Assignment4

SIYU LIU(U74835706)

2/24/2020

Q1

ab

```
setDT(replies)
setDT(mentions)


daily_replies <- replies[,.N,by = format(replies$created_at,'%Y-%m-%d')]
daily_mention <- mentions[,.N,by = format(mentions$created_at,'%Y-%m-%d')]

mean(daily_replies$N)
```

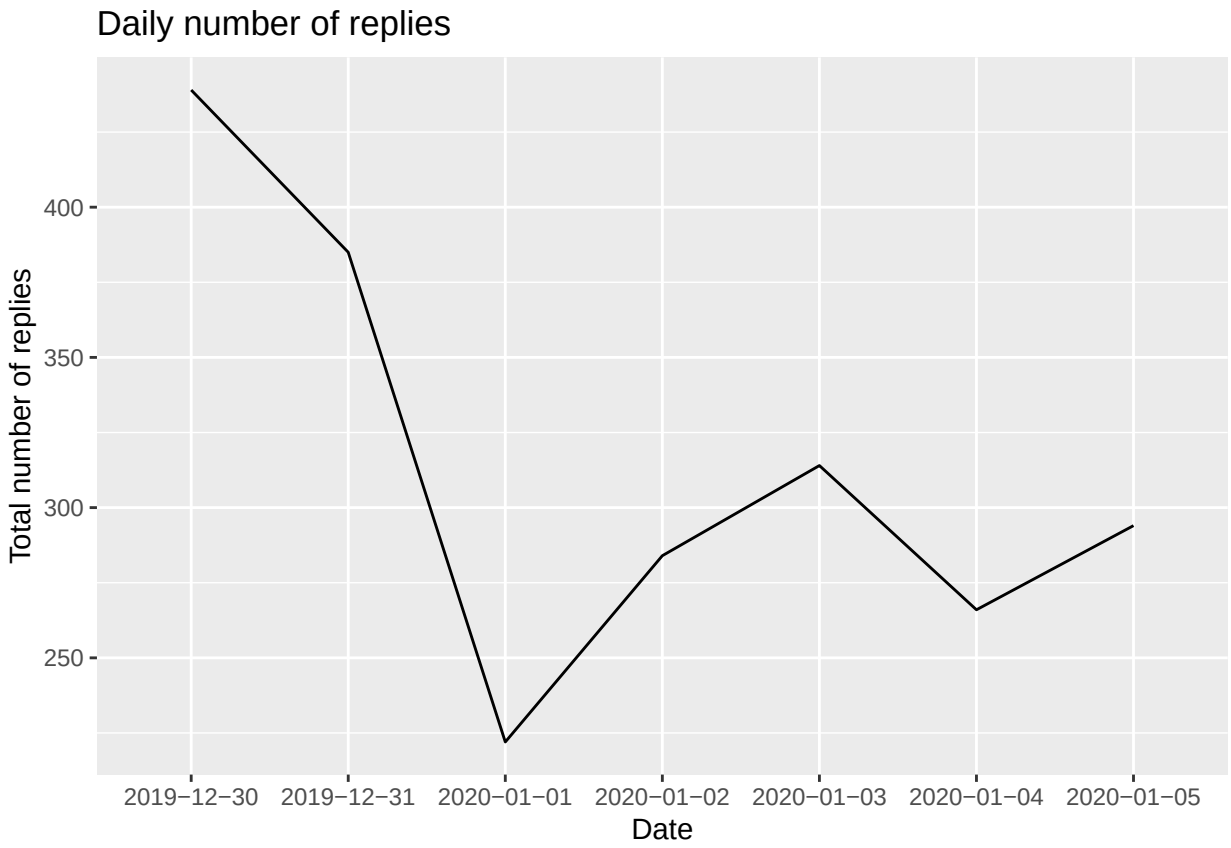
```
## [1] 314.8571
```

```
mean(daily_mention$N)
```

