Hazel Erickson (57958128), Nam Vu (54781288),
Linh Luu (68697438), Priscilla Ursua (23457841)

DATA 201 Project Report

A. What data sources you used

1. http://data.un.org/
2. http://hdr.undp.org/en/data
3. https://www.imf.org/external/pubs/ft/weo/2019/01/weodata/index.aspx

B. Why you chose those data sources

The first 2 sources were chosen because of their wide range of statistical data on social, economic, environment and infrastructure indicators, and are credible sources associated with the United Nations (UN).

UNdata (first source) was launched as part of a project in 2005, called "Statistics as a Public Good", whose objectives was to provide free access to global statistics, to educate users about the importance of statistics for evidence-based policy and decision-making and to assist National Statistical Offices of Member Countries to strengthen their data dissemination capabilities. All data and metadata provided on UNdata's website are available free of charge and may be copied freely, duplicated and further distributed provided that UNdata is cited as the reference.
(From: http://data.un.org/Host.aspx?Content=About)

The aim of the Human Development Report (second source) is to stimulate global, regional and national policy-relevant discussions on issues pertinent to human development. Accordingly, the data in the Report require the highest standards of data quality, consistency, international comparability and transparency. The UNDP website

states that we are free to copy, redistribute in any format and build upon the material for any purpose, even commercially, as long as we give appropriate credit.
(From: http://hdr.undp.org/en/statistics/understanding)

The 3rd source was chosen so that our group could compare our own predicted data with actual prediction data from a reliable source. This data is from the World Economic Outlook (WEO) and they state that we are welcome to use WEO data for written work as long as we cite the publication/database accordingly.

C. What target you chose (i.e., what is the intended use of the data, ...)

Is Human Development Index (HDI) a good indicator of how well countries in the world are doing socially or economically?

The whole scope of our project will mainly be targeted at governments or other authoritative bodies around the world since the topics and data discussed are very relevant to policy-making and overall human development. The data that our group handles could stimulate debate on whether HDI is a useful statistical tool and whether it can be used to measure a country's overall achievement in its social and economic dimensions.

An emphasis of our project was comparing a country's Human Development Index (HDI) with other social and economic factors, to see whether there was a correlation. According to the UN, The Human Development Index (HDI) is a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and have a decent standard of living. The HDI is the geometric mean of normalized indices for each of the three dimensions.

By comparing the correlation between HDI and these factors, it can show us whether HDI is a reliable indicator for how well countries are doing socially or economically, for

factors other than the three used to form the index (Life expectancy, Education Index and GNI Index)

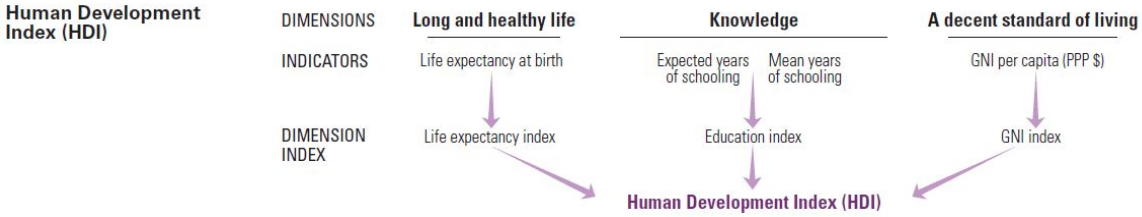


Image Source: http://hdr.undp.org/en/content/human-development-index-hdi

The UN also states that HDI only captures part of what human development entails. It does not reflect on other factors such as inequality, poverty, human security and empowerment. We will be taking a closer look at factors such as inequality, to see if there is some form of correlation between those countries that are considered more developed, and these factors that the HDI has not taken into account.

D. What difficulties you had to overcome to wrangle the data sources into the target data model

It was messy attempting to plot all the countries in the data from the UN website using ggplot and difficult to find a meaningful way to display correlation between the HDI index. For this reason, we chose to select specific countries to provide a representation of the differences between higher and lower HDI indexes.

40 countries were chosen, and to pick these, we used the 10 countries with the highest HDI index, 10 countries from the top of the 66th percentile, 10 countries from the top of the 33rd percentile and the 10 countries with the lowest HDI index. We wrangled the data for these four groups, so we could graph them individually and use different colors to distinguish between these groups, and show correlation between higher and lower HDI indexes.

Many of the data from the UN had missing countries or values, so we had to omit these from the source data before graphing. Also, it was quite difficult for us to summarize the topic based on figures.

Some categories had data collected in different time gaps so we had to use the years that they had in common to compare them.

We were unable to open some csv files if we kept its title, so we had to remove the first row of each csv file.

E. What techniques you did use

The UN website has its data available to download in CSV format, so we downloaded and imported these files into jupyter lab using read_csv. For each dataset we used, the data was read into an R dataframe, sorted by HDI rank and converted into a long dataframe using the gather function. From here, the data could be graphed using ggplot(). The ggplot2 cheat sheet provided a lot of information on how to build our graphs and change fill colors, add layers and change the graph theme. (https://rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf)

For the data where it had been gathered over time, we used line graphs to show the change in the y-values over time, against the years on the x-axis. The four groups have been distinguished using a different color for each group, with a legend to interpret. For some other data via the prediction of unemployment next four years, even though they have the csv file to download, we actually used scraping tools to collect the data instead and implemented several utilities in rvest package to extract data and clean up errors.

We used github to collaborate and track changes to the code while we wrangled our data and produced the graphs.

