Saciva: University Clustering 1B
AI Studio Final Presentation

Break Through Tech Los Angeles @UCLA
December 6th, 2024

Name Meaning	Close friends or minister in Sanskrit
Logo Symbolism	Abstract representation of embracing hands
Mission	Guide international students, foster inclusion

SACIVA: Empowering International Students

A secure platform addressing networking, housing, and roommate challenges for over 1 million international students in the USA.

Meet Our Team!

Nam Vien
Statistics & Data Science @ UCLA

Parini Gandhi
Computer Science @ USC

Joann Sun
Computer Science @ CSUF

Jiarui Song
Statistics & MathEcon @ UCLA

Anita Demirci
Computer Science @CSULB

AI Studio Teaching Assistant & Challenge Advisor

Dominic Diaz
Fall Studio TA

Applied Mathematics
PhD Candidate @
Cornell University

Abhi Medikonduri
Challenge Advisor

CEO of Saciva
DePaul University Alum
in Business Analytics

Presentation Agenda

1. **AI Project Overview**
2. **Data Understanding & Data Preparation**
    a. Weather
    b. Cost of Living & Personal income
    c. U.S. Universities
    d. Campus Safety
3. **Modeling & Evaluation**
    a. Clustering
    b. Profiling
    c. Key Findings
4. **Conclusion and Q&A**

AI Studio Project Overview

Ever spent hours toggling filters, struggling to find what you need?

Traditional filters require users to specify preferences or navigate irrelevant options, causing frustration. For international students, this is worsened by unfamiliarity with their environment, making it harder to find trustworthy resources, compatible roommates, and local networks.

Our solution: Saciva combines machine learning with user-centric design to streamline the process:

- **Security:** Student email verification ensures authenticity, eliminating fake profiles.
- **Privacy:** In-platform messaging safeguards personal information.
- **Wide Network:** Expands connections beyond single universities, fostering city-wide networks.
- **Smart matching:** Uses clustering to group students based on geographic proximity and preferences, reducing the reliance on manual filtering.

AI Studio Project Overview

Ever spent hours toggling filters, struggling to find what you need?

Imagine being an international student...

Our Solution: SACIVA

Saciva combines machine learning with user-centric design to streamline the finding community, network, and more for international students:

- **Security:** Student email verification ensures authenticity, eliminating fake profiles.
- **Privacy:** In-platform messaging safeguards personal information.
- **Wide Network:** Expands connections beyond single universities, fostering city-wide networks.
- **Smart matching:** Uses clustering to group students based on geographic proximity and preferences, reducing the reliance on manual filtering.

Why Clustering Over Filters?

Filters are inherently limited by their reliance on predefined categories and user input.

1. **Fragment Results:** Users may miss connections or opportunities outside their specific filter criteria.
2. **Require Prior Knowledge:** International students may not know the best neighborhoods or universities to connect with.
3. **Time-Consuming:** Manually adjusting and toggling through filters is tedious and inefficient.

By clustering universities geographically:

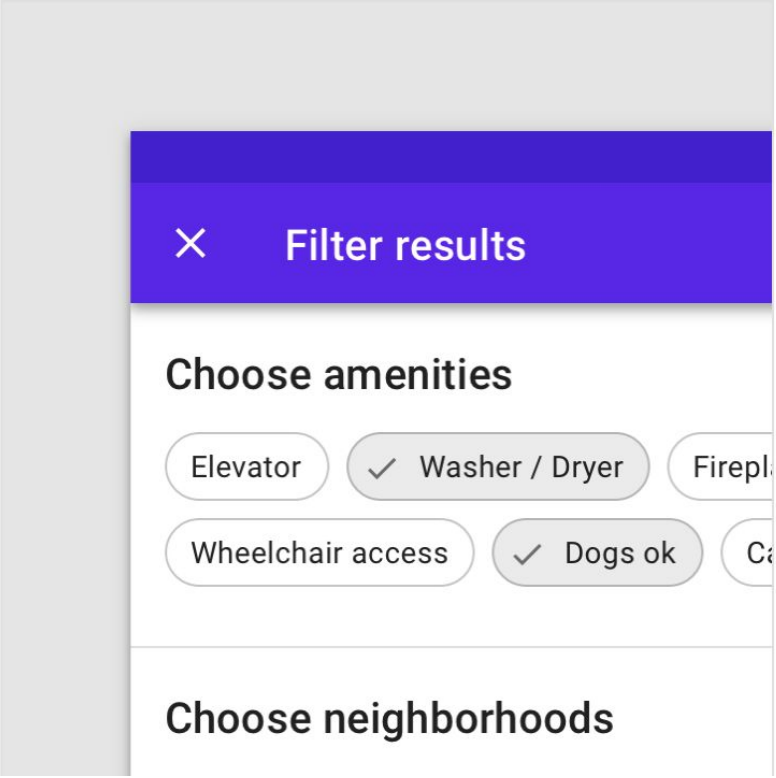
- **Broader Connections:** Students are grouped with others within a practical commuting radius (e.g., 1 to 1.5 hours), promoting regional collaboration and resource sharing.
- **Inclusive Matching:** Groups are formed organically, ensuring no one is excluded due to rigid filter parameters.
- **Simplified Navigation:** Students can easily explore options within their cluster, reducing decision fatigue.

Saciva leverages clustering to create intuitive, location-based networks that make finding housing, roommates, and local connections **seamless** and **stress-free**.

“

The goal of our project is to build a clustering model to help international students in the U.S. network, find housing, and connect with roommates, without having to use filters.

Business Impact



Why Clustering Over Filters?

- Filters are limited by reliance on predefined categories and user input
  - Users may miss connections or opportunities outside their specific filter criteria
  - Requires prior knowledge
  - Time-consuming
- Creates natural regional connections within commute radius
- Helps users discover relevant nearby opportunities
- Especially benefits international students with limited local knowledge
- Accounts for complex commuting patterns between cities and suburbs

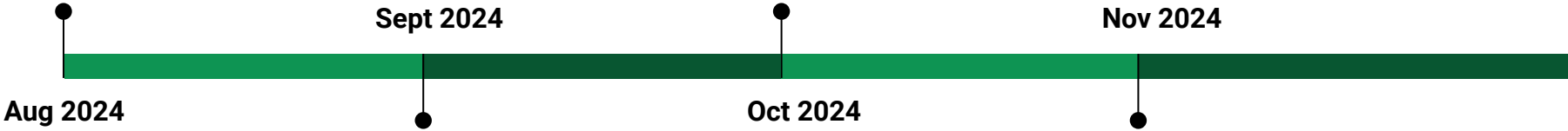
Our Approach and Timeline

Data Preparation/Data Understanding

Gathering, cleaning, and exploring the data to make sure it's ready for analysis and clustering.

