Laboratorio de Datos: Migración y Representación Argentina

Grupo: *inlineSQL*		
Apellido	Nombre	LU
Wencelblat	Agustin	878/23

1. Resumen:

A lo largo de este informe, buscamos investigar si existe algún tipo de relación entre flujos migratorios de argentinos hacia diversos destinos y la cantidad de entidades de representación que nuestro país mantiene allí. A modo de ejemplo, esta información puede resultar útil a la hora de determinar si existen destinos populares entre migrantes argentinos que carecen del nivel de representación y apoyo necesario, o viceversa. Tras limpiar y analizar las bases de datos pertenecientes a entidades de representacion argentinas y migracion mundial, analizamos su informacion haciendo uso de reportes y visualizaciones. Mediante esto llegamos a la conclusion que existe una al menos una corrrelación entre el flujo migratorio de países extranjeros hacia Argentina y las entidades de representacion presentes en el pais destino.

2. Introduccion:

En un mundo cada vez más globalizado, los movimientos migratorios juegan un papel clave en las relaciones internacionales. Argentina, como muchos otros países, ha experimentado flujos migratorios importantes a lo largo del tiempo, con ciudadanos buscando nuevas oportunidades en diferentes regiones del mundo. A la par de estos movimientos, el gobierno argentino ha desarrollado una red de representaciones diplomáticas y consulares en el exterior, con el fin de atender las necesidades de sus ciudadanos y fortalecer sus vínculos con otros países. Estas representaciones varían en cantidad y distribución según el país de destino, lo que nos lleva a cuestionarnos si la magnitud del flujo migratorio influye en dicha distribución.

En este informe, nos proponemos analizar la posible relación entre los flujos migratorios de argentinos hacia distintos países y la cantidad de entidades de representación que la República Argentina mantiene en esos destinos. Partimos de la hipótesis de que existe una correlación directa entre ambos factores: a mayor flujo migratorio de argentinos hacia un país, mayor será la cantidad de entidades de representación establecidas en dicho territorio. Por el contrario, en aquellos países donde los flujos migratorios desde Argentina son menores, anticipamos una menor presencia de entidades diplomáticas o consulares.

Para probar esta hipótesis, utilizaremos datos provenientes de cuatro tablas que contienen información sobre los flujos migratorios globales y las representaciones gubernamentales argentinas en el exterior. A través del análisis de estos datos, buscaremos determinar si la cantidad de argentinos que emigran a un país influye directamente en la cantidad de entidades de representación establecidas allí.

3. Procesamiento de Datos:

3.1 Introduccion y Formas Normales:

A primera vista, las tablas no están en ninguna forma normal, pues no podemos conocer las dependencias ni las claves con tan solo mirar las tablas. Sin embargo, es posible operar bajo algunas suposiciones para determinar en que forma normal esta cada una de las tablas de representacion argentina:

- lista-sedes:

    - Podemos ver que cumple 1FN pues no tiene valores multivaluados y todos los valores son atómicos,
    - Si tomamos sede_id como clave primaria puede ser 2FN (aunque hay atributos como pais_iso_2, pais_iso_3 que podrian depender de pais_castellano o pais_ingles).
    - Cumplen también con 3FN si elegimos las dependencias funcionales tal que ninguna de ellas sea transitiva.
- lista-secciones:
    - Este caso es casi identico al anterior, cumple 1FN pues no tiene valores multivaluados ni valores que no sean atomicos.
    - Para que cumplan 2FN y 3FN, tomamos a sede_id combinada con tipo_seccion como clave primaria.
- lista-sedes-datos:
    - No cumple con 1FN pues la columna “redes_sociales” tiene atributos multivaluados, pues hay mas de 1 link en algunas celdas (separados con “//”).
    - Luego, al no cumplir 1FN, no es 2FN ni 3FN.

3.2 Calidad y Procesamiento de Datos:

Haciendo uso de la técnica GQM, buscamos mejorar la calidad de los datos para poder realizar nuestro informe apropiadamente. Es importante notar que algunos de estos problemas de calidad de datos no resultan relevantes para nuestro analisis. Sin embargo, decidimos incluirlos A continuación documentamos algunos de los problemas con la calidad de cada una de las bases de datos y que hicimos para mitigar su efecto en nuestro análisis.

