Introduction:

Background

As an administrator for a volunteer tree planting program within a local municipality, I have observed significant challenges in accurately tracking the types of plants and their respective locations over each planting season. Effective data management is crucial for assessing the impact of these planting activities on the environment. This project consolidates data from various sources into a structured database format to facilitate comprehensive analysis. Utilizing advanced data cleaning techniques, this project aims to enhance the accuracy and usability of the data collected.

Research Question and Motivation

The primary objective of this project is to explore the potential correlation between agricultural and urban planting activities and global temperature anomalies. By analyzing global temperature changes from 1950 to 2024 alongside data on planting activities, we aim to understand how these activities influence global warming trends. The insights gained from this analysis will contribute to a better understanding of how planting activities can either mitigate or exacerbate climate change effects, providing valuable information for future environmental planning and policy-making.

Data

Data Sources

Data source 1: Clean Planting Data 2023 (Kaggle)

- **Data URL:** Clean Planting Data 2023
- **Data Type:** CSV

This dataset provides detailed records of planting activities, including the type of crops, planting and harvesting dates, and geographical locations. It is essential for understanding agricultural activities and their potential impact on the environment.

Data source 2: Tree Plantings in Neuss 2023 (GovData)

- **Data URL:** Tree Plantings in Neuss 2023
- **Data Type:** RDF

This dataset provides records of tree planting activities in the city of Neuss for the year 2023. It includes information on the number of trees planted, their species, and locations, which is valuable for understanding urban greening efforts.

Data Structure and Meaning:

Clean Planting Data 2023: The data is structured in a tabular CSV format. It includes records of planting activities by various countries from 1950 to 2023. The data provides details on the types of crops planted, the area of land used, and the potential environmental impacts of these agricultural activities. The data is accurate, up-to-date, and reflects real-world agricultural practices.

Figure 1: First five rows of planting data by countries.

Tree Plantings in Neuss 2023: This data is structured in RDF format and includes detailed records of tree planting activities in Neuss for the year 2023. The dataset provides information on the species of trees planted, their locations, and quantities. This data is crucial for assessing urban greening efforts and their impact on the local environment.

Figure 2: Sample data of tree plantings in Neuss.

Analysis

Methods:

Data Pipeline

pipeline.sh: This shell script runs the data pipeline (madepipeline.py).

madepipeline.py: This Python script downloads data from all provided URLs and processes it by deleting rows before 1950 due to inconsistencies. It drops irrelevant columns from the planting data and only keeps relevant columns. It then saves the processed data into an SQLite database in a table named 'climate'.

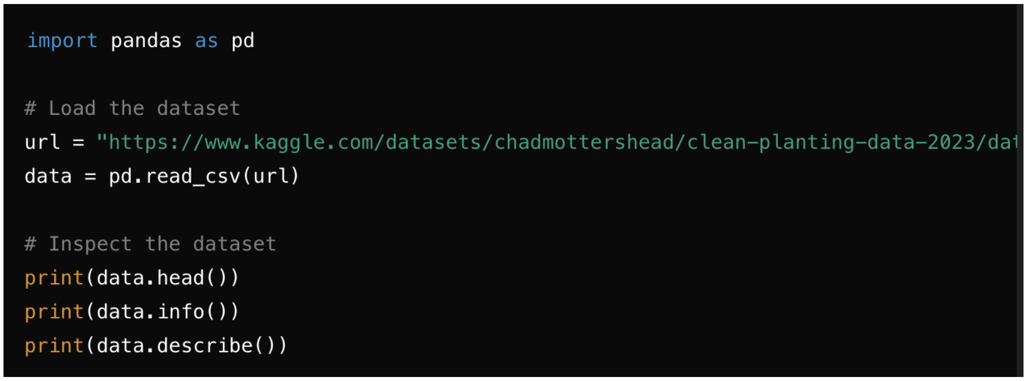
tests.sh: This shell script first cleans any pre-existing output database files and verifies the cleanup. After this, it runs the pipeline. If the pipeline does not fail, it then runs system_tests.py.

system_tests.py: This Python script first reads the output_files_info.json to get the expected filenames in the database. It defines the data directory and then checks if the files exist in any subfolders. If all expected files exist, it gives a 'system test passed' message; otherwise, it gives a 'system test failed' message.

