Austin Real Estate Analysis

Team #65: Team Austin

Jianpang Luo, Panithan Floyd, Robert Kendall, Tony ElHabr, Xikai Zhao
November 30, 2019

1 Table of Contents

2 Introduction and Motivation

Four of the five members of our group reside in Austin, Texas, and each of the current residents would like to purchase a personal residence or rental property in the future. Although there are several popular websites dedicated to real estate listings, none to our knowledge provide long term estimates of future home values or provided interactive visualizations focused on comparing relative future growth between zip codes. The goal of our project is to fill this gap through an intuitive and interactive visualization so that future homebuyers and investors in the Austin real estate market can make better decisions based on data-driven analysis.

2.1 Problem Definition

How can we help buyers and investors make more informed, cost-effective decisions regarding purchases of real estate purchases in Austin, Texas?

2.2 Survey

With the median home price in the US now above $320,000 as of Q2 2019, it is important for future buyers to find a home that fits their needs and more importantly, budget. Real estate price prediction has been a topic of study for decades, but only recently has publicly available housing-related data from Zilllow, Trulia, and Redfin democratized this research. According to Lowrance (2015), multiple linear regression models using descriptive predictors—number of bedrooms and bathrooms, days for sale [1], etc.— was once state-of-the-art. In his study of Los Angeles, however, he found random forests models outperformed previous baseline models [2]. Nguyen (2018) found support vector regression, leveraging similar features were also superior to baseline models [3]. For our purposes, these studies demonstrate that descriptive features and well-crafted regression models should provide a reasonable prediction accuracy for our target city of Austin, Texas.

Unsurprisingly, geo-spatial information supplements descriptive features in real estate price assessment. Using street maps of houses in Los Angeles, Bin et al. (2019) developed a convolutional neural network to derive features quantifying the graphical presentation of houses for their random forest model [4]. Heyman and Sommervoll (2019) explored relative location, i.e. distances to notable places like parks, public transportation, etc, and found typical geographical factors like zip code and census tracts add much less explanatory power in the presence of additional features describing relative position to areas of interest [5]. While the actual numerical findings of these studies cannot be compared directly with our results (because they did not study Austin), they do validate the significance of geography for price prediction.

Several researchers have found that economic variables bolster models. Park and Bae (2014) incorporated federal housing price indices, historical mortgage rates, and public school ratings into a model for house appraisal in Fairfax County, Virginia [6]. Rafiei and Adeli (2015) explored time-dependent and seasonality in similar economic indices [7].

Other collaborators leveraged non-traditional features to model housing prices. Ahmed and Moustafa (2016) used a neural network incorporating visual information—photos of the same locations in each house—along with traditional descriptive information to outperform baseline models [8]. Poursaeed et. al. (2019) developed a convolutional neural network using a larger data set with less structure to outperform Zillow's estimates [9]. Lynn Wu and Erik Brynjolfsson (2013) used Google searches to predict changes in housing prices and sales [10]. While these approaches are worthy of inspiration, they have shortcomings. For example, the image-based work of Ahmed and Moustafa and Poursaeed et. al. suffers from the problem of "equality" among observations, i.e. sub-par image quality may deflate a model's evaluation. Nonetheless, we recognize the ingenuity of these methods.

Finally, we note that there is an abundance of open-source examples for visualization and machine learning for real estate prices prediction. Nagel (2019) focused on differences between scientific and informational scopes [11] and Agnone (2019) highlighted the importance of interactivity [12]. Alejandro (2018), Jermain (2019), and Haseeb, Durrani, et al. (2019) shared code for various machine learning techniques, highlighting the advantages and disadvantages of each. In particular, Alejandro evaluated regression trees, k-nearest neighbors, and neural networks for listings in Spain [13], Jermain used bootstrapped regression for predictions for Kaggle's 2017 Zillow-based competition [14], and Durrani, et al. employed gradient boosting for Russian housing prices [15]. While such work cannot be directly applicatied to ours due to the differences in data sets, their code has provided a guide for model experimentation and implementation.

3 Proposed Method

We developed an interactive dashboard that visualizes historical and predicted monthly changes of real estate value, quantified by Zillow Home Value Index (ZHVI), for 42 zip code areas in Austin, Texas. The visualization is built using Tableau, while all programming tasks—data scraping, data cleaning, model fitting, model evaluation, and evaluation visualization (see the Experimentation and Evaluation section)—were performed with the R programming language.

We explored using several data types and models, such as using various economic and geographical factors to build a predictive regression or random forest mode. However, the biggest challenge with this approach was obtaining enough data for a viable model. For example, U.S. census data only has economic data by zip code for 2012 and 2014. After a period of research, we decided on using a time series predictive model based on the ZHVI due to its simplicity and ease of use.

The historical data for our application is sourced from Zillow, an online leader in real estate marketing. Specifically, we used eight different time series of ZHVI values. The ZHVI represents a dollar value estimate of the median home value across a given region and housing type. Zillow claims ZHVI is the most accurate measures of residential real estate prices available [16]. The eight series used were 1-bedroom, 2-bedroom, 3-bedroom, 4-bedroom, 5 or more bedrooms, condominiums, and two aggregate categories — single family residence and “All”. Each time series contained monthly ZHVI values for Austin area zip codes dated between April 1996 and September 2019. Altogether, the scraped raw data consists of 516,355 observations across eight files that total over 140 MB in size. When cleaned and filtered (to the 42 Austin zip codes), the data consists of just over 49 thousand records.

