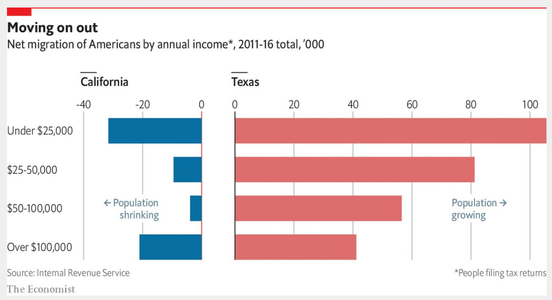
Ranking Best Cities for Tech Worker Homeownership

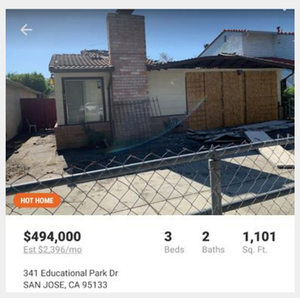
Team 27 CSE6242 Data and Visual Analytics

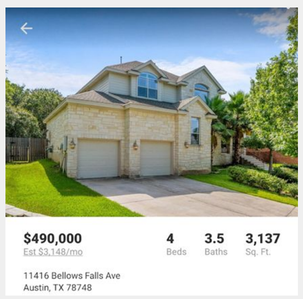
Yujing Huang, Vernon Liang, Weijia Zhan, Yichao Zhang, Ling Zhou

Motivation and Introduction

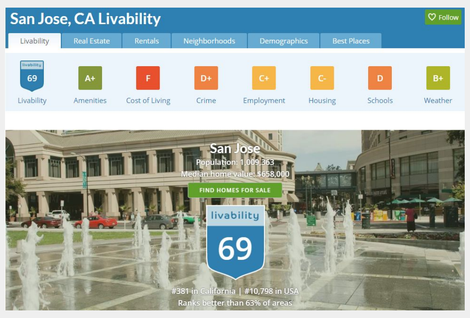
In 2019, housing prices in the major tech hub cities such as San Jose, CA have skyrocketed to more than 40% higher than peak of the mid-2000s housing bubble. In response, many tech workers in search of better value for their money are deciding to move to lower cost locations such as Austin, TX to buy a home and raise their families. Between 2007 and 2016, a net 1M residents left California for another state; Texas was the most popular destination.



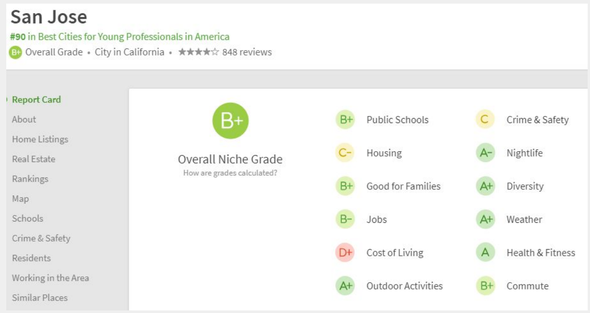




Tech workers considering this move would **want a way to evaluate the cost benefit** of relocation. **The problem is, current city livability rankings are inadequate**: they are subject to editorial bias and don't factor in the purchasing power of highly salaried tech workers.



Areavibes.com



Niche.com

We propose to solve this problem by introducing our own visualization tool that leverages machine learning algorithms (Random Forest, Feature Engineering) and statistical analysis (correlation, SHAP) on a large, real combined dataset of weather, education, crime, salaries, and house prices over the following **5 different cities in CA and TX.**

California: San Jose, Irvine **Texas:** Austin, Dallas, San Antonio

Approach and Innovation

Visualization Innovation

To address the problems with the current city livability rankings we proposed to create an innovative visualization tool that will allow:

- User Customization: the user can rank the 5 major features (house price, weather, education, crime, salaries) in order of own preference to weigh the Best Homeowner Value Index score for each city differently.
  - This **innovation** addresses the editorial bias in current solutions. By default, the features will be ranked House Price > Salary > Weather > Education > Crime, which was determined by a correlation matrix of feature importance to house price (see Experiment Section).
- Traceable Real Data on a Geographic Map: the user can go into each city view and view each data point for the 5 major features on a geographic map. Recent real tech salaries were also aggregated to increase the relevance of the rankings for the target user demographic: Software Engineers.
  - This **innovation** addresses the problem of current city livability rankings being too general to be useful for tech workers, provides traceability to real data (as opposed to current ranking solutions whose data and sometimes methodology is a black box), and superimposes that data onto a geographic map to allow new spatial/location insights into the 5 major features.

