WHERE SHOULD I LIVE?

Pairing User Preferences and Forecasting to Narrow Down the Search

Photo credit: https://i.imgur.com/ANLZPH0.jpg

CSE6242 Spring 2019 – Team 47: Ajay Rana, Jack McPherson, Matthew Morton, Sai Balagi Thalanayar Swaminathan, William Jackson

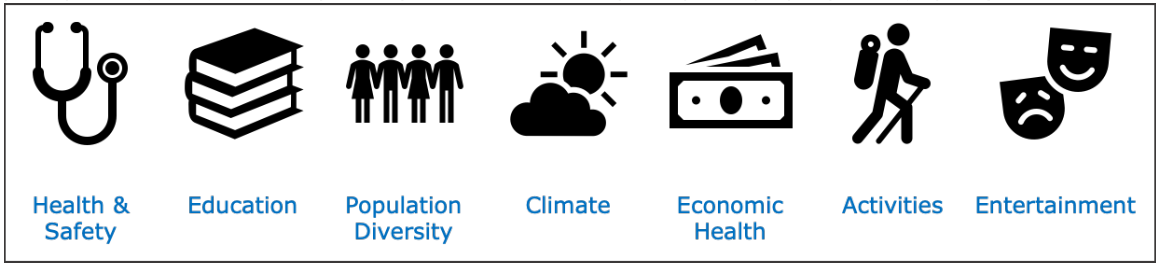
TABLE OF CONTENTS

INTRODUCTION

Moving to a new place is a prospect many will face at least once if not several times during their lives. These changes can represent long term emotional and financial investment risk. Choosing well could lead to happy times and investment growth. But, choosing poorly could lead to despair and financial ruin.

There are many tools (e.g. Zillow, Redfin, RealEstate.com) with helpful interfaces to search for places to live using multiple factors. Where these websites sometimes fail is only focusing on the for-sale-property attributes without considering other quality features important to the buyer. Features such as crime, congestion, and educational opportunities are commonly overlooked.

(Example categorical factors buyers may want to consider beyond specific property demographics)



In our ever increasing Big Data world, consumers expect customizable and specific information to aid their major life decisions, especially when the consequences and risks can persist for many years.

PROBLEM DEFINITION

Property search tools typically focus on housing or vague community characteristics. There is not a convenient solution that provides the following capabilities:

1. Buyer factor selection, priority, and weighting for other quality features important to them when choosing where to live

2. Multi-year factor trend projections to aid this decision that carries long lasting risks and consequences

Our project targeted these gaps to improve the customer experience.

SURVEY SUMMARY

Our research focused on human tendencies in decision making, leading factors to consider when choosing where to live, and analytics suited to forecast these factor trends.

Following are key survey insights that shaped our project design and approach:

- It is difficult to create a subset of factors to meet collective needs of all individuals
- Having too many factor choices becomes overwhelming for the user and can be a deterrent to using an information interface
- Machine Learning, time series, and collaborative filtering using cosine similarity techniques appear to be good options to trend key factors and group qualifying values to the user preferences

Further survey detail can be found in the reference section.

METHOD AND INNOVATIONS

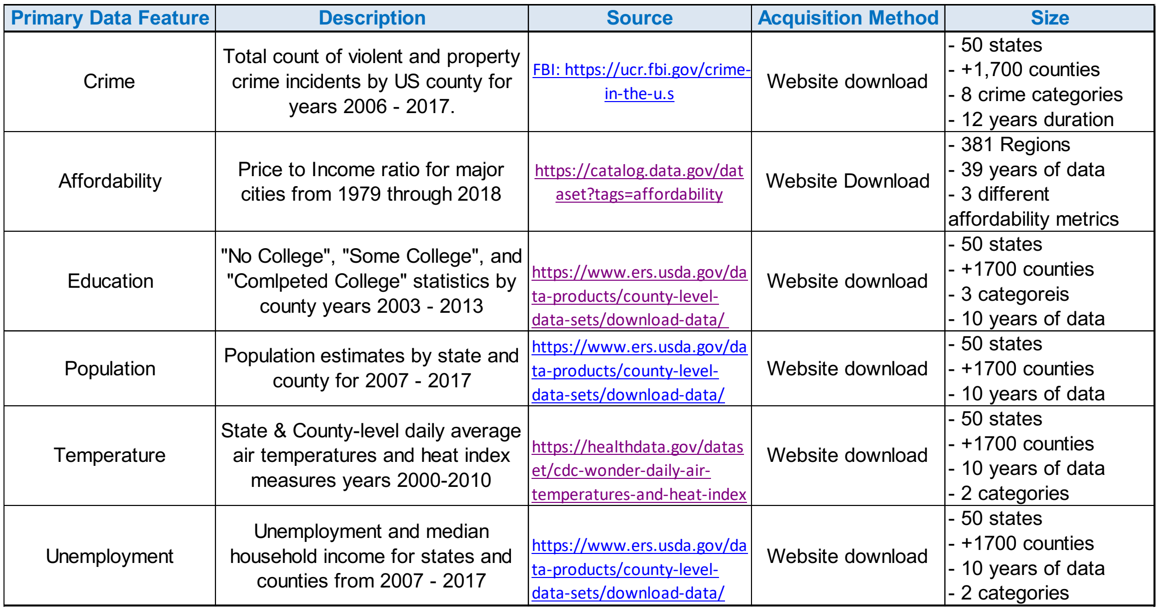
The project deliverable is a tool that takes user selected factors and weights then returns a visual representation of how US counties match their preferences. The user may then drill down and hover over each county for further details on factor values.

