Exposing Reddit

How to create a Reddit post that will get the most engagement?

Ed Eldepssi

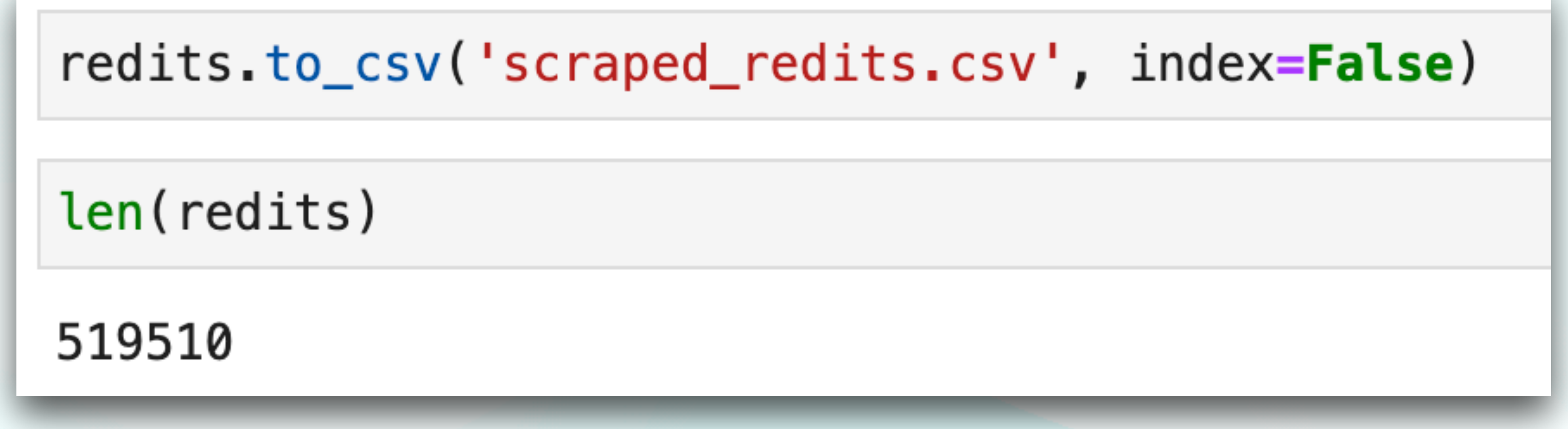
Problem Statement

What characteristics of a post on Reddit are most predictive of the overall interaction on a thread (as measured by number of comments)?

Data Collecting:

Scraping data from Reddit homepage using python and libraries as pandas, BeautifulSoup and Selenium:

Before deleting duplicates and ads

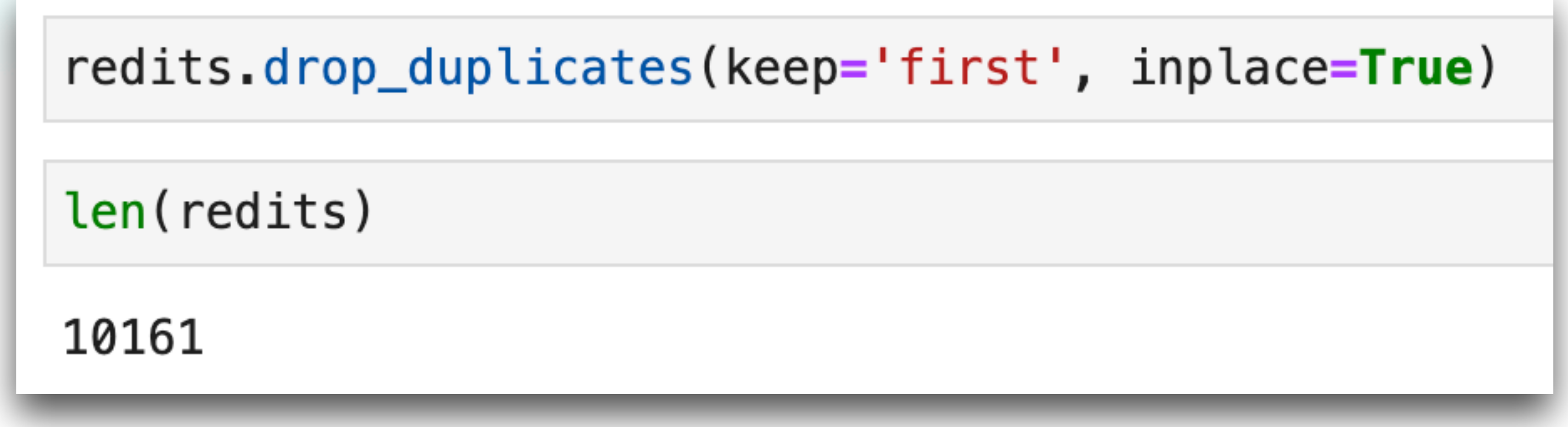


```
redits.to_csv('scraped_redits.csv', index=False)
```

```
len(redits)
```

```
519510
```

After deleting duplicates and ads

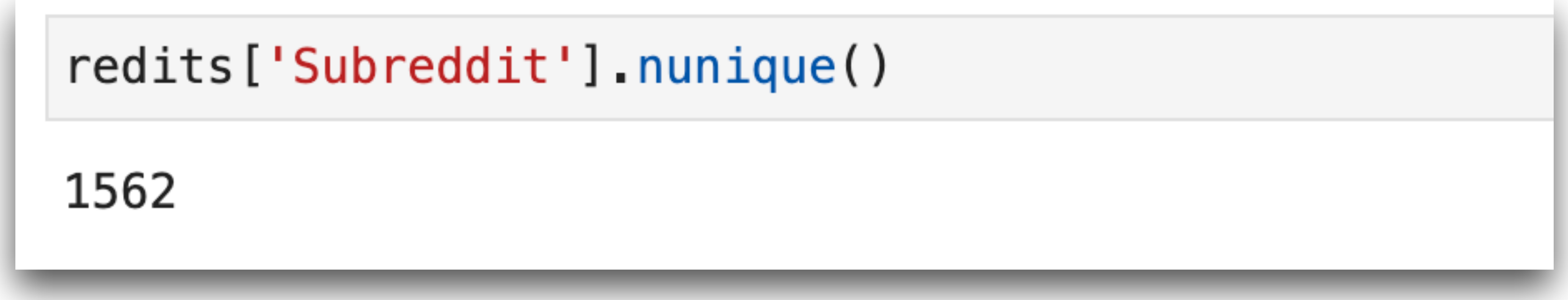


```
redits.drop_duplicates(keep='first', inplace=True)
```

```
len(redits)
```

```
10161
```

Number of Unique Subreddits

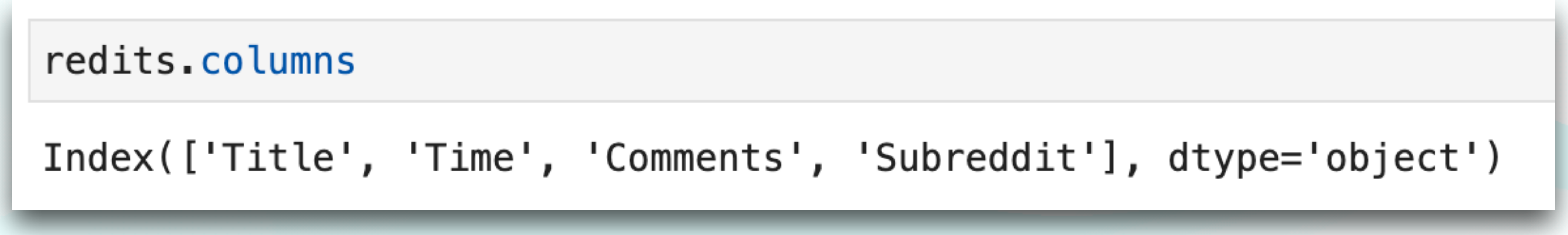


```
redits['Subreddit'].nunique()
```

```
1562
```

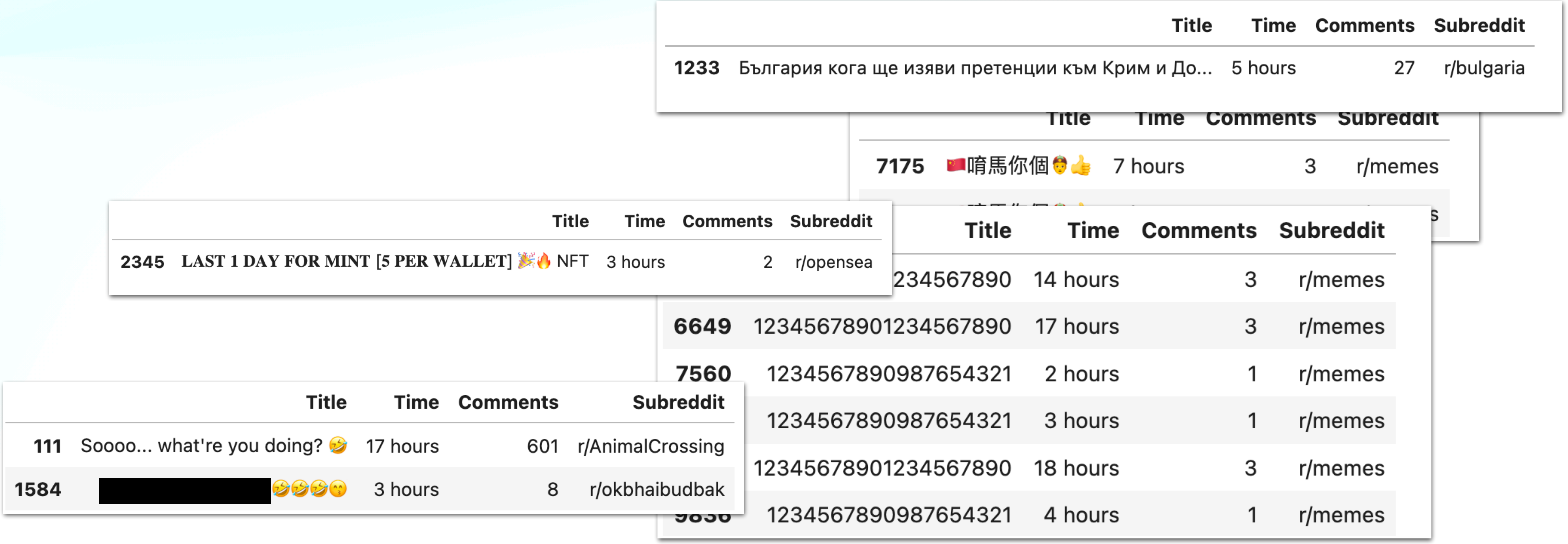
Scrapped Data:

Data Collected



```
redits.columns

Index(['Title', 'Time', 'Comments', 'Subreddit'], dtype='object')
```



	Title	Time	Comments	Subreddit
1233	България кога ще изяви претенции към Крим и До...	5 hours	27	r/bulgaria

	Title	Time	Comments	Subreddit
7175	🇨🇳唷馬你個👲👍	7 hours	3	r/memes

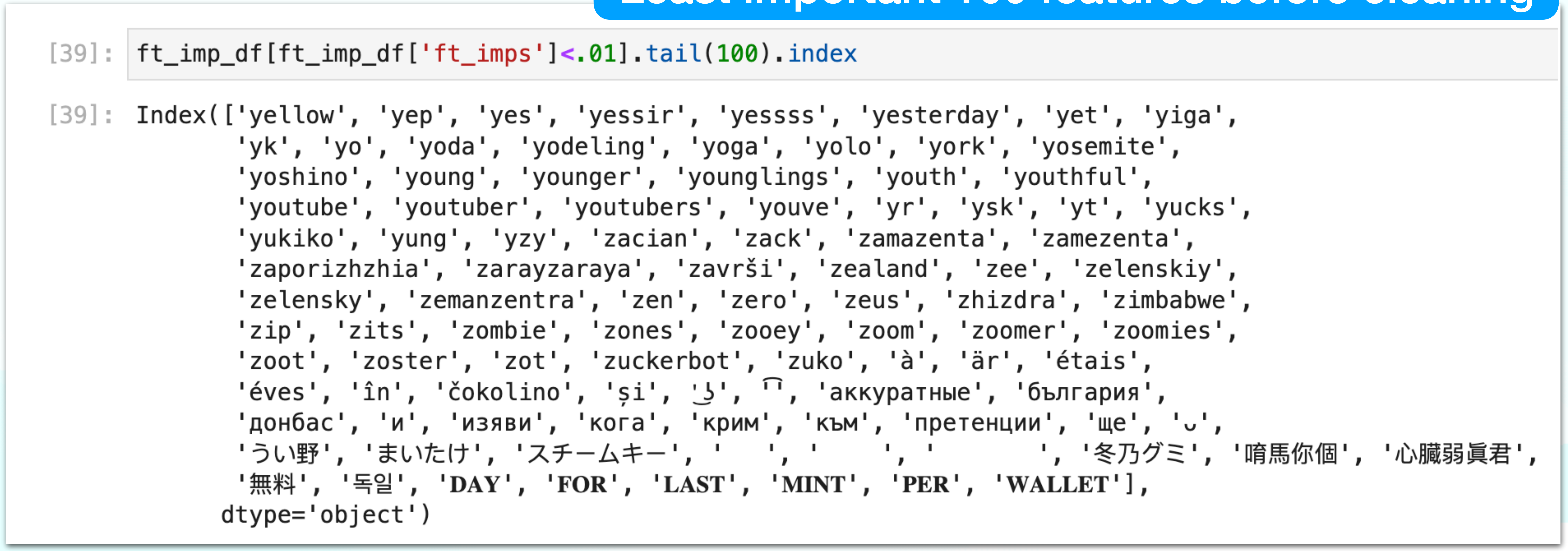
	Title	Time	Comments	Subreddit
2345	LAST 1 DAY FOR MINT [5 PER WALLET] 🎉🔥 NFT	3 hours	2	r/opensea

	Title	Time	Comments	Subreddit
	234567890	14 hours	3	r/memes
6649	12345678901234567890	17 hours	3	r/memes
7560	12345678900987654321	2 hours	1	r/memes
	12345678900987654321	3 hours	1	r/memes
	12345678901234567890	18 hours	3	r/memes
9836	12345678900987654321	4 hours	1	r/memes

	Title	Time	Comments	Subreddit
111	Soooo... what're you doing? 🤣	17 hours	601	r/AnimalCrossing
1584	[redacted]🤣🤣🤣😚	3 hours	8	r/okbhaibudbak

Data Cleaning:

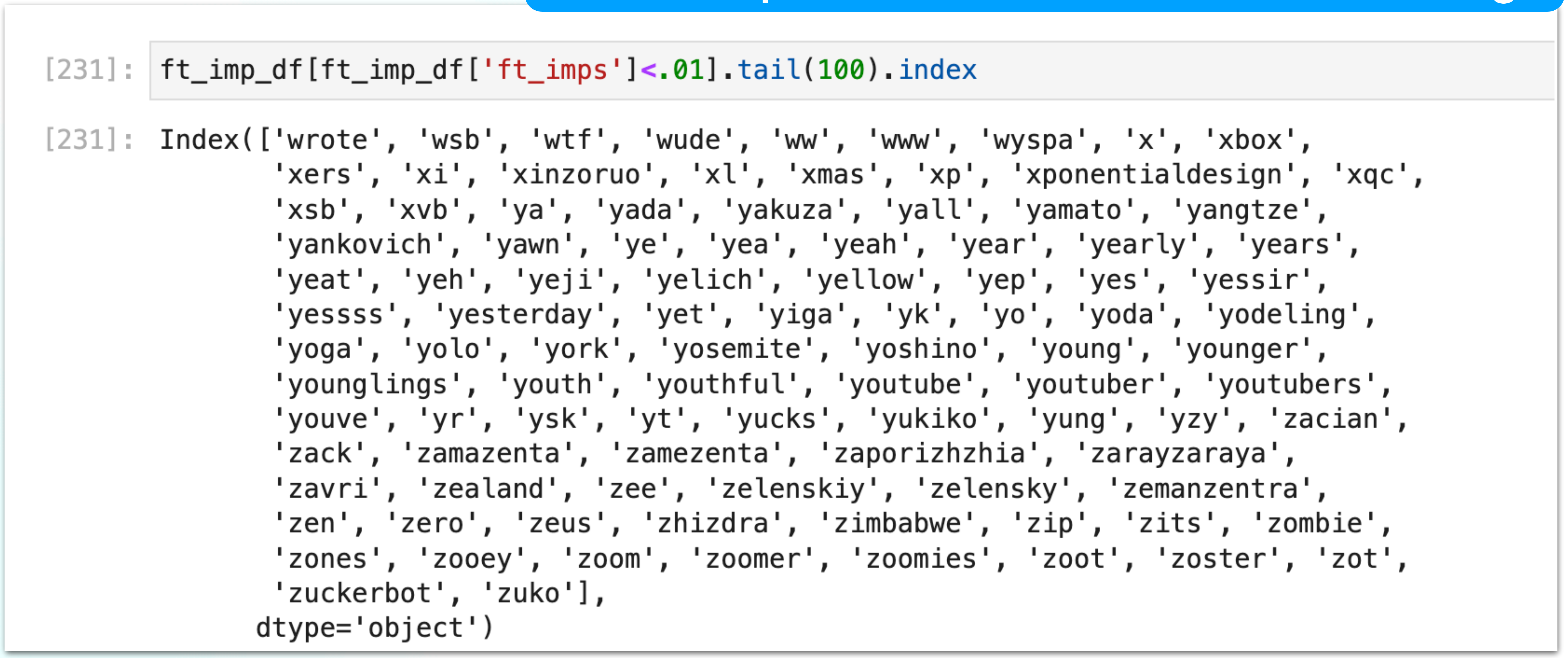
Least important 100 features before cleaning



```
[39]: ft_imp_df[ft_imp_df['ft_imps']<.01].tail(100).index

[39]: Index(['yellow', 'yep', 'yes', 'yessir', 'yessss', 'yesterday', 'yet', 'yiga',
             'yk', 'yo', 'yoda', 'yodeling', 'yoga', 'yolo', 'york', 'yosemite',
             'yoshino', 'young', 'younger', 'younglings', 'youth', 'youthful',
             'youtube', 'youtuber', 'youtubers', 'youve', 'yr', 'ysk', 'yt', 'yucks',
             'yukiko', 'yung', 'yzy', 'zacian', 'zack', 'zamazenta', 'zamezenta',
             'zaporizhzhia', 'zarayzaraya', 'završi', 'zealand', 'zee', 'zelenskiy',
             'zelensky', 'zemanzentra', 'zen', 'zero', 'zeus', 'zhizdra', 'zimbabwe',
             'zip', 'zits', 'zombie', 'zones', 'zooey', 'zoom', 'zoomer', 'zoomies',
             'zoot', 'zoster', 'zot', 'zuckerbot', 'zuko', 'à', 'är', 'étais',
             'éves', 'în', 'čokolino', 'ṣi', 'ڕ', '͡', 'аккуратные', 'българия',
             'донбас', 'и', 'изяви', 'кога', 'крим', 'към', 'претенции', 'ще', 'ᴗ',
             'うい野', 'まいたけ', 'スチームキー', '   ', '    ', '       ', '冬乃グミ', '唷馬你個', '心臟弱眞君',
             '無料', '독일', 'DAY', 'FOR', 'LAST', 'MINT', 'PER', 'WALLET'],
            dtype='object')
```

Least important 100 features after cleaning



```
[231]: ft_imp_df[ft_imp_df['ft_imps']<.01].tail(100).index

[231]: Index(['wrote', 'wsb', 'wtf', 'wude', 'ww', 'www', 'wyspa', 'x', 'xbox',
              'xers', 'xi', 'xinzoruo', 'xl', 'xmas', 'xp', 'xponentialdesign', 'xqc',
              'xsb', 'xvb', 'ya', 'yada', 'yakuza', 'yall', 'yamato', 'yangtze',
              'yankovich', 'yawn', 'ye', 'yea', 'yeah', 'year', 'yearly', 'years',
              'yeat', 'yeh', 'yeji', 'yelich', 'yellow', 'yep', 'yes', 'yessir',
              'yessss', 'yesterday', 'yet', 'yiga', 'yk', 'yo', 'yoda', 'yodeling',
              'yoga', 'yolo', 'york', 'yosemite', 'yoshino', 'young', 'younger',
              'younglings', 'youth', 'youthful', 'youtube', 'youtuber', 'youtubers',
              'youve', 'yr', 'ysk', 'yt', 'yucks', 'yukiko', 'yung', 'yzy', 'zacian',
              'zack', 'zamazenta', 'zamezenta', 'zaporizhzhia', 'zarayzaraya',
              'zavri', 'zealand', 'zee', 'zelenskiy', 'zelensky', 'zemanzentra',
              'zen', 'zero', 'zeus', 'zhizdra', 'zimbabwe', 'zip', 'zits', 'zombie',
              'zones', 'zooey', 'zoom', 'zoomer', 'zoomies', 'zoot', 'zoster', 'zot',
              'zuckerbot', 'zuko'],
             dtype='object')
```

