Comparing Kenyan Health facility location data from different sources

Andy South @afrimapr

2021-05-07

DRAFT Code here:https://github.com/afrimapr/afrimapr_dev/blob/master/compare_moh_kenya.Rmd

We currently have access to the following sources of health facility locations for Kenya.

1. downloaded from Ministry of Health website (no coordinates) http://kmhfl.health.go.ke/#/facility_filter/results
2. a file created by WiGISKe 2021-04 from multiple sources
3. healthsites.io in OpenStreetMap
4. collated data from published paper Maina(2019)

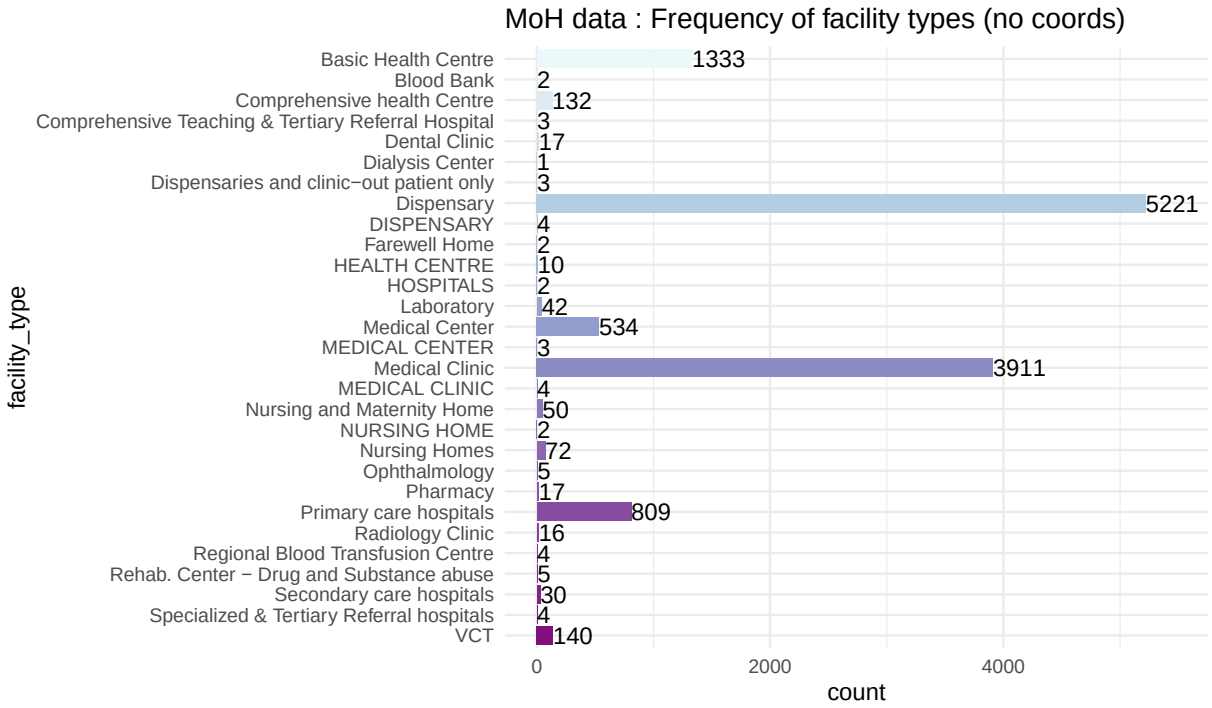
Following discussion 2021-04-29 suggestion that modelling work at the ward level uses the MoH data that has advantage of having a clear provenance (and also has the most facilities and attributes).

