**CSE3040 – EXPLORATORY DATA ANALYSIS**

**NAME: DHINESH KUMAR M**

**REGNO: 23MIA1102**

**DIGITAL ASSIGNMENT-1**

**SDG 10: REDUCED INEQUALITIES**

This goal aims to reduce income disparities within and among countries by promoting equal opportunities and addressing economic imbalances.

**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.

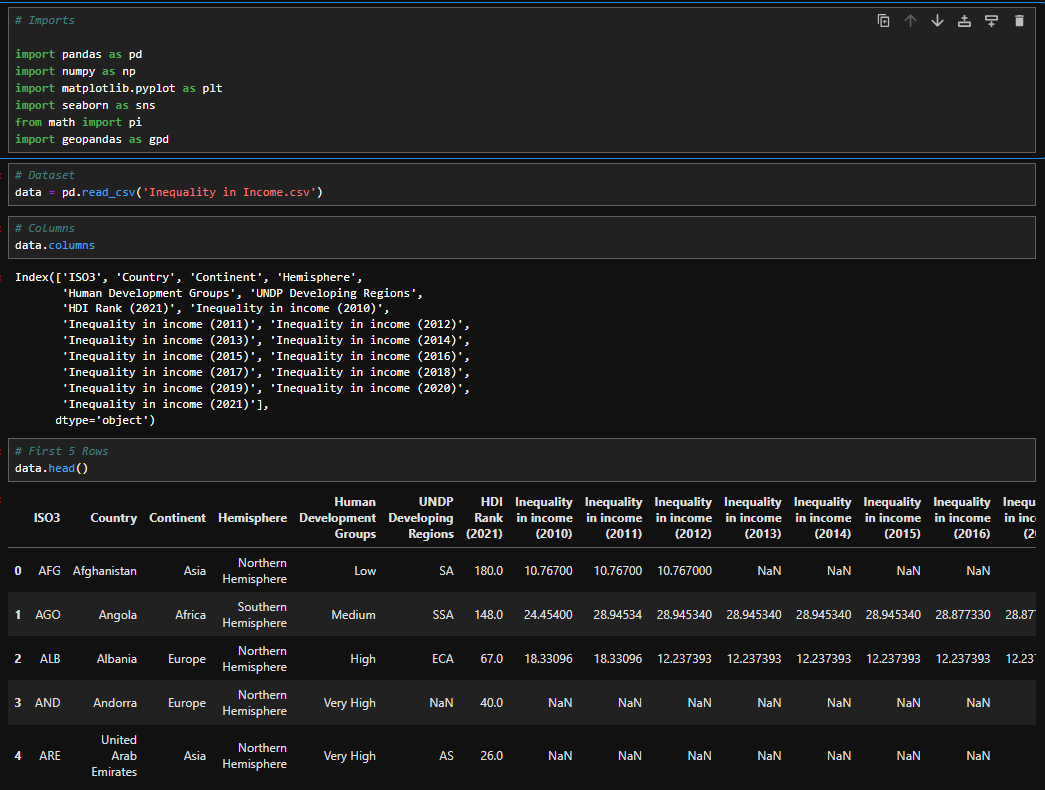
**DATASET GLOSSARY (COLUMN-WISE)**

* 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

**INITIALIZATIONS**

****

**GRAPHICAL ANALYSIS**

**1. HISTOGRAM**

**CODE:**

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

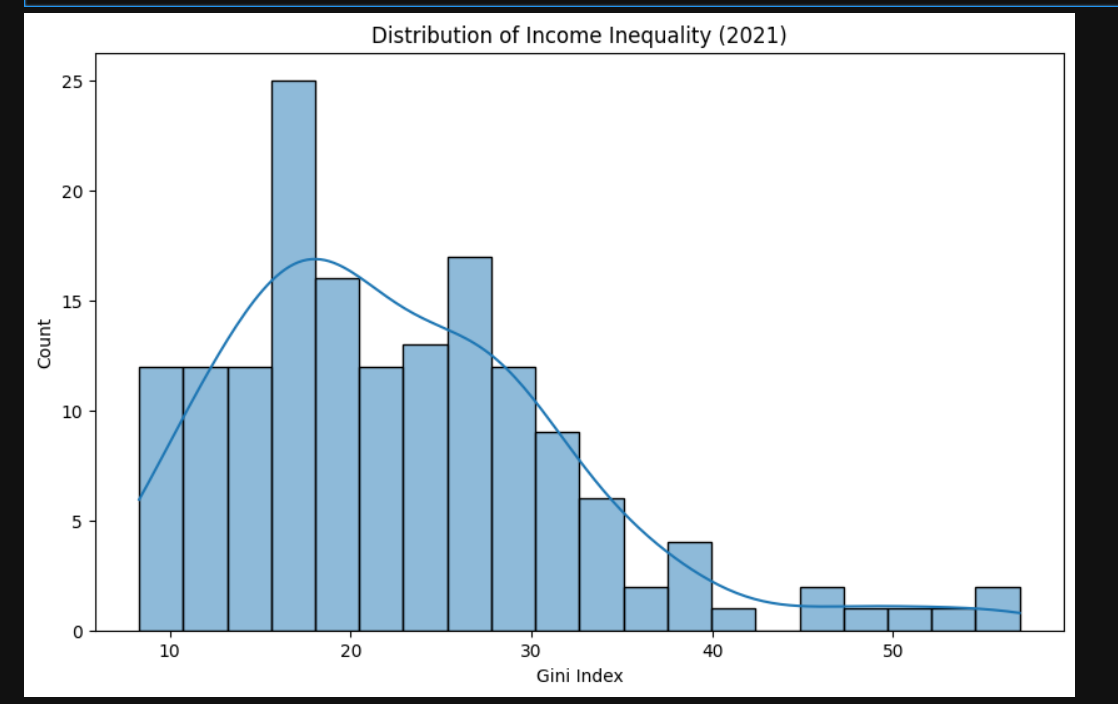
sns.histplot(data['Inequality in income (2021)'], bins=20, kde=True)

