Assignmet4\_1

Dipesh Poudel

10/5/2021

## Web Scrapping. Getting data from Table in a website

### Getting the Data from Website

library(rvest)  
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

wiki\_link <- "https://en.wikipedia.org/wiki/COVID-19\_pandemic\_in\_Nepal"  
# Reading the whole website  
wiki\_page <- read\_html(wiki\_link)  
# Getting only the tables from the whole website  
tables <- wiki\_page %>% html\_table(fill = TRUE)  
# Printing a sample table  
print(tables[[2]])

## # A tibble: 16 × 2  
## `COVID-19 pandemic in Nepal` `COVID-19 pandemic in Nepal`   
## <chr> <chr>   
## 1 Disease COVID-19   
## 2 Virus strain SARS-CoV-2   
## 3 Location Nepal   
## 4 First outbreak Wuhan, Hubei, China   
## 5 Index case Kathmandu, Bagmati Province   
## 6 Arrival date 9 January 2020(1 year, 8 months, 1 we…  
## 7 Date 23 January 2020   
## 8 Confirmed cases 798,766[1] (updated 6 October 2021)   
## 9 Active cases 18,115 (1 October)[2]   
## 10 Recovered 766,696 (1 October)[2]   
## 11 Deaths 11,180[1] (updated 6 October 2021)   
## 12 Fatality rate 1.4% (1 October)   
## 13 Territories Active cases in out of 77 districts   
## 14 Vaccinations 7,913,555[1] (total vaccinated)6,409,…  
## 15 Government website Government website   
## 16 COVID-19 Dashboard (MoHP) Nepal COVID… COVID-19 Dashboard (MoHP) Nepal COVID…

# Getting the Required Table  
covid\_table <- wiki\_page %>%html\_nodes("table") %>% .[16] %>%html\_table() %>% .[[1]]  
print(head(covid\_table))

## # A tibble: 6 × 14  
## Date `Confirmed case… `Confirmed case… `Confirmed case… Recoveries Recoveries  
## <chr> <chr> <chr> <chr> <chr> <chr>   
## 1 Date Total New Active Total New   
## 2 23 Jan 1 +1 1 0 0   
## 3 24 Jan 1 0 1 0 0   
## 4 25 Jan 1 0 1 0 0   
## 5 26 Jan 1 0 1 0 0   
## 6 27 Jan 1 0 1 0 0   
## # … with 8 more variables: Deaths <chr>, Deaths <chr>, RT-PCR tests <chr>,  
## # RT-PCR tests <chr>, TPR <chr>, RR <chr>, CFR <chr>, Ref. <chr>

### Getting the Structure of the dataframe

str(covid\_table)

## tibble [496 × 14] (S3: tbl\_df/tbl/data.frame)  
## $ Date : chr [1:496] "Date" "23 Jan" "24 Jan" "25 Jan" ...  
## $ Confirmed cases: chr [1:496] "Total" "1" "1" "1" ...  
## $ Confirmed cases: chr [1:496] "New" "+1" "0" "0" ...  
## $ Confirmed cases: chr [1:496] "Active" "1" "1" "1" ...  
## $ Recoveries : chr [1:496] "Total" "0" "0" "0" ...  
## $ Recoveries : chr [1:496] "New" "0" "0" "0" ...  
## $ Deaths : chr [1:496] "Total" "0" "0" "0" ...  
## $ Deaths : chr [1:496] "New" "0" "0" "0" ...  
## $ RT-PCR tests : chr [1:496] "Total" "" "" "" ...  
## $ RT-PCR tests : chr [1:496] "New" "" "" "" ...  
## $ TPR : chr [1:496] "TPR" "" "" "" ...  
## $ RR : chr [1:496] "RR" "0%" "0%" "0%" ...  
## $ CFR : chr [1:496] "CFR" "0%" "0%" "0%" ...  
## $ Ref. : chr [1:496] "Ref." "[169]" "" "" ...

As we can see all the variables (columns) have char data type and also the first row and second row are heading. We need to resolve this. This process of cleaning up data is called scrubbing or Tiding the data.

