Classifying Neighborhoods Impacted by Upzoning in California

CP 257 Final Project

Jared Nolan

I. Introduction

A. Background

California is in the depths of a protracted housing crisis. Many factors have contributed to a housing shortage on the supply side while demand, driven by California's strong economy, has not let up. The shortage has caused housing prices and rents to rise dramatically across the state, but especially in metropolitan areas around San Francisco and Los Angeles.

One fundamental way to address the problem of high housing costs is to simply build more housing. If supply increases enough, then prices will go down. But where and how can this new housing be built? To mitigate the environmental impact of additional housing and avoid bringing more cars on the road, it makes sense to build dense housing in areas that are already served by high-quality transit and infrastructure.

There are many zoning restrictions and regulatory hurdles, however, that prevent multifamily development from happening around transit. To address those obstacles, State Senator Scott Wiener (D-SF) introduced Senate Bill 50 (SB 50) in December 2018, which is designed to override local zoning laws to allow denser development in the immediate vicinity of transit stops. SB 50 has a few main components. The first part defines what constitutes "high-quality transit." The next part of the bill describes the changes that will happen at different proximities to transit. The rest of the bill lists various mitigations that will protect sensitive communities.

But who exactly are the communities that will be subject to this policy? Housing advocates have expressed intense concern that new construction could displace low-income residents living in areas at risk of gentrification around transit. In light of this concern, this research is motivated by two primary questions:

1. What are the different kinds of neighborhoods that would be impacted by this policy?
2. Focusing on 5 case study areas, what kind and amount of housing will be unlocked by these upzoning measures?

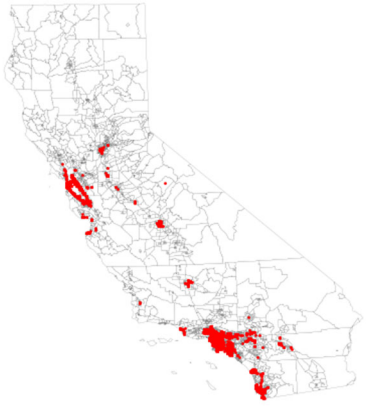
This paper will focus on developing a neighborhood typology that answers the first question and informs the selection of the five case studies.

B. Data

The first step in the process is identifying all of the qualifying transit stations in California. I worked with Simon Hochberg (MCP '19) to create this dataset. Simon downloaded General Transit Feed Specification (GTFS) data from over 100 transit agencies in California and created an algorithm to calculate headways for each of the service periods. This work resulted in 10,550 transit stations that qualified under the SB 50

criteria. Figure 1 provides a sense for the location of these stations across California. The vast majority of stations are concentrated around Los Angeles and the San Francisco Bay Area.

Figure 1: Qualifying Transit Across California



To understand the different kinds of neighborhoods that would be impacted by SB 50, we identified 23 characteristics that are important determinants of the character of the neighborhood as well as measures that indicate residents' vulnerability to displacement. Many of the characteristics were modeled after the methodology of Salon (2015).[1] A list of the characteristics can be found in Table 1. All of these variables were obtained at the census-tract level and came from the ACS/Census data source. They are the 2013-2017 ACS 5-Year Estimates.

Table 1: Characteristics Used when determining Neighborhood Typology

Population Characteristics	Economic Characteristics	Built Form Characteristics
• Share of households that rent • Racial breakdown • Poverty status by race • Households below 200% of poverty line • Households living with children • Median age of population • Population with bachelor's degree	• Median rent compared to the county median rent • Vacancy rate of housing units • Unemployment rate • Number of jobs within commuting distance	• Share of housing units that are detached single-family homes vs. multifamily • Share of housing units that were built before 1950 and since 2000 • Population density

The final dataset consists of 10,550 rail and bus stations with 23 characteristics. To turn this information into a typology of neighborhoods I used a clustering procedure. First I reduced the dimensionality of the dataset by applying Principal Components Analysis (PCA). I used PCA because it can account for a high degree of correlation between the demographic I used the top 10 new dimensions, or components, that PCA identified to run the clustering procedure. To cluster the stations I used the k-means algorithm and specified five clusters.

