# CX 4242 Report

## United States Real Estate Visualization

### Team 4: The Underdogs

**Intro and Motivation (Heilmeier Q's 2, 4, 5)**

Housing is a huge driving factor of the economy. In 2011, the total value of residential properties in the US was $16 trillion [13]. Any American who owns a home would be interested in how the market will change in their area. Furthermore, investors looking for the hottest place to buy real estate want to know where they should grow their business and where they should avoid.

When an individual wants to sell a house, much of this decision making is done by the real estate agent that sellers hire. The agent will generally conduct a Comparative Market Analysis (CMA) which is based largely on subjective features of the house, leading to inconsistencies in valuation across agents. Furthermore, the S&P Case-Shiller Index represents the prices of homes on a national scale, but these indices are removed from the actual value of any property and do not account for differences in geographic location. In practice, buyers and sellers often are forced to speculate on property prices with little knowledge of how property prices are changing in their area.

Our project will make an impact on the stakeholders by increasing their knowledge of the market and thus increasing their confidence when they make decisions. We can quantitatively measure this impact with user surveys. For example, survey a pool of home sellers and ask how confident they are in their list price; then give the seller our tool to use for five minutes and ask how confident they are again. This is one procedure that we could use to quantify the impact of our tool.

**Literature Survey**

As a housing price prediction model is so important, there has been a significant amount of research done, with a rapidly growing focus on using Machine Learning.

**Non machine learning prediction**

[8] uses a macroeconomic dynamic stochastic general equilibrium model (DSGE) based on behavioral economics. The model relies on a multitude of variables, and, as such, relies on many assumptions about the relationships between those variables. [13] is a review of many other papers that attempt to predict price using variables such as rent-to-price ratio, vacancy rates, and monetary policy and concludes that although these may be accurate for capturing the current state of the market, it struggles to predict future prices. We will design a more simplistic model that only relies on price, encapsulating all the variables and assumptions used in these non-machine learning algorithms. These papers do not give our group much guidance on implementation but give us the opportunity to discuss the alternatives to our method and their shortcomings.

**Machine learning prediction**

Robey et. al. found that a multiple linear regression was able to predict housing prices with an $r^2$ of over 90% [1]. Also, certain regression methods such as Support Vector Regression[15] or Fuzzy Least-Squares Regression[16] are found to generate accurate and efficient real estate price prediction models that do not use excessive amounts of variables. By comparing MSE (mean squared error), we can choose the best model[11]. With various regression models, it's effective to take the weighted mean of different models to get more accurate results[10]. Multiple other papers such as [7] - which studied housing prices in Beijing using a number of physical attributes about the house - and [3] - which used macroeconomic variables to determine house price in London - found that a Random Forest Model most accurately predicted the price in their holdout set, however it may be prone to overfitting. [9] used 27 different variables to run multiple machine learning algorithms to determine closing price of houses. They determined that RIPPER was the best classifier to determine if the closing price was above or below the listed price.

Papers in this section provide us with several different machine learning methods from general, like random forest and XGB[14], to advanced, like merged models and weighted combinations, to adhoc, like RIPPER which is designed for predicting house price. These papers give us ideas for which models can accurately predict real estate data but fall short because we will do time-series analysis on price alone.

### Visualization

One of the important aspects of our project will be the visualization of the predictions from the model. Bao et. al[2] and M.Li[12] created a visualization tool called *HomeSeeker* for users to interactively search for properties based on user-specified requirements.

[4] is about balance between information loss and complexity of the graph when visualizing a large data set. Some papers introduce mainstream tools such as excel and tableau[6]. Other papers go further in detail, like interactive user interface and cluster analysis[5]. There are some powerful visualization tools available as well that only specialize in geographic visualization, such as ArcGis or GeoVista that could provide effective visualization for housing price comparison using representations such as Choropleth map [17]. One powerful visualization method is NDS, an interactive web-based system built with D3 and Mapbox GL JS[18]. Those resources would help us perform a logical and concise visualization for real estate. However, those papers are not designed for time series regression visualization, so some ideas may not be the most appropriate in this project.

## Project Summary (Heilmeier Q's 1, 3, 6, 7, 8, 9)

Our team will collect real estate data to perform time-series regression in order to predict future changes in property values across the United States, and create a visual tool to illustrate our findings.

This project will create a tool that people can use to easily assess how property prices have changed in the past and get an estimate for how much they will change in the future. By leveraging big data visualization, consumers will have a revolutionary way of getting a "feel" for the real estate market in their state over time, with the ability to refine their results based on the type of property that they are concerned with. Ultimately, the tool is unique in that our users should be able to draw conclusions from our data in less than 60 seconds and come out with a better understanding of the real estate market.

