

Data Visualization

MSIS 2629

Individual Project

Living Abroad:
Choose the Country best for you!

Submitted By

Ekta Ratanpara
W1275326

Table of Contents

Problem Statement	3
Introduction	3
Audience	3
Claim.....	3
Raw Data	3
Data Cleaning and Transformation using Pandas in Python	5
Data Cleaning.....	6
Data Transformation	6
Data Categorization	7
Visualizations	8
Storytelling using dashboards.....	11
Timeline	13
Discussion and conclusion	14
Conclusion:	14
Critique	14
Reference.....	14

Problem Statement

The number of Indian tech workers living abroad has steadily grown over the past decade as more people are being interested in exploring another culture or finding better career prospects and lifestyle. This project will help you find a country that is most appropriate for you based on your need.

Introduction

Globalization has revolutionized the global economy. It has expanded employment opportunities in the international level and opened the door of free mobilization for skilled tech workers. Plenty of new job opportunities are being generated every day in the developed countries because of their technological advancements, and growing industrialization. Immigration is an inevitable part of globalization and in the perspective of global economy it has a great importance.

According to data from United Nations, 244 million people were recorded as international migrant and around 16 million people were Indians who moved out of country. This project will focus on which country is good option for you being an Indian tech worker.

People move to other countries for many different reasons. The reason can be classified as economic, social, political and environmental. we will compare all the countries based on Indicators like Income, Safety, Life expectancy, Innovation, Salary etc. The questions while comparing countries are “Better at what?”, “Better for what?” and “Better for whom?”. There is no universal answer to this question. We must drill down into the details to determine the answer.

Audience

The visualizations in this project are created for Indian tech people who want to move to another country for various reasons like better job opportunities or happy life.

Claim

For Indian tech people, is united states the best country to move to?

To find the best country for tech people, I searched and came up with the 10 most developed countries which provides good job opportunities and has good indicator values which Indian tech immigrants are looking for.

According to numerous news articles and my personal experience, indicators most people look while moving to a different country are Employment rate, Long-term unemployment rate, Life satisfaction, Life expectancy, Political Stability/Absence of Terrorism, Housing expenditure, Salary/Median pay, Ease of starting a new business, Global Innovation Index and R&D Spend - % of GDP.

Raw Data

I collected raw data from three different sources each consisting different indicators.

1) OECD ([Organization for economic co-operation and development](#))

To compare the countries, I used better life index data from the OECD which has data for all the countries based on different indicators like Income, housing, jobs etc. Below is the list of the indicators I chose to compare countries

Indicator	Meaning
Air pollution	Air pollution is important factor that affects the life expectancy and have adverse effect on health of the people.
Employees working very long hours	Important aspect of work-life balance and shows the amount of time a person spends at work. Long work hours may impair personal health and increase stress level.
Employment rate	Work has clear financial advantages, yet having a Job additionally helps people boost confidence, and adapt new aptitudes and capabilities. Societies with high levels of employment are wealthier, politically steady and more advantageous.
Housing expenditure	Housing expenditure take up a large share of the family spending plan and speak to the biggest single expenditure for some people and families
Life expectancy	Life expectancy is widely used to measure the health of people health, it only considers the length of people's life and not their quality of life.
Life satisfaction	Life satisfaction shows the happiness of the people.
Long-term unemployment rate	Long-term unemployment can have a large negative effect on well-being of human and result in a loss of skills, further reducing employability.
Time devoted to leisure and personal care	The amount and quality of leisure time is important for people's overall well-being, and can bring additional physical and mental health benefits
Water quality	Access to clean water is fundamental to human well-being. Managing water to meet that need is a major – and growing – challenge in many parts of the world. Many people are suffering from inadequate quantity and quality of water.

2) The Global Innovation Index ([Link](#))

The Global Innovation Index (GII) is an annual ranking of countries by their capacity and success in innovation. The indicators examined include factors such as research and development, technology, human capital, tax policy, trade barriers and intellectual property protections. Some indicators I used to compare countries based on global innovation index are

Indicator	Meaning
Ease of starting a new business	Ease of doing business index is an index created by the World Bank Group. Higher rankings (a low numerical value) indicate better, usually simpler, regulations for businesses and stronger protections of property rights

Global Innovation Index	Global Innovation Index is a combination of more than 20 indicators which we use to compare and rank countries
Political Stability/Absence of Terrorism	Political Stability and Absence of Violence/Terrorism measures perceptions of the likelihood of political instability and/or politically motivated violence, including terrorism
R&D Spend - % of GDP	Gross domestic spending on R&D is defined as the total expenditure (current and capital) on R&D carried out by all resident companies, research institutes, university and government laboratories, etc., in a country.

