Analysis of World Wide Income Inequality

Data Science 1 Project Report
Winter 2021

Group 19: Fuad, S. ; Muyoba, M. ; Page, H. ; Syed, A.; Valavan, K.; Waas, T.

Introduction

As the world becomes more global and the distribution of wealth continues to diminish, it is important to understand how the inequality of income has changed over time and space. We chose to use the World Income Inequality Database (WIID) to analyze the trends and patterns to determine whether income inequality exists in developing and developed nations alike despite the global economic growth and that the gap is diverging.

In order to validate our hypothesis, that income inequality does exist around the world, we decided to take a three pronged approach to investigate global wealth distribution over time, analyze income inequality over geographic regions as well as compare and contrast the top and bottom wealthiest populations within each country over time.

The basis of our analysis revolves around Gini coefficients calculated for each country in a specified year. The higher the value, the larger the inequality of income or wealth is and the lower the value, the lesser the inequality of income or wealth exists within that country. This can be calculated based on a few different resource types including consumption, net income and net/gross income.

In addition to the Gini coefficients, the group analyzed the share of per capita net and gross income among population among various geographic regions and subregions to demonstrate this inequality. This analysis highlights that the wealth/income controlled by the top 10% of the population in any given region is higher than the bottom 40% of the population in all areas and even greater in developing regions.

Objective

The primary objective of this analysis is to explore the World Wide Income Inequality Database (WIID) to demonstrate the income inequality in various geographic regions and subregions through the following:

1. Comparing and contrasting income of top x% with bottom y% population in various nations across the world.
2. Discovering any changes in global wealth distribution over time and its correlation with income inequality
3. Analyzing income inequality by geographical region

Data Preparation

This data used in this study is collected by the United Nations University World Institute of Development and Economic Research (UNU-WIDER). The UNU-WIDER World Income Inequality Database(WIID) provides information on income inequality for 189 developed, developing, and transition countries (including historical entities) in an organized and user-friendly manner. The data source is available for download as an excel file from the UNU-WIDER website. This version of the dataset (WWID4) is the most comprehensive income inequality database in the world and contains more than 11,000 observations for 189 countries ranging from 1867 to 2017.

WIID combines information coming from many sources, including historical compilations with updated information from the most salient data repositories, as well as from national statistical offices, and independent research papers. This information is captured in the source feature of the database. To eliminate duplicates and low confidence data, the group ranked various sources based on the quality attribute and the prevalence of data collected. Eliminating duplicates through this approach ensures that we have the highest quality available for each observed country in any given year. The available data sources were ranked as below:

1. Luxembourg Income Study (LIS)
2. World Bank
3. United Nations
4. Research study
5. National statistical authority
6. Organisation for Economic Co-operation and Development (OECD)
7. Socio-Economic Database for Latin America and the Caribbean (SEDLAC)
8. Eurostat
9. Other international organizations

WIID4 dataset is accompanied by a user guide that helps understand the context and definition of all data attributes. In addition it provides suggestions on data interpretation e.g. using income as an indicator of inequality rather than consumption. While there are several opinions supporting both aspects, income based analysis is favoured among the research community.

In order to obtain consistent results, the team decided to filter out some of the records based on the filters below:

1. Sharing unit : ‘Household’ was used as the statistical/income sharing unit
2. Reference unit: ‘Person’ was used to ensure the data used is weighted at individual level. This is a critical element in using the equivalence scale when evaluating income share allocations.
3. Population coverage: ‘All’ was selected to equally represent all population regardless of their respective economic activities
4. Area coverage: ‘All’ was selected to include all of the areas (urban and rural) as indicated in the specific report

The WIID comprises 11,826 observations. The following table summarizes the number of observations for different time periods:

Time span	Number of observations
Total observations	11,826
Before 1960	311
1960–69	689
1970–79	849
1980–89	1,440
1990–99	2,630
2000–09	3,219
2010–19	2,688

(source: WIID4 User Guide.pdf)

Given the low representation in earlier years the focus of this analysis was limited to the data available since1980's. Each study objective (use case) has had specific needs for data elements and further data filtering and aggregation approaches have been used in preparation of this report. However, the consensus has been to use the UN region divisions rather than the world bank to maintain the report consistency.

While analysing the data, one of our subgroup noted the following observations pertaining to income inequality by geographical region:

1. There are records where gini reported and resource columns are not reported. Since the basis of the analysis is gini coefficient as a representation of inequality of wealth, these rows were eliminated from the analysis.
2. By reviewing the data dictionary and inspecting the data using df.info() and df.head(), we observed that resource_un & resource_un_sub provides the categorization of countries belonging to a region and sub-region. Hence these columns were used to categorize countries into regions/sub-regions.
3. By plotting a cross tab of resource vs country, we observed the count of records available for each country per year. This helped to scope the years used in the analysis by geographical region.
4. The crosstab of resource vs country also showed that most countries report gini coefficient for resource = Consumption or Income. Very few countries reported gini coefficient for both resources.
5. By plotting a cross tab of resource vs year, we observed that the number of records for data available for resource = Earnings and Income (gross) is low. Hence these resource types were considered as non representative of the population and eliminated from the analysis by one particular subgroup. The first use case described below uses both gross and net income to analyze the income gaps and these are profoundly different for these two resource categories.

Analysis

Use Case 1: Compare and contrast income of top x% with bottom y% population in various nations across the world

Gini index, due to its mathematical properties, downplays the role of top and bottom indicators. Therefore, analysing the income distribution among the richest 10% and the poorest 40% allows us to inspect information at the lower level to investigate the trends. This part of the analysis compares the net income (earnings received after tax deductions) distribution and gross income (income before deductions) distribution split between 10% of the top earning population and 40% of the lowest earning population in the regions of Oceania, Europe, Asia, Africa and the Americas between the years of 1990 and 2016. The rationale to opt for 10% and 40% was merely based on the ease of demonstration rather than any other factor. We would have liked to use the pareto distribution (80-20) but the data observed in various regions would not have been very comprehensible.

Between all the regions, Africa displays the highest level of income inequality. The net income gap grew steadily between 2002 and 2010 reaching the largest split where less than 5% of the income belonged to the bottom 40% and close to 65% belonged to the top 10% of the population. However between 2010 and 2012 the income split for the top 10% population drastically decreased to just below 50% but is still the largest gap for all the regions.

Europe is the only region which had time periods where 40% of the bottom earning population were capturing more net income than the top 10%. In Europe between 1990 and 1995 the percent share of net income for the bottom 40% population started at 18% surpassed the top 10% in 1993 reaching close to 25% and then back down to 21% in 1995. Onwards from 1995 to 2015 the top 10% maintain a higher share of income however the gap remains between 26% and 19%. The data for gross income also shows similar trend where between 1990 and 1995 the gap lessens for a short period of time and then increases again and maintains a gap onwards between 28% and 19% except in 2001 when the top 10% share of gross income shot up to 35% and the bottom 40% was near to 15%. When looking further into the dataset by dividing it into subregions of Northern, Western, Southern and Eastern Europe, Eastern Europe prevailed as having the least income inequality and Western Europe's inequality significantly started to rise after 1997.

