

EMPLOYMENT STATISTICS VISUALIZATION:

Team Members:

Sai Pavan Choudary

Sai Pramod Yerubandi

Table of Contents

ABSTRACT.....	3
INTRODUCTION.....	3
DATA PREPROCESSING	3
TOOLS.....	4
TECHNIQUES	4
TECHNIQUE DESCRIPTION	4
VISUALIZATION TECHNIQUES & RESULTS	6
GEO VISUALIZATION.....	6
DENDROGRAM	7
RADAR CHART	8
LINE GRAPH.....	8
BAR GRAPH	9
INFERENCES	10
CONCLUSION	11
REFERENCES.....	11

ABSTRACT

In today's technological world analyzing the data is one of the major challenge. Analyzing this data manually is tiring procedure and will not give accurate results every time, Visualization is the best solution to overcome this problem. Visualizing the data helps to identify the different patterns and analyze them in a more effective way by representing in figures. Our main aim of this project is to design an interactive data visualization which helps the user to analyze and compare the global employment statistics data and its dependent factors.

INTRODUCTION

Employment statistics is a user friendly web page which displays interactive visualizations. Apart from visualizing the data alone we also tried to embed the visualization techniques in a web page using the concepts of web designing, scripting and styling. Initially, we built a geo-visualization and a dendrogram to display the employment rate and its dependent factors for 40 different countries from the year 2000-2014. We also used line chart, bar graph and radar chart visualization techniques to compare and contrast the data between 2 countries, which changes dynamically by selecting a different country from the provided dropdown box. The purpose of this whole project is to corelate the employment rate and its dependent factors globally from the year 2000-2014.

DATA PREPROCESSING

The data for this project is collected from <http://www.oecd.org/>. We have selected 3 employment dependent factors to correlate them between 40 different countries. For this purpose, we have chosen 4 different datasets which has information about 240 different countries from the year 1950-2014. The chosen datasets for this project are Employment dataset, Population dataset, Finance dataset and Education dataset.

The data formats of the selected statistical data are in the CSV file format. Data for the geo-visualization is taken in CSV file format and Dendrogram is taken in JSON format. The original data downloaded from the website had a lot of missing data in them, so we had to preprocess the data by using appropriate filtering techniques to eliminate the missing values. Data preprocessing was done in python and after preprocessing the data by applying certain filters we selected the original data of 40 different countries from the year 2000-2014. After preprocessing the data, we had some null values in our data and those null values where handled by equating to 0.

For geo-visualization and Dendrogram we created two separate datasets which is the weighted average of all the attributes i.e. education, population, finance, employment datasets from the year 2000 to year 2014.

TOOLS

Tools used for this course project are D3 programming for visualizations, HTML5 and CSS for interactive user interface. D3 which is a JavaScript is used to produce the data visualizations, that helped to achieve the main purpose of this project. HTML5 (Hyper Text Markup Language) and CSS (Cascading Style Sheets) are used for web page presentation and styling.

TECHNIQUES

Data representation is done in 2 diverse ways in this project, one of them is the statistical representation and the other being the global representation. Techniques used for global representation are Dendrogram and Geo visualization which plots the weighted-average of all attributes from the year 2000 to the year 2014. Statistical data is represented using Bar chart, Line graph and the Radar chart.

Dendrogram was used to represent the countries and their average values of population, employment, finance and education attributes from the year 2000 to 2014. Similarly, other visualization used to portray the average values of the 4 attributes from 2000 to 2014 is Geo visualization, in which all the countries are represented on the world map so that it makes it easy for the user to see the data for a specific country. To select a country, a dropdown box was provided, the input from the dropdown box is fed into other visualization techniques which are mentioned above. Line graph is a type of visualization technique which displays the information as a series of data points called as markers which are connected by straight lines. Radar chart is a graph which is used to represent multivariate data in graphical format. A bar graph is used to represent the data with rectangular bars with heights. In this project Line chart, Radar chart and Bar chart are used to plot the data of a single country from the year 2000 to the year 2014 and they are also used to correlate the data of 2 different countries from the year 2000 to the year 2014.

