Logotipo, nombre de la empresa

Descripción generada automáticamente

MÁSTER EN DATA SCIENCE Y BUSINESS ANALYTICS PRESENCIAL

Predicción del número de inmigrantes extranjeros en España mediante modelo de aprendizaje automático para prever flujos migratorios internacionales

**TFM elaborado por:**

Cristian de Andrade Correia

Erick Ernesto Hernández Lara

**Tutor/a de TFM:**

Miguel Martín

Madrid, 30 de octubre de 2024

# Índice

[Índice 1](#_Toc179489124)

[RESUMEN 2](#_Toc179489125)

[1. INTRODUCCIÓN 3](#_Toc179489126)

[2. ANTECEDENTES 6](#_Toc179489127)

[3. OBJETIVO 8](#_Toc179489128)

[4. MATERIALES Y MÉTODOS 8](#_Toc179489129)

[4.1 Adquisición de fuentes de datos 8](#_Toc179489130)

[4.2 Descripción de fuentes de datos 9](#_Toc179489131)

[4.3 Resumen de herramientas e infraestructura empleada 10](#_Toc179489132)

[4.4 Procedimiento 10](#_Toc179489133)

[5. RESULTADOS 12](#_Toc179489134)

[5.1 Etapa 1: Limpieza, preprocesamiento y análisis de datos de inmigración 12](#_Toc179489135)

[5.4 Etapa 2: Limpieza, preprocesamiento y análisis de variables explicativas 19](#_Toc179489136)

[5.4.1 Continentes y Subregiones 20](#_Toc179489137)

[5.4.2 Padrón de Inmigrantes Residentes en España 21](#_Toc179489138)

[5.4.3 Índices de Desarrollo 21](#_Toc179489139)

[5.4.4 Régimen Político 24](#_Toc179489140)

[5.4.5 Índices de Democracia 25](#_Toc179489141)

[5.4.6 Índices de Libertad y otros 29](#_Toc179489142)

[5.4.7 Tasa de Homicidios 29](#_Toc179489143)

[5.4.8 Conflictos Armados 32](#_Toc179489144)

[5.4.9 Turistas Anuales 34](#_Toc179489145)

[5.5 Etapa 3: Unión y selección de variables 34](#_Toc179489146)

[5.6 Etapa 4: Prueba y comparación de modelos de *machine learning* 34](#_Toc179489147)

[Análisis Exploratorio de Datos o EDA 34](#_Toc179489148)

[Conclusiones 40](#_Toc179489149)

[Referencias bibliográficas 41](#_Toc179489150)

[Anexos 44](#_Toc179489151)

# RESUMEN

# INTRODUCCIÓN

Según el Informe de Migraciones de 2022 de la Organización Internacional de Migraciones (OIM)1, de la Organización de las Naciones Unidas (ONU), la cantidad de migrantes internacionales ha incrementado de 153 millones en 1990 a 281 millones en 2020, acercándose a duplicar dicho valor, y que es, a su vez, más del triple de lo estimado en 1970.

De estos 281 millones de migrantes en el 2020, Europa y Asia acogieron alrededor de 87 y 86 millones respectivamente, que representa el 61% de la población mundial de migrantes, seguidos de Estados Unidos de Norte América con 21%.

Ya en el 2016 el Banco Europeo de Inversiones (BEI) menciona en su análisis de desafíos y oportunidades de la migración en Europa2, que la Unión Europea (UE) fue el mayor aportador durante la escalada de inmigrantes y aplicantes de asilo durante los años 2015 y 2016, con una asignación de 10,1 billones de euros en ambos años, indicando que una deficiente integración de los migrantes podría traducirse en una reducción del Producto interno bruto (PBI) per cápita en los años subsiguientes. Sin embargo, mantuvo la visión que, a largo plazo, las migraciones internacionales tendrían un efecto positivo en Europa y resaltaba la necesidad de modificar y mejorar los marcos legislativos en torno a las políticas de integración y optimización del flujo de migrantes.

El éxito de estos esfuerzos recae en el sinergismo de los diversos procesos que regulan el ritmo migratorio y la integración de los inmigrantes, que a su vez están asociados a la capacidad económica, judicial, logística y de defensa de los países receptores para dar respuesta a las exigencias que plantea el continuo aumento de estos movimientos migratorios internacionales.

Por otra parte, desde el punto de vista del individuo, también tenemos numerosos factores que impactan en su decisión por embarcarse en una travesía migratoria. La Organización Internacional para las Migraciones expone un diagrama (Figura 1) que hace un excelente sumario de la relación entre variables “macro”, “meso” y “micro” que afectan la decisión de las personas en migrar, las cuales involucran variables como la seguridad, libertad, empleo, precios, servicios básicos y salud (macro), actuando en conjunto con las condiciones, posibilidades y ventajas personales (micro y meso).

A diagram of a decision

Description automatically generated

[**Figura 1**](https://emm.iom.int/es/handbooks/contexto-global-de-la-migracion-internacional/factores-que-propician-la-migracion). Factores que propician la migración internacional. Factores que propician la migración internacional (Organización Internacional de Migraciones1).

Así, podemos observar que el estudio del proceso migratorio es un reto complejo debido a la diversidad de factores que lo afectan y el dinamismo de nuestro actual entorno globalizado. Esto se evidencia también en la creciente implementación de políticas en relación con la movilidad humana promovida por la ONU a partir de 2010 que incluye áreas como el cambio climático, desastres y desarrollo sustentable3. Si, además, consideramos eventos de origen bélico, como el actual conflicto entre Ucrania y Rusia, o Israel y el grupo terrorista HAMAS, tendremos una visión general de la complejidad del fenómeno de las migraciones internacionales.

Indudablemente, el mayor esfuerzo y responsabilidad recae sobre los países con mayor recepción de dichos inmigrantes internacionales. De acuerdo con la OIM, tres países de la UE formaban parte del top 10 de países destino para migrantes internacionales a nivel global en el 2020: Alemania, Francia y España, siendo España el país que ocupaba el décimo lugar a nivel global y el tercero de la UE1, que, además, alcanzó en el 2022 un saldo migratorio de 727.005 personas, el máximo nivel en 10 años4.

Aunque el fenómeno de la migración también atiende a decisiones personas que pueden estudiarse desde un punto de vista psicosocial, en este proyecto mantendremos nuestro enfoque en variables macro (por encima del individuo).

A razón de lo expuesto, este proyecto busca aprovechar estas tecnologías para estudiar los datos oficiales de inmigrantes en España y desarrollar un modelo predictivo como herramienta de estrategia y planificación que permita estimar el número de inmigrantes extranjeros al territorio español. Y, aunque el fenómeno de la migración también atiende a decisiones personas que pueden estudiarse desde un punto de vista psicosocial, en este proyecto mantendremos nuestro enfoque en variables macro (por encima del individuo).

# 2. ANTECEDENTES

Entre la diversidad de aplicaciones en los avances de la inteligencia artificial “clásica” o generativa, tenemos que los desarrollos en Aprendizaje Automático (AA), uno de los subcampos de la Inteligencia Artificial (IA), han permitido abordar problemáticas como la expuesta mediante algoritmos que permitan, por ejemplo, generar predicciones de inmigrantes y grupos específicos5, o incluso prever flujos migratorios relacionados con asilo mediante alertas usando datos de eventos y tendencias de Google6.

Aydemin et al. (2022)5 optaron por emplear un enfoque de predicción directa planteándose dos objetivos: i) predecir el grupo de ingresos (bajo, medio bajo, medio alto y alto) con base al porcentaje de inmigrantes en relación a la población y regiones, en conjunto con datos de variables de desarrollo del Banco Mundial –los primeros en emplear este enfoque– como grupos de edad, nivel de industrialización, áreas de agricultura, tasas de mortalidad, indicadores de servicios, etc., y ii) predecir el stock de inmigrantes de una nacionalidad en base al total a nivel mundial. Sus mejores resultados de tasa de éxito para predecir grupos de ingresos los obtuvieron con Regresión Logística (86.04%) y SVM/KNN (83.72% para ambos), mientras que para su estudio de regresión, mencionan una tasa de éxito de 98.37% (XGBoost) y 96.42% (*Random Forest*), pero no aclaran la métrica ni mencionan otras; se asume que se trata del R2.

Es importante mencionar que Aydemin et al. (2022)5 la relevancia de sus modelos considerando que uno de los intereses principales al proyectar migraciones es hacerlo para un amplio período de tiempo (años), sin embargo, su enfoque se ve limitado por la naturaleza de sus datos, los cuales implican el usado de variables de desarrollo, en consecuencia, estimar el futuro estado político, social y económico de las naciones es, en sí mismo, un desafío. En contraste, Carammia et al. (2022)7 optaron por enfoque de *rolling window* sobre datos históricos para su estudio predictivo de migraciones relacionadas a pedidos de asilo en la UE, enfocándose en predecir el número de solicitudes con una, dos, tres y cuatro semanas de anticipación, con la finalidad de proveer a las autoridades de una ventana de preparación en términos operacionales. Su estudio usó datos de eventos y búsquedas de internet de los países de origen, detección de inmigrantes en las fronteras de la UE / UE+ (país de origen seguro) y las tasas de aplicación y reconocimiento de asilos en ambos casos, observando que, en casi todas las semanas, sus predicciones se mantuvieron dentro de las bandas de ± 2 errores estándar.

La migración es un fenómeno que sin dudas puede explorarse desde distintos puntos de vista, así como componerlo con diferentes variables que pueden darnos un enfoque variado y concluyente. Tomemos por ejemplo el trabajo de Imán H.S y Tarasyev, A (2018)10 quienes observan que las decisiones sobre migrar o no dependen en gran medida de la posibilidad de obtener un salario satisfactorio. Los migrantes, bajo la óptica de las teorías de capital humano, la teoría neoclásica o la teoría estándar de equilibrio, tomarán decisiones racionales con base en la satisfacción salarial, mismas que pueden considerar la calidad de los mercados laborales en los países de origen y destino, la posible reducción de los salarios de las personas locales en el país destino como resultado de los flujos migratorios descontrolados o la suposición de encontrar un mercado laboral perfecto que supla la diferencia salarial del migrante.

Micevska, M. (2021)11 va incluso más allá e introduce una variable como el *internet* como un factor que afecta en gran medida la decisión de migrar, así como ahondar en variables más “tradicionales” como lo son presencia de conflictos en el país de origen y condiciones macroeconómicas generales. Internet se coloca como una variable de peso al influir en la capacidad de solicitar asilo y obtener información detallada antes de migrar Micevska, (2021). También se especifica que el cambio climático se ha convertido en una razón por sí sola a la hora de tomar la decisión de migrar, siendo básicamente estas personas en “migrantes climáticos”, Micevska, (2021).

El enfoque de nuestro estudio, si bien sigue una lógica empírica acerca de la disponibilidad de variables macroeconómicas, sociales, demográficas y políticas, no nos centraremos en encontrar estadísticamente todas las variables disponibles que influyan en menor o mayor medida al fenómeno de inmigración, sino en explicar la relación que tienen las variables que seleccionamos desde un punto de vista más conservador y reducido, haciendo énfasis en los residentes en España de origen extranjero y la situación macroeconómica y socio-política de los países de origen.

# 3. OBJETIVO

* Predecir el número de inmigrantes extranjeros en España mediante modelo de aprendizaje automático para prever flujos migratorios internacionales.

Objetivos específicos:

* Lista de nacionalidades mas importantes...
* Obtener un modelo predictivo con coeficiente de determinación cercano o superior al 0.80 (funcional).

# 4. MATERIALES Y MÉTODOS

### 4.1 Adquisición de fuentes de datos

Nuestro punto de partida fueron los datos de inmigración obtenidos del Instituto Nacional de Estadística de España5, los cuales constaban de un grupo de datos anuales de inmigrantes por nacionalidad y variables demográficas (sexo y grupo de edad) desde el año 2008 al 2021 (*24287.csv*), y otro grupo de datos entre 2021-2022 (*61623.csv*)12.

Adicionalmente, con base a nuestro enfoque macro, se realizó un exhaustivo trabajo de investigación durante tres semanas para obtener datos en torno a los factores macroeconómicos, sociales, políticos, libertades, seguridad, entre otros, a considerar para definir nuestras variables predictoras (Figura 2).

El principal desafío durante la investigación fue encontrar bases de datos que abarcaran nuestros años de interés (2008-2022) y que, además, estuviesen completos (ausencia de datos nulos). De manera que, para algunos casos, como para las variables económicas, se usaron varias fuentes en conjunto.

**Figura 2.** Fuentes de información, elaboración propia

A grandes rasgos, las fuentes de los datos consultadas y empleadas comprenden organismos no gubernamentales, estudios y trabajos de investigación, reportes periodísticos y organismos multinacionales, para los cuáles haremos referencia en el siguiente apartado.

### 4.2 Descripción de fuentes de datos

Adicional a los datos centrales de inmigración de España, nuestros datos comprenden una amplia diversidad de variables. Contamos con, por ejemplo, información poblacional de número de residentes, valores booleanos para representar ausencia o presencia de alguna característica, índices económicos, índices de libertad/democracia y también valores estandarizados de 0 a 1 en algunas variables para representar su grado de solidez relacionada al país.

La Tabla 1 (en Anexos) muestra el resumen de los distintos conjuntos de datos utilizados, la dimensión que abarcan, su descripción y/o justificación de su uso, variables que incluyen y enlaces web de referencia (si aplica). Adicionalmente, se agrega en la Tabla 2 (ver Anexos) definiciones relevantes de algunas variables que son intuitivas.

Según cada caso, los conjuntos de datos fueron sujetos a un preprocesamiento, análisis exploratorio y reducción dimensional previo a su uso para el entrenamiento de modelos de aprendizaje automático.

### 4.3 Resumen de herramientas e infraestructura empleada

**Base de datos**: Los datos recopilados y usados en este proyecto fueron datos estructurados/relacionales y, siendo un estudio temporal de inmigración, las claves centrales con los que se relacionaron los distintos ficheros empleados se basan principalmente en: año y nombre del país/código, pero pueden variar según los datos que aporta cada fichero.

**Repositorio**: Se creó un repositorio de GitHubx para los ficheros, notebooks, *exports* resultantes y el escrito final, trabajando progresivamente en el proyecto mediante ramas a nivel local, haciendo *commits* de trabajoy revisiones conjuntas.

