

Mapping COVID19 Hotspots with Social Media

Dana Hackel, Jenny James,
Emmanuel Akindele, Zakaria Zerhouni

Executive Summary

- ▶ Problem Statement
- ▶ Data Collection and Exploration
- ▶ Developed Product and Assessment
- ▶ Future Directions

Problem Statement

The data released regarding instances of the COVID-19 pandemic is aggregated before it is released to (legally and ethically) protect the privacy of those involved. Unfortunately, this takes away some of the utility of the data. Using social media, the location of cases can be narrowed further while still protecting privacy rights. Social media data can provide a heat map for *potential* COVID-19 risk.

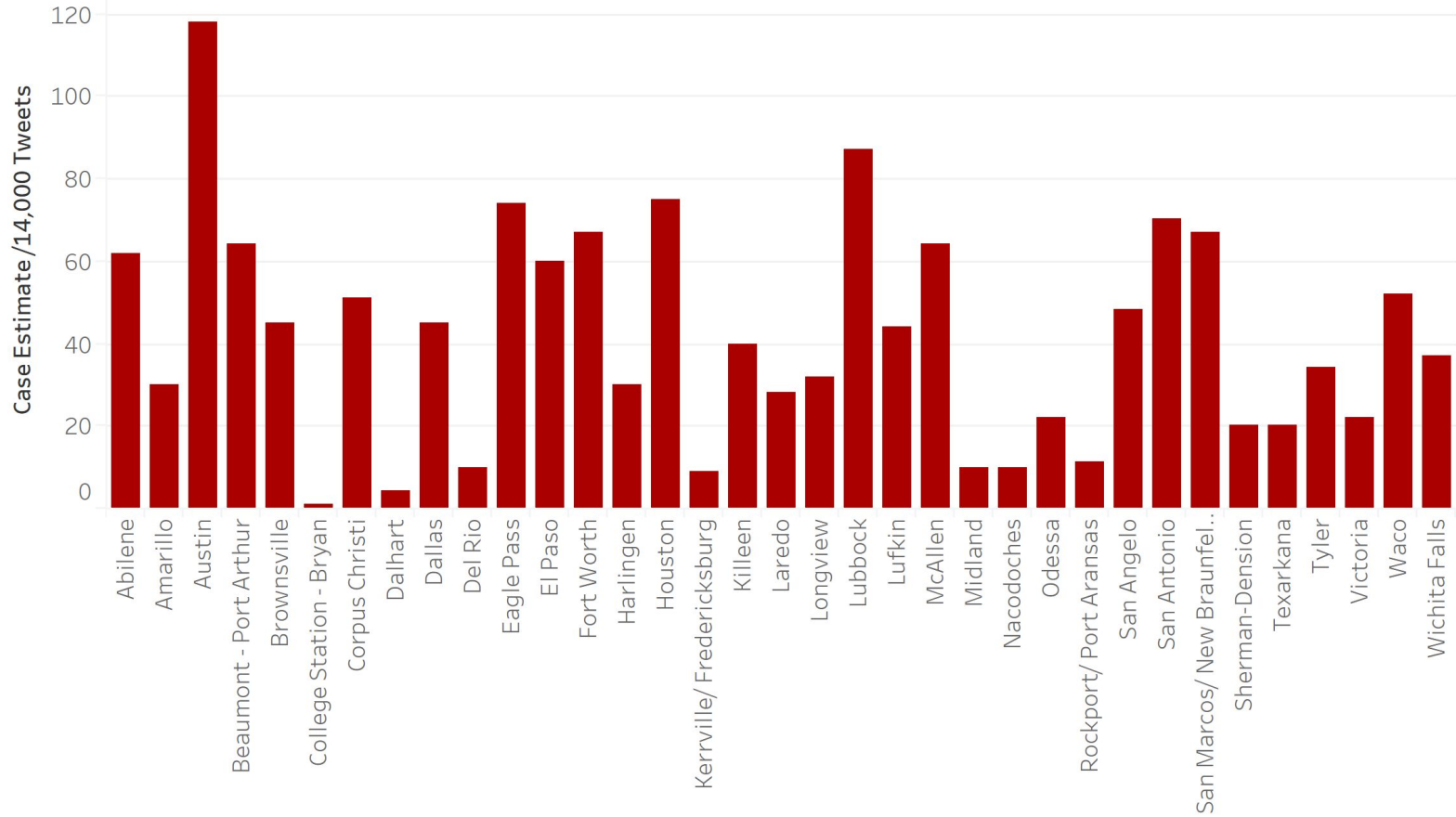
Data Collection

- ▶ GetOldTweets3
- ▶ 8-30-2020 to 9-5-2020
- ▶ 2,000 tweets per day, per city
 - ▷ 14,000 per city (or less)

