PROBLEM STATEMENT 1

EXPLORATORY DATA ANALYSIS

PROBLEM STATEMENT :

Do an Exploratory data analysis on the given data sets (country-wise-average.csv and malnutrition-estimates.csv) without using any in build packages and give insights from the data and plot the analysis which you did.

DATASET

The given dataset contains country-wise-average.csv and malnutrition-estimates.csv Country-wise-average dataset consists of a list of countries, severe wasting, wasting, overweight, stunting, underweight, income classification. Where as malnutrition-estimates consists of Income, severe wasting - the skinny situation faced by people , wasting - unable to move parts, Stunting, LIFD - Low income food deficit, LDC - Low Developed Countries, LLDC - Landlocked Developing Countries, SIDS - Small Island Developing States.

OVERVIEW OF THE DATA SET :

country-wise-average has 152 rows and 8 columns. Malnutrition-estimates has 924 rows and 20 columns.

```
       Country  Income Classification  Severe Wasting    Wasting  Overweight  \
0  AFGHANISTAN                    0.0        3.033333  10.350000    5.125000
1      ALBANIA                    2.0        4.075000   7.760000   20.800000
2      ALGERIA                    2.0        2.733333   5.942857   12.833333
3       ANGOLA                    1.0        2.400000   6.933333    2.550000
4    ARGENTINA                    2.0        0.200000   2.150000   11.125000

    Stunting  Underweight  U5 Population ('000s)
0  47.775000    30.375000            4918.561500
1  24.160000     7.700000             232.859800
2  19.571429     7.342857            3565.213143
3  42.633333    23.600000            3980.054000
4  10.025000     2.600000            3613.651750


   Unnamed: 0 ISO code      Country Survey Year  Year  Income Classification  \
0           0      AFG  AFGHANISTAN        1997  1997                      0
1           1      AFG  AFGHANISTAN        2004  2004                      0
2           2      AFG  AFGHANISTAN        2013  2013                      0
```

DATA PREPROCESSING :

- ❖ DEALING WITH MISSING VALUES

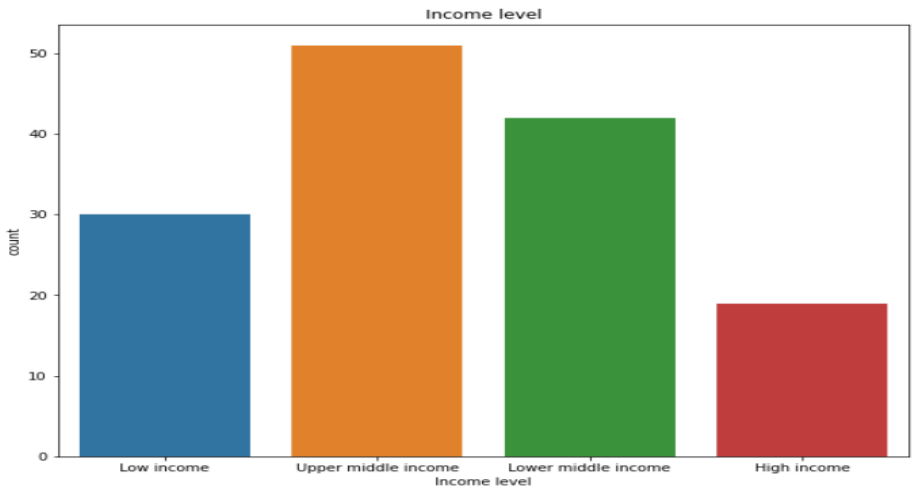
    - The missing values are identified and treated with the median of the attribute.
    - I had removed rows with NaN values if the percentage of NaN values are below 5%
    - For data, we can remove NaN values in wasting, overweight, stunting and underweight. However for severe wasting, I will replace them with median values.
    - For data1, we can remove NaN values in stunting, survey sample (N), notes and underweight. The rest will be replaced with median values.

- ❖ DEALING WITH OUTLIERS

    - Outliers are detected in each attribute and plotted using the box plot.
    - To deal with outliers, we use Z-score.
    - A Z-score is a numerical measurement used in statistics of a value's relationship to the mean (average) of a group of values, measured in terms of standard deviations from the mean.
        - If a Z-score is 0, it indicates that the data point's score is identical to the mean score.
        - A Z-score of 1.0 would indicate a value that is one standard deviation from the mean.
        - Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean.
    - In most of the cases a threshold of 3 or -3 is used i.e if the Z-score value is greater than or less than 3 or -3 respectively, that data point will be identified as outliers.

