Ciudad de México, a 5 de agosto de 2020

Asunto: Limpieza de base de datos de Interrupción legal del embarazo

---

Contexto:

- 4 años de datos (2016-2020)
- Más de 70,000.
- Campos abiertos

Consideraciones:

- La interpretación debe de considerar la totalidad de datos. Por ejemplo, si en la mayoría ponen “sí” y “no” el 1 será interpretado como sí y 0 como no. En caso de controversia, podemos acudir con la Secretaría de Salud o el diccionario de datos. El objetivo es conservar la mayor cantidad de datos vivos.
- Tratar de encontrar el común agrupable. Por ejemplo, en derechohabiente “IMSS”, “IMSS gratuito”, “IMSS Estado de México” se tomará como IMSS.
- Eliminar resultados atípicos. Por ejemplo, en fecha de menstruación un año 2035. O 130 semanas de gestación. Es mejor tener un NA.
- Documentar el proceso para futuras aclaraciones.
- Usar en todos los casos mayúsculas.
- Hacer la limpieza en una columna nueva. Por ejemplo, la columna edad, con **edad_limpia**
- Los missing values siempre deben de ser valores en blanco.
    - NE
    - NA
    - Sin dato

Solicitud de limpieza:

Columnas	**Problema detectado**[1]	**Posible solución**
mes	Hay observaciones con mayúsculas y minúsculas	Pasar todo a mayúsculas
hospital, hospital 2	Hay dos columnas con hospitales. Por motivos de seguridad es mejor que no sea público.	Juntarlas. Agrupar. Después crear ID para que no sea identificable.
fecha ingreso, fecha menstrua, c fecha, h f	Dos o más formatos de fecha. “42446” ,	Hacer todos numéricos. Separar por distintos tipos

[1] Podría haber más problemas.

egreso, h f ingreso, fecha cierre	“dd/mm/aaaa”, “aaaa-mm-dd”. Comentarios escritos.	de estructura: los que tienen 5 dígitos y los que tienen 10. A cada uno darle tratamiento distinto. Establecer aaaa-mm-dd como formato único. Juntar.
autoref, descripción de estado civil, derechohabiencia, nivel educativo, ocupación, religión, parentesco, entidad, delegación o municipio, consejería, anticonceptivo, motiles, descripción del servicio, proc. ile, se complica, pastilla anticonceptiva, medicamento	No están agrupados. No están en lenguaje femenino (por ejemplo contador, pasar a contadora). Usan distintas formas para decir sí o no. Quitar números tipo formulario.	1. Usar case when, grepl y agrupar todas las opciones 2. Volverlos factores, declarar los niveles. (Problema con NA).
f sexual, menarca, semanas de embarazo, número de hijos, gesta, número de aabortos, número de partos, número de cesáreas, c num, semanas de gestación, días de gestación	Son datos numéricos pero hay algunos casos escritos con letra. Los missing values son no están	Convertir los casos con letra a número. Quitar la palabra de medida de unidad (como “años”). Hacer NA valores que no no fueron registrados. Redondear.