Inequality in Education Around the World

This dataset contains historical data of educational inequality on global scale. Components of this dataset include - ISO3, ISO3, Human Development Groups, UNDP Developing Regions, HDI Rank (2021), Inequality in Education spanning the years 2010 TO 2021

First we will import the necessary Python libraries

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Now, let's see what each library is used for :

"pandas" is used for data manipulation and analysis. It provides data structures like DataFrames for working with tabular data.

"numpy" is for numerical operations and array handling. It's often used for numerical computations.

"matplotlib.pyplot" is used for creating various types of plots and charts.

"seaborn" is a data visualization library built on top of matplotlib. It makes it easier to create visually appealing statistical graphics.

In [2]:
```
df = pd.read_csv("Inequality in Education.csv")
df.head()
```

Out[2]:

	ISO3	Country	Human Development Groups	UNDP Developing Regions	HDI Rank (2021)	Inequality in Education (2010)	Inequality in Education (2011)	Inequality in Education (2012)	Inequality in Education (2013)	Inequality in Education (2014)	Inequality in Education (2015)	Inequality in Education (2016)	Inequality in Education (2017)	Inequality in Education (2018)	In E
0	AFG	Afghanistan	Low	SA	180.0	42.809000	44.823380	44.823380	44.823380	44.823380	45.365170	45.365170	45.365170	45.365170	4
1	AGO	Angola	Medium	SSA	148.0	NaN	NaN	NaN	NaN	NaN	34.171440	34.171440	34.171440	34.171440	3
2	ALB	Albania	High	ECA	67.0	11.900000	11.900000	11.900000	11.900000	11.900000	11.900000	11.900000	12.333440	12.333440	1
3	AND	Andorra	Very High	NaN	40.0	15.160302	15.160302	15.160302	15.160302	9.965681	10.083815	10.008154	10.008154	10.008154	1
4	ARE	United Arab Emirates	Very High	AS	26.0	NaN	NaN	NaN	NaN	NaN	NaN	18.241437	14.475335	12.634355	1

In this step, we are reading data from a CSV file named "Inequality in Education" using "pd.read_csv" and storing the data in variable named df.

