Bay Area Housing Data Analysis

Group 5: Xinhao Li, Kentaro Ino, Yunshan Guo, Kevin Khuu

August 9, 2016

Table of Content

1. Introduction

Housing is a necessity for all individuals. Substantial variation exists across neighborhoods in the type of housing available, the quality of public services, the level of tax burdens, and the quality of life. Consequently, prospective buyers/renters must confront important tradeoffs between different types of housing, neighborhood characteristics, and commute times when choosing suitable housing. Since housing expenditures are a large component of every household's budget, the availability of housing and its price assume considerable importance to a household's livelihood.

Our research is intended to present housing market conditions and trends in the San Francisco Bay Area in the context of historic economic events such as the Dot Com Boom, the Great Recession, and the more recent Tech Boom. Specifically, this research aims to identify changes in the housing market in terms of significant characteristics of housing and their corresponding effect on home values. After collecting a large set of relevant data from sources, we focused our analysis on three different parts of the housing market: the major growth tendency of housing prices prominent in each county, the difference in housing prices between different counties, and the relationship between population and income averages and housing prices. With this analysis, we could clearly and comprehensively map the housing environment of the Bay Area. Housing market conditions could then be a good metric to extrapolate the health of the Bay Area economy.

We surveyed several websites which contained straightforward housing price information on specific locations such as Zillow,Homefinder and Trulia. However, these sites only recorded a limited, non-representative amount of data and would've taken too much time to manually extract and clean. Fortunately, through our chosen research method, we manage to glean a substantial amount of housing data from each of our chosen counties to precisely capture the average housing price for each county, presenting an overview of residential markets in the Bay Area.

2. Data Collection and Wrangling process

2.1 Raw Data Collection

We collected our data through Quandl, a search engine primarily for numerical data, which offered access to several million financial, economic and social datasets We used the Housing API to glean housing data from Zillow and Economic data from the Federal Reserve Economic Data. By inputting the code indicating the area category and area code number for the relevant US counties, Quandl compiled a dataset containing all of the included area categories and code numbers. In order to obtain a large representative sample size in a timely manner, instead of manually inputting codes, we constructed a lookup code script only selecting the information about the Bay Area to do a loop for searching and avoid repeating the input process. Similarly, we used the same technique to obtain GDP and population growth data on Bay Area. We also included Sacramento County as a foil to Bay Area counties to see if trends from certain economic events were also replicated on any scale outside the Bay Area.

Below is a sample of raw data obtained from Quandl. *Value stands for the value of the variable Type. Type A stands for average price for all homes)*

```
## Source: local data frame [6 x 7]
##
##               Date  Value       City     County      Metro  Type
##   <int>     <date>  <dbl>      <chr>      <chr>      <chr> <chr>
## 1     1 2016-05-31 272200 Sacramento Sacramento Sacramento     A
## 2     2 2016-04-30 269000 Sacramento Sacramento Sacramento     A
## 3     3 2016-03-31 266600 Sacramento Sacramento Sacramento     A
## 4     4 2016-02-29 262300 Sacramento Sacramento Sacramento     A
## 5     5 2016-01-31 259600 Sacramento Sacramento Sacramento     A
## 6     6 2015-12-31 256600 Sacramento Sacramento Sacramento     A
```

2.2 Aggregation Process

Since we wanted to present the growth tendency of the Bay Area housing market during the past two decades, we inner joined all the separate datasets by year and county to create a master data file containing all of the relevant housing data needed to conduct our analysis.

Below is a sample cleaned data (twoB stands for the average price of two bedroom properties for that specific county, while threeB stands for the average price of three bedroom properties)

Population is in thousands

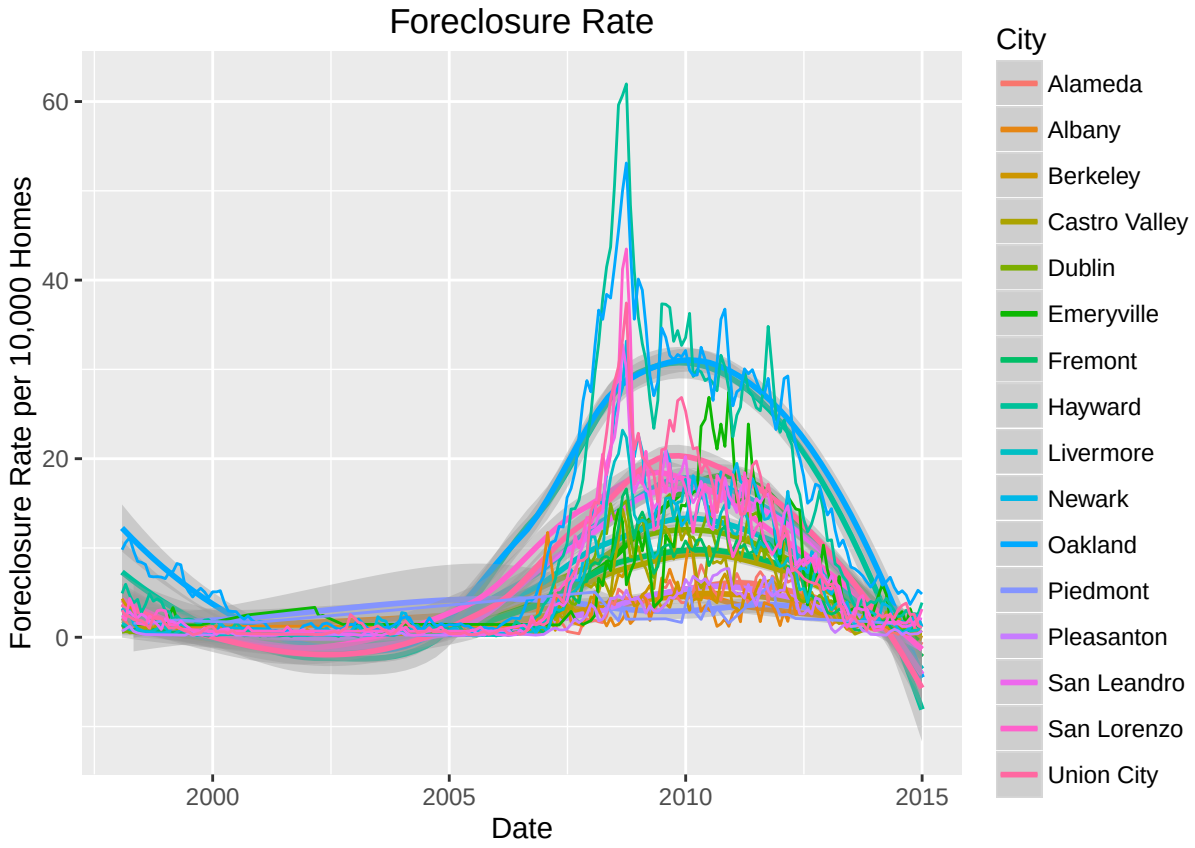
```
## Source: local data frame [6 x 37]
##
##        long     lat group order     region subregion  year      Pop Income
##       <dbl>   <dbl> <int> <int>      <chr>     <chr> <int>    <dbl>  <int>
## 1 -121.4785 37.4829   157  6965 california   alameda  1996 1359.099  28535
## 2 -121.4785 37.4829   157  6965 california   alameda  1997 1380.383  29971
## 3 -121.4785 37.4829   157  6965 california   alameda  1998 1405.903  32234
## 4 -121.4785 37.4829   157  6965 california   alameda  1999 1427.114  34513
## 5 -121.4785 37.4829   157  6965 california   alameda  2000 1450.086  39093
## 6 -121.4785 37.4829   157  6965 california   alameda  2001 1468.652  38991
## Variables not shown: twoB <dbl>, threeB <dbl>, fourB <dbl>, A <dbl>, BT
##   <dbl>, DV <dbl>, FR <dbl>, HF <dbl>, HR <dbl>, IV <dbl>, LPC <dbl>, MLP
##   <dbl>, MLPSF <dbl>, MPC <dbl>, MSP <dbl>, MSPSF <dbl>, MT <dbl>, MVSF
##   <dbl>, PRR <dbl>, RAH <dbl>, RMP <dbl>, RZSF <dbl>, SF <dbl>, SFG <dbl>,
##   SFL <dbl>, SLPR <dbl>, SPY <dbl>, TT <dbl>.
```

