Education Statistics Visualization

Process Book

30th November 2018

Greeshma Mahadeva Prasad, u1141804, greeshma@cs.utah.edu
Tanvi Gangadhar, u1205740, gangta@cs.utah.edu

Project Repository-
https://github.com/magsheer/education-statistics-visualization
Visualization Link-
https://magsheer.github.io/education-statistics-visualization/

Project Proposal

BACKGROUND AND MOTIVATION

When we started looking for project ideas, we came across many that seemed interesting but these did not have good datasets, for example, visualizing black friday shopping trends across various age groups and cities. We could not find an appropriate dataset for it.

Later we started looking for datasets that seemed interesting and we came across the Education statistics dataset by World Bank. This dataset has indicators that describe literacy rate, enrolment rate, illiterate population across education levels (primary/ secondary/ tertiary), age groups and gender. We decided to use this dataset as education can be associated with various parameters such as a country's economic growth, crime rate etc. We also wanted to draw parallels between the economic state of the country (low income vs high income) and education rate to understand the factors influencing the latter over the years.

PROJECT OBJECTIVES

We have wish to try and answer a set of questions as a part of our analysis through our data visualization. The primary questions being:

- How indicators such as literacy rate among males/females, illiterate population etc, have changed in countries over the last five decades
- Which are the most literate/illiterate age groups
- What's the highest level of education in specific countries
- Does the income of the country impact its literacy rate

As mentioned earlier, education can be related to various factors that influence our society. Visualizing such a dataset might help analyze the impact education has, on the growth of a country/region. In more developing countries, it would be beneficial to see the gender parity in education and whether it is moving towards equality in education.

DATA & DATA PROCESSING

We got out dataset from https://datacatalog.worldbank.org/dataset/education-statistics, The World Bank's Data Catalog.

The dataset includes multiple files and is quite large. It consists of over 4,000 internationally comparable indicators that describe education access, progression, completion, literacy, teachers, population, and expenditures. We do not intend to use all of them for the purpose of this project as this might cause a lag in the webpage. We plan to use indicators that fall under the three essential categories: gender, age group and education level. This will allow us to effectively visualize the Gender Parity Index, highest level of education obtained across regions/countries and drop out rate amongst different age groups with respect to the economic state of the country. We also intend to show a global trend by comparing country statistics with the global data.

Since the data we want to use for this visualization is scattered over multiple files, we need to extract only the data we require depending on the indicators, region and countries. We plan on doing this using Python or Javascript.

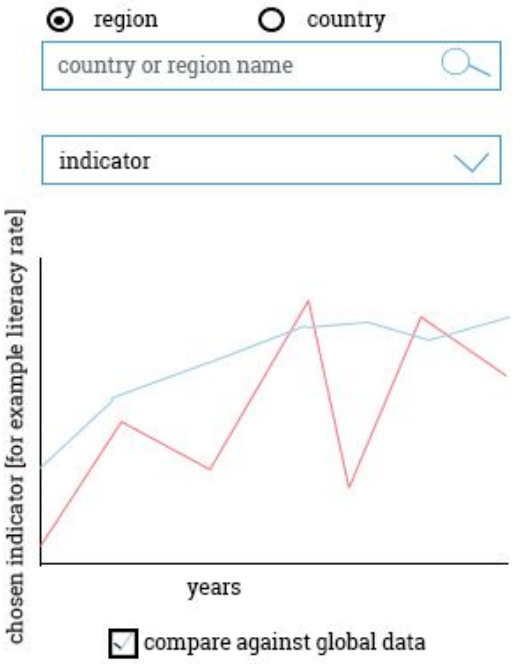
VISUALIZATION DESIGN

Prototype 1

We used some ideas from the assignment 4 for our first prototype. The features include:

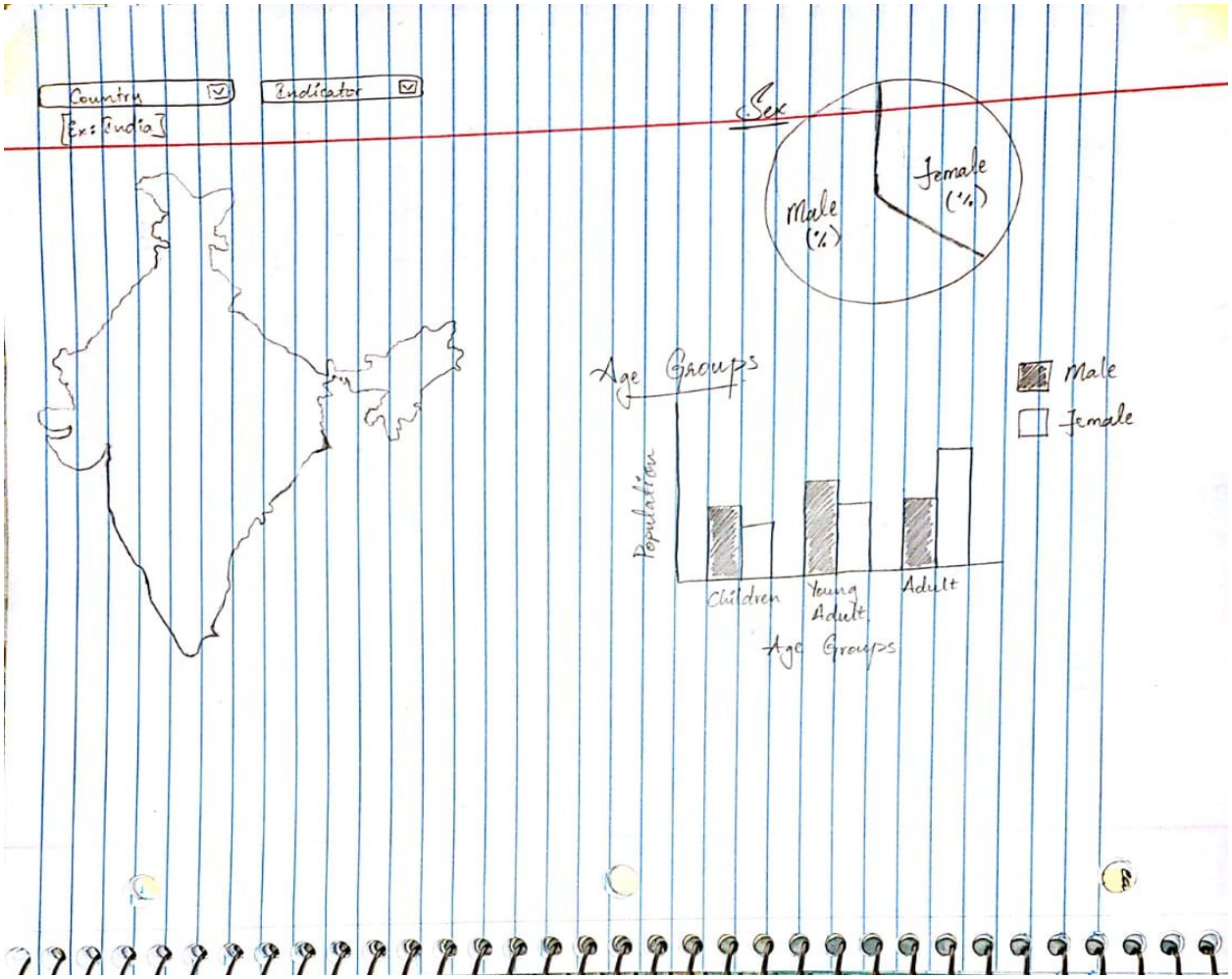
Alternative 1:

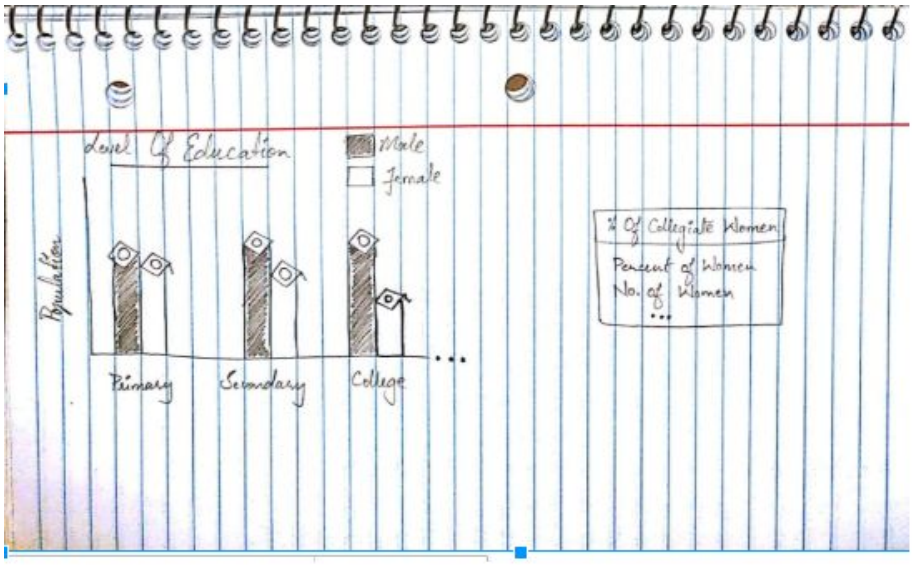
- Select whether a region or country has to be visualized using the radio buttons
- Select the particular country/region using the map or typing it in
- Select the indicator to observe the trend over years
- Check the compare option to compare the global and country/region specific trends.



