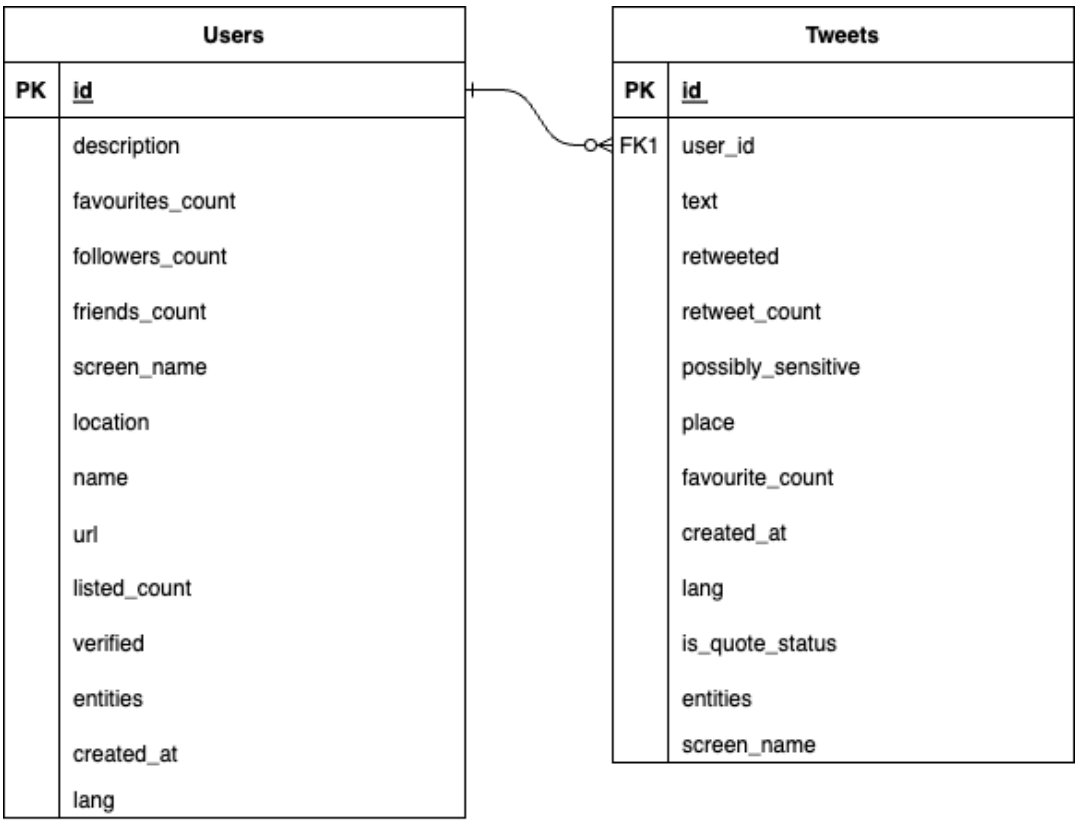
Step 1: Preparing for Your Proposal

1. Which client/dataset did you select and why?
   Client 2: Lobbyists4America (Congressional Tweets Dataset -- 2008-2017 data)
   This dataset contains congressional tweets from 2008-2017, and we can analyze this dataset to understand how different features of tweets (such as the number of words per tweet) affect the favourite count of the corresponding tweet and whether it is a key feature that affects the number of followers of the corresponding politician. I found this topic very interesting and I can get more key insights into how tweets affect the popularity of politicians. We could even try to forecast future popularity of a tweet or politician by some key features.
2. Describe the steps you took to import and clean the data.
   1. I used pyspark library to import the json file to a pyspark data frame
   2. Dropped duplicated rows
   3. Drop empty rows that contains null value in all columns
   4. Fill rows with missing values in some column with proper value
   5. Check the datatype for some column
   6. Convert the created_at column to year, month, hour columns
3. Perform initial exploration of data and provide some screenshots or display some stats of the data you are looking at.
   some screenshots and some stats visualisations are shown below.
4. Create an ERD or proposed ERD to show the relationships of the data you are exploring.