I modeled the methodology for this research after Ibes (2015).[2] Ibes uses factor analysis and k-means clustering to classify the public park system in Phoenix, Arizona. The dataset in that paper covered 162 public parks with 14 related characteristics. Using PCA, Ibes reduced the number of factors to five components that explained 77.2% of the total variation. Ibes applied k-means to the five factors and identified five different park types. Song & Knaap (2007) follow a similar approach for classifying

[1] Salon, Deborah (2015), "Heterogeneity in the relationship between the built environment and driving: Focus on neighborhood type and travel purpose," *Research in Transportation Economics*, 52 (2015), 23–45.
[2] Ibes, Dorothy (2015), "A multi-dimensional classification and equity analysis of an urban park system: A novel methodology and case study application," *Landscape and Urban Planning*, 137 (2015), 122–137.

neighborhoods in the Portland, Oregon metropolitan area.[3] Their dataset consisted of 6,788 parcels with 21 urban form characteristics. Using factor analysis they reduced the dataset to eight factors that explained 82% of the total variation. Song and Knaap applied k-means cluster analysis to the reduced dataset and found six different clusters.

C. Summary

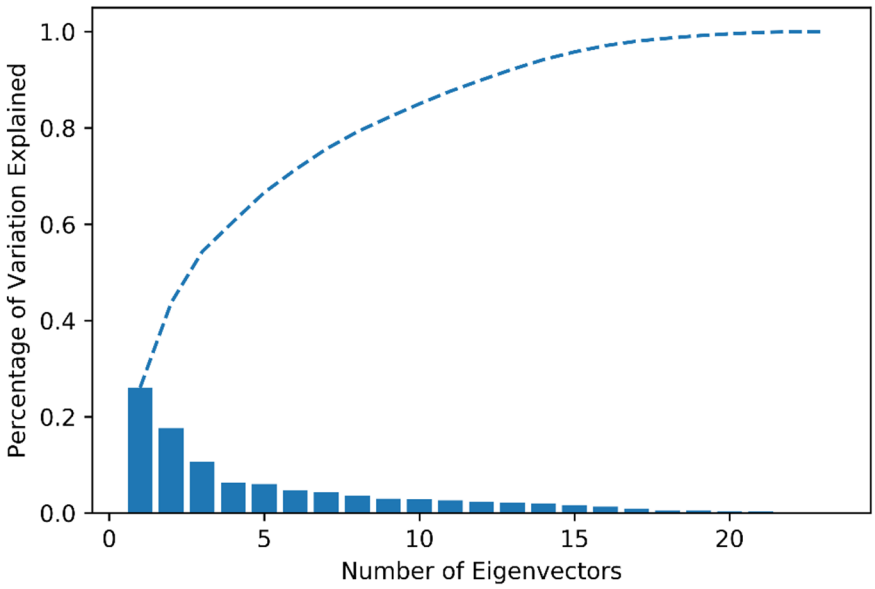
Section II discusses the methods and key results and presents visualizations of the results. Section III draws conclusions and proposes next steps. Section IV lists the references for this paper. This research is part of a project sponsored by the Terner Center for Housing Innovation and the Center for Community Innovation and led by UC Berkeley professors Carolina Reid and Karen Chapple.

II. Methods and Results

A. Principal Component Analysis

The first step of the analysis is to apply Principal Components Analysis (PCA) to the dataset. PCA is necessary in this case because many of the neighborhood characteristics in the dataset are highly correlated. For example, poverty and race are often positively correlated. Similarly, income is positively correlated with education. As a result, it is more efficient to collapse these correlated variables into a smaller set of factors that removes redundancy and accounts for the correlation. PCA accomplishes this exact task by determining new factors that explain the most variation in the data. In the case, PCA was applied to the dataset using the scikit-learn Python library. Figure 2 shows what percentage of the variation that each resulting factor explains, and the cumulative amount of variation explained.