Alternative 2:

- Select the country you want visualize the data for
- Select a single indicator from the provided dropdown
- Based on the country and indicator selected, the visualization will show, the trend for the particular indicator for last five decades.
- For example, choosing India as the country and sex as the indicator, we show a pie chart with the ratios since there are only two values whereas selecting India as the country and education level as the indicator, will show adjacent bar charts.
- On clicking the bars or sections, details will be shown

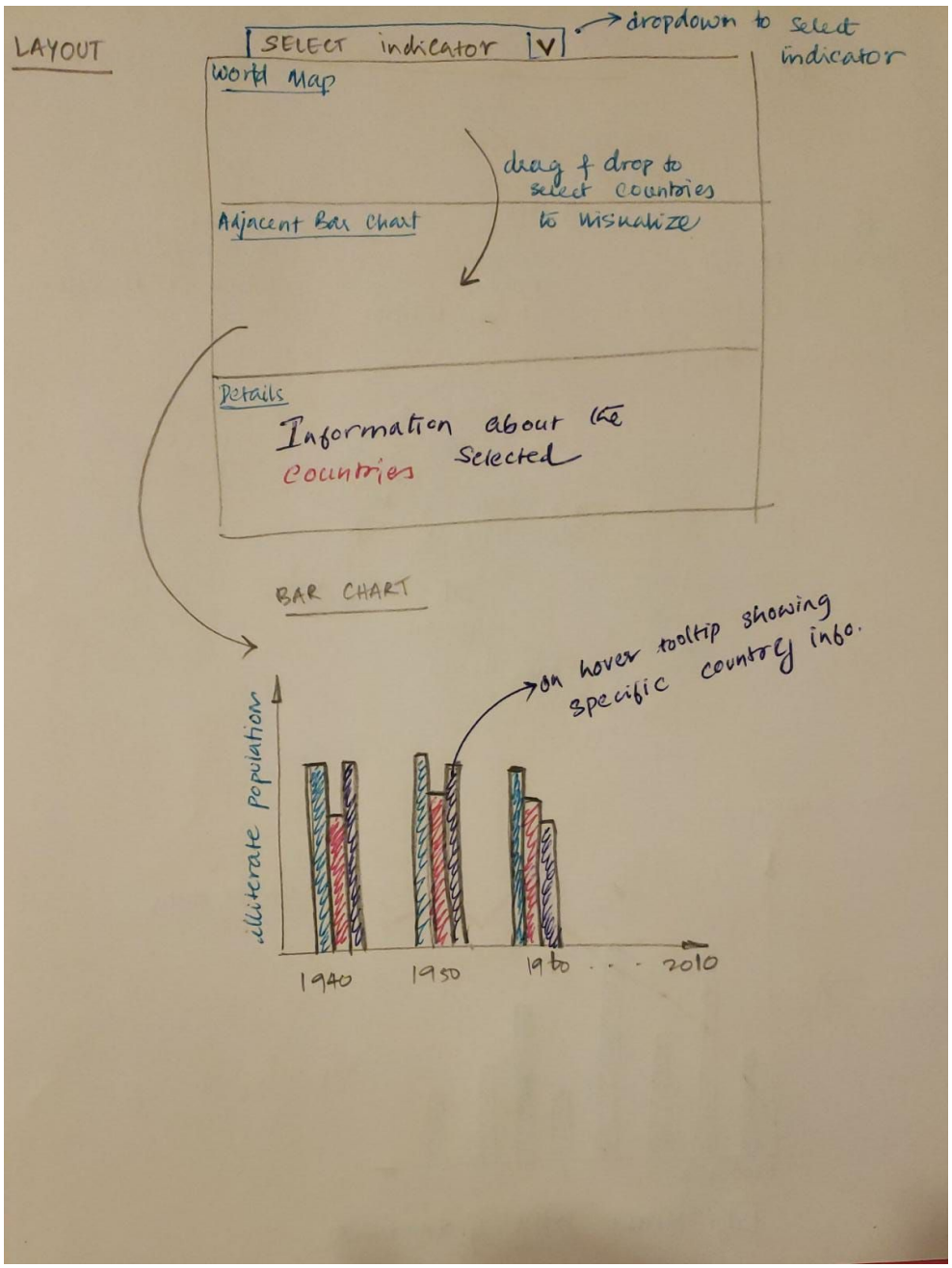

Country
[Ex: India]
Indicator
Sex
Male (%)
Female (%)
Age Groups
Male
Female
Population
Children
Young Adult
Adult
Age Groups


Level Of Education
Male
Female
Population
Primary
Secondary
College
% Of Collegiate Women
Percent of Women
No. of Women

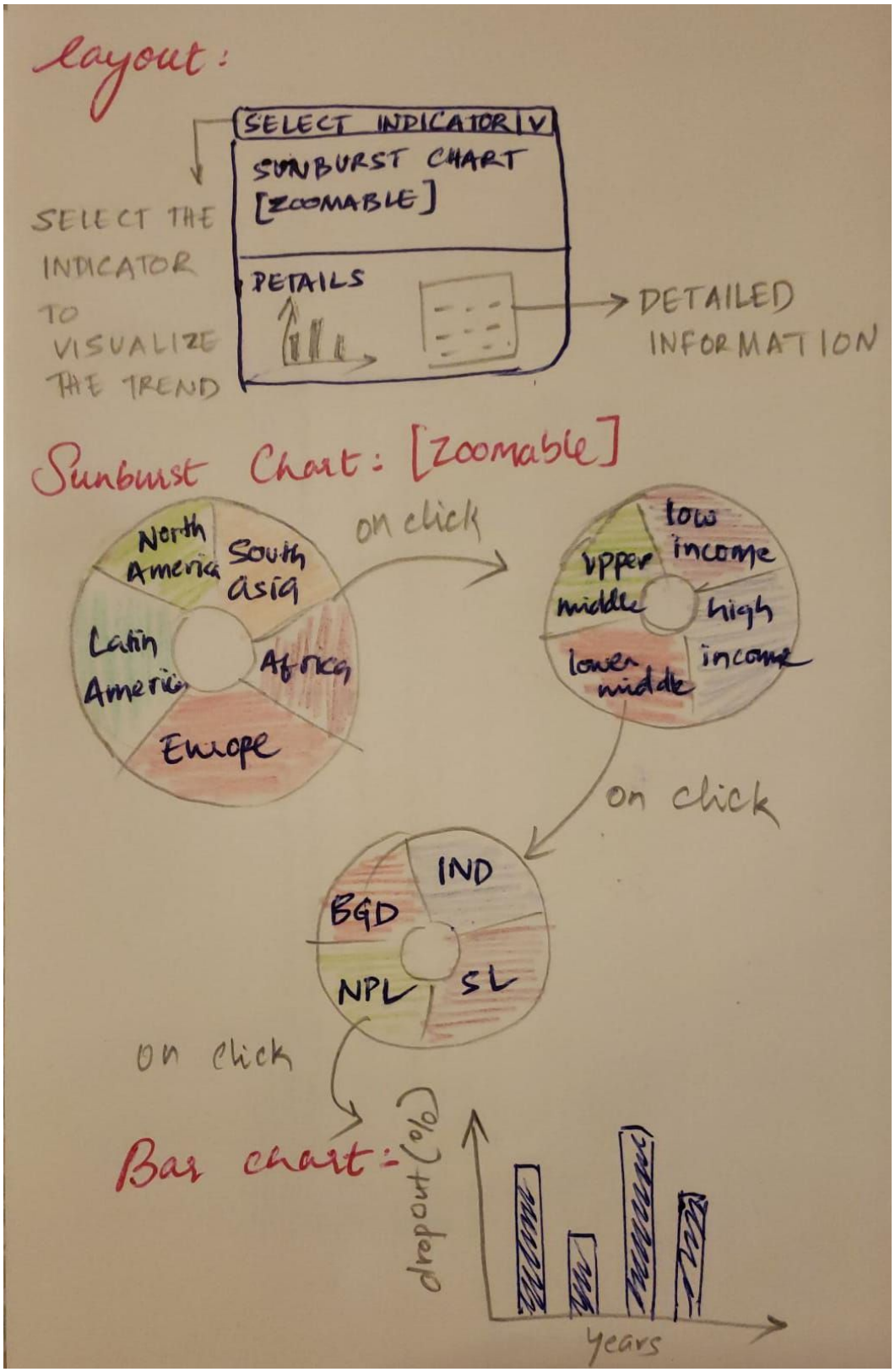
Prototype 2

Compare indicators for all years among countries

- World Map that lets you drag and drop countries (max 5) from it to the side pane
- Bar chart with adjacent bars showing the trend for each country
- Select the indicator to observe the trend over years



