ggplot lecture - solutions

datalab

2022-05-20

0. load libraries and import data

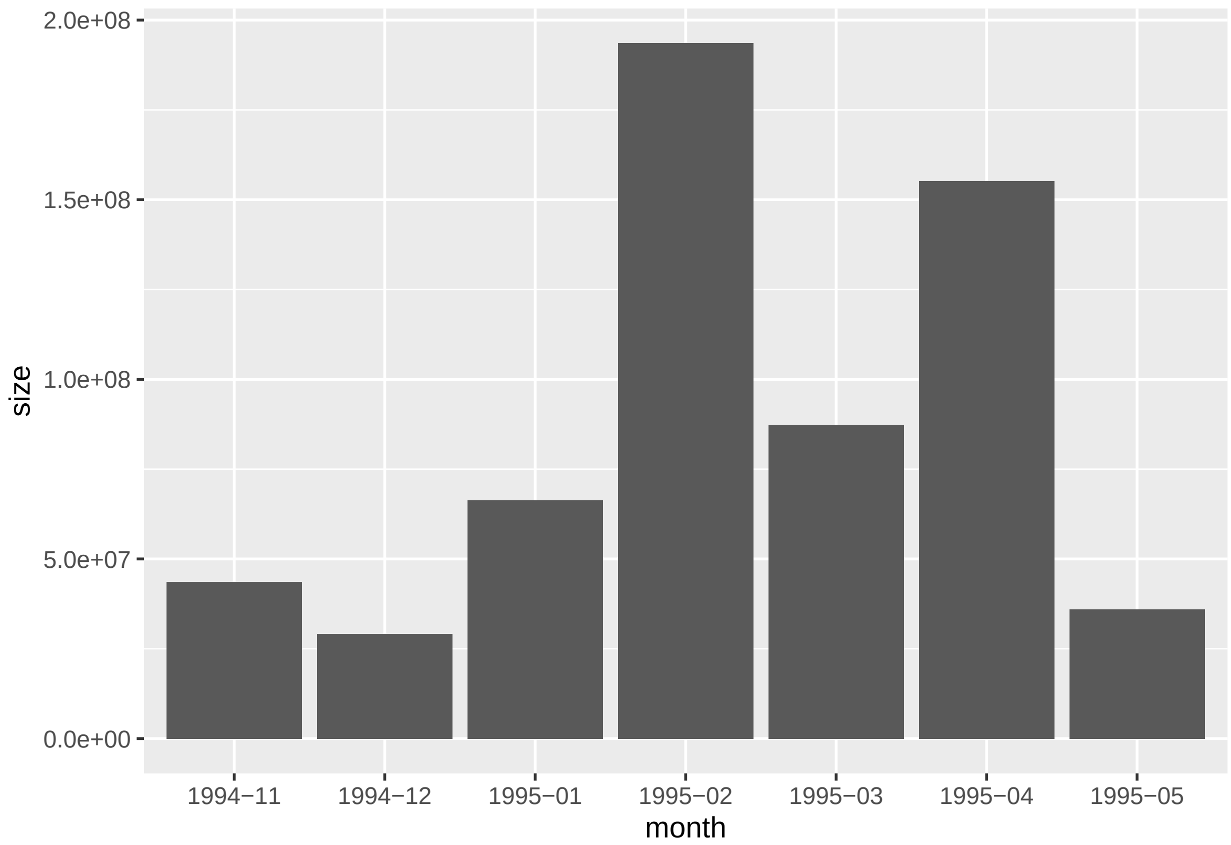
```
library(tidyverse)
library(readxl)
downloads <-
  read_excel("../Presentations/downloads.xlsx") %>%
  filter(size > 0)
downloads
```

```
## # A tibble: 36,708 x 6
##    machineName userID  size  time date                month
##    <chr>        <dbl> <dbl> <dbl> <dttm>              <chr>
##  1 cs18        146579  2464 0.493 1995-04-24 00:00:00 1995-04
##  2 cs18        995988  7745 0.326 1995-04-24 00:00:00 1995-04
##  3 cs18        317649  6727 0.314 1995-04-24 00:00:00 1995-04
##  4 cs18        748501 13049 0.583 1995-04-24 00:00:00 1995-04
##  5 cs18        955815   356 0.259 1995-04-24 00:00:00 1995-04
##  6 cs18        596819 15063 0.336 1995-04-24 00:00:00 1995-04
##  7 cs18        169424  2548 0.285 1995-04-24 00:00:00 1995-04
##  8 cs18        386686  1932 0.286 1995-04-24 00:00:00 1995-04
##  9 cs18        783767  7294 0.397 1995-04-24 00:00:00 1995-04
## 10 cs18        788633  4470 3.41  1995-04-24 00:00:00 1995-04
## # ... with 36,698 more rows
```

