AIT 580
Fall 2016

Strategy to Combat HIV/AIDS

Yousef Hanif, Ashok Kari, Rama Chandra
Raju Rajasagi, Kevin Zhang

Contents

Introduction

The Human Immunodeficiency Virus (HIV) infection and acquired immune deficiency syndrome (AIDS) is a spectrum of conditions caused by infection with the human immunodeficiency virus. It spreads to people through unprotected sex, contaminated blood transfusions, and hypodermic needles among other methods. As it infects a host, it interferes with the immune system, increasing the risk of common infections like tuberculosis, as well as other opportunistic infections. Without treatment, the average survival time after infection is 11 years. As of today, HIV/AIDS is a worldwide pandemic in which there is no known cure or vaccine. In 2015, it is estimated that 36.7 million people were living with HIV and it resulted in 1.1 million deaths.

Despite the large amount of financial resources being poured into HIV/AIDS funding, there is no end in sight for the HIV/AIDS epidemic, as resources are divided up between different ways of advocating awareness and prevention, providing treatment to those already infected, and discovering a cure. Each organization is tackling HIV/AIDs prevention and treatment in individual silos, with minimal insight into other efforts, and are met with varying rates of success.

Through this study, we hope to track the spread of HIV/AIDS throughout the world and analyze the allocation of resources used to prevent and treat HIV/AIDS over time. We would consider this project a success if we can properly prescribe the best method for spending HIV/AIDS funding in order to help bring an end to the HIV/AIDS pandemic.

Team Planning

Role	**Team Members**
Project Manager	Yousef Hanif
Analysts	Yousef Hanif, Ashok Kari, Rama Chandra Raju Rajasagi, Kevin Zhang
Architect	Kevin Zhang
Statistician	Rama Chandra Raju Rajasagi

Developer	Ashok Kari

Problem Definition

How can we best allocate resources and funds to combat HIV/AIDS? Our plan is twofold. First, we will analyze HIV infections in countries across the world over time. We would like to capture which countries have contained HIV with decreasing infection rates and which countries have proliferating HIV rates. Secondly, we will investigate HIV Spending as well as HIV treatment in countries over time. This will provide insight into how countries are tackling the issue of HIV. Our evaluation criteria are measuring the number of cases of HIV within each country, calculating the rate of change in number of new cases over time, discovering the amount of funds allocated, and determining HIV prevention methods (treatment/lifestyle). We believe that if we can discover an allocation of resources that leads to a steadily decreasing rate of new HIV cases, we will be able to recommend this approach to organizations around the world in an effort to eradicate the AIDS virus.

Our biggest external constraints are resource constraints. Not all countries make their HIV funding statistics available. And even for those that do, it can be hard to track exactly how the funding is being spent. Additionally, there will be structural constraints as the data we collect must be cleaned before it can be used. However, we believe that once the data is cleaned, we will be able to leverage it to find the correlation between HIV spending and HIV spread. Since we are merely gathering publicly available data, there are no legal or cultural constraints.

Dataset Discovery

We have 3 main sources of data. There was no acquisition cost as we only used publicly available data. All data was of the text data type and already structured in the form of tables, so we could easily download them as csv files. Each data set was downloaded, cleaned, and stored in excel form in a shared drive for all team members to access. The actual collection cost was minimal, but significant data cleaning had to be performed, as detailed in the Solution Design section. The reputable data sources are as follows:

UNAIDS

UNAIDS is an innovative partnership that leads and inspires the world in achieving universal access to HIV prevention, treatment, care, and support. We utilized multiple datasets from UNAIDS, since it was the most comprehensive collection of publicly available data that we found, including data points from 162 countries around the world from a time period of 2004 – 2015. The data sets include: Number of people living with HIV, Number of new HIV infections, Coverage of people receiving antiretroviral therapy, HIV spending from domestic and international sources, HIV prevalence in inmates/detainees, as well as multiple datasets on prevention. This data was well-documented as all tables were properly labelled and formatted. Reasons, such as "Estimates were unavailable at time of publication" or "Child estimates were not published due to small numbers", were often given for missing data points. This data has a history of successful prior use, as it helps UNAIDS in its goals of reducing sexual transmission, preventing HIV among drug users, and eliminating new HIV infections across the globe.