- datos_migraciones:

- Notamos que existen muchas filas con valores “..” en los números migratorios, que interpretamos como nulos (es decir, no hay data). Este problema afecta a la completitud de los datos y se corresponde a la instancia, pues esos datos no han podido ser recolectados. Utilizando el método GQM, planteamos lo siguiente: como objetivo, queremos que todos los datos pertenecientes a los flujos migratorios estén completos, es decir: ¿Cuál es la proporción de celdas pertenecientes a datos de flujo migratorio con el valor “..”? (que interpretamos como nulo). Para eso, calculamos: cantidad de celdas con “..” como valor *100 / cantidad de celdas correspondientes a flujo migratorio, que resulta ser 37740*100/803880 = 4.695%. Así, vemos que falta casi un 5% de los datos de migración, lo cual es algo que no podemos corregir, pues no tenemos forma de recopilar toda esta información.

- lista-secciones:

- Hay filas en las cuales la columna “correo electrónico” está vacía, lo cual podría ocasionar problemas pues es el medio de contacto más accesible. Este problema también afecta a la completitud de los datos, y está asociado a la instancia. Utilizando la técnica GQM, planteamos: como objetivo, que queremos que todos los correos electrónicos estén completos, es decir ¿Cuál es la proporción de celdas pertenecientes a correo electrónico que están vacías?. Así, calculamos: cantidad de celdas vacías *

100/cantidad de celdas de correo electrónico, que resulta ser 14*100/516 = 2.702%. Al igual que anteriormente, puesto que los datos están incompletos, nos es imposible volver a recopilar esa información.

- Por otro lado, podemos notar que muchos de los horarios de atención están ubicados en la columna “comentario_del_horario” en vez de las 2 columnas correspondientes (“atencion_hora_desde” y “atencion_hora_hasta”). Este es un problema que afecta la disponibilidad/accesibilidad de los datos y su completitud, ya que están ordenados de forma inconsistente e inesperada, y es un problema asociado al modelo, pues la información fue recopilada incorrectamente. Usando la técnica GQM, planteamos lo siguiente: Queremos que los horarios de atención están presentados en las columnas correspondientes (“atencion_dia_desde”, “atencion_dia_hasta”, “atencion_hora_desde” y “atencion_hora_hasta”), es decir, ¿Qué proporción de datos correspondientes tienen el horario cargado de forma “incorrecta”? (Consideramos que este cargado de forma incorrecta si alguna de las 4 casillas mencionadas anteriormente está vacía) Notar que excluimos las filas en las que el horario está presente en las 4 columnas correctas y en el comentario a la vez, pues en ese caso el único problema es de redundancia. Así, calculamos: filas en las que falta 1 de las 4 columnas mencionadas anteriormente * 100 / cantidad de filas = 254*100/516 = 49.225% Así, vemos que en casi la mitad de los casos, la información está o mal cargada (según el estándar que establecimos anteriormente) o ausente. Para corregir esto, utilizamos los valores presentes en la columna “comentario_del_horario” para completar las otras columnas, llevando el valor a 228*100/516 = 44.186%, que son solo las filas en las cuales falta alguna parte de la información del horario de atención, ya sea dentro de los comentarios o las 4 columnas apropiadas.

- lista-sedes-datos:
  - Podemos notar que la fila 16 está “corrida” una celda hacia la derecha a partir de la columna “codigo postal”. Esto parece ser un problema de procesos, ya que puede haber sido ocasionada por la carga incorrecta de datos, o alguna modificación accidental. Esto afecta la validez de los datos, pues hay información absurda en algunas columnas, como por ejemplo “Viernes” en “atencion_hora_desde”. Aunque podría resultar innecesario, usando el método GQM, planteamos lo siguiente: Queremos que la información de las filas esté en las celdas correspondientes, es decir ¿Qué proporción de filas tiene la información en celdas incorrectas? Así, calculamos: filas con corrimientos *100 / filas totales = 1*100/164 = 0.610%. Para arreglar este problema, decidimos incluir “Edificio Centro Plaza"”, que es el valor presente en la columna “codigo_postal” en la columna “direccion” y mover todos los valores hacia la derecha en esa fila una celda hacia la izquierda. Esto lleva nuestra métrica a 0% de las filas afectadas por información en celdas incorrectas, pues la fila 16 era la única afectada.
  - La columna “codigo postal” tiene valores nulos distintos como por ejemplo . “0”, “no existe”, vacío o “-----”. Este problema afecta la consistencia/vigencia de los datos y está asociado a la instancia, pues son datos cargados de forma inconsistente. Utilizando la técnica GQM, planteamos lo siguiente: Queremos unificar los valores nulos a un solo

valor, el “--”, por lo que preguntamos: ¿Qué proporción de celdas en la columna “código postal” tiene un valor nulo distinto a ”--”? Así, calculamos: número de celdas con valor nulo distinto a “--” en “código postal” *100 / número de celdas con valores nulos en “codigo postal” = 16*100/19 = 84.211%. Luego, unificamos el valor nulo al “--” llevando este número a 0*100/19 = 0%.

