Ranking Best Places for Tech Worker Homeownership

Team 27

1 Introduction

Since the Great Recession of 2008, Software Engineer salaries have exploded relative to other professions. This can be attributed to a couple of factors: the settlement of the High-Tech Employee Antitrust Case [21] and IPO of Facebook and rise of "Unicorn" (private companies worth over $1B) companies. The side effect of this is skyrocketing housing prices in the traditional tech hubs. In response, many tech workers are deciding to move to lower cost locations such as Austin, Texas to buy a home[23]. According to the Economist, "between 2007 and 2016 a net 1m American residents ... left California for another state. Texas was the most popular destination"[22]. Software Engineers contemplating a move from CA to TX would want a way to evaluate the cost benefit of relocation.

2 Problem Definition

Current methods and solutions of ranking cities by "Livability" used by popular websites today (Niche.com, Areavibes.com) are subject to editorial bias and are too general to be useful for a specific demographic like Software Engineers. We propose to use machine learning algorithms to aggregate 5 features (Housing Prices, Schools, Crime, Weather, Salaries) into a normalized index called "Best Homeowner Value" (BHV) to compare and rank the best cities for tech worker homeownership. The scope of this project will be limited to California and Texas and to the tech worker job description "Software Engineer." A final visualization tool will be produced to show geographically which areas provide the best value.

3 Survey

3 current examples of City Livability Rankings can be in the screenshots below from Niche.com, AARP.com, and Areavibes.com.

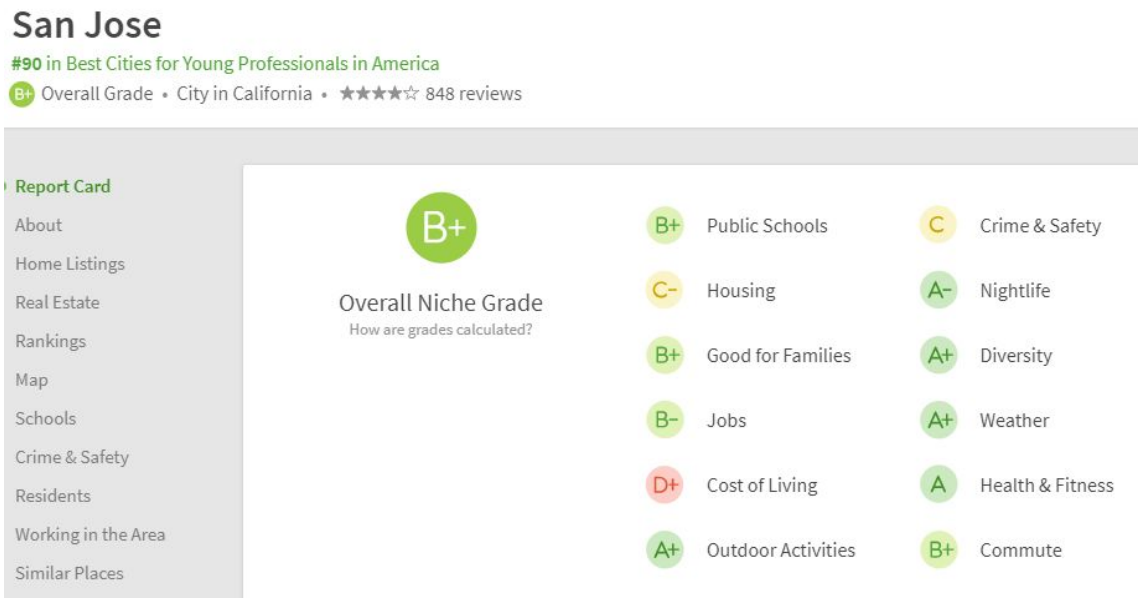

Figure 1: Niche.com

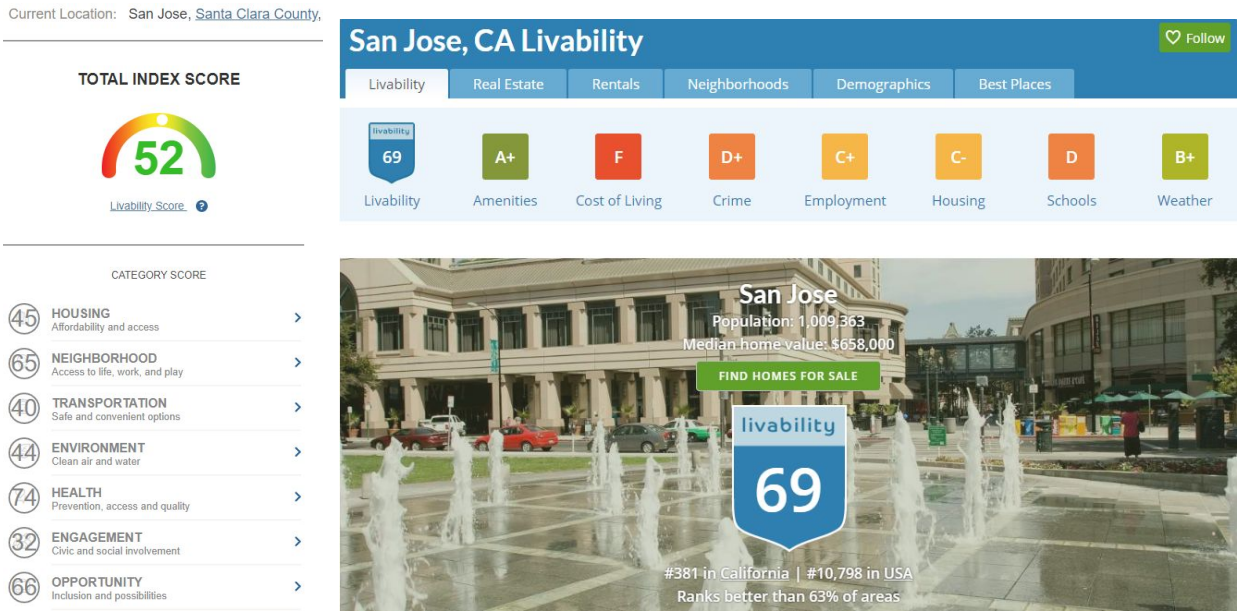

Figure 2: AARP Livability Index (left), AreaVibes.com (right)

AARP's "Livability Index"[1] is based upon an unweighted average of scores across 7 categories. In our study, the living scores will be determined based upon housing price per square, salary, school ratings, crime rates, and we will study the correlation between the factors and housing prices. AreaVibes and Niche are websites that aggregate data across multiple public sources, normalizes it, and applies an arbitrary weighted average of sub-features for a particular feature [2, 3]. The limitations of this practice is that the proportions of the weighted averages are subject to editorial bias.

