**CSE3040 – EXPLORATORY DATA ANALYSIS**

**NAME: DHINESH KUMAR M**

**REGNO: 23MIA1102**

**DIGITAL ASSIGNMENT-1**

**SDG -**

**DATASET**

**LINK**

[**https://www.kaggle.com/datasets/iamsouravbanerjee/inequality-in-income-across-the-globe**](https://www.kaggle.com/datasets/iamsouravbanerjee/inequality-in-income-across-the-globe)

**DESCRIPTION**

This dataset comprises historical information encompassing various indicators concerning Inequality in Income on a global scale. The dataset prominently features: ISO3, Country, Continent, Hemisphere, Human Development Groups, UNDP Developing Regions, HDI Rank (2021), and Inequality in Income from 2010 to 2021.

**DATASET GLOSSARY (COLUMN-WISE)**

* ISO3 - ISO3 for the Country/Territory
* Country - Name of the Country/Territory
* Continent - Name of the Continent
* Hemisphere - Name of the Hemisphere
* Human Development Groups - Human Development Groups
* UNDP Developing Regions - UNDP Developing Regions
* HDI Rank (2021) - Human Development Index Rank for 2021
* Inequality in Income from 2010 to 2021 - Inequality in Income from year 2010 to 2021

**DATA DICTIONARY**

* UNDP Developing Regions:
  + SSA - Sub-Saharan Africa
  + LAC - Latin America and the Caribbean
  + EAP - East Asia and the Pacific
  + AS - Arab States
  + ECA - Europe and Central Asia
  + SA - South Asia

**GRAPHICAL ANALYSIS**

**1. HISTOGRAM**

**CODE:**

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

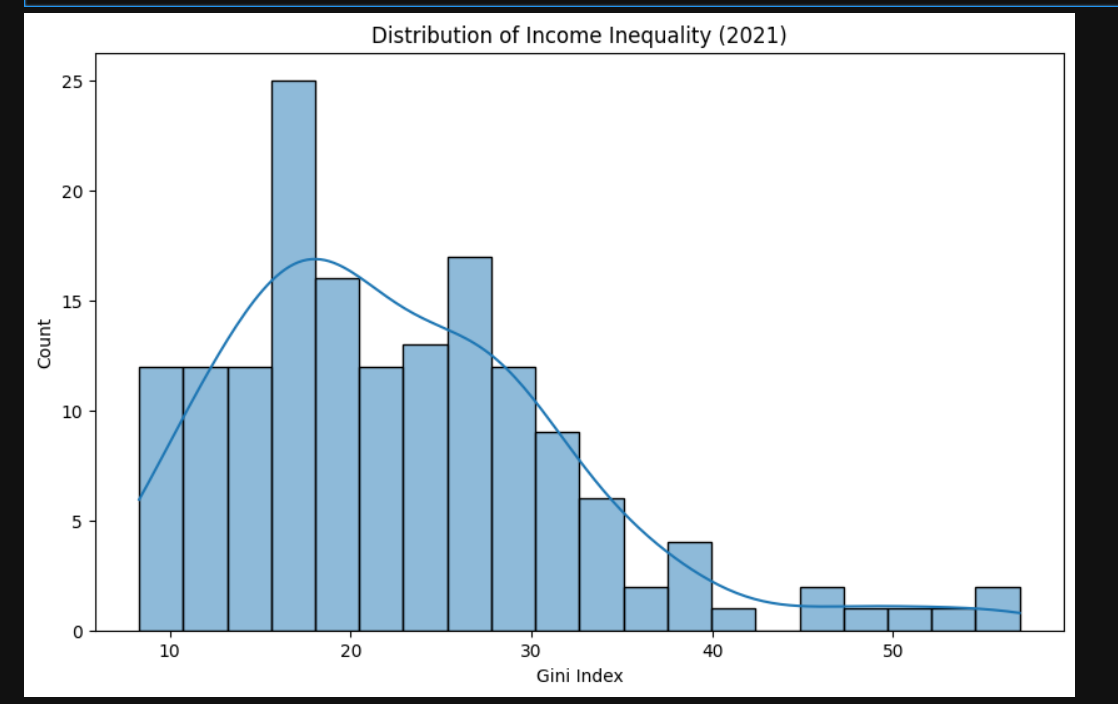
sns.histplot(data['Inequality in income (2021)'], bins=20, kde=True)

plt.title('Distribution of Income Inequality (2021)')

plt.xlabel('Gini Index')

plt.show()

