Deliverable 2

Long Zhang G01233832

1. Background

1.1. Dataset Description

The dataset that I select is ***Health Nutrition and Population Statistics database*** [1] provided by the World Bank. This dataset contains key health, nutrition and population statistics gathered from a variety of international and national sources [1]. Themes include global surgery, health financing, HIV/AIDS, immunization, infectious diseases, medical resources and usage, noncommunicable diseases, nutrition, population dynamics, reproductive health, universal health coverage, and water and sanitation [1].

1.2. Schema

The schema of the original database is shown in **Fig. 1.** The size of the whole dataset is 76Mb. The metadata is shown in **Appendix 1**.

All information is recorded in the dataset "HNP_StatsData." To be specific, it records the values of 403 indicators of each year of 258 countries and regions [1], totally 10,437,764 samples. Moreover, the temporal coverage of each sample is from 1960 to 2018 [1].

Besides this, "HNP_StatsSeries" and "HNP_StatsCountry" provide more descriptions about indicators and countries, such as the topic and definition of each indicator.

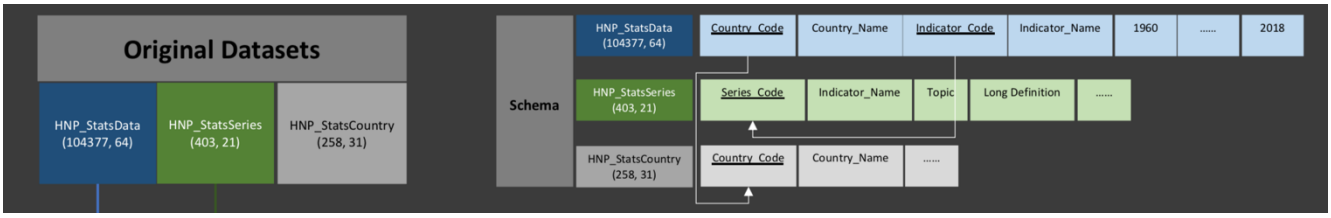


Fig. 1

1.3. Who Collected the Data?

This dataset is collected by World Bank Group. It is a bank in a unique partnership to reduce poverty and support development [2]. The purpose of it is to provide developing countries with financial and technical assistance [2].

1.4. Why did They Collect the Data?

With the purpose of promoting the development of developing countries, World Bank Group collects comprehensive and numerous data about development in countries around the globe [2]. People can use these data to address the world's development challenges [2], including the challenges in health field.

1.5. Privacy, Quality, Ethical, or Other Issues with This Dataset

There is no privacy and ethical issues in this dataset because the data in this dataset is about the description of a country. However, there are some missing values in this dataset, which influences the quality of the dataset.

1.6. What Potential Value Can Be Obtained by Studying This Data?

The indicators in this database involve the topics of health and society. Particularly, this project will focus on the topic of HIV/AIDS. Based on the exploration of this database, I will try to extract valuable information to describe relationships between HIV/AIDS and society characteristics, such as labor force, urban population, gross national income per capita, etc.

1.7. Resources

To study this data, I need to use Excel to have a initiatory understanding of the data. And then, I will use Python to clean, explore, visualize and analyze the data.

1.8. Prior Studies

The website of UNAIDS shows the global HIV& AIDS statistics in 2018 [3]. It has an overview of the HIV infection in 2018 [3].

2. Data Cleaning

This project will focus on the topic of HIV/AIDS and try to extract valuable information to describe relationships between HIV/AIDS and society characteristics. To this end, we need to extract all samples related to HIV/AIDS and society characteristics for each country. The whole process is shown in **Fig. 2**.

2.1. Select Indicators

We can extract indicators that relate to HIV/AIDS and society characteristics in dataset "HNP_StatsSeries."

To be specific, HIV/AIDS indicators can be retrieved via matching keywords "HIV" or "AIDS" in the column of "Indicator_Name." In addition, society characteristics indicators can be retrieved via excluding the keyword "health" in the column of "Topic."

Hence, we get two subsets from dataset "HNP_StatsSeries" which are "HIV_Indicators" and "Society_Indicators." The process and the schema are shown in **Fig. 2**. The number of indicators that relate to HIV is 19. And the number of indicators that relate to society is 52.

2.2. Extract Samples

We can extract all samples that relate to indicators obtained in last step via matching "Series_Code" in dataset "HNP_StatsData." In order to analyze the data easily, I delete all columns that are years before 2018.

