CSE 6242 Fall 2022 – Team 163 *(bbeattie3, mmaleki3, nsengupta7, pdongare6, pkalan3, yfu345)*
Final Report: A Recommendation Tool for Real-Estate Neighborhood Search

[5%] INTRODUCTION

Background

Accurate information is essential to good decision-making. This is especially true when relocating or investing in real estate. We propose a recommendation tool to aid in selecting a neighborhood. When relocating to a new city, what tools are available to aid in selecting a neighborhood that suits personal preferences? How can one sort through the various attributes that could impact that decision and create a rational analysis in an unfamiliar place? This question and our project could potentially affect anyone involved in searching for a new place to live, investing in real estate, or researching these topics.

From a buyer's standpoint there are several other critical factors that influence a buying choice depending on their own needs. For example, a) a new family would look for the neighborhood demographics, school rating and crime rate b) a fresh college graduate would look for the commute time whereas c) a senior couple might look for a locale close to hospitals and other facilities. For the purposes of this project, our scope is the entire country of the United States of America.

Currently available commercial tools (Zillow, Realtor, Trulia, Apartments.com) focus on available inventory and provide the user with some selection of limited attributes closely related to those properties, but fail to consider these other essential decision attributes. Academic studies typically offer a static visualization of a single attribute for a single locale in detail. A single tool that allows these parameters to be dynamically visualized together based on user-selected priorities is not yet available.

Our Innovative Approach

We propose to create a novel tool offering a user three benefits: it allows a user to select multiple attributes, prioritize them, then dynamically visualize their tailored results on a map. We will create indices for each of the decision attributes then visualize the resulting weighted index based on the user's preferences.

[5%] LITERATURE SURVEY

Resident Demographics: Cellmer, Cichulska and Bełej[1] highlight the correlation of different socioeconomic factors based on the mixed geographic regression methods. Reed[2] also confirms that the demographic profile of a household is related to their house prices using principal components analysis and regression analysis models. These two papers highlighted the close relationship between house values, income, and age. In Renwick[3] the representatives from U.S. Census Bureau, Bureau of Labor Statistics, U.S. Department of Health and Human Services et.al. used the American Community Survey to create the housing price index. We likewise reference this data set in our work.
School Ratings: The quality of a public school in Ohio is rated by Haurin and Brasington[4] using specific test scores for a particular age group. As this data is not consistently available across the United States, we follow Hasan and Kumar[5] referencing GreatSchools.org school ratings (GS) to get school ratings based on zip codes for all states. GS computes ratings based on standardized test scores but also considers student progress, college readiness, and equity.
Access to Grocery: Lee and Lim[6] suggest that we consider both proximity and the supply and demand of each geographic region to determine which areas have adequate grocery supply and which are underserved. Widener[7] reminds us that a simple distance metric is not an adequate measure – that true accessibility is determined by both distance and the mode of transportation that one can afford.

Crime: Tita, Petras and Greenbaum[8] indicate that an appropriate crime index must consider the property value and the crime rate – that the crime rate alone will provide disproportionate results relative to property value.
Healthcare: Rivas et al[9] finds a positive and strong correlation between proximity to hospitals and rent costs and housing prices. Van der Zwart, van der Voordt and de Jonge[10] confirm these results in the European market and on private investment in hospitals. Additionally, Boussabaine, Sliteen and Catarina[11] confirmed a positive effect of the usage and availability of healthcare centers on housing and other expenses in a neighborhood.
***Diversity**:* Maly[12] defines a racial diversity index, and Farrell and Lee[13] alternatively use an Entropy Index specifically noting the importance of the change and the rate of change in addition to the multi-group makeup at a specific point in time. We use the simpler notion of the diversity index.
Parks and Green Space: De Bruyne and Van Hove[14] references a satisfaction index involving proximity and access to parks and green space and highlights the effect of the travel time to these amenities on housing prices. Williams et al[15] explicitly emphasize the need to distinguish "safe" green spaces with low incidence of crime.
Sustainability: Cloutier, Jambeck and Scott[16] suggest an index to measure a municipality's or community's environmental sustainability by combining various key sustainability efforts. Ultimately we find that this index is impractical to reproduce as it requires multiple disparate data sources that would require significant effort to gather and aggregate.

