Country Twenty-Three

Insights into a Bright Future

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusions

Executive Summary

The purpose of this analysis was to determine which economic factors most effectively predict high Gross Domestic Product (GDP) values, and lower poverty rates across countries. Using exploratory data analysis (EDA) and machine learning, the following key findings were identified:

- **Low poverty** was most strongly associated with higher export values, a larger middle class, and increased spending on education.
- **High GDP** was most strongly associated with export values, followed by higher college enrollment and an ideal range of governmental transparency and accountability.

This study enables Country23's parliament and economic leaders to prioritize and align policy reforms with their national vision. It also provides an opportunity to consult or partner with peer countries that exemplify these outcomes

Introduction

- Country23 (for the sake of discretion) and the surrounding region have experienced many years of political and economic instability. Following recent elections, foreign investment, and a renewed commitment to reform, senior officials have requested an economic analysis to guide policy decisions aimed at improving national development. Country23, currently ranked in the lower third globally for GDP per capita and high poverty levels, seeks actionable insights into the key factors driving these outcomes. This study will use robust, data-driven methods to identify and explain the economic drivers most critical to GDP growth and poverty reduction, with the goal of supporting informed discussions and effective decision-making among parliamentary leaders

Methodology

Methodology

DATA COLLECTION

DATA WRANGLING

EXPLORATORY DATA ANALYSIS

PREDICTIVE ANALYTICS

[Link to Country23's github repository]

Data Collection

Data Collection

IDENTIFY

- Find sources of economic data
- Identify data relative to client request

REVIEW

- Review and vet data with client
- Manually download tables of interest

LOAD

- Upload data to repository
- Read dataframes into Jupyter notebook

UNdata
A world of information

WORLD BANK GROUP

Data Collection – Target Data

- **Gross Domestic Product per capita (df_gdp)** [$US/Capita] - Monetary value of all goods and services produced within a country's borders - seen as a key indicator of economic health

- **Poverty Rate (df_pov)** [%Population] - The percentage of the country's population that is at or below the poverty line as measured by the United Nations

Data Collection – Feature Dataframes (CPIA Scores)

CPIA (Country Policy and Institutional Assessment) is a rating system (1–6) used to evaluate how well a country's policies and institutions support sustainable growth, poverty reduction, and effective use of development resources

- **CPIA – Business Regulation (df_reg)** [Rating 1-6] - Assessment rating that measures how conducive a country's policies are for private sector development (e.g. Ease of operating a business, Regulatory framework, Property rights)
- **CPIA - Gender Equity (df_gender)** [Rating 1-6] - Assessment rating that measures the extent to which a country's policies promote gender equity and empower women
- **CPIA - Social Inclusion (df_social)** [Rating 1-6] - Assessment rating that measures how well everyone, regardless of background, can participate fully in society
- **CPIA - Transparency Accountability and Corruption (df_tac)** [Rating 1-6] - Assessment that measures how open governments operate, the mechanisms in place to hold public officials responsible, and the prevalence of corrupt practices in the public sector
- **CPIA - Public Resource Equity (df_pre)** [Rating 1-6] - Assessment that measures how well governments allocates its public resources so that all segments of society benefits
- **CPIA - Trade (df_trd)** [Rating 1-6] - Assessment rating that measures how supportive a country's trade policies are of integration into the global economy (e.g. Tariff barriers, Customs efficiency, Trade openness).

Data Collection – Feature Dataframes (Financial)

Trade

- **Commodity Import Value (df_trdc)** [$US/Capita] - Value of goods imported into a country divided by the population of that country
- **Commodity Export Value (df_trdc)** [$US/Capita] - Value of goods exported out of a country divided by the population of that country

Income

- **Income distribution to 2nd Quintile (df_inc2q)** [%Income] - Percentage of a country's total income that is earned by the second lowest earning quintile (20% segment) of the population
- **Income distribution to 3rd Quintile (df_inc3q)** [%Income] - Percentage of a country's total income that is earned by the third lowest earning quintile (20% segment) of the population
- **Income distribution to 4th Quintile (df_inc4q)** [%Income] - Percentage of a country's total income that is earned by the fourth lowest (or 2nd highest) earning quintile (20% segment) of the population
- **Income distribution to 5th Quintile (df_inc5q)** [%Income] - Percentage of a country's total income that is earned by the highest earning quintile (20% segment) of the population
- **Income distribution to Top 10% (df_inc4q)** [%Income] - Percentage of a country's total income that is earned by the top 10% of the population

Data Collection – Feature Dataframes (Other)

- **Healthcare expenditures (df_health)** [$US/Capita] – The value of a country's total expenditures on healthcare related goods and services divided by that country's population
- **Education expenditures (df_edu)** [$US/Capita] – The value of a country's total expenditures on educational goods and services divided by that country's population
- **Gross College Enrollment (df_college)** [%Population] - The number of a country's population enrolled in secondary education divided by number of college aged citizens of that country
- **Ease of Doing Business (df_edb)** [Rating 0 -100] – Measures how a country's policy's and practices support the ability to start, operate, and close a business
- **Population (df_pop) [Count]** – The number of citizens in a country. This data will be used to convert other variable's absolute values to per capita values

Data Wrangling

Data Wrangling – Overview

Clean → Merge → Reduce → Impute → Transform

[Link to Data Wrangling Jupyter Notebook]

Data Wrangling – Clean Data (1 of 3)

The purpose of Data Wrangling to prepare disparate dataframes to be seamlessly merged into one dataframe for the purpose of future analysis

- Remove white space characters

“\s\sBrazil\s” → “Brazil”

- Lower case column names

“Country or Area” → “area”

- Column names common across dfs

	Reference Area	Time Period
0	Afghanistan	2014

	Country or Area	Year
0	Afghanistan	2019

→

Table 1 - First five

	area	year
0	Afghanistan	2021

Table 2 - First five

	area	year
0	Afghanistan	2016

- Remove non-informative columns

	Country or Area	Year	Value	~~Value Footnotes~~
0	Afghanistan	2021	1673.964059	~~NaN~~
1	Afghanistan	2020	2078.595086	~~NaN~~

→

Table 1 - First five rows of tl

	area	year	gdp
0	Afghanistan	2021	1673.96
1	Afghanistan	2020	2078.60

Data Wrangling - Clean Data (2 of 3)

- Manage footers & headers where present

	T11	Expenditure on health	Unnamed: 2	Unnamed: 3	Unnamed: 4
0	Region/Country/Area	NaN	Year	Series	Value
1	4	Afghanistan	2005	Current health expenditure (% of GDP)	9.9
2	4	Afghanistan	2010	Current health expenditure (% of GDP)	8.6
3	4	Afghanistan	2015	Current health expenditure (% of GDP)	10.1
4	4	Afghanistan	2019	Current health expenditure (% of GDP)	14.8

```
Table 9 - First five rows of the dataframe
0        area  year  healthcare$
0  Afghanistan  2005          9.9
1  Afghanistan  2010          8.6
2  Afghanistan  2015         10.1
3  Afghanistan  2019         14.8
4  Afghanistan  2020         15.5
```

- Adjust data types as needed

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2265 entries, 0 to 2264
Data columns (total 7 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   T11                    2265 non-null   object
 1   Expenditure on health  2264 non-null   object
 2   Unnamed: 2             2265 non-null   object
 3   Unnamed: 3             2265 non-null   object
 4   Unnamed: 4             2265 non-null   object
 5   Unnamed: 5             1736 non-null   object
 6   Unnamed: 6             2265 non-null   object
dtypes: object(7)
memory usage: 124.0+ KB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1132 entries, 0 to 1131
Data columns (total 3 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   area         1132 non-null   object
 1   year         1132 non-null   int64
 2   healthcare$  1132 non-null   float64
dtypes: float64(1), int64(1), object(1)
memory usage: 26.7+ KB
```

- Adjust decimals for floats

area	year	comm_export_value
Brazil	2014	56892.22356

area	year	comm_export_value
Brazil	2014	56892.22

Data Wrangling – Clean Data (3 of 3)

- Convert long form data into tidy form where necessary

Country Name	1960.0	1961.0	1962.0	1963.0
Aruba	54608.0	55811.0	56682.0	57475.0
Africa Eastern and Southern	130692579.0	134169237.0	137835590.0	141630546.0
Afghanistan	8622466.0	8790140.0	8969047.0	9157465.0
Africa Western and Central	97256290.0	99314028.0	101445032.0	103667517.0

area	year	population
Aruba	1960	54608.0
and Southern	1960	130692579.0
Afghanistan	1960	8622466.0
and Central	1960	97256290.0
Angola	1960	5357195.0
...	...	...

- Convert tidy form data into long form where necessary

Country or Area	Year	Commodity	Flow	Trade (USD)
Afghanistan	2019	All Commodities	Export	8.704885e+08
Afghanistan	2019	All Commodities	Import	8.568014e+09
Afghanistan	2019	All Commodities	Re-Export	6.655197e+06
Afghanistan	2018	All Commodities	Import	7.406590e+09
Afghanistan	2018	All Commodities	Re-Export	9.263097e+06

Table 13 - First five rows of the dataframe for trade after i

	area	year	comm_import_capita	comm_export_capita
0	Afghanistan	2019	226.850080	23.047394
1	Afghanistan	2018	201.887152	24.109622
2	Afghanistan	2017	218.626623	23.340264
3	Afghanistan	2016	188.650576	17.220573
4	Afghanistan	2015	228.801910	16.928762

- Combine column data to make new features

