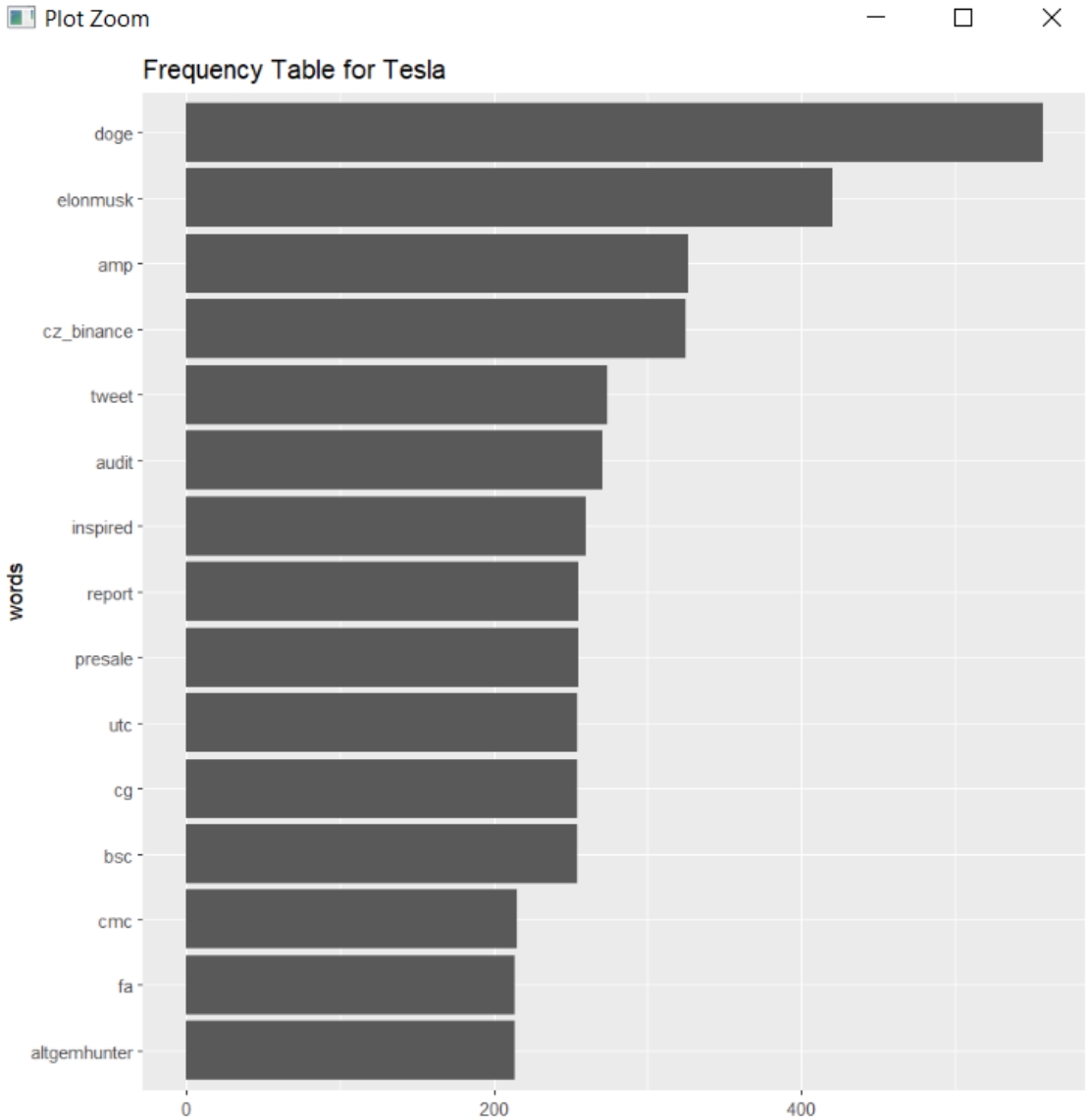
Name: Tien-Wei Hsu

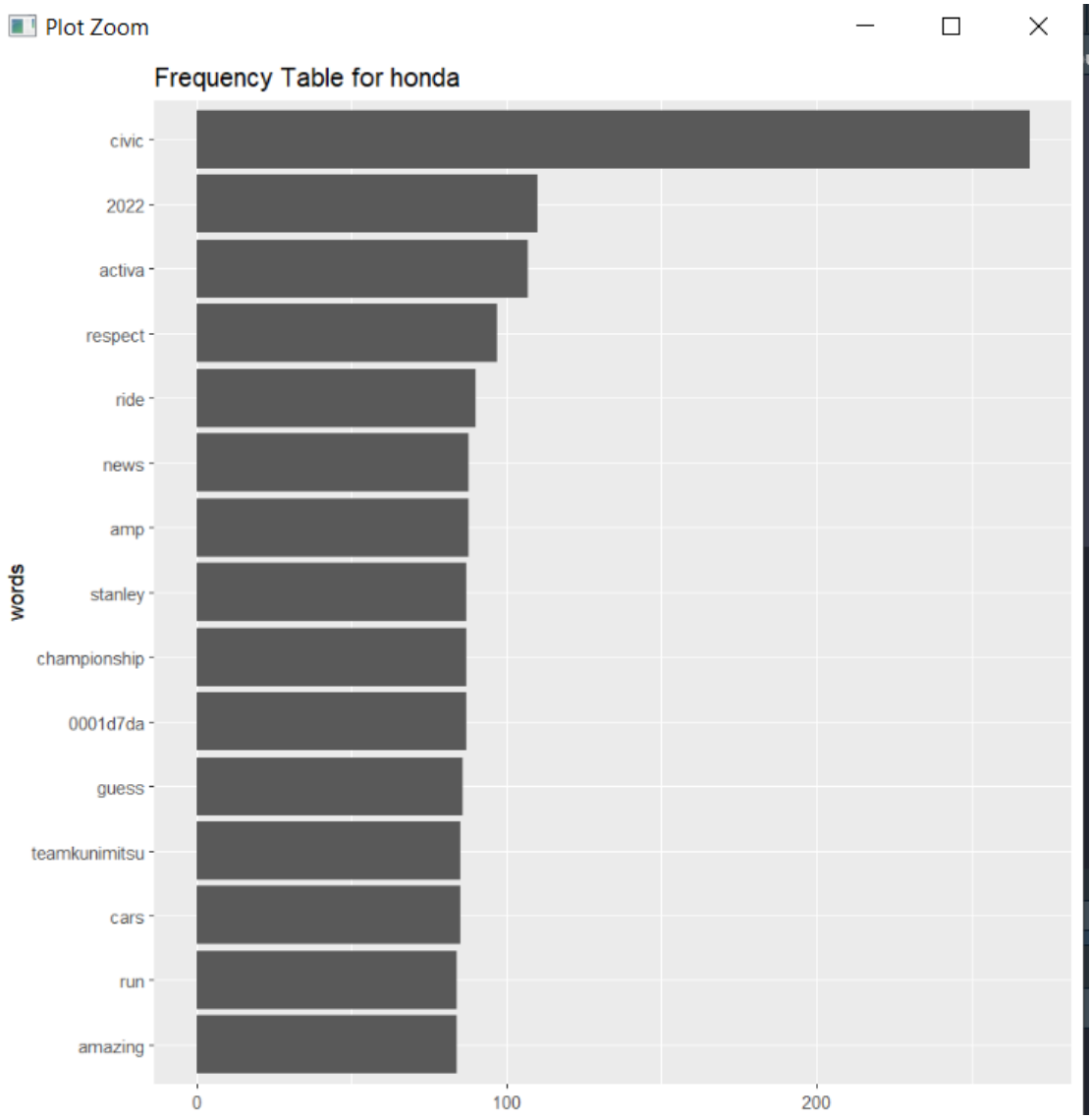
Class: NLP

Assignment: Business Insight

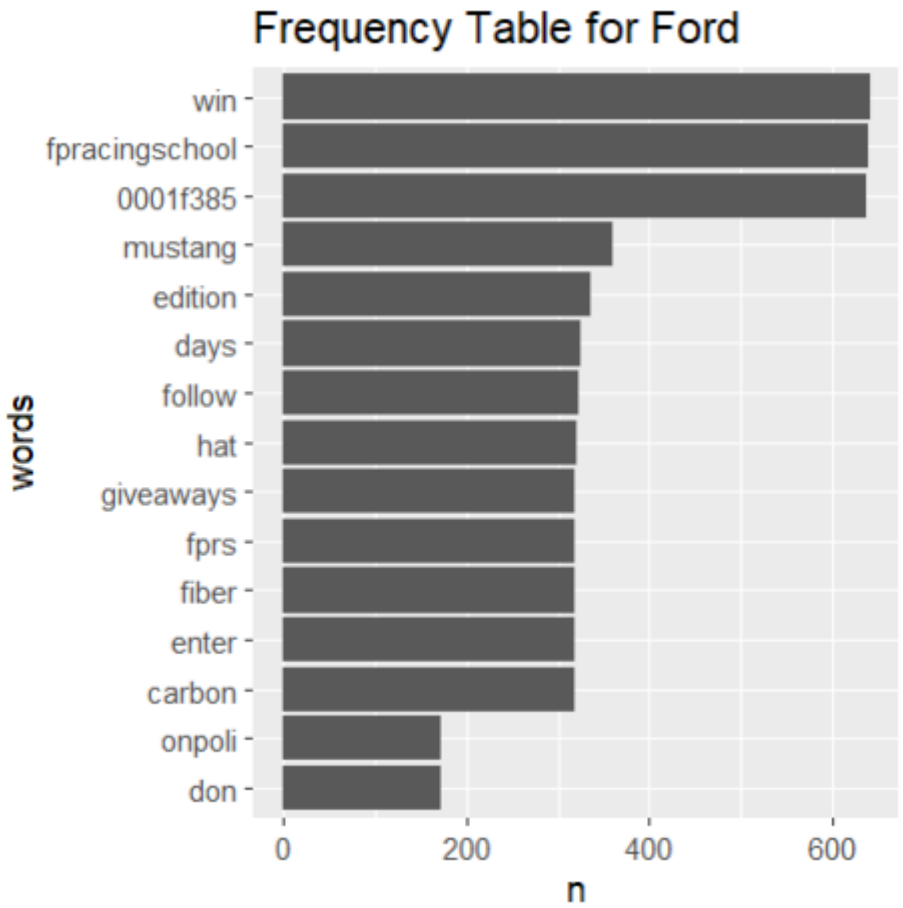
The topic of this Business Insight is about how to use the Twitter's Hashtag of Tesla, Honda, and Ford to find which part of Car area investment will be more valuable in the future. I chose the search tag words are "tesla", "honda", "ford". The data is using setup_twitter_oauth to download the data from twitter developer. In this Assignment, I will show the result.