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

plt.xlabel('Gini Index')

plt.show()

**PLOT:**

****

The histogram shows a roughly normal distribution of income inequality (Gini Index) with a peak around 20. However, there's a positive skew with some countries exhibiting higher inequality (Gini Index above 40). This suggests that while most countries have moderate income inequality, a notable minority experience significantly higher levels.

**2. BOX PLOT**

**CODE:**

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

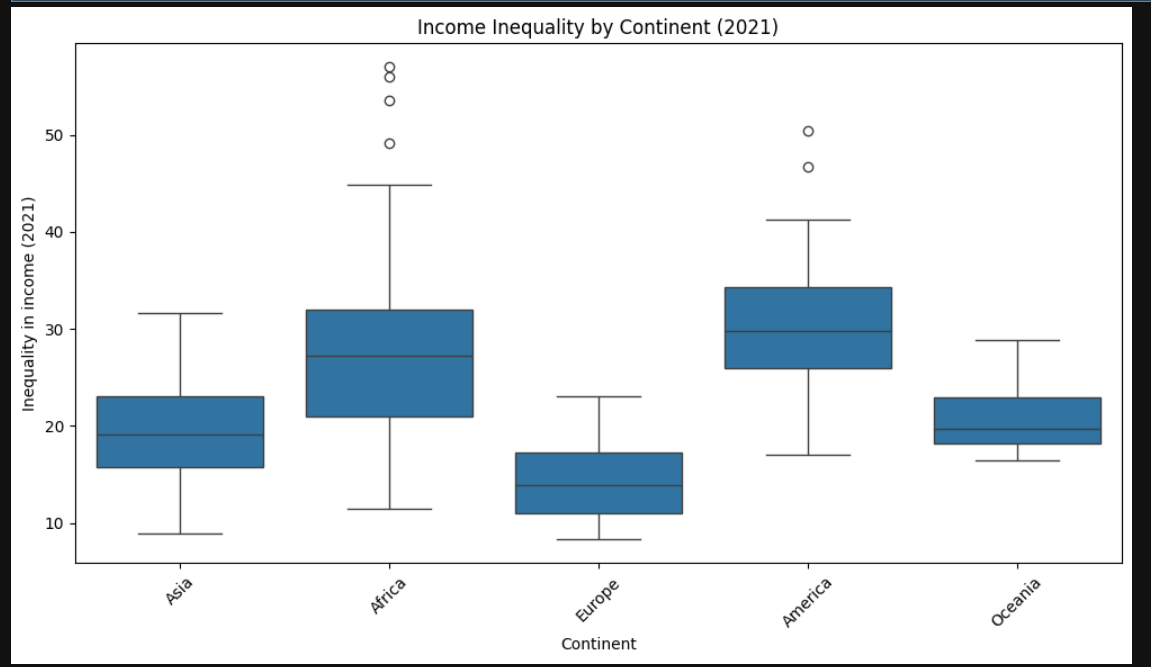
sns.boxplot(x='Continent', y='Inequality in income (2021)', data=data)

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

plt.xticks(rotation=45)

plt.show()

**PLOT:**

****

African nations exhibit the widest range and highest median income inequality, while European countries show the lowest and least variable inequality. Asia and the Americas demonstrate considerable income disparity within their regions, including notable outliers indicative of extreme inequality in specific countries. Oceania presents a low median inequality, but with a relatively limited data range.

