Legacy Document Processing Tool - User Guide [*Team Legacy Hackers*]

1. Contact Details of Team Members

Name	Email	Mobile	Course	Contribution
Arunangshu Karmakar	sutapak2013@gmail.com	9123003808	Mathematics and Computing	Pipeline Ideation, Database, Frontend Design
Malyadip Pal	malyadippaljee2023@gmail.com	8927087113	Instrumentation Engg.	Document Parsing, Data Extraction, UI Design
Aditya Debnath	adityadebnath0202@gmail.com	8191885726	Instrumentation Engg.	Backend (LLM, RAG and Agents Integration)
Akshat Jiwrajka	akshat2306jwr@gmail.com	8420028041	Mathematics and Computing	Backend (LLM, RAG and Agents Integration)
Shaunak Majumdar	onlyshaun17@gmail.com	9836404544	Mechanical Engg.	AI/ML Integration, Query Processing

Institute: Indian Institute of Technology Kharagpur

2. Hardware and Software Requirements

Hardware Requirements

- **Processor**: Intel Core i5 or equivalent (quad-core recommended)
- **RAM**: 8GB minimum, 16GB recommended
- **Storage**: 10GB free disk space
- **Internet Connection**: Required for API access and database operations

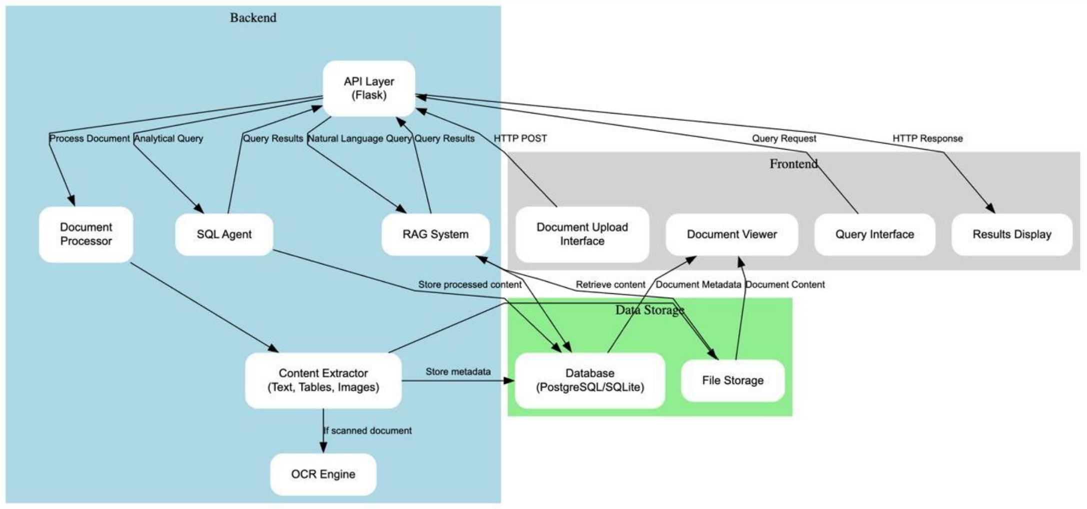
Software Requirements

- **Operating System**: Windows 10/11, macOS 10.15+, or Linux (Ubuntu 20.04+)
- **Python**: Version 3.8 or higher
- **Node.js**: Version 14.0 or higher
- **PostgreSQL**: Version 13.0 or higher (optional, SQLite can be used for development)
- **Web Browser**: Chrome (latest), Firefox (latest), or Edge (latest)

2. Setup Guide

Refer to our public Github repository:
https://github.com/Legacy-Hackers/Legacy-Document-Processing-Tool

3. Flow Diagram



4. List of Python Libraries

- **Web Framework**: Flask (2.2.3), flask-cors (3.0.10)
- **Environment Management**: python-dotenv (1.0.0)
- **Database**: SQLAlchemy (2.0.4), psycopg2-binary (2.9.5)
- **PDF Processing**: PyMuPDF (1.21.1), pdfplumber (0.7.6)
- **OCR**: pytesseract (0.3.10)
- **Image Processing**: opencv-python (4.7.0.72), Pillow (9.4.0)
- **AI/ML**: google-generativeai (0.3.1)
- **Data Processing**: numpy (1.24.2)

5. Environment Details

Backend Environment

- Flask server running on port 5000
- Development mode: Debug=True
- Database options:
  - PostgreSQL (recommended for production)
  - SQLite (for development)
- Environment variables configured in `.env` file

Frontend Environment

- React application running on port 3000
- TypeScript for type-safe code
- Material-UI for component styling
- React Router for navigation

6. List of APIs Used

- **Internal APIs**:
  - `/api/health` - Health check endpoint
  - `/api/documents` - Document management (GET, POST, DELETE)
  - `/api/query` - Query processing (RAG and SQL)
  - `/api/upload` - Document upload
  - `/api/documents/suggestions` - Document search suggestions
  - `/api/documents/<id>/tables` - Extract tables from documents
  - `/api/admin/cleanup` - Admin cleanup operations
- **External APIs**:
  - Google GenerativeAI API for natural language processing
  - OCR services for document text extraction

