Instituto Politécnico Nacional

Escuela Superior de Cómputo

Data Mining

Profesor: Zagal Flores Roberto Eswart

Autor: Velasco Huerta Angel Eduardo

Práctica no. -3

26/09/2021

Proyecto Data Mining.

Dataset seleccionado:

Tras buscar por muchos datasets alojados en el portal de datos abiertos de la Ciudad de México, se encontraron algunos, que cumplen con los requisitos especificados para una correcta exploración de datos, además, tienen ciertas aplicaciones o temas de interés actual. Los datasets son los siguientes:

- Victimas en carpeta de investigación.
- Interrupciones legales del embarazo.
- Carpetas de investigación de la FGJ.

Finalmente, se selecciono el dataset de Interrupciones legales del embarazo, el dataset de carpetas de investigación, también cumplía con los requerimientos, sin embargo, era excesivamente grande, y la mayoría de las muestras representaban un periodo temporal no mayor a 5 años, motivo por el cual, aunque se limpiaran registros antiguos, seguirían quedando aproximadamente 850,000 registros, lo cual es aproximadamente 20 veces mas grande que el dataset utilizado previamente. En cuanto al de Victimas en carpeta, no se contaba con ningún tipo de dimensión de espacio.

Dataset: Interrupción legal del embarazo en la Zona Metropolitana.

3 - Puntos definidos para seleccionar el dataset:

3.1 -Tener al menos 1 año de registros o tuplas: El dataset, cuenta con registros desde 2016-2020

3.2 -El dataset debe contener al menos en la dimensión del tiempo: Se cuenta con año, mes y día de interrupción, hospitalización, y salida.

3.3 -La dimensión de espacio: Se tiene información respecto a la alcaldía y la entidad.

3.4 -Que la cantidad de registros mínima del dataset debe ser 1.5 veces mayor al de incidentes viales usado en prácticas anteriores: Se cuenta con 79,384 registros, lo que representa prácticamente 2.5 veces la cantidad de registros.

3.5 -Buscar una aplicación o caso de estudio de valor adicional del dataset elegido: Otra de las razones por las cuales se escogió este dataset, es porque el tema del aborto, es muy controvertido últimamente, el objetivo del proyecto no es dar una opinión del mismo, sino obtener datos importantes con base en las estadísticas presentes, datos que ayuden a informar del tema, asimismo, existen muchos datasets con los cuales se puede cruzar información, como los relevantes a información por delegación, o una aplicación en concreto que se podría dar, es la relacionada a registros de ocupación y recursos hospitalarios, al cruzar esta información (misma que se encuentra en el portal de datos abiertos) podríamos relacionar los hospitales de las delegaciones donde se dan más casos de aborto, para así determinar si la distribución de recursos y hospitales, cubren este servicio de salud pública correctamente. El dataset, cuenta con un gran número de columnas, en las cuales tenemos mucha información que nos permita realizar comparaciones o exploración de datos, tales como la religión, la edad, la escolaridad, etc.

Dimensiones temáticas:

Este dataset viene muy completo (Incluso algunos datos podrían ser irrelevantes dependiendo la aplicación), afortunadamente, contamos con un diccionario de datos:

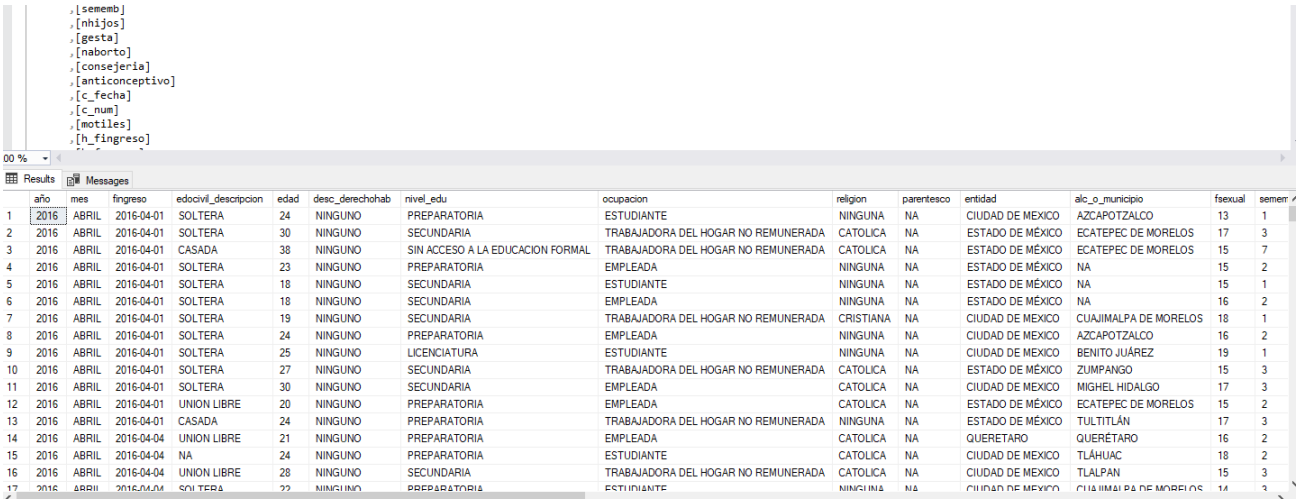
Variable	Descripción	Tipo de dato
año	Año en que se realizó el procedimiento	numérico
mes	Mes en que se realizó el procedimiento	texto
clues_hospital	Clave Única de Establecimientos de Salud (CLUES)	texto