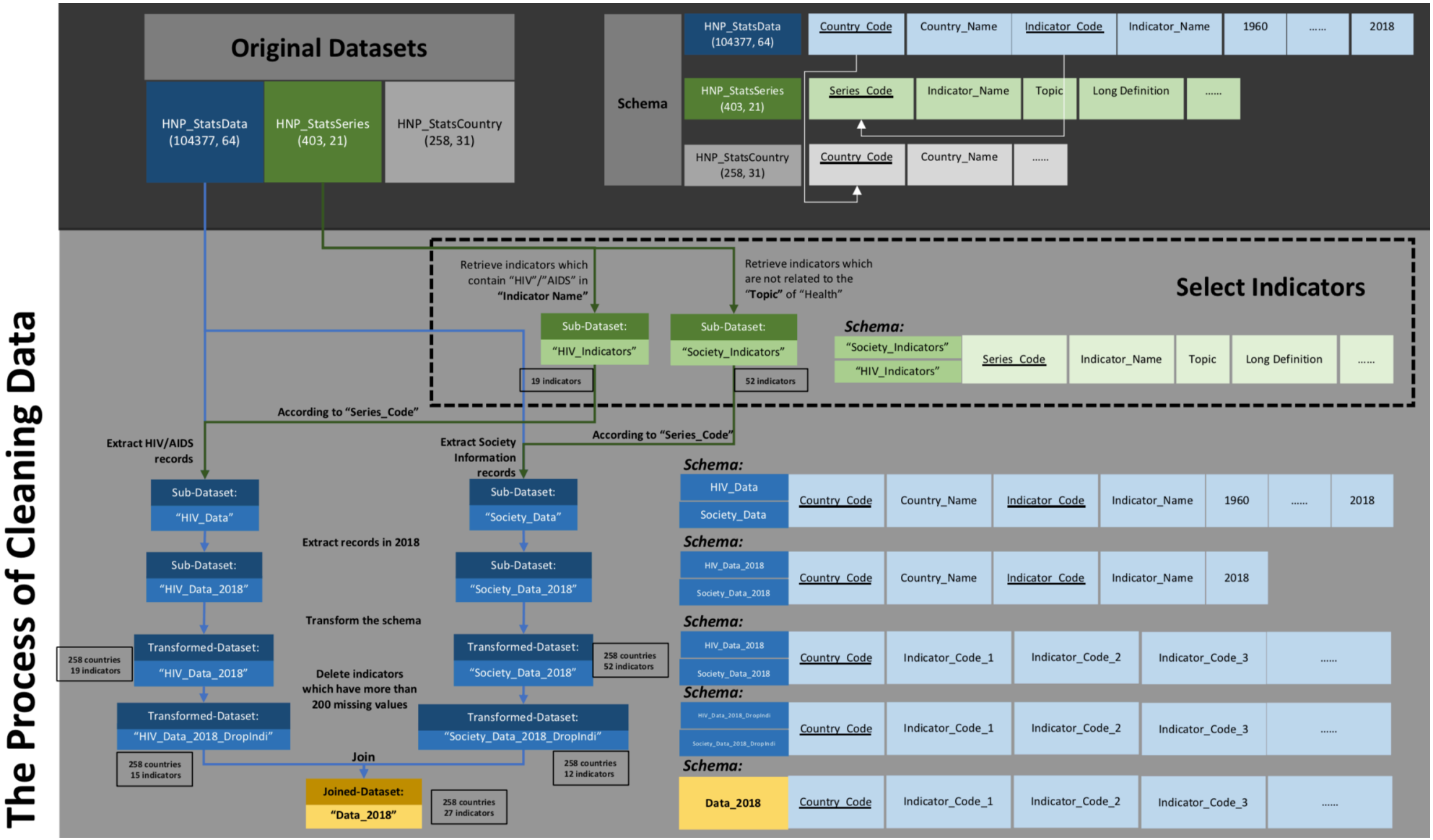

The Process of Cleaning Data
Original Datasets
HNP_StatsData (104377, 64)
HNP_StatsSeries (403, 21)
HNP_StatsCountry (258, 31)
Schema
HNP_StatsData (104377, 64)
Country_Code
Country_Name
Indicator_Code
Indicator_Name
1960
.....
2018
HNP_StatsSeries (403, 21)
Series_Code
Indicator_Name
Topic
Long Definition
.....
HNP_StatsCountry (258, 31)
Country_Code
Country_Name
.....
Select Indicators
Retrieve indicators which contain "HIV"/"AIDS" in "Indicator Name"
Retrieve indicators which are not related to the "Topic" of "Health"
Sub-Dataset: "HIV_Indicators"
19 indicators
Sub-Dataset: "Society_Indicators"
52 indicators
Schema:
"Society_Indicators"
"HIV_Indicators"
Series_Code
Indicator_Name
Topic
Long Definition
.....
According to "Series_Code"
According to "Series_Code"
Extract HIV/AIDS records
Extract Society Information records
Sub-Dataset: "HIV_Data"
Sub-Dataset: "Society_Data"
Schema:
HIV_Data
Society_Data
Country_Code
Country_Name
Indicator_Code
Indicator_Name
1960
.....
2018
Extract records in 2018
Sub-Dataset: "HIV_Data_2018"
Sub-Dataset: "Society_Data_2018"
Schema:
HIV_Data_2018
Society_Data_2018
Country_Code
Country_Name
Indicator_Code
Indicator_Name
2018
Transform the schema
258 countries 19 indicators
Transformed-Dataset: "HIV_Data_2018"
Transformed-Dataset: "Society_Data_2018"
258 countries 52 indicators
Schema:
HIV_Data_2018
Society_Data_2018
Country_Code
Indicator_Code_1
Indicator_Code_2
Indicator_Code_3
.....
Delete indicators which have more than 200 missing values
Transformed-Dataset: "HIV_Data_2018_DropIndi"
Transformed-Dataset: "Society_Data_2018_DropIndi"
Schema:
HIV_Data_2018_DropIndi
Society_Data_2018_DropIndi
Country_Code
Indicator_Code_1
Indicator_Code_2
Indicator_Code_3
.....
258 countries 15 indicators
Join
258 countries 12 indicators
Joined-Dataset: "Data_2018"
258 countries 27 indicators
Schema:
Data_2018
Country_Code
Indicator_Code_1
Indicator_Code_2
Indicator_Code_3
.....

Fig. 2

Hence, we obtain all records that relate HIV and society characteristics in 2018 from dataset "HNP_StatsData." We call the two subsets as "HIV_Data_2018" and "Society_Data_2018." The process and the schema are shown in **Fig. 2**.

2.3. Transform the Schema

We have obtained necessary samples last step.

As shown in **Fig. 2,** the primary keys of datasets "HIV_Data_2018" and "Society_Data_2018" are the combination of "Country_Code" and "Indicator_Code."

However, what we need is to obtain all relative indicators for each country, which means we have to transform the primary key of the datasets to be only "Country_Code" and make all Indicator_Codes as attribute in the schema.

The schema of transformed datasets is shown in **Fig. 2**.

2.4. Delete Indicators that Have too many Missing Values

First of all, "HIV_Data_2018" contains 258 countries and 19 indicators; "Society_Data_2018" contains 258 countries and 52 indicators**. Fig. 3** shows the number of missing values in each column in "HIV_Data_2018"; **Fig. 4** shows the number of missing values in each column in "Society_Data_2018." We can see several indicators have at least 200 missing values. These indicators have too many missing values to be extracted valuable information.

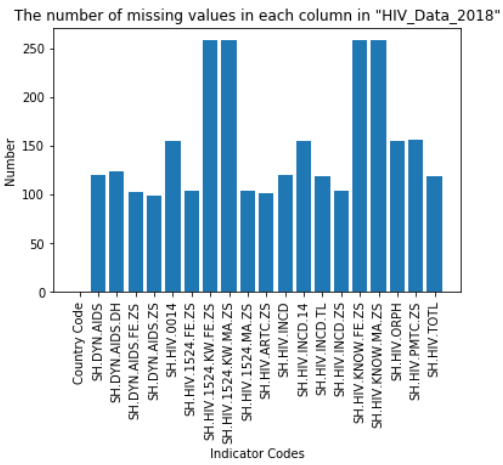


Fig. 3

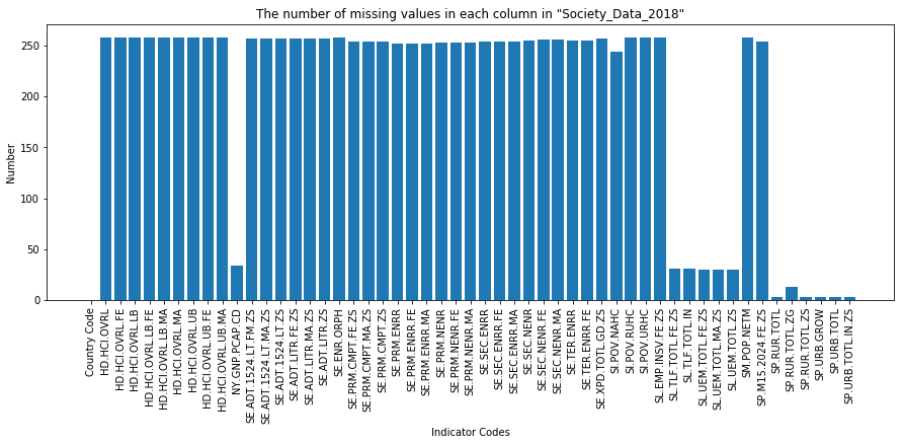


Fig. 4

Hence, we delete all indicators that have more than 200 missing values. After that, the number of indicators that relate to HIV is 15; the number of indicators that relate to society is 12. The specific information of selected indicators is shown in **Table 1** and **Table 2**.

Table 1: Indicators that relate to HIV

	Series_Code	Topic	Type	Indicator_Name	Long Definition
1	SH.DYN.AIDS	Health: Risk factors	Ratio	Adults (ages 15+) living with HIV	Adults living with HIV refers to the number of people ages 15-49 who are infected with HIV.
2	SH.DYN.AIDS.DH	Health: Risk factors	Ratio	AIDS estimated deaths (UNAIDS estimates)	AIDS deaths are the estimated number of adults and children who died due to AIDS-related causes.
3	SH.DYN.AIDS.FE.ZS	Health: Risk factors	Ratio	Women's share of population ages 15+ living with HIV (%)	Prevalence of HIV is the percentage of people who are infected with HIV. Female rate is as a percentage of the total population ages 15+ who are living with HIV.
4	SH.DYN.AIDS.ZS	Health: Risk factors	Ratio	Prevalence of HIV, total (% of population ages 15-49)	Prevalence of HIV refers to the percentage of people ages 15-49 who are infected with HIV.
5	SH.HIV.0014	Health: Risk factors	Ratio	Children (0-14) living with HIV	Children living with HIV refers to the number of children ages 0-14 who are infected with HIV.
6	SH.HIV.1524.FE.ZS	Health: Risk factors	Ratio	Prevalence of HIV, female (% ages 15-24)	Prevalence of HIV, female is the percentage of females who are infected with HIV. Youth rates are as a percentage of the relevant age group.
7	SH.HIV.1524.MA.ZS	Health: Risk factors	Ratio	Prevalence of HIV, male (% ages 15-24)	Prevalence of HIV, male is the percentage of males who are infected with HIV. Youth rates are as a percentage of the relevant age group.
8	SH.HIV.ARTC.ZS	Health: Risk factors	Ratio	Antiretroviral therapy coverage (% of people living with HIV)	Antiretroviral therapy coverage indicates the percentage of all people living with HIV who are receiving antiretroviral therapy.
9	SH.HIV.INCD	Health: Risk factors	Ratio	Adults (ages 15+) newly infected with HIV	Number of adults (ages 15+) newly infected with HIV.
10	SH.HIV.INCD.14	Health: Risk factors	Ratio	Children (ages 0-14) newly infected with HIV	Number of children (ages 0-14) newly infected with HIV.
11	SH.HIV.INCD.TL	Health: Risk factors	Ratio	Adults (ages 15+) and children (ages 0-14) newly infected with HIV	Number of adults (ages 15+) and children (ages 0-14) newly infected with HIV.
12	SH.HIV.INCD.ZS	Health: Risk factors	Ratio	Incidence of HIV (per 1,000 uninfected population ages 15-49)	Number of new HIV infections among uninfected populations ages 15-49 expressed per 1,000 uninfected population in the year before the period.
13	SH.HIV.ORPH	Health: Risk factors	Ratio	Children orphaned by HIV/AIDS	Number of children orphaned by HIV/AIDS is the estimated number of children who have lost their mother or both parents to AIDS before age 15 since the epidemic began. Some of the orphaned children included in this cumulative total are no longer alive; others are no longer under age 15.

