Battle of the Neighborhoods

Finding a new home with data

Brian Bloom

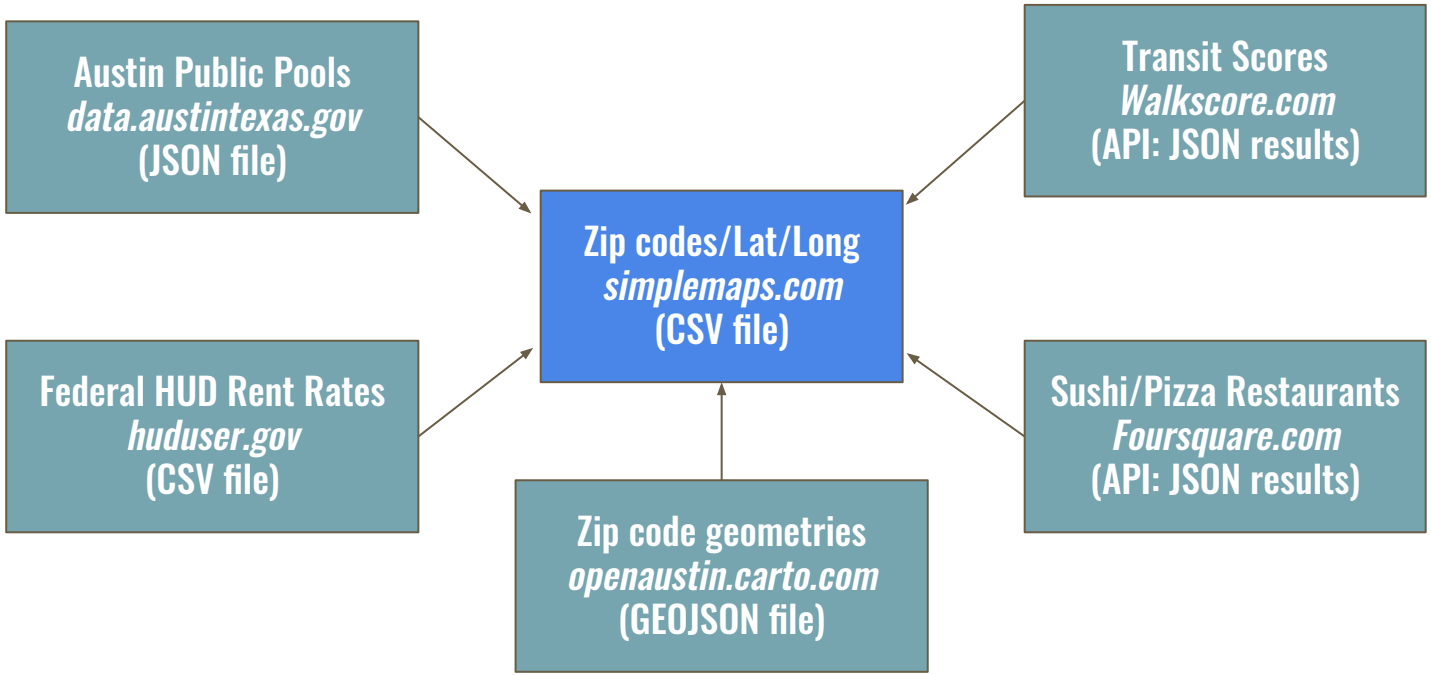
The Problem to solve

- Just returned from 3 years in Europe
- Need to find a home for my family in Austin, Texas area that satisfies:
  - Needs to be in our budget (max rent *$1800/mo*)
  - Daughter is a swimmer and needs a nearby pool
  - Wife relies on mass transit so needs good local public transit options
  - We all enjoy sushi and pizza and would like some nearby options

Plan:

Use newly-acquired data science skills to find ideal neighborhoods using available data and Python, so can give list of top zip codes to a realtor for housing search.

Data Sources



Data Merging Process

1. Filter zip code file to just Austin, TX and remove unneeded columns
2. Merge in 3 bedroom rents per zip code, make normalized rent score and “isAffordable” flag for places under my budget.
3. Determine closest public pool per zip code, make normalized pool score and “nearPool” flag for pools < 10 miles away.
4. Skipping !isAffordable and !nearPool zips, query transit score API per zip code, make normalized transit score.
5. Skipping !isAffordable and !nearPool zips, query API for number of sushi/pizza places within 5km of each zip code, make normalized harmonic mean score for both types.

