U.S. Real Estate Market Visualization and Prediction

Qianyun Chen, Chenyu Dai, Fei Ding, Shuge Fan, Lintong Han and Shaotong Sun
Georgia Institute of Technology
{qchen336,cdai38,fding33,sfan72,lhan63,sunshaot}@gatech.edu

1 INTRODUCTION

Real estate plays an integral role in the U.S.economy. People have a lot of demand to buy houses, both for living and investment. At the individual level, about 65 percent of homes are owned by owners, and housing is often a significant source of household wealth in the United States. The advantages of investing in real estate include passive income, stable cash flow, tax advantages, diversification, and leverage. Finding a suitable house to invest in or live in is a complicated process, and people need to consider many factors. Two of the most important factors are location and price. From an investment and home purchase perspective, we thought it would be beneficial for users to have a more comprehensive view of real estate markets in all states and counties and be able to analyze the timing, price, and location of investments through historical prices.

Therefore, we decide to develop an application that allows users to more intuitively observe the price distribution of state-level and county-level real estate markets, and the price trend of each state. The user could visualize the housing sources within a certain distance according to geographical location with hot/cold spots. We will use the historical data to make price forecasts in the future with linear regression and additive regression models. In addition, our tool calculates the Sharpe ratio that assists investors to understand the predicted return of an investment compared to its risk.

2 PROBLEM DEFINITION

There are several existing real estates visualization websites embedded data visualization such as Zillow, Redfin, and Trulia on the market. However, there are some common shortcomings of these systems for the aspects of visualization. One is that they do not allow users to filter based on geographical data such as distance. There is no effective feature that visualizes state real estate market prices and helps users compare them across states. There is not an efficient approach to finding properties that could meet users' demands within a certain distance. Predicting the real estate price trend is also not precise at the county/zip code level. The purpose of this project is to build a robust visualization system to tackle these problems.

3 LITERATURE SURVEY

Inspired by Sarker [16]'s advanced analytical method, our team decides to combine data mining, visualization, and machine learning modeling to create a web-based project. Similarly, Cheryshenko et al. [2] find how big data with an analytical data streaming method can overall better ensure satisfying decisions for investors. In particular, our web platform [18] will integrate with Google Maps API for geographical visualization; however, instead of using their counter-intuitive stacked graph view, we are more inclined to adopt Havre et al.'s [8] ThemeRiver prototype system, which effectively discerns real estate trends by analyzing texts. We will need to continue exploring that thread with numerical data in our case. Li et al. [12] introduce Homeseeker to interactively gather all information and filter them based on zip code, transportation, and price. We will build a similar filter to allow users to pinpoint their needs. Hong's [9] team develop a system called Real estate Visualizer for exploring the real estate property listings on the web and visualizing them using map-based color-coding techniques. Da'anna et al. [4] introduce a novel approach of integrating a sensitivity analysis method and an interactive visualization to a visual interactive sensitivity analysis environment, which is applied to a real estate prediction system represented in tornado graphs, radar graphs, scatter plots and parallel coordinate plots. However, we find that none of the above representations can be suitably applied to our dataset to serve our project purposes; therefore, we may stick with the simple bar chart design for visualizing price distributions. To explore a high-dimensional dataset (with multiple feature columns) like ours, Ge et al. [6] introduces Parallel Coordinates to visualize trends based on different pairs of factors and/or their ratios, using which researchers can further back their analysis and hypothesis. In addition, Agarwal et al. [1] create 3-D interactive heat maps with the semi-parametric model to dynamically visualize the real estate price affected by multi-factors. To create smooth-looking continuous representations of real estate price trends across regions, Koramaz et al. [10] apply interpolation and regression model to analyze how spatial characteristics affect the housing price trends, which will be helpful as we have a very discrete and sparse dataset.

