# An Exploratory Statistical Analysis of Gentrification and Neighborhood Change in Seattle

Sarah Rothschild

Whitman College
Mathematics Department
May 2019

# Abstract

This study examines changing patterns of urban characteristics in Seattle, Washington from 2000 to 2017, focusing on patterns of urban displacement and segregation through statistical techniques and exploratory data analysis. Utilizing Census data and statistical software, we will analyze how Seattle neighborhoods have changed throughout the past few years with regard to socioeconomic status, race, and class. This project further employs data visualization and geospatial analysis, seeking to draw conclusions about the nature of gentrification in the city over recent decades.



# Acknowledgements

I would like to thank Professor Ptukhina for her support and guidance throughout the execution of this project. I greatly appreciated her encouragement and engagement with my work from start to finish. The helpful comments and suggestions from my peer, Yurixy Lopez Martinez, also contributed to the completion of this study. I would also like to acknowledge my cousin, Brook Frye, for offering her guidance during the initial stages of this project. Finally, I would like to thank my fellow peers in the Senior Project class. Your energy, thoughtfulness, support, and humor made for an enjoyable and memorable semester.



Sarah Rothschild
Whitman College
May 2019

# Contents

# 1 Introduction

Located in the heart of the Pacific Northwest, Seattle, Washington is known for its booming high-tech sector, sustainable efforts, economic vitality, and abundant recreational activities. These qualities have made Seattle a premier travel and tourist destination, as well as an attractive place to live for prospective residents. As a result, the city has seen an influx of new inhabitants in the late 20th century, with a 14.1% increase in population from 1980 to 2000 [16]. In fact, Seattle experienced record population growth in recent years, as the fastest growing big city in the nation for the second time this decade, with nearly 21,000 residents moving to the city between 2015 and 2016 [12]. With this boom, the city has encountered an increasingly well-educated, highly-skilled, and professionally employed population. These new residents brought with them heightened social, cultural, and economic capital that would play an important role in changing the neighborhoods' characteristics in the years to come [16]. Like many other cities across the country, Seattle witnessed an upward trend in the socioeconomic status of its residents within neighborhoods, which sparked a rise in housing and rent prices and the subsequent displacement of lower-income communities that could no longer afford the newly instituted living costs.

The term *gentrification* was coined nearly a half-century ago to describe this phenomenon and is regarded as a "tool, goal, outcome, or unintended consequence of revitalization processes in declining urban neighborhoods, which are defined by their physical deterioration, concentrations of poverty, and racial segregation of people of color" [17]. Thus, gentrification concerns the forced migration of communities due to expensive housing prices and their replacement by wealthier inhabitants. As a process, gentrification disproportionately affects communities of color, which results in limited diversity and racial segregation throughout a city's neighborhoods. Since gentrification has become an increasingly more relevant issue in cities around the world, there is a great need for a method of measuring and evaluating this phenomenon in order to understand the underlying processes behind the topic. This problem raises the question: how might we develop a technique for analyzing social issues such as gentrification to better comprehend the nature of these processes?

## 1.1 Related Research

We will begin by reviewing other similar studies that have been conducted on this topic. Considering that gentrification and rising housing prices are becoming increasingly relevant and widespread, there have been numerous studies done on this subject in Seattle and other cities around the world. However, scholars have pointed out that while there is a great breadth of existing qualitative gentrification literature, there is little research done analyzing the subject through purely quantitative measures regarding questions about how

gentrification has affected cities over time. This may result from skepticism surrounding the usage of freely accessible data, such as Census data. Many critics maintain that Census data is not capable of expressing important factors associated with gentrification. Scholars have noted that Census measures often mislabel middle and upper-class neighborhoods as gentrifying, and that the data overlooks details necessary to fully understand gentrification, such as local housing and business developments [8]. Critics also argue that gentrification results in quantitative and qualitative changes in urban design, so both types of investigation are necessary to achieve a comprehensive analysis. While it is important to acknowledge these criticisms and recognize that exclusively Census-based studies are unlikely to perfectly capture the processes of gentrification, quantitative approaches can offer new ways to look at existing problems and provide a more rigorous method of examining variable change over time.

## 1.2 Project Description

In this project, we will use Seattle as a case study to examine trends in neighborhood change and evaluate the severity of gentrification and urban segregation within the city between 2000 and 2017. Even though displacement is difficult to track and identify, demographic changes at the neighborhood level suggest when and where it has occurred. We will employ statistical techniques and exploratory data analysis on Census data to understand how Seattle neighborhoods have changed over time in an attempt to draw conclusions about the nature of gentrification and assess its presence in the city. We will begin by discussing the research questions that we seek to answer. We will go on to give an overview of Census data and terminology as well as explain how the data sets and variables used in this study were selected. Furthermore, we will explore the theory behind the statistical techniques, apply these methods to our data sets, and discuss the results. Next, we will employ data visualization through maps and plots to further analyze the data. Finally, we will draw conclusions about our findings and offer suggestions for future research on this topic.

## 1.3 Research Questions

Our interest in this topic stems from the fact that gentrification has become a widespread social issue throughout recent decades, which negatively impacts residents in cities across the globe. Furthermore, gentrification perpetuates inequitable urban models that favor wealthier populations, while marginalizing and isolating other communities. Given that Seattle has experienced rapid growth that consists of a professionalized workforce with high socioeconomic status, the city serves as an interesting setting to study the processes of gentrification. As a method of examining the complexities of neighborhood change, we identify the following research questions that we hope to draw conclusions about:

1. What is the nature of gentrification in Seattle and what neighborhoods and regions in the city have experienced the greatest effects of the phenomenon?

2. How does socioeconomic status affect diversity and segregation of Seattle residents?

3. Can we detect patterns of displacement and urban migration due to gentrification in the city of Seattle?

These questions will help frame our variable selection and inform the statistical techniques that are appropriate for analyzing our data. We will return to these questions in the conclusion section of the study.

# 2 Framework and Approach

## 2.1 Census Data

The data in this study were collected through the *United States Census Bureau*, which is the federal government's largest statistical agency. The first Census was inaugurated in 1790, and there has been a Decennial Census every decade since then [3]. The Census is mandated by the Constitution, which was initially established to enumerate the population and determine representation in Congress. Throughout the 20th century, most addresses received a "short form" of the questionnaire, while approximately 1 in 6 households were sent a more detailed "long form." The short form was designed to collect basic demographic and housing information, such as age, race, and sex. On the other hand, the long form collected more detailed social, economic, and housing information, such as citizenship, educational attainment, disability status, employment status, income, and housing costs. However, in the early 1990s, the need for more nationally consistent statistics prompted the federal government policymakers to discuss the possibility of collecting long form data more regularly throughout each decade [3].

## 2.2 American Community Survey

The benefits of providing more frequent statistics and more efficient procedures led the Bureau to consider an ongoing measurement, which later became the *American Community Survey* (ACS), initially released in 2005. This shift resulted in the establishment of the 2010 Census as a short form only questionnaire [2]. The ACS has 3 different data sets available: 1-year, 3-year, or 5-year estimates. The 1-year estimates provide the most current data, but has a small sample size and is less reliable than the 3-year or 5-year estimates. On the other hand, the 5-year estimates are the most reliable and have the largest sample size. For the purposes of this study, we will be using the 5-year estimates because they are best suited for analyzing data at more detailed geography levels, such as *tracts*, which are areas

approximately equivalent to a neighborhood [15]. The spatial size of Census tracts vary widely depending on the density of the region, with a maximum of 8,000 inhabitants and a minimum of 1,200 people. Tract boundaries were established with the intention of being maintained over decades to allow statistical comparisons to be made between Censuses, serving as a useful tool for studying neighborhood change [14]. However, physical changes in street patterns caused by new developments and population growth or decline may require occasional boundary revisions. The tract boundaries from the 2010 Census in Seattle are illustrated in Figure 1.

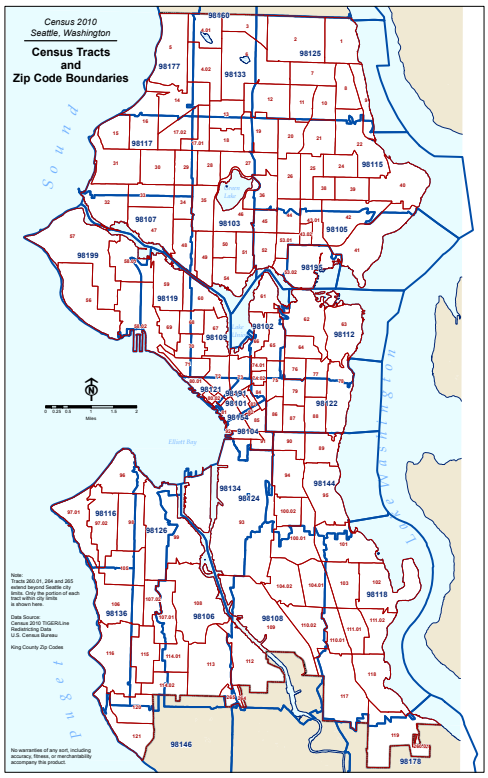



Figure 1: Census Tract Boundaries in Seattle [1]

## 2.3 American FactFinder

We used the *American FactFinder* (AFF) resource available through the Census to create our data sets. The data in AFF come from several censuses and surveys including the 2000 and 2010 Decennial Censuses and the American Community Survey. AFF serves as the Census Bureau's free online, self-service tool designed to investigate a variety of population, economic, geographic, and housing information. It allows users to customize and download data sets for a particular year based on their selected variables. Since we wish to analyze neighborhood change over time, we will focus on three distinct data sets: the 2000 Decennial Census, the 2010 ACS 5-year estimate, and the 2017 ACS 5-year estimate. After choosing our

variables, we downloaded individual data sets for each variable by year, and then compiled comprehensive data sets for the years 2000, 2010, and 2017.

## 2.4 Layout of the Study

In order to select the variables to use in this study, we must return to our initial research questions described in Section 1.3. Considering that gentrification is often associated with the displacement of communities of color, we will examine racial demographics. Furthermore, we are interested in studying rising housing and rent prices, so these variables will play an important role in our analysis. It is also necessary for us to consider the relationship between impoverished and wealthier communities, which will allow us to identify socioeconomic divisions across neighborhoods. Finally, we are also interested in analyzing educational characteristics, since processes of gentrification involve more educated populations moving into urban spaces. Hence, the variables we select to study will allow us to examine how the relationships between race, socioeconomic status, access to education, and housing prices have changed over time, which are all factors rooted in the social or economic aspects of gentrification. Taking these factors into account, the variables selected were also based on previous research done in the field of quantitative gentrification analysis [16]. Based on our research questions and prior studies, we identify the following 9 numerical variables that we will use in our analysis, which can be divided into 3 categories based on their characteristics:

**Population**

- Percentage of total population White alone
- Percentage of total population Black alone
- Percentage of total population Asian alone
- Percentage of total population Hispanic or Latino

**Socioeconomic**

- Percentage of total population age 25+ with bachelor's degree or higher
- Median household income
- Percentage of total population living in poverty

**Housing**

- Median house value
- Median gross rent

It is also important to acknowledge that gentrification is distinct from certain trends in neighborhood change. Other quantitative studies have developed methods for separating neighborhoods that have the potential to gentrify from those that do not. Gentrification is defined as a process in which wealthier residents move into previously low-income neighborhoods. Therefore, gentrification is only appropriate to study neighborhoods that initially housed lower-income populations. Prior research has implemented the following procedure: an eligible "gentrifiable" tract must have an income level that is less than 80% of the metropolitan area's median income, which is consistent with the criteria used by the U.S. Department of Housing and Urban Development [8].

We are interested in developing a measure for differentiating gentrification from other forms of neighborhood change. More specifically, we seek to build groupings of "similar" neighborhoods and identify variables and indicators that illustrate distinctions between the groups. In order to accomplish this, we will employ the statistical techniques *principal components analysis* (PCA) and *clustering*, which are described in Section 3. All figures displayed in Section 4 and Section 5 were created using the statistical software *JMP*. We will use PCA on our 3 distinct data sets to discover underlying factors in our data, and examine how the relationship among our variables has changed across the separate data sets. We will also utilize clustering to group together tracts that share similarities based on the results of PCA, and proceed by analyzing the data geospatially for the 3 data sets. We will then create a combined data set that aggregates the separate years into 1 data set, in order to more directly examine how individual variables have changed over time. Before applying the statistical methods, we will inspect data summaries of the 3 data sets to better understand our data.

Table 1: Data Summaries

| | 2000 Decennial Census | 2010 ACS 5-Year Estimate | 2017 ACS 5-Year Estimate |
|---|---|---|---|
| Percentage of total population White alone | 71.1 | 70.1 | 68.4 |
| Percentage of total population Black alone | 8.2 | 7.9 | 7.2 |
| Percentage of total population Asian alone | 12.4 | 13.8 | 14.4 |
| Percentage of population Hispanic or Latino | 5.4 | 6.4 | 6.7 |
| Percentage of total population age 25+ with bachelor's degree or higher | 46.9 | 54 | 60.2 |
| Median household income (dollars) | 47,901 | 64,570 | 84,215 |
| Percentage of total population living in poverty | 12.6 | 13.5 | 12.7 |
| Median house value (dollars) | 278,257 | 461,169 | 534,636 |
| Median gross rent (dollars) | 753 | 1012 | 1406 |

From the results in Table 1, we note that the percentage of the White alone population has slightly decreased between 2000 and 2017, as has the percentage of the Black alone population. However, the percentages of the Asian and Hispanic/Latino populations have slightly increased over the years. We also observe that Seattle has become steadily more educated over time, increasing by 28.4 percent from 2000 to 2017. The median household income has also increased by 75.8 percent, while the percentage of people living in poverty has remained

relatively constant. Not surprisingly, the median house value has significantly changed, increasing by 92.2 percent from 2000 to 2017. Finally, the median gross rent dramatically increased by 86.7 percent. While racial makeup has remained relatively fixed over the time period, we conclude that Seattle has become substantially more expensive and associated with wealthier populations from 2000 to 2017.