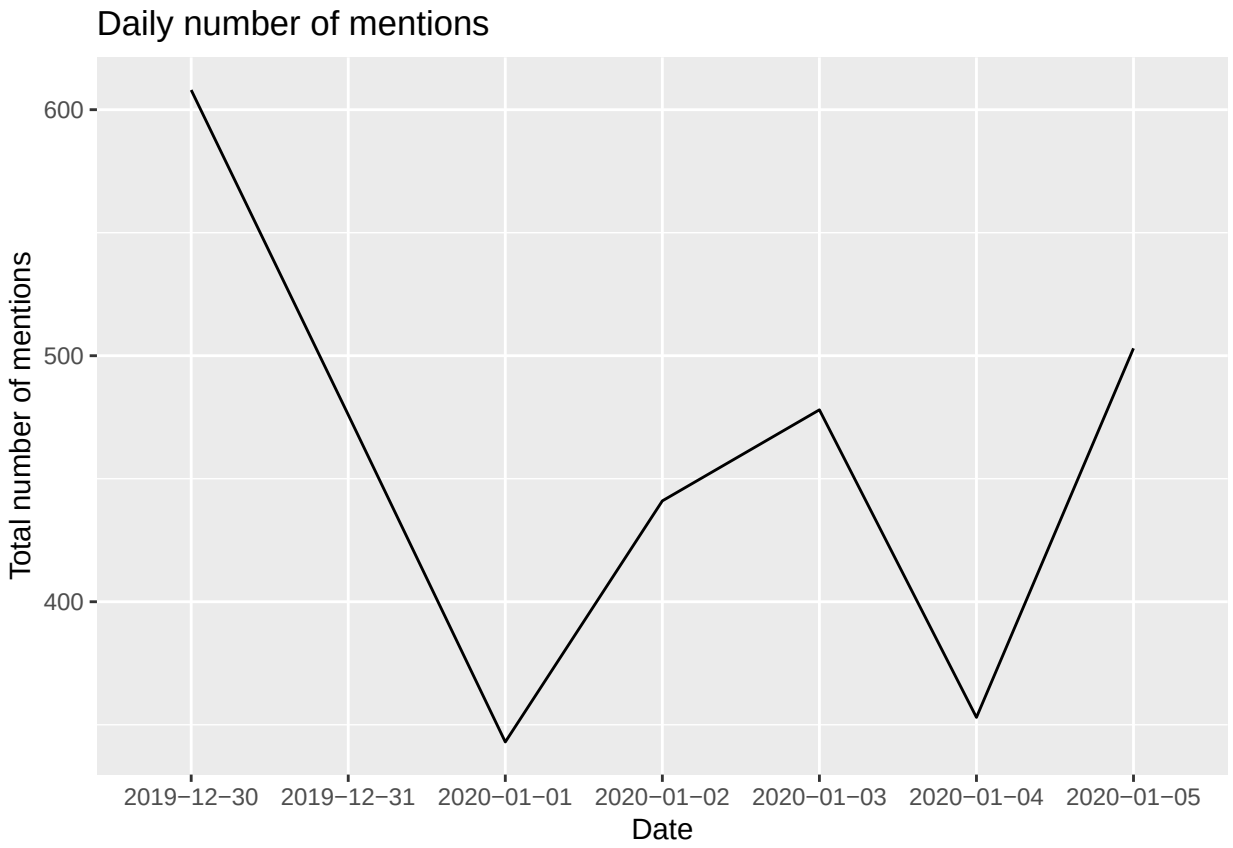
```
## [1] 457.4286
```

The average number of daily reply is 314.86 and the average number of daily mention is 457.42

```
ggplot(daily_replies)+
  geom_line(aes(x = format, y  = N,group = 1))+
  ylab('Total number of replies')+
  xlab("Date")+
  labs(title = 'Daily number of replies')
```




```
ggplot(daily_mention)+
  geom_line(aes(x = format, y  = N,group = 1))+
  ylab('Total number of mentions')+
  xlab('Date')+
  labs(title = 'Daily number of mentions')
```



Q2

a.

```
mentions %>% select(user_id,followers_count) %>% unique() %>% summarise(median = median(followers_count
```

```
##   median
## 1    325
```

```
mentions %>% select(screen_name,followers_count) %>% unique() %>% arrange(desc(followers_count)) %>% hea
```

```
##   screen_name followers_count
## 1  jdickerson         2007122
## 2   NYRangers         1456024
## 3         ajc         1045514
```

a.i The median number of followers in Delta's mention is 325. a.ii The screen name of user with the #3 most follower who mention delta is ajc

b.

```
mentions %>% select(favorite_count) %>% summarise(avg = mean(favorite_count), max = max(favorite_count)
```

```
##        avg max
## 1 3.271081 994
```

```
mentions$text[mentions$favorite_count == max(mentions$favorite_count)]
```

```
## [1] "Game winning moment from 30,000 feet. Thanks \u2066@Delta\u2069 &amp;amp; congrats \u2066@Vikings\u
```

