WORLDWIDE MIGRATION PATTERNS ANALYSIS

ANALYTIC HOUND'S DATA DEPARTMENT | FEBRUARY 2023

DISCLAIMER:

Analytic Hound ® is a consulting company based in Buenos Aires (Argentina) with no socio-political or economic conflict of interest regarding the subject of study. The following documentation is published in good faith and for general information purposes only. We do not make any warranties about the completeness, reliability and accuracy of this information. Any action you may take upon the information here present is strictly at your own risk. We are not liable for any losses and damages in connection with the use of this document.

Opinions that may be expressed along this documentation are strictly the author's responsibility. Henry's company ® , project owner or any other entity that could be related to this analysis have no relationship whatsoever.

The Analytic Hound team ®

TEAM MEMBERS:

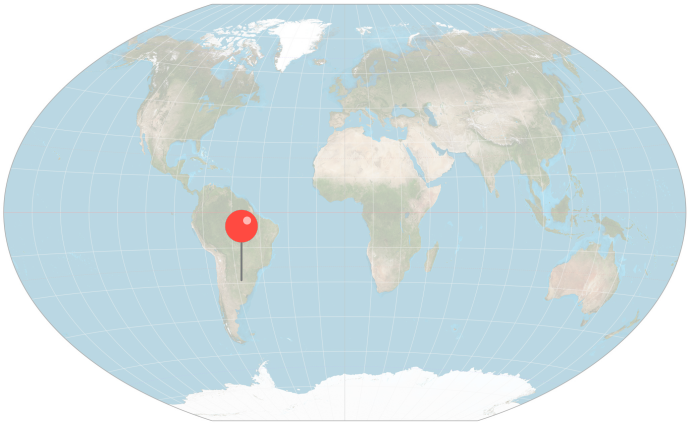
MYSLER, ALAN	***ZAPATA, BELEN***	***RODRIGUEZ, LUCAS***	***BELL, EUGENIA***	***PINI, MAURO***
DATA SCIENTIST	*DATA ANALYST*	*DATA ENGINEER*	*DATA ENGINEER*	*FUNCTIONAL ANALYST*





ABOUT US:

Analytic Hound ® is a traditional highly specialized consulting firm based in Buenos Aires, Argentina. Our mission is to improve the society in which we live through consultancy and, to achieve this, we study, design, execute and evaluate projects and actions that can improve society as we know it. This goal is achieved whether they are promoted by the public sector (NGO, governments, etc.) or arise from the private sector.

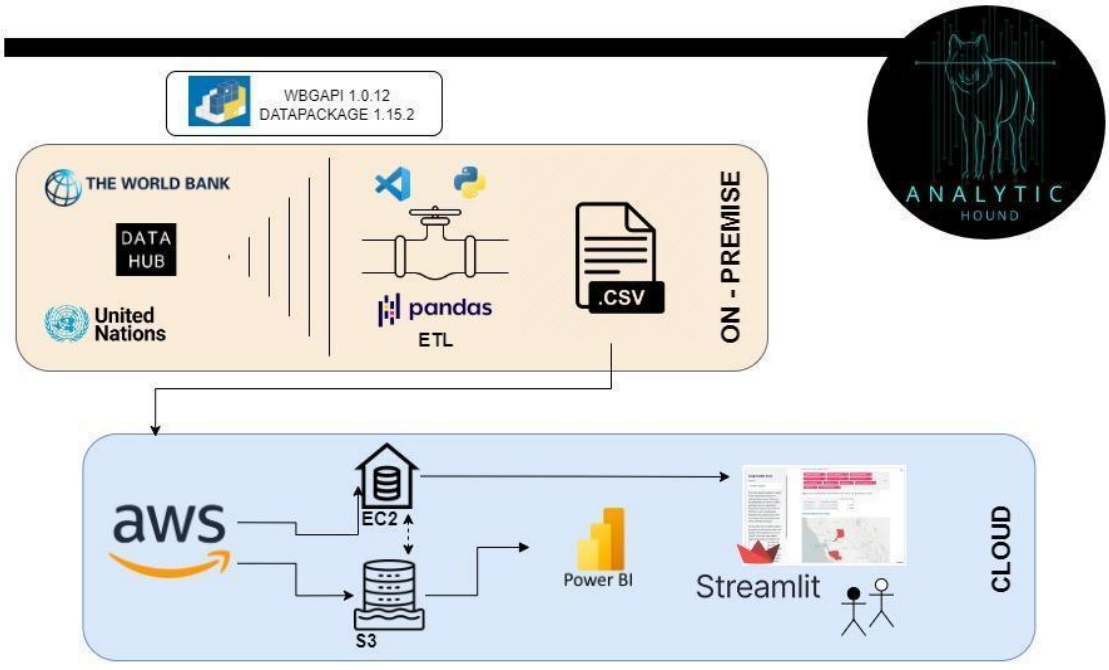


ABSTRACT:

Throughout history, worldwide migratory flows have undergone multiple changes. These changes are the translation of the different sociopolitical, demographic, environmental and economic aspects among others. The ***Analytic Hound ®*** consulting team, with the contribution of data provided by ***Data hub, World bank group*** and ***United nations***, has developed an interactive platform for the dynamic analysis of human movement around the world starting from the year 2000.

The following graph schematically reflects the synthesis of the entire workflow related to the "**Worldwide migration patterns analysis**" project. It should be used as a reference map, which can be consulted throughout all this documentation. Its purpose is to bring the reader a roadmap that facilitates understanding of the correct order and assembly of the different gears that come into play at the time of this project's development.

The detail and description of each stage will be addressed in the corresponding section. The reader is invited to make use of the index of this instrument to quickly locate any specific doubts that may arise.

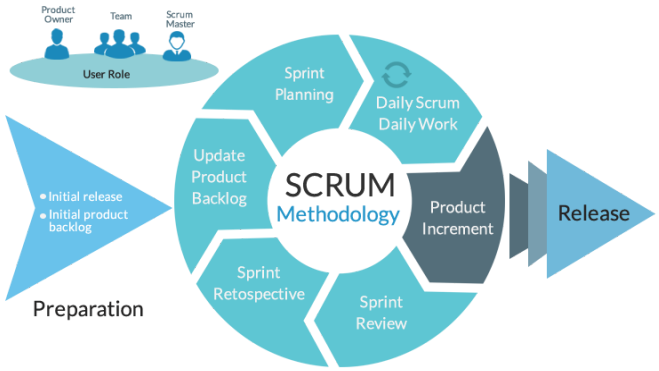


INDEX

1. INTRODUCTION

1.1 WORK METHODOLOGY

The ***Analytics Hound*** ® team has knowledge and training in the working methodology popularly known as **SCRUM**. The tools provided by this modality will serve as a framework to effectively coordinate and expedite the different tasks related to this project's development.

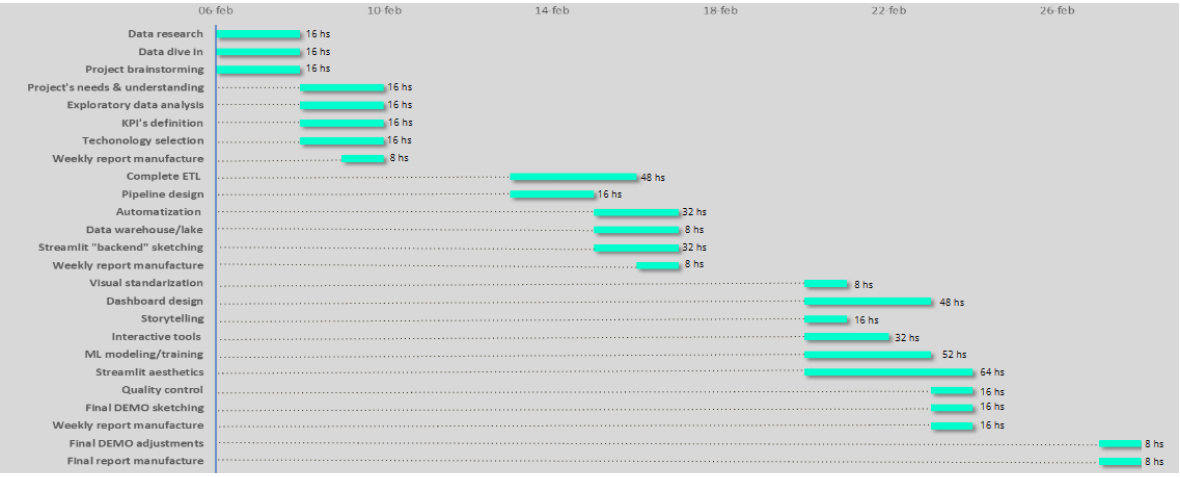


Mainly, the focus will be oriented to:

- ***Sprint planning meeting:*** At the beginning of each weekly sprint with the entire team, perform an inspection of the product backlog (tasks, requirements and functionalities that the project requires for that specific week), selecting the corresponding items that are required to work on during the weekly Sprint.
- ***Daily Scrum:*** Daily, in the morning prior to the meeting with our external advisor *Gonzalo Posse*, the team will meet in this space. A maximum of 15 minutes will be allocated to monitor the project and control compliance with the assumed tasks. In addition, the daily work objectives will be agreed and possible problems that have directly limited or prevented the fulfillment of the objectives will be analyzed.
- ***Sprint Review***: At the end of the weekly Sprint, after meeting with our project owner (PO) *Jonathan Deiloff*, the team will come together to retrospectively

evaluate performance. The goal is to reflect on the performance of the Sprint and identify opportunities for improvement for the next Weekly Sprint.

In order to obtain a global idea of working effort and time distribution for each different task, we've designed a ***gantt chart***. This requirement, explicitly made by the client, is destined to achieve a correct understanding of the management of time and tasks linked to the totality of this project.



1.2 SITUATION ANALYSIS

On this occasion, the report requested by our client covers the subject related to Migration Flows (MF). MF refers to the number of migrants arriving in or leaving an area in the course of a specified period of time. These zones can be countries, regions, cities or continents. The term arose to replace other terms with a more derogatory connotation, such as "mass" or "wave" of immigrants.

