



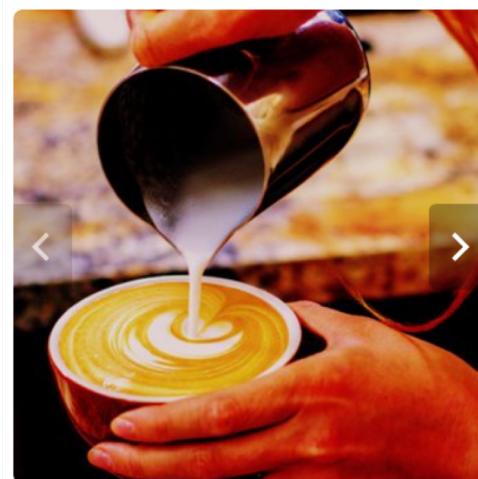
Estimating Neighborhood Affluency with Yelp

Shannon Bingham, Roy Kim

January 18, 2019

Problem Statement

The goal of the project is to estimate the affluence of a neighborhood based on the number of \$ of businesses and services (according to Yelp) in a given neighborhood.



7. JBC Coffee Roasters

17 reviews

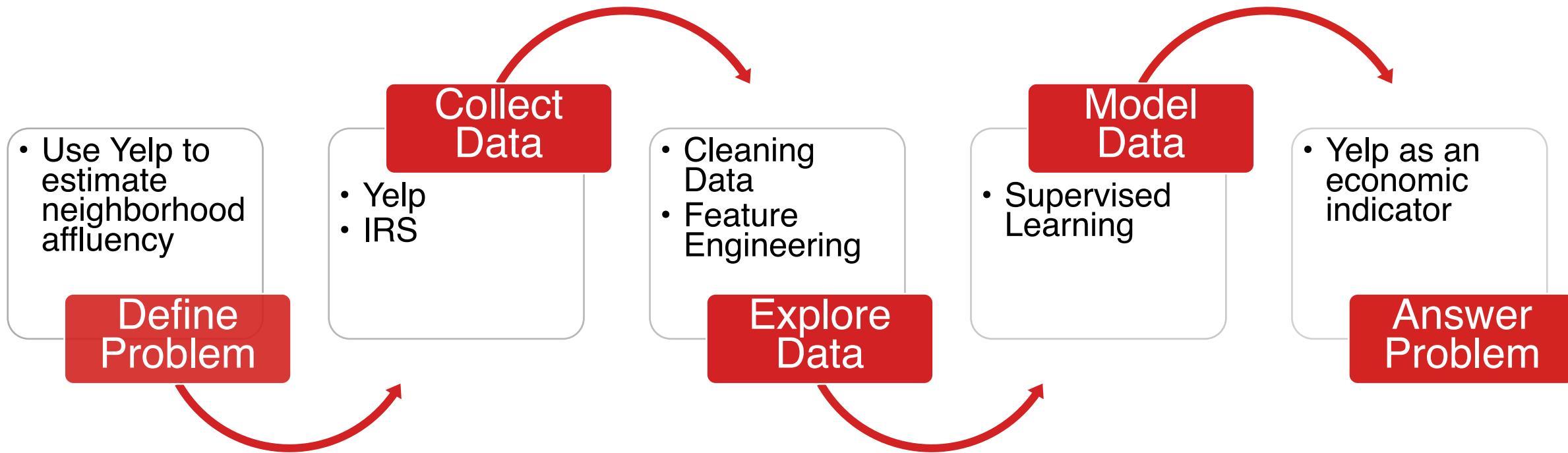
\$\$ · Coffee Roasteries

(608) 256-5282

5821 Femrite Dr

"This **coffee** roaster is the best, not only do they have the best **coffee**, they also have the best customer service. Keep it up Johnson brothers **coffee**. Your the..." [read more](#)

Data Science Process





1

Define Problem

How can we approach
the problem to quantify
success?

Introduction

- The rise of social apps related to commercial activity has given birth to a wealth of data with many applications.
- One of such social apps is Yelp, and the data collected by Yelp is descriptive of businesses by area and location. Specifically, Yelp uses certain markers (\$, \$\$, \$\$\$, \$\$\$\$) to indicate the relative cost of the services of a business.
- Can this data be used to estimate the affluence of a neighborhood (AGI), or at least be an added indicator to what is already available?