fingreso	Fecha de Interrupción Legal del Embarazo	fecha
edocivil_descripcion	Estado civil	texto
edad	Edad cumplida en años	numérico
desc_derechohab	Especificar la institución que otorga la derechohabiencia	texto
nivel_edu	Último nivel escolar acreditado	texto
ocupacion	Ocupación	texto
religion	Religión	texto
parentesco	Parentesco del responsable con la paciente solamente para menores de edad obligatorio	texto
entidad	Entidad de residencia	texto
alc_o_municipio	Alcaldía o municipio de residencia	texto
fsexual	Edad de inicio de vida sexual activa	numérico
sememb	Semanas de embarazo por fecha de última menstruación	numérico
nhijos	Número de hijos	numérico
gesta	Número de embarazos (Incluyendo abortos)	numérico
naborto	Número de abortos (Sin contar ILE)	numérico
consejeria	Especificar si la paciente recibió consejería sobre la ILE	texto
anticonceptivo	Especificar si la paciente utiliza de forma habitual método anticonceptivo	texto
c_fecha	Fecha de primera valoración o atención	fecha
c_num	Número de consultas previas al ILE	numérico
motiles	Motivo por el cual se desea la interrupción del embarazo	texto

h_fingreso	En caso de hospitalización, fecha de ingreso	fecha
h_fegreso	En caso de hospitalización, fecha de egreso	fecha
desc_servicio	Servicio en el que se otorgó la interrupción	texto
p_semgest	Semanas de gestación por USG	numérico
p_diasgesta	Días de gestación por USG	numérico
s_complica	Se presentaron complicaciones por el procedimiento	texto
c_dolor	Se presentó dolor posterior al procedimiento	texto
fecha_cierre	Fecha de cierre del procedimiento de ILE	fecha
resultado_ile	Resultado del procedimiento de ILE	texto

Importación de los registros (Limpieza)

Algo importante que noté en este dataset, es que existen muchos registros con nulos, o información que no corresponde, afortunadamente, estos solo se encontraban en el final del dataset, por lo que se eliminaran, asimismo, en los registros muchas veces se encuentra NA, lo cual se tomará como que no se respondió, pero para evitar conflictos con el tipo de datos, se elimino para obtener mejores resultados en las consultas.



```
,[sememb]
,[nhijos]
,[gesta]
,[naborto]
,[consejeria]
,[anticonceptivo]
,[c_fecha]
,[c_num]
,[motiles]
,[h_fingreso]
```

