What are the relationships between Socioeconomic Status factors and HIV/AIDS infection?

A data wrangling project report

By project group DW007

1. Junye Qu [jqu200] [2575022]
2. Onur Mavitas [oms270][2675618]
3. Megan van den Brink [mbk357][2605809]

Research question

Some research claims that it is possible to observe correlations between wealth, household income, education and HIV prevalence. It is also mentioned that conditions that minorities lacking in certain socioeconomic factors have to deal with create inadequacies for HIV prevention. (Pellowski et al., 2013). Socioeconomic status factors (SES) are what makes up a person their status socially and economically. Those factors include but are not limited to income, education, occupation, gender, sexuality, wealth ethnicity and social class.
Therefore, the research question of this project is:
What are the relationships between Socioeconomic Status factors and HIV/AIDS infection?
The project is focused on region, country, wealth, sexuality, availability of healthcare, gender and their correlation to HIV.

Data sources

AIDS | UNAIDS last accessed on : 16-01-2020

- Men who have sex with men
- People who know their status (%)

World Health organization – HIV/AIDS last accessed on : 16-01-2020

- Number of people (all ages) living with HIV Estimates by country
- Number of new HIV infections Estimates by country
- Number of deaths due to HIV/AIDS Estimates by country

The World Bank last accessed on: 22-01-2020

- GDP (current US$)
- Prevalence of HIV, total (% of population ages 15-49)
- Population
- Prevalence of HIV, total (% of population ages 15-49)

- GINI index for income Inequality (World Bank estimate)
- Incidence of tuberculosis (per 100,000 people)
- Unemployment, total (% of total labor force) (modeled ILO estimate)
- School enrollment, secondary (% gross)

OECD| Data|Violence against women last accessed on: 20-01-2020

Maps - Sexual orientation laws last accessed on: 24-01-2020

Data wrangling methods

Data acquisition

For data acquisition multiple research papers were read, analyzed for import factors that have an impact on the odds of acquiring HIV. It was important to know what other data analyses were done on the subject to have an idea of what other types of connections could be discovered. Those same papers documented what datasets or websites were used, out of which reliable sources could be chosen.

Because the websites used were public, from institutions that are deemed trustworthy, and formatted well, it was not needed to use techniques involving JSON, XML or web crawling. Most of the datasets used were in Comma Separated Values format, except for the dataset stating countries their legal stance on same sex marriage.

Data cleaning

String manipulation

The first data cleaning method executed was string manipulation. What was necessary was to turn records that were strings into actual integers. This was because if a column used special symbols, e.g. <500, all of the records within that column are strings. Hence, to clean this data, it was necessary to get rid of symbols and whitespace; and then make those strings the appropriate data type.

	Country	2018	2010
0	Afghanistan	7200 [4100–11 000]	4200 [2500–6200]
1	Albania	No data	No data
2	Algeria	16 000 [15 000–17 000]	7100 [6600–7600]
3	Angola	330 000 [290 000–390 000]	220 000 [180 000–250 000]
4	Argentina	140 000 [130 000–150 000]	110 000 [96 000–120 000]
5	Armenia	3500 [3000–4400]	3300 [2800–4100]
6	Australia	28 000 [23 000–31 000]	21 000 [17 000–23 000]
7	Austria	No data	No data
8	Azerbaijan	No data	No data
9	Bahamas	6000 [5300–6700]	5800 [5100–6600]
10	Bahrain	No data	No data

Figure 1. Number of people suffering from HIV per country per year

Replacing values

In the records there were estimations, e.g. [1000-1500], of numbers. Because the dataset had the median out of those estimations in the record, a function was written to omit the estimation and make the record the median.
Other types of values that had to be replaced were None values, as can be seen in figure one those were written down as No Data, or in other datasets as '…', and thus had to be replaced.

Lastly, it was needed to give the dataframe proper column names, as can be seen above the columns used were semantically insufficient for a dataframe encompassing more records.