	Series_Code	Topic	Type	Indicator_Name	Long Definition
14	SH.HIV.PMTC.ZS	Health: Risk factors	Ratio	Antiretroviral therapy coverage for PMTCT (% of pregnant women living with HIV)	Percentage of pregnant women with HIV who receive antiretroviral medicine for prevention of mother-to-child transmission (PMTCT).
15	SH.HIV.TOTL	Health: Risk factors	Ratio	Adults (ages 15+) and children (0-14 years) living with HIV	Adults and children living with HIV refers to the number of people ages 0-49 (adult ages 15-49 and children ages 0-14) who are infected with HIV.

Source: "HNP_StatsSeries"

Table 2: Indicators that relate to society characteristics

	Series_Code	Topic	Type	Indicator_Name	Long Definition
1	NY.GNP.PCAP.CD	Economic Policy & Debt: National accounts: Atlas GNI & GNI per capita	Ratio	GNI per capita, Atlas method (current US$)	GNI per capita (formerly GNP per capita) is the gross national income, converted to U.S. dollars using the World Bank Atlas method, divided by the midyear population. GNI is the sum of value added by all resident producers plus any product taxes (less subsidies) not included in the valuation of output plus net receipts of primary income (compensation of employees and property income) from abroad. GNI, calculated in national currency, is usually converted to U.S. dollars at official exchange rates for comparisons across economies, although an alternative rate is used when the official exchange rate is judged to diverge by an exceptionally large margin from the rate actually applied in international transactions. To smooth fluctuations in prices and exchange rates, a special Atlas method of conversion is used by the World Bank. This applies a conversion factor that averages the exchange rate for a given year and the two preceding years, adjusted for differences in rates of inflation between the country, and through 2000, the G-5 countries (France, Germany, Japan, the United Kingdom, and the United States). From 2001, these countries include the Euro area, Japan, the United Kingdom, and the United States.
2	SL.TLF.TOTL.FE.ZS	Social Protection & Labor: Labor force structure	Ratio	Labor force, female (% of total labor force)	Female labor force as a percentage of the total show the extent to which women are active in the labor force. Labor force comprises people ages 15 and older who supply labor for the production of goods and services during a specified period.
3	SL.TLF.TOTL.IN	Social Protection & Labor: Labor force structure	Ratio	Labor force, total	Labor force comprises people ages 15 and older who supply labor for the production of goods and services during a specified period. It includes people who are currently employed and people who are unemployed but seeking work as well as first-time job-seekers. Not everyone who works is included, however. Unpaid workers, family workers, and students are often omitted, and some countries do not count members of the armed forces. Labor force size tends to vary during the year as seasonal workers enter and leave.

	Series_Code	Topic	Type	Indicator_Name	Long Definition
4	SL.UEM.TOTL.FE.ZS	Social Protection & Labor: Unemployment	Ratio	Unemployment, female (% of female labor force) (modeled ILO estimate)	Unemployment refers to the share of the labor force that is without work but available for and seeking employment.
5	SL.UEM.TOTL.MA.ZS	Social Protection & Labor: Unemployment	Ratio	Unemployment, male (% of male labor force) (modeled ILO estimate)	Unemployment refers to the share of the labor force that is without work but available for and seeking employment.
6	SL.UEM.TOTL.ZS	Social Protection & Labor: Unemployment	Ratio	Unemployment, total (% of total labor force) (modeled ILO estimate)	Unemployment refers to the share of the labor force that is without work but available for and seeking employment.
7	SP.RUR.TOTL	Environment: Density & urbanization	Ratio	Rural population	Rural population refers to people living in rural areas as defined by national statistical offices. It is calculated as the difference between total population and urban population. Aggregation of urban and rural population may not add up to total population because of different country coverages.
8	SP.RUR.TOTL.ZG	Environment: Density & urbanization	Ratio	Rural population growth (annual %)	Rural population refers to people living in rural areas as defined by national statistical offices. It is calculated as the difference between total population and urban population.
9	SP.RUR.TOTL.ZS	Environment: Density & urbanization	Ratio	Rural population (% of total population)	Rural population refers to people living in rural areas as defined by national statistical offices. It is calculated as the difference between total population and urban population.
10	SP.URB.GROW	Environment: Density & urbanization	Ratio	Urban population growth (annual %)	Urban population refers to people living in urban areas as defined by national statistical offices. It is calculated using World Bank population estimates and urban ratios from the United Nations World Urbanization Prospects.
11	SP.URB.TOTL	Environment: Density & urbanization	Ratio	Urban population	Urban population refers to people living in urban areas as defined by national statistical offices. It is calculated using World Bank population estimates and urban ratios from the United Nations World Urbanization Prospects. Aggregation of urban and rural population may not add up to total population because of different country coverages.
12	SP.URB.TOTL.IN.ZS	Environment: Density & urbanization	Ratio	Urban population (% of total population)	Urban population refers to people living in urban areas as defined by national statistical offices. The data are collected and smoothed by United Nations Population Division.

Source: "HNP_StatsSeries"

2.5. Join the Datasets

To analyze the data easily, we join the two datasets ("HIV_Data_2018" and "Society_Data_2018") on "Country_Code." It has 258 countries and 27 indicators. The specific schema is shown in **Fig. 2**. In addition, several countries do not have records for some indicators. Hence, we only keep 85 countries that have all records. We call the final dataset as "Data_2018." The overview of the dataset "Data_2018" is shown in **Table 3**.

Table 3: The summary of dataset "Data_2018"

Society	NY.GNP.PCAP.CD	SL.TLF.TOTL.FE.ZS	SL.TLF.TOTL.IN	SL.UEM.TOTL.FE.ZS	SL.UEM.TOTL.MA.ZS	SL.UEM.TOTL.ZS
count	85.00	85.00	85.00	85.00	85.00	85.00
mean	3,588.09	42.22	53,920,642.45	8.31	6.03	6.81
std	3,653.82	8.39	373,951,461.48	7.37	4.98	5.64
min	280.00	7.77	179,356.00	0.16	0.36	0.27
25%	960.00	39.37	2,507,512.00	3.12	2.45	2.84
50%	2,030.00	44.41	7,043,698.00	5.63	4.65	4.95
75%	4,990.00	48.35	15,392,266.00	10.79	8.24	9.37
max	15,650.00	55.76	3,456,043,460.00	29.27	25.05	26.96

Society	SP.RUR.TOTL	SP.RUR.TOTL.ZG	SP.RUR.TOTL.ZS	SP.URB.GROW	SP.URB.TOTL	SP.URB.TOTL.IN.ZS
count	85.00	85.00	85.00	85.00	85.00	85.00
mean	57,097,251.09	0.86	50.58	2.80	63,838,730.16	49.42
std	367,447,538.72	1.24	19.65	1.39	454,132,715.54	19.65
min	160,944.00	-3.05	4.67	-0.47	175,155.00	13.03
25%	2,034,309.00	0.05	36.18	1.83	2,469,421.00	35.92
50%	8,372,370.00	0.97	50.05	2.97	7,455,039.00	49.95
75%	18,056,198.00	1.71	64.08	3.89	16,635,227.00	63.82
max	3,396,031,144.00	3.73	86.97	6.18	4,196,356,648.00	95.33