Pérez-Rave[14] proposed a machine learning approach to the regression analysis of big data for both inferential and predictive purposes and introduced methodologies for the analysis of regression, which demonstrates to us on constructing the machine learning model including the process of sampling, training and validating. But this paper lacks insight on analyzing prediction algorithms. Therefore, we find several references to decide which algorithm best fits our data. Sharma et al[17] demonstrate data mining and predictive analytics using real estate data, applying decision trees and random forests for the prediction, obtaining 80% accuracy. Their findings help us understand the lifestyle of a data mining project. Karshiev[15] also uses a random forest algorithm to predict future prices, with a 92.01% accuracy. Besides, Karshiev[15] used KNN to handle missing data in the large dataset, bringing great referencing value while processing the data. Ghosalkar[7] conducted linear regression instead, giving the minimum prediction error of 0.3713. As a result, since our data has many features with lower noise, we decide to use the liner model, which may outperform the random forest algorithm. Krishna[11] introduces a methodology called CRISP-DM, and visualizes it through multiple algorithms where gradient boost regression is optimal. However, the gradient regression does not fit our project since it is optimal for the smaller dataset. DeLisle[5] utilizes big data about regime shift in real estate to lead enhance the real estate decision through Spatio-temporal Choropleth and categized by multiple factors.

In addition, Pai et al.[13] employ real transaction data and apply various machine learning algorithms to predict real estate prices, verifying least-squares SVR is most suitable for their task. Chiu et al.[3] incorporate spatial and temporal features for a lightweight CNN-LSTM network as the predictor model. However, their dataset is tailored to emphasize their used feature fields. Our dataset consists of aggregate transaction histories and coarse locations and time, so we decide the best attempt is to design a similar sequence neural model using Spatiotemporal features, using a simple linear regression as the baseline. According to the literature description, it's beneficial for our project since they provide the general direction of completion. Specifically, we apply world map and line chart visualization that rely on region and period differences. After that, we will perform the machine learning model linear regression, prophet model/addictive regression model to predict the future estate prices. At this point, we imitate and extend those authors' operations to finish the project.

4 PROPOSED METHODS

Our overall method is to use a combination of visualization techniques and modules to create a informative and insightful real estate discovery tool.

4.1 A choropleth map of the average estate price for every state in the U.S. The choropleth displays the relative estate prices for each geographical region in color gradations where higher values correspond to more intense colors. This is useful for comparing two states' prices and discovering hot spots. Users can also use the time slider to select a particular month of interest, and our map will update accordingly. We create dedicated APIs to filter and return monthly data per user request for improved performance. The data will be used as input to the geoAlbers projection from d3-composite-projections in the d3.js library. We then compute the max/min average prices and percentiles to create color gradations. The color of a certain state depends on within which color group its average price falls. A tool-tip following the user's mouse will also show detailed information about the state estate market. Figure 1 is an example of such choropleth map for July 2021.

Besides visualizing the absolute price values, we also would like to highlight price changes, so the choropleth map also supports "visualizing changes" option where all states are colored to their monthly percentage change in price values. We also create a sorted bar chart ranking all states according to the price percentage changes from the max increment to max decrement. See Figure 2 for an example during June to July 2021.

Another auxiliary visualization tool we create worth of mentioning is a pie chart displaying the top 10 states with the most listing counts. This amount is an indicator of the activeness of the state's estate market and may be useful to the investors. See Figure 3 for an example of July 2021.

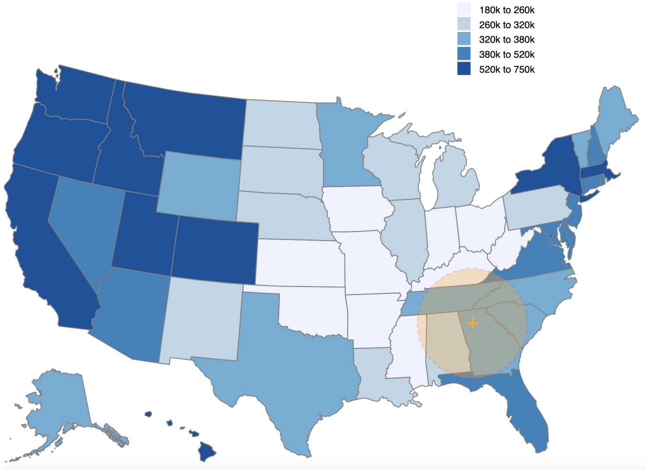


Figure 1: Choropleth Price Visualization.