7. Test Cases Used for Training Model and Checking

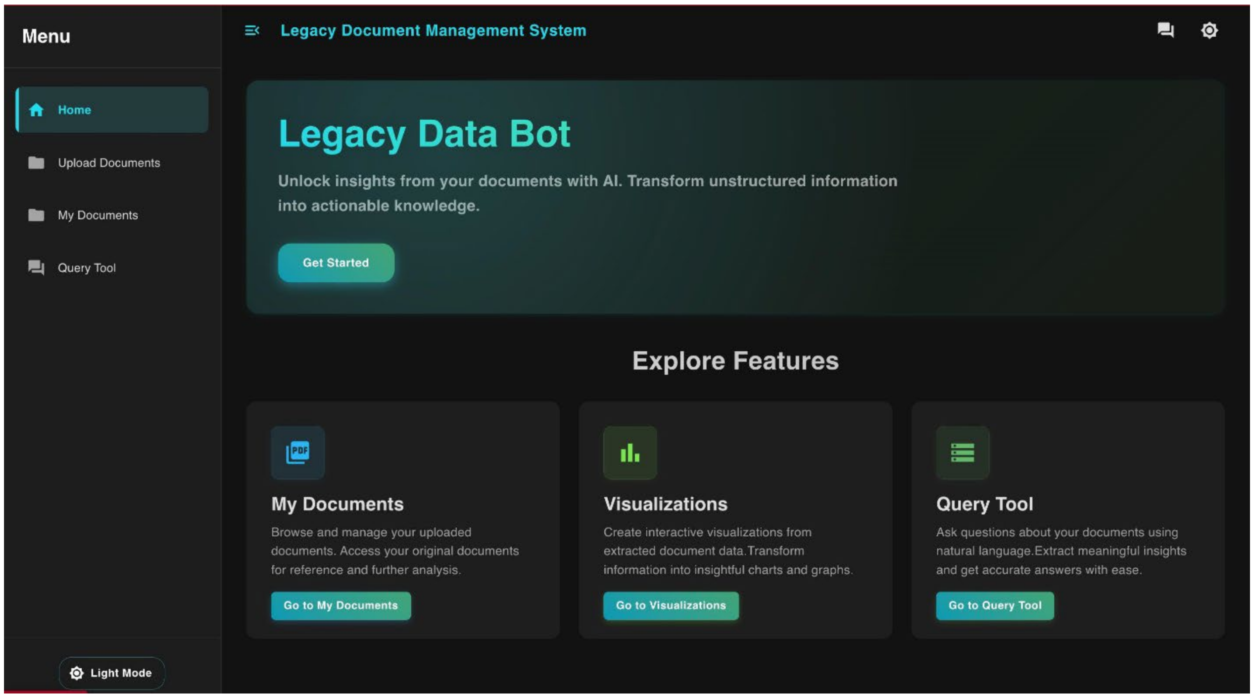
1. **Document Processing Tests**:
   - PDF with text only
   - PDF with mixed content (text, tables, images)
   - Scanned documents (low quality)
   - Documents with multiple tables
   - Multi-page documents
2. **Query Processing Tests**:
   - Simple text queries
   - Complex analytical queries
   - Table extraction requests
   - Document comparison requests
3. **Performance Tests**:
   - Large document processing (>200 pages)
   - Concurrent query handling
   - Database scaling tests

8. List of Required Files

- **Backend**:
  - `app.py` - Main Flask application
  - `services/database_manager.py` - Database connection and operations
  - `sql_agent.py` - SQL query processing
  - `RAG_system.py` - Retrieval-Augmented Generation system
  - `app/Content_Extractors/PdfContentExtractor.py` - PDF content extraction
  - `utils/` - Utility functions
  - `requirements.txt` - Python dependencies
  - `.env` - Environment configuration
- **Frontend**:
  - `package.json` - NPM dependencies
  - `src/` - React application source code
  - `public/` - Static assets

9. Screenshots of Execution

Dark Mode



Legacy Data Bot

Unlock insights from your documents with AI. Transform unstructured information into actionable knowledge.

Explore Features

My Documents

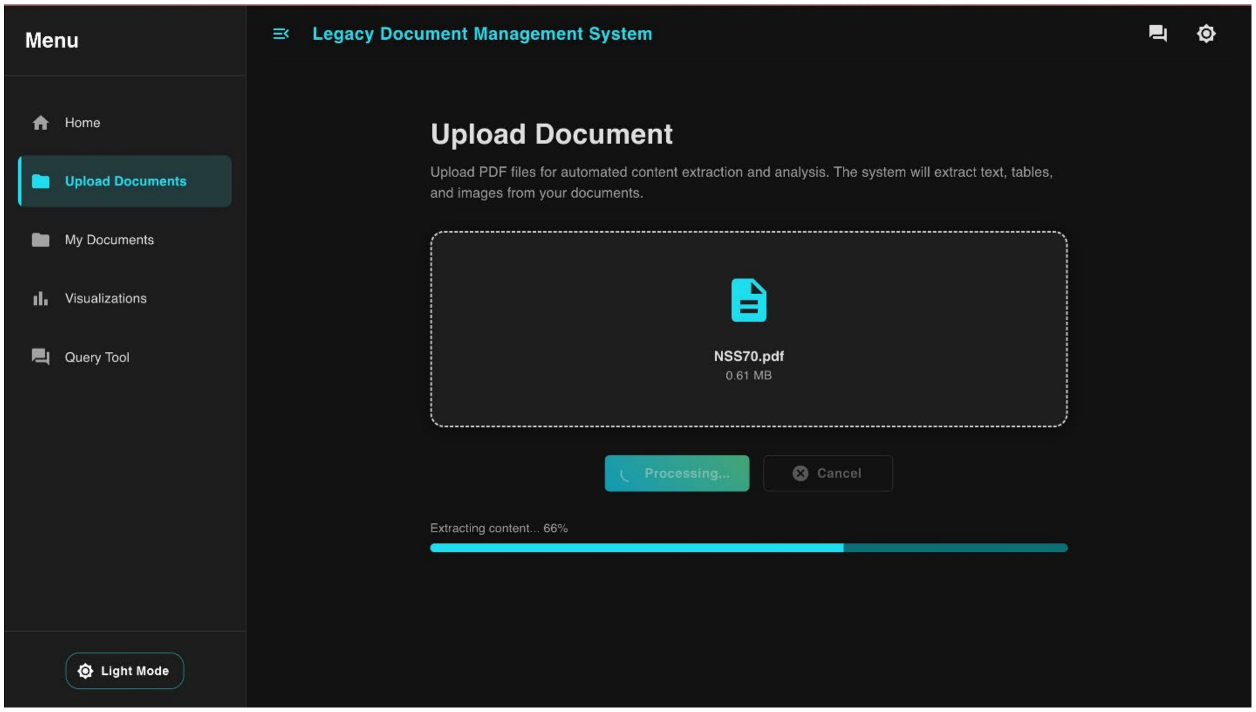
Browse and manage your uploaded documents. Access your original documents for reference and further analysis.

Visualizations

Create interactive visualizations from extracted document data.Transform information into insightful charts and graphs.