The phenomena of migration and human mobility are an innate characteristic of the human being, of an ancient nature and have affected almost without exception all societies in the world. Migration is a complex topic, and as such it can be distorted to alarming degrees by misinformation and politicization.

The understanding of MF from the perspective that Data Science provides us will be the characteristic criterion of our approach, without falling into ideological biases or opinions of a personal nature[1].

We believe that exploring and exposing MF in a clear, precise and scientific way is enormously useful for society in general, companies, policy makers and all the agents that are part of this world, which are directly or indirectly affected by the phenomenon.

In our preliminary research we have been able to determine the characteristics of MF variations depending on the places, times, country policies, demographic, sociopolitical, environmental and economic aspects, among others.

Our mission will be to achieve a detailed explanation of each feature's impact and value, as well as its corresponding relevance in relation to the dynamics of the MF.

1.3 OBJECTIVES

At first instance, we will conduct an exploration of different data sources provided by the client, as well as those that may arise from our preliminary investigation. Once data is finally gathered, we will try to explain the determining features that affect the MF, understanding them in depth and determining the most important ones. We will carry out flow analysis by zones and continents, to broadly understand the scale, direction and frequency of the movements.

Subsequently, we will sift said data at the country level, and we will make an exhaustive study of their characteristic features, classifying them into various categories according to their net amount of migration.

This information will be stored in a Database (DB) created by our engineering team, which will contain tabular data related to the distinct features involved in this study, and their relation with the different countries along time. The mentioned DB will have the ability to automatically update information[2], in such a way that the project structure may be

[1] For more information, check the "*Disclaimer*" section

[2] For more information, check the *"Incremental load"* section

scalable, and that our DB can be aligned to the different variations that used sources may suffer in the future.

Additionally, we will carry out deep analytical tasks and apply Machine Learning (ML) Models trying to answer, among others, the following questions:

¿What makes an emigrant leave a country?
¿What are the predominant characteristics of the countries most chosen by immigrants?
¿What consequences does FM generate in both types of countries?
¿How can we measure that effect?
¿What are the characteristic features of countries with net positive and negative FM?
¿What makes a country have net FM close to zero?

A technological interface will be developed. This will permit access not only to the database, but also to metrics, statistics and specific indicators that we will chase according to our investigation and data comprehension.

As a result of all this work, we will generate a detailed report of the discoveries that our team of analysts will make, with the respective conclusions and insights provided by the data.

1.4 REACH AND OUT OF REACH:

Initially, the scope of the project is extended, upon reaching the agreed deadline on 2/28, to deliver the following products to the client:

- Detailed report containing analysis, development and deliverables.
- Access link to github private repository with all files and documents related to the project
- KPIs (Key performance indicators) considered decisive by the analytics department, with their explanation and information on use and interpretation.

- ❖ Interactive dashboard filled with schemes and filters that facilitate the comprehension of data through different metrics, statistics and KPIs.
- ❖ Trained ML learning model, with access to the code and step-by-step instructions for the process.
- ❖ Report meeting with the entire ***Analytic Hound ®*** team addressing the final deliverables. We will provide a step by step presentation of the entire process guided by our analysts.

Prior agreement with the client, development and management of this project does not include the following:

- ❖ The rigorous forecast of future FM flows. Under no circumstances the ***Analytic Hound ®*** team pretends to generate predictions that may be used for decision making.
- ❖ Modifications in the requirements made beyond the second week of work.
- ❖ Provide information to refugees, asylum seekers and those displaced by political and armed conflicts regarding courses of action in their situation. The ***Analytic Hound ®*** team takes advantage of this instance to express its support for human rights and the need to extend them to all people regardless of their ideals, religions or orientations of any kind.

1.5 KPIs (KEY PERFORMANCE INDICATORS)

Key performance indicators, commonly known as **KPIs**, are (as their name supposes) key performance indicators used to assess the success of actions and/or processes carried out in order to obtain certain objectives. In this way it is possible to determine if the actions taken gave the expected results or if instead it is necessary to make corrections in the approach.

The importance of KPIs lies in their key characteristic of allowing real-time measurement of the operation of business, marketing or sales strategies, providing valuable information for strategic decision making.

KPIs also play an important communicative role as they inform managers, employees and investors about the evolution of the company with respect to the established objectives, so that everyone can work with a general vision and goal in common.

From the in-depth exploration of different variables and data obtained during the first stage of the project, the analysts team proposes the formulation of the following KPIs described here.

The methodology used for determining the KPIs follows the concept of the recognized ***SMART*** objectives. The acronym stands for "Specific", "Measurable", "Attainable", "Relevant" and "Time-bound".

In the next paragraphs we will dedicate ourselves to the explanation and analysis of the selected KPIs. Also, we provide the design of each KPI in python code language.

KPI 1

Expresses the incidence of the migratory flow over the total population of the chosen country.

E.g. A value of 0.2 implies that migratory flow contributed 0.2% to increase the population. Instead, negative numbers such as -0.6 implies that migratory flow contributed to a decrease of 0.6% from the total population.

```python
df['KPI_1'] = round((df['Net Number of Migrants (thousands)']*1000)/(df['Population Total']) * 100 , 2)
```

KPI 2

Represents the variation in the net flow of migrants respect the previous year.

- **Positive** values: more people immigrated or fewer emigrated.
- **Negative** values: more people emigrated or less immigrated.

- Values of **zero:** there were no changes in the annual flow, (E.g. Net Migration was the same as the past year.

```python
# Group the data by country
grouped = df.groupby('Country Name')

# Calculate the difference in 'Net Number of Migrants (thousands)' between consecutive years for each country
diff = grouped['Net Number of Migrants (thousands)'].diff()

# Add the difference as a new column to the original DataFrame
df['KPI_2'] = round(diff,2)

# Replace the NaN values with 0 (since the first row for each country will have a NaN value)
df['KPI_2'].fillna(0, inplace=True)
```

Python

KPI 3

Expresses the percentage variation in migrants net flow respect to previous year.

E.g. ***Positive*** *value of 20.36 implies that net migration increased 20.36% annually.*

```python
# Create a new column 'KPI_3'
df['KPI_3'] = round((np.sign(df['KPI_2']) * abs(df['KPI_2'] / df['Net Number of Migrants (thousands)'].shift()) * 100) , 2)
# Replace the NaN values with 0 (since the first row for each country will have a NaN value)
df['KPI_3'].fillna(0, inplace=True)
```

Python

The following KPIs take for their measurement some specific features, which arise as a result of the implementation of the ML models.[3] For the construction of KPI 4 & 5, the "m = 3" most important features will be used. In this way, authorities and policy makers will know concisely which variables to direct their actions in order to maximize the result on the target variable.

Some of the features in our regression have negative coefficients, meaning an inverse relationship with the target variable. It is so that as the values of these features increase, the target variable is expected to decrease, and vice versa. This situation has been taken into account, and it can be handled in very different ways. For the purpose of dealing with this issue, we have chosen a criteria that provides the client a simple and intuitive final interpretation of the KPI, without explaining in detail certain technicisms.

[3] For more information, check the 3.3 *"Feature selection"* section

There are many considerations regarding the specificity of developing a KPI and much more specific metrics can be thought of based on what the future needs of the client are. This KPI provides a global and general approach, whose intention is to show the client that information can be managed and processed in different ways.

These KPIs are precisely adapted to the characteristics of each group, and their utility and functionality is maximized by following the guidelines below.

KPI 4 will be more useful for countries that have similar characteristics to the countries of Group 1 (large and positive Net Migration Flow)

KPI 5 is most useful for group 5 countries (negative influx of migrants)

Later we will explain the method used so that each country has a point estimate of which group it belongs to.

KPI 4:

This index is a grouping of the characteristics considered most relevant according to our ML model to explain the reception of migrants in the countries and years that receive the largest amount of net migration.

This KPI is the weighted average of the "m" indicators chosen, but normalized. We do this transformation in order to avoid magnitude bias, improve interpretation of the KPI, improve the model performance and decrease the impact of outliers in the data.

This KPI provides an estimate of the measure of the attractiveness of the country to receive migrants. Higher levels of KPI_4 are related to greater positive migratory flows.

According to our ML model, the top features for Group 1 are:

- Exports of goods and services (PCT of GDP)
- Gross Domestic Product
- Refugee population by country or territory of asylum
- Refugee population by country or territory of origin
- Infant Mortality Rate (infant deaths per 1,000 live births)

```python
# List of the selected features
SK_4 = ['Exports of goods and services (PCT of GDP)',
 'Gross Domestic Product',
 'Refugee population by country or territory of asylum',
 'Refugee population by country or territory of origin',
 'Infant Mortality Rate (infant deaths per 1,000 live births)']

# Standardize the columns
df[SK_4] = (df[SK_4] - df[SK_4].mean()) / df[SK_4].std()

# Assign a negative sign for the features with negative coefficient
df['Exports of goods and services (PCT of GDP)'] = -df['Exports of goods and services (PCT of GDP)']
df['Infant Mortality Rate (infant deaths per 1,000 live births)'] = -df['Infant Mortality Rate (infant deaths per 1,000 live births)']


# Create the KPI column
df['KPI_4'] = df[SK_4].sum(axis=1)/5
```
Python

Beyond the usefulness of KPI_4, the performance and effectiveness of the policies implemented in a given country can also be measured by evaluating the performance of the indicator in several periods.

In this way, VAR_KPI_4 shows us the percentage variation of KPI_4 between the period n and period n-1. Provides an effective estimate of the variation in the degree of desirability of the country to receive FM between the current period and the next.

In practical terms, it is a quick and easy way to understand indicators that allows us to assess at a glance the degree of effectiveness of the implemented policies.

E.g. VAR_KPI_4 = 0.12 implies that the objective to be achieved is that there is a percentage increase

12% per year of KPI_4 from one year to the next.

```python
df['VAR_KPI_4'] = round(df['KPI_4'].pct_change()*100,2)

# Replace the NaN values with 0 (since the first row for each country will have a NaN value)
df['VAR_KPI_4'].fillna(0, inplace=True)
```
Python

KPI 5:

The logic is similar to that of KPI_4, except that this basket measures the features that most explain migratory flows in countries where net migration is negative.