```
df_income['income_quintile2'] + df_income['income_quintile3'] + df_income['income_quintile4']
```

income_middle60%
53.5
52.7
53.8
51.8
51.4

Data Wrangling – Merge Data

Table 1 - First five rows of the

	area	year	gdp
0	Afghanistan	2021	1673.96
1	Afghanistan	2020	2078.60
2	Afghanistan	2019	2168.13
3	Afghanistan	2018	2110.24
4	Afghanistan	2017	2096.09

7728 Rows , 3 Columns

Table 2 - First five rows of

	area	year	%pov
0	Afghanistan	2016	54.5
1	Afghanistan	2011	38.3
2	Afghanistan	2007	33.7
3	Albania	2020	22.0
4	Albania	2019	21.8

1012 Rows , 3 Columns

Table 9 - First five rows of the dat

0	area	year	healthcare$
0	Afghanistan	2005	9.9
1	Afghanistan	2010	8.6
2	Afghanistan	2015	10.1
3	Afghanistan	2019	14.8
4	Afghanistan	2020	15.5

1132 Rows . 3 Columns

Table 11 - First five rows of the dataf

	area	year	coll_enrollment
0	Afghanistan	2014	55.65616
1	Afghanistan	2013	56.68866
2	Afghanistan	2012	56.67734
3	Afghanistan	2011	54.61618
4	Afghanistan	2010	53.24683

5989 Rows . 3 Columns

Table 13 - First five rows of the dataframe for trade after i

	area	year	comm_import_capita	comm_export_capita
0	Afghanistan	2019	226.850080	23.047394
1	Afghanistan	2018	201.887152	24.109622
2	Afghanistan	2017	218.626623	23.340264
3	Afghanistan	2016	188.650576	17.220573
4	Afghanistan	2015	228.801910	16.928762

4014 Rows . 4 Columns

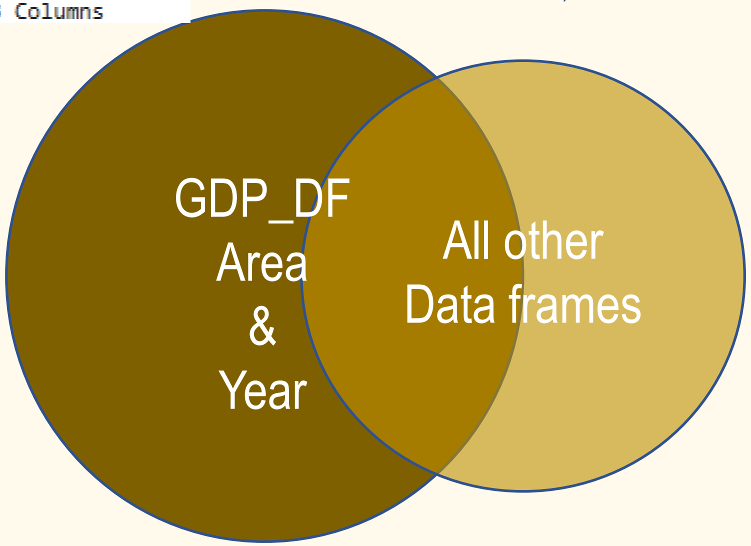


Table 14 - First five rows of the Master Dataframe_v1

	area	year	gdp	%pov	cpia_regulation	cpia_gender	\
0	Afghanistan	2002	943.12	NaN	NaN	NaN	
1	Afghanistan	2003	970.65	NaN	NaN	NaN	
2	Afghanistan	2004	971.81	NaN	NaN	NaN	
3	Afghanistan	2005	1075.67	NaN	NaN	NaN	
4	Afghanistan	2006	1120.89	NaN	2.5	2.0	

	cpia_resources	cpia_transparency	cpia_inclusion	cpia_trade	...	\
0	NaN	NaN	NaN	NaN	...	
1	NaN	NaN	NaN	NaN	...	
2	NaN	NaN	NaN	NaN	...	
3	NaN	NaN	NaN	NaN	...	
4	2.5	2.5	2.3	3.0	...	

	coll_enrollment	income_quintile2	income_quintile3	income_quintile4	\
0	NaN	NaN	NaN	NaN	
1	13.31708	NaN	NaN	NaN	
2	18.66479	NaN	NaN	NaN	
3	19.78370	NaN	NaN	NaN	
4	29.93046	NaN	NaN	NaN	

	income_quintile5	income_top10%	income_middle60%	\
0	NaN	NaN	NaN	
1	NaN	NaN	NaN	
2	NaN	NaN	NaN	
3	NaN	NaN	NaN	
4	NaN	NaN	NaN	

	income_difference_top-mid60	comm_import_capita	comm_export_capita
0	NaN	NaN	NaN
1	NaN	NaN	NaN
2	NaN	NaN	NaN
3	NaN	NaN	NaN
4	NaN	NaN	NaN

[5 rows x 22 columns]
5533 Rows , 22 Columns

Data Wrangling – Additional Reduction and Maintenance

- Standardize names of countries
  - Identify variations of same country
  - Run area names through dictionary to standardize those variations

Bolivia
Bolivia (Plurin. State of)
Bosnia and Herzegovina
Botswana

Burundi
CÃ´te dâ€™Ivoire
CÃ´te d'Ivoire
CÃƒÂ´te d'Ivoire
Cabo Verde

Cyprus
Czech Republic
Czechia
Dem. Rep. Congo
Dem. Rep. of the Congo
Democratic Republic of the C
Denmark
Djibouti

- Remove non-geographical records

High income
Holy See
IBRD only
IDA & IBRD total
IDA blend
IDA only
IDA total

Low & middle income
Low income
Lower middle income

OECD members
Other small states
Post-demographic dividend
Pre-demographic dividend

- Reassign region level redundant data to country level missing data
  - Add regional column (tidy to long form)
  - Impute regional data to their country's missing data
  - Rename 'area' column to 'country'
  - Drop region specific data

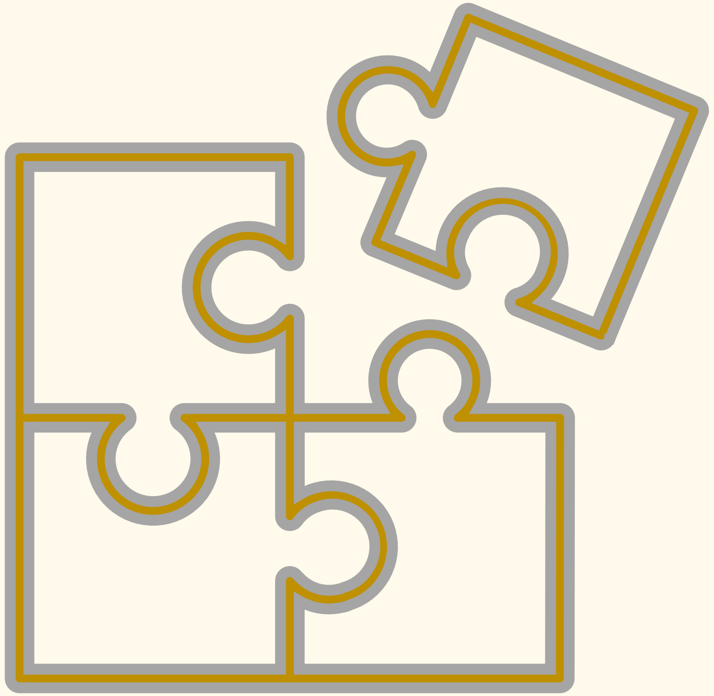
area	cpia_reg	healthcare$
Germany	NaN	5.6
France	NaN	6.6
Europe & Central Asia	3.8	NaN
Ghana	3	4.2
D.R. Congo	NaN	NaN
Western & Central Africa	2.2	4.5

country	region	un_region	cpia_reg	healthcare$
Germany	Central Europe	Europe & Central Asia	3.8	5.6
France	Europe	Europe & Central Asia	3.8	6.6
Ghana	Western Africa	Western & Central Africa	3	4.2
D.R. Congo	Central Africa	Western & Central Africa	2.2	4.5

Data Wrangling – Manage Missing Data (1 of 3)

1. Split single dataframe into GDP and Poverty dfs

2. Figure out some derivation of those dfs with:
    1. Only 15% missing data in any feature
    2. Still an acceptable number of records in df



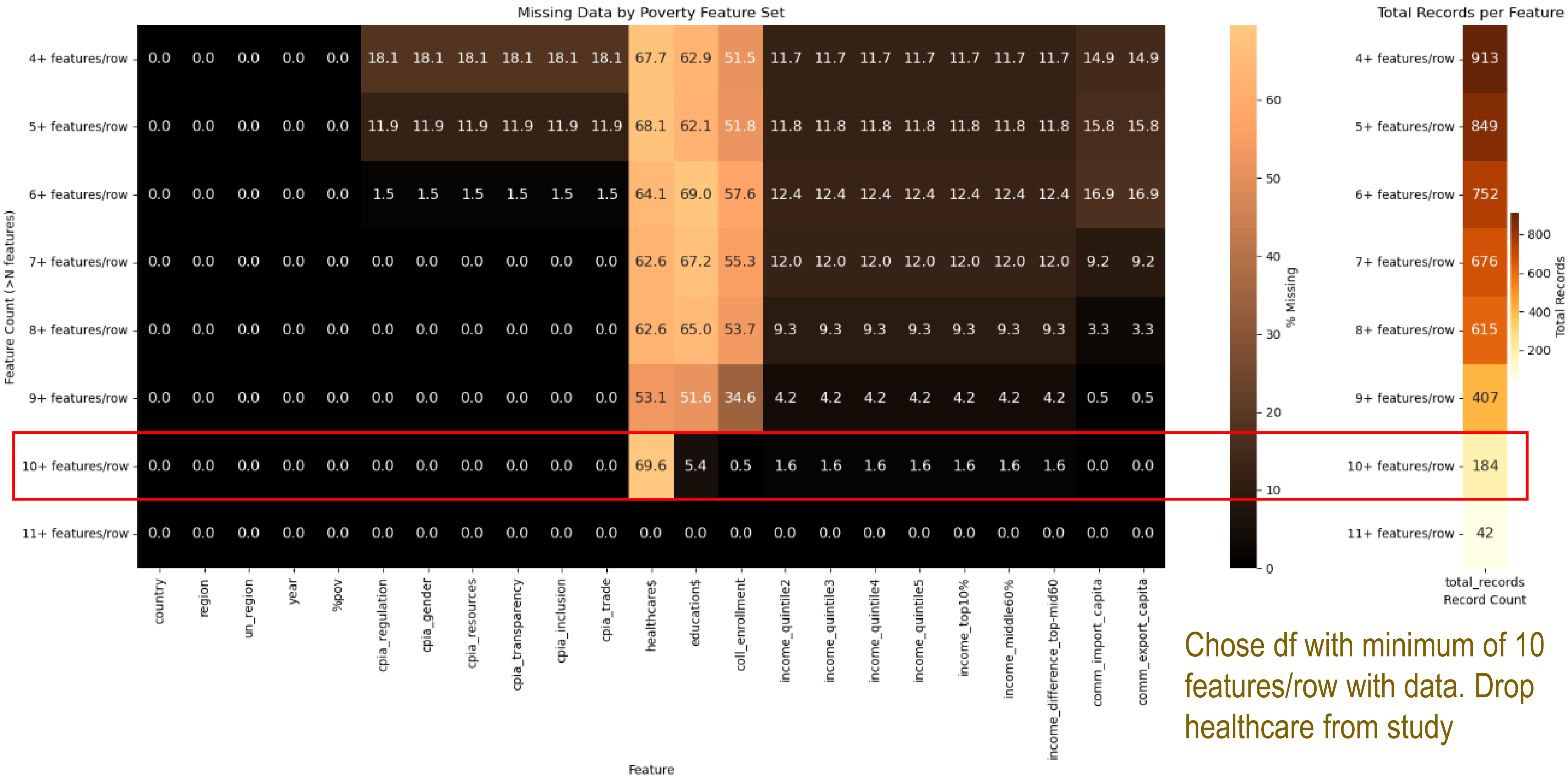