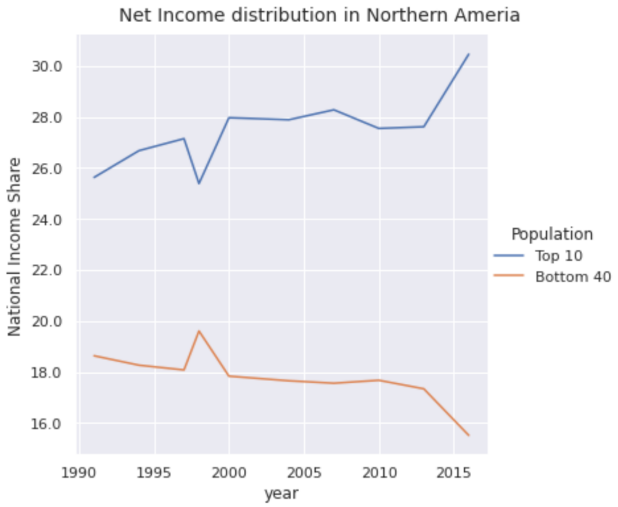
Oceania shows the least volatility, from 1996 to 2010 the bottom 40% 's share of net income remains between 18% to 20%. The top 10% population stays steady at capturing 25% of net income from1996 to 2003 until it increases to 27% in 2008 creating the largest gap of 8%. The gross income of Oceania has the largest gap in 1996 of 14% and the smallest gap in 1995 of 11%.

Asia shows strong fluctuations between 1990 and 2000 for both net and gross income ranging from 45% for the top 10% population and near 5% for the bottom 40% population. The large gaps start to diminish after 2000 and the trend shows that income inequality starts to decline as time progresses. After 2000 net income for the top 10% does not exceed 32% and the bottom 40% does not go below 15%. For gross income the most recent data shows a gap of less than 5%.

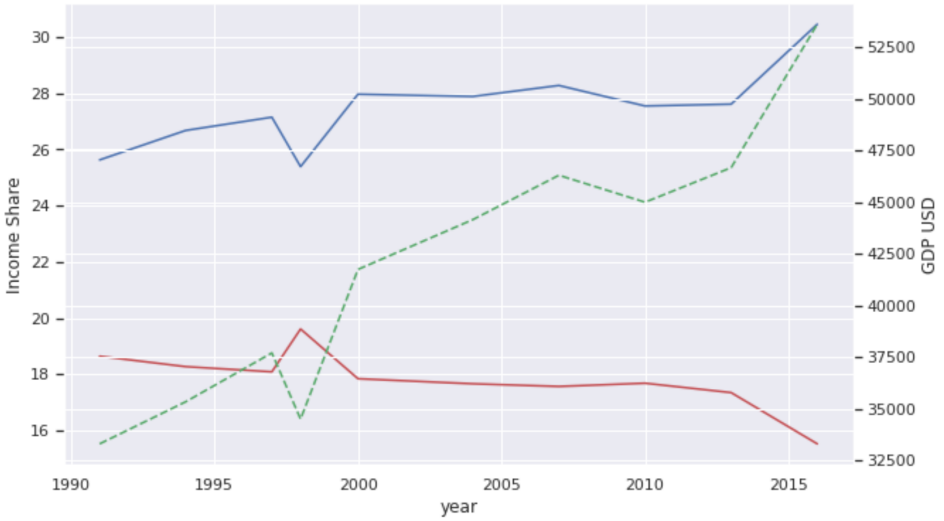
The net income gap for the Americas has stayed steady with the top 10% population maintaining a share between 37% and 45% until 2015 when it drops to its lowest point of 29%. The bottom 40% population maintains between 10% and 15% from 1995 to 2015 and goes slightly above 15% in 2016. For gross income the Americas show strong fluctuations for the top 10% from 1990 up until 2006, with the largest portion of gross income being almost

50% and the lowest point at 38%. After 2006 there is less volatility and a gradual decrease in the gross income inequality gap. When dividing into subregions of Northern America, South America, Central America and Caribbean, the top 10% population for South America, Central America and Carribean captured between 35%-47% of income between 1995 and 2013 wheres the bottom 40% had between 8%-15% during that time frame and in general showed a slight decreasing trend in income inequality. Whereas North America's inequality gap was smaller with the top 10% ranging from 25%-31% and the bottom 40% ranging between 15%-20% however between 2013-2015, North America's income inequality gap starts widening with the top 10% having income distribution increase from 28% to above 30% and the bottom 40% having theirs decrease from 18% to 15%.

Based on these comparisons, income inequality is evident throughout the world in both developed and developing countries. All regions have unbalanced distributions of income with the bottom 40% receiving close to 5% of income share at some points (relevant figures are included in appendix and more in the actual python notebook). However the world wide income trend especially after 2013 is that income distribution for the top 10% is decreasing except for North America where the top 10% population is increasing their income and the bottom 40% is decreasing as shown below.



Further review of GDP confirmed that the income share of the richest 10% of the population is closely related to economic growth, which has indeed contributed to the income disparity.



The green dashed line represents GDP whereas blue and red represent income distribution of the richest 10% and poorest 40% respectively

Use Case 2: Discover any changes in global wealth distribution over time and its correlation with income inequality

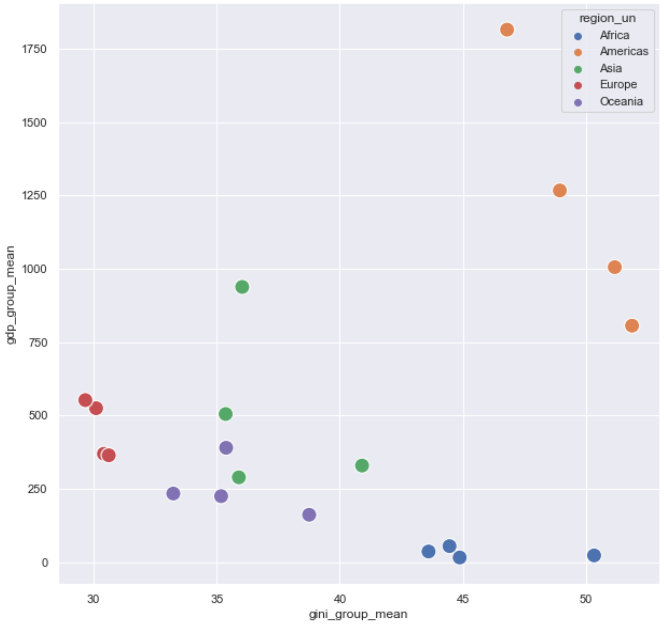
Further Data preparation for time series analysis

The main challenge of the analysis was that the temporal distribution of many countries is not uniform. (Each country does not have a Gini reported for every year). Therefore, a further analysis had to be carried forward to prepare data for the time series analysis.There were 177 countries reported in the dataset. However, when analyzing data points over years, each year had a different count of data points. Therefore, the analysis was done by choosing years which had at least 50 or more countries in the dataset. (from 1995 – 2016). Even after choosing the year, each country does not have data points available for each year. Therefore, for our overall analysis, we wanted at least 100 countries in each bin. Therefore, bins were created in the dataset so that each country will have at least one data point within the bin. 5 year bins were created and each bin represented more than 100 countries. The analysis dropped all missing values since one data point per bin from each country would be sufficient. For countries which had multiple data points, the median was taken as the boxplot showed few outliers when taking the mean. A grouped median for each bin of the Gini and GDP was calculated and merged to the dataframe as a new column. After grouping, an overall Gini and GDP was calculated by grouping according the bins and taking the mean of it. Any missing values were dropped as some we would not be expecting all countries to be represented in each bin.

