Determinants Of Migration

Eda Kal Arıca-2595692
Middle East Technical University
Ankara, Turkey
eda.arica@metu.edu.tr

Esra Şekerci-2141992
Middle East Technical University
Ankara, Turkey
esra.sekerci@metu.edu.tr

Abstract— This comparative study utilizes panel data analysis to investigate net international migration patterns across 140 countries. All procedures are conducted within the R environment. Using an eight-year dataset, we examine the determinants and implications of net migration flows. The valuable insights gained from our study would be helpful in advancing our understanding of the intricate interplay of factors that shape net migration decisions.

Keywords— Migration, statistical modeling, panel data analysis

I. INTRODUCTION

Migration is a complex global phenomenon that carries profound social, economic, and political implications. People migrate to different countries for diverse reasons, including seeking better economic opportunities, escaping conflict or persecution, reuniting with family, or pursuing education and cultural exchange. Consequently, comprehending the factors and consequences of net international migration becomes imperative for policymakers and researchers.

The dataset encompasses variables that capture crucial aspects related to migration, such as economic conditions (unemployment rates and consumer price indices), political stability, and subjective well-being (happiness scores). Analyzing the interrelationships between these variables and net migration enables us to discern the significant factors shaping migration patterns.

In summary, this study offers a comprehensive analysis of net international migration, shedding light on its determinants and implications. Through the utilization of panel data analysis and an extensive dataset, we endeavor to contribute significant insights for researchers interested in understanding and effectively managing migration dynamics.

II. LITERATURE REVIEW

Numerous studies have examined the determinants of migration inflows. For example, Mayda, A. M. (2010) investigates migration inflows into fourteen OECD countries between 1980 and 1995. The study analyzes income factors, geographical and cultural influences, and policy changes. The findings align with theoretical predictions, providing insights into international migration patterns.

Another study focuses on the Eastern enlargements of the European Union, exploring East-West migration flows from 2000 to 2017. The research identifies the impact of GDP per capita and youth unemployment rate on emigration rates. This contributes to understanding labor supply adjustments through international migration (Franc, S., Časni, A. & Barišić, A., 2019).

These studies are part of a larger literature that employs panel data analysis. Economic factors, policy changes, and non-economic factors shape migration flows. Panel data analysis captures within-country and cross-country variations, revealing the dynamics of migration patterns. However, further research is needed.

III. METHODOLOGY

A. *Dataset*

The dataset used in this study was created by merging multiple series sourced from the World Bank, a renowned institution known for its extensive collection and publication of global development data. Additionally, data from the World Happiness Report were included. After conducting a thorough analysis, we identified key factors that potentially influence migration. With the goal of achieving comprehensive coverage, our focus was to include data from as many countries as possible during the dataset construction phase.

To retrieve data from the World Bank, we utilized the R programming language. By specifying the desired indicators, countries, and time range, we leveraged R's packages and functions to access and retrieve the relevant data from the World Bank's database. Moreover, to facilitate exploratory analysis and gain further insights, we applied coding to create a "continent" variable within our dataset.

Now, let us delve into the meaning and parameters of each index. Index parameters refer to the variables, factors, or indicators used to calculate or determine the index value.

The political stability index evaluates factors such as violent crime, government stability, risks of unconstitutional changes or violent overthrows, political violence, and terrorism. It measures stability and the absence of violence in a country's political environment, ranging from -2.5 (indicating weak political stability) to 2.5 (indicating strong political stability). This index provides valuable insights into public trust in governance and the associated risks of political instability.

The Women Business and the Law Index Score assesses how laws and regulations impact women's economic opportunities. It examines the extent to which legal frameworks support or hinder women's participation in economic activities. This index comprises multiple parameters that collectively evaluate the influence of laws and regulations on women's economic empowerment. The overall score is calculated by averaging scores across eight key areas, including mobility, employment, compensation, marital rights, reproductive rights, business ownership, property rights, and pension access. A score of 100 represents the highest rating, indicating robust legal provisions that promote women's economic opportunities and ensure comprehensive protections.

The Consumer Price Index (CPI) considers various elements when calculating price changes. These factors encompass a range of goods and services representing typical consumer purchases, including their prices and the weights assigned to each based on their relative importance in consumer spending. The CPI specifically tracks fluctuations in costs related to categories such as food, housing, transportation, healthcare, and education. By monitoring price changes within these specific parameters, the CPI provides insights into overall inflationary trends and reflects the impact on the average consumer's expenses over time.

The happiness score index measures subjective well-being and happiness levels using parameters including GDP per capita, healthy life expectancy, social support, generosity, perceptions of corruption, and freedom to make life choices. This index helps policymakers and researchers understand and enhance happiness, contributing to overall well-being and quality of life.

Feature name	Explanation
country	chr: country names (having 140 categories)
year	int: from 2009 to 2016
migration	num: net migration
pol.stability	num: political stability and absence of violence/ terrorism (ranging from approximately -2.5 to 2.5)
unemp	num: unemployment, total (percent of total labor force)
women.bl.scr	num: business and law index score measures how laws and regulations affect women's economic opportunity (scale 1-100)
price.idx	num: consumer price index reflects changes in the cost to the average consumer of acquiring a basket of goods and services that may be fixed or changed at specified intervals, such as yearly.
happiness.scr	num: happiness scores (1-10)
continent	chr: continent names

Our final dataset, derived from the combination of various sources, consists of 9 variables and 1120 observations, representing 140 countries. These observations span the years 2009 to 2016. However, during the dataset compilation process, we encountered 407 missing values, which amount to approximately 4.04 percent of the data. To address this issue, we employed a method to generate corresponding values for the missing years, ensuring consistency and comparability across all countries. In the subsequent part of the analysis, we will delve into the details and implications of these missing observations, allowing for an extensive exploration of the dataset.

B. *Descriptive Statistics*

At the beginning of our analysis, we created a descriptive statistics table to gain an initial understanding of the dataset. This table enables us to find out key statistical measures, such as mean, median, standard deviation, and quartiles, which offer insights into the central tendencies and variability of the data. These descriptive statistics serve as a foundation for formulating research questions and guiding further analysis. Table 1 presents the descriptive statistical summary specifically for the continuous variables in the dataset, allowing us to examine their distribution and characteristics.

	migration	**pol.stability**	**unemp**	**women.bl.scr**	**price.idx**	**happiness.scr**
Min.	-2290411	-2.8608	0.250	26.25	78.01	2.662
1st Qu.	-29821	-0.8615	4.055	63.75	100.00	4.615
Median	-2551	-0.1415	6.490	77.19	107.89	5.373
Mean	2698	-0.2088	7.878	73.84	119.12	5.453
3rd Qu.	19608	0.5889	10.088	88.75	118.23	6.284
Max.	1449371	1.5255	33.130	100.00	2740.27	7.858
NA's	64	64	64	72	79	64

Table 1 Descriptive Statistical Summary of Numerical Data

The variable "migration" represents net migration and serves as the response variable in our analysis. It exhibits a wide range of values, ranging from -2,290,411 to 1,449,371. The skewness of -

2.62 indicates a left-skewed distribution, implying that there are a few countries with significantly high positive migration values.