Summary of Key Attributes: From the work above, we identify data sources for the key attributes that we will use in our solution: demographics (age, population, education, income), racial diversity, crime, grocery access, parks and green space, and school ratings. We select data sources that use U.S. Census tract GEOIDs as the common location identifier.

[45%] METHODS

With the data sources identified, we downloaded the data using API or scraped the website. The data was cleaned in OpenRefine to resolve issues like missing values and data type conversions. Since the granularity of each dataset was different, PySpark and SQL were used to integrate all the datasets with the common location identifier.

DATA

Data Attribute	Data Source	Number of Records (Rows x Columns)		Year
Population	Census	242341	12	2020
Sex	Census			
Race	Census			
Home price	Census			
Crime	FBI	3137	24	2021
Grocery	census www.openicpsr.org	1036463	31	2017
Parks & Greenspace	Census	73058	11	2018

School	Census (unified school districts wholly or partially within each county)	16394	4	2020
	Niche www.niche.com (rating for each unified school district)	13822	32	2022
	combined census and niche data (average school district rating for each county)	3067	4	2020 & 2022

Table 1- Dataset sources and record count

Note – All of our data sources were free for public use, so only a few sources are most up to date. For a more commercial project, we could have purchased the latest datasets to run this analysis.

ALGORITHM

Using our pre-processed data containing our feature metrics for each GEOID, we implement our recommendation algorithm.

Feature value calculation and normalization: Individual values are calculated for each feature based on the desired target value. The target value is taken as an input from the user (e.g. median home price of $300k to $500k) or predefined (crime rate should be minimum). The values are normalized to arrive at a scaled index in the range [0,1]:

#	Features	Desired Target	Calculated value (normalized) I_{Attr}
1	Sex Ratio	User selects upper limit and lower limit from the available range shown on a slider	if ($I_{lower\ limit} < I < I_{upper\ limit}$) then 1 else 0
2	Median Age		
3	Population		
4	Median Home Prices		
5	Income Per Capita		
6	Racial Diversity	Racial diversity is predefined to be maximum from the available range. So it is calculated as an inverse of the standard deviation of the population of four major races.	$I_{diversity} = \frac{1}{\sigma(I_{white}, I_{black}, I_{asian}, I_{hispanic})}$ $Normalized\ value = \frac{I_{diversity}}{(I_{diversity})_{max}}$
7	Grocery availability	Predefined as maximum from the available range	$\frac{I - I_{min}}{I_{max} - I_{min}}$
8	Parks availability		
9	School Rating		
10	Crime Rate	Predefined as minimum from the available range	$1 - \frac{I - I_{min}}{I_{max} - I_{min}}$

User Weights: The user can select a preferred categorical weight of "Low", "Medium" or "High" for each feature. We define "Low" = 0.0, "Medium" = 0.5, "High" = 1.0

Individual user weight values are normalized so that all weights sum to 1:

$$\omega_{User\,n} = \frac{\omega_{User\ selected\ n}}{\sum_{i=1}^{n} \omega_{User\ selected\ n}}$$

User Weighted Average: Our final recommendation index, I_{rec} displayed in the visualization is the weighted average of the individual feature indices,

$$I_{rec} = \omega_{User1} I_{Attr1} + \omega_{User2} I_{Attr2} + \ldots + \omega_{User\,n} I_{Attr\,n},$$

where each $I_{Attr\,n}$ is the unity-based normalized feature index and each $\omega_{User\,n}$ is the user specified weight for that feature (where $\sum_{i=1}^{n} \omega_{User\,n} = 1$).

USER INTERFACE

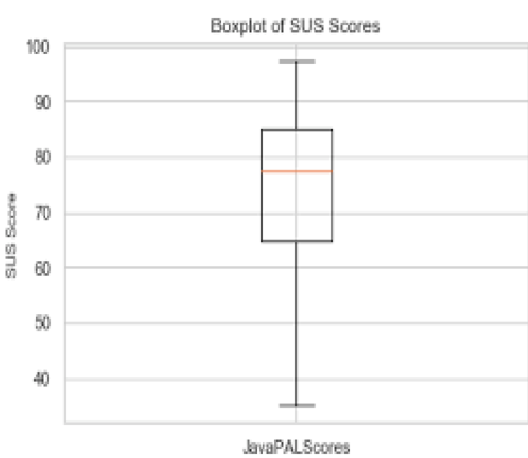
We utilize Tableau to implement our user interface and interactive visualization. We plot a choropleth map using the GEOID references to reference the geometric boundaries for the census tract areas and the recommendation index I_{rec} as the resulting value for that area. The tool is hosted on neighborhoodsearch.net to share with potential users and receive their feedback.

