DOPP 2019W Exercise 3
Group 32

Analysis of flows of refugees between countries

Eszter Katalin Bognar - 11931695
Luis Kolb - 01622731
Alexander Leitner - 01525882

Objectives of the analysis

- What is the most accurate overview of flows of refugees between countries that can be obtained?
- Are there typical characteristics of refugee origin and destination countries?
- Are there typical characteristics of large flows of refugees?
- Can countries that will produce large numbers of refugees be predicted? Can refugee flows be predicted?

Data sources

- OECD International Migration Database data (https://stats.oecd.org/Index.aspx?DataSetCode=MIG)
- Gross Domestic Product per Capita data (https://data.worldbank.org/indicator/NY.GDP.PCAP.CD)
- Human Development Index data (http://hdr.undp.org/en/data)
- World Governance Index data (https://datacatalog.worldbank.org/dataset/worldwide-governance-indicators)

Data preprocessing

Each dataset were loaded and formatted including:

- reshaping columns (changing rows to columns or columns to rows when necessary),
- getting rid of unwanted columns,
- renaming columns,
- setting proper data types,
- setting country-year multiindex to facilitate future data merge.

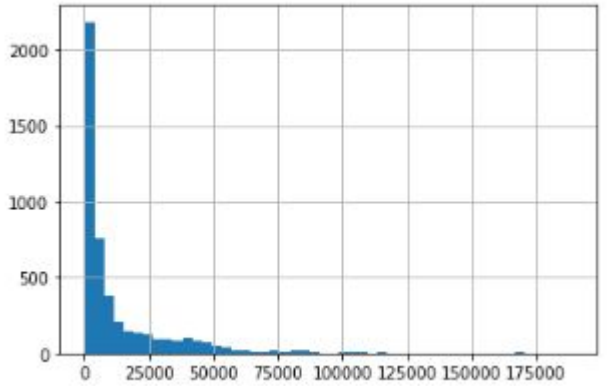
Resolving Country name inconsistency

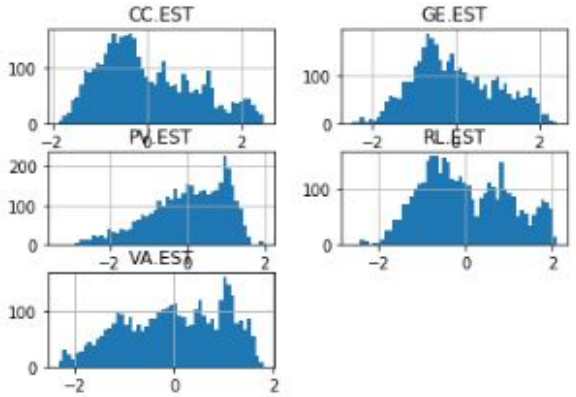
- We searched for different usage and typos in county names.
- We selected oecd_df as base dataframe so we compare the country names in the oecd_df to the country names in the hdi_df, gdp_df, wgi_df datasets.
- For making the country names consistent, we first tried out the fuzzy search method of the fuzzywuzzy library.
- Due to errors, we finally decided to manually create a dictionary of country names to replace or delete them.

