HIVDATASET Project HarvardX PH125.9x

Muzukhona Magagula

20 June 20202

Introduction

The HIV Dataset contains no. of people living HIV. Data from WHO and UNESCO Websites. In the time of epidemics, what is the status of HIV AIDS across the world, where does each country stands, is it getting any better. The dataset was more helpful. This HIV dataset have 170 countries with South Africa having maximum of 4,788,000 number of people on ART followed by Mozambique with a total of 2,700,000 and 1,700,000 from Tanzania respectively. While some countries have some countries have no number provide it is also noted 100 was a minimum number of people record provided in the dataset.

Purpose of the project

As a person working with real health data it is important for me to understand how to analyse, visualise and tweak data using machine learning platform. In this project I was able to get top 20 countries have highest number of people living on ART, minimum number of people, and I was able to rank countries based on the number of people it has.

Instructions and procedures used

Steps Importing HIV Dataset

```
library(readxl)
> hivdataset <- read_excel("Capstone/hivdataset.xlsx",
+     col_types = c("text", "numeric", "text",
+         "text", "numeric", "numeric", "numeric",
+         "numeric", "numeric", "numeric",
+         "text"))
View(hivdataset)
```

Describing the hivdataset table

```
 str(hivdataset)
```

Classes ‘tbl_df’, ‘tbl’ and 'data.frame': 170 obs. of 11 variables:

 $ Country   : chr "Afghanistan" "Albania" "Algeria" "Angola" ...

 $ Reported number of people receiving ART   : num 920 580 12800 88700 85500 1900 22800 NA 4400 3100 ...

 $ Estimated number of people living with HIV   : chr "7200[4100â\200“11000]" "NA" "16000[15000â\200“17000]" "330000[290000â\200“390000]" ...

$ Estimated ART coverage among people living with HIV (%) : chr "13[7â\200“20]" "NA" "81[75â\200“86]" "27[23â\200“31]" ...

$ Estimated number of people living with HIV_median : num 7200 NA 16000 330000 140000 3500 28000 NA NA 6000 ...

$ Estimated number of people living with HIV_min : num 4100 NA 15000 290000 130000 3000 23000 NA NA 5300 ...

$ Estimated number of people living with HIV_max : num 11000 NA 17000 390000 150000 4400 31000 NA NA 6700 ...

$ Estimated ART coverage among people living with HIV (%)_median: num 13 NA 81 27 61 53 83 NA NA 52 ...

$ Estimated ART coverage among people living with HIV (%)_min : num 7 NA 75 23 55 44 70 NA NA 45 ...

$ Estimated ART coverage among people living with HIV (%)_max : num 20 NA 86 31 67 65 93 NA NA 58 ...

$ WHO Region : chr "Eastern Mediterranean" "Europe" "Africa" "Africa" ...

Number of rows the table have

```
nrow(hivdataset)
```

```
[1] 170
```

Number of columns the table have

```
ncol(hivdataset)
```

```
[1] 11
```

What is the minimum value of Reported People living with ART?

```
min(hivdataset[,2], na.rm=T)
```

```
[1] 100
```

What is the maximum value of Reported People living with ART?

```
max(hivdataset[,2], na.rm=T)
```

```
[1] 4788000
```

Change column name Reported number of people receiving ART

```
colnames(hivdataset)[2] <- "Reported"
```

Select all Maximum values within columns

```
Apply (hivdataset, MARGIN = 2, function(x) max(x, na.rm=TRUE))
```

find the 20 countries with the high number of people Reported taking ART ,plus the Zimbabwe and South Africa with their ranks

```
high_hiv <- hivdataset %>%
  arrange(desc(Reported)) %>%
  mutate(rank = c(1:170)) %>%
 filter(rank <= 20 | grepl("Zimbabwe|South Africa", Country))
 high_hiv
```

A tibble: 20 x 12

	Country	Reported	`Estimated numb…`	`Estimated ART …`	`Estimated numb…`	`Estimated numb…`	`Estimated numb…`
	<chr>	<dbl>	<chr>	<chr>	<dbl>	<dbl>	<dbl>
1	South …	4788000	7700000[7100000…	62[57â€“66]	7700000	7100000	8300000
2	Mozamb…	1213000	2200000[1700000…	56[44â€“68]	2200000	1700000	2700000
3	Zimbab…	1151000	1300000[1100000…	88[77â€“95]	1300000	1100000	1500000
4	United…	1109000	1600000[1400000…	71[64â€“78]	1600000	1400000	1700000
5	Kenya	1068000	1600000[1300000…	68[58â€“82]	1600000	1300000	1900000
6	Nigeria	1016000	1900000[1400000…	53[40â€“71]	1900000	1400000	2600000
7	Uganda	1004000	1400000[1300000…	72[68â€“78]	1400000	1300000	1500000
8	Zambia	965000	1200000[1100000…	78[69â€“88]	1200000	1100000	1400000
9	Malawi	814000	1000000[940000â…	78[70â€“84]	1000000	940000	1100000
10	China	718000	NA	NA	NA	NA	NA
11	Brazil	593000	900000[690000â€…	66[51â€“82]	900000	690000	1100000
12	Ethiop…	450000	690000[530000â€…	65[50â€“85]	690000	530000	900000