Evaluating House Price

To evaluate house prices in different areas as well as nationally, repeat sales housing price indexes such as S&P/Case-Shiller are used. However, the limitations of those indexes are "that repeat sales indices omit homes that sell only once from analysis … [like] new home sales." [4]

Evaluating Livability (Crime, School, Commute)

Another way to look at the desirability of the city is to look at the prices of comparable houses (similar age, sq ft, rooms). Multiple studies show that there is multicollinearity between the price of the home and features such as crime through linear hedonic models [5,6,7,8]. One study concluded that violent crime had the most effect on decreasing housing prices, with 1 additional violent crime per 1000 residents lowers property value by 3.6%. [9]

Furthermore, collinearity between house prices and quality of public schools through linear [11,12,13] and non-linear 3rd order quadratic models [10] have also been demonstrated. One study indicated that secondary school quality is the largest school related factor affecting housing prices [14].

The relationship between income and median house prices has been studied, where the obvious conclusion is that unemployment negatively affects house prices [15,16,17] and historically, house prices are 3-4x median annual income [18]. "House prices and regional labor markets"[17] used an error correction model to explain/discuss interactions between house prices and labor markets (average manufacturing wage, unemployment rate, and the labor force). Similarly in our project, we will examine the interactions between the median house prices of single-family homes in the metropolitan area of Texas and California and the median salary of Software Engineers.

The positive relationship between house prices and short commute/access to high speed rail (HSR) [19], has been established in China. Other studies have constructed a linear model to account for the wage premium associated with a longer commute [20], but these models from 2001 predate tools like Google Maps, which we can now use to accurately assess varying commute times.

The limits of the current practices of rating the desirability of the city is that they do not take into account this multicollinearity and presents housing price as an independent variable from the other features [1,2,3], when it is not. In addition, almost all of the studies in our literature review are limited to pairwise relationship of 2 variables (House Price vs Crime, House Price vs Schools, etc) to simplify analysis.

4 Proposed Methods

Intuition

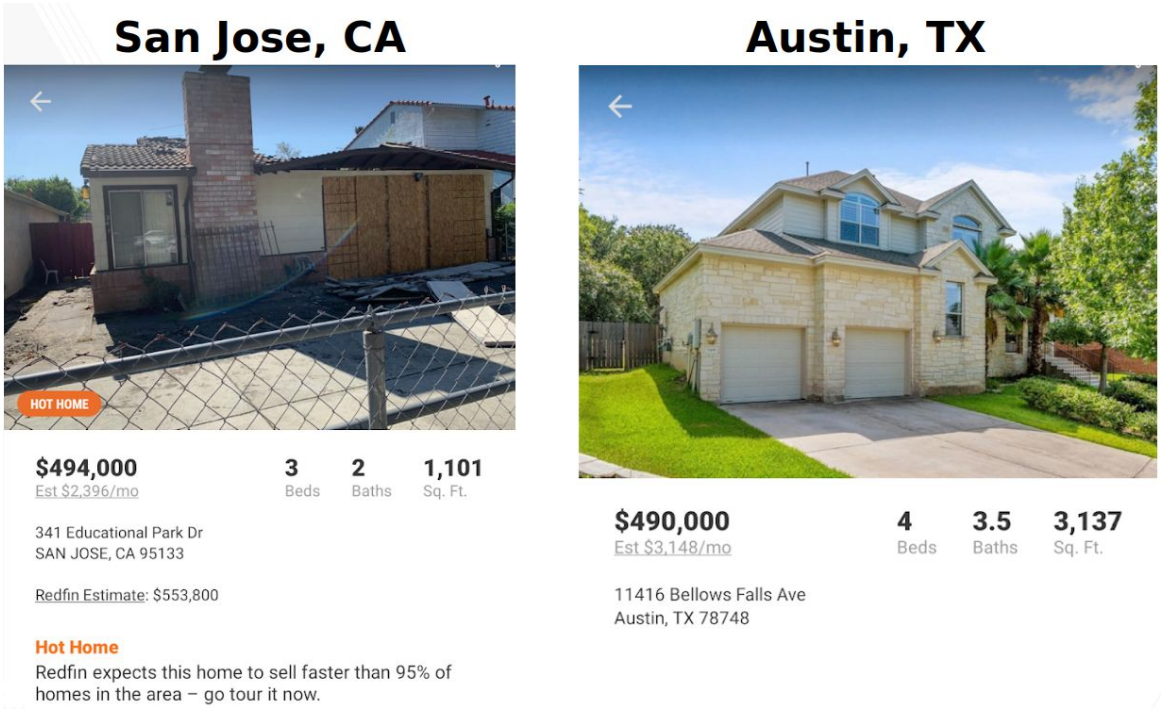

Figure 3: Comparison of Houses at Similar Price Point in San Jose, CA vs Austin, TX

We hypothesize that there are imbalances in "homeowner value" across the nation, that is, the living standard that comes with buying a house is not equal, despite normalizing for house features, prices, property taxes, and income.

Current "Livability" indexes are based on linear models with arbitrary weights assigned to them with no justification of why the weights are assigned. For example, in Areavibes.com, there is a crime score that is calculated with the 65% weight assigned to "Violent Crime" and 35% weight assigned to "Property Crime."

Instead of linear models subject to editorial bias that exist today, our approach will use machine learning ranking algorithms to analyze the raw data of our 5 features to generate scores for cities considered to be tech hubs for software engineers.

Our first key innovation is the following insight: at the heart of the model is the idea that the home price encapsulates what the market already thinks the location and property is worth in terms of quality of schools, crime rate, commute to major employment hubs, salaries in those hubs, and weather. However, it is hard to just compare home prices from one state to another. As seen in Figure 3 above, just on house features (sq ft, #of bd/ba), it is obvious that the house

on the right would be a better value than the house on the left. However, this analysis does not account for the possibility that the house on the left may be in a better school district or in proximity to employers that have a much higher salary than the house on the right. So we need to quantify how much impact each of the 4 features (salary, weather, schools, crime) has on home price. Once each feature is quantified and scored through machine learning algorithms, we can use this information to weigh each score to build a "Best Homeowner Value" metric.

The second key innovation is the following insight: Users have different preferences for city features. For example, maybe one person cares more about an area with better schools over weather, whereas another cares more about an area with higher salaries over crime. So while we will offer our machine learning informed weighted ranking of the cities as a default, we will also offer a way for different users to inject their custom preferences into our visualization tool, perhaps changing the default ranking. This is better than current city livability ranking solutions that don't give users a way to personalize their feature preferences.

