In [ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import plotly.express as px
%matplotlib inline
```

In [ ]:
```
pd.set_option('display.max_rows', 10)
pd.options.display.max_columns = 100
pd.set_option("display.precision", 2)
```

Data Loading and cleaning

Initial Analysis of the Datasets

1. Income inequality

In [ ]:
```
income = pd.read_csv('../CSV/income_per_person_gdppercapita_ppp_inflation_ad
```

In [ ]:
```
print("number of rows: ", income.shape[0])
print("number of columns: {}".format(income.shape[1]))
print("number of duplicates: {}".format(income.duplicated().sum()))
print("datatypes:\n")
print(income.dtypes)
income.head(3)
```

```
number of rows:  193
number of columns: 242
number of duplicates: 0
datatypes:

country    object
1800        int64
1801        int64
1802        int64
1803        int64
            ...
2036        int64
2037        int64
2038        int64
2039        int64
2040        int64
Length: 242, dtype: object
```

Out[ ]:

	country	1800	1801	1802	1803	1804	1805	1806	1807	1808	1809	1810	1811
0	Afghanistan	603	603	603	603	603	603	603	603	603	603	604	604
1	Albania	667	667	667	667	667	668	668	668	668	668	668	668
2	Algeria	715	716	717	718	719	720	721	722	723	724	725	726

3 rows × 242 columns

The results below show that there no nulls.

In [ ]:
```
income_null = income.isnull().sum()/income.shape[0]
income_null.to_frame().transpose()
```

Out[ ]:

	country	1800	1801	1802	1803	1804	1805	1806	1807	1808	1809	1810	1811	18
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

1 rows × 242 columns

2. Tax Revenue as a Percent of GDP

In [ ]:
```
tax = pd.read_csv('../CSV/tax_revenue_percent_of_gdp.csv')
```

In [ ]:
```
print("number of rows: ", tax.shape[0])
print("number of columns: {}".format(tax.shape[1]))
print("number of duplicates: {}".format(tax.duplicated().sum()))
print("datatypes:\n")
print(tax.dtypes)
print("\nSample:")
tax.head(3)
```

```
number of rows:  161
number of columns: 47
number of duplicates: 0
datatypes:

country     object
1972       float64
1973       float64
1974       float64
1975       float64
            ...
2013       float64
2014       float64
2015       float64
2016       float64
2017       float64
Length: 47, dtype: object

Sample:
```

Out[ ]:

	country	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983
0	Afghanistan	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	Albania	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	Algeria	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

An initial analysis revealed that about half the years with > 0.50 missing values. See below:

In [ ]:
```
tax_null = tax.isnull().sum()/tax.shape[0]
tax_null.to_frame().transpose()
```

Out[ ]:

	country	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	19
0	0.0	0.76	0.72	0.72	0.71	0.71	0.7	0.7	0.7	0.68	0.68	0.7	0.69	0

3. Gini dataset

In [ ]:
```
gini = pd.read_csv('../CSV/gini.csv')
```

In [ ]:
```
print("number of rows: ", gini.shape[0])
print("number of columns: {}".format(gini.shape[1]))
print("number of duplicates: {}".format(gini.duplicated().sum()))
print("datatypes:\n")
print(gini.dtypes)
print("\nSample:")
gini.head(3)
```

```
number of rows:  195
number of columns: 242
number of duplicates: 0
datatypes:

country     object
1800       float64
1801       float64
1802       float64
1803       float64
            ...
2036       float64
2037       float64
2038       float64
2039       float64
2040       float64
Length: 242, dtype: object

Sample:
```

Out[ ]:

	country	1800	1801	1802	1803	1804	1805	1806	1807	1808	1809	1810	1811
0	Afghanistan	30.5	30.5	30.5	30.5	30.5	30.5	30.5	30.5	30.5	30.5	30.5	30.5
1	Albania	38.9	38.9	38.9	38.9	38.9	38.9	38.9	38.9	38.9	38.9	38.9	38.9
2	Algeria	56.2	56.2	56.2	56.2	56.2	56.2	56.2	56.2	56.2	56.2	56.2	56.2

3 rows × 242 columns

The results below show that there were no nulls in the dataset.

In [ ]:
```
gini_null = gini.isnull().sum()/gini.shape[0]
gini_null.to_frame().transpose()
```

Out[ ]:

	country	1800	1801	1802	1803	1804	1805	1806	1807	1808	1809	1810	1811
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

1 rows × 242 columns

4. Investment Percent of GDP

Below are results of an initial analysis:

In [ ]:
```
invest = pd.read_csv('../CSV/investments_percent_of_gdp.csv')
```

In [ ]:
```
print("number of rows: ", invest.shape[0])
print("number of columns: {}".format(invest.shape[1]))
print("number of duplicates: {}".format(invest.duplicated().sum()))
print("datatypes:\n")
print(invest.dtypes)
print("\nSample:")
invest.head(3)
```

```
number of rows:  177
number of columns: 59
number of duplicates: 0
datatypes:

country     object
1960       float64
1961       float64
1962       float64
1963       float64
            ...
2013       float64
2014       float64
2015       float64
2016       float64
2017       float64
Length: 59, dtype: object

Sample:
```

Out[ ]:

	country	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	19
0	Afghanistan	16.1	16.6	19.1	14.2	13.9	11.3	8.41	5.18	6.47	6.47	5.46	5.
1	Albania	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
2	Algeria	42.2	47.2	35.4	28.9	21.8	22.6	17.30	23.40	27.90	32.40	36.60	35.

In [ ]:
```
invest_null = invest.isnull().sum()/invest.shape[0]
invest_null.to_frame().transpose()
```

Out[ ]:

	country	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971
0	0.0	0.63	0.63	0.63	0.63	0.62	0.54	0.53	0.51	0.51	0.51	0.44	0.44

5. EIU Democracy Index

Below are the results of the initial analysis:

In [ ]:
```
demo = pd.read_csv('../CSV/demox_eiu.csv')
```

In [ ]:
```
print("number of rows: ", demo.shape[0])
print("number of columns: {}".format(demo.shape[1]))
print("number of duplicates: {}".format(demo.duplicated().sum()))
print("datatypes:\n")
print(demo.dtypes)
print("\nSample:")
demo.head(3)
```

```
number of rows:  164
number of columns: 14
number of duplicates: 0
datatypes:

country     object
2006       float64
2007       float64
2008       float64
2009       float64
            ...
2014       float64
2015       float64
2016       float64
2017       float64
2018       float64
Length: 14, dtype: object