Data merging & aggregation

	CODE	Population 2000	Population 2010	Population 2018	Country	Number of people (all ages) living with HIV in 2018	Number of people (all ages) living with HIV in 2010	Number of people (all ages) living with HIV in 2000	Number of new HIV infections in 2018	HIV incidence rate (per 1000 uninfected population) in 2018	Estimated antiretroviral therapy coverage among people living with HIV (%) in 2018	Reported number of people receiving antiretroviral therapy in 2018	Number of deaths due to HIV AIDS in 2018
0	AFG	20779953.0	29185507.0	37172386.0	Afghanistan	7200.0	4200.0	1600.0	840.0	0.02	13.0	60.0	500.0
1	AGO	16395473.0	23356246.0	30809762.0	Angola	330000.0	220000.0	87000.0	28000.0	1.01	27.0	4800.0	14000.0
4	ARG	36870787.0	40788453.0	44494502.0	Argentina	140000.0	110000.0	64000.0	6500.0	0.15	61.0	1700.0	1700.0
5	ARM	3069591.0	2877319.0	2951776.0	Armenia	3500.0	3300.0	950.0	200.0	0.06	53.0	NaN	200.0
7	AUS	19153000.0	22031750.0	24992369.0	Australia	28000.0	21000.0	13000.0	1000.0	0.04	83.0	NaN	200.0
10	BDI	6378871.0	8675602.0	11175378.0	Burundi	82000.0	93000.0	130000.0	1700.0	0.16	80.0	3400.0	1900.0
12	BEN	6865951.0	9199259.0	11485048.0	Benin	73000.0	61000.0	47000.0	3800.0	0.34	61.0	2000.0	2200.0
13	BFA	11607942.0	15605217.0	19751535.0	Burkina Faso	96000.0	110000.0	140000.0	2400.0	0.12	62.0	1900.0	3300.0
14	BGD	127657854.0	147575430.0	161356039.0	Bangladesh	14000.0	7700.0	940.0	1600.0	0.01	22.0	130.0	580.0

Figure 2. partial dataframe after cleaning and merging multiple datasets

The first idea was to use the country names as the column to merge the datasets on. After realizing that a significant amount of effort will be required to unify country names written differently in each dataset, three letter country codes were merged to the main data frame to be used as a basis for the merge operation. Inner and outer merging were done depending on the needs of the project. Useful features out of thirteen datasets from the above mentioned sources were picked and merged. Besides that certain features such as population and number of people living with HIV were used to create new, more useful features such as percentage of people in a country with HIV or the fraction of people living with HIV who die from HIV related diseases.

The concat method along an axis was used to create data frames with different shapes out of the main data frame in order to create useful graphs and explore the data. Another that was found very useful throughout the project was groupby. It was used to categorize different features based on region, income group, size and the mean values were easily generated by applying the function on the sliced data frame. In data visualization those aggregations are shown and explained.

Data visualization

After the data was merged and aggregated in such a way conclusions could be drawn, visualization was done with seaborn, matplotlib and pandas.

