# 🚀 Military Base Clustering and Analysis using Big Data Tools

**Capstone Project – INSY 8413 | Introduction to Big Data Analytics**

**Academic Year**: 2024–2025, Semester III

**Prepared by**: Milindi Shema David (**Student ID**: 25914)

# 📘 Project Overview

This project analyzes U.S. Department of Defense (DoD) military base data using **Python** and **Power BI**.

1

### Geographical & Operational Patterns

Explore military installation distribution.

2

### Unsupervised Machine Learning

Apply KMeans clustering to group similar bases.

3

### Interactive Dashboard Design

Create a visually appealing Power BI dashboard.

4

### Real-World Application

Demonstrate Big Data tools for government problems.

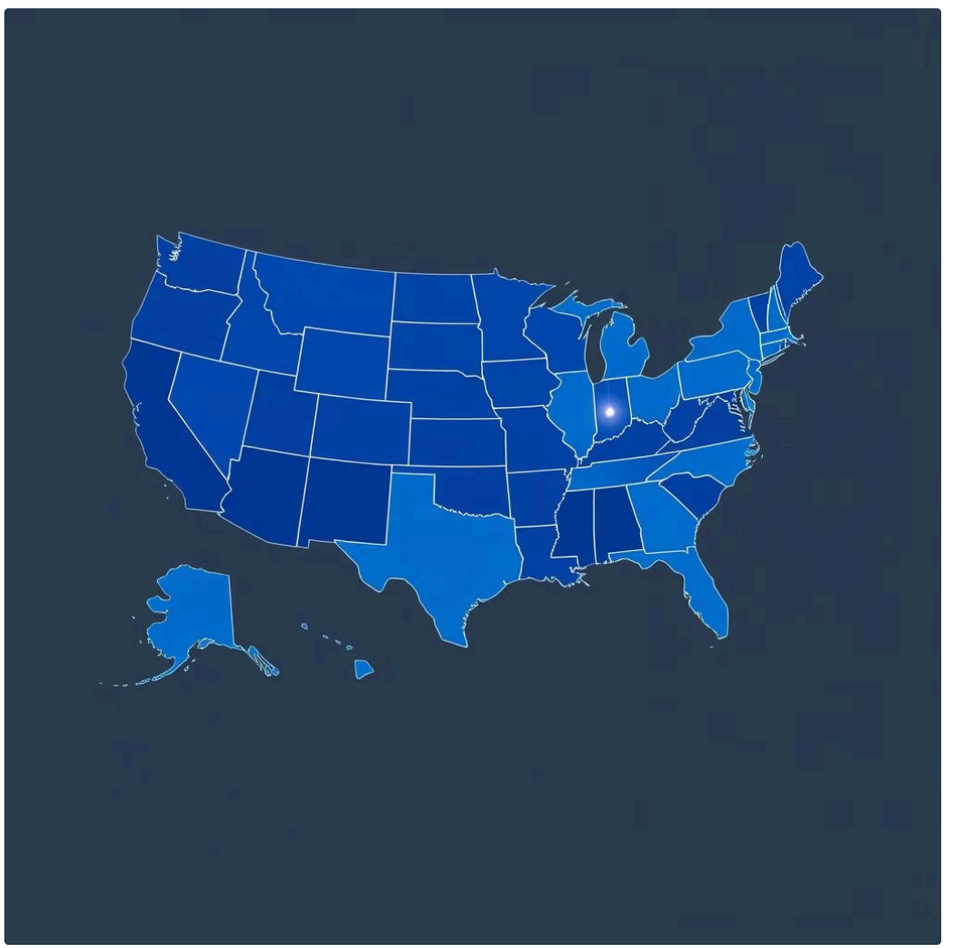
# 🏛️ Selected Sector & Problem Statement

## Selected Sector: Government

This dataset belongs to the transportation/government domain, offering insights into the spatial and administrative structure of military sites.

## Problem Statement

> How can clustering and exploratory data analytics help understand the operational distribution and geographical spread of U.S. military installations across different states and military branches?



# 📂 Dataset Information

- **Dataset Title**: Military Bases
- **Source**: [USDOT Open Data - Military Bases]
- **File Format**: CSV
- **Rows & Columns**: ~2,000 rows × 6 main attributes used
- **Data Type**: Structured
- **Preprocessing Required**: ✅ Yes