- Al igual que con “lista_secciones”, vemos que muchos de los horarios de atención están ubicados en la columna “comentario_del_horario” en vez de las 2 columnas correspondientes (“atencion_hora_desde” y “atencion_hora_hasta”). Como vimos anteriormente, es un problema que afecta la disponibilidad/accesibilidad de los datos y su completitud, ya que están ordenados de forma inconsistente e inesperada, y es un problema asociado al modelo, pues la información fue recopilada incorrectamente. Usando el método GQM, planteamos: Queremos que los horarios de atención estén presentados en las columnas correspondientes, es decir, ¿Qué proporción de datos correspondientes tienen el horario cargado de forma “incorrecta”?. Así, calculamos: filas en las que falta 1 de las 4 columnas mencionadas anteriormente * 100 / cantidad de filas = 69*100/164 = 42.073%. Así, vemos que en esta proporción de casos, la información está o mal cargada (según el estándar que establecimos anteriormente) o ausente. Para corregir esto, utilizamos los valores presentes en la columna “comentario_del_horario” para completar las otras columnas, llevando el valor a 46*100/164 = 28.049%, que son solo las filas en las cuales falta alguna parte de la información del horario de atención.

- lista-sedes:

- Podemos notar que los valores de la columna “país castellano” no están estandarizados, la gran mayoría están en mayúscula pero algunos en minúscula, y algunas excepciones en otros idiomas. Este problema afecta la consistencia de los datos y está asociado a la instancia, pues suponemos que los datos fueron recopilados de distintas fuentes sin un criterio unificado. Utilizando la técnica GQM, planteamos lo siguiente: Queremos unificar todos los valores en la columna “pais castellano” a mayúscula y palabras del abecedario, preguntando ¿Qué proporción de celdas en la columna “pais castellano no estan en mayuscula o en el abecedario? Tomamos la siguiente métrica: número de celdas en la columna “pais castellano” que no estan en mayuscula o en el abecedario *100 / número de celdas la columna “pais castellano” y calculamos 16*100/164 = 9.756%. Luego, unificamos todos los valores a mayúsculas en el abecedario, reduciendo este número a 0*100/164 = 0%.

3.3 Modelado Entidad Relacion:

Como herramienta para organizar y formalizar nuestros datos, construimos el siguiente diagrama entidad relación. Notar que las motivaciones detrás de nuestras elecciones de atributos están en la sección 4.2.

