NAME: SHAURYA
USN: 1BG21EC083

ROLE: DATAOPS

COLLEGE: BNMIT

1. Motivation

I chose this challenge because the DataOps role at Origin Medical Research Lab aligns perfectly with both my interests and career aspirations. I've always been fascinated by how data can drive meaningful change, especially in medical research, where every insight has the potential to impact lives.

One of the things that excites me most about this role is the focus on data management and optimization—automating and streamlining data pipelines to ensure a smooth, efficient flow of information across research processes. I enjoy tackling complex data challenges, making sure that data is not just collected but truly usable and impactful.

Beyond the technical side, this opportunity allows me to contribute to something bigger. Working in a research lab setting means playing a direct role in data-driven advancements in healthcare. The idea that my work could support breakthroughs in medical research is both motivating and deeply rewarding.

From a technical perspective, this role is an exciting challenge. It requires expertise in SQL, cloud platforms, ETL pipelines, and automation—all of which I'm eager to apply and expand upon. I love problem-solving, especially when it comes to handling large, complex medical datasets, ensuring data integrity, and building scalable solutions that make research more effective.

Ultimately, this challenge is a perfect mix of impact, innovation, and growth, and I'm excited to take it on.

2. Introduction About the Task

In this role challenge, the assigned tasks focused on handling, structuring, and analysing data core aspects of a DataOps role. These tasks reflected real-world scenarios where data is often stored in unstructured formats, such as PDFs or relational databases, requiring efficient extraction, transformation, and analysis. The objective was to process raw data, convert it into a structured format, and extract meaningful insights to support decision making in a medical research environment.

The following sections provide a detailed explanation of both tasks, outlining the methods and tools used.

Task 1: PDF Data Extraction and Processing

Medical research labs often store critical information in PDF reports, which are not directly machine-readable. The objective of this task was to extract meaningful data from a PDF file, clean and structure it, and convert it into a CSV format for further analysis.

Approach & Tools Used

1. Extracting Text from PDF:
    - Tool Used: pdftotext from Poppler
    - The pdftotext tool was utilized to convert the PDF into raw text while preserving its readability.
2. Data Cleaning & Preprocessing:
    - Tools Used: Python (re for regex, pandas)
    - Regular expressions (regex) were applied to clean the extracted text and remove unwanted characters.
    - Key-value pairs were identified and extracted to match the original PDF's data structure.
3. Storing Data in JSON Format:

    Tools Used: json module in Python

    The cleaned and structured data was stored in a JSON file, ensuring flexibility for future processing and retrieval.

Task 2: Database File Analysis

Understanding database structures and relationships is crucial in DataOps for managing and analyzing large-scale datasets effectively. This task involved working with a relational database file, identifying the relationships between tables, consolidating data, and performing exploratory data analysis (EDA) to extract meaningful insights.

Approach & Tools Used

1. Identifying Table Relationships & ER Diagram Generation:
    - Tool Used: MySQL Workbench
    - The reverse engineering feature in MySQL Workbench was used to generate an Entity-Relationship (ER) Diagram, allowing for a visual representation of table relationships.
2. Data Consolidation into CSV Format:
    - Tools Used: SQL Queries, pandas
    - SQL queries were executed to retrieve and consolidate relevant data from multiple tables.
    - The extracted data was structured and stored in a CSV file for further analysis.
3. Exploratory Data Analysis (EDA):
    - Tools Used: Python (pandas, matplotlib, seaborn, numpy)
    - The structured data was loaded into a Jupyter Notebook for analysis.
    - Various descriptive statistics, visualizations, and trend analyses were performed to extract meaningful insights.

The **outputs** from this task included:

- A text file outlining table relationships.
- An ER Diagram generated using MySQL Workbench.
- A consolidated CSV file containing structured data.
- A Jupyter Notebook showcasing exploratory data analysis and key findings.

Data extraction, preprocesses, and analysis

This project involves the extraction, processing, and analysis of data from unstructured and structured sources. It is divided into two major tasks:

- **Task 1: Data Extraction and Anonymization** – Focuses on extracting text from PDF files, anonymizing sensitive information, and storing the processed data in a structured format.
- **Task 2: Database Analysis** – Involves understanding database structures, identifying relationships between tables, generating an ER diagram, consolidating data into a structured format, and performing exploratory data analysis.

