AA VALIDA LA DATA DE ACUERDO CON EL PROCESO ANÁLISIS DE EXPLORACIÓN DE DATOS (TIII)

IDENTIFICACIÓN DE LA GUÍA DE APRENDIZAJE

- Denominación del Programa de Formación: TÉCNICO PARA LA PROGRAMACIÓN PARA ANALÍTICA DE DATOS.
- Código del Programa de Formación: 228117
- Nombre del Proyecto: 1901119 APLICAR BUENAS PRACTICAS PARA PREPARAR, LIMPIAR, REFINAR Y EXPLORAR GRANDES VOLÚMENES DE DATOS EN EL SECTOR PRODUCTIVO.
- Fase del Proyecto:
  - EXPLORACIÓN DE DATOS
  - ADQUISICIÓN DE DATOS
- Actividad de Proyecto:
  - EXPLORAR DATA
  - GENERAR FICHA TÉCNICA DE LOS DATOS
- Competencia:
- INTEGRACIÓN DE DATOS SEGÚN TÉCNICAS DE VISUALIZACIÓN Y METODOLOGÍAS DE ANÁLISIS.
- PROCESO DE DATOS DE ACUERDO CON PROCEDIMIENTO TÉCNICO Y METODOLOGÍA ESTADÍSTICA. 150H 210601026
- Resultados de Aprendizaje Alcanzar:
  - VALIDAR LA DATA DE ACUERDO CON EL PROCESO ANÁLISIS DE EXPLORACIÓN DE DATOS.
  - ELABORAR INFORMES SEGÚN LA NECESIDAD DEL CLIENTE.
- Duración de la Guía: 198H (presenciales + trabajo autónomo)

2. PRESENTACIÓN

La validación de la data es uno de los procesos más importantes dentro del análisis de exploración de datos, ya que esto permite tener la certeza de que los datos que se están utilizando son precisos y confiables. Al hacer esto, se reduce el riesgo de análisis erróneos y se puede obtener información precisa para la toma de decisiones informadas. Para validar la data, se debe seguir un proceso riguroso que incluye la revisión de datos erróneos, datos faltantes, valores extremos y la verificación de que los datos se han ingresado de manera correcta. Al utilizar el proceso de análisis de exploración de datos, se pueden identificar las mejores prácticas para validar la data y asegurar la fiabilidad de los resultados.

3. FORMULACIÓN DE LAS ACTIVIDADES DE APRENDIZAJE

3.1 Actividades de Reflexión inicial.

En el proceso de análisis de exploración de datos, es fundamental validar la data para asegurarnos de que los resultados obtenidos sean confiables y precisos. Para ello, podemos seguir los siguientes pasos:

1. Revisión de la calidad de los datos: antes de comenzar el análisis, es necesario revisar la calidad de los datos para identificar posibles errores, omisiones o inconsistencias. Es importante asegurarnos de que los datos estén completos, sin duplicados y sin valores atípicos.
2. Verificación de la consistencia de los datos: una vez que hemos revisado la calidad de los datos, es necesario verificar la consistencia de los mismos. Para ello, podemos comparar la información en diferentes campos y verificar que no existan contradicciones.
3. Validación de la corrección de la información: en caso de encontrar algún error o inconsistencia, es fundamental corregirlo antes de continuar con el análisis. Para ello, podemos revisar la fuente de los datos o contactar a quienes proporcionaron la información para asegurarnos de que se trata de datos precisos y confiables.
4. Verificación de la precisión de los datos: además de la calidad, consistencia y corrección de los datos, es necesario asegurarnos de que sean precisos. Para ello, podemos realizar pruebas estadísticas o comparaciones con datos de referencia para corroborar que los valores sean los correctos.
5. En resumen, validar la data en el proceso de análisis de exploración de datos es fundamental para garantizar la precisión y confiabilidad de los resultados obtenidos. Para ello, es necesario revisar la calidad, consistencia, corrección y precisión de los datos antes de continuar con el análisis.

Desarrolle un ensayo sobre los cincos planteamientos anteriores, resaltando su importancia para proyectos de analítica de datos.