Moving on to the other variables, the political stability index ranges from -2.8608 to 1.5255, with a skewness of -0.40, indicating a slightly left-skewed distribution. This suggests that most countries have relatively higher levels of political stability, while a few countries exhibit lower stability. The unemployment rates have a skewness of 1.51, indicating a heavily right-skewed distribution. The range of unemployment rates extends from 0.25 to 33.13, with a concentration of countries experiencing relatively higher rates. The Women Business and Law Index Score has a skewness of -0.82, indicating a left-skewed distribution. This suggests that most countries have relatively higher scores, indicating better support for women's economic opportunities, while a few countries have lower scores. The Consumer Price Index exhibits a highly right-skewed distribution with a skewness of 24.83, indicating a small number of countries experiencing significant price increases in specific categories. Finally, the happiness score index ranges from 2.662 to 7.858 and has a skewness of -0.40, indicating a slightly left-skewed distribution. Most countries have relatively higher happiness scores, with a few countries having lower scores. These observations provide a deeper understanding of the distributional characteristics of each variable, which will guide further analysis and interpretation.

Following this description, a histogram table is presented to visually observe the distributional form of each variable.

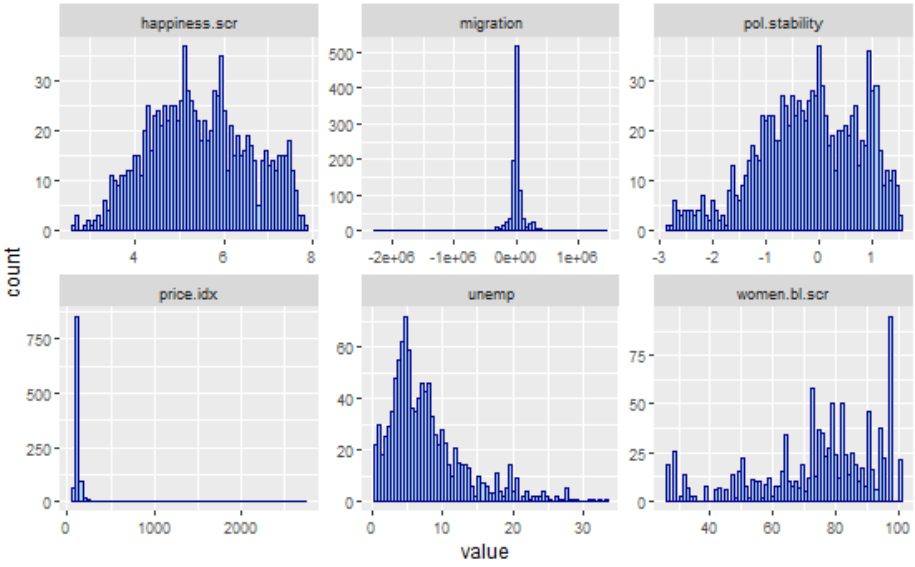

Figure 1 Histograms for Numerical Data

In the modeling part, we explored the effects of applying both transformation and scaling techniques to the variables in our dataset.

C. *Exploratory and Confirmatory Analysis*

In this section, we perform exploratory and confirmatory data analysis on our dataset, aiming to uncover patterns, relationships, and key insights. We address a series of five research questions throughout the analysis to guide our investigation and provide comprehensive insights into the data.

RQ1: How does net migration related with other regressors?

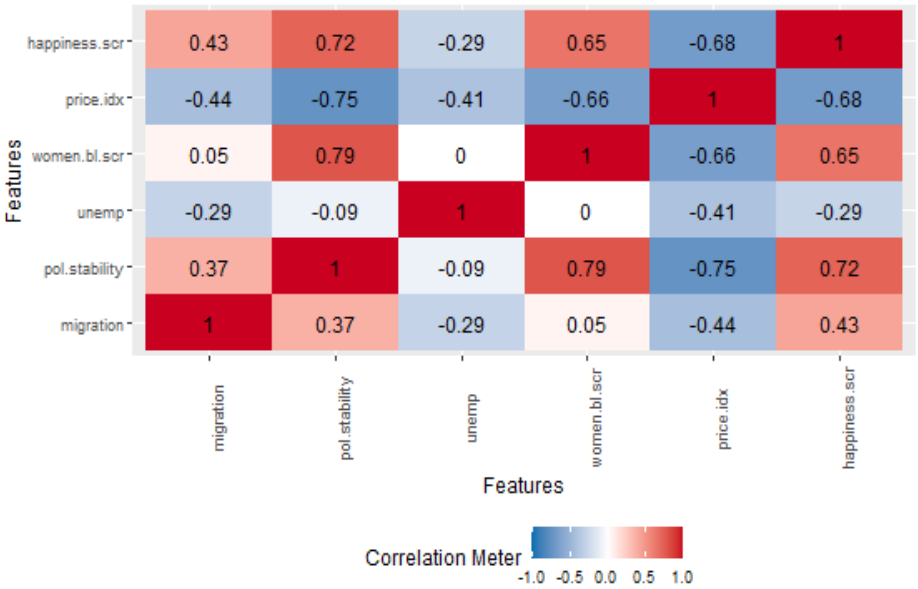

Figure 2 Correlation Matrix

The correlation matrix shows the Kendall-rank correlation coefficients between the variables. We observe that migration, our response variable, shows positive correlations with pol.stability (0.37) and happiness.scr (0.43), indicating that higher levels of political stability and happiness score are associated with higher net migration. On the other hand, migration has negative correlations with unemp (-0.29) and price.idx (-0.44), suggesting that higher unemployment rates and consumer price index are associated with lower net migration.

Pol.stability shows positive correlations with women.bl.scr (0.79) and happiness.scr (0.72), indicating that higher levels of political stability are associated with better scores for women's business and law index and higher happiness scores.

Price.idx, representing the consumer price index, has negative correlations with pol.stability (-0.75) and unemp (-0.41). This means that as the consumer price index increases, indicating higher prices, there is a tendency for political stability to decrease and unemployment rates to increase. The negative correlations indicate an inverse relationship between these variables, supporting the statement that higher consumer price index values are associated with lower levels of political stability and higher unemployment rates.

Happiness.scr, representing the happiness score index, shows negative correlations with price.idx (-0.68), indicating that lower consumer price index is associated with higher happiness scores.

Overall, these correlation coefficients highlight the interdependencies among the variables in our dataset, providing valuable insights into their relationships and potential impacts on each other. Additionally, when interpreting the results and drawing conclusions from the analysis, it is crucial to carefully consider the correlations among the variables and their potential impact on standard errors. The presence of strong correlations suggests the possibility of multicollinearity, which can affect the accuracy and stability of coefficient estimates. It is important to consider these relationships and explore strategies to address multicollinearity, ensuring robust and reliable findings in the analysis.

RQ2: How does net migration vary over time for different countries?