**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')

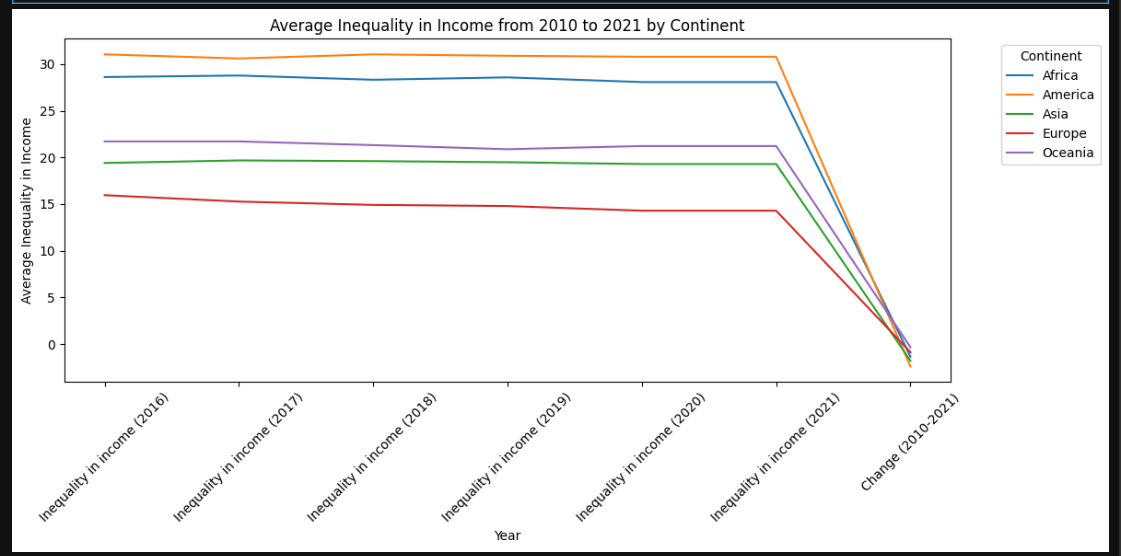
plt.legend(title='Continent', bbox\_to\_anchor=(1.05, 1), loc='upper left')

plt.xticks(rotation=45)

plt.tight\_layout()

plt.show()

**GRAPH:**

****

From 2010 to 2021, income inequality remained relatively stable in Africa, America, Asia, and Europe, while Oceania saw a noticeable decrease.

**4. BAR GRAPH**

**CODE:**

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

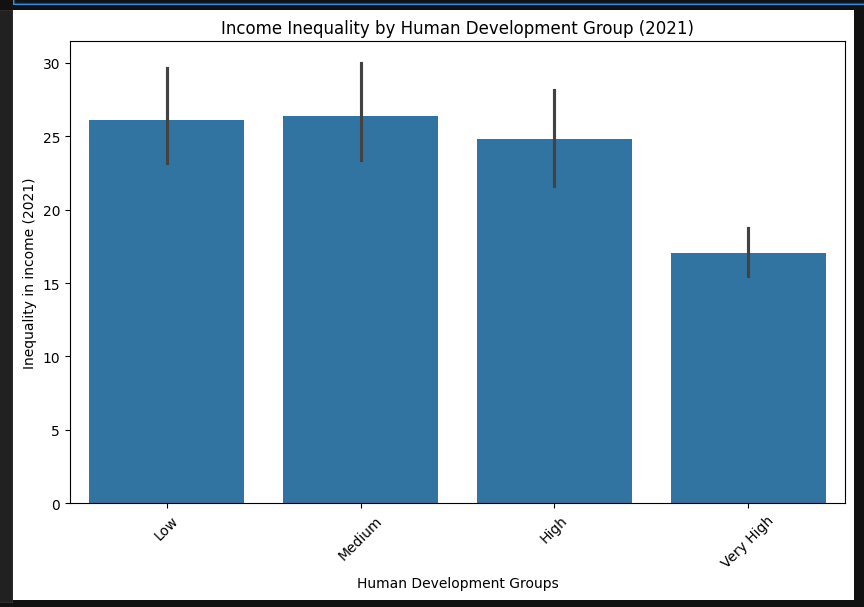
sns.barplot(x='Human Development Groups', y='Inequality in income (2021)', data=data)

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

plt.xticks(rotation=45)

plt.show()

**GRAPH:**

****

Income inequality tends to decrease as the human development index increases, with "Very High" development countries exhibiting the lowest average inequality. While there's a noticeable difference between "Very High" and the other groups, the difference in inequality between "Low," "Medium," and "High" development groups is relatively small.