In [3]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 195 entries, 0 to 194
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   ISO3                            195 non-null    object
 1   Country                         195 non-null    object
 2   Human Development Groups        191 non-null    object
 3   UNDP Developing Regions         151 non-null    object
 4   HDI Rank (2021)                 191 non-null    float64
 5   Inequality in Education (2010)  137 non-null    float64
 6   Inequality in Education (2011)  150 non-null    float64
 7   Inequality in Education (2012)  157 non-null    float64
 8   Inequality in Education (2013)  165 non-null    float64
 9   Inequality in Education (2014)  168 non-null    float64
 10  Inequality in Education (2015)  168 non-null    float64
 11  Inequality in Education (2016)  168 non-null    float64
 12  Inequality in Education (2017)  168 non-null    float64
 13  Inequality in Education (2018)  172 non-null    float64
 14  Inequality in Education (2019)  174 non-null    float64
 15  Inequality in Education (2020)  176 non-null    float64
 16  Inequality in Education (2021)  176 non-null    float64
dtypes: float64(13), object(4)
memory usage: 26.0+ KB
```

Understanding Given Data -

- ISO3 - ISO3 for the Country. This column have 195 non-null entries and datatype of this column is "object".
- Country - Name of the Country. This column have 195 non-null entries and datatype of this column is "object".
- Human Development Groups - Human Development Groups. This column have 191 non-null entries, which implies there are 4 null/missing values and datatype of this column is "object".
- UNDP Developing Regions - UNDP Developing Regions. This column have 151 non-null entries, which implies there are 44 null/missing values and datatype of this column is "object".
- HDI Rank (2021) - Human Development Index Rank for 2021. This is a numeric column with 191 non-null entries. There are 4 null/missing values, and it's of the "float64" data type.
- Inequality in Education (2010) - Inequality in Education for 2010. This is a numeric column with 137 non-null entries. There are 58 null/missing values, and it's of the "float64" data type.
- Inequality in Education (2011) - Inequality in Education for 2011. This is a numeric column with 150 non-null entries. There are 45 null/missing values, and it's of the "float64" data type.
- Inequality in Education (2012) - Inequality in Education for 2012. This is a numeric column with 157 non-null entries. There are 38 null/missing values, and it's of the "float64" data type.
- Inequality in Education (2013) - Inequality in Education for 2013. This is a numeric column with 165 non-null entries. There are 30 null/missing values, and it's of the "float64" data type.
- Inequality in Education (2014) - Inequality in Education for 2014. This is a numeric column with 168 non-null entries. There are 27 null/missing values, and it's of the "float64" data type.
- Inequality in Education (2015) - Inequality in Education for 2015. This is a numeric column with 168 non-null entries. There are 27 null/missing values, and it's of the "float64" data type.
- Inequality in Education (2016) - Inequality in Education for 2016. This is a numeric column with 168 non-null entries. There are 27 null/missing values, and it's of the "float64" data type.
- Inequality in Education (2017) - Inequality in Education for 2017. This is a numeric column with 168 non-null entries. There are 27 null/missing values, and it's of the "float64" data type.
- Inequality in Education (2018) - Inequality in Education for 2018. This is a numeric column with 172 non-null entries. There are 23 null/missing values, and it's of the "float64" data type.
- Inequality in Education (2019) - Inequality in Education for 2019. This is a numeric column with 174 non-null entries. There are 21 null/missing values, and it's of the "float64" data type.
- Inequality in Education (2020) - Inequality in Education for 2020. This is a numeric column with 176 non-null entries. There are 19 null/missing values, and it's of the "float64" data type.
- Inequality in Education (2021) - Inequality in Education for 2021. This is a numeric column with 176 non-null entries. There are 19 null/missing values, and it's of the "float64" data type.

In this step, we are calculating percentage of null values

In [4]:
```
df.isnull().sum()/len(df)*100
```

Out[4]:
```
ISO3                              0.000000
Country                           0.000000
Human Development Groups          2.051282
UNDP Developing Regions          22.564103
HDI Rank (2021)                   2.051282
Inequality in Education (2010)   29.743590
Inequality in Education (2011)   23.076923
Inequality in Education (2012)   19.487179
Inequality in Education (2013)   15.384615
Inequality in Education (2014)   13.846154
Inequality in Education (2015)   13.846154
Inequality in Education (2016)   13.846154
Inequality in Education (2017)   13.846154
Inequality in Education (2018)   11.794872
Inequality in Education (2019)   10.769231
Inequality in Education (2020)    9.743590
Inequality in Education (2021)    9.743590
dtype: float64
```

- ISO3: There are no null values in the "ISO3" column, meaning it's complete.
- Country: There are no null values in the "Country" column, indicating it's also complete.
- Human Development Groups: About 2.05% of the values in the "Human Development Groups" column are null.
- UNDP Developing Regions: Approximately 22.56% of the values in the "UNDP Developing Regions" column are null.
- HDI Rank (2021): About 2.05% of the values in the "HDI Rank (2021)" column are null.
- Inequality in Education (2010) to Inequality in Education (2021): These columns show the percentage of null values for each respective year, ranging from 9.74% to 29.74%. The percentage of null data decreases over time, but there are still some null values in each of these columns.

In this step, we will handle null values which are present in numerical columns.

In [5]:
```
numerical_cols = ['HDI Rank (2021)','Inequality in Education (2010)','Inequality in Education (2011)',
                  'Inequality in Education (2012)','Inequality in Education (2013)','Inequality in Education (2014)',
                  'Inequality in Education (2015)','Inequality in Education (2016)','Inequality in Education (2017)',
                  'Inequality in Education (2018)','Inequality in Education (2019)','Inequality in Education (2020)',
                  'Inequality in Education (2021)']
df.fillna(df[numerical_cols].mean(),inplace=True)
```

- we created a list called "numerical_cols" containing the column names of interest, which are all related to education inequality across different years.
- Then, by using the "fillna()" method, we are filling null values in the DataFrame "df" with the mean value of each respective numerical column from "numerical_cols".

In this step, we will handle null values which are present in categorical columns.

In [6]:
```
categorical_cols = ['ISO3','Country','Human Development Groups','UNDP Developing Regions']

for i in categorical_cols:
    mode = df[i].mode()[0]

