Project 2: Unit 2

Suman Paudel 33

2024-03-26

Contents

Part 1 **1**

Part 2 **10**

Part 3 **26**

Part 1

Task 1

Load all of the necessary packages need for task 1.

```
# Load the packages
library(foreign)
library(gt)
library(tidyverse)
library(magrittr)
library(readxl)
```

Load the Data using CSV module from base R

```
# load the data using Base R read.csv
data <- read.csv("covnep_252days.csv")

summary(data$totalCases)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       2     963   13376   19341   77816
```

Since we need value as 1 instead of zero We can achieve this using multiple ways like ifelse or pmax or subsetting

Using `ifelse`

```
# using ifelse
totalCases_ifelse <- ifelse(data$totalCases < 1, 1, data$totalCases)
summary(totalCases_ifelse)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1       2     963   13377   19341   77816
```

Using pmax

```
# using pmax
totalCases_pmax <- pmax(data$totalCases, 1)
summary(totalCases_pmax)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1       2     963   13377   19341   77816
```

Using subsetting

```
# subsetting
totalCases_subsetting <- data$totalCases
totalCases_subsetting[totalCases_subsetting < 1] <- 1
summary(totalCases_subsetting)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1       2     963   13377   19341   77816
```

Task 2

Read the .sav file using foreign library's read.spss function

For q01

```
# read the .sav file using read_sav function from haven
saq_data <- read.spss("SAQ8.sav",to.data.frame=TRUE)

# for q1
q01 <- saq_data$q01

# computer mathematical operations
datalevels_q01 <- levels(q01)
freq_q01 <- as.numeric(table(q01))
percent_q01 <- as.numeric(round(prop.table(freq_q01) * 100, 1))
valid_percent_q01 <- as.numeric(round(prop.table(freq_q01) * 100, 1))
cum_percent <- cumsum(percent_q01)

# Create data frame
data <- data.frame(
  Levels = datalevels_q01,
  Freq = freq_q01,
  Percent = percent_q01,
  Val_Percent = valid_percent_q01,
  Cum_Percent = cum_percent
)

head(data)
```

```
##              Levels Freq Percent Val_Percent Cum_Percent
## 1    Strongly agree  270    10.5        10.5        10.5
## 2             Agree 1338    52.0        52.0        62.5
## 3           Neither  735    28.6        28.6        91.1
## 4          Disagree  187     7.3         7.3        98.4
## 5 Strongly disagree   41     1.6         1.6       100.0
```

```
# final version of calculated table for q01
data <- data %>% add_row(Levels = "Total", Freq = sum(data$Freq),
                 Percent = sum(data$Percent),
                 Val_Percent = sum(data$Val_Percent),
                 Cum_Percent = NULL)

# aethetics table using gt
data %>% gt(rowname_col = 'Levels') %>%
  tab_header(title = md("Statistics makes me cry")) %>%
  cols_label(Freq = "Frequency",
             Percent = "Percent",
             Val_Percent = "Valid Percent",
             Cum_Percent = "Cumulative Percent") %>%
  sub_missing(missing_text = "")
```

Statistics makes me cry

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly agree	270	10.5	10.5	10.5
Agree	1338	52.0	52.0	62.5
Neither	735	28.6	28.6	91.1
Disagree	187	7.3	7.3	98.4
Strongly disagree	41	1.6	1.6	100.0
Total	2571	100.0	100.0	

For q03

```
# extract q03

q03 <- saq_data$q03
datalevels_q03 <- levels(q03)
freq_q03 <- as.numeric(table(q03))
percent_q03 <- as.numeric(round(prop.table(freq_q03) * 100, 1))
valid_percent_q03 <- as.numeric(round(prop.table(freq_q03) * 100, 1))
cum_percent_q03 <- cumsum(percent_q03)

# convert the computed values into dataframe
data_q03 <- data.frame(
  Levels = datalevels_q03,
  Freq = freq_q03,
  Percent = percent_q03,
  Val_Percent = valid_percent_q03,
  Cum_Percent = cum_percent_q03
)

head(data_q03)
```

```
##              Levels Freq Percent Val_Percent Cum_Percent
## 1    Strongly agree  497    19.3        19.3        19.3
## 2             Agree  672    26.1        26.1        45.4
## 3           Neither  878    34.2        34.2        79.6
## 4          Disagree  448    17.4        17.4        97.0
## 5 Strongly disagree   76     3.0         3.0       100.0
```

```
# add row for total
data_q03 <- data_q03 %>% add_row(Levels = "Total",
                         Freq = sum(data_q03$Freq),
                         Percent = sum(data_q03$Percent),
                         Val_Percent = sum(data_q03$Val_Percent),
                         Cum_Percent = NULL)

# final version of calculated table
data_q03 %>% gt(rowname_col = 'Levels') %>%
  tab_header(title = md("Statistic makes me cry")) %>%
  cols_label(Freq = "Frequency",
             Percent = "Percent",
             Val_Percent = "Valid Percent",
             Cum_Percent = "Cumulative Percent") %>%
  sub_missing(missing_text = "")
```

Statistic makes me cry

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly agree	497	19.3	19.3	19.3
Agree	672	26.1	26.1	45.4
Neither	878	34.2	34.2	79.6
Disagree	448	17.4	17.4	97.0
Strongly disagree	76	3.0	3.0	100.0
Total	2571	100.0	100.0	

For q06

```
# extract q06
q06 <- saq_data$q06

# mathematical computation
datalevels_q06 <- levels(q06)
freq_q06 <- as.numeric(table(q06))
percent_q06 <- as.numeric(round(prop.table(freq_q06) * 100, 1))
```

```r
valid_percent_q06 <- as.numeric(round(prop.table(freq_q06) * 100, 1))
cum_percent_q06 <- cumsum(percent_q06)

# convert into dataframe
data_q06 <- data.frame(
  Levels = datalevels_q06,
  Freq = freq_q06,
  Percent = percent_q06,
  Val_Percent = valid_percent_q06,
  Cum_Percent = cum_percent_q06
)

# add row for total
data_q06 <- data_q06 %>% add_row(Levels = "Total",
                         Freq = sum(data_q06$Freq),
                         Percent = sum(data_q06$Percent),
                         Val_Percent = sum(data_q06$Val_Percent),
                         Cum_Percent = NULL)

# final version of calculated table
data_q06 %>% gt(rowname_col = 'Levels') %>%
  tab_header(title = md("I have little experience of computer")) %>%
  cols_label(Freq = "Frequency",
             Percent = "Percent",
             Val_Percent = "Valid Percent",
             Cum_Percent = "Cumulative Percent") %>%
  sub_missing(missing_text = "")
```

I have little experience of computer

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly agree	702	27.3	27.3	27.3
Agree	1127	43.8	43.8	71.1
Neither	344	13.4	13.4	84.5
Disagree	252	9.8	9.8	94.3
Strongly disagree	146	5.7	5.7	100.0
Total	2571	100.0	100.0	

For q08

```r
# for q08
q08 <- saq_data$q08

# mathematical computation
datalevels_q08 <- levels(q08)
freq_q08 <- as.numeric(table(q08))
percent_q08 <- as.numeric(round(prop.table(freq_q08) * 100, 2))
valid_percent_q08 <- as.numeric(round(prop.table(freq_q08) * 100, 2))
cum_percent_q08 <- cumsum(percent_q08)

# convert into dataframe
data_q08 <- data.frame(
  Levels = datalevels_q08,
  Freq = freq_q08,
  Percent = round(valid_percent_q08,1),
  Val_Percent = round(valid_percent_q08,1),
  Cum_Percent = round(cum_percent_q08,1)
)

# add row for total
data_q08 <- data_q08 %>% add_row(Levels = "Total",
                         Freq = sum(data_q08$Freq),
                         Percent = sum(data_q08$Percent),
                         Val_Percent = sum(data_q08$Val_Percent),
                         Cum_Percent = NULL)


# final version of calculated table
data_q08 %>% gt(rowname_col = 'Levels') %>%
  tab_header(title = md("I have never been good at mathematics")) %>%
  cols_label(Freq = "Frequency",
             Percent = "Percent",
             Val_Percent = "Valid Percent",
             Cum_Percent = "Cumulative Percent") %>%
  sub_missing(missing_text = "")
```

I have never been good at mathematics

	Frequency	Percent	Valid Percent	Cumulative Percent
Strongly agree	383	14.9	14.9	14.9
Agree	1487	57.8	57.8	72.7
Neither	482	18.8	18.8	91.5
Disagree	147	5.7	5.7	97.2
Strongly disagree	72	2.8	2.8	100.0
Total	2571	100.0	100.0	

Task 3

```
mr_drugs <- read_xlsx("MR_Drugs.xlsx")

inco <- mr_drugs %>% select(starts_with('inco'))

transform_inco <- mr_drugs %>% select(starts_with('inco')) %>%
  colSums() %>%
  enframe("income", "N") %>%
  mutate(Percent = round(N / sum(N) * 100, 1))

transform_inco
```

```
## # A tibble: 7 x 3
##   income      N Percent
##   <chr>  <dbl>   <dbl>
## 1 inco1    226    12.8
## 2 inco2    607    34.5
## 3 inco3    293    16.6
## 4 inco4     50     2.8
## 5 inco5     82     4.7
## 6 inco6    151     8.6
## 7 inco7    352    20
```

```
# get the frequencies of 0 and 1 and convert to dataframe
income_frequencies <- apply(inco, 2, table) %>%
  t() %>% as.data.frame()
income_frequencies
```

```
##         0   1
## inco1 746 226
## inco2 365 607
## inco3 679 293
## inco4 922  50
## inco5 890  82
## inco6 821 151
## inco7 620 352
```

```
transform_inco <- transform_inco %>%
  mutate(`Percent of Cases` =
           round(transform_inco$N / (transform_inco$N + income_frequencies[, 1]) * 100, 1))

transform_inco
```

```
## # A tibble: 7 x 4
##   income     N Percent `Percent of Cases`
##   <chr>  <dbl>   <dbl>              <dbl>
## 1 inco1    226    12.8               23.3
## 2 inco2    607    34.5               62.4
## 3 inco3    293    16.6               30.1
## 4 inco4     50     2.8                5.1
## 5 inco5     82     4.7                8.4
## 6 inco6    151     8.6               15.5
## 7 inco7    352    20                 36.2
```

