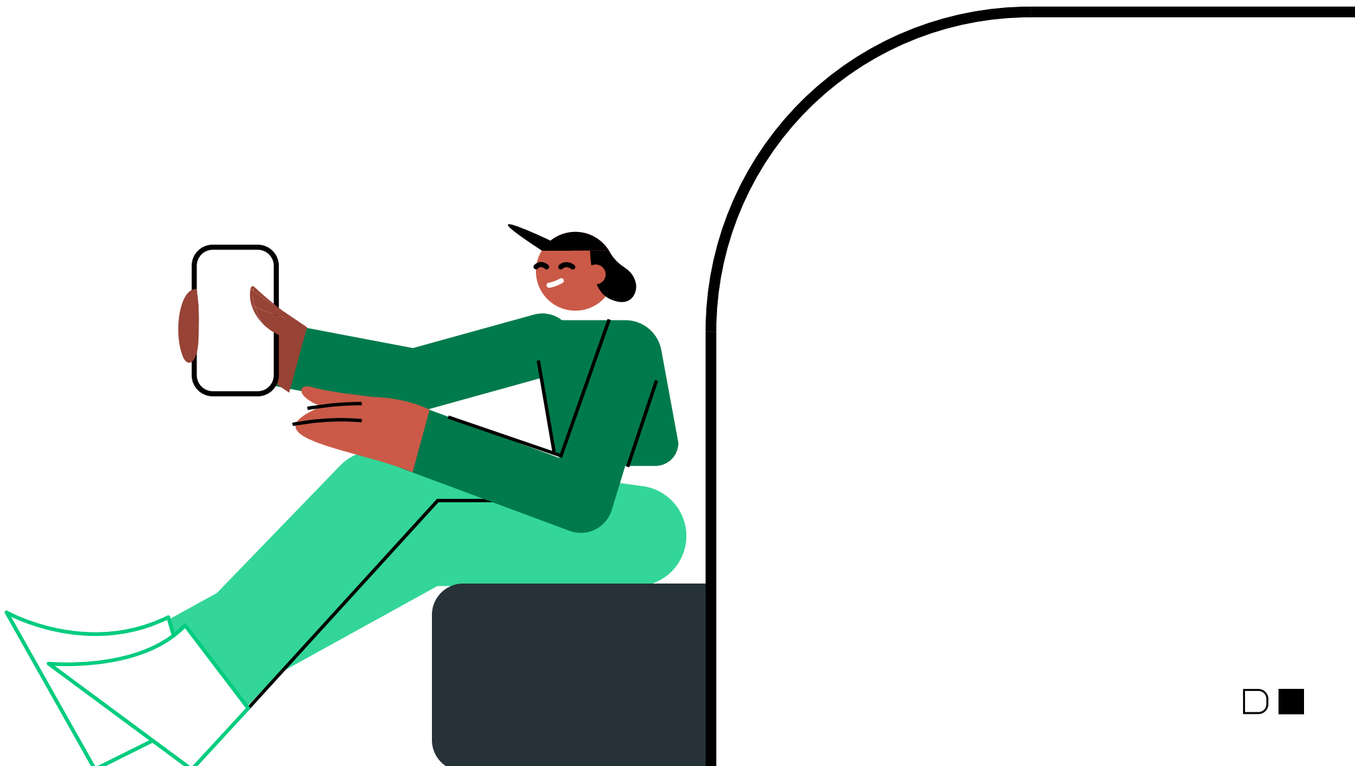
DIGITALHOUSE

Etapa II

Limpieza y Transformación de Datos

Índice



01 Objetivos

Objetivos

- **Calidad de los Datos:** El objetivo principal de esta etapa es asegurar que los datos sean de la más alta calidad posible, eliminando inconsistencias, manejando valores faltantes, y resolviendo cualquier problema de integridad referencial.
- **Preparación para el Análisis:** Se espera que los estudiantes transformen los datos de manera que estén listos para un análisis más profundo y para la creación de visualizaciones avanzadas en la siguiente etapa.
- **Documentación y Reproducibilidad:** Es crucial que cada paso del proceso de limpieza y transformación esté bien documentado para garantizar que el trabajo sea reproducible.

02

Contenidos y Herramientas Clave

Contenidos y Herramientas Clave

SQL y Python

- **SQL:** Para la extracción de datos, creación de nuevas tablas, y realización de consultas que apoyen la limpieza de datos.
- **Pandas:** Para la manipulación y transformación de datos, manejo de valores nulos, y creación de nuevas columnas derivadas.

