```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
```

```
demo_files = [
    'api_data_aadhar_demographic_0_500000.csv',
    'api_data_aadhar_demographic_500000_1000000.csv',
    'api_data_aadhar_demographic_1000000_1500000.csv',
    'api_data_aadhar_demographic_1500000_2000000.csv',
    'api_data_aadhar_demographic_2000000_2071700.csv'
]
```

```
df = pd.concat([pd.read_csv(f) for f in demo_files], ignore_index=True)
df['date'] = pd.to_datetime(df['date'], dayfirst=True)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2071700 entries, 0 to 2071699
Data columns (total 6 columns):
 #   Column         Dtype
---  ------         -----
 0   date           datetime64[ns]
 1   state          object
 2   district       object
 3   pincode        int64
 4   demo_age_5_17  int64
 5   demo_age_17_   int64
dtypes: datetime64[ns](1), int64(3), object(2)
memory usage: 94.8+ MB
```

```
df.columns
```

```
Index(['date', 'state', 'district', 'pincode', 'demo_age_5_17',
       'demo_age_17_'],
      dtype='object')
```

```
df.shape
```

```
(2071700, 6)
```

```
total_kids = df['demo_age_5_17'].sum()
total_adults = df['demo_age_17_'].sum()
total_updates = total_kids + total_adults
```

```
pincode_data = df.groupby(['pincode', 'state', 'district']).agg({
    'demo_age_5_17': 'sum',
    'demo_age_17_': 'sum'
}).reset_index()
pincode_data['total'] = pincode_data['demo_age_5_17'] + pincode_data['demo_age_17_']
```

```
scaler = StandardScaler()
features = pincode_data[['demo_age_5_17', 'demo_age_17_']]
scaled_features = scaler.fit_transform(features)

kmeans = KMeans(n_clusters=3, random_state=42, n_init=10)
pincode_data['cluster'] = kmeans.fit_predict(scaled_features)
```

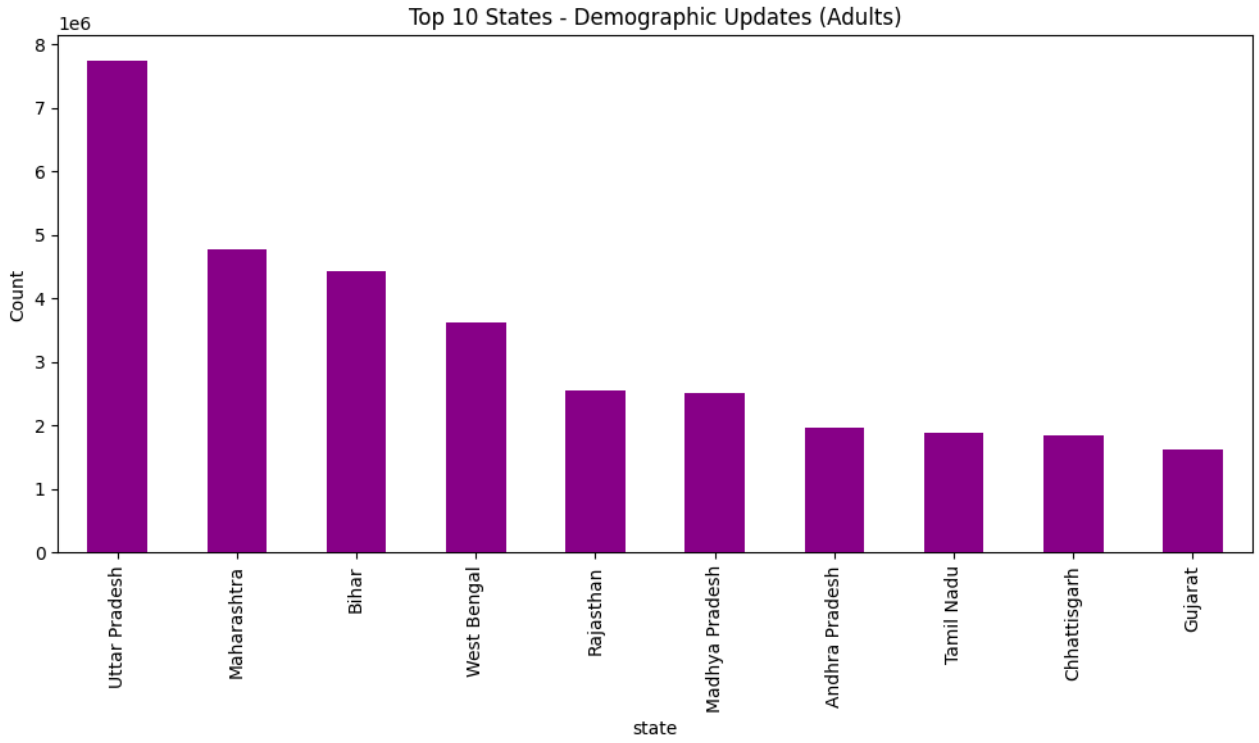
```
cluster_map = {
    pincode_data.groupby('cluster')['total'].mean().idxmin(): 'Low Activity',