The most important risk to note is that users might take our predictions too seriously and potentially lose money, thus getting our team in "trouble." However, the payoff is a potentially huge user-base who are interested in our tool and our predictions. Our data sources are free to use and the prototype should be free to deploy for beta testing. We should have a prototype in less than two months.

## Proposed Method

Prediction is conducted using some sort of time-series regression. Different prediction methods are still being compared. We have so far tried using basic Linear Regression and ARIMA (autoregressive integrated moving average). Visualization will be created using D3.js. We will create a choropleth map giving a visual of real estate prices in each on a color gradient with features like a tooltip and parameter selection.  This approach will be unique in that we combine macro real estate data visualization with advanced machine learning techniques, allowing users to quickly gain insight from historical and forecasted trends in the real estate market.  The state of the art techniques are either report prices on a local level and miss national trends or they report national trends but fail to visualize the data well and incorporate machine learning techniques.

1. Download data from kaggle

We will use this data set:
https://www.kaggle.com/paultimothymooney/zillow-house-price-data?select=City_Zhvi_5BedroomOrMore.csv. To be specific, only 6 files will be used, which are from City_zhvi_1bedroom to City_zhvi_5bedroomOrMore and City_zhvi_AllHomes.
The data set after merging all 6 files would have about 130,000 rows and 300 columns.

2. Organize data

291 columns of the data set is the price at a certain date, and the remaining columns are information about the region, such as region id and region name. Since what we want to predict is the average price of a state, we would only keep the column “State” and ignore the rest.

3. Build a forecasting model

One of the models we are going to use is ARIMA, auto regressive integrated moving average, which is a common model in time series forecasting. ARIMA is composed of three parts: auto regressive(AR), integrated(I), and moving average(MA). AR and MA are two simple time series models, which use the past data to forecast; (I) is a tool to smooth the time series.

4. Train model

For the ARIMA model, we would use the ARIMA package in python. The model would forecast the price of every single sample or the average of samples in every state (it will be decided based on the training time and evaluation, since time consuming maybe unacceptable if we train a model for every sample or the result may not be good enough. )

5. Tune model

It will be important to backtest our models to validate them. For the ARIMA model, we can tune the parameters based on line chart, ACF, and PACF graph of the sample:
d(parameter of I): based on the line chart, we can see the data is smooth. We decided to use $d = 1$ as $d = 0$ is not smooth enough, while $d = 2$ has no significant improvement compared to $d = 1$.
p(parameter of AR): we drawed a ACF chart, and we observed tailing off, which means 0 should be a good option for this parameter. However, $p = 0$ is not performing well in our experiment, so finally we decided to use $p = 1$.
q(parameter of MA):  similar to tuning of p, but this time we need to look at the PACF chart. We observed cutting off at the second lag, so we decided to use $q = 2$.
The final parameters of arima is $p = 1$, $d = 1$, and $q = 2$

6. Build a user interactive web app using d3.js.

An interactive visualization will be created using HTML, Javascript and the D3 Javascript library. We will create a choropleth US map with several features. We will use GeoJSON to generate the map. Feature 1 is a dropdown list for the users to select the level of exploration: 1 bedroom, 2 bedrooms, 3 bedrooms and

4+ bedrooms. Feature 2 is hovering on one state should show an interactive line chart below it showing the median prices in the last 10 years, and the predictions. Feature 3 is the color scale of each state will be based on how much the property's prices are expected to increase in percentage with respect to last year.

7. Connect webapp to historical and predicted data.

Organize csv files into a folder containing the original information from Kaggle. Use our trained model to generate organized csv files representing future predicted real estate data for each state and home type. Load this file into the d3 code using the d3 dsv methods. DSV library contains useful methods for visuals.

8. UI/UX revision

After getting feedback from users we will update the visualization to generate better user insights.

List of innovations:

- Time series machine learning analysis on macro real estate prices on a state level.
- Interactive choropleth map with user parameters incorporating historical and future predicted data published in a web tool.

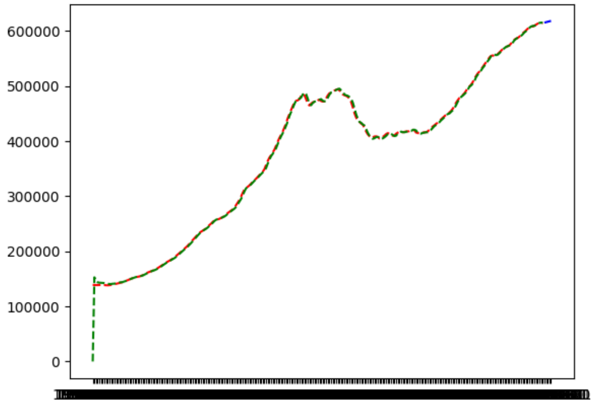
## Experiments

### First Experiment: Proof of Concept

Questions:

1. Is the ARIMA model a good option to pursue for the project?
2. How hard is it to train the ARIMA model to our data?

In our first test we conducted the Autoregressive Integrated Moving Average (ARIMA). The following graph is an example of ARIMA forecasting(blue at the end) and prediction(green) of one sample.