Técnicas de Limpieza de Datos:

- Detección y manejo de valores faltantes.
- Identificación y tratamiento de outliers.
- Normalización y escalado de variables.

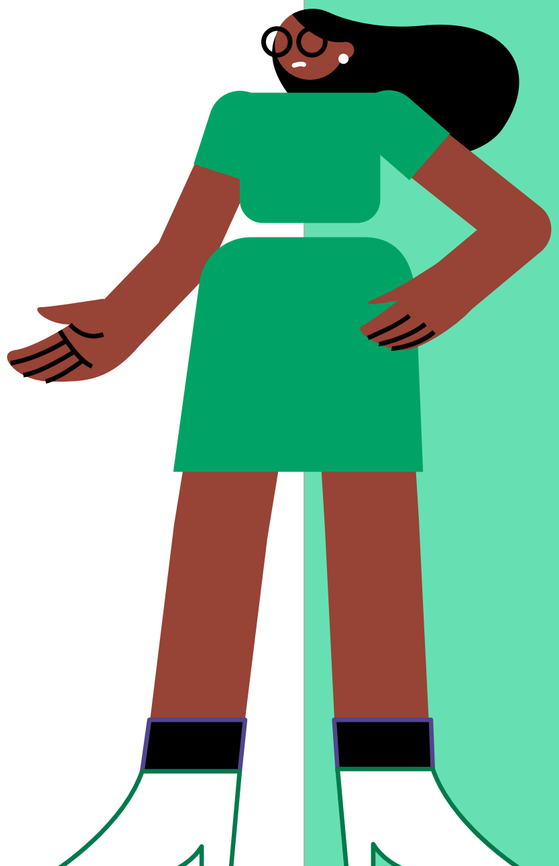
Tips

Revisión de Consistencia

Asegúrate de revisar la consistencia entre diferentes tablas o fuentes de datos.

Automatización

Cuando sea posible, automatiza procesos de limpieza para aplicarlos a futuras versiones de los datos.



03

Proceso de Limpieza de Datos

Proceso de Exploración de Datos

Identificación de Valores Faltantes

- Localiza y maneja valores nulos o faltantes en el dataset.

Detección y Tratamiento de Outliers y Duplicados

- Usa técnicas estadísticas y visuales para identificar outliers.
- Eliminación de duplicados o corrección de errores básicos en los datos.

Normalización y Escalado de Datos

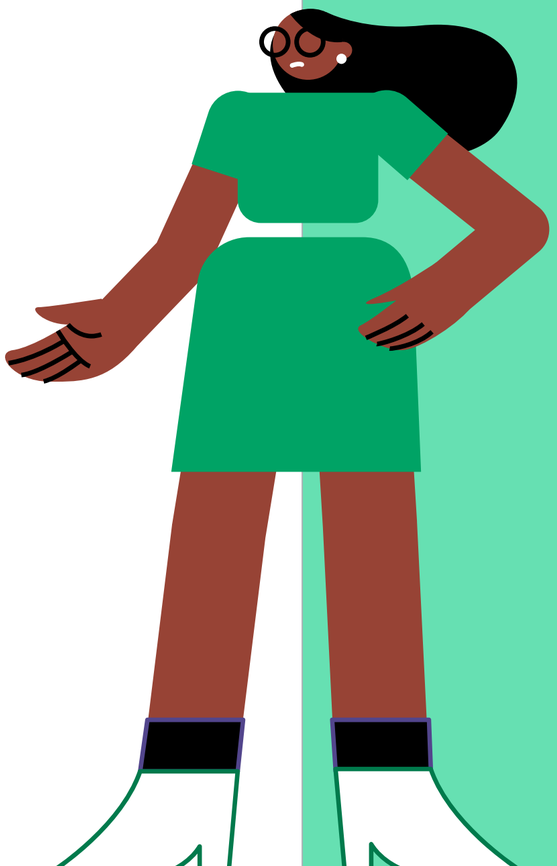
- Ajusta las variables numéricas para que estén en la misma escala, si es necesario para el análisis posterior

Transformación de Variables

- Creación de nuevas variables derivadas, agregación de datos, o cambios en el formato de los datos (por ejemplo, fechas).

Tips

- Utiliza df.isnull().sum() para identificar rápidamente las columnas con valores faltantes y df.dropna() o df.fillna() para tratarlos según el contexto.
- Los boxplots y Z-scores son herramientas útiles para detectar outliers. Decide si eliminarlos, transformarlos o tratarlos de alguna otra manera.
- Utiliza técnicas como Min-Max o Estandarización (Z-score) según el tipo de análisis que realizarás.
- Usa pd.to_datetime() para manejar fechas y groupby() para agregaciones.