3) Salary: Software engineering salary by country ([Link](#))

As the audiences of this project is tech workers, I gathered the salary data for software engineer in all the country as salary can highly motivates the people to move to different country along with different other factors and can provide better life to people.

I choose 3 different indicators as these indicators shows the average annual income of the people live in the country, median annual pay of software engineer and ratio of median annual pay of software engineer and average annual income which shows how much more software engineer is getting paid compare to all other jobs.

Indicator	Meaning
Median annual pay of software engineer	Median salary is a common statistic used by organizations such as the Bureau of Labor Statistics and the Social Security Administration as an estimate of the amount of money earned by a typical worker.
Average annual income	Average income earned per person in each country in a specified year. It is calculated by dividing the area's total income by its total population.
Ratio of median annual pay of software engineer and average annual income	This ratio shows the difference in the salary of software engineer and non- software engineer people

Data Cleaning and Transformation using Pandas in Python

Data cleaning is very important part of any visualization project. After extracting the raw data collected in CSV format from above mentioned sources, I used python to clean the required data for the project and transform it into the format which provide better insights of the data.

Please refer to [this Jupyter notebook](#) for the complete cleaning and transformation script. Below are the steps followed for data cleaning and transformation of the data.

Data Cleaning

- 1) **OECD:** Source data had details for more than 100 countries and around 15 indicators. A different file was available for data of each year. Steps followed to clean and combine historical data in one file are as below:
 - a. Read csv files as pandas data frame
 - b. Filter data for selected countries and indicators and add a year column
 - c. Merge data frames generated for every year into one and write to a csv file
- 2) **Immigration:** Source data file had the data of all the immigrants moved to all the countries across the globe. Hence data was filtered for Indian people moving to selected countries. (Note: As data for Indian tech people moved to different country were not available, count indicating all Indian moved to different country is considered.)
- 3) **Global innovation Index:** For each indicator, a separate source file was available containing data for all countries in the world. So data was filtered for selected countries and then merged into a single output file containing all indicators.

Data Transformation

To compare all countries on various indicators, the value of each indicator was needed in same scale as all indicators had values in different units and scales. For example, Air quality data was in ppm while salary data was in thousand USD. In addition to different units/scales, there were two types of indicators – one with positive values and one with negative values. To maintain consistency in analysis, I changed negative indicators to positive and scaled values of all indicators to have value between 10 and 100. The scaling is done based on relative distance between two values of the same indicator. Below are the snippets used for scaling the OECD dataset. In similar way, immigration, global innovation index and salary data is also scaled.

```
import pandas as pd
from sklearn import preprocessing

def scale_df(df):
    x = df.values #returns a numpy array
    # scikit-learn MinMaxScaler is used to scale the values between 10 and 100.
    # Min value will get 10 and max will get 100.
    # All other values will be assigned based on relative distance
    min_max_scaler = preprocessing.MinMaxScaler(feature_range=(0.1, 1))
    x = x.reshape(-1,1)
    x_scaled = min_max_scaler.fit_transform(x)
    return x_scaled * 100

df = pd.read_csv("./processed_files/oecd_clean_data.csv")

# indicators => key = indicator, value = is key a positive indicator?
indicators = {
    'Housing expenditure': False,
    'Employment rate': True,
    'Long-term unemployment rate': False,
    'Air pollution': False,
```

```

    'Water quality': True,
    'Life expectancy': True,
    'Life satisfaction': True,
    'Time devoted to leisure and personal care': True,
    'Feeling safe walking alone at night': True,
    'Self-reported health': True
}

oecd_years = [2013, 2014, 2015, 2016]
df_dict_indicator = {}
df_with_scaled_value = []
for indicator, is_positive in indicators.iteritems():
    for year in oecd_years:
        tmp_df = df.loc[df['Indicator'] == indicator].loc[df['year'] == year]
        if not tmp_df.empty:
            if not is_positive:
                original_values = tmp_df['Value'].copy(deep=True)
                tmp_df['Value'] = tmp_df['Value'].apply(lambda x: 100-x)
            tmp_df['ScaledValue'] = scale_df(tmp_df['Value'])
            tmp_df['ScaledValue'] = tmp_df['ScaledValue'].round(2)
            if not is_positive:
                tmp_df['Value'] = original_values
            df_with_scaled_value.append(tmp_df)

df_final = pd.concat(df_with_scaled_value)
df_final.to_csv("./processed_files/scaled_oecd_clean_data.csv", index=False)

```

After scaling, I merged all the dataframes containing indicator, actual value and scaled value to create a master dataframe.