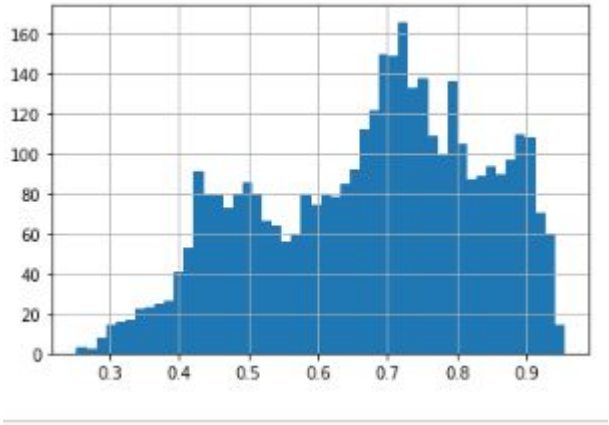
```
country_dict = {
    'Bahamas, The' :"Bahamas",
    'Bolivia (Plurinational State of)' :"Bolivia",
    'Cabo Verde' :"Cape Verde",
    'Congo, Rep.' :"Congo",
    'Czechia' :"Czech Republic",
    "Cote d'Ivoire" :"Côte d'Ivoire",
    'Congo, Dem. Rep.' :"Democratic Republic of the Congo",
    'Congo (Democratic Republic of the)':"Democratic Republic of the Congo",
    'Egypt, Arab Rep.' :"Egypt",
    'Gambia, The' :"Gambia",
    'Iran (Islamic Republic of)' :"Iran",
    'Iran, Islamic Rep.': "Iran",
    'Korea (Republic of)' :"Korea",
    'Korea, Rep.':"Korea",
    'Kyrgyz Republic' :"Kyrgyzstan",
    "Lao People's Democratic Republic" :"Laos",
    "Lao PDR":"Laos",
    'Micronesia (Federated States of)' :"Micronesia",
    'Micronesia, Fed. Sts.':'Micronesia',
    'Moldova (Republic of)' :"Moldova",
    'Russian Federation' :"Russia",
    'St. Kitts and Nevis' :"Saint Kitts and Nevis",
    'St. Lucia' :"Saint Lucia",
    'St. Vincent and the Grenadines' :"Saint Vincent and the Grenadines",
    'Slovakia' :"Slovak Republic",
    'Syrian Arab Republic' :"Syria",
    'Eswatini' :"Swaziland",
    "Eswatini (Kingdom of)": "Swaziland",
    'Tanzania (United Republic of)' :"Tanzania",
    'Venezuela (Bolivarian Republic of)' :"Venezuela",
    "Venezuela, RB":"Venezuela",
    'Vietnam' :"Viet Nam",
    'Yemen, Rep.' :"Yemen"
}
```

Outlier detection

We checked the distribution of hdi, wgi and gdp data showing there are more poor than wealthy countries...We can not see any outliers. GDP is skewed towards zero, HDI is in the range of [0-1], wgi metrics are in the range of [-3-3].





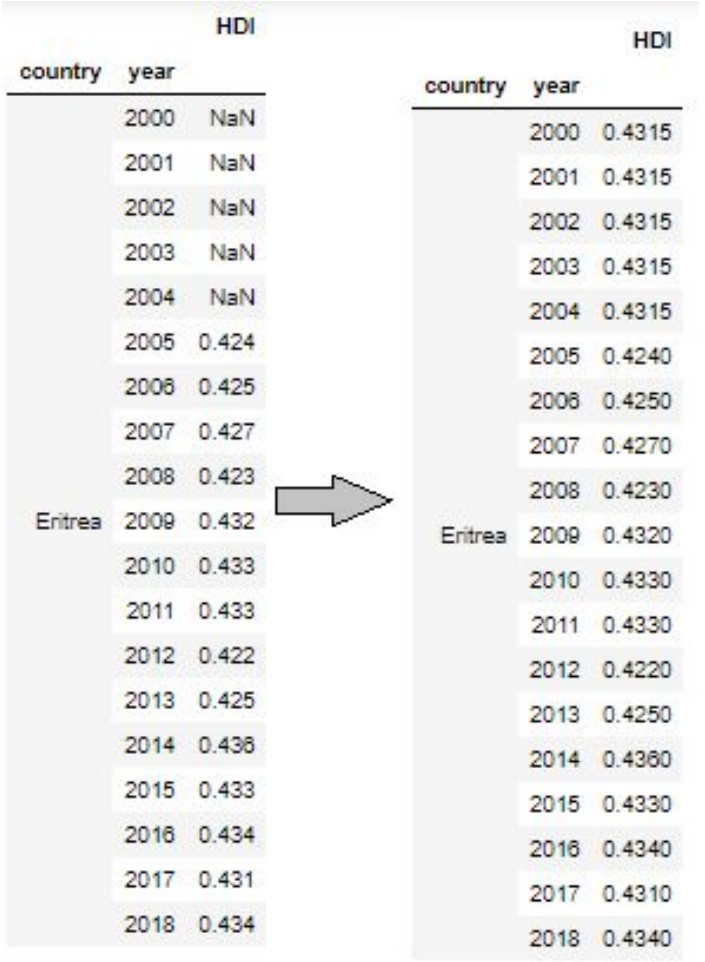


Missing value handling

Missing values in the hgi, gdp and wgi data

Since the hdi, gdp and wgi indicators can be treated equally and the values don't change rapidly from year to year we replace the missing data with the median of the data for the given country pairs.

