

FINAL REPORT FOR TEAM07: Shouldweinvest.com

Stephen Wagner (swagner34), Kayla Looney (klooney3)

Jacob Pierson (jpierson7), Sam Bryan (sbryan35), David Jesse Moody (dmoody6)

INTRODUCTION

Real estate investors have an overwhelming amount of information available to use in helping to determine where to allocate their investment dollars. Having noticed that sifting through this data is a struggle, our team was determined to use the tools of the data analytics space to improve this process for investors.

PROBLEM DEFINITION

Identifying attractive investment opportunities is a challenge faced by real estate investors. Available tools focus on identifying individual properties within a geographic market. However, for investors that want to identify geographic markets before focusing on individual properties, available tools are limited and it can be overwhelming trying to identify a market that fits their investment criteria. Our solution to this problem is a website, shouldweinvest.com, tailored to investors trying to identify attractive geographic markets for investment. Shouldweinvest.com innovates by identifying investment opportunities, based on custom investment personas, for ZIP Codes™ rather than individual properties and presents recommendations through an interactive user interface. The recommendations are determined by aggregating and analyzing data predictive of investment property returns. Value is provided to users by streamlining the process of identifying investment opportunities - increasing users ability to identify high ROI investments and decreasing time spent researching opportunities.

SURVEY

Shouldweinvest.com provides investment recommendations using a custom algorithm that incorporates factors that are predictive of property prices. Last period returns were found to be predictive of future housing prices, at the metropolitan statistical area (MSA) level [1], [2]. Knowing this, our algorithm will include lagged price movements. A shortcoming of [1] and [2] is the focus on only lagged prices. Regional differences in pricing momentum were noted in [3]. A critique of [3] is that it relied on a small sample of only fourteen cities in Finland. Our algorithm was expected to also include Google Trends data reflecting real estate interest in an MSA; shown by [4] to be positively correlated with home prices. However, Google Trends data wasn't available. Additional recommendation factors will address the limited analysis of [4].

[5] along with [6], [7] provide a foundation of metrics to take into account when assessing investment attractiveness. Factors to incorporate into our algorithm may include home value movement, positive net "migration" into the area, declining rental availability, overall aging population, and a strong school district [5], [6], [7]. Some shortcomings include that [5] only analyzes one city and [6], [7] focus on smaller regions than we will, but we will improve on this by analyzing the entire US [5], [6], [7].

[8] and [9] investigate the correlation between stock market prices and home values. There is a discount factor for less desirable neighborhoods that may not fit long-term

sentiment [8], and some areas may correlate more closely with the stock market. Shortcomings of these papers are that they focus on the list price over sale price [8] and lack analysis on whether that correlation hurts diversification [9]. [10] explains that closing costs go beyond sales price and can be originating from the loan as well, diminishing returns. [10] also does not describe what average costs might be.

Other data mining applications have been created and published [11], [12] in Europe and Asia but do little more than display the data. One application simply aggregated data from four data stores [12] while another discovered long-term trends [11]. The overall shortcoming of current applications is that they do not make recommendations to the user. The user is left to draw his/her own conclusions based on the data.

[13] suggested analyzing individual properties based on the net cash flow and ROI that they would provide [13]. One key metric when doing this analysis is the rent to property price metric, where the monthly rent should be at least 1.5% of the property's value [14]. Multi-family homes typically achieve this threshold more easily but tend to be less liquid than single-family homes [15]. Shouldweinvest.com will help investors identify ZIP Codes™ that achieve the highest rent to price ratio, which is a shortfall of [13], [14], [15]. A shortcoming of [13], [14], [15] is that it does not describe tools or strategies to quickly compare multiple markets.

PROPOSED METHOD

Shouldweinvest.com innovates in two ways over existing solutions. The site allows prospective investors to perform market level vs. individual property searches for potential investment opportunities. Also, the site provides predefined investor personas, that also allow additional customizability, that tailors results to investor goals and objectives.

