Social Media Analytics

Delta social care analytics

Sherry Zuo, Yishuang Song

2/27/2020

The data "Delta_social_media-TERM-SECTION" includes a subsample of real Twitter data relating to Delta's social care activity. We see both twitter user mentions of "Delta" as well as Delta's replies to users. A brief description of the individual datasets follows.

```
load("~/Desktop/BU/BA 860/Delta_social_media-W20-MSBA-TTh.RData")
```

```
dim(mentions)
```

```
## [1] 3169   96
```

```
dim(replies)
```

```
## [1] 2204   90
```

```
mentions<-data.table(mentions)
replies<-data.table(replies)
```

1. To begin, we want to understand the scale of Delta's social care activities.

a. (5 pts) What is the average number of daily replies (in replies data)?

```
replies$date<-date(replies$created_at)
nrow(replies)/length(unique(replies$date))
```

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

Answer: The average number of daily replies is **315**

b. (5 pts) What is the average number of daily mentions (in mentions data)?

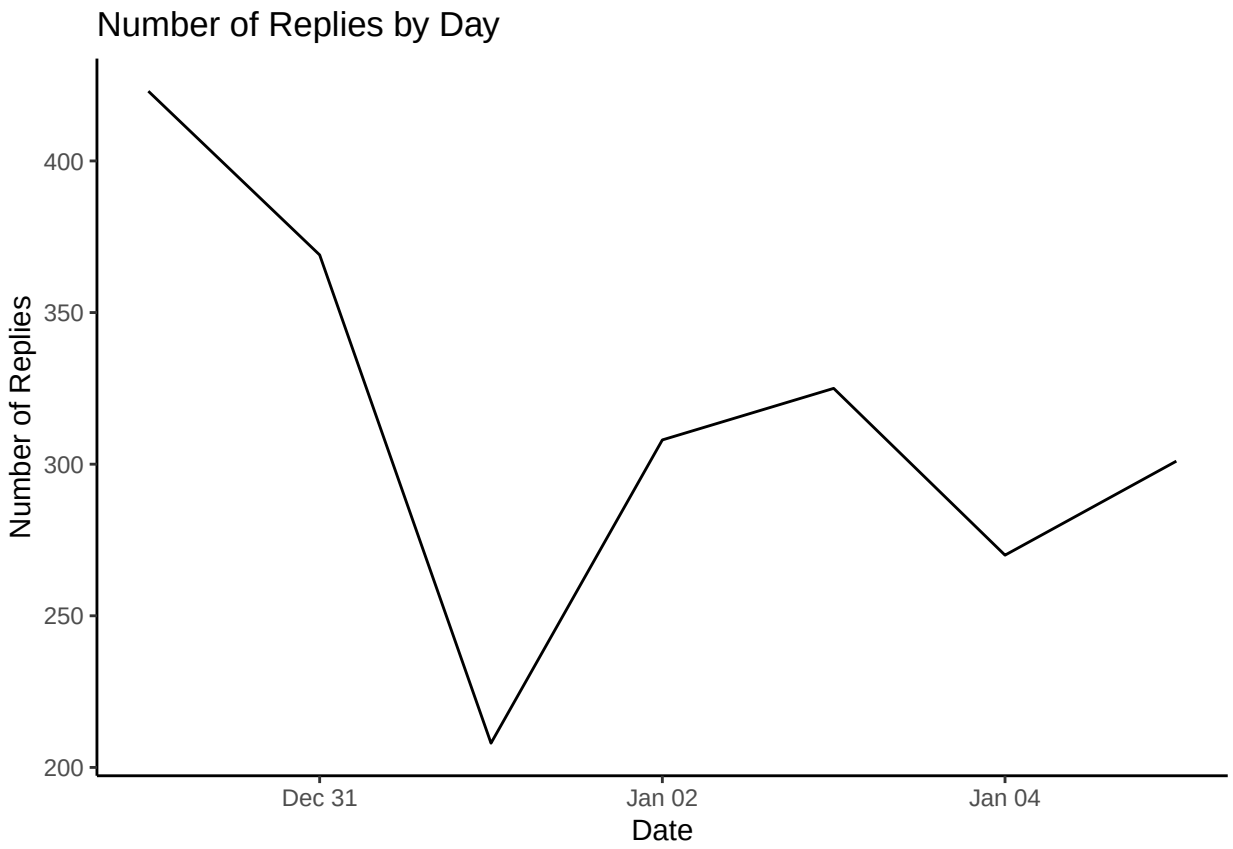
```
mentions$date<-date(mentions$created_at)
nrow(mentions)/length(unique(mentions$date))
```

```
## [1] 452.7143
```