ERD of the congressional tweets dataset



```
1  df_users.summary().show()
2
```

✓ 0.7s

summary	created_at	description	favourites_count	followers_count	friends_count	id	lang	listed_count	location	name	screen_name	url
count	548	548	548	548	548	548	548	548	548	548	548	513
mean	1.3201181857229357E9	NULL	413.9124087591241	163433.90875912408	2033.7317518248176	7.236303389301189E16	NULL	1340.647810218978	NULL	NULL	NULL	NULL
stddev	7.915176373679677E7	NULL	965.1514395532447	1597357.0212540831	6278.436076002725	2.312212864947402...	NULL	3567.588266896501	NULL	NULL	NULL	NULL
min	1177689952		0	4	0	5558312	en	0		(((Rep. Nadler)))	AkGovBillWalker	http://stewart.ho...
25%	1.247616643E9	NULL	32	8862	365	56864092	NULL	426	NULL	NULL	NULL	NULL
50%	1.296736815E9	NULL	120	16690	748	246769138	NULL	749	NULL	NULL	NULL	NULL
75%	1.361487544E9	NULL	379	33042	1668	1209417007	NULL	1257	NULL	NULL	NULL	NULL
max	Tue May 25 16:02:...	proud husband, fa...	12507	31712585	92934	854715071116849157	en	70660	ÜT: 41.052377,-82...	tiberipress	virginiafoxx	https://t.co/zwcW...

User table basic stats summary

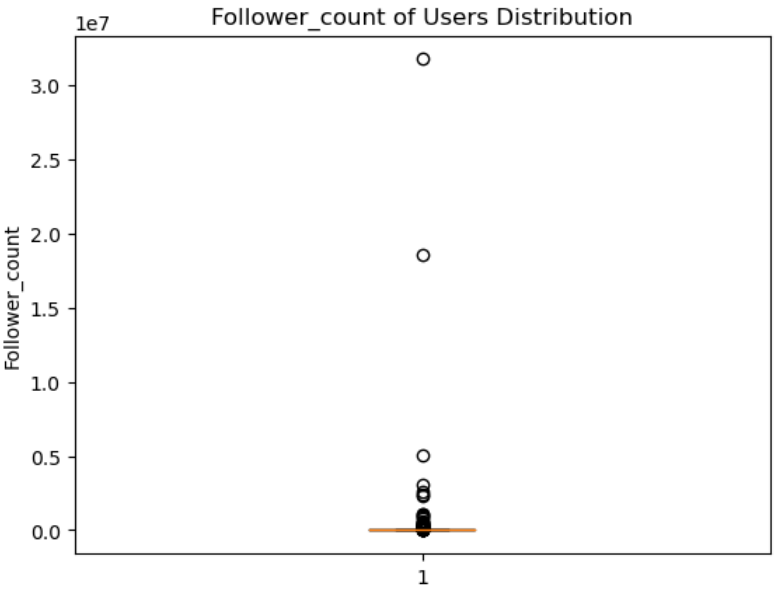
```
1  df_tweets.summary().show()
2
3
```