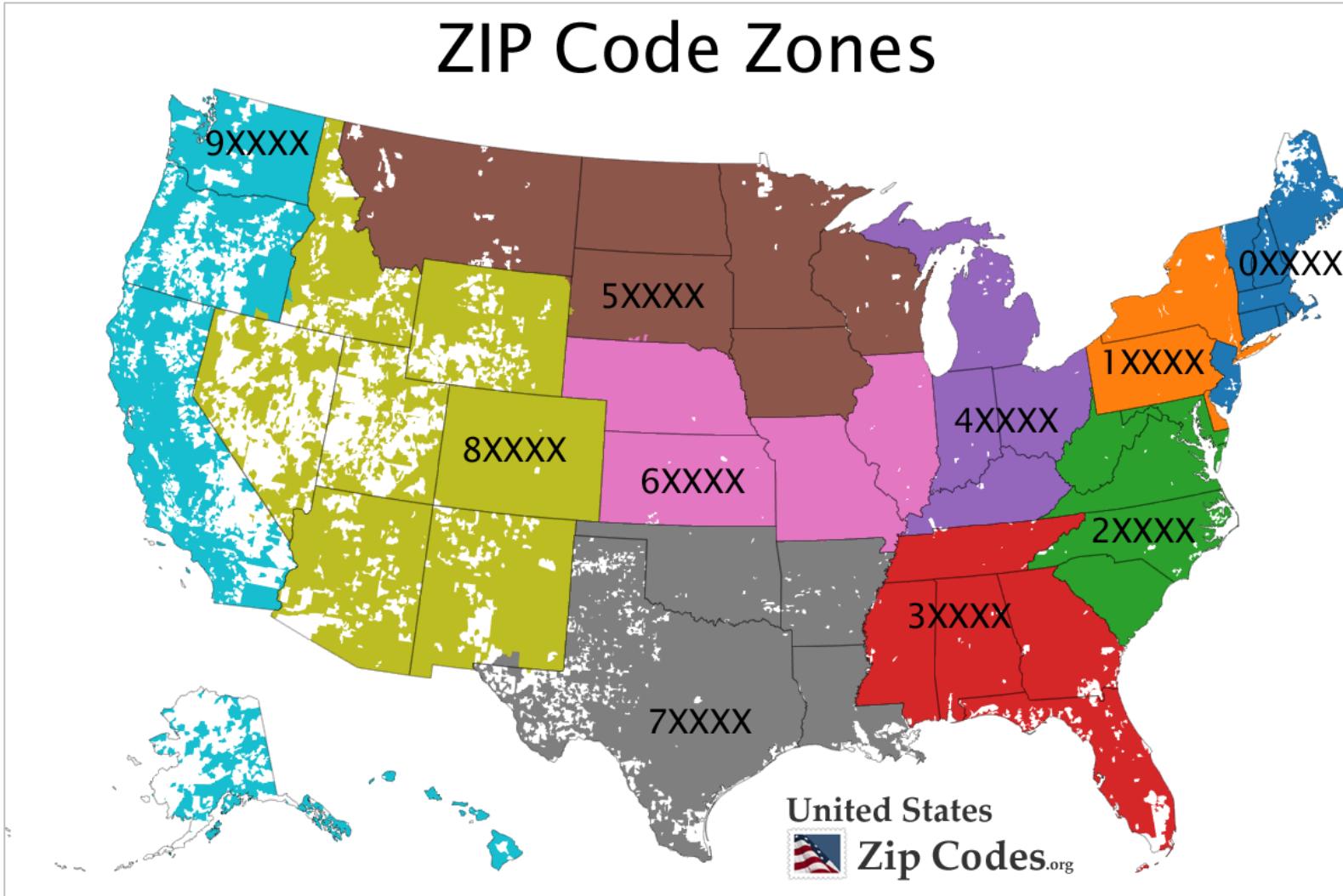


2

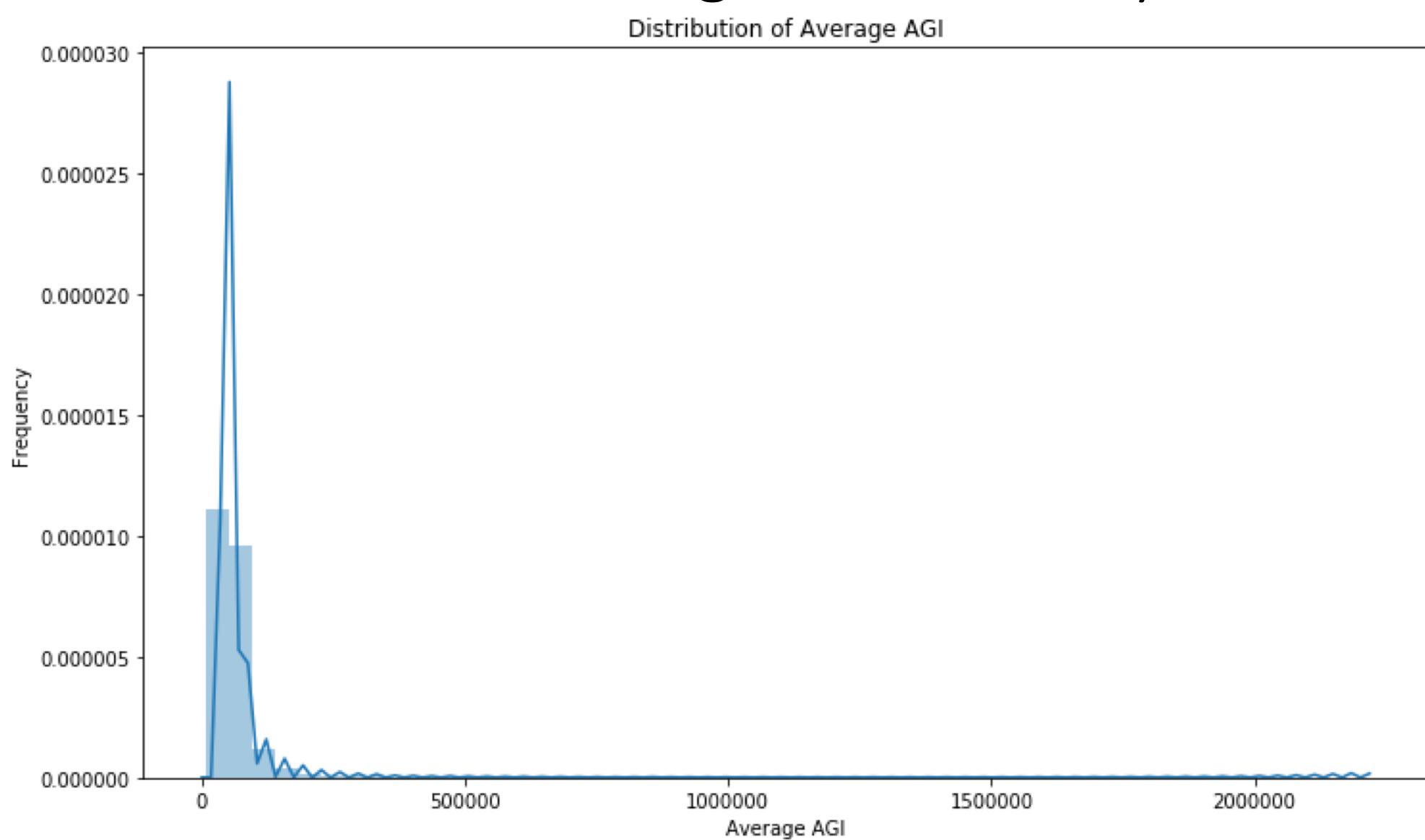
Collect
Data

Where are we looking
for data and how are we
getting it?