Profiling

Analyzing the characteristics and features such as climate, cost of living, and safety scores for each cluster to understand their significances.



Clustering

Applying clustering algorithms to segment the data into clusters based on geographic location.

Testing

Validating clustering and profiling results through performance metrics to ensure accuracy and robustness.

Resources We Leveraged

Project Management & Shared Resources:

- Google Colaboratory
- Google Drive
- GitHub
- Discord/Google Meet/Zoom

Python Libraries:

- Scikit-learn
- Numpy
- Pandas
- Matplotlib
- Unsupervised learning libraries
- Folium

Summary of Insights & Key Findings

Primary Objective:

Clustering U.S. universities based on geographic proximity (latitude and longitude) to assist international students in networking, finding roommates, and accessing resources.

Key Findings:

- Clusters based on **geography**
- Additional profile factors:
  - **Cost of Living Analysis**
  - **Climate**
  - **Safety/Crime rates**

Technical Insights:

We practiced clustering with various models and learned how to evaluate their performance.

What is Clustering?

Technique to automatically group similar items based on their characteristics.

In Our Case: Grouping universities by geographic location to create natural networking zones

Algorithms We Explored:

- **DBSCAN (Density-Based Spatial Clustering):** Ideal for geographic data, finds natural groupings without predefined cluster count
- **Mean Shift:** Centers clusters around density peaks, adapts to data distribution
- **Agglomerative:** Builds hierarchical clusters from the bottom up

Evaluation Metrics:

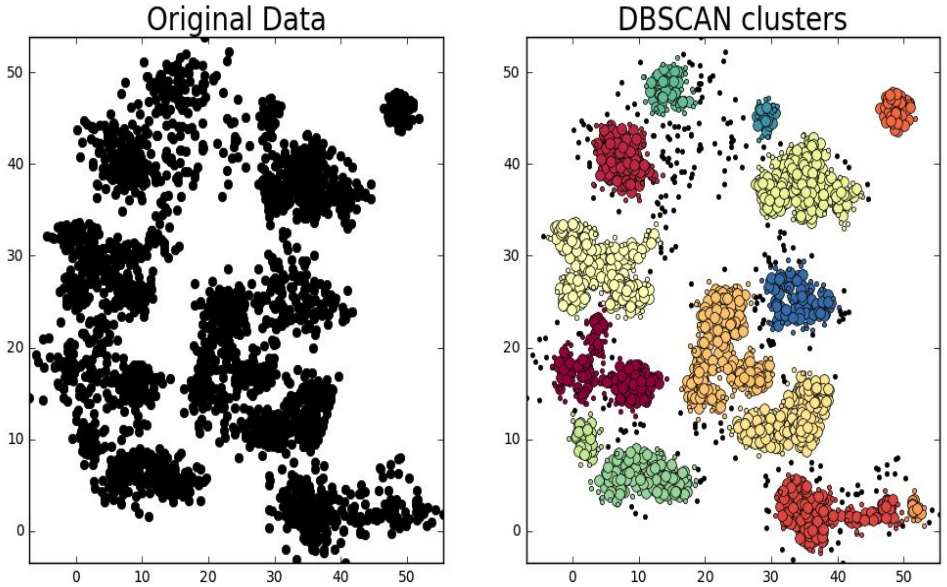
- **Silhouette Score (ranges from -1 to 1):** Measures how similar points are to their own cluster vs. other clusters.
  - ❑ **Close to 1**: The item fits well in its group.
  - ❑ **Close to 0**: The item could belong to another group.
  - ❑ **Negative**: The item is in the wrong group.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

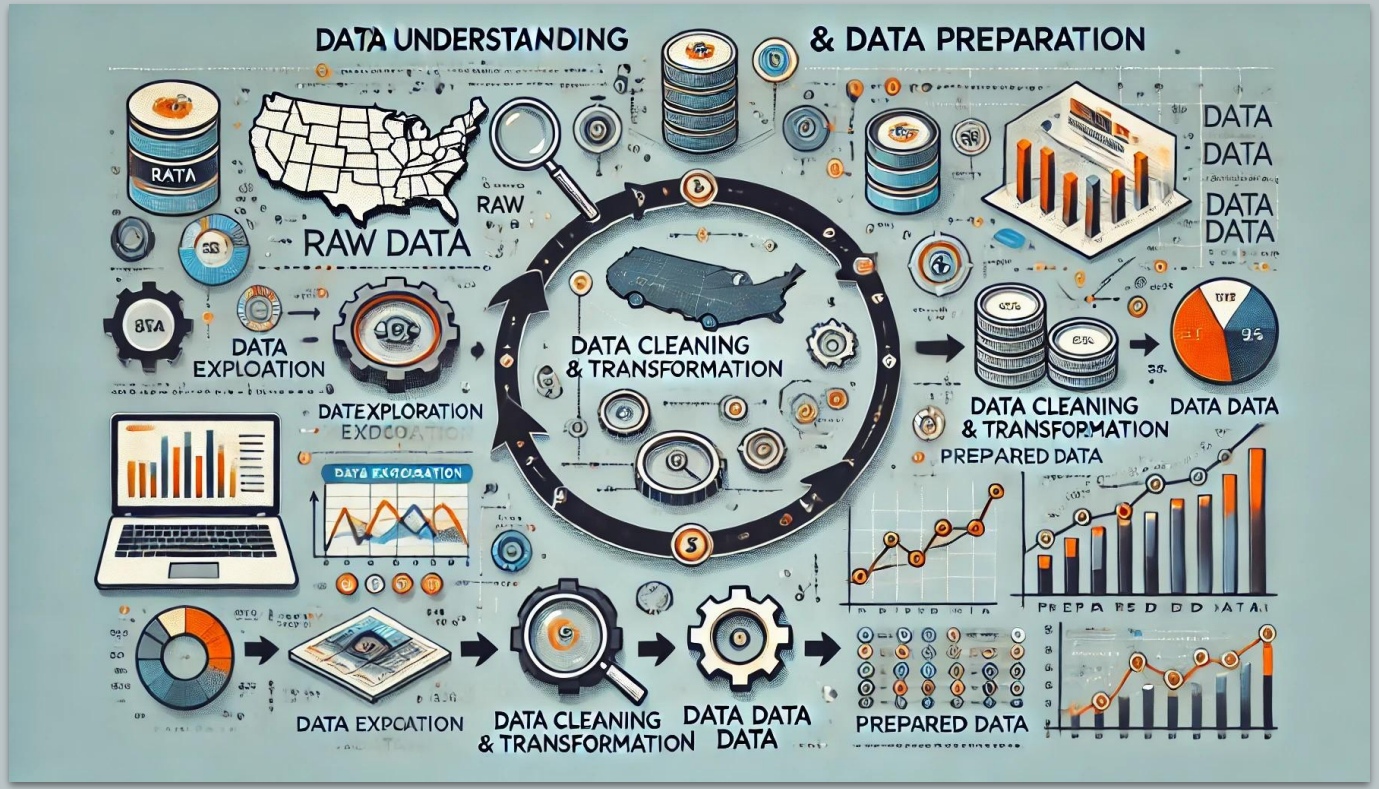
An algorithm that groups points (e.g., universities or people) based on density.