00 %

Results | Messages

	año	mes	fingreso	edocivil_descripcion	edad	desc_derechohab	nivel_edu	ocupacion	religion	parentesco	entidad	alc_o_municipio	fsexual	semem
1	2016	ABRIL	2016-04-01	SOLTERA	24	NINGUNO	PREPARATORIA	ESTUDIANTE	NINGUNA	NA	CIUDAD DE MEXICO	AZCAPOTZALCO	13	1
2	2016	ABRIL	2016-04-01	SOLTERA	30	NINGUNO	SECUNDARIA	TRABAJADORA DEL HOGAR NO REMUNERADA	CATOLICA	NA	ESTADO DE MÉXICO	ECATEPEC DE MORELOS	17	3
3	2016	ABRIL	2016-04-01	CASADA	38	NINGUNO	SIN ACCESO A LA EDUCACION FORMAL	TRABAJADORA DEL HOGAR NO REMUNERADA	CATOLICA	NA	ESTADO DE MÉXICO	ECATEPEC DE MORELOS	15	7
4	2016	ABRIL	2016-04-01	SOLTERA	23	NINGUNO	PREPARATORIA	EMPLEADA	NINGUNA	NA	ESTADO DE MÉXICO	NA	15	2
5	2016	ABRIL	2016-04-01	SOLTERA	18	NINGUNO	SECUNDARIA	ESTUDIANTE	NINGUNA	NA	ESTADO DE MÉXICO	NA	15	1
6	2016	ABRIL	2016-04-01	SOLTERA	18	NINGUNO	SECUNDARIA	EMPLEADA	NINGUNA	NA	ESTADO DE MÉXICO	NA	16	2
7	2016	ABRIL	2016-04-01	SOLTERA	19	NINGUNO	SECUNDARIA	TRABAJADORA DEL HOGAR NO REMUNERADA	CRISTIANA	NA	CIUDAD DE MEXICO	CUAJIMALPA DE MORELOS	18	1
8	2016	ABRIL	2016-04-01	SOLTERA	24	NINGUNO	PREPARATORIA	EMPLEADA	NINGUNA	NA	CIUDAD DE MEXICO	AZCAPOTZALCO	16	2
9	2016	ABRIL	2016-04-01	SOLTERA	25	NINGUNO	LICENCIATURA	ESTUDIANTE	NINGUNA	NA	CIUDAD DE MEXICO	BENITO JUÁREZ	19	1
10	2016	ABRIL	2016-04-01	SOLTERA	27	NINGUNO	SECUNDARIA	TRABAJADORA DEL HOGAR NO REMUNERADA	CATOLICA	NA	ESTADO DE MÉXICO	ZUMPANGO	15	3
11	2016	ABRIL	2016-04-01	SOLTERA	30	NINGUNO	SECUNDARIA	EMPLEADA	CATOLICA	NA	CIUDAD DE MEXICO	MIGHEL HIDALGO	17	3
12	2016	ABRIL	2016-04-01	UNION LIBRE	20	NINGUNO	PREPARATORIA	EMPLEADA	CATOLICA	NA	ESTADO DE MÉXICO	ECATEPEC DE MORELOS	15	2
13	2016	ABRIL	2016-04-01	CASADA	24	NINGUNO	PREPARATORIA	TRABAJADORA DEL HOGAR NO REMUNERADA	NINGUNA	NA	ESTADO DE MÉXICO	TULTITLÁN	17	3
14	2016	ABRIL	2016-04-04	UNION LIBRE	21	NINGUNO	PREPARATORIA	EMPLEADA	CATOLICA	NA	QUERETARO	QUERÉTARO	16	2
15	2016	ABRIL	2016-04-04	NA	24	NINGUNO	PREPARATORIA	ESTUDIANTE	CATOLICA	NA	CIUDAD DE MEXICO	TLÁHUAC	18	2
16	2016	ABRIL	2016-04-04	UNION LIBRE	28	NINGUNO	SECUNDARIA	TRABAJADORA DEL HOGAR NO REMUNERADA	CATOLICA	NA	CIUDAD DE MEXICO	TLALPAN	15	3
17	2016	ABRIL	2016-04-04	SOLTERA	22	NINGUNO	PREPARATORIA	ESTUDIANTE	NINGUNA	NA	CIUDAD DE MEXICO	CUAJIMALPA DE MORELOS	14	3

Después de una breve limpieza de nulos, se terminó con 71,108 registros.

Otro dato importante para considerar es que, en la base, es que se encuentran muchos datos con NA, esto pues en los registros, las personas prefieren no contestar cierta información o simplemente no se registró correctamente. Para la presente práctica no se eliminarán los registros completos por contar con NA, pero si basándonos en la primera exploración derivada de esta primera fase, se nota que los datos no se ven tan claros por eso, se desecharán registros no completos.