The interface for shouldweinvest.com will be a single-page application that is split into a landing page and results page. Technologies used in the interface include HTML, CSS, JavaScript, and embedded Tableau Public dashboards. The landing page introduces users to the objectives for using the site, the UI elements for configuring the investor personas, and the results of the analysis. On the landing page, users have the option to select from two predefined personas, Landlord and Flipper. Site visitors are provided descriptions for the personas but are expected to determine which persona fits their investment goals and criteria. The Landlord persona is tailored to investors who want to hold a property long-term and be a lessor. The Flipper persona is tailored to investors who want to sell a property shortly after purchasing - likely after making further investments, remodeling the property. After selecting a persona, users make further refinements through adjustment of the recommendation algorithm's factor coefficients. Factor adjustments are done through radio button UI elements that correspond to factor importance levels. While refining the recommendation algorithm, the analysis is performed dynamically to display recommendation results. The analysis results provide

five recommended markets for investment. It also includes multiple visual elements: a map highlighting the location of the five recommended markets, details for each of the recommendations, and bar charts that highlight how each factor compares with the broader dataset.

Our datasets were obtained from three primary sources: Experian, Zillow, and GreatSchools. We started with data from the US Census but since the last census was in 2010 we augmented it with Experian data. The Experian dataset is 7GBs, with 238 columns and 248,013 rows, and includes data from the American Community Survey 2018 and projections for 2023. After aggregating all of the Experian data to the ZIP Code™ level, we identified key variables and used SQL to join the Experian data with the key variables our survey identified from Zillow. Ten different datasets, from Zillow, were necessary to build our scoring algorithm. Zillow datasets ranged in size and had roughly 200 columns and 20,000 rows, on average.

GreatSchools data was harvested using a Python script that accessed their API. The script used ZIP Codes™ from the Zillow data to pull school-related data using the API. GreatSchools rates each school on a scale from 1-10; which we averaged for each ZIP Code™ and included in our algorithm. The original dataset was for 74,000 schools and was cleaned down to 57,000 schools because many did not have a usable rating.

We had to be systematic in how we analyzed the data because it frequently pushed the computation limits of our hardware. For data cleansing, we started with OpenRefine but for better performance, switched to Dataiku. A mix of R and Python was used to calculate the algorithm's custom metrics. Our analysis resulted in a concise, aggregated dataset extracted as a Tableau Hyper Database file.

The Tableau Hyper Database file was used within a Tableau Dashboard to feed two recommendation algorithms. Each algorithm was tailored to the goals and objectives specific to a broad investor persona - Flipper or Landlord. The Flipper persona is targeted at investors who want to re-sell investment property shortly after purchase vs. the Landlord persona which is targeted at investors who want to hold a property long-term and be a lessor. Both algorithms recommend markets that are healthy with declining home vacancy and good schools. The Flipper algorithm also prioritizes markets where properties are older and sell quickly. The Landlord algorithm recommends ZIP Codes™ that are considered a "buyer's market" and have a high income to rent ratio. The importance of each factor used in the algorithms can be updated in the site interface - using a "Not Important" to "Most Important" scale (i.e. 0 to 4). The algorithm details are displayed in Figure 0.