Exercise A: 10 mins

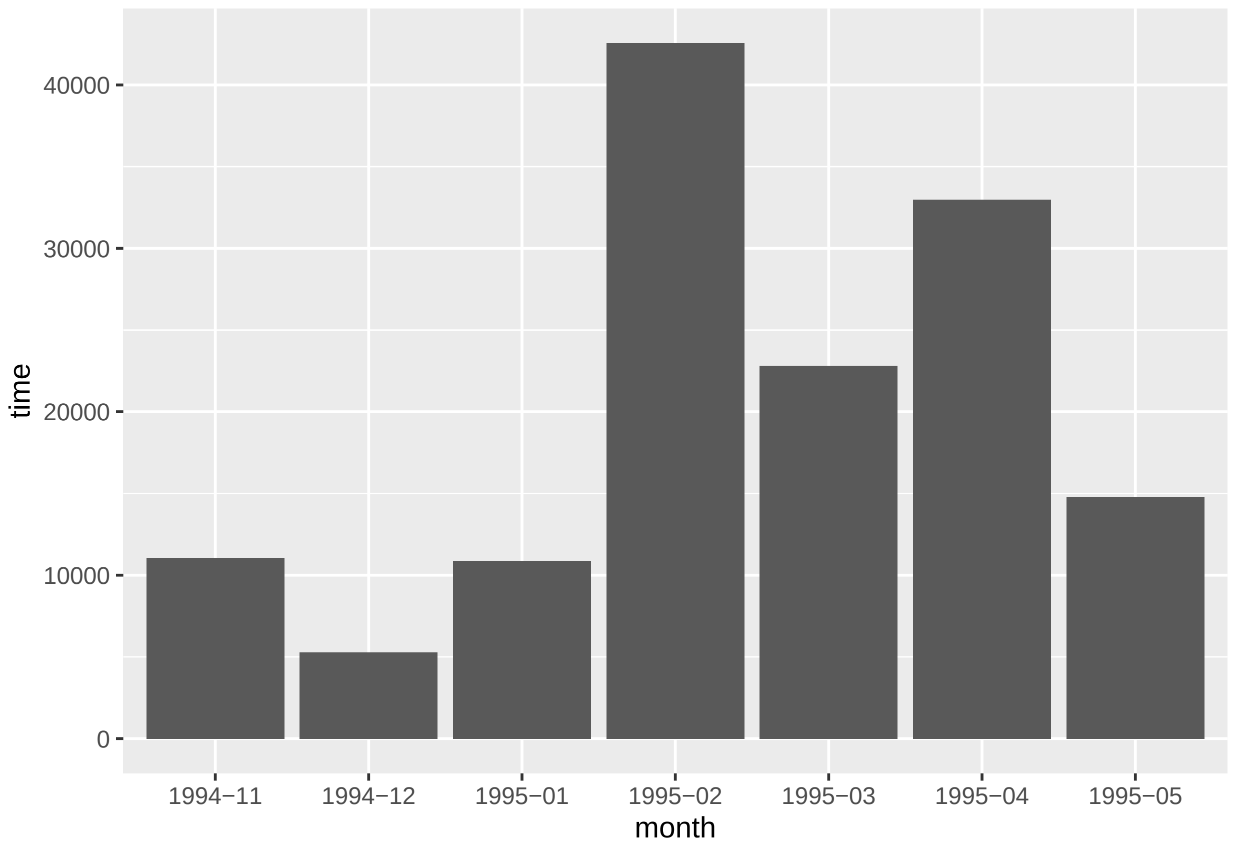
1. Make a bar chart of the downloads data showing the total download size per month. Hint: Very similar to the first example shown during the lecture

```
ggplot(downloads,aes(x=month,y=size)) +
  geom_col()
```