We selected this method because interpolation can't work properly where there are many missing values one after another. Where there weren't any data available for the given country pairs, we simply dropped the rows.



country	year	HDI
Eritrea	2000	NaN
	2001	NaN
	2002	NaN
	2003	NaN
	2004	NaN
	2005	0.424
	2006	0.425
	2007	0.427
	2008	0.423
	2009	0.432
	2010	0.433
	2011	0.433
	2012	0.422
	2013	0.425
	2014	0.436
	2015	0.433
	2016	0.434
	2017	0.431
	2018	0.434

country	year	HDI
Eritrea	2000	0.4315
	2001	0.4315
	2002	0.4315
	2003	0.4315
	2004	0.4315
	2005	0.4240
	2006	0.4250
	2007	0.4270
	2008	0.4230
	2009	0.4320
	2010	0.4330
	2011	0.4330
	2012	0.4220
	2013	0.4250
	2014	0.4360
	2015	0.4330
	2016	0.4340
	2017	0.4310
	2018	0.4340

Missing value handling

Missing values in the oecd dataset

- There are a number of country pairs (e.g. Albania-Chile) where none of the years have inflows of asylum seekers between countries. Since we could not find any similar data source where there was appropriate data available for cold deck imputation. We assume that migration is not considerable between these countries and we decided to delete these rows from the final dataset.
- For the remaining missing values, we calculated the median of asylum_seekers for the given country pairs and filled the holes with this value.

Final dataset validation

	source	destination	asylum_seekers	year	s_GDP	s_HDI	s_PV.EST	d_GDP	d_HDI	d_PV.EST
0	Syria	Germany	266248.0	2016	1335.216773	0.539	-2.916323	42098.92045	0.936	0.681363
1	Syria	Germany	158657.0	2015	1335.216773	0.540	-2.974081	41139.54457	0.933	0.699824
2	Afghanistan	Germany	127011.0	2016	547.228110	0.491	-2.671054	42098.92045	0.936	0.681363
3	Iraq	Germany	96115.0	2016	4776.726499	0.672	-2.313588	42098.92045	0.936	0.681363
4	Afghanistan	Turkey	66459.0	2017	556.302138	0.493	-2.800609	10513.64842	0.805	-1.788224
5	Syria	Hungary	64081.0	2015	1335.216773	0.540	-2.974081	12651.56834	0.835	0.746298
6	Afghanistan	Turkey	63292.0	2015	578.466353	0.490	-2.571222	10948.72461	0.800	-1.493914
7	Iraq	Turkey	56332.0	2015	4989.803075	0.665	-2.260352	10948.72461	0.800	-1.493914
8	Albania	Germany	53805.0	2015	3952.829458	0.788	0.346129	41139.54457	0.933	0.699824
9	Syria	Sweden	50909.0	2015	1335.216773	0.540	-2.974081	51397.19176	0.932	0.947106

- **We can see that the most remarkable refugee inflow was in 2015-2017 from Syria, Afghanistan, Iraq and Albania and Serbia to Germany, Turkey and Hungary. This coincide with the recent news about massive influx of refugees from the Middle East to Europe through the Balkan route.**
- **We can see that the HDI and GDP indexes are very low of the refugee producer countries and also the Political Stability and Absence of Violence/Terrorism indicator shows problems.**
- **Destination countries usually have higher gdp, hdi values and more political stability.**
- **We can see that our dataset is comprehensive and good source for further analysis.**

