EDA

EDA

We first load in necessary data frames. We will use `allcombined` as our first data frame, which includes original text and emotion, sentiment, toxicity, and violence analysis.

We also set colors for each author. Biden will be dark blue, and Trump will be red.

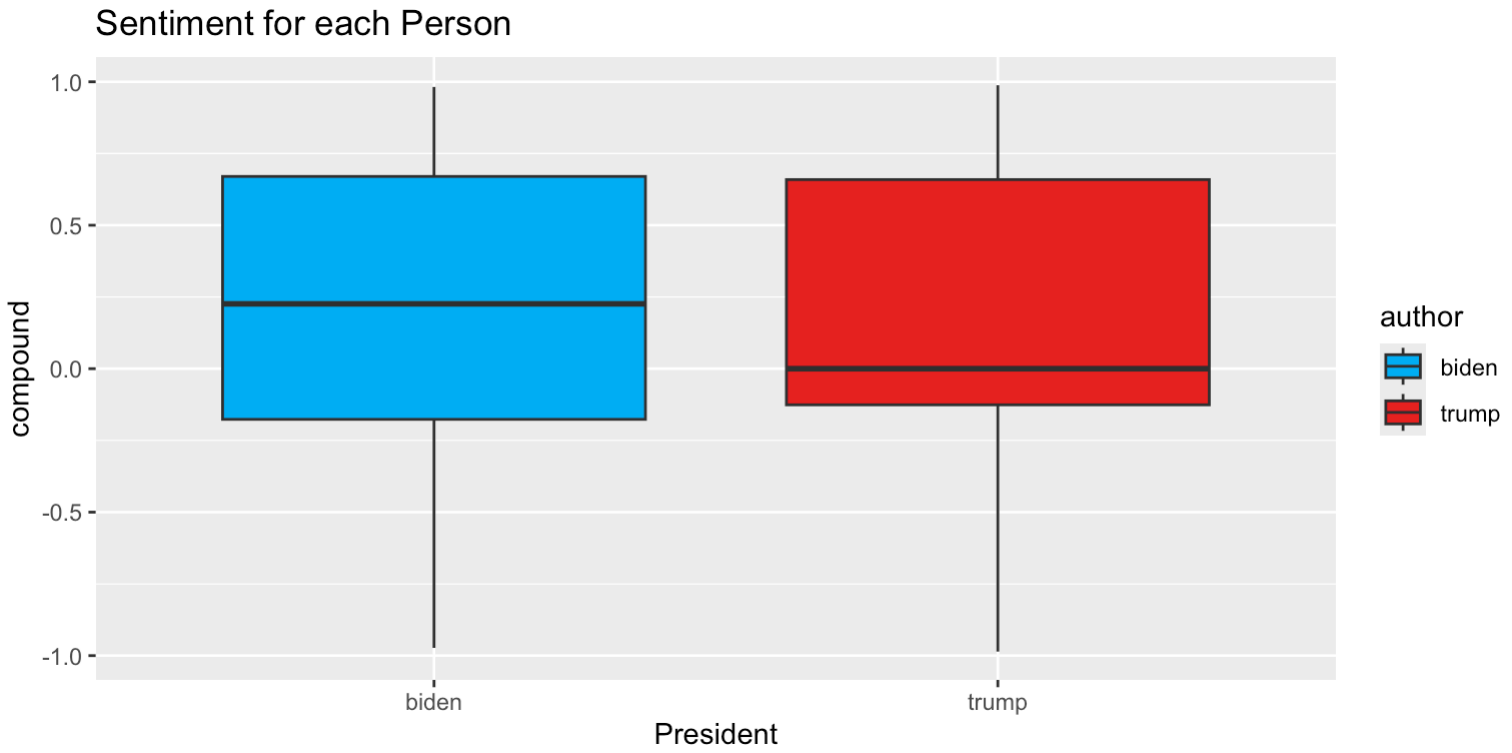
For our other categorical variable, `emotion_label`, we make JOY yellow, ANGER pink, and SADNESS cyan.

Selecting Validation Data

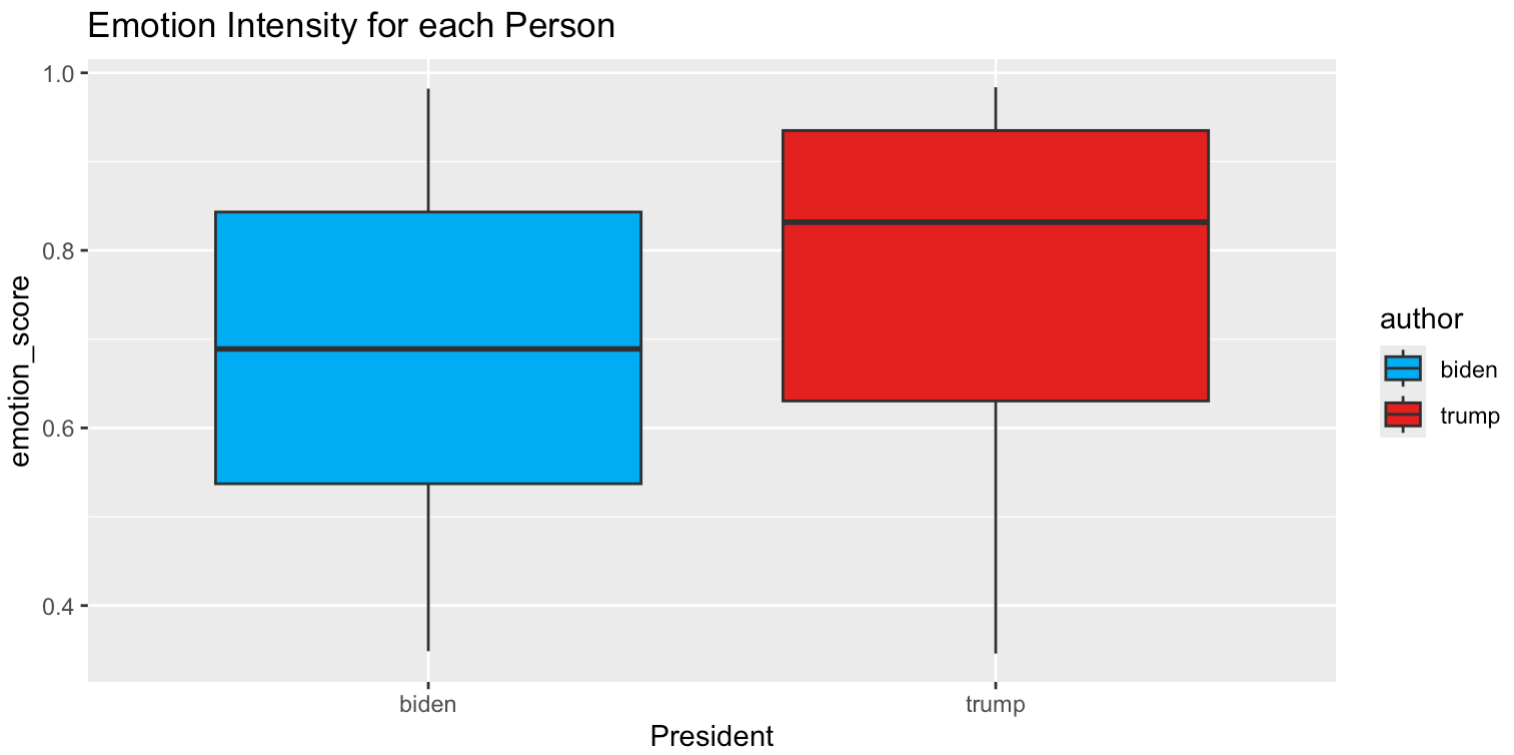
We must select validation data first. Then, we will save `cv_data` (used for training/testing) and `valid_data` as .csv files to use.

