Overall Summary

- Some columns were weird, a chunk of columns can be ignored
-

ProPublica	Ad Observer	description
political		
not_political		
message	Ad text	The detail of ad → sentiment analysis and such
created_at	observed_at/v9	Need to choose format to include time or not
updated_at	v10	Need to choose format to include time or not
impressions		
Political probability	Political value	Probability that it is political
targetings/targets	targetings	Need lots of cleaning Clean age Clean location Clean interests
advertisers		
entities		
paid_for_by		
targetedness		
list_building_fundraising_proba		

- Don’t trust DataExplorer, seems like it registers empty string as input as well

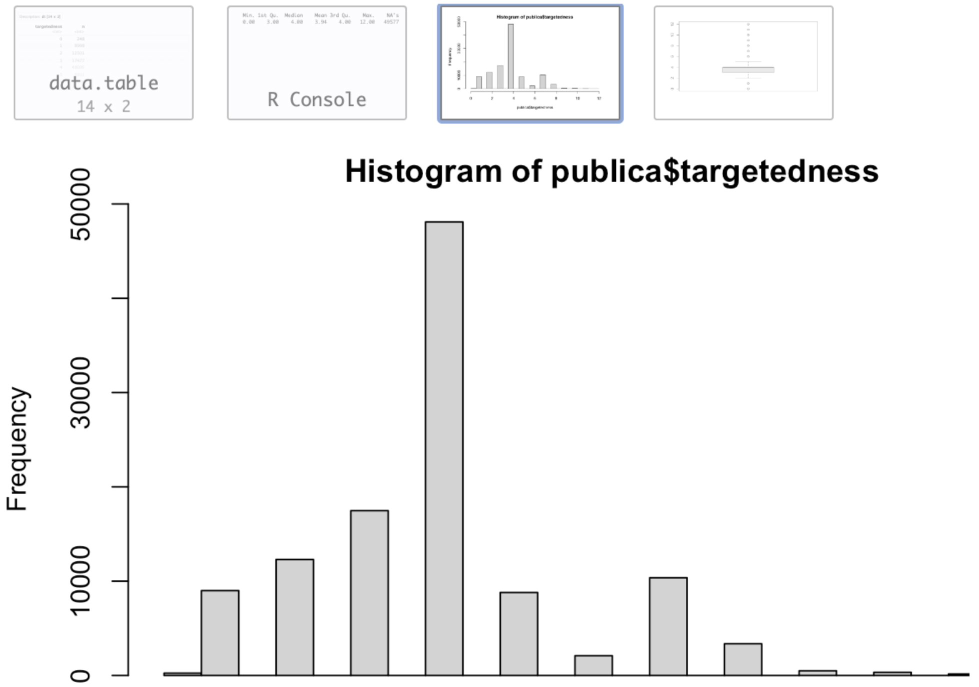
ProPublica

- id: post id number on facebook
    - Just classic id → nothing much
- html: HTML of the ad as collected by the Political Ad Collector
    - Just html link → nothing much

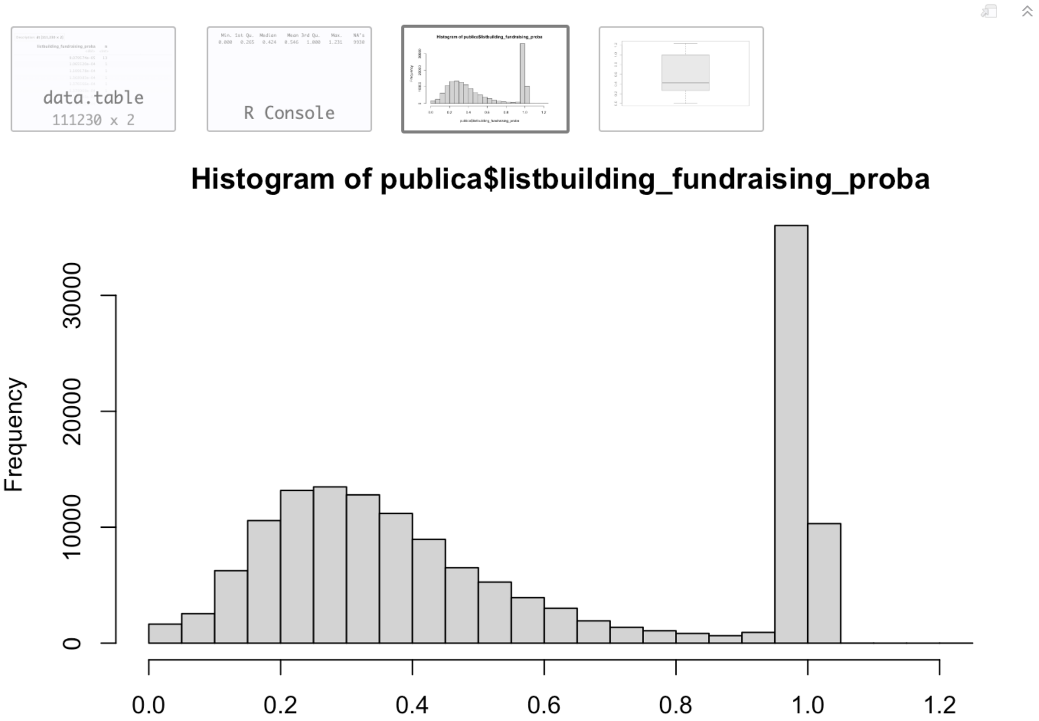
- political: number of Political Ad Collector users who have voted that the ad is political
    - Min of 0, median 1, max of 488
    - Heavily right skewed
    - Possibly useful, though need thought into how to include
- not_political: number of Political Ad Collector users who have voted that the ad is not political
    - Min of 0, median 0, max of 330
    - Heavily right skewed
    - Possibly useful, though need thought into how to include
- title: ad title
    - Title of ad
    - Nothing to be done for eda, though prob nice to do things along length of title and what not later
- message: ad content
    - The full message/content
    - Prob nice to do word analysis and sentiment analysis
- thumbnail: link for a thumbnail of the profile image (of the advertiser)
    - A link → nothing much
- created_at: date ad was first collected by the Political Ad Collector
    - This is the same as v9 of adobserver
- updated_at: the most recent time that it got an impression OR the most recent time it was voted on
    - This is the same as v10 of adobserver
- lang: language of the ad. always en-US.
    - Nothing much → they are all us, so not important
- images: link for images included in the ad
    - Nothing much and probably not important, unless we are willing to do some convolutional neural net, but IDK how we would proceed with that
- impressions: number of times the ad has been seen by the Political Ad Collector
    - Min 0, median 1, max 575
    - Heavily right skewed
- political_probability: calculated by the classifier. data only includes ads with a probability >=0.7
    - SO, this is pretty important
    - Also the probability ranges from 0 to 1. I presume that the description of >= 0.7 refer to the confidence? Not too sure
- targeting: Facebook's “Why am I seeing this?” disclosure provided to Political Ad Collector users
    - This is in weird format that I haven't seen before
    - Need to segment out many things
        - Age stored in two different ways: 18 to 54 vs 18 and older, so need to parse them and make sure they are in same format
            - Treat them as min age and max age and set max as 99 for “and older”