HIV	SH.DYN.AIDS	SH.DYN.AIDS.DH	SH.DYN.AIDS.FE.ZS	SH.DYN.AIDS.ZS	SH.HIV.0014	SH.HIV.1524.FE.ZS
count	85.00	85.00	85.00	85.00	85.00	85.00
mean	745,995.29	15,730.94	47.12	2.83	38,722.82	1.35
std	3,992,358.32	83,721.85	15.03	5.32	186,824.70	2.74
min	4,700.00	100.00	6.00	0.10	100.00	0.10
25%	22,000.00	580.00	34.00	0.30	500.00	0.10
50%	68,000.00	1,800.00	47.40	0.90	3,200.00	0.30
75%	230,000.00	6,100.00	61.00	2.00	12,000.00	1.00
max	36,200,000.00	770,000.00	72.30	27.30	1,700,000.00	15.90

HIV	SH.HIV.1524.MA.ZS	SH.HIV.ARTC.ZS	SH.HIV.INCD	SH.HIV.INCD.14	SH.HIV.INCD.TL	SH.HIV.INCD.ZS
count	85.00	85.00	85.00	85.00	85.00	85.00
mean	0.63	52.13	31,616.35	3,680.59	34,723.88	1.49
std	0.99	19.40	174,828.80	17,566.35	186,027.58	2.79
min	0.10	9.00	100.00	100.00	200.00	0.02
25%	0.10	36.00	970.00	100.00	1,100.00	0.20
50%	0.20	54.00	3,300.00	500.00	3,600.00	0.40
75%	0.50	65.00	11,000.00	1,300.00	12,000.00	1.30
max	4.90	92.00	1,600,000.00	160,000.00	1,700,000.00	15.40

HIV	SH.HIV.ORPH	SH.HIV.PMTC.ZS	SH.HIV.TOTL
count	85.00	85.00	85.00
mean	317,742.82	68.66	784,216.47
std	1,622,508.95	27.64	4,177,496.77
min	940.00	5.00	4,900.00
25%	7,500.00	46.00	22,000.00
50%	37,000.00	80.00	71,000.00
75%	110,000.00	93.00	240,000.00
max	14,900,000.00	95.00	37,900,000.00

3. Explore the Data

In this project, we pursue to find some relationships between HIV/AIDS situation and society characteristics. To this end, we draw the scatter pair-plot **Fig. 5** of dataset “Data_2018” and the heatmap **Fig. 6** of correlation matrix of dataset “Data_2018” to have a preview of the relationships between HIV indicators and society indicators.

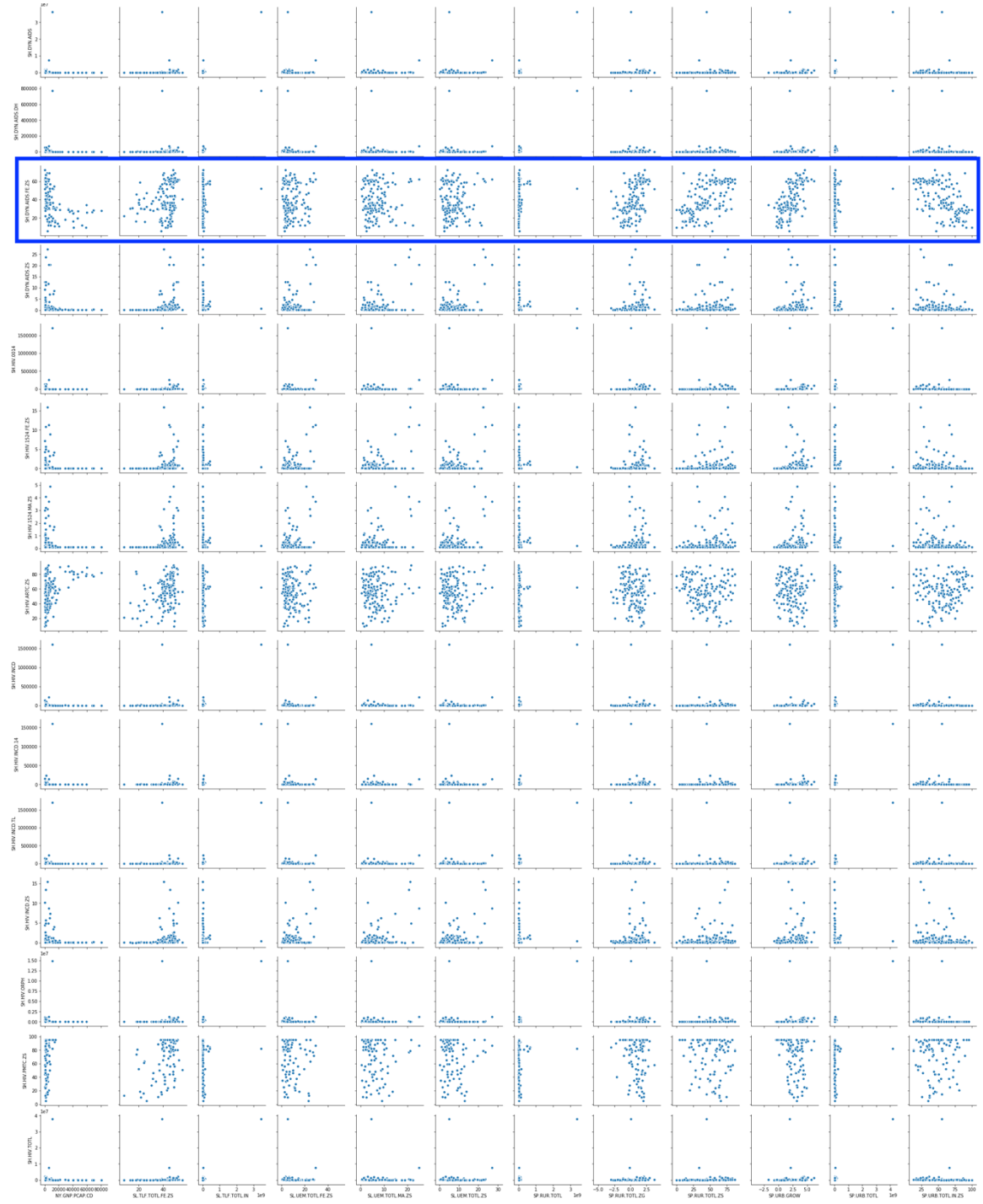


Fig. 5: indicators on x-axis are society indicators; indicators on y-axis are HIV indicators

First of all, the scatter plots in the blue box are the relationships between SH.DYN.AIDS.FE.ZS("women's share of population ages 15+ living with HIV (%)") and 12 society indicators. We can see that several scatter plots in the blue box have potential linear relationships.

From now, indicator "SH.DYN.AIDS.FE.ZS" is the example to analyze in this project. We will investigate what the relationships between indicator "SH.DYN.AIDS.FE.ZS" and society indicators are.

And then, we can follow the process of the analyzation of indicator "SH.DYN.AIDS.FE.ZS" to analyze more other HIV indicators.

