Project code

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

# Set aesthetics for plots

sns.set\_theme(style="whitegrid", palette="muted")

plt.rcParams.update({'figure.autolayout': True})

# Load the dataset

file\_path = r"C:\Users\USER\Downloads\realtime\_data.csv"

df = pd.read\_csv(file\_path)

# -------------------------

# 1. DATA CLEANING & VISUALIZATION

# -------------------------

# Check for missing values

print("Missing values per column:\n", df.isnull().sum())

# Convert date

df['lastUpdated'] = pd.to\_datetime(df['lastUpdated'], format='%d-%m-%Y', errors='coerce')

# Drop duplicates

df = df.drop\_duplicates()

# Total beneficiaries by state

state\_beneficiaries = df.groupby('state\_name')['total\_beneficiaries'].sum().reset\_index().sort\_values(by='total\_beneficiaries', ascending=False)

# Bar plot - Total Beneficiaries by State

plt.figure(figsize=(14, 6))

#sns.barplot(data=state\_beneficiaries, x='state\_name', y='total\_beneficiaries', palette='Spectral', dodge=False)

sns.barplot(data=state\_beneficiaries, x='state\_name', y='total\_beneficiaries', hue='state\_name', palette='Spectral', dodge=False, legend=False)

plt.xticks(rotation=90)

plt.title("Total Beneficiaries by State")

plt.xlabel("State")

plt.ylabel("Total Beneficiaries")

plt.tight\_layout()

plt.show()

# -------------------------

# 2. EDA & STATISTICAL ANALYSIS

# -------------------------

# Correlation Matrix Heatmap

plt.figure(figsize=(10, 7))

correlation = df[['total\_beneficiaries', 'total\_aadhar', 'total\_mobileno', 'total\_sc', 'total\_st', 'total\_gen', 'total\_obc']].corr()

sns.heatmap(correlation, annot=True, cmap='coolwarm', linewidths=0.5)

plt.title("Correlation Heatmap")

plt.tight\_layout()

plt.show()

# Scatterplot - Aadhaar vs Beneficiaries

plt.figure(figsize=(10, 6))

sns.scatterplot(data=df, x='total\_aadhar', y='total\_beneficiaries', hue='state\_name', palette='tab20', alpha=0.7)

plt.title("Aadhaar Linked vs Total Beneficiaries")

plt.xlabel("Aadhaar Linked")

plt.ylabel("Total Beneficiaries")

plt.legend(loc='center left', bbox\_to\_anchor=(1, 0.5), fontsize='small')

plt.tight\_layout()

plt.show()

# Correlation between aadhar and beneficiaries

cor\_val = df['total\_aadhar'].corr(df['total\_beneficiaries'])

print(f"\nCorrelation between Total Aadhar and Total Beneficiaries: {cor\_val}")

# -------------------------

# 3. CREATIVITY & INNOVATION

# -------------------------

# Inclusion Score = average of Aadhaar + Mobile coverage

df['inclusion\_score'] = (df['total\_aadhar'] + df['total\_mobileno']) / (2 \* df['total\_beneficiaries'])

df['inclusion\_score'] = df['inclusion\_score'].clip(upper=1.0)

# Top 5 inclusive districts

top\_districts = df.groupby('district\_name')['inclusion\_score'].mean().reset\_index().sort\_values(by='inclusion\_score', ascending=False).head(5)

print("\nTop 5 Districts by Inclusion Score:\n", top\_districts)

# Top 5 inclusive states

top\_states = df.groupby('state\_name')['inclusion\_score'].mean().reset\_index().sort\_values(by='inclusion\_score', ascending=False).head(5)

print("\nTop 5 States by Inclusion Score:\n", top\_states)

# Plot top 5 states by Inclusion Score

'''plt.figure(figsize=(10, 5))

sns.barplot(data=top\_states, x='state\_name', y='inclusion\_score', palette='coolwarm', dodge=False)

plt.title("Top 5 States by Inclusion Score")

plt.ylabel("Inclusion Score")

plt.xlabel("State")

plt.tight\_layout()

plt.show()'''