F. What you managed to achieve and what you failed to do

a. Economic Indicators

Original data for income index

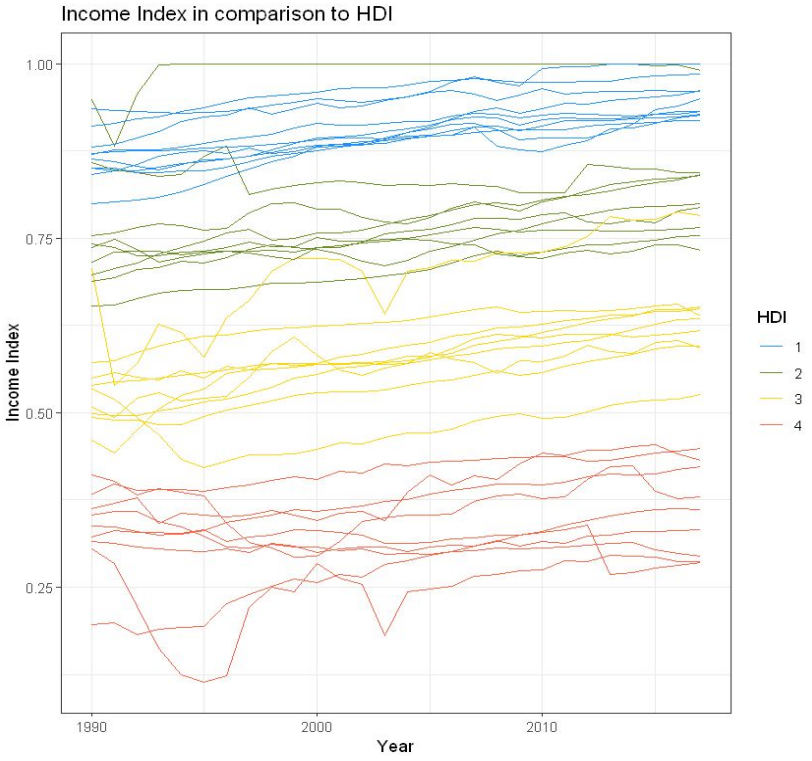
HDI	Country	1990	1991	1992	1993	1994	1995	1996	1997	...	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
168	Afghanistan	NA	NA	NA	NA	NA	NA	NA	NA	...	0.388	0.413	0.421	0.426	0.442	0.443	0.441	0.439	0.438	0.439
68	Albania	0.575	0.523	0.510	0.531	0.547	0.570	0.585	0.568	...	0.683	0.686	0.693	0.699	0.700	0.705	0.707	0.711	0.717	0.722
85	Algeria	0.694	0.686	0.686	0.681	0.676	0.677	0.680	0.680	...	0.730	0.730	0.734	0.734	0.734	0.734	0.737	0.739	0.744	0.744
35	Andorra	0.904	0.902	0.897	0.891	0.890	0.891	0.897	0.910	...	0.931	0.924	0.916	0.910	0.910	0.914	0.920	0.924	0.927	0.931
147	Angola	0.456	0.453	0.450	0.403	0.363	0.443	0.468	0.485	...	0.583	0.613	0.603	0.601	0.602	0.613	0.620	0.625	0.617	0.613
70	Antigua and Barbuda	0.755	0.762	0.762	0.769	0.775	0.764	0.770	0.776	...	0.822	0.801	0.790	0.784	0.786	0.787	0.792	0.797	0.803	0.806

Wrangled long dataframe for the most developed countries

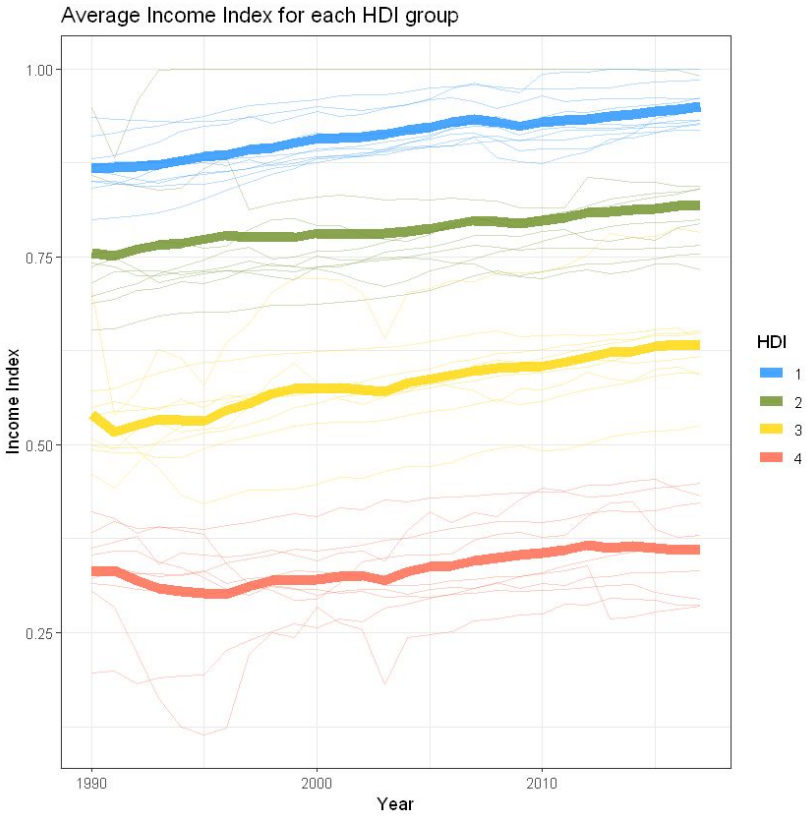
Country	Year	Factor
Norway	1990	0.911
Switzerland	1990	0.936
Australia	1990	0.850
Ireland	1990	0.799
Germany	1990	0.870
Iceland	1990	0.850

The first thing we investigated was whether there was a correlation between income index and HDI. The data was downloaded from the UN website.

The UN website describes income index as the GNI per capita (2011 PPP International $, using natural logarithm) expressed as an index using a minimum value of $100 and a maximum value $75,000. The GNI represents the value produced by a country's economy in a given year, regardless of whether the source of the value created is domestic production or receipts from overseas. We would expect that there would be a clear correlation between a country's HDI and its income index, especially considering that one of the three components of calculating the HDI index is GNI per capita, however these countries will still differ in terms of life expectancy and education.

