EDA PROJECT - EDUCATION INEQUALITY

INTRODUCTION

In this project, we will conduct exploratory data analysis (EDA) to explore the issue of education inequality on a global scale. The data set is created from Human Development Reports.

First, import basic libraries for processing data and visualisation

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

DATA EXPLORATION

In [2]:
```
edu = pd.read_csv('Inequality in Education.csv')
edu.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)	Inequal Educati (20
0	AFG	Afghanistan	Low	SA	180.0	42.809000	44.823380	44.823380	44.8233
1	AGO	Angola	Medium	SSA	148.0	NaN	NaN	NaN	N
2	ALB	Albania	High	ECA	67.0	11.900000	11.900000	11.900000	11.900(
3	AND	Andorra	Very High	NaN	40.0	15.160302	15.160302	15.160302	15.1603
4	ARE	United Arab Emirates	Very High	AS	26.0	NaN	NaN	NaN	N

In [3]:
```
edu.tail()
```

Out[3]:

	ISO3	Country	Human Development Groups	UNDP Developing Regions	HDI Rank (2021)	Inequality in Education (2010)	Inequality in Education (2011)	Inequality in Education (2012)	Inequ in Educa (2
190	WSM	Samoa	High	EAP	111.0	NaN	NaN	NaN	
191	YEM	Yemen	Low	AS	183.0	48.09012	48.09012	48.09012	46.1
192	ZAF	South Africa	High	SSA	109.0	NaN	NaN	16.06077	16.0
193	ZMB	Zambia	Medium	SSA	154.0	23.76000	23.76000	23.76000	23.7
194	ZWE	Zimbabwe	Medium	SSA	146.0	17.82500	17.82500	17.82500	17.8

In [4]:

```
#Distribution of countries among different Human Development Groups
edu['Human Development Groups'].value_counts()
```

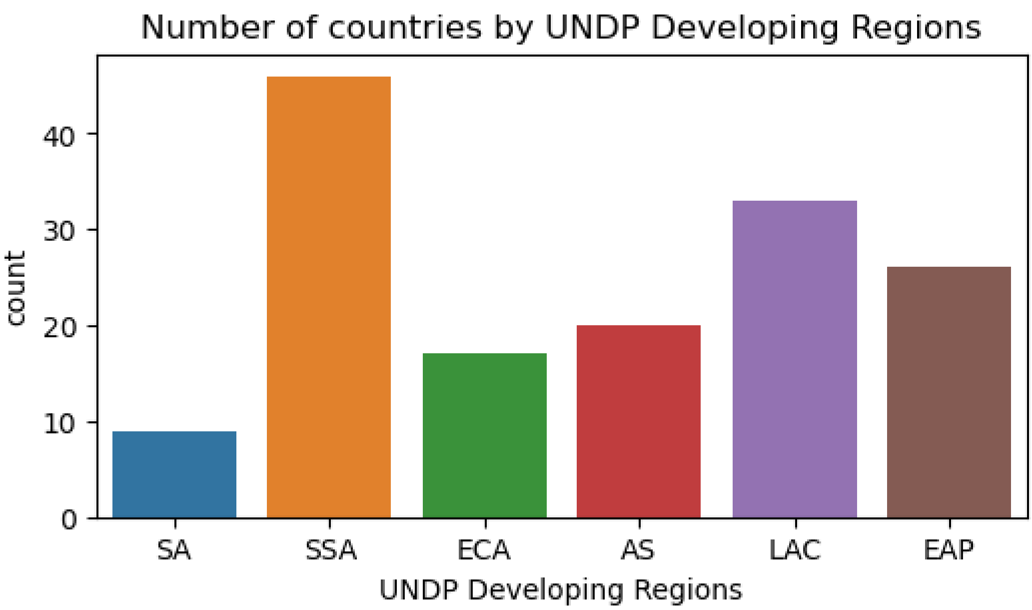
Out[4]:

```
Very High    66
High         49
Medium       44
Low          32
Name: Human Development Groups, dtype: int64
```

Most of the countries included in this report belong to high and very high development group (~60%)

In [5]:

