Best Visualization for migration data

Student Names: Anoop Raj, Akash Srivastava, Arpan Tiwari

Student Numbers: 18200172, 17201082, 18204032

1 Proposed Hypothesis (What Question Are You Asking?)

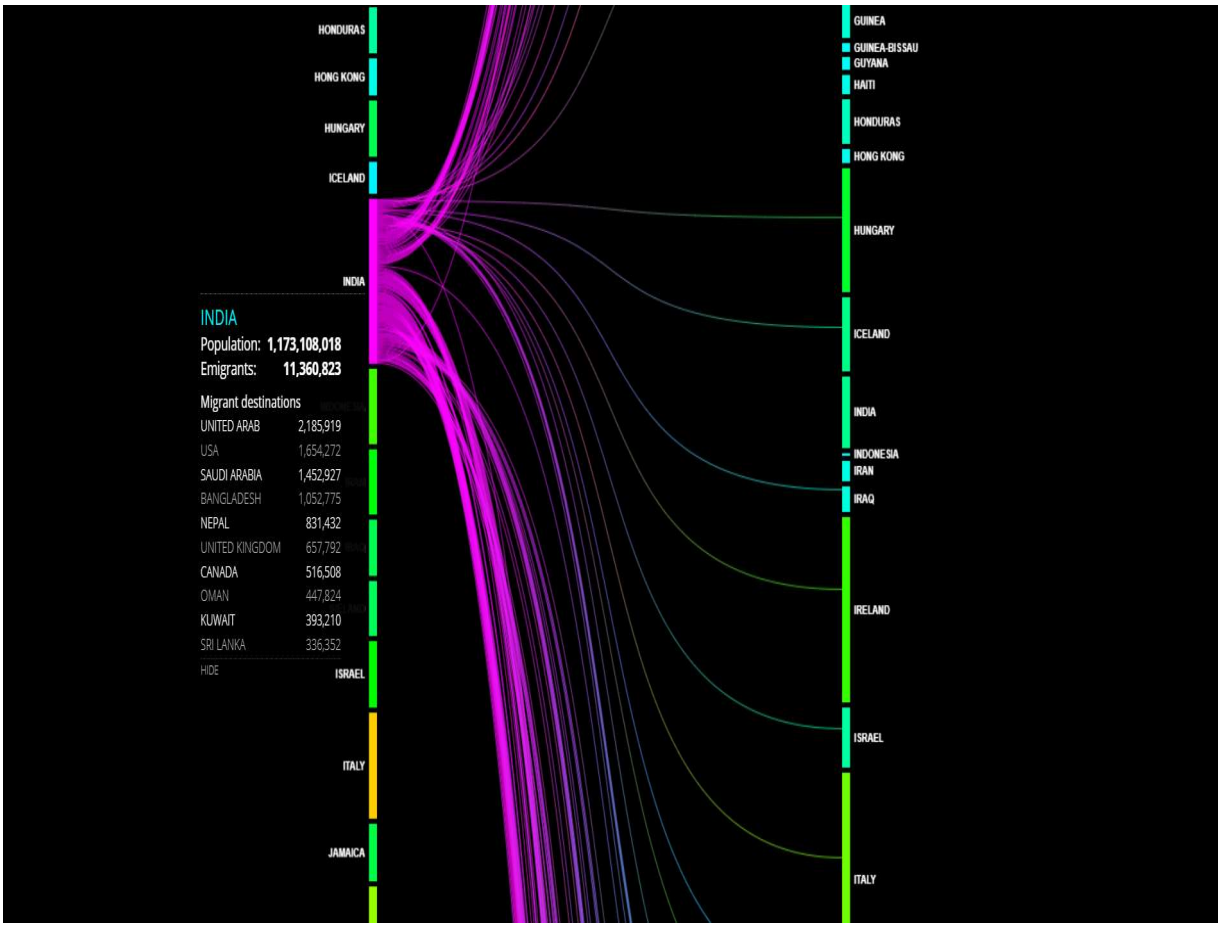
The primary purpose of this hypothesis is to assess which data visualization technique is good for people to visualize data of people migrating from India to other parts of the world due to various reasons. This visualization can help in evaluating the brain drain pattern of Indian citizens, which in turn tells a lot about the economic situation and standard of living in that country. Brain drain is a very serious issue for India, as every year a huge portion of skilled section of population moves out of India for better education or better job opportunities. This can be used to analyse which countries Indians are migrating to and what is it that they find better in that country than India. This can also be used by the Government of India to keep a check on the problem of brain drain by providing similar opportunities for education and employment to skilled population in India. This can also be used by the government of India to provide facilities or schemes at par with other countries to non-Indians to attract them to India.

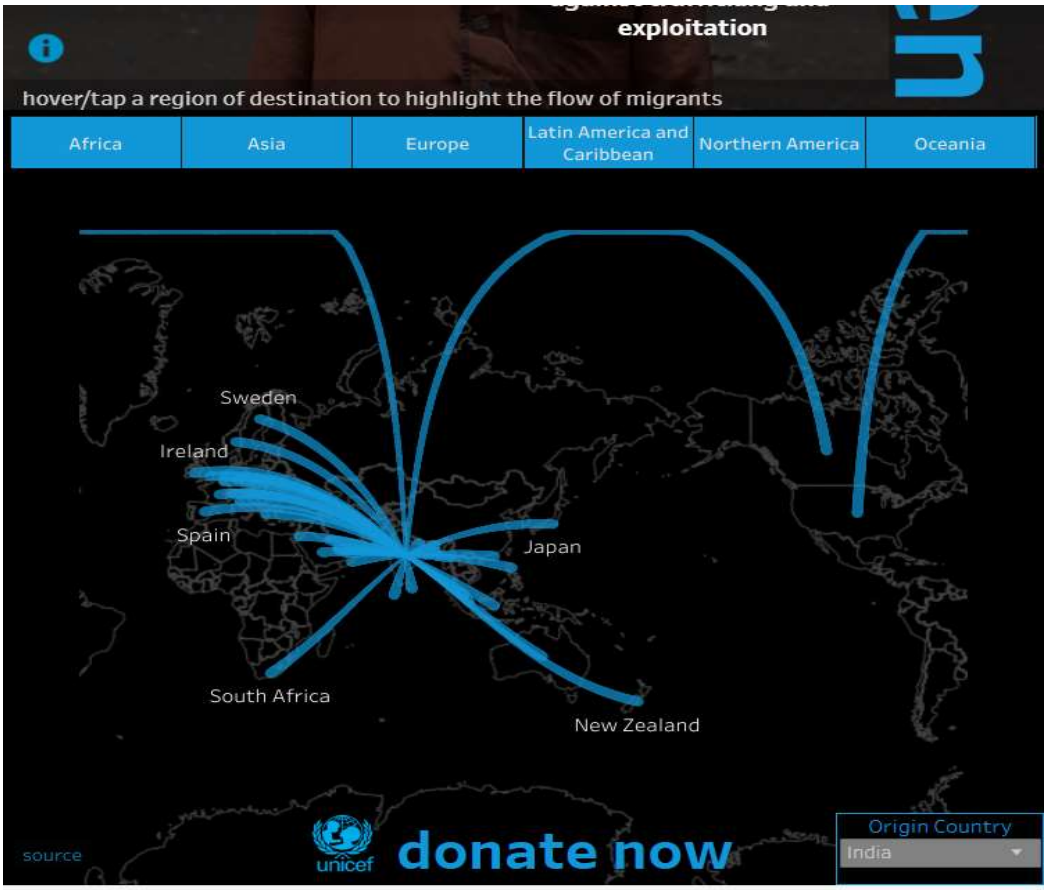
Considering the problem statement above we can visualize it using various types of graphs based on the number of people migrated, the year of migration, purpose and any specific chronological event. Multiple graphs can be used for this purpose starting with basic dual axis slope graph with lines shows the migration as well as other basic graphs such as bar chart and weighted pie chart. We can implement them using various colours and line densities. When we say migration or spatial analysis always Choropleth maps come into picture which is a thematic map where geographic regions are shaded, coloured and used with texture based on the migration. There are many variants of the Choropleth maps with heat maps which is used for the intensity of dataset and Proportional symbol maps which is used for the aggregated values. Animation can also be in the graphs which is indicated by lines and gradient of the graphs based on the Variables used in the data like time series analysis graph. Every graph for migration may not show a map, the best example for the same is Sankey diagram in which the width of the outward arrow is proportional to the flow of data; this graph can also be most relevant to our problem statement. When we are considering spatial graphs there are many attributes and variables that come into the picture which alone cannot be visualized in one graph, we might need to split the graph into many, and while visualizing data we need to handle these discrepancies as well. Graphs can be further enhanced for drill through or drill across of the data in temporal or hierarchical ways.