# Trend in scheme reporting over time

monthly\_report = df.copy()

monthly\_report['month'] = monthly\_report['lastUpdated'].dt.to\_period('M')

monthly\_trend = monthly\_report.groupby('month').size().reset\_index(name='report\_count')

plt.figure(figsize=(10, 5))

sns.lineplot(data=monthly\_trend, x='month', y='report\_count', marker='o', color='purple')

plt.title("Monthly Scheme Data Reporting Trend")

plt.xlabel("Month")

plt.ylabel("Reports Count")

plt.xticks(rotation=45)

plt.tight\_layout()

plt.show()

# District-wise beneficiary analysis (Top 20)

top\_districts\_beneficiaries = df.groupby('district\_name')['total\_beneficiaries'].sum().reset\_index().sort\_values(by='total\_beneficiaries', ascending=False).head(20)

'''plt.figure(figsize=(12, 6))

sns.barplot(data=top\_districts\_beneficiaries, x='district\_name', y='total\_beneficiaries', palette='viridis')

plt.title("Top 20 Districts by Total Beneficiaries")

plt.xlabel("District")

plt.ylabel("Total Beneficiaries")

plt.xticks(rotation=45, ha='right')

plt.tight\_layout()

plt.show()'''

plt.figure(figsize=(12, 6))

sns.barplot(data=top\_districts\_beneficiaries, x='district\_name', y='total\_beneficiaries', hue='district\_name', palette='viridis', dodge=False, legend=False)

plt.title("Top 20 Districts by Total Beneficiaries")

plt.xlabel("District")

plt.ylabel("Total Beneficiaries")

plt.xticks(rotation=45, ha='right')

plt.tight\_layout()

plt.show()

# -------------------------

# Optional: Save static plot as PDF

# -------------------------

plt.figure(figsize=(12, 6))

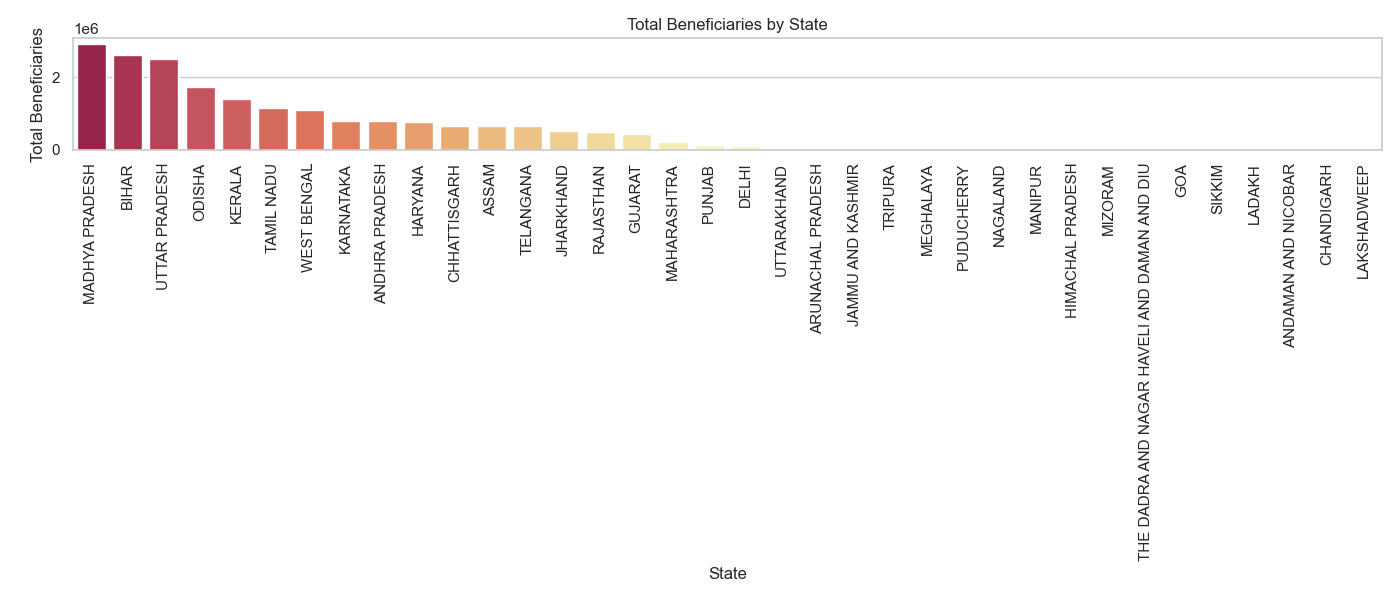
sns.boxplot(data=df[['total\_beneficiaries', 'total\_aadhar', 'total\_mobileno']], palette='Set2')