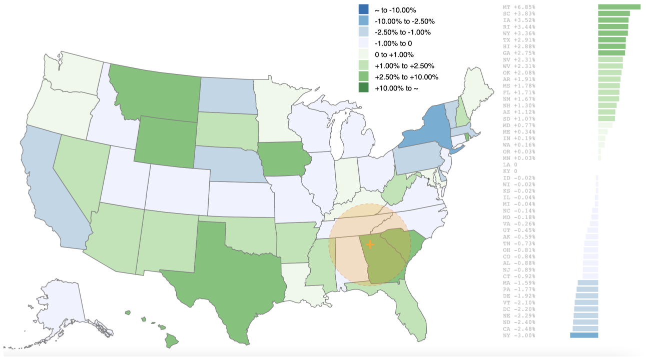


Figure 2: Monthly Price Change Visualization.

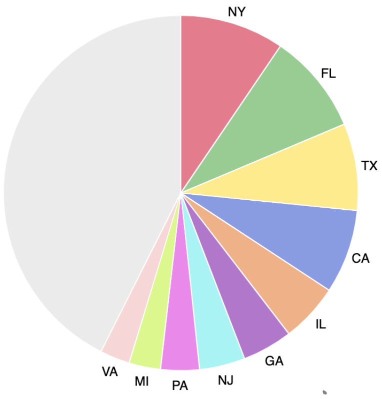


Figure 3: Listing Count Distribution.

4.2 Estate prices nearby visualization using concentric circles. The user can right-click a certain county on the map to highlight a pin. The slide bars on the right allow users to pick the maximum radius and number of concentric circles. Our tool will calculate the average prices of all counties that fall within that circular area and specifically which proximity group each county falls in based on its distance from the highlighted pin: we use the county center's coordinates and the cursor's coordinates to calculate it. After filtering all counties, we average the prices for each group, draw concentric circles on the right, and display these average of average values. At the same time, The color patches of the concentric circles also conform to these values to help the user discover price hot spots and regional differences intuitively.

Figure 4 is an example user interface for the proximity price visualization near Atlanta, Georgia.

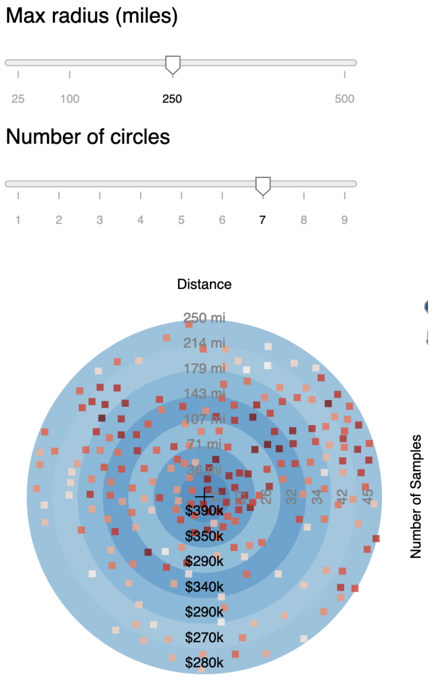
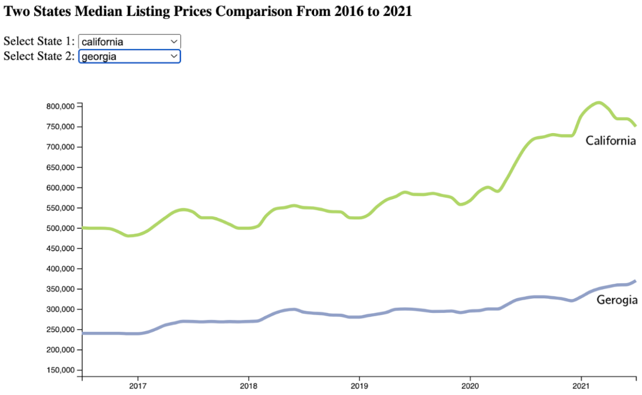


Figure 4: Proximity Price Visualization.