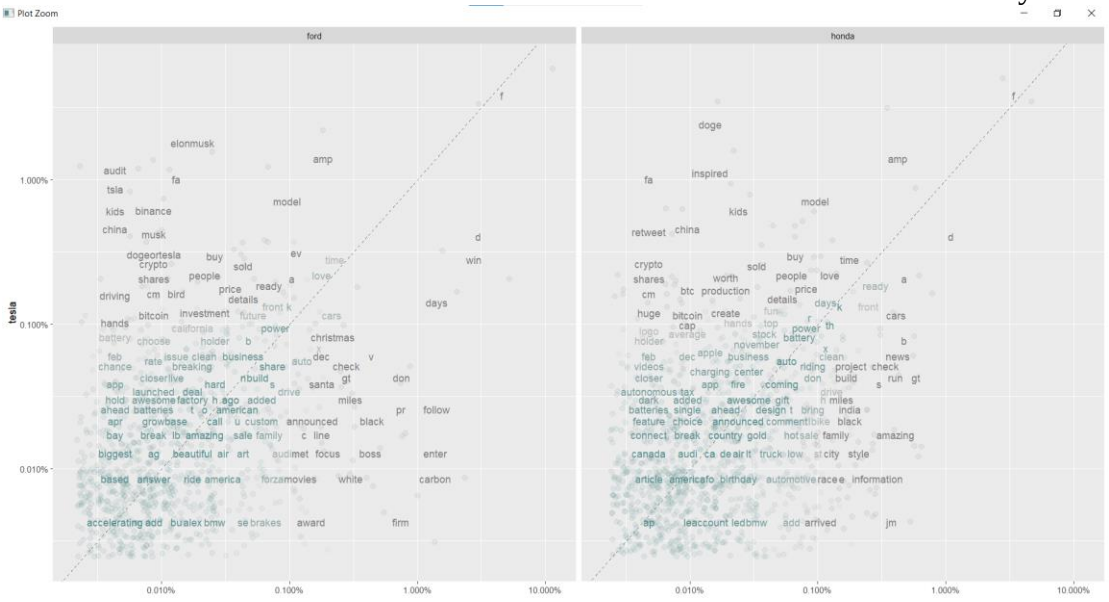
This is the top 15 words in the frequency table of Tesla. It is really interesting that the top word is Doge, which is a new bitcoin, then second is Elon Musk. I think Tesla has accepted the bitcoin payment but they canceled because the value of bitcoin is very unstable, however, one day Tesla may accept the blockchain again. I will consider to invest the dogecoin.



In Honda, we can see the top word is Civic, which is a very high C/P basic car. My first car is also Honda Civic, which was really nice. I think even electric car will be the main strain of future, the Honda Civic will still be popular.



In Ford, we can see that the top word is win, I think which means Ford still want to win the market. The second word is FP racing school, which is a driving school, I guess they really put a lot money on Twitter advertisement. Mustang and Carbon are the most important I think, because which means people is also focusing on electric car. MUSTANG MACH-E will be the first electric car of Ford next year.