Figure 2: Variation Explained by Principal Components



The additional percentage of variance explained decreases rapidly as more factors are added, so only the most important factors should be retained. Ibes selected the top five factors that explained 77.2% of the

[3] Song, Yan and Gerrit-Jan Knaap (2007), "Quantitative classification of neighborhoods: The neighborhoods of new single-family homes in the Portland metropolitan area," *Journal of Urban Design*, 12:1 (2007), 1–24.

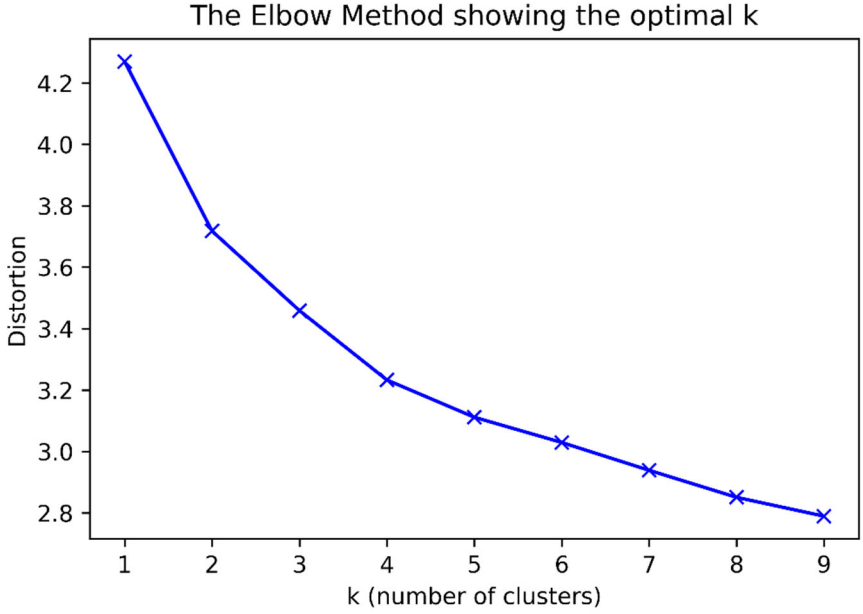
total variation. Song and Knaap selected the top eight factors that explained 82% of the total variation. In this case, I retain the top ten factors that explain 85% of the total variation in the dataset.

B. Clustering Procedure

The next step in the process is to apply a clustering algorithm to the dataset. Both Ibes and Song and Knapp use the k-means clustering algorithm, which "is a simple unsupervised machine learning algorithm that groups a dataset into a user-specified number (k) of clusters."[4] Since the user specifies the number of clusters, it is necessary to determine what the appropriate number of clusters is. There are a few methods for identifying the appropriate number of clusters. One of the most popular approaches is the "elbow method." In the elbow method, k-means is used to calculate clusters with a range of values for k. For each value of k, the method calculates the "distortion," which is essentially the sum of squared errors between all of the points in a cluster and the cluster centroid.[5] A lower value is better and means the observations are more closely clustered around the centroid. To find the best k, look for the "elbow" in the plot, which is where adding one more cluster has the most impact in reducing the distortion.

Figure 3 shows the elbow results for this dataset. Unfortunately, there is no obvious elbow. There are some small kinks at k = 2 and k = 4, but in general there is a smooth decrease in distortion as k increases.