**5. SCATTER PLOT**

**CODE:**

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

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

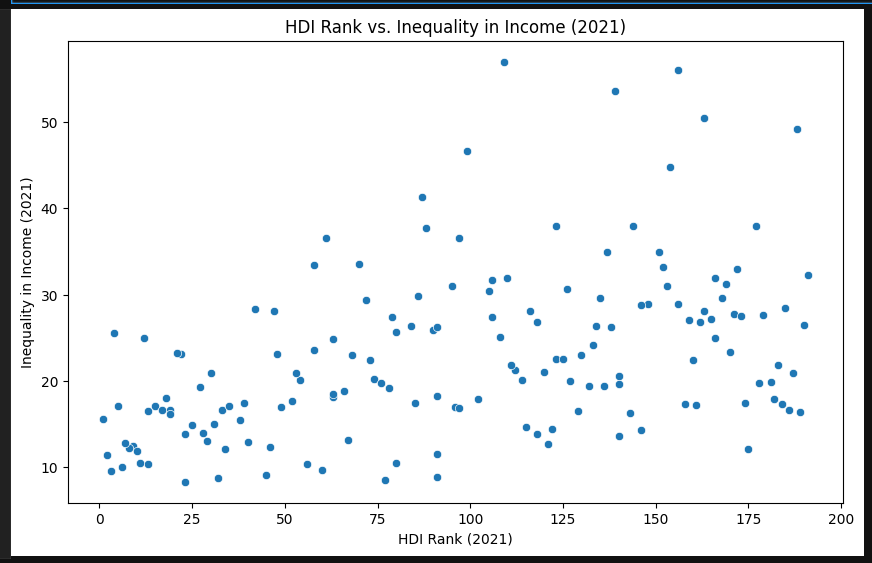
plt.title('HDI Rank vs. Inequality in Income (2021)')

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

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

plt.show()

**PLOT:**

****

A weak negative correlation appears to exist between HDI rank and income inequality, meaning countries with higher human development tend to have lower income inequality, but the relationship is not strong. There's a wide range of income inequality at all HDI ranks, suggesting other factors significantly influence income distribution beyond just human development.

**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]

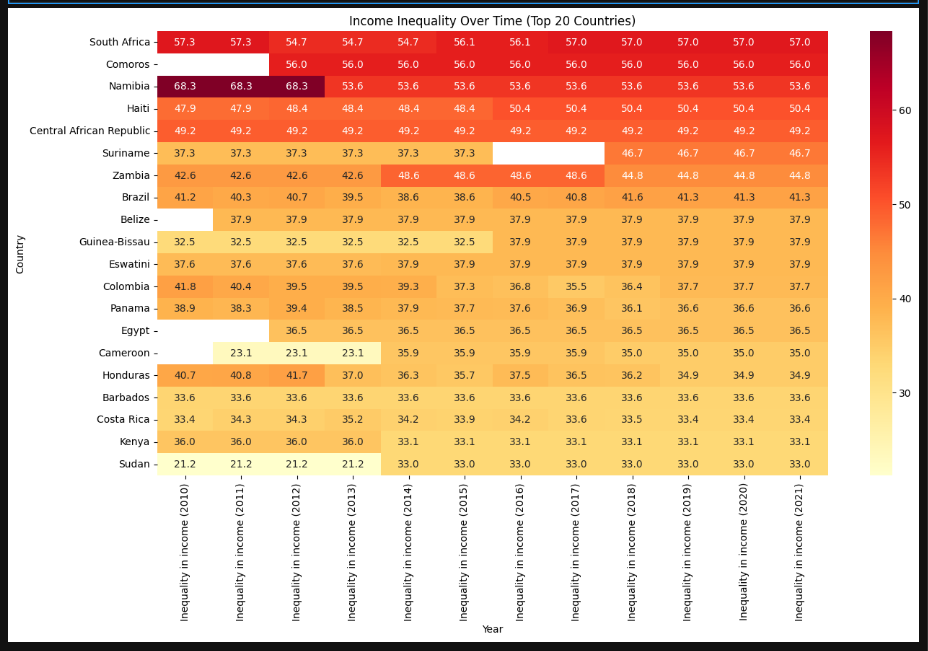
sns.heatmap(top\_countries, annot=True, cmap='YlOrRd', fmt=".1f")

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

plt.xlabel('Year')

plt.show()

**PLOT:**

****

The heatmap visualizes income inequality trends from 2010 to 2021 for the 20 countries with the highest inequality in 2021. Darker shades indicate higher income inequality, revealing persistent high inequality in countries like South Africa and Namibia, while some countries show fluctuations or slight decreases over the 12-year period. A clear pattern of consistent reduction in inequality is not broadly evident across the top 20 countries during this timeframe.

**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]

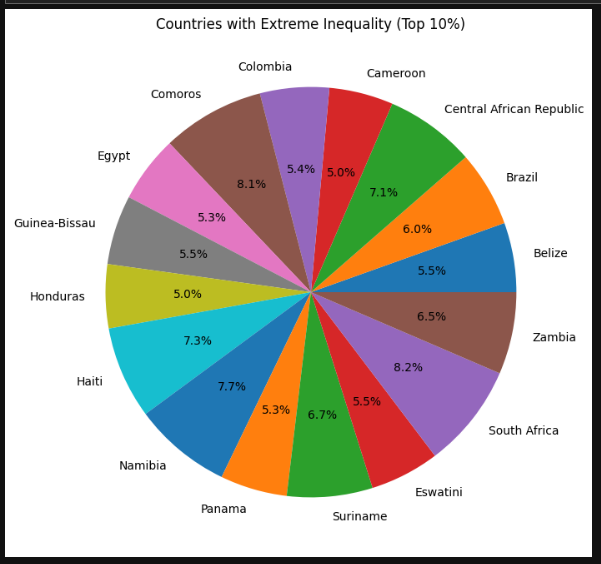
plt.figure(figsize=(8, 8))

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:**

****

The pie chart displays the income inequality distribution among the top 10% most unequal countries. South Africa represents the largest slice, indicating it has the highest level of income inequality within this extreme group, while other countries contribute smaller, varying proportions to the overall high-inequality landscape.