How it works?

- It connects points within a certain distance and expands groups by adding nearby points if they are dense enough.
- Points too far from any group are marked as noise.
- Unlike other methods, it doesn't require deciding the number of groups in advance.



Source: Image by Preethi Thakur



Data Understanding & Data Preparation

Data Understanding & Data Preparation

1. **Data Exploration and Characteristics:**
    - Explored geographic data of U.S. universities (latitude, longitude) and cost of living indicators.
    - Conducted Exploratory Data Analysis (EDA) to examine distribution and relationships, identifying minimum and maximum values (e.g., cost of living extremes).
    - Used visualization techniques to map geographic locations alongside cost data for better insights.
2. **Data Preparation Steps:**
    - Handled missing values, specifically for geographic and cost data, and normalized cost of living where needed.
    - Identified and managed outliers to ensure data quality and consistency.
    - Merged datasets.

NOAA Weather Datasets

Overview – Precipitation Data:

- First two columns indicate the latitude and longitudes of surveyed locations
- The following 3-14 columns indicate the total precipitation level in millimeters for that month.

Overview – Temperature Data:

- Similar structure where first two columns indicate the latitude and longitudes of surveyed locations
- The following 3-14 columns indicate the temperature in Celsius for that month

Latitude	Longitude	Month_1	Month_2	Month_3	Month_4	Month_5	Month_6	Month_7
24.5625	-81.8125	46.25	116.59	133.16	26.15	67.56	242.49	93.17
24.5625	-81.7708	44.64	115.39	131.32	25.83	70.88	240.73	92.03
24.5625	-81.7292	40.23	108.72	127.26	20.68	84.64	231.77	82.67
24.5625	-81.6875	36.86	102.91	123.56	16.22	96.79	225.29	74.02
24.6042	-81.6458	32.68	94.9	120.42	10.27	115.61	221.17	63
24.6458	-81.5625	32.31	94.1	120.79	10.47	122.62	228.93	61.75
24.6458	-81.4792	32.46	92.88	122	10.63	128.52	233.98	59.24
24.6875	-81.5625	32.48	94.48	121.6	10.81	123.06	233.99	63.19
24.6875	-81.3958	32.73	91.43	124.25	10.7	136.66	243.78	57.83

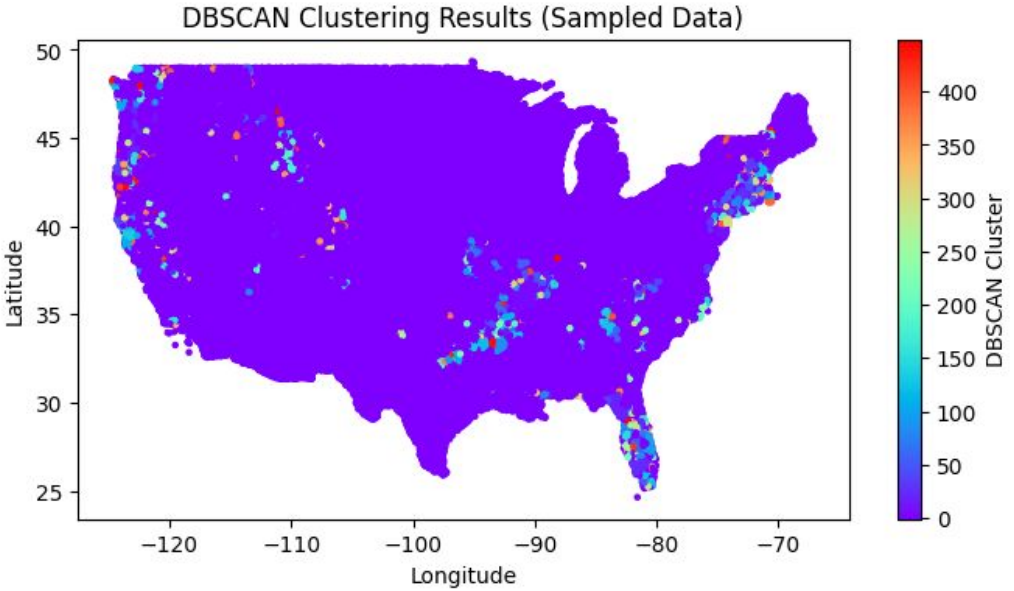
temp_dataset

Latitude	Longitude	Month_1	Month_2	Month_3	Month_4	Month_5	Month_6	Month_7	Month_8	Month_9	Month_10	Month_11	Month_12
24.5625	-81.8125	23.04	25.29	25.49	26.71	27.14	29.25	30.58	30.4	29.37	27.43	25.35	21.86
24.5625	-81.7708	23.04	25.29	25.56	26.71	27.14	29.25	30.57	30.42	29.37	27.44	25.38	21.86
24.5625	-81.7292	23.16	25.4	25.6	26.76	27.19	29.25	30.55	30.51	29.48	27.46	25.47	21.96
24.5625	-81.6875	23.26	25.49	25.64	26.76	27.19	29.25	30.54	30.58	29.54	27.51	25.54	22.04
24.6042	-81.6458	23.39	25.62	25.68	26.67	27.1	29.2	30.57	30.73	29.65	27.54	25.55	22.13
24.6458	-81.5625	23.26	25.47	25.58	26.68	27.06	29.2	30.5	30.68	29.58	27.47	25.52	22.04
24.6458	-81.4792	23.18	25.44	25.57	26.63	27.02	29.2	30.5	30.61	29.55	27.47	25.52	21.99
24.6875	-81.5625	23.14	25.4	25.52	26.58	27.07	29.15	30.46	30.66	29.57	27.42	25.52	21.98

NOAA Weather Datasets

Data Analysis:

- Map – purple clusters indicate the most frequently occurring weather patterns across the U.S., while the red clusters indicate the least frequent.
- Distinct areas show high or low annual precipitation, defining climate zones.
- Most regions experience moderate rainfall, with few outliers on either end.
- The average annual precipitation value is 167 , and the highest precipitation value is 504, which occurs during the 1st month, and the lowest precipitation value is 0 during Month 7.