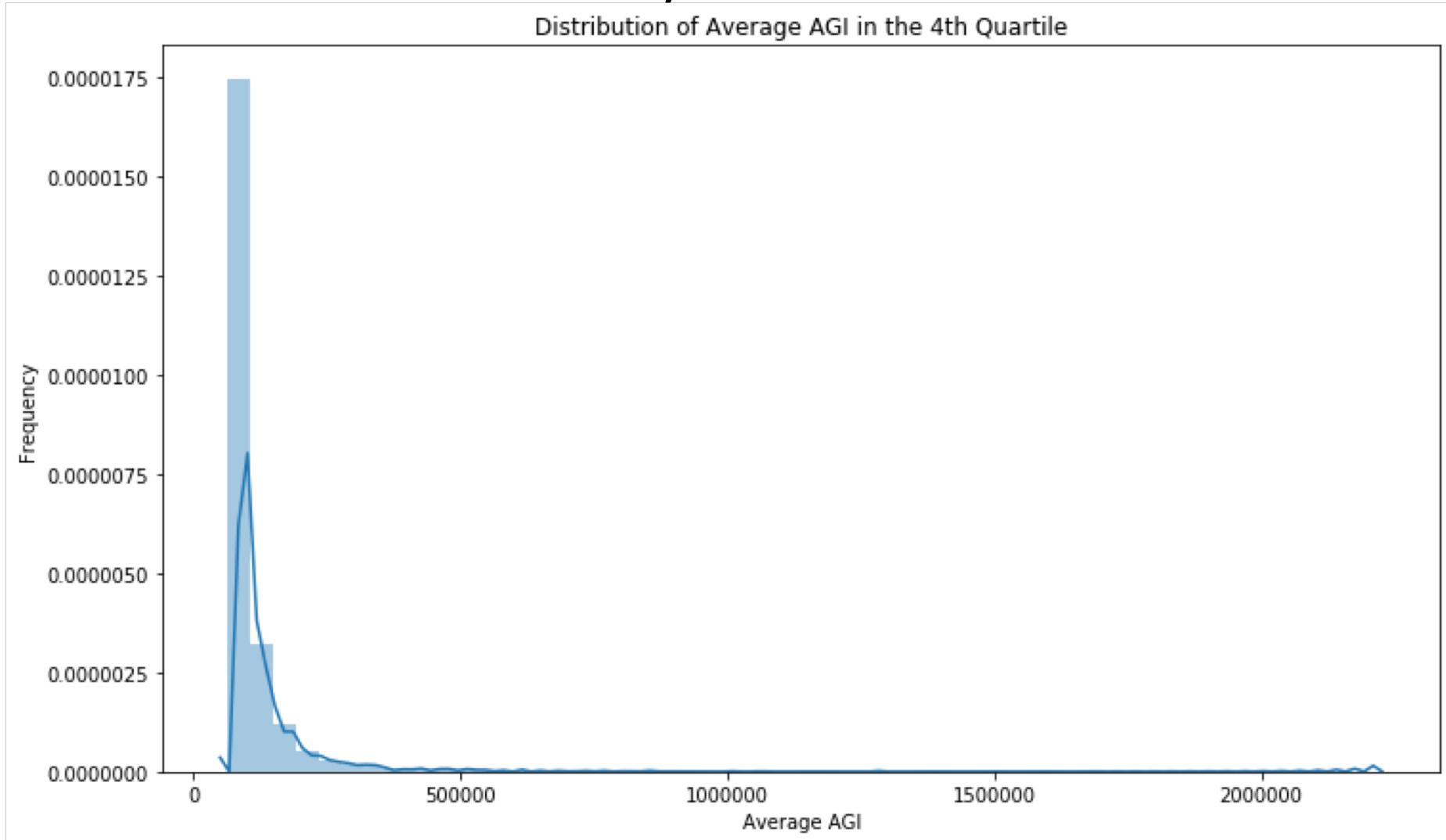
There are 29872 unique zip codes in the U.S.