df.fillna(mode,inplace=True)
```

- First, we created a list called "categorical_cols", containing the names of categorical columns in our DataFrame.
- Then, we are using a for loop to iterate through each column in "categorical_cols".
- Inside the loop, we are calculating the mode (most frequent value) for the current categorical column i using "df[i].mode()[0]".
- After that, we are filling null values in the "categorical columns" with the mode value which we calculated in the loop.

again, we will calculate percentage of null values.

In [7]:
```
df.isnull().sum()/len(df)*100
```

Out[7]:
```
ISO3                              0.0
Country                           0.0
Human Development Groups          0.0
UNDP Developing Regions           0.0
HDI Rank (2021)                   0.0
Inequality in Education (2010)    0.0
Inequality in Education (2011)    0.0
Inequality in Education (2012)    0.0
Inequality in Education (2013)    0.0
Inequality in Education (2014)    0.0
Inequality in Education (2015)    0.0
Inequality in Education (2016)    0.0
Inequality in Education (2017)    0.0
Inequality in Education (2018)    0.0
Inequality in Education (2019)    0.0
Inequality in Education (2020)    0.0
Inequality in Education (2021)    0.0
dtype: float64
```

as we can see, now we don't have null values in our data.

Analysis

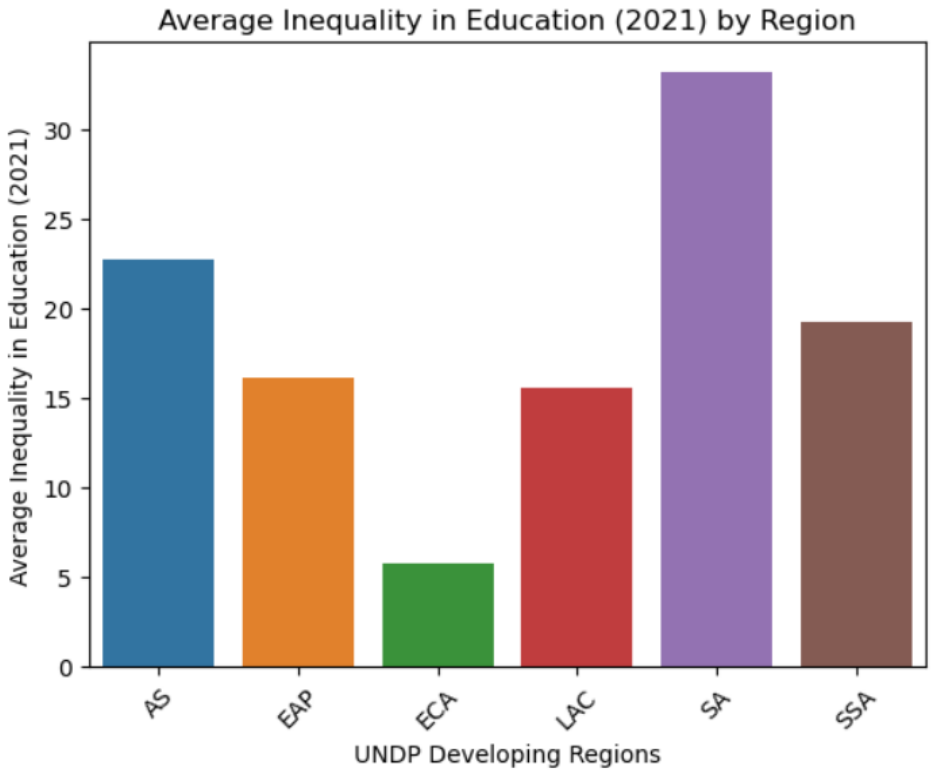
How does the average Inequality in Education (2021) compare between regions?

In [25]:
```
average_inequality_2021_by_region = df.groupby('UNDP Developing Regions')['Inequality in Education (2021)'].mean().reset_index()
average_inequality_2021_by_region
```

Out[25]:

	UNDP Developing Regions	Inequality in Education (2021)
0	AS	22.717882
1	EAP	16.111824
2	ECA	5.790199
3	LAC	15.563934
4	SA	33.189693
5	SSA	19.191285

In [26]:
```
#plot
sns.barplot(x='UNDP Developing Regions', y='Inequality in Education (2021)', data=average_inequality_2021_by_region)
plt.xticks(rotation=45)
plt.xlabel('UNDP Developing Regions')
plt.ylabel('Average Inequality in Education (2021)')
plt.title('Average Inequality in Education (2021) by Region')
plt.show()
```

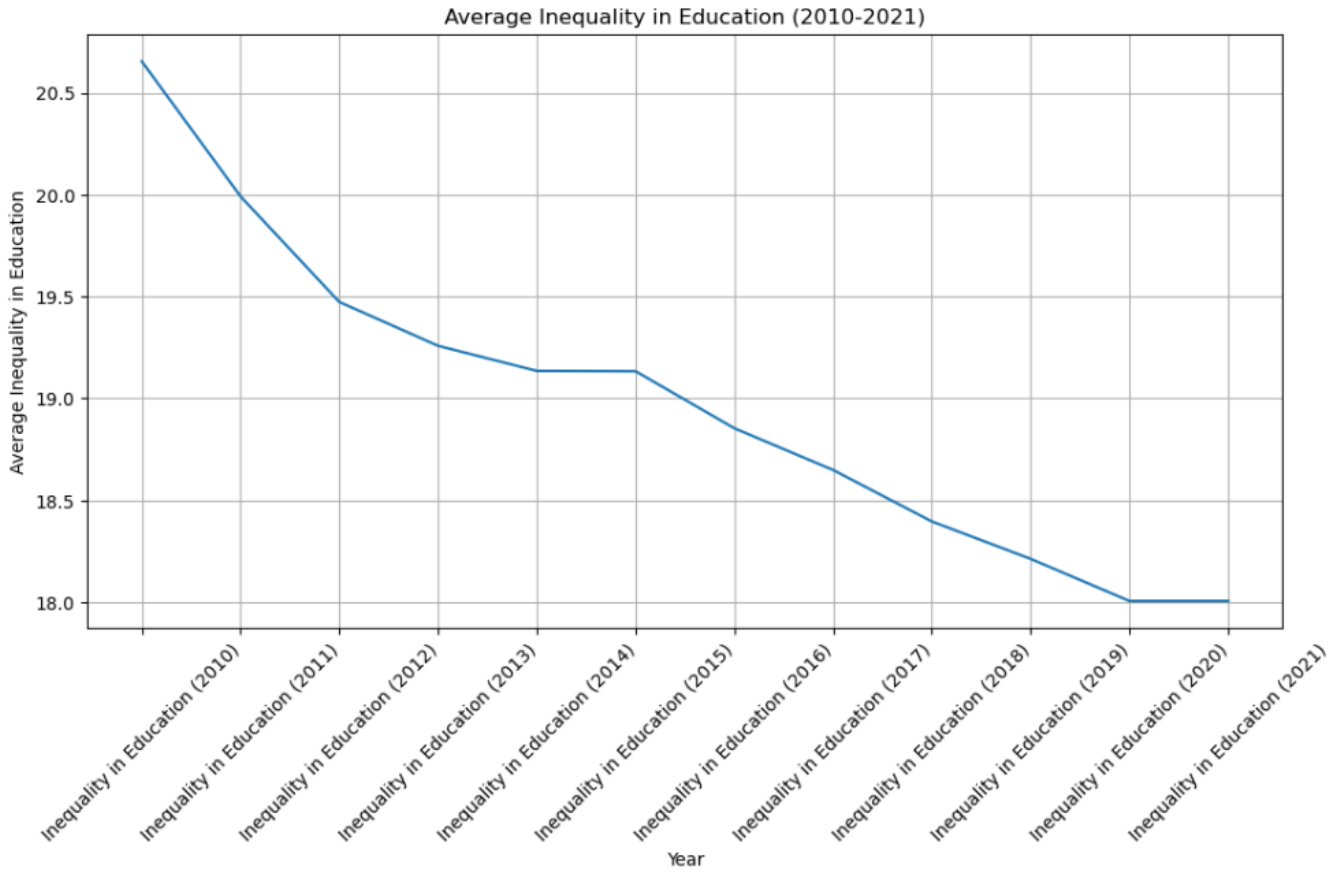


Average Inequality in Education for each year from 2010 to 2021

In [10]:
```
ai_per_year = df.loc[:, 'Inequality in Education (2010)': 'Inequality in Education (2021)'].mean()
ai_per_year
```

Out[10]:
```
Inequality in Education (2010)    20.654419
Inequality in Education (2011)    19.991823