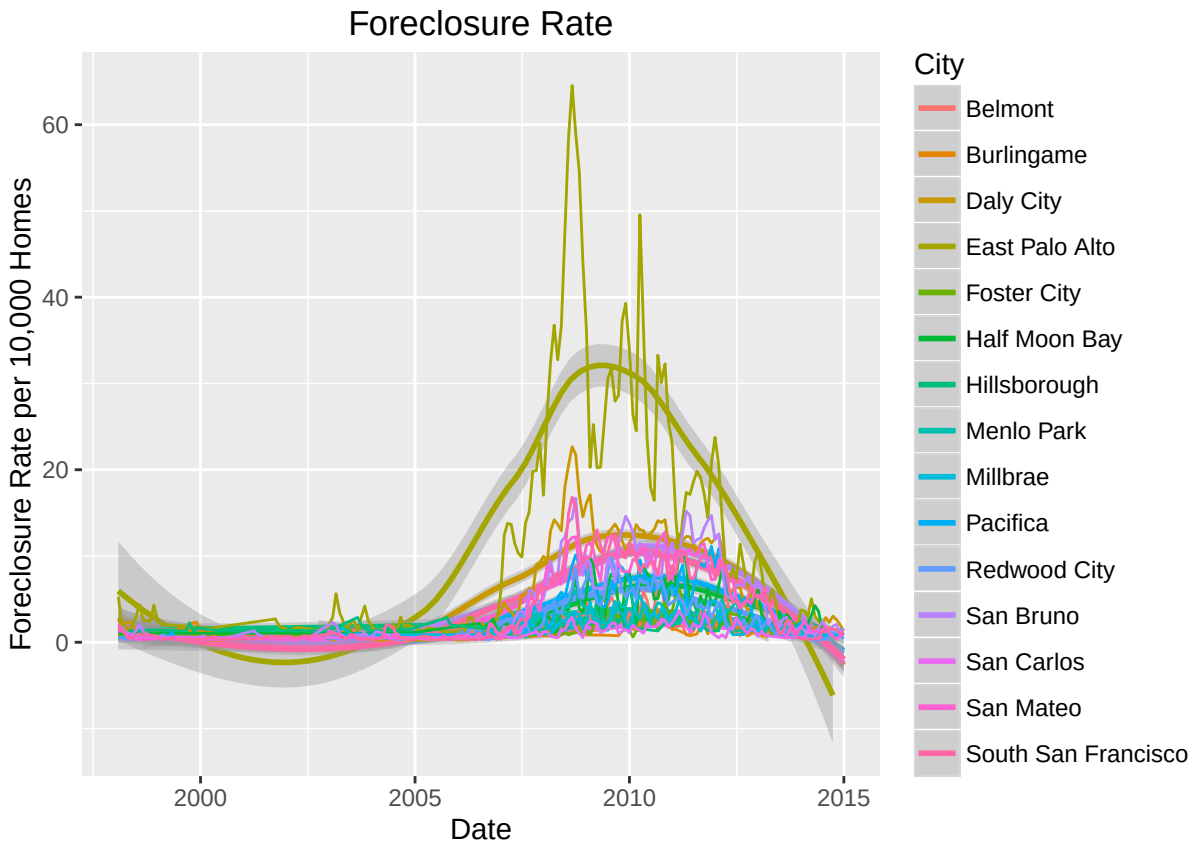
3. Analysis and Data Visualization

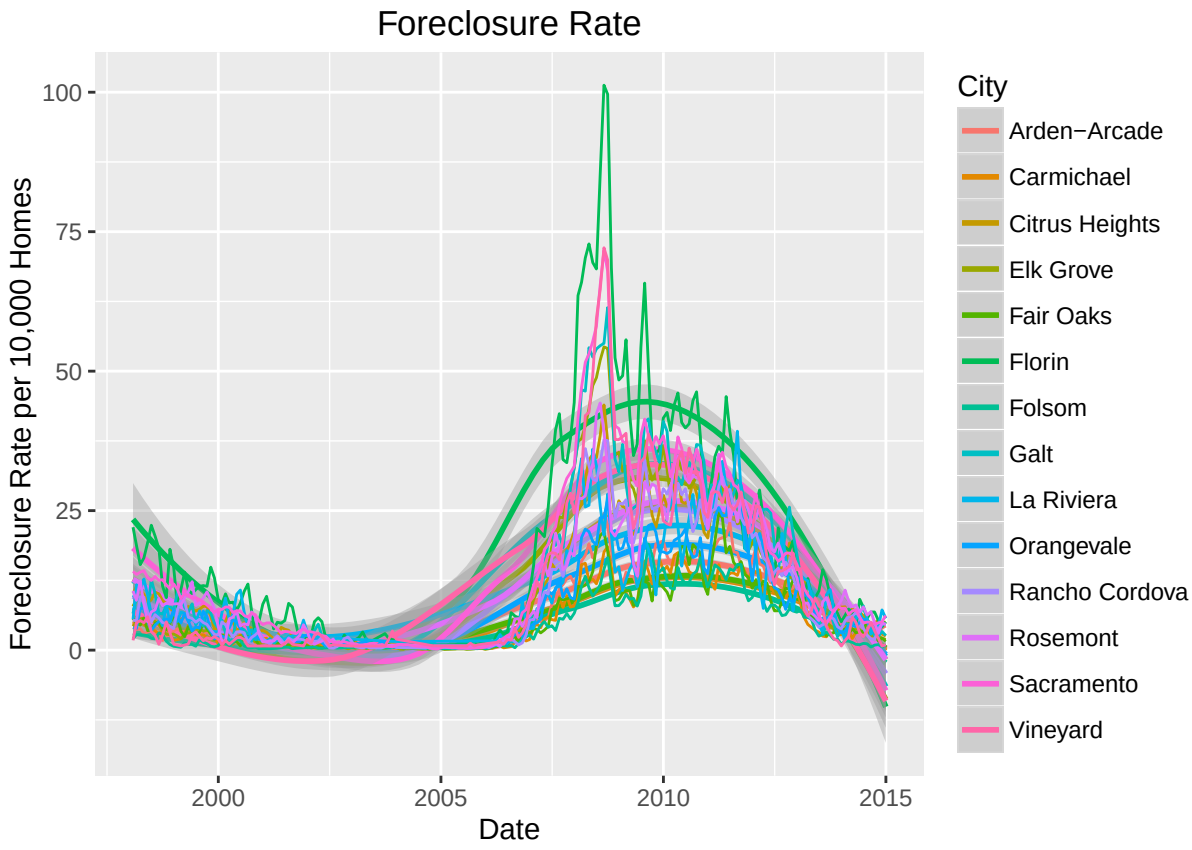
3.1 Identifying Major Factors Affecting Housing Price

*I.Home Foreclosure Rates

Home Foreclosure Rates are an important factor in reflecting the health of housing markets, and the general economy, as mortgage-backed securities are prominent goods in the financial market. We can see that home foreclosure rates remained relatively low and stable up until the 2008 Recession. This spike happens to coincide with the end of two-year teaser rates for Adjustable Rate Mortgages signed in 2005, leading to massive increases in rent and consequently, default rates. Fortunately,Home foreclosure rates for every county have since lowered sigficantly from their Great Recession numbers. We picked plots of Alameda, San Mateo, and Sacramento county to show the similarity of such a trend, despite location in proximity to Silicon Valley. However, we can see that home foreclosure rates were much lower in San Mateo than Alameda or Silicon Valley during the Great Recession, suggesting that higher-income counties were not as affected as lower income counties.