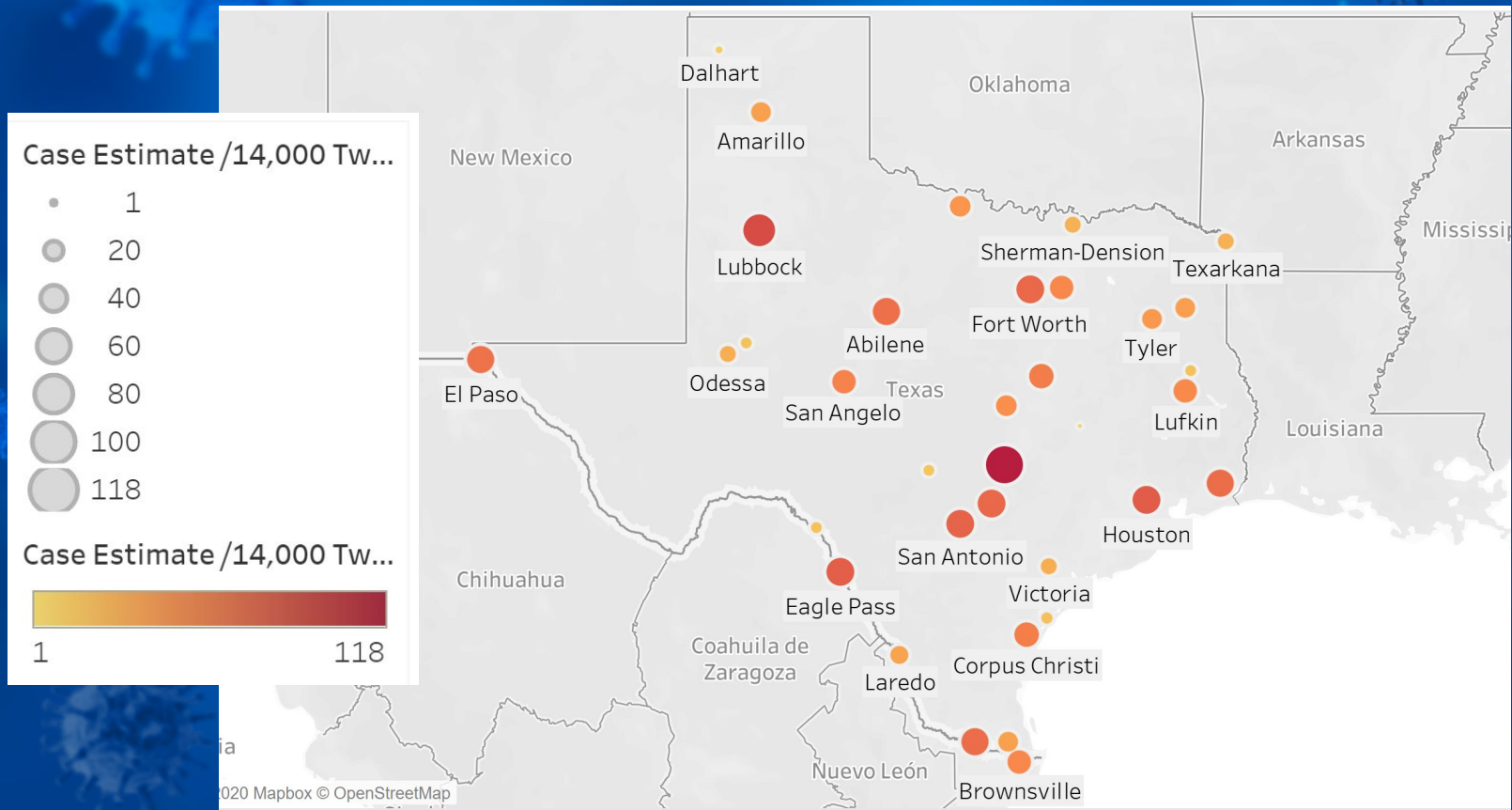
Cleaning and Exploration

- ▶ Double-check dataframe and datatypes
- ▶ Keep “empty” tweets, or other languages
 - ▷ Still represents a tweet not mentioning covid