Query Tool

Ask questions about your documents using natural language.Extract meaningful insights and get accurate answers with ease.

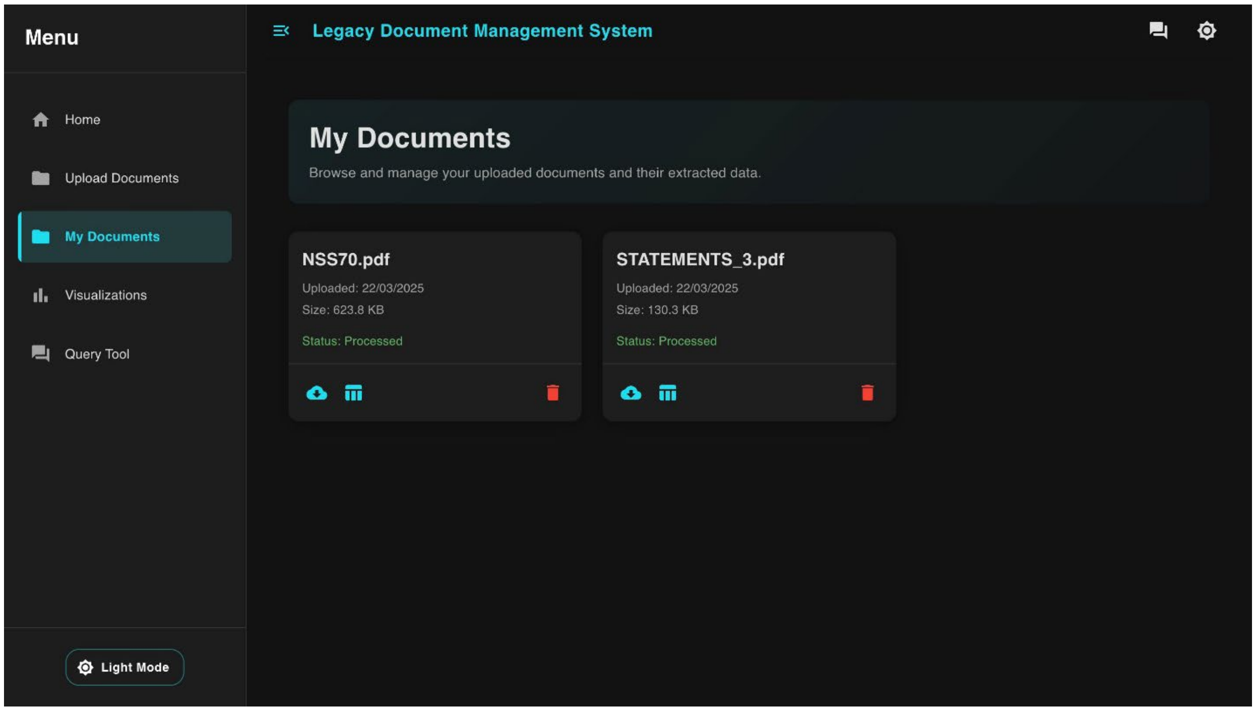


Upload Document

Upload PDF files for automated content extraction and analysis. The system will extract text, tables, and images from your documents.

NSS70.pdf
0.61 MB

Extracting content... 66%



Menu

Home

Upload Documents

My Documents

Visualizations

Query Tool

Light Mode

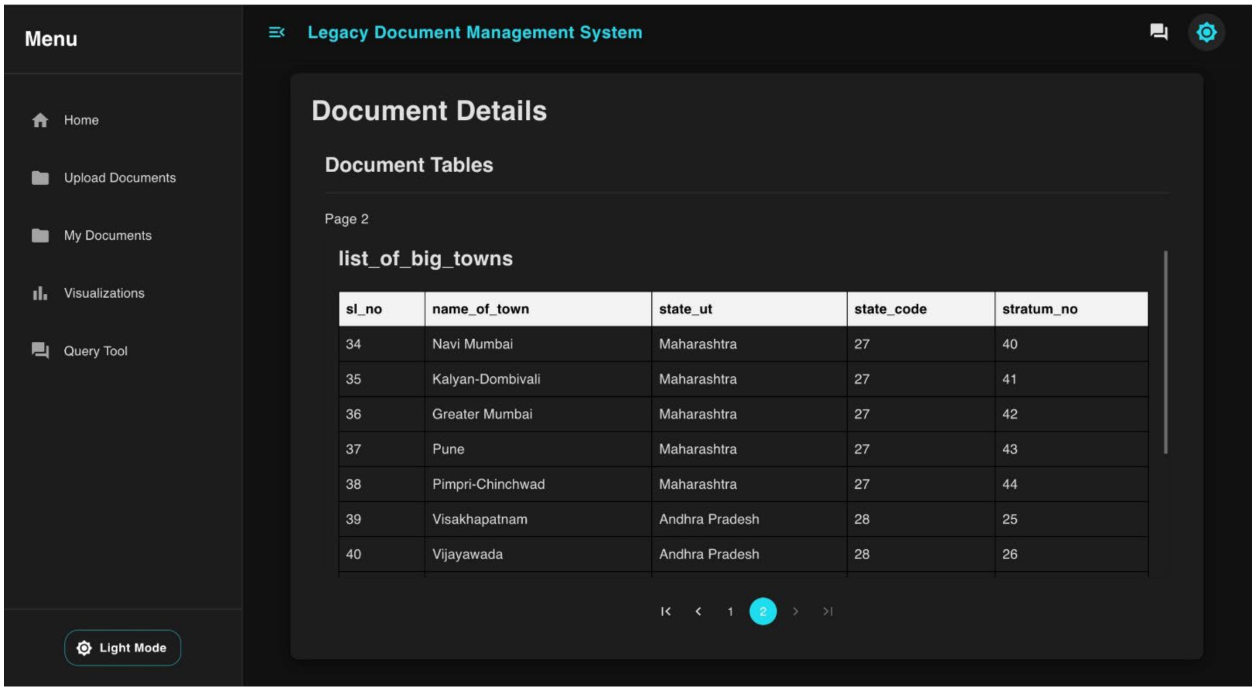
Legacy Document Management System

My Documents

Browse and manage your uploaded documents and their extracted data.