Figure 3: Using the Elbow Method to Determine Clusters

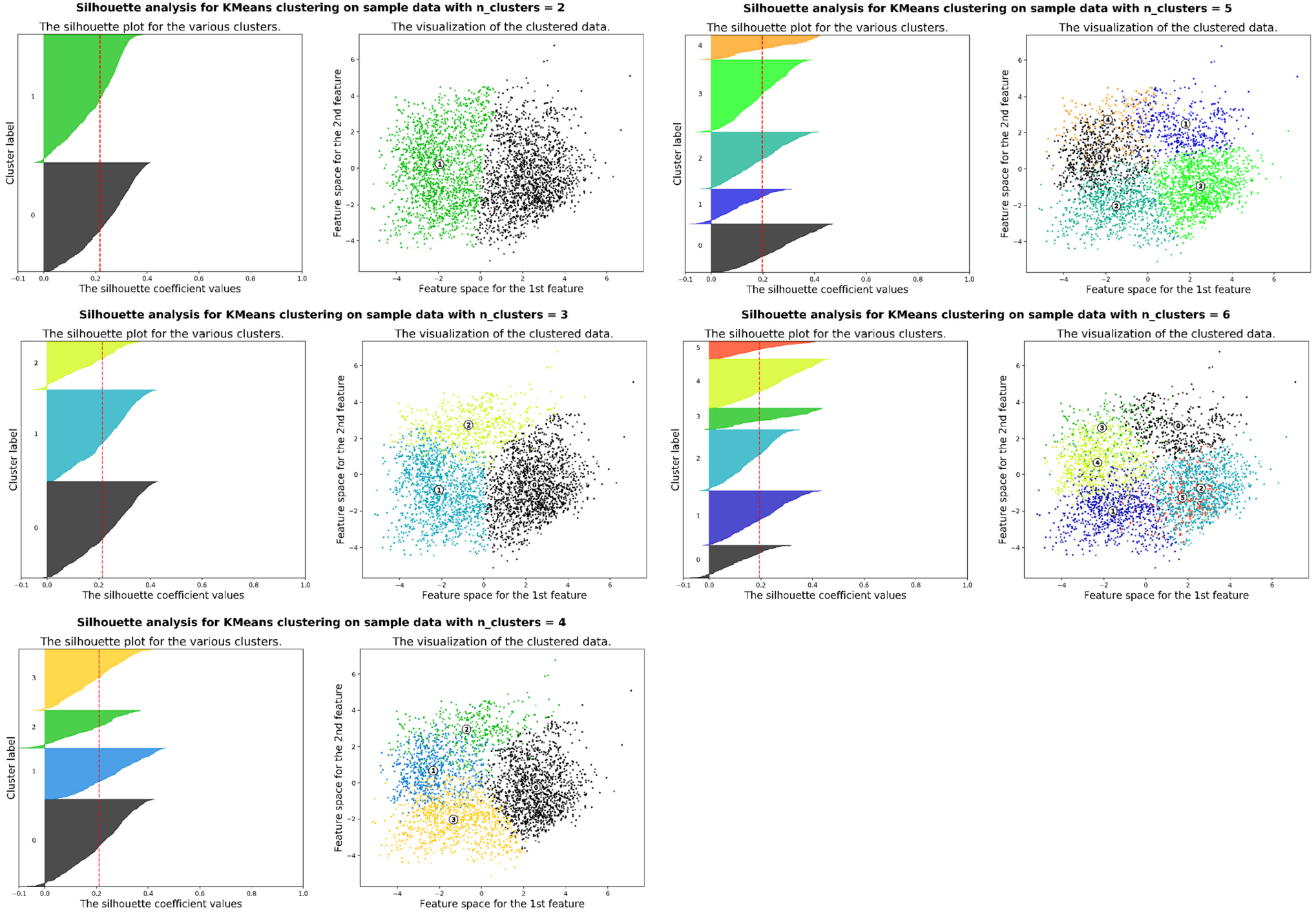


Since the elbow method does not produce a conclusive result, it is necessary to consider another method for identifying the appropriate number of clusters. Figure 4 shows the results for the "Silhouette Method."

Rather than studying the compactness within each cluster like the elbow method, silhouette analysis measures the separation of clusters from each other:

[4] "Using the elbow method to determine the optimal number of clusters for k-means clustering," Robert Gove's Block, https://bl.ocks.org/rpgove/0060ff3b656618e9136b.
[5] "kmeans elbow method," Python, https://pythonprogramminglanguage.com/kmeans-elbow-method/.

