Basic analysis on the behavior of SARS-COV-2 in Costa Rica - Data Science: Capstone - HarvardX PH125.9x

Agustín Villalobos

18/2/2022

```
#Install necesary libraries
#install.packages("readr")
#install.packages("dplyr")
#install.packages("purrr")
#install.packages("magrittr")
#install.packages("ggplot2")
#install.packages("hrbrthemes")
#install.packages("viridis")
#install.packages("viridisLite")
#install.packages("deSolve")
```

```
#OBTAIN THE OFFICIAL SARS-COV-2 DATA FOR COSTA RICA (2022-02-11)
#Load necessary libraries
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(purrr)
options(readr.show_col_types = FALSE)

#Download the csv file and create a local copy
url <- "https://geovision.uned.ac.cr/oges/archivos_covid/2022_02_11/02_11_22_CSV_GENERA
L.csv"
dat <- read_csv(url)
```

```
## Warning: One or more parsing issues, see `problems()` for details
```

```
wd <- getwd()
wd <-  file.path(wd, "CovidCR.csv")
download.file(url, wd)
```

```
#Obtain the file ubication and fill the dataset
wd <- getwd()
wd <-  file.path(wd, "CovidCR.csv")

#Read the csv file from an online source (A local copy is created since inconsistencies
 are found in the online source)
df_covidcr <- read.table(wd, header = TRUE, sep = ";")

df_covidcr
```

FECHA <chr>	… <int>	positivos <int>	nue_posi <int>	conf_lab <int>	conf_nexo <int>	muj_posi <int>	hom_p… <int>	extranj_posi <int>	cos
6/3/2020	10	2	2	*NA*	0	*NA*	*NA*	*NA*	
7/3/2020	10	7	5	*NA*	0	*NA*	*NA*	*NA*	
8/3/2020	11	10	3	*NA*	0	*NA*	*NA*	*NA*	
9/3/2020	11	12	2	*NA*	0	*NA*	*NA*	*NA*	
10/3/2020	11	13	1	*NA*	0	7	6	3	
11/3/2020	11	22	9	*NA*	0	14	8	3	
12/3/2020	11	23	1	*NA*	0	14	9	3	
13/3/2020	11	26	3	*NA*	0	14	12	3	
14/3/2020	11	27	1	*NA*	0	15	12	4	
15/3/2020	12	35	8	*NA*	0	19	16	5	

1-10 of 708 rows | 1-10 of 50 columns Previous **1** 2 3 4 5 6 … 71 Next

```
#EXTRACT AND CLEAN THE DATA
#The data to be used in the analysis are selected and the blank rows that do not corresp
ond to a day of the pandemic are eliminated.
#There is an error in the DIA_COVID19 field, where the value 2267 is wrongly entered, th
e correct one being 457. I have corrected this problem from my base file CovidCR.csv.

#Load necessary libraries
library(dplyr)
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
df_covidcr <- df_covidcr %>%
  select("FECHA","DIA_COVID19","SE","positivos","nue_posi", "conf_lab","conf_nexo",
         "muj_posi","hom_posi","extranj_posi","costar_posi","investig_posi","adul_posi",
"am_posi","menor_posi","eda_ignor_posi","descartados","nue_descar",
         "fallecidos","nue_falleci","muj_fall","hom_fall","adul_fall","am_fall","menor_f
all","eda_igno_falle",
         "RECUPERADOS","NUE_RECUP", "MUJ_RECUP","HOM_RECUP","ADUL_RECUP","AM_RECUP","MEN
OR_RECUP","EDA_IGNO_RECUP",
         "hospital","nue_hospi","salon","nue_salon","UCI","nue_UCI","activos","nue_acti"
,   "muj_acti", "hom_acti", "adul_acti",    "am_acti",  "menor_acti",   "eda_igno_acti")
%>%
  filter(!(is.na(df_covidcr$DIA_COVID19) | df_covidcr$DIA_COVID19=="")) %>%
  mutate(Pandemic_date = as.Date(as.character(FECHA), format="%d/%m/%Y"),month_year=form
at(as.Date(as.character(FECHA), format="%d/%m/%Y"), "%Y-%m"))

df_covidcr
```

