Final Project - Real Talk data

Frank Xu

11/19/2018

Data collection:

“Real Talk” is an online platform based in UNC aiming for teenagers to talk about their own experiences about relationships, puberty, bullying, etc. Teenagers from all over US post their own stories on the platform, which could be as short as a sentence, or as long as a paragraph. Not all stories are available to the public, but my team managed to make contact to them, get the data directly from them and do some data analysis for them.

Question on data:

1. Is there any association between positive/negative sentiment stories and demographic information that come with the stories?

2. Is the positive/negative sentiment stories associate with some outer source data such as teenager pregnancy/suicide and political status of each states?

```
# Reading samples
sample <- read.csv("Real_Talk_Data.csv", stringsAsFactors = FALSE)
sample <- sample[1:811,]
colnames(sample)[13] <- "zip"
sample$zip[sample$zip < 10000 & is.na(sample$zip) == F] <-
  sprintf("%05d", sample$zip[sample$zip < 10000 & is.na(sample$zip) == F])
sample <- sample[ , colSums(!is.na(sample)) != 0]
sample$Story <- do.call(paste, c(sample[18:127], sep=" "))
sample <- sample[,c(1:17,128)]

# Category Data Cleaning up
sample$LGBTQ.[sample$LGBTQ. == "" | is.na(sample$LGBTQ.) == T] <- "Prefer not to shar
e"
sample$LGBTQ.[sample$LGBTQ. == "nyes" | sample$LGBTQ. == "yes "] <- "yes"
sample$LGBTQ.[sample$LGBTQ. == "on"] <- "no"
sample$Gender[sample$Gender == "" | is.na(sample$Gender) == T] <- "Prefer not to shar
e"
sample$Gender[sample$Gender == "abstain" | sample$Gender == "abstain "] <- "Prefer no
t to share"
sample$Gender[sample$Gender == "Female"] <- "female"
sample$Gender[sample$Gender == "Male" | sample$Gender == "male "] <- "male"
sample$Gender[sample$Gender == "Other" | sample$Gender == "non-binary" | sample$Gende
r == "Non-binary"] <- "other"
subtype <- factor(c("Mobile App", "Website", "in person"), levels = c("Mobile App", "
Website", "in person"))
sample$Submission..Type[sample$Submission..Type == "squarespace"] <- subtype[1]
sample$Submission..Type[sample$Submission..Type == "online"] <- subtype[2]
sample$Submission..Type[sample$Submission..Type == "peers"] <- subtype[3]
```

```
# Adding Sentiment score
sample$sentiment <- sentiment_by(sample$Story)[,4]
sample$sentiment <- 10*apply(sample$sentiment,2,as.numeric)

# Adding zipcode, states, count, and other predictors
data(zipcode)
zip <- zipcode %>%
  dplyr::select(zip, state) %>%
  distinct(zip, .keep_all = TRUE)
samplestate <- merge(sample, zip, by = "zip")
samplecount <- samplestate %>%
  group_by(state) %>%
  count(state)
realtalk <- merge(sample, zip, by = "zip", all.x = TRUE)
regdata <- realtalk[c(-3,-10,-11,-17,-131,-133,-272,-411,-449,-614,-622,-769,-777,-80
6), c(1,19,4,10:12,17:18,20)]
sentimentmean <- regdata %>%
  group_by(state) %>%
  summarise(mean_sent = mean(sentiment, na.rm = T))

# Adding election data
sentimentmean <- merge(sentimentmean, samplecount, by = "state")
sentimentmean$red <- 0
sentimentmean$red[c(1:3,5,7,8,10,11,13:16,19:28,33,34,36,38:42,44:46)] <- 1
```

Reading in data.

The original data has 18 columns and 811 rows.

Code Book:

1. Submission..code: the code id generated for each submission.
2. Date.of..submission: the exact data for each submission.
3. Submission..Type: the platform used for each submission.
4. Location.of..Submission: the location for each submission, only applicable for part of “in person” type submissions. The capital letters stand for middle school names.
5. Score: scored by certain Real Talk employees deciding whether they should publish the data.
6. Scored.By: the certain employee that scored the story.
7. Published: the final decision on whether they will publish the story.
8. Published.Date: the date of publish, only applicable for stories that already published.
9. Gender: the gender of the user who published the story. Male/female/other/prefer not to share.
10. LGBTQ.: the LGBTQ. status of the user who published the story. Yes/no/prefer not to share.
11. Age: the age of the user who published the story.
12. Race: the race of the user who published the story.

13. zip: the zip code users submitted when signing up for the account.
14. Location: the location users submitted when signing up for the account.
15. Phone..: the phone numbers users submitted when signing up for the account.
16. Email.Address: the email address users submitted when signing up for the account.
17. Subject: the subject of the each story posted. Users had to choose one subject before submitting each stories.
18. Story: the exact text of story of the each story posted. Can be as shorted as several words, or several sentences.

Missing data:

There are a bunch of missing data in this dataset. Some missing data is due to they did not collect those data or the interviewee did not want to share when gathering stories, such as "Data of submission", "location of submission", "LGBTQ.", "Race", "Zip", "Location", "Phone", "Email".

Other missing data is due to the meaning of the certain columns, such as "Score", "Scored by", "Published", "Published Date". In these columns, blanks were left if the data is not applicable.

```
# Reading in teen preg data
teenpreg14 <- read.csv("TEENBIRTHS2014.csv")
teenpreg15 <- read.csv("TEENBIRTHS2015.csv")
teenpreg16 <- read.csv("TEENBIRTHS2016.csv")
teenpreg <- merge(teenpreg14, teenpreg15, by = "STATE")
teenpreg <- merge(teenpreg, teenpreg16, by = "STATE")
teenpreg$rate <- rowMeans(teenpreg[,c(2,4,6)])
teenpreg <- teenpreg[,c(1,8)]
colnames(teenpreg) <- c("state", "teen_preg_rate")
```

```
# Reading in teen preg data
suicide <- read.table("Compressed Mortality, 1999-2016.txt", sep = "", header = T, fi
ll = T, nrows = 50)
suicide <- suicide[,c(1,3:5)]
colnames(suicide) <- c("State", "death", "population", "suicide_rate")
for (i in 1:50){
  suicide$state[i] <- state.abb[grep(suicide[i, 1], state.name)]
}
```

```
## Warning in suicide$state[i] <- state.abb[grep(suicide[i, 1], state.name)]:
## number of items to replace is not a multiple of replacement length
```

```
suicide <- suicide[,c(5,4)]
```

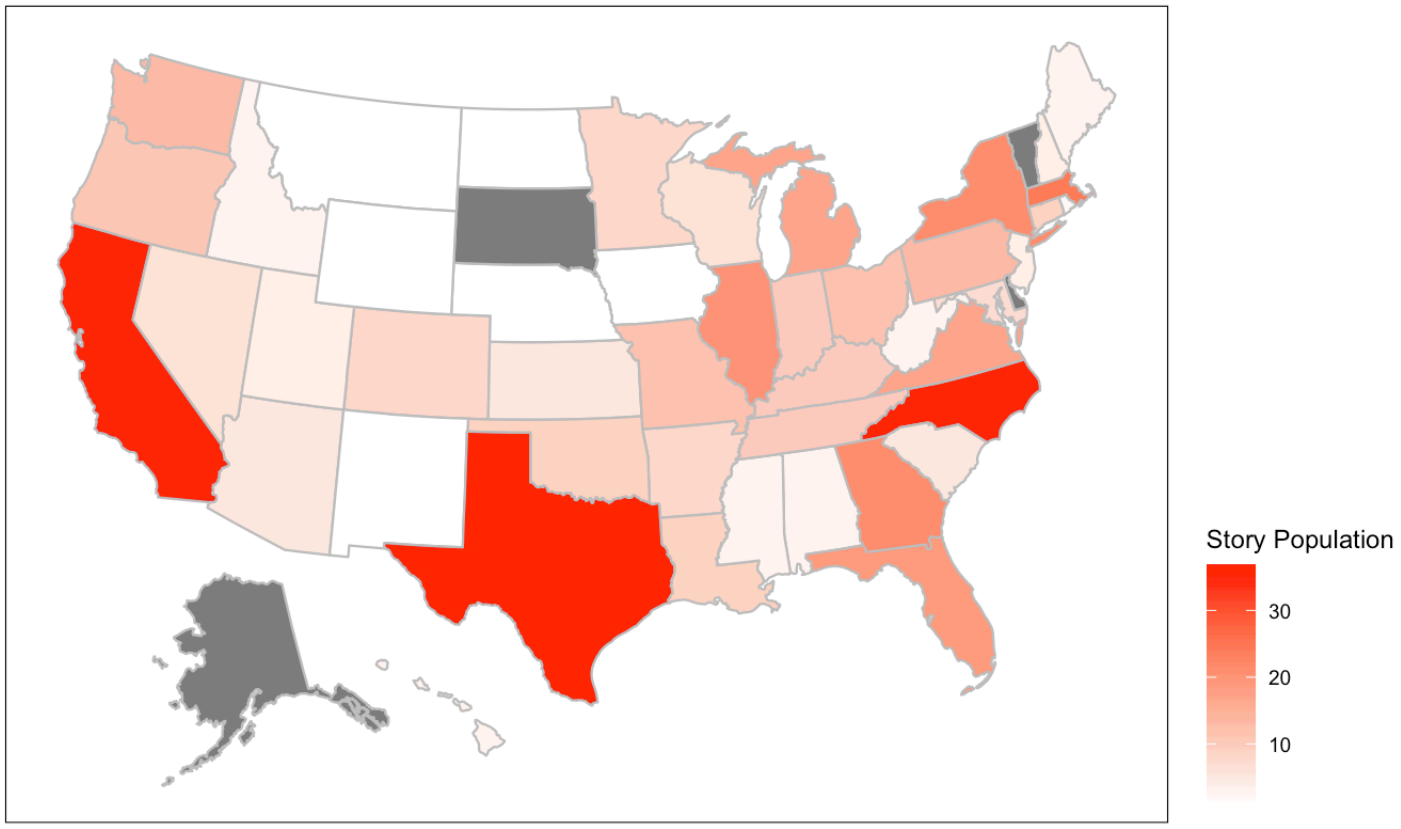
Also including the recent election data, teenager pregnancy data and the teenager suicide

