Modelling Homeless Population in Los Angeles

Ryan Du, Xiang Yang Ng, Hannah Ross, Sanjay Shukla

Math 199, Spring 2018
Advisor: Michael Lindstrom

Abstract

Homelessness is a prevalent issue in the Los Angeles area, and it is an issue that touches many different people in society. As such, we want to understand the underlying features that would explain homelesssness in the area. In this exploratory analysis, we used topic modeling, local variance analysis, and cluster analysis to identify the city features that correlate with homeless population count. We used statistical models and neural networks to find factors that contribute to the rise and decline of homeless populations and to build models that predict the change in homeless population. We also used Earth Mover's distance to analyze the nature at which homeless populations move within Los Angeles.

Contents

1 Introduction

In order to identify why homelessness is so rampant in certain areas of Los Angeles, it is vital to investigate and compare the features that constitute locations of high and low homeless populations. By analyzing these features, we can construct a clearer idea as to why certain areas are more susceptible to homelessness than others. In addition to the effects of certain location-specific features, there are also a variety of forces that play a role in the movement of homeless individuals including planned physical and social change programs such as urban renewal, slum clearance, highway construction, 'Greyhound therapy', and police sweeps of encampments[1]. In understanding the fluctuations and patterns of homelessness, we will be in a better position to inform impactful and efficient uses of resources that are being directed towards tending to the homeless population.

The homeless migrant population that we intend to analyze includes a variety of unique sub-populations, like homeless young adults and homeless veterans. Previous research surrounding these groups offer insights into their own specific causes for homelessness and motivations for moving. In acknowledging these sub-populations of the greater homeless population, we want to recognize dimensionalities of homelessness that constitute the larger categories that our study chooses to discuss, which are namely homeless people living on the street, homeless people living in cars, and homeless people living in homeless shelters.

In one investigation into the correlates of homelessness in adolescents, 37 percent of homeless young adults are reported to of left home due to parent disapproval, 33 percent due to parental sexual abuse, and 51 percent were thrown out by their parents. Over 47 percent of homeless adolescents report a history of sexual abuse with parents, and nearly 37 percent self-identify as LGBTQ in orientation [2].

These statistics remind us that the causes of homelessness can stretch beyond high costs of living or lack of employment. They serve as evidence that homelessness is the culmination of both economic and sociological circumstances. Furthermore, HYA were identified as a subset of the population that were most receptive to care provided by agencies[3]. This motivates our systematic study of homelessness which can serve to better inform agency initiatives that serve the homeless young adult population.

The homeless veterans of Los Angeles are another significant sub-population of the greater Los Angeles homeless population. In interviews conducted by the American Journal of Public Health with 33 chronically and 26 acutely homeless veterans receiving transitional housing services in LA from 2003 to 2005, both groups had substantial physical, psychiatric, and social impairment. Their answers to interview questions revealed how health and substance abuse interacted with loss of support and eviction to exacerbate homelessness. In understanding this sub-group of the homeless population, it reminds us to consider alternative factors that will need to be treated and addressed in order to ammend the chronic issue of homelessness.

Parts of our research focuses on identifying the optimal transport of homeless people year to year. A study into the migration of veterans who received homeless services from the Department of Veterans Affairs analyzed the migration patterns for 113,000 homeless veterans who initiated use of the VA homeless services in 2011 or 2013. The study's model showed an estimated 0.9 percent gain in homeless veterans for every 10 degrees of average temperature gain [4]. Interestingly, in-migration tended to roughly balance out-migration in a region. Our research intends to further explore alternative explanations for how homeless people relocate to new locations by leveraging distance as a cost for moving between locations.

This paper is organized as follows: Section 2 explains how the data was obtained and organized for the research work; Section 3 discusses the techniques and models that we used in our analysis of the data; Section 4 discusses the results and findings from the analyses; finally work is concluded

in Section 5 where summary and directions for future research are discussed.

2 Data Management

The data was collected by the previous groups of students that were working on this project from fall of 2017 [5]. So the steps for collecting the data, except for information on the libraries and the stores and ACS PUMA(which will be described later), is summarized for the rest of this section. The main data sources that they collected are Los Angeles Homeless Services Authority (LAHSA) which collects homelessness data, Zillow website which collects data on housing rent and home value, DataLA which provides analytics on Los Angeles, the Los Angeles County Metropolitan Transportation Authority which collects transportation-related data, and the 2016 American Community Surveys which conducts surveys to obtain information about the people and economy.

Note that they wanted data on census track form, instead of data in zip code form, since census track-level data is more granular than zip code-level data, and thus provide better analysis. But some of the data that they collected are not presented in census track form and thus some ingenuity is required to convert those data.

Homelessness data, which are data that concerns us the most, were collected from LAHSA. The data are the estimates of homeless population by census track in the years 2015, 2016 and 2017. Not only that, some subcategories of the homeless population data were also collected, such as the unsheltered and sheltered population, as well a more specific subcategories such as the number of people living in cars, vans, tents and makeshift shelters.

In order to define the boundaries for the census tracks, they obtained shapefiles from the Los Angeles Geohub, which is a website that provides location-based open data.

In order to understand how rent and home value of a census track would influence homelessness, data were collected from Zillow, such as the Zillow Rent Index(ZRI) and the Zillow Home Value Index(ZHVI). The original data format was given in zip codes and in monthly form, which they averaged for 1-year period. After obtaining the data, they locally averaged over zip code to convert them into census track form.

Some information on features such as restaurants, overnight parking citations and crime, total populations for 2015 are obtained from the DataLA. Information on affordable housing units was also obtained from DataLA. So first they found the street addresses for all these features. After that, they converted the addresses into longitude and latitude using Google Maps Geocoding API using Python. Subsequently, the coordinates were then converted to California State Planar feet using MATLAB code "SP Proj" [6]. Finally, using the aforementioned LA shapefiles, they assigned the California State Planar feet coordinates into a census track if it falls within the census track. The whole procedure thus provides the count for each of the features in all of the census track.

Using Google Places API, street addresses of coffee shops were obtained, while bus stop street addresses came from the Los Angeles County Metropolitan Transportation Authority. Additionally, latitude and longitude of homeless shelters were obtained from Los Angeles Geohub. They then assigned the street addresses and latitude and longitude points to census tracts using a similar process as above.

They also obtained five-year estimates of aggregated personal and economic data such as the median household incomes, poverty rates and total population per census tract as of 2014, 2015, and 2016 from the American Community Survey(ACS). Additionally, Public Use Microdata Area (PUMA) data, such as median family income and percentage of population who has active duty military at one point but no longer, were also collected from ACS. The PUMAs are special non-overlapping areas that would collectively and contiguously cover each state. Since the data is in

geodatabase format, some processing is needed to convert it into spreadsheet. Thus they used QGIS, which is a free and open-sourced geographic information system (GIS) application, to conduct the conversion.

For information on grocery store locations, we, the current group of students working on this project, used the address locations listed on YellowPages.com for all Ralphs and Trader Joes locations in the greater Los Angeles area. The street addresses from the site were web scraped using the open source SelectorGadget tool to extract CSS features from the web site, and using the rvest library in R Studio to integrate the information into data objects. The web-scraped street address locations were then converted into latitude and longitude geocode locations using the ggmap library in R studio, and the Google Maps Geocoding API.

The distances to the nearest Ralphs and nearest Trader Joes were calculated for each census tract by finding the store locations with the minimum haversine distance to the centroid of each census tract. This was computed using the haversine distance function from the geosphere library for R studio.

Similarly, the information for the proximity of each census tract to the nearest public library was obtained by using the latitude and longitude locations for all public library locations available on DataLA. Public library locations data were available on DataLA as a dataset with observations for each public library listed in Los Angeles. The street addresses of these public libraries were converted into latitude/longitude points with the Google Maps Geocoding API. We identified the minimum haversine distances from the centroid of each census tract to each public library location.

The data collected are not without their flaws. Some data, such as the restaurant, crime data came from a citywide database, while other data, like LAHSA data, came from a countywide database. In order to conduct a uniform analysis on the given dataset with many more features, some analyses, such as neural networks or topic modelling, are done only on city data. But when they tried to limit the data only from city, some centroid of the census tracks would fall outside of the city boundary, which creates erroneous distribution of data, especially counts of features from DataLa. Data obtained from Google Places or Yellow Pages might not be comprehensive. Some data, notably from DataLA, are also processed, which might induce some inconsistencies and errors. Moreover, some census tracks might also not have the data desired. So interpolation or local averaging is done from the neighbouring census tracks. Finally, for the LAHSA data, there were some coding error done from University of Southern California(USC) that caused an overestimation of the unsheltered youth count in 2017, which might contribute to senseless inflation of total homeless population in that year [7].

3 Techniques and Models

3.1 Percentile Rank Normalization

Los Angeles had certain features that exhibited huge disparities among census tracts such as Median Household Incomes and Zillow Home Value and Rent Index. In order to account for these large disparities, we normalized the data according to percentile ranks using MATLAB's `tiedrank` function [8]. First, suppose we have a data matrix, $D \in \mathbb{R}^{nxm}$ where the rows represent the census tracts and the columns represents the features associated with the census tracks. Then, we rank all the elements in each column from 1 to the number of observations in that column. If some elements have the same value, then they have the average rank. For example, suppose our column has 3 elements, which are 1,1 and 100, they will be ranked as 1.5, 1.5 and 3 by the tiedrank function. Denote the census track i for feature j converted to its percentile rank be entries, R^{ij}, where $i \in \{1, 2, ..., n\}$ and $j \in \{1, 2, ..., m\}$. We'll divide all of the entries by the number of census tracks,

such that the new normalized matrix be **A**, where for all i,j,

$$A_{ij} = \frac{R^{ij} - 1}{n - 1} \tag{1}$$

Note that the difference of 1 is done for both the numerator and denominator in order to let the ranking starts from 0.

3.2 Z-score normalization

Some algorithm we run requires the data to be on the same scale to perform well. One method we used to normalize the data is Z-score normalization. Z-score normalization makes each feature of all the entries has 0 mean and 1 standard deviation. For example, for a data matrix X that is of dimension $\mathbb{R}^{m \times n}$ (each row is a data entry (m of them) and each entry has n features), each of the normalized matrix Z's column has 0 mean and 1 standard deviation.

We first calculated the mean of of each column:

$$M_j = \frac{\sum_i X_{ij}}{m} \tag{2}$$

Then we calculated the standard deviation of the column:

$$S_j = \sqrt{\frac{\sum_i (X_{ij} - M_j)^2}{m - 1}} \tag{3}$$

The Z-score of the data in a column would be:

$$Z_{ij} = \frac{X_{ij} - M_j}{S_j} \tag{4}$$

We do this calculation for each column in our matrix X, we can get the Z-score normalized data matrix Z, where the data in each column has a mean of 0 and the standard deviation of 1.

3.3 Topic Modelling

In order to obtain a sense of clarity as to how the features in a census track could explain homeless population, we conduct non-negative matrix factorization (NMF) on the data to form topics or themes that extracts meaningful interpretations of certain features that could help us explain homelessness [9].

A simple illustration of NMF is as follows.

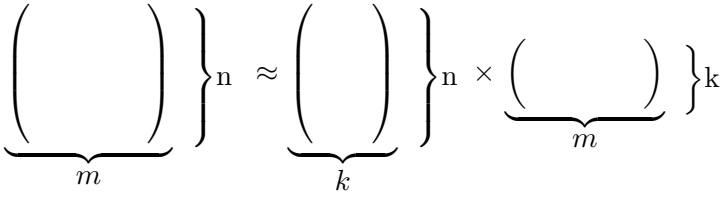


Figure 1: The matrix on the left hand side is the original matrix. The first matrix on the right is the weight matrix and the second matrix is the topic matrix, and k is specified number of topics.

Through NMF, a given matrix would be approximated to a weight matrix and a topic matrix, where k is specified number of topics and all entries of W and H are positive. Each row of the topic matrix represents a unique topic, while each column represents a variable or feature. Moreover, each column of the weight matrix corresponds to a topic, e.g the first column of the weight matrix corresponds to the first topic. We can thus approximate each column of the matrix as a linear combination of the weight matrix weighted by a column of the topic matrix [10]. This means that the weight matrix is a basis that linearly approximates to the given matrix.

To relate to our research matter, denote our data matrix that contains the features of all the census tracks excluding the homeless counts as $A \in \mathbb{R}^{nxm}$ matrix, which is normalized into percentile ranks (See section 4.1). The normalization is done because we want to obtain a more interpretative topic matrix since each column would have the same scaling.

Then, NMF on our normalized data to form a weight matrix, $W^{n\times k}$ and a topic matrix, $H^{k\times m}$ which would approximate to our data matrix such that k is the number of topics.

$$A \approx WH \tag{5}$$

In order to calculate the NMF using Matlab, we use the `nnmf` function[8]. The function finds the optimal W and H by minimizing the root mean square residual in (6). The function would repeatedly factorize using different initial values for W and H to take into account that the residual has local minima.

$$\frac{\|A - WH\|_F}{n \times m} \tag{6}$$

We varied the rank k from $k = 3$ to $k = 5$ in order to test whether a different number of topics could yield weight and topic matrices that would better represent the data. For each non-negative matrix factorization with varying rank k, we selected the top 10 census tracts with the highest weights for each topic and plotted the centroids over a map of Los Angeles in order to visualize the clusters of data.