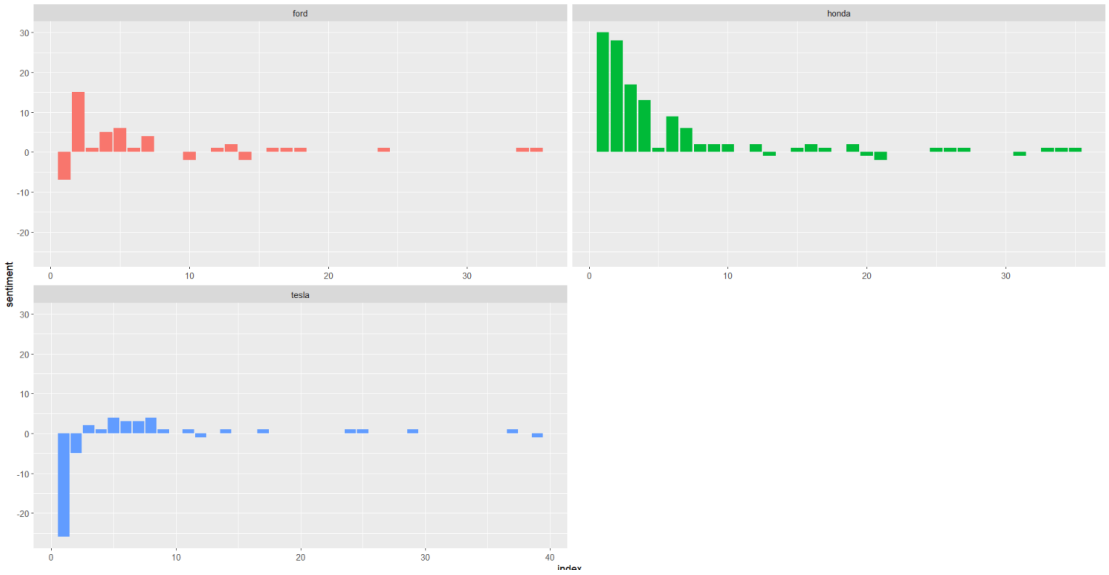


We can see in the Comparing the word frequencies of Ford and Tesla are time, love, power. In the other part, tesla and Honda table shows days, power, battery. I think the power and battery should be analysis because which means people now are more

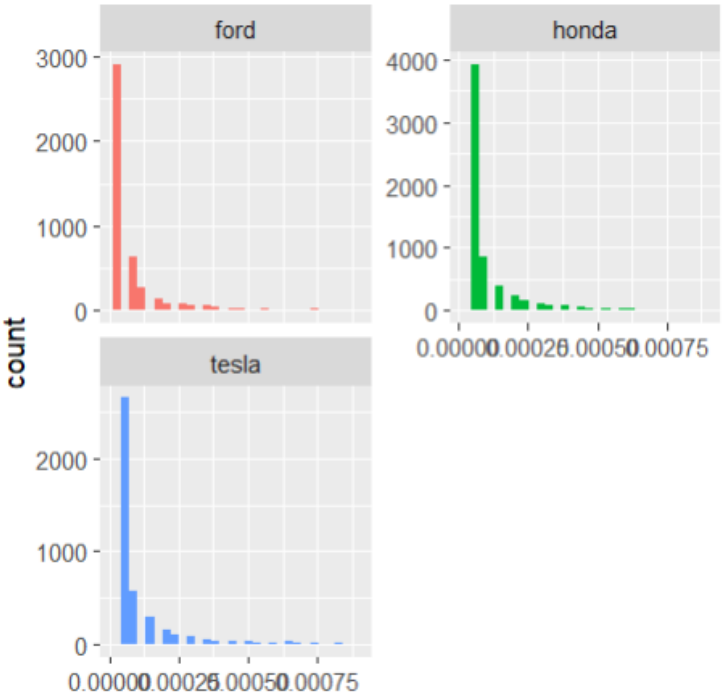
focusing on battery and car's power. Definitely, we will still see more and more people focus on electric cars in the future.

```
> cor.test(data = frequency[frequency$tag == "honda",],
+          ~ proportion + `tesla`)

        Pearson's product-moment correlation

data:  proportion and tesla
t = 37.369, df = 998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7365495 0.7883436
sample estimates:
      cor
0.7636726

>
> cor.test(data = frequency[frequency$tag == "ford",],
+          ~ proportion + `tesla`)

        Pearson's product-moment correlation

data:  proportion and tesla
t = 29.858, df = 853, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6804316 0.7461497
sample estimates:
      cor
0.7148656
```