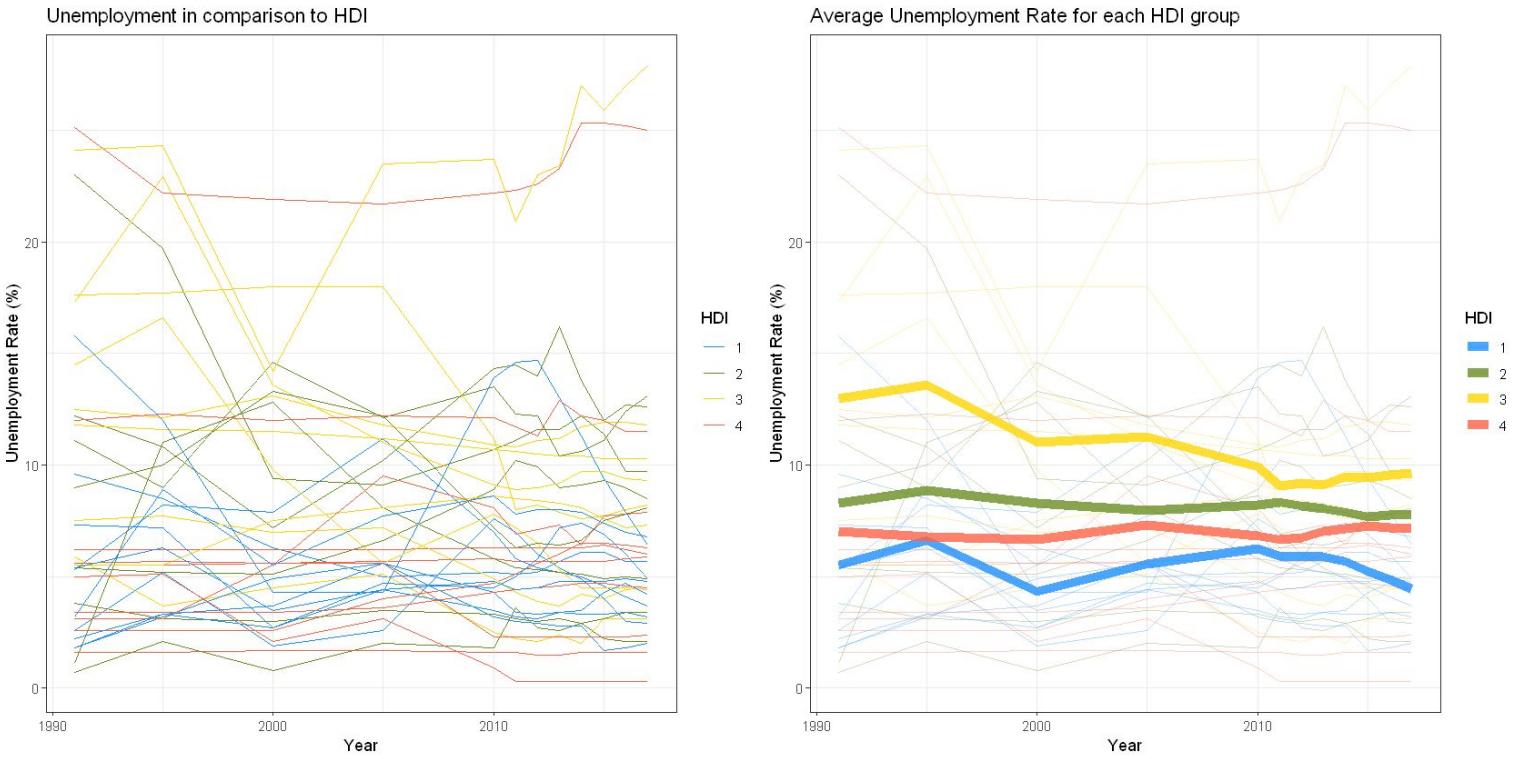


From plotting the data in R, there is a visible correlation between the Income Index and the HDI, as expected. The Income Index appears to generally linearly increase over time, which is expected as GNI increases, and we would expect this trend to continue. However, the countries in the lowest HDI bracket appear to not have as much as an increase in growth as the more developed countries.



The second graph shows the HDI as an average for each group using a bold line. It shows the strong correlation that the countries with higher HDI have higher GNI per capita.

Another factor we chose to look at was Unemployment. The data used here was from a csv file from the UN website. This data indicates the percentage of the labour force population ages 15 and older that is not in paid employment or self-employed but is available for work and has taken steps to seek paid employment or self-employment. We took this data and plotted it for the same 40 countries, but for the unemployment rate rather than the income index. This generated a plot that didn't show a clear pattern between the groups.

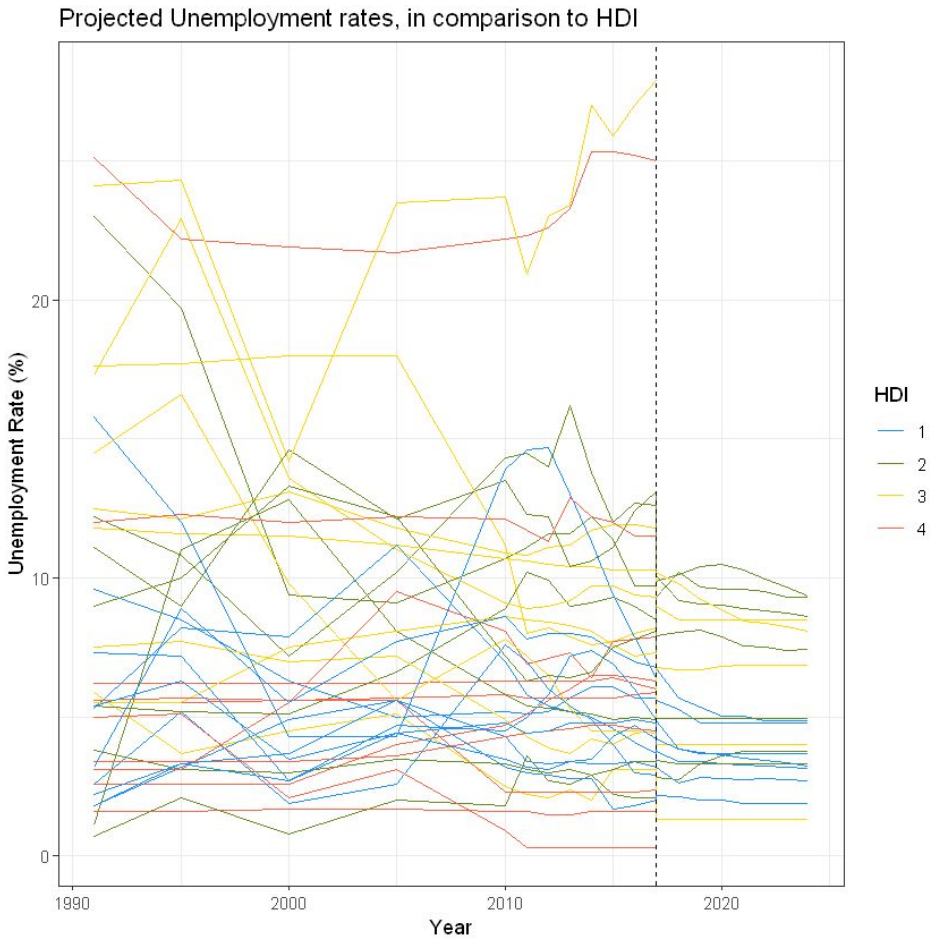


The plot which shows the average Unemployment rate for each HDI group also shows that there is no major link between HDI and Unemployment rates.

