Midterm Project

Chuning Yuan

10/19/2019

The Question

Is there a correlation between GDP per Capita and prevalence of HIV in the 15–49 age bracket? And if yes, how strong is that correlation? How about the correlation between GDP per Capita and prevalence of HIV in female?

The Data

In this data analysis, I use the data available on World Bank webpage.Here are a brief description of each data set:

GDP/capita : GDP per capita is gross domestic product divided by midyear population. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. Data are in current U.S. dollars.

Prevalence of HIV among people aged 15–49 (%): Prevalence of HIV refers to the percentage of people ages 15-49 who are infected with HIV.

Prevalence of HIV, female among people aged 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.

The following code shows how I download the dataset:

```
#getdata
new_wdi_cache <- WDIcache()
WDIsearch("Prevalence of HIV, total")
```

```
##      indicator
## [1,] "SH.DYN.AIDS.ZS"
## [2,] "HF.DYN.AIDS.ZS.Q5"
## [3,] "HF.DYN.AIDS.ZS.Q4"
## [4,] "HF.DYN.AIDS.ZS.Q3"
## [5,] "HF.DYN.AIDS.ZS.Q2"
## [6,] "HF.DYN.AIDS.ZS.Q1"
## [7,] "HF.DYN.AIDS.ZS"
##      name
## [1,] "Prevalence of HIV, total (% of population ages 15-49)"
## [2,] "Prevalence of HIV, total (% of population ages 15-49): Q5 (highest)"
## [3,] "Prevalence of HIV, total (% of population ages 15-49): Q4"
## [4,] "Prevalence of HIV, total (% of population ages 15-49): Q3"
## [5,] "Prevalence of HIV, total (% of population ages 15-49): Q2"
## [6,] "Prevalence of HIV, total (% of population ages 15-49): Q1 (lowest)"
## [7,] "Prevalence of HIV, total (% of population ages 15-49)"
```

```
WDIsearch("Prevalence of HIV, female")
```

```
##                                  indicator
##                        "SH.HIV.1524.FE.ZS"
##                                       name
## "Prevalence of HIV, female (% ages 15-24)"
```

```
WDIsearch("gdp.*capita.*US\\$", cache = new_wdi_cache)
```

```
##      indicator        name
## [1,] "NY.GDP.PCAP.CD" "GDP per capita (current US$)"
## [2,] "NY.GDP.PCAP.KD" "GDP per capita (constant 2010 US$)"
```

Combing dataframe & data cleaning

We remove all entries that are aggregated regional values and then we rename the indicators.Then we combine the three dataframes to allow us to compare GDP per capita and HIV prevalence and HIV prevalence, female.

```
#cleandata
wdi_data <- WDI(indicator = c("NY.GDP.PCAP.CD","SH.DYN.AIDS.ZS","SH.HIV.1524.FE.ZS"), start = 1960, end
names(wdi_data)
```

```
##  [1] "iso2c"             "country"           "year"
##  [4] "NY.GDP.PCAP.CD"    "SH.DYN.AIDS.ZS"    "SH.HIV.1524.FE.ZS"
##  [7] "iso3c"             "region"            "capital"
## [10] "longitude"         "latitude"          "income"
## [13] "lending"
```

```
wdi_data <- subset(wdi_data, region != "Aggregates")
names(wdi_data)[which(names(wdi_data) == "NY.GDP.PCAP.CD")] <- "GDP"
names(wdi_data)[which(names(wdi_data) == "SH.DYN.AIDS.ZS")] <- "HIV_total"
names(wdi_data)[which(names(wdi_data) == "SH.HIV.1524.FE.ZS")] <- "HIV_female"

data=na.omit(wdi_data)
names(data)
```

```
##  [1] "iso2c"      "country"    "year"       "GDP"        "HIV_total"
##  [6] "HIV_female" "iso3c"      "region"     "capital"    "longitude"
## [11] "latitude"   "income"     "lending"
```

Datasets often feature missing data.So we need to take a look at the percentage of missing data in the combined dataframe.

About 24.3% of the GDP per Capita column in the combined dataframe have missing data. This is quite substantial and is most likely due the fact that consistent measurements of GDP are costly and have only started in the last few decades according to the World Bank webpage (It was found in 1945 after WWII). And we have an at 68.5% for HIV Prevalence missing data and 69.1% for HIV Prevalence of female in the combined dataframe. The lag in consistent measurements of HIV associated metrics have only really been performed on a large scale from the early-1980s when HIV/Aids became a recognised major health crisis.