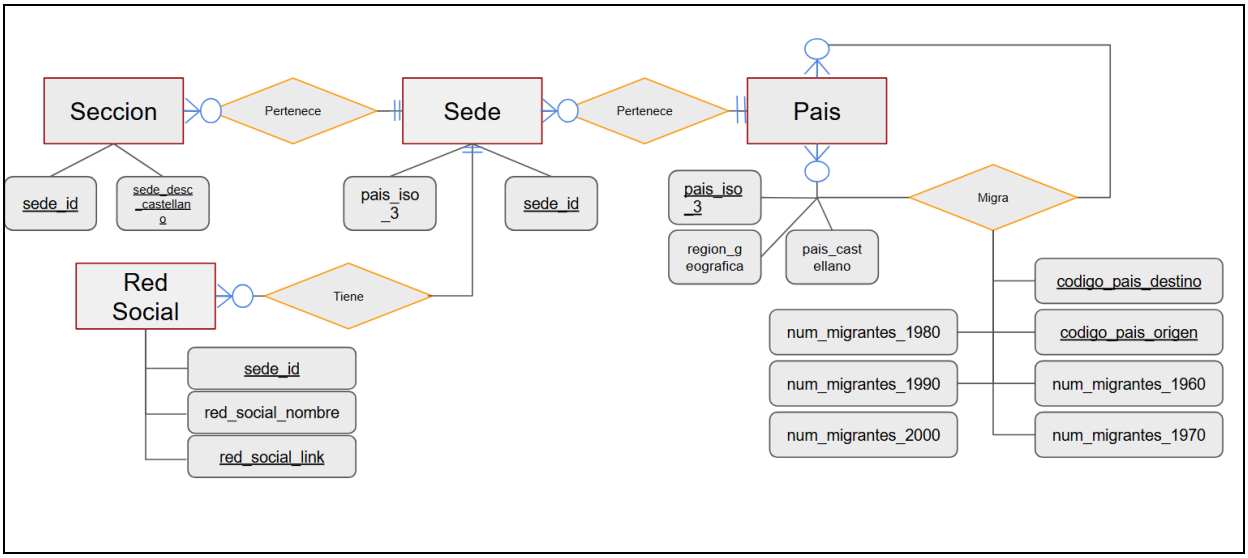

Figura 1: Diagrama Entidad Relación para el modelado de flujos migratorios y representación.

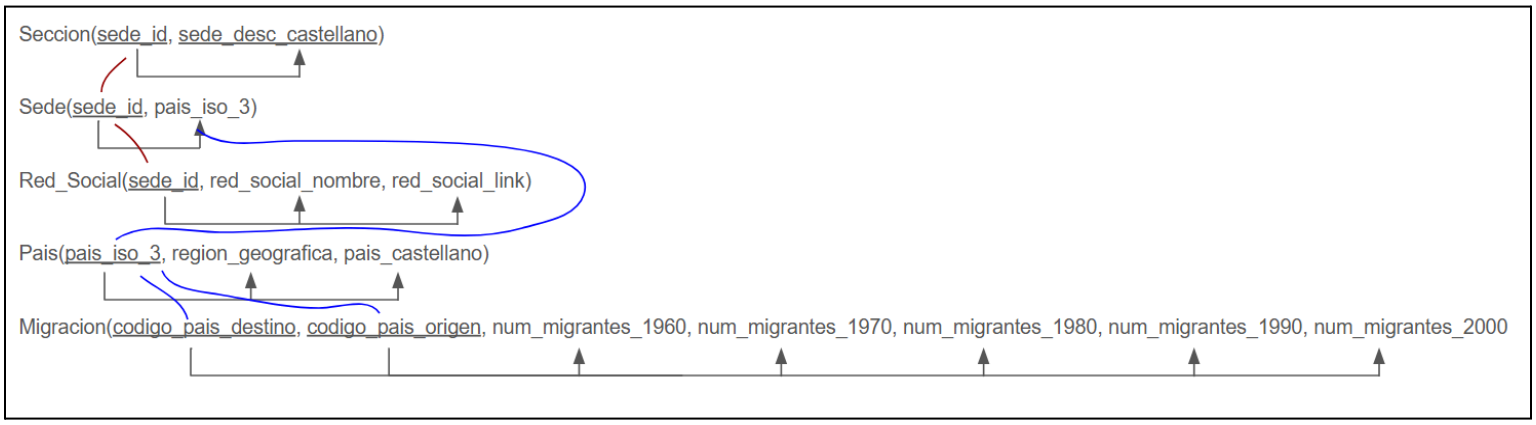

Figura 2: Modelo Entidad-Relacion con las dependencias funcionales y foreign keys en color.

Luego, pasamos estos datos a 4 archivos, uno por cada entidad. Para seccion, utilizamos los datos de lista-secciones, para sede y pais las de lista-sedes-datos y finalmente para migracion, importamos la información de datos_migraciones. El pasaje de datos en sí a estas nuevas tablas puede encontrarse en el archivo .py adjuntado.

4. Decisiones Tomadas:

4.1 Calidad de Datos:

Aparte de las decisiones mencionadas en la sección 3.2, tomamos otras medidas para la estandarización y limpieza de nuestros datos. Notar que los archivos limpios estan ubicados en la carpeta “Tablas Limpias”:

- lista-secciones:

    - En la columna “sitio_web” hay direcciones de correo escritas, y algunos sitios cuyo url está mal escrito, movemos las direcciones de correo a “correos_electronicos_adicionales” y arreglamos los urls que no funcionan correctamente.
    - En la columna “nombre_titular” no hay un valor nulo unificado, pues hay algunas celdas con valor “0” o “test”, por lo cual unificamos todas a la celda vacía.
    - En las columnas “nombre_titular” y “apellido_titular” dejamos todos los nombres escritos con la primera letra mayúscula y el resto minúscula.
- lista-sedes-datos:
    - En la columna “codigo_postal” unificamos los valores nulos al “--”. Los valores interpretados como nulos son los siguientes: "0","s/c","-----","no existe", "----------", "--", "-", "N/A",".","CP".
    - Unificamos los valores de “pais castellano” a mayusculas con el abecedario estándar.
- lista-sedes:
    - Unificamos los valores de “pais castellano” a mayúsculas con el abecedario estándar.

