Proposal: Migration: A story of a mobile species

DATA 450 Capstone

Nischal Bhandari

February 8, 2024

1 Introduction

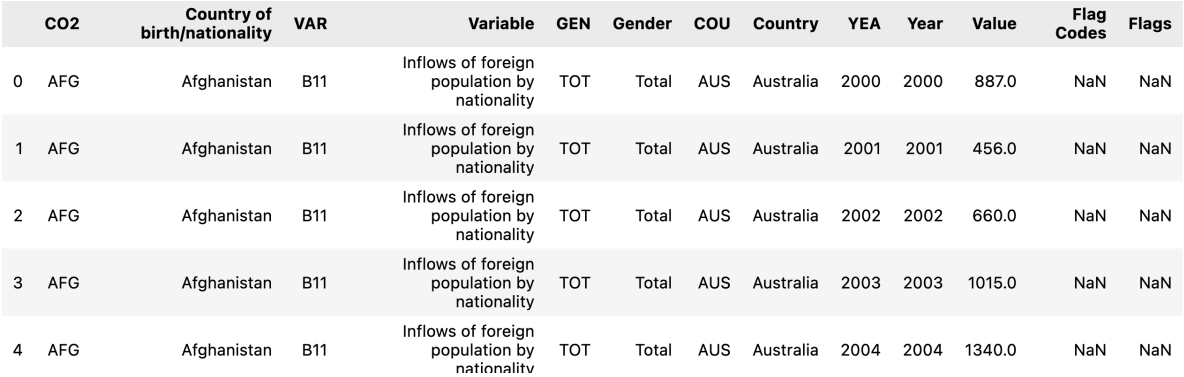
Migration seems to be ingrained in the bluprint of our existence. From the dawn of our species, humans have migrated from Africa to Asia to Eurasia to America. Early ancestors moved from places to places in search of better habitats and food. The pattern of migration not only ensured the survival of our species through suitable climate and food but also changed the behaviors and lifestyle of our species as a whole. The hunter-gatherers tribe settled into agricultural societies, and there are several revolutionary tranformations to the structures of our society that can be ascribed to migration–rise and fall of kingdoms, industrial revolution, displacement of indegenous population, changes in the population charactristics (gender, religion, age groups), to mention some. While the nature of migration and its consequences have changed along with times, it is important that the patterns of migration are scrutunized to understand the socio-economic drivers of migration. This project will study the mobility of human from different countries across two different decades, 2000-2010 and 2010-2020. As stated before, humans have always been mobile species, but the motives have evolved over time. So it will also be equally important to shed the driving force of migration, rather than bluntly mentioning the statistics. In order to achieve this aim, this project will look migration patterns through the lenses of economy, war, epidemics, and human rights.

2 Dataset

The datasets that will be used to study the drivers and pattern of migration will be obtained from "Organization of Economic Cooperation and Development (OECD)" and "The Global Knowledge Partnership on Migration and Development (KNOMAD)". The dataset from OECD will contain the migration rate–recored in the number of individuals migrating from one country to another–across 2002-2022 where as KNOMAD's datasets will contain the capital outflows from and inflows of a specific country. ** how the data was obtained?