[30%] EXPERIMENTS / EVALUATION

METRICS

Based on our research we identified two criteria that yield measurable results: *user satisfaction* and *time to reach a decision*. We evaluated these criteria as illustrated below -

User satisfaction. We utilize Brooke's standardized usability assessment (SUS)[17] to provide an empirical measure of usability. At scale, we would want to recruit a large sample size and utilize multiple evaluations, but given the limitations of time we recruited about 25 family and friends. It is likely that our respondents are not necessarily our target users since most are not actively in the process of searching for a new neighborhood.



We found that in our sample group, the SUS testing yields scores (mean SUS = 65) that indicate *"marginal" acceptability* based on the adjective ratings summarized by Bangor, Kortum and Miller.[18]

We created two cohorts of users, one as a control group using existing tools like Zillow and Redfin and our solution was provided as a treatment. We recorded the responses and concluded that the treatment reached higher statistical significance measured by the score of overall likeability.

Time to reach a decision. At scale, we would want to empirically measure "does the solution reduce the time required to make a decision?"

Measuring "time on task" for our project is a complex endeavor. While tasks such as making an airline reservation or picking a movie to watch are relatively short tasks with well-defined completion criteria, exploring neighborhood-level data to identify a preferred place to live does not have such clear boundaries. Because of the complex nature of this task, it may not be completed in a single sitting. Because the selection of a neighborhood is related to ultimately selecting a house or apartment the process is dynamic – constantly changing with market inventory, external constraints (budget, availability), as well as intangible emotional constraints.

Based on our user cohort we assumed that the recruited users are currently searching for a neighborhood and in need of a tool to assist. We provided the users with guidelines around how to conduct their search and under what conditions they should consider an initial search "*complete*". Our success criteria was that the user would atleast be able to identify a county that met their requirements by looking into our dashboard for a period of 14 days. 80% of the users agreed that they were able to reach a decision.

OTHER OBSERVATIONS

We note several key observations after using our solution and watching others use it:

- Most interesting feature attributes
- Surprising visualization results
- Urban/rural differences

An ideal solution might combine our visualization feature with one of the commercial solutions (Zillow, Apartments.com, etc.) such that either the available inventory of properties could be viewed as a layer on top of our visualization, or our visualization could be used as an additional filter used to select target properties in the commercial solution.

[5%] CONCLUSIONS / DISCUSSION

CONCLUSIONS

NOTE: All team members have contributed a similar amount of effort.

Our work has resulted in an interesting exploratory tool that has the potential for further refinement especially on the number of dimensions considered. We are of the opinion that the techniques used demonstrated a measurable improvement in *user satisfaction* and the *time to reach a decision are done* to the best of our efforts and limited time. Even though it's not ready for General Availability(GA) this tool could provide insights to a subset of renters and homebuyers that none of the existing market tools possess.

We operated with a philosophy that more the data insights more would be the informed decision making process. Mellander, Florida and Stolarick[25] indicate that a more informed choice of location can increase the length of stay.

Additionally, our research indicated that there could be a potential unintended negative social impact resulting from self-segregating into areas of similar socio-economic status which can serve to worsen disparities that already exist[19]. Those that cannot afford access to the best schools,[20-22] healthy food options[23], parks, and cultural experiences, are more likely to have neighbors, friends, and family with the same disadvantages. Van Ham et al[24] extend this notion to describe and visualize the history of neighborhoods and how disadvantage is propagated over generations.

NEXT STEPS / AREAS FOR IMPROVEMENT

We identify several areas for continued improvement.