NSS70.pdf
Uploaded: 22/03/2025
Size: 623.8 KB
Status: Processed

STATEMENTS_3.pdf
Uploaded: 22/03/2025
Size: 130.3 KB
Status: Processed



Menu

Home

Upload Documents

My Documents

Visualizations

Query Tool

Light Mode

Legacy Document Management System

Document Details

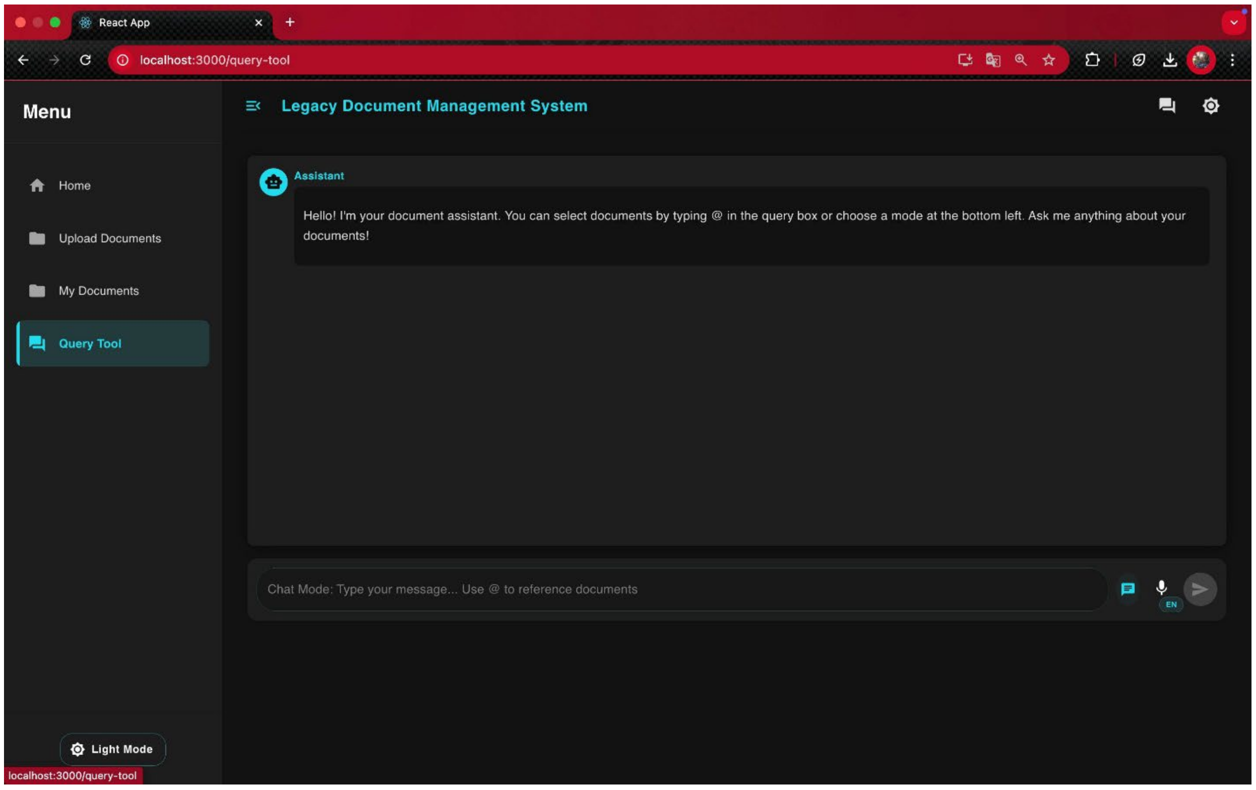
Document Tables

Page 2

list_of_big_towns

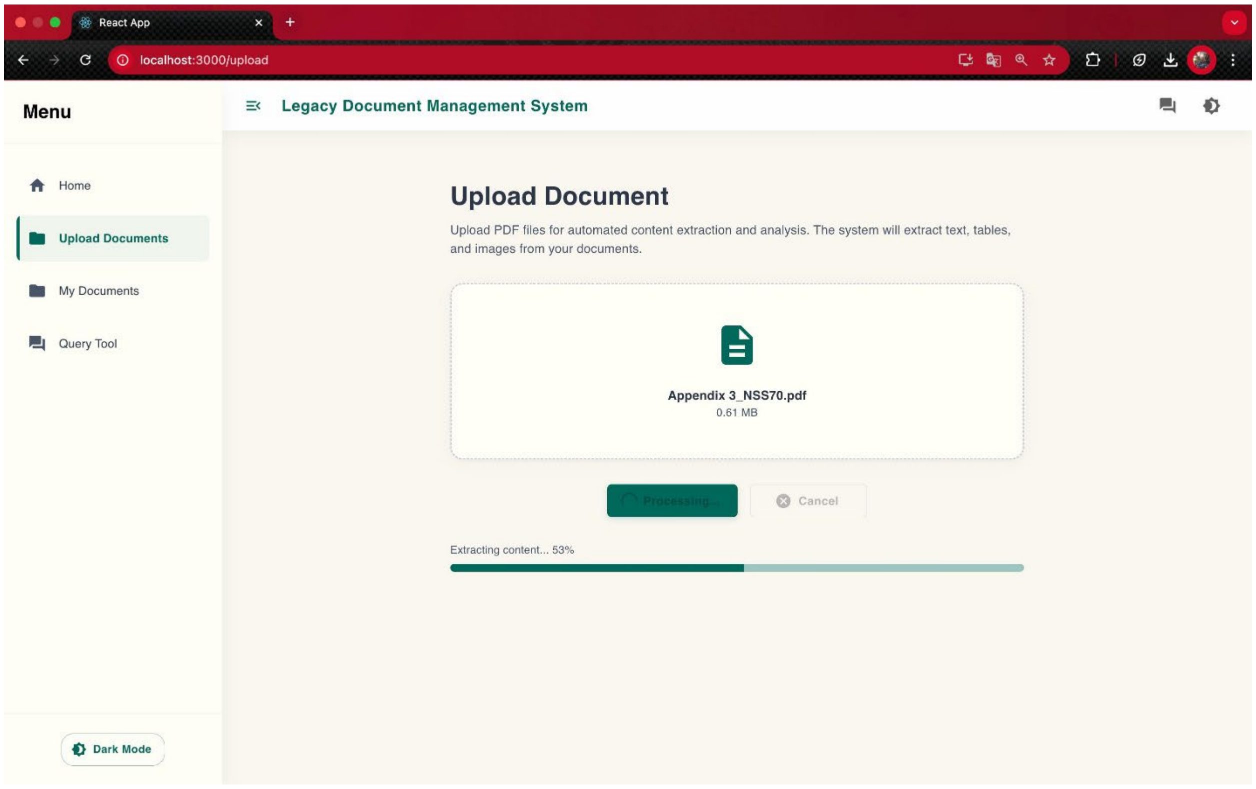
sl_no	name_of_town	state_ut	state_code	stratum_no
34	Navi Mumbai	Maharashtra	27	40
35	Kalyan-Dombivali	Maharashtra	27	41
36	Greater Mumbai	Maharashtra	27	42
37	Pune	Maharashtra	27	43
38	Pimpri-Chinchwad	Maharashtra	27	44
39	Visakhapatnam	Andhra Pradesh	28	25
40	Vijayawada	Andhra Pradesh	28	26