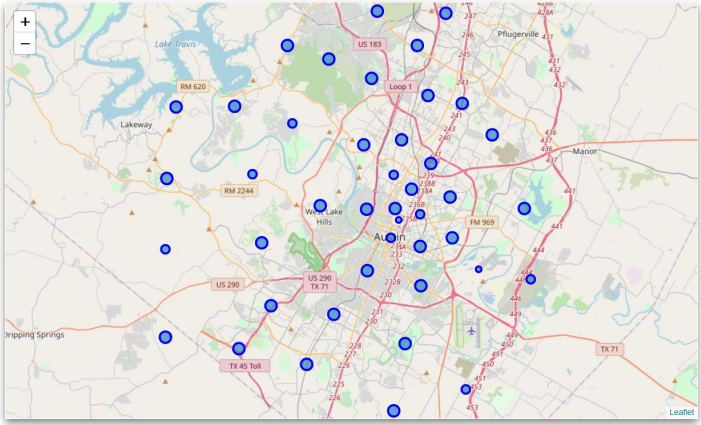
Normalized scores

Each variable's normalized score was computed using a linear scaling formula for that zip code relative to the range of values found in that dataframe column (*e.g.*: *rent, distance, transit score, and # of restuarants*):

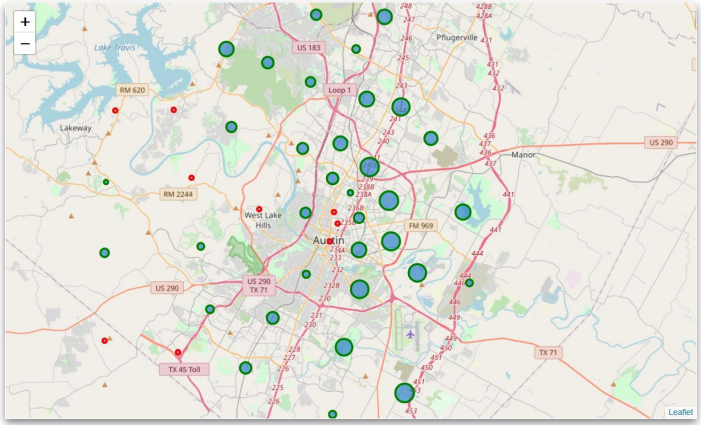
$$\textbf{normscore}_{i} = (\text{Value}_{i} - \text{Value}_{min}) \ / \ (\text{Value}_{max} - \text{Value}_{min})$$

Visualizations of Interim Results

Starting zip codes
(disc size = population)

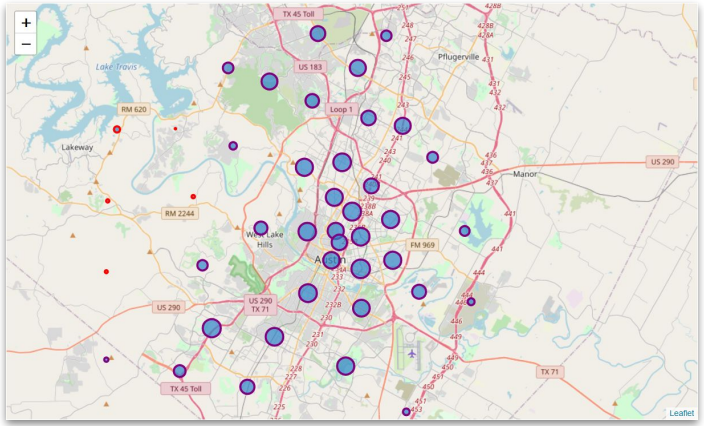


Affordability of each zip code
(large disc = affordable, red = $$$)

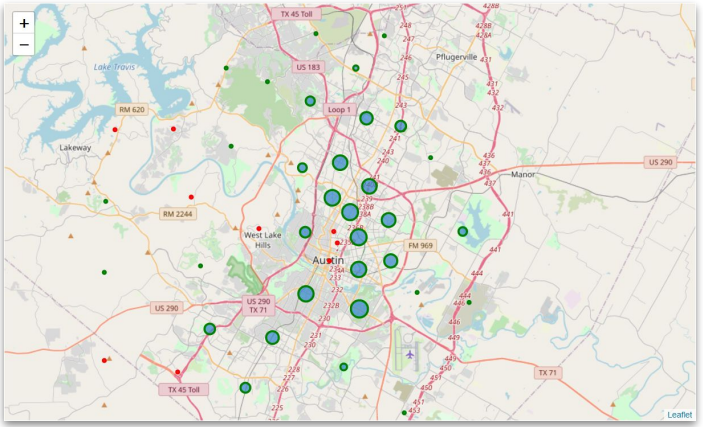


Visualizations continued

Proximity to nearest public pool
(larger disc = close, red = too far)



Availability of public transit
(larger disc = good, red = too $$$)



Calculating an overall score per zip code

For each of the zip codes, generate a weighted average calculated as:

$$\frac{\mathbf{3} \times \text{Rentscore} + \mathbf{2} \times \text{Poolscore} + \mathbf{2} \times \text{Transitscore} + \mathbf{1} \times \text{Foodscore}}{\mathbf{8}}$$

This places more emphasis on rent, which is a limiting factor, while “diluting” the relevance of nearby restaurants, which are just a “nice to have”.

