BA860 Assignment 4

Qiaoling Huang (U20421641)

2/24/2020

Paired with Salina(Ziqin Ma)

```
load("Downloads/Delta_social_media-W20-MSBA-MW.RData")
```

1_a: What is the average number of daily replies (in replies data)?

```
g<-replies %>% mutate(Date_create = format(created_at, "%Y-%m_%d")) %>% group_by(Date_create) %>% summa
mean(g$n)
```

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

Answer: the average number of daily replies is 314.8571.

1_b: What is the average number of daily mentions (in mentions data)?

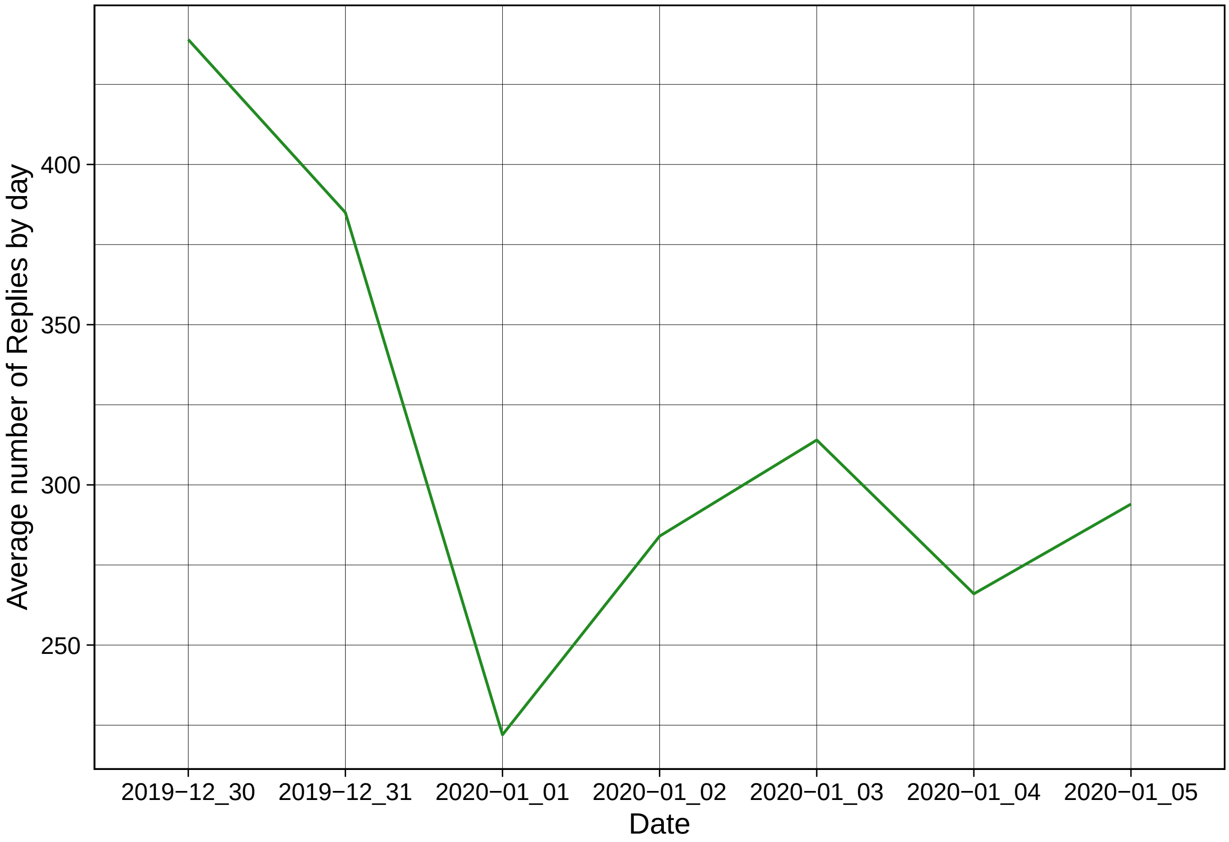
```
m<-mentions %>% mutate(Date_create = format(created_at,"%Y-%m_%d")) %>% group_by(Date_create) %>% summa
mean(m$n)
```

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

Answer: the average number of daily replies is 457.4286.