Cost of Living & Personal Income Dataset

Cost of Living Dataset Overview:

- 510 cities total
- Cost of Living Index (CLI) is based on six major expense categories:
  - Food: 16.1%
  - Housing: 23.2%
  - Utilities: 10.1%
  - Transportation: 18.6%
  - Healthcare: 9.6%
  - Consumer Discretionary Spending: 22.3%
- CLI of 100 = average cost of living in the United States

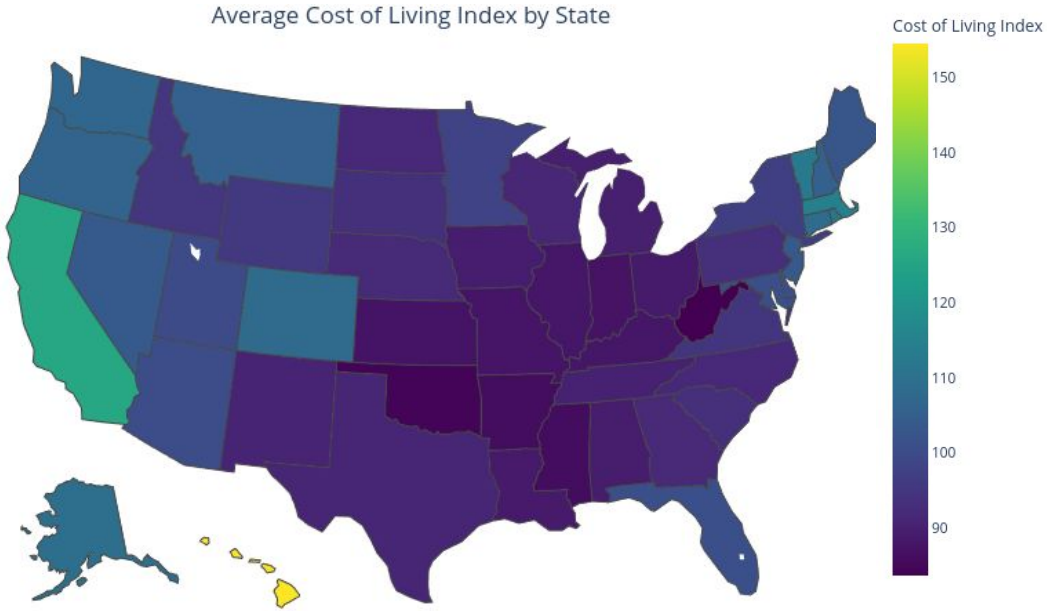
Personal Income Dataset Overview:

- Groups counties by state
- Provides average state income and average income in each of its counties from 2021-2023
- Provides percent change from previous years

Cost of Living & Personal Income Dataset

Data Analysis:

- West Coast (particularly California) and Hawaii dominate the high-cost areas
- Most cities in 80-100 range CLI
- Most outliers are concentrated in:
  - California (11 cities)
  - Hawaii (4 cities)
  - Massachusetts (2 cities)
- Cities are the centroid when clustering
- For Personal Income, we focus on Average Income from 2023 (latest in the dataset)



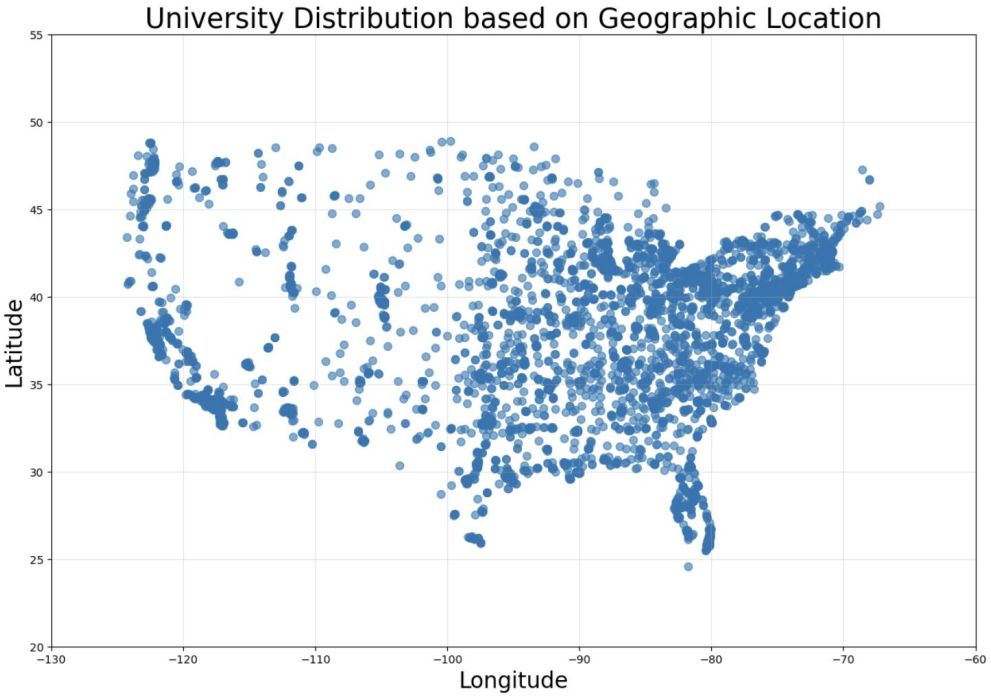
U.S. Universities Dataset

Overview:

- 6,559 universities and 22 features per university, including:
  - location, enrollment housing, and employment data
- Provide university names and location indicators (latitude and longitude).

Data Analysis:

- High-cost cities have dense clusters of institutions.
- California, New York, and Texas host the largest student clusters.
- Public institutions dominate, followed by private non-profits, highlighting accessibility.
- Higher enrollment correlates with larger staff sizes.
- California and Texas excel in total enrollment and dorm capacity, showcasing robust on-campus resources.