4.3 Past and future estate price trends with line charts. The line chart not only shows the price history but also serves as a forecast. We will discuss how ML can help with prediction later. After the user's mouse hovers on a state, a line chart will appear under the U.S. Map. The x-axis indicates the months of years, and the y-axis shows prices. We will use a solid line to represent original prices followed by a dashed line representing predicted prices for clear distinction. One of our innovations is the ability to compare two state prices, where we can also keep one state fixed and compare it with others (as specified by the user's mouse). Figure 5 is an example of comparing the estate price trends in Georgia against in California. Figure 6 is an example of visualizing the price history and forecast for the state of New York.



Two States Median Listing Prices Comparison From 2016 to 2021

Select State 1: california

Select State 2: georgia

Figure 5: Compare Two States' Price Trends.

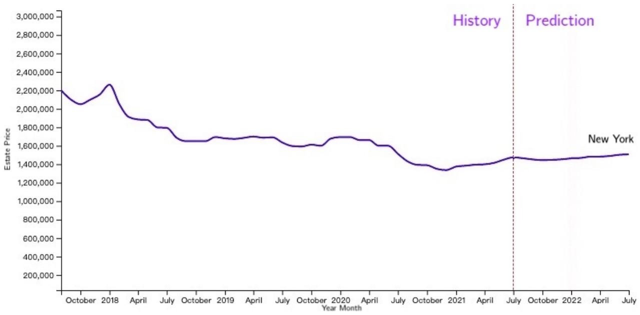
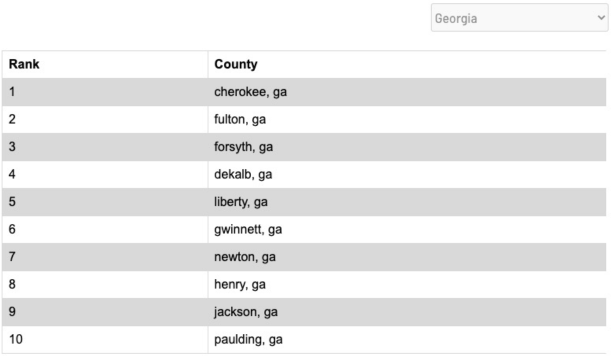


Figure 6: New York Price History & Forecast.

4.4 ML model predictions of future estate prices. We will experiment two tracks: the linear regression model and the addictive regression model (Prophet). Both methods have different sets of merits despite the fact they may use the same set of features: prices in the past X months, where X is an adjustable hyperparameter we choose to minimize the prediction errors (MSE), and we justify our choice by cross-validation. We generate our data by random sampling time intervals from the entire price history of a particular state/county/zip-code region, where all price values except for the last 12 months are features, and therefore the last few prices will be the targets. Linear regression treats the input features as independent variables, while the Prophet model considers the time sequence using non-parametric methods and smoothers, so we expect better results from the Prophet because the price trend indeed forms a sequence where the future depends on the past. However, we need to validate this conjecture by doing experiments and obtain comparable MSE metrics.

4.5 Analysis of estate investment prospects in certain areas. We will use the predicted future estate value to sort out the most valuable areas. To achieve this goal, we will utilize two major indicators: expected price percentage increase and typical mean-squared error (MSE) from our machine learning model's prediction. The larger the price increasing percentage, the more attractive the estate is to the investors. On the contrary, a larger MSE typically indicates low confidentiality from our model which can be attributed to market volatility, so we should be dubious about these results. We will divide price increase percentage factor by the MSE factor to get a final score, which is very similar to the idea of Sharpe Ratio. Our investment prospect rankings of the regions are determined by sorting these scores. Figure 7 is an example of visualizing Georgia's top 10 counties that are worth investing in provided by our tool.



Georgia

Rank	County
1	cherokee, ga
2	fulton, ga
3	forsyth, ga
4	dekalb, ga
5	liberty, ga
6	gwinnett, ga
7	newton, ga
8	henry, ga
9	jackson, ga
10	paulding, ga

Figure 7: Georgia's Top 10 Counties to Invest in.