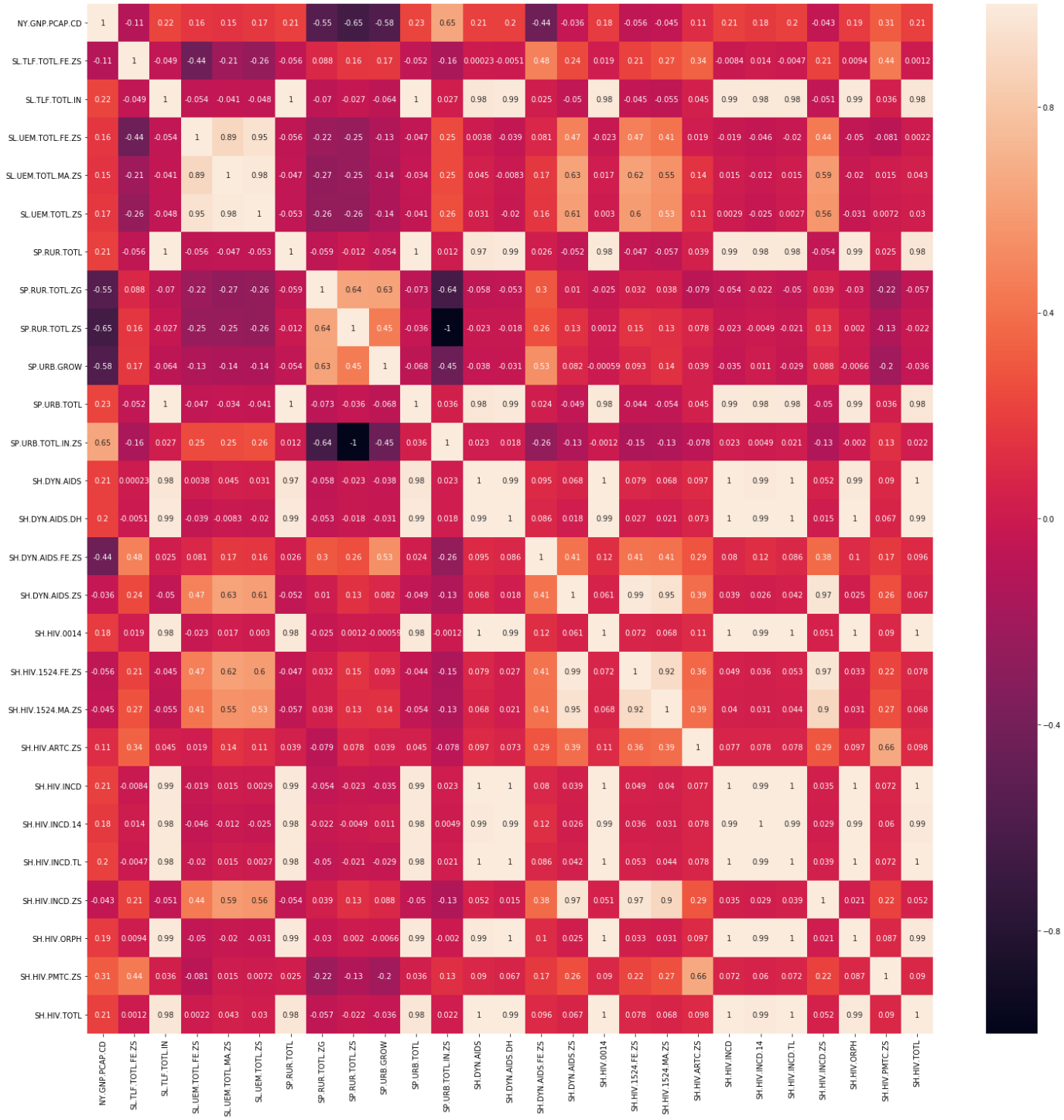


Fig. 6: Heatmap of correlation matrix for dataset "Data_2018"

In Fig. 6 we can see that "SH.DYN.AIDS.FE.ZS" has high correlation with indicator "SP.URB.GROW" which is urban population growth. The correlation is 0.53. However, other society indicators have low correlation with "SH.DYN.AIDS.FE.ZS." Hence, we will build more regression models to analyze it later.

Before do it, we can see how the distribution of indicator "SH.DYN.AIDS.FE.ZS" changes with time going by in boxplot **Fig. 7**. Moreover, **Fig .8** shows the tendency and the linear regression of mean of "SH.DYN.AIDS.FE.ZS" of all countries. It is easy to find that women's share of population age 15+ living with HIV is slowly growing.

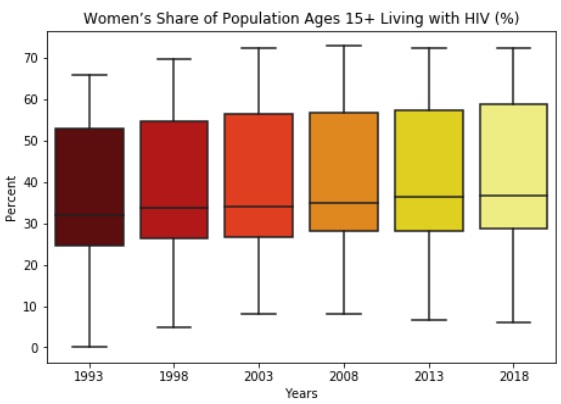


Fig. 7

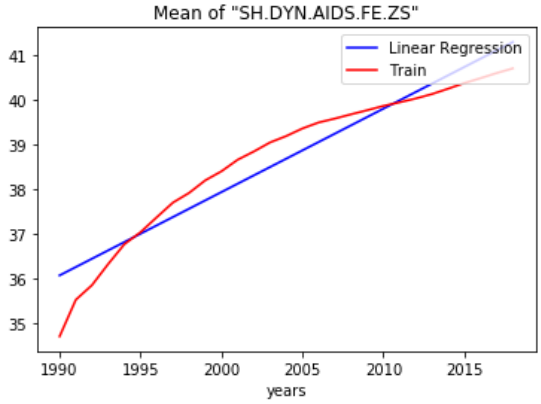


Fig. 8

4. Multivariate Regression Model for "SH.DYN.AIDS.FE.ZS"

4.1. Method

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + + \beta_\rho X_\rho + \sigma$$

In a multiple linear regression model, $X_1, X_2, \dots, X_\rho$ are ρ independent variables, Y is the dependent variable. β_0 is the intercept. $\beta_i (i = 1,2, \dots, \rho)$ is the slope coefficient for each of the independent variables. σ is the error. In this case, the dependent variable is the "SH.DYN.AIDS.FE.ZS" ("women's share of population ages 15+ living with HIV (%)").

$\beta_i (i = 1,2, \dots, \rho)$ equals the mean increase in Y per unit increase in $X_i (i = 1,2, \dots, \rho)$, which other $X_j (j \neq i)$ are kept fixed.

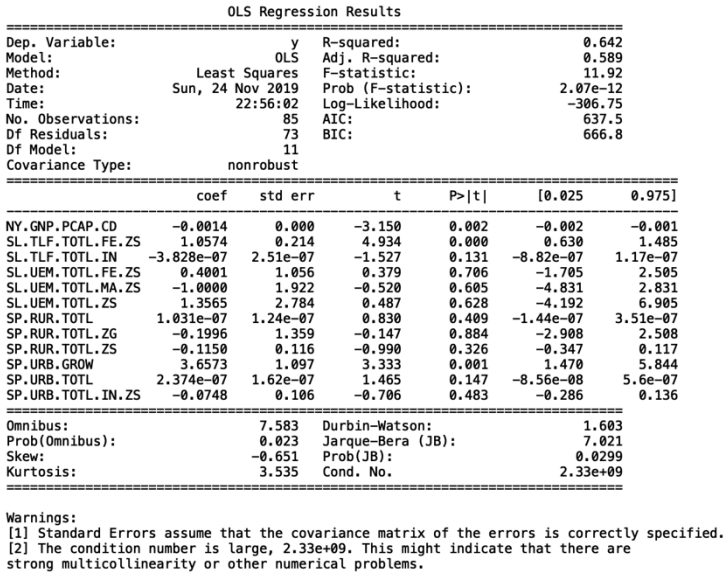
R-squared is a statistical approach to measure how the regression model fit the data. Usually, the higher the R-squared, the better the model fits the data. In addition, Adj. R-squared is the value of adjusted R-squared which is based on the number of observations and the degrees-of-freedom of the residuals.

The OLS regression model in the library "statsmodels.api" of python also provides the hypothesis testing to check how statistically significant between the dependent variable and independent variables.

4.2. Experiment and Evaluation

I use the library "statsmodels.api" in python to make the multivariate regression for "SH.DYN.AIDS.FE.ZS."

First of all, I make the all 12 society indicators to be independent variables. The results of this model are shown in **Fig. 9.** We can see that the R-squared is only 0.642 and the Adj. R-squared is 0.589, which means the regression model does not fit the data well. In addition, the p-values of "NY.GNP.PCAP.CD," "SL.TLF.TOTL.FE.ZS," and "SP.URB.GROW" are 0.002, 0.000, and 0.001 which are smaller than 0.05. This indicates that the three indicators the relationships between "SH.DYN.AIDS.FE.ZS" and these three indicators are statistically significant. Hence, we will adjust the regression model to make it only involve these three indicators.