Sentiment for each Person

(Sentiment is expressed through the `compound` column)

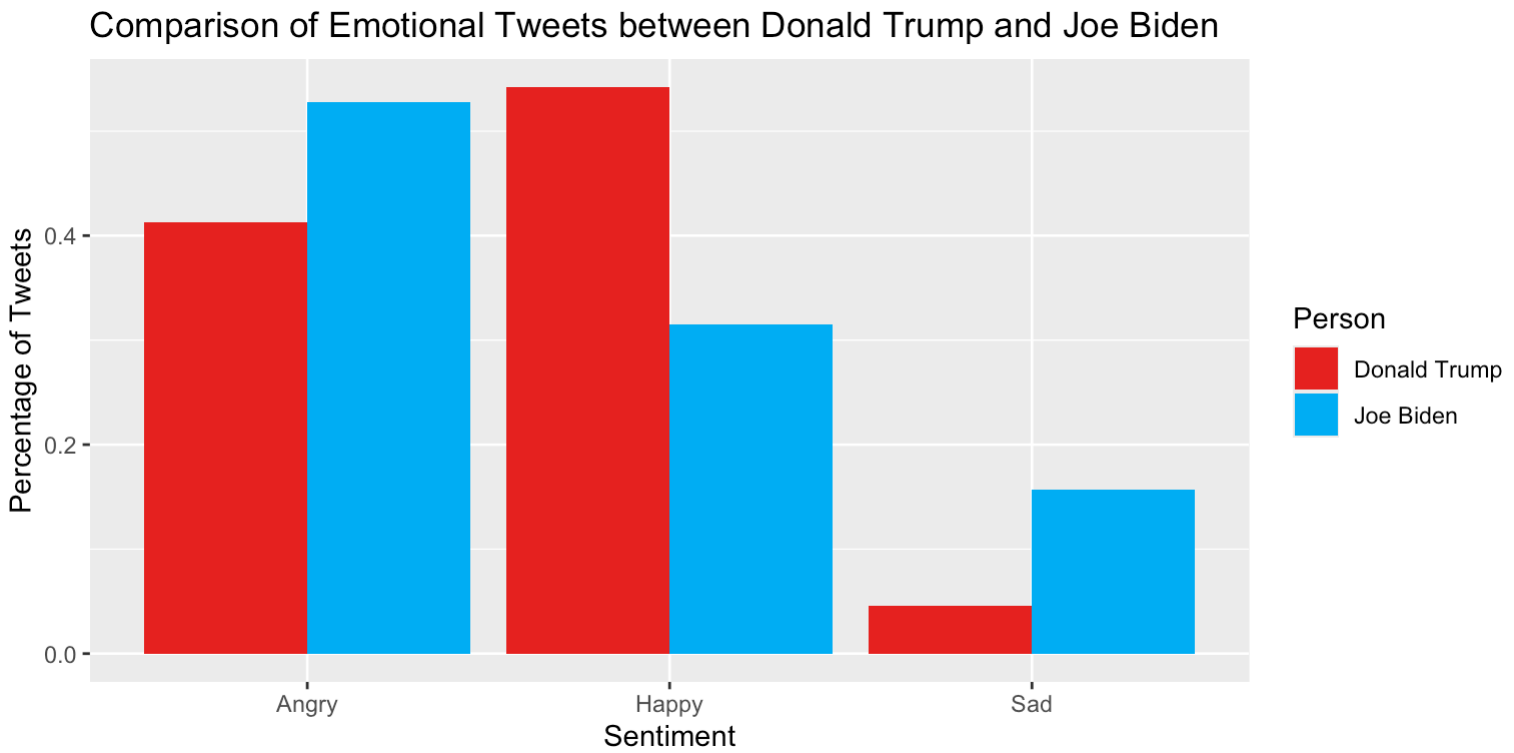


Emotion Intensity for each Person



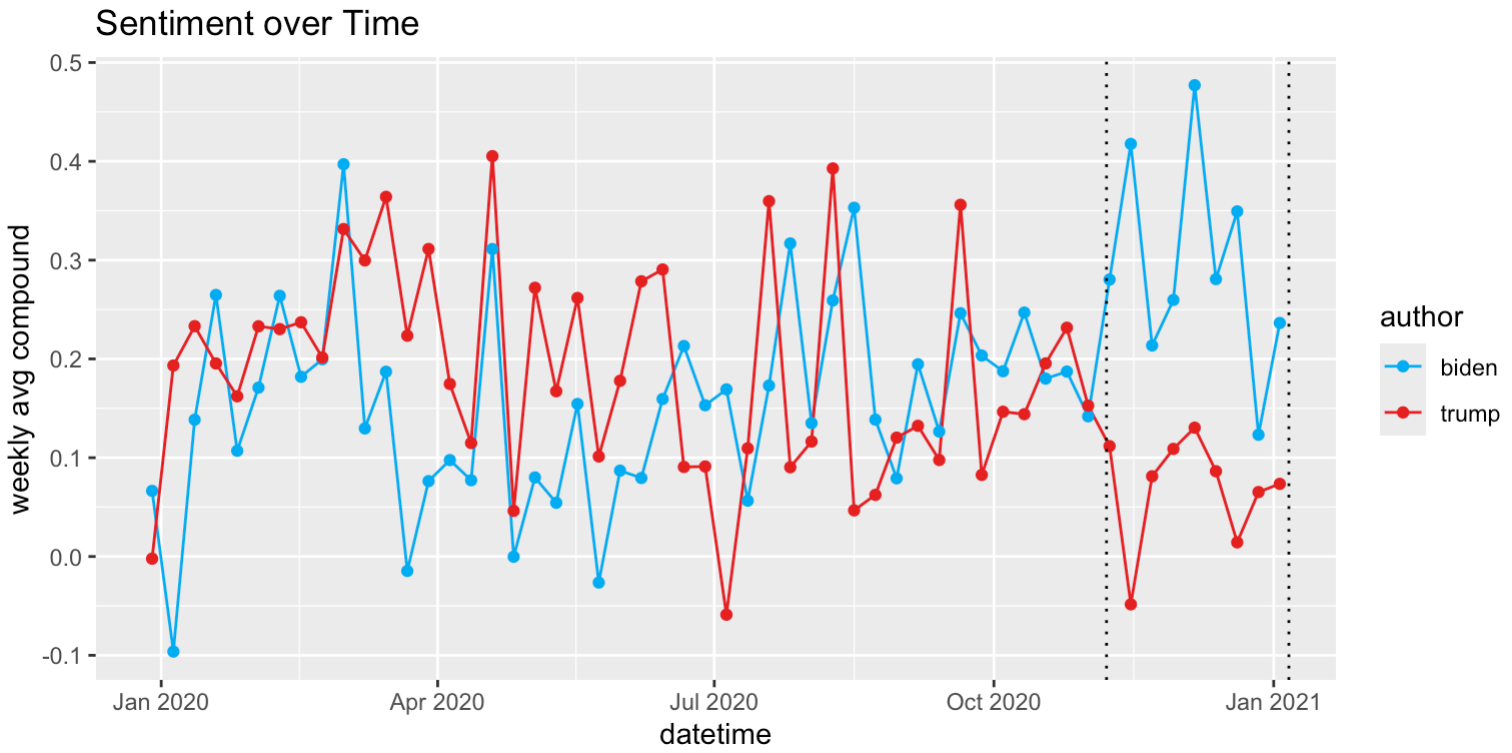
Frequency of Each Emotion (per person)