FECHA	DIA_COVI…	…	positivos	nue_posi	conf_lab	conf_nexo	muj_posi	hom_p…	extr
<chr>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	
6/3/2020	0	10	2	2	*NA*	0	*NA*	*NA*	
7/3/2020	1	10	7	5	*NA*	0	*NA*	*NA*	
8/3/2020	2	11	10	3	*NA*	0	*NA*	*NA*	
9/3/2020	3	11	12	2	*NA*	0	*NA*	*NA*	
10/3/2020	4	11	13	1	*NA*	0	7	6	
11/3/2020	5	11	22	9	*NA*	0	14	8	
12/3/2020	6	11	23	1	*NA*	0	14	9	
13/3/2020	7	11	26	3	*NA*	0	14	12	
14/3/2020	8	11	27	1	*NA*	0	15	12	
15/3/2020	9	12	35	8	*NA*	0	19	16	

1-10 of 708 rows | 1-10 of 50 columns Previous **1** 2 3 4 5 6 … 71 Next

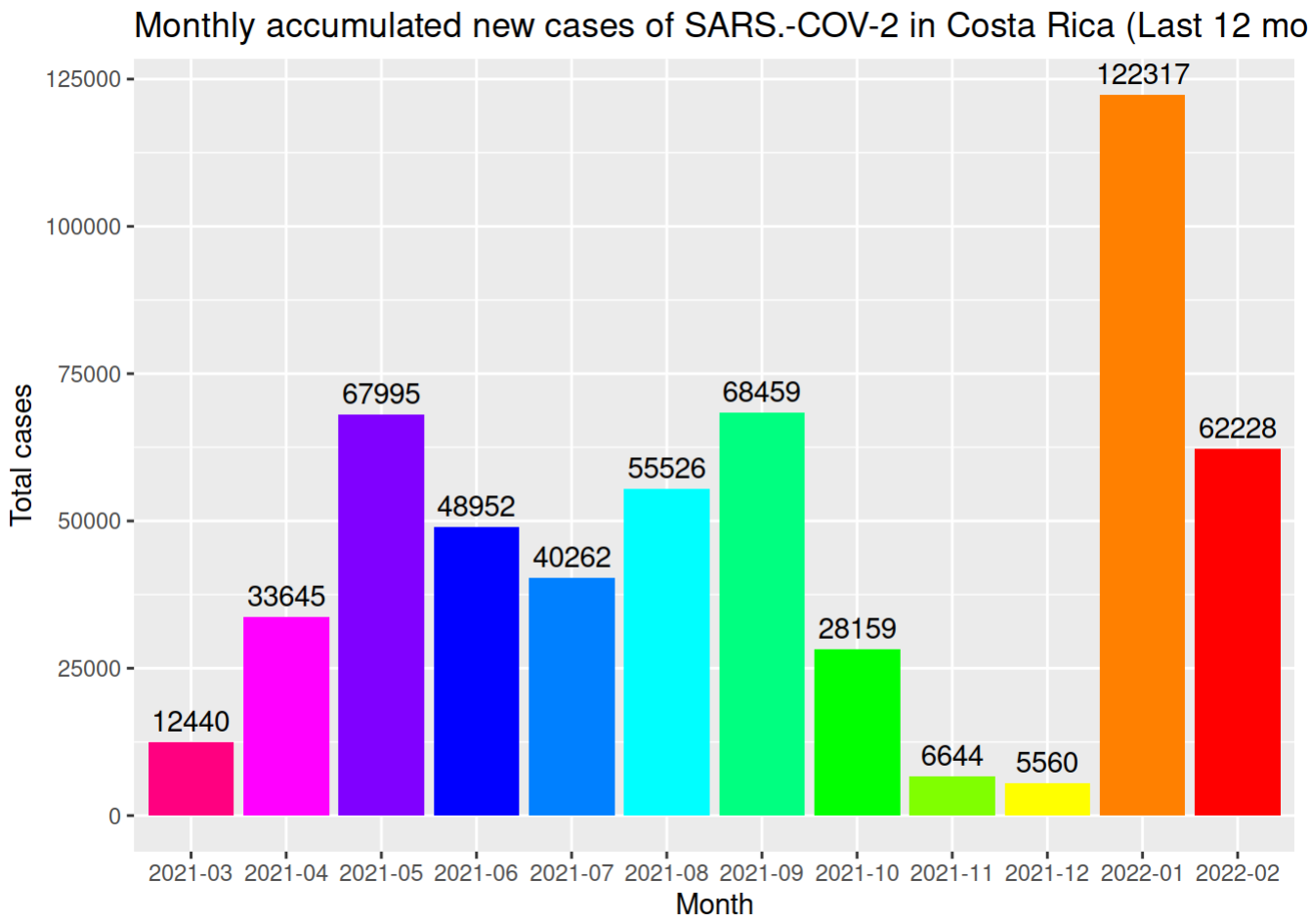
```
#Get day with latest data
i_max <- which.max(df_covidcr$DIA_COVID)

#Official data is obtained for the behavior of the pandemic on the last reported day
vr_Offial_Data <- c(last_data= df_covidcr$FECHA[i_max],
                    total_cases = as.numeric(df_covidcr$positivos[i_max]) ,
                    total_deceased = as.numeric(df_covidcr$fallecidos[i_max]) ,
                    total_recovered=as.numeric(df_covidcr$RECUPERADOS[i_max]))

#Show the obtained data
noquote (vr_Offial_Data)
```

```
##       last_data     total_cases  total_deceased total_recovered
##       11/2/2022          757093            7772          591090
```

```
#Summarise the total for nexus
total_nexus <- sum(df_covidcr$conf_nexo,na.rm=TRUE)

#Obtain the data for the last day
vr_Offial_Data <- c(last_data= df_covidcr$FECHA[i_max],
                    total_cases = as.numeric(df_covidcr$positivos[i_max]) ,
                    total_fornexus = total_nexus,
                    new_cases=as.numeric(df_covidcr$nue_posi[i_max]),