1. **Increased concurrency and a more scalable solution.**
2. **Additional grains of data.**
3. **Page load responsiveness**

Citations

1. Cellmer R, Cichulska A, Bełej M. Spatial analysis of housing prices and market activity with the geographically weighted regression. ISPRS International Journal of Geo-Information. 2020 Jun 9;9(6):380, DOI: 10.3390/ijgi9060380.
2. Reed R. The relationship between house prices and demographic variables: An Australian case study. International Journal of Housing Markets and Analysis. 2016 Oct 3;9(4):520-37, DOI: 10.1108/IJHMA-02-2016-0013.
3. Renwick T. Geographic adjustments of supplemental poverty measure thresholds: Using the American Community Survey five-year data on housing costs. In Conference Paper prepared for Allied Social Science Associations Annual Meeting–Denver, CO 2011 Jan. [link]
4. Haurin DR, Brasington D. The impact of school quality on real house prices: Interjurisdictional effects. Journal of Housing Economics. 1996 Dec;5(4):351-68.
5. Hasan S, Kumar A. Digitization and divergence: Online school ratings and segregation in America. Available at SSRN 3265316. 2019 Jul 23, DOI: 10.2139/ssrn.3265316.
6. Lee G, Lim H. A spatial statistical approach to identifying areas with poor access to grocery foods in the city of Buffalo, New York. Urban Studies. 2009;46(7):1299–315, DOI: 10.1177/0042098009104567
7. Widener MJ. Comparing measures of accessibility to urban supermarkets for transit and auto users. The Professional Geographer. 2017 Jul 3;69(3):362-71, DOI: 10.1080/00330124.2016.1237293.
8. Tita GE, Petras TL, Greenbaum RT. Crime and residential choice: a neighborhood level analysis of the impact of crime on housing prices. Journal of quantitative criminology. 2006 Dec;22(4):299-317, DOI: 10.1007/s10940-006-9013-z.
9. Rivas R, Patil D, Hristidis V, Barr JR, Srinivasan N. The impact of colleges and hospitals to local real estate markets. Journal of Big Data. 2019 Dec;6(1):1-24. DOI: 10.1186/s40537-019-0174-7
10. van der Zwart J, van der Voordt T, de Jonge H. Private investment in hospitals: A comparison of three healthcare systems and possible implications for real estate strategies. HERD: Health Environments Research & Design Journal. 2010 Apr;3(3):70-86. DOI: 10.1177/193758671000300308
11. Boussabaine H, Sliteen S, Catarina O. The impact of hospital bed use on healthcare facilities operational costs: The French perspective. Facilities. 2012 Jan 27. DOI: 10.1108/02632771211194266
12. Maly MT. The neighborhood diversity index: a complementary measure of racial residential settlement. Journal of Urban Affairs. 2000 Mar 1;22(1):37-47, DOI: 10.1111/0735-2166.00038.
13. Farrell CR, Lee BA. Racial diversity and change in metropolitan neighborhoods. Social Science Research. 2011 Jul 1;40(4):1108-23, DOI: 10.1016/j.ssresearch.2011.04.003.
14. De Bruyne K, Van Hove J. Explaining the spatial variation in housing prices: an economic geography approach. Applied Economics. 2013 May 1;45(13):1673-89., DOI: 10.1080/00036846.2011.636021
15. Williams TG, Logan TM, Zuo CT, Liberman KD, Guikema SD. Parks and safety: a comparative study of green space access and inequity in five US cities. Landscape and urban planning. 2020 Sep 1;201:103841, DOI: 10.1016/j.landurbplan.2020.103841

16. Cloutier S, Jambeck J, Scott N. The Sustainable Neighborhoods for Happiness Index (SNHI): A metric for assessing a community's sustainability and potential influence on happiness. Ecological Indicators. 2014 May 1;40:147-52, DOI: 10.1016/j.ecolind.2014.01.012

17. Brooke J. SUS-A quick and dirty usability scale. Usability evaluation in industry. 1996 Jun 11;189(194):4-7.

18. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the system usability scale. Intl. Journal of Human–Computer Interaction. 2008 Jul 29;24(6):574-94, DOI: 10.1080/10447310802205776

19. Charles CZ. The dynamics of racial residential segregation. Annual review of sociology. 2003 Dec 31:167-207, DOI: 10.1146/annurev.soc.29.010202.100002.

20. Morrissey TW, Vinopal KM. Neighborhood poverty and children's academic skills and behavior in early elementary school. Journal of Marriage and Family. 2018 Feb;80(1):182-97, DOI: 10.1111/jomf.12430

21. Roda A, Wells AS. School choice policies and racial segregation: Where white parents' good intentions, anxiety, and privilege collide. American Journal of Education. 2013 Feb 1;119(2):261-93, DOI: 10.1086/668753.