Description of Approaches: Dataset Retrieval and Cleaning

Data	Retrieval Method	Size on Disk	Table Representation (row/col)
Home Price and Features	Python script to scrape house price and house feature (Zestimate price, finished sq ft, house age, #bed, #bath) from Zillow.com Deep Search API for the selected 5 cities	55.5MB (clean)	After Cleaned: 219 columns (features) 52715 rows (each house)
Weather	Downloaded datasets from NOAA.gov of historical weather station data in the major airport of each city from 2012-2019	1.5MB (clean)	After Cleaned: 9 columns (features) 14350 rows (each day weather)
Education (School Quality)	Python script to scrape from schooldigger.com API. Use another script to assign each house to a school by shortest distance lat/long	93MB (raw) 200KB (clean)	After Cleaned: 14 columns (features) 1215 rows (each school)
Salary	Downloaded H-1B salary datasets from Kaggle for total US, needed to clean to narrow down to the selected 5 cities	469MB (raw) 240KB (clean)	After Cleaned: 9 columns (features) 5429 rows (each employee salary)
Crime	Downloaded from FBI crime data repo using crime-data-explorer.fr.cloud.gov/api over the selected 5 cities	15KB (clean)	After Cleaned: 6 columns (features) 421 rows (crime type, # instances)

Figure 4: Summary of Dataset Retrieval and Techniques

We obtained 5 different datasets for Home Price, Weather, Education (School Quality), Salary, Crime using retrieval techniques and sources listed in Figure 4. The final cleaned and combined dataset is 55.5MB large, with 219 Features (columns), and 52715 Home Data points (rows)

across 5 cities (TX: Austin, Dallas, San Antonio; CA: Irvine, San Jose). Associating a particular home to a school was challenging in that we had to create a Python script to assign the nearest public schools to a particular home based on calculated latitude/longitude distances. The Python Pandas library was used to combine the data, with different Python scripts used to clean each dataset before the final join of the 5 datasets.

Description of Approaches: Creating Best Homeowner Value Metric

The third key innovation is after collecting and cleaning our dataset, we quantify which features have stronger correlation to the housing price and generate a normalized index of "best homeowner value" to compare between cities in different states, namely California vs Texas.

To develop our "Best Homeowner Value" (BHV) metric, we did the following:

- Since the home price encapsulates the market desirability of the location, measure the impact or feature importance of the 4 other features (salary, weather, schools, crime) on home price through Random Forest (Ensemble machine learning algorithm) regression.
- Based on a pairwise Correlation Matrix for the 219 features, prune redundant features (features with high multicollinearity) from consideration in the final BHV metric.
- Once we are able to quantify and rank the impact each non-redundant feature has on home price, we can create an index on which to rank the BHV. From our experiments, we found that Price > Salary/Wage > Weather > Education/School Quality > Crime.
    - For each feature, normalize those values to a standard scale (50-100, worst to best) as a normalized feature score.
- Based on the ranked impact, weigh the normalized scores and average them to come up with a default BHV score for each city.
    - Sort the cities based on BHV, in descending order. (see the **Conclusion** section below for the final default ranking)

Description of Approaches: Visualization Tool in Tableau

Another key innovation is the visualization tool built in Tableau that users (prospective homeowners working in tech) can use to easily identify and exploit "homeowner value" arbitrage opportunities between various cities in California and Texas. We collected information from over 52,000 houses to build our housing dataset. Then we aggregated those data points with 4 other datasets that we collected (Weather, School Quality, Crime, Salaries) to build a combined dataset that encapsulates the major factors influencing the price for each house. Our visualization tool provides full traceability of each data point on a geographic map, so the user can mouse over each dot to view where each house is located and the details of each feature associated with the house. See the below 5 screens.

Figure 5: Visualization Homepage with Default Ranking (For Better Readability, Zoom View to 200%)

The interface allows the user to input his/her preferences for each feature on a Scale from 0-5 (0 being Not Important at all, 5 being Most Important) to recalculate the rankings. The default setting of House Price > Wage > Weather > Education > Crime is informed by our feature importance discovery with our Random Forest House Price prediction model (see Section 5 Experiments for more detail).

In Figure 6 through 8 below, the user mouseover each datapoint to explore detailed breakdowns of Home Price vs Geographic location, Education & Crime, Wage & Weather, respectively, for each of the 5 cities by changing the selection in the dropdown box "Change City" in the upper right corner of the screen. Figure 9 is the master summary of all the major feature information for all 5 cities. The user can also mouseover the lines and bar charts to see pop-ups that show the detailed breakdown of data.

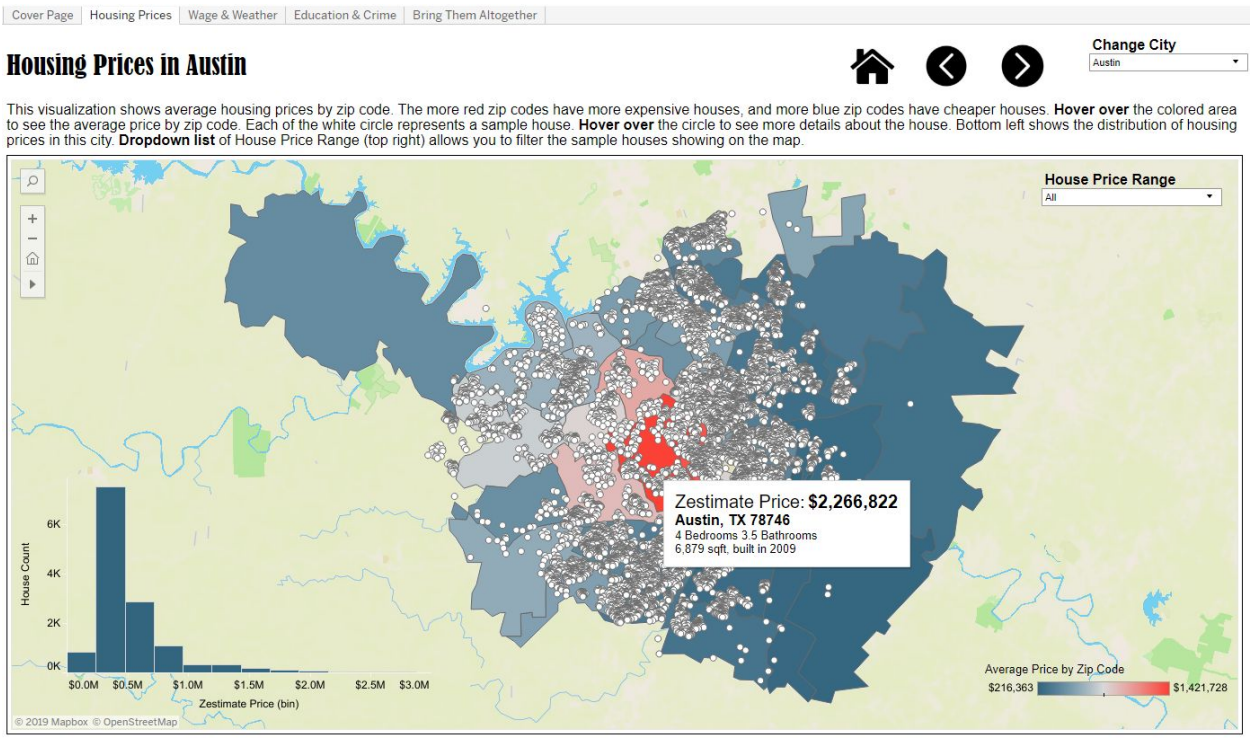

Figure 6: Visualization Austin, Home Price (For Better Readability, Zoom View to 200%)