                    new_fornexus=as.numeric(df_covidcr$conf_nexo[i_max]),
                    total_nationals=as.numeric(df_covidcr$costar_posi[i_max]),
                    total_foreigns=as.numeric(df_covidcr$costar_posi[i_max])
                    )

#Show the obtained data
noquote (vr_Offial_Data)
```

```
##       last_data     total_cases  total_fornexus       new_cases    new_fornexus
##       11/2/2022          757093          117192            5488             644
## total_nationals  total_foreigns
##          668473          668473
```

```
#Monthly accumulated new cases of SARS.-COV-2 in Costa Rica (Last 12 months)

#Load necessary libraries
library(ggplot2)

#Get the total number of positive cases for each month of the pandemic perio
month_total <- df_covidcr %>%
  group_by(month_year) %>%
  summarize(total = sum(nue_posi))

#Obtain the top 12 months with the highest number of new infections
month_total  <- month_total %>% select("month_year","total")  %>%
  arrange(desc(month_year)) %>%  head(n = 12L)

#Obtain the total and date for the last of each month in the period of the pandemic
df_covidcrmonth <- data.frame(
  name=month_total$month_year ,
  value=month_total$total
)

#Generates a barplot with the total cases for the last 12 months of the pandemic period
ggplot(df_covidcrmonth, aes(x=name, y=value),) +
  geom_bar(stat = "identity", fill=rainbow(12)) +
  geom_text(aes(label=value),vjust=-0.5) +
  labs(title = "Monthly accumulated new cases of SARS.-COV-2 in Costa Rica (Last 12 mont
hs)",x = "Month",y="Total cases")
```