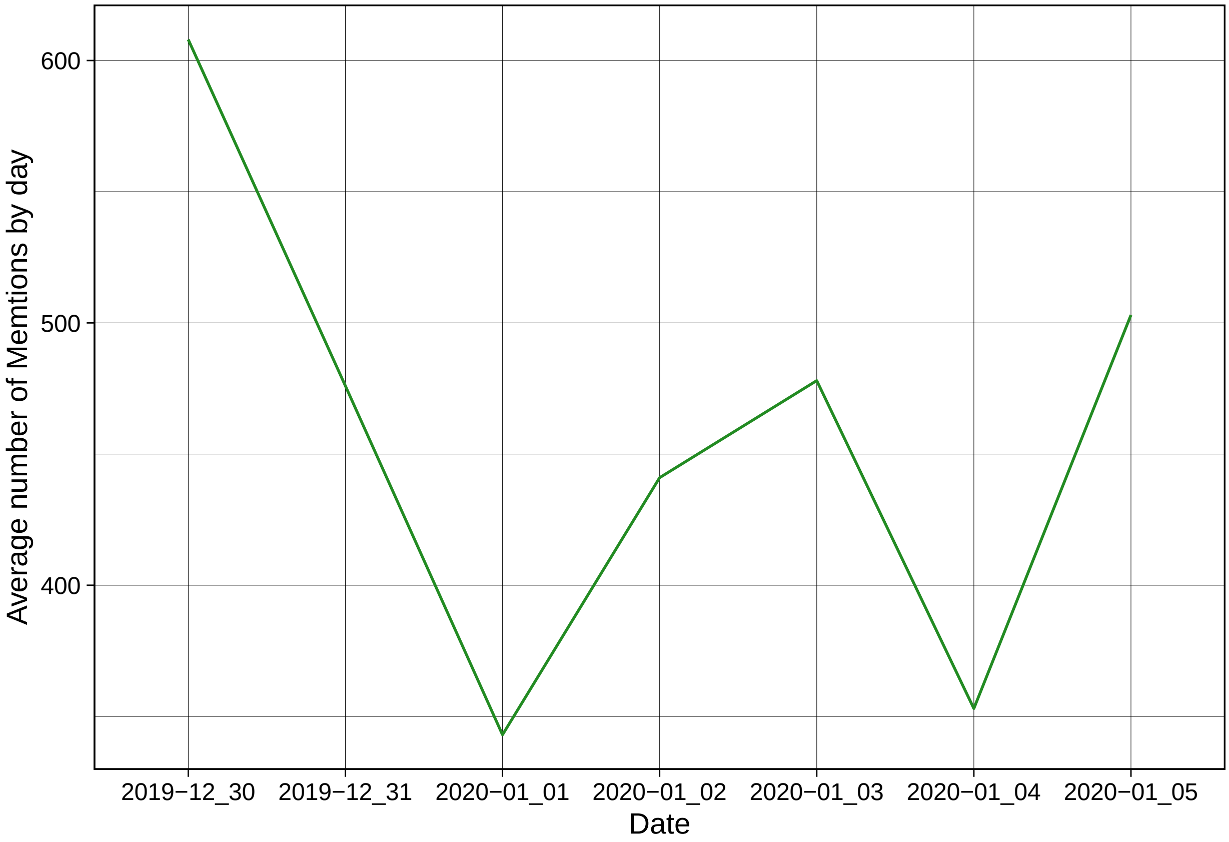
1_c: Using a line chart, plot the number of replies by day (in replies data).

```
g %>% ggplot(aes(Date_create, n, group = 1))+
  geom_line(colour = "forestgreen")+
  xlab("Date")+
  ylab("Average number of Replies by day")+
  theme_linedraw()
```



1_d:Using a line chart, plot the number of mentions by day (in mentions data).

```
m %>% ggplot(aes(Date_create, n ,group = 1))+
  geom_line(colour = "forestgreen")+
  xlab("Date")+
  ylab("Average number of Memtions by day")+
  theme_linedraw()
```