```
[1] 0.315

[1] 0.527

[1] 0.157

[1] 0.542

[1] 0.413

[1] 0.0454
```

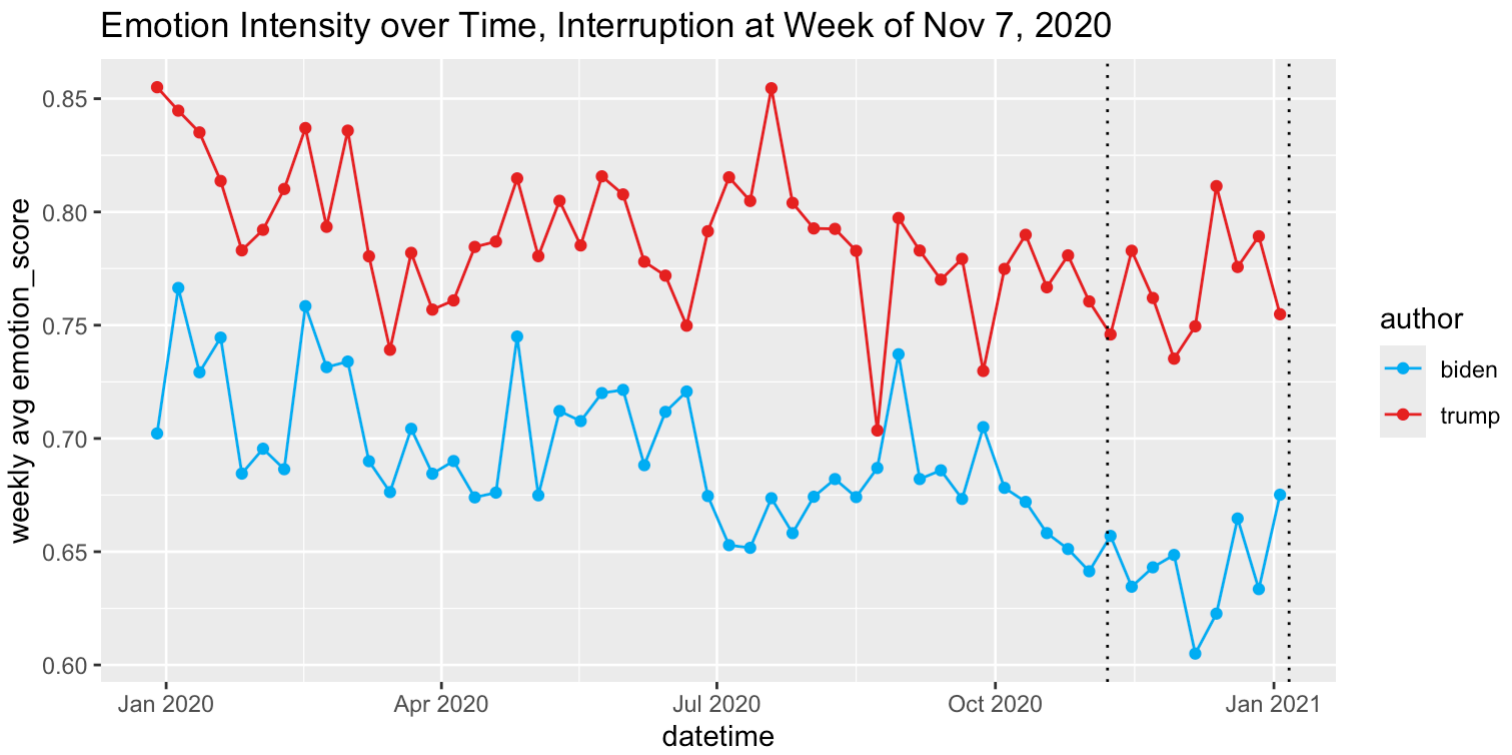


Sentiment over Time

```
`summarise()` has grouped output by 'week'. You can override using the
`.groups` argument.
```