2. Make a bar chart of the downloads data showing the total time spend on downloads per month.

```
ggplot(downloads,aes(x=month,y=time)) +
  geom_col()
```



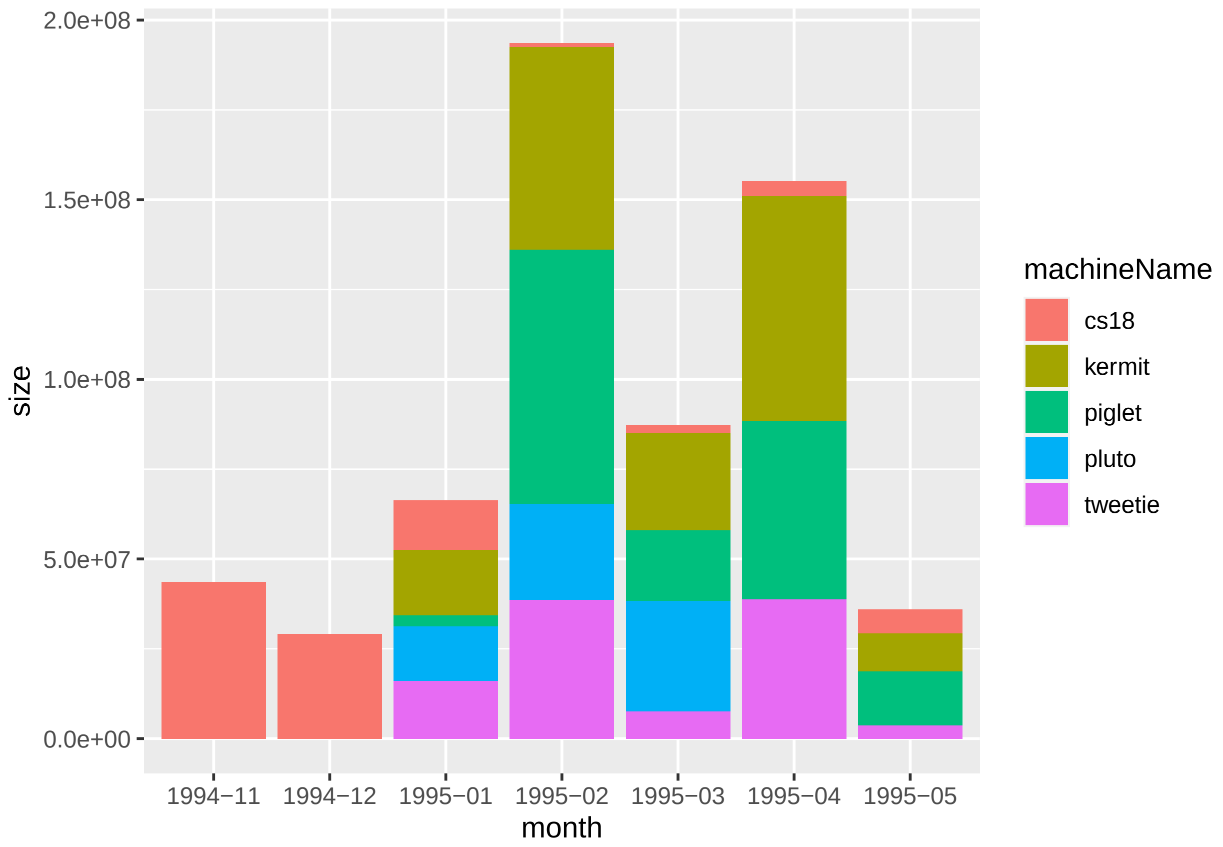
3. Assign the plot you made in 1. to the variable p_size_month .

```
p_size_month <- ggplot(downloads,aes(x=month,y=size)) +
  geom_col()
```

