Assignment 2

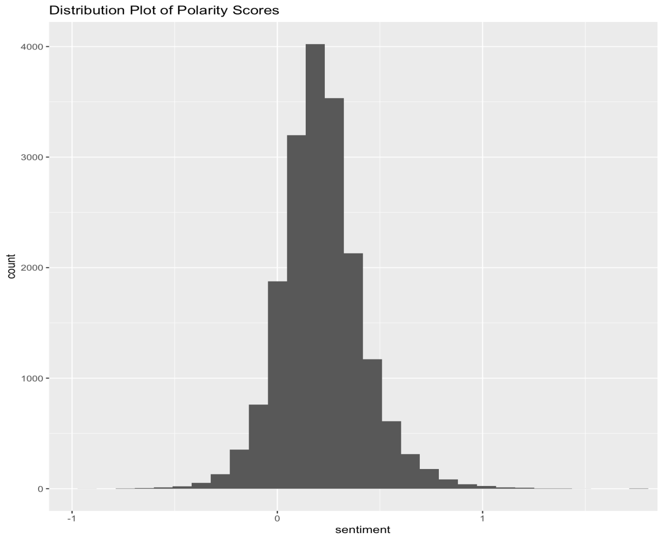
Mandi Zhu

To maximum the value of data about CRM, I am going to conduct sentiment analysis and explore a topic model to the customer reviews in this report and finally give my suggestions on its business based on valid insights from data.

- **Sentiment Analysis**

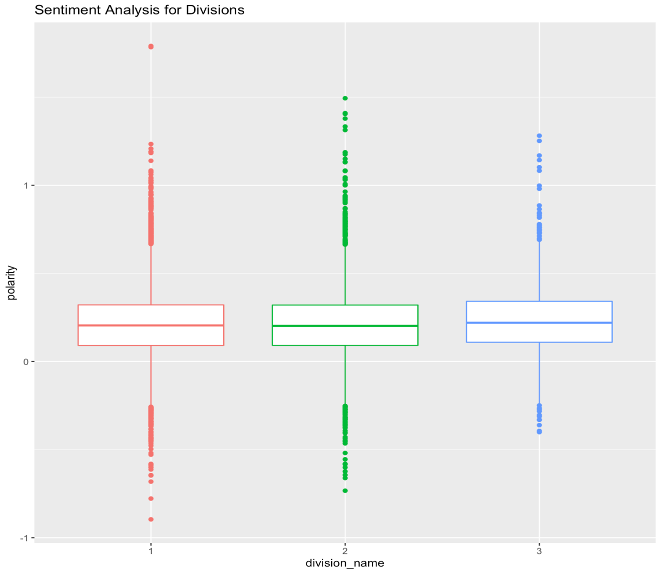
To conduct sentiment analysis, I use the "sentimentr" package and assign a polarity score for each text review given by a customer. I choose this method because I think it reflects customers' "real sentiment" better than some other methods (e.g. Bing tidy polarity). I measure customers' "real sentiment" by variable *ratings* in the dataset and I find the distributions of ratings and polarity scores for both departments and divisions are quite similar.

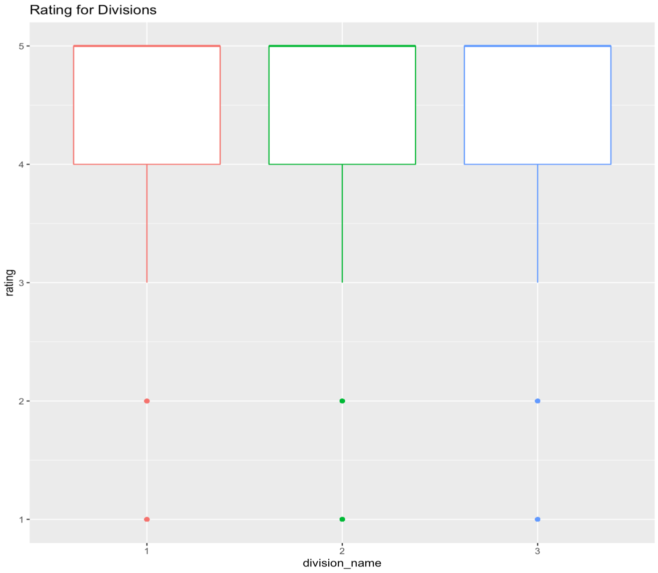
○ Distribution of sentiment scores overall



To get a general idea about how customers feel about the products, I draw a distribution plot of overall polarity scores, which is shown on the left**.** I also get the 25th, 50th and 75th percentile of polarity scores, which are 0.092, 0.205 and 0.322 respectively. From the plot and the percentiles**,** we know that majority of reviews given by customers imply positive sentiment and thus customers think positively and are generally satisfied with products they purchased.

