Data Incubator Project

Shouman Das

February 4, 2019

First we load the two data sets and useful packages. It takes a little time, as the data set is big.

```
library(dplyr)
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
linkedin = read.csv("temp_datalab_records_linkedin_company.csv")
facebook = read.csv("temp_datalab_records_social_facebook.csv")
```

Next we do a simple summary of the interesting columns of the two datasets.

```
summary(linkedin[,c(2,3,4,5,7)])
```

```
##       as_of_date                  company_name     followers_count
##  2018-02-17:   4430   City National Bank:   1605   Min.   :      0
##  2018-02-16:   4429   American Airlines :   1029   1st Qu.:   2148
##  2018-02-13:   4427   Apple             :   1025   Median :   9335
##  2018-02-14:   4427   Activision        :   1024   Mean   :  71677
##  2018-02-15:   4427   Amgen             :   1024   3rd Qu.:  38642
##  2017-12-19:   4426   Cisco             :   1024   Max.   :7833967
##  (Other)   :2399630   (Other)           :2419465
##  employees_on_platform               industry
##  Min.   :     0        Banking           : 168364
##  1st Qu.:   218        Biotechnology     : 152710
##  Median :  1083        Financial Services: 148143
##  Mean   :  7587        Oil & Energy      : 116830
##  3rd Qu.:  4513        Retail            :  95384
##  Max.   :577952        Pharmaceuticals   :  92107
##                        (Other)           :1652658
```