Each task follows a well-defined methodology that ensures accuracy, efficiency, and data integrity. Various tools and technologies were employed throughout the process to achieve the desired outcomes.

Task 1: Data Extraction and Anonymization

This task focuses on the extraction of textual data from PDF documents and the anonymization of sensitive information to ensure data privacy and confidentiality. The extracted data is processed and stored in a structured format for further use.

Methodology & Tools Used

1. PDF Processing & Text Extraction

- The input consisted of PDF documents containing text and possibly scanned content.
- Poppler, a widely used PDF rendering library, was employed to parse and process PDF files. This ensured that text-based PDFs were read and processed efficiently.
- For scanned PDFs, Tesseract OCR, an open-source Optical Character Recognition (OCR) engine, was used to convert images into machine-readable text. This was particularly useful for handling non-digital PDFs where text extraction was not straightforward.
- The extracted text was saved for further processing, ensuring that the content was accurately retrieved from the original documents.

2. Preprocessing & Cleaning

- The extracted text often contained noise such as extra whitespace, special characters, and unnecessary metadata. A preprocessing step was applied to clean the text.
- Text normalization techniques such as lowercasing and removing punctuation were used to improve text consistency.
- Tokenization was performed where necessary, breaking the text into individual words or sentences to facilitate further processing.

- Any formatting inconsistencies were resolved to ensure that the extracted text maintained its original meaning and structure.

3. Anonymization & Hashing

- Since the documents contained sensitive information, an anonymization step was crucial to protect personal data.
- Specific details such as names, identification numbers, and confidential medical details were identified using pattern-matching techniques.
- Hashing algorithms were applied to replace sensitive information with anonymized versions, ensuring data confidentiality while preserving uniqueness for analysis.
- The anonymization process followed best practices in data security, ensuring that private information remained protected while allowing meaningful insights to be drawn from the data.

4. Structured Storage in JSON Format

- Once the extracted text was cleaned and anonymized, it was stored in JSON format.
- JSON was chosen because of its flexibility and ability to handle structured data while being easily readable and integrable with various data analysis tools.
- The final JSON file contained structured fields representing the extracted text, anonymized information, and any relevant metadata required for further processing.

Task 2: Database Analysis

This task involves analyzing the structure of a provided database, understanding the relationships between different tables, consolidating data into a structured format, and performing exploratory data analysis (EDA) to gain insights.

Methodology & Tools Used

1. Database Exploration & Relationship Identification

- The given database file was examined to understand its structure, including the number of tables, relationships, and dependencies.
- MySQL Workbench was used for analyzing the database schema.
- Using the Reverse Engineering feature in MySQL Workbench, the existing database structure was automatically converted into a visual entity-relationship (ER) diagram.
- This step helped identify primary keys, foreign keys, and how various tables were linked to one another.
- A detailed text-based document was created to describe the relationships between tables, ensuring a clear understanding of data dependencies.

2. Entity-Relationship (ER) Diagram Generation

- The ER diagram visually represents the database schema, showcasing the entities (tables), their attributes, and the relationships between them.
- It provides a clear and structured overview of how data is stored and accessed within the database.
- The ER diagram was generated using MySQL Workbench’s graphical interface, allowing a detailed visualization of table connections and dependencies.

