NNDSS
DATA
PIPELINE
PROJECT

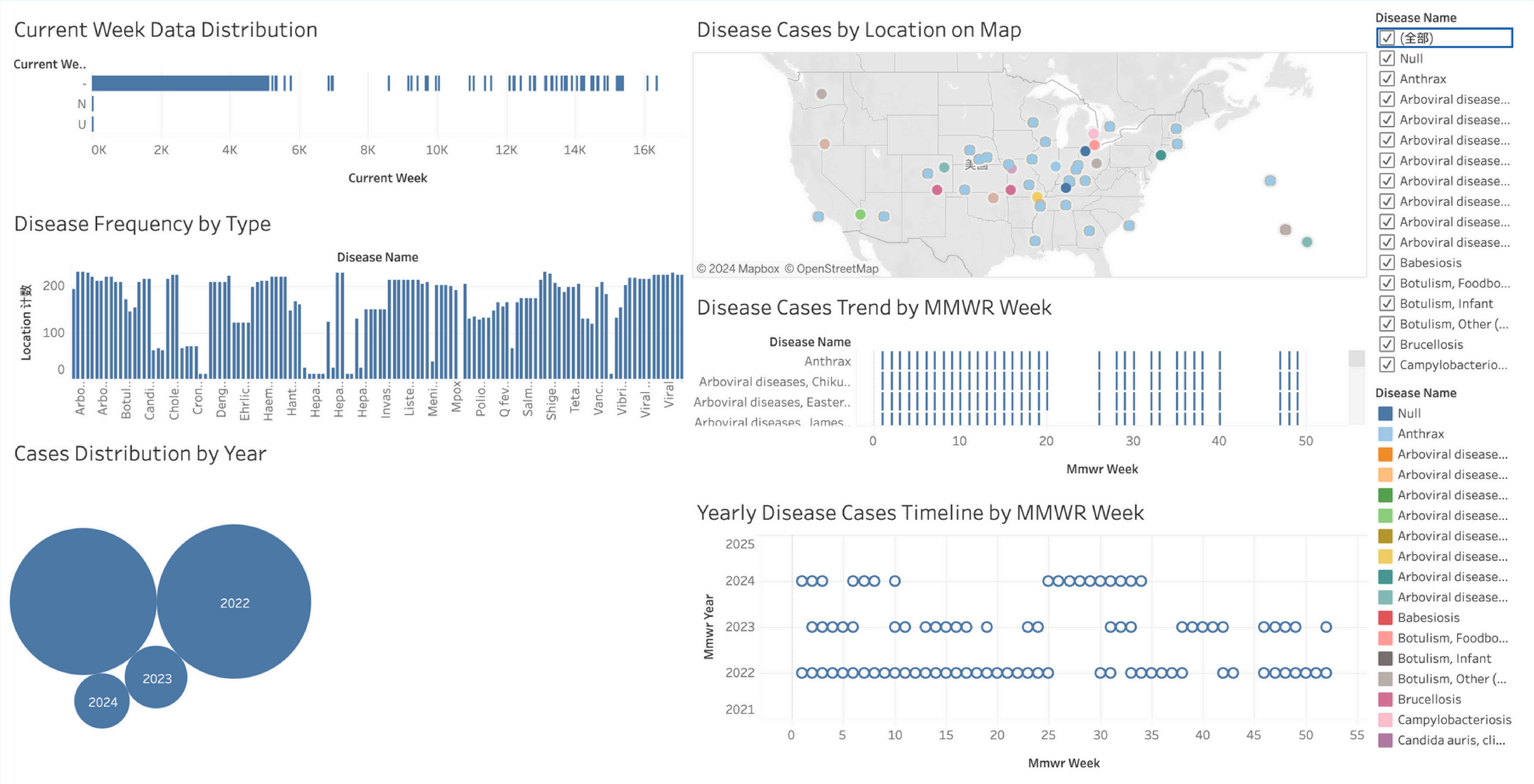
RECAP

- Extracted the data from NNDSS website using API calls and Postman
- Data transformation using BigQuery
- EDA
- Set up the automation using Google cloud Scheduler

