REICE
Revista Electrónica de Investigación en Ciencias Económicas
Abriendo Camino al Conocimiento
Área de Conocimiento de Ciencias Económicas y Administrativas
Universidad Nacional Autónoma de Nicaragua, Managua (UNAN-Managua)

Edición Especial, IIS septiembre 2024 **REICE ISSN: 2308-782X**
https://revistas.unan.edu.ni/index.php/reice
revista.reice@unan.edu.ni

Limpieza de datos en el análisis financiero

Data cleaning in financial analysis

Fecha de recepción: julio 18 de 2024
Fecha de aceptación: agosto 30 de 2024
DOI: https://doi.org/10.5377/reice.v1i2.18869

Orlando José Cruz Orozco
Docente del Departamento de Economía
Área de Conocimiento de Ciencias Económicas y Administrativas
Universidad Nacional Autónoma de Nicaragua, Managua (UNAN-Managua)
Correo: orlando.cruz@unan.edu.ni
ORCID: https://orcid.org/0000-0002-4896-6301

Resumen

La limpieza de bases de datos fue crucial para un análisis financiero efectivo, mejorando la calidad de los datos, reduciendo errores y optimizando recursos. Este proceso incluyó la identificación y eliminación de datos irrelevantes y duplicados, así como la corrección de errores y datos inconsistentes. Se usaron herramientas de gestión de bases de datos, duplicadores y validadores de datos, respetando la privacidad y protección de datos personales según regulaciones aplicables. Los resultados mostraron que la limpieza de bases de datos es esencial para la integridad y calidad de los datos en el análisis financiero, fundamental para la sostenibilidad del negocio. Al finalizar, se mejoró la comodidad con el análisis y la información obtenida, aumentando la eficiencia y productividad en los análisis financieros. Se descubrió que muchos datos no estaban estructurados, dificultando la optimización del procesamiento. La generación continua de datos complicó la recopilación y administración. El análisis permitió descomponer y sintetizar la definición de limpieza de datos, métodos y técnicas para su ejecución y mecanismos de procesamiento. También se presentaron herramientas diversas para administrar la información, aportando ideas para innovar, crecer a largo plazo y crear estrategias más efectivas, mejorando la eficiencia y productividad en los análisis financieros.

Palabras claves: Herramientas de análisis, Métodos y Técnicas de Limpiezas de datos, Análisis Financiero.

Summary

Cleaning databases was crucial for effective financial analysis, improving data quality, reducing errors and optimizing resources. This process included identifying and removing irrelevant and duplicate data, as well as correcting errors and inconsistent data. Database management tools, data duplicators and validators were used, respecting the privacy and protection of personal data according to applicable regulations. The results showed that database cleaning is essential for the integrity and quality of data in financial analysis, fundamental for business sustainability. Upon completion, comfort with the analysis and information obtained was improved, increasing efficiency and productivity in financial analysis. Much data was found to be unstructured, making it difficult to optimize processing. Continuous data generation complicated collection and management. The analysis allowed us to decompose and synthesize the definition of data cleaning, methods and techniques for its execution, and processing mechanisms. Various tools were also presented to manage information, providing ideas to innovate, grow in the long term and create more effective strategies, improving efficiency and productivity in financial analysis.

Keywords

Analysis tools, Data Cleaning Methods and Techniques, Financial Analysis

Introducción

En la era digital actual, la labor del economista ha evolucionado notablemente debido al crecimiento exponencial de la información disponible a través de Internet y redes sociales. Tradicionalmente, los economistas se centraban en el estudio de modelos económicos y escuelas de pensamiento, utilizando estadísticas y matemáticas para sus análisis. Sin embargo, el entorno moderno exige que los economistas actúen como filtros y emisores de información, organizando e interpretando datos financieros, a través de aplicaciones, software e instrumentos computacionales.

La conectividad de dispositivos ha generado un flujo constante de información en bases de datos, convirtiendo a los datos en un insumo fundamental para las estrategias de empresas, gobiernos e individuos.

La gestión adecuada de estos datos es crucial para obtener análisis y conclusiones precisas que impactan directamente en la economía. No obstante, a pesar de la abundancia de datos disponibles a nivel mundial, la calidad de estos datos puede ser deficiente, especialmente en situaciones de emergencia como pandemias o desastres naturales. Los datos incompletos, inconsistentes o erróneos dificultan la toma de decisiones rápidas y precisas, lo que puede resultar en pérdidas económicas significativas e ineficiencias.

Históricamente, los economistas se han centrado en la teoría económica y el análisis estadístico. Con la revolución digital y el auge del big data, ha surgido una nueva dimensión en la que los datos deben ser gestionados de manera técnica y precisa. La calidad y fiabilidad de los datos siguen siendo un desafío constante, y la limpieza de datos se ha convertido en un proceso esencial para quienes manejan grandes volúmenes de información. Este proceso implica identificar y corregir datos erróneos, incompletos, duplicados o irrelevantes, asegurando que los análisis financieros sean precisos y confiables.

El problema central es que, a pesar de la abundancia de datos disponibles a nivel mundial, la calidad de estos datos puede ser deficiente, especialmente en situaciones de emergencia como pandemia o desastres naturales. Los datos incompletos, inconsistentes o erróneos dificultan la toma de decisiones rápidas y precisas, lo que puede resultar en pérdidas económicas significativas, ineficiencias y decisiones equivocadas. La necesidad de limpiar y gestionar adecuadamente los datos se vuelve imperativa para evitar consecuencias negativas y maximizar los beneficios de la información disponible.

La limpieza de datos es un proceso esencial para quienes manejan grandes volúmenes de información. Este proceso implica identificar y corregir datos erróneos, incompletos, duplicados o irrelevantes, asegurando que los análisis financieros sean precisos y confiables. Subirats et al (2019)

La gestión de datos debe enfocarse en generar, almacenar, procesar, analizar y comprender datos mediante herramientas y procesos ágiles. Además, se destaca la importancia de desarrollar una cultura de datos que mejore la toma de decisiones y reconozca la importancia organizativa de los datos para agilizar decisiones mediante un gobierno de información oportuno.

El presente artículo busca informar sobre las mejores prácticas de limpieza de datos y su aplicación en análisis financieros. Pretende evaluar rendimientos, identificar fortalezas y debilidades, y desarrollar estrategias fundamentadas en datos limpios y confiables, esenciales para la democratización de la información, automatización y desarrollo de habilidades que conduzcan a resultados transformadores y a la sostenibilidad empresarial y así alcanzar la mayor exactitud posible, que permita determinar ciertas decisiones de consistencia y validez de la información. (Müller et al, 2003).

Materiales y métodos

En este artículo, se examinaron y analizaron diversas fuentes de información previamente existentes, incluyendo libros, artículos académicos, investigaciones previas y documentos relevantes. El interés fue comprender, sintetizar y evaluar el estado actual del conocimiento sobre la gestión de limpieza de datos en el análisis financiero.

Por tanto, esta investigación sería de carácter aplicada y descriptiva. Es aplicada porque busca resolver problemas prácticos relacionados con la calidad de los datos en el análisis financiero, y es descriptiva porque se enfoca en describir las mejores prácticas, técnicas y herramientas utilizadas para la limpieza de datos. Además, podría involucrar un enfoque cuantitativo para evaluar el impacto de la limpieza de datos en la precisión y eficacia de los análisis financieros.

El "universo" de estudio comprende todas las fuentes de información disponibles sobre el tema, mientras que la "muestra" incluye una selección de las fuentes más pertinentes y representativas. Estas fuentes fueron elegidas para proporcionar una visión completa y precisa del estado de la gestión de limpieza de datos en el ámbito financiero.

Se utilizaron bases de datos académicas, como el número de estudiantes matriculados y egresados en los últimos años en economía agrícola, que desarrollaron temas en seminarios de graduación, basado en estadísticas, y bases de datos institucionales y de investigación reconocidas, (Ministerio Agropecuario, Banco Central, entre otros) así como motores de búsqueda académicos. Estas fuentes fueron utilizadas por profesores y estudiantes de tercero, cuarto y quinto año de la carrera de economía agrícola.

