Exploration of Four Different Income Levels: On Topics of Health and Poverty INSY 6500 - Term Project

Bekircan Kirkici

December 6, 2019

Contents

1 Introduction

Since beginning of this class, we have focused on many datasets. They were almost always very specific parts of something bigger, such as the shopping data, the famous iris data set etc. but never looked at information on a macro level. Hence, we wanted to focus on a macro scale of information. Our understanding of economical data is better than any other subject thus, we proceeded with economical data.

At start we wanted to focus on the economic definition of "Core - Periphery"Wallerstein (1974), but current world economists still argue about the definition and which countries fit in theirs or do not even fit anywhere. Thus, we proceed to use a more analytical approach; income levels of countries.

World Bank classifies all the countries in four different income levels; high, upper middle, lower middle and low income. Because the distinction is clear, we will use this classification method.

Our report will, first do the initialization of the project as presenting our toolbox and intent. Then, we proceed to follow the chronological steps through the project work: finding interesting data, explaining the data to reader/user and exporting it. Afterwards, we will clean and tidy the data as much as possible and explore. Our exploration pattern is general to specific, hence, topics that are first to be discussed are more general compared to the later discussions. Finally, we will summarize our work and finish the report.

We will also address some of the code used in our provided Jupyter Notebook file, for those, we will also use a sectioning and will refer to its place. We also will try to construct this report as a guide to follow while reading the Jupyter Notebook, thus we will explain the results/codes/etc. parallel to the Notebook file. One should also keep in mind that, to run the Notebook provided, they need to install the *world_bank_data* package. While there are other ways to install, we used *pip* to install it, as just running the following code in the Anaconda Prompt at any workspace.

```
pip install world-bank-data
```

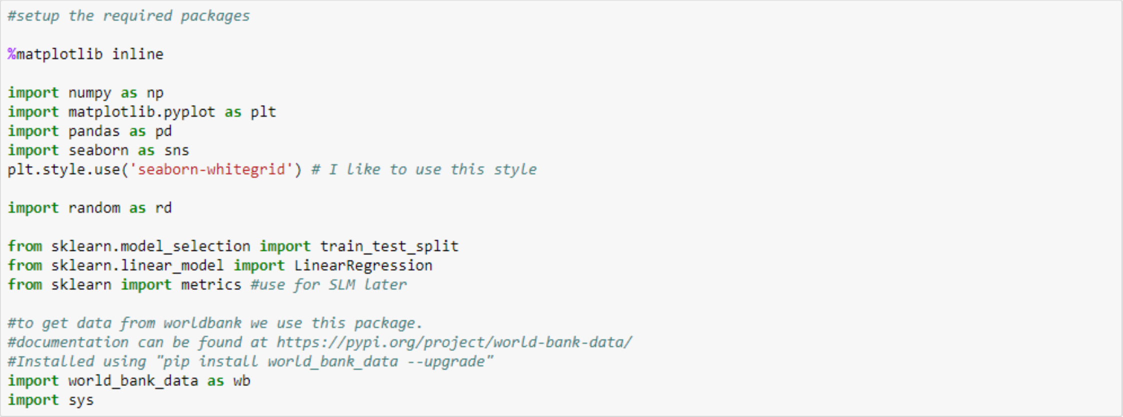
1.1 Toolbox

We used Python, with version 3.7.x, and Jupyter Notebook as the interpreter for this project and the course book VanderPlas (2016) as our main guide and written source. We also used many `stackoverflow.com` discussions to for help and links are provided within the Notebook cells when used and a hex color generator at `https://www.hexcolortool.com` is used to generate color codes for our graphs.

The packages used within Python are; *NumPy*, *Pandas*, *Matplotlib*, *Seaborn*, as usual, and "*World Bank Data* Wouts (2019)" as extra. While first four packages need not introduction, the last one should be addressed.

Although we also import *SkLearn* package and some of its contents, they will not be used. This is due to the reason that the package is used for testing purposes and results are deleted from the final

Notebook as their content is not in the scope of this course.



```
#setup the required packages

%matplotlib inline

import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
plt.style.use('seaborn-whitegrid') # I like to use this style

import random as rd

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn import metrics #use for SLM later

#to get data from worldbank we use this package.
#documentation can be found at https://pypi.org/project/world-bank-data/
#Installed using "pip install world_bank_data --upgrade"
import world_bank_data as wb
import sys
```

Figure 1: Import Section

Data required for this project is accumulated in World Bank Database (WBD). One can access their website and query their data there manually, but to get multiple variables for multiple years, using a simple loop would be more efficient, thus, we found this package online.

World Bank Data package complies the wanted parameters for data sets, such as desired indicators, years, simplifications etc., and creates multiple URLs, then, accesses WBD and converts the data into *Pandas* dataframes.

Having *Pandas* dataframes without reading in a .csv file is obviously more advantageous since we do not have to deal with reading-in complications, such as encoding, delimiter type etc.

1.2 Intent

Our intent for this project is to identify possible research topics and predictive model ideas. Hence, we will not be delivering any conclusions or building predictive models but point out to interesting findings within the scope of this project. Thus, there will not be any literature review to check missing areas of research. We will try to point the reader to at least read more about the topics that will be discussed in Section 2 and demonstrated in 3.

We will also try to provide a small literature suggestions and examples about topics as we go along. While these are not *"must reads"*, they are certainly a start point for a future researcher in those areas.

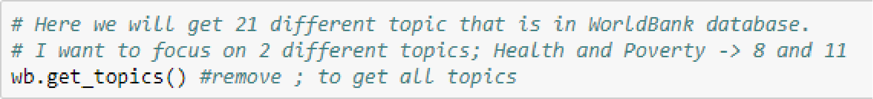
1.3 Randomness

In project's version 1.0.0 and above, we have implamented a randomness for country selection. One can see the instruction in Notebook **Section 3.0** and **3.1**. While most of the graphs will be generated, for reasons we have not explored, yet, pie charts will cause an error. We believe this is because of the manual selection of data rather than a relative selection.

Reader is encouraged to take a look while keeping in mind that the randomness feature is not in its final form.

2 Data Explanation & Exporting

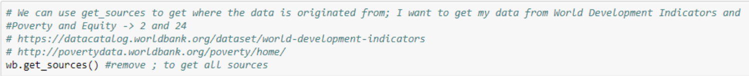
To start with, we look at the subjectively interesting topics in WBD, with the command in Figure 2. The output is given in the **Section 1.2** of the Jupyter Notebook provided.



```
# Here we will get 21 different topic that is in WorldBank database.
# I want to focus on 2 different topics; Health and Poverty -> 8 and 11
wb.get_topics() #remove ; to get all topics
```

Figure 2

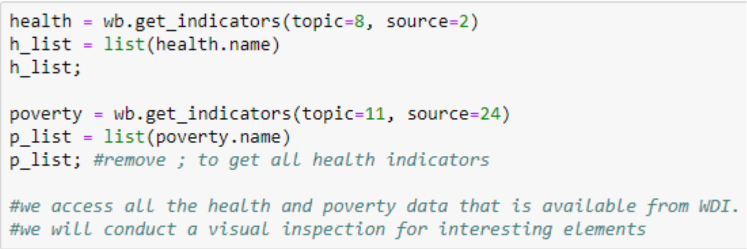
For us, most interesting topics were, Health and Poverty, which correspond to the implicit index of 8 and 11. Thus, we should look for an appropriate source to get this data. We run the code in Figure 3 and get all the 70+ sources. From them, we choose World Bank's own sources; "World Development Indicators" and "Poverty and Equity", which correspond to the implicit index of 2 and 24 respectively.



```
# We can use get_sources to get where the data is originated from; I want to get my data from World Development Indicators and
#Poverty and Equity -> 2 and 24
# https://datacatalog.worldbank.org/dataset/world-development-indicators
# http://povertydata.worldbank.org/poverty/home/
wb.get_sources() #remove ; to get all sources
```

Figure 3

Now that we have the indexes for topics and sources, we explore into the topics' offerings. First, we export the topic data from the corresponding source, put it in a list and manually review it to see interesting variables. Thus, we run the codes below.



```
health = wb.get_indicators(topic=8, source=2)
h_list = list(health.name)
h_list;

poverty = wb.get_indicators(topic=11, source=24)
p_list = list(poverty.name)
p_list; #remove ; to get all health indicators

#we access all the health and poverty data that is available from WDI.
#we will conduct a visual inspection for interesting elements
```

Figure 4: Health and Poverty Indicators

From health topics, our interest falls under; smoking, alcohol consumption and child mortalities. Thus, we gather the following variables:

- Smoking prevalence, total, male, female (% of adults),
- alcohol consumption per capita, total, male, female (liters, for adults),
- infant mortality rates, under-5 mortality rates,
- out-of-pocket health expenditures and,
- skilled staff that attend births.