As before, an increasing KPI over time is desirable because it indicates that the adverse conditions are less and net migration is increasing (or has a less negative number).

According to our ML model, the top features for Group 5 are:

- ❖ Imports of goods and services (PCT of GDP)
- ❖ Population Density
- ❖ Refugee population by country or territory of asylum
- ❖ Labour force, total
- ❖ Infant Deaths, under age 1 (thousands)

```python
# List of the selected features
SK_5 = ['Imports of goods and services (PCT of GDP)',
 'Population Density',
 'Refugee population by country or territory of asylum',
 'Labour force, total',
 'Infant Deaths, under age 1 (thousands)']

# Standardize the columns
df[SK_5] = (df[SK_5] - df[SK_5].mean()) / df[SK_5].std()

# Assign a negative sign for the features with negative coefficient
df['Imports of goods and services (PCT of GDP)'] = -df['Imports of goods and services (PCT of GDP)']

# Create the KPI column
df['KPI_5'] = -df[SK_5].sum(axis=1)/5
```

Python

As we saw in the previous section, in this case we present VAR_KPI_5, which shows us the percentage variation of KPI_5 between the period n and period n-1.

Example: VAR_KPI_5 = 0.08 implies that the objective to be achieved is that there is a percentage increase of 8% year-on-year of the KPI_5. The client can compare the value of the KPI that he wants to achieve with the data provided by the table.

```python
df['VAR_KPI_5'] = round(df['KPI_5'].pct_change()*100,2)

# Replace the NaN values with 0 (since the first row for each country will have a NaN value)
df['VAR_KPI_5'].fillna(0, inplace=True)
```

Python

1.6 TECHNOLOGY STACK

Python: A high-level interpreted programming language whose philosophy emphasizes the readability of its code, it is used to develop applications of all kinds. Multiparadigm programming language. It partially supports object orientation, imperative programming and, to a lesser extent, functional programming. It is an interpreted, dynamic and cross-platform language. Managed by the Python Software Foundation, it is licensed under an open source license, called the Python Software Foundation License. It consistently ranks as one of the most popular programming languages.

Visual studio code (VSC): Source code editor developed by Microsoft for Windows, Linux, macOS and Web. It includes support for debugging, integrated Git control, syntax highlighting, smart code completion, among other features suitable for building this project. Libraries:

- ***Pandas***: A library written for the python programming language designed to allow data manipulation and analysis. It offers data structures and operations for manipulating number tables and time series.
- ***WBGAPI*** 1.0.12 and ***DATAPACKAGE*** 1.15.2 libraries are described "Incremental data loading" section

Amazon Web Services (AWS): A collection of public cloud computing services (also called web services) that together form a cloud computing platform, delivered over the Internet by Amazon.com.

Microsoft Power BI: Interactive data visualization software developed by Microsoft with a primary focus on business intelligence.

Streamlit: Open source Python library that makes it easy to build custom web applications for machine learning and data science.

2. DATA PROCESSING

2.1. SOURCES

During the exploration for different data sources that would contribute to the elaboration of this project, a detailed examination of sources proposed by the client was carried out in a preliminary way. Considering the finding of higher quality data, variety and temporal frequency compared to those indicated by the client, the ***Analytic Hound ®*** team opted for the use of their own sources. They are listed below with their respective links that lead to the main reference page:

- ***World Bank:*** https://data.worldbank.org/
- ***United Nations:*** https://www.un.org/development/desa/pd/data-landing-page
- ***Accesorio de data de UN:*** https://data.un.org/

2.2. INCREMENTAL LOAD

The incremental data load application refers to updating the data as the main sources are modified over time.

In the context of the migratory analysis of the project in question, said update is expected to occur with an average periodicity of the semi- annual/annual type.

For the correct implementation of the incremental load, the official documentation for developers within the selected databases was referred. With their support, the pipeline corresponding to said task is established.

As can be seen in the attached links, the following python libraries were used for the direct extraction of data from the different sources:

- **WBGAPI 1.0.12** | https://pypi.org/project/wbgapi/
- **DATAPACKAGE 1.15.2** [4] | https://pypi.org/project/datapackage/

In short, the code used for this task has the following order of execution:

1. Library import
2. Data extraction directly through python libraries.
3. Saving them in defined variables that are later raw material for the ETL process that will happen immediately after the load is finished.

2.3. REPORT & PIPELINE | COMPLETE ETL

The analysis and report corresponding to data extraction, transformation and load process (ETL) with its respective pipeline can be consulted directly in the file ***"ETL.ipynb"***, hosted on our official github repository. Access to it is facilitated through the following public link: ***https://github.com/Analytic-Hound-Consulting/ONG-Henry***

DATA ORIGIN

Data used in this project comes from multiple CSV files containing information on various features from all countries and regions. These features are studied according to their influence on the positive and negative migration of each country. The data is extracted from world websites, which make such information available.

TRANSFORMATION PROCESSES

Once the data from the CSV file is imported, several transformations are performed to prepare it for use. These included:

Rename columns for clarity.
Fill in the missing data judiciously.

[4] Note: datahub operates within a type of data "*package*" allowing its management with the aforementioned library.

Review and remove duplicate rows.
Remove unnecessary columns.
Eliminate data corresponding to years prior to 2000.
Rearrangement of columns for uniformity.

To carry out these transformations, two pipelines were designed that group the various transformations that must be applied to the different datasets. The language used is Python, and the main libraries were *Pandas, wbgapi, datapackage*, and *scikit learn*.

FINAL DESTINATION

Transformed data is exported to a CSV file for further use and evaluation for the learning models. The entire pipeline was handed over to the engineering team in order to lift it up in the cloud.

CONCLUSIONS

The ETL process was successful, both in the collection and in the transformation of data from different sources. The transformations performed allow the data to be effectively analyzed to obtain valuable information about the effect that different variables have on migration.

2.4. FEATURES DICTIONARY

CATEGORY	FEATURE	DEFINITION	MEASURE
MACROECONOMY	**Gross Domestic Product (GDP)**	Total gross value added by all resident producers in the economy of each country in U$D.	(U$D)
	GDP growth:	Annual percentage growth rate of GDP.	(annualy %)
	Final consumption expenditure:	SUM of household final consumption expenditure (private consumption) and general government final consumption expenditure.	(% of GDP)
	GDP per capita:	Total gross value added by one producer in the economy	(US$)
	GNI per capita (atlas method):	Gross national income (converted to dollars using the world bank atlas method) divided by the midyear population.	(US$)
	Gross savings:	Calculated as gross national income - total consumption.	(% of GDP)
	Consumer price index:	Cost of the average consumer of acquiring a basket of goods and services that may be fixed or changed at specified intervals, such as yearly.	(2010 = 100)
	Unemployment, total:	Share of the labor force that is without work but available for and seeking employment	(% of total labor force)
	Labour force	People ages 15 and older who supply labor for the production of goods and services during a specified period. It includes people who are currently employed and people who are unemployed but seeking work as well as first-time job-seekers	(Total)

	Total reserves	Comprise holdings of monetary gold, special drawing rights, reserves of IMF members held by the IMF, and holdings of foreign exchange under the control of monetary authorities	(gold + US$)
	Exports of goods and services	Value of all goods and other market services provided to the rest of the world. (Exclude: compensation of employees and investment income and transfer payments)	(PCT of GDP)
	Imports of goods and services	Value of all goods and other market services received from the rest of the world.	(PCT of GDP)
MICROECONOMY	**Foreign direct investment, net inflows**	Net inflows of investment to acquire a lasting management interest (10% or more of voting stock) in an enterprise operating in an economy other than that of the investor. It is the sum of equity capital, reinvestment of earnings, other long-term capital, and short-term capital as shown in the balance of payments. This series shows net inflows (new investment inflows less disinvestment) in the reporting economy from foreign investors, and is divided by GDP	(PCT of GDP)
	Total tax and contribution rate	Amount of taxes and mandatory contributions and exemptions as a share of commercial profits. Taxes withheld (such as personal income tax) or collected and remitted to tax authorities (such as value added taxes, sales taxes or goods and service taxes) are excluded	(PCT of profit)
	Time required to start a business	Number of calendar days needed to complete the procedures to legally operate a business (if a procedure can be speeded up at additional cost, the fastest procedure is chosen)	(days)

	Research and development expenditure	Gross domestic expenditure on research and development. Include capital and current expenditure in 4 main sectors: business enterprise, government, higher education and private non-profit. Covers basic research, applied research, and experimental development	(PCT of GDP)
SOCIAL DEVELOPMENT	**Government expenditure on education:**	General government expenditure on education expressed as a percentage of GDP.	total (% of GDP)
	Primary completion rate, total	Number of new entrants (enrollments minus repeaters) in the last grade of primary education, regardless of age, divided by the population at the entrance age for the last grade of primary education	(PCT of relevant age group)
	Intentional homicides	Unlawful homicides purposely inflicted as a result of domestic disputes, interpersonal violence, violent conflicts over land resources, intergang violence over turf or control, and predatory violence and killing by armed groups.	(per 100,000 people)
	Access to electricity:	Percentage of population with access to electricity. Electrificacion data is collected from industry, national surveys, and international sources.	(% of population)
	People using at least basic sanitation services in urban areas:	Improved sanitation facilities that are not shared with other households. This indicator encompasses both people using basic sanitation services as well as those using safely managed sanitation services.	(PCT of population)
	Mobile cellular subscriptions:	Subscriptions to a public mobile telephone service that provide access to the PSTN (public switched telephone network) using cellular	(per 100 people)

		technology. The indicator includes (and is split into) the number of postpaid subscriptions, and the number of active prepaid accounts. The indicator applies to all mobile cellular subscriptions that offer voice communications.	
DEMOGRAPHY	**Population Total:**	Bases on the facto definition of population, which counts all residents regardless of legal status or citizenship.	(Total)
	Population Density :	Amount of people per square meter of land area	(people per sq. km of land area)
	Natural Change:	Births minus Deaths	(thousands)
	Rate of Natural Change	Birth rate minus the death rate of a particular population, over a particular time period	(per 1,000 population)
	Population Growth Rate	Exponential rate of growth of midyear population from year t-1 to t. (see “population total” definition)	(Annual percentage)
HEALTH	**Crude Birth Rate**	Number of live births occurring during the year.	(births per 1,000 population)
	Median Age, as of 1 July	Age that divides the population in two parts of equal size, that is, there are as many persons with ages above the median as there are with ages below the median.	(years)
	Life Expectancy at Birth, both sexes	Number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth where to stay the same throughout its life	(years)
	Infant Mortality Rate	Number of infant deaths for every 1,000 live births.	(infant deaths per 1,000 live births)