Predictions for ZHVI values beyond September 2019 were made for every combination of zip code and ZHVI series [17] using Facebook’s open-sourced prophet package. Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality. We chose to use the prophet package because it can generate robust, automated (i.e. no need for manual parameter tuning) time series forecasts. The package is also able to account for missing data, and outliers effectively [17]. Overall, 174 separate prophet models were trained “offline” using the `prophet` function and default settings except for the `daily.seasonality` and `weekly.seasonality` parameters which were explicitly set to `FALSE`. Note that not every zip code has data for each ZHVI category;

otherwise we would expect there to be 42 * 8 = 336 models. The model predictions and historic data were stored in static CSV to be used in for the user-facing deliverable, i.e. the interactive dashboard.

The dashboard features two “linked” charts that are responsive to user input. The first chart is a heatmap that illustrates predicted percent change in ZHVI values for each Austin zip code. If the user hovers or clicks their mouse on a specific zip code on the map, the area is emphasized and a tool-tip popup will display details such as predicted percent change in value since September 2019 and how the predicted percentage change ranks compared to all other zip codes. Additionally, the user can filter down to specific property types (eg. 3-bedroom vs. 4-bedroom) and different investment time horizons(eg. 2020 up to 2030) to meet their specific needs. All aspects of the visualization update based on user inputs and changes.

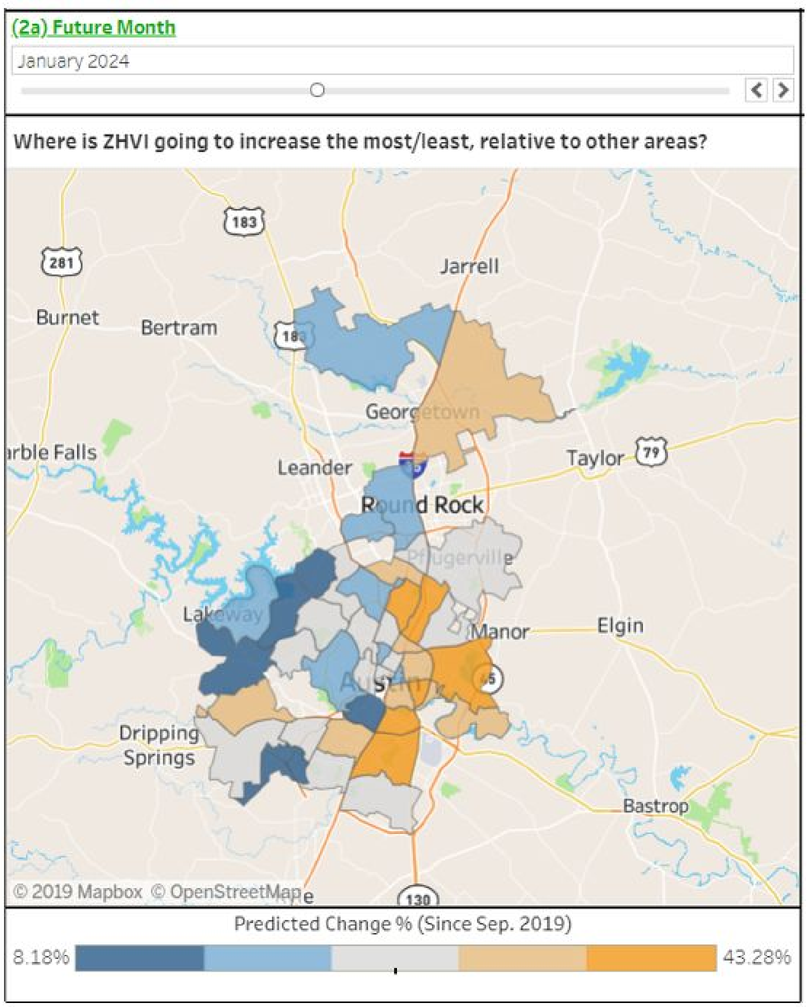

Figure 1: Heat Map Visual - Predicted ZHVI change at January 2024

The second visual is a line chart displaying a time series of the selected home type for all zip codes. If the user hovers their mouse over a zip code in the map, the time series of that zip code is also emphasized via color differentiation and increased line width. This chart gives users a perspective on how prices changed in the past and how we predict they will trend in the future.