data, in order to see if there is any connection among them. The recent election data is added manually based on the 2018 election state results. The teenager pregnancy data is an average result of 2014, 2015, 2016, which all come from CDC. The teenager suicide data is an average result from 2012-2016, also from CDC. Those data all have a value for every state, either binary (for election) or numeric (for teen pregnancy or suicide rates).

```
# Preprocessing regression data
if (!"red" %in% colnames(regdata)){
  regdata <- merge(regdata, sentimentmean[,c(1,4)], by = "state", all.x = TRUE)
}
regdata$Gender <- as.factor(regdata$Gender)
regdata$LGBTQ. <- as.factor(regdata$LGBTQ.)
regdata$state <- as.factor(regdata$state)
regdata$Submission..Type <- as.factor(regdata$Submission..Type)
regdata$Subject <- as.factor(regdata$Subject)
regdata$wordcount <- str_count(regdata$Story, '\\w+')
regdata$wordcountc <- regdata$wordcount - mean(regdata$wordcount)
regdata$sentiment2 <- 0
regdata$sentiment2[regdata$sentiment > 0] <- 1

# Creating Regions
regdata$Reg[regdata$state %in% c("WA", "OR", "CA", "NV", "ID", "MT", "WY", "UT", "CO"
, "AK", "HI")] <- "west"
regdata$Reg[regdata$state %in% c("ND", "MN", "WI", "MI", "OH", "IN", "IL", "IA", "MO"
, "SD", "NE", "KS")] <- "midwest"
regdata$Reg[regdata$state %in% c("AZ", "NM", "OK", "TX")] <- "southwest"
regdata$Reg[regdata$state %in% c("AR", "LA", "MS", "AL", "GA", "FL", "SC", "NC", "KY"
, "TN", "WV", "VA", "MD", "DE")] <- "southeast"
regdata$Reg[regdata$state %in% c("ME", "VT", "NH", "MA", "NY", "RI", "CT", "NJ", "PA"
)] <- "northeast"

regdata <- merge(regdata, teenpreg, by = "state", all.x = T)
regdata <- merge(regdata, suicide, by = "state", all.x = T)

# Regdata2 is the datasets with no NA rows
regdata2 <- regdata[is.na(regdata$state) == F & is.na(regdata$Age) == F,]
```

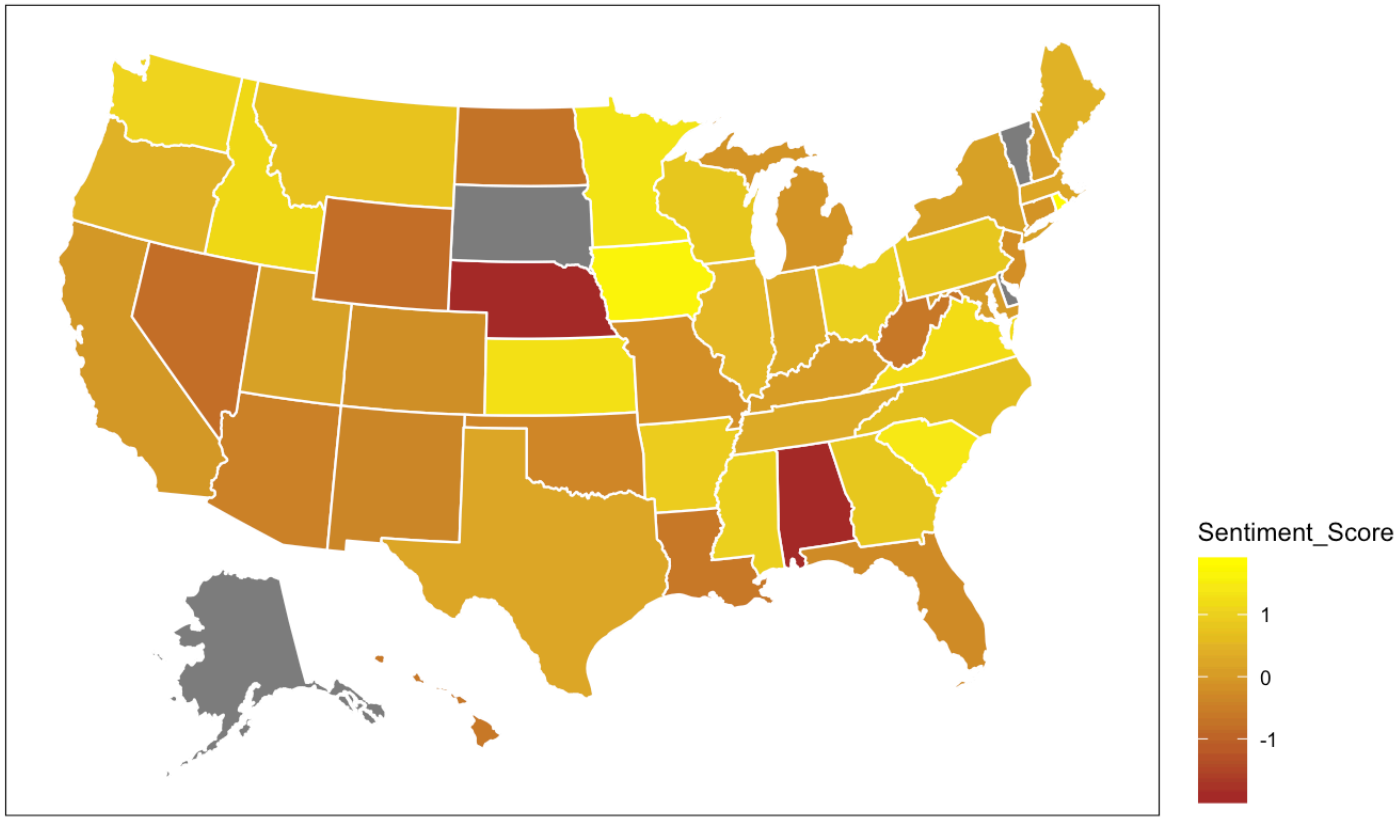
```
# Basic population visualization
plot_usmap(regions = "states", data = samplecount, values = "n", lines = "grey") +
  scale_fill_continuous(low = "white", high = "red", name = "Story Population", label
= scales::comma) +
  labs(title = "Real Talk Sample Density in US") +
  theme(panel.background = element_rect(colour = "black", fill = "white"), legend.pos
ition = "right")
```

Real Talk Sample Density in US




```
plot_usmap(regions = "states", data = sentimentmean, values = "mean_sent", lines = "w
hite") +
  scale_fill_continuous(low = "brown", high = "yellow", name = "Sentiment_Score", lab
el = scales::comma) +
  labs(title = "Real Talk Sample Sentiment in US") +
  theme(panel.background = element_rect(colour = "black", fill = "white"), legend.pos
ition = "right")
```