Mathematical Computation

```
# final version of calculated table
final_inco <- transform_inco %>% add_row(
  income = "Total",
  N = sum(transform_inco$N),
  Percent = round(sum(transform_inco$Percent),2),
  "Percent of Cases" = round(sum(transform_inco$`Percent of Cases`),2),)

# converting into percentage
final_inco$Percent <- paste0(sprintf("%.1f", final_inco$Percent),"%")
final_inco$`Percent of Cases` <- paste0(sprintf("%.1f", final_inco$`Percent of Cases`),"%")
final_inco
```

income	N	Percent	Percent of Cases
inco1	226	12.8%	23.3%
inco2	607	34.5%	62.4%
inco3	293	16.6%	30.1%
inco4	50	2.8%	5.1%
inco5	82	4.7%	8.4%
inco6	151	8.6%	15.5%
inco7	352	20.0%	36.2%
Total	1761	100.0%	181.0%

Final Table using gt table

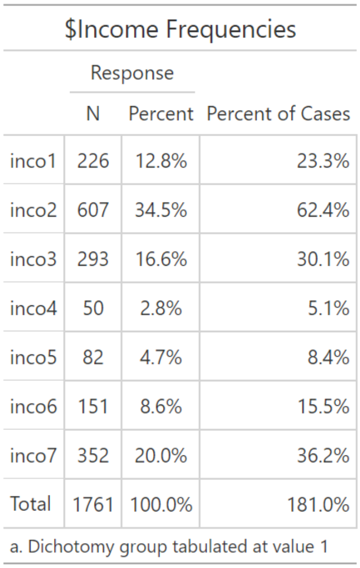
```
final_inco %>% gt(rowname_col = 'income') %>%
  tab_spanner(label='Response',columns = c('N','Percent')) %>%

  tab_header(title = md("$Income Frequencies")) %>%
  tab_footnote(footnote = "a. Dichotomy group tabulated at value 1",
               placement = c('auto')) %>% tab_options(footnotes.multiline = FALSE)
```

$Income Frequencies

	Response		
	N	Percent	Percent of Cases
inco1	226	12.8%	23.3%
inco2	607	34.5%	62.4%
inco3	293	16.6%	30.1%
inco4	50	2.8%	5.1%
inco5	82	4.7%	8.4%
inco6	151	8.6%	15.5%
inco7	352	20.0%	36.2%
Total	1761	100.0%	181.0%

a. Dichotomy group tabulated at value 1

```
knitr::include_graphics('inco.png')
```



$Income Frequencies

	Response		
	N	Percent	Percent of Cases
inco1	226	12.8%	23.3%
inco2	607	34.5%	62.4%
inco3	293	16.6%	30.1%
inco4	50	2.8%	5.1%
inco5	82	4.7%	8.4%
inco6	151	8.6%	15.5%
inco7	352	20.0%	36.2%
Total	1761	100.0%	181.0%

a. Dichotomy group tabulated at value 1

Part 2

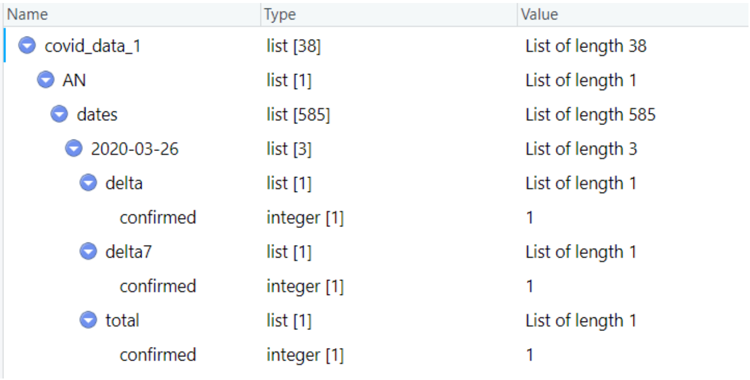
Task 1

Load the necessary library needed for Part 2

```
library(jsonlite) #for working with json data
library(RSelenium) #for web scraping of dynamic table
library(rvest) #scraping the webpage into tibble or df
library(netstat) #for selenium driver
library(stringr) #string manipulation
```

```
data_1 = 'https://data.covid19india.org/v4/min/timeseries.min.json'
data_2 = 'https://data.covid19india.org/v4/min/data.min.json'
covid_data_1 <- jsonlite::fromJSON(data_1)
covid_data_2 <- jsonlite::fromJSON(data_2)
```

```
knitr::include_graphics('cov.png')
```

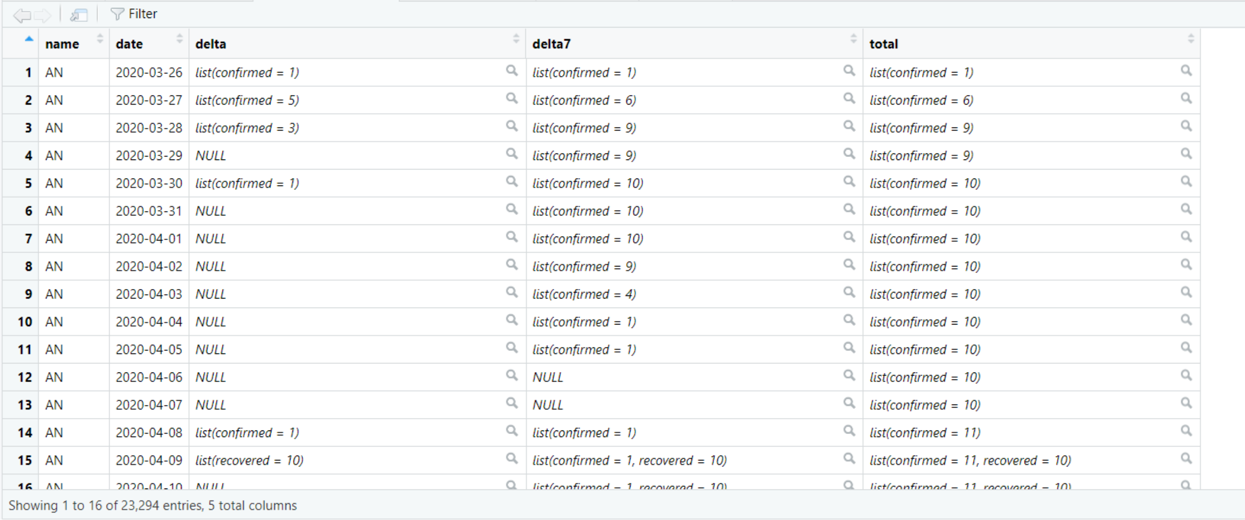


Name	Type	Value
covid_data_1	list [38]	List of length 38
AN	list [1]	List of length 1
dates	list [585]	List of length 585
2020-03-26	list [3]	List of length 3
delta	list [1]	List of length 1
confirmed	integer [1]	1
delta7	list [1]	List of length 1
confirmed	integer [1]	1
total	list [1]	List of length 1
confirmed	integer [1]	1

Sample of raw json data for first record

```
covid_1_parsed <-
  covid_data_1 %>% enframe() %>% unnest_wider(value) %>% unnest_wider(dates) %>%
  pivot_longer(cols = !name,
               names_to = 'date',
               values_to = "value") %>% unnest_wider(value)
```

```
knitr::include_graphics("covid.png")
```

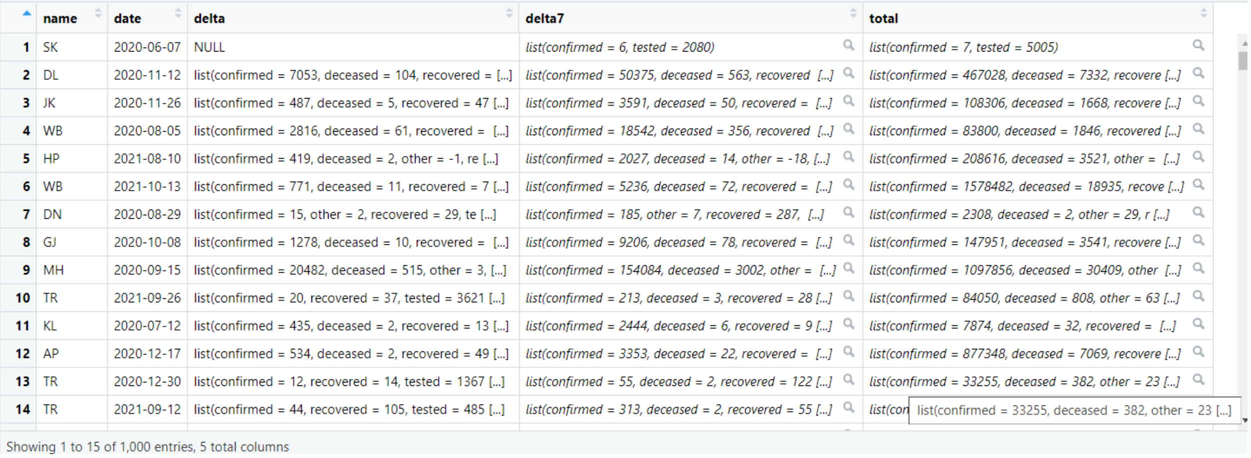


	name	date	delta	delta7	total
1	AN	2020-03-26	list(confirmed = 1)	list(confirmed = 1)	list(confirmed = 1)
2	AN	2020-03-27	list(confirmed = 5)	list(confirmed = 6)	list(confirmed = 6)
3	AN	2020-03-28	list(confirmed = 3)	list(confirmed = 9)	list(confirmed = 9)
4	AN	2020-03-29	NULL	list(confirmed = 9)	list(confirmed = 9)
5	AN	2020-03-30	list(confirmed = 1)	list(confirmed = 10)	list(confirmed = 10)
6	AN	2020-03-31	NULL	list(confirmed = 10)	list(confirmed = 10)
7	AN	2020-04-01	NULL	list(confirmed = 10)	list(confirmed = 10)
8	AN	2020-04-02	NULL	list(confirmed = 9)	list(confirmed = 10)
9	AN	2020-04-03	NULL	list(confirmed = 4)	list(confirmed = 10)
10	AN	2020-04-04	NULL	list(confirmed = 1)	list(confirmed = 10)
11	AN	2020-04-05	NULL	list(confirmed = 1)	list(confirmed = 10)
12	AN	2020-04-06	NULL	NULL	list(confirmed = 10)
13	AN	2020-04-07	NULL	NULL	list(confirmed = 10)
14	AN	2020-04-08	list(confirmed = 1)	list(confirmed = 1)	list(confirmed = 11)
15	AN	2020-04-09	list(recovered = 10)	list(confirmed = 1, recovered = 10)	list(confirmed = 11, recovered = 10)
16	AN	2020-04-10	NULL	list(confirmed = 1, recovered = 10)	list(confirmed = 11, recovered = 10)