Under poverty topics:

- Income share held by the five quantiles,
- GINI index and,
- 3 different poverty lines.

and lastly we get some general population data:

- Population total and growth and,
- population distribution within ages and sexes.

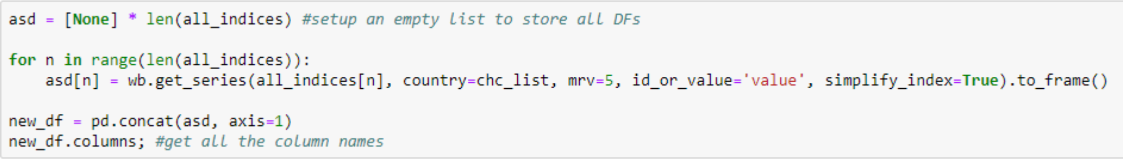
All of the codes can be found in whole **Section 2** of the Jupyter Notebook provided.

Now that we have our variables of interest, we proceed to first, choose countries to explore from four different levels of income then get all the variables under one *Pandas* dataframe.

We first extracted all the available countries in the dataset. We saw that an area/political/trade aggregate, such as OECD countries, is also technically classified as countries. Thus, we first filtered out the aggregates then listed all the income level countries and printed the number of countries available.

After seeing our options, we, with no specific purpose choose, with high to low income, with no inner grouping, choose the following countries to explore: “United States (US), Sweden, Austria, France, Switzerland, Turkey, Russian Federation (Russia), Mexico, Colombia, China, Vietnam, Moldova, Senegal, Bangladesh, Egypt, Afghanistan, Somalia, Ethiopia, Haiti and Central African Republic (CAR)”. We also put their names and three letter abbreviations in a list for easy access.

These lists can be viewed in **Section 3.1** of the Jupyter Notebook provided.



```
asd = [None] * len(all_indices) #setup an empty list to store all DFs

for n in range(len(all_indices)):
    asd[n] = wb.get_series(all_indices[n], country=chc_list, mrv=5, id_or_value='value', simplify_index=True).to_frame()

new_df = pd.concat(asd, axis=1)
new_df.columns; #get all the column names
```

Figure 5: Loop to get the datframe

After choosing the countries and putting all the variables in one list, we create the main dataframe in Figure 5 with an outer join of all the individual dataframes. One can see that this is a simple loop with five parameters in the function. They are, in respect to their position in the code; all the indicators, all the country abbs, five most recent values, requesting value (name) of the country and simplifying the indices. Simple indices only uses countries and years as indices, otherwise the data will have the variables in indices and years in columns, thus creating a complexity during joining and also creating many more *NaN* values than there actually is.

We also change the column names to a more readable version of the original and put all the related columns in specific lists for easier access. These can be seen in **Sections 3.3.1** and **3.3.2** of the Notebook. Now we have slightly cleaned out dataset but did not get rid of the NaNs.

The reason we did not clean NaNs is that, because we choose five most recent values (MRV) to export, rather than specific dates, countries' MRV's may not match. We choose to use MRV instead of specific years because NaN values in specific years create an HTML error. As we mentioned in Section 1.1, that the package creates and accesses an URL, thus a NaN value leads to a non-existent URL, causing major problems. At the end, we will work mostly with data from 2014 to 2018, and in some cases only a specified year.

We are now fully ready to work with our dataframe.

3 Exploration & Visualization

In this section, we will use our variables of interest and look at them in an appropriate graph and conduct some visual inspection then point out to some interesting information. Our general approach is to, first get the desired row-column combination from the dataset, then put it in a seperate list or dataframe to work with. After that we will visualize and edit the features of graphs.

First we will start with the population data, then continue into health data and finalize with poverty indicators.

3.1 Population

We have three different aspect to look at; total population, population growth and population pyramids.

3.1.1 Total Population

In Figure 6 we can see the total population distribution of the chosen countries as well as the rest of the world, so that the pie chart is the world population as a whole.

We only colored the highest five countries' populations for a cleaner view. The graph generation code is in **Section 4.1.1** of the Notebook. One can change the *“colors=colors2”* parameter in the code, to *“colors=colors_not_gray* so that it will be 21 distinct colors.

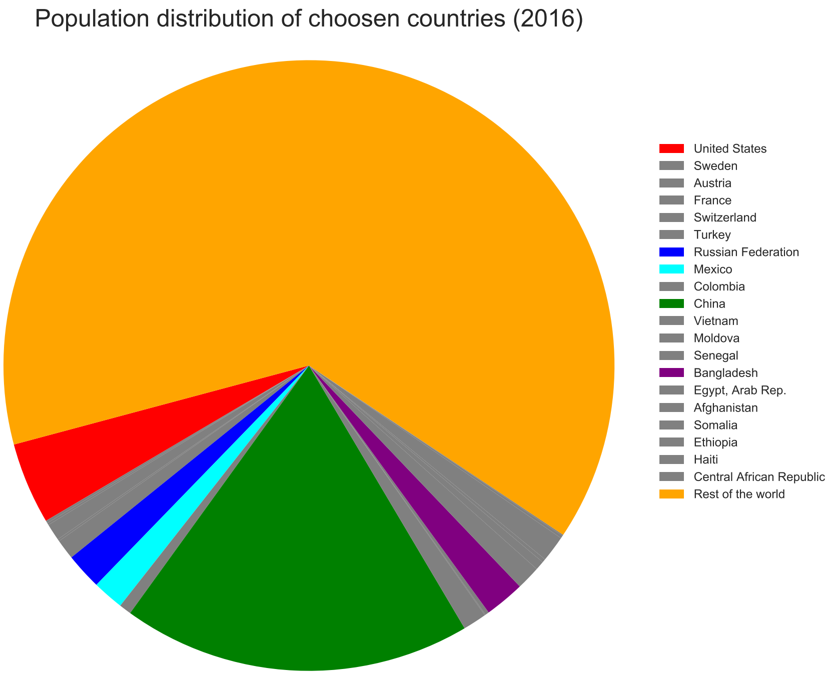


Figure 6

Here we can see that, if we didn't already know, China makes up a big chunk of world's total population and our 20 countries cover 36.44% of the whole world population.

3.1.2 Population Growth

In **Section 4.1.2** of the Notebook file, one can get five graphs; four different population growth graphs, with respect to income levels and a compiled one at the end. The reason is to show/inspect them at different times to see clearly. Figure 7 shows all countries' population growths, from 2014 to 2018.

Most interesting outcome is that Moldova has a negative population growth. This fact can be explored further if one seeks the reasons. Other interesting points are; stable growth in middle income countries, flactuated growth in high income countries and a big variance in lower income countries' population growth. Also, one should notice the rapid increase in CAR's growth and rapid decrease in Afghanistan's population growth.

While there can be little to say about population growth, Ehrlich and Holdren (1971) and Coale and Hoover (2015) are valuable guides for a curious reader's work.