	Infant Deaths, under age 1	Death of an infant before his or her first birthday	(thousands)
	Maternal mortality ratio:	Number of women who die from pregnancy-related causes while pregnant or within 42 days of pregnancy termination. The data are estimated with a regression model using information on the proportion of maternal deaths among non-AIDS deaths in women ages 15-49, fertility, birth attendants, and GDP.	(per 100,000 live births.)
MIGRATION	**Net Number of Migrants**	Net total of migrants during the period, that is, the number of immigrants minus the number of emigrants, including both citizens and noncitizens	(thousands)
	Net Migration Rate	Difference between the number of immigrants and the number of emigrants (people leaving an area) throughout the year	(per 1,000 population)
	Refugee population by country or territory of origin	Refugees are people who are recognized as refugees under the 1951 Convention Relating to the Status of Refugees or its 1967 Protocol, the 1969 Organization of African Unity Convention Governing the Specific Aspects of Refugee Problems in Africa, people recognized as refugees in accordance with the UNHCR statute, people granted refugee-like humanitarian status, and people provided temporary protection. Asylum seekers--people who have applied for asylum or refugee status and who have not yet received a decision or who are registered as asylum seekers--are excluded. Palestinian refugees are people (and their descendants) whose residence was Palestine between June	(Annual, SUM)

		1946 and May 1948 and who lost their homes and means of livelihood as a result of the 1948 Arab-Israeli conflict. Country of origin generally refers to the nationality or country of citizenship of a claimant.	
	Refugee population by country or territory of asylum	Refugees are people who are recognized as refugees under the 1951 Convention Relating to the Status of Refugees or its 1967 Protocol, the 1969 Organization of African Unity Convention Governing the Specific Aspects of Refugee Problems in Africa, people recognized as refugees in accordance with the UNHCR statute, people granted refugee-like humanitarian status, and people provided temporary protection. Asylum seekers -- people who have applied for asylum or refugee status and who have not yet received a decision or who are registered as asylum seekers--are excluded. Palestinian refugees are people (and their descendants) whose residence was Palestine between June 1946 and May 1948 and who lost their homes and means of livelihood as a result of the 1948 Arab-Israeli conflict. Country of asylum is the country where an asylum claim was filed and granted.	(Annual, SUM)

2.5. DATA STRUCTURE (DW, DL)

In line with the client's requirements, and opting for the most effective alternative, the engineering department decided to set up a ***Data Lake*** (DL) and ***Data Warehouse*** (DW) in the Cloud services of*Amazon Web Services*.

The main reasons behind this choice are related to the fact that both services DL ***(S3)*** and DW ***(EC2)*** are offered on the same AWS platform

First of all, it is essential to understand that both concepts (*DL and DW*) refer to tools used to store large amounts of data in an organized manner. However, there are some key differences between them.

A ***Data Lake*** is a data repository used to store large amounts of information, ***regardless of its format or structure***. In a Data Lake, structured, semi-structured, and unstructured data can be stored, allowing analysts to explore and analyze the data in its original form.

On the other hand, a ***Data Warehouse*** is a data repository used to store ***structured and preprocessed data***, with the aim of facilitating its analysis and consultation. The data stored in a Data Warehouse has been previously organized and structured so that it can be used more efficiently.

It is important to take into account that both tools can be used ***complementary***. That is, a Data Lake can be used to store the data in its original form and then process and transform it to load it into a Data Warehouse, where it can be analyzed more efficiently. This last format of use in a complementary way responds to how they were used for the development of this particular project.

Some of the advantages of using ***Data Lake*** are:

- *Data storage in its original form*: Allows you to store large amounts of data without having to structure it beforehand, which can be useful for exploratory analysis and data mining.
- *Scalability*: Data Lakes are highly scalable and can easily grow as more data is added.
- *Flexibility*: Allows you to store data of different formats and structures, which makes it ideal for storing unstructured or semi-structured data.

On the other hand, some of the advantages of using a ***Data Warehouse*** are:

- *Ease of analysis*: The data is pre-structured so that it can be analyzed and consulted more efficiently.

- *Integration with analysis tools*: Data warehouses are often integrated with data analysis tools, allowing users to analyze data more efficiently.
- *Greater security*: The data in a Data Warehouse is usually more secure than in a Data Lake, since measures have been taken to guarantee the integrity and confidentiality of the data.

In short, both ***DL*** and ***DW*** are important tools for data storage and analysis. The choice of which and how to use them will depend on the specific needs of each particular project and the objectives of the data analysis.

Regarding ***Amazon Elastic Compute Cloud (EC2)***, it is an AWS service that provides scalable computing capacity in the cloud. It allows users to launch and manage virtual server instances in the cloud, allowing complete control over their computing environment.

Some of the features and benefits of Amazon ***EC2*** are:

- *Scalability*: One of the main advantages of EC2 is that it is highly scalable. Users can increase or decrease compute capacity based on their needs at any time, allowing them to adjust their compute capacity to meet the demand of their application or service.
- *Diversity of instances*: Amazon EC2 offers a wide variety of instance types, from General Instances to Specialized Instances, each designed to meet specific needs. This allows users to choose the instance that best suits their computation needs.
- *Flexibility*: Users have full control over their computing environment. They can select the operating system, network configuration, and the amount of storage capacity they need for their instance.
- *"Pay as you go"*: EC2 uses a pay-as-you-go model, which means that users only pay for the compute capacity they use. This allows them to reduce their costs and optimize their IT budget.
- *Integration with other AWS services*: Amazon EC2 integrates with other AWS services, such as Amazon Simple Storage Service (S3), Amazon Relational Database

Service (RDS), and Amazon Elastic Load Balancing, allowing users to build more comprehensive and complex IT solutions.

- *Security*: Amazon EC2 integrates with other AWS security tools, such as AWS Identity and Access Management (IAM) and Amazon Virtual Private Cloud (VPC), helping users protect their computing environment and data.

In short, Amazon *EC2* is a powerful tool that allows users to launch and manage virtual servers in the cloud in a scalable and flexible way. Its pay-as-you-go model, with the wide variety of instance types, and its integration with other AWS services make it an ideal solution for this project due to its efficient and scalable compute power. In this case, said service will be used for hosting ***streamlit***.

Finally, ***Amazon Simple Storage Service (S3)*** is an AWS cloud object storage service. *S3* provides a highly scalable, secure, and durable storage solution that enables users to store and retrieve large amounts of data from anywhere in the world.

Some of the features and benefits of Amazon*S3* are:

- *Scalability*: Amazon S3 is highly scalable and allows users to store and retrieve large amounts of data in the cloud without worrying about storage capacity. Users can scale their storage capacity as needed and only pay for the capacity they use.
- *Durability and availability*: S3 is highly durable and designed to provide 99.99% availability. This means that users can be sure that their data will be available at all times and will be protected against data loss.
- *Security*: Amazon S3 uses SSL/TLS encryption to protect data in transit and AES-256 encryption to protect data at rest. Users can configure access policies and role-based access control to protect their data and keep it safe.
- *Flexibility*: S3 supports a wide variety of file types and formats, making it ideal for a variety of use cases, from media file storage to database backup and application logs.

- *Integration*: Amazon S3 integrates with other AWS services such as Amazon EC2, Amazon Lambda, and Amazon CloudFront, allowing users to build more comprehensive and complex IT solutions.
- *Analytics*: Amazon S3 offers a variety of analytics tools that allow users to better understand their data and gain valuable insights. For example, S3 can be used with Amazon Athena to perform ad-hoc queries on data stored in S3.

In short, Amazon *S3* is a highly scalable, secure, and durable object storage service that provides users the flexibility to store and retrieve large amounts of data in the cloud. Its durability, availability, and security features, its integration with other AWS services, and its analysis tools make it an ideal solution for enterprises and developers looking for a secure and scalable way to store and analyze large amounts of data in the cloud.

2.6 IMPLEMENTATION

The procedure established for the structuring of the cloud service was as follows.

First of all, it was decided to create an instance in *EC2* and perform a manual file upload from Visual Studio Code. This was aimed to achieve a better understanding of the behavior of said instance. Subsequently, a deposit was created in *S3* to store the information and then perform a manual test of uploading files from AWS.

Since mobilization of information from one to the other was necessary, a role was provided for the connection between both *(EC2 y S3).*It is important to clarify that it does not prevent individual access to any of them, nor does it prevent the fact that one can delete data without affecting the other.