```
summary(facebook[,c(2,3,4,7,8,9)])
```

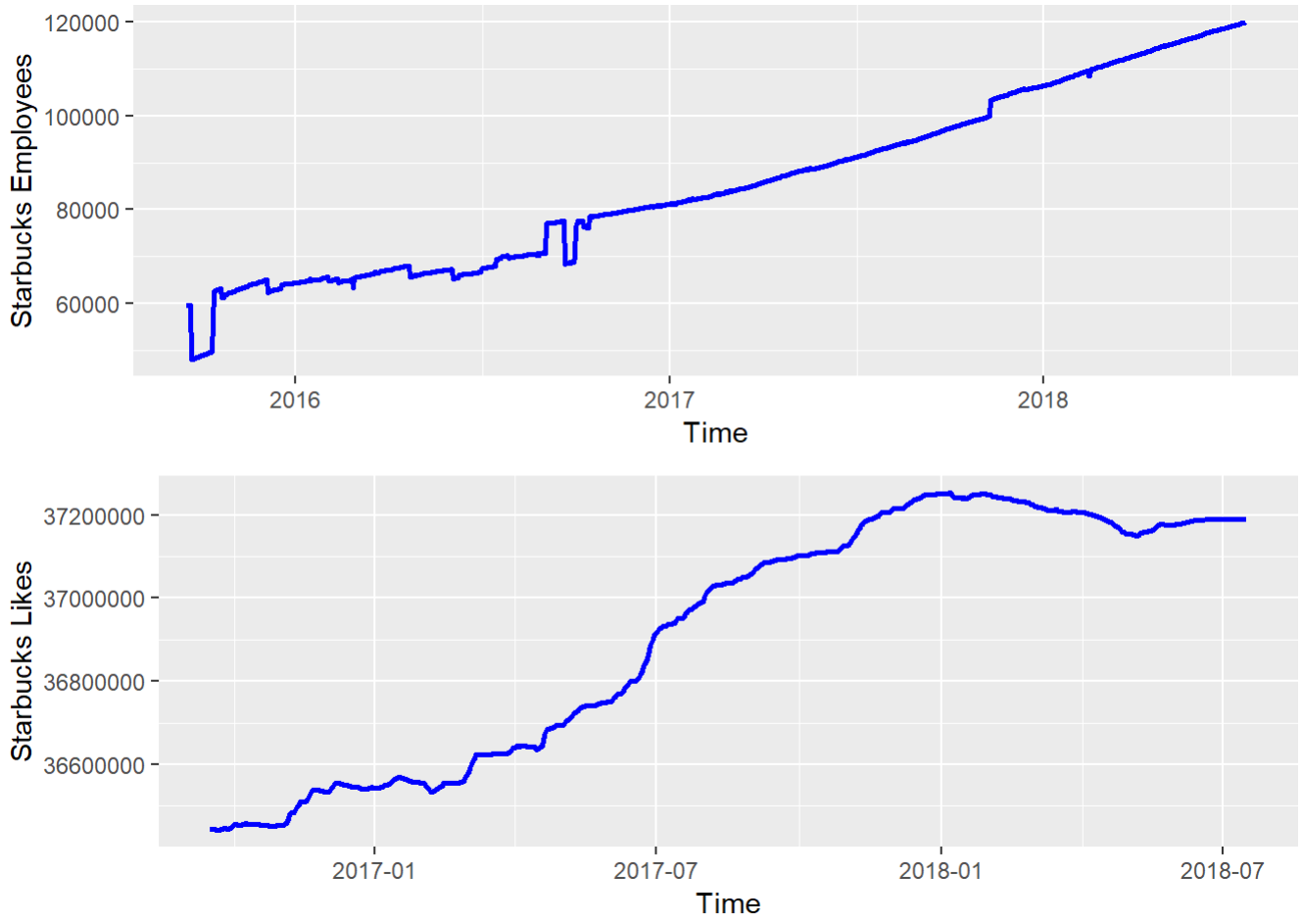
```
##                      time               username          checkins
##  2018-03-20 04:00:00+00:   4502             : 120929   Min.   :       0
##  2018-03-21 04:00:00+00:   4502   2u        :   1222   1st Qu.:       0
##  2018-03-22 04:00:00+00:   4502   aflacduck :   1222   Median :      13
##  2018-03-23 04:00:00+00:   4502   ModelNInc :   1222   Mean   :   14170
##  2018-02-27 05:00:00+00:   4501   RedRobin  :   1222   3rd Qu.:     286
##  2018-02-28 05:00:00+00:   4501   shutterfly:   1222   Max.   :17290550
##  (Other)               :3594381   (Other)   :3494352
##      likes           talking_about_count  facebook_id
##  Min.   :        1   Min.   :      0     Min.   :5.182e+09
##  1st Qu.:     2500   1st Qu.:     27     1st Qu.:9.448e+10
##  Median :    20477   Median :    251     Median :1.123e+14
##  Mean   :   816625   Mean   :  10043     Mean   :1.738e+14
##  3rd Qu.:   217579   3rd Qu.:   2474     3rd Qu.:1.941e+14
##  Max.   :210641077   Max.   :5747010     Max.   :1.015e+16
##
```