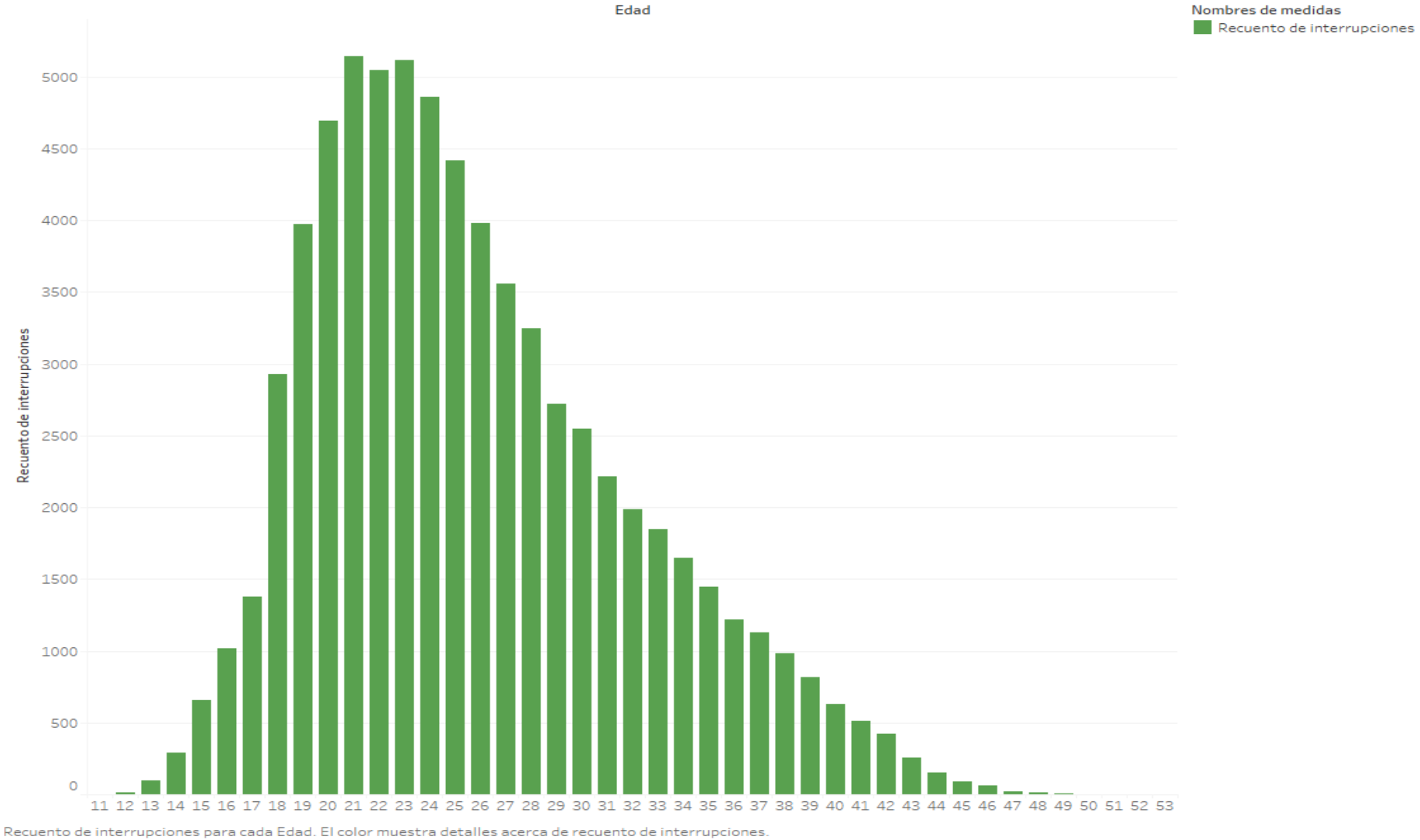
Análisis exploratorio con Tableau:

¿Cual es la distribución de la dimensión categórica o temática (el tema del dataset) más importante (del fenómeno que es descrito por el dataset)?

Como primera exploración, encontré dos dimensiones que me pareció importante destacar, las de edad y la de delegación, siento que ambas brindarian mas información para una aplicación.