DATASET

- **Dataset Source and Purpose:** National Notifiable Diseases Surveillance System (NNDSS) by the CDC, which tracks weekly reported cases of various diseases across different U.S. locations.
- Its primary purpose is to monitor disease trends, detect outbreaks, and inform public health responses.
- **Time Span and Granularity:** The dataset spans from 2022 onwards, and is organized by weekly reports (MMWR weeks). Each record provides detailed information on disease counts per location, allowing for both short-term and seasonal analysis of trends.
- **Key Variables:**
  - Disease Name: Specifies the type of disease (e.g., influenza, measles).
  - Location Information: Includes state-level data, latitude, and longitude coordinates.
  - Case Counts: Current weekly counts, along with cumulative counts and rolling averages, which are useful for trend and anomaly analysis.

DASHBOARD

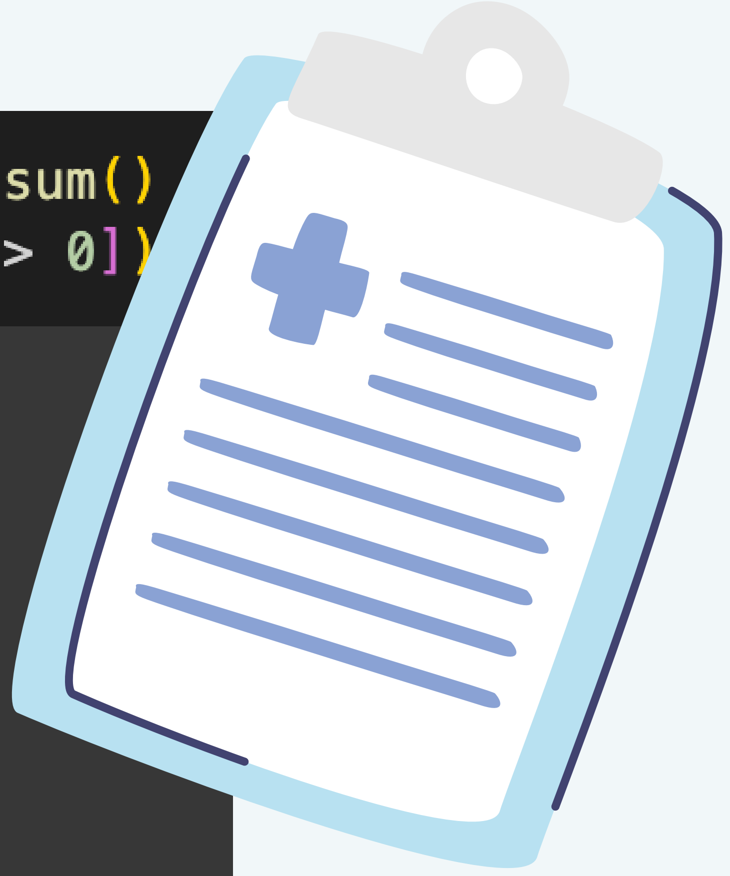

Current Week Data Distribution
Current We..
-
N
U
0K
2K
4K
6K
8K
10K
12K
14K
16K
Current Week
Disease Frequency by Type
Disease Name
Location 计数
200
100
0
Arbo..
Arbo..
Botul..
Candi..
Chole..
Cron..
Deng..
Ehrlic..
Haem..
Hant..
Hepa..
Hepa..
Hepa..
Invas..
Liste..
Meni..
Mpox
Polio..
Q fev..
Salm..
Shige..
Teta..
Vanc..