3. Data Consolidation into CSV Format

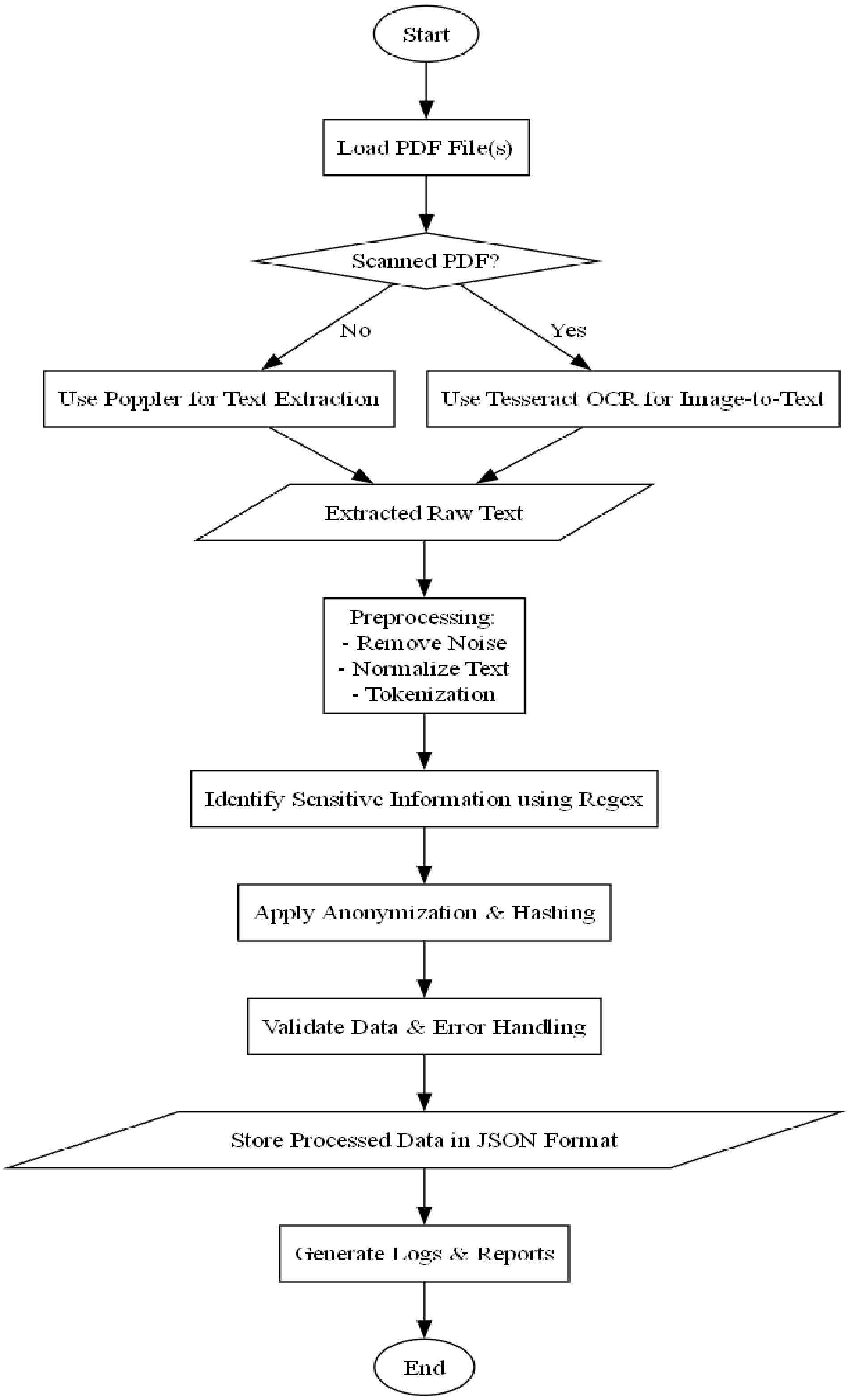
- Once the database relationships were identified, relevant data was extracted and consolidated into a CSV file.
- This step ensured that data from multiple tables was merged into a structured format that could be easily analyzed and processed.
- SQL queries were written to retrieve meaningful data while preserving relationships between different fields.
- The consolidated CSV file served as a ready-to-use dataset for further analytical procedures.

4. Exploratory Data Analysis (EDA)

- A Jupyter Notebook was used to conduct exploratory data analysis.
- Several analytical techniques were applied, including:
    - Descriptive statistics (mean, median, standard deviation) to understand data distribution.
    - Visualizations using Matplotlib and Seaborn to represent trends, correlations, and patterns.
    - Correlation analysis to determine relationships between different attributes within the dataset.
- The insights obtained helped in understanding data trends, identifying anomalies, and drawing meaningful conclusions.