- ▶ Count all tweets which mention “covid” or “coronavirus”

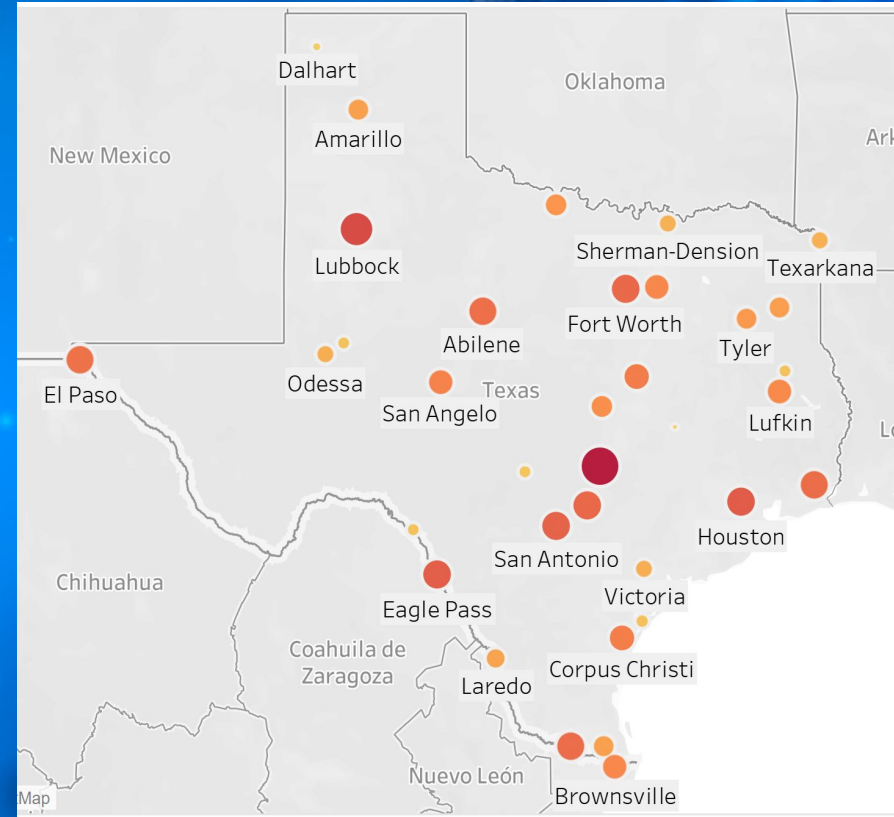
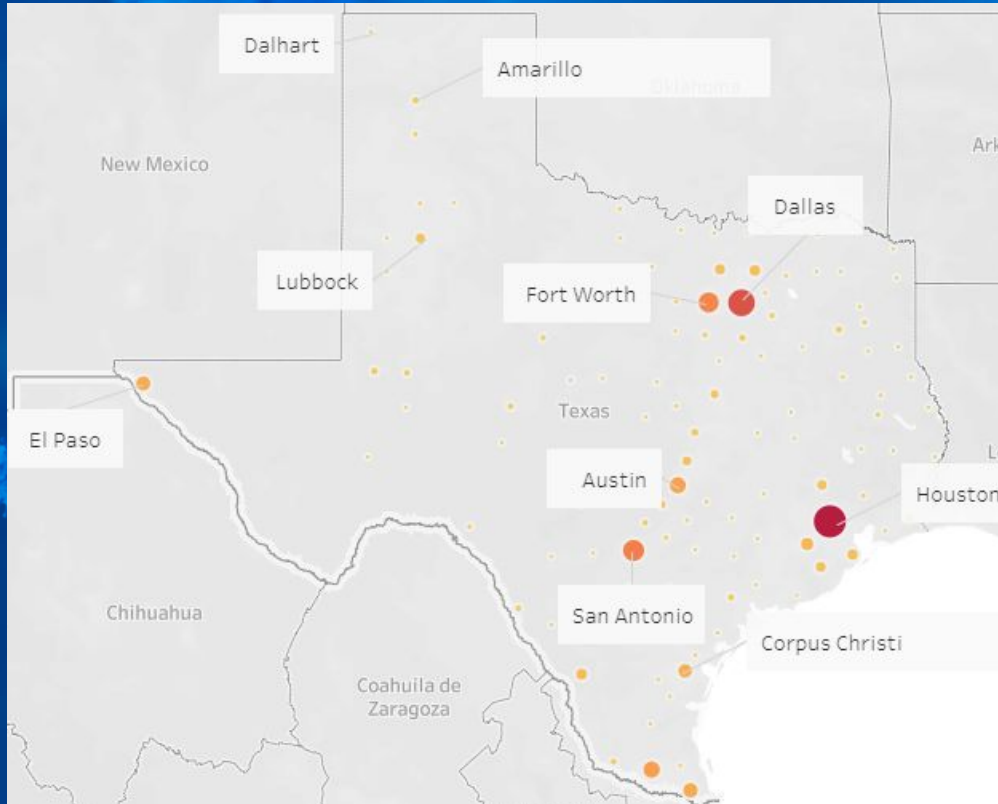
COVID Mentions per City



