

```
!pip install pandas numpy matplotlib seaborn plotly scipy scikit-learn

Requirement already satisfied: pandas in c:\users\sl\anaconda3\lib\site-packages (2.3.3)
Requirement already satisfied: numpy in c:\users\sl\anaconda3\lib\site-packages (2.3.5)
Requirement already satisfied: matplotlib in c:\users\sl\anaconda3\lib\site-packages (3.10.6)
Requirement already satisfied: seaborn in c:\users\sl\anaconda3\lib\site-packages (0.13.2)
Requirement already satisfied: plotly in c:\users\sl\anaconda3\lib\site-packages (6.3.0)
Requirement already satisfied: scipy in c:\users\sl\anaconda3\lib\site-packages (1.16.3)
Requirement already satisfied: scikit-learn in c:\users\sl\anaconda3\lib\site-packages (1.7.2)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\sl\anaconda3\lib\site-packages (from pandas) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in c:\users\sl\anaconda3\lib\site-packages (from pandas) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in c:\users\sl\anaconda3\lib\site-packages (from pandas) (2025.2)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\sl\anaconda3\lib\site-packages (from matplotlib) (1.3.3)
Requirement already satisfied: cycler>=0.10 in c:\users\sl\anaconda3\lib\site-packages (from matplotlib) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\sl\anaconda3\lib\site-packages (from matplotlib) (4.60.1)
Requirement already satisfied: kiwisolver>=1.3.1 in c:\users\sl\anaconda3\lib\site-packages (from matplotlib) (1.4.9)
Requirement already satisfied: packaging>=20.0 in c:\users\sl\anaconda3\lib\site-packages (from matplotlib) (25.0)
Requirement already satisfied: pillow>=8 in c:\users\sl\anaconda3\lib\site-packages (from matplotlib) (12.0.0)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\sl\anaconda3\lib\site-packages (from matplotlib) (3.2.5)
Requirement already satisfied: narwhals>=1.15.1 in c:\users\sl\anaconda3\lib\site-packages (from plotly) (2.7.0)
Requirement already satisfied: joblib>=1.2.0 in c:\users\sl\anaconda3\lib\site-packages (from scikit-learn) (1.5.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in c:\users\sl\anaconda3\lib\site-packages (from scikit-learn) (3.5.0)
Requirement already satisfied: six>=1.5 in c:\users\sl\anaconda3\lib\site-packages (from python-dateutil>=2.8.2->pandas) (1.17.0)

import pandas as pd
import numpy as np

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

```
import plotly.express as px
import plotly.graph_objects as go

from scipy import stats
from sklearn.ensemble import IsolationForest

import os
os.listdir()

['.ipynb_checkpoints',
 'Analysis.ipynb',
 'api_data_aadhar_biometric',
 'api_data_aadhar_demographic',
 'api_data_aadhar_enrolment']

os.listdir('api_data_aadhar_biometric')

['api_data_aadhar_biometric_0_500000.csv',
 'api_data_aadhar_biometric_1000000_1500000.csv',
 'api_data_aadhar_biometric_1500000_1861108.csv',
 'api_data_aadhar_biometric_500000_1000000.csv']

os.listdir('api_data_aadhar_demographic')

['api_data_aadhar_demographic_0_500000.csv',
 'api_data_aadhar_demographic_1000000_1500000.csv',
 'api_data_aadhar_demographic_1500000_2000000.csv',
 'api_data_aadhar_demographic_2000000_2071700.csv',
 'api_data_aadhar_demographic_500000_1000000.csv']

os.listdir('api_data_aadhar_enrolment')

['api_data_aadhar_enrolment_0_500000.csv',
 'api_data_aadhar_enrolment_1000000_1006029.csv',
 'api_data_aadhar_enrolment_500000_1000000.csv']

import pandas as pd

files = os.listdir('api_data_aadhar_biometric')

biometric = pd.concat(
    [pd.read_csv(f'api_data_aadhar_biometric/{f}') for f in files],
    ignore_index=True
)

biometric.head()
```

```

          date      state    district  pincode bio_age_5_17
\ 0 01-03-2025      Haryana Mahendragarh  123029      280
1 01-03-2025        Bihar   Madhepura  852121      144
2 01-03-2025 Jammu and Kashmir       Punch  185101      643
3 01-03-2025        Bihar   Bhojpur  802158      256
4 01-03-2025      Tamil Nadu  Madurai  625514      271