Exercise B: 7 mins

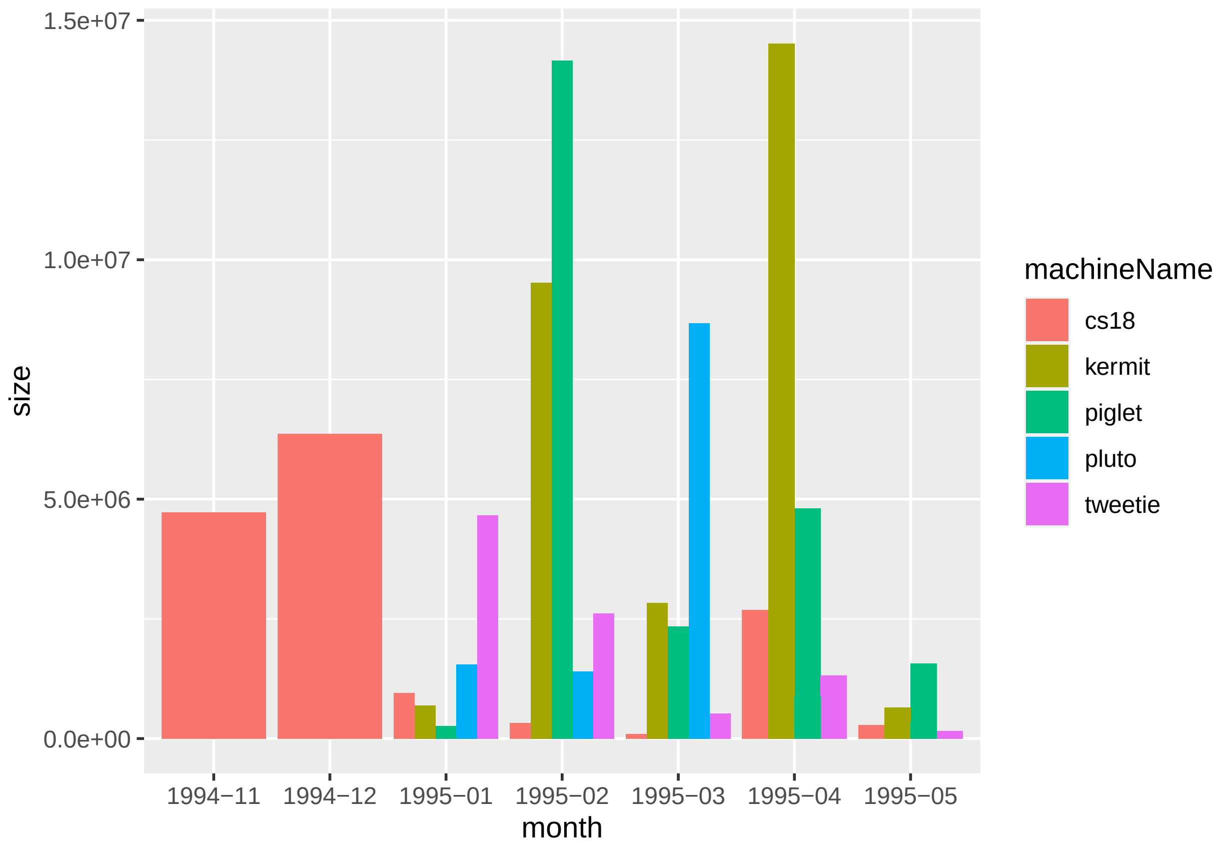
1. On the bar chart you made in A3 (p_size_month), add coloring by the machineName by using the ‘fill’ keyword in the aes.

```
p_size_month <- ggplot(downloads,aes(x=month,y=size, fill = machineName)) +
  geom_col()
p_size_month
```



