Eluvio Report

Hao Tong

2021/4/19

Modeling

(Our data is already processed and cleaned.)

Our purpose is to tell if `years` has played an important role in the relationship between `up_votes` and `country`. So, we deployed two regression models and use anova to compare the fit between the models.

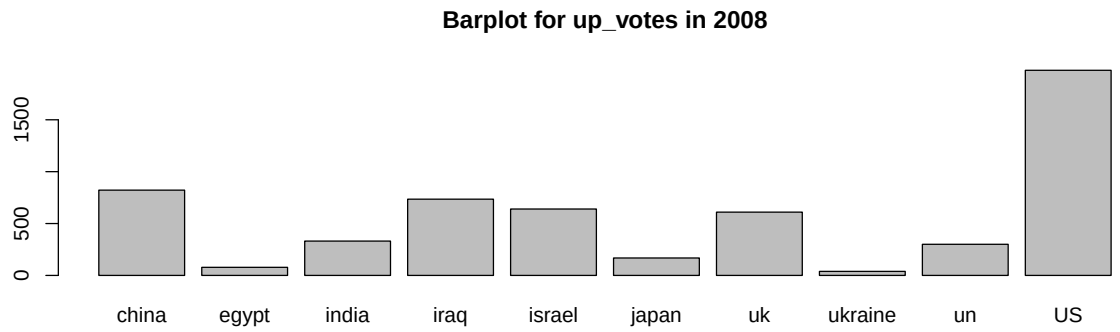
```
# model1 without `year`, model2 with `year`
model1 = lm(up_votes~label,data = df)
model2 = lm(up_votes~label+years, data = df)

anova(model1,model2)
```

```
## Analysis of Variance Table
##
## Model 1: up_votes ~ label
## Model 2: up_votes ~ label + years
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1 141705 3.5127e+10
## 2 141704 3.4927e+10  1 199514305 809.46 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

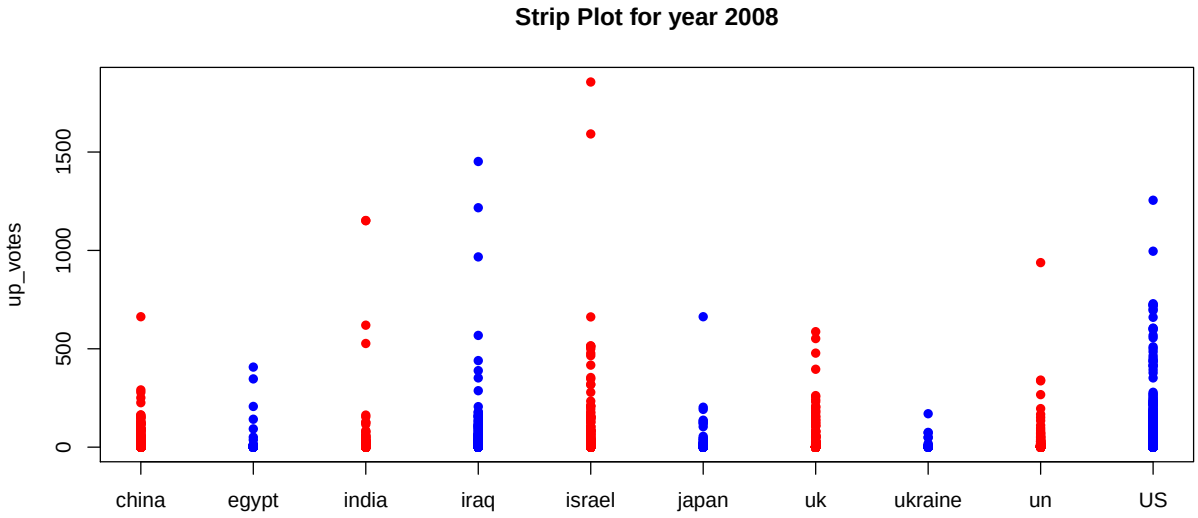
Our assumption assumes there's no effect of `years`, but we get a p-value far less than 0.001, indicating adding `years` did improve the performance,so we reject the null hypothesis. `years` does have a significant effect.

Plotting and Analysis

Now, let's draw some plots for a better look at the pattern and relationship. Since the date started from 2008-01-25, we drew the plot for `year` 2008 first.

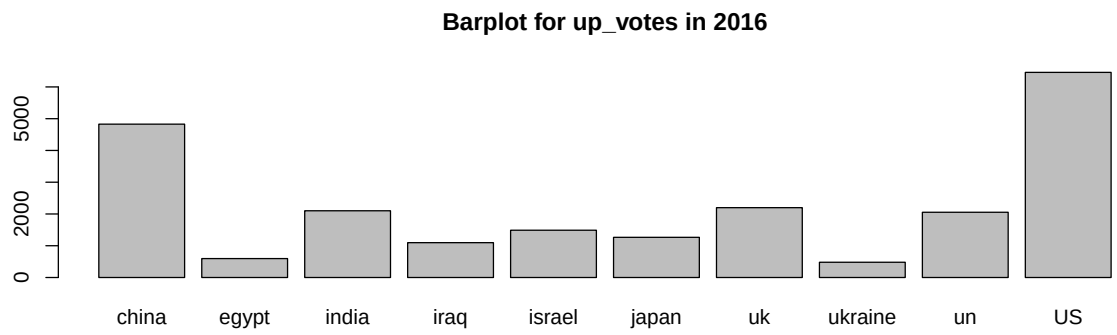


Then, we drew a strip plot for `year` 2008.

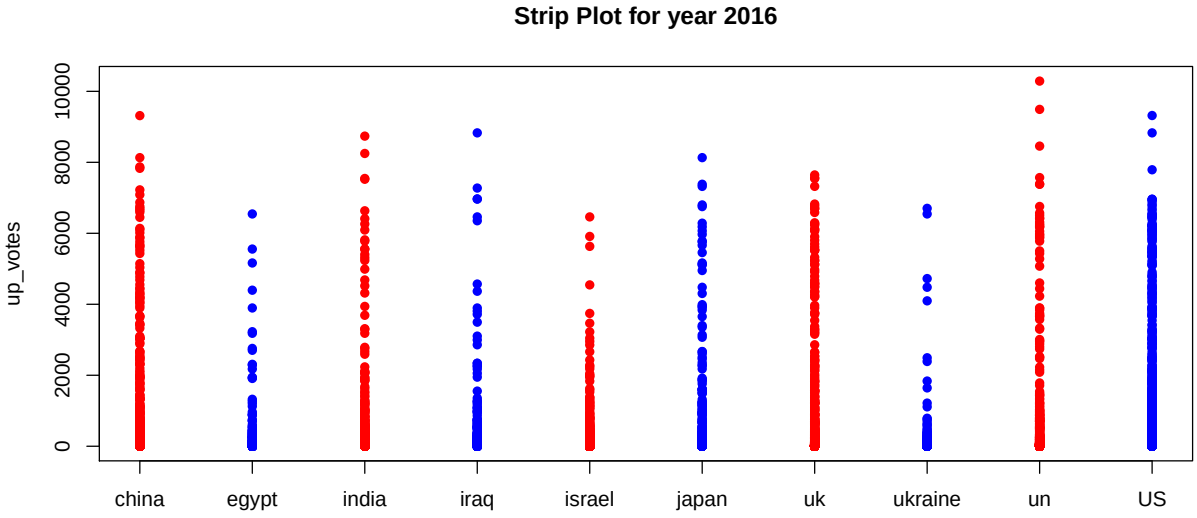