Observed the distribution of each data sets

```
p1=ggplot(data, aes(x = GDP) ) +
  geom_histogram(bins = 100,fill="darkblue")
summary(data$GDP)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##     95.19    663.73   1943.92   7749.73   6651.29 118823.65
```

```
p2=ggplot(data, aes(x = HIV_total)) +
  geom_histogram(bins = 20,fill="darkblue")
```

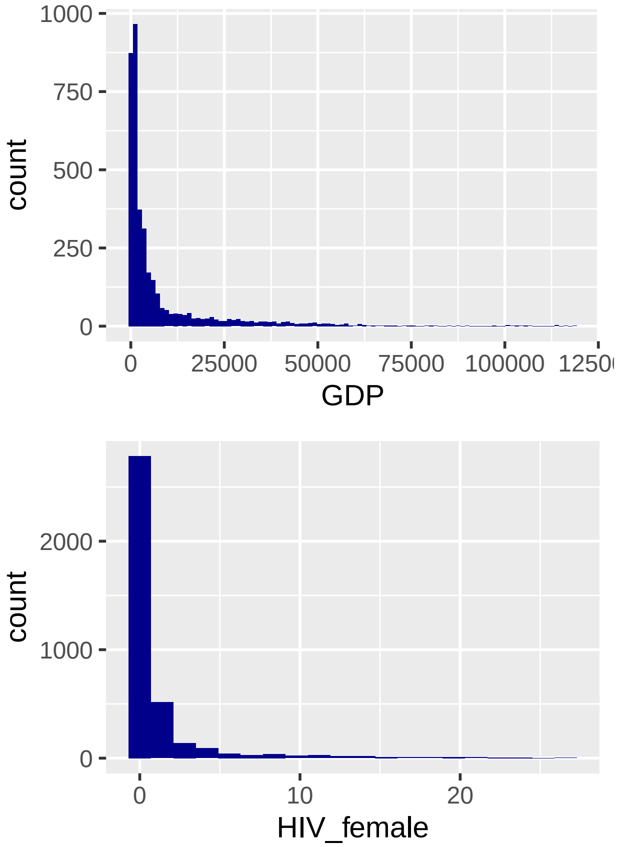
```
summary(data$HIV_total)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.100   0.100   0.300   1.977   1.500  28.200
```

```
p3=ggplot(data, aes(x = HIV_female)) +
  geom_histogram(bins = 20,fill="darkblue")
summary(data$HIV_female)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.100   0.100   0.100   1.314   0.900  26.700
```

```
gridExtra::grid.arrange(p1,p2,p3,ncol= 2)
```



The results above shows that some countries in our dataframe are considerable pulling the distribution's mean up (compared to the median). This is particularly the case with the HIV Prevalence data.

Big picture of the data

```
#data transformation
growth <- data %>% group_by(year) %>%
summarize(HIV_total=mean(HIV_total), HIV_female=mean(HIV_female), GDP=mean(GDP))


kable(growth, digits = 4, align = "c",booktabs=TRUE ,caption = "WorldwideTrend ",col.names = c("year",
```

Table 1: WorldwideTrend

year	HIV_female	HIV_total	GDP
1990	1.0066	0.9426	4115.235
1991	1.2132	1.1248	4240.756
1992	1.4240	1.3025	4433.751
1993	1.6016	1.4512	4241.299
1994	1.7750	1.5863	4485.679
1995	1.8608	1.6369	5089.958
1996	1.9962	1.7192	5242.818
1997	2.1085	1.7662	5176.325
1998	2.1908	1.7708	5118.483
1999	2.2392	1.7492	5249.898
2000	2.2398	1.6820	5123.061
2001	2.2474	1.6233	5063.179
2002	2.2157	1.5403	5309.017
2003	2.1925	1.4687	6230.764
2004	2.1597	1.3955	7183.352
2005	2.1328	1.3284	7911.749
2006	2.1075	1.2687	8606.707
2007	2.0858	1.2179	9949.055
2008	2.0881	1.1851	11002.773
2009	2.0754	1.1515	9719.329
2010	2.0716	1.1239	10275.142
2011	2.0657	1.1000	11387.624
2012	2.0586	1.0752	11383.704
2013	2.0343	1.0478	11600.025
2014	2.0231	1.0269	11628.993
2015	1.9985	0.9993	10320.637
2016	1.9754	0.9731	10366.040
2017	1.9459	0.9421	11128.021
2018	1.9570	0.9352	11433.067