Data Cleaning:

- Other languages
- Emojis
- Numbers
- Profanity
- Punctuations
- Special characters

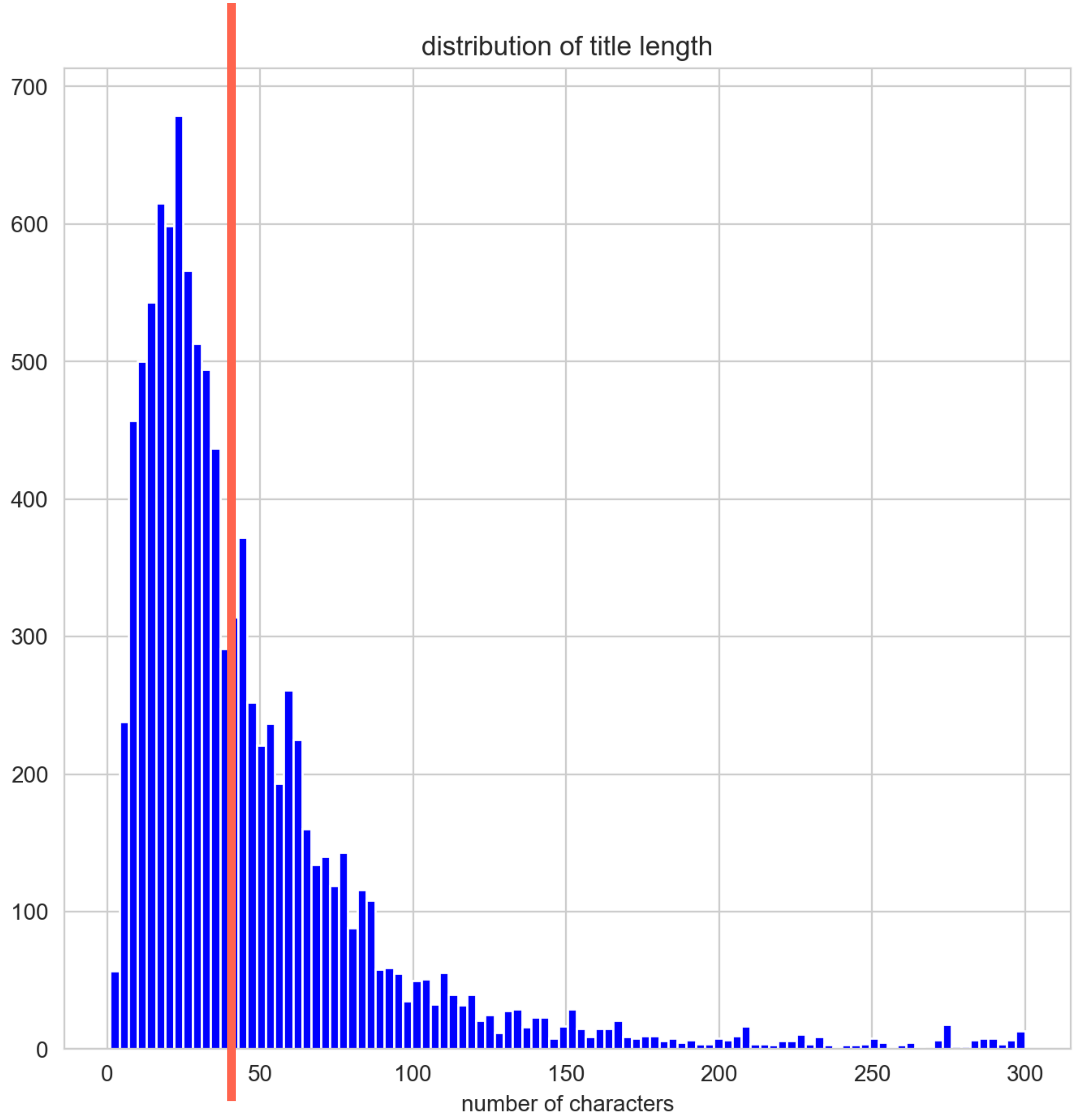
All stripped

EDA on row data:

Distribution of length of title:

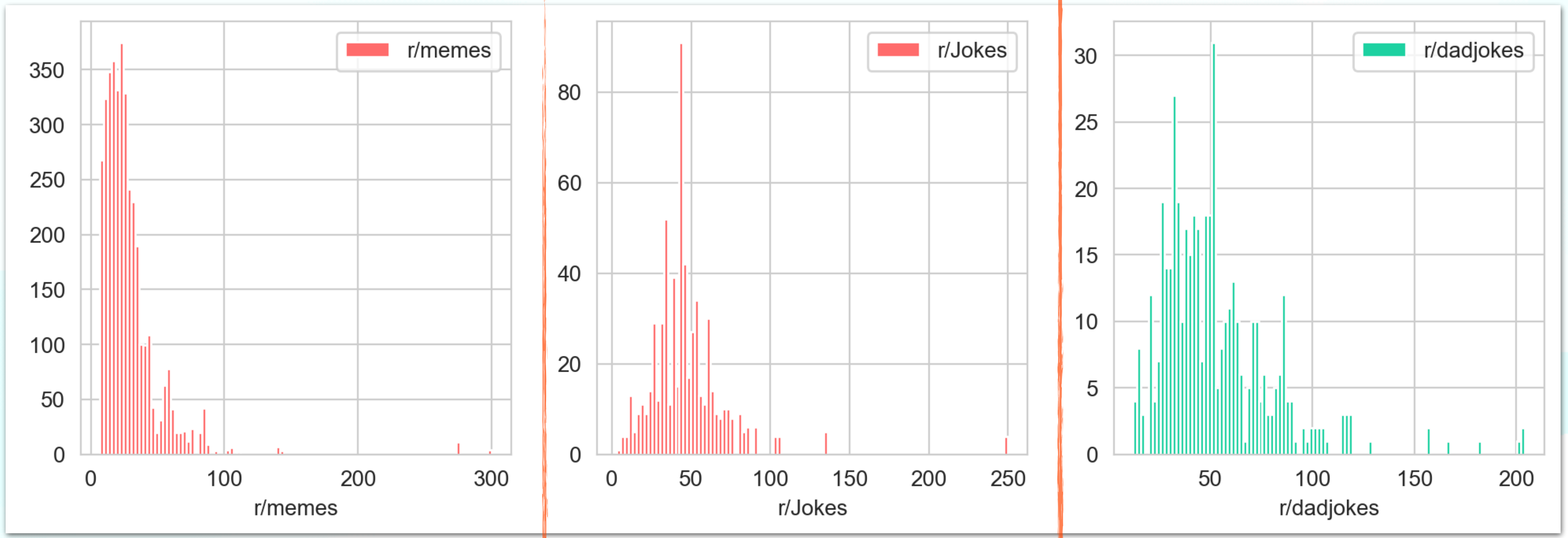
Mean
46 Characters

Median
32 Characters

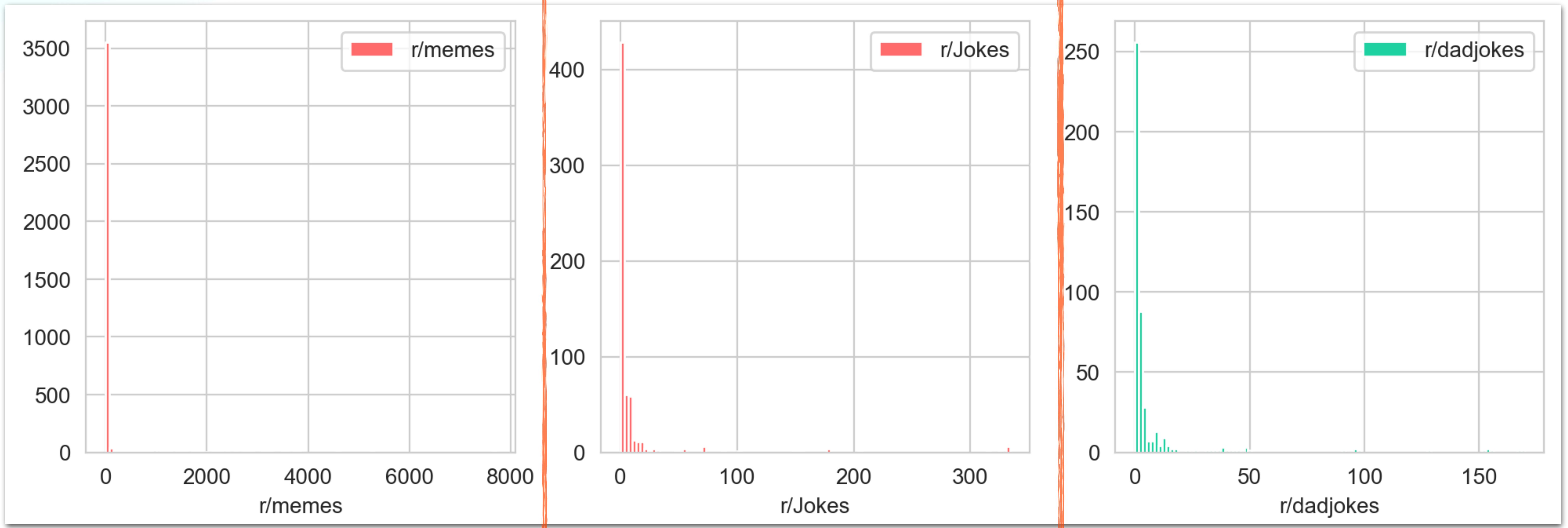


EDA on row data:

Distribution of length of title



Distribution of number of comments



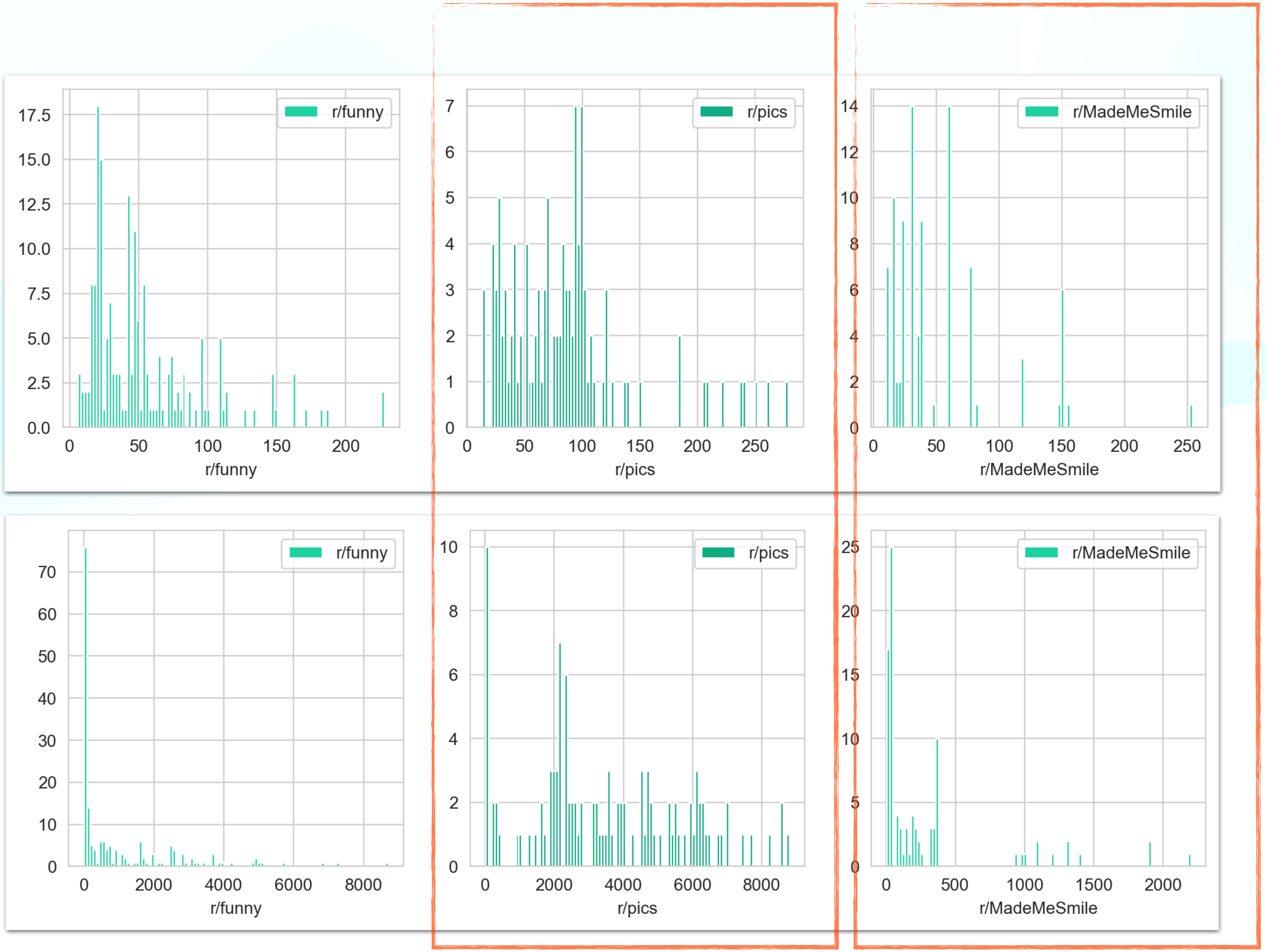
Top Subreddit Engagement:

- r/jokes
- r/dadjokes

EDA on row data:

Distribution of length of title

Distribution of number of comments



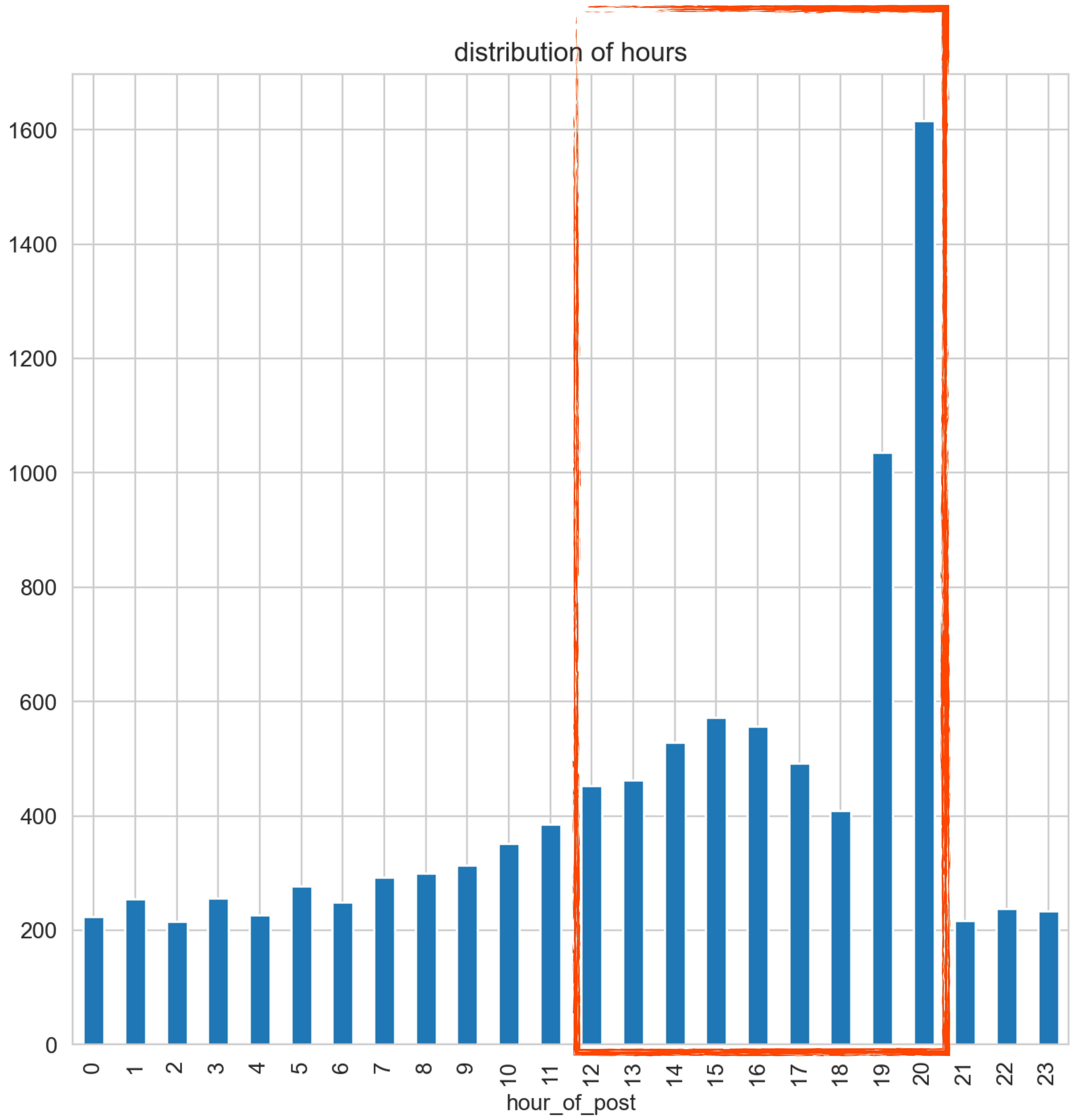
Top Subreddit Engagement:

- r/pics
- R/MadeMeSmile
- r/jokes
- r/dadjokes

EDA on row data:

Distribution of post hours:

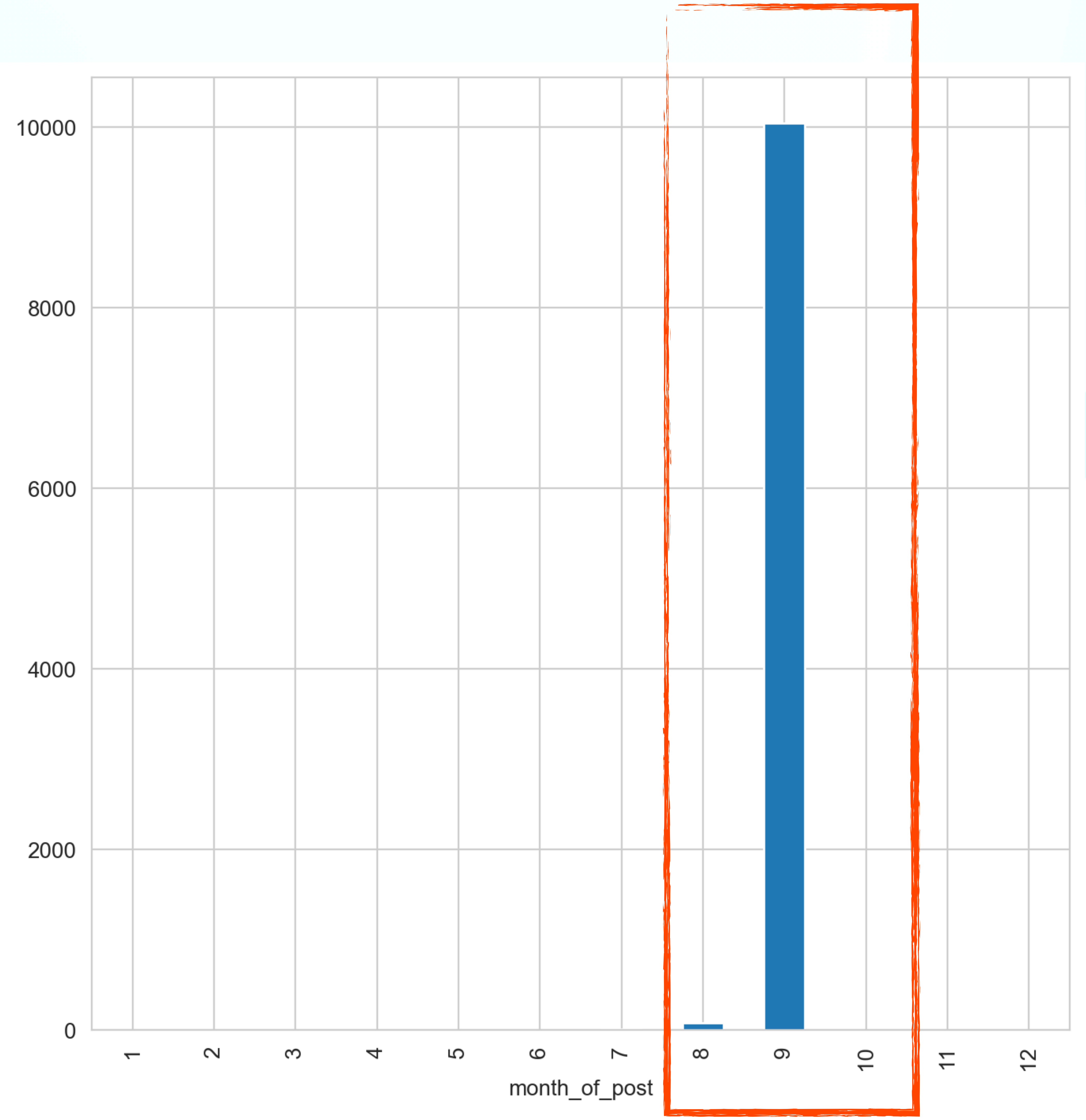
More posts
between:
12 pm to 8 pm



EDA on row data:

Distribution of months of post:

My data collection was on August 8th and August 9th.

