Analysis using R.

There will be 3 data sets. Two are parsed in different ways and the third one contains all the tweets collected.

```
library("readxl")
parsed_data <- read_excel(file.choose()) #parsed_twt_data.xlsx
```

```
New names:
* `` -> `...1`
* `` -> `...12`
* `` -> `...13`
* `` -> `...15`
```

```
general_data <- read_excel(file.choose()) # what.xlsx
maduro_data <- read_excel(file.choose()) # maduro_tweets.xlsx
```

Assigning variables: for the each of the data sets there will be a variable containing all the sentiment values, a variable for each of the 12 months,and a variable with the average sentiment value for tweets from the respective months.

```
par_val <- parsed_data$sen
par_months <- parsed_data$month
avg_months <- parsed_data$avg_month
par_avg_sen <- parsed_data$avg_sen

gen_val<- general_data$'sentiment number' #all the sentiment values for all data collected
gen_month <- general_data$'month'
gen_avg_months <- general_data$month_num # relevant months with avgs
gen_avg_sen <- general_data$avg_sen # avg gen sen for each month

maduro_val<- maduro_data$'sentiment values'
maduro_months<- maduro_data$'month_avg'
maduro_avg_sen<- maduro_data$'avg_sen'
```

This third data set only contains tweets that have the word "Maduro" in them. Only having one qualifier creates a data set with a very narrow scope of analysis – in this case, we can use it as a baseline to compare to the more general data set that was parsed using the python code above. In fact, here's how narrow that scope is:

```
length(gen_val)
```

```
[1] 2179
```

```r
length(par_val)
```

```
[1] 1659
```

```r
length(maduro_val)
```

```
[1] 110
```

The Maduro data set has 94.5% fewer data points than the general data set. Compare this to the parsed data set which only has 23.9% fewer data points. It is clear which data set was better able to represent sentiments of the general data set while also removing irrelevant data points that clouded up the data.

Now, before visualization begins, we have to ensure that removing the 'irrelevant tweets' actually resulted in a more focused data set with fewer outliers.

```r
sd(gen_val, na.rm=TRUE)
```

```
[1] 0.2986169
```

```r
sd(maduro_val, na.rm=TRUE)
```

```
[1] 0.1949718
```

```r
sd(par_val, na.rm=TRUE)
```

```
[1] 0.2447397
```

When comparing the Maduro data set to the other two, the standard deviation is so low that it cannot be representative of the majority of the data points. But, in the case of the data set with more general qualifiers, the standard deviation decreased by ~18%. This is a dramatic decrease but it is not so much of a decrease that the parsed data set is inadequate like the Maduro data set is.

Additionally, since the sd is lower for the parsed data set, it is more proof that the parsed data set is more focused than the general data set and we can now use it for further analysis.

Now, to gain a general understanding of the data:

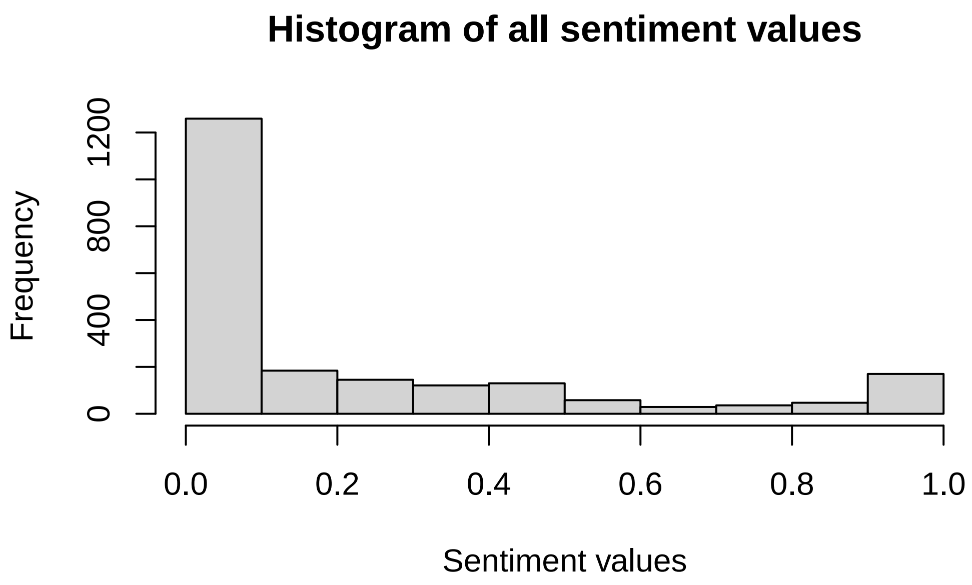
```
summary(gen_val, na.rm=TRUE)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
0.000000 0.009571 0.073617 0.224591 0.350372 0.999880 
```