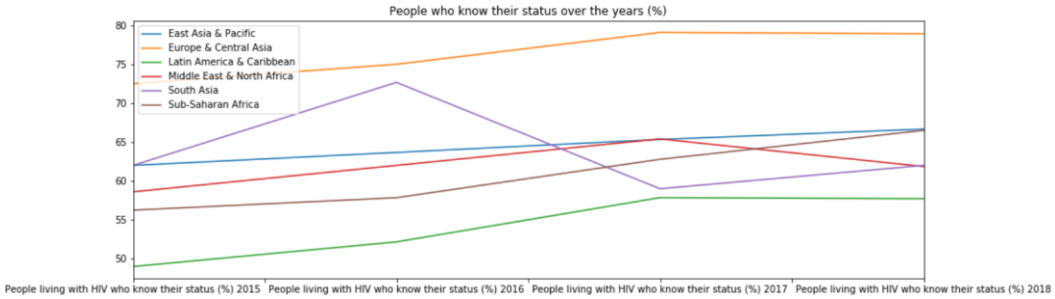
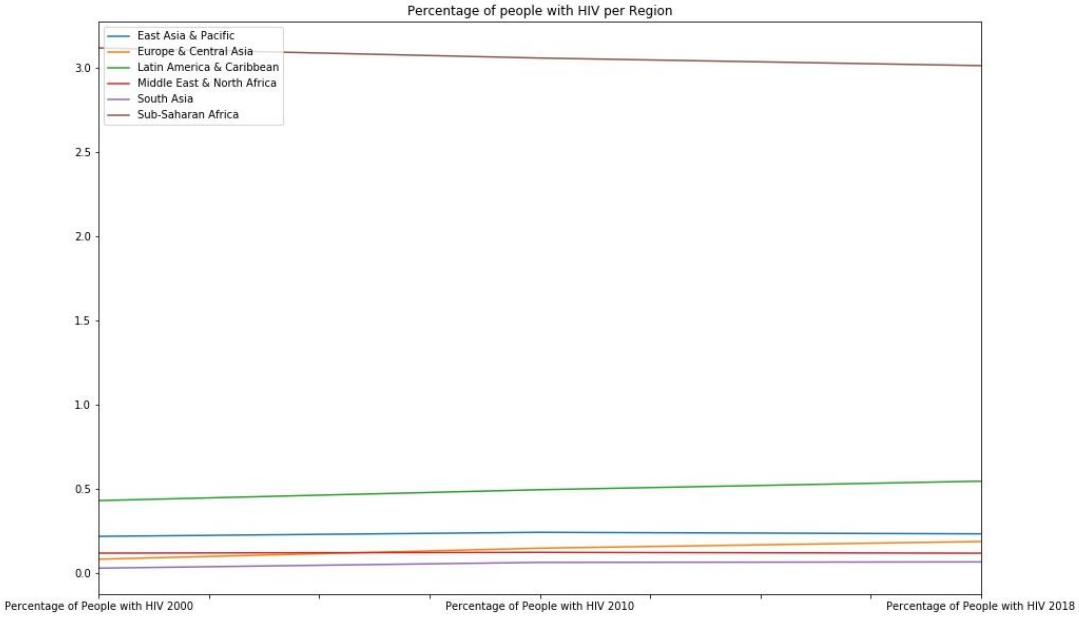


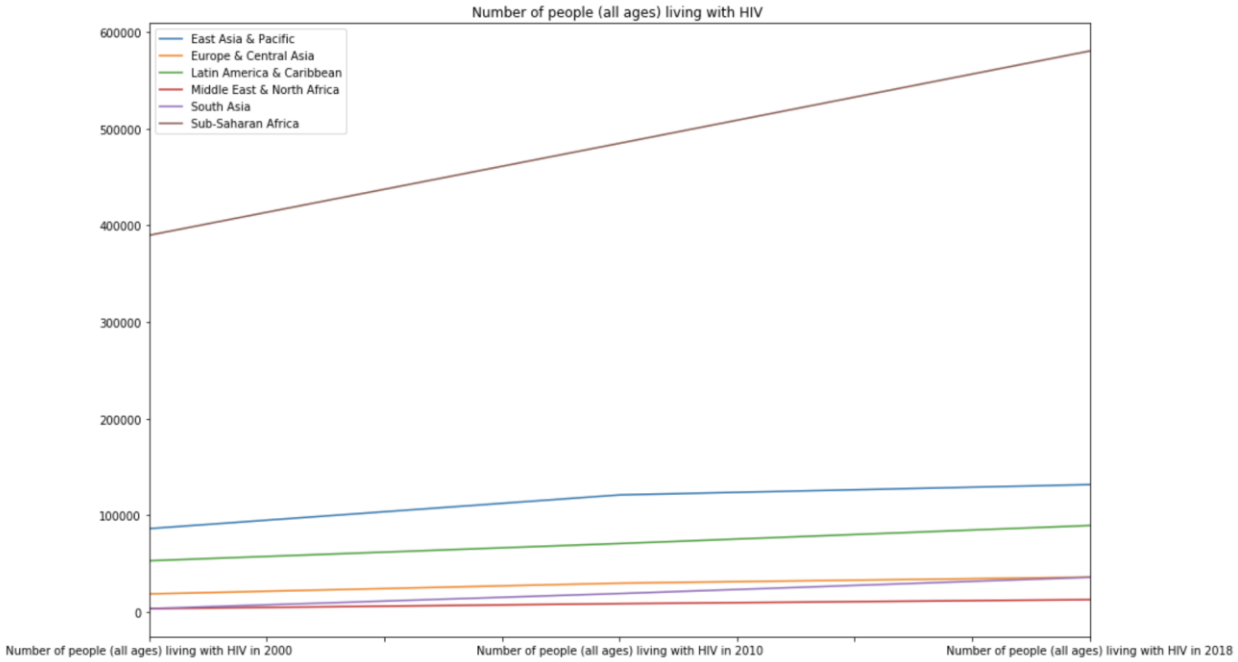
Figure 3. Percentage of people who know their status from 2015 -2018

It can be seen that over the past 5 years, the percentage of people knowing their status seems to rise.



