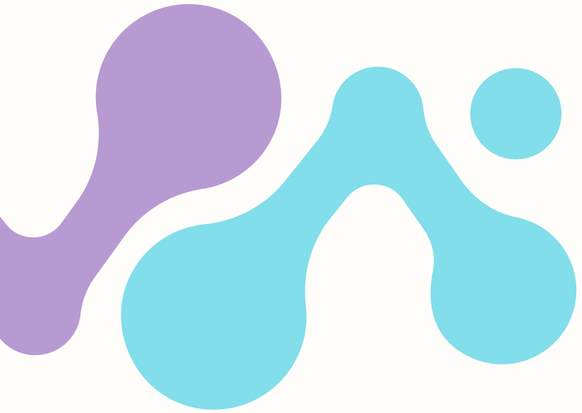
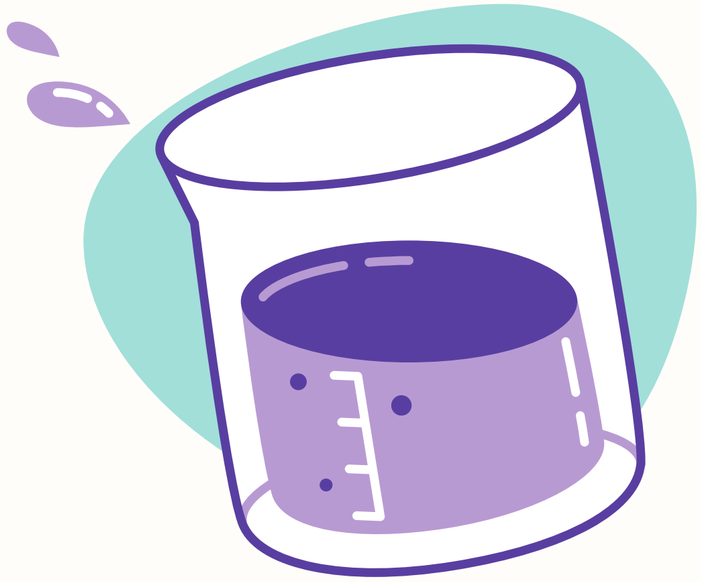
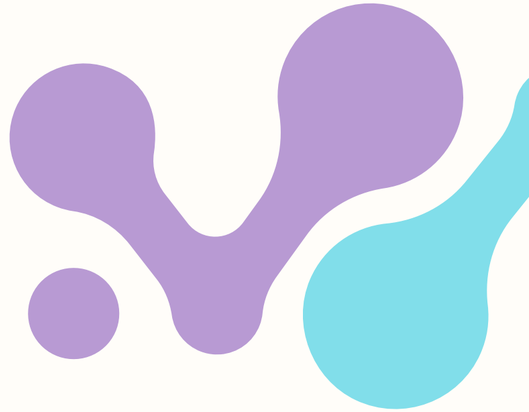