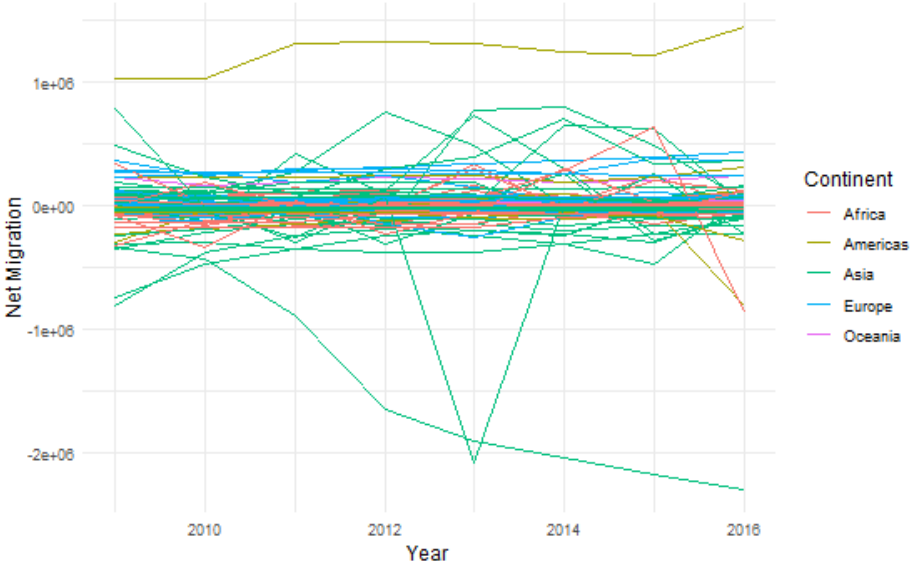

Figure 3 Net Migration Over Time: Spaghetti Plot

Figure 3 provides an overview of net migration levels across 140 countries over the years. The plot reveals that the net migration levels for most countries do not vary significantly, as indicated by the relatively straight lines observed within a certain range. Upon closer examination of the fluctuating lines in the plot, it becomes evident that the United States stands out as a country with consistently high net migration levels, positioned above most of the observations. On the other hand, several countries, primarily in Asia, including Pakistan, Bangladesh, and Syria, exhibit fluctuating net migration levels that are consistently lower than the overall trend.

This finding highlights the need for further investigation into potential factors that could shape migration patterns, such as specific years or countries. In the subsequent modeling phase, we will explore these factors in more detail to assess their impact on net migration.

RQ3: What is the difference in net migration between Turkey and Syria?

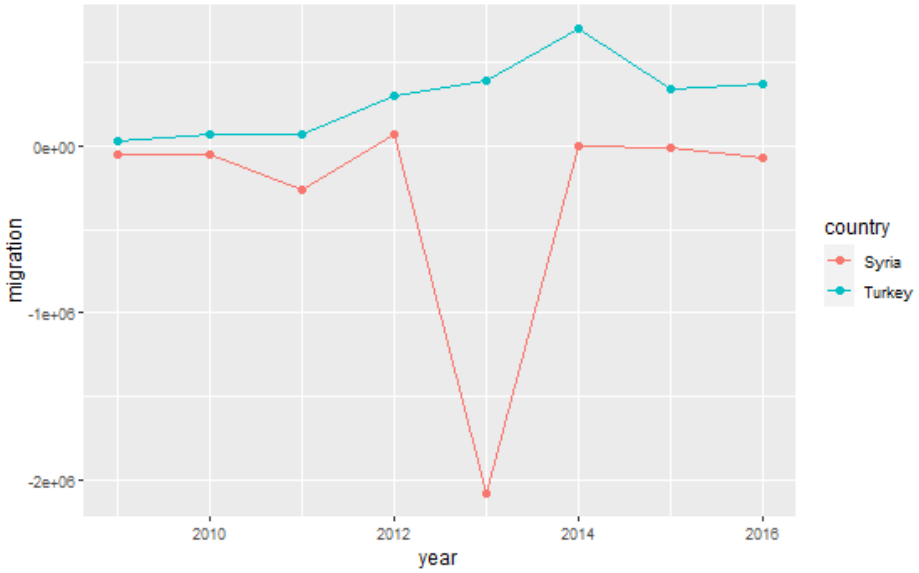

Figure 4 Net Migration Comparison for Turkey and Syria

Although not the focus of our study, we recognize the significant influence of the Syrian Civil War on migration patterns during the period from 2009 to 2016. The conflict, which began in 2011, had a notable impact on population movements in Syria and Turkey. In Syria, the effects of the war became evident in the subsequent years, leading to a substantial increase in emigration as individuals sought refuge from violence and instability. Many Syrian refugees found shelter in neighboring countries and across Europe. Conversely, Turkey experienced a significant rise in immigration, particularly from Syria, as it became a primary destination for Syrian refugees due to its geographic proximity and relatively stable environment. Turkey implemented an open-door policy to facilitate the entry and settlement of these refugees. These migration patterns highlight the profound consequences of the Syrian Civil War, with Syria witnessing substantial emigration and Turkey experiencing notable immigration.

To conduct a comprehensive case study, it is important to acknowledge that migration is a complex phenomenon influenced by various factors, and its outcomes can take time to fully materialize in response to changing conditions. According to Bozcaga, Christia, Harwood, Daskalakis, and Papademetriou (2019), their comprehensive analysis of Syrian refugee integration in Turkey, based on data from the Data for Refugees Challenge, revealed significant impacts of economic activity, availability of health facilities and charity foundations, network centrality, and district location on social integration. This suggests that migration is indeed influenced by these factors, also highlighting their importance in shaping the process of integration for Syrian refugees in Turkey.

RQ4: How does individual happiness and women's economic presence is affected by CPI for different countries for a randomly chosen sample?

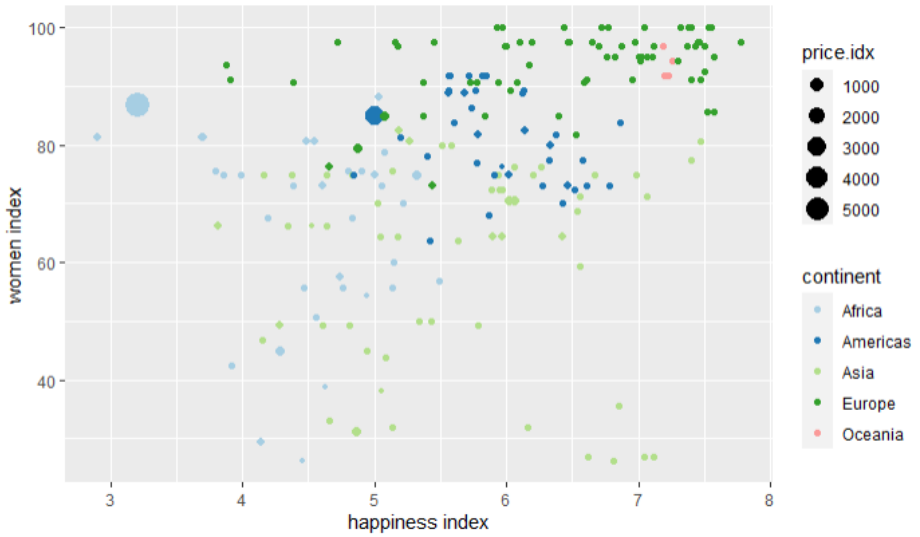

Figure 5 Comparison of Women and Happiness Indexes for Continents based on Customer Price Index

In Figure 5, we present a visualization of three numeric features: women's business and law index, happiness index, and customer price index. Each data point is color-coded based on its continent. To ensure clarity and avoid clutter, a random sample of observations is selected for the plot. By examining the plot, we can simultaneously observe multiple aspects. It becomes evident that countries with higher women's business and law index scores tend to have higher happiness index rankings and lower customer price index values. Furthermore, these countries are primarily

concentrated in the continents of Europe, America, and Oceania. The plot reinforces our earlier observation regarding the correlation between these variables, providing visual evidence of the relationship.

RQ5: How does net migration is affected by political stability considering countries' unemployment level?