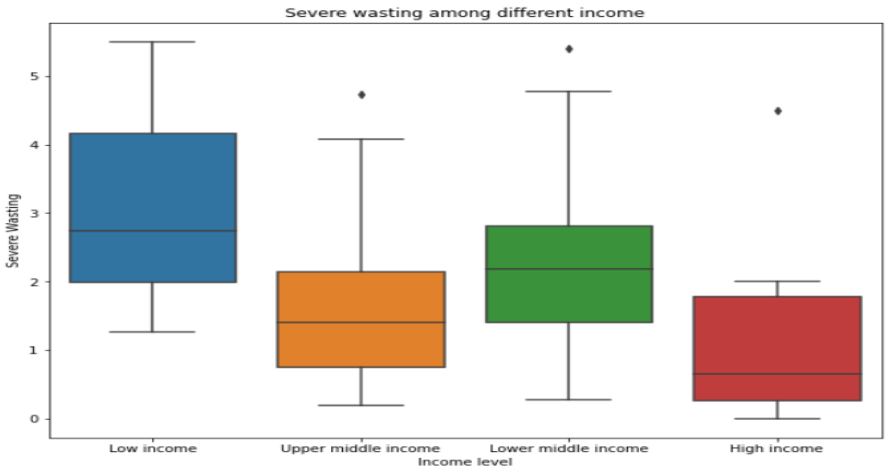
Income level

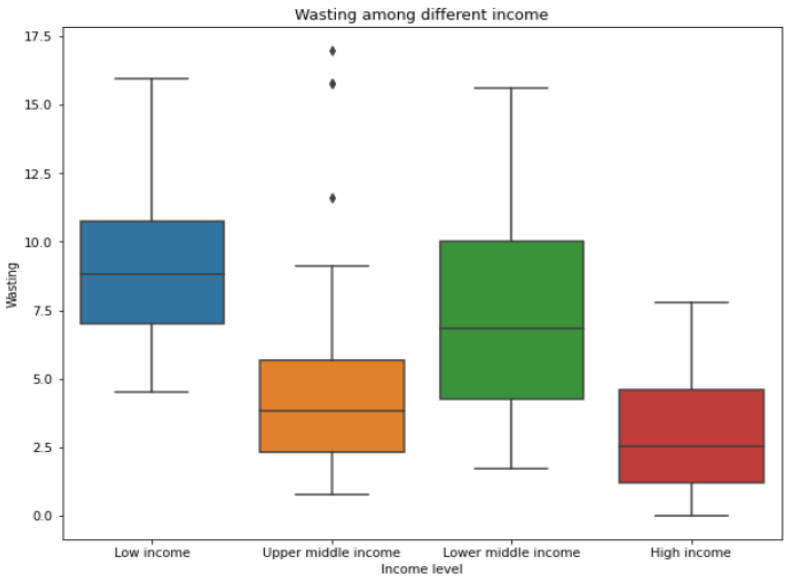
Here, we are converting those numbers in income classification to their real meaning. This allows readers to understand the context without having to constantly go back to the description.



Look into the extend wasting among income level

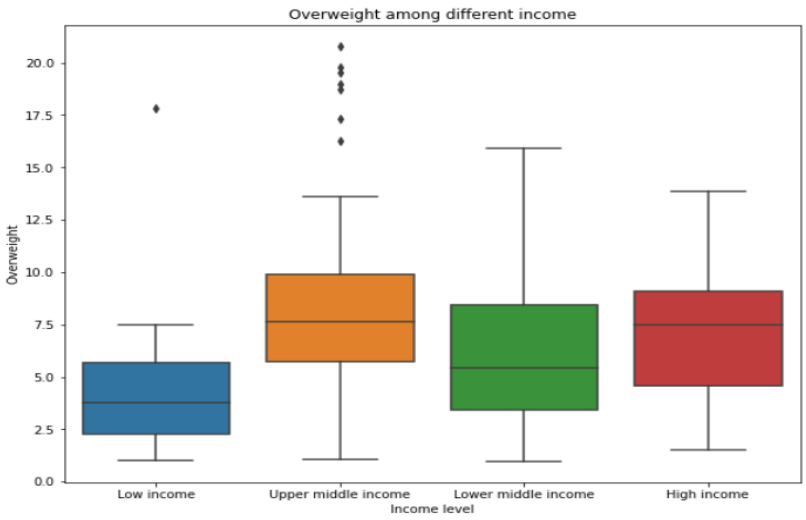
- Wasting refers to the process by which a debilitating disease causes muscle and fat tissue to "waste" away.
- This could be due to the lack of food or proper nutrition.





Low income countries tend to have a higher level of wasting. As we all know, wasting is caused by numerous factors and one of them is low energy intake. People in these countries tend to survive on less than 3 meals and hence suffer from severe malnutrition. Their bodies switch to survival mode and their muscles are striped off to preserve energy.

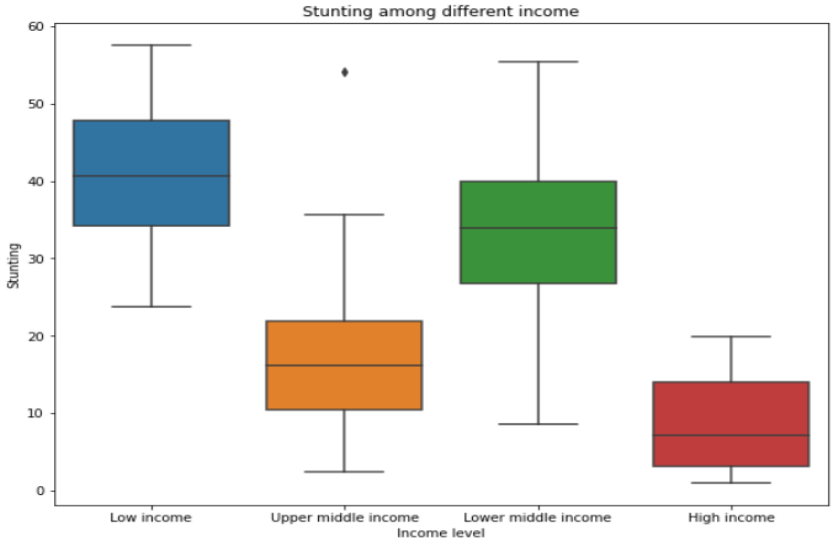
OVERWEIGHT



Unsurprisingly, we see a high level of overweight issues in high income and upper middle income countries. As more people are able to afford their meals in these countries, we will tend to see a greater level of overweight issues.

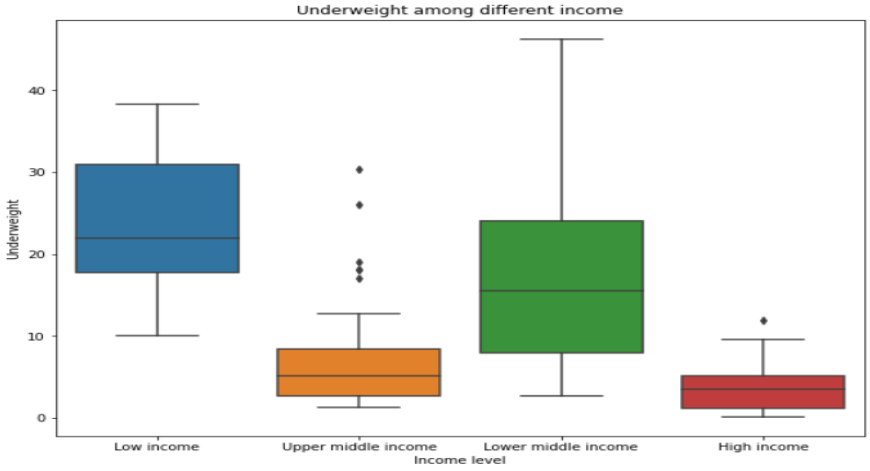
Stunting

Stunting is the impaired growth and development that children experience from poor nutrition, repeated infection, and inadequate psychosocial stimulation.



Both low and lower middle income countries have high median stunting rates. Similar reasons like the ones above.

Underweight



Both low and lower middle income have high rates of underweight issues. Similar reason to the one above.