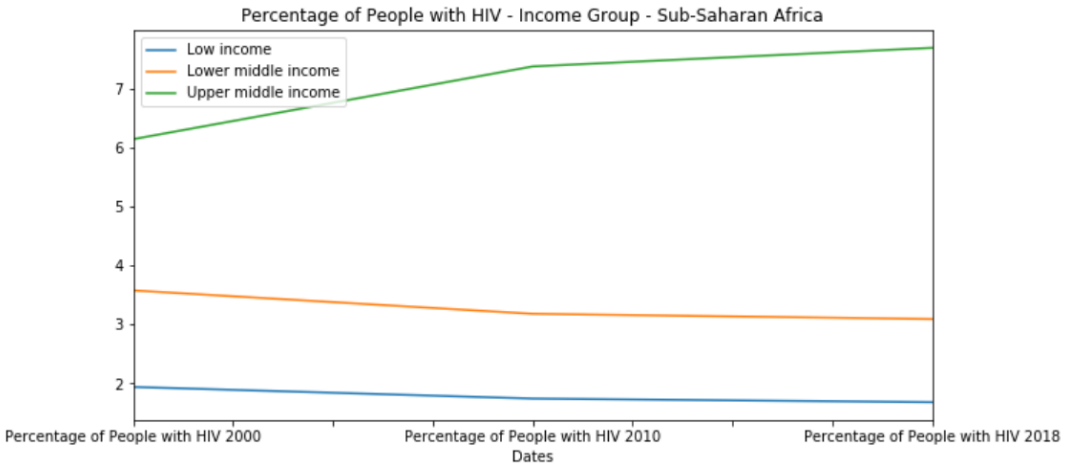
(Figure 4. Percentage of people with HIV per region from 2000 -2018)

UNAIDS and other institutions have set a target for increasing the amount of people who know their status to 90% by 2020 (Bain, 2017). Most regions seem to be gradually increasing, except for the Middle East and North Africa which were declining for a year. While Europe & Central Asia is closest, the target seems to be unrealistic for 2020. A line plot with colored legend was chosen to make the trends easy to see. For this visualization, the percentage per country was aggregated to region.



(Figure 5. Number of people living with HIV from 2000 -2018)

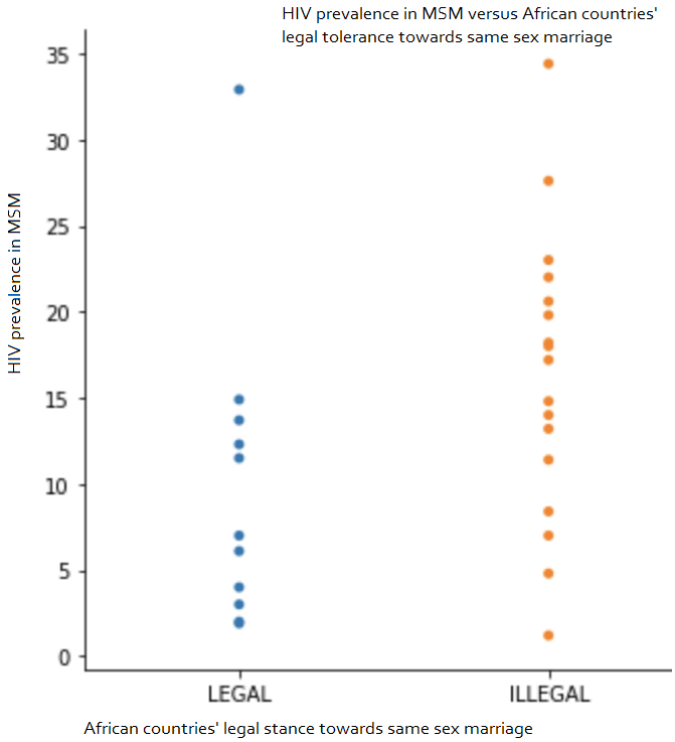
The dataset used has values of certain features for the years 2000, 2010 and 2018. Thus, we wanted to use it to compare values from different years. By using the features "Number of people (all ages) living with HIV" and "Population", we created the feature "Percentage of People with HIV per Region". As shown in the graph above, the Sub Saharan region seems to increase with almost 20% in people suffering from HIV in 2018 compared to 2010. In figure number 4 it can be seen that although there is this increase in numbers, perceptually people living with HIV in the Sub-Saharan region decreases.



