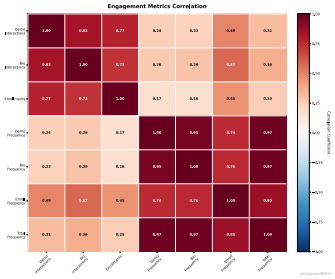
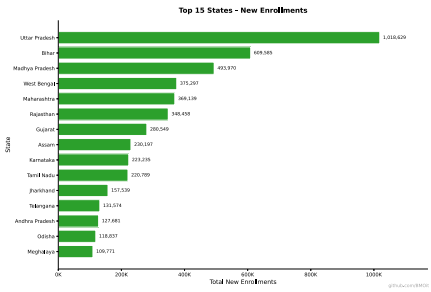
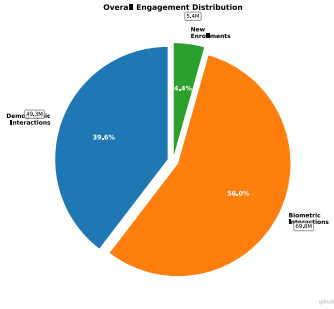
https://github.com/BMOit/UIDAI-Data-Hackathon-2026

UIDAI DATA HACKATHON 2026







Understanding Repeated Aadhaar Engagement Patterns and Their Impact on Users

Author: Rajneesh

University: Jaypee University of Information Technology, Solan, Himachal Pradesh

https://aadhaar.rajneesh.blog

Project Theme: Data Driven Insights for Better Aadhaar Services

This project analyzed 1.5 million Aadhaar engagement records to understand patterns of repeated user interactions with UIDAI systems and provide actionable recommendations for service improvement.

The Research Objectives of this Project

1. Identify repeated engagement patterns across India.
2. Understand geographic and temporal variations.
3. Discover user personas through machine learning
4. Provide data-driven recommendations for UIDAI.

Tech Stack Used:

Python is used as the main programming language has it has really good libraries for data analytics.

For Data Processing, Pandas and Numpy is used.

For Visualization, Matplot aand Seaborn is used.

We used Scikit-learn 1.3.0, K-means clustering, pca and StandardScaler preprocessing.

The used datasets were provided on the hackathon's page.

Github Repository: https://github.com/BMOit/UIDAI-Data-Hackathon-2026

Email: rtbmo23@gmail.com

Listed on https://aadhaar.rajneesh.blog

This project will be proudly open source and open for everyone to see and learn from.

Anyone can start with it and usage commands are mentioned in the Github Readme.

Dataset Overview

Demographics Example

Column	Type	Description	Example
date	Date	Update date	01-03-2024
state	String	State name	Uttar Pradesh
district	String	District name	Gorakhpur
pincode	Integer	6-digit pincode	273213
demo_age_5_17	Integer	Updates for age 5-17	49
demo_age_17_	Integer	Updates for age 18+	529

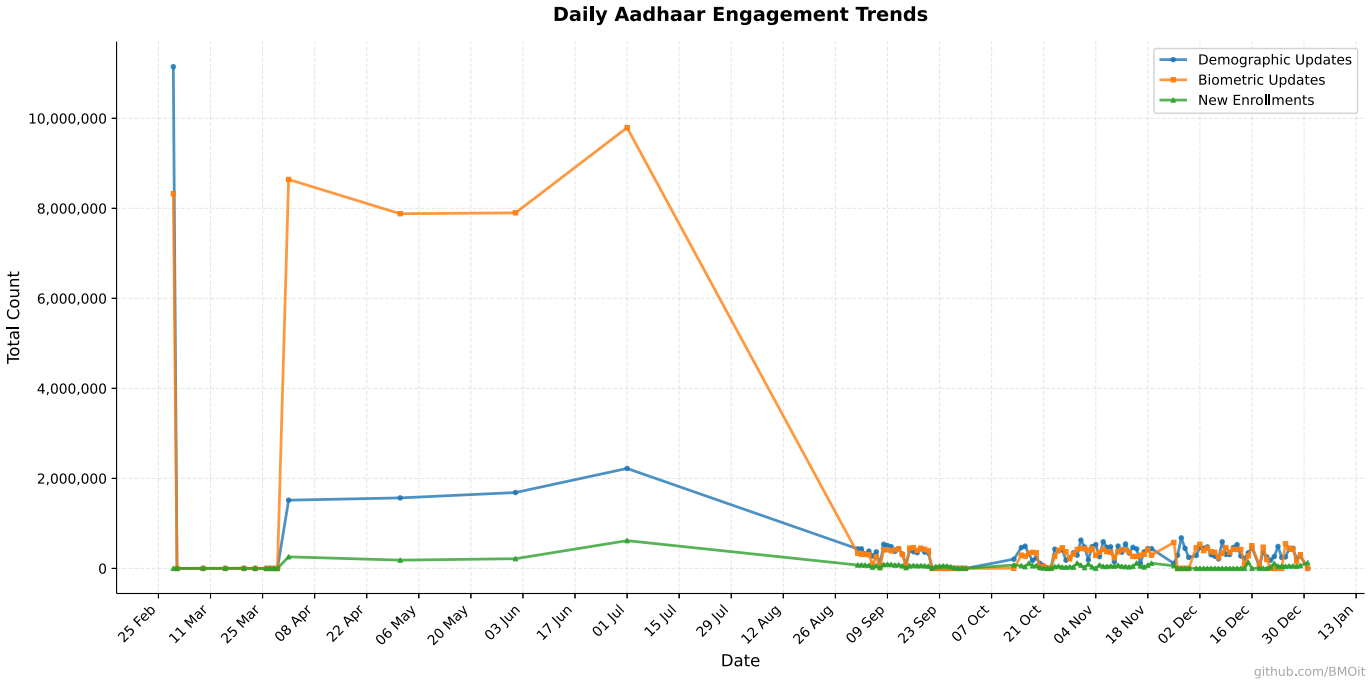
Biometrics Example

Column	Type	Description	Example
date	Date	Update date	01-03-2025
state	String	State name	Haryana
district	String	District name	Mahendragarh
pincode	Integer	6-digit pincode	123029
bio_age_5_17	Integer	Biometric updates age 5-17	280
bio_age_17_	Integer	Biometric updates age 18+	577

Enrollment Example

Column	Type	Description	Example
date	Date	Enrollment date	02-03-2025
state	String	State name	Karnataka
district	String	District name	Bengaluru Urban
pincode	Integer	6-digit pincode	560043
age_0_5	Integer	Enrollments age 0-5	14
age_5_17	Integer	Enrollments age 5-17	33
age_18_greater	Integer	Enrollments age 18+	39

The datasets are stored in the github repository at https://github.com/BMOit/UIDAI-Data-Hackathon-2026/tree/main/Datasets.



LINE 1 - DEMOGRAPHIC (blue):

X-axis: demo['date'] (grouped by day)

Y-axis: SUM(demo['demo_age_5_17'] + demo['demo_age_17_']) per day

LINE 2 - BIOMETRIC (orange):

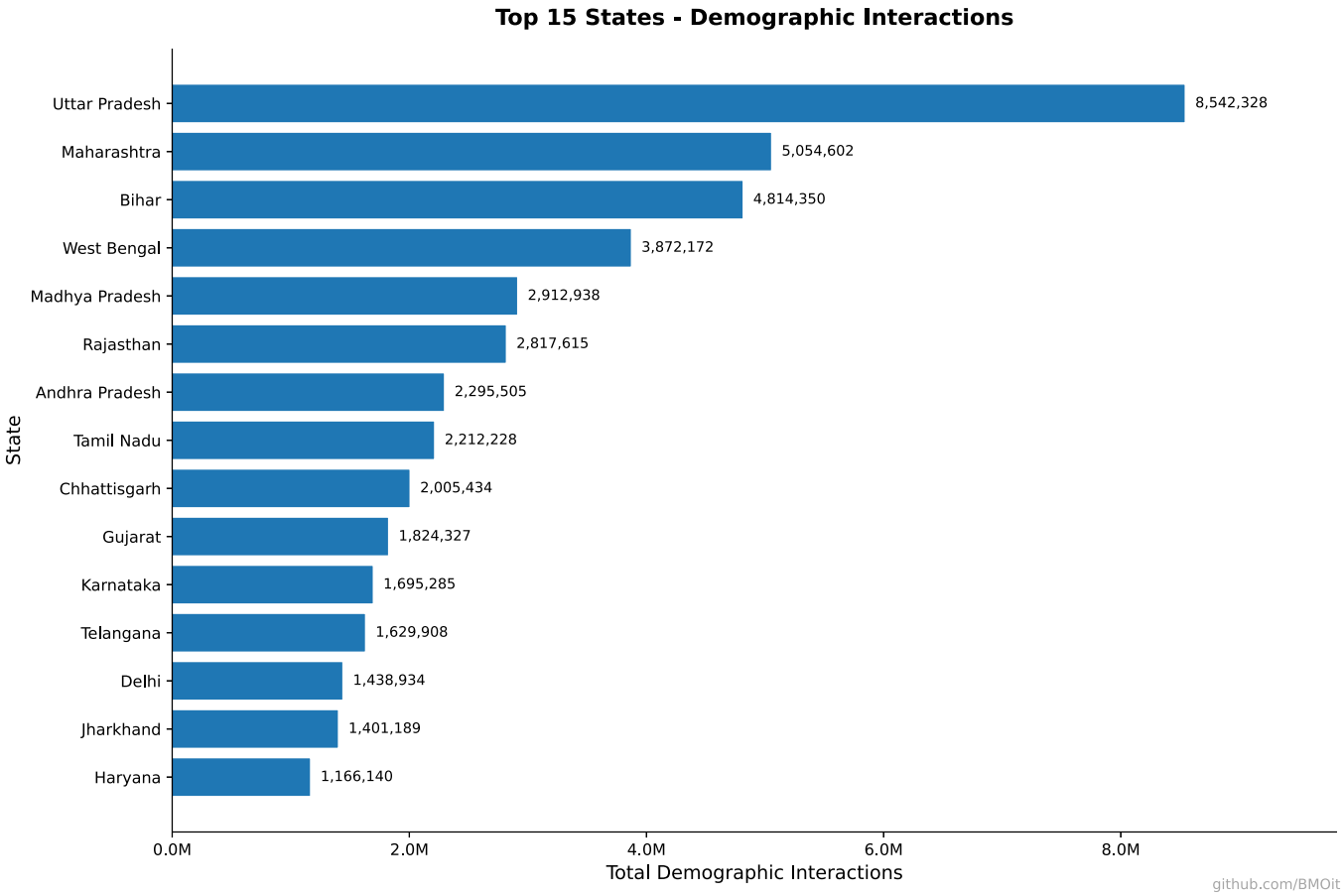
X-axis: bio['date'] (grouped by day)

Y-axis: SUM(bio['bio_age_5_17'] + bio['bio_age_17_']) per day

LINE 3 - ENROLLMENT (green):

X-axis: enroll['date'] (grouped by day)

Y-axis: SUM(enroll['age_0_5'] + enroll['age_5_17'] + enroll['age_18_greater']) per day



X-axis: Total demographic updates per state

Y-axis: State names (top 15)

Calculation:

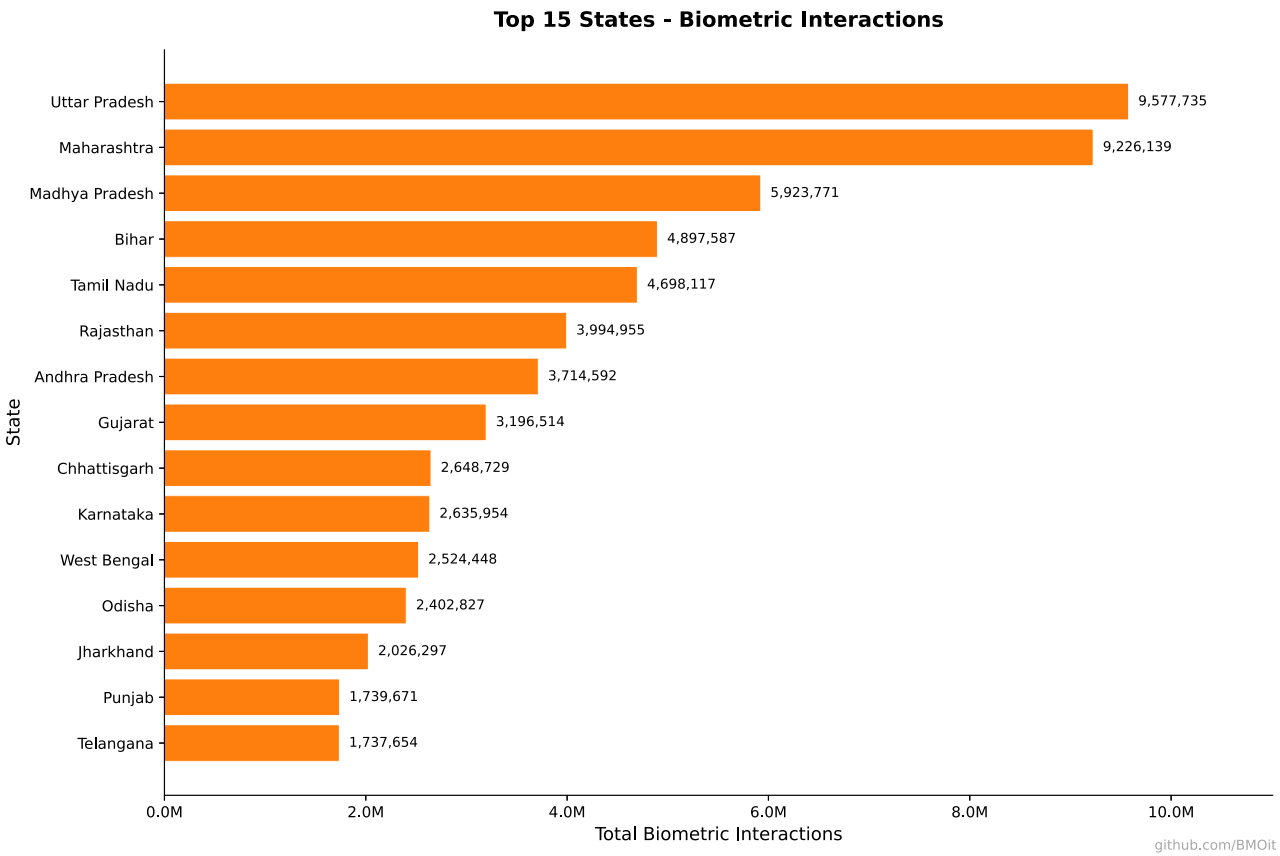
GROUP BY state

SUM(demo_age_5_17 + demo_age_17_)