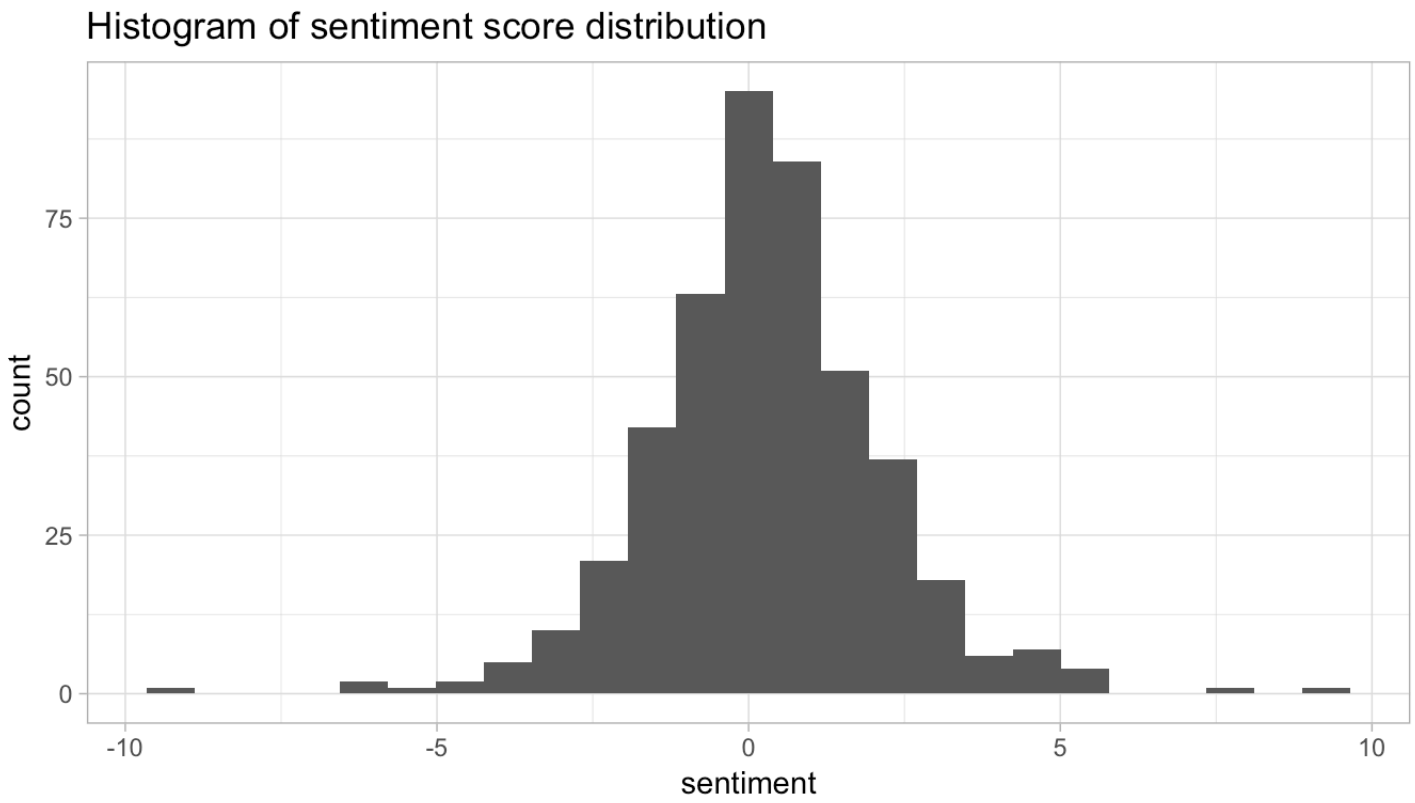
Real Talk Sample Sentiment in US



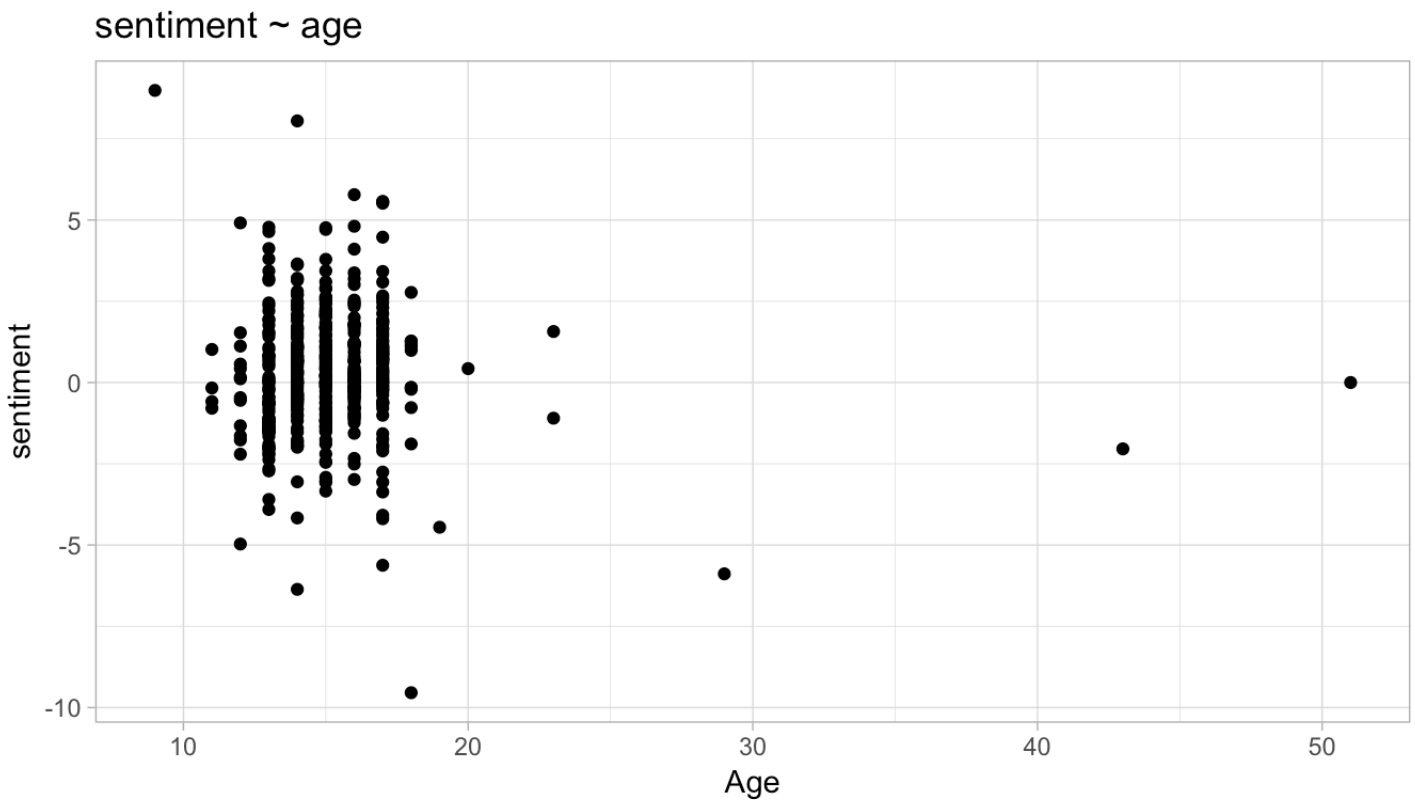
This is a data overview of sample density in states and mean sentiment score for each state.

Exploratory Data Analysis:

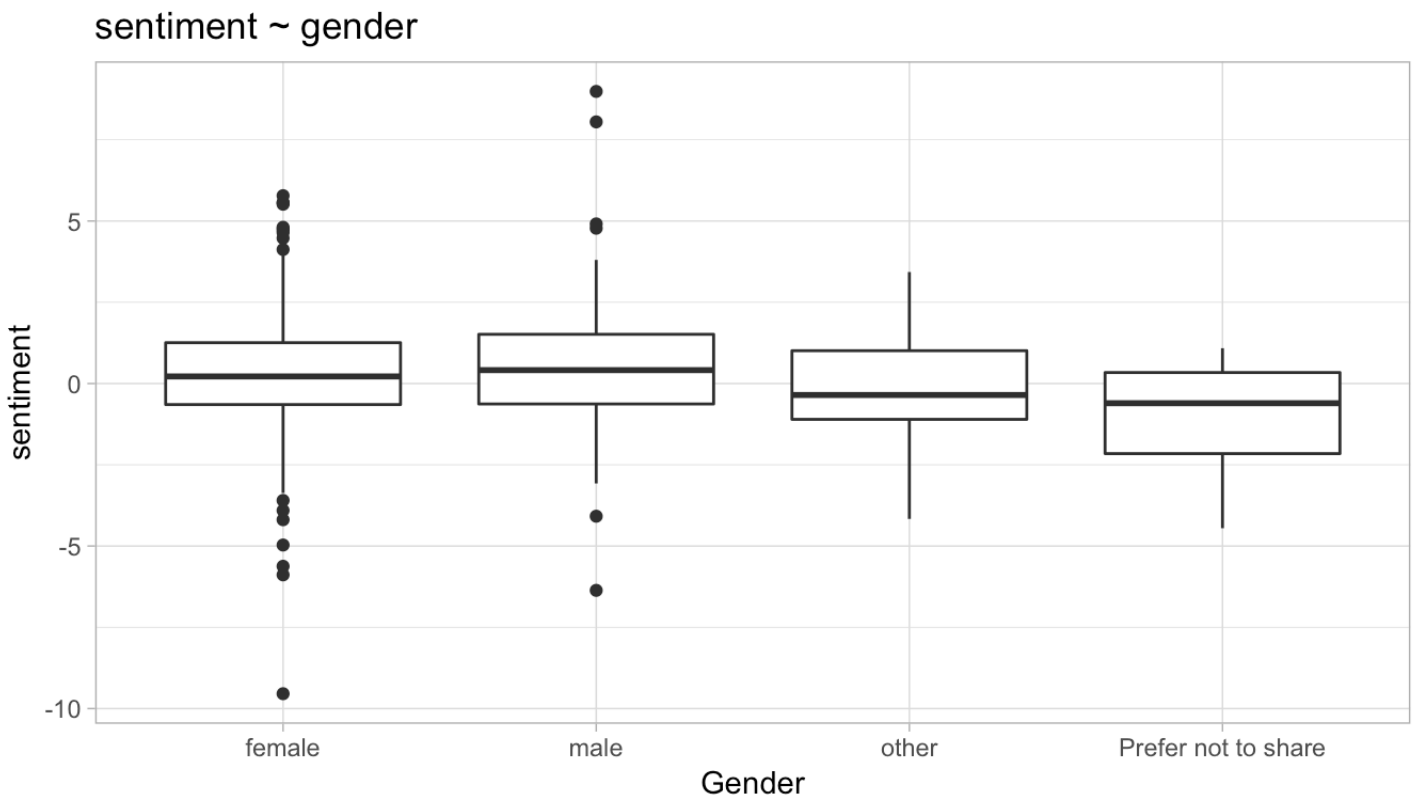
```
# EDA
gp <- ggplot(data = regdata2)
gp + geom_histogram(aes(x = sentiment), bins = 25) + labs(title = "Histogram of senti
ment score distribution") + theme_light()
```