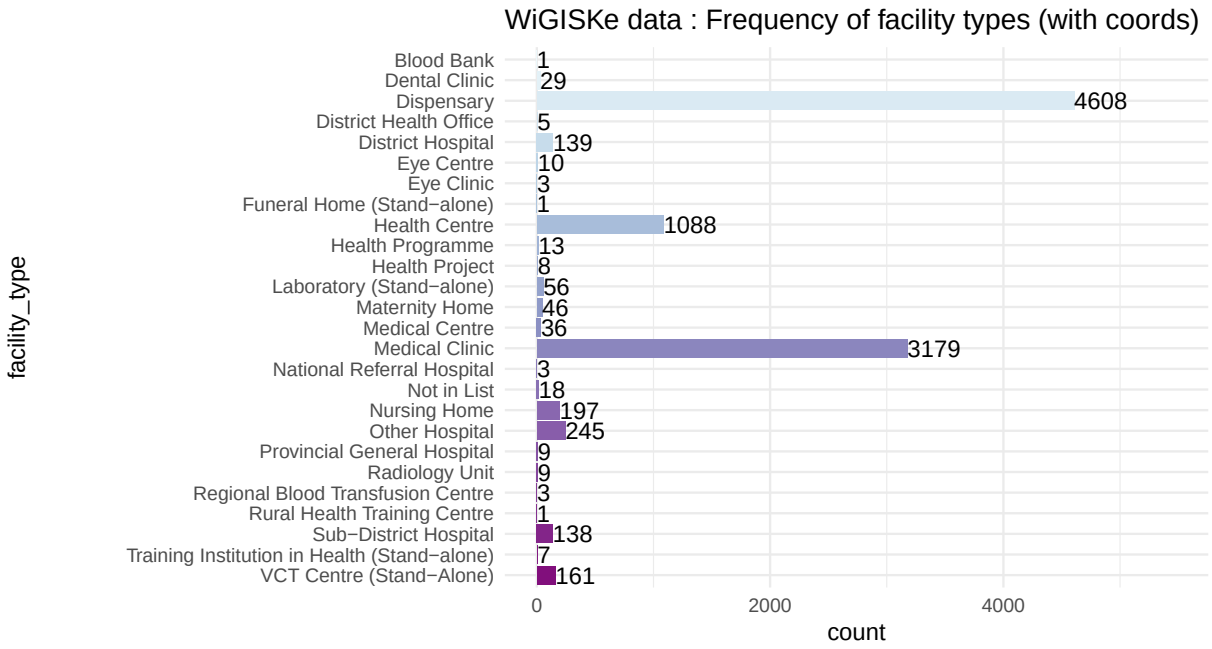
Also note that the MoH data has a column named 'Code' which matches with the 'MFL CODE' column in the pdf of MoH approved Vaccination posts so will allow us to link the two.

Numbers of locations per source.

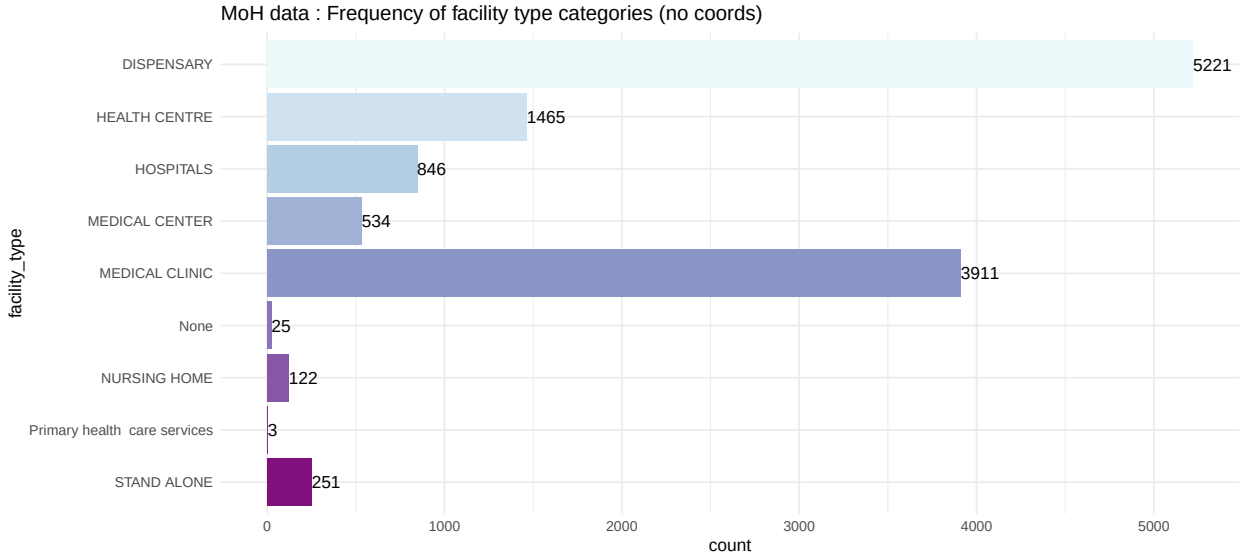
source	number facilities	spatial data	useful attributes
1. MoH	12378	Ward	Registration_number,Keph level,Owner type,Beds,Cots,Operation status,Open_weekends
2. WiGISKe	10013	coords	Owner
3. OSM	2032	coords	(few complete) beds,staff_doctors,staff nurses

