Tercera Entrega Proyecto Ingeniería de Datos

González Laura, Mora Diryon, Rincón Laura

Jueves, 12 de mayo de 2022

1. Base de datos

Colombianos registrados en el exterior.

Como se puede apreciar, la base de datos seleccionada incluye la información de los colombianos que viven en el extranjero, los datos son totalmente anónimos e incluye información desde el 2016 hasta inicios del 2022. Sin embargo, no cuenta con la fecha en la que decidieron emigrar. Por otro lado, con el fin de mantener la información, se cuenta con el nombre del país donde viven y su respectivo código ISO, junto a la ciudad donde están registrados con su respectiva coordenada geográfica.

Se tiene la edad del colombiano, como también el área en la que trabaja y qué rama comprende esta. De manera independiente a lo anterior se incluye el nivel académico del individuo, junto a otros datos personales como lo son el género y la estatura. Para finalizar los datos enumeran la cantidad de personas que cumplen con exactamente las mismas condiciones mencionadas anteriormente.

2. Diagrama Entidad Relación

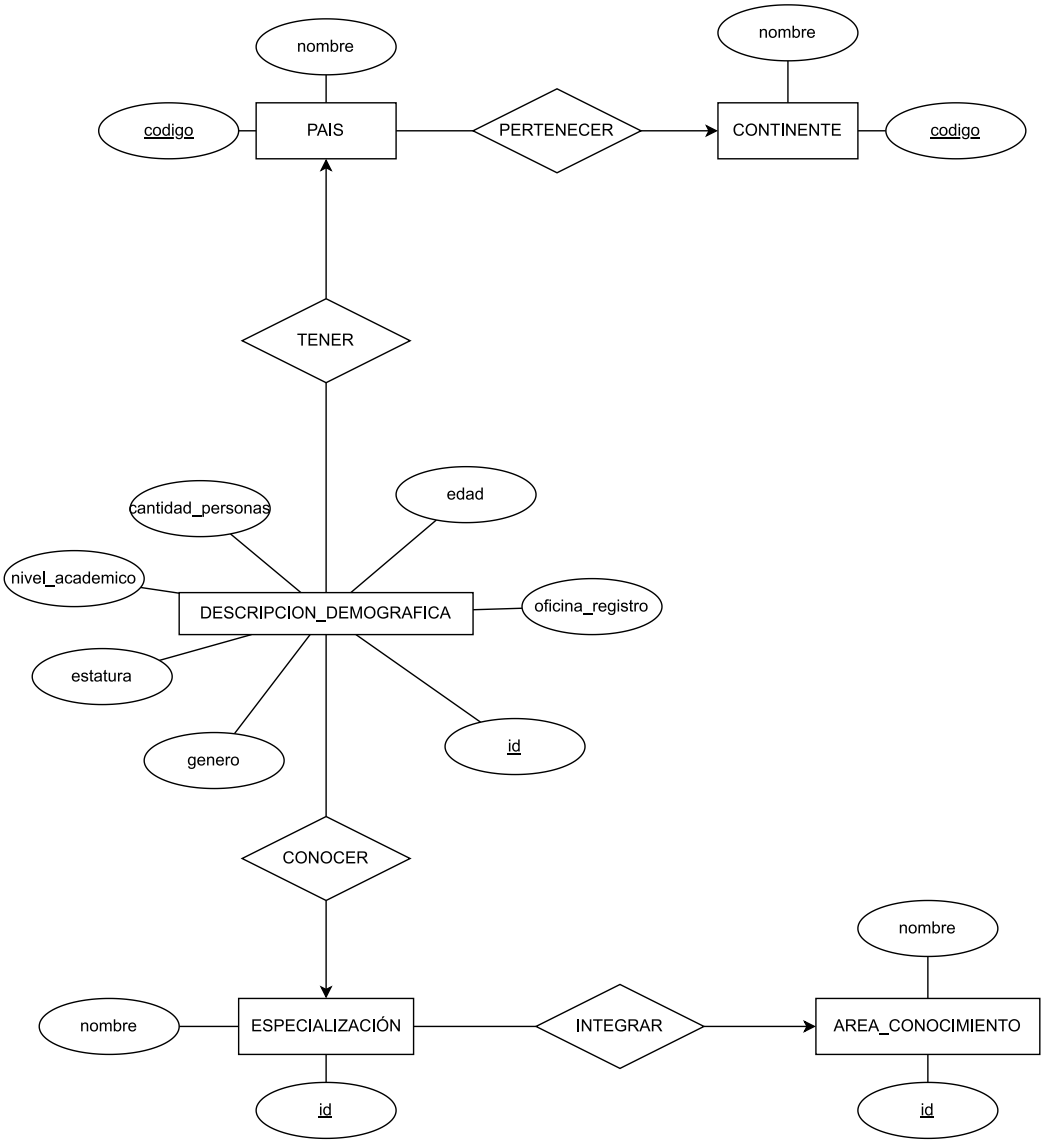

Figura 1: ER

3. Diagrama Relacional Normalizado

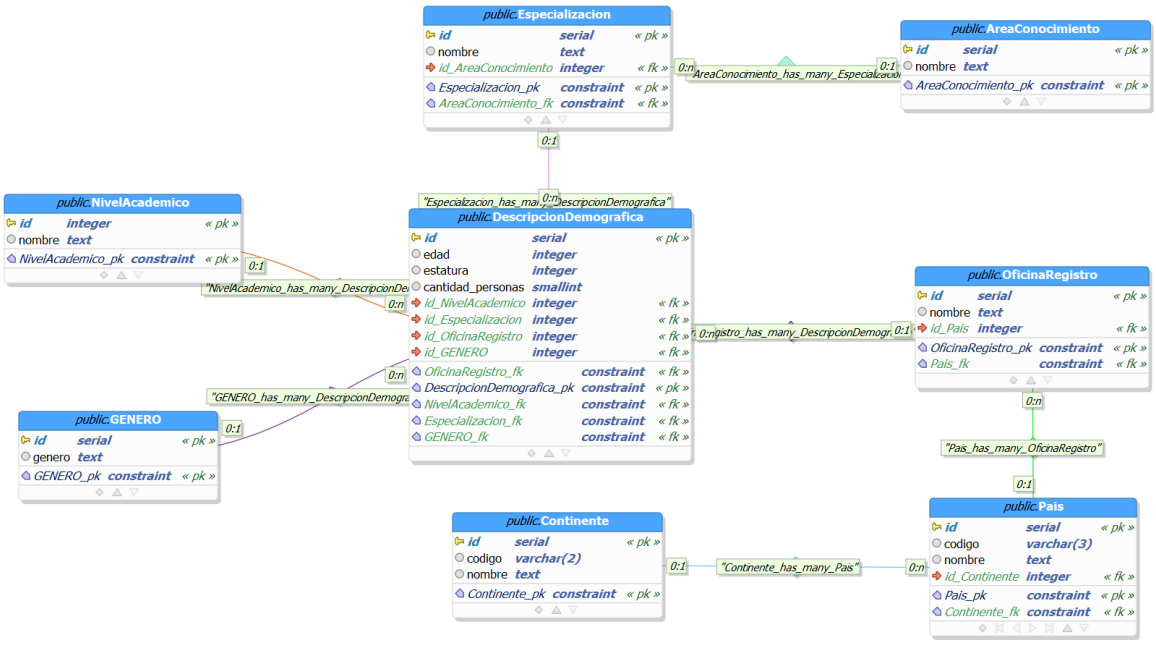

Figura 2: Diagrama Relacional en su forma Normalizado

4. Implementación