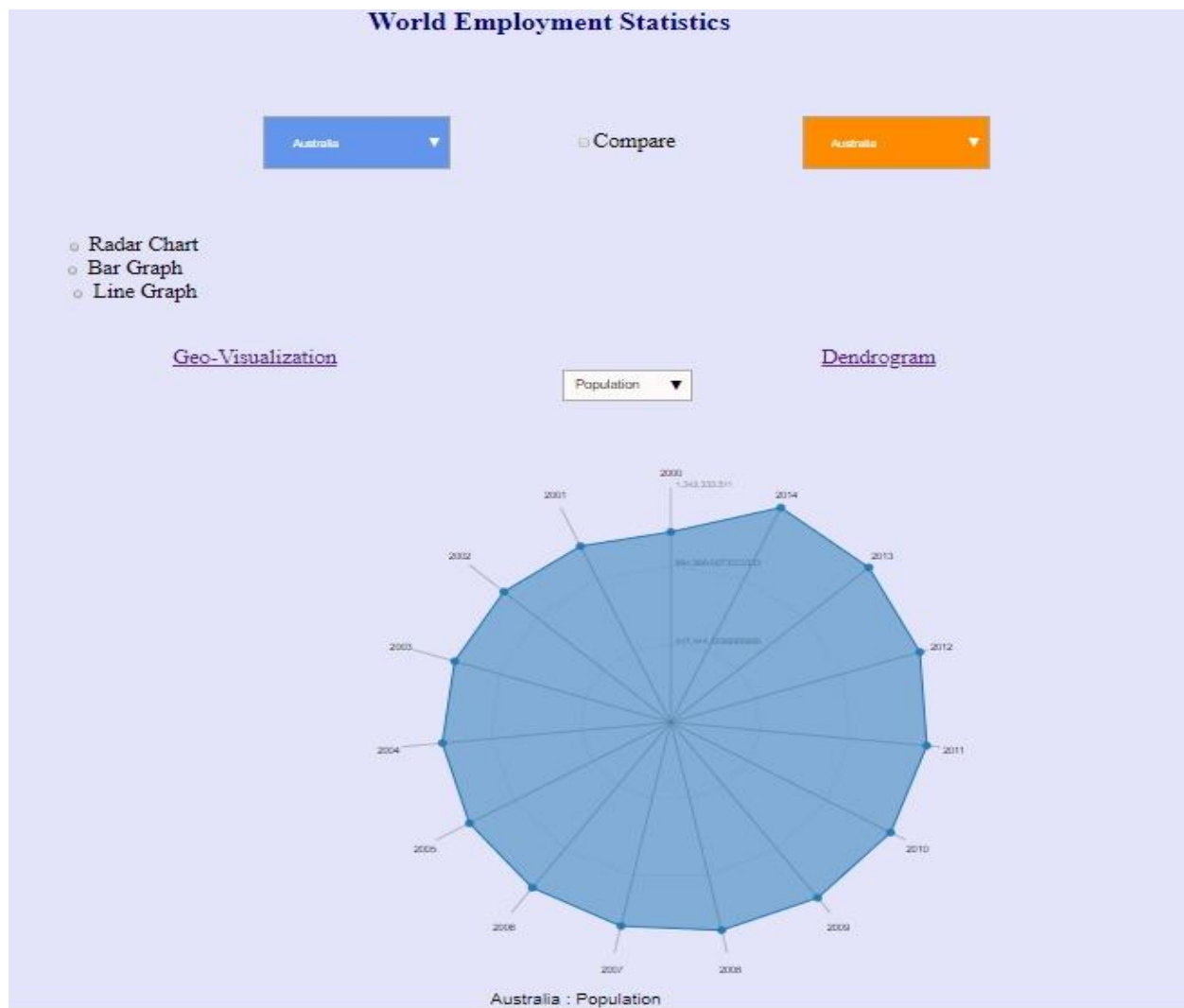
TECHNIQUE DESCRIPTION

For our project we have followed a normal data visualization process,

- Collecting the data
- Data preprocessing

- Plotting the data

Apart from only providing the visualizations of the data we also thought of providing an interactive user interface which can be visual treat to the user, so our initial idea on the web page was to provide a dropdown box for the user to select the country and provide five visualization options for the user and allow the user to select their desired visualization technique. We also provided a compare option button which on selection plots the data of selected country on the same visualization which makes it easy for the user to compare the statistics of two different countries. We have also provided hyperlinks to display geo visualization and dendrogram which on selection redirects to another page where the visualization is displayed, a back button is provided on the geo visualization and dendrogram page to return to the home page.