Figure 4: Using Silhouette Analysis to Determine Clusters



> Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.[6]

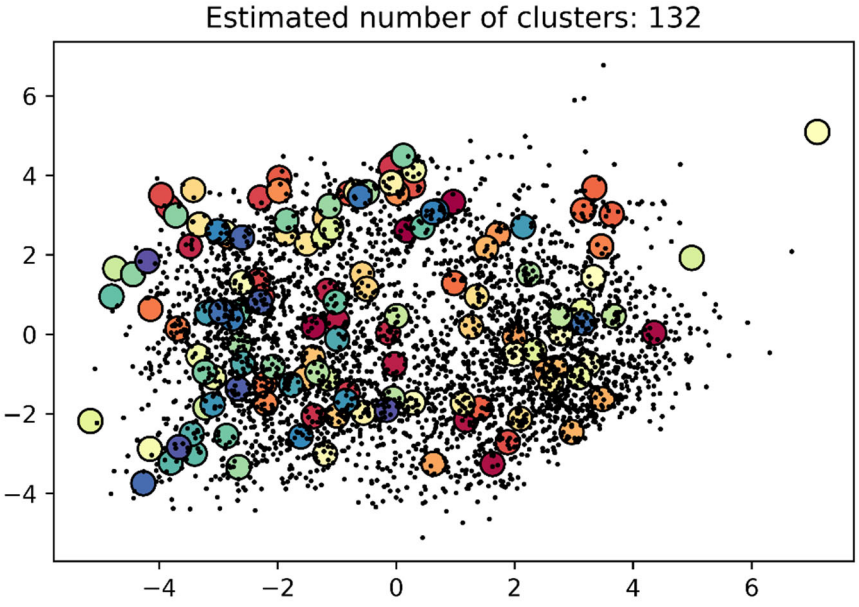
Figure 4 shows that the average silhouette coefficient is pretty consistently around 0.2 for values of k from 2 to 6. The iteration with six clusters produces the worst results out of all of the iterations. There is not much separating the other four clustering results, although it appears that k = 3 and k = 4 might be slightly superior from a visual inspection.

Since there is still not a conclusive result, it is worth trying one more method. Density-based spatial clustering of applications with noise (DBSCAN) is an alternative clustering method to k-means that does not require the user to specify the number of clusters. Instead, the algorithm groups together points that are located around core points of high density. [7] The algorithm requires two parameters, eps and min_points:

- eps is the minimum distance between two points. A smaller values is preferable since it produces denser clusters, but it means that the algorithm is more strict in grouping together points.
- min_points is the minimum number of points necessary to form a dense region. This parameter is generally derived from the number of dimensions in the dataset.

Figure 5 shows the results from the DBSCAN procedure when using eps = 0.3 and min_points = 10.

Figure 5: Using DBSCAN to Identify Clusters



[6] "Selecting the number of clusters with silhouette analysis on KMeans clustering," scikit-learn, https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html#sphx-glr-auto-examples-cluster-plot-kmeans-silhouette-analysis-py.

[7] "Demo of DBSCAN clustering algorithm," scikit-learn, https://scikit-learn.org/stable/auto_examples/cluster/plot_dbscan.html#sphx-glr-auto-examples-cluster-plot-dbscan-py.

The algorithm identifies 132 clusters, which is too many to be intelligible. Increasing both parameters lowers the threshold for combining points into clusters, but does not significantly reduce the number clusters that the algorithm produces.

None of the three approaches produces an emphatically conclusive result for the number of clusters to choose. Because of this finding, I will use the k-means result with k = 5. I chose this result for two reasons. The first reason is that the results with five clusters make intuitive sense and match my prior understanding of neighborhoods in California (see the next section for more details). The other reason is that the next phase of the analysis will use five case studies, so one case study can come from each cluster.

C. Cluster Results

The next step in the analysis is analyzing the cluster results to define the neighborhood typology. After running the clustering algorithm and choosing the appropriate set of results, I returned to the original 23 characteristics and calculated averages for each of the clusters. Table 2 presents these averages.

