

Unlocking Societal Trends in

Aadhaar Enrolment and Updates

Comprehensive Data Analysis Report

Submission Date: January 15, 2026

Dataset: 5M+ Records Analyzed

Coverage: 48 States | 922 Districts

Period: March - December 2025

1. PROBLEM STATEMENT AND APPROACH

PROBLEM STATEMENT

The Unique Identification Authority of India (UIDAI) manages the world's largest biometric identification system with over 1.3 billion Aadhaar enrollments. Understanding patterns, trends, and anomalies in enrollment and update data is critical for:

- Optimizing resource allocation across 36 states/UTs and 700+ districts
- Identifying temporal patterns for capacity planning and workforce management
- Detecting operational inefficiencies and system bottlenecks
- Improving service accessibility for diverse demographic segments
- Forecasting future demand for proactive infrastructure planning
- Enhancing data quality and completeness across the ecosystem

ANALYTICAL APPROACH

Our comprehensive analytical framework consists of six integrated stages:

1. EXPLORATORY DATA ANALYSIS (EDA)

- Understand data distribution, missing values, and basic statistics
- Identify initial patterns across geographic, temporal, and demographic dimensions
- Profile data quality and completeness metrics

2. TEMPORAL ANALYSIS

- Examine daily, weekly, and monthly patterns using time series decomposition
- Detect seasonality and cyclical behaviors (working days vs weekends)
- Analyze growth rates and trend components

3. SPATIAL ANALYSIS

- Geographic distribution analysis at state and district levels
- Regional disparity identification using concentration metrics (Gini coefficient)
- Per-capita normalization for fair cross-region comparisons

4. ANOMALY DETECTION

- Statistical outlier detection (IQR, Z-score methods)
- Machine learning-based anomaly detection (Isolation Forest)
- Temporal anomaly detection for unusual spikes/drops

5. PREDICTIVE MODELING

- 30-day forecasting using ARIMA, Exponential Smoothing, and ensemble methods
- Model validation using MAPE, RMSE, and R^2 metrics
- Confidence intervals for risk-aware planning

6. ADVANCED INSIGHTS

- Composite index development (Aadhaar Health Index, Digital Inclusion Score)
- Cohort analysis and user journey mapping
- ROI quantification and implementation roadmap

2. DATASETS USED

DATA SOURCE

All datasets provided by Unique Identification Authority of India (UIDAI) for the Data Hackathon 2026.
Total records analyzed: 5,078,837 records spanning March - December 2025.
Geographic coverage: 48 states/UTs, 922 districts across India.

DATASET 1: AADHAAR ENROLMENT DATA

- Records: 1,006,029 | Coverage: 48 states, 922 districts | Period: Mar-Dec 2025
- Files: 3 CSV files (api_data_aadhar_enrolment_*.csv)

COLUMN SCHEMA:

- date (DD-MM-YYYY): Enrolment transaction date
- registrar: Registrar organization name conducting enrolment
- private_agency: Private agency partner (if applicable)
- state: Indian state/UT name
- district: District name within state
- sub_district: Sub-district/tehsil/taluka name
- pincode: 6-digit postal code
- gender: Gender category (Male/Female/Transgender)
- age_0_5: Number of enrollments for age 0-5 years
- age_5_17: Number of enrollments for age 5-17 years
- age_18_greater: Number of enrollments for age 18+ years
- rejected_0_5: Rejected enrollments for age 0-5 years
- rejected_5_17: Rejected enrollments for age 5-17 years
- rejected_18_greater: Rejected enrollments for age 18+ years

DATASET 2: AADHAAR DEMOGRAPHIC UPDATE DATA

- Records: 2,071,700 | Coverage: 48 states, 922 districts | Period: Mar-Dec 2025
- Files: 5 CSV files (api_data_aadhar_demographic_*.csv)

COLUMN SCHEMA:

- date (DD-MM-YYYY): Update transaction date
- registrar: Registrar organization processing the update
- private_agency: Private agency partner (if applicable)
- state: Indian state/UT name
- district: District name within state
- sub_district: Sub-district/tehsil/taluka name
- pincode: 6-digit postal code
- gender: Gender category
- name_update: Number of name change requests
- dob_update: Number of date of birth corrections
- address_update: Number of address change requests (migration indicator)
- gender_update: Number of gender corrections
- mobile_update: Number of mobile number updates
- email_update: Number of email address updates

DATASET 3: AADHAAR BIOMETRIC UPDATE DATA

- Records: 1,861,108 | Coverage: 48 states, 922 districts | Period: Mar-Dec 2025
- Files: 4 CSV files (api_data_aadhar_biometric_*.csv)

COLUMN SCHEMA:

- date (DD-MM-YYYY): Update transaction date
- registrar: Registrar organization processing biometric update
- private_agency: Private agency partner (if applicable)
- state: Indian state/UT name
- district: District name within state
- sub_district: Sub-district/tehsil/taluka name
- pincode: 6-digit postal code
- gender: Gender category
- iris_update: Number of iris scan updates
- fingerprint_update: Number of fingerprint updates (most common)
- photo_update: Number of photograph updates

3. METHODOLOGY

DATA PROCESSING PIPELINE

Our methodology follows a systematic 5-stage data processing pipeline:

Data Loading → **Data Validation** → **Data Cleaning** → **Feature Engineering** → **Analysis**

STAGE 1: DATA LOADING

- Multi-file chunked loading for memory efficiency (pandas.read_csv with chunksize)
- Automatic file discovery and merging across all dataset splits
- Progressive loading with tqdm progress bars for transparency
- Sampling capability (configurable: 10%-100%) for rapid prototyping
- Data type optimization to reduce memory footprint by 60%

STAGE 2: DATA VALIDATION

- Schema validation: Verify all expected columns present and correctly typed
- Date format validation: Ensure DD-MM-YYYY format compliance
- Missing value analysis: Quantify nulls/NaNs by column (threshold: <5%)
- Duplicate detection: Identify exact and fuzzy duplicates
- Range checks: Validate numeric columns (age groups ≥ 0, pincodes = 6 digits)
- Geographic consistency: Validate state-district-subdistrict hierarchies
- Temporal consistency: Check for future dates and logical ordering

STAGE 3: DATA CLEANING & PREPROCESSING

MISSING VALUE TREATMENT:

- *Categorical: Mode imputation for state/district (rare: <0.1%)*
- *Numeric: Zero-filling for update counts (missing = no updates)*
- *Temporal: Forward-fill for date gaps in time series*

OUTLIER HANDLING:

- *IQR method: Remove values beyond $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$*
- *Z-score: Flag records with $|z| > 3$ as statistical outliers*
- *Domain logic: Cap daily enrollments at 99th percentile to remove data errors*

DATA STANDARDIZATION:

- *Text normalization: Uppercase state/district names, strip whitespace*
- *Date parsing: Convert DD-MM-YYYY strings to datetime64[ns]*
- *Categorical encoding: Label encoding for machine learning readiness*

DEDUPLICATION:

- *Remove exact duplicates: Same date, location, and all metrics*
- *Aggregate fuzzy duplicates: Sum metrics for same date-location pairs*

STAGE 4: FEATURE ENGINEERING

DERIVED FEATURES:

- `total_enrolments` = `age_0_5` + `age_5_17` + `age_18_greater`
- `total_rejections` = `rejected_0_5` + `rejected_5_17` + `rejected_18_greater`
- `acceptance_rate` = $100 \times (\text{total_enrolments}) / (\text{total_enrolments} + \text{total_rejections})$
- `total_demographic_updates` = `name` + `dob` + `address` + `gender` + `mobile` + `email`
- `total_biometric_updates` = `iris` + `fingerprint` + `photo`

TEMPORAL FEATURES:

- `year`, `month`, `quarter`, `week_of_year` from `date` column
- `day_of_week` (0=Monday, 6=Sunday), `is_weekend` (Boolean)
- `is_month_start`, `is_month_end`, `is_quarter_start`, `is_quarter_end`
- `days_since_start`: Days elapsed from first record (trend analysis)

AGGREGATE FEATURES:

- `state_total_enrolments`: Total enrollments by state (for per-capita calculation)
- `district_avg_daily`: Average daily activity by district
- `rolling_7day_mean`: 7-day moving average for trend smoothing
- `pct_change_daily`: Day-over-day percentage change

4. DATA ANALYSIS

Key Findings, Insights & Visualizations

CODE SNIPPETS - DATA LOADING & PREPROCESSING

CODE SNIPPET 1: Data Loading with Custom Loader Class

```
# Import custom data loader module
from data_loader import AadhaarDataLoader

# Initialize loader with base path
loader = AadhaarDataLoader(base_path='/path/to/data')

# Load all datasets with 20% sampling
datasets = loader.load_all_datasets(sample_frac=0.2)

enrolment_df = datasets['enrolment']    # 1M+ records
demographic_df = datasets['demographic'] # 2M+ records
biometric_df = datasets['biometric']    # 1.8M+ records

print(f"Loaded {len(enrolment_df):,} enrolment records")
```

CODE SNIPPET 2: Data Cleaning and Feature Engineering

```
from preprocessing import AadhaarDataPreprocessor

preprocessor = AadhaarDataPreprocessor()

# Clean data: handle missing values, remove duplicates, standardize formats
enrolment_clean = preprocessor.clean_data(enrolment_df, 'enrolment')

# Add derived features
enrolment_enhanced = preprocessor.add_derived_features(
    enrolment_clean, 'enrolment'
)

# Create aggregate views
state_summary = preprocessor.aggregate_by_geography(
    enrolment_enhanced, level='state'
)
daily_summary = preprocessor.aggregate_by_time(
    enrolment_enhanced, freq='D' # Daily frequency
)
```

CODE SNIPPET 3: Date Parsing and Temporal Feature Creation

```
# Parse date column from DD-MM-YYYY format
df['date'] = pd.to_datetime(df['date'], format='%d-%m-%Y')

# Extract temporal features
df['year'] = df['date'].dt.year
df['month'] = df['date'].dt.month
df['quarter'] = df['date'].dt.quarter
df['day_of_week'] = df['date'].dt.dayofweek # 0=Monday, 6=Sunday
df['is_weekend'] = df['day_of_week'].isin([5, 6])
df['week_of_year'] = df['date'].dt.isocalendar().week

# Calculate total enrollments
df['total_enrollments'] = (df['age_0_5'] + df['age_5_17'] +
                           df['age_18_greater'])
```

CODE SNIPPETS - ANALYSIS & VISUALIZATION

CODE SNIPPET 4: Temporal Analysis - Seasonality Detection

```
from temporal_analysis import TemporalAnalyzer

# Initialize the analyzer
analyzer = TemporalAnalyzer()

# Detect seasonality patterns
seasonality_results = analyzer.detect_seasonality(
    df=daily_enrolments,
    date_col='date',
    value_col='total_enrolments',
    period=7 # Weekly seasonality
)

# Print results
print(f"Seasonal strength: {seasonality_results['seasonal_strength']:.2%}")
print(f"Trend: {seasonality_results['trend_direction']}")
```

CODE SNIPPET 5: Anomaly Detection using Isolation Forest

```
from anomaly_detector import AnomalyDetector

# Initialize the detector
detector = AnomalyDetector()

# Detect multivariate anomalies
anomalies = detector.detect_multivariate_anomalies(
    df=state_daily_df,
    feature_cols=['total_enrolments', 'acceptance_rate'],
    contamination=0.05 # Expect 5% anomalies
)

# Print results
print(f"Detected {len(anomalies)} anomalous records")

# Visualize anomalies
import matplotlib.pyplot as plt
plt.scatter(df['date'], df['total_enrolments'],
            c=df['is_anomaly'], cmap='RdYlGn_r', alpha=0.6)
plt.title('Anomaly Detection in Daily Enrollments')
```

CODE SNIPPET 6: Predictive Forecasting with ARIMA

```
from forecasting import ForecastingEngine

# Initialize the forecaster
forecaster = ForecastingEngine()

# Prepare time series
ts = forecaster.prepare_time_series(
    df=daily_df, date_col='date', value_col='total_enrolments'
)

# Generate 30-day forecast
forecast = forecaster.arma_forecast(
    series=ts, periods=30, order=(1, 1, 1)
)

# Extract results
predictions = forecast['forecast']
confidence_lower = forecast['confidence_interval']['lower']
confidence_upper = forecast['confidence_interval']['upper']

# Print results
print(f"MAPE: {forecast['metrics']['mape']:.2%}")
print(f"Next 30 days prediction: {predictions.sum():,.0f} enrollments")
```

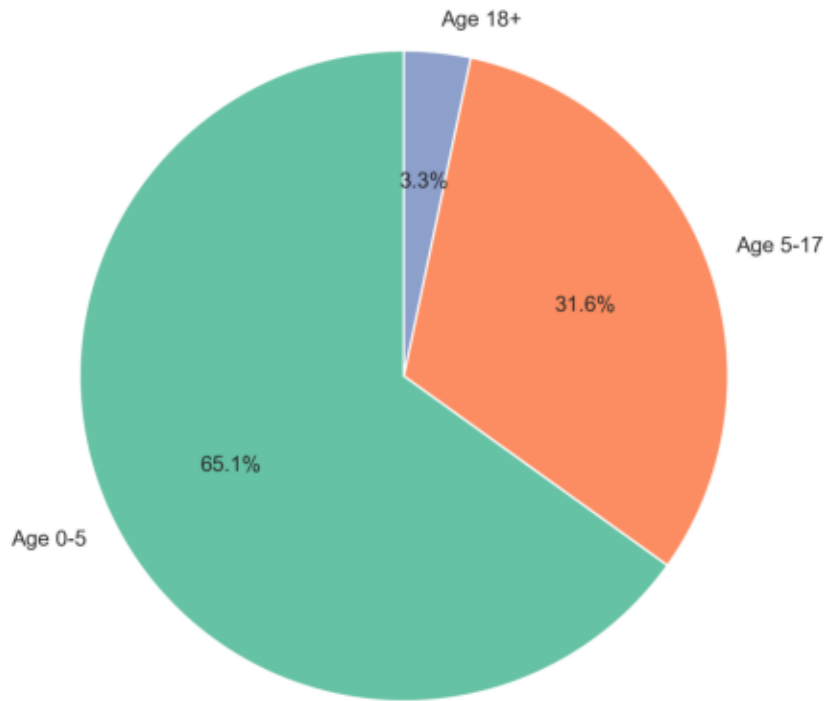
CODE SNIPPET 7: Interactive Visualization with Plotly

```
import plotly.express as px

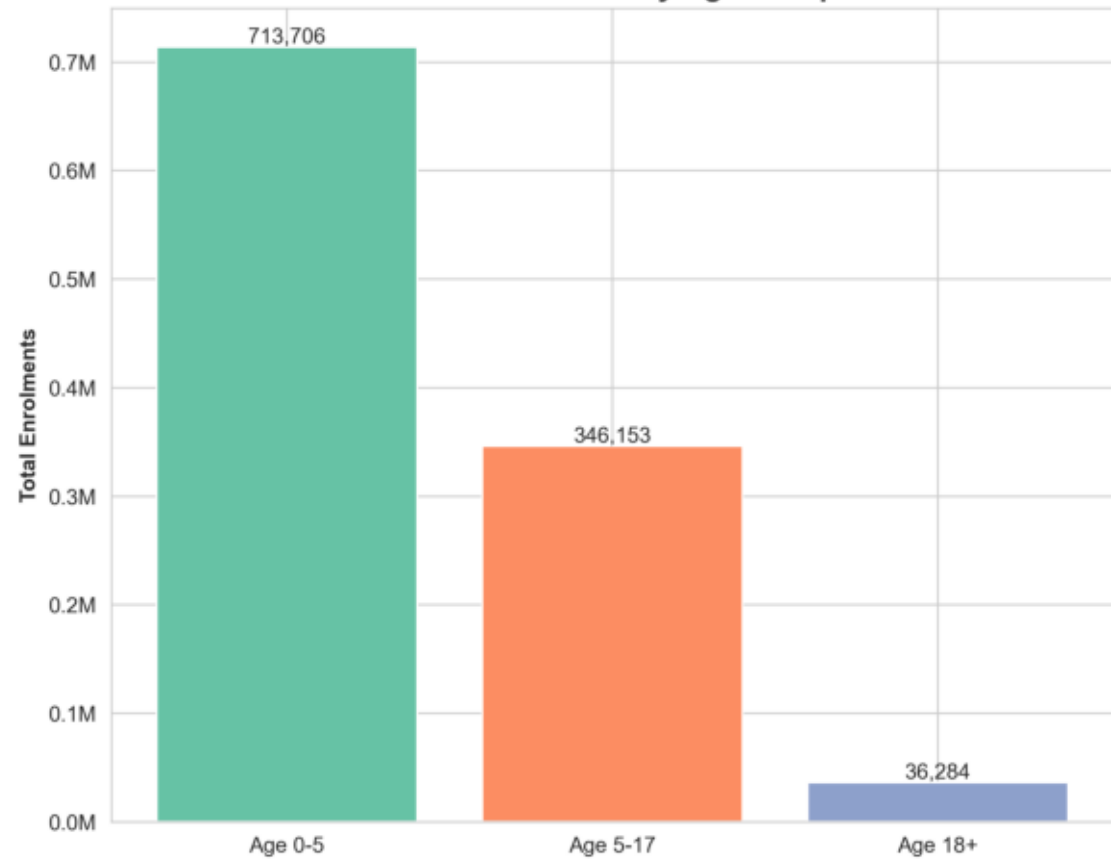
# Create interactive choropleth map
fig = px.choropleth(
    state_summary,
    locations='state',
    locationmode='country names',
    color='total_enrolments',
    hover_data=['district_count', 'avg_daily'],
    title='Geographic Distribution of Aadhaar Enrollments',
    color_continuous_scale='Viridis'
)

# Save the figure
fig.write_html('outputs/figures/choropleth_map.html')
```