Our approach employs SQL, Tableau, Python, Microsoft Azure and introduces two innovations not available in current state-of-the-art solutions:

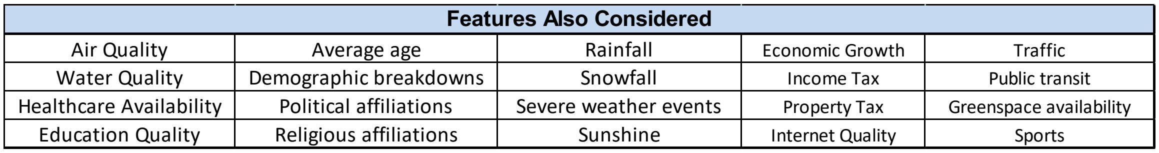
1. Factor weighting- user manages emphasis on each factor importance ranging from low to high.
2. Future year projection- user may view results for any year between current and five years ahead where results may vary depending on forecasted factor values.

This approach is expected to better customize the information relevant to the user and support better decision making on where they should live. The following sections provide further detail on the tool data, algorithm, and user interface.

Data: Over 25 features were researched and considered. Data availability left 10 reasonable possibilities. After thorough research, 6 features made our final product. Final and alternate features considered are listed below:



Primary Data Feature	Description	Source	Acquisition Method	Size
Crime	Total count of violent and property crime incidents by US county for years 2006 - 2017.	FBI: https://ucr.fbi.gov/crime-in-the-u.s	Website download	- 50 states - +1,700 counties - 8 crime categories - 12 years duration
Affordability	Price to Income ratio for major cities from 1979 through 2018	https://catalog.data.gov/dataset?tags=affordability	Website Download	- 381 Regions - 39 years of data - 3 different affordability metrics
Education	"No College", "Some College", and "Comlpeted College" statistics by county years 2003 - 2013	https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/	Website download	- 50 states - +1700 counties - 3 categoreis - 10 years of data
Population	Population estimates by state and county for 2007 - 2017	https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/	Website download	- 50 states - +1700 counties - 10 years of data
Temperature	State & County-level daily average air temperatures and heat index measures years 2000-2010	https://healthdata.gov/dataset/cdc-wonder-daily-air-temperatures-and-heat-index	Website download	- 50 states - +1700 counties - 10 years of data - 2 categories
Unemployment	Unemployment and median household income for states and counties from 2007 - 2017	https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/	Website download	- 50 states - +1700 counties - 10 years of data - 2 categories



Features Also Considered				
Air Quality	Average age	Rainfall	Economic Growth	Traffic
Water Quality	Demographic breakdowns	Snowfall	Income Tax	Public transit
Healthcare Availability	Political affiliations	Severe weather events	Property Tax	Greenspace availability
Education Quality	Religious affiliations	Sunshine	Internet Quality	Sports

Additional steps to cleanse the data were taken to ensure we had accurate data for all county/year combinations to feed into our forecasting and mapping software. Average values of the five closest counties were imputed where data was missing. We obtained a data source of all county latitudes and longitudes and created a 10M row matrix of distances between counties using the Haversine equation to support the effort to impute missing data. Outliers were then capped at +/- 3x the standard deviation from the mean.