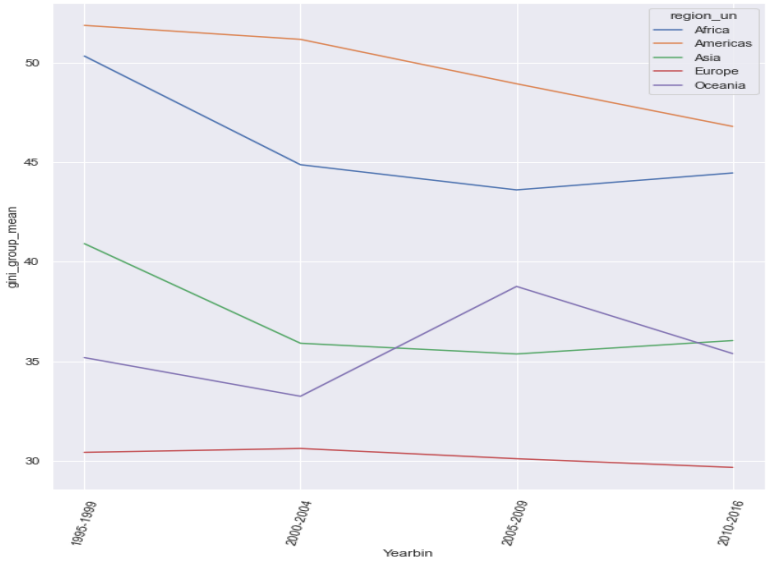
Analysis on by region (Defined by UN) wealth distribution over time and its correlation with income inequality

When analysing gini spread regionally and its correlation to GDP, all regions showed a negative correlation. Americas recorded the highest Gini rates. However, with economic growth Gini has been improving i.e. trending down. America's Gini coefficient showed a strong negative correlation, indicating that in the Americas region the income inequality has increased with economic development and will continue on this trend. . Europe shows the second best correlation statistics in reducing income equality with economic growth. Africa, Ocenia and Asia highlights a slow pace in improving income inequality with economic growth.

Region	Correlation
Americas	-0.986
Europe	-0.929
Africa	-0.422
Oceania	-0.391
Asia	-0.377



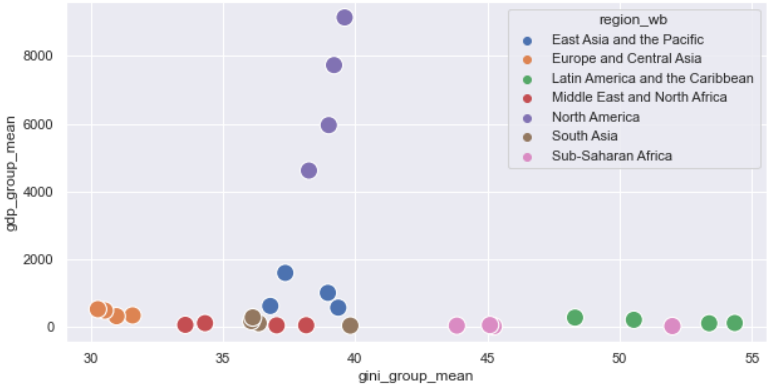
However, when comparing the Gini's regionally, Europe has been maintaining a low gini which indicates since 1995 the income distribution in the region has been stable. Even though with



economic growth Americas shows a strong correlation, still Americas region contributes to the highest income inequality in the globe followed up with Africa. Asia and Africa shows a momentum towards a worsening Gini since 2005-2010 to 2010-2016.

Further to the UN regional analysis, a further analysis was carried out on the regional classification according to the world bank classifications. The world bank classifications further breaks regions to give clarity on the income distribution.

Region	Correlation
Latin America and the Caribbean	-0.983
Europe and Central Asia	-0.859
South Asia	-0.758
Middle East and North Africa	-0.617
Sub-Saharan Africa	-0.392
East Asia and the Pacific	-0.271
North America	0.957

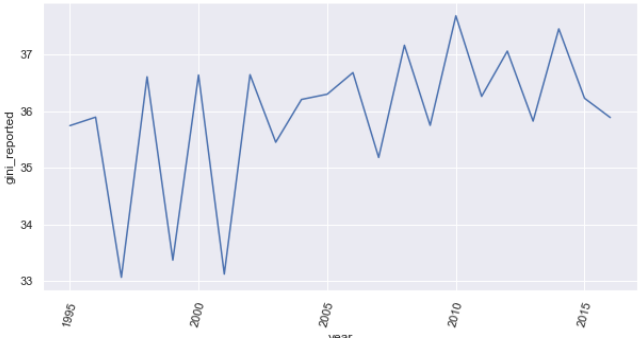


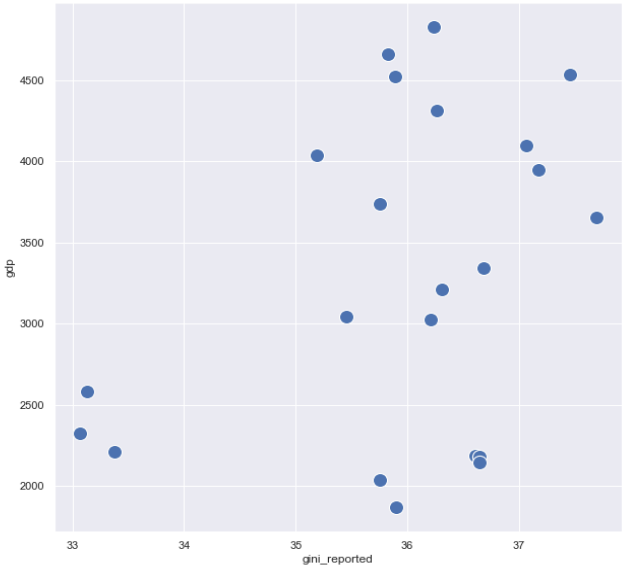
According to the world bank regional divisions, Europe and Central Asia records and best Gini and further will improve with the growth in GDP. North America on the other hand shows evidence of a worsening Gini with GDP growth.

Analysis on the top 10 economies in the world in 2016 with inequality data

Since the temporal distribution of the countries restricted the analysis to be grouped in bins, the further analysis investigated how major economies are performing as a whole on income

inequality. The database did not provide sufficient information about the top 10 major economies in the world. Therefore, with data availability the analysis was carried out in finding the top 10 economies in 2016 which had income inequality data. The United States, China, Germany, UK, France, Italy, Canada, Spain, Mexico and Indonesia were selected. India and Japan, which are major economies in the world, did not have sufficient income data for analysis. The top 10 economies selected had sufficient data for a year by year analysis since 1995 and at least there was data available in 7 countries out of 10 each year.