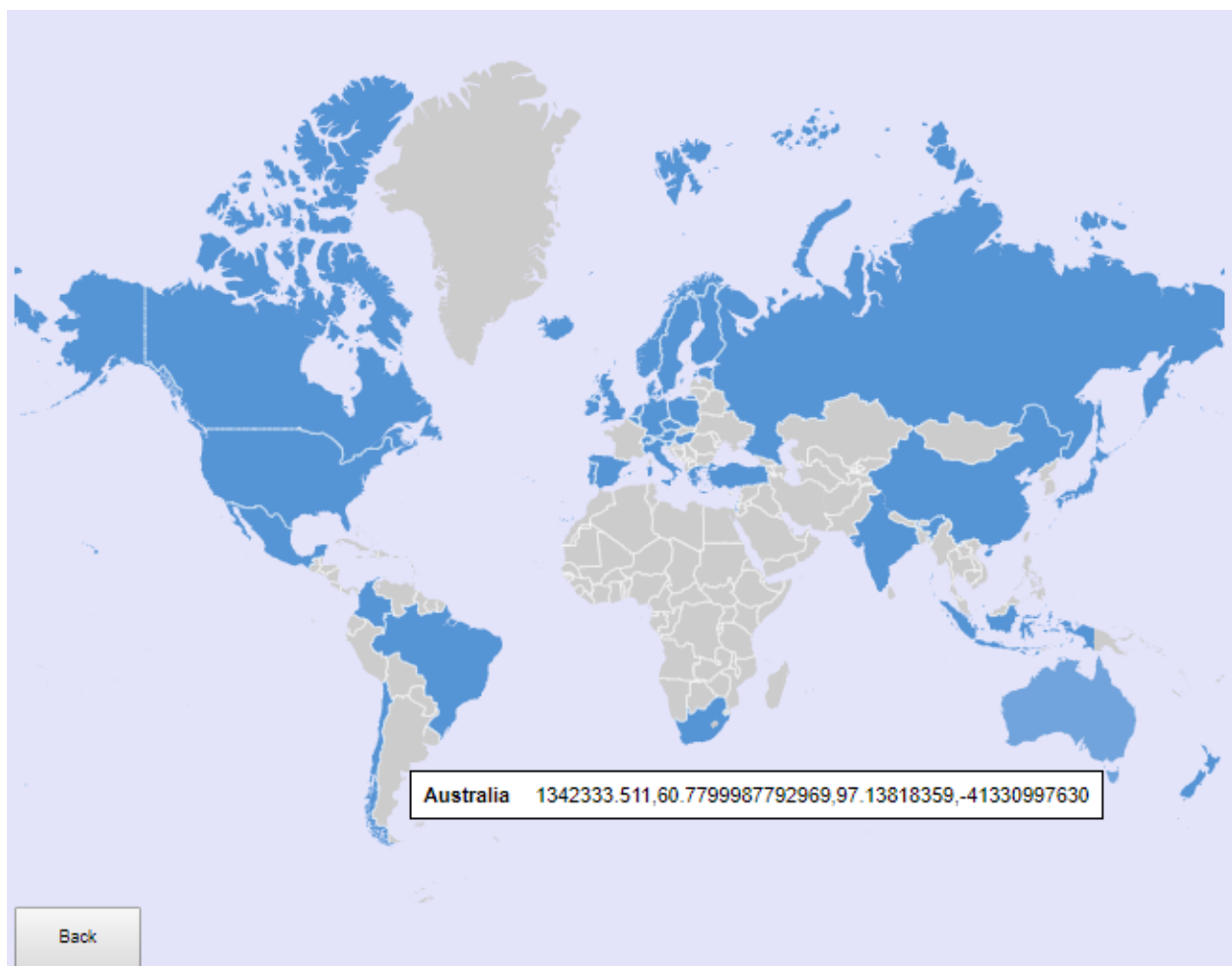
This screenshot depicts how our home page looks like as soon as the page is loaded. We have selected Radar Chart by default for the visualization and we have also chosen country Australia and the dataset to be Population as default.