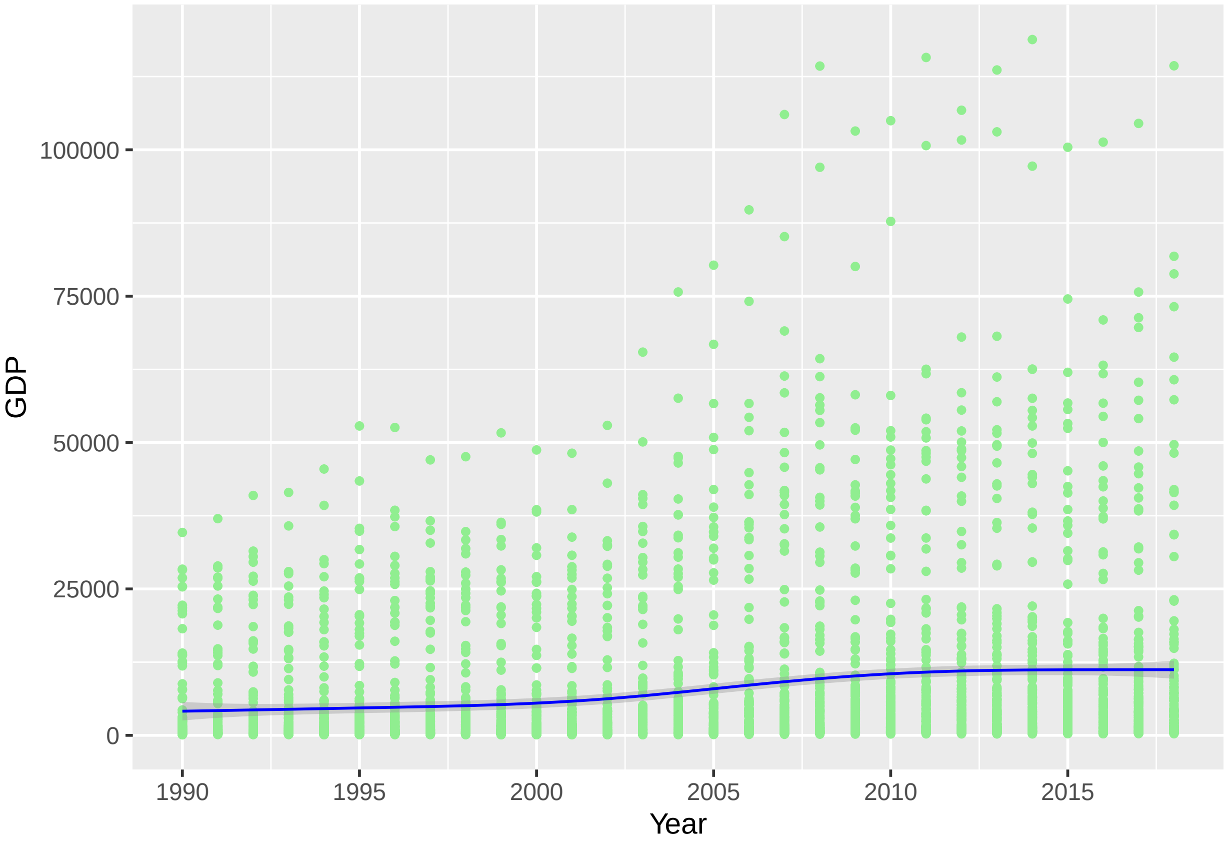
The table above listed all the averge GDP and average HIV prevalence in total and female in each year. From the table we can see the basic trend of the GDP growth and the precentage of the prevalence of HIV growth from 1990-2018. After 2009 the prevalence HIV in total and in felmale has goes down slightly each year while the GDP are still keep growing,in my opinion, this might be the new invention of the hpv immunizations begin to be popular.