“Todo debe hacerse lo más simple posible. Pero no más sencillo.”
Albert Einstein

3.2 Actividades de contextualización e identificación de conocimientos necesarios para el aprendizaje.

El proceso de análisis de exploración de datos es una etapa crucial en el análisis de datos, donde se busca comprender, validar y obtener información relevante a partir de los datos disponibles. Durante este proceso, se aplican diversas técnicas y herramientas para explorar la estructura y características de los datos, identificar patrones, tendencias y anomalías, y generar insights que respalden la toma de decisiones informada.

Para validar la data de acuerdo con el proceso de análisis de exploración de datos, se deben seguir varios pasos:

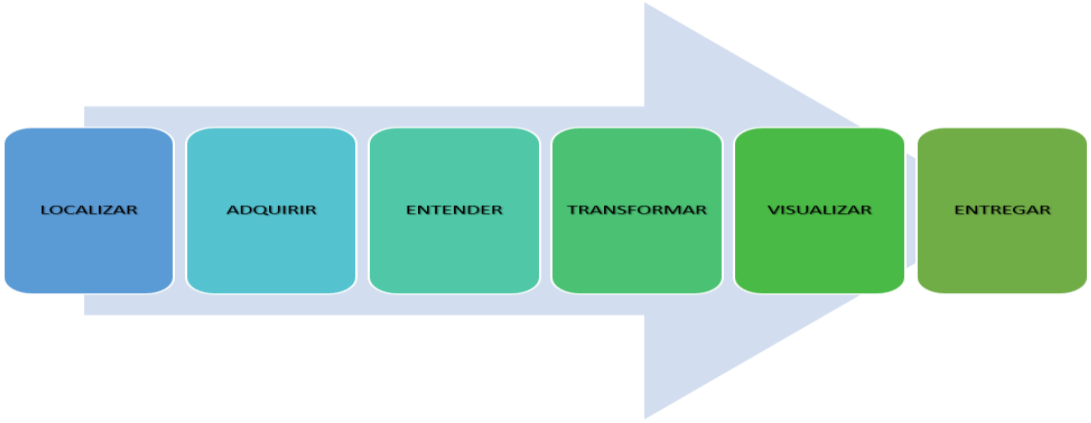
1. Recopilación de datos: Se debe asegurar que los datos utilizados sean completos, relevantes y representativos de la población o fenómeno que se está estudiando. Esto implica tener claridad sobre la fuente de los datos, su calidad y confiabilidad.
2. Limpieza de datos: Antes de comenzar el análisis, es fundamental realizar una limpieza de los datos para eliminar valores atípicos, datos faltantes o inconsistentes. Esto garantiza que los resultados obtenidos sean precisos y confiables.

3. Exploración inicial: En esta etapa, se realiza un análisis descriptivo básico para comprender la distribución de los datos, identificar estadísticas resumidas y detectar posibles errores o inconsistencias. Se pueden utilizar gráficos, tablas y resúmenes estadísticos para visualizar y resumir los datos.
4. Identificación de patrones y relaciones: Mediante técnicas de visualización y análisis estadístico, se busca identificar patrones, tendencias y relaciones entre variables. Esto puede incluir la exploración de correlaciones, análisis de series de tiempo, análisis de agrupamiento (clustering) o cualquier otra técnica pertinente para el conjunto de datos específico.
5. Validación de la consistencia y precisión: Durante el proceso de exploración, se deben realizar diferentes pruebas y validaciones para asegurar que los datos sean consistentes y precisos. Esto puede implicar verificar la coherencia de los datos con respecto a conocimientos previos o fuentes externas confiables, así como identificar posibles errores o valores atípicos que puedan afectar los resultados.
6. Documentación de hallazgos: Es importante documentar todos los hallazgos y resultados obtenidos durante el proceso de exploración de datos. Esto incluye la descripción de las técnicas utilizadas, los patrones identificados, las relaciones encontradas y cualquier conclusión relevante derivada de los análisis realizados.

En resumen, validar la data de acuerdo con el proceso de análisis de exploración de datos implica garantizar la calidad de los datos utilizados, realizar una limpieza adecuada, explorar y analizar los datos en busca de patrones y relaciones, validar la consistencia y precisión de los datos, y documentar los hallazgos obtenidos.

