An Interactive Rent vs. Buy Calculator

Team Tufte Love

I. INTRODUCTION

Each year, one of the many decisions that 40 million Americans who move must make is whether to buy or rent [1]. To some, the question hardly seems worth asking because few ideals are as ingrained in the American mentality as the dream of homeownership [2]. Homeownership has been found to enhance civic pride [3], contribute to less crime, and to create better familial environments and higher educational outcomes [4]. From a financial perspective, buying a house is often seen as a wise investment that builds wealth over time through home appreciation and generous tax breaks [5].

Recent research, however, is mixed. Statistical studies suggest that homeownership has a causal effect (at least Granger causal) on unemployment [6]. Financially, renting can be a better investment than homeownership [7]–[10]. Although increases in holding periods, inflation, and the spread between imputed rent (the amount you would pay to rent a house like the one you own) and the opportunity cost of household savings can favor ownership, renters save on mortgage payments, maintenance, and improvement costs. Further, whether buying or renting is more favorable depends, in part, on interest rates and the actual holding period [8].

The many factors necessary to determine whether buying or renting is the better financial choice can be overwhelming. This has motivated our team to design an interactive application that can help movers decide which zip codes to target and whether buying or renting is the better choice in those specific zip codes.

II. PROBLEM DEFINITION

Our aim was to build an interactive visual analytics tool to assist movers by predicting and contrasting wealth accumulation over time in both rent and buy scenarios for various zip codes in the state of California. We aimed to use data and machine learning techniques to supply users with starting points for choosing the required input parameters. The information generated by the tool should help users determine which zip codes better fit their budgets and whether buying or renting is a better financial decision in those zip codes. Due to considerations discussed below, we limited our original objectives by focusing on buying a house versus renting an apartment when both are located in the same California zip code.

III. SURVEY

Various frameworks have been proposed to determine the optimal housing decision. Beracha and Johnson [11] presented the first “horse race” between renting or buying and compared both scenarios using internal rate of return (IRR). Taubner proposed the use of the net present value (NPV) and found that households often need holding periods of between 5 and 10 years to achieve a break-even NPV [7]. The rent or buy decision problem is noted as equivalent to a disk spindown problem [12]. Krishnan, Long, and Vitter [12] have optimally solved that problem, but only using probability distributions under the assumption that resource use times are a priori unknown. Since our product will make recommendations based on anticipated durations at a property, this solution is not applicable.

The number of frequent movers has inspired the development of web applications to inform housing decisions. *The New York Times* published an interactive web application that keeps a running tally of the most common expenses, taking into account initial costs, recurring costs, and opportunity costs; but the application requires users to enter home price, holding period, mortgage details, tax rates, closing costs, and estimated interest rates [13]. That application does not make use of any market data to help users estimate prices or interest rates. Real estate applications such as Zillow.com [14] and Realtor.com [15] provide similar services and also rely on users to estimate interest and equity appreciation rates. Realtor.com provides initial estimates for costs and rates, but these values reflect a nationwide average, despite the fact that they significantly vary by state, city, and neighborhood [11].

IV. Proposed Method

Our approach addresses the shortcomings identified in the previous section. We understand that the benefit derived from a financial model is highly dependent on its inputs. Many movers may not be well-enough informed to make educated estimates of sales and rental prices, real estate appreciation rates, and mortgage interest rates. If inadequate estimates are given to financial calculators, predictions are likely to be incorrect, and the user may be given a wrong recommendation. One innovation to our approach is the use of data and machine learning techniques to supply inputs otherwise required from the user. This makes our application more intuitive and less prone to erroneous user-provided inputs regarding market dynamics than the currently available financial calculators.

Another key innovation of our product is that we walk the user through every step of the rent-versus-buy decision. Alternatives like Realtor.com force the user to get price estimates and run financial analyses in separate views, with no means of comparing results from buying to that of renting on a single webpage.