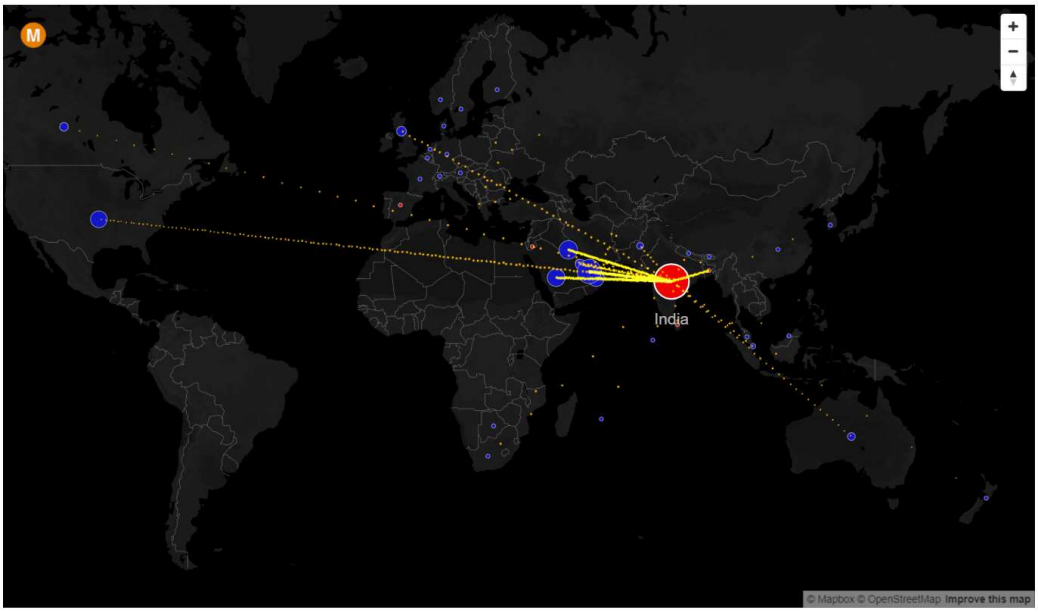
Each and every graph representation depends on each person's perception of how a decision or information in visualization is rendering. This hypothesis will primarily cover the problems related to choosing improper graphs for visualizing migration data and will try to identify the graphs that an render such information most accurately. The graphs we will be using for the experiment will be:-

- World map with pie chart and color intensity
- World map with pie chart
- World heat map
- Map with lines
- Bar chart
- Pie chart on XY plan
- Sankey diagram

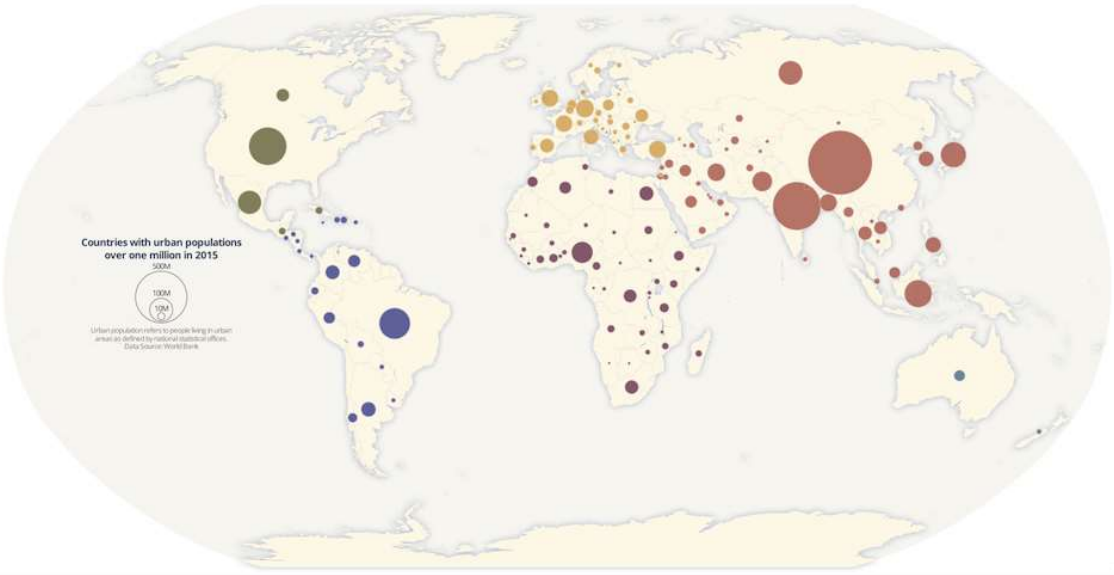
Below are some example graphs that visualize migration data pretty well:




exploitation
hover/tap a region of destination to highlight the flow of migrants
Africa
Asia
Europe
Latin America and Caribbean
Northern America
Oceania
Sweden
Ireland
Spain
Japan
South Africa
New Zealand
source
unicef
donate now
Origin Country
India


India
© Mapbox © OpenStreetMap Improve this map


In 2015 , 15,580,000 people born in India were living in other countries
< 1000
1k-10k
10k-90k
100k-240k
250k-490k
500k-990k
1m or greater


Countries with urban populations over one million in 2015
500M
100M
10M

2 Experimental Method

2.1 Overview

While performing the Experiment the participants of the experiment will be briefed about what the experiment is about and how it will be performed. After briefing, the experiment will be performed in a way that makes the whole process interesting for the participants so that they are pretty comfortable performing the experiment.