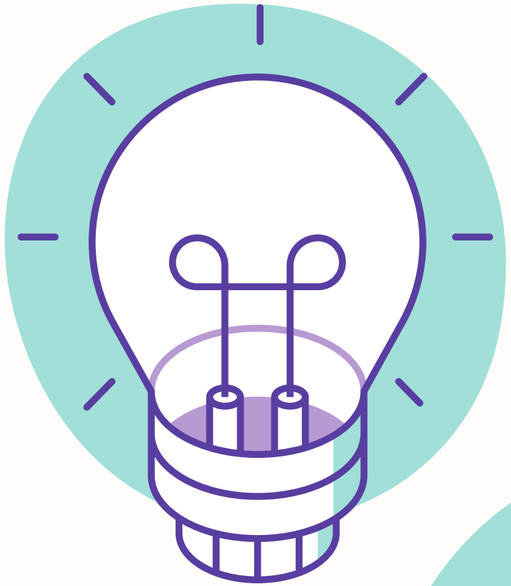
NNDSS DATA PIPELINE PROJECT

Group 3





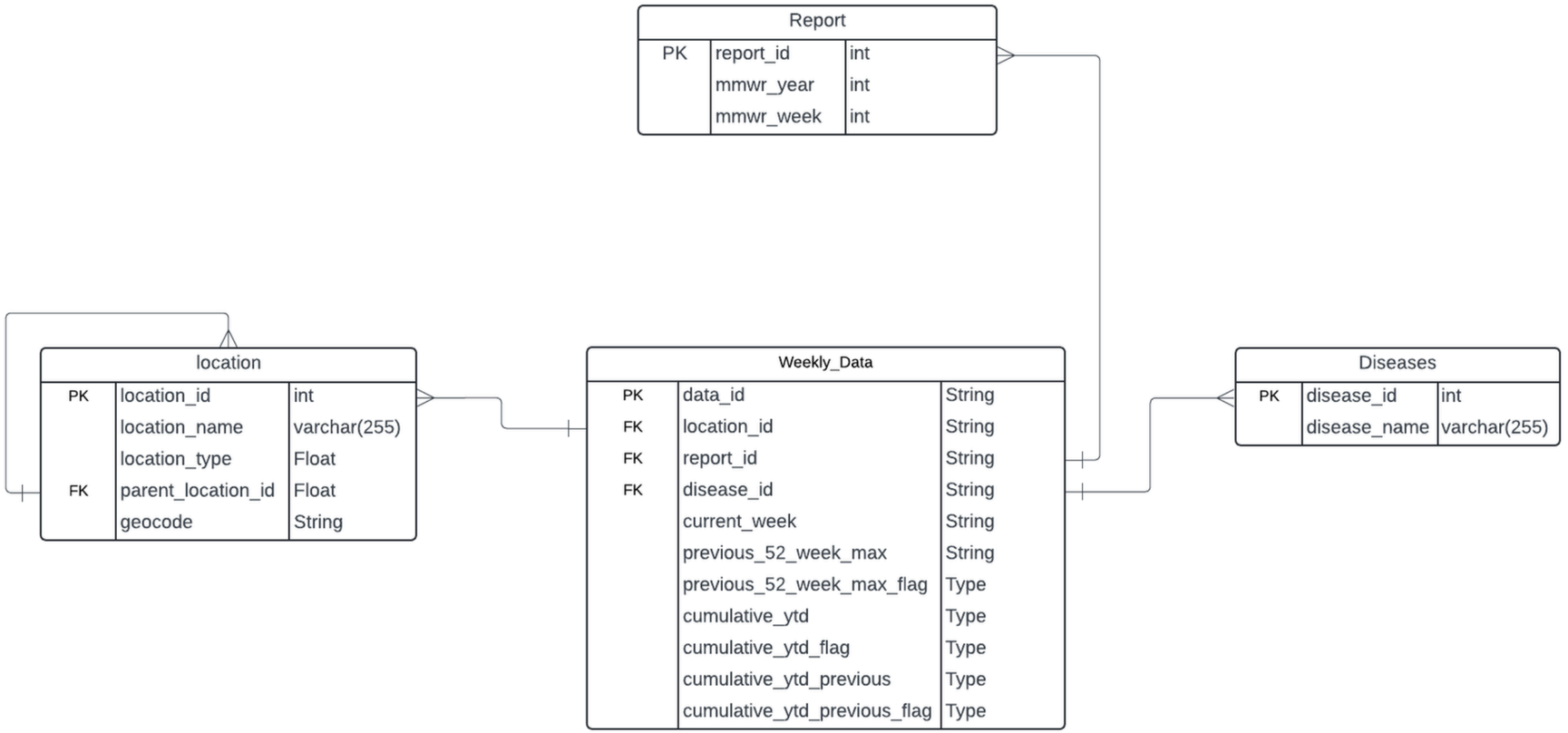
What's NNDSS?