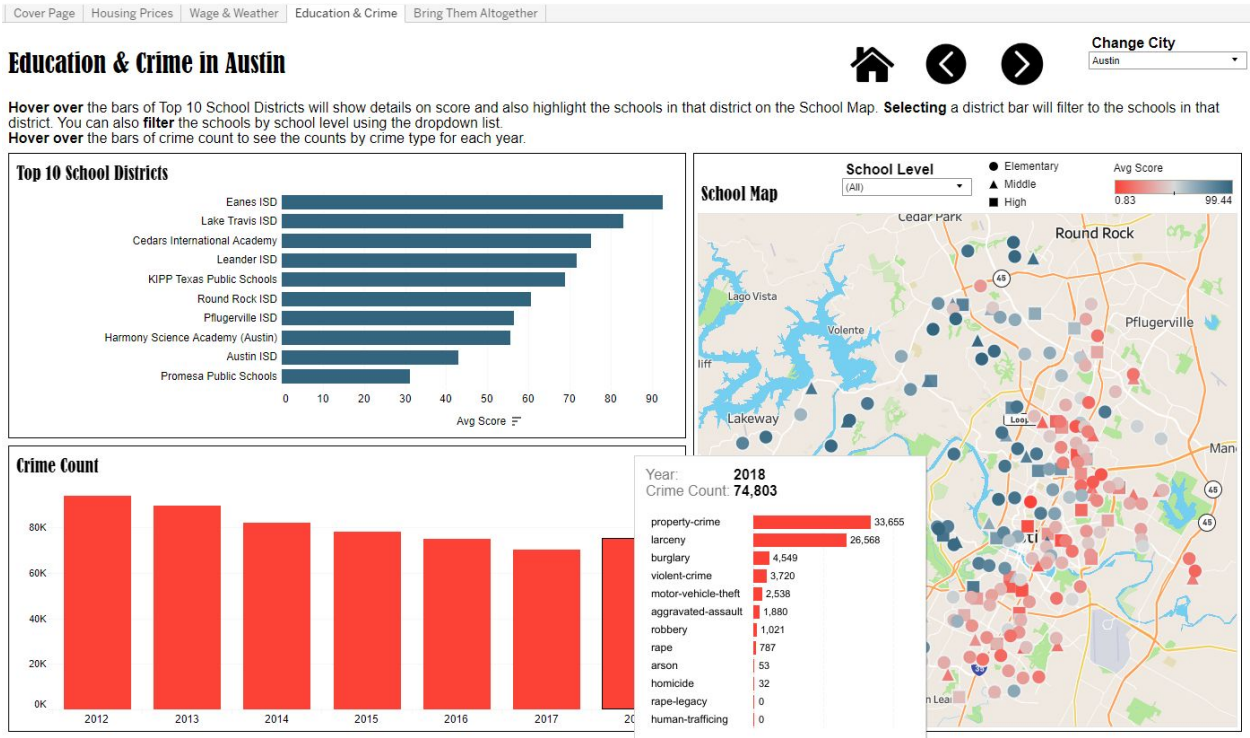

Figure 7: Visualization Austin, TX Education and Crime (For Better Readability, Zoom View to 200%)

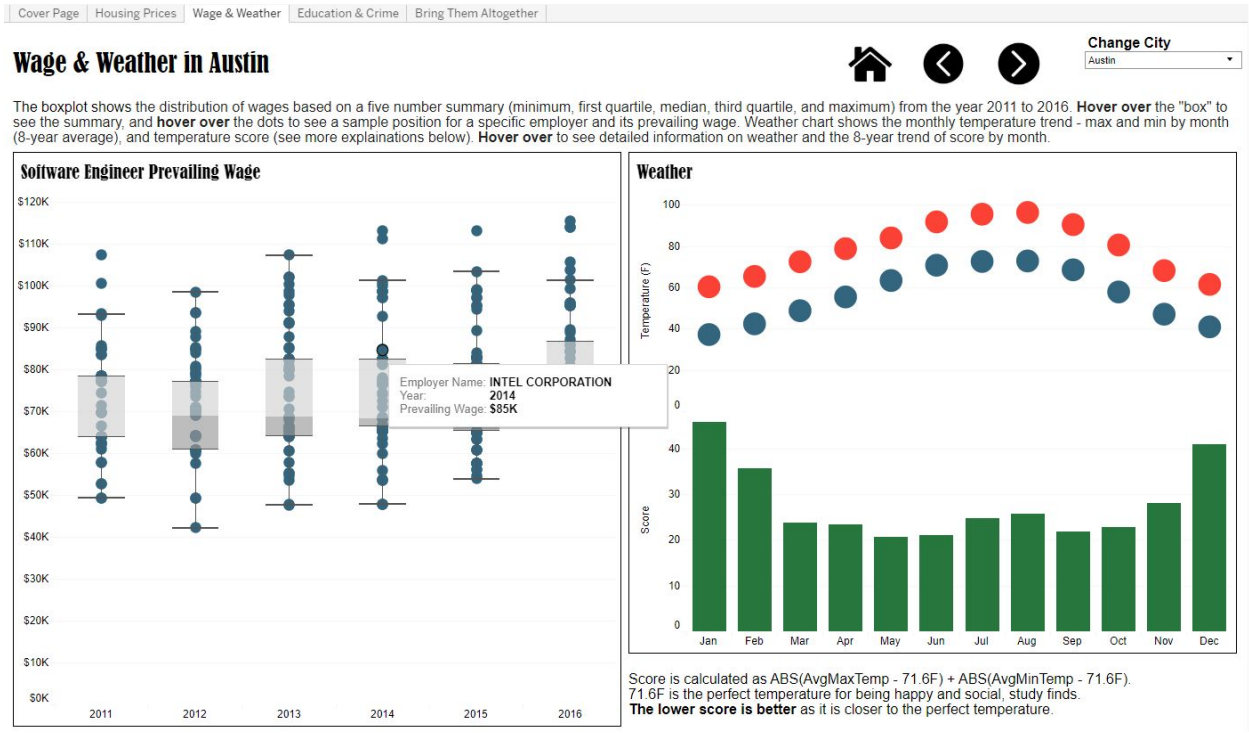

Figure 8: Visualization Austin, TX Wage and Weather (For Better Readability, Zoom View to 200%)