Our approach was prototyped for the state of California, the most populous state in the U.S. and the state with the third highest number of out-of-state movers [16]. We utilize publicly available data from Kaggle [17] of historical housing prices in the U.S. from 1996 to 2017 to estimate median sales and rental prices in the zip codes that surround the area of interest. The same data set is used to estimate the rate of appreciation in that specific zip code. There is typically less inventory for house rentals than multifamily/apartment rentals. Since there is typically less inventory for house rentals than multifamily/apartment rentals, our project focuses on two options: (1) buy a single family residence or (2) rent an apartment or multifamily dwelling in the same area.

Current home listing and rental prices (as of March 2020) for each zip code were forecast from historical data using the Prophet library [18]. Prophet is a nonlinear time series library that builds generalized additive models (GAMs) with time as a regression variable to account for trends (non-periodic changes), seasonality (periodic changes), and the effect of non-periodic holidays. Prophet was selected instead of traditional time series analysis techniques (e.g., ARIMA) due to Prophet's superior handling of outliers and changepoints, which are representative of external shocks in the real estate market (e.g., the Great Recession of 2008). The training set was defined as the median home listing and rental price for each California zip code in the historical data set.

The same data set was used to estimate the appreciation rate of real estate in each zip code, which was extracted from the slope of the low-frequency trend forecasted by our time series model to mitigate the effect of seasonal variations. Figure 1 shows an example of low-frequency trend and seasonal variation in Culver City, CA 90232.

Rent and buy scenarios are evaluated based on wealth accumulation. Our model is based in part on the financial calculations proposed by Khan [19]. For the current iteration of our project, we have built in the following assumptions, which could change or be made variable in future iterations:

- The spread (i.e., the net monthly outflow difference between the buy and rent scenarios) is assumed to be invested with an annual return rate of 4.0% (modeling what might be obtained, for example, in 401(k) account investments).
- All payments in both scenarios are modeled as processed monthly.
- Interest is compounded monthly.
- Closing and selling costs (in the buy scenario) are assumed to be 6.0% of the sales price.
- Selling costs are calculated at the end of each compared duration (modeling total costs as if one were to move after a given duration).
- Inflation is assumed to be constant at 2.0%.
- Marginal income tax rate is assumed to be 30.0%.
- Mortgage interest rate is a function of the buyer's credit score [20].

Our user interface (UI) was designed to be intuitive and to prevent the user from entering inadequate estimates of real estate information, relying instead on the Prophet-forecast estimates based on Zillow time series data. The UI, shown in Fig. 2, has four sections:

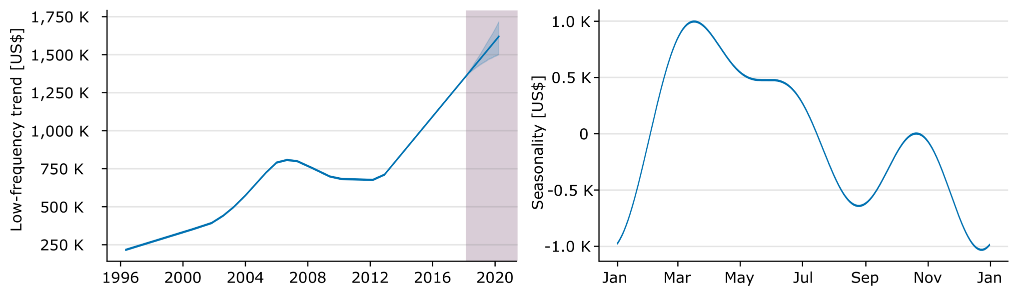

Fig. 1: Sales prices show seasonality with slightly higher prices in the spring and summer. Prices drop at the beginning of the school year and are significantly lower in the winter.

