Milestone 1: Peoject Proposal and Data Selection

Daniel Baquero

Preparing for the proposal
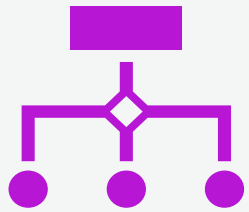
Dataset:

Lobbyists4America.

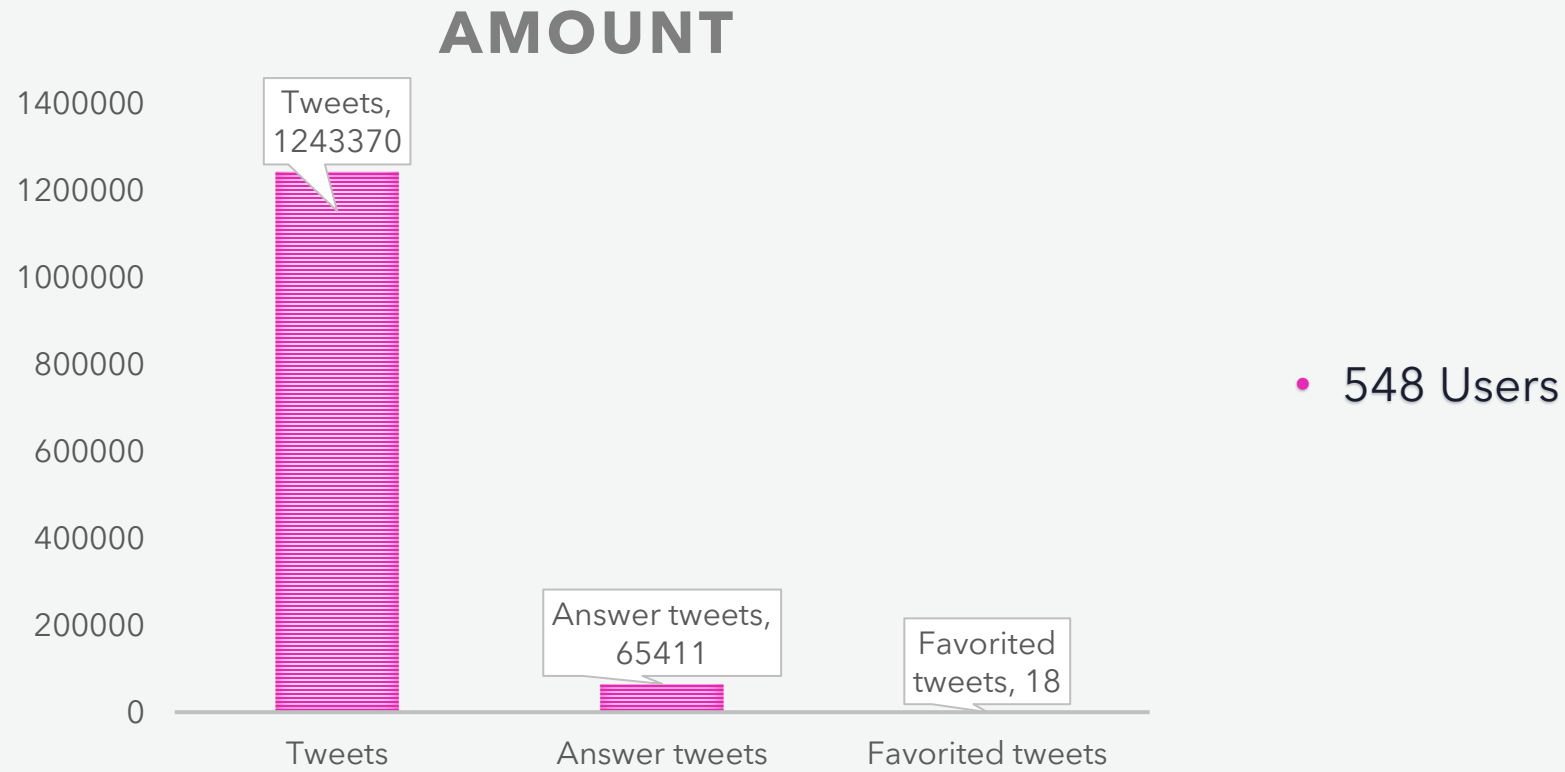Using data in social applications is a complicated, yet fulfilling, branch of data science.

Import and Cleaning:

Unpack the compressed files.

