Final Report: UNESCO Data Analysis

Team Rocket: Shubhi Agarwal, Aaron Mardis, Hasan Qadri, Rhiya Sharma, Priya Narayanan, Jesse Huang

Introduction

UNdata was launched as part of a project in 2005, called "Statistics as a Public Good", whose objective was to provide free access to global statistics, to educate users about the importance of statistics for evidence-based policy and decision-making. In this project, we focus specifically on the UNESCO dataset. There are multiple tools available to visualize some parts of this data, but no integrated place for a user to view and understand the trends in the overall dataset. Our aim is to provide such a framework that gives ease of handling and enough flexibility for users to answer the questions they are interested in. The project will help users make sense of the extensive UN Data quickly and effectively. It may also help policy research groups and advisors to back their studies by data without spending hours in extracting the relevant information. We will try to measure our success by conducting user studies. Feedback on questions like if they learned something new or educational that can help them make informed opinions, would be used as a benchmark for success.

Problem definition

Entire UNESCO data is currently available in the form of raw tables. Focussed case studies determine the trends for specific use-cases. We want to provide an integrated data analysis framework for the UNESCO data, with the following objectives:

1. Extract meaningful trends and insights on each indicator.
2. Ability to generate dynamically created trends between user given metrics.

Survey

There are multiple case studies sponsored by civil organizations each year, that analyze UNESCO data to monitor the implementation and impact of policies, like the Global Report on Reshaping Cultural Policies [1], or to measure the progress towards sustainable development goals, like that done in studies to analyze education access for the disabled [2] and role of education in the eradication of poverty [3]. Case studies in Australia [4] and India [5,6] have shown analysis of literacy levels directly lead to changes in the government budget and establishment of extensive programs. Another case study investigates the differences in labor productivity growth in Arabian Gulf countries in comparison to other developing or developed countries [16]. These studies help us to understand the ways in which policy researchers use and analyse data and the common patterns of the questions that they need answers for. We can use this information to build our data models and structures for faster searches.

Gapminder [7] created a powerful tool - Trendalyzer, to analyse and display world development data. We can derive extensively from their presentation techniques and build upon them for the specific case of UNESCO data. Gapminder gives the ability to see global correlations between various indicators that have been collected by combining multiple data sources. For our specific case, we want to go one step further by providing the users more flexibility with the questions they want to answer, and also the ability to drill-down within one country at a time, which is not something that gapminder currently facilitates.

Data sets facilitate improved decision making and richer analytics. However, the quality of data seems to be a major concern which if treated poorly can lead to incorrect decisions and unreliable analysis. It is important to structure the available data [8]. Gibson's paper discusses an approach at clustering data based on categorical co-occurrence and association rules, patterns we as developers can follow when implementing our own visualizations. The model falls short of providing a step for evaluating a data vis system, there is a step from schema to hypothesis, but that is a large jump [9]. One direction we are thinking

about exploring is finding relationships between different factors in culture/economics/demographics across countries using a self-organizing map which is an unsupervised neural network used for clustering. This kind of hierarchical map requires less training time though can struggle to form isolated clusters, which is slightly problematic for our use case [10].

Since the target users are policy makers, they will likely want to extrapolate from the UNESCO data to predict, we want to operate with the Visualization Information Seeking Mantra: "Overview first, Zoom and filter, then details on demand" as mentioned by Perer and Shneiderman [11] and borrow from the four nested level of predictive visual analytics of J. Lu et. al [12]. Users should explore predictive analytics which takes what *has* happened to produce knowledge about what *will* happen with statistically valid reasons for sensemaking [13].

The notational model of sensemaking for intelligence analysis describes the process from converting extracted data to resolving hypotheses and presenting findings. The way our activity plan is structured is inspired by the notional model. The model falls short of providing a step for evaluating a data visualization system based on insight [14]. Users are enabled to see things they were not aware of, the insight helps them define new questions, hypotheses and models of the data. Great importance is placed on evaluation and it might be difficult to narrow down on an effective visualization strategy [15].

In order to facilitate the notational model of sensemaking, our visualization must enable the user to easily categorize the data that we present to them. An approach is to depict multi-value data. There are multiple approaches, including a parametric approach, a shape descriptor approach, and an operator approach [17]. Lastly, in order to paint a clear picture, we follow the steps of comparing categories, assessing hierarchies to evaluate relationships, showing changes over time, plotting connections and relationships, and then mapping geospatial data [18]. In the end, users should be able to visualize stories with the representations we provide.