Algorithm: The analytics methods included forecasting trend factor values over the coming five years and grouping to display the matching level (i.e. best-to-least-best fit) of each county to the user preferences.

Forecasting was performed to predict the future values of all the features. Data was split into training and testing and machine learning and time series techniques were applied to measure which technique produced the lowest RMSE values. The best technique for each feature was applied to obtain the final forecasts.

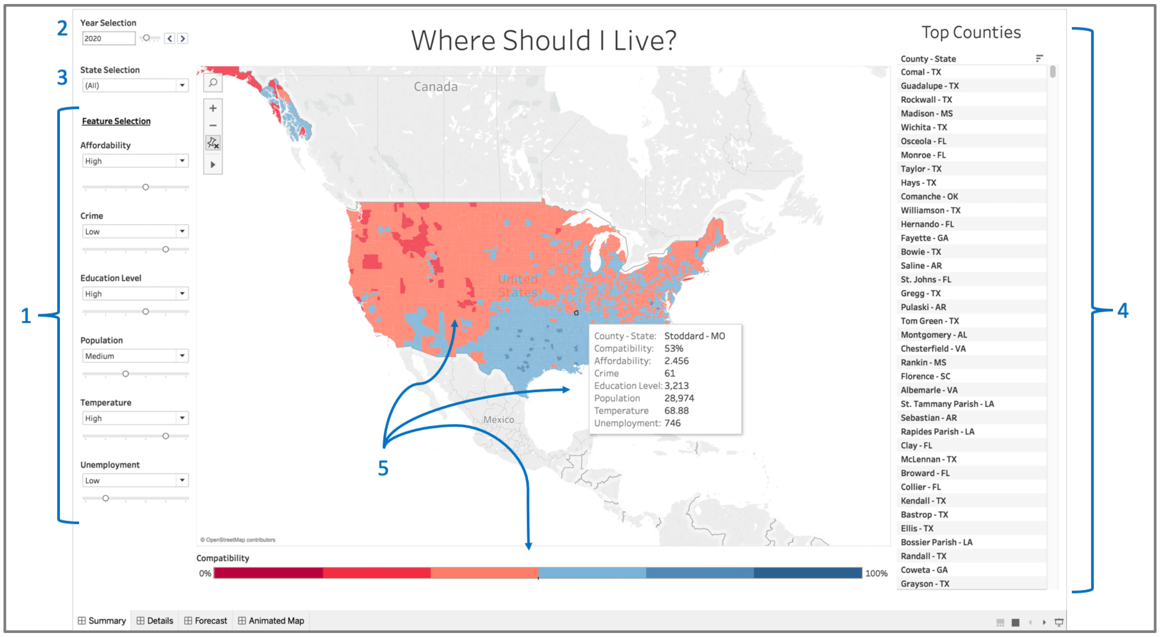
User preferences and weighting were considered in the user interface to rate and rank each county and return the list of most-to-least compatible options.

Interactive Visualization: Users will make their selections and view the results through an interactive Tableau dashboard. The dashboard is organized into the four areas named SUMMARY, DETAILS, FORECAST, and ANIMATED MAP:

SUMMARY SECTION: main landing area where user selections and weighting are entered. Results are displayed on US map visual with hover-over tool tips and county compatibility listing.

1. User selects feature preference (e.g. low, med, or high) from the dropdown then sets level of importance (e.g. 0-least important to 5-most important) on the slider
2. Output year is user selectable
3. Optional filter to select a state for which county results are displayed
4. Output list of counties best matching user preferences in descending order
5. Visual displays identifying degree of compatibility heatmap and tool tip of factor values on county hover over

(Summary Section)



Where Should I Live?

Year Selection: 2020

State Selection: (All)

Feature Selection

Affordability: High

Crime: Low

Education Level: High

Population: Medium

Temperature: High

Unemployment: Low

County - State: Stoddard - MO
Compatibility: 53%
Affordability: 2.456
Crime: 61
Education Level: 3,213
Population: 28,974
Temperature: 68.88
Unemployment: 746