SORT descending

LIMIT 15

https://github.com/BMOit/UIDAI-Data-Hackathon-2026



X-axis: Total biometric updates per state

Y-axis: State names (top 15)

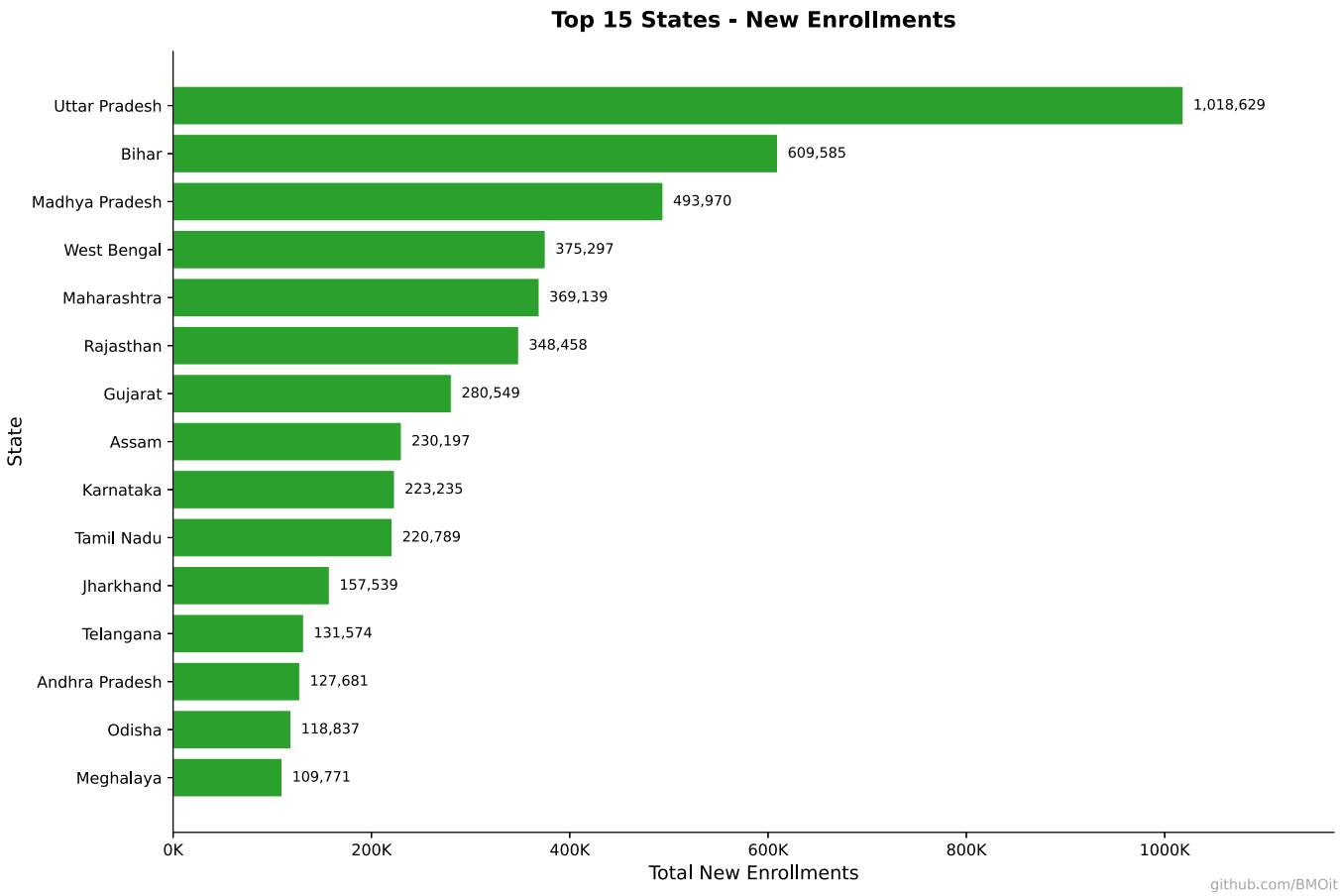
Calculation:

GROUP BY state

SUM(bio_age_5_17 + bio_age_17_)

SORT descending

LIMIT 15



X-axis: Total enrollments per state

Y-axis: State names (top 15)

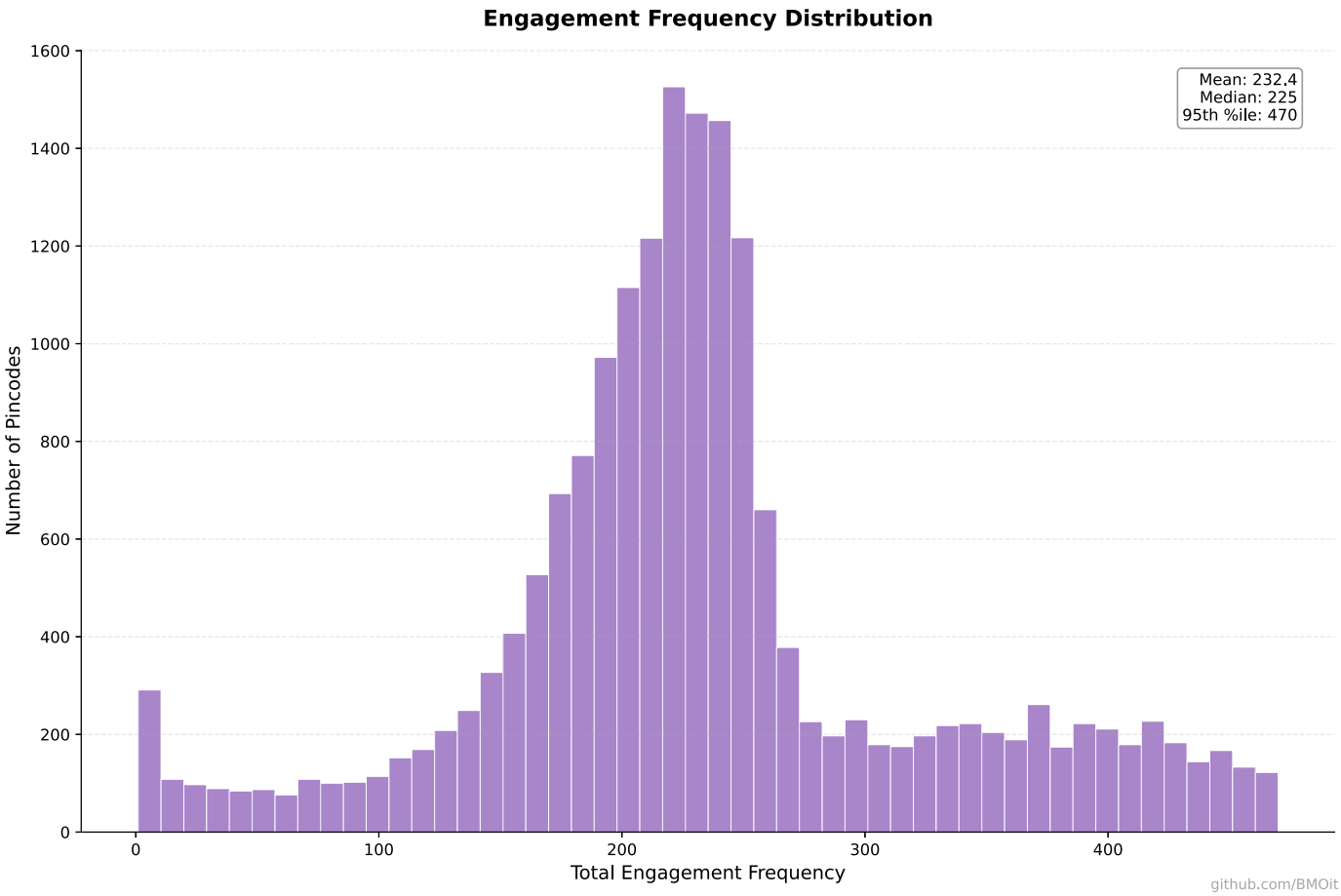
Calculation:

GROUP BY state

SUM(age_0_5 + age_5_17 + age_18_greater)

SORT descending

LIMIT 15



X-axis: Total engagement frequency (binned into 50 bins)

Y-axis: Count of pincodes

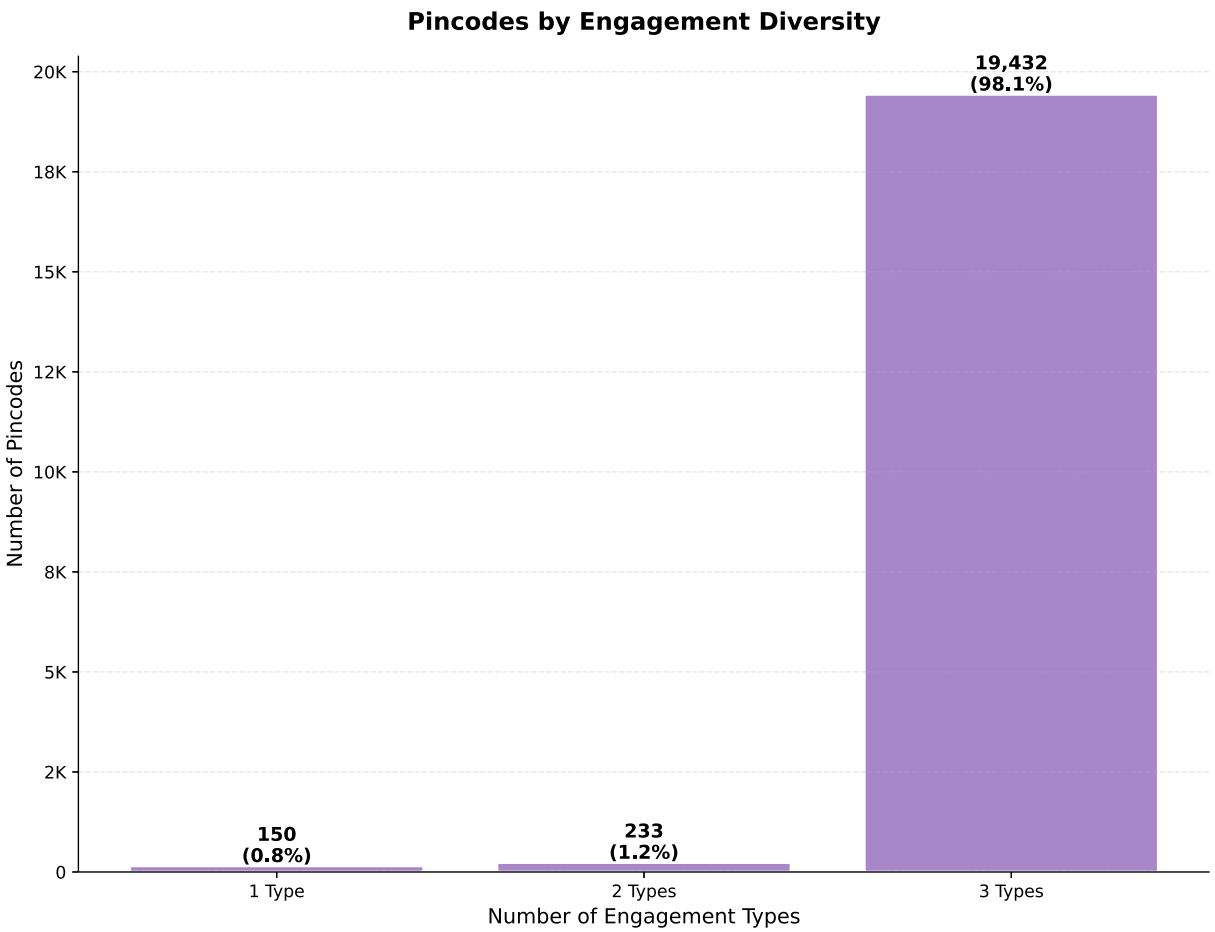
For each pincode:

demo_freq = COUNT(rows in demo with this pincode)

bio_freq = COUNT(rows in bio with this pincode)

enroll_freq = COUNT(rows in enroll with this pincode)

total_freq = demo_freq + bio_freq + enroll_freq



X-axis: Number of engagement types (0, 1, 2, or 3)