Inequality in Education (2012)    19.473658
Inequality in Education (2013)    19.258472
Inequality in Education (2014)    19.135457
Inequality in Education (2015)    19.133751
Inequality in Education (2016)    18.853851
Inequality in Education (2017)    18.649140
Inequality in Education (2018)    18.396401
Inequality in Education (2019)    18.212993
Inequality in Education (2020)    18.006314
Inequality in Education (2021)    18.006314
dtype: float64
```

In [11]:
```
# Plot
plt.figure(figsize=(12, 6))
sns.lineplot(x=ai_per_year.index, y=ai_per_year.values)
plt.xticks(rotation=45)
plt.title('Average Inequality in Education (2010-2021)')
plt.xlabel('Year')
plt.ylabel('Average Inequality in Education')
plt.grid(True)
plt.show()
```



- The line plot shows the average inequality in education for each year from 2010 to 2021. It indicates a gradual decrease in the average inequality in education over this period.

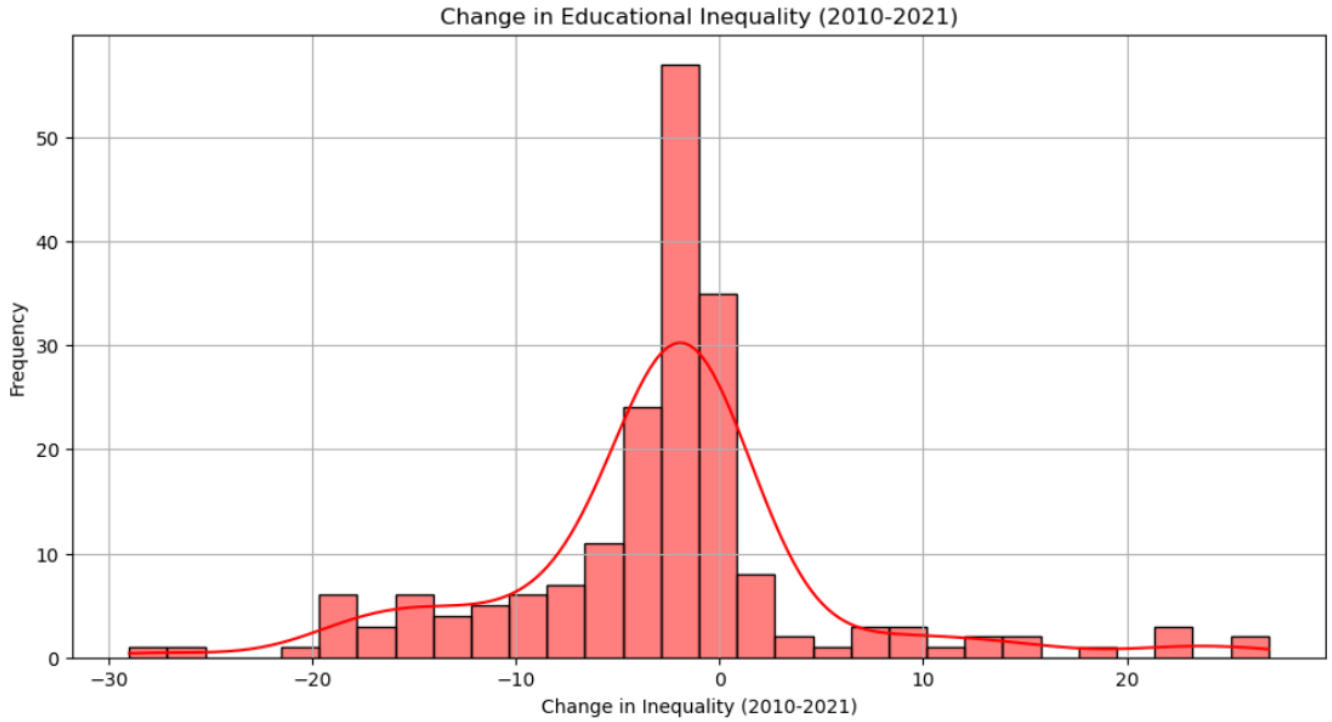
Change in Inequality in Education for each country from 2010 to 2021

In [28]:
```
# Calculate the change in Inequality in Education from 2010 to 2021
df['Change in Inequality (2010-2021)'] = df['Inequality in Education (2021)'] - df['Inequality in Education (2010)']