**PLOT:**

****

**2. BOX PLOT**

**CODE:**

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

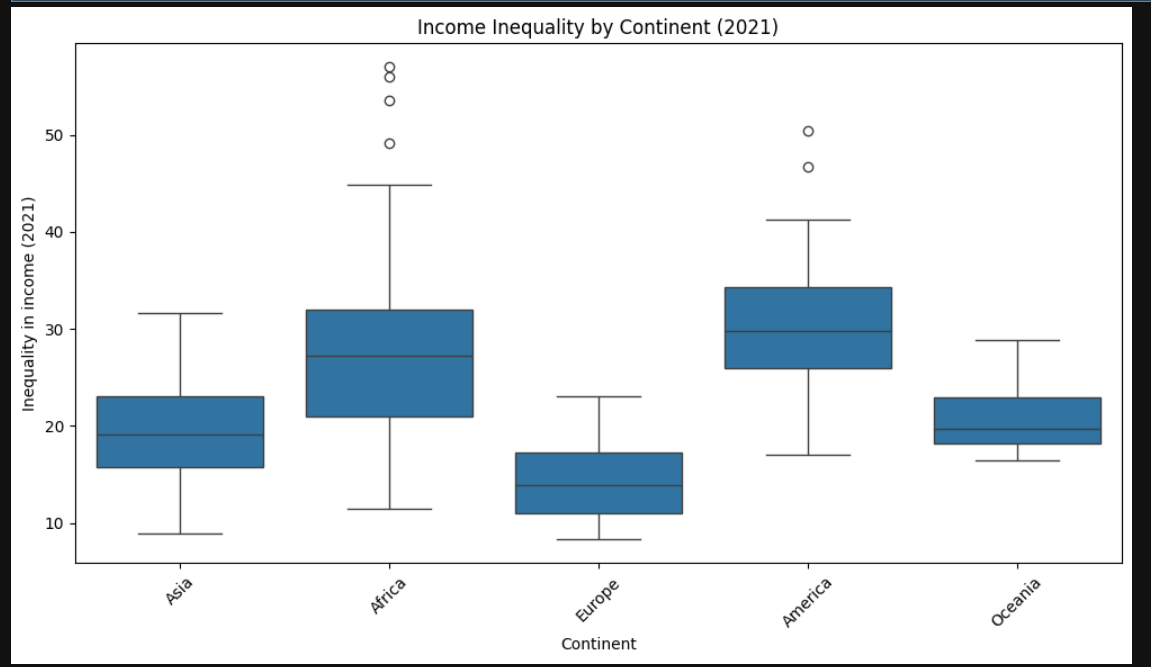
sns.boxplot(x='Continent', y='Inequality in income (2021)', data=data)

plt.title('Income Inequality by Continent (2021)')

plt.xticks(rotation=45)

plt.show()

**PLOT:**

****

**3. LINE GRAPH**

**CODE:**

# Select only numeric columns before computing mean

numeric\_cols = data.select\_dtypes(include=['number']).columns

# Group by 'Continent' and compute the mean only for numeric columns

income\_inequality = data.groupby('Continent')[numeric\_cols].mean().iloc[:, 7:].T

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

for continent in income\_inequality.columns:

sns.lineplot(x=income\_inequality.index, y=income\_inequality[continent], label=continent)

plt.title('Average Inequality in Income from 2010 to 2021 by Continent')

plt.xlabel('Year')

plt.ylabel('Average Inequality in Income')

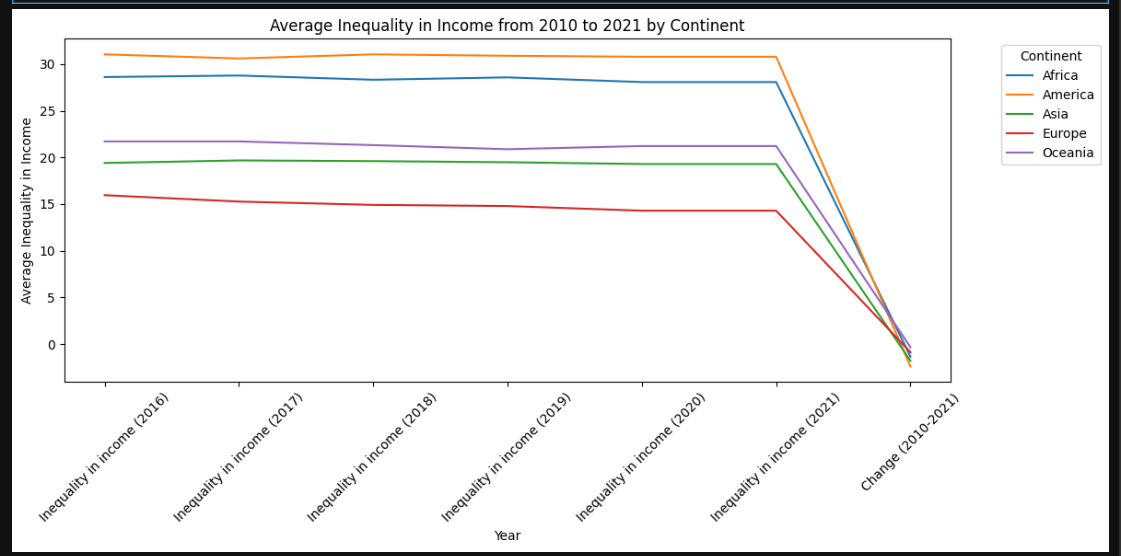
plt.legend(title='Continent', bbox\_to\_anchor=(1.05, 1), loc='upper left')