Answer: The average number of daily mentions is **453**

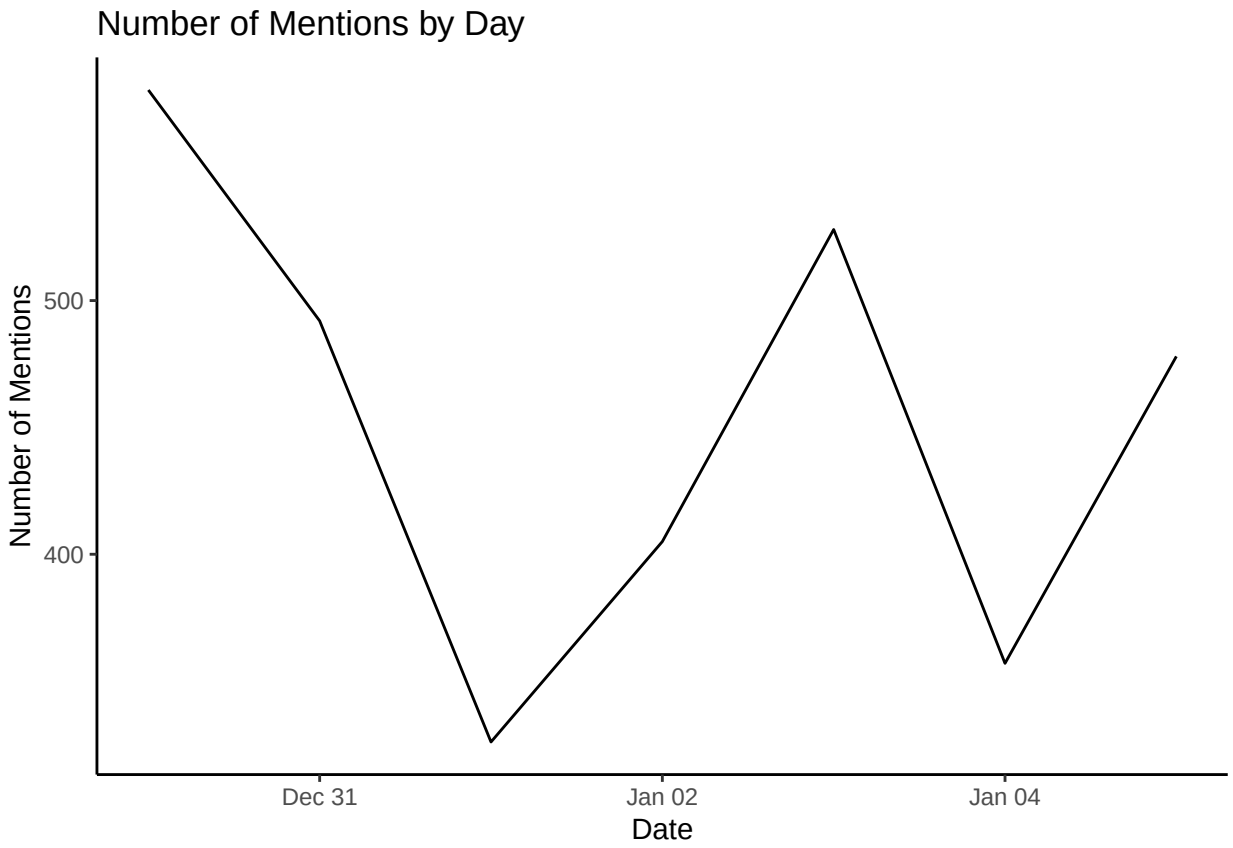
c. (10 pts) Using a line chart, plot the number of replies by day (in replies data).

```r
replies%>%
  group_by(date)%>%
  summarize(n_replies=n())%>%
  ggplot(aes(x=date, y=n_replies))+
  geom_line()+
  labs(title="Number of Replies by Day", x="Date", y="Number of Replies")+
  theme(panel.background = element_rect(fill="transparent"),
        axis.line = element_line(colour = "black"))
```