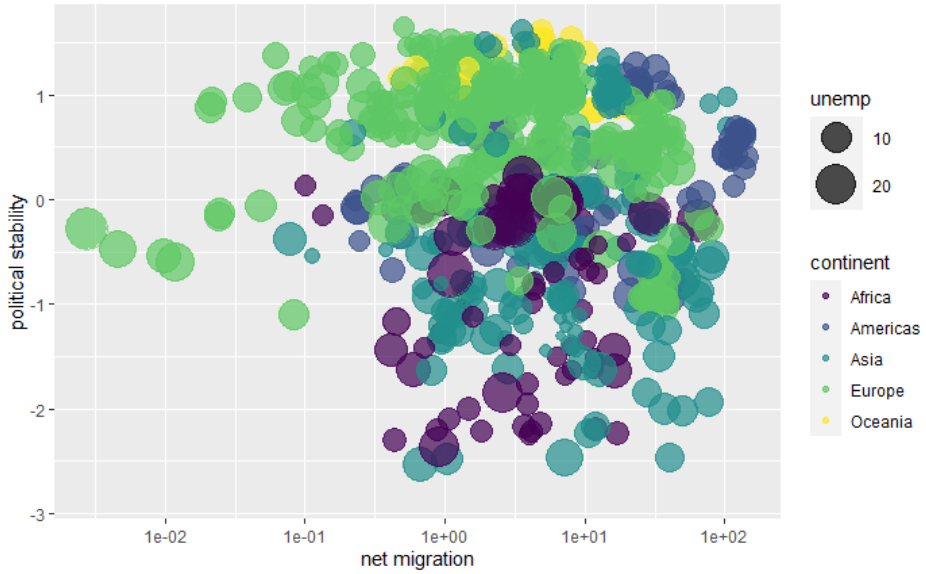

Figure 6 Political Stability vs. Net Migration with Unemployment Level (by Continent)

In the plot displayed, the y-axis represents political stability, the x-axis represents net migration, and the size of the dots indicates the level of unemployment. The colors of the dots correspond to different continents. From the plot, we can observe several trends. Countries with higher political stability and lower unemployment levels tend to have positive net migration, primarily seen in Oceania and Europe. However, if we focus on Africa and Asia, we notice that the representative dots for unemployment are larger, indicating higher unemployment levels, and political stability is relatively lower. As a result, the net migration levels for these regions tend to be negative. The plot highlights the relationship between political stability, unemployment, and net migration, providing insights into how these factors vary across different continents.

D. Missingness

The dataset used in our study contains missing values, indicating the absence of certain data points. Addressing missing data is crucial to ensure the validity and reliability of our analysis. To handle missing data, we employed the mice (Multivariate Imputation by Chained Equations) package in R. We applied multiple imputation methods, such as mean imputation, regression imputation, and predictive mean matching, among others. Through iterative imputation steps, the mice package generated multiple imputed datasets, capturing the uncertainty associated with missing values (Young & Johnson, 2015).

To determine the most appropriate imputation method, we compared the performance of various techniques. After careful evaluation, we selected the CART (Classification and Regression Trees) method as it demonstrated the best performance in terms of imputation accuracy and variability.

To assess the missingness pattern, we examined the percentage of missing values for each variable. Out of the entire dataset, only 4.04 percent of the values are missing. The variable with the highest percentage of missing values is the customer price index, accounting for 7.05 percent of missing values.

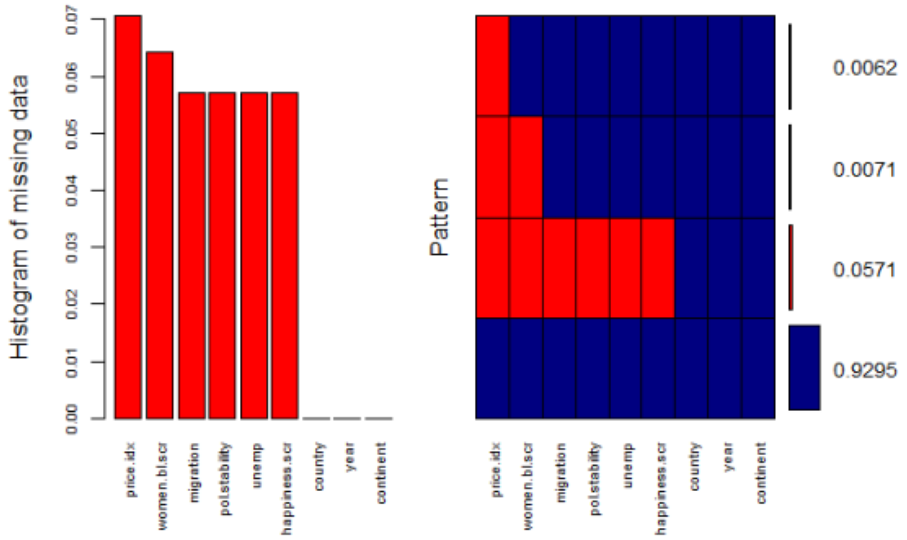

Figure 7 Aggregation Plot of Missing Values

To ensure that our imputations are suitable for further analysis, we conducted a check to compare the distribution of the imputed data with the original data. It is important that the imputed data closely resemble the original data in terms of distribution and visual presentation.

Through our imputation process, we aimed to retain the same distribution and characteristics as the original data. This ensures that any subsequent analysis conducted on the imputed dataset will not produce significantly different results or interpretations compared to the original dataset. By carefully selecting appropriate imputation methods and techniques, we strived to minimize any potential bias introduced by the missing data and maintain the integrity of our analysis.

E. Modelling

We searched for different model types to employ for our data. Random effects models, marginal models, linear regression, and machine learning models were chosen. The following parts separately will inform us about the steps we followed.

"The rationale behind random effects model is that, unlike the fixed effects model, the variation across entities is assumed to be random and uncorrelated with the predictor or independent variables included in the model." [5] In random effects models the regression coefficients measure more direct influence of explanatory variables on the responses for heterogeneous individuals. [6] We chose this model structure to understand how does net migration's components vary when the countries differ. Based on our research questions country differences were vital for us to understand net migration foundations.

Secondly, we checked model assumptions for our data. The response (net migration) for each subject were independent from each other, identically distributed. We checked the structure of response distribution for different time points. (The results are shown in the appendix section1.) The

suggested response distribution for random effects models is normal.[6] We understood that our response is not normally distributed (Shapiro-Wilk Test, *p-value<0.05*). We employed several transformations. For the positive distributed years, we employed log transformation, but in 2011 we failed. For negative distributed ones we employed square, cube and forth power transformations. After that we realized that there should be a general formula for all time points to derive inferences from our results accurately. We used bestNormalise package. The order norm transformation was selected from that algorithm for our response. When we look deep in to details of transformation technique,

$$g(x)= \Phi^{-1} ((\text{rank}(x)-0.5)/\text{length}(x))$$

This kind of transformation will be challenging for us to interpret results from. We then think that since our mean and median values are close to each other (appendix table1) we think that net migration distributes almost normally.

Also, in the literature review we depicted that random effects models can be employed on non-normally distributed responses.

1. Random Effects Models

We started with a full model where we include all our regressors in the model. We start with the raw variables, afterwards for ease of calculation and comments we divided migration to thousand and scaled all regressors for a more reliable model. We constructed 18 models at this part. We also tried the badly transformed response to see if the standard errors would drop, we saved those models (namely data set3) again and compare it with our data set namely data set 2. When examining outputs, we understood that all our models resulted with high standard errors for estimated coefficients and called for the models those models as not reliable.