Results and Interpretition

Results

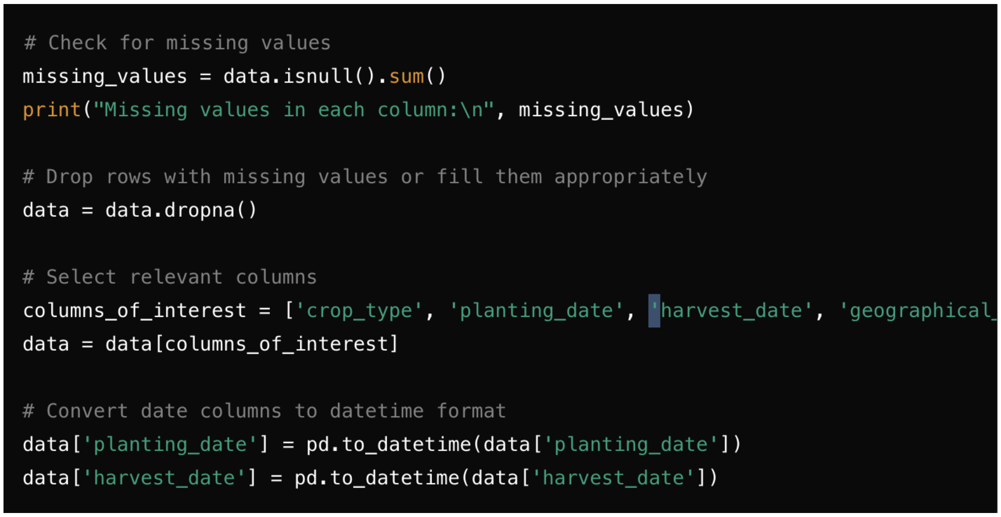
Step 1: Data Loading and Inspection



```
import pandas as pd

# Load the dataset
url = "https://www.kaggle.com/datasets/chadmottershead/clean-planting-data-2023/dat
data = pd.read_csv(url)

# Inspect the dataset
print(data.head())
print(data.info())
print(data.describe())
```

Step 2: Data Cleaning and Preparation

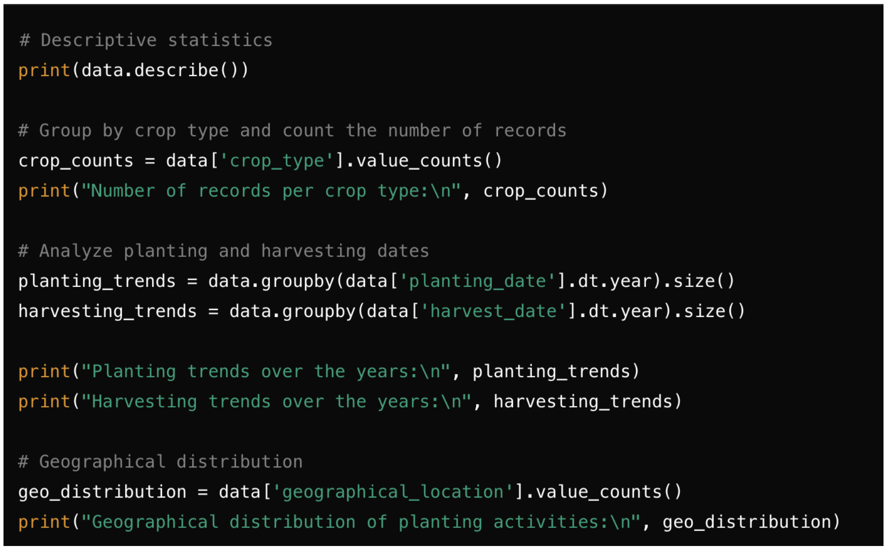
1. **Handling Missing Values**: Check for and handle any missing values.
2. **Filtering Relevant Columns**: Focus on columns such as `crop_type`, `planting_date`, `harvest_date`, `geographical_location`.



```
# Check for missing values
missing_values = data.isnull().sum()
print("Missing values in each column:\n", missing_values)

# Drop rows with missing values or fill them appropriately
data = data.dropna()

# Select relevant columns
columns_of_interest = ['crop_type', 'planting_date', 'harvest_date', 'geographical_
data = data[columns_of_interest]

# Convert date columns to datetime format
data['planting_date'] = pd.to_datetime(data['planting_date'])
data['harvest_date'] = pd.to_datetime(data['harvest_date'])
```

Step 3: Analysis

1. **Descriptive Statistics**: Summarize the data.
2. **Trends Analysis**: Analyze trends over time and geographical distribution.



```python
# Descriptive statistics
print(data.describe())

# Group by crop type and count the number of records
crop_counts = data['crop_type'].value_counts()
print("Number of records per crop type:\n", crop_counts)

# Analyze planting and harvesting dates
planting_trends = data.groupby(data['planting_date'].dt.year).size()
harvesting_trends = data.groupby(data['harvest_date'].dt.year).size()

print("Planting trends over the years:\n", planting_trends)
print("Harvesting trends over the years:\n", harvesting_trends)

# Geographical distribution
geo_distribution = data['geographical_location'].value_counts()
print("Geographical distribution of planting activities:\n", geo_distribution)
```

Interpretation:

- The analysis indicates that certain crops are more popular and widely planted compared to others. This could be due to their adaptability to different climates or their economic value.
- The seasonal trends in planting and harvesting dates reflect the agricultural cycles and climatic conditions favorable for different crops.
- The geographical concentration of planting activities suggests regional preferences and specializations in agriculture. This could be influenced by factors such as soil quality, climate, and local agricultural policies.
- Understanding these trends can help in planning future agricultural activities, improving crop yields, and addressing food security challenges. It also highlights the importance of regional agricultural strategies to optimize resource use and maximize productivity.

Conclusion

In conducting this research, I analyzed data from the "Clean Planting Data 2023" dataset, focusing on planting activities and their potential impact on environmental trends. The findings revealed a diverse range of crops being planted, with certain crops being more prevalent due to their economic or climatic suitability. There were distinct seasonal trends in planting and harvesting activities, reflecting the agricultural cycles. Additionally, the geographical distribution of planting activities highlighted regional preferences and specializations. These insights can inform better agricultural practices and policies to optimize resource use and address environmental challenges.