DATA VISUALIZATION :

SORTING OF DATA

The data is sorted to analyse the highest and the lowest order of countries suffering from the malnutrition.

```
In [12]: data.sort_values(by=['Severe Wasting','Stunting','Wasting','Underweight','Overweight']).head(6)
```

Out[12]:

	Country	Income Classification	Severe Wasting	Wasting	Overweight	Stunting	Underweight	U5 Population ('000s)	Income level
6	AUSTRALIA	3	0.00	0.000000	13.875000	1.000000	0.10	1443.074500	High income
143	UNITED STATES OF AMERICA (THE)	3	0.04	0.514286	7.414286	2.914286	0.88	20077.909571	High income
53	GERMANY	3	0.10	0.650000	3.350000	1.500000	0.80	3641.690000	High income
108	REPUBLIC OF KOREA (THE)	3	0.15	1.050000	6.750000	2.500000	0.80	2487.097000	High income
66	JAPAN	3	0.20	2.300000	1.500000	7.100000	3.40	5554.144000	High income
4	ARGENTINA	2	0.20	2.150000	11.125000	10.025000	2.60	3613.651750	Upper middle income

```
In [13]: data.sort_values(by=['Severe Wasting','Stunting','Wasting','Underweight','Overweight'], ascending =False).head(10)
```

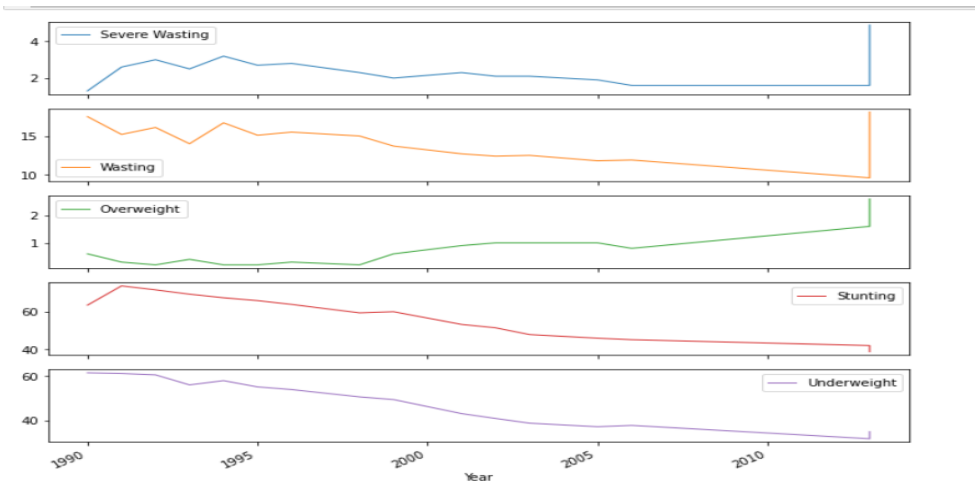
Out[13]:

	Country	Income Classification	Severe Wasting	Wasting	Overweight	Stunting	Underweight	U5 Population ('000s)	Income level
28	CHAD	0	5.500000	15.240000	3.040000	41.260000	31.400000	2024.290400	Low income
32	COMOROS (THE)	1	5.400000	10.125000	12.466667	39.125000	19.575000	90.287750	Lower middle income
149	YEMEN	0	5.300000	15.242857	4.842857	51.900000	37.785714	3269.476000	Low income
129	SYRIAN ARAB REPUBLIC (THE)	0	5.050000	9.566667	17.825000	28.550000	10.050000	2476.209333	Low income
132	TIMOR-LESTE	1	4.775000	15.620000	3.650000	55.380000	42.800000	160.444800	Lower middle income
127	SUDAN (THE)	1	4.766667	15.400000	2.900000	36.866667	29.933333	5551.013000	Lower middle income
81	MALDIVES	2	4.725000	15.780000	4.660000	32.600000	30.320000	36.045400	Upper middle income
46	ERITREA	0	4.633333	14.600000	1.666667	53.225000	37.550000	407.654500	Low income
100	PAKISTAN	1	4.500000	14.314286	4.757143	46.671429	33.737500	21774.897875	Lower middle income
115	SAUDI ARABIA	3	4.500000	7.350000	3.650000	15.350000	9.400000	2744.397000	High income

TIME SERIES ANALYSIS

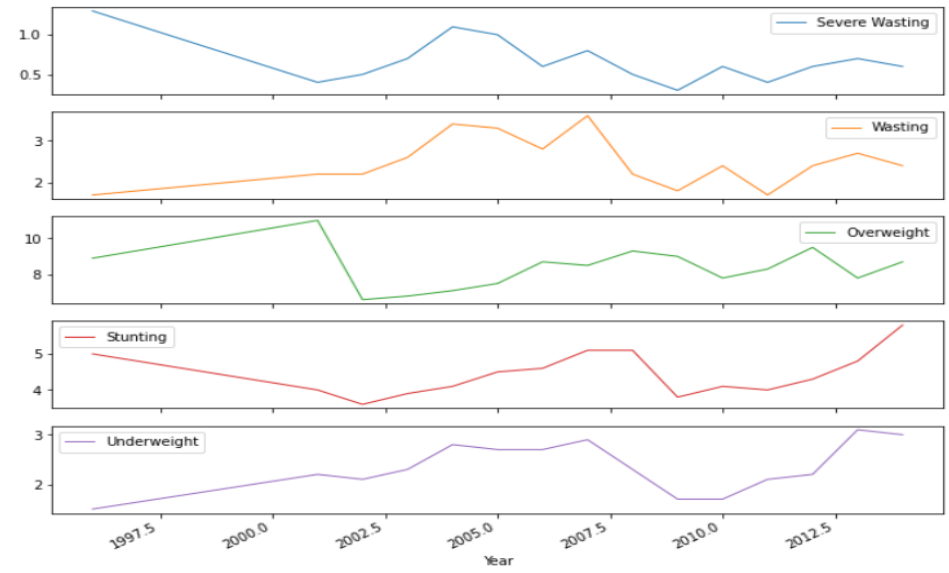
1. Have malnutrition issues improved over the years for Bangladesh?

Here, I will pick countries with more survey years since it will be clearer for us to see the difference over the years.