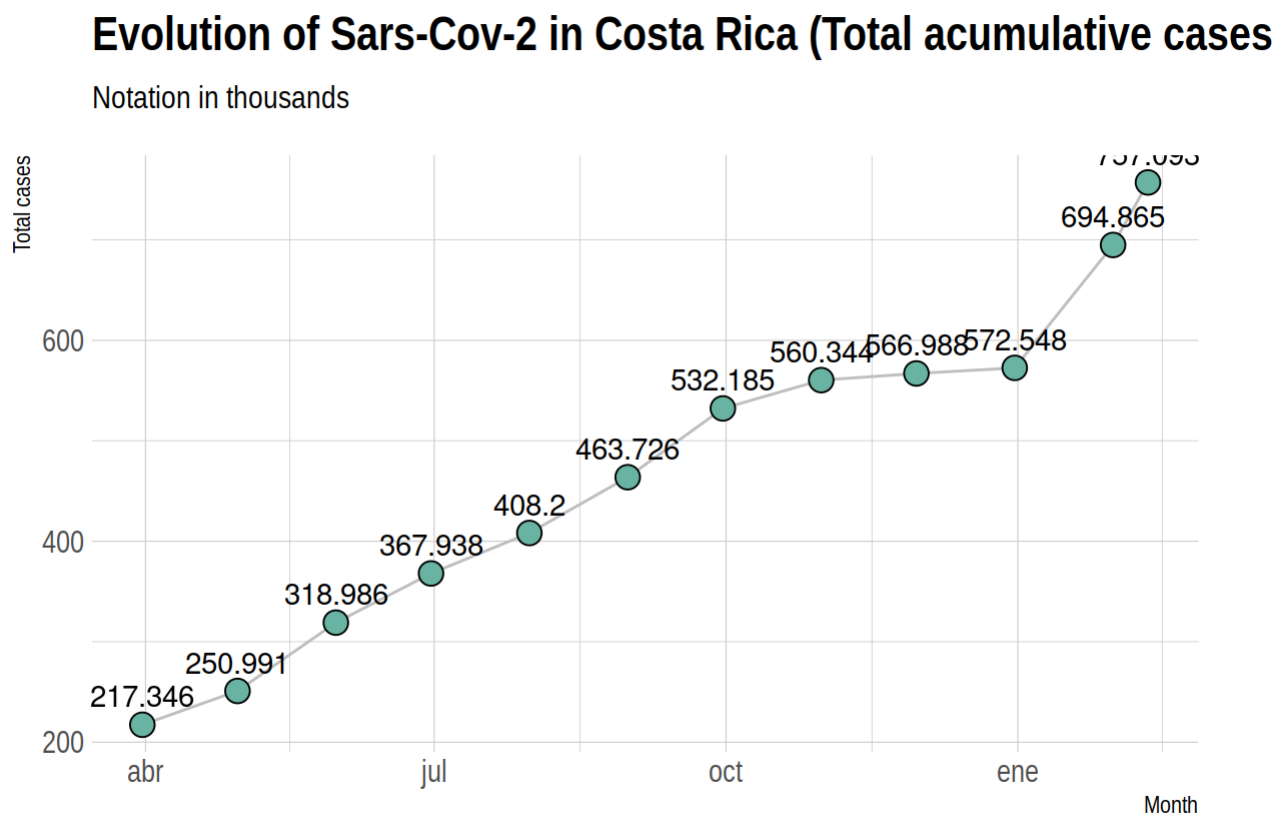

```
#Evolution of Sars-Cov-2 in Costa Rica

#Load necessary libraries
library(ggplot2)
library(dplyr)
library(hrbrthemes)
```

```
## NOTE: Either Arial Narrow or Roboto Condensed fonts are required to use these themes.
```

```
##       Please use hrbrthemes::import_roboto_condensed() to install Roboto Condensed an
d
```

```
##       if Arial Narrow is not on your system, please see https://bit.ly/arialnarrow
```