The participants for the experiment will be selected who have prior knowledge of geography and have performed similar experiments previously, as participants with prior knowledge will be able to perform experiment more efficiently and the chances of getting desired results from participants with prior knowledge is more as compared to participants without the knowledge.

In the experiment independent variables will be the different visualizations used to show the migration data of population from India. These visualizations will include pie chart with varying size, lines with varying thickness, colour intensity and density to depict the number of people migrating to a particular country. The dependent variables will be the accuracy with which a particular participant answers the question, the time taken to respond to the questions asked for a particular visualization and the feedback about how useful the particular visualization was to show the migration data. The confounding variables for the experiment will be the knowledge a participant gains from the previous questions and the prior knowledge of the trend about the migration data. To overcome the effect of these confounding variables, the participant will face different questions for different visualizations so that the effect of learning from previous questions is minimized and the data used in these visualizations will be random so that the bias of previous knowledge of data is also eliminated and that the participants see visualizations that are different from the general trend.

In order to get sufficient responses so as to come to a conclusion from the experiment, we will need a minimum of 100 participants to perform the experiment. Each participant will face all the questions designed, hence the experiment will be within group. Within group is chosen as the experimental method as it requires less number of participants to perform the experiment to get sufficient number of responses, in order to come to the conclusion. Also, the learning effect from the previous question is minimized by asking different questions for different visualizations, therefore within group experimental method seems to be the better approach.

2.2 Data collection

The hypothesis is to assess which data visualization technique is good for visualizing data of people migrating from India to other parts of the world. Therefore in the experiment we will collect the responses made by the participant for a particular visualization. In addition to the responses given by them, we will also record the time taken by them to answer each questions pertaining to each visualization. The accuracy and the time taken by the participant will tell how efficient the visualization is for depicting migration data. In addition to this, we will consider the time taken by the participant to assess

whether the response made by the participant is genuine or random. For example, if the participant responds to a question in few seconds then it will be considered as a random response and it will not be considered for our analysis. The above listed measurements will be the objective measurements and the subjective measurements will be the responses to the questions that will help us compare the two visualizations and the responses to the questions based on the Likert scale, where we will assess how sure the participant is about the answer to the question.

The data collected will tell how easy it was for the participant to read the visualization and answer the questions asked. Also, asking questions on different types of visualization (MAPS and NON MAPS) will enable us to conclude which type of visualization is better to visualize data of people migrating from India to different parts of the world.

2.3 Selected subjects

For the purpose of our pilot experiment, we will be selecting a few fellow students from our batch in the University who would be representing the target audience for whom our hypothesis is intended. The sample size we have chosen for the pilot study is 3. The participants will have basic understanding of Geography (i.e. should be able to interpret basic political maps) and should also have a decent understanding of English, as it is the medium of communication in the experiment. The small number of participants will primarily help us test out our experimental setup so that we can fine tune any parameters and then go ahead with conducting the full experiment.

For our full experiment, we will be sourcing 100 subjects from the Amazon Mechanical Turk platform, which should give us a more accurate result than the pilot experiment.

Our target audience is not a niche category of people with very specific backgrounds, but it is basically any person with a decent understanding of maps. For participating in our experiment, the subjects will also need to have some command over the English language. As our target audience is quite general, we believe a sample size of 100 participants should be a fair representation of the same. Also, the subjects will be chosen from multiple countries so as to negate the effect of visual parameters (e.g. colours) having cultural connotations specific to any country, as well as to make the feedback more universal.

2.4 Data analysis

We will be assessing the collected data on the following three parameters:

Accuracy: We will measure the accuracy of specific answers submitted by the subjects pertaining to the data being represented in the charts. This accuracy information will tell us if the subjects are able to interpret the data from each chart correctly or not, and whether one chart is easier to interpret than others.

For example, we would ask the subjects a question on a qualitative/quantitative feature represented in the chart and on the basis of their answer we would be able to conclude how effective the current chart for answering our question is.

Opinion: The feedback questions that we're additionally asking the subjects pertaining to various charts will give us a direct and high level idea of the subjects' personal preferences regarding which chart they would prefer using for answering specific types of questions.

For example, we would ask the subjects if they would prefer seeing a pie chart or a Sankey diagram for understanding a particular qualitative/quantitative piece of information about the data, and their answer would help us identify which chart is better out of the two.

Time: Lastly, we will also assess the time taken by the subjects to answer each question for each chart. This information will tell us how quickly the audience is able to interpret each chart, and whether a particular type of chart is too difficult to interpret than other charts.

We will analyse the above three parameters for the purpose of answering the proposed question in our hypothesis, since the information will allow us to do a comparative analysis of all the charts for representing immigration data on the basis of the subjects' ease of understanding of the charts, as well as how correctly and quickly can the charts be interpreted by them. We would thus be able to conclude which type (or types) of charts are most suitable for answering questions pertaining to immigration data.

