Extracting Data from pdf file

Sang Kim

May 1, 2019

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------------
---------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------------
----------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(purrr)
library(pdftools)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
fn <- system.file("extdata", "RD-Mortality-Report_2015-18-180531.pdf", package="dslabs")
dat <- map_df(str_split(pdf_text(fn), "\n"), function(s){
  s <- str_trim(s)
  header_index <- str_which(s, "2015")[1]
  tmp <- str_split(s[header_index], "\\s+", simplify = TRUE)
  month <- tmp[1]
  header <- tmp[-1]
  tail_index  <- str_which(s, "Total")
  n <- str_count(s, "\\d+")
  out <- c(1:header_index, which(n==1), which(n>=28), tail_index:length(s))
  s[-out] %>%
    str_remove_all("[^\\d\\s]") %>%
    str_trim() %>%
    str_split_fixed("\\s+", n = 6) %>%
    .[,1:5] %>%
    as_data_frame() %>%
    setNames(c("day", header)) %>%
    mutate(month = month,
           day = as.numeric(day)) %>%
    gather(year, deaths, -c(day, month)) %>%
    mutate(deaths = as.numeric(deaths))
}) %>%
  mutate(month = recode(month, "JAN" = 1, "FEB" = 2, "MAR" = 3, "APR" = 4, "MAY" = 5, "JUN" = 6,
                        "JUL" = 7, "AGO" = 8, "SEP" = 9, "OCT" = 10, "NOV" = 11, "DEC" = 12)) %>%
  mutate(date = make_date(year, month, day)) %>%
  filter(date <= "2018-05-01")

head(dat)
```

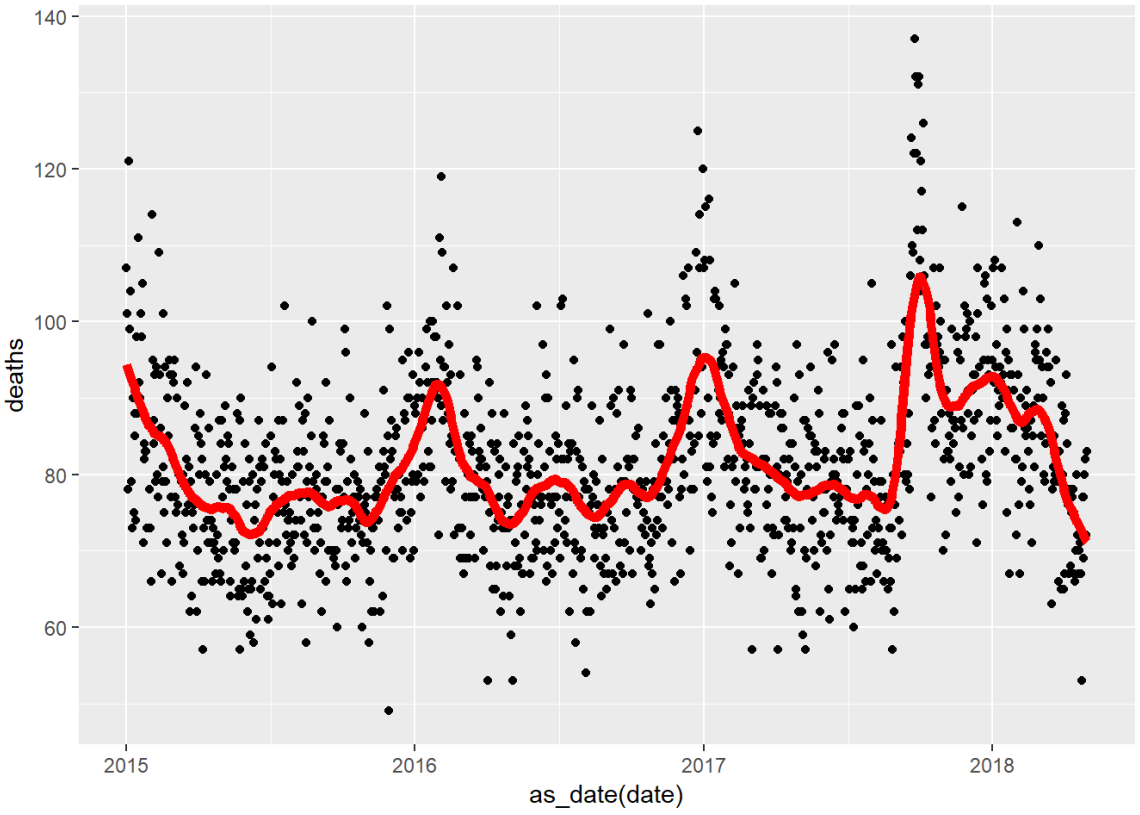
```
## # A tibble: 6 x 5
##     day month year  deaths date
##   <dbl> <dbl> <chr>  <dbl> <date>
## 1     1     1 2015     107 2015-01-01
## 2     2     1 2015     101 2015-01-02
## 3     3     1 2015      78 2015-01-03
## 4     4     1 2015     121 2015-01-04
## 5     5     1 2015      99 2015-01-05
## 6     6     1 2015     104 2015-01-06
```