The variation between countries was high and variation for years was too low, so we kept focusing on random intercept models. For significance we based our interpretation on the assumption that, for large samples t-distribution approximates to z that's why we could been able to use summary of lmer function and compare it with Z-critical value for several significance levels (the significance values were; 0.01, 0.05, 0.1, 0.2). We can observe the details in following table,

Model	**Regressors**	**Random Part**	**Reliable**	**Significance**
Model 3	all	intercept	No	Pol.stab&happiness(0.05)
Model 4	all	With year	No	Pol stab &happiness (0.05)
Model 5	All+GDP	intercept	No	Pol. stab, GDP, happiness (0.05)
Model 6	price*unemp	intercept	No	Pol.stab,happiness,interaction(0.05) Price index (0.2)
Model 7	happ*price	intercept	No	Pol.stab and happiness (0.05)
Model 8	happ*price	With year	No	Pol. stab & happiness (0.05)
Model 9	Woman*unemp	intercept	No	Pol. stab &happiness (0.05)
Model 10	Woman*price	intercept	No	Pol. stab&happiness (0.5)
Model 11	Added year	intercept	No	Pol. stab &happiness (0.05)

Model 12	pol.stab + happiness	intercept	No	Both significant
Model 13	Price*unem Women*year	intercept	No	All significant
Model 13.1	Price*unemp + women	intercept	No	Only women insignificant

Table 2 Models Descriptions

We compared four models which explains best our response by eliminating step by step using ANOVA since we couldn't obtain any function like in linear regression "step". And do a model adequacy check by residuals versus fitted values plot. For the models we thought having best performance which are model 6 and model 13 with smallest model deviances which are depicted in table 2, their residual plots nearly the same with some outliers. We thought if we would expand the x-axis limit and omit the visualization of outliers maybe it could be possible to observe an approximate random distributed residual.

Model	Deviance
Model 3	14,728
Model 6	14,725
Model 12	14,731
Model 13	14,720

Table 3 Deviance Comparison for Selected Models

We conducted a last ANOVA to select between model 6 and 13. Since the deviance is small, and p-value was smaller than 0.05 we concluded model 13 as the best model for random effects.

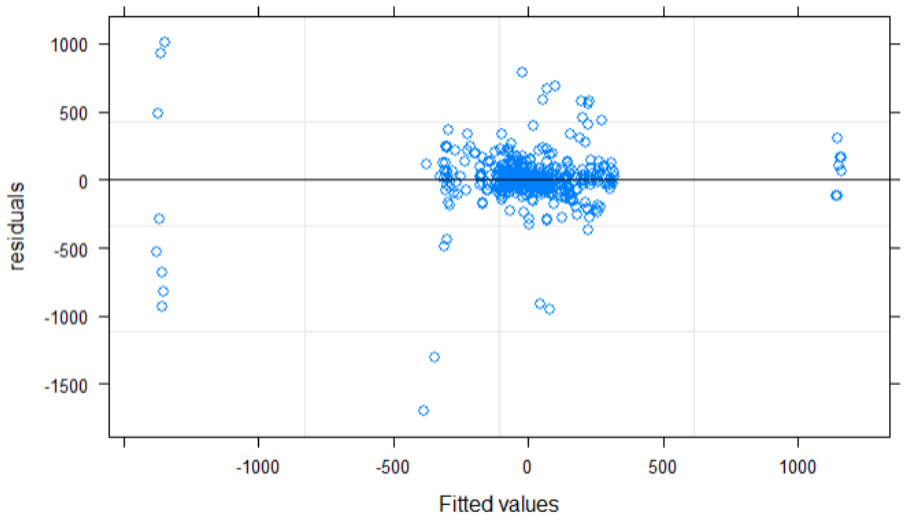

Figure 8 Model 6 Residuals versus Fitted Values

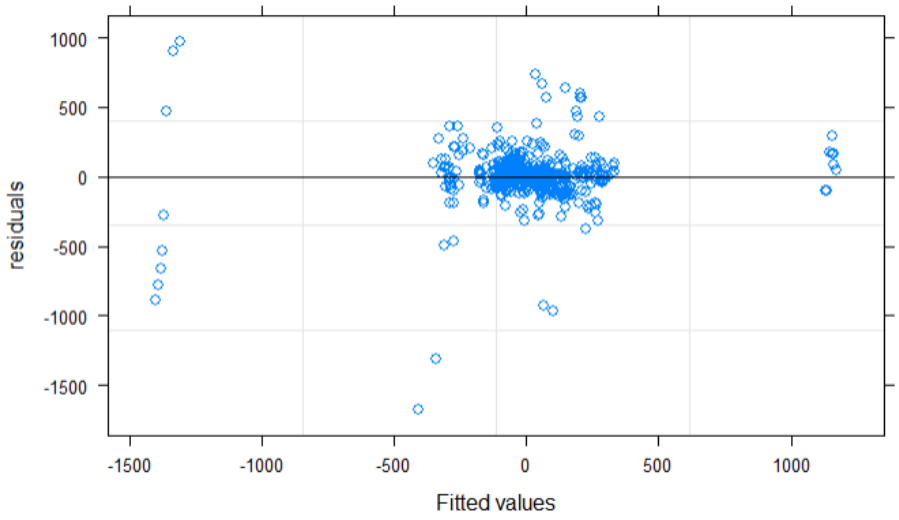

Figure 9 Model 13 Residuals versus Fitted Values

As a last note for this part we tried to employ two R functions suggested for random effects like glmer() and nlme(). The glmer() gave same results as lmer() so we didn't double checked our results and nlme() was a it complicated to use in a short time so we took it as a note for a further work.

The fitted model for random effects models is:

$$\hat{Y}_{it} = \left(\frac{mig}{1th}\right) it$$
$$= 50830 - 57.4\, pol_{it} - 0.69 price_{it} - 13.3 unemp - 682.1 women - 25.2\, years_{it}$$
$$- 0.114(price * unemp)_{it} + 0.39(women * years)_{it}$$

2. Marginal Model

Marginal Models were selected since it enables us to compare groups. The family of response is selected as gaussian, and we select the covariance structure as unstructured. (Further details, see appendix section2). We observed different values for different years in covariance matrix and a decreasing correlation pattern in correlation matrix. Our models didn't converge in this part. Due to time constraint, we are keeping this note for further work.

3. Linear Regression

This part we focused on the last 4 models selected in random effects part. We used multiple linear regression models for those 4 models with same regressor structure meaning the formula of regressors. Our process outputs are summarized in table 4.