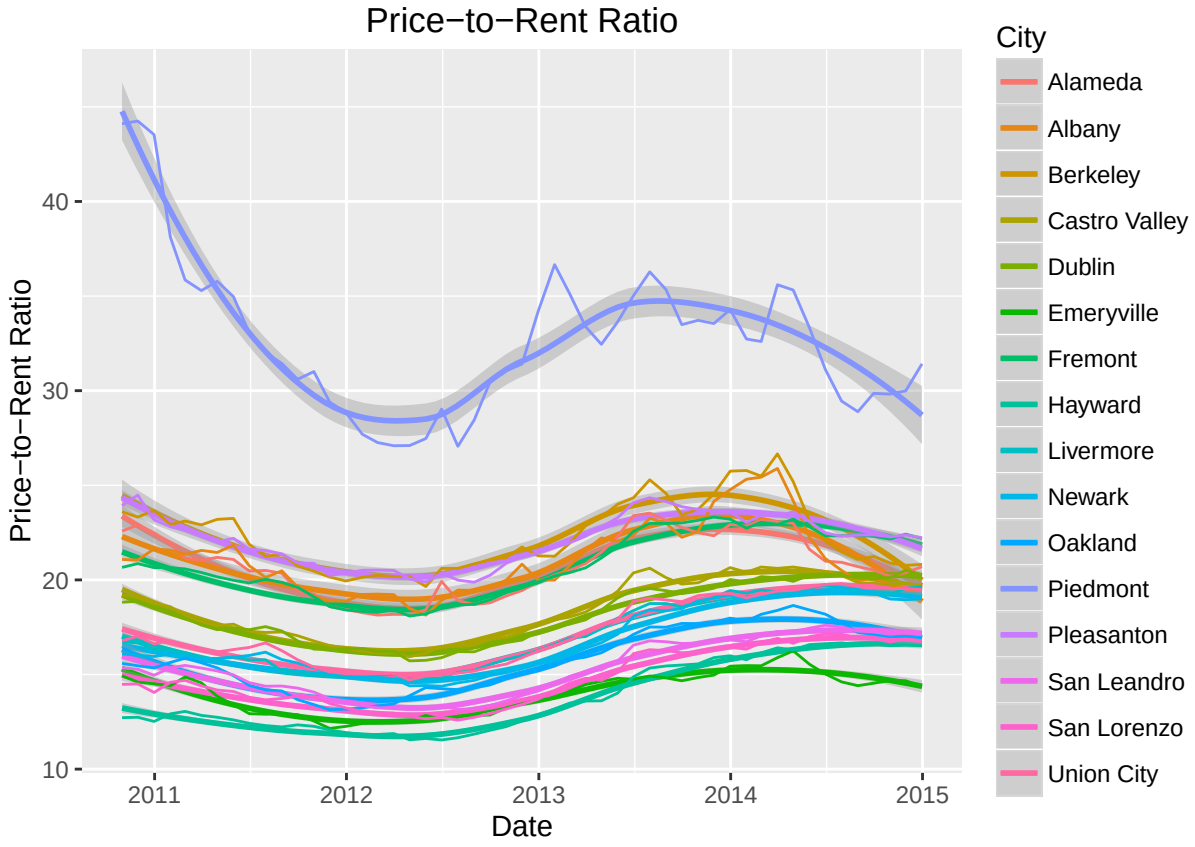


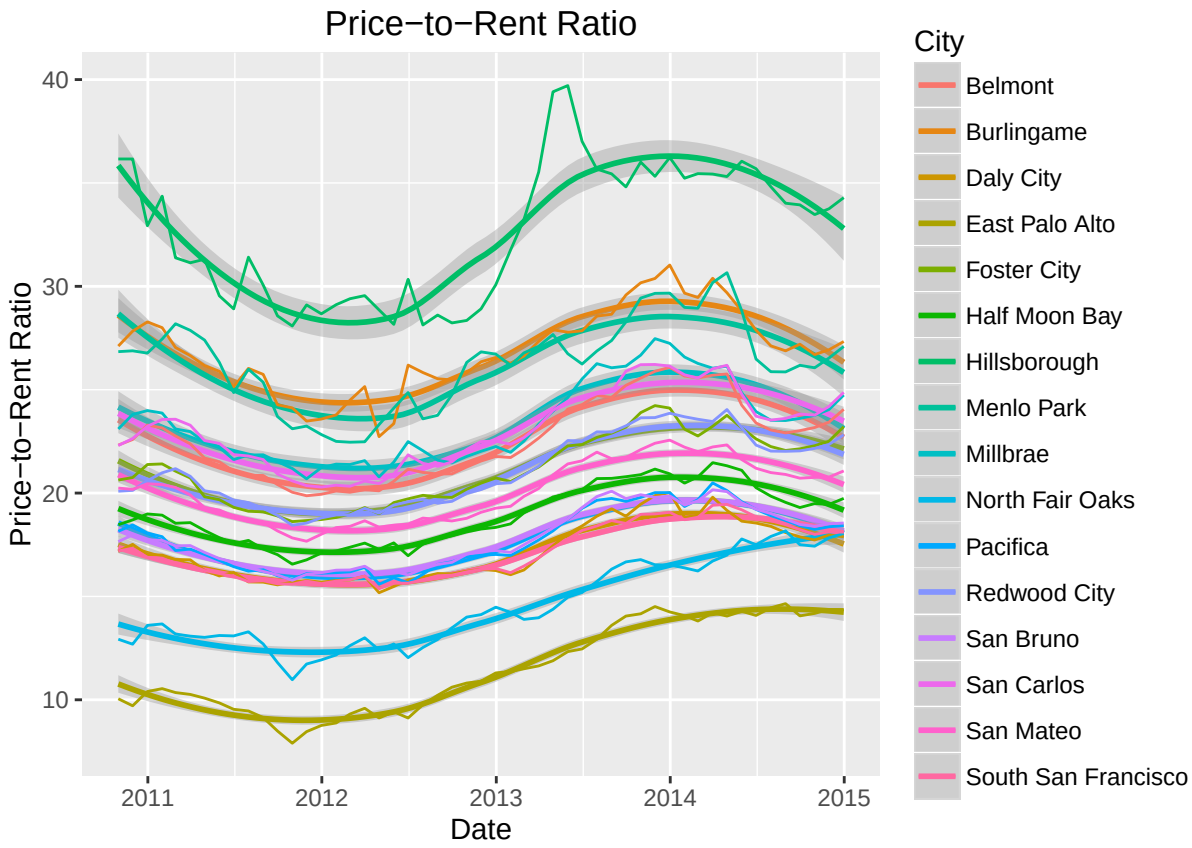

Foreclosure Rate
Foreclosure Rate per 10,000 Homes
60
40
20
0
2000
2005
2010
2015
Date
City
Belmont
Burlingame
Daly City
East Palo Alto
Foster City
Half Moon Bay
Hillsborough
Menlo Park
Millbrae
Pacifica
Redwood City
San Bruno
San Carlos
San Mateo
South San Francisco