Predictive Heat Map



Actual Heat Map



<https://public.tableau.com/profile/emmanuel.akindele#!/vizhome/TexasCovid/Sheet2?publish=yes>

Conclusions

- ▶ COVID mentions can be representative of COVID cases in some cities, but not others
 - ▷ College towns often overrepresented

Future Directions

- ▶ Pull more tweets
- ▶ Pull tweets from other states
- ▶ Incorporate other languages
- ▶ Predictive model - can we predict risk of an area based on tweet content?

Predicting Covid Risk

- ▶ Data: tweets and metadata from Twitter combined with COVID-19 counts of an area.
- ▶ Natural Language Processing: a model can be built to predict a value based on textual and numerical data pulled from twitter.
- ▶ Data from the Dallas/Ft. Worth Texas and achieved the best results with a Random Forest Tree Regressor.
- ▶ Additional preprocessing of the data and benchmarking would provide actionable data to provide risk assessment for an area using social media data.

References

- ▶ <https://github.com/Mottl/GetOldTweets3>
- ▶ <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.str.contains.html#pandas.Series.str.contains>
- ▶ <https://public.tableau.com>
- ▶ https://en.wikipedia.org/wiki/List_of_Texas_metropolitan_areas
- ▶ <https://www.gps-coordinates.net/>
- ▶ https://www.google.com/search?q=covid+cases+in+texas+by+city&rlz=1C1CHBF_enUS906US906&oq=covid+cases+in+texas+by+&aqs=chrome.2.0j69i57j0l6.8686j0j9&sourceid=chrome&ie=UTF-8
- ▶ <https://github.com/nytimes/covid-19-data>
- ▶ <https://covidtracking.com/data/state/texas>