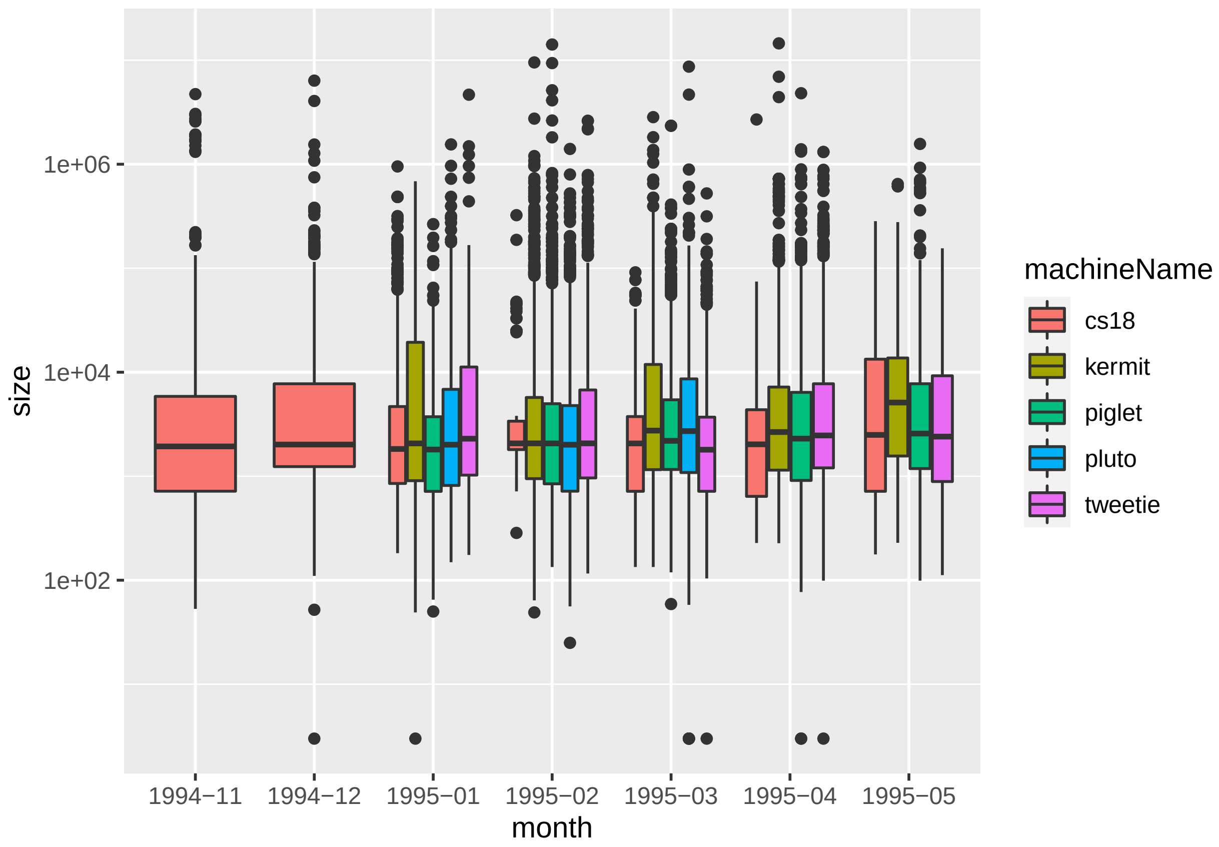
2. Now, position the bars for the different machines next to each other instead of stacked. Hint: Use the ‘position’ keyword.

```
p_size_month <- ggplot(downloads,aes(x=month,y=size, fill = machineName)) +
  geom_col(position = 'dodge')
p_size_month
```



3. Now turn it into a boxplot instead. If it's hard to see the boxes try to make the scale of the size axis logarithmic.

```
p_size_month <- ggplot(downloads,aes(x=month,y=size, fill = machineName)) +
  geom_boxplot() + scale_y_log10()
p_size_month
```



Exercise C

0. Create daily_downloads dataframe (from lecture).

```
daily_downloads <- downloads %>%
    group_by(machineName, date) %>%
    summarize(dl_count = n(), size_mb = sum(size)/10^6) %>%
    mutate(total_dl_count = cumsum(dl_count))
```