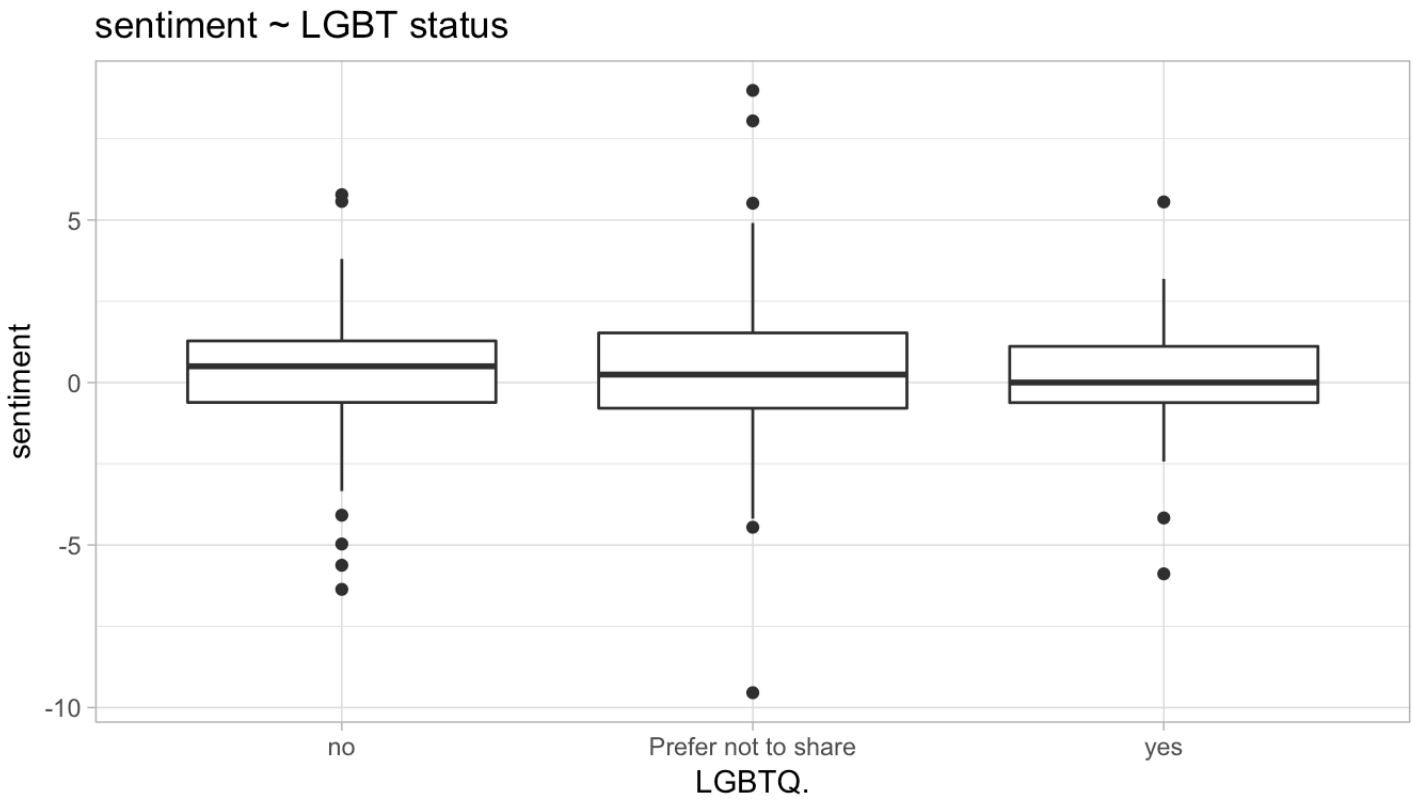

```
gp + geom_point(aes(x = Age, y = sentiment)) + labs(title = "sentiment ~ age") + them
e_light()
```



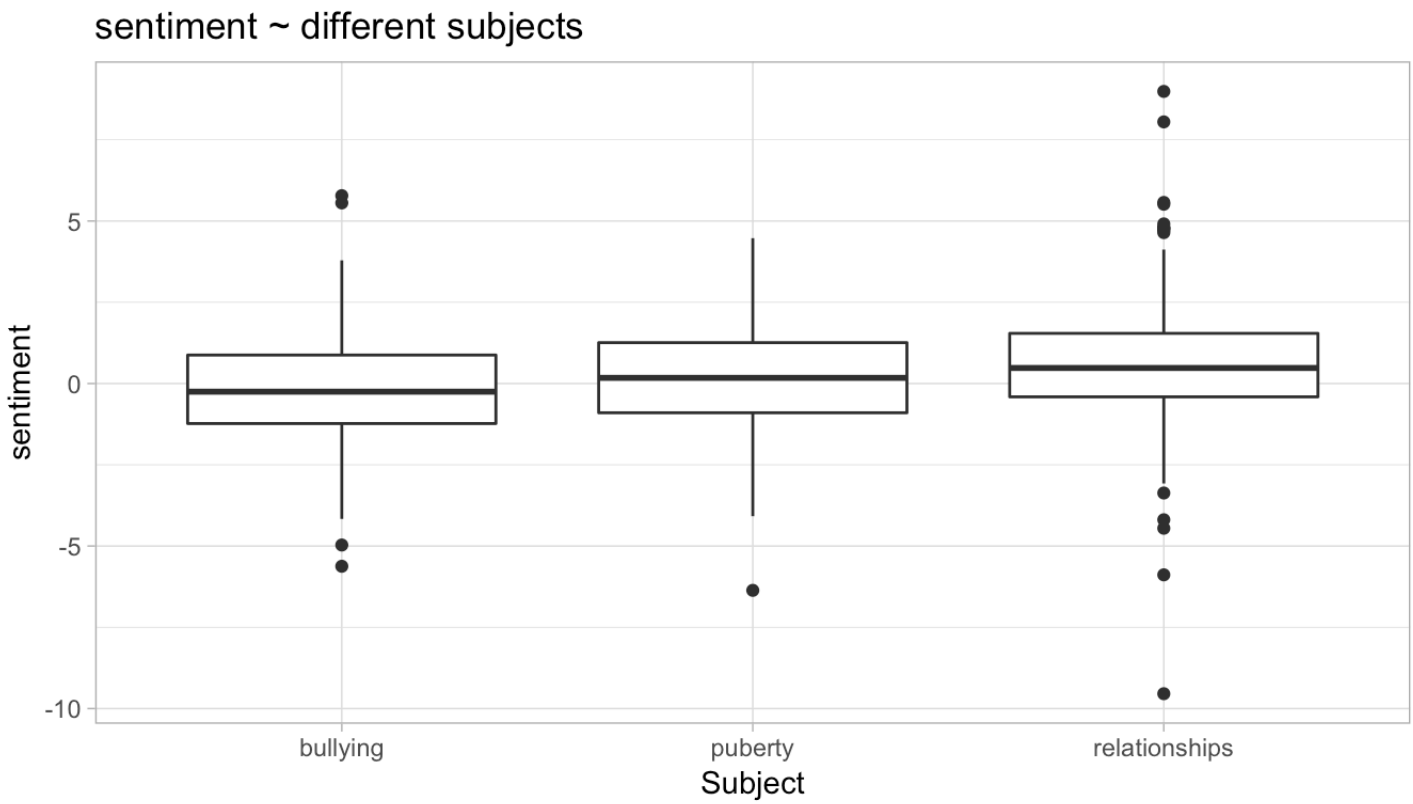

```
gp + geom_boxplot(aes(x = Gender, y = sentiment)) + labs(title = "sentiment ~ gender"
) + theme_light()
```



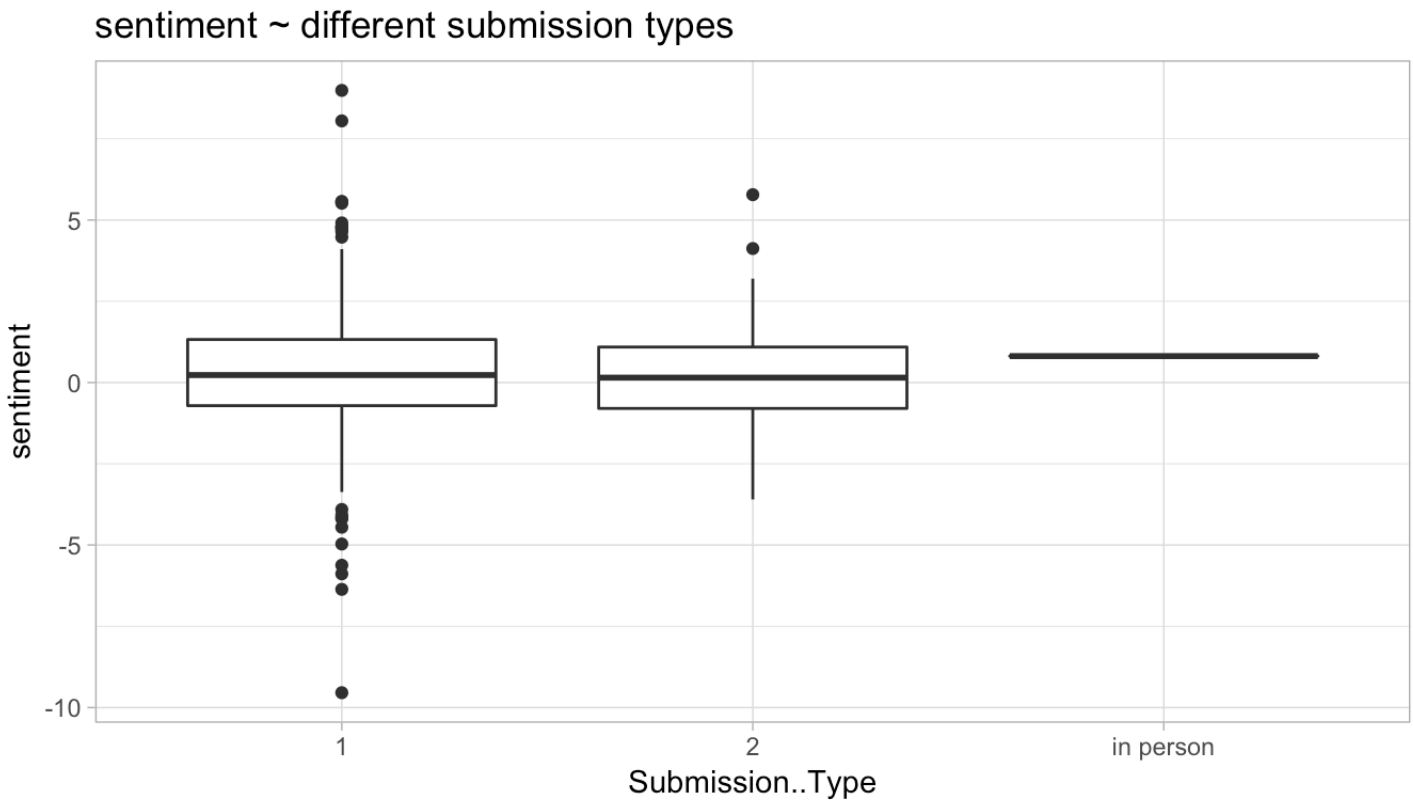

```
gp + geom_boxplot(aes(x = LGBTQ., y = sentiment)) + labs(title = "sentiment ~ LGBT st
atus") + theme_light()
```