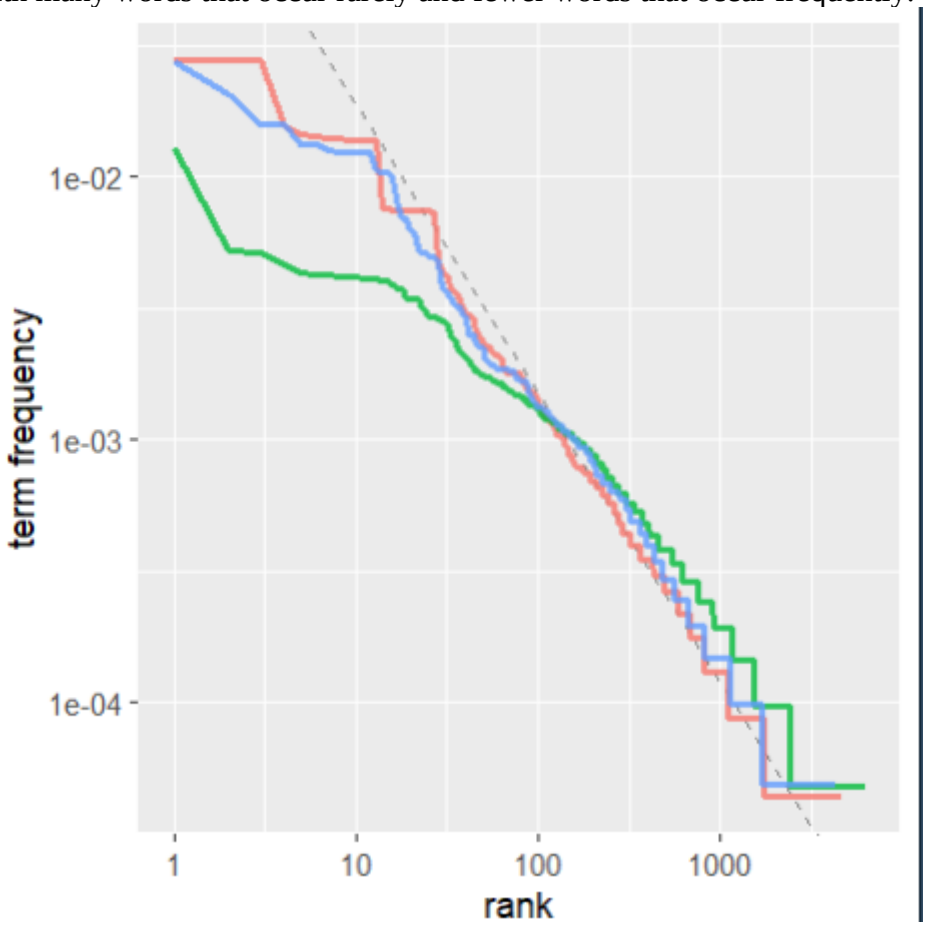
The correlation of frequency of Honda and Ford compare to Tesla.



I tried to use the sentiment analysis, I found that Ford sentiment are very close to 0, which means people opinion Ford is not negative or positive. Most people give a high positive opinion to Honda but Tesla got a very low opinion. I think the reason is because Tesla is still not really profitable and many cars accident from its auto pilot. So many people give a negative opinion in Twitter.