    pincode_data.groupby('cluster')['total'].mean().idxmax(): 'High Activity (Hotspots)'
}
```

```
all_ids = {0, 1, 2}
middle_id = list(all_ids - set(cluster_map.keys()))[0]
cluster_map[middle_id] = 'Medium Activity'
pincode_data['Activity_Category'] = pincode_data['cluster'].map(cluster_map)
```

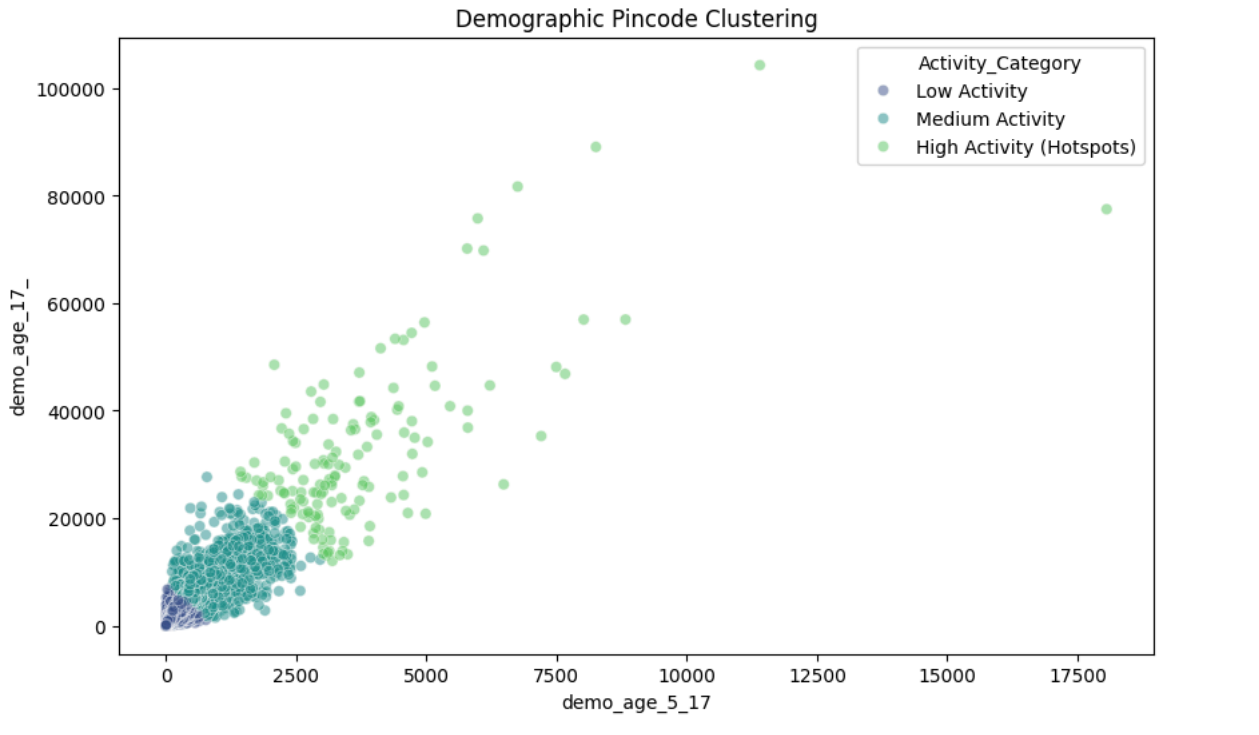
```
pincode_data['zone_digit'] = pincode_data['pincode'].astype(str).str[0]
zone_map = {
    '1': 'North-1', '2': 'North-2 (UP/UK)', '3': 'West-1 (RJ/GJ)',
    '4': 'West-2/Central', '5': 'South-1 (AP/KA)', '6': 'South-2 (TN/KL)',
    '7': 'East-1', '8': 'East-2 (BR/JH)', '9': 'Army'
```

```
}
pincode_data['Zone'] = pincode_data['zone_digit'].map(zone_map)
pincode_data.to_csv('final_demographic_clustering.csv', index=False)
```

```
plt.figure(figsize=(10, 6))
top_states = df.groupby('state')['demo_age_17_'].sum().nlargest(10)
top_states.plot(kind='bar', color='darkmagenta')
plt.title('Top 10 States - Demographic Updates (Adults)')
plt.ylabel('Count')
plt.tight_layout()
plt.savefig('demo_top_states.png')
```




```
plt.figure(figsize=(10, 6))
sns.scatterplot(data=pincode_data, x='demo_age_5_17', y='demo_age_17_', hue='Activity_Category', palette='viridis', alpha=(
plt.title('Demographic Pincode Clustering')
plt.savefig('demo_clustering_scatter.png')
```




```
plt.figure(figsize=(6, 6))
plt.pie([total_kids, total_adults], labels=['Kids (5-17)', 'Adults (17+)'], autopct='%1.1f%%', colors=['#ffcc99','#66b3ff'],
```

```
plt.title('Demographic Age Distribution')
plt.savefig('demo_age_split.png')

print(f"Total Records: {len(df)}")
print(f"Total Updates: {total_updates}")
print(pincode_data['Activity_Category'].value_counts())
```

```
Total Records: 2071700
Total Updates: 49295187
Activity_Category
Low Activity                28914
Medium Activity              2448
High Activity (Hotspots)      146
Name: count, dtype: int64
```