VISUALIZATION TECHNIQUES & RESULTS

GEO VISUALIZATION

Geo-visualization is a visualization technique that can be used to plot geospatial data. In our project, for geo-visualization we have created a dataset in csv format which consists of the weighted average of the values of all the 4 attributes – population, education, employment, finance from the year 2000 to 2014 for all the 40 different countries. The visualization has been color coded which makes it easy for the user to identify the countries which are plotted on the world map. A function was also written to display the country name and the values of the 4 attributes of that country when mouse cursor is placed over it.

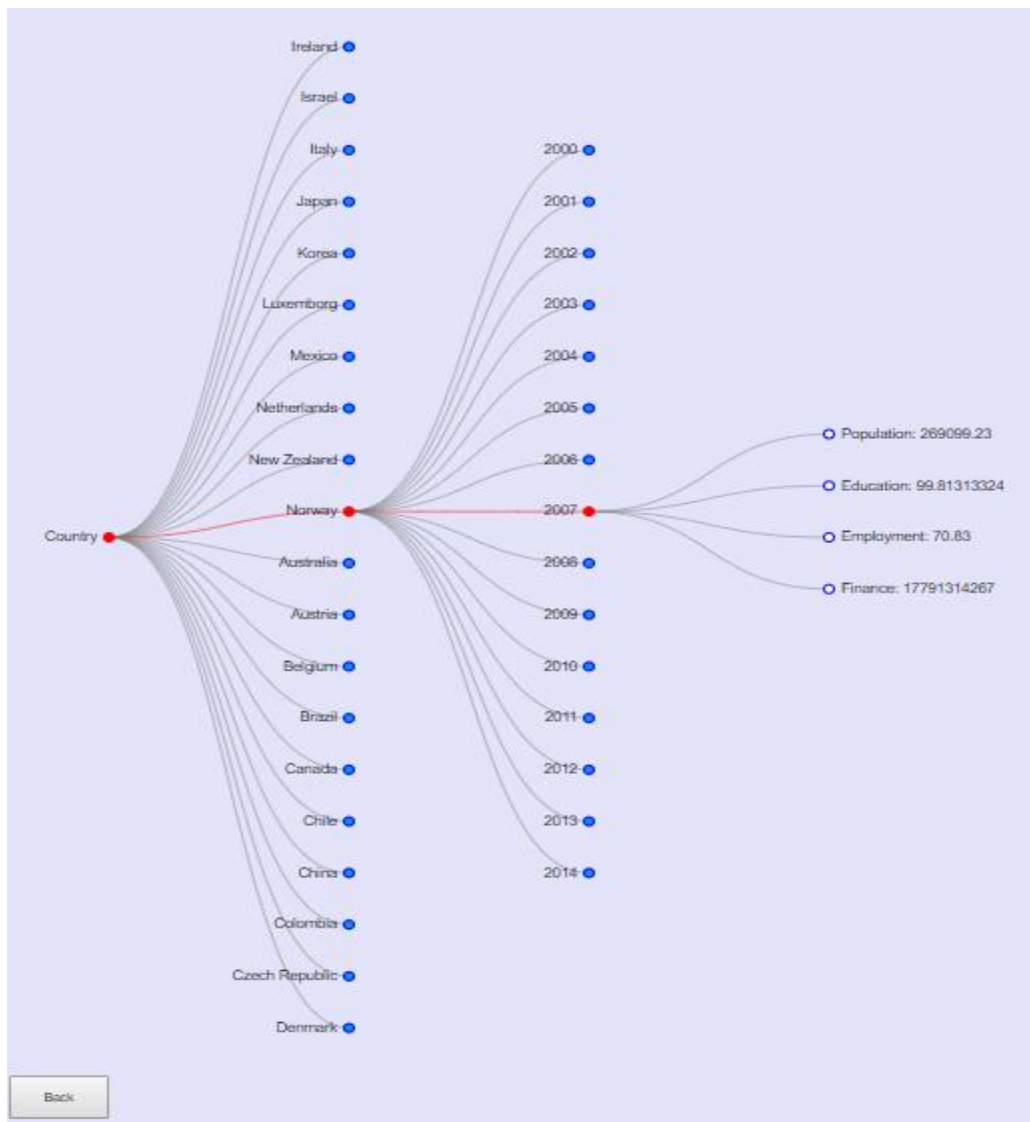
In the below screenshot the dark blue colors represents the countries plotted on the world map and when we hover mouse over a country, the data gets displayed.