Proposed method

Data Collection:

Initially, we believed that we would need an API for our data collection. However, after reaching out to UNESCO Institute for Statistics (UIS) co-ordinators, we were able to obtain bulk data downloads for each archive (Education, Science, Culture, and External) which greatly decreased our collection effort. The archive contained a link to 10 zip files which stored the datasets as well as a README file explaining how the data is organized.

UNESCO data that we are currently using belongs to the categories of SDG4 (Sustainable Development Goal 4), Demographic, Education and Equity datasets, ranging across countries approximately for the years 1970- 2018. These datasets will be useful in our correlation studies.

Visualization:

At the top of our UI, the user sees a drop-down selector where they can choose the indicator and year for which they want to see the data. We have then split this visualization into two major components:

1. Global distribution of the chosen indicator depicted using a choropleth world map.
    a. Clicking on any country in the world map, will result in a bar graph for countries belonging to the same region as the country that was clicked. It helps in understanding the regional details of an indicator. On the other hand, this chart also shows the gender-based distinction between values, for whichever indicators such distinction is applicable.
    b. Another sub-component is a trendline graph which appears on clicking on a bar in the regional chart view. This graph shows the change in indicator value for the selected country

over time. Better interpretation of individual variables like educational attainment of each country over the years can be made based on the level of fluctuations.

2. Correlation graph between multiple indicators.
    a. The graph can be adjusted based on the user's choice of target variable. The aim is to show the correlations between input metrics like government funding and household expenditure on education and output indicators like Mean Literacy Rate or even GDP per capita. This graph provides an overview of the variables that most significantly affect or correlate with the target variable. The correlations have been calculated using Pearson correlation metric on the common data points between every two indicators. Common data points refer to the combinations of country and year that have data for both the indicators.
    b. Clicking on one of these input variables gives a further drilldown of the trend between the chosen and target variable for all countries (that have valid data for the chosen metric) for the selected year. This is represented in the form of a bubble chart inspired by [7].

To make the most of our data, we have stored it using a Firebase Cloud Firestore Database. We chose this database because it was advertised as the best option for our data since it allows for data to be organized at scale, offers faster return on queries than many other alternatives, and can make manipulations whether on- or offline. If we were to scale this as intended for our target users, we would, of course, need to pay for this service as we found that, due to having datasets over 2 million rows with some larger than 6 million, the daily allotment of 50,000 calls was too low of a threshold for testing and would not be sustainable for our target users.. However, we were still able to take advantage of performing computations under the payment threshold, so we did not need to spend any money for this project.

List of Innovations

1. **World map view:** Choropleth map of the world based on the selected factor from the UNESCO data. Gives clear insight into how the world is divided based on any chosen UNESCO variable.
2. **Regional comparisons:** This view compares countries in a region based on a given UNESCO variable in a bar chart form.
3. **Country view:** This is the most specific drill down in our system and gives how for each country, how some UNESCO variable has changed over time
4. **Correlation graph:** The aim is to show the correlations between input metrics like government funding and household expenditure on education and output indicators like Mean Literacy Rate or even GDP per capita. This graph provides an overview of the variables that most significantly affect or correlate with the target variable.

Experiments/ Evaluation

Questions we hoped to answer through experiments:

1. How well does our visual system allow flexibility and data navigation?

2. Does it make it easier to answer data related questions we encountered in different case studies?
3. Does it make a policy researcher's job faster?
4. Does it provide useful insights in considerably less time?

We have tried to measure our success primarily by conducting user studies. Feedback on questions like if they learned something new or educational that can help them make informed opinions, is being used as a benchmark for success. In our user studies, we have collected responses from 40 individuals, including students and professionals working in varied fields. We gave all users 10-15 minutes to use our tool and then the following questionnaire for feedback on their experience. The aggregated responses to each question have been tabulated below in the Survey Results table.