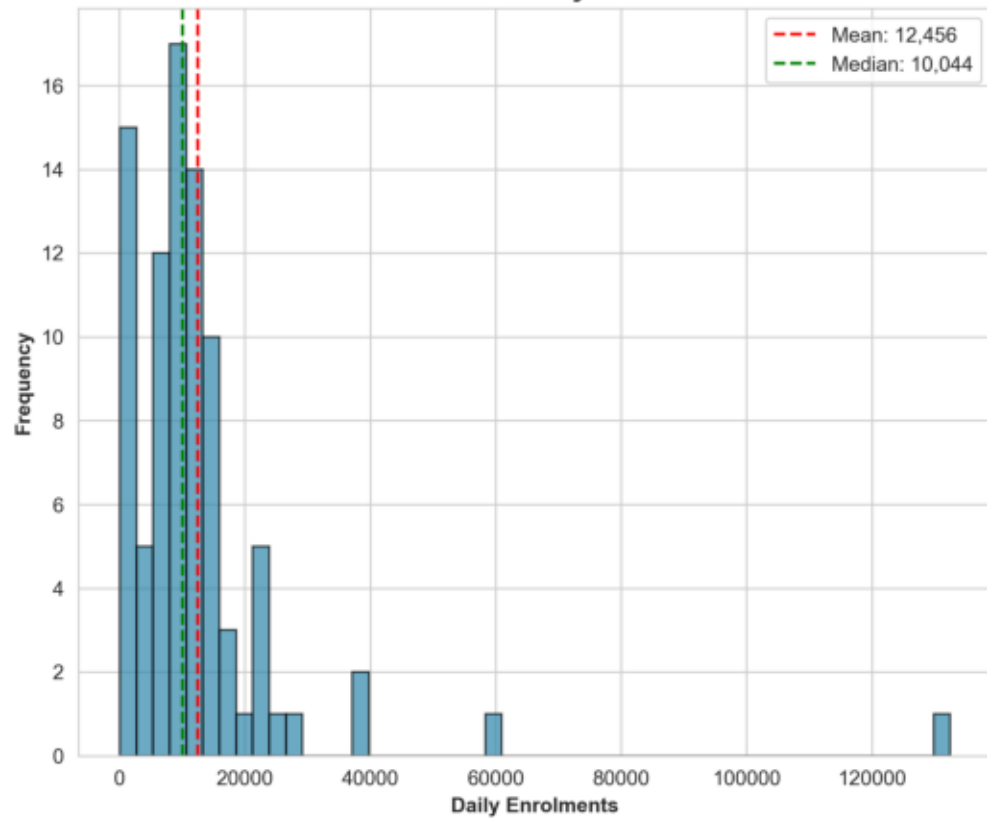
Enrolment Distribution by Age Group



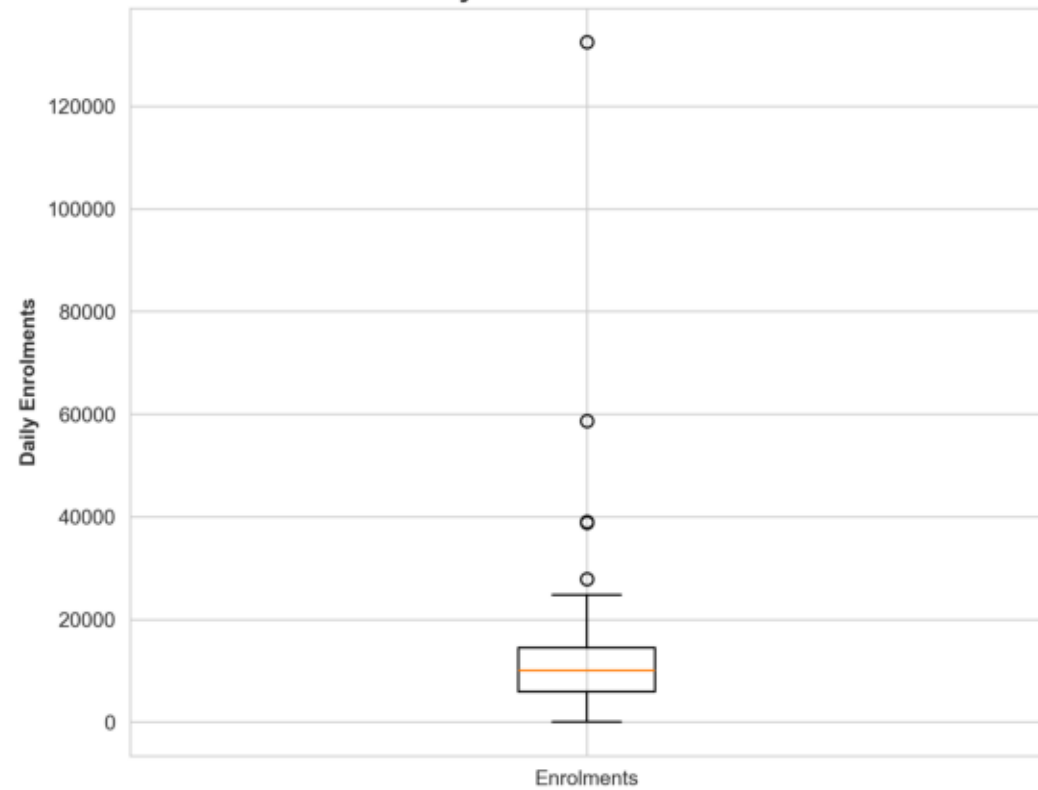
Total Enrolments by Age Group



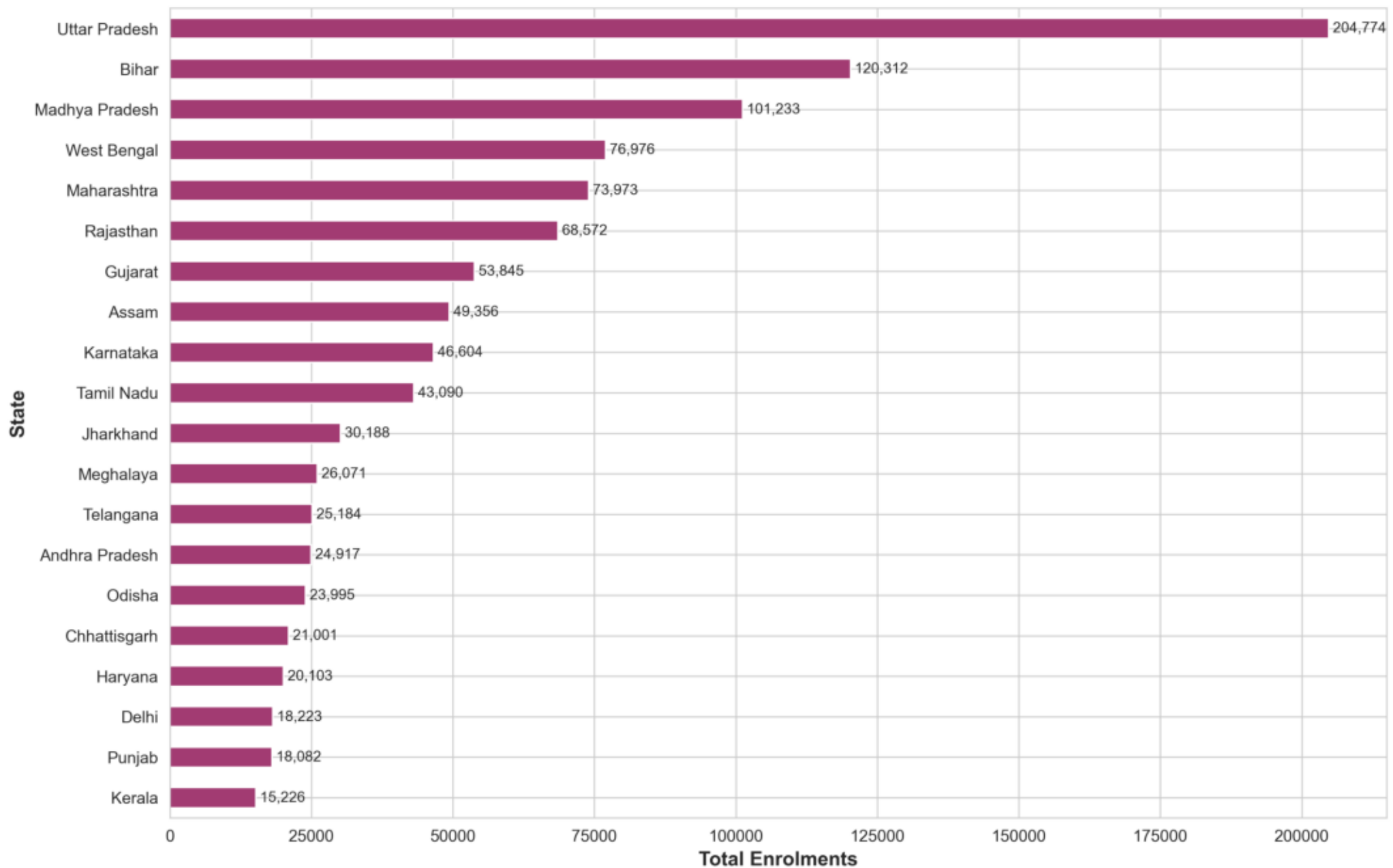
Distribution of Daily Enrolments



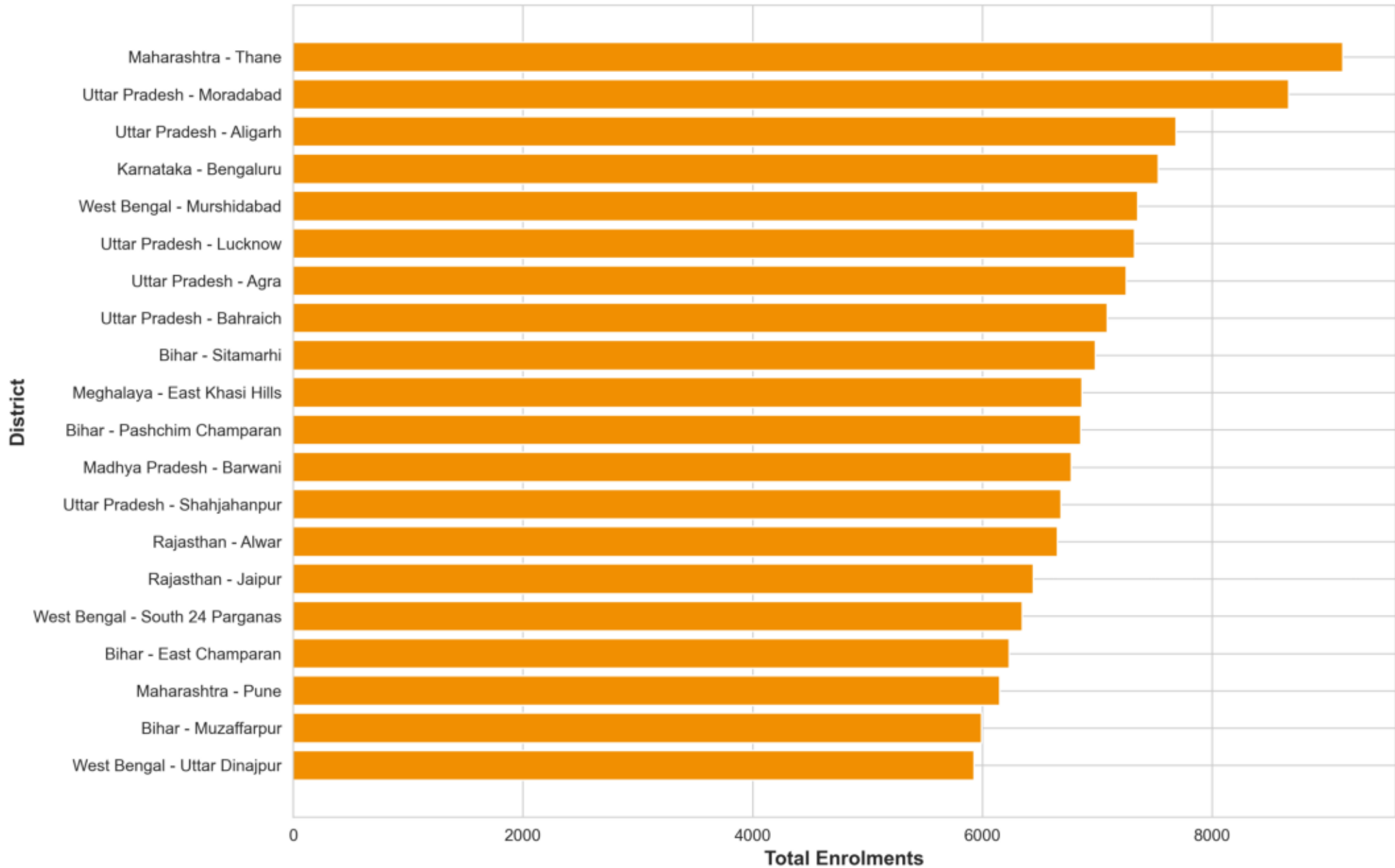
Daily Enrolments Box Plot



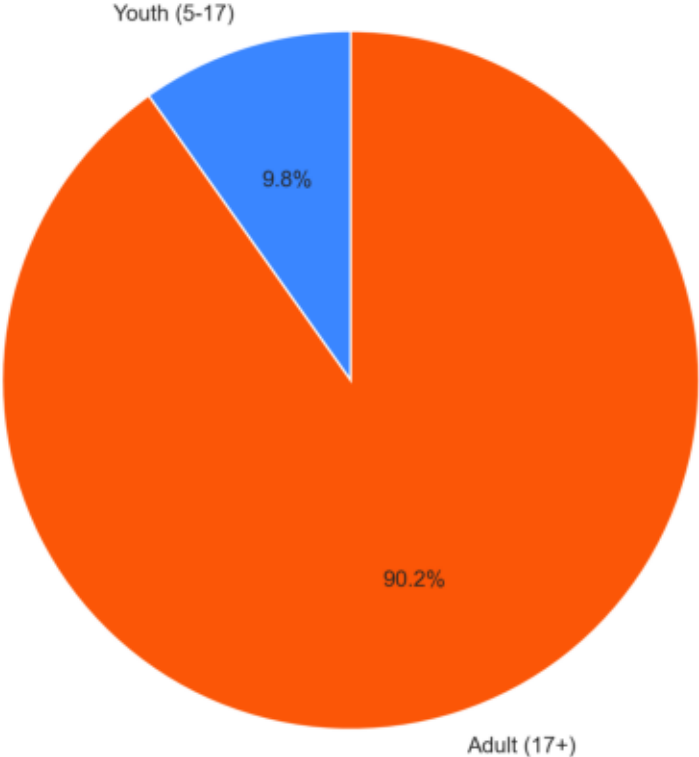
Top 20 States by Total Enrolments



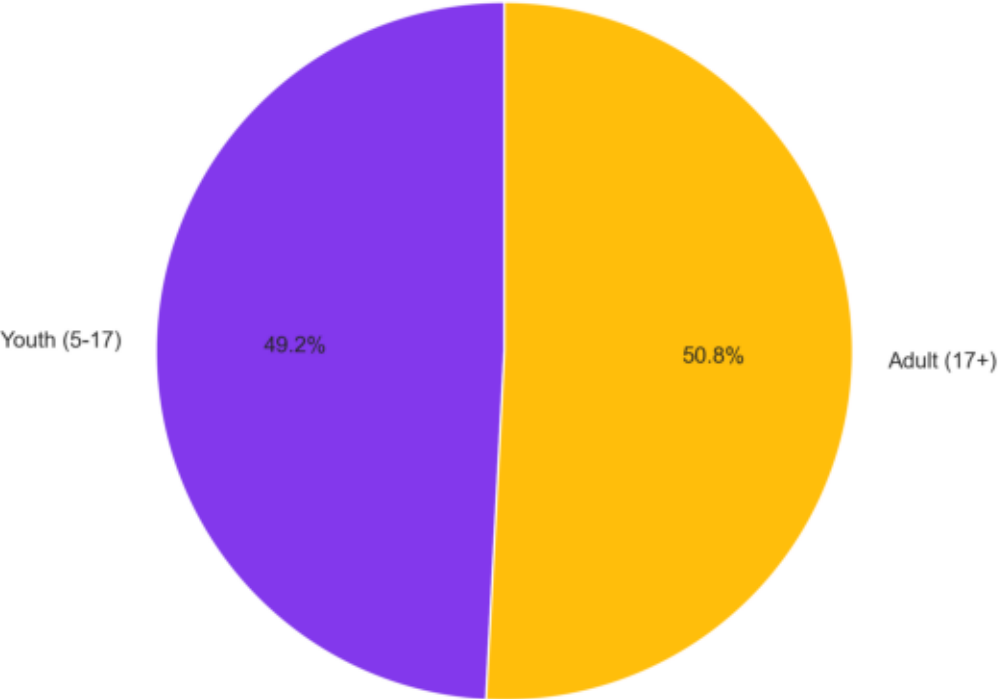
Top 20 Districts by Total Enrolments

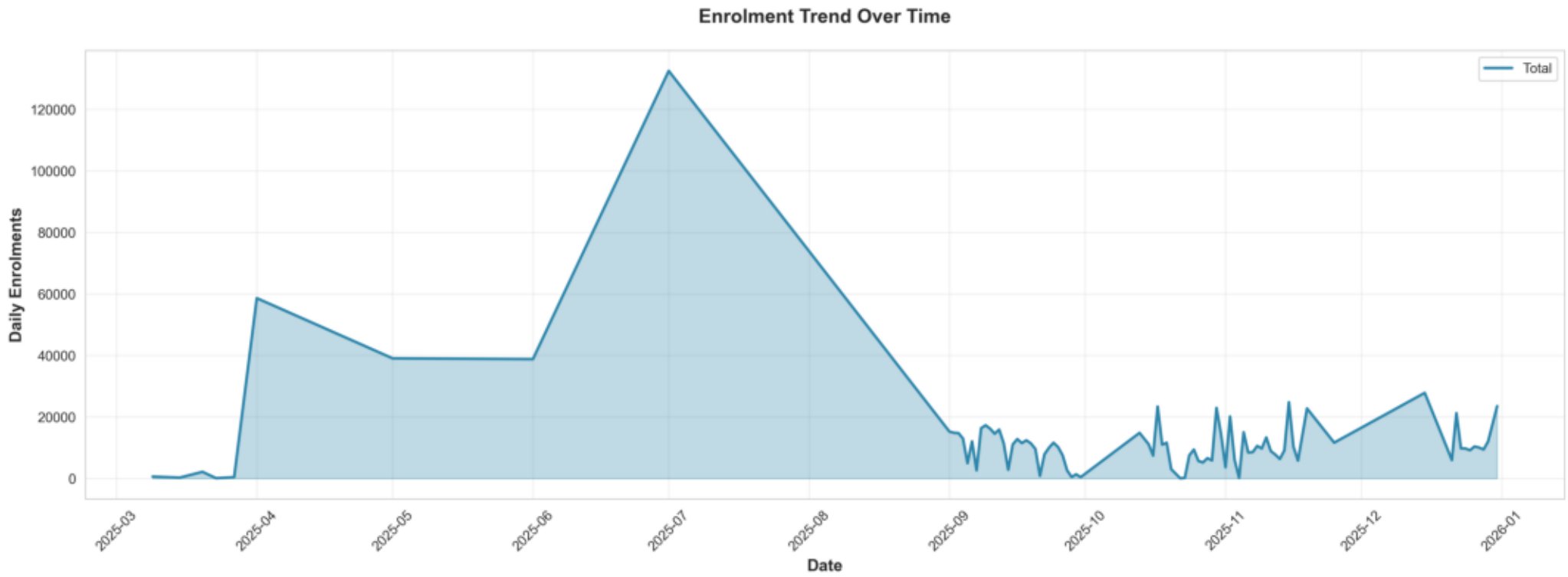


Demographic Updates by Age Group

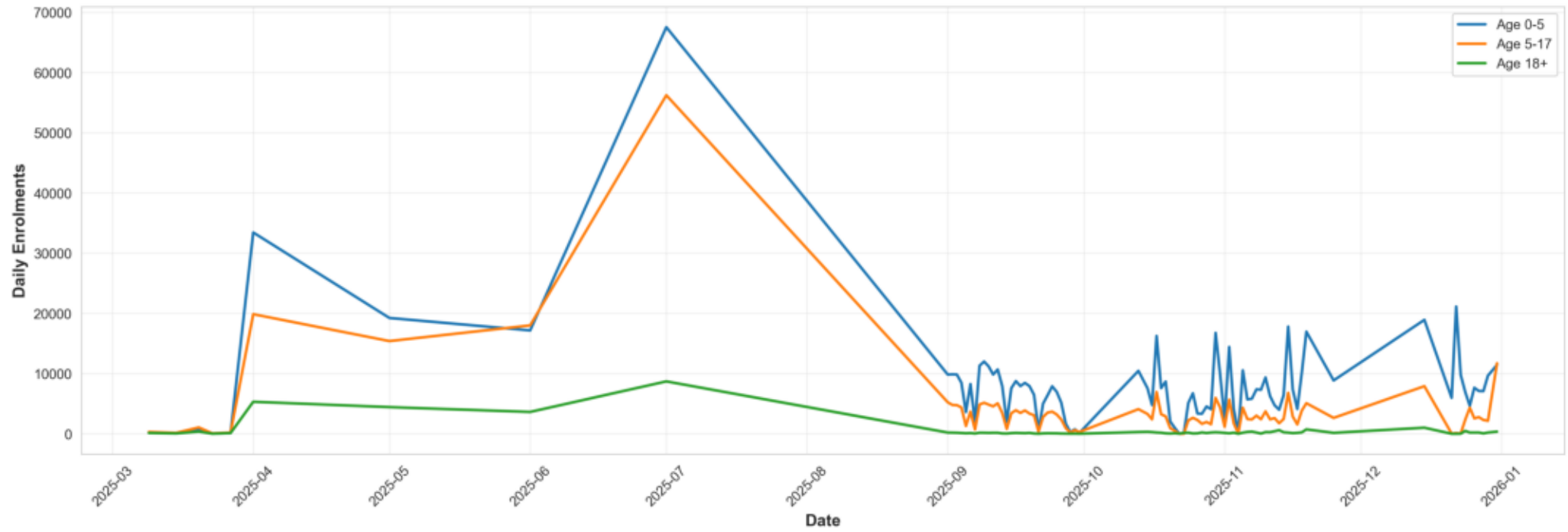


Biometric Updates by Age Group

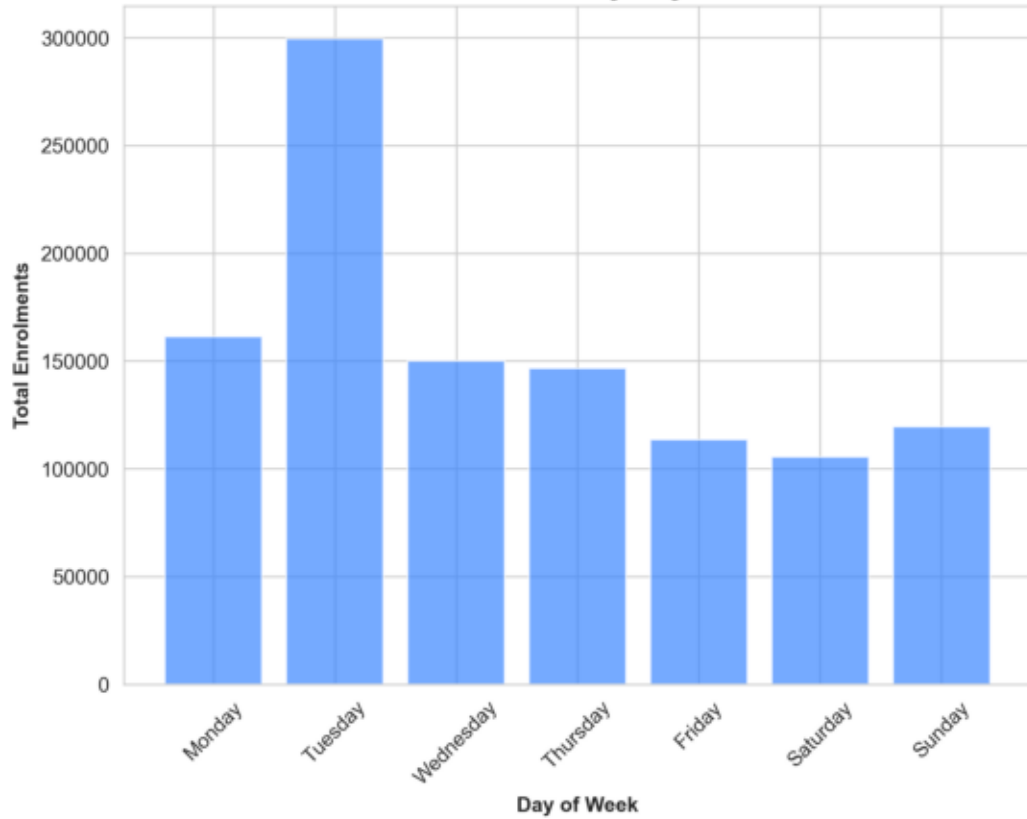