DENDROGRAM

Like geo visualization even Dendrogram is used to plot the weighted average value of the 4 attributes for 20 distinct countries from the year 2000-2014. Dendrogram is a visual representation of compound correlation data. Dendrogram is a tree data structure representing different data in the form of branches. Country is the root node and the child nodes are the 20 countries present, these 20 countries have 14 years from 2000 – 2014 as child nodes and each year node will have 4 distinct values regarding Population, Education, Employment and Finance. Plotting all 40 countries in a dendrogram at the same time is not visually effective. Hence, we proceeded to visualize top 20 countries in the dendrogram. We have also provided the features like zoom, path highlight, expansion and compression for the dendrogram.

In the below screenshot the data of the country Norway for the year 2007 is being displayed in the Dendrogram visualization.



RADAR CHART

A radar chart is a graphical method of displaying the multivariate data in the form of 2-dimensional chart. In our project radar chart is used to plot the data of the selected country for a selected attribute from the year 2000 to 2014. On clicking the compare button and selecting the country to be compared, the data of the selected country will be plotted on to the radar chart, thereby making it easy for the user to correlate between the two countries. Correlation can be done for any attribute between any 2 countries.

Figure 1.a displays the radar chart visualization of Australia for the population dataset and Figure 1.b displays the radar chart visualization of Australia and Chile for the employment dataset where blue represents the Australia data and the orange represents Chile data.

Fig 1.a.

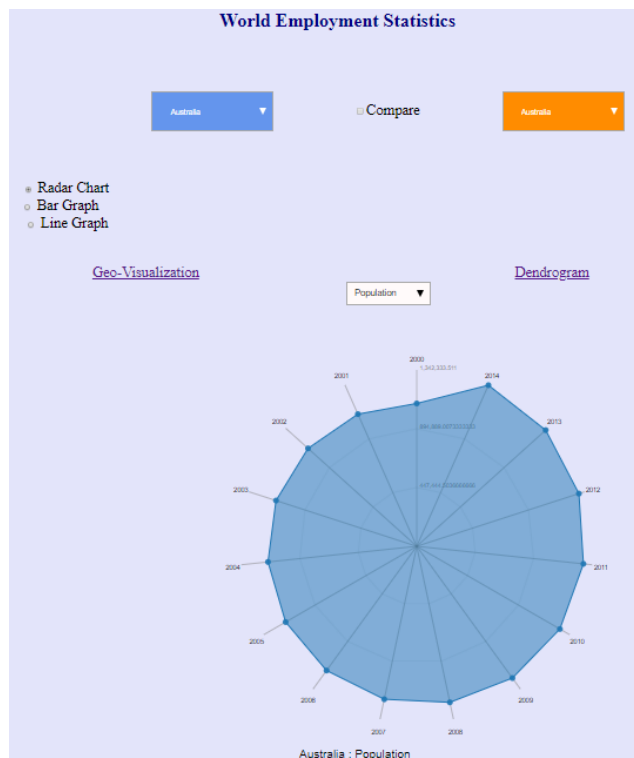
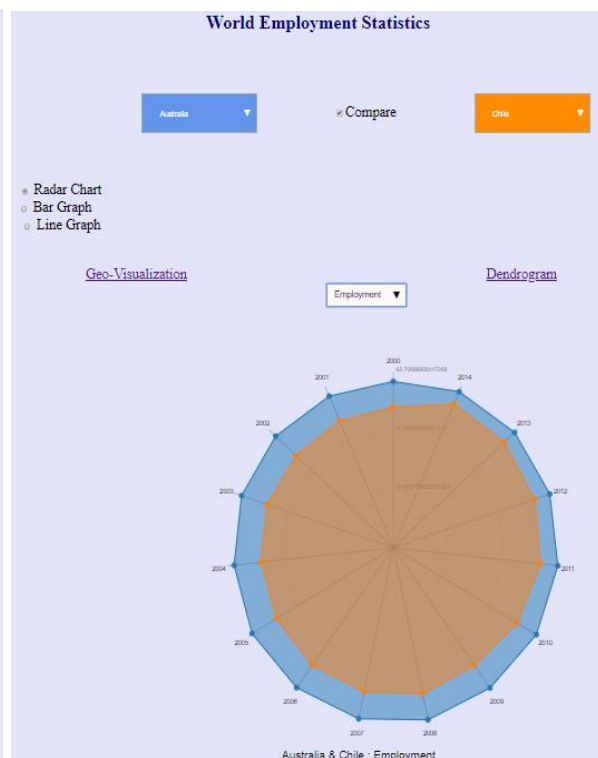