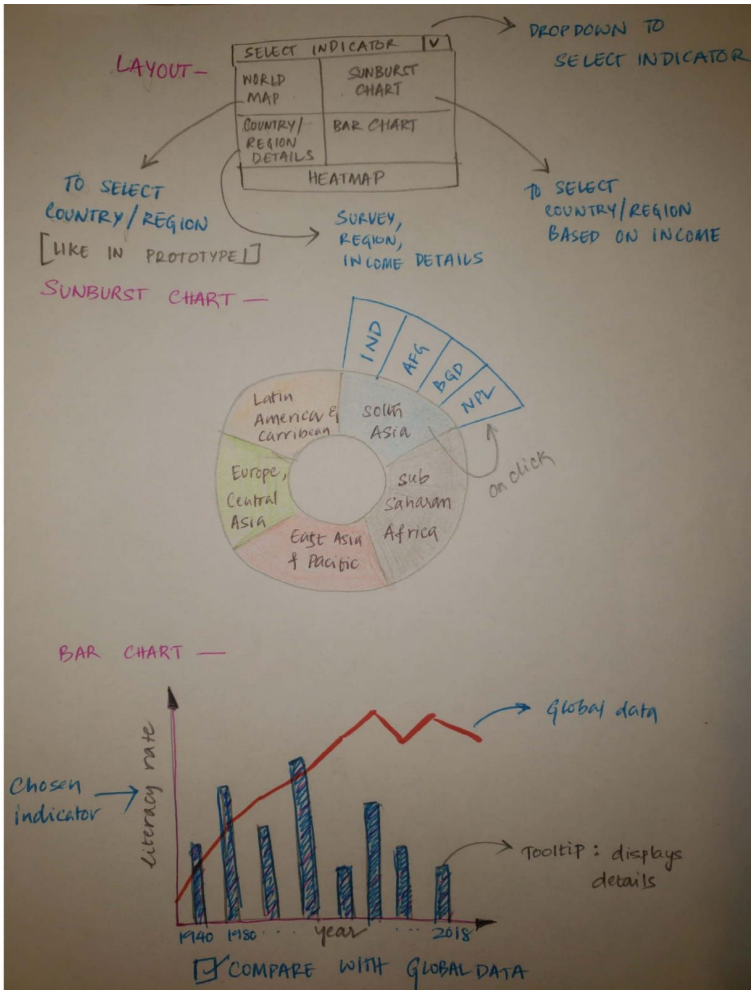
Prototype 3

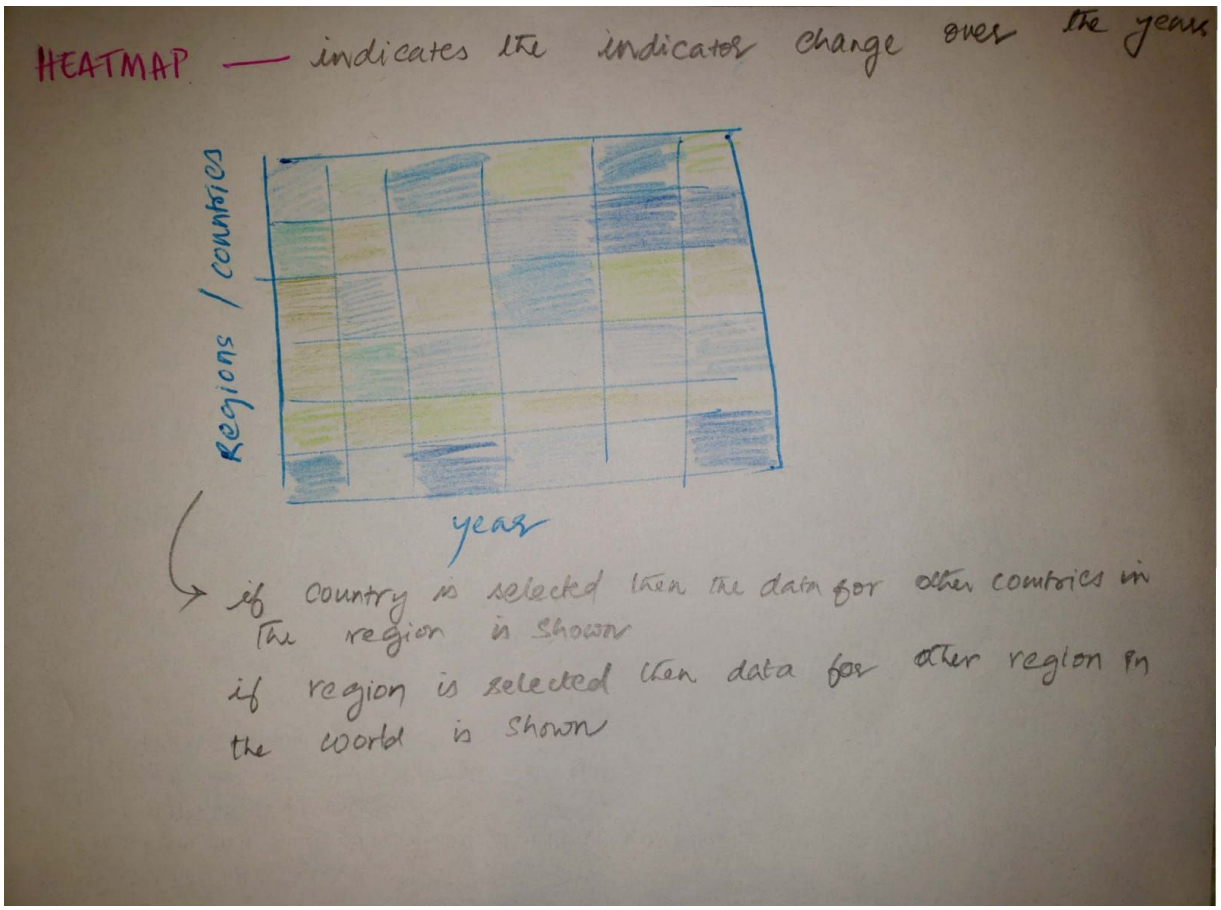

layout:
SELECT INDICATOR
SUNBURST CHART [ZOOMABLE]
DETAILS
SELECT THE INDICATOR TO VISUALIZE THE TREND
DETAILED INFORMATION
Sunburst Chart: [zoomable]
North America
South asia
Latin America
Africa
Europe
on click
low income
upper middle
high income
lower middle
on click
IND
BGD
NPL
SL
on click
Bar chart:
dropout (%)
years

Final Design

We picked up aspects from each of the prototypes that will help us tell the story and visualize education statistics trend.

- World Map that is colored based on the regions such as Sub Saharan Africa, East Asia & Pacific etc
- Sunburst chart with distortion that displays hierarchical data. First level of the hierarchy is represented by regions which drill down to countries, with the World being the innermost circle at the top of the hierarchy
- Bar chart that shows the trend for the selected indicator over the last 5 decades
- Select the region/country that you want to visualize the trend for
- Select the indicator to observe the trend over years
- Check the compare option to compare the global and country/region specific trends with a Line Chart.





MUST-HAVE FEATURES

- Being able to visualize data region-wise and country-wise for at least 5 indicators
- Being able to compare the trend for a particular country with the world data
- Sunburst chart or the world map to select a country with ease

OPTIONAL FEATURES

Heat Map that shows :

- How a selected country compares over countries in the same region as the selected country?
- How a selected region compares with the other regions in the world?

PROJECT SCHEDULE

Week 1 [Oct 29 - Nov 4]

- Do the data processing required for the visualization
- Host the project on github
- Significant work done on getting a working prototype

Week 2 [Nov 5 - Nov 11]

- Have the working prototype ready

Week 3 [Nov 12 - Nov 18]

- Work on incomplete functionalities of the prototype
- Have the basic visualizations working without any glitches

Week 4 [Nov 19 - Nov 26]

- Enhance the visualizations
- Add rest of the planned functionalities
- Fix bugs

Week 5 [Nov 26 - Nov 30]

- Test the visualization
- If there is time, add more extra features and work on making the project better

Project Milestone

OVERVIEW

The main objective of the project is to view how the education statistics indicators have changed over the years and compare different countries falling under different income groups to see the impact of income of a country over education. The visualization consists of four components-

- World map
- Sunburst chart
- Bar/Line chart
- Heatmap

After the peer feedback (Described below and in FeedbackExercise.pdf) and receiving the feedback from the TA, we decided to make some changes to the project.

DATA & DATA PROCESSING

Although the data does not need much processing, since it comprises of several indicators. We decided to extract a csv for each indicator to make data handling simpler. We also had to merge certain files to get the region and income group information. We did some of this work manually and some using a python script.

VISUALIZATION DESIGN

Project Proposal

After considering all the prototypes(Described in ProjectProposal.pdf) we had decided on the design shown below.

- World Map that is colored based on the regions such as Sub Saharan Africa, East Asia & Pacific etc
- Sunburst chart with distortion that displays hierarchical data. First level of the hierarchy is represented by regions which drill down to countries, with the World being the innermost circle at the top of the hierarchy
- Bar chart that shows the trend for the selected indicator over the last 5 decades

- Select the region/country that you want to visualize the trend for
- Select the indicator to observe the trend over years
- Check the compare option to compare the global and country/region specific trends with a Line Chart.

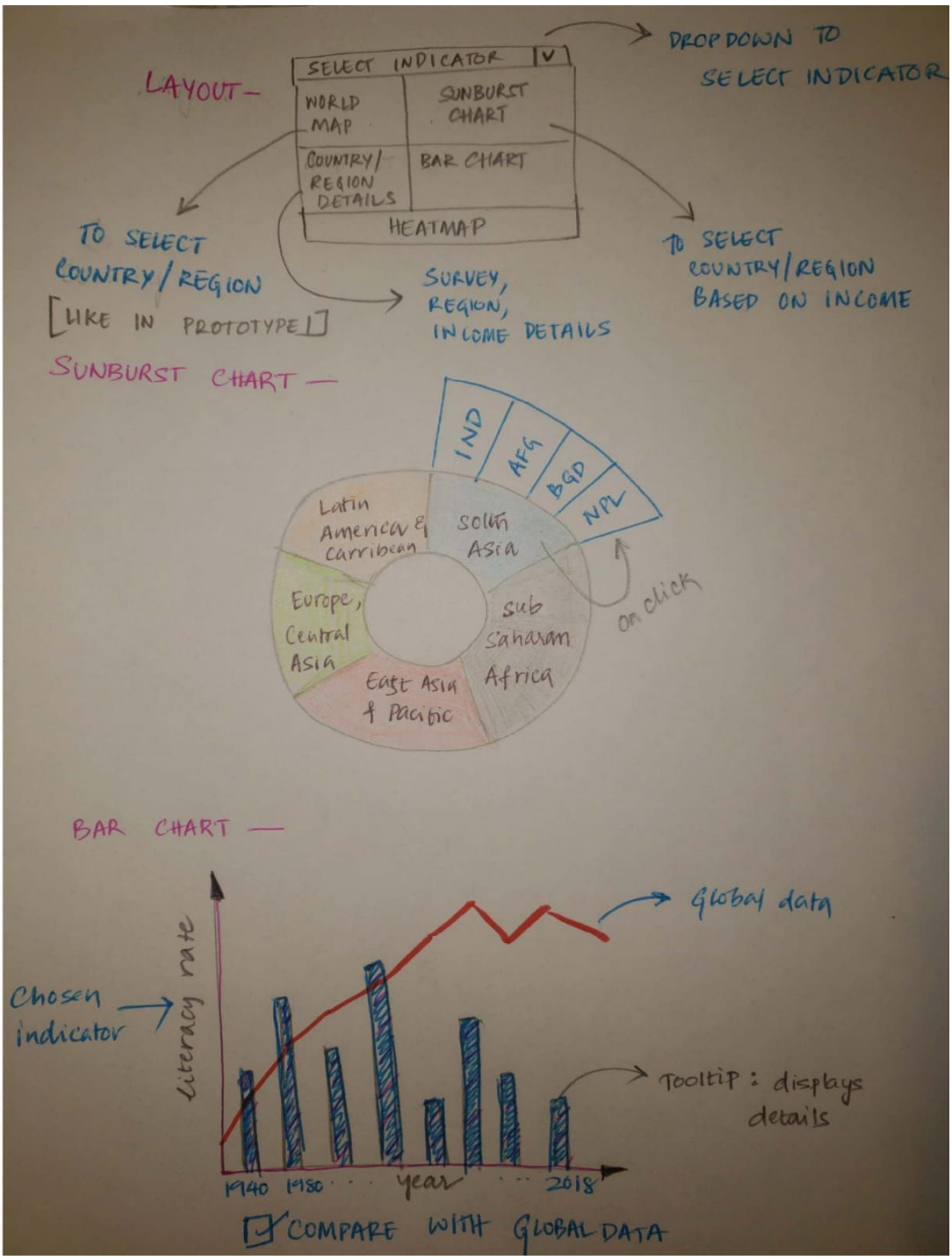
What changed?