b.i The average favorite by mentions is 3.27 while the maximun favorite by mentions is 994

b.ii The text with most favorite number by mention is “Game winning moment from 30,000 feet. Thanks Ž066@DeltaŽ069 & congrats Ž066@VikingsŽ069. #SKOL https://t.co/mjlHb5pAez”

c i

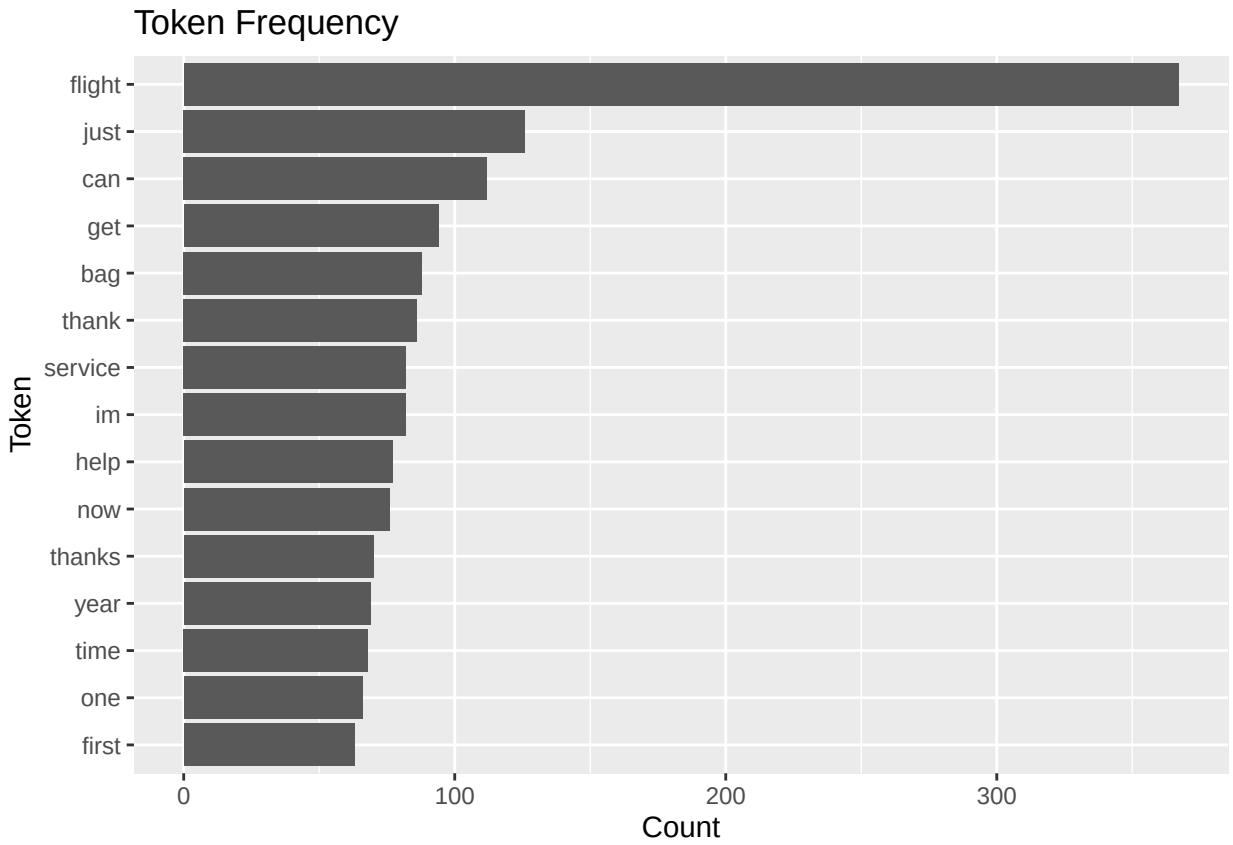
```
text <- mentions %>% filter(delta_responded ==T) %>% select(text)

clean_text <- text %>% unnest_tokens(token, text, token="tweets",
                                     to_lower = T,
                                     strip_punct = T,
                                     strip_url = T)
```

```
## Using `to_lower = TRUE` with `token = 'tweets'` may not preserve URLs.
```

```
sw =get_stopwords()
sw = c(sw$word,'@delta','delta')
tidy_text <- clean_text %>% filter(!token %in% sw)
top <- tidy_text%>% count(token,sort = T) %>% top_n(15)
```

```
## Selecting by n
```

```
ggplot(top)+
  geom_col(aes(x = reorder(token,n),y = n))+
  coord_flip()+
  xlab('Token')+
  ylab('Count')+
  labs(title = 'Token Frequency')
```