Fig 1.b.



LINE GRAPH

A line graph is a type of visualization technique which displays the information as a series of data points which are connected by straight line. In our project line chart is used to plot the data of any selected country for a selected dataset from the year 2000 to 2014. On clicking the compare button and selecting the country to be compared, the data of the selected country will be plotted

on to the line graph, thereby making it easy for the user to correlate between the two countries. Correlation can be done for any selected attribute between any 2 countries.

Below Figure 2.a displays the line graph visualization of Australia for the employment dataset and Figure 2.b displays the line graph visualization of Australia and Czech Republic for the population dataset where blue represents the Australia data and the orange represents data of Czech Republic.

Fig 2.a.

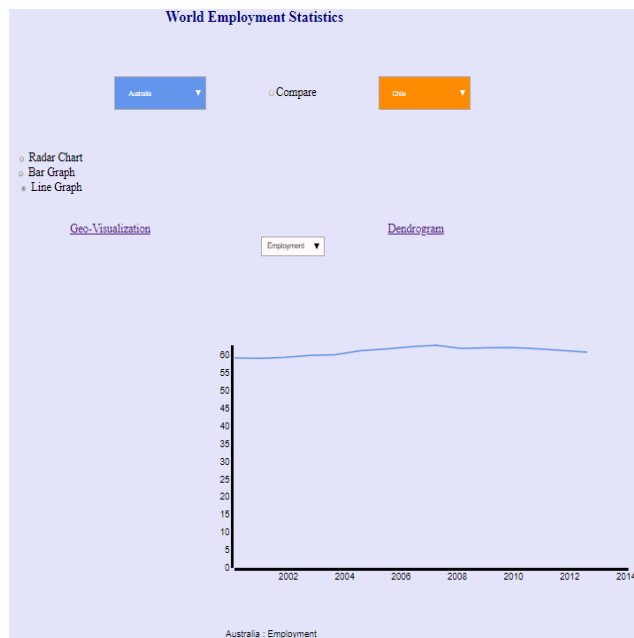


Fig 2.b.



BAR GRAPH

A bar graph is a pictorial representation of statistical data with rectangular bars with heights. In our project bar chart is used to plot the data of any selected country for any selected dataset from the year 2000 to 2014. On clicking the compare button and selecting the country to be compared, the data of the selected country will be plotted on to the bar graph, thereby making it easy for the user to correlate between the two countries. Correlation can be done for any selected attribute between any 2 countries.

Below Figure 3.a displays the bar graph visualization of Australia for the education dataset and Figure 3.b displays the bar graph visualization of Australia and Canada for the finance dataset where blue represents the Australia data and the orange represents data of Canada.

Fig 3.a.