Showing 1 to 16 of 23,294 entries, 5 total columns

Sample parsed till dates

```
num_rows <- nrow(covid_1_parsed)
selected_rows <- sample(1:num_rows, 1000)
covid_1_parsed_subset <- covid_1_parsed[selected_rows, ]

knitr::include_graphics("covid2.png")
```



	name	date	delta	delta7	total
1	SK	2020-06-07	NULL	list(confirmed = 6, tested = 2080)	list(confirmed = 7, tested = 5005)
2	DL	2020-11-12	list(confirmed = 7053, deceased = 104, recovered = [...]	list(confirmed = 50375, deceased = 563, recovered [...]	list(confirmed = 467028, deceased = 7332, recovere [...]
3	JK	2020-11-26	list(confirmed = 487, deceased = 5, recovered = 47 [...]	list(confirmed = 3591, deceased = 50, recovered = [...]	list(confirmed = 108306, deceased = 1668, recovere [...]
4	WB	2020-08-05	list(confirmed = 2816, deceased = 61, recovered = [...]	list(confirmed = 18542, deceased = 356, recovered [...]	list(confirmed = 83800, deceased = 1846, recovered [...]
5	HP	2021-08-10	list(confirmed = 419, deceased = 2, other = -1, re [...]	list(confirmed = 2027, deceased = 14, other = -18, [...]	list(confirmed = 208616, deceased = 3521, other = [...]
6	WB	2021-10-13	list(confirmed = 771, deceased = 11, recovered = 7 [...]	list(confirmed = 5236, deceased = 72, recovered = [...]	list(confirmed = 1578482, deceased = 18935, recove [...]
7	DN	2020-08-29	list(confirmed = 15, other = 2, recovered = 29, te [...]	list(confirmed = 185, other = 7, recovered = 287, [...]	list(confirmed = 2308, deceased = 2, other = 29, r [...]
8	GJ	2020-10-08	list(confirmed = 1278, deceased = 10, recovered = [...]	list(confirmed = 9206, deceased = 78, recovered = [...]	list(confirmed = 147951, deceased = 3541, recovere [...]
9	MH	2020-09-15	list(confirmed = 20482, deceased = 515, other = 3, [...]	list(confirmed = 154084, deceased = 3002, other = [...]	list(confirmed = 1097856, deceased = 30409, other [...]
10	TR	2021-09-26	list(confirmed = 20, recovered = 37, tested = 3621 [...]	list(confirmed = 213, deceased = 3, recovered = 28 [...]	list(confirmed = 84050, deceased = 808, other = 63 [...]
11	KL	2020-07-12	list(confirmed = 435, deceased = 2, recovered = 13 [...]	list(confirmed = 2444, deceased = 6, recovered = 9 [...]	list(confirmed = 7874, deceased = 32, recovered = [...]
12	AP	2020-12-17	list(confirmed = 534, deceased = 2, recovered = 49 [...]	list(confirmed = 3353, deceased = 22, recovered = [...]	list(confirmed = 877348, deceased = 7069, recovere [...]
13	TR	2020-12-30	list(confirmed = 12, recovered = 14, tested = 1367 [...]	list(confirmed = 55, deceased = 2, recovered = 122 [...]	list(confirmed = 33255, deceased = 382, other = 23 [...]
14	TR	2021-09-12	list(confirmed = 44, recovered = 105, tested = 485 [...]	list(confirmed = 313, deceased = 2, recovered = 55 [...]	list(con list(confirmed = 33255, deceased = 382, other = 23 [...]

Showing 1 to 15 of 1,000 entries, 5 total columns

```
covid_1_parsed_subset <- covid_1_parsed_subset %>%
  mutate(across(c(delta, delta7, total), ~ map(., ~ set_names(  as_tibble(.x), paste0(cur_column(), "_"
  unnest_wider(c(delta, delta7, total))
covid_1_parsed_subset
```

```
## # A tibble: 1,000 x 23
##    name  date       delta_confirmed delta_deceased delta_recovered delta_tested
##    <chr> <chr>                <int>          <int>           <int>        <int>
##  1 TG    2021-10-08             201              1             220        47465
##  2 AN    2021-01-06              NA             NA               3         1236
##  3 AS    2021-03-05              29             NA              20        13551
##  4 MH    2021-06-28            6727            287           10812       166163
##  5 MN    2021-01-31              16             NA               8         1310
##  6 DL    2021-06-15              228            12              364         71291
##  7 UN    2021-09-16               NA            NA               NA            NA
##  8 MN    2021-01-20               19            NA               10          1248
##  9 UT    2020-06-29                8             1               93          1412
## 10 MP    2021-06-13              277            21              780         76880
## # i 990 more rows
## # i 17 more variables: delta_vaccinated1 <int>, delta_vaccinated2 <int>,
## #   delta_other <int>, delta7_confirmed <int>, delta7_deceased <int>,
## #   delta7_recovered <int>, delta7_tested <int>, delta7_vaccinated1 <int>,
## #   delta7_vaccinated2 <int>, delta7_other <int>, total_confirmed <int>,
## #   total_deceased <int>, total_recovered <int>, total_tested <int>,
## #   total_vaccinated1 <int>, total_vaccinated2 <int>, total_other <int>
```

```
# delta parsed
covid_1_parsed_subset[80:150,] %>% select(starts_with('delta'))
```

delta_confirmed	delta_deceased	delta_recovered	delta_tested	delta_vaccinated1	delta_vaccinated2	delta_other	delta7_confirmed	delta7_deceased	delta7_recovered	delta7_tested	delta7_vaccinated1	delta7_vaccinated2	delta7_other
NA	NA	1	64	NA	NA	NA	18	NA	30	970	2882	NA	1
2046	20	2426	21833	NA	NA	NA	14375	297	18094	150433	NA	NA	NA
17	NA	25	1140	470	NA	NA	213	1	188	8195	1945	NA	NA
38	NA	8	396	NA	NA	NA	254	2	105	2047	NA	NA	NA
109	NA	113	21818	19	NA	NA	762	11	1281	154195	4414	NA	NA
14233	173	15355	107096	105664	12846	2	101741	1294	141300	729062	986706	85055	14
3509	58	3612	75374	NA	NA	3	22161	463	23675	406429	NA	NA	12
198	NA	67	2542	NA	NA	NA	912	10	467	16679	NA	NA	NA
2	NA	5	499	10244	1479	NA	33	NA	28	5207	63364	5814	NA
1758	15	2287	32677	NA	NA	NA	12685	64	9114	349857	NA	NA	NA
NA	NA	NA	NA	1658	6809	NA	4	NA	2	NA	7810	25994	NA
4	NA	NA	947	NA	NA	NA	29	NA	2	2558	NA	NA	NA
2918	24	4303	61330	NA	NA	NA	27099	197	35820	496805	NA	NA	NA
2177	11	1006	36750	NA	NA	NA	11659	81	6399	349693	NA	NA	NA
332	11	515	79177	145609	295097	NA	2924	61	5077	742757	866835	1626293	NA
108	1	85	18205	15704	NA	NA	689	9	854	123741	192272	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
70	NA	43	2843	NA	NA	NA	585	1	329	17744	NA	NA	-1
14120	174	8595	173909	68627	84394	NA	98114	1090	47959	1228730	536501	579938	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
9	NA	29	682	NA	NA	NA	130	1	113	5123	NA	NA	NA
3	NA	3	NA	435	62	NA	20	NA	32	NA	1320	290	NA
993	9	1417	112982	194173	29362	NA	7207	68	10516	795693	1231962	120505	NA
1186	24	1776	132192	268899	101431	NA	11427	205	11925	1040939	1262623	619740	2
3178	11	2201	28705	138650	4949	NA	18168	82	10936	174658	989498	49275	NA
2667	50	1909	50697	NA	NA	NA	15608	264	10078	325814	NA	NA	NA
4178	61	4389	102922	21908	2855	NA	33206	472	34557	760599	237386	26714	NA
NA	NA	NA	115	NA	NA	NA	9	NA	1	1142	NA	NA	NA
582	NA	NA	NA	NA	NA	NA	3149	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
485	5	365	3981	10409	1710	NA	3095	35	3073	27826	62789	6443	NA
361	9	324	22211	NA	NA	NA	2094	88	3079	141983	NA	NA	NA
46	1	5	NA	NA	NA	NA	262	1	57	5315	NA	NA	NA
23	NA	28	67851	57415	24511	NA	224	4	233	385266	557419	183540	NA

