Limpieza y adecuación de datos

¿Qué es? La limpieza de datos es el proceso por el cual descubrimos los registros erróneos (incompletos, incorrectos, inexactos, no pertinentes...) de tablas y bases de datos para luego corregirlos o eliminarlos

¿Para qué? Necesaria para hacer la base de datos útil, coherente y compatible con otras bases de datos similares en el sistema.

34 casos que es necesario contemplar:

1. **Faltan valores:** Si tienes **valores en blanco o "null**" debes conocer su significado o preguntar a tu fuente. Debe haber una opción para los que no hayan querido contestar
   Las hojas de cálculo antiguas tienen **65.536 filas:** Si recibes un dataset con ese número de filas es casi seguro que recibiste datos truncos. Llama de vuelta y pide el resto. En una versión de Excel más recientes permitían **1.048.576 filas**

2. **Las categorías fueron mal elegidas:** No están definidas respuestas como "no quiso contestar" y se ponen a cambio valores válidos y con significado. Las malas categorías también pueden excluir datos artificialmente. Revisa los **cambios de categoría y las definiciones** que tienden a ser **arbitrarias**, como la raza o etnicidad.

3. **Valores faltantes reemplazados:** Peor practica que no poner el valor es ponerlo arbitrariamente.
   Revisa que los **0 sean ceros** (**-1** también se usa a veces) y no tengan otro significado.
   Lo mismo para cualquier **secuencia de números** como 000 o **99999** o números como 2,147,483,647(máximo valor para un entero con signo en los sistemas computacionales con arquitectura de 32 bits) 4,294,967,295 (máximo para un entero sin signo de 32 bits) o teléfonos con el prefijo 555
   Revisa las f**echas** representadas **en los años:**
   1900 (la fecha predeterminada de Excel Windows, es el 1 de enero de 1900)
   1904 (la fecha predeterminada de Excel Mac, es el 1 de enero de 1904)
   1969 o 1970 (como 1970-01-01T00:00:00Z o 1969-12-31T24:59:59Z, ya que es el comienzo del registro de tiempo en Unix)
   Hay una diversidad de formas en que los datos en Excel puedan ser ingresados o calculados de manera incorrecta y terminen por dar una de estas dos fechas. Si las ves entre tus datos, probablemente se trate de un problema. En otras palabras, esto es lo que sucede cuando un sistema trata de mostrar un valor nulo o un **valor de 0 como una fecha**.
   Revisar las **ubicaciones:**
   Null Island o 0°00'00.0?N + 0°00'00.0?E o simplemente 0°N 0°E
   Código postal estadunidense 12345 (Schenectady, New York)
   Código postal 90210 (Beverly Hills, California)
   Hay valores sospechosos: Si ves alguno de los siguientes datos asegúrate de que su significado sea real. Siempre debes conocer su significado o preguntar a tu fuente.

4. **Números que fueron guardados como texto:** Hojas de cálculo con números almacenados como texto: "1,000,000" o "1 000 000" o "USD 1,000,000" con el formato de comas, unidades y espacios ingresados como caracteres.

5. **Texto que fue convertido a números**: Códigos numéricos para identificar cada sitio en el país como el 037 para el condado de Los Ángeles. Cuidado si tienen distintas longitudes y no empiezan por 0.

6. **La ortografía no es consistente:** Los datos editados a mano son los más proclives a fallas y hay que corregirlos manualmente o publicarlos como errores en tu reporte.
   Lugares donde los nombres de estados o ciudades no sean consistentes (como Los Angelos) se detectan con **Open Refine** (sugeridor de coincidencias cercanas entre valores inconsistentes en una columna). **Siempre documenta** los cambios para garantizar un buen origen de datos.

7. **El orden de las palabras no es consistente**: Con la globalización los nombres con los que trabajamos proceden de muchos países. Nos aseguraremos de que los datos de nombre y apellidos estén en el orden correcto

8. **Formatos de fecha no consistentes**: Las fechas escritas Europa o Latinoamérica siguen el formato DD/MM/AAAA mientras que los estadunidenses usan MM/DD/AAAA. Para el análisis, todos los datos deben estar en el mismo formato.