(Figure 6. Percentage of people with HIV group by Income in Sub-Saharan Africa 2000-2018)

Not only were the values for the countries in Sub-Saharan Africa almost complete, here HIV was actually prevalent in such a way theory could be tested against it. After grouping and visualizing the percentage of HIV per Sub Saharan country by income group, it was observed that the values for upper middle income increased over the years. At the same time there are less countries in that income group than others so the mean value for upper middle income might have changed more easily. Nonetheless, correlating wealth

directly with HIV prevalence does not accurately reflect other underlying social drivers and structural factors in a country or a region.



'Homophobic violence and conservative attitudes can prevent sexual minorities to get the healthcare they need.' (HIV and AIDS in South Africa, 2019)

HIV is most prevalent in men having homosexual contact. According to literature, this is because they are a minority discriminated heavily against in the form of violence and healthcare denial. (HIV and AIDS in South Africa)

Therefore, a hypothesis was that there is a correlation between LGBT rights in the country and HIV prevalence among Men having Sex with Men (MSM).

(*Figure 7. HIV prevalence in MSM versus African countries' legal tolerance same sex marriage*)

Research in this phenomenon showed that the countries with the lower percentages of HIV among MSM seemed to be pioneers on LGBT rights and the people in countries with high numbers are heavily discriminated against. A dataset containing data on whether or not same sex marriage was legal within a country was added as a way to visualize the most progressive countries concerning LGBT was plotted against percentages. This visualization seemed to support the hypothesis.

Conclusion

Out of the visualizations made from the datasets it is clear that socioeconomic factors have substantial influence on a group of people their odds of contracting and surviving HIV.

The biggest factors seem to be location; region and country. The odds of contracting HIV in the Sub Saharan region versus those in a western European country is a lot higher, no matter what income is had in either region. Out of those two, the country has the highest influence. Even within the sub Saharan region there are still major differences between countries.

As mentioned above, because the odds of contracting HIV are rather low around the world, to see the influence of Socioeconomic factors it is most useful to look at the sub Saharan region for reliable results. HIV is most prevalent there.

To prove minorities that are discriminated against have higher odds of getting HIV, gay rights and HIV prevalence were looked at. Those showed an irrefutable correlation between tolerance and lower prevalence of HIV.

Male HIV prevalence is more strongly correlated with income inequality compared to female HIV prevalence and this is confirmed by Buot et al. (2014), along with our results.

Notes for future research

Another minority researched for this project were women who were facing violence, but no direct correlation was found. It has to be said that this was not a focus of the research, and maybe with more in depth literature research and looking further for more relevant datasets could have come to a well worked out conclusion.

It was a goal to correlate level of education versus HIV prevalence globally, but then it was realized that level of education does not mean the people were educated about sexually transferable diseases because of culture or religion. Hence, this socioeconomic factor was barely researched.

References

Bain, L. E., Nkoke, C., & Noubiap, J. J. N. (2017). UNAIDS 90–90–90 targets to end the AIDS epidemic by 2020 are not realistic: comment on "Can the UNAIDS 90–90–90 target be achieved? A systematic analysis of national HIV treatment cascades". BMJ global health, 2(2), e000227.