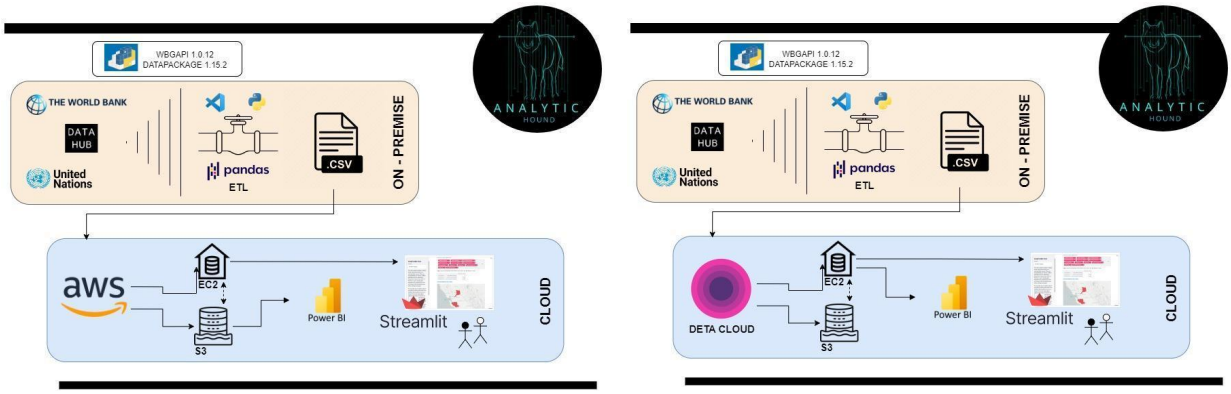
Initially, the original idea was to automate the link between GitHub and the repository. This was abandoned in favor of a faster and easier option. Modifications were made on the ETL *pipeline* so that, once finished, it directly executes the sending of the final CSV file (with all the data processed) to the deposit, to later carry out its automatic upload to the data warehouse.

Finally, the implementation of ***CRON*** allows the execution of a command responsible for updating the data file on which the analysis of this project is carried out. In this way we will have the entire automated process running itself once a month.

2.6. DW AUTOMATION SOLUTION

In the following diagram we detail the workflow that satisfies the requirement made by the product owner at the end of *Sprint #1*. Said modification of the deliverables consisted of considering an alternative plan for compliance with the requirements in the event that for some eventuality the primary option could not be implemented.

In this alternative we decided to use the ***DETA cloud service*** for data storage and for deploying streamlit. This version works as a PLAN B which has already been tested and implemented.



3. MACHINE LEARNING

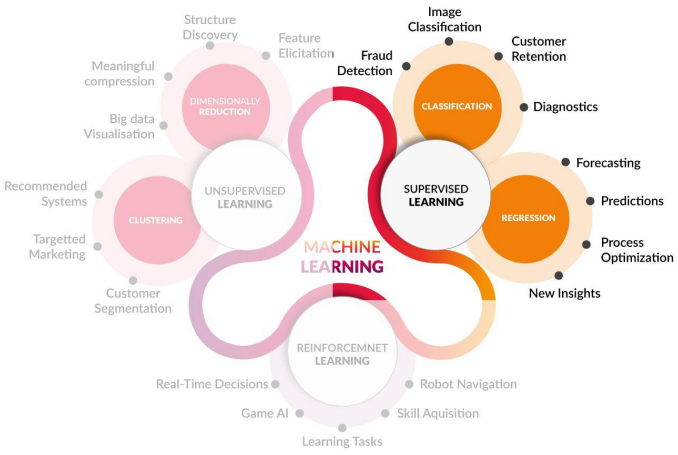
3.1. MACHINE LEARNING - REGRESIÓN LINEAL

Starting from the technical definition, the concept of Machine Learning (ML) focuses on the use and development of algorithms that use data to mimic the way humans learn, gradually improving their accuracy. Considering that for the purposes of this document, the development of this subject would be excessive or insufficient, a complementary reading material is suggested for those wishing to know more about this wonderful world.[5]

In a schematic way and without intending to delve into what the subject of machine learning covers, as well as its applications by which it has managed to revolutionize the way of developing technology in recent years, a brief introduction to the model used for the project "***Worldwide migration patterns analysis***" developed by ***Analytic Hound ®*** is presented here.

According to the ***label*** type or ***output variable*** (commonly called "***Y***"), ***supervised*** learning models can be classified into:

- **Classification:** in these models the label is a type of categories (eg sick/healthy, cat/dog/bird, spam/not spam)
- **Regression:** The output variable is a numeric value. (eg price, quantity, temperature)



[5] Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems - https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/

In summary, supervised learning models use the different variables or ***features*** (only tolerated in numerical format) to determine the patterns (***algorithms***) that allow predicting a selected output variable or label.

In general, it is considered a good practice to allocate a considerable percentage of data (≈80%) to model training and then use the rest for model testing. All this happens under the premise of "More training data means better prediction".

Finally, once the model has been developed, it can be evaluated using different error quantification formulas. Among them are the ***MAE*** (mean absolute error), ***MSE*** (mean square error), ***RMSE*** (square root mean square error)) and ***R square*** (coefficient of determination) among others.

Having briefly commented on the details that make up the supervised ML models, in the following section we will limit ourselves to exposing the model developed for this specific project.

3.2. WORK HYPOTHESIS

The interest in the use of automated learning tools is oriented towards the discovery of those features that are most important when it comes to explaining migratory phenomena. This happens in order to provide different KPIs to governments and third party individuals that may be interested, in such a way that they are used as instruments for monitoring and measuring the migratory situation within their borders. Likewise, the KPIs will turn out to be powerful tools for the development of government policies and decision-making on the matter.

The ***working hypothesis*** is based on the assumption that those characteristics that best explain net migratory flows may differ depending on whether it is a country with positive or negative net migration. In order to support this hypothesis, we will run a ML model using ***SelectKBest*** features for different registers depending on the year and country to which the migratory flow corresponds.

A series of tests were carried out to find out the relationships between our variables and compare the results.

First, *SciKitLearn's* ***SelectKBest*** module will perform a selection of the top ***"k"*** features with the highest scores obtained from a scoring function called "***f_regression***". Said function will calculate the existing correlation between each feature and the target variable, returning the regression coefficient and its ***p-value*** for each feature. For our approach, we will consider the target variable as the '***Net number of migrants (thousands)***'.

SelectKBest will then select the top *k* features with the highest *F-values*, indicating the strongest linear relationship linked to the target variable. In summary, the selected characteristics will then be those with the greatest influence when it comes to explaining the migratory dynamics.

In order to achieve a better understanding of the potential effect of each feature on the phenomenon of net migration, the coefficients of the linear regression model are used to account for the direction as well as the impact of each characteristic in relation to the target.

This methodology is applied to all countries, to obtain an overview of this idea.

In the first instance, the 5 best characteristics are selected to explain migratory flows worldwide:

- Population Total
- Natural Change, Births minus Deaths (thousands)
- Population Growth Rate (percentage)
- Net Migration Rate (per 1,000 population)
- Infant Deaths, under age 1 (thousands)

These 5 variables and their respective coefficients imply that, considering all countries, Net Migration tends to be globally explained by them.

Observing the features listed here, it may seem intuitive at first glance to consider them crucial for migratory flows. We will delve into the more detailed analysis once the complete data is obtained to compare, infer patterns and draw conclusions.

A *p-value* is a statistical measure of the evidence against the null hypothesis that the coefficient for a given feature is zero. As a practical matter, a *p-value* less or equal to the 5% significance level suggests that the coefficient for that characteristic is statistically significant at a 95% confidence level.

This translates into strong evidence linked to the relationship between each of the variables and the "Net Number of Migrants (thousands)" in the specific population from which the used sample was drawn. This means that there is less than a 5% chance that the observed relationship between the predictor and the response variable is due to chance.

Based on the information obtained from the tests performed, the *p-values* would appear to be extremely low (much less than 0.05), indicating that the corresponding coefficients are statistically significant with a confidence level of 95%. Therefore, it is possible to affirm with a high degree of certainty that the characteristics have a significant impact on the target variable.

3.3. FEATURE SELECTION

Multicollinearity refers to a situation in which two or more independent variables, in a regression model, show a high degree of correlation. As a result, the coefficients of the model variables can become unstable and difficult to interpret, and the overall performance of the model can be affected. This effect can be a problem for several reasons which lie beyond the scope of this documentation.

One way to verify multicollinearity is based on the use of the ***variance inflation factor (VIF)*** for each characteristic. Calculating the *VIF* allows us to measure how much the variance of the estimated regression coefficient for a given characteristic increases due to multicollinearity with the other characteristics. A *VIF* value of 1 indicates no multicollinearity, while higher values relate to increasing levels of multicollinearity.

After the calculations, the following variables present a *VIF* greater than 5:

- Access Elect.
- Crude Birth Rate (births per 1,000 population)
- Natural Change, Births minus Deaths (thousands)
- Population Growth Rate (percentage)
- Population Total
- Rate of Natural Change (per 1,000 population)

Following the *VIF* concept, these features will be eliminated and not taken into account for the analysis developed here. Finally, the rest of the variables will be used to work with the ML model.

3.4. IMPLEMENTATION

To test the proposed hypothesis, we proceed to divide and classify the data based on net migratory flows.

All records are sorted by net migration in descending order and classified into 5 categories. Group 1 will include those with the highest positive Net Migration while group 5 will involve those with a higher negative value for the same target variable. In this way, a deeper understanding about migration in different scenarios and times is obtained.

It is noteworthy that in this experiment groups will be formed by different years and countries. This situation is due to the fact that we will not be studying the behavior of migratory flows for a particular country. Rather, we will try to understand migration as a social phenomenon inherent to human nature, regardless of pre-established political borders, which are subject to constant modification.

Once the data has been divided, those with values 1, 3 and 5 are grouped establishing the following distribution: records with high positive net migration (Group 1), records with net migration "with a tendency towards neutral values" (Group 3), and a last group with negative net migration (Group 5).

For each of these groups, a selection of the 5 features with the greatest impact is made and the corresponding coefficients are calculated. *P-values* and some relevant statistical data are verified.

For the purposes of this ML development, those non-numerical characteristics, the target 'Net number of migrants (thousands)' and 'Net migration rate (per 1,000 people)', are dispensed.

This last variable is highly correlated with the target, so we will remove it from the sample as we have strong reasons to believe it could bring to the model problems of endogeneity and autocorrelation.