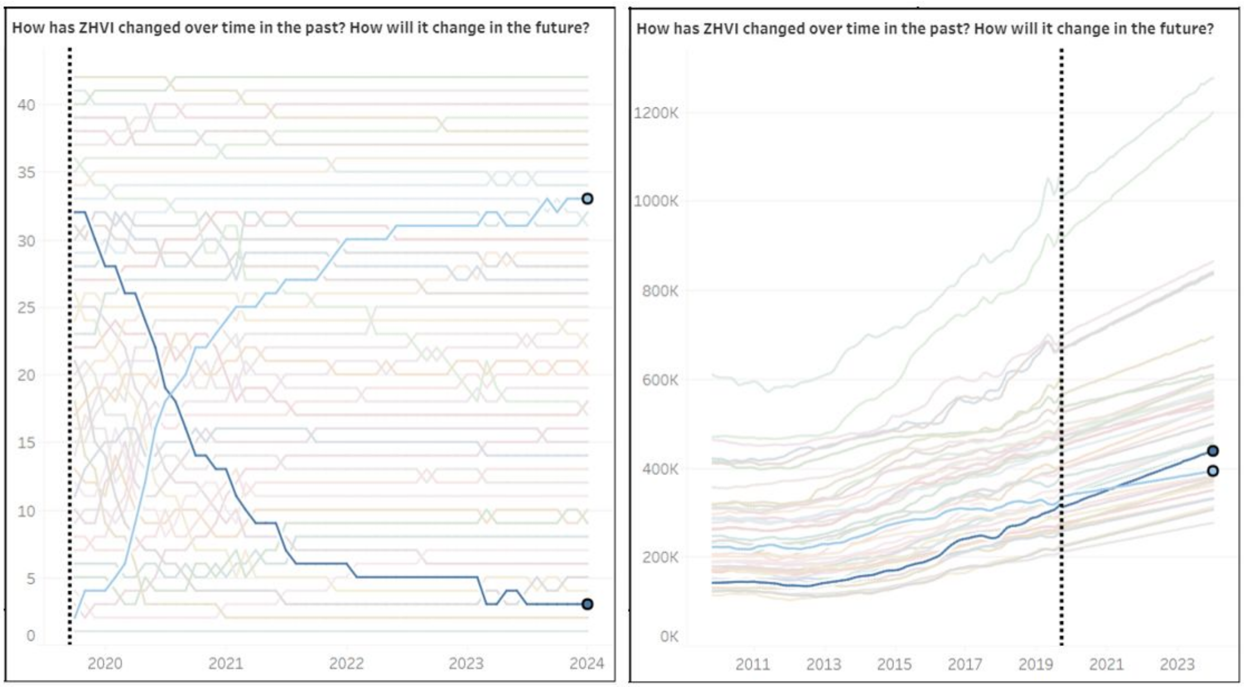

Figure 2: Line Chart Visual - Comparing a high growth and low growth zip code

3.1 Innovations

While there are several real estate marketplaces, e.g. Zilllow, Trulia, and Redfin, that provide similar services, i.e. web interfaces for visualizing price change for real estate, our approach is novel in several ways. First, instead of focusing on individual listings, our method aggregates real estate features to the zip code level. The idea is to point users to a general area—as opposed to suggesting individual properties—where real estate investment would make the most economic sense. Second, we are analyzing a specific city (Austin, Texas) which should make our solution more robust than a state of the art model generalizing to multiple areas. Based on our literature survey, many findings suggest that models specializing in a specific region tend to perform better than a general model. Austin's rapid population growth, low housing inventory, growing tech sector, limited public transportation options, etc. are all reasons why a Austin specific model is needed. Moreover, our team's familiarity with the area would provide domain knowledge that a more general model may be lacking.

Overall, we believe that our approach benefits from the limited scope of the study. Our aggregation approach and focus on a specific city strikes a balance between granularity and optimizing for scalability. We simultaneously avoid the pitfalls that may come with becoming fixated with individual listings (i.e. the issue of granularity) and the pitfalls that may come with an approach that works well in a broader scope, e.g. for the entire United States, but not for a given smaller scope, e.g. for Austin (i.e. the optimizing for scalability).

4 Experiments and Evaluation

To better gage the design, usability, and usefulness of our dashboard, we surveyed six other CSE2642 teams during development. Specifically, we created an online survey and asked the other teams to provide feedback after hands on use of our dashboard. Tables in the Appendix summarize the questions asked and the feedback received. The overall response from the survey was positive (average score of 3.8/5.0), with most respondents indicating that the dashboard is easy to use while also providing value. Based on the responses, we paid particular attention to the free-form feedback in improving our final deliverable.

Beyond user feedback, we also quantified the robustness of the time series models predictions shown in the dashboard using the prophet package's `cross_validation()` and `performance_metrics()` functions. `cross_validation()` was called twice for each series: (1) with a forecast horizon of 1 year (i.e. `horizon = 365.25` with `units = 'days'`) and defaults for all other parameters (`init = 3 * horizon` and `period = 0.5 * horizon)`; and (2) with a forecast horizon of 10 years and an initial time training period of 1 year (i.e. `init = 365)`. The plots and tables below illustrate the results of the mean absolute percentage error (MAPE) of the cross validation (CV) results, as calculated by `performance_metrics()` with `metrics = 'mape'` and `rolling_window = -1` (i.e. no averaging across data points). (When leaving `rolling_window` at its default value of `0.1`, i.e. averaging across 10 points, similar results were found.)

To provide more clarity about how these function calls and parameters should be interpreted, take the following example. In the case of (1) above—where `horizon = init = 3 * horizon`, etc.—a prophet model is first trained on the first 36 months (i.e. three years) of data, and predictions are made for the month that is 12 months in the future (i.e. 48 months from the first month in the data set). Next, the prophet model is trained on the the 36 months that follow the first 6 months of data (excluding the first 6 months), and the predictions is made for the month that is 54 months (i.e. 48 + 6 months) from the first month in the data set. This process is repeated until all data is processed.

The “ridge” plots below are functionally similar to histograms—they show the profile of a distribution. They are ordered to provide intuition about how CV MAPE changes with increasing levels of aggregation (or, put another way, with decreasing layers of filtering).

1. Figure 3 shows the distribution of CV MAPE by forecast horizon for one zip code and one ZHVP series.
2. Figure 4 is the same as Figure 3, except all eight ZHVI series are included for only one zip code. (The case where all zip codes are shown for one ZHVI series would also be in this category of aggregation.)
3. Figure 5 shows CV MAPE for all 42 zip codes and all eight ZHVI series.