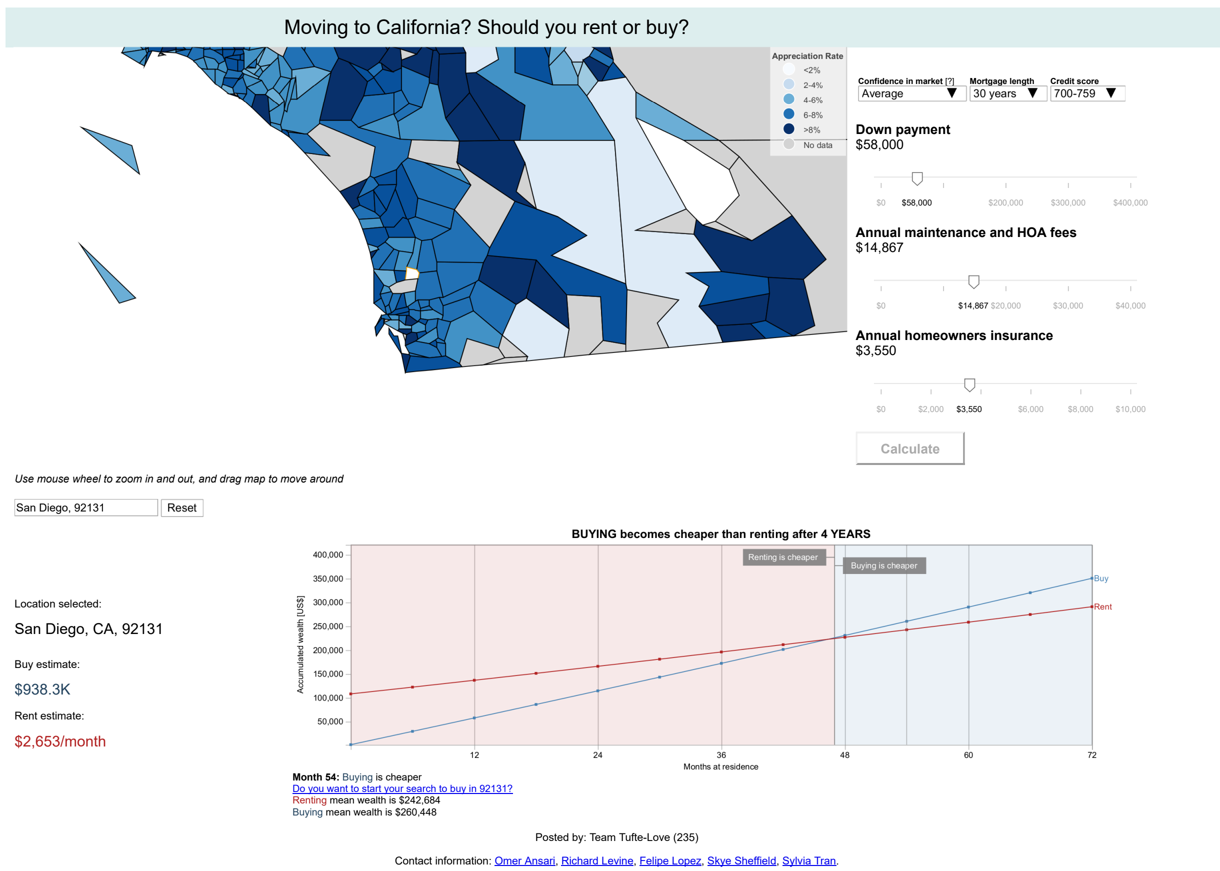

Fig. 2: The user interface has four sections: an interactive choropleth of California where the color of each zip code corresponds to its rate of housing appreciation (top left), a set of sliders for the user to enter financial information (top right), information on the zip code that has been selected for analysis (bottom left), and a figure showing the estimated accumulation of wealth in rent and buy scenarios (bottom right).