World Health Organization

The World Health Organization is a non-profit organization that works with governments in over 150 different countries and other partners to ensure the highest attainable level of health for all people. We used the Health Financing data set to view health expenditure ratios by county. This data has a history of successful prior use as WHO continually pushes out publications in a variety of fields, from Eastern Mediterranean Health to Epidemiological Records.

UNICEF

UNICEF is a United Nations program that provides humanitarian and developmental assistance to children and mothers in developing countries. We used the global regional trends data to determine the top 20 burden countries for new HIV infections amongst adolescents. Their data is well documented, as it includes their estimate methodology.

Experimental Design

We aim to take HIV data from countries all over the world in the year 2015 and perform linear regressions as well as determine the correlations between HIV prevalence and a variety of factors. These factors include HIV funding, knowledge of HIV status, knowledge of HIV prevention, intimacy before the age of 15, condom usage, and HIV treatment. The goal is to find a model that will be able to determine the percentage of HIV prevalence within a given population. The independent variable is countries around the world, so we are conducting an experimental design with independent measures. The system is stable and finite, but not scalable. The lifecycle is annual, since 2016 will provide a new set of data.

We plan to reincorporate what we learn into our program by taking the top 20 countries for HIV prevention that we found and performing a time-series analysis to evaluate HIV prevalence vs HIV funding. The goal of this test is to determine how the best countries fair at preventing HIV over time. By introducing time as the independent variable, we measure the decrease of HIV prevalence in the top 20 countries. We will consider the project a success if we can conclude which country has lowered HIV prevalence the best over the years and prescribe their HIV funding allocation as the most favorable method for preventing HIV.

Solution Design

Design Overview

Data Collection, Storage, and Retrieval

We downloaded the data in the form of excel spreadsheets from various HIV data sources. We considered storing the data via a shared database, but after discussing implementations and security risks, decided to keep the data stored in excel files. We created a shared folder in Google Drive to put these data files where only team members had access.

Data Curation and Quality Assurance

The datasets used in this project were mainly from the UNAIDS source. The data was imported in excel from the website. Multiple excel functions were utilized to clean the data. Upon initially importing the data there was unnecessary space between numbers, parentheses, brackets, and text in the columns. The find and replace function was useful for cleaning the data so that it could be analyzed. Furthermore, the vlookup function was used to match data from multiple variables with the appropriate countries. When creating a regression and time series model, multiple datasets were combined into one excel sheet.

Statistical and Analytical Plan/Approach

One of the greatest challenges we faced was to ensure the quality of data when creating a regression and time series model. In order to create a regression model we had to ensure that there were no missing values. As a result, the number of countries had to be significantly decreased to be able to ensure all values for variables were present. The first linear regression model that will be explained in more detail in the evaluation methodology section utilized data from 49 different countries to create a model with funding as a predicting variable. The second linear regression model with multiple prevention factors as the predicting variables utilized data from 25 different countries.