delta_confirmed	delta_deceased	delta_recovered	delta_tested	delta_vaccinated1	delta_vaccinated2	delta_other	delta7_confirmed	delta7_deceased	delta7_recovered	delta7_tested	delta7_vaccinated1	delta7_vaccinated2	delta7_other
24	NA	9	401	NA	NA	NA	139	5	35	2117	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
12	NA	NA	177	NA	NA	NA	16	NA	13	875	NA	NA	NA
1	NA	NA	NA	NA	NA	NA	3	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
119	NA	226	NA	296	509	NA	1303	NA	1640	NA	2812	3204	2
166	NA	92	837	NA	NA	NA	877	16	505	4903	NA	NA	NA
554	20	510	9701	NA	NA	NA	3628	111	2766	126425	NA	NA	NA
1307	7	2561	33346	NA	NA	NA	9398	88	12020	215656	NA	NA	NA
713	5	668	136770	NA	NA	NA	4306	37	4099	868502	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
22823	401	52252	140783	238867	29237	NA	179079	3599	317345	886286	1089463	97983	NA
34	NA	38	61094	41239	60681	NA	225	1	208	427343	292189	450927	NA
34	1	36	21875	261016	85019	NA	137	1	158	106887	419311	359592	NA
215	2	95	11094	5465	45644	NA	1229	22	1260	62193	34641	195881	NA
7	NA	5	824	76	217	NA	38	1	60	4677	800	1014	NA
4797	130	3710	199963	157847	60812	NA	39332	1043	45545	1391211	1752705	906524	35
1501	27	1889	54741	66911	52872	NA	11998	219	13917	377804	1127080	559365	NA
113	4	109	2618	527	5361	NA	839	11	1067	16404	4847	45708	NA
44	NA	105	4852	72	443	NA	313	2	559	36672	6991	49575	NA
NA	NA	NA	76	NA	NA	NA	10	NA	20	991	NA	NA	NA
25	1	16	1592	NA	NA	NA	169	8	65	10626	NA	NA	NA
128	2	165	1628	10435	1969	1	698	18	909	8819	63446	6164	8
3	NA	30	3953	NA	NA	NA	121	4	350	8848	NA	NA	NA
12	NA	10	11832	2222	9728	NA	77	2	146	99533	51203	180295	3
6	NA	4	1357	9514	2911	NA	24	NA	30	10866	48438	23242	NA
92	1	76	2233	1234	1220	7	446	15	727	11053	7005	12119	18
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
9	NA	5	288	113	206	NA	34	NA	27	2895	653	3708	NA
938	1	334	4619	99	97	NA	4930	14	2209	33153	17280	44126	NA
1085	15	1410	27523	NA	NA	NA	7818	77	9152	199118	NA	NA	NA
20295	832	31964	258759	272501	34154	14	159990	6730	228743	1885584	1374190	135403	191
NA	NA	NA	99	72	1168	NA	3	NA	11	3967	829	6996	1
NA	NA	NA	NA	1397	4015	NA	2	NA	3	NA	8612	20060	NA
37	1	43	32157	1866	7101	2	259	14	360	44474	8818	40003	23
NA	NA	NA	NA	NA	NA	NA	10	NA	NA	NA	NA	NA	NA

```
# delta7 parsed
covid_1_parsed_subset[345:451,] %>% select(starts_with('delta7'))
```

delta7_confirmed	delta7_deceased	delta7_recovered	delta7_tested	delta7_vaccinated1	delta7_vaccinated2	delta7_other
383	1	337	5392	NA	NA	NA
1171	7	484	32867	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA
351	11	357	4143	NA	NA	NA
7168	158	18631	330482	224293	23765	NA
150	9	124	33294	NA	NA	NA
1091	10	984	13444	7856	1859	NA
225841	2191	142902	1123025	215951	162973	NA
3577	102	2341	34202	NA	NA	NA

delta7_confirmed	delta7_deceased	delta7_recovered	delta7_tested	delta7_vaccinated1	delta7_vaccinated2	delta7_other
5169	56	2348	63646	NA	NA	NA
605	24	1134	273919	988834	529076	NA
458	NA	408	20082	NA	NA	NA
16966	202	19962	530128	NA	NA	NA
702	6	1393	111907	130197	NA	NA
75	2	170	20276	3590	70366	NA
9753	49	13226	274733	NA	NA	NA
86	NA	NA	3702	NA	NA	NA
1242	9	799	295280	231146	46426	NA
650	3	214	21872	NA	NA	NA
104700	716	90720	889771	667212	823020	10
151	1	122	334410	386320	1505840	NA
36	1	4	5311	288473	11290	NA
25602	577	35423	503317	NA	NA	NA
64528	1680	39610	302872	NA	NA	4
2401	19	2997	293384	NA	NA	NA
13	NA	21	90796	442036	989147	NA
1842	21	2496	42975	63140	514	NA
6171	82	7328	191196	NA	NA	NA
3	NA	10	4006	NA	NA	NA
2144	21	2905	18293	NA	NA	NA
18	NA	NA	764	NA	NA	NA
877	38	925	41790	NA	NA	NA
49	NA	83	12402	818	10348	NA
91642	1467	86048	415629	43724	145603	NA
22	NA	NA	1963	NA	NA	NA
42	NA	56	1024	NA	NA	1
3328	93	4811	515836	NA	NA	NA
87508	904	34408	1244085	574192	525139	NA
1709	12	3438	127767	32379	NA	NA
18988	273	19945	343355	NA	NA	NA
3	NA	13	NA	22083	6190	NA
47	1	59	449953	3734178	1182129	NA
5	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA
54	NA	10	1487	NA	NA	NA
45031	168	19206	298112	541700	239833	NA
1034	6	356	26877	NA	NA	5
NA	NA	NA	NA	NA	NA	NA
1461	2	535	NA	3199	9349	NA
256	NA	5	13321	NA	NA	NA
624	13	11	9088	NA	NA	NA
291382	3281	226137	933431	215160	484149	NA
515	8	299	3559	NA	NA	NA
163	2	144	128317	350930	194752	NA
20010	89	13674	355876	691641	55854	3
64174	884	121247	487806	1679827	76815	NA
4304	40	5128	879608	NA	NA	NA
8208	79	7567	192674	NA	NA	NA
236	NA	128	4205	NA	NA	NA
3782	101	5625	27535	12944	140	NA
135679	2512	142007	519944	419813	148730	NA

delta7_confirmed	delta7_deceased	delta7_recovered	delta7_tested	delta7_vaccinated1	delta7_vaccinated2	delta7_other
3	NA	2	2499	128	1499	NA
35423	111	13392	815424	866522	117883	NA
111	8	116	173627	941734	498355	NA
111137	1295	123086	5512176	NA	NA	40
4	NA	4	NA	7189	23339	NA
937	11	1486	44077	16451	21321	NA
9313	79	8720	208669	NA	NA	NA
14676	94	13601	348157	754711	28356	NA
5560	82	5192	235532	4202570	1381391	NA
388	8	657	6625	NA	NA	NA
37	NA	27	875209	902794	1013026	NA
2	NA	NA	2270	10639	2954	NA
1856	47	1368	128179	20562	3051	NA
1079	3	862	376660	344823	394682	NA
15593	64	12200	247221	NA	NA	1
NA	NA	NA	NA	NA	NA	NA
18190	494	15150	191448	NA	NA	NA
2186	46	3426	740195	2076536	159220	NA
NA	NA	NA	NA	NA	NA	NA
241	10	199	1391	NA	NA	NA
25639	249	21938	373970	NA	NA	NA
460	4	668	34398	12897	28807	NA
NA	NA	NA	NA	NA	NA	NA
5901	72	4910	25372	21441	14128	NA
88694	524	57217	692722	476502	222916	NA
426	40	696	258648	185307	185111	NA
580084	7533	598151	8183750	NA	NA	90
15	NA	11	5954	4985	3263	NA
2863	50	3663	413905	965266	853508	NA
9165	369	2703	36011	NA	NA	NA
54	NA	39	12199	44742	25773	NA
3	NA	2	NA	2224	749	1
3019	19	1046	260593	169056	30688	NA
692	11	719	38415	26745	9548	NA
14998	117	4570	219459	462783	37626	NA
1191	8	368	18916	NA	NA	NA
437	3	385	10403	19540	4318	NA
3202	24	3594	738434	NA	NA	NA
5534	46	7599	163267	NA	NA	NA
NA	NA	NA	330	NA	NA	NA
34115	323	53940	610489	1737293	171653	NA
36972	522	22384	342218	471399	101084	NA
1049	12	587	4193	NA	NA	-2
962	13	1207	21298	6902	36507	NA
213070	702	333416	863113	341236	74495	15
80567	1356	85624	433687	45331	164115	NA

```
# total parsed
# for delta variants
covid_1_parsed_subset[789:885,] %>% select(starts_with('total'))
```

total_confirmed	total_deceased	total_recovered	total_tested	total_vaccinated1	total_vaccinated2	total_other
64420	1043	60023	1272632	505998	89694	16
2837206	34836	2704755	33971945	18564563	3639500	23
639928	8924	603495	35754807	6478775	1162710	NA
19243	184	15460	358887	NA	NA	NA
10668	4	10631	72410	621727	159109	31
10642	4	10563	72410	538592	72646	31
10004827	145171	9549923	160090514	NA	NA	2662
1411	13	714	39133	NA	NA	7
3615	83	2570	219528	NA	NA	NA
1393	1	1092	64478	NA	NA	NA
10678	4	10640	72410	654800	335427	31
20090	202	19626	359314	171445	57372	NA
NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA
1439358	25089	1413943	28608161	12725811	6830497	NA
13651	143	13356	310487	NA	NA	NA
450	6	389	7938	NA	NA	NA
622851	10453	605685	8351048	NA	NA	NA
54525	734	46186	814616	NA	NA	284
718711	23089	522427	3798306	NA	NA	322
239	1	58	10986	NA	NA	NA
1958	NA	1548	76976	NA	NA	NA
45697	729	37029	577386	NA	NA	44
1708208	22750	1684601	64277972	37897452	7335881	NA
4469488	23296	4256697	33944832	23623801	9672550	528
1893	43	207	22283	NA	NA	NA
6283	9	3959	336091	NA	NA	3
140471	453	110883	2702280	NA	NA	3
2	NA	1	311	NA	NA	NA
316	NA	233	27527	2343	621	3
69	NA	NA	888	746	NA	NA
341772	1974	337430	9103948	2309764	389688	NA
65087	813	64236	665644	793611	294231	NA
1263	1	904	57753	NA	NA	NA
47	NA	NA	NA	NA	NA	NA
55676	803	54169	509776	48076	8767	NA
NA	NA	NA	NA	NA	NA	NA
52633	599	43506	447896	195168	69769	NA
334780	1959	331667	7635887	194058	NA	NA
34231243	456418	33606777	604498405	723497151	317002722	13182
10662	4	10623	72410	607359	109613	31
428	2	201	23217	NA	NA	1
10645	4	10568	72410	542315	73151	31
862804	11541	825141	9568625	NA	NA	19
31725455	425227	30888805	471294789	372626926	105917188	12559
NA	NA	NA	NA	NA	NA	NA
3390	2	3353	72410	8915	1090	30
348	4	6	1800	NA	NA	NA
208389	938	200381	4811501	NA	NA	3
119041	1836	114991	1278530	1216639	429011	NA
1709954	22896	1686917	80311528	90815820	24272961	NA
NA	NA	NA	NA	NA	NA	NA

