Task 9 – Data Preprocessing & Cleaning

Topics Covered

Removing duplicates and irrelevant data
Handling inconsistent data formats
Renaming and organizing columns for clarity

Program 1 – Data Cleaning using Pandas

```
import pandas as pd

# Create sample dataset
data = pd.DataFrame({
    'Name': ['Alice', 'Bob', 'Alice', 'David'],
    'Age': [25, 30, 25, 35],
    'City': ['New York', 'Los Angeles', 'New York', 'Chicago']
})

# Remove duplicates
data.drop_duplicates(inplace=True)

# Standardize column names
data.columns = [col.strip().lower() for col in data.columns]

# Reorder columns
data = data[['name', 'city', 'age']]

print(data)
```

Key Takeaways

Learned to clean data and remove duplicates.
Improved dataset consistency and readability.
Understood preprocessing steps essential for modeling.
Developed a structured approach to preparing raw data.