Se enfocaron en encontrar literatura relevante relacionada con la gestión de limpieza de datos en la economía y el análisis financiero, para poder aprender la importancia de la información organizada y ordenadas además se enumeran las principales técnicas de análisis que permiten explorar los datos, con el objetivo de identificar tendencias y patrones de interés. Desde el análisis estadístico descriptivo e

inferencial, pasando por la regresión, la correlación, los análisis de supervivencia y enumerando algunos modelos supervisados y no supervisados. Asimismo, se comenta brevemente la representación visual de los resultados como método adicional de análisis. Subirats Laia et al (2019)

Se incluyeron fuentes que abordan directamente la gestión de limpieza de datos y su aplicación en el análisis financiero, igualmente se priorizaron publicaciones en revistas académicas de alto impacto y libros de editoriales reconocidas, y encontramos que cuando los datos no cumplen con criterios de calidad, se consideran “sucios” debido a que cualquier información inválida, inconsistente, incompleta o incorrecta, atenta contra la integridad de cualquier tipo de base de datos (Kitlas, 2012).

Aunque son difíciles de medir, los costos de datos erróneos o inexactos son generalmente mayores a los costos de ingreso de la información, sobre todo cuando se procesan y convierten en conocimiento para la toma de decisiones (Ahmed et al, 2010).

También se vieron estudios y artículos publicados en los últimos quince años para asegurar la relevancia y actualidad de la información, que permitió tener mayores conocimientos de los temas ligados a las limpiezas de datos.

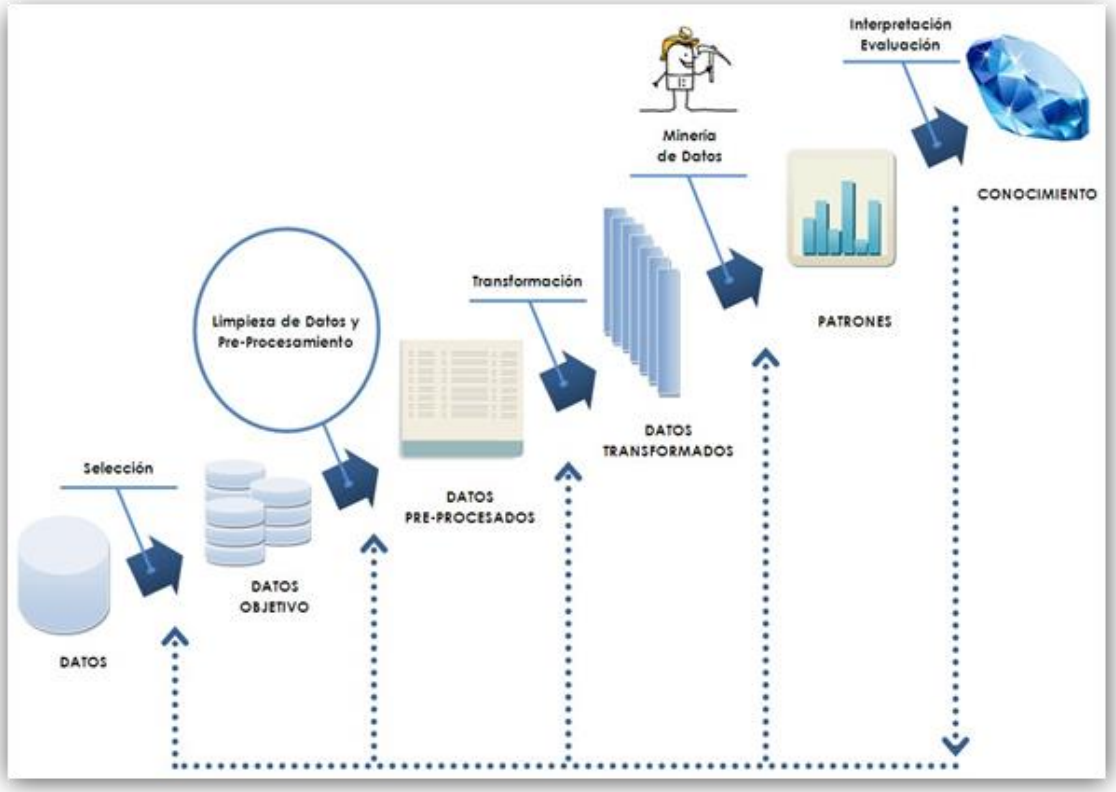
Los contenidos de las fuentes seleccionadas fueron analizados y sintetizados para identificar tendencias, enfoques comunes y desafíos en la gestión de limpieza de datos en el ámbito financiero, Es importante recordar que la limpieza trata con problemas de datos una vez que estos han ocurrido. No compensa problemas de muestreo, diseños metodológicos inadecuados, ni pobres objetivos de análisis Castillo Noguera María et al 2015

Si se identifica que parte de las deficiencias de la base de datos se relacionan con procesos anteriores al ingreso de la información, también será importante retroalimentar a ese nivel y documentarlo para que se tome en cuenta en las siguientes fases del proceso de generación de conocimiento.

La identificación de valores atípicos es importante, ya que sigue siendo un problema complejo. Las técnicas estadísticas y de aprendizaje automático son útiles, pero aún requieren mejoras para ser completamente efectivas, donde se requiere la protección de la privacidad de los datos económicos, especialmente en un entorno de creciente regulación de la privacidad. Esto requiere un equilibrio entre la transparencia y la protección de la información sensible.

Metodología de Limpieza de Datos

Figura 1. Proceso de descubrimiento de conocimientos en bases de datos



Modificado de: Fayyad, Piatetsky-Shapiro, & Smyth, 1996

La limpieza de datos es mucho más que una actualización de la base con elementos correctos; se debe observar el objetivo por el cual se hace esto (se muestra en la Figura 1). Castillo Noguera María et al 2015. Implica también descomponer y volver a ensamblar cierta información (Maletic & Marcus, 2000). Van den Broeck, Argeseanu Cunningham, Eeckels y Herbst (2005) plantean que tres fases componen

el proceso de limpieza de datos: exploración, diagnóstico y edición. Estas fases se retroalimentan entre sí en ciclos que pueden repetirse. En la exploración o cribado se procura identificar falta o exceso de información, inconsistencias, datos extremos, patrones extraños y resultados sospechosos.

Posteriormente, se procede a identificar la naturaleza del problema de ciertos puntos, patrones y estadísticas. Se diagnostican errores, datos ausentes, valores atípicos genuinos y datos normales auténticos, señalando o marcando la información que requiere atención. Con base en este diagnóstico, se editan las bases de datos mediante la corrección, eliminación o mantenimiento de los datos, y se retroalimenta el sistema con estas modificaciones.

El marco general del proceso de limpieza de datos implica:
Definición y Determinación de Errores

- Identificar los tipos de errores posibles en los datos.
- Establecer criterios y estándares para detectar errores.
- Crear una lista de control de errores comunes y sus características.

Búsqueda e Identificación de Casos de Error

- Utilizar herramientas y técnicas de análisis de datos para detectar errores.
- Implementar procesos de auditoría y revisión periódica de los datos.
- Marcar y categorizar los casos de error identificados.

Corrección de Errores

- Aplicar métodos de limpieza de datos para corregir errores.
- Reemplazar datos incorrectos con valores precisos y verificables.
- Eliminar datos duplicados o irrelevantes.

Documentación de Casos y Tipos de Error

- Registrar todos los errores identificados y las correcciones aplicadas.
- Mantener un historial de cambios y ajustes realizados en la base de datos.
- Crear informes detallados sobre los tipos de errores y su frecuencia.

Actualización de Mecanismos de Entrada para Evitar Errores Futuros (Ahmed & Aziz, 2010).