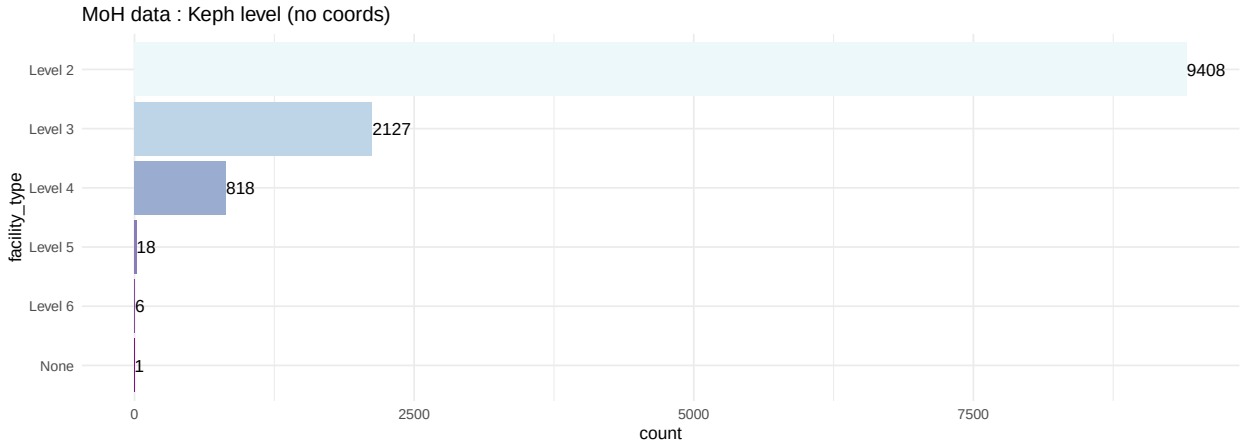
Comparing frequency of facility types in MoH & WiGISKe data