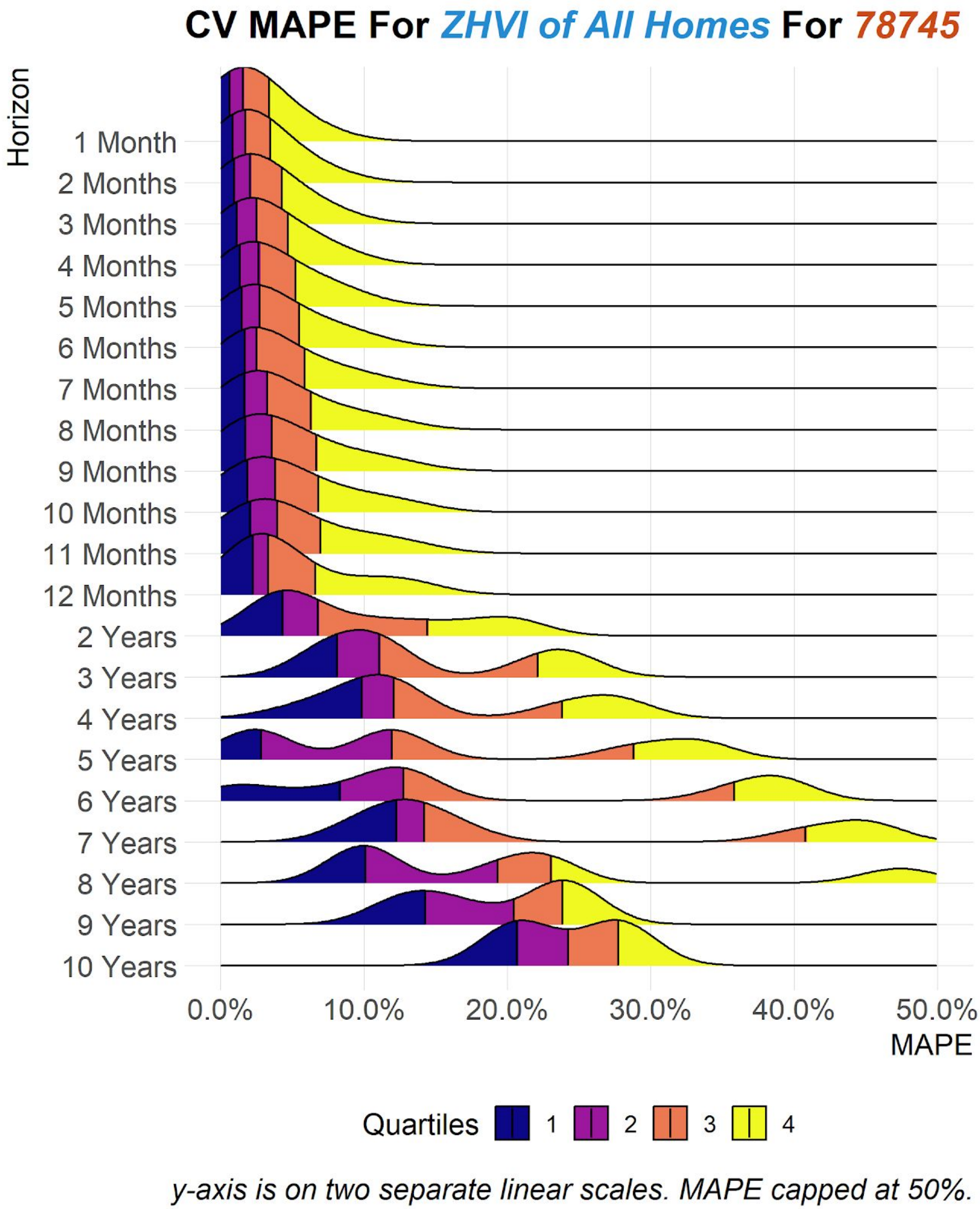

Figure 3: CV MAPE for One Zip Code and One ZHVI Series

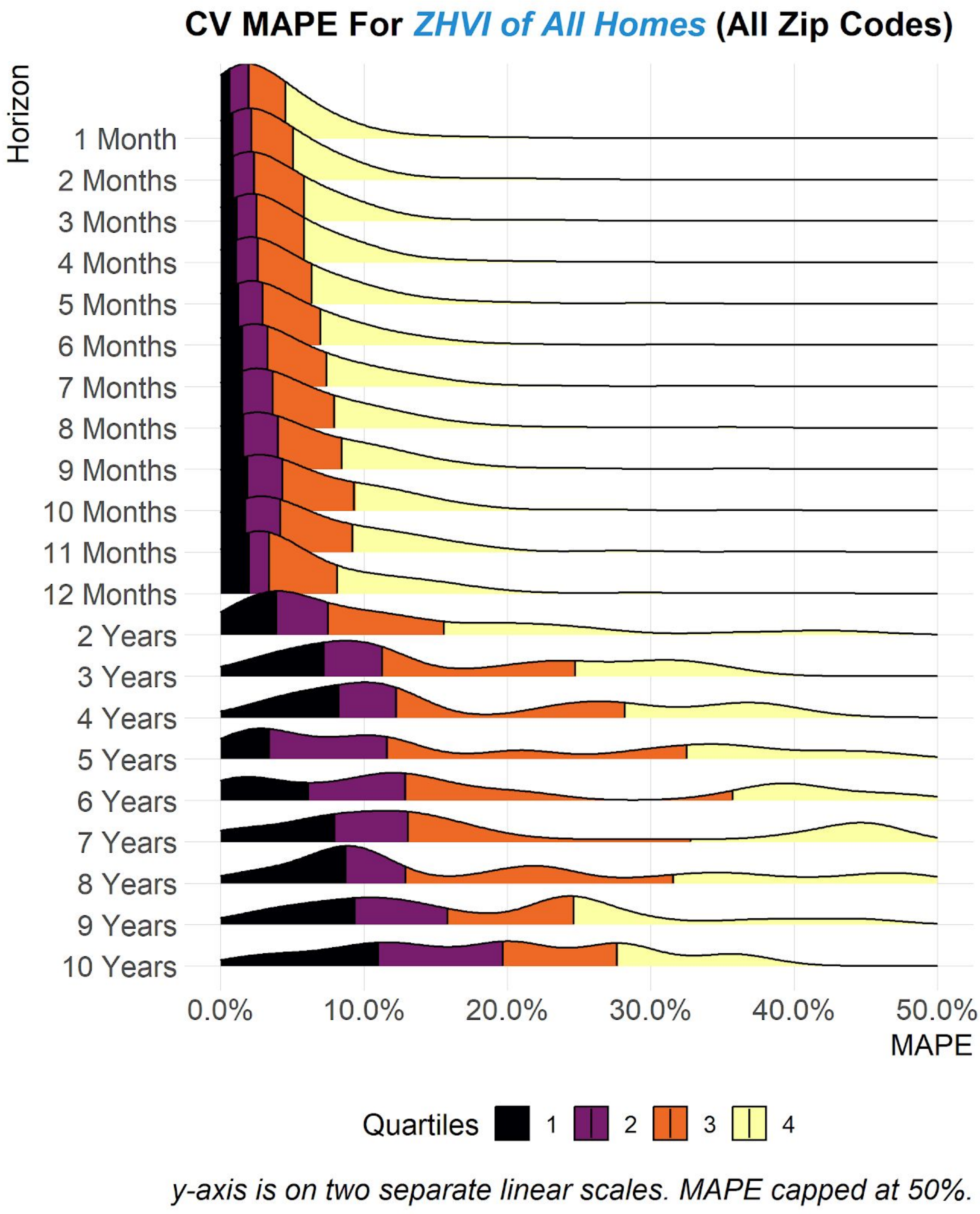

Figure 4: CV MAPE for All Zip Codes and One ZHVI Series

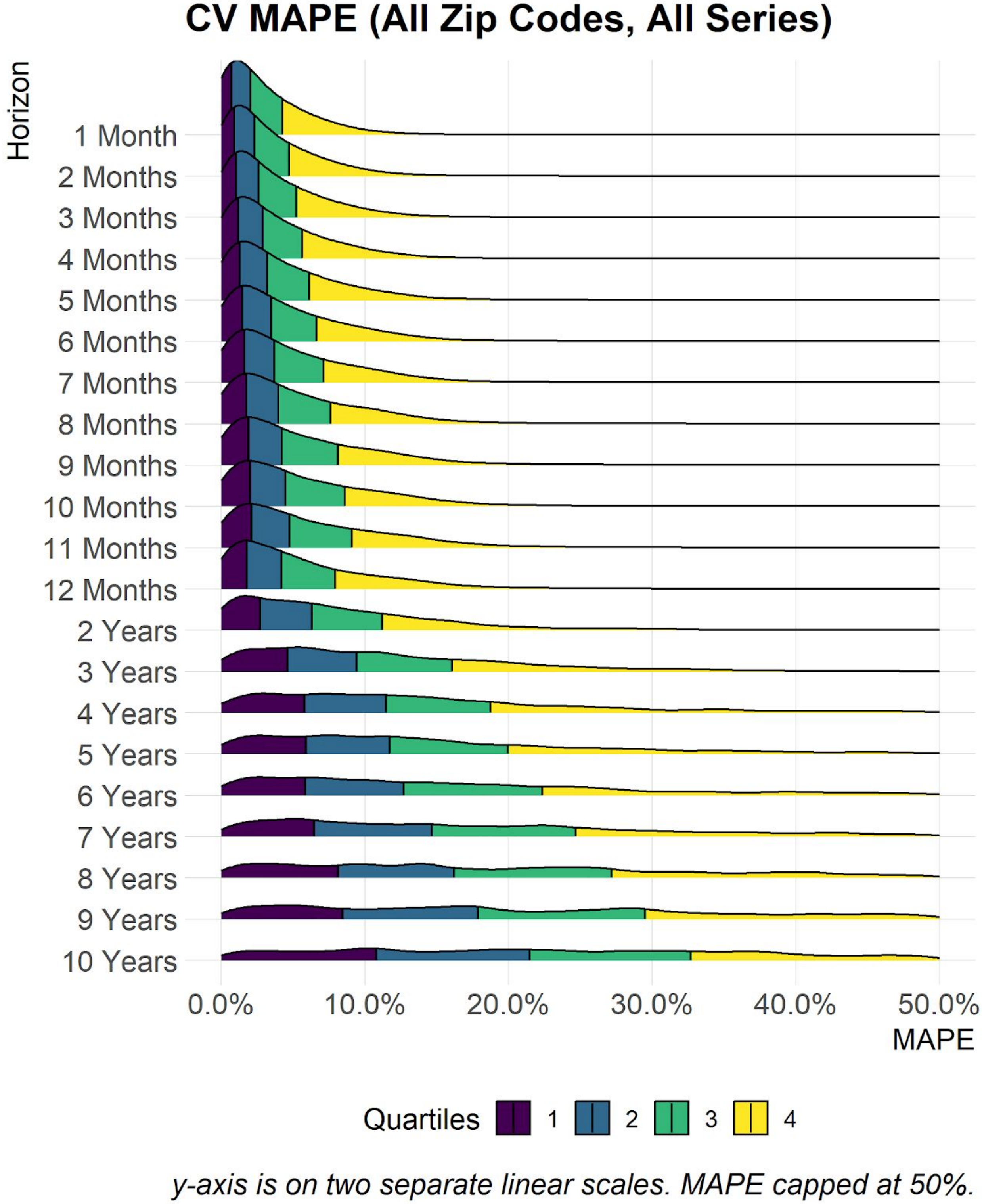

Figure 5: CV MAPE for All Zip Codes and All ZHVI Series

We observe that the MAPE of the forecasts becomes much more variable with increasing time horizons and we expect this remain true for each housing type and aggregate group. In particular, note that the unimodal profile of MAPE distributions for horizons within one year is distorted for horizons beyond a year. Additionally, It is evident that the distributions become more "stable" as the level of aggregation increases, i.e. the distributions become “tighter” when going from Figure 3 to Figures 4, then again when going from Figure 4 to Figure 5.