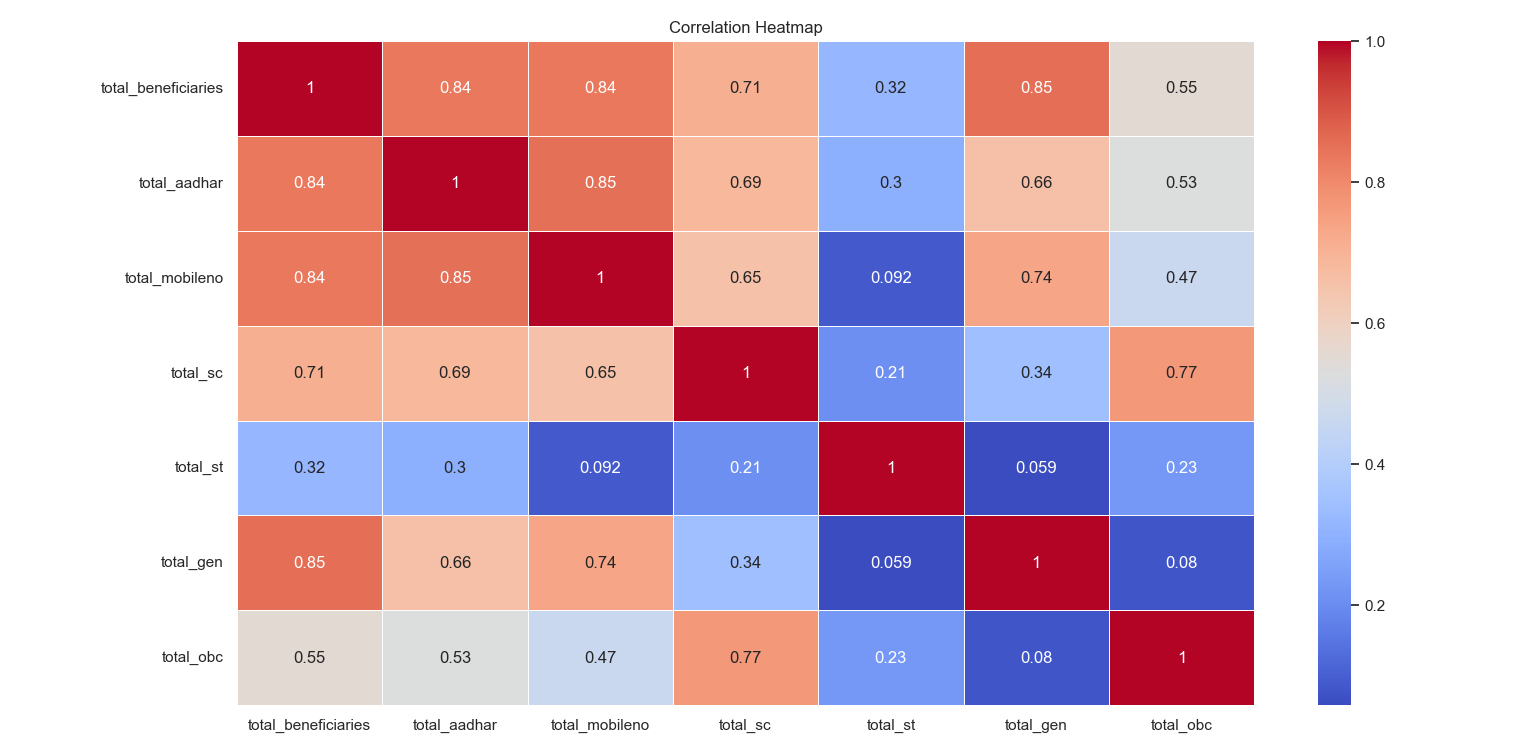
plt.title("Boxplot for Key Metrics")

plt.tight\_layout()