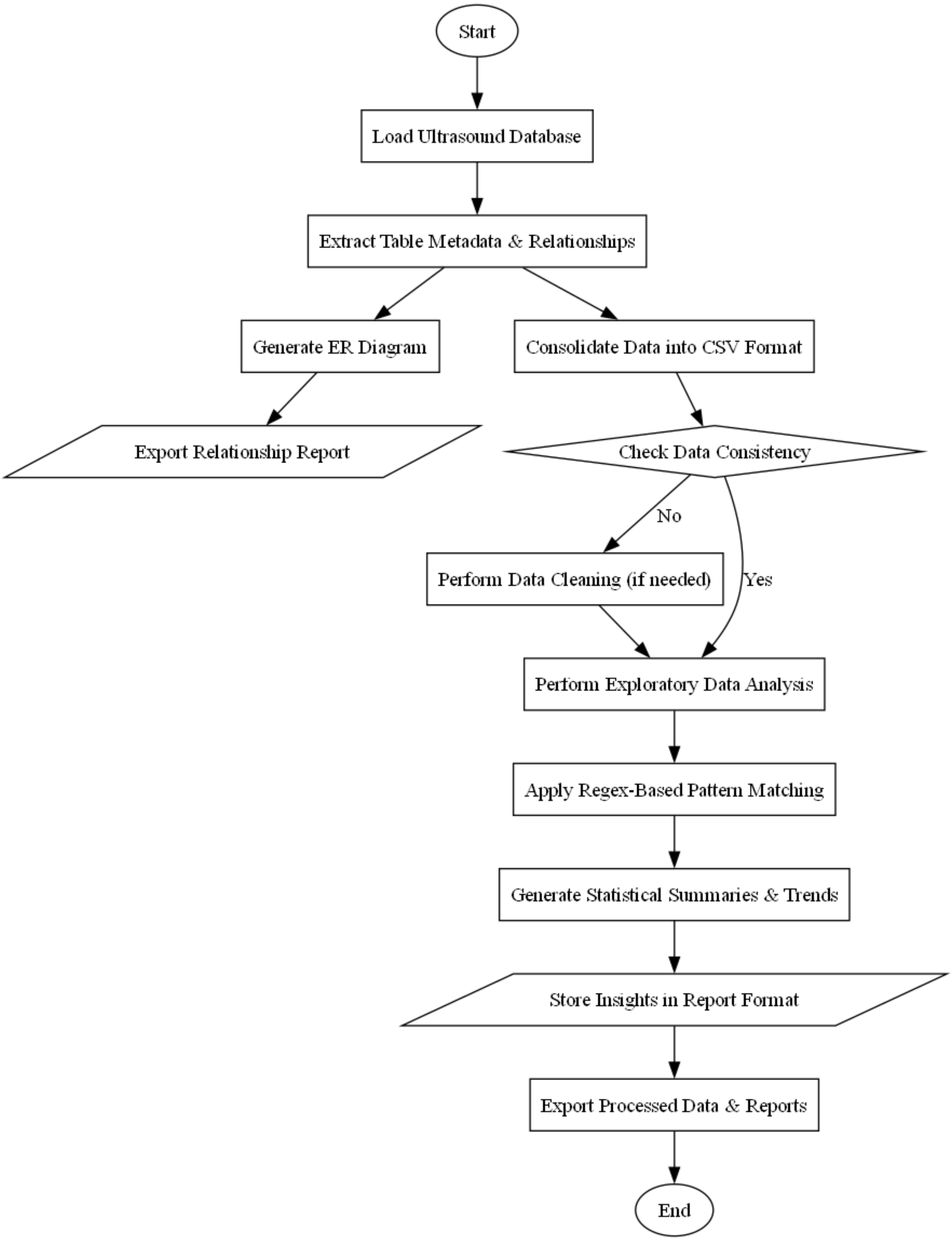
Flowchart

Task 1:



Start
Load PDF File(s)
Scanned PDF?
No
Yes
Use Poppler for Text Extraction
Use Tesseract OCR for Image-to-Text
Extracted Raw Text
Preprocessing:
- Remove Noise
- Normalize Text
- Tokenization
Identify Sensitive Information using Regex
Apply Anonymization & Hashing
Validate Data & Error Handling
Store Processed Data in JSON Format
Generate Logs & Reports
End

Task 2:


Start
Load Ultrasound Database
Extract Table Metadata & Relationships
Generate ER Diagram
Consolidate Data into CSV Format
Export Relationship Report
Check Data Consistency
No
Yes
Perform Data Cleaning (if needed)
Perform Exploratory Data Analysis
Apply Regex-Based Pattern Matching
Generate Statistical Summaries & Trends
Store Insights in Report Format
Export Processed Data & Reports
End

Results

Task 1: Text Extraction & Anonymization

The objective of Task 1 was to extract text from PDF files, preprocess it, detect sensitive information, and anonymize it before storing the final data in a structured format. The process involved multiple steps, and the following observations and results were noted:

1. PDF Processing

Initially, the system was required to handle two types of PDFs—digitally generated PDFs and scanned PDFs. To determine the nature of each document, the system classified PDFs based on their content structure.