plt.xticks(rotation=45)

plt.tight\_layout()

plt.show()

**GRAPH:**

****

**4. BAR GRAPH**

**CODE:**

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

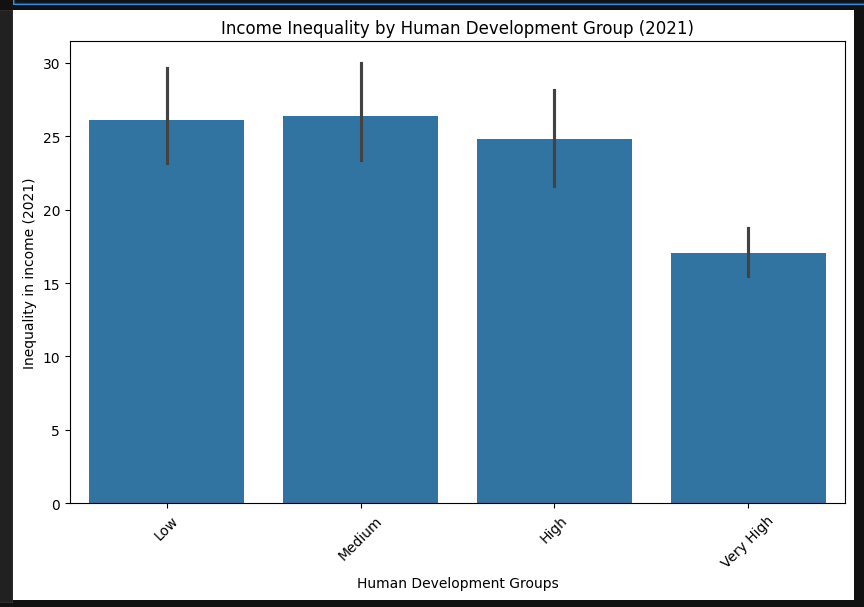
sns.barplot(x='Human Development Groups', y='Inequality in income (2021)', data=data)

plt.title('Income Inequality by Human Development Group (2021)')

plt.xticks(rotation=45)

plt.show()

**GRAPH:**

****

**5. SCATTER PLOT**

**CODE:**

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

sns.scatterplot(x='HDI Rank (2021)', y='Inequality in income (2021)', data=data)

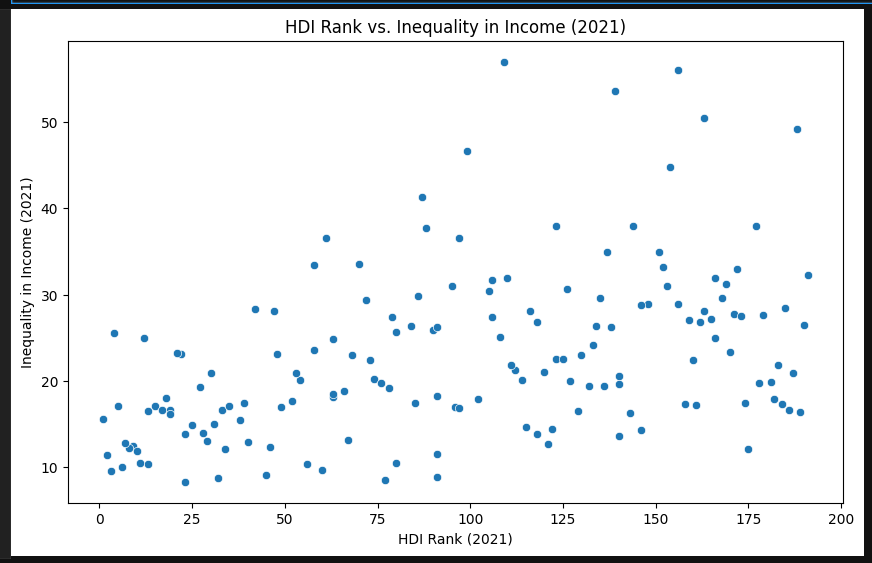
plt.title('HDI Rank vs. Inequality in Income (2021)')

plt.xlabel('HDI Rank (2021)')

plt.ylabel('Inequality in Income (2021)')

plt.show()