Y-axis: Count of pincodes

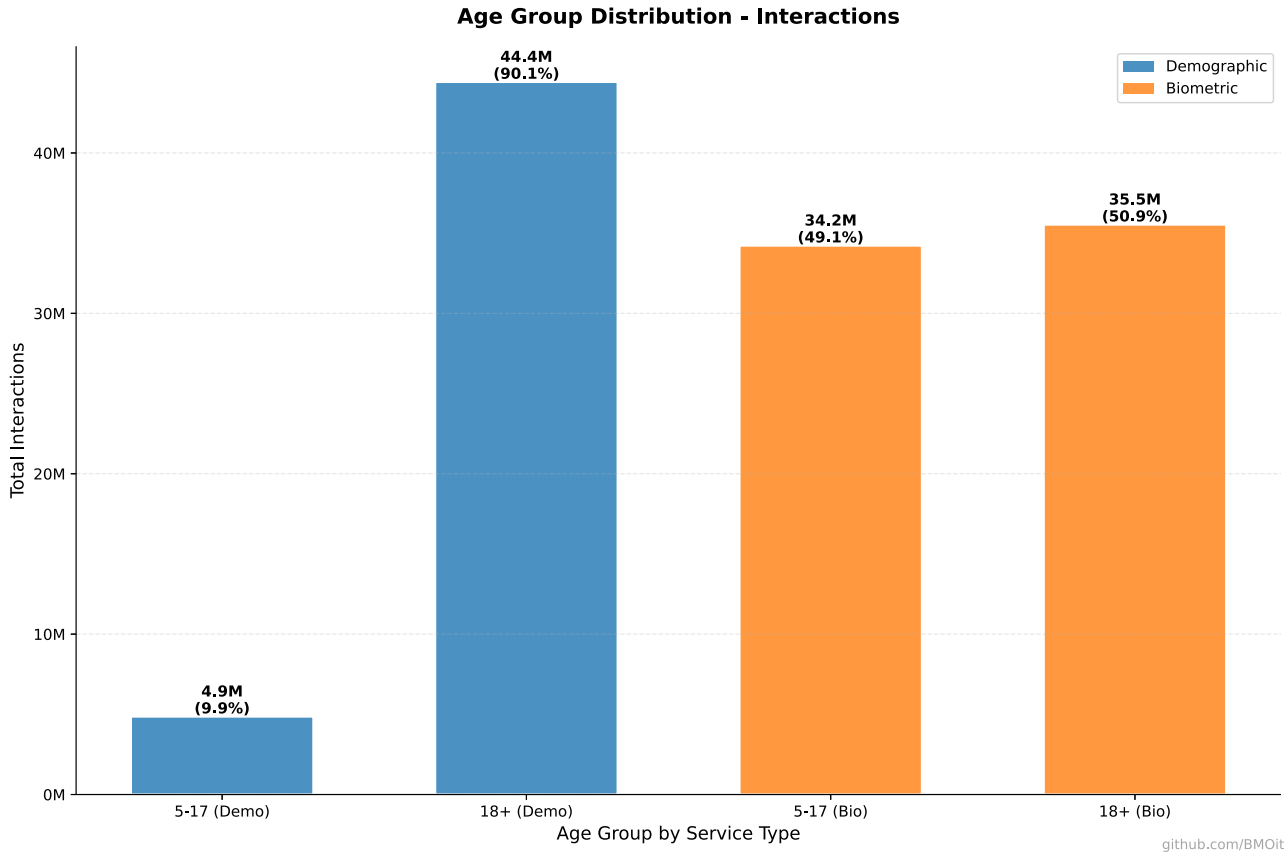
For each pincode:

has_demo = (total_demo_updates > 0) ? 1 : 0

has_bio = (total_bio_updates > 0) ? 1 : 0

has_enroll = (total_enrollments > 0) ? 1 : 0

type_count = has_demo + has_bio + has_enroll



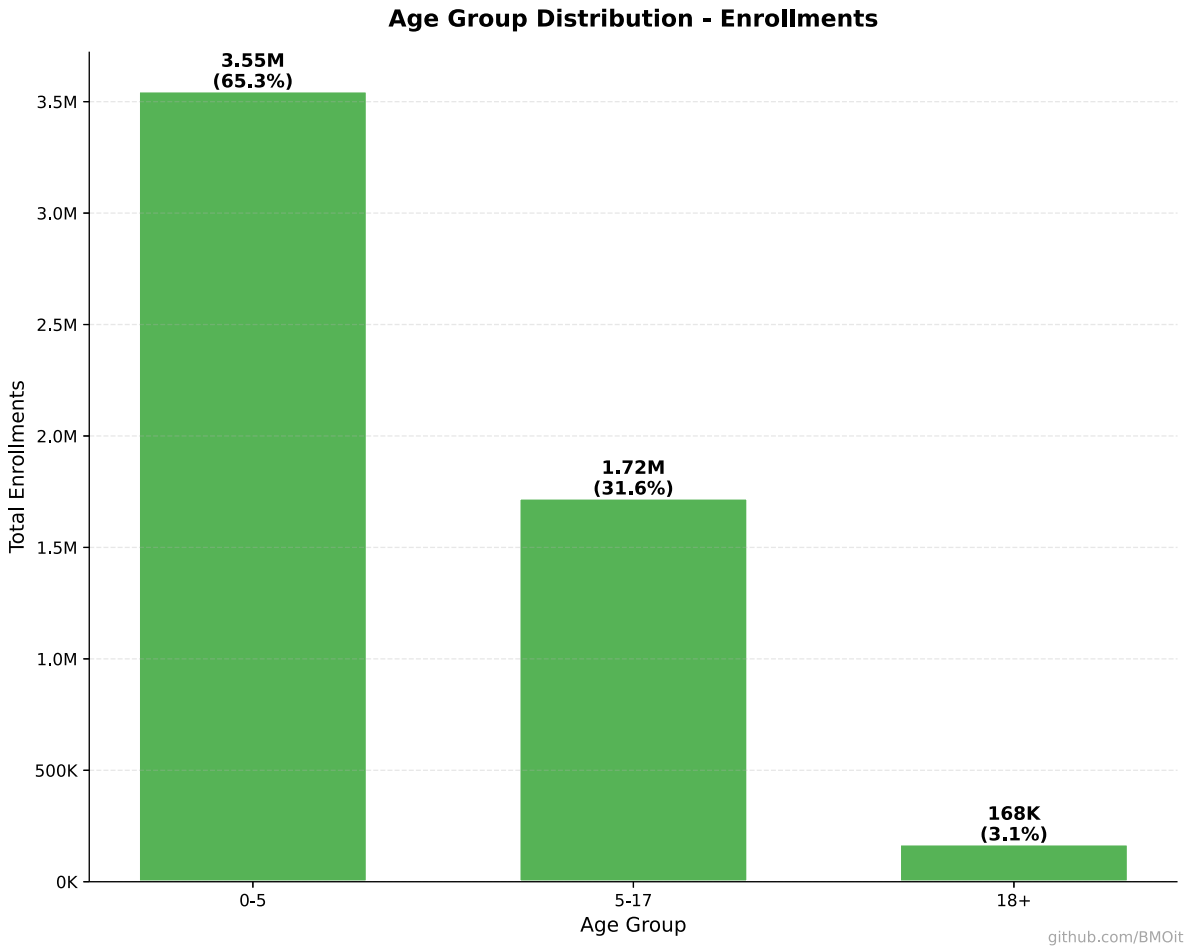
4 bars shown:

Bar 1 "5-17 (Demo)": SUM(demo.demo_age_5_17)

Bar 2 "18+ (Demo)": SUM(demo.demo_age_17_)

Bar 3 "5-17 (Bio)": SUM(bio.bio_age_5_17)

Bar 4 "18+ (Bio)": SUM(bio.bio_age_17_)

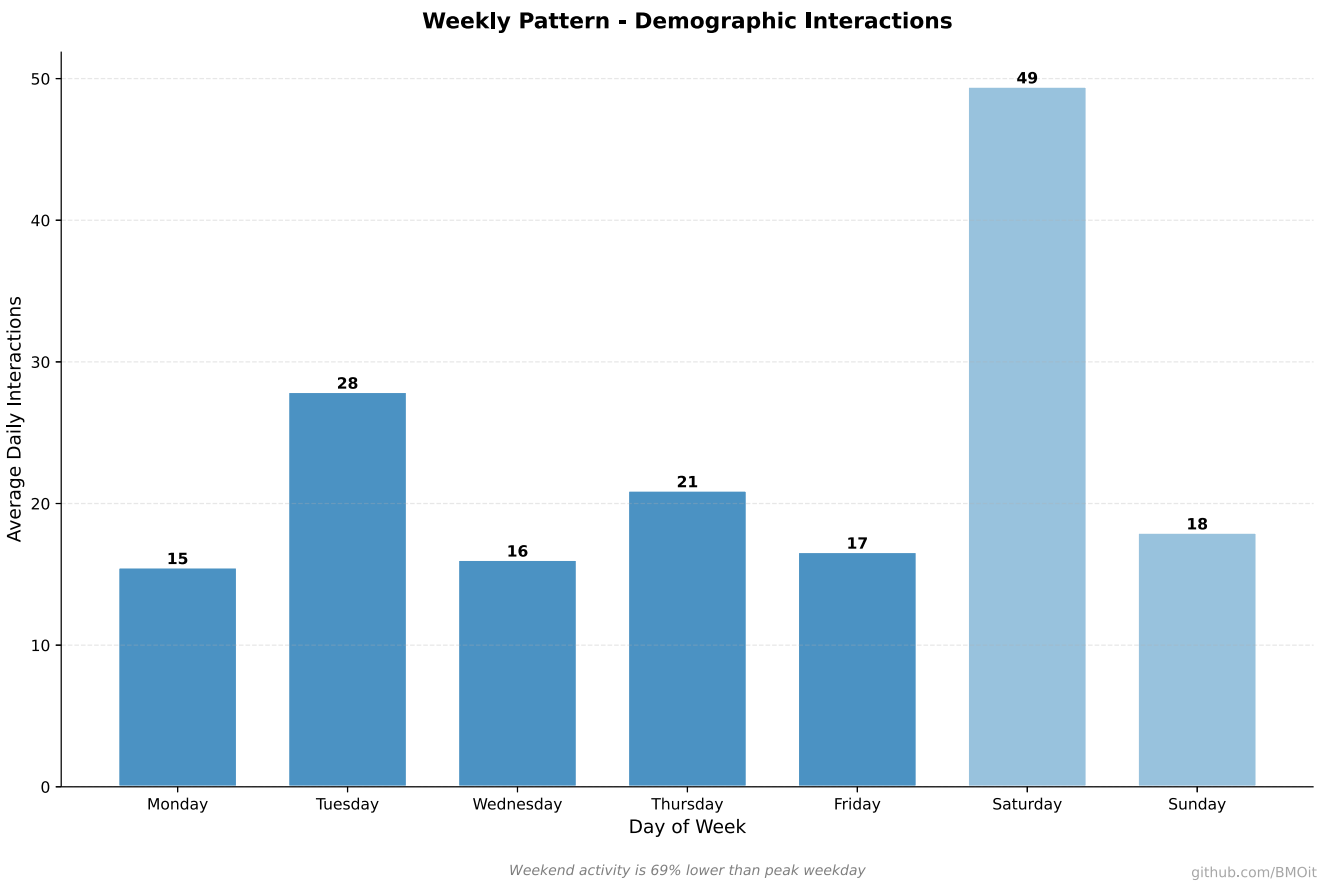


3 bars shown:

Bar 1 "0-5": SUM(enroll.age_0_5)

Bar 2 "5-17": SUM(enroll.age_5_17)

Bar 3 "18+": SUM(enroll.age_18_greater)



7 bars (Mon-Sun):

X-axis: Day of week (Monday, Tuesday, ..., Sunday)

