Lista de Control: Auditoría de Calidad de Datos

Objetivo: Evaluar todos los datos recibidos para su nivel de lo completo, validez, consistencia, exactitud, y interpretabilidad en la forma oportuna para que:

- Se pueda establecer una comprensión básica y amplia de los datos lo antes posible.
- Ante problemas con los datos, estos puedan ser comunicados al cliente tan pronto como sea posible, para permitir una flexibilidad suficiente en el cronograma en caso de que se requiera una re-extracción de los datos.
- Se pueda gestionar de forma efectiva la expectativa del cliente desde el inicio.

Es necesario tener en cuenta lo siguiente:

- Asumir que todos los datos son erróneos hasta que se demuestre lo contrario. Cuestionar todo.
- No tratar de interpretar los datos en esta etapa. Éste es un ejercicio basado en hechos.
- No eliminar los registros o alterar los contenidos (con muy pocas excepciones), hasta que las frecuencias finales se confirman con el cliente y la disposición de los casos en cuestión esté acordada mutuamente. Absolutamente ninguna interpolación en esta etapa.
- Siempre considerar el impacto eventual de los detalles del manejo de datos. Tener justificación para todo lo que se hace.

De la recepción[1]:	**A lo mínimo[2]:**

Las Primeras 24 Horas: Verificar los hechos *muy* obvios.	• Verificar los archivos y las fuentes con las especificaciones en el documento de diseño del proyecto.
	• Verificar que los archivos se puedan descomprimir y abrir sin errores.
	• Verificar la presencia del diccionario de datos para cada archivo según sea necesario.
	  ○ OJO: las estructuras no son estrictamente necesarias a menos que las columnas en el archivo no tengan cabezales; por lo cual, la falta de la estructura o del diccionario de datos no necesariamente previene la carga de los datos. Sin embargo, los contenidos de los campos pueden no ser discernibles sólo por los nombres (esto es otro tema).
	• Identificar cualquier otro problema realmente obvio.
	• Reconocer la recepción de los datos y comunicar cualquier problema al cliente a través del correo electrónico:
	  ○ Confirmar los archivos recibidos por nombre.
	  ○ Listar todos los archivos que faltan (comparar con los requerimientos de datos según lo documentado).
	  ○ Listar todos los diccionarios de datos necesarios que falten.
	  ○ Indicar los problemas descomprimiendo y abriendo los archivos.
	  ○ Indicar cualquier otro problema *muy* obvio.
	  ○ De lo contrario, indicar al cliente que los resultados del análisis de calidad de los datos se entregarán dentro de los 7 días, si no hay problemas críticos y que cualquier duda o problema que surja en el ínterin será comunicado.

[1] Del respectivo archivo. Estos son los hitos mínimos que deben alcanzarse al final de la hora señalada, y se espera que algún trabajo para los hitos de las siguientes horas se realice en paralelo, según corresponda.
[2] En función de los datos y/o el proyecto, puede haber otras tareas por añadir a la lista.