**PLOT:**

****

**6. HEAT MAP**

**CODE:**

years = [f'Inequality in income ({year})' for year in range(2010, 2022)]

plt.figure(figsize=(15, 8))

top\_countries = data.nlargest(20, 'Inequality in income (2021)').set\_index('Country')[years]

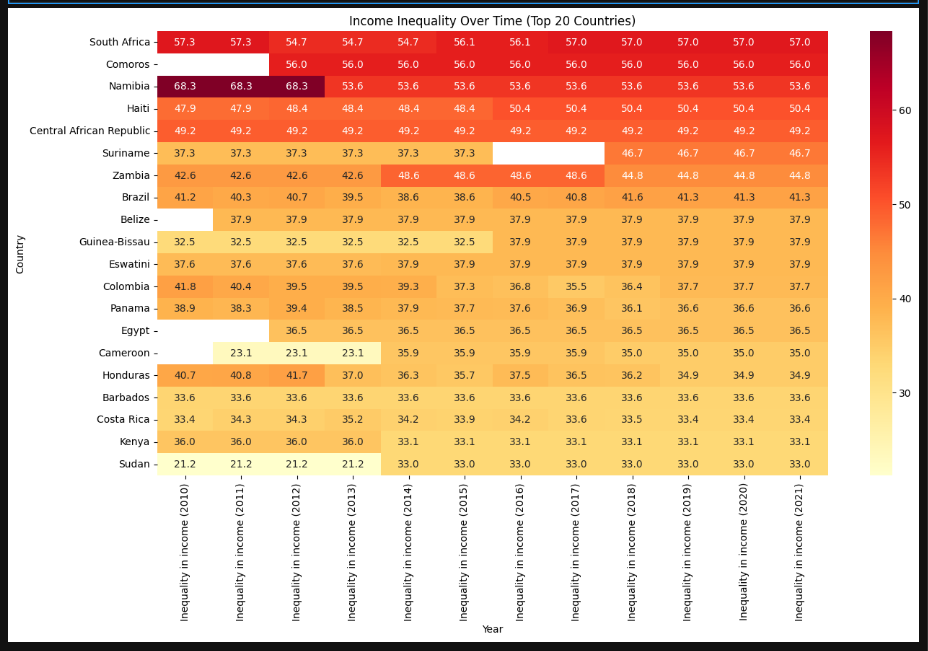
sns.heatmap(top\_countries, annot=True, cmap='YlOrRd', fmt=".1f")

plt.title('Income Inequality Over Time (Top 20 Countries)')

plt.xlabel('Year')

plt.show()

**PLOT:**

****

**7. PIE CHART**

**CODE:**

threshold = data['Inequality in income (2021)'].quantile(0.90) # Top 10% threshold

extreme\_inequality = data[data['Inequality in income (2021)'] >= threshold]

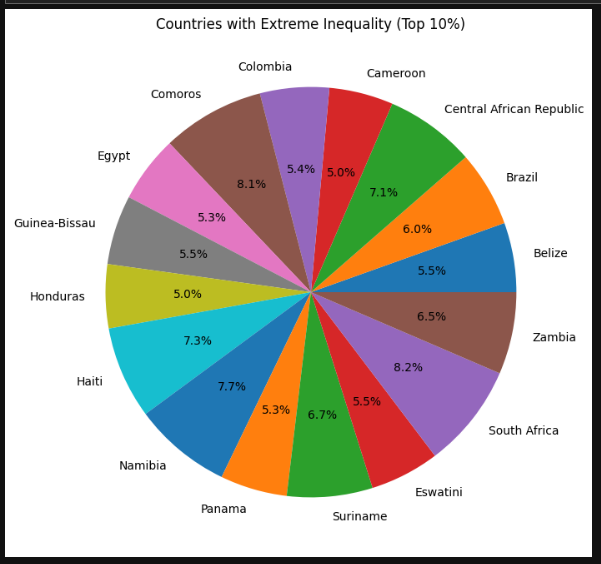
plt.figure(figsize=(8, 8))

plt.pie(extreme\_inequality['Inequality in income (2021)'], labels=extreme\_inequality['Country'], autopct='%1.1f%%')

plt.title('Countries with Extreme Inequality (Top 10%)')

plt.show()

**PLOT:**

****

**8. AREA CHART**

**CODE:**

selected\_countries = data[data['Country'].isin(['United States', 'India', 'Russian Federation', 'Japan', 'United Kindom'])]

plt.figure(figsize=(40, 10))

for country in selected\_countries['Country']:

country\_data = selected\_countries[selected\_countries['Country'] == country]

plt.fill\_between(years, country\_data[years].values.flatten(), alpha=0.3, label=country)

plt.title('Income Inequality Trends in Selected Countries')