We point out that the data sets created from American FactFinder included a small number of missing observations, due to limited data available in some of the tracts. In order to not greatly affect our analyses, we performed *linear regression* on these tracts to predict the value of the missing observations. Since some of these values are predictions, we acknowledge that they may not be entirely accurate and representative of the data in these tracts. Furthermore, we have removed the tract that surrounds the University of Washington, since it is not an appropriate neighborhood to study for the effects of gentrification, due to the majority student population. We also note that because Census tracts change minimally over time, there are 125 tracts in the 2000 data set and 134 tracts in the 2010 and 2017 data sets. For the purposes of mapping and preserving continuity among the 3 data sets, we have slightly modified the 2000 data to account for the additional tracts. We also recognize that these changes may affect the full accuracy of our data and could contribute to potential sources of error.

# 3 Description of Statistical Methods Used in the Study

In this study, we will be focusing on two main statistical techniques: principal components analysis (PCA) and cluster analysis, or clustering. Both PCA and clustering are examples of *unsupervised learning*, which refers to the situation in which for every observation $i = 1, \ldots, n$, we observe a vector of measurements $x_i$, but no associated response $y_i$. On the other hand, traditional statistical techniques, such as *linear regression*, are classified as *supervised learning*, because there is an associated response measurement $y_i$ for each predictor measurement(s) $x_i, i = 1, \ldots, n$. In unsupervised learning, it is not possible to fit a linear regression model, since there is no response variable to predict. Rather, unsupervised learning techniques such as PCA and clustering allow us to understand relationships between the variables or observations and offers an informative way to visualize the data, enabling us to discover subgroups among the variables [4]. We will proceed by describing each of these statistical techniques and will then apply these methods to our data sets.

## 3.1 Principal Components Analysis

PCA is concerned with explaining the *variance-covariance* relationship of a set of variables through a linear combination of these variables. When handling a large set of possibly cor-

related variables, (PCA) allows us to explain this set with a smaller number of uncorrelated variables that collectively explain most of the variability in the original data set. Since PCA is an unsupervised approach, we are not interested in prediction, because we do not have an associated response variable [4]. PCA serves as a valuable data visualization tool, and can help identify patterns and trends in a data set. Furthermore, an analysis of principal components often reveals relationships that were not previously suspected and thus allows interpretations that would not otherwise be obvious [5].

We will briefly define some fundamental mathematical terms that we will use in our explanation of PCA.

**Variance**: The spread of data around its mean value [13].

**Correlation**: A demonstration of how strongly two variables are related to one another [13].

**Covariance**: A measure of the strength of the correlation between two or more sets of random variables [13].

**Eigenvector**: A nonzero vector $\mathbf{x}$ of an $n \times n$ matrix $\mathbf{A}$ such that $\mathbf{Ax} = \lambda\mathbf{x}$ for some scalar $\lambda$ [7].

**Eigenvalue**: A scalar $\lambda$ such that there is a nontrivial solution $\mathbf{x}$ of $\mathbf{Ax} = \lambda\mathbf{x}$ [7].

Further explanation of *linear algebra* concepts described in this section can be found in [7].

### 3.1.1 How does PCA work?

Suppose that we want to examine $n$ observations with measurements on a set of $p$ features, $X_1, X_2, \ldots, X_p$ as part of our data analysis. One way to do this may be to examine two-dimensional scatterplots of the data, each of which displays the $n$ observations' measurements on two of the features. However, we note that there are $\binom{p}{2}$ such scatterplots, which may be quite cumbersome to analyze. Hence, if $p$ is a large number, then often it is not practical or possible to look at all of the graphs. Additionally, it is probable that none of them would be explanatory, since they each exhibit a fraction of the total information present in the data set. Thus, it is clear that we need a better method to find a low-dimensional representation of the data that captures as much of the information as possible [4].

PCA offers a solution to this problem by finding a low-dimensional representation that contains as much as possible of the variation. Each of the $n$ observations exists in a $p$-dimensional

space, but not all of these dimensions are equally meaningful. Thus, PCA seeks a small number of dimensions that are as meaningful as possible, which is measured by the amount that the observations vary along each dimension. Each of the dimensions identified by PCA is a linear combination of the $p$ features. We also note that PCA does not require a multivariate normal assumption. We will discuss the way that these dimensions, or *principal components* (PC) are found [4].

Suppose we have a data set with $p$ numerical variables which each contain $n$ observations. These data values define $p$ $n$-dimensional vectors $x_1, \ldots, x_p$, or an $n \times p$ data matrix $\mathbf{X}$, whose $j$th column represents the vector $x_j$ of observations on the $j$th variable. In PCA, we seek a linear combination of the columns of matrix $\mathbf{X}$ with maximum variance. Such linear combinations are given by $\sum_{j=1}^{p} \mathbf{v}_j \mathbf{x}_j = \mathbf{Xv}$, where $\mathbf{v}$ is a vector of constants $v_1, v_2, \ldots, v_p$. The variance of any such linear combination is described by var($\mathbf{Xv}$)=$\mathbf{v'Sv}$, where $\mathbf{S}$ denotes the sample covariance matrix associated with the dataset and $'$ indicates transpose. Thus, finding the linear combination with maximum variance is identical to producing a $p$-dimensional vector $\mathbf{v}$ which maximizes $\mathbf{v'Sv}$. To solve this problem, we must impose an additional restriction which requires that $\mathbf{v'v} = 1$. This problem is equivalent to maximizing the equation $\mathbf{v'Sv}$ - $\lambda(\mathbf{v'v}$ - $\mathbf{1})$, where $\lambda$ is a Lagrange multiplier. If we differentiate with respect to the vector $\mathbf{v}$ and equate to the null vector, we find that

$$\mathbf{Sv} - \lambda\mathbf{v} = 0 \iff \mathbf{Sv} = \lambda\mathbf{v}. \tag{1}$$

Hence, $\mathbf{v}$ must be a unit-norm eigenvector, and $\lambda$ denotes the corresponding eigenvalue of the covariance matrix $\mathbf{S}$. We are particularly interested in the *largest* eigvenvalue $\lambda_1$ and the corresponding eigenvector $\mathrm{v}_1$, since the eigenvalues are the variances of the linear combinations demonstrated by the corresponding eigenvector $\mathbf{v}$: var($\mathbf{Xv}$) = $\mathbf{v'Xv} = \lambda\mathbf{v'v} = \lambda$. We also note that equation (1) remains valid if the eigenvectors are multiplied by -1, so the signs of the principal components are scores and only their relative magnitudes and sign patterns are important.

A Lagrange multipliers approach, with the restrictions of *orthogonality* of different coefficients vectors, can be used to show that the full set of eigenvectors of $\mathbf{S}$ are the solutions to the problem of producing up to $p$ new linear combinations $\mathbf{Xv}_k = \sum_{j=1}^{p} v_{jk}\ \mathbf{x}_j$, which has maximal variance subject to *uncorrelatedness* with previous linear combinations. Uncorrelatedness stems from the fact that the covariance between two such linear combinations, $\mathbf{Xv}_k$ and $\mathbf{Xv}_{k'}$ is given by $\mathbf{v'}_{k'}\mathbf{Sv}_k = \lambda_k\ \mathbf{v'}_{k'}\ \mathbf{v}_k = 0$ if $k' \neq k$.

The linear combinations $\mathbf{Zv}_k$ are referred to as the *principal components* of the data set. Furthermore, the elements of the eigenvectors $\mathbf{v}_k$ are called the principal component *load-*

*ings* and the elements of the linear combinations $\mathbf{Xv}_k$ are called the principal component *scores*, since they are the values that each individual would score on a given PC.

In standard approaches, it is customary to define principal components as the linear combinations of the centered variables $\mathbf{x}_j^*$ with the equation $\mathbf{x}_{ij}^* = \mathbf{x}_{ij} - \bar{\mathbf{x}}_j$, where $\bar{\mathbf{x}}_j$ represents the mean value of the observations on variable $j$. We note that this technique does not change the solution, but rather it allows us to understand a more geometric approach to PCA. By labeling $\mathbf{X}^*$ as the $n \times p$ matrix whose columns are the centered variables $\mathbf{x}_j^*$ we have

$$(n-1)\mathbf{S} = \mathbf{X}^{*'}\mathbf{X}^*. \tag{2}$$

Equation 2.2 connects the *eigendecomposition* of the covariance matrix $\mathbf{S}$ with the *singular value decomposition* of the column-centered matrix $\mathbf{X}^*$. Any arbitrary matrix $\mathbf{Y}$ of dimension $n \times p$ and rank $r$ can be expressed as

$$\mathbf{Y}=\mathbf{UDA}', \tag{3}$$

where $\mathbf{U,A}$ are $n \times r$ and $p \times r$ matrices with *orthonormal* columns. Hence, $\mathbf{U}'\mathbf{U} = \mathbf{I}_r = \mathbf{A}'\mathbf{A}$, where $\mathbf{I}_r$ represents the $r \times r$ identity matrix and $\mathbf{D}$ is an $r \times r$ diagonal matrix. The diagonal elements of the matrix $\mathbf{D}$ are called the *singular values* of $\mathbf{Y}$, which denote the non-negative square roots of the non-zero eigenvalues of the matrices $\mathbf{Y}'\mathbf{Y}$ and $\mathbf{YY}'$. Because of the orthogonality of the columns of $\mathbf{A}$, the columns of the matrix product $\mathbf{X}^*\mathbf{A} = \mathbf{UDA}'\mathbf{A} = \mathbf{UD}$ represent the principal components of $\mathbf{X}^*$. The variances of these principal components are given by the squares of the singular values of $\mathbf{X}^*$, divided by $n-1$. It follows that

$$(n-1)\mathbf{S} = \mathbf{X}^{*'}\mathbf{X}^* = (\mathbf{UDA}')'(\mathbf{UDA}') = \mathbf{ADU}'\mathbf{UDA}' = \mathbf{AD}^2\mathbf{A}', \tag{4}$$

where $\mathbf{D}^2$ is the diagonal matrix with the squared singular values (the eigenvalues of $(n-1)\mathbf{S}$). Equation 4 gives the *spectral decomposition*, or *eigendecomposition*, of matrix $(n-1)\mathbf{S}$. Thus, PCA is equivalent to the singular value decomposition (SVD) of the column-centered data matrix $\mathbf{X}^*$ [6].

Now that we have explored the theory behind PCA, we also must address the question of how many components are necessary to include in the analysis of our data. There is no definitive answer to this problem, but there are factors we can consider when making our selections. Aspects to consider include the amount of total variance explained and the relative sizes of the eigenvalues. An informative visual aid to assessing the appropriate number of components is a *scree plot*. With the eigenvalues ordered from largest to smallest, a scree plot shows how much variation each principal component captures from the data. To determine the adequate number of components, we look for an elbow (bend) in the scree

plot. Hence, the appropriate number of components is illustrated by the point at which the remaining eigenvalues are relatively small and at a constant level. We recall that eigenvalues measure the amount of variation in the total sample accounted for by each factor. Hence, if a factor has a low eigenvalue, then it is contributing little to the explanation of variances and may be ignored. According to the *Kaiser's rule*, we should only retain eigenvalues that are greater than 1 [10].

## 3.2 Hierarchical Clustering

The term clustering refers to a broad range of techniques for finding *subgroups*, or *clusters* within a data set. When clustering the observations of a particular data set, the objective is to partition them into distinct groups so that the observations within each group are similar to each other, while observations in other groups are different from each other. However, we must define what it means for two or more observations to be *similar* or *different*. Like PCA, clustering is an unsupervised approach because we do not have a particular response variable $Y$ that we are modeling. While both PCA and clustering are data reduction techniques, their mechanisms are distinct. The goal of PCA is to find a low-dimensional representation of the observations that explain a significant percentage of the variance. On the other hand, clustering seeks to determine, on the basis of $x_1, \ldots, x_n$, whether the observations fall into relatively distinct groups [4].

When handling a large number of observations, we can rarely examine all grouping possibilities, even with advanced technologies. Hence, a wide range of clustering algorithms have been developed that seek to find "reasonable" clusters without having to analyze all combinations. Thus, there exist a great number of of clustering techniques, but the most popular approaches are *K-means clustering* and *hierarchical clustering*. In $K$-means clustering, the goal is to partition the observations into a pre-specified number of clusters. On the other hand, hierarchical clustering does not require us to indicate the number of clusters in advance. Rather, it results in a tree-like two-dimensional visual representation of the observations, which is referred to as a *dendrogram*. The dendrogram displays the fusions or divisions that have been performed at any given level. For the purposes of this study, we will be focusing on hierarchical clustering.

We will examine *bottom-up* or *agglomerative* clustering, which is the most common form of hierarchical clustering and begins with the individual objects. Each *leaf* of the dendrogram represents one of the $n$ observations. As we move further up the tree, some leaves begin to fuse into *branches*, which correspond to observations that are similar to each other. As we continue to move up the dendrogram, the branches begin to fuse either with other leaves or branches. The earlier the fusions occur, the more similar the observations are to each other.

Thus, observations that fuse further up the dendrogram can actually be quite different from one another. The height of the fusion, as measured on the horizontal axis, tells us how different any two observations are [4].

### 3.2.1 The Hierarchical Clustering Algorithm

In order to apply hierarchical clustering, we must first begin by defining some measure of *dissimilarity* between every pair of observations. *Euclidean distance* is most often used, which is computed using the formula $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$, where $(x_1, y_1)$ denotes the first point and $(x_2, y_2)$ denotes the second point. The algorithm proceeds as follows:

1. Begin with $n$ observations and a measure (such as Euclidean distance). Starting at the bottom of the dendrogram, each of the $n$ observations is treated as its own cluster.

2. For $i = n, n-1, \ldots, 2$ :

   - Examine all pairwise inter-cluster dissimilarities among the $i$ clusters and identify the pair of clusters that are the most similar. Proceed by merging these clusters according to their similarities.
   - Compute the near pairwise inter-cluster dissimilarities among the remaining $i-1$ clusters [4].

Hence, as the similarities decrease, all subgroups are eventually fused into a single cluster.

### 3.2.2 Ward's Hierarchical Clustering Method

In this study, we will be using *Ward's hierarchical clustering procedure*, which seeks to minimize the total within-cluster variance. Specifically, Ward's method states that the distance between two clusters $A$ and $B$ is how much the *error sum of squares* (ESS) will increase when merged. For a given cluster $k$, let $ESS_k$ denote the sum of the squared deviations of each item in the cluster from the cluster mean, which is referred to as the *centroid*. If we suppose there are $K$ clusters, then $\text{ESS} = \text{ESS}_1 + \text{ESS}_2 + \ldots + \text{ESS}_k$.