Table 2: Average Characteristics for Clusters

Cluster	0	1	2	3	4
number of stops	963	3,305	1,557	2,186	2,539
avg population	9,231	10,699	12,104	11,692	9,280
percent renters	74.7%	69.6%	92.0%	71.1%	40.1%
percent NH white	46.0%	7.7%	20.7%	57.0%	32.9%
percent hispanic	16.8%	66.8%	41.0%	14.8%	27.6%
percent black	7.6%	15.7%	9.7%	5.1%	7.1%
percent asian	25.4%	7.3%	25.2%	17.9%	27.9%
percent below 200% of poverty rate	31.4%	60.4%	61.2%	24.2%	25.8%
percent hispanic in poverty	23.4%	29.7%	38.0%	13.9%	12.7%
percent black in poverty	28.1%	33.5%	44.8%	20.8%	15.0%
percent asian in poverty	18.7%	22.4%	30.4%	12.9%	9.6%
percent white in poverty	14.7%	25.9%	28.5%	9.5%	9.1%
percent single-family detached house	6.2%	41.7%	6.5%	17.6%	57.2%
percent small multifamily (2-4 units)	4.1%	16.2%	8.0%	25.6%	8.9%
percent medium multifamily (5-18 units)	10.2%	18.6%	22.5%	30.9%	8.8%
percent big multifamily (20+ units)	75.7%	12.2%	59.2%	19.1%	10.1%
percent vacant	12.6%	5.9%	9.0%	7.2%	5.1%
percent of units built before 1950	17.9%	40.4%	41.4%	50.5%	33.4%
percent of units built after 2000	36.5%	5.8%	13.1%	4.9%	6.0%
percent with bachelor's degree	60.9%	12.2%	29.6%	62.7%	39.0%
percent of households with children	12.4%	45.9%	20.7%	16.8%	33.1%
unemployment rate	6.4%	11.9%	10.8%	6.1%	7.4%
density (population/square mile)	11,639	15,634	26,631	21,620	11,142
median tract rate / median county rent	1.32	0.81	0.76	1.14	1.12
jobs within commuting distance	1,092,714	1,187,058	1,465,269	1,093,013	790,501

Patterns emerge when comparing these averages across clusters. I defined clusters by visually inspecting this table and noting when a cluster had an average that was well above or below the other clusters. Table 3 presents the results of this process. The colors in the column headers match the colors of the circles in the visualization section.

Table 3: Cluster Descriptions

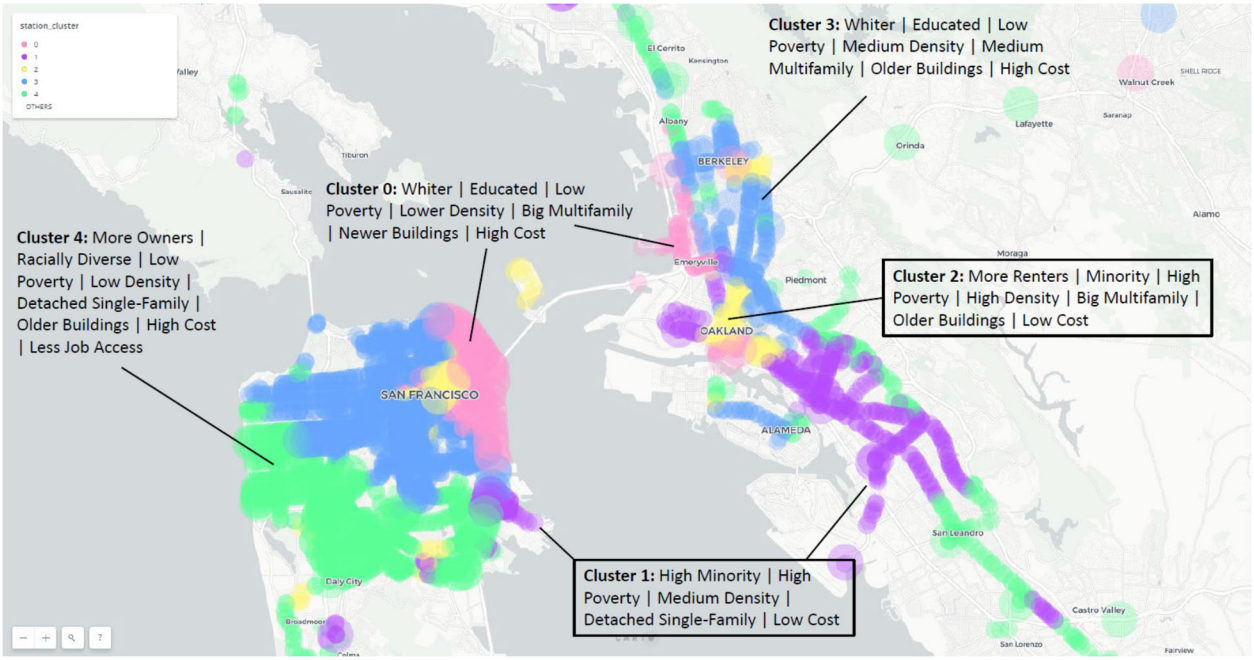
Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Whiter	High Minority	High Minority	Whiter	Racially Diverse
Low Poverty	High Poverty	High Poverty	Low Poverty	Low Poverty
Lower Density	Medium Density	High Density	Medium Density	Low Density
Big Multifamily	Single-Family	Big Multifamily	Medium Multifamily	Single-Family
High Cost	Low Cost	Low Cost	High Cost	High Cost
Newer Buildings	Older Buildings	Older Buildings	Older Buildings	Older Buildings
High Education	Low Education		High Education	
		More Renters		More Owners
		More Job Access		Less Job Access