Choropleth: The choropleth of zip codes in the state of California enhanced with zoom and drag capabilities was written in d3/js. It is set up using a mercator projection for the state of California and restricted to California-only zip codes for faster rendering. By default it shades zip codes by their appreciation rates, as predicted by our back end algorithms. Certain zip codes did not hold sufficient data either to impute or forecast appreciation rates; those zip codes regions are gray. A tooltip appears when the user hovers over any particular zip code, displaying the corresponding city name, purchase and rental prices, and appreciation rate. A zip code can be selected by clicking on it, or by entering either the zip code or the city's name. Since some cities can have multiple zip codes, a drop down appears under the field allowing the user to select the desired zip code. When the user clicks on (or enters) a particular zip code, the choropleth approximately zooms into that zip code's latitude and longitude and colors the zip code orange. It also prints the buy and rent prices and renders the other calculation options on the screen. A reset button unselects any previously selected zip code and zooms back to the original projection.

Purchase and rental price estimates: The selected zip code is shown to the user in the bottom left section of the UI alongside estimated purchase and rental prices for the user's reference. If the estimates do not match the user's budget, a new zip code can be selected.

Financial inputs: Once a zip code has been selected, the section for financial information becomes visible. The user may enter a mortgage term and a credit score range, as well as indicate a level of confidence that the real estate market will appreciate as predicted.[1]. This section also includes sliders to select the intended down payment, estimates for annual maintenance and HOA fees, and annual homeowners insurance. At any point, the user can click on the *Calculate* button to run a Fetch request to the Flask server to run financial calculations for wealth accumulation.

Financial plot: The request message is then parsed by the browser and shown as an interactive figure that presents the estimated accumulated wealth (as NPV) for the rent and buy scenarios. The background in the figure is shaded to illustrate the period in which it is better to rent or buy. The figure also has tooltips that provide the user with accurate estimates with six-month intervals. The user can click on any point to indicate the intended length of stay, at which point a detailed description of both scenarios is shown below the figure alongside a link to start the search of a house in the targeted zip code using *Zillow.com*.

V. EVALUATION

A. Usability

The ease of use of our web application was measured by the number of inputs that the user must enter manually. This approach resembles the DPU (defect opportunity per unit) metric used in Six Sigma [21]. The rationale of this choice is that every parameter that must be entered manually by the user is a defect opportunity. Table I compares our web application with Realtor.com and *The New York Times*. Our application requires fewer input variables and reduces the probability of user-provided inputs that may yield misleading information that might affect a user's decisions. Furthermore, our application does not require users to have a firm grasp of real estate or rental market dynamics. Note that while Realtor.com and *The New York Times* provide default values for mortgage interest rate and home appreciation rate, those defaults are static and do not capture variability due to the buyers' credit scores or market differences between zip codes.

App	No. inputs
Realtor.com	21
New York Time	21
Our application	6

TABLE I: Benchmarking our application with other financial calculators based on the number of financial inputs required from the user.

B. Model evaluation

The machine learning algorithms used to predict sales prices were evaluated on a test set that consists of five zip codes (90292, 96146, 96161, 90044, and 90003). Time series data sets were split into training and validation sets at the November 2015 mark. The sequence prior to the split date was taken as the training set to predict the median home listing and multifamily rental, respectively. The mean absolute percentage error (MAPE) was adopted as the metric for predictive power of the model regardless of the variability in the magnitude of the predicted variables (see Table II).

Zip code	Sale MAPE (%)	Rent MAPE (%)
90292	13.47	5.19
96146	14.73	NA
96161	5.93	NA
90044	18.68	NA
90003	3.09	NA

TABLE II: Mean absolute percentage error for time series forecasts of sale and rent price in five California zip codes. NA indicates that there was no data on rental prices to support the analysis.

[1]Three levels of confidence on the real estate market allow the user to run "what if" simulations. In the *optimistic* scenario, the appreciation rate is assumed to continue as it has been observed since the recovery from the Great Recession. The *average* scenario multiplies our estimated appreciation rate by 0.75 to match average estimates from the past 10 years. The *conservative* scenario multiplies our estimated appreciation rate by 0.5 for an even more conservative estimate.