During each step in the analysis, the union of every possible pair of clusters is considered, and the two clusters whose fusion yields the smallest increase in ESS (the minimum loss of information) are merged. At the beginning of the process, each cluster consists of a single observation, so the value of $\text{ESS} = 0$. When all of the clusters are combined into a single group of $n$ observations, the value of ESS is computed as:

$$\text{ESS} = \sum_{j=1}^{n} (x_j - \hat{x})'(x_j - \hat{x}),$$

where $x_j$ represents the value for the multivariate measurement associated with the $j^{th}$ observation and $\hat{x}$ represents the mean of all the items [5]. We will utilize Ward's hierarchical clustering method in our analysis to identify clusters of Census tracts that are similar.

# 4 Data Analysis

We will apply the statistical techniques discussed in the previous section to our data sets, beginning with PCA. After performing the initial PCA on our data, we will use hierarchical clustering to group together Census tracts that share similar principal component scores and display these clusters geospatially. This will allow us to identify tracts that are associated with particular variables, as well as discuss how the relationships between our variables have changed over time.

## 4.1 Analysis of Census Data by Year

We will use PCA as a data reduction technique for the entire dataset of Census tracts in the 2000, 2010, and 2017 data sets. PCA is appropriate for the purposes of this study because it allows us to maximize the amount of common variance explained by the whole dataset through the fewest number of components, while at the same time maximizing the amount of unique variance explained by individual variables on each component [16]. The dimensions social status, family status, and ethnic status were commonly examined in principal components analyses of previous studies. PCA is a valuable tool in the field of quantitative gentrification research because it allows us to identify components that distinguish the phenomenon from other types of neighborhood change, offering a foundation for interpreting patterns of gentrification in urban environments [16].

We note that PCA works best with numerical data, which all of our variables are. Before performing the PCA, we standardize the data to have mean 0 and standard deviation 1, which is necessary because our data involves two different units (estimates and percentages). The data are normalized using $z$-scores to ensure that each variable is presented in terms of standard deviations from its mean. When scaling our variables, the data can be transformed as $z_i = \frac{x_i - \mu}{\sigma}$, where $\mu$ represents the mean of the $x$ values, and $\sigma$ represents the standard deviation.

Table 2 denotes the variables examined in this study with their corresponding abbreviations, which we will use to refer to them throughout the remainder of this study. Figure 2 represents a map of Seattle with neighborhood names, which we will also be referencing in later sections.

Table 2: Variable Names

| | |
|---|---|
| Percentage of total population White alone | EP_WHITE |
| Percentage of total population Black alone | EP_BLACK |
| Percentage of total population Asian alone | EP_ASIAN |
| Percentage of total Hispanic or Latino | EP_HISP |
| Percentage of total population age 25+ with a bachelor's degree or higher | EP_BACH25 |
| Median household income | MEDHI |
| Percentage of total population living in poverty | EP_POV |
| Median house value | MEDHV |
| Median gross rent | MEDGR |

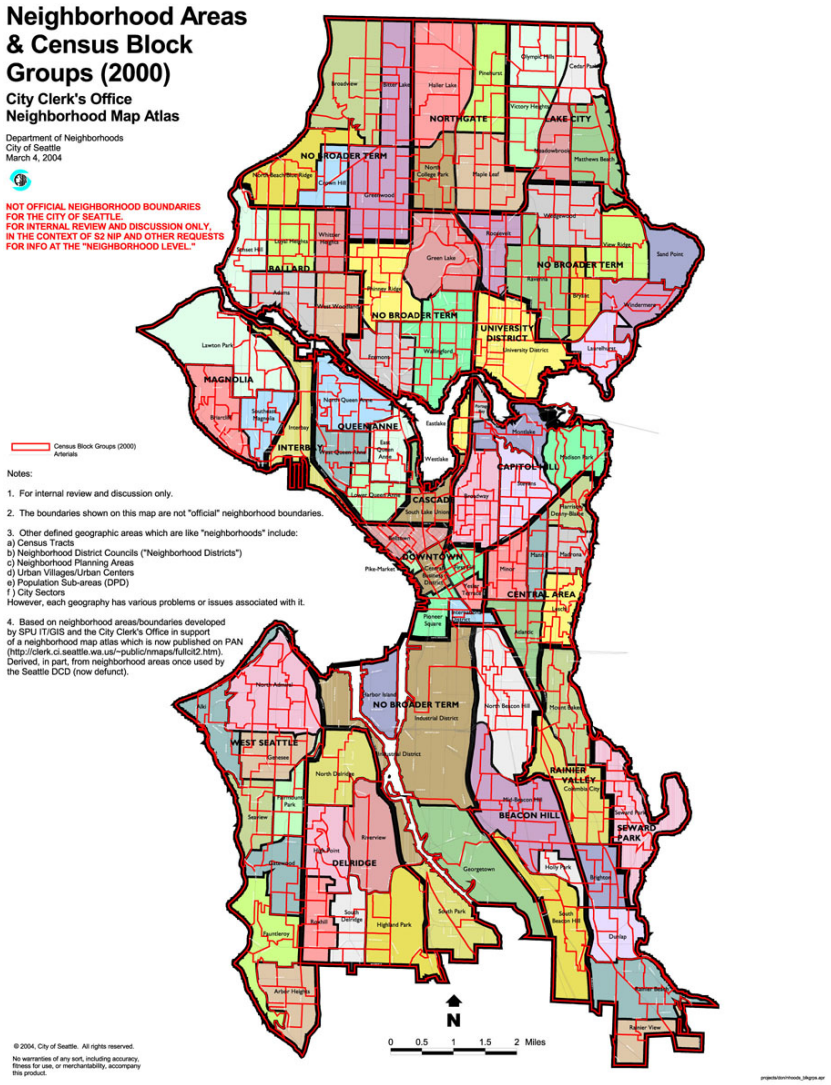


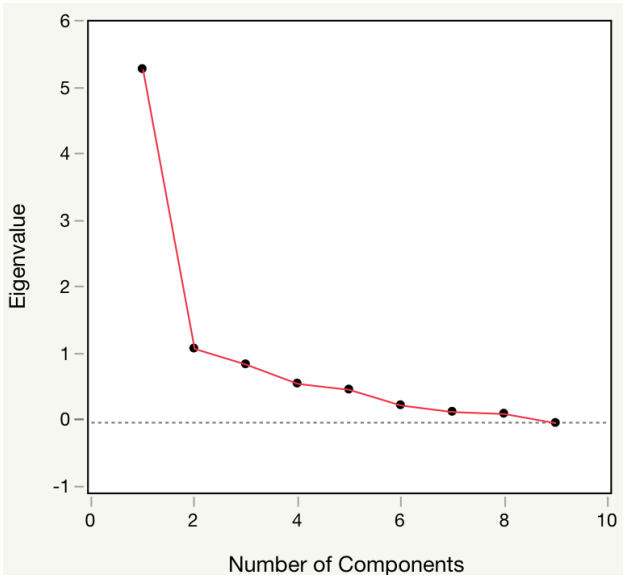

Figure 2: Map of Seattle Neighborhoods [9]

### 4.1.1 2000 Census Results

We will begin by performing PCA on the 2000 Census data set, which consists of data on 125 census tracts with the 9 variables indicated previously. After standardizing our data, we display the variance explained by each of the components and examine the resulting scree plot.

| Number | Eigenvalue | Percent | Cum Percent |
|---|---|---|---|
| 1 | 5.3214 | 59.127 | 59.127 |
| 2 | 1.1648 | 12.942 | 72.068 |
| 3 | 0.8709 | 9.677 | 81.745 |
| 4 | 0.5843 | 6.492 | 88.237 |
| 5 | 0.4863 | 5.403 | 93.640 |
| 6 | 0.2665 | 2.961 | 96.601 |
| 7 | 0.1623 | 1.803 | 98.404 |
| 8 | 0.1416 | 1.573 | 99.978 |
| 9 | 0.0020 | 0.022 | 100.000 |

(a)




(b)

Figure 3: Scree Plot of 2000 Data

From (a) in Figure 3, we obtain 9 principal components, each of which explains a percentage of the total variation in the dataset. The first principal component (PC1) explains 59.1% of the total variance, the second principal component (PC2) explains 12.9% of the total variance, and so on. Thus, the first two principal components, collectively, explain 72.1% of the total variance. Consequently, we can conclude that the variation is well summarized by two principal components and a reduction in the data from 125 observations on 9 variables to 125 observations on 2 principal components is reasonable.

To determine the appropriate number of components to keep in our study, we will examine the scree plot (b) in Figure 3. We observe that an elbow occurs when the number of components is equal to 2. That is, the eigenvalues after PC2 are relatively similar and small, so we conclude that 2 principal components effectively summarize the total variance.

Next, we will examine the principal component loading vectors for the 2000 Census data, which is displayed in Table 3. From the signs of the loadings, we can see that MEDHI, EP_WHITE, MEDHV, MEDGR, and EP_BACH25 all have positive loadings for PC1. We

can further identify that MEDHI and MEDGR share very similar PC1 loadings of 0.801 and 0.788, respectively. Additionally, EP_WHITE and EP_BACH25 also share similar PC1 loadings of 0.892 and 0.884, respectively. We note that the PC1 loadings for EP_POV, EP_BLACK, EP_ASIAN, and EP_HISP are all negative.

Table 3: Loading Matrix

| | PC1 | PC2 |
|---|---|---|
| MEDHI | 0.8012804 | 0.502793874 |
| EP_POV | -0.7466165 | -0.480274977 |
| EP_WHITE | 0.8924837 | -0.371435396 |
| EP_BLACK | -0.6866631 | 0.337992184 |
| EP_ASIAN | -0.7400126 | 0.353323718 |
| EP_HISP | -0.6631633 | 0.007615572 |
| MEDHV | 0.6787695 | -0.194956613 |
| MEDGR | 0.7884089 | 0.440692534 |
| EP_BACH25 | 0.8839708 | -0.165700341 |

We can confirm these observations visually by examining a *biplot*, which allows us to represent both the principal component scores and the loading vectors in a single plot. Figure 4 plots the first two principal components of these data.

Using the biplot, we can draw conclusions about which variables are similar and which are different, while also understanding how each variable contributes to each principal component. The red arrows indicate the first two principal component loading vectors. In Figure 4, we see that the first loading vector places approximately equal weight on MEDHI, MEDGR, EP_BACH25, MEDHV, and EP_WHITE. Each of the points represents a Census tract, which have been grouped into 7 distinct clusters using hierarchical clustering. Hence, each cluster illustrates a group of individual tracts that share similar profiles.

As we observed from the loading matrix, we note that MEDHI, MEDGR, EP_BACH25, MEDHV, and EP_WHITE are located close to together, indicating that these variables are correlated with each other. Hence, from the biplot we can conclude that Census tracts with higher median household income and median gross rent tend to have higher percentages of white, well-educated residents, as well as higher median house values. Similarly, we can also infer that tracts that house higher percentages of people of color (Asian, Black, and Hispanic communities), tend to have higher rates of poverty.

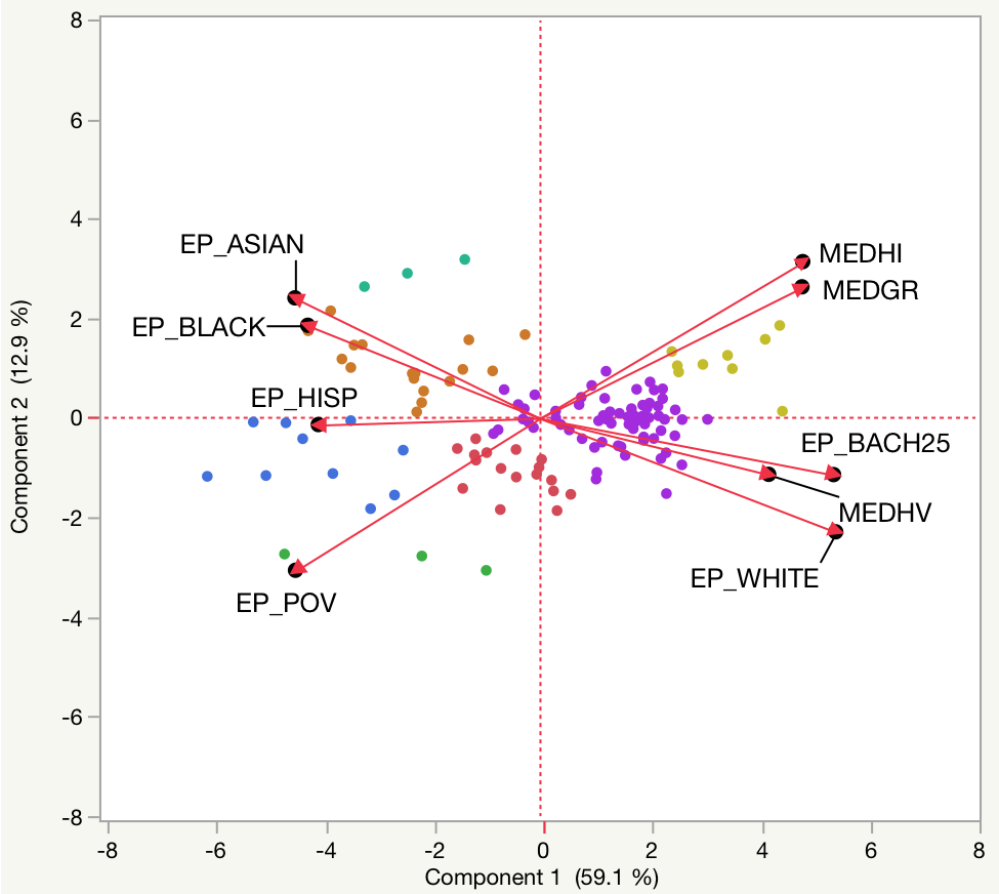



Figure 4: Biplot of 2000 Data

We can also examine differences between the tracts using the 2 principal component score vectors shown in Figure 4. Our discussion of the loading vectors suggests that tracts with large positive scores on the first component, such as the yellow cluster, have a high median household income and median gross rent, while tracts such as the cluster in dark-blue, with negative scores on the first component, display lower median household incomes and median gross rent. Hence, PC1 highlights disparities in socioeconomic status, and the variables that exhibit positive PC1 loadings emphasize a clear division between tracts with high and low socioeconomic status. We can also analyze the PC loadings geospatially by examining the clusters on a map of Seattle, as illustrated in Figure 5.

