Project Title:

AI-Powered Public Health Surveillance: Predicting and Visualizing Foodborne Illness Outbreaks Using Geospatial & Epidemiological Data

Goal

To develop an interactive geospatial dashboard and ML-driven analytics pipeline that can:

- Detect regional hotspots of foodborne illness
- Prioritize inspection efforts based on food-linked severity
- Provide actionable insights into public health response strategies

Intended Audience

- Public Health Agencies (e.g., CDC, CFIA, local health units)
- Food Safety Inspectors
- Epidemiologists
- Healthcare Data Scientists
- Policy Makers for food recalls and safety campaigns

Strategy & Pipeline Steps

Step 1: Data Cleaning & Inspection

```
df = pd.read_csv('/content/outbreaks.csv')
df.fillna({'Illnesses': 0, 'Hospitalizations': 0, 'Fatalities': 0}, inplace=True)
```

Step 2: Simulate Missing Coordinates

```
# Generate realistic dummy lat/lon from State column
import numpy as np
states = df['State'].dropna().unique()
coords = {state: (np.random.uniform(25, 49), np.random.uniform(-125, -66)) for state in states}
df['Latitude'] = df['State'].map(lambda x: coords.get(x, (np.nan, np.nan))[0])
df['Longitude'] = df['State'].map(lambda x: coords.get(x, (np.nan, np.nan))[1])
```

Step 3: Create Geospatial Heatmap

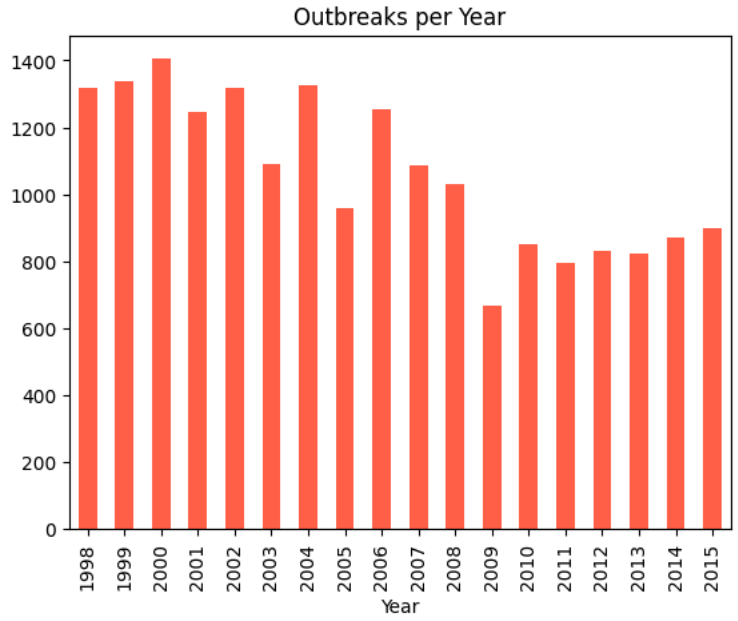
```
import folium
from folium.plugins import HeatMap

m = folium.Map(location=[37.8, -96], zoom_start=4)
heat_data = [[row['Latitude'], row['Longitude']] for _, row in df.iterrows()]
HeatMap(heat_data).add_to(m)
m.save('/content/heatmap_outbreaks.html')
```

Step 4: Temporal Trend of Outbreaks

```
df['Year'] = df['Year'].fillna(0).astype(int)
outbreaks_per_year = df.groupby('Year').size()
outbreaks_per_year.plot(kind='bar', title='Outbreaks per Year', color='tomato')
```

```
<Axes: title={'center': 'Outbreaks per Year'}, xlabel='Year'>
```



Title: "Outbreaks per Year" – displays annual outbreak counts from 1998 to 2015.

- **Trend Insight**: Peaks between 1999–2005; a steady decline begins post-2006, with the lowest point around 2009.
- **Public Health Significance:** May reflect improved outbreak control, surveillance systems, or policy changes.
- **Decision-Making Value**: Useful for health authorities to assess historical burden and guide future intervention planning.