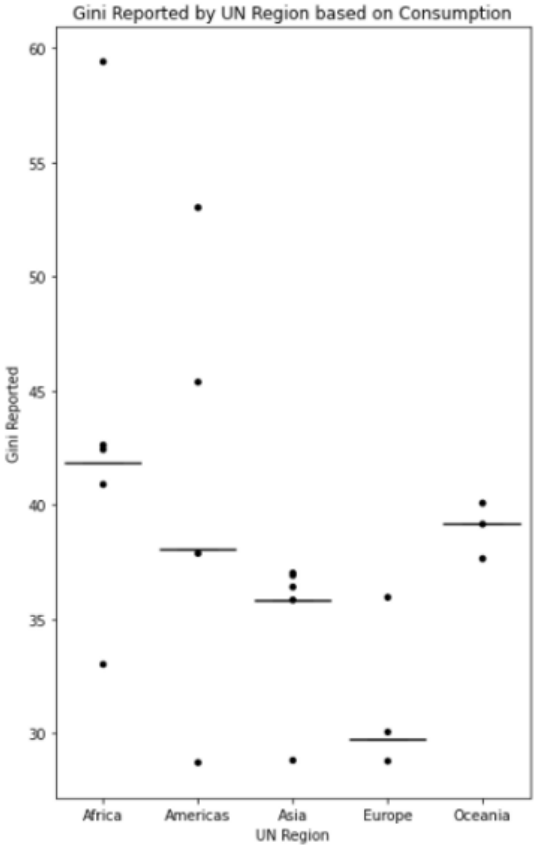
The countries were grouped and analysed according to the Gini performance and its correlation with GDP. There was significant economic growth in the top 10 economies over time. When comparing the overall Gini reported as a group, the Gini reported an upward momentum since 1995. The analysis is further supported by the scatter plot. The top 10 economies show a weak positive correlation of 0.38 which indicates the income distribution worsening with GDP growth.

Use Case 3: Analyze income inequality by geographical region

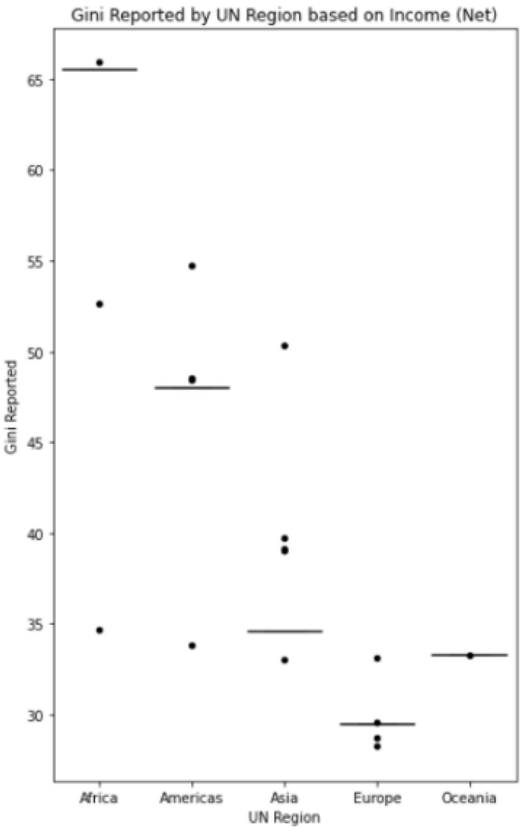
In examining the Gini coefficients to represent income inequality, we also took the approach of analyzing the regions and subregions of each geographic area to determine how they compare to the Gini coefficient of each other and the geographic region respectively. We made the decision to examine sub regions as opposed to individual countries as we noticed most of the analysis available for income inequality stopped at the country level.

As mentioned previously, the decision was made to keep analysis of Gini coefficients based on consumption, income (net) and income (net/gross) separate, therefore all analysis based on geographic region was conducted for each resource type. This ensured that the value calculated for the gini coefficients were based on similar methods of determining wealth.

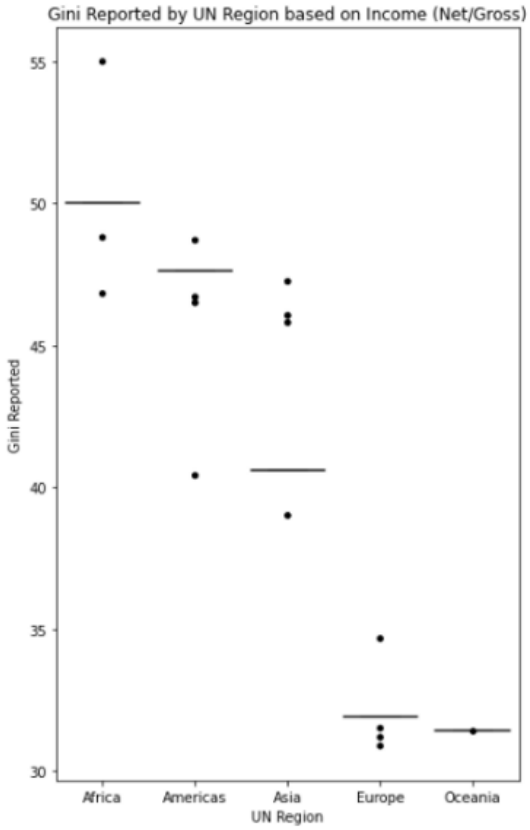
The following three graphs depict the median gini coefficients for each UN sub region within each UN region (points) and the median gini coefficient for each UN region (horizontal line). This allowed us to visually analyze how the UN sub regions compared to each other and the median regional gini coefficient. (analysis follows the charts on the next page)



region_un_sub	region_un	gini_reported	gini_reported_region	diff_from_region
Eastern Africa	Africa	42.600	41.785	0.815
Middle Africa	Africa	42.410	41.785	0.625
Northern Africa	Africa	33.000	41.785	-8.785
Southern Africa	Africa	59.385	41.785	17.600
Western Africa	Africa	40.880	41.785	-0.905
Caribbean	Americas	37.860	38.030	-0.170
Central America	Americas	45.360	38.030	7.330
Northern America	Americas	28.700	38.030	-9.330
South America	Americas	53.000	38.030	14.970
Central Asia	Asia	28.800	35.780	-6.980
Eastern Asia	Asia	35.820	35.780	0.040
South-eastern Asia	Asia	37.000	35.780	1.220
Southern Asia	Asia	36.390	35.780	0.610
Western Asia	Asia	36.900	35.780	1.120
Eastern Europe	Europe	28.765	29.730	-0.965
Northern Europe	Europe	35.935	29.730	6.205
Southern Europe	Europe	30.040	29.730	0.310
Melanesia	Oceania	37.630	39.140	-1.510
Micronesia	Oceania	40.060	39.140	0.920
Polynesia	Oceania	39.140	39.140	0.000