	Mean	Std. dev.	Minimum	25%	50%	75%	Maximum
Purchase price [1000 US$]	764.7	729.0	-16.2	315.6	582.8	905.6	8,395.6
Rental price [US$/month]	2,601.1	949.0	590.0	2,148.0	2,592.8	2,852.8	7,645.0
Appreciation rate [%]	6.46	3.59	-14.37	5.27	6.50	7.90	103.01

TABLE III: Statistics of real estate data before removal of outliers.

Based on the initial predictions generated by Prophet, there were zip codes that exhibited abnormally high appreciation rates, and in some cases, negative buy values. Upon closer inspection, it was discovered that these zip codes had outliers that caused Prophet to generate such predictions.

The team collectively decided that, for these zip codes, an additional data preprocessing step was required: the removal of outliers. The revised preprocessed data for these zip codes were then used to generate revised predictions. Criteria used to select zip codes that exhibited abnormal predictions included the following:

- *Appreciation Rates*: Outliers were determined by finding abnormally high or low appreciation rates that were further than two standard deviations from the mode (see Fig. 3). The mode was selected based on the right-skewed distribution of the initial appreciation rates generated by Prophet for all California zip codes.
- *Buy and Rent Values*: Outliers were determined as points below the lower bound (US$0) and above the upper bound (10 times the mean predicted purchase price).

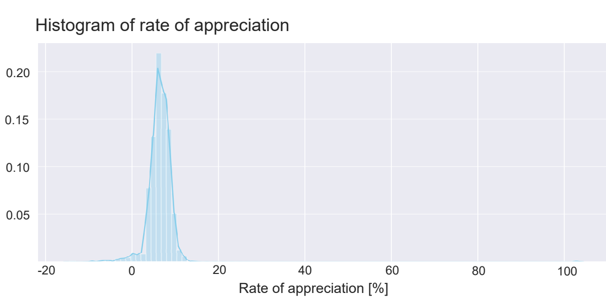

Fig. 3: Normalized histogram of estimated rates of appreciation for zip codes in California.

As observed in Tables III and IV, our results significantly improved and now reflect predicted figures that are consistent with market dynamics.

However, outliers were not the only data challenge encountered. As shown in Figure 4, some rent versus buy ratios were inconsistent with ratios seen in most other zip codes. Some zip codes had identical buy values even though their rent prices were dramatically different. Conversely, there were zip codes with matching rental values but wildly different buy values.

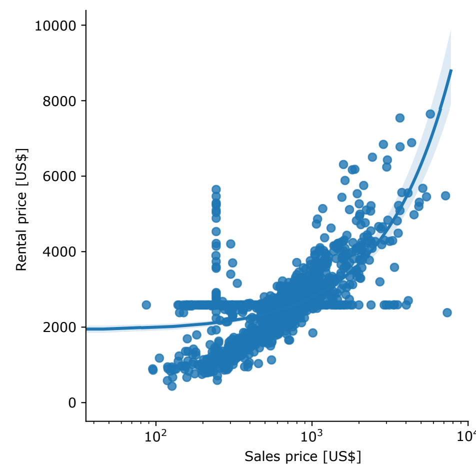

Fig. 4: Comparison of forecast rental prices and forecast purchase prices for zip codes in California.

Financial predictions from our application were also compared to those obtained with Realtor.com to ensure that predicted payoff time, i.e., the holding horizon after which it is preferable to buy instead of rent, is relatively close between the two solutions. Table V compares the payoff times estimated with both applications. Note that the two applications follow different but related approaches. Realtor.com is based on net annual cost, while we compare the accumulation of wealth. Regardless of the approach, results obtained are comparable, but the difference increases with larger rates of appreciation.