c ii

```
sen_bing <-inner_join(tidy_text, get_sentiments("bing"),
                      by=c("token" = "word"))

sen_bing %>% group_by(sentiment, token) %>% count(sort = T) %>% slice(1:15)
```

```
## # A tibble: 437 x 3
## # Groups:   sentiment, token [437]
##    sentiment token         n
##    <chr>     <chr>     <int>
##  1 negative  absurd        1
##  2 negative  addict        1
##  3 negative  allergic      3
##  4 negative  allergies     1
##  5 negative  allergy       1
##  6 negative  angry         1
##  7 negative  annoying      2
##  8 negative  anxiety       1
##  9 negative  ashamed       1
## 10 negative  atrocious     1
## # ... with 427 more rows
```

```
sen_bing_n <- sen_bing %>% count(sentiment, token,sort = T) %>%  arrange(desc(n)) %>% group_by(sentimen
```

```
## Selecting by n
```

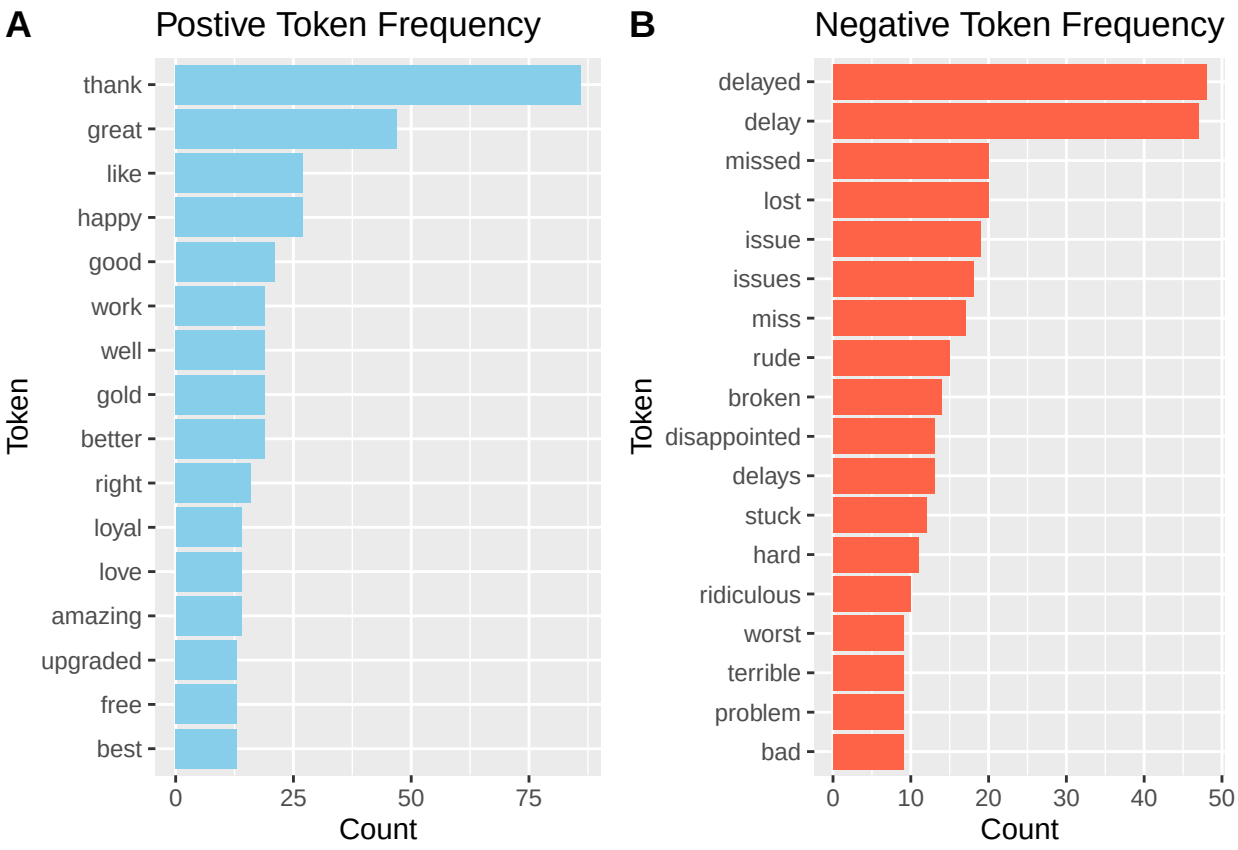
```
sen_bing_n_pos <- sen_bing %>% filter(sentiment == 'positive') %>% count(sentiment, token,sort = T) %>%
```

```
## Selecting by n
```