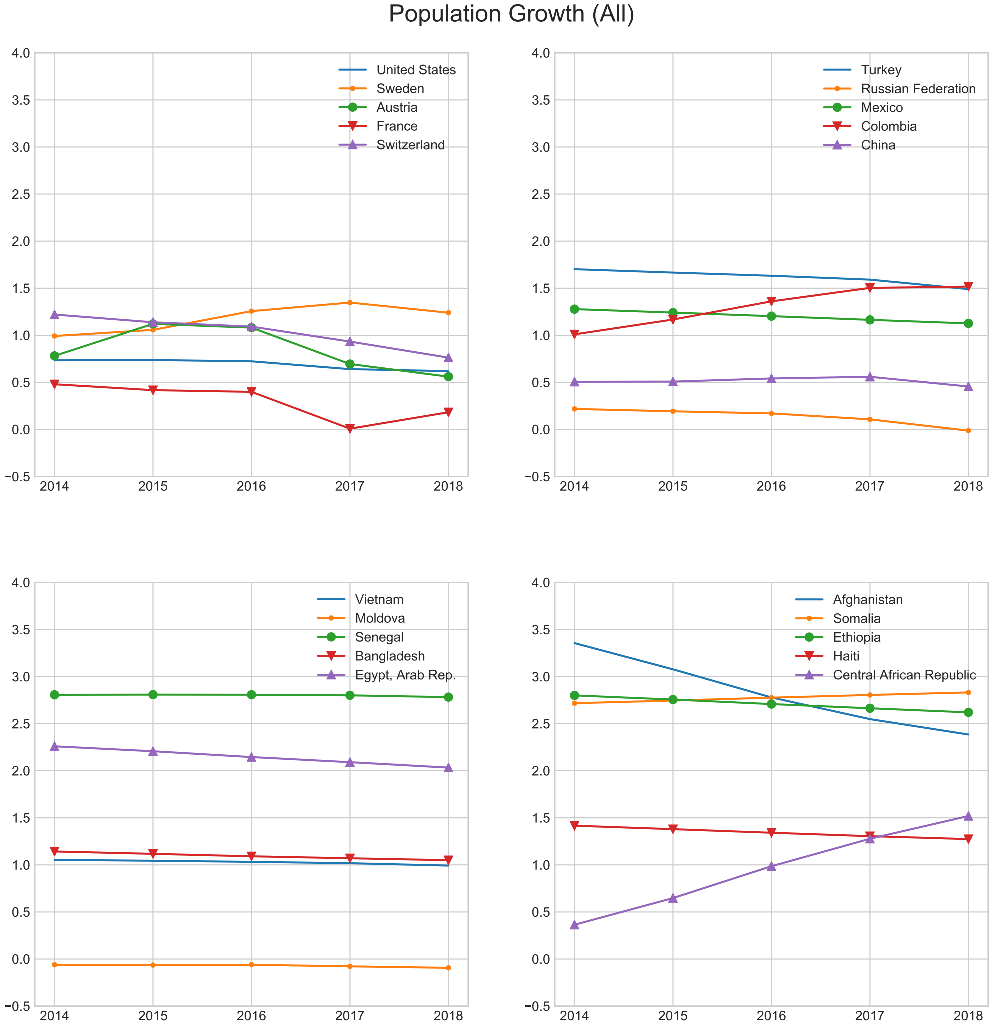


Figure 7

3.1.3 Population Pyramids

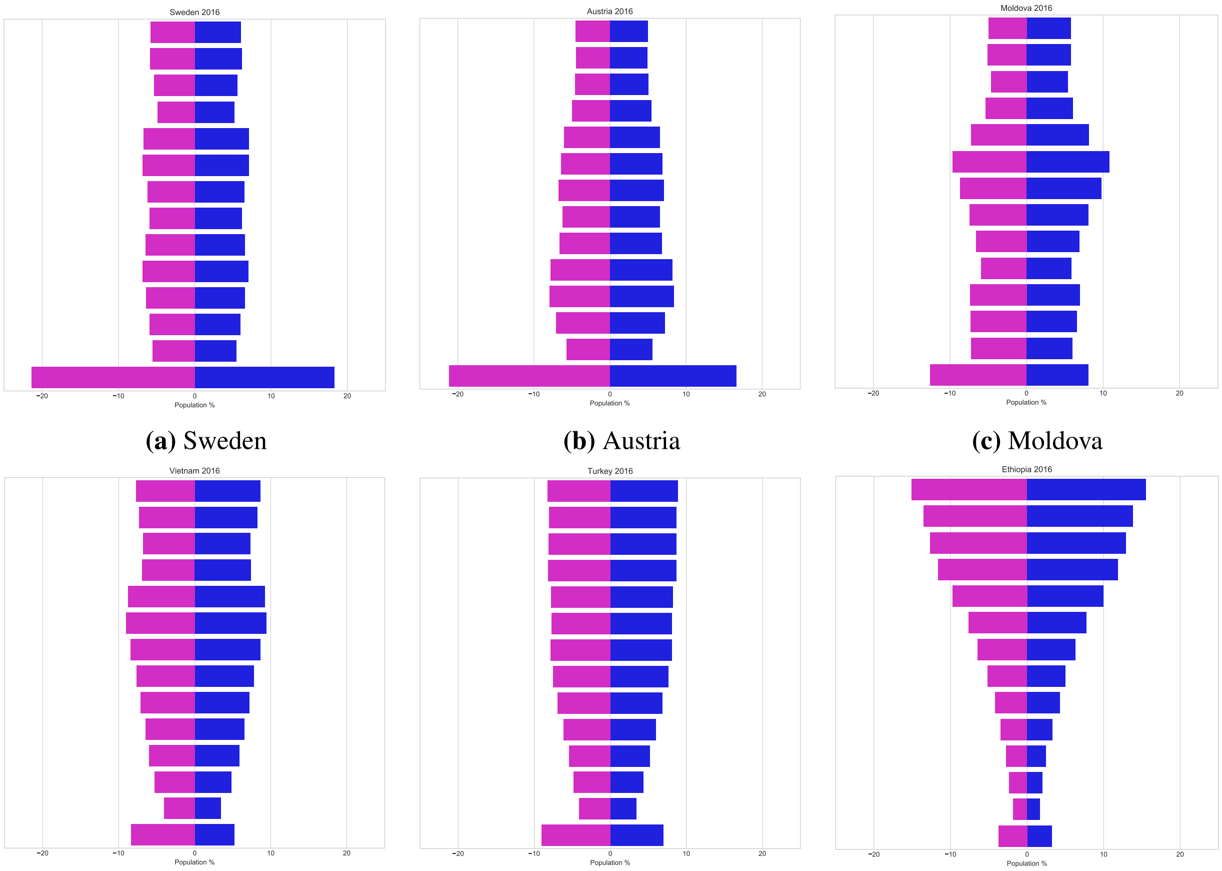
In **Section 4.1.3** of the Notebook, one can access all the population pyramids generated through the loop given there in two cells, and will also save the figures in their working folder. Here we will only visualize top three interesting findings for the sake of simplicity.

In Figure 8, from top to bottom, the age groups are increments of 5, thus first row is 0-4, second

5-9 and so on until the last being ages 65+. Color pink represents the females and blue represents the males. The statistics are with in the sex group. Thus, the third row in Figure 8f, for example, show the % of females of age 10-14 in total females and males in the same sense.

Figure 8 shows some handpicked countries' population pyramids. First, Sweden in Figure 8a, is a sample for showing the high income countries because all of them have a similar pyramid, where most of the population is 65+. In Figure 8f, we see a representation of low income countries where the opposite occurs; a younger generation is prevalent.

Also when we look at 8b,8c and 8d, we can see that the older female population is dominant. This obviously shows there is less males in 65+ age group than females, but the reasoning is left for the reader. Finally a more balanced distribution happens at upper middle income countries, such as in Figure 8e.



(a) Sweden

(b) Austria

(c) Moldova

(d) Vietnam

(e) Turkey

(f) Ethiopia

Figure 8

3.2 Health

Now that we have established some population demographics, we will look at the health indicators that are chosen; smoking, alcohol consumption and child mortalities.

3.2.1 Smoking Prevalence

We usually wonder which part of the society is addicted to cigarettes, and this distinction is usually made using income levels, thus creates the perfect basis for our data. In Figure 9, we can see four subplots for smoking prevalence in total adult population of the given country. Figure 9 is generated using **Section 4.2** of the Notebook.