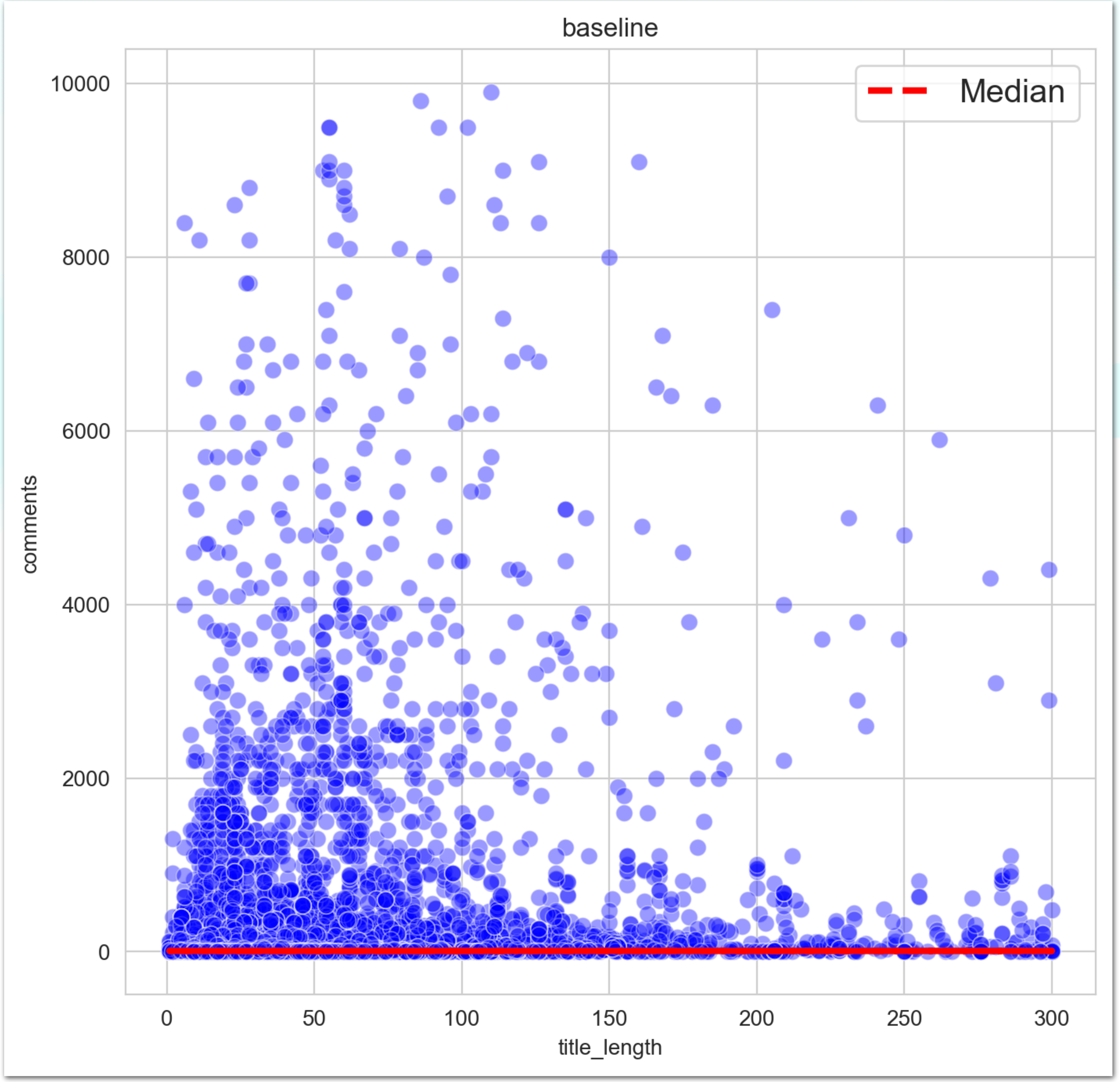


EDA on row data:

Exploring Target Label:

Median Comments:

9



EDA on row data:

Exploring Target Label:

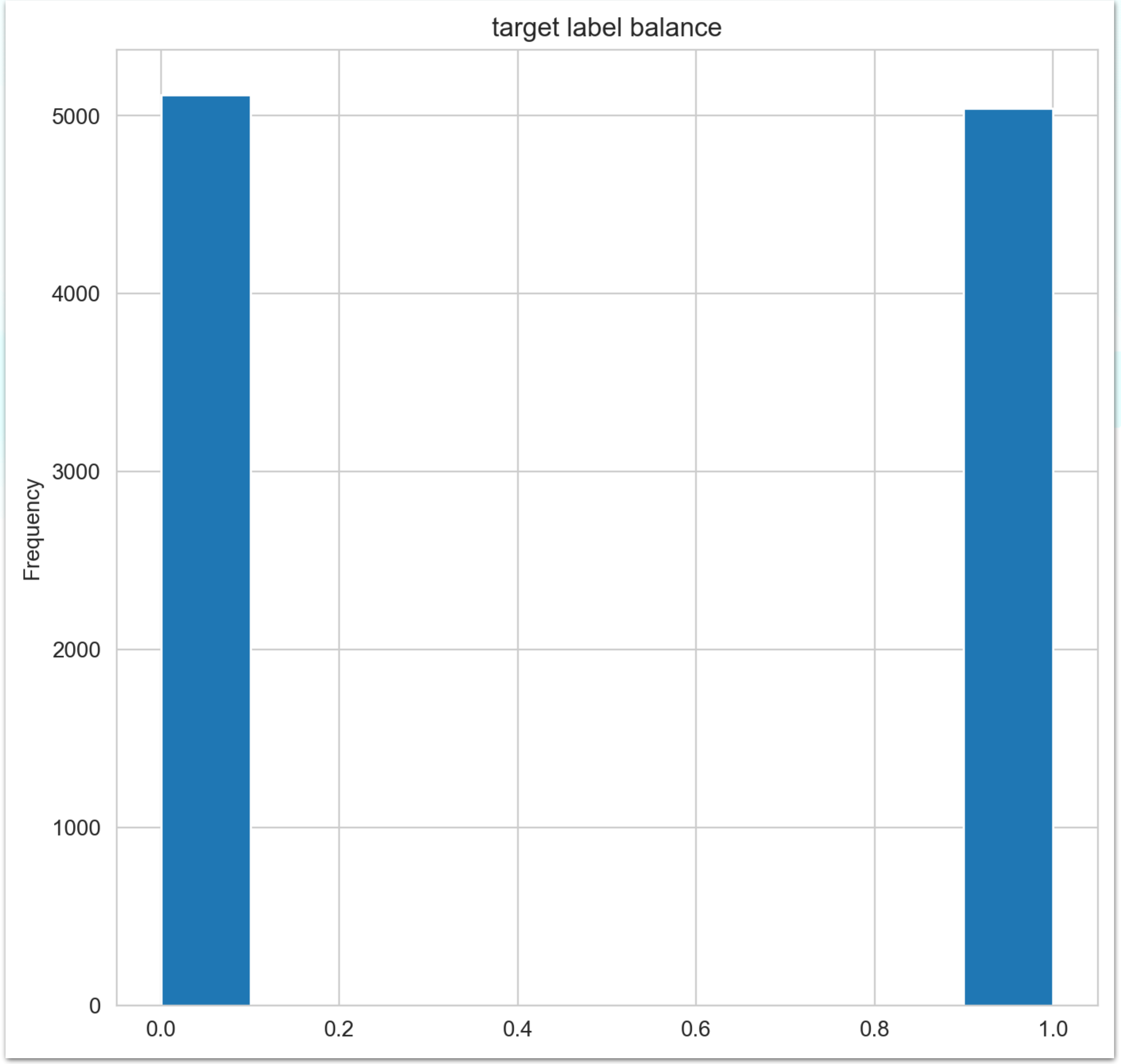
Median Comments:

9

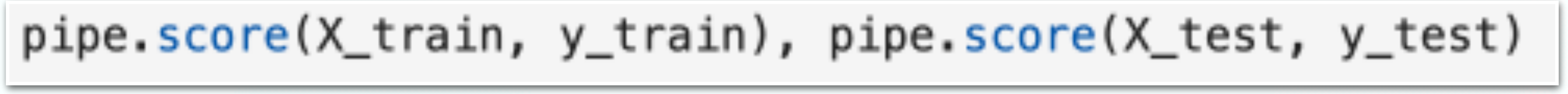
Target Label Balance:

50.3% Low engagement

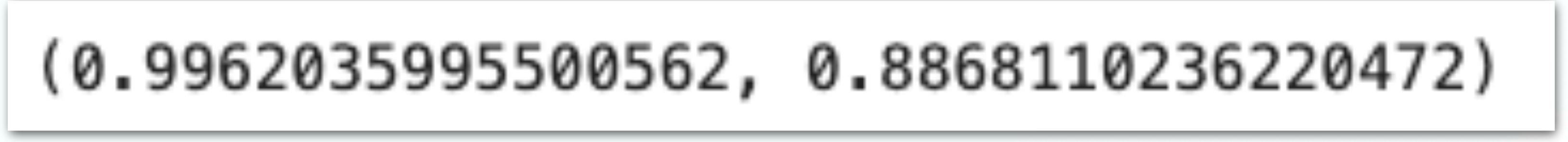
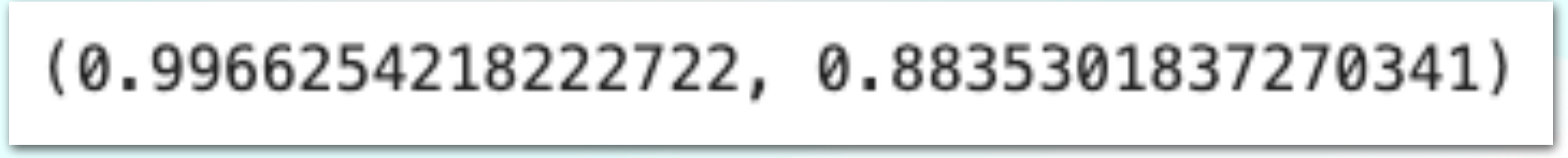
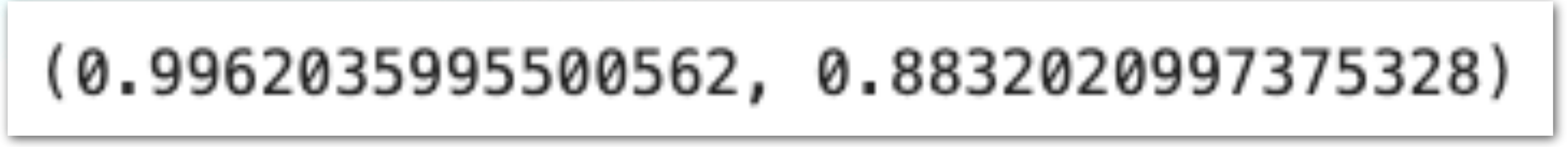
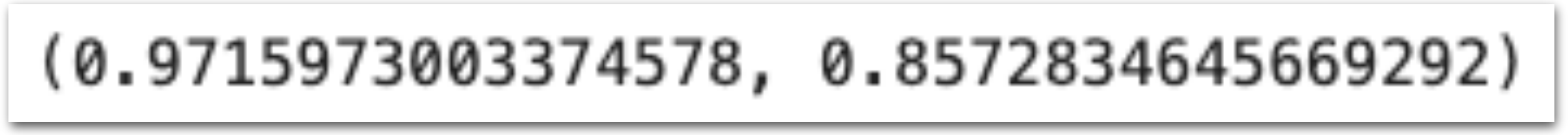
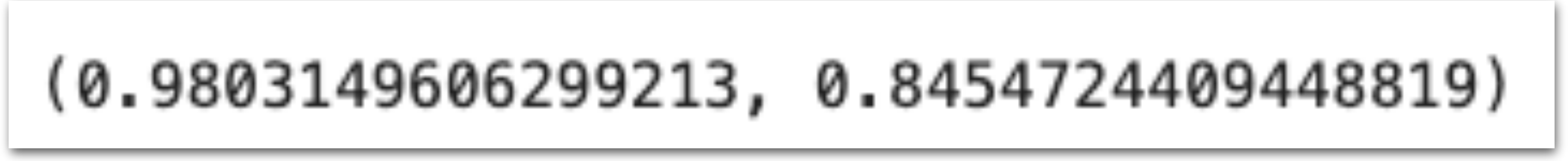
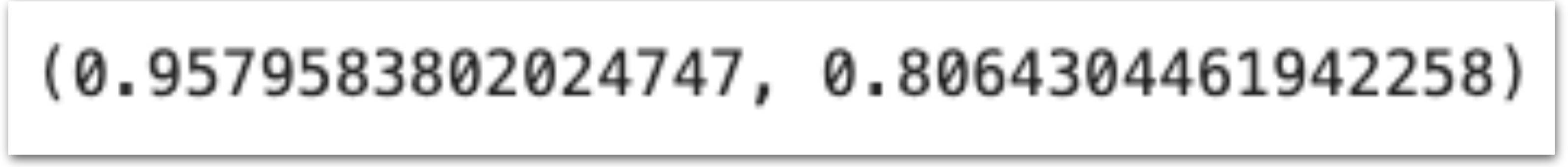
49.6% High engagement



Modeling



```
pipe.score(X_train, y_train), pipe.score(X_test, y_test)
```

Features Used		
length_of_title hours_of_post Text: Title	RandomForestClassifier	 (0.9962035995500562, 0.8868110236220472)
length_of_title hours_of_post Text: Title	ExtraTreeClassifier	 (0.9966254218222722, 0.8835301837270341)
length_of_title hours_of_post Text: Title	DecisionTreeClassifier	 (0.9962035995500562, 0.8832020997375328)
length_of_title hours_of_post Text: Title	AdaBoostClassifier	 (0.9715973003374578, 0.8572834645669292)
Text: Title	RandomForestClassifier	 (0.9803149606299213, 0.8454724409448819)
Text: Title	LogisticRegression	 (0.9579583802024747, 0.8064304461942258)

Findings - Most Important Features

Best Subreddits:

-r/memes

-r/dadjokes

-r/jokes

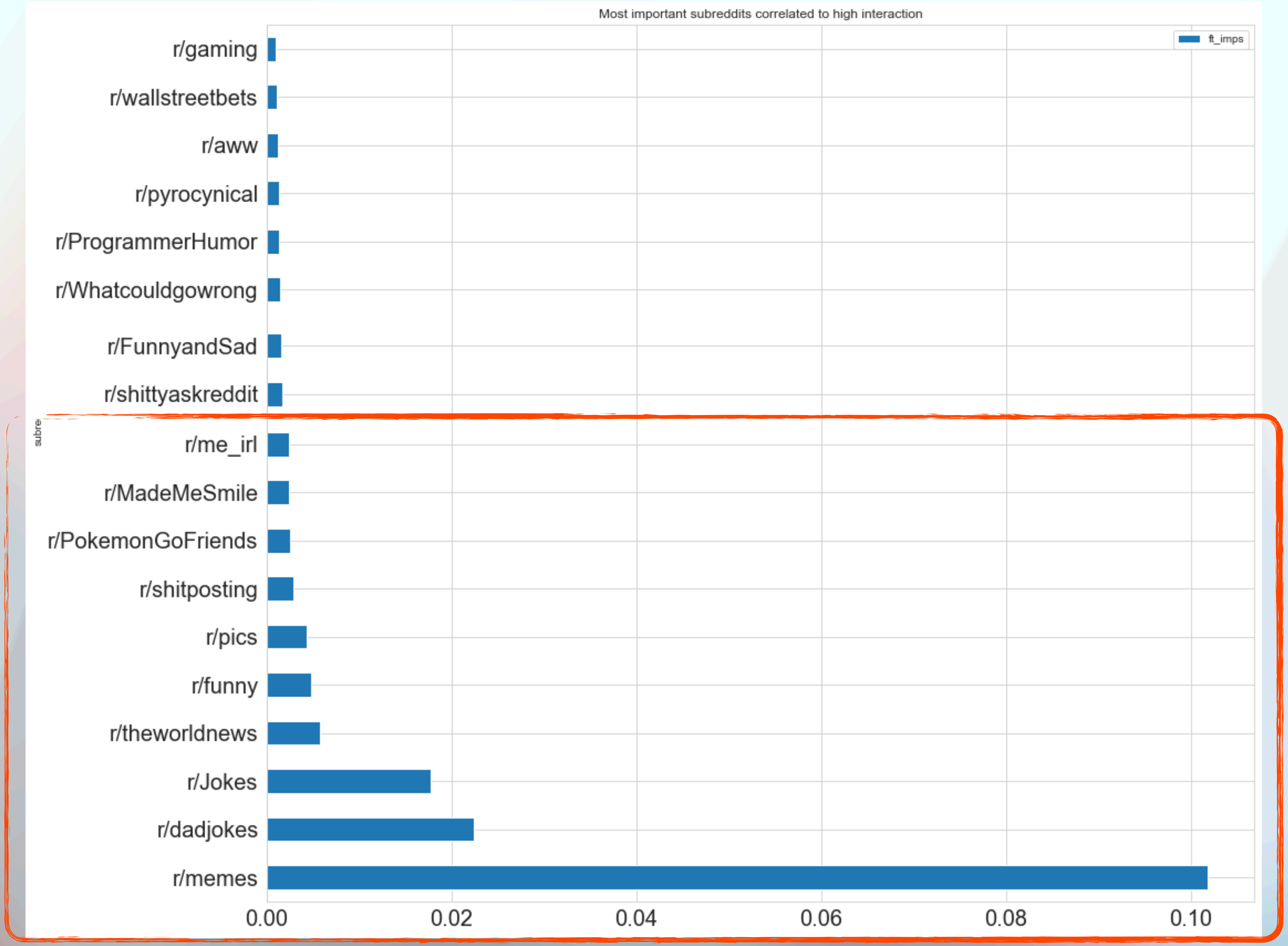
-r/theworldnews

-r/funny

-r/pics

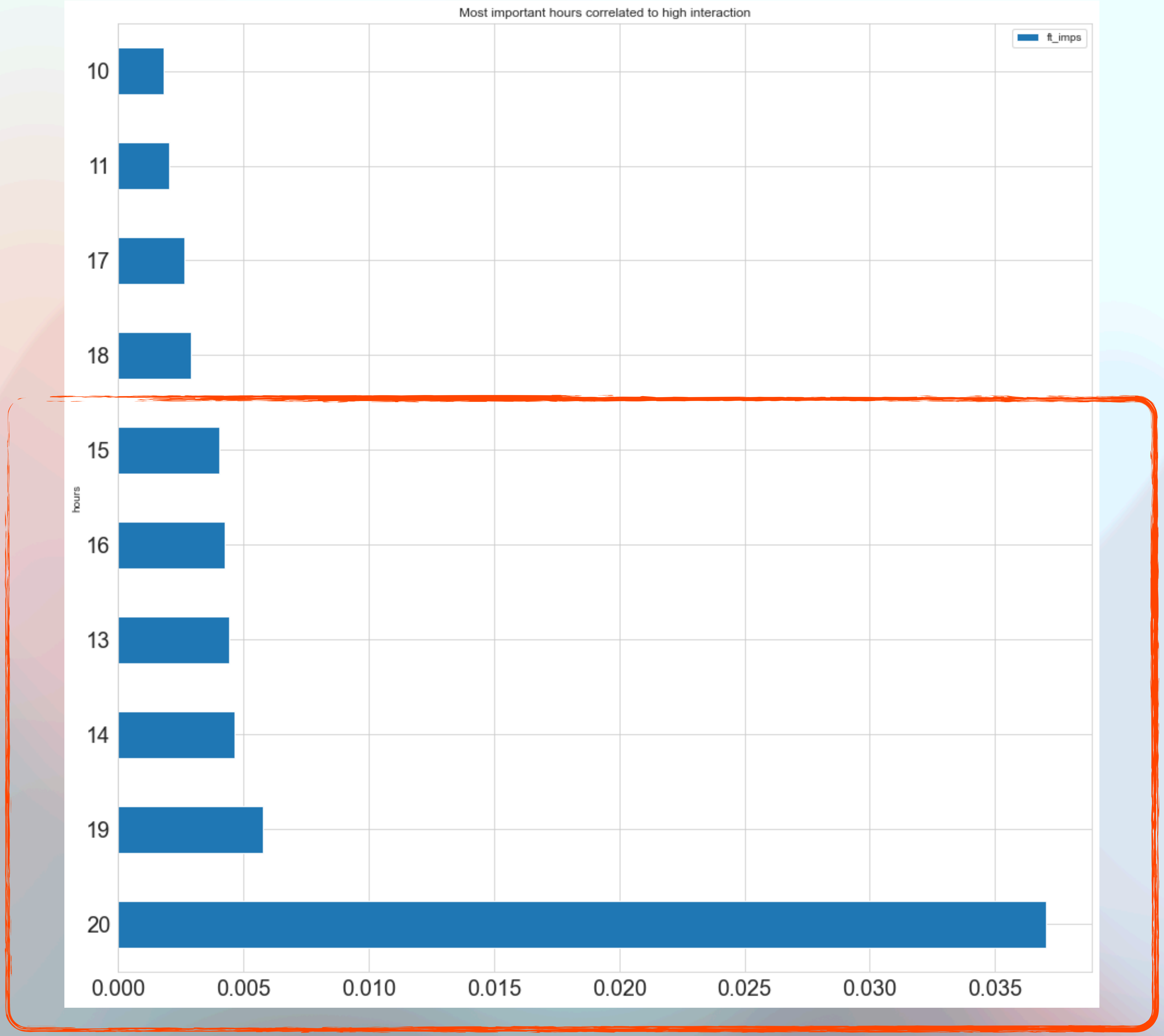
-r/PokemonGoFriends

-r/me_irl



Findings - Most Important Features

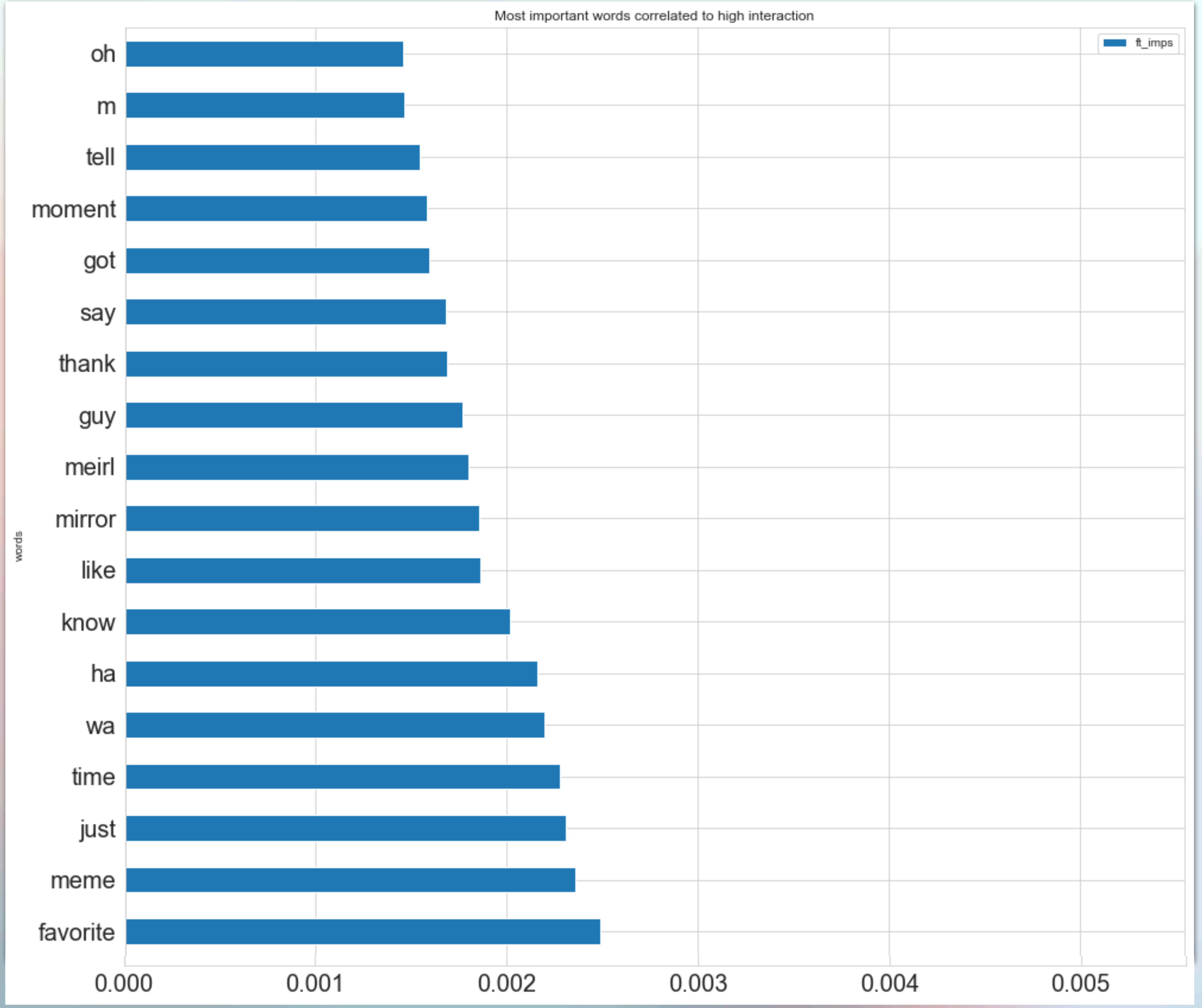
Best time to post:
Between 3pm to
8pm



Findings - Most Important words with ExtraTreeClassifier

Best Words:

- Favorite
- Meme
- Just
- Time
- Wa
- Ha
- Know
- Like
- Mirror
- Meirl: me in real life
- Guy
- Thank
- Say
- Got
- Moment
- Tell
- M
- Oh



Findings - top 20 words with highest coefs:

Best Words:

✓ Meme
- Game

✓ Oc:

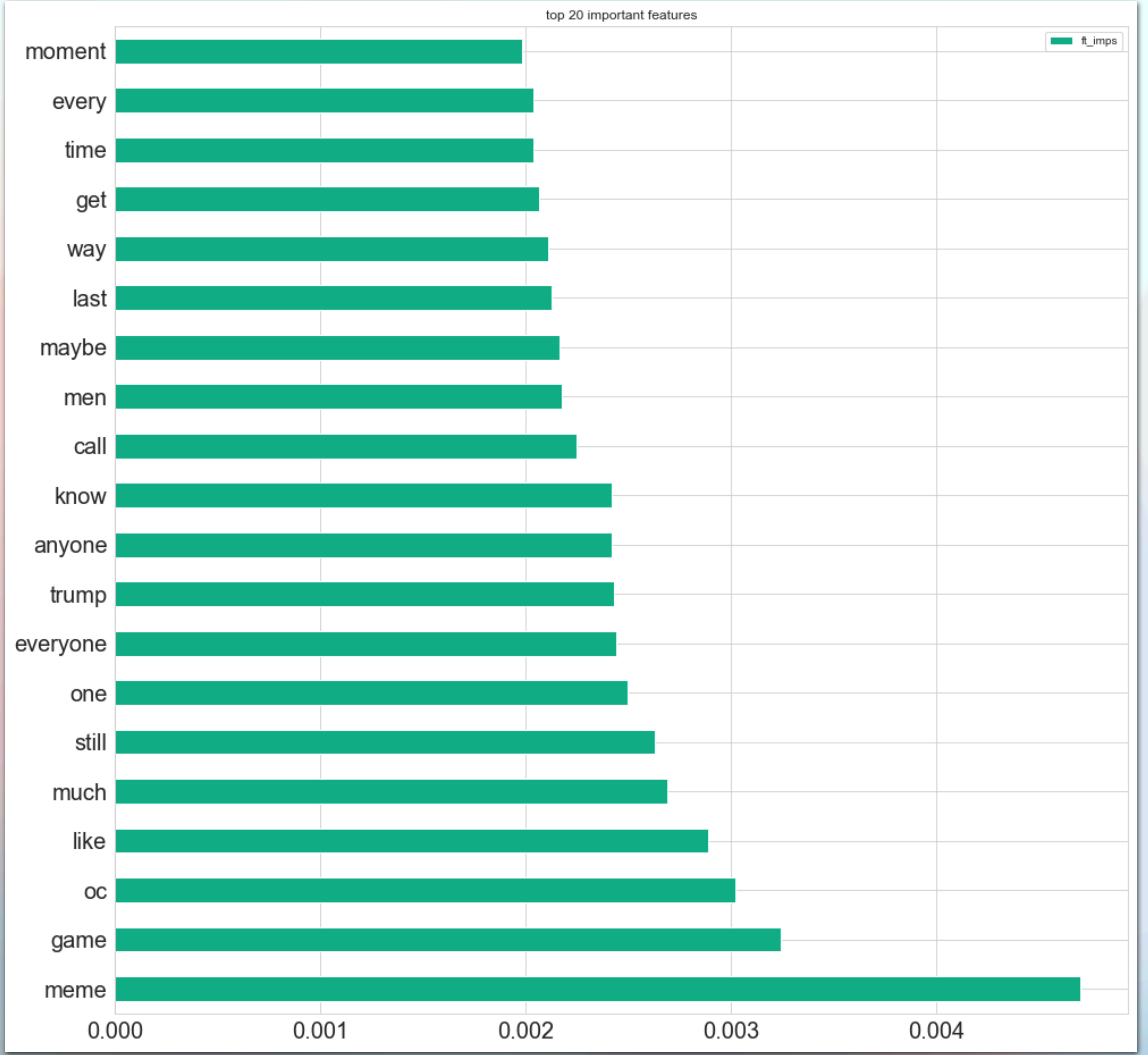
✓ Like
- Much
- Still
- One
- Everyone

✓ Trump
- Anyone

✓ Know
- Call
- Men
- Maybe
- Last
- Way
- Get

✓ Time
- Every

✓ Moment



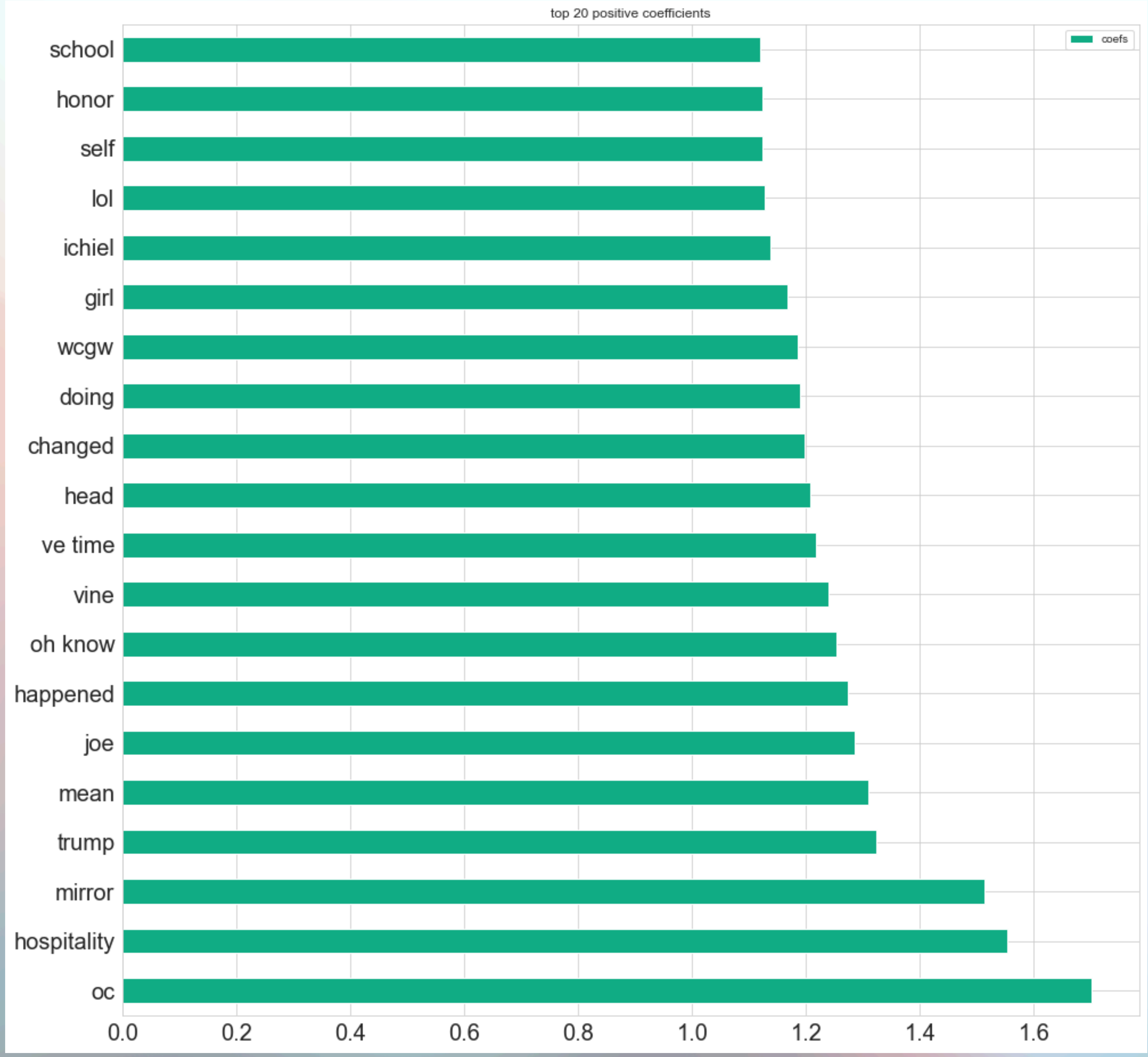
Findings - top 20 words predicted with RandomForestClassifier:

Best Words:

- ✓ Meme
- - Game
- ✓ Oc
- ✓ Like
- - Much
- - Still
- - One
- - Everyone
- ✓ Trump
- - Anyone
- ✓ Know
- - Call
- - Men
- - Maybe
- - Last
- - Way
- - Get
- ✓ Time
- - Every
- ✓ Moment

Best Words:

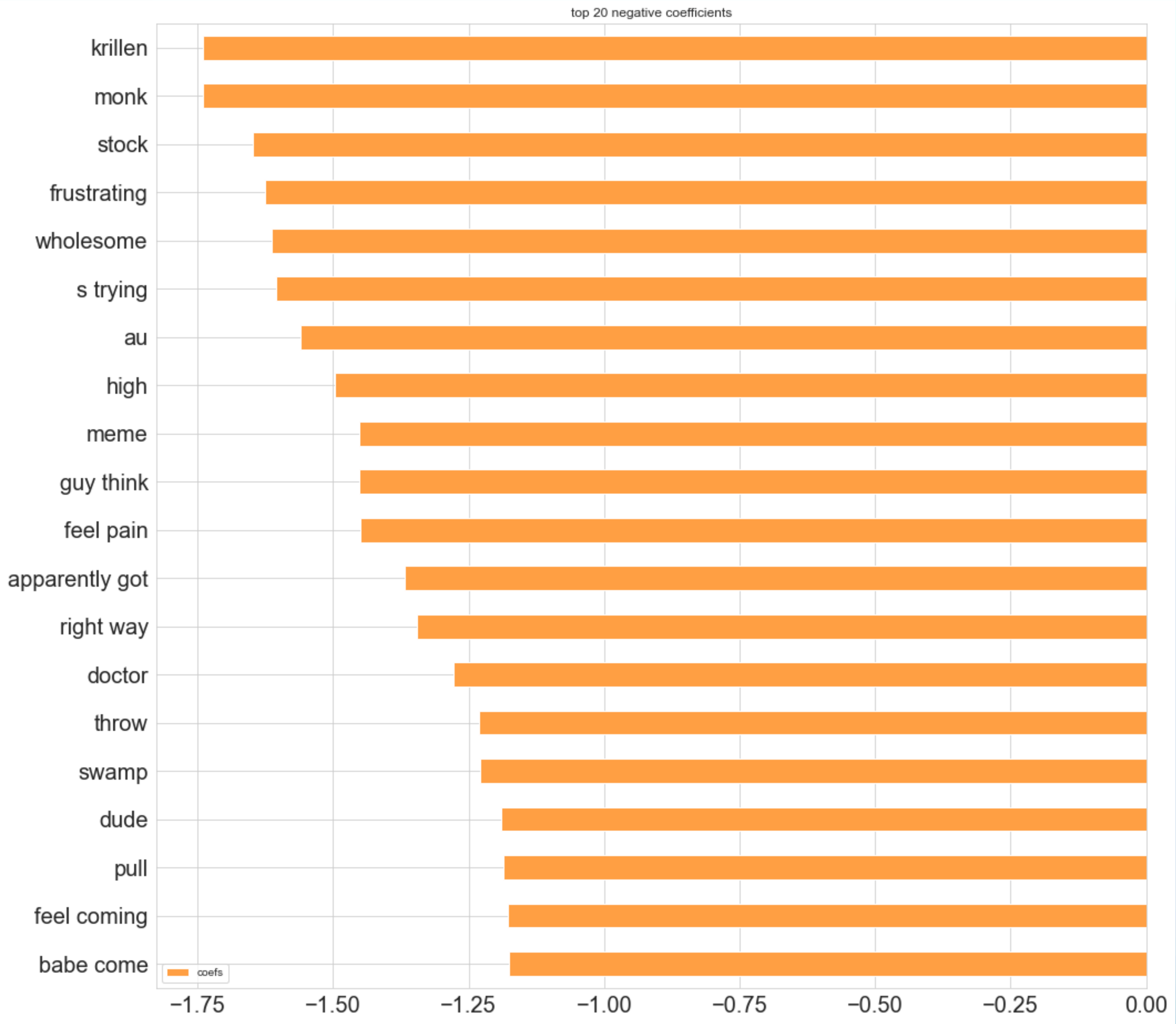
- ✓ Oc
- - Hospitality
- - Mirror
- ✓ Trump
- - Joe
- - Happened
- ✓ Oh know
- - Ve time
- - Head
- ✓ Anyone
- - Changed
- - Doing
- - Wcgw
- - Girl
- - Ichiel
- - lol
- - Self
- ✓ Time
- - Honor
- - School



Findings - top 20 words with lowest coefs:

Worst Words:

- Wow year
- Throw
- Animal
- Meme
- Talk
- Face
- War
- ll
- Night
- Work
- Say
- Happens
- Wasted
- Walk
- Wrong
- Hear
- Fact
- Story
- Anime_irl



Recommendations

Best Words:

- ✓ Meme
- Game
- ✓ Oc
- ✓ Like
- Much
- Still
- One
- Everyone
- ✓ Trump
- Anyone
- ✓ Know
- Call
- Men
- Maybe
- Last
- Way
- Get
- ✓ Time
- Every
- ✓ Moment

Best Words:

- ✓ Oc
- Hospitality
- Mirror
- ✓ Trump
- Joe
- Happened
- ✓ Oh know
- Ve time
- Head
- ✓ Anyone
- Changed
- Doing
- Wcgw
- Girl
- Ichiel
- lol
- Self
- ✓ Time
- Honor
- School

Best time to post:

Between 3pm to 8pm

Best Subreddits:

-r/memes
-r/dadjokes
-r/jokes
-r/theworldnews
-r/funny
-r/pics
-r/PokemonGoFriends
-r/me_irl

Sample Post:

I still think the best way to get anyone to like your oc meme is to get everyone to know about the game.

Thank you.

Thank you.