VI. CONCLUSIONS AND DISCUSSION

The application developed by our team was successful in leveraging real estate data to greatly simplify the experience of selecting target zip codes in California and determining whether to rent or buy in them. Compared to available solutions, our application comes preloaded with historical real estate data for the state in California and does not ask the user to enter it manually. Our

	Mean	Std. dev.	Minimum	25%	50%	75%	Maximum
Purchase price [1000 US$]	761.8	721.4	0.0	312.2	581.6	903.9	7,306.6
Rental price [US$/month]	2,583.7	965.1	0.0	2,148.5	2,592.0	2,841.8	7,645.0
Appreciation rate [%]	6.37	2.20	-3.31	5.23	6.48	7.90	12.3

TABLE IV: Statistics of real estate data after removal of outliers.

Zip code	Financial variable	Optimistic	Average	Conservative
Santa Monica 90405	Appreciation rate	8.0%	6.0%	4.0%
	Estimated payoff time [years]	1.5	2.0	4.5
	Realtor.com estimate [years]	2.0	4.0	8.0
San Diego 92131	Appreciation rate	4.0%	3.0%	2.0%
	Estimated payoff time [years]	3.0	3.0	5.0
	Realtor.com estimate [years]	5.0	8.0	N/A
Los Angeles 90002	Appreciation rate	1.0%	0.8%	0.5%
	Estimated payoff time [years]	1.5	1.5	1.5
	Realtor.com estimate [years]	2.0	2.0	2.0

TABLE V: Comparison of payoff time of the investment for our application and Realtor.com

application greatly reduces the opportunities for a layperson to produce errors in the calculation by requiring only 5 financial inputs as compared to the 21 requested in alternative solutions. Despite its simplicity, our application predicts a pay off time for the purchasing decision that is consistently in the vicinity of the Realtor.com predictions [15].

Despite the data having zip code level granularity, there were several zip codes that were surprisingly sparse. In fact, data sparsity influenced which features we selected for model prediction purposes (median home values and multifamily rental prices were the most data rich).

The following techniques were explored but not pursued as ways to impute the sparser zip codes that exhibit similar traits to other less sparse zip codes: shapelets [22] and time series clustering [23]. More time would need to be spent more closely analyzing the data within and across zip codes (each a time series) to better understand which of these techniques, if any, should be implemented. Based on preliminary research, the use of either shapelets or clustering for calculating time-series similarity to impute more sparse zip codes by finding other zip codes that exhibit similar traits may pose new problems. In lieu of this more technical effort to apply imputation methods on sparse zip codes based on finding comparable and less sparse zip codes, we implemented interpolation techniques to fill null values to mitigate Prophet's generating estimates that were wildly inaccurate. Furthermore, we excluded zip codes where the only available data was between 2008 and 2010 (during the height of the Great Recession). Provided more time, we would explore various time-series feature extraction and engineering techniques to enrich the sparse dataset for purposes of generating revised time series estimates from our model.

A possibly more important source of uncertainty for our financial predictions is the current economic climate, which has recently received the shock of the Covid-19 pandemic and has been predicted to be heading into a long recession that is likely to affect the real estate market [24].

VII. DISTRIBUTION OF TEAM MEMBER EFFORT

O.A., R.L., F.L., S.S., and S.T. conceived and planned the project. S.T. built the data ETL pipeline, and together with S.S., analyzed the data and built the machine learning algorithms to estimate rent and sale prices, and the rate of appreciation of real estate. R.L. and F.L. programmed the financial calculations for wealth accumulation in the buy and rent scenarios. O.A., R.L., and F.L. built the front end. O.A. added the choropleth with search capabilities and auto-zoom for all zip codes in California. F.L. designed the wealth accumulation plot and connected the financial calculations to the front end. The architecture and CI/CD pipeline were managed by O.A. R.L. lead the quality assurance efforts for all front end development. All team members have contributed a similar amount of effort.

References

[1] M. Sauter, "Population migration patterns: US cities we are flocking to," *USA Today*, 2018.