Intuitions

We think our approach is better than the state-of-the-art methods in the following ways:

- Use of concentric circles to help users visualize the average prices within some distance.
- Compare the estate price trend (current and future) of multiple states in the same line chart.
- Experiment multiple ML models (linear regression and Prophet model/Addictive regression model) to predict the future price trend and use various metrics to validate the predictions.
- Use of ML results to derive weighted scores besides visualizing trends to help users make investment plans. Human evaluations will prove the effectiveness.

5 EXPERIMENTS

We would like to quantify the performance of our platform and the effectiveness of our machine learning models. We design the following questions to answer:

1. How much time does our system spend fetching data before displaying them?
2. How accurate are our ML models in terms of predicting estate prices in the future?
3. Is our investment recommendation / ranking scheme effective?

Thus, we define the project's success based on factors below.

5.1 System Scalability

Given that there are enormous regions in the country and we have hundreds of timestamps (all months between 2015 and 2021), the tool needs to deal with millions of rows of raw data. We tackle the challenge of good scalability by properly partitioning using months and regions and creating API calls between the visualization end and the data processing end. To verify our idea, we measure the average latency of typical user interactions with our final web application.

As we see from Table 1, requesting the whole table for a visualization sub-task is both wasteful and impractical when interacting with users. Proper partitioning often results in an acceptable latency that is several magnitudes faster than the whole table call, improving overall data use efficiency at the same time.

Task	Mean Time
Whole table (states)	1510 ms
Whole table (counties/zipcodes)	> 1 min
Price history for one region	38 ms
All state prices of a month	36 ms
Monthly county prices in a state	57 ms
Monthly zipcode prices in a state	105 ms

Table 1: API call latency with&without Partitions

5.2 Price Prediction Error

There is an ML component in our project which predicts the future real estate price trend for a particular region: state or county levels. Although we have no knowledge of how the housing market will behave in the future, we can still utilize the existing data in the dataset: we may ask the models to forecast the median house price of Georgia in September 2020, given the price history and some metrics since September 2019. We then compare the model's predicted data with the existing data to examine how well our models perform. We trained two different ML models in this project: linear regression model and additive regression model. We obtained the following metrics for both models shown in Table 2.

Since the Additive Regression model has a lower mean error and a higher R^2, it seems to have a better performance in predicting the future house prices. However, to further examine two models, we conduct the following experiments.

5.3 Recommendation Effectiveness

One major functionality of our tool is to help people make real estate investment decisions. Our recommendation system considers a region with good investment potential if it has a high score, which is computed using the price increase and model prediction error as we mentioned in section 4.5. We compare this strategy to how people actually make investment decisions: we first hide the model's recommendation list but only display the relevant visualizations, and then we ask the users to jot down their own list of region choices. We let both ML models evaluate and sort the first 10 counties for each state that has the highest real estate investment potential and compare these counties with the rank in Niche.com, a reliable ranking and reviewing site. The accuracy is measured by how well our list intersects with the list provided in Niche.com. The results are shown in Table 3.

Model	**Avg. MSE**	**Mean Error**	R^2
Train			
Linear	928.47 M	30.47 k	0.9445
Additive	191.51 M	13.84 k	0.9677
Validation K = 4			
Linear	950.49 M	30.86 k	0.9133
Additive	249.01 M	15.78 k	0.9478

Table 2: Evaluation metrics of ML models

Model	**Matches**	**Total**	**Accuracy**
Linear	303	500	60.6%
Additive	355	500	71.0%

Table 3: Recommendation match accuracy

6 CONCLUSION

In this project, all team members have contributed a similar amount of effort. The project gathers five years of estate prices from 2016 to 2021 covering most of the regions in the United States. Furthermore, the predicted 2022 prices of those estates are reasonable. Generally, this project gives the users an excellent intuitive feeling. Users are able to consider different angles through the different types of diagrams. Specifically, users can filter to monthly change together with the sharp ratio for investment purposes or sift to the absolute median price and the concentric circle diagram for other purposes related to a limited budget. The series of diagram displays give users more mature thinking and analysis, which dramatically reduces the research time required. Overall, the project provides users great convenience and offers more time for other essential life tasks.