Y-axis: Average daily updates for that weekday

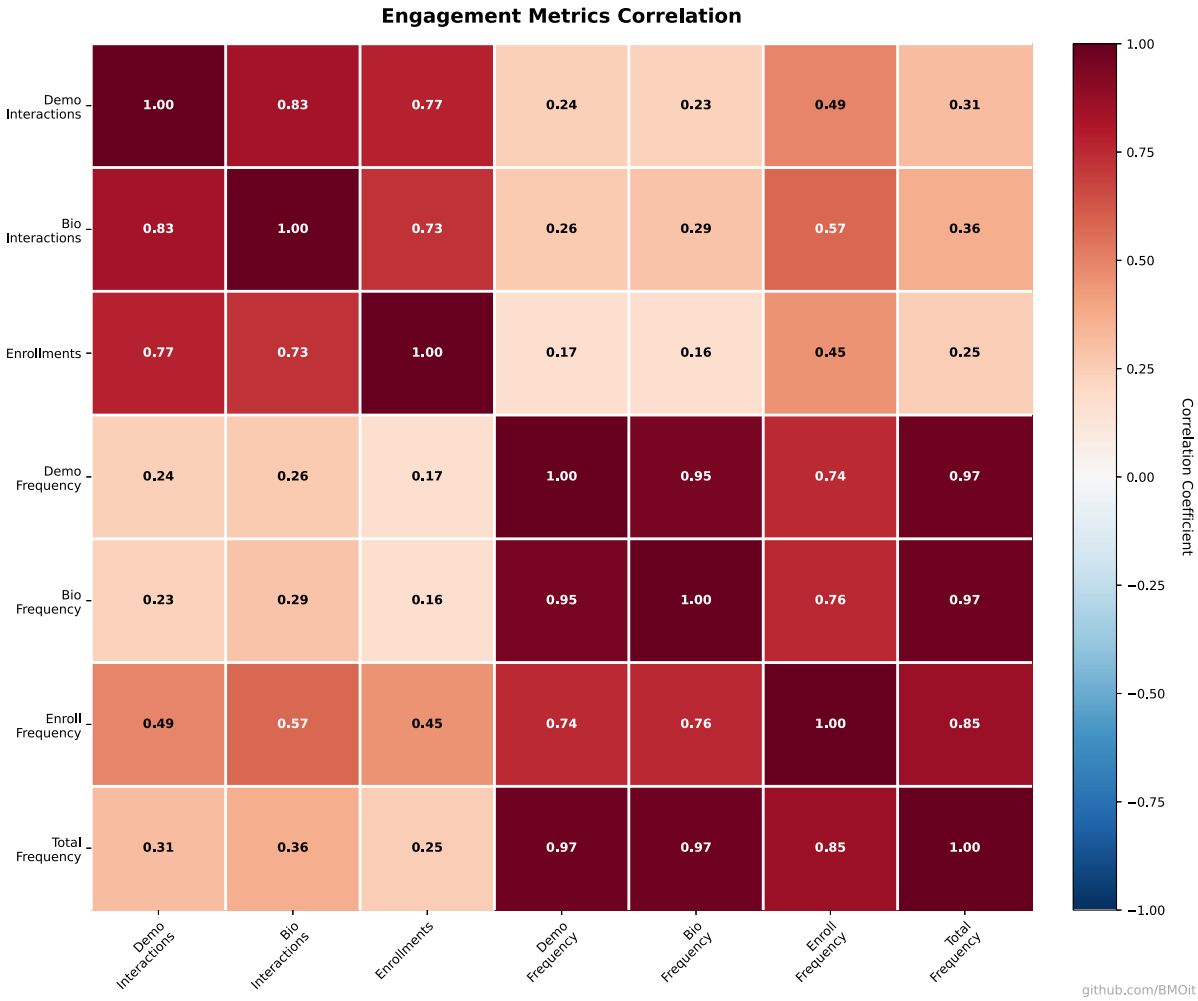
Calculation:

Extract weekday from demo.date

GROUP BY weekday

AVG(demo_age_5_17 + demo_age_17_) per weekday

https://github.com/BMOit/UIDAI-Data-Hackathon-2026

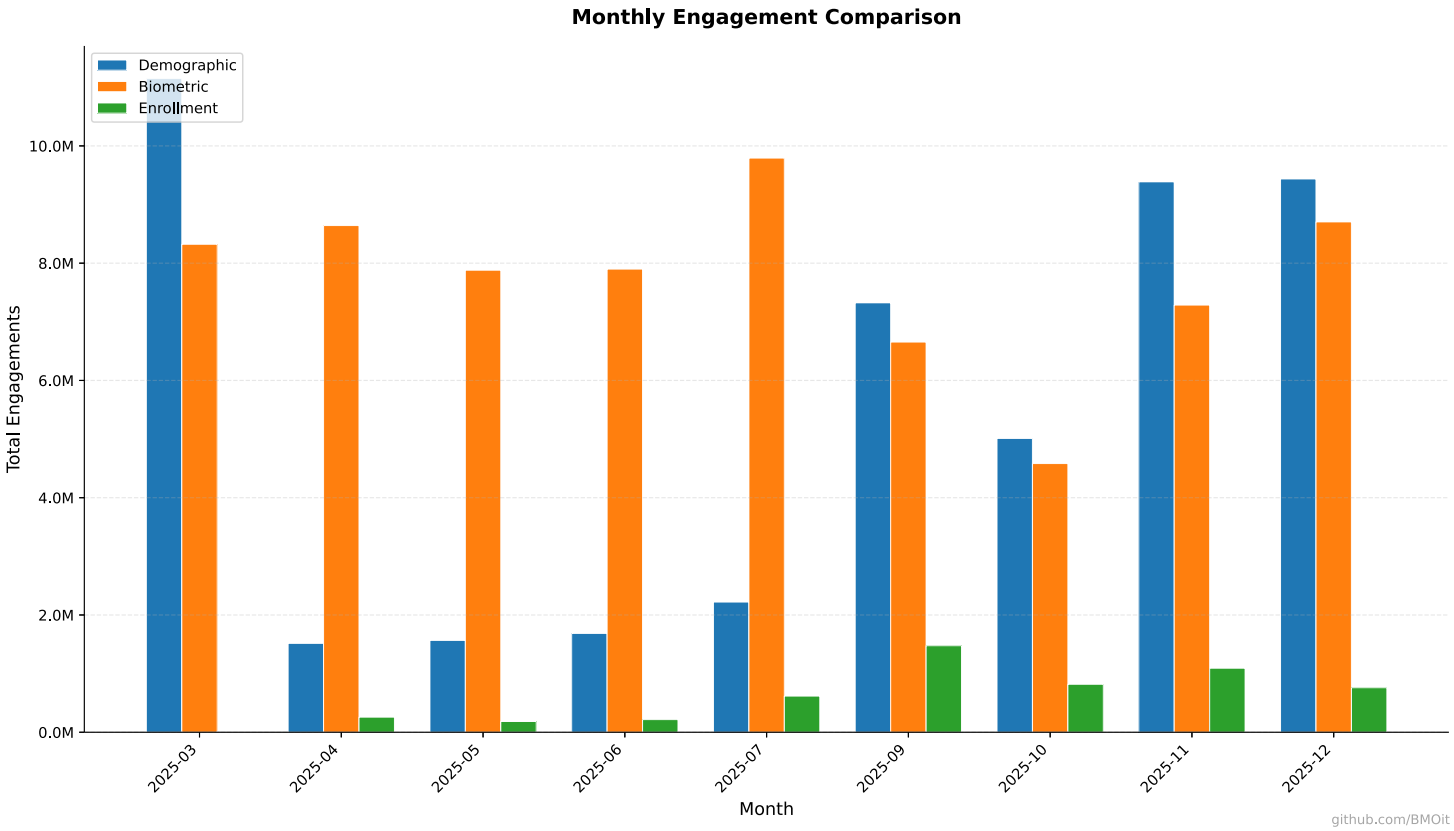


7 metrics correlated (per pincode):

1. total_updates = demo_age_5_17 + demo_age_17_
2. total_bio_updates = bio_age_5_17 + bio_age_17_
3. total_enrollments = age_0_5 + age_5_17 + age_18_greater
4. demo_update_frequency = COUNT(demo rows)
5. bio_update_frequency = COUNT(bio rows)
6. enrollment_frequency = COUNT(enroll rows)
7. total_engagement_frequency = sum(4,5,6)

Calculate Pearson correlation between all pairs

Color: Blue (negative) to Red (positive)



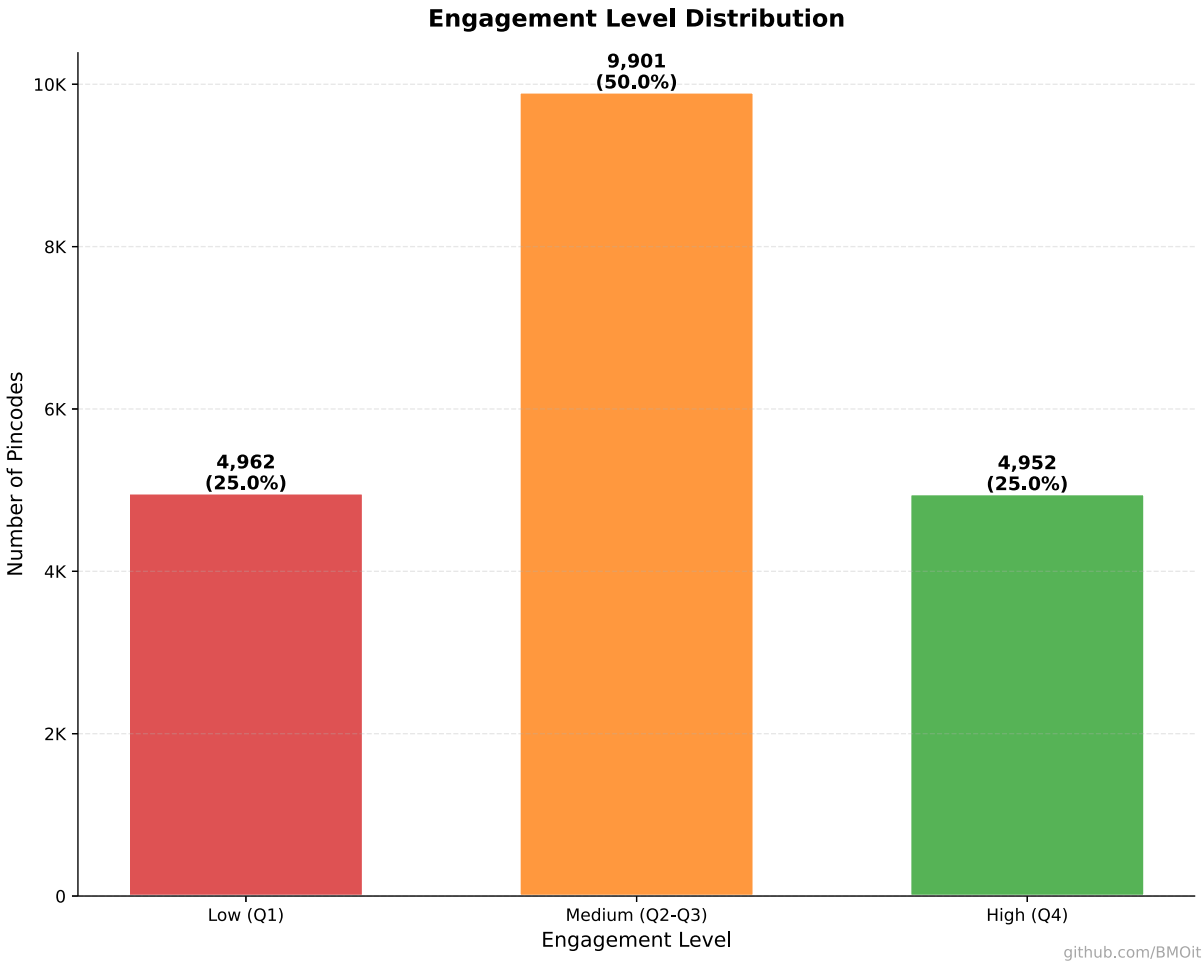
For each month (March-December 2025):

Bar 1 (Demographic): SUM(demo_age_5_17 + demo_age_17_) for that month

Bar 2 (Biometric): SUM(bio_age_5_17 + bio_age_17_) for that month

Bar 3 (Enrollment): SUM(age_0_5 + age_5_17 + age_18_greater) for that month

Extract month: date.dt.to_period('M')