**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')

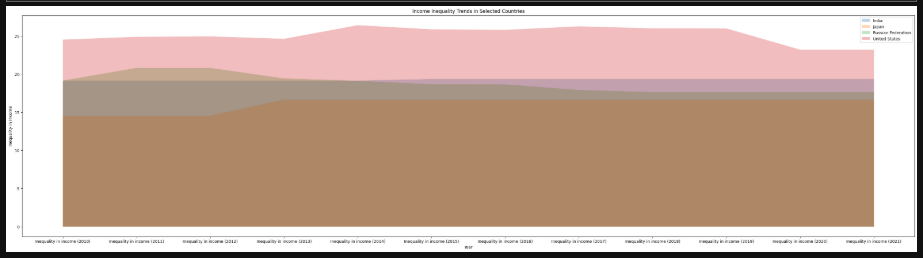
plt.xlabel('Year')

plt.ylabel('Inequality in Income')

plt.legend()

plt.show()

**PLOT:**

****

The area chart compares income inequality trends over time for five selected countries. The United States consistently exhibits the highest level of income inequality among the group, while India shows a notable upward trend over the period. The remaining countries, while displaying some fluctuations, generally maintain lower and relatively stable levels of inequality compared to the US and India.

**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)

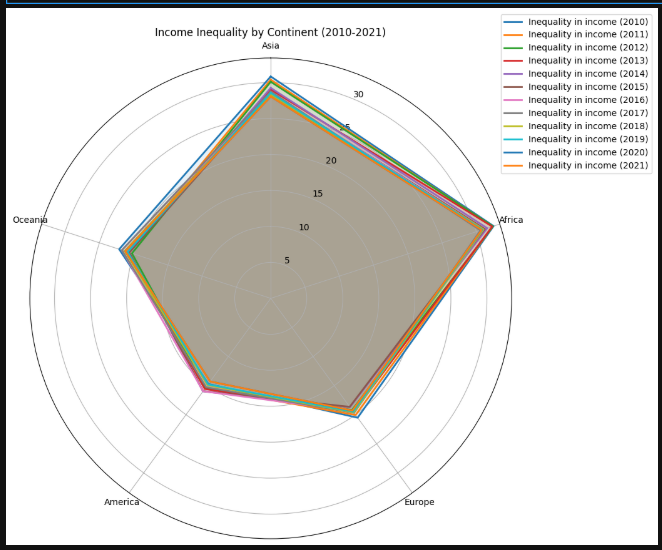
plt.xticks(angles[:-1], labels)

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

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

plt.show()

**PLOT:**

****

This radar chart compares income inequality across continents from 2010 to 2021. Each spoke represents a continent, and the distance from the center indicates the level of income inequality. While there are slight variations across years, the chart suggests that income inequality remained relatively stable for each continent during this period, with Africa consistently showing the highest levels and Oceania the lowest.

**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'})

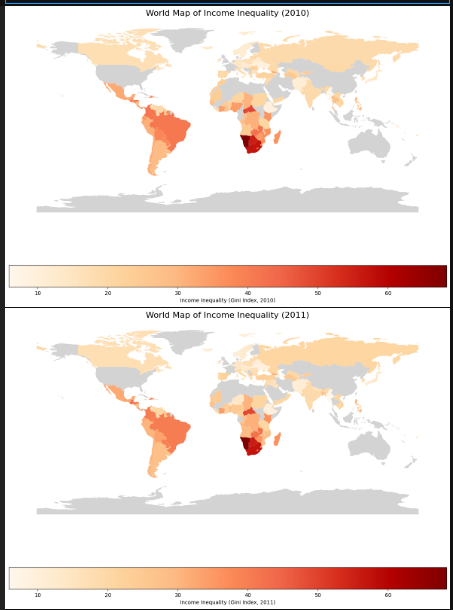
# Title and labels

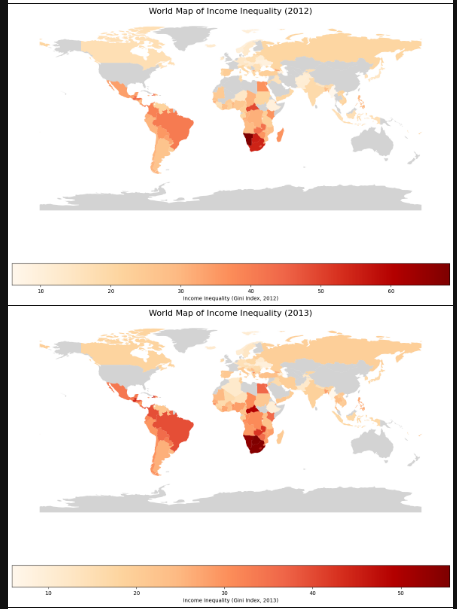
plt.title(f'World Map of Income Inequality ({year})', fontsize=16)