Data Categorization

Finally, I assigned categories to all indicators in the master dataframe file using python based on the table below and wrote to a CSV file:

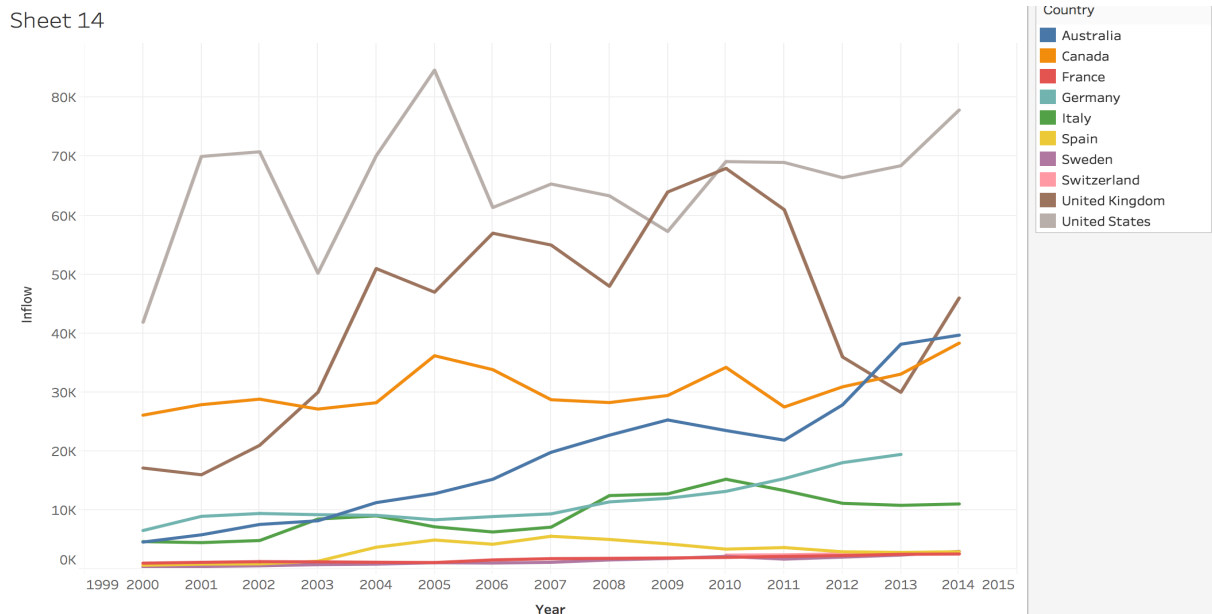
Indicator	Category
R&D Spend - % of GDP	Business & Innovation
Ease of starting a new business	Business & Innovation
Global Innovation Index	Business & Innovation
Immigration Inflow	Ease of Immigration
Immigration Nationality	Ease of Immigration
Employment rate	Employment
Long-term unemployment rate	Employment
Air pollution	Environment Health
Water quality	Environment Health
Median Pay	Financial Health

Housing expenditure	Financial Health
Time devoted to leisure and personal care	Happiness
Life satisfaction	Happiness
Self-reported health	Health
Life expectancy	Health
Feeling safe walking alone at night	Safety
Political Stability/Absence of Terrorism	Safety