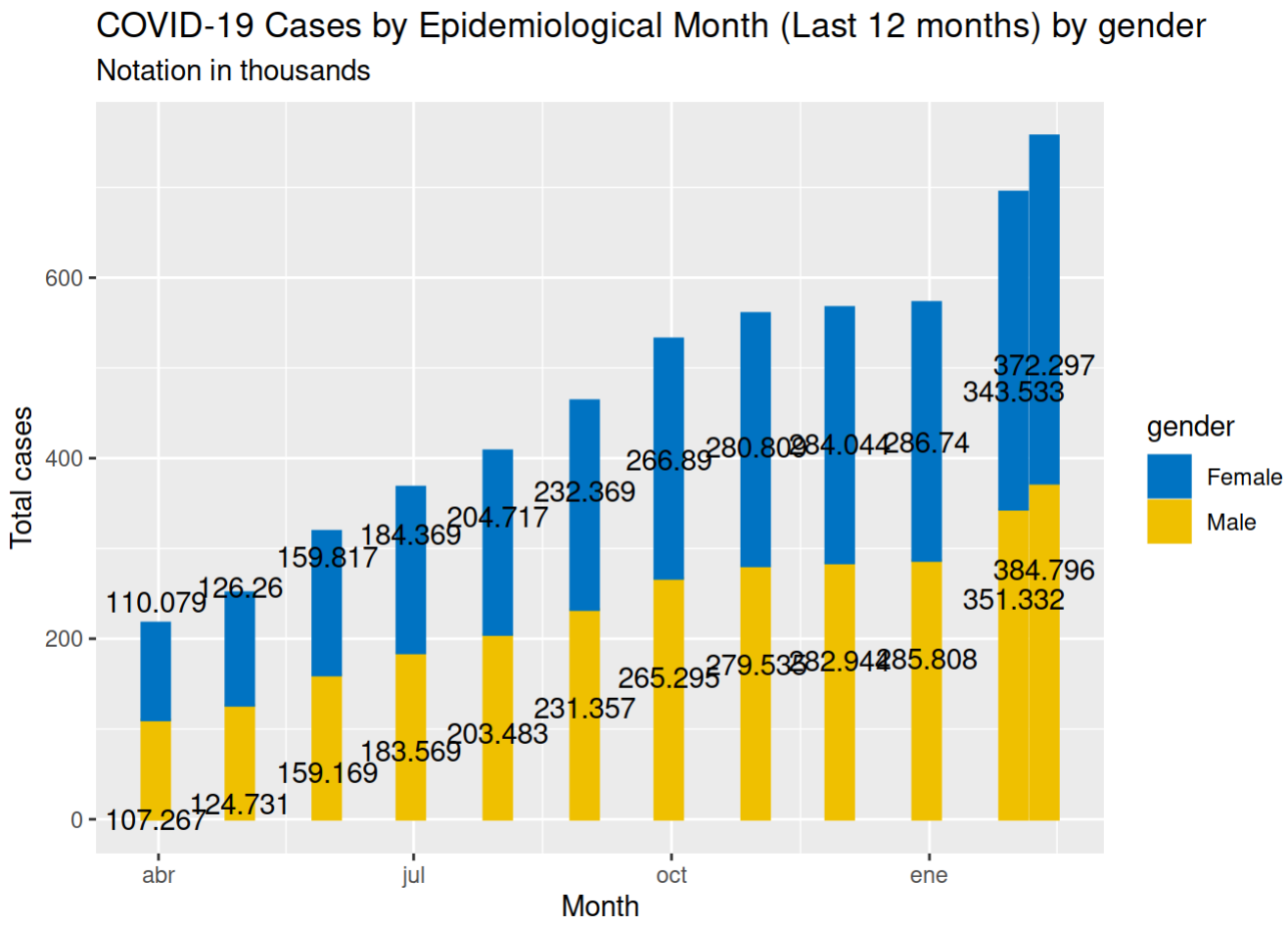
```
#Get the last day of the month with accumulated data
df_maxdates <- df_covidcr %>%
   group_by(month_year) %>%
  summarize(Pandemic_date = max(Pandemic_date))

#Extract only the final accumulated data of each month
df_maxdates <- merge(x=df_maxdates,y=df_covidcr,by="Pandemic_date",all.x=FALSE, all.y=FA
LSE)

#Generate a line plot with evolution of Sars-Cov-2 in Costa Rica
df_maxdates %>%
  tail(12) %>%
  ggplot( aes(x=Pandemic_date, y=positivos/1000)) +
  geom_line( color="grey") +
  geom_point(shape=21, color="black", fill="#69b3a2", size=4) +
  theme_ipsum() +
  geom_text(aes(label=positivos/1000),vjust=-0.8) +
  ggtitle("Evolution of Sars-Cov-2 in Costa Rica (Total acumulative cases per month for
 the last 12 months)")+
  labs(x = "Month",y="Total cases",subtitle ="Notation in thousands")
```



```
#COVID-19 Cases by Epidemiological Month (Last 12 months) by gender

#Obtains the monthly cumulative total of cases by gender
month_total <-rbind(select(df_maxdates,Pandemic_date, total = hom_posi) %>% mutate(gende
r="Male"),
      select(df_maxdates,Pandemic_date, total = muj_posi) %>% mutate(gender="Female"))

#Obtain the last 12 months with data. A control field gender_count2 is created to contro
l the position of the data labels on the chart
month_total <- month_total %>%
  mutate(vertical = case_when(gender == "Female" ~ 5,
                                  TRUE ~ as.numeric(-5))) %>%
  arrange(desc(Pandemic_date)) %>%  head(n = 24L)

#Load necessary libraries
library(ggplot2)
library(babynames)
library(dplyr)
library(hrbrthemes)
library(viridis)
```

```
## Loading required package: viridisLite
```

```
library(viridisLite)

#Generates a bar graph with the total number of cases accumulated in the last 12 months,
grouped by gender
ggplot(month_total, aes(x = Pandemic_date, y = total/1000)) +
  geom_bar(stat = "identity", position = 'dodge') +
  geom_col(aes(color = gender, fill = gender), position = position_stack()) +
  scale_color_manual(values = c("#0073C2FF", "#EFC000FF"))+
  scale_fill_manual(values = c("#0073C2FF", "#EFC000FF")) +
  labs(title = "COVID-19 Cases by Epidemiological Month (Last 12 months) by gender",x =
"Month",y="Total cases",subtitle ="Notation in thousands") +
  geom_text(aes(label = total/1000,vjust=vertical))
```


