
# CBS Boston Census Analysis

-- CS 506 Project Report

Team members:

Chen Dong

Kaijun Wang

# Abstract

Our client WBZ-TV (CBS Boston) is a local news station that provides stories about the local community and events surrounding them. Our team was tasked to analyze US Census data and answer questions relating to vacant houses within the Boston area as well as perform further analysis to answer more questions relating to housing. This analysis could be used to generate story leads to share with our client's viewers. We analyzed the housing trend over the years in the Boston area from different aspects including occupancy, housing value, household income, mortgage status, and occupancy type (renters/owners occupied). To further strengthen and confirm our trend analysis we conducted similar analysis and compared our findings on cities similar to Boston, aided by our K-means clustering results. As an extension of our research analysis, we also explored geographic mobility and jobs offering to help us understand whether the population in Boston was increasing or decreasing that could give further insight into the housing trend we studied.

# Data

Most of the analysis was carried out using the US Census Data as it provided reliable large scale housing information. The data we used is from 2010 to 2019. Some specific topics pulled from the census data for our analysis include: occupancy, housing costs, number of residents, mean and median wage, on-sale households, mortgage prices, housing tenure, and geographic mobility.

With the help of our Project Manager Kamran, we were able to research and compare the data from Center for Housing Data and DataCommon, which were used by the previous teams who did similar projects, for more information about the housing situations of Massachusetts and Boston.

Moreover, we pulled out the data from simplemaps, which is the United States Cities database for our clustering analysis. We were only able to use the free version of the database. The data we pulled out was still accurate but we were unable to access all potentially useful features to do comparison among cities.

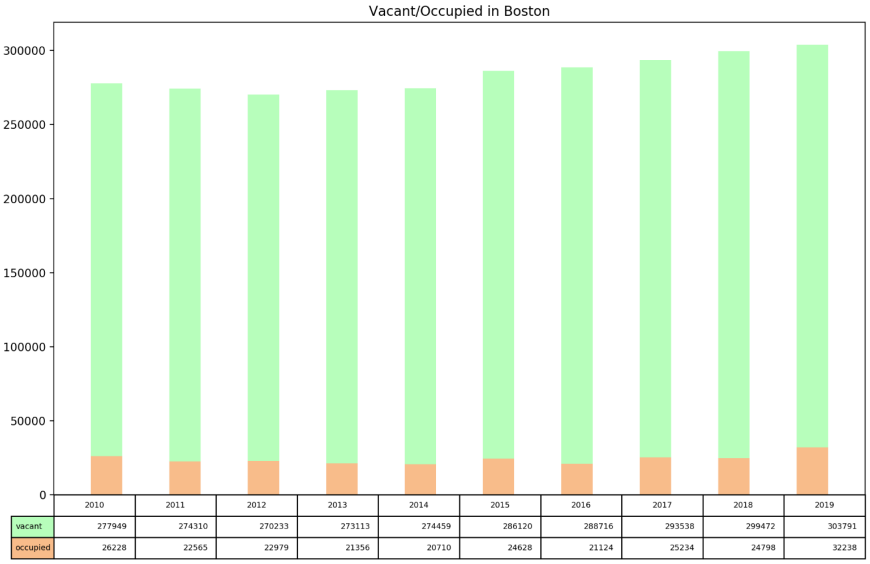
# Key Questions

As the project progressed, we tried to answer the following questions. The analysis and results of these questions will be elaborated in detail in the later section.

1) How many housing units are in Boston and how many of them are occupied? (Are there many “zombie” houses, and did that number change over years?)
2) How many housing units have mortgages and what is the owner's cost for housing units with and without a mortgage?
3) How does Boston monthly housing costs as percentages of income change over time?
4) How do Boston housing prices change over time?
5) Did the cities that are similar to Boston have the same housing features?
6) Did the population change in Massachusetts or in Boston? (Did we have any “flight”, or did we attract more people to move here?)

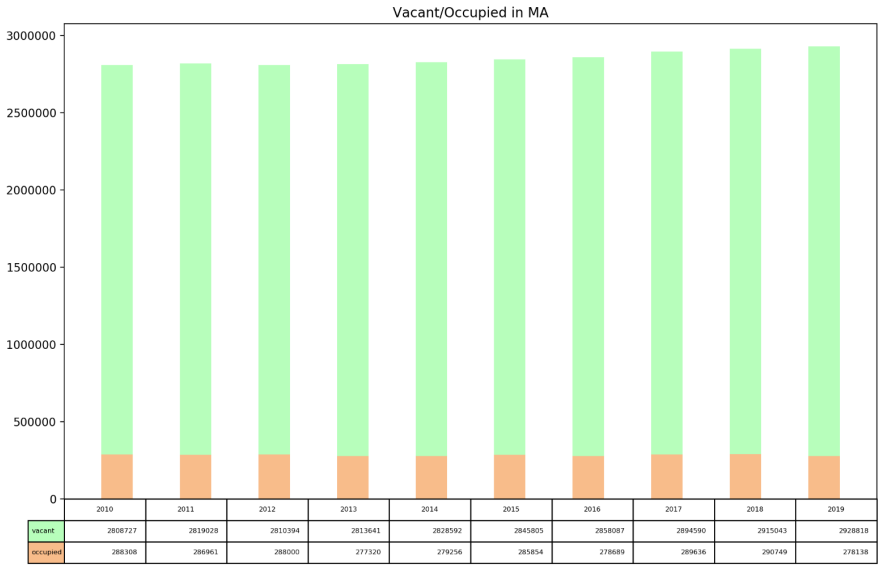
# Analysis - Boston

The first step we took is to analyze the occupancy in Boston and Massachusetts (MA). Figure 1 and 2 are the plots for the occupancy situations for Boston and MA over the past 10 years respectively. The occupancy rate in Boston has increased to 9.6% in 2019 as compared to 8.6% in 2010 and the lowest occupancy rate is approximately only 6.8% in 2016. Similar trend is observed in MA. The occupancy rate has increased to 9.3% in 2019 as compared to 8.7% in 2010. Although both the occupancy rate of Boston and MA have increased throughout the years, the number of vacant units also increased especially in Boston. From figure 1, the number of vacant units is 277,949 in 2010 and 303,791 in 2019, which is approximately 8.5% increase.




| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 |
|---|---|---|---|---|---|---|---|---|---|---|
| vacant | 277949 | 274310 | 270233 | 273113 | 274459 | 286120 | 288716 | 293538 | 299472 | 303791 |
| occupied | 26228 | 22565 | 22979 | 21356 | 20710 | 24628 | 21124 | 25234 | 24798 | 32238 |

Figure 1.




| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 |
|---|---|---|---|---|---|---|---|---|---|---|
| vacant | 2808727 | 2819028 | 2810394 | 2813641 | 2828592 | 2845805 | 2858087 | 2894590 | 2915043 | 2928818 |
| occupied | 288308 | 286961 | 288000 | 277320 | 279256 | 285854 | 278689 | 289636 | 290749 | 278138 |

Figure 2.

In order to have a better understanding of the vacancy trend as mentioned above, we explored different features of Boston housing. We firstly looked at the housing values in Boston. As shown in figure 3, the median of housing values in Boston has increased largely from 2010 to 2019 and it shows a stable increasing trend from 2013 onwards. We used median instead of average here to mitigate the impact of outliers. Figure 4 shows the detailed housing value range breakdown. From 2013 onwards, there is an increase in the more costly housing units as compared to the lower-price units. And this explains why the median of housing values in Boston shows an increasing trend. Overall, the increase in housing value shows a non-decreasing

interest towards Boston housing. The steady increase in housing value makes Boston housing a good investment option which aligns with the increasing vacancy rate we noticed.

