scidata

Gestión de los Datos

Instructor: Juan Luis Palacios Soto

palacios.s.j.l@gmail.com

Contenido

Recolección de datos

Por lo general, buscamos que nuestros datos cumplan las siguientes características:

- Temporalidad. Es decir, que sean relativamente recientes. Si los datos no son relativamente recientes, difícilmente se podrán usar para realizar predicciones.
- Adecuados. También llamada representatividad. Si eliges muy poca cantidad de muestra o una cantidad demasiado grande, puedes terminar provocando un análisis inadecuado de los datos.
- Válidos. Validar los datos es un proceso que comprueban y aseguran que un programa funcione de manera correcta y segura con datos limpios y útiles.
- Medibles. Que tus datos sean cuantitativos.

Métodos de recolección de datos

Anteriormente ya hemos visto que existen varias fuentes de datos entre primarias y secundarias. A saber:

1. Encuestas
2. Censos
3. Sondeos
4. Casos particulares de estudio
5. Agencias recolectoras de datos
6. Fuentes públicas
7. Experimentos
8. Redes sociales e internet

Encuestas y censos

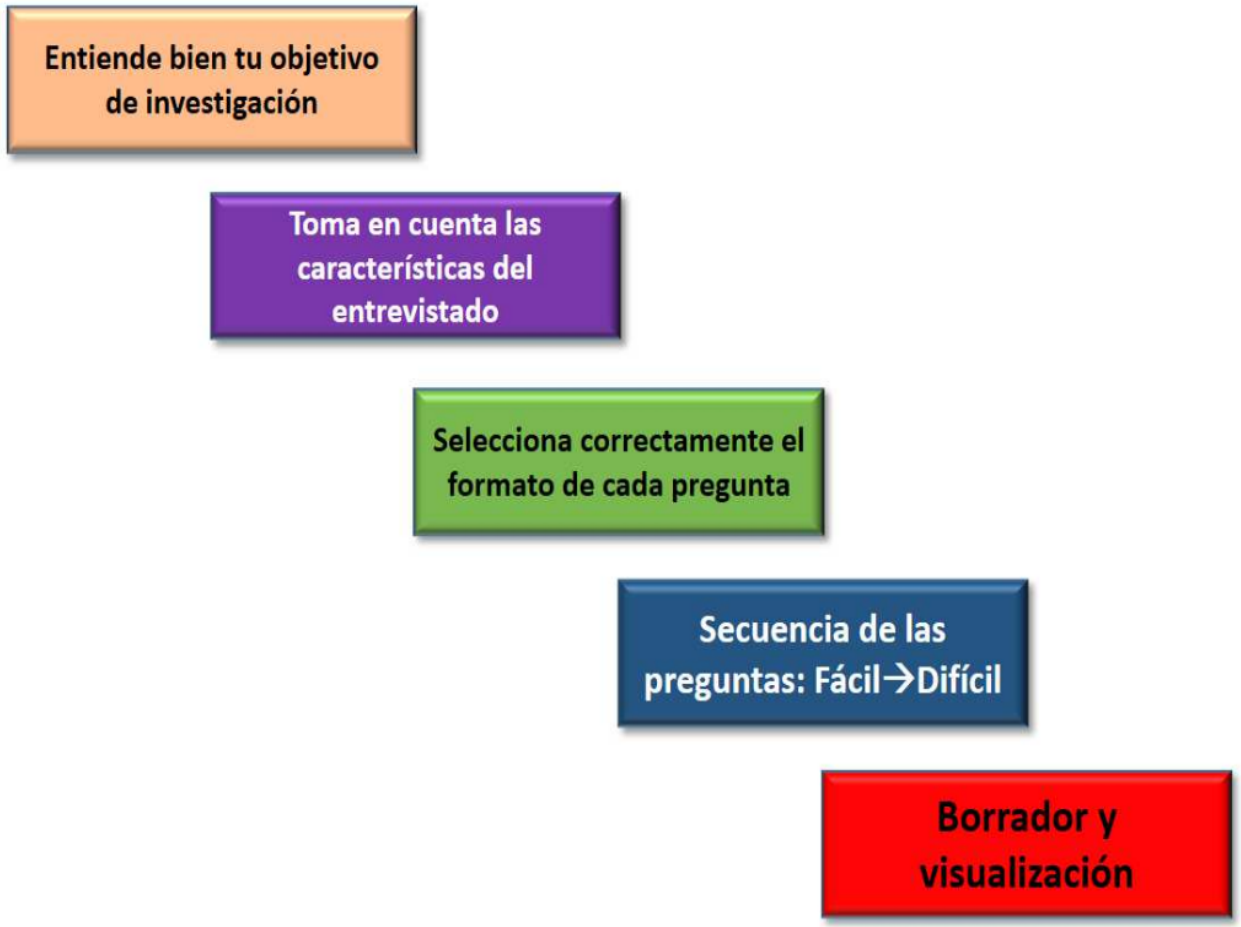
Encuestas	Censos
- Se ocupan sobretodo para el análisis cuantitativo - Son mucho menos caros y ahorran tiempo - Las respuestas se pueden ver afectadas por el entrevistador	- Se ocupan para tener un conocimiento general del ambiente poblacional - Son costosos tanto económica como temporalmente - Las respuestas no representan gran problema