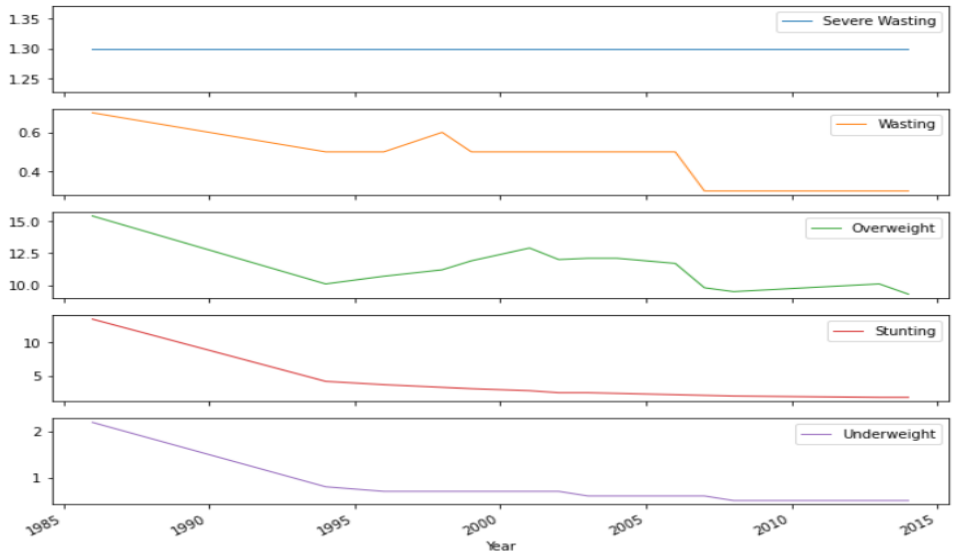
We observed an improvement in severe wasting, wasting, overweight, stunting and underweight issues over the years.

2. Have malnutrition issues improved over the years for Kuwait?



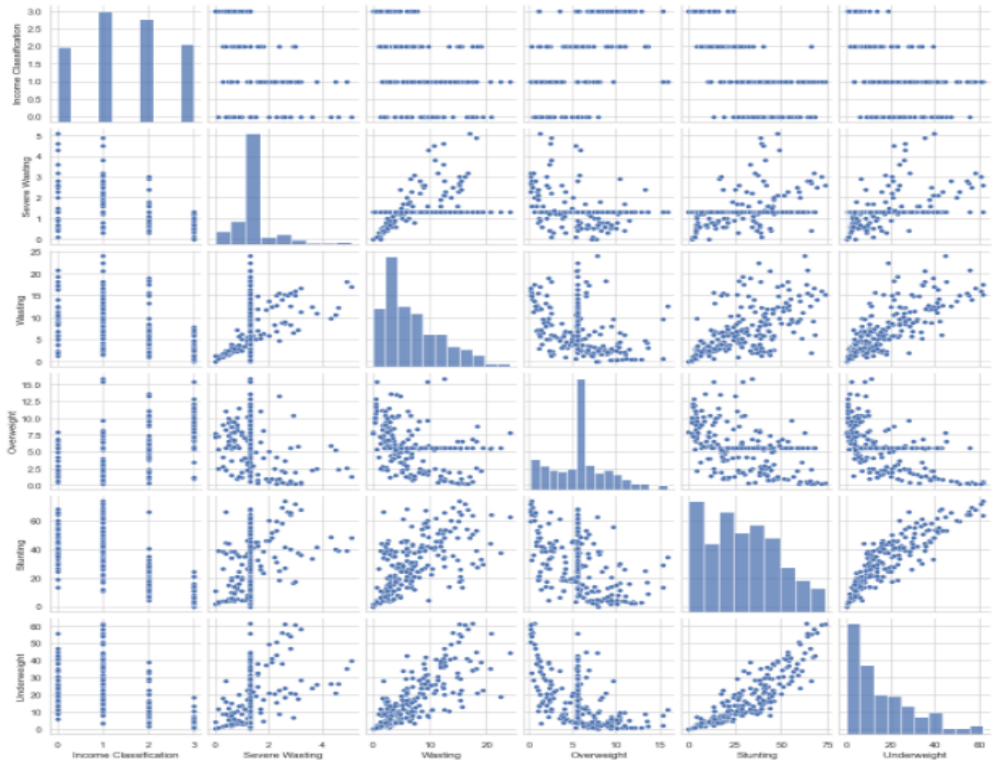
For Kuwait we observed a fluctuation of values. We see an improvement in 2007 onwards. Seems like it took Kuwait quite some time to recover from the war.

3. Have malnutrition issues improved over the years for Chile?



We observed a decrease in malnutrition issues in Chile over the years. That was mainly due to the fact that Chile is a high income country.

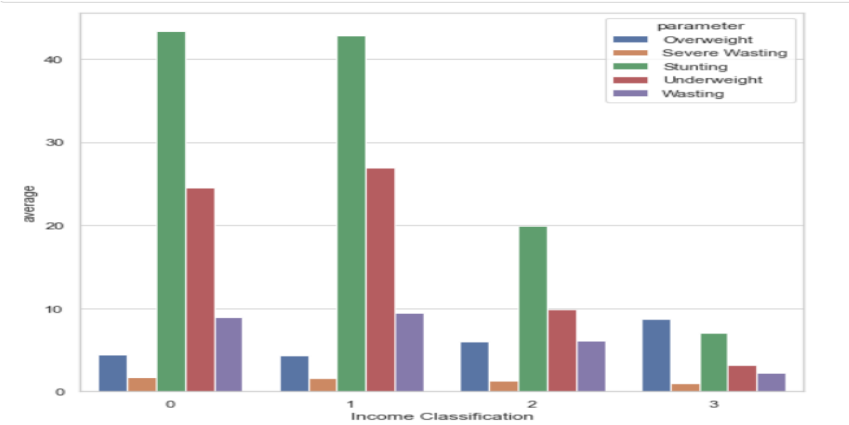
PAIRPLOT



MALNUTRITION ACROSS DIFFERENT INCOME GROUPS

Undernutrition categories :
stunting, wasting, overweight, Underweight, and severe wasting

Income Classification:
Low income = 0, Lower middle income = 1, Upper middle income = 2, High income = 3,



Observations from this plotting:

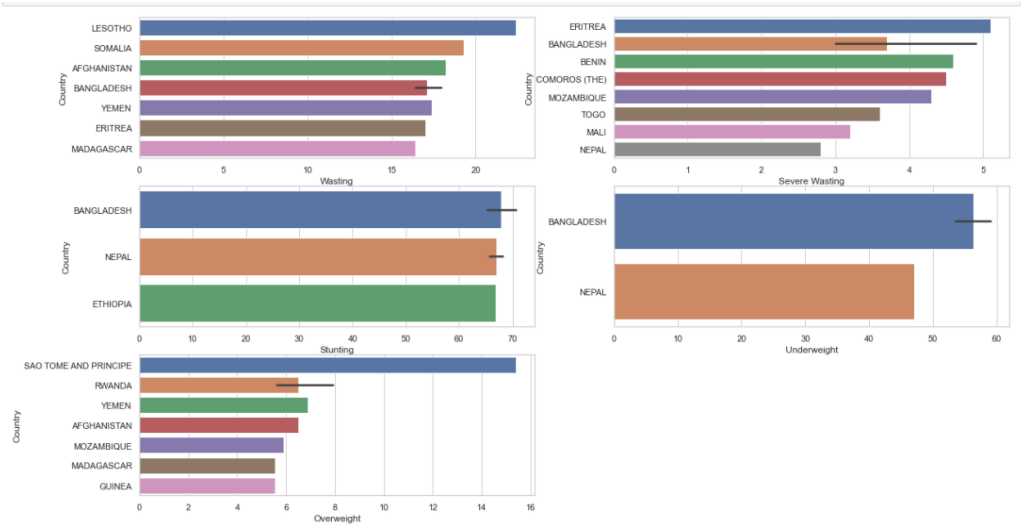
- Every country in the world is affected by one or more forms of malnutrition.
- We can observe that 'Stunting %' and 'Underweight %' is very high in LOW and LOWER MIDDLE income countries. Whereas 'Overweight %' is comparatively higher in UPPER MIDDLE and HIGH income countries.
- 'Wasting %' and 'Severe Wasting %' is also higher in LOW and LOWER MIDDLE income countries.
- Overall we can see that LOW and LOWER MIDDLE income countries are the ones which are most affected by malnutrition.

Analysis of countries which are both LDC and LIFD

Based on recent survey (survey year= 2018 or 2019)

Low Income Food Deficit (LIFD)(true=1, false=0) , Least Developed Countries (LDC)(true=1, false=0)

Our data, malnutrition-estimates.csv has a recent survey of only 32 countries.



Observations from this plotting:

- countries which are most affected from malnutrition are:

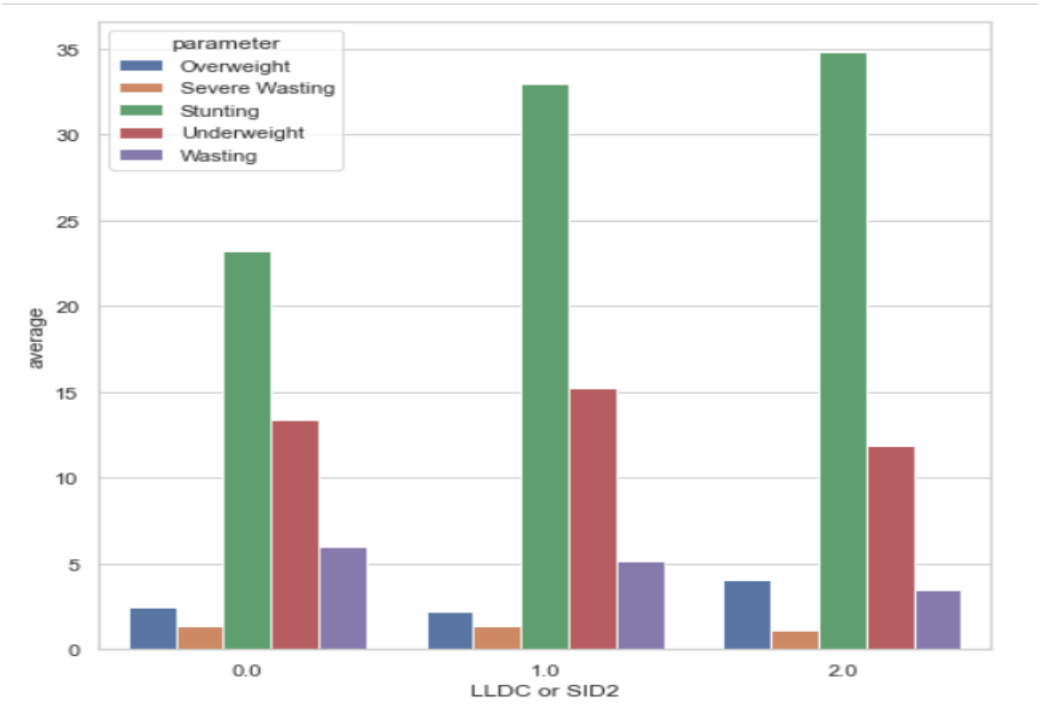
Niger, Ethiopia, Afghanistan, Mauritania, Guinea, Mali, Burkina Faso, Bangladesh, Senegal, Central African Republic, Madagascar

- Highest % of undernutrition categories:

Stunting: 56%, Wasting: 14%, Severe Wasting: 36%, Underweight: 37%, Overweight: 6.8%
Overweight % is very less in such countries(<=8%)

Malnutrition in LLDCs and SIDSs:

Landlocked Developing Countries (LLDC) ('LLDC or SID2' =1), Small Island Developing States (SIDS) ('LLDC or SID2'= 2),