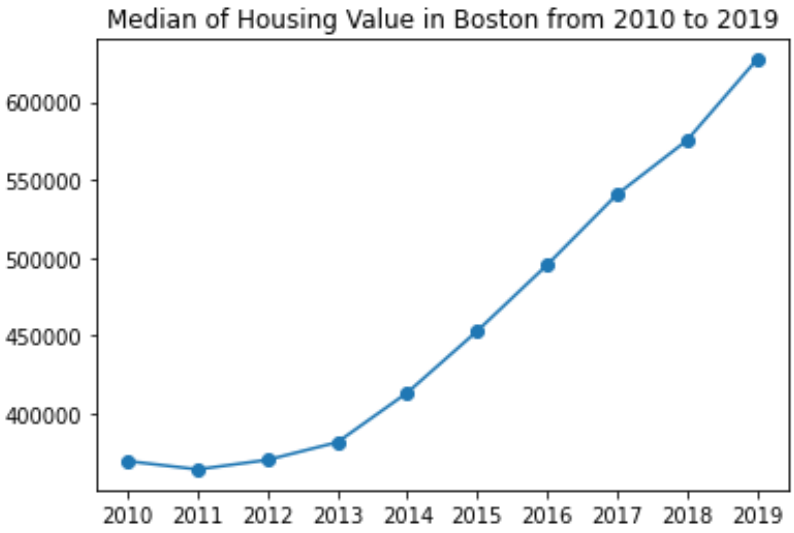



Figure 3.

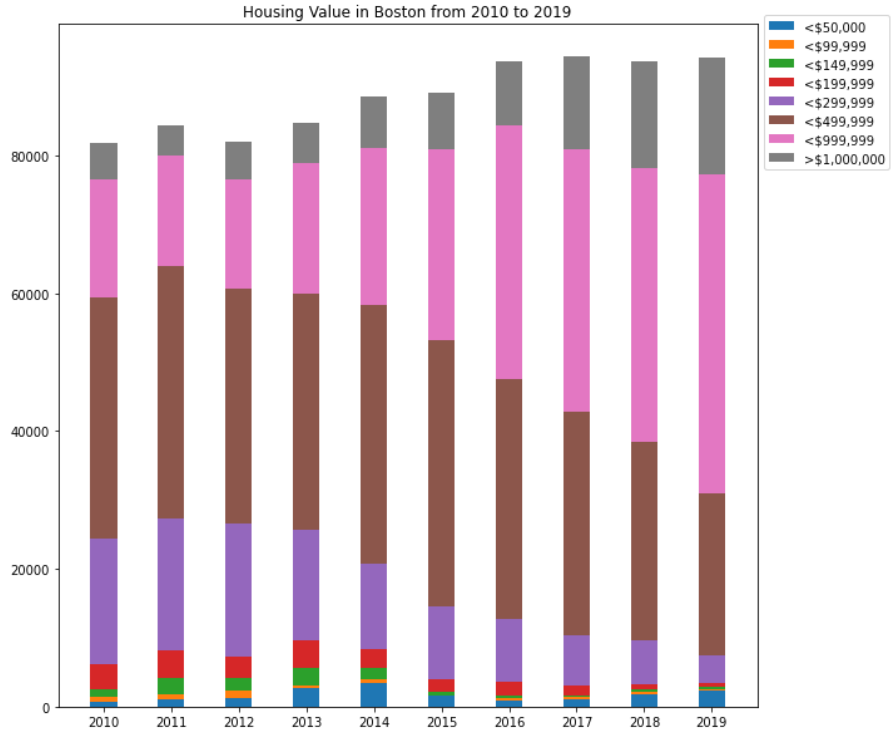
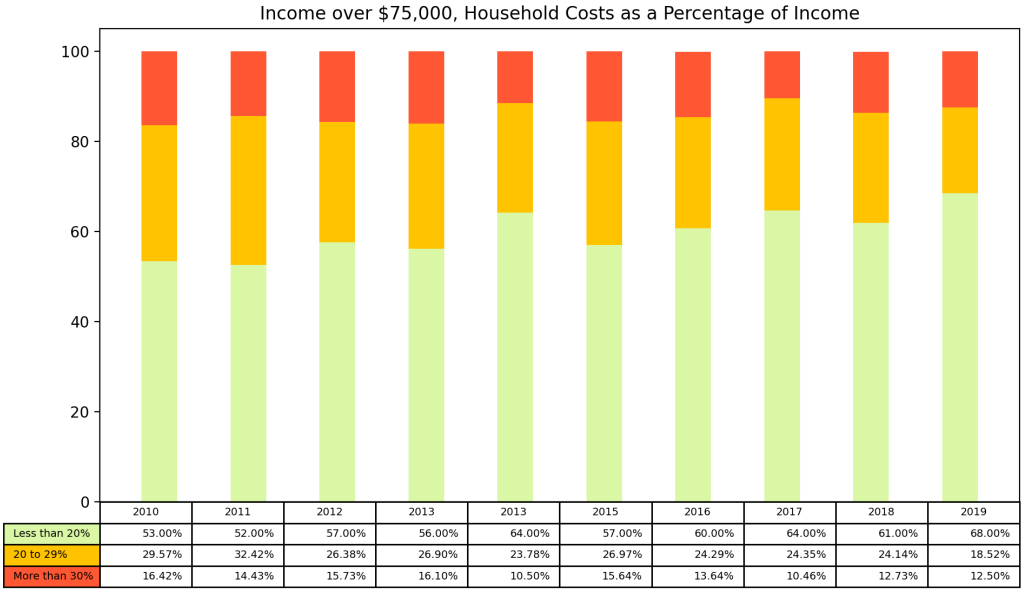



Figure 4.

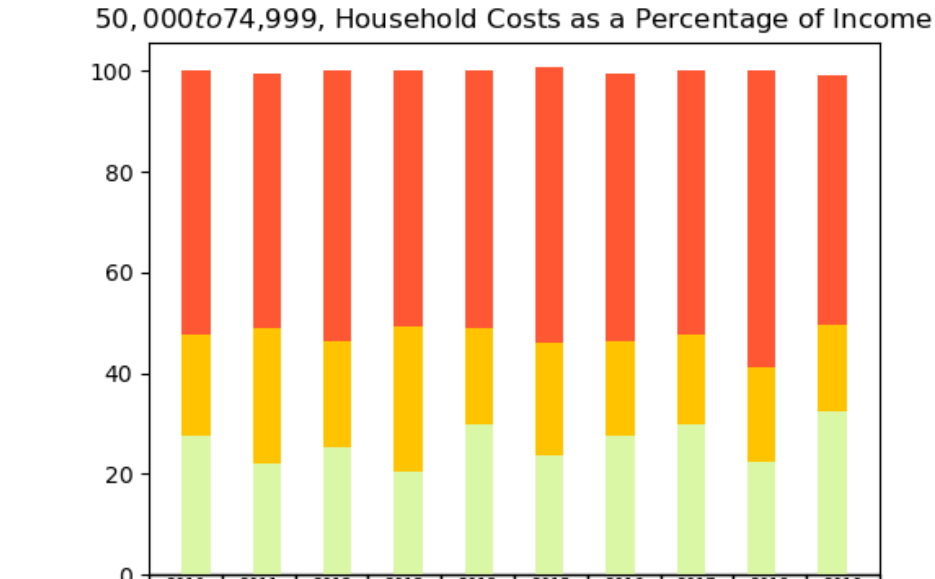
The next feature we considered is housing costs as a percentage of income. As shown in the bar charts below (figure 5 to figure 7), the percentage of households with incomes over $75,000 spending less than 20% on house costs has steadily decreased. More than 50% of households

with incomes below $75,000 spend more than 30% of their income on housing costs. This percentage has fluctuated but overall did not trend upward or downard. There are cases where the bars do not go up to 100%. This is due to a small (~2-3%) of households not having rent or not having any income.