Question to answer:

What minimum number of features per record having data would accomplish the above goal?

- 4 feature per record/row?, 5?...

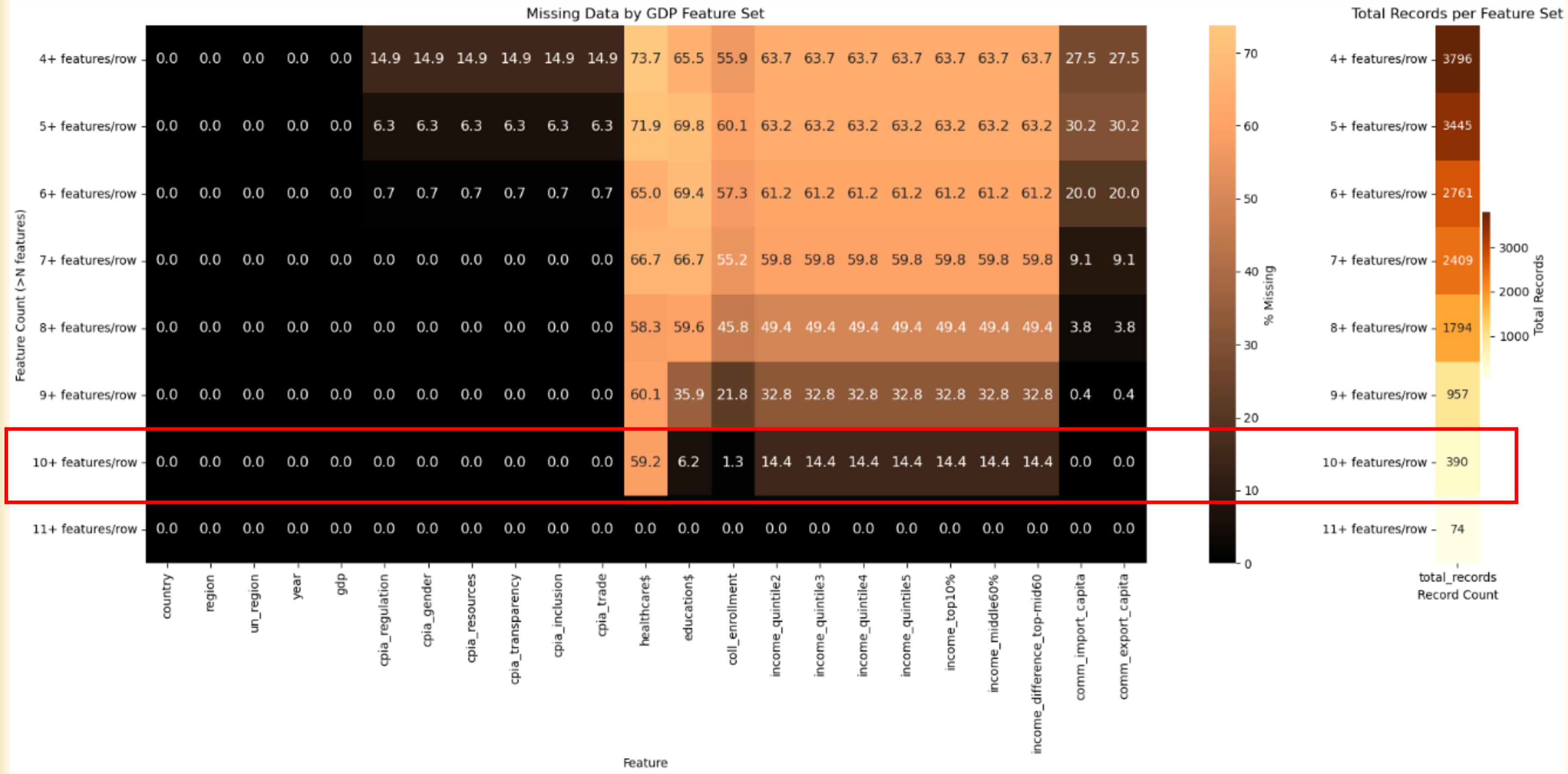
Link to Missing Data Jupyter Notebook

Data Wrangling – Manage Missing Data (2 or 3)



Chose df with minimum of 10 features/row with data. Drop healthcare from study

Data Wrangling – Manage Missing Data (3 or 3)



GDP Results Identical to Poverty

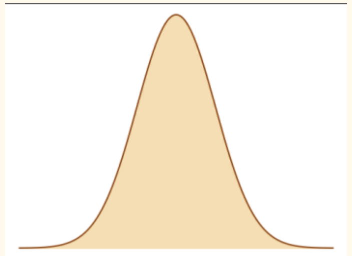
Drop healthcare expenditure feature

Impute the median (robust to outliers, and skewness) into missing values

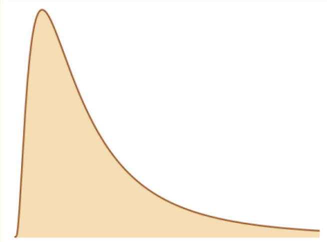
- Median approach based on histograms that can be viewed in missing data notebook link

Data Wrangling – Transform Data

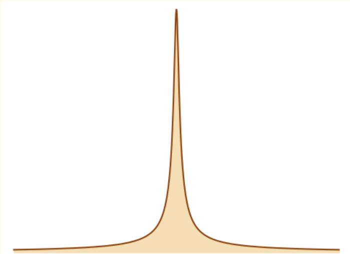
Regression models expect normality



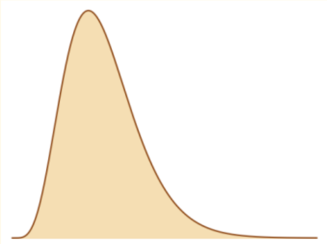
- Find out which (if any) transformation is needed
  - How skewness and kurtosis
  - Skewness > 1 or Kurtosis > 10:
    - ❖Log transformation



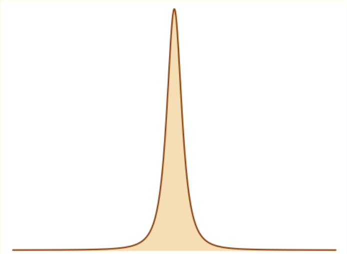
OR



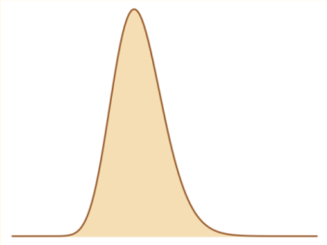
  - Skewness > 0.5 or Kurtosis > 5:
    - ❖Square root transformation



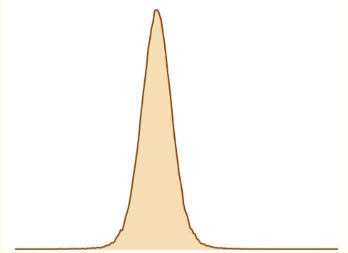
OR



  - Skewness > 0.25 or Kurtosis > 3:
    - ❖Box-Cox transformation



OR



Data Wrangling – Transformation Results

Transformation Results for Poverty Dataframe

Study Variables	Transformations			
	log	sqrt	boxcox	none
%pov	X			
cpia_regulation	X			
cpia_gender				X
cpia_resources		X		
cpia_transparency		X		
cpia_inclusion				X
cpia_trade		X		
education$				X
coll_enrollment				X
income_quintile2				X
income_quintile3				X
income_quintile4		X		
income_quintile5	X			
income_top10%	X			
income_middle60%				X
comm_import_capita		X		
comm_export_capita	X			

Transformation Results for GDP Dataframe

Study Variables	Transformations			
	log	sqrt	boxcox	none
gdp	X			
cpia_regulation			X	
cpia_gender				X
cpia_resources			X	
cpia_transparency		X		
cpia_inclusion				X
cpia_trade		X		
education$		X		
coll_enrollment				X
income_quintile2				X
income_quintile3				X
income_quintile4		X		
income_quintile5	X			
income_top10%	X			
income_middle60%				X
comm_import_capita	X			
comm_export_capita	X			

Only minor differences between Poverty and GDP dataframes

Exploratory Data Analysis (EDA)

Exploratory Data Analysis Approach using SQL

GENERAL INFORMATION

RANKING

CHANGES AT THE EXTREMES

Link to Exploratory Data Analysis Jupyter Notebook

EDA – General Information about Poverty Dataframe

Number of Countries and Regions

	Count of Countries	Count of Regions
0	51	12

Timespan of Study

	Year of Earliest Record	Year of Latest Record
0	2005	2015

Count of Countries in each Region with additional Poverty Information

	Regions	Count of Countries per Region	Count of Countries with Poverty over 30%	Count of Countries with Poverty under 15%
0	Western Africa	2	2	0
1	Southern Africa	1	1	0
2	South Asia	5	2	0
3	Pacific	1	0	0
4	Northern Africa	1	0	0
5	Middle East	1	0	0
6	Latin America	8	6	1
7	Europe	19	1	9
8	Eastern Africa	8	6	1
9	East Asia	3	0	2
10	Central Africa	1	1	0
11	Caribbean	1	0	1

- North America, and Australia have no records in poverty study
- Europe has more than twice as many countries than next highest region
- 18% of poverty data is at the extremes (above 30% or below 15%)
- 80% of poverty data under 15% is from Europe and East Asia
- 84% of poverty data over 30% is from Africa and Latin America

Count of Records per Year

	Year	Records per Year
0	2005	26
1	2006	14
2	2007	18
3	2008	16
4	2009	20
5	2010	29
6	2011	24
7	2012	19
8	2013	12
9	2014	5
10	2015	1

- Smaller number of records after 2012

EDA – Countries with Highest and Lowest Poverty Rates

Recorded Years of Poverty

15 Highest Recorded Years

	Country	Year	% in Poverty
0	Madagascar	2005	73.2
1	Madagascar	2012	70.7
2	South Africa	2005	66.6
3	Burundi	2013	64.9
4	South Africa	2008	62.1
5	Peru	2005	55.6
6	South Africa	2014	55.5
7	South Africa	2010	53.2
8	Guatemala	2006	51.0
9	Malawi	2010	50.7