COVID-19 Cases by Epidemiological Month (Last 12 months) by gender
Notation in thousands
Total cases
600
400
200
0
abr
jul
oct
ene
Month
110.079
126.26
159.817
184.369
204.717
232.369
266.89
280.802
284.044
286.74
343.533
372.297
107.267
124.731
159.169
183.569
203.483
231.357
265.295
279.538
282.944
285.808
351.332
384.796
gender
Female
Male

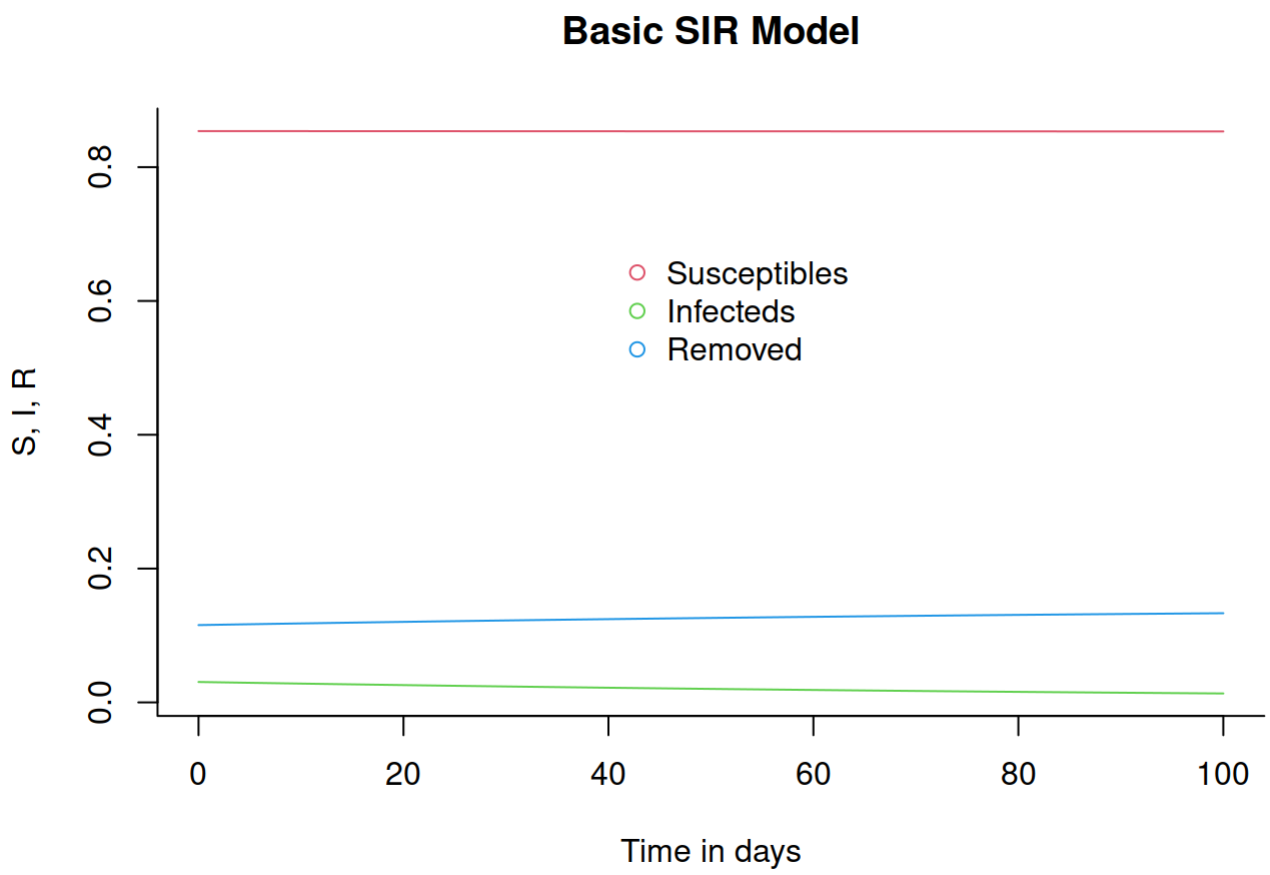
```
#Projection of the behavior of COVID within the next 100 days, using the SIR model
#Based on the example of D. S. Fernández del Viso (https://rpubs.com/dsfernandez/422937)

#Define the Costarrican population
cr_population = 5182351

#Obtain the last variables for the SIR model
cr_SIRData<-select(df_covidcr,Pandemic_date,Infecteds=activos,fallecidos,RECUPERADOS,nue
_posi,positivos)  %>% filter(Pandemic_date=="2022-02-11") %>%
  mutate(Removed=(fallecidos+RECUPERADOS),   susceptibles =(cr_population - (Infecteds +
Removed)))

#Define the variables to use in the model
susceptibles<-(cr_SIRData$susceptibles / cr_population)
Infecteds<-(cr_SIRData$Infecteds / cr_population)
Removed<-(cr_SIRData$Removed / cr_population)
prediction_days<-100
beta_coef<-Infecteds/prediction_days
gamma_coef<-susceptibles/prediction_days

#Load necessary libraries
library(deSolve)

#population size
N = 1

#Initial state for the SIR variables
init <- c(S = susceptibles,
          I = Infecteds,
          R = Removed)

#variable coefficients
param <- c(beta = beta_coef,