The Plots

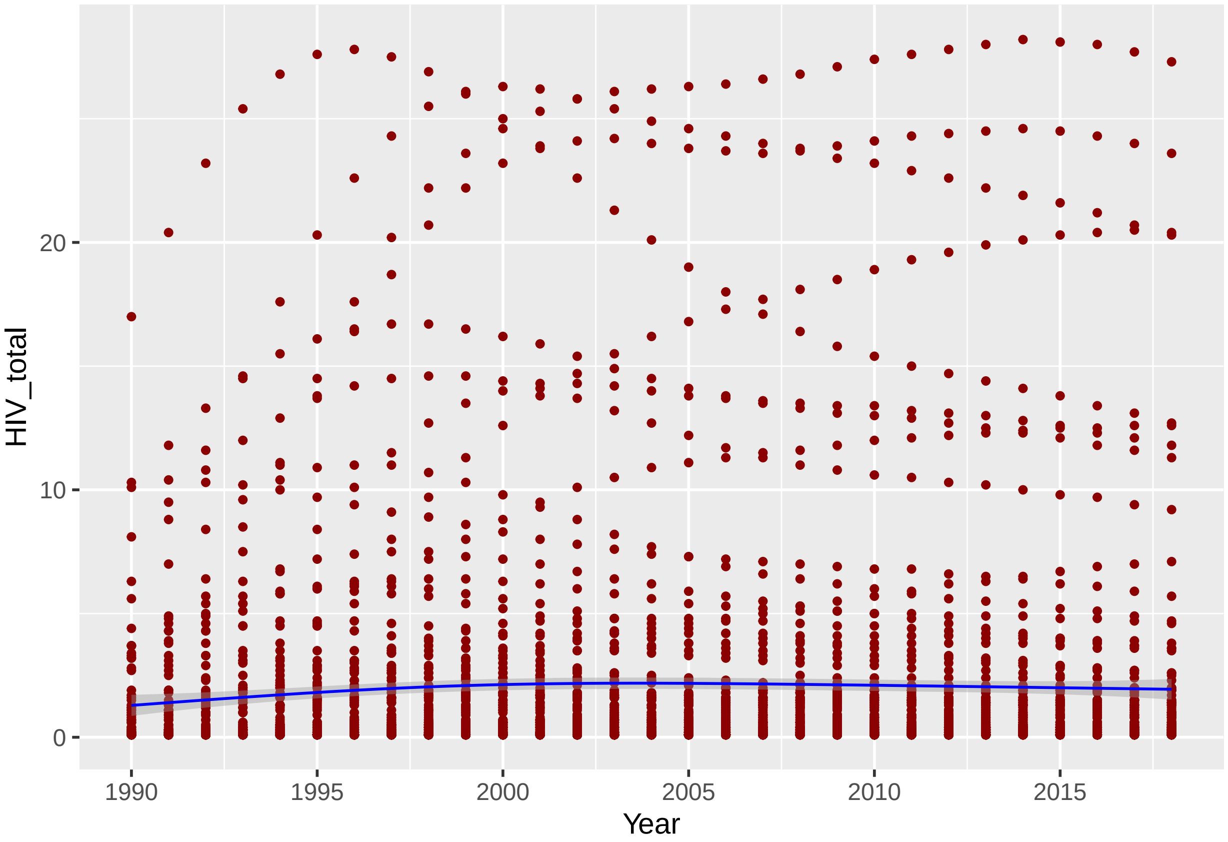
Scatter plots

I generate various plots to get an overview of the distribution and attempt to identify trends and patterns. First, we look at the overall data set and generate a scatter plot of GDP per Capita and HIV Prevalence in for the 136 countries listed in our dataset from 1990 to 2018.

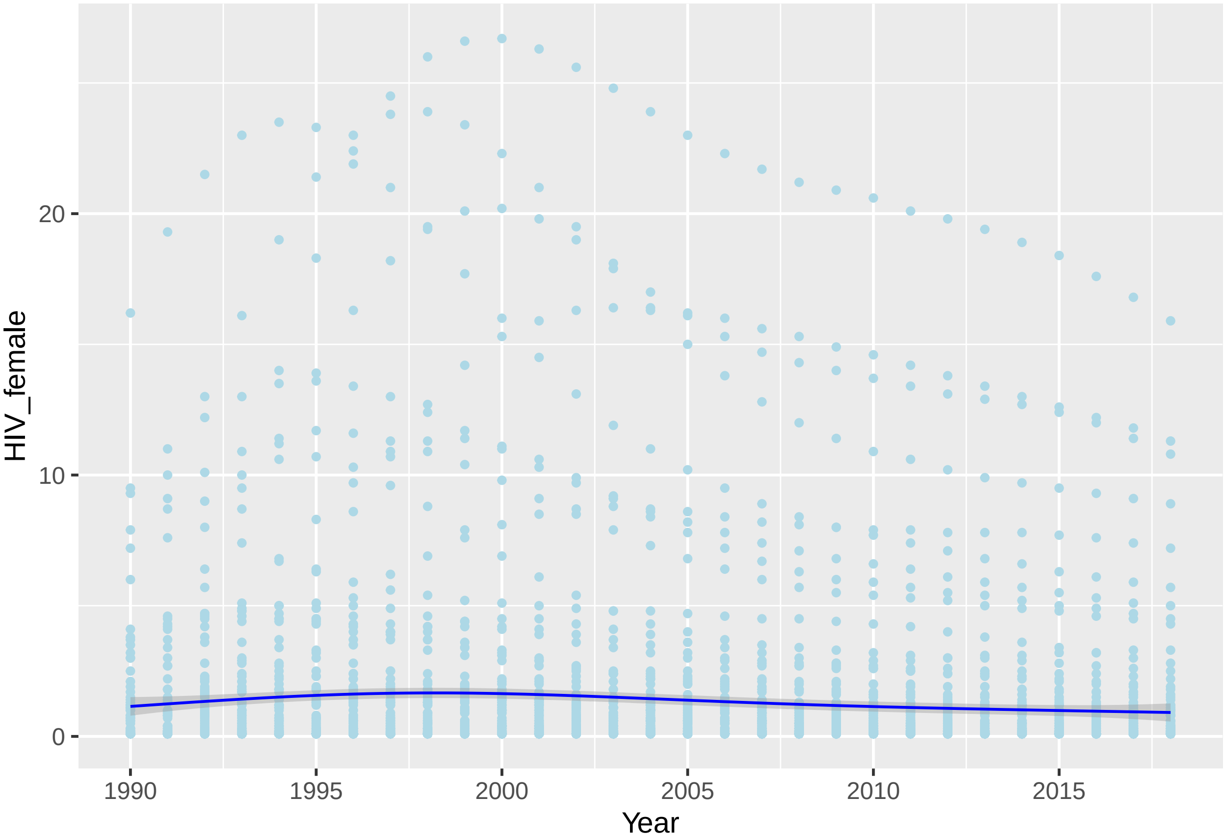
```
#scatter plot of GDP per Capita
ggplot(data, aes(x = year, y = GDP), title = "GDP")+
  geom_point(size = 1,color="lightgreen") + scale_x_continuous("Year",breaks = seq(1990,2018,5)) + geom_
```



```
#scatter plot of Prevalence of HIV total
ggplot(data, aes(x = year, y = HIV_total), title = "HIV_total")+
  geom_point(size = 1,color="darkred") + scale_x_continuous("Year",breaks = seq(1990,2018,5)) + geom_sm
```



```
#scatter plot of Prevalence of HIV female
ggplot(data, aes(x = year, y = HIV_female), title = "HIV_female")+
  geom_point(size = 1,color="lightblue") + scale_x_continuous("Year",breaks = seq(1990,2018,5)) + geom_s
```



