, as we had stipulated before.

As for the outlier: Cuba, we found  data compiled by the Inter-Parliamentary Union on the basis of information provided by National Parliaments by 1st February 2019, that says that in 2019, Cuba had 53% female representation in the lower house of parliament. This may not have been captured in our current dataset.

Ultimately, running a regression model for the purposes of inferencing or prediction may have been easier if we had used cross-sectional data as opposed to panel data.
Cross-Sectional data comprises many observations at the same point of time whereas, Panel data consists of the number of variables and of multiple time periods.

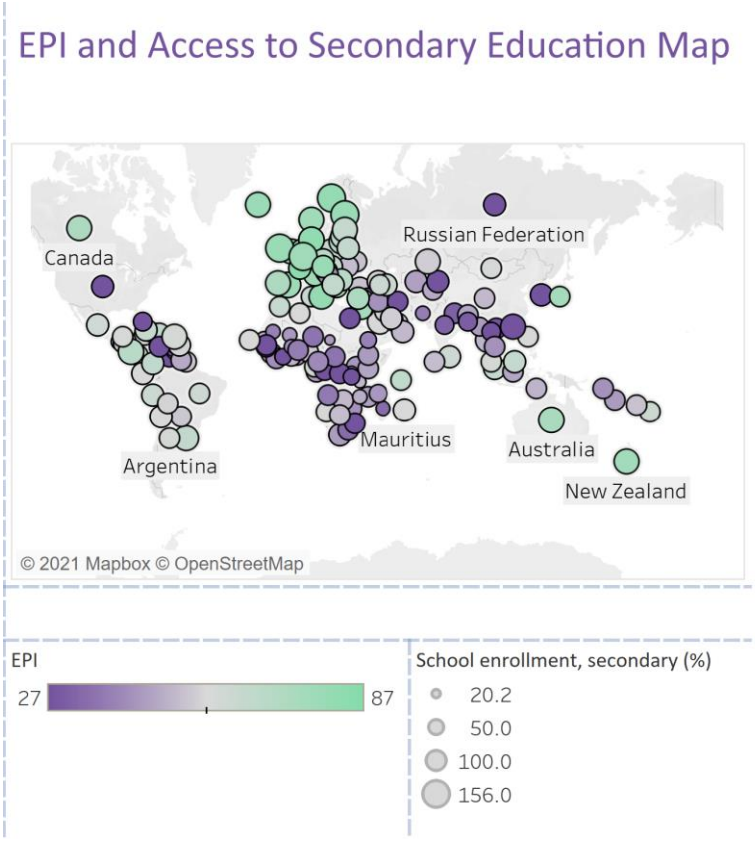
Using R, we built a regression model on EPI and:

- Pupil-Teacher Ratio,
- Compulsory Education,
- Access to Electricity, and
- Women in Law and Business

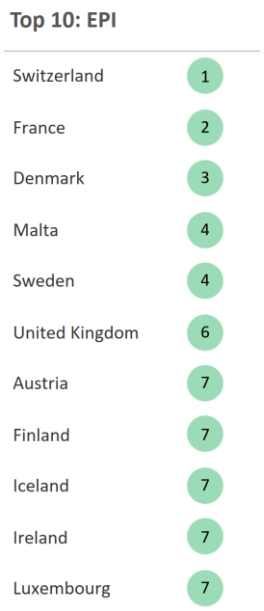
We concluded the following:

- Parameters are individually statistically significant at the 1% significance level (pupil-teacher ratio significant at 5%)
- There is slight variation in the error terms, resulting in lower R squared. (**Recommendation-** could add quadratic term for better fit, treat outliers, or explore other relevant variables)
- There may be omitted variables that could provide a better fit, however, at a glance, there was a definite scope to improve this model to estimate the relationship between EPI and Education + Gender)
- We concluded that it would be better to use cross-sectional data (data at a point in time) vs panel data (data over a period of time) for this kind of inferencing.

**World Map Visual:**




Our next visual mapped and detailed the relationship between the Variables: **% of Students enrolled in Secondary Education and EPI.** We found again that countries with higher enrollment in High School and presumably in Higher Education Institutions, had better environmental performance, with the following countries being in the top and bottom ten performers:





**Conclusion:**

These were some of our key findings in the data, although a lot more information is yet to be gleaned.
Our recommendations after this analysis were as follows:

Education at all levels can shape the world of tomorrow, equipping individuals and societies with the skills, perspectives, knowledge, and values to live and work in a sustainable manner. It directly affects sustainability plans in the following areas:

1. Implementation

2. Decision-making

3. Quality of life

Education for Sustainable Development, therefore, is focused on giving people knowledge and skills for lifelong learning to help them find new solutions to their environmental, economic, and social issues.
We would therefore ask our governments to place education as a priority in both policy and practice. Additionally, lobbying for reforms towards providing free or subsidized school education of good quality to all, would be a good idea (especially in developing countries, policy catering to vulnerable or marginalized groups would be beneficial)

Another aspect would be to invest in Educational Training programs, to improve the quality of teaching and instruction in many countries, and to make online access to educational tools easier for communities that are underserved. Finally, investing in Educating Women has a direct impact on a country's environmental health, so why not start now?