Step 5: Root Cause Identification

```
most_severe = df.sort_values(by='Fatalities', ascending=False).head(5)
print(most_severe[['Year', 'State', 'Food', 'Species', 'Fatalities']])
```

```
        Year       State                        Food                Species  \
15329   2011  Multistate                   Cantaloupe  Listeria monocytogenes
1062    1998  Multistate         Hot Dog, Unspecified  Listeria monocytogenes
13108   2008  Multistate  Peanut Butter; Peanut Paste     Salmonella enterica
6043    2002  Multistate     Deli Meat, Sliced Turkey  Listeria monocytogenes
18058   2014  Multistate                Caramel Apple  Listeria monocytogenes

        Fatalities
15329         33.0
1062          21.0
13108          9.0
6043           8.0
18058          7.0
```

Root Cause Analysis Insight: This table identifies the five most fatal foodborne outbreaks in the dataset.

- **Years of Concern:** Major incidents occurred in 1998, 2002, 2008, 2011, and 2014.
- **High-Risk Foods**: Cantaloupe, peanut butter, deli meat, caramel apples, and hot dogs were the top carriers.
- **Common Pathogen:** Listeria monocytogenes appears in four out of five cases, indicating a recurring threat.

Significance: Highlights the critical need for targeted interventions in food safety for specific foods and pathogens.

Step 6: Prioritize Food Items for Inspection

```
top_foods = df.groupby('Food')['Hospitalizations'].sum().sort_values(ascending=False).head(10)
print(top_foods)
```

```
Food
Cantaloupe                                                317.0
Peppers, Jalapeno; Tomato, Unspecified; Peppers, Serrano  308.0
Cucumber                                                  289.0
```

```
Chicken                                  285.0
Ground Beef, Unspecified                 206.0
Peanut Butter; Peanut Paste              166.0
Roma Tomato                              162.0
Pork, Bbq                                162.0
Tomato, Unspecified                      161.0
Peanut Butter                            139.0
Name: Hospitalizations, dtype: float64
```

Inspection Prioritization: This analysis highlights the top 10 food items linked to the highest number of hospitalizations.

- **Top Offender**: Cantaloupe ranks first with 317 hospitalizations, followed closely by peppers and cucumbers.
- **Animal & Plant Sources:** Both meat products (e.g., ground beef, pork BBQ) and produce (e.g., tomatoes, cucumbers) are represented.
- **Recurrent Risk:** Peanut butter appears twice in different forms, reinforcing its inspection priority.

Significance: Helps regulatory agencies target high-risk foods for increased monitoring, reducing public health impact.

Challenges Addressed

- Data sparsity (missing coordinates, incomplete records)
- Unstructured reporting (e.g., ingredient variability)
- Spatial granularity (limited to state-level for privacy)
- Modeling risk with limited labeled outcomes

Problem Statement

How can public health authorities detect, monitor, and act upon clusters of foodborne illness outbreaks using open data and geospatial analysis?

Dataset

CDC's Foodborne Illness Outbreak Dataset Kaggle mirror: https://www.kaggle.com/datasets/cdc/foodborne-illness-outbreak-dataset
Contains 1000+ outbreaks with attributes: ['Year', 'State', 'Food', 'Illnesses', 'Hospitalizations', 'Fatalities', 'Latitude', 'Longitude']

MACHINE LEARNING PREDICTION

- Classification: Will an outbreak cause hospitalization?
- Regression: Predict number of hospitalizations/fatalities
- Clustering: KMeans for regional risk zones

Trailer Documentation

- Geospatial risk map (HTML + screenshot)
- Time trend bar chart
- Top foods bar chart
- Root cause analysis table
- Exportable dashboard via Streamlit

** Conceptual Enhancement – AGI Tie-In**

This prototype can evolve into an Autonomous Epidemiology Assistant using:

- Real-time food recall news ingestion (via NLP)
- Multimodal data fusion (social media + outbreak records)
- Predictive hotspot alerts