d. (10 pts) Using a line chart, plot the number of mentions by day (in mentions data).

```r
mentions%>%
  group_by(date)%>%
  summarize(n_mentions=n())%>%
  ggplot(aes(x=date, y=n_mentions))+
  geom_line()+
  labs(title="Number of Mentions by Day", x="Date", y="Number of Mentions")+
  theme(panel.background = element_rect(fill="transparent"),
        axis.line = element_line(colour = "black"))
```



2. Let's explore the mentions data.

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). (5 pts) By unique user, what is the median number of followers in Delta's mentions?

```
unique_user<-mentions[!duplicated(user_id)]
median(unique_user$followers_count)
```

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

Answer: By unique user, the median number of followers in Delta's mentions is **325**

ii). (5 pts) Among unique users who mention Delta, what is the screen name of the user with the #3 most followers?

```
desc_followers<-unique_user%>%
  select(user_id, followers_count, screen_name)%>%
  arrange(desc(followers_count))
desc_followers[3,"screen_name"]
```

```
## [1] "jdickerson"
```

Answer: Among unique users who mention Delta, the screen name of the user with the #3 most followers is **jdickerson**

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

i). (5 pts) What is the average & maximum number of favorites by mention?

```
mean(mentions$favorite_count)
```

```
## [1] 3.282108
```

```
max(mentions$favorite_count)
```

```
## [1] 569
```

Answer: The average number of favrites by mention is **3**, the maximum number of favorites by mention is **569**

ii). (5 pts) What is the text of the mention that receives the highest number of favorites?

```
high_favorites<-mentions%>%
  select(text,favorite_count)%>%
  arrange(desc(favorite_count))
cat(high_favorites[1,"text"])
```

```
## Hi @delta can you guys just put up little signs that say "please don't scroll the TV with your feet u
##
## This is where we are now this is 2020
```

Answer: The text of the mention that receives the highest number of favorites shows above.