```
glimpse(dat)
```

```
## Observations: 1,205
## Variables: 5
## $ day    <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, ...
## $ month  <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ year   <chr> "2015", "2015", "2015", "2015", "2015", "2015", "2015",...
## $ deaths <dbl> 107, 101, 78, 121, 99, 104, 79, 73, 90, 75, 88, 85, 74,...
## $ date   <date> 2015-01-01, 2015-01-02, 2015-01-03, 2015-01-04, 2015-0...
```

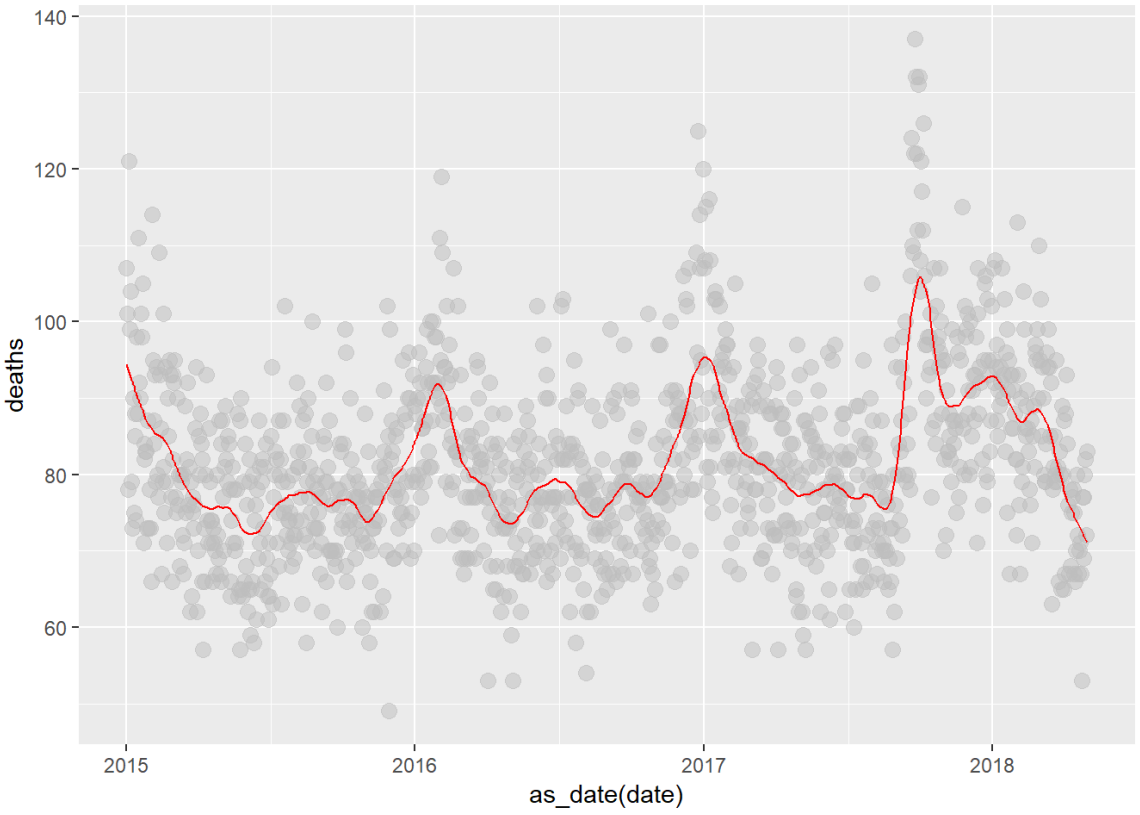
```
# Q1
span <- 60 / as.numeric(diff(range(dat$date)))
fit <- dat %>% mutate(x = as.numeric(date)) %>% loess(deaths ~ x, data = ., span = span, degree = 1)
dat %>% mutate(smooth = predict(fit, as.numeric(date))) %>%
  ggplot() +
  geom_point(aes(as_date(date), deaths)) +
  geom_line(aes(as_date(date), smooth), lwd = 2, col = 2)
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```




```
dat <- dat %>% mutate(date=as.numeric(date)) %>% filter(!is.na(deaths))

total_days <- diff(range(dat$date))
span <- 60/total_days
fit <- loess(deaths ~ date, degree=1, span = span, data=dat)
dat %>% mutate(smooth = fit$fitted) %>%
  ggplot(aes(as_date(date), deaths), color=year) +
  geom_point(size = 3, alpha = .5, color = "grey") +
  geom_line(aes(as_date(date), smooth), color="red")
```


140
120
100
80
60
deaths
2015
2016
2017
2018
as_date(date)