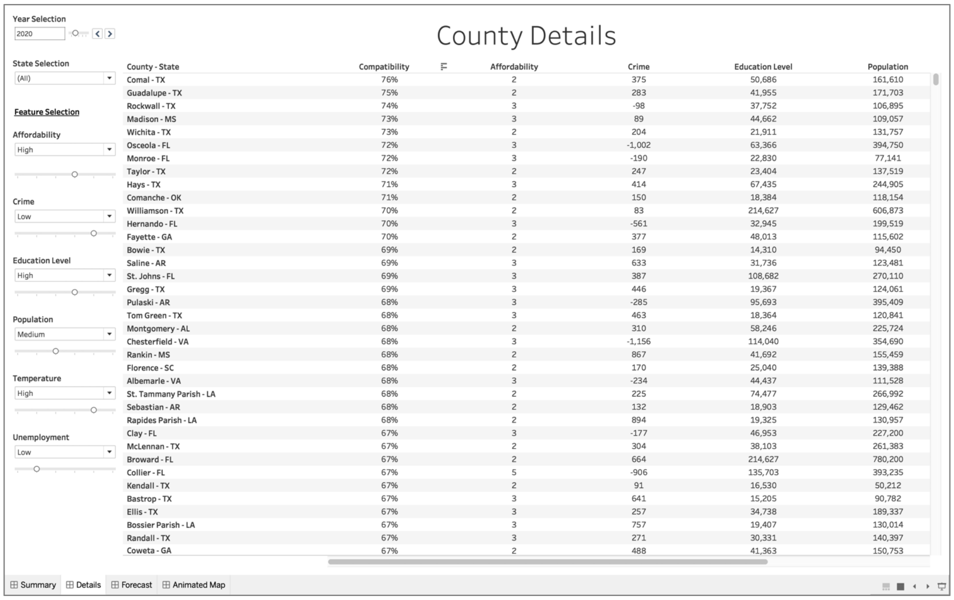
Compatibility: 0% – 100%

Top Counties

County - State
Comal - TX
Guadalupe - TX
Rockwall - TX
Madison - MS
Wichita - TX
Osceola - FL
Monroe - FL
Taylor - TX
Hays - TX
Comanche - OK
Williamson - TX
Hernando - FL
Fayette - GA
Bowie - TX
Saline - AR
St. Johns - FL
Gregg - TX
Pulaski - AR
Tom Green - TX
Montgomery - AL
Chesterfield - VA
Rankin - MS
Florence - SC
Albemarle - VA
St. Tammany Parish - LA
Sebastian - AR
Rapides Parish - LA
Clay - FL
McLennan - TX
Broward - FL
Collier - FL
Kendall - TX
Bastrop - TX
Ellis - TX
Bossier Parish - LA
Randall - TX
Coweta - GA
Grayson - TX

DETAILS SECTION: matrix view of county compatibility degree and feature values, based on user selections on left side of the screen.

(Details Section)



Year Selection: 2020

State Selection: (All)