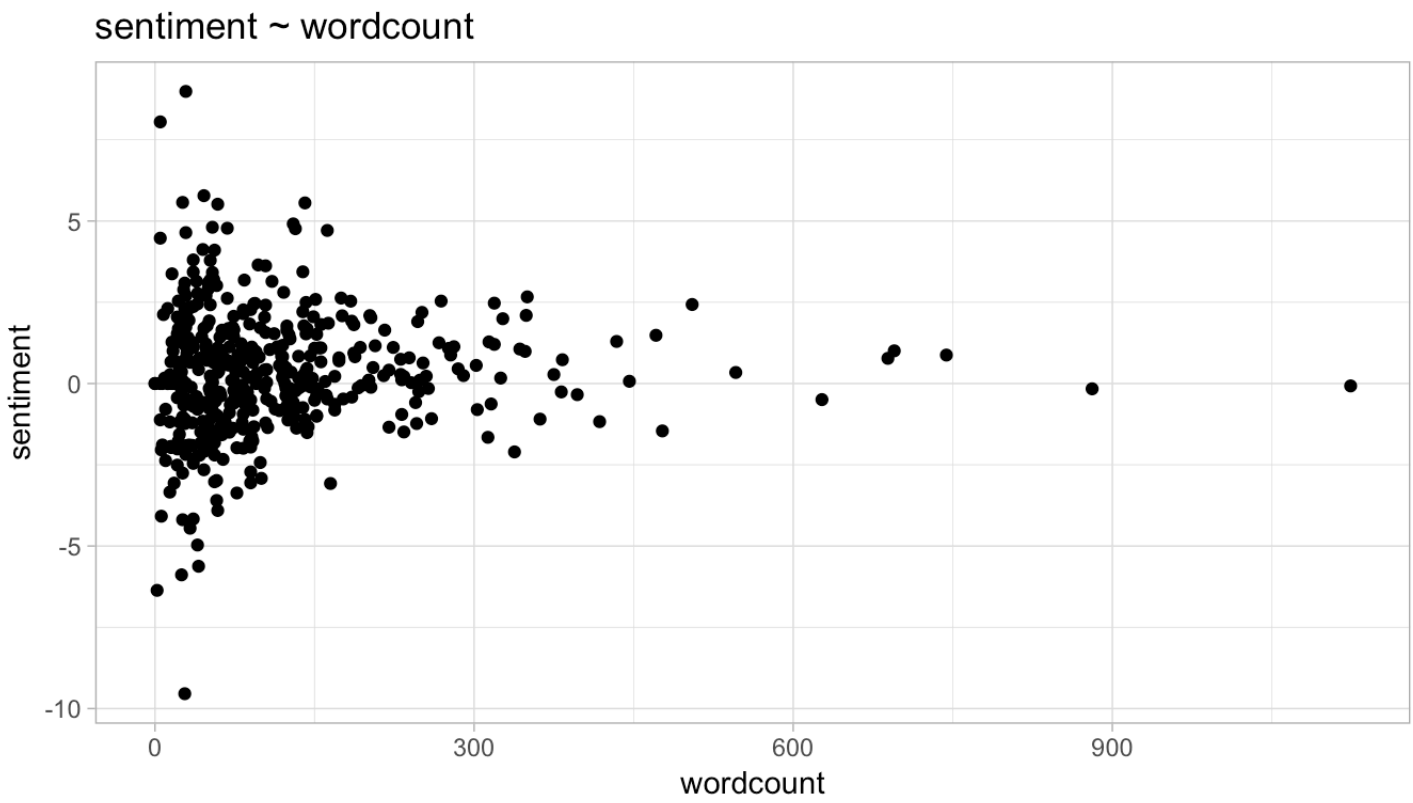

```
gp + geom_boxplot(aes(x = Subject, y = sentiment)) + labs(title = "sentiment ~ differ
ent subjects") + theme_light()
```



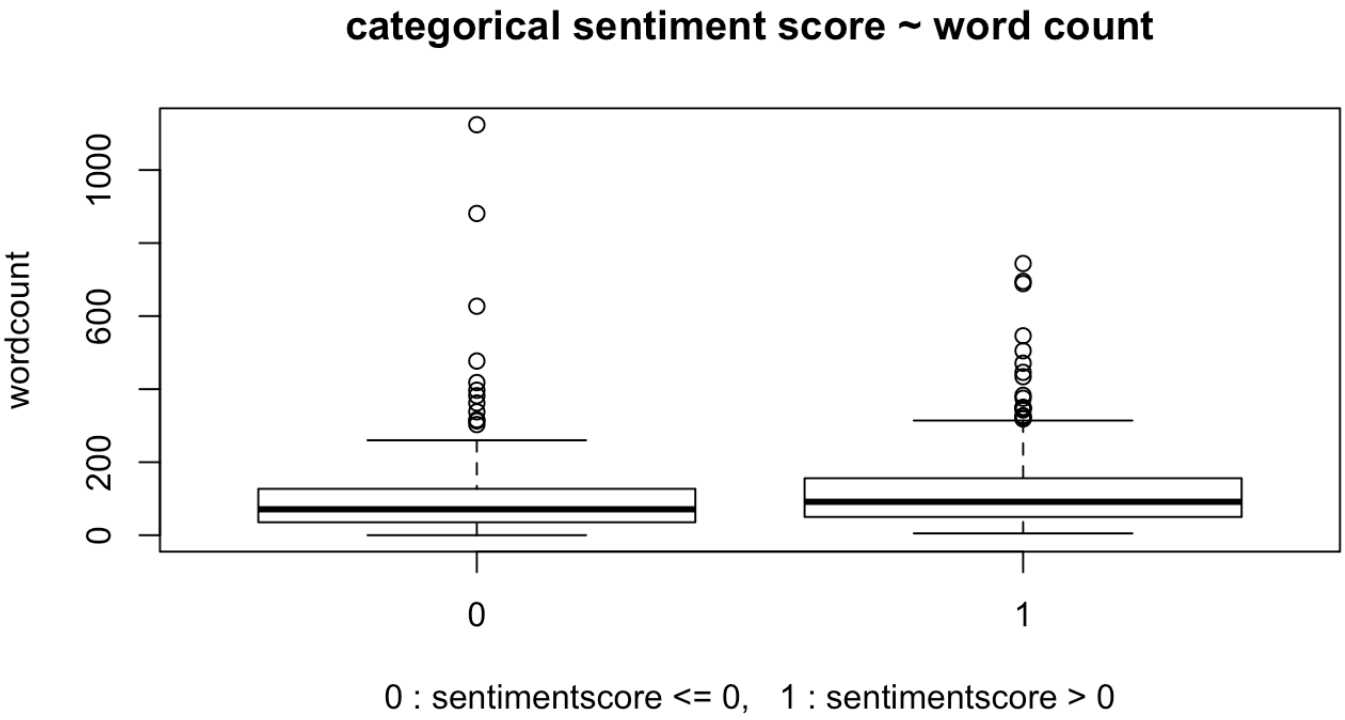

```
gp + geom_boxplot(aes(x = Submission..Type, y = sentiment)) + labs(title = "sentiment
~ different submission types") + theme_light()
```



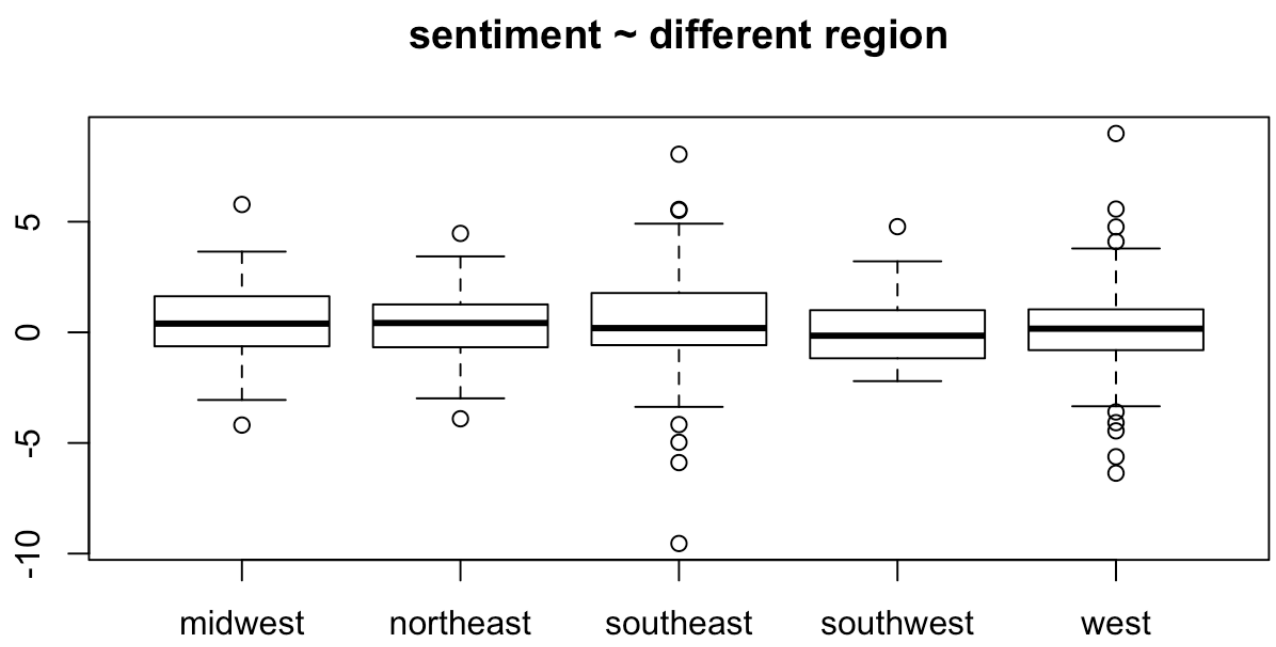

```
gp + geom_point(aes(x = wordcount, y = sentiment)) + labs(title = "sentiment ~ wordco
unt") + theme_light()
```