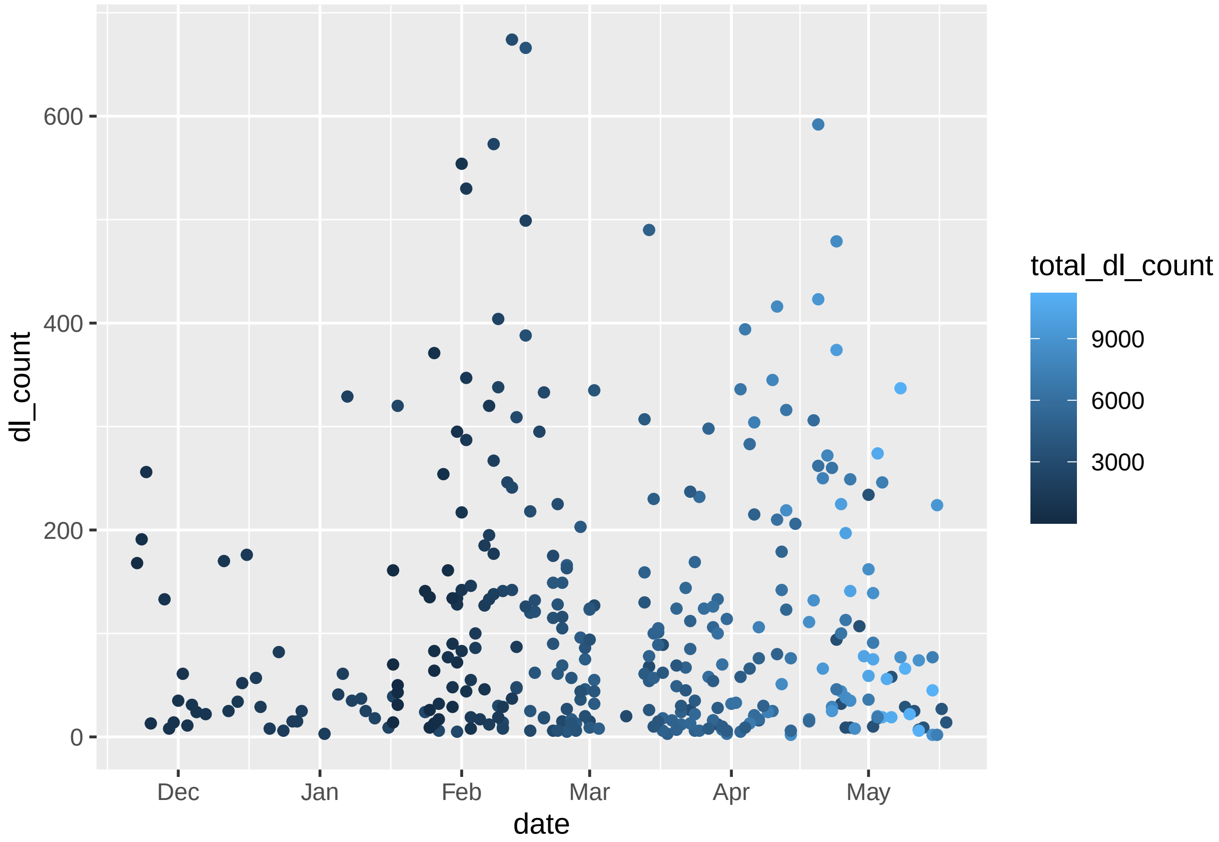
```
## `summarise()` has grouped output by 'machineName'. You can override using the
## `.groups` argument.
```

```
daily_downloads
```

```
## # A tibble: 337 x 5
## # Groups:   machineName [5]
##    machineName date                dl_count size_mb total_dl_count
##    <chr>       <dttm>                 <int>   <dbl>          <int>
##  1 cs18        1994-11-22 00:00:00      168 22.4               168
##  2 cs18        1994-11-23 00:00:00      191 12.2               359
##  3 cs18        1994-11-24 00:00:00      256  8.05              615
##  4 cs18        1994-11-25 00:00:00       13  0.0655            628
##  5 cs18        1994-11-28 00:00:00      133  0.625             761
##  6 cs18        1994-11-29 00:00:00        8  0.0201            769
##  7 cs18        1994-11-30 00:00:00       14  0.209             783
##  8 cs18        1994-12-01 00:00:00       35  0.631             818
##  9 cs18        1994-12-02 00:00:00       61  5.67              879
## 10 cs18         1994-12-03 00:00:00       11  0.156              890
## # ... with 327 more rows
```