Machine Learning, Statistics Approach

Along with our innovative visualization tool, we conducted various experiments with Machine Learning Algorithms and statistical analysis to operate on the large dataset that we obtained. We employed:

- Random Forest/Decision Tree: to utilize the features (210 quantifiable features over Year 2012 to 2018) we collected.
  - **This works** by building a Random Forest Classifier to predict the housing price from the features and rank feature importance in terms of determining housing price.
  - What is new about this approach is that housing prices are usually obtained by a linear hedonic model, which is a linear weighted equation of various features. The Random Forest/Decision Tree helps to quantify any non-linear effects between features and housing price.
- Shapely Additive Explanation (SHAP) Analysis: to interpret the feature importance with high or low feature value
  - **This works** by using game theory compute Shapely values to explain a target model. One can see the relationship between the value of feature and its impact on the prediction.
- Correlation Matrix: to perform feature multicollinearity analysis, discovery of feature correlation to housing price.
  - **This works** by creating various feature matrices with each cell the computed correlation between each feature row/col.

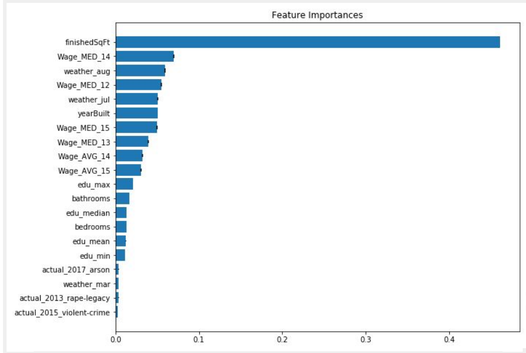
Intuition

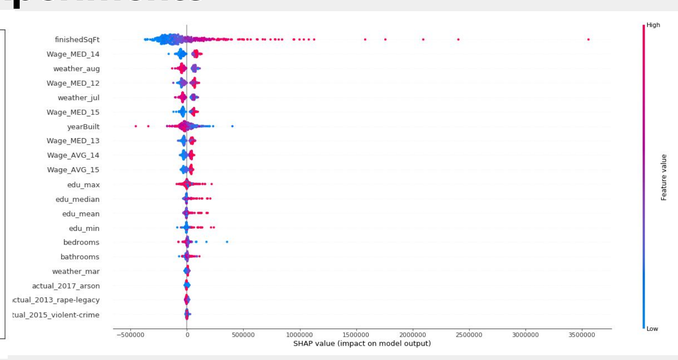
- "Location, Location, Location." -- *an old real estate proverb*
- The **intuition** behind our approaches is key insight that the housing price should encapsulate the desirability of a location due to proximity to companies paying higher salaries, low crime, better schools, better weather).
  - Need to quantify which features have a stronger effect on housing price to enable a better informed ranking of cities based on those features.
  - Intuitively, we don't believe that each of those features have an equal effect on housing prices (as some of the current livability rankings do). We also need to normalize those features across 5 different cities in different states.

Data

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)

Experiments



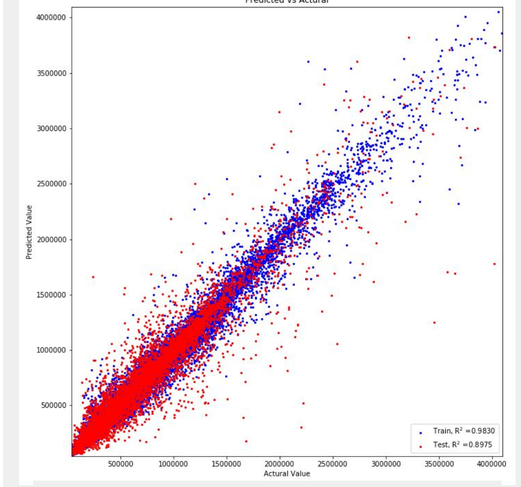


Using Random Forest, the top features above are:

1. Finished Square Feet
2. Median Wage 2014
3. August Weather
4. Year Built (after Wage 2012 and July Weather)
5. Education Score of Best Schools in Area

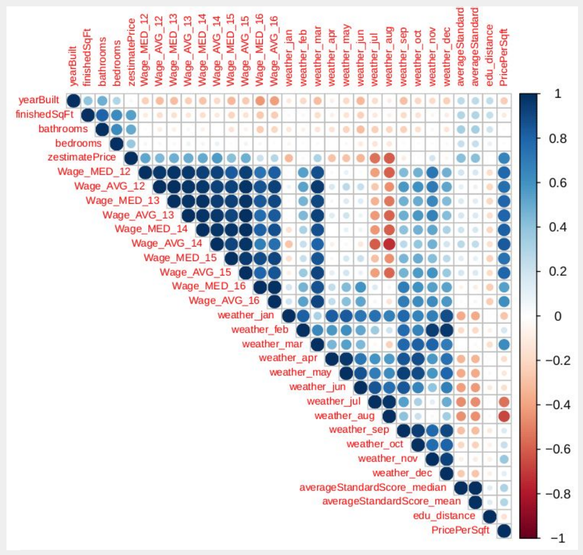
Using SHAP analysis above, Home Price is most impacted by:

- Larger finishedSqFt number (larger the house -> higher home price)
- Larger Wages in area (higher salaries -> higher home price)
- Larger Weather Scores (more deviation from 71.6F temp -> lower price). Less comfortable, especially in Summer months.
- Year Built (surprisingly newer -> lower price), data skewed by San Jose - older houses are in the best areas and more expensive, also newer houses have smaller lots



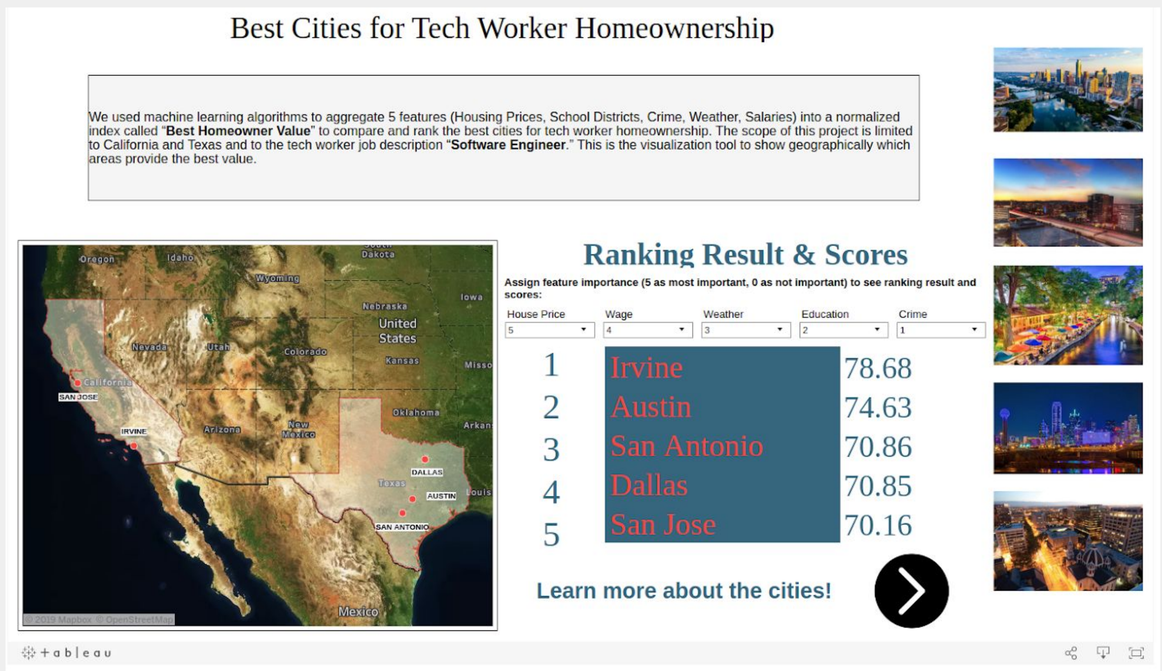
Random Forest model using Features above