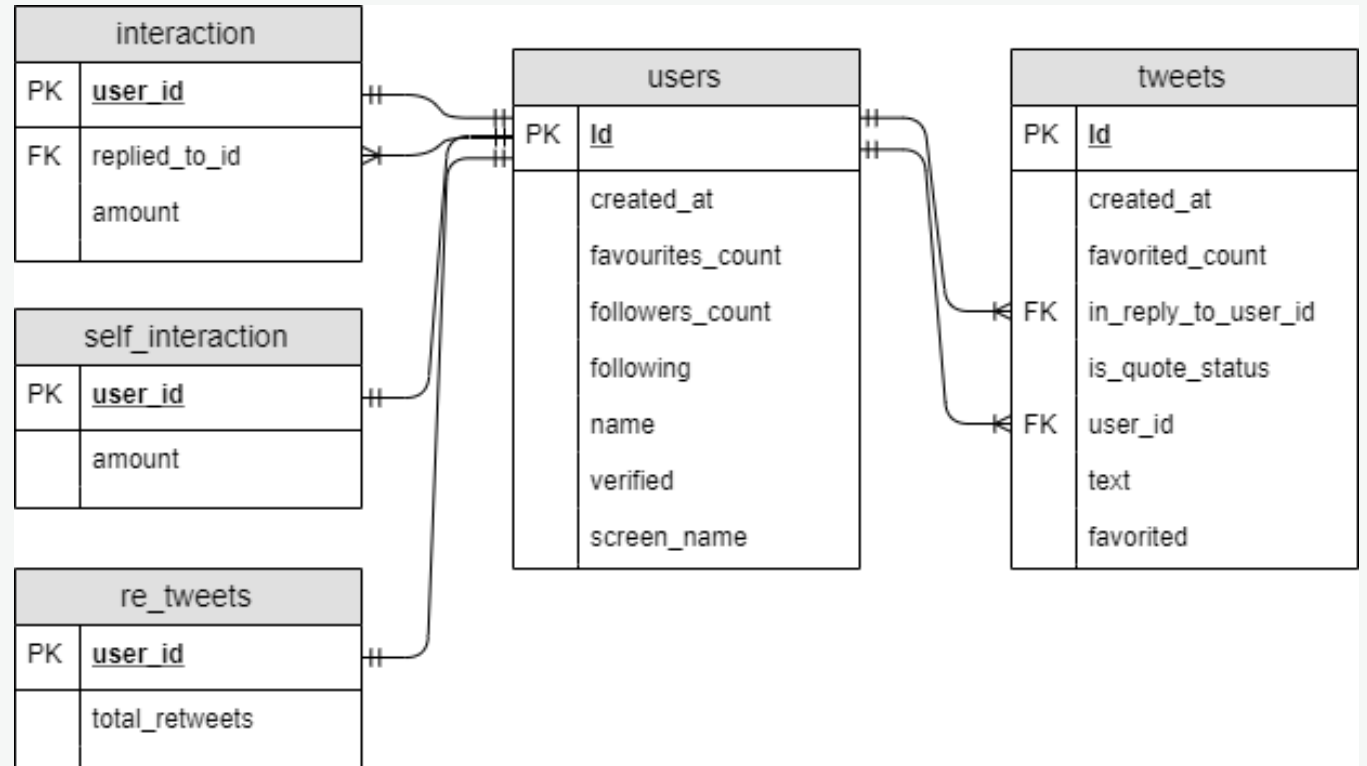Use DataBricks table creator to import the JSON files as tables.

Look for principal key.

Divide tables in useful parts.

Preparing for the proposal: EDA

Preparing for the proposal: ER diagram

Milestone 1:

Project Proposal Development

Description

- Lobbyst4America objective is to provide insights for people who aim to affect legislation. A good strategy and an influential objective is the heart of lobbying. However, selecting the objective can be difficult. By analyzing the tweets data it is possible to find the most influential person within the congress for specific topics. The project aim is to define who is the most influential person related to the congress in Tweeter.
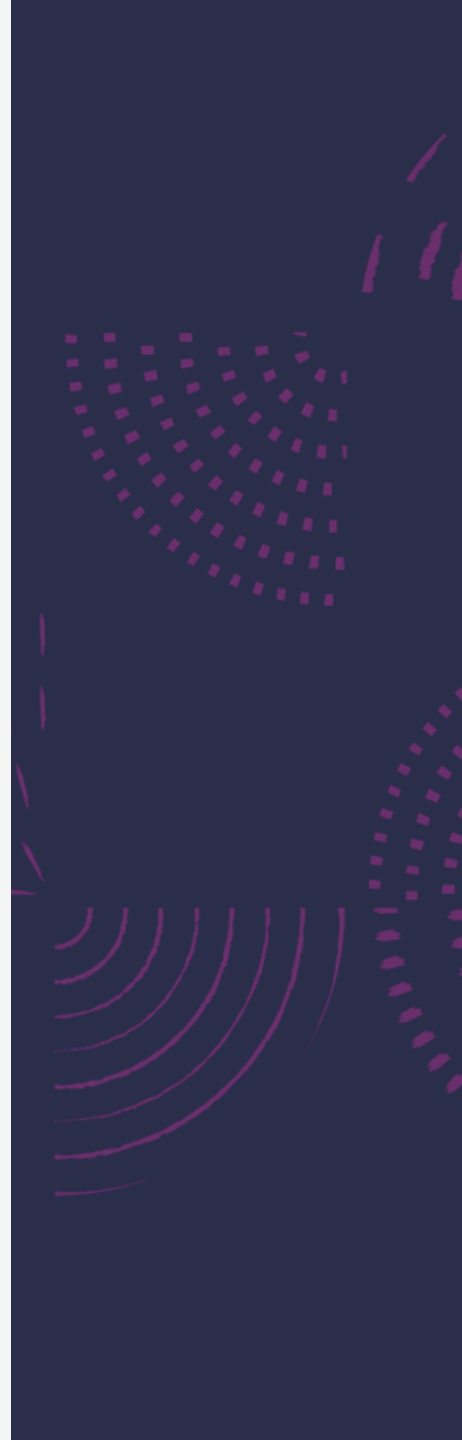
Questions

Who has the most followers in tweeter?

Who is the most active person in tweeter?

Is there any unusual activity at the most recent times in tweeter?

Been a verified user has any impact on followers?

Hypothesis

- People who retweet or answer to themselves don't have as many followers as people who interact with others.

- People with favorited tweets have the most followers.

- The more tweets activity, the more followers.

Approach

- Every question can be answered with existing fields such as:
  - Favorities count
  - Followers count
  - Favortied
  - Total retweets
- Metrics are given as the total number or count of boolean fields.
- By plotting the relation between different metrics the most influential person can be found.
- The aim is to predict who can be influential in the future by comparing similar activity profile with influential people in the present.