  bio_age_17_
0 577
1 369
2 1091
3 980
4 815

files = os.listdir('api_data_aadhar_enrolment')

enrolment = pd.concat(
    [pd.read_csv(f'api_data_aadhar_enrolment/{f}') for f in files],
    ignore_index=True
)

enrolment.head()

          date      state    district  pincode age_0_5
age_5_17 \
0 02-03-2025      Meghalaya East Khasi Hills  793121      11
61
1 09-03-2025      Karnataka Bengaluru Urban  560043      14
33
2 09-03-2025      Uttar Pradesh  Kanpur Nagar  208001      29
82
3 09-03-2025      Uttar Pradesh       Aligarh  202133      62
29
4 09-03-2025      Karnataka Bengaluru Urban  560016      14
16

  age_18_greater
0            37
1            39
2            12
3            15
4            21

files = os.listdir('api_data_aadhar_demographic')

```

```

demographic = pd.concat(
    [pd.read_csv(f'api_data_aadhar_demographic/{f}') for f in files],
    ignore_index=True
)

demographic.head()

      date      state   district  pincode  demo_age_5_17 \
0  01-03-2025  Uttar Pradesh  Gorakhpur    273213          49
1  01-03-2025  Andhra Pradesh  Chittoor    517132          22
2  01-03-2025        Gujarat    Rajkot    360006          65
3  01-03-2025  Andhra Pradesh  Srikakulam   532484          24
4  01-03-2025  Rajasthan     Udaipur    313801          45

      demo_age_17_
0            529
1            375
2            765
3            314
4            785

print("Enrolment:", enrolment.shape)
print("Demographic:", demographic.shape)
print("Biometric:", biometric.shape)

Enrolment: (1006029, 7)
Demographic: (2071700, 6)
Biometric: (1861108, 6)

enrolment.columns = enrolment.columns.str.strip().str.lower()
demographic.columns = demographic.columns.str.strip().str.lower()
biometric.columns = biometric.columns.str.strip().str.lower()

enrolment['date'] = pd.to_datetime(enrolment['date'], errors='coerce')
demographic['date'] = pd.to_datetime(demographic['date'],
errors='coerce')
biometric['date'] = pd.to_datetime(biometric['date'], errors='coerce')

print(enrolment.isnull().sum())
print(demographic.isnull().sum())
print(biometric.isnull().sum())

date           682238
state             0
district           0
pincode            0
age_0_5              0
age_5_17              0
age_18_greater         0
dtype: int64

```

```
date           1187968
state            0
district          0
pincode           0
demo_age_5_17      0
demo_age_17_       0
dtype: int64
date           944100
state            0
district          0
pincode           0
bio_age_5_17      0
bio_age_17_       0
dtype: int64

enrolment['date'] = pd.to_datetime(enrolment['date'], errors='coerce')
demographic['date'] = pd.to_datetime(demographic['date'],
errors='coerce')
biometric['date'] = pd.to_datetime(biometric['date'], errors='coerce')