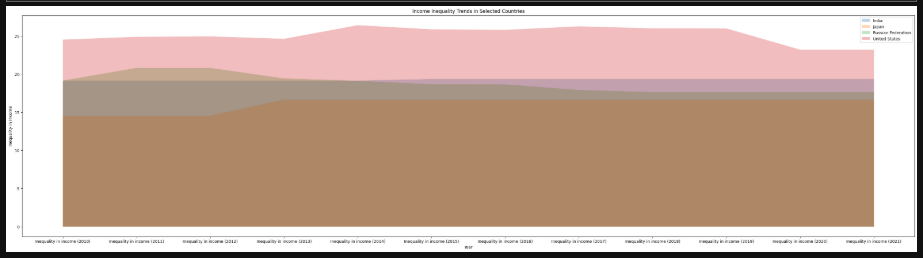
plt.xlabel('Year')

plt.ylabel('Inequality in Income')

plt.legend()

plt.show()

**PLOT:**

****

**9. RADAR CHART**

**CODE:**

years = [f'Inequality in income ({year})' for year in range(2010, 2022)]

continents = data['Continent'].unique()

fig, ax = plt.subplots(figsize=(10, 10), subplot\_kw={'projection': 'polar'})

# Define angles for each continent

labels = data['Continent'].unique()

angles = [n / len(labels) \* 2 \* pi for n in range(len(labels))]

angles.append(angles[0])

# Plot data for each year

for year in years:

if year in data.columns:

values = data.groupby('Continent')[year].mean().fillna(0).tolist()

values.append(values[0])

ax.plot(angles, values, linewidth=2, linestyle='solid', label=year)

ax.fill(angles, values, alpha=0.1)

# Adjust chart settings

ax.set\_theta\_offset(pi / 2)

ax.set\_theta\_direction(-1)

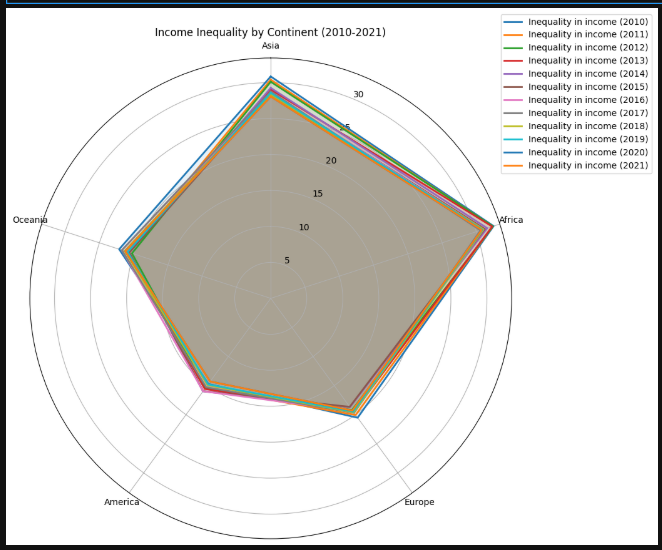
plt.xticks(angles[:-1], labels)

plt.title('Income Inequality by Continent (2010-2021)')

plt.legend(loc='upper right', bbox\_to\_anchor=(1.3, 1.1))

plt.show()

**PLOT:**

****

**10. GEO PLOT**

**CODE:**

world = gpd.read\_file('110m Cultural Vectors/ne\_110m\_admin\_0\_countries.shp')

# Merge datasets using ISO3 codes

merged = world.merge(data, left\_on='SOV\_A3', right\_on='ISO3', how='left')

# List of years to visualize

years = [str(year) for year in range(2010, 2022)]

# Create a world map for each year

for year in years:

plt.figure(figsize=(15, 10))

# Plot the world map with inequality data for the specific year

merged.plot(column=f'Inequality in income ({year})',

cmap='OrRd',

ax=plt.gca(),

legend=True,

legend\_kwds={'label': f'Income Inequality (Gini Index, {year})',

'orientation': 'horizontal'},

missing\_kwds={'color': 'lightgrey'})