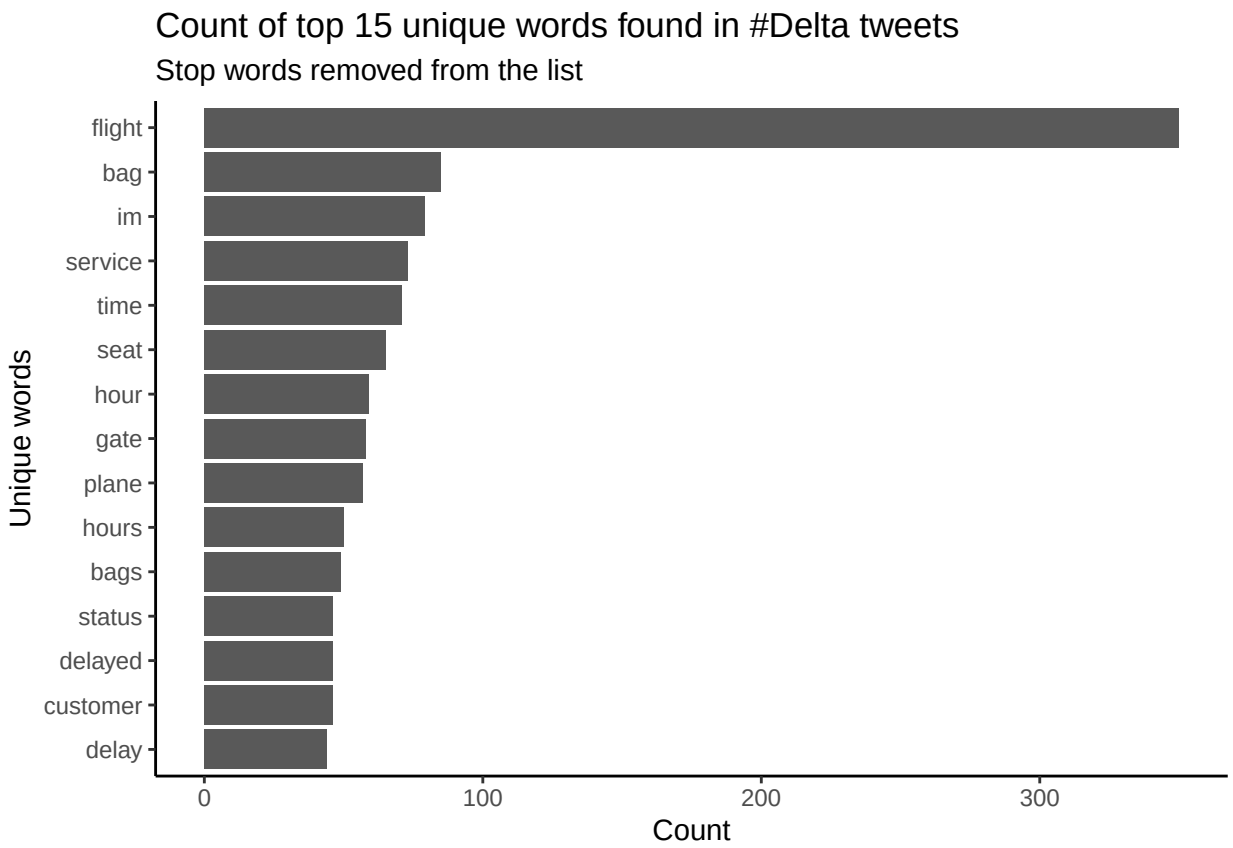
c. We wish to better understand the reasons why customers reach out to Delta for social care by analyzing the text content of Delta's mentions.
In part (c), you must only examine the content of the tweets that get a response from Delta (delta_responded ==TRUE). Before analyzing the word content (of tweets that get a response), you must clean the text data in several cleaning steps

```
tweets_get_response<-mentions%>%filter(delta_responded==TRUE)
```

```
tweets_get_response$text<-str_to_lower(tweets_get_response$text)
tweets_get_response$text <- gsub("@delta","",  tweets_get_response$text)
tweets_get_response$text <- gsub("delta","",  tweets_get_response$text)
tweets_get_response<-data.frame(tweets_get_response)
Delta_tweets_words <- tweets_get_response %>%
  select(text)%>%
  unnest_tokens(word, text, token="tweets",
                strip_url=TRUE,
                strip_punct=TRUE,
                )%>%
  anti_join(stop_words)
```

i). (10 pts) 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.