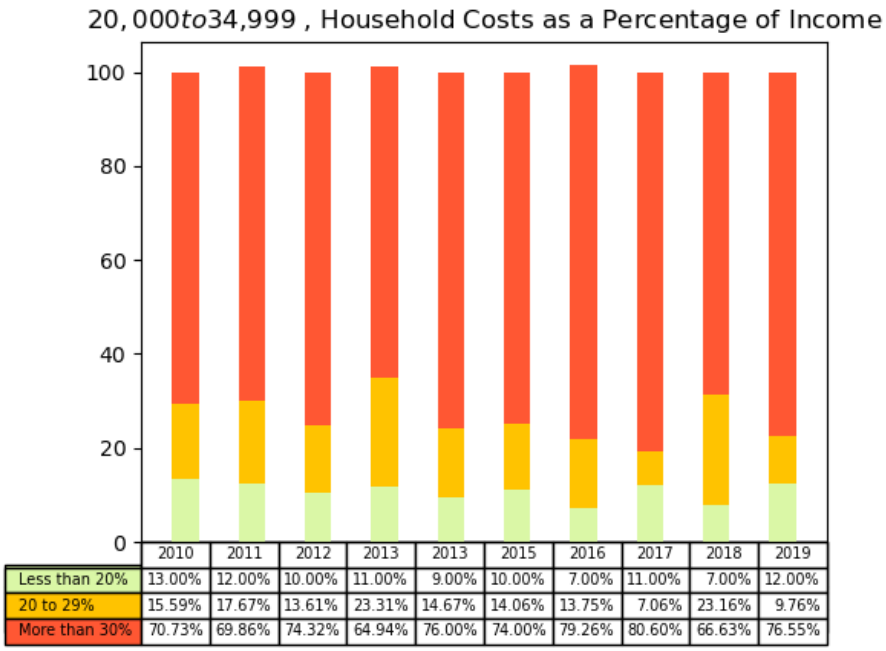

| | 2010 | 2011 | 2012 | 2013 | 2013 | 2015 | 2016 | 2017 | 2018 | 2019 |
|---|---|---|---|---|---|---|---|---|---|---|
| Less than 20% | 53.00% | 52.00% | 57.00% | 56.00% | 64.00% | 57.00% | 60.00% | 64.00% | 61.00% | 68.00% |
| 20 to 29% | 29.57% | 32.42% | 26.38% | 26.90% | 23.78% | 26.97% | 24.29% | 24.35% | 24.14% | 18.52% |
| More than 30% | 16.42% | 14.43% | 15.73% | 16.10% | 10.50% | 15.64% | 13.64% | 10.46% | 12.73% | 12.50% |

Figure 5.




| | 2010 | 2011 | 2012 | 2013 | 2013 | 2015 | 2016 | 2017 | 2018 | 2019 |
|---|---|---|---|---|---|---|---|---|---|---|
| Less than 20% | 27.00% | 22.00% | 25.00% | 20.00% | 29.00% | 23.00% | 27.00% | 29.00% | 22.00% | 32.00% |
| 20 to 29% | 19.56% | 25.84% | 20.65% | 28.63% | 19.20% | 22.22% | 18.41% | 17.31% | 18.57% | 16.72% |
| More than 30% | 52.44% | 50.14% | 53.52% | 50.62% | 50.99% | 53.85% | 52.79% | 51.34% | 58.88% | 49.39% |

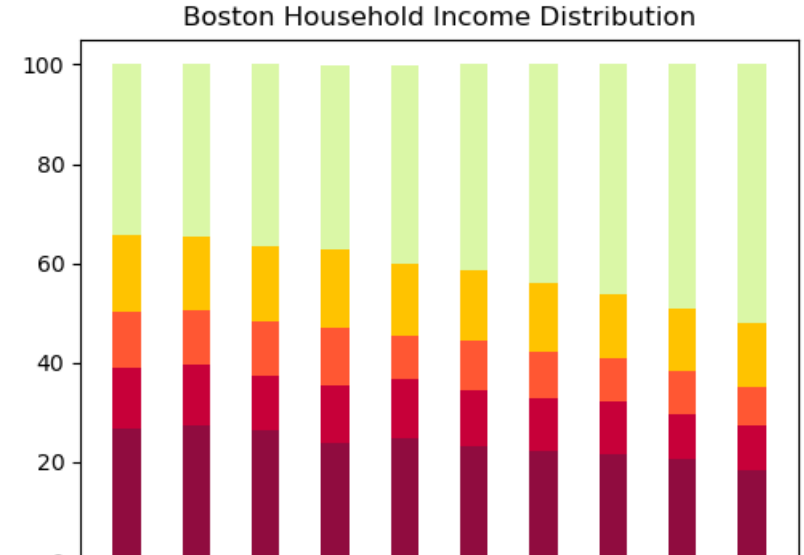
Figure 6.




| | 2010 | 2011 | 2012 | 2013 | 2013 | 2015 | 2016 | 2017 | 2018 | 2019 |
|---|---|---|---|---|---|---|---|---|---|---|
| Less than 20% | 13.00% | 12.00% | 10.00% | 11.00% | 9.00% | 10.00% | 7.00% | 11.00% | 7.00% | 12.00% |
| 20 to 29% | 15.59% | 17.67% | 13.61% | 23.31% | 14.67% | 14.06% | 13.75% | 7.06% | 23.16% | 9.76% |
| More than 30% | 70.73% | 69.86% | 74.32% | 64.94% | 76.00% | 74.00% | 79.26% | 80.60% | 66.63% | 76.55% |

Figure 7.

As can be seen from figure 8, household income distribution has been skewing higher over the years. Our theory is that this is due to lower-income households being priced out and moving out, while higher income households move into Boston. Although inflation should be considered before completely backing this theory.




| | 2010 | 2011 | 2012 | 2013 | 2013 | 2015 | 2016 | 2017 | 2018 | 2019 |
|---|---|---|---|---|---|---|---|---|---|---|
| <20K | 26.00% | 27.00% | 26.00% | 23.00% | 24.00% | 23.00% | 22.00% | 21.00% | 20.00% | 18.00% |
| 20K-34K | 12.30% | 11.60% | 10.70% | 11.30% | 11.30% | 10.90% | 10.70% | 10.56% | 8.45% | 8.58% |
| 35K-49K | 10.90% | 10.50% | 10.50% | 10.70% | 8.40% | 9.60% | 9.00% | 7.93% | 8.31% | 7.51% |
| 50K-74K | 14.90% | 14.50% | 14.60% | 15.10% | 14.70% | 13.60% | 13.90% | 12.03% | 12.64% | 13.00% |
| >=75K | 34.40% | 34.60% | 36.60% | 36.20% | 39.00% | 41.50% | 44.00% | 46.13% | 48.97% | 51.96% |

Figure 8.

We also considered the features of the mortgage in Boston. Figure 9 and 10 show the number of housing units based on mortgage status. On average, there are 65,713 housing units with mortgages and 22,955 without mortgages. There is a 10% increase in the number of housing units with mortgages from 2010 to 2019. This observation also aligns with the housing value trend in Boston shown in figure 3. The median housing prices increase over the years which might be one of the reasons that more households choose to have mortgages.

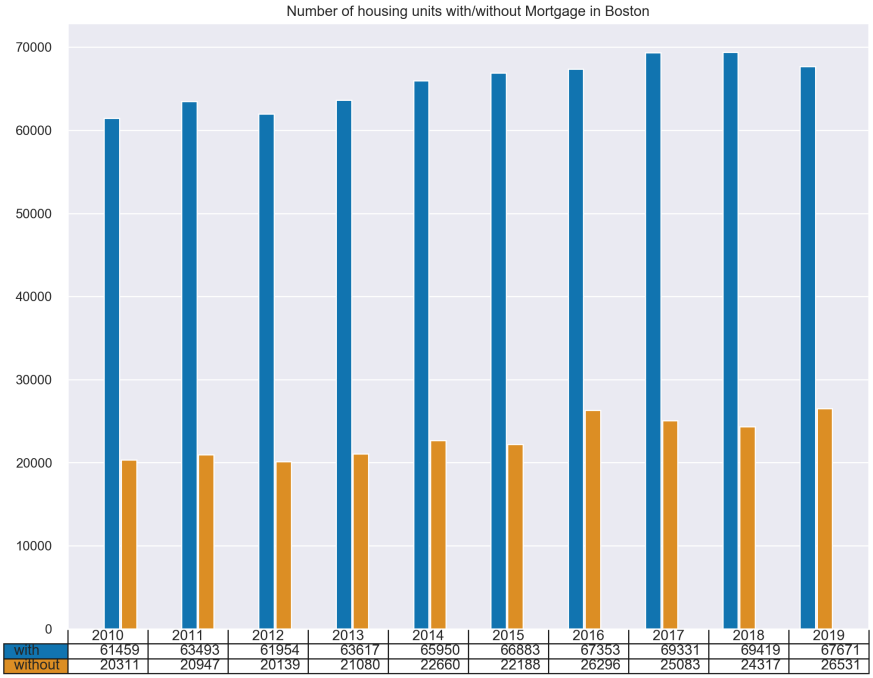



Figure 9.