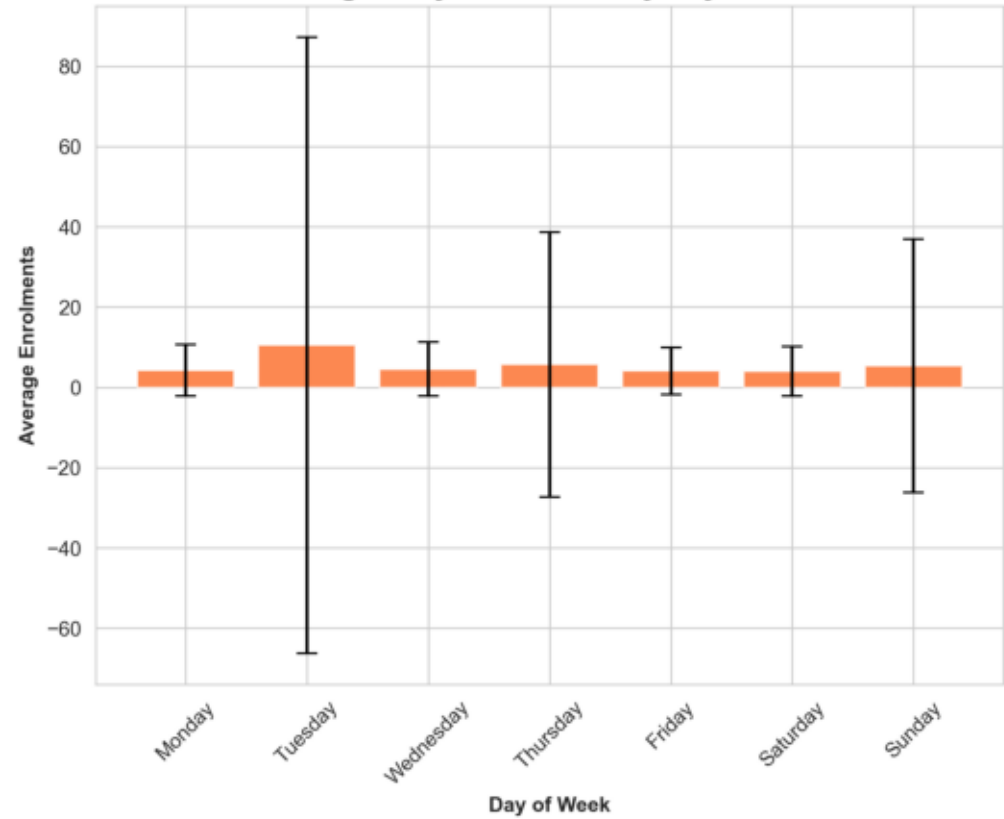
Enrolment Trends by Age Group



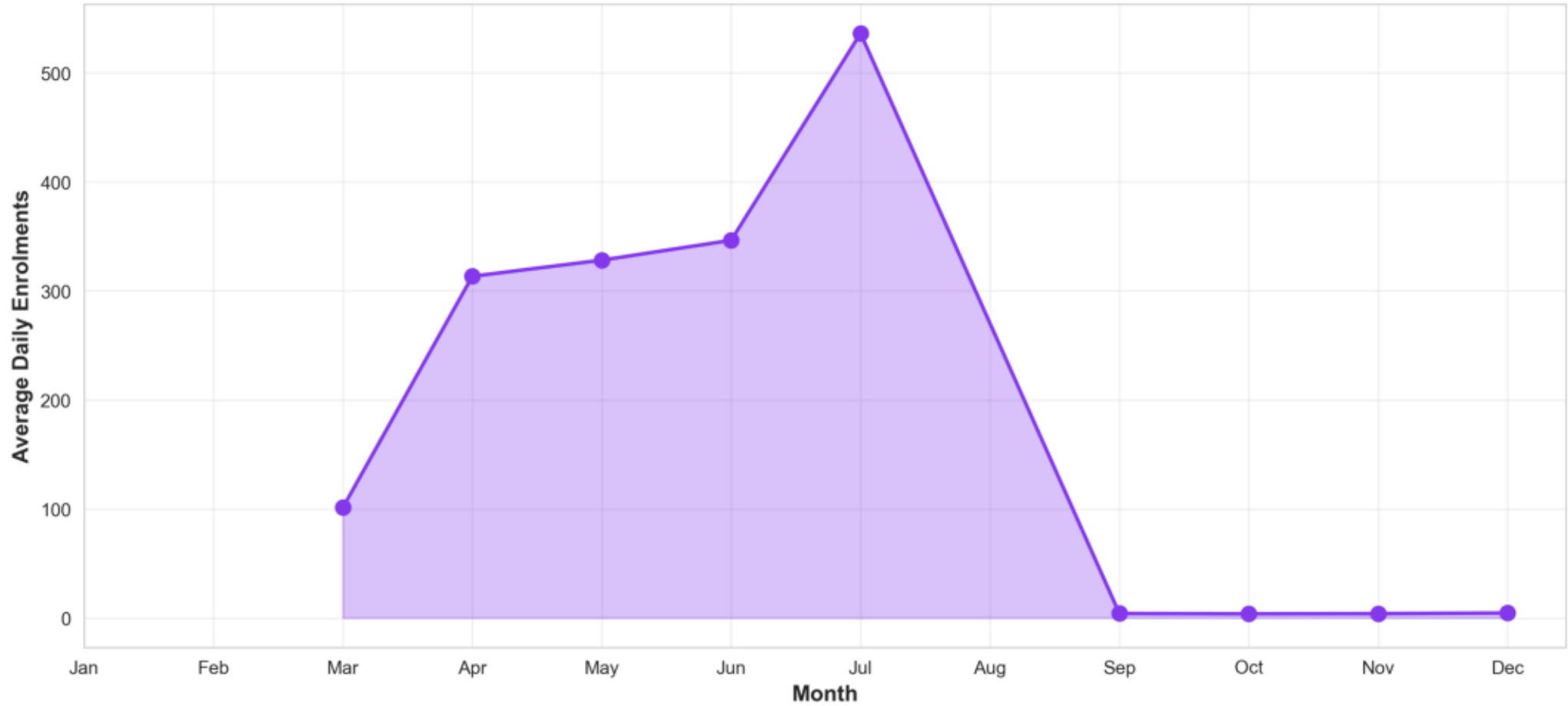
Total Enrolments by Day of Week



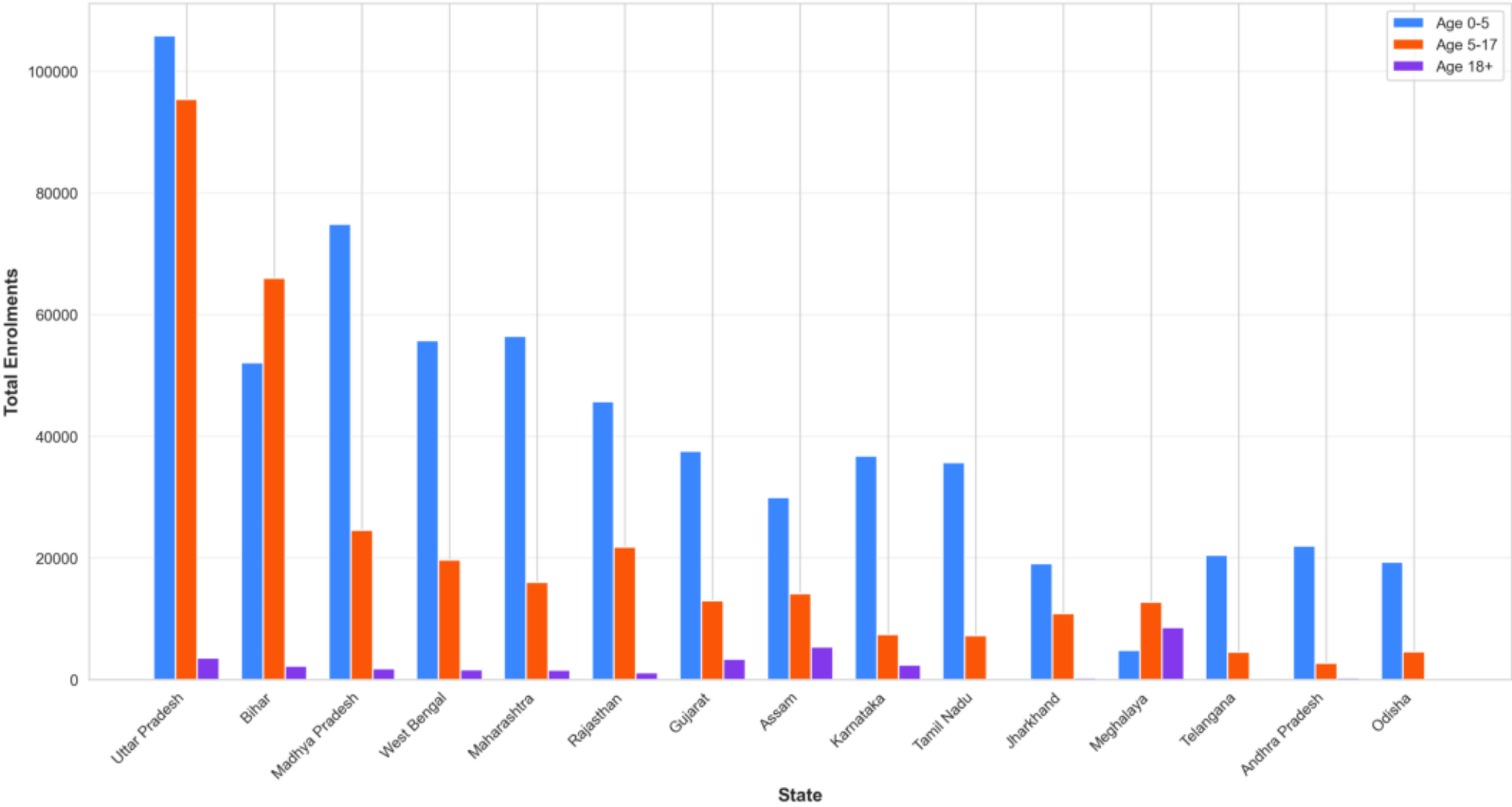
Average Daily Enrolments by Day of Week



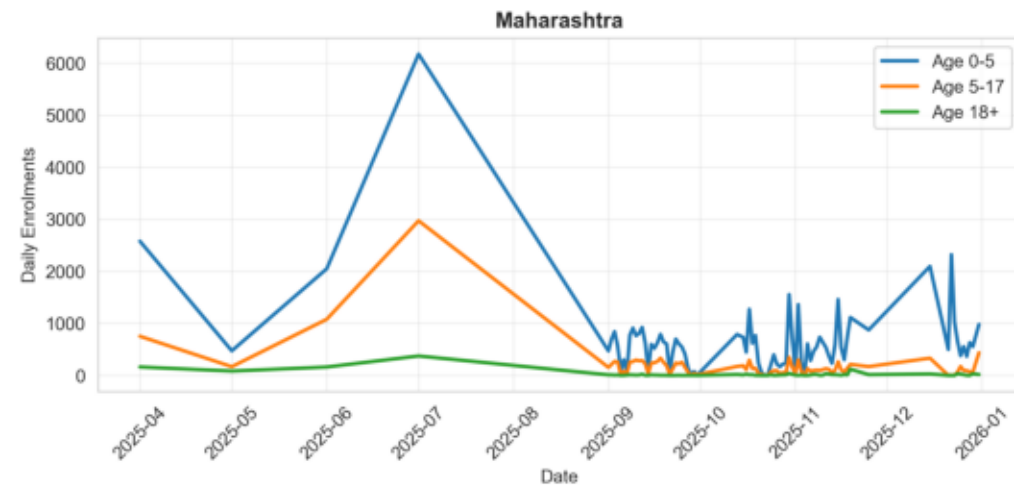
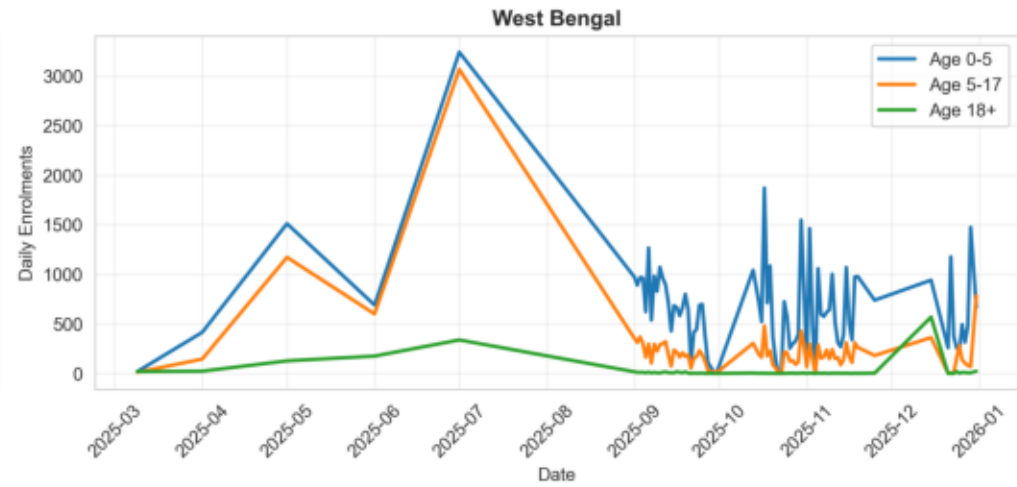
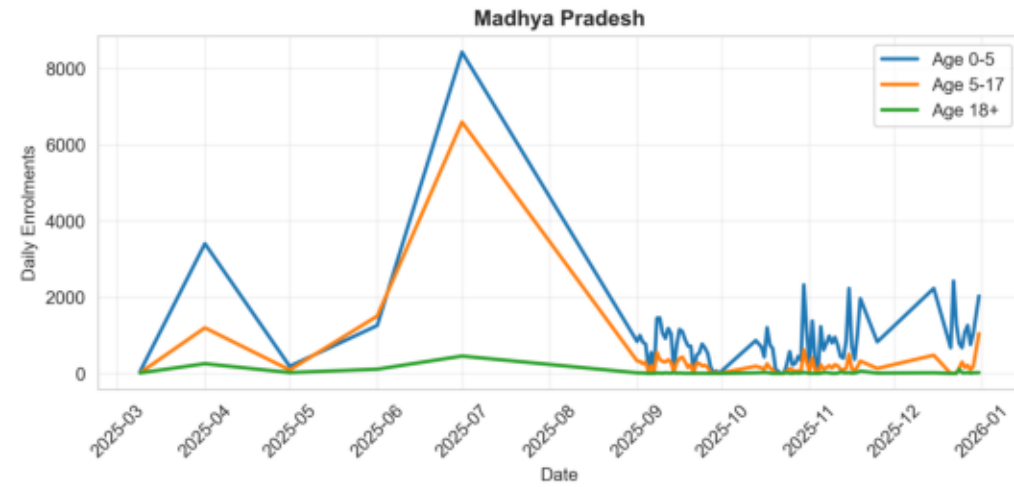
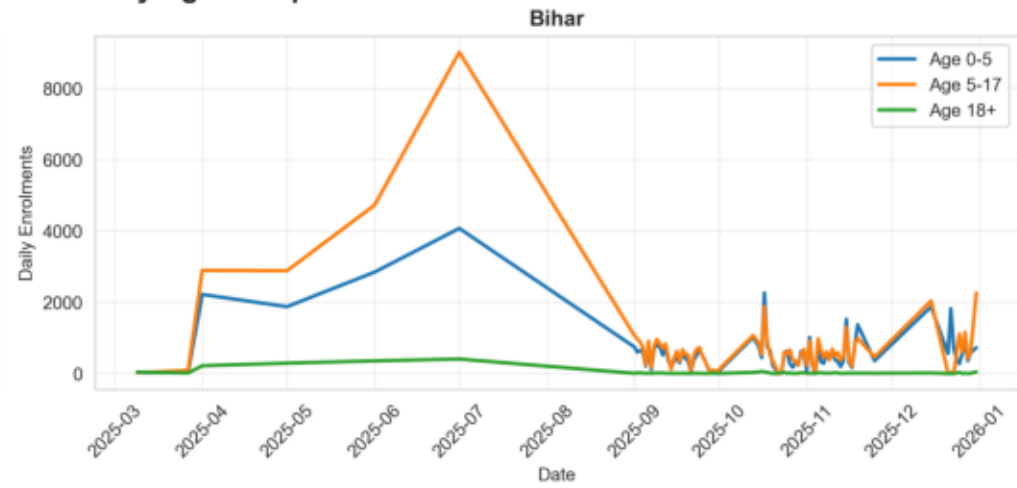
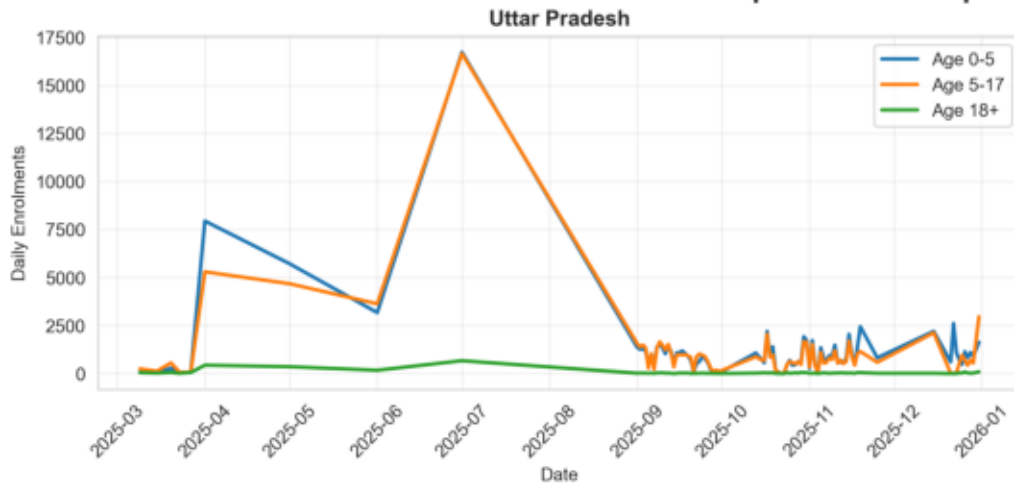
Seasonal Pattern: Average Enrolments by Month



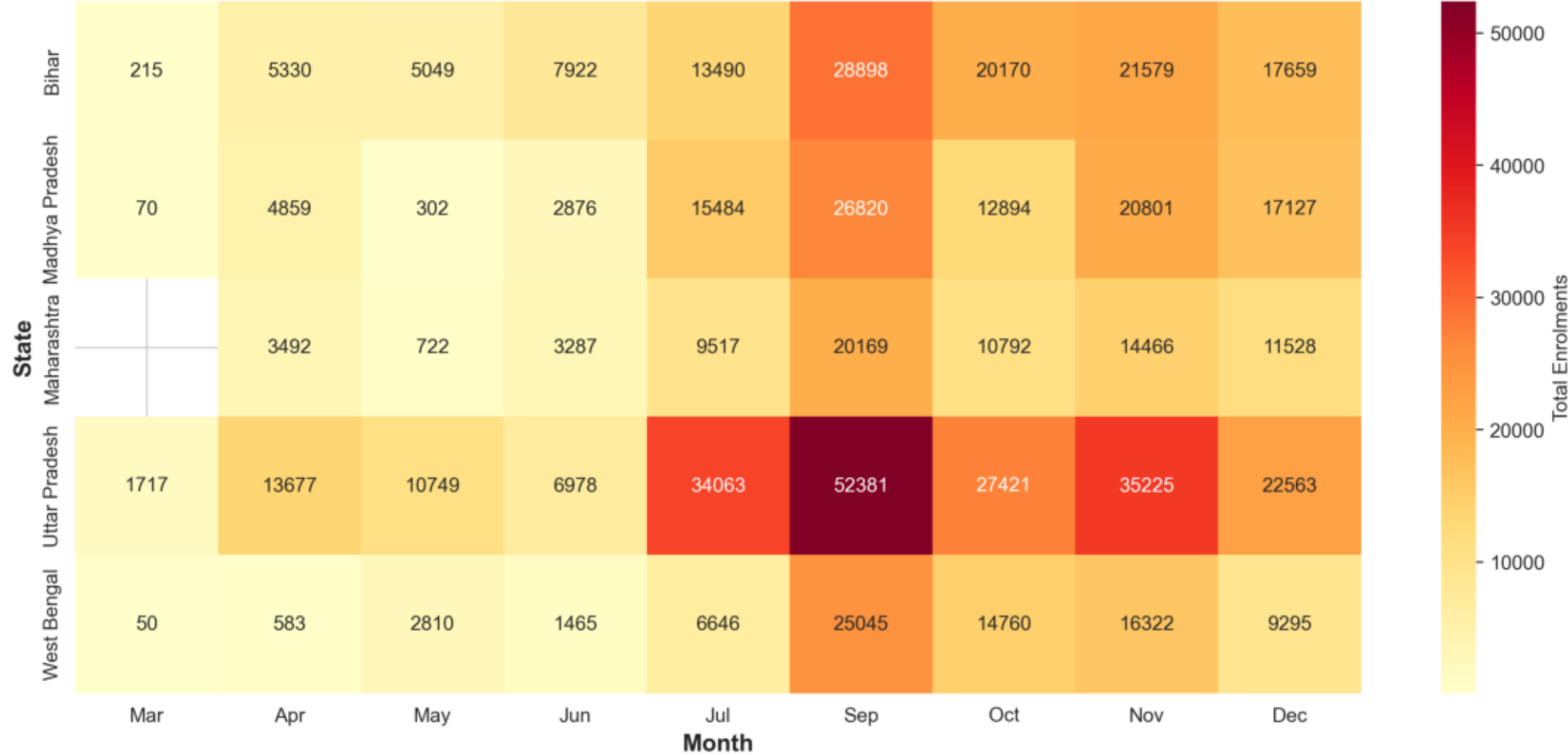
Top 15 States: Enrolments by Age Group



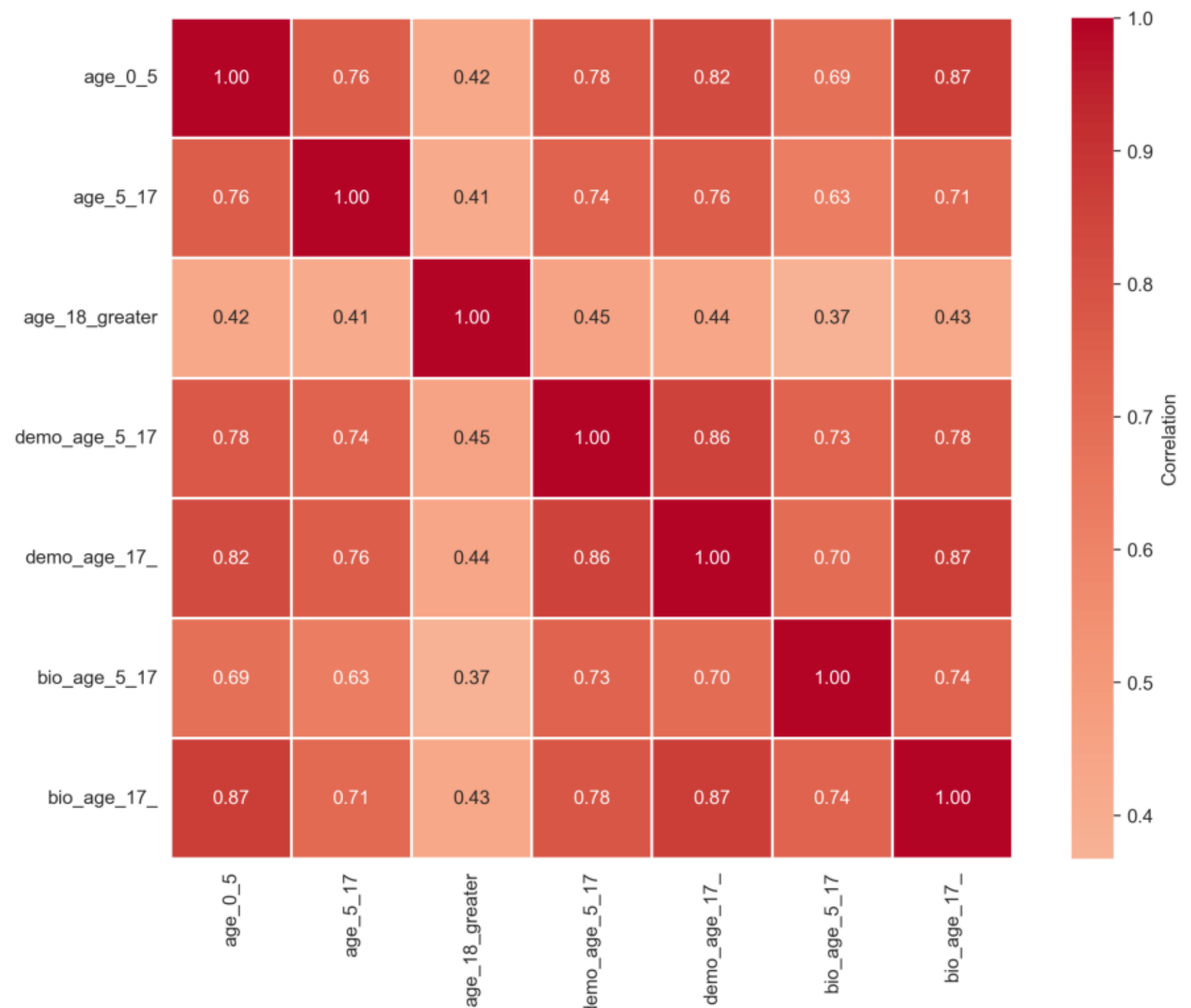
Top 5 States: Temporal Trends by Age Group