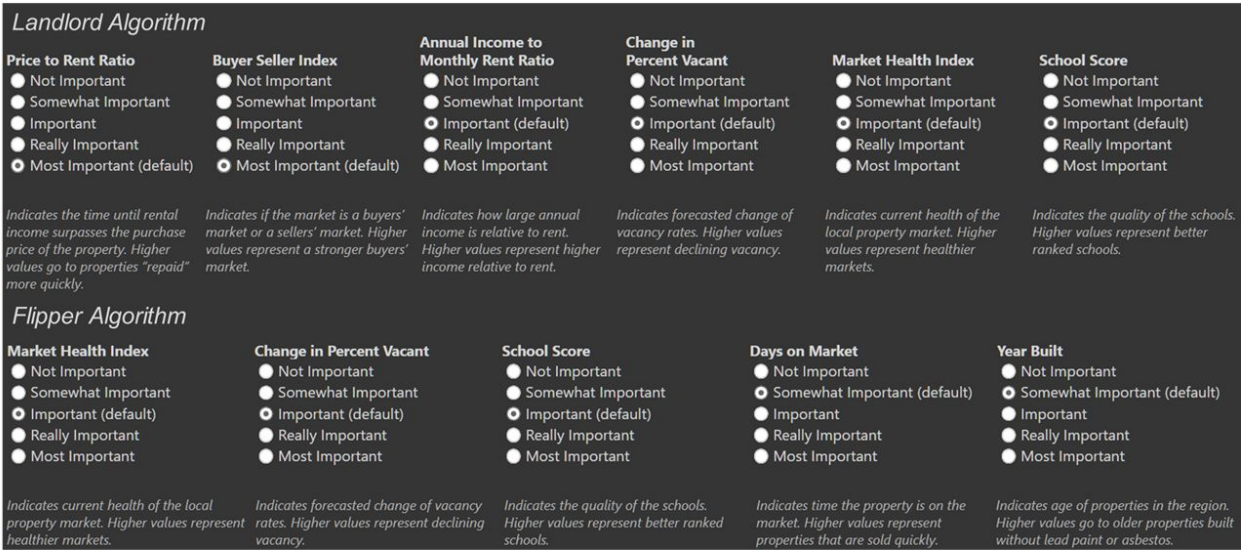


Figure 0: Recommendation Algorithm Details

Shouldweinvest.com includes the following architectural elements:

Component	Location	Details
Domain	www.shouldweinvest.com	Address to access server
Hosting	Digital Ocean	Hosted Server
Service	Flask/Python	HTTP Service for displaying website
Database	Tableau Public	Hyper Database file containing aggregated data necessary to display analysis results
Data Visualizations	Tableau Public	Dashboards that display recommendations
Code Repository	github.com	Private Github repository for version control

Figure 1: Shouldweinvest.com System Architecture

DISTRIBUTION OF TEAM ACTIVITIES

Team members have contributed a similar amount of effort across the activities displayed in the table below.

Activity Area	Sub-activity	Detail	Effort	Team Member Responsible
Database	Creation	Extract data into Hyper Database file	2 hours	Jesse
Infrastructure	Provision Server	Server provisioning	5 hours	Jesse
	Repository Creation	Github repository	1 hours	Jesse
	Installation	Installation and updates of server infrastructure	10 hours	Jesse
	Domain	Domain purchase and DNS record updating	2 hours	Jesse
	Access	Provide server access for all users	2 hours	Jesse
	HTTP Service	Setup of server for HTTP service	10 hours	Jesse
Data Collection	Acquire Data	Identify appropriate datasets to use	15 hours	Kayla, Sam
	Data Cleansing and Calculation	Clean data and calculate key scoring metrics	20 hours	Kayla
	Data Investigation and Algorithm Creation	Create scoring algorithm and select key variables	15 hours	Kayla
	Algorithm Experimentation	Experiment with different scoring algorithms	10 hours	Kayla
	School Data Collection	Use an API to generate a datafile of school level performance	20 hours	Sam
Interface	Site Landing Page	Create and style HTML landing page	15 hours	Jacob
	Site Results Page	Create and style HTML results page	15 hours	Jacob
	Configuration Visual	Create Tableau Public dashboard that allows user to adjust algorithm	15 hours	Steve
	Results Visual	Create Tableau Public dashboard that displays analysis results	15 hours	Steve
	Design Iteration	Focus test the design of the site and revise based on feedback	15 hours	Steve, Jacob
	Integration	Integrate Tableau Public dashboards with site pages	20 hours	Steve, Jacob
	Deploy	Deploy page to shouldweinvest.com	10 hours	Jesse

Figure 2: Revised Plan of Activities

EXPERIMENTS AND EVALUATION

An initial algorithm was created using a subset of available variables. After implementation, we iteratively enhanced our scoring algorithm by selectively increasing the number of variables our model relied on for scoring ZIP Codes™. Predictive models were built in Microsoft Azure Machine Learning Studio to identify variable importance. The findings were incorporated into our scoring algorithm.

The “Flipper” version of our algorithm includes factors that relate to how fast properties resell and property demand for a ZIP Code™. Additionally, we strive to identify markets that are considered “healthy” relative to other markets in the US. A healthy market is when recent homes sold for a gain, had positive year-over-year home value change, and few delinquent homes are in the ZIP Code™ [6]. Lastly, ZIP Codes™ with older homes are favored, as [7] identified older homes as being a good candidate for flipping. In order to verify that this algorithm is working, we manually researched the top results returned by the flipper persona (currently Gilroy CA, Mission Viejo CA, Grand Prairie TX, Pleasanton CA, and Gallatin TN) and verifying that they met the criteria for our algorithm and were generally rated as good markets for flipping.