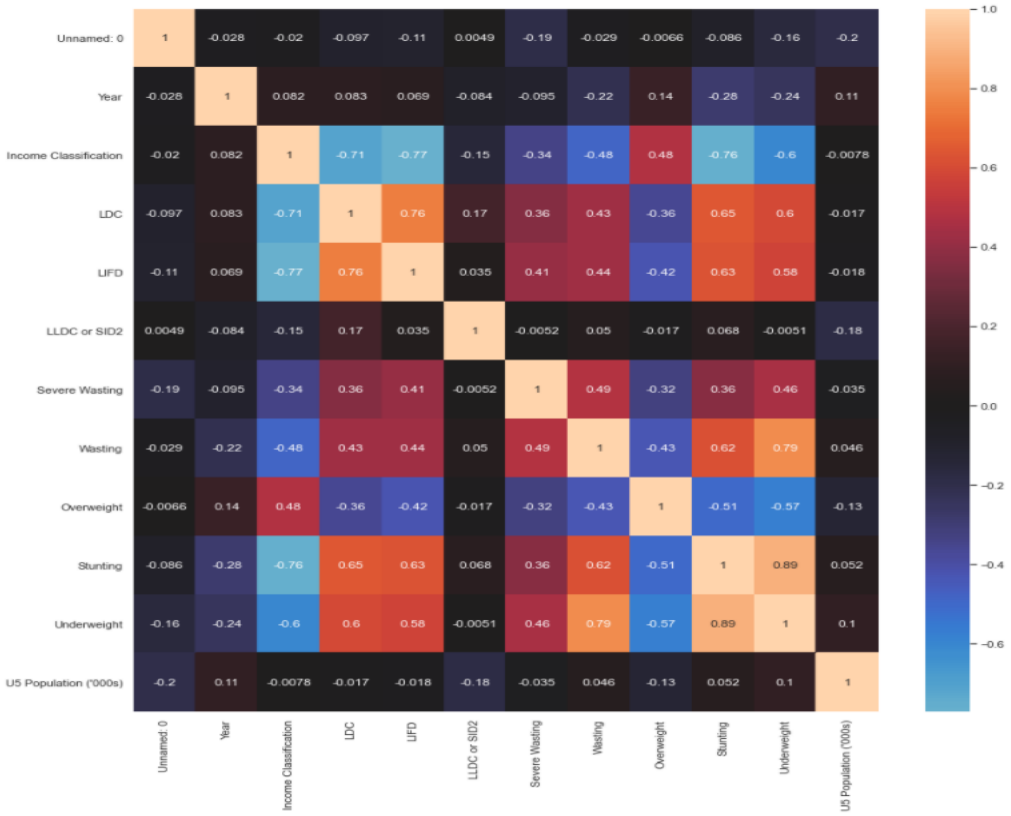


Observations from this plotting:

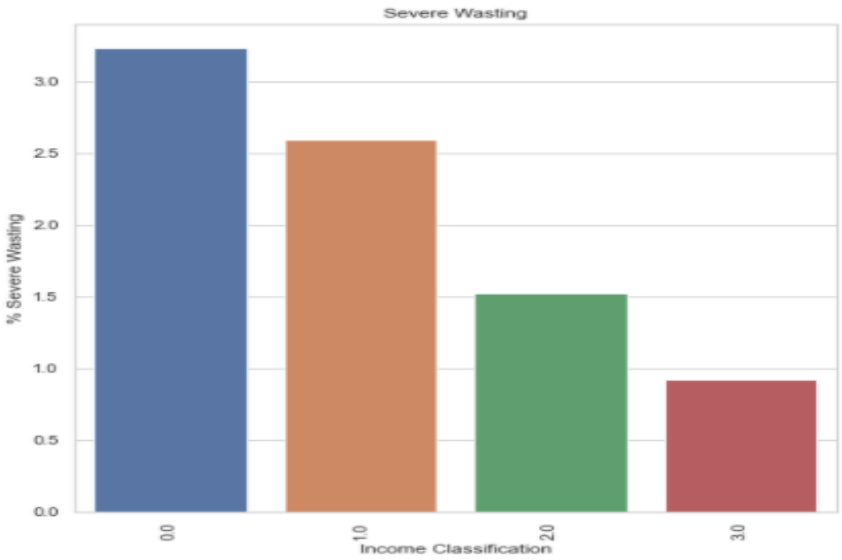
LLDC countries have a higher percentage of 'Stunting' undernutrition (about 30% of the children suffer from 'Stunting'). 'Severe Wasting %' and 'Overweight %' are very less in LLDC countries. About 16% of children are 'Underweighted'.

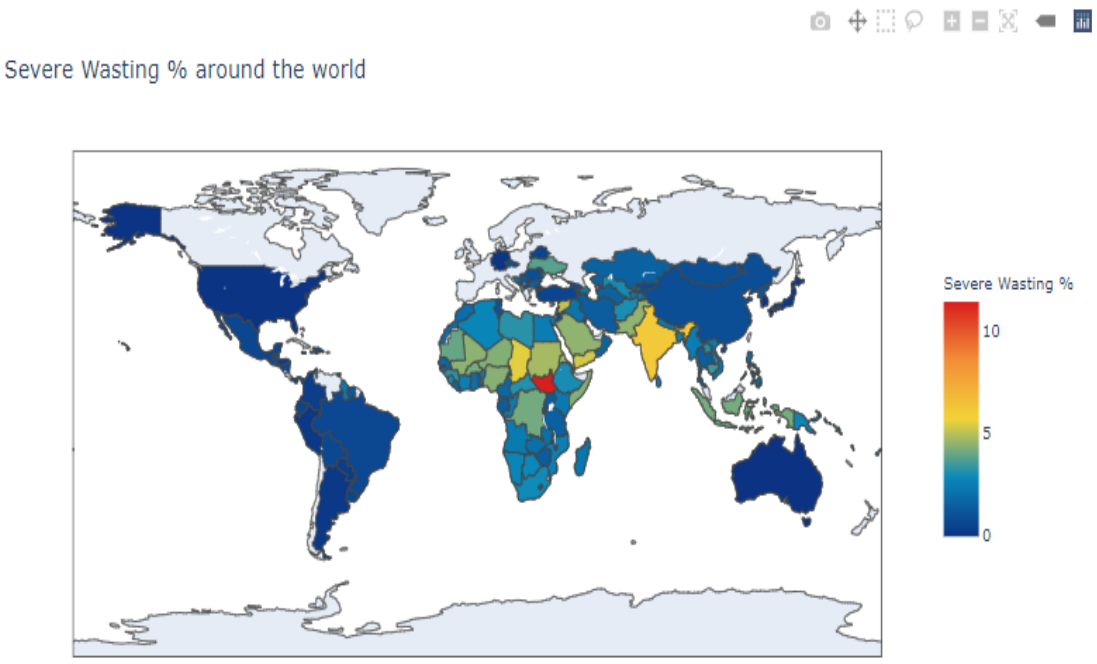
CORRELATION USING HEATMAP

Each square shows the correlation between the variables on each axis. Correlation ranges from -1 to +1. Values closer to zero means there is no linear trend between the two variables. The closer to 1 the correlation is the more positively correlated they are; that is as one increases so does the other and the closer to 1 the stronger this relationship is. A correlation closer to -1 is similar, but instead of both increasing one variable will decrease as the other increases. The diagonals are all pale because those squares are correlating each variable to itself (so it's a perfect correlation). For the rest the larger the number and darker the color the higher the correlation between the two variables. The plot is also symmetrical about the diagonal since the same two variables are being paired together in those squares.