From the map in Figure 5, we note that tracts with positive PC1 scores indicating higher socioeonomic status (the yellow tracts), tend to be located near the waters of Lake Washington, Lake Union, and Green Lake, as well as in Magnolia and Queen Anne. These neighborhoods are associated with excellent views of Puget Sound and the Olympic Mountains. Conversely, areas of low performance, with negative factor scores that signify lower socioeconomic status (the dark-blue and green clusters), are primarily situated near downtown in the International

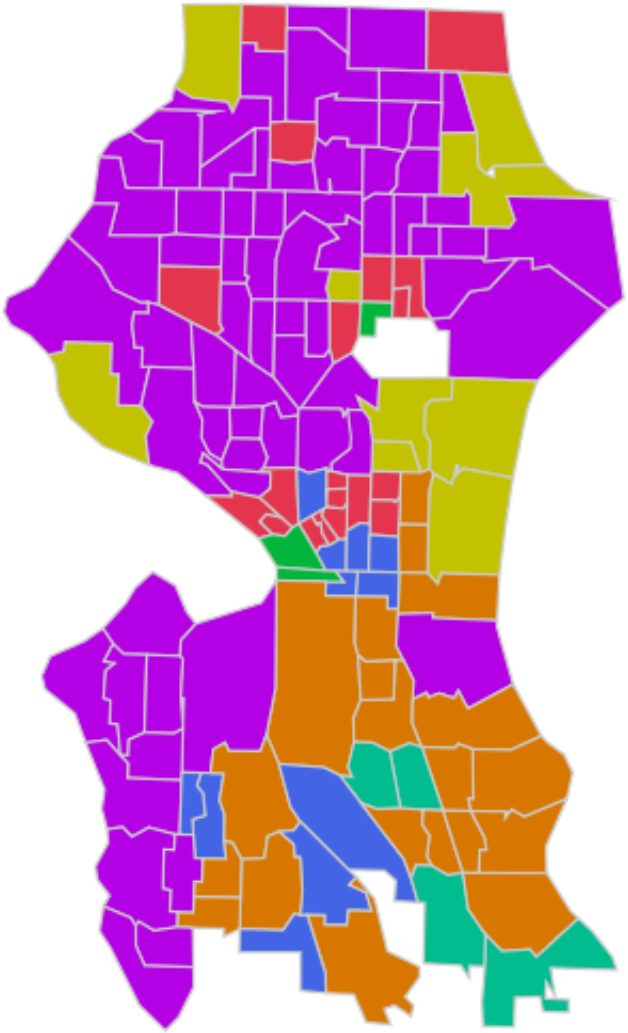


Figure 5: Clustered Map of Principal Component Scores

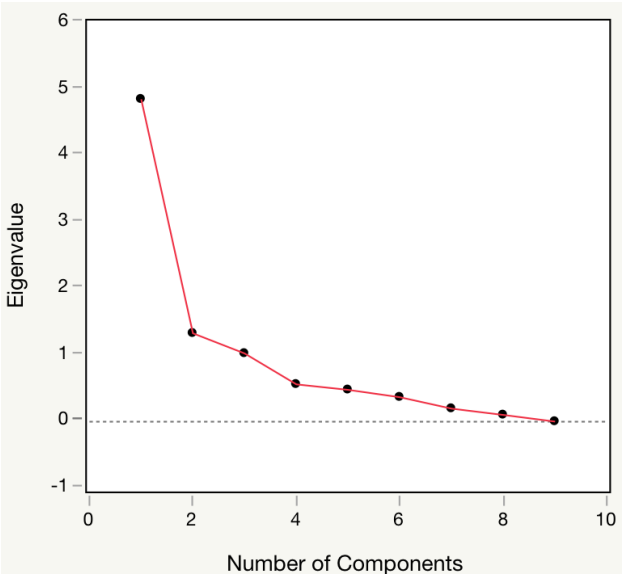
and Central Districts, as well in the southern neighborhoods of South Park, Georgetown, and Delridge. We will proceed by analyzing the 2010 and 2017 data sets.

#### 4.1.2 2010 ACS Results

From (a) in Figure 6, we note that PC1 explains 53.9% of the total variance and PC2 explains 14.9% of the total variance. Hence, PC1 and PC2 combined can explain 68.8% of the total variance. In order to determine the appropriate number of principal components to retain, we will examine the scree plot (b) in Figure 6. We note that the elbow occurs at PC2, so we conclude that the first 2 principal components are sufficient to explain the total variance.

| Number | Eigenvalue | Percent | Cum Percent |
|---|---|---|---|
| 1 | 4.8547 | 53.941 | 53.941 |
| 2 | 1.3385 | 14.872 | 68.813 |
| 3 | 1.0358 | 11.508 | 80.322 |
| 4 | 0.5716 | 6.351 | 86.673 |
| 5 | 0.4891 | 5.434 | 92.107 |
| 6 | 0.3785 | 4.206 | 96.312 |
| 7 | 0.2071 | 2.301 | 98.614 |
| 8 | 0.1122 | 1.246 | 99.860 |
| 9 | 0.0126 | 0.140 | 100.000 |

(a)




(b)

Figure 6: Scree Plot of 2010 Data

We will proceed by examining the loading matrix for PC1 and PC2, shown in Table 4. Similar to our observations from the 2000 data set, we note that the PC1 displays positive loadings for MEDHI, EP_WHITE, MEDHV, MEDGR, and EP_BACH25. On the other hand, EP_POV, EP_BLACK, EP_ASIAN, and EP_HISP have negative loadings for PC1. From Table 4, we see that EP_WHITE and EP_BACH25 possess very similar positive loadings for PC1 of 0.872 and 0.854, respectively, suggesting that these variables are highly correlated. Similarly, EP_BLACK and EP_ASIAN also exhibit very similar negative loadings for PC1 of -0.733 and -0.704, respectively, indicating that these variables are correlated as well.

Table 4: Loading Matrix

| | PC1 | PC2 |
|---|---|---|
| MEDHI | 0.8179987 | -0.41009227 |
| EP_POV | -0.6529669 | 0.26623823 |
| EP_WHITE | 0.8719558 | 0.46448681 |
| EP_BLACK | -0.7327120 | -0.41691669 |
| EP_ASIAN | -0.7040159 | -0.55791089 |
| EP_HISP | -0.4059518 | 0.48921386 |
| MEDHV | 0.7743020 | -0.25873736 |
| MEDGR | 0.6871389 | -0.30348775 |
| EP_BACH25 | 0.8543215 | 0.01580913 |

Next, we will visually analyze the biplot of the 2010 data, which is illustrated in Figure 7. It is important to note that the colors of these clusters are not related to the ones in Figure 4. We note that MEDGR, MEDHI, and MEDHV are close together and point in virtually the same direction, suggesting that they are positively correlated. Additionally, we observe that the variables EP_POV and EP_HISP are more closely related than in the 2000 biplot. We also observe that there is an inverse relationship between the variables EP_WHITE and EP_ASIAN/EP_BLACK, which highlights a racial divide in these tracts.

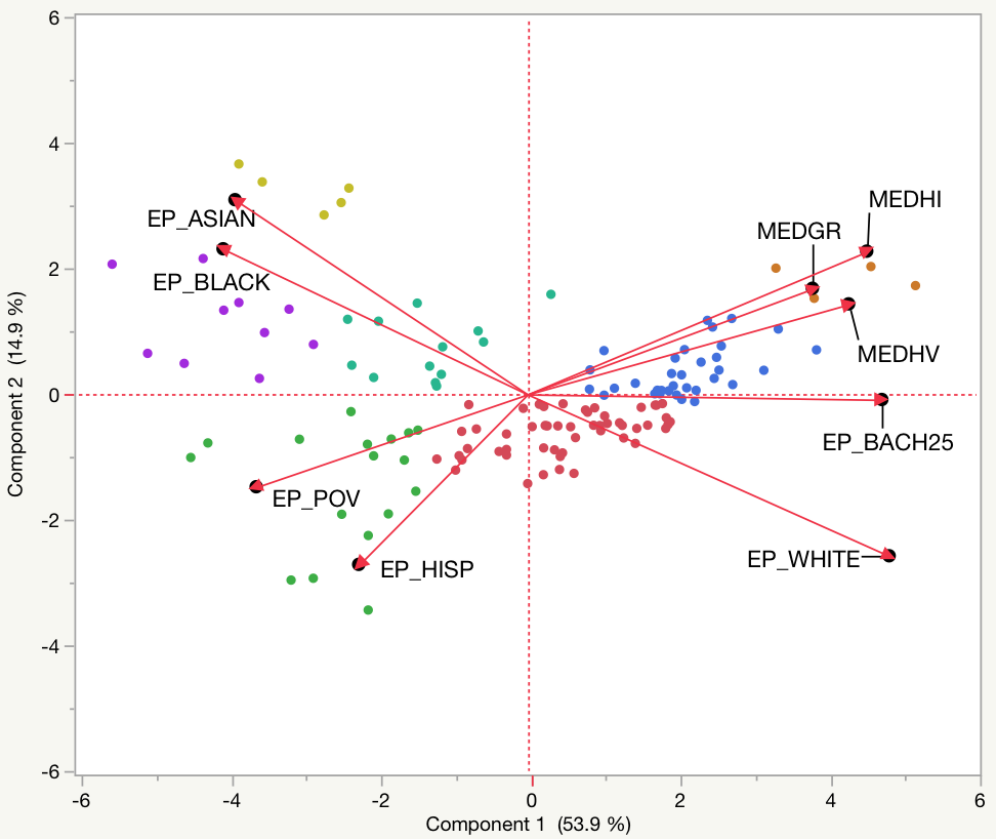



Figure 7: Biplot of 2010 Data

Furthermore, we note that there are 3 clusters that have mostly positive loadings along PC1 (red, dark-blue, and orange), which corresponds to 89 Census tracts. However, in the biplot in Figure 4, we see that there are mainly 2 clusters with positive loadings for PC1 (yellow and purple), corresponding to 76 tracts. This signifies that there has been an increase in tracts displaying higher socioeconomic status between 2000 and 2010, indicating that Seattle houses more wealthy, white, and educated residents in 2010 compared to the rest of the population than it did in 2000. The red, dark-blue, and orange clusters are mainly located in the northern half of the city and in West Seattle. We also note that the clusters associated with lower socioeconomic status and Hispanic populations (green), are mostly

located in the International and Central Districts, as well as in southern regions of the city. The southeastern area of Seattle is home to tracts that are associated with high populations of color, as well as lower-income communities, as indicated by the clusters in green, purple, yellow, and teal-green. We will conclude our analysis of this section by examining the 2017 data set.

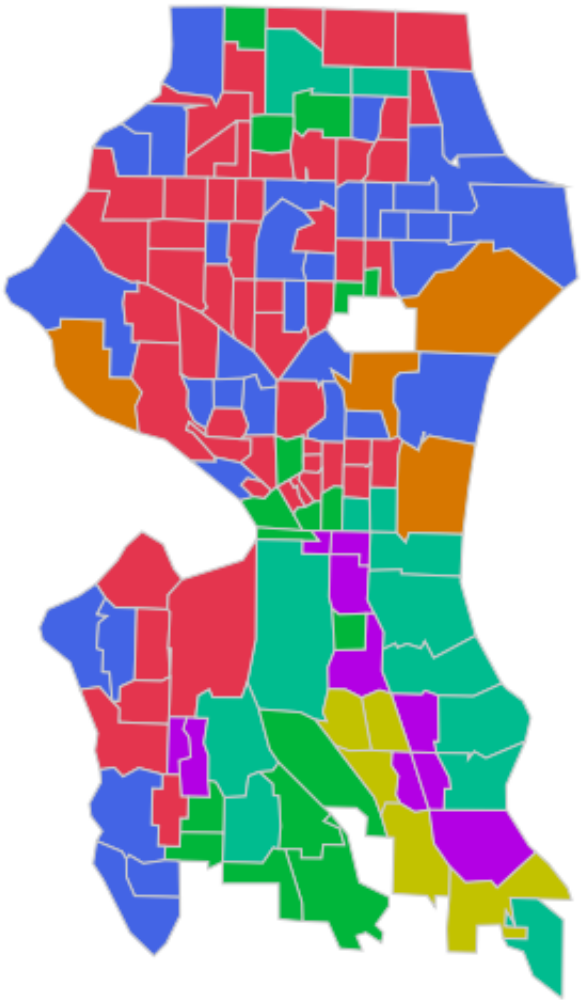


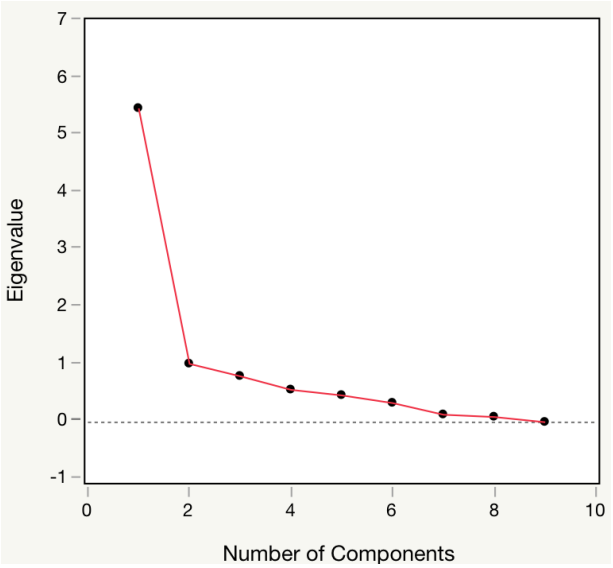
Figure 8: Clustered Map of Principal Component Scores

### 4.1.3 2017 ACS Results

From (a) in Figure 9, we note that PC1 explains 61% of the total variance and PC2 explains 11.5% of the total variance, so together PC1 and PC2 can explain 72.4% of the total variance. The scree plot (b) in Figure 9 suggests that the first 2 principal components are sufficient to summarize the total variance.

| Number | Eigenvalue | Percent | Cum Percent |
|---|---|---|---|
| 1 | 5.4877 | 60.974 | 60.974 |
| 2 | 1.0309 | 11.455 | 72.429 |
| 3 | 0.8147 | 9.052 | 81.482 |
| 4 | 0.5776 | 6.418 | 87.899 |
| 5 | 0.4810 | 5.345 | 93.244 |
| 6 | 0.3470 | 3.855 | 97.099 |
| 7 | 0.1439 | 1.599 | 98.698 |
| 8 | 0.1040 | 1.155 | 99.853 |
| 9 | 0.0132 | 0.147 | 100.000 |

(a)




(b)

Figure 9: Scree Plot of 2017 Data

Examining the loading matrix in Table 5, we observe that MEDHV and MEDGR display very similar positive loadings for PC1, as do EP_WHITE and EP_BACH25. Furthermore, EP_BLACK, EP_ASIAN, and EP_POV exhibit very similar negative loadings.