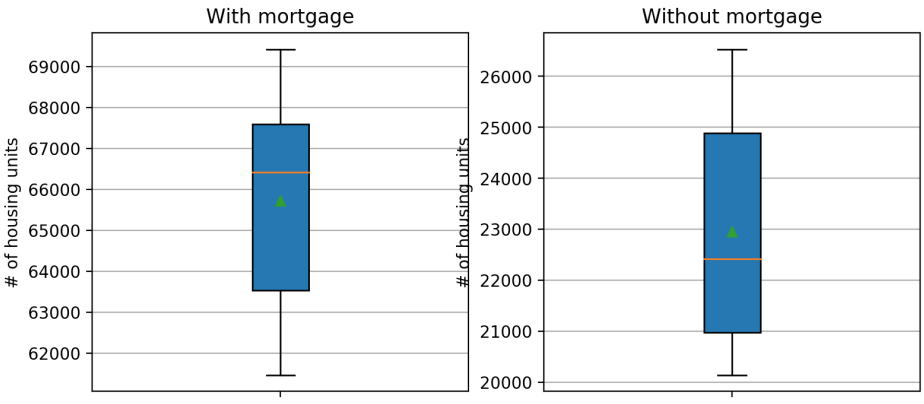



Figure 10.

Our team is also interested in that among those occupied units, how many percent are occupied by the owners and how many percent are occupied by the renters. Figure 11 is the plot for the

proportion of housing tenure between owner-occupied and renter-occupied through the years. In Boston most of the occupied housing units are occupied by the renters. In Figure 12, it shows that there are no big changes in the ratio in the past ten years. The largest difference is less than 3%. On average, only approximately 34.08% is occupied by the owners and 65.92% is occupied by the renters. However, there is a slightly increasing trend for occupancy by owners.

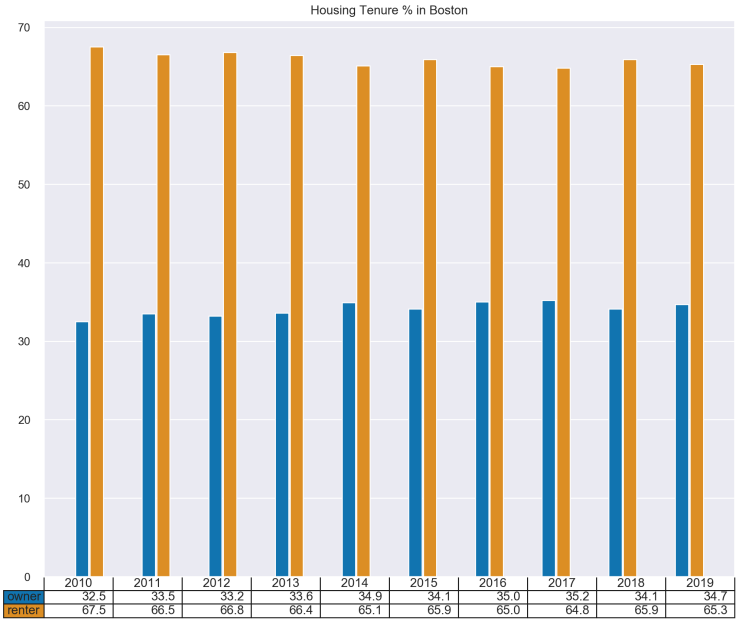



Figure 10.

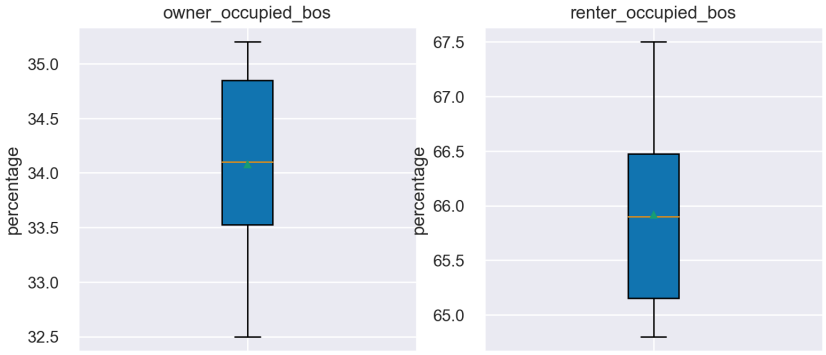



Figure 11.

# Comparisons

Besides analysing the housing situations in Boston and MA, our team also did comparisons with similar cities as Boston for some important features. The section will elaborate in detail.

*Similar cities*

We wanted to compare our trend analysis in Boston with other similar cities to see if our local trends still apply. As a first step to selecting cities similar to Boston, we used K-means clustering with the population density as the feature to cluster on. Figure 12 shows the distribution of the clustering result. Boston is within the red cluster. The red (Boston) cluster also contains two of our selected cities (Seattle and Philadelphia), indicating these two cities as being the most similar to Boston within our selected criteria. Portland is within the green cluster while San Francisco is within the blue cluster. We can see that both of these two cluster groups are close to the red target cluster (Boston).

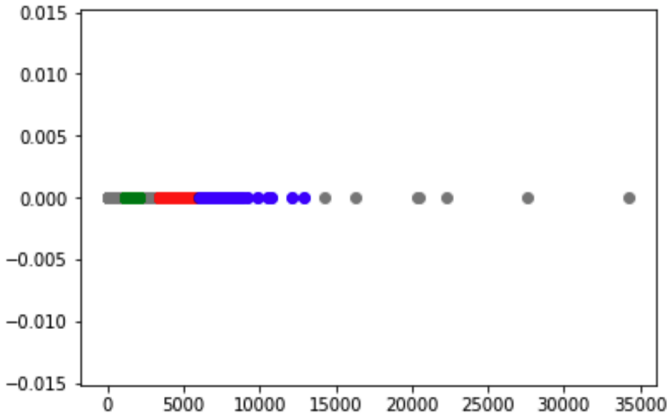



Figure 12.

Because relevant data was not all available to us, we chose to continue with our subjective selection to close the gap between the cities not within Boston's cluster. The rest of the clusters are all colored gray in Figure 12. Eventually, we defined the similar cities as Seattle, Portland, San Francisco, Philadelphia. The criteria for similarity being general population density, average income and some features shown in table 13. These four cities align with our findings using the clustering method.

| City | GDP per capita 2017 | Population 2017 | Population 2020 | Cost of living index |
|---|---|---|---|---|
| **Boston** | 78,465 | 687,788 | 675,647 | 87.05 |
| **Seattle** | 80,833 | 728,661 | 737,015 | 88.04 |
| **San Francisco** | 89,978 | 878,040 | 873,965 | 97.84 |
| **Philadelphia** | 63,519 | 1,580,863 | 1,603,797 | 83.46 |
| **Portland, OR** | 63,817 | 647,924 | 652,503 | 86.19 |

Table 13.

*Occupancy Status*

Figure 14 shows the number of occupied units over the years and Figure 15 shows the year to year rate of change in the number of occupied units. From Figure 14 we can see that, unlike Seattle which shows a noticeable stable upward trend towards the number of occupied houses over the years, Boston growth is less apparent and stays relatively flat. This is expected across all five cities. If we take a look at Figure 15, we can see that for all five cities, there is not a significant acceleration in the growth rate of the housing occupancy, among which, Boston's rate is the most stable.