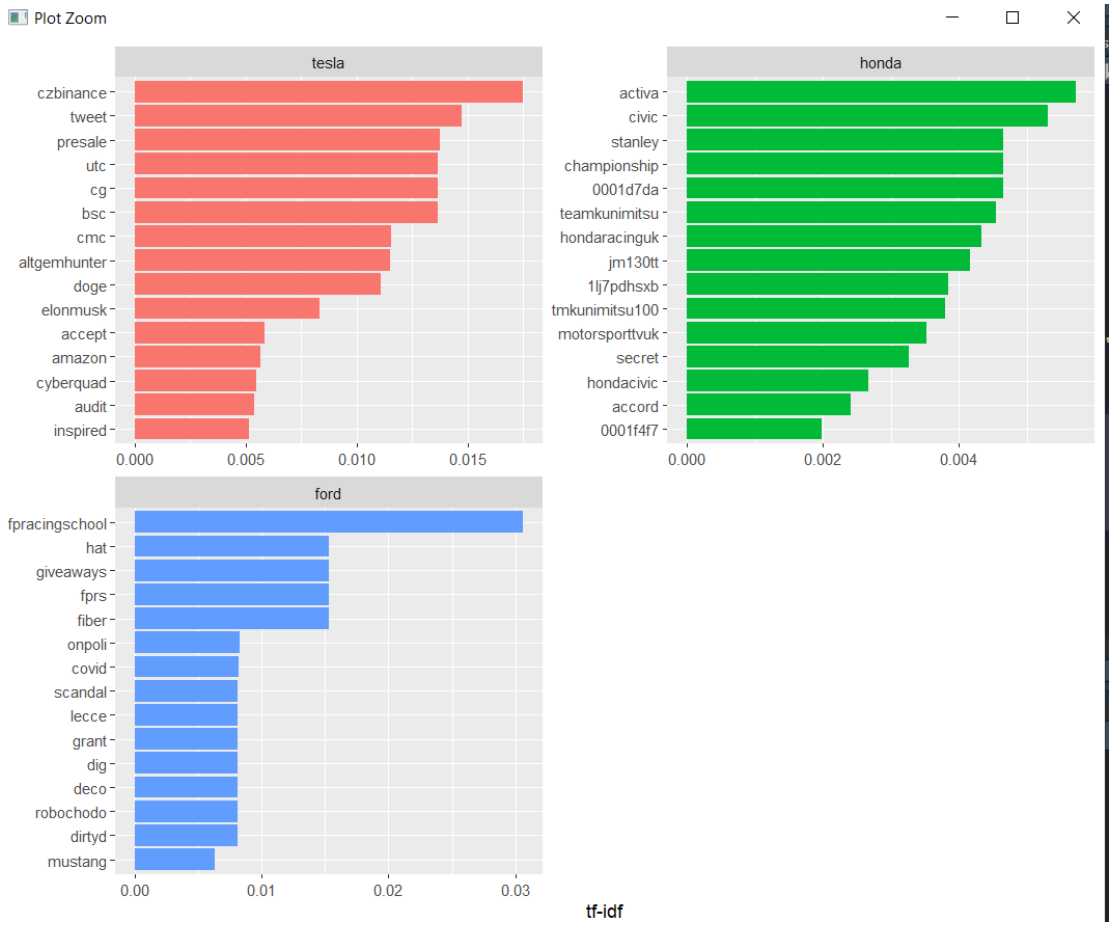


There are very long tails to the right for these words, which extremely rare that we have not shown in these plots. These plots exhibit similar distributions for all the Cars, with many words that occur rarely and fewer words that occur frequently.



Zipf's law for Cars tag and fitting an exponent

We can find that in the deviations in low rank, Tesla and Ford are more similar, but Honda are a little different, I think the reason is Honda is a Japanese company, so many words might twitter in Japanese, but the deviations we see here at high rank are not uncommon for many kinds of language.