- Revisar y mejorar los procesos de ingreso de datos para reducir errores.
- Implementar validaciones y controles automáticos en los sistemas de entrada de datos.
- Capacitar al personal en buenas prácticas de gestión de datos y prevención de errores.

La documentación de las rutinas de limpieza es parte del manejo transparente y apropiado de datos. Idealmente debería contarse con un registro de herramientas de tamizaje, procedimientos y diagnósticos utilizados para distinguir errores de valores verdaderos, formas de señalar o marcar elementos a revisar, reglas de decisión aplicadas para la edición de los datos, tipos y tasas de error encontrados, tasas de eliminación de valores y de corrección al menos para las principales variables–, justificación de cambios, impacto en los resultados al quitar valores extremos o fuera del patrón esperado, fechas de trabajo y personal involucrado (Van den Broeck et al., 2005).

La limpieza de datos requiere al menos a una persona o a un equipo para que lea y verifique la exactitud de la información. Estas personas deberían ser expertas en el campo específico del conjunto de datos, para que puedan realizar con mayor confiabilidad correcciones en bases de gran tamaño. Hacerlo manualmente toma un tiempo considerable. Algunas rutinas pueden hacerse de modo semiautomático utilizando paquetes estadísticos. Castillo Noguera María et al (2015)

Se han desarrollado herramientas como Centrus Merge/Purge, Data Tools Twins, Data Cleansing DataBlade, DataSet V, DATASTAGE, DECISIONBASE, DeDuce, DeDupe, dfPower, DoubleTake, ETI Data Cleanse, Holmes, i.d. Centric, Integrity, matchIT, matchmaker, NADIS Merge/Purge Plus, NoDupes, POWERMART, Pure Integrate, PureName PureAddress, Quick Address Batch, reUnion and MasterMerge, SAGENT SOLUTION PLATFORM, SSA-Name/Data Clustering Engine, Trillium Software System, TwinFinder, Ultra Address Management y WAREHOUSE ADMINISTRATOR, que dependiendo del tipo de datos, facilitan la validación de información personal y la identificación de casos duplicados (Ahmed & Aziz, 2010 ;

Maletic & Marcus, 2010 ; Maydanchik, 2007 ; Müller & Freytag, 2003 ; Rahm & Do, 2000).

También, algunos procedimientos de limpieza pueden llevarse a cabo utilizando *software* que permite la programación de reglas y configuración de rutinas. Con un algoritmo simple pueden corregirse errores similares simultáneamente. Aunque esto ahorra tiempo, el reto está en el esfuerzo analítico requerido para entender los patrones y determinar las soluciones, así como en el hecho de que las reglas de calidad de datos son interdependientes. Corregir un elemento de un conjunto de valores, puede resultar en la violación de otras reglas establecidas. En un intento por corregir un error original pueden inducirse nuevos errores, si se falla en ver las distintas relaciones entre los valores. Entre los programas utilizados para corregir anomalías y realizar procesos de limpieza en las bases de datos están AJAX, FraQL, Potter'sWheel, ARKTOS e IntelliClean (Ahmed & Aziz, 2010; Maletic & Marcus, 2010; Maydanchik, 2007; Müller & Freytag, 2003).

Por tanto, una limpieza completa incluye un agregado de operaciones que pueden resumirse de la manera siguiente;

Auditoría de datos: análisis de atributos de los datos y de la colección de datos de la cual se genera información como mínimos y máximo rango, frecuencia de valores, varianza, unicidad, ocurrencia de valores, nulos patrones, típicos específicos y generales. Los resultados contribuyen a la definición de criterios de consistencia, formatos de dominio e indicadores de posibles errores y las características con las que se presentan.

Especificación del flujo de trabajo: determinación de los procedimientos específicos a aplicar para modificar errores, así como para detectar y eliminar anormalidades en los datos. La especificación de métodos de limpieza y corrección debe considerar las posibles causas de los valores erróneos. (Ahmed et al, 2010)

Ejecución del flujo del trabajo: verificación e implementación del flujo de trabajo; el objetivo es alcanzar la mayor exactitud posible. La aplicación eficiente las distantes

operaciones y las reglas, deben ser factibles independientemente del tamaño del conjunto de datos. La interacción con expertos resulta fundamental puesto que el criterio profesional determinará ciertas decisiones de consistencia y validez de la información. (Müller et al, 2003)

Post procesamiento/control: comprobación de los resultados obtenidos para verificar la exactitud de las operaciones ejecutadas en la base de datos. Se llevan a cabo procedimiento de control para localizar información no corregida. El conjunto de datos resultante es objeto de un nuevo ciclo de limpieza.

Müller y Freytag (2003) señalan que la limpieza de datos nunca termina por completo, algunas anomalías son difíciles de localizar y muchas veces se identifican hasta que las bases de datos se encuentran en procesos posteriores de análisis. Castillo Noguera María et al 2015

Según las aplicaciones previstas de la información y según los recursos disponibles, se debe determinar la cantidad de tiempo y esfuerzo que se invertirá en este proceso, considerando los criterios mínimos de calidad esperados y los criterios mínimos para asegurar la compatibilidad con otros conjuntos de información en el sistema.

Recientemente están las aplicaciones, software, que relacionadas a las estadísticas y a las matemáticas pueden mejorar, limpiar y hasta ordenar los datos y son las siguientes:

Para la eliminación de Valores Atípicos, se utilizaron técnicas estadísticas como el cálculo de desviaciones estándar y visualización gráfica (por ejemplo, diagramas de caja y bigotes) para identificar y eliminar valores atípicos que podrían distorsionar los análisis.

En la normalización de Datos, se aplicaron métodos de normalización, incluyendo la normalización min-max y Z-score, para asegurar que los datos siguieran un rango específico o una distribución particular, facilitando comparaciones y análisis consistentes.

Igualmente, en las Corrección de Errores, se identificaron y corrigieron errores comunes en los datos, como registros duplicados y formatos inconsistentes. Se utilizó una combinación de revisión manual y algoritmos de corrección automática para asegurar la calidad de los datos.

En Cuanto a las Herramientas Utilizadas, tenemos:

Software de Limpieza de Datos:

- OpenRefine: Herramienta de código abierto utilizada para la limpieza y transformación interactiva de datos.
- Trifacta Wrangler: Utilizada para la preparación de datos con técnicas de aprendizaje automático.
- Talend Data Quality: Proporciona una amplia gama de funcionalidades para la limpieza y preparación de datos.

Bases de Datos y Motores de Búsqueda:

- Bases de datos académicas y motores de búsqueda como Google Scholar y JSTOR para identificar y seleccionar literatura relevante.

Lenguajes de Programación y Librerías:

- Python con bibliotecas como Pandas para la manipulación y limpieza de datos.
- SSPS

Resultados y discusión

Resultados

La revisión bibliográfica sobre la gestión de limpieza de datos en el análisis financiero ha arrojado varios hallazgos significativos que subrayan la importancia de este tema en la economía moderna, proporcionando una visión integral de este campo en constante evolución. Los resultados más destacados:

Se han identificado diversos enfoques para la limpieza de datos, desde técnicas estadísticas avanzadas hasta métodos de aprendizaje automático. Estas técnicas son esenciales para mejorar la calidad de los datos económicos y financieros. Por ejemplo, técnicas como la normalización de datos, la eliminación de valores atípicos y la corrección de errores son ampliamente utilizadas para asegurar la integridad y precisión de los datos.

La calidad de los datos económicos es crucial para la toma de decisiones precisas en áreas como la planificación financiera, la evaluación de riesgos y el análisis de tendencias económicas. La revisión destaca que los datos limpios y confiables son fundamentales para apoyar decisiones informadas y evitar errores costosos.

La falta de limpieza y calidad en los datos económicos puede llevar a decisiones erróneas y costosas tanto en el ámbito empresarial como gubernamental. La revisión del estudio previo revela que cuando los datos inexactos o incompletos, los análisis resultantes pueden ser incorrectos. Esto significa que las conclusiones y recomendaciones derivadas de esos análisis no son confiables. En un contexto financiero, donde las decisiones deben basarse en información precisa y completa, la presencia de datos erróneos puede llevar a decisiones equivocadas.