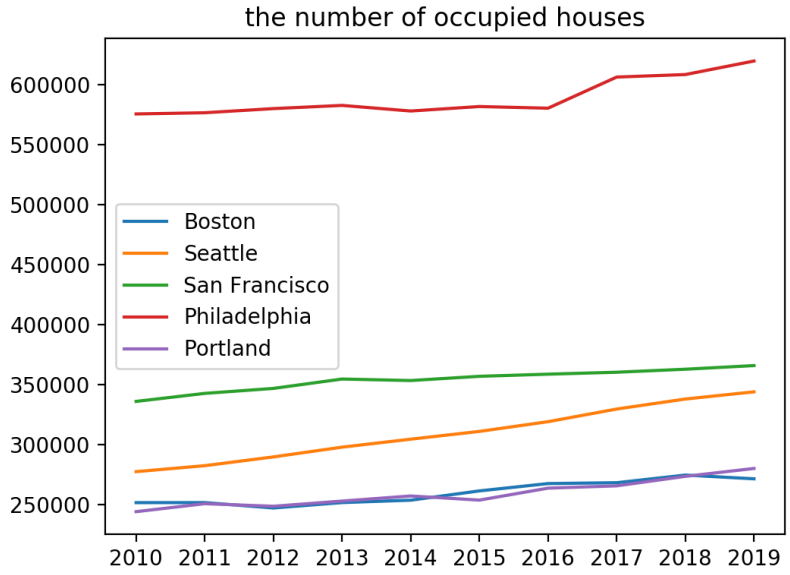



Figure 14.

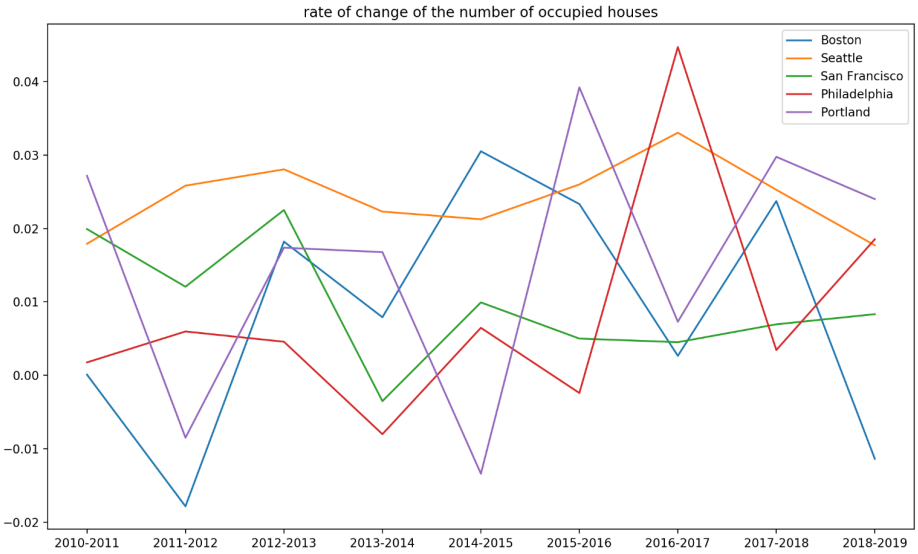



Figure 15.

Figure 16 and 17 provides the same analysis but for the number of vacant units. We can see for most cities including Boston, there is a decrease in the number of vacant units from 2010-2013. From 2013-2019, there is another noticeable uptick for all selected cities except for Philadelphia.

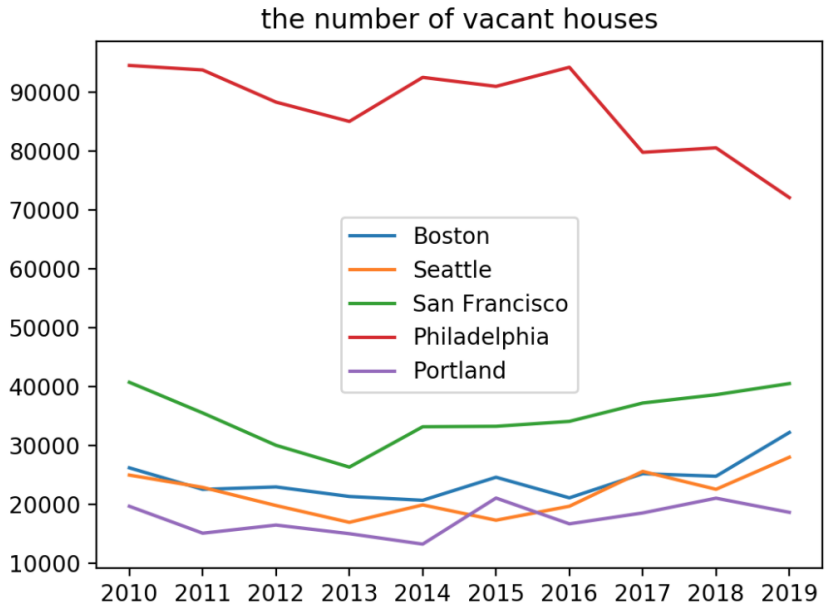



Figure 16.

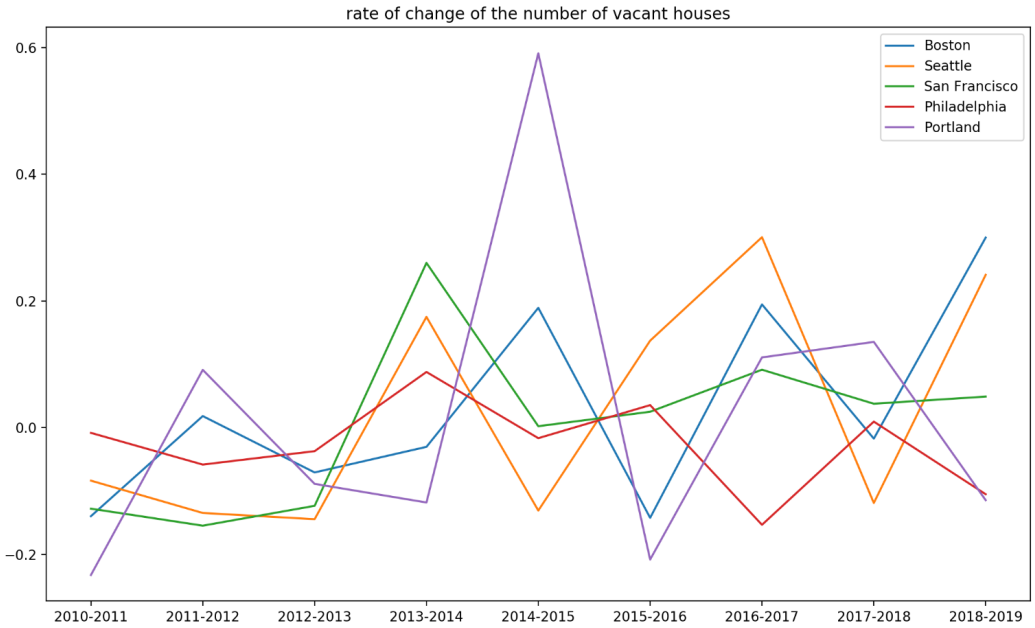



Figure 17.

*Housing prices*

With the analysis on the number of vacant "zombie" houses complete, we wanted to further examine the potential factors for the increase over the recent years. So we analyzed data that could potentially affect this phenomenon. One potential factor we identified was house prices. In Figure 18, we show the number of housing units in various price ranges for our selected cities. We can observe that the majority of Boston house prices cluster around $500,000. Portland and Seattle also feature similar distributions within that price range. Meanwhile Philadelphia features a much wider range of prices that extends much lower than any other city. San Francisco prices swing the other direction, with most of the prices skewing to the higher side.