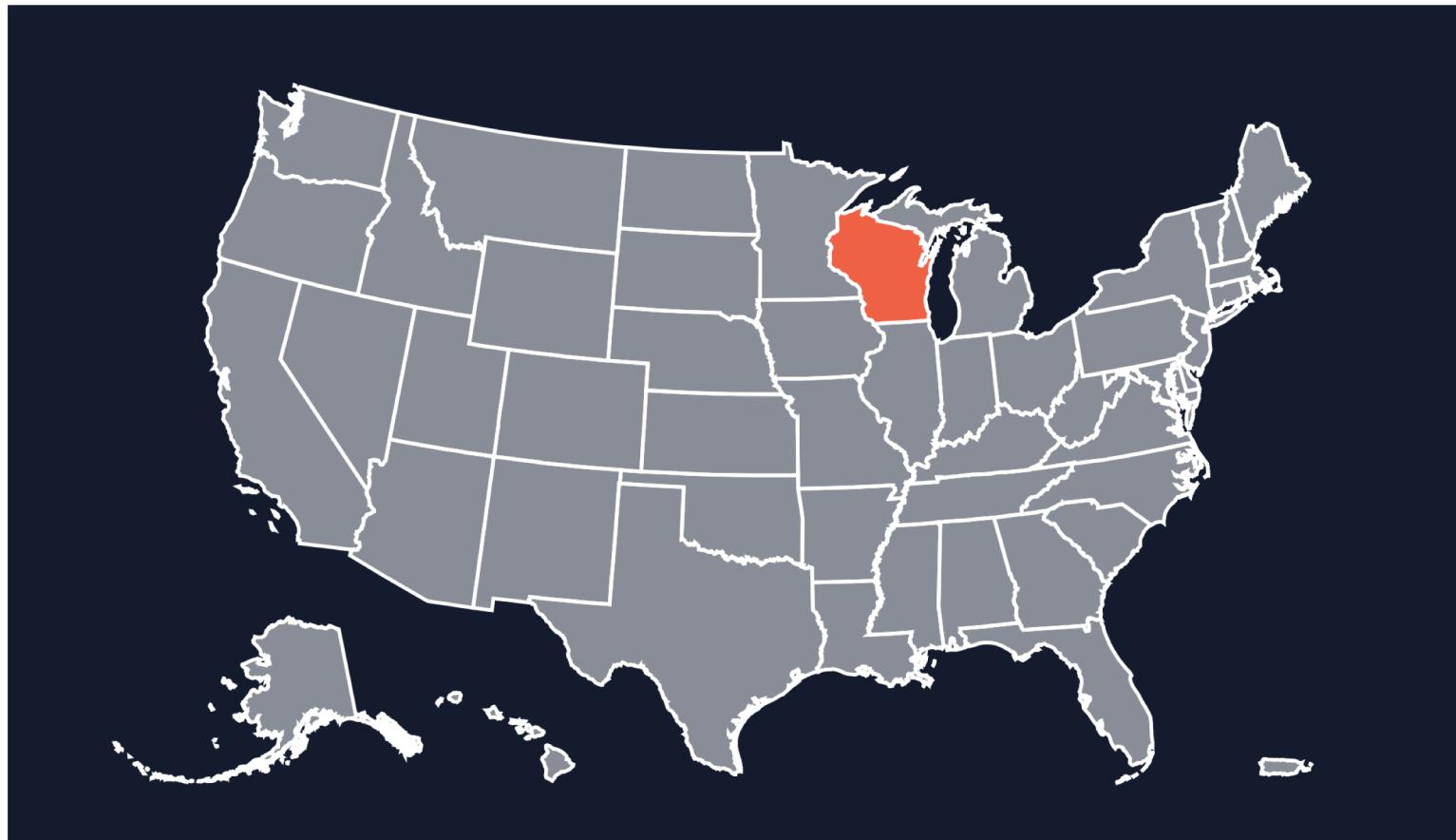
The distribution of average AGI is not symmetrical.



Even among those with the highest average AGI,
the distribution is not symmetrical.



Sample data was selected from Wisconsin.



<https://datausa.io>

Average AGI near
the median for U.S.

882 zip codes.

Zip codes represent
rural and urban
areas.

Used the basic Yelp interface to get sample data.



Advantages:

Free
Real-time information

Disadvantages:

Limitations on retrieval



3

Explore
Data

What does the data
tell us?

Many records did not have price/rating data.



4. Fromagination

199 reviews

\$\$ • [Cheese Shops](#), [Sandwiches](#),
[Cheese Tasting Classes](#)

(608) 255-2430

12 S Carroll St

Capitol

“ a wide array of delicious cheeses from Wisconsin and around the world, and samples are always available. Additionally, they sell other **specialty food** items...” [read more](#)