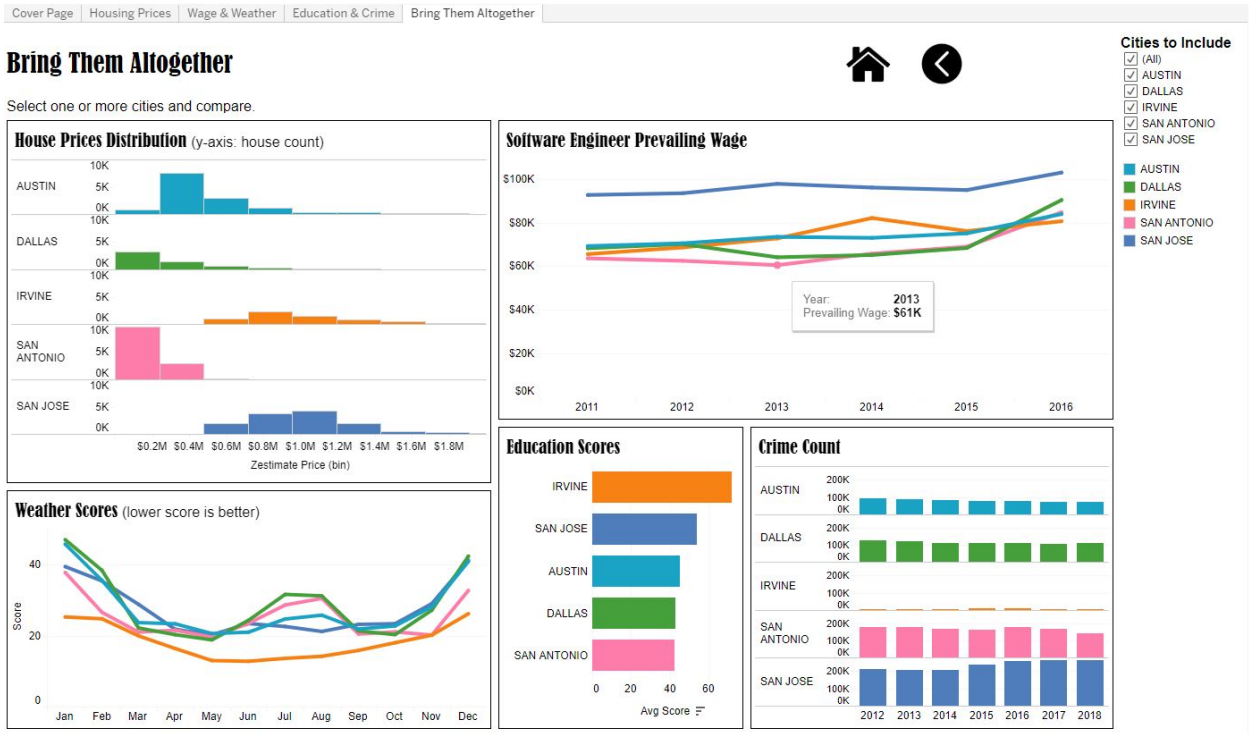

Figure 9: Visualization Summary of all 5 Cities (For Better Readability, Zoom View to 200%)

5 Experiments and Evaluations

List of Questions to Answer

1. Which Features (School Quality, Crime Rate, etc) are more strongly correlated with Housing Price?
2. Are there multicollinearity issues between features?
3. Do we need to engage in further feature engineering to eliminate highly correlated features?
4. What is the best way to visualize "Best Homeowner Value" Rankings across different states?

Description of Testbed

We can measure success in 2 ways:

1. Heuristically comparing our rankings with existing Livability rankings to see if ours was able to identify significant real estate arbitrage opportunities not captured by the existing rankings.
2. Surveying a sample group of people to rank the places by desirability themselves and measuring the deviation of these user rankings to our generated rankings.

Details of Experiments, Observations

Feature Correlation with Housing Prices

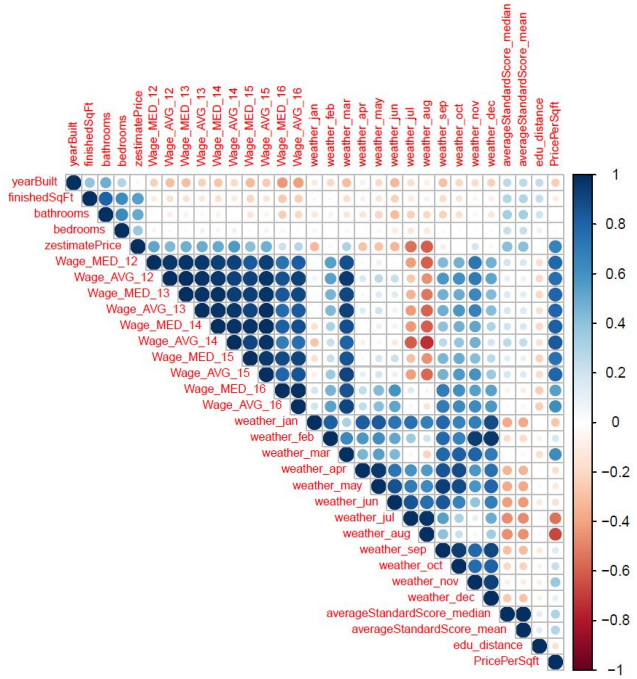

Figure 10: Correlation Matrix of 31 Most Important Features (For Better Readability, Zoom View to 200%)

Observations for Correlation Matrix of 31 Features:

- We find that there are multicollinearity issues with wages for different years from 2012-2016. Wages are highly correlated with zestimatePrice (home price), but we do not need every single year, so we can eliminate earlier years from our Random Forest prediction model.
- The weather of different months also exhibit multicollinearity. However, with respect to home price, the weather in August and July have the most impact, where the smaller value of weather (smaller deviation from the “golden” temp of 71.6F is better), correlates to higher zestimatePrice. This is due to the contribution of the Irvine dataset, which has the 2nd highest median home prices, but the best weather (smallest weather deviation scores). The conclusion is that we could eliminate weather scores outside of August and July.
- For internal home features, overall home finishedSqFt has the most impact on home price, followed by number of bathrooms, then number of bedrooms.
- For external home features, wage has the most impact on home price, followed by weather, then school education score (averageStandardScore).