total_confirmed	total_deceased	total_recovered	total_tested	total_vaccinated1	total_vaccinated2	total_other
27586	552	24910	259295	608661	145059	774
7564	129	7430	486944	254459	104851	NA
41	NA	12	6536	NA	NA	NA
716	4	102	26951	NA	NA	3
1028819	8276	1016165	20295168	22932463	9039027	NA
119324	1083	117778	5352653	206182	2209	NA
6609906	140196	6449186	62559171	67145633	30943704	3619
2947255	37278	2891193	43194662	31045670	10001488	26
12099	88	11810	125075	3998	NA	147
596550	16122	578310	11107570	6639425	1234983	NA
1723135	45325	1577322	9482940	NA	NA	924
7560	129	7428	484869	251619	104010	NA
1083531	4343	1043473	12060313	1056499	232717	301
33380535	444278	32590885	549229149	582606905	189818839	12925
3147	105	1587	23388	NA	NA	NA
10392	183	7135	101732	164479	57477	208
3271530	16035	3114716	26248280	12972163	5622271	507
5665	66	5467	363056	87865	9527	NA
1005872	13572	992066	13584411	14550032	6865942	NA
30630	639	28439	345892	667893	255765	934
139985	2228	114793	769184	362946	94502	NA
117249	1749	113146	1316296	440326	67068	NA
1330	1	1062	60199	NA	NA	NA
3314	54	1078	43370	NA	NA	NA
2980170	37866	2930867	48554234	40038832	18567279	29
19345	314	18686	174395	NA	NA	NA
931997	12166	911232	15983473	13594	NA	19
NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA
34735	391	33457	691184	856929	257028	57
126737	3980	116165	2274772	NA	NA	NA
104	NA	1	3152	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA
910	1	510	31193	NA	NA	NA
770915	9874	760755	12487626	16840564	6559654	NA
374277	5366	307611	9145828	NA	NA	NA
73238	2697	56516	1058881	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA
322642	1847	317870	6281162	NA	NA	NA
2117	3	1430	104833	NA	NA	NA
261766	3864	255117	5786018	650684	160632	NA
115529	1035	112893	4838094	NA	NA	NA
204224	3488	199582	2648371	3498848	1001930	18
331062	4426	325793	15675779	9404423	4335339	NA
179	2	36	3663	NA	NA	NA

```r
covid_2_parsed <-
  covid_data_2 %>% enframe() %>% unnest_wider(value) %>%
  unnest_wider(c(delta, delta21_14, delta7, total), names_sep = "_") %>% select(-c(districts, meta))

# for delta
covid_2_parsed %>% select(starts_with('delta'))
```

```
## # A tibble: 37 x 15
##    delta_tested delta_vaccinated1 delta_vaccinated2 delta_confirmed
##           <int>             <int>             <int>           <int>
##  1         1376                 3                13              NA
##  2        39848             20497             24137             385
##  3          334                42               195               1
##  4        15060             19124             37463             212
##  5       226443            114694            145827               8
##  6         1403               211              1282               5
##  7        11869             21312             39393              32
##  8        56751             12482             11839              45
##  9           NA                 3                20              NA
## 10         2361              2572             12404              23
## # i 27 more rows
## # i 11 more variables: delta_deceased <int>, delta_recovered <int>,
## #   delta_other <int>, delta21_14_confirmed <int>, delta7_confirmed <int>,
## #   delta7_recovered <int>, delta7_tested <int>, delta7_vaccinated1 <int>,
## #   delta7_vaccinated2 <int>, delta7_deceased <int>, delta7_other <int>
```

```
# for delta7
covid_2_parsed %>% select(starts_with('delta7'))
```

```
## # A tibble: 37 x 7
##    delta7_confirmed delta7_recovered delta7_tested delta7_vaccinated1
##               <int>            <int>         <int>              <int>
##  1                3                5          8936                884
##  2             2873             3590        254532            1223010
##  3               66               97          4788               3312
##  4             2056             2215        269097             274869
##  5               40               31       1378539            1286708
##  6               28               20         10726               3680
##  7              205              103        147451             379374
##  8              267              239        395086             160323
##  9               NA                2            NA               2802
## 10              222              409         19026               8418
## # i 27 more rows
## # i 3 more variables: delta7_vaccinated2 <int>, delta7_deceased <int>,
## #   delta7_other <int>
```

```
# for delta21-_14
covid_2_parsed %>% select(starts_with('delta2'))
```

```
## # A tibble: 37 x 1
##    delta21_14_confirmed
##                   <int>
##  1                    9
##  2                 3220
##  3                   87
##  4                 1499
##  5                   30
##  6                   23
##  7                  124
##  8                  195
##  9                    4
## 10                  409
## # i 27 more rows
```

```
# for total
covid_2_parsed %>% select(starts_with('total'))
```

```
## # A tibble: 37 x 7
##    total_confirmed total_deceased total_recovered total_tested total_vaccinated1
##              <int>          <int>           <int>        <int>             <int>
##  1            7651            129            7518       598033            294001
##  2         2066450          14373         2047722     29518787          32976969
##  3           55155            280           54774      1185436            771875
##  4          610645           5997          600974     24712042          20172463
##  5          726098           9661          716390     50531824          49874828
##  6           65351            820           64495       792851            926035
##  7         1006052          13577          992159     13709510          14851682
##  8         1439870          25091         1414431     29427753          13055636
##  9           10681              4           10644        72410            660753
## 10          178108           3364          174392      1468399           1262568
## # i 27 more rows
## # i 2 more variables: total_vaccinated2 <int>, total_other <int>
```

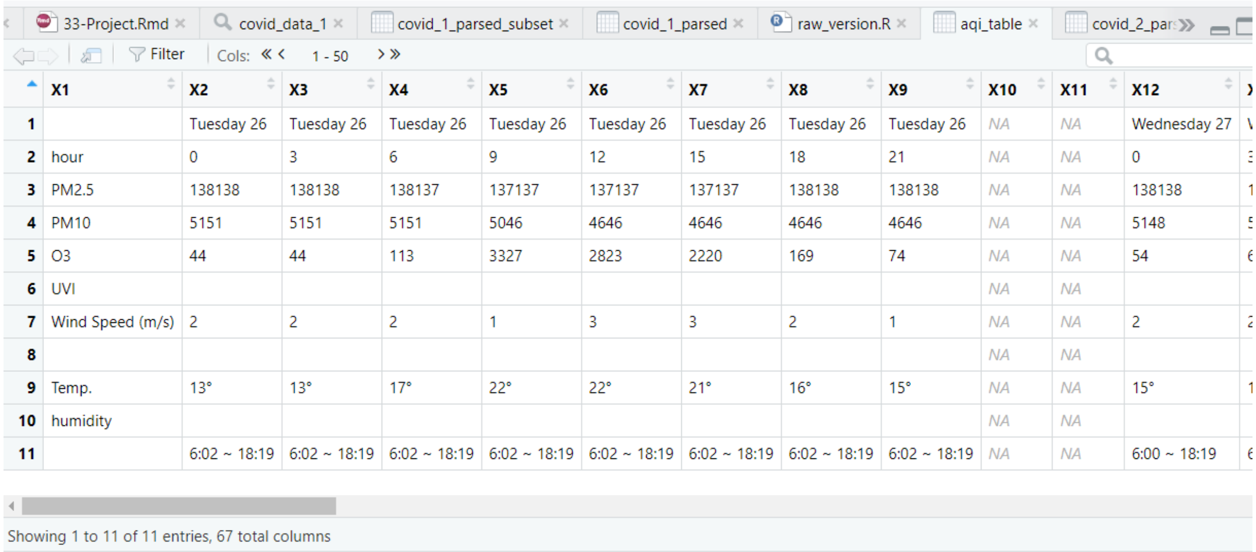
```
# merge into single file
merged_df <- merge(covid_1_parsed_subset,
    covid_2_parsed,
    by.x = "name",
    by.y = "name",
    sort = T,
    all = F)

head(merged_df)
```

```
##   name       date delta_confirmed.x delta_deceased.x delta_recovered.x
## 1   AN 2020-10-25                20               NA                15
## 2   AN 2021-04-25                51               NA                57
## 3   AN 2021-08-23                 2               NA                NA
## 4   AN 2020-07-29                65                1                 5
## 5   AN 2020-11-19                11               NA                20
## 6   AN 2021-07-31                 2               NA                 2
##   delta_tested.x delta_vaccinated1.x delta_vaccinated2.x delta_other.x
## 1            746                  NA                  NA            NA
## 2           1462                1055                 228            NA
## 3           1862                3012                1104            NA
## 4            292                  NA                  NA            NA
## 5           1626                  NA                  NA            NA
## 6           1231                3665                 697            NA
##   delta7_confirmed.x delta7_deceased.x delta7_recovered.x delta7_tested.x
## 1                137                 2                115            7596
## 2                275                 2                267           10970
## 3                 10                NA                  1           13244
## 4                207                 2                 38            2174
## 5                 97                NA                116            9193
## 6                 12                NA                 18            8243
##   delta7_vaccinated1.x delta7_vaccinated2.x delta7_other.x total_confirmed.x
## 1                   NA                   NA             NA              4245
## 2                16508                 2648             NA              5665
## 3                 7028                 3462             NA              7559
## 4                   NA                   NA             NA               428
## 5                   NA                   NA             NA              4604
## 6                16978                 5085             NA              7537
##   total_deceased.x total_recovered.x total_tested.x total_vaccinated1.x
## 1               58              3983          82626                  NA
## 2               66              5467         363056               87865
## 3              129              7420         474665              241644
## 4                2               201          23217                  NA
## 5               61              4398         112792                  NA
## 6              129              7400         440870              209696
##   total_vaccinated2.x total_other.x delta_tested.y delta_vaccinated1.y
## 1                  NA            NA           1376                   3
## 2                9527            NA           1376                   3
## 3              101276            NA           1376                   3
## 4                  NA             1           1376                   3
## 5                  NA            NA           1376                   3
## 6               91562            NA           1376                   3
##   delta_vaccinated2.y delta_confirmed.y delta_deceased.y delta_recovered.y
## 1                  13                NA               NA                NA
## 2                  13                NA               NA                NA
## 3                  13                NA               NA                NA
## 4                  13                NA               NA                NA
## 5                  13                NA               NA                NA
## 6                  13                NA               NA                NA
##   delta_other.y delta21_14_confirmed delta7_confirmed.y delta7_recovered.y
## 1            NA                    9                  3                  5
## 2            NA                    9                  3                  5
## 3            NA                    9                  3                  5
## 4            NA                    9                  3                  5
## 5            NA                    9                  3                  5
## 6            NA                    9                  3                  5
##   delta7_tested.y delta7_vaccinated1.y delta7_vaccinated2.y delta7_deceased.y
## 1            8936                  884                10640                NA
## 2            8936                  884                10640                NA
## 3            8936                  884                10640                NA
## 4            8936                  884                10640                NA
## 5            8936                  884                10640                NA
## 6            8936                  884                10640                NA
##   delta7_other.y total_confirmed.y total_deceased.y total_recovered.y
## 1             NA              7651              129              7518
## 2             NA              7651              129              7518
## 3             NA              7651              129              7518
## 4             NA              7651              129              7518
## 5             NA              7651              129              7518
## 6             NA              7651              129              7518
##   total_tested.y total_vaccinated1.y total_vaccinated2.y total_other.y
## 1         598033              294001              200157            NA
## 2         598033              294001              200157            NA
## 3         598033              294001              200157            NA
## 4          598033             294001             200157           NA
## 5          598033             294001             200157           NA
## 6          598033             294001             200157           NA
```