Vibri..
Viral ..
Viral
Cases Distribution by Year
2022
2023
2024
Disease Cases by Location on Map
美国
© 2024 Mapbox © OpenStreetMap
Disease Cases Trend by MMWR Week
Disease Name
Anthrax
Arboviral diseases, Chiku..
Arboviral diseases, Easter..
Arboviral diseases, James..
0
10
20
30
40
50
Mmwr Week
Yearly Disease Cases Timeline by MMWR Week
Mmwr Year
2025
2024
2023
2022
2021
0
5
10
15
20
25
30
35
40
45
50
55
Mmwr Week
Disease Name
(全部)
Null
Anthrax
Arboviral disease...
Arboviral disease...
Arboviral disease...
Arboviral disease...
Arboviral disease...
Arboviral disease...
Arboviral disease...
Arboviral disease...
Babesiosis
Botulism, Foodbo...
Botulism, Infant
Botulism, Other (...
Brucellosis
Campylobacterio...
Disease Name
Null
Anthrax
Arboviral disease...
Arboviral disease...
Arboviral disease...
Arboviral disease...
Arboviral disease...
Arboviral disease...
Arboviral disease...
Arboviral disease...
Babesiosis
Botulism, Foodbo...
Botulism, Infant
Botulism, Other (...
Brucellosis
Campylobacteriosis
Candida auris, cli...

Detecting Anomalies

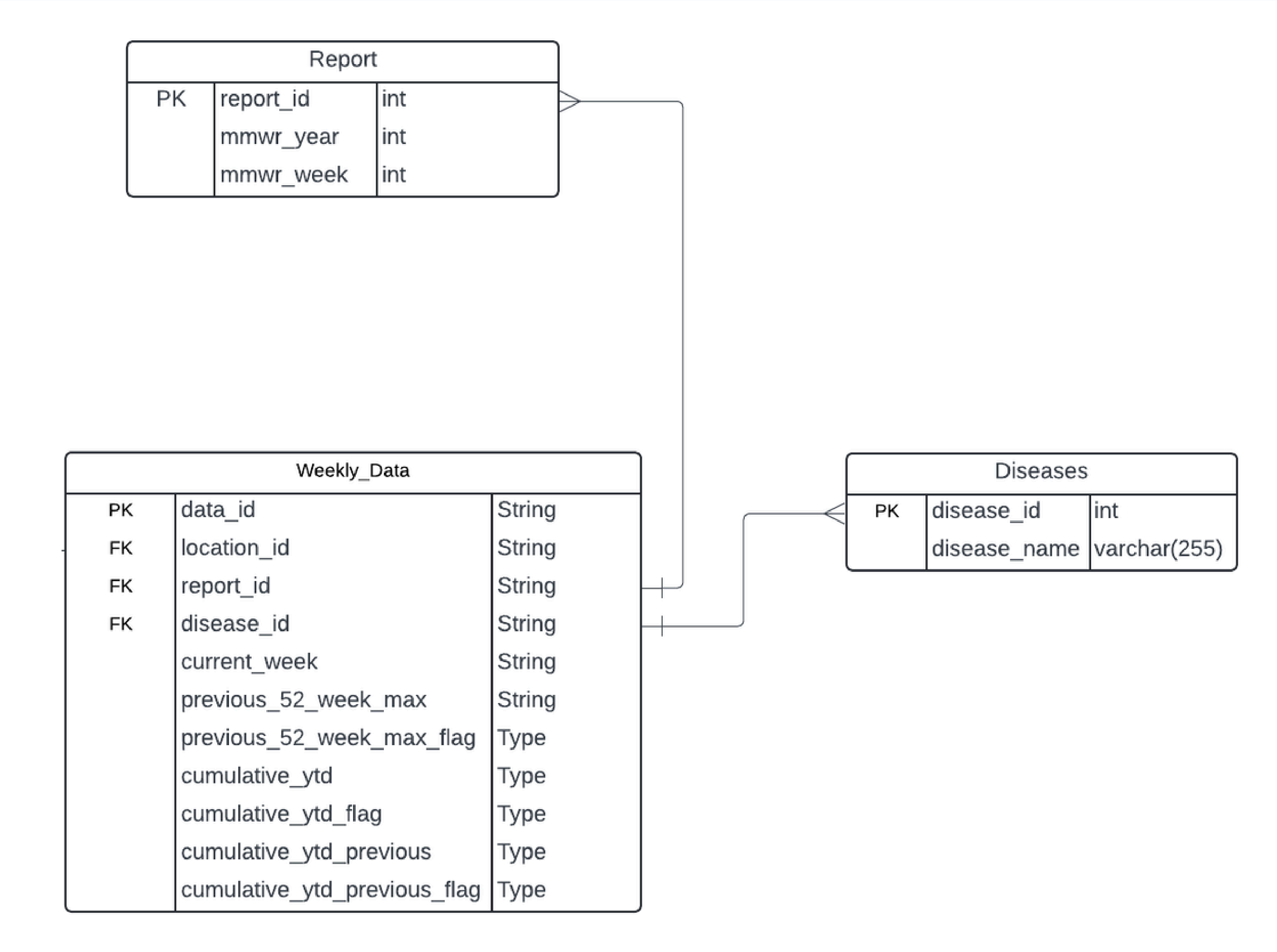
- Data Cleaning
- Trend Analysis
- Preprocessing
- Anomaly Detection

```
missing_values = merged_df.isnull().sum()
print(missing_values[missing_values > 0])
```