La implementación de la base de datos en PostgreSQL y los módulos de Python están disponibles en el repositorio de github. En este se encuentra un archivo README.es.md con las instrucciones de que contiene cada carpeta y la forma de guiarse. También se encuentra un archivo README.md en inglés.

5. Escenarios de Análisis

Para este proyecto, se ha decidido plantear los siguientes escenarios para su posterior análisis y planteamiento de conclusiones:

5.1. Oferta Laboral

Determinar las ciudades y los países con mayor oferta laboral. Para esto se planea mostrar la cantidad de personas que se encuentran en cada zona, ayudándose de graficas de barras, circulares y de mapa.

5.2. Demanda sobre las áreas y las especializaciones

Analizar las áreas y especializaciones con mayor demanda. Esto se observara en las ciudades y países, haciendo una división por el género. Esto se implementara a partir de gráficos de barras, circulares y de mapa.

5.3. Años de las personas

Teniendo en cuenta las descripciones demográficas de los colombianos, se observara los rangos de edades más y menos comunes para emigrar. Esto considerando los años que más se repiten, los datos atípicos y partir de cuál edad se tiene un 25 %, 50 % y 75 % de personas. Todo esto a partir de diagramas de cajas y bigotes e histogramas ('box' y 'hist' en la librería pandas).

