Analysis on Teenager Sentiments

- Based on data from Real Talk

By Frank Xu
Fall 2018
IDS-702

Quick Fact About the Data

- *"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.

Question on Data

- Is there any association between positive/negative sentiment stories and demographic information that come with the stories?
- Is the positive/negative sentiment stories associate with some outer source data such as teenager pregnancy/suicide and political status of each states?

Code Book

1. Submission..Type: the platform used for each submission.

2. 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.

3. Gender: the gender of the user who published the story. Male/female/other/prefer not to share.

4. LGBTQ.: the LGBTQ. status of the user who published the story. Yes/no/prefer not to share.

5. Age: the age of the user who published the story.

6. Race: the race of the user who published the story.

7. zip: the zip code users submitted when signing up for the account.

8. Subject: the subject of each story posted. Users had to choose one subject before submitting each story.

9. Story: the exact text of story of each story posted. Can be as shorted as several words, or several sentences.

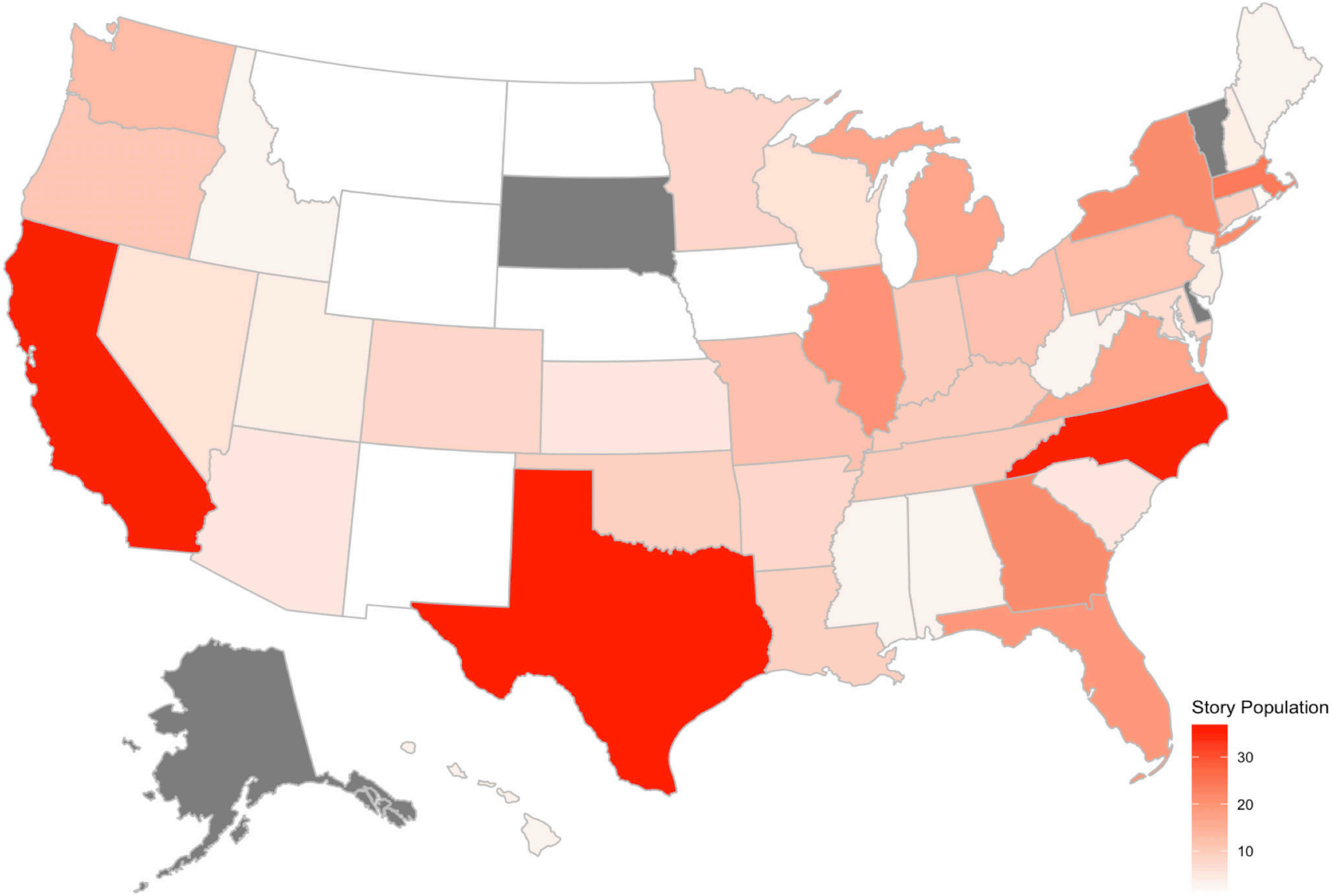
Missing data

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

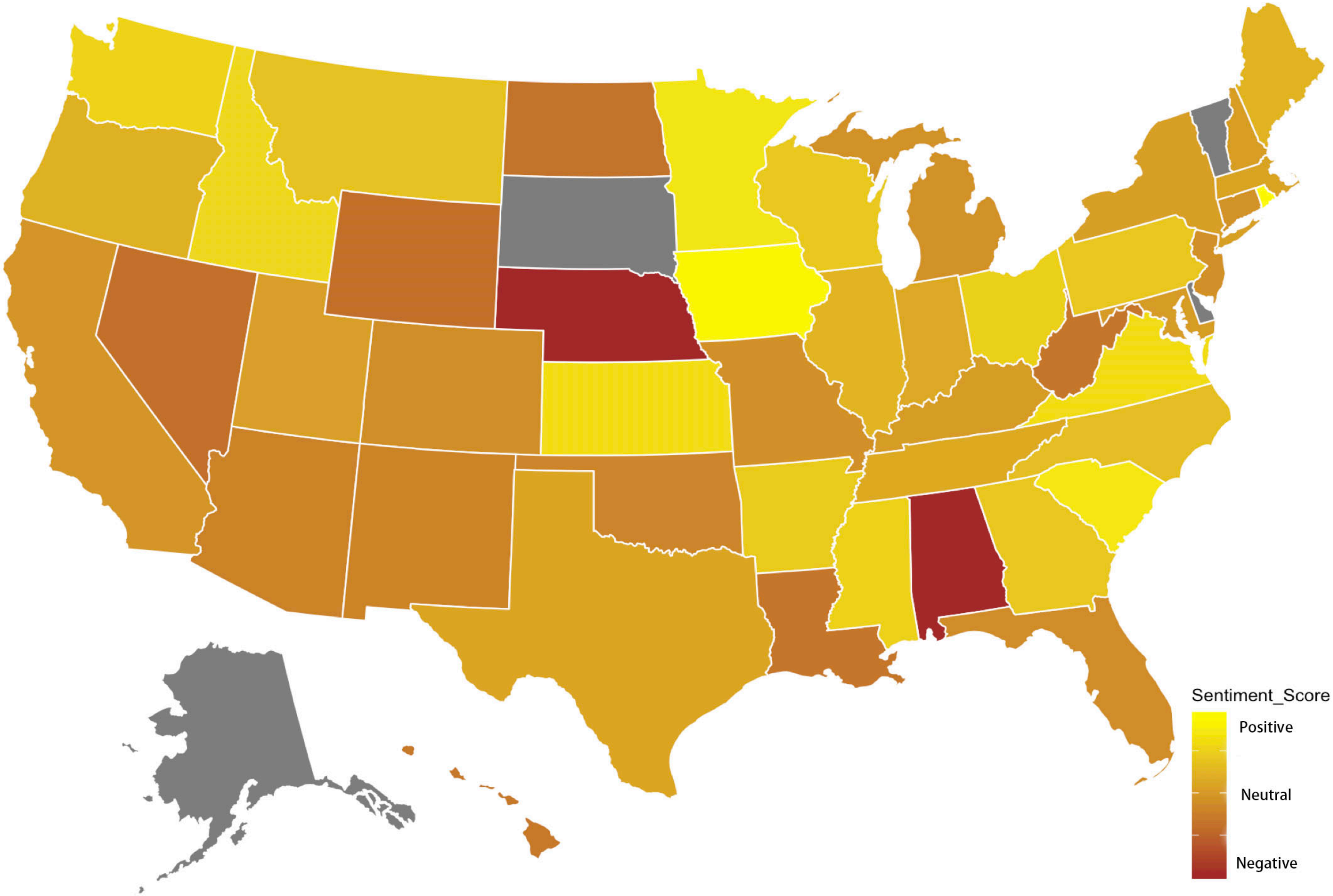
Outer Sources

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).

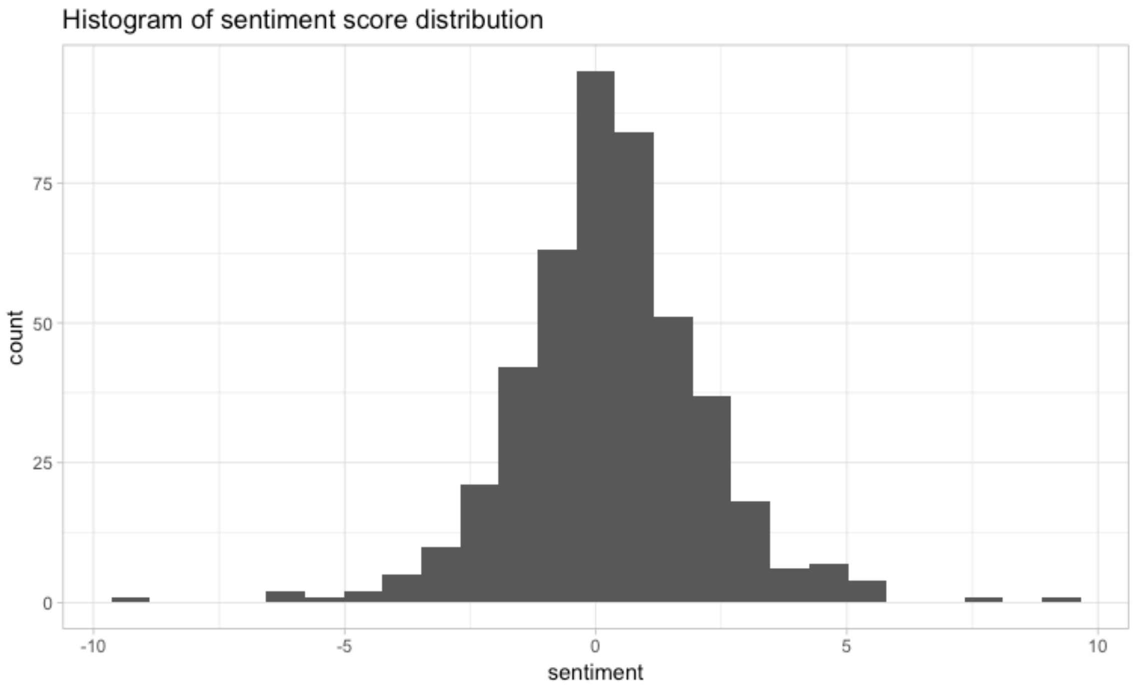
Visualization of overall data distributed over states:

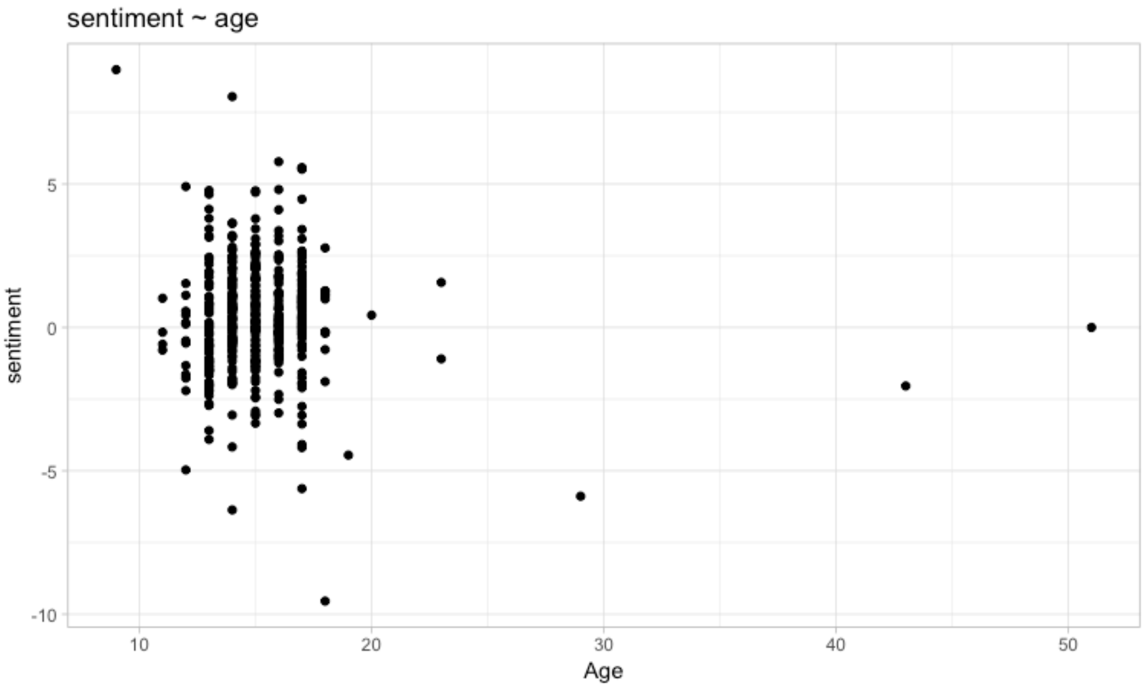


Visualization of average sentiment score over states:

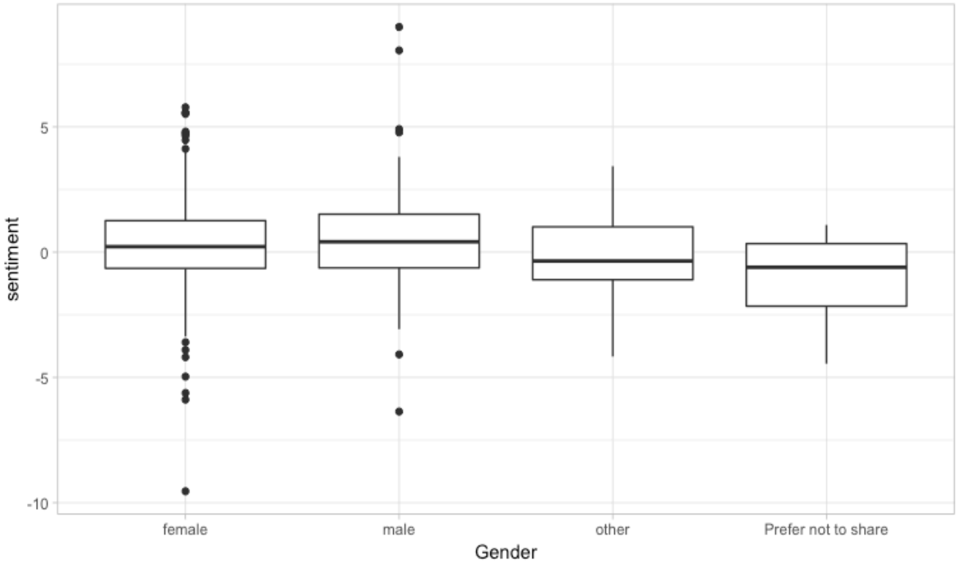


Exploratory **D**ata **A**nalysis

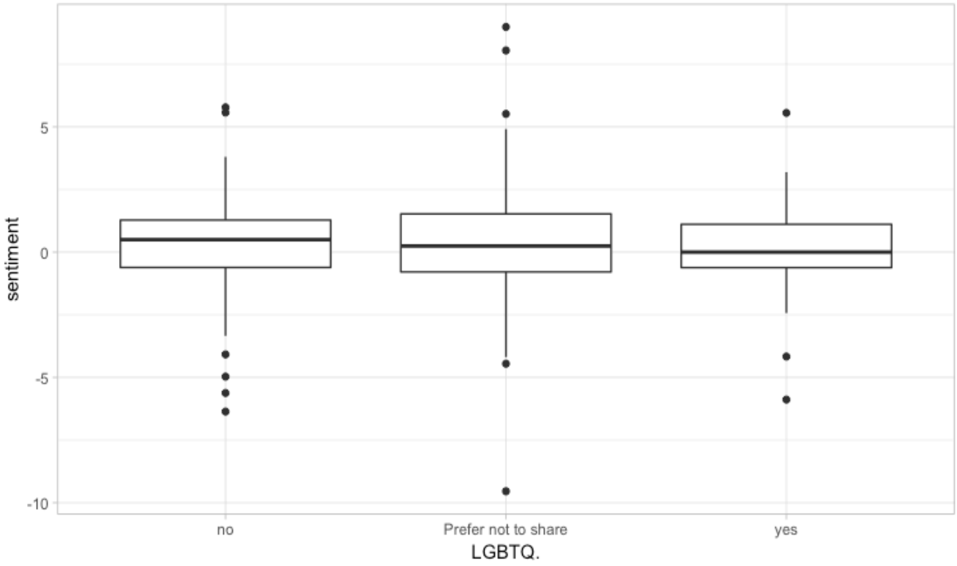




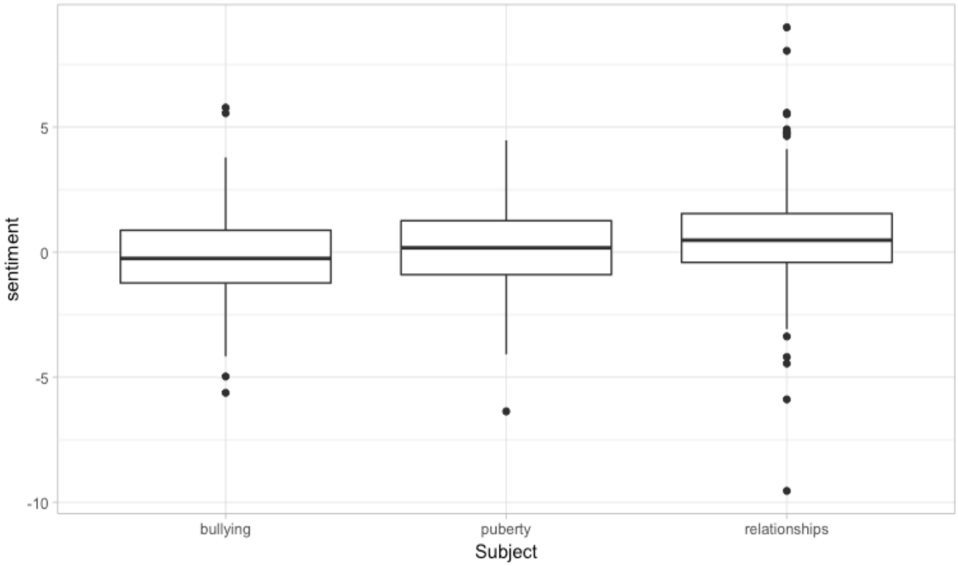
sentiment ~ gender



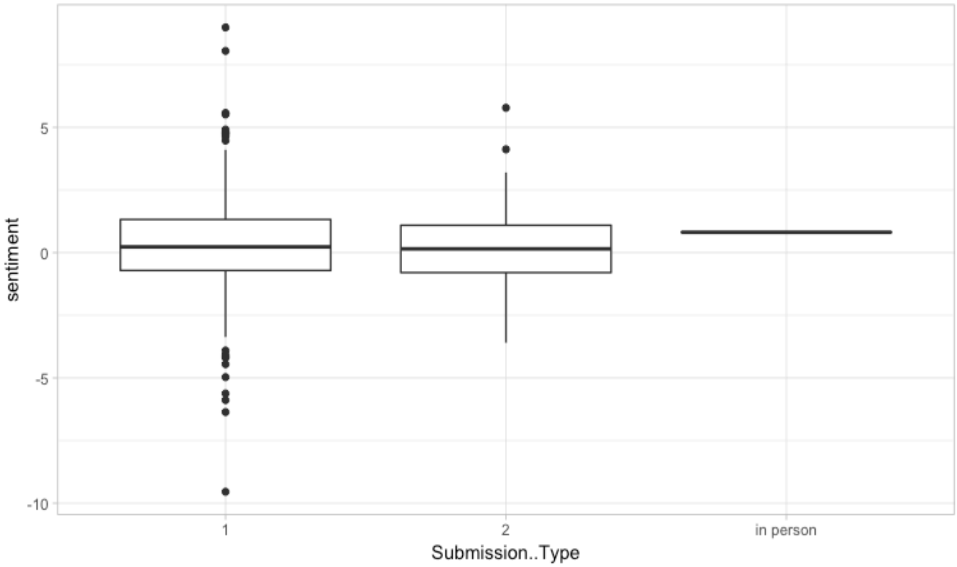
sentiment ~ LGBT status

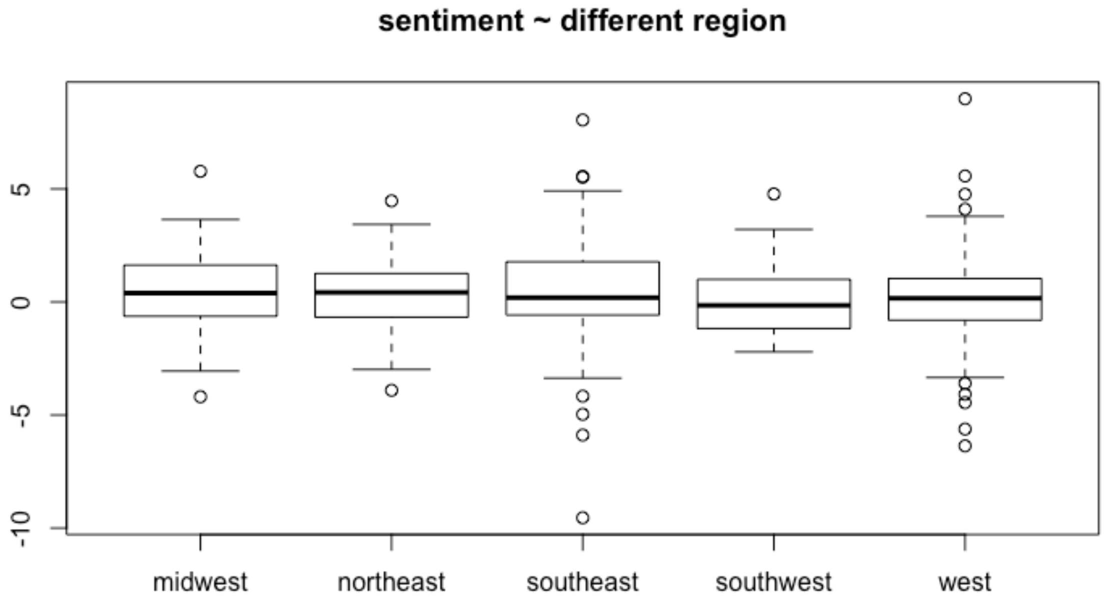


sentiment ~ different subjects

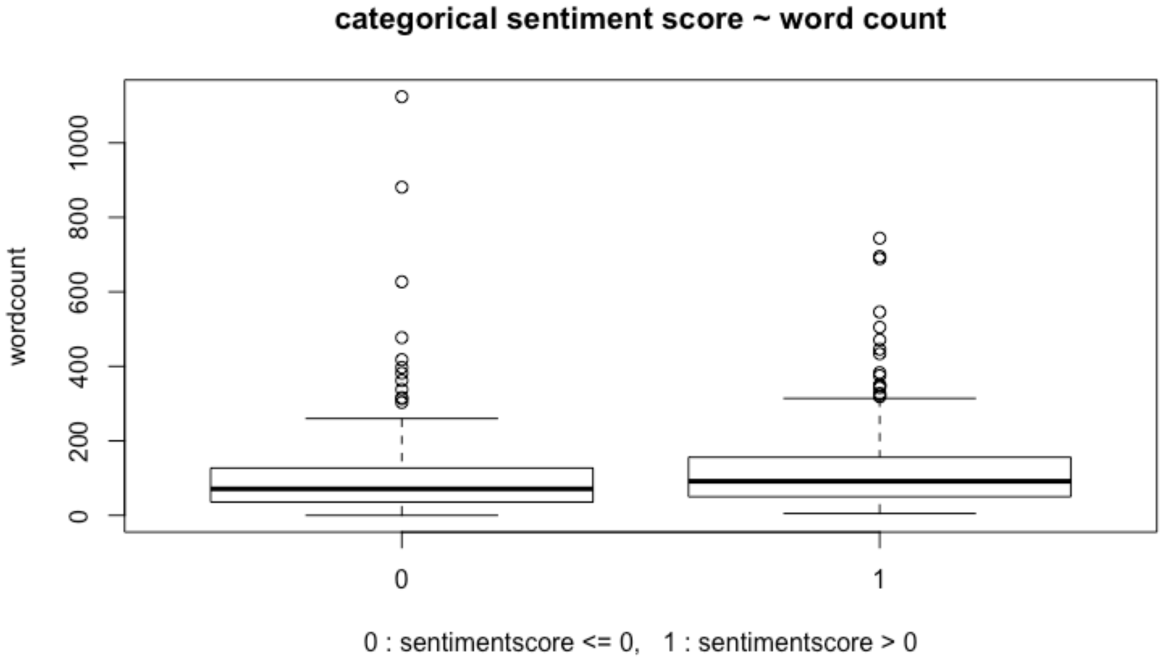


sentiment ~ different submission types

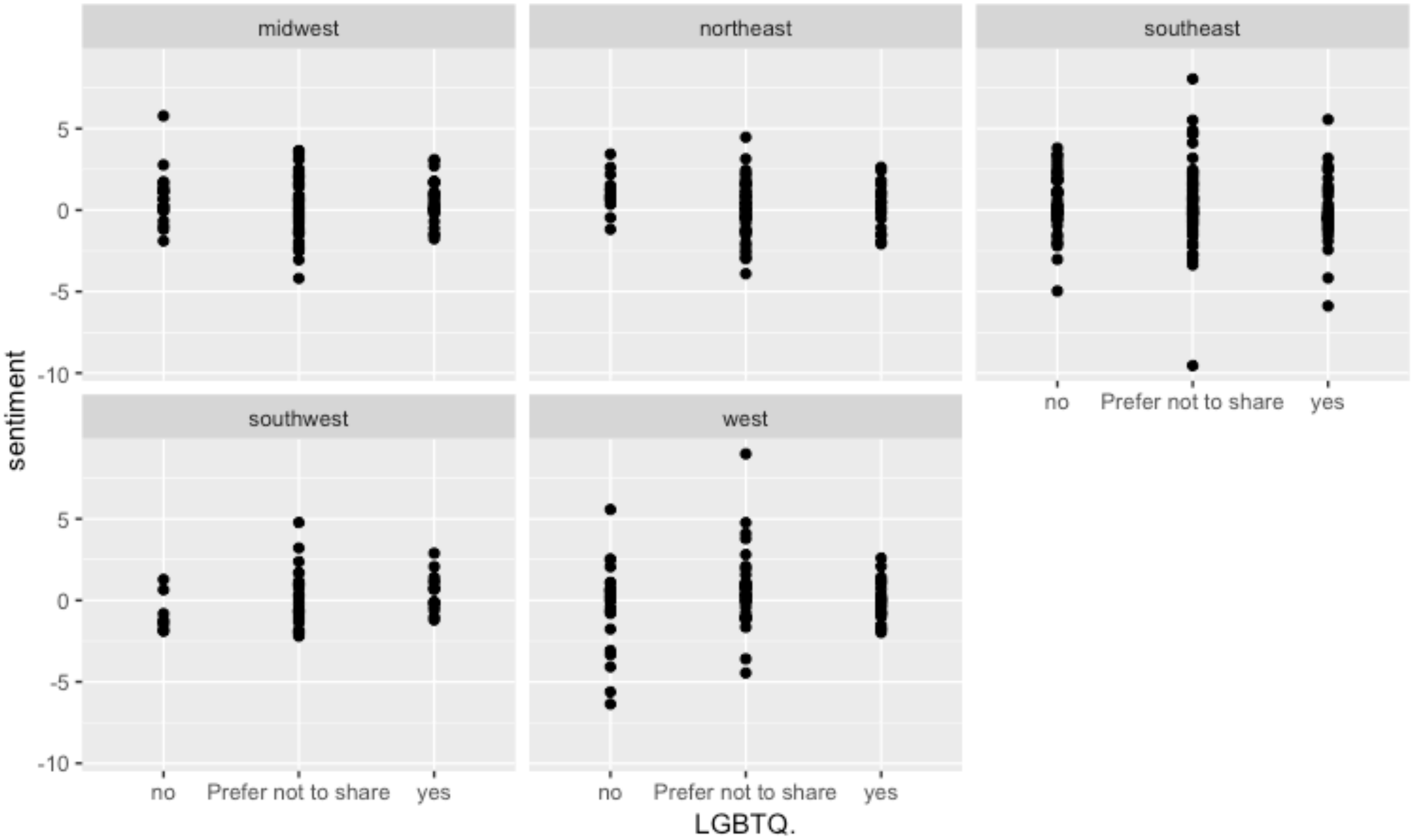




sentiment ~ different region



categorical sentiment score ~ word count


midwest
northeast
southeast
southwest
west
sentiment
5
0
-5
-10
5
0
-5
-10
no
Prefer not to share
yes
no
Prefer not to share
yes
no
Prefer not to share
yes
LGBTQ.

```
# 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")
anova(reg1, reg3, test= "Chisq")
anova(reg1, reg4, test= "Chisq")
```

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)
```

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 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:

- 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)).
- 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.
- 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.

- 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.
- 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

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.

Limitation and Discussion

- Sample size is too small: only 451 rows of data with full demographic information.
- Sample data has almost no correlation with outer sources.
- Sample data shows that the older you get, the more positive sentiment you can have (about teenager issues). Subject Bullying has most negative sentiment stories.
- Overall speaking, male has higher odds of positive sentiment than female. Non-LGBTQ people tend to have higher odds of positive sentiment, followed by people prefer not to share, then LGBTQ people.
- Different regions of the states treat LGBTQ teenagers differently. In Southwest, LGBTQ people have the most positive sentiment, while in most other region LGBTQ people have most negative sentiment.