13 Thaila… 359000 480000[420000â€… 75[66â€“86] 480000 420000 550000

14 Botswa… 307000 370000[330000â€… 83[75â€“90] 370000 330000 400000

15 Camero… 281000 540000[470000â€… 52[46â€“57] 540000 470000 590000

16 Democr… 256000 450000[370000â€… 57[47â€“67] 450000 370000 530000

17 CÃ´te … 252000 460000[360000â€… 55[44â€“70] 460000 360000 580000

18 Lesotho 206000 340000[320000â€… 61[57â€“65] 340000 320000 360000

19 Rwanda 194000 220000[200000â€… 87[76â€“95] 220000 200000 250000

20 Namibia 184000 200000[190000â€… 92[84â€“95] 200000 190000 220000

… with 5 more variables: `Estimated ART coverage among people living with HIV (%)_median` <dbl>, `Estimated ART

coverage among people living with HIV (%)_min` <dbl>, `Estimated ART coverage among people living with HIV

(%)_max` <dbl>, `WHO Region` <chr>, rank <int>

Grouping Data by WHO Region

```
hivdataset <- hivdataset %>%
+     select(Reported, `WHO Region`) %>%
+     group_by('WHO Region')
```

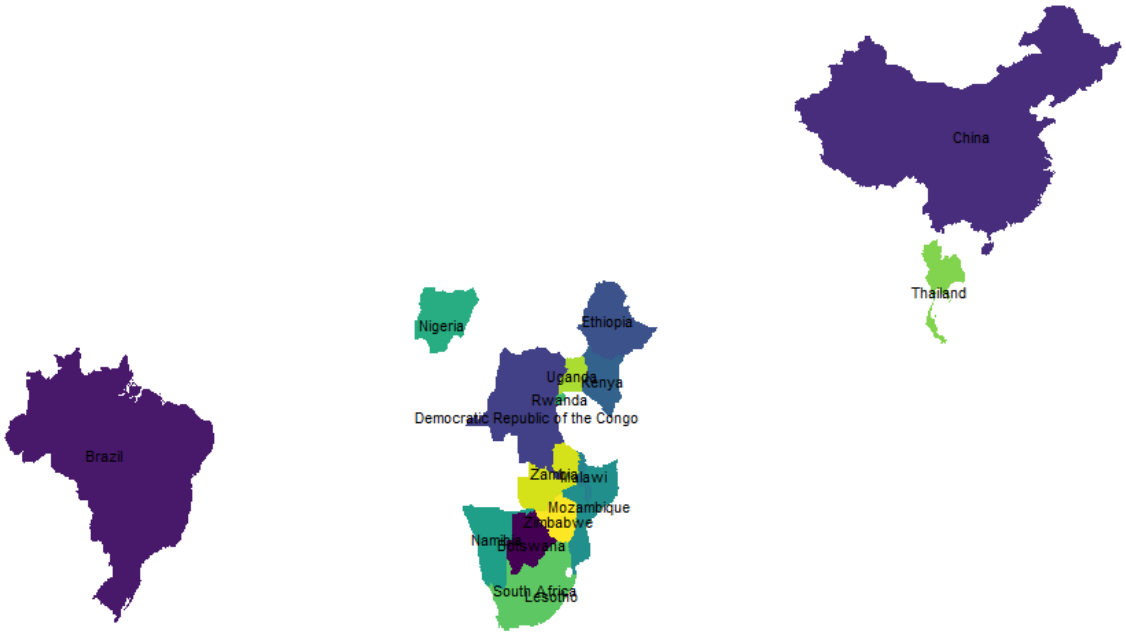
Regionalised Top 20 countries with most people on ART

 # TO 2O Countries with high number of people on ART

```
> some.eu.countries <- c(
+     "South Africa","Mozambique","Zimbabwe",
+       "United Republic of Tanzania",
"Kenya","Nigeria","Uganda","Zambia","Malawi","China","Brazil",
```

```
+      "Ethiopia","Thailand","Botswana","Camerooon","Democratic Republic of the Congo","CÃ´te
d'Ivoire","Lesotho","Rwanda","Namibia")
>
> # Retrievethe map data
> some.eu.maps <- map_data("world", region = some.eu.countries)
>
> # Compute the centroid as the mean longitude and lattitude
> # Used as label coordinate for country's names
> region.lab.data <- some.eu.maps %>%
+     group_by(region) %>%
+     summarise(long = mean(long), lat = mean(lat))
>
> ggplot(some.eu.maps, aes(x = long, y = lat)) +
+   geom_polygon(aes( group = group, fill = region))+
+   geom_text(aes(label = region), data = region.lab.data,  size = 3, hjust = 0.5)+
+   scale_fill_viridis_d()+
+   theme_void()+
+   theme(legend.position = "none")


ggplot(hivdataset, aes(x=Reported))+
+     geom_histogram(color="black", fill="lightblue",
+                    linetype="dashed")
```