9. **Unidades sin especificar:** Indicar y no asumir las unidades de medida. Hay que preguntar a tu fuente si los datos no explicitan sus unidades (tonelada corta no es una tonelada imperial, ni una tonelada). También hay que tener en cuenta que en unidades monetarias el valor es relativo a un momento y pueden cambiar con el tiempo o estar en su propia moneda local.

10. **La inflación distorsiona los datos:** La inflación monetaria implica que con el tiempo el valor del dinero cambia. No hay manera de saber si los números fueron ajustados a la inflación sólo con mirarlo. Si obtienes datos y no estás seguro de que hayan sido ajustados, verifícalo con tu fuente.

11. **Filas o valores que están duplicados**: Necesario averiguar el por qué. ¿Es una 'corrección' que usa los mismos identificadores únicos que la transacción original?

12. **Los datos fueron capturados o editados por humanos sin validación alguna**: Para la captura solo queda tomar medidas para que no se tomen malos datos, mediante **validaciones**. Para los datos editados manualmente se empiezan a filtrar problemas cuando la persona que está editando no tiene total conocimiento de los datos originales. Para ello asegúrate de que tus datos tengan un **origen bien documentado.** La falta de éste puede ser un buen indicador de que alguien haya estado jugando con ellos (Los académicos y analistas de políticas públicas obtienen datos del gobierno con frecuencia, los manosean y luego se los redistribuyen a periodistas. Sin ningún registro de los cambios que hacen es imposible saber si son justificados) Siempre que sea posible **trata de acceder a la fuente primaria o al menos a la versión más antigua disponible** y haz tu propio análisis a partir de ella

13. **Los nombres de los campos son ambiguos:** La residencia ¿es el lugar donde vive alguien o el lugar donde paga sus impuestos? ¿Es una ciudad o un condado? Hay que tener cuidado con aquellos que significan dos o más cosas para que desde la persona que compila los datos no ingrese el valor incorrecto y el destinatario lo analice correctamente

14. **Los datos son muy burdos:** Quizás tus datos que **han sido agregados demasiado para lo que se necesita**. Tienes países, empleadores y agregados por años, pero necesitas ciudades, empleados y agregados por meses, así que pide a tu fuente algo más específico. Estos datos se agregan para proteger la privacidad de los individuos que podrían estar identificados de manera única por esos mismos datos.
    No se debe **dividir un valor anual entre 12** y llamarlo "promedio por mes". Este número no tendrá significado, puesto que los datos no suelen ser lineales, tienen ciclos y siguen tendencias.

15. **Los agregados fueron calculados con valores que faltan:** Deja fuera las filas faltantes, pero no compares agregados de dos columnas con filas faltantes sin saber si los valores faltantes pueden ser interpretados legítimamente como 0. Cuidado si tus datos ya vienen agregados.

16. **Los totales difieren de los agregados publicados:** Asegurarte que los números publicados empatan con los totales de los datos que te dieron o conocer porque difieren.

17. **El origen de los datos no fue documentado:** Saber de dónde provienen tus datos te permite conocer sus límites ya que estos están creados por variedad de individuos y organizaciones, reunidos de diferentes maneras y en diferentes formatos. Los datos de encuestas rara vez son exhaustivos. Los sensores tienen diferencias en precisión. Los gobiernos usualmente dan información sesgada. Para empeorar esta situación, estas fuentes distintas entre sí están habitualmente encadenadas.

18. **El autor no es confiable:** No publiques datos de una fuente sesgada a menos que tengas evidencia sustancial que la corrobore.

19. **El proceso de recolección es opaco:** Para evitar falsas suposiciones y errores es importante que los métodos de recogida de datos sean transparentes. Comprueba que el origen de los datos no sea sospechoso (cifras incluidas sean de una precisión real o que los datos no sean demasiado buenos para ser verdad)

20. **Los datos están mezclados con diferentes formato o anotaciones: Identifica el problema.** Asegúrate de que no hay filas extra de encabezados u otros caracteres de formato insertos entre los datos como una llave o diccionario de datos a media hoja, encabezados de filas repetidos, o la hoja de cálculo puede incluir tablas con diferente longitud en la misma página en lugar de estar separadas en distintas páginas.