- The National Notifiable Diseases Surveillance System (NNDSS) is a public health surveillance system run by the CDC.
- It tracks notifiable diseases that healthcare providers must report to public health authorities.
- NNDSS data helps detect disease outbreaks, monitor trends, and guide public health responses.
- It plays a crucial role in controlling epidemics and preparing for emerging health threats.

About the Data

Source:
CDC NNDSS API (updates weekly)
Size:
Rows: 1.12 million,
Columns: 16
Data Type:
Structured data in JSON format



The dataset provides a detailed, time-series view of notifiable disease trends across the United States, making it an essential resource for monitoring public health and identifying potential outbreaks.

Objective

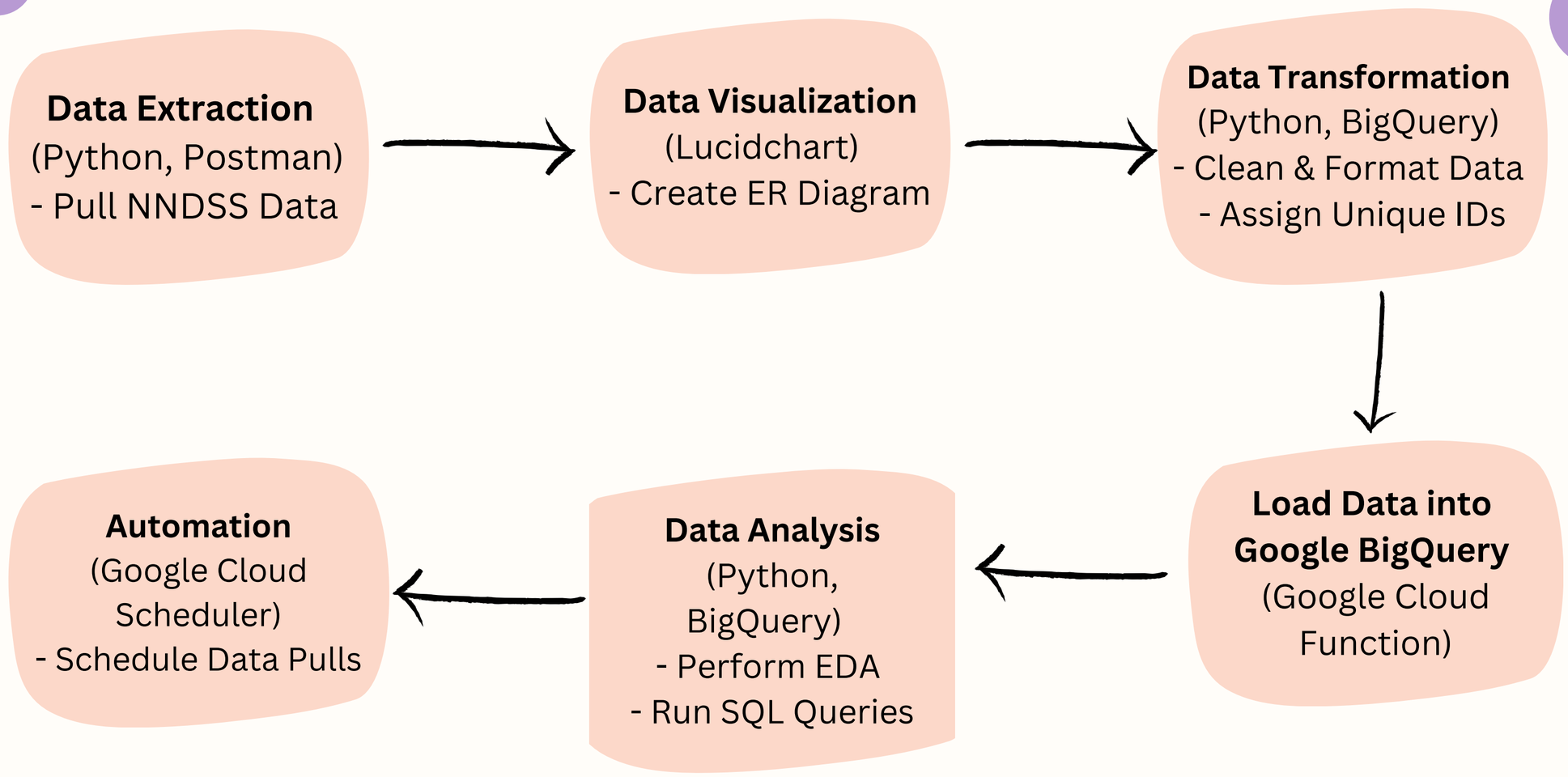
Data Extraction: Efficiently pull large datasets from the CDC using APIs.