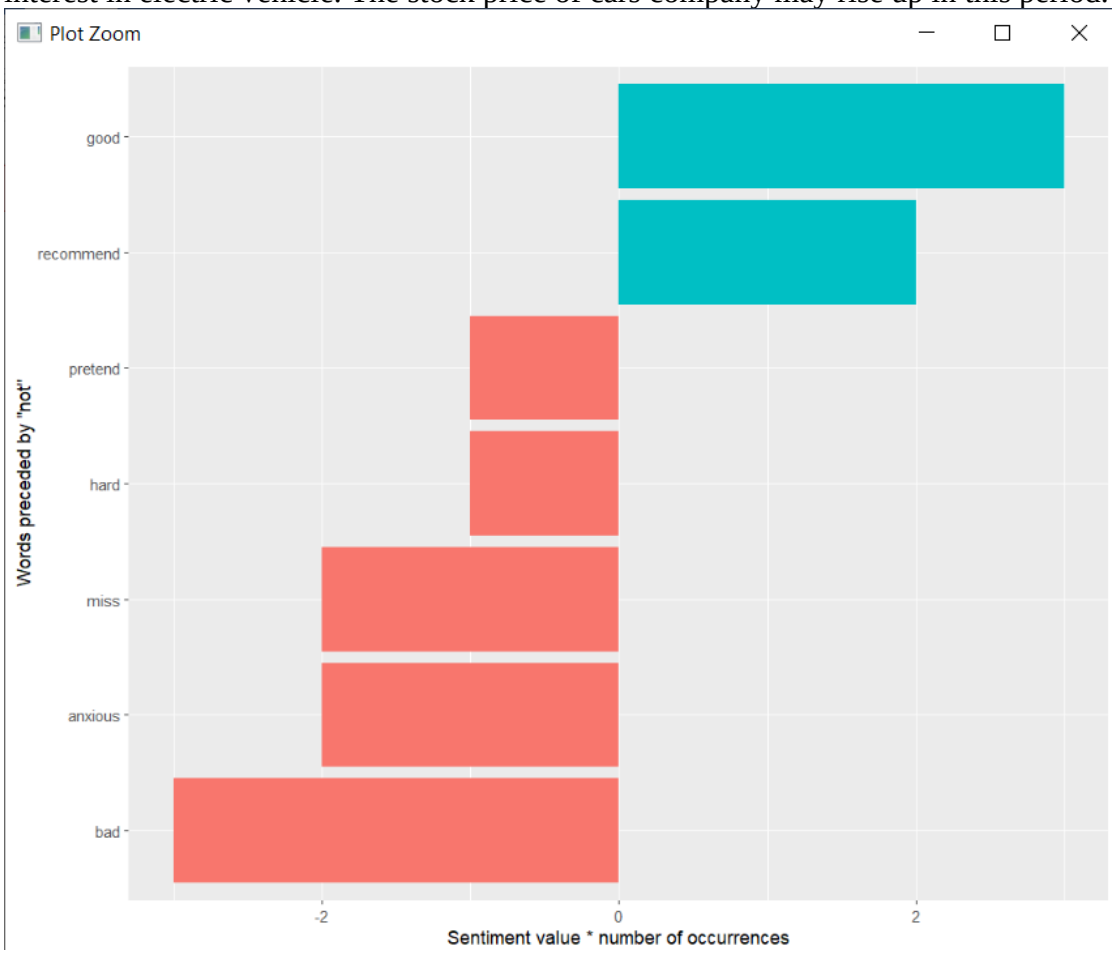


The visualization for these high tf-idf words.

Here, the plot shows that 3 hashtags of car are very different, the point of tf-idf; it identifies words that are important to Twitter data within a collection of different data. Surprising me is that Ford has scandal and covid. It might means people has more concern about the Ford and the influence of Covid. I will be more less interest about investing Ford.

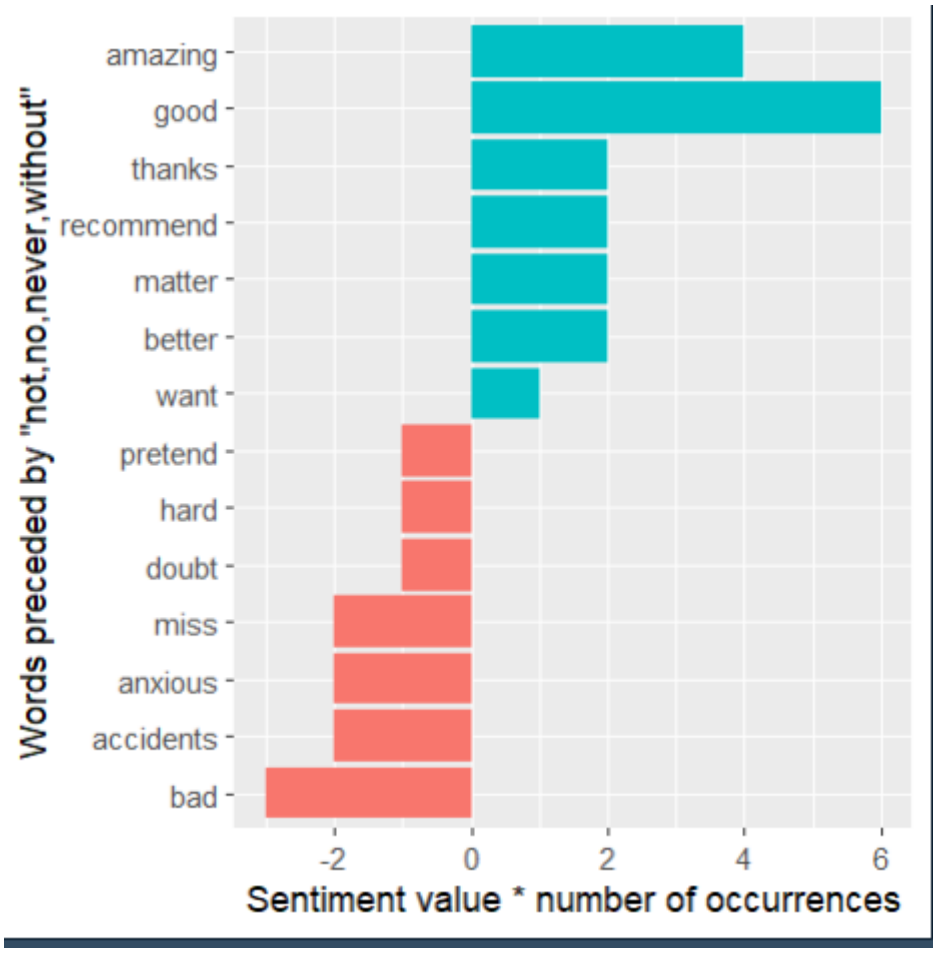
```
> bigram_counts
             word1          word2 n
1        christmas           gift 3
2           cruise        control 3
3         electric        vehicle 3
4           forza        horizon 3
5          holiday         season 3
6         0001f1fa        chassis 2
7         0001f4e6    zoommeeting 2
8         0001f914           zoom 2
9             0151            523 2
10             083            790 2
11             129            bhp 2
12            2.5t            awd 2
13            2022        hyundai 2
14            2022            kia 2
15 2022hondacivic          civic 2
```

In bigram_count, we can find that Christmas gift is very popular, and electric vehicle. I think it shows that many people prepare to buy car in Christmas, and many people interest in electric vehicle. The stock price of cars company may rise up in this period.