We would use some common evaluation criterion on regression, such as MSE or R square, on the model to tune our model.

### Second Experiment: Model Validation

Questions:

1. How accurate is the ARIMA model at predicting new data?
2. How can we tune the hyperparameters of the model we chose?

In order to validate our model we will train it without some time series data and then see how it performs with that missing data. For example we could hold back data from 2020 and then test the model on that data. Then we compare predictions and actual results to measure the accuracy of our model.

### Third Experiment: User Impact

Questions:

1. How much value do users derive from our tool?

2. What are users' feedback from our visualization?
3. How much do users learn by using our tool?

In order to measure the impact of our product, our group conducted a user survey with students at Georgia Tech. The procedure was as follows: a new user will be asked on a scale from 1 to 10 how confident they are in their understanding of the national real estate market; then the user will be given our tool to use however they like for 10 minutes; then they will be asked to rate their confidence again on the same scale. This will measure the impact of our tool on the user and serve as a proof-of-concept for our product.

The drawback of this method is that Georgia Tech students are not our primary stakeholders and they are not really our intended market for the product. However, for a proof-of-concept and with the time constraint of the class, this survey base is the most accessible for our team.

**User Impact Survey Results:**

Our group performed the user impact survey as described above. We surveyed 16 participants and measured the difference in their knowledge of the U.S. real estate market before and after using the web tool. On average, users reported their knowledge of the market increasing by 3.72 points on the 1-10 scale. The majority of participants reported their knowledge increased by 2 or 3 out of 10 points. Before using the tool, participants had an average knowledge of 2.7 points out of 10 and after using the tool participants reported an average knowledge of 6.4 points out of 10.

This user impact survey proves that our tool provides valuable insights into the United States real estate market. Every user reported an increase of their knowledge in this market which is proof that this tool could be valuable to a wide range of potential users.

<u>Plan of activities</u>

| Name | Has done | Has done (revised) | Will do | Will do (revised) |
|---|---|---|---|---|
| William Lusty | Research, Helmeier questions writing | Created the scikit learn Linear Regression Model | Database design, model training, data wrangling | Web tool/visualization, generating predicted data, model testing |
| Jun Hyuk Jeon | Research, Proposal Writing | Worked on the report | Model training and visualization | Visualization |
| Matthew Pleskow | Research, Proposal Writing | Worked on the report, created graphs for LinearRegression | Model training and visualization | Continue working on model training |
| Haoxuan Huang | Research, helmeier question | Created the base ARIMA model, worked on the report. | Model training and visualization | Model training and tuning |
| Leonardo Calizaya | Research, literature review | Worked on the report | Model training and visualization | Visualization |
| Michael Cho | Research, Proposal | Worked on the report | Visualization | Visualization |

*All team members have contributed similar amount of effort"

## Conclusion

Overall, this project was challenging but successful. Our team had a lofty goal in mind, which was to build a visualization that leverages machine learning to provide insights in the real estate market. We are proud of the final product and the feedback we have received from user testing has been positive. verall, we created an interactive web tool that allows users to explore historical and forecasted trends in the U.S. real estate market on a macro level, accomplishing most of our vision in doing so.

Our team overcame many obstacles in completing the project. We learned about developing code in a medium-sized team, which was challenging at first. We faced tough decisions when it came to dividing and conquering this project. Furthermore, due to the pandemic our team was collaborating 100% remotely which presented unique challenges. As full-time students, we had to juggle differing schedules and meet at unusual times over Discord throughout the semester. Our team learned quickly that we needed to leverage each of our individual strengths to tackle each phase of the project efficiently. We also shifted our approach to starting each phase of the project several weeks before the deadline to give our group ample time to complete our tasks.

So what's next with this project? Our team has already created a GitHub repository that we've been using to manage the project. We have decided to open-source the project and make the codebase available to the public so that we may show off our tool and results. We believe that real estate data visualization has a strong demand and not many companies are supplying these types of insights. Even if our tool is not downloaded and used millions of times, we believe that the project is a great example of the potential that machine learning and data visualization tools have in the real estate space. The fact that our team was able to create such a powerful and insightful tool in just one semester shows how much potential there is for data scientists to create products in the real estate segment.