Data Transformation: Clean and format the data before loading it into BigQuery.

Data Analysis: Perform exploratory data analysis (EDA) using SQL queries.

Automation: Set up an automated data pipeline using cloud technologies.

Pipeline Overview



API Testing and Data Loading

1. To access the NNDSS data, we configured the API by obtaining the necessary authentication tokens and API keys from the CDC website.
2. We utilized Postman as a powerful tool to test and verify our API requests. This step was crucial for ensuring our configurations were correct and that we could successfully retrieve data.
3. One of the challenges we faced was the 1000-row limit imposed by the Socrata platform when querying datasets. To overcome this limitation, we implemented the following strategies:

- Utilized pagination in our API requests, allowing us to specify parameters that retrieve data in batches.
- Conducted tests in Postman to ensure our pagination implementation worked correctly and retrieved the complete dataset efficiently.

Exploratory Data Analysis

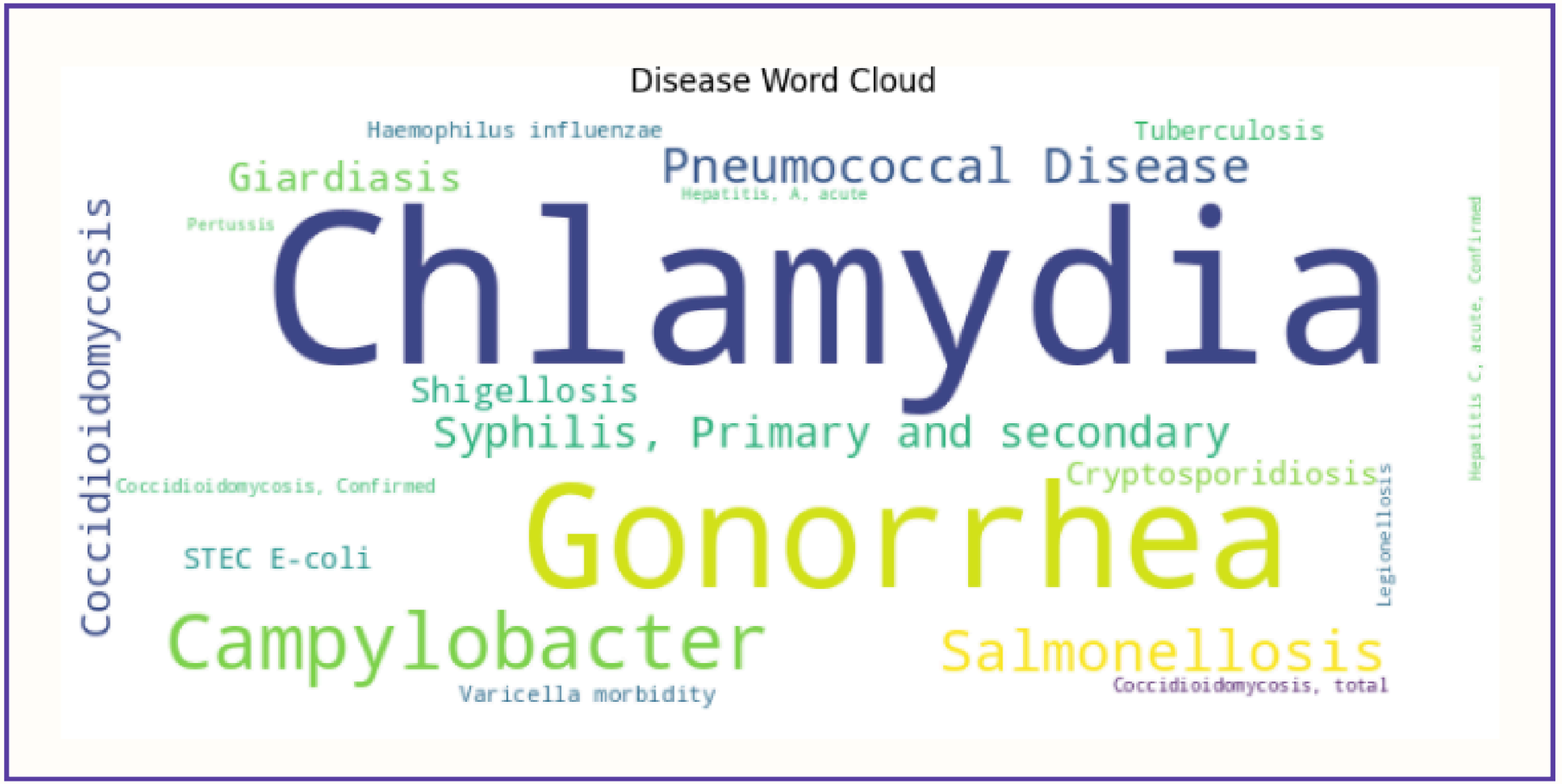
Year	Distinct Reports	Total Cases
2022	3565	12227115
2023	148	1001438
2024	63	607624