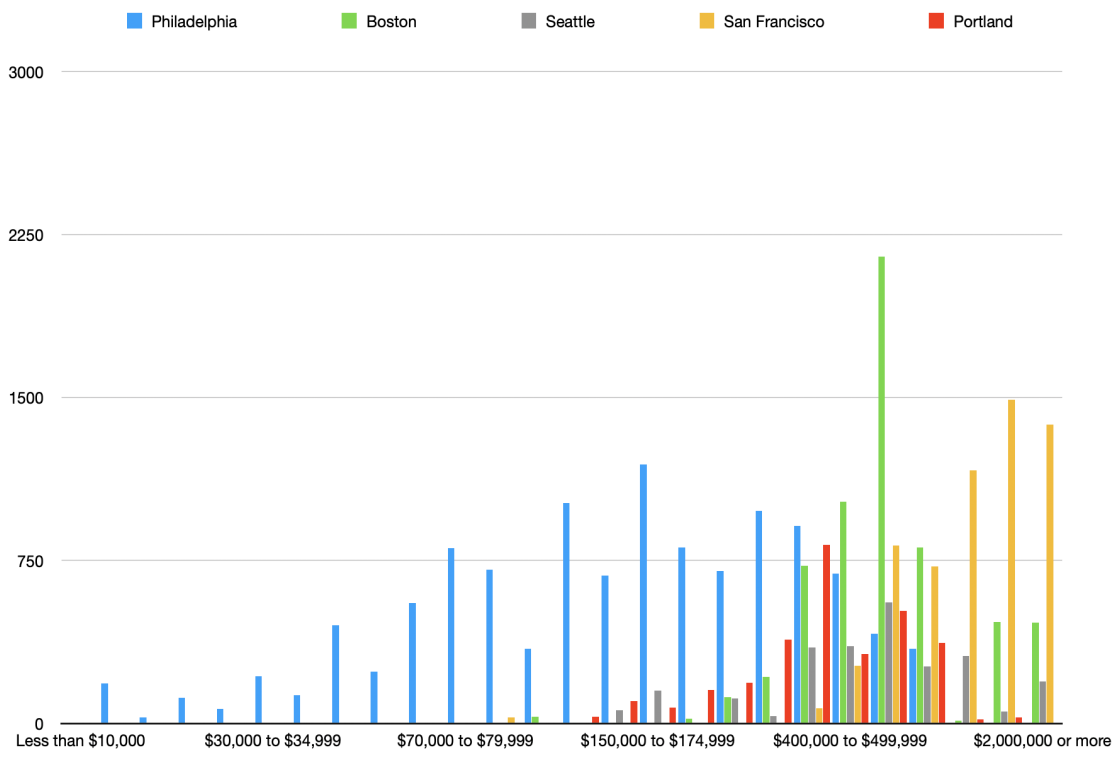



Figure 18.

*Income*

And this also can be explained by the mean income of these 5 cities. San Francisco has the highest median wage, hence it tends to have higher housing prices while Philadelphia has the lowest median income and hence a more left-skewed price distribution for their on-sale households.

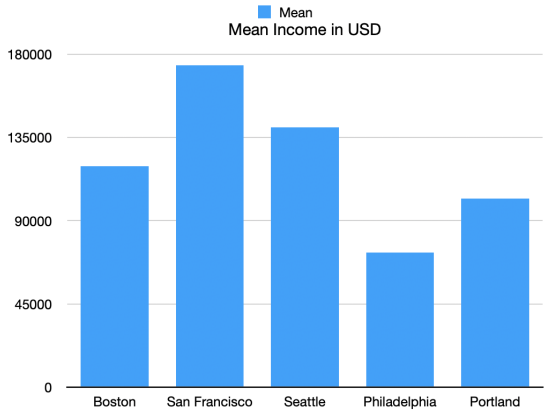
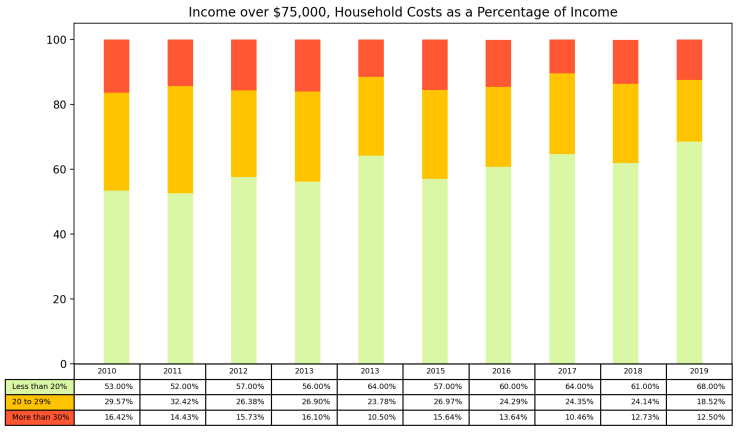



Figure 19.

Additional data analysis needed to be conducted in order to inform the previous finding. House prices needed to be put in context with income. So we visualized housing costs as a percentage of income with these other cities as well. Figure 20 shows the visualization for households with a total income of $75,000 and above. The percentages are color coded with higher percentages spent in the red, and lower percentages in the light green. As we can see Boston (the first chart) and San Francisco have proportionally more households (larger orange and reds) spending more of their income on household costs in comparison to a city like Philadelphia (small orange and reds).




| | 2010 | 2011 | 2012 | 2013 | 2013 | 2015 | 2016 | 2017 | 2018 | 2019 |
|---|---|---|---|---|---|---|---|---|---|---|
| Less than 20% | 53.00% | 52.00% | 57.00% | 56.00% | 64.00% | 57.00% | 60.00% | 64.00% | 61.00% | 68.00% |
| 20 to 29% | 29.57% | 32.42% | 26.38% | 26.90% | 23.78% | 26.97% | 24.29% | 24.35% | 24.14% | 18.52% |
| More than 30% | 16.42% | 14.43% | 15.73% | 16.10% | 10.50% | 15.64% | 13.64% | 10.46% | 12.73% | 12.50% |

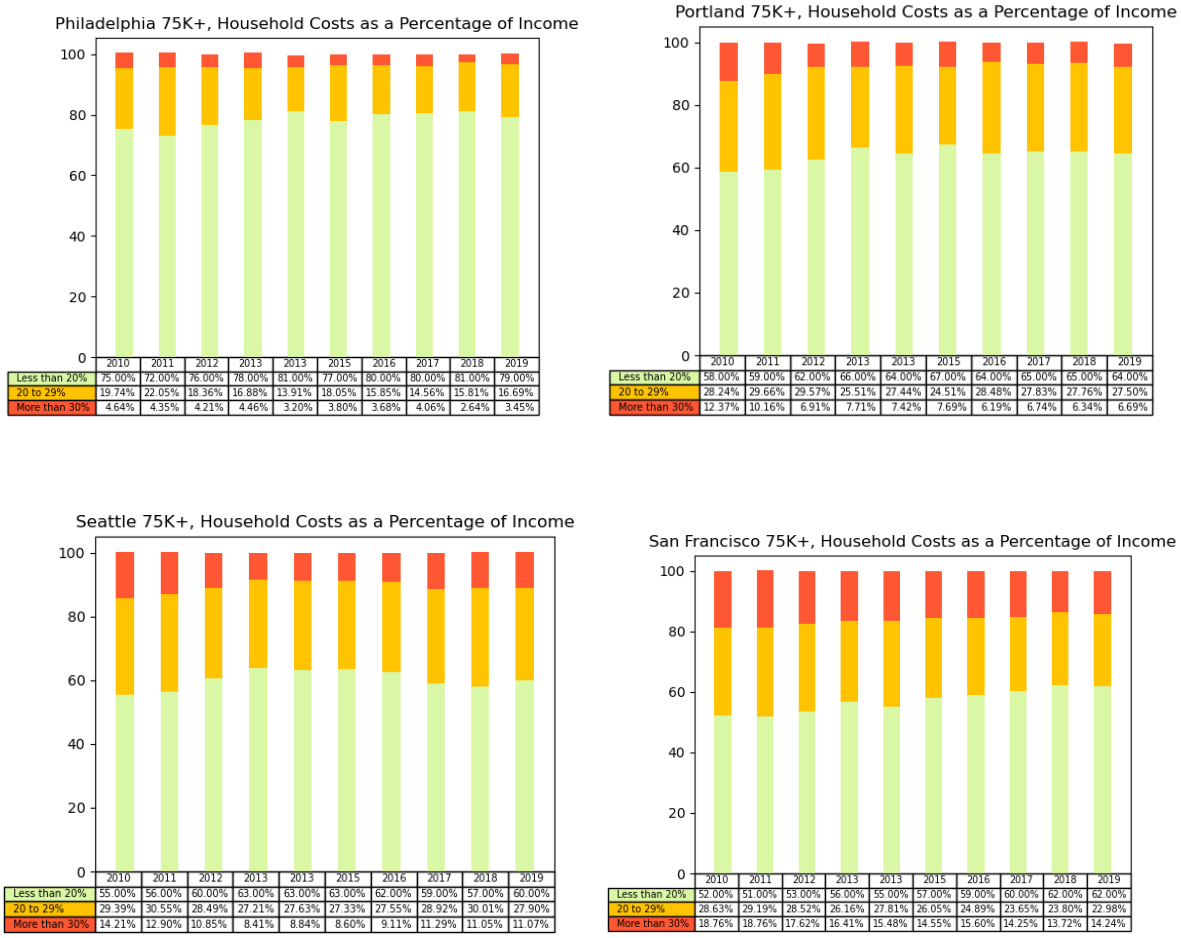



Figure 20.