Visualization and Reporting

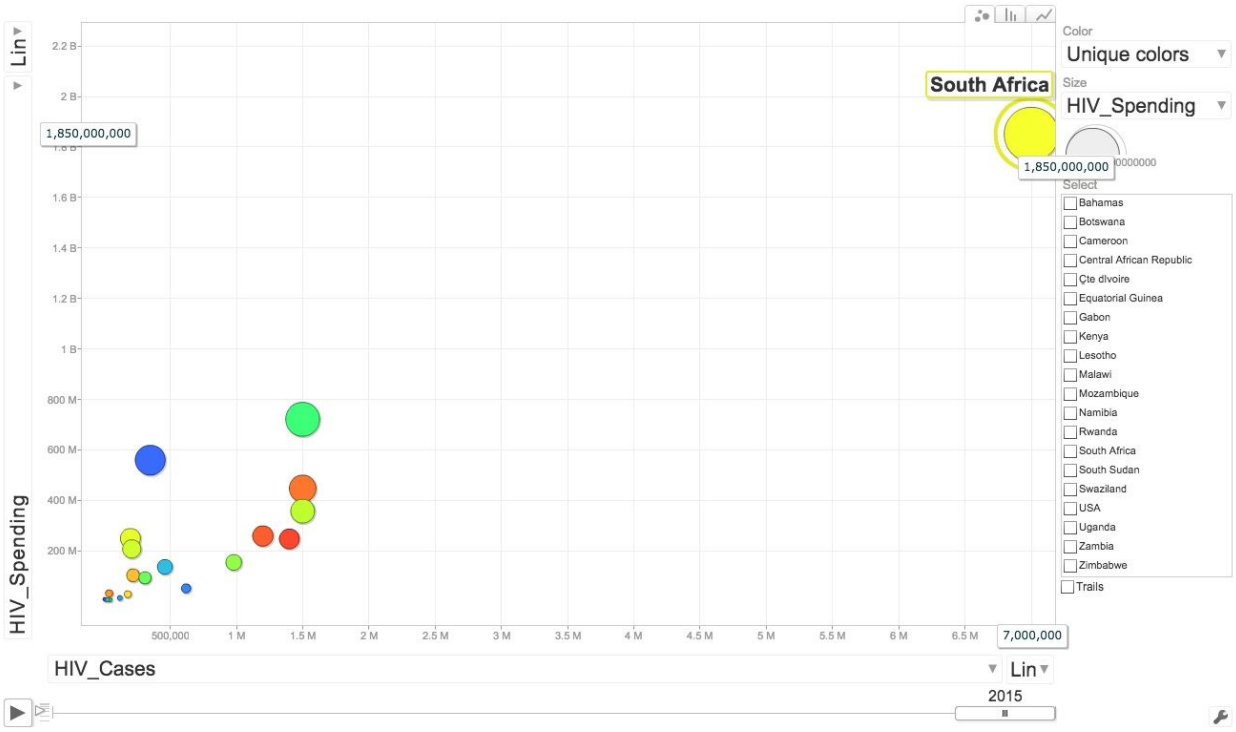
Linear regressions and correlations were created to determine the relationship between various variables and the ability to predict the percentage of HIV prevalence within a given population. Furthermore, the linear regression method was determined as a great method of predicting the relationship between funding and HIV prevalence to help address our problem statement. The first linear regression that was created with funding as a predictor variable did not appear to be a good model as the r-squared value was low. As a result, the analyzing modified to include different predictor variables. The second linear regression model created focused on data collected for various prevention measures.

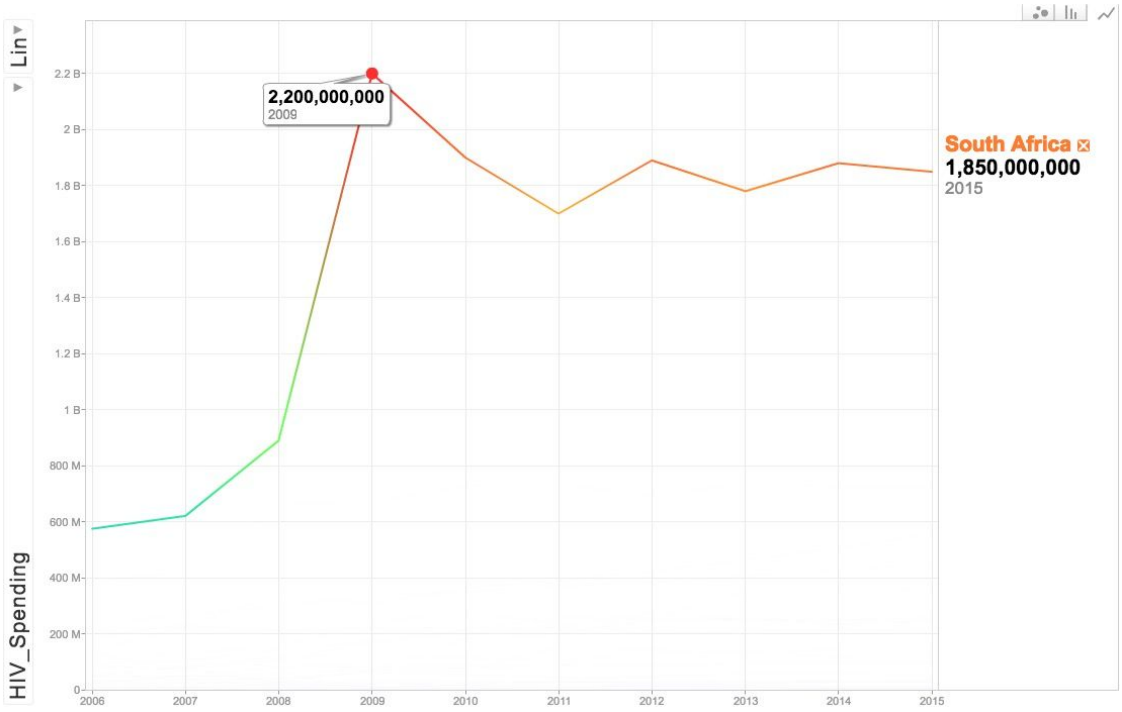
Time-Series Analysis

We conducted time series analysis using Google Visualization package in R. Data from sources listed above is used for this visualization. Total number of HIV cases and Spending for HIV related issues every year is taken for 20 countries which has the highest HIV Prevalence. In the plot X-Axis represent total number of HIV cases and Y-Axis represent HIV Spending. Each circle represents a country and all countries are color encoded. These are the key takeaways from this time-series analysis.