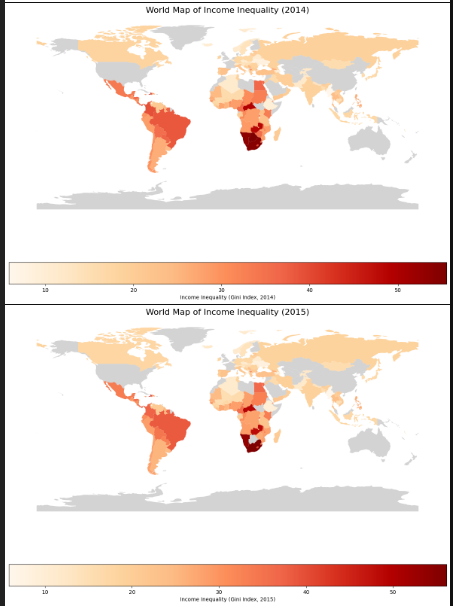
plt.axis('off')

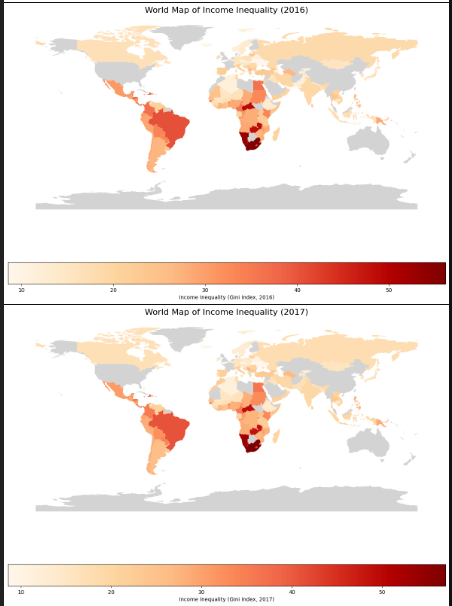
plt.show()

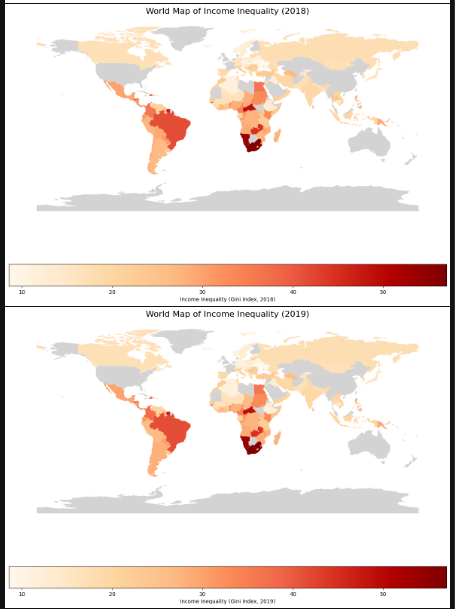
**PLOT:**

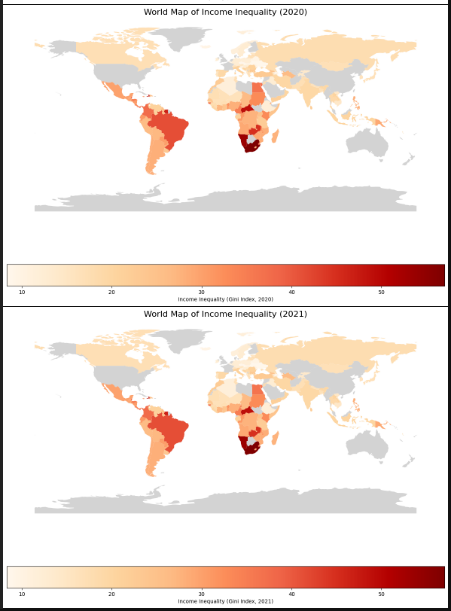
****

****

****

****

****

****

Income inequality varies across continents, with Africa consistently exhibiting the highest levels and Oceania the lowest. While fluctuations exist, the general trend suggests relative stability in inequality levels for most continents over the 12-year period. This visualization highlights the persistent disparities in income distribution globally, emphasizing the need for targeted interventions.