10	Pakistan	2005	50.4
11	Peru	2006	49.2
12	Nicaragua	2005	48.3
13	Senegal	2005	48.3
14	Kenya	2005	46.8

15 Lowest Recorded Years

	Country	Year	% in Poverty
0	Belarus	2014	4.8
1	Belarus	2010	5.2
2	Belarus	2009	5.4
3	Belarus	2013	5.5
4	Belarus	2012	6.3
5	Belarus	2011	7.3
6	Iceland	2011	7.9
7	Mauritius	2012	7.9
8	Azerbaijan	2010	9.1
9	Iceland	2010	9.2
10	Iceland	2005	9.6
11	Iceland	2009	9.8
12	Norway	2011	10.0
13	Iceland	2006	10.1
14	Iceland	2007	10.1

- Repeat Actors
- Lowest %pov – Belarus?

Average Poverty% (2005-2015)

Highest Avg Poverty% by Country

	Country	Avg Poverty %	Rank
0	Madagascar	71.950	1
1	Burundi	64.900	2
2	South Africa	59.350	3
3	Guatemala	51.000	4
4	Malawi	50.700	5
5	Senegal	48.300	6
6	Nicaragua	48.300	6
7	Kenya	46.800	8
8	Cameroon	39.900	9
9	Paraguay	39.520	10
10	Pakistan	39.420	11
11	Rwanda	39.100	12
12	Georgia	36.375	13
13	Bangladesh	35.750	14
14	Peru	34.900	15

Lowest Avg Poverty% by Country

	Country	Avg Poverty %	Rank
0	Belarus	6.742857	1
1	Mauritius	7.900000	2
2	Azerbaijan	9.100000	3
3	Iceland	9.557143	4
4	Norway	11.200000	5
5	Slovenia	12.625000	6
6	Finland	13.066667	7
7	Indonesia	13.662500	8
8	Sweden	14.985714	9
9	Malta	15.340000	10
10	Cyprus	15.414286	11
11	Jamaica	16.200000	12
12	Thailand	16.662500	13
13	China	17.200000	14
14	Portugal	18.114286	15

- 1 out of 2 citizens
- Vs 1 out of every 8

EDA – Fringe Poverty Values vs Feature Values

CPIA Scores

Poverty rate (15 highest values) and CPIA scores

		Avg Poverty% (15 Highest)	Regulation	Gender	Resources	Transparency	Inclusion	Trade	Combined CPIA Score
0	Total Average	47.1	3.6	3.6	3.6	3.0	3.5	4.0	3.6

Poverty rate (15 lowest values) and CPIA scores

		Avg Poverty% (15 Lowest)	Regulation	Gender	Resources	Transparency	Inclusion	Trade	Combined CPIA Score
0	Total Average	13.2	3.5	3.9	3.7	2.8	3.6	4.1	3.6

-3% Δ 8% Δ 3% Δ -7% Δ 3% Δ 2% Δ 0% Δ

Education

Poverty rate (15 highest values) and education factors

		Avg Poverty% (15 Highest)	Education $	College Enrollment
0	Total Average	47.1	4.3	52.0

Poverty rate (15 lowest values) and education factors

		Avg Poverty% (15 Lowest)	Education $	College Enrollment
0	Total Average	13.2	5.3	97.2

+23% Δ +87% Δ

Income

Poverty rate (15 highest values) and income distribution

		Avg Poverty% (15 Highest)	Quintile2	Quintile3	Quintile4	Quintile5	Middle60%	Top10%
0	Total Average	47.1	9.6	13.7	20.3	50.8	43.6	35.4

Poverty rate (15 lowest values) and income distribution

		Avg Poverty% (15 Lowest)	Quintile2	Quintile3	Quintile4	Quintile5	Middle60%	Top10%
0	Total Average	13.2	12.5	16.5	22.0	40.9	51.1	26.1

+16% Δ -26% Δ

Trade Values

Poverty rate (15 highest values) and commodity trade values

		Avg Poverty% (15 Highest)	Commodity Import Value	Commodity Export Value
0	Total Average	47.1	542.8	348.4

Poverty rate (15 lowest values) and commodity trade values

		Avg Poverty% (15 Lowest)	Commodity Import Value	Commodity Export Value
0	Total Average	13.2	8117.8	7704.7

+1,395% Δ +2,114% Δ

- Significant changes in feature values (especially trade)
- Top 10% have smaller share of income for low poverty countries

EDA – General Information about GDP Dataframe

Number of Countries and Regions

Count of Countries	Count of Regions
98	12

Timespan of Study

Year of Earliest Record	Year of Latest Record
2005	2015

Count of Countries in each Region with additional GDP Information

	Regions	Count of Countries per Region	Count of Countries with GDP over $10k	Count of Countries with GDP under $3k
0	Western Africa	10	0	8
1	Southern Africa	3	2	1
2	South Asia	7	1	4
3	Pacific	5	2	1
4	Northern Africa	1	1	0
5	Middle East	6	5	0
6	Latin America	16	10	0
7	Europe	30	27	0
8	Eastern Africa	9	1	8
9	East Asia	6	4	0
10	Central Africa	3	0	2
11	Caribbean	2	1	0

- Europe has nearly twice as many countries as next highest region
- 21% of GDP data is at the extremes (above $10k and below $3k)
- 50% of GDP data above $10k is from Europe
- 80% of GDP data below $3k is from Africa

Count of Records per Year

	Year	Records per Year
0	2005	77
1	2006	32
2	2007	33
3	2008	36
4	2009	35
5	2010	80