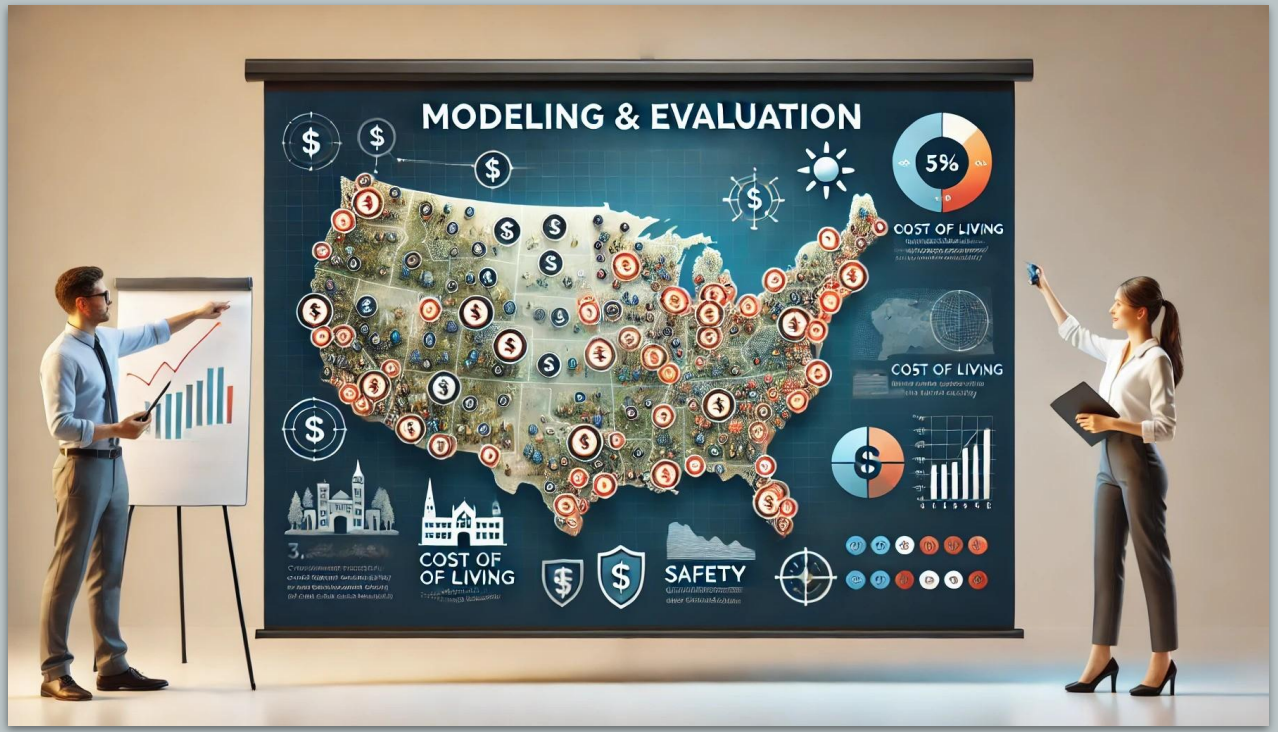
Campus Safety Dataset

Overview:

- 10,499 universities with 49 features for each university documented:
  - Campus Address & ID
  - Type of University (Private/Public)
  - Student enrollment count
  - Type of Crime and its count

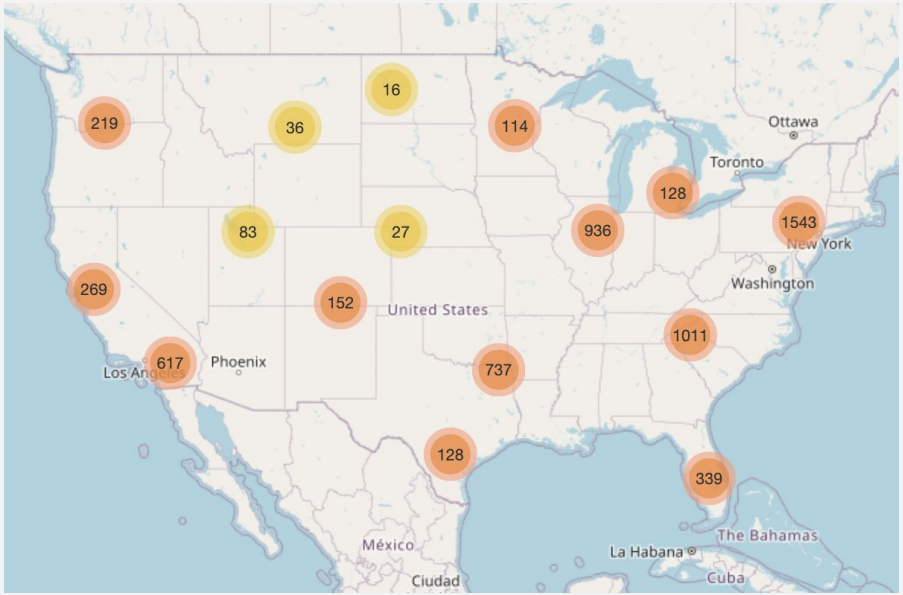
Data Analysis:

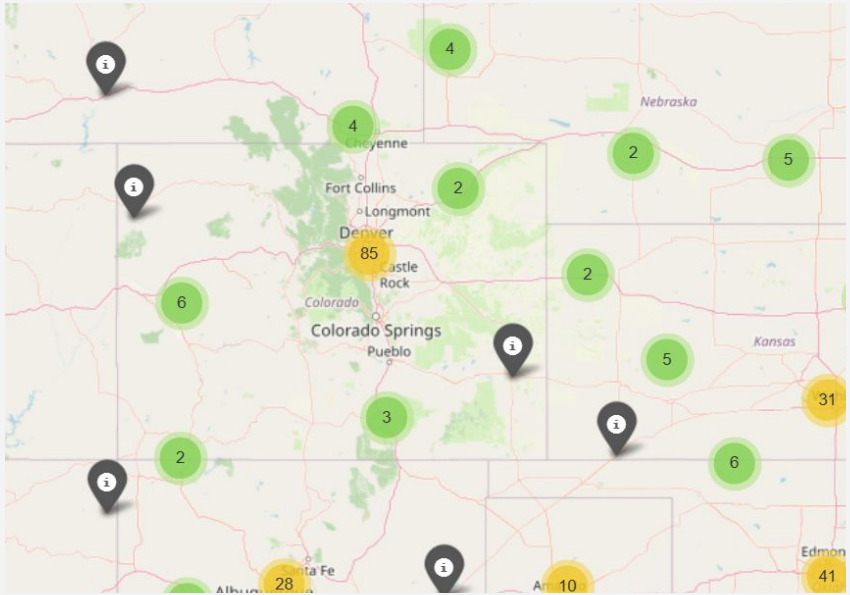
- We total the crimes into a single crime counter/safety score for each campus
- Most common reported incidents:
  - Burglary and aggravated assault
- Strong correlation between campus size (enrollment) and total incident reports
- Public universities typically report more incidents, possibly due to larger average size
- Data suggests many smaller institutions report zero incidents across all categories
- Data quality considerations:
  - Reporting standards may vary between institutions
  - Some campuses show incomplete data across years

