MPs Parliament Attendance Data collation

true

Y

```
library(dplyr)
library(magrittr)
```

16th Loksabha Sessions

Session dates and attendance of all members taken from the official website.

```
## return the number of weekdays between dates d1 and d2}
num_weekdays <- function(d1, d2){
    weekend  <- c("Saturday", "Sunday")
    all_days <- seq(d1, d2, "1 day") %>% weekdays
    sum(!(all_days %in% weekend))
}

num_weekdays_v <-Vectorize(num_weekdays)
```

I copied the session dates into a csv file. Two sessions (IV and XIV sessions) had a break. We calculate the number of weekdays in each session.

```
session_dates <- here::here("data/16th", "loksabha_session_dates.csv")
session_dates <- readr::read_csv(session_dates)
```

```
## Parsed with column specification:
## cols(
##   session = col_character(),
##   start = col_character(),
##   end = col_character()
## )
```

```
session_dates$session %<>% as.roman %>% as.integer

session_dates$start %<>% lubridate::dmy(.)
session_dates$end   %<>% lubridate::dmy(.)

session_dates$session_length <- num_weekdays_v(session_dates$start,
                                               session_dates$end)

sessions_durations <- session_dates %>%
                      group_by(session) %>%
                      summarise_if(is.numeric, sum)
```

The attendance information is unfortunately made available in doc format. There were 14 files (one

for each session). I converted each doc file to docx and used the docxtractr package to extract the tables. I know I've to extract the 4th table from each document because I had queried the table structure using the $docx_describe_tbls$ from the package.

```
infile     <- here::here("data/16th", paste0(1:14, ".docx"))
attendance <-lapply(infile, function(fname){
                                docxtractr::read_docx(fname) %>%
                                docxtractr::docx_extract_tbl(4)})
```

Stacking together the information from all sessions, I get a usable data frame after having cleaned up the column names and fixed column classes.

```
names(attendance) <- 1:14
all_sessions <- bind_rows(attendance, .id = "session")

all_sessions %<>% janitor::clean_names()
all_sessions$session %<>% as.integer
all_sessions$no_of_days_member_signed_the_register %<>% as.integer
```

Some of the column names are really clunky, so lets fix that.

```
all_sessions %<>% dplyr::rename(mp_name = name_of_member,
                                days_attended =
                                no_of_days_member_signed_the_register)
```

Finally, append the length of the session to this data frame and add a column for percentage attendance.

```
all_sessions %<>% left_join(sessions_durations)
```

```
## Joining, by = "session"
```

```
all_sessions %<>%
    mutate(percent_attendance = 100 * days_attended /session_length)
```

Features

What we have is a bare bones data set that records the name of each MP, their constituency and the number of days they attended the parliament (or at least, signed the register). I will now add some features such as gender, education etc. that I care about to this information. The source for this information is PRS Legislative Research, specifically this. While the data is well-formatted and this file also includes percent attendance, I chose not to use it as it was not clear how this percent has been calculated i.e., what was the denominator in each case. It was also becoming a tad annoying to extract the dates hidden in the notes.

```
infile <- here::here("data", "MPTrack-16.xls")

mps_attributes <- readxl::read_excel(infile) %>%
                  janitor::clean_names()
```

We can use skimr to take a quick look at the data to see if there is anything missing. Unfortunately the output of skimr doesn't play well with tex.

Data clean-up

Fix the classes for some columns. For instance, the number of term and the political party should be a factor. Also, get rid of some messy columns we won't use.

```
mps_attributes %<>% select(-notes)

term_levels <-c("First", "Second", "Third", "Fourth","Fifth",
                "Sixth", "Seventh", "Eighth", "Ninth")
mps_attributes$no_of_term %<>% factor(term_levels)

mps_attributes$political_party %<>% factor
```

The really tricky bit is this - we now have to join this set of attributes with the data on attendance using name and/or constituency. There are so many ways in which this can go wrong - variation on spelling, using or not using titles etc. We will do the best we can! For starters, make everything lowercase, get rid of space and hope for the best.

```
clean_up_strings <-function(str){
    str %>%
    tolower %>%
    gsub(' ','', .)
}

mps_attributes$mp_name %<>% clean_up_strings
all_sessions$mp_name %<>% clean_up_strings

mps_attributes$constituency %<>% clean_up_strings
all_sessions$constituency   %<>% clean_up_strings
```

Also, manually fix up some spelling differences between the two files.

```
## constituency as spelt in mp_attributes = as in all_sessions
mps_attributes$constituency %<>%
 forcats::fct_recode(ramanathapuram = "ramanthapuram",
                     mahesana = "mehesana",
                     kushinagar = "khushinagar",
                     krishnagiri = "krisnagiri",
                     bangaon = "bongaon")

levels(mps_attributes$constituency) %<>% c("nominatedanglo-indian")

## set this to nominatedanglo-indian
mps_attributes[mps_attributes$nature_of_membership == "Nominated",
               "constituency"] <-  "nominatedanglo-indian"
```

Join time!

```
attendance_with_attributes <- left_join(all_sessions, mps_attributes,
                                        by = "constituency")
```

```
## Warning: Column `constituency` joining character vector and factor,
## coercing into character vector
```

```
##
```

No complaints from R. Check data using skimr and a few rows.