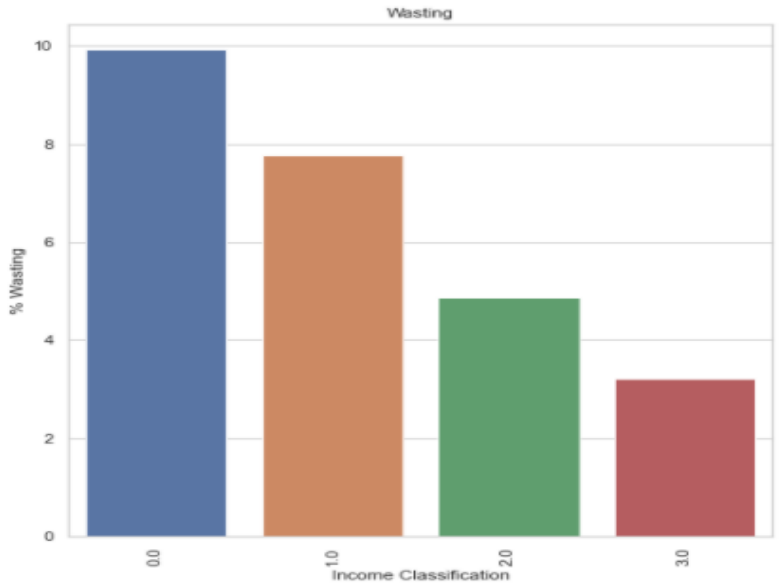


SEVERE WASTING ACROSS INCOME CLASSIFICATION

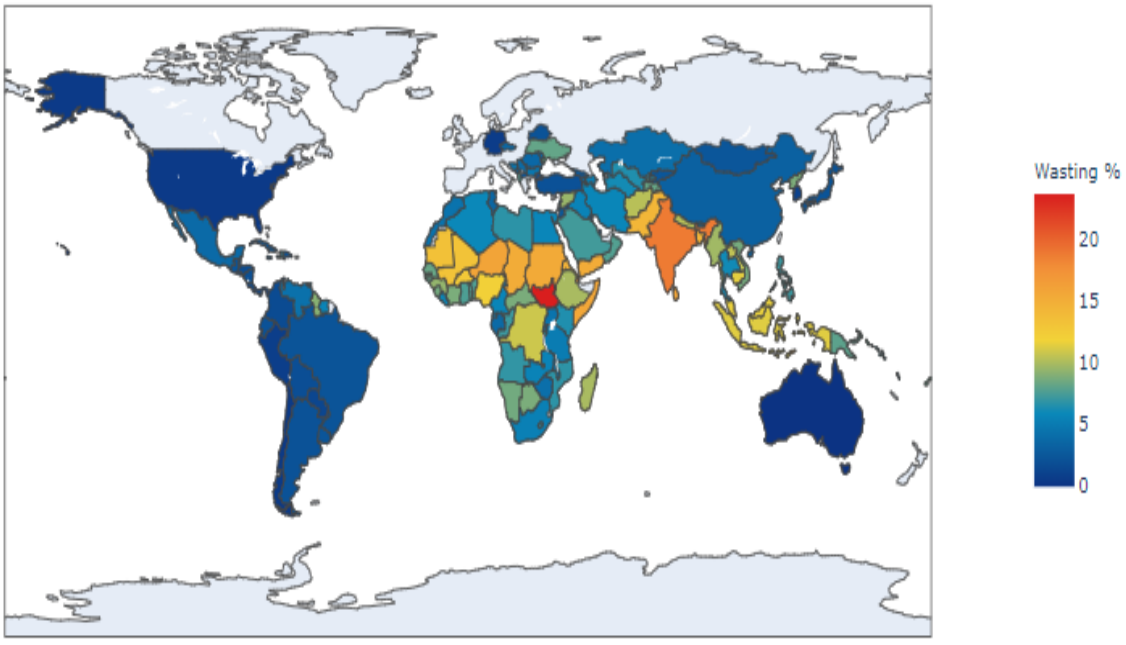




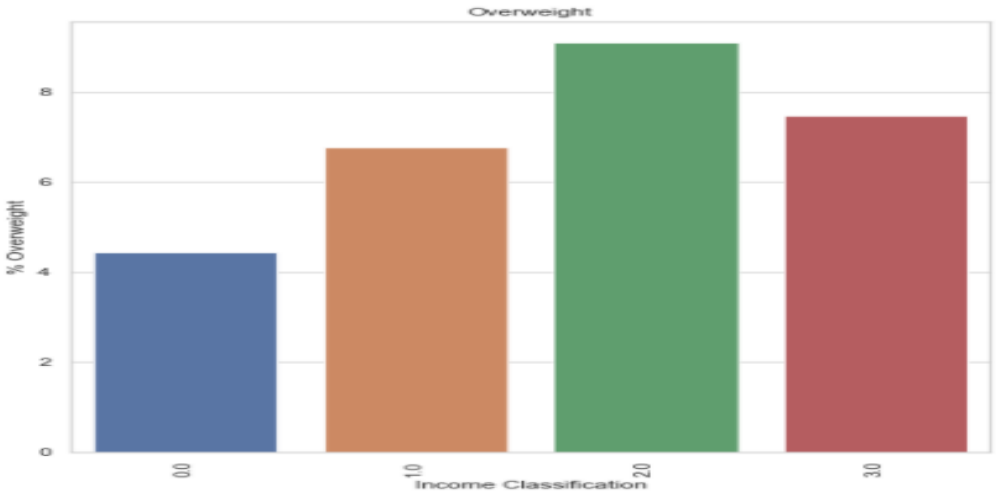
WASTING ACROSS INCOME CLASSIFICATION



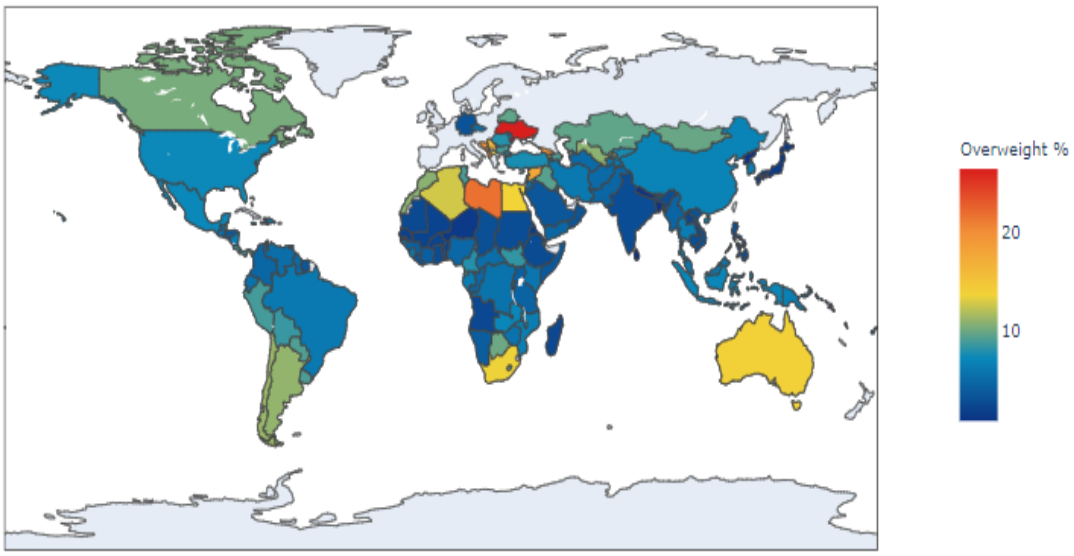
Wasting % around the world



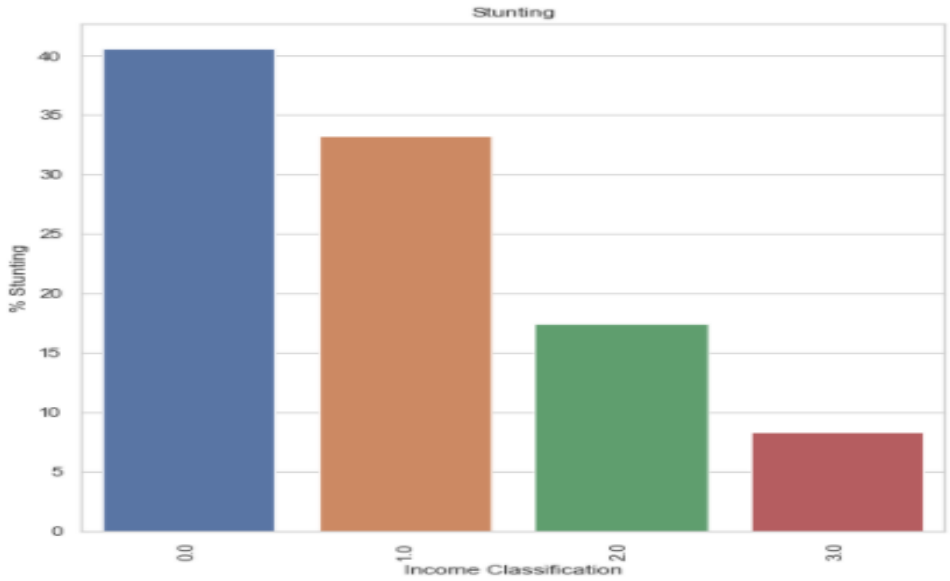