Sample:
```

Out[ ]:

	country	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	201
0	Afghanistan	30.6	30.4	30.2	27.5	24.8	24.8	24.8	24.8	27.7	27.7	25.5	25.
1	Albania	59.1	59.1	59.1	58.9	58.6	58.1	56.7	56.7	56.7	59.1	59.1	59.
2	Algeria	31.7	32.5	33.2	33.8	34.4	34.4	38.3	38.3	38.3	39.5	35.6	35.

The results below show that there were no nulls in the dataset.

In [ ]:
```
demo_null = demo.isnull().sum()/demo.shape[0]
demo_null.to_frame().transpose()
```

Out[ ]:

	country	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Because of the results below, the years with the lowest null value percentage were in the 10 year range from 2006-2016. As a result, this yearly range was selected for the rest of the datasets.

Slicing and Reorganizing the Datasets

Income Per Person (GDP per Capita)

In [ ]:
```
income_last_10 = income.iloc[:, np.r_[:1, 207:218]]
income_last_10
```

Out[ ]:

	country	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
0	Afghanistan	1120	1250	1270	1500	1670	1630	1770	1810	1800	1770
1	Albania	7910	8450	9160	9530	9930	10200	10400	10500	10700	11000
2	Algeria	12400	12600	12700	12700	12900	13000	13200	13300	13500	13800
3	Andorra	42700	43400	41400	41700	39000	42000	41900	43700	44900	46600
4	Angola	5500	6040	6470	6290	6360	6350	6650	6730	6810	6650
...	...	...	...	...	...	...	...	...	...	...	...
188	Venezuela	16400	17600	18200	17400	16900	17300	18000	18000	17100	15600
189	Vietnam	3630	3850	4030	4210	4430	4660	4860	5070	5310	5610
190	Yemen	4270	4290	4320	4360	4570	3880	3860	3940	3830	3110
191	Zambia	2650	2800	2930	3120	3340	3420	3570	3630	3690	3680
192	Zimbabwe	1890	1810	1480	1630	1930	2170	2490	2490	2510	2510

193 rows × 12 columns

In [ ]:
```
income_last_10 = income_last_10.melt(id_vars=['country'], var_name='year', v
income_last_10.sort_values(['country', 'year'], inplace=True)
```

In [ ]:
```
income_last_10.head(3)
```

Out[ ]:

	country	year	income_per_person
0	Afghanistan	2006	1120
193	Afghanistan	2007	1250
386	Afghanistan	2008	1270

Investment Percent of GDP

In [ ]:
```
invest_last_10 = invest.iloc[:, np.r_[:1, 47:58]]
invest_last_10 = invest_last_10.melt(id_vars=['country'], var_name='year', v
invest_last_10.sort_values(['country', 'year'], inplace=True)
invest_last_10.head(3)
```

Out[ ]:

	country	year	invest_%_gdp
0	Afghanistan	2006	23.4
177	Afghanistan	2007	19.9
354	Afghanistan	2008	18.9

Tax Revenue Percent of GDP

In [ ]:
```
tax_last_10 = tax.iloc[:, np.r_[:1, 35:46]]
tax_last_10 = tax_last_10.melt(id_vars=['country'], var_name='year', value_r
tax_last_10.sort_values(['country', 'year'], inplace=True)
tax_last_10.head(3)
```

Out[ ]:

	country	year	tax_%_gdp
0	Afghanistan	2006	6.88
161	Afghanistan	2007	5.23
322	Afghanistan	2008	6.04

Gini Index

In [ ]:
```
gini_last_10 = gini.iloc[:, np.r_[:1, 207:218]]
gini_last_10 = gini_last_10.melt(id_vars=['country'], var_name='year', value
gini_last_10.sort_values(by=['country', 'year'], inplace=True)
gini_last_10.head(3)
```

Out[ ]:

	country	year	gini_index
0	Afghanistan	2006	36.8
195	Afghanistan	2007	36.8
390	Afghanistan	2008	36.8

EIU Democracy Index

In [ ]:
```
demo_last_10 = demo.iloc[:, :-2]
demo_last_10 = demo_last_10.melt(id_vars=['country'], var_name='year', value
demo_last_10.sort_values(['country','year'], inplace=True)
demo_last_10.head(3)
```

Out[ ]:

	country	year	demox_eiu
0	Afghanistan	2006	30.6
164	Afghanistan	2007	30.4
328	Afghanistan	2008	30.2

Merging the Datasets

In [ ]:
```
combined = demo_last_10.merge(income_last_10, left_on=['country', 'year'], r
combined = combined.merge(invest_last_10, left_on=['country', 'year'], righ
combined = combined.merge(tax_last_10, left_on=['country', 'year'], right_o
combined = combined.merge(gini_last_10, left_on=['country', 'year'], right_
combined
```

Out[ ]:

	country	year	demox_eiu	income_per_person	invest_%_gdp	tax_%_gdp	gini_ir
0	Afghanistan	2006	30.6	1120	23.4	6.88	
1	Afghanistan	2007	30.4	1250	19.9	5.23	
2	Afghanistan	2008	30.2	1270	18.9	6.04	
3	Afghanistan	2009	27.5	1500	17.9	8.44	
4	Afghanistan	2010	24.8	1670	17.9	9.12	
...	...	...	...	...	...	...	
1524	Zimbabwe	2012	26.7	2490	11.8	21.40	
1525	Zimbabwe	2013	26.7	2490	11.4	NaN	
1526	Zimbabwe	2014	27.8	2510	11.8	NaN	
1527	Zimbabwe	2015	30.5	2510	12.3	NaN	
1528	Zimbabwe	2016	30.5	2490	12.2	NaN	

1529 rows × 7 columns

In [ ]:

```
cont = pd.read_csv('../CSV/continent_country.csv')
cont
```

Out[ ]:

	continent	country
0	Africa	Congo, Dem. Rep.
1	Africa	Congo, Rep.
2	Africa	Algeria
3	Africa	Angola
4	Africa	Benin
...	...	...
165	Europe	Ukraine
166	Europe	United Kingdom
167	Oceania	Australia
168	Oceania	New Zealand
169	continent	country

170 rows × 2 columns

Matching Country with Continent

In this step, we match each country with its continent. This will enable analysis at the continent level for broader trend detection.

In [ ]:

```
combined_final = cont.merge(combined, left_on=['country'], right_on=['count
combined_final
```

Out[ ]:

	continent	country	year	demox_eiu	income_per_person	invest_%_gdp	tax_%_gdp
0	Africa	Congo, Dem. Rep.	2006	27.6	605	14.6	6.83
1	Africa	Congo, Dem. Rep.	2007	25.2	623	13.7	6.99
2	Africa	Congo, Dem. Rep.	2008	22.8	640	10.9	8.97
3	Africa	Congo, Dem. Rep.	2009	22.1	637	14.6	7.89
4	Africa	Congo, Dem. Rep.	2010	21.5	660	28.8	8.35
...	...	...	...	...	...	...	...
1524	Oceania	New Zealand	2012	92.6	33300	20.9	26.80
1525	Oceania	New Zealand	2013	92.6	33900	22.0	26.60
1526	Oceania	New Zealand	2014	92.6	34600	22.9	26.80
1527	Oceania	New Zealand	2015	92.6	35200	23.4	27.40
1528	Oceania	New Zealand	2016	92.6	35700	24.4	27.20

1529 rows × 8 columns

Data Cleaning

Below are the steps taken to ensure quality of the dataset:

Missing Values

Below are is a summary of missing values (nulls) in the dataset:

In [ ]:
```
combined_final.isna().sum()
```

Out[ ]:
```
continent              0
country                0
year                   0
income_per_person      0
tax_%_gdp            292
gini_index             0
dtype: int64
```

One option for handling the missing 'tax_%_gdp' values would be to replace them with the country's mean. However, some of the countries have all nulls and some have mostly nulls for this column.

A second option is to drop the rows with nulls. In the interest of simplicity, we will use this option.

In [ ]:
```
combined_final.dropna(inplace=True)
combined_final.isna().sum()
```

Out[ ]:
```
continent           0
country             0
year                0
income_per_person   0
tax_%_gdp           0
gini_index          0
dtype: int64
```

Duplicates

There are no duplicates in the dataset:

In [ ]:
```
combined_final.duplicated().sum()
```

Out[ ]:
```
0
```

Descriptive Statistics

Below are descriptive statistics of the dataset. A review of the values indicates that the min, max and mean values appear to be reasonable.

In [ ]:
```
combined_final.describe()
```

Out[ ]:

	demox_eiu	income_per_person	invest_%_gdp	tax_%_gdp	gini_index
count	1529.00	1529.00	1498.00	1244.00	1529.00
mean	58.54	18514.42	24.42	16.79	38.46
std	21.15	19857.67	8.43	7.65	8.19
min	14.30	605.00	0.00	0.04	24.40
25%	39.50	4160.00	19.40	12.40	32.40
50%	61.50	11400.00	23.20	15.90	37.10
75%	76.10	27500.00	27.90	21.32	42.80
max	99.30	124000.00	70.70	62.90	63.90

Save the Cleaned Dataset

In [ ]:
```
combined_final.to_csv('combined_final_last_10_years.csv', index=False)
```

Exploratory Data Analysis

Research Question 1 - Is Income Inequality Getting Worse or Better in the Last 10 Years?

Better means the Gini Index is going down.

Global Gini Mean By Year

In [ ]:
```
columns = ['year', 'gini_index']
gini = combined_final[columns]
gini
```

Out[ ]:

	year	gini_index
0	2006	42.2
1	2007	42.1
2	2008	42.1