1 2


React App
localhost:3000/query-tool
Menu
Home
Upload Documents
My Documents
Query Tool
Light Mode
localhost:3000/query-tool
Legacy Document Management System
Assistant
Chat Mode: Type your message... Use @ to reference documents
EN

Hello! I'm your document assistant. You can select documents by typing @ in the query box or choose a mode at the bottom left. Ask me anything about your documents!

Light Mode

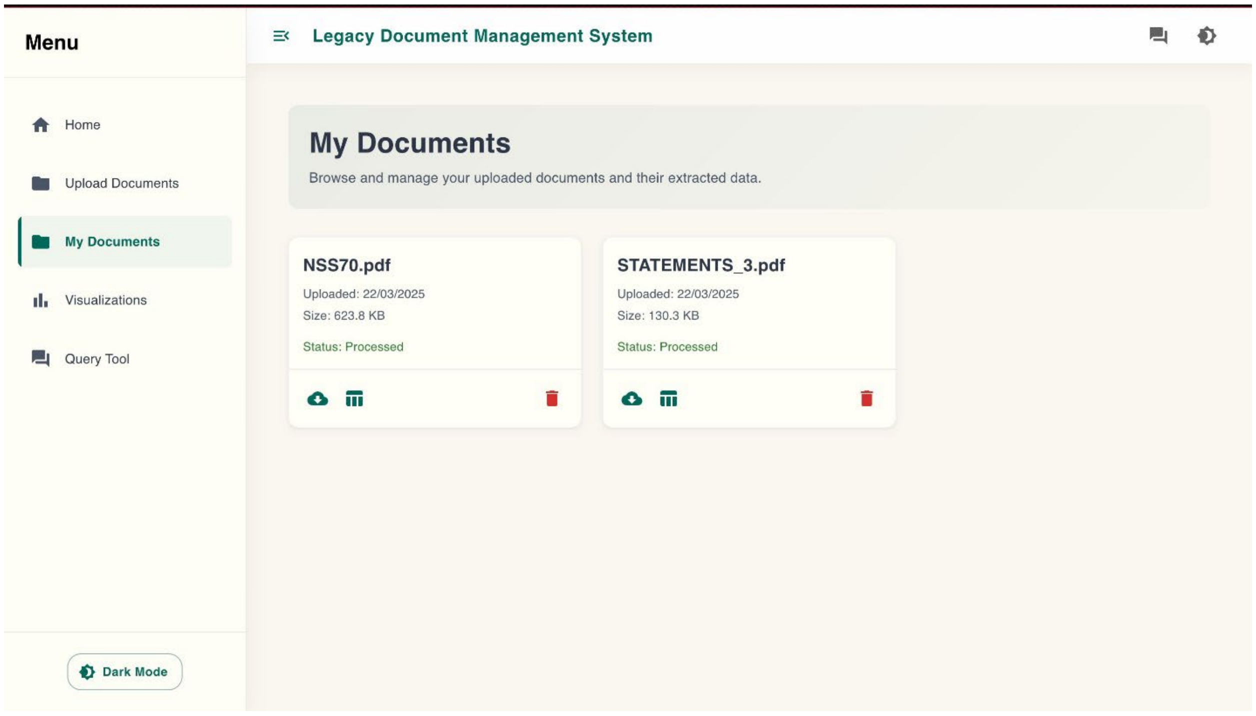


Upload Document

Upload PDF files for automated content extraction and analysis. The system will extract text, tables, and images from your documents.