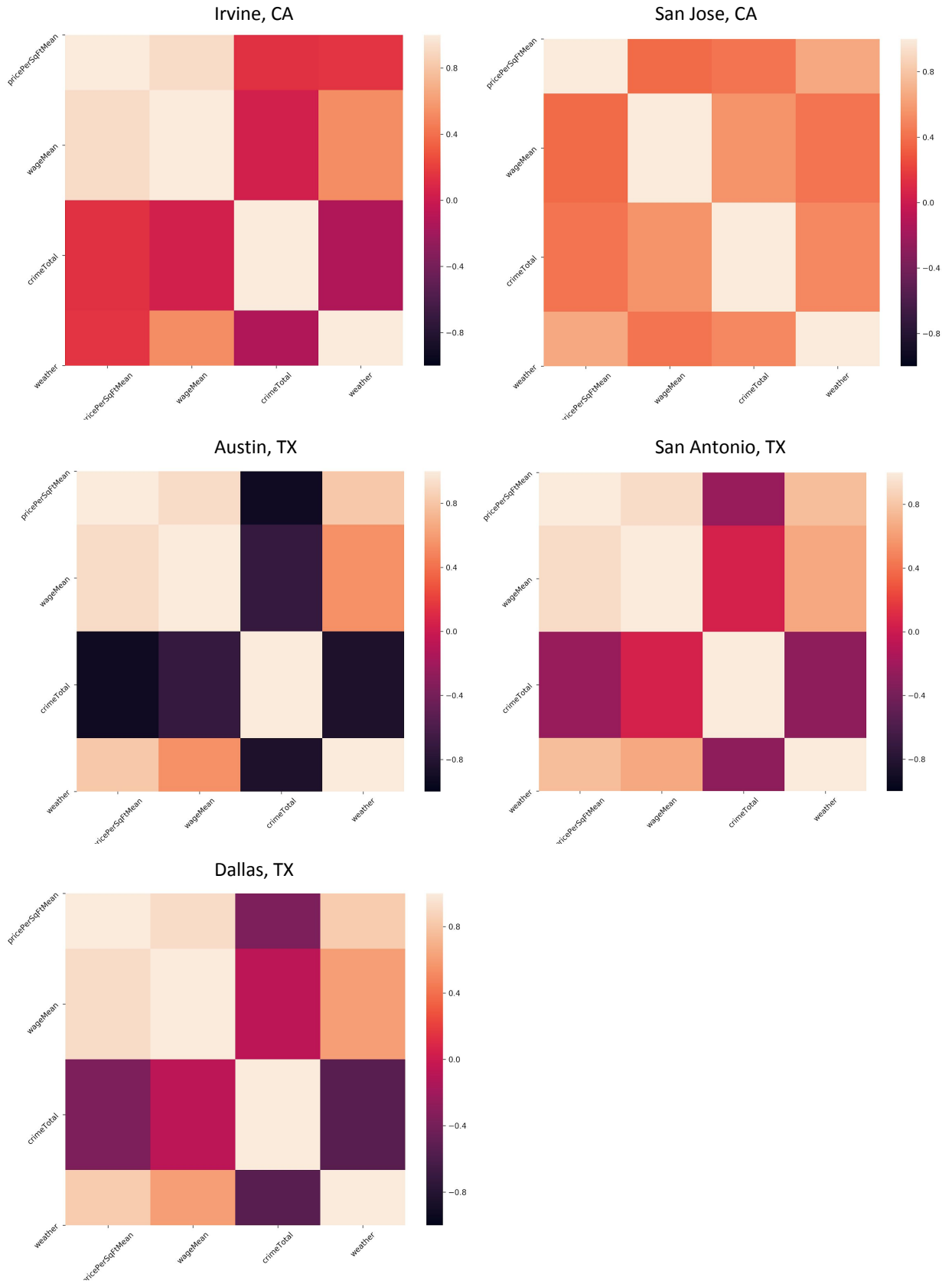

Figure 11: Correlation Matrix of Important Features for the 5 Cities (For Better Readability, Zoom View to 200%)

Observations for Correlation Matrices for the 5 cities:

- In Texas, crime is negatively correlated with home price. Less crime, higher the home price. In California, especially San Jose, there is a slight positive correlation for price, partially due to higher crime rates overall.
- Wage is highly positively correlated with home price for 4 out of 5 cities. The only exception is San Jose, CA, where there is a weak positive correlation. Perhaps because the median salaries in San Jose, CA do not allow for purchasing a home at the high prices in Silicon Valley. This shows the housing affordability problem in San Jose, CA.
- The relationship between crime and weather is negatively correlated for 4 out of 5 cities. That is because the weather score (measured by total max and min temperature deviation from the "golden temperature" 71.6F) is worse (high) in Texas, but with comparatively low crime rates. The outlier is San Jose, CA, with a positive correlation because of the relatively high crime rate.

Feature Importances derived from Random Forest Home Price Model (CA + TX)