Model	Significant regressors	Model Significance	R^2 adjusted	RSE
Lm3	Pol.stab,unemp,women, Happiness, Regions (Asia,Europe)	Yes	0.09	221
Lm6	All	Yes	0.08	221.4
Lm 12	All	Yes	0.07	222.7
Lm 13	All	yes	0.081	222.4

Table 4 Linear Regression Comparison

Only model 13 have high variance inflation values for regressors (vif scores>5). That's why we omitted that model. The model with lowest RSE and highest R2 adjusted value model lm3 was selected in this section. Model adequacy check for model lm3 is as below:

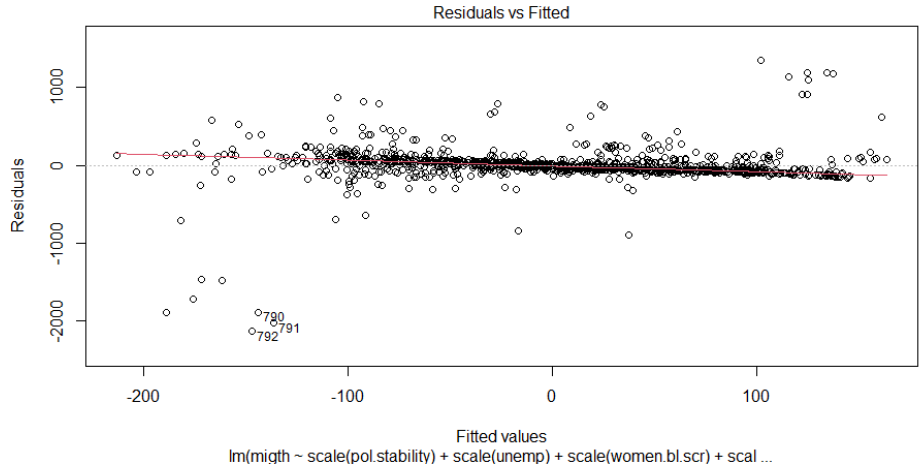

Figure 10 Model lm3 Residuals versus Fitted Values Plot

Like in random effect models' outliers are detected. We noted overcoming this problem as further work.

4. Machine Learning Model

This part was a bit time consuming to find the appropriate machine learning algorithm and employ it to the model. We chose RMtree() function. We splitted our data as train and test. We constructed two models in this section due to time constraint. The first one is for random intercept and the other was for random effect where time is considered. The train MSE was lower for the first model, but test MSE was lower for first model. (for details please see codes section). For different regressor formulas machine learning models are noted for further work.

Also we tried REEMforest() function but the algorithm didn't converge. This should be examined in detail for further work.

IV. CONCLUSION

In this study we wanted a modeling challenge for us since the concept was new and we needed to do research in detail. We constructed our data by aggregating variables from different sources. It was an educating project, and we expanded our knowledge and experience.

We concluded that random effects models are useful for our data structure since they explain response better than other modeling methods. The outcomes we derived from our random effects models are compatible with the literature.

V. ACKNOWLEDGE

For creating a great environment for research and participation we are so glad to accomplish this project under supervision of Assist. Prof. Fulya Gökalp Yavuz. She supported our ideas and motivation. This made us learn better with applications.

VI. FURTHER WORK

- Better transformation for response
- Seeking a better modeling algorithm since the data is a socio-economic data.
- Lowering standard errors for random effects models and obtaining a more reliable model
- Outlier structure and detailed analysis is demanded.
- Employing different regressor formulas for machine learning section is preferred and examining the harmony with random effects models is needed.

VII. APPENDIX

Section1: Response distribution and structure

Response skewness, mean, median for different time points:

year	skewness	mean	median
2009	0.7	-4,601.05	-6,096.5
2010	2.46	-3,194.62	-5,859
2011	2.38	5,576.271	-1,772.5
2012	-1.24	6,813.95	-3,416.5
2013	-3.07	5,072.27	-1,540
2014	-2.67	12,253.95	-1,654.5
2015	-3.59	-8,346.77	-1,964
2016	-3.9	4,435.49	-641

Table 5 Skewness for net migration for different time points

Section 2: Covariance and correlation structure for response

2009	2010	2011	2012	2013	2014	2015	2016
-58.881	-141.355	418.796	105.905	48.076	255.611	-281.739	-90.238
-34.116	-29.330	-24.465	-19.946	-16.845	-14.265	-12.240	-10.887
-33.857	-33.071	-16.833	-35.119	-29.092	-36.088	-36.371	-36.227
77.264	90.016	94.709	96.056	77.738	120.311	3.983	-24.720
-7.869	-4.970	-1.011	3.127	5.369	5.801	5.522	5.215
-32.000	-31.991	-29.214	-28.731	-28.257	-27.763	-27.283	-26.804

Table 6 Covariance Matrix for Response

	2009	2010	2011	2012	2013	2014	2015	2016
2009	1.000	0.713	0.604	0.511	0.344	0.439	0.459	0.369
2010	0.713	1.000	0.837	0.782	0.510	0.667	0.684	0.611
2011	0.604	0.837	1.000	0.888	0.649	0.742	0.720	0.728
2012	0.511	0.782	0.888	1.000	0.623	0.829	0.818	0.818
2013	0.344	0.510	0.649	0.623	1.000	0.655	0.629	0.569
2014	0.439	0.667	0.742	0.829	0.655	1.000	0.921	0.799
2015	0.459	0.684	0.720	0.818	0.629	0.921	1.000	0.799
2016	0.369	0.611	0.728	0.818	0.569	0.799	0.799	1.000

Table 7 Correlation Matrix for Response

VIII. REFERENCES

1. Bozcaga, T., Christia, F., Harwood, E., Daskalakis, C., & Papademetriou, C. (2019). Syrian Refugee Integration in Turkey: Evidence from Call Detail Records. Data for Refugees Challenge.
2. Franc, S., Časni, A. & Barišić, A. (2019). Determinants of Migration Following the EU Enlargement: A Panel Data Analysis. Southeast European Journal of Economics and Business,14(2) 13-22. https://doi.org/10.2478/jeb-2019-0010
3. Mayda, A. M. (2010). International migration: a panel data analysis of the determinants of bilateral flows. Journal of Population Economics, 23(4), 1249–1274. http://www.jstor.org/stable/40925859
4. Young, R., & Johnson, D. R. (2015). Handling Missing Values in Longitudinal Panel Data With Multiple Imputation. Journal of marriage and the family, 77(1), 277–294. https://doi.org/10.1111/jomf.12144
5. Reyna O., Panel Data Analysis Fixed and Random Effects Using Stata Lecture Notes, 2007,Princeton University https://www.princeton.edu/~otorres/Panel101.pdf
6. Anlaysis of Longitudinal Data, Digle P.,Heagerty P., Liang K., Zeger S., 2002,Oxford Statistical Science Series
7. Diggle, P. J. (1988). An Approach to the Analysis of Repeated Measurements. *Biometrics*, *44*(4), 959–971. https://doi.org/10.2307/2531727
8. Testing The Assumptions of Multilevel Models, Palmeri M., https://ademos.people.uic.edu/Chapter18.html#611_how_do_you_test_this_assumption
9. Lecture Notes University of Texas at Austin https://web.ma.utexas.edu/users/mks/384Esp08/randeff.pdf
10. ILK DAĞ Ö., Lecture Notes for Panel Data Analysis, 2022-2023 Semester