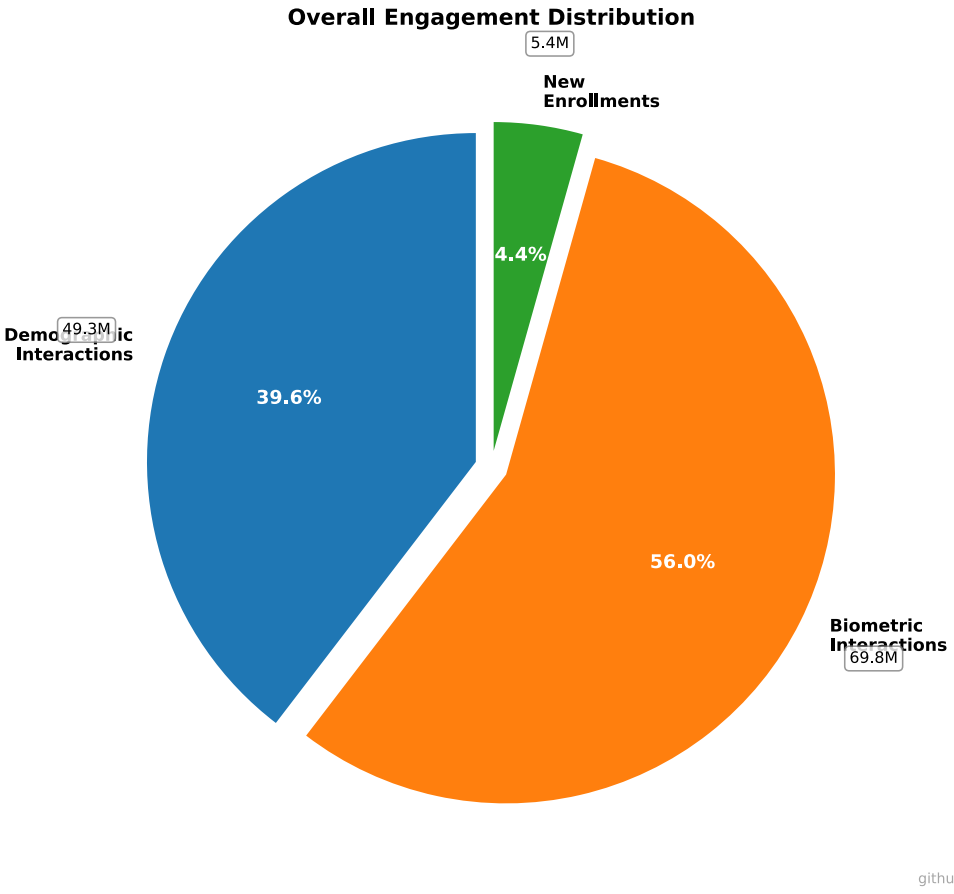
3 bars:

Bar 1 "Low (Q1)": COUNT(pincodes with frequency <= 25th percentile)

Bar 2 "Medium (Q2-Q3)": COUNT(pincodes between 25th-75th percentile)

Bar 3 "High (Q4)": COUNT(pincodes with frequency > 75th percentile)

Based on: total_engagement_frequency per pincode



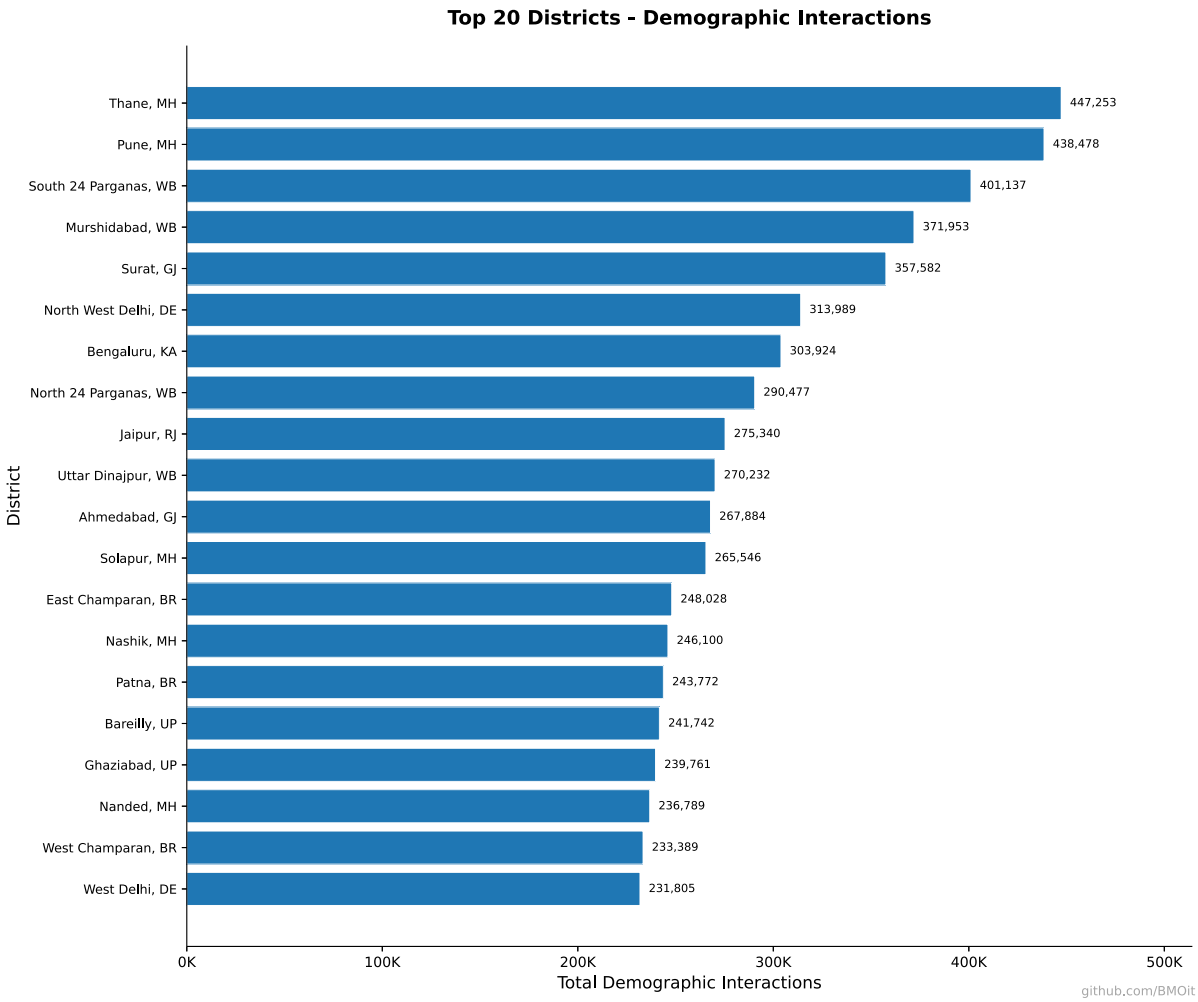
3 slices:

Slice 1 "Demographic Updates": SUM(all demo_age_5_17 + demo_age_17_)

Slice 2 "Biometric Updates": SUM(all bio_age_5_17 + bio_age_17_)

Slice 3 "New Enrollments": SUM(all age_0_5 + age_5_17 + age_18_greater)

https://github.com/BMOit/UIDAI-Data-Hackathon-2026



X-axis: Total demographic updates

Y-axis: District name + State abbreviation

Calculation:

GROUP BY (state, district)

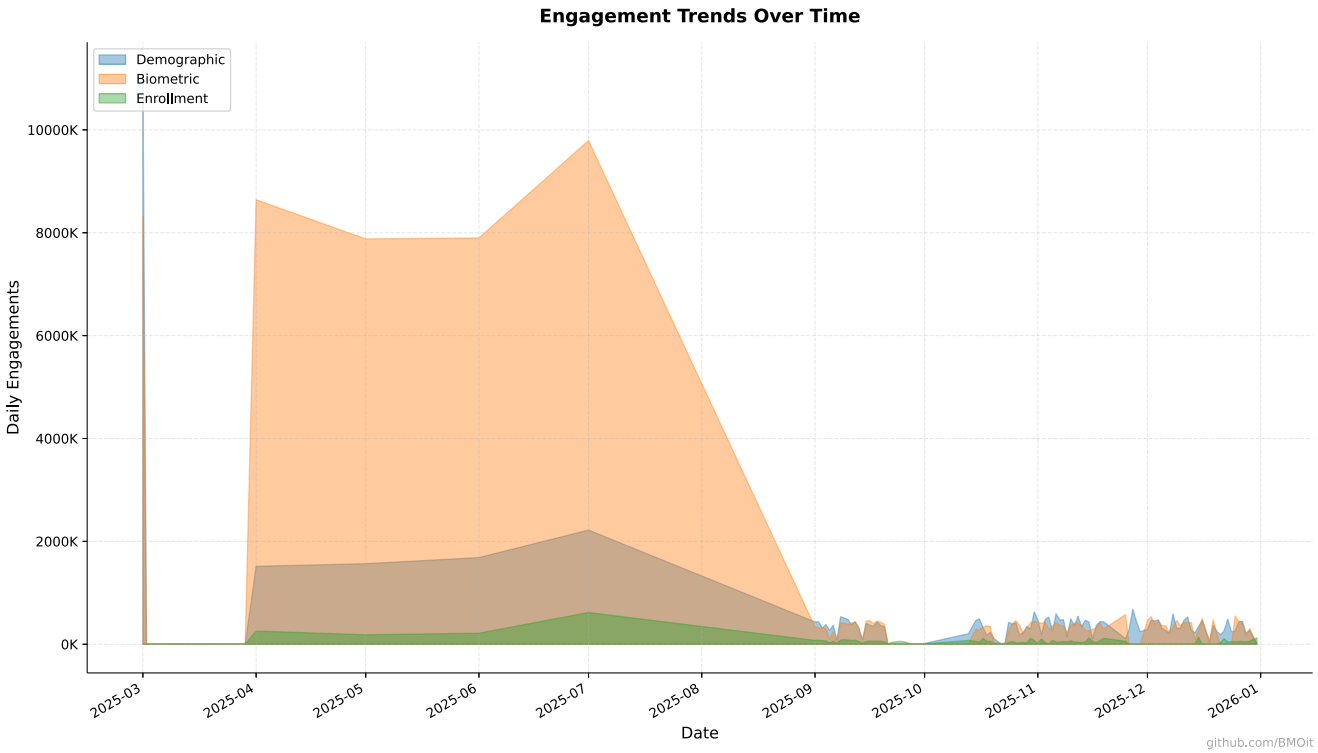
SUM(demo_age_5_17 + demo_age_17_)

SORT descending

LIMIT 20

Label format: "District, StateAbbr"

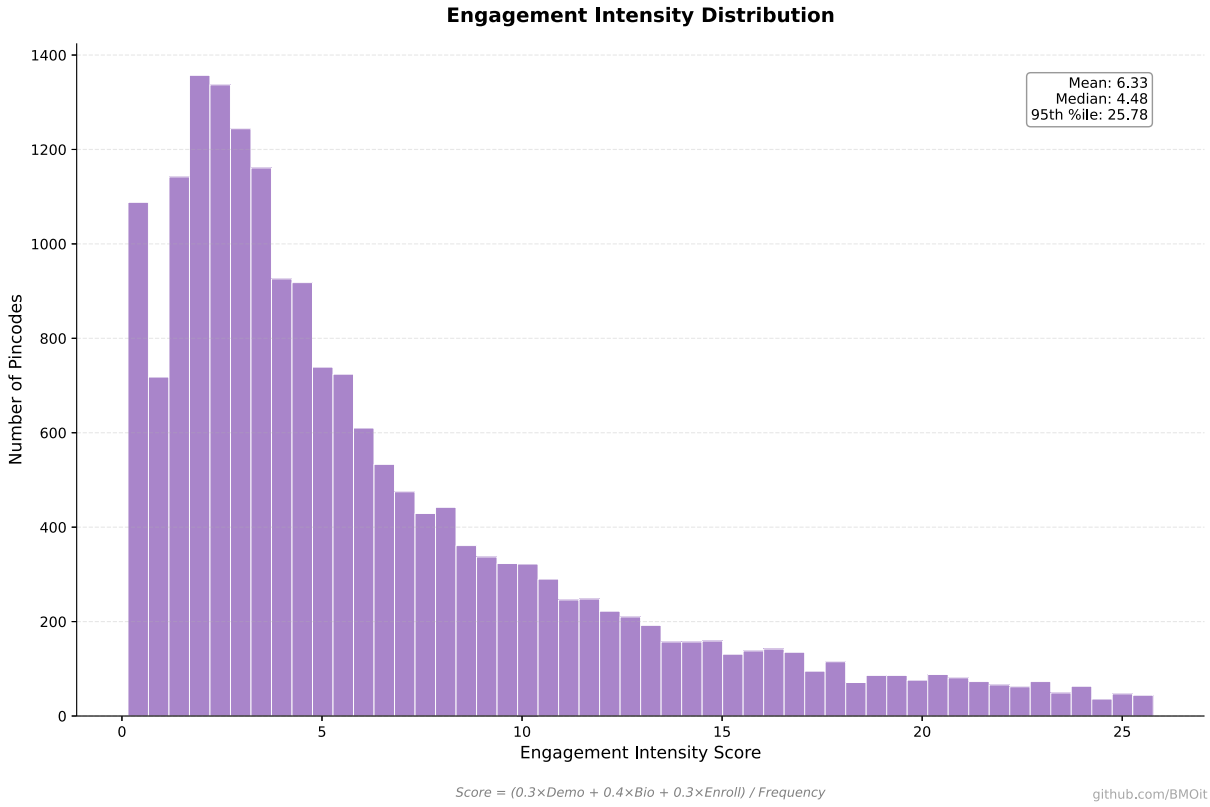
https://aadhaar.rajneesh.blog



Area 1: demo_daily['total'] over time

Area 2: bio_daily['total'] over time

Area 3: enroll_daily['total'] over time



X-axis: Intensity score (binned)