Table 1: Disease reportings each year

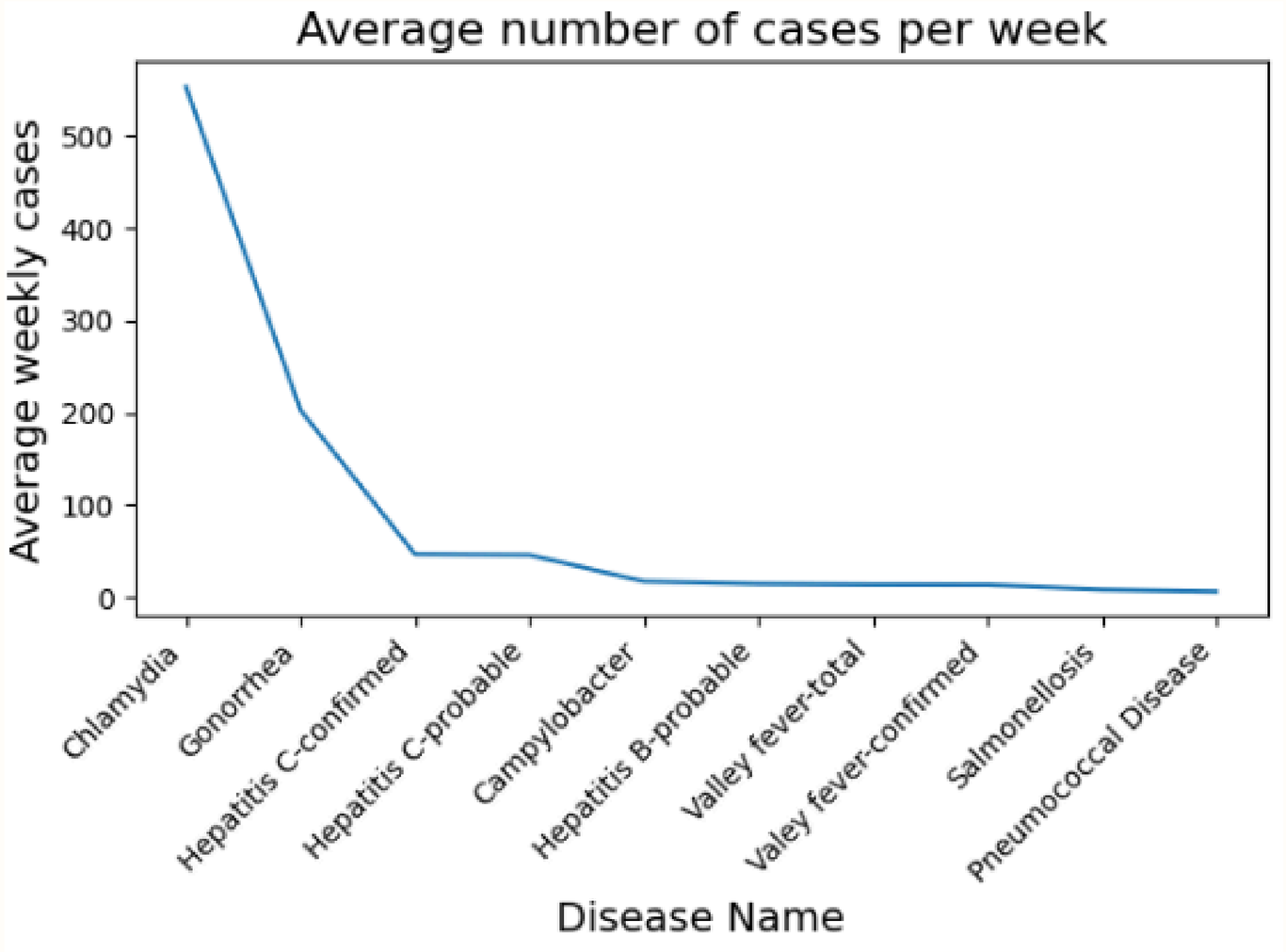
mmwr_year	disease_name	total_cases
2022	Chlamydia trachomatis infection	8326897
2023	Chlamydia trachomatis infection	588675
2024	Chlamydia trachomatis infection	227296