P.D. Se anexan capturas de los libros de tableau para una mejor visualización

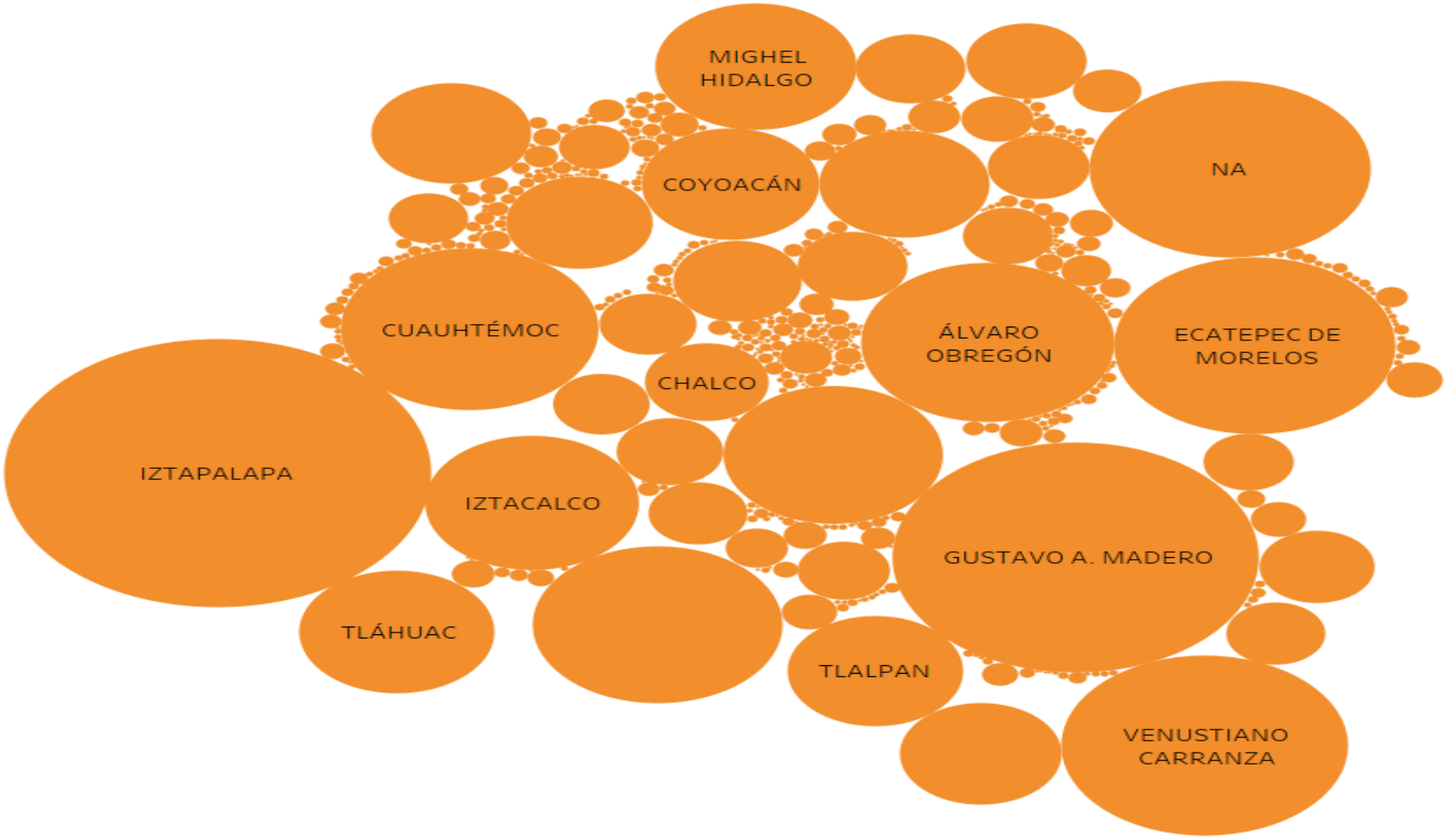
EDAD



Recuento de interrupciones para cada Edad. El color muestra detalles acerca de recuento de interrupciones.