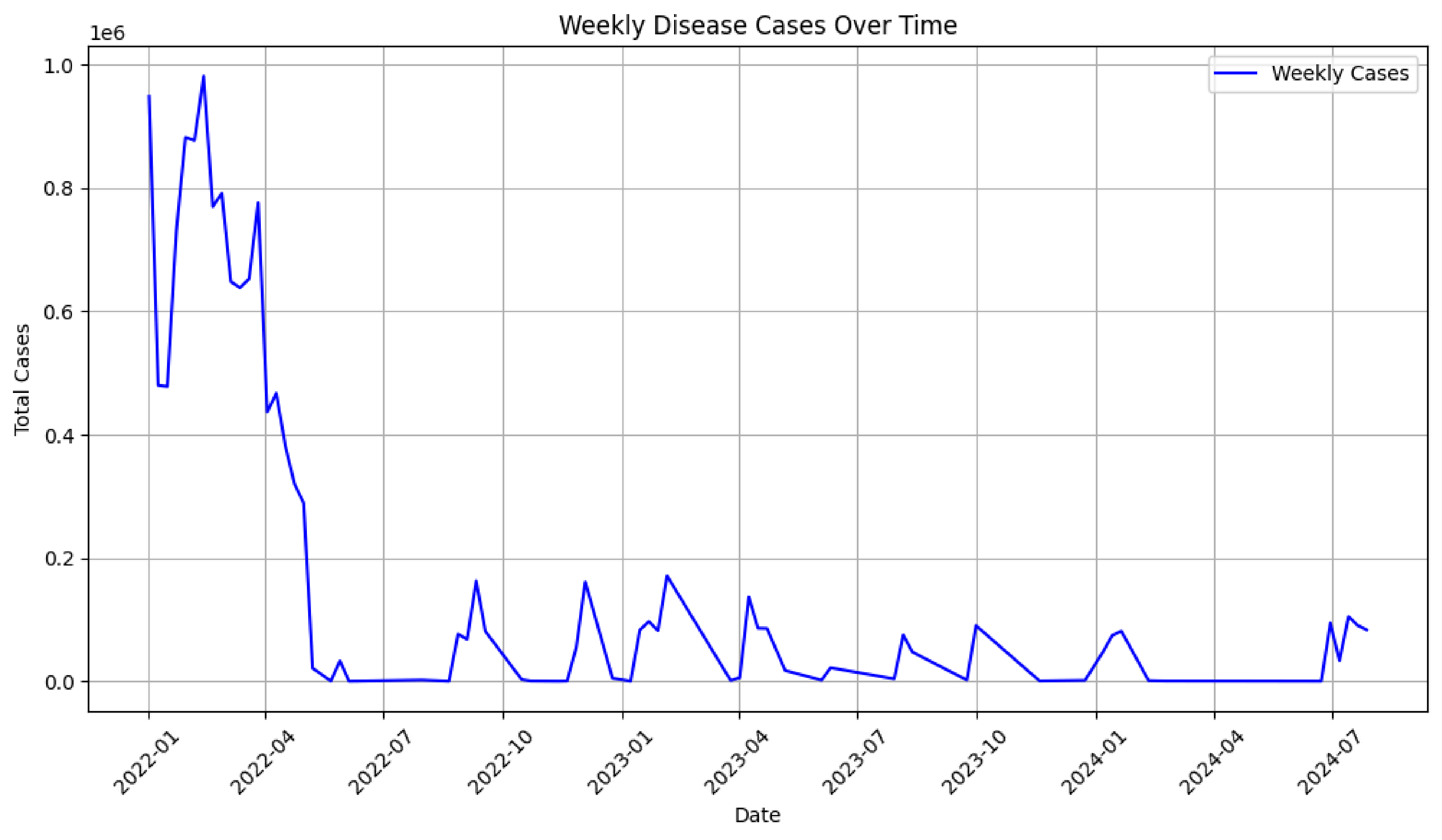
```
disease_name        66036
location_name     2040783
states_x          2040783
location2         2040783
longitude         2055217
latitude          2055217
mmwr_year          291068
mmwr_week          291068
states_y           291068
label              291068
dtype: int64
```