Affordability: High

Crime: Low

Education Level: High

Population: Medium

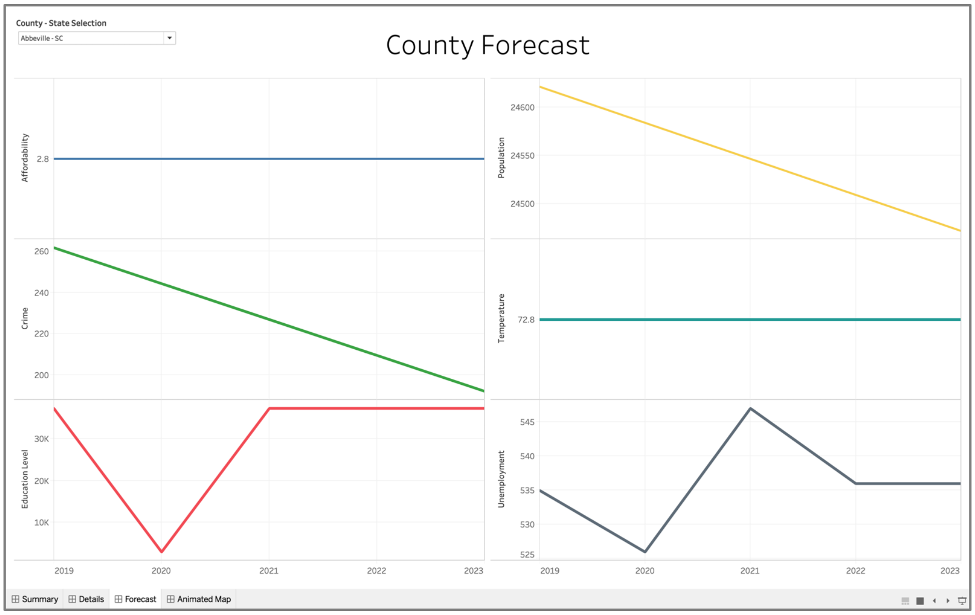
Temperature: High

Unemployment: Low

County - State	Compatibility	Affordability	Crime	Education Level	Population
Comal - TX	76%	2	375	50,686	161,610
Guadalupe - TX	75%	2	283	41,955	171,703
Rockwall - TX	74%	3	-98	37,752	106,895
Madison - MS	73%	3	89	44,662	109,057
Wichita - TX	73%	2	204	21,911	131,757
Osceola - FL	72%	3	-1,002	63,366	394,750
Monroe - FL	72%	3	-190	22,830	77,141
Taylor - TX	72%	2	247	23,404	137,519
Hays - TX	71%	3	414	67,435	244,905
Comanche - OK	71%	2	150	18,384	118,154
Williamson - TX	70%	2	83	214,627	606,873
Hernando - FL	70%	3	-561	32,945	199,519
Fayette - GA	70%	2	377	48,013	115,602
Bowie - TX	69%	2	169	14,310	94,450
Saline - AR	69%	3	633	31,736	123,481
St. Johns - FL	69%	3	387	108,682	270,110
Gregg - TX	69%	3	446	19,367	124,061
Pulaski - AR	68%	3	-285	95,693	395,409
Tom Green - TX	68%	3	463	18,364	120,841
Montgomery - AL	68%	2	310	58,246	225,724
Chesterfield - VA	68%	3	-1,156	114,040	354,690
Rankin - MS	68%	2	867	41,692	155,459
Florence - SC	68%	2	170	25,040	139,388
Albemarle - VA	68%	3	-234	44,437	111,528
St. Tammany Parish - LA	68%	2	225	74,477	266,992
Sebastian - AR	68%	2	132	18,903	129,462
Rapides Parish - LA	68%	2	894	19,325	130,957
Clay - FL	67%	3	-177	46,953	227,200
McLennan - TX	67%	2	304	38,103	261,383
Broward - FL	67%	2	664	214,627	780,200
Collier - FL	67%	5	-906	135,703	393,235
Kendall - TX	67%	2	91	16,530	50,212
Bastrop - TX	67%	3	641	15,205	90,782
Ellis - TX	67%	3	257	34,738	189,337
Bossier Parish - LA	67%	3	757	19,407	130,014
Randall - TX	67%	3	271	30,331	140,397
Coweta - GA	67%	2	488	41,363	150,753

FORECAST SECTION: graphical view of selected county forecasted feature values.

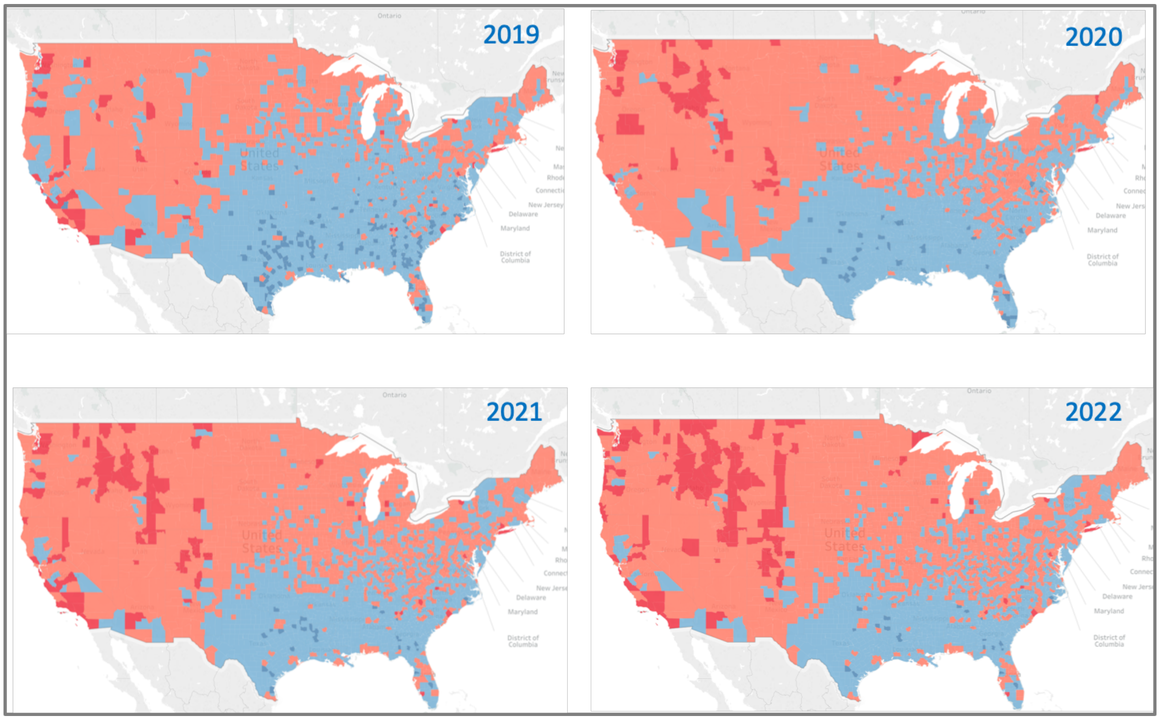
(Forecast Section)