Moreover, for all k, we also find the correlation between the weights of each topic and the percentile-ranked homeless population density, which is done to analyze the relationship between the topic matrix and the density to see if we could identify dynamics in the homeless population that could be explained by the topics. Note that the homeless population density is also percentile-ranked to create consistency with previous normalization on the data matrix. Denote $W_{.i}$, where i is the i^{th} column of W and D be percentile-ranked homeless population density for all of the census tracks. The correlation between the weights of each topic and the homeless population density, $\rho(D, W_{.i})$:

$$\rho(D, W_{.i}) = \frac{1}{n-1} \sum_{k=1}^{n} \left(\frac{D_k - \mu_D}{\sigma_D} \right) \left(\frac{W_{ki} - \mu_{W_{.i}}}{\sigma_{W_{.i}}} \right) \tag{7}$$

where μ_D is the mean of D, σ_D is the standard deviation of D, $\mu_{W_{.i}}$ is the mean of $W_{.i}$, $\sigma_{W_{.i}}$ is the standard deviation of $W_{.i}$

After finding the correlation, heat map graphs of weight/density for the topics that have the highest positive and negative correlations with the homeless population density and the density itself are plotted on the census track's longitude-latitude plane. That's because the higher the weights of a topic in a census track, the stronger the presence of the topic in the census track. So we expect that the large presence of the positively and negatively correlated topics would encourage and discourage respectively homeless population counts in the census track.

3.4 Local Variance

We also want to understand if a large variance of the features in a neighbourhood of a census track would contribute to high homeless population.

First normalize the columns of our data matrix, A, as z-scores (See section 4.2). The reason for that normalization is to create a consistent distribution for each feature. Then for each census tract with its corresponding features, $A_{i.}$, obtain the 5 nearest census tracks based on their Euclidean distance, i.e the ℓ^2-norm of the pair of longitude and latitude of the census track and that of the other census tracks. Denote $A_{k_j.}$ to be one of the nearest census tracks, where $k_j \neq i$ and $j \in \{1, 2, 3, 4, 5\}$. Then, we form a local data matrix with all the features of the census tracks and its neighbours, i.e form a matrix $B^{(i)} \in \mathbb{R}^{jxm}$, such that

$$B^{(i)} = \begin{pmatrix} -A_{i.}- \\ -A_{k_1.}- \\ -A_{k_2.}- \\ -A_{k_3.}- \\ -A_{k_4.}- \\ -A_{k_5.}- \end{pmatrix} \tag{8}$$

At last, from (8), we take the sum of squared deviations from the mean of each column of $B^{(i)}$, which we denote $SSB^{(i)}$:

$$SSB^{(i)} = \sum_{j=1}^{6} \sum_{t=1}^{m} \|B_{jt}^{(i)} - \mu_t\|^2 \tag{9}$$

Where μ_ℓ is the mean of the column t in B.

3.5 Principal Component Analysis

Principal component analysis (PCA) aims to find an orthogonal transformation for the data matrix to denoise or reduce the dimension of the data. The transformed matrix T's column are the principle components. Matrix T has the property that the first column has the largest variance possible, accounts for the most variability in the data. The following columns have the largest variance, given that they need to be orthogonal to the previous columns. PCA is sensitive to the scaling of the original variables, therefore we use the normalized Z matrix derived from Z-score normalization above [11].

It is common nowadays to use single value decomposition (SVD) to derive the principal components. Suppose our data matrix is $Z \in \mathbb{R}^{n \times m}$, we can decompose it via SVD to obtain the eigenvalue matrix Σ, eigenvector matrix W, and the corresponding unitary matrix U. Since each eigenvector points to the direction that accounts for the largest variance while being orthogonal to each other, we can then obtain T by matrix multiply Z and W.

$$Z = U\Sigma W^T \tag{10}$$

$$T = ZW \tag{11}$$

And each column of T is the principal component of Z.

To choose the optimal number of principal components, since each eigenvalue in the eigenvalue matrix Σ represents variance explained by each principal component, the total variance explained

by the first k principal components can be computed by simply taking the sum of the first k eigenvalues divided by the total sum of the eigenvalues [12]. As such, suppose we choose the k eigenvectors that corresponds the proportion of variance in the data we want explained, and form a matrix W_k, the corresponding transformed matrix T':

$$T' = ZW_K \tag{12}$$

3.6 Cluster analysis

Clustering census tracks based on the features presents a clear idea on the distribution of the homeless population over the years and also how the features would group census tracks that would explain homelessness in the tracks.

For our analysis, clustering based on the Gaussian Mixture model is done on the features of the census tracks through the Expectation-Maximization (EM) method [13]. The model works as follows: Given the data matrix, it clusters each census track into a pre-determined number of groups such that each group will have a Gaussian distribution with a mean and covariance that is different from that of other groups.

For our calculation, first the optimal Gaussian Mixture model is created by using the `fitgmdist` function from MATLAB [8]. The function requires 2 inputs: the data matrix and the number of clusters. It then performs an iterative EM algorithm to maximize the likelihood of the model. Then, fit the data and the model into the `cluster` function, which would then output the cluster index for each census track [8].

Before the clustering is done, many features are removed from the data in order to avoid the 'curse of dimensionality; in order for the clustering to work, there must be a high similarity between each member of a group and low similarity between members of different groups, but high dimensional data, such as ours, distorts that similarity, since census tracks very similar to each other with large number of features, undermining the differences between the census tracks [14]. As such, to determine the features that are kept for the analysis, first take the correlation between the homeless population counts and all the features and keep only the features that have correlation greater than 0.4 as a benchmark [15]. The reason why the number is chosen is because most features have low correlation with homeless population counts. Moreover, features that are significant from the highest positively and negatively correlated topic with homelessness are also kept, since these features provide information about homelessness in the census track (See section 4.4). Features that have strong correlation with each other are also removed, since the Gaussian Mixture model would fail to produce consistently good clusters.

After removing the features, principal component analysis is also done on the remaining features to further denoise the data (See section 4.3). Then, choose the first few principal components such that the variance explained by these principal components is more than 90% and then perform the clustering on them.

After the clustering, the probability density function (PDF) of homeless population density for each cluster is plotted for years 2015 to 2017. The probability density function is estimated using the kernel density estimation [16]. For our calculation, the homeless population density for each cluster is inputted into MATLAB's `ksdensity` function to obtain an estimation of the PDF[8]. Bar charts of the homeless population density as well as its changes are plotted for each year.

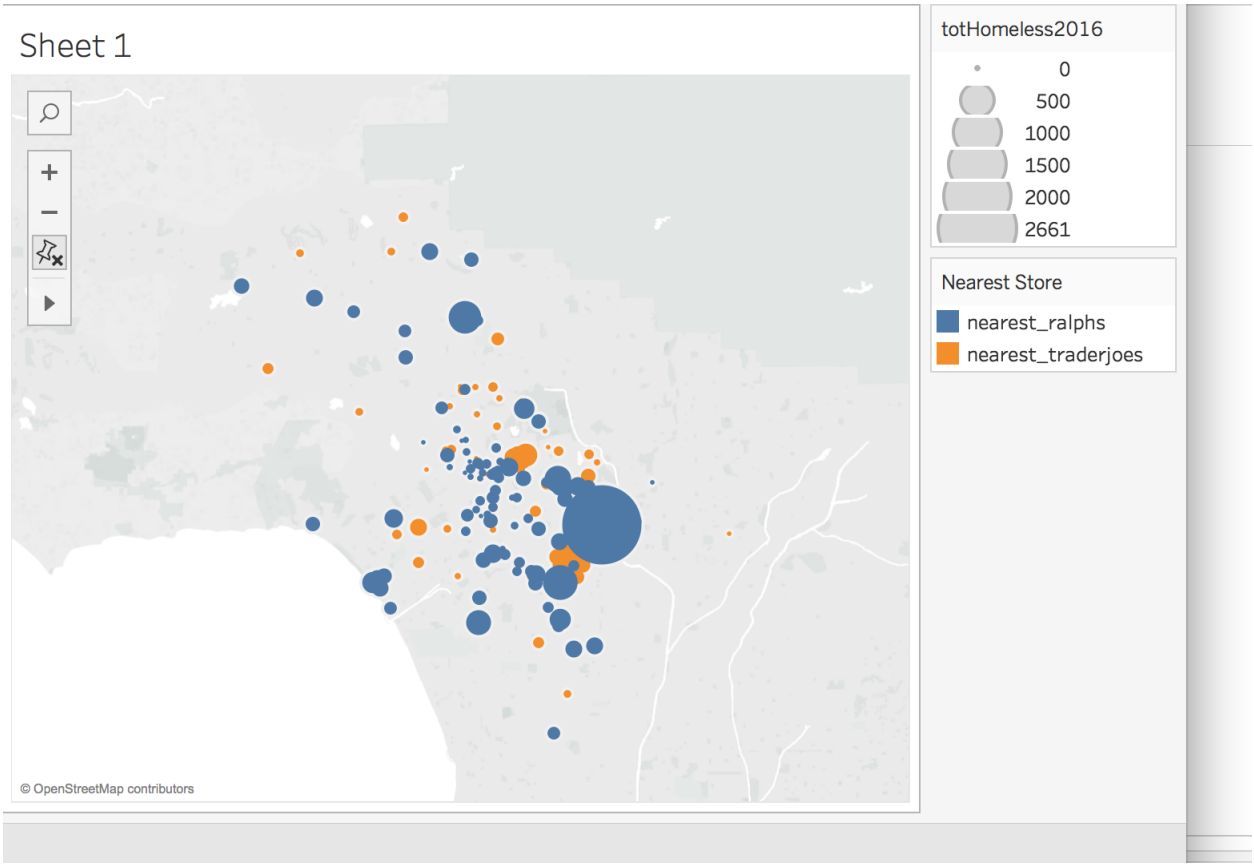
To interpret the features that would explain the clustering of the census tracks, we look at the features of the census track that belongs to each of the clusters. To do that, we first normalize the resulting data before PCA but after removing the variables by Z-score (See section 4.2), then we take the mean of the processed data to determine if the features are important in explaining

the census track belonging in that cluster. So for example, suppose for the census tracks in the first cluster, mean of some features would be positive while the rest of mean would be negative. Positive mean would indicate that the features are highly-valued in these census tracks, and thus contribute to the census track being in that cluster. If the mean is negative, it would in turn mean that the features are mostly very small-valued, and thus contribute little as to why the census track is assigned that cluster.

3.7 Statistical Models

We focused on creating namely three categories of linear models to predict the variation in three different categories of the homeless population. These three distinct categories of the greater homeless population are the shelter homeless population, the homeless population living in vehicles, and the homeless population living on the street. In distinguishing these three categories, we intend to explore which features of a census tract are specific to explaining each sub-population of its homeless population, and we hope to see if variation in a given homeless population category of a census tract can explain variation in the census tract's other homeless population categories. In addition to using the other sub-populations to explain each category, we also created predictors from a variety of collected census tract features including distance to the nearest Ralphs, distance to the nearest Trader Joes,distance to the nearest Whole Foods, distance to the nearest public library, the number of citations, the number of coffee shops, the number of restaurants, the number of shelters, the total number of cars, the 2015 ZRI, the 2016 ZRI, the 2017 ZRI, the 2015 ZHVI, the 2016 ZHVI, the 2017 ZHVI, the median home income, the number of affordable housing units, the number of bus stops , the 2015 Housed Population, the total 2015 homeless population, the total 2016 homeless population, and census tract square millage. Additionally, we were able to construct models that incorporate variables from the Public Use Microdata Sample which contributes features detailing information about various military involvement measures of each census tract's population.

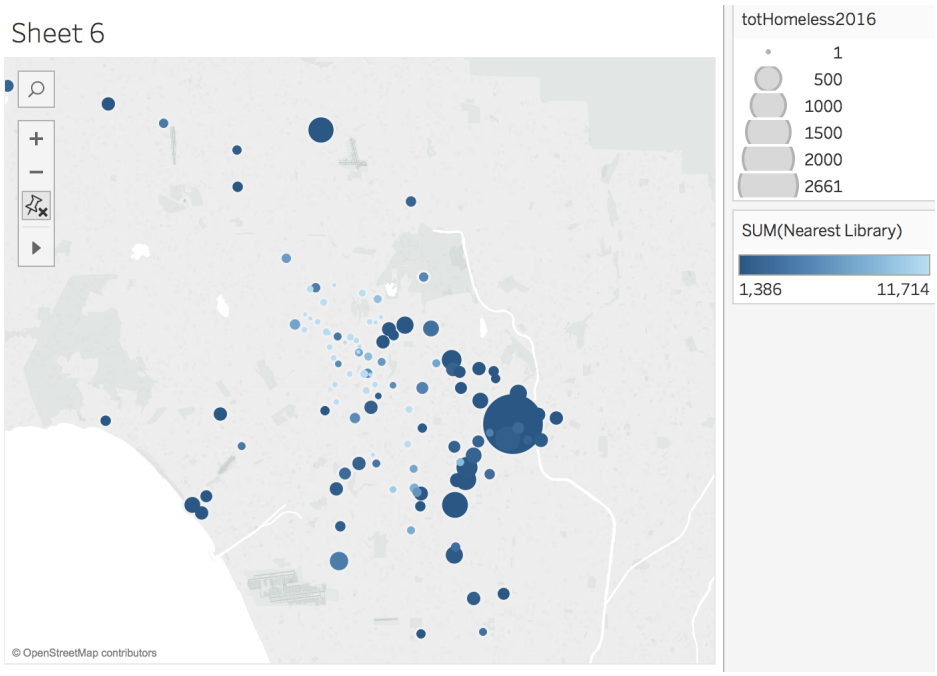
Figure 2: Census Tract 2016 Homeless Population Size with Brand of Nearest Grocery Store



Visualizing the size of census tract homeless populations with the brand of their nearest grocery store

It appears that census tracts with larger homeless populations tend to be in closer proximity to Ralphs, while census tracts with smaller homeless populations tend to be in closer proximity to Trader Joes.