[2] C. C. Tu and M. J. Eppli, "Valuing new urbanism: The case of kentlands," *Real Estate Economics*, vol. 27, no. 3, pp. 425–451, 1999.

[3] W. M. Rohe, S. Van Zandt, and G. McCarthy, "Home ownership and access to opportunity," *Housing Studies*, vol. 17, no. 1, pp. 51–61, 2002.

[4] D. R. Haurin, T. L. Parcel, and R. J. Haurin, "Does homeownership affect child outcomes?," *Real Estate Economics*, vol. 30, no. 4, pp. 635–666, 2002.

[5] Z. X. Di, E. Belsky, and X. Liu, "Do homeowners achieve more household wealth in the long run?," *Journal of Housing Economics*, vol. 16, no. 3-4, pp. 274–290, 2007.

[6] D. G. Blanchflower and A. J. Oswald, "Does high home-ownership impair the labor market?," tech. rep., National Bureau of Economic Research, 2013.

[7] I. T. Tabner, "Buying versus renting–determinants of the net present value of home ownership for individual households," *International Review of Financial Analysis*, vol. 48, pp. 233–246, 2016.

[8] "Better off renting? owner-occupation is not always a better deal than renting," *The Economist*, Jan 2020.

[9] E. Beracha and K. H. Johnson, "Lessons from over 30 years of buy versus rent decisions: is the American dream always wise?," *Real Estate Economics*, vol. 40, no. 2, pp. 217–247, 2012.

[10] C. E. Herbert, D. T. McCue, and R. Sanchez-Moyano, "Is homeownership still an effective means of building wealth for low-income and minority households? (Was it ever?)," *Homeownership Built to Last*, 2013.

[11] E. Beracha, M. Seiler, and K. Johnson, "The rent versus buy decision: Investigating the needed property appreciation rates to be indifferent between renting and buying property," *Journal of Real Estate Practice and Education*, vol. 15, no. 2, pp. 71–87, 2012.

[12] P. Krishnan, M. P. Long, and S. J. Vitter, "Adaptive disk spindown via optimal rent-to-buy in probabilistic environments," tech. rep., USA, 1995.

[13] "Is it better to rent or buy?." `https://www.nytimes.com/interactive/2014/upshot/buy-rent-calculator.html`. Accessed: 2020-02-27.

[14] "Zillow." `https://www.zillow.com`. Accessed: 2020-02-27.

[15] "Rent vs. buy calculator." `https://www.realtor.com/mortgage/tools/rent-or-buy-calculator/`. Accessed: 2020-02-27.

[16] "Where americans are moving." `https://www.usnews.com/news/best-states/slideshows/americans-are-moving-to-these-state`. Accessed: 2020-03-26.

[17] "Zillow economics data." `https://www.kaggle.com/zillow/zecon`. Accessed: 2020-02-27.

[18] S. J. Taylor and B. Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018.

[19] "Renting versus buying a home." `https://www.khanacademy.org/economics-finance-domain/core-finance/housing/renting-v-buying/v/renting-versus-buying-a-home`. Accessed: 2020-03-26.

[20] "Mortgage rate comparison." `https://www.myfico.com/loan-center/home-mortgage-rate-comparison/default.aspx`. Accessed: 2020-04-04.

[21] M. J. Harry, "Six sigma: a breakthrough strategy for profitability," *Quality progress*, vol. 31, no. 5, p. 60, 1998.

[22] L. Ye and E. Keogh, "Time series shapelets: a new primitive for data mining," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 947–956, 2009.

[23] J. Serra and J. L. Arcos, "An empirical evaluation of similarity measures for time series classification," *Knowledge-Based Systems*, vol. 67, pp. 305–314, 2014.

[24] "IMF says the world will "very likely" experience worst recession since the 1930s." `https://www.cnbc.com/2020/04/14/imf-global-economy-to-contract-by-3percent-due-to-coronavirus.html`. Accessed: 2020-04-15.