**Infraestructura computaciona**l: Se trabajo en local mediante equipos personales y software de uso libre y/o estándar: VS Code, Git/GitHub, Jupyter Notebooks, MS Excel, MS Word y MS PowerPoint (diagramas). Adicionalmente, las librerías principales que se usaron en lenguaje python durante el proyecto fueron: pandas, numpy, matplotlib, altair y scikit-learn (*machine learning* y redes neuronales) ~~y tensorflow (redes neuronales)~~. Las versiones usadas fueron las siguientes:

Visual Studio Code 1.93.1

Jupyter 3.6.0

Python 3.11.5

Pandas 2.0.3

Numpy 1.24.3

Altair 5.0.1

Matplotlib 3.7.2

Seaborn 0.12.2

Plotly 5.9.0

Ipywidgets 8.0.4

Scipy 1.11.1

Scikit-posthocs 0.9.0

Scikit-learn 1.2.2

Xgboost 2.0.3

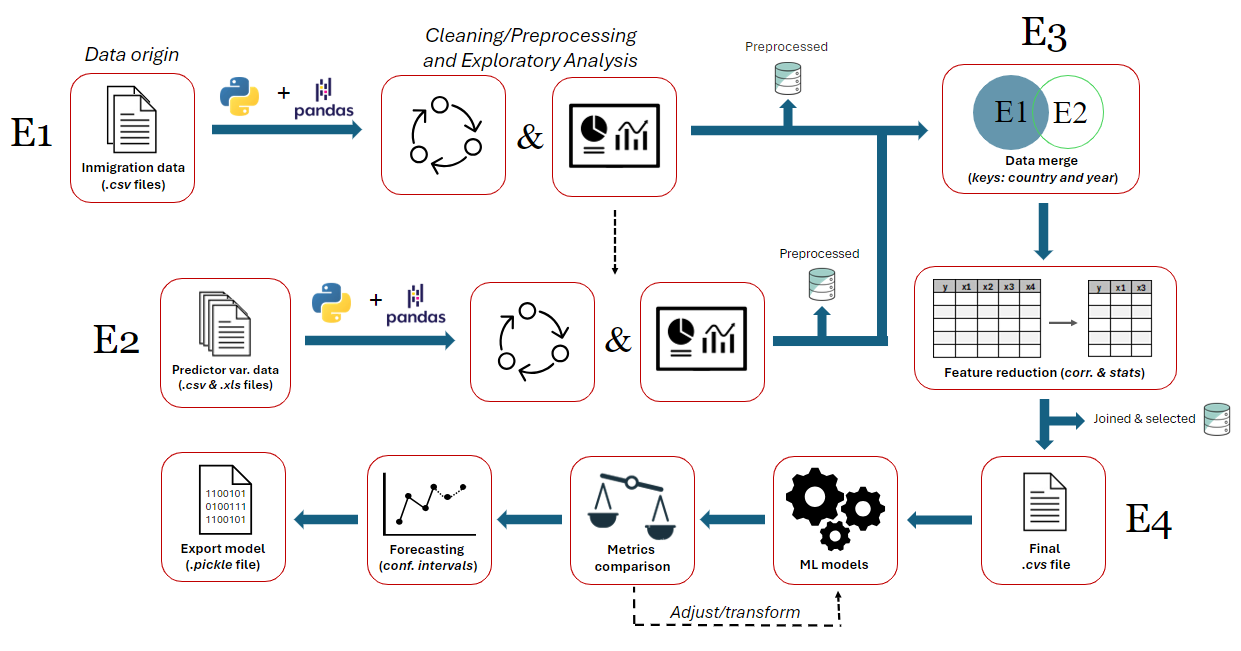
Tensorflow 2.16.1 ~~/ Keras 3.3.3~~

### 4.4 Procedimiento

Teniendo en cuenta la variedad de conjuntos de datos en nuestro estudio y la necesidad particular de limpieza/preprocesamiento de cada uno, se decidió dividir el proyecto en 4 fases (Figura 3) que proporcionaran una secuencia ordenada de limpieza, análisis y prueba de algoritmos de *machine learning*:

* **Etapa 1**: Centrada en el trabajo con los datos de inmigración de España segregados por variables demográficas.
* **Etapa 2**: Enfocada en las variables predictoras.
* **Etapa 3:** Unión de todos los conjuntos de datos para análisis de selección de variables.
* **Etapa 4:** *Machine learning*.

En la Etapa 1 (E1), se realizó la limpieza, preprocesamiento y exploración inicial de los conjuntos de datos de inmigración, obteniendo los primeros *insights* y el top países en número de inmigrantes a incluir en el modelo. Luego, a partir de las observaciones y top países obtenidos en la Etapa 1, se inició la limpieza, preprocesamiento y análisis de los distintos conjuntos de datos de las variables predictoras (Etapa 2 - E2), finalizando con la exportación de los conjuntos de datos preprocesados de ambas etapas.

**Figura 3**. Esquema de las etapas del proyecto, teniendo: i) Etapa 1 (E1) para la limpieza, preprocesamiento y análisis de los datos de inmigración de España, ii) Etapa 2 (E2) orientada a una limpieza, preprocesamiento y análisis orientado según los *insights* de la etapa 1, iii) Etapa 3 (E3) de unión y selección de variables y iv) Etapa 4 (E4) para la prueba y comparación de modelos de *machine learning* con nuestro conjunto de datos final para realizar predicciones.

Posteriormente, se procedió con la unión de todas las variables predictoras al conjunto de datos central de datos de inmigración (Etapa 3 - E3), la inclusión de tres variables categóricas adicionales relacionadas a identificar los países de habla hispana, y el período de pandemia/postpandemia (2020-2021 y 2022, respectivamente), así como la reducción de variables mediante el estudio de correlaciones y contrastes de hipótesis.

Finalmente, con los datos exportados en la Etapa 3, se probó algoritmos basados en modelos lineales y de árboles (Etapa 4 - E4), comparando distintas métricas para observar el rendimiento de los modelos, identificar los mejores y realizar ajustes/transformaciones de éstos para disminuir el error hasta obtener el modelo con el mejor rendimiento posible. Luego se procedió a exportarlo y realizar predicciones, incluyendo la estimación de un intervalo de confianza del 90%.

Cada etapa se desarrolló por separado en un *jupyter notebook*, los cuales están disponibles en la carpeta “Notebooks” asociada con este proyecto.

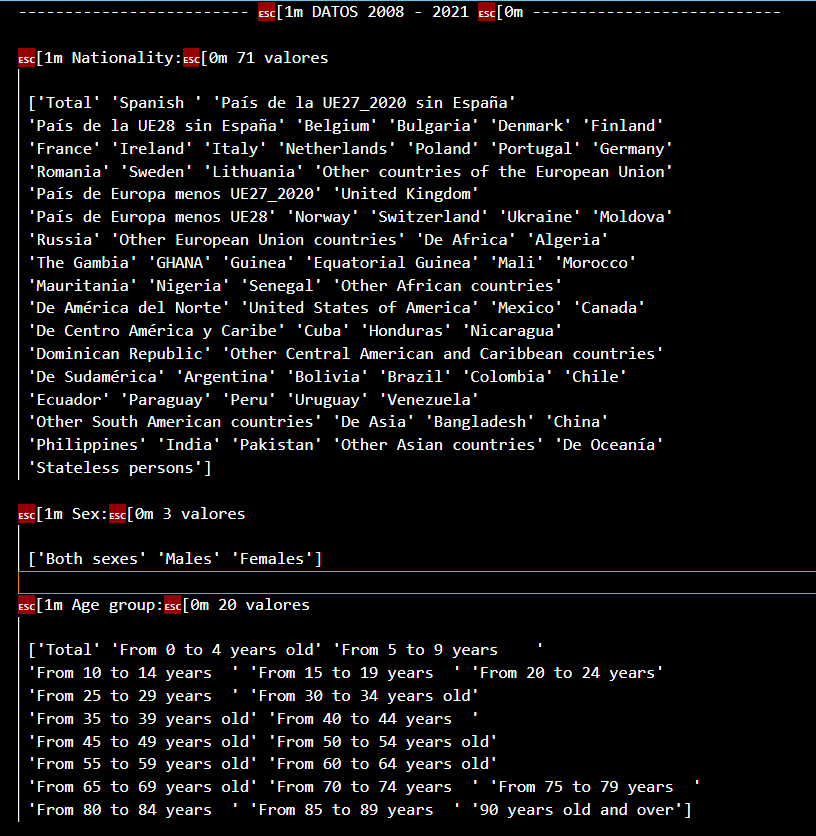
# 5. RESULTADOS

### 5.1 Etapa 1: Limpieza, preprocesamiento y análisis de datos de inmigración

Inicialmente, como ambos conjuntos de datos de inmigración (2008 - 2021 y 2021-2022) presentaban información del año 2021, se filtró el segundo de los conjuntos de datos de inmigración para tener únicamente los datos que corresponden al año 2022 (referirse al *jupyter notebook* “Etapa 1” para mayores detalles de esta etapa). Luego, se compararon los valores que tomaban las variables categóricas (Figura 4), observando que:

* El conjunto de 2008-2021 estaba en idioma inglés, mientras que el de 2022 estaba en español.
* Había una cantidad muy grande de nacionalidades, y también observamos agrupaciones de regiones/continentes como: "UE27\_2020 sin España", "Otro país de Asia", "América del Norte", entre otros.
* Los grupos de edades son diferentes entre los *datasets*, en los datos de 2008 – 2021 hay rangos de edades que aumentan intervalos de cinco años ("From 0 to 4 years old", "From 5 to 9 years", ...), mientras que en el segundo los rangos son más amplios ("De 0 a 15 años", "De 16 a 24 años", ...). Y también se observó que algunas de las categorías del 2008-2021 tenían espacios en blanco.

**(A)**

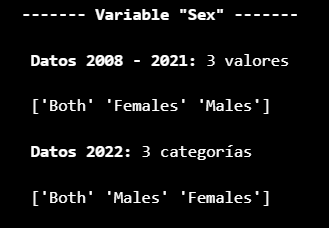
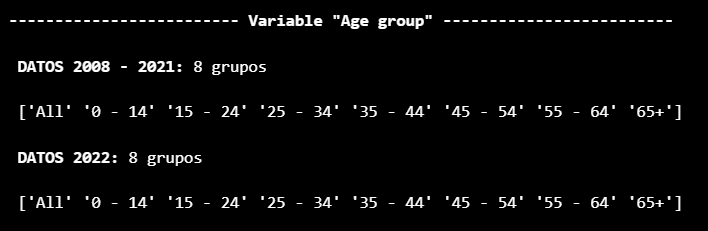


**(B)**



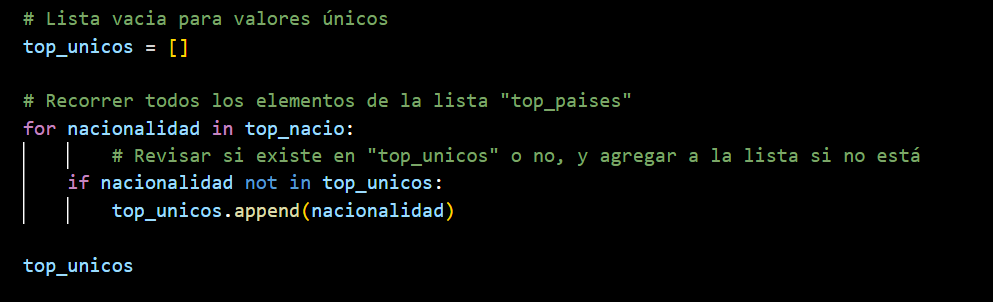
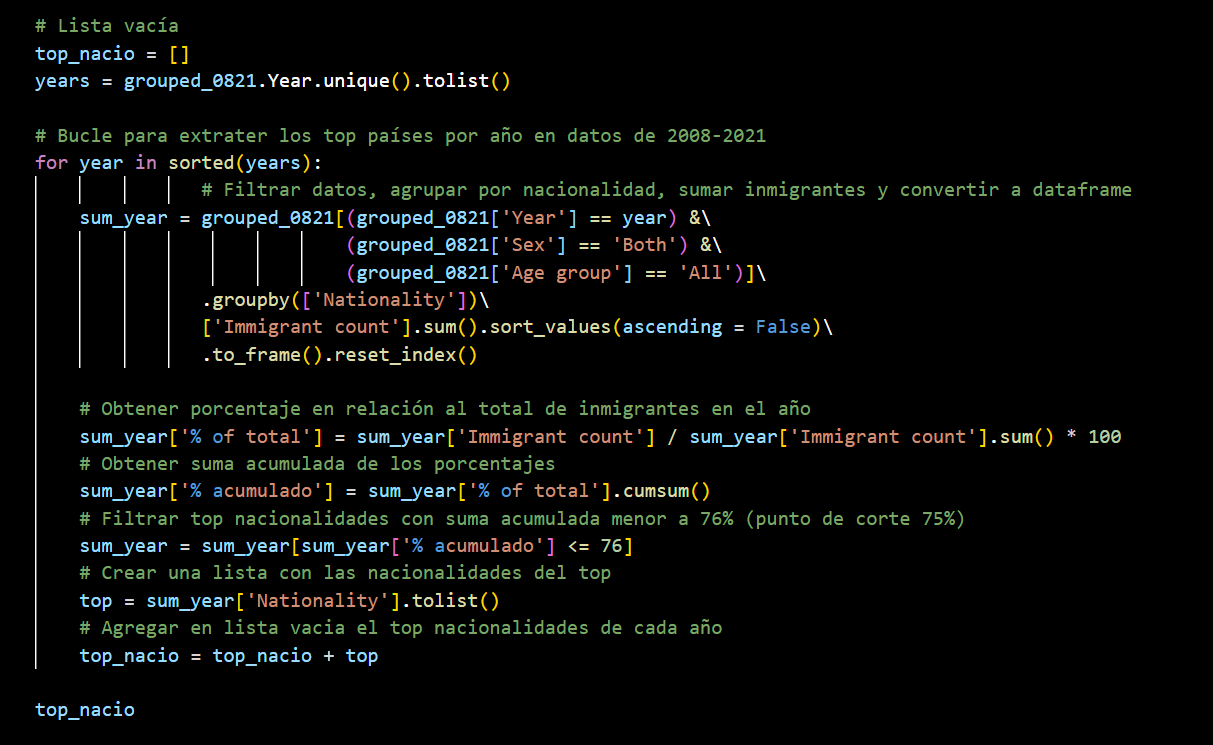
**Figura 4.** Valores que toman las categorías de nacionalidades, sexo y grupos de edad para los *datasets* de inmigración de 2008-2021 (A) y 2022 (B).

A partir de estas observaciones iniciales, se unificaron las categorías de sexo en inglés y se estandarizaron los grupos de edades (Figura 5).

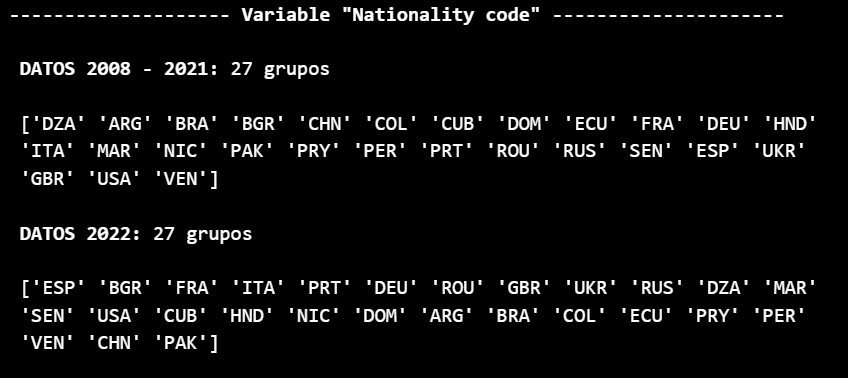
 

**Figura 5**. Valores las categorías de sexo (izquierda) y grupos de edad (derecha) post-limpieza/estandarización.

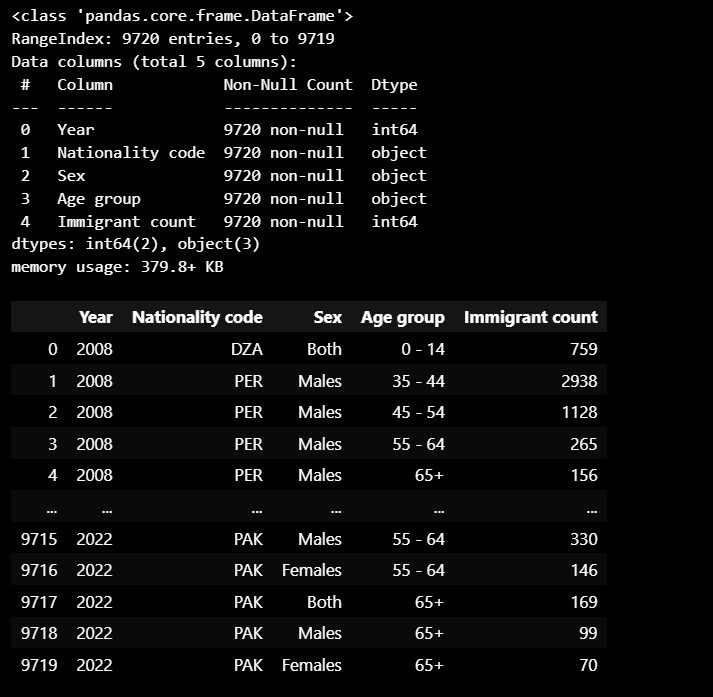
En cuanto a las nacionalidades, se removieron aquellas categorías relacionadas a regiones/continentes y se hizo una reducción del número de nacionalidades. Para ello, se consideraron como “top nacionalidades” a aquellas que, en conjunto y de forma decreciente, englobasen alrededor del 75% de los inmigrantes en cada año (Figura 6), seleccionando los valores únicos en una lista final para filtrar cada *dataset* con sus top, obteniendo un total de 27 nacionalidades (Figura 7).