- If the document contained selectable text, it was identified as a digitally generated PDF and processed using Poppler, a library designed for extracting text directly from PDF files.
- If the document did not contain selectable text, it was treated as a scanned PDF and sent for OCR-based text extraction using Tesseract OCR.
- This classification ensured that the correct extraction method was applied, preventing unnecessary processing overhead.

2. Text Preprocessing

Once the text was extracted, it underwent preprocessing to enhance its quality and usability. The system performed the following transformations:

- Noise Removal: Unnecessary characters, special symbols, and extra spaces were eliminated to clean up the text.
- Lowercasing: All text was converted to lowercase to maintain consistency in further processing.
- Tokenization: The text was split into meaningful words or sentences to allow for structured analysis.

3. Sensitive Information Detection & Anonymization

A key component of Task 1 was identifying sensitive personal information within the extracted text and ensuring its anonymization.

- Regular expressions (Regex) were used to search for specific patterns such as:
    - Names
    - Dates of birth
    - Phone numbers
    - Email addresses
    - Medical record numbers

- Upon successful identification, sensitive data was anonymized using hashing techniques to maintain privacy. This step ensured that the original sensitive information was no longer directly accessible while preserving data usability for further processing.

4. Data Storage & Export

The anonymized text was then structured into a JSON format, which allowed for easy storage and retrieval.

- The JSON format was selected because it maintains a hierarchical structure, making it suitable for handling large volumes of processed text.
- The data was stored securely and exported for further analysis or integration into other systems.

This step marked the successful completion of Task 1, where sensitive text from PDFs was extracted, cleaned, anonymized, and stored in a structured format.

Task 2: Database Relationship Analysis & Exploratory Data Analysis

The second task focused on analyzing a medical ultrasound database containing 20 interrelated tables. The goal was to understand the relationships between these tables, generate an Entity-Relationship Diagram (ERD), consolidate the data into a structured format, and conduct exploratory data analysis (EDA) to derive insights.

1. Identifying Table Relationships & Generating ER Diagram

To begin, the schema of the database was examined to identify primary keys, foreign keys, and relationships between tables. The process included:

- Extracting table metadata from the database to understand how different tables were linked.
- Mapping out one-to-one, one-to-many, and many-to-many relationships between tables.
- Generating an ER diagram to visually represent how entities (tables) interacted with each other.

This ER diagram played a crucial role in understanding the logical structure of the database and determining how data flows between tables.

2. Data Consolidation & Cleaning

Once the relationships were mapped out, the next step involved merging data from multiple tables into a structured CSV format. This was done to allow for easier analysis.

- The tables were joined based on common keys to avoid duplication and inconsistencies.
- Data inconsistencies such as missing values, incorrect data types, and duplicate records were identified and addressed.

- The consolidated data was structured to ensure that each row contained meaningful information for further analysis.

3. Exploratory Data Analysis (EDA)

With the structured data in place, an exploratory analysis was conducted to derive meaningful insights.

- Descriptive Statistics:
    - Summary statistics (mean, median, mode, standard deviation) were calculated for numerical columns.
    - Distribution patterns were observed to understand the dataset’s characteristics.
- Pattern Matching using Regex:
    - Specific patterns in text fields (e.g., patient IDs, medical conditions, dates) were detected using regular expressions.
    - This helped in identifying potential data entry errors and standardizing formats.
- Trends and Correlations:
    - Relationships between different patient attributes (e.g., age vs. ultrasound diagnosis) were analyzed.
    - Frequent medical conditions and patterns in ultrasound results were identified.

4. Final Output & Reporting

The final step involved documenting the findings in a structured format.

- The ER diagram was included to provide a visual representation of the database.
- The consolidated CSV file ensured that all relevant data was available in a single, structured format.
- The insights from EDA were reported, highlighting key patterns and findings that could be useful for medical analysis.