The OECD dataset has measures of inflows and outflows of foreign population for a country. The measures are based on population registers, residance and/or work permits, and estimation from surveys. The general view of the the data in OECD dataset is shown in Figure 1.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 464640 entries, 0 to 464639
Data columns (total 13 columns):
 #   Column                         Non-Null Count   Dtype  
---  ------                         --------------   -----  
 0   CO2                            464640 non-null  object 
 1   Country of birth/nationality   464640 non-null  object 
 2   VAR                            464640 non-null  object 
 3   Variable                       464640 non-null  object 
 4   GEN                            464640 non-null  object 
 5   Gender                         464640 non-null  object 
 6   COU                            464640 non-null  object 
 7   Country                        464640 non-null  object 
 8   YEA                            464640 non-null  int64  
 9   Year                           464640 non-null  int64  
 10  Value                          464637 non-null  float64
 11  Flag Codes                     3 non-null       object 
 12  Flags                          3 non-null       object 
dtypes: float64(1), int64(2), object(10)
memory usage: 46.1+ MB
```



	CO2	Country of birth/nationality	VAR	Variable	GEN	Gender	COU	Country	YEA	Year	Value	Flag Codes	Flags
0	AFG	Afghanistan	B11	Inflows of foreign population by nationality	TOT	Total	AUS	Australia	2000	2000	887.0	NaN	NaN
1	AFG	Afghanistan	B11	Inflows of foreign population by nationality	TOT	Total	AUS	Australia	2001	2001	456.0	NaN	NaN
2	AFG	Afghanistan	B11	Inflows of foreign population by nationality	TOT	Total	AUS	Australia	2002	2002	660.0	NaN	NaN
3	AFG	Afghanistan	B11	Inflows of foreign population by nationality	TOT	Total	AUS	Australia	2003	2003	1015.0	NaN	NaN
4	AFG	Afghanistan	B11	Inflows of foreign population by nationality	TOT	Total	AUS	Australia	2004	2004	1340.0	NaN	NaN

Figure 1: Migration Dataview

From this dataset, the study will use all the variables except `Flage Codes` and `Flags`. Furthermore, some of the redundant variable will be deleted from the dataset. The `Variable` coulumn has information on inflows, outflows, and citizenship acquisition which will be used in this study. OECD has separate dataset for population characteristics like gender, employment, education levels, and so on for migrating population. The research will delve into these characteristics for some years and countries.

The other data that will aim to study one of the potential drive of migration is remittance database. The remittance are in two categories: i. capital money outflowing from a country into a foreign country and ii. capital gains of a country from abroad. The variables in the data are shown below:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 226 entries, 0 to 225
Data columns (total 26 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   Remittance inflows (US$ million)  224 non-null    object
 1   2000                             199 non-null    float64
 2   2001                             199 non-null    float64
 3   2002                             199 non-null    float64
 4   2003                             199 non-null    float64
 5   2004                             199 non-null    float64
 6   2005                             199 non-null    float64
 7   2006                             199 non-null    float64
 8   2007                             198 non-null    float64
 9   2008                             198 non-null    float64
 10  2009                             198 non-null    float64
 11  2010                             198 non-null    float64
 12  2011                             197 non-null    float64
 13  2012                             197 non-null    float64
 14  2013                             197 non-null    float64
 15  2014                             197 non-null    float64
 16  2015                             197 non-null    float64
 17  2016                             197 non-null    float64
 18  2017                             196 non-null    float64
 19  2018                             196 non-null    float64
 20  2019                             196 non-null    float64
 21  2020                             196 non-null    float64
 22  2021                             196 non-null    float64
 23  2022                             196 non-null    float64
 24  2023e                            196 non-null    float64
 25  % of GDP in 2023                 190 non-null    float64
```

```
dtypes: float64(25), object(1)
memory usage: 46.0+ KB
```

This is the dataset for the flow of capital in a specific country–column 1–and there will be another similar dataset that will contain flow of capitals out of a specific country. They will have the same data structure and composition but will serve different purpose to understand the pattern of migration: are countries receiving the highest numbers of migration sending out the highest remittance to the country of permanent residence? If not, does the former residence have any other drivers: war, political unstability, and so on?

The World Bank launched the project "KNOMAD" to track the data on remittance flows. KNOMAD coordiantes with organizations like International Monetary Fund, United Nations, and EuroStat to obtain the up-to-date information on remittance flows.

3 Data Acquisition and Processing

The datasets will be downloaded from the OECD and KNOMAD's websties. The migration dataset has variables like 'Flag' and 'Flag Codes' and other redundant variables like 'Var', 'Gen', and 'Year' which will be eliminated from the dataset. The naming of the first column will be changed for the remittance dataset for clarity. In addition, for the sake of consistency, all the variables names will be lower-cased across datasets. The datasets will be split into two decades: 2000-2010 and 2010-2020. The mean migration rate and mean remittance will be calculated in a new column for each decade. Only a little time will be invested in imputing missing values as this research plans to study only the top ten or twenty country with highest migration and capital flows. Another valid reason of not imputing null values with mean or median of the row or column is that null value can either be actual measurement or also a missing measurement.

However, for the machine learning modelling, the missing values will be imputed either by mean or median based on the distribution within variables.

4 Research Questions and Methodology

The following topics will be covered by this study:

[In this section, list each of the questions you will explore. Following each question, provide a detailed and specific plan for how you plan to answer the question. Include the specific steps you will take, what form the answer will take (a number? table? visualization? model? Give all the specifics), and estimate how many hours each question will take to complete.]

1. **What is the general trend of migration across countries and time? Is there any change in pattern after the inception and during COVID pandemic?**

To explore the general trend of migration over time, descriptive statistics will be applied to discern global migration patterns, focusing on changes post-inception and during the COVID-19 pandemic. Visualizations, specifically line graphs and time-series charts, will be used to depict the evolution of migration dynamics across countries and decades. It will take around three hours.

2. **What are the top 10 countries with most immigration inflows across two decades? For the year with most migration, what is the composition of genders and education levels of migrants?**

First of all, top 10 countries will be calculated based on the mean migration flows within each decade. The result will be presented in tables and bar charts. Besides that, a year with peak migration (based on total migrants) will be selected within each decade and will be studied for the population characteristics of the immigrants. Pie charts or stacked bar charts will be used for illustration. It will take approximately two to three hours.

