Instituto Politécnico Nacional Escuela Superior de Cómputo

Data Mining

Práctica 06 - Definición del proyecto

Graciano Herrera Gabriel
Meza Zamora Abraham Manuel

31 de mayo de 2022

Índice

1. Objetivo

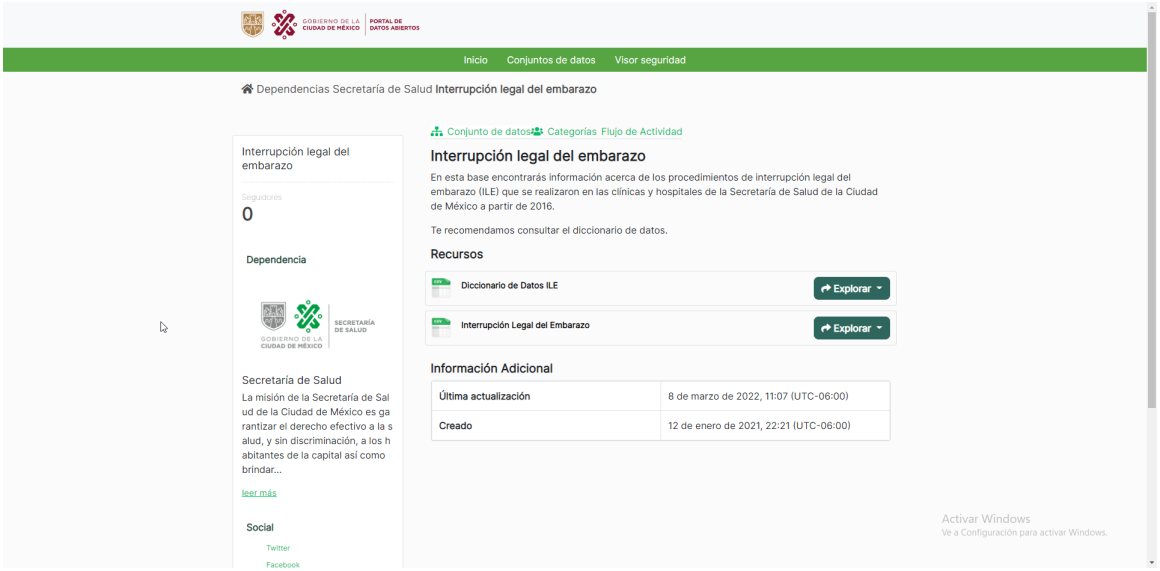
Verificar la factibilidad del proyecto semestral de datos.

2. Procedimiento

Definir el alcance del proyecto semestral de datos, realizando un primer reconocimiento a la muestra de datos a elegir obtenida en el repositorio de la Ciudad de México.

1. Revise la clase que corresponda al tema ”exploración básica de datos con Tableau”, el tema de ”limpieza de datos” y ”CRISP DM”.

2. Explore las diferentes categorías de los conjuntos de datos abiertos de la Ciudad de México: https://datos.cdmx.gob.mx/



Se eligió el conjunto de dataset de interrupción legal del embarazo (ILE) que se realizaron en las clínicas y hospitales de la Secretaría de Salud de la Ciudad de México a partir de 2016.

3. Seleccione un conjunto de datos (dataset) que cumpla con las siguientes condiciones:

   3.1 Tener al menos 1 año de registros o tuplas. Que el dataset se pueda reducir su tamaño haciendo filtros por año (filtrando al año más reciente), en caso que no sea posible procesar todos los registros.

   El dataset elegido cuenta con los datos apartir del año 2016 hasta el año en curso.

   3.2 El dataset debe contener al menos en la dimensión del tiempo “año” y “mes” como dimensión mínima de temporalidad.

   El dataset elegido cuenta con las dimensiones de año, mes y día

   3.3 La dimensión de espacio, al menos deben contener “delegación o alcaldía” y “coordenadas (latitud-longitud)”.

El dataset elegido cuenta con las dimensiones de delegación y alcaldía, pero no las dimensiones de coordenadas de las clínicas, para complementar estas dimensiones sabemos el Clave Única de Establecimientos de Salud (CLUES) por lo que con el catalogo de CLUES que tiene el gobierno federal podremos complementar estos datos.