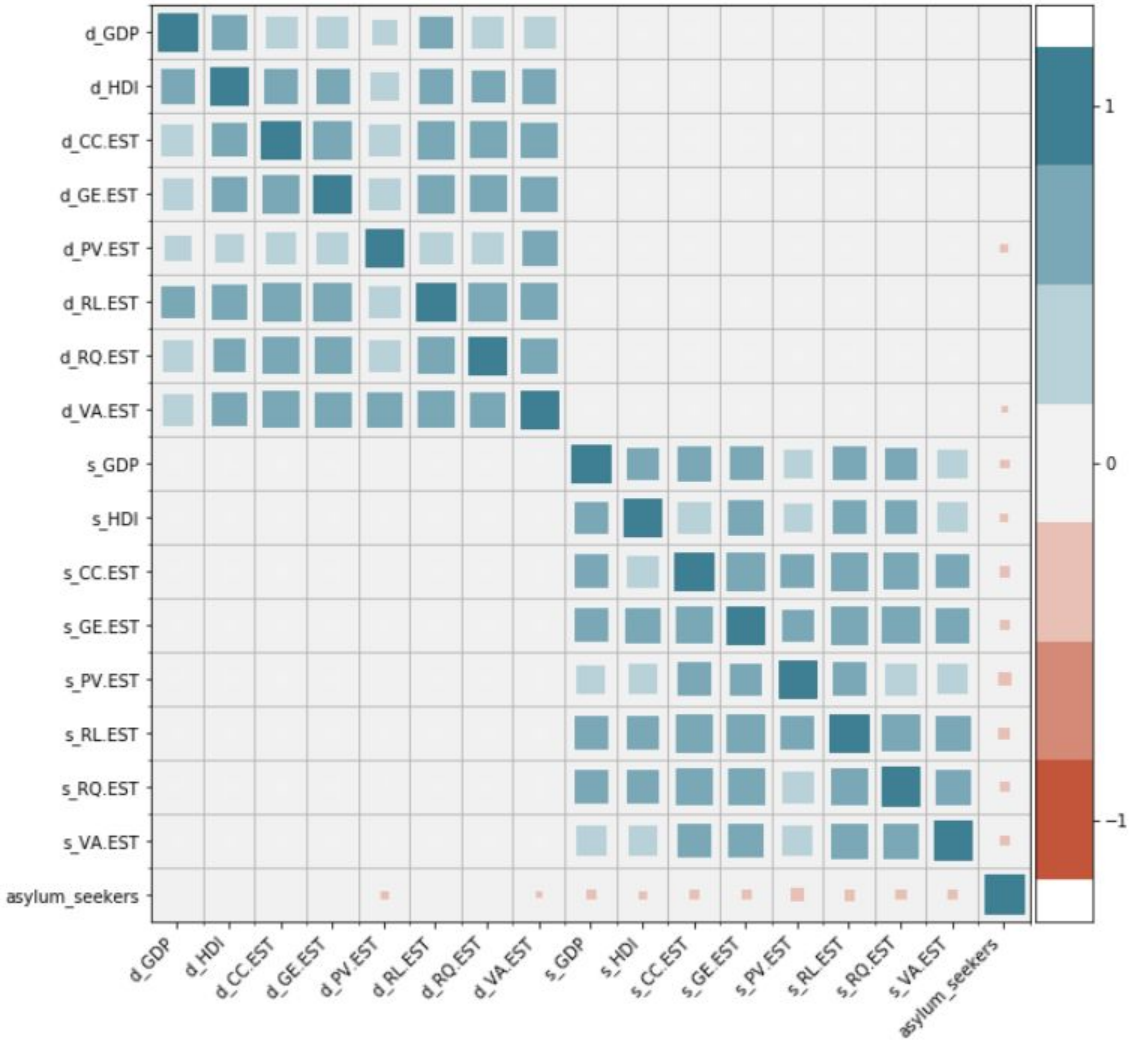
Correlation between refugees and Indicators