Features not taken into account:

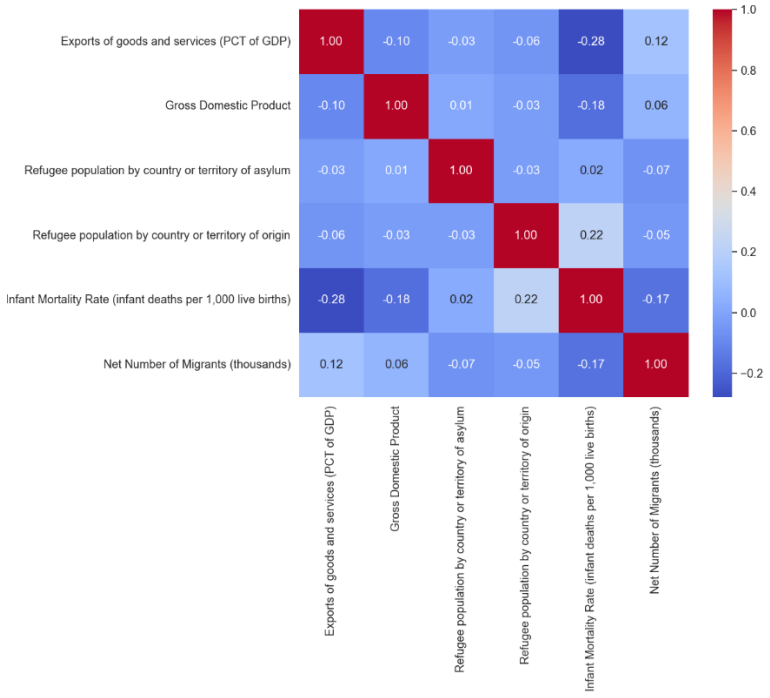
- Country Code
- Country Name
- Year
- Net Number of Migrants (thousands)
- Net Migration Rate (per 1,000 population)
- Access Elect.
- Crude Birth Rate (births per 1,000 population)
- Natural Change, Births minus Deaths (thousands)
- Population Growth Rate (percentage)
- Population Total
- Rate of Natural Change (per 1,000 population)

GROUP 1 (Positive net migration)

The 5 most decisive characteristics corresponding to this group are the following:

- ❖ Exports of goods and services (PCT of GDP)
- ❖ Gross Domestic Product
- ❖ Refugee population by country or territory of asylum
- ❖ Refugee population by country or territory of origin
- ❖ Infant Mortality Rate (infant deaths per 1,000 live births

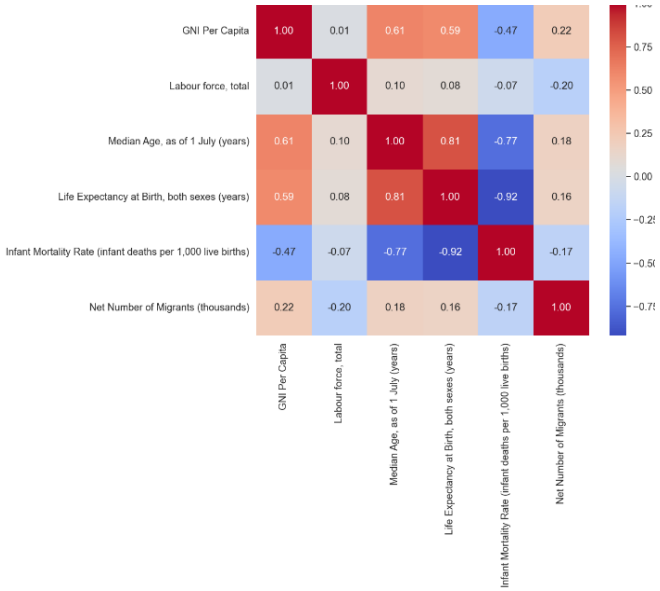
A correlation matrix of the features in question is attached



GROUP 3 (Net migration "with a tendency towards neutral values)

The 5 most decisive characteristics corresponding to this group are the following:

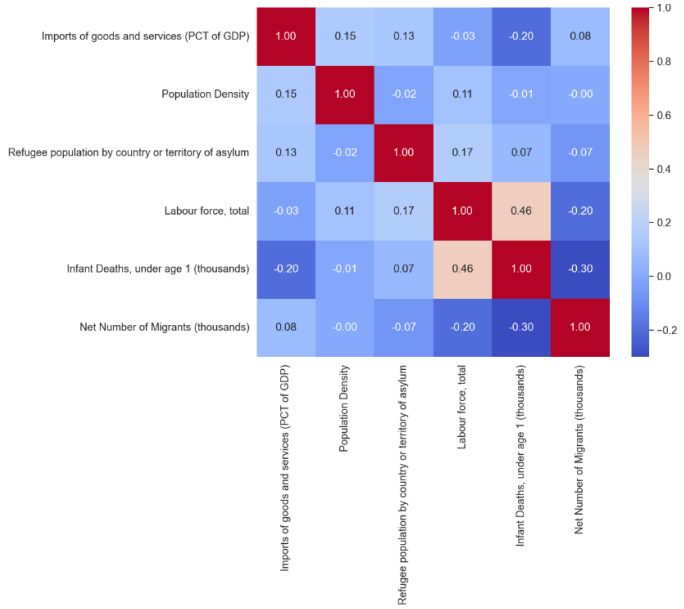
- ❖ GNI Per Capita
- ❖ Labour force, total
- ❖ Median Age, as of 1 July (years)
- ❖ Life Expectancy at Birth, both sexes (years)
- ❖ Infant Mortality Rate (infant deaths per 1,000 live births)



GROUP 5 (Negative net migration)

The 5 most decisive characteristics corresponding to this group are the following:

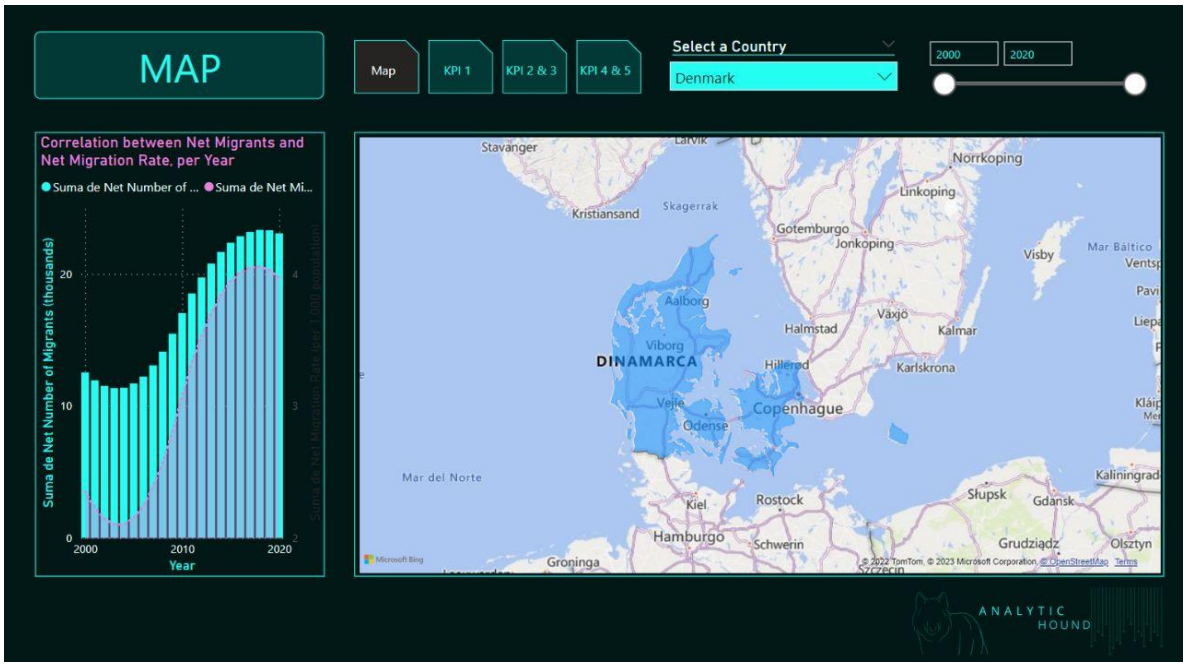
- ❖ Imports of goods and services (PCT of GDP)
- ❖ Population Density
- ❖ Refugee population by country or territory of asylum
- ❖ Labour force, total
- ❖ Infant Deaths, under age 1 (thousands)