2_a:User's number of followers (followers_count). Note that users may appear multiple times in the mentions data. For part (a), examine distinct users: that is, pivot the data by user rather than by tweet.

i:By unique user, what is the median number of followers in Delta's mentions?

```
c<-mentions %>% group_by(user_id) %>% summarise(followers_by_users = median(followers_count))
median(c$followers_by_users)
```

```
## [1] 325
```

Answer: the median number of followers in Delta's mentions is 325 by unique user.

ii:Among unique users who mention Delta, what is the screen name of the user with the #3 most followers?

```
d<-c %>% mutate(followers_by_users = sort(followers_by_users, decreasing = T))
e<-mentions %>% group_by(screen_name) %>% summarise(followers_by_users = sum(followers_count))
length(d$user_id) == length(e$screen_name)
```

```
## [1] TRUE
```

```
left_join(d,e)
```

```
## Joining, by = "followers_by_users"
```

```
## # A tibble: 7,063 x 3
##    user_id             followers_by_users screen_name
##    <chr>                               <dbl> <chr>
##  1 1000727934393638912               2007122 <NA>
##  2 1001797649685778433               1456024 <NA>
##  3 1003010232979738624               1045514 ajc
##  4 100393969                          774091 KTLA
##  5 100527340                          412892 benhiggi
##  6 1006297513329156097                395595 willam
##  7 1007104583582093312                380445 <NA>
##  8 1007413741                         329732 mrjaxtaylor
##  9 1009183002318995456                296919 JaValeMcGee
## 10 101093155                          238986 FOX5Vegas
## # ... with 7,053 more rows
```

Answer: the screen name of the user with the #3 most followers are NYRangers, jdickerson, ajc.

2_b: We now examine the engagement that the mention tweets receive in terms of the number of favorites/likes (favorite_count).

i: What is the average & maximum number of favorites by mention?

```
## check if each row include hashtag delta
mentions$text <- tolower(mentions$text)
g<-mentions %>% filter(str_detect(text, "@delta"))
nrow(g) == nrow(mentions)
```

```
## [1] TRUE
```

```
## calculate the average and max number of favorite count
mentions %>% filter(str_detect(text, "@delta")) %>% summarise(avg_favorite = mean(favorite_count),
                                                              max_favorite = max(favorite_count))
```

```
##   avg_favorite max_favorite
## 1     3.271081          994
```

Answer: the average number of favorites by mention is 3.271081, and the maximum number of favorites by mention is 994.

ii:What is the text of the mention that receives the highest number of favorites?

```
mentions %>% summarise(max_favorite = max(favorite_count))
```

```
##   max_favorite
## 1          994
```

Answer: the text of the mention that receives the highest number of favorites is 994.

c: We wish to better understand the reasons why customers reach out to Delta for social care by analyzing the text content of Delta's mentions. In part (c), you must only examine the content of the tweets that get a response from Delta (delta_responded ==TRUE). Before analyzing the word content (of tweets that get a response), you must clean the text data in several cleaning steps: • Remove web links (http elements) • Extract words (remove punctuation, convert to lowercase) • Remove stop words (e.g. "the", "a") • Classify words by positive vs. negative sentiment.