6	2011	40
7	2012	27
8	2013	20
9	2014	8
10	2015	2

- Large changes in available records year to year

EDA – Countries with Highest and Lowest GDP Values

Recorded years of GDP

15 Highest Recorded Years

	Country	Year	GDP
0	Qatar	2010	143070.21
1	Qatar	2005	112073.10
2	Luxembourg	2010	90357.10
3	Luxembourg	2005	68787.85
4	Norway	2012	65774.35
5	Norway	2011	62460.09
6	Norway	2008	62072.75
7	Norway	2010	58226.71
8	Norway	2007	56175.66
9	Norway	2009	55620.84
10	Norway	2006	54366.01
11	Netherlands	2013	49241.52
12	Norway	2005	47966.86
13	Netherlands	2012	47272.10
14	Ireland	2007	46779.40

- Repeat actors again

15 Lowest Recorded Years

	Country	Year	GDP
0	Burundi	2005	567.70
1	Burundi	2010	630.36
2	Burundi	2013	696.50
3	Mozambique	2005	707.53
4	Niger	2005	881.85
5	Rwanda	2005	916.98
6	Niger	2007	948.69
7	Ethiopia	2010	1010.02
8	Niger	2010	1051.09
9	Niger	2011	1057.84
10	Niger	2014	1134.74
11	Burkina Faso	2005	1176.36
12	Rwanda	2010	1315.03
13	Togo	2005	1346.02
14	Madagascar	2005	1368.62

Average GDP (2005-2015)

Highest Avg GDP by Country

	Country	Avg GDP	Rank
0	Qatar	127571.7	1
1	Luxembourg	79572.5	2
2	Norway	57832.9	3
3	Saudi Arabia	46012.8	4
4	Netherlands	44634.8	5
5	Ireland	43974.7	6
6	Oman	43198.0	7
7	Sweden	41544.0	8
8	Iceland	40684.6	9
9	Denmark	39929.7	10
10	Finland	38293.5	11
11	Belgium	37377.8	12
12	Australia	36633.2	13
13	United Kingdom	36090.6	14
14	Japan	35545.3	15

- Top 15 is 36x wealthier than lowest 15

Lowest Avg GDP by Country

	Country	Avg GDP	Rank
0	Burundi	631.5	1
1	Mozambique	707.5	2
2	Ethiopia	1010.0	3
3	Niger	1014.8	4
4	Rwanda	1245.3	5
5	Burkina Faso	1423.0	6
6	Madagascar	1429.1	7
7	Guinea	1438.6	8
8	Malawi	1468.1	9
9	Togo	1476.6	10
10	Mali	1607.3	11
11	Lesotho	1660.1	12
12	Afghanistan	1771.2	13
13	Uganda	1869.6	14
14	Benin	2011.4	15

EDA – Fringe GDP Values vs Feature Values

CPIA Scores

+6% Δ +15% Δ +3% Δ -7% Δ +6% Δ +11% Δ +3% Δ

Poverty rate (15 highest values) and CPIA scores

		Avg GDP (15 Highest)	Regulation	Gender	Resources	Transparency	Inclusion	Trade	Combined CPIA Score
0	Total Average	49926.4	3.5	3.8	3.6	2.7	3.5	4.2	3.5

Poverty rate (15 lowest values) and CPIA scores

		Avg GDP (15 Lowest)	Regulation	Gender	Resources	Transparency	Inclusion	Trade	Combined CPIA Score
0	Total Average	1384.3	3.3	3.3	3.5	2.9	3.3	3.8	3.4

Education

+17% Δ +248% Δ

Poverty rate (15 highest values) and education factors

		Avg GDP (15 Highest)	Education $	College Enrollment
0	Total Average	49926.4	5.6	108.4

Poverty rate (15 lowest values) and education factors

		Avg GDP (15 Lowest)	Education $	College Enrollment
0	Total Average	1384.3	4.8	30.8

Income

+8% Δ -16% Δ

Poverty rate (15 highest values) and income distribution

		Avg GDP (15 Highest)	Quintile2	Quintile3	Quintile4	Quintile5	Middle60%	Top10%
0	Total Average	49926.4	13.0	13.0	13.0	13.0	52.4	24.6

Poverty rate (15 lowest values) and income distribution

		Avg GDP (15 Lowest)	Quintile2	Quintile3	Quintile4	Quintile5	Middle60%	Top10%
0	Total Average	1384.3	11.2	11.2	11.2	11.2	48.1	29.9

Trade Values

+10,589% Δ +28,500% Δ

Poverty rate (15 highest values) and commodity trade values

		Avg GDP (15 Highest)	Commodity Import Value	Commodity Emport Value
0	Total Average	49926.4	15819.9	18876.0

Poverty rate (15 lowest values) and commodity trade values

		Avg GDP (15 Lowest)	Commodity Import Value	Commodity Export Value
0	Total Average	1384.3	148.9	66.4

- GDP Feature changes aligned with poverty feature changes in direction
- Larger than expected changes in trade values
- More people enrolled in college than there are college aged citizens in high GDP countries

Data Visualization

Visualization Approach

SPREAD OF DATA

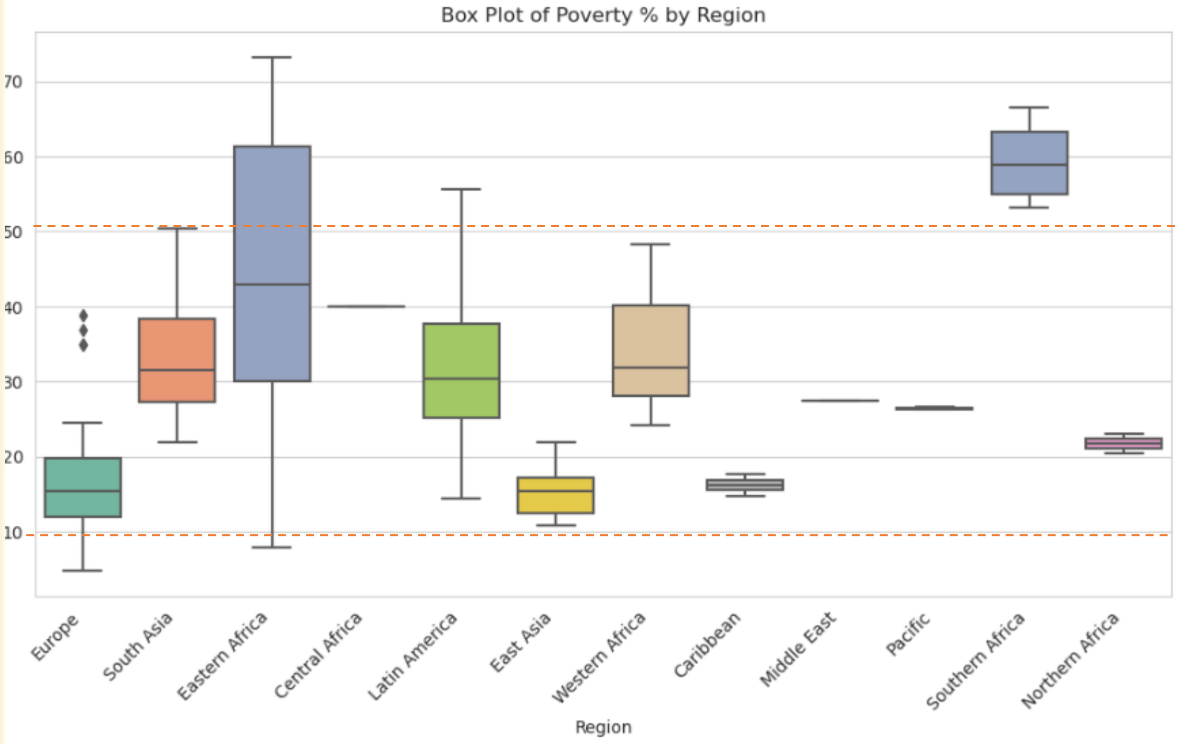
MAGNITUDE BY LOCATION

CHANGE OVER TIME

RELATIONSHIPS BETWEEN VARIABLES

Link to Visualization Jupyter Notebook

Visualization – Box & Bubble Plot of Poverty by Region

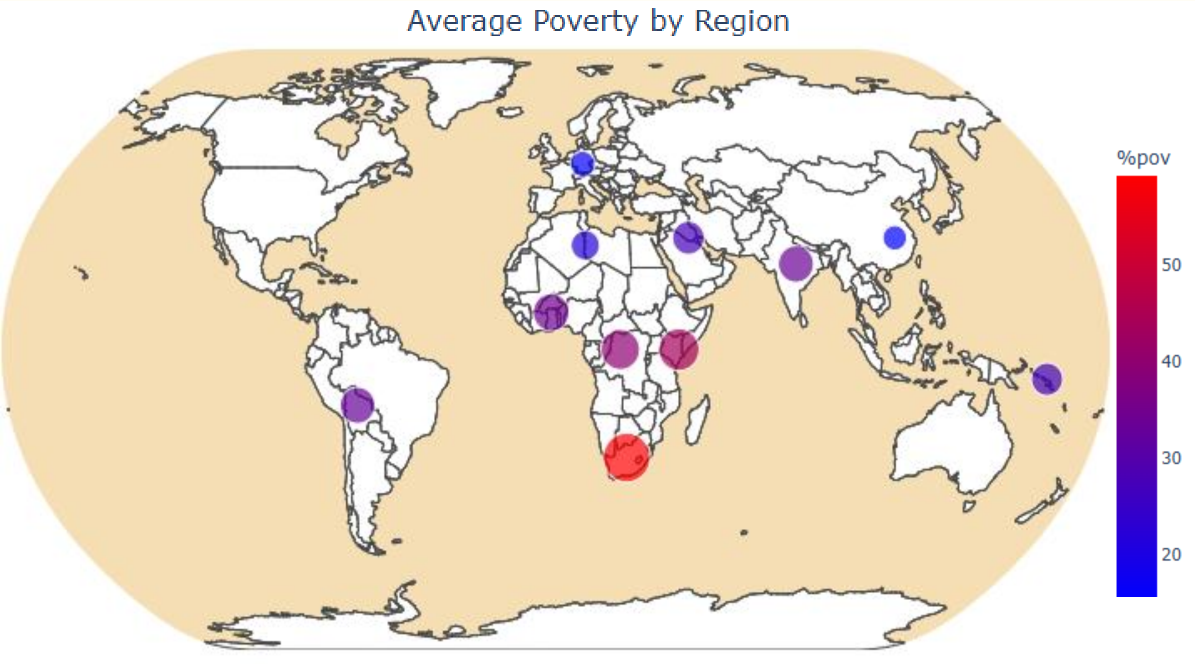


Eastern Africa's distribution spans the full range of overall Poverty distribution

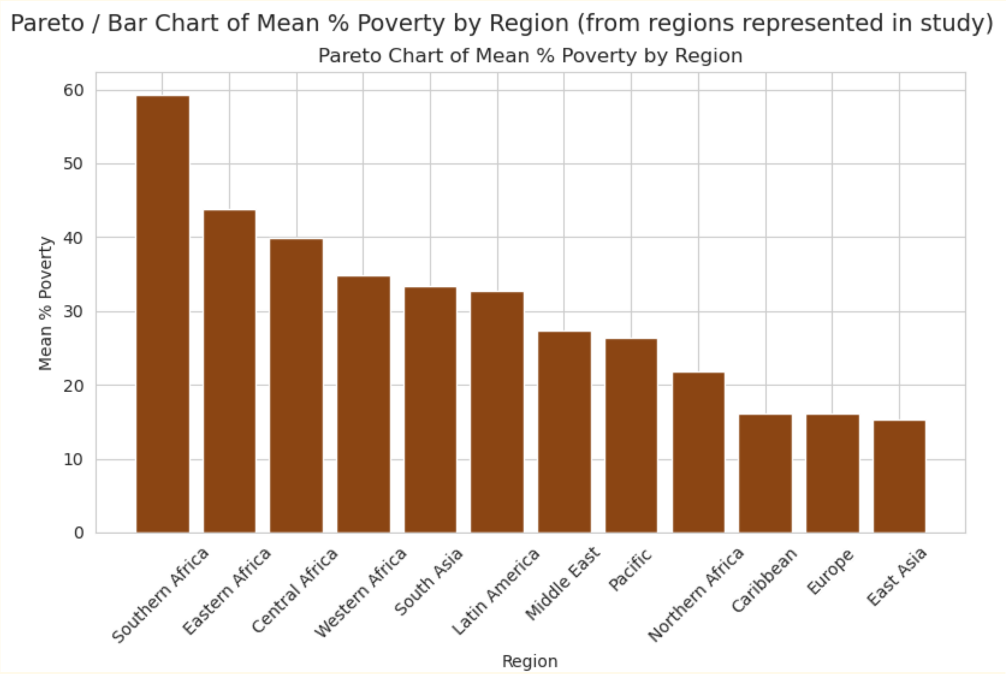