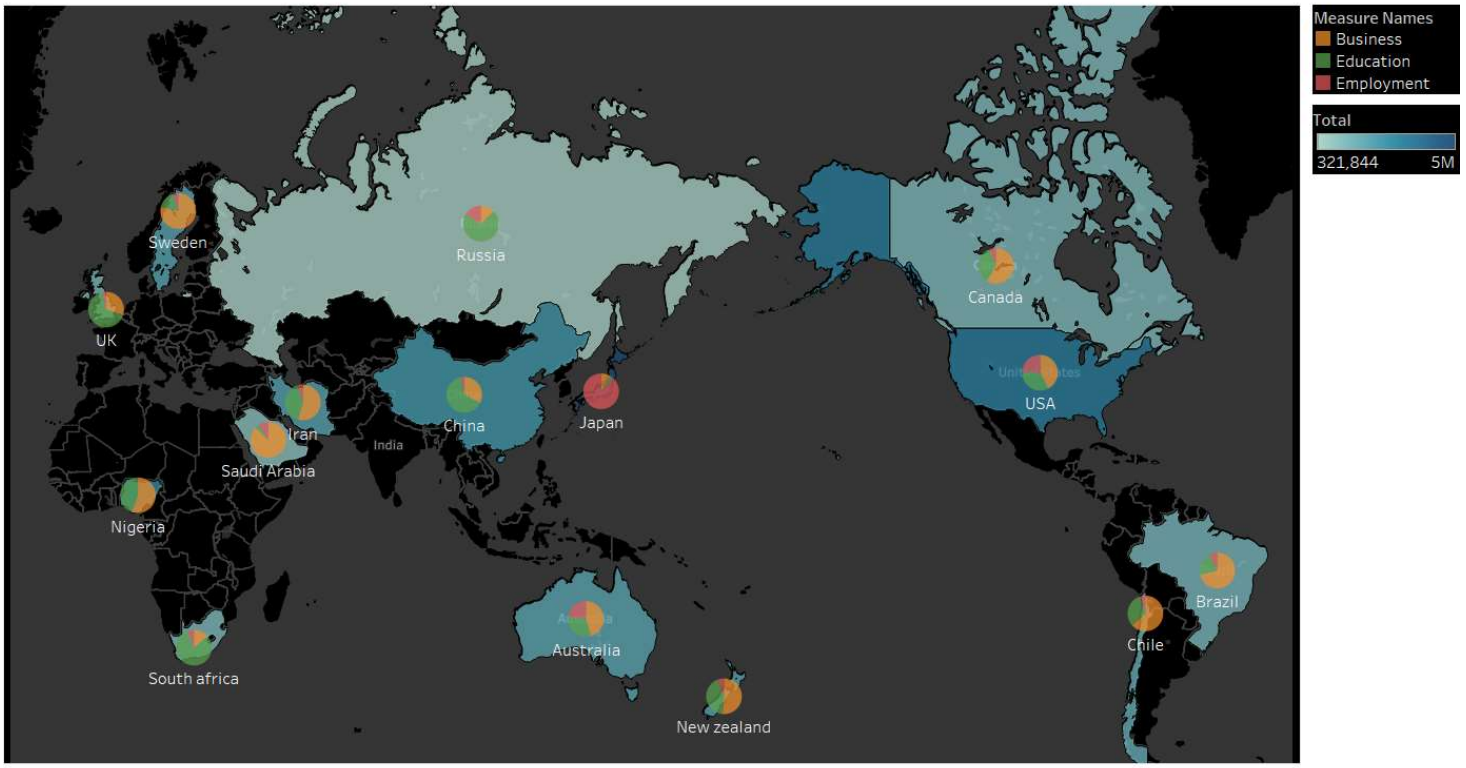
2.5 Practical setup

Practical setup for the experiment is considered with random values of the dataset of migration of population from India, and the purpose of the migration due to employment, education and business. We have used Tableau and Sankey Built in tools to build graphs with available data. There were 7 graphs grouped into maps and non-maps, for each graph relevant 3 question were asked. Google survey was used for the survey in which Graph and question sequence were uploaded. Each participant was given the online link of the survey in which they had to provide their names and answer all the questions which were mandatory. All participants needed laptops to use the link and were given instructions accordingly. Each participant were given clear instructions to analyze the graph properly as migration Data from India is very predictable and participants might ignore the data visualization .So the dataset that we created were random and graph needs to be analysed thoroughly to answer the question provided . Each graph had enough details which enabled the participants to analyse the visualization and any concerns were addressed if they find the questions or graphs were not clear .There were 3 participants and each participant was given timeslots in which they had to be present for the experiment. Everyone had the laptops in which the links were provided and manually experiment conductors noted the time taken for each graph and their responses are captured by Google surveys response section for further analysis.

3 Data Visualisations

Below are the visualization along with the questions asked for the each visualization:

Indian Immigration Around the World



Map based on Longitude (generated) and Latitude (generated) and Latitude (generated). For pane Latitude (generated): Color shows sum of Total. Details are shown for F2. For pane Latitude (generated) (2): Color shows details about Business, Education and Employment. The marks are labeled by F2. The view is filtered on Longitude (generated), which keeps non-Null values only.

Which Country received the highest number of Indian immigrants for business?

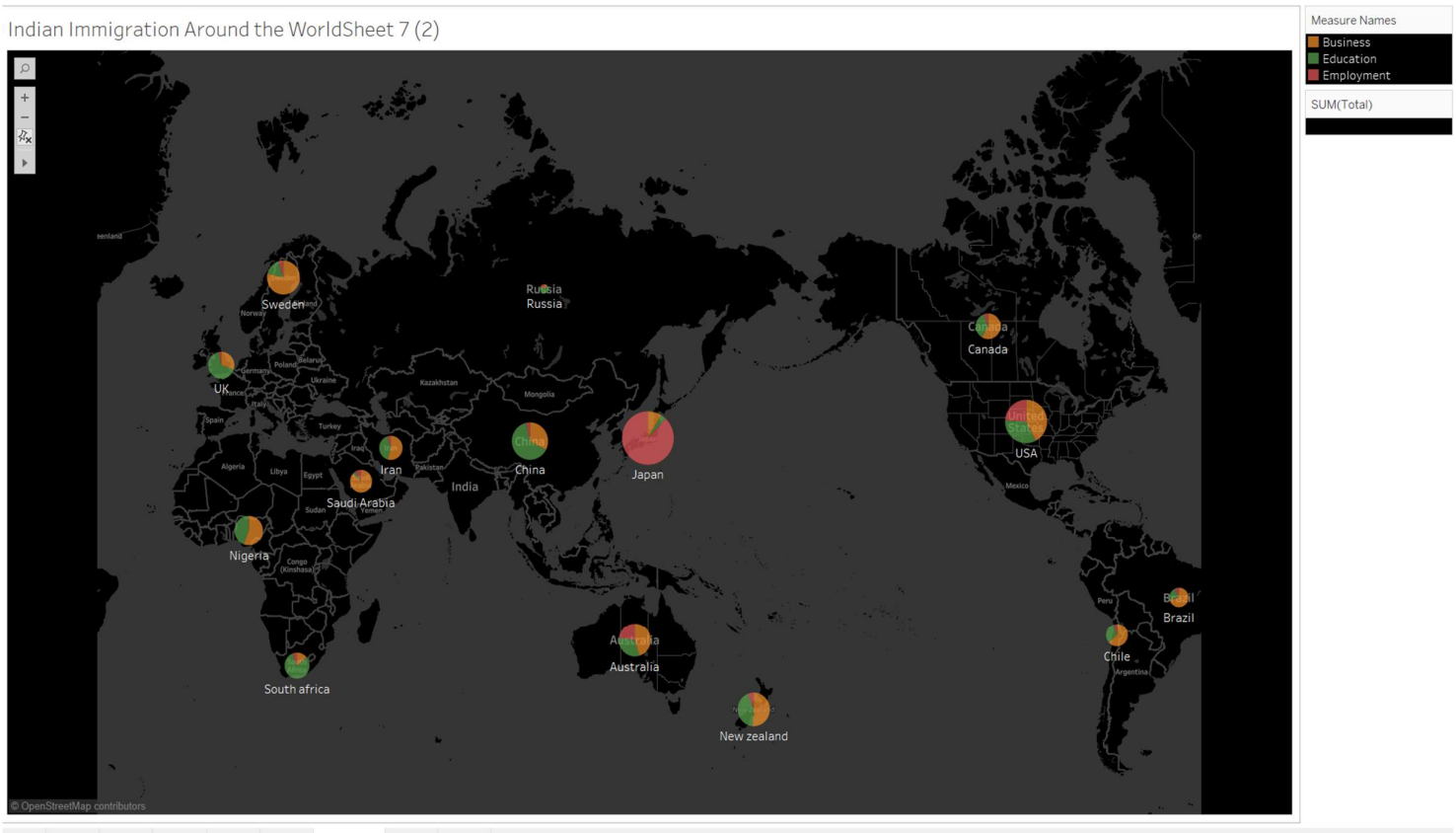
- Japan
- Sweden
- USA
- China

Which Country received the highest number of Indian immigrants overall?

- USA
- Australia
- China
- Japan

China received more number of Indian students in comparison to Iran.

- Strongly Agree
- Somewhat agree
- Somewhat disagree
- Strongly disagree
- Can't say



Note: The size of the pie charts indicates total number of Indian immigrants in a particular country

Which Country received the lowest number of Indian immigrants for employment?

- Russia
- Nigeria
- Chile
- Brazil

Between New Zealand and Australia, which country received the least number of Indian students?