Unemployment is not taken into account when calculating the HDI index, however it is an extremely important factor for the country's economy. Lower unemployment means optimal levels of production, easier job access and less government borrowing. In fact, one of the UN's targets is by 2030, to achieve full and productive employment and

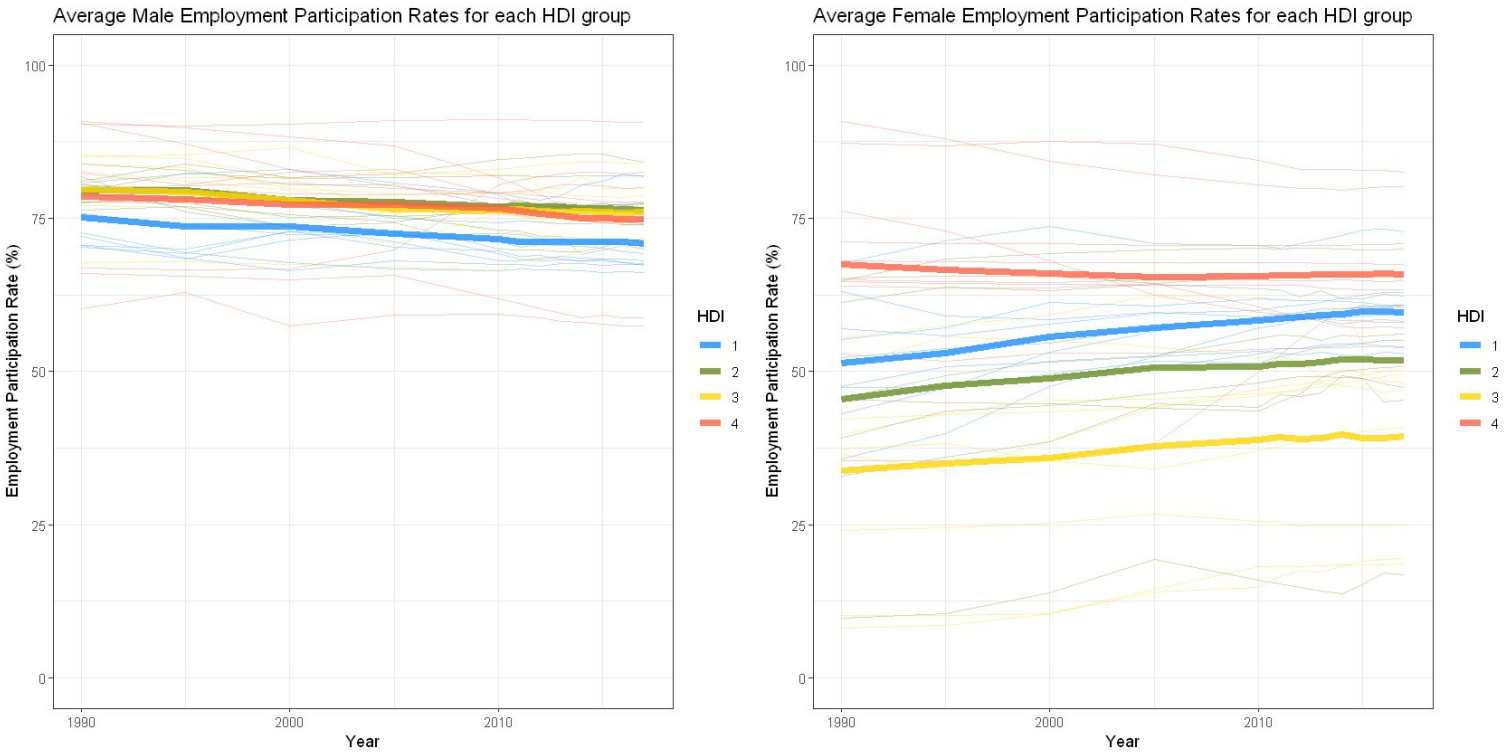
decent work for all women and men, including for young people and persons with disabilities, and equal pay for work of equal value.

It is clear that this is a major issue the UN wants to focus on, but the Human Development Index fails to take unemployment into account. Since there is no clear correlation, we might assume that a major key reason each country's unemployment differs is based on the local government, as unemployment is impacted by factors such as raising the minimum wage, job outsourcing and cyclical unemployment during times of economic crisis.



We also scraped data from the World Economic Database to join to our Unemployment Dataframe as a side project to our main question, then used this information to graph the projected unemployment rates until 2024. Although there was a lot of missing data it shows a projection for unemployment rates in the future, especially for countries with higher HDI.

We also graphed the labour force participation rate for males and for females.



There is not a visible correlation between HDI and each genders workforce participation rate here. It is clear that the overall trend is lower female workplace participation rates in many countries. Equal workplace participation and equality is important as it is associated with improved national productivity, economic growth, increased organisational performance, enhanced ability of companies to attract talent and retain employees and enhanced organisational reputation.

The next factor we chose to graph was income inequality. Income inequality is the measure of the deviation of the distribution of income among individuals or households within a country from a perfectly equal distribution. A value of 0 represents absolute equality, a value of 100 absolute inequality. We chose The Gini index as it's the most used measure of inequality. It looks at the distribution of a nation's income or wealth, where 0 represents complete equality and 100 total inequality.

As mentioned earlier, HDI does not reflect on inequality or the distribution of wealth throughout the country. In fact, many developed countries have issues with income inequality and a growing gap between the country's richest and poorest. Many people living in developed countries are living in poverty while the wealthiest people continue to get richer. Our graph for income inequality is a bar graph that shows the slight correlation between countries that are considered more developed and having less inequality. However, there are particular countries in each group which stand out for having a much greater level of inequality, and this is not something that is considered in the HDI.

Inequality is a major issue - Unequal societies are less functional, less cohesive and less healthy than their more equal counterparts. So it is surprising that the most widely used and accepted index used to rank countries by human development does not use income inequality as a composite of the index.