### Removing the first row and last column

#Changing column names: Column Underscore Row 1  
names(covid\_table) = paste(names(covid\_table), covid\_table[1, ], sep = "\_")  
#Removing first row  
covid\_table = covid\_table[-1, ]  
#Removing last column  
covid\_table <- covid\_table[,-14]  
#Viewing the data  
View(covid\_table)  
#Checking structure of the data  
str(covid\_table)

## tibble [495 × 13] (S3: tbl\_df/tbl/data.frame)  
## $ Date\_Date : chr [1:495] "23 Jan" "24 Jan" "25 Jan" "26 Jan" ...  
## $ Confirmed cases\_Total : chr [1:495] "1" "1" "1" "1" ...  
## $ Confirmed cases\_New : chr [1:495] "+1" "0" "0" "0" ...  
## $ Confirmed cases\_Active: chr [1:495] "1" "1" "1" "1" ...  
## $ Recoveries\_Total : chr [1:495] "0" "0" "0" "0" ...  
## $ Recoveries\_New : chr [1:495] "0" "0" "0" "0" ...  
## $ Deaths\_Total : chr [1:495] "0" "0" "0" "0" ...  
## $ Deaths\_New : chr [1:495] "0" "0" "0" "0" ...  
## $ RT-PCR tests\_Total : chr [1:495] "" "" "" "" ...  
## $ RT-PCR tests\_New : chr [1:495] "" "" "" "" ...  
## $ TPR\_TPR : chr [1:495] "" "" "" "" ...  
## $ RR\_RR : chr [1:495] "0%" "0%" "0%" "0%" ...  
## $ CFR\_CFR : chr [1:495] "0%" "0%" "0%" "0%" ...

### Renaming the columns

#Renaming the column names with underscore between spaces  
colnames(covid\_table)

## [1] "Date\_Date" "Confirmed cases\_Total" "Confirmed cases\_New"   
## [4] "Confirmed cases\_Active" "Recoveries\_Total" "Recoveries\_New"   
## [7] "Deaths\_Total" "Deaths\_New" "RT-PCR tests\_Total"   
## [10] "RT-PCR tests\_New" "TPR\_TPR" "RR\_RR"   
## [13] "CFR\_CFR"

#Checking column names to do correct coding below  
names(covid\_table)[names(covid\_table) == "Date\_Date"] = "Date"  
names(covid\_table)[names(covid\_table) == "Confirmed cases\_Total"] = "Confirmed\_Cases\_Total"  
names(covid\_table)[names(covid\_table) == "Confirmed cases\_New"] = "Confirmed\_Cases\_New"  
names(covid\_table)[names(covid\_table) == "Confirmed cases\_Active"] = "Confirmed\_Cases\_Active"  
names(covid\_table)[names(covid\_table) == "RT-PCR tests\_Total"] = "PCR\_Total"  
names(covid\_table)[names(covid\_table) == "RT-PCR tests\_New"] = "PCR\_New"  
names(covid\_table)[names(covid\_table) == "TPR\_TPR"] = "TPR"  
names(covid\_table)[names(covid\_table) == "RR\_RR"] = "RR"  
names(covid\_table)[names(covid\_table) == "CFR\_CFR"] = "CFR"  
str(covid\_table)

## tibble [495 × 13] (S3: tbl\_df/tbl/data.frame)  
## $ Date : chr [1:495] "23 Jan" "24 Jan" "25 Jan" "26 Jan" ...  
## $ Confirmed\_Cases\_Total : chr [1:495] "1" "1" "1" "1" ...  
## $ Confirmed\_Cases\_New : chr [1:495] "+1" "0" "0" "0" ...  
## $ Confirmed\_Cases\_Active: chr [1:495] "1" "1" "1" "1" ...  
## $ Recoveries\_Total : chr [1:495] "0" "0" "0" "0" ...  
## $ Recoveries\_New : chr [1:495] "0" "0" "0" "0" ...  
## $ Deaths\_Total : chr [1:495] "0" "0" "0" "0" ...  
## $ Deaths\_New : chr [1:495] "0" "0" "0" "0" ...  
## $ PCR\_Total : chr [1:495] "" "" "" "" ...  
## $ PCR\_New : chr [1:495] "" "" "" "" ...  
## $ TPR : chr [1:495] "" "" "" "" ...  
## $ RR : chr [1:495] "0%" "0%" "0%" "0%" ...  
## $ CFR : chr [1:495] "0%" "0%" "0%" "0%" ...

In the data we can see that there are “+” and “%” sign. We need to remove those before analysis.