OVERWEIGHT ACROSS INCOME CLASSIFICATION



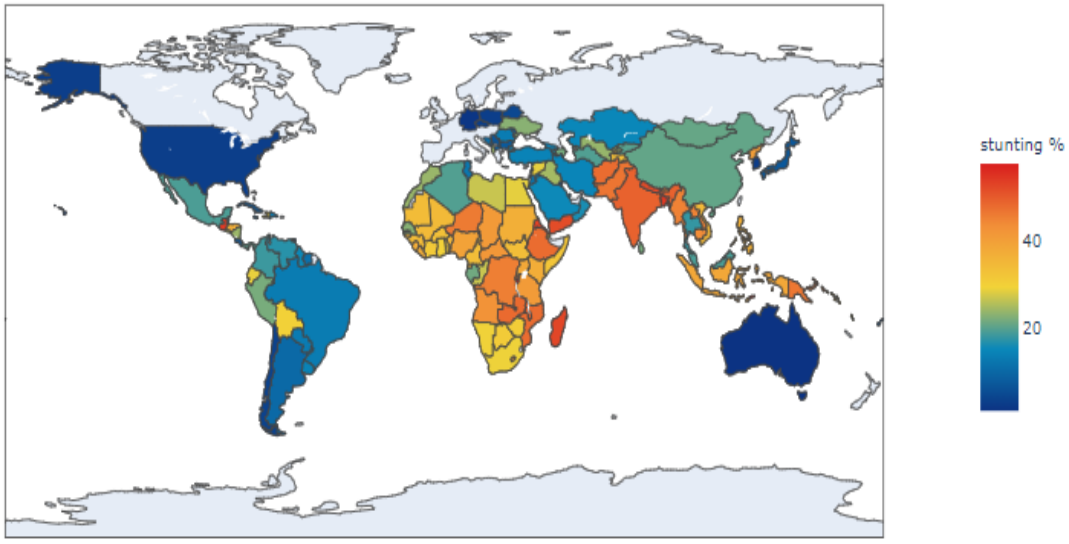
Overweight % around the world



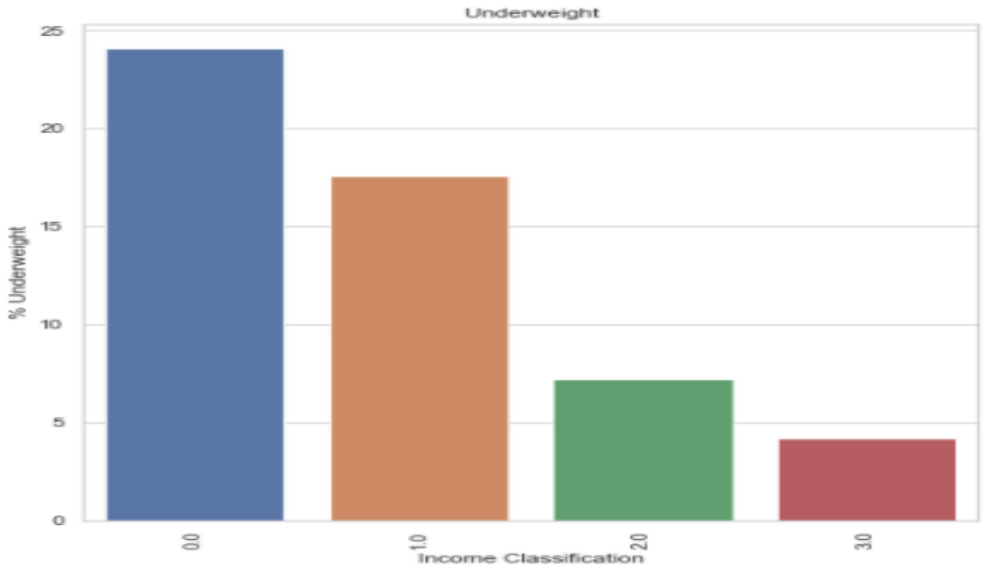
STUNTING ACROSS INCOME CLASSIFICATION



stunting % around the world



UNDERWEIGHT ACROSS INCOME CLASSIFICATION



Underweight % around the world