4.2 Elección de Datos:

Segun lo pedido en las consignas del trabajo práctico, para realizar nuestro análisis, solo tomamos en cuenta la siguiente información. Notar que los archivos correspondientes estan ubicados en la carpeta “Tablas Post-Importacion”:

De las secciones, solo necesitamos una forma de poder identificarlas, pues lo que nos importa de ellas es su cantidad y ubicación geográfica (país). Así, los únicos datos que tomamos de ellas fueron sede_id y sede_desc_castellano (para usar como PK), esta información proviene de lista-secciones.

Luego, en cuanto a las sedes, al igual que las secciones, lo que más nos interesa para nuestro análisis es su cantidad y ubicación, por lo que tomamos sede_id (para utilizar como PK) y pais_iso 3 (Pues datos_migraciones utiliza ISO 3 para los códigos de los países origen y destino), información la cual provino de lista-sedes-datos. De esa misma tabla, extrajimos la información para cada pais, donde utilizamos solo pais_iso_3 (Para identificar cada pais), region_geografica (para el desarrollo de algunas consignas) y pais_castellano (para identificarlos visualmente) y la informacion para la entidad “Red Social”, pues debido a que su columna correspondiente tenia valores multivaluados. Para red social, nos quedamos con red_social_nombre (Twitter, Instagram, etc.) y red_social_link y sede_id para usar como PK.

Finalmente, para las migraciones, decidimos quedarnos solo con la tupla pais origen y destino, para identificar cada “migracion” individual, y los valores correspondientes a las migraciones de cada una de las décadas. Por otro lado, no mantuvimos la información de género de cada migracion pues no la consideramos relevante para nuestro análisis.

5. Análisis de Datos:

5.1: Reportes:

Confeccionamos una serie de reportes pertenecientes a distintos tópicos de interés. Aquellos cuyo tamaño es demasiado grande, dejamos solo las primeras 6 filas de contenido disponibles. Sin embargo, comentamos lo observado en cada uno de ellos. Cada uno de estos reportes esta disponible de forma completa en la carpeta “Anexo Reportes”.

Tabla 1: Países, cantidad de sedes por pais, el promedio de secciones por sede y flujo migratorio neto de cada país en el año 2000.

pais	cantidad_sedes	promedio_secciones	flujo_migratorio_neto
REPÚBLICA FEDERATIVA DEL BRASIL	11	3.6	-453481
ESTADOS UNIDOS DE AMÉRICA	9	7.5	6.52623e+07
REPÚBLICA ORIENTAL DEL URUGUAY	8	4	-294260
ESTADO PLURINACIONAL DE BOLIVIA	7	8	-545568
REINO DE ESPAÑA	7	5.33333	1.28399e+06
REPÚBLICA DE CHILE	7	14	-659976

Podemos ver que la gran mayoría de las sedes, 62 en total (representado como proporción del total son el 76.54%), son la única sede en su pais. Esto último hace que muchos países tengan un promedio de secciones por sedes mucho más alto, como es el caso para Chile (14) y Rusia (13). Por otro lado, con tan solo examinar la tabla, resulta difícil extraer cualquier tipo de conclusión sobre si existe algún tipo de relación entre los niveles de representación y el flujo migratorio neto de cada uno de los países.

Tabla 2: Cantidad de paises con sede argentina y promedio de flujo migratorio en el 2000 por region geografica

region	cantidad_paises_con_sede	promedio_flujo_migratorio_2000
AMÉRICA DEL NORTE	3	53030.1
AMÉRICA DEL SUR	11	19517.7
EUROPA OCCIDENTAL	18	15353.7
OCEANÍA	2	4862.22
ASIA	23	2078.44
AMÉRICA CENTRAL Y CARIBE	14	297.524
ÁFRICA SUBSAHARIANA	7	269.048
EUROPA CENTRAL Y ORIENTAL	8	90.1875
ÁFRICA DEL NORTE Y CERCANO ORIENTE	5	52.3077

Notamos que, las regiones geográficas con mayor promedio de flujo migratorio son aquellas que están ubicadas más cerca de nuestro pais, como América del Sur y América del Norte, mientras que América Central, al ser una región comparativamente mucho más pequeña y menos poblada, con muchos mas paises, tiene un menor promedio de flujo migratorio.