**NON - GRAPHICAL ANALYSIS**

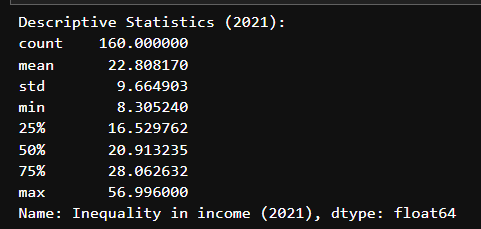
**1. DESCRIPTIVE STATISTICS**

**CODE:**

print("Descriptive Statistics (2021):")

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

**OUTPUT:**

****

In 2021, the average income inequality (Gini index) across the 160 countries in the dataset was approximately 22.8. However, there's substantial variation, with some countries exhibiting very low inequality (minimum of 8.3) and others very high (maximum of 57), as indicated by the large standard deviation of 9.7. This suggests a wide disparity in income distribution globally.

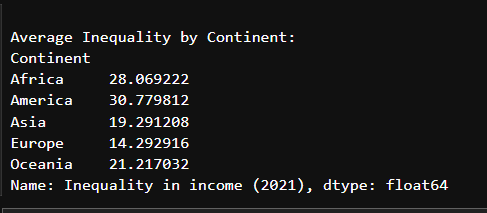
**2. AVERAGE INEQUALITY BY CONTINENT (2021)**

**CODE:**

print("\nAverage Inequality by Continent:")

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

**OUTPUT:**

****

Africa and the Americas exhibit the highest average income inequality among the continents, while Europe demonstrates the lowest. Asia and Oceania fall in the middle range, suggesting moderate levels of inequality compared to the extremes. This highlights significant regional disparities in income distribution across the globe.

**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()

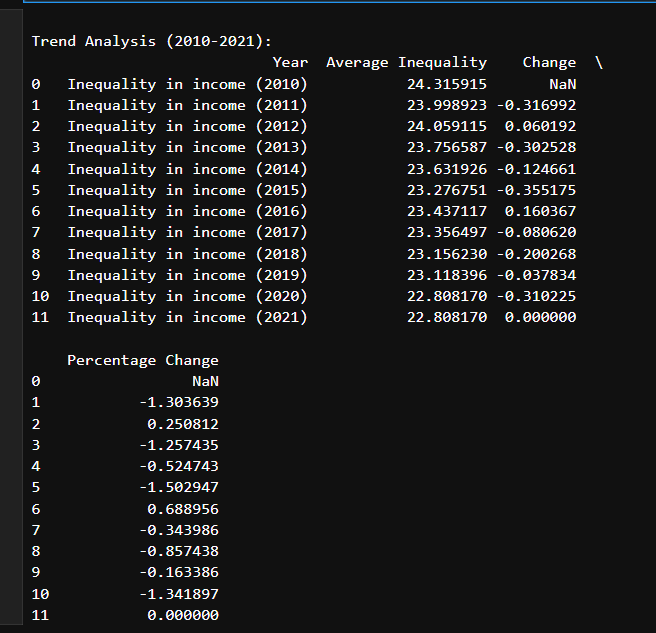
trend\_analysis.columns = ['Year', 'Average Inequality']

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:**

****

Global average income inequality fluctuated slightly between 2010 and 2021, showing no consistent upward or downward trend. While some years saw small increases or decreases, the overall change in average inequality during this period is minimal, suggesting relative stability in global income disparities. A notable drop occurred between 2020 and 2021, but without more context, it's difficult to determine the cause or significance of this change.

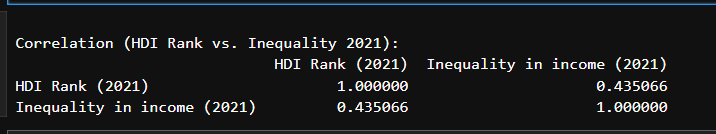
**4. CORRELATION: HDI RANK VS INEQUALITY (2021)**

**CODE:**

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

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

**OUTPUT:**

****

There is a moderate positive correlation (0.44) between HDI rank and income inequality in 2021. This suggests that countries with lower HDI ranks (less developed) tend to have higher income inequality, though the relationship is not extremely strong, implying other factors also play a significant role. It's important to note that a positive correlation with HDI *rank* means a negative relationship with HDI score.

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

**CODE:**

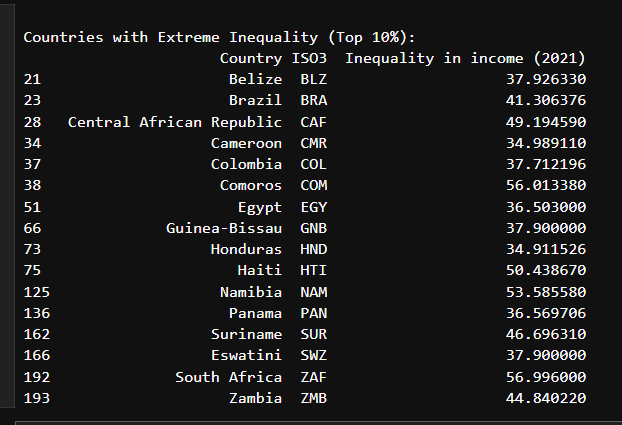
threshold = data['Inequality in income (2021)'].quantile(0.90)

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:**

****

The list identifies 16 countries that represent the top 10% of income inequality globally in 2021. South Africa exhibits the highest level of inequality within this group (57.0), while Belize has the lowest (37.9) among these high-inequality nations. This underscores that even within the most unequal countries, there is a range of disparity levels.