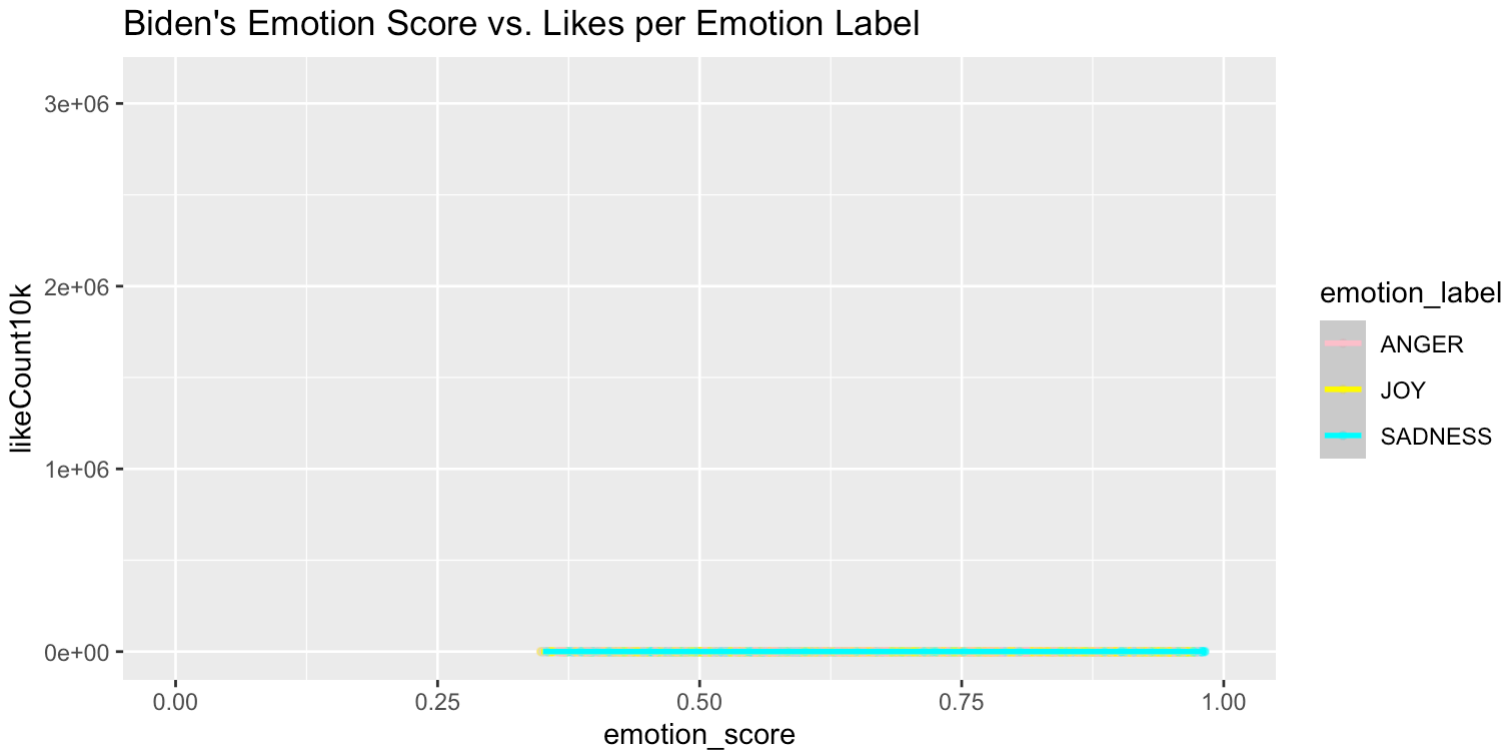


Emotional Intensity over Time

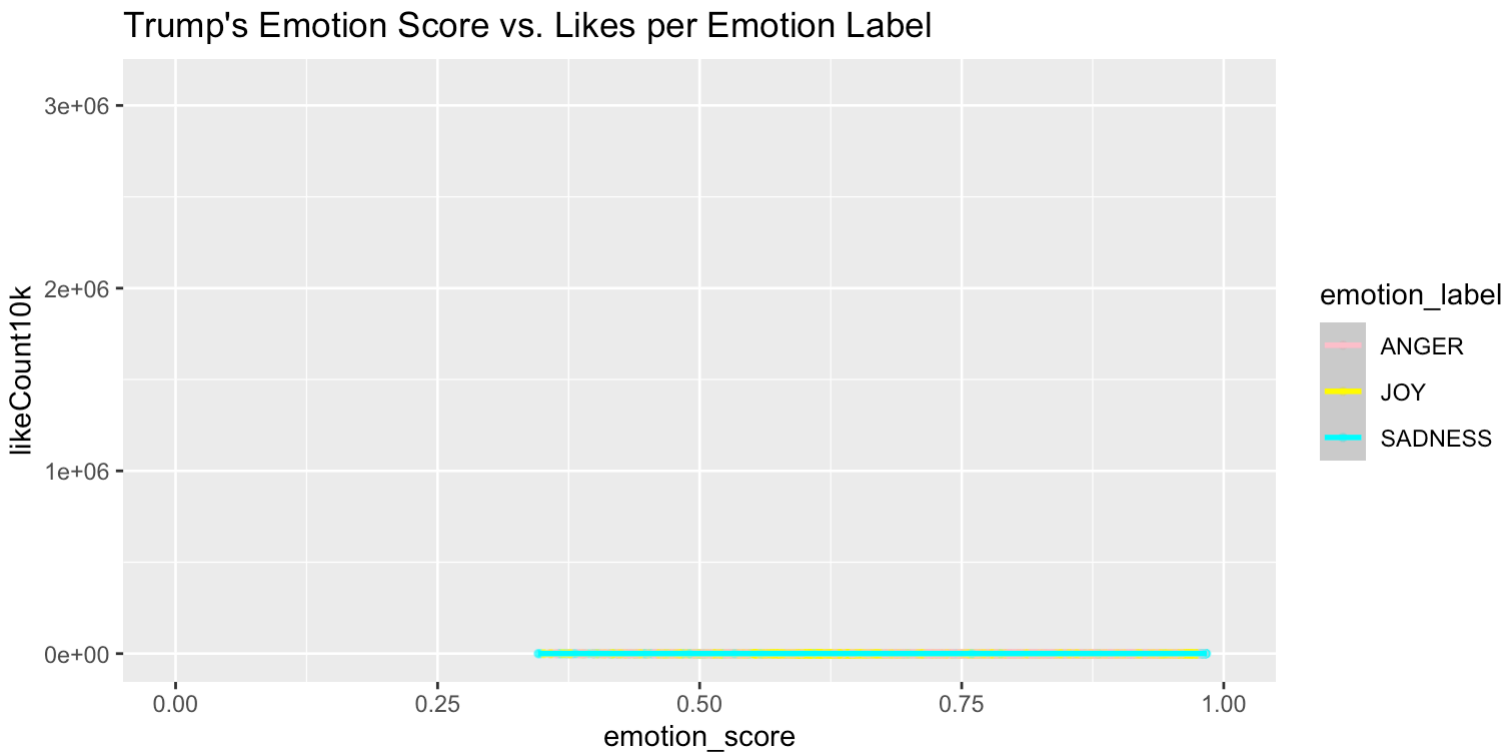


Likes vs. Emotional Intensity