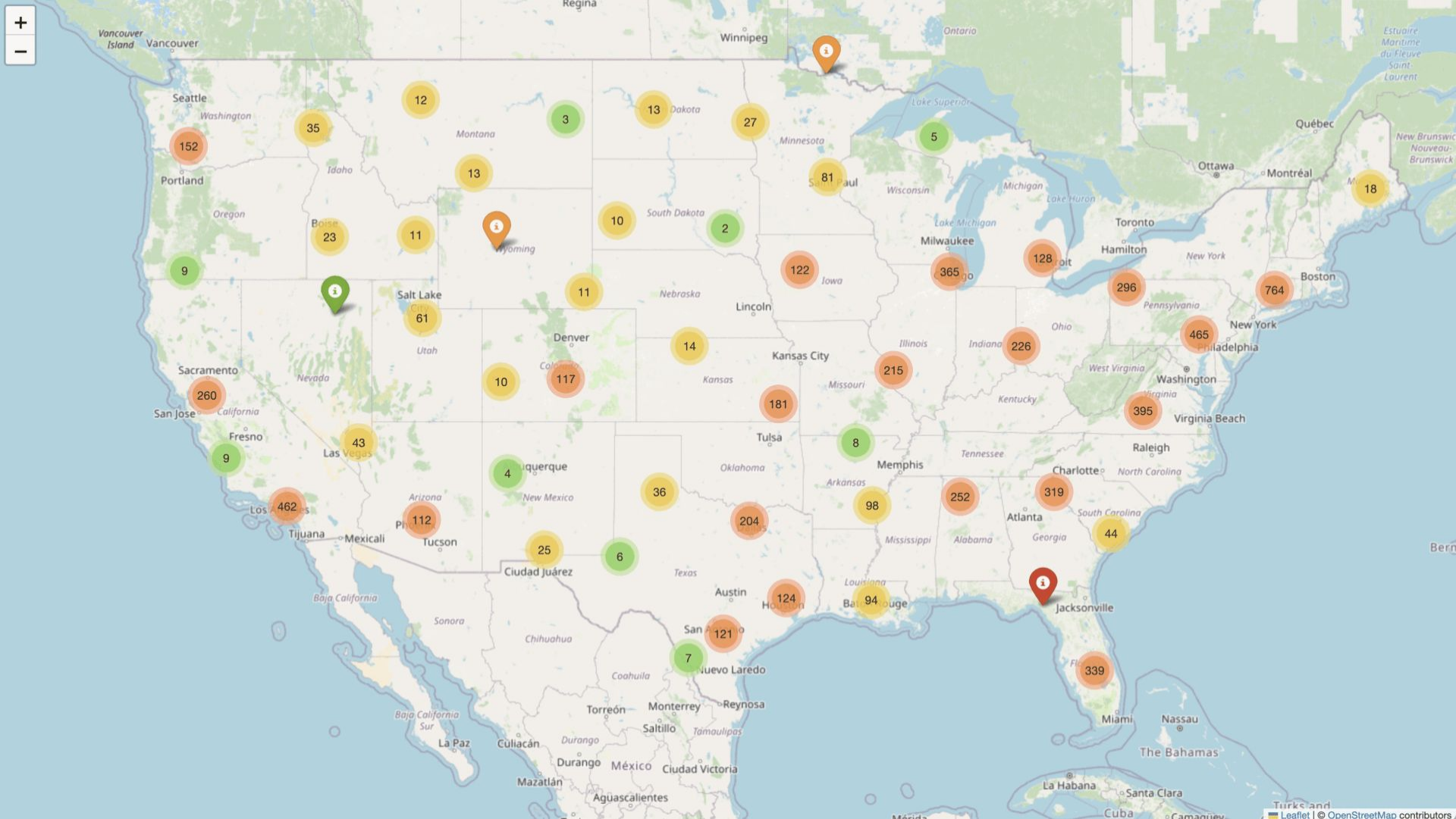


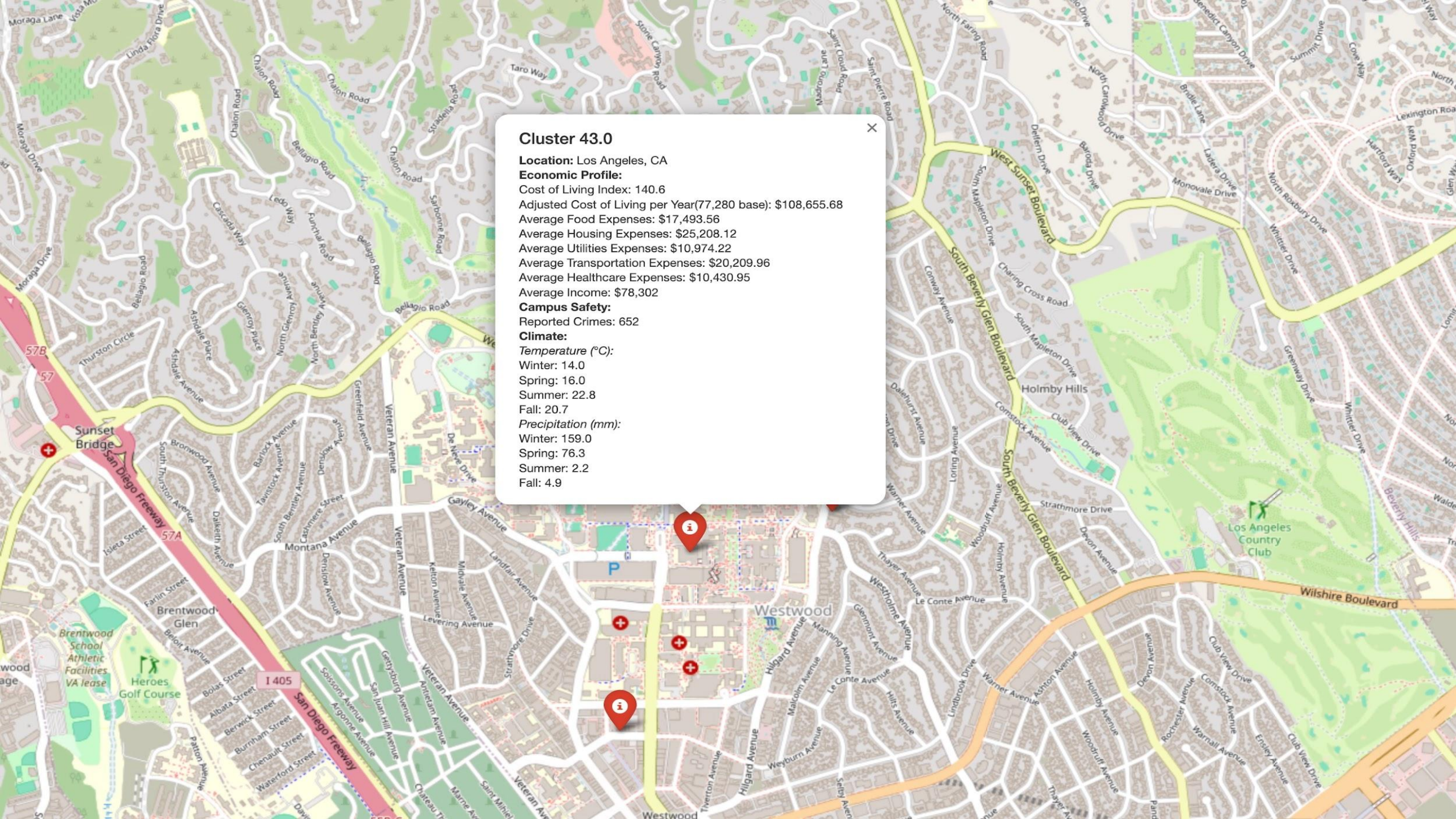
Modeling & Evaluation

Clustering Map






Vancouver Island
Vancouver
Regina
Winnipeg
Ontario
Seattle
Washington
Montana
Dakota
Minnesota
Lake Superior
Québec
Portland
Oregon
Idaho
Wisconsin
Michigan
Lake Huron
Ottawa
Montréal
Boise
Wyoming
South Dakota
Lake Michigan
Toronto
Milwaukee
Hamilton
New York
Salt Lake
Nebraska
Iowa
Lincoln
Boston
Pennsylvania
Utah
Denver
Ohio
New York
Philadelphia
Sacramento
Nevada
Kansas City
Illinois
Indiana
Kansas
Missouri
West Virginia
Washington
San Jose
California
Kentucky
Virginia
Virginia Beach
Fresno
Las Vegas
Tulsa
Raleigh
Tennessee
Memphis
Oklahoma
Charlotte
North Carolina
Arkansas
Arizona
New Mexico
Atlanta
South Carolina
Los Angeles
Phoenix
Tucson
Tijuana
Mexicali
Mississippi
Alabama
Georgia
Ciudad Juárez
Texas
Austin
Houston
Baton Rouge
Jacksonville
Baja California
Sonora
Chihuahua
San Antonio
Nuevo Laredo
Coahuila
Miami
Nassau
Torreón
Monterrey
Reynosa
Saltillo
Tamaulipas
The Bahamas
La Paz
Culiacán
Durango
México
Ciudad Victoria
Mazatlán
La Habana
Santa Clara
Aguascalientes
Cuba
Leaflet | © OpenStreetMap contributors



Cluster 43.0

Location: Los Angeles, CA
Economic Profile:
Cost of Living Index: 140.6
Adjusted Cost of Living per Year(77,280 base): $108,655.68
Average Food Expenses: $17,493.56
Average Housing Expenses: $25,208.12
Average Utilities Expenses: $10,974.22
Average Transportation Expenses: $20,209.96
Average Healthcare Expenses: $10,430.95
Average Income: $78,302
Campus Safety:
Reported Crimes: 652
Climate:
Temperature (°C):
Winter: 14.0
Spring: 16.0
Summer: 22.8
Fall: 20.7
Precipitation (mm):
Winter: 159.0
Spring: 76.3
Summer: 2.2