IX. CODES

1. Codes for Exploratory data analysis

a. Data installing and feature engineering:

```
#install.packages("WDI")
library(WDI)
#install.packages("plm")
library(plm)
library(readr)
library(tidyverse)
library(dplyr)
library(mice)
library(DataExplorer)
library(tibble)
library(VIM)
library(ggplot2)
library(plotly)
#install.packages("devtools")
library(devtools)
#devtools::install_github("hrbrmstr/streamgraph")
library(streamgraph)
#install.packages("rgl")
library(rgl)
```

```
#install.packages("gganimate")
library(gganimate)
library(remotes)
#install_github('vincentarelbundock/countrycode')
library(countrycode)
library(naniar)
library(e1071)
library(lme4)
library(tidyr)

wdi<-WDI

wdi<-WDI(country = c("USA","CHN","HKG","JPN","DEU",
                     "IND","GBR","ARE","FRA","CAN",
                     "RUS","ITA","IRN","BRA","PRK",
                     "AUS","AUT","MEX","ESP","IDN",
                     "SAU","TUR","TKM","NLD","CHE",
                     "POL","ARG","SWE","BEL","THA",
                     "ISR","IRL","NOR","NGA","EGY",
                     "BGD","MYS","SGP","VNM","ZAF",
                     "PHL","DNK","PAK","COL","CHL",
                     "ROU","CZE","IRQ","FIN","PRT",
                     "NZL","PER","KAZ","GRC","QAT",
                     "UKR","DZA","HUN","KWT","MAR",
                     "AGO","PRI","ECU","KEN","SVK",
                     "SVN","DOM","ETH","OMN","CUB",
                     "GTM","BGR","LUX","VEN","BLR",
                     "UZB","TZA","GHA","COD","SRB",
                     "MMR","UGA","JOR","TUN","CMR",
                     "BHR","BOL","SDN","PRY","LBY",
                     "LVA","EST","NPL","ZWE","SLV",
                     "PNG","HND","TTO","KHM","ISL",
                     "YEM","SEN","ZMB","CYP","GEO",
                     "BIH","MAC","GAB","HTI","AFG",
                     "GIN","BRN","MLI","BFA","ALB",
                     "LBN","BWA","MOZ","ARM","BEN",
                     "MLT","GNQ","LAO","JAM","KOR",
                     "MNG","NIC","SYR","MDG","GUY",
                     "LKA","URY","PAN","AZE","HRV",
                     "CRI","LTU","NER","MKD","MDA",
                     "TCD","BHS","NAM","RWA","MWI",
                     "MUS","MRT","TJK","KGZ","NCL",
                     "BLZ","BTN","BDI","CAF","COM",
                     "CIV","DJI","SWZ","GMB",
                     "LSO","LBR","MNE","SLE","SOM",
                     "SSD","SUR","TGO"),
        indicator = c("SM.POP.NETM", "PV.EST", "SL.UEM.TOTL.ZS", "SG.LAW.INDX",
"FP.CPI.TOTL"),
        start=2009,
        end=2016,
        extra=FALSE,
        cache=NULL)

#View(wdi)
dim(wdi)
```

```
sapply(wdi, function(x) sum(is.na(x)))
## Checking the percentages of NA values
(colMeans(is.na(wdi)))*100

(sum(is.na(wdi))/prod(dim(wdi)))*100

glimpse(wdi)

summary(wdi)

# Remove columns using select()
wdi <- wdi %>% select(-c(iso2c, iso3c))

arcp<- c(100,100,122.06,97.95,156.35,316.34,435.38,567.96)
wdi[33:40,7]<-arcp
head(wdi)

happ <- read_csv("happiness-cantril-ladder.csv")
head(happ)

happ <- happ %>% select(c(Entity, Year, 'Cantril ladder score'))

colnames(happ)[1] <- "country"
colnames(happ)[2] <- "year"

length(unique(wdi$country))
length(unique(happ$country))

unique(happ$country)
unique(wdi$country)

wdi[wdi$country == "Congo, Dem. Rep.", "country"] = "Democratic Republic of Congo"
wdi[wdi$country == "Egypt, Arab Rep.", "country"] = "Egypt"
wdi[wdi$country == "Gambia, The", "country"] = "Gambia"
wdi[wdi$country == "Hong Kong SAR, China", "country"] = "Hong Kong"
wdi[wdi$country == "Iran, Islamic Rep.", "country"] = "Iran"
wdi[wdi$country == "Kyrgyz Republic", "country"] = "Kyrgyzstan"
wdi[wdi$country == "Lao PDR", "country"] = "Laos"
wdi[wdi$country == "Russian Federation", "country"] = "Russia"
wdi[wdi$country == "Slovak Republic", "country"] = "Slovakia"
wdi[wdi$country == "Syrian Arab Republic", "country"] = "Syria"
wdi[wdi$country == "Turkiye", "country"] = "Turkey"
wdi[wdi$country == "Venezuela, RB", "country"] = "Venezuela"
wdi[wdi$country == "Yemen, Rep.", "country"] = "Yemen"

df <- merge(wdi, happ, by=c("country", "year"))
dim(df)
(sum(is.na(df)/prod(dim(df)))*100)
head(df)

# assigning new names to the columns of the data frame
colnames(df) <- c('country','year','migration', 'pol.stability', 'unemp', 'women.bl.scr', 'price.idx',
'happiness.scr')

glimpse(df)
```

```
summary(df)

unique(df[c("country", "year")])

uni <- df %>%
  group_by(country) %>%
  summarise(n_unique = n_distinct(unlist(year))) %>%
  ungroup()
uni

sum(uni$n_unique < 4)
dplyr::filter(uni, n_unique < 4)

df %>% group_by(year) %>% summarize(count=n())

# adding rows for missing years

library(tidyr)
df <- df %>%
   complete(country, year = 2009:2016,
            fill = list(incidents = 0)) %>%
   as.data.frame()

# removing countries with less observation

df<-df[-c(105:112,313:320,361:368,857:864,1057:1064,
          121:128,177:184,401:408,209:216,249:256, 633:640,
          785:792,921:928,1009:1016,1025:1032,1121:1128),]
length(unique(df$country))

df$continent <- countrycode(sourcevar = df[, "country"],
                            origin = "country.name",
                            destination = "continent")

(sum(is.na(df)/prod(dim(df)))*100)

df %>% group_by(year) %>% summarize(count=n())
```

b. Missing value imputation:

```
missplot<-ggplot(df,
       aes(x = year,
           y = migration)) +
  geom_miss_point() +
  facet_wrap(~continent)
missplot #for migration

df %>%
  select(-c('year')) %>%
  keep(is.numeric) %>%
```

```
  gather() %>%
  ggplot(aes(value, colour = "orange")) +
    facet_wrap(~ key, scales = "free") +
    geom_density()

df %>%
  select(-c('year')) %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram(color="darkblue", fill="lightblue",bins=70)

plot_missing(df)

aggr_plot <- aggr(df, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(df),
cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Pattern"))

init = mice(df, maxit=0)
init

meth = init$method # shows the method of imputation for each variable
meth

predM = init$predictorMatrix
predM

set.seed(12)
imputed.wo.age = mice(df, method="cart", predictorMatrix=predM, m=5,  maxit = 0)

imputed <- complete(imputed.wo.age)

colSums(is.na(imputed))

df <- imputed
summary(df)
```