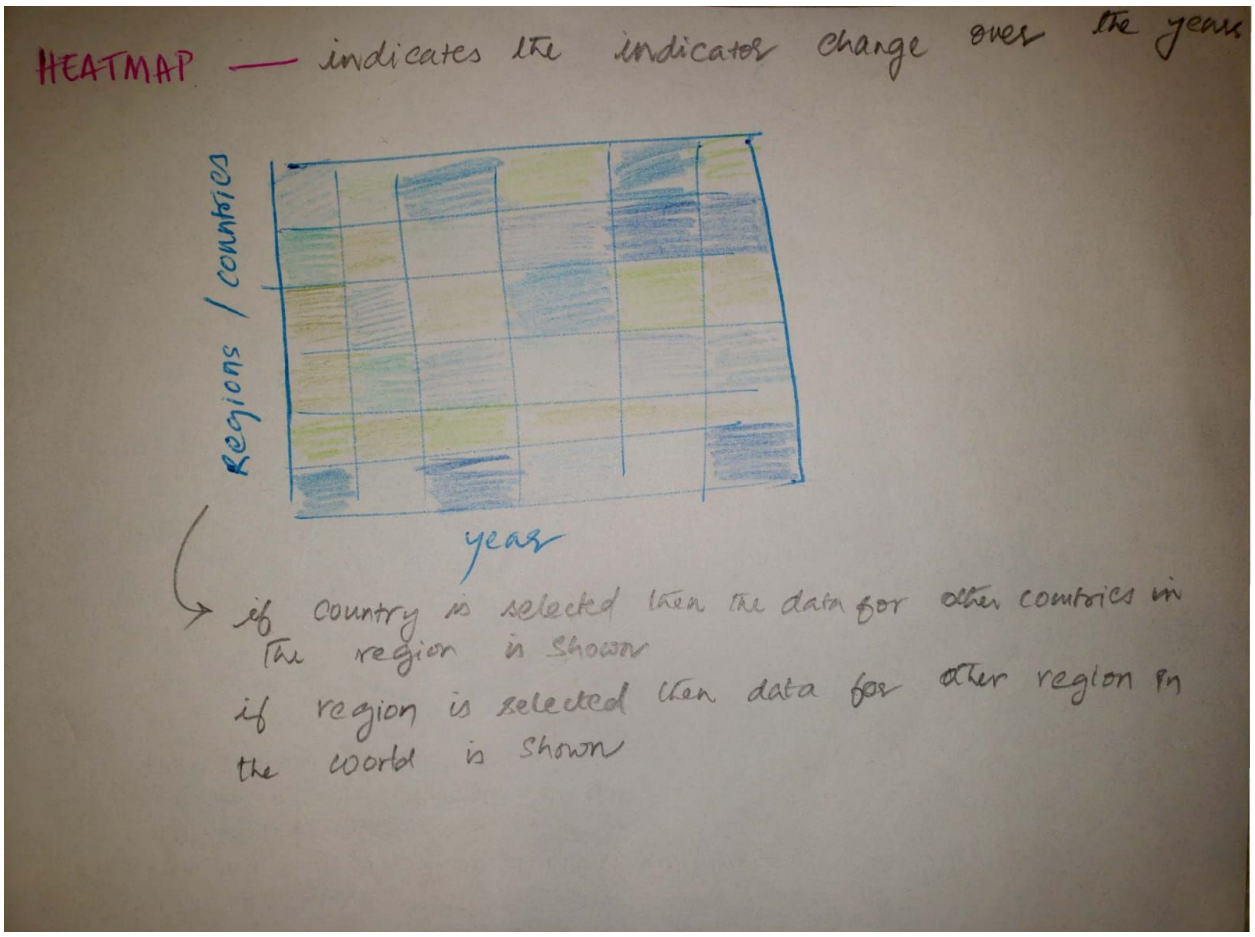
Based on the feedback we received from the peer group: The functionalities of the sunburst chart and world map were almost the same, that is, to select the country or data to visualize. Instead we decided to color code the countries based on the suggestion we got. The countries are now colored based on the indicator selected.

The heatmap representing the changes in the region and country was an optional feature. We decided to make it a must-have feature as it's helpful in visualizing the change over the years and compare the indicator data to other countries.

Based on the feedback we received from the TA: It was unclear how we were going to visualize various income groups impact on education. To address this, we decided to have the line chart or bar chart indicate the trend for all the countries in the selected income group (Region can be selected using sunburst chart/Map).

Also, we decided to encode the most literate/illiterate age group using a tooltip for the both map and line chart.


LAYOUT –
SELECT INDICATOR
WORLD MAP
SUNBURST CHART
COUNTRY/ REGION DETAILS
BAR CHART
HEATMAP
DROPDOWN TO SELECT INDICATOR
TO SELECT COUNTRY/REGION
[LIKE IN PROTOTYPE]
SURVEY, REGION, INCOME DETAILS
TO SELECT COUNTRY/REGION BASED ON INCOME
SUNBURST CHART –
IND
AFG
BGD
NPL
South Asia
Latin America & Carribean
Europe, Central Asia
East Asia & Pacific
Sub Saharan Africa
on click
BAR CHART –
Chosen indicator
Literacy rate
Global data
Tooltip : displays details
1940
1980
year
2018
COMPARE WITH GLOBAL DATA

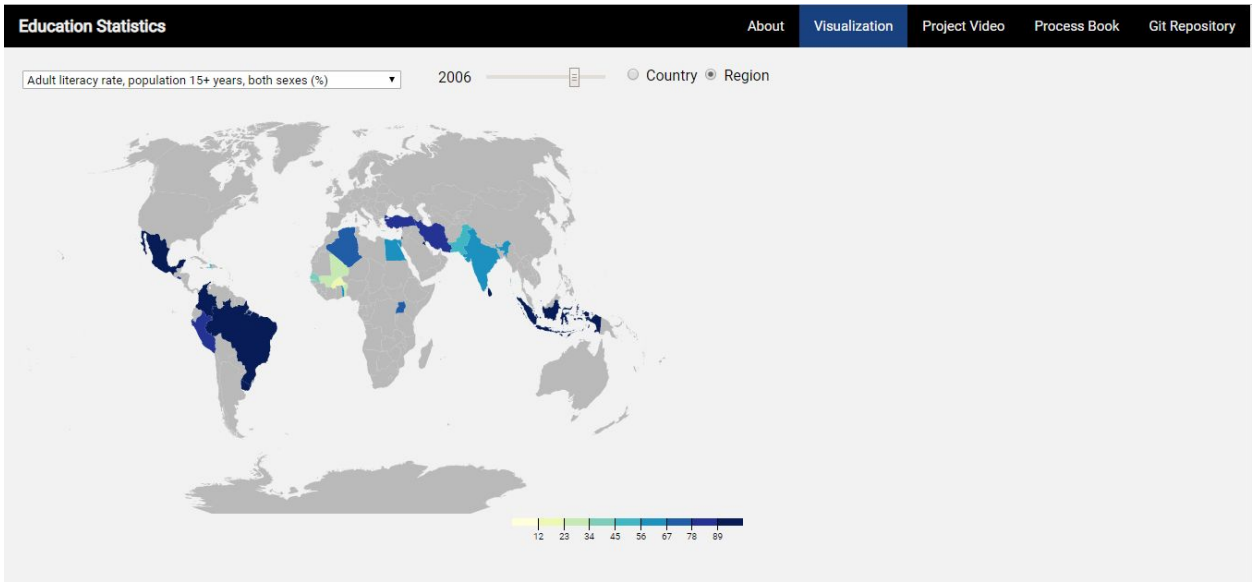


Project Milestone

The screenshot below shows the landing page for our visualization. Currently we have three selections on which the visualizations are based on.

- Dropdown: the indicator which has to be visualized
- Year slider: the year based on which the countries are colored
- Country/Region radio buttons: whether the world map and heatmap should be encoded based on country or region

Implementation of the landing page was straight forward and it is hosted using github pages. Right now, we have based the prototype on one indicator only. We have add support for other indicators. The framework for the map view is in place.



Implementation

Choropleth world map

We had to consider a lot of factors while deciding based on which indicator the countries should be colored or whether we need different colors based on region. In the end we decided to have a country and region option and encode it based on that.

For the project milestone, we have only the country view working. We have the data required for the region view, but there is some processing left before we can complete the visualization.

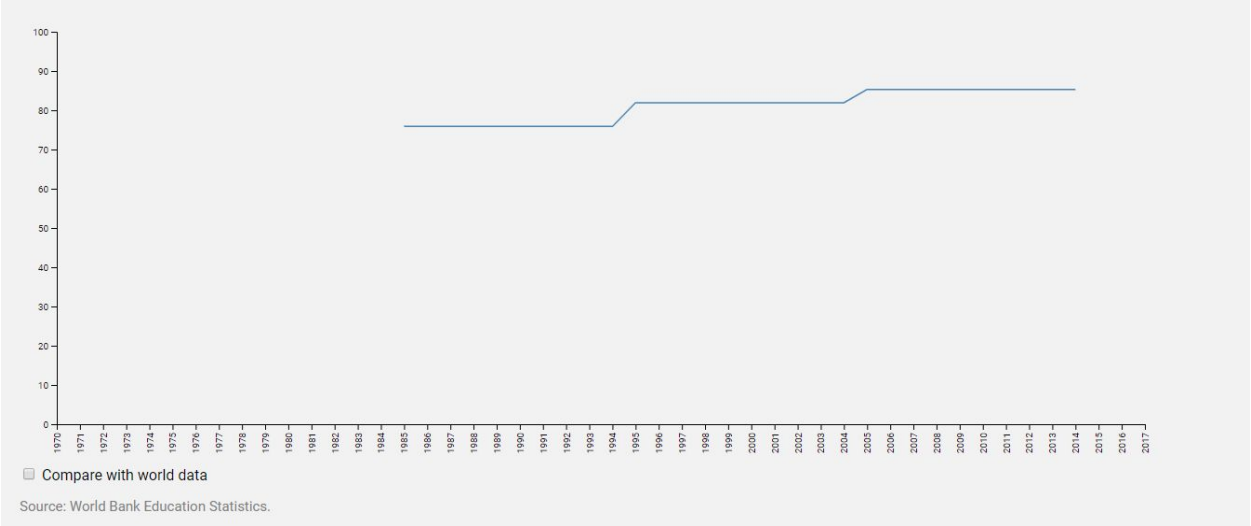
We chose the color scheme to suit our landing page.

Line chart

Our dataset has a lot of missing values, this is because the data is collected using several surveys and not all of them are conducted annually. After discussing with the TAs and going over several ideas, we decided to do something similar to https://bocoup.com/blog/showing-missing-data-in-line-charts. We are still experimenting whether line chart or bar charts can be used to interpret the data well.

Currently, our line chart shows the global data and the checkbox is meant to toggle whether to compare against the global data or not. Our other alternative for this is to

also, compare against the countries region data for country view. Also, if a income group is selected then options to compare against countries in the same income group or region. There are ideas we are still thinking of, for the line chart implementation.



PROJECT SCHEDULE

Work completed so far

As per our schedule mentioned in the project proposal, we are on track and we have completed :

- Data processing. Some of the processing was done manually and some csv's were generated using python (data_processing.py). Our data is arranged in 4 different csv's namely adult_literacy_rate_both_sexes, adult_literacy_rate_gpi, adult_literacy_rate_male, adult_literacy_rate_female.
- We have our landing page ready and hosted.
- We have also completed a working prototype with our world map and line chart.

Team member contributions

Worked together on building and hosting the landing page, data processing. We constantly communicated about what each of us was doing and came up with implementation ideas together.

Greeshma: Worked on getting the world map working.

Tanvi: Worked on line chart and how to represent missing data.