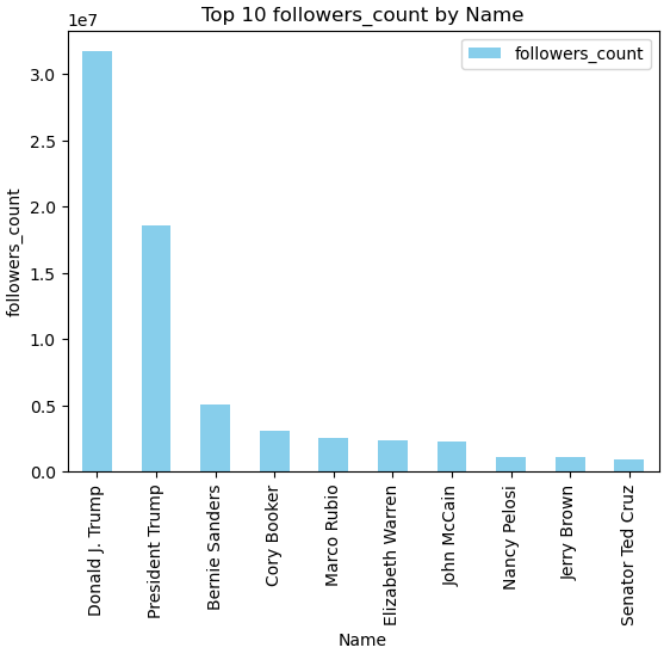
✓ 56.2s

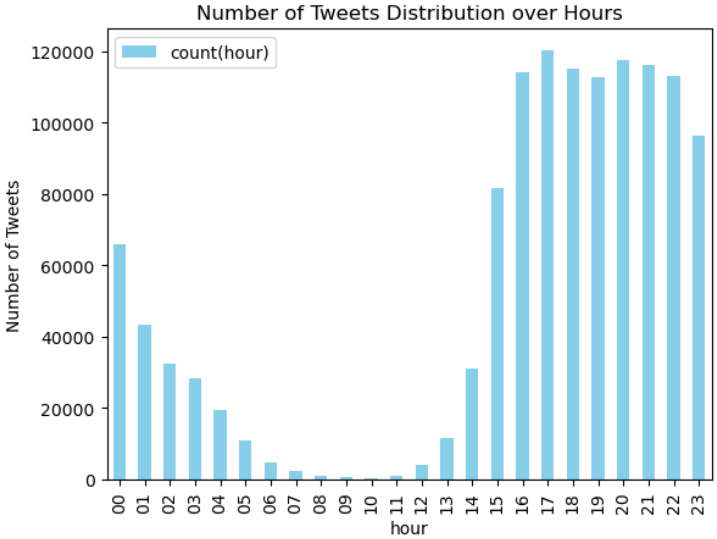
[Stage 25:> (0 + 1) / 1]

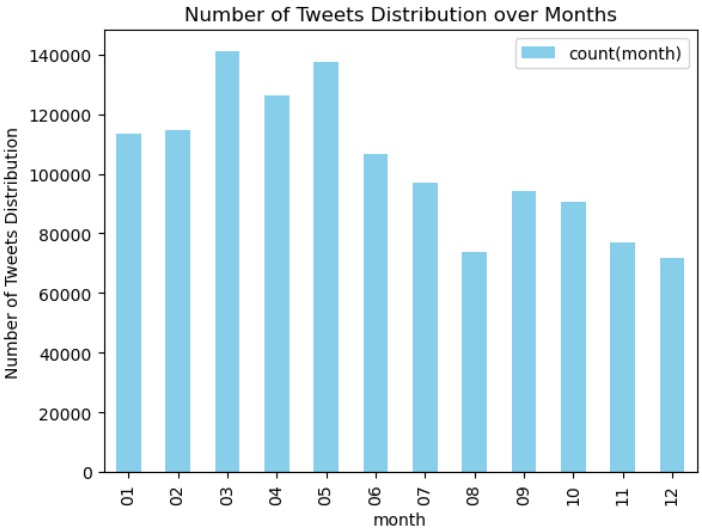
summary	created_at	favorite_count	id	lang	quoted_status_id	retweet_count	screen_name	text	user_id
count	1243370	1243370	1243370	1243370	56418	1243370	1243370	1243370	1243370
mean	1.4337994173263252E9	200.8516837305066	6.096595883812902E17	NULL	7.751529168911256E17	190.06283005058833	NULL	45.380530884529925	1.397404793854793...
stddev	5.226324353104003E7	3545.4052124108666	2.140925153240984...	NULL	7.897866282630867E16	9944.391751135185	NULL	37.15300697329736	1.053382691017018E17
min	1217870931	0	877418565	cs	90957762	0	AkGovBillWalker	https://t.co/OJ5...	5558312
25%	1402512021	0	476796091975733248	NULL	725038291448418304	1	NULL	3.141592653589793	33750798
50%	1446758247	2	662378213931360256	NULL	794546014644555776	4	NULL	60.0	234022257
75%	1475096503	8	781237476388179968	NULL	839263333295484928	10	NULL	73.0	993153006
max	1496769360	984832	872140026737336320	zh	872138292199923718	3637896	virginiafoxx	👱 https://t.c...	854715071116849157