3. **Are the country receiving the most immigrants also the ones that send out the most remittances and vice versa? What percentage of GDP is contributed by remittance to a country receiving incomes from citizens working abroad? What about the countries that receives or sends the lowest remittance: are they independent, isolated, politically different..?**

This research will integrate the mean migration and mean remittance data calculated on top of the OECD and KNOMAD datasets to explore potential correlations between immigration inflows, remittance patterns, and their impact on the GDP of receiving countries. This will perform the downstream analysis based on the findings of the previous research question. The results will be communicated through a correlation matrix, bar charts for top countries, and a stacked bar chart illustrating remittance contributions to GDP. This question will also focus on the country receiving the least remittance and will see their pattern in migration, politics, and global engagement.It will take around three hours.

4. **What are the top 10 countries based on the outflow of the foreign population from a country? Is outflow related to the inflows–a lot of people emigrate from a country X and a lot of people return to country X? Or are there certain countries where a large population emigrate from a country but only minimal proportion returns? Are there any other explanation: war, human rights issues, etc?**

This research seeks to understand relationships between outflows and inflows, especially in identifying countries with significant emigration but minimal returns. The question will extend on the research question #2. Visualizations in the form of a world map and side-by-side bar charts will be employed to illustrate outflow patterns, shedding light on the dynamics of population movements. Plotly will be used to animate the inflows and outflows across top 10 countries and compare them side by side. This question also demands extensive research beyond dataset to understand why migrants from certain countries look for permanent escape from their residence. It will take around four hours.

5. **And what are the top 10 countries that have the highest acquistion of forein population and how does the acquistions change over time? Is it different among Schengen countries who adopted open broder policies among its countries–some of which are prosperous and some are not (and significantly)?**

Line charts will be used to show the acquisition rate for top 10 countries (based on the mean acquisition rate). This will be studied in contrast with the Schengen-countries as they have adopted open-border policies among its members. Mostly, line charts will be used. Bar graphs may also be used to illustrate the acquistion for countries. It will take around four hours.

6. **Is it possible to cluster countries based on their migration and remittance flow?**

Clustering algorithms, specifically k-means, will be employed on migration and remittance data to categorize countries based on their migration and remittance flows. Features such as immigration inflow, outflow, and remittance sent/received will be inputted for training the model. The results will be communicated through a cluster dendrogram illustrating country groupings and a chart mentioning the characteristics of each cluster. Overall, this modelling aims to see if there is a nuanced pattern in global migration dynamics. It will take around six hours.

5 Work plan

Week 4 (2/12 - 2/18):

- Data tidying (1 hours)
- Question 1 (3 hours)
- Question 2 (3 hours)

Week 5 (2/19 - 2/25):

- Question 3 (2 hours)
- Question 4 (1 hours)
- Question 5 (2 hours)
- Question 6 (2 hours)

Week 6 (2/26 - 3/3): * Question 3 (1 hours) * Question 4 (2 hours) * Question 5 (2 hours) * Question 6 (2 hours) **Week 7 (3/4 - 3/10):**

- Question 6 (3 hours)
- Presentation prep and practice (4 hours)

Week 8 (3/11 - 3/17): *Presentations given on Wed-Thu 3/13-3/14. Poster Draft due Friday 3/15 (optional extension till 3/17).* * Question 6 (1.5 hours) * Poster prep (4 hours) * Presentation peer review (1.5 hours)

Week 9 (3/25 - 3/31): *Final Poster due Sunday 3/31.*

- Peer feedback (3.5 hours)
- Poster revisions (3.5 hours)

Week 10 (4/1 - 4/7):

- Code revisions (5 hours)
- Literature reviews (2 hours)

Week 11 (4/8 - 4/14):

- try other machine learning models, time-series forecasting (7 hours)

Week 12 (4/15 - 4/21):

- Sync results, name files, and organize directories in Github (3 hours)
- Read blogs from Our World In Data (2 hours)
- Re-read datascience blogs (2 hours)

Week 13 (4/22 - 4/28): *Blog post draft 1 due Sunday night 4/28.* [All project work should be done by the end of this week. The remaining time will be used for writing up and presenting your results.]

- Draft blog post (4 hours).

Week 14 (4/29 - 5/5):

- Peer feedback (3 hours)
- Blog post revisions (4 hours)
- [Do not schedule any other tasks for this week.]

Week 15 (5/6 - 5/12): *Final blog post due Weds 5/8. Blog post read-throughs during final exam slot, Thursday May 9th, 8:00-11:20am.*

- Blog post revisions (2 hours)
- Peer feedback (2 hours)
- [Do not schedule any other tasks for this week.]

6 References

1. OECD. (n.d.). Migration Statistics. International Migration Database. https://stats.oecd.org/Index.aspx?
2. World Bank. (n.d.). Remittances. KNOMAD. https://www.knomad.org/data/remittances
3. Ratha, D. (2019, February 5). Remittances: Funds for the Folks Back Home. IMF.
https://www.imf.org/en/Publications/fandd/issues/Series/Back-to-Basics/Remittances#:~:text=When%