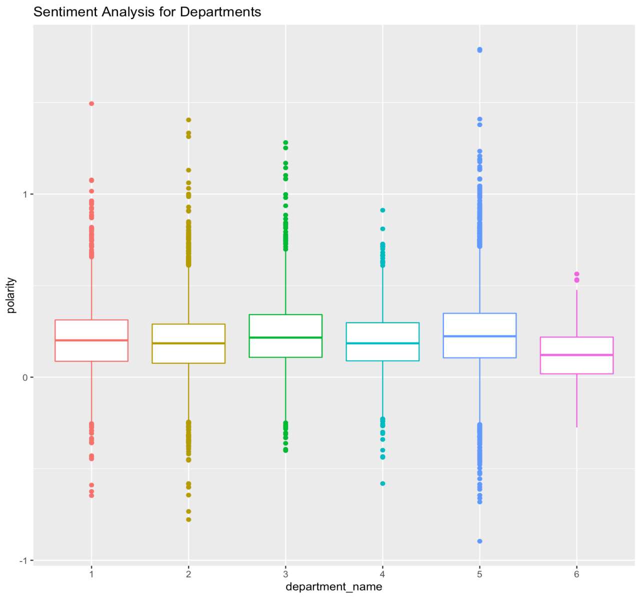
○ Sentiment analysis for divisions

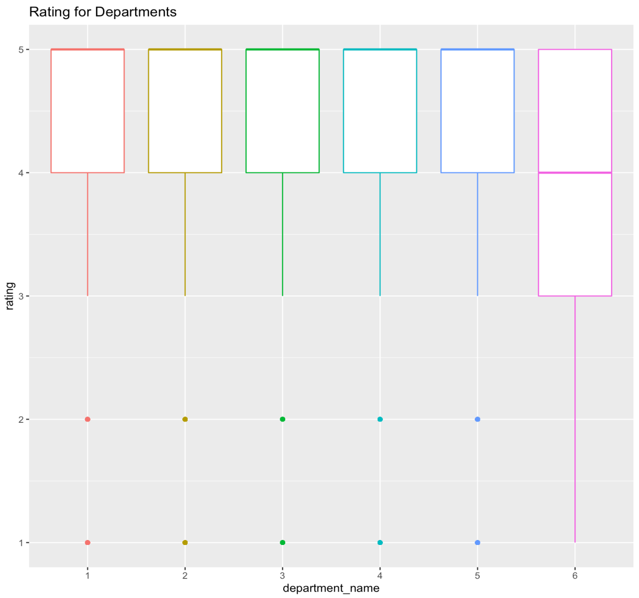




After the overview of customer sentiment, I conduct another sentiment analysis specifically for divisions. As shown in the above, three divisions do not show significant different performance on average according to both ratings and polarity scores, but Division 3 performs moderately better than the other two in terms of extreme negative review from customers.