Final Design

OVERVIEW AND MOTIVATION

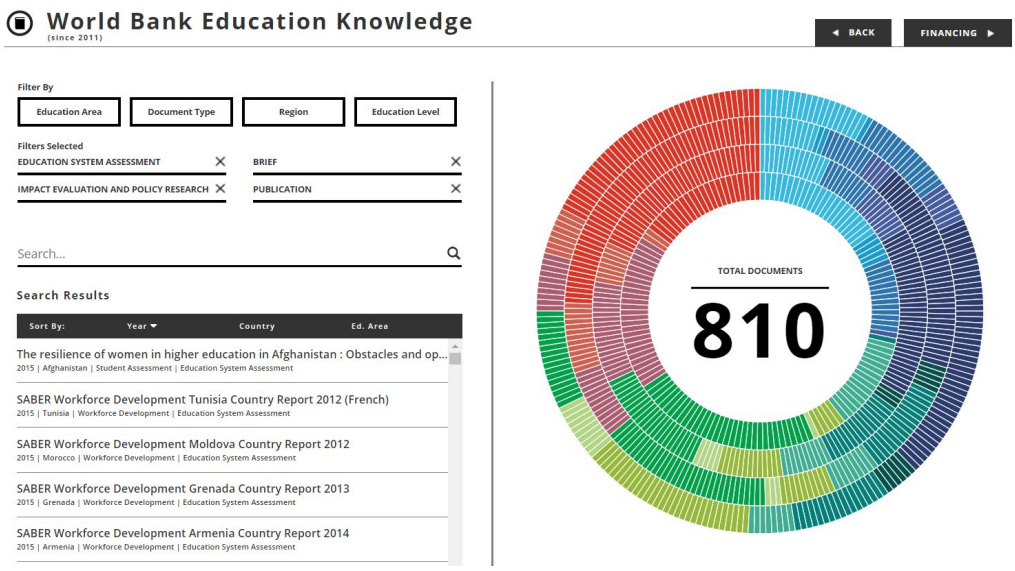
We started looking for datasets that seemed interesting and we came across the Education statistics dataset from World Bank datasets. This data has indicators that describe literacy rate, enrolment rate, illiterate population across education levels (primary/ secondary/ tertiary), age groups and gender. We decided to use this dataset as education can be associated with various parameters such as a country's economic growth, crime rate etc. We also wanted to draw parallels between the economic state of the country (low income vs high income) and education rate to understand the factors influencing the latter over the years.

The main objective of the project is to view how the education statistics indicators have changed over the years and compare different countries falling under different income groups to see the impact of income of a country over education. The visualization consists of four components-

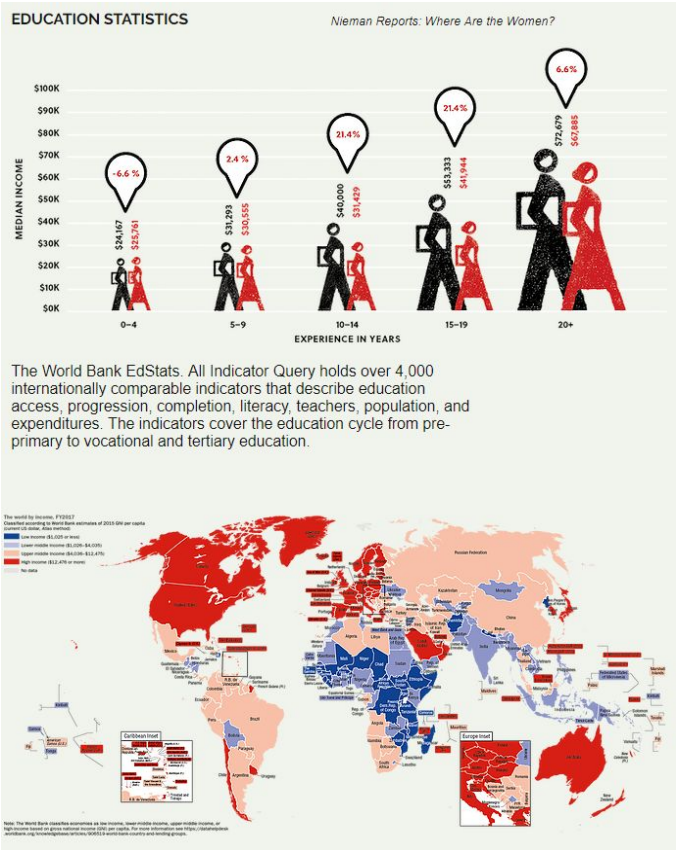
- World map
- Sunburst chart
- Bar/Line chart
- Heatmap

RELATED WORK

We took inspiration from websites such as columnfive which conceptualize and design interactive, online tools to visualize the World Bank Group's education lending trends and contributions to education from 2000–2015, emphasizing how the World Bank Group has supported country reform of education systems from 2011–2015 based on evidence of what matters.



Another effective visualization we found with respect to world data was [awdcglobal] where visualizations are used for similar datasets to tell a story about the advancements achieved in public health, gender equality, and economic development.



QUESTIONS

We tried to answer a set of questions as a part of our analysis through our data visualization. The primary questions being:

- How indicators such as literacy rate among males/females, illiterate population etc, have changed in countries over the last five decades
- Which are the most literate/illiterate age groups
- What's the highest level of education in specific countries
- Does the income of the country impact its literacy rate

As mentioned earlier, education can be related to various factors that influence our society. Visualizing such a dataset might help analyze the impact education has, on the growth of a country/region. In more developing countries, it would be beneficial to see the gender parity in education and whether it is moving towards equality in education.

DATA

We got out dataset from https://datacatalog.worldbank.org/dataset/education-statistics, the World Bank's data datalog.

The dataset includes multiple files and is quite large. It consists of over 4,000 internationally comparable indicators that describe education access, progression, completion, literacy, teachers, population, and expenditures. We do not intend to use all of them for the purpose of this project as this might cause a lag in the webpage. We plan to use indicators that fall under the three essential categories: gender, age group and education level. This will allow us to effectively visualize the Gender Parity Index, highest level of education obtained across regions/countries and drop out rate amongst different age groups with respect to the economic state of the country. We also intend to show a global trend by comparing country statistics with the global data.

Since the data we want to use for this visualization is scattered over multiple files, we need to extract only the data we require depending on the indicators, region and countries. We plan on doing this using Python or Javascript.

Although the data does not need much processing, since it comprises of several indicators. We decided to extract a csv for each indicator to make data handling simpler. We also had to merge certain files to get the region and income group information. Some of the processing was done manually and some csv's were generated using

python (data_processing.py). Our data is arranged in 4 different csv's namely adult_literacy_rate_both_sexes, adult_literacy_rate_gpi, adult_literacy_rate_male, adult_literacy_rate_female.

EXPLORATORY DATA ANALYSIS

We used the Databank query tool provided by the World Bank for exploratory data analysis. The World Bank EdStats Query holds around 2,500 internationally comparable education indicators for access, progression, completion, literacy, teachers, population, and expenditures. The indicators cover the education cycle from pre-primary to tertiary education. The query also holds learning outcome data from international learning assessments (PISA, TIMSS, etc.), equity data from household surveys, and projection data to 2050.

We examined the data using the query tool for various indicators and regions and decided to use three important indicators that had the least missing data. The query tool also helped us understand some of the indicators better like what adjusted net enrolment rate, lower secondary, both sexes meant.

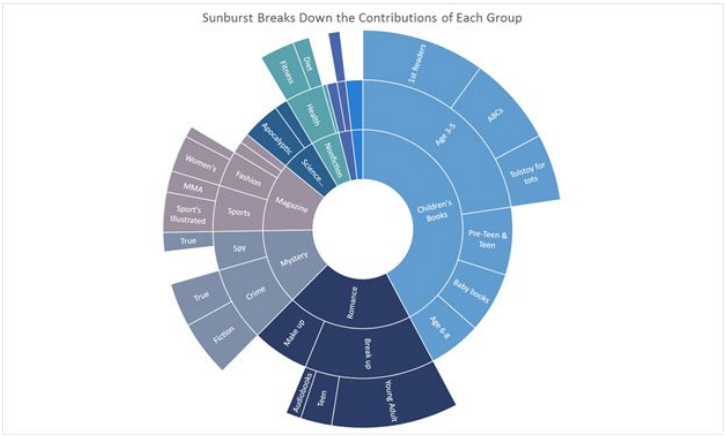
We also saw the need for a visualization for this data after we spent hours using the query tool to better understand the data.

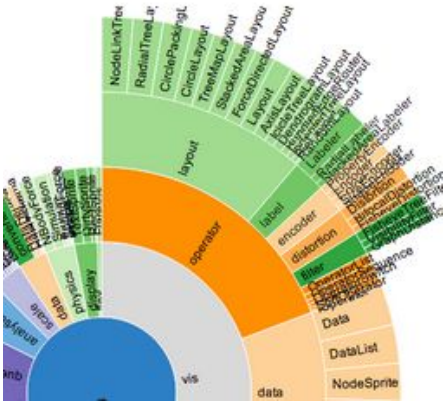
DESIGN EVOLUTION

After considering all the prototypes(Described in ProjectProposal.pdf) we had decided on the design described below:

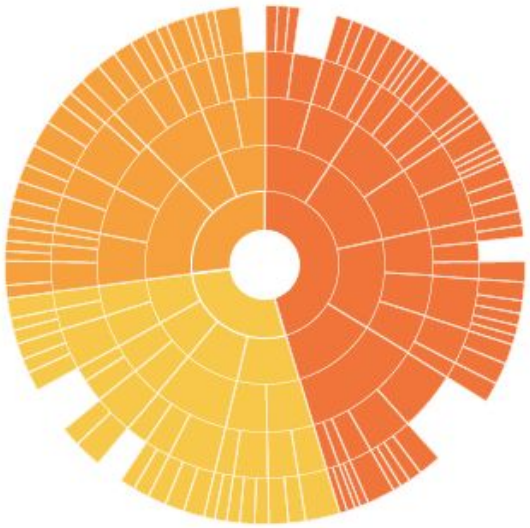
- World Map that is colored based on the regions such as Sub Saharan Africa, East Asia & Pacific etc
- Sunburst chart that displays hierarchical data. First level of the hierarchy is represented by regions which drill down to countries, with the World being the innermost circle at the top of the hierarchy
- Bar chart that shows the trend for the selected indicator over the last 5 decades
- Select the region/country that you want to visualize the trend for
- Select the indicator to observe the trend over years
- Check the compare option to compare the global and country/region specific trends with a Line Chart.

Distorted Sunburst chart :





We initially started with the implementation of a [distorted sunburst] to represent the region and country hierarchy as we thought it would help us imagine the dataset better. The standard distorted sunburst charts are displayed above. We decided against this version of sunburst as it takes up more screen space and we were trying to optimize our screen usage to fit in all visualizations. We also found it difficult to implement and debug in D3 v5 as we couldn't find any relevant examples for the same.

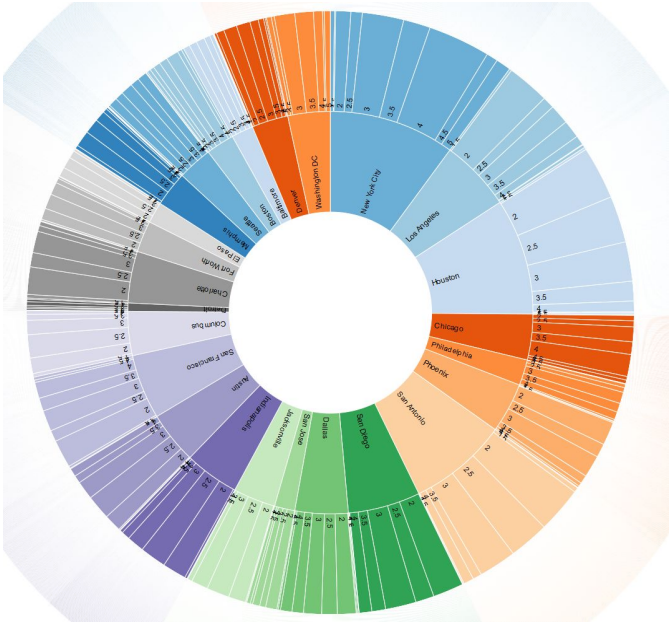


What we choose instead: Zoomable sunburst

Some of the standard examples are below. This type of visualisation shows hierarchy through a series of rings, that are sliced for each category node. Each ring corresponds to a level in the hierarchy, with the central circle representing the root node and the hierarchy moving outwards from it.