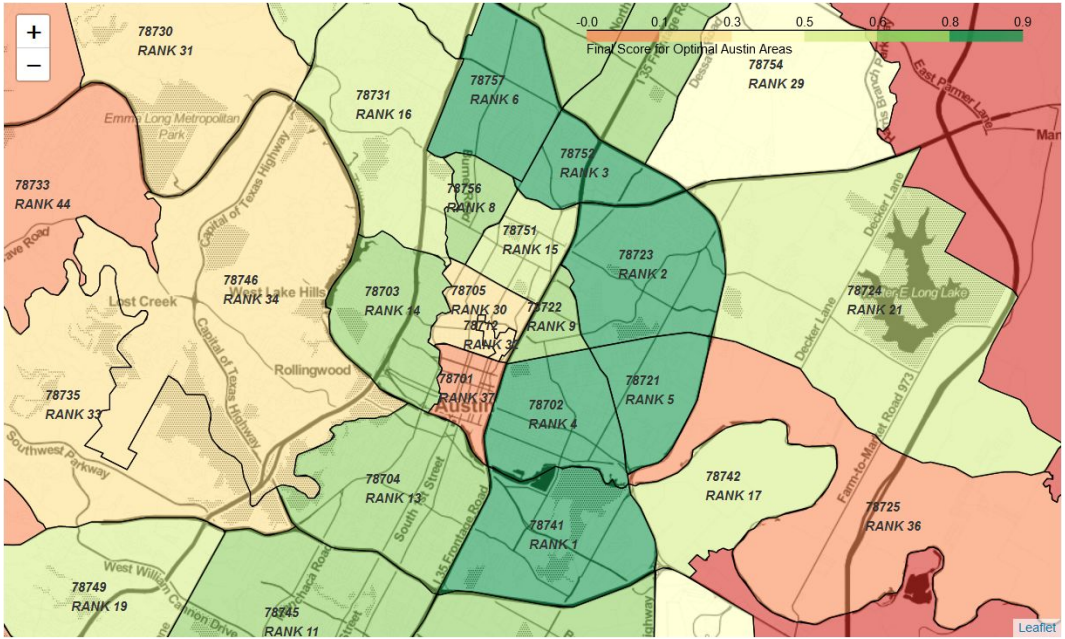
And the winner is...

Outcome of calculating the composite scores and ranking the results:

	zip	city	rent3br	rentscore	nearestPool	poolscore	transitscore	foodscore	overallscore	finalrank
27	78741	Austin	1640	0.931507	Montopolis	0.866505	1.000000	0.882824	0.926294	1
10	78723	Austin	1590	1.000000	Bartholomew	0.898224	0.766667	0.712105	0.880236	2
37	78752	Austin	1600	0.986301	Brentwood	0.737103	0.816667	0.908961	0.871926	3
1	78702	Austin	1790	0.726027	Parque Zaragoza	1.000000	0.866667	0.968812	0.860028	4
8	78721	Austin	1620	0.958904	Givens	0.898096	0.716667	0.549722	0.831995	5
41	78757	Austin	1830	0.671233	Northwest	0.936935	0.816667	0.956283	0.809648	6
38	78753	Austin	1660	0.904110	Walnut Creek	0.818021	0.550000	0.666122	0.764312	7
40	78756	Austin	1950	0.506849	Ramsey	0.891479	0.883333	0.968812	0.754873	8
9	78722	Austin	2020	0.410959	Patterson	0.940369	0.916667	0.967125	0.739259	9
42	78758	Austin	1780	0.739726	Walnut Creek	0.717830	0.666667	0.901141	0.736164	10

Heatmap of good and bad areas *for us*

Green = areas that best fit our needs based on composite score



Observations

Using public or API data for zip codes, pools, rents, transit scores, and restaurants, I was able to determine optimal starting places in Austin for a place to live for my family.

Those results point to East and North Austin, particularly in zip codes 78741, 78723, 78752, 78702, and 78721.

These calculations are preserved in a Jupyter notebook so that if I want to add new variables (*like crime or traffic*), or update with newer data, I can continue with the current data and rerun the code.

Conclusion

Full report for project is available at:
https://github.com/brianbloom/Coursera_Capstone/raw/master/The%20Battle%20of%20Neighborhoods%20report.pdf

Or it can be downloaded with this QR code →

Jupyter notebook with code and data is available at:

https://github.com/brianbloom/Coursera_Capstone/blob/master/Brians_Capstone_Final.ipynb