OLS Regression Results

Dep. Variable:	y	R-squared:	0.642
Model:	OLS	Adj. R-squared:	0.589
Method:	Least Squares	F-statistic:	11.92
Date:	Sun, 24 Nov 2019	Prob (F-statistic):	2.07e-12
Time:	22:56:02	Log-Likelihood:	-306.75
No. Observations:	85	AIC:	637.5
Df Residuals:	73	BIC:	666.8
Df Model:	11		
Covariance Type:	nonrobust		

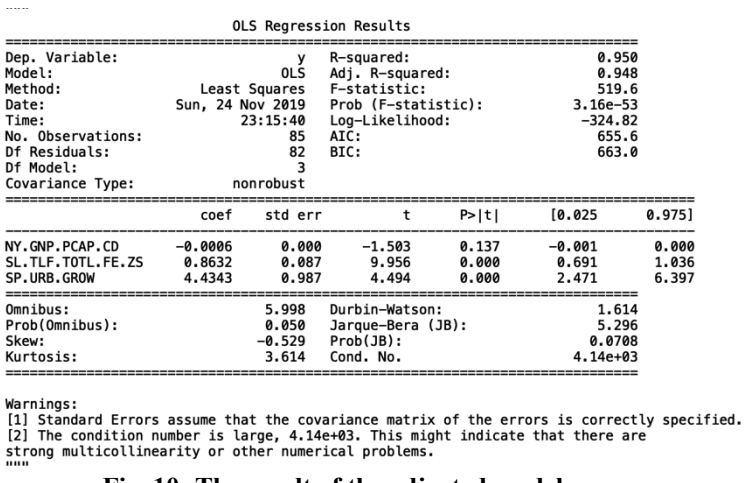
	coef	std err	t	P>|t|	[0.025	0.975]
NY.GNP.PCAP.CD	-0.0014	0.000	-3.150	0.002	-0.002	-0.001
SL.TLF.TOTL.FE.ZS	1.0574	0.214	4.934	0.000	0.630	1.485
SL.TLF.TOTL.IN	-3.828e-07	2.51e-07	-1.527	0.131	-8.82e-07	1.17e-07
SL.UEM.TOTL.FE.ZS	0.4001	1.056	0.379	0.706	-1.705	2.505
SL.UEM.TOTL.MA.ZS	-1.0000	1.922	-0.520	0.605	-4.831	2.831
SL.UEM.TOTL.ZS	1.3565	2.784	0.487	0.628	-4.192	6.905
SP.RUR.TOTL	1.031e-07	1.24e-07	0.830	0.409	-1.44e-07	3.51e-07
SP.RUR.TOTL.ZG	-0.1996	1.359	-0.147	0.884	-2.908	2.508
SP.RUR.TOTL.ZS	-0.1150	0.116	-0.990	0.326	-0.347	0.117
SP.URB.GROW	3.6573	1.097	3.333	0.001	1.470	5.844
SP.URB.TOTL	2.374e-07	1.62e-07	1.465	0.147	-8.56e-08	5.6e-07
SP.URB.TOTL.IN.ZS	-0.0748	0.106	-0.706	0.483	-0.286	0.136

Omnibus:	7.583	Durbin-Watson:	1.603
Prob(Omnibus):	0.023	Jarque-Bera (JB):	7.021
Skew:	-0.651	Prob(JB):	0.0299
Kurtosis:	3.535	Cond. No.	2.33e+09

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.33e+09. This might indicate that there are
strong multicollinearity or other numerical problems.

Fig. 9: The result of the model that have 12 independent variables

The result of the adjusted model is shown in the **Fig. 10**. We can see that the R-squared is 0.95 and the Adj. R-squared is 0.948, which means that the adjusted regression model has good performance on fitting the datasets.



OLS Regression Results

Dep. Variable:	y	R-squared:	0.950
Model:	OLS	Adj. R-squared:	0.948
Method:	Least Squares	F-statistic:	519.6
Date:	Sun, 24 Nov 2019	Prob (F-statistic):	3.16e-53
Time:	23:15:40	Log-Likelihood:	-324.82
No. Observations:	85	AIC:	655.6
Df Residuals:	82	BIC:	663.0
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
NY.GNP.PCAP.CD	-0.0006	0.000	-1.503	0.137	-0.001	0.000
SL.TLF.TOTL.FE.ZS	0.8632	0.087	9.956	0.000	0.691	1.036
SP.URB.GROW	4.4343	0.987	4.494	0.000	2.471	6.397

Omnibus:	5.998	Durbin-Watson:	1.614
Prob(Omnibus):	0.050	Jarque-Bera (JB):	5.296
Skew:	-0.529	Prob(JB):	0.0708
Kurtosis:	3.614	Cond. No.	4.14e+03

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.14e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Fig. 10: The result of the adjusted model.

In addition, we can draw the plot to show the relationships between these indicators.

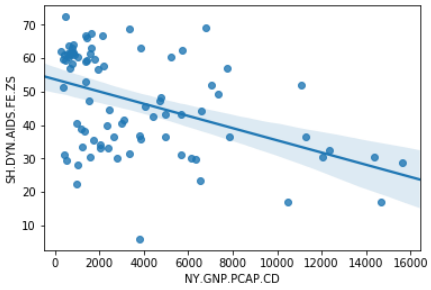


Fig. 11

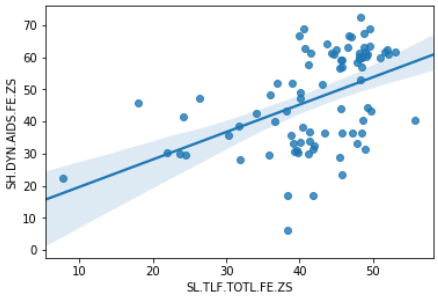


Fig. 12

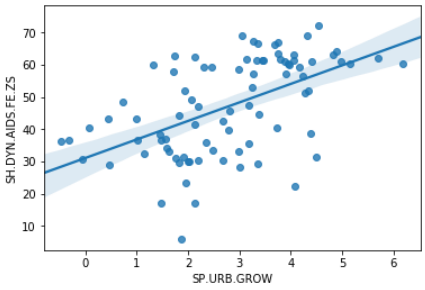


Fig. 13

5. Conclusion

According to the adjusted model, we can predict the "SH.DYN.AIDS.FE.ZS" ("women's share of population ages 15+ living with HIV") by indicators "NY.GNP.PCAP.CD", "SL.TLF.TOTL.FE.ZS", and "SP.URB.GROW" which are "GNI(gross national income) per capita", "Labor force, female (% of total labor force)", and "Urban population growth (annual %)".

The correlations between "SH.DYN.AIDS.FE.ZS" and "NY.GNP.PCAP.CD", "SH.DYN.AIDS.FE.ZS" and "SL.TLF.TOTL.FE.ZS", and "SH.DYN.AIDS.FE.ZS" and "SP.URB.GROW" are statistically significant.

In addition, according to the coefficient of "SH.DYN.AIDS.FE.ZS", the relationship between "SH.DYN.AIDS.FE.ZS" and "NY.GNP.PCAP.CD" is negative correlation. Keep "SL.TLF.TOTL.FE.ZS", and "SP.URB.GROW" fixed, per unit increase in "NY.GNP.PCAP.CD" will make "SH.DYN.AIDS.FE.ZS" reduce 0.0006. Moreover, the relationships between "SH.DYN.AIDS.FE.ZS" and "SL.TLF.TOTL.FE.ZS", and "SH.DYN.AIDS.FE.ZS" and "SP.URB.GROW" are both positive correlation.

6. Limitations and Future Works

This project only focus on the relationship between HIV indicator "women's share of population ages 15+ living with HIV" and society indicators. Based on the cleaned dataset "Data_2018", We can analyze the relationships between other HIV indicators and society indicators to extract more information via the process in the project.

In addition, the society indicators in this project are only retrieved from this health database. We can collect more society indicators from other database to make more comprehensive analysis.

Moreover, the model in this project only extracts the data in 2018. we can consider the time series analysis on the model.

Lastly, as for the models in the future, we need to think about more academic theories of current situations of HIV/AIDS. Based on this, we can collect more relevant data to make more precise analysis.

References

[1] The World Bank, "Health Nutrition And Population Statistics," The World Bank, 2019. [Online]. Available: https://datacatalog.worldbank.org/dataset/health-nutrition-and-population-statistics.