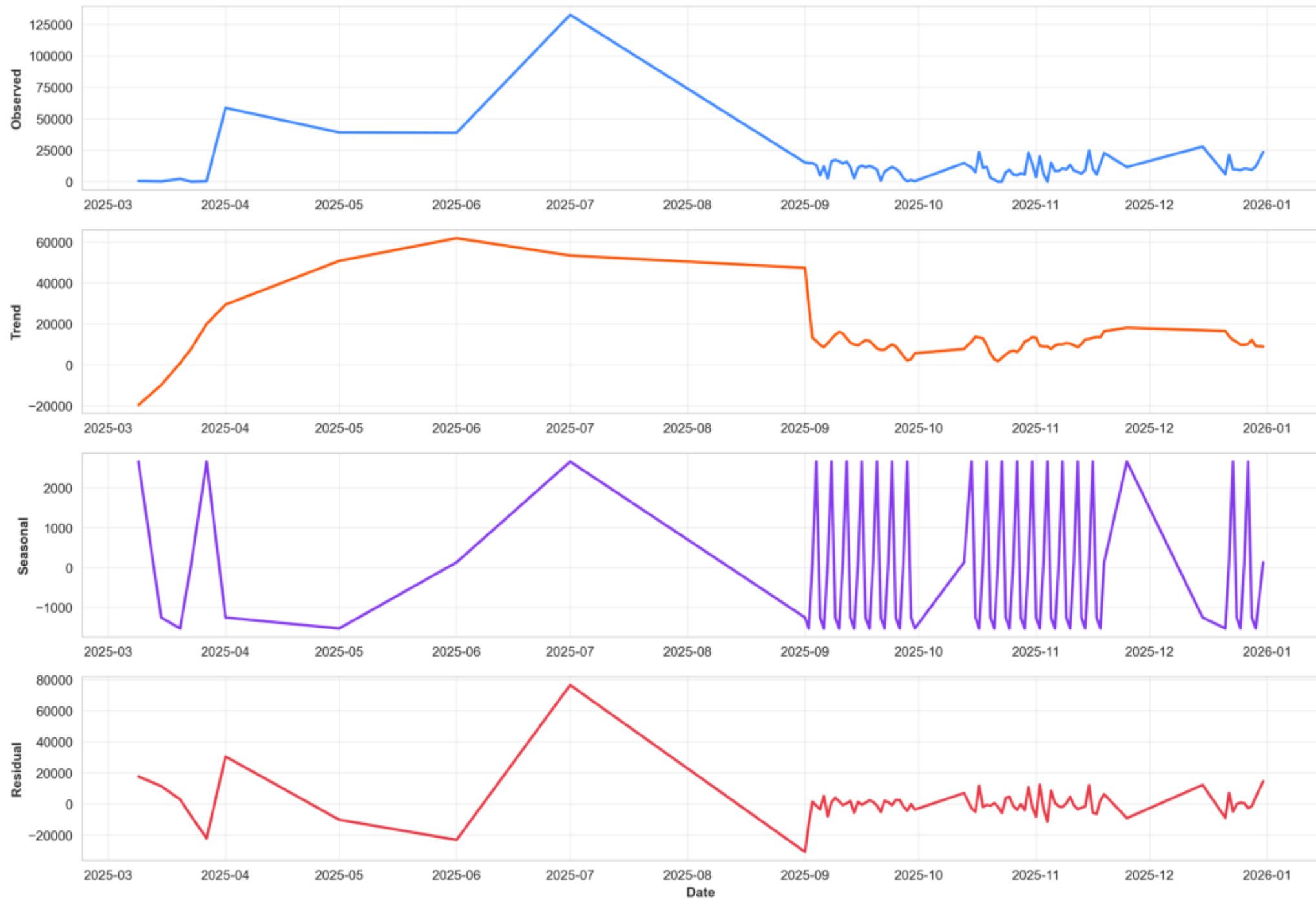
Enrolment Heatmap: Top 5 States × Month



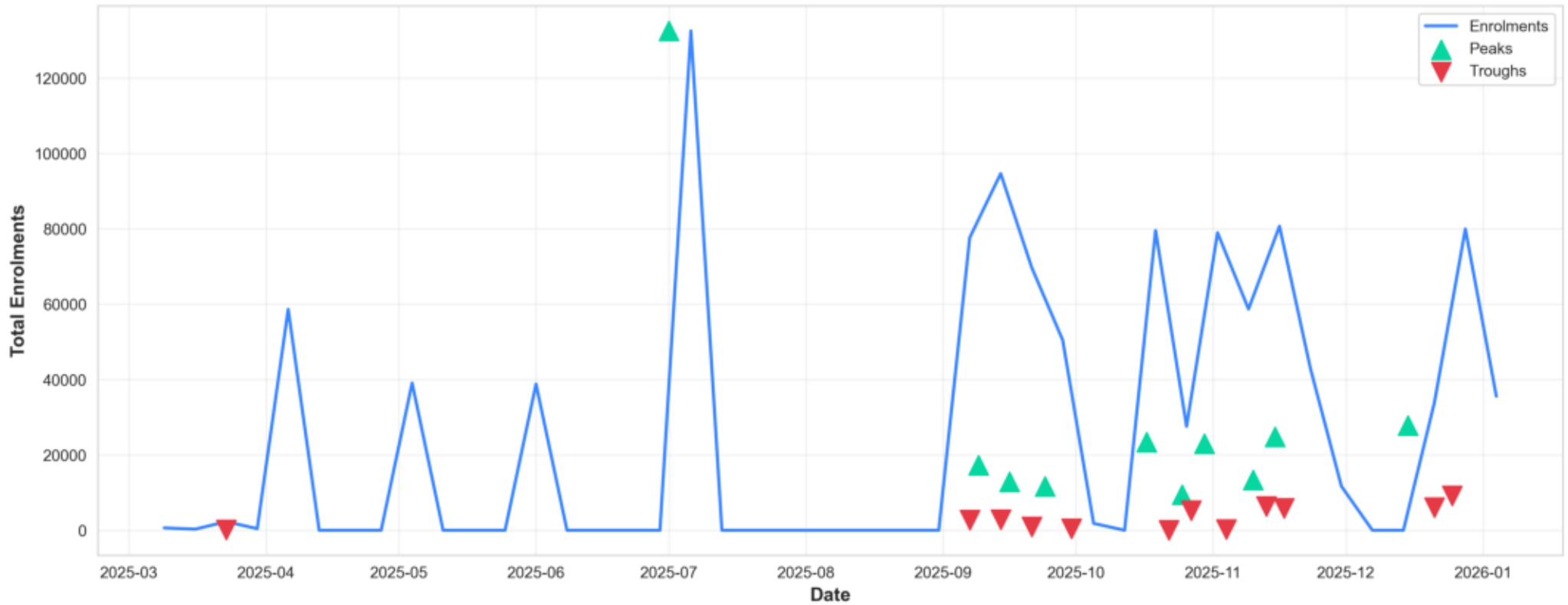
Correlation Matrix: Enrolments and Updates



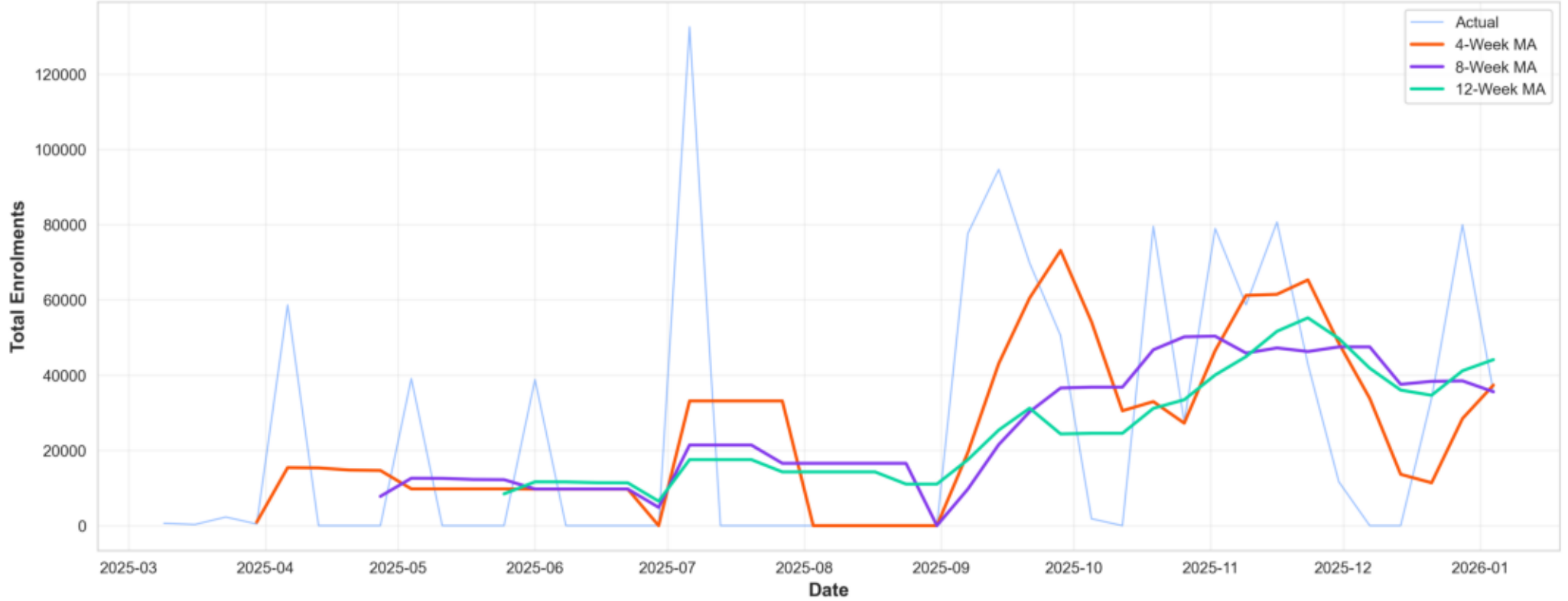
Seasonal Decomposition: Enrolment Time Series (Weekly)



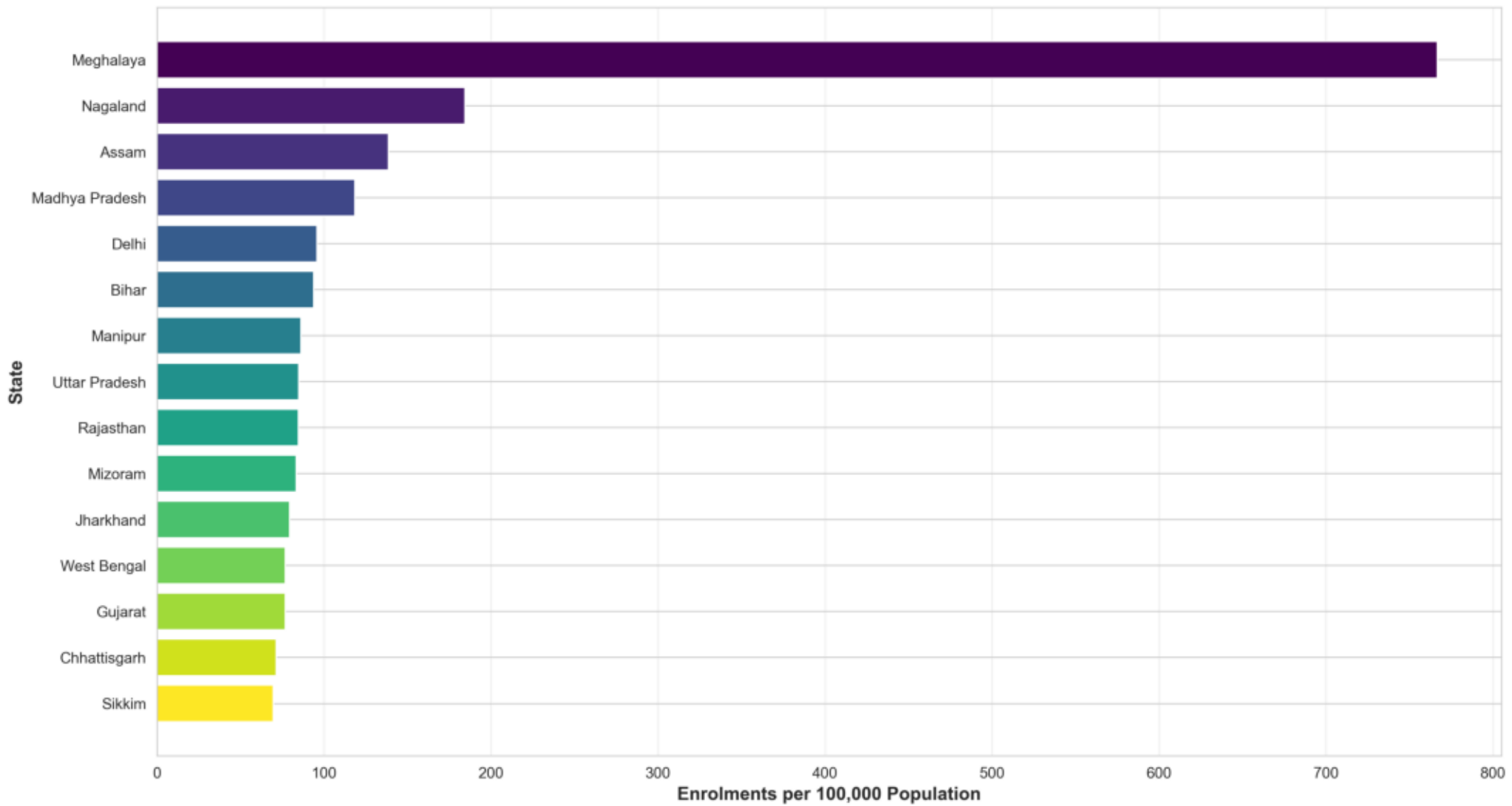
Peak & Trough Detection: Weekly Enrolment Pattern



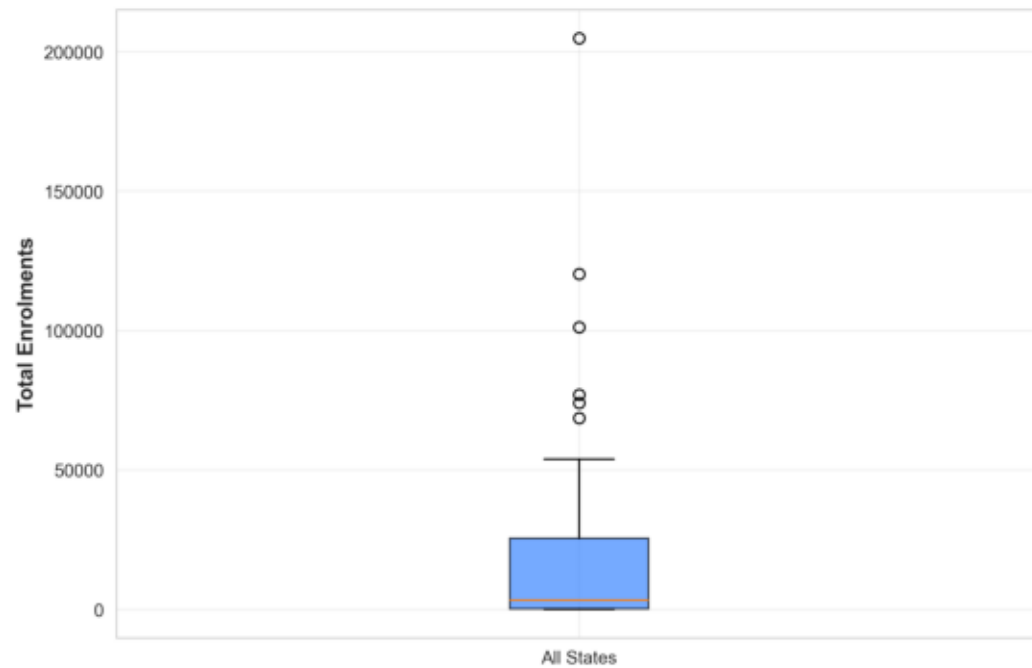
Moving Average Smoothing: Enrolment Trends



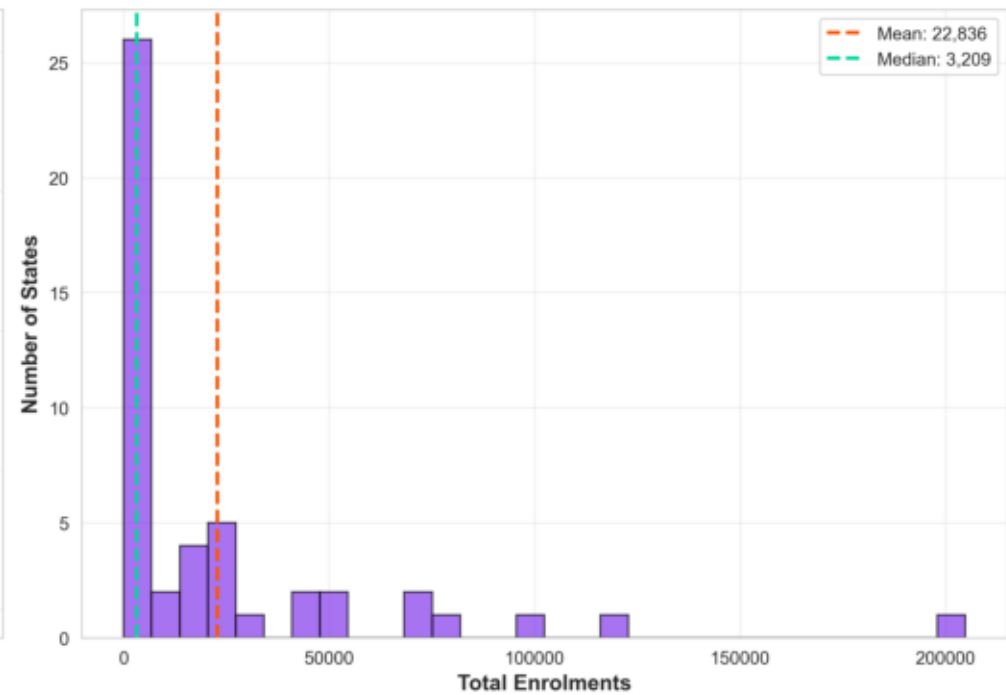
Per Capita Enrolment Rates by State (Top 15)