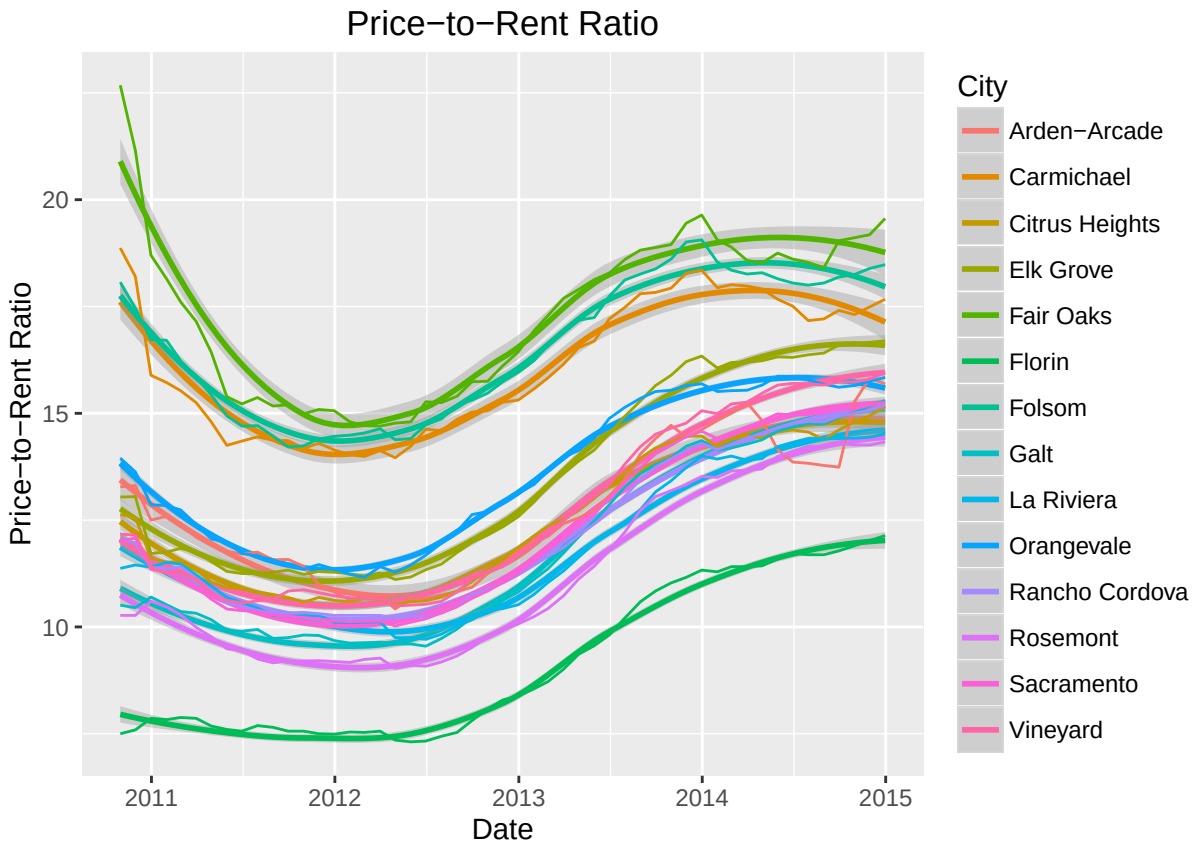

Foreclosure Rate
Foreclosure Rate per 10,000 Homes
100
75
50
25
0
2000
2005
2010
2015
Date
City
Arden−Arcade
Carmichael
Citrus Heights
Elk Grove
Fair Oaks
Florin
Folsom
Galt
La Riviera
Orangevale
Rancho Cordova
Rosemont
Sacramento
Vineyard

- II. Price to Rent Ratio

Price-to-Rent Ratio(or PRR) is another important factor. A price-to-Rent Ratio of a property is the value of the property divided by the annual rent that could be gained by that property. A high Price-to-Rent Ratio is an indication that it is better to rent a property than to buy it. The Bay Area has always had a higher price-to-rent ratio relative to the U.S. average, which is 19. Silicon Valley counties and San Francisco have higher than average Price-to-Rent Ratios with average PRR ratios in the East Bay and Sacramento lower than the national average. However, Price-to-Rent-Ratio just measures the relative cost/benefit between buying or renting a property and nothing about the absolute cost of living in the Bay Area, which is still one of the highest in the country.This data is of Alameda,San Mateo and Sacramento county for similar reasons to the Home Foreclosure analysis.




Price−to−Rent Ratio
Price−to−Rent Ratio
40
30
20
10
2011
2012
2013
2014
2015
Date
City
Belmont
Burlingame
Daly City
East Palo Alto
Foster City
Half Moon Bay
Hillsborough
Menlo Park
Millbrae
North Fair Oaks
Pacifica
Redwood City
San Bruno
San Carlos
San Mateo
South San Francisco


Price−to−Rent Ratio
Price−to−Rent Ratio
20
15
10
2011
2012
2013
2014
2015
Date
City
Arden−Arcade
Carmichael
Citrus Heights
Elk Grove
Fair Oaks
Florin
Folsom
Galt
La Riviera
Orangevale
Rancho Cordova
Rosemont
Sacramento
Vineyard

3.2 Comparisons and Analysis of Housing Market trends during Past Two Decades

Shiny Map Application(https://jeromexlee.shinyapps.io/map_graph/)

Use the Shiny App above to find plots depicting the average price of ALL homes in each different county, selected by specific years to show the historical economic events and their effect on the housing market. Change the factor to "A" which represents the average prices of all homes to follow closely with the analysis in "General Trends"

General Trends

From 1990 to 1995, housing prices moved significantly. They increased dramatically during the Dot Com Boom before dropping immensely during the Great Recession. However, we can observe that housing prices have recovered and even surpassed pre-Great Recession levels by 2014.

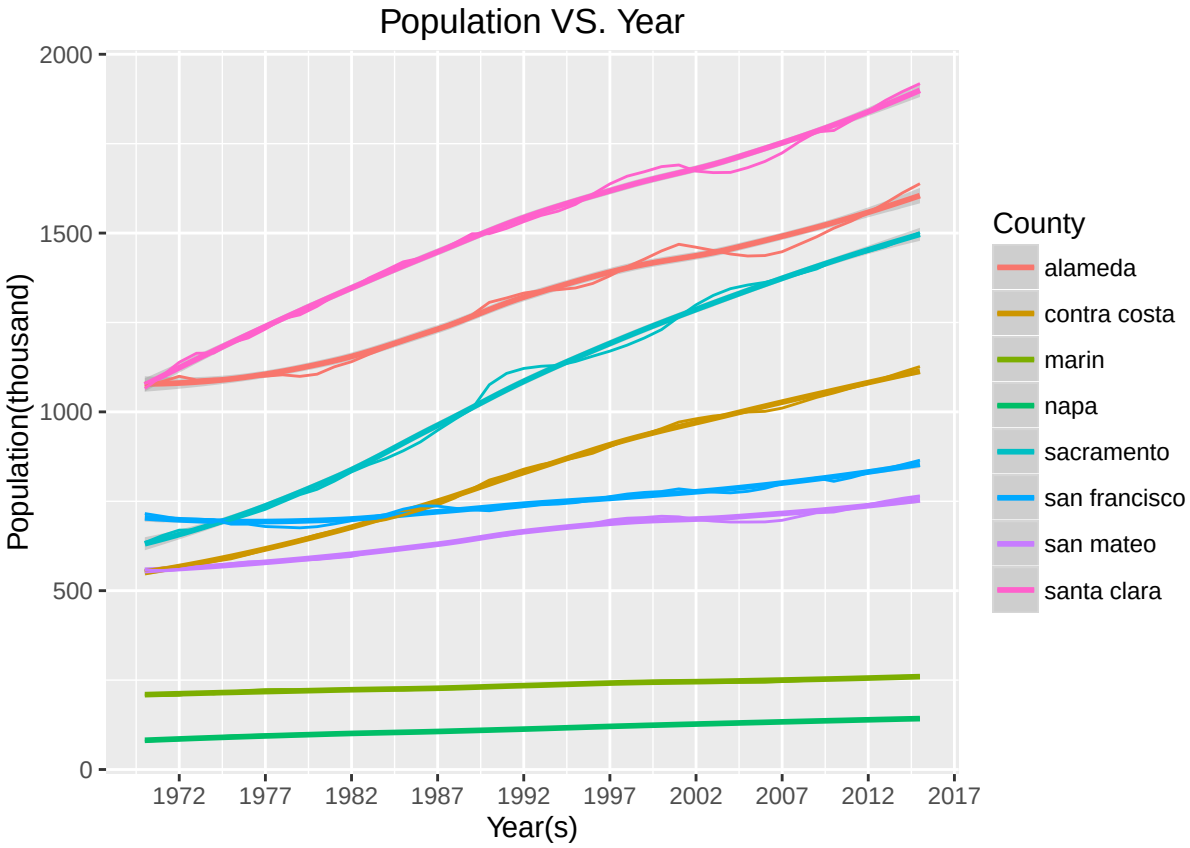
When analyzing the present Bay Area housing market, one of the more intuitive questions that an individual might ponder is: Which counties are the most expensive to live in? Before using real data to illustrate, an individual might suspect that San Francisco and Santa Clara would have higher average incomes and property values than counties in the North and East Bay, given the Tech Boom's concentration around Silicon Valley.