¿Qué es un cuestionario?

Primero que todo, debemos entender las características principales de un buen cuestionario.

- Debe tener un conjunto de preguntas formales y específicas. Esto es, preguntas que se alineen únicamente con el objetivo de nuestra investigación y que formulados con un formato correcto.
- Las preguntas pueden ser o no estructuradas. Esto significa que pueden seguir una estructura de opciones múltiples o bien, preguntas cuyas respuestas pueden ser de cualquier tipo.
- Algo que siempre debes considerar es que tus preguntas deben incluir situaciones u objetivos específicos de tu interés de estudio. Recuerda: enfócate en los objetivos de tu investigación.

Proceso del diseño de encuestas



Tipo de preguntas I

Existen tres tipos básicos

1. Final abierto
2. Preguntas dicotómicas
3. Opción múltiple

Tipo de preguntas II

Dependiendo del formato y la seriación de las preguntas, también tenemos:

- Detección. Preguntas que sirven para saber si el entrevistado está o no en condiciones de responder todo o parcialmente el cuestionario.
- Apertura. Son preguntas iniciales que sirven para dirigir al entrevistado. Deben ser atrevidas, pero sencillas y fáciles de responder.
- Declaraciones de transición. Sirven para hacer que el entrevistado se mueva de una sección a otra del cuestionario.
- Difíciles. Se trata de preguntas sensibles o difíciles de responder y deben dejarse para el final del cuestionario.

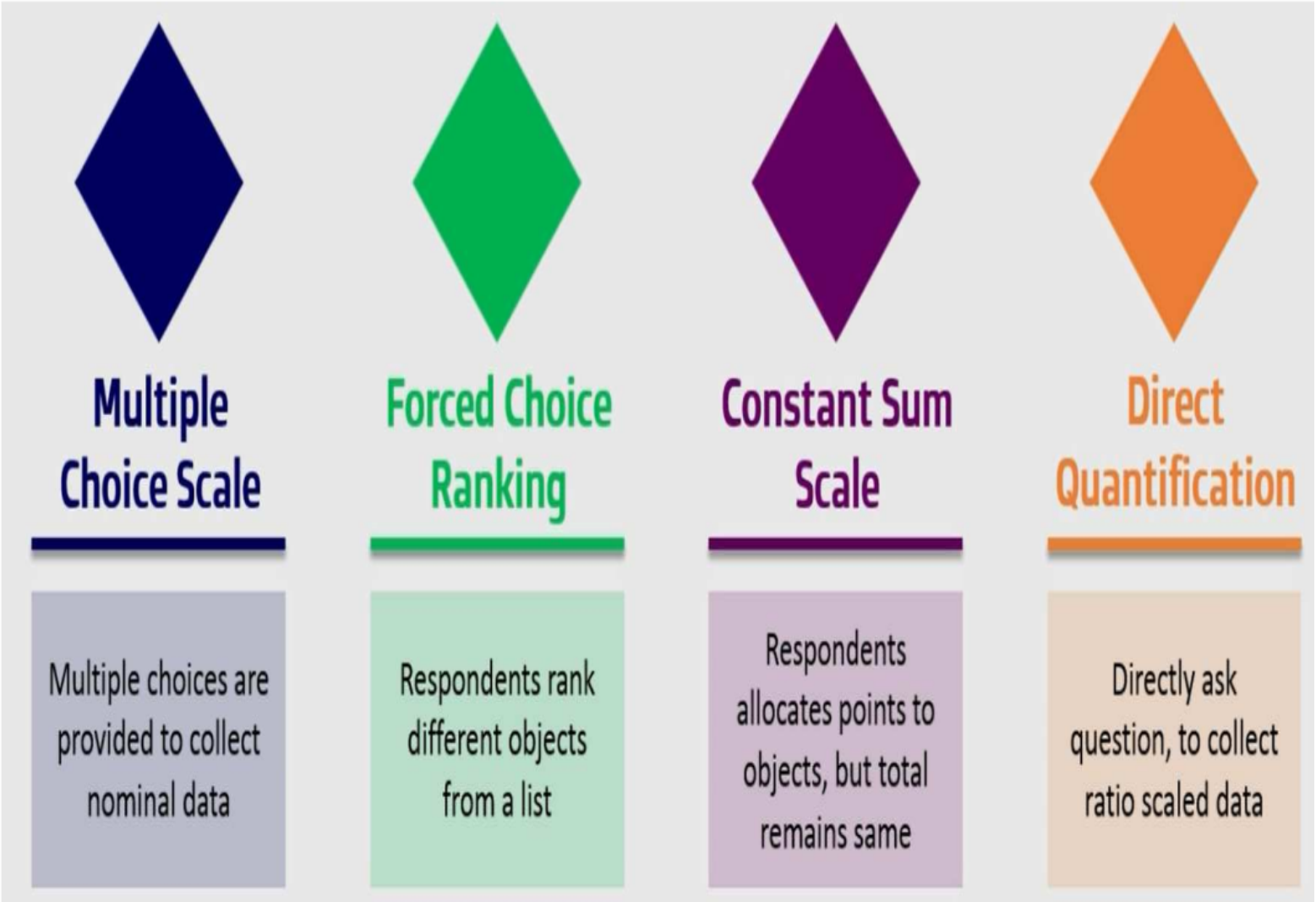
`https://www.inegi.org.mx/programas/envipe/2023/#documentacion`