Appendix 3_NSS70.pdf
0.61 MB

Extracting content... 53%

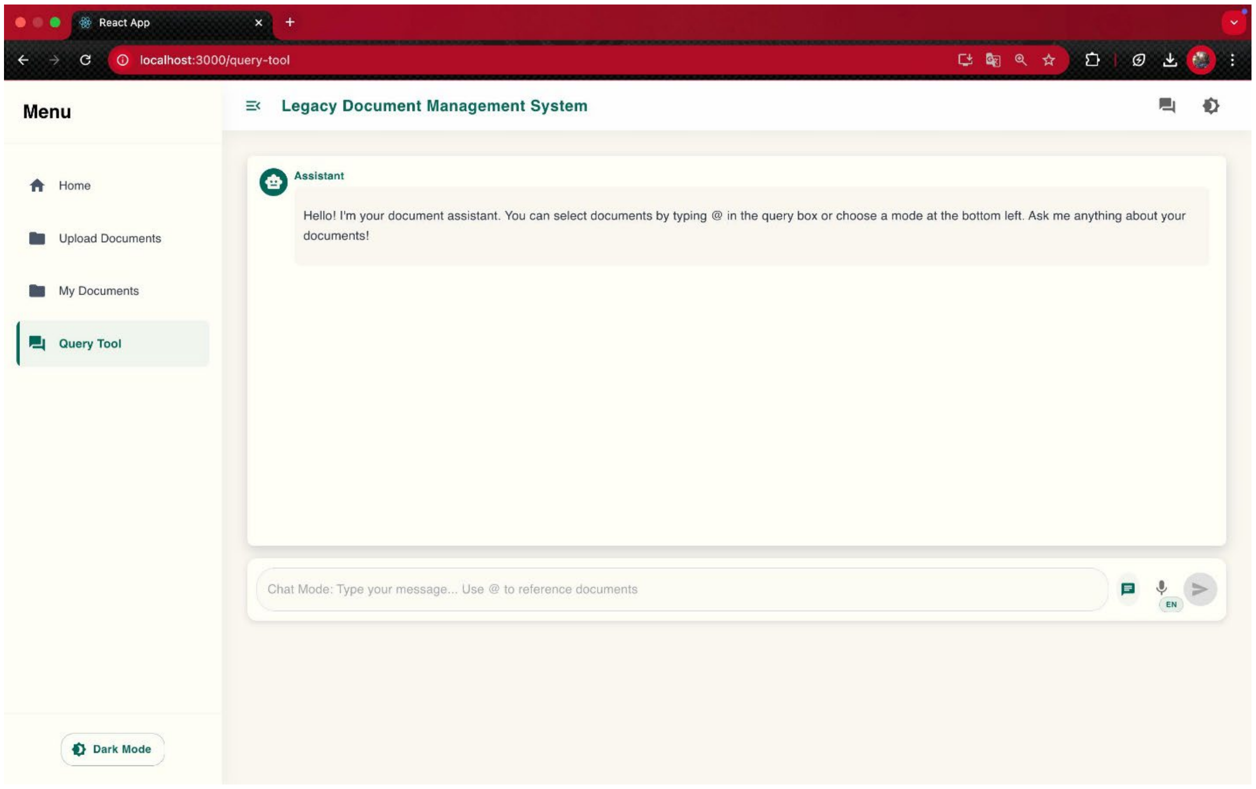


My Documents

Browse and manage your uploaded documents and their extracted data.

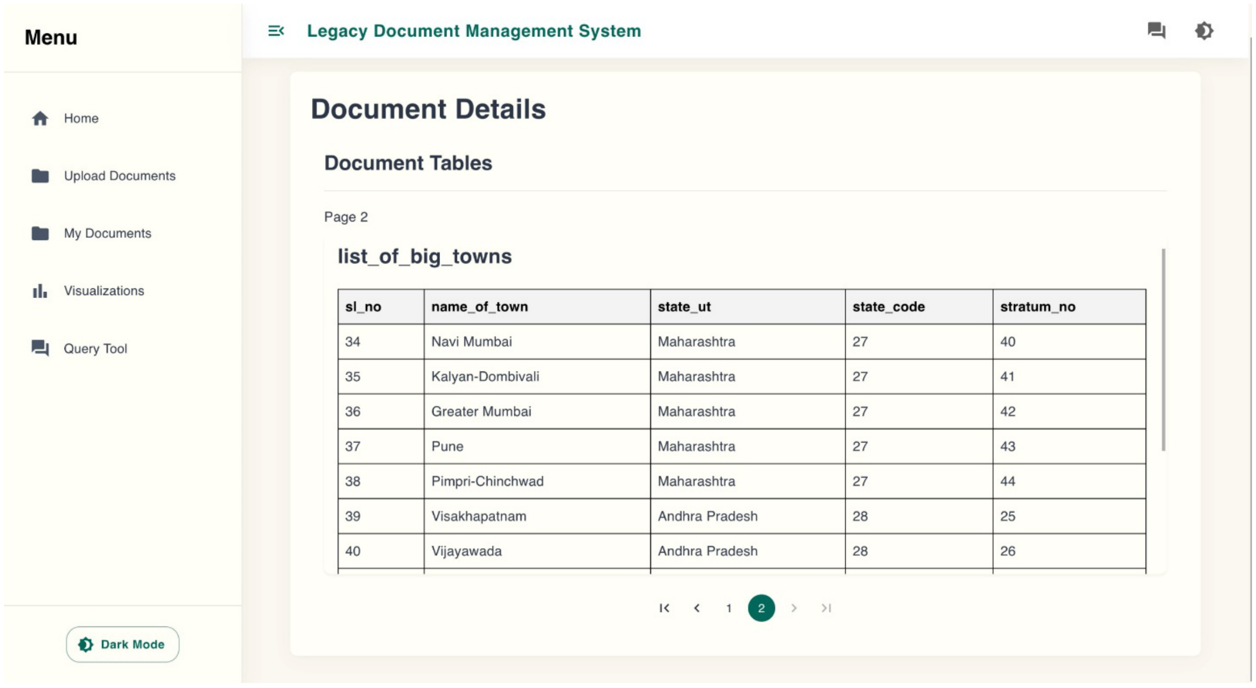
NSS70.pdf
Uploaded: 22/03/2025
Size: 623.8 KB
Status: Processed

STATEMENTS_3.pdf
Uploaded: 22/03/2025
Size: 130.3 KB
Status: Processed



Hello! I'm your document assistant. You can select documents by typing @ in the query box or choose a mode at the bottom left. Ask me anything about your documents!