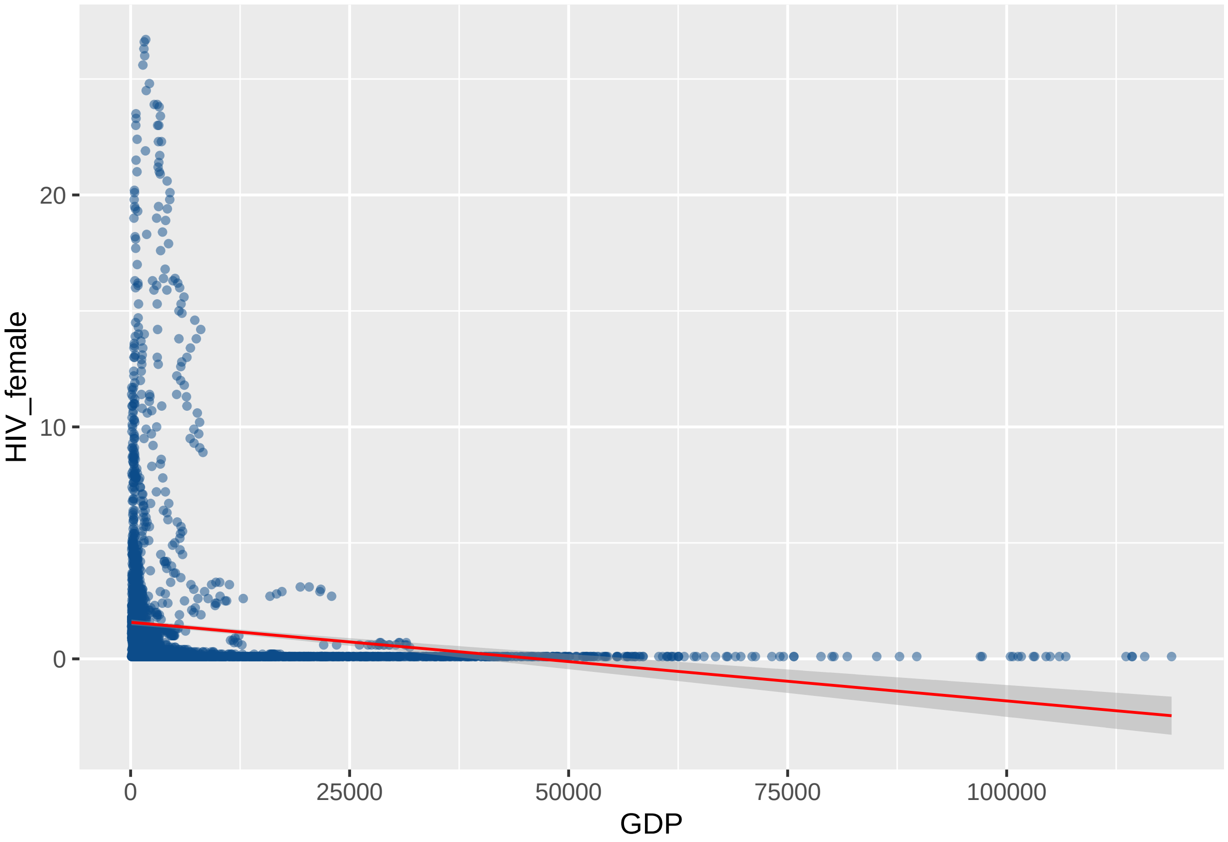
According to our data, the rate of HIV prevalence has slight increased between 1990 and 2010 with a stagnation in the mean from the early 2000s and slight decline since 2005. This would correspond to advances in preventative measures to reduce the incidence and likelihood of contracting HIV.