3.4 Que la cantidad de registros mínima del dataset debe ser 1.5 veces mayor al de incidentes viales usado en prácticas anteriores; es decir aproximadamente mayor a 40 mil registros. En caso de que el dataset en su tamaño original no pueda ser procesado, filtre los datos hasta que el dataset cumpla con este requisito. ES IMPORTANTE IMPORTAR LOS DATOS AL MANEJADOR DE SU PREFERENCIA PARA CONOCER SI ESTE REQUISITO SE CUMPLE.

El dataset cuenta con más de 90 mil registros, haciendo un análisis exploratorio de los datos nos dimos cuenta que hay valores complementarios que son inconsistentes, por lo que esos datos no podrían ser usados para un análisis muy especifico, por ejemplo la edad en la que iniciaron su actividad sexual, en este caso habría que hacer una limpieza de los datos, pero se asegura que al final de esta limpieza contaremos con más de 40 mil registros.

3.5 Buscar una aplicación o caso de estudio de valor adicional del dataset elegido si este se complementa o se le integra información sobre el perfil de la población en la CDMX, está información será obtenida desde el sitio oficial del INEGI.

El dataset en su mayoría cuenta con datos de residentes de la CDMX, por lo que complementándolo con información para las delegaciones de residencia podemos llegar a un análisis más profundo como puede ser la relación que tiene el estatus económico con la cantidad de abortos o nivel educativo entre otros enfoques.

4. Explique en el reporte escrito, cómo el dataset elegido cumple cada uno de los requisitos del punto anterior.

4.1 Adicionalmente, explique cuántas dimensiones temáticas identificó en el dataset. Es importante identificar si el dataset cuenta con diccionario de datos. Por ejemplo, tipo de incidente vial, clasificación del origen del reporte, etc.

El dataset cuenta con un diccionario de datos, las dimensiones identificadas son las temporales, espaciales, de credo, de ocupación, de nivel educativo, estado civil, relacionados a su vida sexual y relacionados al procedimiento.

4.2 Ponga especial atención en describir a detalle la granularidad temporal y espacial. Por ejemplo, el nivel de descripción del tiempo: día, mes, año minuto, segundo, etc.

En la granularidad temporal la dimensión más baja que tenemos es el día, mes, año de cuando se inició el proceso de ILE y cuando finalizó

4.3 Explique a detalle el caso de estudio adicional donde el dataset elegido se pueda complementar con datos del perfil población de la CDMX. Por ejemplo, en el dataset de incidentes viales integrando el perfil poblacional, podemos conocerla relación entre la cantidad de incidentes y la cantidad de población que vive en la delegación Coyoacán.

Como se mencionó en la respuesta: El dataset en su mayoría cuenta con datos de residentes de la CDMX, por lo que complementándolo con información para las delegaciones de residencia podemos llegar a un análisis más profundo como puede ser la relación que tiene el estatus económico con la cantidad de abortos o nivel educativo entre otros enfoques.

4.4 Explique las razones o los motivos por las que ha elegido el dataset.

Revisando el catalogo de datos encontramos este primeramente lo elegimos porque cumplía con los requisitos del punto 3, después de realizar un análisis exploratorio de los datos nos dimos cuenta que el dataset podía ser analizado con diferentes enfoques con los cuales podríamos llegar a conclusiones interesantes acerca del aborto en la CDMX, además que el tema del aborto es un tema importante para el libre desarrollo de las mujeres en el país

4.5 Explique el problema que quiere resolver al explorar los datos. El alcance del proyecto: es decir explicar cual es el conocimiento que espero descubrir al estudiar el dataset.

Tenemos distintos enfoques o problemas que nos gustaría explorar con los datos, a continuación se listan algunos de ellos.

- **Delegación que tiene más abortos.**
- **Edad promedio de abortos en la CDMX.**
- **Posible relación entre el nivel económico o social y la decisión de realizarse un procedimiento de ILE.**
- **Posible relación entre el credo y la decisión de realizarse un procedimiento de ILE.**
- **Posible relación entre el uso de métodos anticonceptivos y el embarazo.**
- **Procedimiento más común con el que se realiza el ILE.**
- **Mes en el que más se aborta.**
- **Tendecia del aborto en la CDMX.**
- **Impacto de la cuarentena por COVID-19 en la cantidad de abortos.**
- **Motivo más común para realizar un ILE.**

5. Realice el análisis exploratorio básico usando Tableau, contestando las siguientes preguntas generales. Responda aplicando su propio criterio, es decir filtrando la información como considere conveniente. Agregue los resultados en el reporte.