```
#Distribution of countries among different UNDP Developing Regions
regions = edu['UNDP Developing Regions'].value_counts()
plt.figure(figsize = (6,3))
sns.countplot(x = 'UNDP Developing Regions', data = edu)
plt.title('Number of countries by UNDP Developing Regions')
plt.show()
```



The distribution of countries across different UNDP Developing Regions:

- Sub-Saharan Africa (SSA): 46 countries
- Latin America and the Caribbean (LAC): 33 countries
- East Asia and the Pacific (EAP): 26 countries

- Arab States (AS): 20 countries
- Europe and Central Asia (ECA): 17 countries
- South Asia (SA): 9 countries

Overall, the dataset contains these columns:

- ISO3: ISO code for the country/territory
- Country: Name of the country/territory
- Human Development Groups: Very High, High, Medium, Low
- UNDP Developing Regions: SSA, LAC, EAP, AS, ECA, SA
- HDI Rank (2021): Human Development Index Rank for 2021
- Inequality in Education (2010 - 2021): Inequality in education for reported countries from 2010 - 2021

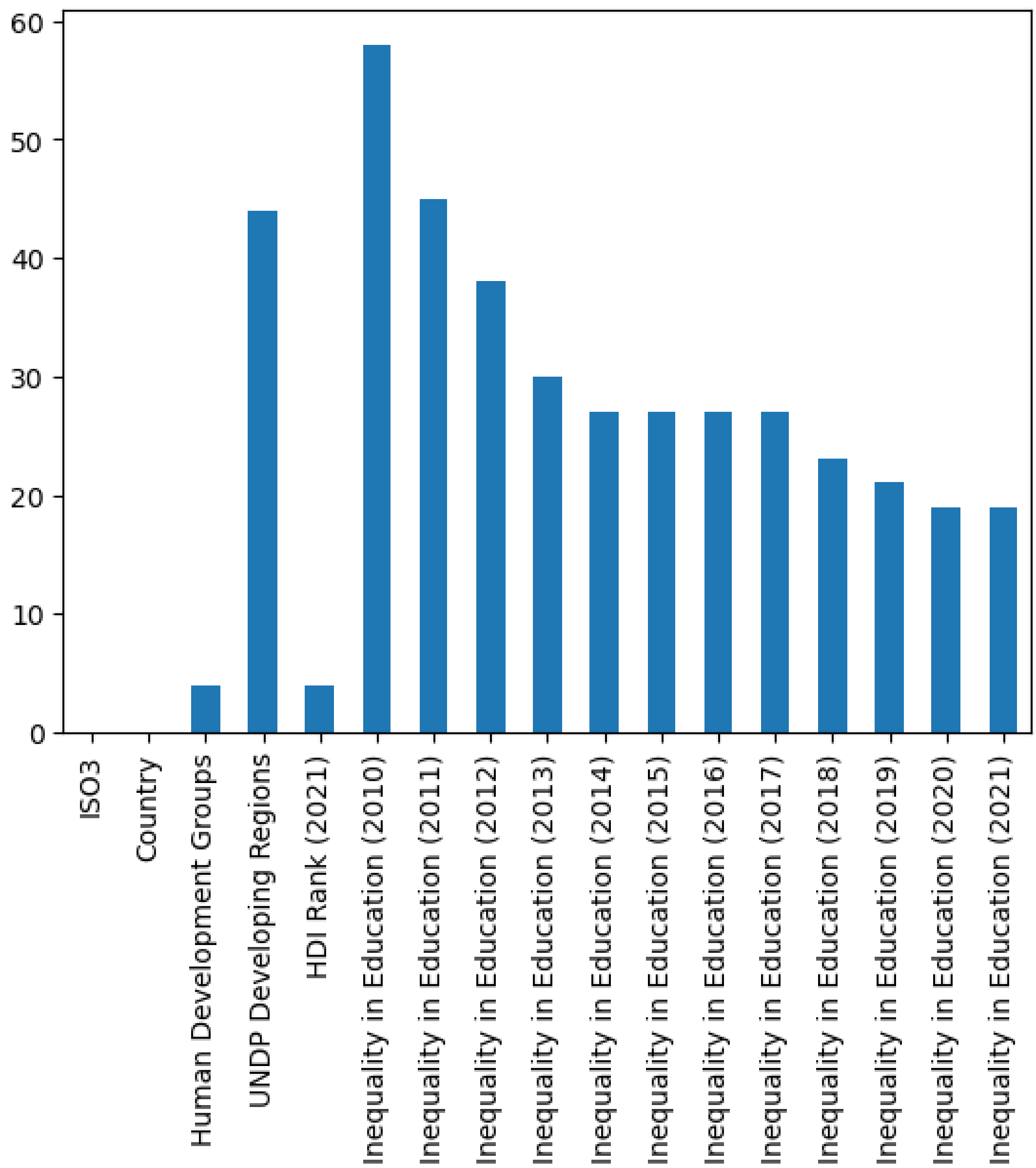
In [6]:
```
edu.info()
```