#Removing + from four variables  
covid\_table$Confirmed\_Cases\_New = gsub('[+]', '',  
covid\_table$Confirmed\_Cases\_New)  
covid\_table$Recoveries\_New = gsub('[+]', '', covid\_table$Recoveries\_New)  
covid\_table$Deaths\_New = gsub('[+]', '', covid\_table$Deaths\_New)  
covid\_table$PCR\_New = gsub('[+]', '', covid\_table$PCR\_New)  
#Removing % from three variables  
covid\_table$TPR = gsub('[%]', '', covid\_table$TPR)  
covid\_table$RR = gsub('[%]', '', covid\_table$RR)  
covid\_table$CFR = gsub('[%]', '', covid\_table$CFR)  
str(covid\_table)

## tibble [495 × 13] (S3: tbl\_df/tbl/data.frame)  
## $ Date : chr [1:495] "23 Jan" "24 Jan" "25 Jan" "26 Jan" ...  
## $ Confirmed\_Cases\_Total : chr [1:495] "1" "1" "1" "1" ...  
## $ Confirmed\_Cases\_New : chr [1:495] "1" "0" "0" "0" ...  
## $ Confirmed\_Cases\_Active: chr [1:495] "1" "1" "1" "1" ...  
## $ Recoveries\_Total : chr [1:495] "0" "0" "0" "0" ...  
## $ Recoveries\_New : chr [1:495] "0" "0" "0" "0" ...  
## $ Deaths\_Total : chr [1:495] "0" "0" "0" "0" ...  
## $ Deaths\_New : chr [1:495] "0" "0" "0" "0" ...  
## $ PCR\_Total : chr [1:495] "" "" "" "" ...  
## $ PCR\_New : chr [1:495] "" "" "" "" ...  
## $ TPR : chr [1:495] "" "" "" "" ...  
## $ RR : chr [1:495] "0" "0" "0" "0" ...  
## $ CFR : chr [1:495] "0" "0" "0" "0" ...

All the variables have char as data type. We need to convert the data type to numeric or integer so that we can perform analysis.

### Converting chr variables as numbers and integers

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.5 ✓ purrr 0.3.4  
## ✓ tibble 3.1.4 ✓ stringr 1.4.0  
## ✓ tidyr 1.1.3 ✓ forcats 0.5.1  
## ✓ readr 2.0.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x readr::guess\_encoding() masks rvest::guess\_encoding()  
## x dplyr::lag() masks stats::lag()

covid\_table<-covid\_table%>%mutate\_at(vars(Confirmed\_Cases\_Total,Confirmed\_Cases\_New,Confirmed\_Cases\_Active,Recoveries\_Total,Recoveries\_New,Deaths\_Total,Deaths\_New,PCR\_Total,PCR\_New),as.integer)

## Warning in mask$eval\_all\_mutate(quo): NAs introduced by coercion  
  
## Warning in mask$eval\_all\_mutate(quo): NAs introduced by coercion  
  
## Warning in mask$eval\_all\_mutate(quo): NAs introduced by coercion  
  
## Warning in mask$eval\_all\_mutate(quo): NAs introduced by coercion  
  
## Warning in mask$eval\_all\_mutate(quo): NAs introduced by coercion  
  
## Warning in mask$eval\_all\_mutate(quo): NAs introduced by coercion  
  
## Warning in mask$eval\_all\_mutate(quo): NAs introduced by coercion  
  
## Warning in mask$eval\_all\_mutate(quo): NAs introduced by coercion

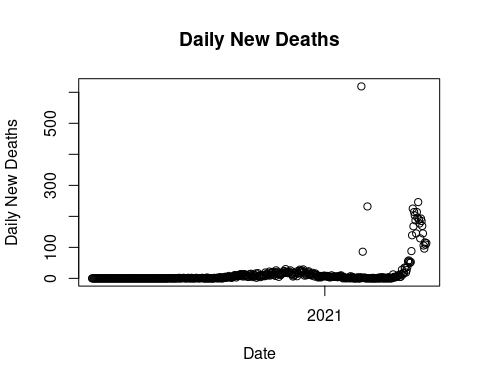
covid\_table<-covid\_table%>%mutate\_at(vars(TPR,RR,CFR),as.numeric)  
str(covid\_table)