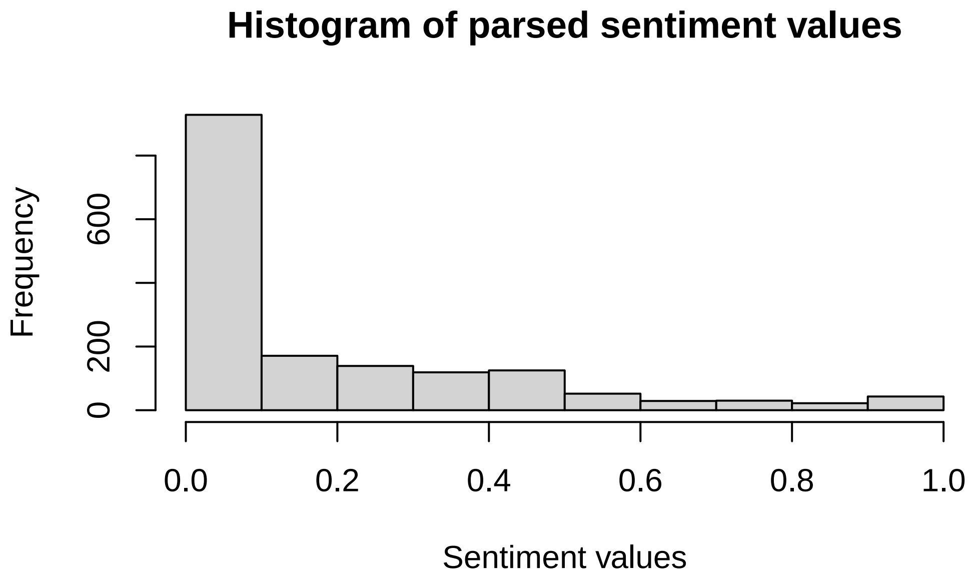
```
summary(par_val, na.rm=TRUE)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
0.000000 0.003572 0.057638 0.185105 0.306765 0.999880        1 
```

```
#Figure 1
hist(gen_val, main = "Histogram of all sentiment values", xlab = "Sentiment values")
```



```
#Figure 2
hist(par_val, main = "Histogram of parsed sentiment values", xlab = "Sentiment values")
```



There are two things to look at here.

First is the mean. A tweet with a sentiment value of 1 is totally positive. A tweets with a score of 0 is totally negative. Both of the mean values are very close to 0, but it is the difference between them is telling.

Since the parsed data set is without many of the irrelevant tweets that are found in the general data set, its average sentiment value is going to be more representative of the sentiments Ecuadorians actually felt towards Venezualens and the immigration crisis in 2015. And since removing those irrelevant data points resulted in a lower average sentiment value, it means that sentiments were potentially far more negative than what the data first showed.

And since these tweets are from 2015, right before the Venezualen migrant crisis truly started, it provides a baseline to compare future tweets to.

Second, the histogram helps visualize the degree that the irrelevant tweets were influencing the data. The amount of very positive tweets,(0.9 to 1.0), went from slightly below 200 to around a quarter as many. However, the positive tweets weren't the only ones removed. If that was the case then of course the average sentiment value would be lower. But since tweets were removed across the board, the lower average sentiment value is a more accurate representation of the data.

Placing qualifiers on the tweets in order to sort through the noise proved to be a positive action as the data would otherwise appear to be far more positive than what it actually was.