Table 5: Loading Matrix

| | PC1 | PC2 |
|---|---|---|
| MEDHI | 0.8542042 | -0.21599599 |
| EP_POV | -0.7228413 | -0.11287385 |
| EP_WHITE | 0.9129275 | 0.29605193 |
| EP_BLACK | -0.7563142 | -0.33938235 |
| EP_ASIAN | -0.7348616 | -0.49788465 |
| EP_HISP | -0.5416754 | 0.63590010 |
| MEDHV | 0.7787791 | -0.22804316 |
| MEDGR | 0.7696083 | -0.23417176 |
| EP_BACH25 | 0.8932201 | -0.09812578 |

From the biplot in Figure 10, we see that now we have 4 clusters displaying mainly positive loadings for PC1 (green, teal-blue, purple, and orange), which corresponds to 99 tracts. Hence, there has been an increase in tracts that display characteristics associated with higher socioeconomic status and whiteness relative to the rest of the population since 2010. We also point out that in the 2010 analysis, we observed a stronger correlation between EP_POV

and EP_HISP. However, in the 2017 data set, we note that the EP_HISP arrow points in a different direction than any other variable, while EP_ASIAN, EP_BLACK, and EP_POV are more correlated with one another compared to the previous 2 analyses.

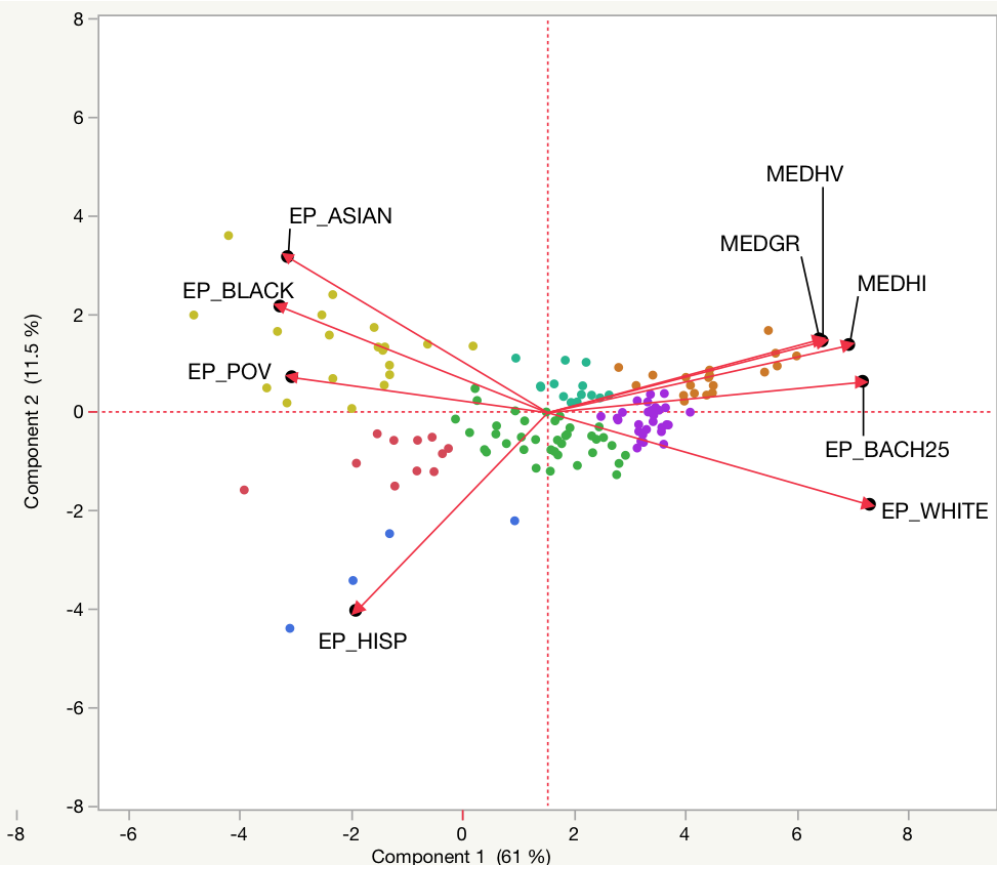



Figure 10: Biplot of 2017 Data

Considering the map in Figure 11, we note that there is an increased prominence of the orange clusters, which represent the wealthiest populations with high positive loadings along PC1 pointing in the direction of MEDGR MEDHV, and MEDHI compared to those in the 2010 data set. If we consider the map of the 2010 data in Figure 8, we note that some of the tracts that were previously associated with EP_POV (the clusters in green), have more positive loadings along PC1 in the 2017 analysis. For example, if we examine the green cluster in Figure 10, we observe that it is situated closer to the center, with a number of the tracts located in the direction of the EP_WHITE arrow. We note that a couple of the previously green tracts in the 2010 data located in the downtown and Central District are represented as green tracts in the 2017 data, suggesting a gradual shift towards the variable EP_WHITE in those areas. We also remark that in the 2010 data set, the teal-blue cluster points toward the variables EP_BLACK and EP_ASIAN, as illustrated in Figure 7. However, in the 2017 biplot, we see that the teal-blue cluster is located to the right of

the center, with mainly positive loadings for PC1 and PC2. Interestingly, we note a shift toward positive PC1 loadings in a number of the tracts south of downtown. For example, the previously teal-blue tract in the Industrial District in Figure 8 has become part of the green cluster in Figure 11, suggesting that this tract is a possible candidate for gentrification. Furthermore, we note that neighborhood Riverview in the Delridge region has transformed from teal-blue to green, indicating that gentrification may have affected this area as well. We note similar transformations taking place in 3 tracts in the southeastern sectors of the city, as neighborhoods that previously housed more Black and Asian communities in 2010 with negative PC1 loadings are gravitating toward more positive loadings in 2017, particularly in Rainier Valley and Columbia City.

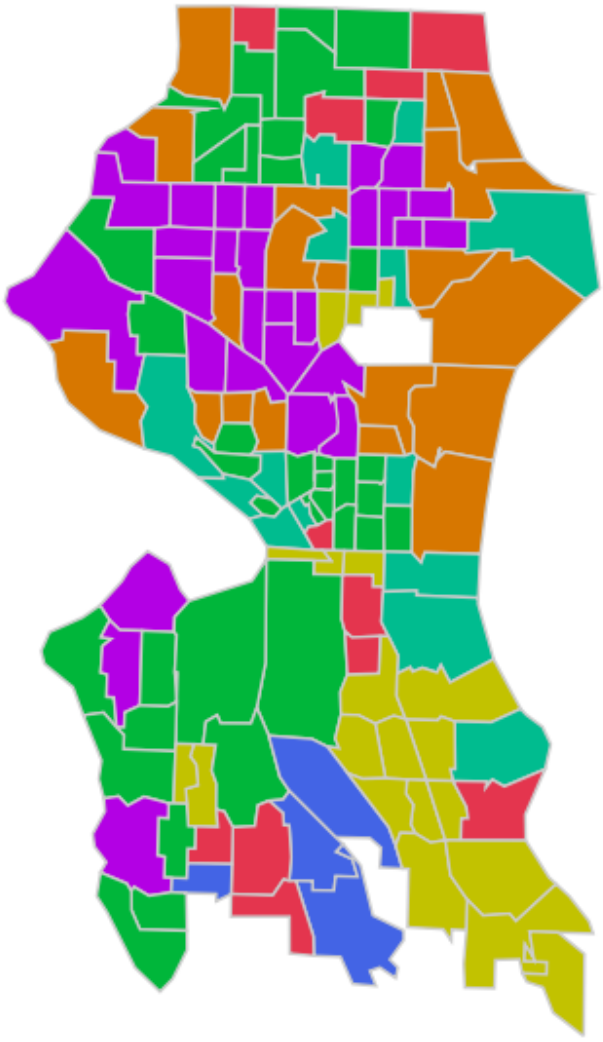


Figure 11: Clustered Map of Principal Component Scores

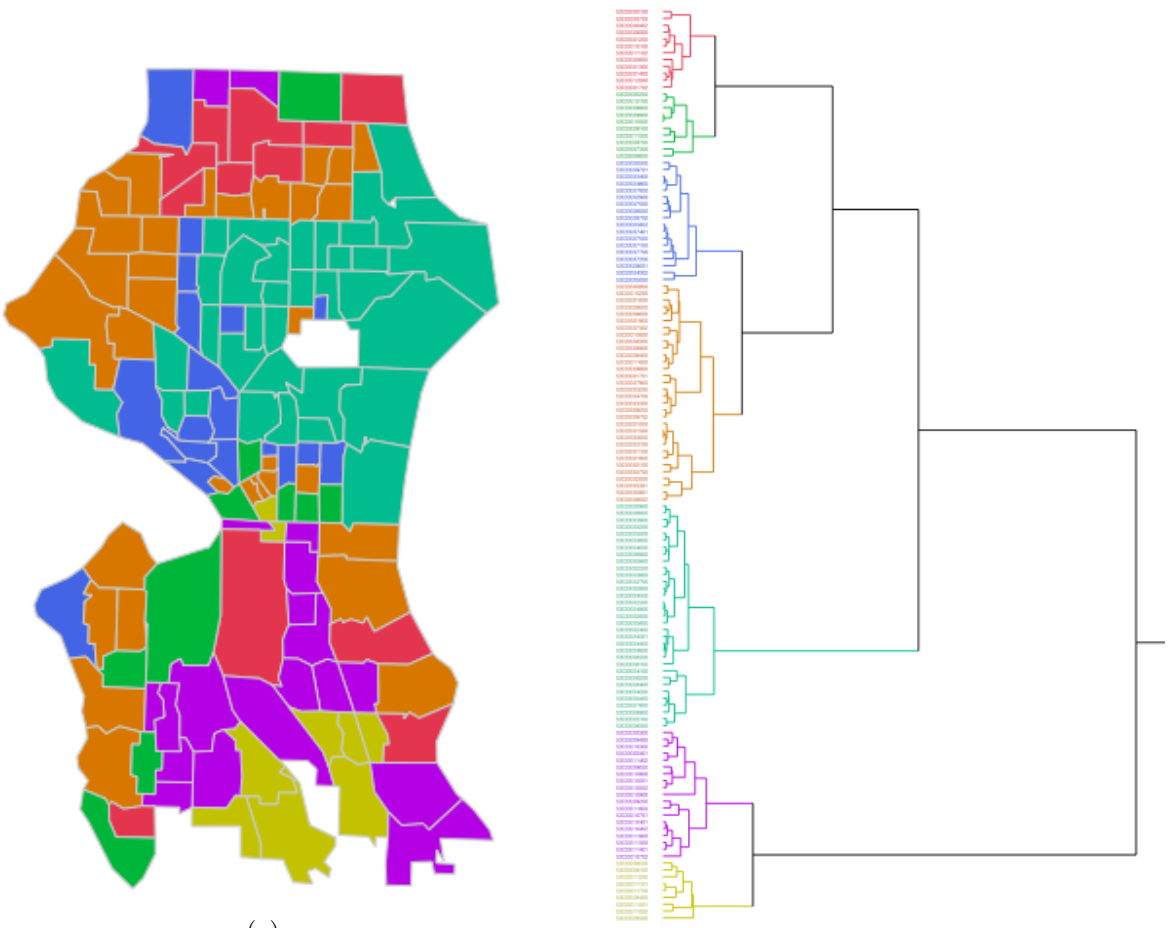
### 4.2 Conclusion

To summarize the results found in Section 4, we noted that there are few tracts representing the most affluent populations with high positive loadings along PC1 in the 2000 data. Instead, the majority of the tracts are located to the right of the center of the biplot, exhibiting low positive loadings for PC1. In 2010, while the clusters were more spread out throughout the city, we observed a notable shift toward higher positive loadings along PC1 for a number of tracts. In 2017, we saw that the percentage of the population living in poverty became more correlated with Asian and Black communities than in either of the previous 2 analyses. Furthermore, we discovered a greater correlation among tracts that exhibit high income, expensive rent and housing prices, highly educated and primarily white populations than in either of the prior years. We also learned that number of tracts that possessed characteristics that are more associated with gentrifying forces (the tracts with positive PC1 loadings) has steadily increased throughout the years, from 76 to 99 between 2000 and 2017. Performing PCA and clustering on our data has allowed us to analyze the 3 separate data sets and compare our observations, but how may we more directly assess neighborhood change over time? We will utilize a combined data set that aggregates all 3 separate years to answer this question, which is described in Section 5.