*Housing tenure*

Figure 21 and figure 22 are the plots for owner-occupied rate and renter-occupied rate of Boston with four other similar cities respectively. Both Boston and San Francisco have much lower owner-occupied rates as compared to Portland, Philadelphia. Seattle is quite evenly splitted as compared to other cities. But it had a decreasing owner-occupied rate in recent years. But in general, there is a slowly increasing trend for the owner-occupied rate in Boston.

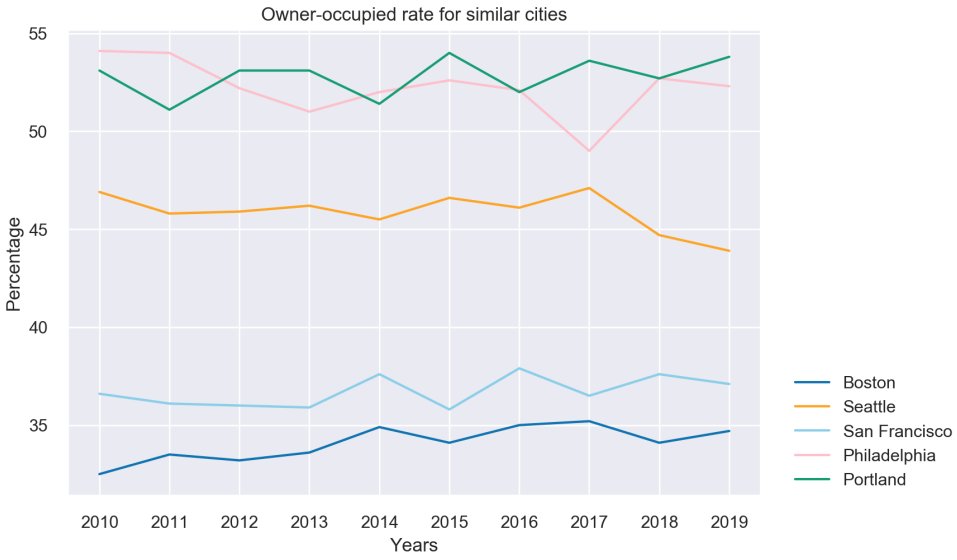



Figure 21.

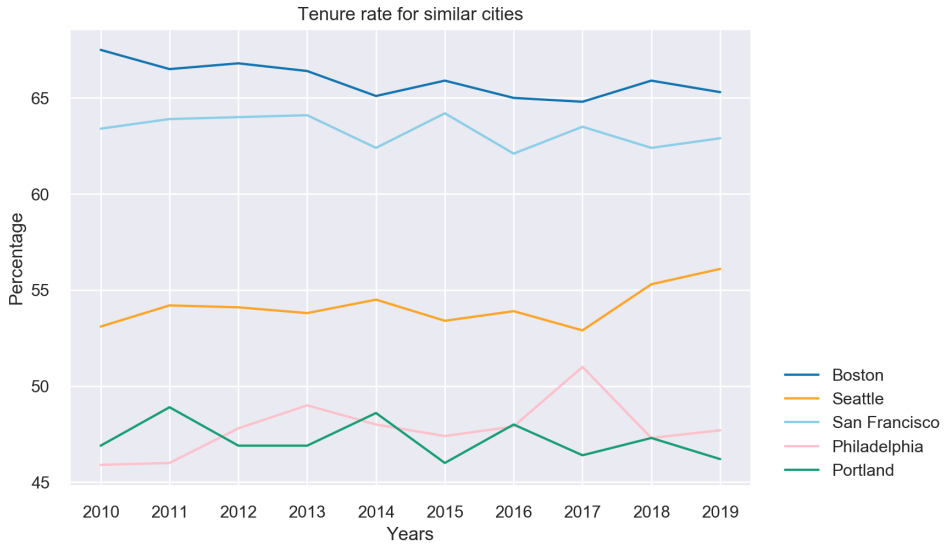



Figure 22.

*Geographic mobility*

Another key question we wanted to answer along the way was whether the population in Boston was increasing or decreasing. Analysis on this subject could potentially inform our broader analysis on the cause of vacancies. However a large limitation of the dataset was that specific information on those moving out of Boston was not available. Only data about those moving into Boston was available. We can observe in Figure 23 that the population of Boston was still increasing from 2010 to 2019. Therefore we can at least rule out that the vacancies were only the result of residents moving out of Boston.

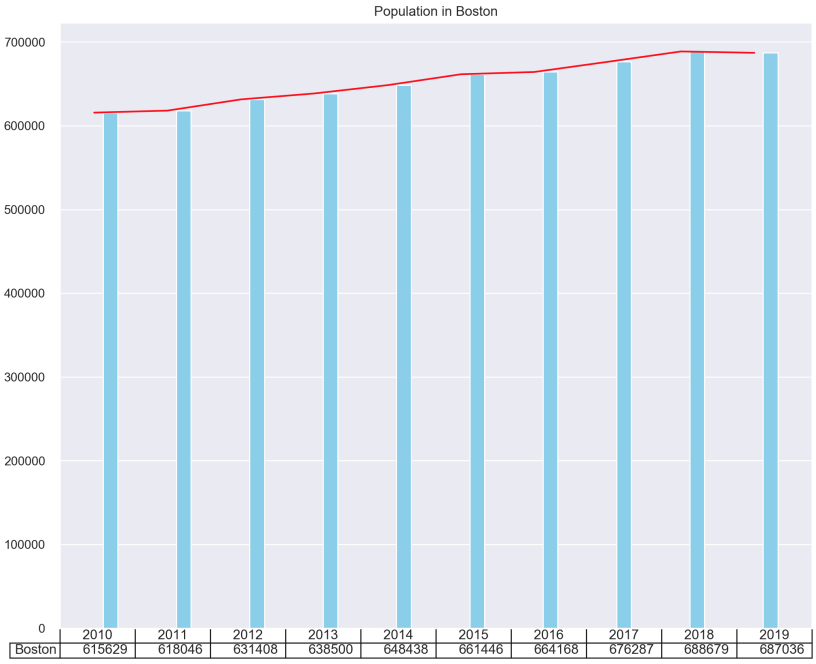



Figure 23.


Figure 24.

Deeper additional population analysis was also performed by separating by age group from 20 to 69. Figure 24 visualizes the population of various age groups in Boston and in Massachusetts as a whole. We can see over all figures that people between the ages of 20 and 29 were generally the most mobile. There is no steady trend year over year, but over the decade, we can see a gradual increase in population matching up with Boston's increase in population.