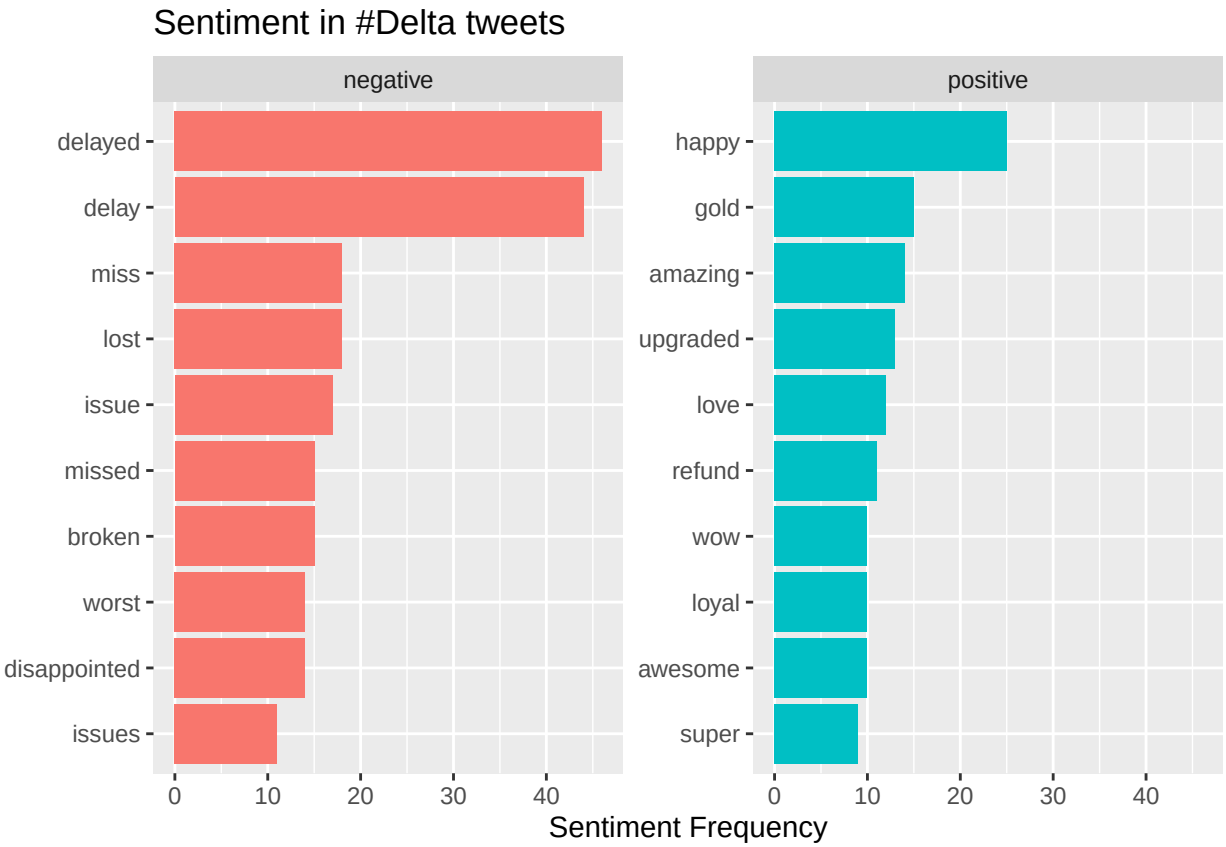
```
Delta_tweets_words %>%
  count(word, sort = TRUE) %>%
  top_n(15) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(x = word, y = n)) +
  geom_col() +
  xlab(NULL) +
  coord_flip() +
  labs(y = "Count",
       x = "Unique words",
       title = "Count of top 15 unique words found in #Delta tweets",
       subtitle = "Stop words removed from the list")+
  theme(panel.background = element_rect(fill="transparent"),
        axis.line = element_line(colour = "black"))
```



ii). (20 pts) Using two bar charts (side-by-side), plot the top 10 (excluding "delta") negative versus positive sentiment words and their frequency.

```
Delta_tweets_words%>%
  inner_join(get_sentiments("bing"))%>%
  count(word, sentiment, sort = TRUE) %>%
  ungroup()%>%
  group_by(sentiment) %>%
  top_n(10) %>%
```

```
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(title = "Sentiment in #Delta tweets",
       y = "Sentiment Frequency",
       x = NULL) +
  coord_flip()
```