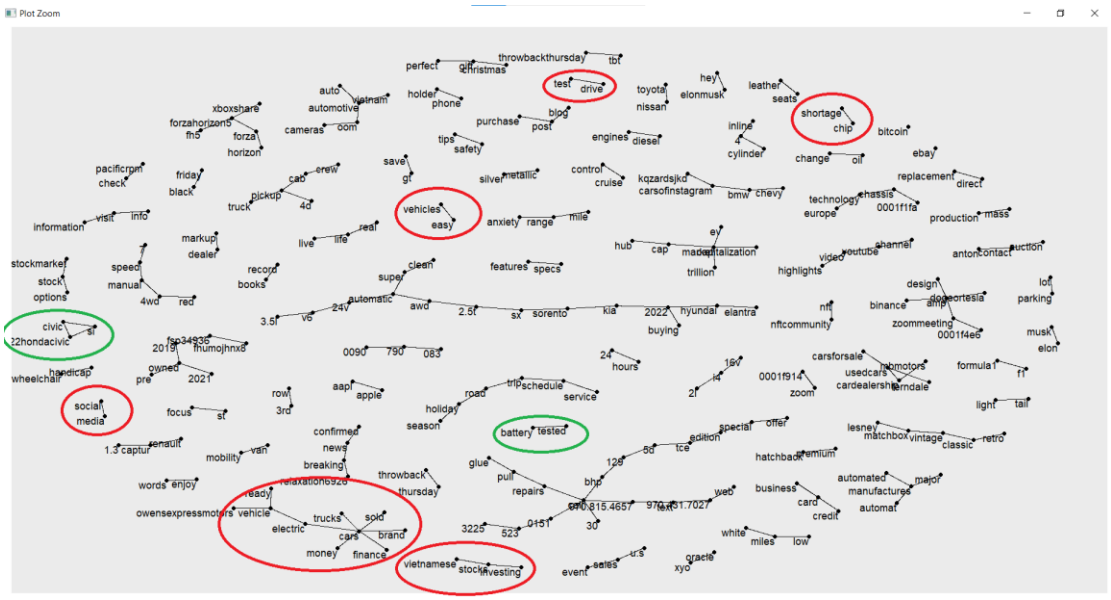


I use the words “not” to predict the sentiment value number of occurrences. The bigrams “not recommend” and “not good” were the largest causes of misidentification, making the tags seem much more positive than they are. But we can

see phrases like “not bad” and “not anxious” sometimes suggest that tags are more negative than they are. As I see on the Twitter, some people really had negative opinion to these brands, the reason I think some people just like to complain on Twitter.



If I use 4 words to predict.



When I visualize some details of the text structure, I found some important words to connect each other. Like “2022 Honda Civic”, “Batter tested”, “electric car track money finance”, “Chip shortage”, “Vietnamese stock investing”.

“2022 Honda Civic” was most popular car of Honda in this period.
“Batter tested” means the car issue also connects to the recently news, which said about that Ford and Tesla all invested more budget to new battery of electric vehicle.
“Vietnamese stock investing” means recently more and more car company invest the new factory in Vietnam due the cheap wages. Some cars companies might list on Vietnam stock market.
“Electric car track” means more and more people focus on Nikola Motor (an electric track company), I also want to buy Nikola shares.
The last one “Chip shortage” also shows that recently advanced chip is shortage because in the Covid period, more and more people use laptop to go work and some tech companies stockpile for future. I think it means the investors will be more interest invested in chip stock TSM, or Intel.

Conclusion:

The analysis shows that more and more people will focus on electric car, which is gradually instead of the traditional car, even though it still has long way to go. Some people might have negative opinions and worry about like Tesla because it have not really realize the positive net profits by only selling the cars (some profits are from allowance of government and selling the carbon credit). Besides, the data also shows the different twitters of three brands concerned the not similar thing, Ford and Tesla are closer but Honda is quietly different. I think it means the decision-making layers of Ford and Tesla will focus on the same road, but Japanese Honda will go the other ways. Finally, I think Batter and electric car will be next big step of cars and also have more increasing value. I will invest in this area and its related companies. Some third world countries like Vietnam also could seem as new star because the cheaper wage also attracts more company to invest, which will also poll their growth GDP. As an investor, I will be definitely focusing in this part.