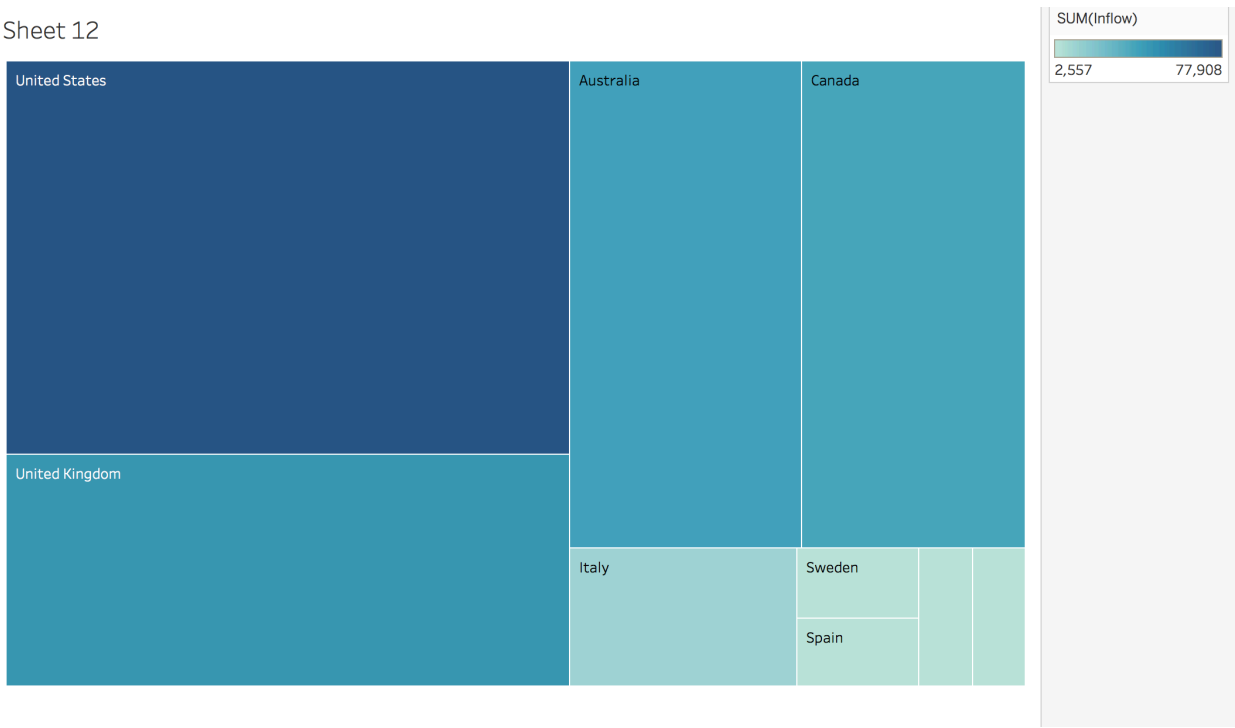
Visualizations

The idea behind this project was to create a story which helps tech people from India to choose best country on the indicators important to consider while moving to another country and shows how well country is doing in comparison of other countries.

I started with the selection of Top 10 countries popular among tech people based on several articles and created visualization of how many Indian people migrated to those country in past few years. This graph was the visual exploration as my actions were unknown at the time and I might change my goal.