The results show that Clusters 1 and 2 are the neighborhoods that are the most vulnerable to displacement. They are areas of high poverty with many people of color. Residents are also mostly renters living in older buildings and the associated rent is low compared to the overall county. Therefore, the older buildings are candidates for demolition and if new development occurs and raises the rents in the area, many of these residents would not be able to afford to live there and would be at risk of displacement. This process of defining the clusters lacks scientific rigor, so in future work this step will be refined. Section IV discusses potential alternative approaches to this step.

D. Visualization

It is helpful to visualize these clusters on a map to test whether they make sense given knowledge of the area. Figure 6 shows the clusters in the Bay Area. These visualizations can also be viewed online as interactive maps at: https://jarednolan.carto.com/builder/291b26bb-c6f9-4ad4-b3c7-3c1468b868f3/embed.

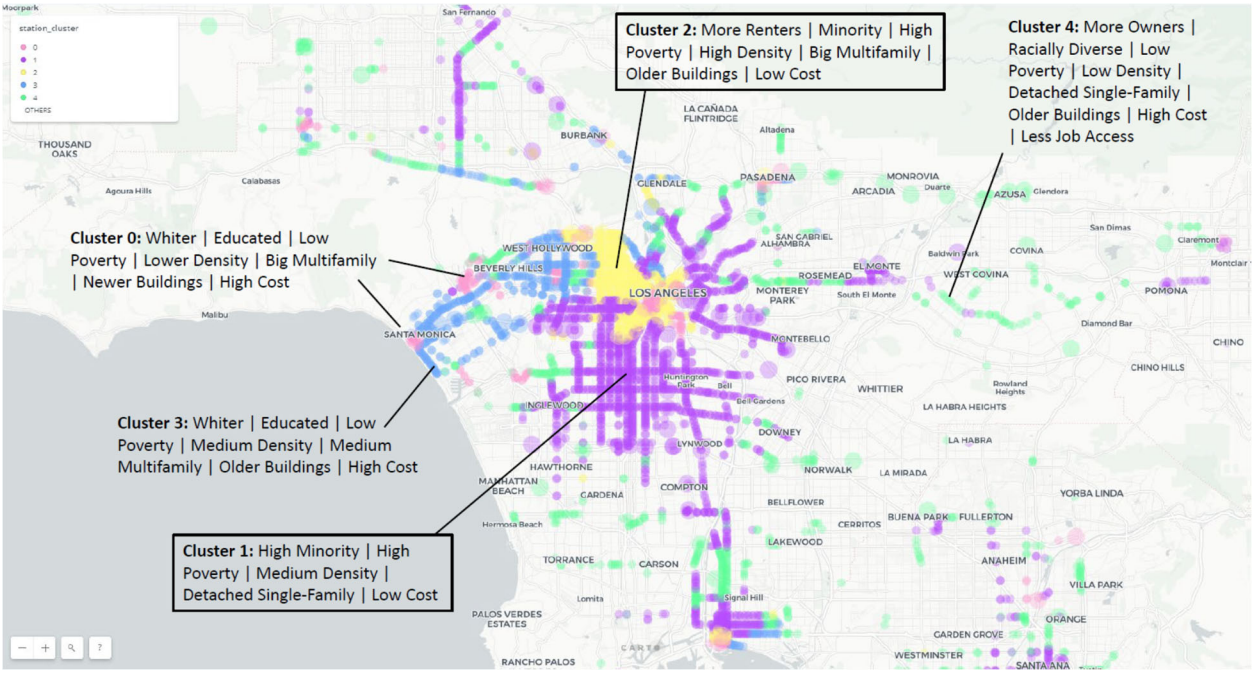
Figure 6: Clusters Visualized in the San Francisco Bay Area



