ADITYA SINGHAL GDG_ML_01
23BHI10065

Workflow Overview:

- **Step 1:** Data Loading and Preprocessing
- **Step 2:** Feature Engineering and Scaling
- **Step 3:** Model Selection (Random Forest, XGBoost)
- **Step 4:** Hyperparameter Tuning
- **Step 5:** Model Evaluation (R^2, MSE, MAE)
- **Step 6:** Answering the Questions Based on Model Results

Q1: Are foodborne disease outbreaks increasing or decreasing?

To determine whether foodborne disease outbreaks are increasing or decreasing, you should analyze the trend of outbreaks over time (assuming your dataset has a date or year column). You can do this by plotting the number of outbreaks for each year. A rising or falling line in the plot will give us insight into whether outbreaks are increasing or decreasing.

```
import matplotlib.pyplot as plt

data['year'] = pd.to_datetime(data['date_column']).dt.year  # Convert date to year

yearly_data = data.groupby('year')['outbreak_column'].sum()  # Group by year and sum the outbreaks

plt.plot(yearly_data)

plt.title('Foodborne Disease Outbreaks Over Time')

plt.xlabel('Year')

plt.ylabel('Number of Outbreaks')

plt.show()
```

- **Explanation:** This code groups the data by year and sums the outbreaks for each year. Then, it plots the number of outbreaks over time. By inspecting the plot, you can determine whether outbreaks are increasing or decreasing.

ADITYA SINGHAL GDG_ML_01
23BHI10065

Q2: Which contaminant has been responsible for the most illnesses, hospitalizations, and deaths?

To answer this, you'll need to group the data by the type of contaminant (such as bacteria, virus, etc.) and then sum the corresponding illnesses, hospitalizations, and deaths. This will allow you to identify which contaminant is associated with the highest number of cases for each of these metrics.

```
contaminant_data = data.groupby('contaminant_column')[['illnesses', 'hospitalizations', 'deaths']].sum()

print(contaminant_data)
```

Explanation: This code groups the dataset by the 'contaminant_column' and sums the illnesses, hospitalizations, and deaths for each contaminant. The output will show which contaminant is most strongly linked to these health outcomes, providing insight into the most dangerous contaminants.

Q3: What location for food preparation poses the greatest risk of foodborne illness?

Here, we group the data by the location where food was prepared (e.g., restaurant, home, factory) and sum the number of illnesses, hospitalizations, and deaths for each location. This will help identify which locations are most frequently associated with foodborne illnesses.

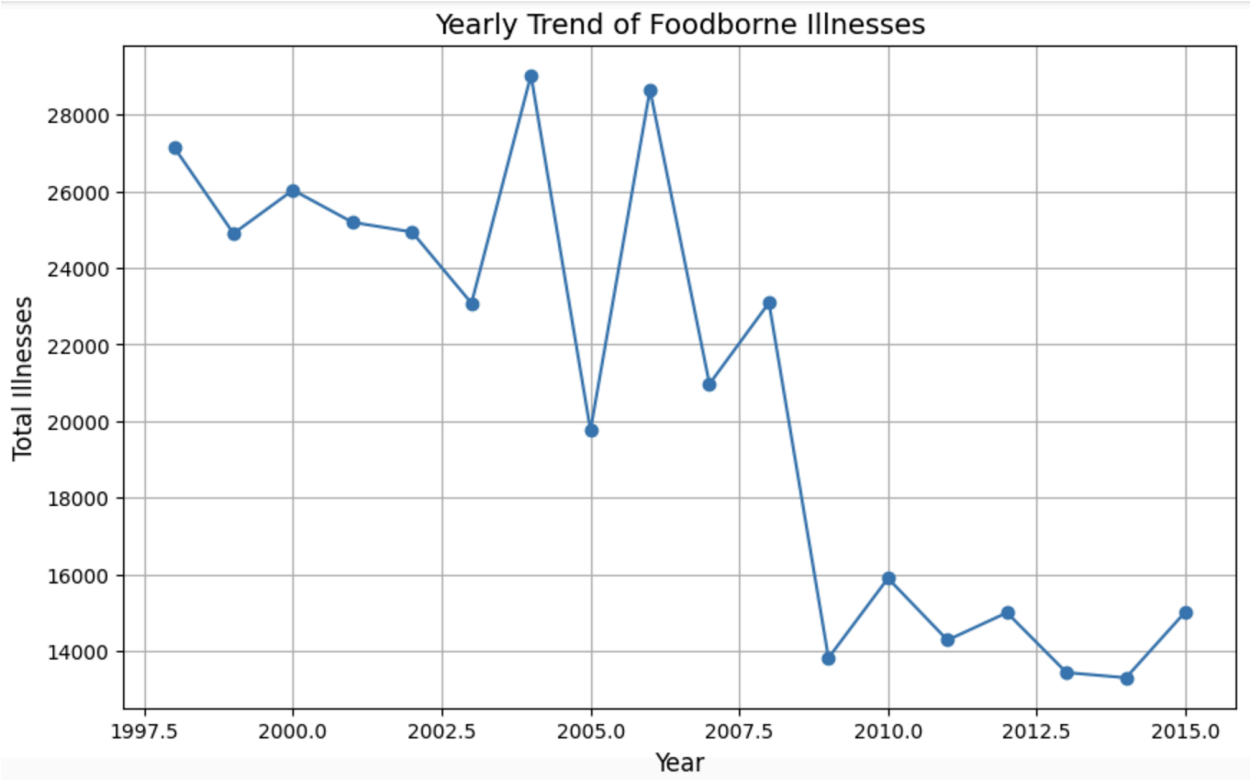
```
location_data = data.groupby('location_column')[['illnesses', 'hospitalizations', 'deaths']].sum()

print(location_data)
```

- **Explanation:** This code groups the dataset by the 'location_column' and sums the illnesses, hospitalizations, and deaths for each food preparation location. The results will help identify which locations pose the greatest risk for foodborne illnesses.

ADITYA SINGHAL GDG_ML_01
23BHI10065

ANSWER-01



Yearly Illnesses Trend:

Year	
1998	27156
1999	24899
2000	26033
2001	25192
2002	24939
2003	23079
2004	29034

ADITYA SINGHAL GDG_ML_01
23BHI10065

2005 19761

2006 28656

2007 20970

2008 23089

2009 13813

2010 15893

2011 14278

2012 14995

2013 13431

2014 13295

2015 15018

Name: Illnesses, dtype: int64

ANSWER-02

Top 5 Contaminants:

Species	Illnesses	Hospitalizations	Fatalities
Unknown	77954	967.0	27.0
Norovirus genogroup I	76406	668.0	2.0
Salmonella enterica	60018	6888.0	82.0
Norovirus genogroup II	38175	518.0	6.0
Clostridium perfringens	28734	106.0	12.0

ADITYA SINGHAL GDG_ML_01
23BHI10065

ANSWER-03

Top 5 High-Risk Locations:

Location

Restaurant 131970

Unknown 66015

Catering Service 36044

Private Home/Residence 22564

Prison/Jail 20608

Name: Illnesses, dtype: int64