region_un_sub	region_un	gini_reported	gini_reported_region	diff_from_region
Eastern Africa	Africa	34.65	65.55	-30.90
Northern Africa	Africa	52.60	65.55	-12.95
Southern Africa	Africa	65.90	65.55	0.35
Caribbean	Americas	54.70	48.00	6.70
Central America	Americas	48.40	48.00	0.40
Northern America	Americas	33.80	48.00	-14.20
South America	Americas	48.50	48.00	0.50
Central Asia	Asia	39.70	34.60	5.10
Eastern Asia	Asia	33.00	34.60	-1.60
South-eastern Asia	Asia	39.00	34.60	4.40
Southern Asia	Asia	50.30	34.60	15.70
Western Asia	Asia	39.10	34.60	4.50
Eastern Europe	Europe	28.25	29.50	-1.25
Northern Europe	Europe	29.55	29.50	0.05
Southern Europe	Europe	33.10	29.50	3.60
Western Europe	Europe	28.70	29.50	-0.80
Australia and New Zealand	Oceania	33.30	33.30	0.00



region_un_sub	region_un	gini_reported	gini_reported_region	diff_from_region
Eastern Africa	Africa	46.820	50.00	-3.180
Middle Africa	Africa	55.000	50.00	5.000
Western Africa	Africa	48.800	50.00	-1.200
Caribbean	Americas	46.500	47.61	-1.110
Central America	Americas	48.700	47.61	1.090
Northern America	Americas	40.410	47.61	-7.200
South America	Americas	46.700	47.61	-0.910
Eastern Asia	Asia	47.250	40.60	6.650
South-eastern Asia	Asia	46.050	40.60	5.450
Southern Asia	Asia	45.800	40.60	5.200
Western Asia	Asia	39.000	40.60	-1.600
Eastern Europe	Europe	31.500	31.92	-0.420
Northern Europe	Europe	31.180	31.92	-0.740
Southern Europe	Europe	34.660	31.92	2.740
Western Europe	Europe	30.875	31.92	-1.045
Australia and New Zealand	Oceania	31.400	31.40	0.000

In examining the range of gini coefficients within each region, we can deduce that there is more variation in income inequality amongst the sub regions in Africa and the Americas whereas the subregions in Asia, Europe and Oceania are for the most part fairly close to the regional median.

Within Africa, Southern Africa showed the least amount of income inequality for the region whereas Eastern and Northern Africa showed the largest amount of income inequality in the region. This may be likely due to the higher instability within some of the countries within those subregions including the civil wars and famines raveging numerous Eastern African countries.

Looking at the Americas, we can see the larger amount of income inequality is found within North America which has been seeing an increasing amount of wealth being accumulated by the few and poverty rates increasing at the same time. Looking at South America (based on consumption), there seems to be much less inequality within the subregion.

The other noteworthy difference in gini coefficients is when looking at Southern Asia based on Income (Net). There appears to be much less income inequality in this subregion however, this may be due to the lack of reporting from lower income areas for net income.

Conclusion

The analysis of our three use cases or objectives demonstrates that income inequality does exist throughout the world and has increased in recent decades, though at varying rates of change. While shares of national income distribution have remained somewhat stable in some regions, the other regions particularly North America indicate a strong bias towards increasing income disparity with economic growth.

Appendix

A: Data Preparation

Starting with the raw data set WIID_19Dec2018.xlsx downloaded from UNU-WIDER

Step 1: Apply obvious filters

```
df = pd.read_excel('WIID_19Dec2018.xlsx')

#The filters I used were (areacovr="All", popcovr="All",
reference_unit="Person", sharing_unit="Household", resource not equal to
"Income (net/gross)").

df2 = df[(df['areacovr']=='All') & (df['popcovr']=='All') &
(df['reference_unit']== 'Person') \
     & (df['sharing_unit'] == "Household") & ~(df['resource'] == "Income
(net/gross)")]
```

Step 2: Select data with highest quality scores

```
#select the dataset with highest quality score if and when there is
collision on year, country, resource and scale
df3 = df2[df2['quality_score'] ==
df2.groupby(['year','country','resource','scale',])['quality_score'].tra
nsform('max')]
```

Step 3: Remove Duplicates

```
# Create the source dictionary for ranking studies; ranks are assigned
based on analysis of data quality scores and availability of various
features
source_dictionary ={'Luxembourg Income Study':1,\
'World Bank': 2, \
'United Nations':3, \
'Research study':4, \
'National statistical authority':5, \
'OECD':6, \
'SEDLAC':7, \
'Eurostat':8, \
'Other international organizations':9}

# Add a new column named 'source_rank' to help eliminate duplicates in
high quality data
```

```
df3['source_rank'] = df3['source'].map(source_dictionary)

# include newly engineered filed 'source_rank' in sorting the dataframe
df3_sorted=df3.sort_values(by = ['country', 'year','quality_score',
'source_rank'], ascending = [True,True,False,True])

# Dedup sorted dataframe
df4 =df3_sorted.drop_duplicates(subset=['country',
'year','resource','scale'])
```

Step 4: Validate no duplication

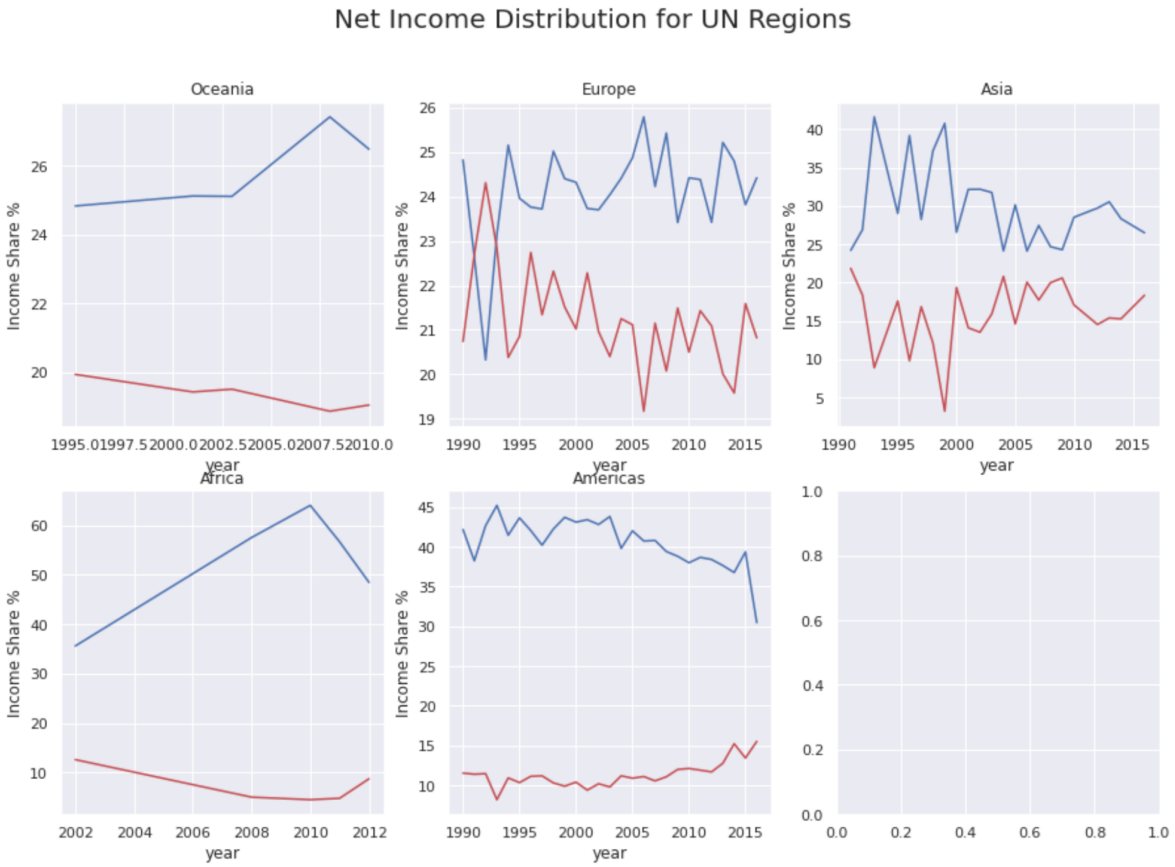
```
df4 =
(df3.groupby(['year','country','resource','scale','quality','quality_sco
re']).size()\
.sort_values(ascending=False) \
   .reset_index(name='count'))

df4[df4['count'] > 1]
```