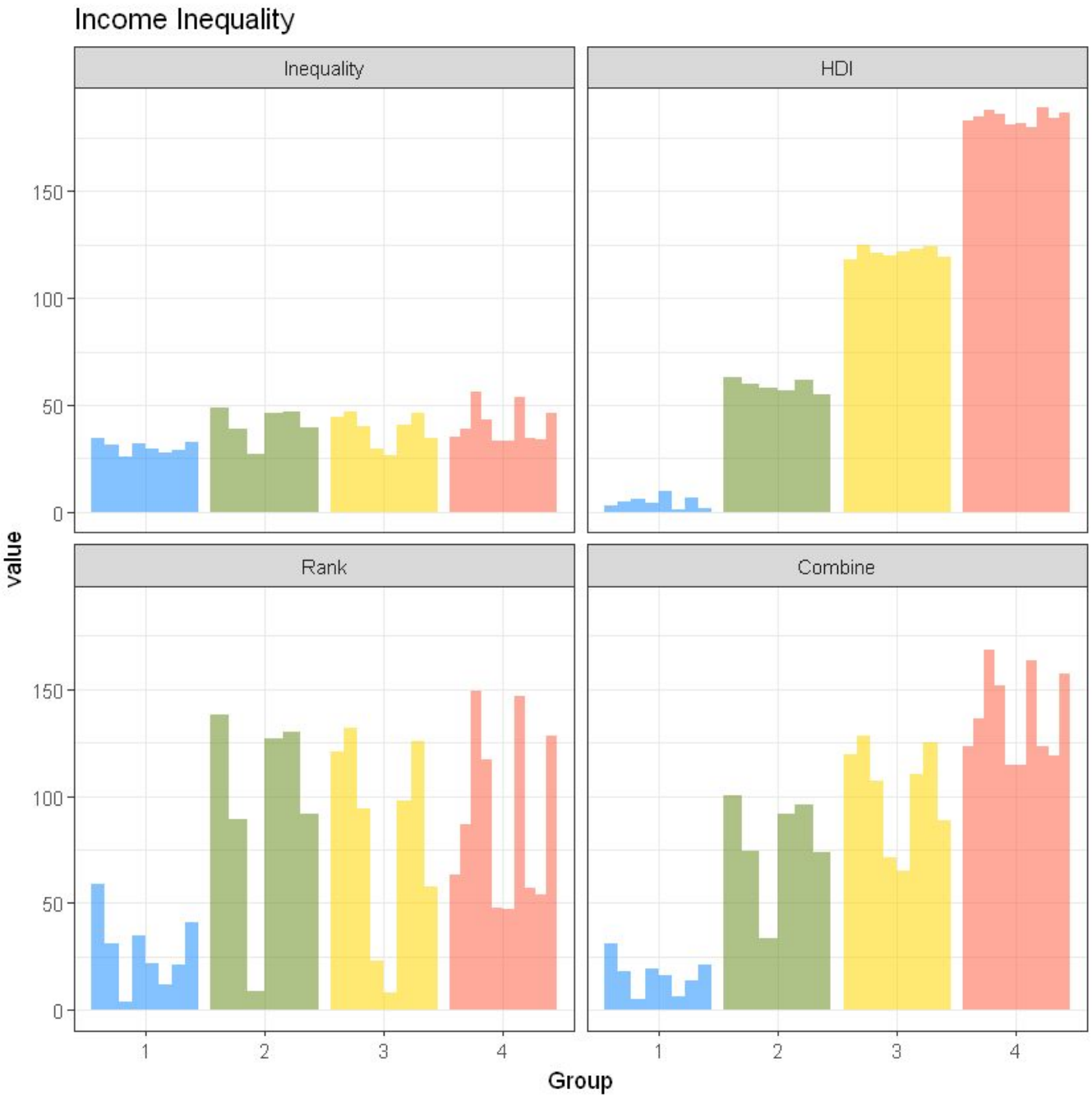
For this reason, we chose to attempt to alter the HDI index for these 40 countries to take into account the inequality index. Shown below is a bar graph for the HDI indexes, followed by a graph for the income inequality. We ranked each country from 1-189 for the lowest to highest income inequality, so we had rankings for each country's HDI and income inequality. Then we took the average of these two ranks, to show an example of what HDI might look like if it had been adjusted for income inequality.
The original and wrangled data frame.

head(income_inequality)

HDI	Country	2010-2017
68	Albania	29.0
85	Algeria	27.6
147	Angola	42.7
47	Argentina	42.4
83	Armenia	32.5
3	Australia	34.7

head(income_inequality_highest)

Country	Inequality	HDI	Rank	Combine	Group
Australia	34.7	3	59	31.0	1
Germany	31.7	5	31	18.0	1
Iceland	25.6	6	4	5.0	1
Ireland	31.9	4	35	19.5	1
Netherlands	29.3	10	22	16.0	1
Norway	27.5	1	12	6.5	1



Inequality - The inequality index for each country (0-100%)

HDI - The HDI rank of each country (Ranking each country from 1-189)

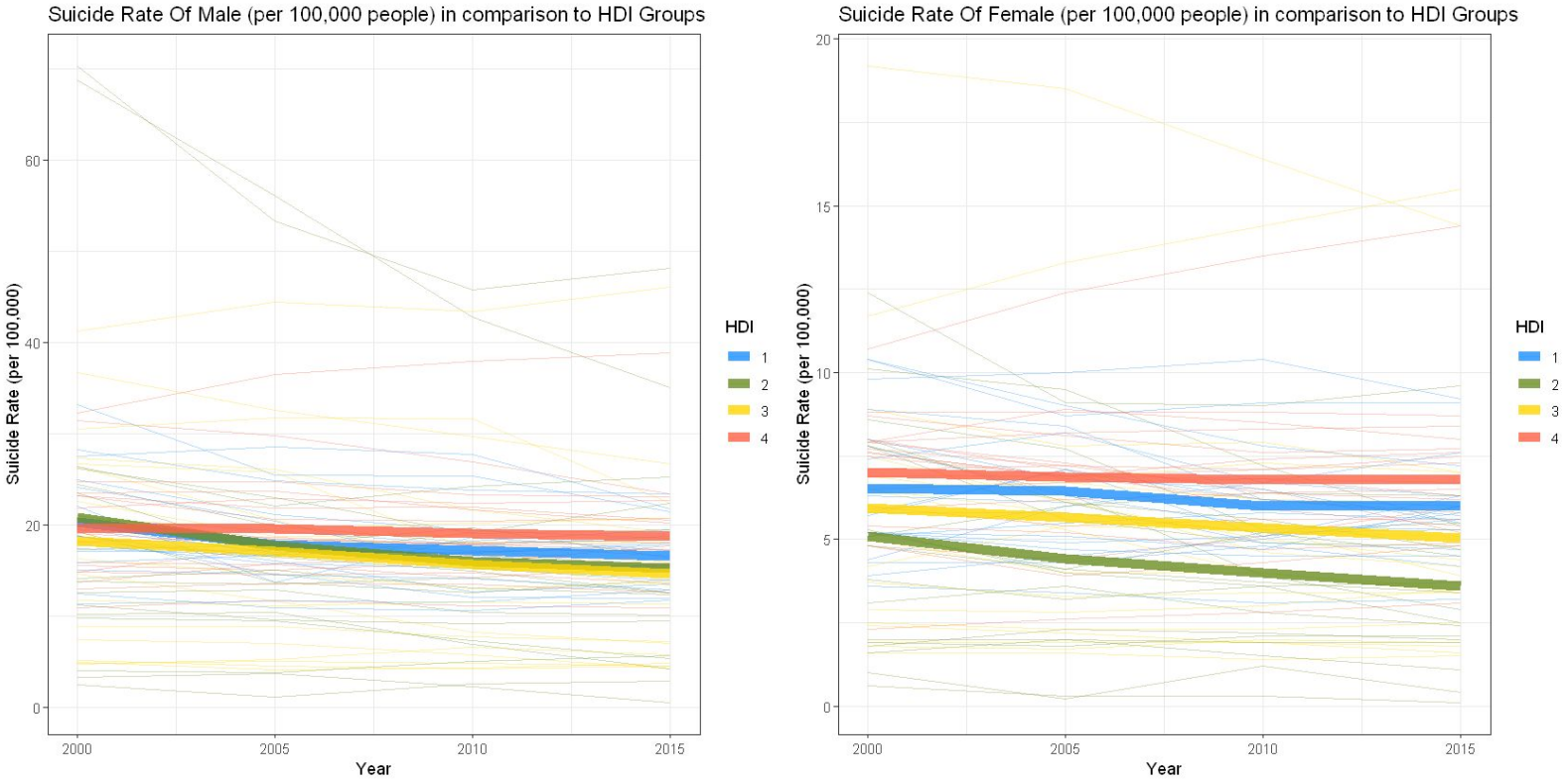
Rank - The Inequality Index rank of each country (Ranking each country from 1-189)