5.1 ¿Cual es la distribución de la dimensión categórica ó temática (el tema del dataset) más importante (del fenómeno que es descrito por el dataset)?. Ej. La distribución general de incidentes viales.

La distribución general de Interrupciones Legales del Embarazo

5.2 ¿Cual es la distribución del fenómeno que mide el dataset en el tiempo?, explorar la mayor cantidad de los niveles de granularidad de tiempo. Ej. La distribución anual de incidentes viales por mes.

La distribución anual de Interrupciones Legales del Embarazo

5.3 ¿Encontró valores atípicos en el dataset o valores inconsistentes? (verificar el diccionario del datos).

Si se encontraron valores atípicos, por ejemplo en los motivos, semanas de gestación, el anticonceptivo usado entre otros.

5.4 Verifique si las preguntas se pueden procesar con todos los registros originales del dataset o explique si el dataset fue recortado o filtrado por tiempo u otra variable.

Dependiendo de la pregunta se deben realizar ciertos cambios en el dataset, además de complementarlo con los catálogos del CLUES y los que vayamos a utilizar relacionados al perfil de población proporcionados por INEGI

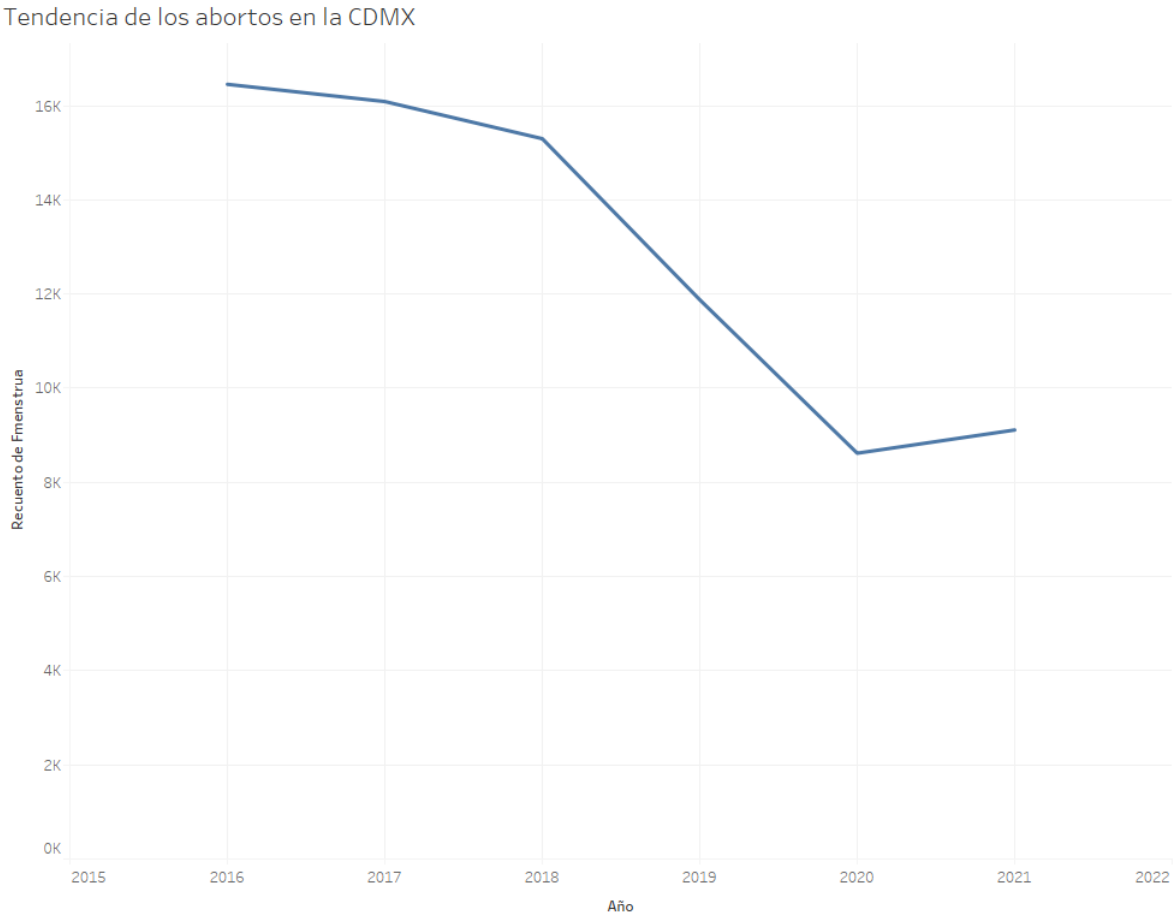


Figura 1: Tendencia de los abortos en la CDMX

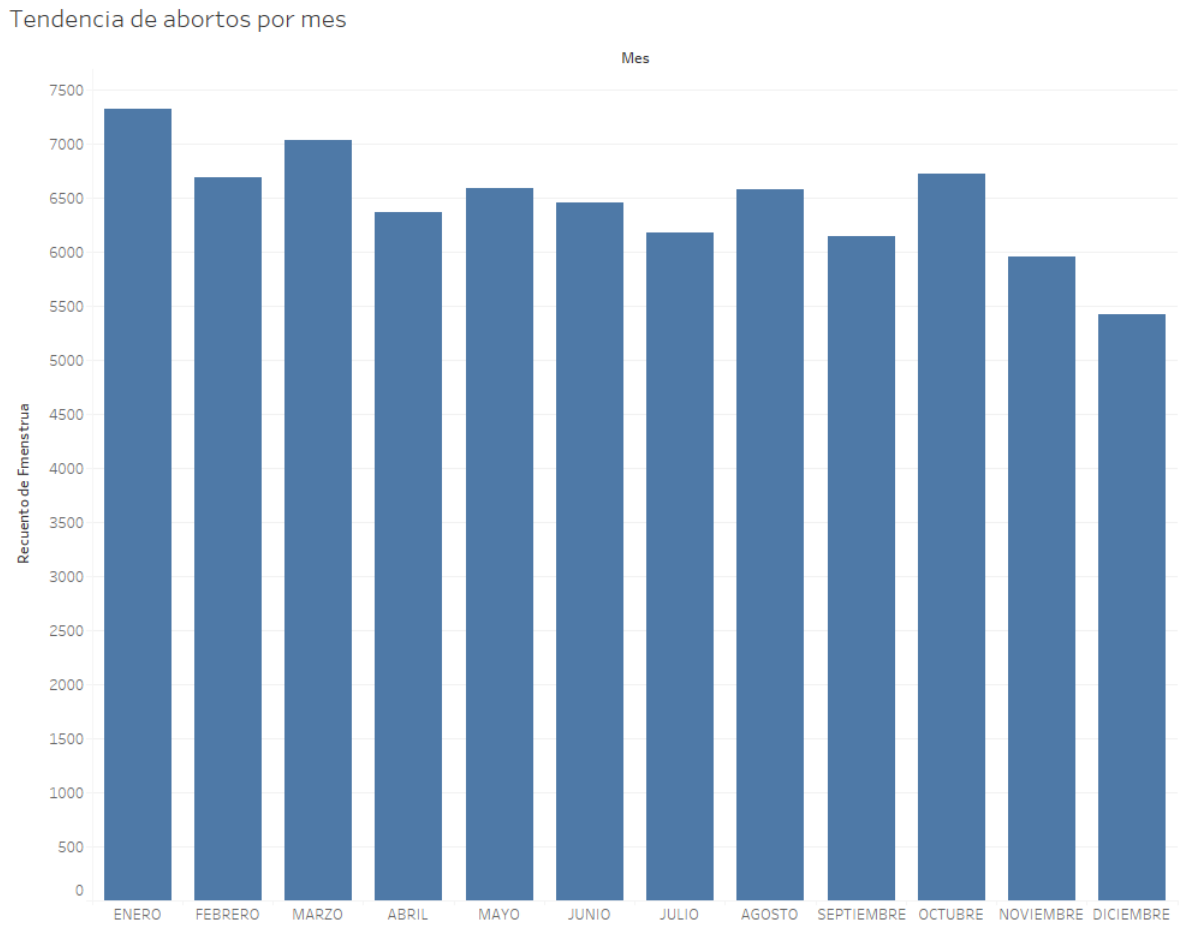


Figura 2: Tendencia de abortos por mes

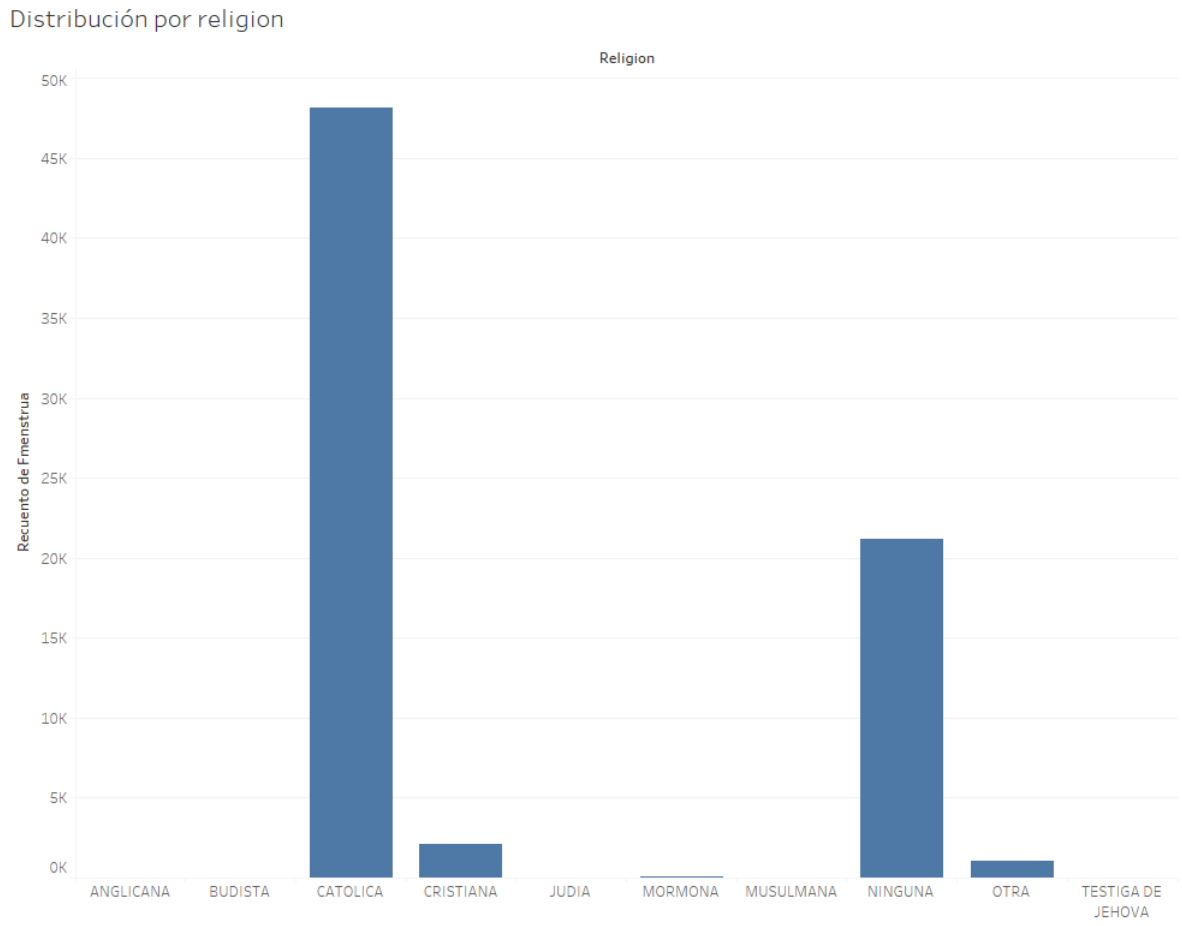


Figura 3: Distribución de abortos por religión.