Task 2

Webscraping of Dynamic Table AQI Kathmandu

```
# load the webdriver for firefox
rD <- rsDriver(browser="firefox",verbose = F, port = 14421L)
remDr <- rD[["client"]]
remDr$navigate("https://aqicn.org/forecast/kathmandu/")
aqi_html  <- read_html(remDr$getPageSource() %>% unlist())

# scrape the needed table for data analysis
aqi_html %>% html_element(".forecast-body-table") %>%
  html_nodes("table") %>%
  html_table() ->
  forecast_table


# since forecast consist the list of dataframe
# extracted first value from the list which conists the required dataframe
aqi_table <- forecast_table %>% .[[1]]

knitr::include_graphics('aqi.png')
```



	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
1		Tuesday 26	Tuesday 26	Tuesday 26	Tuesday 26	Tuesday 26	Tuesday 26	Tuesday 26	Tuesday 26	NA	NA	Wednesday 27
2	hour	0	3	6	9	12	15	18	21	NA	NA	0
3	PM2.5	138138	138138	138137	137137	137137	137137	138138	138138	NA	NA	138138
4	PM10	5151	5151	5151	5046	4646	4646	4646	4646	NA	NA	5148
5	O3	44	44	113	3327	2823	2220	169	74	NA	NA	54
6	UVI									NA	NA	
7	Wind Speed (m/s)	2	2	2	1	3	3	2	1	NA	NA	2
8										NA	NA	
9	Temp.	13°	13°	17°	22°	22°	21°	16°	15°	NA	NA	15°
10	humidity									NA	NA	
11		6:02 ~ 18:19	6:02 ~ 18:19	6:02 ~ 18:19	6:02 ~ 18:19	6:02 ~ 18:19	6:02 ~ 18:19	6:02 ~ 18:19	6:02 ~ 18:19	NA	NA	6:00 ~ 18:19

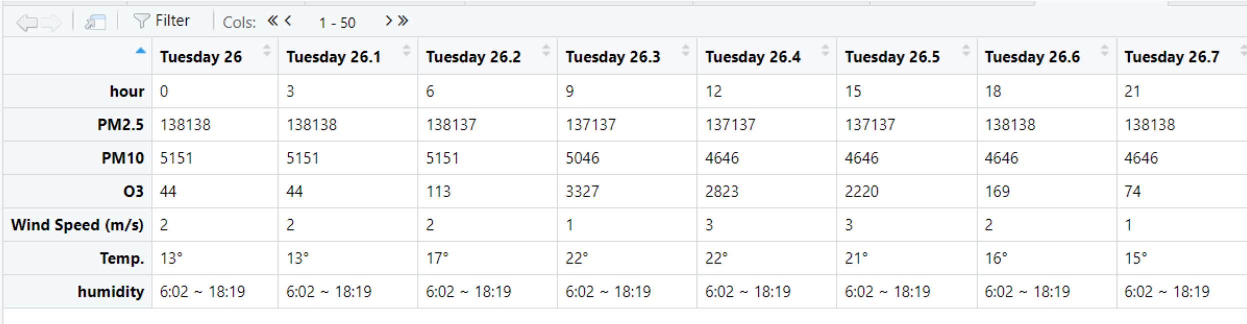
Showing 1 to 11 of 11 entries, 67 total columns

Data Wrangling the forecast Table AQI Kathmandu

- **Step 1**: Remove null columns that came while parsing `aqi_table %>%  select(-c('X10','X11','X20','X21','X30','X31','X40','X41','X50','X51','X60','X61'))`
- **Step 2**: Filter out rows with ‘UVI’ and ‘humidity’ in the ‘X1’ column where entire row is NULL `aqi_table %>% filter(X1 != 'UVI')` and `aqi_table %>% filter(X1 != 'humidity')`

- **Step 3**: Replaced the value in the 'X1' column at index 9 with 'humidity' `aqi_table %>% mutate(X1 = replace(X1, 9, "humidity"))`
- **Step 4**: Replaced the value in the 'X1' column at index 1 with 'Index' `aqi_table %>% mutate(X1 = replace(X1, 1, "Index"))`
- **Step 5**: Filter out rows with empty values in the 'X1' column `aqi_table %>% filter(X1 != '')`
- **Step 6**: Assigned the first row of the data frame as the column headers `headers <-  aqi_table[1,]` `colnames(aqi_table) <- headers`
- **Step 7**: Remove the first row of the data frame `aqi_table <- aqi_table[-1,]`
- **Step 8**: Converted the 'Index' column to row names `aqi_table %>% column_to_rownames(var = 'Index')`

```
# extract first value from list
aqi_table <- forecast_table %>% .[[1]]

# delete null columns
aqi_table <- aqi_table %>%
  select(-c('X10','X11','X20','X21','X30','X31','X40','X41','X50','X51','X60','X61'))

# remove null row 'UVI'
aqi_table <- aqi_table %>% filter(X1 != 'UVI')

# since value of humidity interchange for now I have removed empty row.
aqi_table <- aqi_table %>% filter(X1 != 'humidity')

# now I have assigned the value at 1st column 9th row as 'humidity'
aqi_table <- aqi_table %>% mutate(X1 = replace(X1, 9, "humidity"))

# now I have assigned the value at 1st column 1st row as 'Index'
aqi_table <- aqi_table %>% mutate(X1 = replace(X1, 1, "Index"))

# finally remove the last empty row
aqi_table <- aqi_table %>% filter(X1 != '')

# setting first row as headers
headers <- aqi_table[1,]
colnames(aqi_table) <- headers

# dropping the first row as header has been set.
aqi_table <- aqi_table[-1,]

# now setting the index or row name as 'Index' column
aqi_table <- aqi_table %>% column_to_rownames(var = 'Index')

knitr::include_graphics('aqi_parse.png')
```



	Tuesday 26	Tuesday 26.1	Tuesday 26.2	Tuesday 26.3	Tuesday 26.4	Tuesday 26.5	Tuesday 26.6	Tuesday 26.7
hour	0	3	6	9	12	15	18	21
PM2.5	138138	138138	138137	137137	137137	137137	138138	138138
PM10	5151	5151	5151	5046	4646	4646	4646	4646
O3	44	44	113	3327	2823	2220	169	74
Wind Speed (m/s)	2	2	2	1	3	3	2	1
Temp.	13°	13°	17°	22°	22°	21°	16°	15°
humidity	6:02 ~ 18:19	6:02 ~ 18:19	6:02 ~ 18:19	6:02 ~ 18:19	6:02 ~ 18:19	6:02 ~ 18:19	6:02 ~ 18:19	6:02 ~ 18:19