La revisión de estudios previos revela que cuando los datos son inexactos o incompletos, los análisis resultantes pueden ser incorrectos. Esto significa que las conclusiones y recomendaciones derivadas de esos análisis no son confiables. En un contexto financiero, donde las decisiones deben basarse en información precisa y completa, la presencia de datos erróneos puede llevar a decisiones equivocadas. Esto, a su vez, afecta negativamente tanto la planificación estratégica, que implica la definición de objetivos a largo plazo y la formulación de planes para alcanzarlos, como la planificación operativa, que se refiere a la gestión diaria y la ejecución de tareas específicas dentro de una organización. En resumen, la precisión y completitud de los datos son cruciales para asegurar que las decisiones estratégicas y operativas se fundamenten en análisis fiables y efectivos.

Se han identificado tendencias emergentes en la gestión de limpieza de datos, como el uso de tecnologías blockchain para garantizar la integridad de los datos económicos y la automatización de procesos de limpieza. Estas innovaciones están cambiando la forma en que se manejan los datos, mejorando la eficiencia y reduciendo los errores.

En la limpieza de datos para un análisis financiero, no existe un procedimiento único a implementar. Varias rutinas pueden adaptarse y utilizarse en bases con información similar, pero hay que considerar que cada conjunto de datos es particular. La metodología de limpieza de datos es altamente dependiente del dominio de conocimiento y de naturaleza significativamente exploratoria. Castillo Noguera María et al 2015

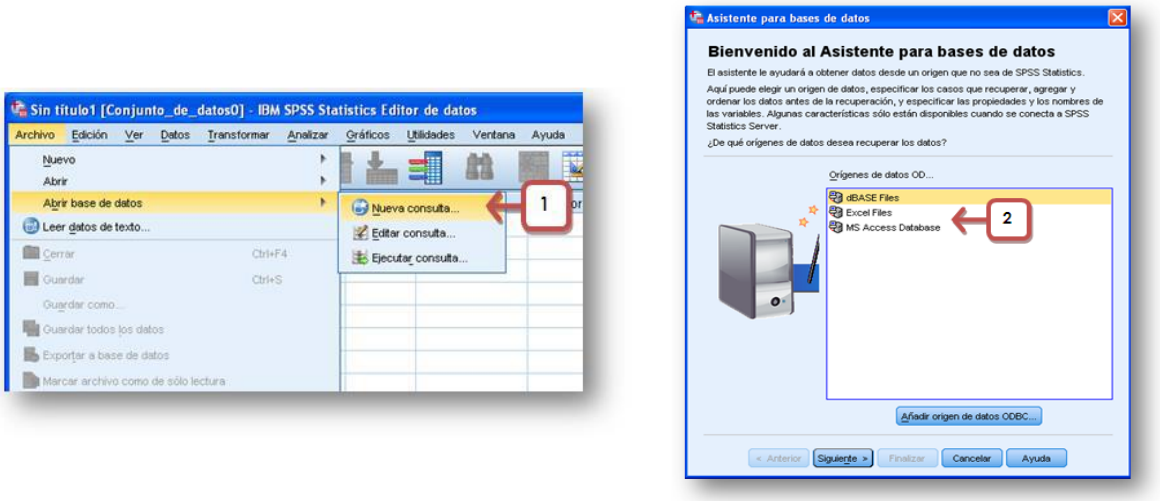
Desde una perspectiva práctica, estos resultados tienen diversas aplicaciones que pueden beneficiar a organizaciones financieras, empresas, gobiernos y analistas económicos. La implementación efectiva de estrategias de limpieza de datos puede llevar a una toma de decisiones más precisa y a una mayor confiabilidad de los informes económicos, y sumado la adopción de tecnología en entornos de aprendizaje ha venido acompañada de una disponibilidad de datos sin precedentes.

Los dos formatos típicos en los que puede recibirse la información son: formato de libro de Excel (*.xls) o formato de base de datos de Microsoft Access (*.mdb). Una vez identificado el conjunto de datos con el que se va a trabajar, se importan y se transforman en un archivo de SPSS (*.sav), como se observa en la [1]Figura 2. Castillo Noguera María et al 2015

Si el conjunto de datos está dividido en varios archivos, se sugiere agregar una variable “marcador” para poder identificar de dónde proviene cada caso cuando la información esté integrada.

[1] Las figuras y tablas que se presentan en adelante son elaboración propia del autor, por lo que no se indicará individualmente su fuente.

Figura 2. Importación de archivos en SSPS



Es importante contar con el libro de códigos de las bases de datos con las que se va a trabajar. Este documento debe describir los ítems que se recolectaron, el proceso de ingreso de datos, la codificación de la información, así como la organización del archivo. También es primordial disponer de los instrumentos utilizados en la evaluación tal y como fueron aplicados. De manera que pueda observarse la estructura de la prueba o cuestionario, los tipos de preguntas y formatos posibles de respuesta.

Una vez que se tienen los datos en un archivo de extensión *.sav, se procede a generar frecuencias para cada una de las variables con el objetivo de conocer qué información nos ha sido trasladada en la base de datos. El contenido de esta información puede editarse desde las variables o desde los casos. Los archivos de IBM® SPSS® Statistics están organizados por filas y columnas. En la vista de datos (ver Figura 3), las filas representan los casos u observaciones, mientras que las columnas representan las variables. En la vista de variables (ver Figura 4), cada fila es una variable y cada columna es un atributo asociado a dicha variable.