HIV & AIDS. The Global Fund. Retrieved January 30, 2020 https://www.theglobalfund.org/en/hivaids/

HIV and AIDS in South Africa. (2019, October 10). Retrieved January 29, 2020, from https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/south-africa

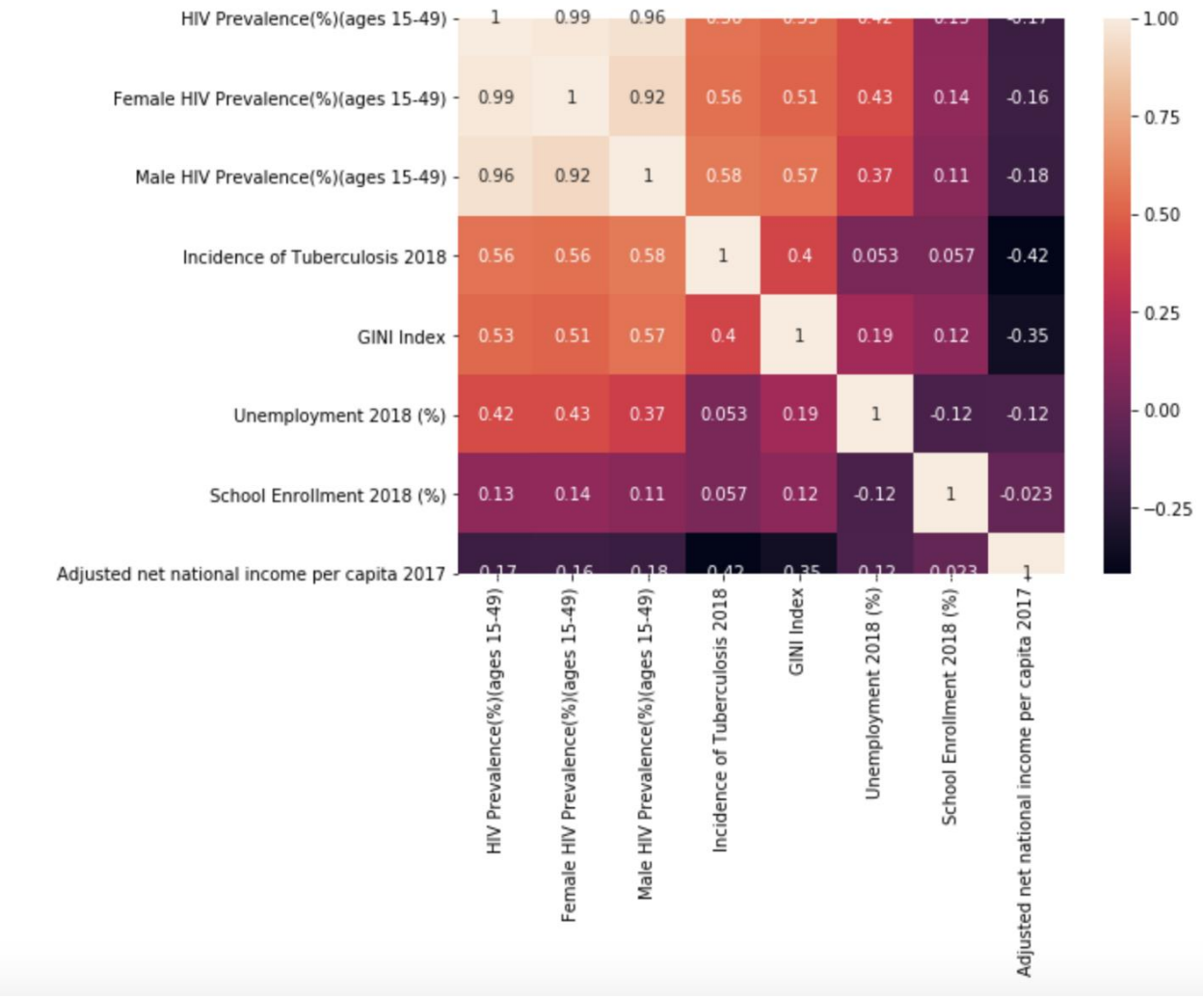
HIV/AIDS and Socioeconomic Status. American Psychological Association. Retrieved January 30, 2020 https://www.apa.org/pi/ses/resources/publications/hiv-aids

Parkhurst, O. J. (2010). Understanding the correlations between wealth, poverty and human immunodeficiency virus infection in African countries. Bulletin of the World Health Organization 2010;88:519-526. doi: 10.2471/BLT.09.070185. https://www.who.int/bulletin/volumes/88/7/09-070185/en/

Pellowski, J. A., Kalichman, S.C., Matthews, K. A., & Adler, N. (2013). A pandemic of the poor: Social disadvantage and the U.S. HIV epidemic. American Psychologist, 68, 197-209.