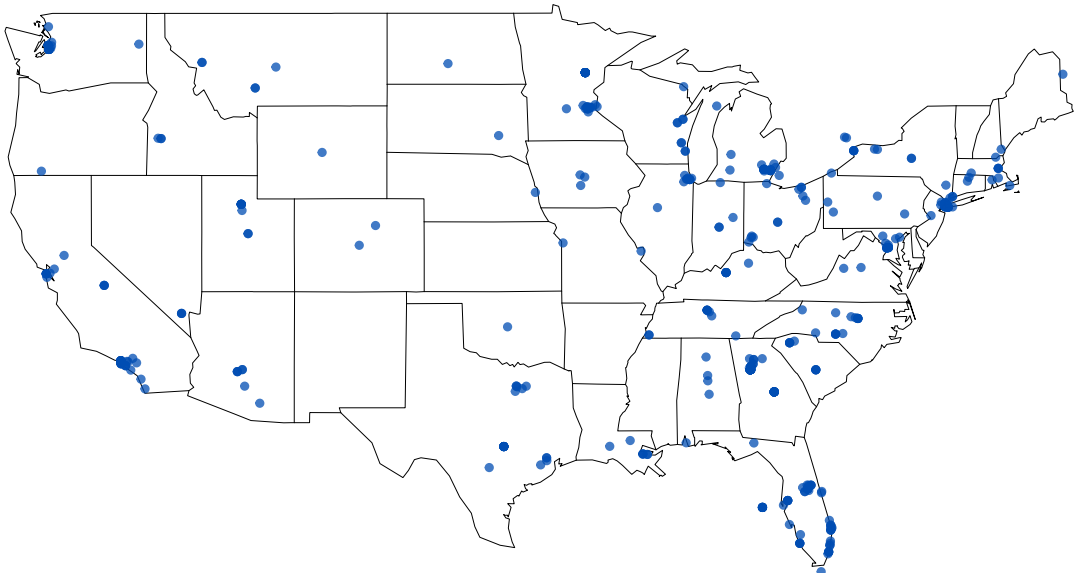
```
sen_bing_n_neg <- sen_bing %>% filter(sentiment == 'negative') %>% count(sentiment, token,sort = T) %>%
```

```
## Selecting by n
```

```
pos <- ggplot(sen_bing_n_pos)+
  geom_col(aes(x = reorder(token,n), y = n),fill = 'skyblue')+
  coord_flip()+
  xlab('Token')+
  ylab('Count')+
  labs(title = 'Postive Token Frequency')

neg <- ggplot(sen_bing_n_neg)+
  geom_col(aes(x = reorder(token,n), y = n),fill ='tomato')+
  coord_flip()+
  xlab('Token')+
  ylab('Count')+
  labs(title = 'Negative Token Frequency')

plot_grid(pos, neg, labels = "AUTO")
```



c iii Apparently, clients value their time and personal belongs so they don't like any delay of their flights as well as the lost of broken of their package. There is no ovbious clue on what customers likes as most positive tokens only express the emotions.



2d

Q3

3.a

```
summary(mentions$delta_reply_favorite_count)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.0000  0.0000  0.2837  0.0000 26.0000    2275
```

The average, median and maximum of the favorites for dealta's replies number is 0.28, 0, and 26

3.b

```
mean(mentions$delta_responded)
```

```
## [1] 0.2895066
```

The response rates of Dealta is 29.0%

3.c

```
mentions %>% mutate(response_time = (as.numeric(delta_reply_created_at - created_at)/60)) %>% summarise
```

```
##    avg_min median_min max_min
## 1 7.541208   4.583333    76.1
```

The average time for Delta to response is 7.54 minutes. median time for Delta to response is 4.58 minutes. max time for Delta to response is 76.1 minutes.

