A Clustering-Based Approach Incorporating Census and Amenities Data to Find the Best Neighborhood to Live

Amelia Bell, Michael Bodie, Aravinda Dassanayake, Joseph Janicki, Riesling Meyer, and Nathan Smooha

RELO Relocation Made Easy

Motivation

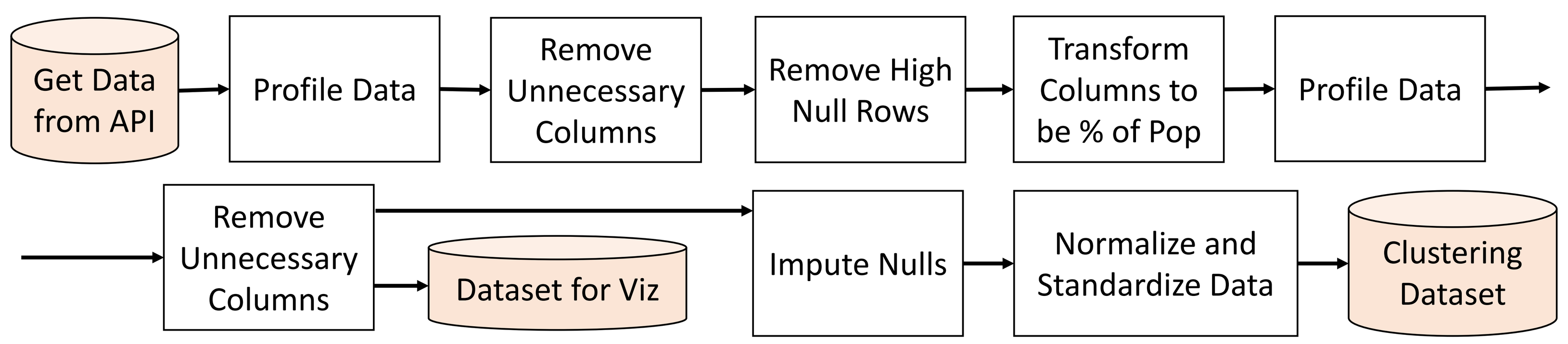
According to Census data, relocation in America is at its lowest rate on record[1]. This is concerning because the lack of people moving for work could negatively impact the economy[2]. Still, trying to find the right place to live can be difficult.

Current tools do not simultaneously consider nuances between neighborhoods and provide personalized suggestions. All tools lack data on essential services (like grocery stores) and interactive visualizations. "Top City" lists are impersonal and too broad[3].

RELO is a tool that will give people the confidence to relocate when they might have been hesitant to before. By using a user's input of an address, **RELO** determines its neighborhood, and presents a list of recommended locations from across the country that are similar to the user's identified neighborhood based on a clustering of Census and amenities data.

Census Data

We used 2018 Census data downloaded from data.census.gov



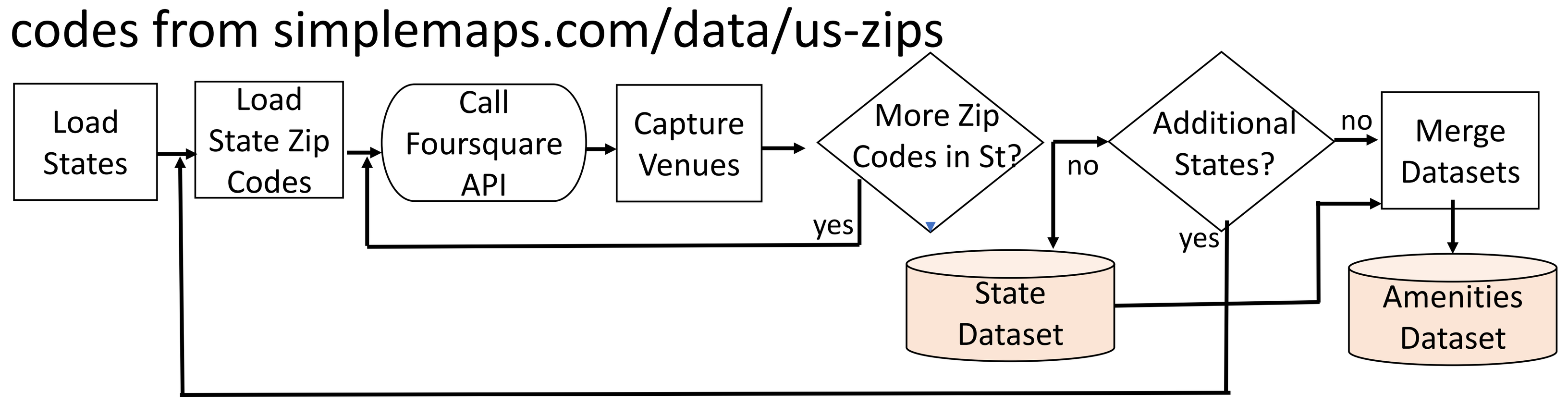
Original Census Dataset	
# Columns	1,316
# Rows	74,001
Size on Disk	1,200 MB