Wolitski, R., Fecik, N. (2017). Knowledge of HIV Status is on the Rise. U.S. Department of Health and Human Services. Retrieved January 28, 2020 https://www.hiv.gov/blog/knowledge-hiv-status-rise

Appendix:



(Figure 8 The heat map - shows how strongly correlated with different features. If the absolute value is higher than 0.5, which indicates there is a correlation.)

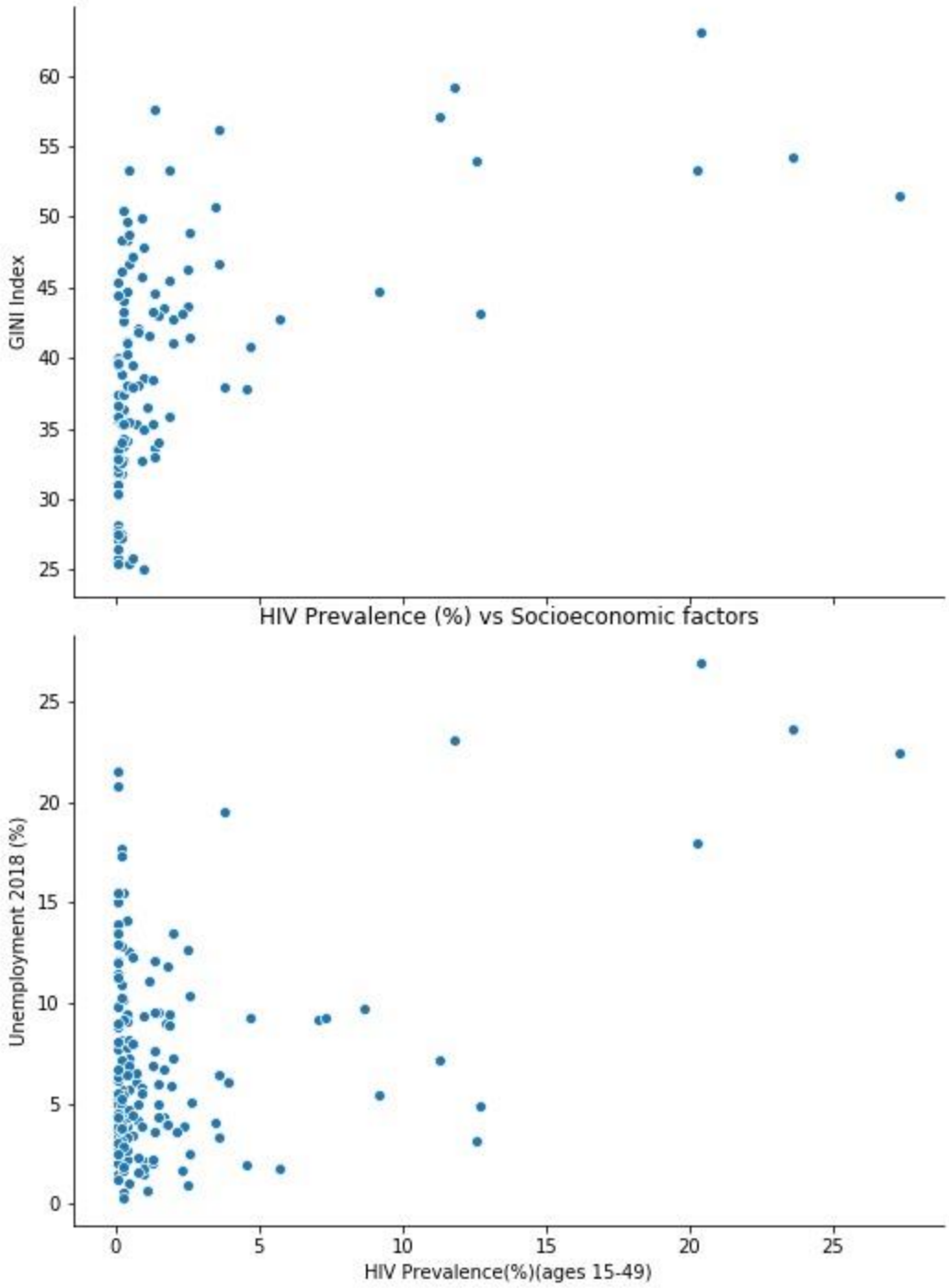


Figure 9 The correction between HIV prevalence % and GINI index. (0.53)