| E | F | G | H |
|---|---|---|---|
| bel:type | label:price | label:address | 1980 |
| ınd | $7,175,000.00 | Wadsen Rd, Montgomery, AL, 36105 | 0.0 |
| ınd | $3,950,000.00 | Lower Buckeye Road, Buckeye, AZ, 85326 | 0.0 |
| ınd | $3,000,000.00 | 21286 I-30, Benton, AR, 72015 | 0.0 |
| etail | $399,900.00 | 485 N Washington Ave, Titusville, FL, 32796 | 0.0 |
| etail | $3,300,000.00 | 4645 E Chandler Blvd, Phoenix, AZ, 85048 | 0.0 |
| etail | $3,175,000.00 | 300 S Gentry Blvd, Gentry, AR, 72734 | 0.0 |
| ınd | $4,807,320.00 | SW of Sheridan Rd. and Hwy 65, Redfield, AR, 72132 | 0.0 |
| etail | $4,210,526.00 | 1740 Scenic Highway, Snellville, GA, 30078 | 0.0 |
| ınd | $5,227,200.00 | River Valley Drive, Fort Smith, AR, 72908 | 0.0 |
| ınd | $13,852,080.00 | 1419 N. 2nd E., Rexburg, ID, 83440 | 0.0 |
| ınd | $3,500,000.00 | 85713, Tucson, AZ, 85713 | 0.0 |
| ınd | $4,850,841.00 | S Gate Dr, Windsor, CO, 80550 | 0.0 |
| ınd | $3,500,000.00 | 3741 E 64th Ave, Commerce City, CO, 80022 | 0.0 |
| dustrial | $7,500,000.00 | 4115 E Speedway Blvd, Tucson, AZ, 85712 | 0.0 |
| etail | $3,975,000.00 | 27596 Clinton Keith Rd, Murrieta, CA, 92562 | 0.0 |
| etail | $3,287,000.00 | 7111 E Thomas Rd, Scottsdale, AZ, 85251 | 0.0 |
| ınd | $3,575,000.00 | 31601 S State Road 85, Buckeye, AZ, 85326 | 0.0 |
| etail | $10,000,000.00 | 13909 Chenal Pky, Little Rock, AR, 72211 | 0.0 |
| ınd | $5,841,000.00 | 31 Towner Lane, Oxford, CT, 06478 | 0.0 |
| etail | $9,350,000.00 | 5755-5761 Mountain Hawk Way, Santa Rosa, CA, 95409 | 0.0 |
| etail | $6,000,000.00 | 425 Buford Hwy, Suwanee, GA, 30024 | 0.0 |
| ınd | $1,250,000.00 | Gatlin Blvd and Savona Blvd, Port Saint Lucie, FL, 34953 | 0.0 |
| etail | $3,900,000.00 | 1416 Skyland Boulevard East, Tuscaloosa, AL, 35405 | 0.0 |
| etail | $3,100,000.00 | 22202 N Cave Creek Rd, Phoenix, AZ, 85024 | 0.0 |
| ffice | $300,000.00 | 446 Magnolia Ave, Merritt Island, FL, 32952 | 0.0 |
| etail | $7,499,000.00 | 185 Vallecitos De Oro, San Marcos (San Diego), CA, 92069 | 0.0 |
| ffice | $4,045,000.00 | 8100 Shaffer Pky, Littleton, CO, 80127 | 0.0 |
| ınd | $4,358,100.00 | 213 LLC Alcovy Road, Covington, GA, 30014 | 0.0 |
| etail | $3,900,000.00 | 760 S San Pedro St, Los Angele | |
| ınd | $3,492,000.00 | 1685 Friendship Road, Hoschton, | |

# 🧪 Python Analytics Workflow (Jupyter Notebook)

All analysis was conducted in `military_bases.ipynb` through a series of structured steps.

### 1. Import Libraries & Load Data

Initial setup and data ingestion.

### 2. Data Cleaning & Selection

Refine dataset for analysis.

### 3. Exploratory Data Analysis (EDA)

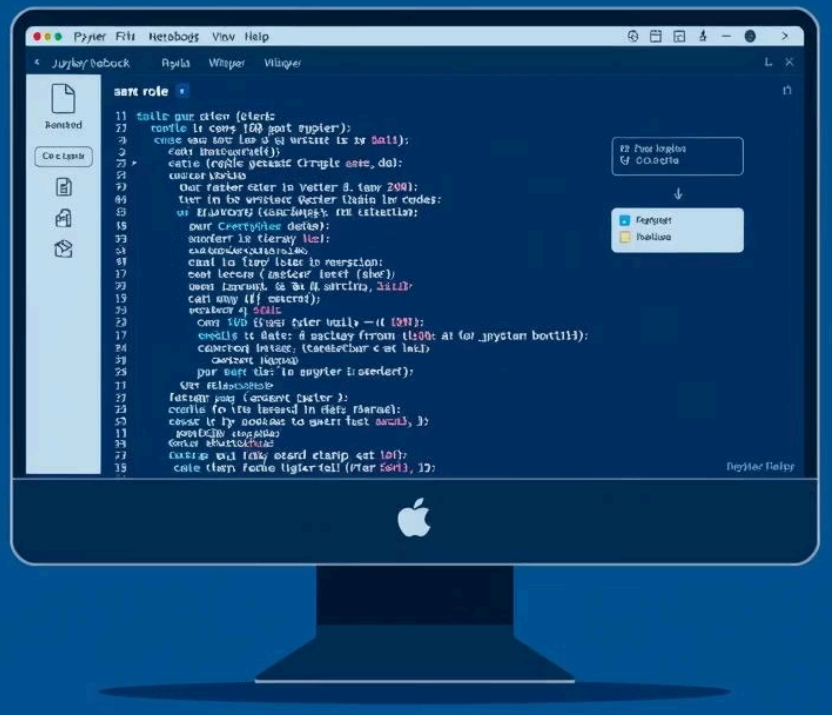
Visualize data patterns and distributions.