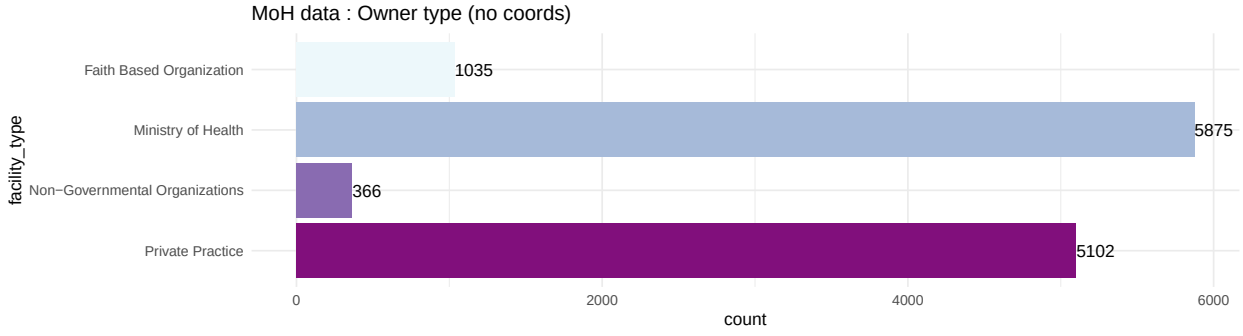


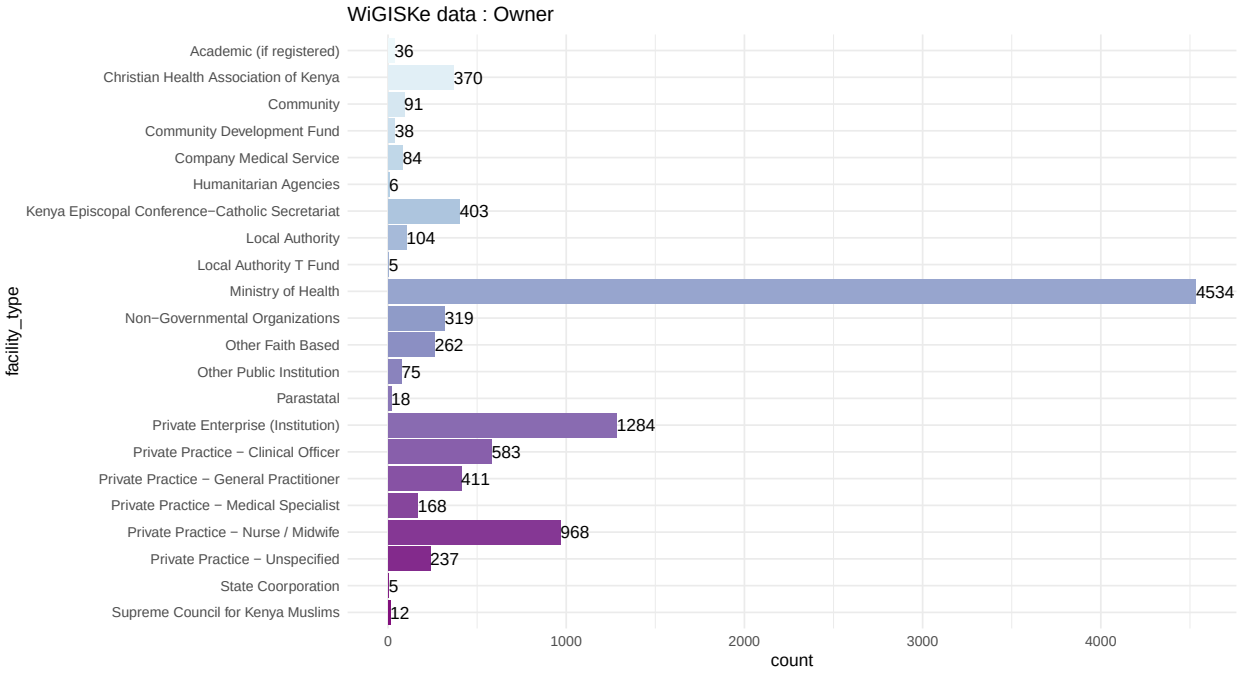


MoH facility attributes







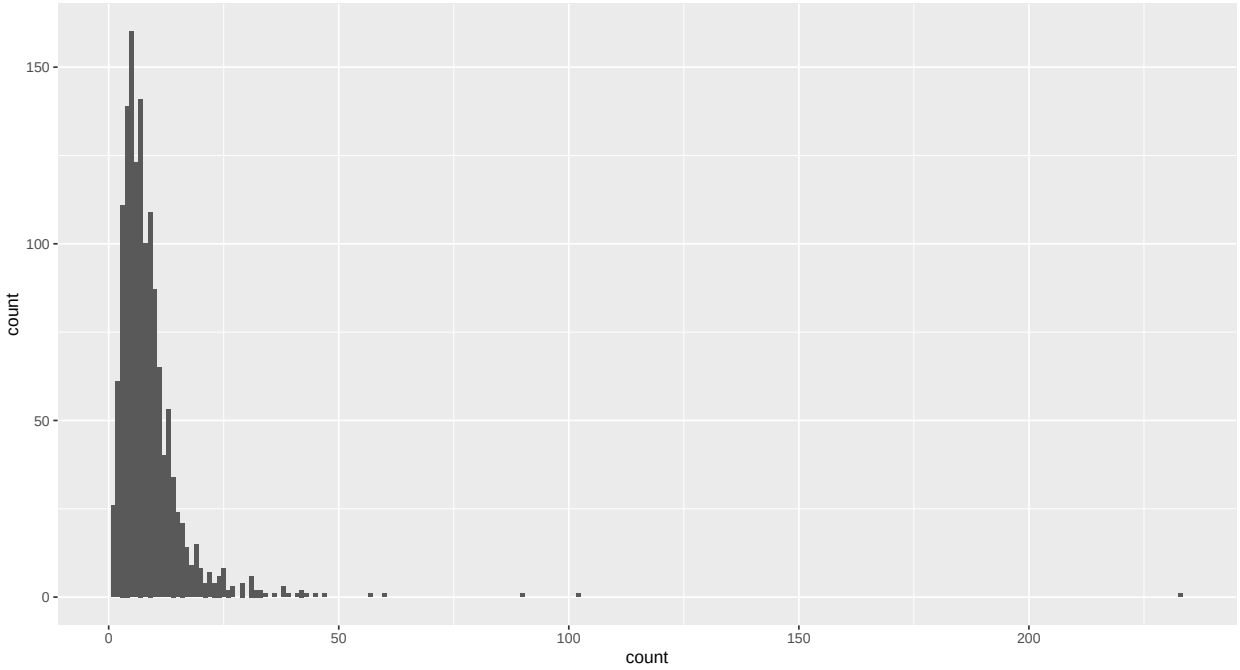


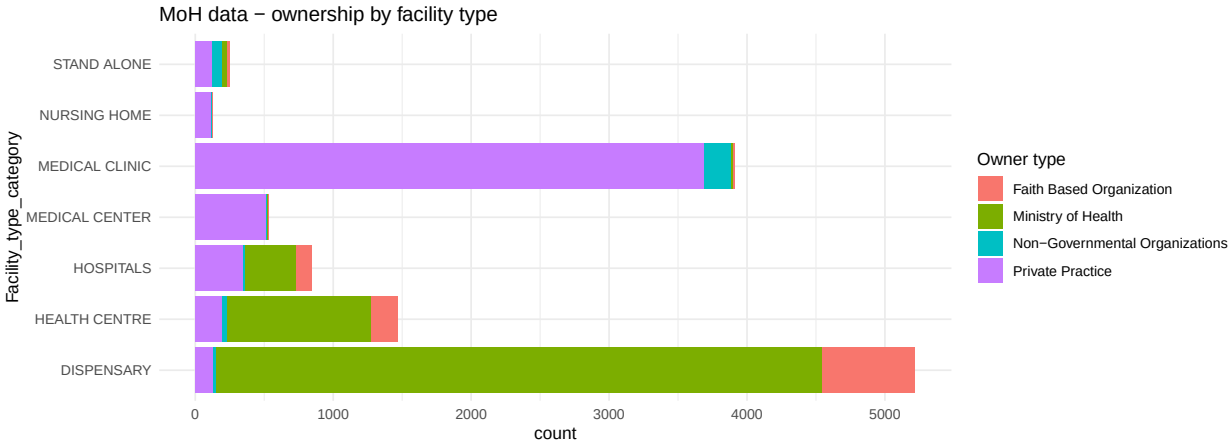
All attribute columns in WiGISKe & MoH data

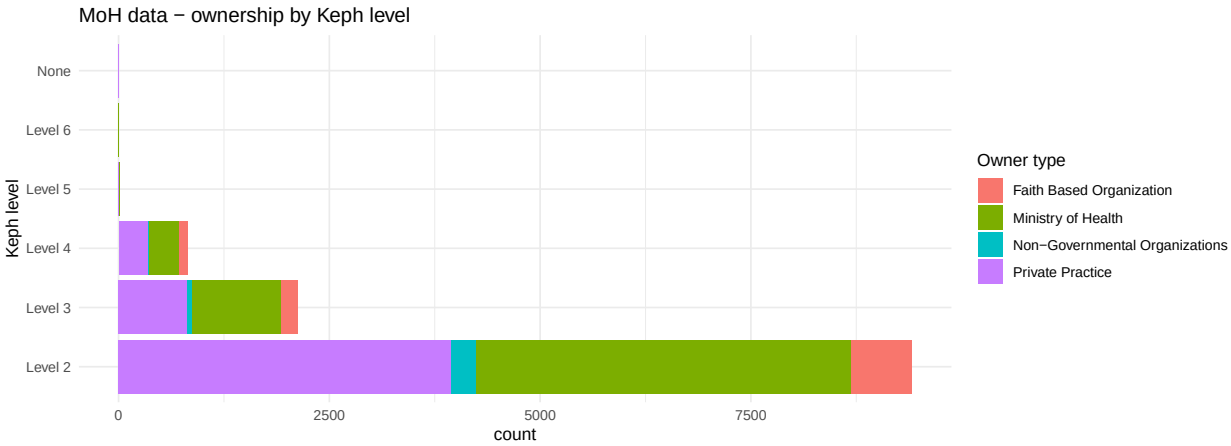
```
##  [1] "OBJECTID"   "Facility_N" "Type"       "Owner"      "County"
##  [6] "Sub_County" "Division"   "Location"   "Sub_Locati" "Constituen"
## [11] "Nearest_To" "Latitude"   "Longitude"  "geometry"
```

```
##  [1] "Code"                   "Name"                   "Officialname"
##  [4] "Registration_number"    "Keph level"             "Facility type"
##  [7] "Facility_type_category" "Owner"                  "Owner type"
## [10] "Regulatory body"        "Beds"                   "Cots"
## [13] "County"                 "Constituency"           "Sub county"
## [16] "Ward"                   "Operation status"       "Open_whole_day"
## [19] "Open_public_holidays"   "Open_weekends"          "Open_late_night"
## [22] "Service_names"          "Approved"               "Public visible"
## [25] "Closed"
```