3	2009	42.1
4	2010	42.1
...	...	...
1546	2012	33.5
1547	2013	34.0
1548	2014	34.0
1549	2015	34.5
1550	2016	34.8

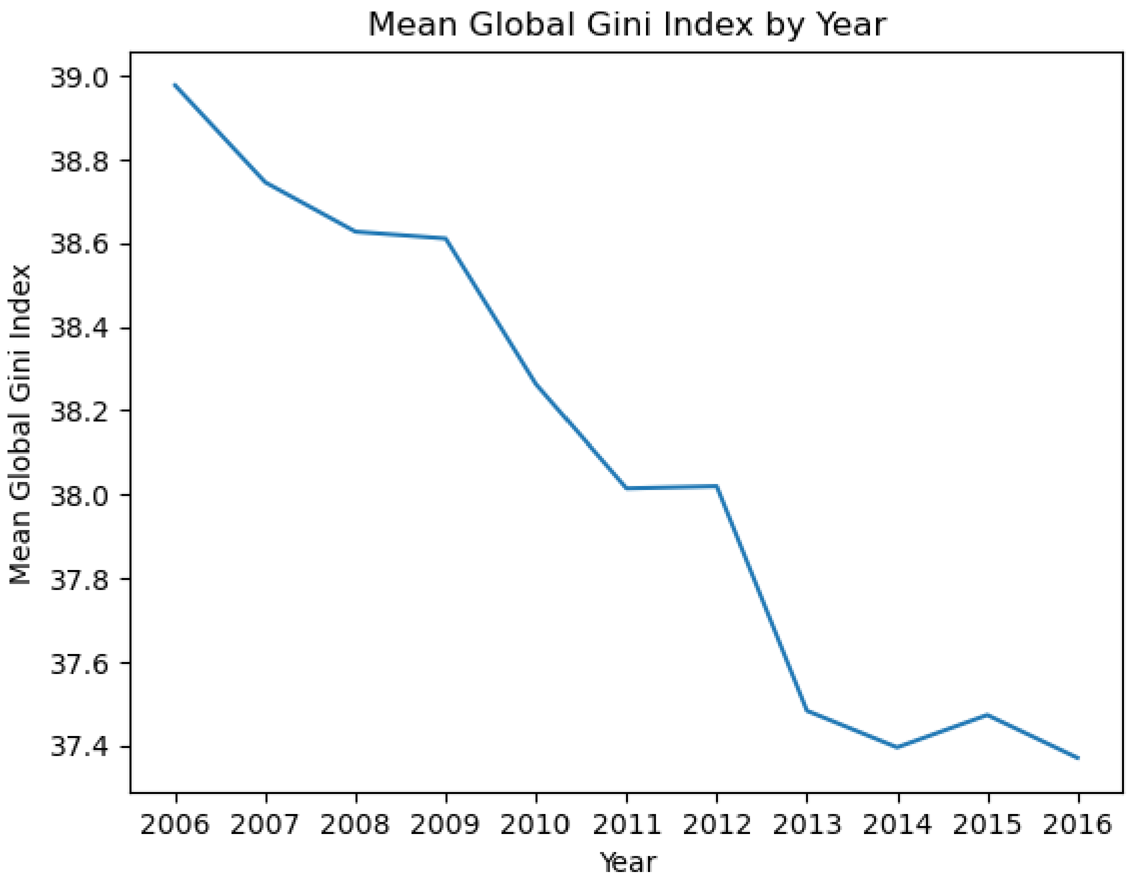
1259 rows × 2 columns

In [ ]:
```
gini_annual_average = gini.groupby('year')['gini_index'].mean()
gini_annual_average
```

Out[ ]:
```
year
2006    38.98
2007    38.75
2008    38.63
2009    38.61
2010    38.26
        ...
2012    38.02
2013    37.48
2014    37.40
2015    37.47
2016    37.37
Name: gini_index, Length: 11, dtype: float64
```

As the plot below shows, the mean global gini index has been going down over the last 10 years, meaning global income inequality is improving.

In [ ]:
```
plt.plot(gini_annual_average.index, gini_annual_average)
plt.title('Mean Global Gini Index by Year')
plt.xlabel('Year')
plt.ylabel('Mean Global Gini Index');
```



Mean Global Gini Index by Continent:

In [ ]:
```
columns = ['year', 'continent', 'gini_index']
gini = combined_final[columns]
gini
```

Out[ ]:

	year	continent	gini_index
0	2006	Africa	42.2
1	2007	Africa	42.1
2	2008	Africa	42.1
3	2009	Africa	42.1
4	2010	Africa	42.1
...	...	...	...
1546	2012	Oceania	33.5
1547	2013	Oceania	34.0
1548	2014	Oceania	34.0
1549	2015	Oceania	34.5
1550	2016	Oceania	34.8

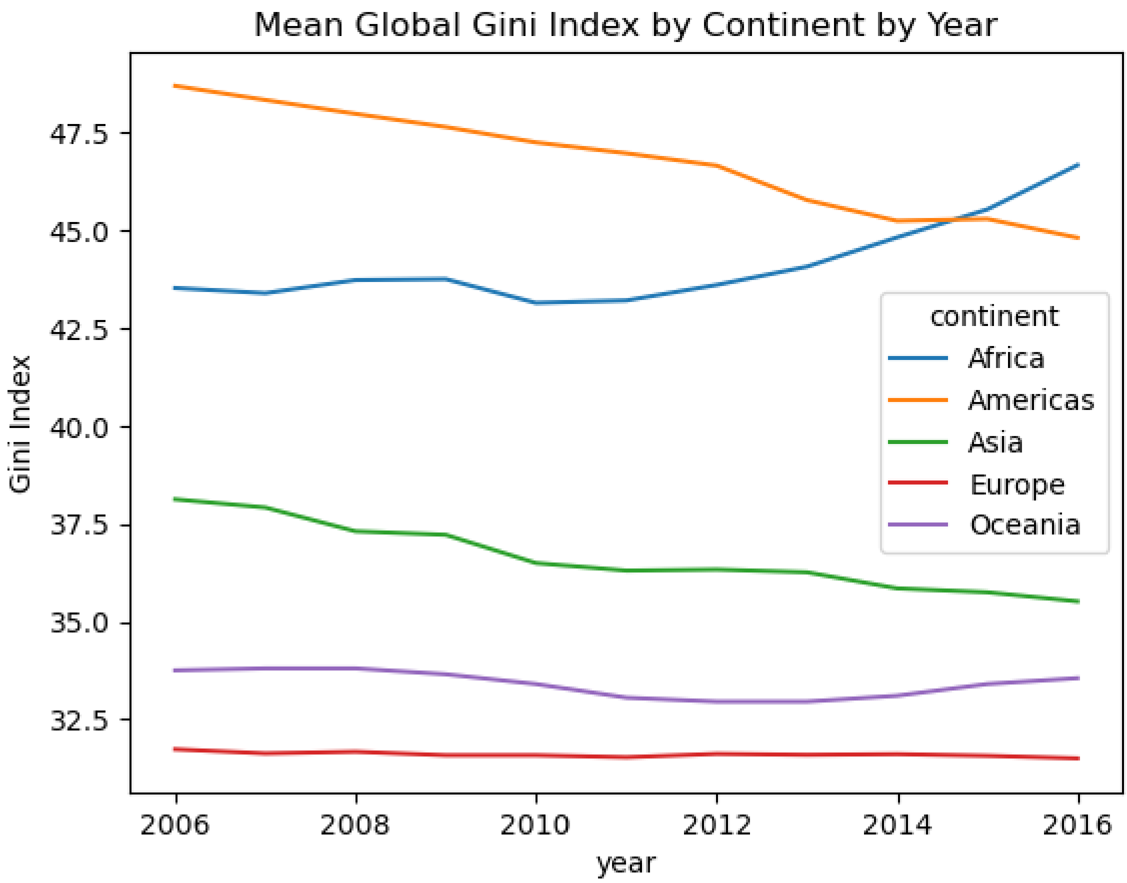
1259 rows × 3 columns

In [ ]:
```
gini_cont_average = gini.groupby(['year', 'continent'])['gini_index'].mean(
gini_cont_average
```

Out[ ]:
```
year  continent
2006  Africa       43.52
      Americas     48.70
      Asia         38.13
      Europe       31.73
      Oceania      33.75
                   ...
2016  Africa       46.67
      Americas     44.81
      Asia         35.52
      Europe       31.50
      Oceania      33.55
Name: gini_index, Length: 55, dtype: float64
```

The chart below reveals that, on a continent basis, all were either declining or mostly flat, except for Africa.

In [ ]:
```
gini_cont_average.unstack(level=1).plot(kind='line', subplots=False, \
                                        title='Mean Global Gini Index by Cor
                                        set_ylabel("Gini Index");
```



In [ ]:
```
columns = ['year', 'continent', 'country', 'gini_index']
gini = combined_final[columns]
gini
```

Out[ ]:

	year	continent	country	gini_index
0	2006	Africa	Congo, Dem. Rep.	42.2
1	2007	Africa	Congo, Dem. Rep.	42.1
2	2008	Africa	Congo, Dem. Rep.	42.1
3	2009	Africa	Congo, Dem. Rep.	42.1
4	2010	Africa	Congo, Dem. Rep.	42.1
...	...	...	...	...
1546	2012	Oceania	New Zealand	33.5
1547	2013	Oceania	New Zealand	34.0
1548	2014	Oceania	New Zealand	34.0
1549	2015	Oceania	New Zealand	34.5
1550	2016	Oceania	New Zealand	34.8

1259 rows × 4 columns

Research Question 2 - What Top 10 Countries Have the Lowest and Highest Income Inequality?

Lowest

Overall, most of the countries with the lowest income inequality are in Europe.

In [ ]:

```
gini.groupby(['country', 'continent'])['gini_index'].mean().to_frame().sort_
```

Out[ ]:

country	continent	gini_index
Slovenia	**Europe**	25.06
Ukraine	**Europe**	25.64
Czech Republic	**Europe**	26.25
Norway	**Europe**	26.75
Slovak Republic	**Europe**	26.79
Denmark	**Europe**	27.16
Kazakhstan	**Asia**	27.44
Finland	**Europe**	27.45
Belarus	**Europe**	27.49
Kyrgyz Republic	**Asia**	27.65

Highest

Overall, most of the countries with the lowest income inequality are in Africa and in Americas.

In [ ]:

```
gini.groupby(['country', 'continent'])['gini_index'].mean().to_frame().sort_
```

Out[ ]:

country	continent	gini_index
South Africa	Africa	63.35
Botswana	Africa	61.09
Namibia	Africa	60.75
Suriname	Americas	60.51
Zambia	Africa	55.76
Central African Republic	Africa	55.70
Bolivia	Americas	54.55
Honduras	Americas	53.94
Lesotho	Africa	53.82
Colombia	Americas	53.16

Research Question 3 - Is a higher tax revenue as a % of GDP associated with less income inequality?

The hypothesis is that countries with higher tax revenue as % of GDP are associated with lower income inequality. The assumption for this is that higher tax revenues are distributed back to lower economic strata in the form of social benefits. Let's see what the data shows.

In [ ]:
```
columns = ['continent', 'country', 'year', 'tax_%_gdp', 'gini_index']
tax = combined_final[columns]
tax
```

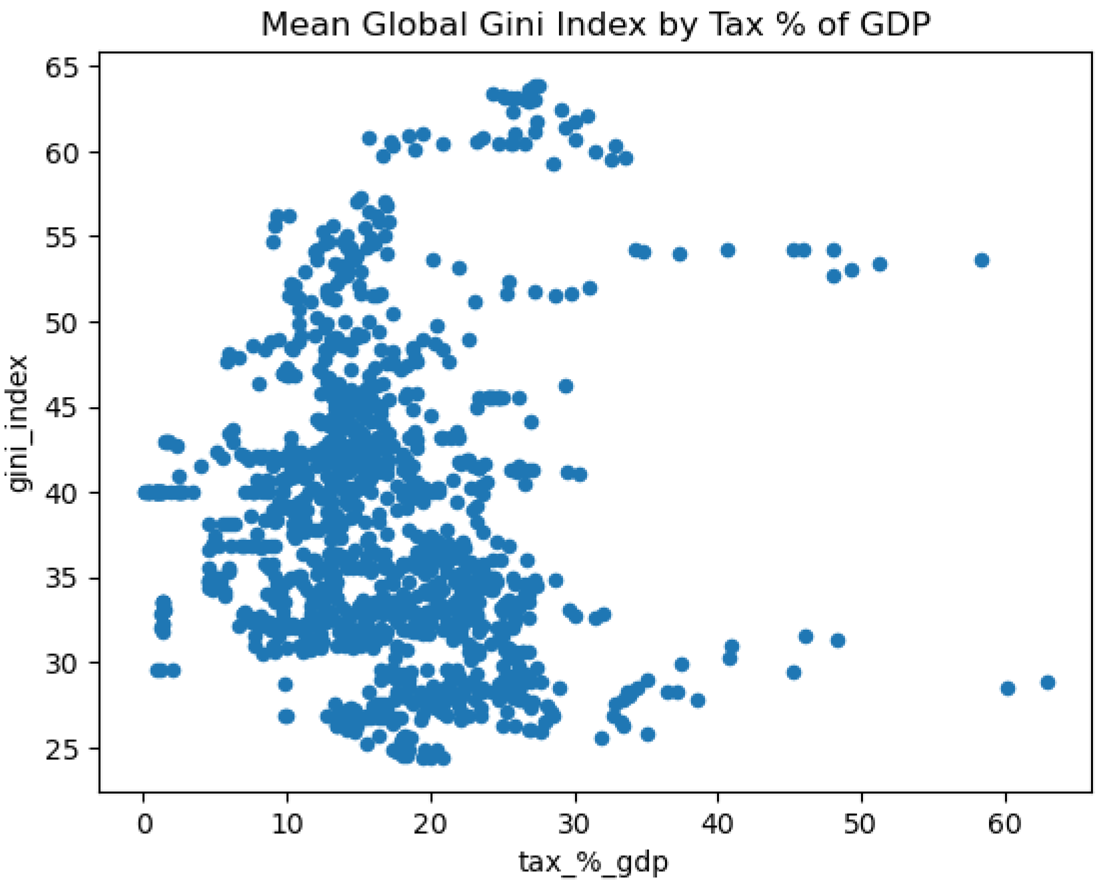
Out[ ]:

	continent	country	year	tax_%_gdp	gini_index
0	Africa	Congo, Dem. Rep.	2006	6.83	42.2
1	Africa	Congo, Dem. Rep.	2007	6.99	42.1
2	Africa	Congo, Dem. Rep.	2008	8.97	42.1
3	Africa	Congo, Dem. Rep.	2009	7.89	42.1
4	Africa	Congo, Dem. Rep.	2010	8.35	42.1
...	...	...	...	...	...
1546	Oceania	New Zealand	2012	26.80	33.5
1547	Oceania	New Zealand	2013	26.60	34.0
1548	Oceania	New Zealand	2014	26.80	34.0
1549	Oceania	New Zealand	2015	27.40	34.5
1550	Oceania	New Zealand	2016	27.20	34.8