- Australia
- New Zealand
- Both are equal

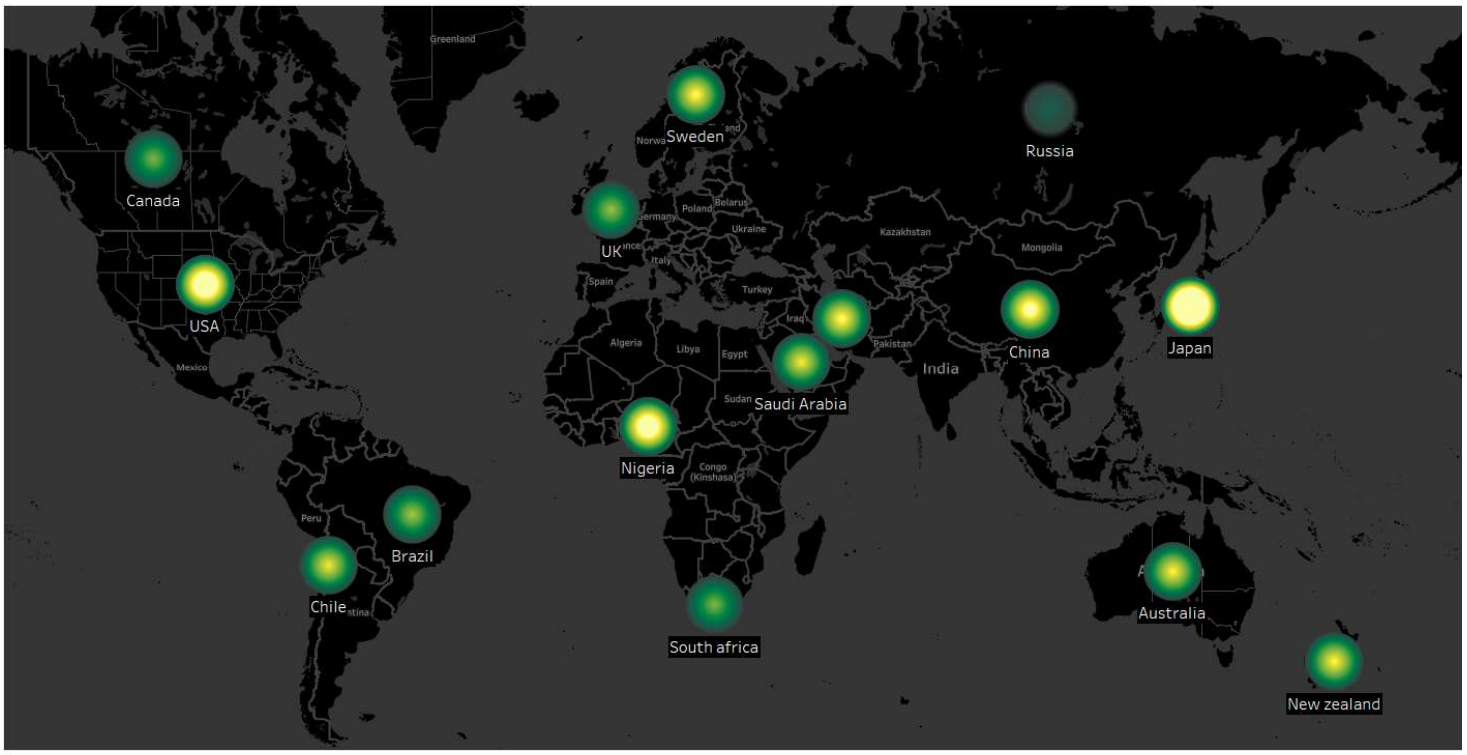
Out of Saudi Arabia, Iran and Canada, Iran received the most number of Indian workers.

- Strongly Agree
- Somewhat agree
- Somewhat disagree
- Strongly disagree
- Can't say

Which of the previous 2 map charts was more effective in conveying immigration information?

- First
- Second
- Both were same

Heat map of Indian immigration around the world



Map based on Longitude (generated) and Latitude (generated). Color shows sum of F6. The marks are labeled by Country.

Which country received the maximum number of immigrants out of the following?

- Chile
- Brazil
- Sweden
- New Zealand

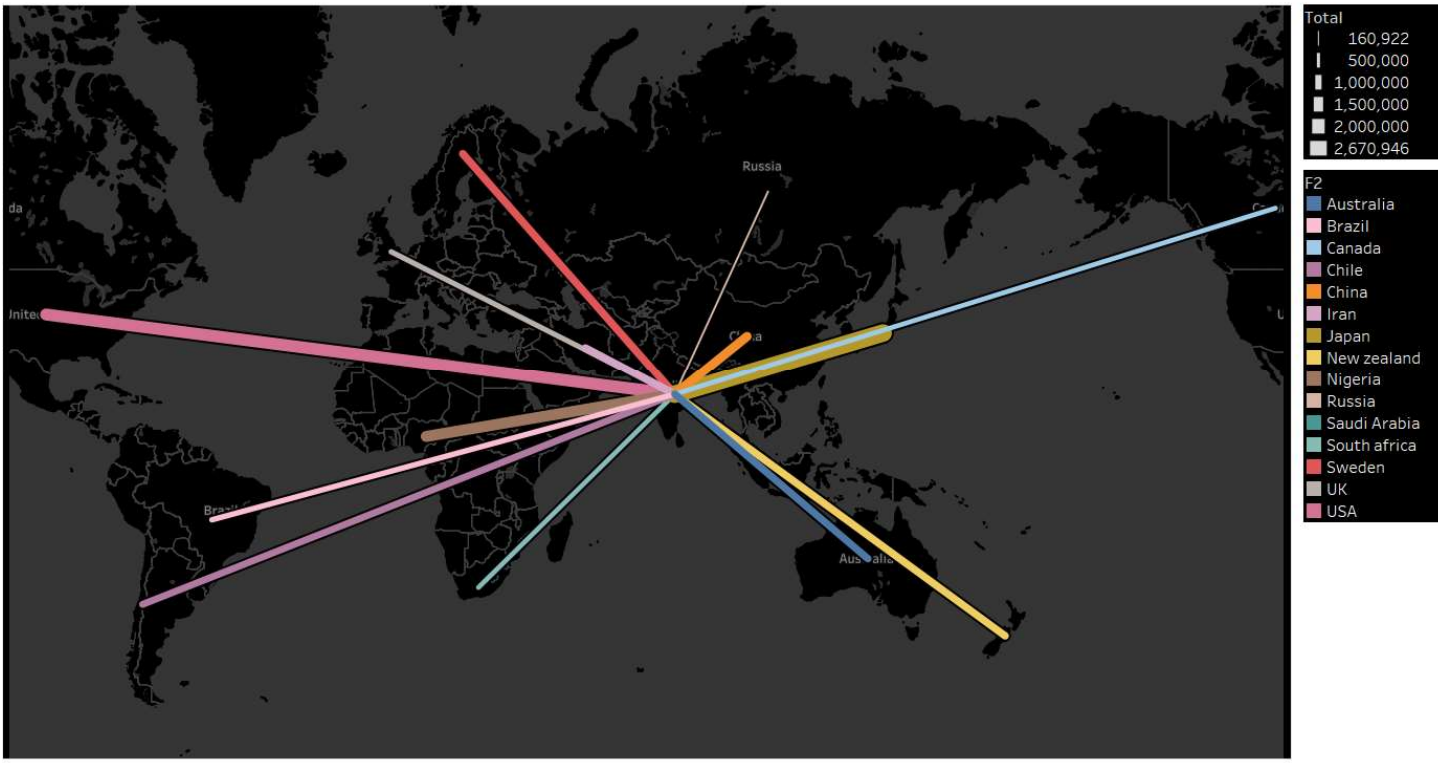
Which country received minimum number of immigrants out of the following list?