```
`geom_smooth()` using formula = 'y ~ x'
```

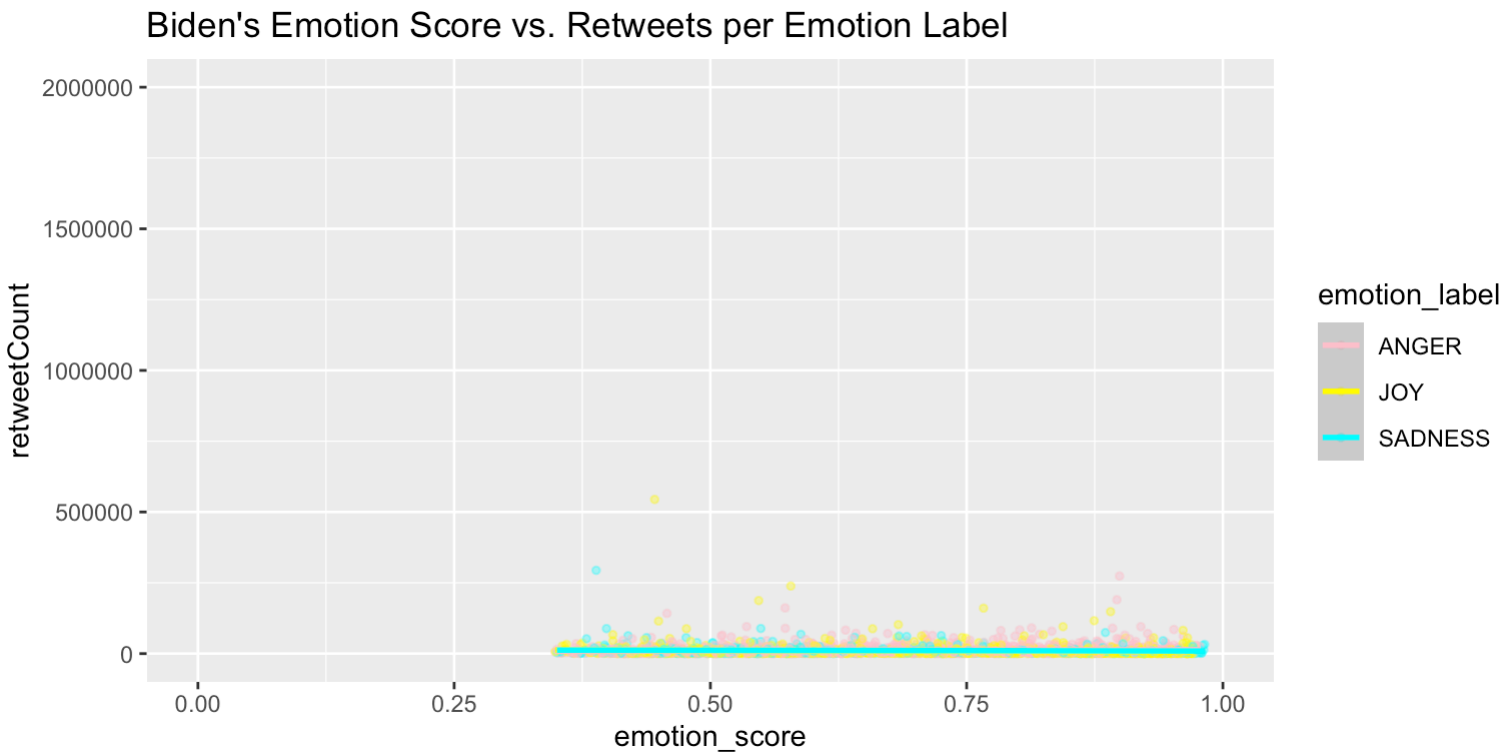



```
`geom_smooth()` using formula = 'y ~ x'
```