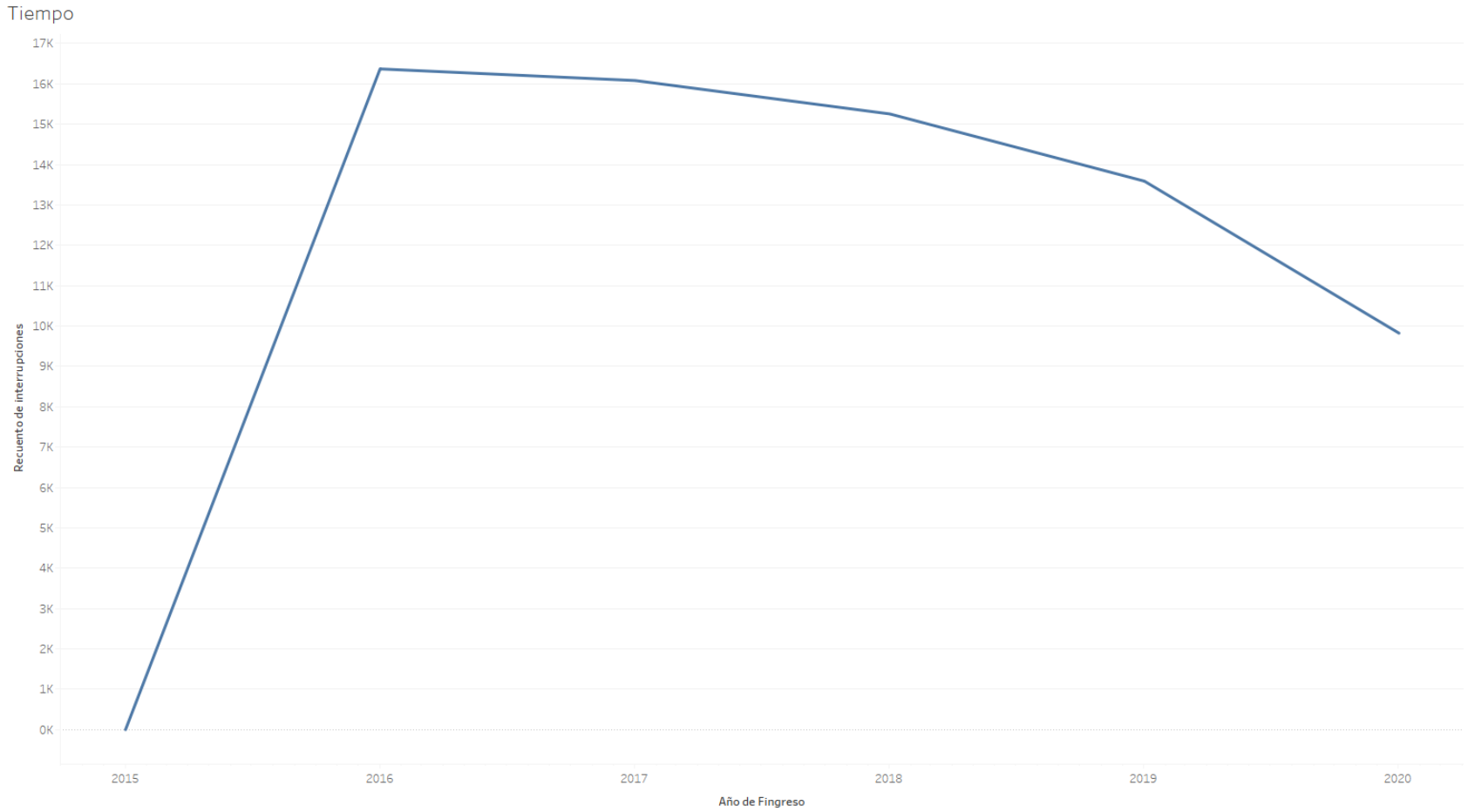
EDAD

ALCALDÍAS



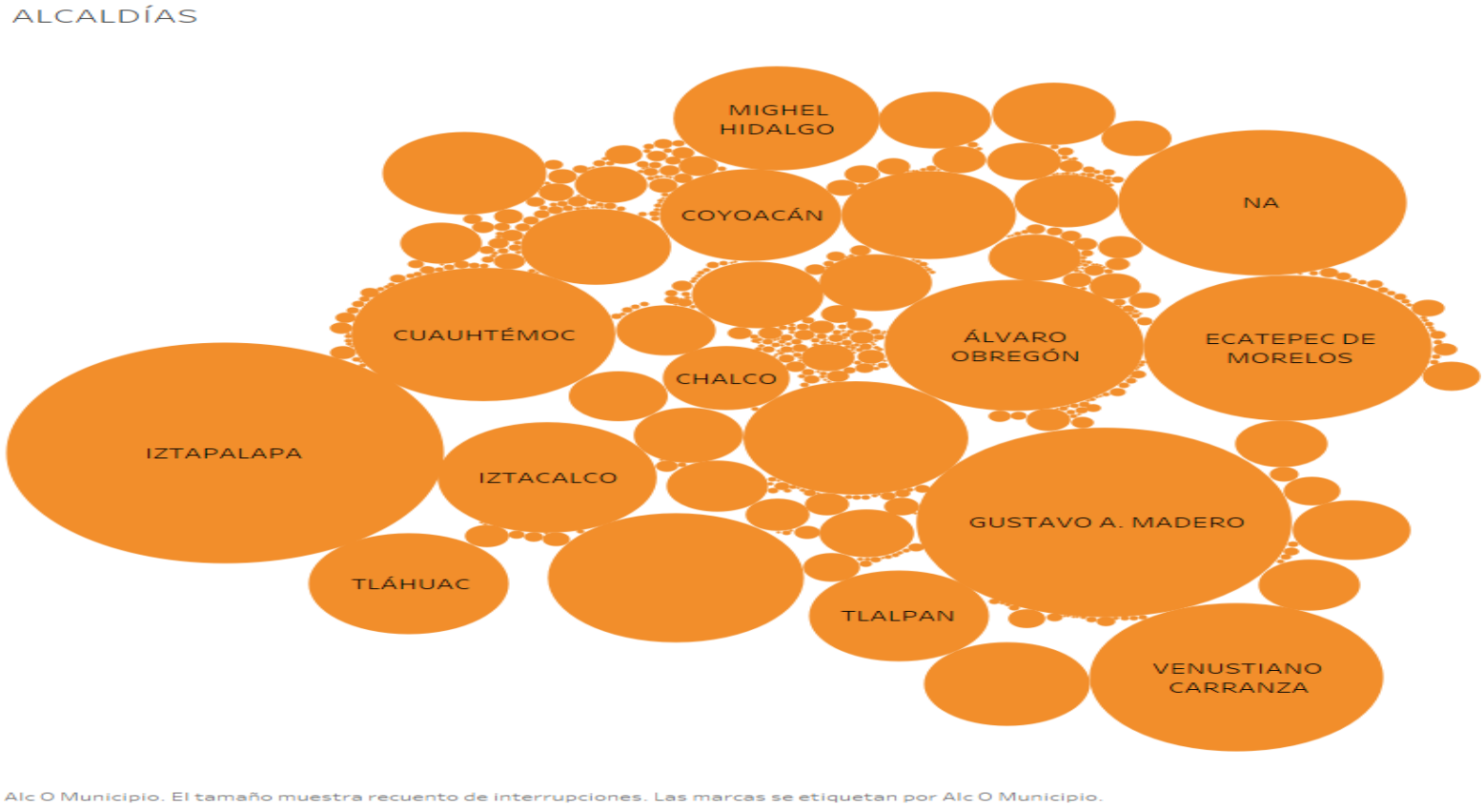
Alc O Municipio. El tamaño muestra recuento de interrupciones. Las marcas se etiquetan por Alc O Municipio.

¿Cuál es la **distribución del fenómeno** que mide el dataset en el tiempo?