Correlation and Linear Regression

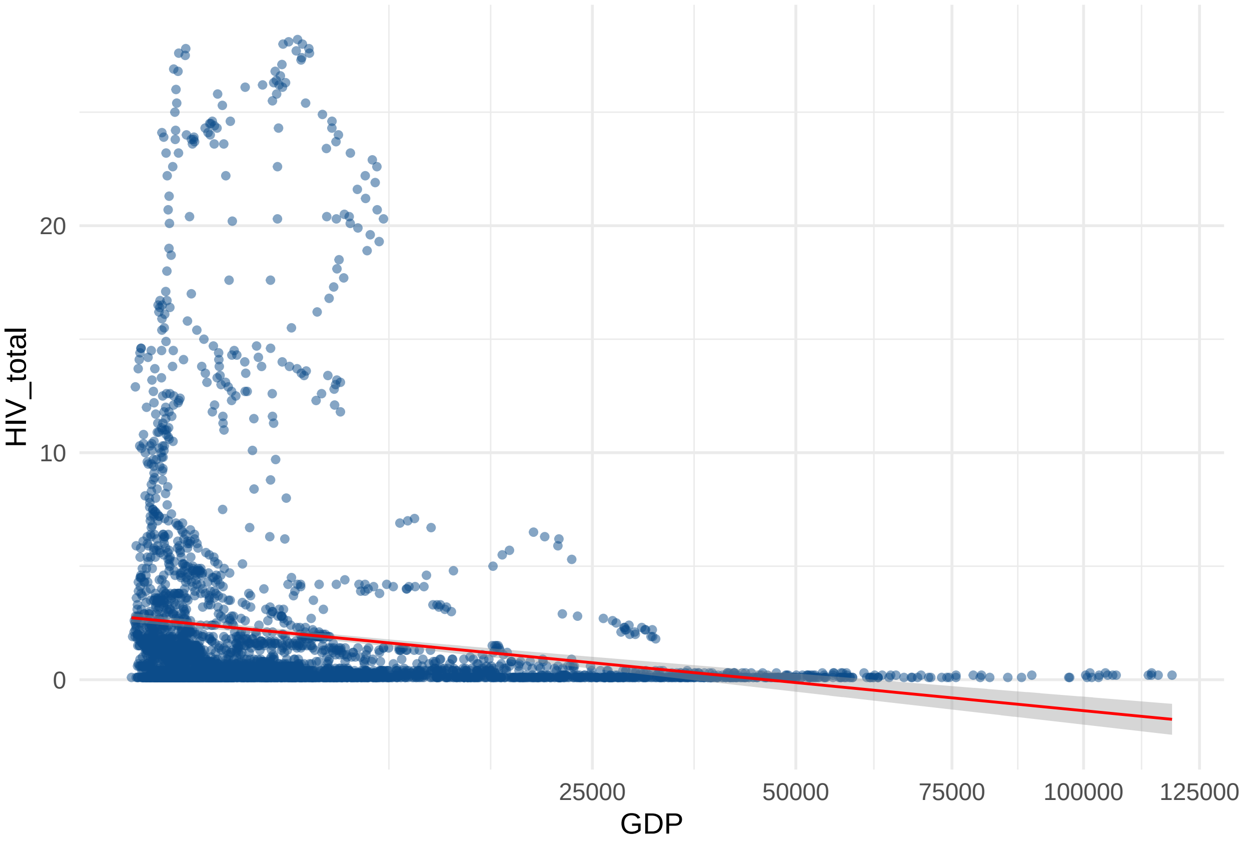
The scatter plot clearly indicates that the lower GDP per Capita data points (i.e. countries) have a much higher HIV prevalence compared countries with higher GDP per Capita.

Let’s take a closer look by creating a plot with a square root scale applied to the x-axis to further emphasise countries with lower GDP per capita. We’ll also use the R function geom_smooth() to perform a simple linear regression to better visualise the relationship between the two variables.

```
#plot HIV Prevalance vs. GDP per Capita.
ggplot(data, aes(x = GDP, y = HIV_female)) + geom_point(size = 1L,alpha=0.5,colour = "#0c4c8a") +geom_sm
```



```r
# square root scale of GDP
ggplot(data) +
 aes(x = GDP, y = HIV_total) +
 geom_point(size = 1L, colour = "#0c4c8a",alpha=0.5) +
 scale_x_continuous(trans = "sqrt") +
 theme_minimal()+geom_smooth(method = 'gam',color="red",size=0.5)
```



The scatter plot above further indicates that countries which smaller GDP per capita have on average higher HIV prevalence.

```
#plot HIV Prevalance female vs. GDP per Capita.
p4=ggplot(data, aes(x = GDP, y = HIV_female)) + geom_point(size = 1L,alpha=0.5,colour = "#0c4c8a") +geom

# square root scale of GDP
p5=ggplot(data) +
 aes(x = GDP, y = HIV_female) +
 geom_point(size = 1L, colour = "#0c4c8a",alpha=0.5) +
 scale_x_continuous(trans = "sqrt") +
 theme_minimal()+geom_smooth(method = 'gam',color="red",size=0.5)
```

Explore more of the data

Futhermore, I want to select top ten HIV prevalance in total and in female countries for the most rencently year(2018) of the data set. So I made a table to arrange the data first, then I compared the GDP of those ten countries in 2018 with their HIV to see if their is any correlation.