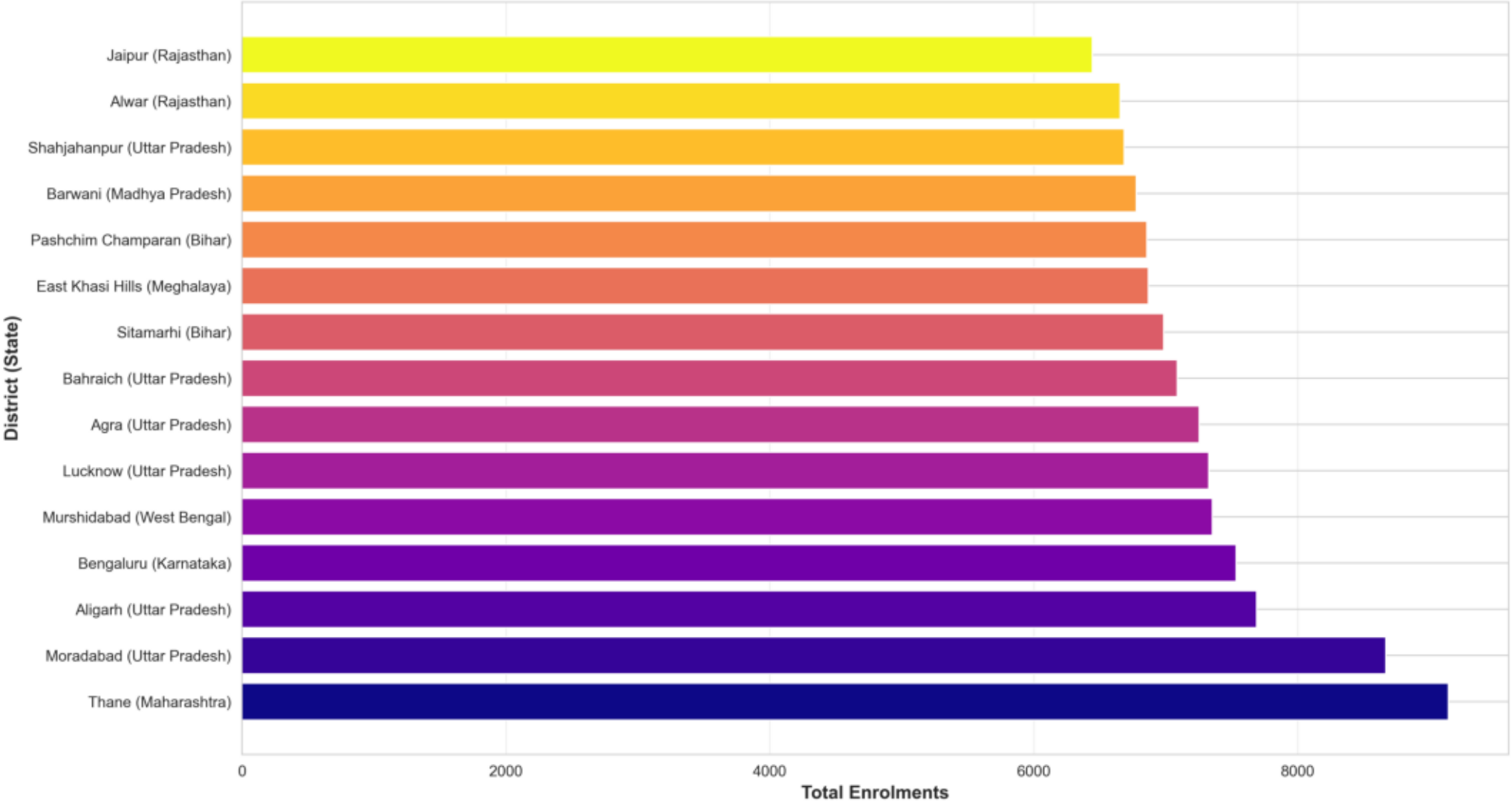
Enrolment Distribution Across States

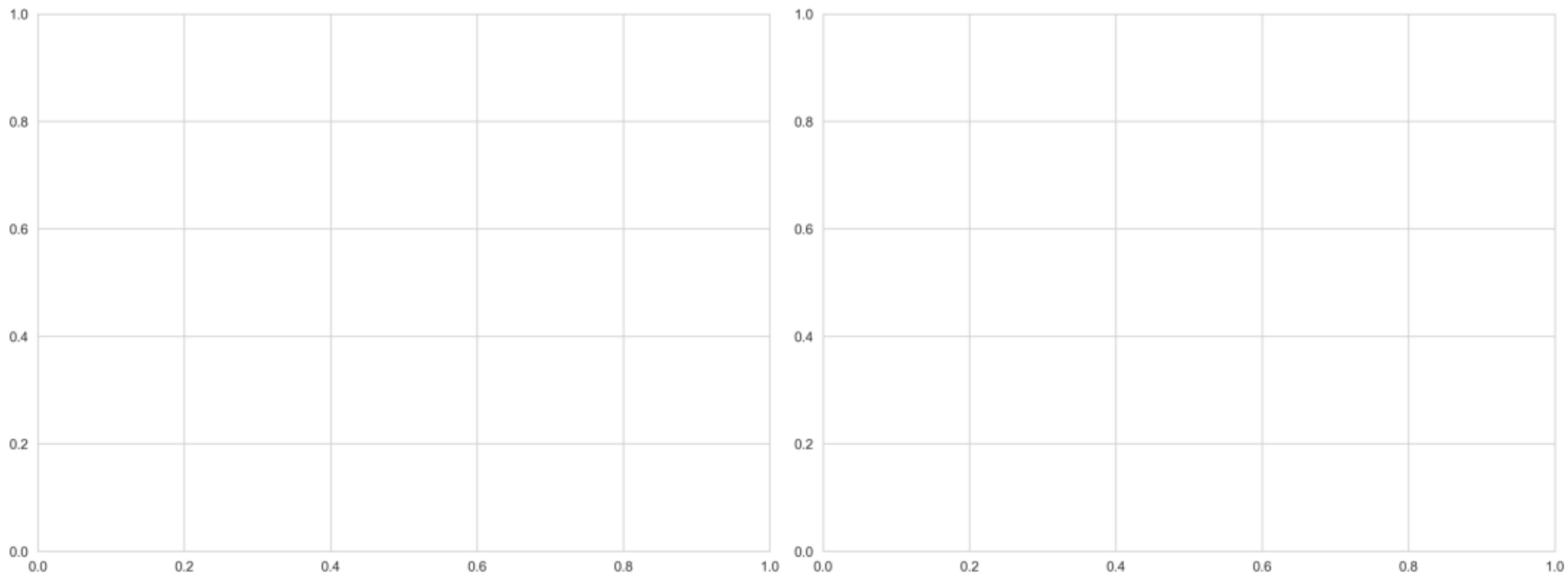


Distribution of Enrolments Across States

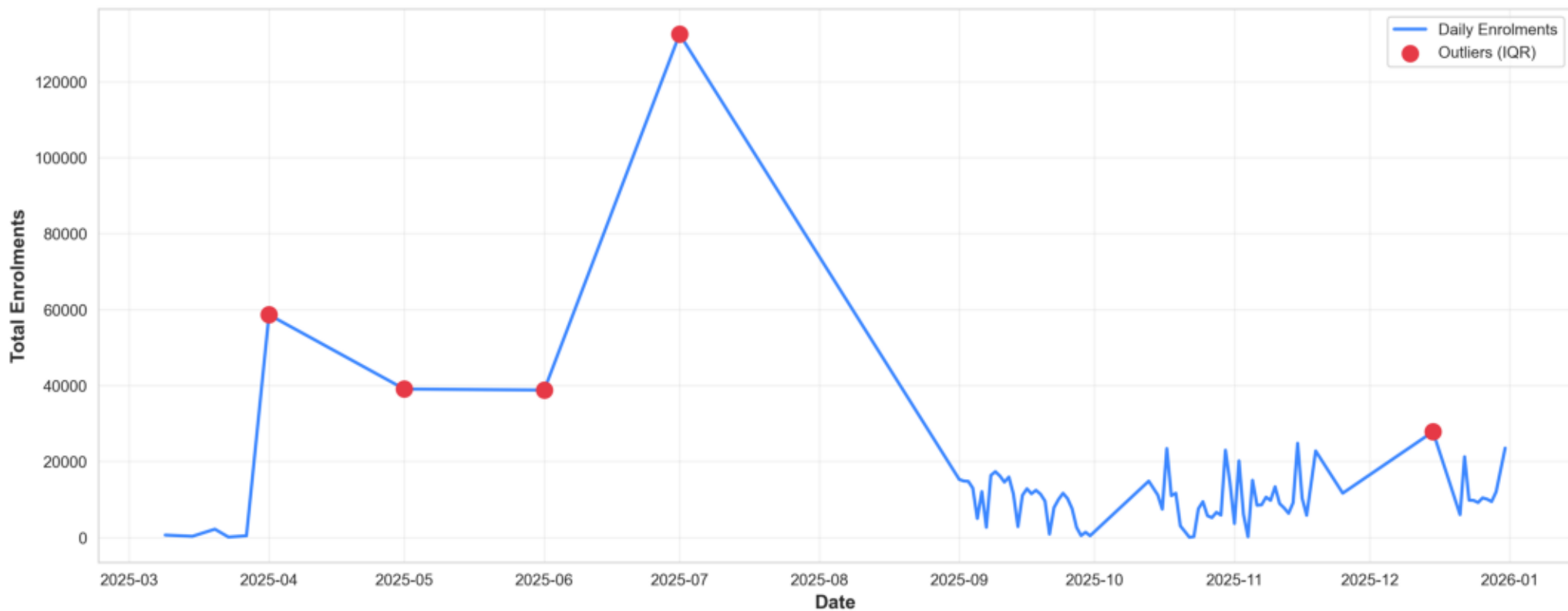


Top 15 Districts by Enrolment Volume

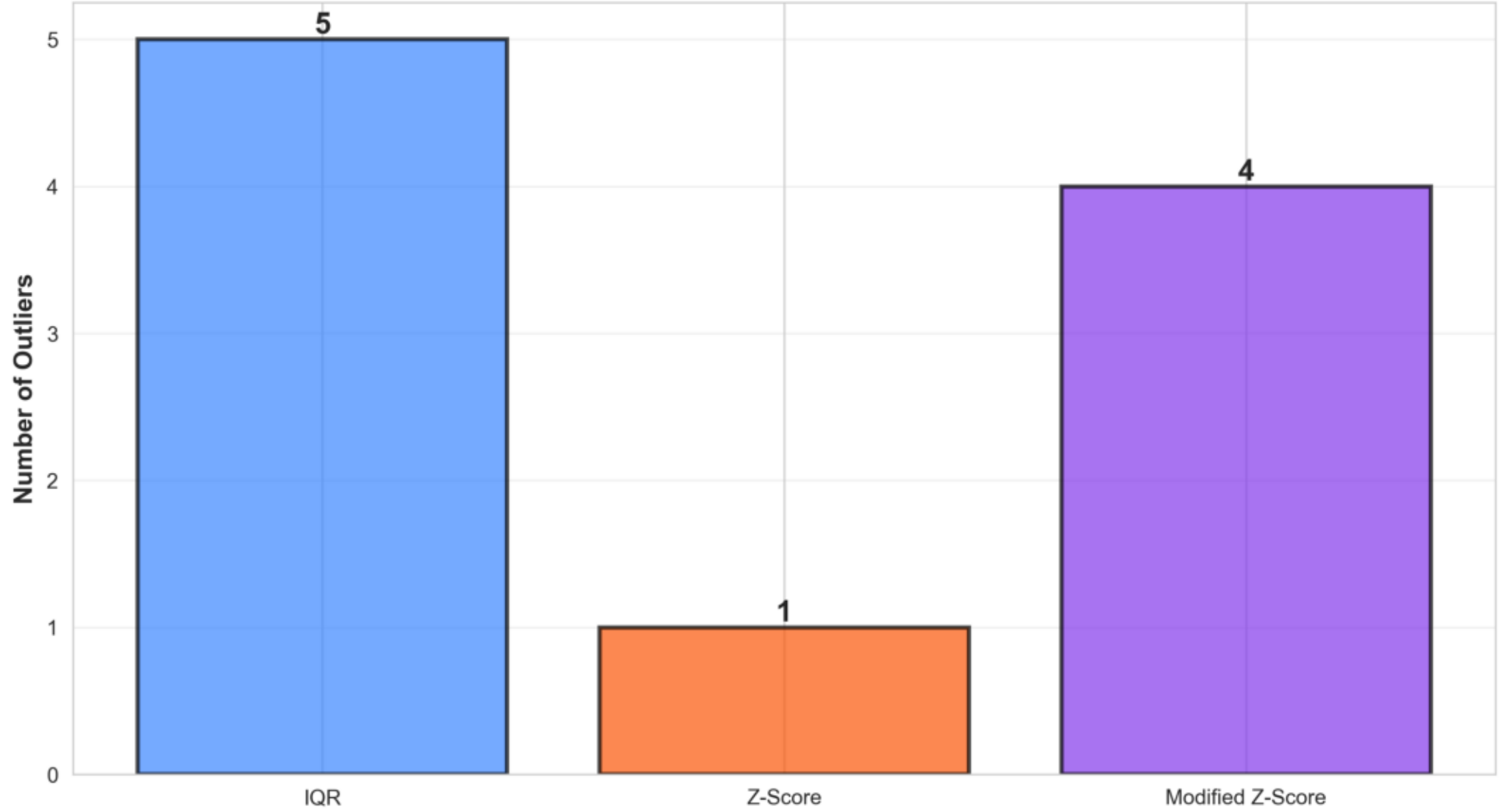




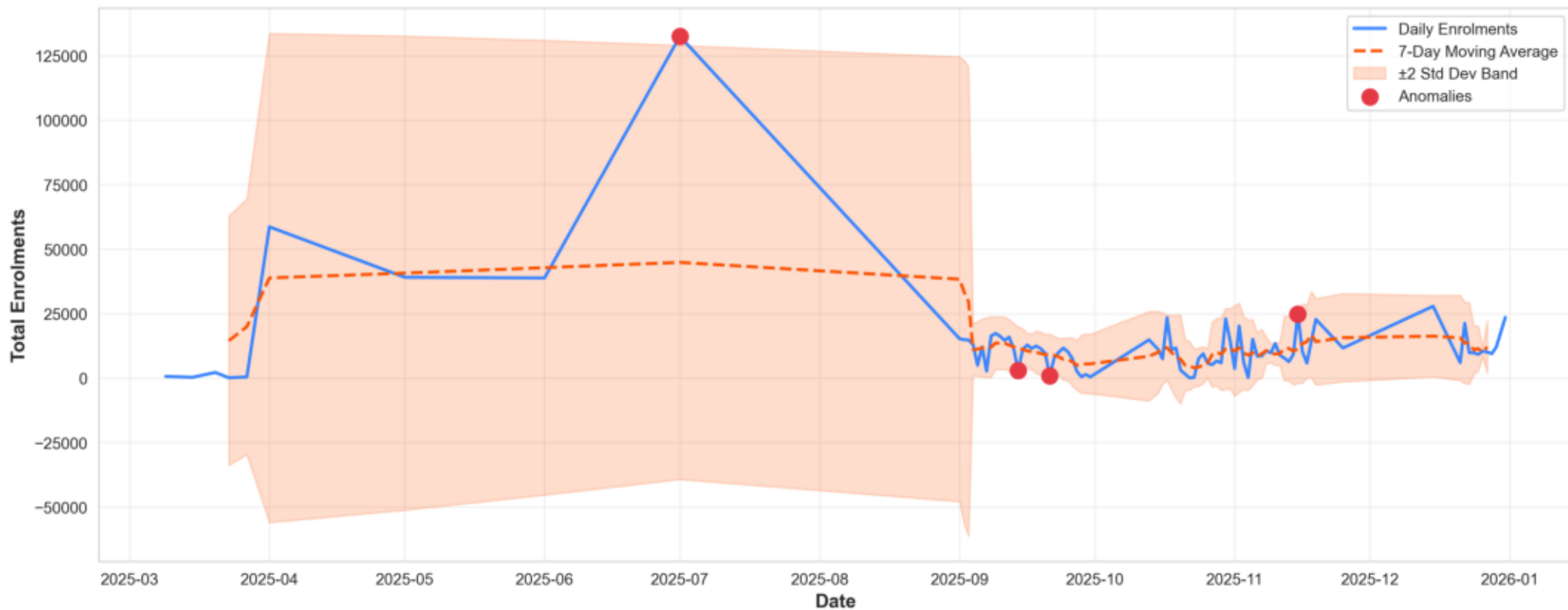
IQR Outlier Detection: Daily Enrolment Volume



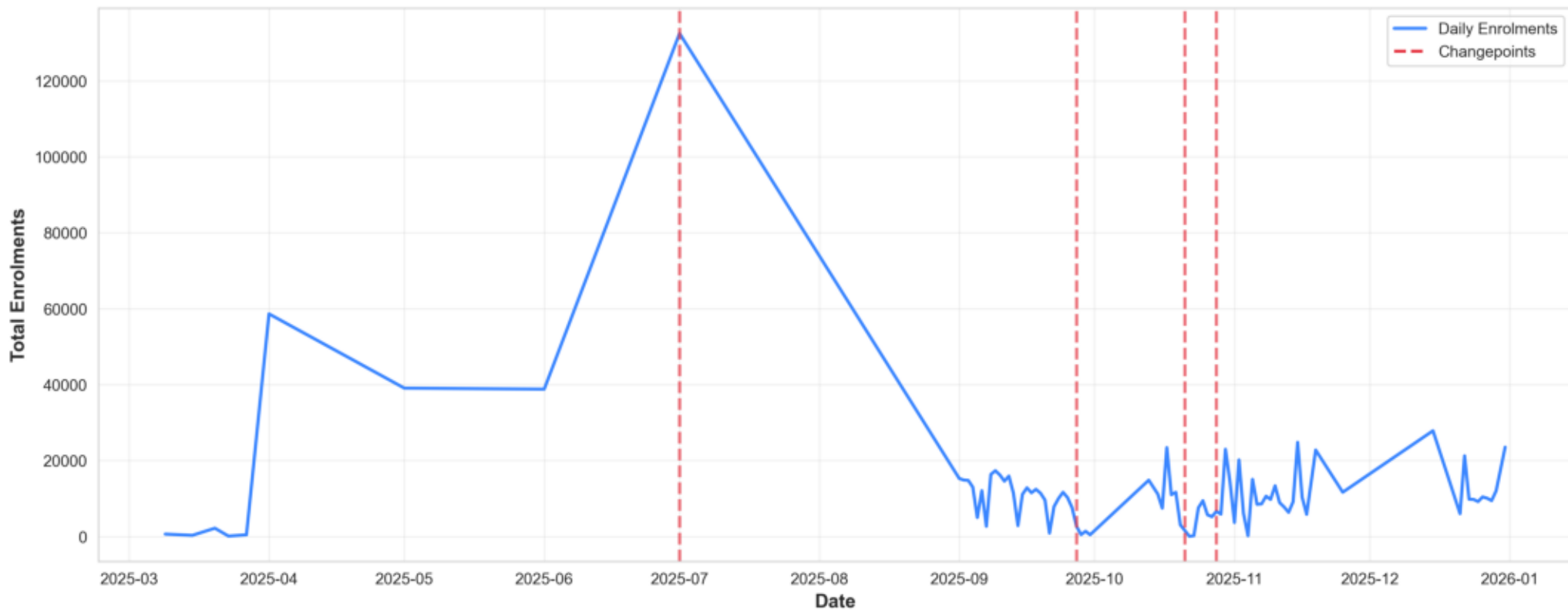
Outlier Detection Method Comparison



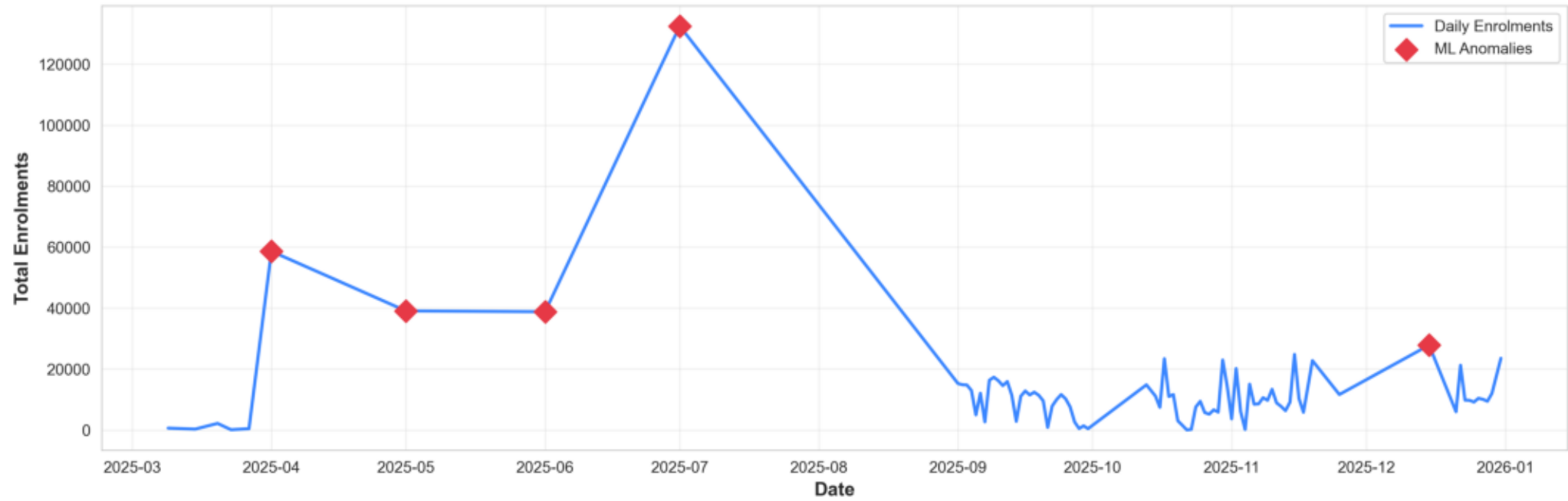
Temporal Anomaly Detection (7-Day Window)