REFERENCES

[1] Sumit Agarwal, Ying Fan, Daniel P McMillen, and Tien Foo Sing. 2021. Tracking the pulse of a city—3D real estate price heat maps. *Journal of Regional Science* 61, 3 (2021), 543–569.

[2] MS Cheryshenko and Yu Yu Pomernyuk. 2021. Integration of big data in the decision-making process in the real estate sector. In *IOP Conference Series: Earth and Environmental Science*, Vol. 751. IOP Publishing, 012096.

[3] Sheng-Min Chiu, Yi-Chung Chen, and Chiang Lee. 2022. Estate price prediction system based on temporal and spatial features and lightweight deep learning model. *Applied Intelligence* 52, 1 (2022), 808–834.

[4] Massara Daana and Mao Lin Huang. 2013. Visual sensitivity analysis in real estate prediction system. In *2013 10th International Conference Computer Graphics, Imaging and Visualization*. IEEE, 100–105.

[5] James R DeLisle, Brent Never, and Terry V Grissom. 2020. The big data regime shift in real estate. *Journal of Property Investment & Finance* (2020).

[6] Xin Ge and Jinson Zhang. 2019. ANALYSE PROPERTY DATA THROUGH VISUALISATION.

[7] Nehal N Ghosalkar and Sudhir N Dhage. 2018. Real estate value prediction using linear regression. In *2018 fourth international conference on computing communication control and automation (ICCUBEA)*. IEEE, 1–5.

[8] S. Havre, B. Hetzler, and L. Nowell. 2000. ThemeRiver: visualizing theme changes over time. In *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings*. 115–123. https://doi.org/10.1109/INFVIS.2000.885098

[9] Theodore Hong. 1999. Visualizing real estate property information on the web. In *1999 IEEE International Conference on Information Visualization (Cat. No. PR00210)*. IEEE, 182–187.

[10] Turgay Kerem Koramaz and Vedia Dokmeci. 2012. Spatial determinants of housing price values in Istanbul. *European Planning Studies* 20, 7 (2012), 1221–1237.

[11] Nikhil Krishna. 2021. Dubai Real Estate Investment: A Predictive and Time Series Analysis. (2021).

[12] Mingzhao Li, Zhifeng Bao, Timos Sellis, Shi Yan, and Rui Zhang. 2018. HomeSeeker: A visual analytics system of real estate data. *Journal of Visual Languages Computing* 45 (2018), 1–16. https://doi.org/10.1016/j.jvlc.2018.02.001

[13] Ping-Feng Pai and Wen-Chang Wang. 2020. Using machine learning models and actual transaction data for predicting real estate prices. *Applied Sciences* 10, 17 (2020), 5832.

[14] Jorge Iván Pérez-Rave, Juan Carlos Correa-Morales, and Favián González-Echavarría. 2019. A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes. *Journal of Property Research* 36, 1 (2019), 59–96.

[15] Karshiev Sanjar, Olimov Bekhzod, Jaesoo Kim, Anand Paul, and Jeonghong Kim. 2020. Missing data imputation for geolocation-based price prediction using KNN–mcf method. *ISPRS International Journal of Geo-Information* 9, 4 (2020), 227.

[16] Iqbal H Sarker. 2021. Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. *SN Computer Science* 2, 5 (2021), 1–22.

[17] Nitin Sharma, Yojna Arora, Priyanka Makkar, Vikas Sharma, and Hardik Gupta. 2021. Real Estate Price's Forecasting through Predictive Modelling. In *Machine Learning for Predictive Analysis*. Springer, 589–597.

[18] GuoDao Sun, RongHua Liang, FuLi Wu, and HuaMin Qu. 2013. A Web-based visual analytics system for real estate data. *Science China Information Sciences* 56, 5 (01 May 2013), 1–13. https://doi.org/10.1007/s11432-013-4830-9