Combine - Our calculated average of HDI rank and Income Inequality Index rank

b. Social Indicators

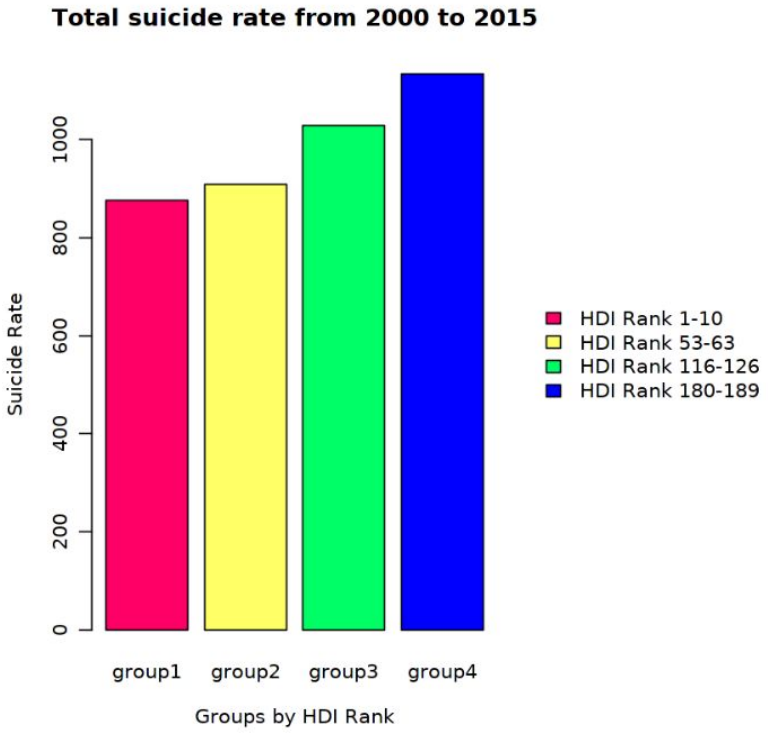
We chose suicide rates and homicide rates as social factors to observe. These factors reflect the reality of the social status of a nation and is one of the key factors to have a look at how well governors handle the human security in their countries. Our target is still based on four groups of human development index in countries from highest to lowest.

In suicide rate, firstly, we explored the relation of the independent variable with gender and HDI. Generally, it was pretty hard to get the overall trend of HDI in these graph, however, we acknowledged that the male suicide rate was significantly higher than the women suicide rate from the scale we got. This may be due to some external effects such as men usually using more lethal methods, unlike religion or race in different societies.



Next, we tried to make a bar chart to see the correlation between HDI levels in selected groups and the incidence of suicide. As the definition of HDI, our expectation was that countries with higher poverty levels and an inadequate social security system in place will have higher suicide rates. In other words, HDI is expected to be inversely propotional to the suicide rate. Due to missing values in some countries, we had to pick another countries belonging to the same group, for example, Denmark replaced to Hong Kong which also has HDI greater than 0.9 percentile. And as expected, from the data frame we sorted, the increase in HDI were associated with decreasing suicide rates.

Category	Suicide Rate
<chr>	<dbl>
group1	876.3
group2	909.1
group3	1028.2
group4	1135.2

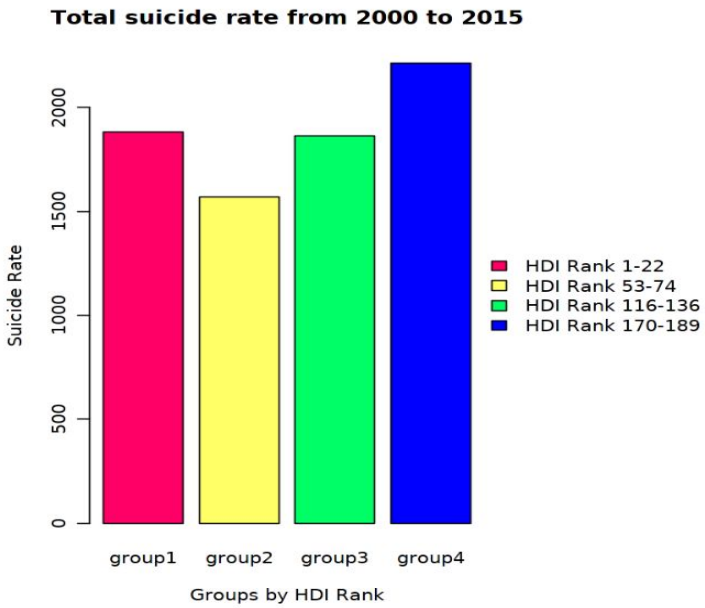


However, surprisingly, when we attempted to plot with groups of larger observations, about twenty and thirty countries per group, we found very interesting information. The first group has a much higher rate than the second and third group, unlike the first plot we did. Therefore, we guessed the loss of values in general and value in Hong Kong (in group one) in particular, might affect the result of the plot and be the reason why this

error occurred (the first plot as an evidence). This can be thought of as one of the failures that we have experienced in this project.