Step 5: Save revised new CSV

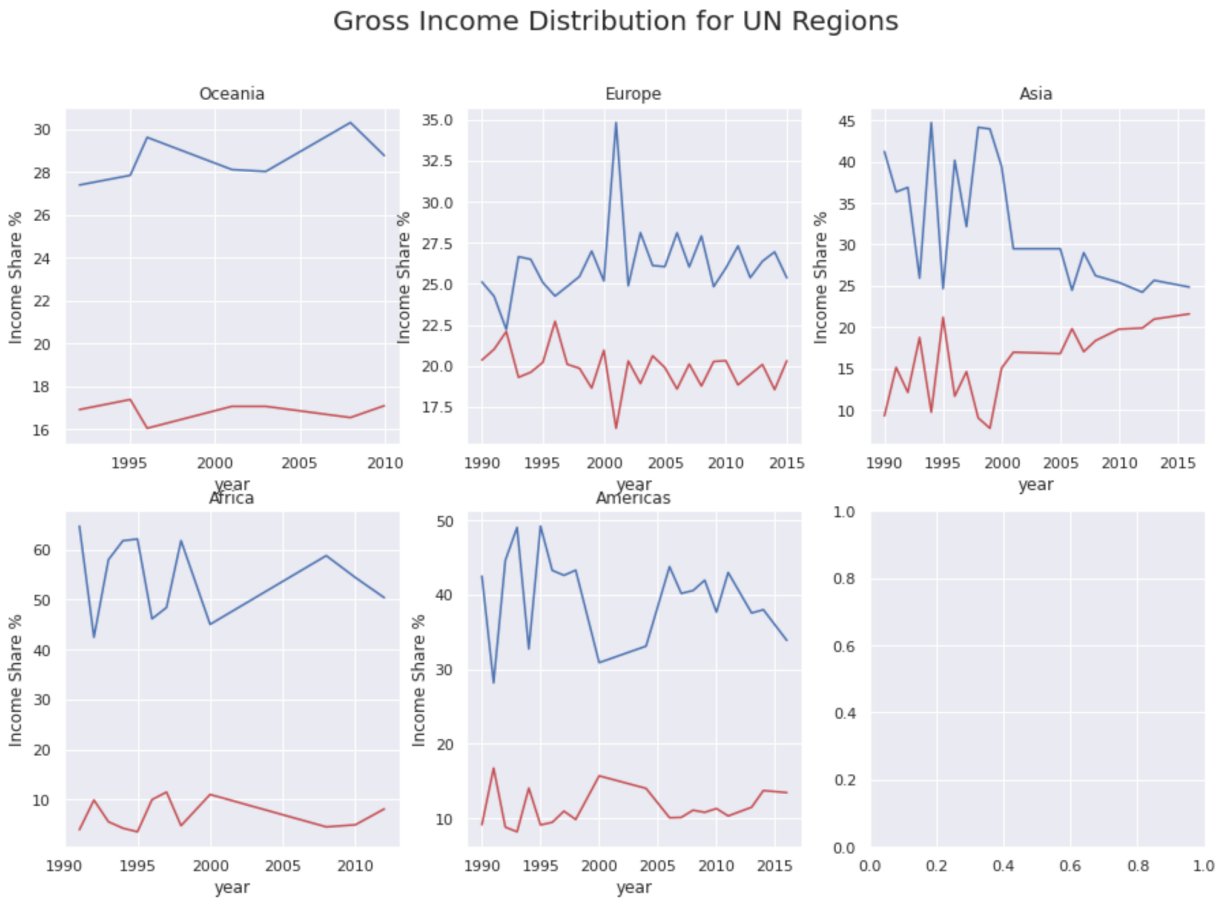
```
#no duplicates found; save the resulting data_set
df4.to_csv("revised_new_wiid.csv")
```

B: Figures pertaining to Use Case #1 (Top 10 vs Bottom 40

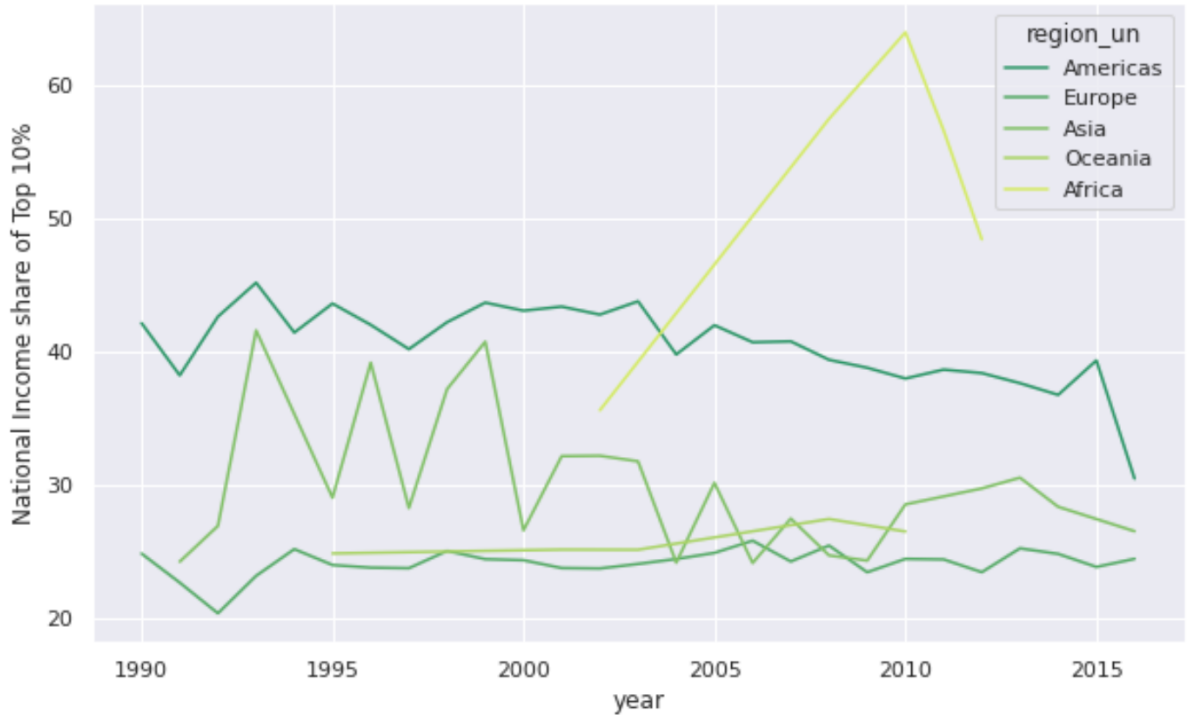
B.1: Net Income Distribution at UN Region level