Housing data of each county on specific year was plotted as a map to visualize the price distribution of all homes between different counties. The result mainly confirmed the previous assumption: Santa Clara county contains the most valuable properties followed by San Francisco, San Mateo and Marin, East Bay counties, and Sacramento. Location is a good causal link for this analysis as Silicon Valley has always been the Mecca of Technology and it's no surprise that closer proximity to Silicon Valley is linked with higher property values.

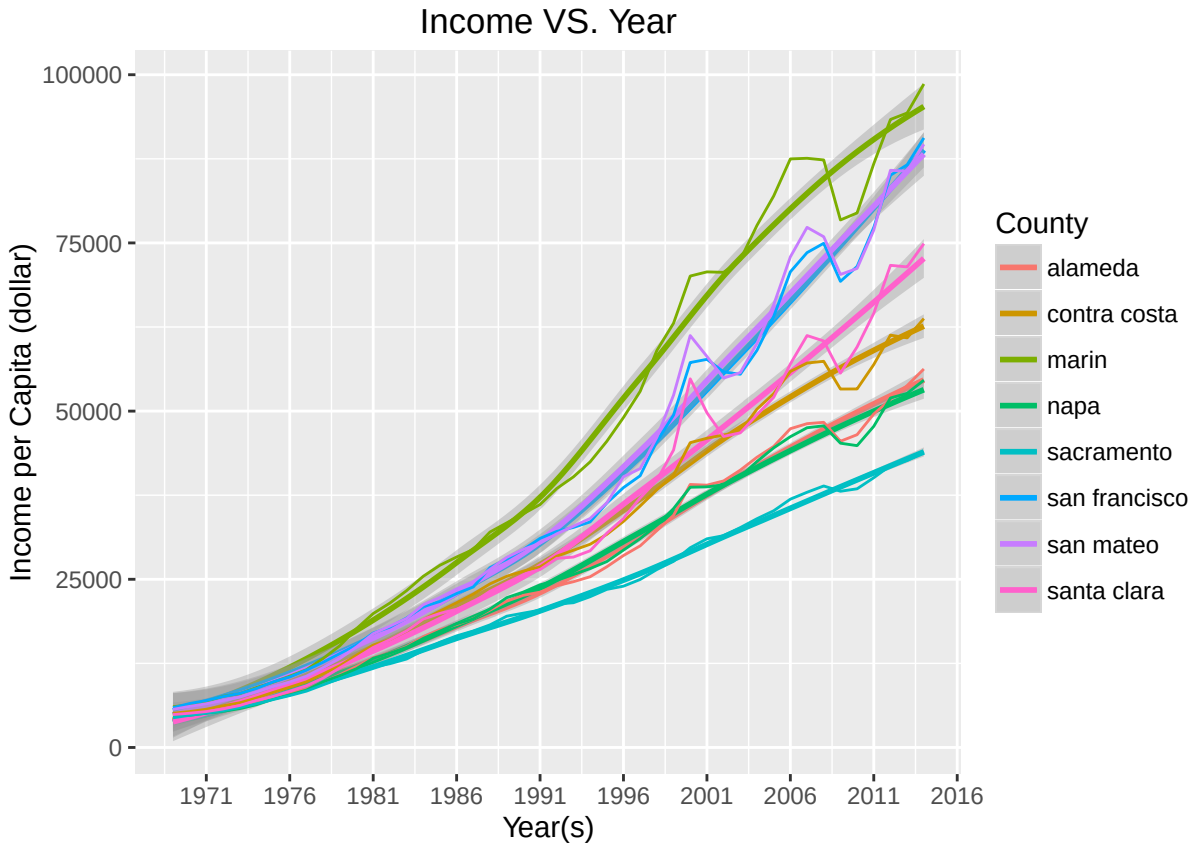
As shown in the Map-App, Santa Clara and Marin counties had the most valuable homes during the 1990s. As time passed, housing prices in Marin County maintained their high values. However, Santa Clara county housing prices have experienced great fluctuation during the past two decades. Santa Clara county dropped from the top position during the Great Recession and has only recently retained its top position, as a result of the Tech Boom.