Visualization using R.

```r
library(ggplot2)
library(dplyr)
```

```
Attaching package: 'dplyr'


The following objects are masked from 'package:stats':

    filter, lag


The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(tidyverse)
```

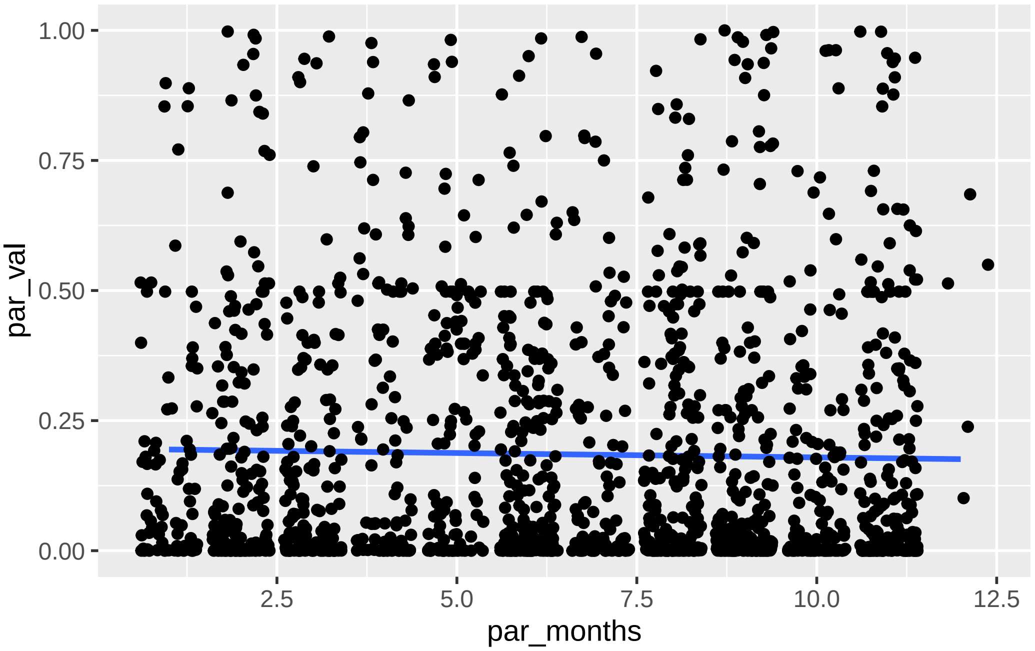
```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v forcats   1.0.0     v stringr   1.5.0
v lubridate 1.9.2     v tibble    3.2.1
v purrr     1.0.1     v tidyr     1.3.0
v readr     2.1.4

-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

Graph of all the parsed data values and graph of all the data points initially collected.

```r
#Figure 3: All parsed sen values
ggplot(data = parsed_data, mapping = aes(x = par_months, y = par_val))+geom_smooth(method
```

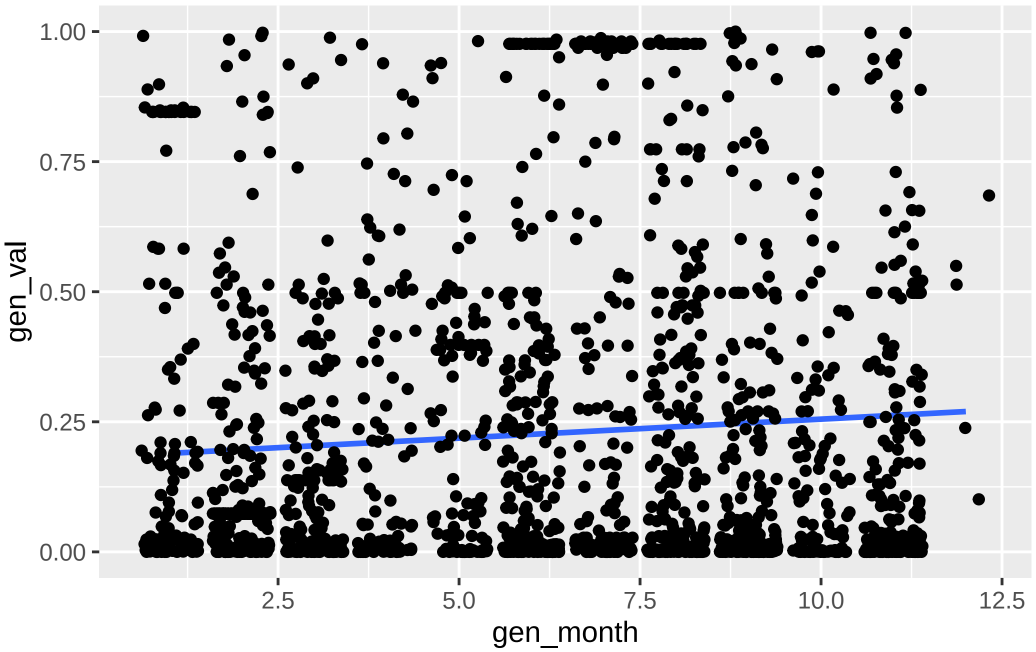
```
`geom_smooth()` using formula = 'y ~ x'