Excluyendo América del Sur, pues es la región más cercana a nuestro pais y 5 de los 9 paises de América del Sur son limítrofes con la República Argentina, podemos ver que las 3 regiones geográficas con mayor promedio de flujo son aquellas que históricamente contaron con un mayor nivel de riquezas y oportunidades, como son America del Norte, Europa Occidental y Oceania. Esto permite visualizar una problemática grande de nuestro pais, donde muchos individuos deciden emigrar en busca de una mejor calidad de vida y mejores oportunidades, con un énfasis particular en el ámbito económico.

Tabla 3: Cantidad de tipos de red social de sedes argentinas en el exterior por pais.

pais	cantidad_tipos_redes
ESTADOS UNIDOS DE AMÉRICA	6
REINO DE BÉLGICA	6
REINO UNIDO DE GRAN BRETAÑA E IRLANDA DEL NORTE	5
REINO DE ESPAÑA	5
REPÚBLICA ITALIANA	5
CANADÁ	5
REPÚBLICA DEL PARAGUAY	4
REPÚBLICA FEDERATIVA DEL BRASIL	4

Podemos ver que la mayor variedad de redes sociales posibles, es decir, la cantidad de redes sociales distintas que tienen las sedes es un pais es de 6, que es el caso para los Estados Unidos y Bélgica. El valor más popular es el 2, que es el número de redes sociales (distintas) que tienen las sedes de 23 pais, seguido por un empate entre el 0 y el 1, con 19 cada uno.

Tabla 4: Pais, sede, nombre y link de las redes sociales de la sede correspondiente.

pais	sede	po_red_soci	url
AUSTRALIA	CSIDN	Facebook	https://www.facebook.com/ArgentinaEnSidney/
AUSTRALIA	EAUST	Facebook	https://www.facebook.com/ArgentinaEnAustralia/
AUSTRALIA	EAUST	Twitter	https://twitter.com/ARGinAustralia
BARBADOS	EBARB	Facebook	https://www.facebook.com/ArgentinaEnBarbados/
CANADÁ	CTORO	Facebook	https://www.facebook.com/ArgentinaEnToronto/
CANADÁ	CTORO	Instagram	https://www.instagram.com/consuladoargtoronto/
CANADÁ	ECANA	Facebook	https://www.facebook.com/ArgentinaInCanada
CANADÁ	ECANA	Instagram	https://www.instagram.com/argentinaincanada/

Podemos observar que la red social más popular es facebook, donde muchos países tienen incluso múltiples cuentas de facebook, distinguidas segun en que ciudad operan, como

es el caso con Australia y Canada. Por otro lado, Instagram y Twitter son las dos siguientes en términos de popularidad.

5.2: Visualizaciones:

En esta sección, veremos distintas visualizaciones de la informacion que recopilamos a lo largo del trabajo practico, con un foco especifico en buscar dilucidar una relacion entre la cantidad de entidades de representacion que mantiene la Republica Argentina en un pais y su flujo migratorio.

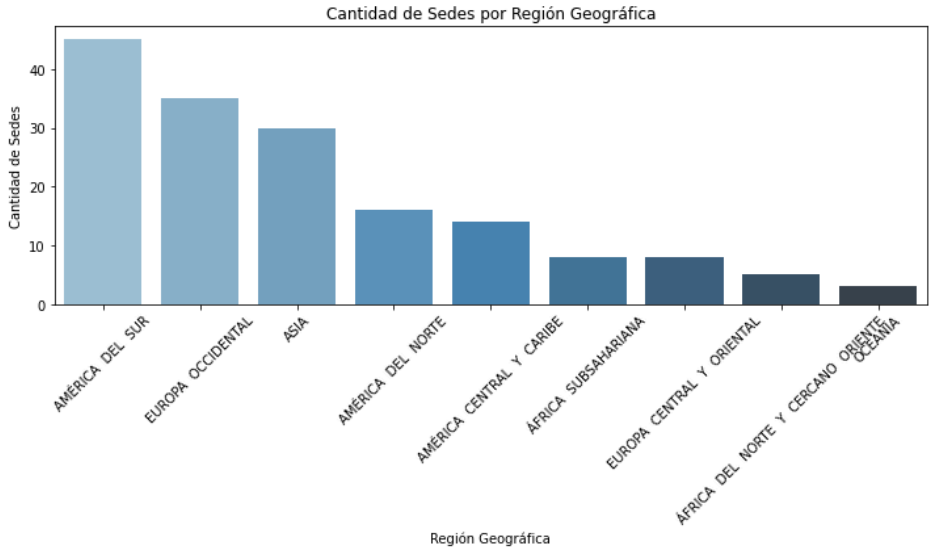

Figura 3: Cantidad de sedes por región geográfica en orden descendiente.