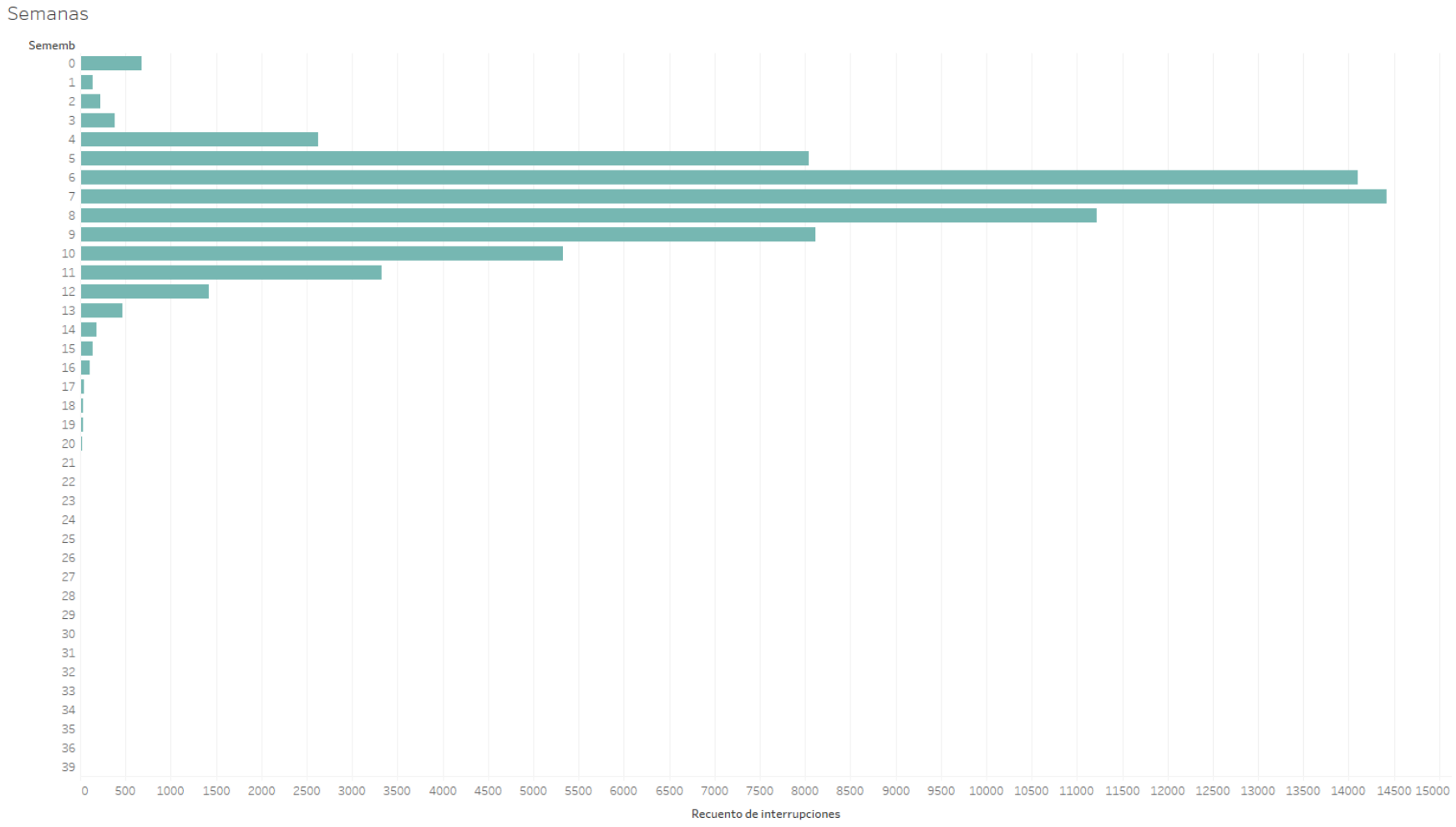


La tendencia de recuento de interrupciones para Fingreso año.

¿Cuál es la distribución del fenómeno que mide el dataset en el espacio?



¿Cuál es la distribución de otras dimensiones temáticas (que consideren importante) del dataset?



Recuento de interrupciones para cada Sememb.

Motivos2

Motiles	Recuento de interrupciones
DECISIONES VÍNCULADAS..	12
FALLA DEL METODO	11
INTERRUPCION VOLUNTA..	69,993
NA	136
OTRA	558
PROBLEMAS DE SALUD	24
PROYECTO DE VIDA	272
SIN APOYO	4
SITUACION ECONOMICA	87
VIOLACION	11

Recuento de interrupci..
4 69,993

Recuento de interrupciones desglosado por Motiles. El color muestra recuento de interrupciones. Las marcas se etiquetan por recuento de interrupciones.

Hijos

Nhijos (color) y recuento de interrupciones (tamaño).

Recuento de interrupciones
71,108

Nhijos
0
1
2
3
4
5
6
7
8
9
10

5.6 y 5.7.

Al realizar una primera exploración de los datos encontramos muchos registros que no contaban con la información necesaria para la exploración, la eliminación de estos ayudó a tener mejores datos, sin embargo, todavía existen algunos registros con NA, (se procuró no usar esas dimensiones) que podrían no representar información de manera tan clara, lo bueno de esto, es que existen muchos campos a explorar, y solo son pocos los que cuentan con este vacío de información.

Conclusiones:

Una vez analizado el dataset, puedo concluir que es trabajable, y que es factible su procesamiento, esto por varios motivos:

- **Cuenta con información suficiente, desde 2015 a la fecha, y con una gran cantidad de registros.**
- **Cuenta con muchas dimensiones de información para explorar, por lo que se pueden realizar distintos análisis.**
- **La información no es tan complicada de limpiar y mantiene un formato consistente.**
- **Representa un tema del cual existe mucha aplicación y se puede obtener información valiosa.**
- **Cumple con las dimensiones básicas especificadas**