Las Primeras 48 Horas: Verificar los hechos obvios.	• Verificar el formato de cada archivo.
	• Crear una salida ditto/xxd (o equivalente) para cada archivo.
	  ○ Esto es necesario, especialmente para identificar los caracteres no-mostrables que puedan impactar la carga de datos.
	• Inspeccionar los datos visualmente.
	  ○ A menos que los archivos sean muy pequeños, crear las muestras de registros (“head” y “tail”) para cada archivo.
	• Verificar la estructura de cada archivo: nunca se debe asumir que la estructura es la correcta.
	  ○ Archivos delimitados: visualmente verificar el orden de los campos con los datos en bruto (no sólo los cabezales; sino también ver los contenidos), y además verificar que la longitud y el formato son correctos para cada campo.
	  ○ Archivos fijos: visualmente verificar las posiciones y los formatos de los campos con los datos en bruto.
	• Verificar la cantidad de registros de cada archivo, utilizando una función propia del sistema operativo.
	  ○ NO utilizar SAS/R/etc. para realizar el conteo en este paso. Las herramientas tienen sus propias interpretaciones de una “línea”, y es importante aislar el efecto de utilizar un software.
	• Verificar la presencia de los campos con los requerimientos de los datos.
	• Comenzar a cargar los datos en el entorno analítico.
	  ○ Al cargarlos, verificar que la cantidad de registros cargados coincida con la cantidad anteriormente obtenida de registros.
	  ○ Determinar el mejor “informat” en base a la información y la verificación de las estructuras, los mensajes que produce SAS en el log, el entendimiento de los datos, y el sentido común.
	  ○ ¡NO utilizar PROC IMPORT!
	  ▪ Se basa en un supuesto en donde el archivo es limpio, consistente, y relativamente pequeño—en práctica, esto es raro.
	  ○ Verificar la longitud y el formato de cada campo, incluso si están dados en la estructura y/o el diccionario de datos.
	  ▪ La información documentada NO siempre corresponde.
	  ▪ El software (como SAS) puede tener su propia interpretación que varía de la definición propia de la fuente de datos.
	  ▪ Programáticamente obtener la longitud para los archivos delimitados.
	  ▪ Las consideraciones incluyen: carácter vs. numérico, cantidad de decimales, decimales implícitos vs. explícitos, decimales empaquetados (esp. los archivos provenientes del mainframe), otros meta-caracteres no-mostrables, etc.
	  ○ Puntos adicionales por considerar:
	  ▪ Si el mismo campo existe en varios archivos (ej. claves para realizar joins), ¿se deberían cargar siempre con el mismo formato y la misma longitud? ¿Y si los campos son similares pero no exactamente iguales?
	  ▪ ¿Cómo impactan los campos de uno a otro a través de los archivos distintos?
	  ▪ ¿Qué implica cargar cada campo en el formato que he elegido?
	  ▪ ¿Por qué cargo esta línea, archivo, variable, etc., de esta manera?
	• Identificar los problemas en la carga de datos.
	• Verificar la presencia de las claves correctas en cada archivo para identificar registros distintos y para realizar los joins (según sea aplicable).
	  ○ En base de las descripciones y la mirada inicial de los campos incluidos, ¿se pueden juntar las tablas (en principio)?
	• Identificar cualquier otro problema obvio.
	• Comunicar al cliente dudas o problemas que han surgido por correo electrónico.

De la recepción[1]:	**A lo mínimo[2]:**

Las Primeras 72 Horas	• Concluir la primera pasada de la carga de los datos.
	• Comprobar si hay duplicados exactos.
	  ○ NO eliminar los duplicados, a menos que se conozca exactamente cómo surgieron y que se confirme que es apropiado eliminarlos.
	• Identificar las distintas claves que identifican los diferentes registros.
	  ○ El propósito de este ejercicio es entender el nivel de registro (i.e., qué concepto representa cada registro), ya que es frecuentemente distinto a lo que se entiende.
	  ○ Comenzar con un entendimiento de cuál combinación de campos razonablemente representaría un registro en el archivo en cuestión y su jerarquía.
	  ○ Utilizar el sentido común. El hecho de que una combinación de campos identifica de forma distinta un registro no significa que los campos sean válidos como las variables de identificación.
	• Cuantificar cuán distintos son los valores en cada variable (o variable candidato) de identificación.
	• Investigar los “duplicados” por varios niveles de identificación.
	  ○ ¡NO eliminarlos!
	  ○ Intentar responder: ¿cuáles son las diferencias entre los “duplicados”?
	• Comprobar si los archivos pueden unirse.
	  ○ ¿Las claves coinciden entre los archivos? ¿Cuán bien?
	  ○ Éste es el próximo paso al día anterior en el cual simplemente se evaluó si es teóricamente posible.
	• Identificar cualquier otro problema en la carga de los datos.
	• Comprobar que los archivos recibidos son generalmente razonables en el contexto del negocio:
	  ○ Volumen
	  ○ Tasa preliminar del evento de interés, o promedio de la métrica de interés, según sea apropiado para el proyecto (u otra indicación relacionada)
	  ○ Conteo de las subpoblaciones
	  ○ Las tareas específicas variarán por proyecto.
	• Comunicar al cliente dudas o problemas que hayan surgido por correo electrónico.