Podemos observar que la region con mas sedes argentinas es América del Sur con 60 sedes y luego Europa Occidental con 35 sedes, seguido de Asia y otros. Las regiones geograficas donde menos sedes hay son África del Norte y Cercano Oriente, y Oceanía con solo 3 sedes

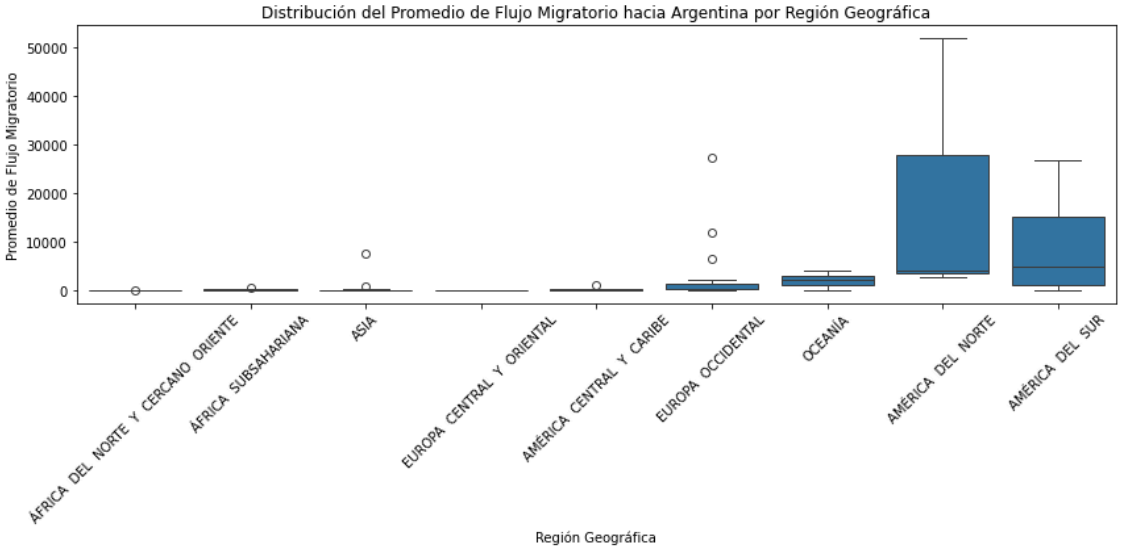

Figura 4: Promedio de flujo migratorio por región geográfica, ordenados segun su mediana.

Podemos observar que las regiones que muestran mayor flujo migratorio son America del Norte y del Sur, que poseen un promedio de flujo migratorio mucho mayor al de otras regiones, mientras que otras regiones se distinguen por algunos de sus outliers, los cuales tienen un promedio de flujo migratorio muy elevado en comparacion a otros paises de la misma region.

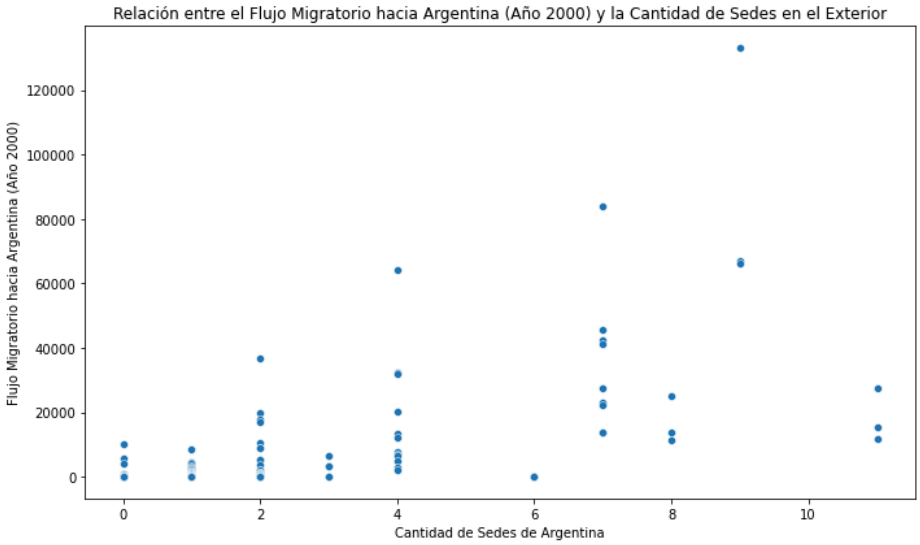

Figura 5: Flujo migratorio desde cada pais hacia Argentina en el año 2000 y la cantidad de sedes que tiene Argentina allí.