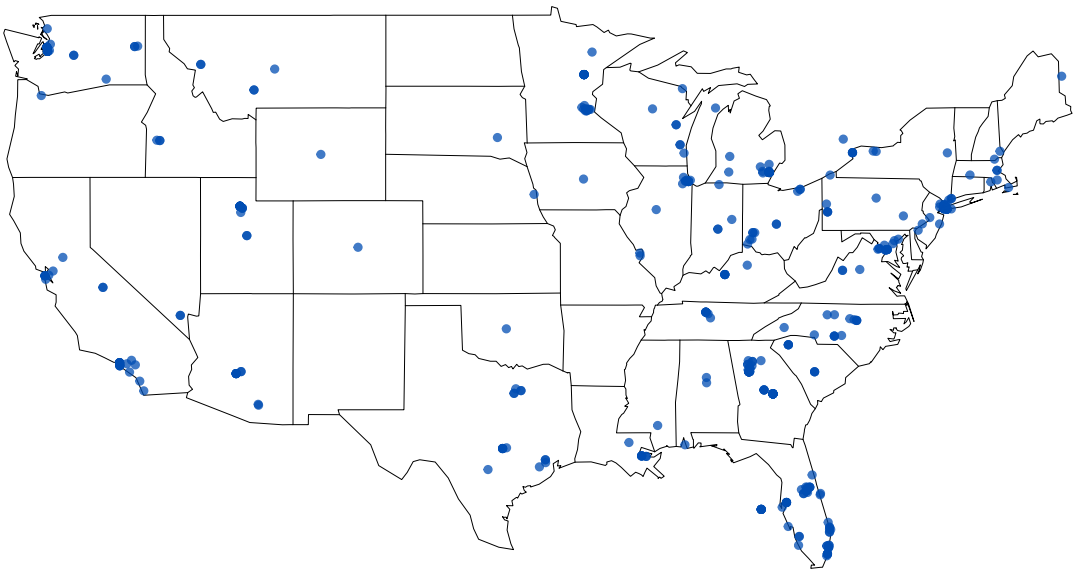


iii). (5 pts) What do you conclude are some of the main recurring customer issues in the mention data?
Answer: The main recurring customer issues in the mention data are problems of flights delay, missed flights, luggage broken, bag missing and lost packages.

d. (10 pts) 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.

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

## plot lat and lng points onto state map
with(mentions, 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.

   a. (5 pts) What is the average, median, and maximum in engagement in terms of the number of favorites for Delta's replies (delta_reply_favorite_count)?

```
mean(mentions$delta_reply_favorite_count, na.rm = T)
```

```
## [1] 0.2691441
```

```
median(mentions$delta_reply_favorite_count, na.rm = T)
```

```
## [1] 0
```

```
max(mentions$delta_reply_favorite_count, na.rm = T)
```

```
## [1] 18
```

Answer: The average in engagement in terms of the number of favorites for Delta's replies is **0.27**, the median is **0**, and the maximum is **18**.

   b. (5 pts) What is Delta's response rates (delta_responded) to its mentions (in percentage)?

```
mentions<-data.table(mentions)
mean(mentions$delta_responded)*100
```

```
## [1] 28.02146
```

Answer: Delta's response rates to its mentions is **28%**

c. (5 pts) What is the average, median, and maximum response time (delta_reply_created_at minus created_at) for Delta's replies in minutes?

```
mentions<-mentions%>%
  mutate(response_time=delta_reply_created_at-created_at)
mean(mentions$response_time, na.rm = T)
```

```
## Time difference of 461.9606 secs
```

```
median(mentions$response_time, na.rm = T)
```

```
## Time difference of 277 secs
```

```
max(mentions$response_time, na.rm = T)
```

```
## Time difference of 5170 secs
```

```
461.9606/60
```

```
## [1] 7.699343
```

```
277/60
```

```
## [1] 4.616667
```

```
5170/60
```

```
## [1] 86.16667
```

Answer: The average response time for Delta's replies in minutes is **7.7 minutes**, the median is **4.62 minutes**, and the maximum is **86.17 minutes**.

d. (15 pts) Provide both one strength and one limitation for using each of these customer success metrics.

Answer:
Metric Engagement:
Strength: It allows Delta to know how the public reacts to their replies of the tweets.
Limitation: Most of the time, people won't like/favorite official replies, therefore, they can't draw meaningful conclusions according to the engagement. Also, a lot of feedback comes directly from customers, within private messages, likes in Twitter do not represent most users.

Metric Response Rate:
Strength: It allows Delta to have an overall idea about how many tweets it replies and to what percentage the mentions are meaningful and worthy of replies.
Limitation: Indicates low response rate compared with other channels, and response rate (quantity) not means response quality.

Metric Response Time:
Strength: It allows Delta to measure the efficiency of tweets response, and to see whether or not they need to expand the team to shorten response time.
Limitation: It takes a long time to reply when the traffic is high, the response time cannot truly reflect and real-time responses, therefore it could not be a standard measurement at all the time. It only considered the time for response tweets, not for all tweets.