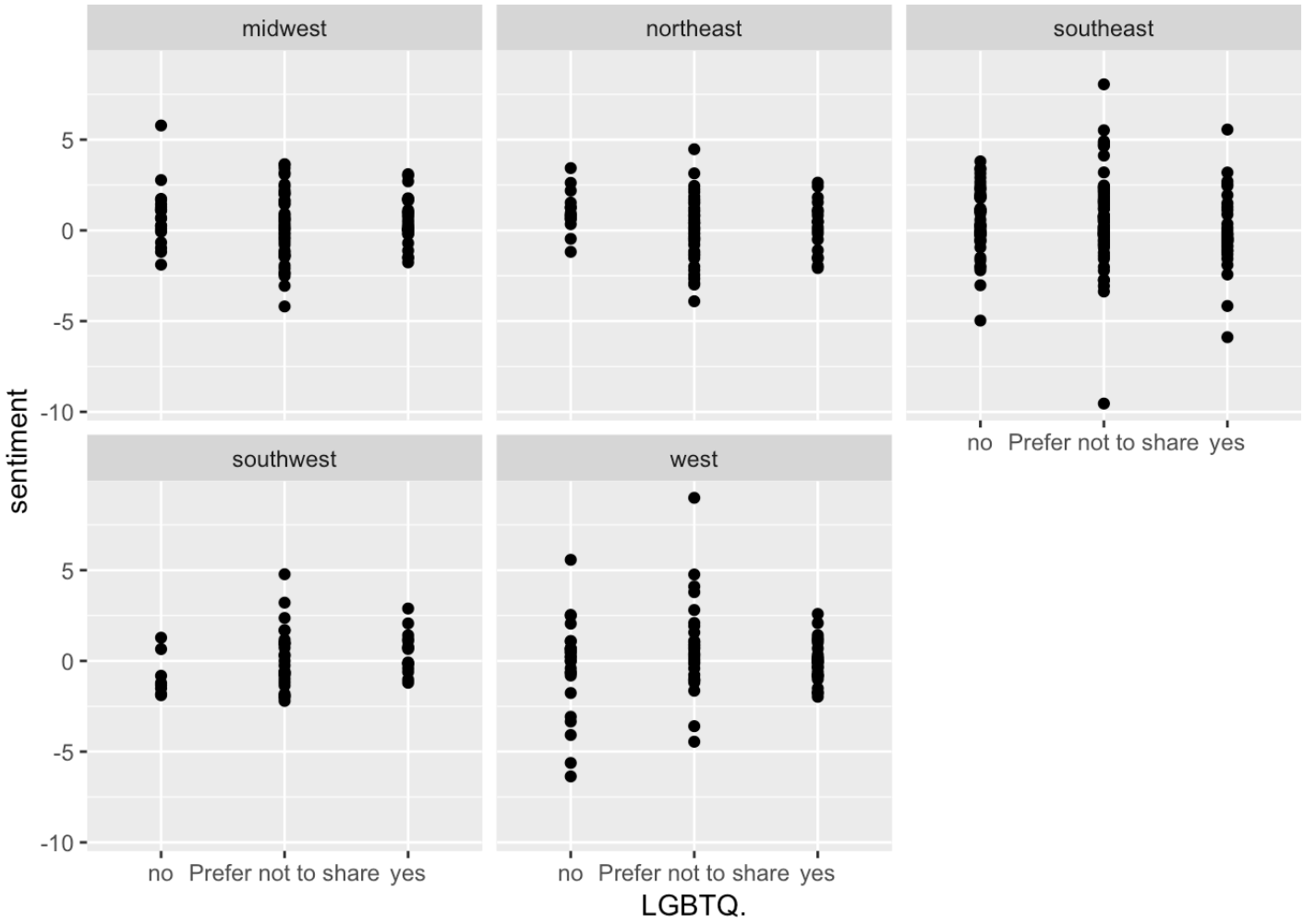

```
boxplot(wordcount~sentiment2, data = regdata2, main = "categorical sentiment score ~
word count",
        xlab = "0 : sentimentscore <= 0,   1 : sentimentscore > 0", ylab = "wordcount
")
```



```
boxplot(sentiment~Reg, data = regdata2, main = "sentiment ~ different region")
```




```
ggplot(data = regdata2)+
  geom_point(aes(y = sentiment, x = LGBTQ.))+
  facet_wrap(~Reg)
```




```
# the regression model with no other outer source
reg1 <- glm(sentiment2 ~ Age + Gender + LGBTQ. + Reg + wordcount + Submission..Type +
Subject, family = binomial, data = regdata2)
# the regression model with teenager pregnancy rate
reg2 <- glm(sentiment2 ~ Age + Gender + LGBTQ. + Reg + wordcount + Submission..Type +
Subject + teen_preg_rate, family = binomial, data = regdata2)
# the regression model with teenager suicide rate
reg3 <- glm(sentiment2 ~ Age + Gender + LGBTQ. + Reg + wordcount + Submission..Type +
Subject + suicide_rate, family = binomial, data = regdata2)
# the regression model with 2018 election result
reg4 <- glm(sentiment2 ~ Age + Gender + LGBTQ. + Reg + wordcount + Submission..Type +
Subject + red, family = binomial, data = regdata2)
# Anova tests to see if these outer sources are useful for predicting sentiment score
anova(reg1, reg2, test= "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: sentiment2 ~ Age + Gender + LGBTQ. + Reg + wordcount + Submission..Type +
##     Subject
## Model 2: sentiment2 ~ Age + Gender + LGBTQ. + Reg + wordcount + Submission..Type +
##     Subject + teen_preg_rate
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       435     592.59
## 2       434     592.14  1  0.45426   0.5003
```

```
anova(reg1, reg3, test= "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: sentiment2 ~ Age + Gender + LGBTQ. + Reg + wordcount + Submission..Type +
##     Subject
## Model 2: sentiment2 ~ Age + Gender + LGBTQ. + Reg + wordcount + Submission..Type +
##     Subject + suicide_rate
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       435     592.59
## 2       434     592.42  1  0.17254   0.6779
```

```
anova(reg1, reg4, test= "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: sentiment2 ~ Age + Gender + LGBTQ. + Reg + wordcount + Submission..Type +
##     Subject
## Model 2: sentiment2 ~ Age + Gender + LGBTQ. + Reg + wordcount + Submission..Type +
##     Subject + red
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       435     592.59
## 2       434     591.69  1   0.9016   0.3424
```

Reg5 is the model with the interaction between LGBTQ status and regions (northeast, southeast, west, midwest, southwest). The following anova shows that the interaction is significant, with a P value of 0.033. The area under curve is 0.6739.