## tibble [495 × 13] (S3: tbl\_df/tbl/data.frame)  
## $ Date : chr [1:495] "23 Jan" "24 Jan" "25 Jan" "26 Jan" ...  
## $ Confirmed\_Cases\_Total : int [1:495] 1 1 1 1 1 1 1 1 1 1 ...  
## $ Confirmed\_Cases\_New : int [1:495] 1 0 0 0 0 0 0 0 0 0 ...  
## $ Confirmed\_Cases\_Active: int [1:495] 1 1 1 1 1 1 0 0 0 0 ...  
## $ Recoveries\_Total : int [1:495] 0 0 0 0 0 0 1 1 1 1 ...  
## $ Recoveries\_New : int [1:495] 0 0 0 0 0 0 1 0 0 0 ...  
## $ Deaths\_Total : int [1:495] 0 0 0 0 0 0 0 0 0 0 ...  
## $ Deaths\_New : int [1:495] 0 0 0 0 0 0 0 0 0 0 ...  
## $ PCR\_Total : int [1:495] NA NA NA NA NA 3 4 5 5 NA ...  
## $ PCR\_New : int [1:495] NA NA NA NA NA NA 1 1 0 NA ...  
## $ TPR : num [1:495] NA NA NA NA NA ...  
## $ RR : num [1:495] 0 0 0 0 0 0 100 100 100 100 ...  
## $ CFR : num [1:495] 0 0 0 0 0 0 0 0 0 0 ...

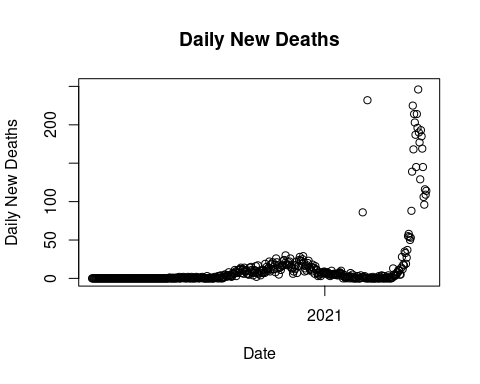
While doing the type casting (conversion) the places where there was no value it is replaced by NA.

# Creating a sequence of date from 2020 Jan 23  
date2<-seq(as.Date('2020-1-23'), by='days', length.out = 495)  
# Adding the sequence of date to the dataframe  
covid\_table<-cbind(covid\_table, date2)

#Plot  
plot(covid\_table$date2, covid\_table$Deaths\_New,main = "Daily New Deaths",xlab = "Date",ylab = "Daily New Deaths")

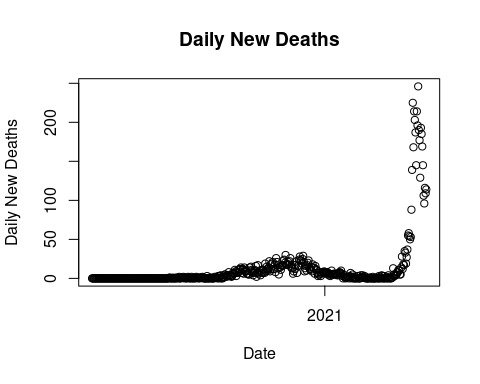


plot(covid\_table$date2, covid\_table$Deaths\_New, ylim =range(0:250),main = "Daily New Deaths",xlab = "Date",ylab = "Daily New Deaths")



As we can see in the first graph, there are three data points that are out of ordinary. For now we will replace these data points using data from other sources.

#Repalce Deaths\_New of 24 Feb as 1 in the data  
covid\_table[covid\_table$date2=="2021-02-24", "Deaths\_New"] = 1  
covid\_table[covid\_table$date2=="2021-02-26", "Deaths\_New"] = 4  
covid\_table[covid\_table$date2=="2021-03-05", "Deaths\_New"] = 0  
  
plot(covid\_table$date2, covid\_table$Deaths\_New, main = "Daily New Deaths",xlab = "Date",ylab = "Daily New Deaths")



summary(covid\_table$Deaths\_New)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 2.00 13.04 10.00 246.00

From summary above we can see that the maximum daily new deaths between 2020-01-23 and 2021-05-31 from covid in Nepal was 246. The mean median and the quartile values are low which means that the new deaths was relatively low for a long period of time and it exploded in the later time period.

summary(covid\_table$Deaths\_Total)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.0 0.0 18.0 142.9 149.0 984.0 210

The total number of deaths in Nepal was 984. However, this does not pain the accurate picture since the Deaths\_Total column has null value from date 2020-11-02.

summary(covid\_table$Deaths\_CFR)

## Length Class Mode   
## 0 NULL NULL

We got this result because we do not have column name Deaths\_CFR. The correct column name is CFR.

summary(covid\_table$CFR)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.2200 0.5800 0.5216 0.7500 1.3200

This shows that the case fatality rate went up to 1.32%. This means that the at max 1.32% of total people who had covid died in Nepal from 2020-01-23 to 2021-05-31.