Este proceso es esencial para obtener información confiable y relevante a partir de los datos disponibles y respaldar la toma de decisiones basada en evidencia.

Se sugiere ver el video "Limpieza de datos" y "Análisis de Datos con R", para contextualizar la actividad de aprendizaje



3.3 Actividades de apropiación

Antes de continuar con el proceso de apropiación se invita al aprendiz a leer el documento **"Análisis exploratorio con R"**

Siguiendo la quinta fase "Visualiza", en compañía del instructor se desarrolla el taller "**GC-F-005_EXPLORARCONR.pdf**", donde se valida la data transformada desde una herramienta informática.

Desarrolla el taller "**GC-F-005_VISUALIZACION.pdf**", donde se utilizan herramientas informáticas para la visualización de la información procesada.

El taller "**GC-F-005_POWERBI.pdf**", ayudará a la visualización de la data procesada.

3.4 Actividades de transferencia del conocimiento

Aplicar las técnicas y métodos de limpieza de datos utilizando un DATASET propuesto por el instructor (SB11-20121-RGSTRO-CLFCCN-V1-0-txt).
Se desea comprobar lo siguiente:

- Quienes se destacaron más en matemáticas en la muestra de población con limitación, si las mujeres o los hombres, teniendo en cuenta:
  - La ciudad
  - Edad de acuerdo al tipo de documento de identidad
  - Tipo de colegio (Oficial, Privado) y caracterización del colegio (ACADEMICO, TECNICO, etc.)
  - Qué nivel de ingles
  - Nacionalidad

Entregue los datos solicitados mediante tablas en un modelo estrella en POSTGRESQL; no olvide adjuntar el código en R y el script SQL, como también el informe de desarrollo que debe contener:
● **Encabezado**: título del informe, nombre del instructor, autor del informe (nombres y apellidos completos), nombre del programa formativo, así como la fecha de realización.
● **Introducción**: describa el tema abordado en dimensiones y medidas.
● **Desarrollo**: corresponde al cuerpo del trabajo, donde se explica con detalle el desarrollo de los aspectos que se mencionan en la introducción. En este apartado deberá incluir:
  ● Informe argumentado, el desarrollo del caso de estudio propuesto.
  ● Pantallazos que demuestren las acciones.
  ● Acta de cambios, eliminaciones o adiciones al DATASET.
  ● Código en R utilizado.
● **Conclusiones**: presente las conclusiones a las que llegó luego de haber realizado el taller y el caso propuesto.

Lineamientos generales para la entrega de la evidencia:
● Productos a entregar: un documento que incluya lo solicitado para el desarrollo del caso de estudio propuesto en el taller.
● Formato: ZIP.
● Para hacer el envío de la evidencia remítase al área de la actividad correspondiente y acceda al espacio de evidencias del LMS.
Nota: Esta evidencia se debe realizar en grupo, pero cada integrante sube individualmente describiendo sus compañeros integrantes del grupo.

Esta evidencia se debe elaborar en grupo de tres aprendices en las fechas acordadas en el plan de trabajo concertado (PTC) y las instrucciones dadas por el instructor.

No olvide guardar la evidencia en el portafolio del aprendiz.

"Nunca consideres el estudio como una obligación sino como una oportunidad para penetrar en el bello y maravilloso mundo del saber"
Albert Einstein

4. ACTIVIDADES DE EVALUACIÓN

Evidencias de Aprendizaje	Criterios de Evaluación	Técnicas e Instrumentos de Evaluación
Evidencias de Conocimiento: **Evidencias de Desempeño.** **EV1 Realiza la validación dentro del análisis de exploración de datos del caso de estudio.** **Evidencias de Producto:** **EV2 Valida el modelo físico de la fuente de datos escogida para la limpieza de aplicando análisis exploratorio de datos.**	CREA NUEVOS DATOS A PARTIR DE OTROS DATOS, MANEJO DE DIFERENTES FUNCIONES PARA LA TRANSFORMACIÓN.	**IEV1. Lista de verificación** **IEV2. Lista de verificación**

5. GLOSARIO DE TÉRMINOS

Acceso
La manera en la cual los archivos o conjunto de datos son referenciados por la computadora**.**
Archivo
Un archivo es un elemento que contiene información y que a su vez se identifica por un nombre y su extensión. Esta última comienza por un punto y determina el tipo de aplicación a la que está asociado el archivo.
Buscadores
O también llamados motores de búsqueda, son herramientas que permiten clasificar la información que existe en la red y hacerla localizable en poco tiempo según las preferencias del usuario.
Base de datos
Una colección de registros o archivos relacionados de manera lógica.