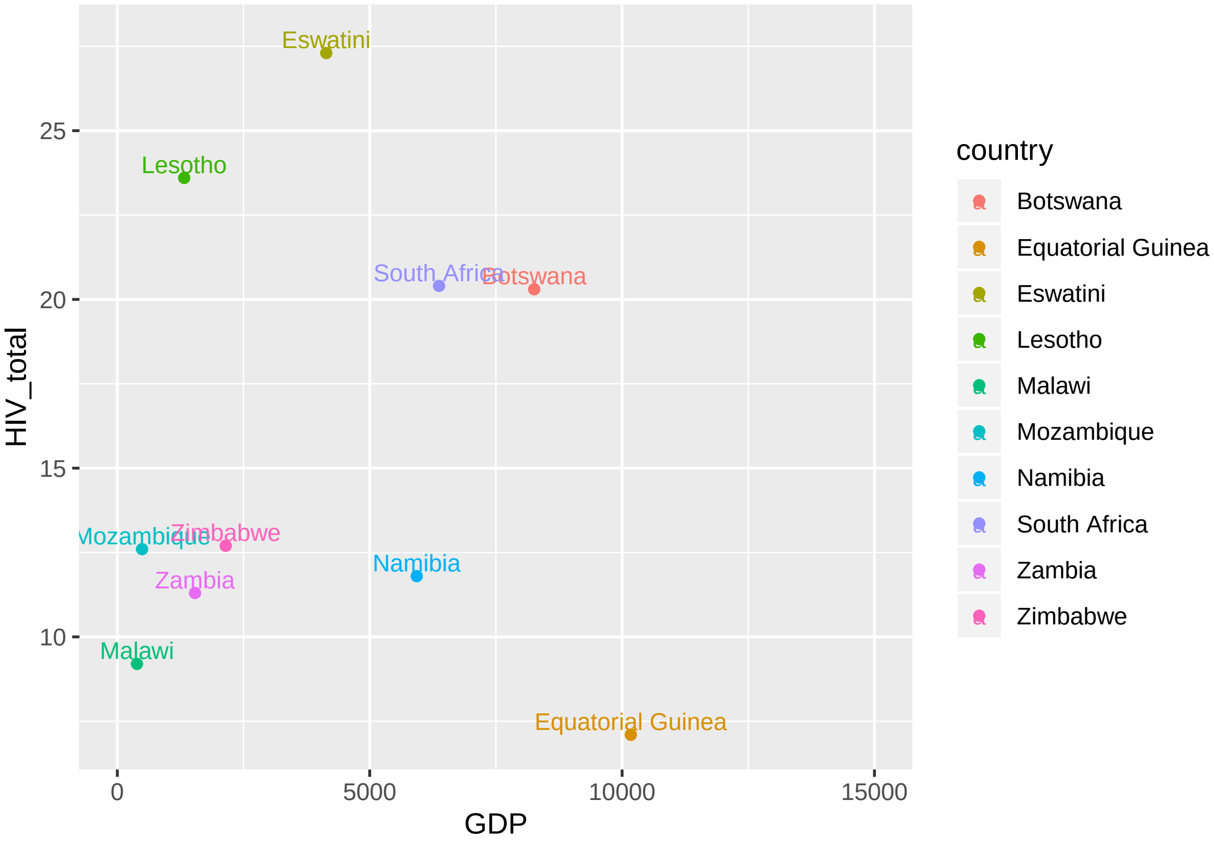
```
#select the top ten HIV Prevalance countries in most recent years
da = data %>% filter(year == 2018) %>% dplyr::select (country,HIV_total,HIV_female)%>%arrange(desc(HIV_
kable(head(da,10),align = "c")%>% kable_styling(latex_options = "HOLD_position")
```

country	HIV_total	HIV_female
Eswatini	27.3	15.9
Lesotho	23.6	10.8
South Africa	20.4	11.3
Botswana	20.3	8.9
Zimbabwe	12.7	5.7
Mozambique	12.6	7.2
Namibia	11.8	4.5
Zambia	11.3	5.0
Malawi	9.2	4.3
Equatorial Guinea	7.1	3.3