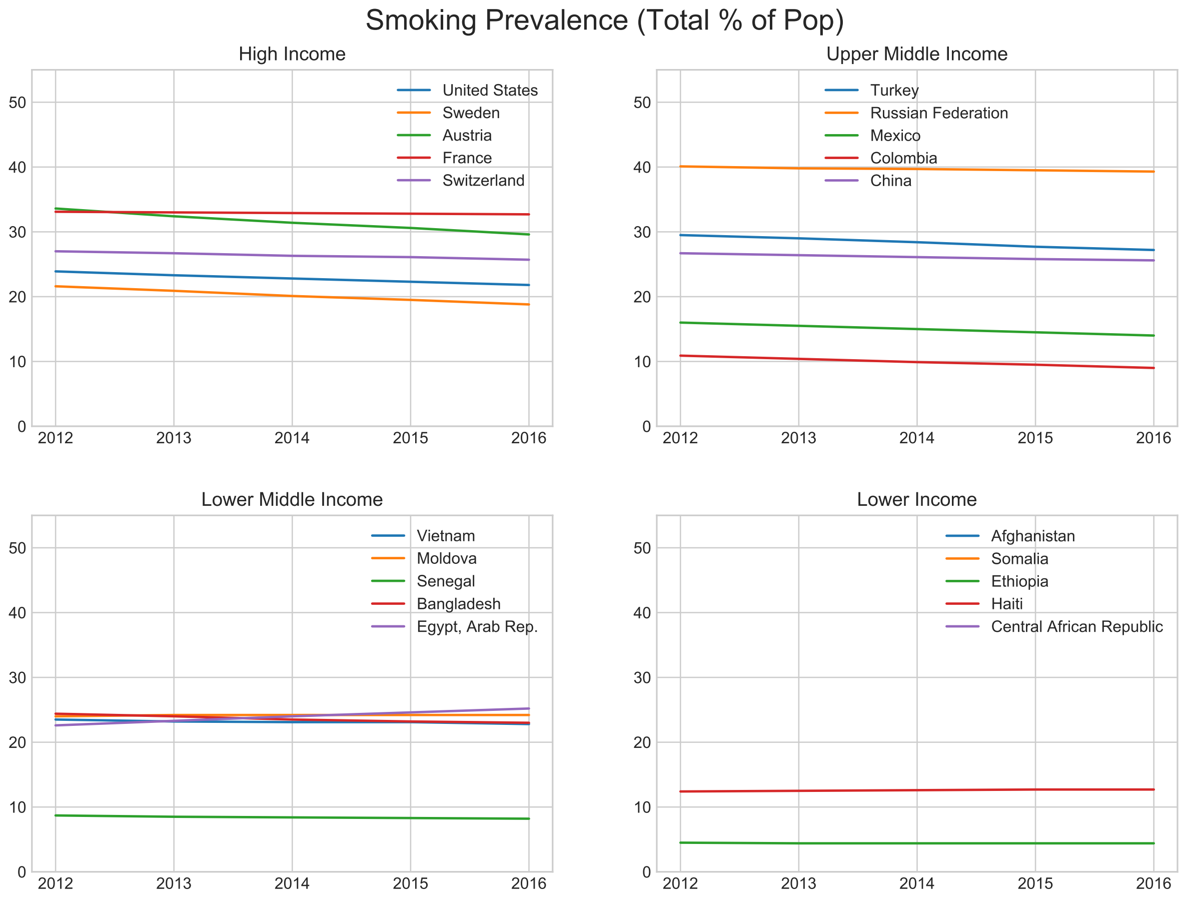


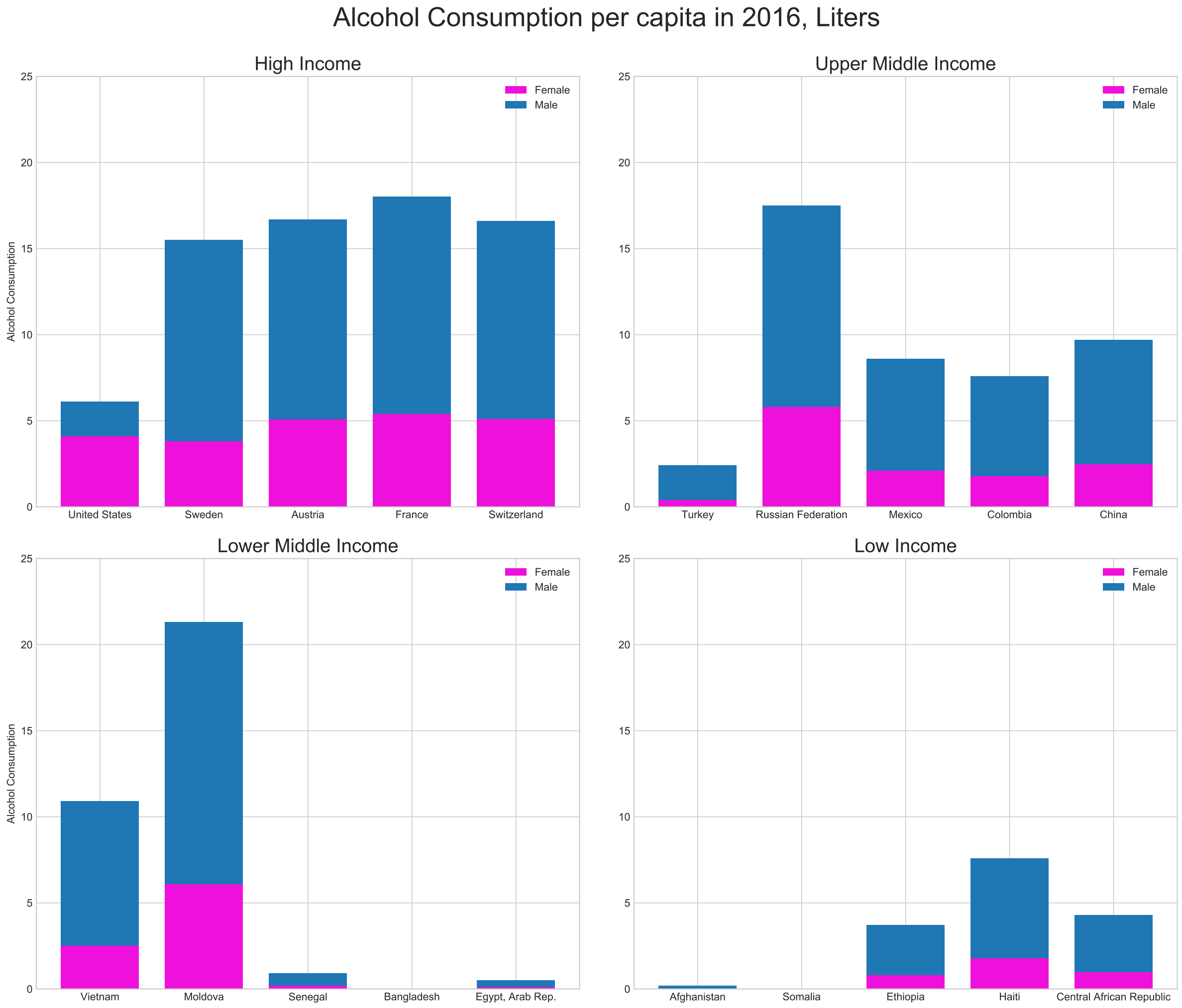
Figure 9

The most interesting findings here are; Russia is the most smoking country among the 17 (since there is no data for CAR, Somalia and Afghanistan), there seems to be a high decline of smoking in high income countries and a small decline in upper middle income countries and finally the other income levels seems to have a stable, or small increasing/decreasing, smoking prevalence. While there are economic studies about the tax effect on tobacco consumption such as Chaloupka et al. (2000) and Bader et al. (2011). We will, again, leave the reader with these facts so that they can arrive to their conclusions in future.