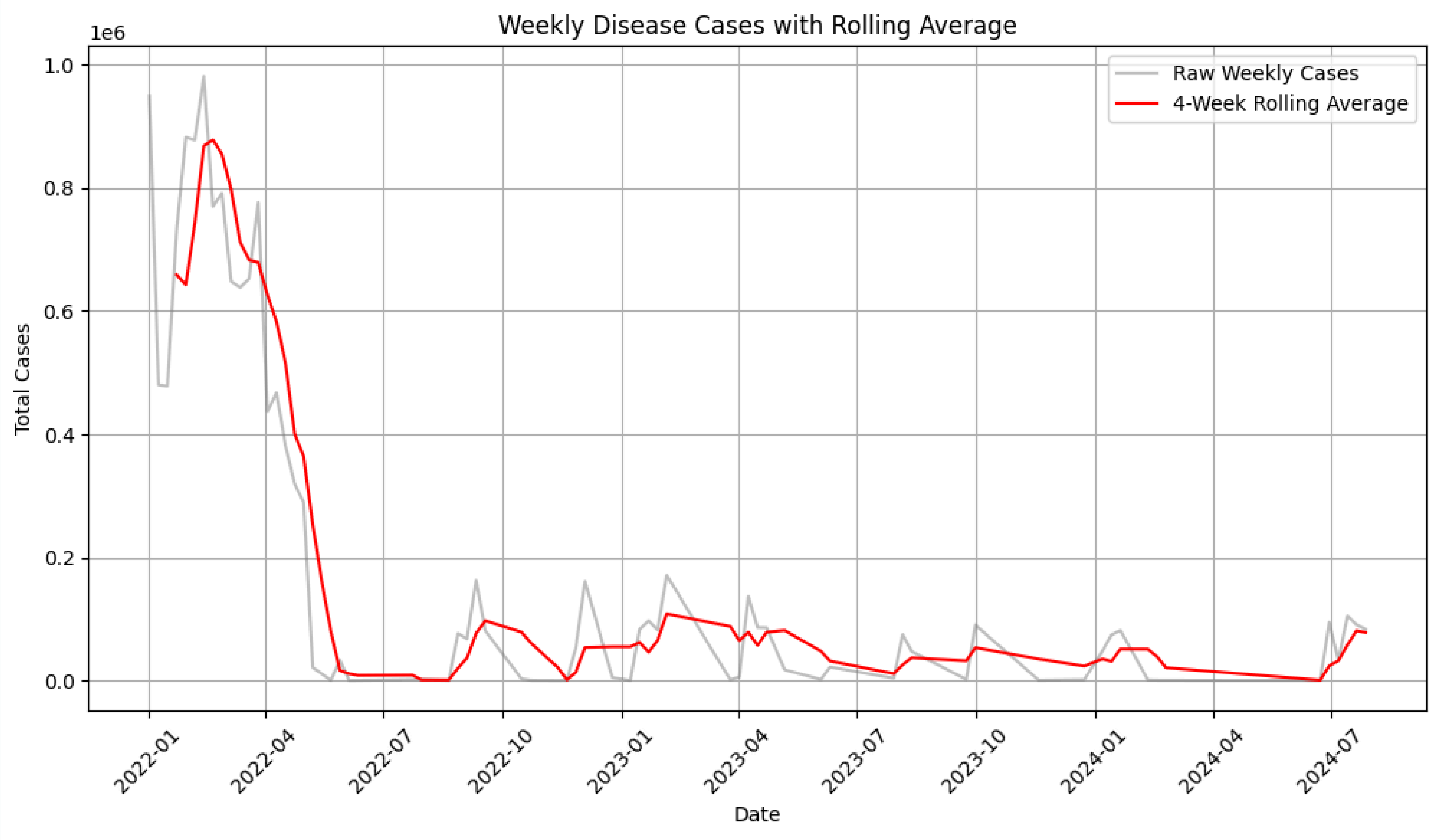


TABLES USED

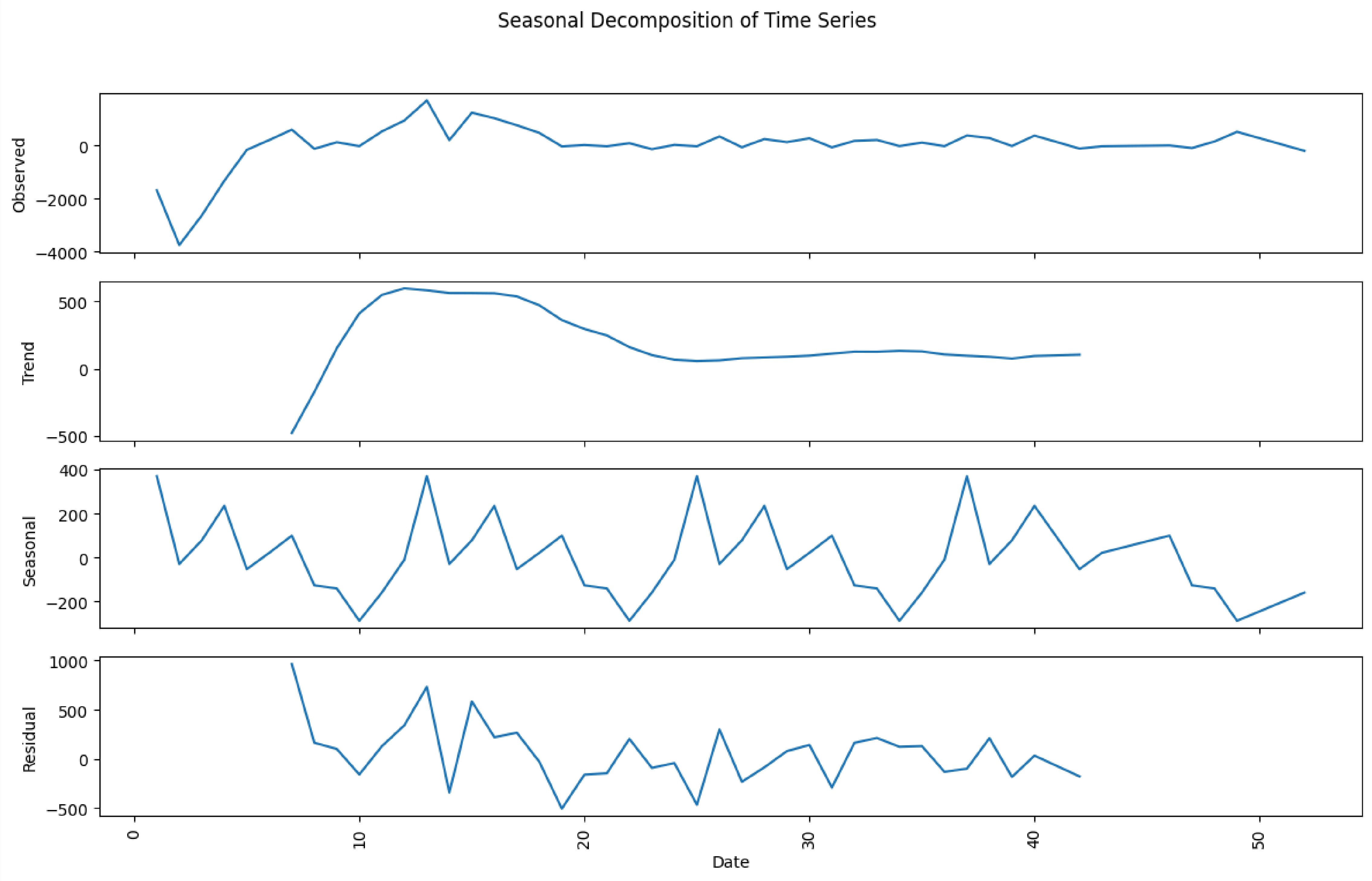

Report
PK report_id int
mmwr_year int
mmwr_week int
Weekly_Data
PK data_id String
FK location_id String
FK report_id String
FK disease_id String
current_week String
previous_52_week_max String
previous_52_week_max_flag Type
cumulative_ytd Type
cumulative_ytd_flag Type
cumulative_ytd_previous Type
cumulative_ytd_previous_flag Type
Diseases
PK disease_id int
disease_name varchar(255)

Trend Analysis





Trend Analysis (Contd.)