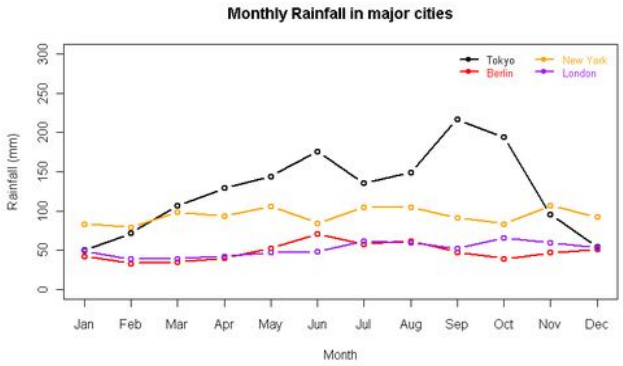
Rings are sliced up and divided based on their hierarchical relationship to the parent slice. The angle of each slice is divided equally under its parent node. Colour is used to highlight hierarchical groupings or specific categories.

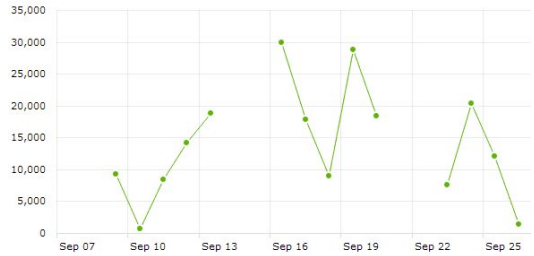
Use of the zoomable sunburst sped up our implementation and also helped us better coordinate our map and sunburst together



Line Chart:

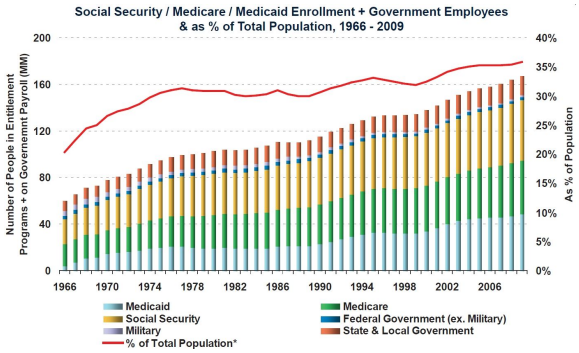
To represent in the indicator data for countries/regions, we initially decided to implement line charts (something similar to the first image below). A line chart or line graph is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments. After realizing that our dataset has a lot of missing values, because the data is collected using several surveys and not all of them are conducted annually. We imagined our line chart with bad/missing data using bocoup. It would look something similar to the second image below and hence we decided against it. Though we are still using a line chart for the world since the data is complete.

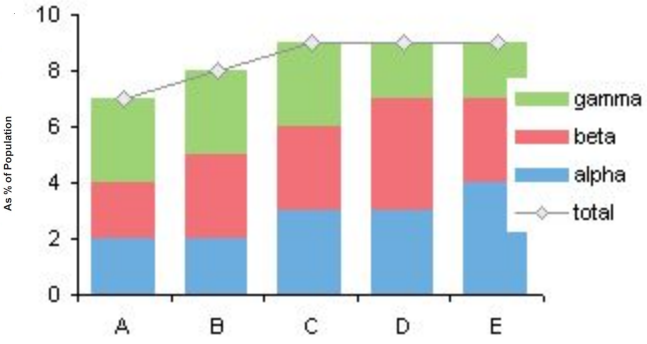




What we choose instead: Bar Chart

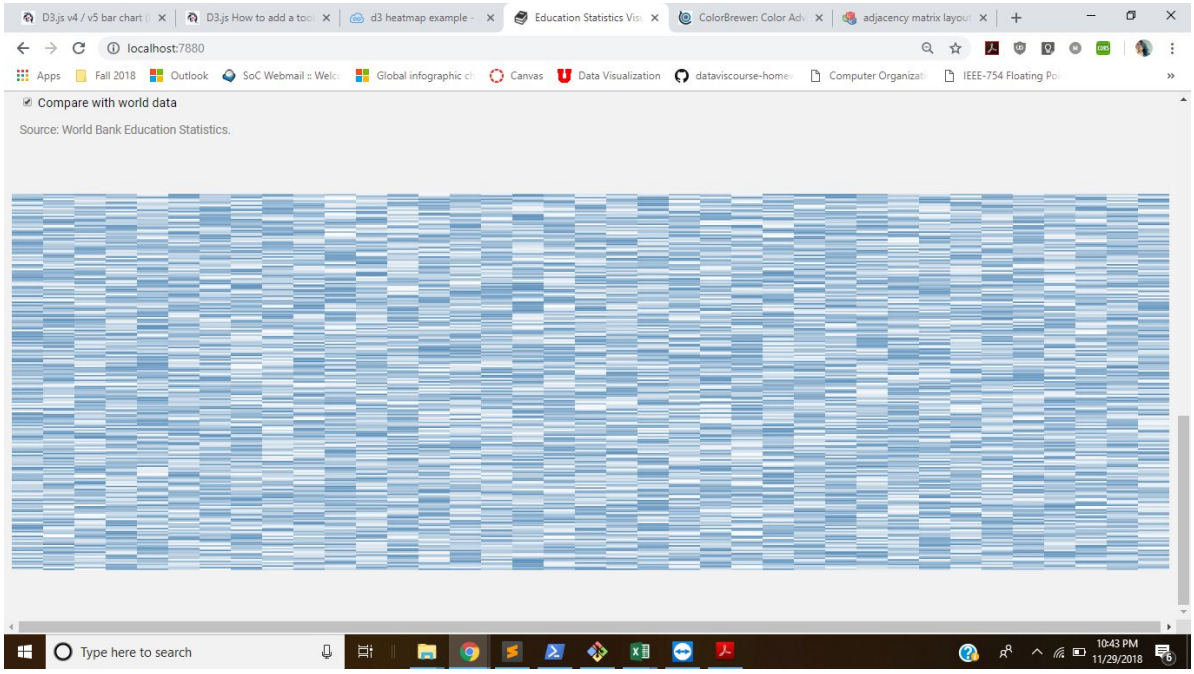
Bar charts or graphs are used to compare things between different groups or to track changes over time. Below are some of the designs we looked at.

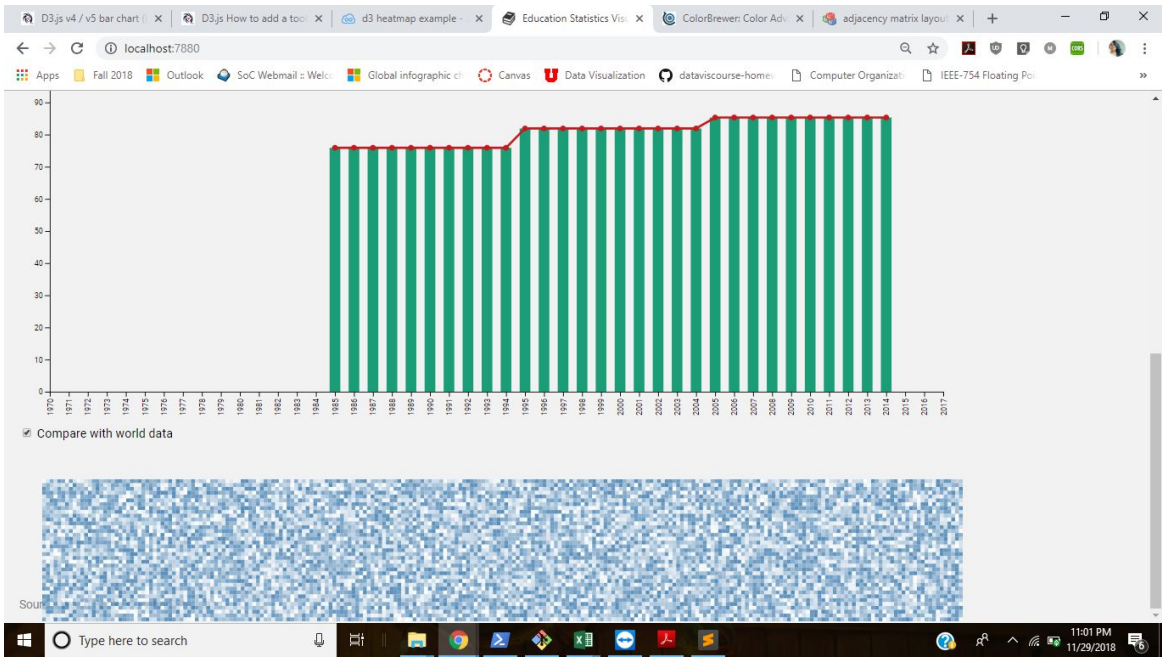




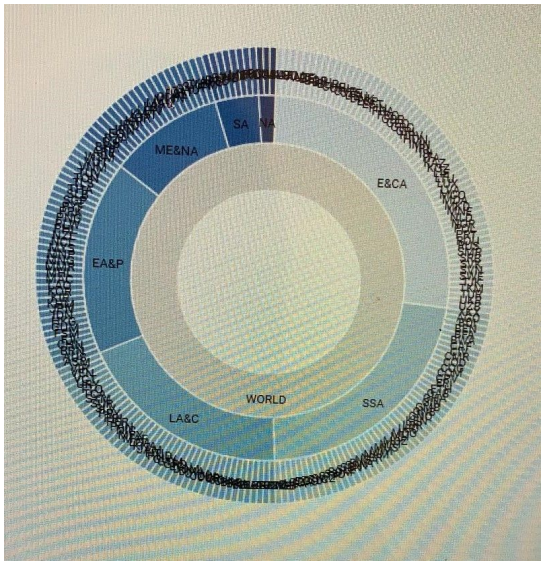
Heat Map:

A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors. We tried various implementations of heat map and some of them did not turn up as we had expected (refer images below). Since the number of entities is huge, the heat map looked like a pixelated image and did not appropriately represent the data.


Education Statistics Vis
localhost:7880
Compare with world data
Source: World Bank Education Statistics.
Type here to search
10:43 PM
11/29/2018


Education Statistics Vis
localhost:7880
Compare with world data
Type here to search
11:01 PM
11/29/2018

We have also made some color changes along the way for the sunburst chart cause we did not want it to be related to the blue hues of the map that are actually determined by the data. The below was our initial choice.