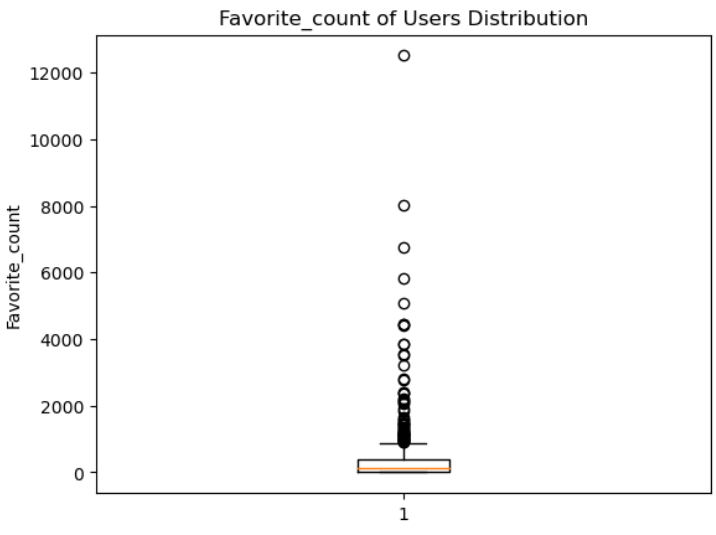
Tweets table basic stats summary

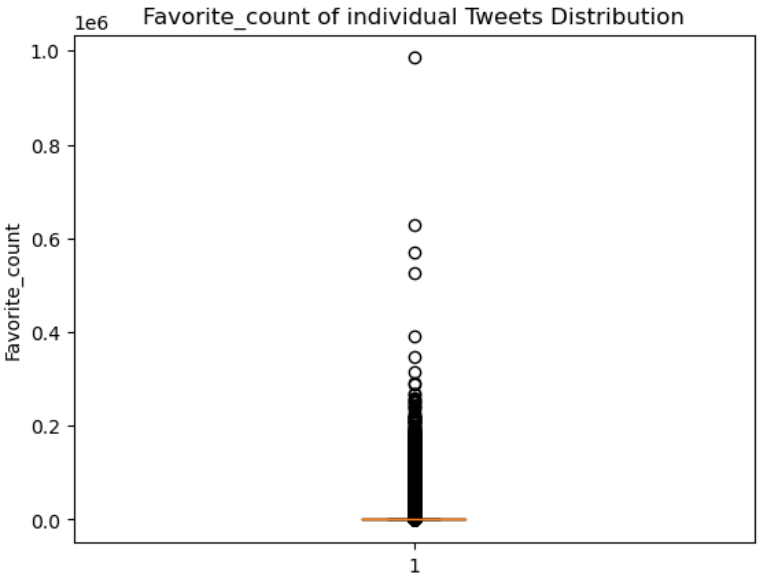

Follower_count of Users Distribution
1e7
Follower_count
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1


Top 10 followers_count by Name
1e7
followers_count
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Donald J. Trump
President Trump
Bernie Sanders
Cory Booker
Marco Rubio
Elizabeth Warren
John McCain
Nancy Pelosi
Jerry Brown
Senator Ted Cruz
Name


Number of Tweets Distribution over Hours
count(hour)
Number of Tweets
120000
100000
80000
60000
40000
20000
0
00
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
21
22
23
hour


Number of Tweets Distribution over Months
count(month)
Number of Tweets Distribution
140000
120000
100000
80000
60000
40000
20000
0
01
02
03
04
05
06
07
08
09
10
11
12
month


Favorite_count of Users Distribution
Favorite_count
12000
10000
8000
6000
4000
2000
0
1


Favorite_count of individual Tweets Distribution
1e6
Favorite_count
1.0
0.8
0.6
0.4
0.2
0.0
1

Step 2: Develop Project Proposal

Description

This project is about analysing congressional tweets data from the USA to determine how tweets, politicians could affect the congress most, so as to understand legislative trends and develop better lobbying strategies.Audiences are companies like Lobbyist4America that try to help their customers who want to affect legislation by using effective lobbying strategy.

Questions

1. What is the average word length of a popular tweet?
2. Whether a positive tweet would be more popular?
3. What is the average tweet count of a popular politician?

Hypothesis

What are your initial hypotheses about the data?

1. The politician who has posted more tweets overall have more followers.
2. The politician who has more friends on twitter have more followers.
3. The longer political tweets (more words) are more popular.
4. More positive tweets get more favourites.
5. Retweeted count is positively correlated with favourite count

Approach

I plan to look at following features at first: text, retweeted_count, favourites_count, followers_count, created_at.

I want to explore whether the positiveness of a tweet affects the popularity of the corresponding tweet, in order to check this relationship, I plan to use NLTK library to analyse the positiveness of a tweet.

Followers_count of a politician and favourite_count of a tweet are metric that I plan to use.