**Figura 6.** Código empleado para encontrar las top nacionalidades del *dataset* 2008 - 2021 considerando cada año. Se siguió un procedimiento similar para el conjunto de datos de 2022.

Después de haber filtrado los datos de inmigración con el top nacionalidades, se estandarizó el formato de estas en una nueva columna “Nationality code” usando la nomenclatura internacional de tres códigos (ALPHA-3), descrita en la ISO 316613,para luego proceder a juntar ambos conjuntos de datos en un *dataframe* (Figura 8).



**Figura 7.** Top nacionalidades (en códigos ISO) de los conjuntos de datos de inmigración de 2008-2021 y 2022.



**Figura 8**. *Dataframe* de la unión de los *datasets* de datos de inmigración post-limpieza/preprocesamiento

Finalmente, continuando con los análisis de los valores totales de inmigración con relación a su evolución en los años (Figura 9) y por grupo de edad/sex (Figura 10), se observó:

* Entre 2008-2013 hay una caída de número de inmigrantes y a partir del 2014 aumenta progresivamente hasta alcanzar un pico de 750,480 inmigrantes en 2019, para luego disminuir en el 2020 y 2021 (período de pandemia y restricciones sanitarias relacionadas al COVID 19) y mostrar un gran salto hasta ≈1.25 millones de inmigrantes en el 2022 (postpandemia), superando al 2019 en aproximadamente 500,000 inmigrantes.
* Observando los inmigrantes según el grupo de edad, vemos que predomina la inmigración de personas jóvenes de entre 25-34 años, seguidos de jóvenes de entre 15-24. A partir de este último grupo, es notable la disminución progresiva del número de inmigrantes con el incremento de la edad.
* En cuanto al sexo, vemos que la cantidad de mujeres y hombres es similar en todos los grupos, especialmente en los grupos de mayor presencia (15-24 y 25-34 años). Únicamente los grupos de entre 0-14 y 55-64 años muestran una mayor diferencia: el primero hacia los hombres y el segundo hacia las mujeres.

A graph with a line and a blue line

Description automatically generated

**Figura 9.** Total de inmigrantes en España desde el año 2008 al 2022.

A graph of a number of people

Description automatically generated

**Figura 10**. Cantidad de inmigrantes en España por grupo de edad y sexo.

Si observamos la distribución por año en mediante gráficos de cajas (Figura 11), vemos que a medida que hay mayor inmigración, también hay una mayor dispersión de los datos y los valores atípicos son más “extremos”. Esto se debe a que el flujo de inmigrantes por nacionalidad no es igual y hay algunas que predominan sobre otras, las cuales generan la dispersión observada. Adicionalmente, observando el rango intercuartílico entre 2008-2021, notamos que la mayoría de los datos se concentran en un rango medio/bajo de número de inmigrantes, alrededor o por debajo de 10,000 inmigrantes.

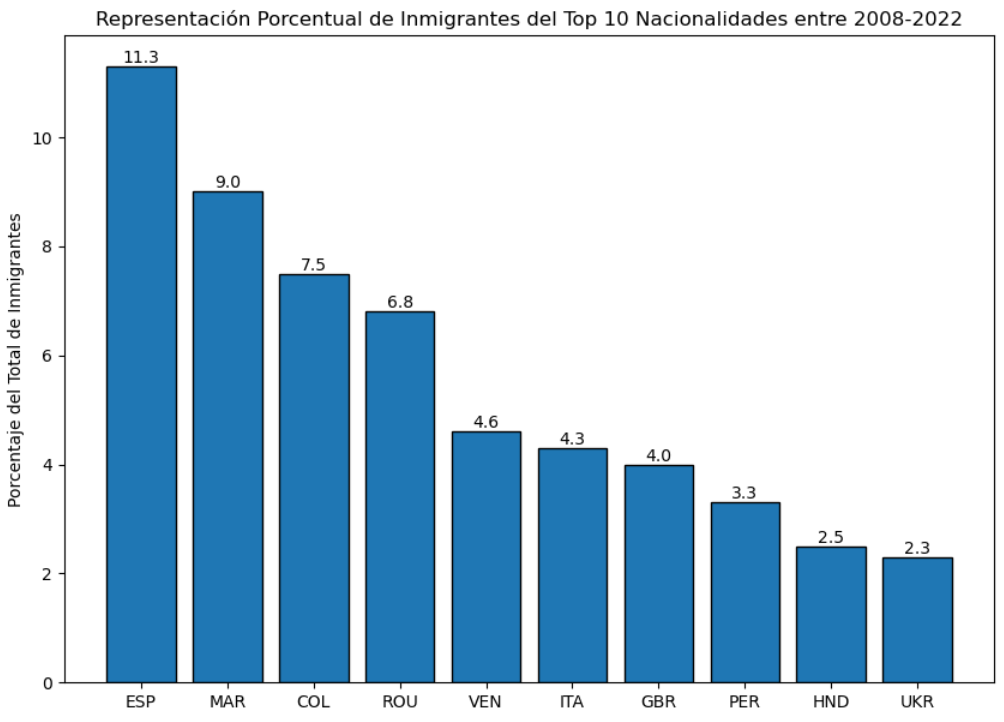
A screenshot of a computer

Description automatically generated**Figura 11.** Distribución de inmigrantes en España por año desde el 2008 a 2022.

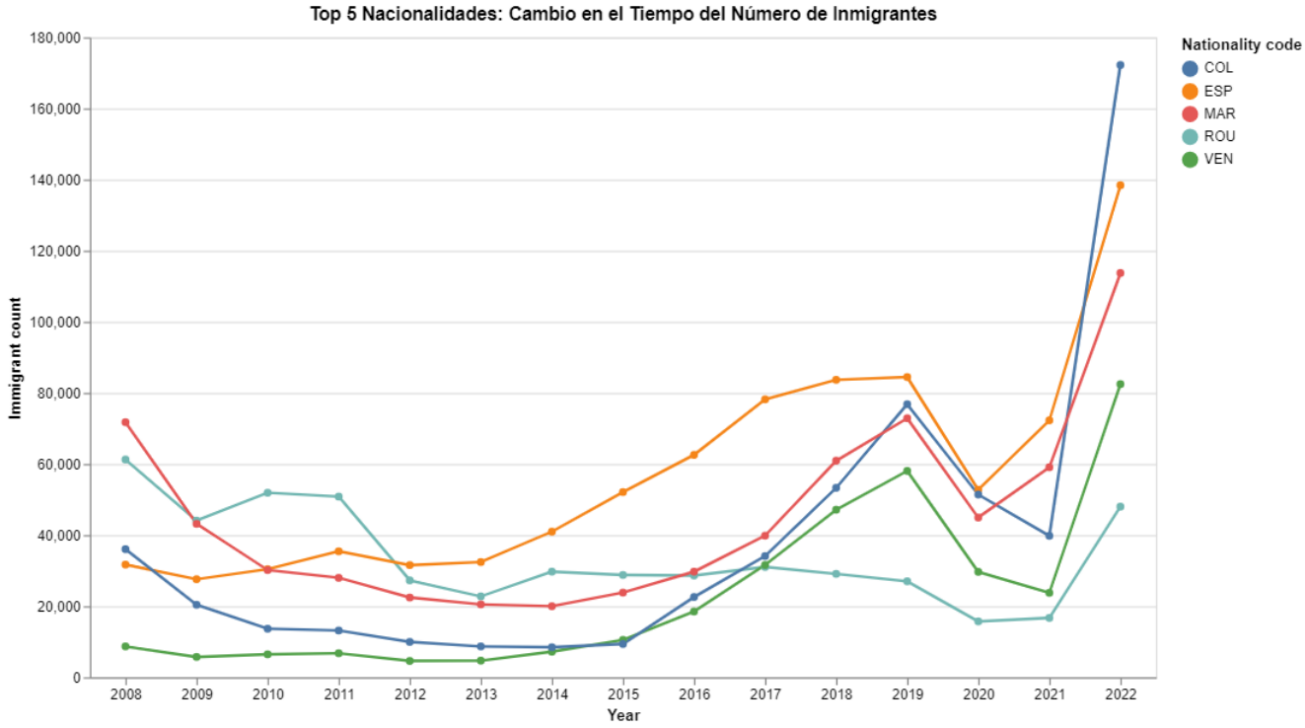
A partir de estas observaciones, se analizó la cantidad total por nacionalidad, observando que los inmigrantes de nacionalidad española, colombiana, marroquí y rumana poseen la mayor representación porcentual en relación con el total de inmigrantes (Figura 12A). Además, si observamos el número de inmigrantes por año dentro del Top 5 (Figura 12B), notamos que hay diferencias entre los mismo. Por ejemplo, los inmigrantes colombianos y venezolanos están en un rango medio/bajo durante 2008-2015, pero a partir del 2016 su número comienza a aumentar drásticamente. Incluso podemos ver como los colombianos se convierten en la nacionalidad con el mayor número de inmigrantes en el 2022 (postpandemia).

Así mismo, vemos como los rumanos presentan un descenso a partir del 2012 y el número de inmigrantes se estabiliza. Por otro lado, en cuento a crecimiento, el incremento de españoles es similar a venezolanos y marroquíes a partir del 2012, pero vemos que su número incrementa a un ritmo diferente.

**(A)**



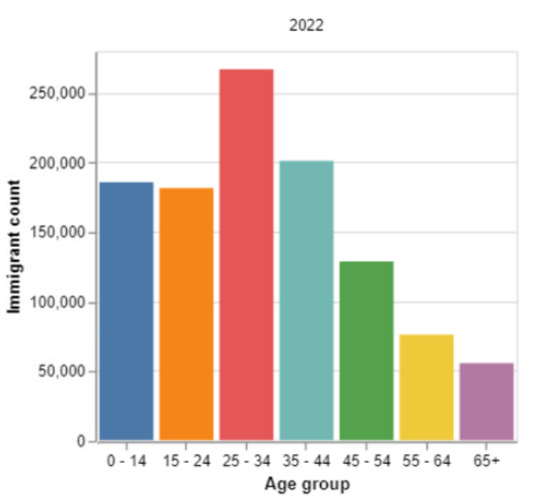
**(B)**



**Figura 12.** Top 10 nacionalidades en porcentaje de inmigrantes en España durante el periodo 2008-2022 (A) y el cambio en el número de inmigrantes en ese período para el Top 5 (B).

Estos datos indican que las condiciones específicas de cada nación, o nación de origen, del inmigrante tiene un efecto importante a tener en cuenta, por lo que, emplear el enfoque macro por país parece ser adecuado.

Otra observación que cabe destacar, relacionada con los grupos de edades, fue que únicamente en el año 2022 se observa un cambio en la tendencia en comparación con la observada previamente en los valores totales de inmigración. Para el 2022, tenemos que el grupo de 15-24 años presenta una cantidad de inmigrantes menor a la del grupo 0-14 años y el grupo de 35-34 supera en cantidad a ambos (Figura 13).



**Figura 13.** Cantidad de número de inmigrantes en España por grupo de edad en el año 2022.

Es importante destacar que, antes de proceder a exportar el *dataset* con los datos de inmigrantes limpios y preprocesados, se procedió a remover a la nacionalidad española del mismo considerando que nuestro enfoque macro implica que usaremos variables/datos de los países de cada nacionalidad. Por lo que, se estarían asociando datos relacionados con España a los inmigrantes. En consecuencia, los datos de inmigrantes de nacionalidad española requieren un estudio y modelado aparte que no se abarca en este proyecto.

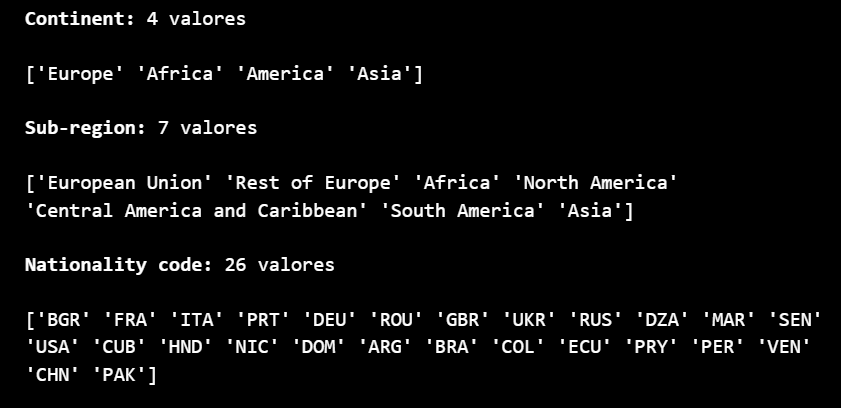
El fichero en formato *.csv* de los datos de inmigración obtenido de este proceso de limpieza y preprocesamiento, puede en la carpeta “13 - Exports (preprocesamiento)” asociada a este proyecto.

### 5.4 Etapa 2: Limpieza, preprocesamiento y análisis de variables explicativas

Los resultados se presentan segmentados por variable, enfocados en los aspectos más relevantes, debido a las particularidades que presentaba cada uno de los conjuntos de datos y con la finalidad de seguir el mismo esquema de trabajo empleado. Puede consultarse el *jupyter notebook* “Etapa 2” en la carpeta “Notebooks” para mayores detalles del proceso de limpieza y análisis, así como los ficheros *.csv* resultantes del preprocesamiento de cada conjunto de datos en la carpeta “13 - Exports (preprocesamiento)”.

#### 5.4.1 Continentes y Subregiones

Los datos de contenientes y sub-regiones14 se filtraron por mapeo con el top nacionalidades obtenidos de la Etapa 1 y se agregaron los códigos de país correspondientes, abarcando cuatro contenientes y siete subregiones (Figura 14).



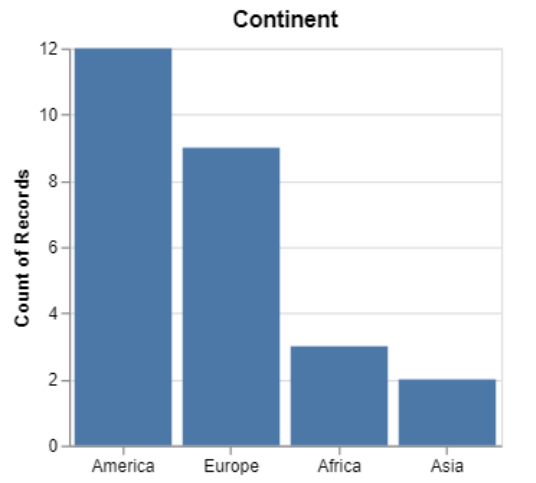
**Figura 14**. Contenientes y subregiones luego de aplicar el filtro del top nacionalidades en inmigración en España.