The “Landlord” version of our algorithm focuses on factors that favor property lessors. We identify ZIP Codes™ that represent a “buyer’s market” to provide the investor beneficial negotiating leverage. Like Flippers, we also look for healthy markets and declining rental availability [5, 6]. Additionally, we take the price of the house into account as well as how much the landlord could charge in rent. ZIP Codes™ with a low home price to rent ratio are preferred, suggesting the landlord could collect enough rent to pay for the house quicker. Percent of vacant homes and declining forecasted vacancy are also positive indicators for positive Landlord returns [6]. Lastly, median household income relative to the cost of rent is factored into the algorithm. In order to verify that this algorithm is working, we manually researched the top results returned by the flipper persona (currently McKinney TX, Celina TX, Frisco TX, Ponte Vedra FL, and Cumming GA) and verified that they met the criteria for our algorithm and were generally rated as good markets for renting.

In order to determine if our algorithm was effective at making recommendations based on the user input, we designed a series of experiments to isolate the variables and verify the results. For example, we marked all the factors as “Not important” except for schools which were marked as “Most Important”. The resulting recommendation, Solebury PA, was then verified that it did, in fact, have an extremely high average school rating. This process was repeated for every algorithm factor and manually verified.

We experimented with the site design through an iterative process. Each iteration, feedback was gathered from team members to identify areas to improve. Wireframes were built to show the overall layout of the application. The team then chose a color

scheme that complimented our “needle in a haystack” theme, used in our presentation video. This color scheme was combined with the wireframe and Tableau graph to create an initial mockup of the application.

LANDING
PAGE

Should We Invest

Should We Invest helps you, as an investor, identify residential real estate markets that provide the best investment opportunities based upon your investment criteria

Select an Investor Profile

Flipper

The Flipper profile is optimized for investors who have a short time-horizon and are looking to re-sell their investment quickly

Landlord

The Landlord profile is optimized for investors who have a long time-horizon and are looking to lease their investments

Custom

The Custom profile is weighted evenly between the investment factors and can be further refined

Adjust the sliders to further refine your criteria importance

Criteria A

Criteria B

Criteria C

Low


Importance

High

View results

RESULTS
PAGE

Recommended Markets



8,728
3,334

Rank 1

55408 – Minneapolis, MN

92

SCORE

Population 25,125

Median Income \$82,135

Avg. Days on Market 32

Percentile

Rank 2

90210 – Beverly Hills, CA

90

SCORE

Population 31,230

Median Income \$162,234

Avg. Days on Market 43

Percentile

Rank 3

55555 – Middle America, MO

89

SCORE

Population 19,325

Median Income \$62,324

Avg. Days on Market 15

Percentile

Figure 3: Initial Wireframes

Feedback on usability and design was received with each design iteration. An example of feedback was, “One recommendation I have would be to maybe use a different color scheme. Because your data is at the zip code level, until I zoomed in a lot I wasn’t sure

if the colors were changing at all because it was a sea of dark blue. Using a lighter color for the majority of the data will help highlight the darker/more intense color used to show the areas of focus, possibly even when zoomed out some.” As a result, the map was replaced with small multiple maps that highlighted the specific recommended locations. After a number of iterations, the final design, shown below, was used.