County - State Selection: Abbeville - SC

ANIMATED MAP SECTION: visual presentation of SUMMARY view over the forecasted period.

(Animated Map Section – 4 of 5 years shown)



DISTRIBUTION OF TEAM MEMBER EFFORT

Assignments were delegated to balance effort, interests, and strengths. The plan has not changed from the original proposal and all milestones and deadlines were met.

		Week Ending							
Milestone / Task	**Leads**	**2-Mar**	**9-Mar**	**16-Mar**	**23-Mar**	**30-Mar**	**6-Apr**	**13-Apr**	**20-Apr**
Data Collection and Cleaning		▨	▨	▨	▨				
Dataset / Variable Selection	Morton	■	■						
Data Collection	Swaminathan		■	■					
Outlier Detection / Data Scrubbing	Morton			■	■				
Analysis				▨	▨	▨	▨		
Clustering	Rana			■	■	■	■		
Forecasting	Swaminathan			■	■	■	■		
Visualization			▨	▨	▨	▨	▨		
User Interface Layout	Jackson		■	■	■				
Output Design	Jackson				■	■			
Integration	McPherson					■	■		
Presentation						▨	▨	▨	▨
Poster Design	Rana					■	■	■	
Poster Script/Presentation	McPherson							■	■
Final Report	Team					■	■	■	■

All members contributed equally to the project.

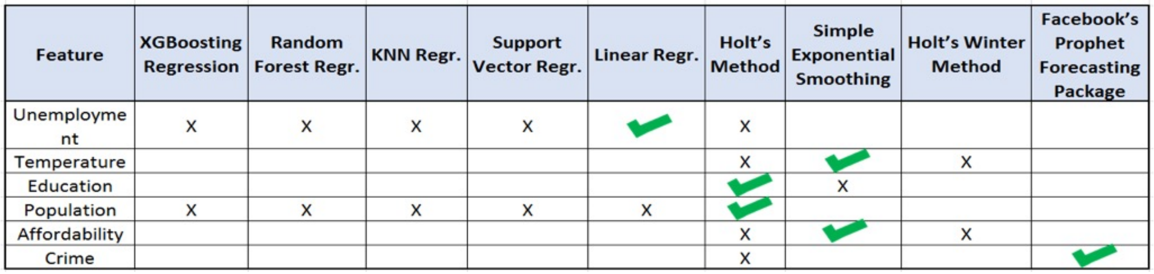
EXPERIMENTS AND EVALUATION

Our project aims to improve user experience through the use of more relevant and customized information. It is not feasible to perform a full-blown user experience experiment, as this would require identifying a pool of individuals changing where they live, having them use the tool in their decision process, and track the results over several years. Time constraints remove this as an option for this project.

Instead, our experiments focused on testing various forecasting techniques for improved prediction accuracy. The choice of mathematical models depended on the kind of data and the information available for each dataset.

Machine learning algorithms and time series techniques were applied to “Unemployment” and “Population”, since each dataset had feature columns to support these methods. For the others, various time series techniques were used. For “Education” and “Population”, we found the Holt’s method performed well. For “Temperature” and “Affordability”, simple exponential smoothing was best, while linear regression was the top choice for “Unemployment”. Finally, Facebook’s prophet package performed best for “Crime”.