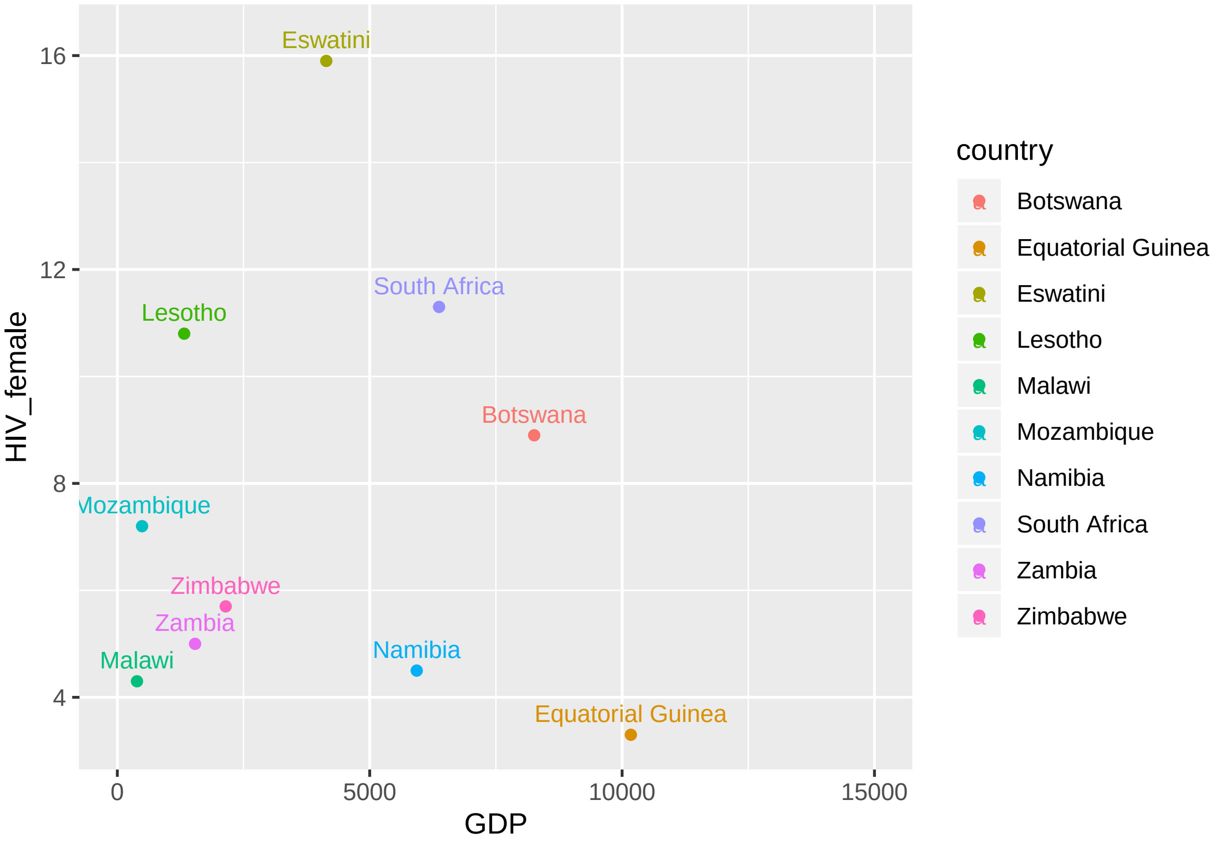
```r
#select the top ten HIV Prevalance of female countries in most recent years
da = data %>% filter(year == 2018) %>% dplyr::select (country,HIV_total,HIV_female)%>%arrange(desc(HIV_
kable(head(da,10),align = "c")%>% kable_styling(latex_options = "HOLD_position")
```

country	HIV_total	HIV_female
Eswatini	27.3	15.9
South Africa	20.4	11.3
Lesotho	23.6	10.8
Botswana	20.3	8.9
Mozambique	12.6	7.2
Zimbabwe	12.7	5.7
Zambia	11.3	5.0
Namibia	11.8	4.5
Malawi	9.2	4.3
Equatorial Guinea	7.1	3.3

```r
sub_HIV_total <- subset(data, country %in% c("Eswatini", "Lesotho", "South Africa",
"Botswana", "Zimbabwe", "Mozambique", "Namibia", "Zambia", "Malawi", "Equatorial Guinea"))
ggplot(subset(sub_HIV_total, year == 2018), aes(x = GDP, y = HIV_total, color = country)) +geom_point()
scale_x_continuous(limits = c(0, 15000))
```



```
sub_HIV_female <- subset(data, country %in% c("Eswatini", "Lesotho", "South Africa",
"Botswana", "Zimbabwe", "Mozambique", "Namibia", "Zambia", "Malawi", "Equatorial Guinea"))
ggplot(subset(sub_HIV_female, year == 2018), aes(x = GDP, y = HIV_female, color = country )) +geom_poin
scale_x_continuous(limits = c(0, 15000))
```



From the plots above I can't tell wether those two variable has linear relationship because some countries who has higher prevalance HIV rate also have higher GDP than some other country (ex. Eswatini), and Equatorial Guinea which has the lowest prevalance HIV rate in these ten countries but also have the highest GDP among them.Therefore this is not consistent with what I have found before. We may consider other factor at this point such as countries enviornment, geogarphical location and their development level of the medical facilities.

Conclusion

In this project, I collected data from public sources (World Bank). And I did an initial exploratory data analysis. Then, I derived a correlation factor and applied linear regression to assess the linear relationship between three interest (GDP per capita, HIV prevalence, HIV prevalence in female). In addition, I also try to see if the current top ten HIV prevalence countries have any representative information to support what I found, but the results turns out not so helpful, which is normal.Therefore, we may need futher analysis on this topic.