3.d Reply favorite count benefit: Reply favorite count can measure the reply quality. We could know whether the public are satisfied with our response. limitation: As most people don't engage, data is limited and biased.

Response rate Benefit: It measures generally how many people get our services. Limitation: we don't know the response content quality at all.

Response Time Benefit: It can measure the efficiency of our service. Limitation: same as the response rate, we have no clue on the quality of the response content.

Q4

a.i

```
df <- mentions %>% select(followers_count,favorite_count,retweet_count,verified,delta_responded)
df$verified <- as.integer(df$verified)
df$delta_responded <- as.integer(df$delta_responded)
df$delta_responded <- as.factor(df$delta_responded)

mylogit <- glm(delta_responded ~ followers_count + favorite_count + retweet_count + verified,
               data = df, family = "binomial")
summary(mylogit)
```

```
##
## Call:
## glm(formula = delta_responded ~ followers_count + favorite_count +
##     retweet_count + verified, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8971  -0.8428  -0.8405   1.5537   2.2343
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -8.523e-01  4.023e-02 -21.187  < 2e-16 ***
```

```
## followers_count  4.961e-07  7.369e-07   0.673 0.500793
## favorite_count  -6.638e-03  5.072e-03  -1.309 0.190605
## retweet_count    1.843e-02  2.522e-02   0.731 0.464936
## verified        -8.258e-01  2.288e-01  -3.609 0.000307 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3853.3  on 3201  degrees of freedom
## Residual deviance: 3831.9  on 3197  degrees of freedom
## AIC: 3841.9
##
## Number of Fisher Scoring iterations: 5
```

a.ii As the summary shows, followers count, favorite count, retweet_count all seems insignificant to the response, while verified is the most significant factor contributing to the response among this 4 factors, which is a little suprising to me as it shows negative relationhip with response.

regression model can only shows linear relationship, while the factors may not effect

b.i

```
replies <- replies %>% mutate(tactic_dm=ifelse(grepl('DM|private message',text,ignore.case = T),1,0))
mean(replies$tactic_dm )
```

```
## [1] 0.3294011
```

32.94% of replies contain DM or private message.

b.ii 1.Delta don't want negative emotion spread among the public hence they prefer the private message. 2.Delta don't want followers know more about their tailored solution and ride on the rules. For example, if they provide some free drink tickets for delayed customers. they will not happy that public know this. 3.Some information Delta required is related to personal information. Direct message could better protect customers' privacy.

Q5

a

```
replies$text <- gsub('http\\S+\\s*', '', replies$text)
replies$employee <- with(replies,substr(text,nchar(text)-3,nchar(text)))
replies$employee <- gsub(" ","",replies$employee)
replies$employee <- gsub("*","",replies$employee,fixed = T)
replies$employee <- gsub('\n',"",replies$employee)
unique(replies$employee)
```

```
##  [1] "HNH" "HSG" "HBD" "HYC" "HWG" "HAV" "HRB" "HMD" "HCA" "HKS" "HDB"
## [12] "HJW" "HSH" "HPF" "HFG" "HJH" "HCJ" "HQB" "HOS" "BCG" "BAW" "HTH"
## [23] "HBN" "HCW" "HJF" "BAM" "HLA" "HSK" "HEC" "HDV" "BAV" "HAL" "HKC"
## [34] "HJC" "HML" "BMC" "HEM" "BAY" "HJZ" "BTB" "HMB" "HAA" "HAC" "BDM"
## [45] "BDS" "BMS" "HRS" "BTA" "HBB" "BCN" "TSR" "HRO" "HSL" "BBW" "BJL"
## [56] "HAN" "BAF" "BLC" "HCM" "HGG" "BAS" "ACB" "BSJ" "HDR" "AJA" "TDL"
## [67] "BAP" "BAH" "ADH" "TMU" "TSL" "TMS" "TPB" "TMG" "TAC" "TLT"
```

```
length(unique(replies$employee))
```

```
## [1] 76
```

76 emplyees in total

5.b

```
top5 <- replies %>% count(employee,sort = T) %>% top_n(5)
```

```
## Selecting by n
```

```
sum(top5$n)/nrow(replies)
```

```
## [1] 0.2554446
```

Top 5 employees contribute to 25.54% total replies.

5.c By knowing which employees reply which tweets. Delta can easily measure the KPI for employees. Also, it is to train employees in individual level.