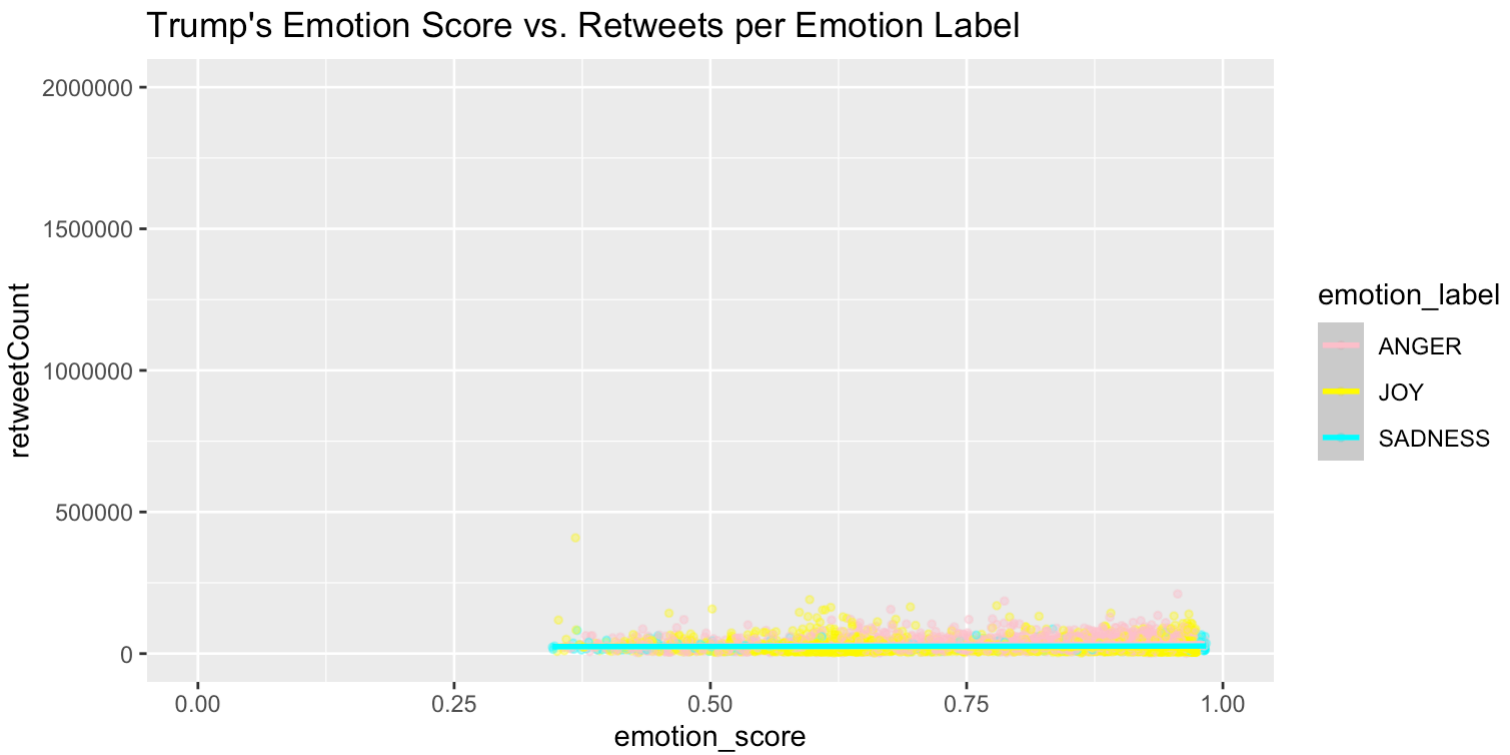


Retweets vs. Emotional Intensity

```
`geom_smooth()` using formula = 'y ~ x'
```

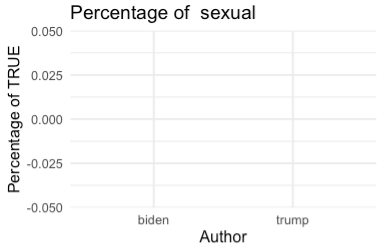


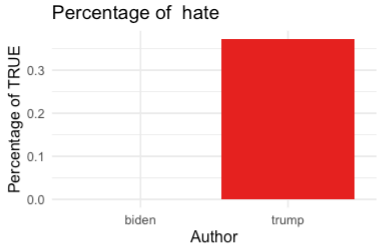

```
`geom_smooth()` using formula = 'y ~ x'
```

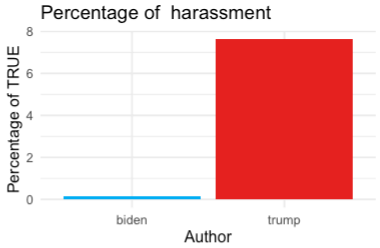


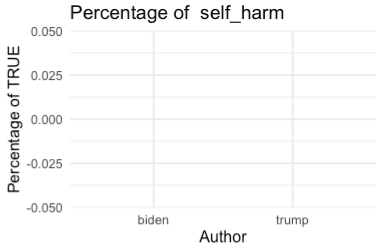
Moderation Data

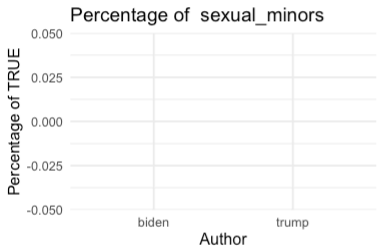
Moderation Flags in Trump's and Biden's Tweets

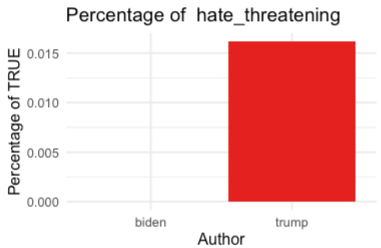

Percentage of sexual
Percentage of TRUE
0.050
0.025
0.000
-0.025
-0.050
biden
trump
Author

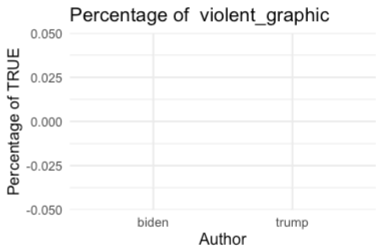

Percentage of hate
Percentage of TRUE
0.3
0.2
0.1
0.0
biden
trump
Author

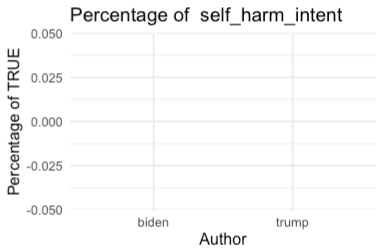

Percentage of harassment
Percentage of TRUE
8
6
4
2
0
biden
trump
Author

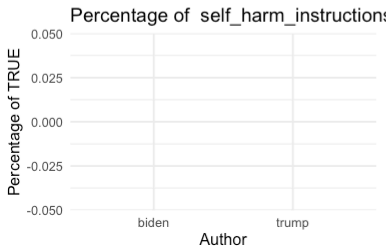

Percentage of self_harm
Percentage of TRUE
0.050
0.025
0.000
-0.025
-0.050
biden
trump
Author

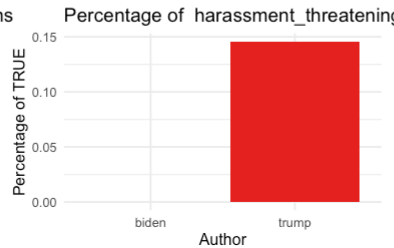

Percentage of sexual_minors
Percentage of TRUE
0.050
0.025
0.000
-0.025
-0.050
biden
trump
Author