Several regions relatively narrow distributions (SQL query analysis showed many regions with single datapoints)

Most of poverty data appears to be between 10% and 50%

Northern hemisphere countries show lower average poverty



Visualization – Bar and Line Chart of Poverty by Region

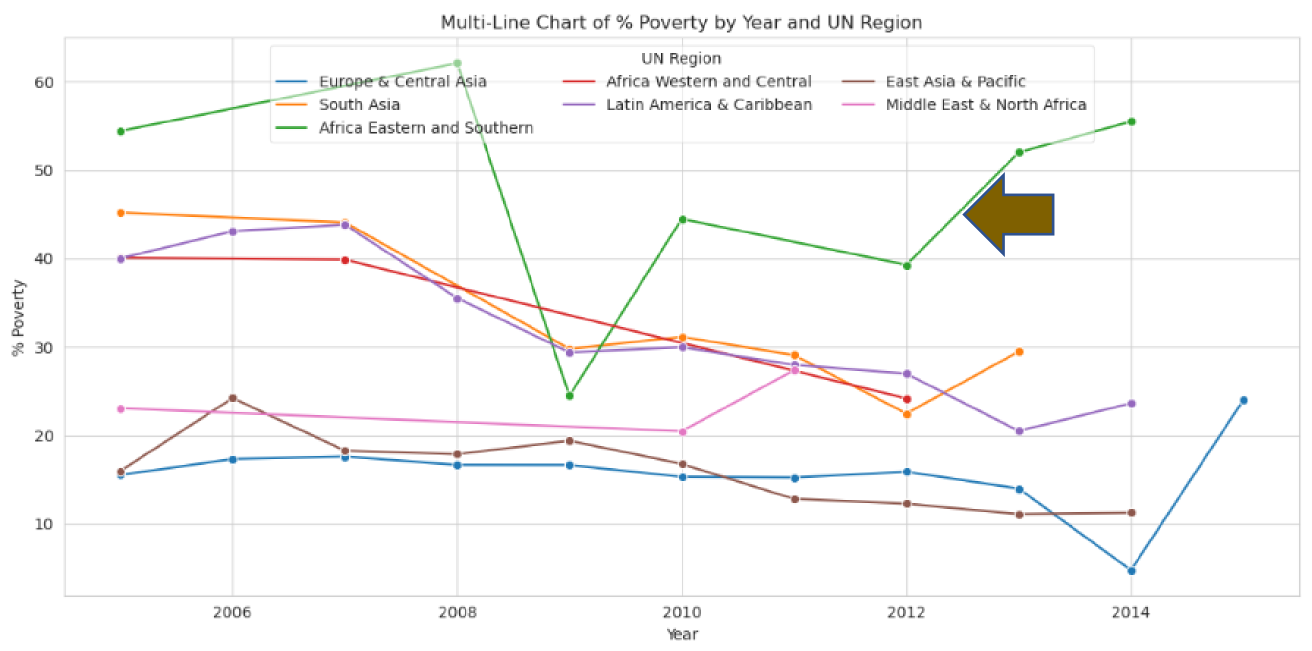


The top 4 regions with the highest mean poverty are in Africa

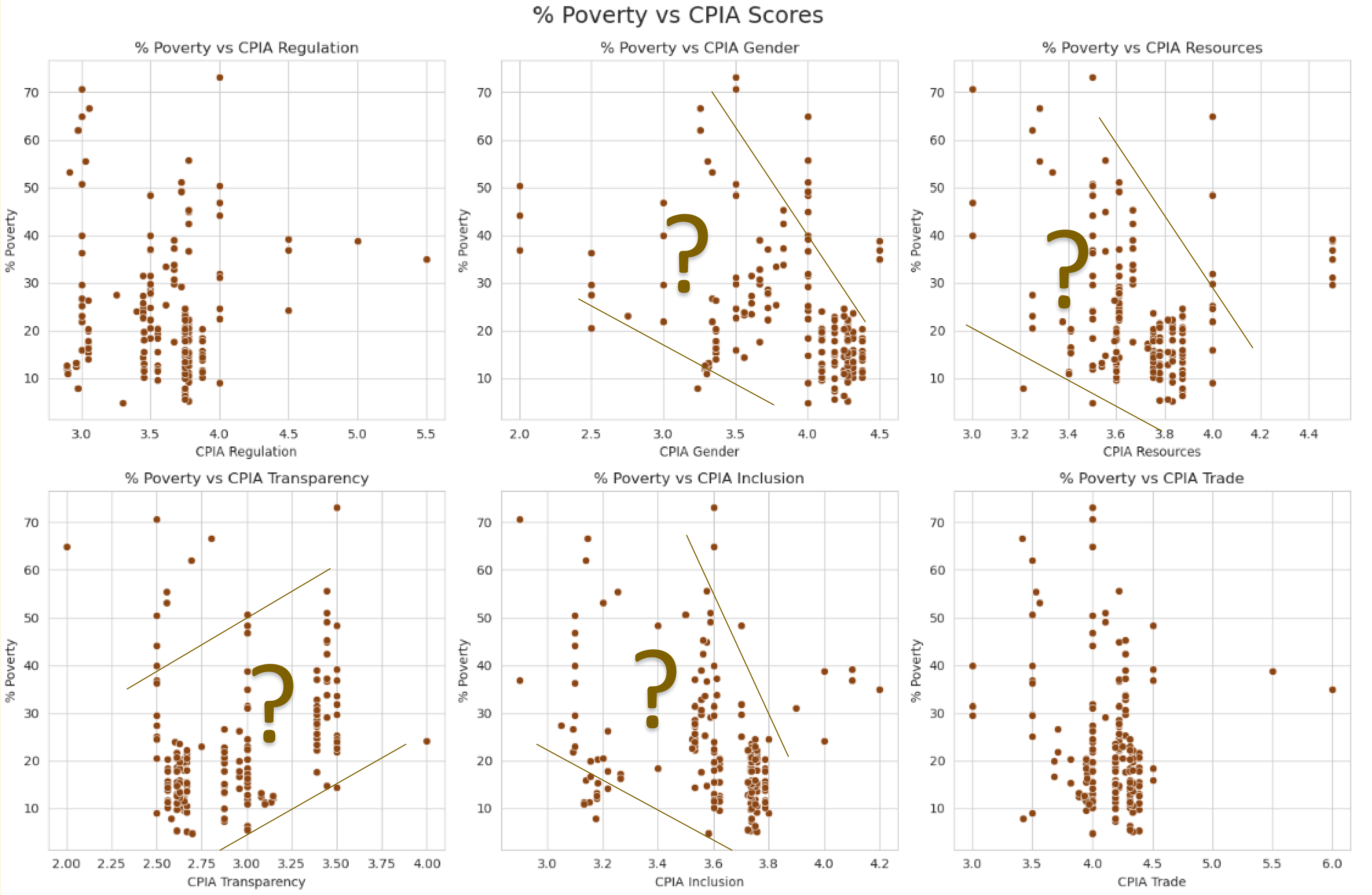
Europe and East Asia have the two lowest mean poverty rates

Most region's poverty rates appear to decrease over study period

East & Southern Africa show large swing in 2011, then steady return to previous highs



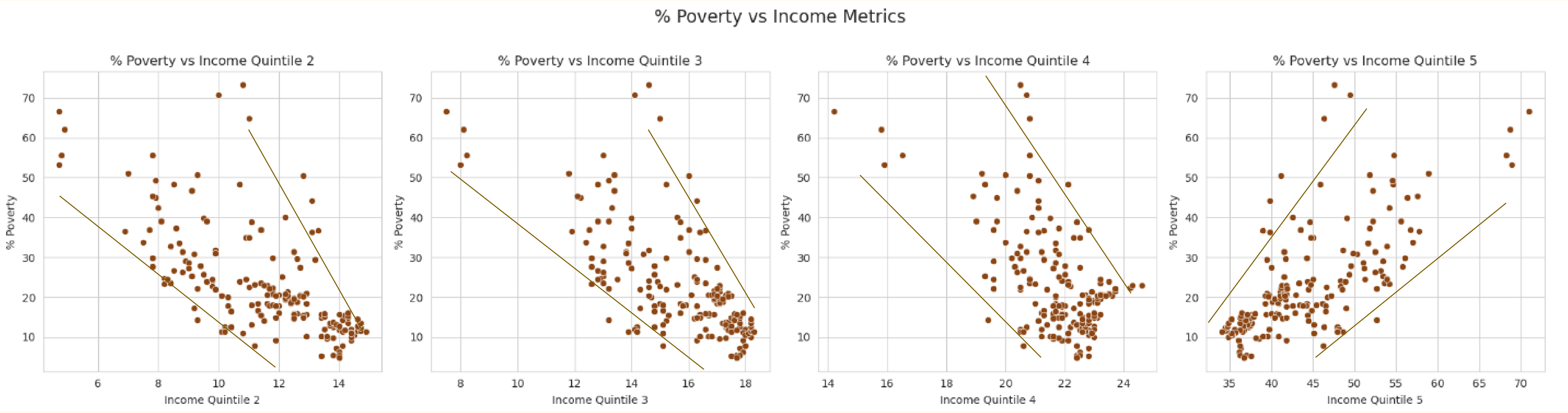
Visualization – Scatter Plot of Poverty vs CPIA Scores

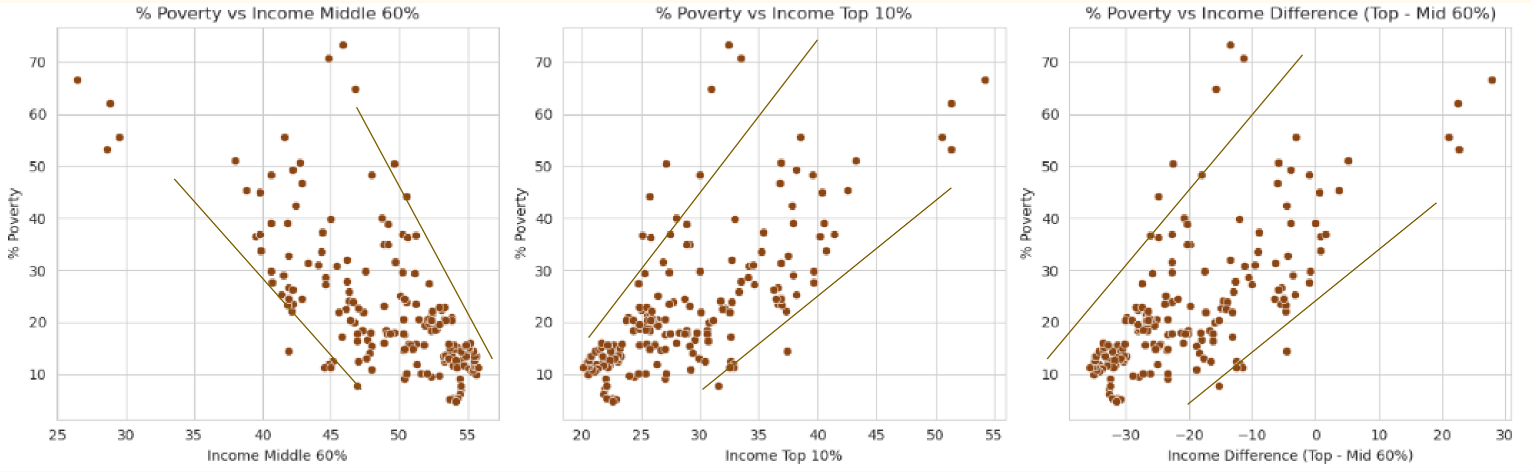


All plots appear similar in shape

Relationships appear mild if at all

Visualization – Scatter Plot of Poverty vs Income



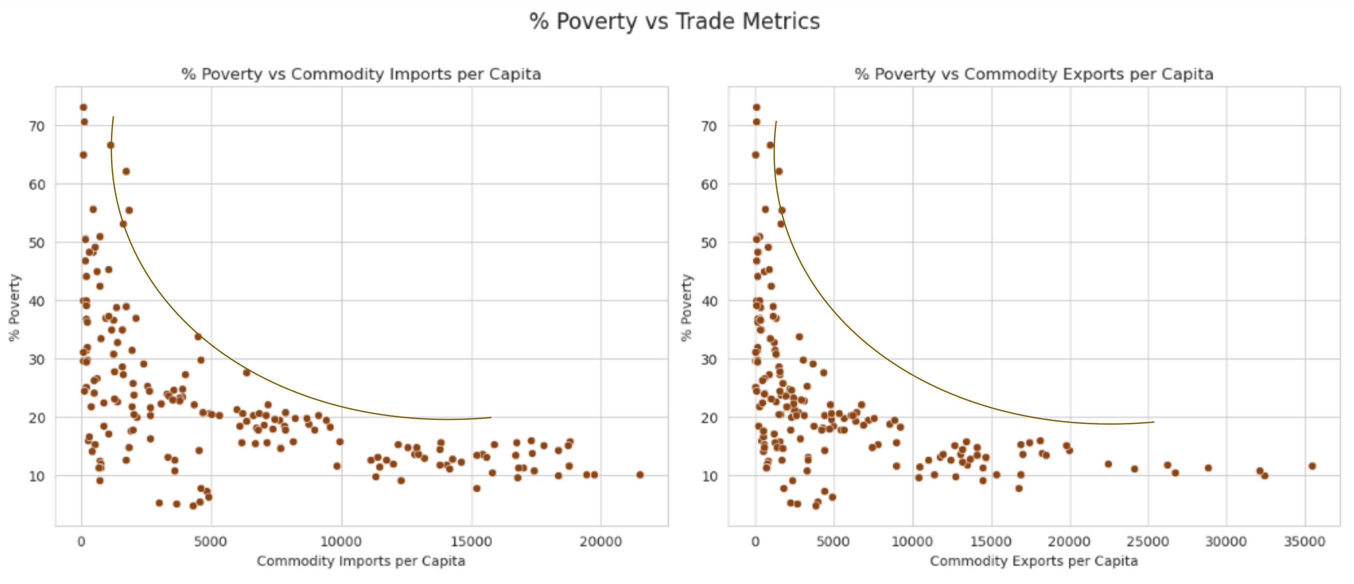


All appear to have moderate correlation

3 positive, 4 negative

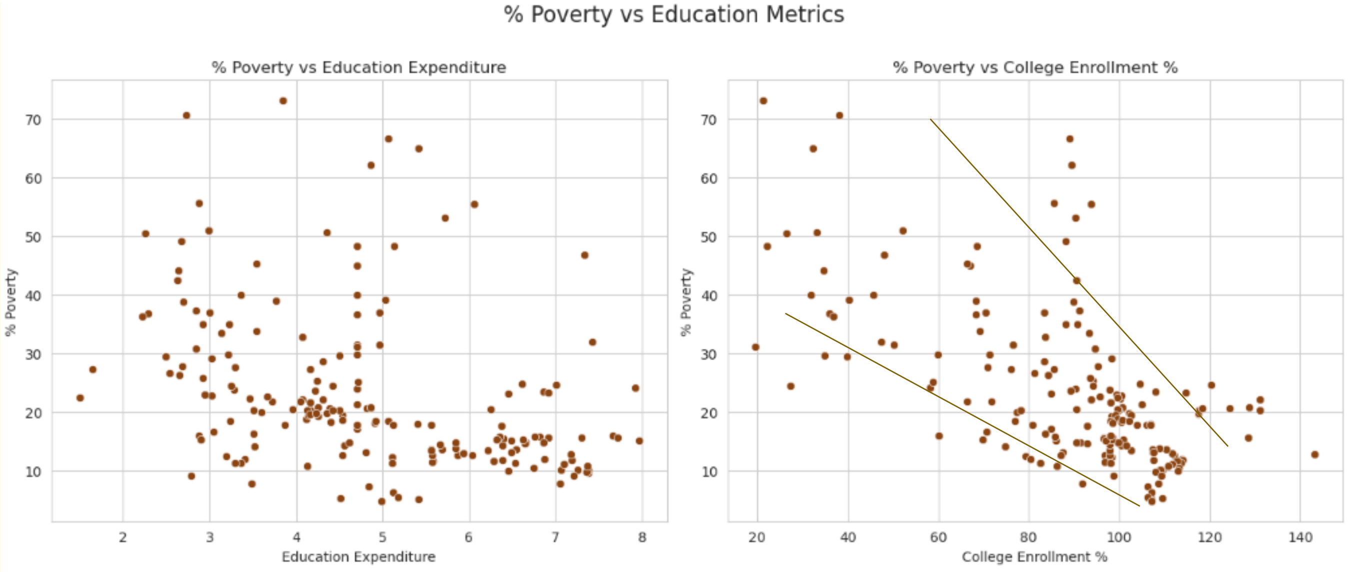
Also appears to be collinearity between them

Visualization – Scatter Plot of Poverty vs Trade & Education



Non-linear moderate negative relationship to poverty

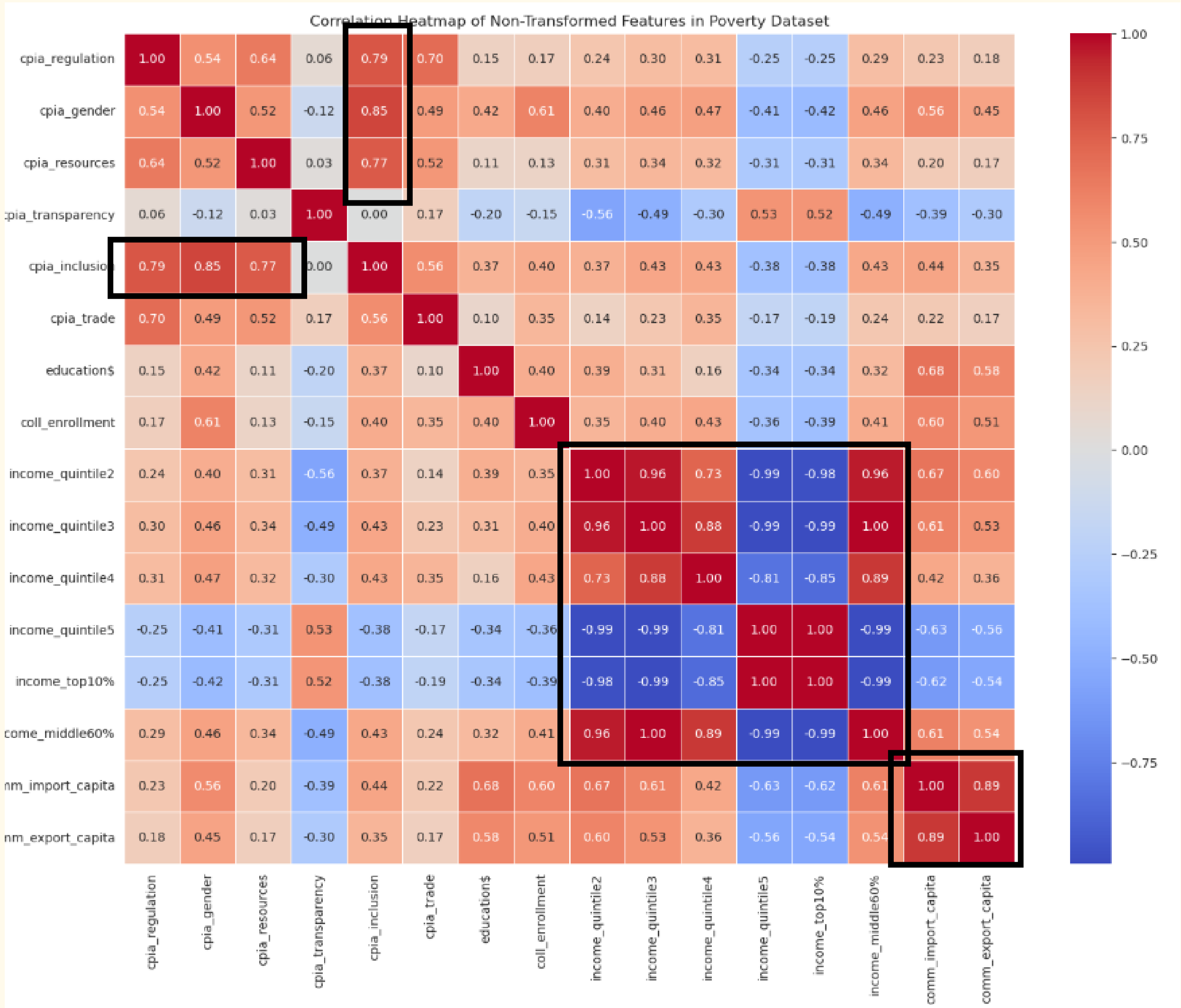
Imports and Export plot looks identical – suspect collinearity



College Enrollment appears to have moderate negative relationship

Education expenditure is not as clear

Visualization – Heatmap of Poverty Feature Collinearity



Darker the color (correlation score closer to 1 or -1), more likely features are correlated to each other

All income related features correlated

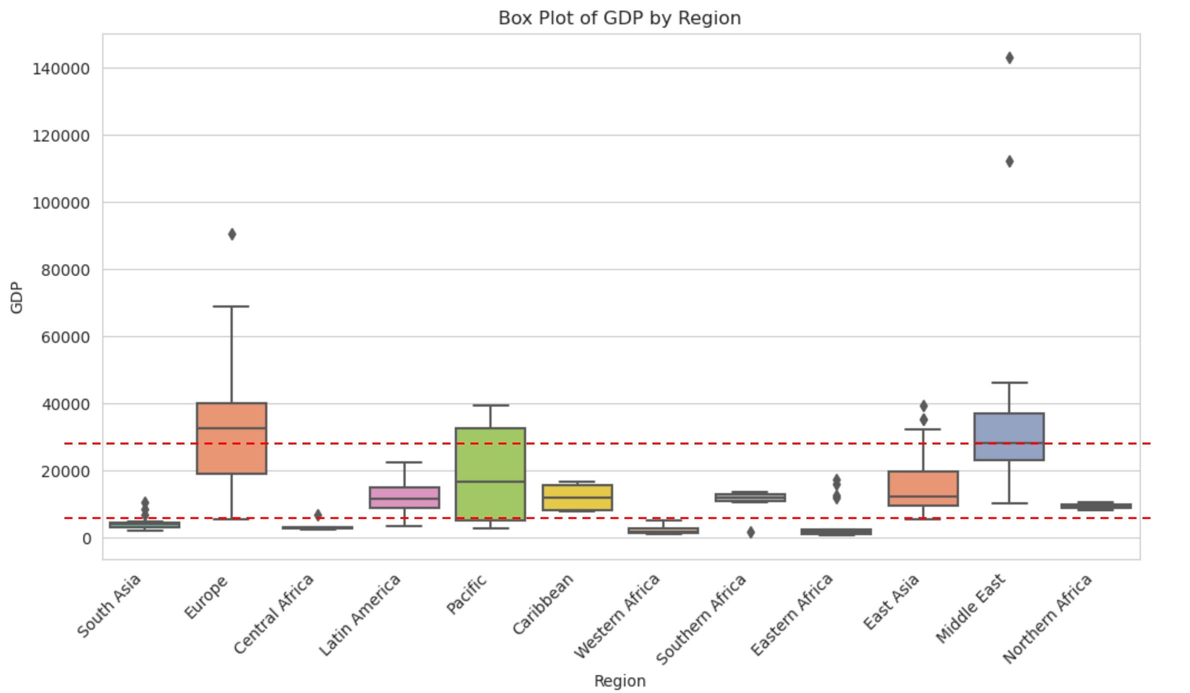
Import, Exports correlated

CPIA Inclusion correlated to 3 of 5 CPIA features

Actions:

- Drop all but middle-class feature as representative of income share
- Drop import feature
- Drop CPIA inclusion feature

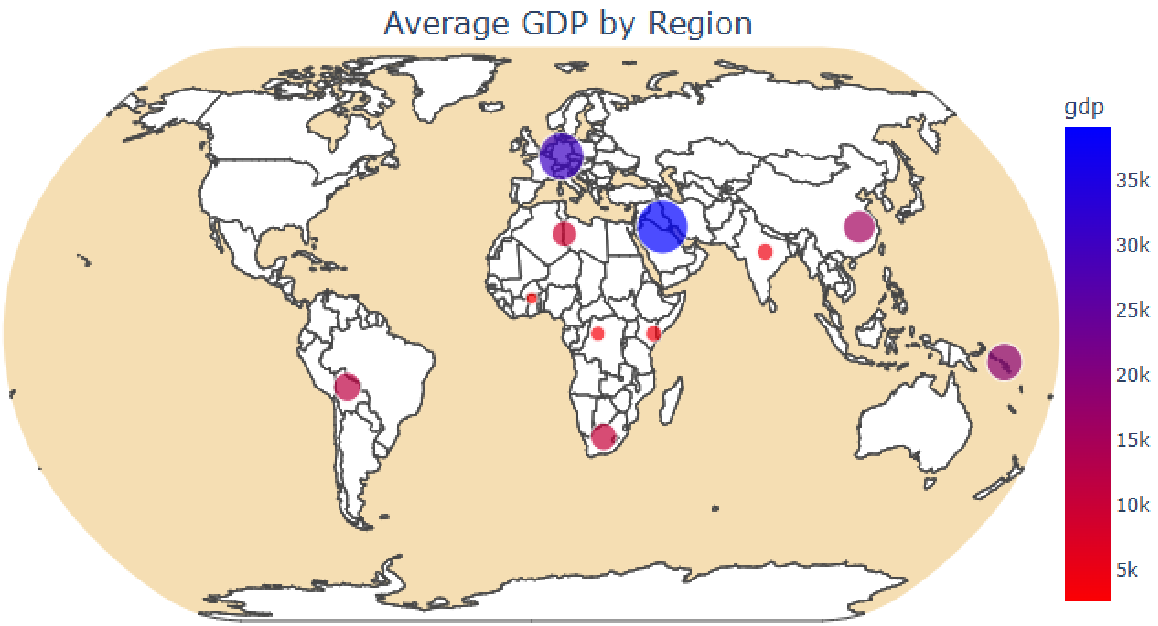
Visualization – Box & Bubble Plot of GDP by Region



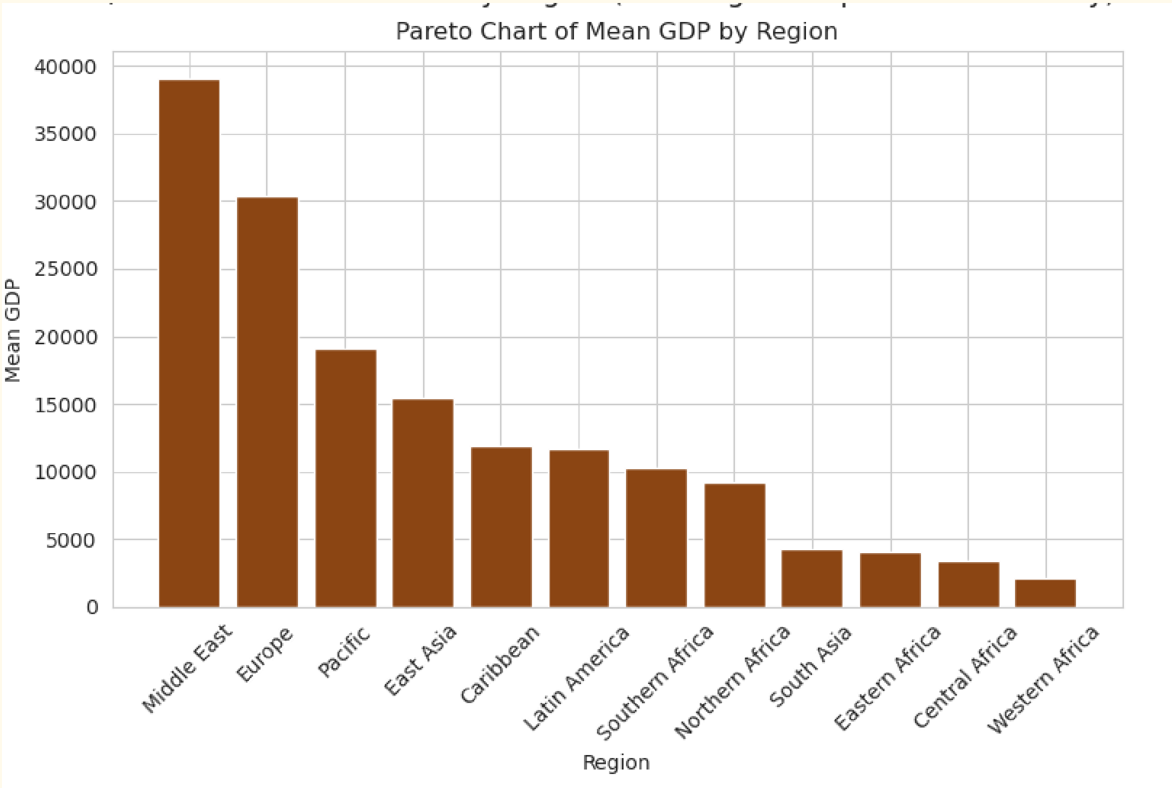
Several regional distributions with outliers

Most of data appear to be within the $5k – $30k range

Appears to be a larger disparity in Average GDP extremes with less transition from high to low than observed for Poverty



Visualization – Bar & Line chart of GDP by Region

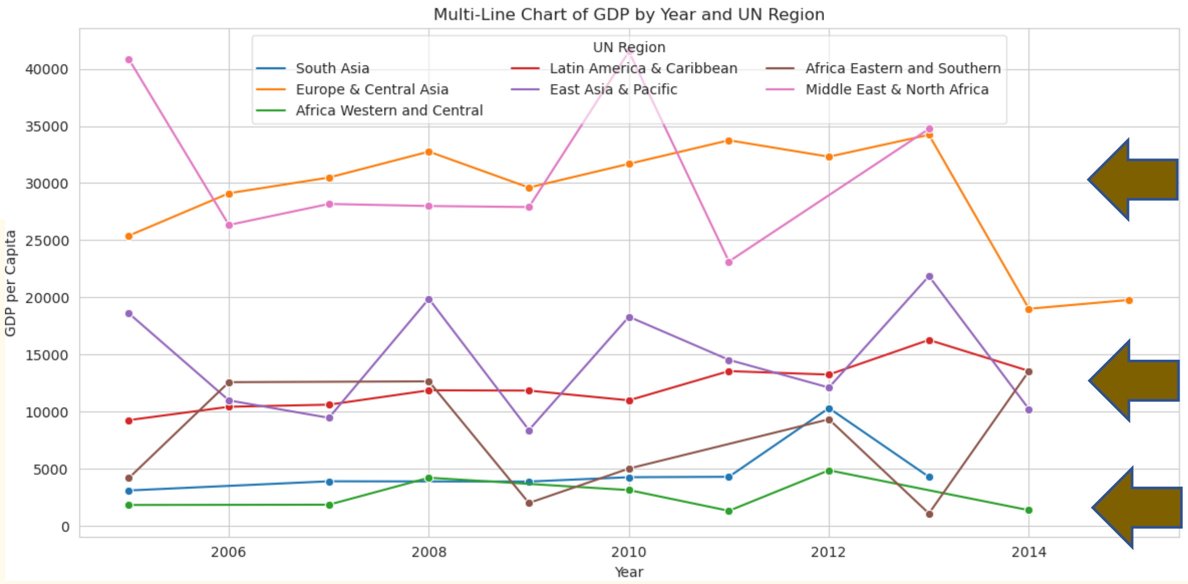


Expected inverse pattern in regions compared to mean Poverty

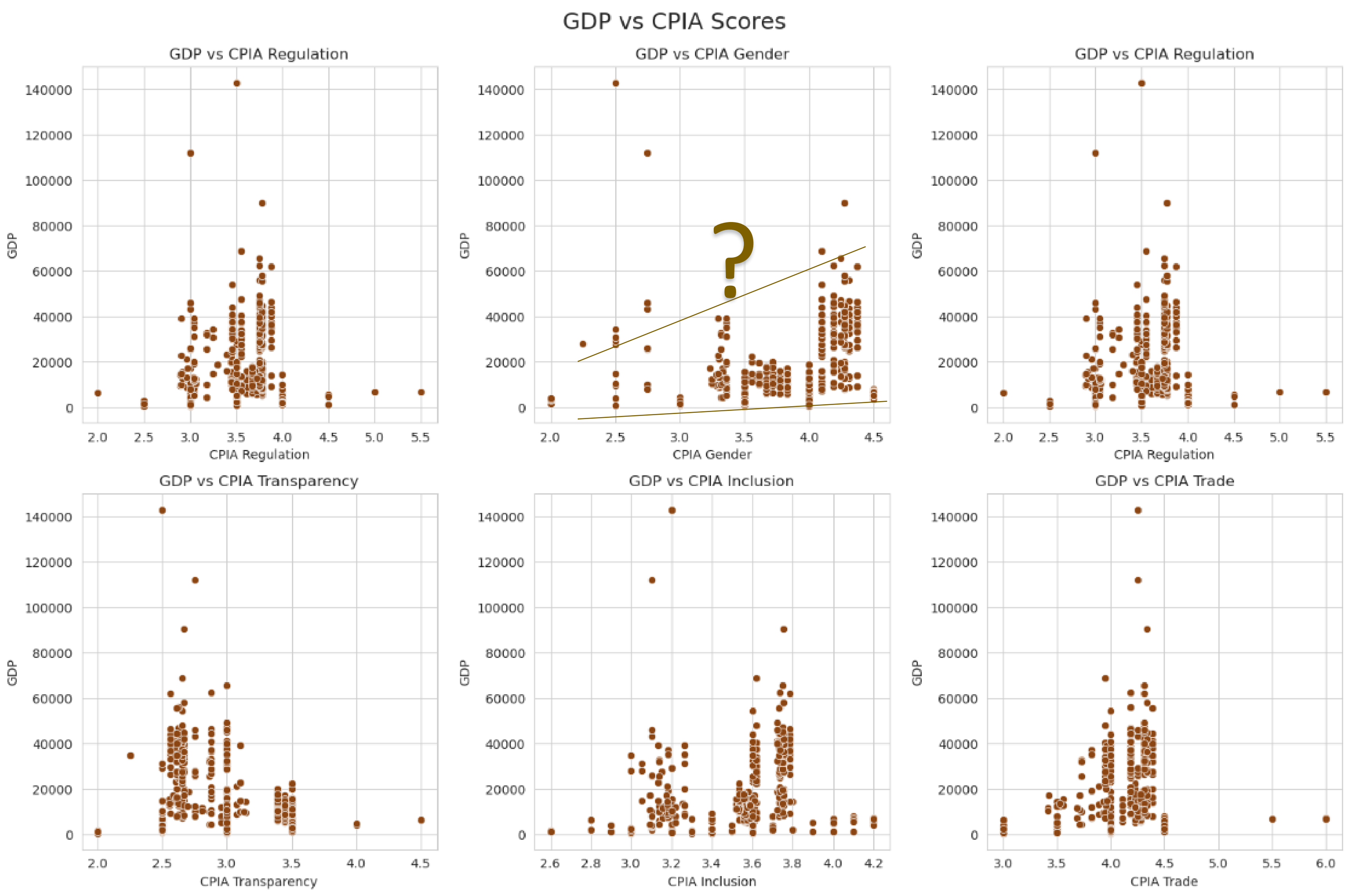
Three clusters of trend lines:

1. Europe with Middle East/North Africa
2. Latin America with East Asia/Pacific
3. West/Central Africa and South Asia

- East/South Africa moves b/w the bottom two



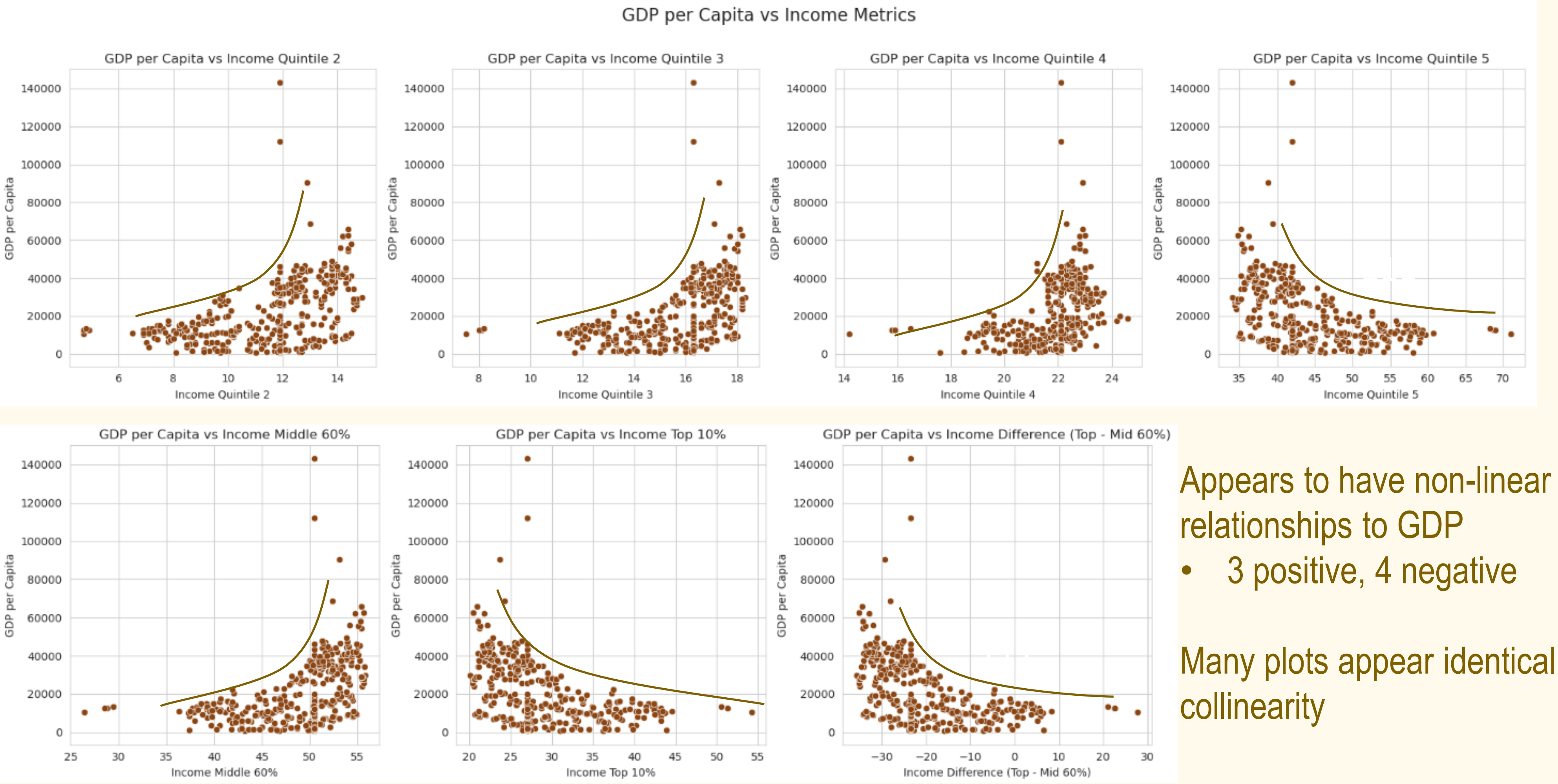
Visualization – Scatter Plot GDP vs CPIA Scores



Most of plots appear centrally oriented like a histogram

Gender – mild positive relationship?

Visualization – Scatter Plot of GDP vs Income

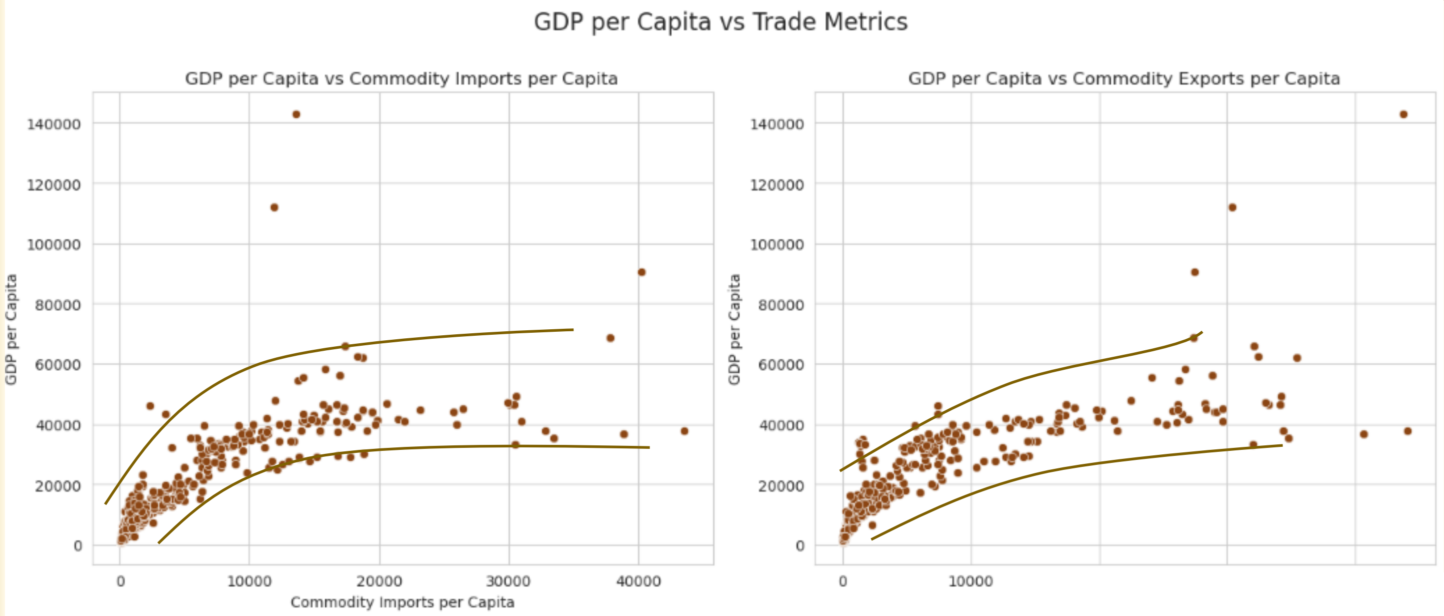


Appears to have non-linear moderate relationships to GDP

- 3 positive, 4 negative

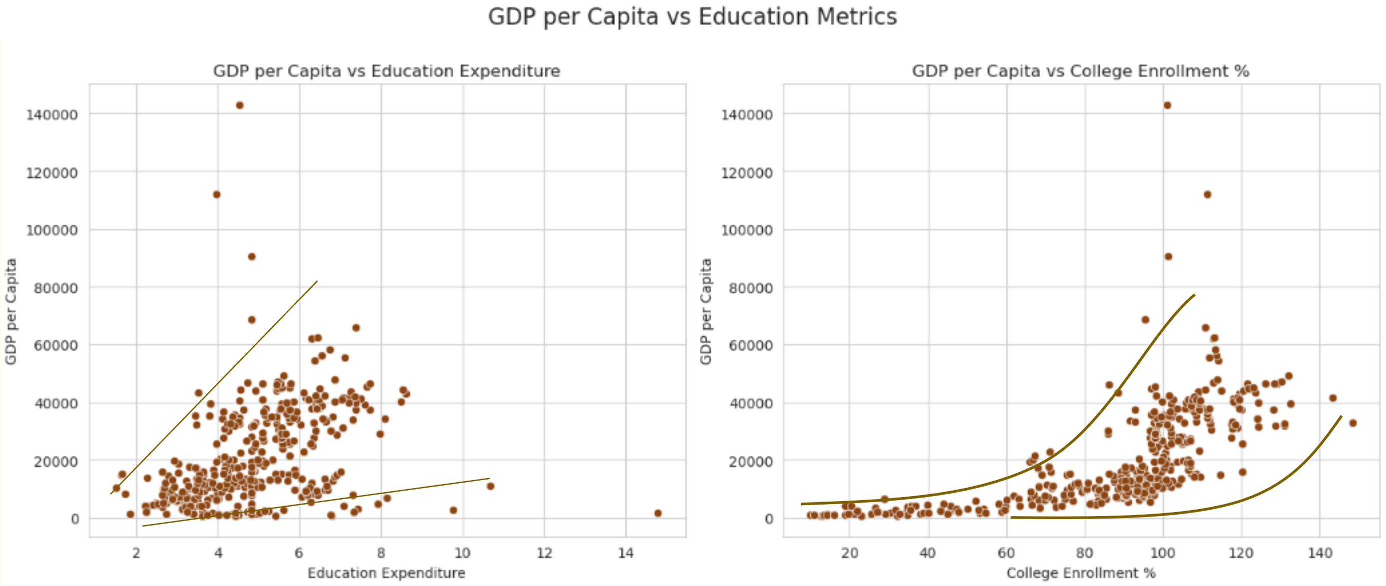
Many plots appear identical – suspect collinearity

Visualization – Scatter Plot of GDP vs Trade & Education



Trade features have strongest visual correlation thus far between feature and target.

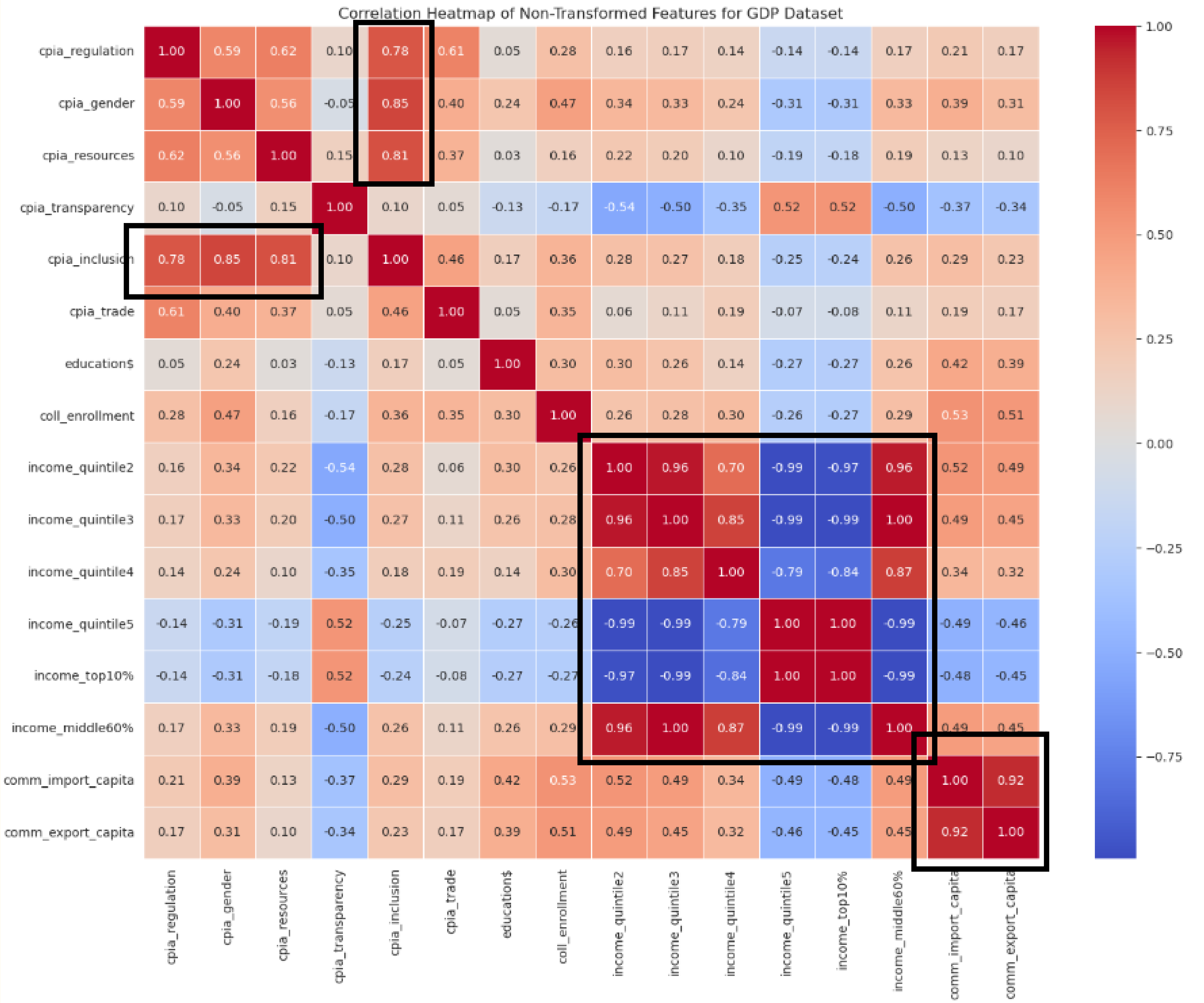
Suspect collinearity here as well



Education expenditures appears to show at least mild correlation to GDP

College enrollment appears almost as strongly correlated to GDP as the trade features

Visualization – Heatmap of GDP Feature Collinearity



Same collinearity as seen between Poverty features

Actions for GDP dataframe:

- Drop all but middle-class feature as representative of income share
- Drop import feature
- Drop CPIA inclusion feature

Predictive Analytics

Predictive Analytics - Preprocessing

1. Separate target from features
2. Normalize features to neutralize unit bias
3. Create categorical target data
   a) Flexibility to use discriminant models also
4. Split datasets into train and test sets

1