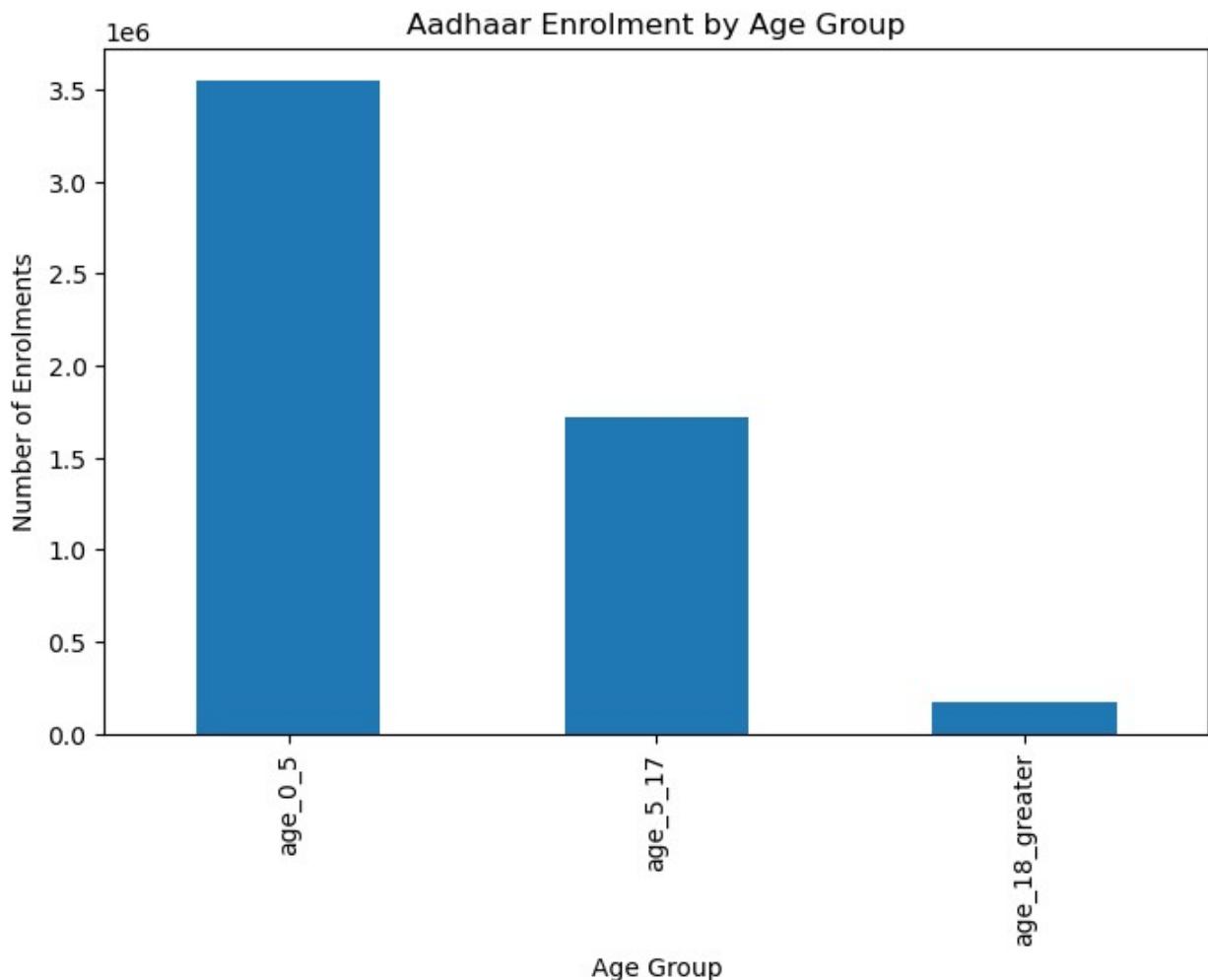
enrolment_full = enrolment.copy()
demographic_full = demographic.copy()
biometric_full = biometric.copy()

enrolment_date = enrolment.dropna(subset=['date'])
demographic_date = demographic.dropna(subset=['date'])
biometric_date = biometric.dropna(subset=['date'])

print(enrolment_date.shape)
print(demographic_date.shape)
print(biometric_date.shape)