Chat Mode: Type your message... Use @ to reference documents



Document Details

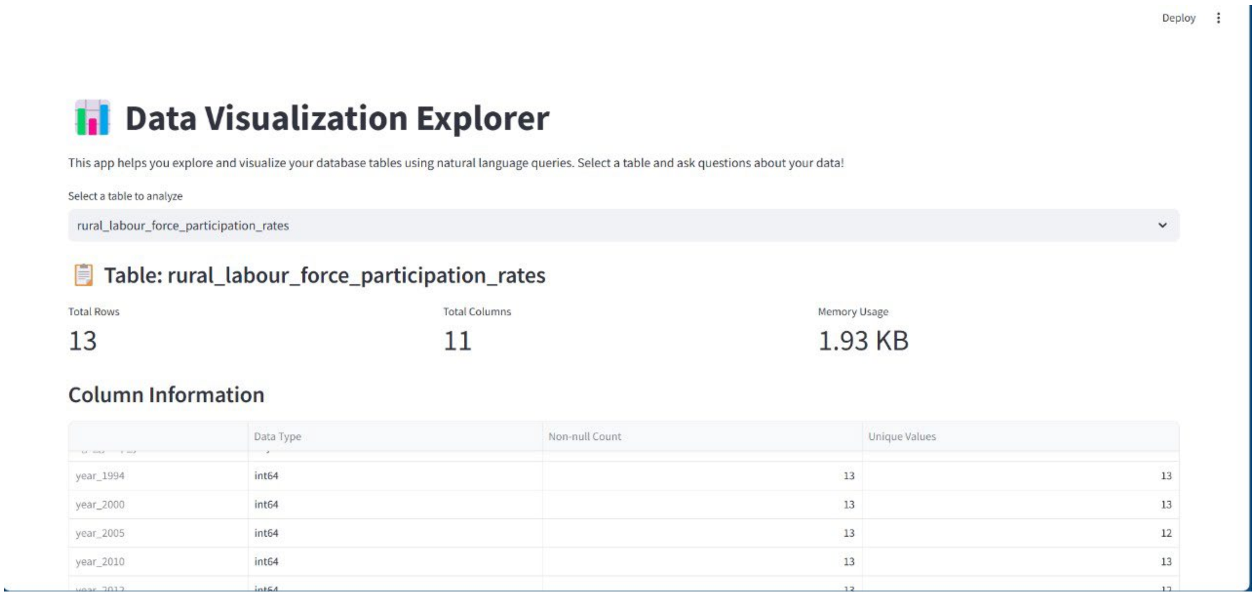
Document Tables

Page 2

list_of_big_towns

sl_no	name_of_town	state_ut	state_code	stratum_no
34	Navi Mumbai	Maharashtra	27	40
35	Kalyan-Dombivali	Maharashtra	27	41
36	Greater Mumbai	Maharashtra	27	42
37	Pune	Maharashtra	27	43
38	Pimpri-Chinchwad	Maharashtra	27	44
39	Visakhapatnam	Andhra Pradesh	28	25
40	Vijayawada	Andhra Pradesh	28	26

Visualizations



Deploy

📊 Data Visualization Explorer

This app helps you explore and visualize your database tables using natural language queries. Select a table and ask questions about your data!

Select a table to analyze

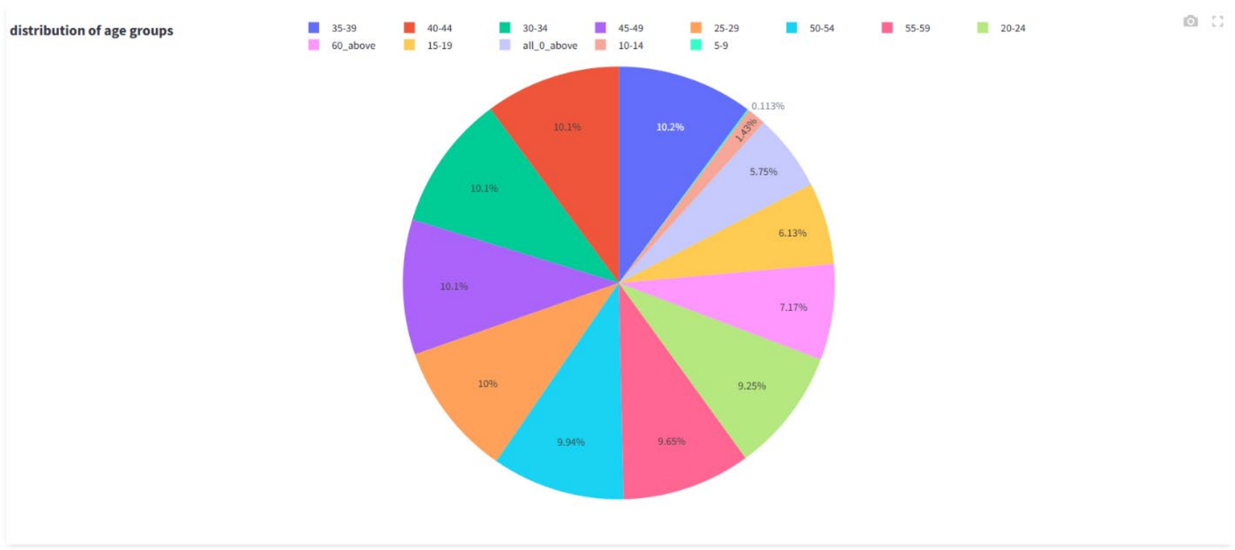
rural_labour_force_participation_rates

📋 Table: rural_labour_force_participation_rates

Total Rows	Total Columns	Memory Usage
13	11	1.93 KB

Column Information

	Data Type	Non-null Count	Unique Values
year_1994	int64	13	13
year_2000	int64	13	13
year_2005	int64	13	12
year_2010	int64	13	13



📈 Insights

Plot Rationale

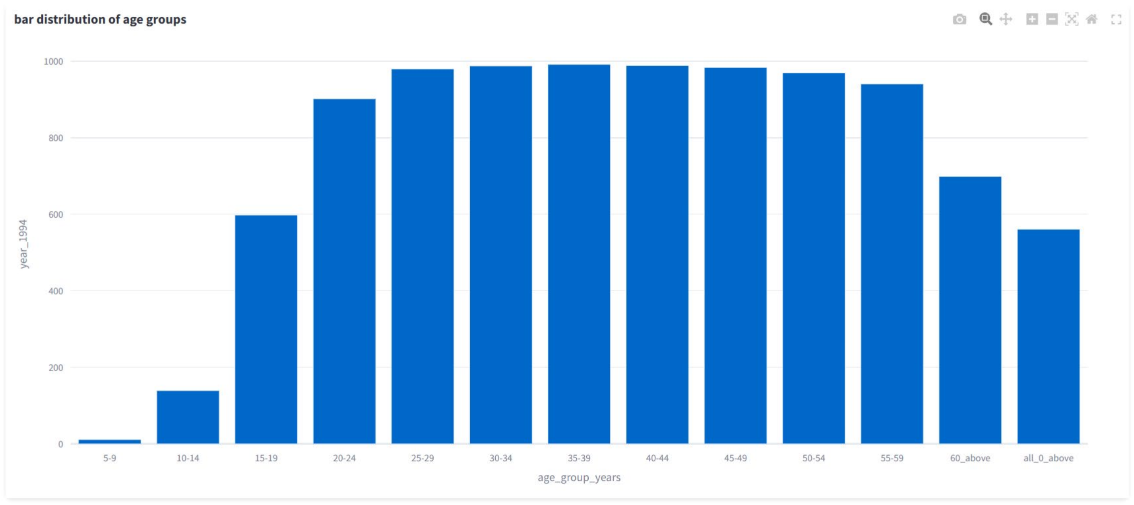
A pie chart is suitable for visualizing the proportion of `year_1994` values across different `age_group_years`, because it effectively displays how a total is divided into parts. Given the objective of understanding the distribution of values in `year_1994` for each `age_group_years`, this plot will show the relative contribution of each age group. The pie chart will reveal which age groups had the largest and smallest values within the year 1994.