Niveles de mediciones de datos

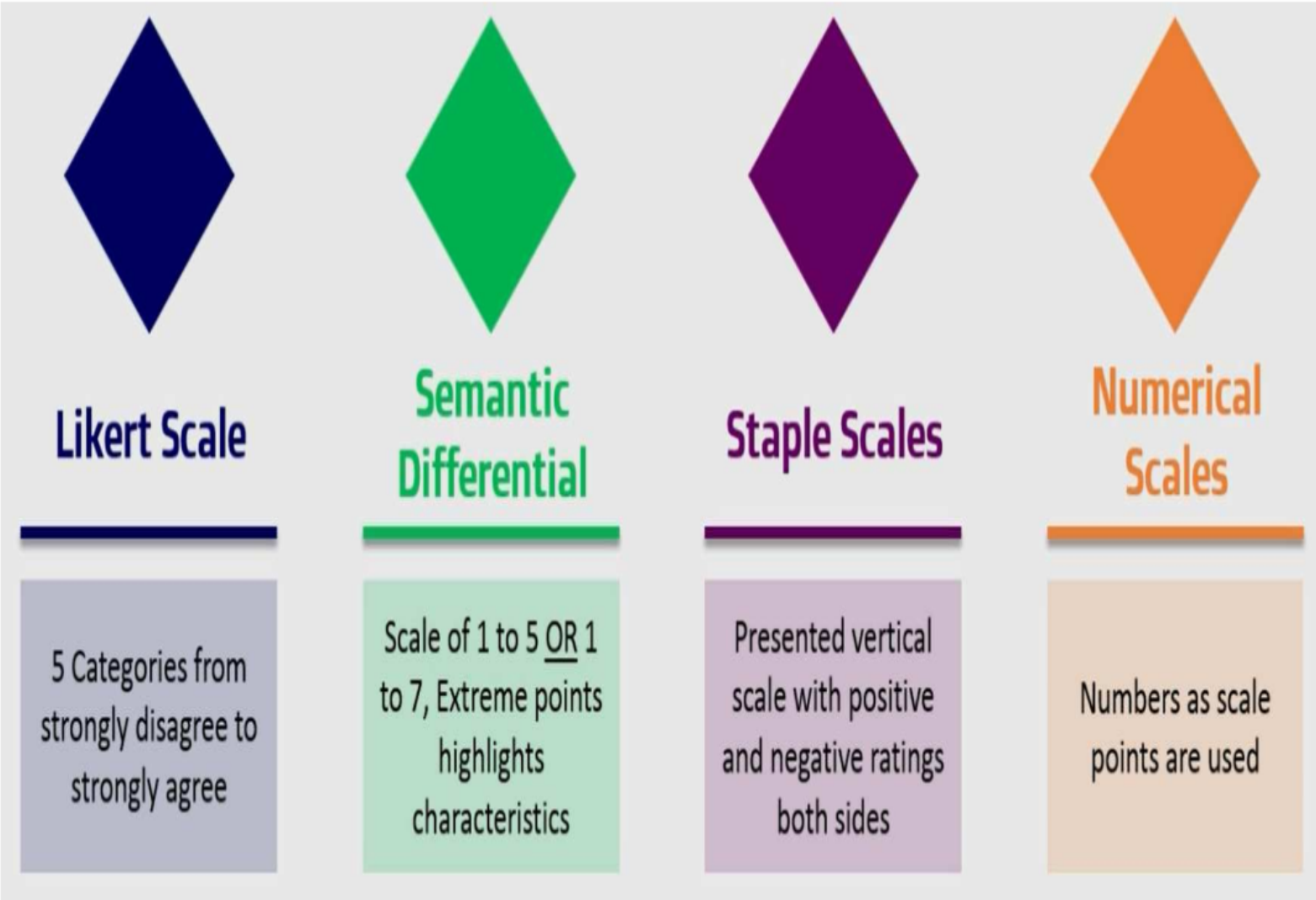
Existen cuatro niveles para clasificar nuestros datos. Ya los hemos comentado antes.