Warning: Removed 1 rows containing non-finite values (`stat_smooth()`).
```

```
Warning: Removed 1 rows containing missing values (`geom_point()`).
```



```
#Figure 4: All general sen values
ggplot(data = general_data, mapping = aes(x = gen_month, y = gen_val))+geom_smooth(method
```

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



In figure 3, the regression line is almost flat, meaning that there was not strong relationship between the sentiment values of tweets and when they were posted. People's opinions did not change as the year progressed. Meanwhile, the regression line for figure 4 shows that the sentiment values are becoming more positive as the year goes on. Which once again highlights how misleading the original data was since it contained so many irrelevant tweets.

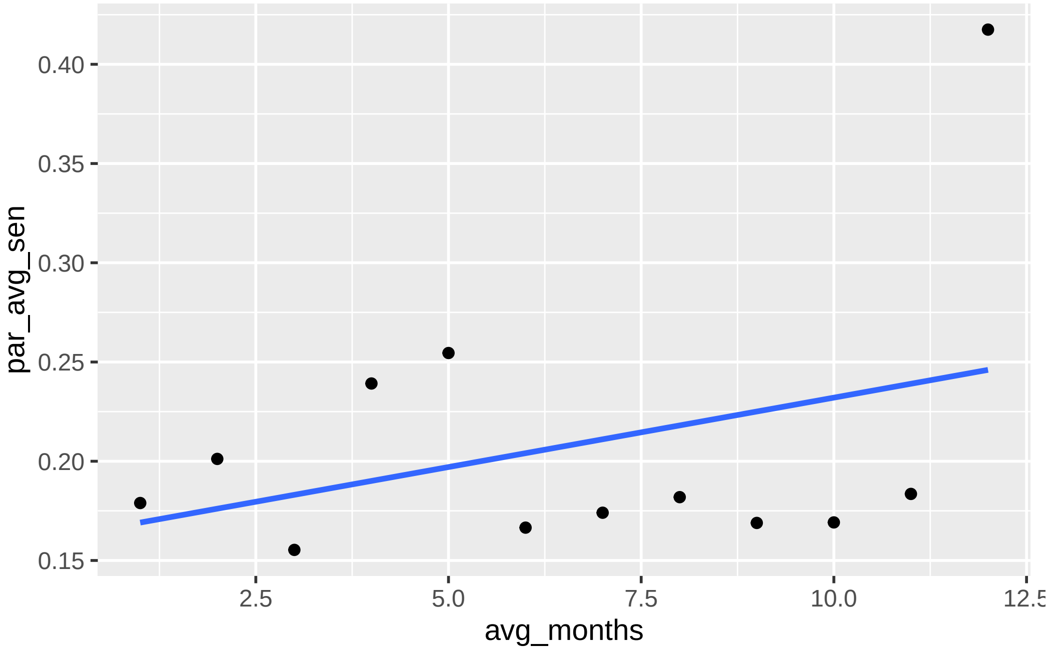
But, when you look at the monthly averages, you see a different story.

```
#Figure 5: Monthly averaged parsed sen values
ggplot(data = parsed_data, mapping = aes(x = avg_months, y = par_avg_sen))+geom_smooth(met
```

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

```
Warning: Removed 1647 rows containing non-finite values (`stat_smooth()`).
```