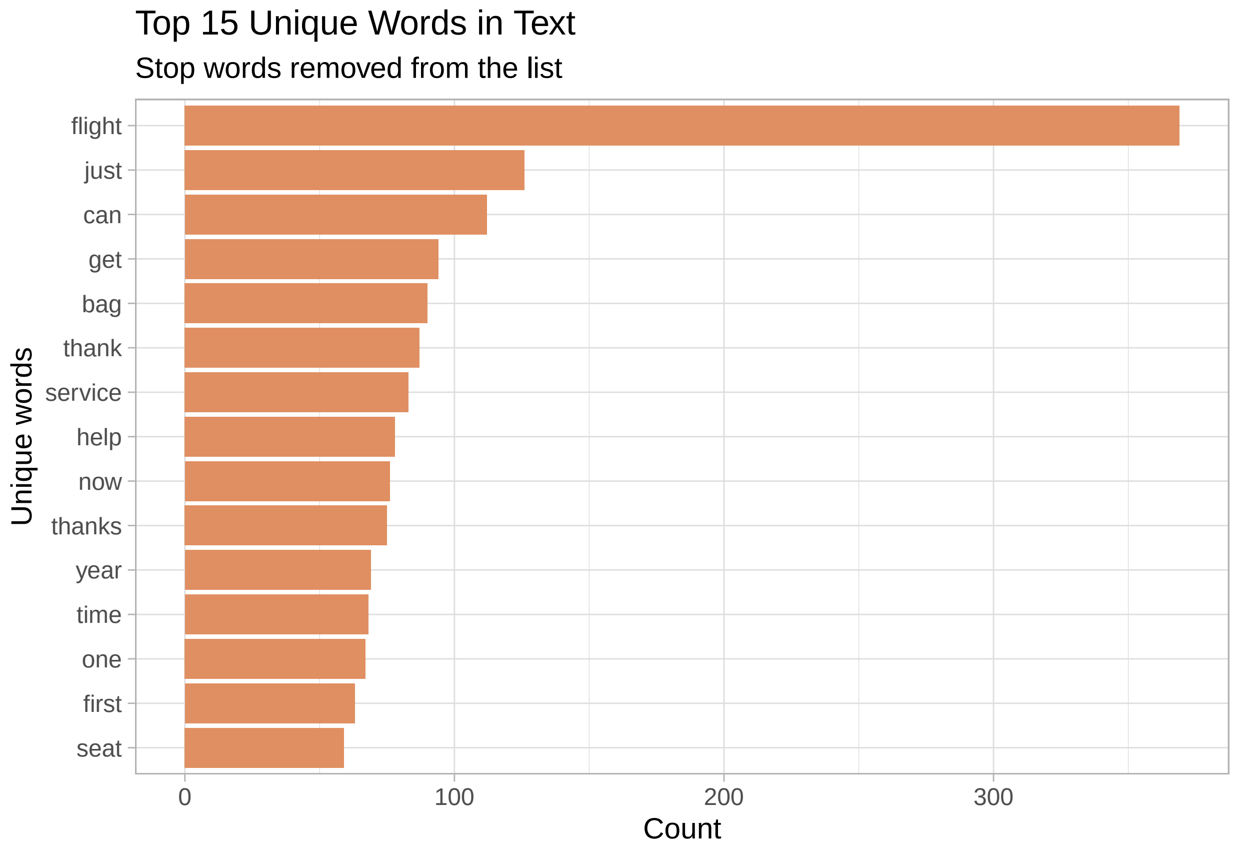
```r
## extract the text from mentions
delta_resp<-mentions %>% filter(delta_responded == TRUE)

## remove web link
delta_resp$text <-gsub("http.*", "", delta_resp$text)
delta_resp$text <-gsub("https.*", "", delta_resp$text)

## untokenize text and remove punctuation, stop words
tidy_mentions <-delta_resp %>% select(status_id, user_id, text) %>%
  unnest_tokens(token, text, "token" = "words", strip_punct = T) %>%
  anti_join(get_stopwords(), by = c("token" = "word"))

## classify words by sentiment
sent_bing<-inner_join(tidy_mentions,
                      get_sentiments("bing"),
                      by = c("token" = "word"))
head(sent_bing,20)
```

```
##              status_id             user_id       token sentiment
## 1  1211438403025289216            17003317     problem  negative
## 2  1211438431383031808          2988350667       delay  negative
## 3  1211438431383031808          2988350667      better  positive
## 4  1211439685790306310           628877003       delay  negative
## 5  1211441133114593281          2196028641      missed  negative
## 6  1211441133114593281          2196028641  unbearable  negative
## 7  1211441133114593281          2196028641      helped  positive
## 8  1211441133114593281          2196028641  incredibly  positive
## 9  1211445045762543617           316790027       delay  negative
## 10 1211445045762543617           316790027      faulty  negative
## 11 1211445045762543617           316790027       right  positive
## 12 1211446660133351424           303596644         mad  negative
## 13 1211447918550540288           210159397    complain  negative
## 14 1211447918550540288           210159397       great  positive
## 15 1211448411729338369           255809120        nice  positive
## 16 1211450315251339269           101328325     delayed  negative
## 17 1211452178034036736           139069140     problem  negative
## 18 1211452178034036736           139069140    problems  negative
## 19 1211456346996006912 1181603236748386304 handicapped  negative
## 20 1211456346996006912 1181603236748386304    delaying  negative
```