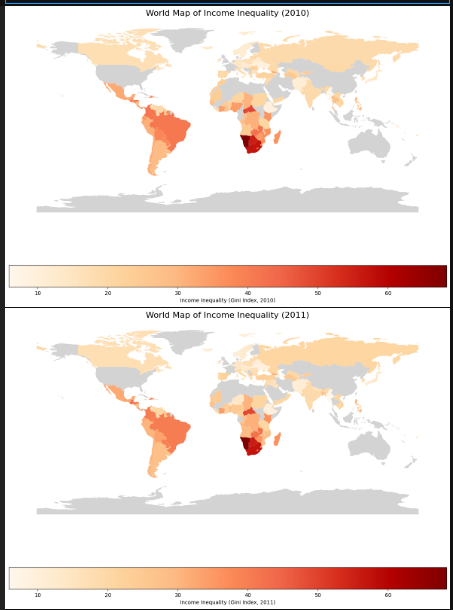
# Title and labels

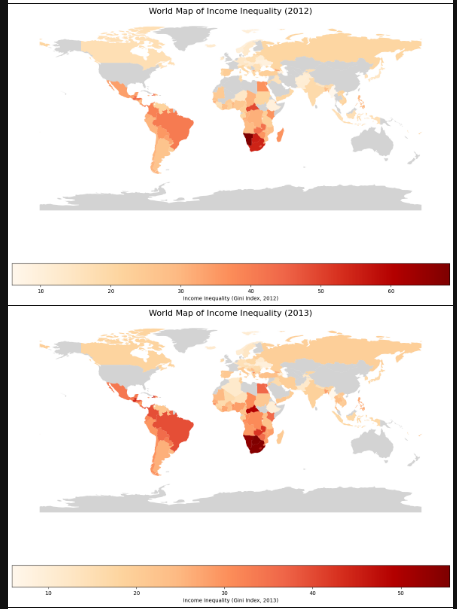
plt.title(f'World Map of Income Inequality ({year})', fontsize=16)

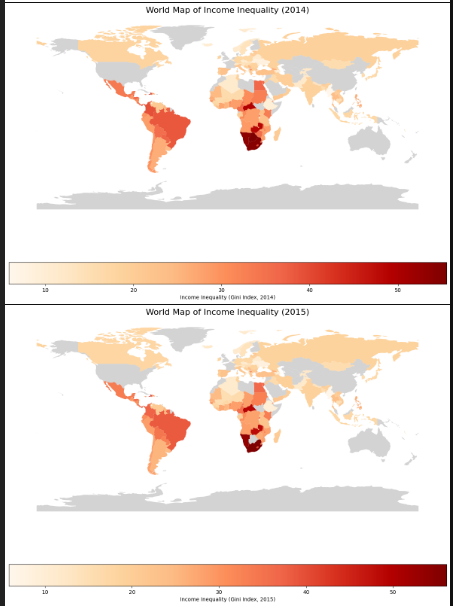
plt.axis('off')

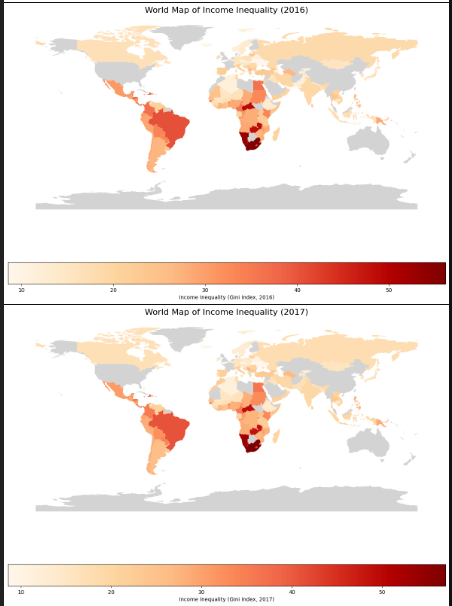
plt.show()