Base de datos relacional
Una colección de relaciones normalizadas en la que cada relación tiene un nombre distintivo.
Bases de datos distribuidas
Son Bases de Datos que no están almacenadas totalmente en un solo lugar físico, (están segmentadas) y se comunican por medio de enlaces de comunicaciones a través de una red de computadoras distribuidas geográficamente.
Campo
Un campo es la unidad básica de una base de datos. Un campo puede ser, por ejemplo, el nombre de una persona. Los nombres de los campos no pueden empezar con espacios en blanco y caracteres especiales. No pueden llevar puntos, ni signos de exclamación o corchetes.
Clave principal
La clave principal en una tabla de una base de datos que se selecciona para identificar de forma unívoca cada registro de la tabla. Por ejemplo, en una tabla de alumnos podría ser su número de expediente académico.
Consulta
Mediante las consultas tendrás la posibilidad de obtener toda la información contenida en las tablas añadiendo interesantes funcionalidades.
DDL
Lenguaje de definición de datos utilizado para describir todas las estructuras de información y los programas que se usan para construir, actualizar e introducir la información que contiene una base de datos.
Diseño de la base de datos
Cuando trabajamos con bases de datos relacionales es habitual distribuir la información en diferentes tablas vinculadas entre sí. Esta característica obliga a un proceso de planificación y diseño previo para obtener el resultado esperado. Piensa que deseas almacenar en la base de datos, qué datos necesitas recuperar y en definitiva, determina el propósito final del proyecto para establecer unos cimientos lo suficientemente sólidos.
DBMS
Conjunto de programas destinados a manejar la creación y todos los accesos a las bases de datos. Se compone de un lenguaje de definición de datos (DDL: Data Definition Language), de un lenguaje de manipulación de datos (DML: Data Manipulation Language) y de un lenguaje de consulta (SQL: Structured Query Language).
ELIMINACION
Es una solicitud de eliminación que se expresa de forma muy parecida a una consulta. Sin embargo, en vez de presentar tuplas al usuario, quitamos las tuplas seleccionadas de la base de datos. Sólo puede eliminar tuplas completas; no se puede eliminar únicamente valores de determinados atributos.
Facilidad de Consultas
Permitir al usuario hacer cuestiones sencillas a la base de datos. Este tipo de consultas tienen como misión proporcionar la información solicitada por el usuario de una forma correcta y rápida.
Formulario
Los formularios resultan útiles principalmente en tareas de introducción de información. Cuando se trata de incluir pocos datos podemos hacerlo directamente sobre las tablas, pero cuando el volumen es importante, este método se vuelve poco eficaz. Para resolver este problema tenemos los formularios donde la inclusión de datos se hace de forma mucho más intuitiva y sencilla.
HTML
Siglas de HyperText Markup Language (Lenguaje de Etiquetas de Hipertexto), es el lenguaje predominante para la construcción de páginas web. Se utiliza para describir la estructura y el contenido en forma de texto, así como para complementar el texto con otros objetos, como por ejemplo: imágenes. Los archivos creados en este lenguaje suelen identificarse por su extensión del tipo: "nombre_archivo.html".
Informe

Los informes tienen como objetivo proporcionar las herramientas necesarias para obtener una copia impresa de los datos existentes en una base de datos, aunque existen otras posibilidades tan interesantes como la generación de archivos en formato PDF. Habitualmente, los informes se suelen construir a partir de los resultados obtenidos de la ejecución de consultas. De esta forma combinamos la posibilidad de seleccionar sólo los datos que deseemos que nos ofrecen las consultas con la ventaja de imprimirlos que aportan los informes.

Independencia de los datos

Se refiere a la protección contra los programas de aplicaciones que pueden originar modificaciones cuando se altera la organización física y lógica de las bases de datos.