```
        gdp  cpia_regulation  cpia_gender  cpia_resources
0   1771.20         2.500000          2.0        3.000000
1   5865.29         3.500000          4.0        3.500000
2   6586.47         2.000000          3.5        2.500000
3  13513.67         3.777778          4.0        3.555556
4  14896.73         3.722222          4.0        3.611111
```

```
        gdp
0   1771.20
1   5865.29
2   6586.47
3  13513.67
4  14896.73
```

```
   cpia_regulation  cpia_gender  cpia_resources
0         2.500000          2.0        3.000000
1         3.500000          4.0        3.500000
2         2.000000          3.5        2.500000
3         3.777778          4.0        3.555556
4         3.722222          4.0        3.611111
```

2

Feature	Value
cpia_gender	2.5
income_middle60%	29.5

$$z = \frac{x - \mu}{\sigma}$$

Feature	Value
cpia_gender	4.7
income_middle60%	5.8

3

```
        gdp
0   1771.20
1   5865.29
2   6586.47
3  13513.67
4  14896.73
```

```
y_t_cat3 distribution:
 1    154
2    151
0     85
Name: gdp, dtype: int64
```

4

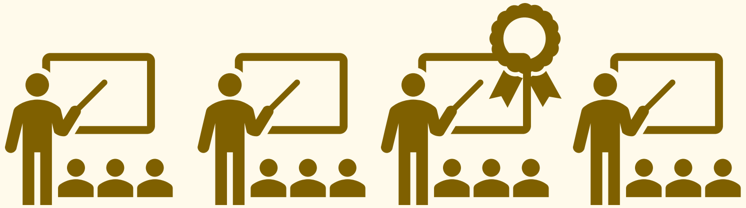
index	y_target	X_gender	X_trade	X_income
0	11705.2	0.1	0.0	-0.9
1	12367.1	0.5	1.2	-1.6
2	13833.7	-0.2	-1.2	0.5
3	14506.2	-2.7	1.5	0.3
4	10728.4	0.7	-0.4	-0.3
5	15035.3	-0.4	0.0	0.8
6	7408.7	2.0	0.9	1.3
7	13178.2	0.9	-0.8	-2.0
8	10063.4	-1.0	-0.3	-0.8
9	10902.4	-0.3	0.6	0.7
10	11393.5	0.5	-0.9	1.8
11	16725.9	-0.3	-1.3	-0.7
12	13382.8	-1.3	0.8	1.3
13	11465.8	-1.3	0.4	-1.2
14	11706.4	-0.4	-0.6	-0.4
15	8927.8	-0.3	-1.0	0.2
16	14226.1	1.2	-1.1	-1.3

index	y_target	X_gender	X_trade	X_income
1	12367.1	0.5	1.2	-1.6
2	13833.7	-0.2	-1.2	0.5
3	14506.2	-2.7	1.5	0.3
4	10728.4	0.7	-0.4	-0.3
5	15035.3	-0.4	0.0	0.8
7	13178.2	0.9	-0.8	-2.0
9	10902.4	-0.3	0.6	0.7
10	11393.5	0.5	-0.9	1.8
11	16725.9	-0.3	-1.3	-0.7
12	13382.8	-1.3	0.8	1.3
14	11706.4	-0.4	-0.6	-0.4
15	8927.8	-0.3	-1.0	0.2
16	14226.1	1.2	-1.1	-1.3

index	y_target	X_gender	X_trade	X_income
0	11705.2	0.1	0.0	-0.9
6	7408.7	2.0	0.9	1.3
8	10063.4	-1.0	-0.3	-0.8
13	11465.8	-1.3	0.4	-1.2

Predictive Analytics Approach – Training & Testing

STEPS FOR EACH MODEL

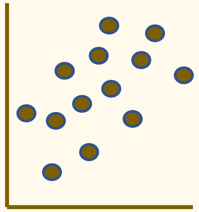


TRAIN AND CROSS VALIDATE TRAINING DATA

PREDICT UNSEEN TEST DATA

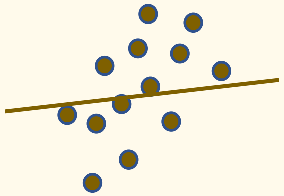
EVALUATE THE MODELS PERFORMANCE

MODELS IN STUDY



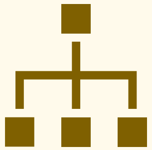
REGRESSION MODELS

- LINEAR / POLYNOMIAL
- SUPPORT VECTOR REGRESSION



DISCRIMINANT MODELS

- LOGISTIC REGRESSION
- SUPPORT VECTOR CLASSIFIER



TREE MODELS

- DECISION TREE
- RANDOM FORREST

Link to Poverty Machine Learning Jupyter Notebook

Link to GDP Machine Learning Jupyter Notebook

Predictive Analytics - Model Evaluation

The following metrics will be evaluated for each model

Regression Models (Continuous data)

- **R^2** - Tells us how much of the change in the outcome (target) is explained by the changes in the input (feature) variables

Discriminant Models (Classification data)

- **Accuracy** – Tells us the **overall correctness** of the model (how often it gets the prediction right)
- **Precision** – Tells us how well the model **predicts the positive class** (avoids model concluding there's an impact when in fact there is not)
- **Recall** – Tells us how well the model **identifies the actual positive class** (avoids model concluding there is no impact when in fact there is)
- **F1 Score** – Tells us the **balance between** Precision and Recall (harmonic mean)
- **AUC** – Tells us how well the model **separates the classes** (overall discriminative power. Helps when comparing different models)

Given the client's objective they prioritize Precision, followed by Accuracy

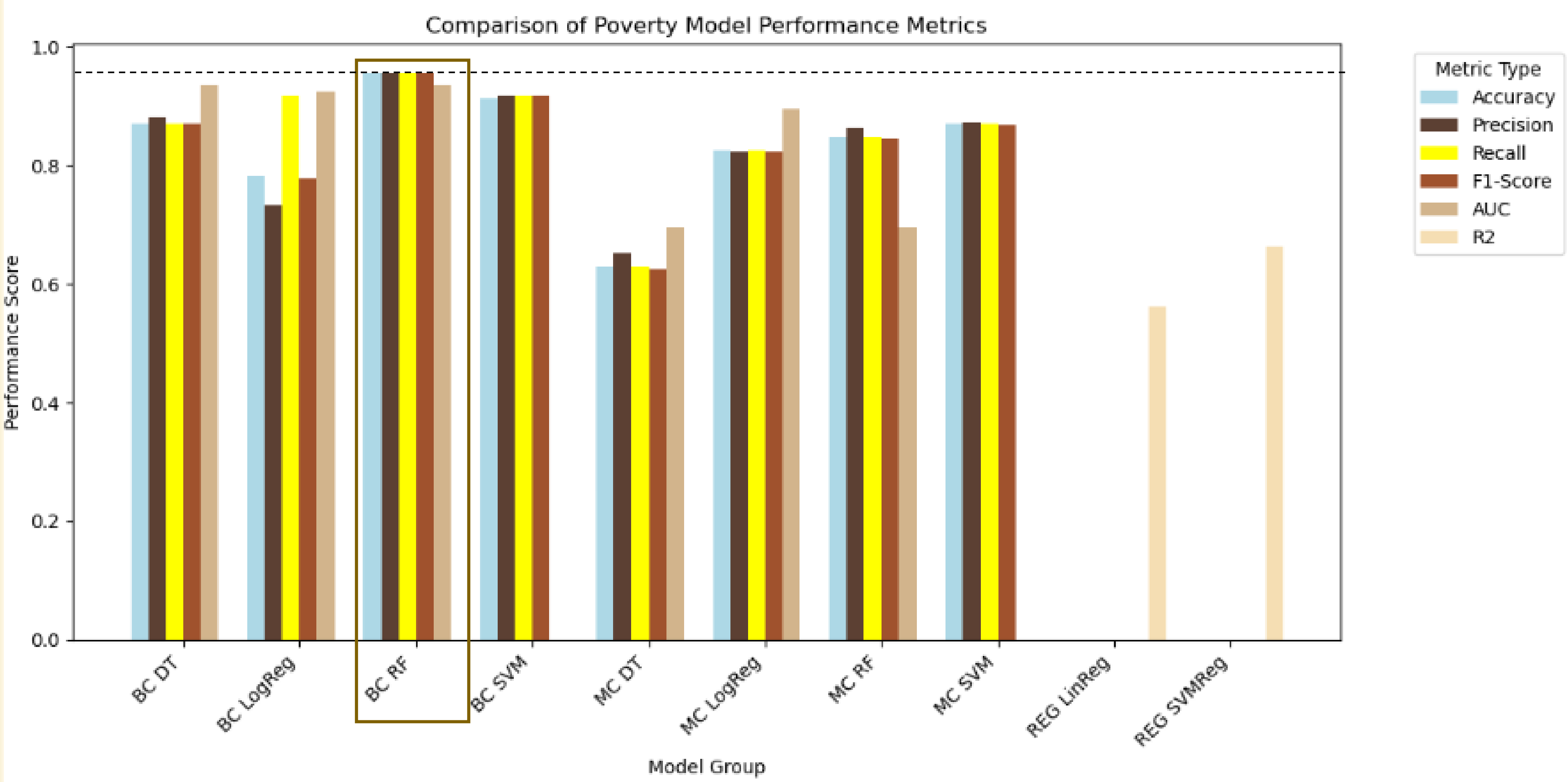
Regression Example