Figure 4: Census Tract 2016 Homeless Population Size with Proximity to Public Libraries



Visualizing the size of census tract homeless populations with their proximity to public libraries (distance to nearest stores measured in haversine distance).

It appears that census tracts with larger homeless populations tend to be in closer proximity to Public Libraries (dark blue), while census tracts that are farther from Public Libraries (lighter blue) tend to have smaller homeless populations.

Through linear regression, we sought out to discern if patterns like these are statistically significant. In the modeling process for each of these three homeless population categories, we express the specified homeless population y_i of census tract i as a linear combination of our data sets p features for census tract i.

$$y_i = \beta_0 1 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \boldsymbol{\varepsilon_i} \qquad i = 1, \ldots, n$$

In order to use linear models to express the data, we must ensure that we can meet the four underlying assumptions for linear regression. The first assumption is an assumption of linearity, we assume a linear relationship between response variable and independent variables. The second assumption is that of multivariate normality; we assume the residuals are distributed normally. Were the residuals not to follow a normal distribution, it would indicate the the true relationship in our data is not being captured in our linear model. The third assumption for linear regression is homoscedasticity; we assume our residuals have constant variance across fitted values such that the residuals for making low estimates are not significantly greater than or less than the residuals for making high estimates. To check that these assumptions are sound, we inspect the diagnostic residual plots for the linear regressions (see Appendix D Figures 39-42). The fourth assumption pertains to the predictors that we include in our linear model. We assume that our explanatory variables are independent of one another. In order to ensure that our variables avoid high multicollinearity, we inspect the Variance Inflation Factor for each variable throughout the model construction. The Variance Inflation Factor detects high collinearity associated with each predictor by capturing how

much variation of a predictor is explained by the other predictors. In order do this, we express each predictor,X_i, as a linear combination of the other predictor variables.

$$X_i = \alpha_0 + \alpha_2 X_2 + \alpha_3 X_3 + \cdots + \alpha_k X_k + \epsilon_i \tag{13}$$

In expressing each predictor in this way, we can calculate each predictor's associated coefficient of determination (14). We will then use this in our calculation for the variable's Variance Inflation Factor (15).

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} \tag{14}$$

$$\text{VIF}_\text{i} = \frac{1}{1 - R_i^2} \tag{15}$$

In accordance with with Hair et. al. (1995), to control for issues of high multicollinearity, we eliminated predictors that had Variation Inflation Factors greater than 10 [17]. This corresponds to predictors whose variations are more than 90% ($R^2 = .90$) explained by the variation of the other predictors (13). For purposes of our model constructions, this is what motivated the removal of various Zillow variables pertaining to property values and rent index's which collectively served to explain too much variation of one another.

In order to meet assumptions for normally distributed residuals, we imposed linear transformations on the response variable. In modeling the 2017 category populations, we imposed square root transformations on the response variables to attain normally distributed residuals with constant variance.

In addition to modeling with all of the features available, we also employed feature selection methods to distill our available predictors into a smaller subset. To identify the features we would include in our regression models, we used a forward step-wise regression process to inform feature selection. In each step, this step-wise regression process will consider a variable for addition from the set of explanatory variables based on some pre-specified criterion. We constructed models using both Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) as criterions. These are criteria for model selection that measure the trade-off between model fit and complexity of the model. Low values for AIC and BIC indicate a better fit. Both AIC and BIC reward goodness of fit, and have penalties for too many predictors. BIC's penalty for adding predictors is greater than AIC's penalty which leads to BIC's tendency to produce underfitting models, and AIC's tendency to produce overfitting models. Below, k is the number of predictors, and $\hat{L}$ is the maximum value of the likelihood function for the model.

$$\text{AIC} = 2k - 2\ln(\hat{L}) \tag{16}$$

$$\text{BIC} = \ln(n)k - 2\ln(\hat{L}). \tag{17}$$

The forward step process begins with no predictors, and then iteratively will add predictors that will benefit the model's AIC and BIC values. Predictors are added individually until adding predictors no longer improves the calculated AIC or BIC for the model.

As another strategy to model the variation in the three categories of the homeless population, we also used R's Random Forest package to construct Random Decision Forests. Random Forest is a statistical learning method that uses bootstrap aggregating to construct an ensemble of decision

trees [18].The trees are constructed on different sub-samples from the data, each attained by randomly sampling from the original data with replacement. Ultimately, random forest is an additive model which will average the decision of each tree in making its prediction for the regression. The benefit of using random forest regression, is that it gives us another means of identifying important features that are unique to explaining each specific category of a census tract's homeless population [18] Important features are identified with random forest by measuring the Mean Decrease Accuracy for when the variable's values are randomly permuted. The importance for a feature is computed as the average decrease in model accuracy when the values of the respective feature are randomly permuted [19]. If a variable is important, we expect that randomly assigning its values to observations would cause our regression model to make predictions with less accuracy and greater mean squared error. This is in contrast to unimportant variables for which randomly permuting values would not greatly effect the regression's prediction accuracy.

3.8 Exponential Distance Decay-function

Homeless populations are mobile, they can migrate from one location (census tract) to another. Therefore, we accounted for this migration effect by adding the data of the local neighborhood surrounding a census tract into the data of the central census tract. The effect of nearby features decay when the census tract is further away from the central census tract, and one of the best functions to account for this distance decay phenomenon for migration analysis is the exponential distance decay-function [20].

$$e^{-d/\lambda} \tag{18}$$

We calculated the distance between two census tracts by calculating the euclidean distance between their center position. d is the distance converted into miles. We experimented on values of λ: 0 (close to 0), 0.1, 0.2, 0.3, ... , 0.9, 1. The shape of the exponential distance decay-function with selected λ is shown below. A smaller λ gives less weight to the local features further away and a larger λ gives a larger weight to local features further away.

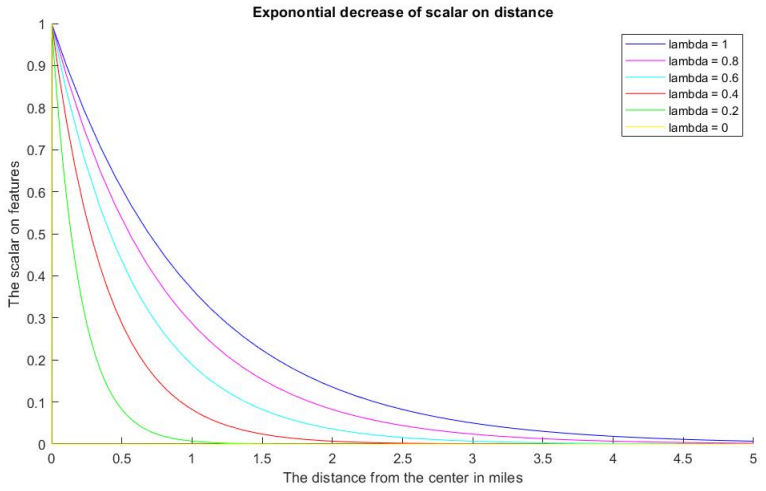


Figure 6: The shape of the exponential distance decay-function function with different λ values.

We summed the central census tract and its 5 local nearest tracts with different weights given by the exponential distance decay-function. X_i denotes a census tract entry in the data matrix (a

row vector) and S_j is a census tract entry in the data matrix after the procedure.

$$S_j = \sum_{i=0}^{5} e^{-d_i/\lambda} X_i \tag{19}$$

We then concatenated all the S_j for all census tracts into a new data matrix S. We took the matrix S generated by different λ, fed them into a classifying neural networks, and checked their average testing accuracy to pick the best λ for our data set.

3.9 Neural Networks

We wish to predict the changes in homeless population from the time-dynamic information about the city. Neural networks construct complex non-linear regression models that fit our purpose well. The basic structure of a neural networks is represented by the figure below [21].

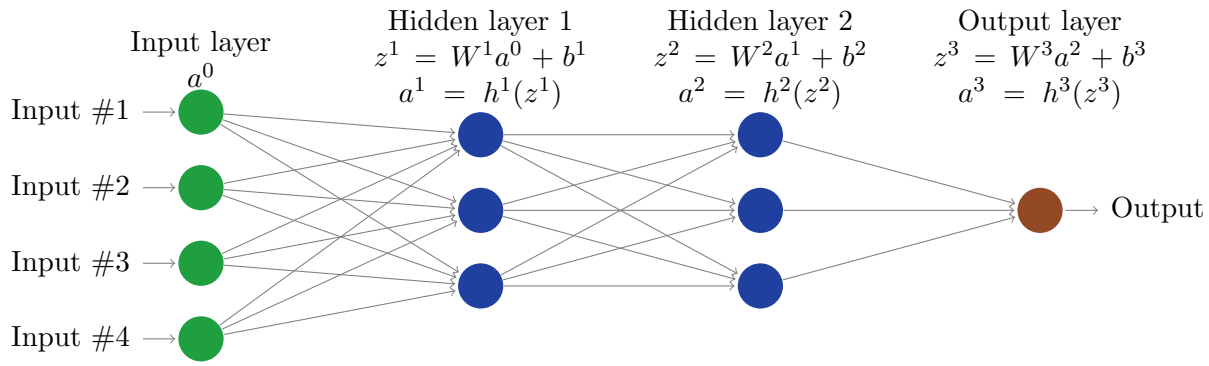


Figure 7: Structure of neural networks

Imagine a size m data-set with n features. The i^{th} entry is (x_i, y_i), where x_i is the vector of all the predictors and y_i is the correct output. We want to approximate y_i from x_i. Each layer of the neural networks learns to approximate the value by taking the linear combination of the input and incorporate some non-linearity by activating the output value through a non-linear activation function.

The neural networks takes the linear combination of the input by multiplying the input with a weight matrix and adding a bias. For the step that connects the $(\ell-1)^{\text{th}}$ layer to the ℓ^{th} layer in the neural networks, W^ℓ denotes the weight matrix and W^ℓ_{jk} denotes the entry at (j,k) position in the weight matrix. b^ℓ denotes the vector of bias and b^ℓ_j denotes the j^{th} entry of the bias. h^ℓ denotes the non-linear activation function of the ℓ^{th} layer. The output after the activation, which is the ℓ^{th} layer vector, is a^ℓ and a^ℓ_j denotes the j^{th} entry of the a^ℓ vector.

With these notations, the equation to calculate the entries of the ℓ^{th} layer from the $(\ell-1)^{\text{th}}$ layer is:

$$z^\ell_j = W^\ell_{jk} a^{\ell-1}_k + b^\ell_j \qquad\qquad a^\ell_j = h^\ell(z^\ell_j) \tag{20}$$

We can rewrite the above equation in vector form:

$$z^\ell = W^\ell a^{\ell-1} + b^\ell \qquad\qquad a^\ell = h^\ell(z^\ell)$$

The activation functions used in our models are: ReLU, leaky ReLU, Sigmoid, and Softmax. ReLU function introduces non-linearity by "turning off" the negative values. The ReLU function outputs the original input if the input is positive or 0, 0 if the input is negative [22].

$$\begin{cases} h(z) = z & \text{if } z \geq 0 \\ h(z) = 0 & \text{if } z < 0. \end{cases} \tag{21}$$

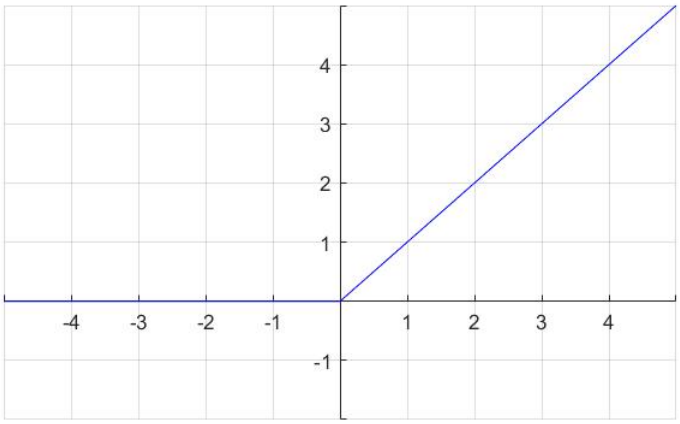


Figure 8: This figure shows the shape of the ReLU function.

The leaky ReLU function is similar to ReLU function. It does not turn off the negative value but makes the negative values grow slower. It outputs the original input if it is positive or 0. If the input is negative, the function outputs the original value multiplied with a scalar k $(0 < k < 1)$. We picked $k = 0.2$ for our leaky ReLU function [22].

$$\begin{cases} h(z) = z & \text{if } z \geq 0 \\ h(z) = kz & \text{if } z < 0. \end{cases} \tag{22}$$

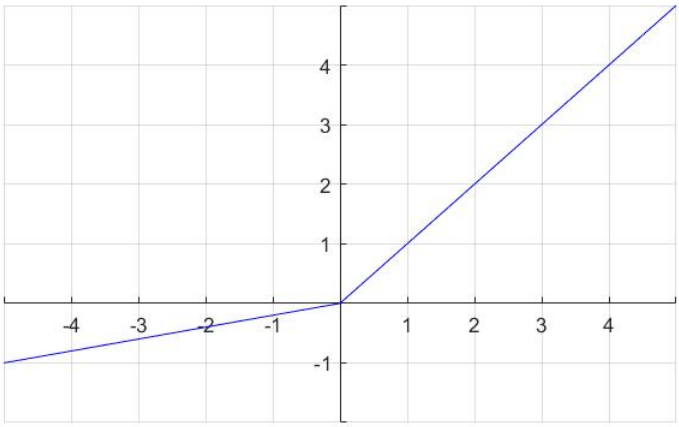


Figure 9: This figure shows the shape of the leaky ReLU function.