Figura 3. Vista de datos en SPSS

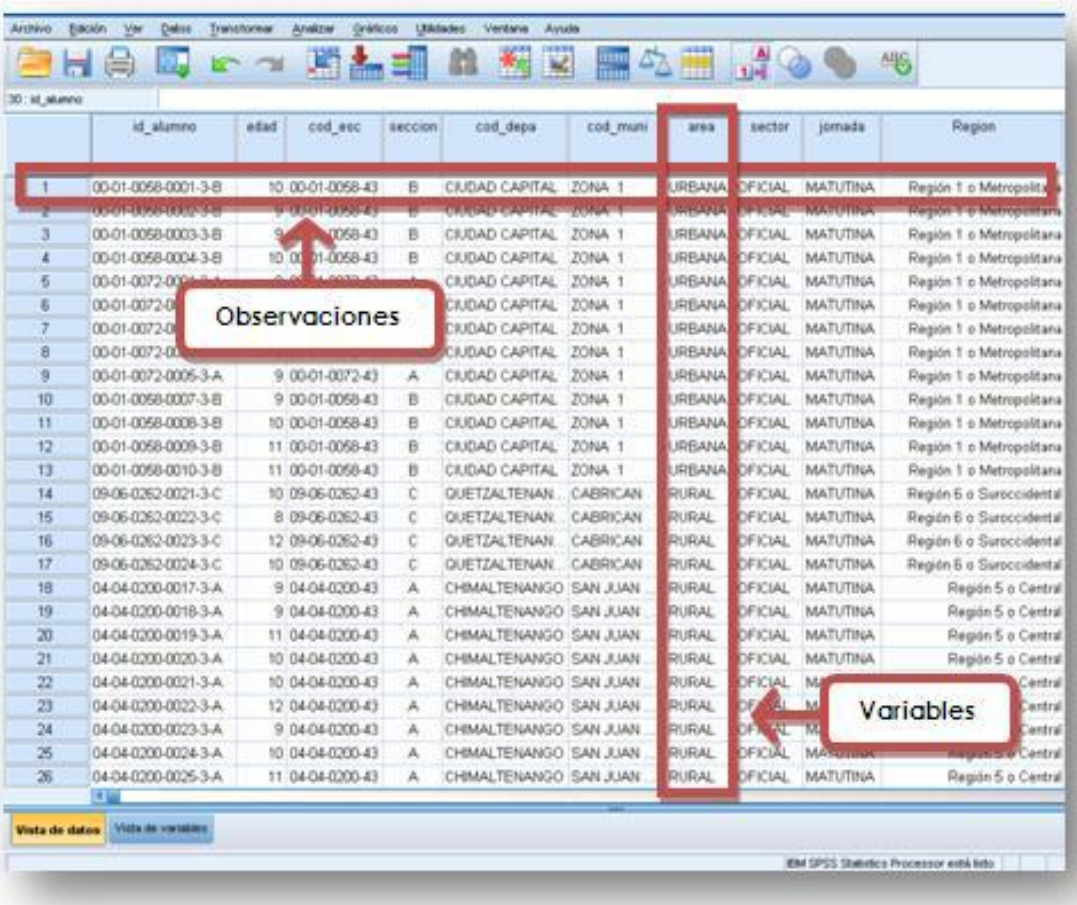
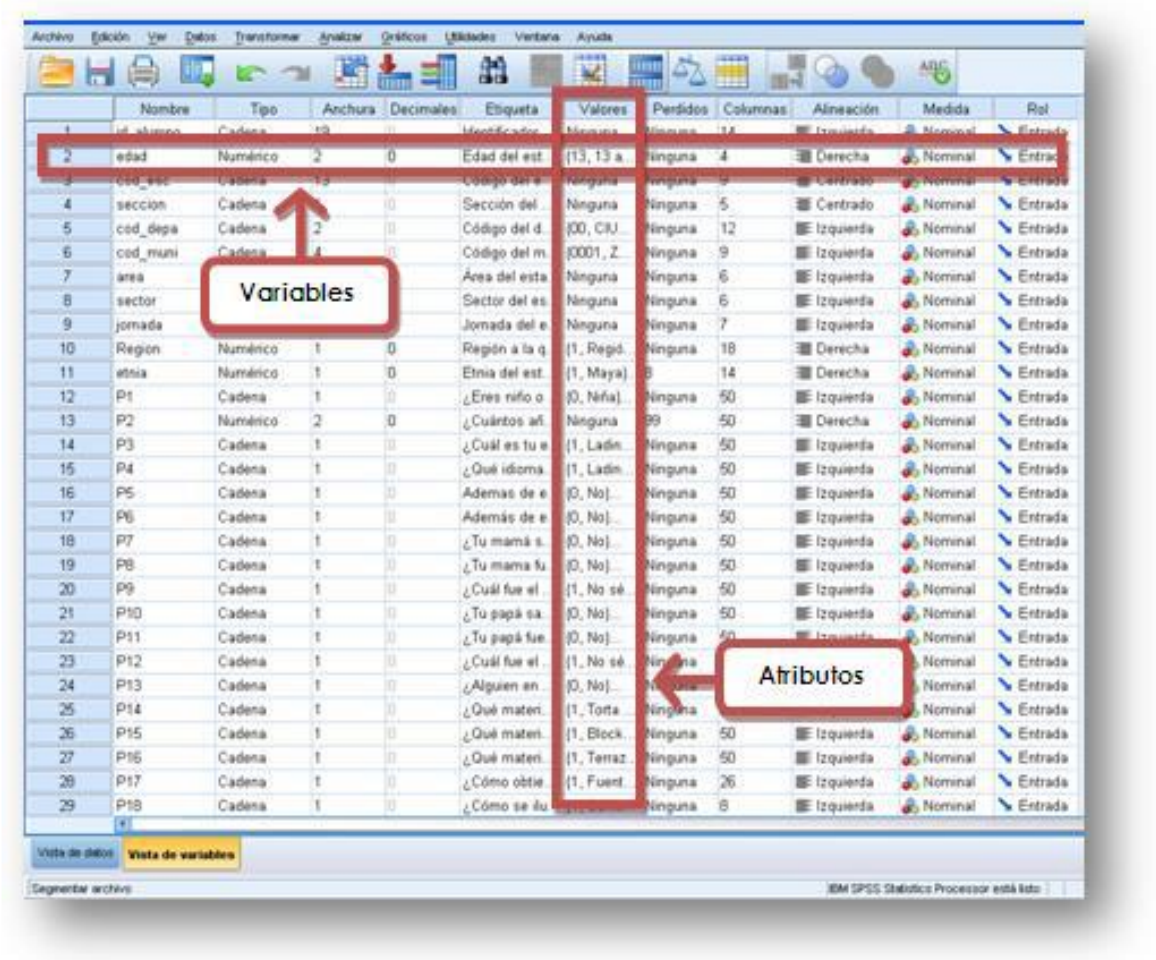


Figura 4. Vista de variables en SPSS



Distintas características pueden definirse para cada una de las variables, como se muestran en la Figura 5: nombre, tipo, anchura, decimales, etiqueta, valores, perdidos, columnas, alineación, medida y rol.

Figura 5. Atributos de las variables en SPSS

	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	VAR1	Numérico	8	2	Variable 1	Ninguna	Ninguna	5	Derecha	Ordinal	Entrada
2	VAR2	Cadena	4	0	Variable 2	Ninguna		8	Centrado	Nominal	Entrada
3	VAR3	Numérico	2	1	Variable 3	Ninguna	999.0, 888.0	8	Izquierda	Escala	Entrada

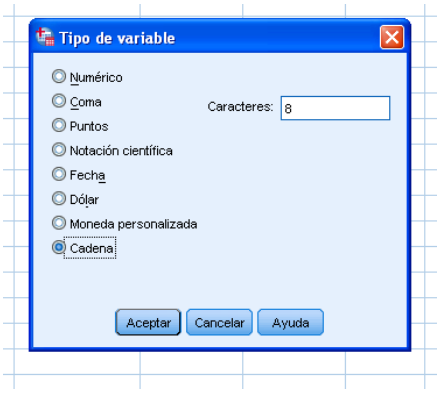
El nombre de la variable debe ser claro, conciso y representativo. Se espera también que, si se determinan algunas normas de codificación, sea congruente con ellas. Pueden utilizarse caracteres alfanuméricos y guion bajo. No debe incluir tildes, espacios o caracteres especiales. Estas consideraciones son importantes para facilitar la migración y comparación de archivos entre distintas versiones de IBM® SPSS® Statistics. Los nombres de cada variable deben ser únicos dentro de cada conjunto de datos, es decir, en una base no pueden existir dos nombres iguales (ver Tabla 1).

Tabla 1. Ejemplos de variables Nombre

Nombre	Explicación de lo que representa la variable
Nº_carnet	(Identificador único del alumno)
cod_depa	(Código de departamento)
nom_depa	(Nombre de departamento)
forma_CL	(Forma del Cuestionario de economía política)
forma_EE	(Forma del Cuestionario de Estrategias Docentes)
PM1	(Pregunta 1 de Matemáticas)
PL3_CHE	(Pregunta 3 de Lectura Historia Económica)
P7_rep_primero	(7. Marque las asignaturas que repitió: primer año)
P7_rep_tercero	(7. Marque las asignaturas que repitió: tercer año)
P7_rep_quinto	(7. Marque las asignaturas que repitió: quinto año)

El **tipo** de la variable puede ser numérico, coma, puntos, notación científica, fecha, moneda y cadena. Dada la información que se maneja en las aulas de clases, los tipos de variable más utilizados son numéricos y de cadena (ver figura N.6).

Figura 6. Variables numérica y variables de cadenas



Las variables numéricas son aquellas que tienen como valores números y por lo tanto permiten la realización de cálculos. Para ellas se definen las cifras decimales y el ancho, que incluye el punto decimal entre los caracteres. Las variables de cadena, también conocidas como variables alfanuméricas, tienen valores no numéricos, de manera que no pueden utilizarse en operaciones matemáticas o estadísticas. Los valores de las variables de cadena pueden contener cualquier carácter siempre que no se exceda la longitud definida. Las mayúsculas y las minúsculas se consideran diferentes.