  - Treat them as “factors,” but would require a bit of discussion in how to format this “factor”
- Location saved in different format, and it goes everywhere
  - Some i saw was region, which referred to country as well as state
  - But then there is also a “state” categorization that categorizes it as state
- Preference
  - Stores information about target’s preference and so forth. Again, another nightmare for munging

- suppressed: value is false. suppressed ads are excluded from this data set because they were misclassified.
  - They are all false and we can ignore
- targets: a parsed version of targeting
  - This seems to be in json format. Use either this or targets, and dont do both since they are same except the format
  - Need to segment out many things
    - Age stored in two different ways: 18 to 54 vs 18 and older, so need to parse them and make sure they are in same format
      - Treat them as min age and max age and set max as 99 for “and older”
      - Treat them as “factors,” but would require a bit of discussion in how to format this “factor”
    - Location saved in different format, and it goes everywhere
      - Some i saw was region, which referred to country as well as state
      - But then there is also a “state” categorization that categorizes it as state
    - Preference
      - Stores information about target’s preference and so forth. Again, another nightmare for munging
- advertiser: the account that posted the ad
  - String values
- entities: named entities mentioned in the ad, extracted using software
  - Think of it as a “focus/target/main subject”
  - Have multiple in this single value
- page: the page that posted the ad
  - Web url
- lower_page: the Facebook URL of the advertiser that posted the ad (the “page” column, lowercased)
  - Web url
- targetings: an array of one or more of Facebook’s “Why am I seeing this?” disclosures provided to Political Ad Collector users
  - Its in that weird format, and I dont really know how to decipher it
- paid_for_by: for political ads, the entity listed in Facebook’s required disclosure as having paid for the ad

    - Its just who paid for it
- targetedness: an internal metric for estimating how granularly an ad is targeted, used for sorting in the ProPublica search interface
    - Not really sure what it means
    - Min 0, med 4, max 12
    - 

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00    3.00    4.00    3.94    4.00   12.00   49577
```

- Listbuilding_fundraising_proba: there is no description of what this was
    - 

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   0.265   0.424   0.546   1.000   1.231    9930
```

AdObserver

- Ad_id. Unique ID generated by Meta for this ad. (There may be multiple duplicate ads with the same content, but different ids.)
    - Just id
- Page_name. Page name of the entity which published the ad.
    - As it says, might be nice to look at
- **Political_value.** A number ranging from 0-1, populated by C4D's political classification model. Ads that Meta indicates are political (because they have disclosures) have a value of -1, as C4D's political classifier did not run on these ads.
    - Min -1, med 0.84, max 1
    - bimodal
    - Most data at -1, but other sets of data starting from 0.8 as minimum
- Paid_for_by. Name of the entity listed in the ad disclosure. This field will be empty if the ad does not have a disclosure (that is, it's not a declared political ad).
    - names
- Ad_text. Text content of the ad.
    - Run some sentiment analysis and such
- **Observed_at.** Date/time when ad was observed.
    - Consider it as "Created_ad" from propublica
- Call_to_action. Text on button in ad, if there is one. (This might be donate, subscribe, etc.)
    - Not really sure what to make out of it. Looked at some which had foreign language, seem to have many different formats that goes from just hashtags to donate messages to something i dont understand
- Targetings. Jsn object containing all of the targeting information collected for this ad. A description of the specific fields in this json object are shown below.
    - This is similar to targets of propublica
    - A bit "cleaner" in terms of cleaning for json
- V9
    - Basically created_date, but without time (hr/min/sec)
- V10
    - Bascially updated_date, but without time (hr/min/sec)