i: Using a bar chart, plot the top 15 unique words (excluding "delta") and their frequency. For an R example, see the Margaret Wanjiru medium article.

```r
## plot the top 10 words
tidy_mentions%>%
  count(token, sort = T) %>%
  filter(token != "delta") %>%
  top_n(15) %>%
  mutate(token = reorder(token, n)) %>%
  ggplot(aes(token, n))+
  geom_col(fill = "#e08f62")+
  coord_flip()+
  theme_light()+
  labs(y = "Count",
       x = "Unique words",
       title = "Top 15 Unique Words in Text",
       subtitle = "Stop words removed from the list")
```

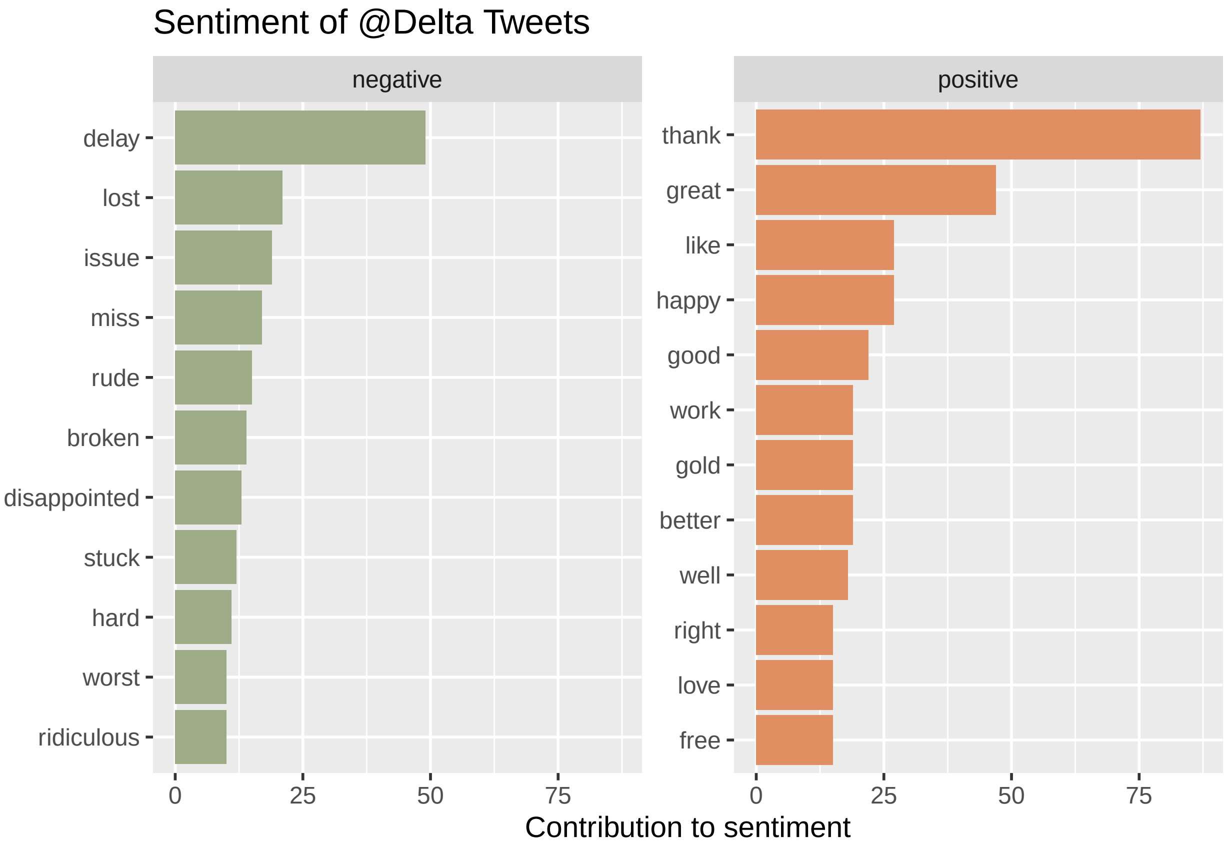
```
## Selecting by n
```



ii: Using two bar charts (side-by-side), plot the top 10 (excluding "delta") negative versus positive sentiment words and their frequency.

```
## extract token and sentiment for plot
a <- sent_bing %>% select(token, sentiment)
## delete duplicated word for top 10 words which is dupicated "delay"
a<-a %>% filter(token!="delayed")
a<-a %>% filter(token!="delays")
a<-a %>% filter(token!="issues")
a<-a %>% filter(token!="missed")