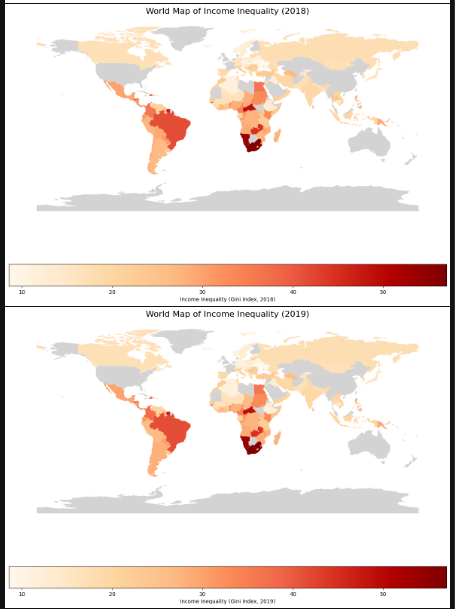
**PLOT:**

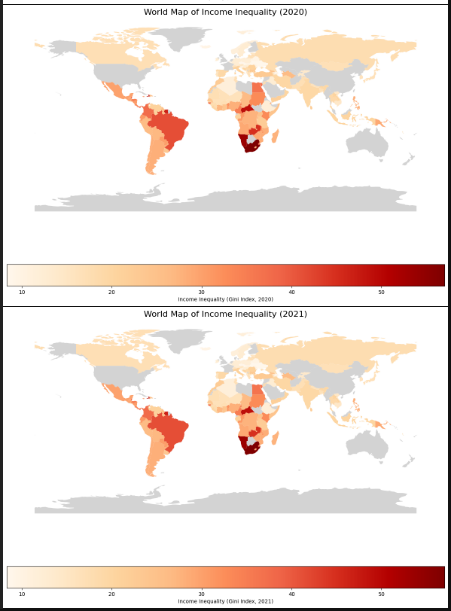
****

****

****

****

****

****

**NON - GRAPHICAL ANALYSIS**

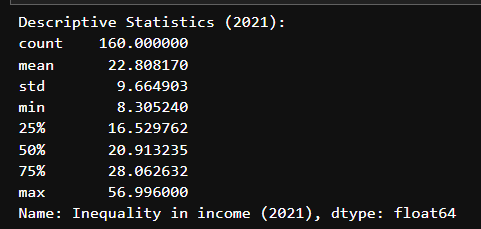
**1. DESCRIPTIVE STATISTICS**

**CODE:**

print("Descriptive Statistics (2021):")

print(data['Inequality in income (2021)'].describe())

**OUTPUT:**

****

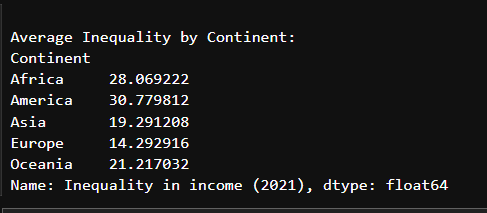
**2. AVERAGE INEQUALITY BY CONTINENT (2021)**

**CODE:**

print("\nAverage Inequality by Continent:")

print(data.groupby('Continent')['Inequality in income (2021)'].mean())

**OUTPUT:**

****

**3. TREND ANALYSIS**

**CODE:**

years = [f'Inequality in income ({year})' for year in range(2010, 2022)]

print("\nTrend Analysis (2010-2021):")

average\_inequality = data[years].mean()

trend\_analysis = pd.DataFrame(average\_inequality).reset\_index()

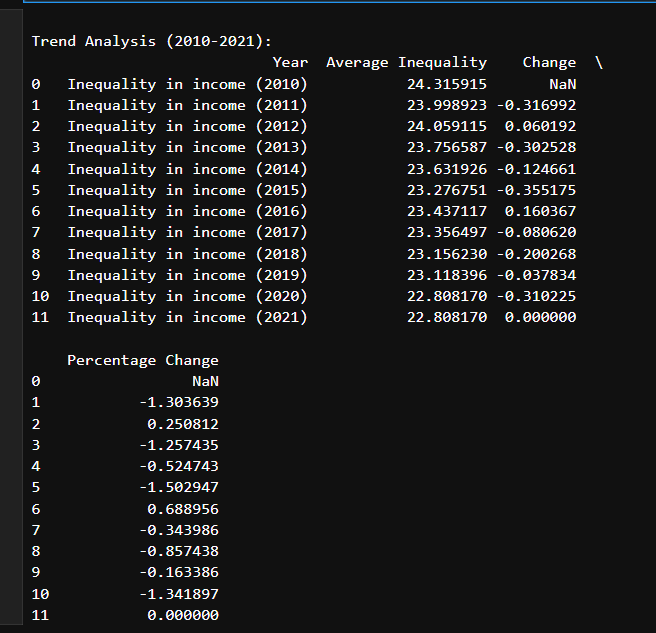
trend\_analysis.columns = ['Year', 'Average Inequality']

trend\_analysis['Change'] = trend\_analysis['Average Inequality'].diff()

trend\_analysis['Percentage Change'] = (trend\_analysis['Change'] / trend\_analysis['Average Inequality'].shift(1)) \* 100

print(trend\_analysis)

**OUTPUT:**

****

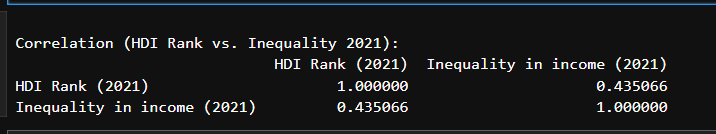
**4. CORRELATION: HDI RANK VS INEQUALITY (2021)**

**CODE:**

print("\nCorrelation (HDI Rank vs. Inequality 2021):")

print(data[['HDI Rank (2021)', 'Inequality in income (2021)']].corr())