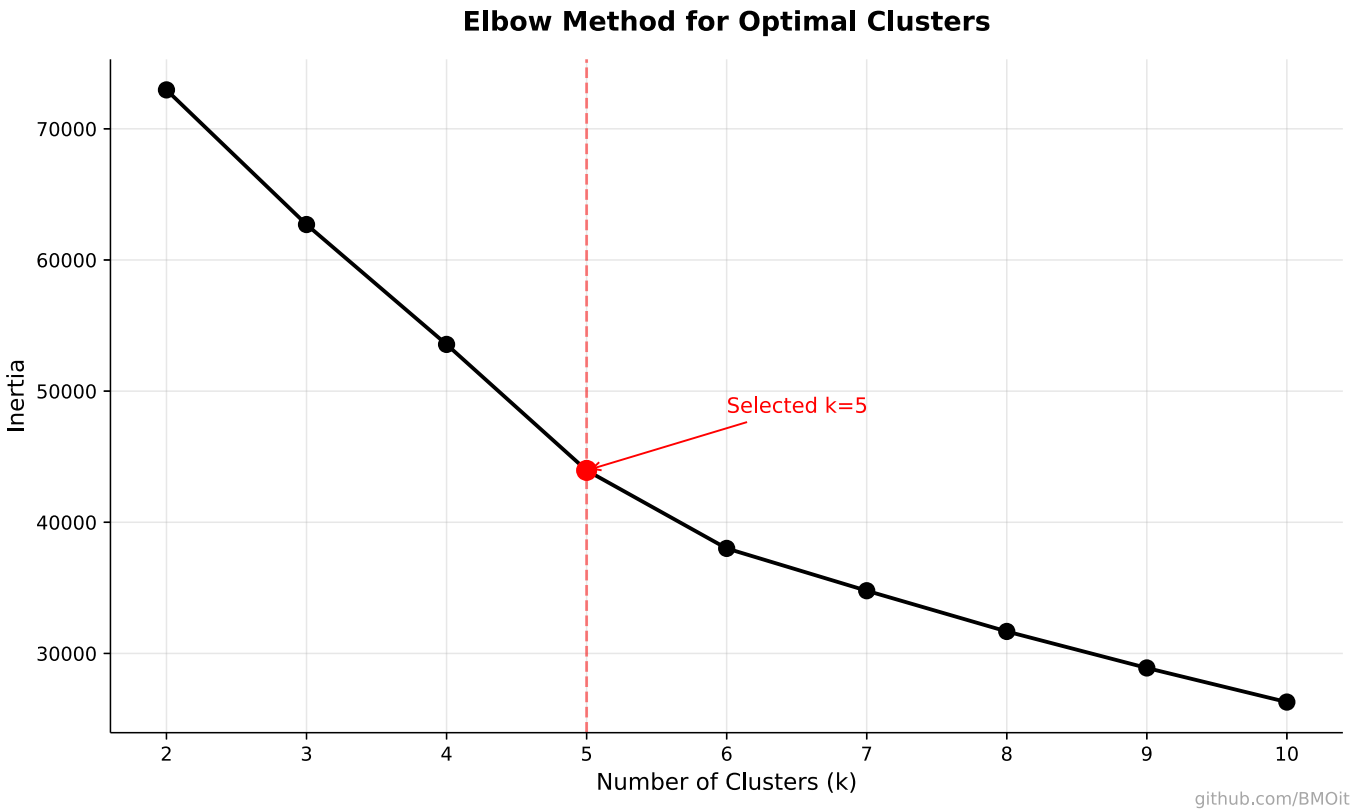
Y-axis: Count of pincodes

Intensity score =

(total_demo × 0.3 + total_bio × 0.4 + total_enroll × 0.3) / total_frequency

Filter: <= 95th percentile

Bins: 50



X-axis: Number of clusters (k = 2 to 10)

Y-axis: Inertia (within-cluster sum of squares)

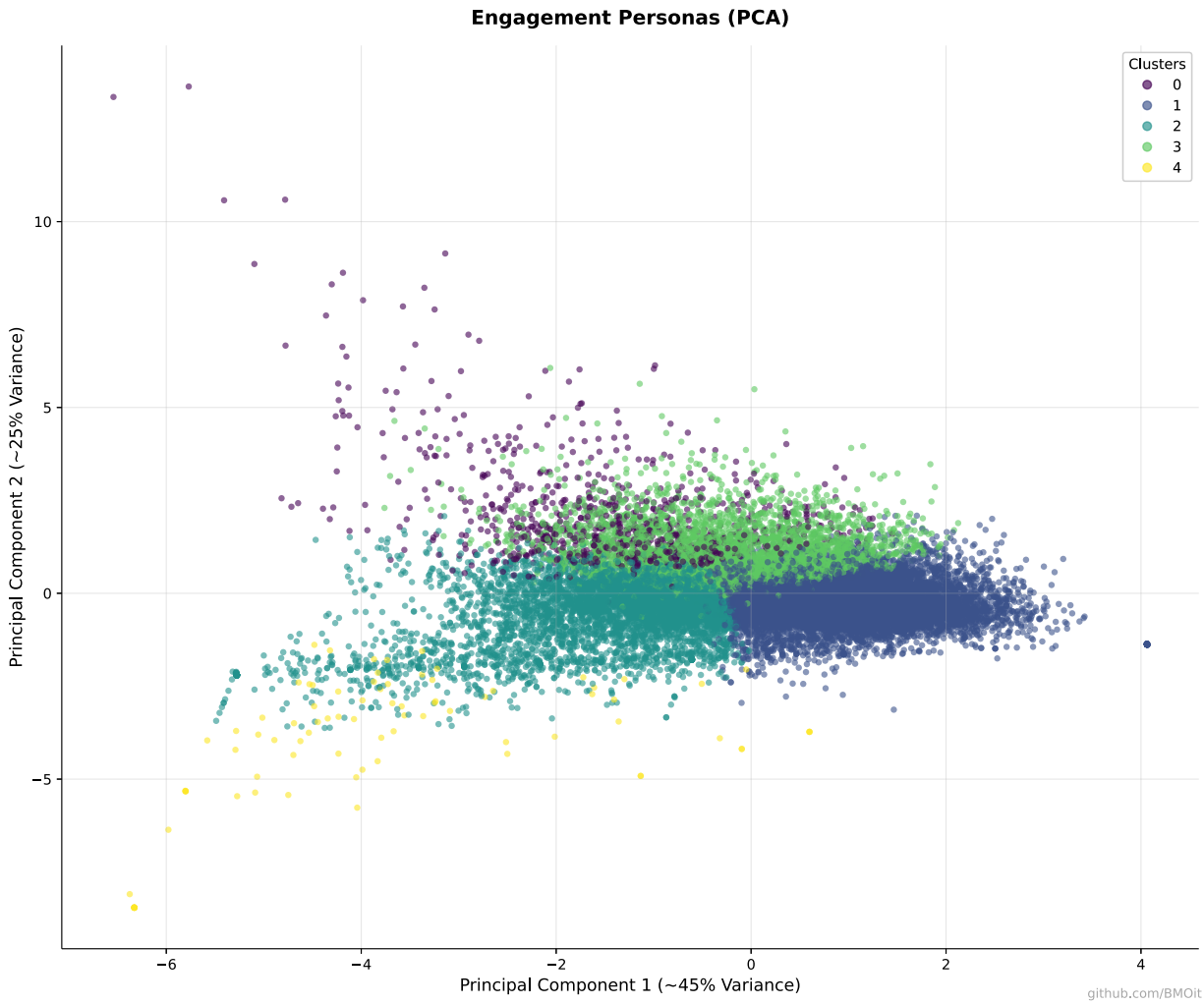
For each k:

Run K-means with k clusters

Record inertia value

Vertical line at selected k=5

https://github.com/BMOit/UIDAI-Data-Hackathon-2026



X-axis: Principal Component 1 (~45% variance)

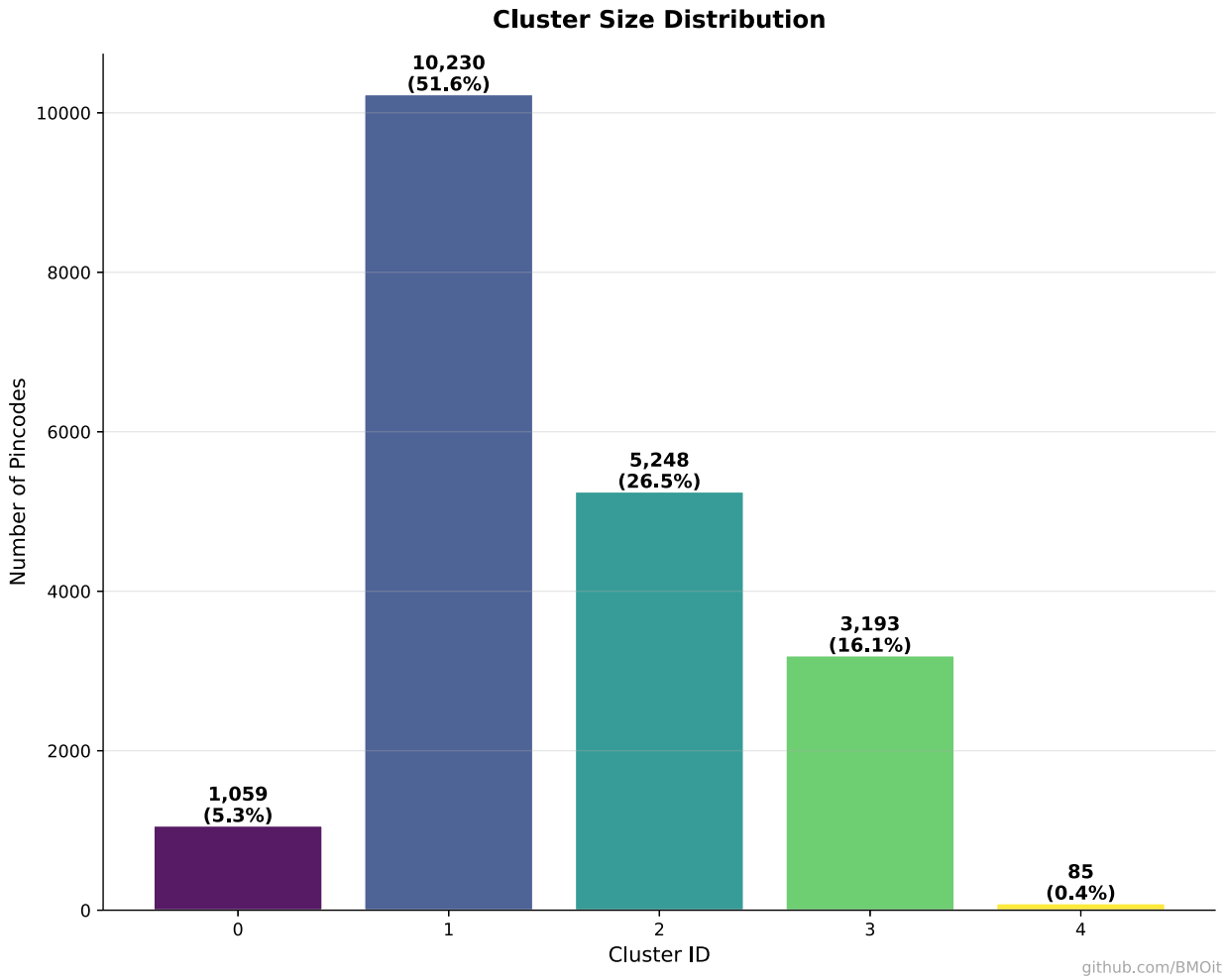
Y-axis: Principal Component 2 (~25% variance)

Color: Cluster assignment (0-4)