Al graficar la distribución de nacionalidades en el top de inmigración en relación con los contenientes y subregiones (Figura 15), se observó que:

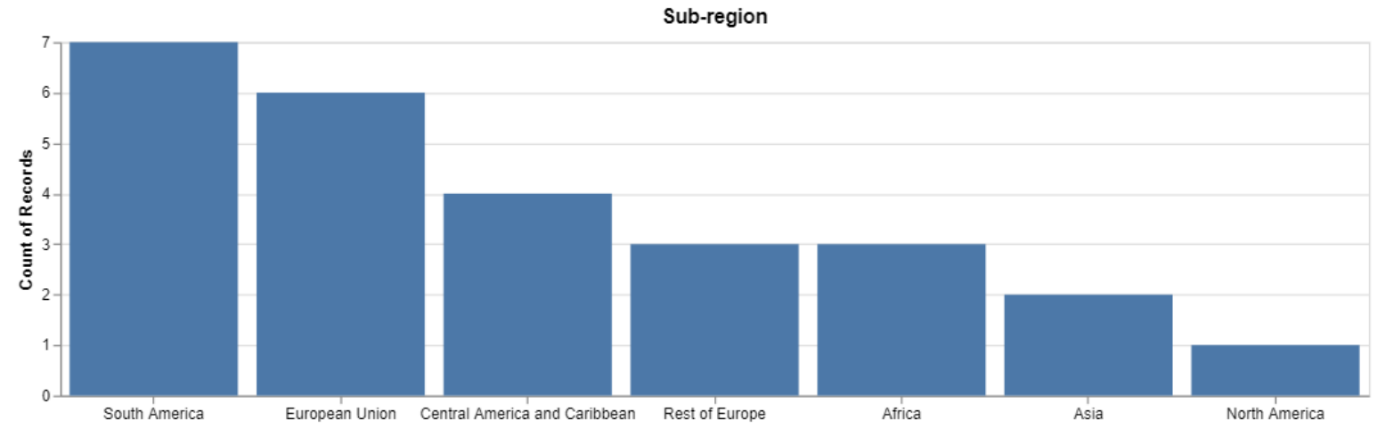
* América es el continente con mayor cantidad de países dentro del top nacionalidades en inmigración (≈ 45%, o 12 de 26), seguido de Europa con ≈35% (9 de 26).
* Si se toma en cuenta las subregiones, vemos que la mayor cantidad dentro de América se ubican en América del Sur (7 de 26), y que los de Europa son, mayoritariamente, parte de la Unión Europea, de la cual España también forma parte.

Esto se contrasta con las observaciones hechas en la Figura 12A, donde incluso se ven a países como Colombia, Venezuela, Perú y Honduras dentro del top 10 de nacionalidades en inmigración durante 2008 - 2022.

**(A)**



**(B)**

**Figura 15.** Cantidad de países por continente (A) y subregión (B) dentro del top nacionalidades en inmigración en España.

#### 5.4.2 Padrón de Inmigrantes Residentes en España

Para los datos del padrón de inmigrantes residentes en España, encontramos situaciones similares a las encontradas en el conjunto de datos de inmigración de 2008 a 2022, principalmente siendo éstos agrupaciones respecto a los países de origen de los inmigrantes

A screenshot of a computer program

Description automatically generated

**Figura XX.** Extracto de grupos de países encontrados en padrón de residentes

A screen shot of a computer

Description automatically generated

**Figura XX.** Extracto de estandarización con ISO3

El resultado sobre el análisis de los datos de los residentes arroja claras tendencias sobre el comportamiento actual del fenómeno de inmigración en España. Es siempre notoria la presencia de ciertos países más que otros en los datos históricos que tomamos como base. También observamos que los niveles de población inmigrante van cambiando conforme pasan los años.

A graph of a number of people

Description automatically generated

**Figura XX.** Suma de residentes por país de origen en el año 2022.

A graph of different colored lines

Description automatically generated

**Figura XX**. Principales 5 países en la inmigración en España, a través del tiempo

A graph of different colored lines

Description automatically generated

**Figura XX.** Bottom 5 países en la inmigración a España, a través del tiempo

Nuestros datos base, cabe destacar, tienen un peso importante en nuestros modelos que más adelante expondremos. Como conclusiones generales, observamos que:

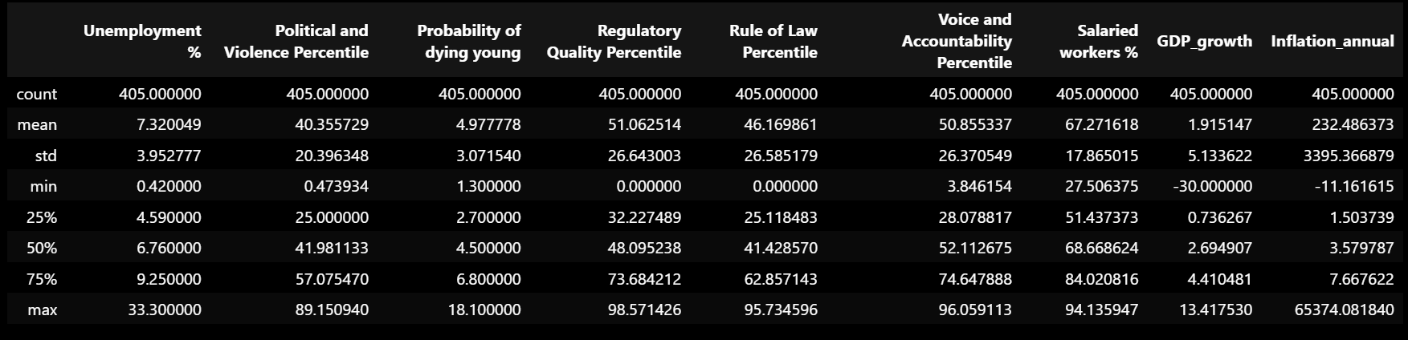
* Marruecos y Rumanía se colocan como los principales países de origen de inmigrantes.
* Las tendencias de inmigración tienen poco que ver con lazos de idioma o históricos

#### 5.4.3 Índices de Desarrollo

El estudio de los diversos índices de desarrollo socioeconómico requirió el uso de diversas fuentesx, y, z, incluyendo su adecuada transformación, para completar la información faltante de algunos dentro del período 2008 – 2022, siendo necesario imputar con la media un solo valor de la variable “Salaried workers %” para Ucrania.

La descripción inicial de los datos de los índices de desarrollo (Figura X) mostró todas las variables presentan alta dispersión (observa la relación de la media/desviación estándar y mínimo/máximo). Lo que se aprecia en mayor magnitud para la variable económica “Inflation\_annual”, que presenta una dispersión muy alta.

Esto nos dice que tenemos países con índices en ambos extremos, tanto niveles bajos como niveles muy altos.

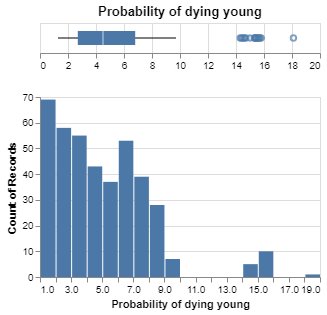
**Figura X**. Tabla de estadísticos descriptivos para las variables de índices de desarrollo.

Al observar la distribución de estas variables y los datos atípicos (Figura X), resaltaron los siguientes aspectos:

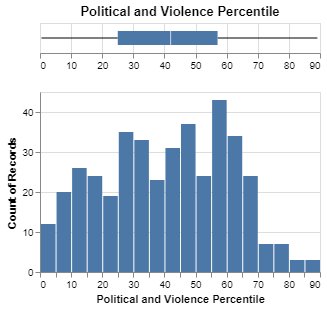
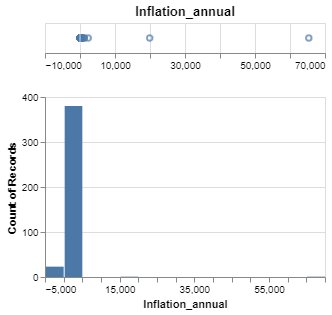
* Los datos de las variables “GDP\_Growth” (Figura XY) y “Unemployment %” (Figura XY) son las que se asemejan más a una distribución normal, pero ambas presentan valores extremos atípicos.
* La inflación anual (Figura XY) presenta datos tan dispersos que un histograma sólo puede agruparlos en valores de inflación negativa (desinflación) y valores positivos (aumento de inflación). Nótese que en los gráficos de cajas solo se distinguen los valores atípicos debido al amplio rango en los valores.
* Para “Probability of dying young” (Figuran XY), la mayoría de los valores están en un rango de probabilidad baja, con algunos datos que muestran una mayor probabilidad, mientras que para "Salaried workers %" (Figura XY) predomina un porcentaje medio/alto.
* El resto de las variables no poseen una tendencia clara, sino que presentan grupos con menor o mayor valores a lo largo del rango de los datos.

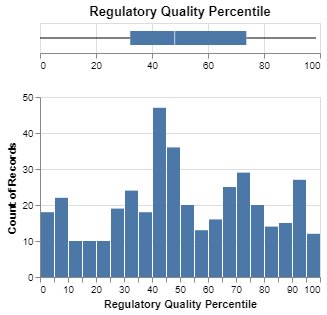
A graph of law percentages

Description automatically generatedA graph of growth and a chart of growth

Description automatically generated

A graph of blue bars

Description automatically generated

A graph of workers and a number of workers

Description automatically generatedA graph of unemployment and unemployment

Description automatically generated

**Figura X.** Distribución de las variables de índices de desarrollo y detección de datos atípicos. (A) “Rule of Law Percentile”, (B) “GDP\_growth”, (C) “Probability of dying young”, (D) “Voice and Accountability Percentile”, (E) “Inflation\_annual”, (F) “Political and Violence Percentile”, (G) “Regulatory Quality Percentile”, (H) “Salaried workers %” y (I) “Unemployment %”.

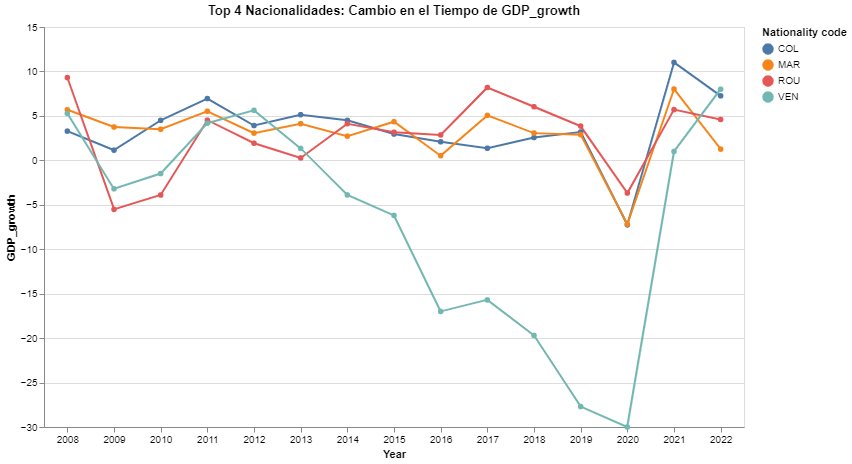
Luego, al comparar el cambio en el tiempo de alguna de estas variables en contraste con la tendencia de los datos de inmigración del Top 4 −Colombia, Marruecos, Rumania y Venezuela− entre las nacionalidades (Figura XY), se observó que:

* Para Venezuela, observando el rango de 2013 a 2019, es evidente la correlación inversa con la inmigración en los datos de PBI (Figura XY vs Figura XY). Sin embargo, dicha relación es de menor magnitud en los otros top países.
* También observamos como, en relación con los percentiles de estado de derecho y estabilidad política/violencia (Figura XY), los países del Top 4 están en los percentiles medios/bajos.
* Es interesante ver que los dos países de América del Sur del Top 4 (Colombia y Venezuela) están en los percentiles más bajos, particularmente a partir del 2017, siendo Venezuela el caso más extremo.
* Nótese también que Rumania mantiene un mejor percentil en el tiempo y es, además, la nacionalidad que muestra la menor variación en el número de inmigrantes por año. Caso contrario para Colombia, Venezuela y Marruecos cuyo número de inmigrantes incrementa significativamente y son, además, los países peor ubicados en los percentiles.

A graph of a number of immigrants

Description automatically generatedA graph of different colored lines

Description automatically generated

A graph of different colored lines

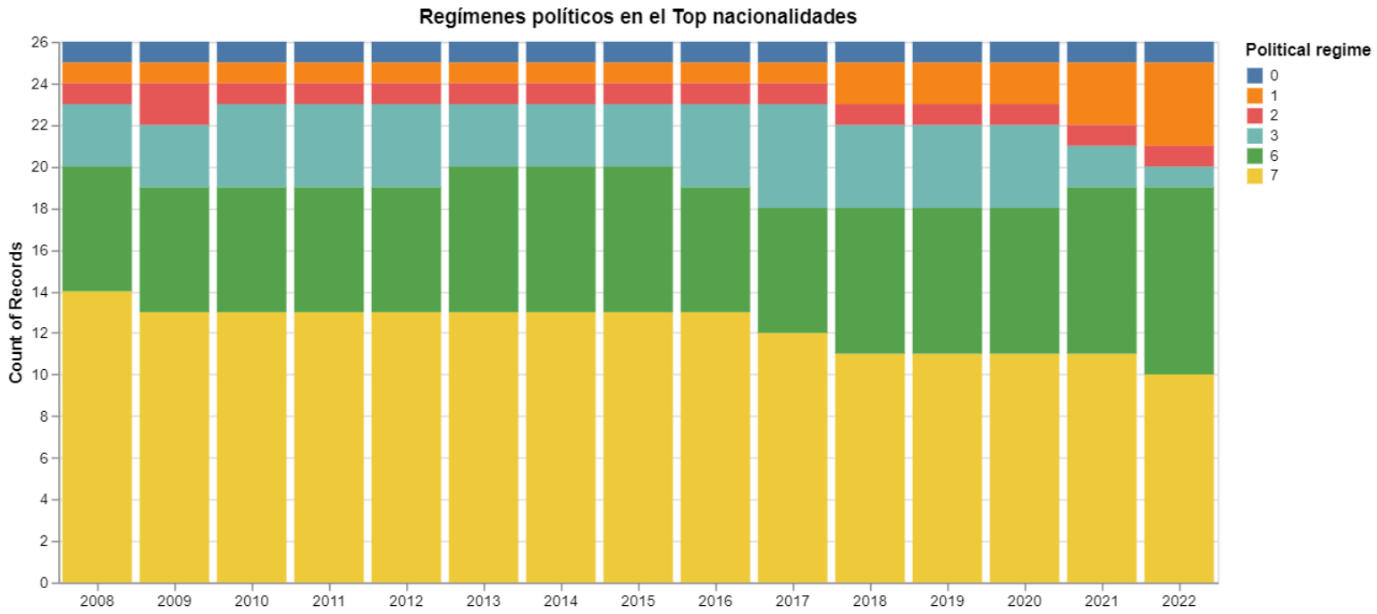
Description automatically generated**Figura X**. Contraste entre la tendencia de cambio en el tiempo de los datos de inmigración del Top 4 nacionalidades (A) y algunos índices de desarrollo: (B) "Political and Violence Percentile”, (C) ”GDP\_growth“, y (D) “Rule of Law Percentile”.