## 5 Analysis of Combined Data

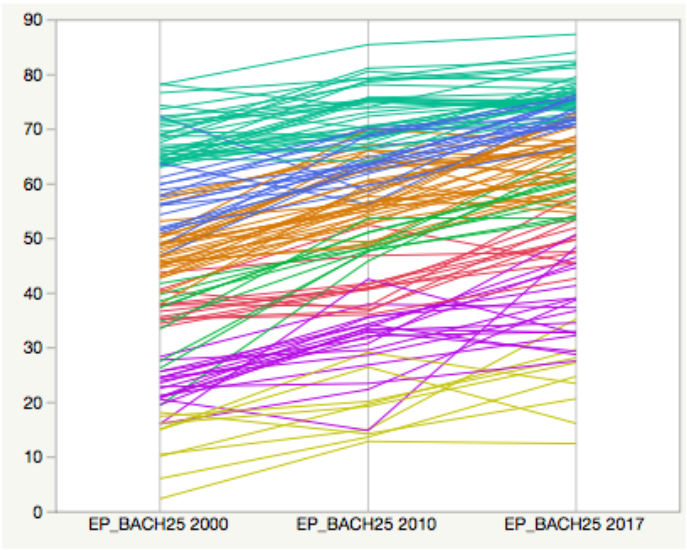
We will employ data visualization techniques to further analyze our data visually on the merged data set, which will allow us to examine the data both graphically and geospatially. In order to better understand variable change over time, we will utilize a resource called *parallel plots*, which enables us to examine different years for a given variable through a single plot. Each line in the parallel plot corresponds to an individual Census tract. We will perform hierarchical clustering on tracts that exhibit similar trends in variable change over time, and plot these clusters geospatially. Parallel plots are useful for analyzing high dimensional data and they allow us to draw conclusions about general trends for a given variable over time, which is very helpful for gentrification research.

We will first examine the parallel plots, maps, and hierarchical clusters for the 4 four variables that demonstrated the most significant changes between 2000 and 2017: 1) percentage of population age 25+ with a bachelor's degree or higher, 2) median household income, 3) median gross rent, and 4) median house value. From the parallel plots (c) in Figures 12, 13, 14, and 15, we note that all of the variables display upward trends over time, which is more apparent in the latter three. For the most part, we note that the parallel plots in Figure 12 and Figure 13 demonstrate more linear trends over time, with a relatively constant rate of change. On the other hand, the majority of the tracts in Figure 14 exhibit a gradual



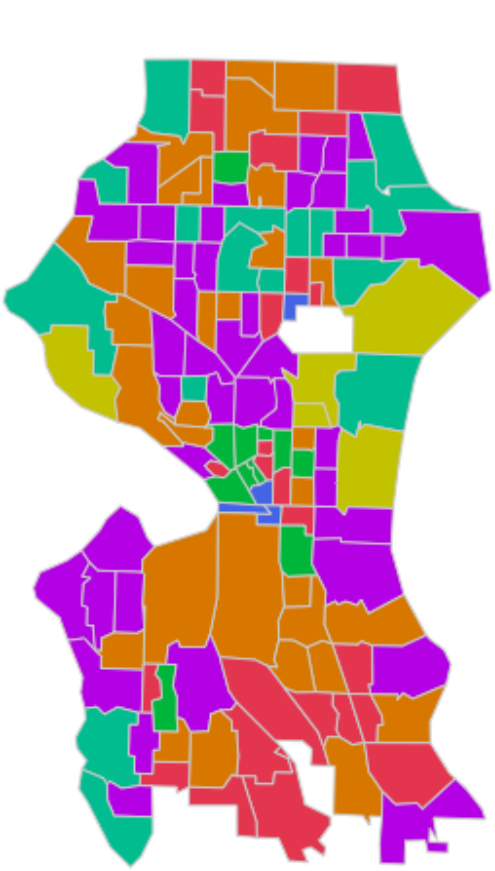
(a)

(b)

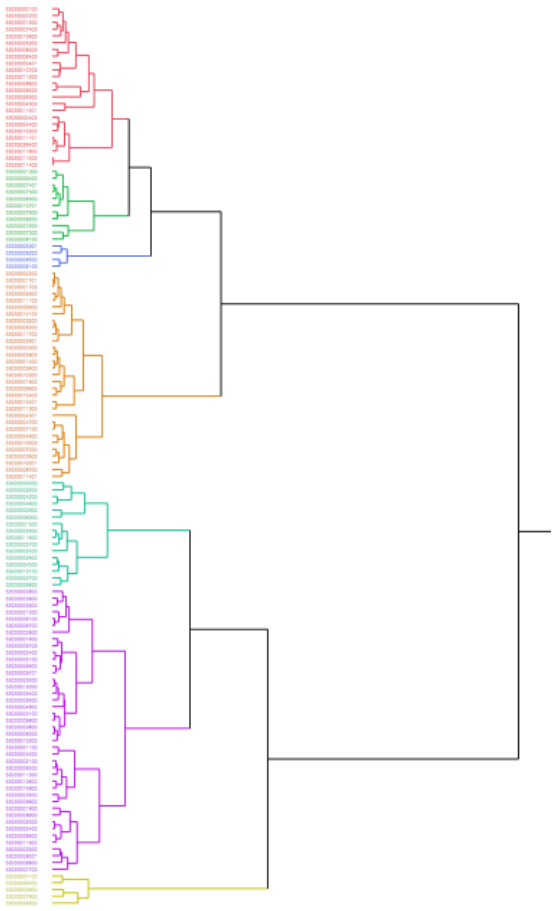



(c)

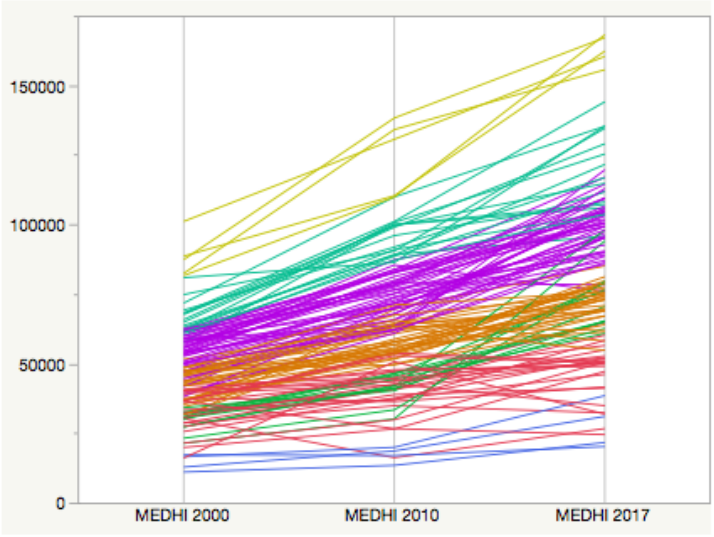
Figure 12: Percentage of Population with a Bachelor's Degree or Higher



(a)



(b)




(c)

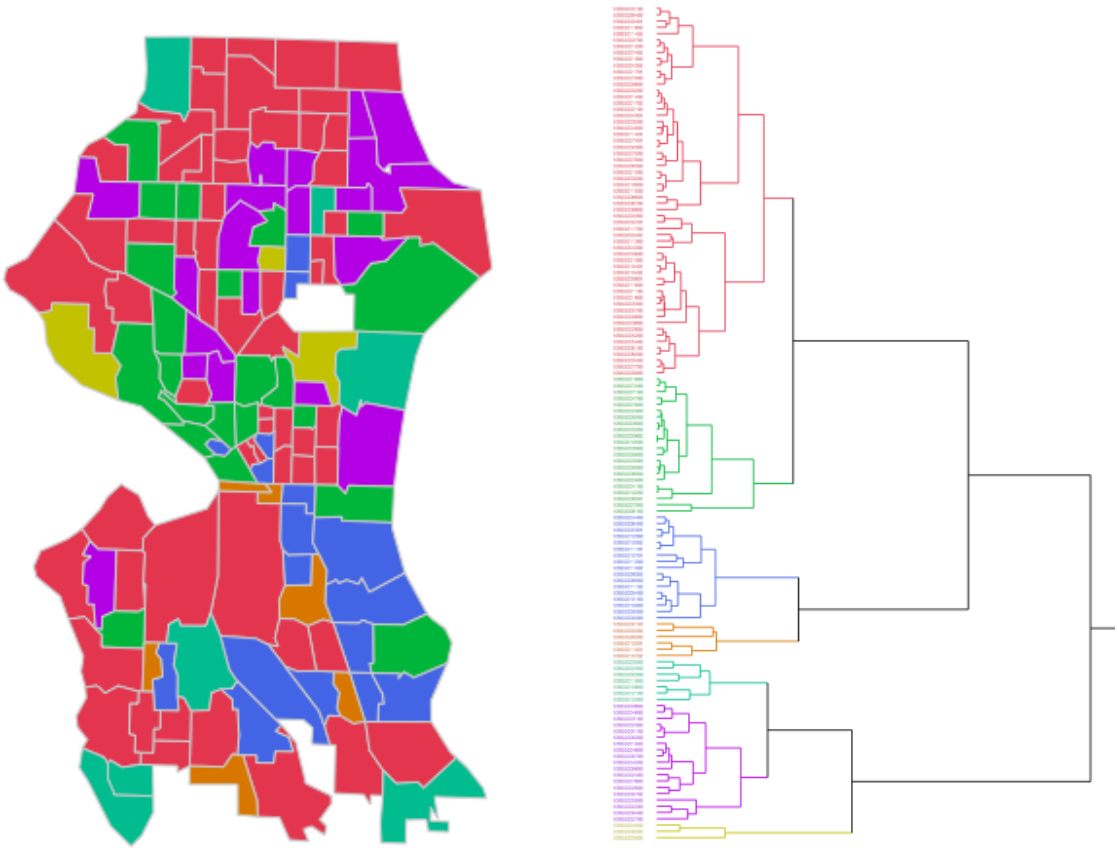
Figure 13: Median Household Income

incline between the years 2000 and 2010, with a notably steeper increase from 2010 to 2017. Furthermore, the bulk of the tracts in Figure 15 reveal a more prominent incline from 2000 to 2010, with a comparatively gradual increase from 2010 to 2017. The sharp increase in median gross rent could be explained by the apartment boom that has hit Seattle in recent years, achieving a record number of occupied units in the year 2015 [11].

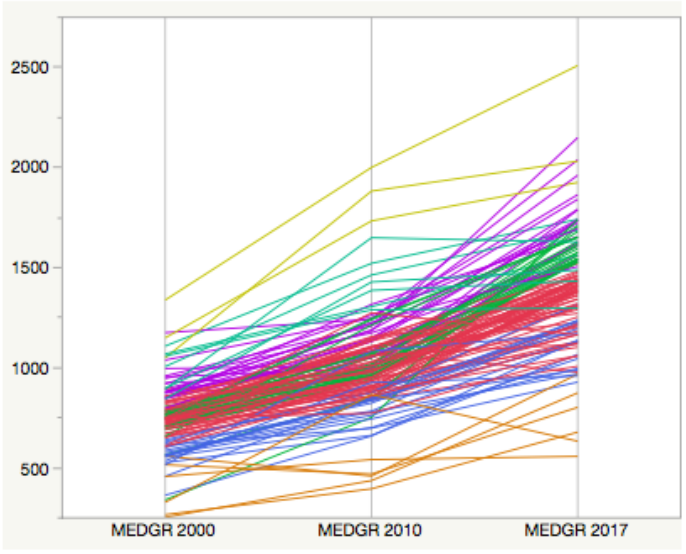
As illustrated in the figures, we can produce maps of the 4 different variables to compare trends in neighborhood change. In Figure 12, we note that the most educated populations are located primarily in the northeastern regions of the city, as indicated by the teal-blue cluster. On the other hand, the least educated tracts are established in the far southern regions, which corresponds to the neighborhoods Delridge, Highland Park, South Park, and South Beacon Hill. We also note that a couple of these tracts are located in the Central and International districts.

In Figure 13, we note that the highest income populations (indicated by the yellow and teal-blue clusters) are primarily located near the coastal regions of the city, which is consistent with the fact that those neighborhoods tend to have a higher cost of living. While the majority of the clusters exhibit rather linear trends from 2000 to 2017, we observe that the clusters indicated in green experienced relatively low median household incomes between 2000 and 2010, with a steep growth from 2010 to 2017. This augmentation suggests that those tracts experienced notable gentrification, because previously lower income communities were replaced by significantly wealthier populations. We note that these populations are located primarily clustered together in downtown, East Queen Anne, South Lake Union, Broadway, and the Central District, with a few outliers in North Beacon Hill, west Delridge, and north Green Lake. The lowest income communities (the dark blue tracts) are grouped together south of downtown with another tract located in the University District.

In Figure 14, we note that the highest rent neighborhoods are spread out across the map, located in Magnolia, Montlake, and Green Lake. The clusters in purple experienced a notable rise in median gross rent, particularly between 2010 and 2017. The majority of these tracts are located north of downtown, primarily in Lake City, Phinney Ridge, North Green Lake, Queen Anne, with some outliers in West Seattle, and the Central District. Furthermore, the clusters in red, dark-blue, and green have faced a relatively steady increase in cost of rent over the years. We observe that these tracts are clustered together throughout the city. The green tracts are mainly situated in the downtown and Queen Anne neighborhoods, with a few scattered areas on the eastern coast of the waterfront. The red tracts are primarily located in the northern regions of Seattle, with a cluster of tracts east of downtown and another grouping in the International District, West Seattle, and Delridge regions. The dark blue clusters, which have a lower median gross rent than the red tracts, are mainly located