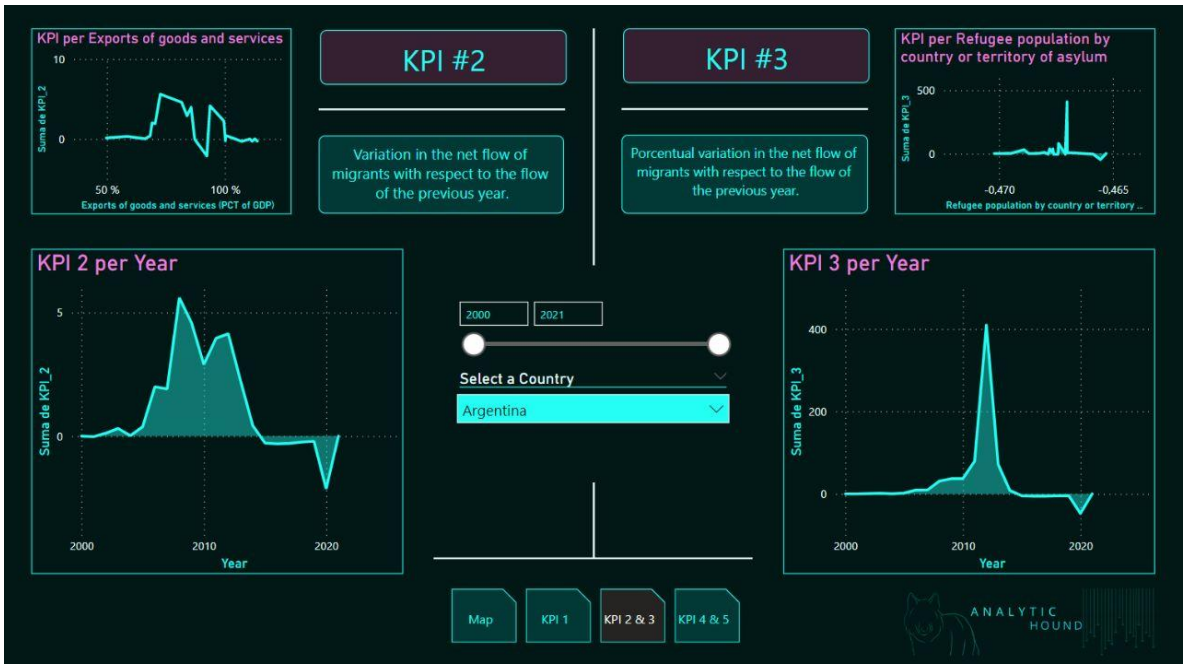
4. VISUALIZATION

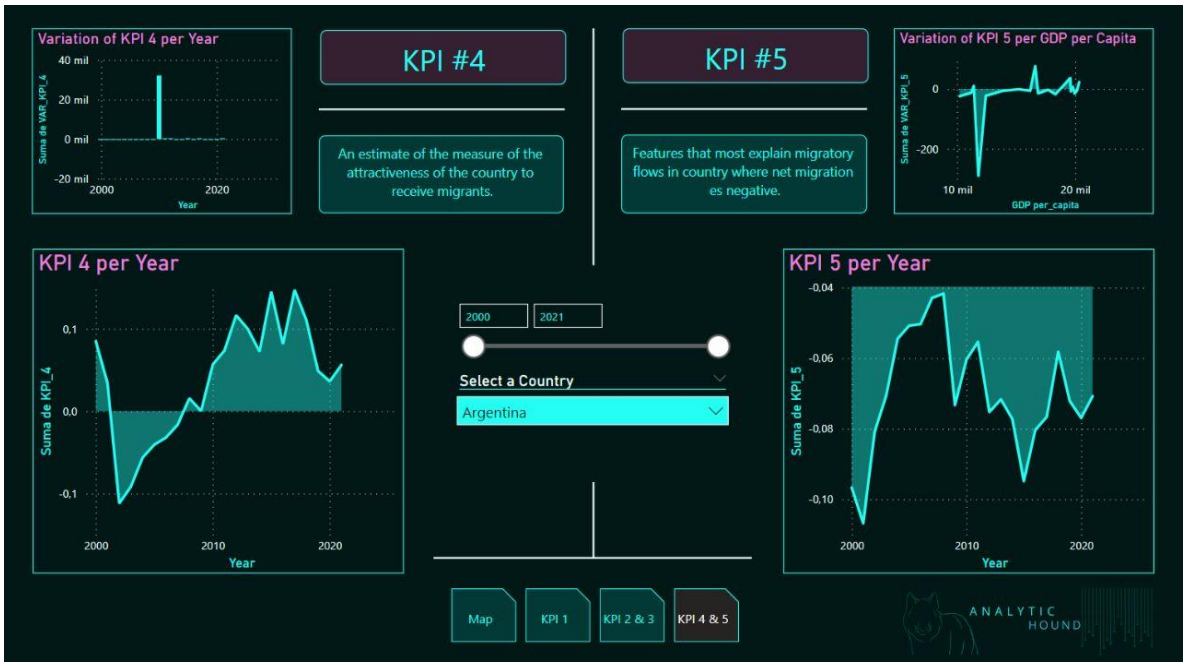
The use of interactive visualization tools allows the user to fully and easily navigate the different data used as well as their different relationships. The way in which these data are linked and the conclusions derived from them are the master key to our project. Below are schematically presented images corresponding to the visualizations developed by the ***Analytic Hound ®*** team of analysts.

4.1 POWER BI

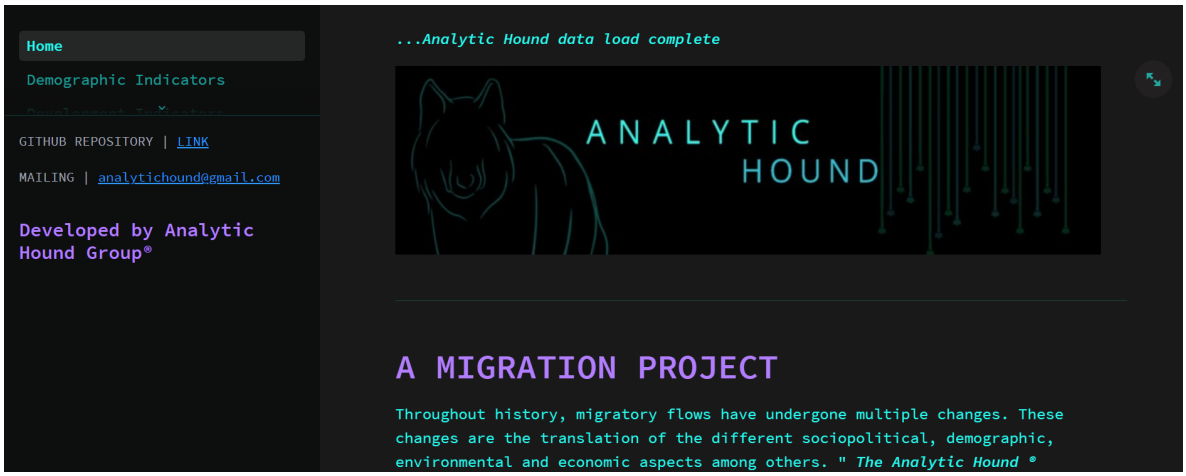


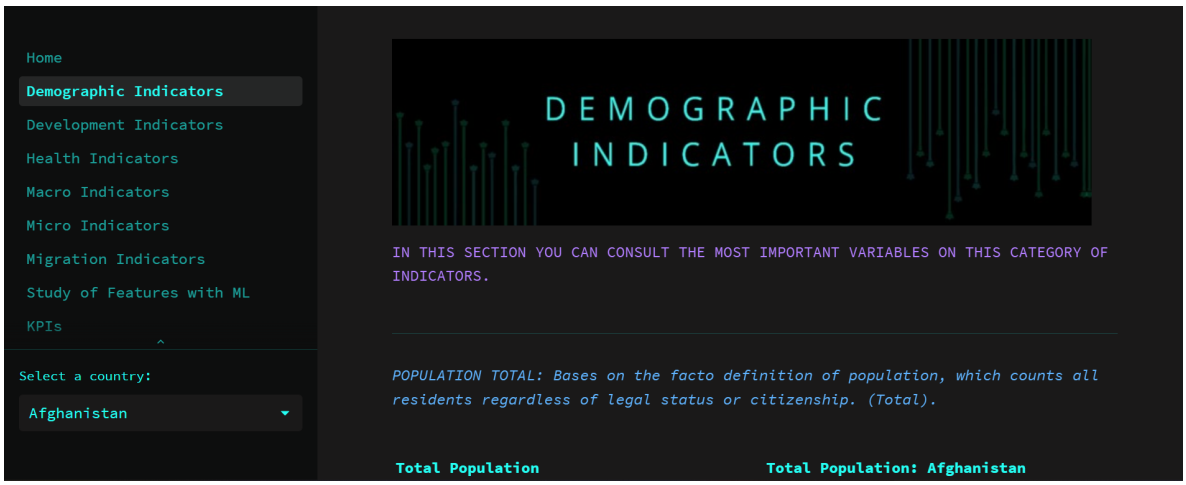

Net Number of Migrants vs Unemployment per Year
Suma de Net Number of Migra...
Suma de Unemploy...
Suma de Net Number of ...
Suma de Unemployment
Year
KPI #1
Incidence of the migratory flow over the total population of the chosen country.
Net Number of Migrants per GDP per Capita
Suma de Net Number of Migr...
GDP per_capita
Net Number of Migrants per Unemployment
Unemployment
2000
2021
Select a Country
Argentina
KPI per Year
Suma de KPI_1
Map
KPI 1
KPI 2 & 3
KPI 4 & 5
ANALYTIC HOUND

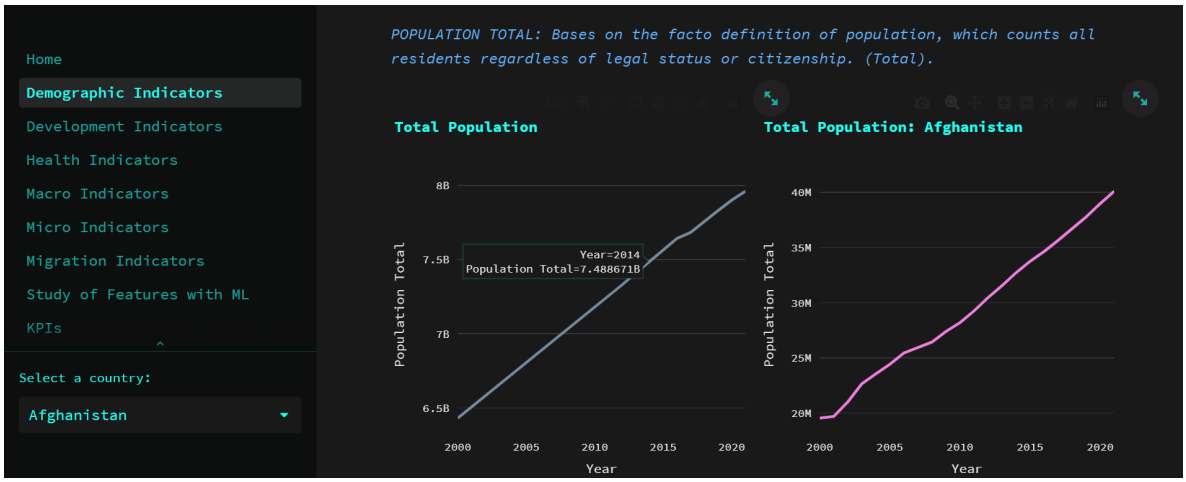

KPI per Exports of goods and services
Suma de KPI_2
Exports of goods and services (PCT of GDP)
KPI #2
Variation in the net flow of migrants with respect to the flow of the previous year.
KPI #3
Porcentual variation in the net flow of migrants with respect to the flow of the previous year.
KPI per Refugee population by country or territory of asylum
Suma de KPI_3
Refugee population by country or territory ...
KPI 2 per Year
Year
2000
2021
Select a Country
Argentina
KPI 3 per Year
Map
KPI 1
KPI 2 & 3
KPI 4 & 5
ANALYTIC HOUND


Variation of KPI 4 per Year
40 mil
20 mil
0 mil
-20 mil
Suma de VAR_KPI_4
2000
2020
Year
KPI #4
An estimate of the measure of the attractiveness of the country to receive migrants.
KPI #5
Features that most explain migratory flows in country where net migration es negative.
Variation of KPI 5 per GDP per Capita
0
-200
Suma de VAR_KPI_5
10 mil
20 mil
GDP per_capita
KPI 4 per Year
0,1
0,0
-0,1
Suma de KPI_4
2000
2010
2020
Year
2000
2021
Select a Country
Argentina
KPI 5 per Year
-0,04
-0,06
-0,08
-0,10
Suma de KPI_5
2000
2010
2020
Year
Map
KPI 1
KPI 2 & 3
KPI 4 & 5
ANALYTIC
HOUND

4.2 STREAMLIT (ACCESS LINK: http://107.21.7.155:8501/)







5. INSIGHTS & CONCLUSIONS

In the previous sections, the content of this document has been fully dedicated to manufacture a precise and detailed manual linked to the different processes and stages that encompass the whole data processing.