04

Documentación del Proceso ETL

Documentación del Proceso ETL

Extract

Detalla cómo se han extraído los datos de las fuentes originales, incluyendo las consultas SQL utilizadas.

Transform:

Documenta cada paso del proceso de transformación, explicando por qué se tomaron ciertas decisiones (e.g., eliminación de outliers, técnicas de normalización).

Load:

Explica cómo se han almacenado los datos transformados, asegurando que están listos para ser utilizados en el análisis y visualización posteriores.

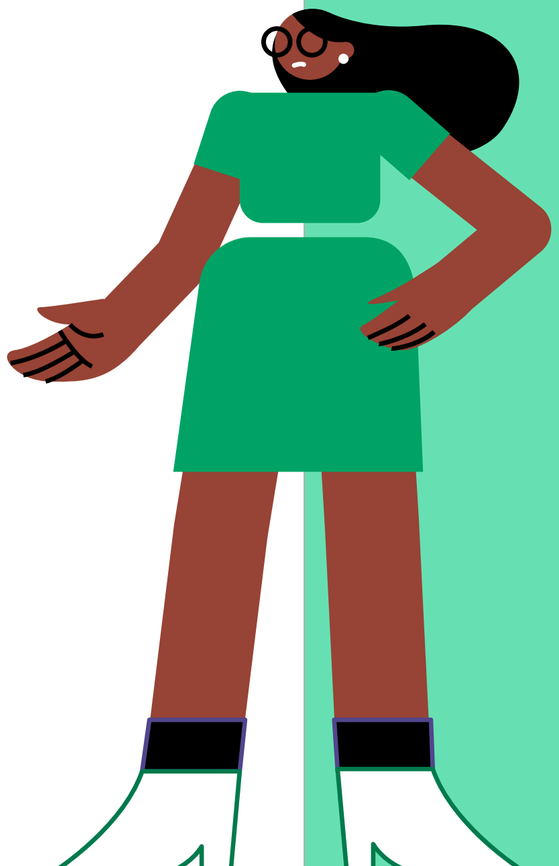
Tips

Clara Justificación

Justifica cada decisión de transformación de manera clara, explicando por qué se eligió una técnica específica para la limpieza o transformación de los datos.

Reproducibilidad

Asegúrate de que todo el proceso esté bien documentado para que pueda ser reproducido por otros.



05 Entregables Esperados

Entregables Esperados

01 **Scripts de Limpieza y Transformación de Datos**

02 **Dataset Final Transformado**

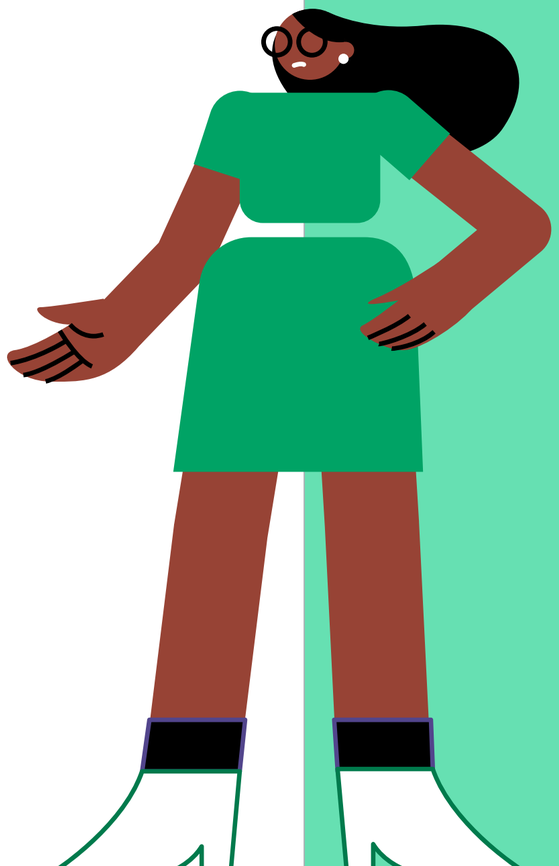
03 **Documentación del Proceso de ETL**

Scripts de Limpieza y Transformación de Datos

Código en Python y consultas SQL que documentan cómo se han limpiado y transformado los datos.