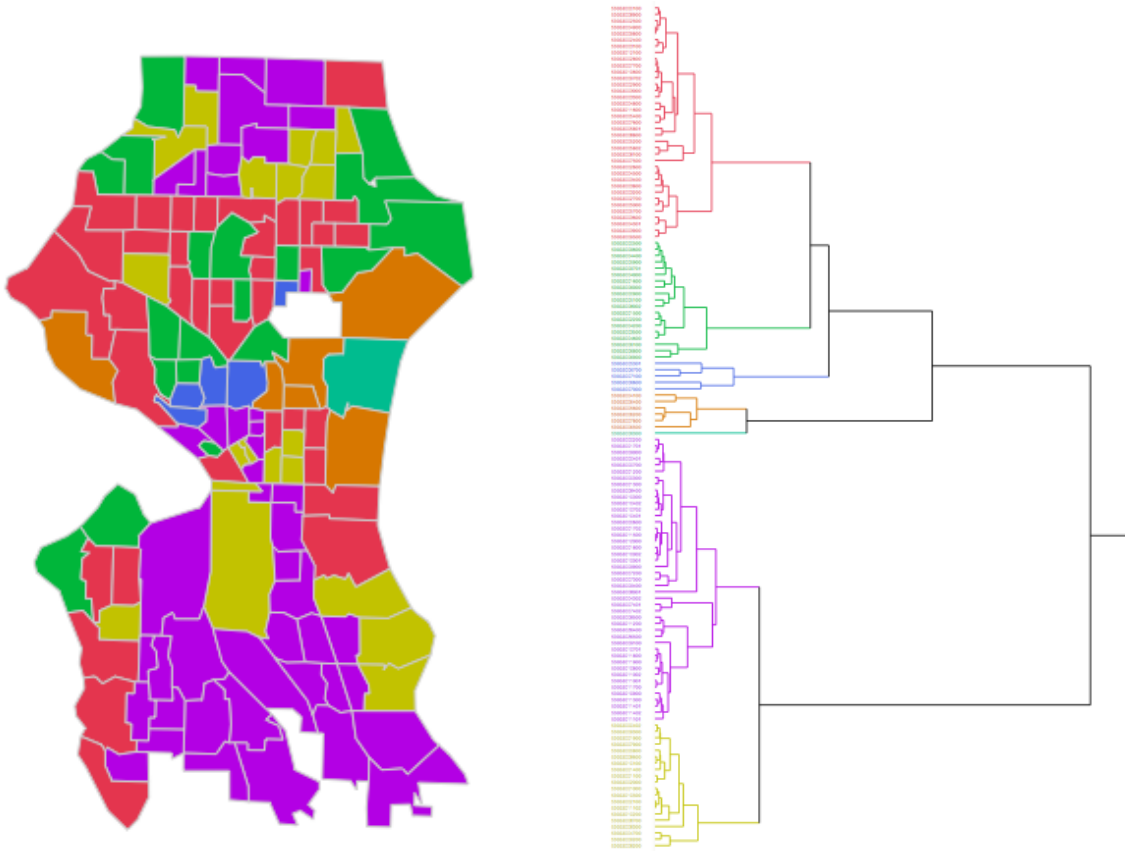


(a) (b)

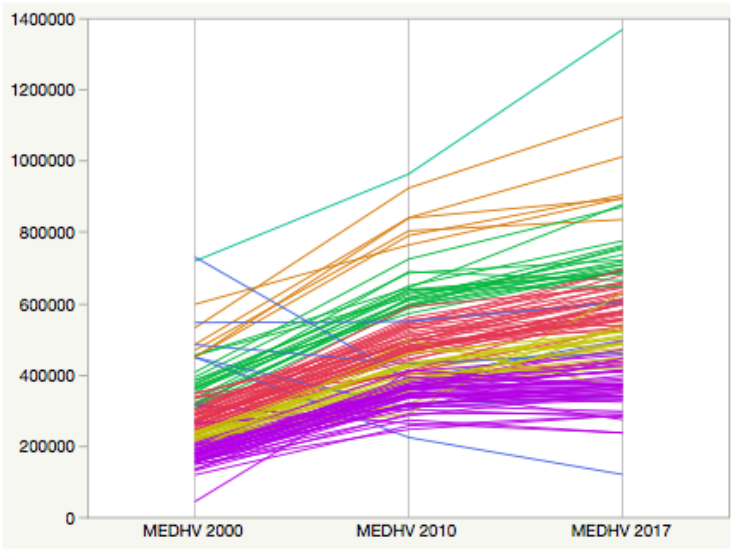



(c)

Figure 14: Median Gross Rent



(a) (b)



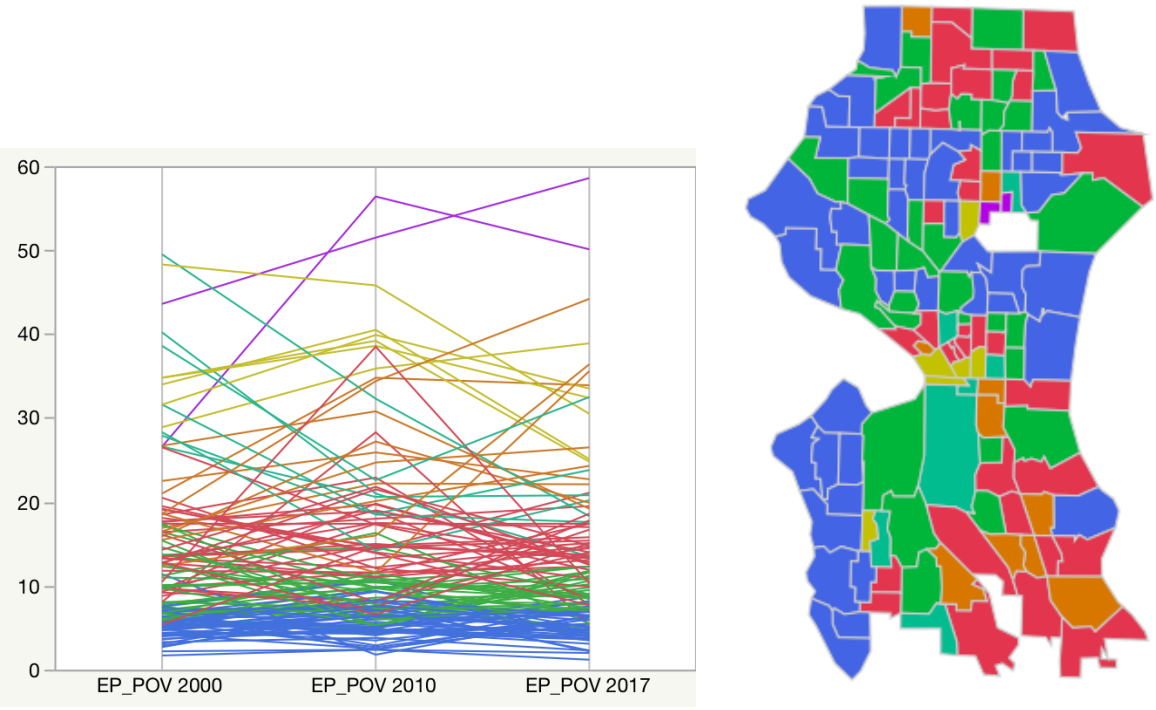

(c)

Figure 15: Median House Value

in the southeastern sectors of the city, in the Mount Baker and Rainier Valley neighborhoods.

Finally, we will interpret the median house value variable, illustrated in Figure 15. We note that the orange cluster has experienced a dramatic increase in house value from 2000 to 2010, leveling out between 2010 and 2017. These tracts are located in Magnolia, Laurelhearst, Windermere, Montlake, and Capitol Hill. We note that the dark blue cluster has faced a notable decline in house value between the years. This trend could be explained by the fact that these tracts are mostly located in the lower Queen Anne and South Lake Union neighborhoods, which have experienced a boom in the construction of apartment buildings in recent years. We also observe that the nearly all of the purple tracts are grouped together, which represent the lowest house value bracket apart from the dark blue clusters. These tracts are mostly located in the non-waterfront neighborhoods south of downtown, and the non-water regions in north Seattle. Additionally, the red tracts are mainly clumped together in the Ballard, Green Lake, Roosevelt, and Ravenna neighborhoods, with other clusters in the southwestern part of the city and Mount Baker.

We will proceed by visually analyzing how the remaining variables have changed over time. For the following variables, the clusters displayed on the maps were established through the same hierarchical clustering methods illustrated in the previous figures, but the dendrograms are not included. In Figure 16, we observe that the tracts exhibiting the lowest percentages of impoverished communities (dark-blue, green, and red clusters), are not surprisingly located in the northern half of the city and in West Seattle. We note that the purple clusters neighbor the University of Washington, which would explain the high rates of poverty in these areas due to large student populations. The yellow and teal-blue clusters demonstrate decreased rates of poverty from 2000 to 2017, suggesting that these tracts may have undergone gentrification processes. The yellow tracts are primarily located downtown, in Pioneer Square, the International District, with an isolated tract in west Delridge. The teal-blue clusters are mainly found in the southern half of the city, in the Industrial District, Central Area, and South Delridge.

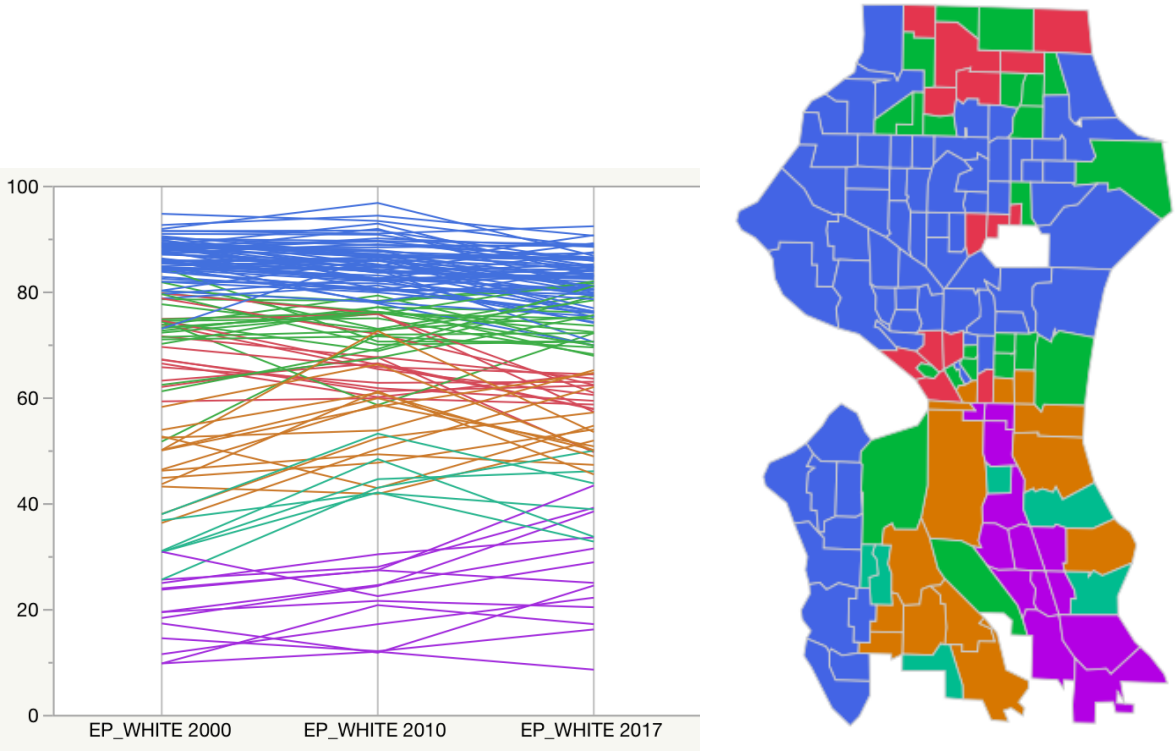



(a) (b)

Figure 16: Percentage of Total Population Living in Poverty

As mentioned previously, the geospatial spread of white populations and communities of color is an important consideration in gentrification studies. We will proceed by analyzing patterns in racial segregation and migration for the 4 race demographic variables selected: EP_WHITE, EP_BLACK, EP_ASIAN, and EP_HISP.

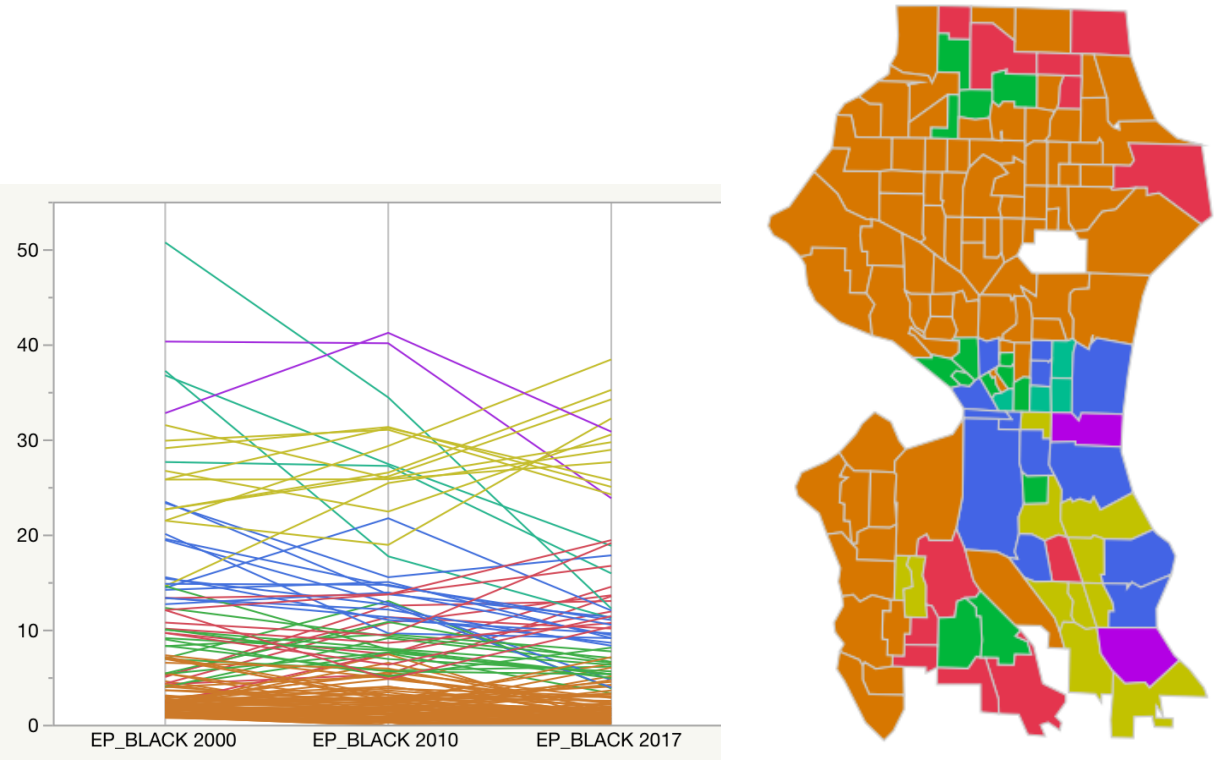
Considering Figure 17, we find that the whitest populations, represented by the dark-blue, green, and red tracts, are almost exclusively located in West Seattle and the neighborhoods north of downtown. Interestingly, these tracts correspond to those in Figure 16 with the lowest rates of poverty. On the other hand, the purple clusters that house more communities of color are mainly grouped together in South Seattle neighborhoods, such as the Central Area, Beacon Hill, Columbia City, and Rainier Beach. However, we observe that these tracts are becoming associated with whiter populations, as illustrated by the upward trend in the purple cluster between 2010 and 2017. We also comment that the teal-blue cluster demonstrates a notable increase in white populations between 2000 and 2010, and a slight decline in this trend between 2010 and 2017. This suggests that the effects of gentrification may have been more severe during 2000-2010 in the neighborhoods Mount Baker, Dunlap, High Point, and Highland Park.