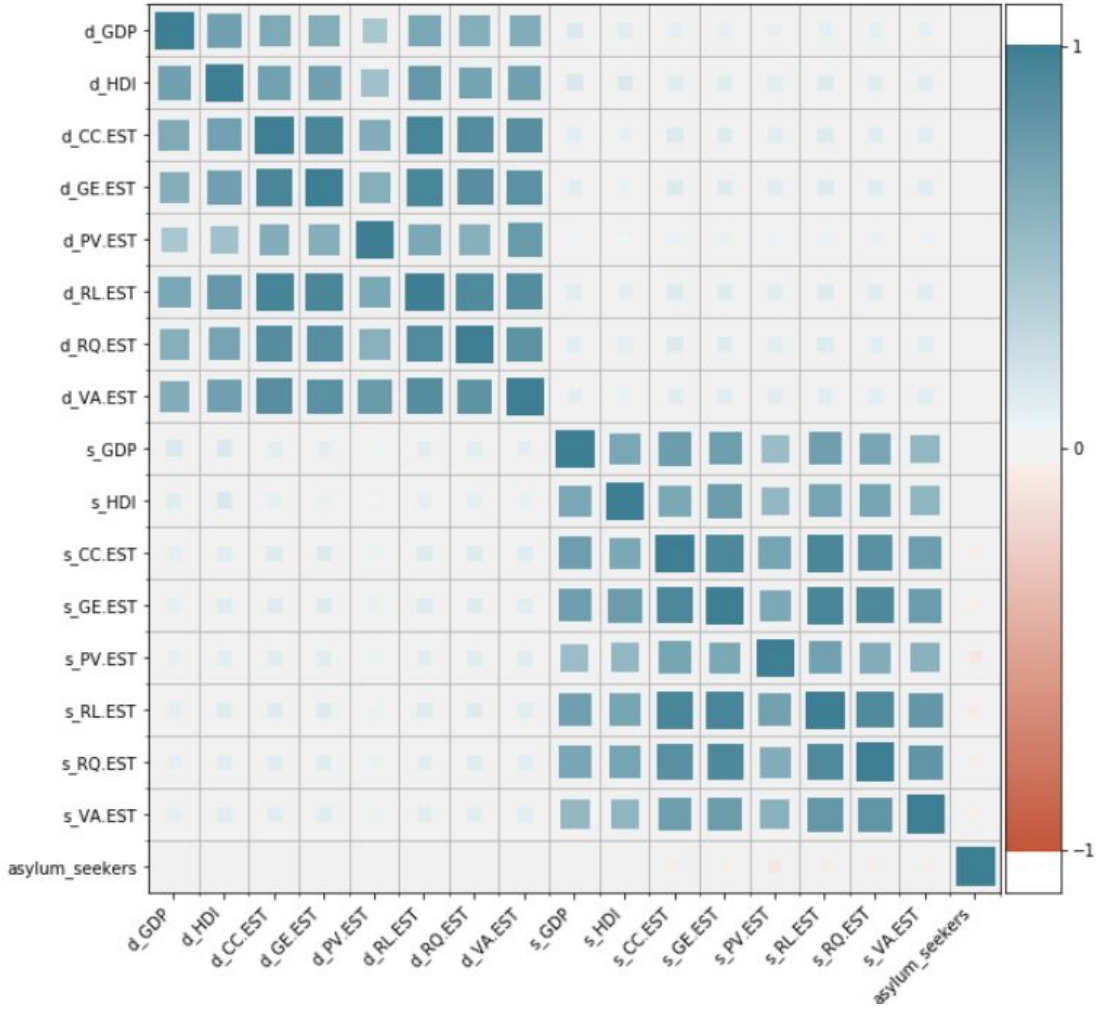
- Are there typical characteristics of refugee origin and destination countries?

- Are there typical characteristics of large flows of refugees?

Are there typical characteristics of refugee origin and destination countries?

The World Governance Indicators of origin countries negatively correlate with the number of asylum seekers originating from there - when some Indicators decrease, the number of refugees originating from that country tends to increase. Important to notice, HDI and GDP have no correlation with the number of refugees fleeing a country or choosing a country to flee to.


d_GDP
d_HDI
d_CC.EST
d_GE.EST
d_PV.EST
d_RL.EST
d_RQ.EST
d_VA.EST
s_GDP
s_HDI
s_CC.EST
s_GE.EST
s_PV.EST
s_RL.EST
s_RQ.EST
s_VA.EST
asylum_seekers
1
0
−1


d_GDP
d_HDI
d_CC.EST
d_GE.EST
d_PV.EST
d_RL.EST
d_RQ.EST
d_VA.EST
s_GDP
s_HDI
s_CC.EST
s_GE.EST
s_PV.EST
s_RL.EST
s_RQ.EST
s_VA.EST
asylum_seekers
1
0
−1

Are there typical characteristics of large flows of refugees?