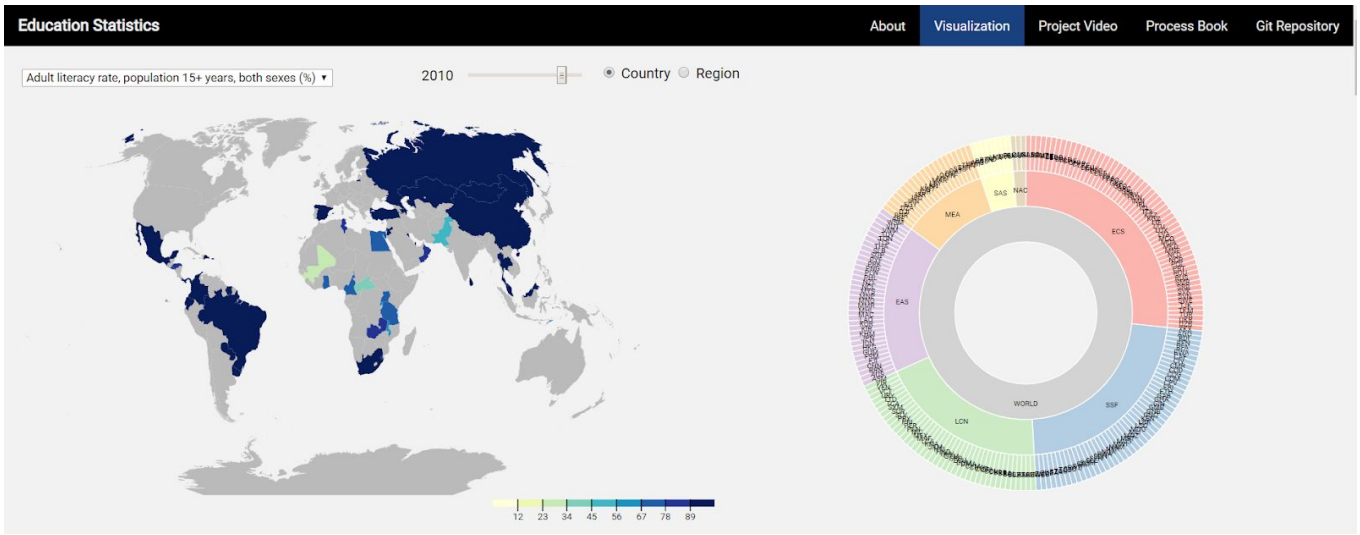


IMPLEMENTATION

The screenshot below shows the landing page for our visualization. We have three selections on which the visualizations are based on.

- Dropdown: the indicator which has to be visualized
- Year slider: the year based on which the countries are colored
- Country/Region radio buttons: whether the world map and heatmap should be encoded based on country or region

Implementation of the landing page was straight forward and it is hosted using github pages.



The visualization consists of four components-

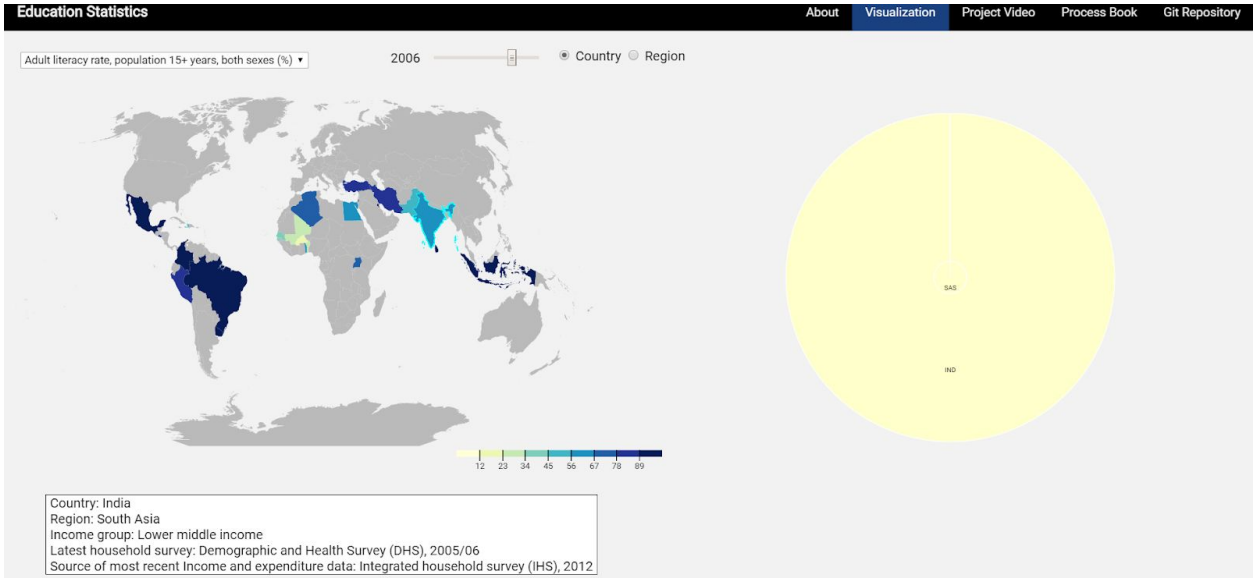
- World map
- Sunburst chart
- Bar/Line chart
- Heatmap

Choropleth world map

We had to consider a lot of factors while deciding based on which indicator the countries should be colored or whether we need different colors based on region. In the end we decided to have a country and region option and encode it based on that. We choose a

world map cause that best represents all countries and makes it easy to identify and select regions for the users.

We chose the color scheme to suit our landing page.

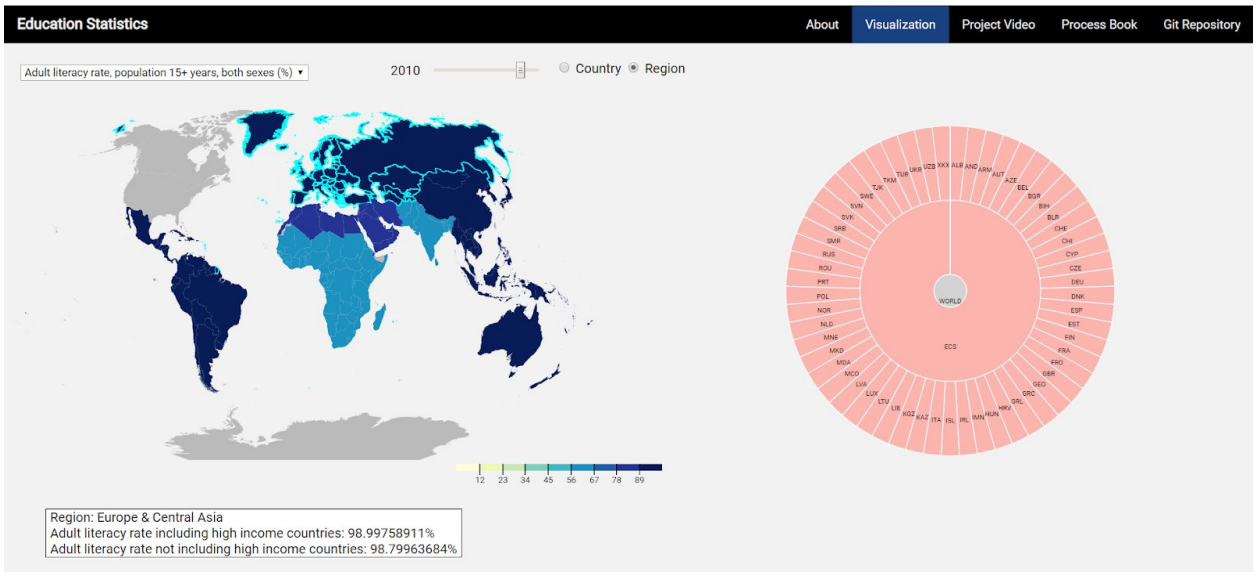


Sunburst

We decided to have a sunburst along with the world map for representing the hierarchical structure between regions and countries better. Suppose, the user is unaware of which region a particular country belongs too, they can pick the region from the sunburst chart which will inturn select the region in the world map.

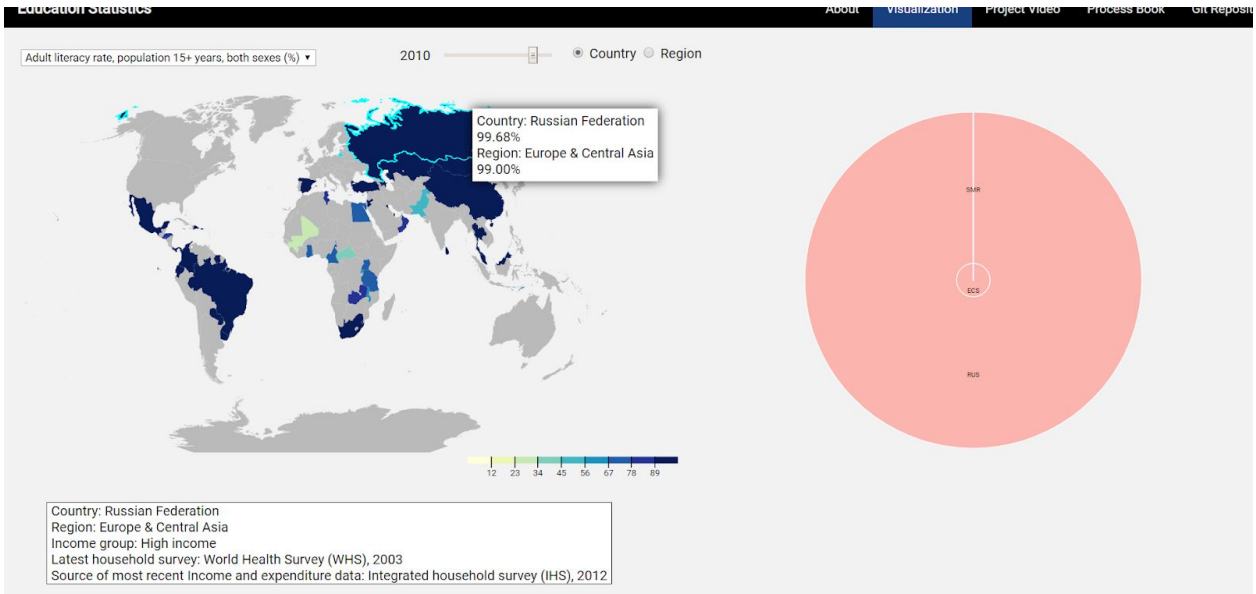
The tooltip on hover for both the world map and the sunburst gives details regarding the selected entity.

The information box tells you more details about the selected data such as the region a country belongs too, the particular surveys that were conducted to collect the data.