To start with some exploratory plots, I pick some well-known established companies which appear in the both data sets and then I plot the timeseries data. For example, first I picked “Starbucks” and plot it’s number of facebook likes and linkedin profile.

```
starbucks.lkd = linkedin[linkedin$company_name=="Starbucks",c(2,5)]
starbucks.fb = facebook[facebook$username=="Starbucks",c(2,7)]

starbucks.lkd$as_of_date = as.POSIXct(starbucks.lkd$as_of_date)
starbucks.fb$time = as.POSIXct(starbucks.fb$time)

# plotting
s.lkd = ggplot(starbucks.lkd, aes(x = as_of_date, y = employees_on_platform)) +
  geom_line(size = 1, color = "blue") + xlab("Time")+ylab("Starbucks Employees")
s.fb = ggplot(starbucks.fb, aes(x = time, y = likes)) +
  geom_line(size = 1, color = "blue") + xlab("Time")+ylab("Starbucks Likes")
grid.arrange(s.lkd, s.fb, nrow = 2)
```



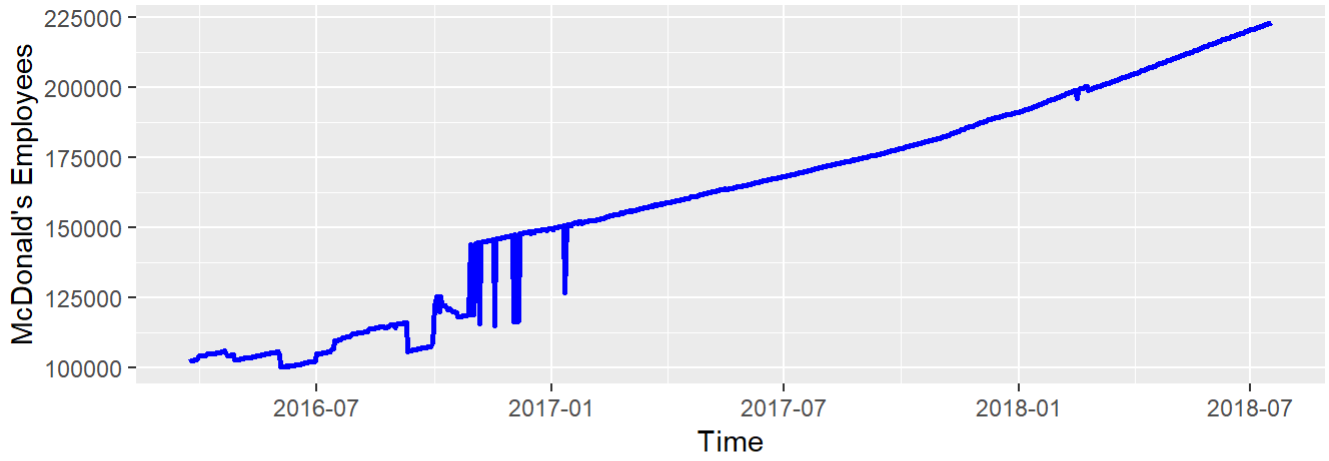
From the plot, we can infer that at the beginning of 2018, starbucks social media popularity didn't perform as well as the number of employees. There might be some unknown variables which has effected badly their social media presence.

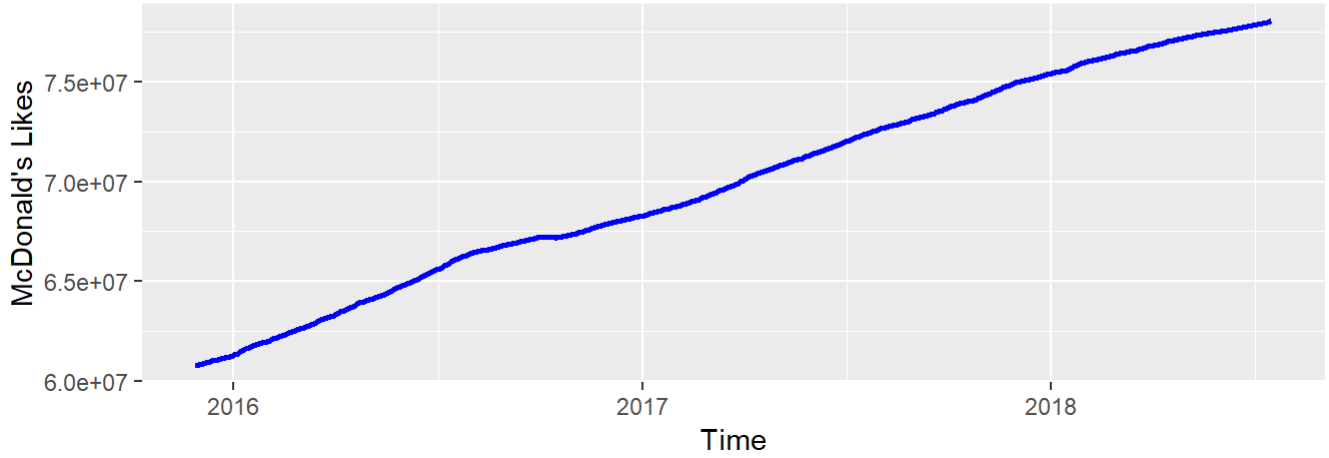
For the next plot we pick another well known company "McDonalds" which has a siginificant presence in the social media. We draw similar plot.

```
mcdonalds.lkd = linkedin[linkedin$company_name=="McDonald's",c(2,5)]
mcdonalds.fb = facebook[facebook$username=="McDonalds",c(2,7)]

mcdonalds.lkd$as_of_date = as.POSIXct(mcdonalds.lkd$as_of_date)
mcdonalds.fb$time = as.POSIXct(mcdonalds.fb$time)

# plotting
m.lkd = ggplot(mcdonalds.lkd, aes(x = as_of_date, y = employees_on_platform)) +
  geom_line(size = 1, color = "blue") + xlab("Time")+ylab("McDonald's Employees")
m.fb = ggplot(mcdonalds.fb, aes(x = time, y = likes)) +
  geom_line(size = 1, color = "blue") + xlab("Time")+ylab("McDonald's Likes")
grid.arrange(m.lkd, m.fb, nrow = 2)
```





For Mcdonald's, the employee growth and social media growth looks highly correlated.

It would be an interesting question how this number of employees and social media popularity are correlated for other medium to small scale companies.