```
Best Polynomial Degree: 1
Best R² Score: 0.5614705537566833
```

Classification Example

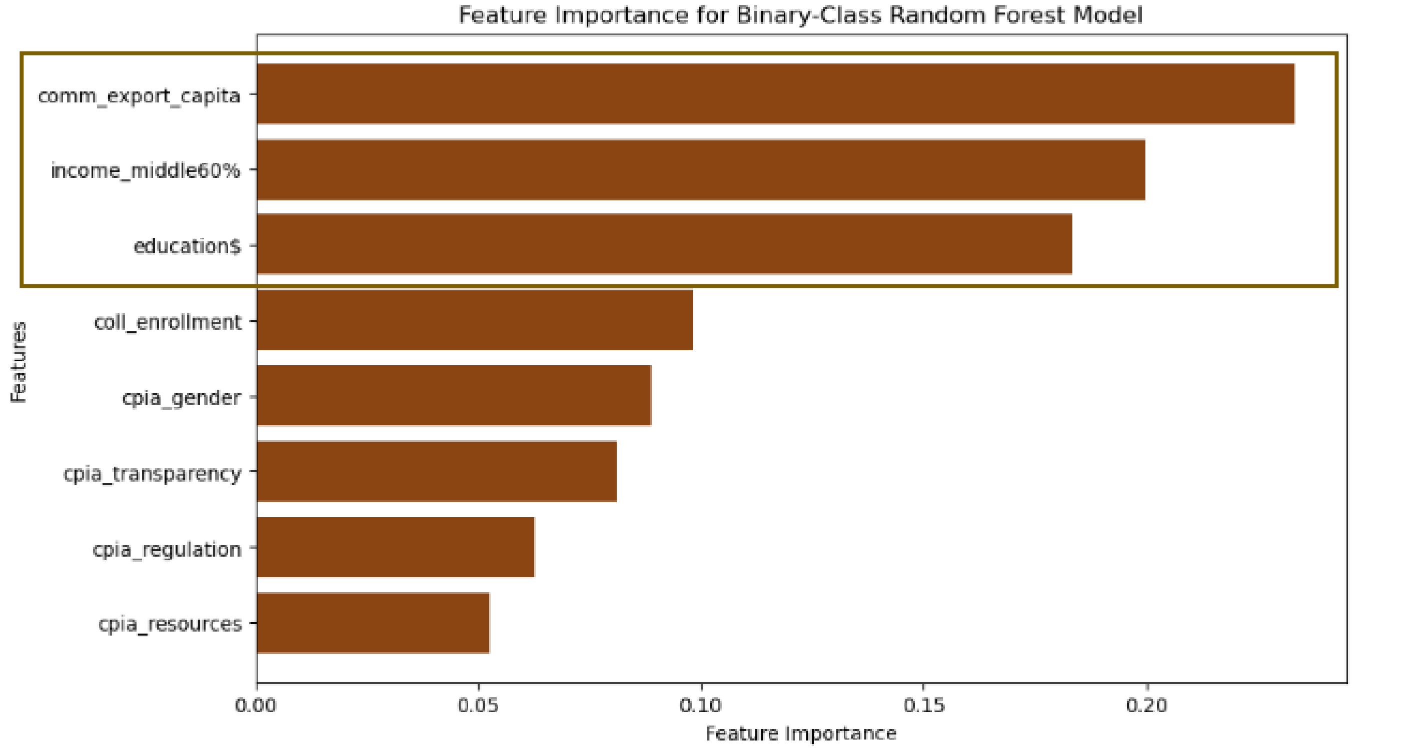
```
Accuracy: 0.8261
Precision: 0.8220
Recall: 0.8261
F1 score: 0.8233
AUC: 0.8952
```

Predictive Analytics – Poverty Model Comparison



Binary-Class Random Forrest is the best performing model for predicting Poverty, with highest Precision and Accuracy

Predictive Analytics – Poverty Feature Importance

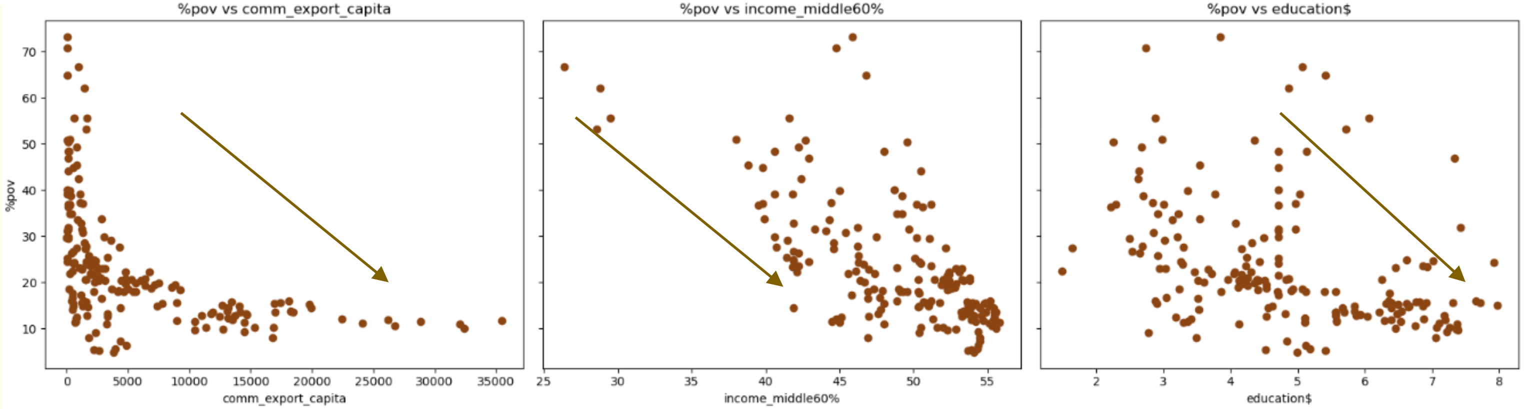


The most important features from the best performing model are commodity exports, income to the middle class, education expenditures

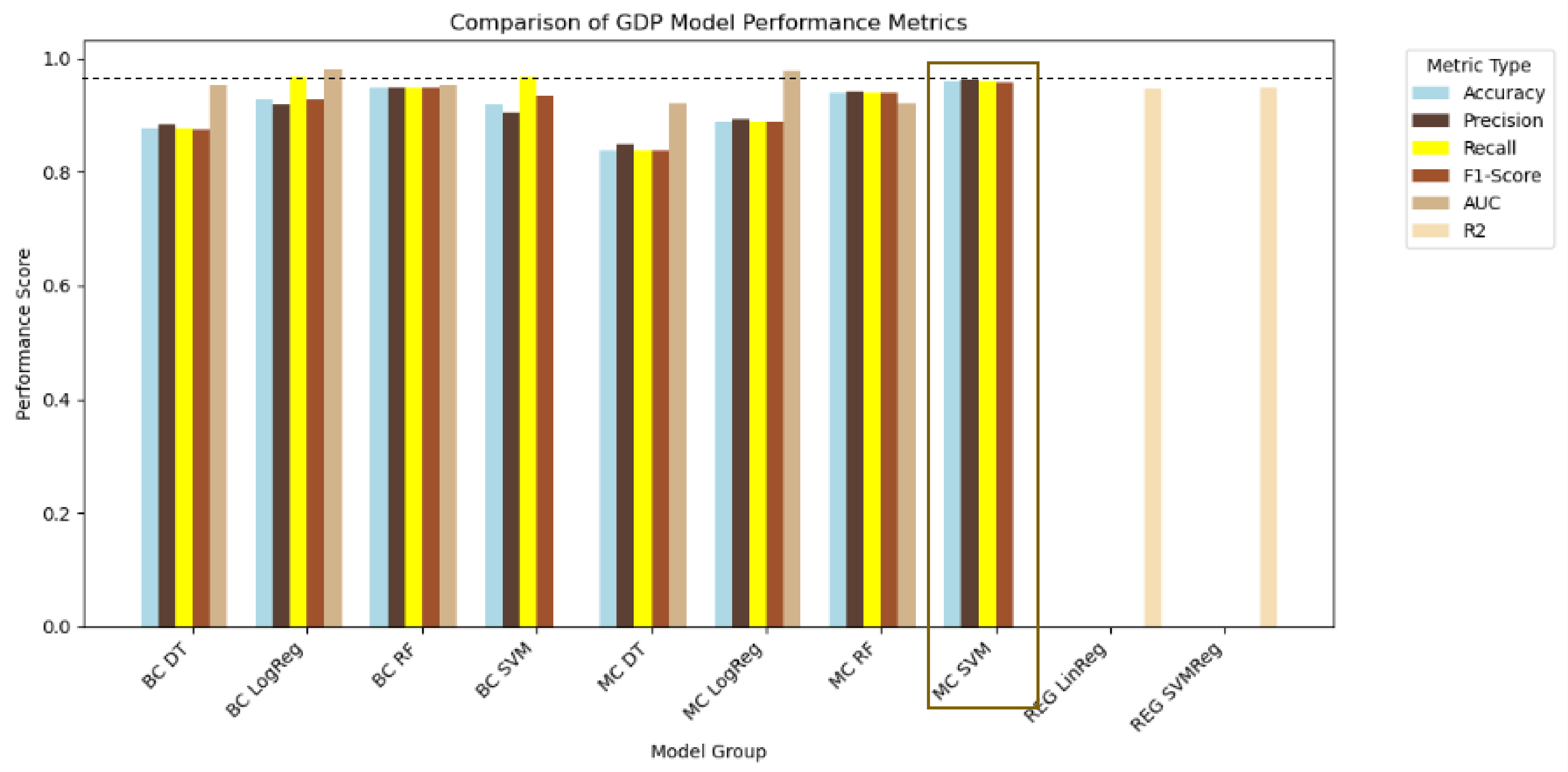
- Step-change difference b/w those features and the remaining

Shows inverse relationship

- As input (feature) variable increases, Poverty (target) value goes down

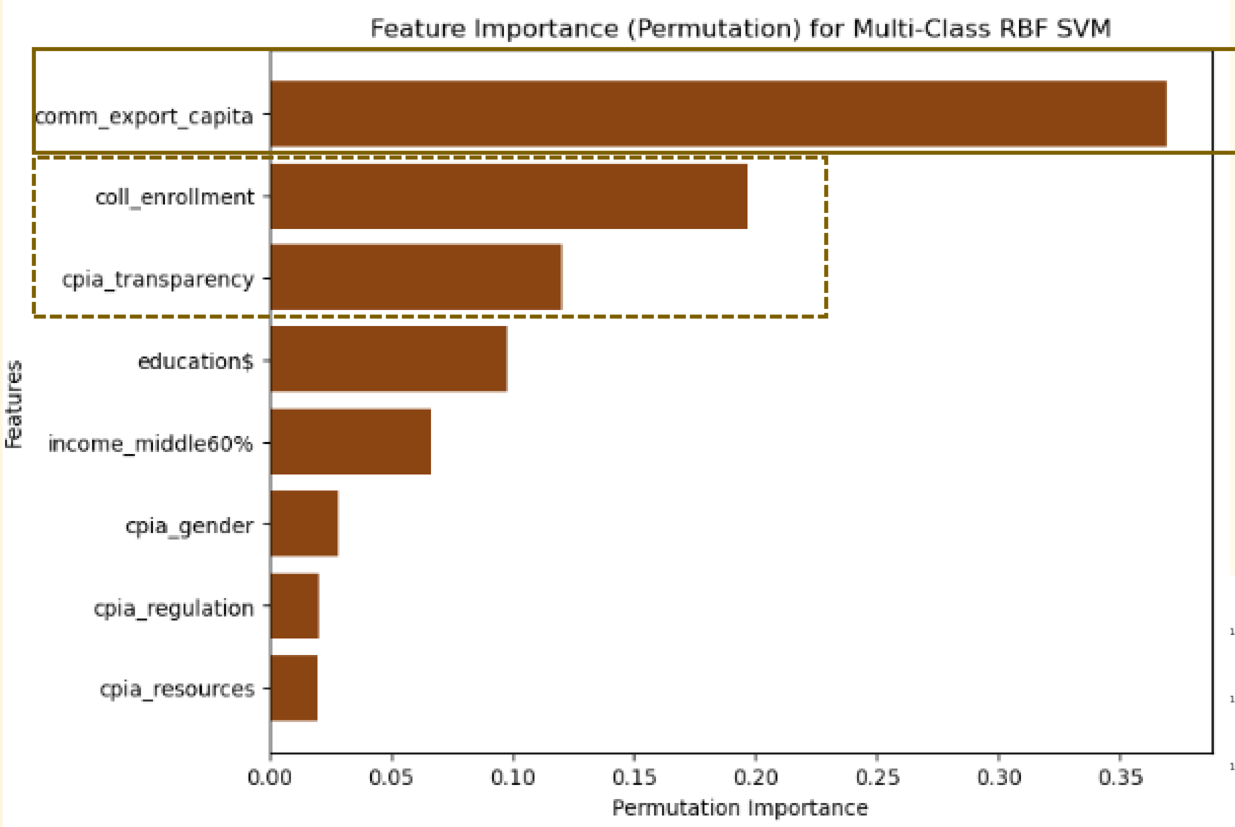


Predictive Analytics – GDP Model Comparison



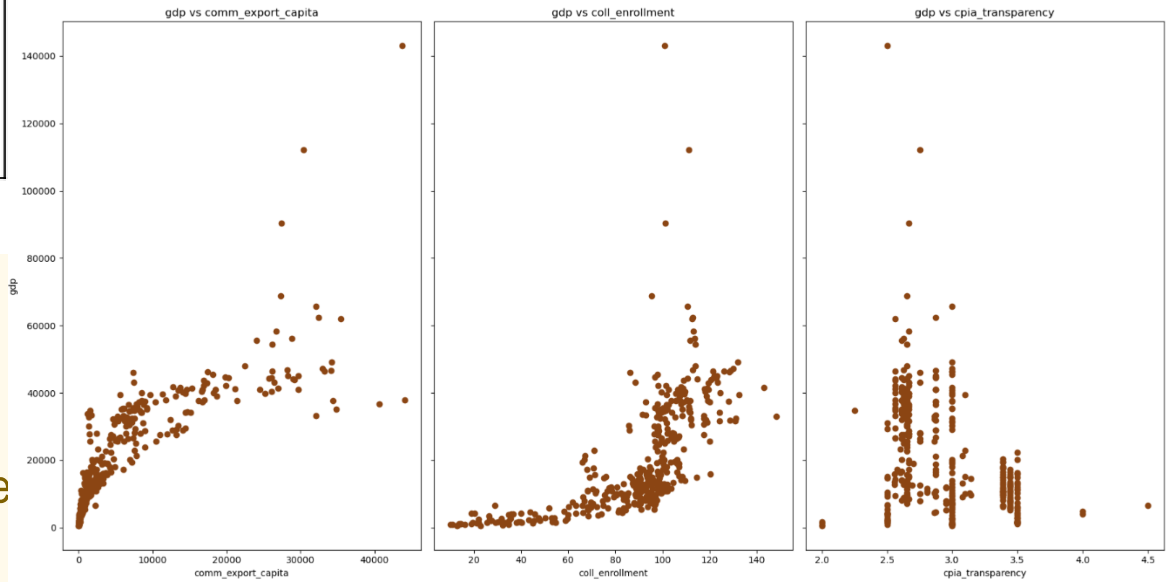
Multi-Class SVM Classifier is the best performing model for predicting GDP with highest Precision and Accuracy

Predictive Analytics – GDP Feature Importance



The most important features from the best performing model are commodity exports, followed by college enrollment and government transparency

- Step-change difference trade and other inputs



Direct relationship with trade and college enrollment

Transparency relationship not as clear – appears to imply a range that's good enough (afterwards law of diminishing returns)

Results & Conclusion

Results from Poverty Data

- Data Imputation and Coverage
  - 82% of records dropped enabling imputation
  - Major economies US, Canada, Australia absent from study as a result
- Extreme Poverty Rates
  - **18%** of countries had poverty >30% or <15%
  - Top 15 poorest countries: **avg poverty 47%**
  - Top 15 least poor: **avg poverty 13%**
- Feature Change Comparing Poverty Extremes (Top 15)
  - CPIA score: **No change**
  - Education spending: **Up 23%**
  - College enrollment: **Up 87%**
  - Middle class income: **Up 16%**
  - Top 10% income: **Down 26%**
  - Trade Activity: **Up > 1,000%**
- Feature Correlations
  - All income features correlated
  - Commodity imports >>> exports
  - CPIA inclusion >>> gender, resources, regulation
- Model Performance
  - Binary Random Forest Classifier
    - **Precision and Accuracy 96%**
    - Top predictors:
      - **Commodity trade**
      - **Middle class size**
      - **Education spending**

Results from GDP Data

- Data Imputation and Coverage
  - 91% of records dropped enabling imputation
  - Major economies US, Canada absent from study as a result
- Extreme GDP Rates
  - **21%** of countries had GDP >$10k or <$3k
  - Top 15 wealthiest countries: **avg GDP $49,927**
  - Top 15 least wealthy: **avg GDP $1,384**
- Feature Change Comparing GDP Extremes (Top 15)
  - CPIA score: **Up 3%**
  - Education spending: **Up 17%**
  - College enrollment: **Up 287%**
  - Middle class income: **Up 8%**
  - Top 10% income: **Down 16%**
  - Trade activity: **Up > 10,000%**
- Feature Correlations
  - All income features correlated
  - Commodity imports >>> exports
  - CPIA inclusion >>> gender, resources, regulation
- Model Performance
  - Multi-Class Support Vector Machine Classifier
    - **Precision and Accuracy 96%**
    - Top predictors:
      - **Commodity trade**
      - **College enrollment**
      - **Governmental Transparency & Accountability**

Conclusion

- Models accurately predicted both **poverty**, and **GDP**, offering valuable guidance for **Countr23's policy decisions**
- **Commodity trade** and **education** emerged as the most influential factors in the study for both target outcomes, and **size of the middle class** was most critical to poverty
- **Government transparency** was a significant predictor of GDP, though its non-linear relationship calls for deeper study
- **Collinearity observed** amongst specific trade and income variables – further analysis recommended
- **Healthcare's** impact on the economy could not be determined due to lack of data; additional data collection and analysis are recommended

Thank you!