## plot two bar charts
a %>% count(token, sentiment, sort = T) %>%
  ungroup() %>%
  group_by(sentiment) %>%
  top_n(10) %>%
  ungroup() %>%
  mutate(token = reorder(token, n)) %>%
  ggplot(aes(token, n, fill = sentiment))+
  geom_col(show.legend = F)+
  facet_wrap(~sentiment,scales = "free_y")+
  scale_fill_manual(values = c("#9dab86", "#e08f62"))+
  labs(title = "Sentiment of @Delta Tweets",
       y = "Contribution to sentiment",
       x = NULL)+
  coord_flip()
```

```
## Selecting by n
```



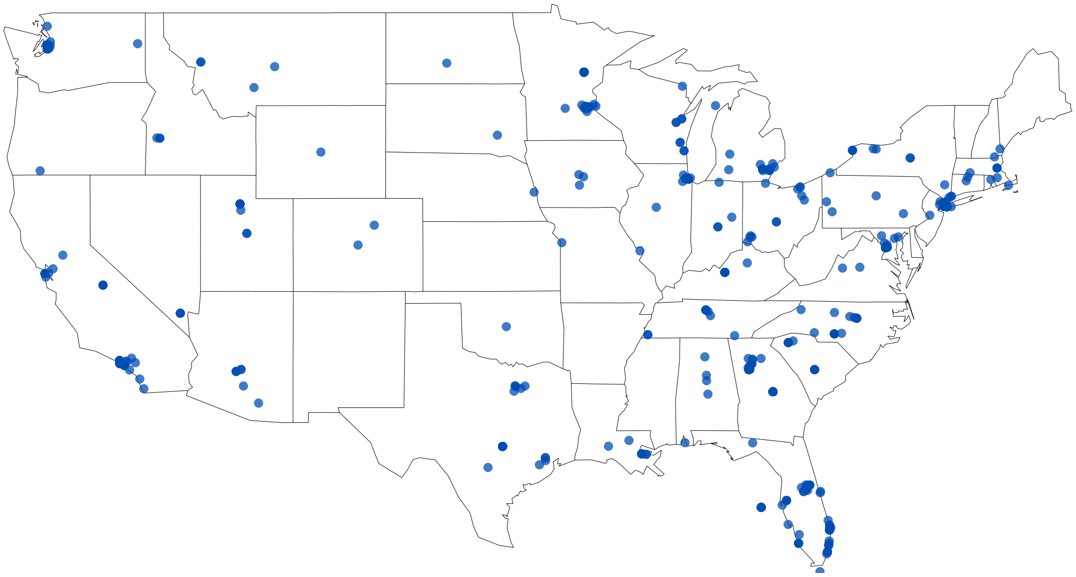
iii:What do you conclude are some of the main recurring customer issues in the mention data?

Answer: Based on the Delta response data, positive sentiment is more than negative which means Delta response mostly in positive sentiment text. Delta also response negative contents which are related to packages loss, missed filghts, and employee being rude things like that. In the social media platform, content can go viral so Delta responds more on positive content in order to maintain a positive brand image.