3.3 Special Case Analysis

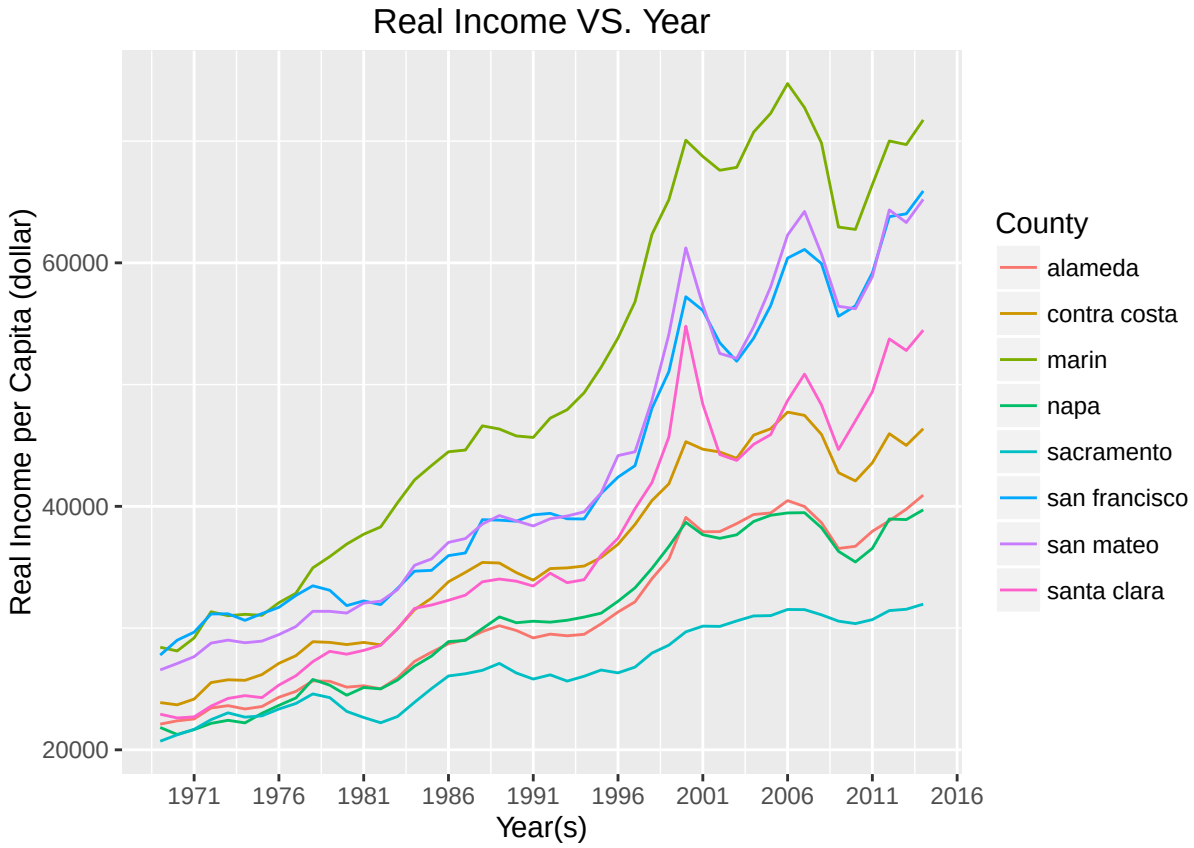
In this section, we analyze Population and Income trends in the Bay Area and Sacramento and then see how these trends in these factors are reflected in their respective housing markets.



We can see that Population has risen in the East Bay and Silicon Valley areas, while remaining stagnant for Napa,Sacramento and most interestingly, Marin county.