It is imperative at this point to clarify what the ***Analytic Hound ®*** team considers the *cornerstone* of this project; we refer here to the "***Transformation of data into information***"

We have carefully examined the process that data undergoes from its obtaining from the sources, through its entire transformation stage, to its visualization in the form of graphs and dashboards, as well as its use in the development of the different machine learning models

Intuitively, it is expected to question the precise moment in which the data leaves its place as “just data” to become valuable information. We refer to information or ***insight*** when, after implementing the different corresponding processes, the data becomes generators of new conclusions or ***knowledge***, allowing its use primarily for strategic decision-making.

Next, we will present the different ***insights*** obtained in relation to the object of study of this project.

Referring to the machine learning section, the result of the analysis shows the 5 most relevant features for each group of records when it comes to explaining net migration. For the development of this introductory paper, in order not to fall into the extension of this document, we will select only two of these groups: Group 1 and Group 5. These groups agglomerate records located at opposite poles with respect to the net migratory flow. Future appendices to this documentation will address the analysis in a more comprehensive and detailed manner.

In relation to the first ***group,*** represented by records with the highest value of ***positive net migration***, we can find ***export of goods and services, GDP and infant mortality*** as the

most outstanding characteristics in relation to the influence that these could exert on a positive migration.

Just as the GDP (gross domestic product) is a clear reflection of the real economic situation of a country, the quality of "exporter" contributed by the corresponding variable allows obtaining a global idea of the degree of industrialization of the country. This industrial development necessarily implies a labor correlation, since those regions with the greatest amount of exports require a greater amount of labor for the production of said goods or services. Based on the above, it can be deduced that the impulse oriented towards the development of the industry, translated into more jobs, a greater quantity of exports and improvements in the situation of the national GDP, turn out to be variables that directly and indirectly affect the dynamics of migratory flows.

According to the empirical evidence from our study, there is an inverse relationship between export levels and net migratory flow. We will elaborate the specific interpretation of this relationship after analyzing the effect of the import of goods and services feature on Group 5, addressing both features together.

Infant mortality is closely linked to health policies and the quality of health services in each country. It is not surprising that the regression coefficient of the variable is negative, indicating that the higher the levels of infant mortality, the lower the net migratory flows there will be. It is easy to infer that a high value of this indicator results in a disincentive when defining a migratory process towards the country in question. The infant mortality rate is nothing more than a reflection of the economic, social and environmental conditions in the health of mothers and babies, as well as the access and effectiveness of health systems.

Analyzing the alternative pole of the migratory flow, in those records belonging to the ***group 5*** we find a high ***negative net migration***. The main *features* with statistical significance linked to this set are the ***import of goods and services, population density, infant mortality (in children under 1 year of age)*** and ***working population***.

As far as the health system is concerned, once again the importance of health emerges through *Child mortality*. In this case, it can be seen that the sign of the regression coefficient is negative, which is why it is established that the higher the value of this indicator, the lower its migratory flow. Given the obviousness of the interpretation, we will not go into details regarding this indicator.

Recalling the definition provided in the dictionary of this document, "*labor force* " is defined as: "People ages 15 and older who supply labor for the production of goods and services during a specified period. It includes people who are currently employed and people who are unemployed but seeking work as well as first-time job-seekers". In this case, a negative correlation is observed between this feature and a negative net migratory flow. Consequently, the existence of a smaller labor force would imply a greater number of immigrants for a given country.

This relationship is expected in terms of the bases of International Trade theories, which consider labor (both skilled and unskilled) as one more factor of production, as well as raw materials and inputs. It is to be expected that there is an inverse relationship between labor and net migration, since when labor is scarce in a country, it is natural that it is compensated by external migratory flows. Conversely, when the labor force grows and tends to full employment of the productive factors, the labor force migratory flows will tend to be reduced, understanding that they would be less demanded.

It would be interesting to investigate this relationship, since it offers many study "edges". For example, in the context of a globalized world and with the advent of "remote" jobs, it will be essential for the correct development of the economy to have precise knowledge of the effects of the labor market to align the incentives of these skilled workers, who are usually the best paid. A good policy in this matter could generate a potential impact on multiple aspects of society, generating "spillovers" at the levels of consumption, collection, domestic investment, etc.

Interestingly, from the analysis of group 5, the variable *population density* also emerges. It presents a high and negative regression coefficient, with a very significant p-value

determining its preponderance in the analysis. In short, the finding shows that the registries with the highest net negative migratory flows are characterized by showing an inverse relationship to population density. This can be interpreted in various ways, and it would be necessary to break down the effect considering the numbers of immigrants and emigrants, but specifically it means that the more densely populated a given area is, the less migrants will enter or, in other words, the higher it will be the number of people who leave compared to the migrants who could enter.

It is opportune to emphasize that this interesting demographic effect presents a greater potential for analysis. In subsequent publications, it will be decided to segment said indicator into different parts in order to achieve a more precise understanding of its consequences.

In relation to imports of goods and services, a greater amount of data is necessary for the correct interpretation of its effect. Specifically, the quality of "importer" is positively related to migratory flows in countries with negative net migratory flows. Various arguments could be put forward regarding the causes behind this effect, such as that the country's productive matrix is purely importing, imports are the only thing that prevents migratory bleeding from being greater, etc. Finally, in this case, it is only possible to assert what was previously postulated as a concrete fact where the positive relationship between the variables is recognized.

In strictness to the accuracy of the interpretation, we must determine the scope of this work. With what has been developed so far, we have managed to categorize the different countries by groups, allowing the client to have a confidence interval of 95% certainty when it comes to knowing those specific issues on which special attention must be paid when making decisions that could influence the migratory flow of the following year.

The thoughtful reader will understand that constructing the variable *Net Exports* could illuminate this dilemma since it would provide another type of information. Finally, the intention at this point is to make it clear that the specific conclusions that explain the

nature behind the effect that we perceive require an alternative work methodology that exceeds the scope of this document.

In later versions we will delve into the detailed study of these situations that arise from the analysis. We will study specific cases that have aroused interest, and we will analyze "outlier" registries and countries.

In conclusion, as it was initially defined in this documentation[6], we have completed the task by delivering a powerful immigration policy tool that provides accurate and relevant information regarding the effect per se. It would be possible to make modifications to the model to adaptate it to the specific needs of a country, but it has generated more interest for us and we have found it more useful to explore the nature of the effect and provide a practical and effective instrument for evaluation and decision-making.

Certainly there are many debates focused on the evolution of migratory flows, many interests and many interested in driving the threads of this purely human phenomenon. It is our greatest desire to have contributed with hard work and effort developing a tool that allows migration policies to advance and be oriented in the future towards a healthier and more controlled practice, allowing the benefit of all human beings on this planet... After all, we all live within the same border.

[6] For more information, go to section 1.4*"Reach and Out of Reach"*

ANNEXES:

1. REPOSITORY:

All information related to this project is publicly accessible through the provided github repository.

https://github.com/Analytic-Hound-Consulting/ONG-Henry

EDA-ETL	Add files via upload
Weekly Sprint #1	Add files via upload
Weekly Sprint #2	Add files via upload
Weekly Sprint #3	Add files via upload
graph	Add files via upload
COPYING	Create COPYING
Dashboard.pbix	Add files via upload
Documentacion (ESP).docx	Add files via upload
Documentation (ENG).docx	Add files via upload
README(ENG).md	Update README(ENG).md
README.md	Update README.md

EDA - ETL	Databases (.csv files) EDA code and report (Exploratory data analysis)
Weekly sprint	Weekly log, graphic support, official report and deliverables
Graph	Graphic support
COPYNG	Licensing
Dashboard	Dashboard - Power BI
Documentación	Final documentation (English and spanish)
Readme	Repository's README (English and spanish)

2. BIBLIOGRAPHIC REFERENCES

- Informe ONU Migración 2022. Organización Internacional para las Migraciones (OIM).
- Presentación del Dossier estadístico anual sobre inmigración italiana. Publicación Actualidad Internacional Sociolaboral nº 228.
- eBrain – Using AI for Automatic Assessment at the Hong Kong Immigration Department. American Association for Artificial Intelligence. Wong, R.W.-M. y A.H.W. Chun
- Econometría, Segunda Edición. Damodar N. Gujarati. Copyright 1986 by McGraw Hill Inc.
- "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition, by Aurélien Géron (O'Reilly). Copyright 2019 Aurélien Géron, 978-1-492-03264-9."

ACKNOWLEDGMENTS:

The ***Analytic Hound ®*** team appreciates the collaboration of *Data World Bank, Data Hub* and *United Nations* in providing the necessary data for the development of this project. The correct analysis of the world migratory situation, as well as the development of the predictions made in this document, would not have been possible without the appropriate support of information provided by said sources.