Key takeaways

- Slight Linear Trend between HIV Cases and HIV Spending
- South Africa is the top country in both no.of cases and spending
- South Africa significantly increased their spending between 2008 and 2009
- Despite the huge increase in the HIV spending by South Africa there is no decline in the total number of cases
- Countries like Central African Republic and USA have decline in the total no.of cases over the years, rest of the countries are on the growth scale
- Central African Republic saw the decline in the number of cases without increasing their spending, there should be more key factors associated with this outcome.





Case Study

One of the risks we ran into was not finding enough data from the United States to incorporate into our linear regressions. This is significant because the United States is one of the leaders in HIV prevention and we wanted to compare the US HIV rates vs those of other countries. We attempted to mitigate this risk by trying to discover other ways to include US HIV data into our study. Therefore, as a part of our analysis we wanted to take a look into a success story which can be shown as an example. During the initial stages of our analysis we have come across some interesting data about HIV in US Federal and State prisons. This following data made us look into HIV in Prisons:

There is huge decrease in AIDS related deaths in US prisons despite of prisons being a place with less AIDS awareness, safety and medical facilities. It went so low that by the end of 2009 it was less than that the general population.
Starting with a vision by National HIV/AIDS Strategy: “The United States will become a place where new HIV infections are rare and when they do occur, every person, regardless of age, gender, race/ethnicity, sexual orientation, gender identity, or socio-economic circumstance, will have unfettered access to high quality, life-extending care, free from stigma and discrimination.”
USA has gone a long way from the rate of HIV/AIDS among state and federal prison inmates 194 cases per 10,000 inmates in 2001 to 146 per 10,000 at yearend 2010. Between 2001 and

2010, the average annual decline of 16% in the national AIDS mortality rate was similar to the decline in small (down 12%), medium (down 17%), and large (down 19%) state prison populations.

There is an average notable decline of 3% each year from 2001 to 2010 in the estimated rate of HIV/AIDS among state and federal prison inmates, and 16% of death rate annually in among the same set. Which is around 24 deaths per 100,000 inmates in 2001 to 5 per 100,000 in 2010.

There a number of reasons for this change, starting with allocation of huge budget, involvement of private organizations, Funding community-based pilot projects, Voluntary rapid HIV testing, recommendation on condom distribution programs along At least 35 states have criminal laws that punish HIV-positive people for exposing others to the virus, even if they take precautions such as using a condom.
So by the end of 2015 we have our jail system with less 1.5% of jail population with HIV AIDS in US

Evaluation Methodology

In order to determine if there is a relationship between percent of population with HIV in a given country and funding, a linear regression model was created based on data from multiple countries. The determining variable in this model was the percentage of the population with HIV in a given country. The predicting variables were the following: percentage of the population affected that had knowledge of their HIV status, HIV funding, and percentage of affected population receiving HIV treatment. The following model was outputted from R:

From this model we can see that the adjusted r-squared value is small indicating that the model created does not have a good fit.

The data was also analyzed for correlations between the various variables. The correlation values show that there is a strong correlation between knowledge of status and treatment. To ensure that there was no multicollinearity, the variance inflation factor (VIF) was calculated to be less than 3 for all variables.

Security

There are several security risks associated with online databases. One of these risks is unencrypted database files. If database files are left unencrypted, they're vulnerable to the same sort of virus and spyware attacks that other computer files are. A common method to account for this is to encrypt database tables. This is a prevailing practice for online databases today. A second security risk is SQL injection attacks. The only effective prevention against this is due diligence on the part of the web programmer. Thirdly, unsecured database backups also pose a security threat. One of the best prevention methods for this is to make the backups offline.
Since we do not control how UNAIDS, WHO, or UNICEF store their data, it's hard for us to personally adjust our methods to account for security risks. However, we mitigated any further security risks by securely storing all our files, including data files as well as documentation, in a shared Google Drive folder with the security settings restricted to just members of our group. Additionally, when we found sets of overlapping data, we cross checked the data between the 2 sets to make sure they matched. This is just a cautionary measure taken to ensure data integrity.

