Project 2 Part 2

Kaushal Khatiwada

2024-03-18

Web scraping and data wrangling

```
library(rvest)
library(jsonlite)
```

https://data.covid19india.org (two JSON files)

Find all the hyperlinks in the site and grep the .json link

```
covid_india_urls <- read_html("https://data.covid19india.org") %>%
  html_elements("a") %>%
  html_attr("href")

# Filter URLs with ".json" extension
urls <- covid_india_urls[grep("\\json$",covid_india_urls)]
urls
```

```
## [1] "https://data.covid19india.org/v4/min/timeseries.min.json"
## [2] "https://data.covid19india.org/v4/min/data.min.json"
```

Extract all the data set from json url.

```
covid_india_timeseries <- lapply(urls[1],fromJSON)
timeseries <- covid_india_timeseries[[1]]
```

```
covid_india_data <- lapply(urls[2],fromJSON)
datas <- covid_india_data[[1]]
```

Datawrangling on json

Data set is in the nested list. Tried to clean and format the data set.

```
# For fist json link
# Tried to get values
timeseries$WB$dates$`2021-10-31`$total$vaccinated1   #56192166
```

```
## [1] 56192166
```

```r
#Converted to data Frame
timeseries_df <- as.data.frame(timeseries)

# Unlisted all the elements from the list
test <- unlist(timeseries, recursive = T)
```

```r
head(test)
```

```
##  AN.dates.2020-03-26.delta.confirmed AN.dates.2020-03-26.delta7.confirmed
##                                    1                                     1
##  AN.dates.2020-03-26.total.confirmed  AN.dates.2020-03-27.delta.confirmed
##                                    1                                     5
## AN.dates.2020-03-27.delta7.confirmed  AN.dates.2020-03-27.total.confirmed
##                                    6                                     6
```

```r
# For fist json link
# Tried to get values
datas$AN$districts$`South Andaman`$total$vaccinated1    #189662
```

```
## [1] 189662
```

```r
datas$WB$districts$Darjeeling$total$vaccinated1         #1324555
```

```
## [1] 1324555
```

```r
data_df <- as.data.frame(covid_india_data)
```

AQI FORECAST TABLE

From https://aqicn.org/forecast/kathmandu (aqi forecast table)

```r
#aqi_forecast_table <- read_html("https://aqicn.org/city/kathmandu") %>%
#  html_nodes("table") %>%
#  html_table() %>% .[[6]]
# aqi_forecast_table
#head(aqi_forecast_table)
```

Note: Could not get appropriate table from Base R package so used "Rselenium" package for web scrapping

```r
library(RSelenium)
library(netstat)
library(rvest)
```

Scrape the table with table class "aqiforecast-table" by using web client

```r
remote_driver <- rsDriver(remoteServerAddr = "localhost", browser = "firefox", port=free_port(), chrome
mybrowser <- remote_driver$client
mybrowser$navigate("https://aqicn.org/forecast/kathmandu/")

aqi_html  <- read_html(mybrowser$getPageSource() %>% unlist())
aqi_html %>% html_element("table.aqiforecast-table") %>% html_table() -> forecast_table
remote_driver[["server"]]$stop()    #Close the session
```

```
## [1] TRUE
```