## Data Modeling

#Simple Moving Average fit for new deaths data!  
library(smooth)

## Loading required package: greybox

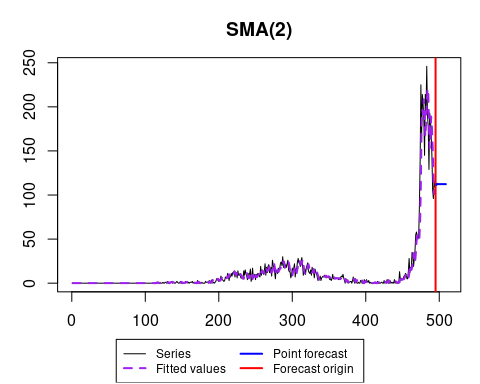
## Package "greybox", v1.0.1 loaded.

##   
## Attaching package: 'greybox'

## The following object is masked from 'package:tidyr':  
##   
## spread

## This is package "smooth", v3.1.3

sma = sma(covid\_table$Deaths\_New, h=14, silent=FALSE)



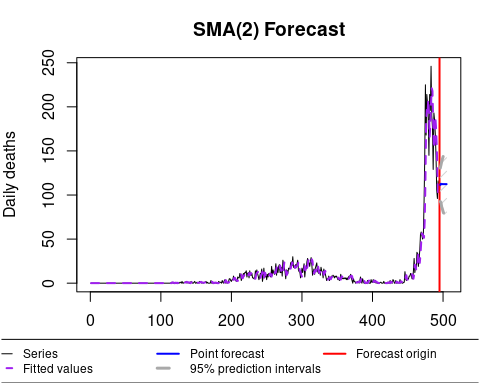
summary(sma)

## Time elapsed: 4.95 seconds  
## Model estimated: SMA(2)  
## Initial values were produced using backcasting.  
##   
## Loss function type: MSE; Loss function value: 92.4722  
## Error standard deviation: 9.6357  
## Sample size: 495  
## Number of estimated parameters: 2  
## Number of degrees of freedom: 493  
## Information criteria:  
## AIC AICc BIC BICc   
## 3649.569 3649.593 3657.978 3658.053

forecast(sma)

## Time Series:  
## Start = 496   
## End = 505   
## Frequency = 1   
## Point forecast Lower bound (2.5%) Upper bound (97.5%)  
## 496 111.5000 92.56783 130.4322  
## 497 112.7500 91.58319 133.9168  
## 498 112.1250 86.63679 137.6132  
## 499 112.4375 84.33662 140.5384  
## 500 112.2812 81.31236 143.2501  
## 501 112.3594 78.99122 145.7275  
## 502 112.3203 76.60990 148.0307  
## 503 112.3398 74.48099 150.1987  
## 504 112.3301 72.41496 152.2452  
## 505 112.3350 70.47560 154.1943

plot(forecast(sma), main = "SMA(2) Forecast", ylab="Daily deaths")



## Importing Data From a Web API

library(jsonlite)

##   
## Attaching package: 'jsonlite'

## The following object is masked from 'package:purrr':  
##   
## flatten

url <- "https://data.askbhunte.com/api/v1/covid/timeline"  
covidtbl <- fromJSON(txt=url, flatten=TRUE)  
colnames(covidtbl)

## [1] "date" "totalCases" "newCases" "totalRecoveries"  
## [5] "newRecoveries" "totalDeaths" "newDeaths"

summary(covidtbl)

## date totalCases newCases totalRecoveries  
## Length:247 Min. : 0 Min. : 0 Min. : 0   
## Class :character 1st Qu.: 4 1st Qu.: 0 1st Qu.: 2   
## Mode :character Median : 1810 Median : 100 Median : 190   
## Mean :14277 Mean : 315 Mean : 9006   
## 3rd Qu.:20208 3rd Qu.: 466 3rd Qu.:14402   
## Max. :77816 Max. :2020 Max. :56282   
## newRecoveries totalDeaths newDeaths   
## Min. : 0.0 Min. : 0.00 Min. : 0.000   
## 1st Qu.: 0.0 1st Qu.: 0.00 1st Qu.: 0.000   
## Median : 4.0 Median : 8.00 Median : 0.000   
## Mean : 227.9 Mean : 72.06 Mean : 2.016   
## 3rd Qu.: 211.5 3rd Qu.: 59.00 3rd Qu.: 2.000   
## Max. :2287.0 Max. :498.00 Max. :16.000