Figure 10 The correction between HIV prevalence% and Unemployment. (0.42)

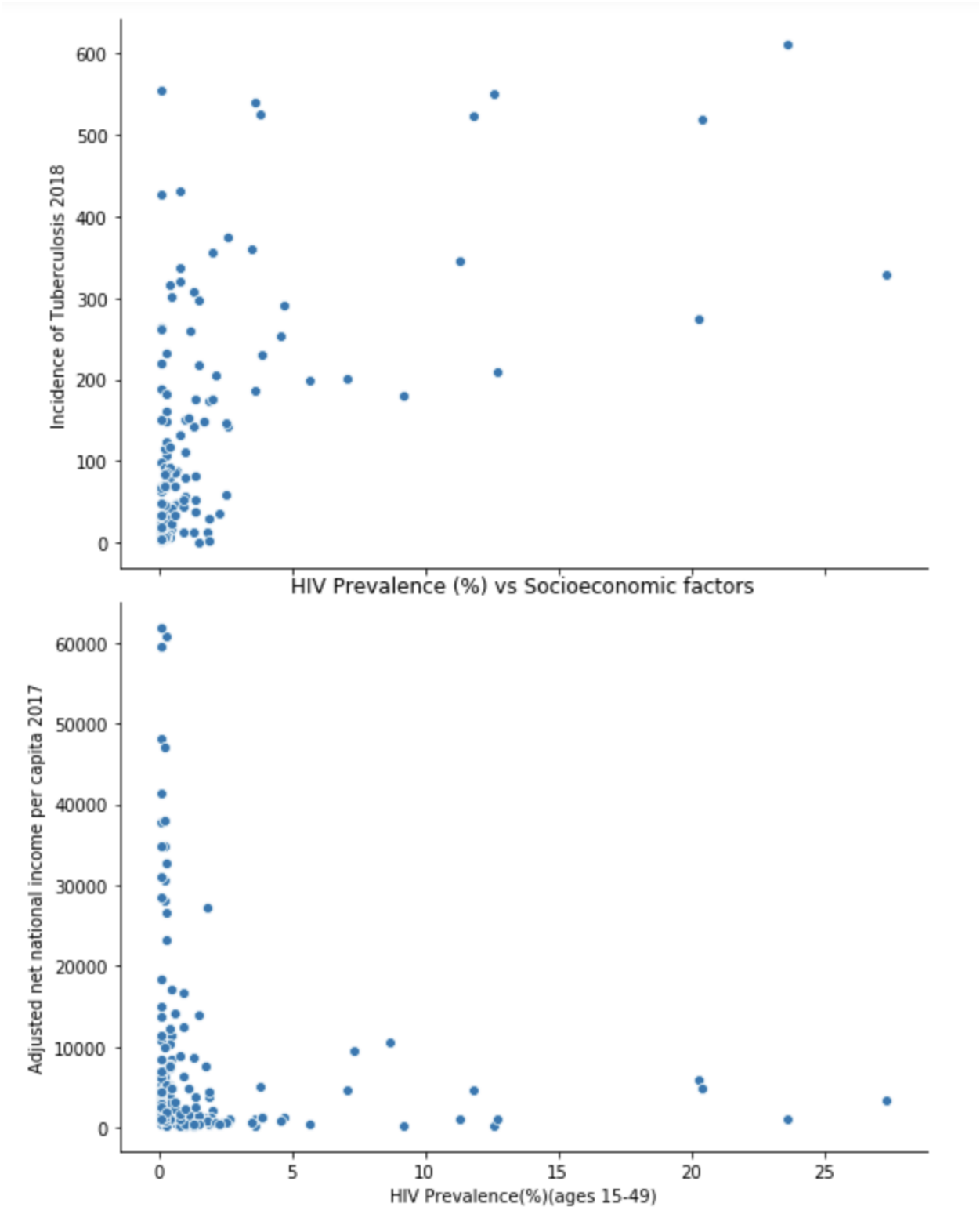


Figure 11 The correction between HIV prevalence % and Incidence of tuberculosis. (which is more correlated with HIV prevalence compared with other factors) (0.56)

Figure 12 The correction between HIV prevalence% and adjusted net national income per capita. (0.17)