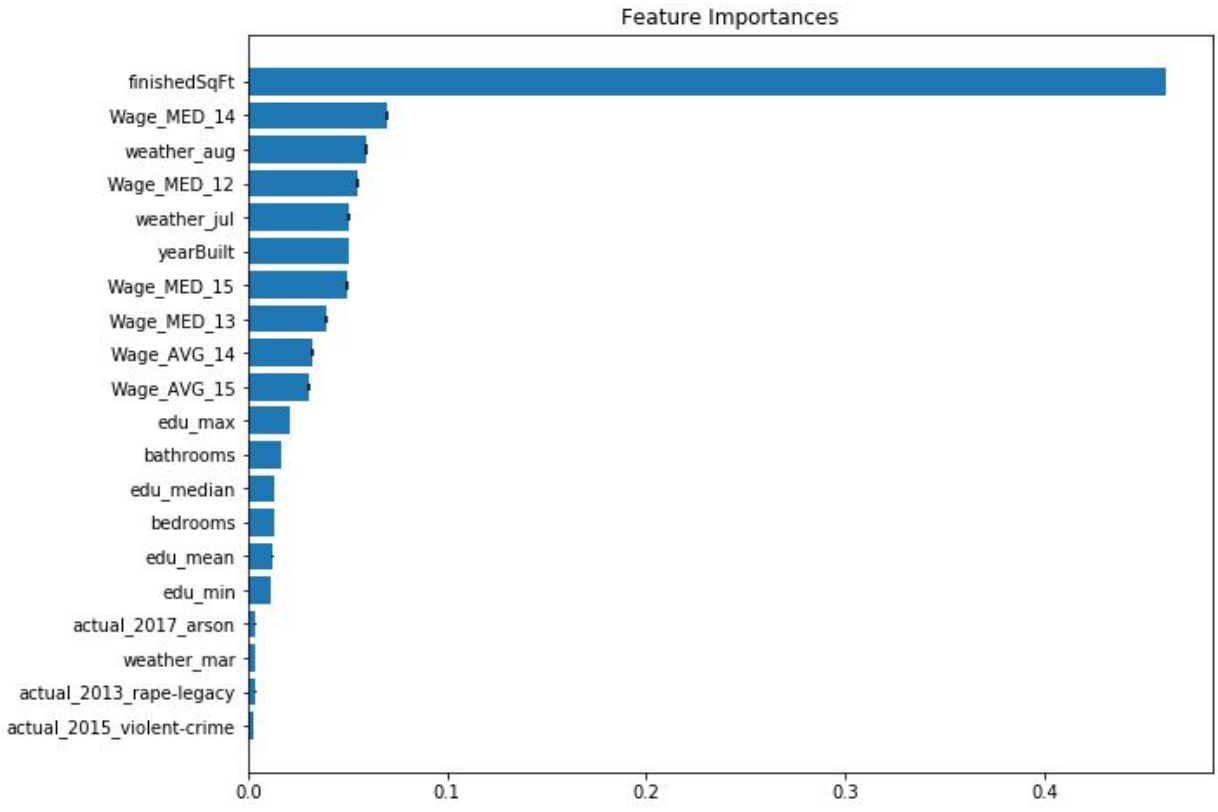

Figure 12: Feature Importance Ranking

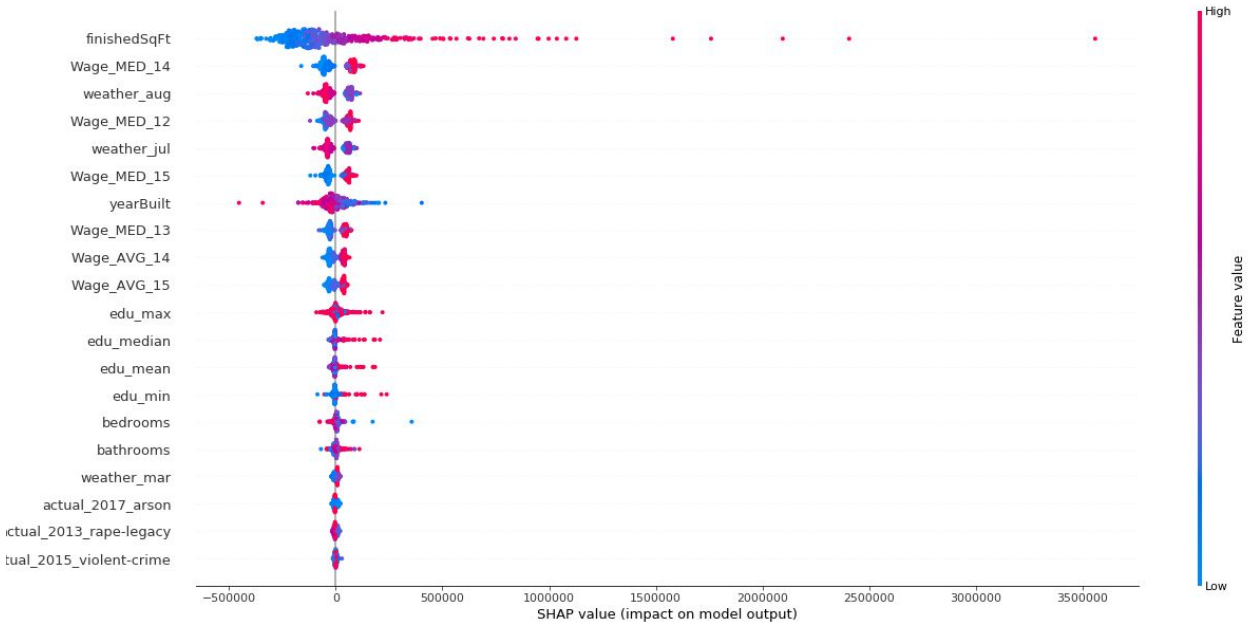

Figure 13: Shapley Additive Explanation (SHAP) Analysis on Feature Importance Ranking

The top 20 features impacting home price as determined by the Random Forest Model is seen on Figure 12. By far the most important is the internal home feature of finishedSqFt (size of the house interior). Next is the Median Wage for 2014, followed by August Weather. Skipping over redundant wage and weather features, the next important feature is yearBuilt then edu_max (maximum school score near the home), then the internal home features of bedroom and bathrooms. The features with the least impact of the top 20 are ones classified as crime (arson, rape, violent crime).

While the impact of feature importances can be seen in Figure 12, we cannot determine whether a high or low number is impacting the house price. That is where Shapley Additive Explanation (SHAP) analysis comes in. From Figure 13, we can determine that high values of finishedSqFt impact the home price the most, which is logical -- the bigger the house, the more expensive it is. Likewise, a high wage in cities also impact higher home prices. For weather, lower values (smaller deviation from the golden temperature of 71.6F) lead to higher home prices. One non-intuitive finding is that lower numbers of YearBuilt (older houses) are associated with higher prices. This can be explained through the fact that the San Jose dataset consists of many older homes (30+ years), but also make up a large portion of the high price homes of the 5 city dataset.

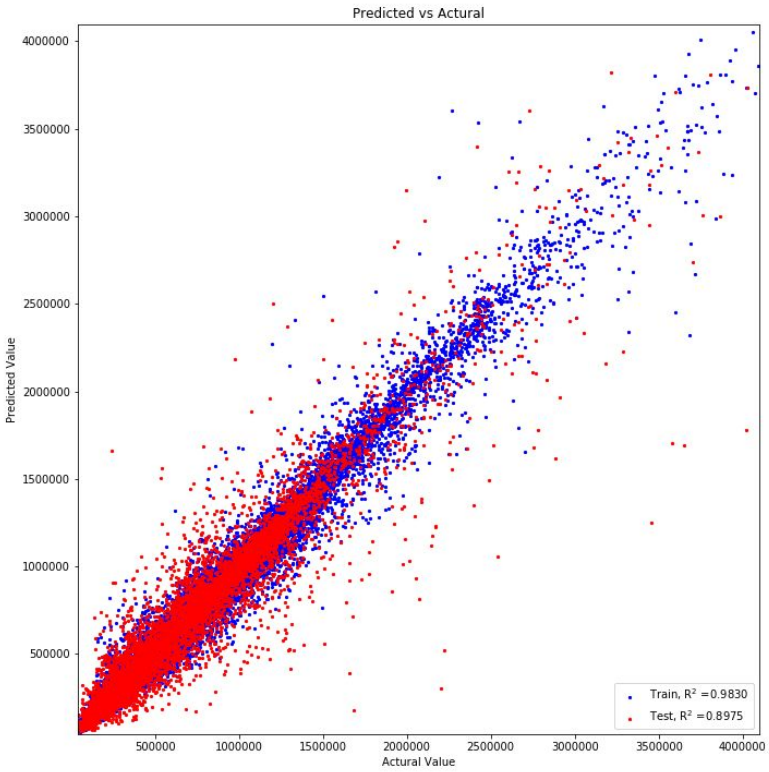

Figure 14: Random Forest Home Price Prediction vs Actual Value (Zoom to 200%)

Our Random Forest was trained on 80% of the dataset and tested on 20% of the dataset. The R^2 value of the in-sample Train dataset is 0.983. The R^2 value of the out-of-sample Test dataset is 0.8975, showing a good performance of our Random Forest Home Price prediction model based on the top 20 features.

Comparison with other Ranking Solutions

Rank	DVA Team27	Niche.com	Areavibes.com	AARP.org
1	Irvine, CA 78.68	Irvine, CA A+	Austin, TX 83	Austin, TX 57
2	Austin, TX 74.63	Austin, TX A+	Irvine, CA 82	Irvine, CA 56
3	San Antonio, TX 70.86	San Antonio, TX A-	San Antonio, TX 80	Dallas, TX 55
4	Dallas, TX 70.85	Dallas, TX A-	Dallas, TX 75	San Jose, CA 52
5	San Jose, CA 70.16	San Jose, CA B+	San Jose, CA 69	San Antonio, TX 50
Kendall τ	1.0	1.0	0.8	0.4

Figure 15: City Ranking Comparison Table

Comparing with the other city livability ranking solutions in the market today, our rankings have a Kendall-Tau rank correlation coefficient of values [1.0, 0.8, 0.4] with the rankings of Niche.com, Areavibes.com, and AARP.org, respectively.