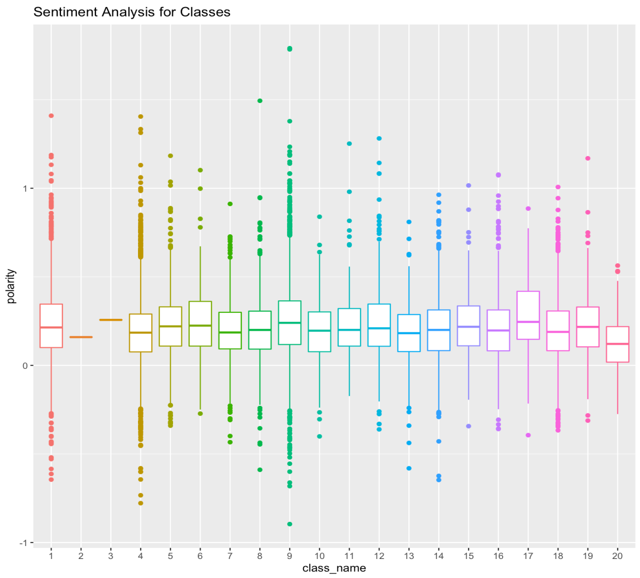
- ○ Sentiment analysis for departments

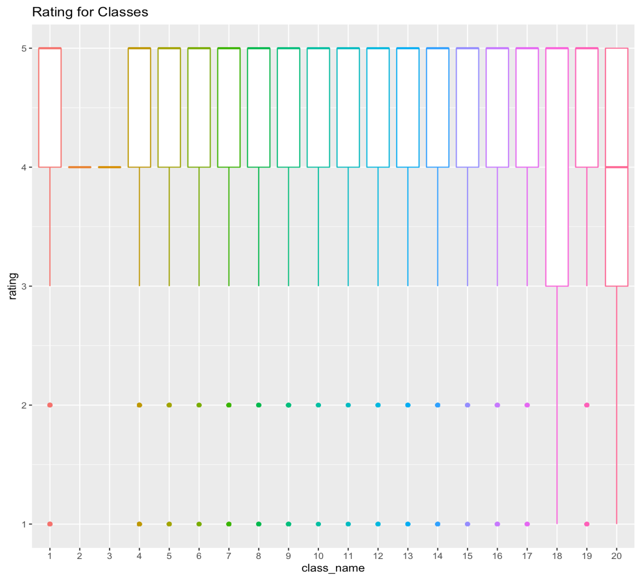




According to boxplots shown above, products from Department 6 obviously underperform in customer satisfaction. In addition, products from Department 2 and 5 get more extremely negative reviews from customers.

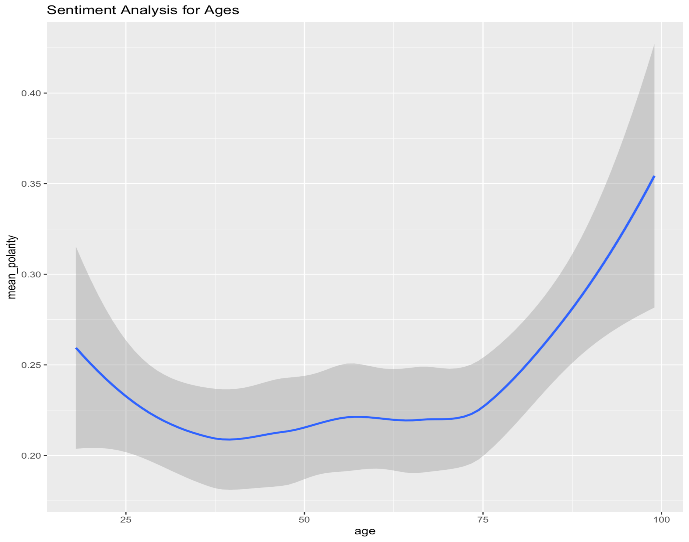
- ○ Sentiment analysis for classes





According to graphs above, customer satisfaction is apparently lower for products from Class 20 as well as Class 18 than for products from other classes. Besides, Class 4 and 9 are shown to receive more extremely negative reviews about their products than other classes.

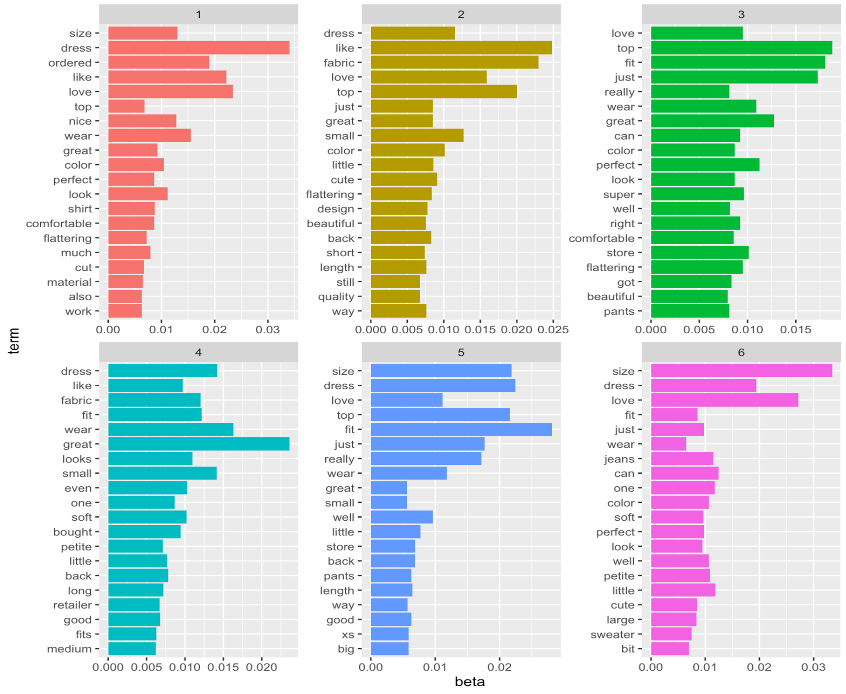
- Sentiment analysis for customers in different ages



The line graph on the right implies the online retailer has a relatively poor performance in satisfying customers aged from 35 to 50 because the mean polarity from this group of customers is low compared with mean polarity from others. It might be because quality of products sold to them does not meet their expectation as much as quality of products sold to other customers in different ages does. Also, it is likely that they have higher expectation on response quality from customer support staff.