```
<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
```

In this dataset, there are 17 columns and 195 entries

Check null & duplicated values

In [7]:
```
edu.isnull().sum().plot.bar()
plt.show()
```



In [8]:
```
round(edu.isnull().sum() / 195, 3)
```

Out[8]:
```
ISO3                              0.000
Country                           0.000
Human Development Groups          0.021
UNDP Developing Regions           0.226
HDI Rank (2021)                   0.021
Inequality in Education (2010)    0.297
Inequality in Education (2011)    0.231
Inequality in Education (2012)    0.195
Inequality in Education (2013)    0.154
Inequality in Education (2014)    0.138
Inequality in Education (2015)    0.138
Inequality in Education (2016)    0.138
Inequality in Education (2017)    0.138
Inequality in Education (2018)    0.118
Inequality in Education (2019)    0.108
Inequality in Education (2020)    0.097
Inequality in Education (2021)    0.097
dtype: float64
```

Approximately 2 - 30% of the data are missing from each column, except for column Country and ISO3

In [9]:
```
edu.duplicated().sum()
```
Out[9]:
```
0
```

There is no duplicated values in the dataset

DATA CLEANING

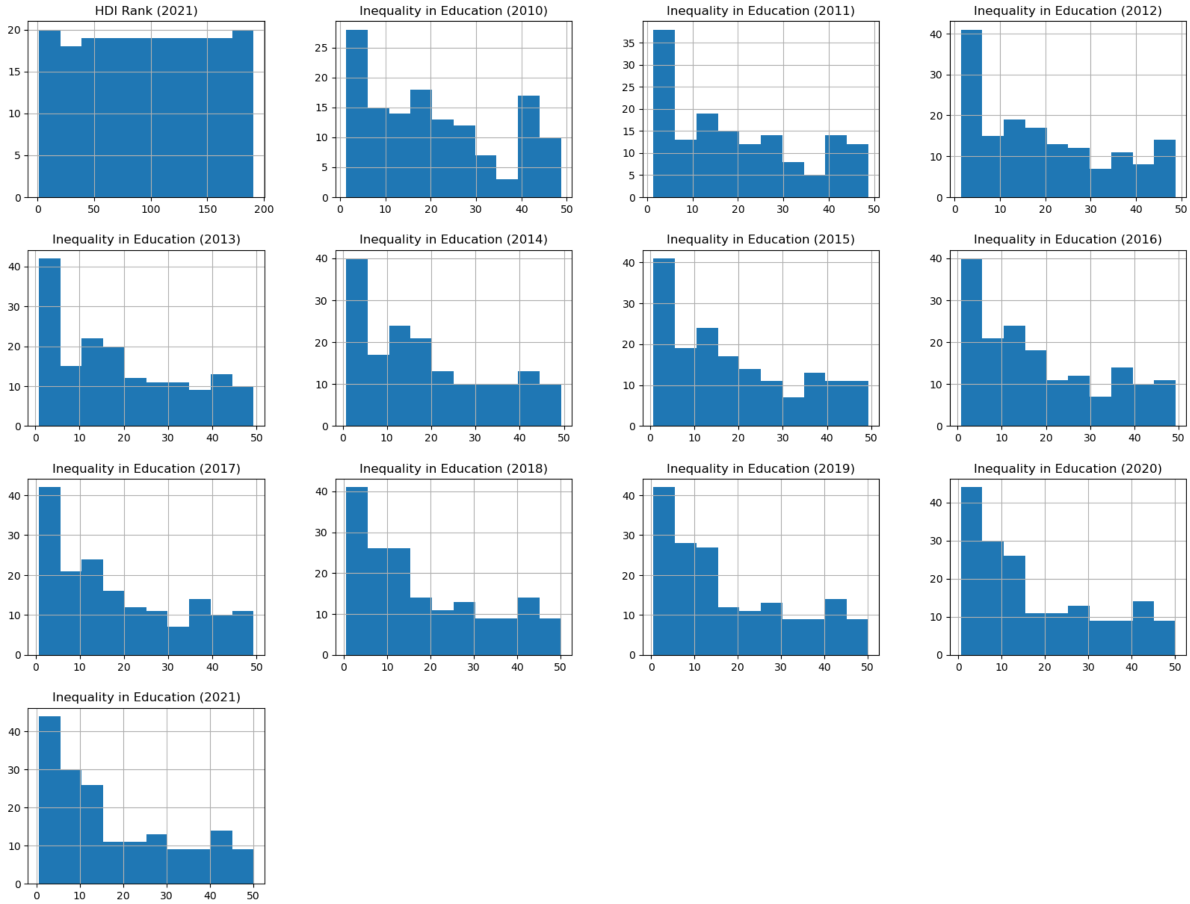
Here is some descriptive statistics for the dataset:

In [10]:
```
edu.describe(include = 'all')
```
Out[10]:

	ISO3	Country	Human Development Groups	UNDP Developing Regions	HDI Rank (2021)	Inequality in Education (2010)	Inequality in Education (2011)	Inequa Educat (20
count	195	195	191	151	191.000000	137.000000	150.000000	157.000(
unique	195	195	4	6	NaN	NaN	NaN	N
top	AFG	Afghanistan	Very High	SSA	NaN	NaN	NaN	N
freq	1	1	66	46	NaN	NaN	NaN	N
mean	NaN	NaN	NaN	NaN	95.811518	20.654419	19.991823	19.473(
std	NaN	NaN	NaN	NaN	55.307333	14.392552	14.342499	14.305
min	NaN	NaN	NaN	NaN	1.000000	1.322970	1.385640	1.390
25%	NaN	NaN	NaN	NaN	48.500000	6.917102	6.119250	6.011
50%	NaN	NaN	NaN	NaN	96.000000	17.825000	17.312742	16.421
75%	NaN	NaN	NaN	NaN	143.500000	30.542861	30.176057	30.201
max	NaN	NaN	NaN	NaN	191.000000	48.723000	48.723000	48.723(

