Comprehension Check: Smoothing

Q1

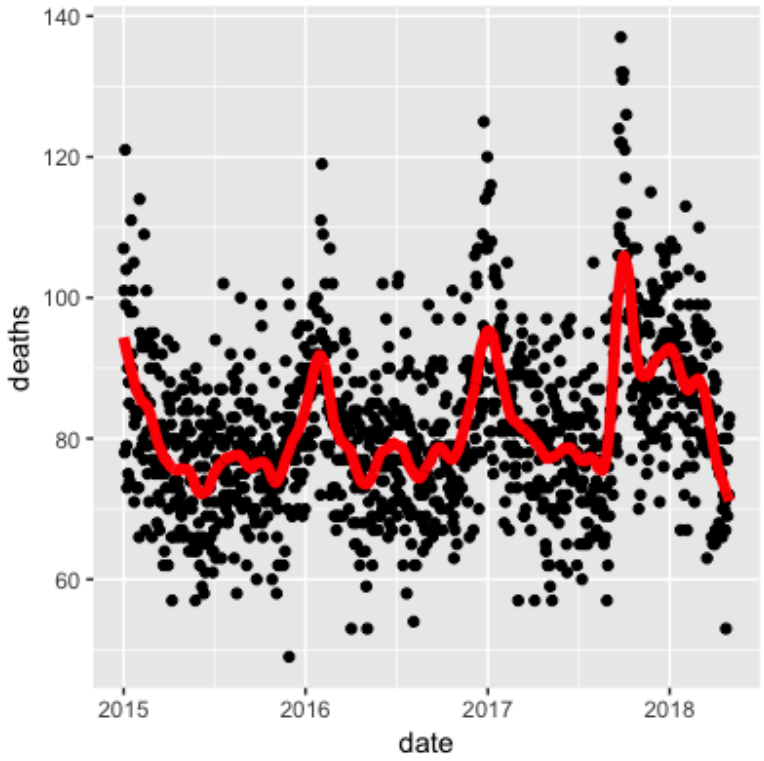
1/1 point (graded)

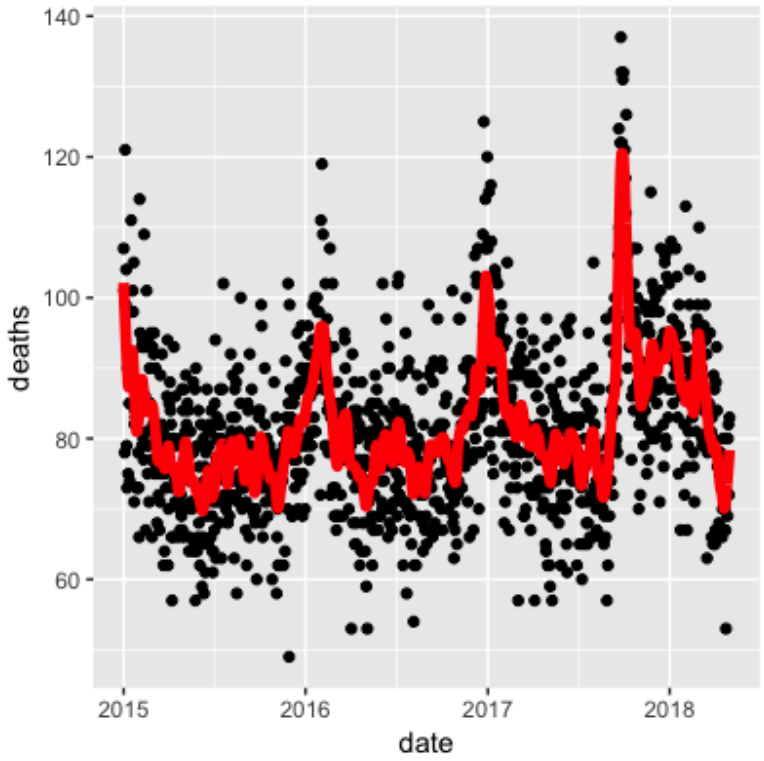
In the Wrangling course of this series, PH125.6x, we used the following code to obtain mortality counts for Puerto Rico for 2015-2018:

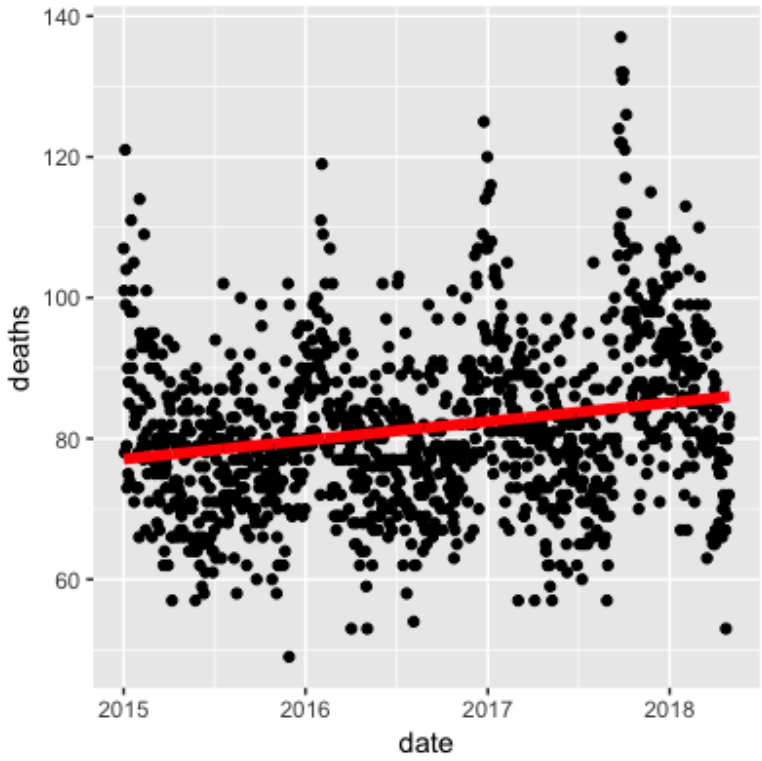
```
library(tidyverse)
library(lubridate)
library(purrr)
library(pdftools)

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" =
                              "JUL" = 7, "AGO" = 8, "SEP" = 9, "OCT" = 10, "NOV" = 11, "DEC" = 12)) %>%
        mutate(date = make_date(year, month, day)) %>%
      filter(date <= "2018-05-01")
```

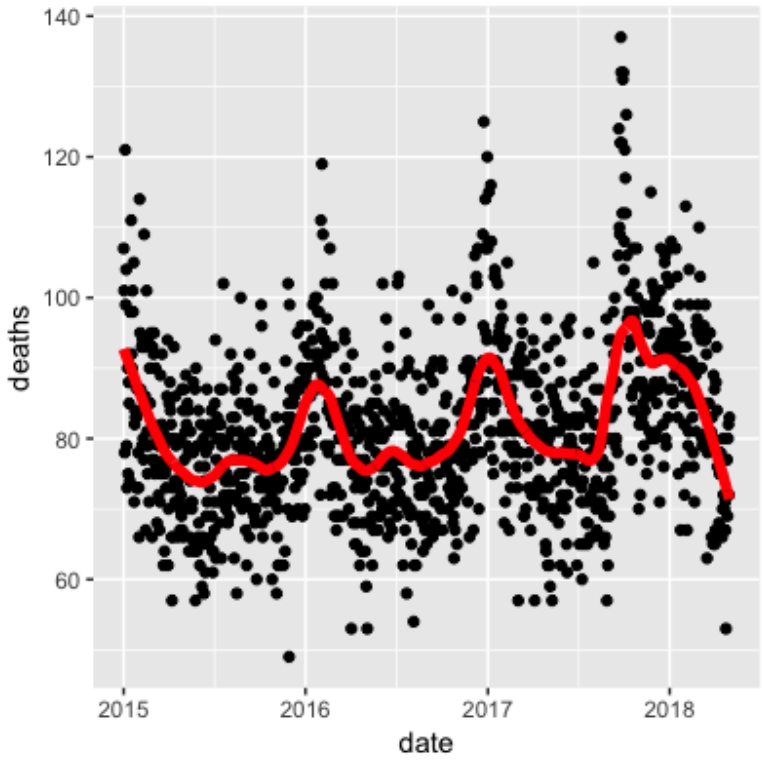
Use the `loess` function to obtain a smooth estimate of the expected number of deaths as a function of date. Plot this resulting smooth function. Make the span about two months long.