21. **Los espacios al final de la línea están mal codificados**: Se debe al sistema operativo con el que se creó el archivo. Normalmente, se resuelve abriendo un archivo con un editor de texto general y volviéndolo a guardar.

22. **El texto es confuso:** Tu fuente debe ser capaz de decirte la codificación de tus datos

23. **Los datos están en PDF**: Usa Tabula o Acrobat Pro para exportar tablas de PDF a Excel.

24. **Los datos están en documentos escaneados:** OCR (reconocimiento óptico de caracteres). Recuerda validar los resultados, que deben acercarse al original.

25. **Los datos son de una precisión irreal:** Fuera del mundo de la ciencia dura, pocas cosas son medidas rutinariamente con mayor precisión que la de dos puntos decimales. Con siete decimales denota que han sido estimados a través de otros valores. Es importante ser transparentes acerca de estimados ya que usualmente están equivocados.

26. **La ley de Benford falla**: La ley de Benford es una teoría que propone que los dígitos pequeños (1, 2, 3) aparecen al comienzo de un número con mucha mayor frecuencia que números más grandes (7,8,9). La Ley de Benford puede usarse para detectar anomalías en conteos o resultados de elecciones, aunque en la práctica se aplica de manera equivocada con frecuencia. Si sospechas que un dataset fue creado o modificado para engañar a la audiencia, la Ley de Benford puede ser un primer test excelente, pero siempre verifica tus resultados con un experto antes de concluir que tus datos fueron manipulados.

27. **Hay p-hacking en los resultados**: P-hacking es la maniobra de alterar datos, cambiar estadísticas o reportar de manera intencional y selectiva los resultados de un análisis con el objetivo de mostrar hallazgos estadísticamente significativos. (Ejemplos: dejar de recolectar datos una vez que tienes un resultado estadísticamente significativo, borrar observaciones para obtenerlo o implementar numerosos análisis, pero sólo publicar aquellos que te den resulten significativos.)

28. **Hay valores atípicos inexplicables:** Echa un vistazo a los valores mínimos y mayores y asegurarte de que están **en un rango razonable**. También hacer un análisis estadístico más riguroso usando **desviaciones estándar o desviaciones medias**. Los valores atípicos pueden fastidiar tu estadística dramáticamente, especialmente si estás usando promedios. (Probablemente deberías estar usando medianas). Los valores atípicos pueden servir para encontrar buenos encabezados: 'Toma 5 mil veces el tiempo que en el resto'

29. **Un índice enmascara variaciones subyacentes:** Un Índice puede combinar varias mediciones, con condiciones que inclinan la balanza del mismo inesperadamente.

30. **La muestra no es aleatoria:** Puede darse porque la base de datos no esté completa.

31. **La muestra está sesgada:** Una muestra sesgada es engañosa y debe ser ponderada para asegurarse que cubre segmentos proporcionales de cualquier población que pudiera sesgar los resultados.

32. **La escala de tiempo fue manipulada:** Para no distorsionar la percepción hay que tener en cuenta el histórico para tener certeza de que no estarás haciendo una comparación que sería invalidada con la adición de un solo punto de datos.

33. **El marco temporal fue manipulado:** Evitar las comparaciones contra años punta, o sesgadas. Mira la evolución.

34. **El margen de error es demasiado amplio:** Lo que causa más errores de información, es el uso irreflexivo de números con márgenes de error demasiado amplios (MOE, margin of error). **Normalmente asociado a datos de encuestas**, el MOE es una medida de rangos de posibles valores verdaderos Categoría A: **nº+-variación de n (+-MOE%)**
Donde "nº" y "variación de n" pueden reportarse con seguridad. "MOE" nunca se debe utilizar para publicaciones y hay que tener **cuidado** al usar cualquier número con un **MOE mayor al 10%.**
Como regla general, cada vez que tenga los datos que son de una encuesta, pregunta por el MOE. Si la fuente no se puede decir, no vale la pena utilizar dichos datos para ningún análisis que sea serio.

Para aclaraciones o consultas contáctenos en el +34 638 729 193 o info@ñdata.com

ñdata a su servicio