The tables below provide summary statistics for the CV MAPE by (1) horizon, (2) series, and (3) zip code, separately. (Maximum and minimum values for each mean are emphasized with red and blue text.) For brevity, only the top and bottom five zip code areas are tabulated in Table 3.

Horizon	Count	Min	Max	Mean
1 Month	1,689	1.20%	6%	**3.00%**
2 Months	1,689	1.39%	7%	3.30%
3 Months	1,521	1.56%	8%	3.70%
4 Months	1,696	1.62%	7%	3.90%
5 Months	1,529	1.97%	8%	4.20%
6 Months	1,696	1.97%	10%	4.50%
7 Months	1,696	1.95%	11%	4.90%
8 Months	1,530	2.03%	11%	5.10%
9 Months	1,696	2.03%	11%	5.50%
10 Months	1,530	2.19%	11%	5.80%
11 Months	1,696	2.27%	12%	6.10%
12 Months	1,698	2.45%	16%	6.20%
2 Years	6,000	2.65%	50%	6.50%
3 Years	6,000	2.82%	85%	10.00%
4 Years	6,000	2.86%	121%	13.40%
5 Years	6,000	2.91%	157%	15.90%
6 Years	6,000	3.08%	192%	17.80%
7 Years	6,000	3.01%	231%	20.70%
8 Years	6,000	3.85%	250%	24.00%
9 Years	6,000	3.70%	271%	26.80%
10 Years	6,000	3.72%	317%	**30.90%**

Table 1: Summary of CV MAPE By Forecast Horizon

Series	Count	Min	Max	Mean
1 Bedroom	485	3.10%	47%	**18.99%**
2 Bedroom	3,565	1.85%	58%	14.95%
3 Bedroom	10,740	1.20%	52%	**13.24%**

4 Bedroom	9,420	1.73%	317%	17.90%
5 Bedroom Or More	2,816	1.48%	84%	16.03%
All Homes	18,480	0.82%	67%	14.63%
Condominium	9,680	2.12%	50%	14.12%
Single Family Residence	18,480	1.15%	64%	14.18%

Table 2: Summary of CV MAPE By ZHVI Series

Zip	Count	Min	Max	Mean
78660	2,200	1.72%	36%	**8.90%**
78754	1,760	2.59%	23%	9.30%
78753	1,760	1.67%	28%	9.50%
78736	880	2.01%	22%	9.50%
78758	2,200	1.15%	31%	9.50%
78747	1,365	1.20%	38%	9.70%
78751	2,200	2.20%	64%	18.90%
78633	880	3.00%	44%	19.30%
78717	1,320	2.34%	49%	19.30%
78739	1,236	1.48%	91%	19.90%
78722	880	2.91%	55%	**22.50%**

Table 3: Summary of CV MAPE By Zip Code (Top and Bottom Five Zip Codes According to Mean)

We make the following observations.

- Table 1 shows that mean MAPE across time horizons increases in a relatively linear fashion between 1 month and 12 months, as well as between 1 year and 10 years.
- Table 2 shows that the 1 bedroom ZHVI series has the largest mean MAPE (i.e. the “worst” accuracy), which may be a consequence of its relatively low count of observations (485). The 3 bedroom ZHVI is found to have lowest mean MAPE (i.e. the “best” accuracy)—even lower than the three ZHVI series having more observations.
- Table 3 indicates that mean MAPE ranges from 8.90% to 22.50% across all zip codes, averaged across ZHVI series. If we had not gone as far as to include prediction horizons of ten years, surely this range would be tighter and closer to 0%.

5 Team Member Contributions

All team members have contributed a similar amount of effort, with specific responsibilities distributed as follows.

Group Member	Responsibilities
Jianping	Tableau dashboard, poster report draft(s)
Panithan	data processing, modeling, report video(s) and presentation(s)
Trey	data processing, modeling, report draft(s)
Tony	data collection, data processing, modeling, report draft(s)
Xikai	Tableau dashboard, report videos and presentation(s)

Table 1: Team Member Contributions

6 Conclusions and Discussion

In summary, we have developed an interactive dashboard that depicts change in Zillow Home Value Index (ZHVI) by zip code in Austin, Texas. This dashboard is fed by monthly historical data (from Zillow) and time series predictions (generated with the prophet package) for each of Austin's 42 zip codes and for eight different categories of ZHVI. Mean absolute percentage error (MAPE) was calculated for each time series model at varying horizons using the prophet package's cross validation functionality.

When using our tool and looking at a 5-year horizon with the "All Home" ZHVI series, we make the following inferences.

1. Zip code 78721 (east downtown Austin) seems to have the best overall prospects. Its percent change from Sep. 2019 is 45.82%, which ranks as the best among all zip codes.

Figure 6: Dashboard Example 1 - A zip code (78721) with sustained high potential.

2. Zip code 78738 (southeast Austin) seems to have the worst overall prospects. Its percent change is 9.21%, which ranks as the lowest (42) among all zip codes.