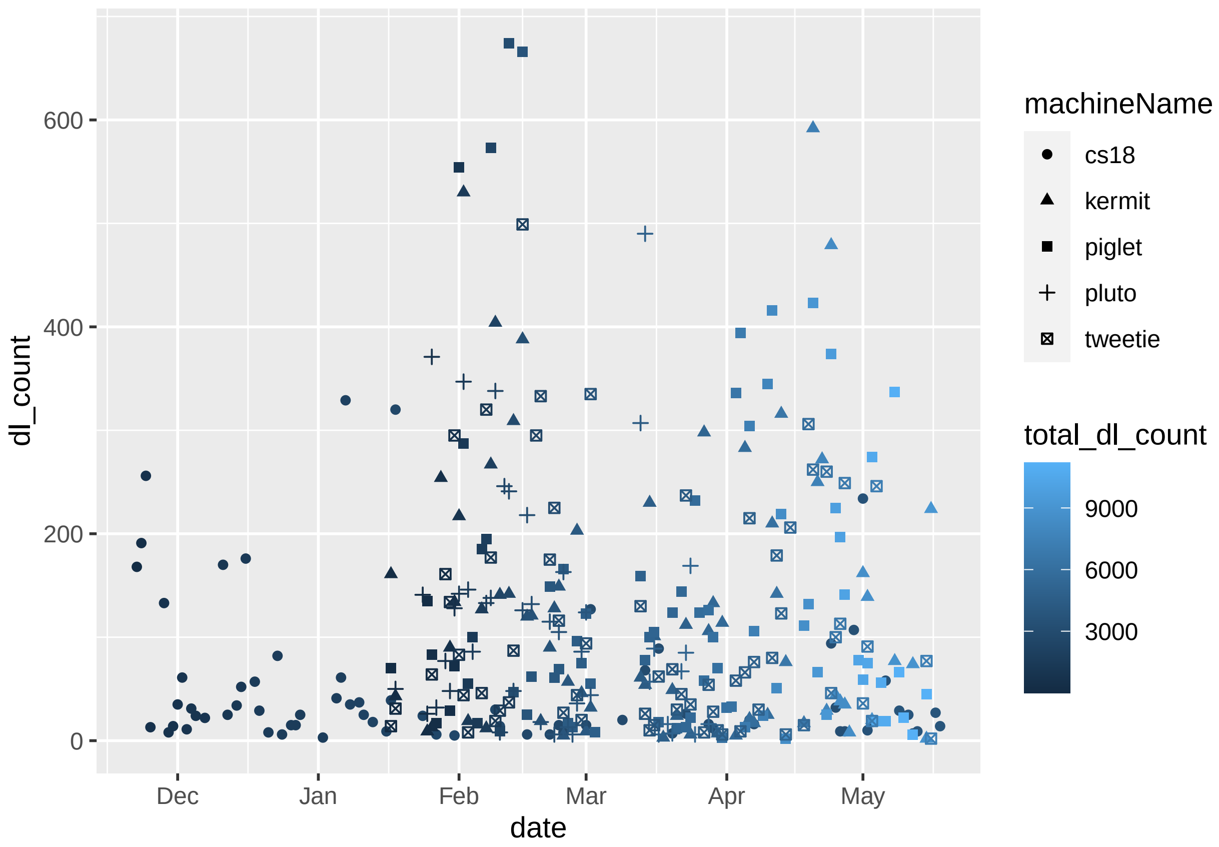
1. Add coloring by the total download count (total_dl_count) to this plot: p <- ggplot(daily_downloads, aes(x = date, y = dl_count)) + geom_point()

```
p <- ggplot(daily_downloads, aes(x = date, y = dl_count, color = total_dl_count)) +
  geom_point()
p
```