3.2.2 Alcohol Consumption

We can ask the same question for alcohol consumption, but this time, we will look at male versus female alcohol consumption, so that we can create more exploration possibilities for the future explorer/researcher.

Figure 10 shows the alcohol consumption of 2016 for our chosen countries and colors according to gender. We create this graphs and its components in **Section 4.3** of the Notebook.



Alcohol Consumption per capita in 2016, Liters

High Income

Upper Middle Income

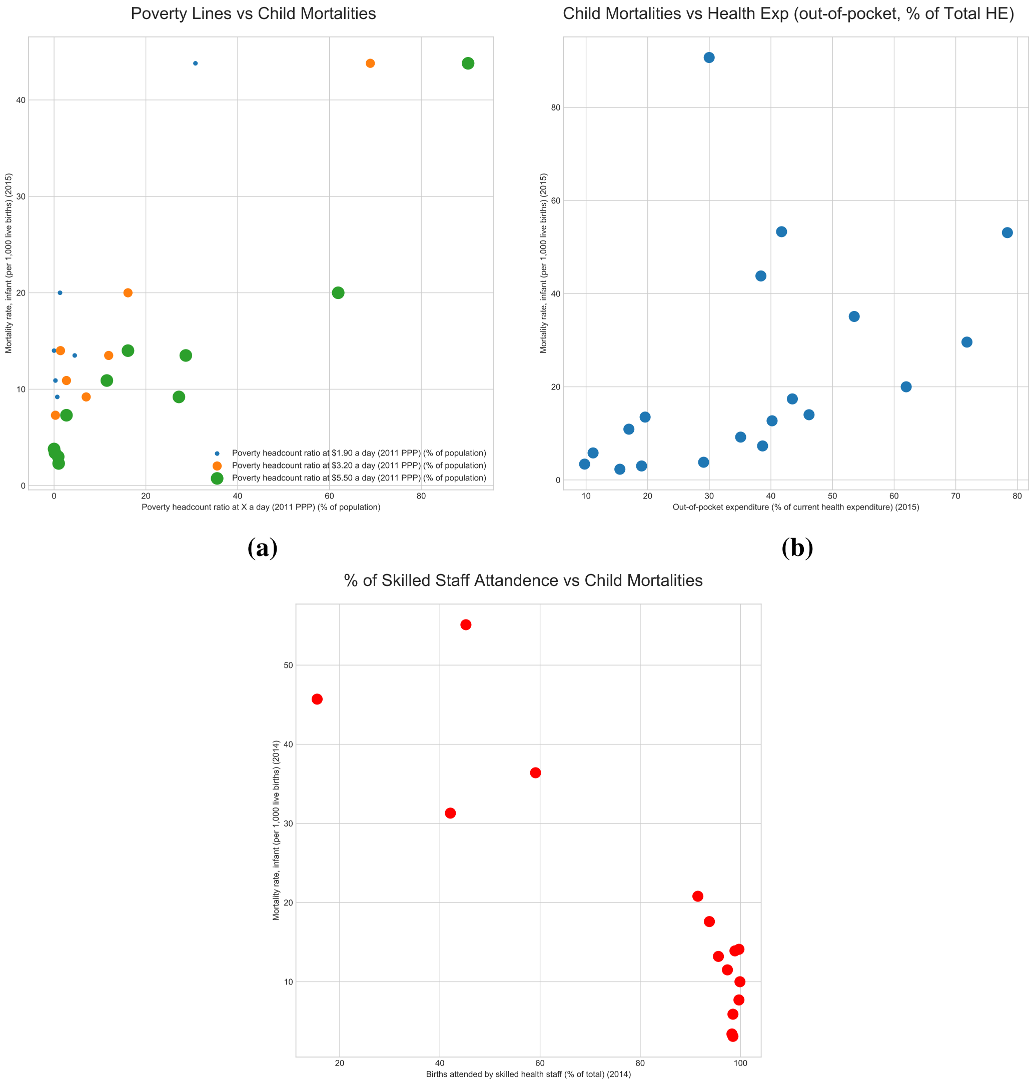
Lower Middle Income

Low Income

Figure 10

Here we can point out to four interesting facts. First, we see that only in US, females are the dominant alcohol consumers. Second, For some countries we have no data, we wonder if this is related to some policy, such as prohibitionism or high taxation on alcohol. Third, Moldova and Russia are the leading countries among the Twenty, one might be interested in the occurrence cirrhosis in those countries. Finally, an interested reader should also know that Russia has an historical alcohol problem McKee (1999), so this year's data can be useful for a very specific article/essay.

3.2.3 Child Mortalities



(a)

(b)

(c)

Figure 11

Child mortality (CM) data was the most interesting variable we saw during the exploration process of our project. Thus, we will show three different perspectives within our data as a potential explanatory variable in CM's explanation and prediction. The three separate graphs' contents are gathered in **Section 4.4** and graphs are created using the code in **Section 4.4.1 to 4.4.3** in the Jupyter Notebook.

Figure 11a shows how child mortalities are mapped against the population below a poverty line. Here we have three predefined poverty lines, \$1.90/day, \$3.20/day and \$5.50/day. Details about these lines will be discussed at Section **??**. One can see that three dots form a vertical line, this is because those three lines that form a perpendicular line are from one country, thus the CM count does not change. We constructed this graph to see the magnitude's' differences.

When we take a look at Figure 11b, this time one point representing on country, we see that there seems to be a steeper placement meaning that a small increase in CM shows that there is a relatively more increase in in out-of-pocket health expenditure (oopHE), in other words it is very elastic.

Last figure, 11c shows that even with a high percentage of skilled staff attendance to births (SSA), there seems to be a very sensitive increase in CM count and as low SSA's show, and as expected, the CM count is increased.

Reader should consider how CM is an indicator for a country's healthiness, such as in Reidpath and Allotey (2003), as well as the other factors which indirectly affect the healthiness. One can also look at a country's chronological data and observe any changes, as in MacDorman et al. (2013) and Bae and Bae (2004).

3.3 Poverty

Structure of this subsection will be almost exactly the same as Section 3.1 and 3.2. We will follow with the flow of the Jupyter Notebook and try to provide useful readings and meaningful questions along the way.

3.3.1 GINI Index

For readers which are unfamiliar with the Lorenz Curve and thus the GINI Index, we suggest reading Gastwirth (1972) before moving forward to maximize understanding in Sections 3.3.1 and 3.3.2.

Readers can find the preparation of Figure 12 within the **Section 4.5.1** of the Notebook.