Privacy and Ethics

The main ethical issue is that big data analytics are not 100 percent accurate, however, unethical actions can sometimes be based on these interpretations. For example, we may find that spreading awareness is overall more effective at lowering HIV prevalence than investing in protease inhibitors. So we might want to prescribe a reallocation of resources from producing protease inhibitors to educating civilians about HIV. However, in the country of South Africa, where there are already millions living with HIV, reducing the funding of protease inhibitors could lead to a lack of protease inhibitors for everyone who needs them, resulting in a greater number of deaths. Therefore, we must be cautious about prescribing a treatment plan and make sure that we take all environmental factors into consideration.
Privacy was not a huge consideration on our part, since we are using exclusively publicly available data. However, the privacy of HIV/AIDS organizations did make it difficult for us to provide a proper prescription. Even though we were able to find HIV Prevalence, treatment, and funding numbers overall, it was very difficult to track exactly which organizations received the funds and what the funds were spent on. HIV organizations were not transparent about the allocation of funds and we were not able to find the information in any publicly available

databases. If we had more time, we might be able to investigate further by reaching out to specific organizations.

Cost and Project Planning/Implementation

Our project plan was to have our problem definition ironed out by October 11 and our solution design and evaluation methodology set by November 1. We would then perform the data cleaning and analysis and have our results November 29. We would spend the remaining time addressing any risks we encountered, writing up the results, and forming a conclusion.
There were no costs for this project. The return on investment is transparency into how HIV funds are spent by countries around the world. This can help open doors for future research into how best to allocate funds for HIV treatment and prevention.

Results

In the correlation plots below Swaziland can be seen as an outlier, however understanding the data gives a clearer picture of why it is not baseline. Swaziland has a high percentage of HIV prevalence within its population and has addressed this through funding and treatment. The funding does not appear significant in relation to other countries due to the relatively low population.
In the second linear regression model the percentage of population with HIV in a given country was still held as the determining variable. The predicting variables changed to the following prevention factors: knowledge about HIV prevention amongst (15-24), intimate before age 15 amongst(15-24), condom use at last intimacy (15-49) with multiple partners, and percentage of affected receiving treatment. In this method the r-squared value was calculated to be 0.31 showing that this model is better for predicting the percentage of HIV prevalence within a population. While a value of 0.31 is still relatively low for goodness of fit in the real world it can still be held significant.

The following correlation plots were created consisting of the variables in the second linear regression model. This correlation plot also shows Swaziland out of population. The positive side of this is that Swaziland is addressing this issue by providing a large amount of treatment and ensuring there population is more knowledgeable about HIV prevention.

Analyzing both linear regression models we can see that second linear regression model with data on variables related to prevention is a better model. From this we can conclude that funding is not a good predictor of HIV prevalence, however we were able to determine specific preventive measures that would provide better use of funds.

Recommendations and Future Directions

While we were not able to correlate HIV funding in general to HIV prevalence, we did find several specific preventative measures that proved effective. It would be worthwhile to take a deeper look into some of these measure, such as knowledge of living with HIV and use of condoms. We hypothesize that the lack of correlation between HIV funding and HIV prevalence could be due to a high percentage of funds going to HIV treatment instead of HIV prevention. It has proved difficult to track exactly how much is allocated to each segment, as the US government alone spreads it funding to dozens of programs. However, we did find that the expenses on medication, housing, and support, treatment takes up a significant portion. Increasing the effectiveness of HIV prevention could result in a substantial decrease in the overall funding required. Future studies could focus on a smaller scope- looking specifically at populations where HIV rate is increasing most rapidly and studying HIV prevention methods in such regions.

Conclusion

The FDA has approved 39 different medicines to treat HIV. The US spent $19 billion in 2015 alone to help combat HIV/AIDS. This is because there is still an HIV epidemic and it remains a major issue for countries across the world. 36.7 million people are currently living with HIV and 18.2 million people are receiving antiretroviral treatment worldwide. Through this project, we have learned the amount of HIV funding does not necessarily correlate to HIV prevalence. In addition, it would be worthwhile to further explore specific prevention factors, as HIV education had the strongest correlation to HIV prevalence among the HIV containment factors we measured. Although we were not able to recommend one specific prevention plan, we hope that we shed light on this very real issue and set the foundation for future research to be able to focus more on specific prevention factors in an effort to reduce new infections, increase access to care

and improve health outcomes for people living with HIV, reduce HIV-related health disparities and health inequities, and achieve a more coordinated global response to the HIV epidemic.