Map Visualisation



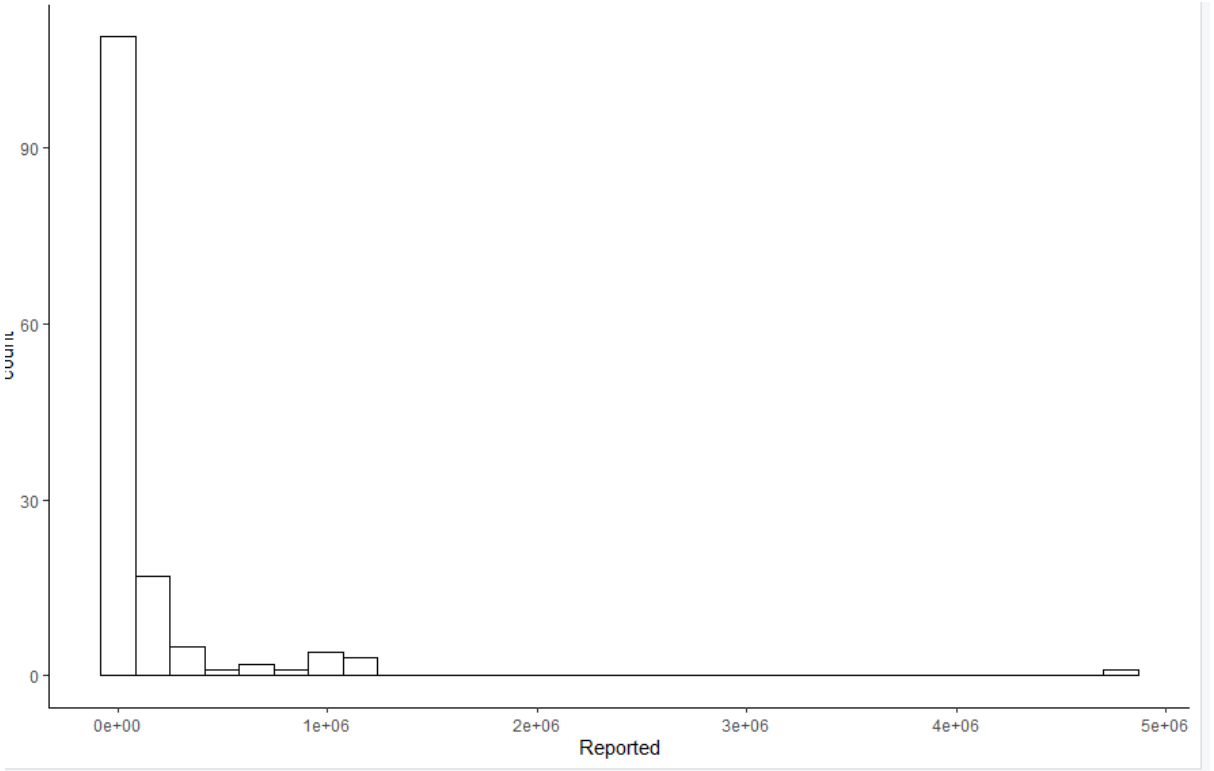
GGPLOT MAPPING

#GGPLOT MAPPING

```
>
> # Add mean lines
> p<-ggplot(hivdataset, aes(x=weight, color=sex)) +
+   geom_histogram(fill="white", position="dodge")+
+   geom_vline(data=mu, aes(xintercept=grp.mean, color=sex),
+              linetype="dashed")+
+   theme(legend.position="top")

> p+scale_color_manual(values=c("#999999", "#E69F00", "#56B4E9"))+
+     scale_fill_manual(values=c("#999999", "#E69F00", "#56B4E9"))
```

```
> # Use grey scale
> p + scale_color_grey()+scale_fill_grey() +
+     theme_classic()


> p + theme(legend.position="top")
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

> p + theme(legend.position="bottom")

> # Remove legend
> p + theme(legend.position="none")

> p+scale_color_manual(values=c("#999999", "#E69F00", "#56B4E9"))+
+     scale_fill_manual(values=c("#999999", "#E69F00", "#56B4E9"))


> # use brewer color palettes
> p+scale_color_brewer(palette="Dark2")+
+     scale_fill_brewer(palette="Dark2")

> # Use grey scale
> p + scale_color_grey()+scale_fill_grey() +
+     theme_classic()
```

GGPLOT VISUALISATION



Conclusion

With the aid of this course I feel fit to work any real-world data, like I said earlier on, my daily work is to work with large health related datasets.