Cluster 1 (purple) appears in the Bayview/Hunter's Point area in San Francisco and in West Oakland. The description for Cluster 1 is high-poverty and non-white with older single-family dwellings. That description matches my knowledge of those areas. Similarly, Cluster 2 (yellow) can be seen in the Tenderloin in San Francisco and downtown Oakland. This finding makes sense because Cluster 2 represents area of high poverty, people of color, and older, bigger multifamily buildings, which is an apt description for both of those areas as well.

Figure 7 shows the clusters in the Los Angeles metro area. The circles are smaller since the map is zoomed further out (the circles correspond to the ½-mile radius around rail stations and the ¼-mile radius around bus stations). An interesting finding is that there are many more neighborhoods in Cluster 1 (purple) and Cluster 2 (yellow) in the Los Angeles area than there were in the Bay Area, suggesting that there is a greater risk of displacement in Los Angeles as compared to the San Francisco Bay Area.

Figure 7: Clusters Visualized in the Los Angeles Metro Area



III. Conclusions and Future Work

Applying PCA and the k-means clustering algorithm to this dataset produced five sensible and informative clusters that will guide the selection of case studies for the next phase of the research. There are a few areas of potential refinement, however. Both Ibes and Song and Knaap use a more sophisticated method for analyzing the cluster results. In both papers they compare the loadings on each of the factors to label the factors with informative names, and then analyze the centroids of each cluster. Ibes also calculates Pearson's correlations between the clusters and the original characteristics to identify which variables are the most correlated with each cluster. Using these methods will allow me to more precisely identify the defining features of each of the clusters.

IV. References

"Demo of DBSCAN clustering algorithm," scikit-learn, https://scikit-learn.org/stable/auto_examples/cluster/plot_dbscan.html#sphx-glr-auto-examples-cluster-plot-dbscan-py.

Ibes, Dorothy (2015), "A multi-dimensional classification and equity analysis of an urban park system: A novel methodology and case study application," Landscape and Urban Planning, 137 (2015), 122–137.

"kmeans elbow method," Python, https://pythonprogramminglanguage.com/kmeans-elbow-method/.

Salon, Deborah (2015), "Heterogeneity in the relationship between the built environment and driving: Focus on neighborhood type and travel purpose," Research in Transportation Economics, 52 (2015), 23–45.

"Selecting the number of clusters with silhouette analysis on KMeans clustering," scikit-learn, https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html#sphx-glr-auto-examples-cluster-plot-kmeans-silhouette-analysis-py.

Song, Yan and Gerrit-Jan Knaap (2007), "Quantitative classification of neighborhoods: The neighborhoods of new single-family homes in the Portland metropolitan area," Journal of Urban Design, 12:1 (2007), 1–24.

"Using the elbow method to determine the optimal number of clusters for k-means clustering," Robert Gove's Block, https://bl.ocks.org/rpgove/0060ff3b656618e9136b.