5.4. Niveles educativos

Examinar los niveles educativos requeridos para trabajar en un país y continente. Considerando las personas que se encuentran en cada una de las ubicaciones. Esto se logrará por medio de tablas de frecuencia, gráfico de mapa, diagrama de barras y de torta.

6. Implementación en Dash

6.1. Oferta Laboral y Migración

Para este análisis se propuso visualizar los datos de cada país por medio de un mapa. Esto principalmente porque permite de una forma mucho más entendible la visualización de los países, junto a la posibilidad de observar que tan alta es la concentración por medio de una escala de colores y la información específica al momento de pasar el cursor por cada país. Sin embargo, este

no permite hacer mucho énfasis en la información más específica de cuáles son las ciudades con más inmigración, ni permite visualizar directamente cuantas son las personas en cada país. Para este punto también se propuso la observación por medio de gráficos de barras, pero estos resultaban terriblemente extensos y no permitían un correcto análisis.

Por otro lado, para saber las ciudades con mayor migración se propuso gráficos el uso de un gráfico de barras y uno circular. El mayor problema era la descomunal cantidad de ciudades, lo que hacía poco factible la creación de un gráfico por cada país, sin contar lo poco eficiente que sería para un análisis. Por lo cual, la solución fue colocar un limitador para la cantidad de ciudades visualizadas, arbitrariamente se colocó como máximo 30. Así pues, el grafico de barras es un excelente grafico para visualizar la cantidad de personas que hay en cada ciudad. Sin embargo, en las ciudades donde la diferencia no es tan grande, se pierde el nivel comparativo. Por otra parte, la gráfica circular permite una visualización mucho más visual, viendo segmentos claramente más grandes que otros. Pero no llega de ser de todo exacta, principalmente al tener una gran cantidad de datos se pierde la exactitud.

Específicamente hablando, se tiene que los países con mayor inmigración son: Venezuela (210.952 k), Estados Unidos (238.720 k) y España (229.617k). Donde la letra k denota miles de personas. Además, las ciudades con mayor migración de colombianos son: Madrid (104.49 k), Maracaibo (61.565 k), Miami (55.374 k), Caracas (42.527 k) y Nueva York (34.358 k). Esto permite reafirma que Venezuela es el país con mayor migración al tener dos de sus ciudades principales entre las 5 con más altos niveles de personas. A su vez, permite analizar que esto se deba posiblemente a la similitud culturar que posee Venezuela con Colombia, más allá de aspectos laborales o económicos, como es posible que sea en el resto de países.

6.2. Demanda sobre las áreas y las especializaciones

Para esta investigación se propuso visualizar los datos de cada área de conocimiento mediante un gráfico de barras dividió por el género. De esta forma se puede observar claramente las

áreas donde hay más personas, junto a la variable género que permite una mayor comparación. Sin embargo, con este tipo de grafico no se puede saber información más detallada.

Por otro lado, para saber los lugares con mayores especializaciones se propuso un gráfico de mapa a nivel de países y un gráfico de barras por continente. El primero de estos es muy visual y fácil de entender, pero no es tan fácil para realizar comparaciones. El segundo es mucho más adecuado para realizar comparaciones, pero con categorías muy dispersas puede generar confusión. De manera adicional, se tiene un diagrama de barras que divide a cada una de las especializaciones entre hombres y mujeres, junto a un diagrama circular con los porcentajes de cada especialización. Ambos son muy intuitivos de comprender, siendo muy visuales en lo que desean informar, pero con categorías demasiado grandes generan confusión y agobio.

Específicamente hablando, se tiene que las áreas con mayor demanda son: ninguna (458.63 k); economía, administración contaduría y afines (81.561 k) e ingeniería, arquitectura y afines (66.565 k). En general las mujeres tienen una mayor presencia que los hombres, exceptuando el área de ingeniería, donde la mayoría son hombres. A nivel de cada especialización se puede hacer el análisis de cuál es el país con mayor preferencia, el continente donde está más presente (junto a una división por genero), pero al ser tanta cantidad de datos quedara plasmada la idea.

6.3. Años de las personas

Para este análisis se propuso visualizar las edades de las personas emigrantes por medio de un diagrama de cajas y bigotes e histograma. Decidimos estas visualizaciones porque nos permite obtener mucha información como lo es la edad que más se repite, que edades son inusuales hasta posibles errores.