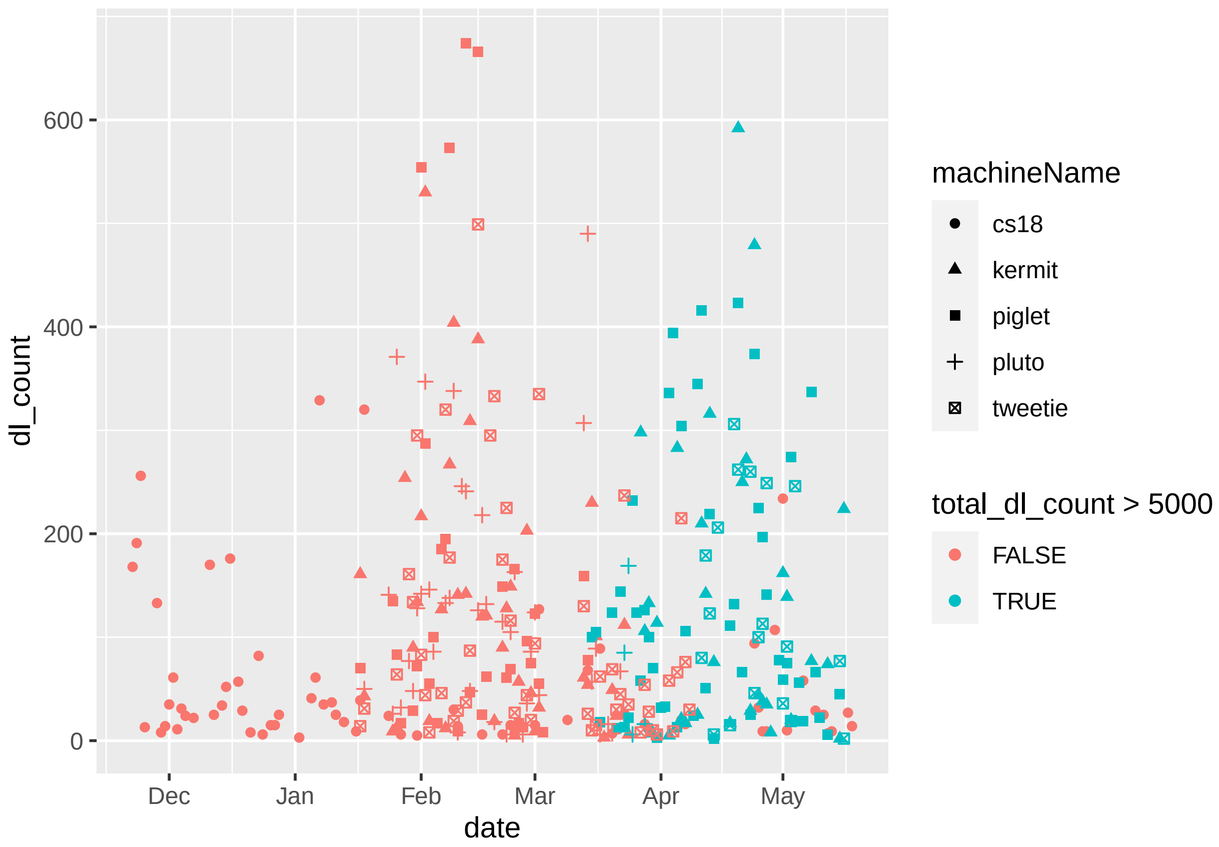
2. Add a different point shape depending on the machine to the same plot.

```
p <- ggplot(daily_downloads, aes(x = date, y = dl_count, color = total_dl_count, shape = machineName)) +
  geom_point()
p
```



3. Change the coloring to be discrete instead of continuous. You can choose total_dl_count > 5000 or any cutoff you like.

```
p <- ggplot(daily_downloads, aes(x = date, y = dl_count, color = total_dl_count > 5000, shape = machineN
  geom_point()
p
```


dl_count
600
400
200
0
Dec
Jan
Feb
Mar
Apr
May
date
machineName
cs18
kermit
piglet
pluto
tweetie
total_dl_count > 5000
FALSE
TRUE