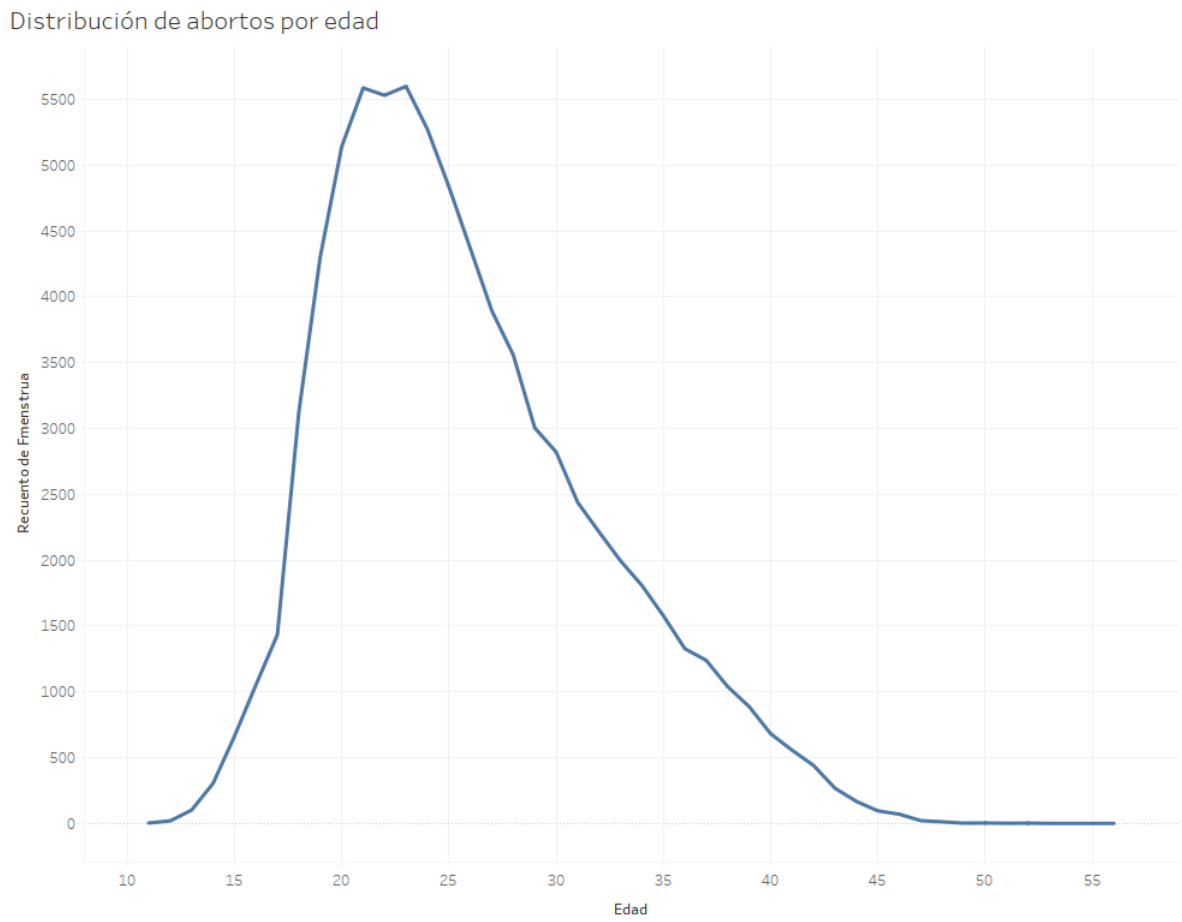


Figura 4: Distribución de abortos por edad.

3. Conclusiones

- Graciano Herrera Gabriel
  Después de realizar el análisis exploratorio de los datos podemos concluir que el dataset elegido es factible para utilizarlo como proyecto semestral, aunque dependiendo del enfoque a analizar habrá que realizar distintos cambios en la estructura del dataset y de los datos para que el análisis sea más fácil de realizar y certero. Lo interesante de este dataset es que tiene muchos enfoques de análisis, mientras realizábamos la exploración se nos ocurrieron bastantes, realizando un análisis más detallado de los datos seguramente se encontrarán más enfoques y se podrán llegar a conclusiones bastantes interesantes.

- Meza Zamora Abraham Manuel
  Con lo aprendido en el curso es fácil ver que dicho conjunto de datos cumple con los requerimientos del proyecto, es interesante analizar este tipo de datos y observar cierto tipo de tendencias, dependiendo del giro que busquemos darle a este proyecto, se podrá complicar el proceso de limpieza de datos. Es de suma importancia mencionar que creemos que los resultados del mismo podrán ayudar a reflejar una parte de la realidad de un tema controversial en la CDMX.