### 4. Feature Engineering

Prepare data for clustering algorithms.

### 5. KMeans Clustering

Group similar military installations.



# 🔷 1. Import Libraries and Load CSV

The first step involves importing necessary libraries and loading the raw CSV data into a Pandas DataFrame.

```
import pandas as pd
df = pd.read_csv("NTAD_Military_Bases.csv")
```

**Why?** To efficiently read and inspect the original structured CSV file into a manageable DataFrame for subsequent operations.

# 🔷 2. Data Cleaning & Column Selection

```
df_clean = df[[ 'Site Name', 'Site Operational Status', 'Site Reporting Component Code', 'State Name Code', 'Shape__Area', 'Shape__Length' ]].copy()
df_clean.columns = ['site_name', 'status', 'component', 'state', 'area', 'length']
df_clean.dropna(inplace=True)
df_clean['component'] = df_clean['component'].str.strip().str.upper()
```

**Why?**

- Reduce to relevant columns only for focused analysis.
- Standardize naming for improved consistency and readability.
- Remove missing data to ensure model reliability and accuracy.

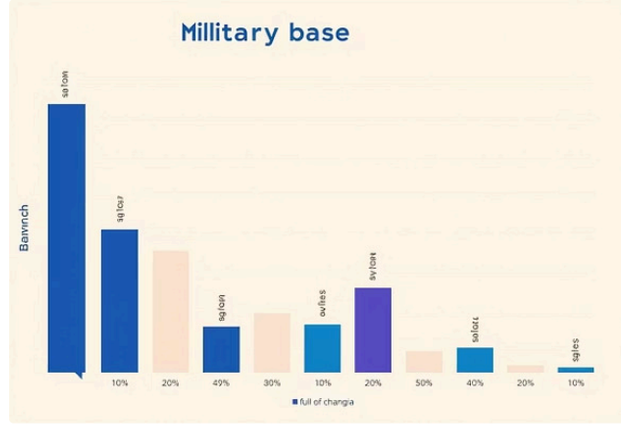
# 🔷 3. Exploratory Data Analysis (EDA)

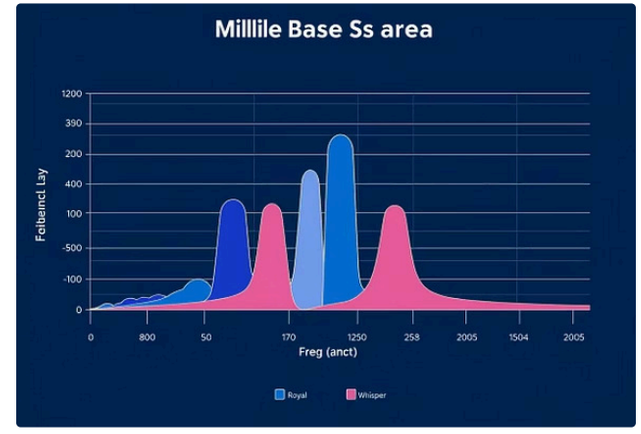
```
import seaborn as sns
import matplotlib.pyplot as plt

# Bar chart: bases by component
sns.countplot(data=df_clean, x='component')

# Histogram: area distribution
sns.histplot(df_clean['area'], bins=30, kde=True)
```

**Why?** To visualize distribution and frequency across different dimensions, such as military branch (`component`), state, and base size (`area`), gaining initial insights.



## 🔷 4. Feature Engineering for Clustering

```
from sklearn.preprocessing import LabelEncoder
df_clean['component_code'] =
LabelEncoder().fit_transform(df_clean['component'])
df_clean['status_code'] =
LabelEncoder().fit_transform(df_clean['status'])
```

**Why?** Machine learning algorithms require numerical input. Label encoding converts categorical values (like military branches and operational statuses) into numerical representations, making them suitable for clustering algorithms.


# Summary & Features Provided

## Key Takeaways

This initial phase of the project has laid a robust foundation for military base analysis:

- Successful data acquisition and cleaning.
- Initial insights gained through EDA.
- Data prepared for advanced clustering.

## Features Provided

The subsequent phases will involve:

- Implementing KMeans clustering algorithm.
- Developing interactive Power BI dashboards.
- Presenting comprehensive findings and strategic implications.