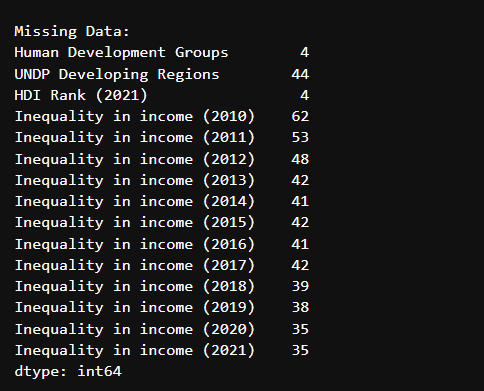
**6. MISSING DATA**

**CODE:**

print("\nMissing Data:")

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

**OUTPUT:**

****

The dataset has missing values across several variables, particularly in income inequality data for earlier years (2010-2014), which have the most missing entries. The presence of missing data, especially in the target variable, could affect the accuracy of analyses and limit the scope of drawing robust conclusions. The missingness in 'Human Development Groups', 'UNDP Developing Regions', and 'HDI Rank' could also introduce bias if not handled appropriately.

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

**CODE:**

def categorize\_inequality(value):

if value < 30: return 'Low'

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

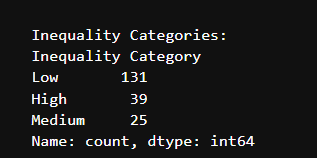
else: return 'High'

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

print("\nInequality Categories:")

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

**OUTPUT:**

****

The majority of countries (131 out of 195) fall into the "Low" income inequality category in 2021, based on the defined thresholds. However, a substantial minority (39 countries) are classified as "High" inequality, while a smaller portion (25) falls into the "Medium" category. This indicates that while low inequality is more common, high income disparities remain a significant global concern.

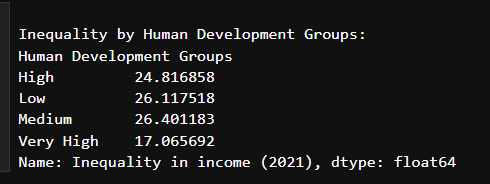
**8. INEQUALITY BY HUMAN DEVELOPMENT GROUPS**

**CODE:**

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

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

**OUTPUT:**

****

Countries with "Very High" human development exhibit significantly lower income inequality compared to all other groups. "Low" and "Medium" human development groups show the highest levels of inequality, while the "High" group falls in between. This suggests a link between human development and income distribution, with higher development associated with lower inequality.

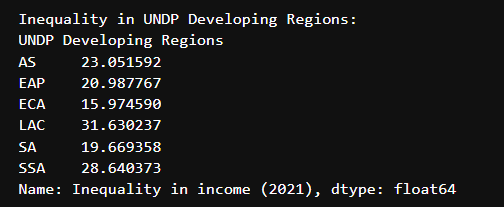
**9. REGIONAL ANALYSIS**

**CODE:**

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

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

**OUTPUT:**

****

Latin America and the Caribbean (LAC) exhibits the highest average income inequality among the UNDP developing regions listed, while Europe and Central Asia (ECA) shows the lowest. Sub-Saharan Africa (SSA) also demonstrates high inequality. East Asia and the Pacific (EAP), South Asia (SA), and Arab States (AS) fall within a moderate range, indicating regional disparities in income distribution across developing regions.

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

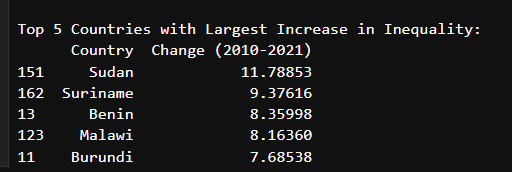
**CODE:**

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:**

****

Between 2010 and 2021, Sudan experienced the largest increase in income inequality among the countries in the dataset, followed by Suriname, Benin, Malawi, and Burundi. These countries represent the top 5 with the most substantial growth in inequality during this period, suggesting significant shifts in their income distribution dynamics. It's important to investigate the specific factors driving these changes in each country.

**CONCLUSION**

The graphical and non-graphical analysis of income inequality over the years reveals trends in various countries and regions. Income inequality has varied across different continents, human development groups, and UNDP developing regions, with certain countries experiencing more significant disparities. By examining the annual data from 2010 to 2021, we can identify whether income inequality has improved or worsened, highlighting regions that need more focused policy interventions to reduce disparities.

**ANALYSIS LINK:**

[**https://github.com/Dhinesh-Fedor/EDA-DA-1.git**](https://github.com/Dhinesh-Fedor/EDA-DA-1.git)