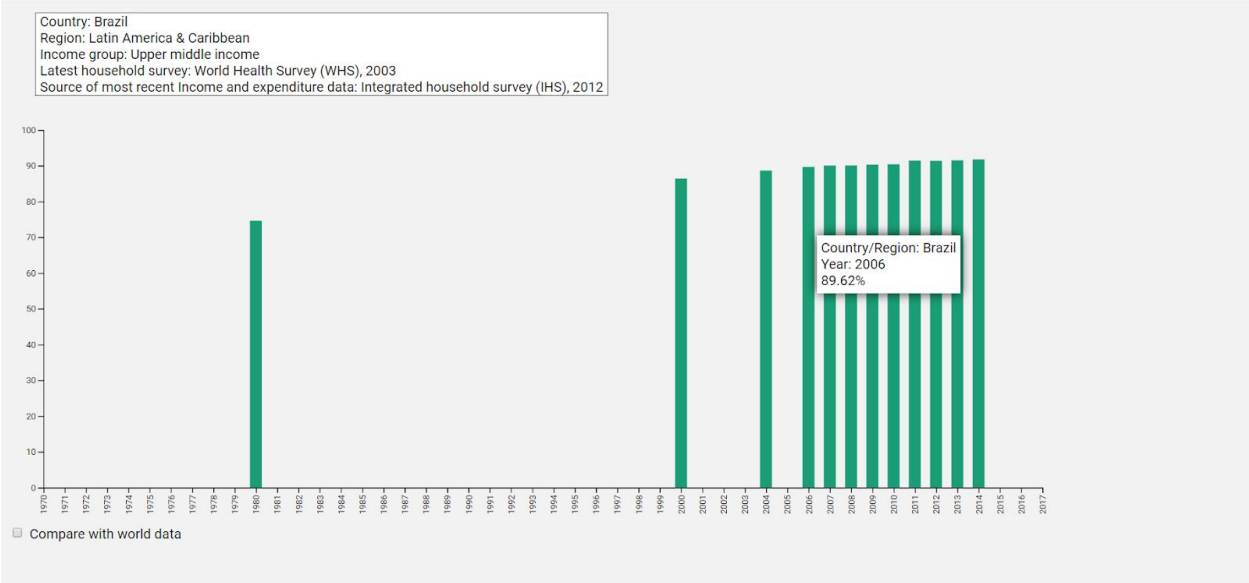
On selecting a particular region in the map, the views that are updated are

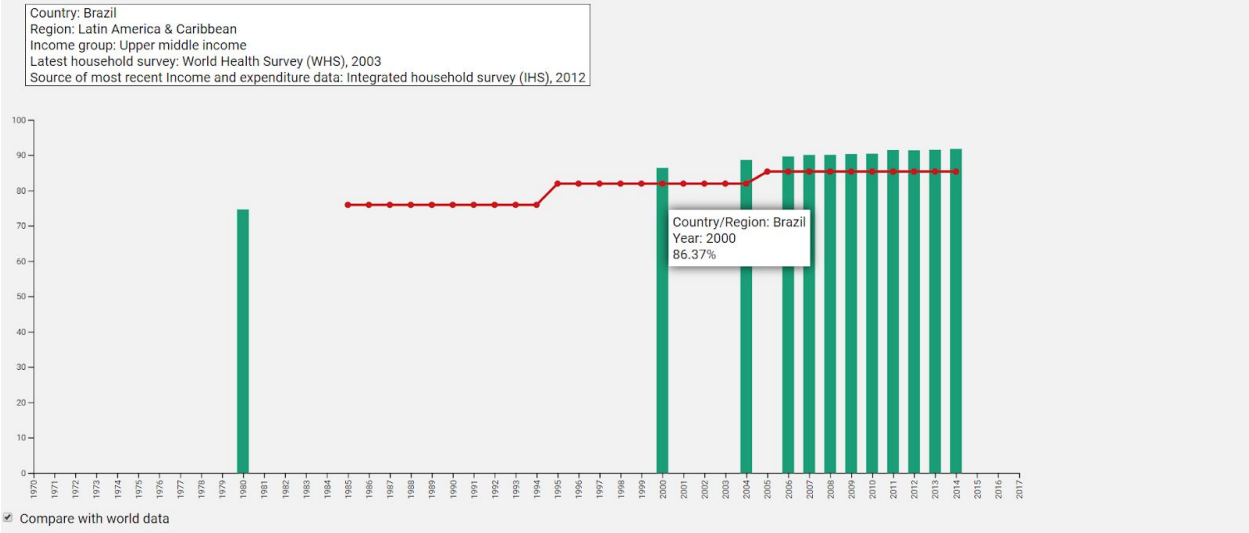
The Sunburst - the region and all the countries in it are shown.
The Bar chart - the data was the chosen indicator and the chosen region is displayed
The Heat Map - The comparison between all regions is shown.



Bar/Line Chart

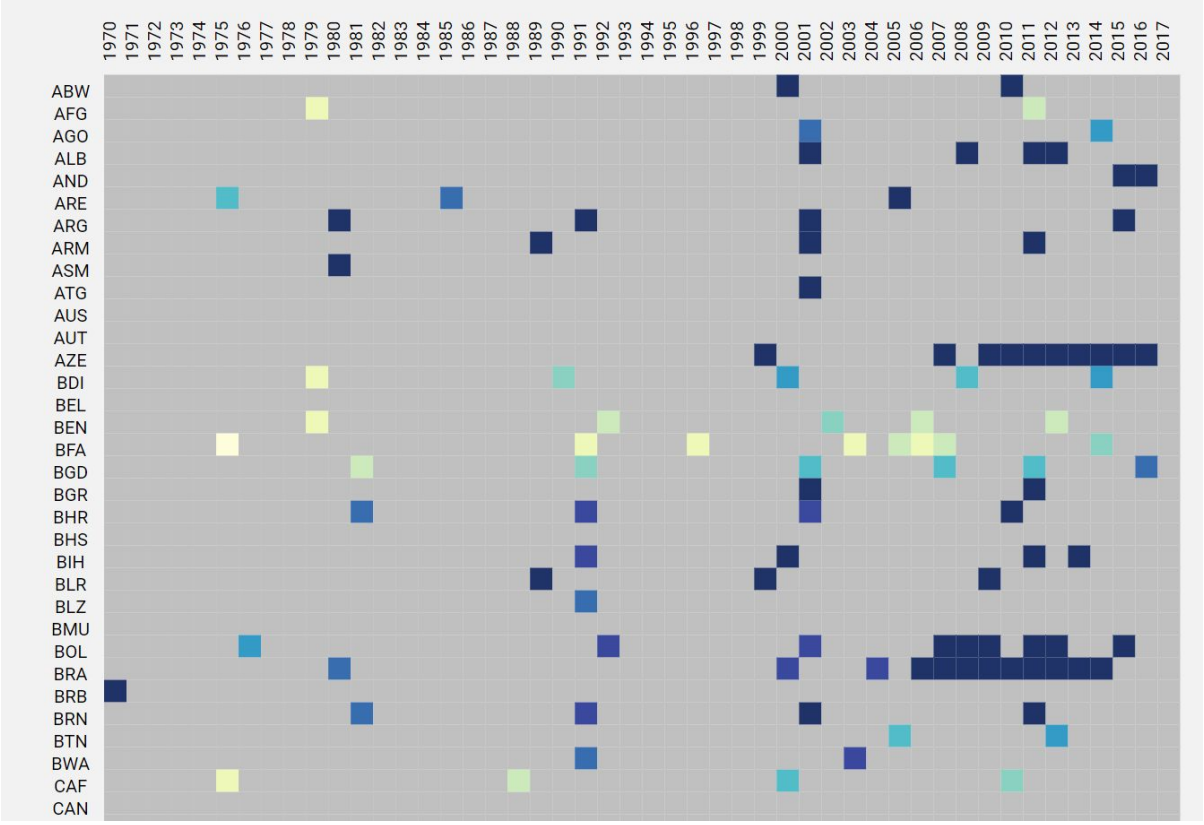
The bar chart displays the data for the indicator and country/region selected for all years. We can also choose to compare a particular region/country with world data by enabling the compare. We are displaying the world data using a line chart since it is complete data and there is no missing/bad data.



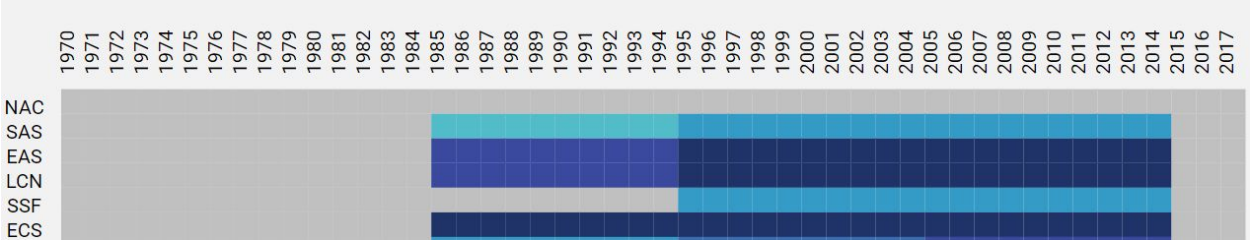


Heat Map

The heat map has two views, one that compares data across years for all the countries for a given indicator. For example, consider two countries Burkina Faso (BFA) and Brazil (BRA), the indicator chosen being adult literacy rate - both sexes. We can clearly see that for year 2006, the rate was much lower for BFA when compared to BRA.



The second view of the heat map compares the data for all the regions, For example, consider two regions SSF and SAS, the indicator chosen being adult literacy rate - both sexes. we can infer from the heat map that from years 1995 to 2014, the regions Sub saharan africa (SSF) and South Asia (SAS) have had similar literacy rate.



EVALUATION

Did you deviate from your proposal?

Yes. Unfortunately, we ran into fairly significant problems in collecting data. The data we were relying on proved to be sparsely populated, and reliable data was not available in any of the sources we looking into. After much debate and collecting data for the initial proposal, we decided that the difficulties in locating data were large enough to merit a switch in the direction we were taking our project in. We limited our focus to Adult Literacy rate. We were unsuccessful in answering some of the questions mentioned during our project proposal like, Which are the most literate/illiterate age groups?]What's the highest level of education in specific countries? This was due to unexpected inconsistencies in the data.

What did you learn about the data by using your visualizations?

- We learnt that the literacy rate in countries that come in the lower income group has started to improve recently.
- There are no countries in the lower or middle income group in regions like North America and regions like South Asia have no higher income groups, this might be a major contributor in the fact that some of the countries have higher literacy rate (US,Canada) and countries like Nepal, India have lower literacy rate.

How did you answer your questions?

The primary questions being:

- How indicators such as literacy rate among males/females have changed in countries over the last five decades

  The year slider and heatmap lets you see the trend across the years for all countries and the two views (Country view and Region view) lets you look at literacy rate in a broader context.

- Does the income of the country impact its literacy rate

  By clicking on the countries we can see it's income group. The general trend we observed was countries with higher literacy rate were falling in a higher income (For example, Brazil) category compared to countries with lower literacy rate (For example, Mail).

How well does your visualization work, and how could you further improve it?

We were able get all the interactive views that we planned on, and it's easy to navigate. However, while working on it we did realize some of the options might not be best suited and we changed it accordingly.

We could add some more features to our visualization such as:

- Compare against the countries region data for country view. Also, if a income group is selected then options to compare against countries in the same income group or region
- Heatmap for countries could only comparisons between other countries in the same region would be more helpful.
- Having a search feature would help find a country easily.