Key Pattern

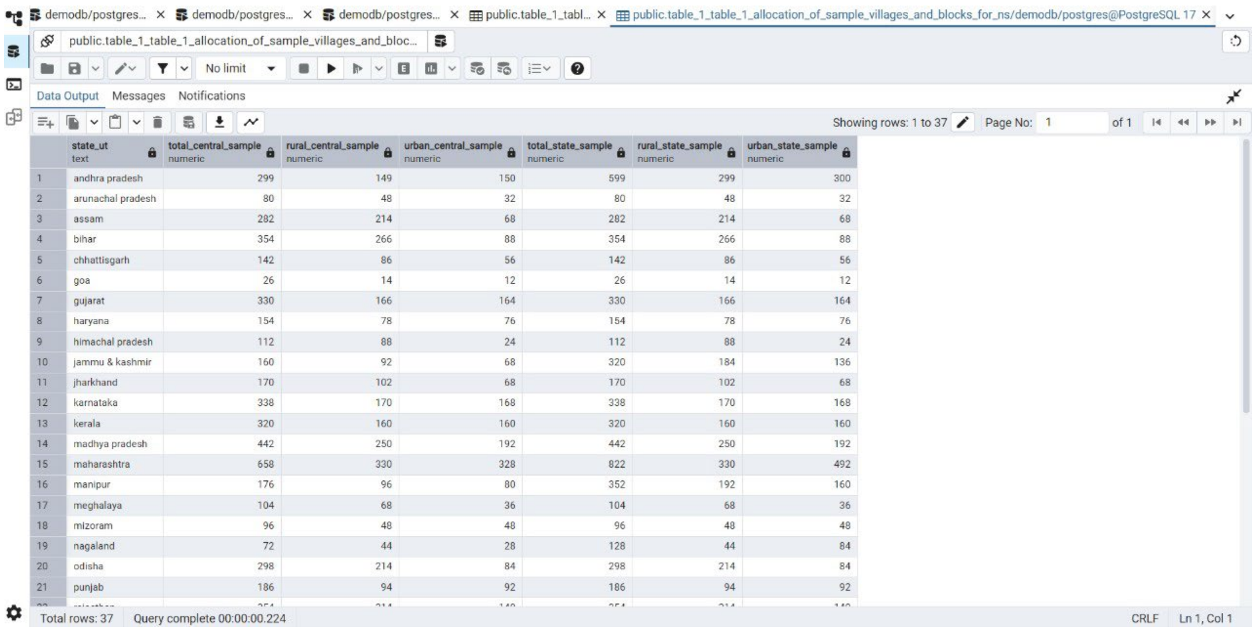
The pie chart will display the relative sizes of `year_1994` values for each `age_group_years`.

Key Finding

The visualization will show the distribution of `year_1994` values across age groups, and the largest slice will represent the age group with the highest value of `year_1994`.



PgAdmin (Postgres Database)


demodb/postgres...
demodb/postgres...
demodb/postgres...
public.table_1_tabl...
public.table_1_table_1_allocation_of_sample_villages_and_blocks_for_ns/demodb/postgres@PostgreSQL 17
public.table_1_table_1_allocation_of_sample_villages_and_bloc...
No limit
Data Output
Messages
Notifications
Showing rows: 1 to 37
Page No: 1
of 1

	state_ut text	total_central_sample numeric	rural_central_sample numeric	urban_central_sample numeric	total_state_sample numeric	rural_state_sample numeric	urban_state_sample numeric
1	andhra pradesh	299	149	150	599	299	300
2	arunachal pradesh	80	48	32	80	48	32
3	assam	282	214	68	282	214	68
4	bihar	354	266	88	354	266	88
5	chhattisgarh	142	86	56	142	86	56
6	goa	26	14	12	26	14	12
7	gujarat	330	166	164	330	166	164
8	haryana	154	78	76	154	78	76
9	himachal pradesh	112	88	24	112	88	24
10	jammu & kashmir	160	92	68	320	184	136
11	jharkhand	170	102	68	170	102	68
12	karnataka	338	170	168	338	170	168
13	kerala	320	160	160	320	160	160
14	madhya pradesh	442	250	192	442	250	192
15	maharashtra	658	330	328	822	330	492
16	manipur	176	96	80	352	192	160
17	meghalaya	104	68	36	104	68	36
18	mizoram	96	48	48	96	48	48
19	nagaland	72	44	28	128	44	84
20	odisha	298	214	84	298	214	84
21	punjab	186	94	92	186	94	92

Total rows: 37 Query complete 00:00:00.224

CRLF Ln 1, Col 1

10. List of Test Cases with Execution Time

Test Case	Description	Execution Time
TC-001	Single page PDF upload and processing	15 seconds
TC-002	Multi-page document (50 pages) processing	40 seconds
TC-004	Natural language query processing	4-5 seconds
TC-005	SQL query generation and execution	1-2 seconds
TC-006	Document search with filters	1.5 seconds
TC-007	Very large document (200+ pages) processing	90 seconds
TC-009	OCR processing of scanned document (non-selectable text)	0.5 secs per page