Table 2: Disease with the highest cases

Exploratory Data Analysis

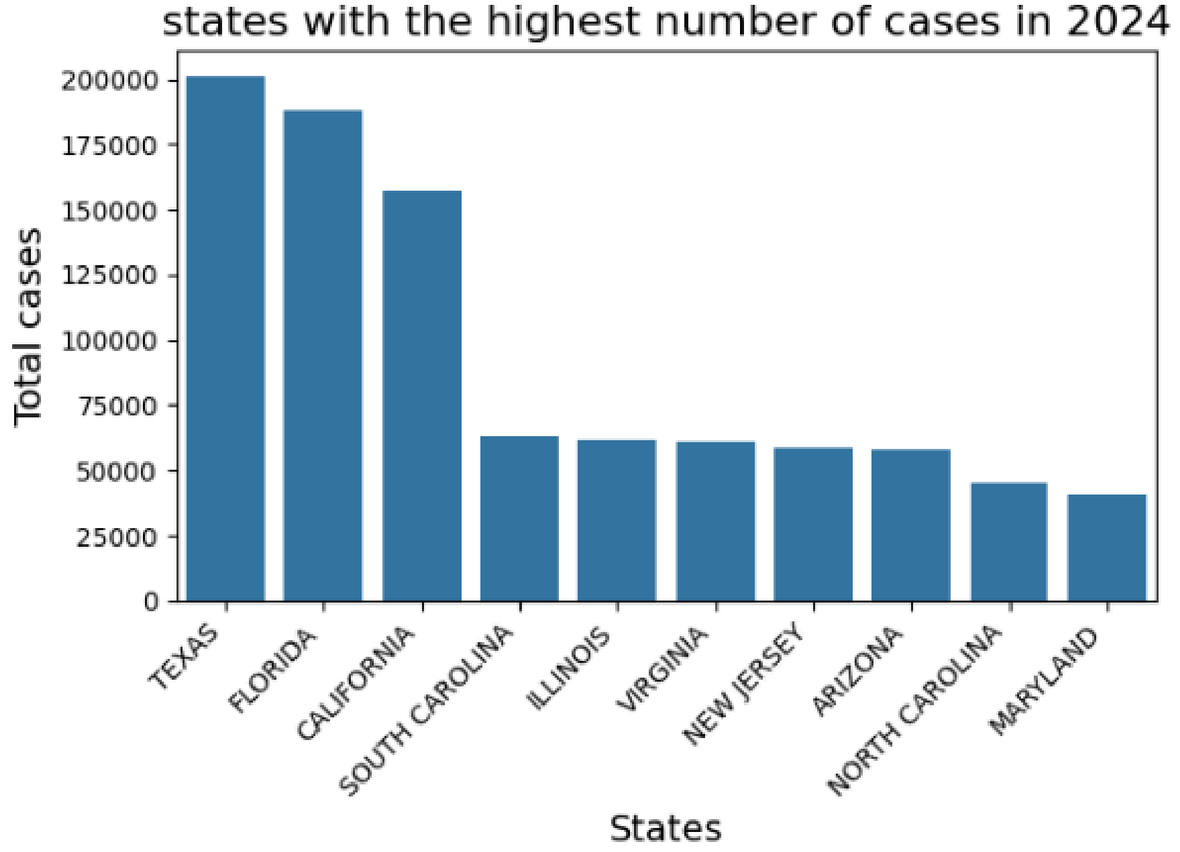


Exploratory Data Analysis



- Chlamydia has an average case of 553 every week.
- In 2024, florida had the highest number of chlamydia cases so far
- However, it decreased by 43.6% in 2024 compared to 2023.
- Gonorrhea increased by 8.8% in 2024 and has an average case of 202

Exploratory Data Analysis



Challenges Faced

Challenge:
Timeout Errors

Solution:
We optimized our API requests by:

- Increasing timeout settings in our requests to allow more time for data retrieval.
- Reducing the size of the data fetched per request by focusing on specific parameters, improving overall efficiency.

Challenge:
Prefect Scheduling Issue - issues with GET/POST requests. Our data retrieval required POST requests, but the Prefect setup was configured for GET requests, leading to failed data pulls.

Solution:
We switched to Google Cloud Scheduler, which allowed us to run our data extraction scripts more flexibly. This change enabled us to easily configure POST requests