Typically, large flows of refugees originate from countries with rather negative Indicators from the WGI, as well as lower HDI and GDP. Destination countries typically register a much higher GDP and HDI, as well as much better/higher Indicators from the WGI.

Again, historical, geographical and cultural events and aspects play a huge part.

	source	s_GDP	s_HDI	s_CC.EST	s_GE.EST	s_PV.EST	s_RL.EST	s_RQ.EST	s_VA.EST	asylum_seekers_produced_sum
161	Syria	1400.40549	0.60629	-1.13345	-1.07885	-1.26378	-0.96129	-1.28780	-1.75172	1058638.01040
0	Afghanistan	439.99645	0.44165	-1.45318	-1.39201	-2.44404	-1.69196	-1.49303	-1.24803	924854.31259
78	Iraq	4590.02118	0.64559	-1.35680	-1.41028	-2.30838	-1.58808	-1.35903	-1.26924	871230.76282
146	Serbia	5055.69663	0.75818	-0.41791	-0.17913	-0.38301	-0.49918	-0.23822	0.10631	517606.84198
35	China	4400.65216	0.68429	-0.43605	0.10478	-0.47740	-0.47327	-0.25674	-1.60329	383557.78544
137	Russia	8949.88269	0.77594	-0.95729	-0.36914	-0.97762	-0.85290	-0.34398	-0.84051	375794.58728
77	Iran	4806.11363	0.74424	-0.61310	-0.47198	-1.02821	-0.85116	-1.43291	-1.41866	347982.90506
126	Pakistan	973.51457	0.51465	-0.92488	-0.61937	-2.22600	-0.83695	-0.65790	-0.89728	341597.82172
122	Nigeria	1924.17722	0.49447	-1.15770	-1.02822	-1.89812	-1.14849	-0.87758	-0.64214	306990.74389

	destination	d_GDP	d_HDI	d_CC.EST	d_GE.EST	d_PV.EST	d_RL.EST	d_RQ.EST	d_VA.EST	asylum_seekers_sum
32	Turkey	9080.75903	0.73371	-0.06707	0.20175	-1.04689	0.02219	0.25859	-0.20267	216991.92134
21	Mexico	9018.07242	0.73982	-0.41828	0.18820	-0.56801	-0.47473	0.33923	0.13974	196712.14677
27	Slovak Republic	14542.94618	0.81776	0.25669	0.80977	0.91969	0.50704	0.97062	0.93680	21735.51872
12	Hungary	11976.43968	0.81565	0.44296	0.71943	0.82006	0.75729	1.00874	0.87135	20600.72086
5	Czech Republic	16607.19451	0.85265	0.38266	0.93870	0.94477	0.95285	1.11204	0.98441	13585.14027
25	Poland	10737.46196	0.82865	0.50028	0.58842	0.69921	0.62926	0.88290	0.98222	10849.94853
34	United States	48362.83819	0.90500	1.47007	1.57840	0.43354	1.59238	1.49090	1.17255	9889.36537
15	Israel	28674.71316	0.88294	0.93755	1.25818	-1.21114	0.96264	1.11999	0.67829	8776.99060

Can countries that will produce large numbers of refugees be predicted?

First we look for the countries which produce the largest number of refugees

Then use Lasso regression to predict the data from 2017 (based) on 2000-2016

```
countries = []
countries = ["Afghanistan","China","Chad","Algeria"]
```

Results of the prediction

```
'Chad', score     0.2991
'China', score     0.1731
'Croatia', score     0.0125
'Algeria', score     0.3709
```

Can refugee flows be predicted?

Difficult to answer

suddenly WAR!

environmental disasters (volcanic eruptions, drought…..)