Esto nos muestra que cada país/variable se comporta y relaciona de forma diferente con la inmigración. Además, esto también deja en claro la necesidad de incluir más variables que describan otros tipos características y escenarios.

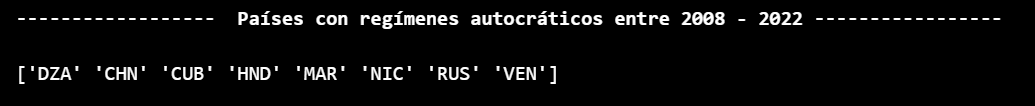
#### 5.4.4 Régimen Político

Los datos de regímenes políticosx están en una escala discreta del cero al siete, yendo desde los tipos de regímenes más autocráticos hasta los más democráticos (ver Tabla 2 en Anexos para mayor detalle).

Luego de filtrar las 26 nacionalidades (top nacionalidades) entre 2008 - 2022 y graficar la cantidad de cada tipo de régimen por año (Figura X), se observó que predominan regímenes políticos del tipo "7" (poliarquías), seguido de tipo "6" (democracias electorales). También vemos la ausencia de regímenes "4" y "5" (democracia excluyente y democracia masculina), y luego tenemos un grupo que son menores o iguales 3 (autocracias). Además, se observa que los conteos se mantienen hasta el 2015 y, a partir de allí, hay algunas variaciones, de las cuales resalta el incremento de la cantidad de regímenes tipo "1" (autocracia unipartidista), hasta llegar a estar presente en 4 países para el año 2022, y la disminución de regímenes tipo "7" e incremento de tipo "6" hasta alcanzar una cantidad similar en el 2022.

**Figura X**. Conteo de países en el top nacionalidades en inmigración por cada tipo de régimen político por año.

Adicionalmente, entre los países con algún tipo de régimen autocrático (0, 1, 2 o 3) entre 2008 – 2022, se identificaron tres países dentro del Top 10: Marruecos, Venezuela y Honduras (Figura X).

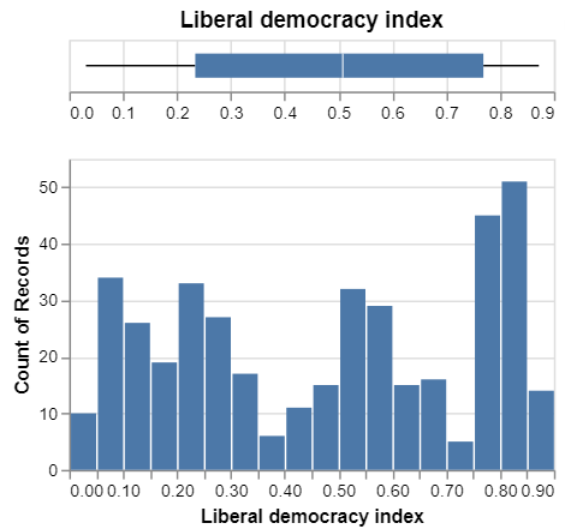
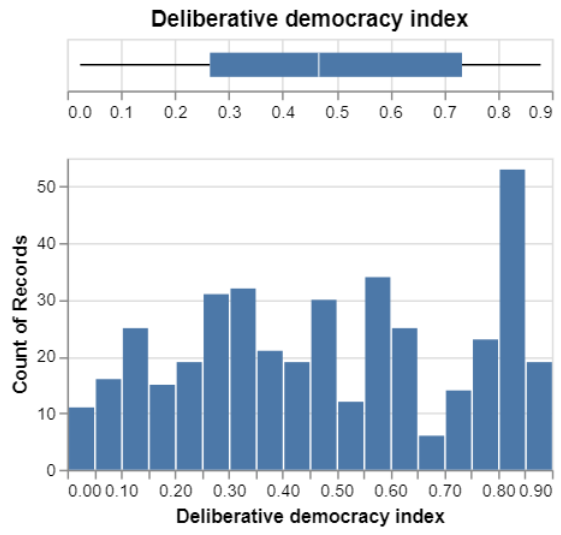


**Figura X**. Países dentro del top nacionalidades en inmigración con regímenes políticos autocráticos entre el año 2008 y 2022.

#### 5.4.5 Índices de Democracia

Los datos de índice de democracia liberal y deliberativa basado en las estimaciones de expertos y el índice de V-Demx muestran una tendencia similar, en cuanto su distribución, con algunas de las variables de índices de desarrollo observadas en el punto 5.4.3 (Figura X), pudiéndose destacar que:

* El índice de democracia liberal presenta tres grupos de concentración de datos entre nuestros top nacionalidades en inmigración: aquellos con índice bajo (< 0.35), otros en la zona media/media-alta y otro con índices de democracia alto (> 0.75).
* Vemos que para el índice de democracia deliberativa no se distinguen tres grupos como en el anterior, sino que observan algunos picos en las tres zonas, como entre 0.25 - 0.35, 0.55 - 0.65 y 0.80 - 0.85.

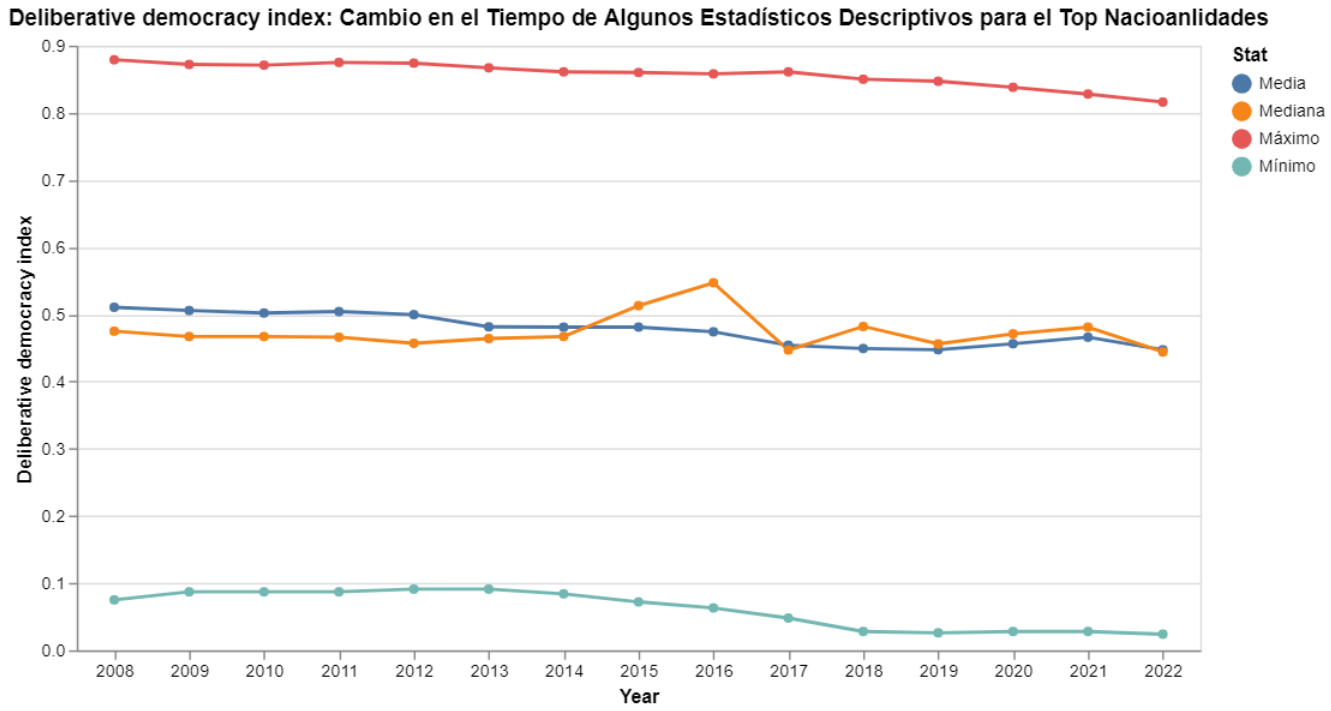
 

**Figura X**. Distribución de las variables de índice de democracia liberal (A) y deliberativa (B), y detección de datos atípicos.

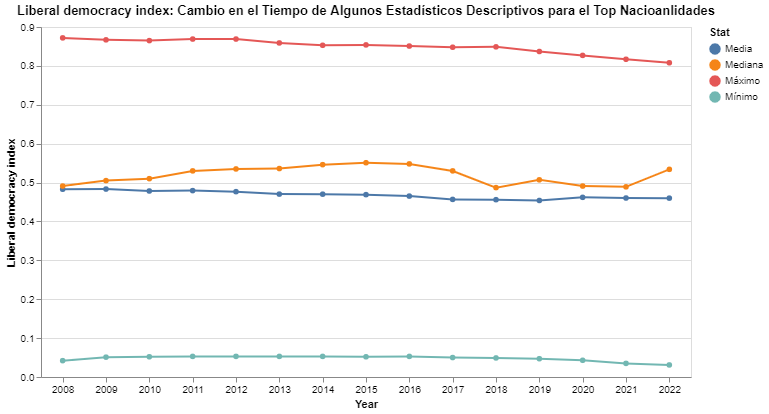
Además, luego de estimar y graficar algunos estadísticos descriptivos por año para ambos índices (Figura X), se observó que:

* Considerando las observaciones en cuanto a la distribución, se esperaba, observar el mínimo en torno a 0 y máximo próximo a 1, pero la similitud entre la media y mediana nos dice que la relación entre los índices de los países en el top de nacionalidades cambia en una forma similar a través del período 2008 - 2022.
* A partir de 2016 hay una leve disminución de los índices de democracia, especialmente notable en el índice de democracia deliberativa.

**(A)**



**(B)**

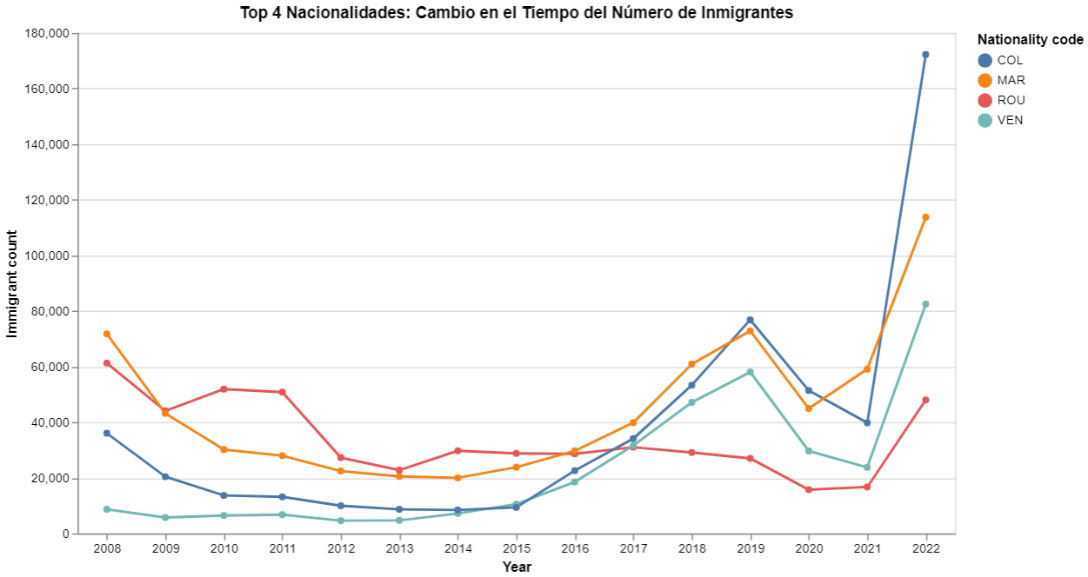


**Figura X**. Variación anual del máximo, mínimo, media y mediana para el índice de democracia deliberativa (A) y liberal (B).

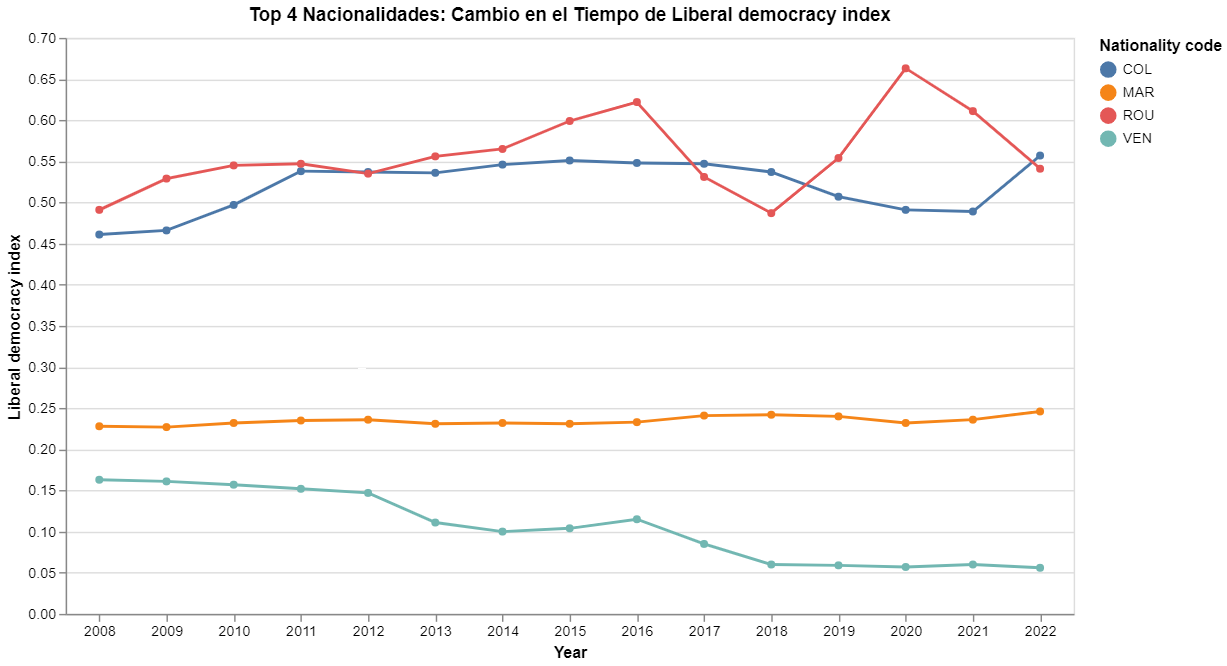
Finalmente, al contrastar el cambio en los años de la cantidad de inmigrantes (Figura XA) con los índices de las Top 4 nacionalidades en inmigración (Figura XB, XC) vemos que:

* Hay mayor variación en el tiempo para el índice de democracia deliberativa.
* Los índices de Colombia y Rumania oscilan en la zona media, mientras que Marruecos y Venezuela se mantienen en la zona baja.
* Previamente, en el análisis de los índices de desarrollo, se observó más similitudes entre Rumania/Marruecos y Colombia/Venezuela (ambos de América del Sur), pero ahora vemos más cercanía entre Venezuela/Marruecos y Rumania/Colombia, por lo también deben considerarse el régimen político en el período 2008 – 2022. Así, tenemos que Colombia y Rumania tienen democracias electorales, mientras que Marruecos y Venezuela presentan regímenes autocráticos.
* Para Marruecos, ambos índices se mantienen alrededor de 0.25 en el tiempo, mientras que en Venezuela se observa una disminución progresiva a partir del año 2013, punto en el cual también empieza a aumentar la inmigración de venezolanos a España.
* Vemos que para Colombia hay un incremento y estabilización entre 2009 - 2017, seguido por una caída que puede relacionarse con incremento de inmigrantes colombianos. Para Rumania, vemos un incremento en los índices entre 2008 - 2016, período en el cual disminuye la inmigración progresivamente hasta estabilizarse.