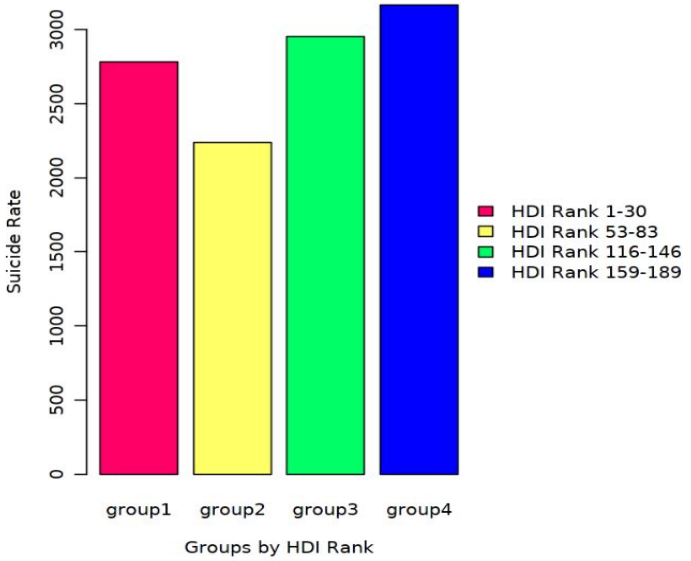
Total suicide with 20 countries per group

Category	Suicide Rate	HDI
<chr>	<dbl>	<chr>
group1	1882.5	HDI Rank 1-22
group2	1567.9	HDI Rank 53-74
group3	1863.4	HDI Rank 116-136
group4	2210.9	HDI Rank 170-189

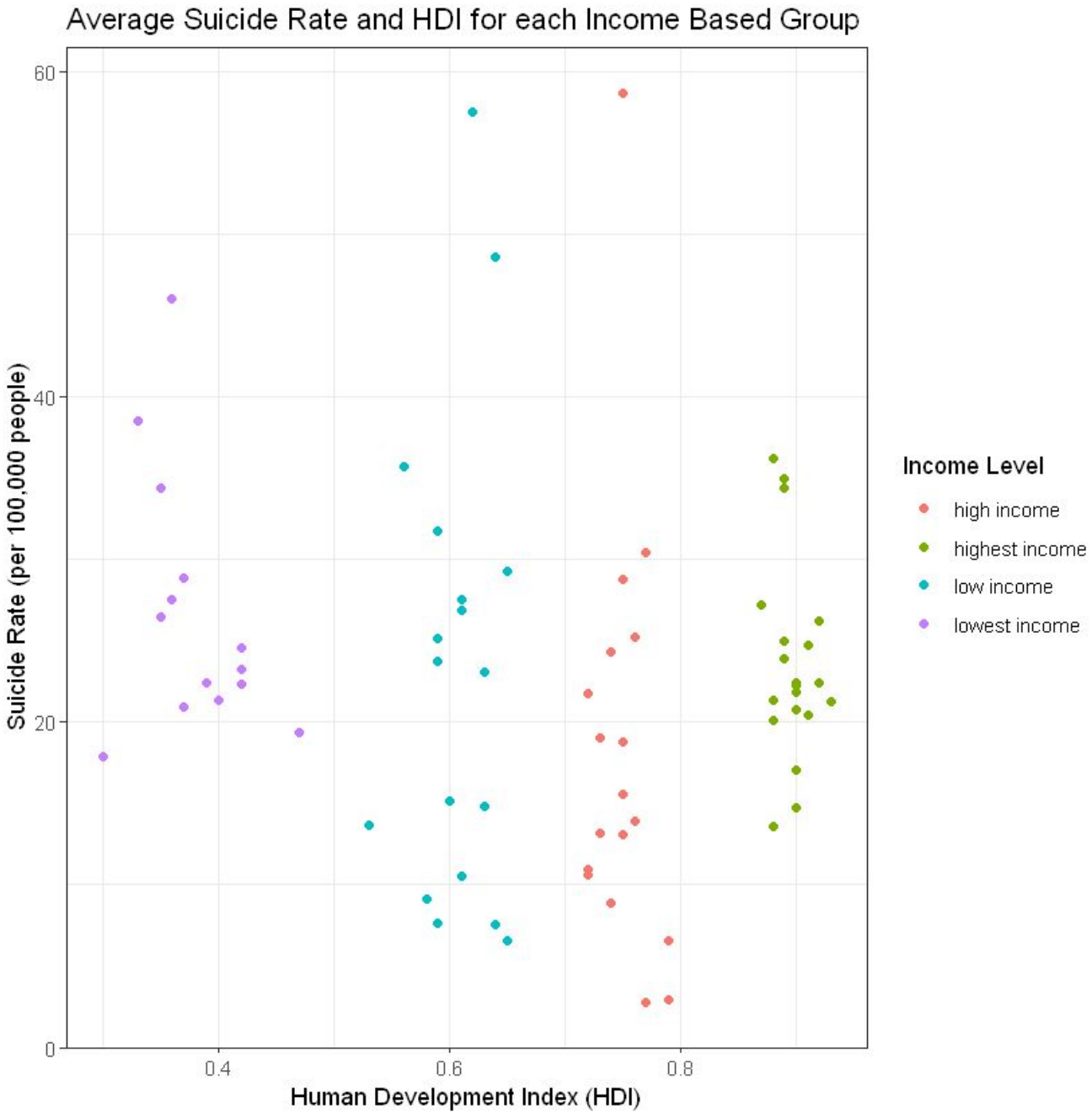


Total suicide with 30 countries per group

Category	Suicide Rate	HDI
<chr>	<dbl>	<chr>
group1	2783.7	HDI Rank 1-30
group2	2237.0	HDI Rank 53-83
group3	2953.0	HDI Rank 116-146
group4	3169.9	HDI Rank 159-189