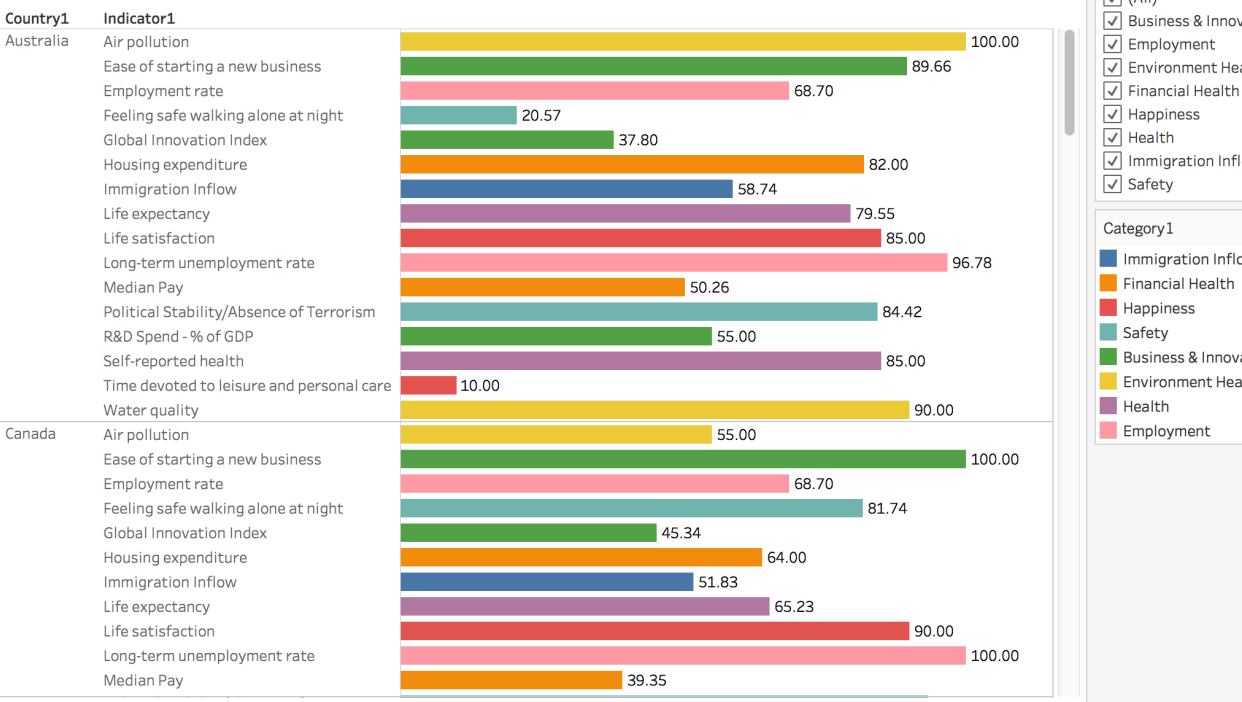


Above graph clearly shows which country is the popular among Indians based on immigration inflow but as I had to display data for 10 countries and year over year analysis was not required so I decided to go with the Tree map which is more clean and easy to understand as size of the rectangle and darkness of color shows the Inflow of Indian immigrants.



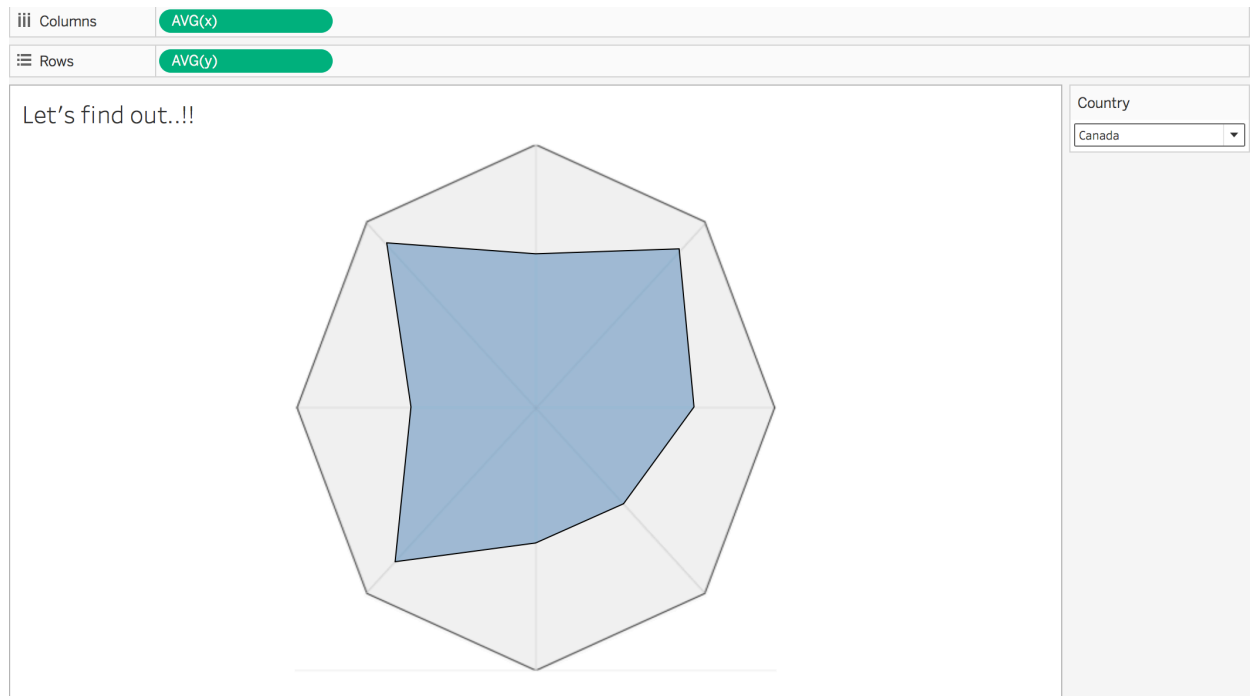
Second, I started with visualization which provides measurement for every indicator for the country and created a bar chart, bar chart worked fine except...