Based upon a survey of the 5 group members and a 10 person test subject group (each member interviewed 2 of their friends/co-workers), 86.67% of the 15 people agreed with the rankings, with 2 dissenting in favor of Austin, TX over Irvine, CA as the top city, same as the rankings provided by Areavibes.com.

6 Conclusion and Discussion

Based on our default feature importances, the BHV rankings for the 5 cities are as follows:

1. Irvine, CA
2. Austin, TX
3. San Antonio, TX
4. Dallas, TX
5. San Jose, CA

One thing is consistent, San Jose, CA usually lands in the bottom 2 in our rankings unless Wage/Salary ranking is the most important feature (value = 5) for the user and all other features are de-emphasized (value ≤ 3). If home affordability factors in the user's preference at all (value ≥ 4), then San Jose will be always ranked below Austin, even if Wage/Salary is at max importance.

This conclusion is consistent with what the demographic trends show, many people, even high salaried tech workers are making the decision to move from the traditional tech hub of Silicon Valley to Austin, TX due to economic factors, primarily home affordability. However, those same people may not have to look outside of California for better homeowner value -- our rankings identified Irvine, CA as a better value than even Austin, TX.

7 Distribution of Team Member Effort

All team members have contributed a similar amount of effort.

References

1. AARP Livability Index - Great Neighborhoods for All Ages. (n.d.). Retrieved October 9, 2019, from https://livabilityindex.aarp.org/how-are-livability-scores-determined
2. Methodology for Niche Places to Live Rankings. (n.d.). Retrieved October 9, 2019, from https://www.niche.com/places-to-live/rankings/methodology/
3. Areavibes. (n.d.). Methodology. Retrieved October 9, 2019, from https://www.areavibes.com/methodology/
4. Nagaraja, C.H., Brown, L.D., & Wachter, S.M. (2011). House Price Index Methodology. Retrieved from https://faculty.wharton.upenn.edu/wp-content/uploads/2013/05/Brown_2013_House_Price_Index_Methodology_1.pdf
5. Pope, D. G., & Pope, J. C. (2012). Crime and property values: Evidence from the 1990s crime drop. Regional Science and Urban Economics, 42(1–2), 177–188. https://doi.org/10.1016/J.REGSCIURBECO.2011.08.008
6. Ihlanfeldt, K., & Mayock, T. (2010). Panel data estimates of the effects of different types of crime on housing prices. *Regional Science and Urban Economics*, *40*(2–3), 161–172. https://doi.org/10.1016/J.REGSCIURBECO.2010.02.005
7. Buonanno, P., Montolio, D. & Raya-Vílchez. (2013). Housing Prices and Crime Perception. *J.M. Empir Econ 45*: 305. https://doi.org/10.1007/s00181-012-0624-y
8. Huang, Y., Spahr, R.W., Sunderman, M.A., & Ozdenerol, E. (2015, October). The Geospatial Impact of Crime on Neighborhood Property Values. Paper submitted for presentation at 2015 FMA Annual Meeting, Orlando, Florida. Retrieved October 9, 2019, from http://www.fmaconferences.org/Orlando/Papers/TheGeospatialImpactofCrimeonNeighborhoodPropertyValues_1-15-15_FMA.pdf
9. Tita, G.E., Petras, T.L. & Greenbaum, R.T. Crime and Residential Choice: A Neighborhood Level Analysis of the Impact of Crime on Housing Prices. (2006, December). *Journal of Quantitative Criminology 22: 299*. https://doi.org/10.1007/s10940-006-9013-z
10. A.J. Chiodo, R.H. Murillo, M.T. Owyang. (2010, May). “Nonlinear Effects of School Quality on House Prices”, Federal Reserve Bank of St. Louis Review, vol.92(3), pp. 185-204, 2010. Retrieved October 8, 2019, from https://files.stlouisfed.org/files/htdocs/publications/review/10/05/Chiodo.pdf
11. Seo, Y. & Simons, R. (2009). The Effect of School Quality on Residential Sales Price. *Journal of Real Estate Research: 2009, Vol. 31, No. 3*, pp. 307-327.
12. Crone, T.M. (1998). House prices and the quality of public schools: what are we buying? *Business Review, Federal Reserve Bank of Philadelphia, issue Sep*, pp. 3-14.

13. Kane, T.J., Riegg, S. K., & Staiger, D. O. (Summer 2006). School Quality, Neighborhoods, and Housing Prices. *American Law and Economics Review, Volume 8, Issue 2*, pp. 183–212. https://doi.org/10.1093/aler/ahl007
14. Ries, J. C., & Somerville, T. (2004, Dec 20). School quality and residential values : evidence from Vancouver zoning. http://dx.doi.org/10.14288/1.0052291
15. Goodman, A.C. (1988).“An econometric model of housing price, permanent income, tenure choice, and housing demand”, *Journal of Urban Economics vol.23(3)*, pp. 327-353. Retrieved October 7, 2019, from https://www.sciencedirect.com/science/journal/00941190/23/3
16. Labor income, housing prices, and homeownership. (2006). Journal of Urban Economics 59, pp. 209-235. https://doi.org/10.1016/j.jue.2005.04.001
17. Johnes, G., & Hyclak, T. House prices and regional labor markets. (1999). *The Annals of Regional Science, Vol.33(1)*, pp.33-49
18. Frank, S. (n.d.). Home Price to Income Ratio. Retrieved October 8, 2019, from https://www.longtermtrends.net/home-price-median-annual-income-ratio/
19. Geng, B., Bao, H., & Liang, Y. A study of the effect of a high-speed rail station on spatial variations in housing price based on the hedonic model. (2015, Oct) *Habitat International. Vol 49*, 2015 pp. 333-339. https://doi.org/10.1016/j.habitatint.2015.06.005
20. So, K.S., Orazem, P.F., & Otto, D.M. (2001, Nov) The Effects of Housing Prices, Wages, and Commuting Time on Joint Residential and Job Location Choices. Retrieved October 7, 2019, from https://www.jstor.org/stable/pdf/1244712.pdf
21. Eash, D.E., Lewin, D., & Wazzan, C.P. (2017). Antipoaching Collusion in the Contemporary Labor Market: Evidence, Analysis, and Implications. *Employee Relations Law Journal; New York Vol. 43, Iss. 2*, (Autumn 2017): 50-71. Retrieved October 8, 2019, from https://search.proquest.com/docview/1925072873
22. Bastone, N. (2019, Aug 19). This former Googler says he's so tired of the astronomical housing prices in San Francisco that he bought land in Austin, Texas, instead. Retrieved October 7, 2019, from https://www.businessinsider.com/former-googler-leaves-bay-area-buys-land-austin-texas-2019-8
23. Many people are moving from California to Texas. (2019, Jun 20). *Economist*. Retrieved October 6, 2019, from https://www.economist.com/special-report/2019/06/20/many-people-are-moving-from-california-to-texas