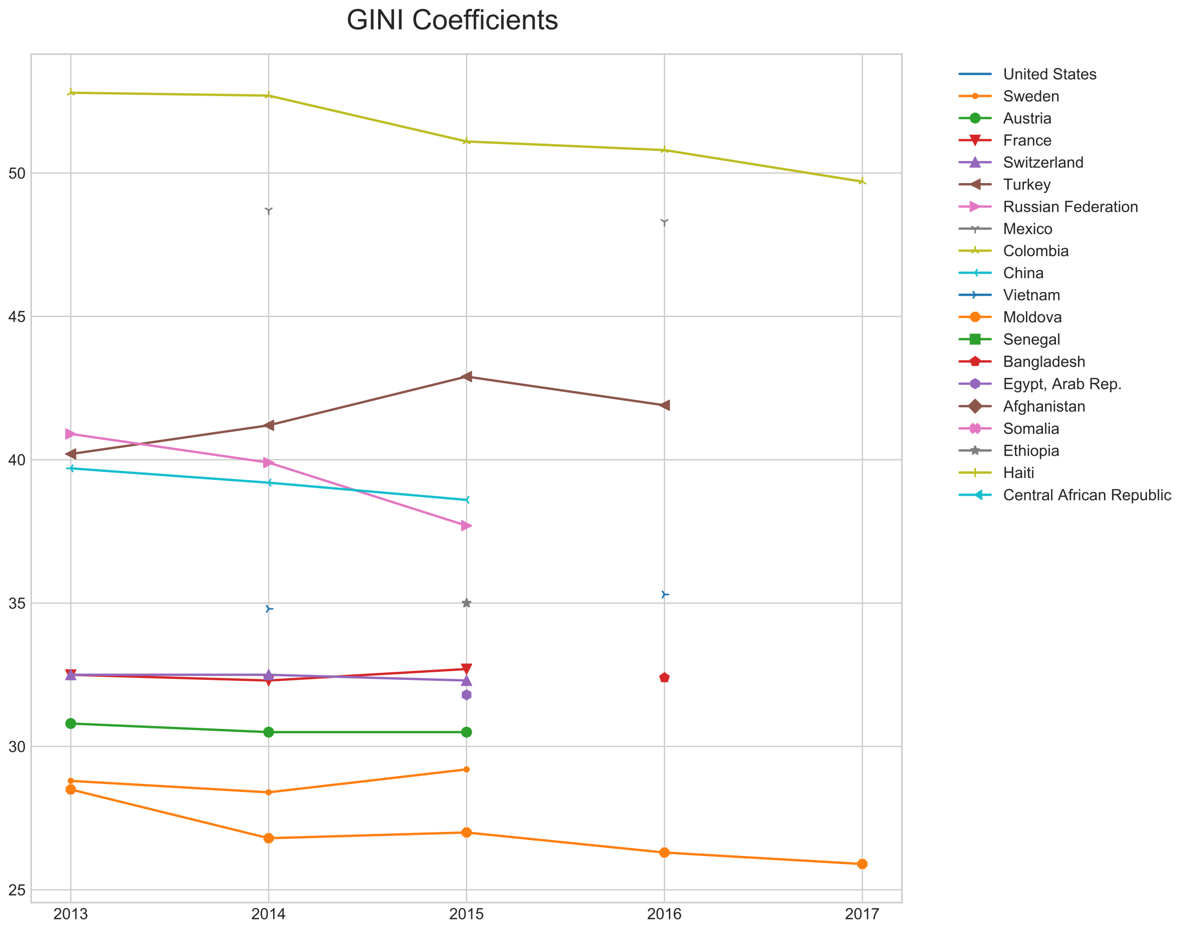


Figure 12

Figure 12 shows the GINI Index movement of the twenty countries we choose. One can easily see that there is missing data. While this is a small disturbance, it won't affect our questions.

We should notice two things at the first glance. First the highest and lowest GINI's; Colombia and Moldova. One can ask "Why is Moldova the lowest?" and that would be a perfectly fine question to ask because throughout this report we have seen Moldova standing out in many cases. Second is the ranking of upper middle income countries. Is this surprising? Is this a coincidence? Readers should look in to this cornerstone by Kuznets (1955) for further explanations.

3.3.2 Income Inequality

Seeing the high income inequality, we were driven to provide a Lorenz Curve for the top four inequal countries; Colombia, Turkey, China and Russia in Figure 13. Then we will provide the most income equal countries' Lorenz; Moldova, Sweden Austria and Switzerland in Figure 14. The red area in Figure 14 represents the difference between Colombia and the country.

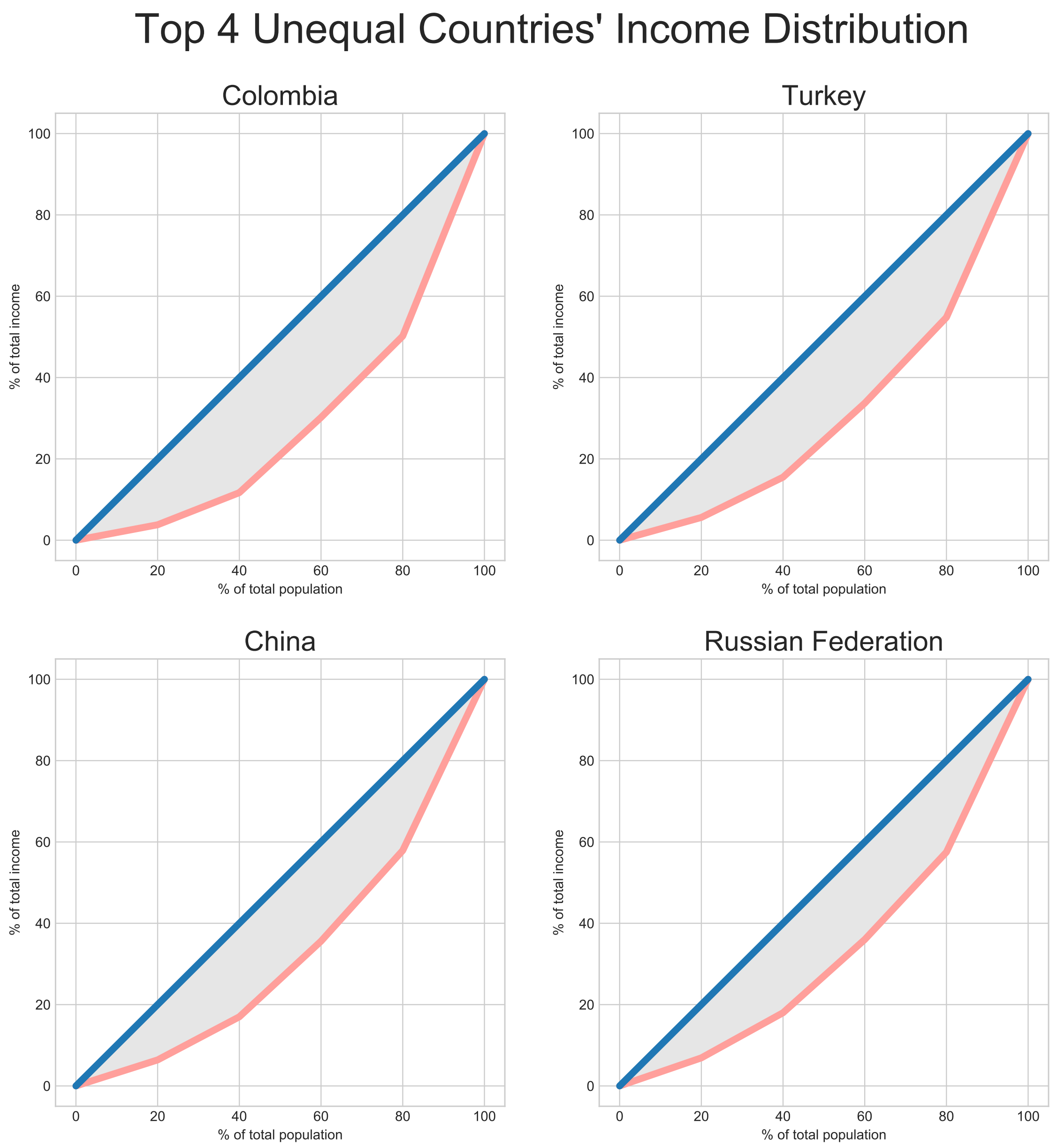


Figure 13

Readers can notice at ease, that, the smaller area between the two curves, the more “equally” distributed the national income is. They can also search the reasons of these inequalities as there are many papers such as, Rosser Jr et al. (2000) for transition economies in general and Duman (2008), Yang (1999) and Bourguignon et al. (2003) for more country specific analyses.