(Methods considered and chosen for each compatibility factor)



Feature	XGBoosting Regression	Random Forest Regr.	KNN Regr.	Support Vector Regr.	Linear Regr.	Holt’s Method	Simple Exponential Smoothing	Holt’s Winter Method	Facebook’s Prophet Forecasting Package
Unemployment	x	x	x	x	✓	x			
Temperature						x	✓	x	
Education						✓	x		
Population	x	x	x	x	x	✓			
Affordability						x	✓	x	
Crime						x			✓

CONCLUSIONS AND DISCUSSION

Following are team member impressions of the project outcome and potential future enhancements to consider.

Jack McPherson: Overall I am satisfied at the final result of the project. The initial vision was ambitious, to build an interactive interface that determines compatible areas to live based on user selected features, preferences and importance. Having recently moved to a new area, the concept is one I wish I had available to me at the time I was making a decision. What we were able to deliver included six important features at the county level with projections out to 2023. The end result is a valuable tool where users can easily gain insights to what areas are a best fit for them and how those may change in the near future. It is a robust proof-of-concept with potential for further enhancements and refinements such as including additional features, drill downs to smaller geographies (e.g. cities, school districts) and a longer forecast time horizon.

Ajay Rana: The topic “Where Should I Live” is very interesting and that led us do a detailed research in functional and technology areas. This project should be very helpful and intuitive to the people who are looking for desired locations to move based on the importance of features criteria. Another novel approach in this project is to get future prediction about the feature results based on the advanced forecasting techniques. For example results show current crime rate as well as crime rate in 2023 for a specific location. We did dataset research on several features but finalized on six important to our team. We can further enhance the project by adding more features, and different location types (city, states, and zip codes). We could also add more UI enhancements and provide cloud hosted solution as time permits. Overall I am very satisfied with the project outcome and learned a great deal working on this project and working with the project team.

William Jackson: We were drawn to this project topic because it has far reaching applicability to a large portion of the US population and the sizeable data / analytics available to support the effort. As we researched novel ways of helping answer the question “Where Should I Live”, it became increasingly evident that there isn’t a “one size fits all” solution to answer this question. In fact, interpretation and personal values heavily influence what one individual may choose compared to someone else. As a result, we expanded the functionality to allow the user to weight the factors and further customize the response to their values. The project has met our original objectives of providing customized decision support. Future enhancements could explore a larger number of factors and allowing the user to pick which to include, longer factor projection periods, and expanded independent variable consideration for forecasting factor trends.

Sai Balagi Thalanayar Swaminathan: I would consider this project to be a success. We had a novel idea of “where a person wants to live” which impacts everyone and very less is known on how to solve it. This gave us an opportunity to look at all possible reasons that a normal person would consider while moving to a new place. Literature review was performed and finally six features were selected. Data was gathered, cleaned and was used to calculate the computability score along with forecasting the future values. A tableau interface was built to have the user choose his/her rating of the features and an interactive map along with summary is displayed to show the ideal places for the user to live based on his/her preferences. Going forward, we should build an interactive website, add in more features to select and have the ability to scale for more people to use.

Matthew Morton: I think the project is a great start at an interesting idea. The problem of a lack of a website or tool out there that can help identify cities or neighborhoods within a city based on customized user preferences is something that I recently experienced when moving into the city. Because of the timeline and availability of opensource data, we used counties as the base unit of data. This is fine for choosing cities, but more granularity is needed to identify specific neighborhoods within a city. This is a large change I would make if we were to take this project further. From a technical perspective, I think tableau was the right tool for the job. It did not have too great a learning curve for development, but had a robust array of capabilities that allowed us a lot of flexibility in how we wanted to display the data.

REFERENCE SECTION: SURVEY DETAILS

Researched by William Jackson

1. **Consumer Behavior - Attitudes**
   a. Idea: Attributes, Beliefs, and Importance weigh heavily into consumer decision making. Attributes are relevant object characteristics considered. Beliefs are feelings about Attribute values. Importance reflects the priority one gives to each Attribute in their final decision making.
   b. Usefulness: Project design will incorporate user factor selection(s) by weighted importance.
   c. Shortcomings: None to be addressed in the project.
2. **Teleport Cities – Where Should I Live? – Compare Cities' Quality of Life**
   a. Idea: Interactive website collects select information about the user then allows them to choose important factors to consider when choosing where to live.
   b. Usefulness: Concept is similar to the proposed project and provides relevant design examples to consider.
   c. Shortcomings: User input does not allow an importance ranking and results do not reflect forecasted measures.
3. **The U.S. City You Should Be Living In**
   a. Idea: Interactive website lets the user pick and weight factors important to their city selection. Further filtering

options are available after results are returned. Includes corresponding cities to avoid.

b. Usefulness: Factor weighting concept example will aid in project user interaction design.

c. Shortcomings: Results do not reflect forecasted measures.