La anchura se refiere al número de dígitos o caracteres de una variable. Este valor no debe ser menor al del dato con mayor longitud observado. Una vez que se reduce el ancho de la variable, los valores que quedan fuera del margen de largo establecido se pierden.

Algunas variables tienen un ancho definido previamente, por ejemplo, código del establecimiento que tiene un largo de 13 caracteres, código del departamento que tiene una anchura de dos o código de municipio con una anchura de cuatro. Sin embargo, en estos casos se recomienda en un inicio, observar la frecuencia de valores para detectar datos incorrectos y para no crear un error cortando información de un valor.

Para formatos numéricos se pueden ingresar valores con decimales. IBM® SPSS® Statistics versión 19 permite hasta 16 dígitos decimales. En la “Vista de datos” se observa solo el número definido de dígitos decimales. Los valores con más decimales de los determinados se redondean para cumplir con el formato; el valor completo se almacena internamente y se utiliza en todos los cálculos.

Variables, también admiten la definición de una etiqueta, que identifica y describe el contenido de los valores de cada variable. Las etiquetas de variable pueden tener una longitud de hasta 256 caracteres, contener espacios y caracteres reservados que no se admiten en los nombres de variable.

Para la mejor compresión de los conceptos vamos a exponer algunos ejemplos simples de Variables Numéricas

- Ventas Mensuales, Ejemplo: 150,000 (ventas en dólares), esto Permite evaluar el rendimiento de una empresa mes a mes, identificar tendencias y tomar decisiones sobre estrategias de marketing y ventas.
- Otro ejemplo seria la Tasa de Interés, Ejemplo: 3.5%, este interés es utilizada para calcular el costo de los préstamos y el rendimiento de las inversiones, esencial para la planificación financiera personal y empresarial.
- Número de Empleados, Ejemplo: 200. Ayuda en la gestión de recursos humanos, planificación de la capacidad y análisis de productividad**.**
- Duración de Llamadas al Servicio al Cliente, si tuviéramos 5 minutos por llamada, Se utiliza para medir la eficiencia del equipo de soporte, identificar áreas de mejora y optimizar la experiencia del cliente.

También facilitamos otros ejemplos de Variables de Cadena

- Nombre del Cliente, Orlando Cruz, es fundamental para personalizar el servicio al cliente, gestionar bases de datos de clientes y realizar marketing segmentado.
- Código de Producto, OCO-1969, esto facilitaría la identificación y seguimiento de productos en inventarios, mejora la gestión logística y el control de calidad.
- Dirección de Correo Electrónico, Orlando.cruz@unan.edu.ni, Utilizada para comunicarse con clientes, enviar boletines informativos y promociones, y para la recuperación de cuentas en plataformas en línea.
- Categoría de Producto, Electrónica, Ayuda en la organización del catálogo de productos, facilita la búsqueda y filtrado en sistemas de comercio electrónico y permite análisis de ventas por categoría.

Un tercer grupo de ejemplo seria las Aplicaciones Prácticas Combinadas

- Análisis de Ventas por Cliente, Variables Involucradas: Ventas Mensuales (numérica) y Nombre del Cliente (cadena). Permite identificar clientes clave, personalizar ofertas y mejorar la retención de clientes.

- Gestión de Inventarios, variables Involucradas: Número de Empleados (numérica) y Código de Producto (cadena). Optimiza la distribución de tareas entre empleados y facilita el seguimiento de inventarios, asegurando la disponibilidad de productos.

Es importante definir el concepto de Anchura de Variables

La anchura de una variable se refiere al rango o al tamaño del espacio de almacenamiento asignado para esa variable en una base de datos o en un sistema de procesamiento de datos. Dependiendo del tipo de variable (numérica, de cadena, etc.), la anchura puede tener diferentes implicaciones:

- Variables Numéricas, la anchura: Define el número de dígitos que puede contener una variable numérica. Por ejemplo, un campo definido con una anchura de 10 dígitos puede almacenar números entre -123456789 y 987654321. Una anchura adecuada asegura que los números se almacenen con precisión sin desbordar el espacio asignado.
- Variables de Cadena, la anchura: Determina el número máximo de caracteres que una cadena puede almacenar. Por ejemplo, una variable de cadena con una anchura de 50 caracteres puede contener hasta 50 caracteres alfanuméricos. La anchura garantiza que las cadenas de texto se almacenen correctamente sin truncar la información**.**

Importancia de Definir Correctamente el Ancho

- Precisión y Exactitud, Variables Numéricas: Definir un ancho insuficiente para una variable numérica puede resultar el ensanchamiento de datos o en errores de desbordamiento, lo que afecta la precisión de los cálculos y análisis.
- Variables de Cadena: Un ancho insuficiente para una cadena puede truncar la información, lo que lleva a la pérdida de datos importantes y a la posible confusión en la interpretación de la información.

Integridad de los Datos

- Variables Numéricas: Si los datos se almacenan en un campo con una anchura menor al requerido, los números pueden ser redondeados o truncados, afectando la integridad y la confiabilidad de los datos.

- Variables de Cadena: La truncación de cadenas puede resultar en datos incompletos, como direcciones de correo electrónico parciales o nombres truncos, afectando la calidad de la información y su utilidad en análisis posteriores.

Ejemplos de Implicaciones de un Ancho Insuficiente

- Caso de Variables Numéricas, Si un campo diseñado para almacenar una cifra de ventas solo tiene 5 dígitos y se intenta ingresar una venta de $120,000, el sistema puede truncar la cifra a $12,000 o generar un error, lo que lleva a datos incorrectos y decisiones basadas en información errónea.
- Caso de Variables de Cadena, Si un campo para nombres de clientes tiene un ancho de 20 caracteres y se ingresa un nombre largo como "Alexander Perezcassar", el nombre podría ser truncado a "Alexander Pérez", lo que resulta en una información incompleta y potencialmente confusa en la base de datos.

Tabla 2. Ejemplos de etiquetas de variable

Nombre de la variable	Etiqueta de la variable
nom_Jornada	Nombre de la jornada de trabajo
cod_jornada	Código de la jornada del establecimiento
PM27	Pregunta 27 de Guía de campo
necesidad_educativa	¿El alumno tiene alguna necesidad educativa especial?
P12_experiencia	Tiempo de experiencia docente (años)
etnia_recode	Etnia del estudiante (re-codificada en Niacarao/NoDiringen)
dominio_idioma_materno	¿Qué destreza domina en su idioma materno?