- Canada
- UK
- South Africa
- Japan

Between Australia and Saudi Arabia, Australia has the higher number of Indian immigrants.

- Strongly Agree
- Somewhat agree
- Somewhat disagree
- Strongly disagree
- Can't say

Indian Immigration Around the World



Map based on F2 (Sheet12) and F1 (Sheet12). Color shows details about F2. Size shows sum of Total. Details are shown for F2.

Which country received maximum number of immigrants out of the following?

- Nigeria
- Sweden
- Sweden
- Canada

To which hemisphere did most Indians immigrate?

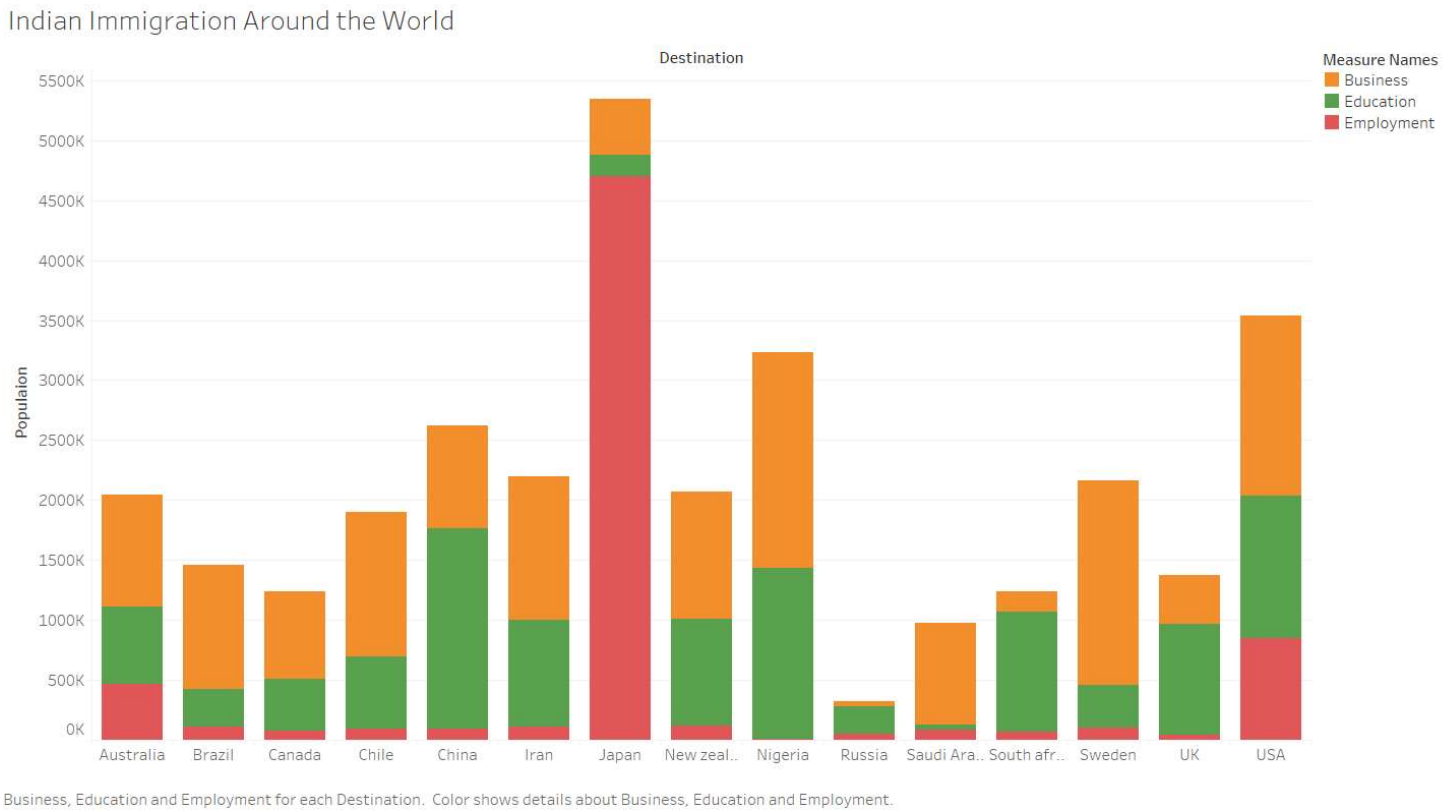
- Northern hemisphere
- Southern hemisphere
- Both are equal

Does distance play an important role in the immigration pattern of Indians?

- Strongly Agree
- Somewhat agree
- Somewhat disagree
- Strongly disagree
- Can't say

Which of the previous 2 map charts was more effective in conveying immigration information?

- First
- Second
- Both were same



Which country received third highest number of immigrants overall?