B.2: Gross Income Distribution for UN Regions

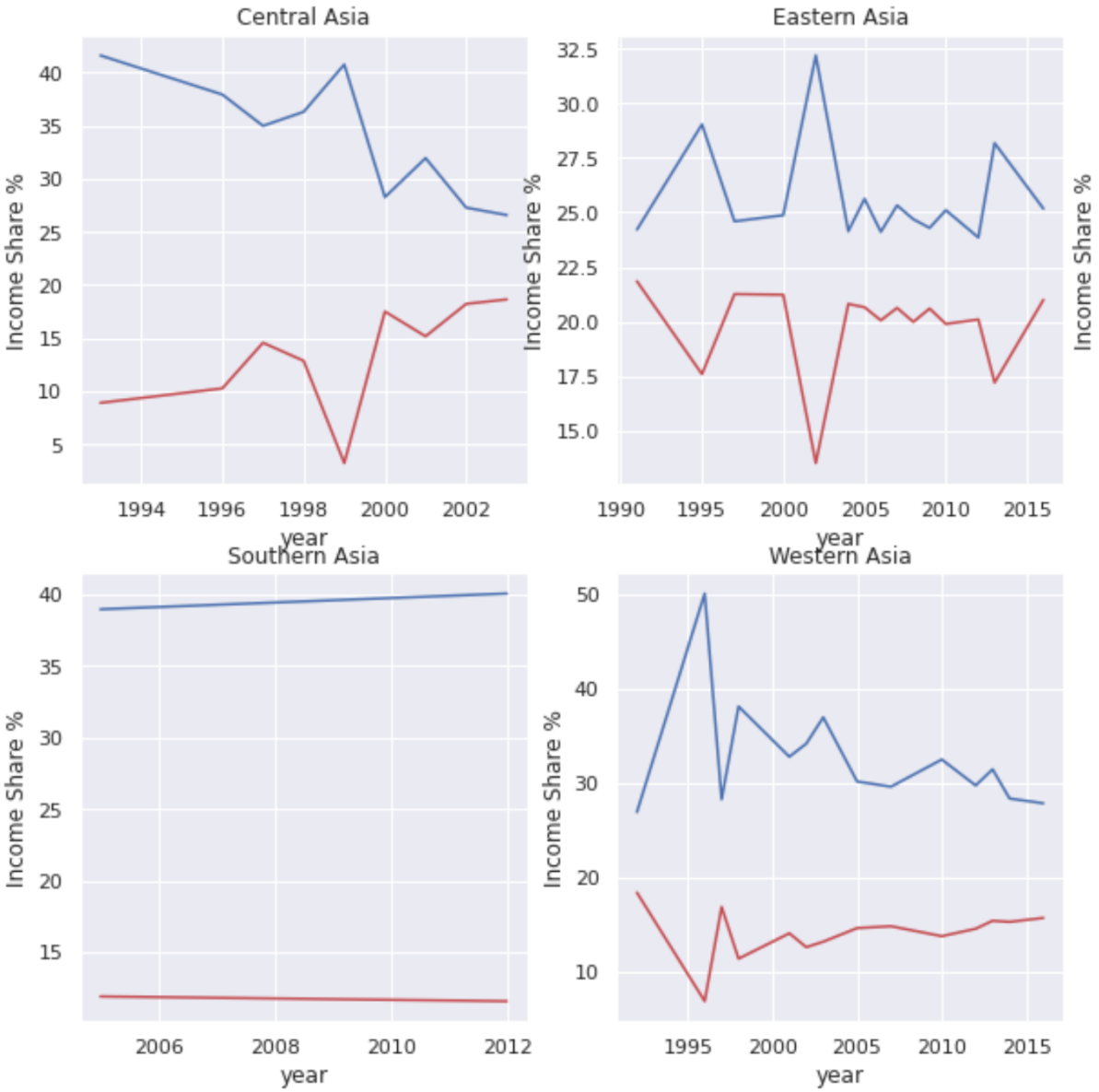


B 3: Regional Comparison of the Net Income share of the Richest 10%

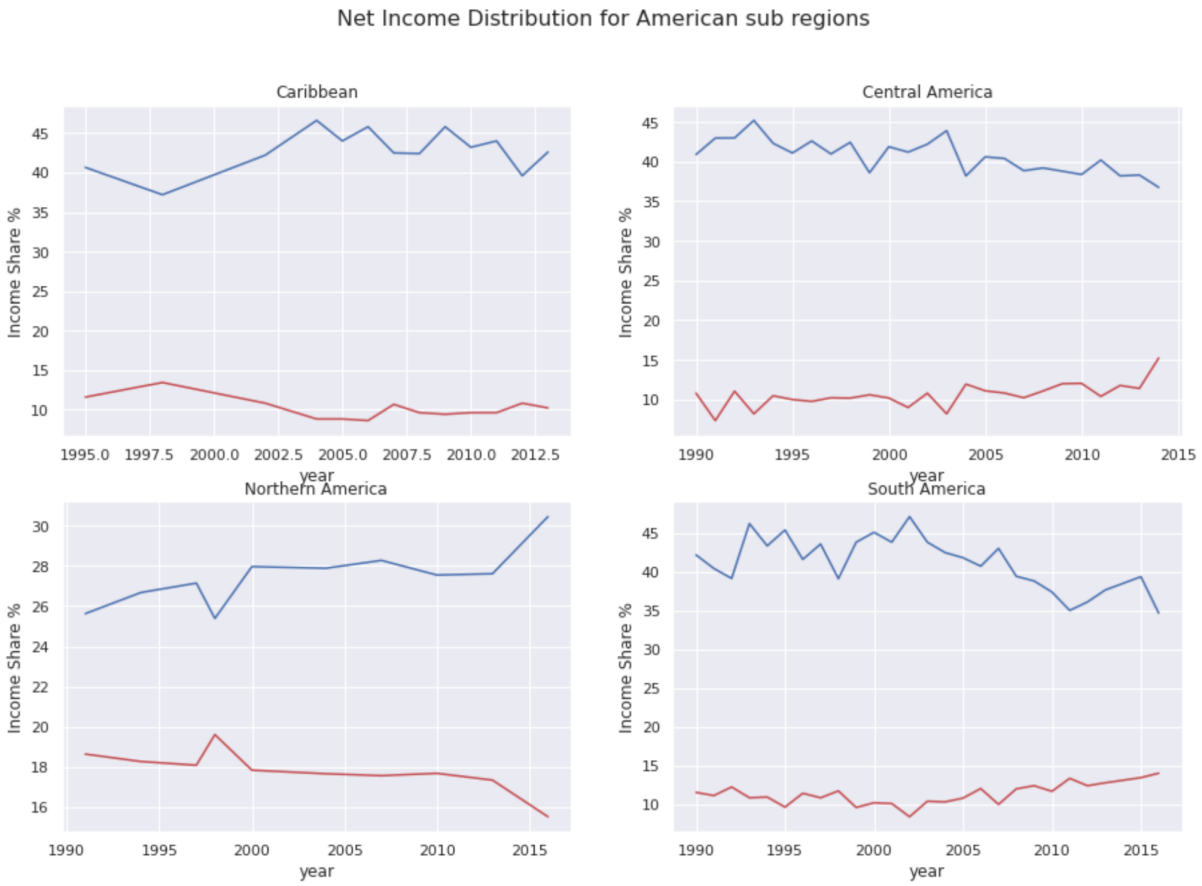


B 4: European Sub Regional Perspective on Net Income Distribution

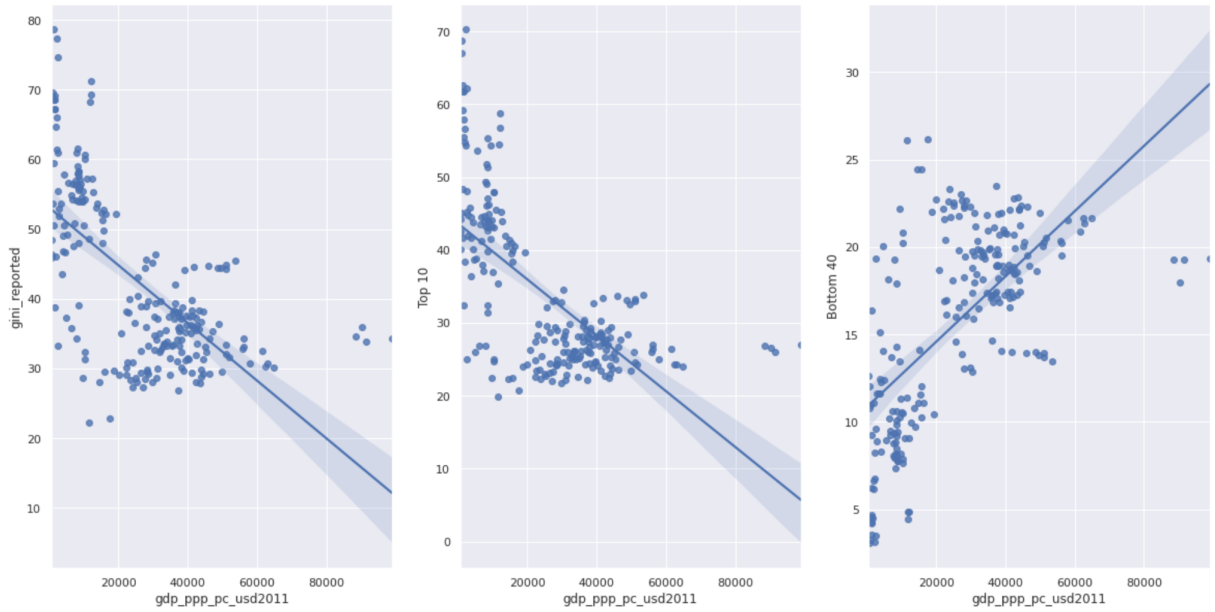
B5. Asia Sub Regional Perspective on Net Income Distribution



B 6: Americas sub Regional Perspective on Net Income Distribution

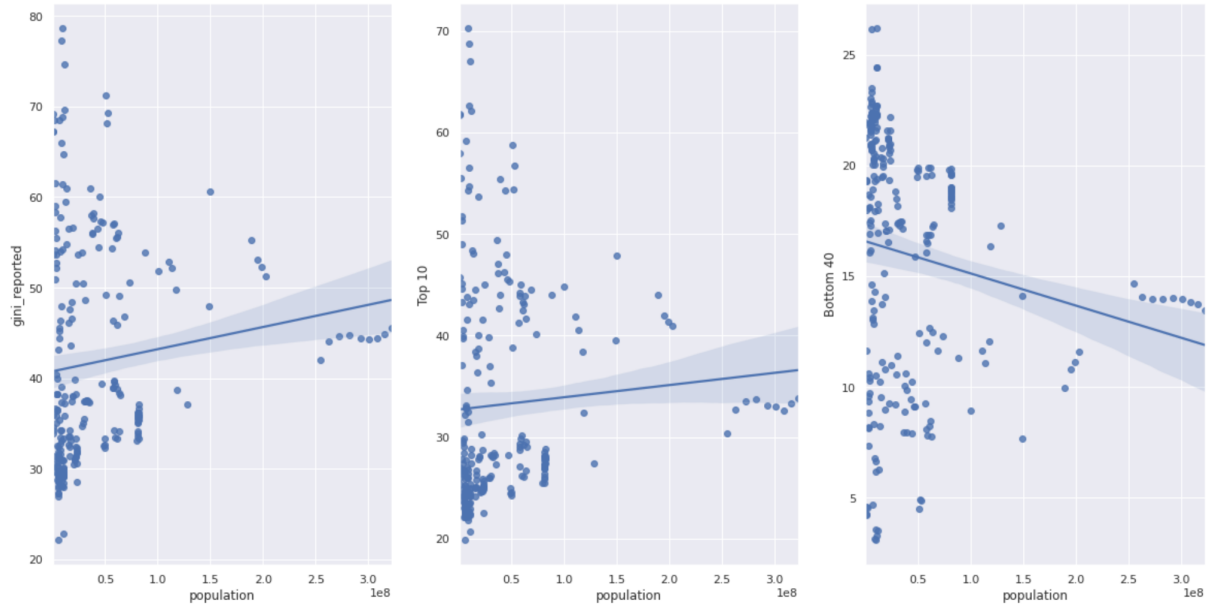


B 7: Regression Plot of GDP vs Gini Index and Gross Income shares of Top 10% and Bottom 40% respectively



Overall trend throughout the world shows a reduction in income disparity with GDP growth. However, as we know certain regions have had negative correlation and have reported increased disparity, case in point being North America as shown in the B6 above. This is further supported by the scatter plot matrix in B9 below.

B 8: Regression Plot of Population vs Gini Index and Gross Income shares of Top 10% and Bottom 40%



B 9: Scatter plot matrix showing correlations among GDP, Gini Index and income share distribution in North America