- Training on 80% of data, Testing on 20%
- Highly predictive of actual Zestimate
- R^2 Train = 0.983, R^2 Test = 0.8975
- Most Red test points fall along "line of equality" Predicted Y = Actual X



Correlation Matrix show similar correlation between features, finishedSqFt highly correlated with Home Price (zestimatePrice).

Results: Visualization Tool in Tableau

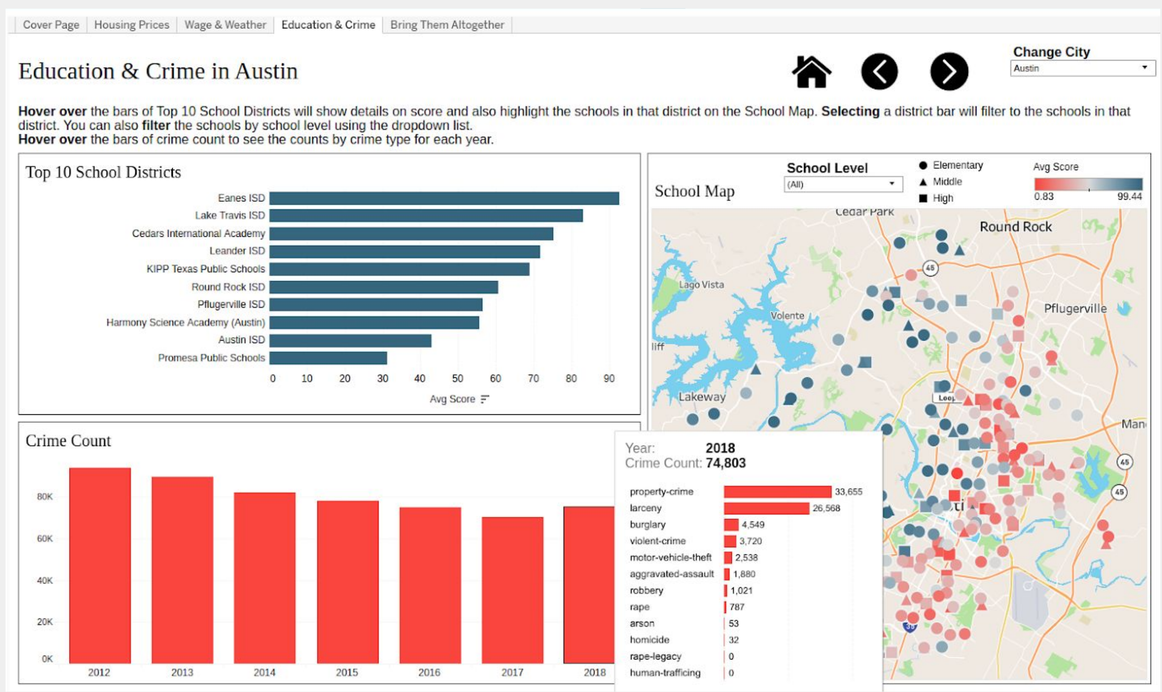


Visualization Home Page (ABOVE):

- Default Rank Home Price > Wage > Weather > Education > Crime
- Results shown above
- New Ranking can result based on User feature importance

Specific City Feature Breakdown with Map (BELOW):

- Austin, with data overlay on geographic map
- Mouseover datapoint show school name, score
- Mouseover bar, pops up breakdown of crime



Conclusion

Our Final Rankings on Best Homeowner Value over the 5 features (Home Price, Salary, Schools, Weather, Crime) based on importance of those features as calculated by our machine learning model**:**

1. Irvine, California
2. Austin, Texas
3. San Antonio, Texas
4. Dallas, Texas
5. San Jose, California

The tool also offers the user the ability to weigh on how important each of those 5 features are to them, which will change the weights on the rankings calculations and may change the rankings above. Try it out!