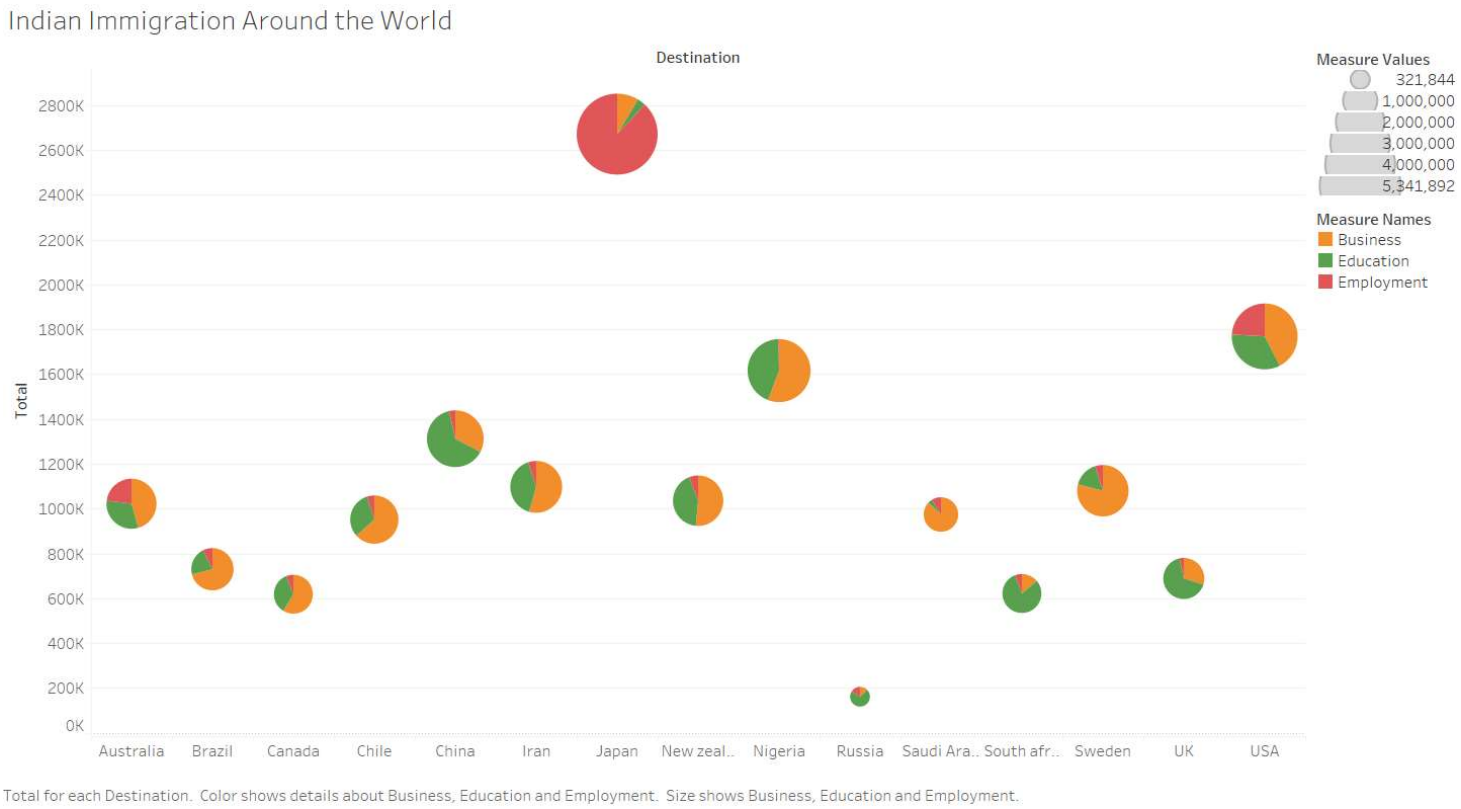
- Brazil
- Nigeria
- USA
- Japan

Which country received least number of immigrants for employment from the following?

- Brazil
- Iran
- New Zealand
- Sweden

The proportion of Indian students and employees is equal in New Zealand.

- Strongly Agree
- Somewhat agree
- Somewhat disagree
- Strongly disagree
- Can't say



For which purpose did most number of Indians immigrate?

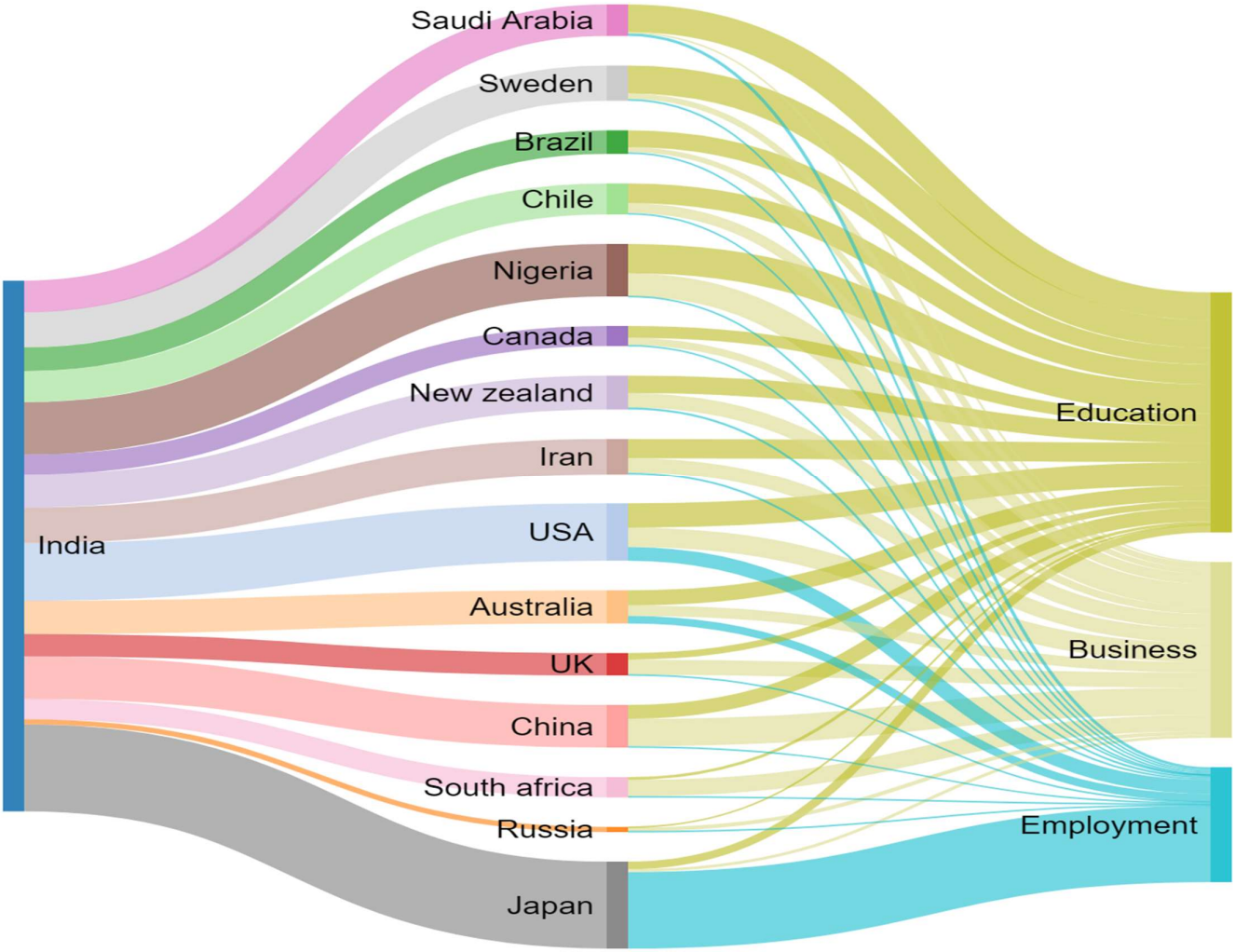
- Business
- Education
- Employment
- Can't say

Which Country received the lowest number of Indian immigrants for employment?

- Russia
- Nigeria
- Chile
- Brazil

Between New Zealand and Australia, which country received the least number of Indian students?

- Australia
- New Zealand
- Both are equal



Between USA and Nigeria, which country received more number of Indian immigrants?

- USA
- Nigeria
- Can't say

Which country received almost equal number of immigrants for each category?

- USA
- Australia
- Russia
- All of the above

For which purpose did most number of Indians immigrate?

- Business
- Education
- Employment
- Can't say

Which of the previous 2 map charts was more effective in conveying immigration information?