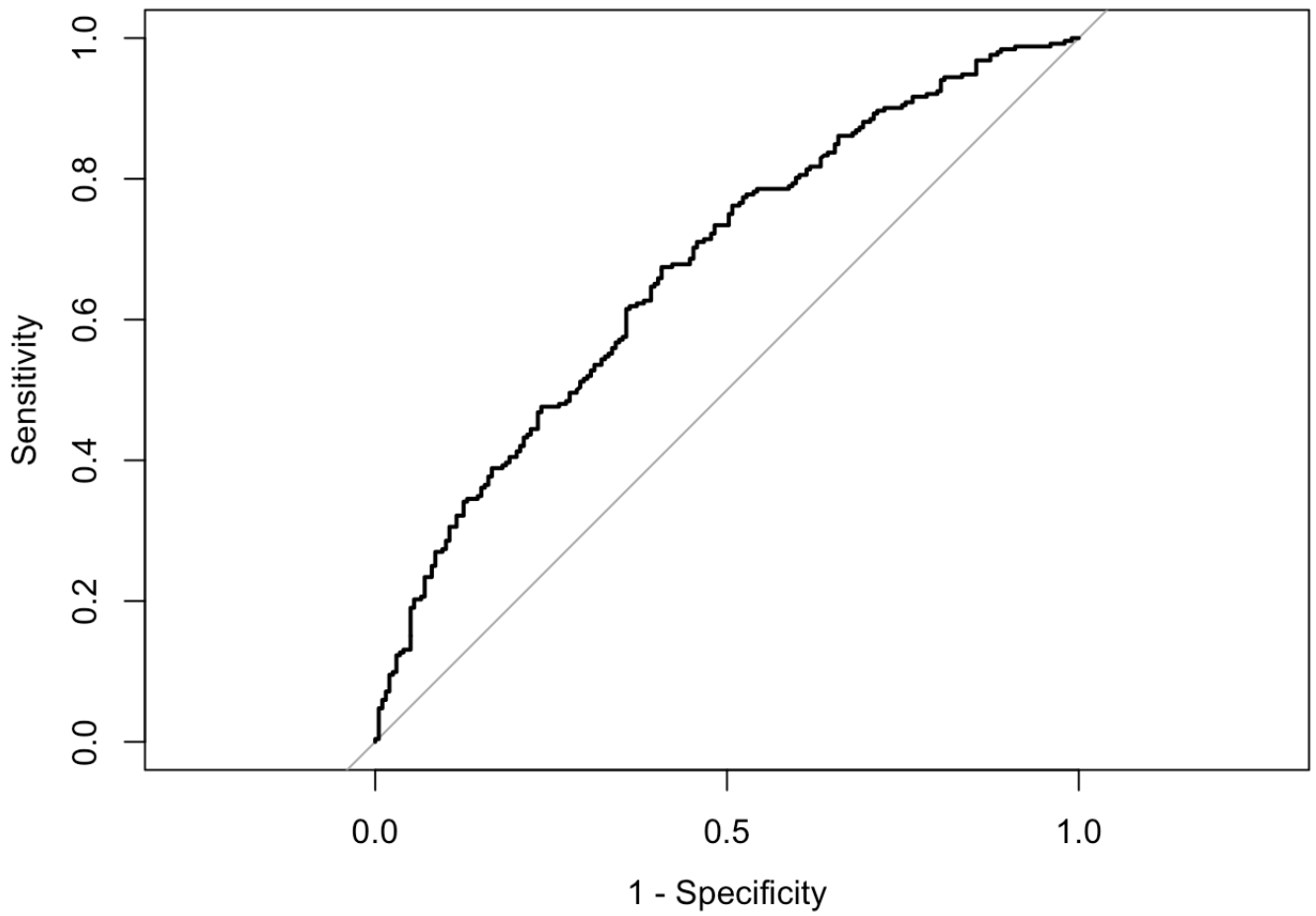
```
reg5 <- glm(sentiment2 ~ Age + Gender + LGBTQ. * Reg + wordcountc + Submission..Type
+ Subject, family = binomial, data = regdata2)
summary(reg5)
```

```
##
## Call:
## glm(formula = sentiment2 ~ Age + Gender + LGBTQ. * Reg + wordcountc +
```

```
##     Submission..Type + Subject, family = binomial, data = regdata2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5599  -1.1795   0.7707   1.0222   1.8551
##
## Coefficients:
##                                          Estimate Std. Error z value
## (Intercept)                               0.769806   0.789108   0.976
## Age                                      -0.018031   0.036402  -0.495
## Gendermale                                0.291073   0.233144   1.248
## Genderother                              -0.805906   0.589482  -1.367
## GenderPrefer not to share                -1.110403   0.921620  -1.205
## LGBTQ.Prefer not to share                -0.825961   0.596908  -1.384
## LGBTQ.yes                                -0.217110   0.660445  -0.329
## Regnortheast                              0.935937   0.923674   1.013
## Regsoutheast                             -0.513266   0.592362  -0.866
## Regsouthwest                             -1.924303   0.972656  -1.978
## Regwest                                  -0.939952   0.648687  -1.449
## wordcountc                                0.002080   0.000921   2.258
## Submission..Type2                         0.193603   0.433792   0.446
## Submission..Typein person                13.457325 535.411271   0.025
## Subjectpuberty                            0.477955   0.333904   1.431
## Subjectrelationships                      0.625207   0.273752   2.284
## LGBTQ.Prefer not to share:Regnortheast   -1.141599   1.021105  -1.118
## LGBTQ.yes:Regnortheast                   -1.373393   1.135620  -1.209
## LGBTQ.Prefer not to share:Regsoutheast    0.701849   0.718773   0.976
## LGBTQ.yes:Regsoutheast                   -0.963122   0.813622  -1.184
## LGBTQ.Prefer not to share:Regsouthwest    1.477889   1.093866   1.351
## LGBTQ.yes:Regsouthwest                    1.214662   1.183029   1.027
## LGBTQ.Prefer not to share:Regwest         1.278302   0.810009   1.578
## LGBTQ.yes:Regwest                        -0.317157   0.878887  -0.361
##                                         Pr(>|z|)
## (Intercept)                               0.3293
## Age                                       0.6204
## Gendermale                                0.2119
## Genderother                               0.1716
## GenderPrefer not to share                 0.2283
## LGBTQ.Prefer not to share                 0.1664
## LGBTQ.yes                                 0.7424
## Regnortheast                              0.3109
## Regsoutheast                              0.3862
## Regsouthwest                              0.0479 *
## Regwest                                   0.1473
## wordcountc                                0.0240 *
## Submission..Type2                         0.6554
## Submission..Typein person                 0.9799
## Subjectpuberty                            0.1523
## Subjectrelationships                      0.0224 *
## LGBTQ.Prefer not to share:Regnortheast    0.2636
```

```
## LGBTQ.yes:Regnortheast                    0.2265
## LGBTQ.Prefer not to share:Regsoutheast   0.3288
## LGBTQ.yes:Regsoutheast                    0.2365
## LGBTQ.Prefer not to share:Regsouthwest   0.1767
## LGBTQ.yes:Regsouthwest                    0.3045
## LGBTQ.Prefer not to share:Regwest        0.1145
## LGBTQ.yes:Regwest                         0.7182
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 618.98  on 450  degrees of freedom
## Residual deviance: 575.88  on 427  degrees of freedom
## AIC: 623.88
##
## Number of Fisher Scoring iterations: 12
```

```
roc(regdata2$sentiment2, fitted(reg5), plot=T, legacy.axes=T)
```



```
##
## Call:
## roc.default(response = regdata2$sentiment2, predictor = fitted(reg5),     plot = T
, legacy.axes = T)
##
## Data: fitted(reg5) in 199 controls (regdata2$sentiment2 0) < 252 cases (regdata2$s
entiment2 1).
## Area under the curve: 0.6739
```

```
anova(reg1, reg5, test= "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: sentiment2 ~ Age + Gender + LGBTQ. + Reg + wordcount + Submission..Type +
##     Subject
## Model 2: sentiment2 ~ Age + Gender + LGBTQ. * Reg + wordcountc + Submission..Type
+
##     Subject
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       435     592.59
## 2       427     575.88  8   16.708   0.0333 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
exp(confint.default(reg5))
```

```
##                                             2.5 %     97.5 %
## (Intercept)                            0.45986694 10.1394060
## Age                                    0.91449959  1.0547630
## Gendermale                             0.84714661  2.1128276
## Genderother                            0.14067979  1.4182971
## GenderPrefer not to share              0.05410957  2.0055909
## LGBTQ.Prefer not to share              0.13589412  1.4105185
## LGBTQ.yes                              0.22056602  2.9368526
## Regnortheast                           0.41709946 15.5849461
## Regsoutheast                           0.18744428  1.9112197
## Regsouthwest                           0.02169497  0.9822278
## Regwest                                0.10955210  1.3929872
## wordcountc                             1.00027440  1.0038923
## Submission..Type2                      0.51860453  2.8400456
## Submission..Typein person              0.00000000         Inf
## Subjectpuberty                         0.83821229  3.1030783
## Subjectrelationships                   1.09271218  3.1955271
## LGBTQ.Prefer not to share:Regnortheast 0.04315631  2.3625211
## LGBTQ.yes:Regnortheast                 0.02734648  2.3452230
## LGBTQ.Prefer not to share:Regsoutheast 0.49316056  8.2533377
## LGBTQ.yes:Regsoutheast                 0.07747555  1.8805201
## LGBTQ.Prefer not to share:Regsouthwest 0.51373421 37.4058532
## LGBTQ.yes:Regsouthwest                 0.33153167 34.2387129
## LGBTQ.Prefer not to share:Regwest      0.73397050 17.5647015
## LGBTQ.yes:Regwest                      0.13006199  4.0772791
```

Data interpretation:

For a youngest(9-year-old) average non-LGBTQ female from midwest with a story of average lenghth, about bullying and submitted on mobile app, the average odds of positive sentiment is 2.16 (exp(0.770) = 2.159, with a 95% CI of (0.36, 8.19)). With every 1 year increase in age for people from the sample, the odds of positive sentiment from their story decreased by a multiplier of 0.98. (exp(-0.018) = 0.9822, with a 95% CI of (0.91, 1.05))

For the gender, as the baseline and as mentioned before, an average non-LGBTQ female from midwest with a story of average lenghth, about bullying and submitted on mobile app, the average odds of positive sentiment is 2.16. The model indicates that people who "Prefer not ot share" their gender has the lowest odds of positive sentiment, which is 67% lower than that of "female" (exp(-1.11) = 0.330, with a 95% CI of (0.05, 2.01)) on average. Followed by people that identify their gender as "other", which is 55% lower than that of "female" (exp(-0.806) = 0.447, with a 95% CI of (0.14, 1.42)) on average. Nontheless, "male" has the highest odds of positive sentiment above all, which is 34% higher than "female" (exp(0.291) = 1.336, with a 95% CI of (0.84, 2.11)), and 199% higher than "other", 305% higher than "Prefer not to share", on average.

There is some interesting interaction with regions and LGBTQ status.

1. In midwest, people not being "LGBTQ" has the highest odds of positive sentiment. People of "LGBTQ" has slightly lower odds of positive sentiment, which is 19% lower than non-LGBTQ (exp(-0.217) = 0.805, with a 95% CI of (0.22, 2.94)). People who "Prefer not to share" their LGBTQ status has the lowest odds of all, which is 54% lower than non-LGBTQ (exp(-0.826) = 0.438, with a 95% CI of (0.14, 1.42)).

2. In northeast, people have highest odds of positive sentiment among all regions. Non-LGBTQ has the highest odds of about 5.51 (exp(0.770 + 0.936) = 5.507), followed by LGBTQ people, and people “Prefer not to share” has the lowest. The odds are 80% and 86% lower, respectively.
3. Following northeast and midwest, southeast has the third highest odds among all. Non-LGBTQ has the highest odds of about 1.29, followed by people “Prefer not to share”, then LGBTQ people, the odds are 12% and 69% lower, respectively.
4. West has the fourth highest odds (2nd lowest odds) among all, where people “prefer not to share” has the highest odds of about 1.32, followed by non-LGBTQ people, and LGBTQ people, the odds are 36% and 55% lower, respectively.
5. Southwest has the lowest odds, where LGBTQ people has the highest odds of about 0.86, followed by people “prefer not to share”, then non-LGBTQ, the odds are 30% and 63% lower, respectively.