The Sigmoid function introduces non-linearity by turning "on" or "off" depending on the input. The Sigmoid function always outputs a number between 0 and 1 and it has a steeper rate of increase when the input is near 0. Therefore a large input would result in an output near 1 (on) and a small input would result in an output near 0 (off) [21].

$$h(x) = \frac{1}{1 + e^{-x}} \tag{23}$$

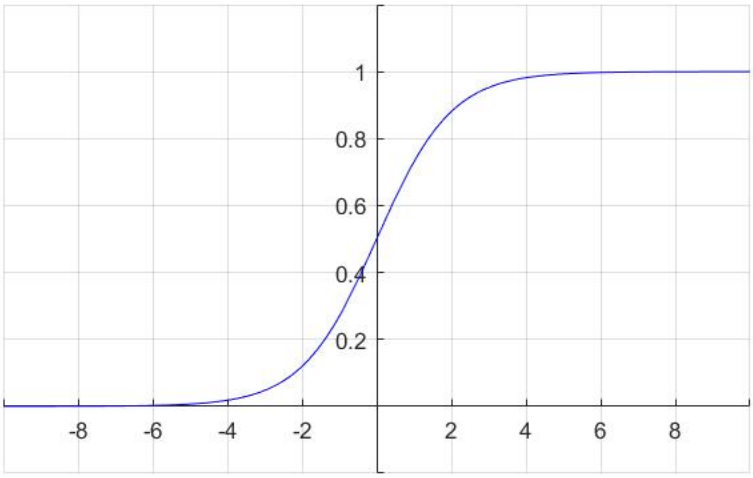


Figure 10: This figure shows the shape of the Sigmoid function.

The Softmax function makes the max value in a layer stands out. The Softmax function reduces a vector of arbitrary real values to a vector of real values where each entry is in the range (0, 1), and all the entries add up to 1 [23].

$$h(z_j) = \frac{e^{z_j}}{\sum_{n=1}^{N} e^{z_n}} \qquad N \text{ is the number of neurons in the layer} \tag{24}$$

The algorithm learns how to best fit the data-set by trying to minimize a cost function. In classifier model, we want the neural networks to learn to classify the input into different categories. Cross-Entropy cost function fits this purpose well [21].

$$C = \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{N} -y_{j_i} \log(a^L(x_j)_i) \qquad N \text{ is the number of output classes} \tag{25}$$

The cost will exponentially increase as the prediction versus actual error increase, therefore it forces the model to learn how to classify the data set.

In regression model, we want the model to minimize the square error of its predictions and we picked Mean Square Error as cost function:

$$C = \frac{1}{2m} \sum_{j=1}^{m} \parallel a^L(x_j) - y_j \parallel^2 \tag{26}$$

The method of minimizing the cost is Gradient Descent [21]. All weights and biases are updated towards the direction of the steepest decrease of the cost by looking at the gradient of the cost function at each iteration. The learning rate α ($\alpha > 0$) is the size of the step each iteration takes. For example, from the i^{th} iteration to the $(i+1)^{\text{th}}$ iteration, the weight and bias should be updated by:

$$W_{i+1}{}^{\ell}_{pq} = W_{i}{}^{\ell}_{pq} - \alpha \frac{\partial C}{\partial W_{i}{}^{\ell}_{pq}} \qquad\qquad b_{i+1}{}^{\ell}_{k} = b_{i}{}^{\ell}_{k} - \alpha \frac{\partial C}{\partial b_{i}{}^{\ell}_{k}} \tag{27}$$

Adam is a variation of gradient descent. The method adjusts the learning rates α for different parameters from estimations of the first and second moments of the gradients [24].

The gradient of the cost function against each elements of the weight and bias in each layer can be calculated by applying the chain rule repeatedly. This way of finding the gradient is called backpropagation [25]. The chain rule can be written out as this:

$$\frac{\partial C}{\partial b^{\ell}_{k}} = \frac{\partial C}{\partial a^{L}} \frac{\partial a^{L}}{\partial a^{L-1}} \frac{\partial a^{L-1}}{\partial a^{L-2}} \cdots \frac{\partial a^{\ell+1}}{\partial a^{\ell}} \frac{\partial a^{\ell}}{\partial b^{\ell}_{k}} \tag{28}$$

$$\frac{\partial C}{\partial W^{\ell}_{pq}} = \frac{\partial C}{\partial a^{L}} \frac{\partial a^{L}}{\partial a^{L-1}} \frac{\partial a^{L-1}}{\partial a^{L-2}} \cdots \frac{\partial a^{\ell+1}}{\partial a^{\ell}} \frac{\partial a^{\ell}}{\partial W^{\ell}_{pq}} \tag{29}$$

Each element of the chain rule is easy to compute. Multiplying them together would give us the gradient needed.

The features we used are in two groups:

- Static: Coffee-shop count, Restaurant count, Shelter count, Library Count, Ralphs Count, Trader Joes Count, Crime Count, Bus Stops count, Tract size in Square Miles;
- Time Dynamic:
  - City Features: Available Housing Units, Total Vacant Housing Units, Unemployment Rate, Below Poverty Rate, Medium Rent As Percent Of Gross Income, Total Population, Medium Household Income, Medium Rent, Medium Home Value, Medium Monthly Housing Costs, ZRI, ZHVI, Population per Square Mile.
  - Homeless Population: Total Homeless Population, Total Homeless Population in Shelter, Total Homeless Population in Tents, Total Homeless Population in Encampments, Total Homeless Population in in Cars, Total Homeless Population in Vans, Total Homeless Population in Campers

For data in the static group, we assumed that there was not much yearly changes. For data in the time dynamic group, we have data from 2014, 2015, and 2016 for the city features and we have 2015, 2016, 2017 for the homeless population data. Because of the limitations of the data we described in the Data Management section, there are only 818 census tracts that contains all the data we desire to use.

We compiled the data that account for changes in homeless population from 2015 to 2016 and from 2016 to 2017 together to form our data set.

We used the data of the static and city feature group of 2014 and 2015 and the homeless population data (with all it's subcategories) of 2015 to predict the changes in homeless population from 2015 to 2016; and we used the data of the static and city feature group of 2015 and 2016 and the homeless population data (with all it's subcategories) of 2016 to predict the changes in homeless population from 2016 to 2017.

We used the data of the static and city feature group of 2014 and 2015, the homeless population data (with all it's subcategories) of 2015, and the homeless population data of homeless population categories not in vehicles (Total homeless population in Shelter, Total Homeless Population in Tent, Total Homeless Population in Encampment) of 2016 to predict the change in homeless population in vehicles from 2015 to 2016; we used similar data-set to predict the changes in homeless population in vehicles from 2016 to 2017.

We used the data of the static and city feature group of 2014 and 2015, the homeless population data (with all it's subcategories) of 2015, and the homeless population data of homeless population categories not on the streets (Total homeless population in Shelter, Total Homeless Population in Cars, Total Homeless Population in Vans, Total Homeless Population in Campers) of 2016 to predict the change in homeless population on the street from 2015 to 2016; we used similar data-set to predict the changes in homeless population in vehicles from 2016 to 2017.

We used the changes of city features of one year before to predict the changes in homeless population because (1) ACS did not release the data for 2017 yet [26], and (2) the homeless population data is gathered during January, changes in city feature in one year before are reflected in the changes of homeless population one year later [27].

We then split the data set up into two groups of 70% and 30%; 70% of the data was used as the training set and 30% of the data was used as the testing set. Therefore, the size of the training set is $818 \times 2 \times 70\% = 1145$ and the size of the testing set is $818 \times 2 \times 30\% = 491$.

We did ternary classifications trying to predict whether the changes of total homeless populations and the subcategories of the homeless population (population in vehicles, population on the streets) in a census track is going down, staying static, or going up. We divided the changes in homeless population into three buckets of equal size (roughly 33% each). The lower bucket represent a significant decrease in homeless population (-), the middle bucket represent a small, insignificant variation in the homeless population (0), and the higher bucket represent a significant increase in homeless population (+).

We also did regression models trying to predict the changes of homeless population number.

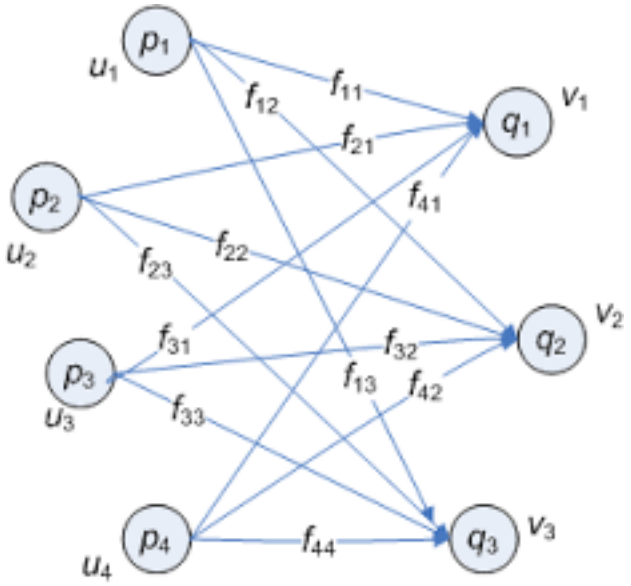
We implemented the neural networks in Tensorflow, a popular machine learning system [28].

We first normalized the data using Z-score normalization. Then we tested different neural net work structures. We did trials on: the number of hidden layers and the number of neurons in each layer. We tried all the activation functions introduced above: ReLU, Leaky ReLU, Sigmoid, Softmax. The optimizer we used is Adam Optimizer.

3.10 Earth Mover's Distance

Homeless populations are not stagnant individuals - they are constantly moving within census tracts of Los Angeles as well as from outside cities. In an attempt to model the flow of Homeless populations between census tracts we employed the Earth Mover's Distance (EMD), or the Wassertein metric. The EMD computes the minimum work, or cost, used from one probability distribution to another probability distribution [29].

Take a cluster of census tracts denoted P such that $P = \{(p_1, w_{p1}), ..., (p_i, w_{pi})\}$ where p_i is the cluster representative and w_{pi} is the weight of clusters. Similarly, take another cluster of points denoted Q such that $Q = \{(q_1, w_{q1}), ..., (q_j, w_{qj})\}$. Take $D = [d_{ij}]$ to be the cost of moving between clusters p_i and q_j



[29]

Our goal is to find flow $F = [f_{ij}]$ with f_{ij} being the flow between p_i and q_j census tracts that minimizes the overall cost. This can be done by minimizing the objective function [29]:

$$\min \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij} \tag{30}$$

In applying this method to our problem of modeling homeless movement, we denoted the flow, f, as the change in population frequencies from one year to the next. In order to account for individuals who entered or left Los Angeles outside of our census tract data we created an external node in the system. The is the same of creating an imaginary census tract that people can flow in and out of when the movement occurs outside of Los Angeles. Let us denote S_P as the total homeless population in one year (2016) and S_Q as the total population in a future year (2017).

$$S_P = \sum_{i=1}^{n} P_i$$

$$S_Q = \sum_{j=1}^{n} Q_j$$

To determine the value of our external node, the migratory patterns of homeless individuals outside of Los Angeles, we took the difference between S_Q and S_P. If the result was positive, that indicated a rise in homeless populations in the future. So our external node for distribution P would be filled with the absolute value of the difference in populations. In distribution Q the external node value was set to zero. In doing so, we conserved the mass, or population counts, within the system. The same process was performed if the difference between S_Q and S_P was negative, indicating a drop in homeless population. In this case, the external node of distribution P would be set to zero and external node of distribution Q would have a value of the absolute difference between the distributions, indicating that people migrated out of the system. f_{ij} would then denote the amount of people moving from tract i to tract j from one year to the next.

We also had to indicate the cost of movement between census tracts. To do this, we used three distance metrics: Haversine distance, Manhattan distance, and Binary cost.

The Haversine distance computes the straight line path on a spherical surface. The haversine distance, d, is computed with the following equation, with r being the radius of the sphere, φ_1 and φ_2, the latitude of point one and point two, in radians, λ_1 and λ_2 the longitude of point one and longitude of point two, in radians [30]:

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \tag{31}$$

The Manhattan distance breaks a straight line into horizontal and vertical components and sums the distances of each component. Since this was done along a horizontal surface, we employed the Haversine equation (31) to compute the distance of each component before summing them to compute the Manhattan distance.

Binary distance was the last cost function that we used. If a homeless person stayed in the same census tract, the cost was computed to be zero. If a homeless person moved between census tracts, the cost was computed to be one.

Using each of these distance metrics, we computed pair-wise distances between each census tracts which become our cost matrix d. With our cost matrix, d, and our population counts in each census tract along with the external node, we used PyEMD, an algorithm for Earth Mover's Distance in Python, to compute the minimal flow, f, between tracts [31].

4 Results

4.1 Topic Modeling

The topic matrices for the non-negative matrix factorization are referred in Appendices A. For each table, the first column are the features, while subsequent columns are the topics. For each topic, the features that are significant are the ones with values larger than 0.1 as a benchmark.