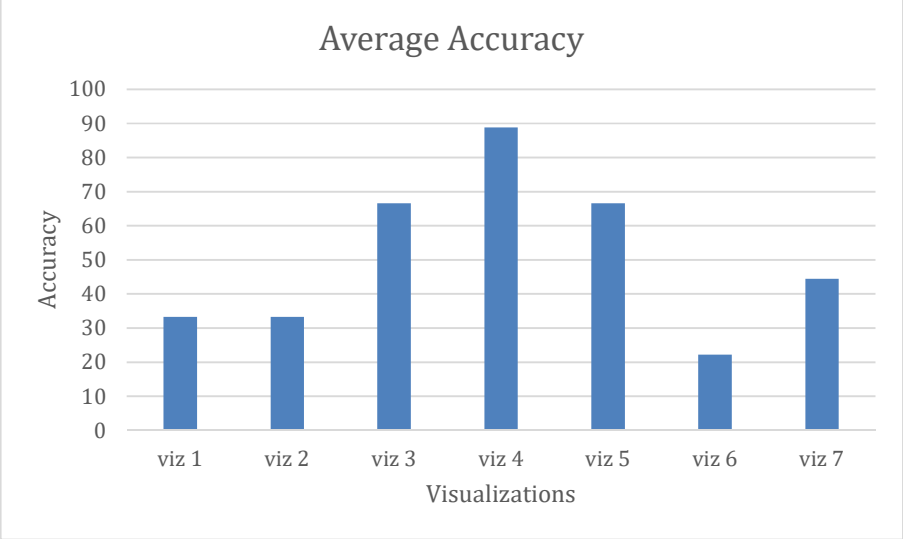
- First
- Second
- Both were same

4 Pilot Experiment

4.1 Data Analysis

As mentioned earlier, we analysed our collected data on the following three parameters:

- **Accuracy:** We measured the accuracy of the charts on the basis of the number of correct answers submitted by the subjects for each visualization. These were the results:



The above bar chart shows the average accuracy obtained in the pilot experiment for the each chart. Here it would not be good to consider the above accuracy as the number of participants for pilot experiment were very less to come to some concrete conclusion.

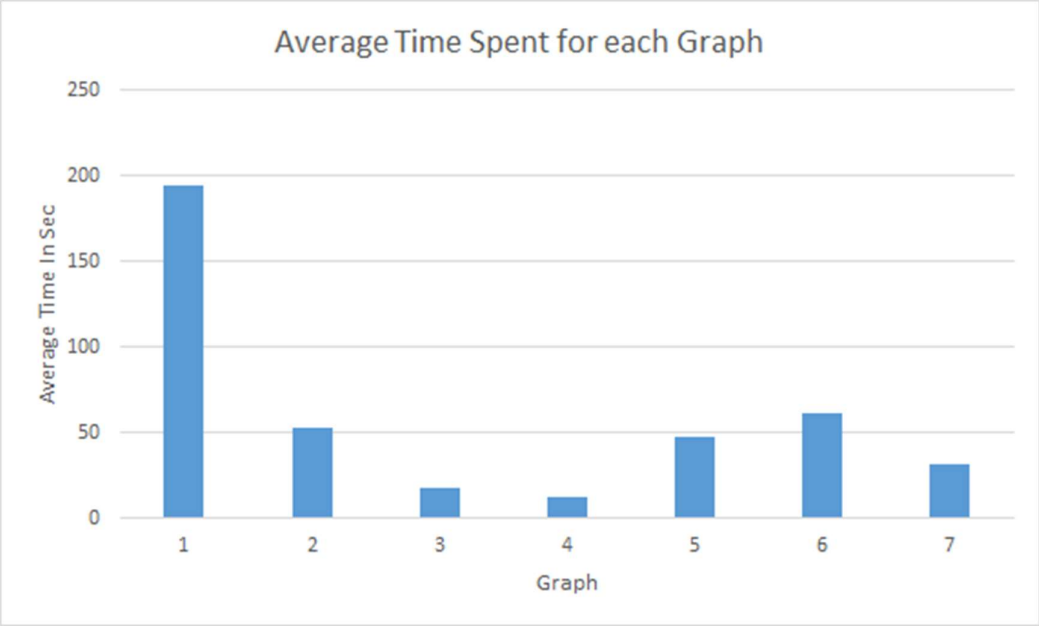
- **Audience Feedback**: We took the feedback from the audience about which chart they found most effective in visualizing migration data and the results are as follow:
  - Between World Map with pie chart and color intensity and World Map with pie, all three participants found the first one better.
  - Between heat map and map with lines, two out of three found heat map more effective.
  - Between bar chart, pie chart and Sankey diagram, all three participants found bar chart most effective in showing migration data.

  The last question we asked was out of maps and non-maps, which was better to visualize migration data and all three participants found non maps more informative in showing migration data.

- **Time taken:** The last measure we analyzed was the time taken by each participant to answer questions pertaining to each chart. This measure should

give us an idea about the level of difficulty in interpreting various graphs, i.e. lesser the time taken, the easier it is to interpret the chart, and vice versa.

Following were the time recordings:



From the above graph, it appears that chart 1 (i.e. map chart with color intensity and pie chart) took the most amount of time to interpret, while chart no. 4 took the least amount of time. On an average, non-map charts took optimal amount of time to interpret as compared to map based charts.

Having said that, since the sample size was too small (i.e. 3), we would need a much higher number of participants in the actual experiment to come to a concrete conclusion.

4.2 Reflections

We can construe the following inferences from the pilot experiment that we conducted for identifying the best visualization for migration data:

- In our first graph participants had to interpret intensity of the population migrated first and then based on the pie chart they had to interpret highest number of migration due to business. Time taken for this graph was highest and results were not very accurate so graph could have been more lighter in much darker room which would have resulted in better accuracy.
- Another major difference that we found from this experiment was for the overall migration .As per our data Japan had highest number of immigrants than USA , however while interpreting the graph since Japan's area in the graph is small compared to USA, participants failed to identify the intensity of the colour and all answered as USA. So pie chart with total population as length of the radius yielded better results .
- Resolution of the graphs used in google survey was not very clear and participants had to either spend more time on the graphs or with the existing low resolution graphs had to approximate the answers. Google survey may not be the best choice for images as participants had to zoom the maps in few questions . So a better survey tool with better resolution would have made this experiment more feasible .

- In the view of this experiment's result we need more participants as answers were biased and Accuracy rate of the maps were less compared to non-maps even though maps were visually more appealing .Non-maps had better results and very optimal time was taken as compared to maps . Simple interpretation like overall population migration were better suited for maps as per the feedback but when parameters increased non maps like Sankey and bar charts had better results .
- From this experiment we can conclude that choropleth/thematic representation of data depends on area of the country/region which will be easy to represent less parameters and visually more appealing however for complex interpretation with more parameters non maps yields more accuracy . Time taken for the maps were less Compared to non-maps as non-maps interpretation needs basic understanding for data visualization. Participants who participated had good knowledge of the data visualization to interpret the graphs but in general maps are more visually engaging with less parameters and non-maps are more accurate for complex interpretation.

References

- http://peoplemov.in/#f_IN
- https://data.unicef.org/resources/migration-refugee-data-visualizations/
- http://metrocosm.com/global-immigration-map/
- https://www.weforum.org/agenda/2017/03/5-facts-that-show-india-is-a-migration-superpower
- https://carto.com/blog/popular-thematic-map-types-techniques-spatial-data/