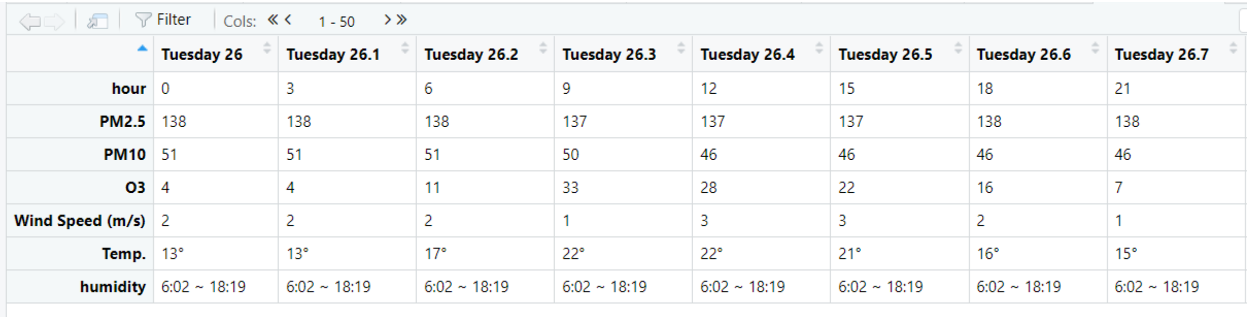
Sample Parsed AQI Table

Note Still some changes needs to be done on `hour`, `PM2.5`, `PM10` and `O3` as values are concatenated wrongly while parsing html.

```
aqi_table[2,] <-
  floor(as.integer(str_extract(as.character(aqi_table[2,]), "\\d+")) / 1000)
aqi_table[3,] <-
  floor(as.integer(str_extract(as.character(aqi_table[3,]), "\\d+")) / 100)

lengths <- as.numeric(nchar(aqi_table[4,]))
aqi_table[4,] <-
  ifelse(lengths == 2, substr(aqi_table[4,], 1, 1),
  ifelse(lengths %in% 3:4, substr(aqi_table[4,], 1, 2), ""))

aqi_table
```

```
##                     Tuesday 26   Tuesday 26   Tuesday 26   Tuesday 26
## hour                         0            3            6            9
## PM2.5                      138          138          138          137
## PM10                        51           51           51           50
## O3                           4            4           11           33
## Wind Speed (m/s)             2            2            2            1
## Temp.                      13°          13°          17°          22°
## humidity          6:02 ~ 18:19 6:02 ~ 18:19 6:02 ~ 18:19 6:02 ~ 18:19
##                     Tuesday 26   Tuesday 26   Tuesday 26   Tuesday 26
## hour                        12           15           18           21
## PM2.5                      137          137          138          138
## PM10                        46           46           46           46
## O3                          28           22           16            7
## Wind Speed (m/s)             3            3            2            1
## Temp.                      22°          21°          16°          15°
## humidity          6:02 ~ 18:19 6:02 ~ 18:19 6:02 ~ 18:19 6:02 ~ 18:19
##                   Wednesday 27 Wednesday 27 Wednesday 27 Wednesday 27
## hour                         0            3            6            9
## PM2.5                      138          151          151          147
## PM10                        51           51           51           51
## O3                           5            6            9           28
## Wind Speed (m/s)             2            2            1            3
## Temp.                      15°          14°          18°          23°
## humidity          6:00 ~ 18:19 6:00 ~ 18:19 6:00 ~ 18:19 6:00 ~ 18:19
##                   Wednesday 27 Wednesday 27 Wednesday 27 Wednesday 27
## hour                        12           15           18           21
## PM2.5                      138          138          138          138
## PM10                       50           46           46           46
## O3                         26           20           17           11
## Wind Speed (m/s)            3            2            1            2
## Temp.                     23°          20°          16°          16°
## humidity         6:00 ~ 18:19 6:00 ~ 18:19 6:00 ~ 18:19 6:00 ~ 18:19
##                   Thursday 28  Thursday 28  Thursday 28  Thursday 28
## hour                        0            3            6            9
## PM2.5                     138          138          125          137
## PM10                       46           46           46           46
## O3                          4            3            4           28
## Wind Speed (m/s)            2            1            1            2
## Temp.                     14°          14°          18°          23°
## humidity         5:59 ~ 18:20 5:59 ~ 18:20 5:59 ~ 18:20 5:59 ~ 18:20
##                   Thursday 28  Thursday 28  Thursday 28  Thursday 28
## hour                       12           15           18           21
## PM2.5                     137          138          138          138
## PM10                       46           46           46           46
## O3                         29           23           16            5
## Wind Speed (m/s)            2            2            1            2
## Temp.                     23°          20°          17°          18°
## humidity         5:59 ~ 18:20 5:59 ~ 18:20 5:59 ~ 18:20 5:59 ~ 18:20
##                     Friday 29    Friday 29    Friday 29    Friday 29
## hour                        0            3            6            9
## PM2.5                     138          138          138          138
## PM10                       46           46           46           46
## O3                          4            3            4           23
## Wind Speed (m/s)            2            1            1            1
## Temp.                     16°          16°          19°          21°
## humidity         5:58 ~ 18:20 5:58 ~ 18:20 5:58 ~ 18:20 5:58 ~ 18:20
##                     Friday 29    Friday 29    Friday 29    Friday 29
## hour                       12           15           18           21
## PM2.5                     138          138          138          138
## PM10                       46           46           46           46
## O3                         24           22           16            7
## Wind Speed (m/s)            1            1            1            2
## Temp.                     21°          21°          18°          16°
## humidity         5:58 ~ 18:20 5:58 ~ 18:20 5:58 ~ 18:20 5:58 ~ 18:20
##                   Saturday 30  Saturday 30  Saturday 30  Saturday 30
## hour                        0            3            6            9
## PM2.5                     138          137          137          137
## PM10                       46           51           51           50
## O3                          4            5           13           27
## Wind Speed (m/s)            2            1            1            2
## Temp.                     17°          16°          20°          26°
## humidity         5:57 ~ 18:21 5:57 ~ 18:21 5:57 ~ 18:21 5:57 ~ 18:21
##                   Saturday 30  Saturday 30  Saturday 30  Saturday 30
## hour                       12           15           18           21
## PM2.5                     137          138          138          138
## PM10                       46           46           46           46
## O3                         22           18           14            5
## Wind Speed (m/s)            4            4            4            3
## Temp.                     28°          26°          16°          14°
## humidity         5:57 ~ 18:21 5:57 ~ 18:21 5:57 ~ 18:21 5:57 ~ 18:21
##                     Sunday 31    Sunday 31    Sunday 31    Sunday 31
## hour                        0            3            9           12
## PM2.5                     138          125          103          103
## PM10                       46           46           46           46
## O3                          5            5
## Wind Speed (m/s)            2            2            4            5
## Temp.                     16°          16°          28°          28°
## humidity         5:56 ~ 18:21 5:56 ~ 18:21 5:56 ~ 18:21 5:56 ~ 18:21
##                     Sunday 31    Sunday 31    Sunday 31   NA     Monday 1
## hour                       15           18           21 <NA>            3
## PM2.5                      10            8            8 <NA>            7
## PM10                       46           46           34 <NA>           34
## O3                                                          N
## Wind Speed (m/s)            1            3            1 <NA>            2
## Temp.                     24°          18°          18° <NA>          16°
## humidity         5:56 ~ 18:21 5:56 ~ 18:21 5:56 ~ 18:21 <NA> 5:55 ~ 18:22
##                      Monday 1     Monday 1     Monday 1     Monday 1
## hour                        6            9           12           15
## PM2.5                       6            7            8            8
## PM10                       32           42           50           51
## O3
## Wind Speed (m/s)            2            2            4            2
## Temp.                     24°          29°          29°          25°
## humidity         5:55 ~ 18:22 5:55 ~ 18:22 5:55 ~ 18:22 5:55 ~ 18:22
##                      Monday 1
## hour                       18
## PM2.5                       8
## PM10                       51
## O3
## Wind Speed (m/s)            2
## Temp.                     19°
## humidity         5:55 ~ 18:22
```

```r
knitr::include_graphics('aqi_final.png')
```



	Tuesday 26	Tuesday 26.1	Tuesday 26.2	Tuesday 26.3	Tuesday 26.4	Tuesday 26.5	Tuesday 26.6	Tuesday 26.7
hour	0	3	6	9	12	15	18	21
PM2.5	138	138	138	137	137	137	138	138
PM10	51	51	51	50	46	46	46	46
O3	4	4	11	33	28	22	16	7
Wind Speed (m/s)	2	2	2	1	3	3	2	1
Temp.	13°	13°	17°	22°	22°	21°	16°	15°
humidity	6:02 ~ 18:19	6:02 ~ 18:19	6:02 ~ 18:19	6:02 ~ 18:19	6:02 ~ 18:19	6:02 ~ 18:19	6:02 ~ 18:19	6:02 ~ 18:19

Final Version of Cleaned AQI Table

Part 3

Load the necessary library needed for Part 3

```
library(pdftools) # for working with pdf files
library(tm) # for text mining
library(wordcloud) # plotting word cloud
library(Rgraphviz) # plotting network like graph for word association
library(graph) # plotting network like graph for word association
library(ggplot2) # for bargraph
```

Load the pdf files and convert it to Corpus

```
# load the file path in list
files <- list.files(pattern = "pdf$")

# load the pdf files into list
pdf_files <- lapply(files, pdf_text)

# create a corpus from vector source i.e from list pdf_files
corpus <- Corpus(VectorSource(unlist(pdf_files)))
# copy the loaded corpus
corpus_copy <- corpus

# inspect first few texts of corpus
inspect(corpus[1:2])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 2
##
## [1] See discussions, stats, and author profiles for this publication at: https://www.researchgate.net
## [2]  Overview\n\n\n\nData mining in education\nCristobal Romero* and Sebastian Ventura\n\n
27 doi: 10.1002/widm.1075\n\n\n\n\nINTRODUCTION                                                     tio
```

Text Mining Preprocessing

- Step 1: Convert all texts to lowercase
- Step 2: Remove punctuation
- Step 3: Remove numbers
- Step 4: Remove stop words or user defined stop words
- Step 5: Stem the corpus
- Step 6: Remove specific words which doesn't help the corpus
- Step 7: Create Term Document Matrix

```
# convert the all texts in lower
corpus <- tm_map(corpus, tolower)
```

```
## Warning in tm_map.SimpleCorpus(corpus, tolower): transformation drops documents
```

```
inspect(corpus[1:2])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 2
##
## [1] see discussions, stats, and author profiles for this publication at: https://www.researchgate.ne
## [2]  overview\n\n\n\ndata mining in education\ncristobal romero* and sebastian ventura\n\n
27 doi: 10.1002/widm.1075\n\n\n\n\nintroduction                                                     tio
```

```
# remove punctuations
corpus <- tm_map(corpus, removePunctuation)
```

```
## Warning in tm_map.SimpleCorpus(corpus, removePunctuation): transformation drops
## documents
```

```
# stem the corpus
corpus <- tm_map(corpus, stemDocument)
```

```
## Warning in tm_map.SimpleCorpus(corpus, stemDocument): transformation drops
## documents
```

```
remove <- function(x) gsub("values","value",x)
corpus <-  tm_map(corpus, remove)
```

```
## Warning in tm_map.SimpleCorpus(corpus, remove): transformation drops documents
```

```
# create Term Document Matrix with word length 1 or many
tdm <- TermDocumentMatrix(corpus, control = list((wordLenghts=c(1,Inf))))
```

Best way to create a Term Document Matrix with preprocessing

```
remove <- function(x) gsub("values","value",x)
corpus_copy <-  tm_map(corpus_copy, remove)
```

```
## Warning in tm_map.SimpleCorpus(corpus_copy, remove): transformation drops
## documents
```

```
my_tdm <- TermDocumentMatrix(
  unlist(corpus_copy),
  control =
    list(
      removePunctuation = TRUE,
      stopwords = TRUE,
      tolower = TRUE,
      stemming = FALSE,
      removeNumbers = TRUE,
      bounds = list(global = c(3, Inf)),
      wordLenghts = c(1,Inf),
      removeWords = (c("can","may","used")))
)
```