- **A Topic Model**

After trying different number of topic to set up an effective topic model, I decide to use a model with k=6 because it performs better in terms of gamma values of documents overall.



The graph on the left shows the top 20 words in 6 topics. Here is my guess about topics based on the top features of topics:

--Topic 1: positive feedback about shirts
--Topic 2: positive feedback about shorts
--Topic 3: positive feedback about pants
--Topic 4: size and materials
--Topic 5: size about pants
--Topic 6: size about jeans or sweaters.

Based on the information from topic modelling, shirts, shorts, pants and jeans are likely to be popular products for customers.

- **Recommendations**

In summary, the online retailer has performed reasonably well in getting customer satisfaction overall based on the polarity scores. However, department 6 and class 20 obviously underperformed in managing their products and customer support teams so I suggest that they review and improve both product qualities and customer service. In addition, the retailer should put more efforts in investigating needs of customers aged between 35 and 50 and improve their service offered to these people accordingly. Last but not least, according to the topic modelling results, shirts, shorts and pants sold by the retailer seem to be much popular among customers compared with other products so the retailer can maintain its management on these products or even put more efforts on them to strengthen its current advantages over competitors.

Code:

```
library(tidyverse)
library(readxl)
library(factoextra)
library(tidytext)
library(wordcloud)
library(quanteda)
library(sentimentr)
library(topicmodels)
library(cluster)
library(streamgraph)
library(plotly)


##read data
raw_review<-readRDS('datasets/hw2.rds')
View(raw_review)
skimr::skim(raw_review) #no missing value; (18555,9)

#########EDA##########

#rating for each department
raw_review %>% group_by(department_name) %>%
  ggplot(aes(x=department_name,y=rating,color=department_name)) +
  geom_boxplot(show.legend = F) +
  ggtitle('Rating for Departments')   ##department 6 shows the worst performance among all deps.
#rating for each division
raw_review %>% group_by(division_name) %>%
  ggplot(aes(x=division_name,y=rating,color=division_name)) +
  geom_boxplot(show.legend = F) +
  ggtitle('Rating for Divisions')  ## 3 divisions are same
#rating for each class
raw_review %>% group_by(class_name) %>%
  ggplot(aes(x=class_name,y=rating,color=class_name)) +
  geom_boxplot(show.legend = F) +
  ggtitle('Rating for Classes')  ## class 18 and 20 obviously underperform compared with others
#how many products under each department
uni_pdt_dep<-raw_review %>% group_by(department_name) %>%
  summarise(uni_pdt=length(unique(product_id)))

#how many products under each division
uni_pdt_div<-raw_review %>% group_by(division_name) %>%
  summarise(uni_pdt=length(unique(product_id)))

#how many products under each class
uni_pdt_class<-raw_review %>% group_by(class_name) %>%
  summarise(uni_pdt=length(unique(product_id)))  #18 and 20

########sentiment analysis#########
```

```
#unnest text
review<-raw_review %>% select(c(crmid,review_text))

review_tidy<-review %>% unnest_tokens(token,review_text,token='words',strip_punct=T,strip_num=T)
head(review_tidy,20)
#remove stop words
sw=get_stopwords()
review_tidy_nosw<-review_tidy %>% anti_join(sw,by=c('token'='word'))
View(review_tidy_nosw)

############
#get sentiment for non-stopwords: sentiment r (sentence base)
text<-review$review_text %>% get_sentences() %>% sentiment()
head(text,20)

senti_afi_2<-text %>% group_by(element_id) %>% summarise(polarity=mean(sentiment))
nrow(senti_afi_2)==nrow(raw_review)

#what is it overall
skimr::skim(senti_afi_2)
#join onto the original data
review_senti2 = raw_review %>% mutate(polarity=senti_afi_2$polarity)
View(review_senti2)