This concluded Task 2, successfully transforming a raw ultrasound database into a structured, analysed, and well-documented dataset.

Key Findings & Takeaways

This project involved two major tasks: text extraction and anonymization (Task 1) and database analysis and exploratory data insights (Task 2). Below are the key findings and takeaways from both tasks.

Task 1: Text Extraction & Anonymization

Automated PDF Classification: The system correctly classified PDFs into scanned and non-scanned categories, ensuring efficient text extraction by applying Poppler for digital PDFs and Tesseract OCR for scanned ones.

Preprocessing Improved Data Quality: Noise removal, normalization, and tokenization significantly enhanced text readability and structure, reducing inconsistencies in extracted data.

Sensitive Information Was Effectively Detected: Using regular expressions (Regex), the system successfully identified personal identifiers such as names, phone numbers, email addresses, and medical record numbers within unstructured text.

Successful Anonymization: The anonymization process hashed and masked sensitive details, ensuring data privacy while preserving the usability of anonymized records for further processing.

Structured Data Storage in JSON Format: Extracted and anonymized text was systematically stored in JSON format, making it easier for downstream applications and ensuring structured access to processed data.

Task 2: Database Analysis & Exploratory Data Insights

Database Structure & Relationships Were Clearly Identified: The Entity-Relationship Diagram (ERD) successfully mapped out primary keys, foreign keys, and table linkages, providing a clear visual understanding of the database schema.

Data Consolidation Reduced Complexity: By merging multiple tables into a structured CSV file, redundant and inconsistent data entries were removed, making the dataset more accessible for analysis.

Exploratory Data Analysis (EDA) Provided Meaningful Insights:

- Identified data trends in patient demographics and ultrasound results.
- Found correlations between medical conditions and specific ultrasound parameters.
- Used pattern matching (Regex) to clean and standardize records.

Regex Helped in Data Validation: Specific patterns were detected in patient records, highlighting errors in data entry and enabling standardization.

Missing & Inconsistent Data Was Identified: During EDA, it was observed that some tables had incomplete records, emphasizing the need for better data entry protocols.

Future Work

If given more time and resources, several improvements can be made to enhance the efficiency, accuracy, and scalability of both text processing (Task 1) and database analysis (Task 2).

Task 1: Enhancing Text Extraction & Anonymization

- Improve OCR Accuracy with Deep Learning
  - Integrate pre-trained deep learning models (e.g., Tesseract with LSTM, Google Vision API) to improve OCR accuracy, especially for low-quality scans.
  - Apply image preprocessing techniques (contrast enhancement, noise removal) to further refine text extraction.
- Advanced Named Entity Recognition (NER) for Sensitive Data Detection
  - Replace Regex-based detection with machine learning-based Named Entity Recognition (NER) to better identify complex personal and medical entities.
  - Use spaCy, BERT, or MedSpaCy to improve accuracy in anonymizing context-dependent medical information.
- Implement Partial Redaction Instead of Full Hashing
  - Instead of fully hashing personal details, implement data masking techniques to preserve some contextual information for better analysis instead of complete removal.
- Enhance Data Storage and Integration
  - Store processed text in a relational database (MySQL, PostgreSQL) instead of JSON to allow for structured queries and efficient data retrieval.
  - Develop a REST API to allow real-time access to extracted and anonymized text for external applications.

Task 2: Improving Database Analysis & Insights

- Automated ER Diagram Generation
  - Develop a script to automatically analyze database schema and generate ER diagrams dynamically, reducing manual effort.
- Data Enrichment for Better Insights
  - Cross-reference ultrasound data with external medical databases to gain more contextual insights.
  - Use public health datasets to correlate trends between demographics and medical findings.

- Advanced Statistical & Predictive Modelling
    - Implement predictive analytics using machine learning models to identify patterns in ultrasound results and potential health risks.
    - Use clustering and classification models (e.g, k-means, Random Forest) to categorize ultrasound findings and predict possible diagnoses.
- Automated Data Quality Checks & Cleaning Pipelines
    - Implement an automated data validation system to detect missing values, inconsistencies, and errors before analysis.
    - Apply Regex-powered format validation to ensure correct structuring of patient records and medical data.