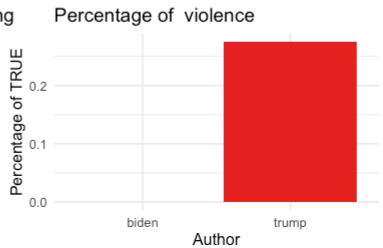

Percentage of hate_threatening
Percentage of TRUE
0.015
0.010
0.005
0.000
biden
trump
Author

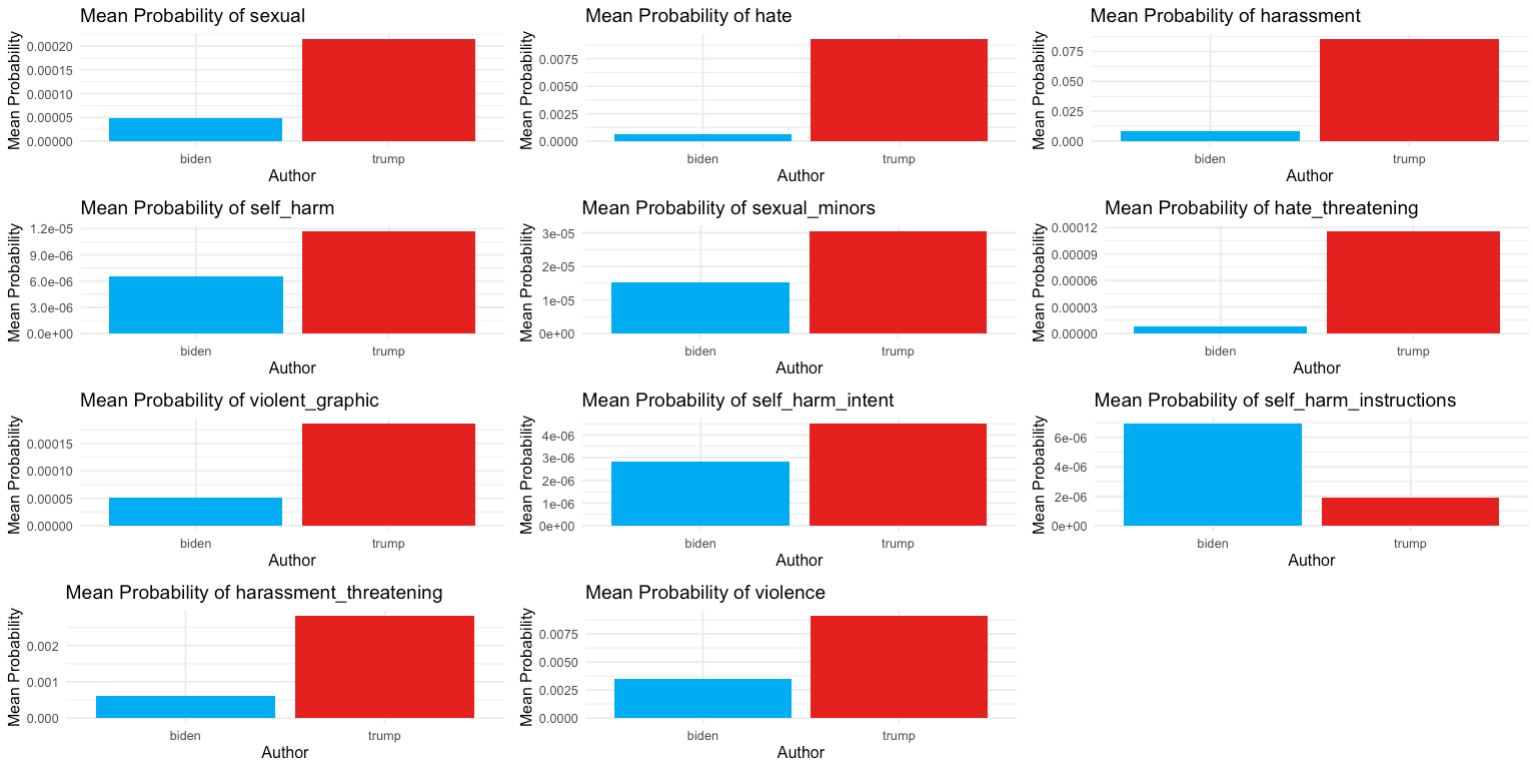

Percentage of violent_graphic
Percentage of TRUE
0.050
0.025
0.000
-0.025
-0.050
biden
trump
Author


Percentage of self_harm_intent
Percentage of TRUE
0.050
0.025
0.000
-0.025
-0.050
biden
trump
Author


Percentage of self_harm_instructions
Percentage of TRUE
0.050
0.025
0.000
-0.025
-0.050
biden
trump
Author


Percentage of harassment_threatening
Percentage of TRUE
0.15
0.10
0.05
0.00
biden
trump
Author