By matching the top ten census tracts with correlating topics to homeless density with the centroid latitude and longitude coordinate, we were able to plot the geospatial locations of the census tracts with for a given topic. For each varying k the maps are as followed:

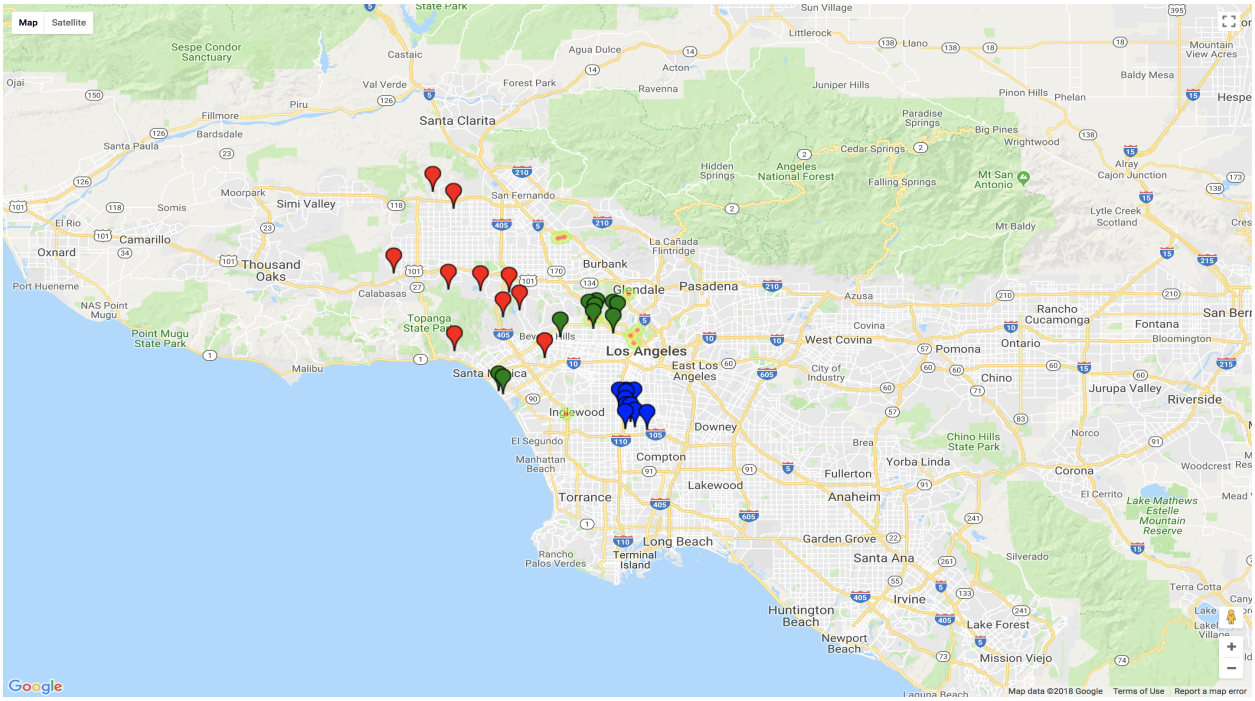


Figure 11a: $k = 3$

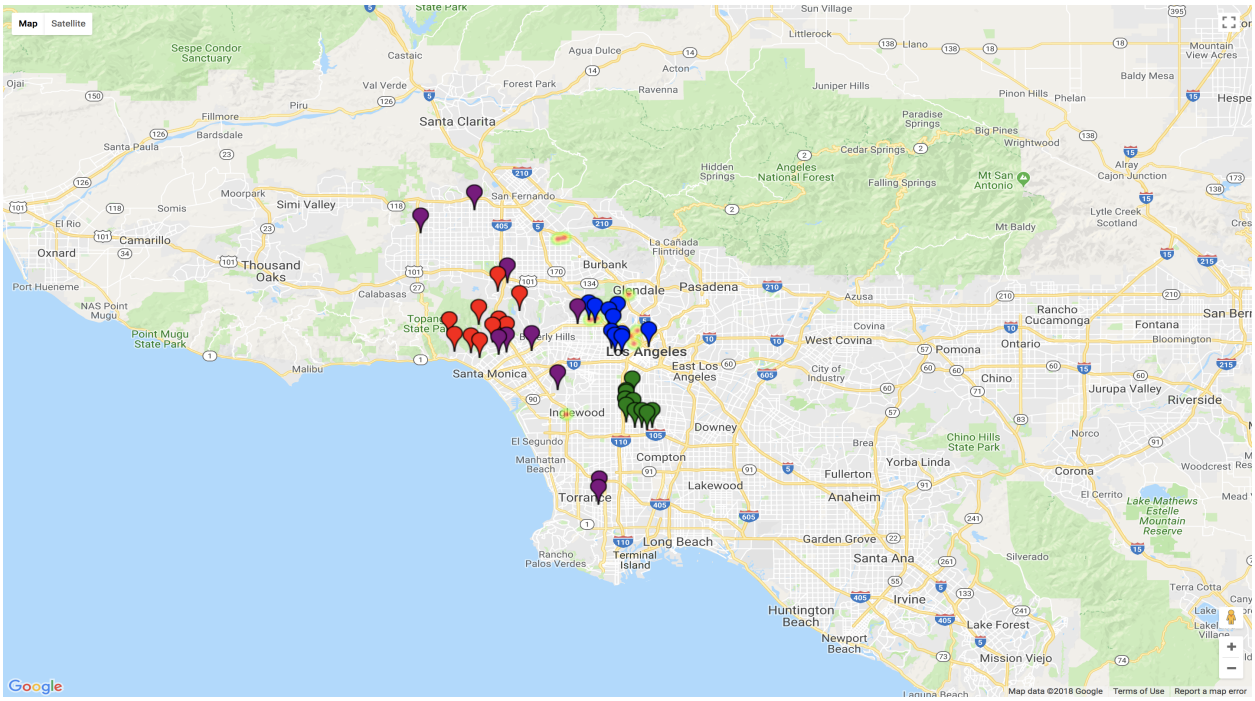


Figure 11b: $k = 4$

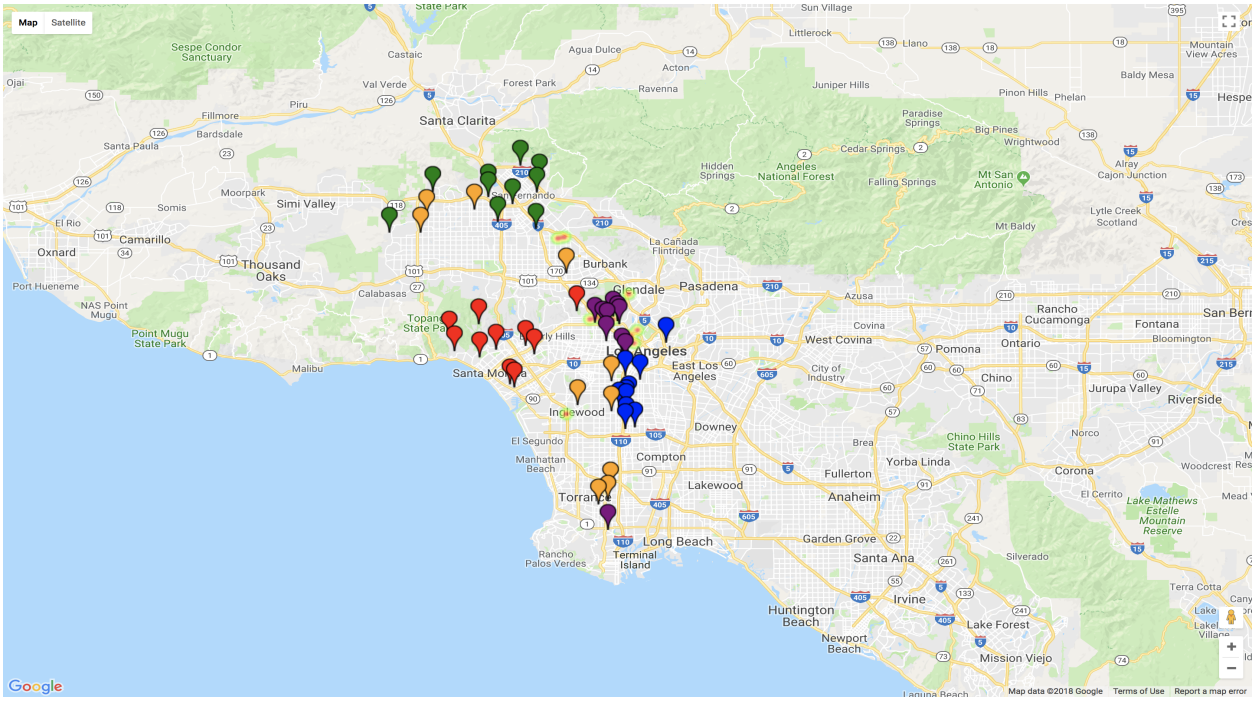


Figure 11c: $k = 5$

The graphs show that the more topics that we choose, the less likely that these topics would give a clear interpretation of features in the census tracks, as the clustering would be much less defined. But nonetheless, we could see that there is at least one cluster around downtown LA, which corresponds high homeless population, which might mean that that topic represents the best features that explain homelessness.

For the correlation between the weights and the homeless population density, heat maps of the coordinates of the census tracks with the highest positively and negatively correlated topics with homeless population density, as well as the density itself are shown from figure 12 to 17. Each figure compares the homeless population density to either the topic of highest positively or negatively correlated to homelessness for different number of topics. For each comparison, there are substantially more overlap between the homeless population density and the highest positively correlated topics. This makes sense since we expect that the census tracks that are high in features that are significant in the highest positively correlated topic would suggest also a high homeless population density, while census tracks that are high in features that are significant in the highest negatively correlated topic would lead to lesser density.

Readers can refer to the highest positively and negatively correlated topics in the Appendices A. The first table corresponds to the 3 topics, the second table corresponds to 4 topics and the third table corresponds to the 5 topics. For each number of topics, the first column represents the highest positively correlated topic, while the second column represents the highest negatively correlated topic. Although there are some variations in how certain features are significant in a topic for different number of topics, there are still some overlaps.

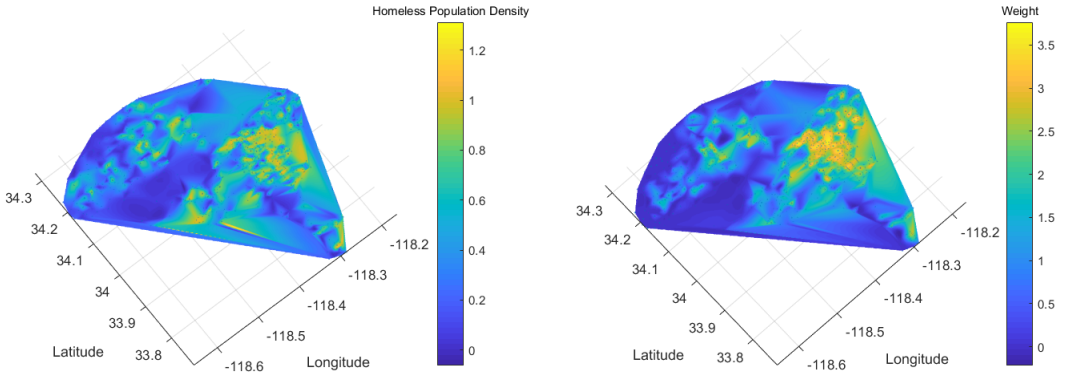
For the highest positively correlated topic, the features that are significant are bus stop density, general population density in 2015, coffee density, restaurant density, crime density,affordable housing unit density, unemployment rate in 2014 and 2015, below poverty rate in 2014 and 2015, median rent as percent of gross income in 2014 and 2015 and most of the changes of features, such as Zillow Rent Index from 2014 to 2016

Some of features are related to amenities, such as restaurant, for which high values would draw homeless people to stay in the census track; some features that can be grouped as measure to poorness, such as unemployment rate, would in turn cause people to be homeless.

For the highest negatively correlated topic, the features that are significant are Zillow Rent Index in 2014 and 2015, Zillow Home Value Index in 2014 and 2015, median household income in 2014 and 2015, median rent 2014 and 2015, median value in 2014 and 2015, median monthly housing costs in 2014 and 2015, some of the changes of features, such as total vacant unit from 2014 to 2016.

Most of these features are related to rent and home value, for which high values would discourage homeless people from coming into the census track.

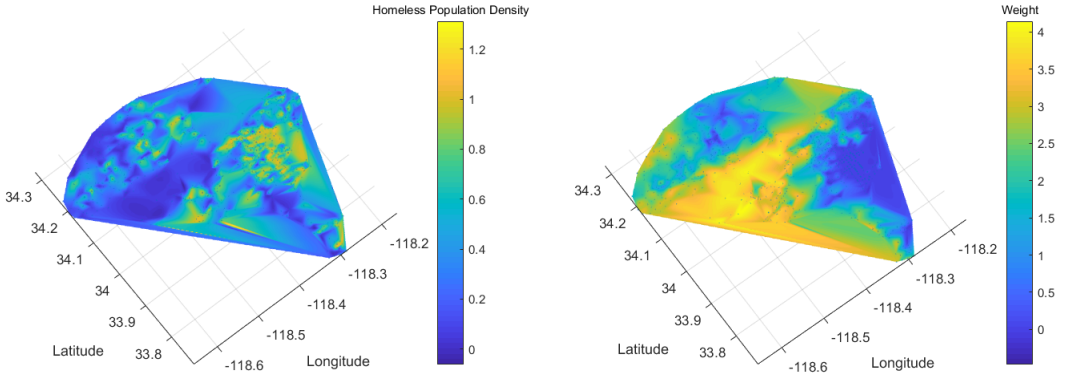
Changes of features, such as the change of the Zillow Rent Index, are present in both positively and negatively correlated topics. This might mean that for some census tracks, a large change in the values would draw homeless people in, while others it would discourage them. So this might mean that depending on the census track's other features, the change in these features would help draw in or discourage homeless people.