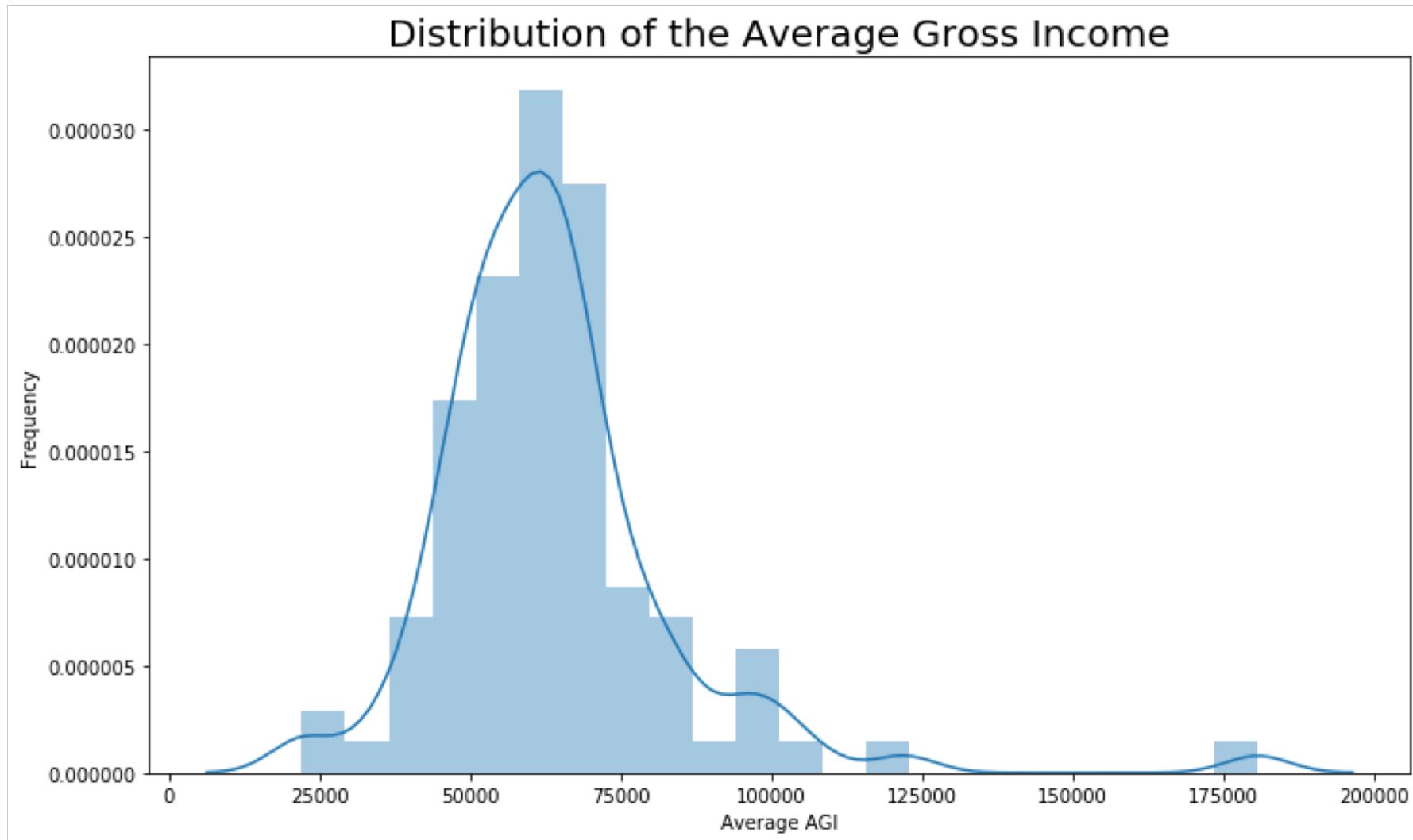


Offers delivery

Start Order

www.yelp.com

The distribution of average AGI in our sample data was similar to the distribution for the U.S.



Our initial focus was on price and ratings.



4. Fromagination



199 reviews

(608) 255-2430

\$\$ • Cheese Shops, Sandwiches,
Cheese Tasting Classes

12 S Carroll St

Capitol

“ a wide array of delicious cheeses from Wisconsin and around the world, and samples are always available. Additionally, they sell other **specialty food** items...” [read more](#)

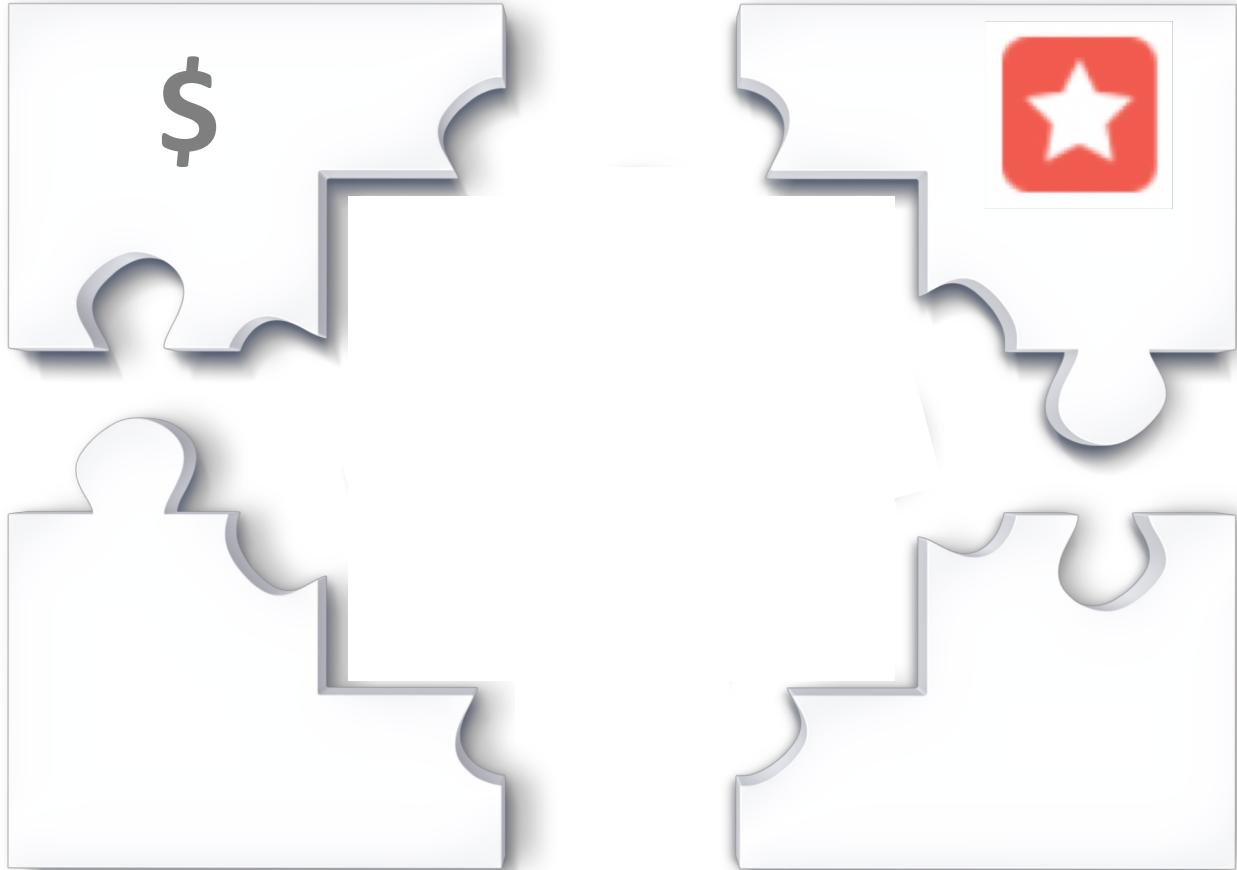


Offers delivery

Start Order

www.yelp.com

Patterns using only price and rating were too weak to connect to average AGI.