Number of facilities per ward in MoH data







Draft code to :

1. scrape vaccination centre locations from PDF
2. join to MoH Master facility list to get ward names
3. join to wards map

```
# looking at getting vaccination sites from pdf
# https://www.brodrigues.co/blog/2018-06-10-scraping_pdfs/

library(pdftools)
library(stringr) #string manipulation
library(mapview) #interactive mapping
library(tmap) #ststic maps

download.file(url = "https://www.health.go.ke/wp-content/uploads/2021/03/MOH-APPROVED-COVID-19-VACCINATI
                    destfile = "data-raw/ken-vacc-posts.pdf", mode = "wb")

filename = "data-raw/ken-vacc-posts.pdf"

#puts the text into 9 char strings, 1 per page
txt <- pdf_text(filename)

#this splits page 1 into lines, puts each into a separate column
txt1 <- stringr::str_split(txt[1], "\n", simplify = TRUE)

#str(txt1)
#chr [1, 1:78]

#on page1 line5 contains the headers
#txt1[1,5]
#[1] "                         Subcounty                    Health Facility Name             MFL CODE
#on page1 lines6 & after contain the data
#luckily columns are separated by multiple spaces
#txt1[1,6]
#[1] "1               Mvita                  Coast General Teaching and Referal Hospital    11289    Pu
#beware some lines contain extra word that is the county
#should be relatively easy to remove later
#txt1[1,10]
#[1] "5   Mombasa Mvita                      Aga Khan Hospital Mombasa                      11203    Pr

#removing all countynames followed by 2 spaces e.g. 'Kwale '
#to avoid problem caused by column2 County names which contains entries for only a few rows
countyregex <- c("Kwale  |Kilifi  |Tana River  |Lamu  |Taita Taveta  |Garissa  |Wajir  |Mandera  " )
txt1 <- str_remove_all(txt1, countyregex)
#txt1 <- str_remove_all(txt1, c("Kwale  "))
#replace more than 2 spaces with '|' then use that as a delimiter to read into a dataframe
txt1columns <- str_replace_all(txt1, "\\s{2,}", "|")


#remove first 4 rows from page1
txt1columns2 <- txt1columns[-c(1:4)]
#add extra item to header row
txt1columns2[1] <- paste("vacc_center_num",txt1columns2[1])
#make text connection to the object
```

```
text_con <- textConnection(txt1columns2)
#read into a dataframe
df1 <- read.csv(text_con, sep = "|") #, skip=4)
#df1 <- readr::read_delim(text_con, delim = "|", skip=4)

#mostly works - this issue now mostly fixed above
#issue with this line where the county (Kwale) gets read in as an extra column
#problem because the county column has cells merged across multiple rows
#I could fix by removing counties at the txt or txt1 stage
#would cause county names to be removed from facility names e.g. Mombasa hospital
#but the MFL code would still allow joining
#could fix by removing e.g. '|Kwale', but that could cause problems by removing later columns
#removing all countynames followed by 2 spaces should sort it e.g. 'Kwale '
#another way could be to count number of | and if > 4 remove the first
#str_count(txt1columns2,"\\|")
#4 4 4 4 4 4 4 4 4 4 4 4 4 5 4 4 4 4 4 4 4 4 4 4 5 4 4 4 4 4 4 4 4 4 4 5 4 4 4 4 1 4 4 4 4 4 4 4 4 4 1 4 4
#txt1columns2[18]
#[1] "13|Kwale|Lungalunga|Lungalunga Sub County Hospital|11526|Public\r"


# checking whether joining this onto the MoH data works
# try left join df1,MoH so it is easier to see
# this does work for ~70 facilities, a few problems
dfjoined <- dplyr::left_join(x=df1, y=dfmoh,
                             by=c("MFL.CODE"="Code")
                            )

#try to join to a ward map to test

#read wards map (downloaded from HDX), could get via rhdx
sfwards <- sf::st_read("data-raw/kenya-wards-hdx/kenya_wards.shp")
```