Los Primeros 7 Días	• Finalizar la carga de los datos.
	  ○ A lo mínimo, el log de SAS debe estar libre de "error", "warning", "not initialized" (comprobar tanto mayúsculas como minúsculas) -ninguno de ellos son aceptables.
	• Frecuencia completa de cada campo.
	  ○ Evitar atajos; NO asumir que un agrupamiento estándar es apropiado ya que muchas veces puede ocultar problemas de datos. En su lugar, correr una frecuencia de los valores brutos con una cantidad limitada de registros (100, 1,000, 10,000, etc.) hasta que tenga una idea de cómo están los datos, y después aplicar un agrupamiento según sea necesario.
	  ○ Trabajar en iteraciones múltiples. Esto es una oportunidad importante para observar los valores brutos de todas las variables.
	• Un resumen de "ocho números" de cada variable numérica *para la cual esto tiene sentido* (#nulos, mínimo, 1ro percentil, 1ro cuartil, mediano, 3ro cuartil, el percentil 99, máximo).
	  ○ Identificar cualquier valor potencialmente por defecto (ej. una serie de 9s, de 0s, etc.)
	  ○ Considerar si los valores negativos tienen sentido para cada variable.
	• Comparar con los requerimientos de datos en mayor detalle.
	• Verificar que los valores en las distribuciones tienen sentido.
	• Verificar que las definiciones de los valores codificados se encuentren documentados para todos los valores tales.
	  ○ Utilizar las frecuencias y etiquetar todos los valores codificados como se encuentren en los diccionarios de datos y otros documentos relevantes.
	  ○ Identificar en cuáles de los valores faltan: descripciones, inconsistencias, etc.
	  ○ Incluir todos los comentarios directamente a la salida de la frecuencia para ser entregada al cliente (ver abajo).
	• Verificar que las distribuciones de las variables importantes sean razonables.
	• Verificar la consistencia de los campos dentro de y a través de los archivos (ej., los campos similares provenientes de las fuentes distintas, la consistencia entre las clases y las subclases, etc.), en la medida posible.
	• Entregar un informe completo de auditoría de calidad de datos, con los comentarios y las consultas identificadas (un archivo Excel es generalmente aceptable):
	  ○ Listado de los nombres de los archivos con la descripción y la cantidad de registros para cada uno.
	  ○ Métricas de duplicados y la calidad de joins.
	  ○ Frecuencia/distribución de cada variable, con la definición de la variable y con la descripción de cada valor.
	  ○ Todas las consultas sobre los datos.
	  ○ Comentarios sobre cualquier cosa incompleta, rara, no-esperada, no-conocida, no-entendida, no-explicada, etc.
	  ○ Comentarios de observaciones iniciales de los datos, especialmente para validar el entendimiento de los datos.
	• Solicitar al cliente la fecha esperada para la entrega de las respuestas a las consultas pendientes.
	• Aceptar la calidad de los datos para el análisis si no hay errores no-corregibles; desarrollar y comunicar un plan de acción con un cronograma si hay errores no-corregibles y/o dudas en la factibilidad de corregir (ej. requiere respuestas del cliente para resolver).
	  ○ Si la aceptación de la calidad del archivo en cuestión depende de la recepción de otros archivos, aceptarla condicionalmente, y desarrollar y comunicar un plan con un cronograma.

De la recepción[1]:	**A lo mínimo[2]:**
Consultas por realizarnos antes del día 7:	• ¿Tengo las descripciones para el 100% de los campos? • ¿Tengo todas las definiciones? • ¿Entiendo cómo los archivos se relacionan uno con otro? • ¿Entiendo todo en los datos? ¿Puedo explicar todo en los datos? • ¿Los valores tienen sentido? • ¿Existen valores raros o no-esperados? • ¿Existen variables que no se deban utilizar desde el punto de vista práctico? • ¿Estoy listo para aceptar la responsabilidad sobre la calidad de datos? (Una vez que la aceptas, todos los errores “fatales” son tus problemas, ¡no del cliente!)
Consultas por realizar al cliente al día 7:	• ¿Respuestas a los comentarios y observaciones (no sólo consultas)? • ¿Existen algunas variables que no tienen sentido? • ¿Existen algunas variables que no se deban utilizar? ¿Por qué?