Answer the following questions on a scale of 1 to 5, with 1 being the least agreeable and 5 being the most agreeable:

1. The website clearly conveys the distribution of any UNESCO data indicators across the globe.
2. It clearly shows the correlation of other indicators with your selected indicator.
3. The style of presentation is helpful in understanding the global picture of progress in education.
4. In your opinion, correlation is a good metric to see interdependence between indicators. It helps to understand which factors most/least affect your target indicator.
5. For any given metric, you can clearly understand gender equity.
6. The drill-down options are sufficient to look at the low-level data.
7. You think this would be a helpful tool to begin any research based on UNESCO data.
8. You gained/expect to gain new/interesting insights from using this interface.
9. The interface is missing some information/analysis that you would have liked to see.
10. If you selected "Yes" in the above question, please elaborate on some other features that you have liked to see.

Survey Results :

Question Number	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)
1	0	0	2.5	35	62.5
2	0	5	5	45	45
3	0	0	5	50	45
4	0	0	5	35	60
5	0	7.5	2.5	55	35
6	0	7.5	10	40	42.5
7	0	5	5	40	50
8	0	2.5	2.5	30	65

9	77.5% users voted “No”

It appears that of our participants, no one felt that they strongly disagreed with any of our questions posed. But from what we see from the table above, questions two and five through seven garnered the strongest negative responses. These four questions focused on navigability of our tool and the effectiveness by which correlation and gender equity is communicated. It appears that a significant number of participants felt that the drill-down options could have been improved more in some form or fashion and that the labels for gender equity and correlation could have been better. There were some indicators that simply didn't have ‘gendered’ data and that may have led to some confusion amongst participants who were expecting to see that on the drill-down but did not. On the other hand, our two questions with the highest agreeability were one and eight, which respectively asked if the website effectively conveyed the distribution of any selected indicator across the globe and if they expect to derive insight from the use of our application. The latter question in particular is perhaps one of the most important asked in our survey for it is from insight that one gains new knowledge and can answer an hypothesis.

The final question on the survey (excluded in the table above) asked what could be improved upon for the interface, and we got suggestions ranging from including more tooltips to encourage drill-down to some users feeling the interface felt a little too cluttered and overwhelming. In response to this feedback, we added certain features like an introductory explanation at the top and changing the cursor to hand pointer to indicate clickability, in order to improve the navigation experience.

We also try to test the efficacy of our tool by a qualitative analysis of the different trends to see if they make realistic sense. For example, when we observe the correlation chart for GDP/capita, we see that it has the highest levels of interdependence with Educational attainment rate for Bachelor's or higher and Mean years of schooling for the population of 25+ years. This makes sense because the progress in education can reflect in GDP only when more students go upto higher education and qualify for professional jobs. Similarly, it makes sense for the Youth literacy rate for the population of 15-24 years to be highly correlated with Completion rate of primary and secondary education because those are the variables directly impacting the Youth Literacy Rate.

Conclusions and Discussion

In conclusion, we can use the experiment results to answer our testbed questions as follows:

1. How well does our visual system allow flexibility and data navigation?
   More than 80% of the users agree that the interface clearly conveys the distribution of indicators across the globe and that the drill-down options are sufficient to look at the low-level data. Some changes as mentioned in Experiments section, have been made to improve navigation experience after user feedback.

2. Does it make it easier to answer data related questions we encountered in different case studies?
   Our qualitative analysis shows that the tool makes it possible and easier to answer such questions.

3. Does it make a policy researcher’s job faster?
   90% of our users agreed that this would be the case, with 50% in strong agreement.

4. Does it provide useful insights in considerably less time?

95% users acknowledged that they gained new or interesting insights, with 65% in strong agreement.

There is still a lot of scope for improvement and expansion in the tool. We see that UNESCO data is incomplete for some countries and years for many indicators. We can integrate data from other sources and even other UN websites like WHO, UNICEF to make it more complete and extensive. There is also scope for adding features like user-generated indicators, and predictions based on past data using Machine Learning models.

Work Distribution: All team members have contributed a similar amount of effort.