(a) Homeless Population Density

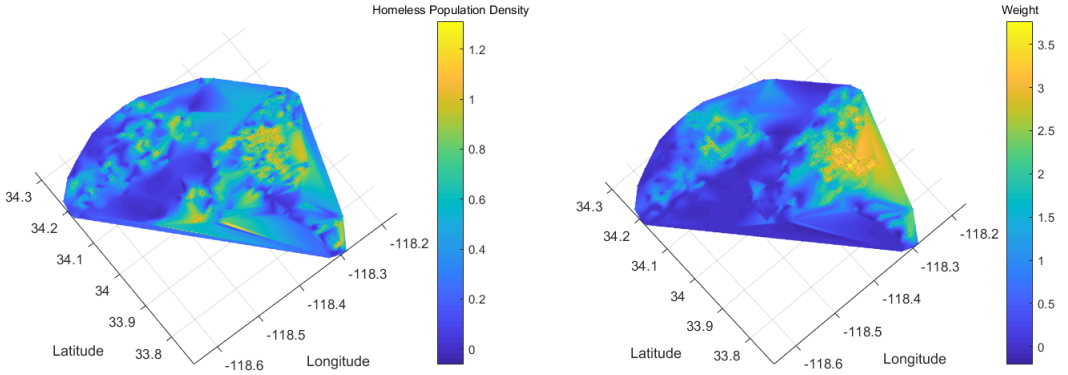
(b) Highest Positive Correlated Topic

Figure 12: Number of topics: 3



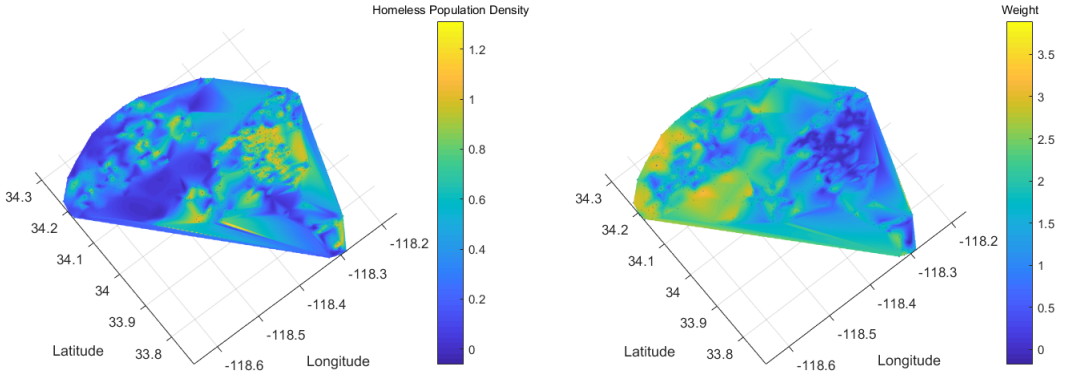
(a) Homeless Population Density (b) Highest Negatively Correlated Topic

Figure 13: Number of topics: 3



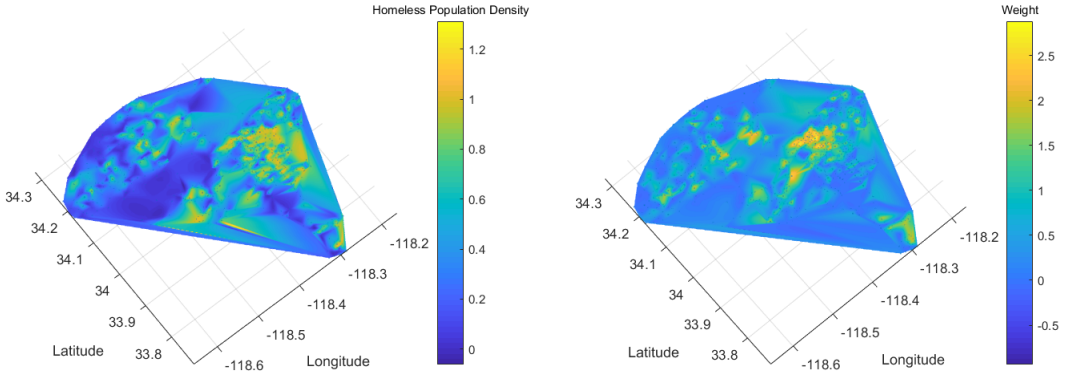
(a) Homeless Population Density (b) Highest Positively Correlated Topic

Figure 14: Number of topics: 4



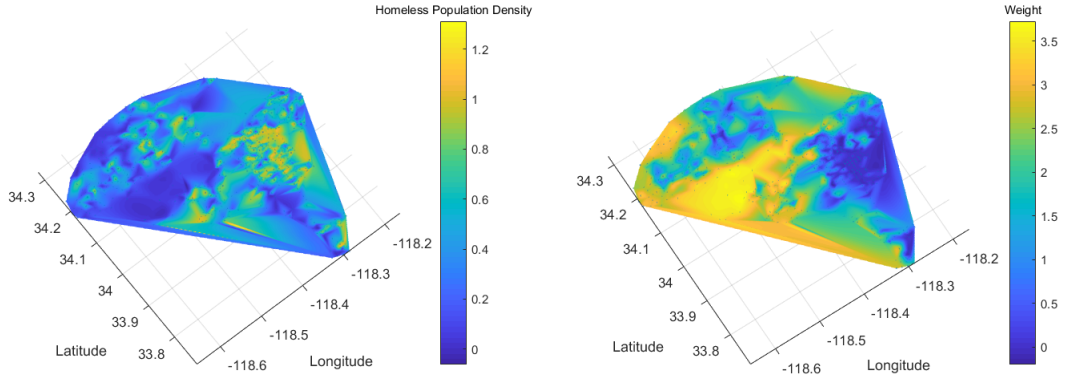
(a) Homeless Population Density (b) Highest Negatively Correlated Topic

Figure 15: Number of topics: 4



(a) Homeless Population Density (b) Highest Positively Correlated Topic

Figure 16: Number of topics: 5

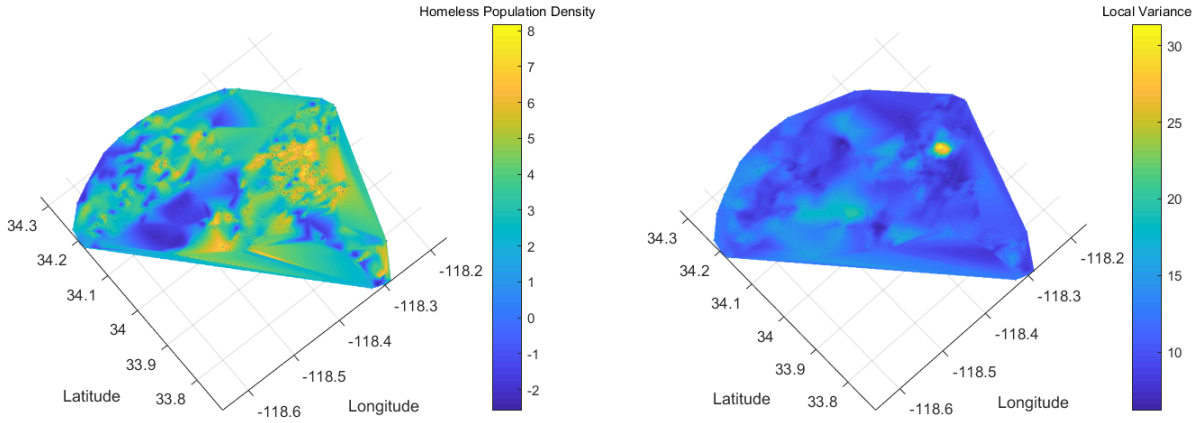


(a) Homeless Population Density (b) Highest Negatively Correlated Topic

Figure 17: Number of topics: 5

4.2 Local Variance

A heat map plot for both the coordinates of the census tracks of the local variance and the homeless population density is in Figure 18. The homeless population density is on log scale to make the heat map more apparent. From the maps, the area with the local variance overlaps with area with the highest homeless population density. The variables that contribute the most to the local variance are bus stop density, general population density 2015, coffee shop density, restaurant density, crime count density, affordable housing density, total housing units 2014/2015, total vacant units 2014/2015, change of affordable housing unit from 2014 to 2016, change of unemployment rate from 2014 to 2016, change of total population from 2014 to 2016, change of median household income from 2014 to 2016, change of median rent from 2014 to 2016 and change of median monthly housing cost from 2014 to 2016.



(a) (1) (b) (2)

Figure 18: (1) homeless population density and (2) local variance

This makes sense because we expect that suppose there is large difference in the amenities and

the conditions to living within a neighbourhood, there would be more homeless people. The changes of the features, representing shocks to the census tracks, also contributes to the local variance, which would mean that the high variance of the changes would lead to an influx of homeless people from the neighbours of a census track into the track.

4.3 Cluster analysis

The number of clusters that we used are 2 and 3, because the more number of clusters we add into the Gaussian Mixture model, the more unstable the results are, as results from 3 clusters are already showing signs of inconsistency through many iterations, while results from 4 or more clusters are just very inconsistent. The reason might be that we specifically choose variables that would highly correlate to homeless population either positively or negatively, and therefore 2 clusters would make the most sense. The variables that are kept are bus stop density, coffee shop density, shelter density, restaurant density, crime count density, below poverty rate 2016, median Household Income 2016, median rent 2016, median value 2016 and median monthly housing cost 2016.

Now we want to interpret the features that contribute to the clustering of the census tracks. The features that contribute to the first cluster are median Household Income 2016, median rent 2016, median value 2016 and median monthly housing cost 2016. The features that contribute to the second cluster are bus stop density, coffee shop density, shelter density, restaurant density, crime count density and below poverty rate 2016.

From results of topic modelling and local variance, we can interpret the features contributing to the second cluster are features that are positively correlated to homeless population and also contribute to the highest local variance. This means that the second cluster captures the census tracks with highest densities of homeless population. From the results of topic modelling, we can see that the these features from the first cluster corresponds to those that negatively correlated to homeless population, which makes sense since it captures census tracks with the least of densities of homeless population.

The following are the plots of the probability density functions (PDF) for the homeless population for each cluster from 2015 to 2017 for 2 clusters:

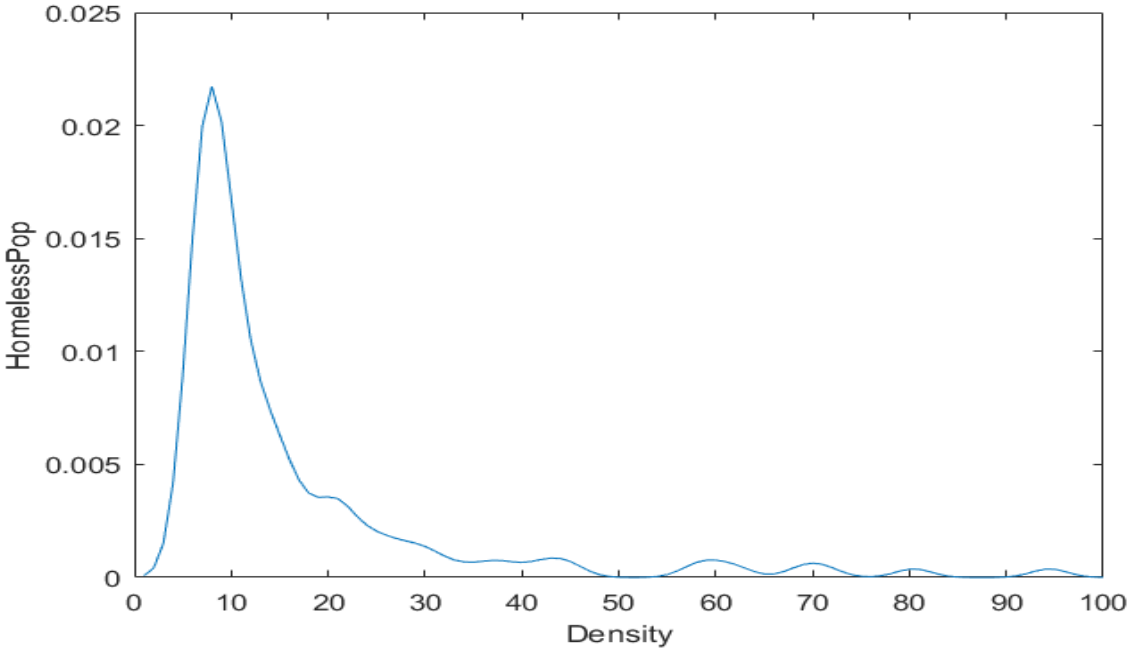


Figure 19a: PDF of homeless population for the first cluster in 2015

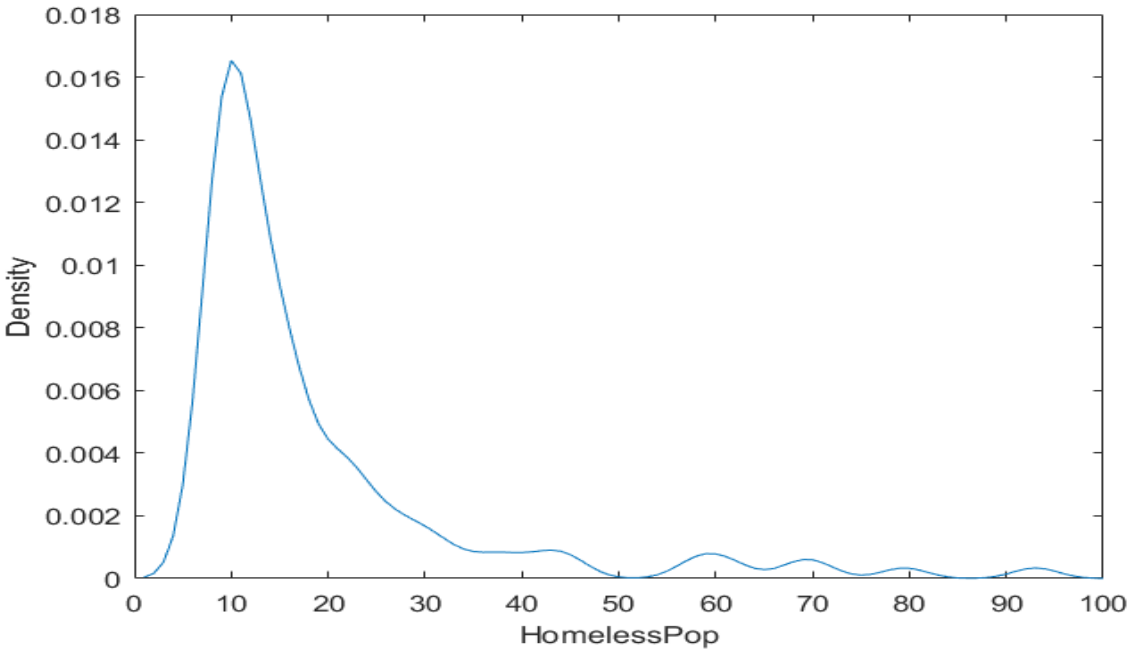


Figure 19b: PDF of homeless population for the second cluster in 2015

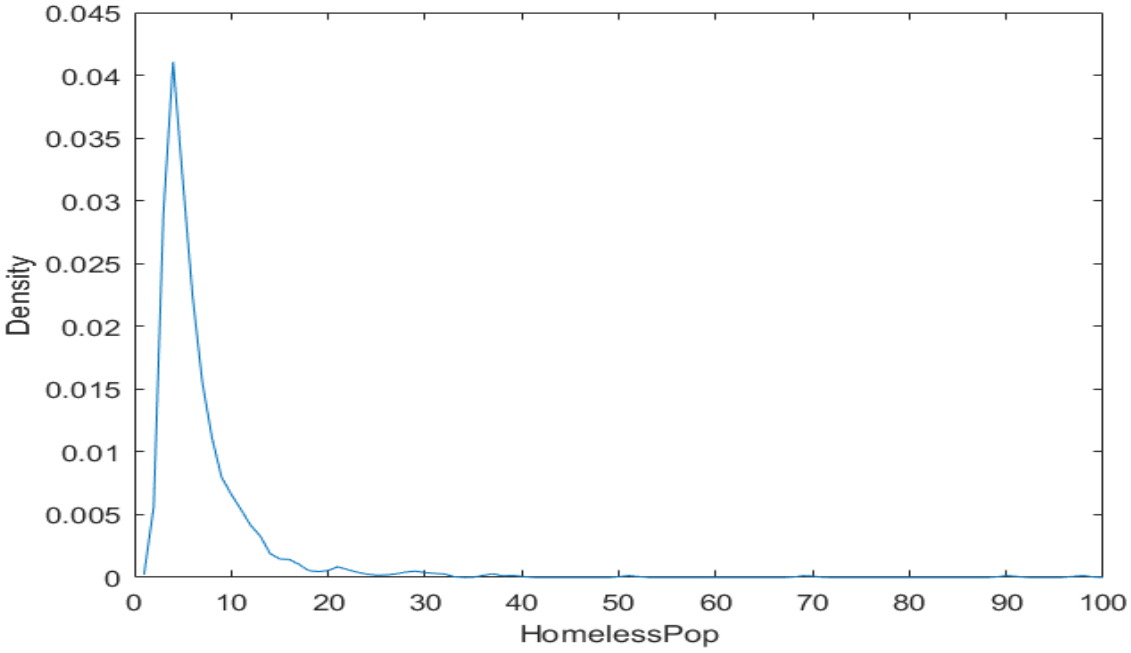


Figure 19c: PDF of homeless population for the first cluster in 2016

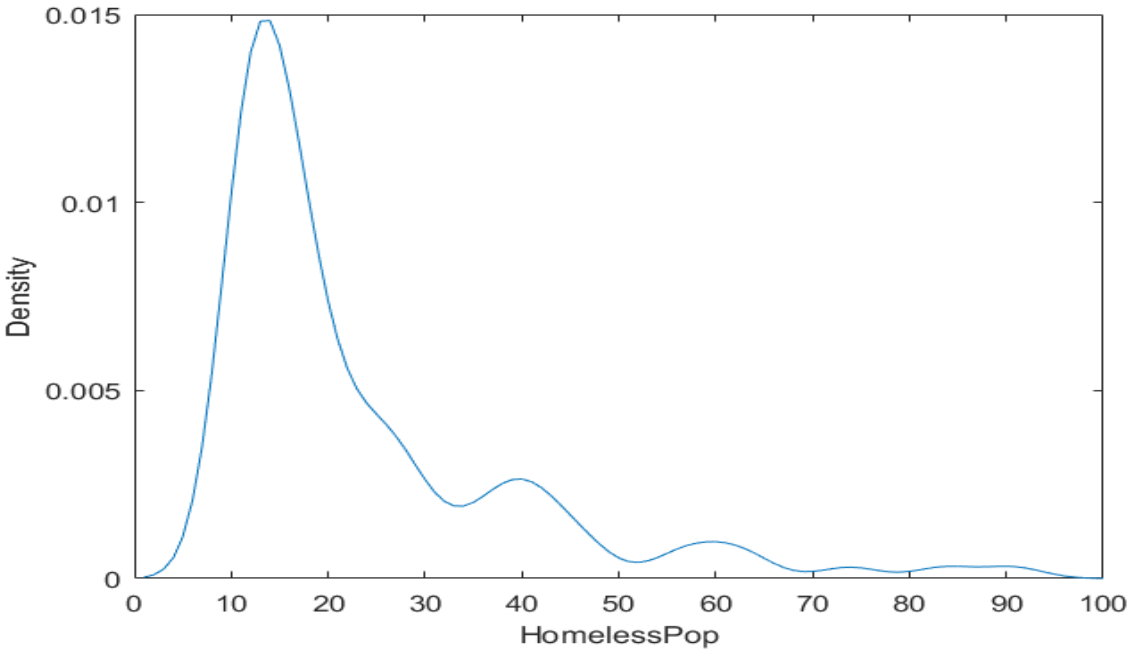


Figure 19d: PDF of homeless population for the second cluster in 2016

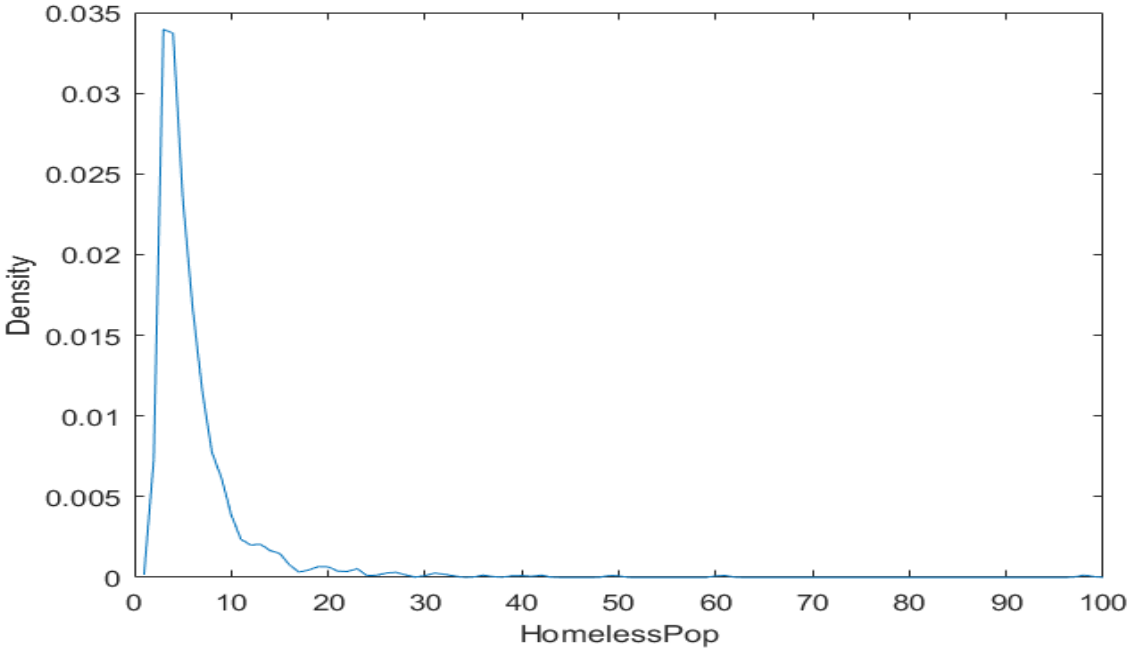


Figure 19e: PDF of homeless population for the first cluster in 2017

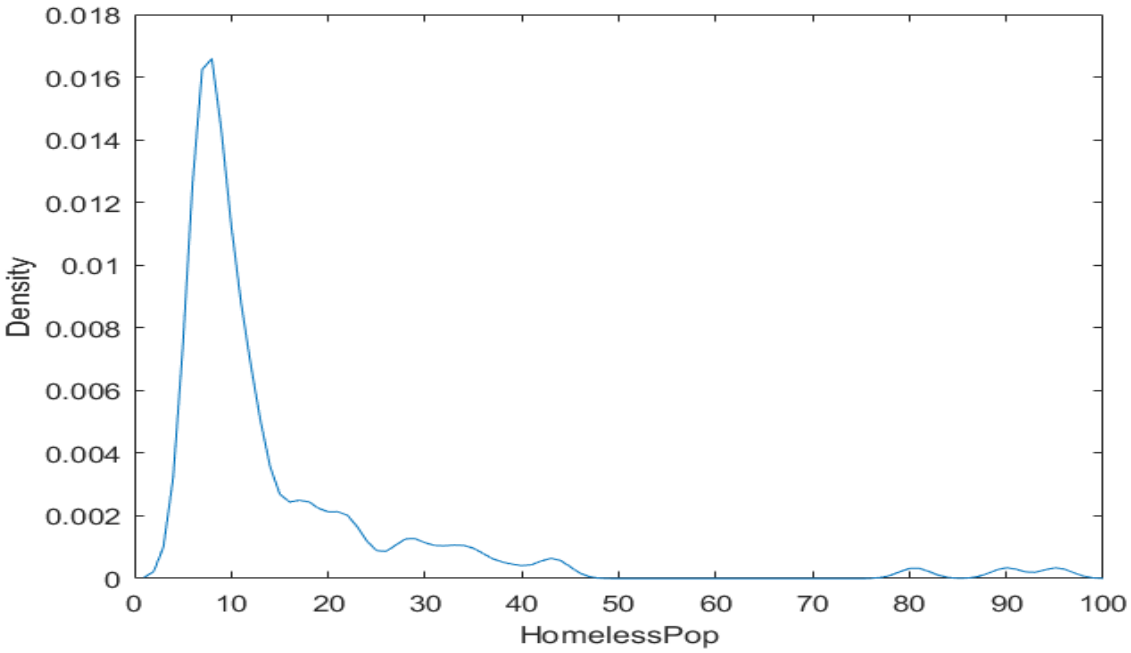


Figure 19f: PDF of homeless population for the second cluster in 2017

From the PDFs, we could see that from the first cluster, the densities are relatively higher but also peaks of each homeless population bin is much smaller than the those of the second cluster. This means that the first cluster would capture more census tracks with smaller number

of homeless population, while the second cluster would capture less census tracks but with high number of homeless population.