Adding the business category showed promise.



Businesses are classified in one or more categories.



4. Fromagination



199 reviews

(608) 255-2430

\$\$ • Cheese Shops, Sandwiches,
Cheese Tasting Classes

12 S Carroll St

Capitol

“ a wide array of delicious cheeses from Wisconsin and around the world, and samples are always available. Additionally, they sell other **specialty food** items...” [read more](#)



Offers delivery

Start Order

www.yelp.com

For our sample, three categories stood out.



“Gourmet”

Grocery Stores

Coffee



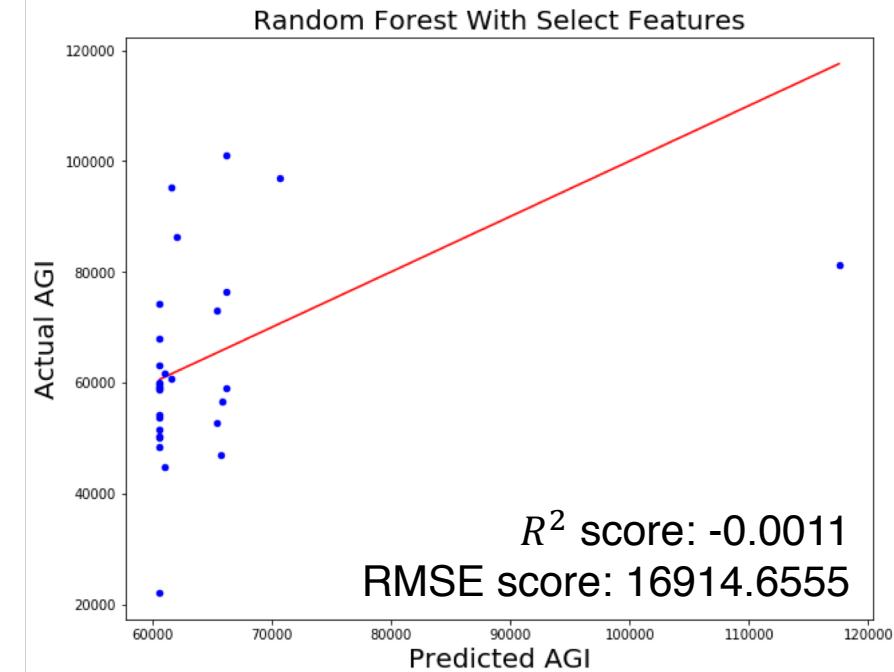
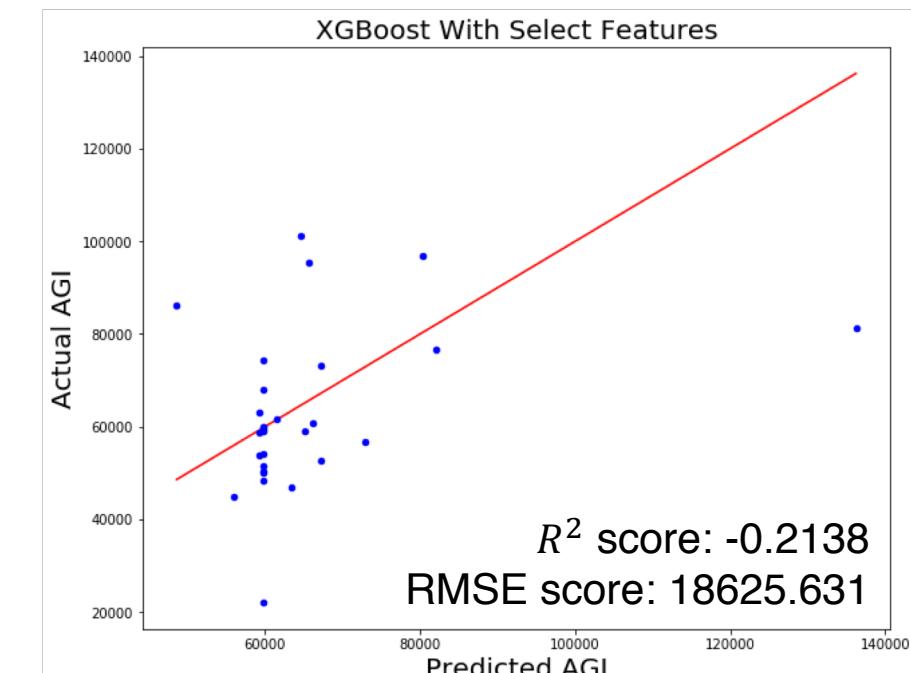
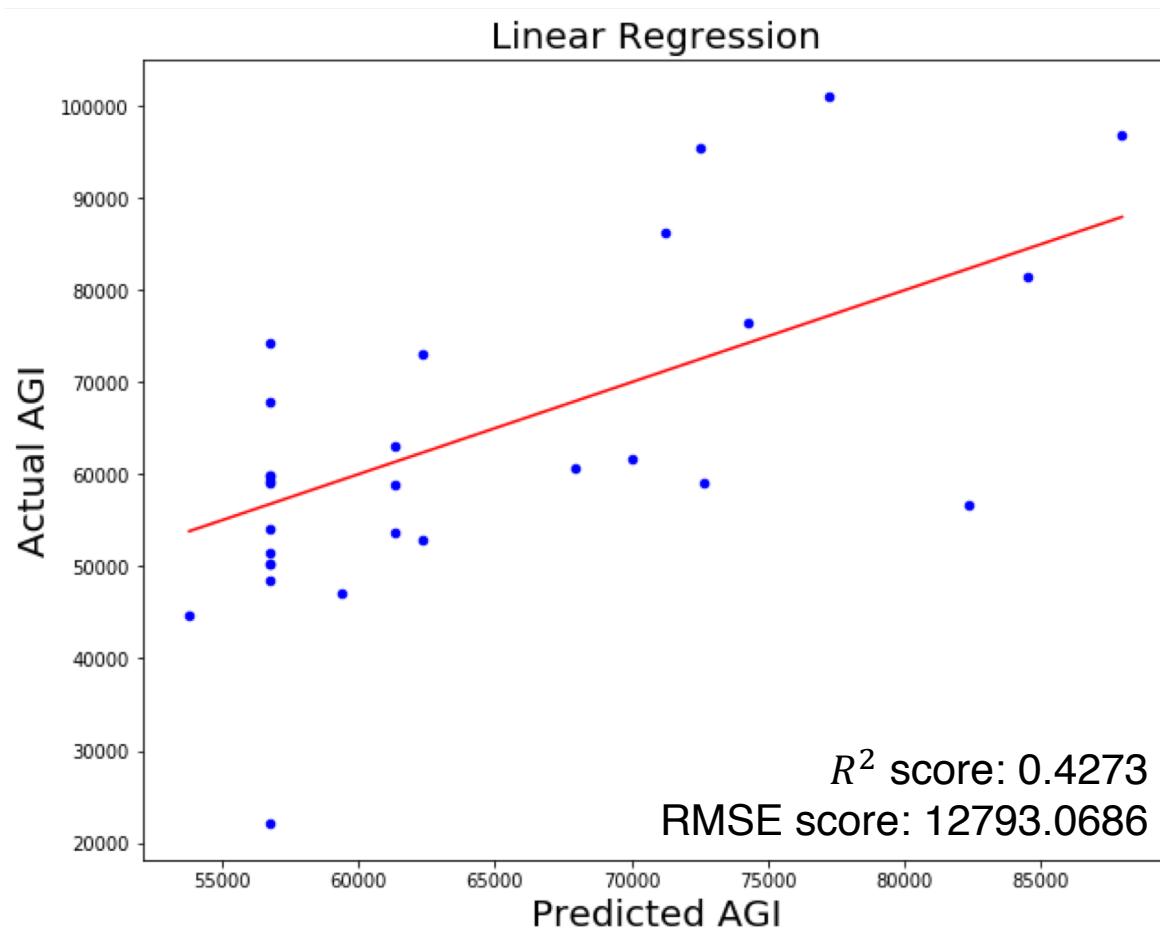
4