Fall: 4.9

Clustering Model Testing

- **Finding the best model involved parameter fine tuning**
  - Used Haversine distance to account for Earth's curvature
  - Set a 20-mile radius for cluster boundaries
  - Optimized parameters using silhouette scores
  - Selected appropriate cluster centroids
  - Epsilon (eps) – using 20 miles (instead of 50) as the max distance for any two universities to be considered clusters
  - Min_samples: the minimum number of points required within a specified radius/epsilon for a point to be classified as a core point
    - contribute to forming a cluster in density-based clustering algorithms like DBSCAN
- **Making a decision about what the cluster centroid would be—geographic mean vs city as centroid**

Clustering Model Comparison

Model Name	Description	Results	Pros	Cons
DBSCAN	Density-based clustering detects core points and noise	Achieved the highest silhouette score of **0.868** with well-separated and cohesive clusters among the three methods.	Handles noise and detects clusters of various shapes.	Struggles with varying densities, sensitive to parameters.
Mean Shift	Clusters around data density peaks	Also performed well with a silhouette score of **0.831**, but slightly less effective than DBSCAN in terms of clustering quality.	No need to specify cluster count; effective with non-linear clusters.	Computationally expensive, sensitive to bandwidth choice.
Agglomerative	Hierarchical merging based on proximity	Resulted in the lowest silhouette score of **0.474**, meaning that clusters are not as well defined as DBSCAN and Mean Shift.	Clear hierarchy, flexible linkage methods.	High time complexity, sensitive to noise and outliers.