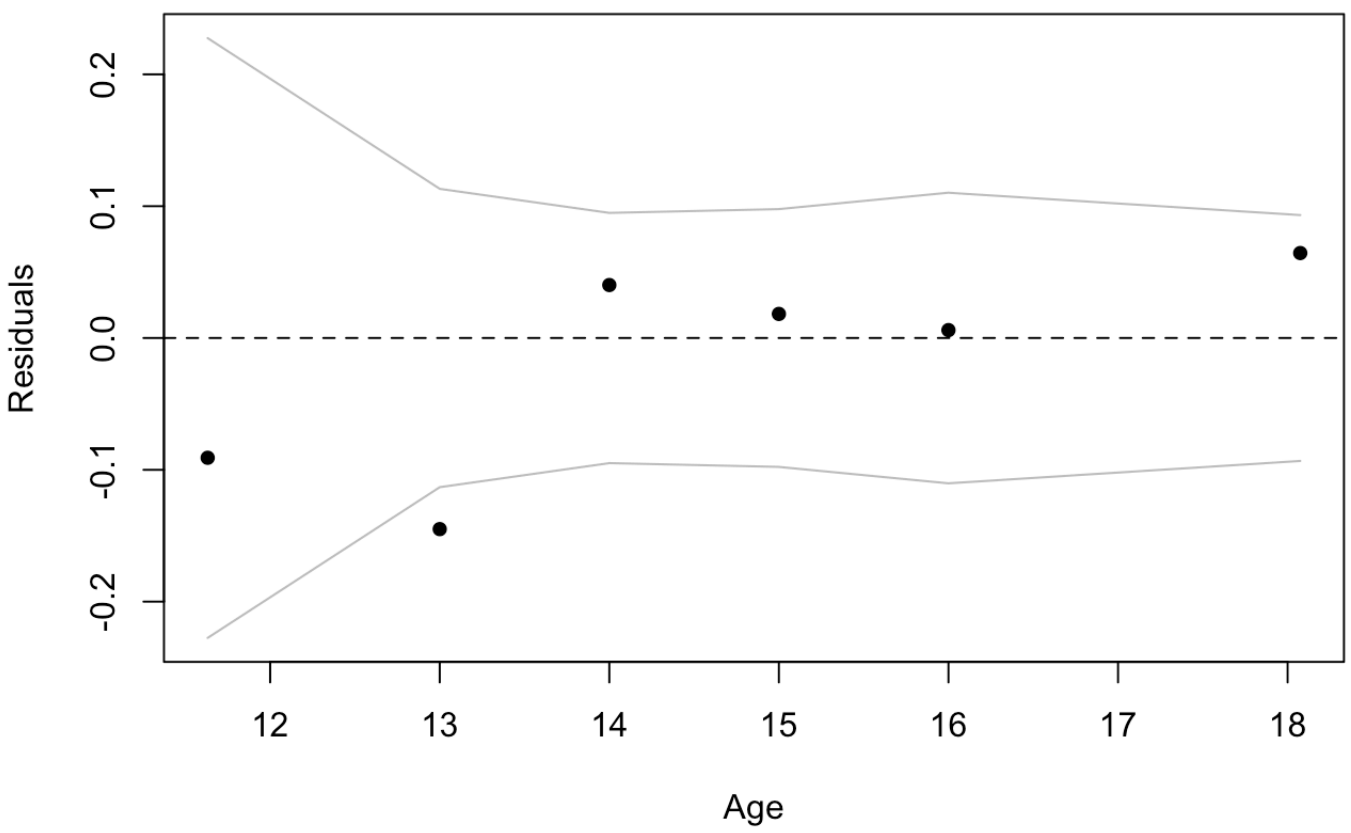
Wordcount shows a slightly positive relation with the odds of positive sentiment. As mentioned before, the baseline person with an average lenghth of wordcount has the odds of 2.16. With per 100 words increase in the story, the odds increased by a multiplier of 1.22. (exp(0.002*100) = 1.221)

Submission type does not really make a statistical difference on the odds, especially for the type “in person”, since there is only one valid row of data of that group. Type “website” (type2) has slightly higher odds (21%) than “mobile app”.

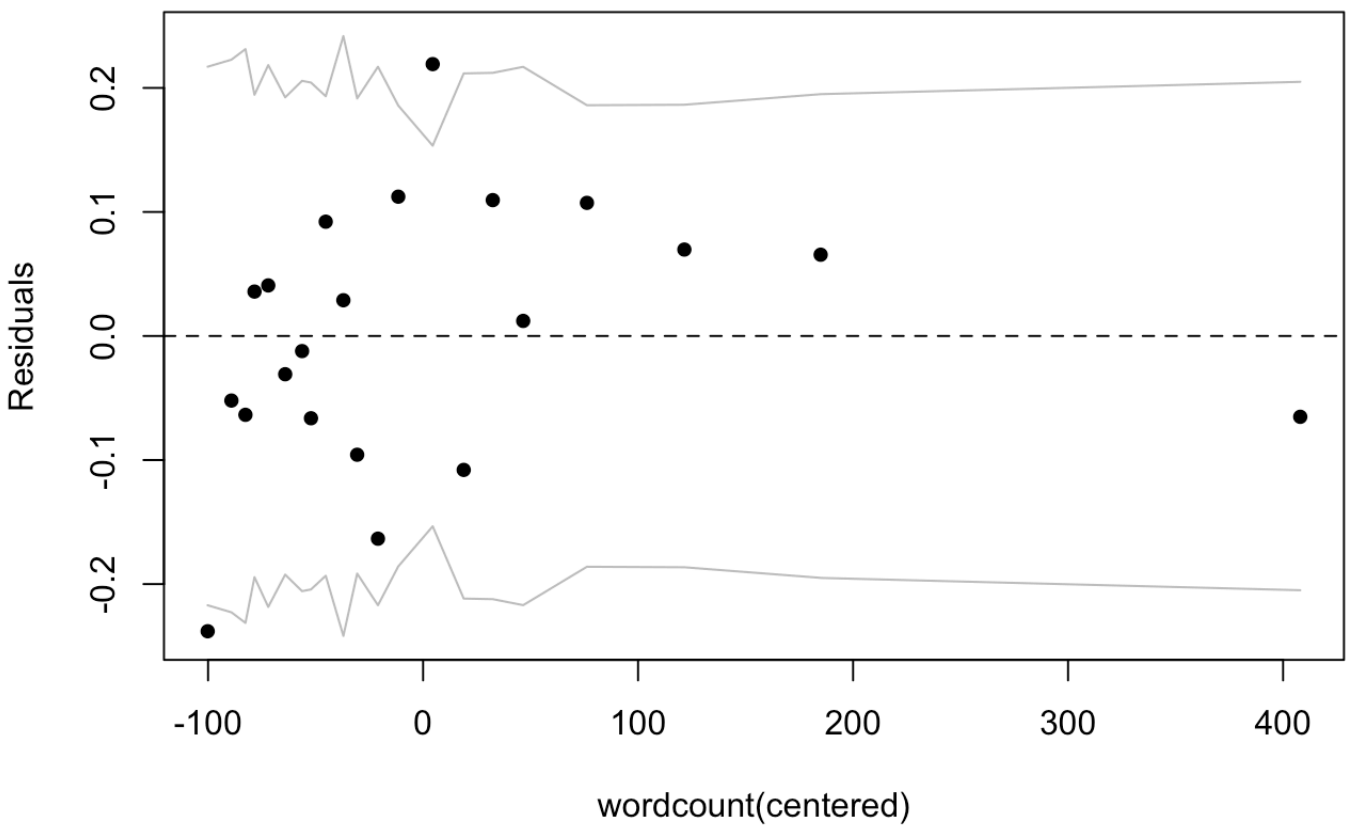
Subject of the story has a strong effect on odds of positive sentiment score. The baseline of “Bullying” has the lowest odds among all, while “Relationships” is the highest, and “puberty” in the middle. The odds of “Relationships” is 87% higher than that of baseline, and “puberty” is 61% higher.(exp(0.625) = 1.87, exp(0.478) = 1.61, 95% CI of (1.09, 3.20) and (0.84, 3.10) repectively)

Residuals Examination:

```
binnedplot(x=regdata2$Age, y = regdata2$sentiment2 - fitted(reg5), xlab = "Age", ylab
= "Residuals", main = "Binned residuals versus Age")
```




```
binnedplot(x=regdata2$wordcountc, y = regdata2$sentiment2 - fitted(reg5), xlab = "wor
dcount(centered)", ylab = "Residuals", main = "Binned residuals versus wordcount")
```




```
resid1 = regdata2$sentiment2 - fitted(reg5)
tapply(resid1, regdata2$Gender, mean)
```

```
##              female                male               other
##        4.335507e-09        4.301511e-17       -6.505213e-17
## Prefer not to share
##       -8.789266e-17
```

```
tapply(resid1, regdata2$LGBTQ., mean)
```

```
##                  no Prefer not to share                     yes
##        2.973400e-17        5.913872e-09       -3.634586e-17
```

```
tapply(resid1, regdata2$Reg, mean)
```

```
##       midwest     northeast     southeast     southwest          west
##  2.356391e-17  1.688566e-08  1.351387e-16 -1.171426e-16 -1.618301e-17
```

```
tapply(resid1, regdata2$Submission..Type, mean)
```

```
##            1            2            3    in person
## 2.761102e-18 1.258524e-16           NA 1.283310e-06
```

```
tapply(resid1, regdata2$Subject, mean)
```

```
##      bullying       puberty relationships
##  1.052099e-16  5.520534e-17  4.824474e-09
```

The Residual binned plots show no clear trends for the used model. The residual tables show no clear bias on the model. The model is reasonable and the assumptions are met.