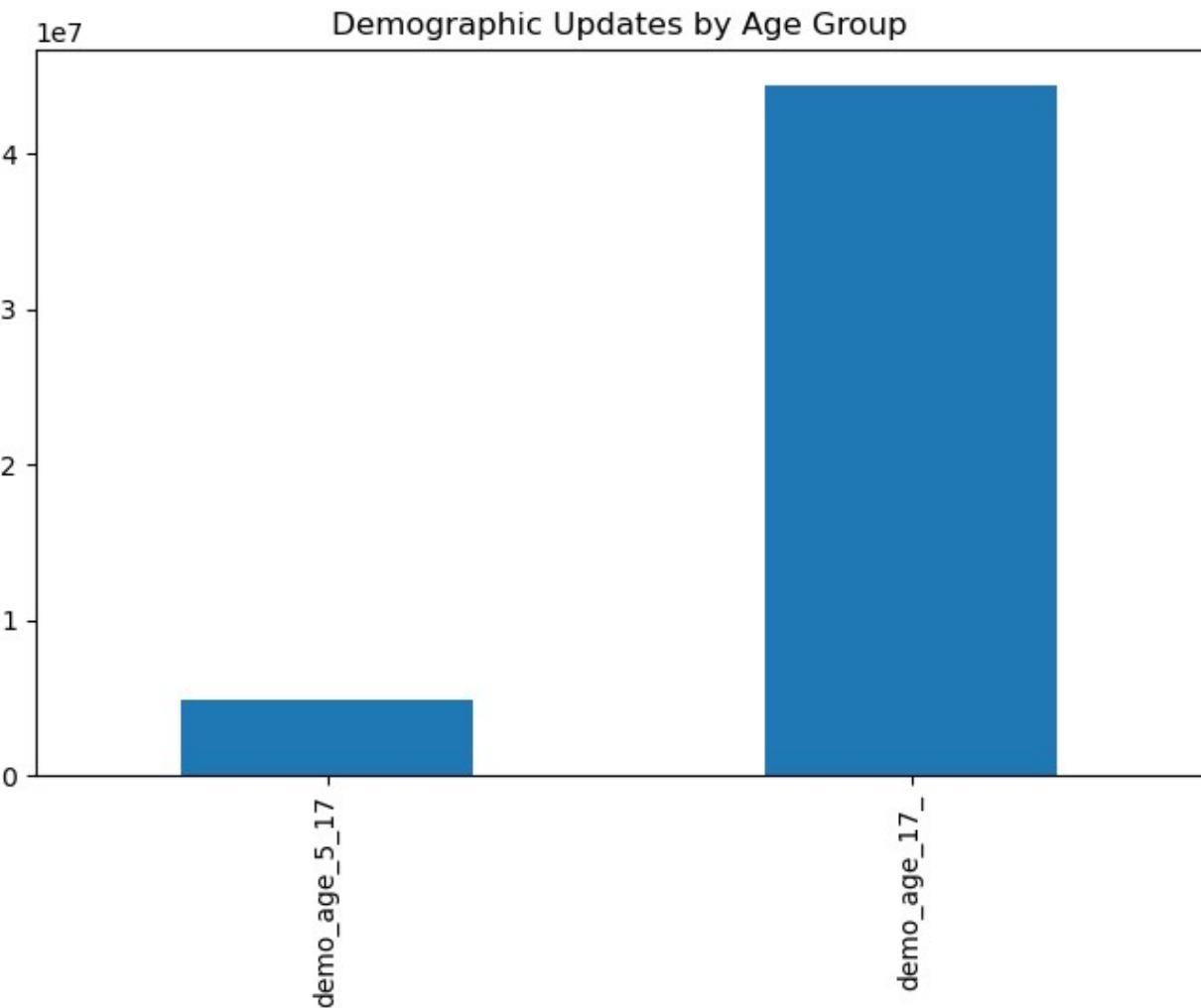
(323791, 7)
(883732, 6)
(917008, 6)