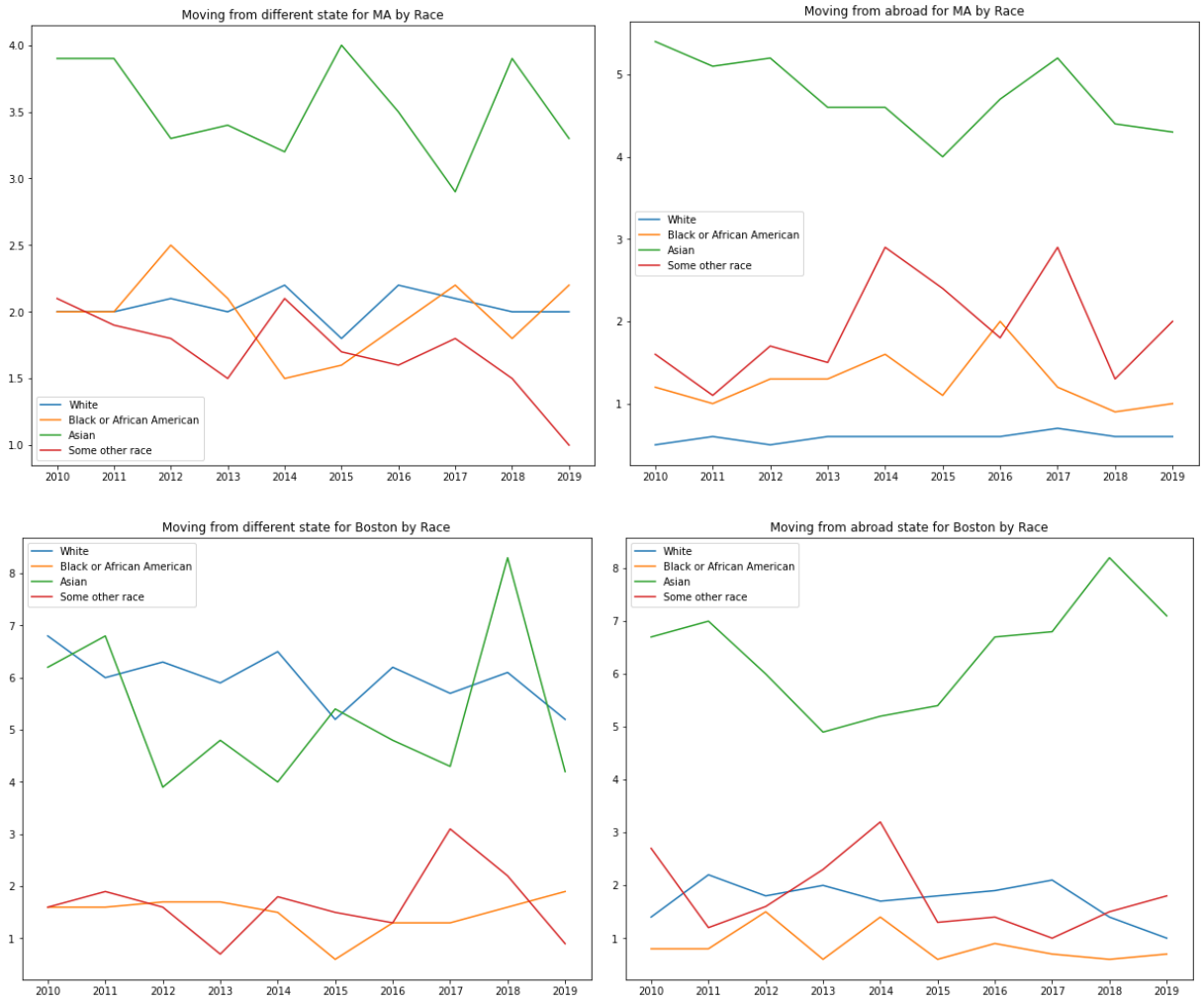



Figure 25.

We also analyze this same trend with racial demographics as shown in figure 25. Since the population of each race varies greatly, we use the percentage of immigrants in the population as a measure. Here we can observe that the Asians were the most populous group moving into Boston both from another state and from abroad. However this trend has seen a slight downward trend over the years.

*Job offerings*

Figure 26 is Boston's total job growth from 2010 to 2017 provided by Boston's economy 2019. We can see that the number of job offerings increases for most of the sectors. Also, based on the latest US Census Data, figure 27 shows that except for the government part, all other fields are providing more jobs in 2021 than the year before. It indicates that more opportunities can be found in Boston which align with our findings that the younger age groups are moving to Boston. That might lead to the increase in occupancy as well as the increase of housing prices.


Figure 26.

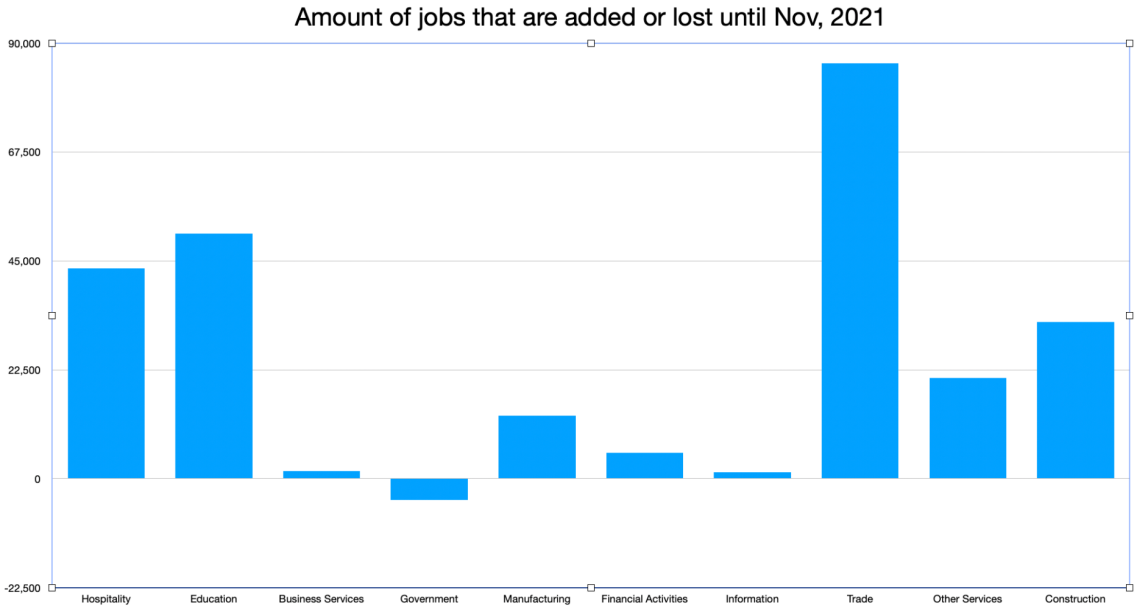



Figure 27.

# Conclusion and possible next steps

*Conclusion*

The number of vacant “zombie” houses is increasing. One potential factor we identified was house prices, which shows an increasing trend for the amount of houses with higher values and a decreasing trend for the amount of houses with lower values. Same behaviours can be observed from those cities that are similar to Boston.

Although housing vacancy has been increasing in Boston in recent years, Boston’s population also increases, which means population does not seem to correlate with increasing vacancies. From geographic analysis, there are a large number of young people (age from 20 to 39) moving from other states or from abroad to Boston and this number gradually increases. At the same time, there is evidence showing that Boston’s job offerings are increasing, which might relate to the geographic mobility of young people.

*Possible Next Steps*

One of the possible next steps is to build a statistical model and analyze the p-values for each feature to see whether it is significant. Also, a correlation matrix can be used to have a better understanding of the pairwise relationships. Lastly, it will be good to consider more attributes that may contribute to the occupancy.

# Challenges and limitations

*Challenges*

Even after researching from different sources and contacting the previous team to share related data sources, we were still not able to find housing data in Seaport, specifically data on housing that is owned by investors but unoccupied, which is the client's main interests in Boston.

*Limitations*

The pulled out data from simplemaps, which is the United States Cities database to perform k-means clustering, contains limited information. We were only able to use these limited features

in the free version. Therefore, we had to also rely on some manual research and selection of similar cities to perform our analysis on. Our manual research is largely informed by our k-means clustering results and there is a clear alliance with the manual selection and the clustering results. Therefore, we are confident in the representativeness of our selected cities. Moreover, the data of 2020 is not released yet. Our team suspects that there will be big changes with trends in different features by including the 2020 data due to the pandemic.

# Appendix

All the datasets and analysis code can be found in separate files. Please refer to the ‘Readme’ document for more information.