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

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

i). (10 pts) 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).

```
responses<-mentions%>%
  select(followers_count, favorite_count, retweet_count, verified, delta_responded)%>%
  mutate(verified=ifelse(verified==TRUE, 1, 0), delta_responded=ifelse(delta_responded==TRUE, 1, 0))
head(responses)
```

```
##   followers_count favorite_count retweet_count verified delta_responded
## 1              60              0             0        0               1
## 2             524              0             0        0               0
## 3             176              0             0        0               0
## 4            2255              2             0        0               0
## 5              73              0             0        0               0
## 6               2              0             0        0               0
```

```
summary(lm(delta_responded~followers_count+favorite_count+retweet_count+verified, data=responses))
```

```
##
## Call:
## lm(formula = delta_responded ~ followers_count + favorite_count +
##     retweet_count + verified, data = responses)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3551 -0.2894 -0.2886  0.7105  0.9223
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.894e-01  8.223e-03  35.193  < 2e-16 ***
## followers_count  4.483e-08  4.828e-08   0.928 0.353264
## favorite_count  -7.681e-04  4.927e-04  -1.559 0.119090
## retweet_count    2.441e-03  4.289e-03   0.569 0.569282
## verified        -1.396e-01  3.613e-02  -3.864 0.000114 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.448 on 3164 degrees of freedom
## Multiple R-squared:  0.006391,   Adjusted R-squared:  0.005135
## F-statistic: 5.088 on 4 and 3164 DF,  p-value: 0.0004372
```

ii). (10 pts) Summarize the regression output. Be sure to mention any limitations of the model you feel are important.
Answer: The p-value is **0.0004372** which is less than 0.05, so we could reject the null-hypothesis, meaning that the linear regression is statistically significant. However, most of those independant variables are not significant except the variable 'verified', also we need to care about the adjusted R-square which is **0.005135** and is too small since it's less than 0.01, meaning that the proportion of the variance for a dependent variable that's explained by those independent variables in this linear regression model is very low, so it may not be vaild. And the standard error is **0.448**, which is very high and means that the data may have some notable irregularities.

Responses should highlight the sign/direction and statistical significance of the relationships. Some model limitations include:
- We lack information on tweet content to infer whether Delta should respond to the tweet.
- Several variables (e.g. follower count) have very large outliers that would have outsized influence in the regression.

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). (5 pts) 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).

```
private_conversation<-replies%>%
  select(text)%>%
  mutate(tactic_dm=ifelse(grepl("private message|\\bDM\\b", ignore.case=T, text), 1, 0))
private_conversation<-data.table(private_conversation)
mean(private_conversation$tactic_dm)*100
```

```
## [1] 33.25771
```

Answer: **33.26%** of delta's replies direct the customer to a private conversation.

ii). (10 pts) Why would Delta wish to direct customers to a private conversation? Provide three reasons.
Answer:
1) To offer better assistance based on specific individual needs.
2) There are words limits in a single tweet, while direct private messages are longer, which allow customers to write down their request in more details than the standard 140 characters per tweet.
3) Protect customers privacy especially their personal information such as name, confirmation number, email address.

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.

a. (5 pts) How many different employees appear in the data?

```
replies$text<-gsub(' http\\S*',"", replies$text)
nchar(replies$text)->N
replies$employee<-substring(replies$text, N-2,N)
#row 1025 does not end with http or initials
replies$employee<-ifelse(str_detect(replies$employee,"^[:upper:]+$"), replies$employee, NA)
```

```
replies%>%
  distinct(employee)%>%
  na.omit()%>%
  tally(name="employees_count")
```

```
##   employees_count
## 1              74
```

Answer: There are **74** different employees appear in the data.

b. (5 pts) What percentage of Delta's replies are written by the top five employees collectively?

```
top5<-replies%>%
  group_by(employee, na.omit=T)%>%
  count()%>%
  arrange(desc(n))%>%
  head(5)%>%
  ungroup()%>%
  summarize(top5=sum(n))
top5*100/nrow(replies)
```

```
##       top5
## 1 26.17967
```

Answer: **26.18%** of Delta's replies are written by the top five employees collectively.

c. (5 pts) Why would Delta want its employees to sign each tweet?
Answer:
First, customers would be aware of the fact that their tweets are being handled and Delta is working on their problems to provide better services. Second, by signing each tweet, Delta would be able to trace back to each tweet and make further follow up to ensure consistency. Whenever there are unsolved issues or Delta wants to know who is responsible for a particular tweet problem, it would become easy to follow. Additionally, it's a better way of counting employee efficiency and improving future customer services.