Most Frequent Terms

```
# finding frequency of words which is at least present 10 times
low_frequent_terms <- findFreqTerms(my_tdm, lowfreq = 10)
head(low_frequent_terms)
```

```
## [1] "article"   "author"    "authors"   "content"   "data"      "discovery"
```

```
# finding frequency of words which is at max present 10 times
high_frequent_terms <- findFreqTerms(my_tdm, highfreq = 10)
head(high_frequent_terms)
```

```
## [1] "cordoba"           "downloaded"        "interdisciplinary"
## [4] "profile"           "profiles"          "publication"
```

Associated terms of the most frequent term

```
# associated terms for mining with correlation 0.3
findAssocs(my_tdm, "mining", 0.3)
```

```
## $mining
##            data       knowledge       databases       discovery         systems
##            0.55            0.54            0.49            0.45            0.45
##        database           kinds        patterns            user           mined
##            0.44            0.42            0.40            0.39            0.39
##     interactive           users        research        analysis     association
##            0.37            0.37            0.36            0.35            0.34
## interestingness           erent       retrieval           rules      multimedia
##            0.33            0.32            0.31            0.31            0.31
##      challenges      techniques
##            0.30            0.30
```

```
# associated terms for mining with learning 0.3
findAssocs(my_tdm, "learning", 0.35)
```

```
## $learning
##       machine  intelligence          arti          cial           vol
##          0.74          0.56          0.52          0.50          0.43
##       shavlik        morgan      kaufmann     michalski    statistics
##          0.43          0.42          0.41          0.41          0.40
##        expert      mitchell         ijcai international      learners
##          0.40          0.40          0.39          0.38          0.38
##       quinlan  decisiontree  bibliography     carbonell        kluwer
##          0.38          0.37          0.36          0.36          0.36
##         neter         mateo
##          0.35          0.35
```

```
# associated terms for mining with data 0.3
findAssocs(my_tdm, "data", 0.4)
```

```
## $data
##      mining    cleaning integration   warehouse  warehouses
##        0.55        0.43        0.42        0.42        0.41
```

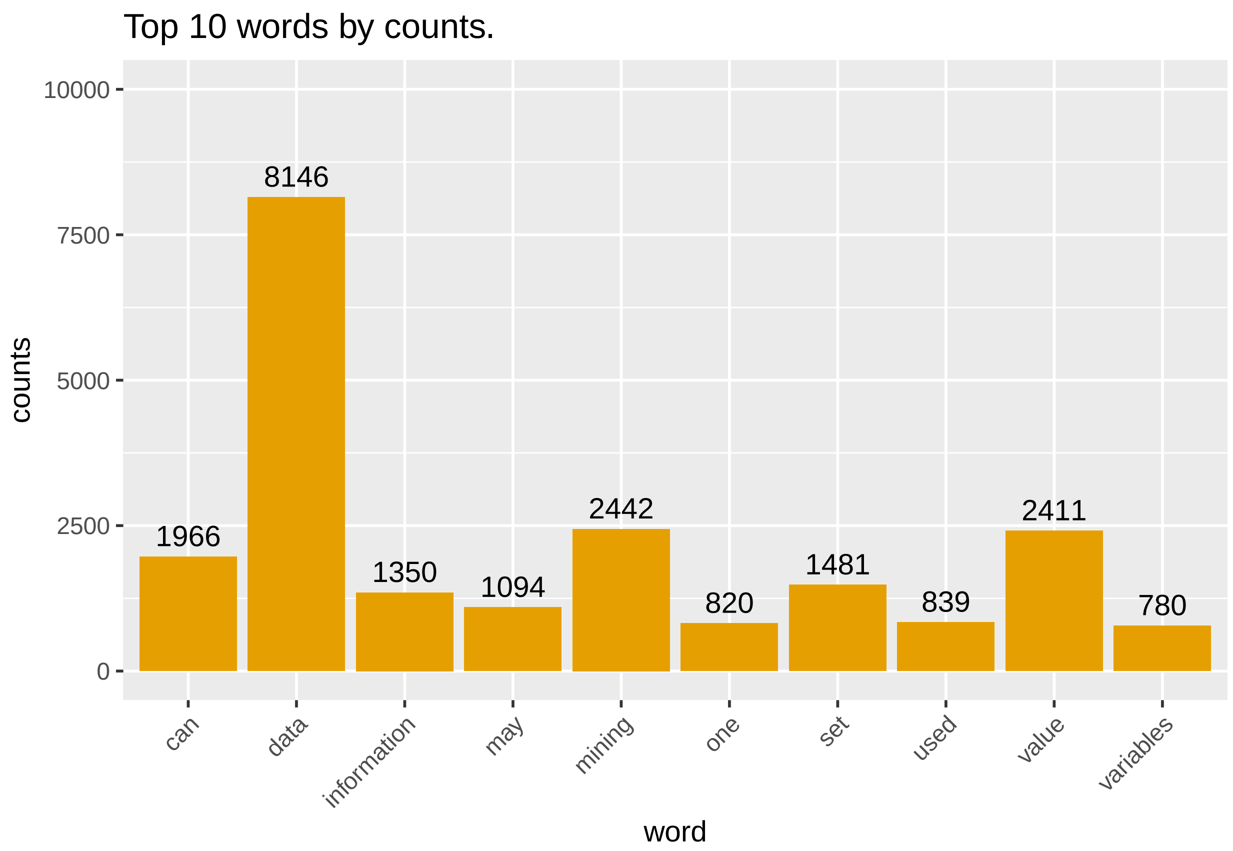
Top 10 words and their respective counts

```
# top 10 words and their respective counts

df <-
  my_tdm %>%
  as.matrix() %>%
  rowSums() %>%
  sort(decreasing = TRUE) %>%
  head(10) %>%
  enframe(name = "word", value = "counts")

df
```

```
## # A tibble: 10 x 2
##    word        counts
##    <chr>        <dbl>
##  1 data          8146
##  2 mining        2442
##  3 value         2411
##  4 can           1966
##  5 set           1481
##  6 information   1350
##  7 may           1094
##  8 used           839
##  9 one            820
## 10 variables      780
```

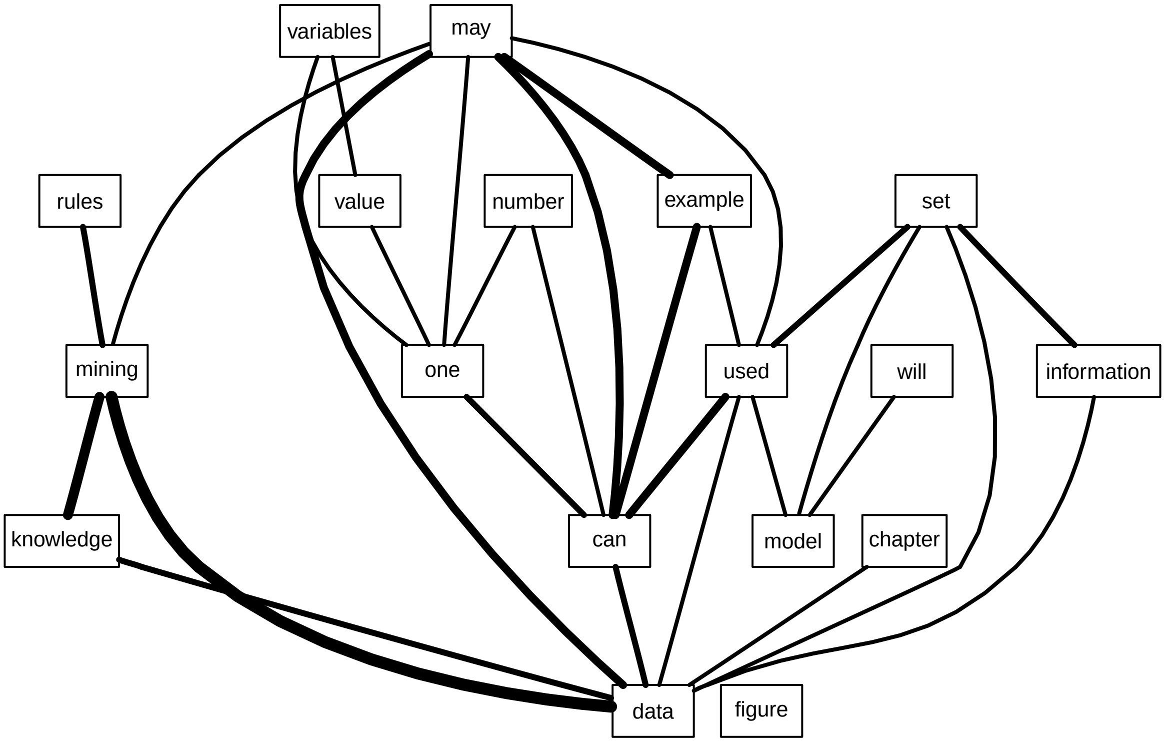
Bargraph of top 10 words and their respective counts

```
# using ggplot2
bargraph <- ggplot(df, aes(word, counts)) +
  geom_bar(stat = "identity", fill = "#E69F00") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Top 10 words by counts.") +
  geom_text(aes(label = counts), vjust = -0.5) +
  ylim(0, 10000L)

bargraph
```



Correlation between top 650 frequent terms

```
top_650_frequent_tems <- findFreqTerms(my_tdm, lowfreq = 650)
plot(my_tdm, terms = top_650_frequent_tems, corThreshold = 0.2, weighting = T)
```



Topic Models

```
# topic models
library(topicmodels)

## Warning: package 'topicmodels' was built under R version 4.3.3

set.seed(123)

lda <- LDA(my_tdm, k=4)

# terms
head(terms(lda,3))

##      Topic 1      Topic 2      Topic 3      Topic 4
## [1,] "content173" "content711" "content15"  "content945"
## [2,] "content171" "content676" "content16"  "content684"
## [3,] "content473" "content713" "content669" "content791"

# topics
head(topics(lda))

## article  author authors content cordoba    data
##       3       3       3       1       3       2
```