Cleaned Census Dataset	
# Columns	28
# Rows	73,056
Size on Disk	20.5 MB

The geometries for Census tracts (e.g. shape files) were downloaded from github.com/loganpowell/census-geojson/tree/master/GeoJSON/500k/2018

Amenities Data

We pulled data on grocery stores, parks, hardware stores, gyms, and medical care facilities from Foursquare's Places API, leveraging zip codes from simplemaps.com/data/us-zips

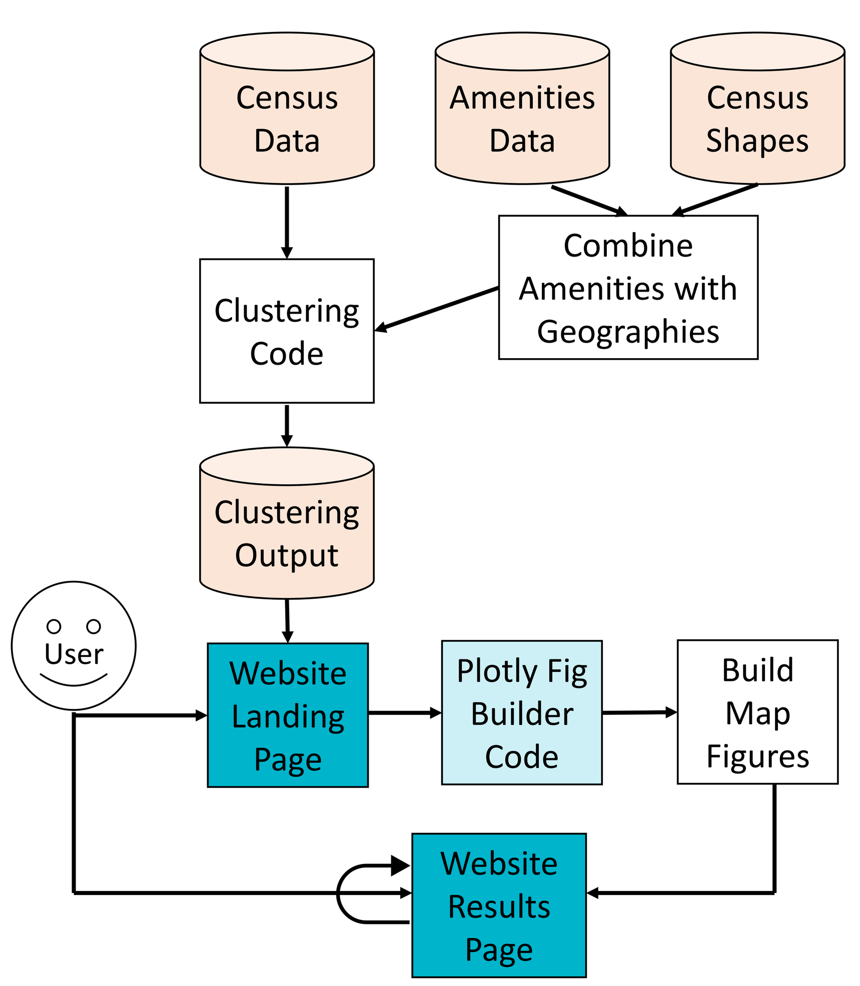


Original Amenities Dataset	
# Columns	8
# Rows	275,641
Size on Disk	32 MB

Cleaned Amenities Dataset	
# Columns	11
# Rows	230,157
Size on Disk	30.7 MB

We identified the amenities within each Census tract, as well as those within 2, 5, 10, 25, and 50 miles of the tract center. The amenities data was weighted and joined to the Census data. For each amenity, the count of entities within 25 miles were features in our clustering dataset, which was based on the data distributions.