Integridad referencial

La integridad referencial es una propiedad imprescindible en cualquier base de datos. Gracias a la integridad referencial se garantiza que un conjunto de datos (registro) siempre se relacione con otros conjuntos válidos, es decir, que existen en la base de datos. Implica que en todo momento dichos datos sean correctos, sin repeticiones innecesarias, datos perdidos y relaciones mal resueltas.

JDBC

La Conectividad de Bases de Datos Java (Java Database Connectivity, JDBC) es una especificación de la interfaz de aplicación de programa (application program interface, API) para conectar los programas escritos en Java a los datos en bases de datos populares.

Lenguaje de consulta

Son los lenguajes en el que los usuarios solicitan información de la base de datos. Estos lenguajes son generalmente de más alto nivel que los lenguajes de programación. Los lenguajes de consulta pueden clasificarse como procedimentales y no procedimentales.

Modelo de base de datos orientado a objetos

Es una adaptación a los sistemas de bases de datos. Se basa en el concepto de encapsulamiento de datos y código que opera sobre estos en un objeto.

Modelos de Red

Este modelo permite la representación de muchos a muchos de una Base de Datos. El modelo de red evita redundancia en la información, a través de la incorporación de un tipo de registro denominado el conector.

Nivel lógico

Definición de las estructuras de datos que constituyen la base de datos.

Reglas de Integridad

Son restricciones que definen los estados de consistencias de las bases de datos.

Registro

Un registro es el conjunto de información referida a una misma unidad.

Relación

El objetivo de estas relaciones sería principalmente evitar la duplicidad de información y en consecuencia, optimizar el rendimiento de la base de datos.

Recuperación

Proporcionar como mínimo el mismo nivel de recuperación que los sistemas de bases de datos actuales. De forma que, tanto en caso de fallo de hardware como de fallo de software, el sistema pueda retroceder hasta un estado coherente de los datos.

Sistema de Administración de Base de Dato

Es el software que controla la organización, almacenamiento, recuperación, seguridad e integridad de los datos en una base de datos.

SISTEMA GESTOR DE BASE DE DATOS

Es un conjunto de programas que permiten crear y mantener una base de datos, asegurando su integridad, confidencialidad y seguridad.

Software

Es un sistema manejador de bases de datos que permite al usuario accesar con facilidad a los datos almacenados o que ande ser almacenados

Tabla

Unidad donde crearemos el conjunto de datos de nuestra base de datos. Estos datos estarán ordenados en columnas verticales. En ella se definen los campos y sus características.

Transacción

Es una secuencia de operaciones de acceso a la base de datos que constituye una unidad lógica de ejecución.

Transacciones compartidas

Las transacciones compartidas soportan grupos de usuarios en estaciones de trabajo, los cuales desean coordinar sus esfuerzos en tiempo real, los usuarios pueden compartir los resultados intermedios de una base de datos. La transacción compartida permite que varias personas intervengan en una sola transacción.

Tupla

También se denomina de este modo a un registro o fila de una tabla.

Usuario final

Es quien accesa a las bases de datos por medio de un lenguaje de consulta o de programas de aplicación.

Valor nulo

Representa un valor para un atributo que es actualmente desconocido o no es aplicable para ese registro.

Vista

El resultado dinámico de una o más operaciones relacionales que operan sobre las relaciones base para producir otra relación. Una vista es una relación virtual que no tiene por qué existir necesariamente en la base de datos, sino que puede producirse cuando se solicite por parte de un usuario concreto, generándose en el momento de la solicitud.

6. REFERENTES BIBLIOGRÁFICOS

- Hawkins, D. M. (1980). Identification of outliers (Vol. 11). London: Chapman and Hall.
- Aldás Manzano, J., & Uriel Jimenez, E. (2017). Análisis multivariante aplicado con R. Ediciones Paraninfo, SA.

7. CONTROL DEL DOCUMENTO

	Nombre	Cargo	Dependencia	Fecha
Autor (es)	**JOSE FERNANDO GALINDO SUAREZ**	**INSTRUCTOR**	**CGMLTI**	**20/01/2023**

8. CONTROL DE CAMBIOS (diligenciar únicamente si realiza ajustes a la guía)

	Nombre	Cargo	Dependencia	Fecha	Razón del Cambio
Autor (es)					