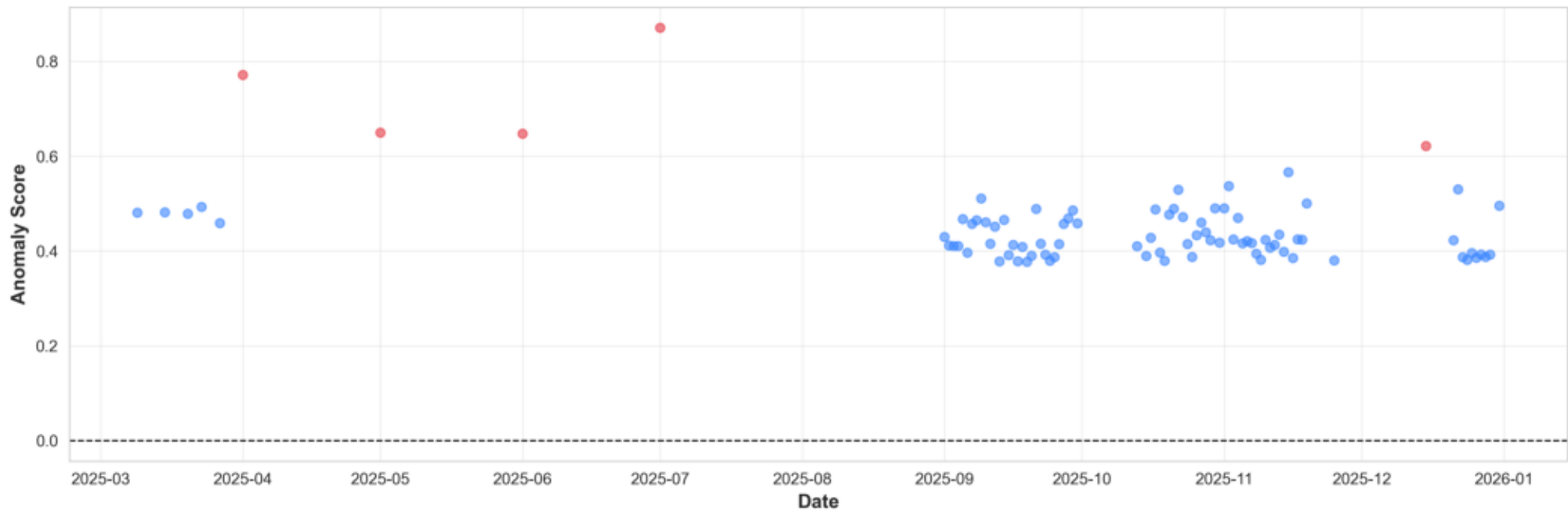
Changepoint Detection: Structural Breaks in Time Series



Isolation Forest Anomaly Detection

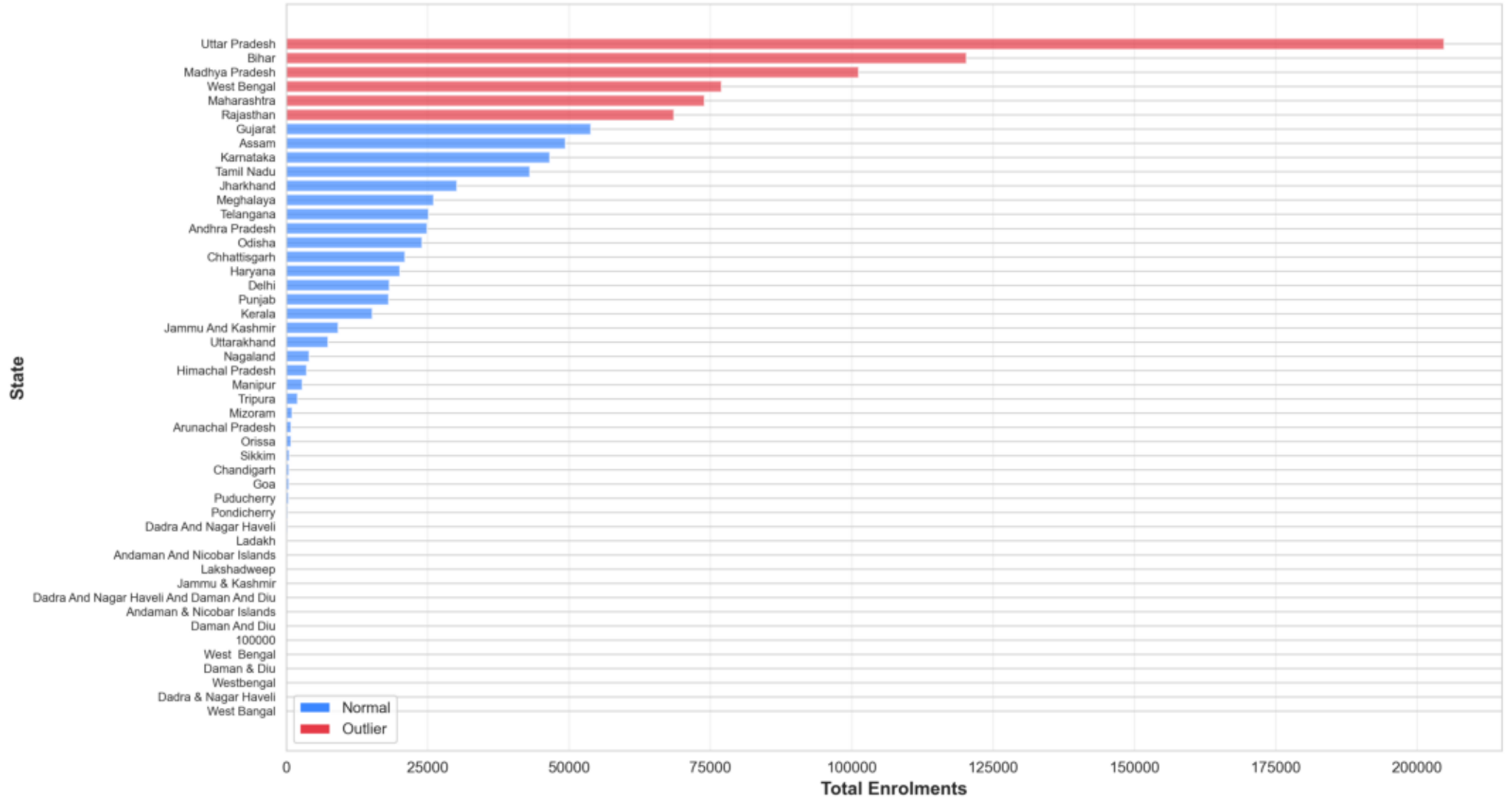


Isolation Forest Anomaly Scores (Higher = More Anomalous)