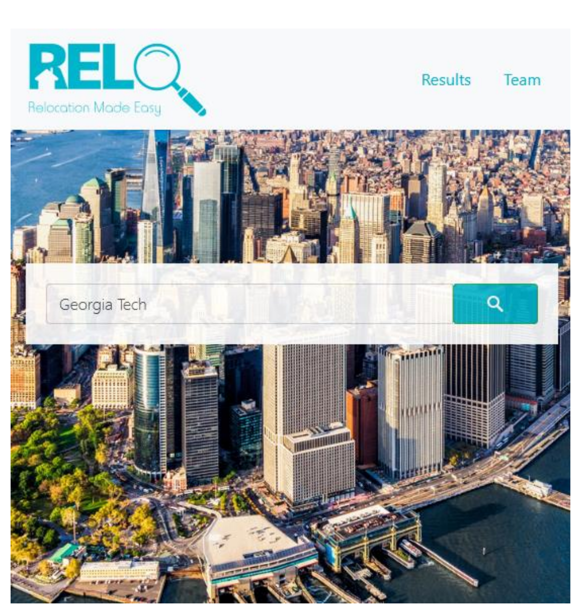
Approach



RELO is an interactive web application. Users input an address of a location they like, and our app finds the closest Census tract center. Then, RELO showcases all Census tracts (i.e., neighborhood proxies) that are similar to users' inputs, and it displays the 3 closest by Euclidian distance in the feature space. Similar neighborhoods belong to the same cluster, generated by a KMeans model with a parameter of 50 clusters. Users can click any featured location to zoom in to see amenities in the area.

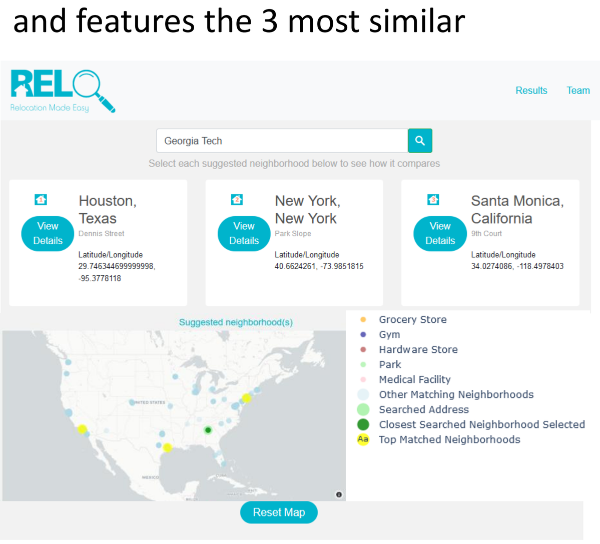
This new approach effectively solves our problem by providing users recommendations based on their input. Users can view suggestions across the US and can click to see more information about specific locations. This gives users confidence in choosing a new neighborhood.

Welcome Search Screen



Search Results

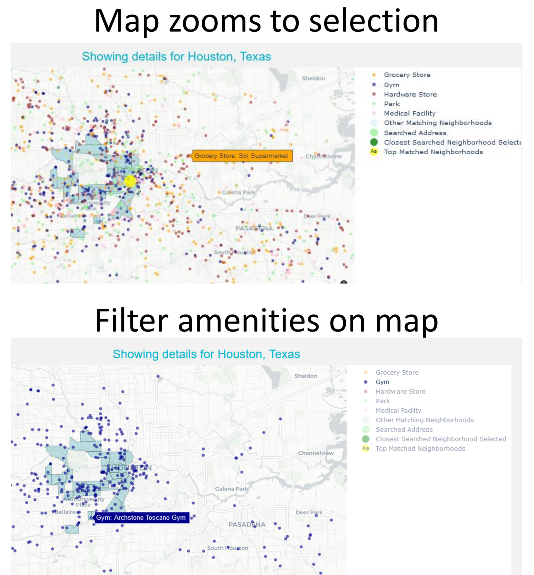
Highlights all neighborhoods in cluster and features the 3 most similar



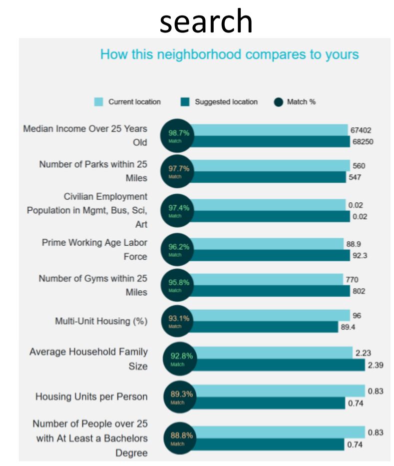
Interactivity (Click on Recommendation)