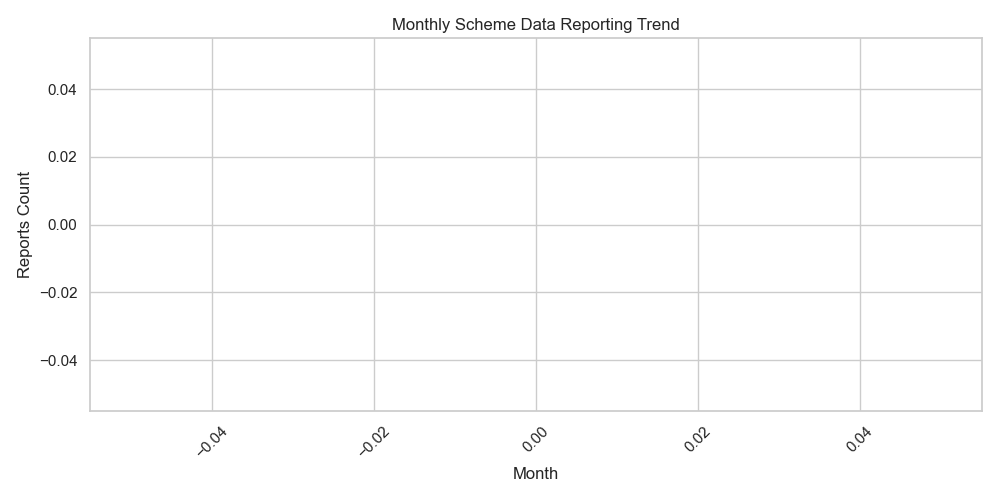
#plt.savefig("key\_metrics\_boxplot.pdf")

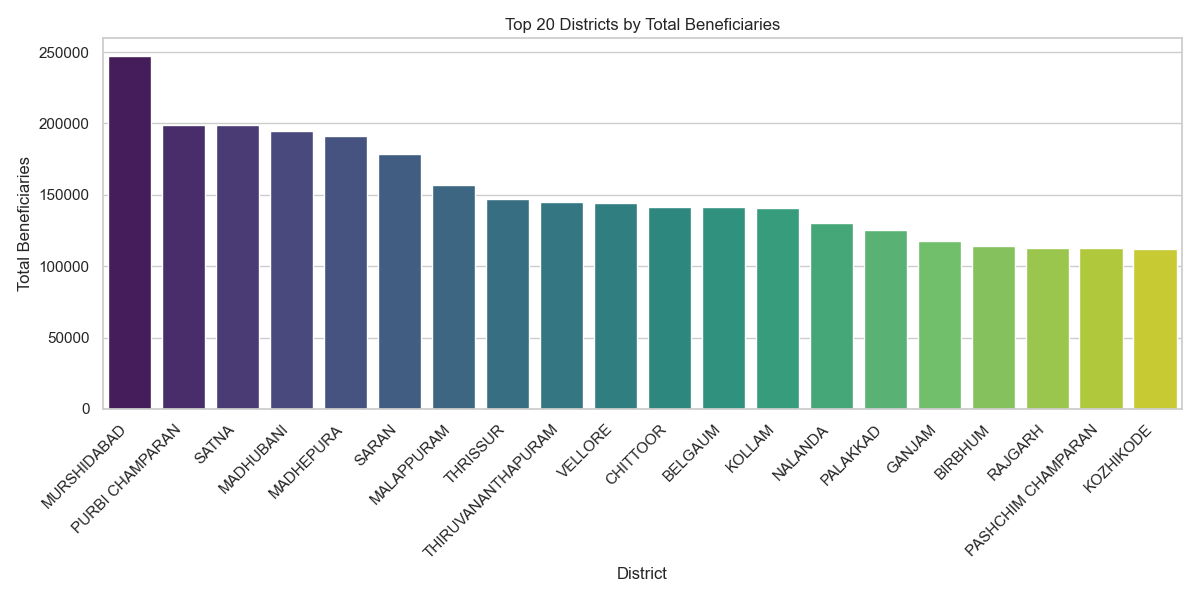
plt.show()