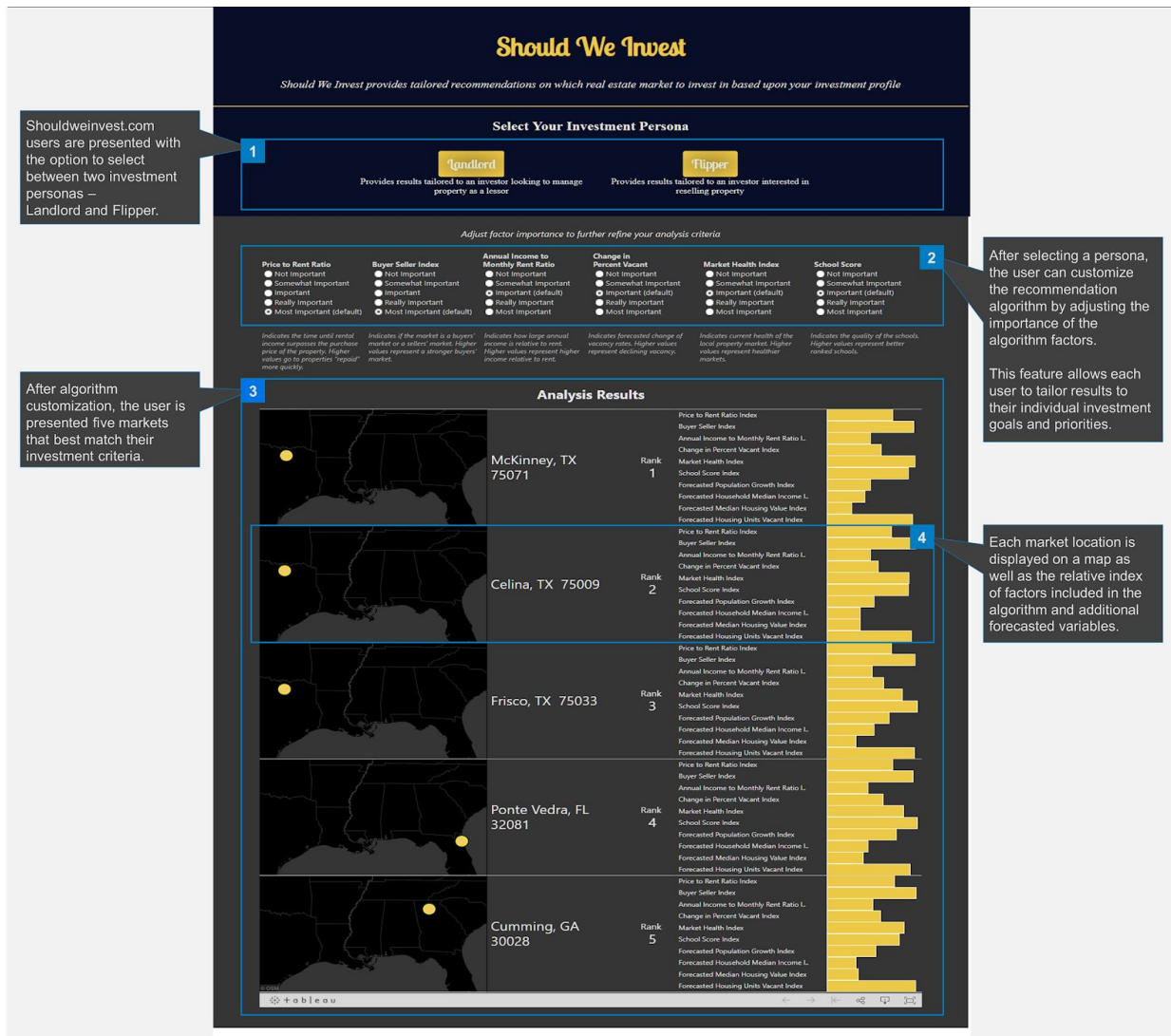


Figure 4: Final version of Shouldweinvest.com

CONCLUSIONS AND DISCUSSION

Shouldweinvest.com solves the challenge of identifying attractive markets for real estate investment by providing recommendations for real estate markets across the United States. These recommendations are generated from two algorithms that utilize over seven gigabytes of data aggregated from three data sources and accepts user customizable parameters that ensure that the market recommendations are specific to the user's unique preferences. The interface for the web application combines modern design principles with fast responsive results to ensure that user experience is an improvement over existing methods. Shouldweinvest.com is ideal for those looking for markets where they want to invest in real estate for short-term or long-term gain or for anyone who is interested in learning more about real estate markets across the United States.

A number of future efforts would be used to improve upon the current algorithms and site design. The first would be to continually gather investment performance feedback from investors to ensure that the algorithms are performing at a high level. Also, continually surveying the research space to ensure the algorithms include factors known to be predictive of investment performance. Lastly, a future action related to the site design would be to perform A/B testing on design elements to continually inform and improve the site's usability.