22. Samuels C. Are GreatSchools Ratings Making Segregation Worse? [Internet]. Education Week. Education Week; 2020 [cited 2022Oct9]. Available from: https://www.edweek.org/leadership/are-greatschools-ratings-making-segregation-worse/2019/12

23. Eisenhauer E. In poor health: Supermarket redlining and urban nutrition. GeoJournal. 2001 Feb;53(2):125-33, DOI: 10.1023/A:1015772503007

24. Van Ham M, Hedman L, Manley D, Coulter R, Östh J. Intergenerational transmission of neighbourhood poverty: an analysis of neighbourhood histories of individuals. Transactions of the Institute of British Geographers. 2014 Jul;39(3):402-17, DOI: 10.1111/tran.12040.

25. Mellander C, Florida R, Stolarick K. Here to stay—the effects of community satisfaction on the decision to stay. Spatial Economic Analysis. 2011 Mar 1;6(1):5-24.

APPENDIX

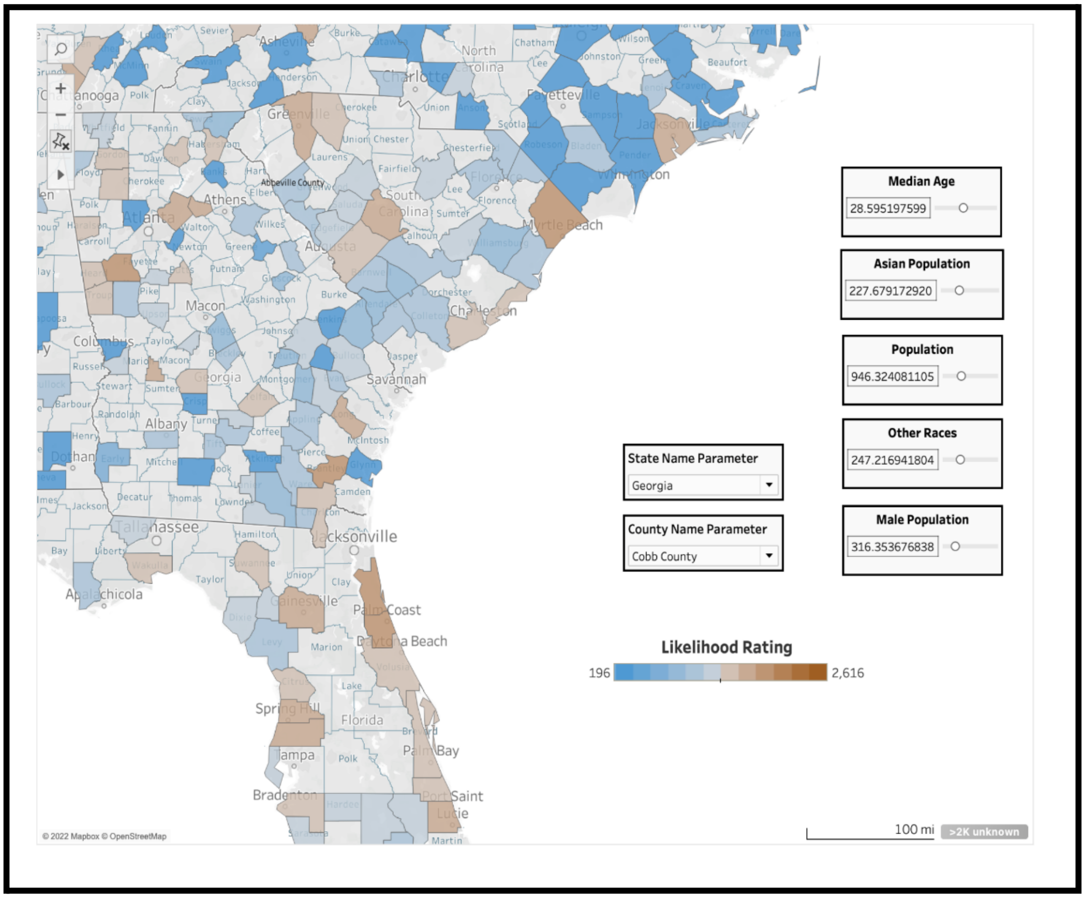


Figure 1. *Screenshot of the User Interface*