How good is the country ?



it didn't show the relative comparison of every indicator with respect to the reference or the highest value. Inspired from the blog "Use radar chat to compare dimension over several metrics" written by

Jonathan Trajkovic ([Blog Link](#)), I decided to create a radar chart to compare and display multiple variable at the same time with reference background image. When I started implementing radar chart, I had around 12 different indicators to compare and the final radar chart was quite messy and confusing so, I decided to group the indicator in categories for e.g. Air quality and water quality both are related to environment so I assigned them to environment category and used categories instead of individual indicator in the final graph.



The final radar graph with 8 categories was better looking and we can easily tell, how country is doing in all the categories by looking at the size of the polygon and comparing each category's axis with reference axis provided by the reference image in the background. "Bigger the polygon better is the country"

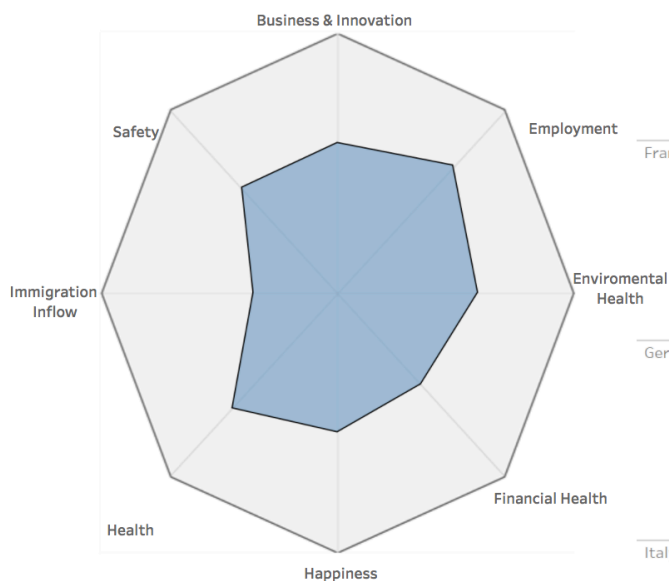
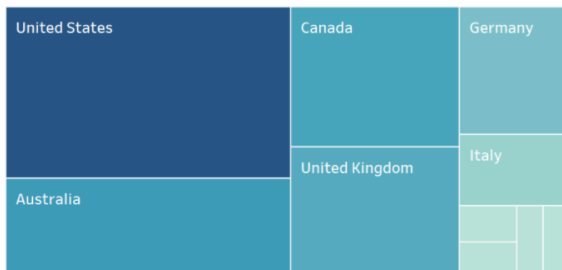
To create a radar chart, we need to create X and Y axis values. Below formula will change based on the number of dimensions want to display in the chart. To create Y-axis, replace SIN with COS.

x


```
case [Category]
when "Business & Innovation" then [Scaled Value] * SIN(0*2*PI()/8)
when "Employment" then [Scaled Value] * SIN(1*2*PI()/8)
when "Environment Health" then [Scaled Value] * SIN(2*2*PI()/8)
when "Financial Health" then [Scaled Value] * SIN(3*2*PI()/8)
when "Happiness" then [Scaled Value] * SIN(4*2*PI()/8)
when "Health" then [Scaled Value] * SIN(5*2*PI()/8)
when "Immigration Inflow" then [Scaled Value] * SIN(6*2*PI()/8)
when "Safety" then [Scaled Value] * SIN(7*2*PI()/8)
end
```