**OUTPUT:**

****

**5. COUNTRIES WITH EXTREME INEQUALITY (TOP 10 %) (2021)**

**CODE:**

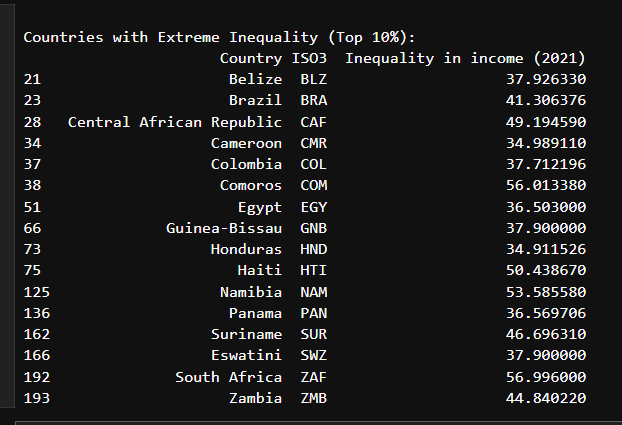
threshold = data['Inequality in income (2021)'].quantile(0.90)

high\_ineq = data[data['Inequality in income (2021)'] >= threshold]

print("\nCountries with Extreme Inequality (Top 10%):")

print(high\_ineq[['Country', 'ISO3', 'Inequality in income (2021)']])

**OUTPUT:**

****

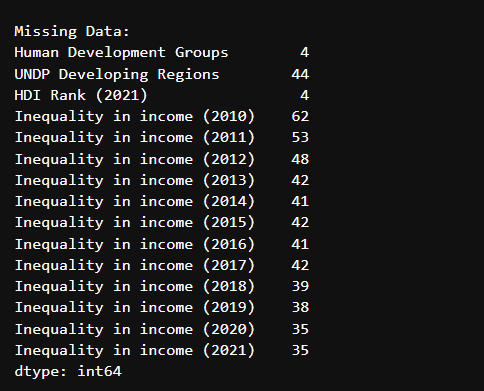
**6. MISSING DATA**

**CODE:**

print("\nMissing Data:")

print(data.isnull().sum()[data.isnull().sum() > 0])

**OUTPUT:**

****

**7. CATEGORIZE COUNTRIES BY INEQUALITY LEVEL (2021)**

**CODE:**

def categorize\_inequality(value):

if value < 30: return 'Low'

elif 30 <= value < 50: return 'Medium'

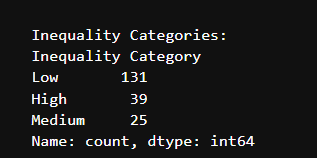
else: return 'High'

data['Inequality Category'] = data['Inequality in income (2021)'].apply(categorize\_inequality)

print("\nInequality Categories:")

print(data['Inequality Category'].value\_counts())

**OUTPUT:**

****

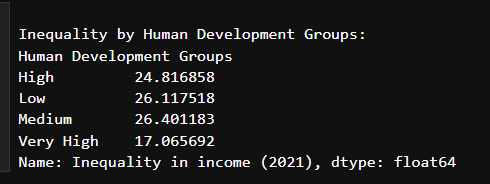
**8. INEQUALITY BY HUMAN DEVELOPMENT GROUPS**

**CODE:**

print("\nInequality by Human Development Groups:")

print(data.groupby('Human Development Groups')['Inequality in income (2021)'].mean())

**OUTPUT:**

****

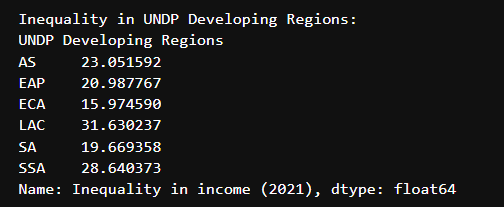
**9. REGIONAL ANALYSIS**

**CODE:**

print("\nInequality in UNDP Developing Regions:")

print(data.groupby('UNDP Developing Regions')['Inequality in income (2021)'].mean())

**OUTPUT:**

****

**10. TEMPORAL CHANGES (2010-2021)**

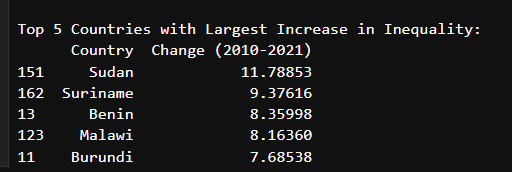
**CODE:**

data['Change (2010-2021)'] = data['Inequality in income (2021)'] - data['Inequality in income (2010)']

print("\nTop 5 Countries with Largest Increase in Inequality:")

print(data[['Country', 'Change (2010-2021)']].nlargest(5, 'Change (2010-2021)'))

**OUTPUT:**

****