# Top 10 countries with the largest increase in Inequality in Education from 2010 to 2021
largest_increase_inequality = df.nlargest(10, 'Change in Inequality (2010-2021)')

# Top 10 countries with the largest decrease in Inequality in Education from 2010 to 2021
largest_decrease_inequality = df.nsmallest(10, 'Change in Inequality (2010-2021)')

largest_increase_inequality[['Country', 'Change in Inequality (2010-2021)']], largest_decrease_inequality[['Country', 'Change in
```

Out[28]:
```
(             Country  Change in Inequality (2010-2021)
 38            Comoros                         26.959961
 65             Gambia                         26.345631
 115              Mali                         23.205231
 13              Benin                         23.039131
 55           Ethiopia                         22.115941
 14       Burkina Faso                         18.186385
 140  Papua New Guinea                         14.997621
 126             Niger                         14.310411
 1              Angola                         13.517021
 49            Algeria                         12.628201,
         Country  Change in Inequality (2010-2021)
 45     Djibouti                         -28.967686
 158     Somalia                         -25.512476
 185   Uzbekistan                        -20.104309
 30   Switzerland                        -18.639849
 134        Oman                         -18.594060
 139       Palau                         -18.473653
 44       Germany                        -17.991219
 20       Belarus                        -17.882389
 62       Georgia                        -17.869614
 74       Croatia                        -16.412449)
```

- These results indicate that while some countries have made significant progress in reducing educational inequality, others have experienced an increase in inequality over the past decade.

In [31]:
```
# Plot
plt.figure(figsize=(12, 6))
sns.histplot(df['Change in Inequality (2010-2021)'], kde=True, bins=30, color='red')
plt.title('Change in Educational Inequality (2010-2021)')
plt.xlabel('Change in Inequality (2010-2021)')
plt.ylabel('Frequency')
plt.grid(True)
plt.show()
```



- This visualization provides insights into the overall change in educational inequality over the past decade. It is clear that while many countries have made progress in reducing educational inequality, there are still a significant number of countries where educational inequality has increased.

Which country has the highest HDI Rank (2021)?

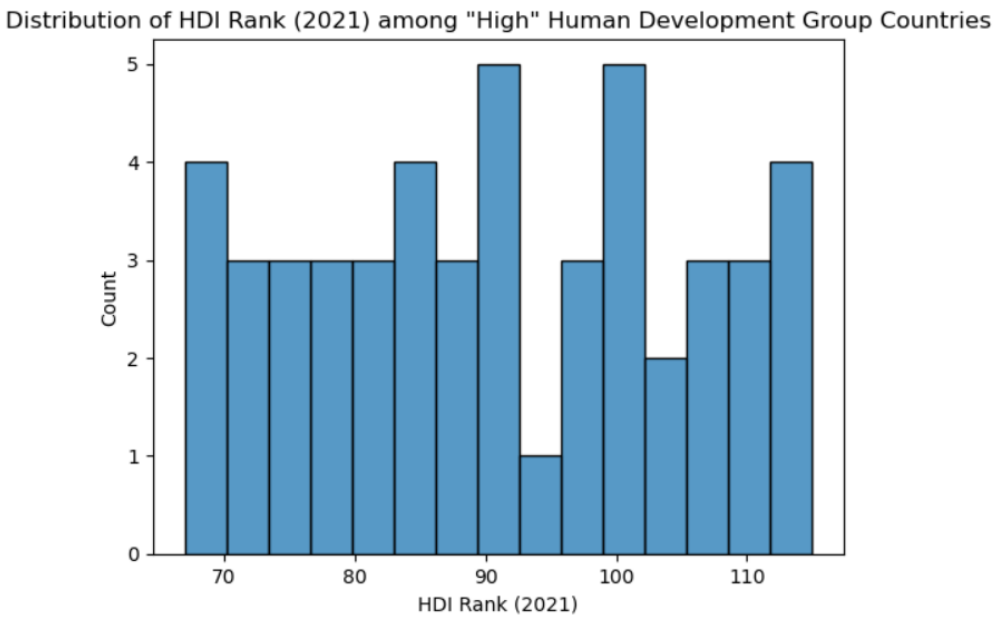
In [53]:
```
country_highest_hdi = df[df['HDI Rank (2021)'] == df['HDI Rank (2021)'].max()]['Country'].values[0]
print(f"The country with the highest HDI Rank (2021) is \n{country_highest_hdi}")
```

```
The country with the highest HDI Rank (2021) is
South Sudan
```

What is the distribution of HDI Rank (2021) among countries with "High" Human Development Groups?

In [60]:
```
high_hdi_group = df[df['Human Development Groups'] == 'High']
```

In [61]:
```
#plot
sns.histplot(data=high_hdi_group, x='HDI Rank (2021)', bins=15)
plt.xlabel('HDI Rank (2021)')
plt.ylabel('Count')
plt.title('Distribution of HDI Rank (2021) among "High" Human Development Group Countries')
plt.show()
```



- This histogram illustrates how the HDI Rank (2021) values are distributed among countries classified as "High" in terms of Human Development Groups.