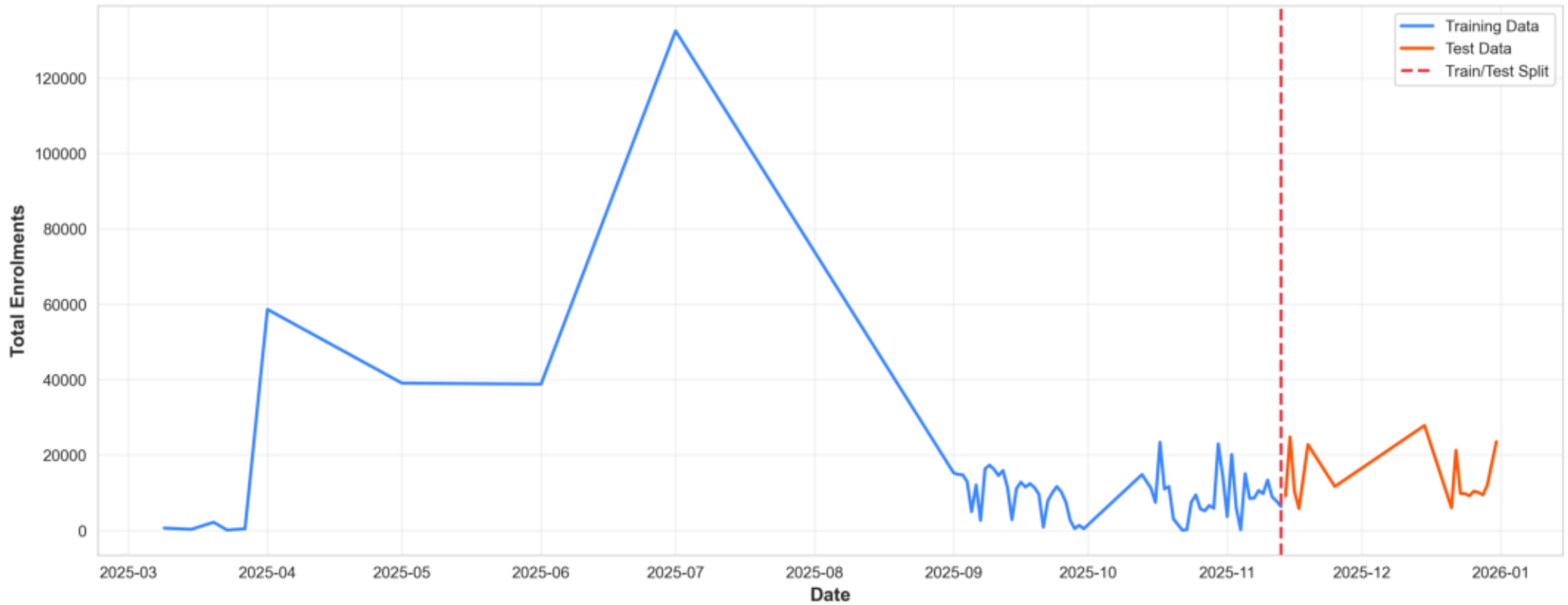


State-Level Anomalies: Enrolment Distribution

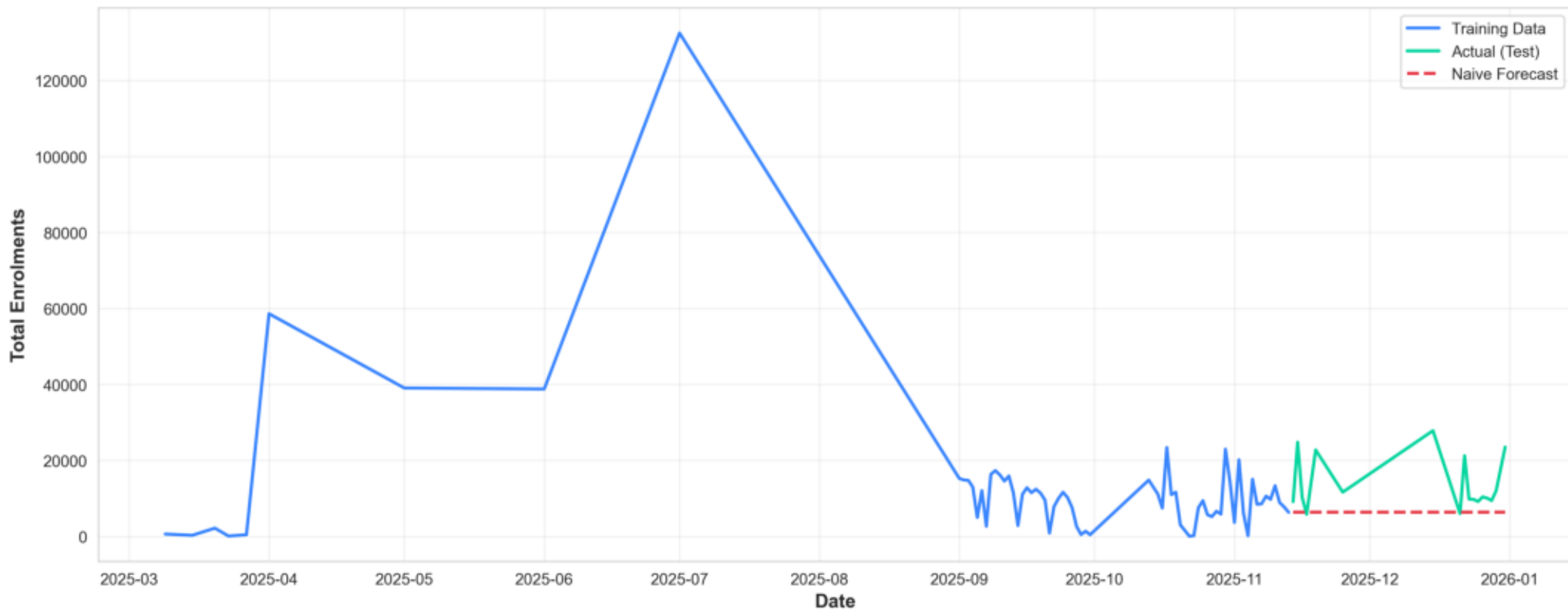
State



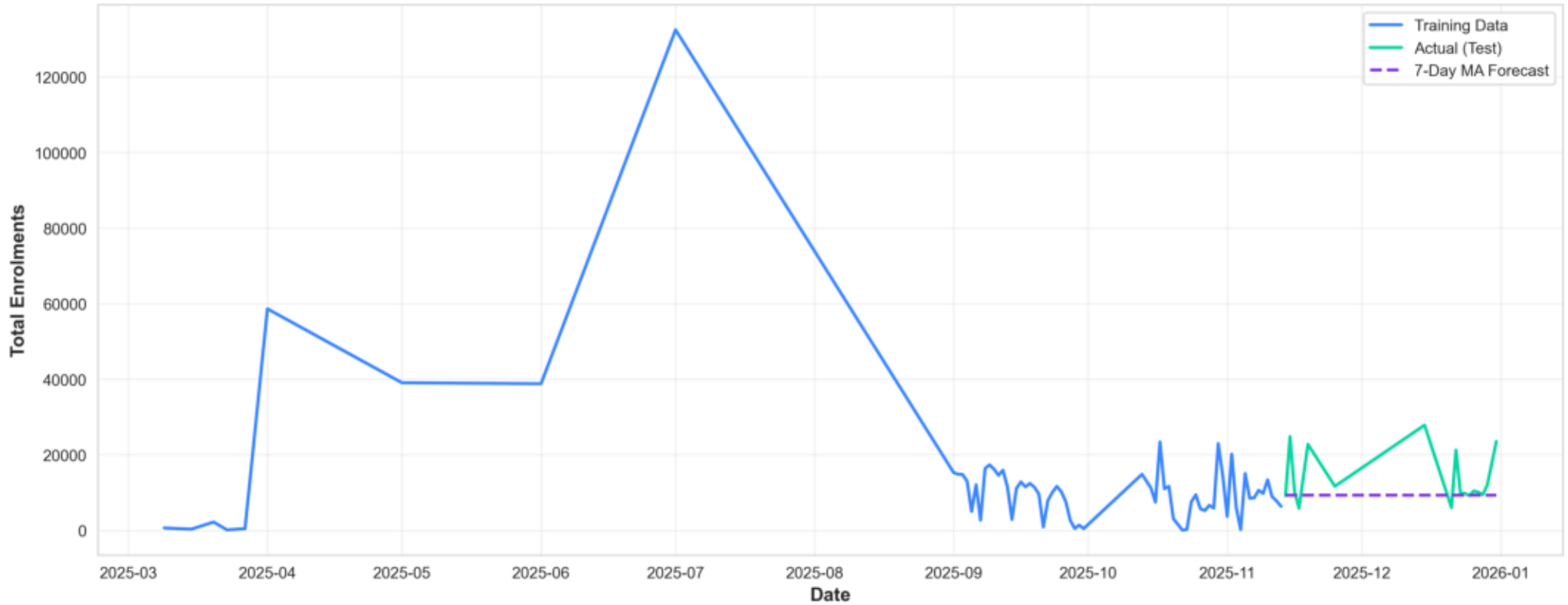
Time Series: Train/Test Split



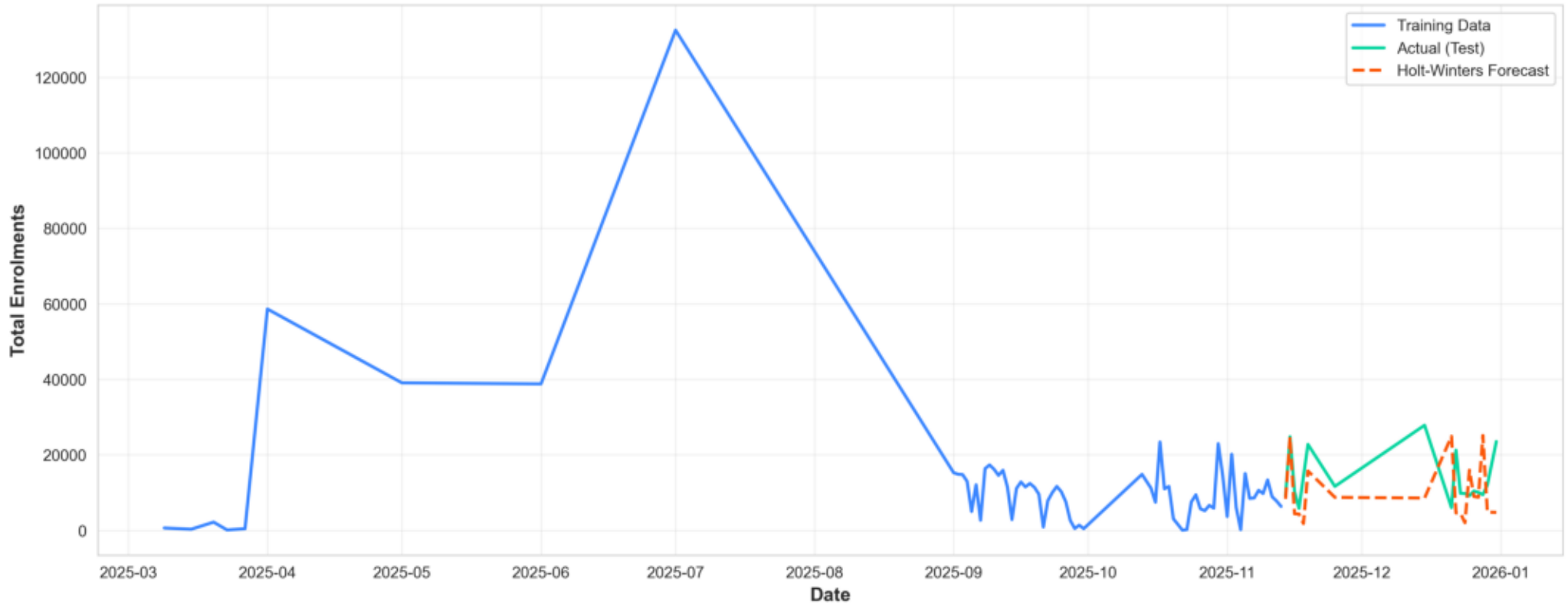
Naive Forecast vs Actual



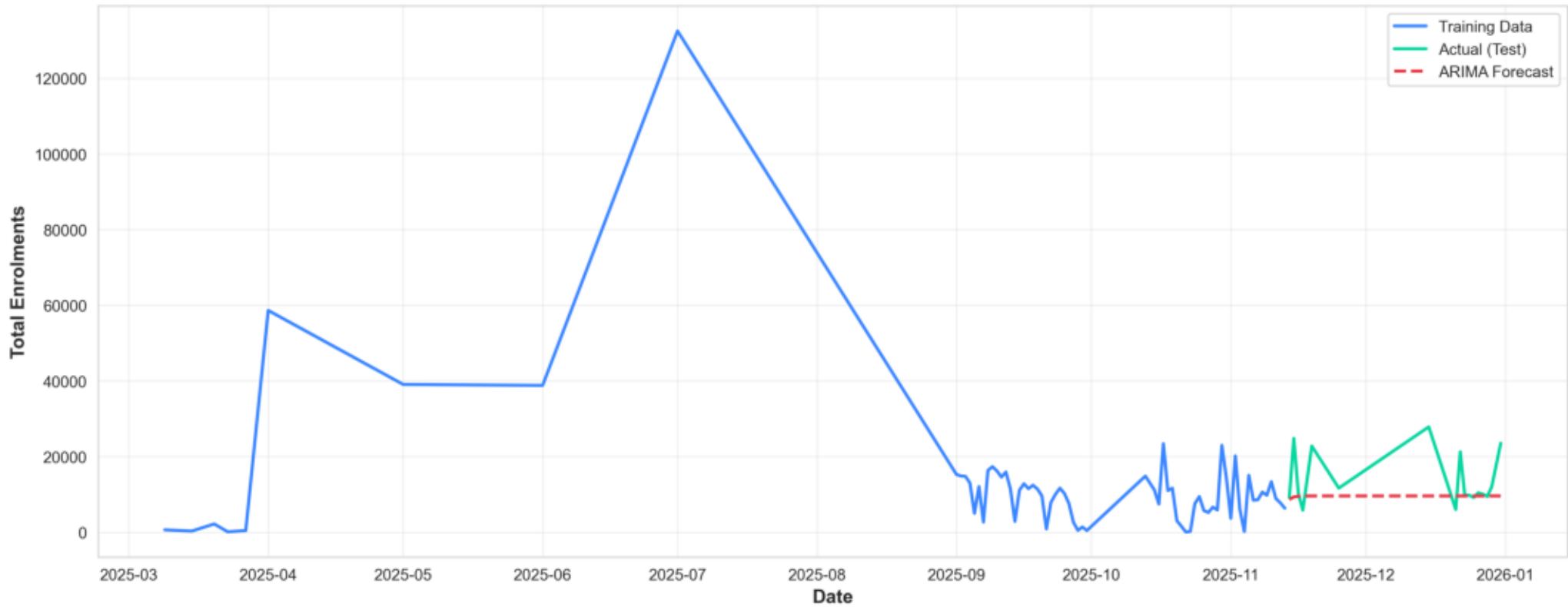
Moving Average Forecast vs Actual

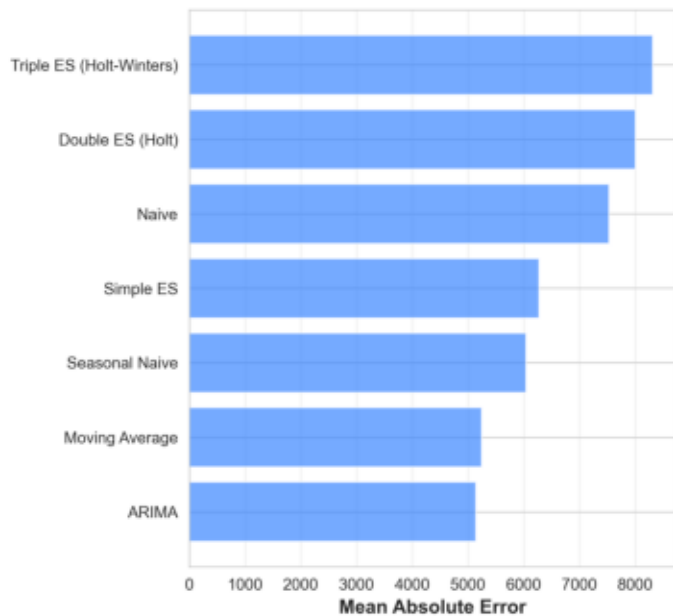
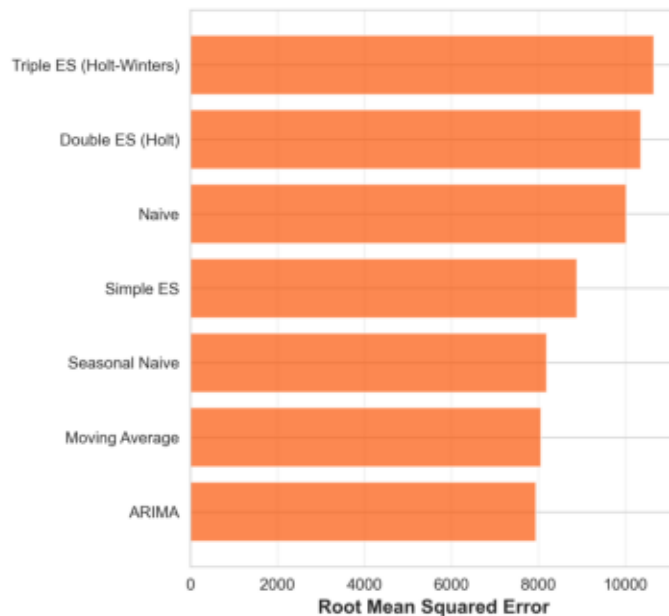
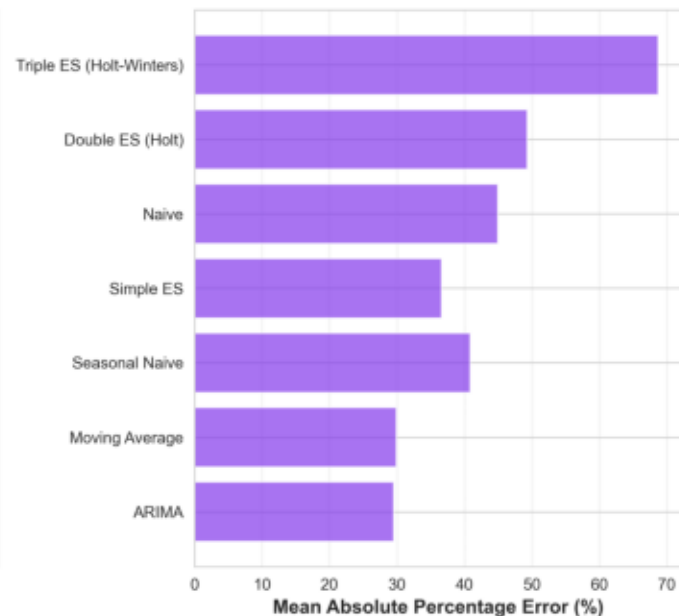


Holt-Winters Forecast with Trend & Seasonality

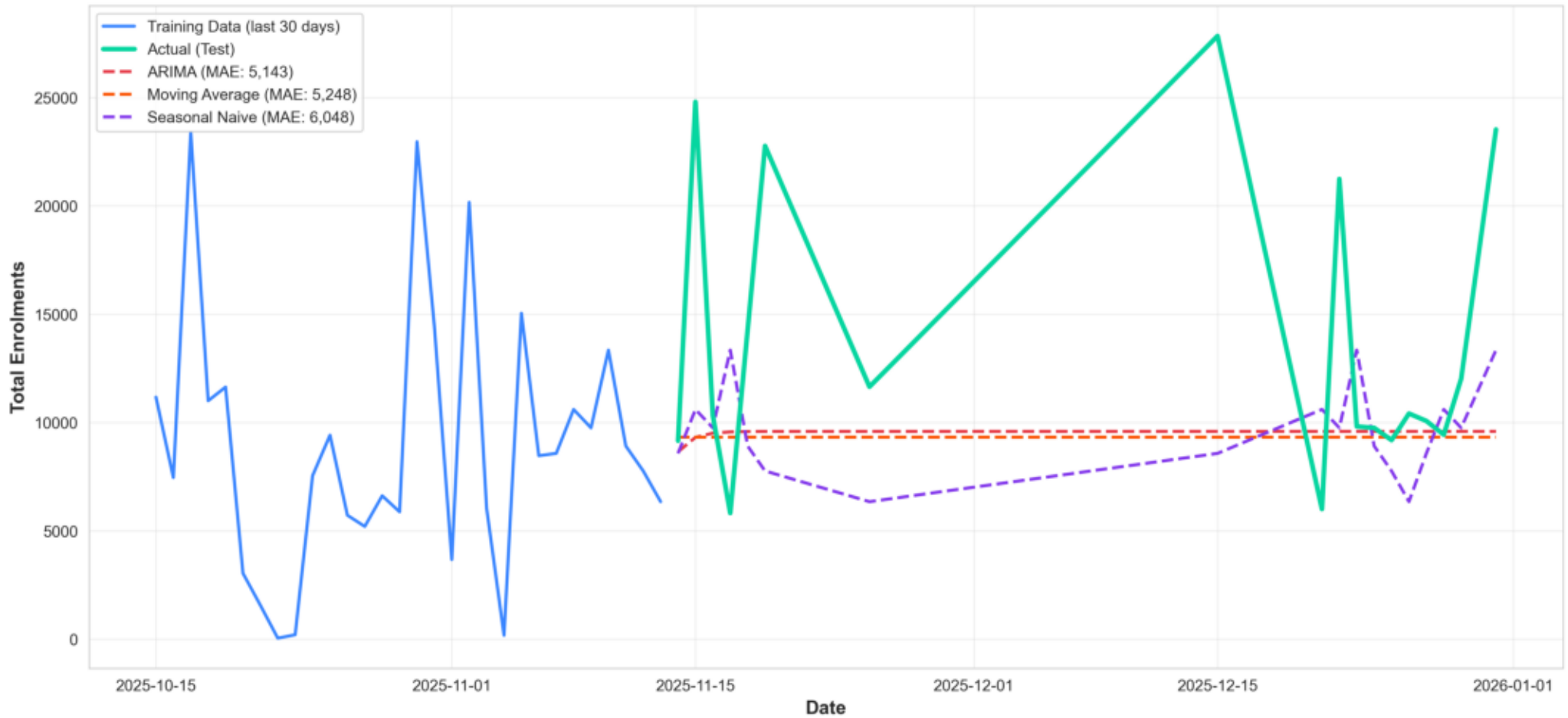


ARIMA(1,1,1) Forecast

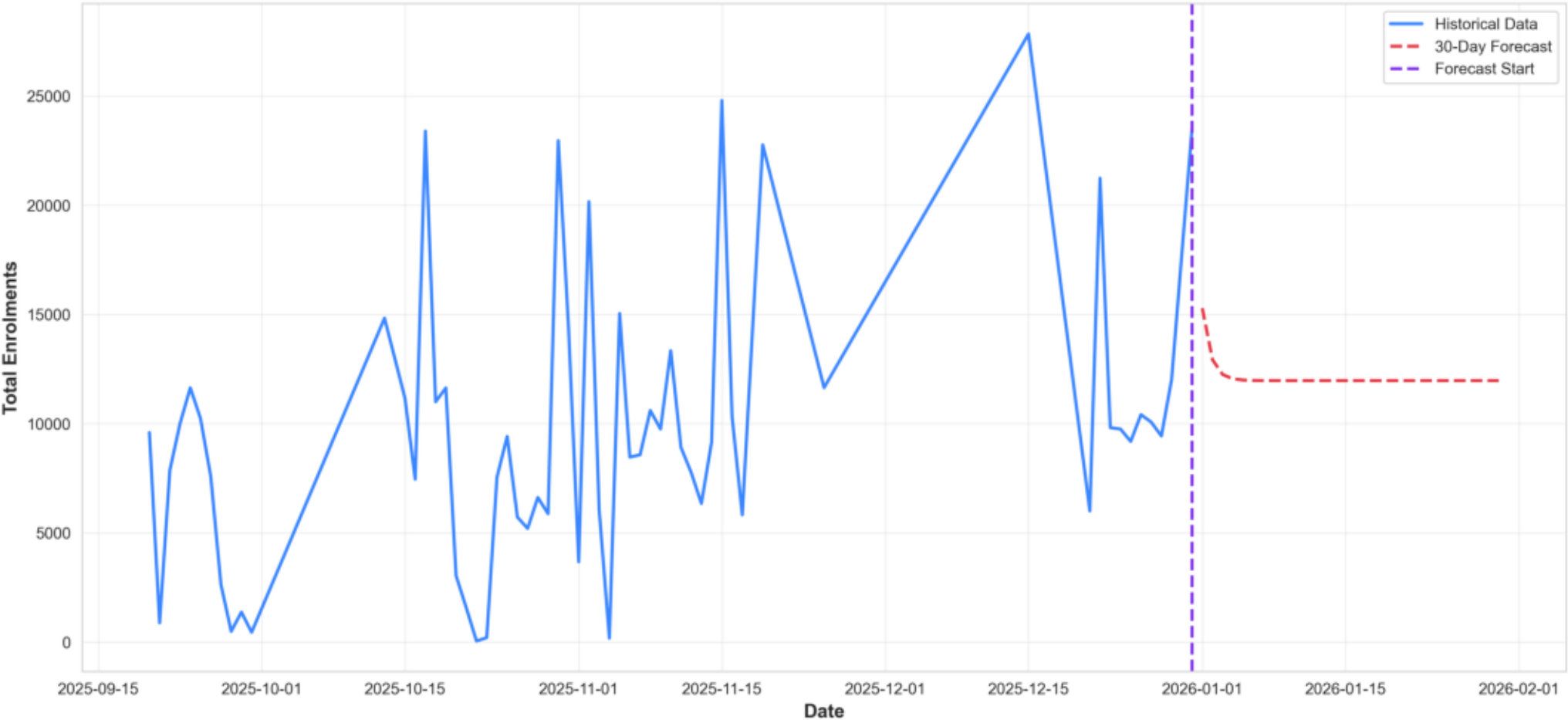


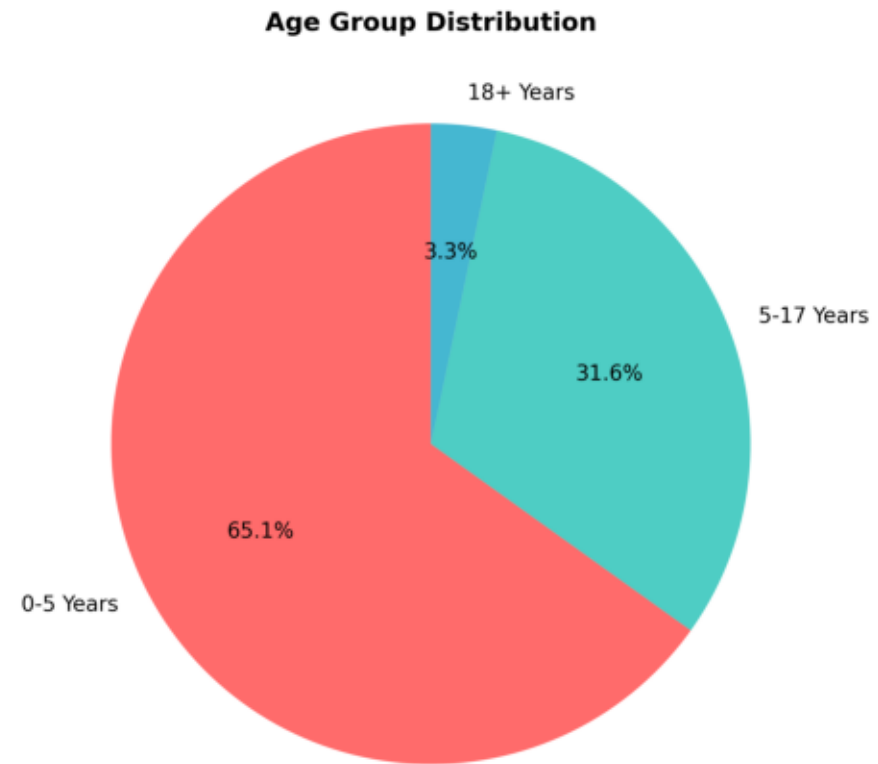
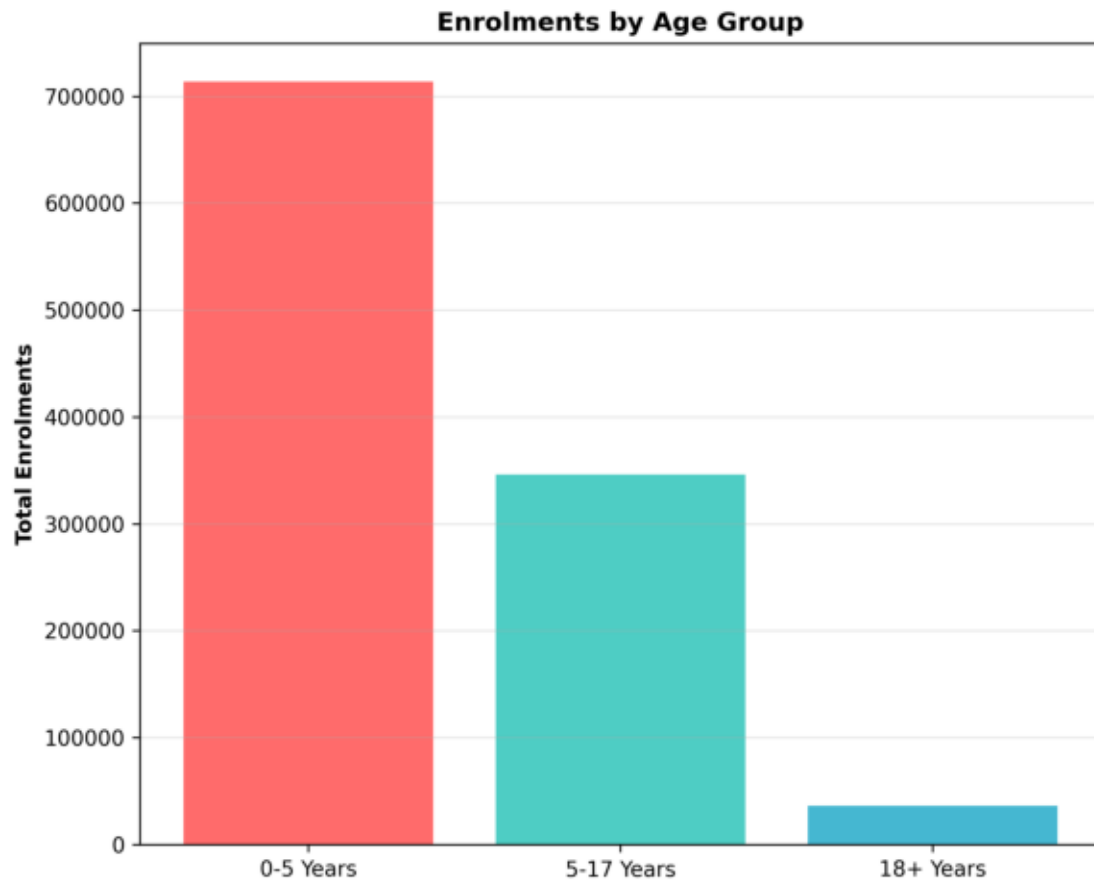
MAE Comparison**RMSE Comparison****MAPE Comparison**