Plan of Activities	**People**	**Time**	**Date of Completion**
Ideas & Proposal	All*	30h	02/24/2020
Data collection & Cloud Setup	Aaron and Priya	30h	02/29/2020
Data cleaning & Categorization	Aaron and Jesse	30h	03/19/2020
Analysis	Shubhi and Rhiya	30h	03/28/2020
Viz Techniques	Hasan, Priya, Rhiya, Shubhi	15h	04/04/2020
UI Development	Hasan and Shubhi	20h	04/11/2020
Evaluation	All	20h	04/18/2020
Report	All	15h	04/20/2020

UNESCO exploration tool can be accessed : https://unesco-interface.herokuapp.com/home.html
Survey Responses :
https://docs.google.com/forms/d/1DYeCPSyVsc9ilAXjgcO0QgTxFkHz6M7qxWy2XxrQ5IQ/edit#responses

Bibliography

1. Azoulay, A (2017), Reshaping cultural policies: advancing creativity for development. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000260592.
2. UIS. Education and Disability: Analysis of Data from 49 Countries. Available at: http://uis.unesco.org/sites/default/files/documents/ip49-education-disability-2018-en.pdf
3. UNESCO Institute for Statistics (UIS), and Global Education Monitoring Report (GEMR) (2017). “Reducing Global Poverty through Universal Primary and Secondary Education.” Policy paper 32/Fact sheet 44. Montreal and Paris: UIS and GEMR. Available at: http://unesdoc.unesco.org/images/0025/002503/250392e.pdf
4. U. Hanemann (Ed.). Last update: 26 July 2017. The Australian Language, Literacy and Numeracy Programme (LLNP), Australia. UNESCO Institute for Lifelong Learning.
5. Chatzigianni, Sofia. Tata Consultancy Services’ Adult Literacy Programme: Computer-Based Functional Literacy, India. UNESCO Institute for Lifelong Learning.
6. Learning and earning: Evidence from a randomized evaluation in India https://doi.org/10.1016/j.labeco.2016.11.007
7. LeBlanc, D. (2012), Gapminder: Using a World of Human Ecology Data to Teach Students Critical Thinking Skills. The Bulletin of the Ecological Society of America, 93: 358-372. doi:10.1890/0012-9623-93.4.358
8. Xu Chu, Ihab F. Ilyas, Sanjay Krishnan, and Jiannan Wang. 2016. Data Cleaning: Overview and Emerging Challenges. In Proceedings of the 2016 International Conference on Management of Data (SIGMOD ’16).

Association for Computing Machinery, New York, NY, USA, 2201–2206. DOI:https://doi.org/10.1145/2882903.2912574

9. Gibson, D., Kleinberg, J. & Raghavan, P. Clustering categorical data: an approach based on dynamical systems. The VLDB Journal 8, 222–236 (2000). https://doi.org/10.1007/s007780050005
10. U. Hanemann (Ed.). Last update: 26 July 2017. The Australian Language, Literacy and Numeracy Programme (LLNP), Australia. UNESCO Institute for Lifelong Learning.
11. Adam Perer and Ben Shneiderman. 2008. Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08). Association for Computing Machinery, New York, NY, USA, 265–274. DOI: https://doi.org/10.1145/1357054.1357101
12. Lu, J., Chen, W., Ma, Y. et al. Recent progress and trends in predictive visual analytics. Front. Comput. Sci. 11, 192–207 (2017). https://doi.org/10.1007/s11704-016-6028-y
13. Lu, Y., Garcia, R., Hansen, B., Gleicher, M., & Maciejewski, R. (2017). The State-of-the-Art in Predictive Visual Analytics. Computer Graphics Forum, 36(3), 539-562. https://doi.org/10.1111/cgf.13210
14. Pirolli, Peter, and Stuart Card. "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis." Proceedings of International Conference on Intelligence Analysis. Vol. 5. McLean, VA: Mitre, 2005.
15. Van Wijk, J.J (2005). The value of visualization. IEEE Visualization, 79–86.
16. Al Raee, Mueid & Ritzen, Jo & Crombrugghe, Denis de, 2017. "Innovation policy & labour productivity growth: Education, research & development, government effectiveness and business policy," MERIT Working Papers 019, United Nations University - Maastricht Economic and Social Research Institute on Innovation and Technology (MERIT).
17. A. L. Love, A. Pang and D. L. Kao, "Visualizing spatial multivalue data," in IEEE Computer Graphics and Applications, vol. 25, no. 3, pp. 69-79, May-June 2005.
18. "Taxonomy of Data Visualization Methods." Data Visualization: a Successful Design Process, by Andy Kirk, CreateSpace Independent Publishing Platform, 2015.