#visualization for interesting patterns
length(unique(raw_review$product_id))   #1083/18555
unique(raw_review$division_name)  #1 2 3
unique(raw_review$department_name)  #1-6


  #distribution of the sentiment overall
ggplot(review_senti2, aes(x=polarity)) + geom_histogram(fill='black') +ggtitle('Distribution Plot of Polarity Scores')
quantile(review_senti2$polarity,probs = c(0.25,0.5,0.75))
postive_senti<-nrow(filter(review_senti2,polarity>0))/nrow(review_senti2)  # 0.8880086
  #sentiment vs age groups

        #ggplot(review_senti2,aes(x=age,y=polarity)) + geom_smooth()
review_senti2 %>% group_by(age) %>% summarise(mean_polarity=mean(polarity)) %>%
  ggplot(aes(x=age,y=mean_polarity)) + geom_smooth() +
  ggtitle('Sentiment Analysis for Ages')


  #sentiment vs devision
ggplot(review_senti2,aes(x=division_name,y=polarity)) +
  geom_boxplot(aes(color=division_name),show.legend = F) +
  ggtitle('Sentiment Analysis for Divisions')

  #sentiment vs department
```

```
ggplot(review_senti2,aes(x=department_name,y=polarity)) +
  geom_boxplot(aes(color=department_name),show.legend = F)+
  ggtitle('Sentiment Analysis for Departments')

  #sentiment vs class
ggplot(review_senti2,aes(x=class_name,y=polarity)) +
  geom_boxplot(aes(color=class_name),show.legend = F)+
  ggtitle('Sentiment Analysis for Classes')


########topic modelling#########
review_cps<-corpus(review$review_text)
summary(review_cps)
docvars(review_cps,'crmid')<-review$crmid
review_dfm<-dfm(review_cps,
                remove=get_stopwords()$word,
                remove_punct=T,
                remove_numbers=T,
                remove_symbols=T,
                remove_twitter=T,
                remove_url=T) %>%
  dfm_trim(min_termfreq = 10,
           termfreq_type = 'count',
           max_docfreq = 0.7,
           docfreq_type = 'prop')
review_m<-convert(review_dfm,'topicmodels')


review_lda<-LDA(review_m,k=6,control = list(seed=820))
summary(review_lda)
#view beta
review_beta = tidy(review_lda, matrix="beta")
head(review_beta,30)

review_top20 = review_beta %>%
  group_by(topic) %>%
  top_n(20, beta) %>%
  ungroup() %>%
  arrange(topic,desc(beta))
review_top20

#view tokens example for each topic
terms(review_lda,20)

review_top20 %>% mutate(term=reorder(term,beta)) %>%
  ggplot(aes(x=term,y=beta,fill=factor(topic))) +
  geom_col(show.legend = F)+
  facet_wrap(~topic,scales='free') +
  coord  flip()
```

```
#view gamma
review_gamma = tidy(review_lda, matrix="gamma")
review_gamma %>%
  arrange(-gamma) %>%
  print(n=30)


skimr::skim(fgamma)

ggplot(review_gamma,aes(x=gamma)) +
  geom_histogram() +
  facet_wrap(~topic)

   #review_gamma %>% group_by(document) %>% summarise(sum=sum(gamma))
##here is the loop to choose the optimal k#########
n_topic<-c()
n_doc<-c()
opt_k<-data.frame(n_topic=n_topic,n_doc=n_doc)
for (i in 2:10) {
  review_lda=LDA(review_m,k=i,control = list(seed=820))
  review_gamma = tidy(review_lda, matrix="gamma")
  ndoc=review_gamma %>%
    arrange(-gamma) %>%
    filter(gamma>1/i) %>% nrow()/nrow(review_gamma)
  opt_k[i-1,1]=i
  opt_k[i-1,2]=ndoc
}

opt_k


##############################################draft ####################################################

#get sentiment for non-stopwords: bing(only 'positive' and 'negative' for each token)
review_tidy_sen_bing<-review_tidy_nosw %>% inner_join(get_sentiments('bing'),by=c('token'='word'))
head(review_tidy_sen_bing,10)

senti_bing<-review_tidy_sen_bing %>% count(crmid,token,sentiment) %>%
  pivot_wider(names_from =sentiment ,values_from = n,values_fill = list(n=0)) %>%
  group_by(crmid) %>% summarise(pos=sum(positive),neg=sum(negative),polarity=pos-neg)

#what is it overall
skimr::skim(senti_bing)
#join onto the original data
review_senti1 = inner_join(review,senti_bing,by = "crmid")
View(review_senti1)
#ggplot(review_senti1, aes(x=airlines, y=polarity)) + geom_boxplot()
```