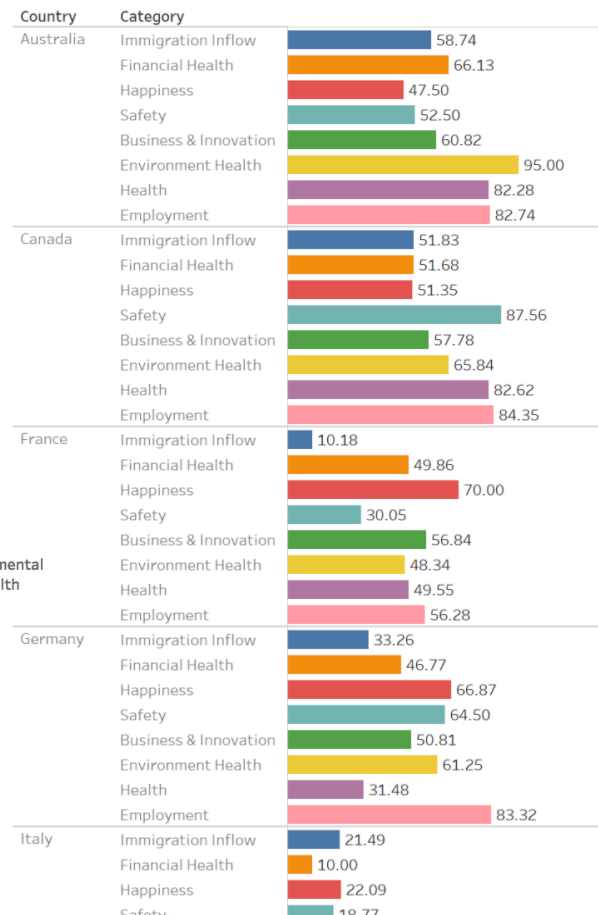
Storytelling using dashboards

Which country is popular among indian people?



How good is the country ?

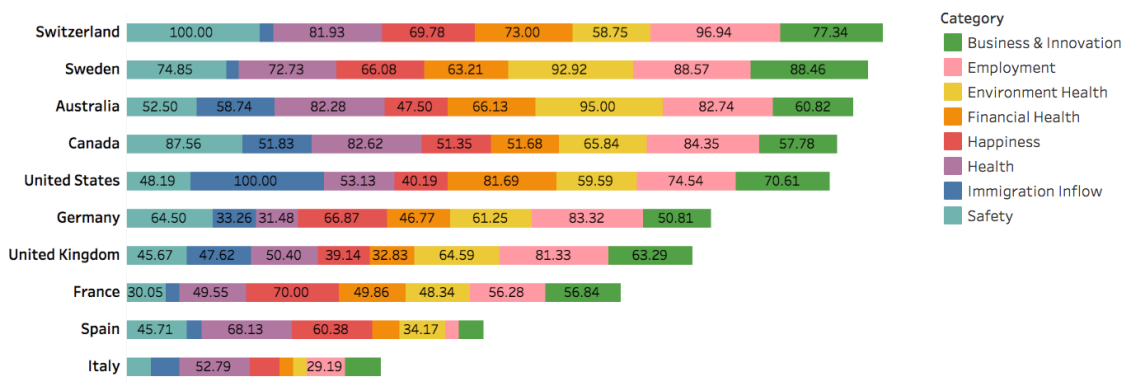
Hover here > 



Choose the country of your interest and click on the country branch on the tree map, bar graph will provide the information of country for different category. Click on country name on bar graph and a Wikipedia link will open so that more about the country can be learn. Radar graph will show the indicator's value with respect to the maximum value.

	Switzerland	Sweden	Australia	Canada	United States	Germany	United Kingdom	France	Spain	Italy
Business & Innovation	3	1	6	5	4	8	2	7	10	9
Employment	1	2	5	6	8	3	4	9	7	10
Environment Health	1	3	5	3	8	2	6	6	10	9
Financial Health	1	5	3	6	2	4	7	8	9	10
Happiness	2	4	7	6	8	3	9	1	5	10
Health	4	5	2	1	3	10	7	8	6	9
Immigration Inflow	8	10	2	3	1	5	4	9	7	6
Safety	1	3	5	2	7	4	8	9	6	10

Comparison with other countries..



Find the perfect country for you based on what matters to you the most

Change the value using slider (in the bottom and on the right) in any of the categories to see which country fits your criteria



This dashboard will show the country ranking in each category and when we click on category name, below bar graph will display the comparison of country across all the selected counties. I added slider filter for all the categories and based on your selection in any filter, list of country will get displayed which has equal to greater value than selected value.

These visualizations are functional as we can compare the countries based on different indicators which is the main purpose of the project. Insightful as it was showing some new information to the audience like how well is the country is doing in business and innovation. And enlightening as well as they can change or make their decision to choose the country based on the information available in the visualizations.