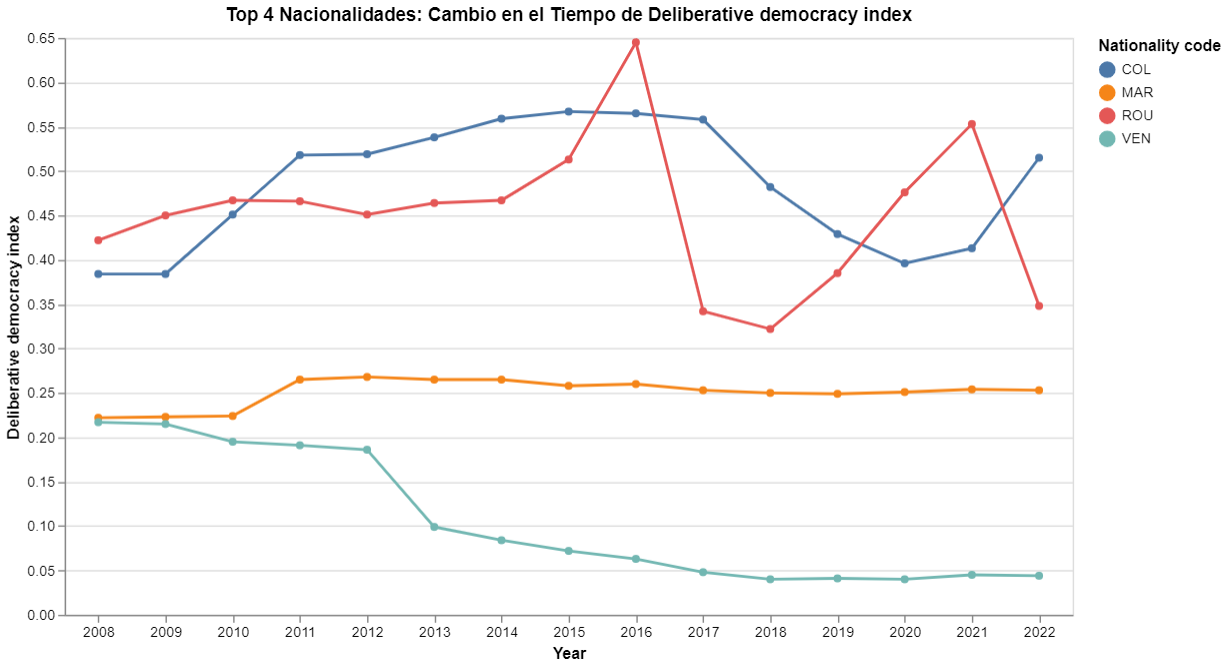
**(A)**



**(B)**



**(C)**



**Figura X**. Variación anual de la cantidad de inmigrantes en el Top 4 nacionalidades (A) en contraste con la variación de sus índices de democracia liberal (B) y deliberativa (C).

#### 5.4.6 Índices de Libertad y otros

Nuestro archivo de trabajo se basa en las investigaciones de *International IDEA,* mismas que se vierten en el compendio de índices *Global State of Democracy* (GSoD)x. En este archivo, venían muchos índices que no nos eran de utilidad todos, por lo que procedimos a realizar un limpiado y estandarización de datos.

Para este *dataframe*, realizamos un *pivot*, eliminamos cadenas de texto innecesarias en la escala de los índices (informativamente, todos los índices traían la leyenda “highest score=1”).

A screenshot of a computer program

Description automatically generated

**Figura xx**. Dataframe antes de preprocesamiento

A screenshot of a computer

Description automatically generated

**Figura XX**. Muestra de indicadores presentes el estudio, suman 172.

A screenshot of a computer

Description automatically generated

**Figura XX**. *Dataframe* final con países selectos y *pivot* aplicado.

Los índices que usados fueron: *"Absence of Corruption", “Public sector corrupt exchanges", "Civil Liberties", "Judicial accountability", "Educational equality" y "Health equality*".

En general, los países tienen datos que varían de acuerdo con su composición política, cultura y economía, aunque en general, la mayoría de ellos tiene un rendimiento por indicador cercano al 0.5 (1 es la calificación más alta).

A graph with a line

Description automatically generated

**Figura XX**. Promedio de todos los países en Ausencia de corrupción.

A graph with a line

Description automatically generated

**Figura XX**. Promedio de todos los países en frecuencia de sobornos.

A graph with blue line

Description automatically generated

**Figura XX**. Promedio de todos los países en grado de respeto a las libertades.

A graph with a line

Description automatically generated

**Figura XX**. Promedio de todos los países en responsabilidad judicial.

A graph with blue line

Description automatically generated

**Figura XX**. Promedio de todos los países en igualdad educacional.

A graph with a line

Description automatically generated

**Figura XX**. Promedio de todos los países en igualdad en salubridad.

A graph of different colored lines

Description automatically generated

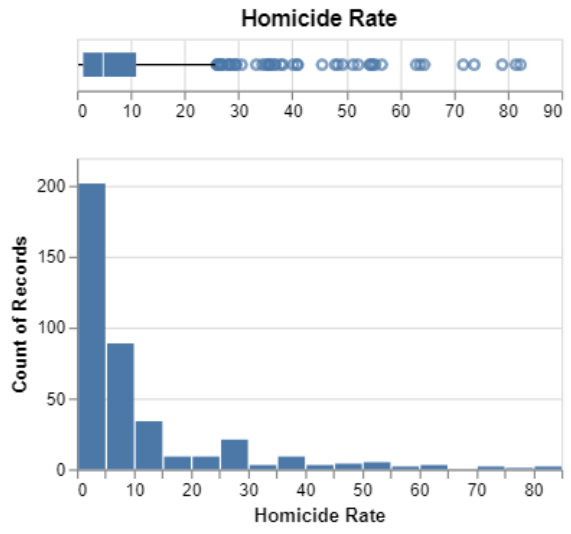
**Figura XX.** Grado de desempeño de indicadores selectos

Los países en general más desarrollados como Gran Bretaña logran calificaciones que rondan los 0.8 puntos, mientras que otros como Marruecos promedian el 0.4 de eficiencia. ¿Puede ser esto indicio de que los países más desarrollados emigren a España por cuestiones de placer o negocios mientras que los países con menor desarrollo lo hagan por cuestiones humanitarias o por persecución?

#### 5.4.7 Tasa de Homicidios

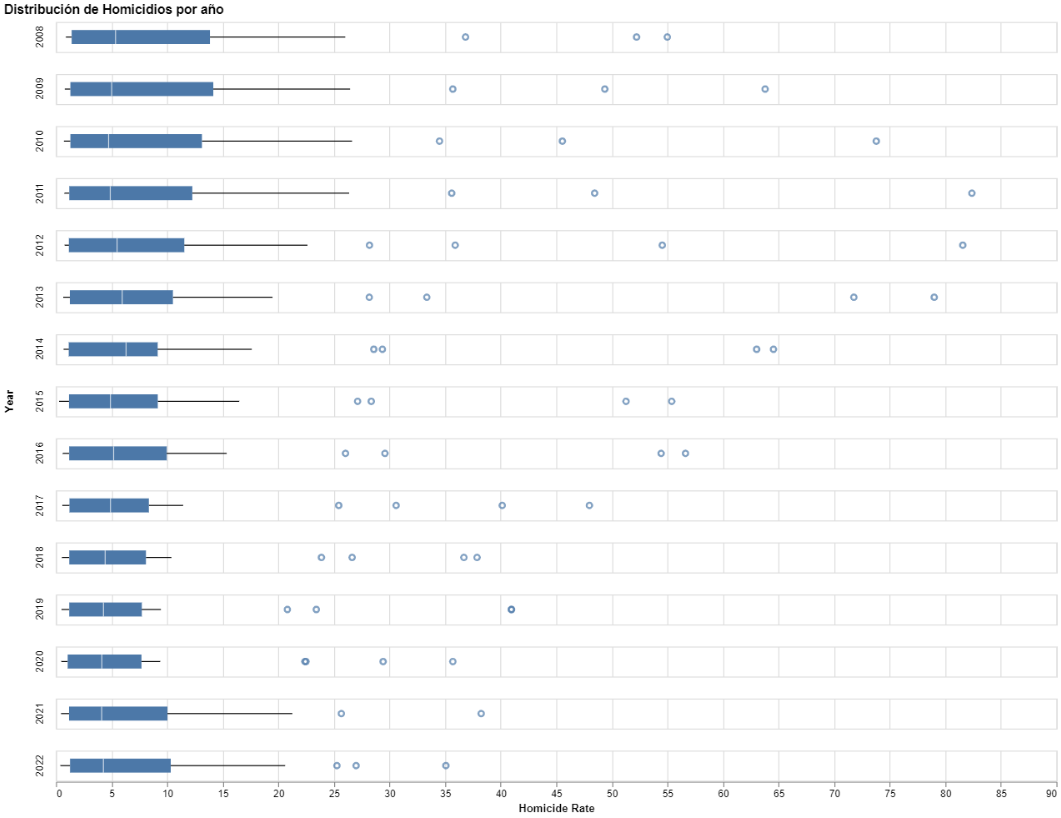
Al igual que en las variables anteriores, los datos de tasa de homicidio por cada 100.000 habitantes por país y añox, y se limpiaron y filtraron para el top nacionalidades y el período 2008 – 2022. Cabe destacar que los datos de Reino Unido estaban informados en tres regiones (Irlanda del Norte, Escocia e Inglaterra-Gales), por lo que el valor final para Reino Unido fue la media de las tres por año.

Posteriormente, se observó mediante un histograma que alrededor del 80% de los datos es inferior a 10 homicidios por 100.000 habitantes (Figura X), y en el 20% restante hay una gran variabilidad. Además, los datos atípicos pertenecen a algunos países de América del Sur, especialmente Venezuela y Honduras.



**Figura X**. Distribución de datos de Tasa de Homicidios y detección de datos atípicos.

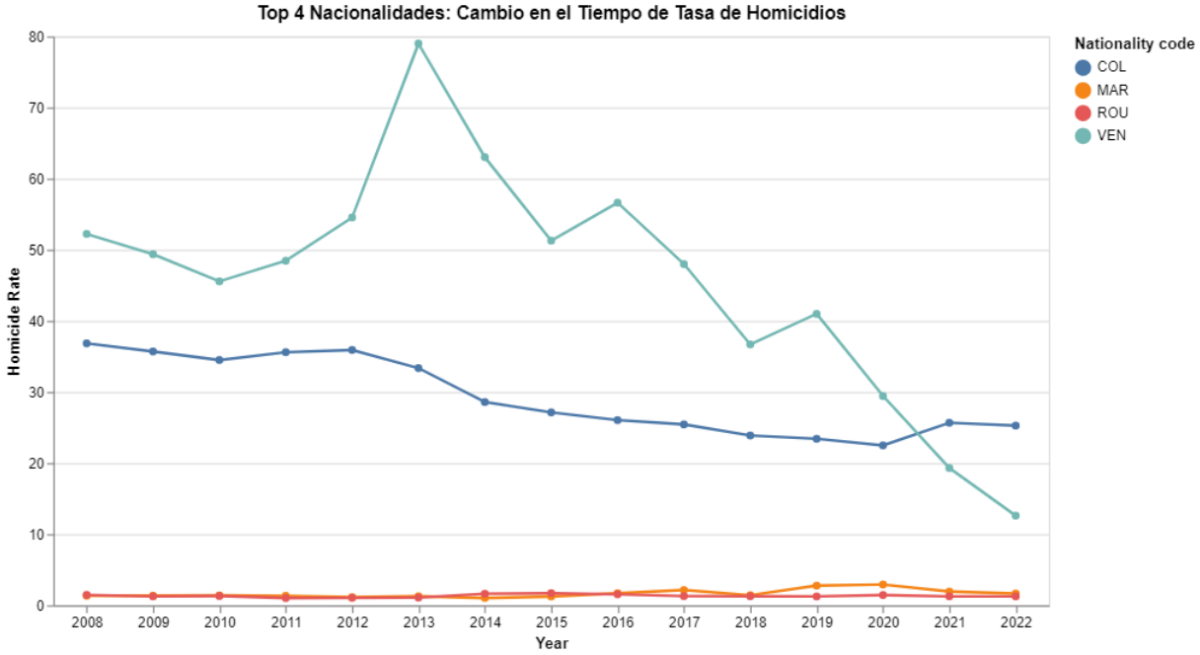
Al analizar la distribución por año, se observó que la distribución de la tasa de homicidios es muy similar entre los años 2008 – 2011 (Figura X) y que partir del 2012 tiende a bajar hasta alcanzar su menor rango en el 2020, año del inicio de la pandemia del COVID-19. Asimismo, se identifica con mayor facilidad que Venezuela, Colombia, y Honduras permanecen con valores tasa de homicidios atípicamente elevados en todo el período (no apreciable en la figura), añadiéndose Brasil a partir del 2012.

**Figura X**. Distribución de Tasa de Homicidios por año desde el 2008 a 2022 para el top de nacionalidades en inmigración en España.

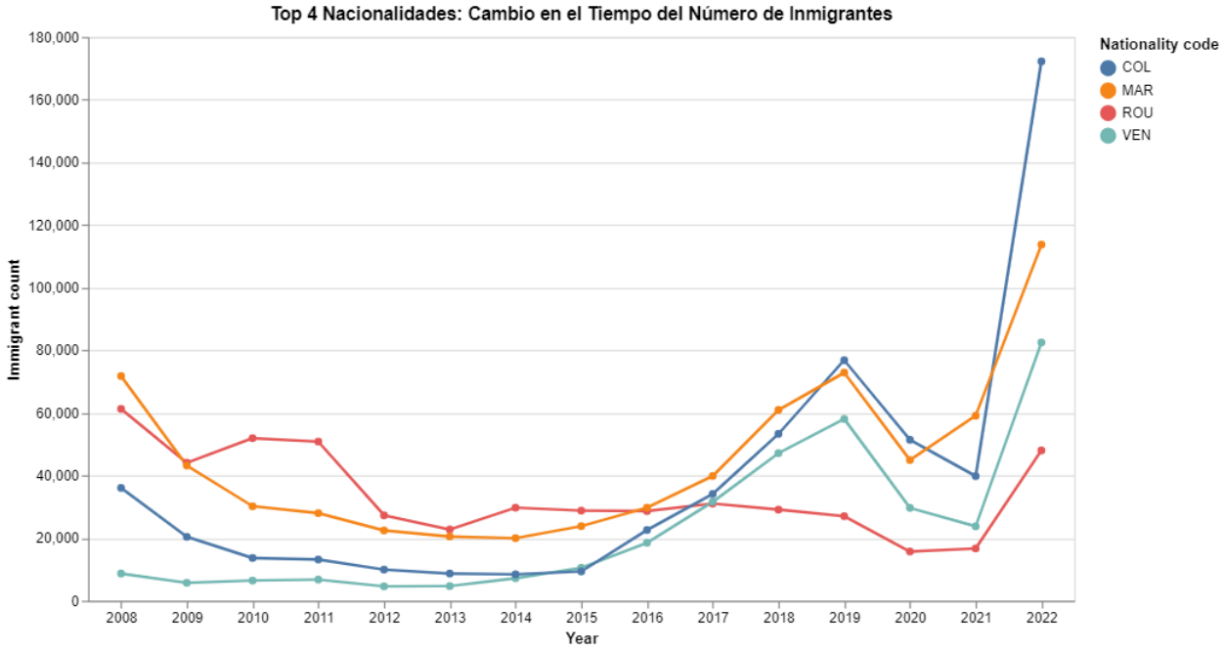
Finalmente, al comparar las tasas de homicidios del Top 4 en inmigración (Figura XA) en contraste con la cantidad de inmigrantes (Figuras XB) se observó que:

* No parece haber una relación clara en cuanto a tendencia entre la inmigración y homicidios para las Top 4 nacionalidades; pero sí notamos que Colombia y Venezuela, ambos de América del Sur, tienen tasas de homicidios elevadas durante amplios períodos de tiempo, especialmente entre 2008 - 2013.
* Marruecos y Rumania, ambos países mucho más próximos geográficamente a España (siendo Rumania parte de la Unión Europea), tienen una tasa de homicidios baja y muy similar entre ellos.