- Nominales
- Ordinales
- Intervalo
- Razón

Escalas con un solo ítem



Escalas con varios ítems



Escalas continuas

¿Qué es el muestreo?

- Se trata de una herramienta (proceso) para obtener la muestra de nuestra población de interés.
- Es bastante útil ya que disminuye el tiempo de recolección de datos y los factores económicos.
- Además, si la población de interés de estudio es muy grande, ofrece ventajas evidentes.

Diseño del muestreo

Los pasos a seguir son:

Métodos de muestreo

Se dividen en dos categorías. Lo que los diferencia es la oportunidad que tiene una unidad para pertenecer o no a la muestra.

Aleatorios	No aleatorios
- Todos tienen la misma oportunidad de ser elegidos. - No se ocupan criterios de selección.	- No todos tienen la misma oportunidad de selección. - Se basan en criterios de selección.

Muestreo aleatorio

- Simple. Todas las muestras posibles tienen la misma oportunidad de ser elegidas (si el tamaño del universo es N , hay 2^{N-1} muestras).
- Estratificado. Antes de seleccionar la muestra, se divide a la población en estratos considerando características homogéneas al interior. Cada grupo se llama estrato.
- Conglomerados. También se divide la población antes de seleccionar la muestra. Sin embargo, se hace mediante marcas o áreas geográficas. Cada grupo se llama conglomerado.
- Sistemático. Es un tipo de muestreo donde la parte aleatoria viene dada al escoger al primer individuo de la muestra.
- Multietápico. Se aplican varios de los muestreos anteriores y se arma una muestra final.

Muestreo NO aleatorio

- Por cuotas. Se seleccionan los elementos con el objetivo de cumplir con una cuota, la cual es establecida por el investigador, quien debe tener en cuenta la proporción de cada grupo. Si su población objetivo tiene un 40 % de mujeres, el muestreo por cuotas también debería incluir elementos en la misma proporción.
- Por conveniencia. El investigador elige a los miembros solo por su proximidad y no considera si realmente estos representan muestra representativa de toda la población o no.
- Por juicio. El investigador elige, a su juicio, la muestra.
- Bola de nieve. El investigador elige la muestra a partir de las sugerencias de alguna observación ya captada.

Errores de muestreo

Los errores más comunes al seleccionar una muestra son:

- Muestreo. No hay representatividad.
- Respuesta. El entrevistado no respondió correctamente.
- Sesgo por no respuesta. El entrevistado no quiso responder.
- Errores del investigador. Errores propios cometidos por el equipo de investigación.

Codificación

Se trata de la codificación en números de cada una de las variables a estudiar. Se asignan valores y códigos a cada respuesta. Por ejemplo, en la escala de Likert para una pregunta en particular, puede ser codificada con los valores 1, 2, 3, etc. Con esto, cumplimos el criterio de medibilidad de nuestros datos.