This project gave our team a great opportunity to apply the tools we learned in CX 4242. It was a great way to wrap up the class we're proud of the final product.

- # References

[1] S. Robey, M. McKnight, M. Price and R. Coleman, "Considerations for a Regression-Based Real Estate Valuation and Appraisal Model: A Pilot Study", *Accounting and Finance Research*, vol. 8, no. 2, p. 99, 2019. Available: 10.5430/afr.v8n2p99 [Accessed 12 October 2021].

[2] M. Li, Z. Bao, T. Sellis and S. Yan, "Visualization-Aided Exploration of the Real Estate Data", *Lecture Notes in Computer Science*, pp. 435-439, 2016. Available: 10.1007/978-3-319-46922-5_34 [Accessed 12 October 2021].

[3] S. Levantesi and G. Piscopo, "The Importance of Economic Variables on London Real Estate Market: A Random Forest Approach", *Risks*, vol. 8, no. 4, p. 112, 2020. Available: 10.3390/risks8040112 [Accessed 12 October 2021].

[4] E. Gorodov and V. Gubarev, "Analytical Review of Data Visualization Methods in Application to Big Data", *Journal of Electrical and Computer Engineering*, vol. 2013, pp. 1-7, 2013. Available: 10.1155/2013/969458 [Accessed 12 October 2021].

[5] M. Khotilin and A. Blagov, "Visualization and cluster analysis of social networks", 2021. .

[6] S. Ali, N. Gupta, G. Nayak and R. Lenka, "Big data visualization: Tools and challenges", *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, 2016. Available: 10.1109/ic3i.2016.7918044 [Accessed 12 October 2021].

[7] Q. Truong, M. Nguyen, H. Dang and B. Mei, "Housing Price Prediction via Improved Machine Learning Techniques", *Procedia Computer Science*, vol. 174, pp. 433-442, 2020. Available: 10.1016/j.procs.2020.06.111 [Accessed 12 October 2021].

[8] Y. He and F. Xia, "Heterogeneous traders, house prices and healthy urban housing market: A DSGE model based on behavioral economics", *Habitat International*, vol. 96, p. 102085, 2020. Available: 10.1016/j.habitatint.2019.102085 [Accessed 12 October 2021].

[9] B. Park and J. Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data", *Expert Systems with Applications*, vol. 42, no. 6, pp. 2928-2934, 2015. Available: 10.1016/j.eswa.2014.11.040 [Accessed 12 October 2021].

[10] A. Varma, A. Sarma, S. Doshi and R. Nair, "House Price Prediction Using Machine Learning and Neural Networks", *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2018. Available: 10.1109/icicct.2018.8473231 [Accessed 12 October 2021].

[11] G. Milunovich, "Forecasting Australia's real house price index: A comparison of time series and machine learning methods", *Journal of Forecasting*, vol. 39, no. 7, pp. 1098-1118, 2020. Available: 10.1002/for.2678 [Accessed 12 October 2021].

[12] M. Li, Z. Bao, T. Sellis, S. Yan and R. Zhang, "HomeSeeker: A visual analytics system of real estate data", *Journal of Visual Languages & Computing*, vol. 45, pp. 1-16, 2018. Available: 10.1016/j.jvlc.2018.02.001 [Accessed 12 October 2021].

[13] E. Ghysels, A. Plazzi, R. Valkanov and W. Torous, "Forecasting Real Estate Prices", *Handbook of Economic Forecasting*, pp. 509-580, 2013. Available: 10.1016/b978-0-444-53683-9.00009-8 [Accessed 12 October 2021].

[14] J. Manasa, R. Gupta and N. Narahari, "Machine Learning based Predicting House Prices using Regression Techniques", *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, 2020. Available: 10.1109/icimia48430.2020.9074952 [Accessed 12 October 2021].

[15] J. Y. Wu, “Housing price prediction using support vector regression,” *San Jose State University*, May 2017. Available: 10.31979/etd.vpub-6bgs [Accessed 12 October 2021].

[16] A. G. Sarip, M. B. Hafez, and M. N. Daud, “Application of fuzzy regression model for real estate price prediction,” *Malaysian Journal of Computer Science*, vol. 29, no. 1, pp. 15–27, 2016. [Accessed 12 October 2021].

[17] M. Nöllenburg, “Geographic visualization,” *Human-Centered Visualization Environments*, pp. 257–294. [Accessed 12 October 2021].

[18] Y. Lan, E. Delmelle, and E. Delmelle, “NDS: An interactive, web-based system to visualize urban neighborhood dynamics in United States,” *Journal of Maps*, vol. 17, no. 1, pp. 62–70, 2021. [Accessed 12 October 2021].