Here is the distribution of numerical variables, mostly contain the education inequality scores between 2010 - 2021:

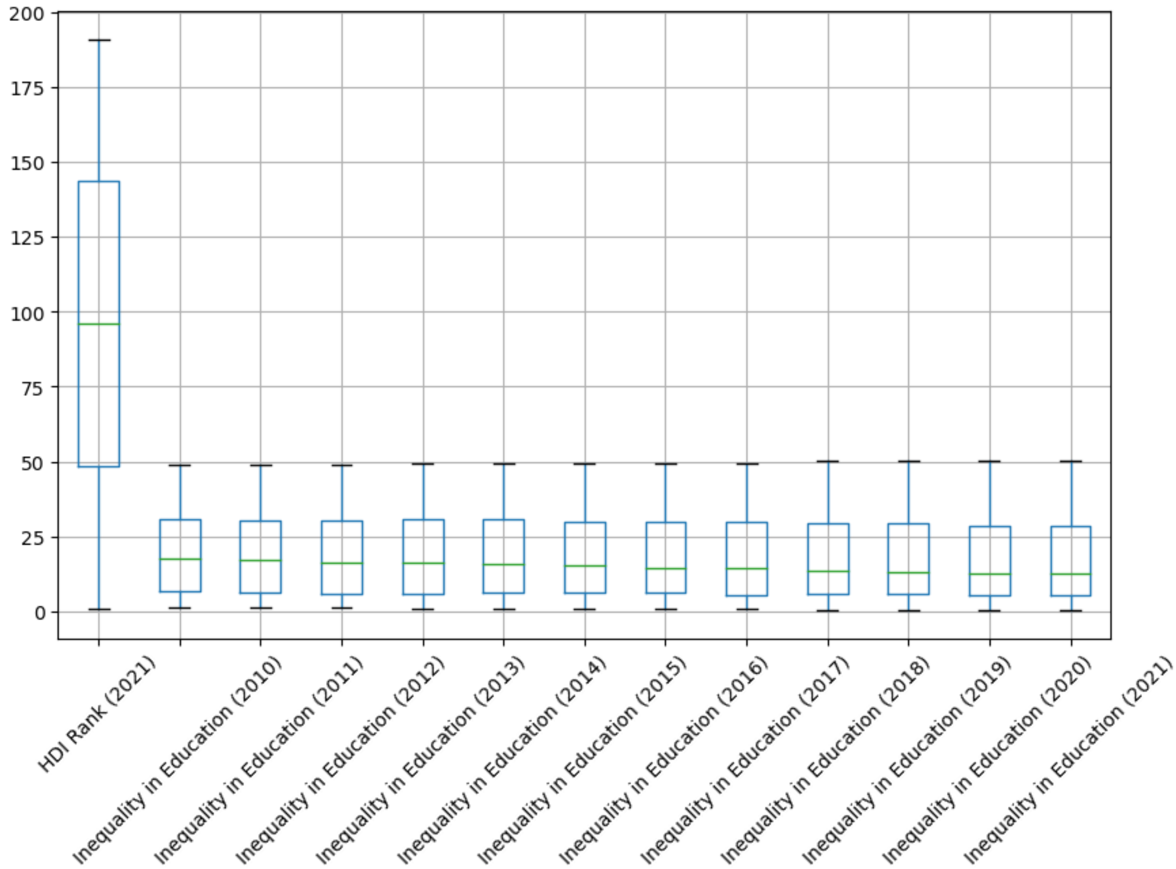
In [11]:
```
edu.hist(figsize = (20,15))
plt.show()
```



The histograms show us that the majority of inequality scores in years from 2010 to 2021 are postively skewed. There are more lower inequality scores (<= 20) as compared to higher inequality scores as recorded in the dataset.

Let's inspect the data range and any anomalies using boxplots

In [12]:
```
edu.boxplot(figsize = (10,6))
plt.xticks(rotation = 45)
plt.show()
```



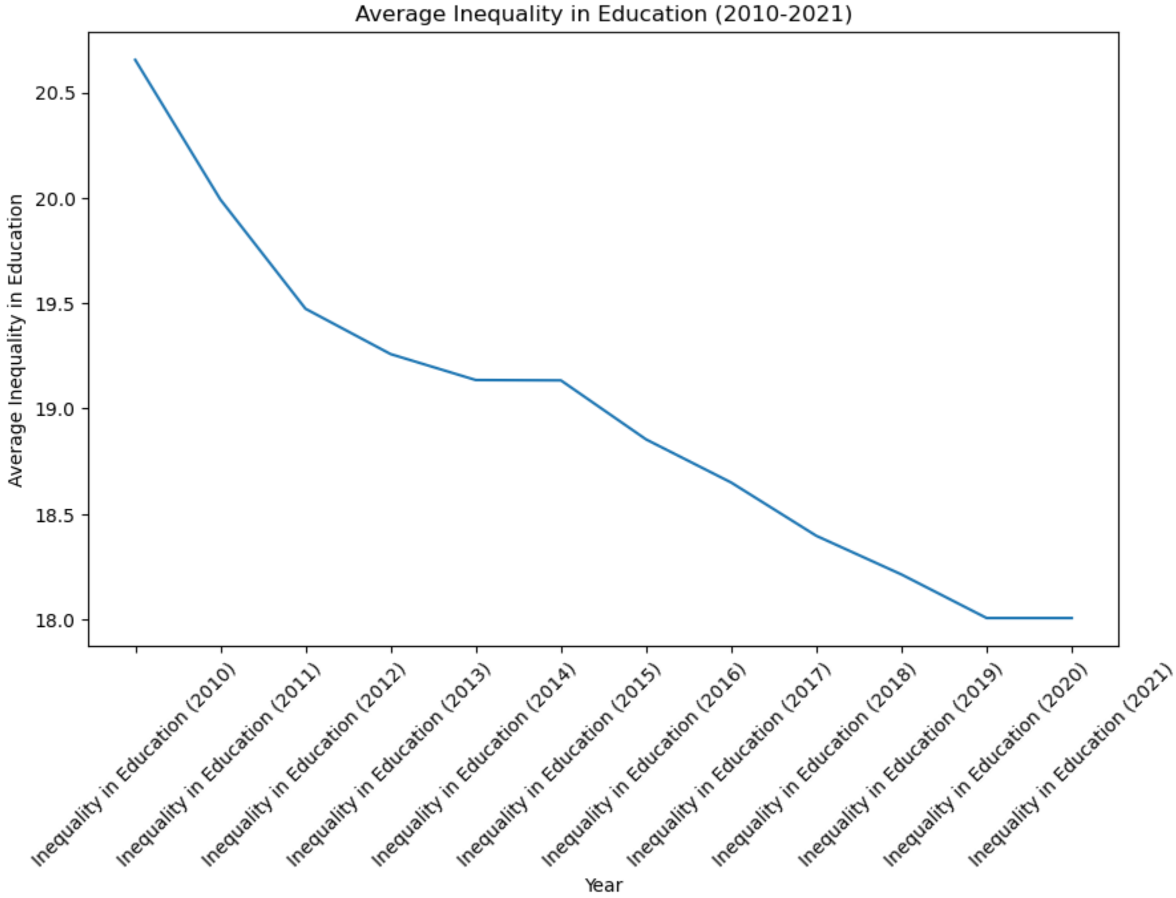
- There seems to be no significant data anomalies in the numerical variables
- However, both the distribution histograms and boxplots suggest that there is no significant change in inequality scores over time. The question is raised regarding the relevancy/accuracy of the dataset "Is this a good thing given all the world context (e.g: technology developments, increase GDP, etc) in recent years?"

DATA ANALYSIS

In [13]:

```
# Average Inequality in Education for each year from 2010 to 2021
mean_inequality_per_year = edu.loc[:, 'Inequality in Education (2010)': 'Inequality

# Plot
plt.figure(figsize=(10,6))
sns.lineplot(x=mean_inequality_per_year.index, y=mean_inequality_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.show()
```



The line plot shows the average inequality in education for each year from 2010 up to 2021. It shows a graduate decrease in the average inequality in education over that period. This may indiciate a favourable trend for global education for the past decade, even though it still remains a significant issue.

Some things to note:

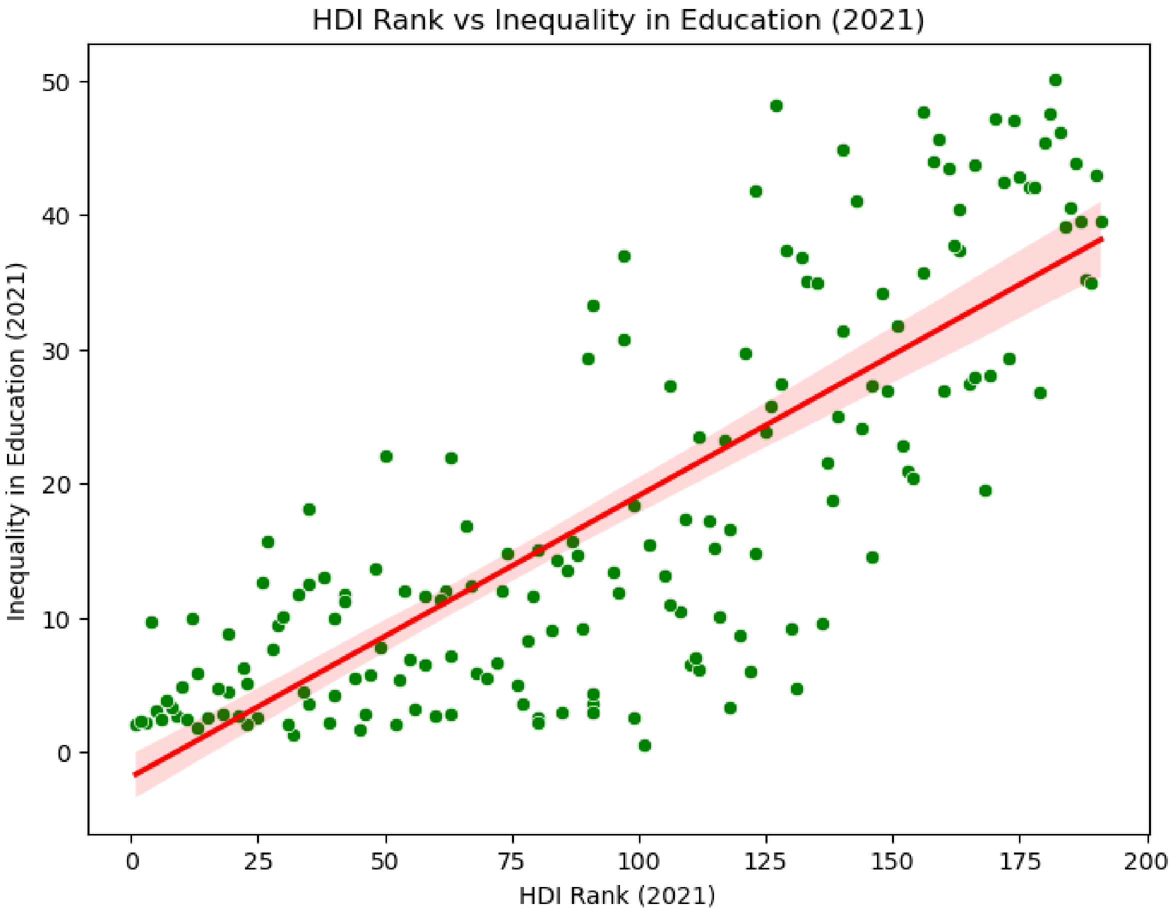
- The overall inequality does not change over time
- Inequality in education declined => This means there might be an increase in inequality in other areas that may be worth digging deeper

In [14]:

```
#Scatter plot for HDI rank(2021) and Inequality in Education in 2021
# Plot
plt.figure(figsize=(8, 6))
sns.scatterplot(x='HDI Rank (2021)', y='Inequality in Education (2021)', data=edu,


# Add a regression line (line of best fit)
sns.regplot(x='HDI Rank (2021)', y='Inequality in Education (2021)', data=edu, scat


plt.title('HDI Rank vs Inequality in Education (2021)')
plt.xlabel('HDI Rank (2021)')
plt.ylabel('Inequality in Education (2021)')
plt.show()
```



The heat map and scatter plot show the relationship between the Human Development Index (HDI) rank and inequality in education score in 2021 for all countries. As indiciated on the scatter plot, an increase in education inequality is positively associated with an increase in HDI score

In [15]:
```
#Change in inequality in education for each country from 2010 to 2021
edu['Change in Inequality'] = edu['Inequality in Education (2021)'] - edu['Inequali
```

In [16]:
```
#Top 10 countries with the highest increase in education inequality
edu.nlargest(10,'Change in Inequality')
```

Out[16]:

	ISO3	Country	Human Development Groups	UNDP Developing Regions	HDI Rank (2021)	Inequality in Education (2010)	Inequality in Education (2011)	Inequality in Education (2012)	Ine Edu
14	BFA	Burkina Faso	Low	SSA	184.0	20.966409	20.966409	20.966409	20.
120	MOZ	Mozambique	Low	SSA	185.0	30.920347	30.920347	30.920347	30.
64	GIN	Guinea	Low	SSA	182.0	42.000000	42.000000	48.265360	48.
117	MMR	Myanmar	Medium	EAP	149.0	19.440000	19.440000	19.440000	19.
26	BTN	Bhutan	Medium	SA	127.0	44.810670	44.810670	44.810670	44.
152	SEN	Senegal	Low	SSA	170.0	44.579000	44.579000	44.579000	44.
0	AFG	Afghanistan	Low	SA	180.0	42.809000	44.823380	44.823380	44.
33	CIV	Ivory Coast	Medium	SSA	159.0	43.200000	43.200000	45.111420	45.
121	MRT	Mauritania	Medium	SSA	158.0	42.060000	40.785160	40.785160	40.
66	GNB	Guinea-Bissau	Low	SSA	177.0	40.315000	40.315000	40.315000	40.

In [17]:
```
edu.nsmallest(10,'Change in Inequality')
```

Out[17]:

	ISO3	Country	Human Development Groups	UNDP Developing Regions	HDI Rank (2021)	Inequality in Education (2010)	Inequality in Education (2011)	Inequality in Education (2012)	Inequ Educ (2
134	OMN	Oman	Very High	AS	54.0	30.542861	30.542861	30.542861	30.54
92	KIR	Kiribati	Medium	EAP	136.0	21.380000	21.380000	21.380000	21.38
111	MDV	Maldives	High	SA	90.0	39.965570	39.965570	39.965570	39.96
17	BHR	Bahrain	Very High	AS	35.0	22.154491	22.154491	22.154491	22.15
57	FJI	Fiji	High	EAP	99.0	11.854388	11.854388	11.854388	11.85
114	MKD	North Macedonia	High	ECA	78.0	17.466000	10.537340	10.537340	10.53
187	VEN	Venezuela	Medium	LAC	120.0	17.532918	13.359263	13.359263	13.35
27	BWA	Botswana	Medium	SSA	117.0	32.082000	32.082000	32.082000	32.08
146	QAT	Qatar	Very High	AS	42.0	19.053505	19.881821	18.427006	14.11
22	BOL	Bolivia	Medium	LAC	118.0	23.671900	21.624900	20.535670	19.90

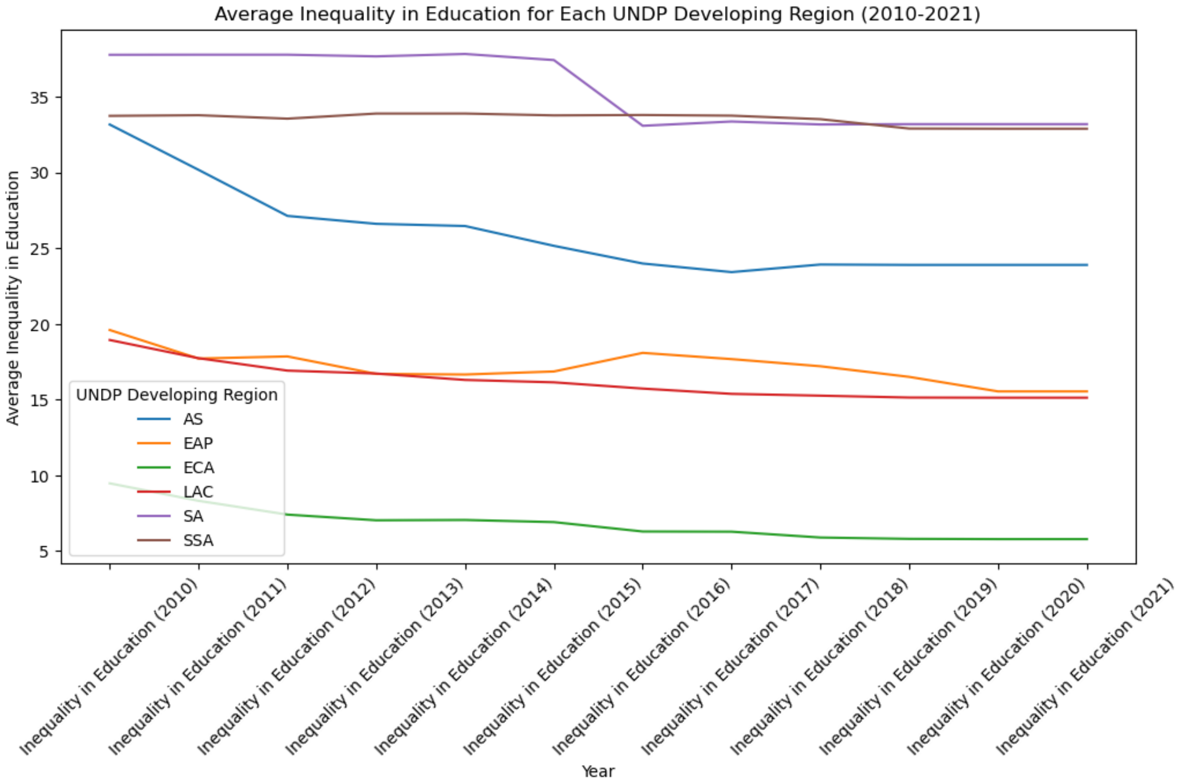
- The development group is determined by considering multiple factors, which include the inequality. As a result, countries with higher inequality will belong to the lower end groups and vice versa.
- Most of the top 10 countries with increasing inequality (70%) are from SSA region. However, the highlight is that Botswana, which is in the same region but got the top improvement in inequality.

In [18]:
```
#Analysis by regions