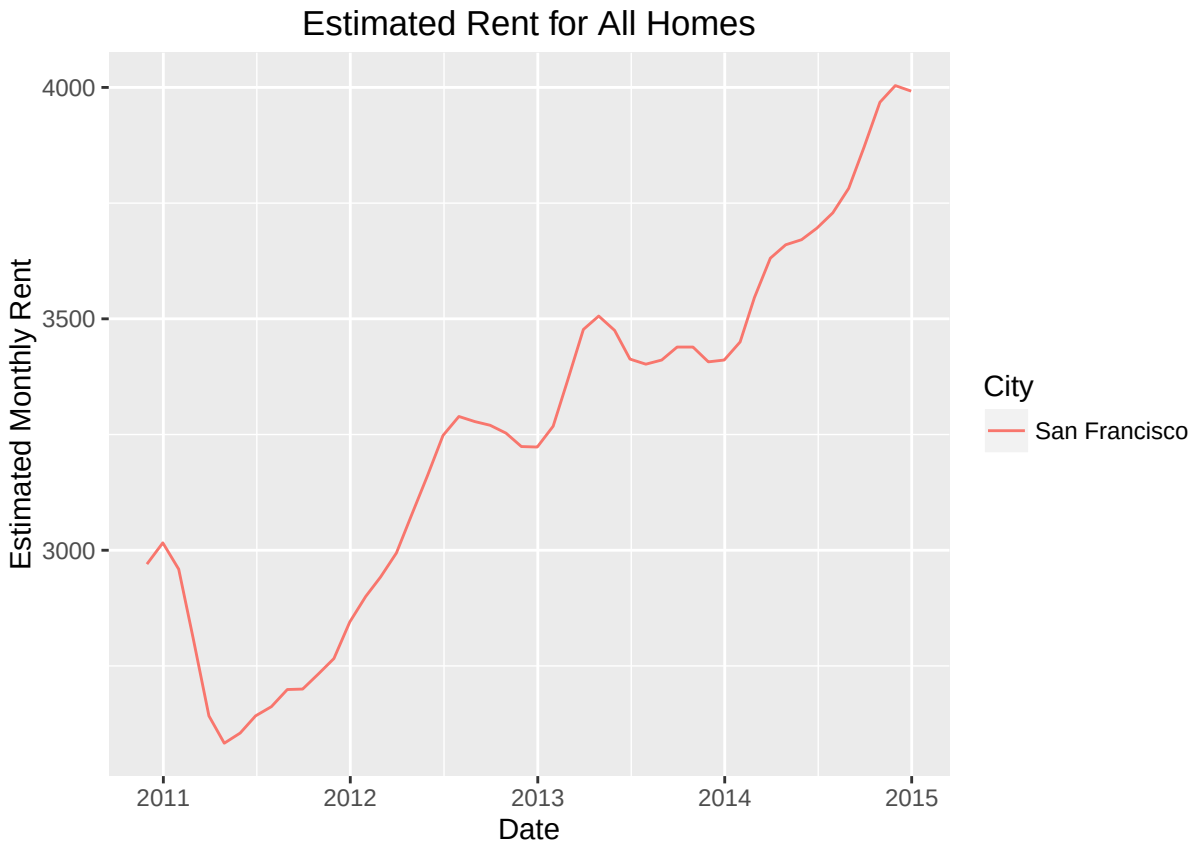


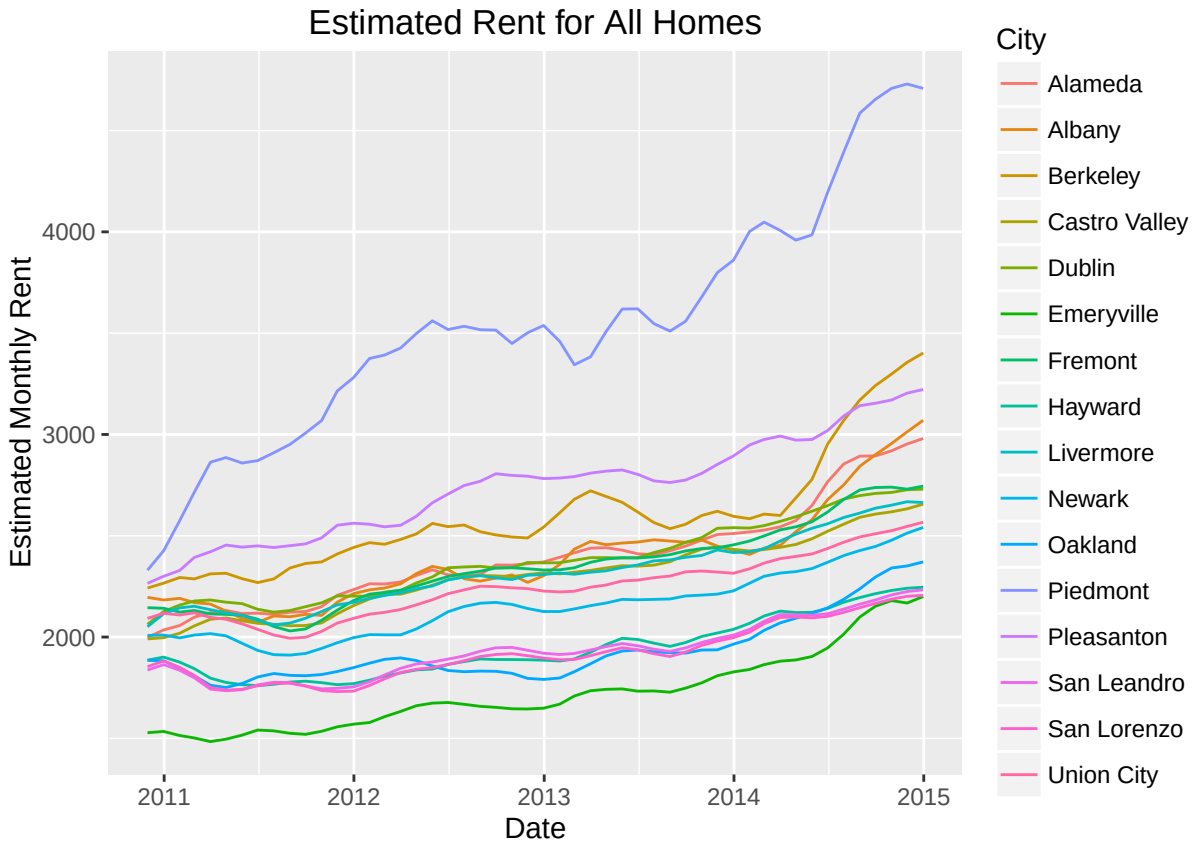
As we can see there has been a sizeable growth in average income across all counties in the Bay area since the 1970s,including heavy growth over the past few years after the recession and during the Tech Boom. However, this picture does not take into account factors such as Inflation and Cost of Living expenses.



This is the Real Income graph, which was created by dividing the original income data with the Consumer Price Index for each year of that data. This adjusted income was benchmarked against Year 2000 dollars.

We can see that average Real Income for most counties have not recovered to their Dot Com Boom levels. Also, given that Income has risen significantly over the past 15 years, why hasn't Real income risen as drastically also? Well, the Consumer Price Index includes Housing costs as one of the many goods in its index.





Estimated rent is one of the most interesting variables to observe in relation to income. We plotted Average

Estimated Rent in San Francisco because it has experienced the most meteoritic of rises in over the last four years, growing from an average of 2500 in 2011 to 4000 in 2014, a 60% rise. You can see the prices of other housing markets in the Bay Area also begin to rise during late 2013-2014. One of the reasons for this is that prospective renters and buyers might have been priced out of the San Francisco market and flowed into other housing markets, increasing the demand in these outside markets and increasing prices.

Bubble_App

Bubble Data App. This app captures a more intimate picture of the relationships of Population or Income with different Types included from the cleaned bubble Data Set, allowing the user to also observe such trends by county.

Link to Bubble_data App(https://jeromexlee.shinyapps.io/bubble/)

4. Limitations and Further Questions:

One glaring omission that readers might notice from this paper is the absence of Santa Clara data from graphs capturing different types of factors including Price-to-Rent Ratio graphs and the Home foreclosure rate graphs. Unfortunately, due to complications from the Quandl API regarding access requests, we had to exclude Santa Clara county data for many of the factors,including home foreclosure rates and estimated rent.

Professor Do suggested that we separate our income data into quintiles, in order to see trends amongst all income brackets. While the overall trend in income was highly positive, he posited that the growth in income for the bottom 80% would've been flat and neglible, with most of the growth being concentrated in the uppermost quintile. Unfortunately, we could not separate the income dataset accurately into quintiles as our dataset consisted of the average income of cities, not individual incomes within every city. A further search into this question with appropriately cleaned individual income data could be conducted to confirm the Professor's hypothesis.

Our datasets are only recent up to 2014 so this we couldn't analyze whether trends continued or whether they stabilized through 2015.

It remains to be seen whether the high growth seen from both income and housing prices from 2011-2015 was only prevalent in the Bay Area. A further study could be a comparative analysis of whether other tech hubs like Seattle, Portland, New York City, and Austin to see if any of these metropolitan areas achieved the same growth seen in the Bay Area or on any comparable scale.

5. Conclusion

Through statistical analysis of housing datasets, we were able to isolate major factors that mirrored economic trends in the housing market. For example, we were able to capture home foreclosure rates and their trends during the Great Recession. In addition, we were able to see the differences in these trends among different Bay Area counties. One of the more intriguing observations to be taken from this paper is that increases in income in the Bay Area have been met with corresponding increases in housing costs and cost of living, meaning that Real Income has not risen quite as dramatically and has only recently recovered to Dot Com Boom levels for most counties in our sample.