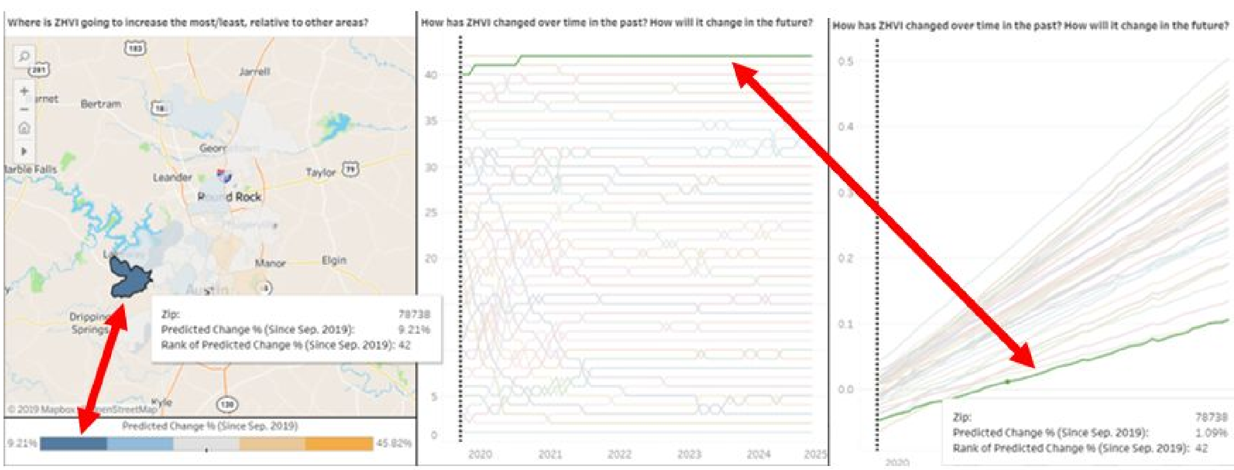

Figure 7: Dashboard Example 2 - A zip code (78738) with sustained low potential.

3. Zip code 78758 (northeast Austin) seems to have the largest increase in potential.

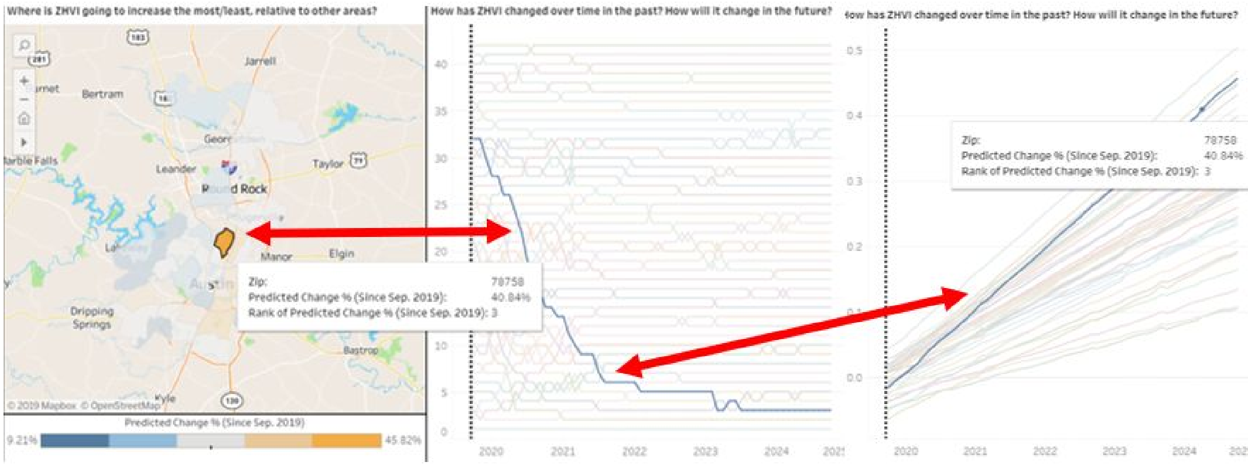

Figure 8: Dashboard Example 3 - A zip code (78758) with a large increase in potential.

4. In general, the east downtown and northeast areas of Austin seem to have the best prospects.

Given our group's personal experience with living in Austin, we have additional (subjective) insight.

- It makes sense that the eastside of downtown Austin is found to be a hotspot. It serves as a perfect settling place for people that either work downtown or work at a tech company, such as Oracle, in that region.
- The north area of Austin is attractive for the same reasons.
- The southwest side of Austin is already a well developed area, so it is not surprising to find that zip codes there seem to have less real estate potential.

We believe that our visualization tool informs users—buyers or investors alike—of potential value they can attain from a real estate purchase in Austin. In particular, the dashboard shows ZHVI change as a percentage (in the heat map) and as a value (in the line chart) for the past and the future. These visuals should help buyers quickly identify areas (by zip code) where they are more likely to find budget-friendly, for-sale properties, and it help investors estimate how much short and long term value they can expect to achieve from a property purchase in a given area.

7 References

[1] Zhu, Hengshu, et al. “Days on Market: Measuring Liquidity in Real Estate Markets.” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16, ACM Press, 2016, pp. 393–402. DOI.org (Crossref), doi:10.1145/2939672.2939686.

[2] Lowrence, Roy E. “Predicting the Market Value of Single-Family Residential Real Estate.” PhD diss., New York University, New York, 2015, https://cs.nyu.edu/media/publications/lowrance_roy.pdf.

[3] Nguyen, An et al. “Housing Price Prediction.” 2018. https://pdfs.semanticscholar.org/782d/3fdf15f5ff99d5fb6acafb61ed8e1c60fab8.pdf.

[4] Bin, Junchi, et al. “Attention-Based Multi-Modal Fusion for Improved Real Estate Appraisal: A Case Study in Los Angeles.” Multimedia Tools and Applications, July 2019. DOI.org (Crossref), doi:10.1007/s11042-019-07895-5.

[5] Heyman, Axel Viktor, and Dag Einar Sommervoll. “House Prices and Relative Location.” Cities, vol. 95, Dec. 2019, p. 102373. ScienceDirect, doi:10.1016/j.cities.2019.06.004.

[6] Park, Byeonghwa, and Jae Kwon 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, Apr. 2015, pp. 2928–34. DOI.org (Crossref), doi:10.1016/j.eswa.2014.11.040.