From the plots above, we can see that although Israel didn't have the most articles (only about 4th among these countries), but `up_votes` are quite high overall. The reasons are probably the good strategies and policies by Israel government of making friends and promote cooperation globally to contribute to globalization. In contrast, China has a high frequency (rank 2nd, only less than US), but hasn't got higher `up_votes` than other countries. This was probably because the Chinese government has faced a hard time in 2008 dealing with Sichuan earthquake and hosting Beijing Olympics.

Now, we take another look at `year` 2016, which is the last year in our dataset.



Then, we drew a strip plot for `year` 2016.



From the plots for `year` 2016, we can see that the frequencies are more variant now, but `up_votes` are more evenly distributed. So from the `up_votes`, we can say the world is more globalised, but from the frequencies, authors still pay more attention to certain countries.

Weaknesses

Due to the large dataset, we only extract some information from the raw data. So the analysis is not adequate. Besides, the data starts from 2008-01-25 and ends at 2016-11-25. Some data is not recorded for a year, which may lead to errors.

Others

Firstly, I have to apologive for only knowing `nltk` about NLP. There's still a lot to learn, what I did is to give it a best try.

I know my work is not perfect, yet I really hope I can get this internship position. I'm sure I can do better and learn more things to help the team and the company.

Thank you for your time to review!