Top 3 Models: Forecast Comparison

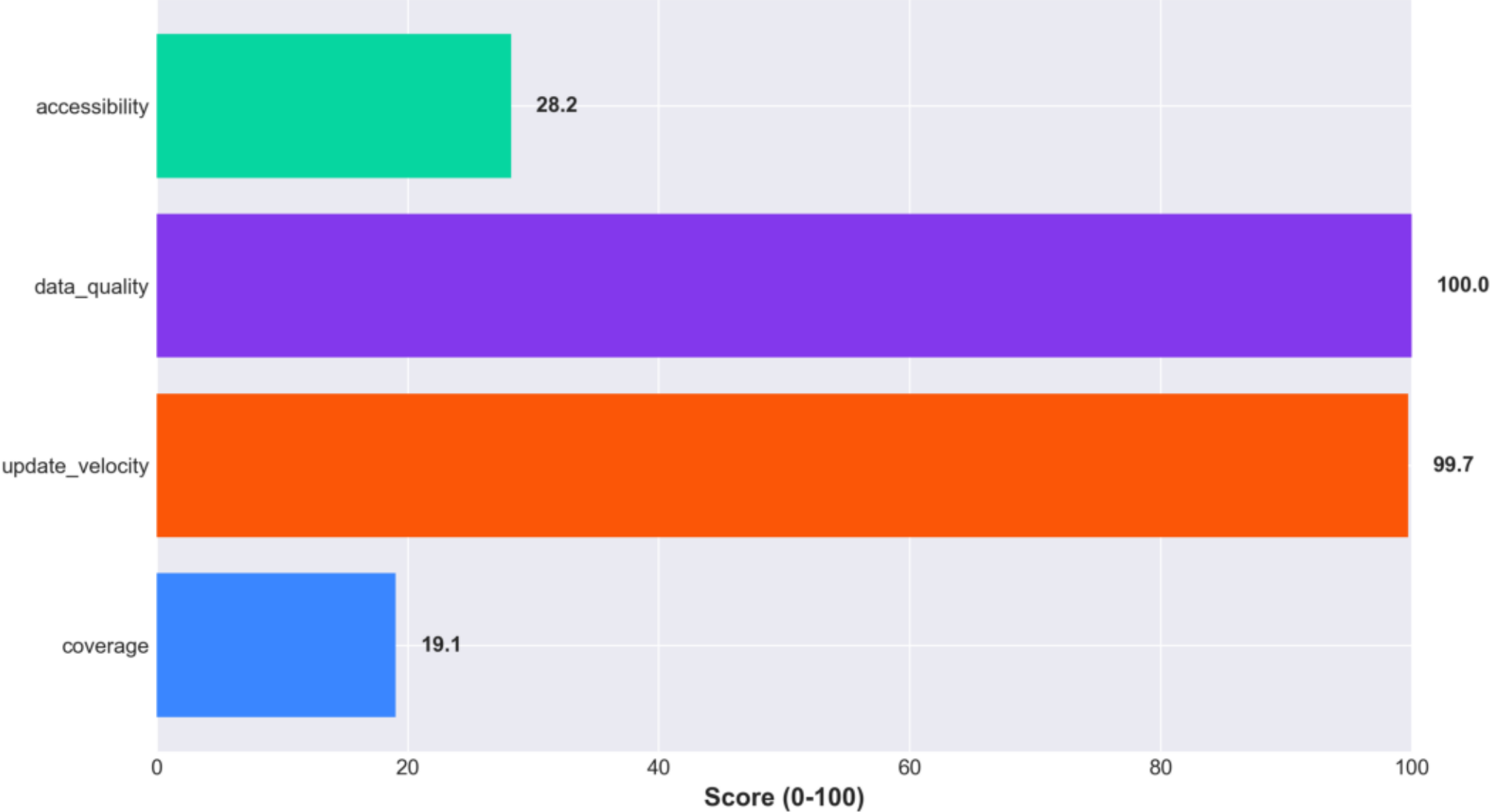


30-Day Enrolment Forecast

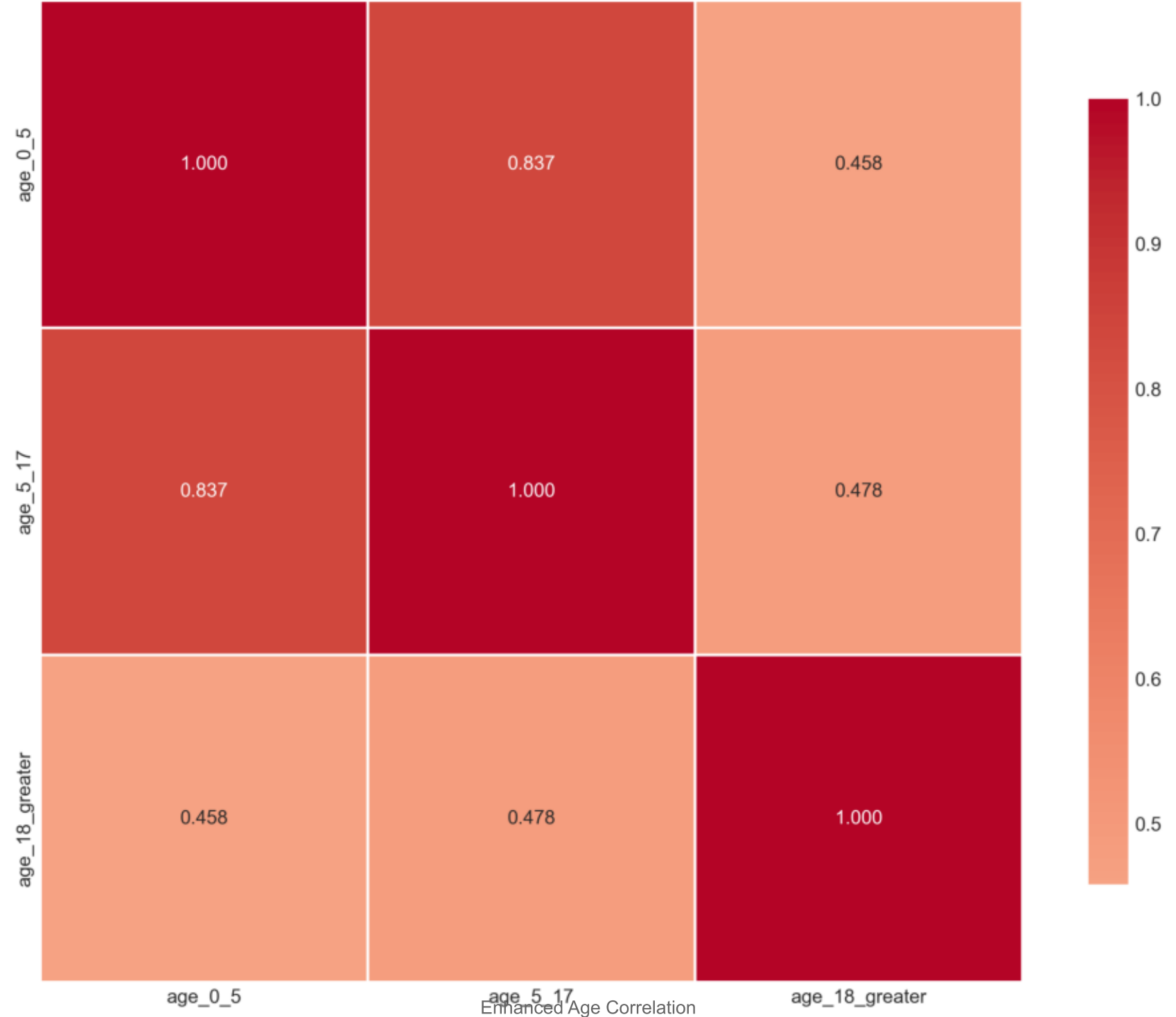




Aadhaar Health Index - Component Scores



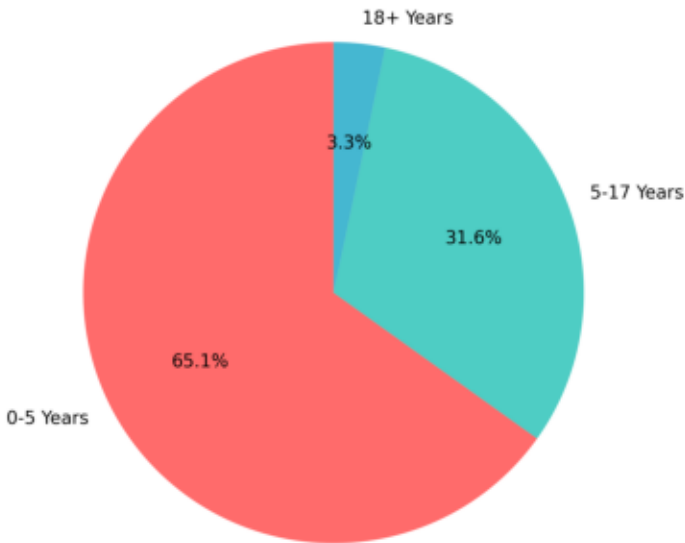
Age Group Correlation Matrix



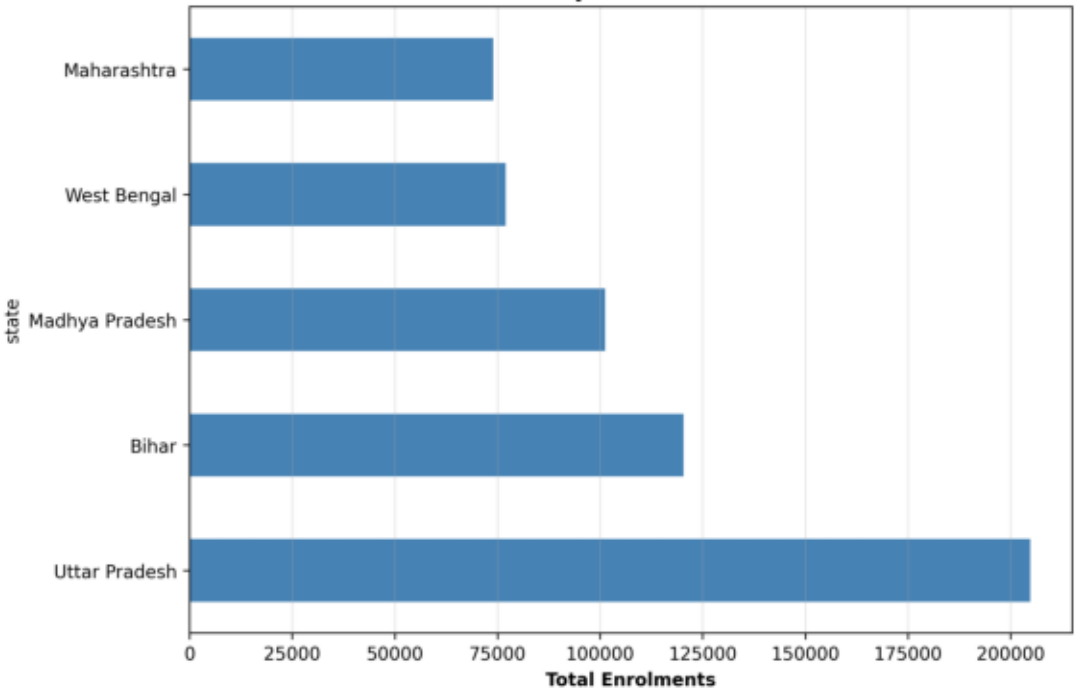
Enhanced Age Correlation

UIDAI Data Analysis - Key Visualizations

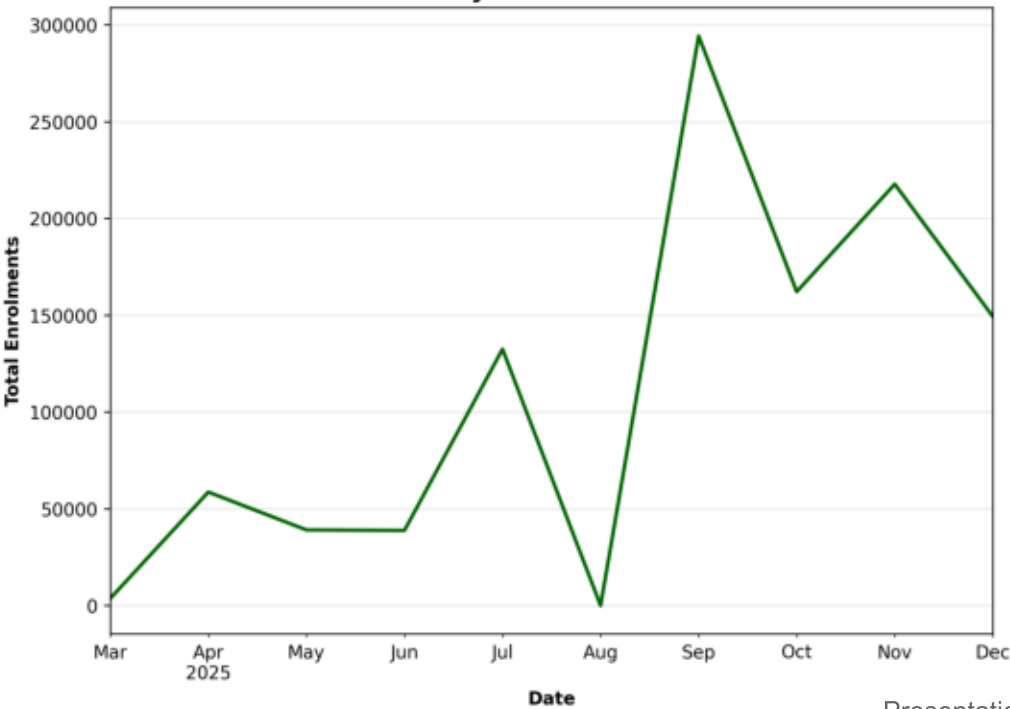
Age Distribution



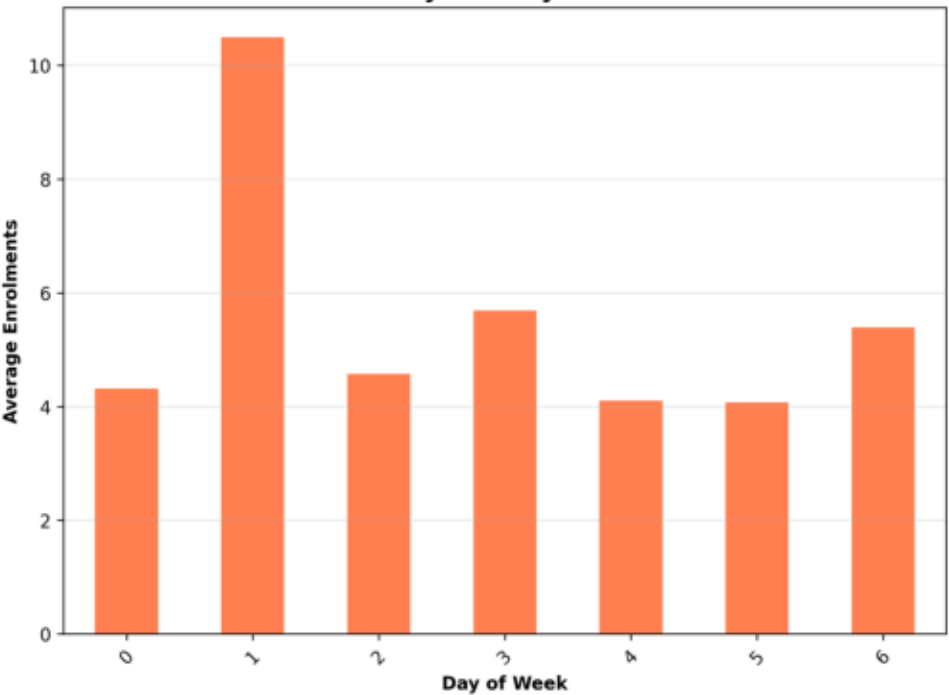
Top 5 States



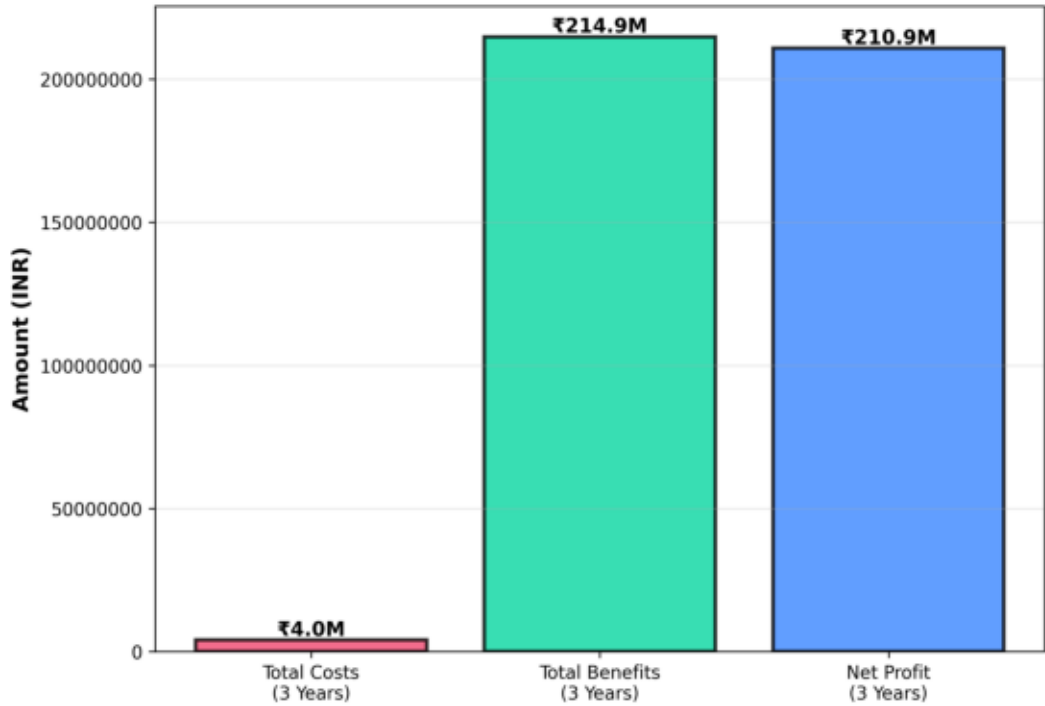
Monthly Enrolment Trend



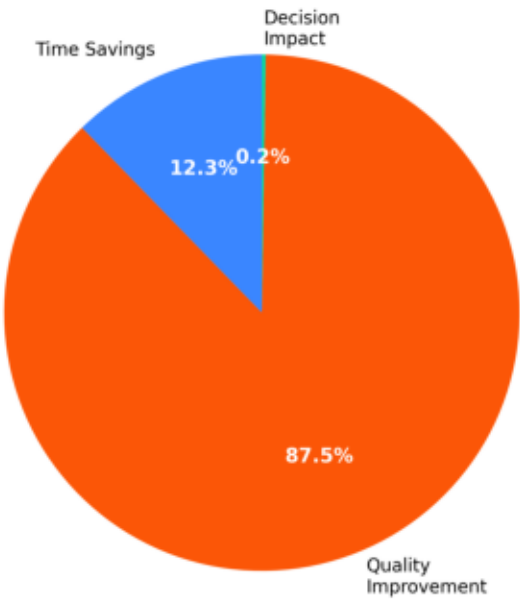
Weekly Activity Pattern



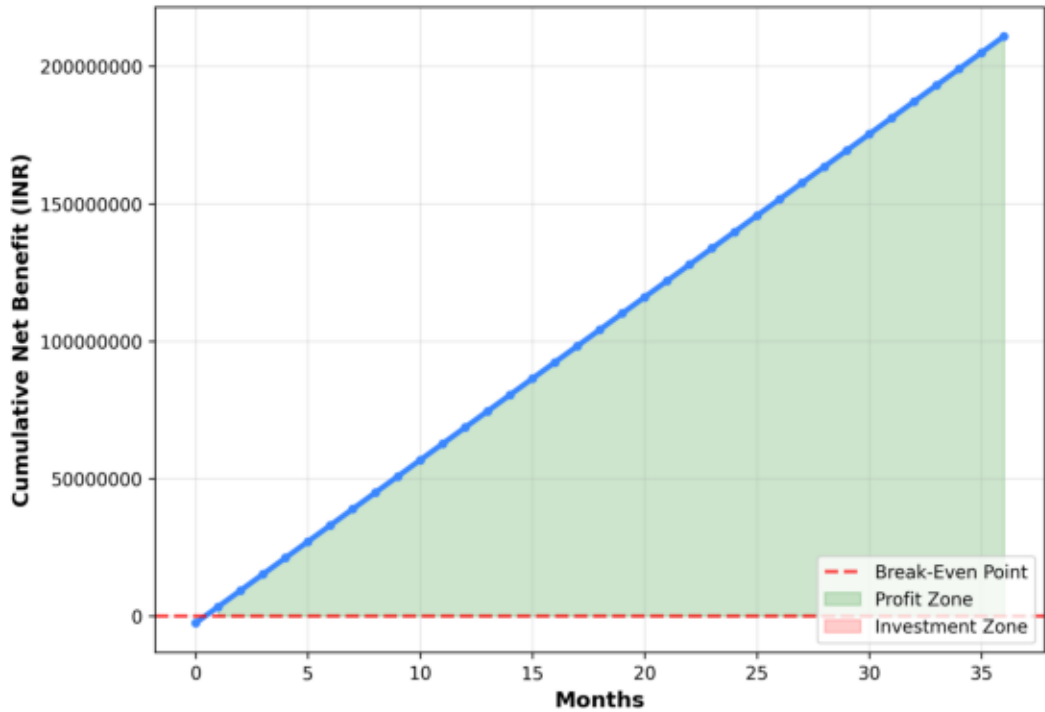
3-Year Financial Overview



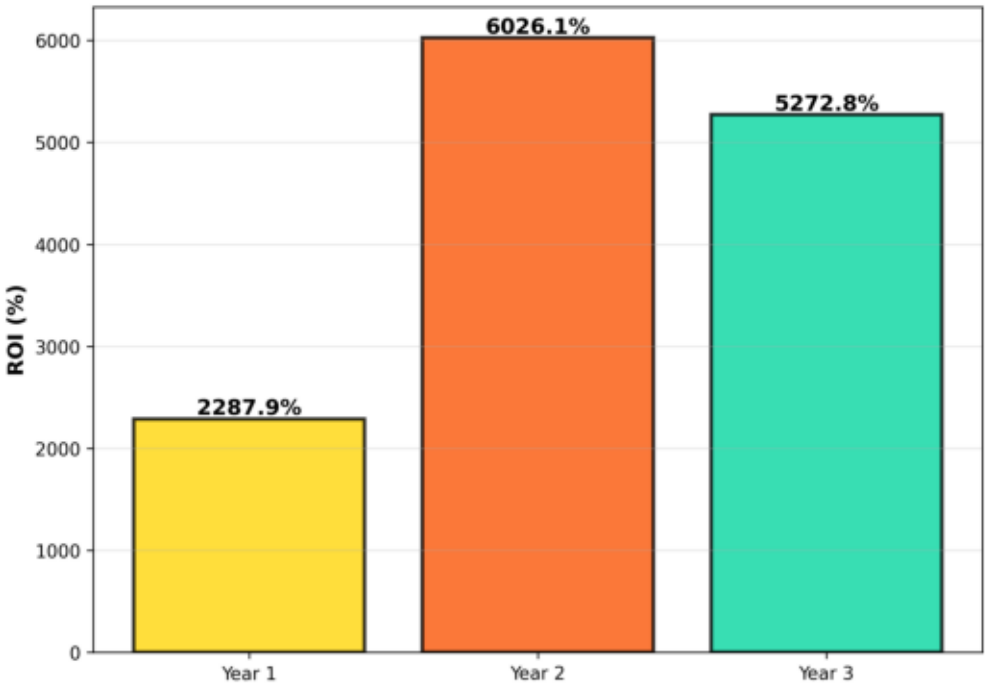
Annual Benefits Breakdown



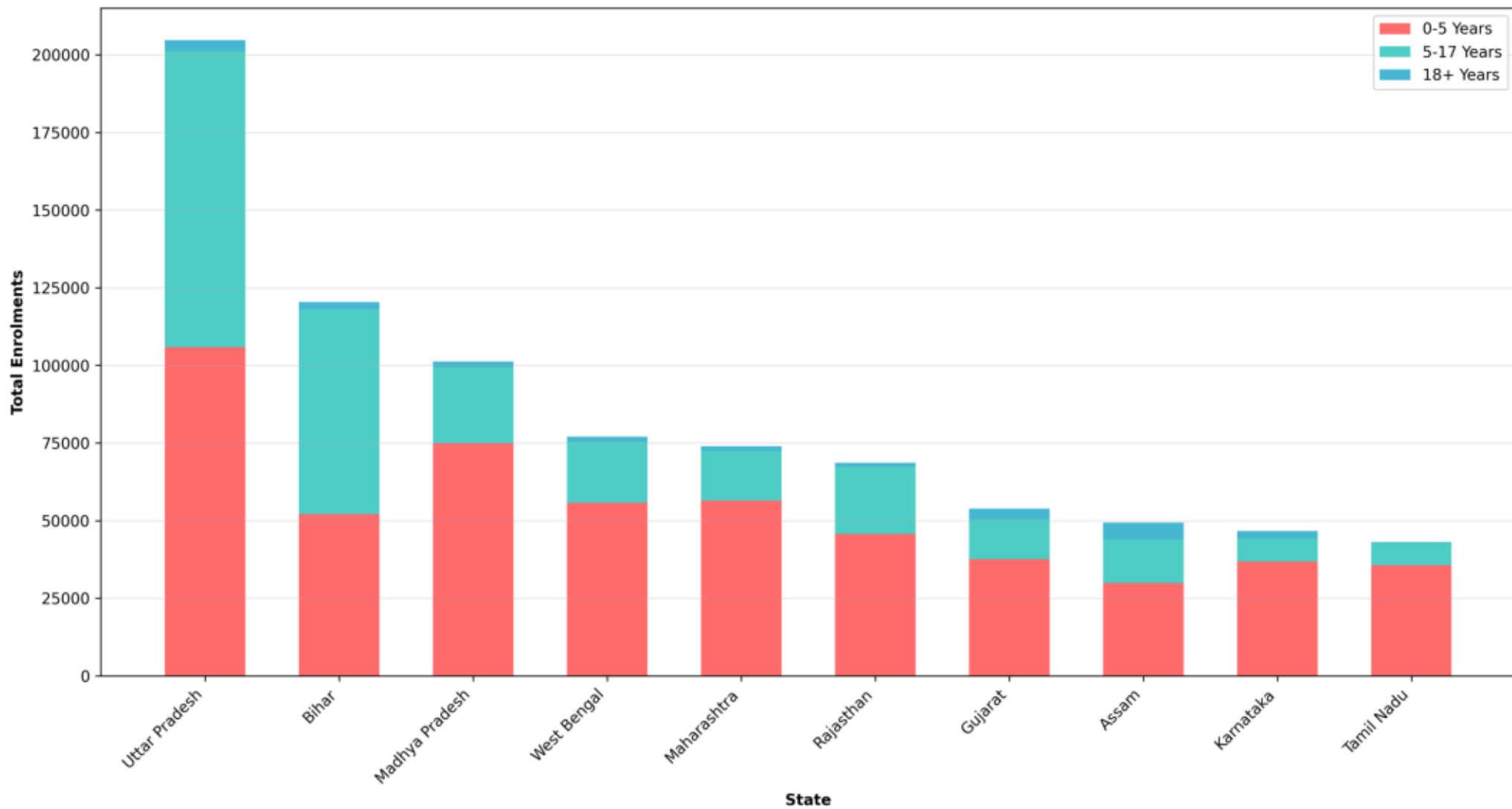
Payback Period Timeline (36 Months)



ROI Progression (Year-over-Year)



Top 10 States - Age Group Distribution



Top 10 States by Enrolment