# Group by UNDP Developing Regions and compute the mean for each year
region_mean = edu.groupby('UNDP Developing Regions').mean()

# Select only the Inequality in Education columns
region_mean = region_mean.loc[:, 'Inequality in Education (2010)': 'Inequality in E

# Plot
plt.figure(figsize=(12, 6))
sns.lineplot(data=region_mean.T, dashes=False)
plt.xticks(rotation=45)
plt.title('Average Inequality in Education for Each UNDP Developing Region (2010-20
plt.xlabel('Year')
plt.ylabel('Average Inequality in Education')
plt.legend(title='UNDP Developing Region', labels=region_mean.index)
plt.show()
```



Over the years, Europe and Central Asia (ECA) shows lowest average inequality score in education, followed by Latin America and The Caribbean region (LAC) Meanwhile, South Asia (SA) and followed by Sub-Saharan Africa (SSA) shows highest average inequality score in education. However, SA and AS show a gradually decreasing trend in average inequality score over the years. Meanwhile, SSA average inequality score in education tends to increase from 2012 as compared to 2010, and remain stable at high score since then till 2021.

Recommendations:

- Policymakers may consider cross-regional collaboration and knowledge exchange to share successful strategies for reducing educational inequality.
- Targeted interventions should address the specific challenges faced by high-inequality regions, focusing on improving access, quality, and inclusivity in education.
- Monitoring and evaluation systems should be strengthened to assess the impact of policies over time and make data-driven adjustments to educational strategies.

- Investment in Research: Policymakers could invest in research to better understand the contextual factors contributing to educational inequality in each region, enabling the development of more targeted and effective interventions.

IT'S THE END OF THE REPORT. THANKS FOR YOUR ATTENTION.