Tips

Asegúrate de que el código esté bien organizado, comentado, y que sea fácil de seguir.

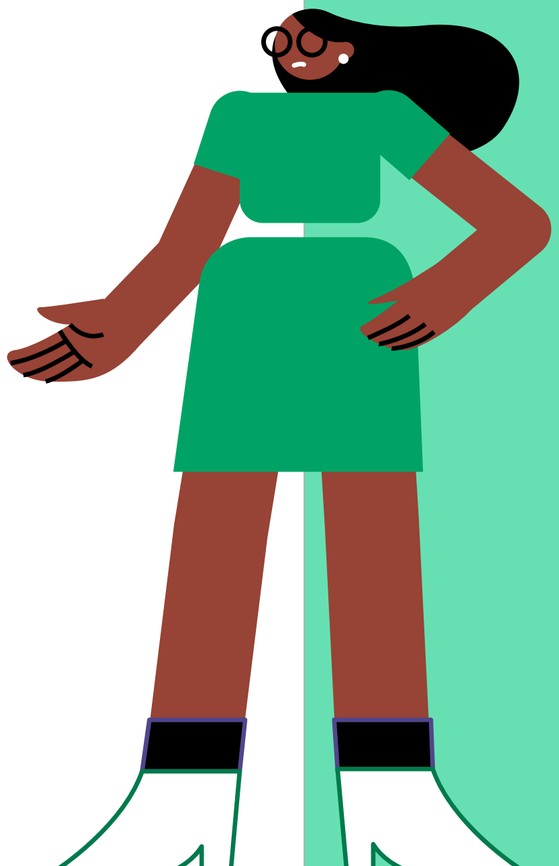


Dataset Final Transformado

Dataset limpio y transformado, listo para ser utilizado en la etapa de visualización.

Tips

Verifica que todos los campos sean consistentes y que no haya errores o incoherencias en los datos finales.

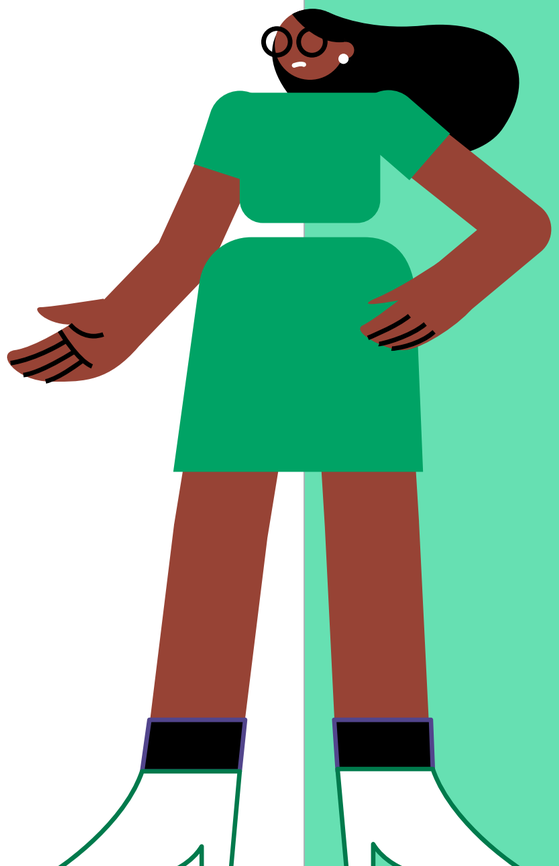


Documentación del Proceso de ETL

Documento que explique detalladamente cada paso del proceso de extracción, transformación y carga (ETL) de los datos.

Tips

Mantén la documentación clara y concisa, pero lo suficientemente detallada para que cualquier persona pueda seguir el proceso.



06 Evaluación

Criterios de Evaluación

- Calidad de los datos procesados, incluyendo la detección y manejo de outliers y valores faltantes.
- Eficiencia en la aplicación de transformaciones y técnicas de normalización.
- Documentación clara y completa del proceso ETL.
- Reproducibilidad del proceso de limpieza y transformación.

Con los datos ahora limpios y transformados, estarás preparado para iniciar la Etapa 3. Asegúrate de que todos los datos estén en las mejores condiciones posibles para facilitar el análisis y la visualización final.

La limpieza y transformación de datos es una de las tareas **más críticas** en cualquier proyecto de análisis de datos. Hazlo bien ahora, y te facilitará enormemente las etapas posteriores del proyecto.

DIGITALHOUSE

¡Muchas gracias!