(a) (b)

Figure 17: Percentage of Total Population White Alone

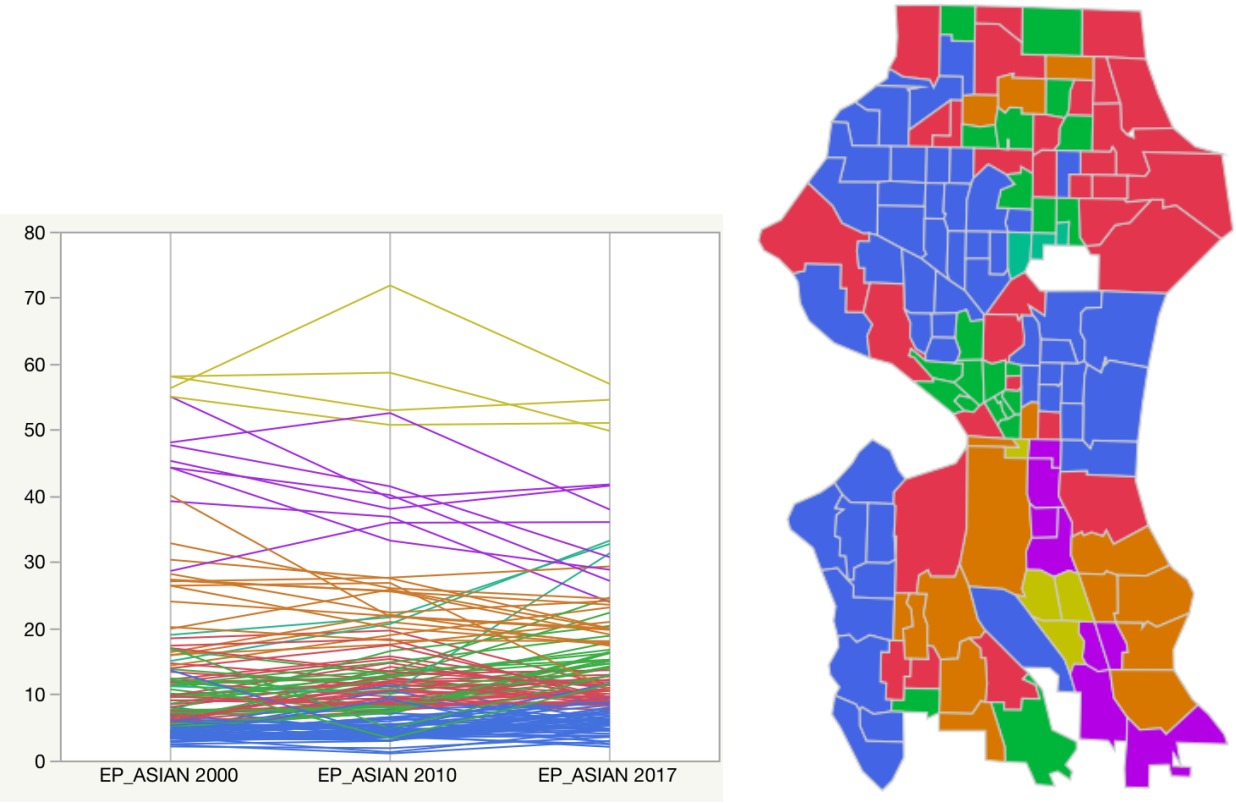
Figure 18 illustrates an inverse relationship between Black and White communities, since the lowest Black populations are mainly located in the tracts that demonstrate high percentages of White residents. Furthermore, we notice a significant decline in some tracts that previously exhibited high rates of Black populations, as shown by the teal-blue cluster. The tracts are mostly located east of downtown in the Central District, and in Yesler Terrace. Furthermore, we observe that the 2 purple tracts have experienced a decline in black populations between 2010 2017, which correspond to the neighborhoods Mount Baker and Rainier Beach. We also observe that the dark-blue cluster illustrates a subtle decline in Black populations between 2000 and 2017, which has occurred in Pioneer Square, the Industrial and International Districts, Beacon Hill, Madrona, Dunlap, and Seward Park.




(a) (b)

Figure 18: Percentage of Total Population Black Alone

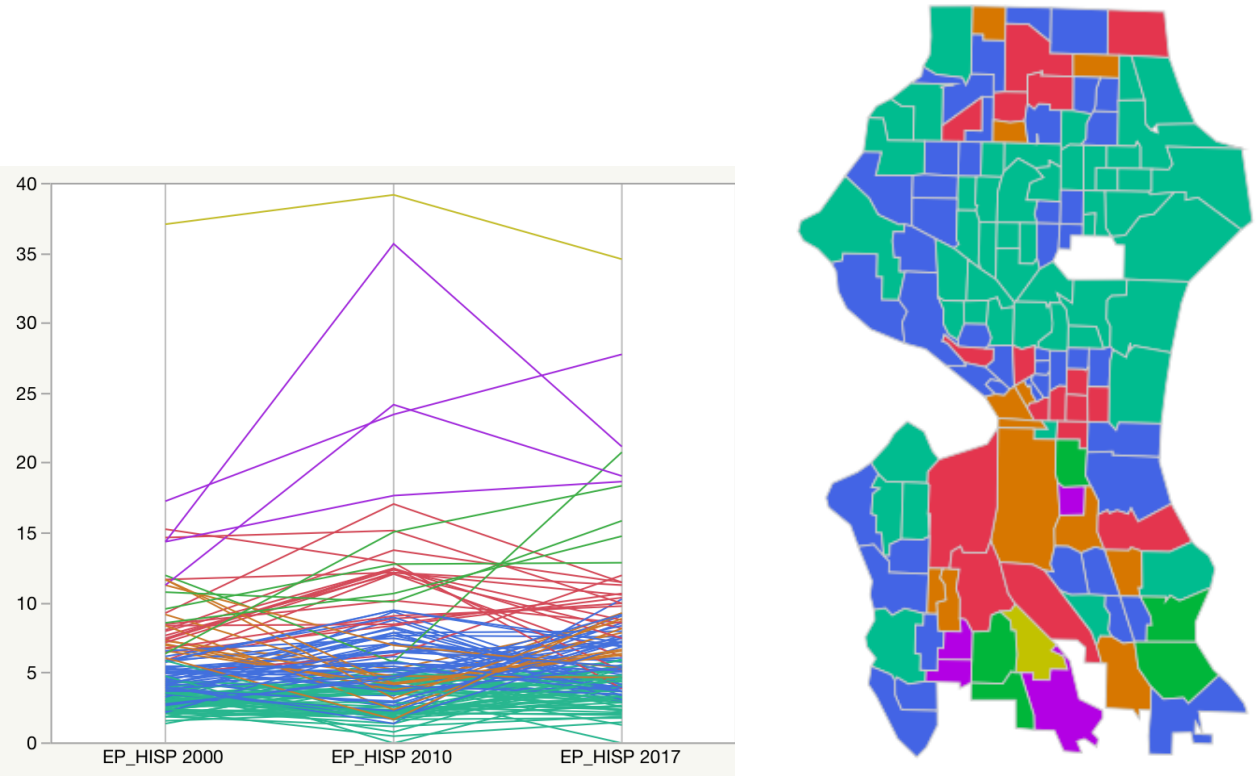
In Figure 19, we note that neighborhoods representing the lowest Asian populations (the dark-blue, red, and green clusters) are mostly located in tracts that display very high percentages of White residents. The tracts that represent high Asian populations are found in Delridge, Beacon Hill, and Rainier Valley. However, even the highest numbers of Asian populations have seen an overall decline from 2000 to 2017, as demonstrated by the yellow and purple clusters. Some of these clusters overlap with ones in Figure 18 that exhibit a decline in Black populations, particularly in the Beacon Hill area.




(a) (b)

Figure 19: Percentage of Total Population Asian Alone

We conclude our analysis of the combined data by analyzing the final variable, the percentage of the total population that is Hispanic or Latino. In Figure 20, we observe that tracts representing low Hispanic/Latino populations (teal-blue, orange, and dark-blue clusters) more spread throughout the city compared to the Black and Asian variables. We are not surprised to see that the lowest Hispanic and Latino tracts are found in the northern and southwestern regions of Seattle, but these neighborhoods are also located south of downtown as well. Rather, it appears that tracts housing high Hispanic and Latino populations are distinct from the other racial groups, suggesting racial segregation within Seattle neighborhoods. The yellow cluster symbolizing the largest Hispanic/Latino population in the city only contains 1 tract, which is located in the neighborhood South Park. The purple cluster also represents substantial Latino/Hispanic communities, which corresponds to tracts located in Delridge and Beacon Hill.




(a) (b)

Figure 20: Percentage of Total Population Hispanic or Latino

# 6 Conclusions and Future Research

We will return to the 3 research questions that we posed in Section 1.3.

1. What is the nature of gentrification in Seattle and what neighborhoods and regions in the city have experienced the greatest effects of the phenomenon?

2. How does socioeconomic status affect diversity and segregation of Seattle residents?

3. Can we detect patterns of displacement and urban migration due to gentrification in the city of Seattle?

To reflect on these questions, we conclude that the processes of gentrification in Seattle have been more severe in recent years, particularly between 2010 and 2017. We also note that Seattle is moving toward becoming a whiter, wealthier, and more educated city. We see that communities of color are being pushed further south in the city, which highlights a clear racial divide in urban spaces. Our analysis suggests that the effects of gentrification are most notable in the Central Area, downtown, the Industrial and International Districts, Delridge,

and Beacon Hill. We also found that the lowest poverty regions are overwhelmingly white, suggesting a significant segregation of Seattle residents by race and socioeconomic status. While patterns of displacement are difficult to trace, we conclude that the effects of it do occur in Seattle, which is illustrated by the fact that even high diversity neighborhoods are becoming whiter and more expensive in recent years.

We will briefly discuss the challenges and limitations of unsupervised learning techniques, such as the ones implicated in this study. On one hand, supervised learning is a well-understood field, with a developed set of tools, in addition to a clear understanding of how to assess the quality of the results found. However, interpreting unsupervised learning results often poses more challenges. The method tends to be more subjective, since there is no straightforward goal for the analysis, such as prediction of a response. Additionally, it can be difficult to evaluate the results obtained from unsupervised learning methods, since there is no broadly accepted procedure for validating results on an independent data set. This is due to the fact that when we wish to fit a predictive model using a supervised learning technique, then it is possible to confirm our results by examining how well our model predicts the response $Y$ on observations or variables not used in fitting the model. However, in unsupervised learning, there is no way of checking our work since we do not have knowledge of the true answer because the problem is unsupervised [4].

This project serves as a first attempt to understand the nature of neighborhood change and equitable urban environments in Seattle. We note that all data sources have limitations and that the findings in this study can inform, but should not determine the comprehensive nature of processes of gentrification in Seattle. Rather, broad historical and qualitative context is necessary to avoid drawing simplistic conclusions. However, we hope that this project can help guide future studies and research on this topic, and help us better understand the underlying forces behind important social issues.

## 6.1 Future Research

If we were to conduct a similar project in the future, it would be interesting to generalize this study to other cities in order to compare the effects of gentrification. We could also develop predictive modeling in an attempt to predict which regions will be most affected by gentrification in the coming years. Furthermore, we could analyze a smaller Census geography level such as *block groups*, in order to capture more detailed shifts in neighborhood change. Finally, it would also be informative to include additional variables in our analysis, such as immigration status and travel time to work.

# References


[1] "2010 Census Tracts and Zip Code Boundaries." *Seattle.gov*, Office of Planning Community Development.

[2] "Decennial Census and the American Community Survey (ACS)." *United States Census Burear*, 5 Sept. 2017.

[3] "History of the Decennial Census." *United States Census Bureau*, 2 Apr. 2018.

[4] James, Gareth, et al. *An Introduction to Statistical Learning with Applications in R.* Springer, 2017.

[5] Johnson, Richard A., and Dean W. Wichern. *Applied Multivariate Statistical Analysis.* Pearson, 2007.

[6] Jolliffe, Ian T., and Jorge Cadima. "Principal component analysis: a review and recent developments." Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374.2065, 2016.

[7] Lay, David C. "Linear Algebra and its Applications, Fourth Edition." Pearson, 2012.

[8] Martin, Richard W. "A Quantitative Approach to Gentrification: Determinants of Gentrification in U.S. Cities, 1970-2010." *Department of Insurance, Legal, Studies, and Real Estate, Terry College of Business, University of Georgia.*

[9] "Neighborhood Areas Census Block Groups (2000)." *City Clerk's Office*, Department of Neighborhoods, City of Seattle, 4 Mar. 2004.

[10] Raiche, G. "Critical Eigenvalue Sizes (Variances) in Standardized Residual Principal Components Analysis. Rasch Measurement Transactions. 2005; 19–1: 1012" 2014.

[11] Rosenberg, Mike. "Renter Boom: Apartments Filling up Faster in Seattle Area than Anywhere in the U.S." *The Seattle Times*, The Seattle Times Company, 20 Apr. 2019.

[12] "Seattle's Population Growth Leads Nation." *KING 5 News*, 25 May 2017.

[13] Smith, Lindsay I. *A tutorial on principal components analysis*, 2002.

[14] "Understanding Geographic Relationships: Counties, Places, Tracts and More." *United States Census Burear*, 31 July 2014.

[15] "When to Use 1-Year, 3-Year, or 5-Year Estimates." *United States Census Burear*, 6 Sept. 2018.

[16] White, Jonah D. "Obscured geographies of the Emerald City: a study on gentrification in Seattle, WA." *WWU Graduate School Collection*: 191, 2012.

[17] Zuk, Miriam, et al. "Gentrification, displacement, and the role of public investment." *Journal of Planning Literature* 33.1 (2018): 31-44.