[2] World Bank Group, "What We Do," World Bank Group, [Online]. Available: https://www.worldbank.org/en/about/what-we-do.

[3] UNAIDS, "Global HIV & AIDS statistics — 2019 fact sheet," UNAIDS, 2019. [Online]. Available: https://www.unaids.org/en/resources/fact-sheet.

[4] M. Hajizadeh, "Socioeconomic inequalities in child vaccination in low/middle-income countries: what accounts for the differences?," Epidemiol Community Health, 2018.

Appendix 1 (Metadata)

1. HNP_StatsData

Attribute Name	Type	Description
Country Code	Nominal	The code of each country, totally 258 countries.
Country Name	Nominal	The name of each country, totally 258 countries.
Indicator Code	Nominal	The code of each indicator, totally 403 indicators.
Indicator Name	Nominal	The name of each indicator, totally 403 indicators.
1960	Raito	The value of each indicator for each country in 1960.
…	Raito	The value of each indicator for each country in years before 2018 after 1960.
2018	Raito	The value of each indicator for each country in 2018.

2. HNP_StatsSeries

Attribute Name	Type	Description
Series Code	Nominal	The code of each indicator, totally 403 countries.
Indicator Name	Nominal	The name of each indicator, totally 403 indicators.
Topic	Nominal	The topic of each indicator.
Long definition	Nominal	The description of each indicator
…		Others

3. **HNP_StatsCountry**

Attribute Name	Type	Description
Country Code	Nominal	The code of each country, totally 258 countries.
Country Name	Nominal	The name of each country, totally 258 indicators.
…		Others

Appendix 2 (Code)

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Fri Nov 22 23:26:50 2019

@author: zhanglong
"""

import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
import statsmodels.api as sm
from sklearn import metrics
from scipy.stats import pearsonr
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures

def get_indicator_info(indicator_code):
    info = pd.DataFrame(columns = ["Series Code","Topic", "Indicator Name", "Long
definition", "Statistical concept and methodology"])
    Series_Code = list(HNP_StatsSeries.loc[:,"Series Code"])
    for code in indicator_code:
        index = Series_Code.index(code)
        info = info.append(HNP_StatsSeries.loc[index,["Series Code","Topic",
"Indicator Name", "Long definition", "Statistical concept and methodology"]])
    return info;

# country_name code
HNP_StatsCountry = pd.read_csv('HNP_StatsCountry.csv')
country_code = list(HNP_StatsCountry.iloc[:,0])

# ============================================================================
# # retrieve HIV data in 2018
# ============================================================================
# retrive indicators which contain HIV or AIDS
HNP_StatsSeries = pd.read_csv('HNP_StatsSeries.csv')
HIV_indicator = pd.DataFrame()
for i in range(len(HNP_StatsSeries)):
    if "HIV" in HNP_StatsSeries.iloc[i,2] or "AIDS" in HNP_StatsSeries.iloc[i,2]:
        HIV_indicator = HIV_indicator.append(HNP_StatsSeries.iloc[i,:])
HIV = list(HIV_indicator["Series Code"])
```

```
# retrive tuples related to HIV in 2018
HNP_StatsData = pd.read_csv('HNP_StatsData.csv')
HIV_Data = pd.DataFrame()
for i in range(len(HNP_StatsData)):
    if HNP_StatsData.iloc[i,3] in HIV :
        HIV_Data = HIV_Data.append(HNP_StatsData.iloc[i,:])
HIV_Data_2018 = HIV_Data.iloc[:,-6:-1]
print(HIV_Data_2018.dtypes)
# transform
ll = ["Country Code"]
ll = ll + HIV
HIV_Data_2018_2 = pd.DataFrame(columns = ll)
tt = [None]*len(ll);
m = 0
for i in country_code:
    HIV_Data_2018_2 = HIV_Data_2018_2.append(tt)
    n = 1
    HIV_Data_2018_2.iloc[m, 0]= i
    a = HIV_Data_2018.loc[HIV_Data_2018['Country Code'] == i]
    for j in HIV:
        for k in range(len(a)):
            if a.iloc[k,-2] == j:
                HIV_Data_2018_2.iloc[m, n] = a.iloc[k,0]
                n = n+1
    m = m+1
HIV_Data_2018_2 = HIV_Data_2018_2.reset_index(drop=True)
HIV_Data_2018_3 = HIV_Data_2018_2.iloc[0:len(country_code),:]
HIV_Data_2018_3 = HIV_Data_2018_3.iloc[:,:-1]
#for i in range(len(country_code)):
#    n = 1
#    HIV_Data_2018_2 = HIV_Data_2018_2.append(tt)
#    HIV_Data_2018_2.iloc[i, 0]= country_code[i]
#    a = HIV_Data_2018.loc[HIV_Data_2018['Country Code'] == country_code[i]]
#    for j in HIV:
#        for k in range(len(a)):
#            if a.iloc[k,-2] == j:
#                HIV_Data_2018_2.iloc[i, n] = a.iloc[k,0]
#                n = n+1

# count missing values for each column
HIV_missing = HIV_Data_2018_3.isnull().sum()

plt.bar(list(HIV_missing.index), list(HIV_missing))
plt.xticks(rotation=90)
plt.title('The number of missing values in each column in "HIV_Data_2018"')
plt.xlabel('Indicator Codes')
plt.ylabel('Number')



# drop indicators which have more than 200 missing values & save as HIV_Data_2018_4
```

```
HIV_missing_index = list(HIV_missing.index) # get the list of indicators' code
HIV_delete = [] # record deleted indicators' code
HIV_Data_2018_4 = HIV_Data_2018_3
for i in range(len(HIV_missing)):
    if HIV_missing[i] > 200:
        HIV_Data_2018_4 = HIV_Data_2018_4.drop(HIV_missing_index[i], axis = 1)
        HIV_delete.append(HIV_missing_index[i])
HIV_new = [] # record the rest indicators' code
for indicator in HIV:
    if indicator not in HIV_delete:
        HIV_new.append(indicator)
HIV_indicator_new = get_indicator_info(HIV_new) # retrive the information of non-
deleted indicators
HIV_indicator_new = HIV_indicator_new.reset_index(drop=True)


# ================================================================================
# # retrive society data in 2018
# ================================================================================
# retrive indicaters whose topic does not contain health
HNP_StatsSeries = pd.read_csv('HNP_StatsSeries.csv')
society_indicator = pd.DataFrame()
for i in range(len(HNP_StatsSeries)):
    if "Health" not in HNP_StatsSeries.iloc[i,1]:
        society_indicator = society_indicator.append(HNP_StatsSeries.iloc[i,:])
society = list(society_indicator["Series Code"])
# retrive tuples related to society indicators in 2018
HNP_StatsData = pd.read_csv('HNP_StatsData.csv')
society_Data = pd.DataFrame()
for i in range(len(HNP_StatsData)):
    if HNP_StatsData.iloc[i,3] in society :
        society_Data = society_Data.append(HNP_StatsData.iloc[i,:])
society_Data_2018 = society_Data.iloc[:,-6:-1]
society_Data_2018 = society_Data_2018.reset_index(drop=True)
print(society_Data_2018.dtypes)
# transform
nn = ["Country Code"]
nn = nn + society
society_Data_2018_2 = pd.DataFrame(columns = nn)
pp = [None]*len(nn);
m = 0
for i in country_code:
    society_Data_2018_2 = society_Data_2018_2.append(pp)
    n = 1
    society_Data_2018_2.iloc[m, 0]= i
    a = society_Data_2018.loc[society_Data_2018['Country Code'] == i]
    for j in society:
        for k in range(len(a)):
            if a.iloc[k,-2] == j:
                society_Data_2018_2.iloc[m, n] = a.iloc[k,0]
                n = n+1
```

```
    m = m+1
society_Data_2018_2 = society_Data_2018_2.reset_index(drop=True)
society_Data_2018_3 = society_Data_2018_2.iloc[0:len(country_code),:]
society_Data_2018_3 = society_Data_2018_3.iloc[:,:-1]

# count missing values for each column
society_missing = society_Data_2018_3.isnull().sum()

plt.figure(figsize=(15,5))
plt.bar(list(society_missing.index), list(society_missing))
plt.xticks(rotation=90)
plt.title('The number of missing values in each column in "Society_Data_2018"')
plt.xlabel('Indicator Codes')
plt.ylabel('Number')