c. Exploratory data analysis:

```
dfx<-df[,3:8]
dfy<- df[,3]
cdf<-cor(dfx,method="spearman")
library(corrplot)
corrplot(cdf)
a <- plot_correlation(cdf)
a


df %>%
  select(-c('year')) %>%
  keep(is.numeric) %>%
```

```
  gather() %>%
  ggplot(aes(value, colour = "purple")) +
    facet_wrap(~ key, scales = "free") +
    geom_density()

df %>%
  select(-c('year')) %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram(color="red", fill="pink",bins=70)

set.seed(300)
s = df[sample(1:nrow(df),200),]
ggplot(s,aes(x=happiness.scr,y=women.bl.scr,size=price.idx,col=continent))+geom_point()+labs(title
="-------")+ scale_color_brewer(palette="Paired")

p <- df %>%
  filter(year==2015) %>%
  ggplot(aes(happiness.scr, women.bl.scr, size = price.idx, color=continent)) +
  geom_point() +
  theme_bw()

ggplotly(p)

ts<-dplyr::filter(df, country %in% c("Turkey", "Syria"))
g7count<-dplyr::filter(df, country %in% c("Germany","Canada","Italy","France","United
Kingdom","Japan","United States"))

pp <- streamgraph(df, key="country", value="migration", date="year", interpolate="cardinal") %>%
  sg_axis_x(1, "year", "%Y") %>%
  sg_axis_y(10) %>%
  sg_legend(show=TRUE, label="names: ")
pp
# save the widget
# library(htmlwidgets)
# saveWidget(pp, file=paste0( getwd(), "/HtmlWidget/streamgraphDropdown.html"))

p <- ggplot(
  df,
  aes(x = migration, y=pol.stability, size = unemp, colour = continent)
  ) +
  geom_point(show.legend = FALSE, alpha = 0.7) +
  scale_color_viridis_d() +
  scale_size(range = c(2, 12)) +
  scale_x_log10() +
  labs(x = "migration", y = "political stability")
p

fig <- df %>%
  plot_ly(
    x = ~price.idx,
    y = ~happiness.scr,
    size = ~migration,
```

```
    color = ~continent,
    frame = ~year,
    text = ~country,
    hoverinfo = "text",
    type = 'scatter',
    mode = 'markers'
  )
fig <- fig %>% layout(
    xaxis = list(
      type = "log"
    )
  )

fig

ggplot(ts, aes(x = year, y = migration, colour = country)) +
  geom_line() +
    geom_point(size = 2)

ggplot(df, aes(x = year, y = migration, group = country, color = continent)) +
  geom_line() +
  labs(x = "Year", y = "Net Migration") +
  scale_color_discrete(name = "Continent") +
  theme_minimal()
```

2. Selected codes for modeling

d. Codes for transformation:

```
library(e1071)
skewness(migdata) #-2.71 left skewed data power transformation needed
logmig<-log(data$migration) #nans produced (-) exists
sqmig<-(migdata)^2
shapiro.test(sqmig) #2.2e-16
hist(sqmig)
skewness(sqmig)
library(MASS)
bxcx<-boxcox(migdata)
library(bestNormalize)
bst<- bestNormalize(migdata)
nwbst<-bst$x.t
hist(nwbst)
shapiro.test(nwbst) #p-value=1
skewness(nwbst) #0.0000525 almost normal?
bst$method
# "Out-of-sample via CV with 10 folds and 5 repeats"
bst #ordernorm is selected interpretation is so hard...
data$nwbst<-nwbst #I don't know how to make inferences out of this
```

e. Codes for random effects

```
library(lme4)
model1 <- lmer(migration ~ pol.stability+unemp+women.bl.scr+price.idx+happiness.scr +
(1|country), data = data2)
summary(model1)
ranef(model1)
VarCorr(model1)
VarCorr(model2)
anova(model4,model6,test="Chisq")
glmer1<-glmer(migth ~ scale(pol.stability)+scale(women.bl.scr)*scale(price.idx)+scale(unemp)+
scale(happiness.scr)+
        (1 | country), data = data,family="gaussian")
summary(glmer1)
#same with lmer()

library(nlme)
nlm1<-#couldn't perform the model became so complicated.

plot(model7)
model7$residual
shapiro.test(model7$residuals)
```

f. Codes for linear regression

```
data3<-data2
data3$year<-as.factor(data3$year)
lm1<- lm(thmig~year+pol.stability+unemp+women.bl.scr+price.idx+happiness.scr,data3)
summary(lm1)
library(car)
vif(lm1) #no vif
```

g. Codes for machine learning

#machine learning part:

```
data2<-read.csv("dflast.csv")
install.packages("LongituRF")
library(LongituRF)
install.packages("groupdata2")
library(groupdata2)
library(xgboost)
library(dplyr)
data2$thmig<-(data2$migration)/10000
data2$country = factor(data2$country)
data2$region=factor(data2$region)
#split train-test data
test.train.d=partition(data2,p = 0.2, id_col = "country")
test.train.d
test.d = test.train.d %>% .[1]
test.d
train.d=test.train.d %>% .[2]
```

```
train.d
#setting my fixed and random effects
X.fixed.effects <- as.data.frame(train.d[[1]][,c(3,5,6,7,8,9)])
z=cbind(rep(1,dim(train.d[[1]])[1]),train.d[[1]]$year)  # tried a random intercept and slope model
y=as.matrix((train.d[[1]]$thmig))
#ml1 model with random slope
ml1<-
REEMtree(Y=y,X=X.fixed.effects,Z=z,id=train.d[[1]]$country,time=train.d[[1]]$year,sto="BM")
ml1$forest #n= 904
#root 904 272749.400   0.4284272 others also


ml1$forest$variable.importance


p.train = predict(ml1$forest, X=X.fixed.effects,Z=z,id=train.d[[1]]$country,
time=train.d[[1]]$year)
p.test=predict(ml1$forest,
test.d[[1]],X=as.data.frame(test.d[[1]][,c(2,3,4,6)]),Z=cbind(rep(1,dim(test.d[[1]])[1])),id=test.d[[
1]]$country, time=test.d[[1]]$year)

sum(((train.d[[1]]$thmig)-p.train)**2)/(dim(train.d[[1]])[1]-dim(X.fixed.effects)[2]) #452.3745
sum(((test.d[[1]]$thmig)-p.test)**2)/(dim(test.d[[1]])[1]-dim(X.fixed.effects)[2]) #961.8957

ml1$random_effects  #too small

######## setting z2:
z2=cbind(rep(1,dim(train.d[[1]])[1]))
ml2 <-
REEMtree(Y=y,X=X.fixed.effects,Z=z2,id=train.d[[1]]$country,time=train.d[[1]]$year,sto="BM"
)
ml2$forest #n= 904
#root 904 524592.700   0.7733656
plot(ml2$forest,ylim=c(-0.5,1)) #works in rscript
text(ml2$forest, use.n=TRUE, all=TRUE, cex=1)

ml2$forest$variable.importance
p.train2 = predict(ml2$forest, X=X.fixed.effects,Z=z,id=train.d[[1]]$country,
time=train.d[[1]]$year)
p.test2=predict(ml2$forest,
test.d[[1]],X=as.data.frame(test.d[[1]][,c(2,3,4,6)]),Z=cbind(rep(1,dim(test.d[[1]])[1])),id=test.d[[
1]]$country, time=test.d[[1]]$year)


sum(((train.d[[1]]$thmig)-p.train2)**2)/(dim(train.d[[1]])[1]-dim(X.fixed.effects)[2]) #442.0176
sum(((test.d[[1]]$thmig)-p.test2)**2)/(dim(test.d[[1]])[1]-dim(X.fixed.effects)[2]) #982.0563


#let's use REEMforest function:
#forest didn't worked.
X.fixed.effects2 <- as.data.frame(train.d[[1]][,c(3,5,6,7,8,9)])
ml3 <-
REEMforest(Y=y,X=X.fixed.effects2,Z=z2,id=train.d[[1]]$country,time=train.d[[1]]$year,sto="B
M",mtry=2)
```