Entrada y edición de datos

Los errores más comunes al seleccionar una muestra son:

- En la entrada de datos trataremos de separar nuestra información. Se eligen, por ejemplo, características demográficas para hacerlo. La elección de las características con las que haremos la separación es muy útil dependiendo del tipo de investigación que hagamos. Luego, se incluyen todos los valores válidos disponibles y se vacían los datos en (este caso) Excel.
- Esto creará lo que se conoce como una base de datos.
- Además, probablemente tengamos que editar algunos datos. Por ejemplo, el trato de datos faltantes o las contradicciones entre los datos debidos a errores del investigador.

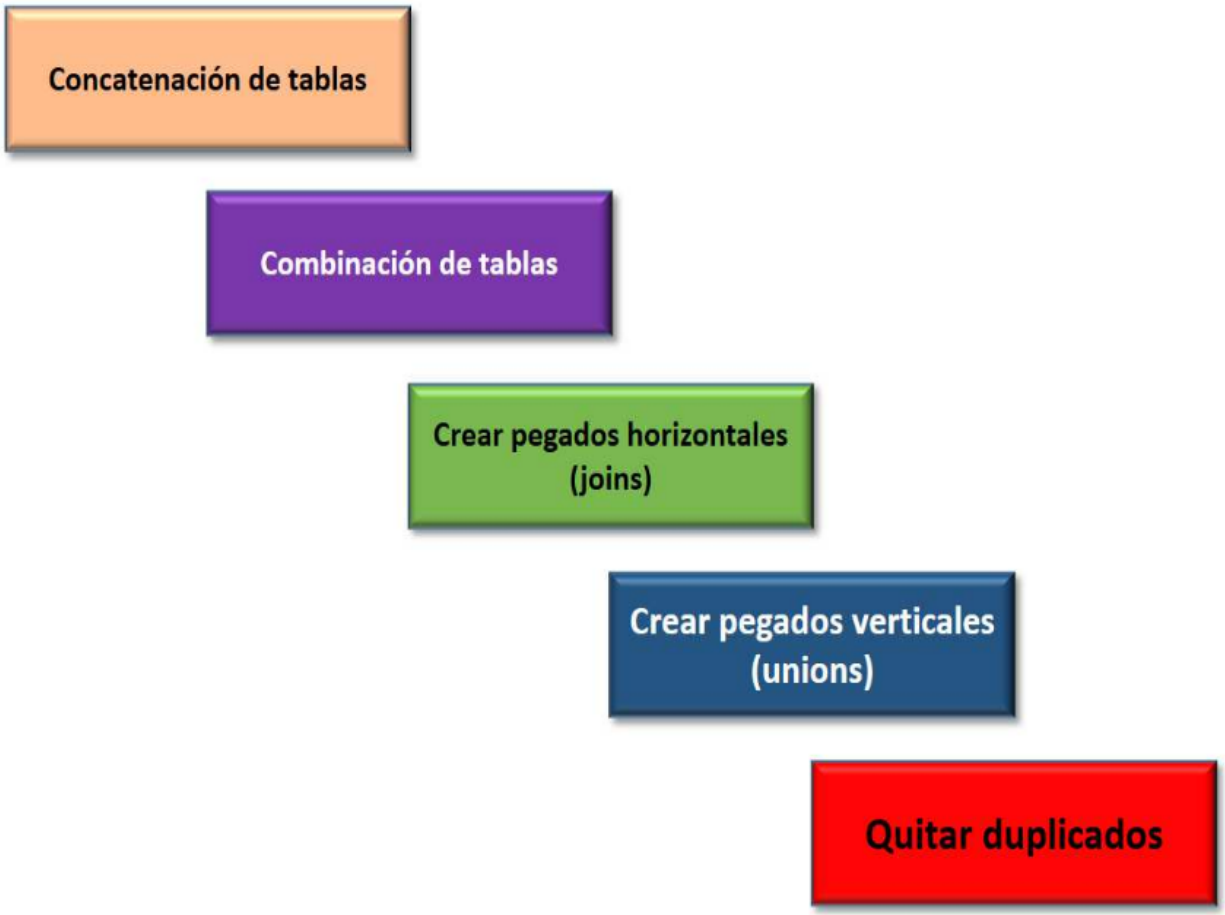
Limpieza

En general no se tienen los datos en forma pura; esto es, se pudieron cometer errores durante la recolección. Para solucionar esto, tenemos técnicas de limpieza de datos. En particular, algunos de los problemas que se buscan resolver son:

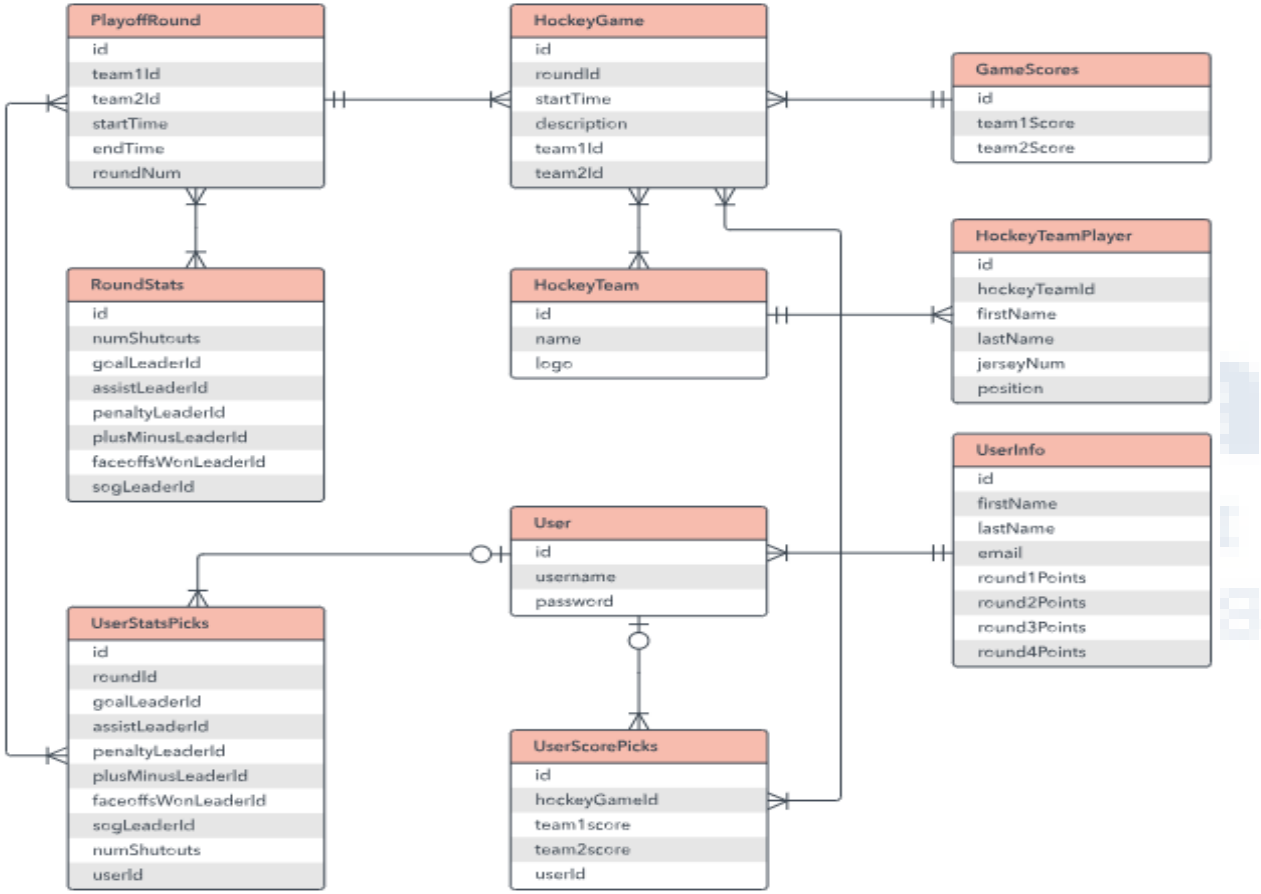
1. **Datos faltantes**
2. **Datos no coincidentes**
3. **Gestión de metadatos**
4. **Filtrado de datos**

Operaciones entre múltiples fuentes

Cuando nuestra información está diseminada en múltiples fuentes, algunas operaciones para concentrar la información son las siguientes:



Tablas relacionales



Herramientas para preparar datos

Algunos programas y lenguajes especializados en la preparación de datos son:

- Tableau
- Excel
- SQL
- R
- Python