Appendix

Graphs and Visualizations

WHO Regional Offices

R Code

Linear Regression Model 1

```
mydata<-read.csv("HIV_Treatment_Funding_Knowledge.csv",header=TRUE)
summary(mydata)
myvars = c(2:5)
mydatanumeric = mydata[myvars]
cor(mydatanumeric)
library(usdm)
vif(mydatanumeric)

par(mfrow=c(1,1))
plot(mydata[,3], mydata[,2], main="KnowStatus vs. HIV Prevalence", xlab="People living with HIV who know Status", ylab="Prevalence")
plot(mydata[,4], mydata[,2], main="Funding vs. HIV Prevalence", xlab="HIV Funding", ylab="Prevalence")
plot(mydata[,5], mydata[,2], main="Treatment vs. HIV Prevalence", xlab="Receving Treatment", ylab="Prevalence")

trainfit<-lm(Prevalence ~., data=mydatanumeric)
trainfit
```

```
plot(trainfit)
summary(trainfit)
```

Linear Regression Model 2

```
mydata<-read.csv("HIVregression.csv",header=TRUE)
summary(mydata)
myvars = c(2:7)
mydatanumeric = mydata[myvars]
cor(mydatanumeric)
library(usdm)
vif(mydatanumeric)

par(mfrow=c(2,2))
plot(mydata[,3], mydata[,2], main="Knowledge about HIV prevention among 15-24 vs. HIV Prevalence", xlab="Knowledge", ylab="Prevalence")
plot(mydata[,4], mydata[,2], main="Intimate before age 15 amongst (15-24) vs. HIV Prevalence", xlab="Intimacy before 15", ylab="Prevalence")
plot(mydata[,5], mydata[,2], main="Condom use last intimacy amongst (15-49) with Multiple Partners vs. HIV Prevalence", xlab="Condom Use", ylab="Prevalence")
plot(mydata[,6], mydata[,2], main="Treatment vs. HIV Prevalence", xlab="Receving Treatment", ylab="Prevalence")

trainfit<-lm(Prevalence ~., data=mydatanumeric)
trainfit
summary(trainfit)
```

Google Visualization

```
install.packages("devtools")
require(devtools)
install.packages('Rcpp')
require(Rcpp)
install_github('ramnathv/rCharts', force = TRUE)
require(rCharts)
install.packages("googleVis")
require(googleVis)

# Custom settings for the motion chart
myStateSettings <-'
{"xZoomedDataMin":1199145600000,"colorOption":"2",
"duration":{"timeUnit":"Y","multiplier":1},"yLambda":1,
"yAxisOption":"4","sizeOption":"_UNISIZE",
"iconKeySettings":[],"xLambda":1,"nonSelectedAlpha":0,
"xZoomedDataMax":1262304000000,"iconType":"LINE",
"dimensions":{"iconDimensions":["dim0"]},
"showTrails":false,"uniColorForNonSelected":false,
"xAxisOption":"_TIME","orderedByX":false,"playDuration":150000,
"xZoomedIn":false,"time":"2010","yZoomedDataMin":0,
"yZoomedIn":false,"orderedByY":false,"yZoomedDataMax":100}'

# Fox News
casesVsSpending <- read.csv("HIV_Cases_Vs_Spending.csv", header = TRUE)
casesVsSpending$Year <- as.numeric(casesVsSpending$Year)
summary(casesVsSpending)

casesVsSpendingVis = gvisMotionChart(casesVsSpending,
                    idvar="Country",
                    timevar="Year",
                    options = list(width=1300, height=760))

# Plotting Google Visualization
plot(casesVsSpendingVis)
```

Bibliography

- AIDS.gov: https://www.aids.gov/hiv-aids-basics/
- Bureau of Justice Statistics: https://www.bjs.gov/index.cfm?ty=pbse&sid=7
- Center for Disease Control and Prevention: https://www.cdc.gov/hiv/
- Qubole: https://www.qubole.com/
- UN AIDS: http://www.unaids.org/
- UNICEF: https://www.unicef.org/
- World Health Organization: http://www.who.int/en/