Which of the following plots is correct?









✔

Explanation

The following code makes the correct plot:

```
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(date, deaths)) +
        geom_line(aes(date, smooth), lwd = 2, col = 2)
```

The second plot uses a shorter span, the third plot uses the entire timespan, and the fourth plot uses a longer span.

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Q2

1/1 point (graded)

Work with the same data as in Q1 to plot smooth estimates against day of the year, all on the same plot, but with different colors for each year.

Which code produces the desired plot?

```
dat %>%
        mutate(smooth = predict(fit), day = yday(date), year =
as.character(year(date))) %>%
        ggplot(aes(day, smooth, col = year)) +
        geom_line(lwd = 2)
```

```
dat %>%
        mutate(smooth = predict(fit, as.numeric(date)), day = mday(date), year =
as.character(year(date))) %>%
        ggplot(aes(day, smooth, col = year)) +
        geom_line(lwd = 2)
```

```
dat %>%
        mutate(smooth = predict(fit, as.numeric(date)), day = yday(date), year =
as.character(year(date))) %>%
        ggplot(aes(day, smooth)) +
        geom_line(lwd = 2)
```

```
dat %>%
        mutate(smooth = predict(fit, as.numeric(date)), day = yday(date), year =
as.character(year(date))) %>%
        ggplot(aes(day, smooth, col = year)) +
        geom_line(lwd = 2)
```

✔

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Q3

1/1 point (graded)

Suppose we want to predict 2s and 7s in the `mnist_27` dataset with just the second covariate. Can we do this? On first inspection it appears the data does not have much predictive power.

In fact, if we fit a regular logistic regression the coefficient for `x_2` is not significant!

This can be seen using this code:

```
library(broom)
mnist_27$train %>% glm(y ~ x_2, family = "binomial", data = .) %>% tidy()
```

Plotting a scatterplot here is not useful since `y` is binary:

```
qplot(x_2, y, data = mnist_27$train)
```

Fit a loess line to the data above and plot the results. What do you observe?

- There is no predictive power and the conditional probability is linear.

- ○ There is no predictive power and the conditional probability is non-linear.
- ○ There is predictive power and the conditional probability is linear.
- ◉ There is predictive power and the conditional probability is non-linear. ✔

Explanation
Note that there is indeed predictive power, but that the conditional probability is non-linear.
The loess line can be plotted using the following code:

```
mnist_27$train %>%
        mutate(y = ifelse(y=="7", 1, 0)) %>%
        ggplot(aes(x_2, y)) +
        geom_smooth(method = "loess")
```

Submit You have used 1 of 2 attempts

ℹ Answers are displayed within the problem

Ask your questions or make your comments about Smoothing here! **Remember, one of the best ways to reinforce your own learning is by explaining something to someone else, so we encourage you to answer each other's questions (without giving away the answers, of course).**

Some reminders:

- Search the discussion board before posting to see if someone else has asked the same thing before asking a new question
- Please be specific in the title and body of your post regarding which question you're asking about to facilitate answering your question.
- Posting snippets of code is okay, but posting full code solutions is not.
- If you do post snippets of code, please format it as code for readability. If you're not sure how to do this, there are instructions in a pinned post in the "general" discussion forum.

Discussion: Smoothing

Show Discussion

Topic: Section 3: Linear Regression for Prediction, Smoothing, and Working with Matrices / 3.2: Smoothing