All the major important factors are added as a tooltip mark and additional info provided in Info section, so that it can be ease of use and provide explanation to the users to see all the important details of these factors. When hover over Info image on each dashboard, additional details provided on how and what is the purpose of each graph and how additional details can be seen. Both visualization provides overview first and later provide the additional details and information.

I validated all the visualizations before releasing by validating domain and data availability.

Timeline

	Task	Comments
Week 1-2	Problem Statement <ol style="list-style-type: none"> Find a topic for a Project Prepare a problem statement 	<ol style="list-style-type: none"> Did research about the topic and worked on finding Datasets from multiple source to support my claim. Met TA and discussed the topic, did little modification in the targeted audience according to TA's valuable comment. Decided on the claim and story I can create based on the topic I choose for the project
Week 3-4	Data Wrangling <ol style="list-style-type: none"> Data Preparation and Cleansing Understanding the meta Data <ol style="list-style-type: none"> Measure Dimensions Format 	<ol style="list-style-type: none"> Started looking for a indicators which will support my claim and searched for datasets required for the project. Started Data Cleaning using Python. Class concepts of python were quite useful to get me started. Removed unwanted data and added few calculated columns (Scaled value) from/to the raw files and kept /add data which were useful to create the KPIs
Week 5-6	Data Visualization <ol style="list-style-type: none"> Explore Multiple Visualization Visualization <ol style="list-style-type: none"> Style Preference Ease of Use 	<ol style="list-style-type: none"> Started working on Tableau and prepared charts to support my claim. I tried to improve my charts based on the KPI concept learnt in the class. Read some articles on internet on how to create a story and how to display the same data differently which provide more insights. Modified graphs based on the valuable inputs provided by Professor and TA
Week 7-8	Argument and Refinement <ol style="list-style-type: none"> Develop and Validate argument Report and add changes in visualization 	<ol style="list-style-type: none"> Started working on the report to document everything I have worked on so far. Added some changes suggested by professor in visualization

Discussion and conclusion

Things that worked well:

- To see country's overall performance in all indicators and compare a country with a reference (maximum value), tableau's calculation field and polygon features works well to provide visualization in radar chart.
- Ranking of the country based on the score with relative score of other contrives.
- Population of the countries, based on selected value by user in slider.
- Tableau's supercool features to add URL and custom images from the web.

Things didn't work:

- I wanted to rank the countries based slider's movement and the country's position will change based on its indicator value. Country values are getting populated correctly but could not implement ranking.
- I wanted to show the savings in each country by taking the difference of salary and expenditure but only house expenditure details was available and we can't make the decision solely based on house expenditure.

Conclusion:

- Python with Pandas library is very powerful combination to clean and transform the data.
- Choose the right KPI matters to convey the right information and support your claim.
- Providing additional information about graph and how does it work in tableau is essential as user might not be aware with the functionality you have implemented.
- Interactive visualization can be very effective in customize the data based on user selected value.

Critique

The visualizations I have created are easy to understand and convey the insights, there are couple of things I would have implemented which make it more insightful and functional.

1. **Display country based on user provided indicator value:** We can add provide sliders for each indicator, user can change the value in slider from 10 to 100 and based on selected value list of countries would populate which has close indicator value a user is looking for.
2. **Immigration rules and policies:** In the visualization, only immigration inflow details is available but not enough information available on why people are moving to United states compared to other countries and less to Switzerland though being the No 1 country in most the indicator. I could have provided the ranking based on the easy of immigration.

Reference

1. Tableau Public link: https://public.tableau.com/profile/publish/DataViz_Ekta/Story
2. Github repo for project artifacts <https://github.com/ektaratanpara/dataviz-indiv-project>
3. OECD <http://stats.oecd.org/Index.aspx?DataSetCode=BLI>
4. Global Innovation Index <https://www.globalinnovationindex.org/analysis-indicator>
5. Software Engineer Salaries <http://swexperts.com/news/software-engineer-salaries-by-country/>

6. How to create radar graph in Tableau <https://www.tableau.com/about/blog/2015/7/use-radar-charts-compare-dimensions-over-several-metrics-41592>
7. Data cleaning notebook <https://nbviewer.jupyter.org/github/ektaratanpara/dataviz-indiv-project/blob/master/DataViz-Ekta-IndivProject.ipynb>