En grupo hemos podido apreciar que estas visualizaciones en conjunto se complementan muy bien para análisis. Sin embargo, si trabajamos con cada visualización por aparte no podríamos tener todos los datos, es decir, si trabajáramos solo con el diagrama de cajas y bigotes no podríamos

saber cuál es la edad que más se repite y si trabajáramos solo con el histograma solo pudiéramos tener datos como la media, la moda, pero información como los datos atípicos no los podríamos tener. Por lo tanto, como ventajas vemos los datos que nos pueden proporcionar ambas en conjunto y como desventajas las visualizaciones por aparte nos harían falta datos para dar respuesta a nuestro análisis.

Si observamos el histograma podremos identificar rápidamente que la edad que más se repite son los 37 años, con una cantidad de 25.032 personas, seguido por la edad de 41 años, con una cantidad de 24.778 personas. Por otro lado, si observamos el diagrama de cajas y bigotes podremos identificar muchos datos atípicos, son datos que se consideran como errores, estos van de un rango de 90-138 años y un solo dato que informa de una persona con 0 años. Ahora bien, las edades anteriormente mencionadas están dentro de un rango de 34 y 56 años, comprendiendo asimismo el 50 % de todas las edades registradas.

6.4. Niveles educativos

Para esta investigación se propuso visualizar los datos de los niveles educativos por país y continente por medio de tablas de frecuencia, grafico de mapa, diagramas de barras y tortas. De esta forma se podría observar claramente en qué lugar predomina más cada nivel académico y con qué frecuencia, lo cual nos brindara información muy detallada.

En grupo hemos podido apreciar que estas visualizaciones nos permiten tener una perspectiva más amplia de que nivel académico es más usual de que las personas posean a la hora de emigrar de Colombia. Por tal motivo pensamos que realmente las visualizaciones que hemos escogido nos han brindado muchas más ventajas que desventajas, lo único que tendríamos por resaltar es que todas se complementan para dar respuesta a nuestro análisis.

Si observamos la visualización de mapa podemos darnos cuenta de que al seleccionar el nivel educativo del que queremos obtener información nos desplegará la frecuencia que posee

dentro del país que seleccionemos. Sin embargo, si lo que queremos es saber qué nivel educativo se repite más en cada continente lo podremos visualizar por medio de una gráfica de barras y de tortas, que está basado en una tabla de frecuencias, brindándonos el primero la frecuencia de cada nivel educativo y el segundo el porcentaje que corresponde a cada uno de estos.

7. Conclusiones

Desde un inicio el manejo de la dimensionalidad del conjunto de datos fue un inconveniente; tan sólo el hecho de tener el archivo CSV en el repositorio era problemático, ¿cómo sería su lectura y división para la creación de las diferentes tablas presentes en el diagrama relacional?

Este único hecho impulsó tal vez el mayor dilema, ¿qué patrón de diseño llevaríamos a cabo? Almacenar todas las consultadas usadas en funciones y hacer las inserciones de manera iterativa, sin ninguna estructura de datos particular; Crear clases que agruparan toda la funcionalidad de un objeto, dígase Tablas SQL que con su nombre de tabla y columnas ya se podrían estructurar las uniones (JOINS) con otras tablas; etc. ¿Qué conllevó esto? Partes de código extenso que no se le sacó el máximo provecho.

En cuanto al desarrollo en Dash, la visualización de los datos fue bastante sencilla de implementar, pero no tanto en el apartado del control de eventos de la página, dígase presionar un botón o manejar la localización actual del URL.

Por ejemplo, al querer cambiar de sección en la página, dígase ir a «Inicio», de la manera tradicional en Dash ocasionaba que todos las gráficas volvieran a cargarse, implicando así conexiones innecesarias a la base de datos y, además, lentitud al obtener de nuevo los datos de manera síncrona. La solución que implementamos fue acudir al manejo de eventos en Javascript para almacenar los gráficos en la página como elementos ocultos.

Por otro lado, debido a que tuvimos cambiar el conjunto de datos, otra problemática fue definir el problema de estos dado que no comenzamos en el «orden natural», reglas de negocio $\rightarrow$ diagrama entidad - relación $\rightarrow$ diagrama relacional $\rightarrow$ diagrama relacional normalizado, sino justamente en el orden contrario.

Finalmente, las diferencias en los sistemas operativos y en las versiones del software usado

nos restringió al momento de trabajar simultáneamente, incluso llegando al punto de no lograr replicar el trabajo hecho entre nosotros. Así, decidimos implementar un contenedor en Docker que nos permitiera ejecutar nuestro proyecto en (casi) cualquier contexto.