enrolment_full[['age_0_5','age_5_17','age_18_greater']].sum().plot(
    kind='bar',
    title='Aadhaar Enrolment by Age Group',
    figsize=(8,5)
)
plt.xlabel('Age Group')
plt.ylabel('Number of Enrolments')
plt.show()
```



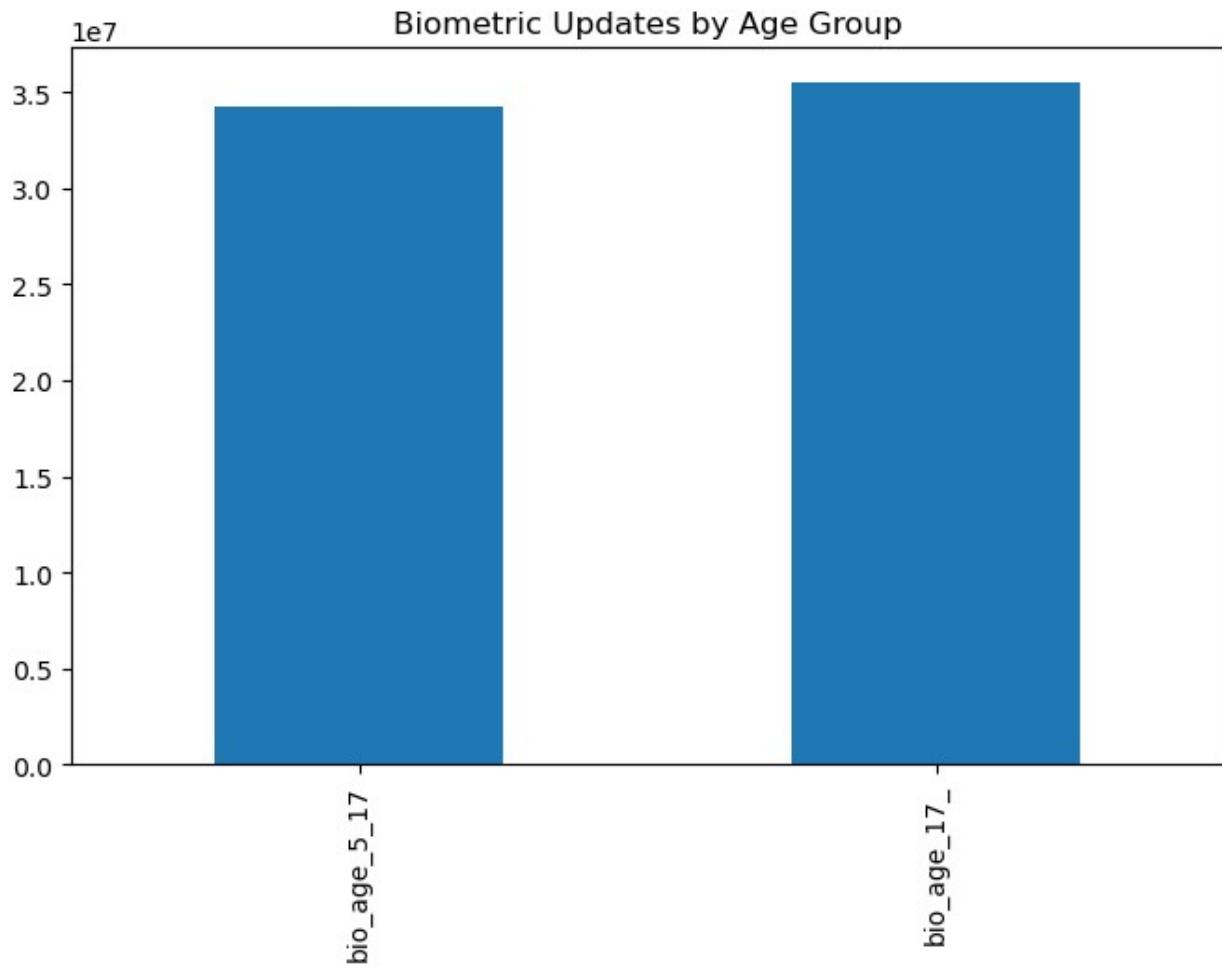
```
demo_cols = ['demo_age_5_17', 'demo_age_17_']

demographic_full[demo_cols].sum().plot(
    kind='bar',
    title='Demographic Updates by Age Group',
    figsize=(8,5)
)
plt.show()
```



```
bio_cols = ['bio_age_5_17','bio_age_17_']

biometric_full[bio_cols].sum().plot(
    kind='bar',
    title='Biometric Updates by Age Group',
    figsize=(8,5)
)
plt.show()
```



```
enrolment_full.groupby('state')
[['age_0_5','age_5_17','age_18_greater']].sum() \
    .sum(axis=1) \
    .sort_values(ascending=False) \
    .head(10)
```

state	
Uttar Pradesh	1018629
Bihar	609585
Madhya Pradesh	493970
West Bengal	375297
Maharashtra	369139
Rajasthan	348458
Gujarat	280549
Assam	230197
Karnataka	223235
Tamil Nadu	220789

dtype: int64

```

demo_state = demographic_full.groupby('state')
[demo_cols].sum().sum(axis=1)
bio_state = biometric_full.groupby('state')
[bio_cols].sum().sum(axis=1)

comparison = pd.DataFrame({
    'demographic_updates': demo_state,
    'biometric_updates': bio_state
})