Notamos que resulta dificil determinar una relacion entre nuestras dos variables utilizando un scatterplot, por lo cual decidimos generar una visualizacion mas apropiada para determinar si existe dicha relacion.

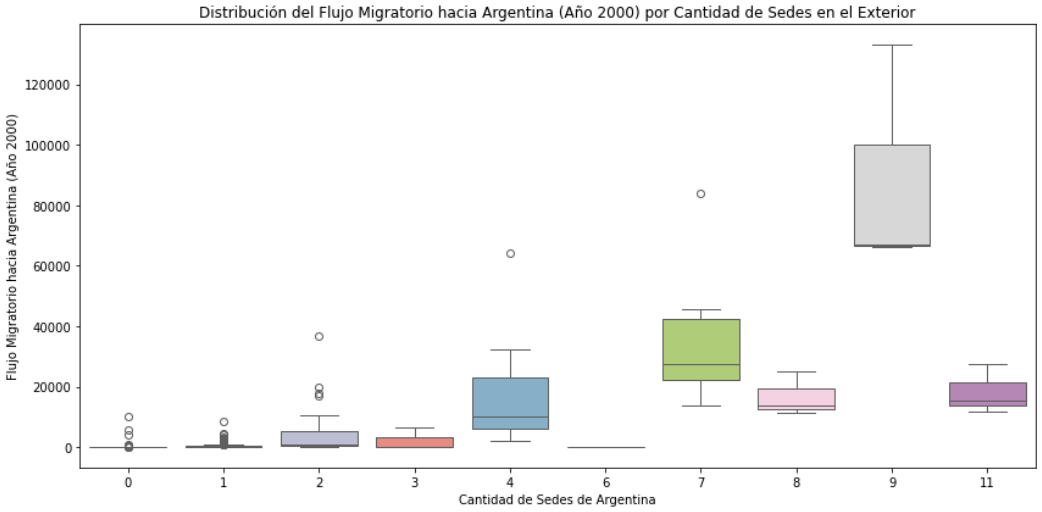

Figura 6: Boxplot de flujo migratorio desde cada pais hacia Argentina en el año 2000 y la cantidad de sedes que tiene Argentina allí.

Haciendo uso de esta visualizacion resulta mas facil observar si existe una relacion entre la cantidad de sedes de la Republica Argentina y el flujo migratorio hacia Argentina. Podemos ver que los paises con mayor cantidad de sedes generalmente muestran un mayor flujo migratorio, con un par de excepciones. Notamos que los paises con 11 sedes resultan ser una excepcion a esta regla, pero es posible que esto se deba a que son solo 3 paises con 11 sedes, por lo cual no podemos decir si contamos con suficiente informacion para determinar si este patron continuaria.

6. Conclusiones:

Para concluir, vamos a observar el desarollo de nuestro trabajo practico y si nuestro objetivo y hipotesis fueron cumplidos o acertados. Por un lado, podemos decir que el objetivo fue cumplido, pues logramos analizar las correlaciones entre las distintas variables como migraciones, migracion neta y la cantidad y densidad de entidades de representacion argentinas a lo largo del mundo. Tambien determinamos que existe al menos una correlacion entre el flujo migratorio entre un pais y el nivel de representacion del estado argentino en el, que podria o no ser una relacion directa.

A modo de perspectiva a futuro, el análisis podría incluir bases de datos que contengan información sobre la situación socioeconómica de los migrantes, para investigar si existe una relación entre el poder adquisitivo de los migrantes y el nivel de representación que tiene nuestro pais en el país destino. Sin embargo, también podríamos haber analizado otras relaciones con la información disponible, como por ejemplo si la distribución en cuanto a género de las migraciones tiene algún tipo de incidencia en la distribución de entidades de representación.

Por otro lado, las bases de datos con las que contamos tenían diversos problemas, algunos de los cuales solucionamos: como información desordenada y en celdas que no corresponden. A modo de mejora, resultaría ideal si la tabla lista-secciones tuviese un único id unico para cada sección, para facilitar su uso, y si las redes sociales en lista-sedes-datos estuviesen en columnas separadas.