REFERENCES

- [1] E. Beracha and H. Skiba, "Momentum in Residential Real Estate," *Journal of Real Estate Finance and Economics*, vol. 43, (3), pp. 299-320, 2011. Available: <http://prx.library.gatech.edu/login?url=https://search.proquest.com/docview/880355946?accountid=11107>. DOI: <http://dx.doi.org/10.1007/s11146-009-9210-2>.
- [2] K. E. Case and R. J. Shiller, "The Efficiency Of The Market For Single-Family Homes," *Am. Econ. Rev.*, vol. 79, (1), pp. 125, 1989. Available: <http://prx.library.gatech.edu/login?url=https://search.proquest.com/docview/233034208?accountid=11107>.
- [3] E. Oikarinen and J. Engblom, "Differences in housing price dynamics across cities: A comparison of different panel model specifications," *Urban Stud.*, vol. 53, (11), pp. 2312-2329, 2016. Available: <http://prx.library.gatech.edu/login?url=https://search.proquest.com/docview/1811897848?accountid=11107>. DOI: <http://dx.doi.org/10.1177/0042098015589883>.
- [4] E. Beracha and M. B. Wintoki, "Forecasting Residential Real Estate Price Changes from Online Search Activity," *The Journal of Real Estate Research*, vol. 35, (3), pp. 283-312, 2013. Available: <http://prx.library.gatech.edu/login?url=https://search.proquest.com/docview/1449565441?accountid=11107>.
- [5] C. A. Depken, II, H. Hollans and S. Swidler, "An Empirical Analysis of Residential Property Flipping," *Journal of Real Estate Finance and Economics*, vol. 39, (3), pp. 248-263, 2009. Available: <http://prx.library.gatech.edu/login?url=https://search.proquest.com/docview/203163757?accountid=11107>. DOI: <http://dx.doi.org/10.1007/s11146-009-9181-3>.
- [6] R. R. Roberts and J. Kraynak, "Scoping Out a Fertile Neighborhood" in *Flipping Houses for Dummies*. Hoboken, NJ: Wiley, 2017, ch.6, pp. 79-92.
- [7] R. R. Roberts and J. Kraynak, "Inspecting the Property with an Eye for Rehab" in *Flipping Houses for Dummies*. Hoboken, NJ: Wiley, 2017, ch.11, pp. 161-176.

- [8] W. Seo, "Does Neighborhood Condition Create a Discount Effect on House List Prices? Evidence from Physical Disorder," *The Journal of Real Estate Research*, vol. 40, (1), pp. 69-87, 2018. Available:
<http://prx.library.gatech.edu/login?url=https://search.proquest.com/docview/2030180351?accountid=11107>.
- [9] C. W. Anderson and E. Beracha, "Home Price Sensitivity to Capital Market Factors: Analysis of ZIP Code Data," *The Journal of Real Estate Research*, vol. 32, (2), pp. 161-185, 2010. Available:
<http://prx.library.gatech.edu/login?url=https://search.proquest.com/docview/578337166?accountid=11107>.
- [10] D. Hymer, "Starting Out" in *Starting Out: The Complete Home Buyer's Guide*. San Francisco: Chronicle Books, 1997, ch.1, pp. 5-34.
- [11] E. Hromada, "Mapping of Real Estate Prices Using Data Mining Techniques," *Procedia Engineering*, vol. 123, pp. 233-240, 2015. Available:
www.sciencedirect.com/science/article/pii/S1877705815031847.
- [12] S. GuoDao, et al., "A Web-Based Visual Analytics System for Real Estate Data," *Science China*, vol. 56, pp. 1-13, 2013. Available:
<http://godoorsun.org/papers/science%20china.pdf>.
- [13] B. Turner, "Analyzing a Rental Property" in *The Book on Rental Property Investing*. Denver: BiggerPockets Publishing LLC, 2016, ch.5, pp.97-120.
- [14] B. Turner, "Investing While Living in an Expensive Area" in *The Book on Rental Property Investing*. Denver: BiggerPockets Publishing LLC, 2016, ch.6, pp.121-132.
- [15] B. Turner, "Types of Rental Properties" in *The Book on Rental Property Investing*. Denver: BiggerPockets Publishing LLC, 2016, ch.7, pp.133-150.