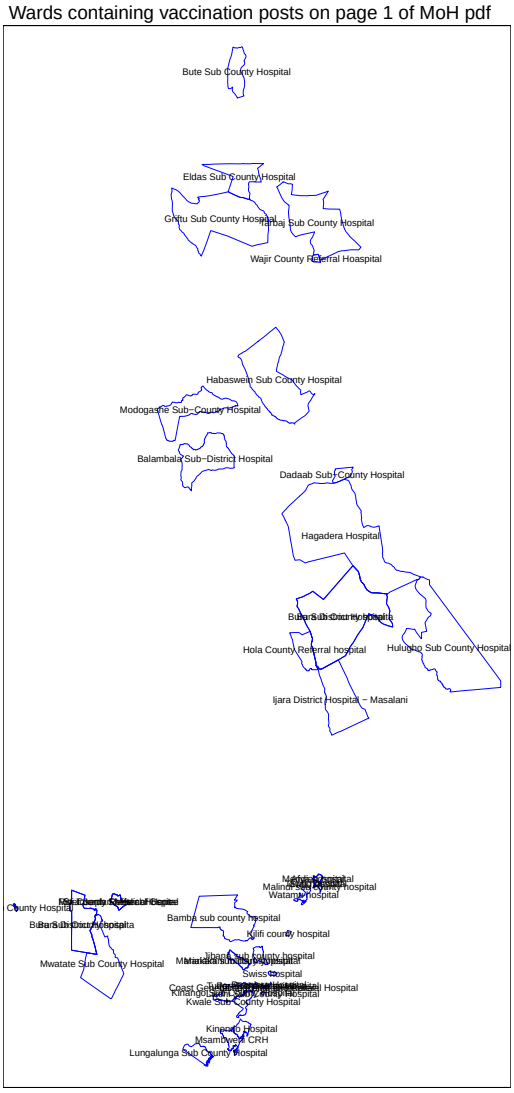
```
## Reading layer `kenya_wards' from data source `C:\rsprojects\afrimapr_dev\data-raw\kenya-wards-hdx\ken
## Simple feature collection with 1450 features and 8 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 33.9105 ymin: -4.67973 xmax: 41.91056 ymax: 5.466979
## Geodetic CRS:  WGS 84
```

```
#remove " Ward" from the ward names in sfwards$ward to allow joining
sfwards$ward2 <- stringr::str_remove(sfwards$ward," Ward")

#join vaccination posts to wards map
#Some fail to join probably due to differences in ward naming
sfwards_vacc <- dplyr::left_join(x=dfjoined, y=sfwards,
                             by=c("Ward"="ward2")
                            )
#re sf the object, which seems to forget it is sf in the join
sfwards_vacc <- sf::st_as_sf(sfwards_vacc)

#view the wards that contain a facility on page1
#mapview::mapview(sfwards_vacc, label="Health.Facility.Name")
```

```
#static map
tmap::tm_shape(sfwards_vacc) +
    tm_borders("blue", lwd = .5) +
    tm_layout(main.title="Wards containing vaccination posts on page 1 of MoH pdf",
              main.title.size=1) +
    tm_text("Health.Facility.Name", size=0.5)
```



#TODO #work out how to deal with few remaining rows that fail on page1 #check on wards that fail to join #try applying to later pages