**(A)**



**(B)**



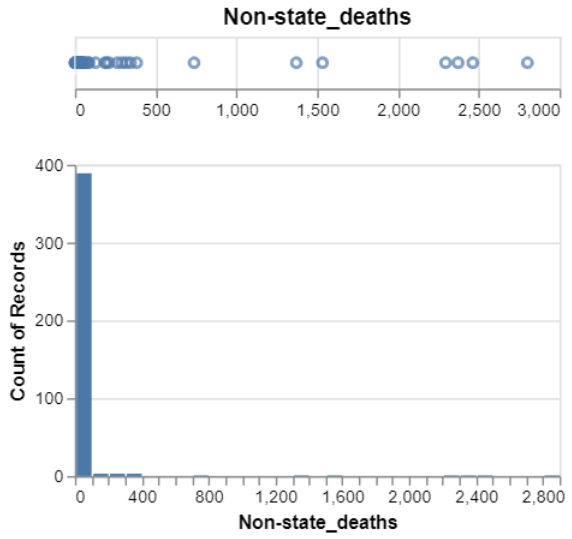
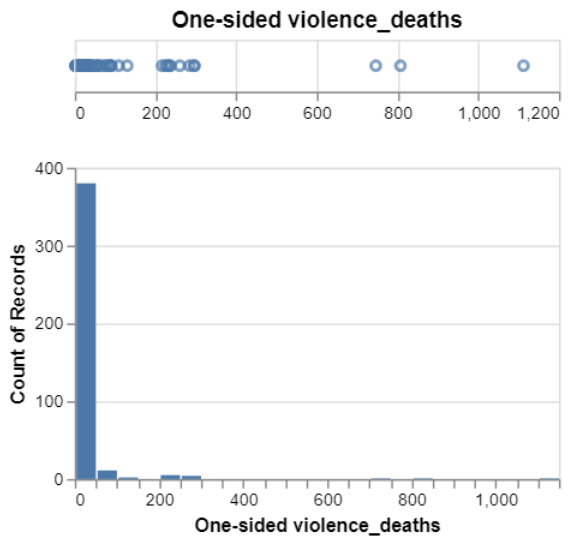
**Figura X**. Contraste entre la tendencia de cambio en el tiempo de la Tasa de Homicidios (A) frente a los datos de inmigración en España del Top 4 nacionalidades (B).

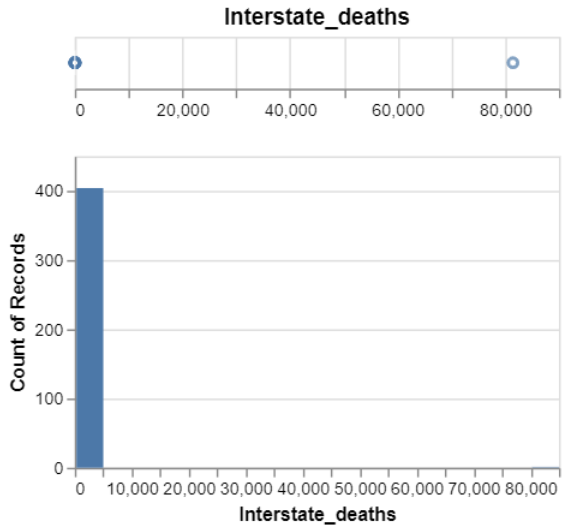
A partir del análisis, puede inferirse que la tasa de homicidios, en vez de tener un efecto inmediato, presenta un efecto acumulativo en el tiempo, en cuyo caso, será importante que los modelos de *machine learning* logren identificar relaciones entre variables, además de considerar las condiciones particulares de cada país bajo este complejo esquema de enfoque macro.

#### 5.4.8 Conflictos Armados

Continuando con las variables relacionadas con la inseguridad, se consideró también incluir y analizar las muertes debidas a cuatro tipos de conflictos armadosx (referirse a la Tabla 2 para las definiciones correspondientes).

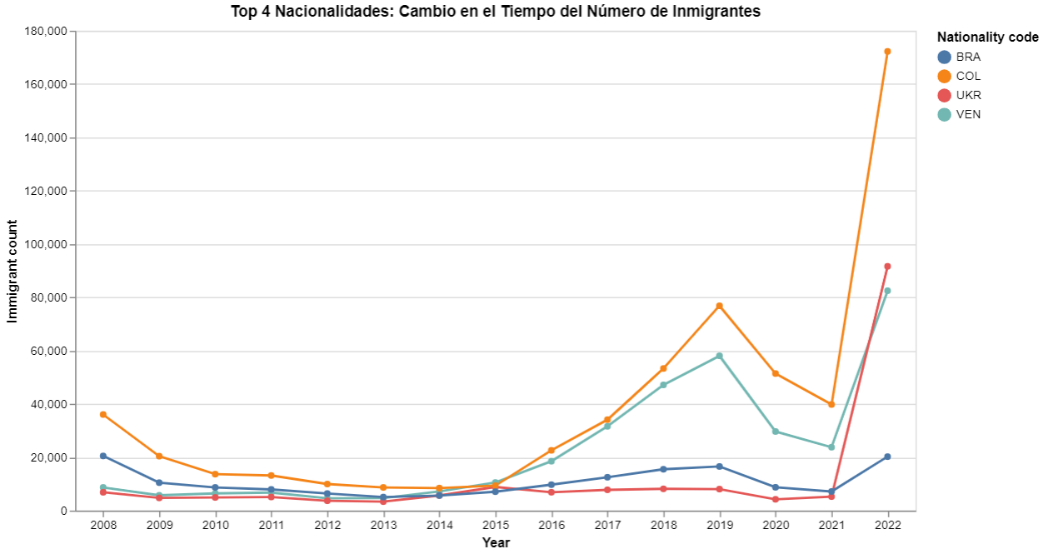
Después de filtrar los datos de nuestro interés, se observó que predomina la baja cantidad o ausencia de muertes dentro del top países en inmigración a España (Figura X), lo que explica porque los gráficos de caja no son visibles y considera como datos atípicos los registros de muertes para Ucrania, Pakistán, Colombia, Brasil y Venezuela (no apreciable en la figura), que son los países que se repiten dentro de los valores atípicos.

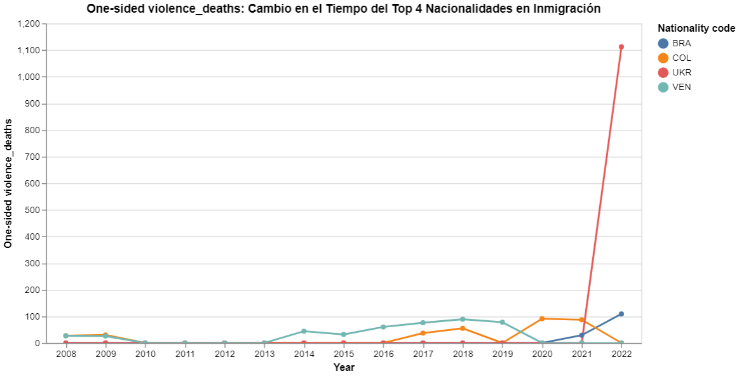
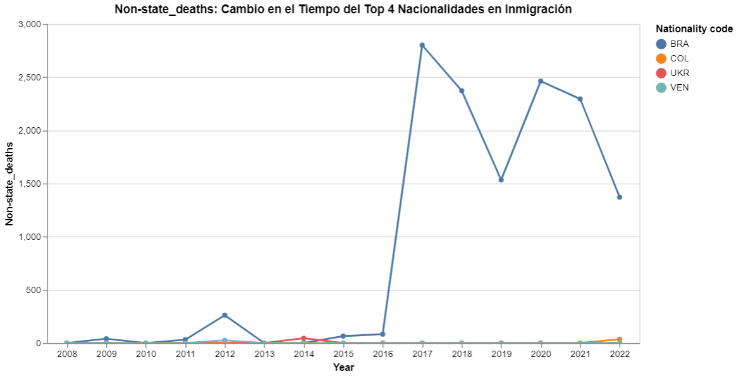


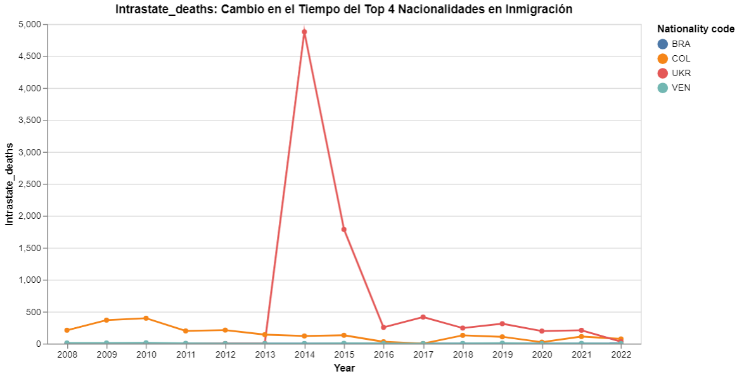
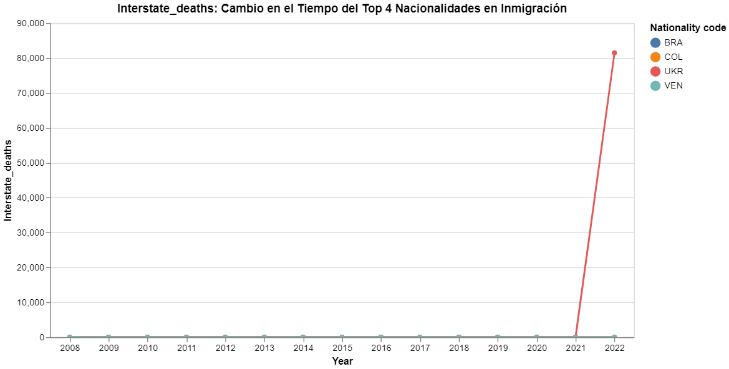
 

**Figura X.** Distribución las variables de conflictos armados y detección de datos atípicos. (A) "One-sided violence\_deaths", (B) “Non-state\_deaths”, (C) "Intrastate\_deaths" y (D) “Interstate\_deaths".

Debido a la dificultad para observar tendencias totales por la predominancia de algunos países, se filtraron los países más relevantes (Brasil, Colombia, Venezuela y Ucrania) para, nuevamente, contrastar la variación de los datos de inmigración con las muertes por estos conflictos con (Figura XB, XC, XD), observando que:



**Figura X**. Variación anual de la cantidad de inmigrantes en el Top 4 nacionalidades (A) en contraste con la variación de las muertes por conflictos armados: (B) "One-sided violence\_deaths", (C) “Non-state\_deaths”, (D) "Intrastate\_deaths" y (E) “Interstate\_deaths".

* Para Ucrania, hay una correlación en el pico de muertes por “One-sided violence\_deaths” (Figura XB) y “Interstate\_deaths” (Figura XE) en 2022, debido al conflicto con Rusia, y el incremento de inmigrantes.
* También vemos que, entre 2014 y 2019, los valores de "One-sided violence\_deaths” tienen un incremento para Colombia y Venezuela, al igual que hay un aumento progresivo de inmigración en ese mismo intervalo. Además, se aprecia algo similar para Brasil en relación a “Non-state\_deaths” (Figura XC) e inmigración a partir del 2016 (Figura XA).

Lo observado con estas cuatro nacionalidades, parece soportar la hipótesis que mientras mayores son las consecuencias de estos conflictos, mayor es su efecto en los desplazamientos migratorios, ya sea por razones humanitarias o nacionales buscando mejores condiciones de seguridad.

#### 5.4.9 Turistas Anuales

En comparación con los *dataframes* anteriores, el conjunto de datos de turistas tiene un volumen menor de datos. Estos datos representan un volumen de personas que llegan a España por medio de aeropuertos.

La finalidad de incluir estos datos es para introducir una variable explicativa que pudiera darnos un grado de magnitud a la hora de modelar.

A screenshot of a number

Description automatically generated

**Figura XX**. Cuenta de turistas anuales en España

Gráfico, Gráfico de líneas

Descripción generada automáticamente

**Figura XX**. Total de turistas que llegaron a España en avión por año.

Como otros indicadores, la presencia de turistas se ve disminuida en el año de 2020, misma que se recupera a partir de 2022.

### 5.5 Etapa 3: Unión y selección de variables

Las variables predictivas (Etapa 2) se añadieron al conjunto de datos de inmigrantes (Etapa 1) con un *merge* (Figura X), teniendo únicamente datos nulos para Senegal en la variable “Homicide Rate” ya que solo se encontraron datos para el año 2015.

Adicionalmente, se agregaron tres variables categóricas: i) una variable de idioma castellano para identificar los países que tienen al castellano como idioma oficial ('Spanish language' = 1), ii) otra para etiquetar a los años 2020 y 2021 por ser los años donde hubieron fuertes restricciones a nivel global debido a la pandemia del COVID 19 ("Restricciones\_pandemia" = 1) y iii) otra para etiquetar el año 2022 como el primer año con flexibilización de restricciones ("Año post\_pandemia" = 1).

Texto

Descripción generada automáticamente

**Figura X**. Código python para unir las variables predictoras al *dataset* “inmigrantes”con datos anuales de inmigración de España.

A partir de estos datos, se observó altas correlaciones entre las variables de desarrollo asociados a estado de derecho (“Regulatory Quality Percentil”, “Rule of Law Percentil” y “Voice and Accountability Percentil”), los índices de democracia y las variables del conjunto de datos relacionados a libertad (Figura X, observar notable la zona cálida en la matriz de correlación). En particular, si consideramos el valor 0.85 como valor de referencia para alta correlación de Pearson, vemos:

* + Alta correlación entre los índices de democracia, “Absence of Corruption”, "Voice and Accountability Percentile" y “Civil Liberties”. En consecuencia, se removió: “Deliberative democracy index”, “Absence of Corruption”, "Voice and Accountability Percentile" y “Civil Liberties”.
  + Alta correlación entre “Health equality” y “Educational equality”; se removió “Educational equality”.
  + Alta correlación entre “Regulatory Quality Percentile”, “Public sector corrupt exchanges” y “Rule of Law”, relacionadas con el sector público y la ley. Se removió “Regulatory Quality Percentile” y “Public sector corrupt exchanges”.
  + No hay correlación lineal con la variable de número de inmigrantes.

Continuando con las variables categóricas, se evaluó su significancia mediante un contraste de hipótesis usando pruebas no paramétricas debido a que ninguna de ellas seguía una distribución normal (Figura X, Figura XY). En consecuencia, se aplicó la prueba de U de Mann-Whitney para dos grupos y Kruskal-Wallis para aquellas variables con múltiples grupos.