[7] Rafiei, Mohammad Hossein, and Hojjat Adeli. “A Novel Machine Learning Model for Estimation of Sale Prices of Real Estate Units.” Journal of Construction Engineering and Management, vol. 142, no. 2, Feb. 2016, p. 04015066. DOI.org (Crossref), doi:10.1061/(ASCE)CO.1943-7862.0001047.

[8] Ahmed, Eman, and Mohamed Moustafa. “House Price Estimation from Visual and Textual Features.” ArXiv:1609.08399 [Cs], Sept. 2016. arXiv.org, http://arxiv.org/abs/1609.08399.

[9] Poursaeed, Omid, et al. “Vision-Based Real Estate Price Estimation.” Machine Vision and Applications, vol. 29, no. 4, May 2018, pp. 667–76. arXiv.org, doi:10.1007/s00138-018-0922-2.

[10] Wu, Lynn, and Erik Brynjolfsson. The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales. Vol. Economic Analysis of the Digital Economy, 2013, https://www.nber.org/chapters/c12994.

[11] Nagel, Henrik R. “Scientific Visualization Versus Information Visualization.” 2019. https://www.hpc2n.umu.se/sites/default/files/events/para06/papers/paper_213.pdf

[12] Agnone, Anthony. “Visualizing House Price Distributions.” *Medium*, Towards Data Science, 19 July 2019, https://towardsdatascience.com/visualizing-house-price-distributions-fb0e6c2c264d.

[13] Baldominos, Alejandro, and Iván Blanco. “Identifying Real Estate Opportunities Using Machine Learning.” Computer Science Department, Universidad Carlos III de Madrid, 28911 Leganés, Spain, Nov. 2018, https://arxiv.org/pdf/1809.04933.pdf.

[14] Jermain, Nate. “Home Value Prediction.” *Medium*, Towards Data Science, 5 Apr. 2019, towardsdatascience.com/home-value-prediction-2de1c293853c.

[15] Haseeb, Durrani, et al. “A Data Scientist’s Guide to Predicting Housing Prices in Russia.” *NYC Data Science Academy Blog*. https://nycdatascience.com/blog/student-works/data-scientists-guide-predicting-housing-prices-russia/.

[16] “Housing Data 101: Why Use the Zillow Home Value Index (ZHVI) Instead of a Median Sale Price Series.” Zillow Research, 21 Dec. 2017, https://www.zillow.com/research/why-zillow-home-value-index-better-17742/.

[17] Taylor, Sean J., and Benjamin Letham. Forecasting at Scale. e3190v2, PeerJ Preprints, 27 Sept. 2017. peerj.com, doi:10.7287/peerj.preprints.3190v2.

8 Appendix

8.1 Survey Results

The tables below summarize the feedback from 16 respondents.

Question	Choice Range	Average Response (1 to 5)
Overall, how helpful do you think the dashboard is for a person (not necessarily yourself) interested in buying real estate in Austin, Texas?	Not at all helpful - Extremely helpful	3.94
Compared to using a popular real estate marketplaces like Zillow, how would you rate the usefulness of this dashboard for exploring real estate valuation in Austin?	Nowhere near as helpful - Much more helpful	3.19
How easy is it to understand the instructions for adjusting the selection filters (i.e. drop-downs) in the dashboard?	Nowhere near as helpful - Much more helpful	3.78
How easy is it to understand the color encoding (i.e. the gradient scheme in the map and the colors in the line chart)?	Very difficult - Very easy	4.07

Table 4: Summary of Closed-Ended Survey Questions

Question	Selected Responses
What do you like most about the dashboard?	“Amount of detail succinctly shown”, “nice idea and easy concept”
	“it predicts future prices!”, “cool idea to see house price go up!”, “Really interesting data and prediction model”, “I liked the prediction that you guys have. It is very helpful and it is very easy to understand”
	“I like the use of the map with a block of color for the specified zip code”, “colorfulness and the combination

	of chart & map”
	“The interactivity of the paired choropleth map and line chart”, “I like the clear color encoding and the mapping between area and corresponding line chart”, “The interactive map visualization”
	“Use of tableau gives it a very professional look”
What do you like the least about the dashboard?	“table view or other visualization than a line chart would have been better”
	“box design”
	“information amount felt limited”, “information is limited but focused”
	“The gradient on the map should be a single hue, or only use a second hue for predicted drops in value. White being higher than blue but lower than yellow while being harder to see than both of them in counterintuitive. Also in the tooltips, I think it would help if the left side text were right aligned. I didn't realize zip corresponded to a value on the right at first glance”
	“A little bit laggy to use”, “Drop downs not responsive on mobile”, “It lags a bit when moving between elements”
	“I think it's better to use dashed line for predicted trend on the line chart”
	“How was the forecasting validated for the future ZHI predictions? There seems to be a near linear trend for the 78705 zip code which is difficult to believe given that the zip code 78704 has more seasonal fluctuations that one would expect in the housing market.
	“I don't understand why there are negative percent changes when all of the lines are trending upward (e.g. 3 bedrooms, June 2020). Is the percent-change adjusted for inflation? That's the best explanation that I can think of, but I would like to see it explained somewhere”

Table 5: Summary of Open-Ended Survey Questions

Question	Yes / No	Selected Comments
Did you find any issues with the selection filters? If yes, please specify.	0 / 16	
Is there a feature that you think should be removed? If yes, please specify.	2 / 14	
Is there a feature that you think should be added? If yes, please specify.	5 / 11	“Search by address”
		“I would like to know the regression error on the historical data to decide the confidence of your prediction model, like you can plot the prediction of historical data using your model against the real data
		“trend comparison. it would be great if you could compare two trends at once, based on addresses entered (e.g. a potential homebuyer comparing two houses in different neighborhoods)”

Table 6: Summary of Hybrid Survey Questions