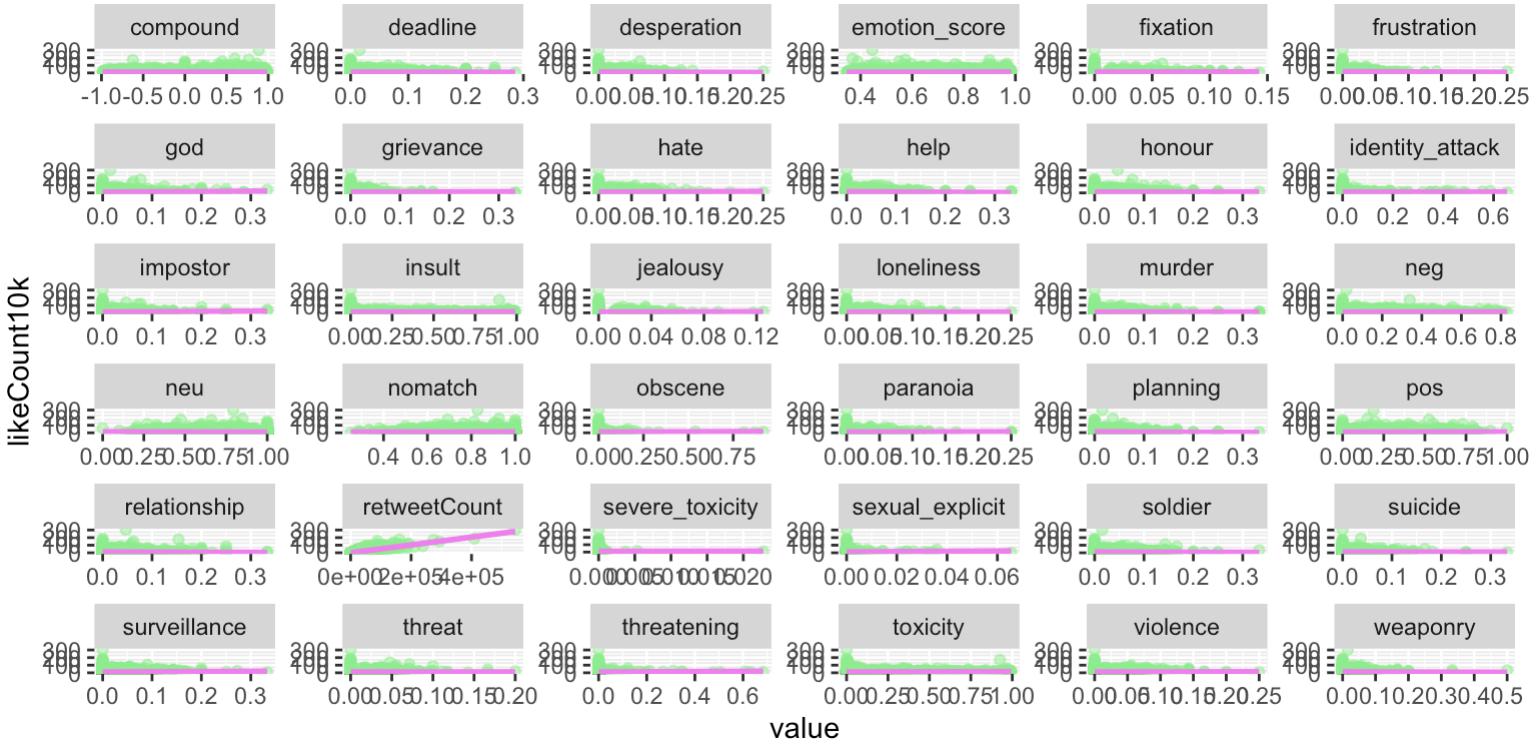

Percentage of violence
Percentage of TRUE
0.2
0.1
0.0
biden
trump
Author



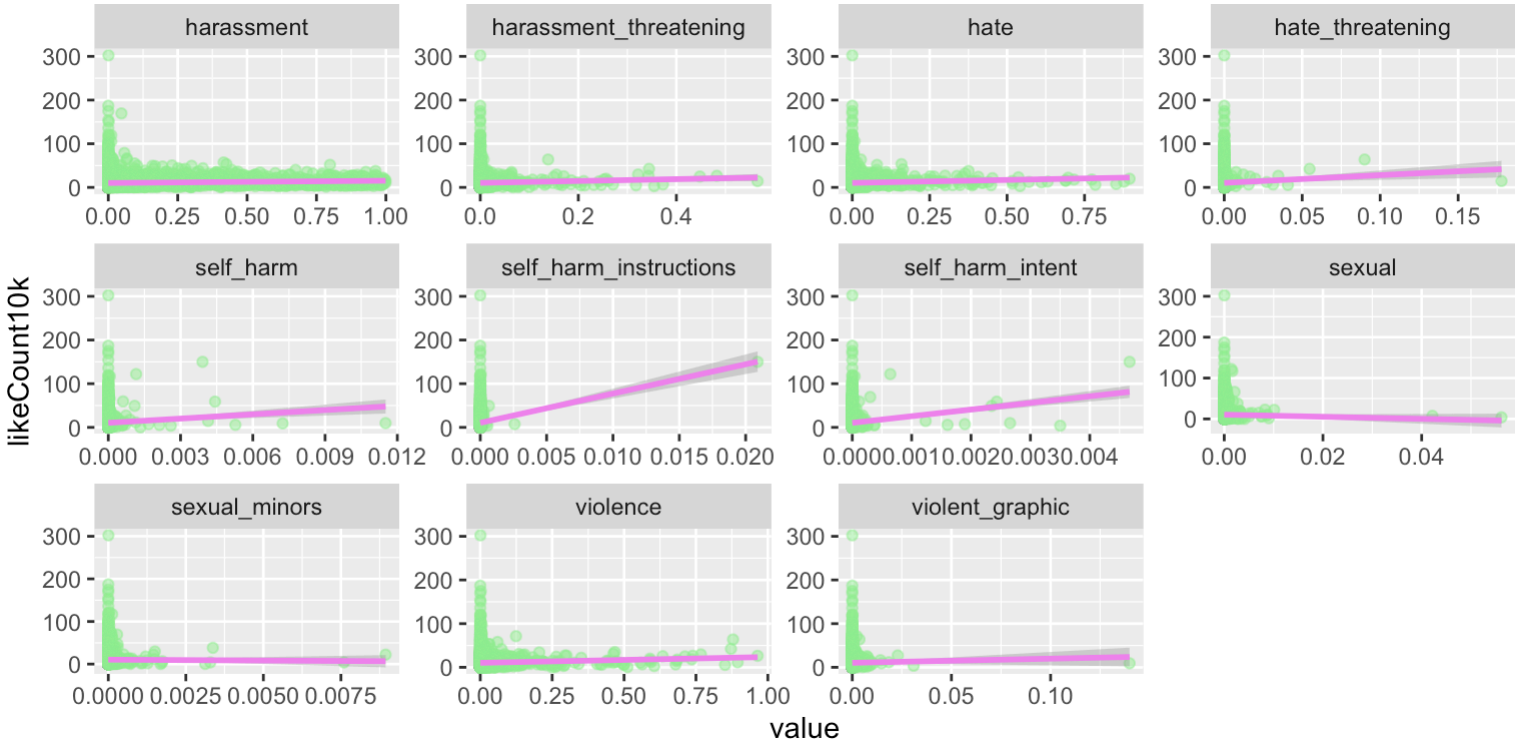
Likes vs. All Variables

We plot scatter plots of all variables against like count.

```
`geom_smooth()` using formula = 'y ~ x'
```




```
`geom_smooth()` using formula = 'y ~ x'
```


harassment
harassment_threatening
hate
hate_threatening
self_harm
self_harm_instructions
self_harm_intent
sexual
sexual_minors
violence
violent_graphic
likeCount10k
value

```
[1] 10.4
```