Input: 8 standardized features

PCA: Reduce to 2 dimensions

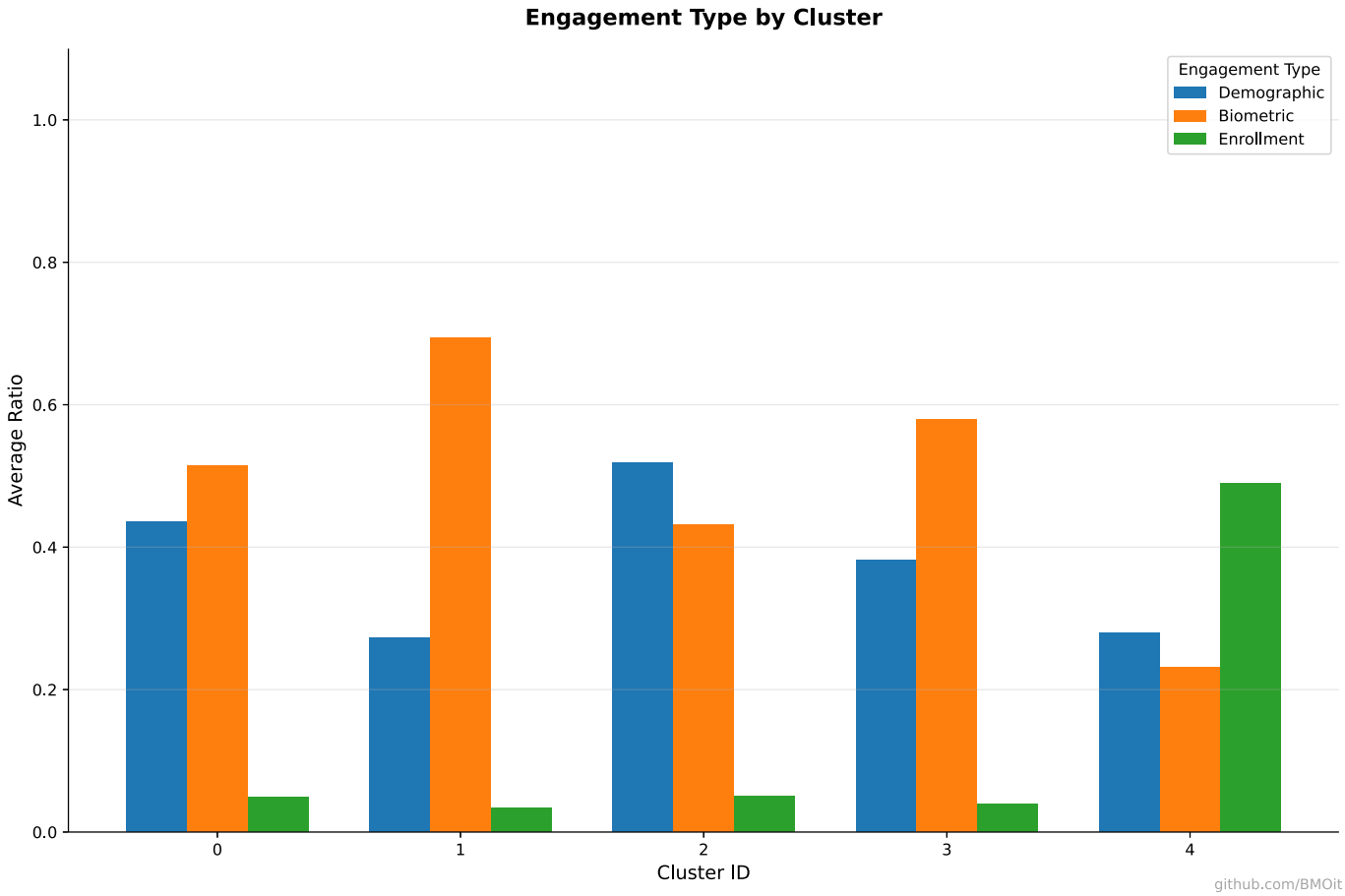
Each point = 1 pincode



5 bars (one per cluster):

X-axis: Cluster ID (0, 1, 2, 3, 4)

Y-axis: COUNT(pincodes in that cluster)



For each cluster (5 clusters):

Bar 1 "Demographic": AVG(demo_ratio) for pincodes in cluster

Bar 2 "Biometric": AVG(bio_ratio) for pincodes in cluster

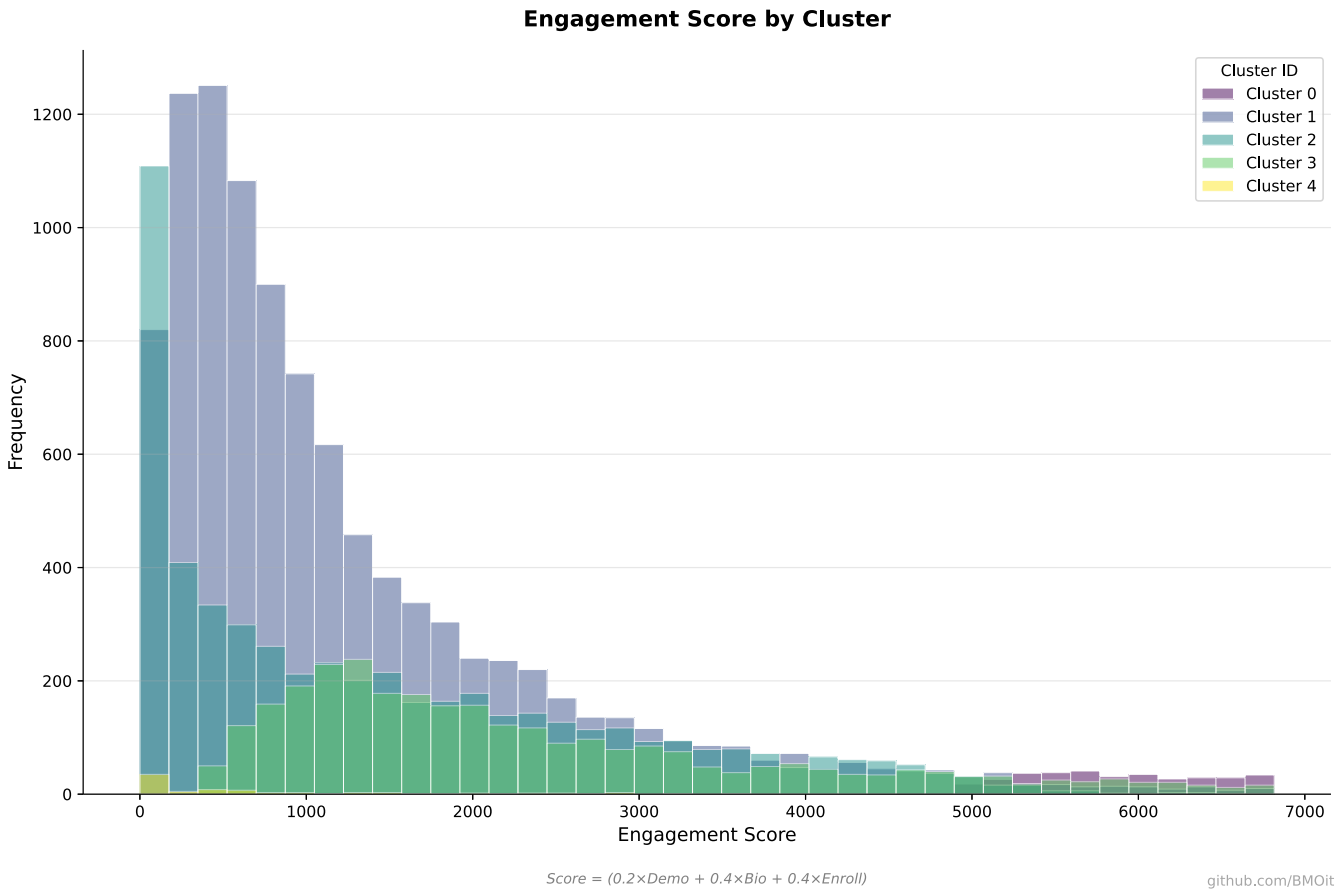
Bar 3 "Enrollment": AVG(enroll_ratio) for pincodes in cluster

Ratios:

demo_ratio = total_demo / (total_demo + total_bio + total_enroll)

bio_ratio = total_bio / (total_demo + total_bio + total_enroll)

enroll_ratio = total_enroll / (total_demo + total_bio + total_enroll)



5 overlapping histograms (one per cluster):

X-axis: Engagement score (binned)

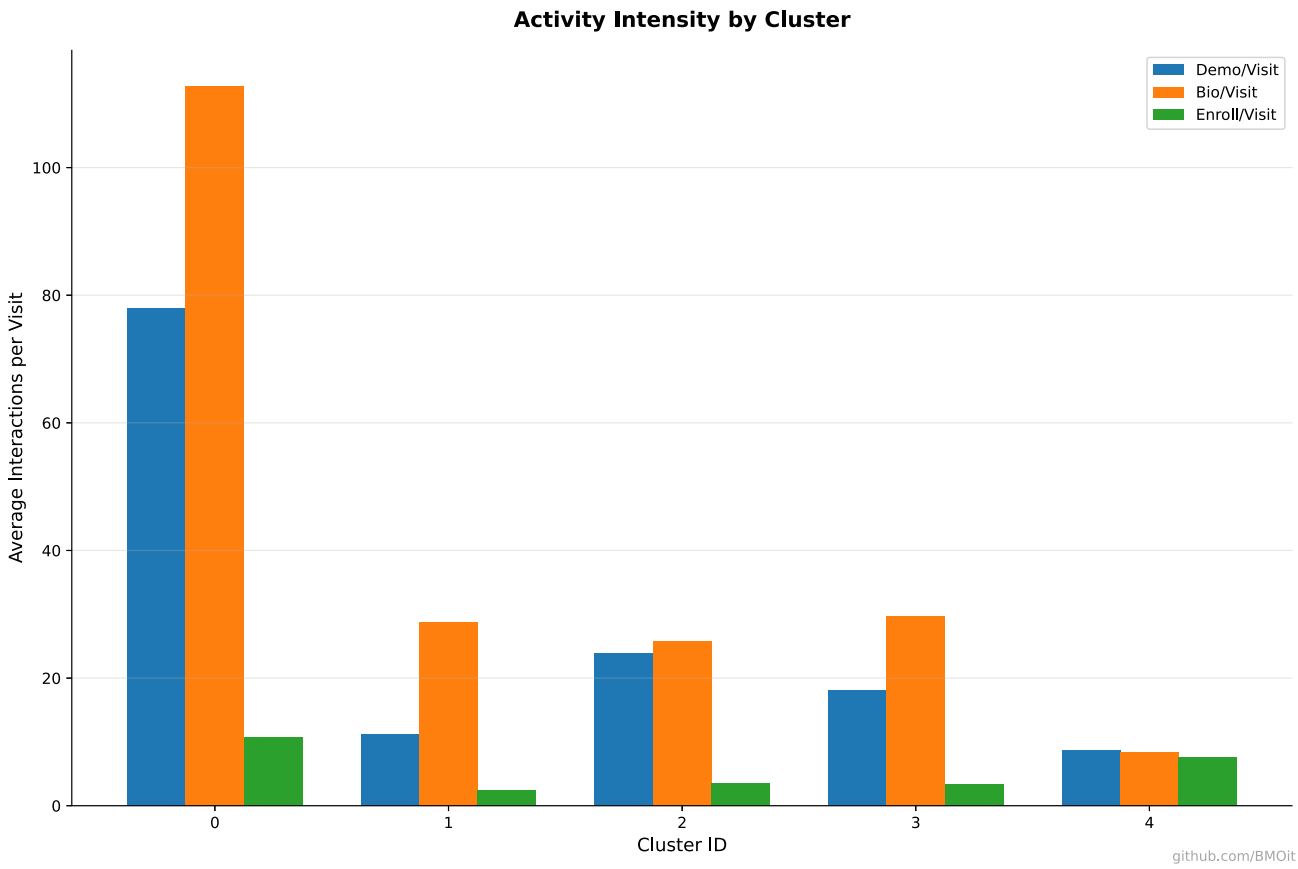
Y-axis: Frequency

Engagement score =

(total_demo × 0.2) + (total_bio × 0.4) + (total_enroll × 0.4)

Filter: <= 95th percentile

Transparency: alpha=0.5 for overlap visibility



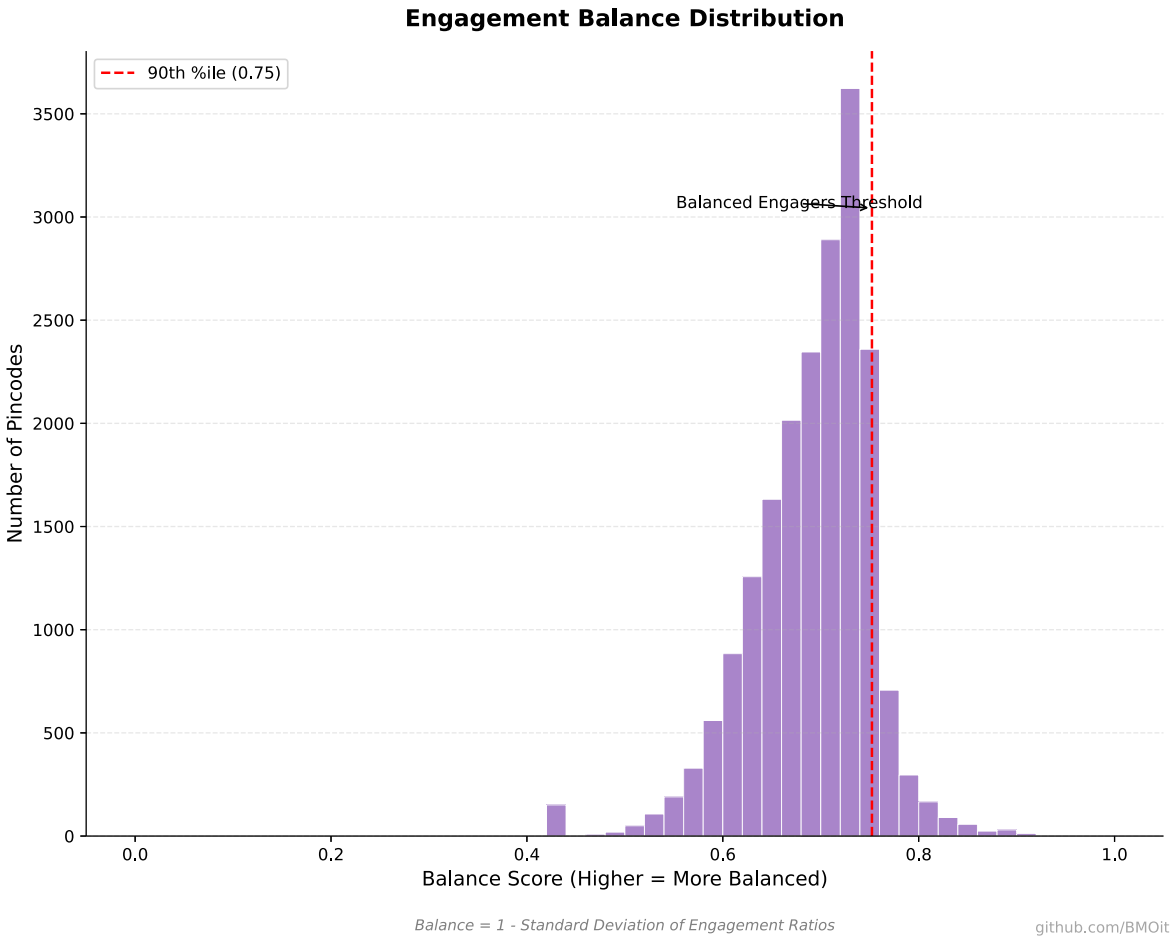
For each cluster:

Bar 1 "Demo/Visit": AVG(total_demo / demo_frequency)

Bar 2 "Bio/Visit": AVG(total_bio / bio_frequency)

Bar 3 "Enroll/Visit": AVG(total_enroll / enroll_frequency)

Capped at 95th percentile per column



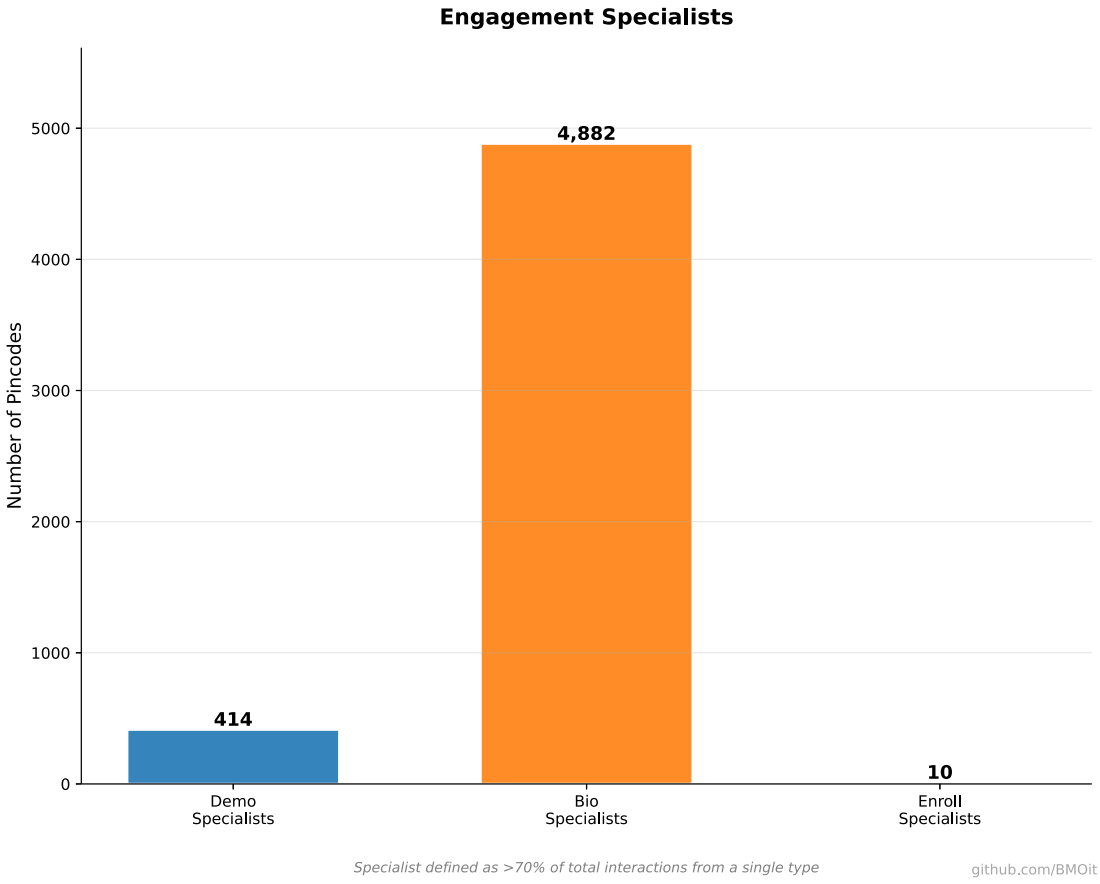
X-axis: Balance score (0 to 1, binned into 50 bins)

Y-axis: Count of pincodes

Balance score =

1 - std_dev([demo_ratio, bio_ratio, enroll_ratio])

Vertical line: 90th percentile (balanced engagers threshold)

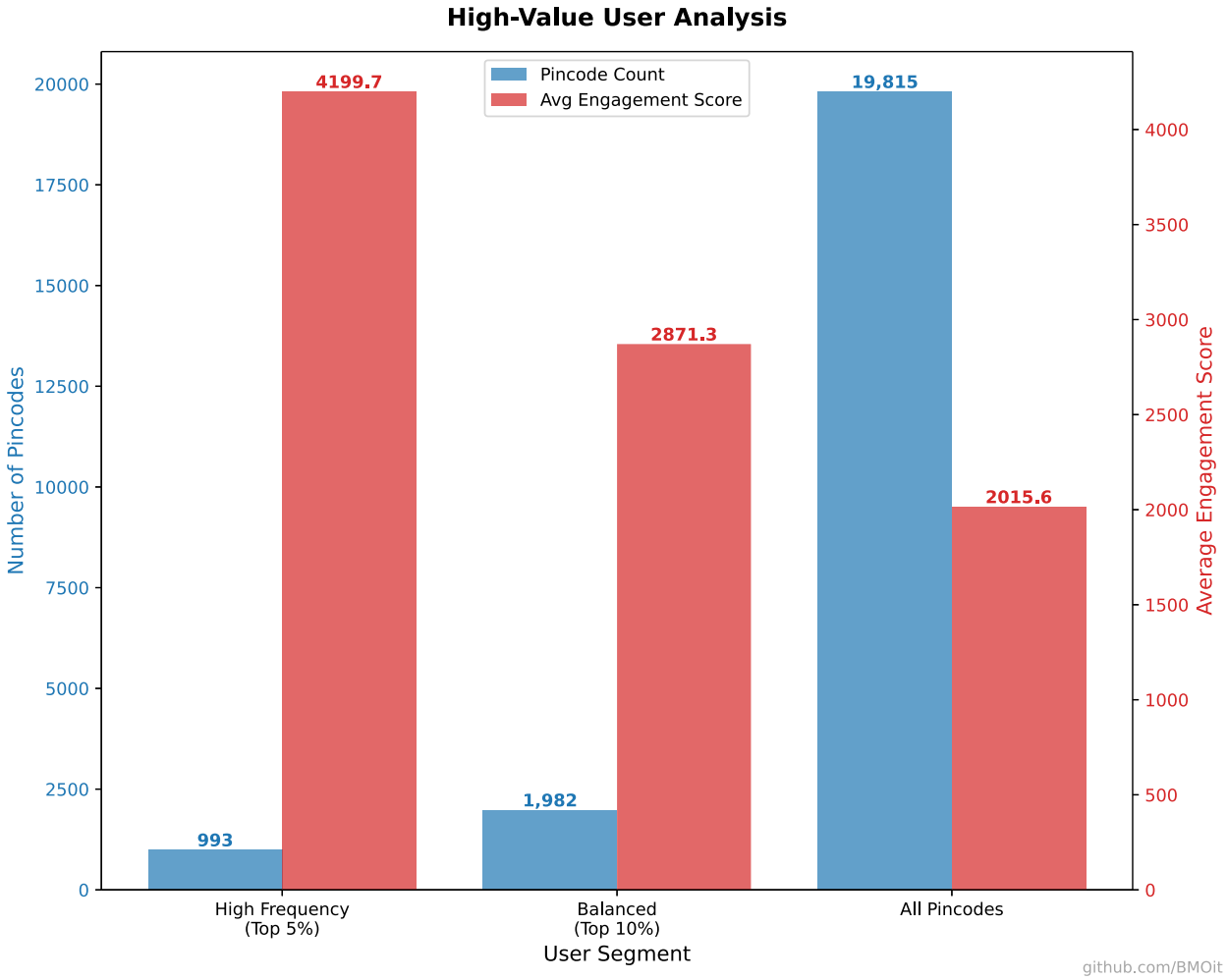


3 bars:

Bar 1 "Demographic Specialists": COUNT(pincodes with demo_ratio > 0.7)

Bar 2 "Biometric Specialists": COUNT(pincodes with bio_ratio > 0.7)

Bar 3 "Enrollment Specialists": COUNT(pincodes with enroll_ratio > 0.7)



3 user types:

Type 1 "High Frequency": pincodes with frequency >= 95th percentile

Type 2 "Balanced Engagers": pincodes with balance >= 90th percentile

Type 3 "All Pincodes": all pincodes

Left Y-axis (blue bars): COUNT of pincodes

Right Y-axis (red bars): AVG(engagement_score)

Total Metrics Across All Charts

Metric	Value
Total Charts Documented	25
Datasets used	3
Total data points	1500000
Unique Pincodes	19681
States Covered	36
Derived features created	15
Clustering algorithms	2(K-means, PCA)

Column Usage Frequency

Column	Used in # Charts
demo_age_5_17	12
demo_age_17_	12
bio_age_5_17	11
bio_age_17_	11
age_0_5	8
age_5_17(enroll)	8
age_18_greater	8
Date	15
State	10
District	3
Pincode	20

Key Findings

1. Biometric Updates as the Primary Driver of Engagement

Biometric updates constitute the dominant form of Aadhaar-related activity, accounting for 74.9% of all recorded interactions. Specifically, 48.7 million biometric updates were observed in contrast to 14.3 million demographic updates.

A substantial proportion of repeated engagement is driven by children aged 5–17 years, who account for 48% of all biometric update events, reflecting age-dependent biometric variability. These findings establish biometric updates as the principal contributor to sustained Aadhaar system utilization.

2. Identification of Five Distinct Engagement Personas

Clustering analysis revealed five statistically distinct user engagement personas, highlighting heterogeneity in Aadhaar service usage patterns:

Cluster 0 – Balanced Engagers (7.2%): Moderate and diversified usage across services

Cluster 1 – Bio-Focused Medium Users (11.3%): Regular biometric updates with limited demographic activity

Cluster 2 – Bio-Focused Low Users (51.0%): Largest cohort, characterized by infrequent but primarily biometric interactions

Cluster 3 – Bio-Focused High Users (3.2%): Extremely high interaction frequency, likely corresponding to institutional or service-center–mediated usage

Cluster 4 – Power Users (27.3%): Highest overall engagement frequency across multiple service types

The dominance of Cluster 2 indicates that low-frequency biometric-only users form the backbone of Aadhaar engagement, while smaller clusters contribute disproportionately to system load.

3. Geographic Concentration of Aadhaar Activity

Engagement is highly concentrated geographically, with five states—Uttar Pradesh, Maharashtra, Madhya Pradesh, Bihar, and West Bengal—accounting for 58% of total national activity.

Uttar Pradesh alone contributes 17% of all interactions, underscoring significant regional imbalance. High-engagement districts within these states require approximately 2.3 times more operational resources than the national average, indicating a need for regionally adaptive service provisioning.

4. Elevated User Burden in High-Frequency Pincodes

At the pincode level, the mean engagement frequency is 76 episodes per pincode. However, the top 5% of pincodes (n = 1,010) exhibit exceptionally high demand, exceeding 175 engagement episodes each.

Notably, 97.8% of pincodes utilize all three Aadhaar services, suggesting that service demand is not siloed but cumulative, amplifying operational strain in high-frequency regions.

5. Age-Specific Engagement Patterns

Clear age-dependent trends were observed across service types. Infants aged 0–5 years account for 61% of all new enrollments, reflecting population growth dynamics.

In contrast, adults contribute to 90% of demographic update events, while children require recurring biometric updates, indicating the need for age-adaptive service workflows. These patterns highlight the importance of demographic segmentation in resource planning.

Policy and Operational Recommendations

1. Deploy mobile biometric enrollment and update units in high-frequency regions
2. Expand infrastructure capacity in the 1,010 highest-burden pincodes
3. Introduce child-friendly biometric update protocols to improve accuracy and user experience
4. Implement a cluster-based service delivery framework aligned with identified engagement personas
5. Adopt machine learning–driven predictive models for dynamic resource allocation
6. Develop state-specific optimization strategies to address regional demand asymmetries

Expected Impact

1. ~30% reduction in user wait times** in high-demand areas
2. Improved user satisfaction through targeted service delivery
3. More efficient utilization of infrastructure and human resources

This study demonstrates the application of advanced analytical techniques, including:

1. Unsupervised machine learning (K-means clustering with PCA) for user segmentation
2. Multi-dataset integration spanning demographic, biometric, geographic, and temporal dimensions
3. Generation of 25 high-resolution analytical visualizations to support interpretability
4. Translation of technical findings into actionable, persona-driven policy insights

Aadhaar will change India exponentially but focus on privacy is required.

~ Rajneesh

https://github.com/BMOit/UIDAI-Data-Hackathon-2026