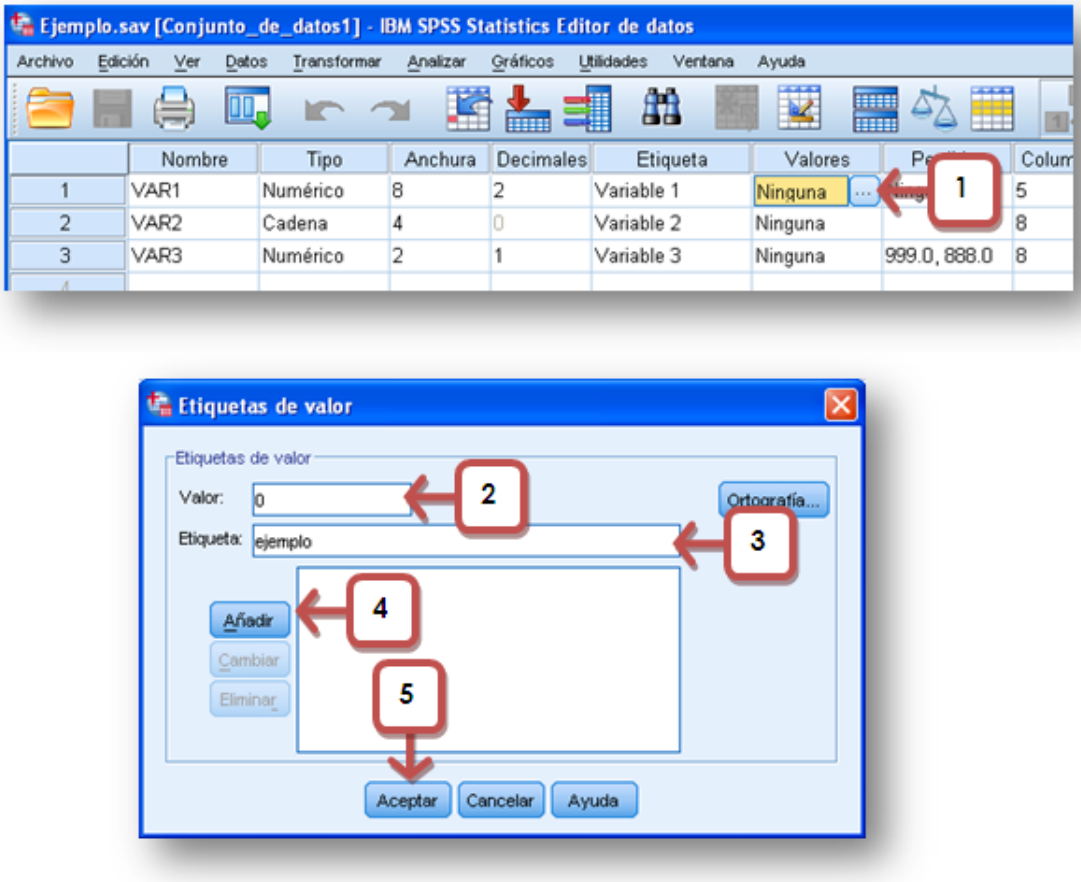
En la característica de **valores**, se pueden asignar etiquetas descriptivas a cada valor de una variable. Esto resulta útil, especialmente cuando una variable utiliza códigos para representar distintas categorías. Una vez definidas las etiquetas de valor para una variable, ya no es necesario volver a indicarlas cada vez que se abre el archivo de datos. Ejemplos de las etiquetas que se asignan a valores de 4 las variables se pueden observar en la Tabla 3.

Tabla 3. Ejemplos de etiquetas de valor

Variable	Valores y etiquetas
sexo	0, Femenino; 1, Masculino
área	11, Urbana; 12, Rural
escalafón salarial	1, Clase A; 2, Clase B; 3, Clase C; 4, Clase D; 5, Clase E; 6, Clase F
idioma materno	1, español; 2, idioma misquito; 3, criol; 4, sumo 5, otro
asistencia a preprimaria	0, No; 2, Sí

En el programa SPSS se pueden definir las etiquetas de valor en la vista de variables de la base de datos, según se observa en la Figura 7, lo cual se puede hacer con la propiedad de la variable o con sintaxis de programación de comandos.

Figura 7. Definición de etiquetas de valor en SPSS

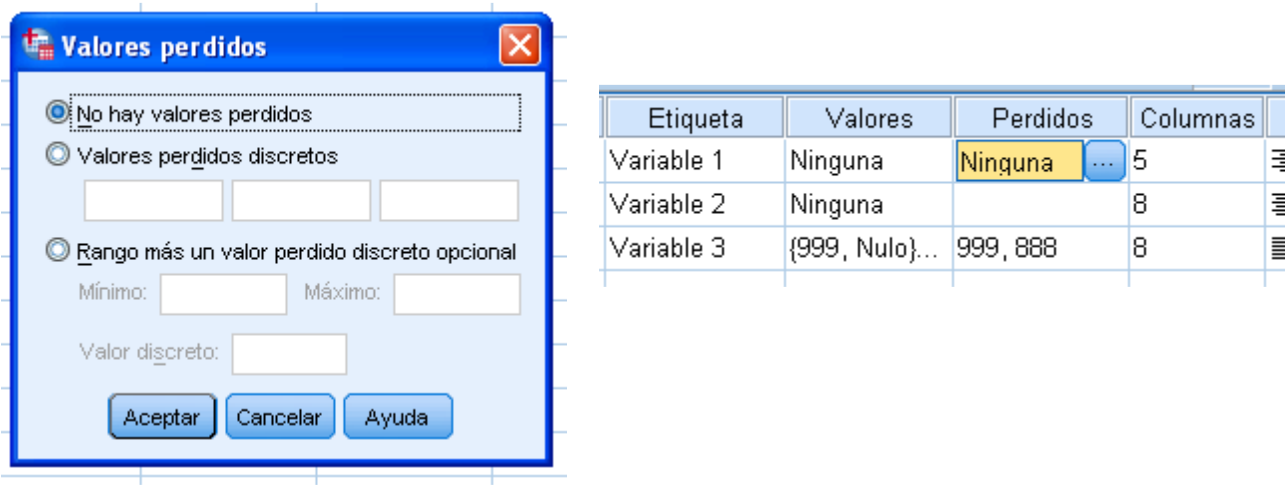


Como se ha mencionado antes, los datos perdidos o no válidos son comunes en las bases de datos. Es probable que las personas que hayan respondido a un cuestionario o participado en alguna evaluación, se nieguen a responder ciertos ítems, no conozcan la respuesta a determinadas preguntas o contesten de forma inesperada. Por ello es necesario identificar y filtrar estos datos perdidos en cada variable, para que los análisis proporcionen resultados exactos. IBM® SPSS® Statistics no utiliza los valores que se definen como perdidos en los cálculos realizados con la variable.

Se pueden señalar hasta tres valores perdidos de tipo discreto, un rango de valores perdidos o un rango más un valor de tipo discreto. Los rangos únicamente se pueden especificar para variables numéricas. Los valores perdidos de las variables de cadena distinguen mayúsculas y minúsculas.

El software considera como válidos todos los valores de cadena incluidos los valores vacíos o nulos, a menos que se definan explícitamente como perdidos. Para definir como perdidos los valores nulos o vacíos de una variable de cadena, se ingresa un espacio en blanco en uno de los campos de la sección valores perdidos discretos (ver Figura 8).

Figura 8. Ventana para definir valores perdidos para las variables en SPSS



Los valores perdidos también pueden tener etiquetas de valor, por ejemplo “Z” puede representar “Respuesta en blanco”; “Y”, “Respuestas múltiples; “E”, “Respuesta sin sentido”; 999, “Valor Nulo”; 888, “Doble Respuesta”.

En la propiedad columnas se especifica el número de caracteres para el ancho de la columna en la "Vista de datos". Esta propiedad puede cambiarse también directamente desde la "Vista de datos", pulsando y arrastrando los bordes de las columnas. El ancho de columna afecta solo la presentación, pero no los valores en el "Editor de datos" de IBM® SPSS® Statistics. Al cambiar el ancho de columna no se cambia el ancho definido de una variable.

Debe procurarse que la vista de los valores de las variables sea cómoda y fácil de manejar para el posterior estudio y análisis de la base de datos. La definición de esta propiedad dependerá tanto del largo de los valores de la variable como del aspecto estético y funcional deseado. La percepción de orden y claridad en la "Vista de variables" se favorece al homogenizar anchos de columna para variables similares.

La definición de formato de variables también considera la alineación. Los valores y etiquetas de datos pueden presentarse alineados a la derecha, a la izquierda o al centro. Por defecto, las variables numéricas tienen alineación hacia la derecha y las variables de cadena hacia a la izquierda. Al ajustar esta propiedad se afecta únicamente la "Vista de datos".

IBM® SPSS® Statistics permite especificar el nivel de medida de cada variable como Nominal, Ordinal o Escala; esta especificación se puede ver en la (Figura N. 8.)

Discusión

La discusión de los resultados revela que la gestión de limpieza de datos en los análisis financieros es un campo interdisciplinario que involucra a profesionales de la informática, la estadística, la economía y la gestión empresarial.