Data Wrangling to remove the NA values, remove unnecssary rows are columns

```
df <- data.frame(forecast_table)

#Remove all column with NA
new_df<- df[,colSums(is.na(df))==0]

#Replace column names with first row of data frame
names(new_df) <- new_df[1,]

#Remove 1st row
forecast_df <- new_df[-1,]

#Remove empty row
forecast_df <- forecast_df[!apply(forecast_df == "", 1, all), ]
forecast_df
```

```
##                       Tuesday 26   Tuesday 26   Tuesday 26   Tuesday 26
## 2              hour            0            3            6            9
## 3             PM2.5       138138       138138       138137       137137
## 4              PM10         5151         5151         5151         5046
## 5                O3           44           44          113         3327
## 6               UVI
## 7  Wind Speed (m/s)            2            2            2            1
## 9              Temp.          13°          13°          17°          22°
## 10          humidity
## 11                   6:02 ~ 18:19 6:02 ~ 18:19 6:02 ~ 18:19 6:02 ~ 18:19
##      Tuesday 26   Tuesday 26   Tuesday 26   Tuesday 26 Wednesday 27
## 2            12           15           18           21            0
## 3        137137       137137       138138       138138       138138
## 4          4646         4646         4646         4646         5148
## 5          2823         2220          169           74           54
## 6
## 7             3            3            2            1            2
## 9           22°          21°          16°          15°          15°
## 10
## 11 6:02 ~ 18:19 6:02 ~ 18:19 6:02 ~ 18:19 6:02 ~ 18:19 6:00 ~ 18:19
##    Wednesday 27 Wednesday 27 Wednesday 27 Wednesday 27 Wednesday 27
## 2             3            6            9           12           15
## 3        151143       151151       147138       138138       138138
## 4          5151         5151         5151         5046         4646
## 5            65           95         2820         2621         2019
## 6
## 7             2            1            3            3            2
## 9           14°          18°          23°          23°          20°
## 10
## 11 6:00 ~ 18:19 6:00 ~ 18:19 6:00 ~ 18:19 6:00 ~ 18:19 6:00 ~ 18:19
##    Wednesday 27 Wednesday 27  Thursday 28  Thursday 28  Thursday 28
## 2            18           21            0            3            6
## 3        138138       138138       138138       138137       125103
## 4          4646         4646         4646         4646         4646
```

```
## 5          1713          115           44           33           43
## 6
## 7             1            2            2            1            1
## 9           16°          16°          14°          14°          18°
## 10
## 11 6:00 ~ 18:19 6:00 ~ 18:19 5:59 ~ 18:20 5:59 ~ 18:20 5:59 ~ 18:20
##     Thursday 28  Thursday 28  Thursday 28  Thursday 28  Thursday 28
## 2             9           12           15           18           21
## 3        137115       137137       138138       138138       138138
## 4          4646         4646         4646         4646         4646
## 5          2819         2925         2321          167           54
## 6
## 7             2            2            2            1            2
## 9           23°          23°          20°          17°          18°
## 10
## 11 5:59 ~ 18:20 5:59 ~ 18:20 5:59 ~ 18:20 5:59 ~ 18:20 5:59 ~ 18:20
##       Friday 29    Friday 29    Friday 29    Friday 29    Friday 29
## 2             0            3            6            9           12
## 3        138138       138138       138138       138138       138138
## 4          4646         4646         4646         4646         4646
## 5            43           32           42         2315         2423
## 6
## 7             2            1            1            1            1
## 9           16°          16°          19°          21°          21°
## 10
## 11 5:58 ~ 18:20 5:58 ~ 18:20 5:58 ~ 18:20 5:58 ~ 18:20 5:58 ~ 18:20
##       Friday 29    Friday 29    Friday 29  Saturday 30  Saturday 30
## 2            15           18           21            0            3
## 3        138138       138138       138138       138137       137137
## 4          4646         4646         4646         4646         5148
## 5          2220          169           74           44           54
## 6
## 7             1            1            2            2            1
## 9           21°          18°          16°          17°          16°
## 10
## 11 5:58 ~ 18:20 5:58 ~ 18:20 5:58 ~ 18:20 5:57 ~ 18:21 5:57 ~ 18:21
##     Saturday 30  Saturday 30  Saturday 30  Saturday 30  Saturday 30
## 2             6            9           12           15           18
## 3        137137       137137       137137       138138       138138
## 4          5151         5046         4646         4646         4646
## 5           134         2725         2219         1818          147
## 6
## 7             1            2            4            4            4
## 9           20°          26°          28°          26°          16°
## 10
## 11 5:57 ~ 18:21 5:57 ~ 18:21 5:57 ~ 18:21 5:57 ~ 18:21 5:57 ~ 18:21
##     Saturday 30    Sunday 31    Sunday 31    Sunday 31    Sunday 31
## 2            21            0            3            9           12
## 3        138138       138137       125103       103103       103103
## 4          4646         4646         4646         4646         4646
## 5            54           55           55
## 6
## 7             3            2            2            4            5
## 9           14°          16°          16°          28°          28°
```

```
## 10
## 11 5:57 ~ 18:21 5:56 ~ 18:21 5:56 ~ 18:21 5:56 ~ 18:21 5:56 ~ 18:21
##       Sunday 31    Sunday 31    Sunday 31     Monday 1     Monday 1
## 2            15           18           21            0            3
## 3         10394         8989         8989         8979         7570
## 4          4646         4638         3434         3434         3430
## 5
## 6
## 7             1            3            1            1            2
## 9           24°          18°          18°          17°          16°
## 10
## 11 5:56 ~ 18:21 5:56 ~ 18:21 5:56 ~ 18:21 5:55 ~ 18:22 5:55 ~ 18:22
##        Monday 1     Monday 1     Monday 1     Monday 1     Monday 1
## 2             6            9           12           15           18
## 3          6868         7268         8475         8989         8989
## 4          3228         4234         5046         5151         5151
## 5
## 6
## 7             2            2            4            2            2
## 9           24°          29°          29°          25°          19°
## 10
## 11 5:55 ~ 18:22 5:55 ~ 18:22 5:55 ~ 18:22 5:55 ~ 18:22 5:55 ~ 18:22
```