2_d: One nice feature of twitter data is that some users share the location that they are tweeting from. Map the location of Delta's mentions in the United States.

```
## filter mentions in the US
s<-mentions %>% filter(country_code == "US") %>%
  select(country_code, coords_coords, geo_coords, bbox_coords)
## extract location from data
rt <- lat_lng(s)

## plot state boundaries
```

```
par(mar = c(0, 0, 0, 0))
maps::map("state", lwd = .25)

## plot lat and lng points onto state map
with(rt, points(lng, lat, pch = 20, cex = .75, col = rgb(0, .3, .7, .75)))
```



3: A virtue of social care is the relative ease of quantifying customer service success. Compared to other customer service channels (e.g. in-person or phone), several quantitative metrics of customer success are readily available on social media. Below, we consider three such metrics: engagement, response rate, and response time. To keep a consistent data across these three metrics, we focus on the mention data.

3_a: What is the average, median, and maximum in engagement in terms of the number of favorites for Delta's replies (delta_reply_favorite_count)?

```
avg_engagement<-mean(mentions$delta_reply_favorite_count, na.rm = T)
median_engagement<-median(mentions$delta_reply_favorite_count, na.rm = T)
max_engagement<-max(mentions$delta_reply_favorite_count, na.rm = T)
data.frame(avg_engagement,median_engagement,max_engagement)
```

```
##   avg_engagement median_engagement max_engagement
## 1      0.2837109                 0             26
```

Answer: the average in engagement is 0.2837109, the median is 0, and the maximum is 26.

3_b:What is Delta's response rates (delta_responded) to its mentions (in percentage)?

```
resp<-subset(mentions, delta_responded == "TRUE")
## Calculate Delta's response rates
resp_rate<-nrow(resp)/nrow(mentions)
resp_rate
```

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

Answer: Delta's response rate is 0.2895066.

3_c:What is the average, median, and maximum response time

```
## Calculate the different between tweet created and response created
diff<-difftime(resp$delta_reply_created_at, resp$created_at, units = "mins")
avg_resp_time <- mean(diff)
median_resp_time<-median(diff)
max_resp_time<-max(diff)
data.frame(avg_resp_time,median_resp_time,max_resp_time)
```

```
##   avg_resp_time median_resp_time max_resp_time
## 1 7.541208 mins    4.583333 mins     76.1 mins
```

Answer: the average response time is 7.541208 mins, median response time is 4.583333 mins, and the maximum response time is 76.1 mins.

3_d:Provide both one strength and one limitation for using each of these customer success metrics.

Answer: 1: Engagement metric: the strength for using engagement metric is to allow company to track how many people engage on the company's responses and to get a sense of what kind issues peole most care about. A limitation is that company cannot know whether their response is good response or bad response for their audience based on the engagement metric. 2: Response rate metric: the strength for using response rate is to allow company to track how many mentions that the company has responded, but a limitation is that company cannot know what kind of content that the company has responded, if the company has responded to the right content. So, based on response rate is hard to track how effective that company has responded. 3: Response time metric: the strength for using response time is to allow company to track how quickly they access to the mentions, if it's efficent enough. A limitation is same as response rate metric, company cannot know what kind of mentions has responded, if the company address to the right content quickly enough.

4: To deliver effective social care, Delta needs to plan when and how to respond to its mentions.

4_a: What dictates which tweets get a response? For part (a), use the mentions data.