Challenge:
Managing the large dataset (1.12M rows) proved challenging in terms of processing time and memory consumption during extraction and loading.

Solution:
We addressed this by:

- Using pagination to handle data in smaller batches, which reduced memory load and improved performance.

Business Motive for Our Project

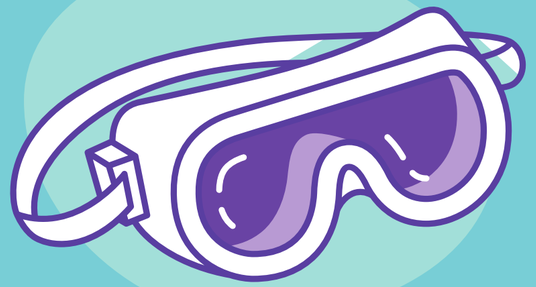
1. Improving Public Health Response
2. Optimizing Healthcare Resources
4. Enhancing Reporting and Transparency
5. Leveraging Machine Learning for Proactive Actions

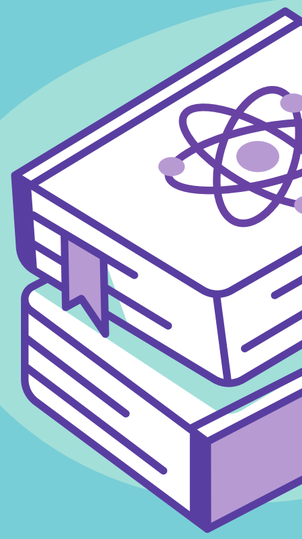
Future Work

Dashboard Development
Create interactive dashboards using tools like Streamlit or Apache Superset to visualize trends and insights from the NNDSS data.

Trend Analysis
Conduct in-depth trend analysis to identify patterns and anomalies in disease reporting.

Machine Learning Applications
Clustering: Apply clustering algorithms to identify groups of similar diseases or regions with comparable outbreak patterns.





THANK YOU