Existen diversos problemas que pueden ser resueltos mediante la limpieza y transformación de datos. Las inconsistencias descubiertas pueden deberse a discordancias con definiciones de datos, fallas en la entrada de información y corrupción en la transmisión o en el almacenaje, y afectar distintos tipos de variables, a partir de esta aseveración los principales puntos de discusión incluyen, que la

diversidad de fuentes de datos plantea un desafío significativo, donde los datos provienen de múltiples sistemas y formatos, lo que complica su integración y limpieza.

Según el origen de los datos, los problemas de calidad pueden ser problemas de una sola fuente o de múltiples fuentes y presentarse a nivel de esquema o a nivel de instancia. Los problemas a nivel de esquema se refieren a pobre diseño de la base y a falta de restricciones de integridad que controlen los datos válidos y permitidos. Por otro lado, los problemas a nivel de instancia se refieren a errores e inconsistencias en el contenido actual, los cuales no son visibles desde el esquema de la base de datos (Rahm et al, 2000).

Por ejemplo, cuando hablamos de Inconsistencias, tenemos que las inconsistencias ocurren cuando hay discrepancias o diferencias en los datos que deberían ser coherentes entre sí. Esto puede manifestarse como datos que no coinciden en diferentes registros o bases de datos que deberían reflejar la misma información. Y esto se debe a Errores Humanos, cuando introducen de forma incorrecta los datos por parte de usuarios, como escribir diferentes formatos de fechas o unidades.

En la Integración de Datos, se combina los datos provenientes de múltiples fuentes sin una estandarización adecuada.

Y en las actualizaciones Concurrentes, se da cuando se modifica simultáneamente los datos en el sistema de diferente manera que no se sincronizan correctamente, lo cual permite, que se den decisiones Erróneas, datos inconsistentes que pueden llevar a conclusiones incorrectas y decisiones empresariales equivocadas.

Confusión en Reportes: Los informes generados pueden ser engañosos, afectando la credibilidad de la información, por las fallas de Entrada, la cuales se producen cuando los datos ingresados en el sistema son incorrectos o incompletos, ya sea por errores en el proceso de entrada o por deficiencias en los mecanismos de validación, y puede darse por errores tipográficos: errores humanos al ingresar datos, como escribir mal una dirección de correo electrónico o número de teléfono.

Formularios Mal Diseñados: Campos de entrada que no tienen validaciones adecuadas para asegurar la precisión de los datos ingresados.

Falta de Validación: Sistemas que permiten la entrada de datos sin comprobar su formato o rango, esto tiene sus consecuencias en los datos Inexactos: Información incorrecta en el sistema que puede afectar la operación y análisis.

Reprocesamiento: Necesidad de corregir los errores manualmente, lo que consume tiempo y recursos.

Los Datos Duplicados, ocurren cuando la misma información se introduce múltiples veces en la base de datos, creando registros redundantes que pueden causar confusión y se da or falta de Restricciones de Unicidad: La base de datos no impone restricciones para evitar la entrada de registros idénticos.

Integración de Sistemas: Al combinar datos de múltiples fuentes sin un control adecuado, se pueden crear duplicados.

Errores en Procesos de Importación: La importación de datos desde otras fuentes puede introducir registros repetidos si no se manejan correctamente, lo cual duplica el Esfuerzos: Los usuarios pueden trabajar con información repetida, llevando a esfuerzos innecesarios y confusión.

Impacto en Análisis: Análisis y reportes pueden ser imprecisos debido a la presencia de datos redundantes, Datos Incompletos

Conclusiones

A partir de la revisión realizada sobre la limpieza de datos, como un aspecto fundamental que requiere un enfoque integral y de mejora continua para sustentar análisis precisos y mitigar riesgos en la toma de decisiones, a partir de aquí se pueden derivar las siguientes afirmaciones como una sola conclusión:

- La limpieza de datos es un proceso fundamental para garantizar la calidad y confiabilidad de la información utilizada en un análisis financiero. Los datos sucios o erróneos pueden llevar a resultados imprecisos y decisiones equivocadas.
- Existen diferentes enfoques, técnicas y herramientas que pueden aplicarse para limpiar y normalizar los datos económicos y financieros. Desde métodos estadísticos hasta aprendizaje automático, cada vez se desarrollan soluciones más sofisticadas.
- La gestión de la calidad de los datos es un aspecto crítico, ya que los análisis financieros sustentan decisiones importantes en áreas como la planificación, evaluación de riesgos y proyección de tendencias. Datos deficientes generan incertidumbre.
- Tanto empresas como gobiernos generan y recopilan grandes volúmenes de información que requieren limpieza constante para asegurar su validez a lo largo del tiempo. Es necesario un enfoque integral y procesos de mejora continua.
- El uso combinado de técnicas estadísticas, herramientas tecnológicas y métodos de aprendizaje automático permite abordar los desafíos de la limpieza de datos de manera más eficiente en el contexto financiero.
- La gestión de los datos debe estar alineada con la cultura organizacional para trascender como una función transversal y contribuir de manera sostenible a la toma óptima de decisiones.

Referencias

Ahmed, I., & Aziz, A. (2010). Dynamic Approach for Data Scrubbing Process. (Enfoque dinámico para el proceso de depuración de datos) International Journal on Computer Science and Engineering, 2 (2), 416-423.

Castillo Noguera María José, Santos Solares José Adolfo. Limpieza de Datos (2015).

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. (De la minería de datos al descubrimiento de

conocimientos en bases de datos) American Association for Artificial Intelligence, 37-54.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. (Minería de datos: conceptos y técnicas) Estados Unidos: Elsevier.

Hernández Orallo José, María José Ramírez Quintana, & María Adela Gutiérrez Díaz. Pearson Educación, 2014. "Minería de datos: Conceptos y técnicas"

Information Builders. (2011). Data Quality: It's no joke. (Calidad de datos: no es una broma) Obtenido de Insights - Trends & Technologies: - http://informationbuilders.co.uk/new/insights/UKNewsletterQ1.html (constructores de información).

Kitlas, J. (2012). Dirty Data. (Datos sucios) Obtenido de Subject Guides - Syracuse University Library: http://researchguides.library.syr.edu/content.php?pid=156265&sid=2578897

Laia Subirats Maté Diego Oswaldo Pérez Trenard Mireia Calvo González (2019). Introducción a la limpieza y análisis de los datos.

Maletic, J. I., & Marcus, A. (2000). Data Cleansing: (Limpieza de datos) Beyond Integrity Analysis. Obtenido de Proceedings of the Conference on Information Quality (IQ2000), Boston, October 2000: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.37.5212

Maletic, J., & Marcus, A. (2010). Data Cleansing: ((Limpieza de datos) A prelude to knowledge discovery. En O. Maimon, & L. Rokach, Data Mining and Knowledge Discovery -Handbook (págs. 21-36). New York: Springer.

Mar Pérez Sanagustín & José A. Ruipérez Valiente. Editorial Universitat Oberta de Catalunya, 2019. "Data Science: Limpieza, análisis y visualización de datos".

Maydanchik, A. (2007). Data Quality Assessment. (Evaluación de la calidad de los datos) Estados Unidos: Technics Publications, LLC. Müller, H., & Freytag, J.-C.

(2003). Problems, Methods, and Challenges in Comprehensive. Data Cleansing. Berlin: Humboldt University Berlin.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data Quality Assessment. Communications of the ACM, 211-218. (Evaluación de la calidad de los datos).

Rahm, E., & Do, H. H. (2000). Data Cleaning: (Limpieza de datos) Problems and Current Approaches. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering , 1-11. Recuperado desde http://www.acm.org/sigs/sigmod/disc/disc01/out/websites/deb_december/rahm.pdf.

Sánchez González Jorge Jesús. Editorial Marcombo, 2018. "Análisis de Datos con Excel: Una guía práctica para la limpieza y análisis de datos".

Van den Broeck, J., Argeseanu Cunningham, S., Eeckels, R., & Herbst, K. (2005). Data cleaning: Detecting,diagnosing, and editing data abnormalities. PLoS Medicine, 2 (10), e267. (Limpieza de datos: detección, diagnóstico y edición de anomalías en los datos).