i.Use a linear probability model to explore this question. That is, regress an indicator of whether Delta replies on the following explanatory variables: followers_count, favorite_count, retweet_count, & verified (convert TRUE/FALSE variables to a 0/1 indicator variables if necessary). Provide the regression output (i.e. coefficient estimates, standard errors, t statistics or p-values).

```
## convert verified variable to 1 and 0
mentions$verified[mentions$verified == TRUE] <- 1
mentions$verified[mentions$verified == FALSE] <- 0
mentions$delta_responded[mentions$delta_responded == TRUE] <- 1
mentions$delta_responded[mentions$delta_responded == FALSE] <- 0
model<-lm(delta_responded ~ followers_count + favorite_count + retweet_count + verified, data = mentions
summary(model)
```

```
##
## Call:
## lm(formula = delta_responded ~ followers_count + favorite_count +
##     retweet_count + verified, data = mentions)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3141 -0.2978 -0.2973  0.7022  0.8863
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.977e-01  8.239e-03  36.136  < 2e-16 ***
## followers_count  3.950e-08  1.188e-07   0.333 0.739517
## favorite_count  -4.233e-04  4.605e-04  -0.919 0.358084
## retweet_count    1.496e-03  4.395e-03   0.340 0.733548
## verified        -1.404e-01  3.781e-02  -3.714 0.000208 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4526 on 3197 degrees of freedom
## Multiple R-squared:  0.005524,   Adjusted R-squared:  0.00428
## F-statistic:  4.44 on 4 and 3197 DF,  p-value: 0.001402
```

ii: Summarize the regression output. Be sure to mention any limitations of the model you feel are important.

Answer: The p-value for the model is 0.001402 which is less than 0.05 meaning it's statistical significant. However, each variables are not significant except 'verified'. The R-square is 0.00428 which is too small and less than 0.01, which means the regression model cannot explain much of the variance and it may not be valid. lastly, the standard error is 0.4526 which is very high meaning the sample data may not reflect the true population.

4_b: One decision in social care is whether to engage the customer publicly or privately via direct message. For part (b), use the replies dataset. i: What percent of delta's replies direct the customer to a private conversation? *As an indicator, define the dummy variable tactic_dm for whether or not Delta's reply contains "DM" or "private message" (ignoring case).*

```
## define the dummy variable tactic_dm
r<-replies %>% mutate(tactic_dm = ifelse(str_detect(text, "DM|private message"), 1,0))
## calculate the percentage
percent_to_direct_rep <- length(r$tactic_dm[r$tactic_dm == 1]) / length(r$tactic_dm)
percent_to_direct_rep
```

```
## [1] 0.3194192
```

Answer: the percent of delta's replies direct the customers to a private conversation is 0.3194192.

ii: Why would Delta wish to direct customers to a private conversation? Provide three reasons.

Answer: The first reason is that Delta doesn't want some sensitive content going viral. The second reason is Delta response customers directly can make customer feel the company care about their issues with Delta. The third one is one o one conversation is more effective because there is no other people involve.

5: Delta uses a team of social care employees to respond to Twitter mentions. Delta has a policy that each member of their team signs each tweet by ending it with their initials. First, construct an "employee" variable for the replies data that extracts the employee name from each tweet. This is the three capital letters at the end of the text.

5_a: How many different employees appear in the data?

```
## clean text data
reply<-replies %>% select(text) %>% mutate(text = str_remove_all(text, "[:punct:]+"),
                                         text = str_remove_all(text, "http*"),
                                         text = str_remove_all(text, " stco*"))
## extract last capital letter and create employee variable
reply$employee <- str_sub(replies$text, -3, -1)
## there are total 4 row missing employee name, replce "c2m" to NA
reply$employee[reply$employee == "c2m"] <- NA
```

```
## calculate the number of unique employee
length(unique(reply$employee))
```

```
## [1] 73
```

Answer: There are total 73 unique employee appear in the data.

5_b:What percentage of Delta's replies are written by the top five employees collectively?

```
top5<-reply %>% group_by(employee) %>%
      summarise(n = n()) %>%
      mutate(persentage = n/sum(n)) %>%
      arrange(desc(n)) %>%
      top_n(5)
```

```
## Selecting by persentage
```

```
## total percentage of Delta's replies are written by the top five employees collectively
sum(top5$persentage) * 100
```

```
## [1] 48.54809
```

Answer: The total percentage of Delta's replies written by the top five employees is 48.54809 which is almost half of total replies

5_c: Why would Delta want its employees to sign each tweet?

Answer: First, Delta can trace employees who handle on each tweet if something go wrong. Second, employee can follow up their replies and it's easy to follow the process of solving problems. Third, it's easy to meansure the efficiency of the employee in order to improve future service.