# drop indicators which have more than 200 missing values & save as HIV_Data_2018_4
society_missing_index = list(society_missing.index)
society_delete = [] # record deleted indicators' code
society_Data_2018_4 = society_Data_2018_3
for i in range(len(society_missing)):
    if society_missing[i] > 200:
        society_Data_2018_4 = society_Data_2018_4.drop(society_missing_index[i],
axis = 1)
        society_delete.append(society_missing_index[i])
society_new = [] # record the rest indicators' code
for indicator in society:
    if indicator not in society_delete:
        society_new.append(indicator)
society_indicator_new = get_indicator_info(society_new) # retrive the information
of non-deleted indicators
society_indicator_new = society_indicator_new.reset_index(drop=True)

# ================================================================================
# # merge the HIV data and society data
# # save as Data_2018
# ================================================================================
Data_2018 = pd.merge(society_Data_2018_4, HIV_Data_2018_4, how='inner', on =
'Country Code')
Data_2018_2 = Data_2018
Data_2018_2 = Data_2018_2.loc[:,society_new + HIV_new].astype(float)
corr = Data_2018_2.corr()

a = sns.pairplot(Data_2018)

Data_2018_dropna = Data_2018_2.dropna(axis=0)
describe = Data_2018_dropna.describe()
corr = Data_2018_dropna.corr(method ='pearson')
plt.subplots(figsize=(25, 25))
sns.heatmap(corr, annot=True)

corr_society = Data_2018_dropna.loc[:,society_new].corr(method ='pearson')
```

```
plt.subplots(figsize=(10, 10))
sns.heatmap(corr_society, annot=True)

a = sns.pairplot(Data_2018_dropna)


# ================================================================================
# # explore SH.DYN.AIDS.FE.ZS
# ================================================================================
HNP_StatsData = pd.read_csv('HNP_StatsData.csv')
HIV_FE = pd.DataFrame()
for i in range(len(HNP_StatsData)):
    m = HNP_StatsData.columns.get_loc("Indicator Code")
    if HNP_StatsData.iloc[i, m] == "SH.DYN.AIDS.FE.ZS":
        HIV_FE = HIV_FE.append(HNP_StatsData.iloc[i, :])
n = HNP_StatsData.columns.get_loc("1986")
HIV_FE = HIV_FE.iloc[:,n:-1]
# a = HIV_FE.iloc[1,:]
HIV_FE = HIV_FE.dropna(axis=0)
HIV_FE = HIV_FE.iloc[:,0:-4]
# a = HIV_FE.iloc[1,:]
plt.figure(figsize=(7.5,5))
sns.boxplot(data=HIV_FE.iloc[:,[-26,-21,-16,-11,-6,-1]], palette='hot')
plt.title('Women's Share of Population Ages 15+ Living with HIV (%)')
plt.xlabel('Years')
plt.ylabel('Percent')

SHDY = pd.DataFrame()
n = HIV_Data.columns.get_loc("Indicator Code")
for i in range(len(HIV_Data)):
    if "SH.DYN.AIDS.FE.ZS" in HIV_Data.iloc[i,n]:
        SHDY =SHDY.append(HIV_Data.iloc[i,:])
m = HIV_Data.columns.get_loc("1990")
SHDY = SHDY.iloc[:,m:]
SHDY = SHDY.iloc[:,:-5]
SHDY =SHDY.T
SHDY = SHDY.dropna(axis=1)
mean = SHDY.mean(axis=1)
x = pd.DataFrame(list(range(1990,2019)))
plt.scatter(x, mean)
lr2 = LinearRegression()
y2 = mean.values.reshape(-1,1)
X2 = x
lr2.fit(X2, y2)
y_pred = lr2.predict(X2)
lr2.score(X2, y2)
lr2.coef_
lr2.intercept_
plt.figure()
plt.plot(x,y_pred,'b',label="Linear Regression")
plt.plot(x,mean,'r',label='Train')
```

```
plt.legend(loc='upper right')
plt.xlabel('years')
plt.title('Mean of "SH.DYN.AIDS.FE.ZS"')



# ============================================================================
# # regression SH.DYN.AIDS.FE.ZS
# ============================================================================
#lr2 = LinearRegression()
#y2 = Data_2018_dropna.loc[:,"SH.DYN.AIDS.FE.ZS"].values.reshape(-1,1)
#X2 = Data_2018_dropna.loc[:,society_new]
#lr2.fit(X2, y2)

y2 = Data_2018_dropna.loc[:,"SH.DYN.AIDS.FE.ZS"].values.reshape(-1,1)
X2 = Data_2018_dropna.loc[:,society_new]
line = sm.OLS(y2, X2)
result_line = line.fit()
result_line.summary()

y2 = Data_2018_dropna.loc[:,"SH.DYN.AIDS.FE.ZS"].values.reshape(-1,1)
X2 = Data_2018_dropna.loc[:,society_new]
line2 = sm.OLS(y2, sm.add_constant(X2)) # add constant
result_line2 = line2.fit()
result_line2.summary()


a = pearsonr(Data_2018_dropna.loc[:,"NY.GNP.PCAP.CD"],
Data_2018_dropna.loc[:,"SL.TLF.TOTL.FE.ZS"])

# adjust the model
X3 = Data_2018_dropna.loc[:,["NY.GNP.PCAP.CD", "SL.TLF.TOTL.FE.ZS", "SP.URB.GROW"]]
y3 = Data_2018_dropna.loc[:,"SH.DYN.AIDS.FE.ZS"].values.reshape(-1,1)
line = sm.OLS(y3, X3)
result_line = line.fit()
result_line.summary()

X3 = Data_2018_dropna.loc[:,["NY.GNP.PCAP.CD", "SL.TLF.TOTL.FE.ZS"]]
y3 = Data_2018_dropna.loc[:,"SH.DYN.AIDS.FE.ZS"].values.reshape(-1,1)
line = sm.OLS(y3, X3)
result_line = line.fit()
result_line.summary()

sns.regplot(x="NY.GNP.PCAP.CD", y="SH.DYN.AIDS.FE.ZS",data = Data_2018_dropna)

sns.regplot(x="SL.TLF.TOTL.FE.ZS", y="SH.DYN.AIDS.FE.ZS",data = Data_2018_dropna)

sns.regplot(x="SP.URB.GROW", y="SH.DYN.AIDS.FE.ZS",data = Data_2018_dropna)

# adjust the model
X4 = Data_2018_dropna.loc[:,["SL.TLF.TOTL.FE.ZS", "SP.URB.GROW"]]
```

```
y4 = Data_2018_dropna.loc[:,"SH.DYN.AIDS.FE.ZS"].values.reshape(-1,1)
line = sm.OLS(y4, X4)
result_line = line.fit()
result_line.summary()



# ==============================================================================
# # regression "SH.HIV.ARTC.ZS"
# ==============================================================================


# SH.HIV.ARTC.ZS: Antiretroviral therapy coverage (% of people living with HIV)
# NY.GNP.PCAP.CD: GNI per capita, Atlas method (current US$)
#lr = LinearRegression()
#y_test = Data_2018_dropna.loc[:,"SH.HIV.ARTC.ZS"].values.reshape(-1,1)
#X_test = Data_2018_dropna.loc[:,society_new]
#lr.fit(X_test, y_test)
#print(lr.intercept_)
#print(lr.coef_)
#zip(society_new,lr.coef_)

y_test = Data_2018_dropna.loc[:,"SH.HIV.ARTC.ZS"].values.reshape(-1,1)
X_test = Data_2018_dropna.loc[:,society_new]
line = sm.OLS(y_test, X_test)
result_line = line.fit()
result_line.summary()

# adjust
X_test2 = Data_2018_dropna.loc[:,["NY.GNP.PCAP.CD", "SP.URB.GROW"]]
y_test2 = Data_2018_dropna.loc[:,"SH.HIV.ARTC.ZS"].values.reshape(-1,1)
line = sm.OLS(y_test2, X_test2)
result_line = line.fit()
result_line.summary()

# RMSE
y_pred = result_line.predict(X_test2)
np.sqrt(metrics.mean_squared_error(y_test2, y_pred))
```