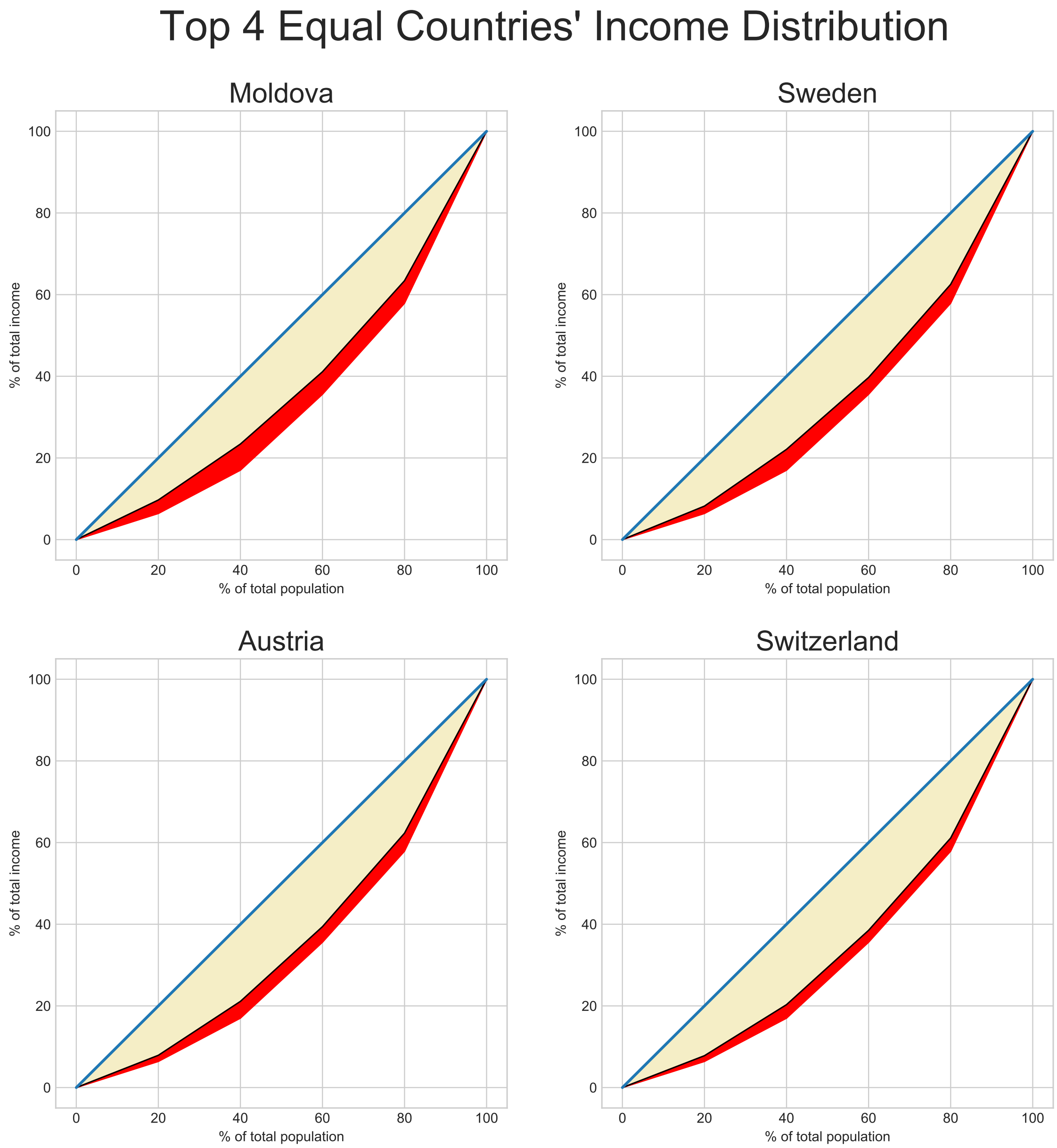


Figure 14

3.3.3 Poverty Lines

While income distribution is important to see a country's social welfare equality, to see their social welfare and their aggregate's welfare, such as throughout European Union (EU), we must look at some degree of measure over all countries in the desired pool.

Thus, we move to poverty lines. These three lines, \$1.9, \$3.2 and \$5.5 average income per day per person, are determined by World Bank itself in its data. One should also keep in mind that every countries' internal poverty lines are different but a common threshold brings relativeness and relatability.

Figures 15, 16 and 17 demonstrate how much of the total population's daily average income falls below the aforementioned thresholds. Reader will notice at first glance, or question, how the numbers are determined and how the translations between exchange rates are determined. To answer those questions we suggest reading about *Purchasing Pover Parity* concept in Shapiro (1983) and Grootaert (1997) as an example the poverty line determining process in Cote d'Ivoire.

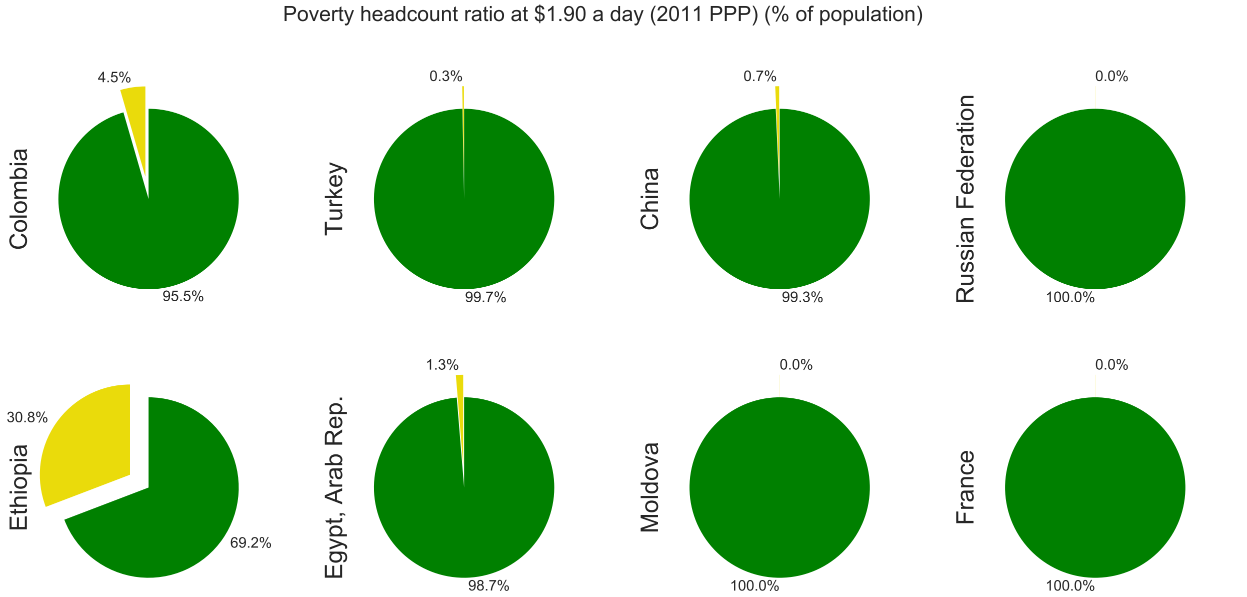


Figure 15

Graphs are generated using **Section 4.5.3** in the Notebook file.

Most valueable inference should be, income inequality does not necessarily mean a population below poverty lines. Because, when we look at Figure 17, we see that only 2.7% of Russia's population is under an average daily income of $5.50 and recall in Figure 12 that we saw Russia's GINI was the fourth highest one.

We should also notice Ethiopia, chosen to represent low income countries, has a very relatively poor population among others that are displayed. While this might be just a case of "poverty", this does not necessarily mean that Ethiopians are all in poverty. Because we have not provided any information on living standards, relative puchasing powers or any other measures for purchasing and living conditions. Thus, an average citizen in Ethiopia might buy a tomato for $0.02 while an average Swiss citizen might pay $3.00. We leave further discussion and research for the interested readers.