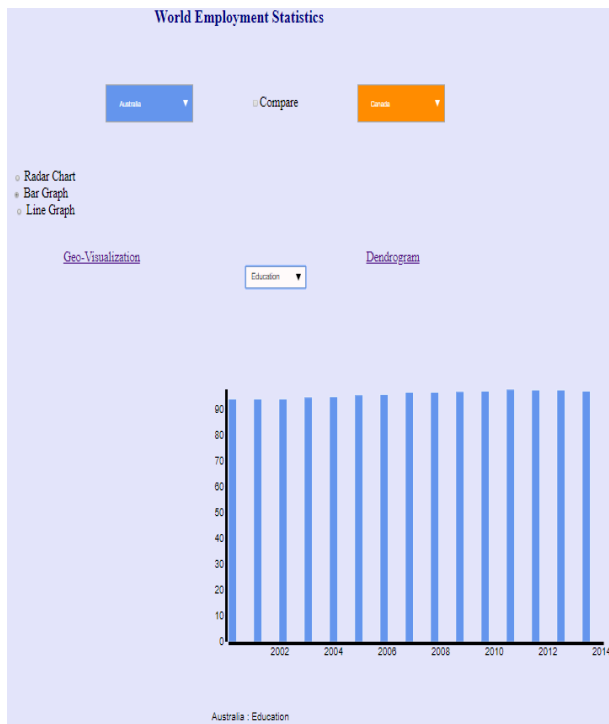
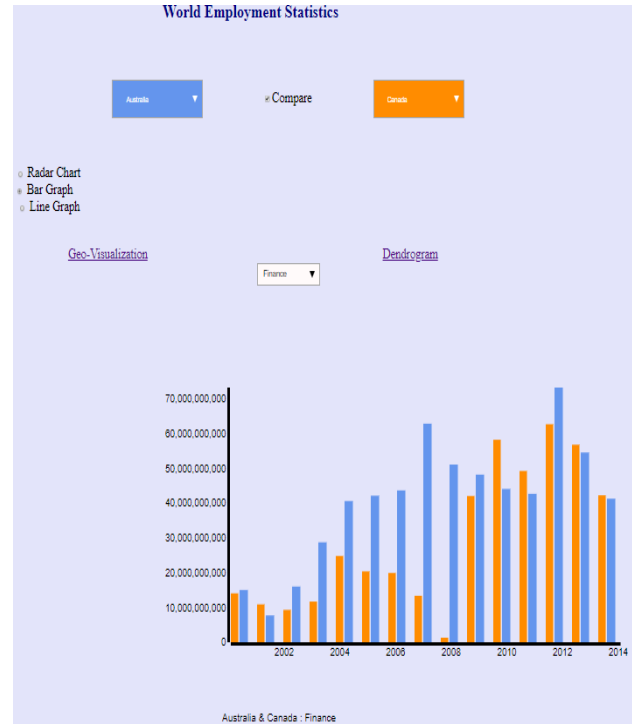


Fig 3.b.



INFERENCES

After collecting, analyzing and visualizing the 4 datasets – Population dataset, Education dataset, Employment dataset and Finance dataset which are dependent on each other we observed that for most of the countries increase in the population lead to increase the education rate from which we can infer that the countries globally are considering as “Education is must for all”. We also observed that as the employment rate increased the finance of the country was also high and similarly vice versa.

Although education levels are raised with increase in the population the employment levels declined accordingly, and the finance decreases as well. In some cases, the employment ratio was declined drastically despite decrease in the population for example in some years like 2009 the employment declined in many countries.

We can also infer that the amount of population in the developed countries is increasing steadily but, on a contrast, the developing countries rate of increase was larger than the ratio of increase in developed countries.

CONCLUSION

Finally, an interactive web page for employment statistics and its dependent factors have been represented through this project. For this project, we have put in the knowledge of web designing and styling along with the learned data visualization techniques in D3 programming. The analysis of about four distinct statistics have been done over a period of 14 years for 40 different countries in the world using five data visualization techniques.

REFERENCES

- [1] <http://www.oecd.org/>
- [2] <http://www.nonlinear.com/support/progenesis/comet/faq/v2.0/dendrogram.aspx>
- [3] <https://d3js.org/>
- [4] <https://github.com/d3/d3/wiki/Gallery>
- [5] <http://www.sthda.com/english/articles/28-hierarchical-clustering-essentials/92-visualizing-dendrograms-ultimate-guide/>