Researched by Jack McPherson

4. Using Ridge Regression with Genetic Algorithm to Enhance Real Estate Appraisal Forecasting

a. Idea: A modified ridge regression with genetic algorithm enhances performance over traditional models.

b. Usefulness: Ridge regression is helpful to alleviate multicollinearity in datasets with many variables.

c. Shortcomings: The added complexity of the genetic algorithm modification may not enhance results significantly.

5. Buying a Home with a Resale Value: Location, Location, Location

a. Idea: This paper explores a multiregional model to evaluate how housing location choice depends on wealth and other area factors.

b. Usefulness: The variables selected for the OLS regression models are useful to note when building the forecast.

c. Shortcomings: The specific model utilized may not have direct applicability.

6. Clustering the U.S. Real Estate Markets

a. Idea: Useful real estate market clusters exist based off economic and geographic structures.

b. Usefulness: How metro areas were grouped in this paper will be useful to consider during clustering analysis.

c. Shortcomings: The model utilized is not entirely applicable.

Researched by Matthew Morton

7. Algorithms to live by: The computer science of human decisions

a. Idea: Discusses how to apply different computer science concepts to every-day problems.

b. Usefulness: We will have to relate every-day human problems to the visualization and CS algorithms that will be used behind the scenes.

c. Shortcomings: Book doesn't focus on any particular implementation of algorithms; only how to think critically about the prospect of doing it.

8. Algorithms – Substring Search

a. Idea: Introduction to the concept of substring search and 4 major algorithms for accomplishing efficient and accurate substring match.

b. Usefulness: Part of our project may need to include matching ill-defined user search input against backend databases.

c. Shortcomings: Doesn't fit exactly into our established goal. We will need to think critically on how to incorporate into our UI.

9. An Efficient Approach to Clustering Real-Estate Listings

a. Idea: Explores a way to integrate data from numerous semi-structured data sources for useful analysis.

b. Usefulness: Much of the data we'll need may be from one-off semi-structured or unstructured data. The techniques described here could potential be of use in pulling and clustering similar data ourselves.

c. Shortcomings: Narrow focus – paper assumes data format that our data is not likely to conform to. May need to be adapted to our use-case.

Researched by Ajay Rana

10. Places rated almanac

a. Idea: Provides the 9 selection factors (Ambience, Housing, Jobs, Crime, Transportation, Education, Health care, Recreation, Climate), in addition, it provides characteristics for the each.
b. Usefulness: Helps with the functional scope of our project.
c. Shortcomings: The book is limited for the interactive visualization and implementation aspects.

11. Where work pays: How does where you live matter for your earnings?

a. Idea: The research paper shows that earnings and cost of living for the similar occupations can vary by location and varies for different age range.
b. Usefulness: The paper puts more emphasis on interactive web tools for better visualization and decision making.
c. Shortcomings: The paper focuses on the only a couple of selection parameters.

12. Going Dutch: How I Used Data Science and Machine Learning to Find an Apartment in Amsterdam — Part I

a. Idea: This article shows data science tools, techniques for finding suitable apartments.
b. Usefulness: Our project could use examples from the technology point of view.
c. Shortcomings: The topic is smaller in scope and does not provide details about time series forecasting and visualization part of our project.

Researched by Sai Balagi Thalanayar Swaminathan

13. A hybrid MCDM approach for agile concept selection using fuzzy DEMATEL, fuzzy ANP and fuzzy TOPSIS

a. Idea: Authors apply a combination of fuzzy MCDM techniques to find the best agile concept.

b. Usefulness: Application techniques used to model linguistic input and quantitative weights to find the importance of each feature.

c. Shortcomings: Does not include content directly applied to neighborhood recommendations.

14. Mining of Massive Datasets

a. Idea: Addresses standard techniques used for recommender systems.

b. Usefulness: We will apply content-based recommendation system to determine neighborhood preference.

c. Shortcomings: Does not include content directly applied to neighborhood recommendation.

15. Zillow Home Value Index and Forecast

a. Idea: Real estate prices forecasts based on SVM and parameter tuned using Particle Swarm Optimization.

b. Usefulness: Our model will forecast neighborhood factors.

c. Shortcomings: Does not involve econometric models to account for economic variability.