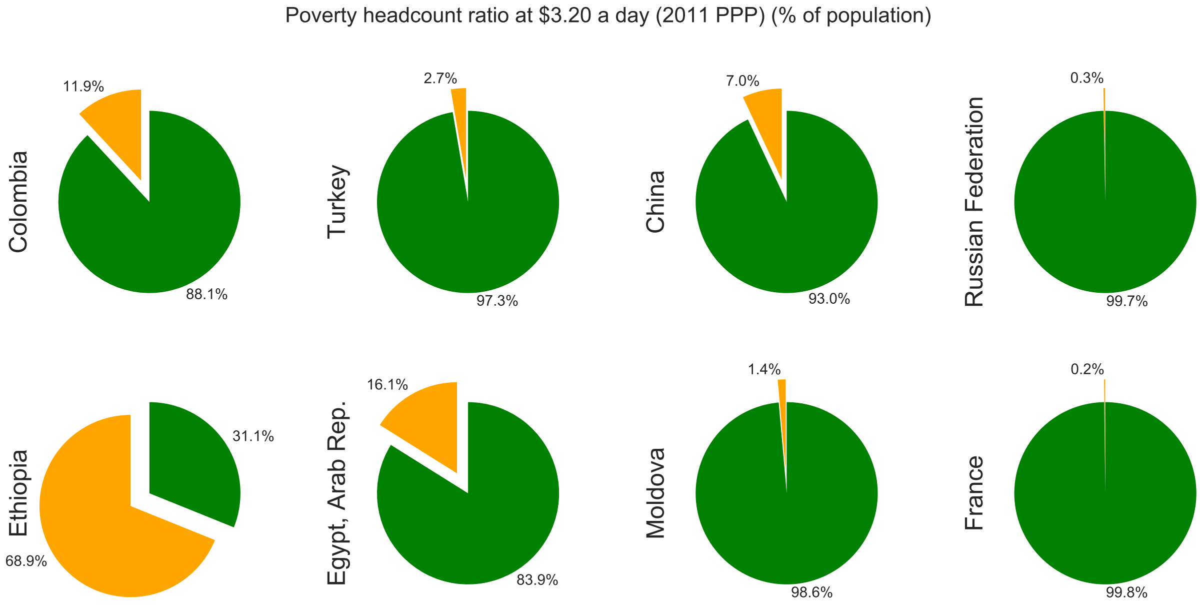


Figure 16

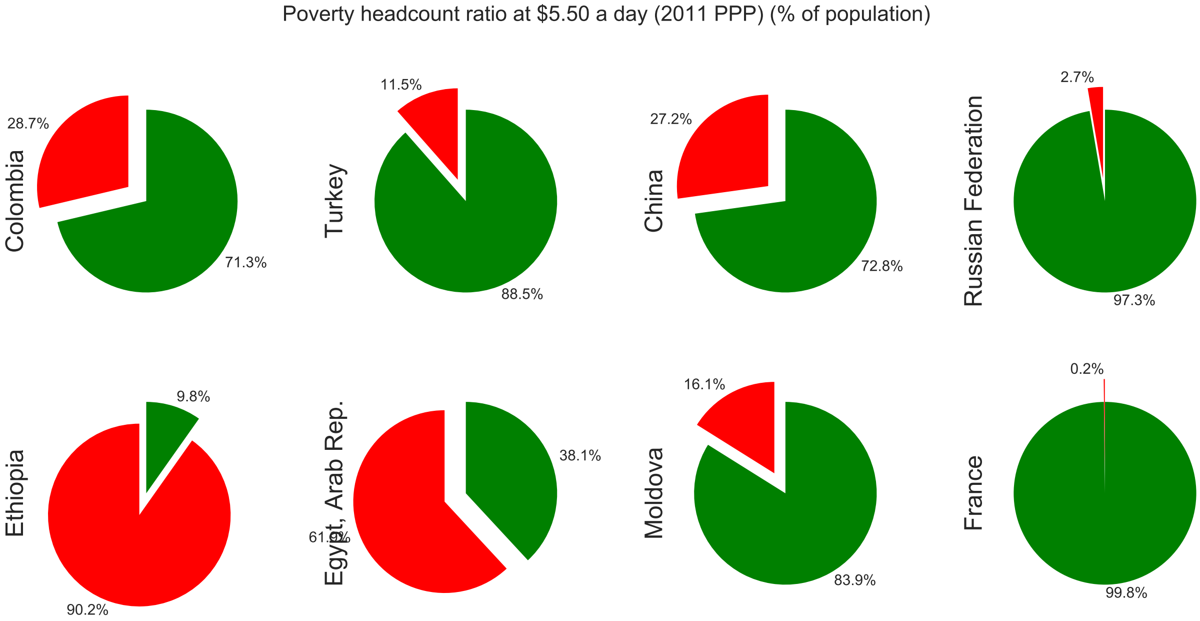


Figure 17

4 Conclusion

The research in the subtopics of economics are always emerging and creating more and more measures to explain and explore quantitative humanities. Our approach was to introduce the reader to our interesting findings, then enable them to conduct their own opinions and research toward their interests. For those purposes we have written a Jupyter Notebook and exported data from World Bank Data.

Our approach along the report was to create a guide to read along while skimming through the provided Notebook. Thus we have provided sectioning in the Notebook and labeled them throughout this report. Smaller code chunks are provided with figures through Section 1.

The casual findings from, population dynamics and pyramids, to more detailed findings, such as child mortalities vs skilled surgeons' attendance to births and GINI indices, are provided throughout Section 3. While provided graphs are visualizations only with no causality or correlation implemented, a reader's attention should be drawn in some topics. Thus, we provided both introductory papers, such as Wallerstein (1974), and more advanced papers, such as Bae and Bae (2004). The reader is recommended to at least skim through the given papers to get the basic understanding for concepts and for more detailed paper our recommendation should be taken as a guide for a researcher that is new in their topics or for those who are just curious.

5 Bibliography

Bader, P., Boisclair, D., and Ferrence, R. (2011) "Effects of tobacco taxation and pricing on smoking behavior in high risk populations: a knowledge synthesis," *International journal of environmental research and public health*, **8** (11), 4118–4139.

Bae, Y.-M. and Bae, C.-W. (2004) "The changes in the mortality rates of low birth weight infant and very low birth weight infant in Korea over the past 40 years," *Journal of Korean medical science*, **19** (1), 27–31.

Bourguignon, F., Nuñez, J., and Sanchez, F. (2003) "A structural model of crime and inequality in Colombia," *Journal of the European Economic Association*, **1** (2-3), 440–449.

Chaloupka, F. J., Hu, T.-w., Warner, K. E., Jacobs, R., and Yurekli, A. (2000) "The taxation of tobacco products," in: "In: Jha, P.; Chaloupka, F.(Eds.). Tobacco Control in Developing Countries," Citeseer.

Coale, A. J. and Hoover, E. M. (2015) *Population growth and economic development*, volume 2319, Princeton University Press.

Duman, A. (2008) "Education and income inequality in Turkey: does schooling matter?" *Financial Theory and Practice*, **32** (3), 369–385.

Ehrlich, P. R. and Holdren, J. P. (1971) "Impact of population growth," *Science*, **171** (3977), 1212–1217.

Gastwirth, J. L. (1972) "The estimation of the Lorenz curve and Gini index," *The review of economics and statistics*, 306–316.

Grootaert, C. (1997) "The Determinants of Poverty in Cote d'Ivoire in the 1980s," *Journal of African Economies*, **6** (2), 169–196.

Kuznets, S. (1955) "Economic growth and income inequality," *The American economic review*, **45** (1), 1–28.

MacDorman, M. F., Hoyert, D. L., and Mathews, T. (2013) *Recent declines in infant mortality in the United States, 2005-2011*, 120, US Department of Health and Human Services, Centers for Disease Control and . . . .

McKee, M. (1999) "Alcohol in Russia," *Alcohol and Alcoholism*, **34** (6), 824–829.

Reidpath, D. D. and Allotey, P. (2003) "Infant mortality rate as an indicator of population health," *Journal of Epidemiology & Community Health*, **57** (5), 344–346.

Rosser Jr, J. B., Rosser, M. V., and Ahmed, E. (2000) "Income inequality and the informal economy in transition economies," *Journal of Comparative Economics*, **28** (1), 156–171.

Shapiro, A. C. (1983) "What does purchasing power parity mean?" *Journal of International Money and Finance*, **2** (3), 295–318.

VanderPlas, J. (2016) *Python data science handbook: essential tools for working with data*, " O'Reilly Media, Inc.".

Wallerstein, I. (1974) *The Modern World-System I. Capitalist Agriculture and the Origins of the European World-Economy in the Sixteenth Century, With a New Prologue*, University of California Press.

Wouts, M. (2019) "PyProject, world-bank-data 0.1.3," `https://pypi.org/project/world-bank-data/`, accessed: 2019-11-25.

Yang, D. T. (1999) "Urban-biased policies and rising income inequality in China," *American Economic Review*, **89** (2), 306–310.