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

**Figura X**. Matriz de correlación de Pearson entre las variables cuantitativas.

Calendario

Descripción generada automáticamente

Gráfico de dispersión

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

Aplicación, Tabla, Excel

Descripción generada automáticamente con confianza media

Imagen que contiene Aplicación

Descripción generada automáticamente

Gráfico de dispersión

Descripción generada automáticamente con confianza baja

**Figura X**. Distribución del número de inmigrantes mediante *boxplots* y detección de datos atípicos para los distintos grupos de cada variable categórica: (A) Sexo, (B) Idioma español, (C) Grupo de edad, (D) Régimen político, (E) Continente y (F) Subregión.

Imagen que contiene texto, tabla

Descripción generada automáticamente

**Figura X**. Resultados de la prueba de Shapiro-Wilk para normalidad. H0: Datos se ajustan a una distribución normal, H1: No se ajustan a una normal. Intervalo de confianza = 0.95.

Los resultados de las pruebas no paramétricas indican que hay diferencias significativas en las variables con dos grupos y, al menos, diferencias significativas entre dos grupos dentro de las variables con múltiples (Figura X).

Texto

Descripción generada automáticamente

**Figura X**. Resultados de las pruebas de U de Mann-Whitney y Kruskal-Wallis. H0: No hay diferencia (en términos de tendencia central) entre los grupos, H1: Existe una diferencia (con respecto a la tendencia central) entre los grupos. Intervalo de confianza = 0.95.

Adicionalmente… observaciones mas interesantes:

### 5.6 Etapa 4: Prueba y comparación de modelos de *machine learning*

### Análisis Exploratorio de Datos o EDA

El análisis exploratorio fue dividido en varios *notebooks* con la finalidad de evitar una confusión al introducir muchas variables en uno solo.

A manera de exposición, exponemos los gráficos más relevantes para comprender el tipo de análisis que ejecutamos, manteniendo el análisis completo en cada *notebook*.

A graph with a line and a blue line

Description automatically generated

Figura 1, total de inmigrantes en España de 2008 a 2022, elaboración propia

Entre 2008-2013 hay una caída de número de inmigrantes y a partir del 2014 aumenta progresivamente hasta alcanzar un pico de 750,480 inmigrantes en 2019, para luego disminuir en el 2020 y 2021 (período de pandemia y restricciones sanitarias relacionadas al COVID 19) y mostrar un gran salto hasta ≈1.25 millones de inmigrantes en el 2022 (post-pandemia), superando al 2019 en aproximadamente 500,000 inmigrantes.

Observando los inmigrantes que llegaron a España durante estos 15 años según la edad, vemos que predomina la inmigración de personas jóvenes de entre 25-34 años, seguidos de jóvenes de entre 15-24. A partir de ese punto, es notable la disminución progresiva del número de inmigrantes con el incremento de la edad.

En cuanto al sexo, vemos que la cantidad de mujeres y hombres es similar en todos los grupos, especialmente en los grupos de mayor presencia (15-24 y 25-34 años). Únicamente los grupos de entre 0-14 y 55-64 años muestran una mayor diferencia, el primero hacia los hombres y el segundo hacia las mujeres.

La distribución por año parece indicar que a medida que hay menor inmigración, la dispersión de los datos es menor y los valores atípicos menos extremos en relación con el rango intercuartílico (caja), mientras que cuando hay aumenta la inmigración en el tiempo, la dispersión también aumenta.

Además, vemos que la mayoría de los datos se concentran en un rango medio/bajo de número de inmigrantes (observar rango intercuartílico entre 2008-2021 que se mantiene alrededor o por debajo de 10,000) y sólo algunas nacionalidades presentan valores "atípicamente" superiores al resto. Y son precisamente estas nacionalidades con grandes valores atípicos los que son de nuestro interés de estudio y modelado, pues en ellas se condensan la mayor masa de inmigrantes hacia España. Más aún, es interesante observar cómo algunas nacionalidades mantienen un número atípicamente alto de inmigración a lo largo de todo el período 2008-2022, como son españoles, colombianos, rumanos y marroquíes. A graph of blue rectangular objects

Description automatically generated with medium confidence

Figura 4, porcentaje por total de inmigrantes en España por país de origen, elaboración propia

Previamente, ya habíamos notado la cantidad de inmigrantes españoles, colombianos, marroquíes y rumanos, y ahora lo observamos también en su distribución porcentual en base al total.

Luego de los top 4, parece variación porcentual menor entre las subsiguientes nacionalidades.

A graph of blue rectangular bars with white text

Description automatically generated

Figura 5, desempeño por indicadores de libertad para Argentina de 2008 a 2022 en promedio, elaboración propia

Para los índices de libertad, obtuvimos gráficas por cada país para conocer el desempeño por cada uno de los indicadores. En el caso de Argentina, observamos que las libertades civiles alcanzan un grado de plenitud (donde 1 significa certeza y 0 ausencia). En general, los países seleccionados obtienen un promedio de desempeño alrededor del 0.5 de todos los indicadores.

A graph with a line

Description automatically generated

Figura 6, desempeño global en indicador de igualdad en salubridad de 2008 a 2022, elaboración propia

Este indicador es particularmente interesante dado el contexto de la pandemia del COVID. El indicador responde a la accesibilidad de los ciudadanos a servicios de salud de alta calidad, para permitirles ejercer sus derechos políticos en la adultez. En franco declive desde 2008, la situación se ve un poco acelerada para 2019, tocando fondo en 2020 y teniendo un rebote. Aunque se aprecia un esfuerzo global por mejorar los servicios de salud en los países, nuevamente se pierde el avance para 2022.

A graph of a graph of a graph

Description automatically generated with medium confidence

Para efectos del estudio, podemos afirmar que los migrantes en España provienen de un trasfondo diverso, muchos marcados por carencias palpables y otros con un nivel de vida alto. Las gráficas anteriores muestran el ejemplo concreto para Pakistán mismo que tiene altibajos de un indicador a otro.

# Conclusiones

# Referencias bibliográficas

1. *Iman, H. S., & Tarasyev, A. (2018). MACHINE LEARNING METHODS IN INDIVIDUAL MIGRATION BEHAVIOR.* Disponible en:<https://elar.urfu.ru/bitstream/10995/68086/1/978-5-8295-0581-3_2018-08.pdf>
2. Micevska, M. (2021). *Revisiting forced migration: A machine learning perspective*. European Journal of Political Economy, 70, 102044. doi:10.1016/j.ejpoleco.2021.102044
3. International Organization for Migration. *Factores que propician la migración internacional*. Año 2011. Recuperado de <https://emm.iom.int/es/handbooks/contexto-global-de-la-migracion-internacional/factores-que-propician-la-migracion> el 19 de septiembre 2024.
4. Organización Internacional de Migraciones (OIM). *Informe sobre las Migraciones en el Mundo 2022*. Geneva. Disponible en: <https://publications.iom.int/books/informe-sobre-las-migraciones-en-el-mundo-2022>
5. Banco Europeo de Inversiones (BEI). *Migración y las naciones europeas. Retos, oportunidades y el papel del BEI*. Marzo 2016. Disponible en: <https://publications.iom.int/books/informe-sobre-las-migraciones-en-el-mundo-2022>
6. CLIMB Database: Human Mobility in the Context of Disasters, Climate Change and Environmental Degradation Database. Red de las Naciones Unidas. Disponible en: <https://migrationnetwork.un.org/climb#accordion>
7. Instituto Nacional de Estadística (INE). *Estadística de Migraciones y Cambios de Residencia (EMCR). Año 2022*. Nota de Prensa. Disponible en: <https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177098&menu=ultiDatos&idp=1254735573002>
8. AYDEMİR Belgin, AYDIN Hakan, ÇETİNKAYA Ali y POLAT Doğan Şafak. *Predicting the Income Groups and Number of Immigrants by Using Machine Learning* *(ML).* International Journal of Multidisciplinary Studies and Innovative Technologies. 2022. 6 (2), 162-168.
9. CARAMMIA Marcello, MARIA IACUS Stefano y WILKIN Teddy*. Forecasting asylum‑related migration flows with machine learning and data at scale*. Nature, Scientific Reports. 2022. 12:1457.
10. Imán H.S y Tarasyev, A (2018)10
11. Micevska, M. (2021)11
12. Instituto Nacional de Estadística (INE). *Estadística de Migraciones y Cambios de Residencia (EMCR). Año 2022*. Nota de Prensa. Disponible en: <https://www.ine.es/dynt3/inebase/index.htm?padre=3694&capsel=1963>
13. ISO <https://www.iso.org/obp/ui/#search>
14. INE Continentes y Sub-regiones <https://www.ine.es/daco/daco42/clasificaciones/paisesyterritorios.xls>
15. Índices de desarrollo - -

<https://databank.worldbank.org/GDP---Inflation---Country-Risk/id/ae8fd58a>#

<https://www.worldbank.org/en/research/brief/inflation-database>

<https://www.imf.org/external/datamapper/NGDP_RPCH@WEO/OEMDC/ADVEC/WEOWORLD/VEN>

1. Regimen político - Skaaning et al. (2023) <https://ourworldindata.org/grapher/political-regime-lexical?tab=table>
2. Indices democracia -

<https://ourworldindata.org/explorers/democracy?tab=table&facet=none&Dataset=Varieties+of+Democracy&Metric=Deliberative+democracy&Sub-metric=Main+index&country=ARG~AUS~BWA~CHN~OWID_WRL>

1. *Global State of Democracy (GSoD)*x. / <https://www.idea.int/theme/democracy-assessment>

<https://prosperitydata360.worldbank.org/en/home>

1. Tasa homicidios - <https://dataunodc.un.org/dp-intentional-homicide-victims> \* y datos adicionales en: <https://datosmacro.expansion.com/demografia/homicidios>
2. Conflictos armados - Link: <https://ourworldindata.org/war-and-peace> bajo la autoria de Bastian Herre, Lucas Rodés-Guirao, Max Roser, Joe Hasell and Bobbie Macdonald (2024)\*

# Anexos

# 

# Anexo 1: Tabla de Variables

| **Dimensión** | **Conjunto de datos** | **Comentario / Descripción** | **Variables** | **Fuente** | **Link(s)** |
| --- | --- | --- | --- | --- | --- |
| Socio-Económica | Índices de Desarrollo | Índices socioeconómicos con los que se busca considerar y analizar el impacto de la condición de vida de los países sobre el incremento o disminución de inmigrantes hacia España. | 1) Porcentaje de Desempleo 2) Percentil - Política y Violencia 3) Probabilidad de morir joven 4) Percentil - Calidad Regulatoria 5) Percentil - Estado de Derecho 6) Percentil - Representatividad y Participación Ciudadana 7) Porcentaje de trabajadores asalariados 8) Crecimiento de PBI (anual) 9) Inflación - deflactor del PBI (anual) | Banco Mundial y Fondo Monetario Internacional | https://databank.worldbank.org/GDP---Inflation---Country-Risk/id/ae8fd58a#  https://www.worldbank.org/en/research/brief/inflation-database  https://www.imf.org/external/datamapper/NGDP\_RPCH@WEO/OEMDC/ADVEC/WEOWORLD/VEN |
| Socio-Política | Índices de Democracia | Índices de democracia basado en las estimaciones de expertos y el índice de V-Dem. Con ellas se busca capturar el efecto de depreciación o mejoría de los derechos democráticos en los países sobre la inmigración. | 1) Índice de democracia liberal 2) Índice de democracia deliberativa | V-Dem (2024) | [https://ourworldindata.org/explorers/democracy?tab=table&facet=none&Dataset=Varieties+of+Democracy&Metric=Deliberative+democracy&Sub-metric=Main+index&country=ARG~AUS~BWA~CHN~OWID\_WRL\*](about:blank) |
| Índices GSOD | Indicadores (*Global State of Democracy Indices - GSoD Índice*s) que miden la tendencia democrática de un país en base a distintas categorías | 1) Ausencia de Corrupción 2) Libertades Civiles 3) Derechos civiles 4) Calidad de educación 5) Igualdad de oportunidades 6) Equidad en salud (acceso) 7) Rendición de cuentas judiciales 8) Intercambios corruptos en el sector público | Banco Mundial / IDEA Internacional | [https://prosperitydata360.worldbank.org/en/home   https://www.idea.int/democracytracker/gsod-indices](about:blank) |
| Política | Régimen Político | Identifica el régimen político de los países dentro de top nacionalidades en inmigración a España y se distingue entre: autocracias no electorales (0), autocracias unipartidistas (1), autocracias multipartidistas sin ejecutivo electo (2), autocracias multipartidistas (3), democracias excluyentes (4), democracias masculinas (5), democracias electorales (6) y poliarquías (7) | Régimen Político | Skaaning et al. (2023) | [https://ourworldindata.org/grapher/political-regime-lexical?tab=table](about:blank) |
| Seguridad | Tasa de Homicidios | Tasa de homicidios por cada 100,000 habitantes por países/año. | Tasa de homicidios por cada 100,000 habitantes | Organización de las Naciones Unida, y otros | [https://dataunodc.un.org/dp-intentional-homicide-victims  https://datosmacro.expansion.com/demografia/homicidios](about:blank) |
| Conflictos Armados | Los conflictos armados reflejan la cantidad de muertes según cuatro tipo de conflictos: *One-sided violence*, *Non-state*, *Intrastate* e *Interstate*.  Con estas variables se incorporar incorporar efecto de los conflictos de tipo bélico en el incremento de movimientos migratorios | *1) One-sided violence\_deaths 2) Non-state\_deaths 3) Intrastate\_deaths 4) Interstate\_deaths* | Bastian Herre, Lucas Rodés-Guirao, Max Roser, Joe Hasell and Bobbie Macdonald (2024) | [https://ourworldindata.org/war-and-peace](about:blank) |
| Geográfica | Continentes y Subregiones | Agrupación oficial de países usada por el Instituto Nacional de Estadística (INE) de España y la Eurostats | 1) Continente 2) Subregión (ej. A. del Sur, UE, Caribe) | Instituto Nacional de Estadística de España | [https://www.ine.es/daco/daco42/clasificaciones/paisesyterritorios.xls](about:blank) |
| Socio-cultural | Idioma Castellano | Variable categórica (1/0) con los que se especifica si el idioma castellano es o no el idioma oficial del país.  Con esta variable se busca incluir el efecto del dominio del mismo idioma hablado en España sobre la decisión preferente de este destino por los inmigrantes. | Idioma Castellano | Elaboración propia | N/A |
| Salud | Pandemia | Variable categórica (1/0) para la presencia de restricciones sanitarias debido a la pandemia de COVID (2020-2021)  Con esta variable se busca incluir el efecto de las presencia de restricciones sanitarias en la variación del número de inmigrantes | Restricciones por Pandemia | Elaboración propia | N/A |
| Post-Pandemia | Variable categórica (1/0) para indicar el primer año después de las restricciones sanitarias del COVID (2022)  Con esta variable se busca considerar la acumulación de inmigrantes durante las restricciones sanitarias entre 2020-2021, los cuáles no lograron viajar a España, y así incluir la particularidad del año 2022 al ser el primer año después de la flexibilización casi total de las restricciones. | Año Post-Pandemia | Elaboración propia | N/A |
| Turismo | Turistas Anuales | Cantidad de Turistas extranjeros que llegaron a España en Avión.  Con esta variable buscamos hacer una relación de inmigrantes-por-turista para proveerle al modelo una variable de magnitud para el efecto de la pandemia | Turistas Anuales | Statista | [https://es.statista.com/estadisticas/474883/llegadas-de-turistas-extranjeros-a-espana-en-avion/](about:blank) |