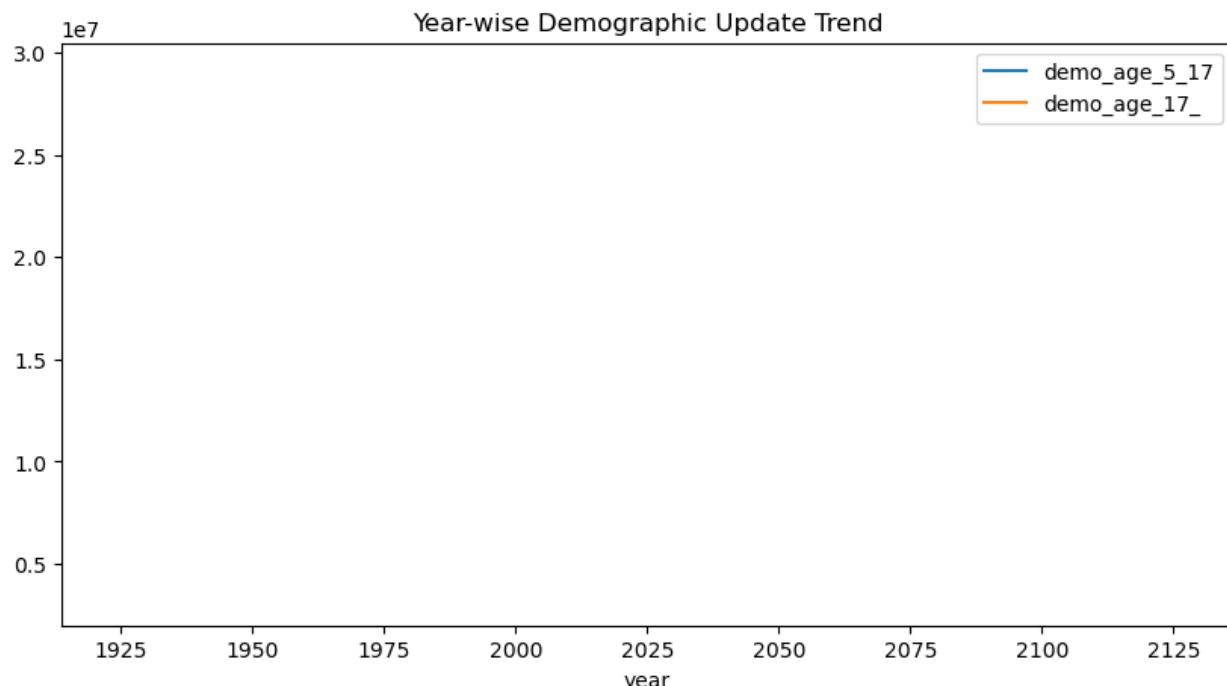
comparison.head()

      demographic_updates  biometric_updates
state
100000                           2.0          NaN
Andaman & Nicobar Islands           1059.0        2384.0
Andaman and Nicobar Islands           6187.0       18314.0
Andhra Pradesh                      2295505.0     3714592.0
Arunachal Pradesh                   36443.0       72394.0

demo_date = demographic_date.copy()
demo_date['year'] = demo_date['date'].dt.year

demo_date.groupby('year')
[['demo_age_5_17','demo_age_17_']].sum().plot(
    title='Year-wise Demographic Update Trend',
    figsize=(10,5)
)
plt.show()

```



```

demo_date['year'].describe()

count    883732.0
mean      2025.0
std        0.0
min       2025.0
25%       2025.0
50%       2025.0
75%       2025.0
max       2025.0
Name: year, dtype: float64

demo_state = demographic_full.groupby('state')
[['demo_age_5_17','demo_age_17_']].sum()