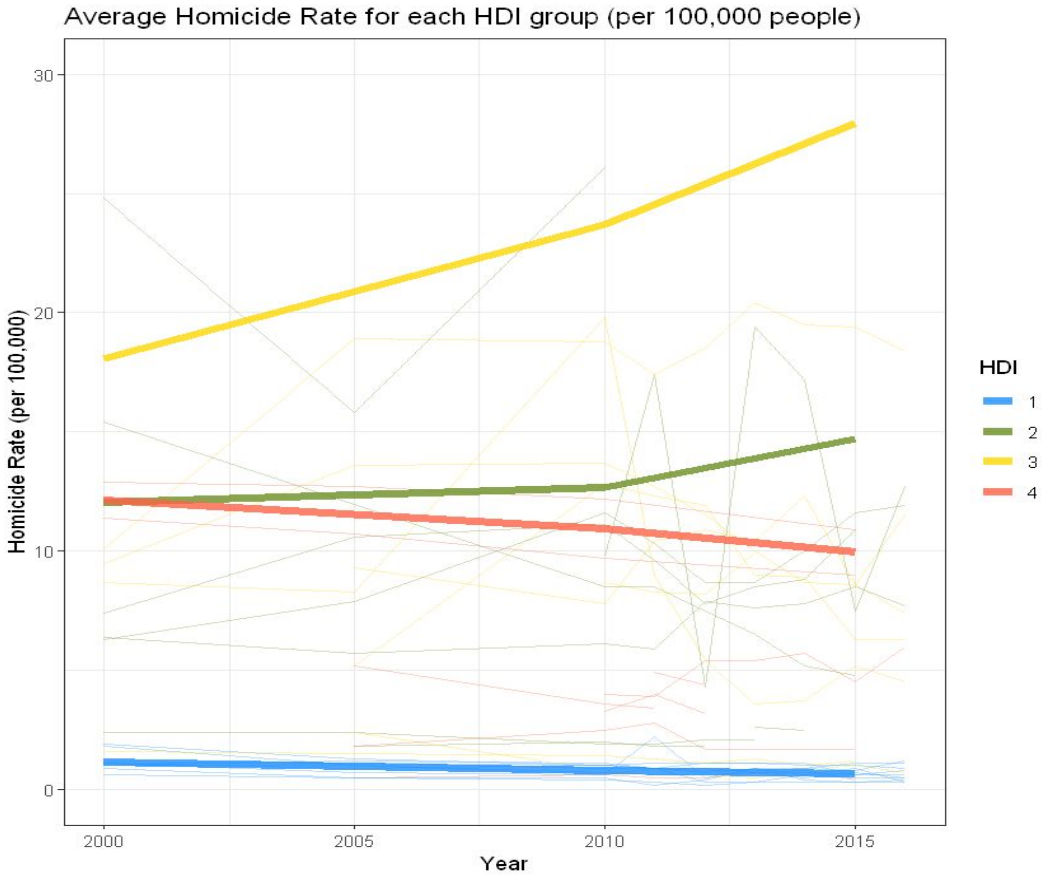
Continously, we tried to discover correlation between suicide rate and HDI factors which are human development and income index by using a scatter-plot in four groups of twenty countries instead of its ranking. We observed that there was a very low correlation in group one which contains countries with the highest income and also three outliers lying outside the range which could be Japan, Korea and Slovenia. So we guessed they likely were an answer for the higher suicide incidence in group one significantly higher than group two and three.



Moving on, another factor we chose to is homicide rate. We will also try to find whether there is a correlation in the relationship between a country's homicide rate and its HDI.

159	Mauritania	13.7	12.4	10.9	..	..	..	..	9.9	..
161	Madagascar	11.4	10.0	9.2	..	..	..	..	7.7	..
162	Uganda	..	8.7	9.3	10.7	11.2	10.3	11.5	..	..
163	Benin	7.3	7.2	6.6	..	..	..	..	6.2	..
164	Senegal	10.0	9.3	8.5	..	..	..	..	7.4	..
165	Comoros	10.4	9.5	8.5	..	..	..	..	7.7	..
165	Togo	10.5	10.7	9.6	..	..	..	..	9.0	..
167	Sudan	..	..	..	..	..	..	..	..	..
168	Afghanistan	..	..	3.4	4.1	6.3	..	..	..	..
168	Haiti	..	..	6.8	9.0	10.0	..	..	..	..

After we had wrangled data from the UN website, we discovered that there was a lot of missing data in low and lowest HDI countries as seen in the dataframe above. Therefore, we decided to take the average homicide rate of each group and remove all missing data to make the graph more accurate.



After that, we tried to make a graph that plots the average homicide rate for each HDI group. As expected, the graph shows that the group with the highest HDI is at the bottom with the lowest average homicide rate. However, there is no relationship between homicide rate and HDI indicator for the less developed countries. The reason for this might be homicide rate strongly depends on the mental health of the people, and security in each country.

G. Conclusion

Only certain indicators that we chose displayed a direct correlation to the countries HDI index. For the economic indicators, there was a strong correlation between Income Index and HDI, as generally countries with a higher HDI have a higher Income Index. This is expected as GNI (Gross National Income) is used as one of the indicators to calculate the HDI. However other important economic indicators such as Unemployment rates, Labour force participation rates and Income Inequality have little to no correlation with HDI. This is important to note as the index is used to determine the level of human development in each country, and it is used widely to tell policymakers and citizens how well a country is doing, but it is not a good indicator for all economic issues.

This also held true for the social indicators as we did not find a correlation between Suicide Rates and HDI, and only a minor correlation between countries that are in the highest range of development having lower Homicide Rates. Human security is however an important and useful concept regarding people's safety and livelihood and should not be overlooked when considering how developed a country may be.

In conclusion, using HDI as an indicator for the economic or social status of a country is advantageous in many contexts but focuses only on certain aspects of development. Researchers must keep in mind that it has a limited range and it cannot provide a complete view of human development and therefore, a more useful measure would be a combination of HDI and other useful indicators to get a more accurate view.

REFERENCE LIST

Amadeo, K (Jan, 2019). *Seven causes of Unemployment*. Retrieved from https://www.thebalance.com/causes-of-unemployment-7-main-reasons-3305596

Australian Government (n.d.). *About Workplace Gender Equality*. Retrieved from https://www.wgea.gov.au/topics/about-workplace-gender-equality

Human Development Index (HDI) (n.d.). Retrieved from http://hdr.undp.org/en/content/human-development-index-hdi

SDG Indicators (n.d.). Retrieved from https://unstats.un.org/sdgs/metadata/?Text=&Goal=8&Target=8.5

The guardian (n.d.). *Inequality index: Where are the world's most unequal countries?.* Retrieved from https://www.theguardian.com/inequality/datablog/2017/apr/26/inequality-index-where-Are-the-worlds-most-unequal-countries

United Nations Development Programme (n.d). *Homicide rate (per 100,000 people).* Retrieved from http://hdr.undp.org/en/indicators/61006

United Nations Development Programme (n.d.). *Labour force participation rate (% ages 15 and older), female.* Retrieved from http://hdr.undp.org/en/indicators/48706

United Nations Development Programme (n.d.). *Labour force participation rate (% ages 15 and older), male*. Retrieved from http://hdr.undp.org/en/indicators/48806

United Nations Development Programme (n.d.). *Income Index Data*. Retrieved from http://hdr.undp.org/en/indicators/103606

United Nations Development Programme (n.d.). *Income inequality, Gini coefficient.* Retrieved from http://hdr.undp.org/en/indicators/67106

United Nations Development Programme (n.d). *Suicide rate, female (per 100,000 people).* Retrieved from http://hdr.undp.org/en/indicators/112606

United Nations Development Programme (n.d). *Suicide rate, male (per 100,000 people).* Retrieved from http://hdr.undp.org/en/indicators/112506

United Nations Development Programme (n.d). *Unemployment, total (% of labour force).* Retrieved from http://hdr.undp.org/en/indicators/140606