           gamma = gamma_coef)

#Create the function with the ODE to evaluate the SIR model
sir <- function(times, init, param) {
  with(as.list(c(init, param)), {
    #Differential equations
    dS <- -beta * S * I
    dI <-  beta * S * I - gamma * I
    dR <-                 gamma * I
    #Exchange rate results
    return(list(c(dS, dI, dR)))
  })
}
#Time range and resolution
times <- seq(0, prediction_days, by = 1)
#Solve system of equations with function 'ode'
out <- ode(y = init, times = times, func = sir, parms = param)
#Set data in the dataframe
out <- as.data.frame(out*N)
```

```
#Remove variable 'time' in out
out$time <- NULL
#Show the last 10 lines
head(out, 10)
```

	S <dbl>	I <dbl>	R <dbl>
1	0.8539094	0.03053267	0.1155580
2	0.8539014	0.03028095	0.1158176
3	0.8538936	0.03003130	0.1160751
4	0.8538858	0.02978372	0.1163305
5	0.8538780	0.02953818	0.1165838
6	0.8538704	0.02929466	0.1168350
7	0.8538628	0.02905315	0.1170841
8	0.8538552	0.02881363	0.1173312
9	0.8538477	0.02857608	0.1175762
10	0.8538403	0.02834049	0.1178192

1-10 of 10 rows

```
#Generate the line plot
matplot(x = times, y = out, type = "l",
        xlab = "Time in days", ylab = "S, I, R", main = "Basic SIR Model",
        lwd = 1, lty = 1, bty = "l", col = 2:4)
legend(40, 0.7, c("Susceptibles", "Infecteds", "Removed"),
       pch = 1, col = 2:4, bty = "n", cex = 1)
```


Basic SIR Model
S, I, R
0.0
0.2
0.4
0.6
0.8
Susceptibles
Infecteds
Removed
0
20
40
60
80
100
Time in days