```
Warning: Removed 1647 rows containing missing values (`geom_point()`).
```

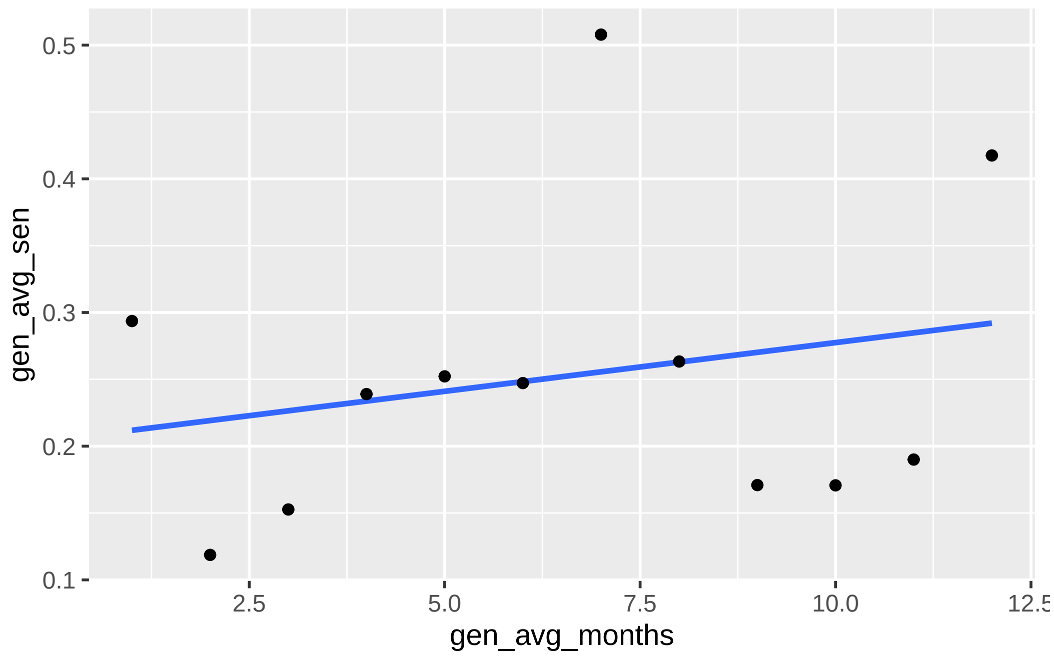


```
#Figure 6: Monthly averaged general sen values
ggplot(data = general_data, mapping = aes(x = gen_avg_months, y = gen_avg_sen))+geom_smoot
```

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

```
Warning: Removed 2167 rows containing non-finite values (`stat_smooth()`).
```

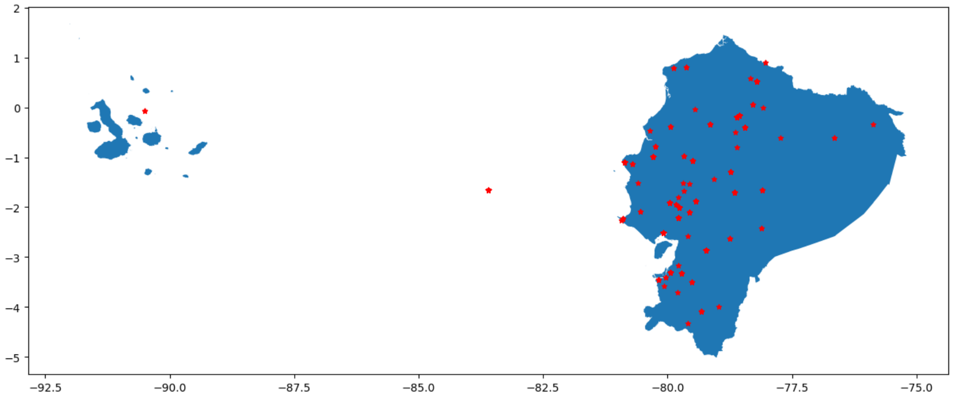
```
Warning: Removed 2167 rows containing missing values (`geom_point()`).
```



In both graphs, the regression line is increasing, implying that there is a strong relationship between the time of the tweet and its sentiment value. However, the points on figure 5 are not close to the regression line, unlike the points on figure 6. Since the points aren't clustered around the line, there is no strong relationship between the sentiment value and when it was tweeted. Esentially, the sentiment values do not necessarily increase as time passes leaving us with the conclusion we drew after viewing figures 3 – there is no strong relationship between the sentiment values and when the tweet was published.

This seems to back up the claim of the paper, ("Migrant Exposure and Anti-Migrant Sentiment: The Case of the Venezuelan Exodus"), that increased exposure to migrants does not necessarily lead to increased negative feelings towards migrants. And since these tweets were published right at the start of the Venezualan exodus, its a microcosm of what is to come since it captures the ramp up to the mass migration.

As a bonus: here's a map of the where the tweets were published, (python code can be found in the github repo):


2
1
0
−1
−2
−3
−4
−5
−92.5
−90.0
−87.5
−85.0
−82.5
−80.0
−77.5
−75.0