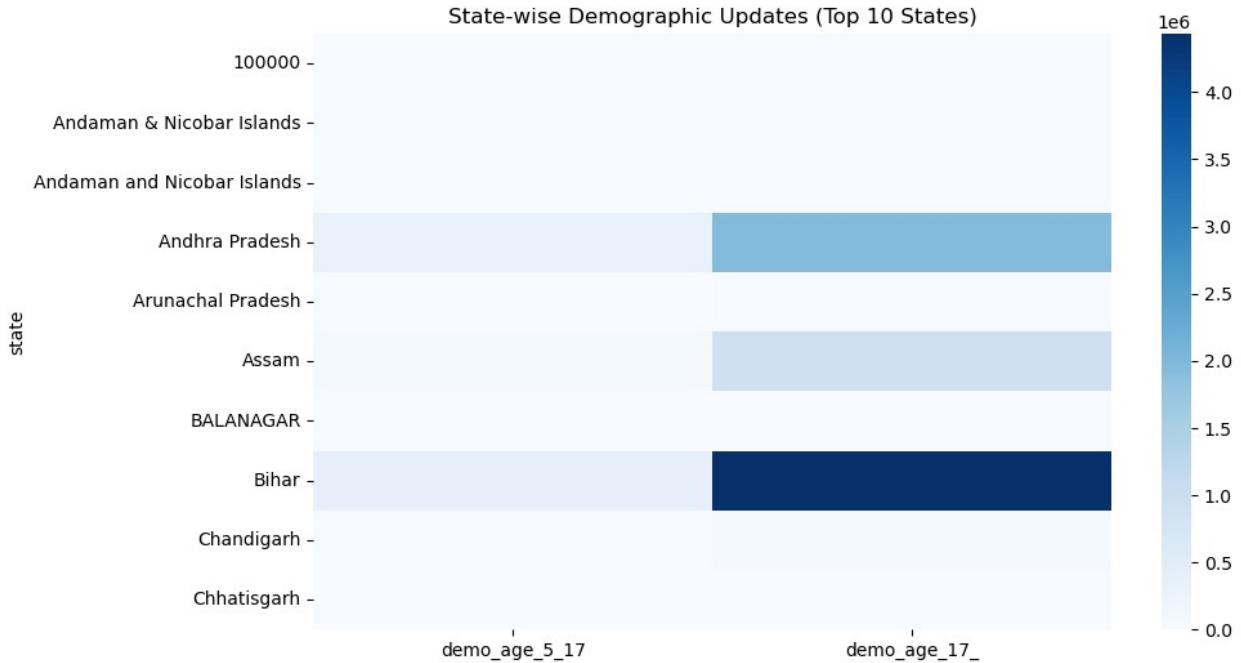
demo_state.sort_values(by='demo_age_17_', ascending=False).head(10)

          demo_age_5_17  demo_age_17_
state
Uttar Pradesh      790308      7752020
Maharashtra        273322      4781280
Bihar              380023      4434327
West Bengal         242549      3629623
Rajasthan           257224      2560391
Madhya Pradesh     407098      2505840
Andhra Pradesh     321143      1974362
Tamil Nadu          315638      1896590
Chhattisgarh        165207      1840227
Gujarat             208474      1615853

heatmap_data = demographic_full.groupby('state')
[demo_cols].sum().head(10)

plt.figure(figsize=(10,6))
sns.heatmap(heatmap_data, cmap='Blues')
plt.title('State-wise Demographic Updates (Top 10 States)')
plt.show()

```



```

state_updates = demographic_full.groupby('state')
[demo_cols].sum().sum(axis=1).reset_index(name='updates')

model = IsolationForest(contamination=0.05, random_state=42)
state_updates['anomaly'] =
model.fit_predict(state_updates[['updates']])

state_updates[state_updates['anomaly'] == -1]

      state  updates  anomaly
7       Bihar   4814350      -1
32  Maharashtra   5054602      -1
51  Uttar Pradesh   8542328      -1
58   West Bengal   3872172      -1

enrol_state = enrolment_full.groupby('state')
[['age_0_5', 'age_5_17', 'age_18_greater']].sum().sum(axis=1)
update_state = demographic_full.groupby('state')
[['demo_age_5_17', 'demo_age_17_']].sum().sum(axis=1)

ratio_df = pd.DataFrame({
    'enrolment': enrol_state,
    'updates': update_state
})

ratio_df['update_ratio'] = ratio_df['updates'] / ratio_df['enrolment']
ratio_df.sort_values('update_ratio').head(10)

      enrolment  updates  update_ratio
state

```

100000	218.0	2.0	0.009174
Meghalaya	109771.0	87378.0	0.796003
Nagaland	15587.0	36791.0	2.360364
Jammu & Kashmir	155.0	426.0	2.748387
Assam	230197.0	1012578.0	4.398745
Lakshadweep	203.0	1176.0	5.793103
Pondicherry	1272.0	7459.0	5.863994
Madhya Pradesh	493970.0	2912938.0	5.896994
Gujarat	280549.0	1824327.0	6.502704
Orissa	4149.0	28758.0	6.931309