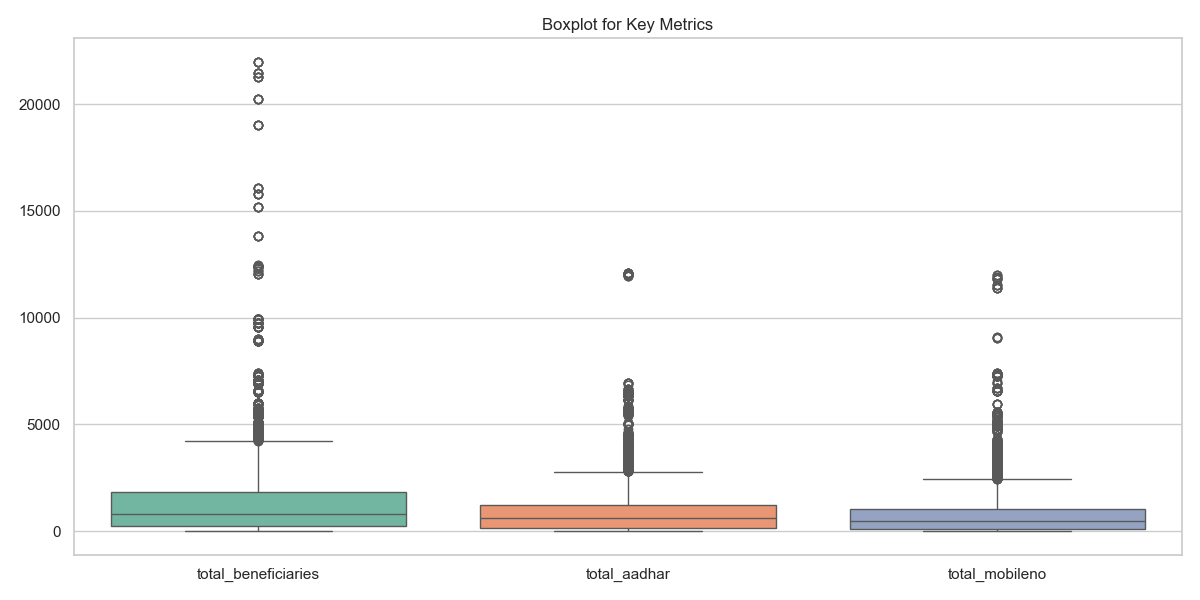












Missing values per column:

fin\_year 0

mnth 0

lgd\_state\_code 0

state\_name 0

lgd\_district\_code 0

district\_name 0

scheme\_code 0

total\_beneficiaries 0

total\_sc 0

total\_st 0

total\_gen 0

total\_obc 0

total\_aadhar 0

total\_mobileno 0

lastUpdated 0

dtype: int64

Correlation between Total Aadhar and Total Beneficiaries: 0.8359695973539879

Top 5 Districts by Inclusion Score:

district\_name inclusion\_score

466 NEW DELHI 0.999143

643 TAPI 0.997176

684 VALSAD 0.994914

146 DANG 0.993853

477 NORTH WEST- I 0.992968

Top 5 States by Inclusion Score:

state\_name inclusion\_score

7 DELHI 0.986018

30 TELANGANA 0.918032

35 WEST BENGAL 0.911258

0 ANDAMAN AND NICOBAR 0.889147

29 TAMIL NADU 0.872944