Model Data

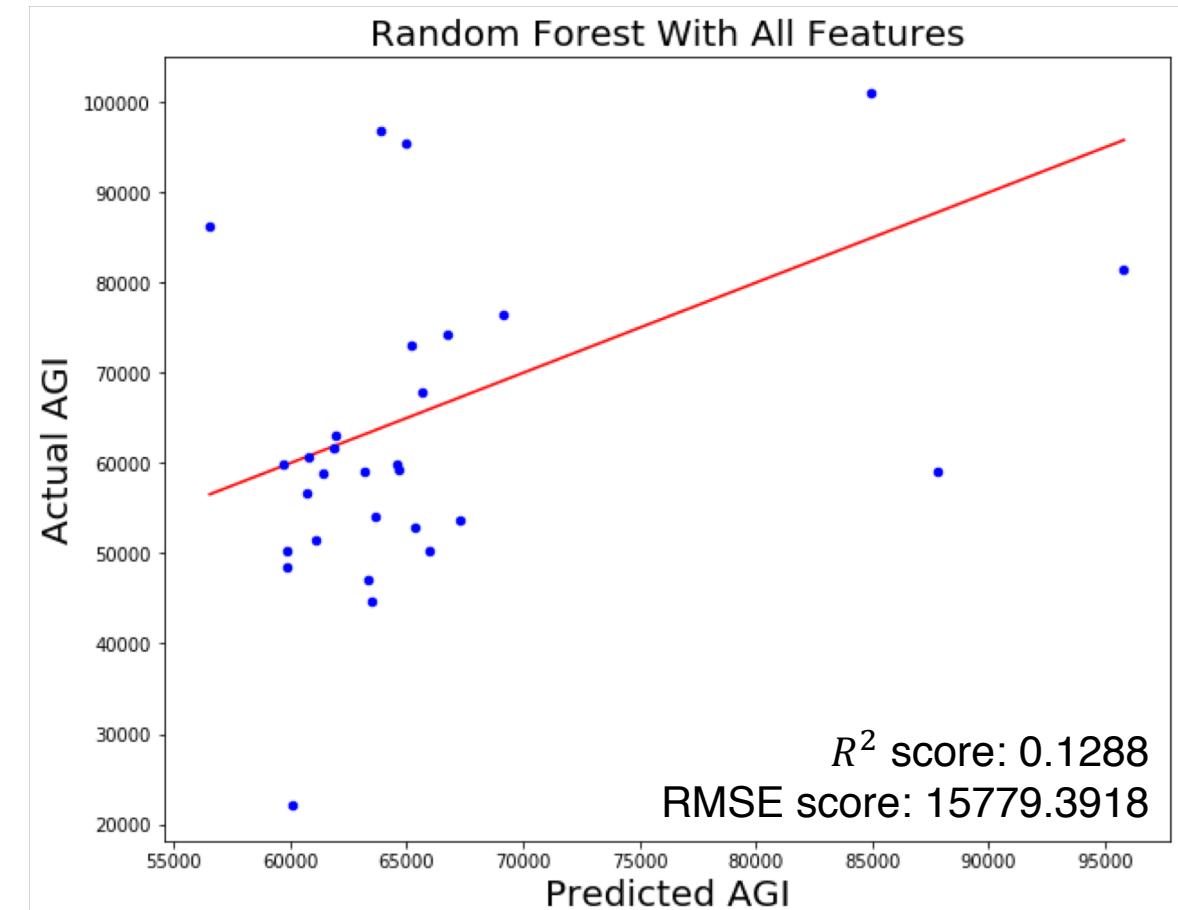
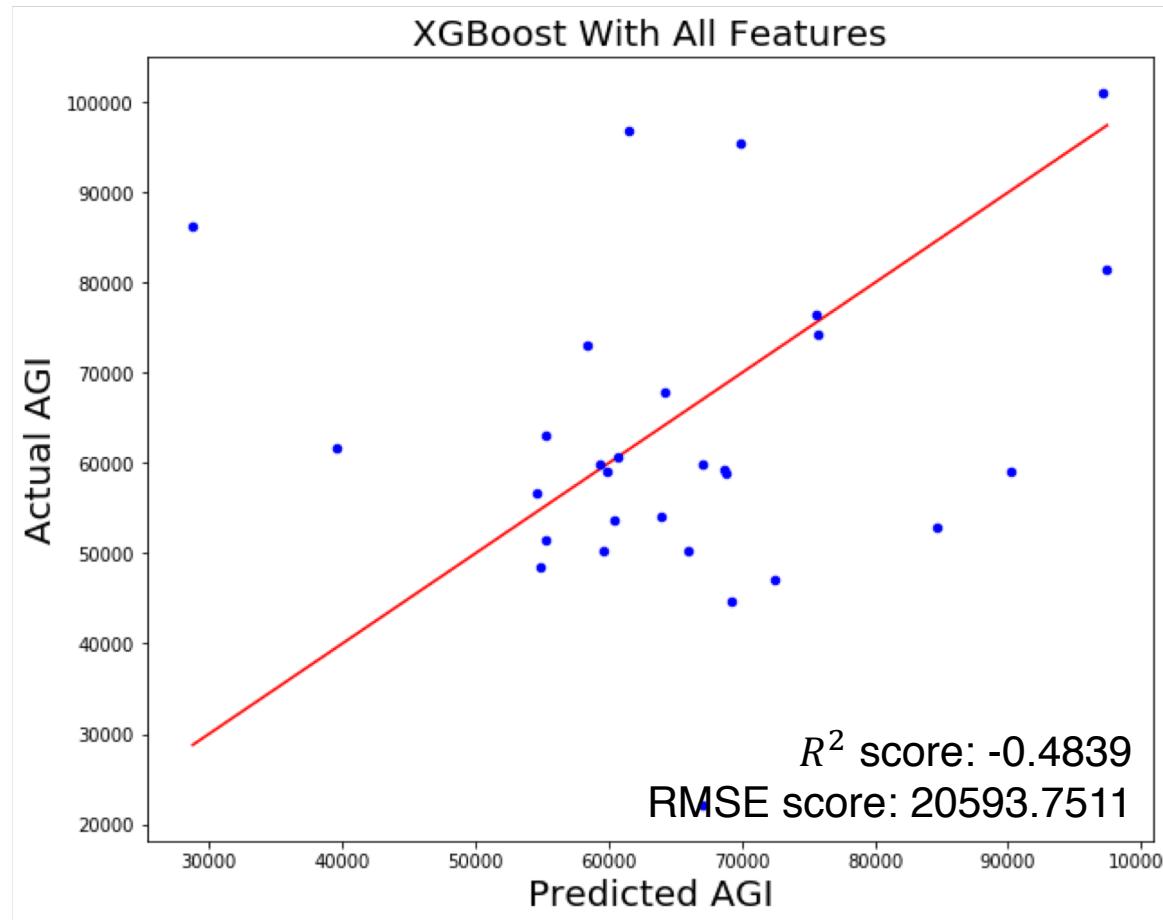
How do supervised and unsupervised models make predictions based on the data?

Supervised Learning

Baseline RMSE: 20425.3965



More Supervised Learning (With all features)





5

Answer Problem

Does the outcomes of
our project address the
problem statement?

Conclusions

- We believe there is **more signal to be seen** as more data is gathered and deeper analysis is done.
- We conclude that Yelp can be used as a **supplementary economic indicator**, one more tool to help determine the wealth of a neighborhood.
- We believe there is **potential for the Yelp data to be used as a measure of economic growth or shrinkage** in an area.



6

Next Steps

Towards a deliverable

Next Steps

- **First, having full access to the Yelp API** is crucial in obtaining the most amount of data. The basic version of the API limits the data that can be retrieved.
- **Second, pulling data from other resources**, such as the Internal Revenue Service or the US Census, will allow a more informed model.
- **Lastly, deploying a functional tool** by automating the whole process would make our work more accessible and instantly useable. Our model can still be refined further.

Thank you!



Shannon
Bingham



Roy Kim

Any questions?