Map zooms to selection

Filter amenities on map



See how selected recommendation compares to original search

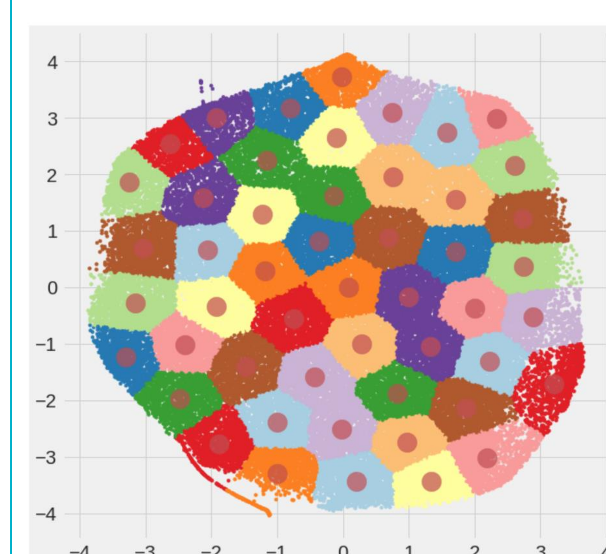


Experiments and Results

Is the data clusterable? We ran a statistical test that used the Hopkins statistic, which measures the probability that a dataset was generated by a uniform distribution[4]. With a threshold of 0.5, the Hopkins statistic returned a result of 0.008, so we rejected the null hypothesis in favor of the alternative that our data contained clusters and was not uniformly distributed.

How many clusters? We balanced choosing the number of clusters based on the relevant internal validity metrics for various numbers of clusters (Fig. 1), and load times of our visualization tool (Fig. 2). Ultimately, we selected 50 clusters for all models in order to improve the user experience of our tool.

Which algorithm to use? We compared the Silhouette, Calinski-Harabasz, and Davies-Bouldin internal validity measures to determine the best clustering algorithm for our dataset with 50 clusters (Fig. 3). Based on the results, we chose KMeans to be our final clustering algorithm for our tool.



2D Visualization of clustering output and cluster centers of final chosen model

What are the features that make each cluster unique? We calculated the variance of means between clusters within each feature and determined the top ones. These were the amenities within 25 miles (std devs ~.2), percent of population that is white (single race) (std dev .200), and % of housing that is multi-unit (std dev .185)

How does this compare to other methods? This clustering and ranking approach represents a first step in providing personalized recommendations on places to live, which is not a feature offered by other tools.

Figure 1

	Silhouette	Calinski-Harabasz	Davies-Bouldin	"Elbow Method"	Gap Statistic
KMeans	3 (.40)	50 (72462.59)	15 or 39 (0.77)	10 (51231.32)	34 (0.47)
Mini-Batch KMeans	3 (.39)	50 (69040.65)	7 (0.78)	10 (55459.62)	6 (0.35)
Spectral	48 (.322)	50 (64344.15)	2 (0.61)		
Ward	2 (.35)	50 (58761.53)	4 (0.87)		
BIRCH	3 (.35)	50 (57272.38)	4 (0.86)		

Figure 2

	15 clusters ~5,000 members per cluster		50 clusters ~1,400 members per cluster	
Sample	Time (in sec.) from app start to server ready	Time (in sec.) from search of "Atlanta, GA" to results page loaded.	Time (in sec.) from app start to server ready	Time (in sec.) from search of "Atlanta, GA" to results page loaded.
1	12.74	31.54	10.94	7.19
2	12.77	31.6	10.74	6.97
3	12.64	29.84	10.38	6.27
4	12.79	33.32	10.5	6.98
5	12.65	29.99	10.39	7.28
Mean	12.7	31.3	10.6	6.9
SD	0.1	1.4	0.2	0.4

Figure 3

	Silhouette	Calinski-Harabasz	Davies-Bouldin
KMeans	0.34	72462.59	0.79
Mini-Batch KMeans	0.33	69040.65	0.83
Spectral	0.32	64344.15	0.76
Ward	0.27	58761.53	0.91
BIRCH	0.26	57272.38	0.93

[1] Tavernise, S. (2019, November 20). Frozen in Place: Americans Are Moving at the Lowest Rate on Record. *The New York Times*. https://www.nytimes.com/2019/11/20/us/american-workers-moving-states-.html
[2] Kopf, D. (2018, November 30). Americans are moving less than ever, and it's bad for the economy. *Quartz*. https://qz.com/1480835/the-share-of-americans-moving-hit-a-record-low/
[3] Okulicz-Kozaryn, A. (2011). City Life: Rankings (Livability) Versus Perceptions (Satisfaction). *Social Indicators Research, 110*, 433-451. https://doi.org/10.1007/s11205-011-9939-x
[4] Lawson, R., & Jurs, P. (1990). New index for clustering tendency and its application to chemical problems. *Journal of Chemical Information and Computer Sciences, 30(1),* 36-41. https://doi.org/10.1021/ci00065a010