1259 rows × 5 columns

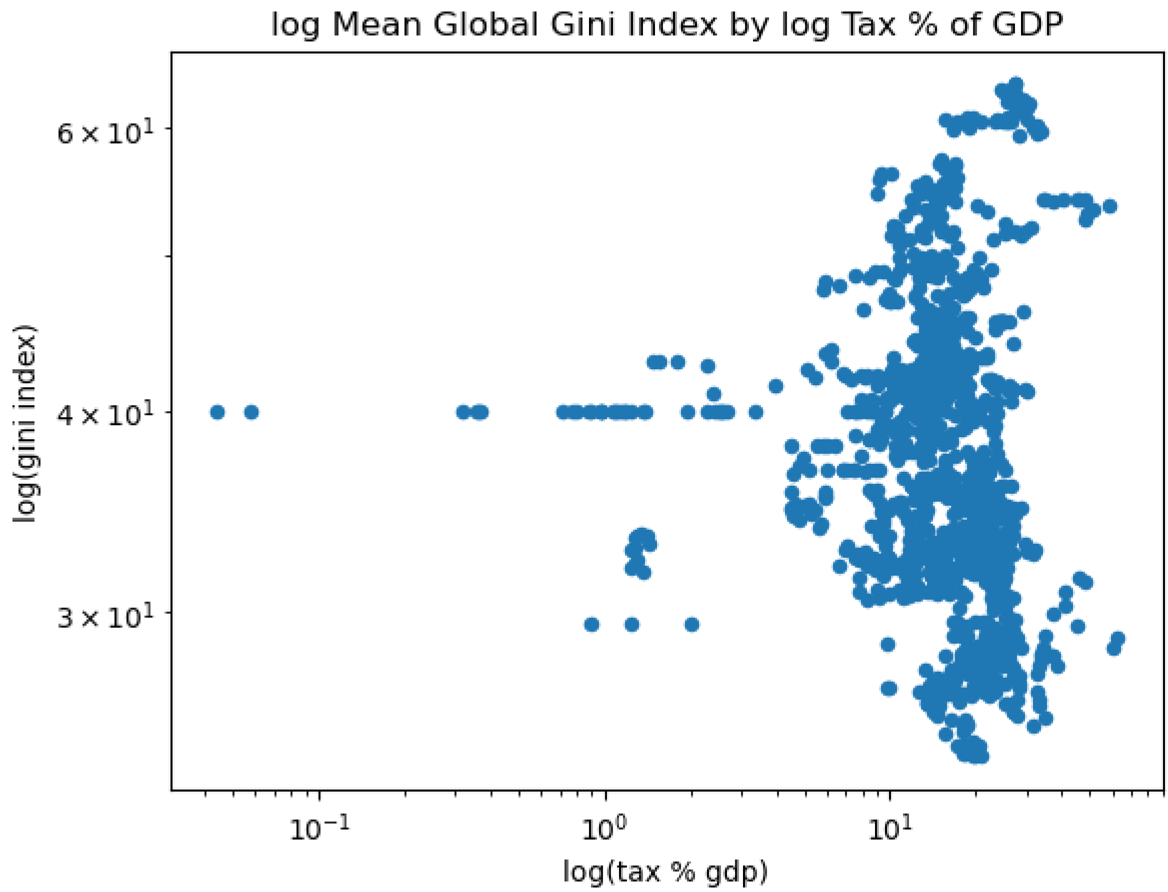
It is difficult to see a trend in the scatter plot below:

In [ ]:
```
tax.plot(x='tax_%_gdp', y='gini_index', kind='scatter', title='Mean Global (
```



Looking at the log of both values reveals that the correlation between the two variables is essentially flat - there is no compelling evidence that higher tax percent of GDP leads to less income inequality.

In [ ]:
```
tax_plot = tax.plot(x='tax_%_gdp', y='gini_index', kind='scatter', loglog=Tr
                    title='log Mean Global Gini Index by log Tax % of GDP')
tax_plot.set_xlabel('log(tax % gdp)')
tax_plot.set_ylabel('log(gini index)');
```



The Pearson correlation is slightly negative at -0.08:

In [ ]:
```
tax_log = np.log(tax['tax_%_gdp']).to_frame()
tax_log['log_gini_index'] = np.log(tax['gini_index'])
tax_log.corr()
```

Out[ ]:

	tax_%_gdp	log_gini_index
tax_%_gdp	1.00	-0.07
log_gini_index	-0.07	1.00

Research Question 4 - Is Higher Income Per Person - GDP Per Capita associated with less income inequality?

The hypothesis is that a higher income per person indicates that more of the country's GDP is being distributed equality among its population.

In [ ]:
```
columns = ['continent', 'country', 'year', 'income_per_person', 'gini_index
income = combined_final[columns]
income
```

Out[ ]:

	continent	country	year	income_per_person	gini_index
0	Africa	Congo, Dem. Rep.	2006	605	42.2
1	Africa	Congo, Dem. Rep.	2007	623	42.1
2	Africa	Congo, Dem. Rep.	2008	640	42.1
3	Africa	Congo, Dem. Rep.	2009	637	42.1
4	Africa	Congo, Dem. Rep.	2010	660	42.1
...	...	...	...	...	...
1546	Oceania	New Zealand	2012	33300	33.5
1547	Oceania	New Zealand	2013	33900	34.0
1548	Oceania	New Zealand	2014	34600	34.0
1549	Oceania	New Zealand	2015	35200	34.5
1550	Oceania	New Zealand	2016	35700	34.8

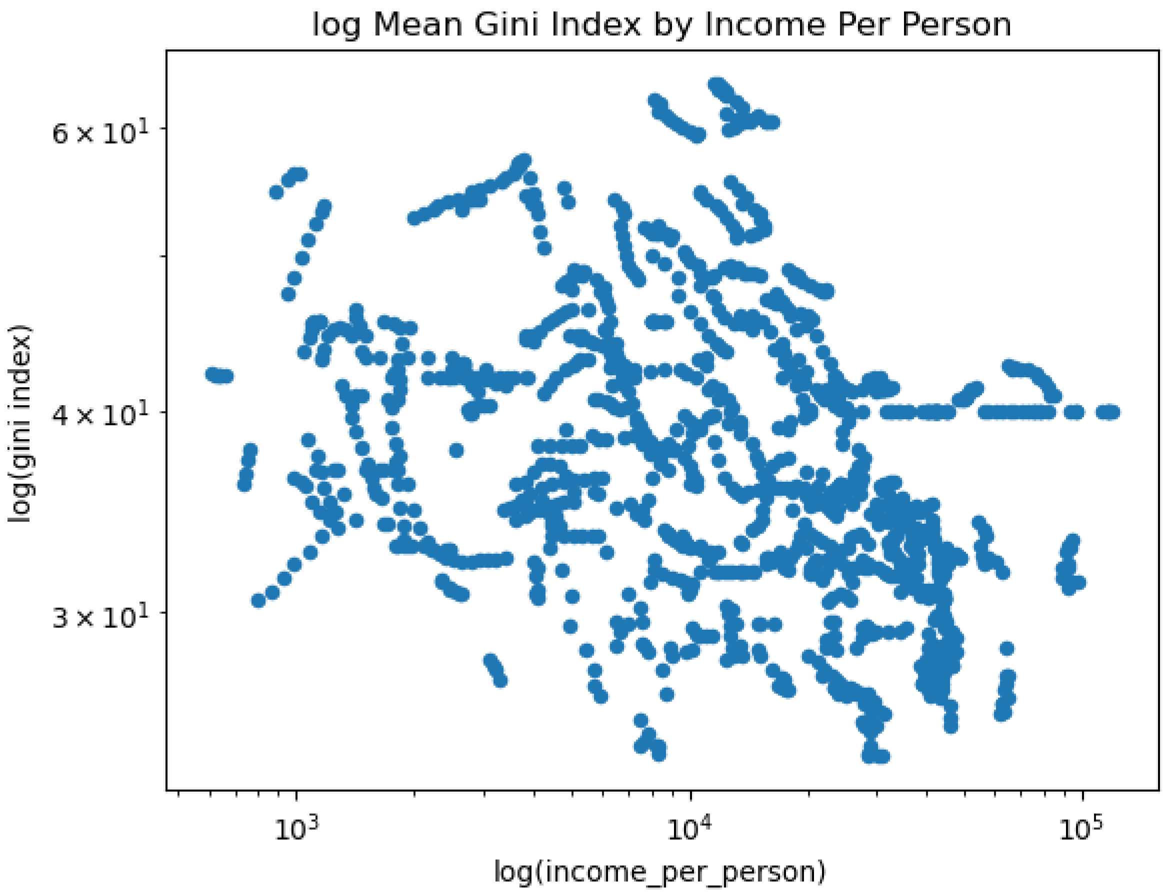
1259 rows × 5 columns

In [ ]:

```
income.plot(x='income_per_person', y='gini_index', kind='scatter', title='Me
```

In [ ]:

```
income_plot = income.plot(x='income_per_person', y='gini_index', kind='scatt
                          title='log Mean Gini Index by Income Per Person')
income_plot.set_xlabel('log(income_per_person)')
income_plot.set_ylabel('log(gini index)');
```



In this case, the Person correlation coefficient is -0.34 indicating that there is weak correlation between log(income_per_person) and the log(gini_index):

In [ ]:
```
income_log = np.log(income['income_per_person']).to_frame()
income_log['log_gini_index'] = np.log(tax['gini_index'])
income_log.corr()
```

Out[ ]:

	income_per_person	log_gini_index
income_per_person	1.00	-0.34
log_gini_index	-0.34	1.00

Research Question 5 - Is Higher Investment as % GDP associated with less income inequality?

The hypothesis is that a higher investment as a percent of GDP indicates that more of the country's GDP is being invested in capital improvements which distributes income benefits across a wide segment of the populcation leading to more equality among its population.

In [ ]:
```
columns = ['continent', 'country', 'year', 'invest_%_gdp', 'gini_index']
invest = combined_final[columns]
invest
```

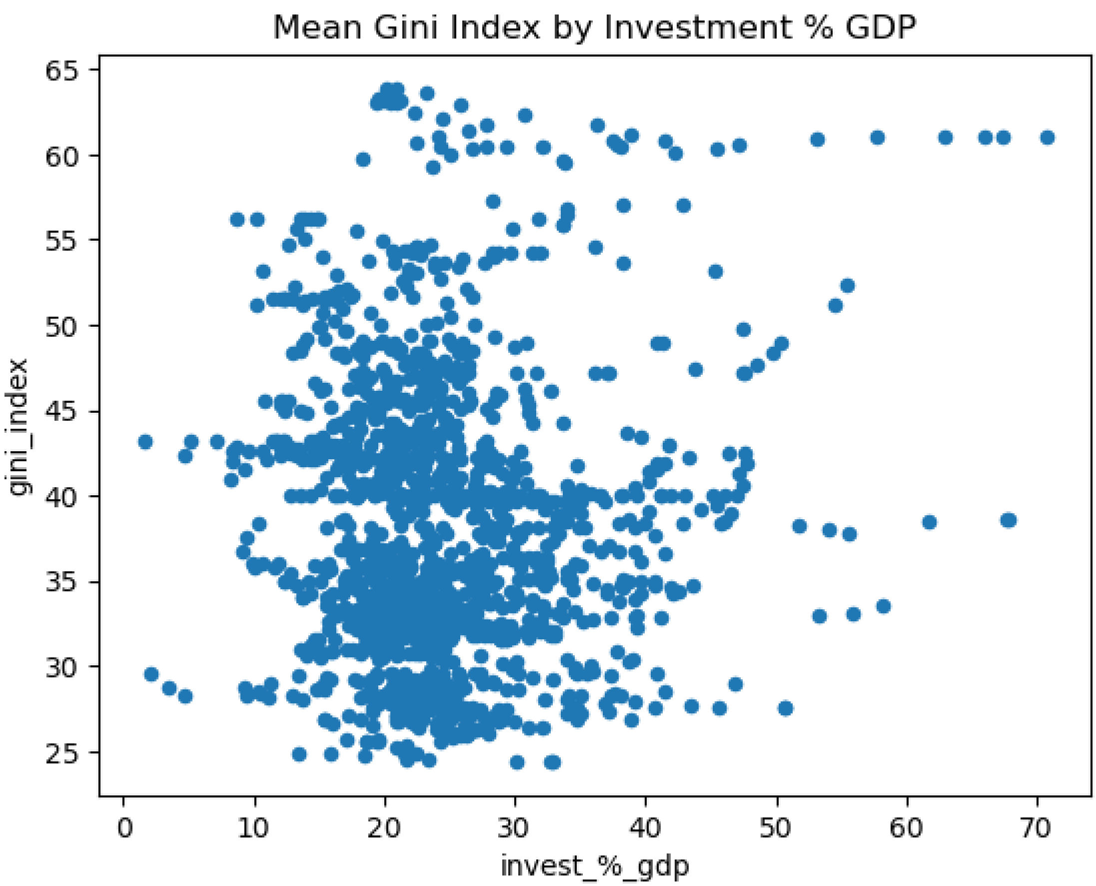
Out[ ]:

	continent	country	year	invest_%_gdp	gini_index
0	Africa	Congo, Dem. Rep.	2006	14.6	42.2
1	Africa	Congo, Dem. Rep.	2007	13.7	42.1
2	Africa	Congo, Dem. Rep.	2008	10.9	42.1
3	Africa	Congo, Dem. Rep.	2009	14.6	42.1
4	Africa	Congo, Dem. Rep.	2010	28.8	42.1
...	...	...	...	...	...
1524	Oceania	New Zealand	2012	20.9	33.5
1525	Oceania	New Zealand	2013	22.0	34.0
1526	Oceania	New Zealand	2014	22.9	34.0
1527	Oceania	New Zealand	2015	23.4	34.5
1528	Oceania	New Zealand	2016	24.4	34.8

1529 rows × 5 columns

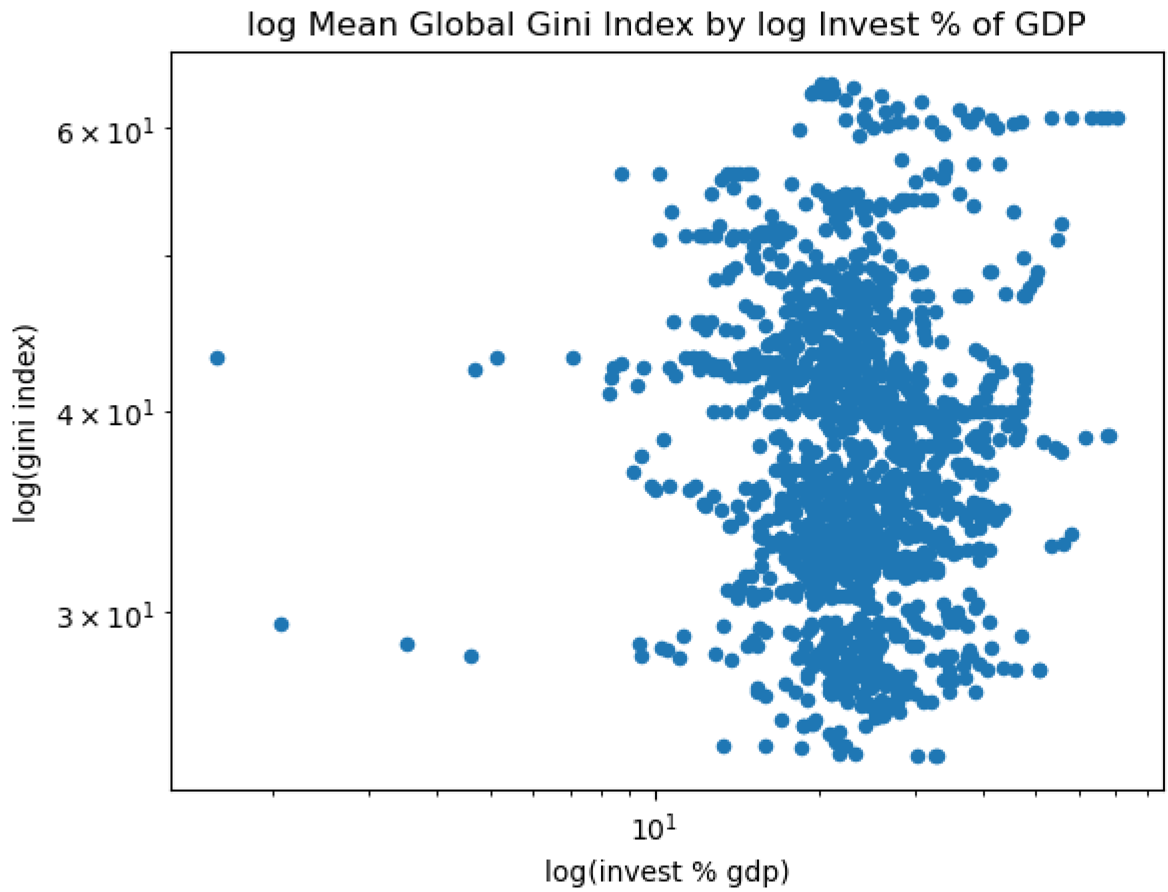
In [ ]:

```
invest = invest[invest['invest_%_gdp'] > 0]
invest.plot(x='invest_%_gdp', y='gini_index', kind='scatter', title='Mean Gi
```



In [ ]:

```
invest_plot = invest.plot(x='invest_%_gdp', y='gini_index', kind='scatter',
                          title='log Mean Global Gini Index by log Invest % of GDI
invest_plot.set_xlabel('log(invest % gdp)')
invest_plot.set_ylabel('log(gini index)');
```



The Pearson corr coefficient of -0.03 indicates no correlation between these two variables.

In [ ]:
```
invest_log = np.log(invest['invest_%_gdp']).to_frame()
invest_log['log_gini_index'] = np.log(tax['gini_index'])
invest_log.corr()
```

Out[ ]:

	invest_%_gdp	log_gini_index
invest_%_gdp	1.00	-0.07
log_gini_index	-0.07	1.00

Research Question 6 - Is Higher EIU Democracy Index associated with less income inequality?

The hypothesis is that countries with higher EIU Democracy Index address the needs of a broader segment of the popluation leading to less income inequality.

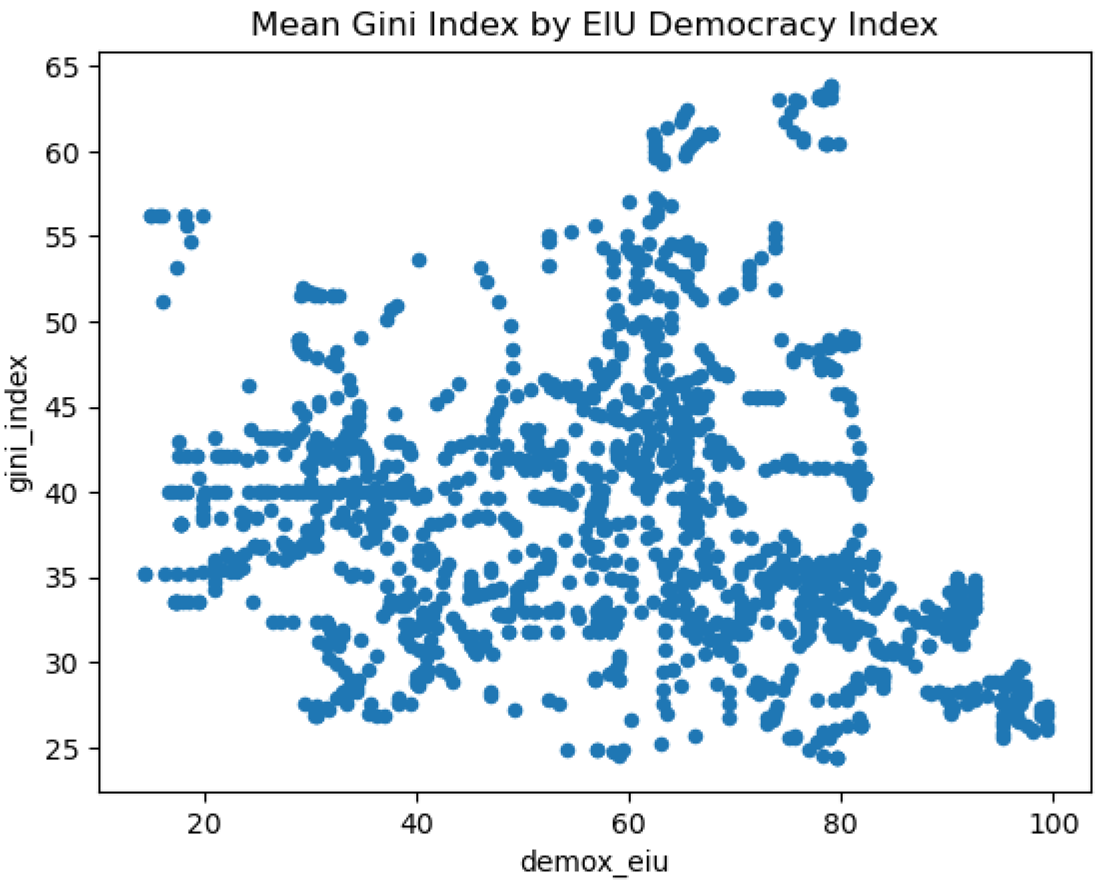
In [ ]:
```
columns = ['continent', 'country', 'year', 'demox_eiu', 'gini_index']
demo = combined_final[columns]
demo
```

Out[ ]:

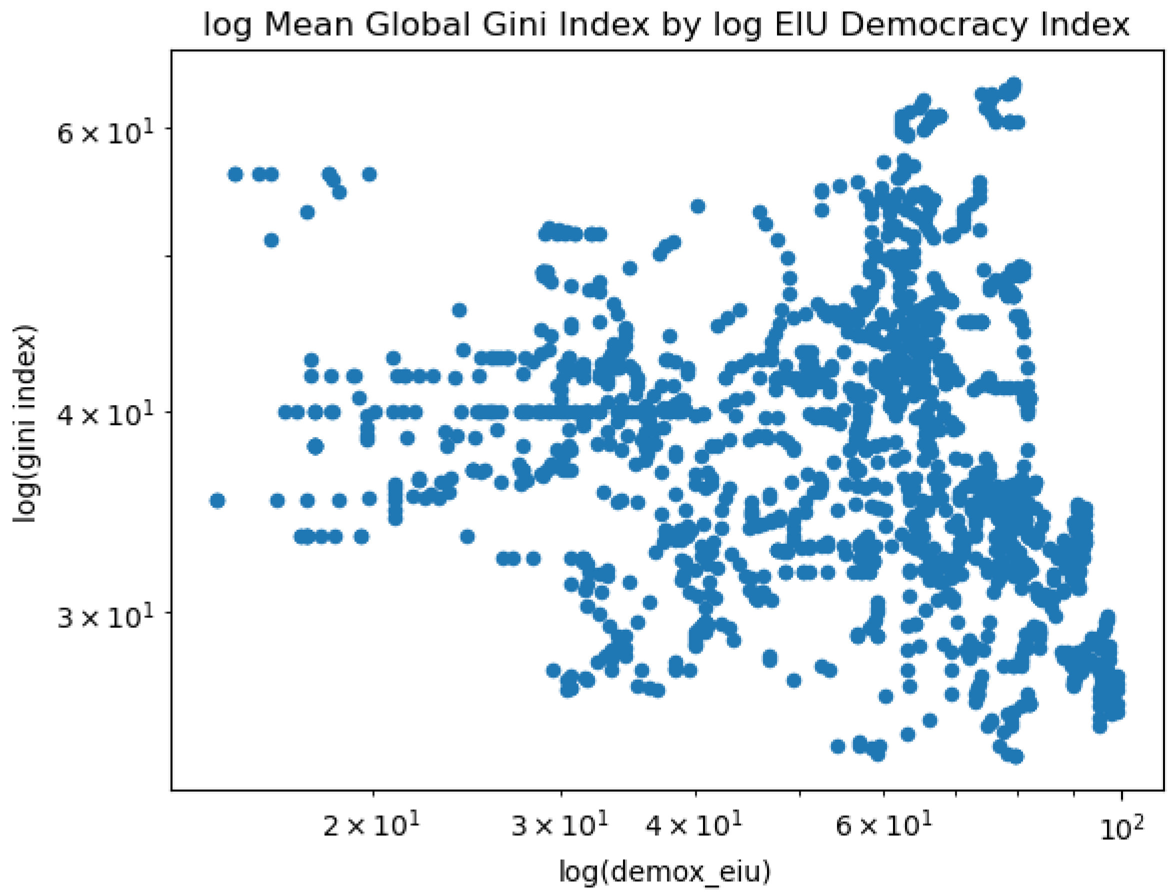
	continent	country	year	demox_eiu	gini_index
0	Africa	Congo, Dem. Rep.	2006	27.6	42.2
1	Africa	Congo, Dem. Rep.	2007	25.2	42.1
2	Africa	Congo, Dem. Rep.	2008	22.8	42.1
3	Africa	Congo, Dem. Rep.	2009	22.1	42.1
4	Africa	Congo, Dem. Rep.	2010	21.5	42.1
...	...	...	...	...	...
1524	Oceania	New Zealand	2012	92.6	33.5
1525	Oceania	New Zealand	2013	92.6	34.0
1526	Oceania	New Zealand	2014	92.6	34.0
1527	Oceania	New Zealand	2015	92.6	34.5
1528	Oceania	New Zealand	2016	92.6	34.8

1529 rows × 5 columns

In [ ]:
```python
demo.plot(x='demox_eiu', y='gini_index', kind='scatter', title='Mean Gini In
```



In [ ]:
```python
demo_plot = demo.plot(x='demox_eiu', y='gini_index', kind='scatter', loglog=
                      title='log Mean Global Gini Index by log EIU Democracy I
demo_plot.set_xlabel('log(demox_eiu)')
demo_plot.set_ylabel('log(gini index)');
```



In this case, the Person correlation coefficient is -0.2 indicating that there is weak correlation between log(demox_eiu) and the log(gini_index):

In [ ]:
```
demo_log = np.log(demo['demox_eiu']).to_frame()
demo_log['log_gini_index'] = np.log(tax['gini_index'])
demo_log.corr()
```

Out[ ]:

	demox_eiu	log_gini_index
demox_eiu	1.00	-0.19
log_gini_index	-0.19	1.00

Conclusions

The following are the conclusions from this analysis:

Research Question 1 - Is Income Inequality Getting Worse or Better in the Last 10 Years?

Answer:

Yes, it is getting better, improving from 38.7 to 37.3

On a continent basis, all were either declining or mostly flat, except for Africa.

Research Question 2 - What Top 10 Countries Have the Lowest and Highest Income Inequality?

Answer:

Lowest: Slovenia, Ukraine, Czech Republic, Norway, Slovak Republic, Denmark, Kazakhstan, Finland, Belarus,Kyrgyz Republic

Highest: Colombia, Lesotho, Honduras, Bolivia, Central African Republic, Zambia, Suriname, Namibia, Botswana, South Africa

Research Question 3 Is a higher tax revenue as a % of GDP associated with less income inequality?

Answer: No

Research Question 4 - Is Higher Income Per Person - GDP Per Capita associated with less income inequality?

Answer: No, but weak negative correlation.

Research Question 5 - Is Higher Investment as % GDP associated with less income inequality?

Answer: No

Research Question 6 - Is Higher EIU Democracy Index associated with less income inequality?

Answer: No, but weak negative correlation.

The above results suggest that there are other drivers for the overall reduction in income inequality. Futher analysis of additional factors should be undertaken.