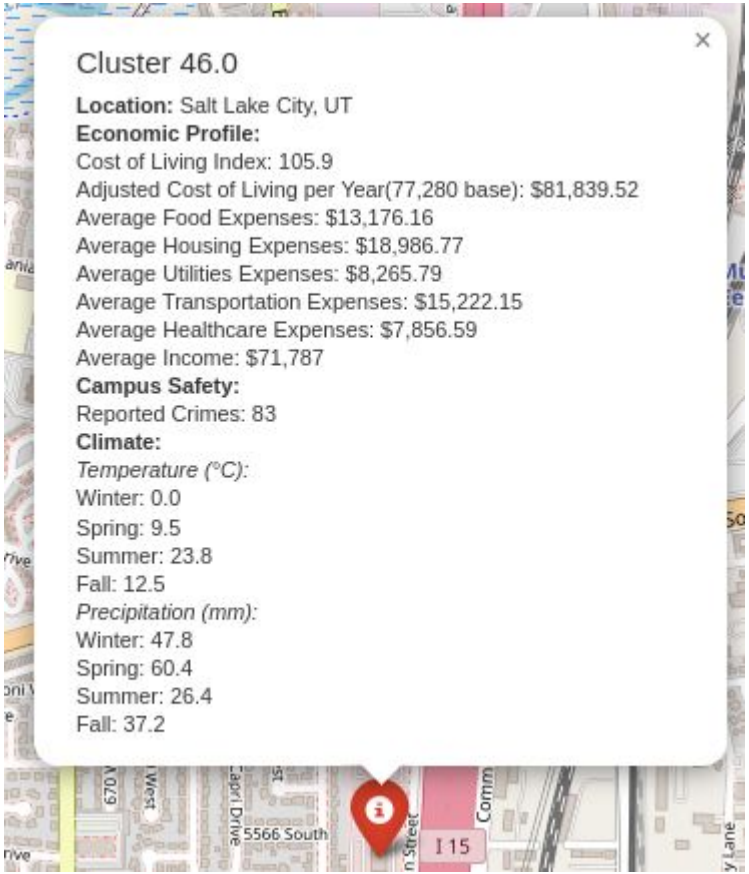
Final Clustering Model Selection: DBSCAN

- DBSCAN achieved the **highest silhouette score of 0.868**
- DBSCAN is also **robust to noise**, which makes it suitable for datasets with varying densities
- The number of clusters identified by DBSCAN (excluding noise) is 294, with 695 data points labeled as noise (points that do not meet the density requirements to form a cluster)

* The primary metric used in this analysis is the *silhouette score* because it effectively measures how similar each point is to other points within the same cluster and how distinct or well-separated a cluster is from other clusters, ensuring cohesive and high-quality clusters.

Profiling

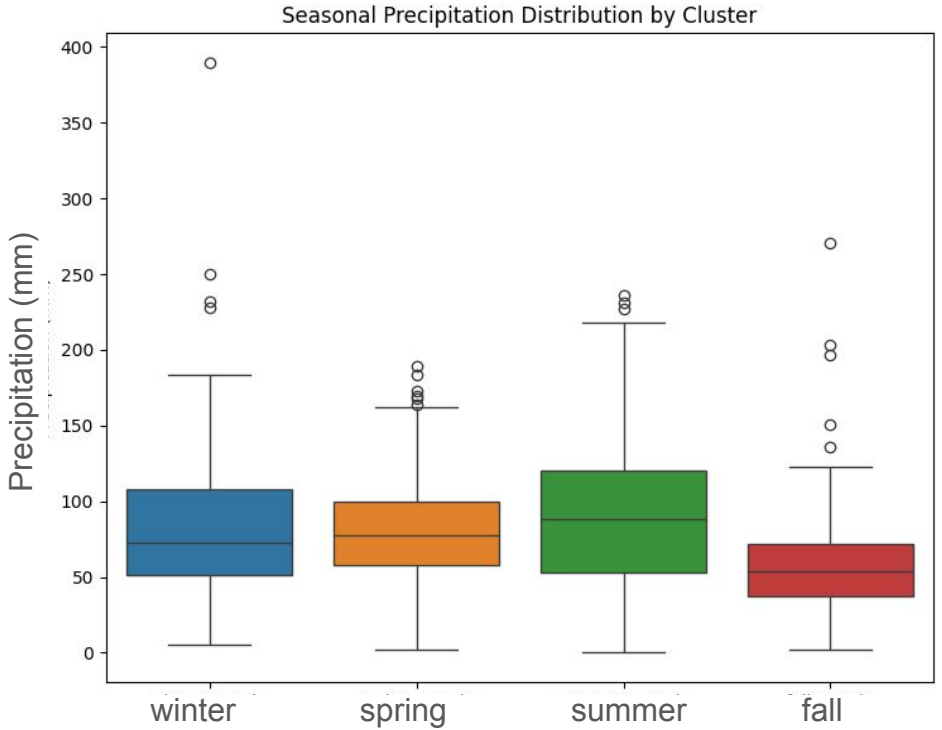
- Merged datasets using geographic location:
  - Temperature & Weather
  - Cost of Living & Personal Income
  - Campus Crime Reports
- Average expenses calculated using Average Consumer Unit Expenditure: [$77,280 per year]
- Libraries/Tools used:
  - Geopy: Geographic Calculations
  - Pandas/Numpy: Data Merging & Distance Calculations
  - Folium: Interactive Map Visualization & Profile→



Profiling continued

Why These Profiles?

- **Seasonal Climate:**
  - Different weather across academic terms
  - Impacts commute and housing decisions
- **Economic Factors:**
  - Regional cost of living affects budget planning
  - Local income levels can indicate job market strength
- **Safety Analysis:**
  - Campus crime statistics for informed decisions
  - Helps evaluate surrounding areas
  - Important for student/parent confidence



Insights and Key Findings

Improved user experience by replacing manual filtering with intelligent cluster-based recommendations:

- Used DBSCAN (0.868 silhouette score) to achieve 294 clusters reflecting student mobility patterns.
- Considered commuter proximity instead of strict city boundaries.
- Combined data sources to profile areas by housing affordability, climate, safety, and network access.
- Enabled informed decisions for international students.
- Established a foundation for automated, preference-based student matching to avoid manual research and navigating complex city boundaries.

Final Thoughts

What We Learned

- **Technical Skills & Takeaways**:
    - Real data needs heavy cleaning unlike neat academic datasets
    - Clustering models, different Python libraries, profiling
    - Patient data exploration helps align with business goals
- **Other Takeaways**:
    - Team collaboration taught us communication and conflict management
    - We learned to balance tasks, deadlines and team accountability
    - Staying flexible and open to feedback improved our problem-solving

Potential Next Steps

- Build preference-based matching system for personalized recommendations
- Develop automated pipeline for real-time cluster data updates
- Expand cluster profiles with detailed cost, climate, and safety metrics
- Add employment rates and job opportunities for each academic major in cluster regions
- Provide insights into dominant industries or sectors within each region, aligning academic majors with local job markets.
- Provide data on scholarships, grants, and other funding opportunities specific to universities and clusters

We are incredibly grateful to **Abhi Medikonduri** for his visionary leadership as the CEO of Saciva, which has greatly inspired and supported our team, and to **Dominic Diaz** for his invaluable guidance and expertise as our Fall Studio TA throughout this project.


Thank you!
Any feedback/questions?