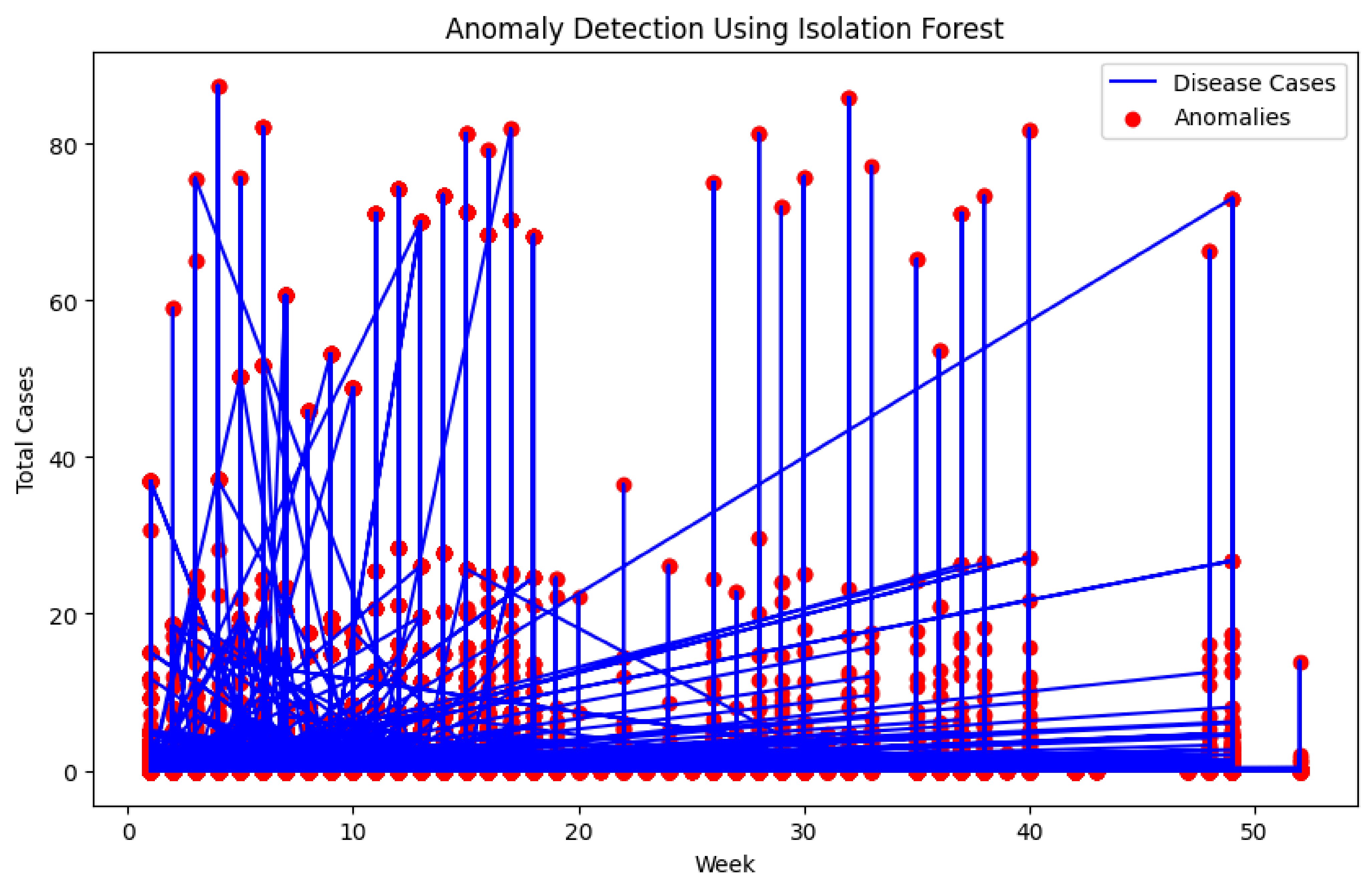
- **Observed:** Shows raw disease case data over time, highlighting overall fluctuations.
- **Trend**: Illustrates the long-term direction of cases, helping to identify sustained increases or decreases.
- **Seasonal**: Captures regular patterns, revealing potential cycles like weekly or yearly spikes in cases.
- **Residual**: Displays remaining variation, potentially indicating rare events, anomalies, or data irregularities.

Anomaly Detection

Methods used:

1. Z-Score
2. Isolation Forest
3. DBSCAN (Did not work)

- The blue line represents the overall weekly trend
- Red dots indicate anomalies—data points flagged as outliers by the model.
- With a contamination setting of 5%, the algorithm expects about 5% of the points to be anomalies, identifying weeks with case counts that significantly differ from the overall pattern.



Model Evaluation

```
Confusion Matrix:
[[1821555   14124]
 [      0    4264]]


Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.99      1.00   1835679
           1       0.23      1.00      0.38      4264

    accuracy                           0.99   1839943
   macro avg       0.62      1.00      0.69   1839943
weighted avg       1.00      0.99      0.99   1839943
```

- The model performs well in identifying normal weeks, with most non-anomalous weeks correctly classified (True Negatives: 1,821,555)
- A small number of normal weeks misclassified as anomalies (False Positives: 14,124).
- Successfully captures all true anomalies (True Positives: 4,264), achieving perfect recall, meaning no actual anomalies were missed.
- The precision for anomalies (0.23) indicates that, while some flagged cases are normal, most are meaningful deviations. This balance shows the model’s ability to capture unusual patterns while keeping false positives low, making it effective for monitoring potential disease outbreaks.

CHALLENGES

- Using Apache Superset for Visualisations
- Dealing with Date column during preprocessing
- DBSCAN - cannot work with large dataset
- Error during cloud function setup for ML deployment

NEXT STEPS

- Correctly deploying our ML using cloud functions
- Attempt neural network for more advanced modeling and prediction

THANK YOU