The following are bar charts of the total number and percentage change of homeless population in each cluster from 2015 to 2017 for 2 clusters:

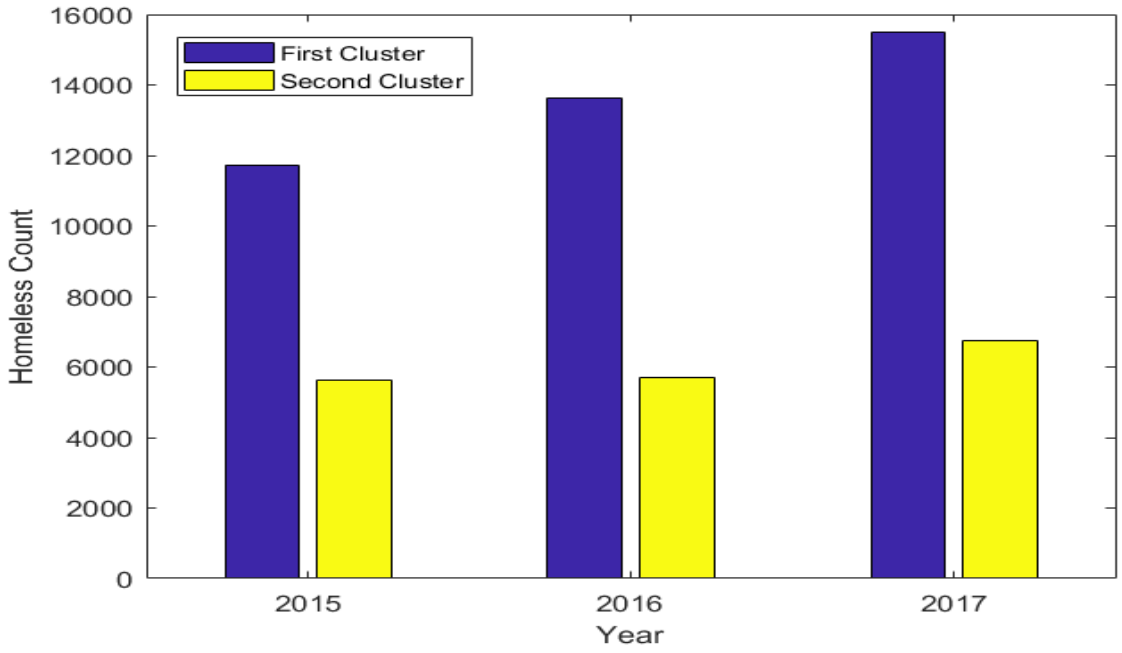


Figure 20a: Total number of homeless population in each cluster from 2015 to 2017

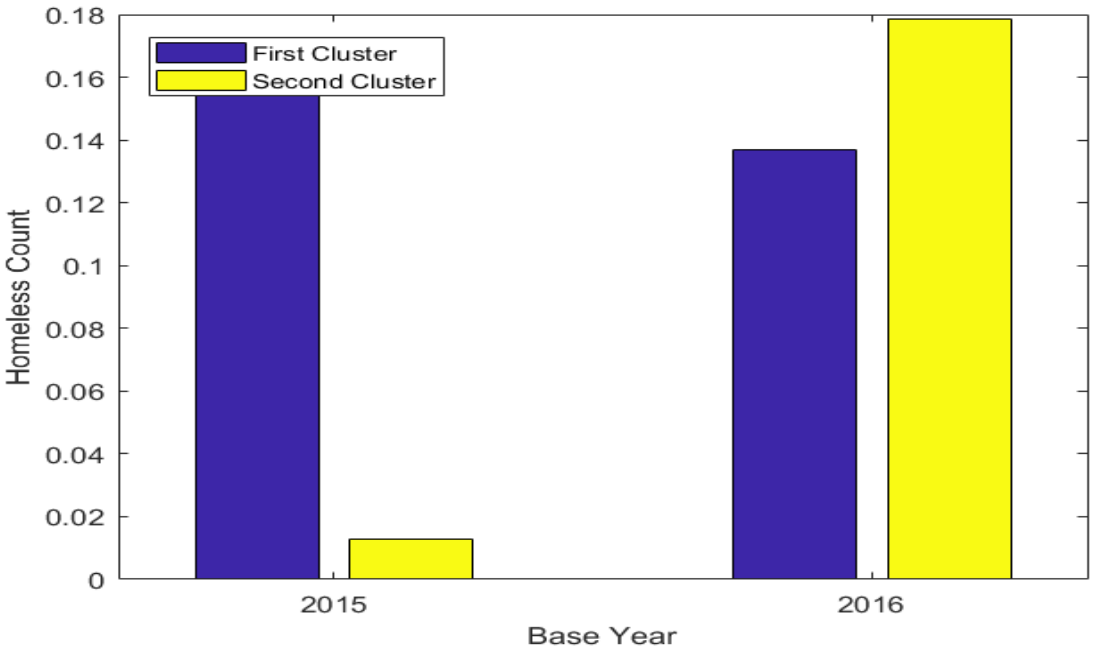


Figure 20b: Percentage changes of homeless population in each cluster from 2015 to 2017

Indeed, from the total number of homeless population chart, the first cluster would encapsulate the highest total number of homeless population, which means that it captures most of the census tracks with smaller number of homeless people, whereas the second one would capture less tracks but with more homeless people, resulting in a lesser total. Perhaps more surprising is the first cluster would at first obtained more homeless people from 2015 to 2016 compared to that of the second cluster, but then the second cluster would capture more homeless people from 2016 to 2017. This might mean that year by year, the census tracks from the first cluster would capture less of the homeless people as the years go on, while the second cluster would increasing attract more homeless people.

The following are the plots of the probability density functions (PDF) for the homeless population for each cluster from 2015 to 2017 for 3 clusters:

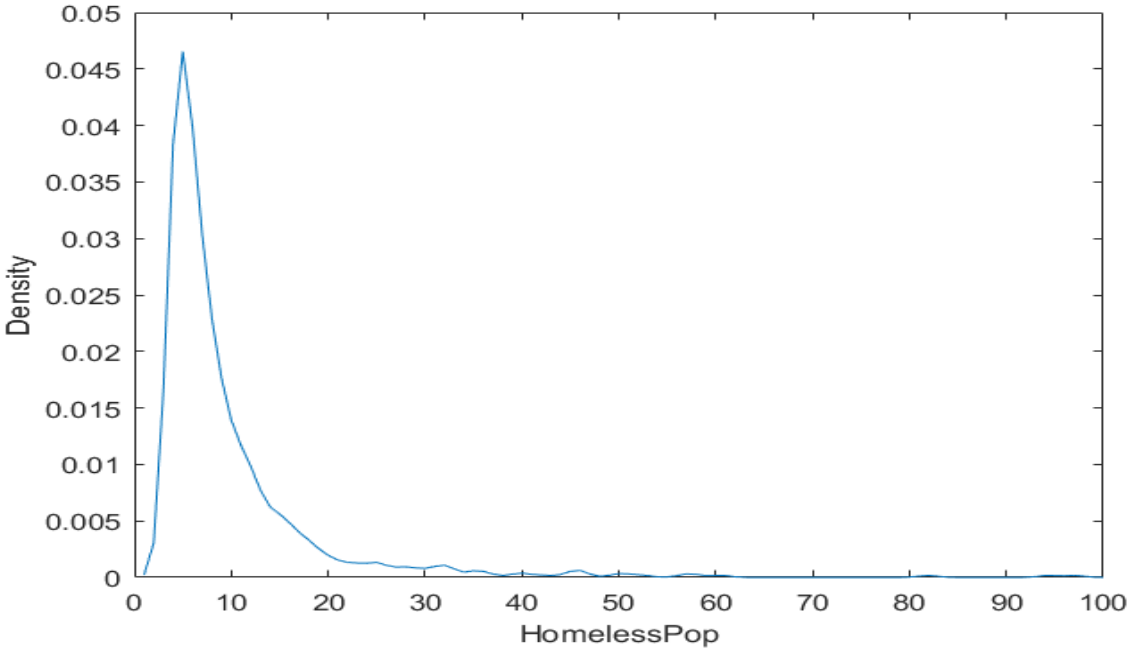


Figure 21a: PDF of homeless population for the first cluster in 2015

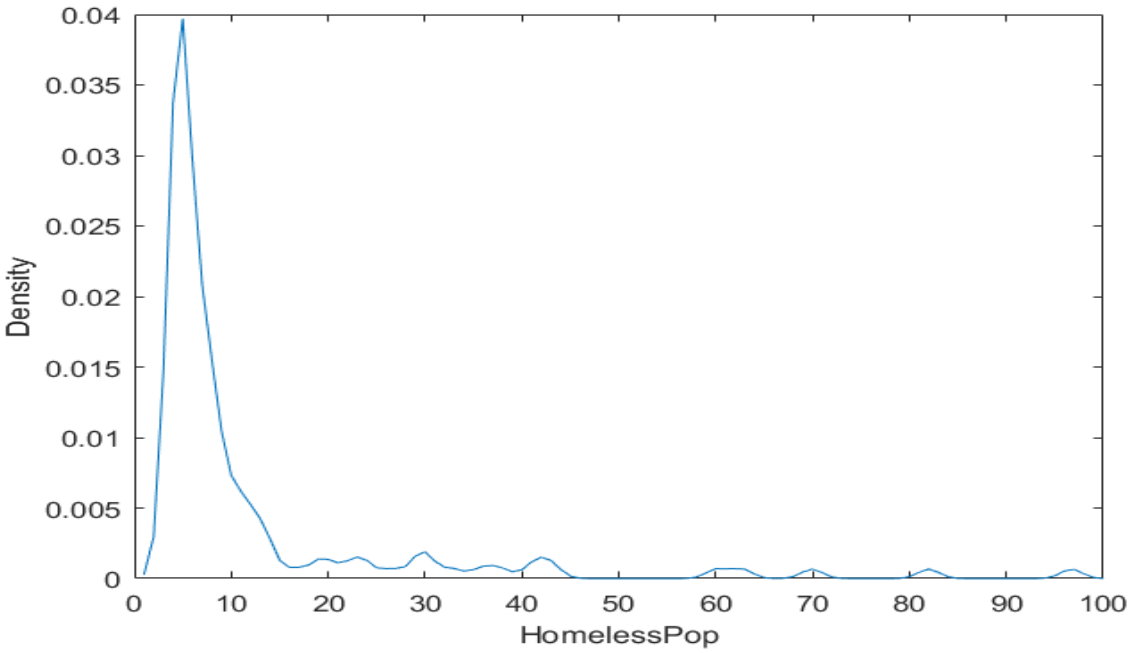


Figure 21b: PDF of homeless population for the second cluster in 2015

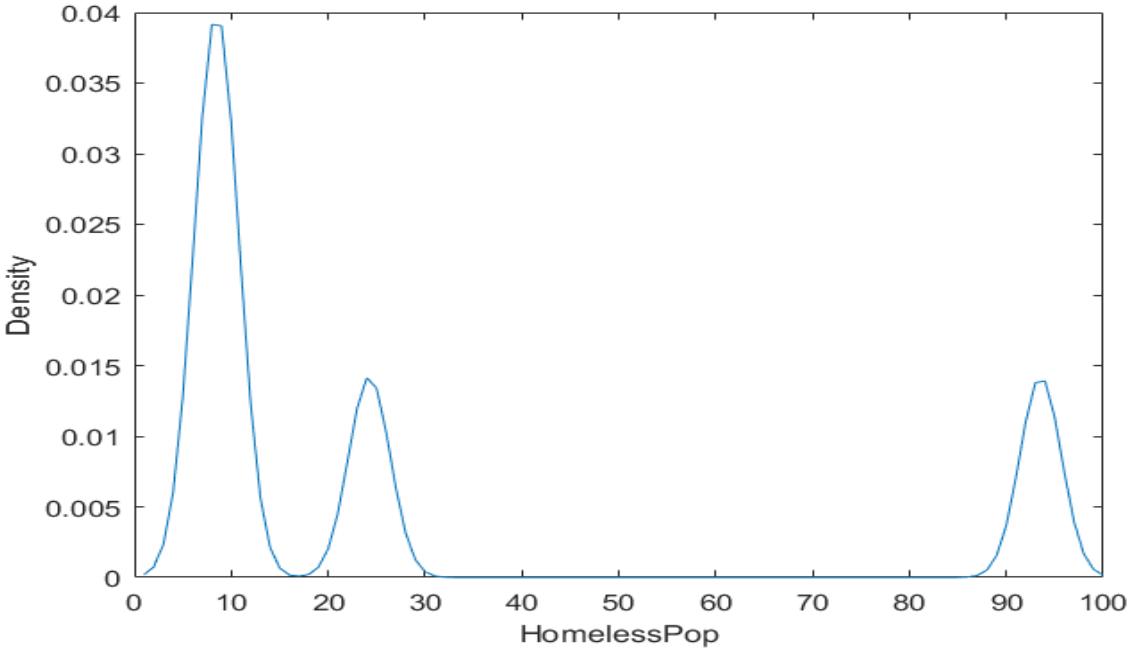


Figure 21c: PDF of homeless population for the third cluster in 2015

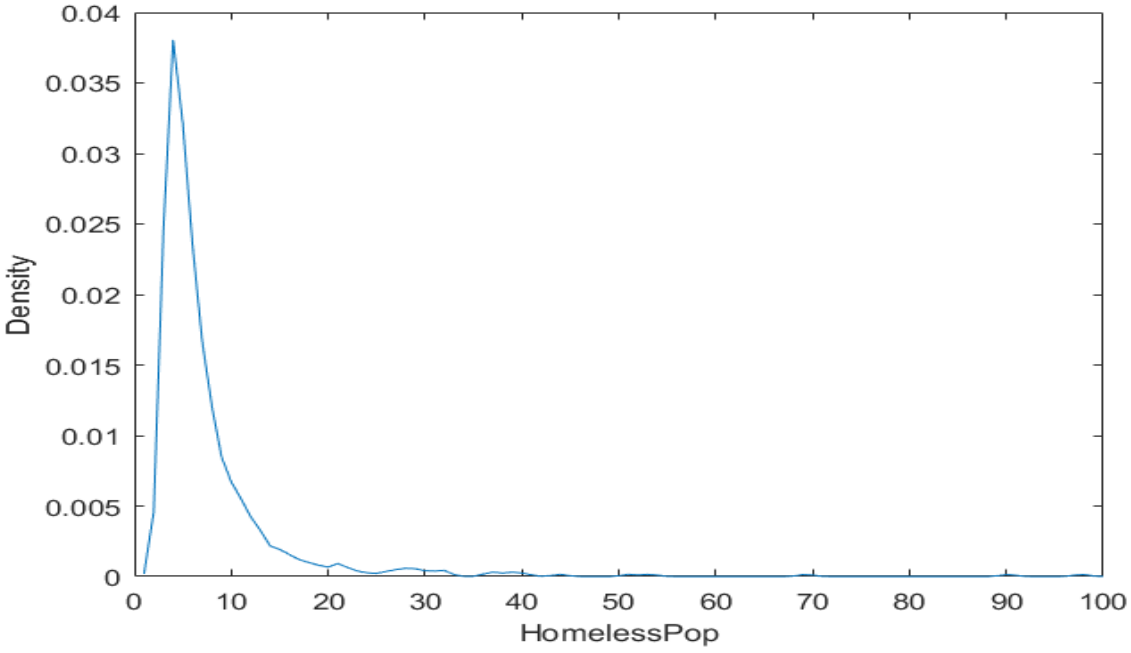


Figure 21d: PDF of homeless population for the first cluster in 2016

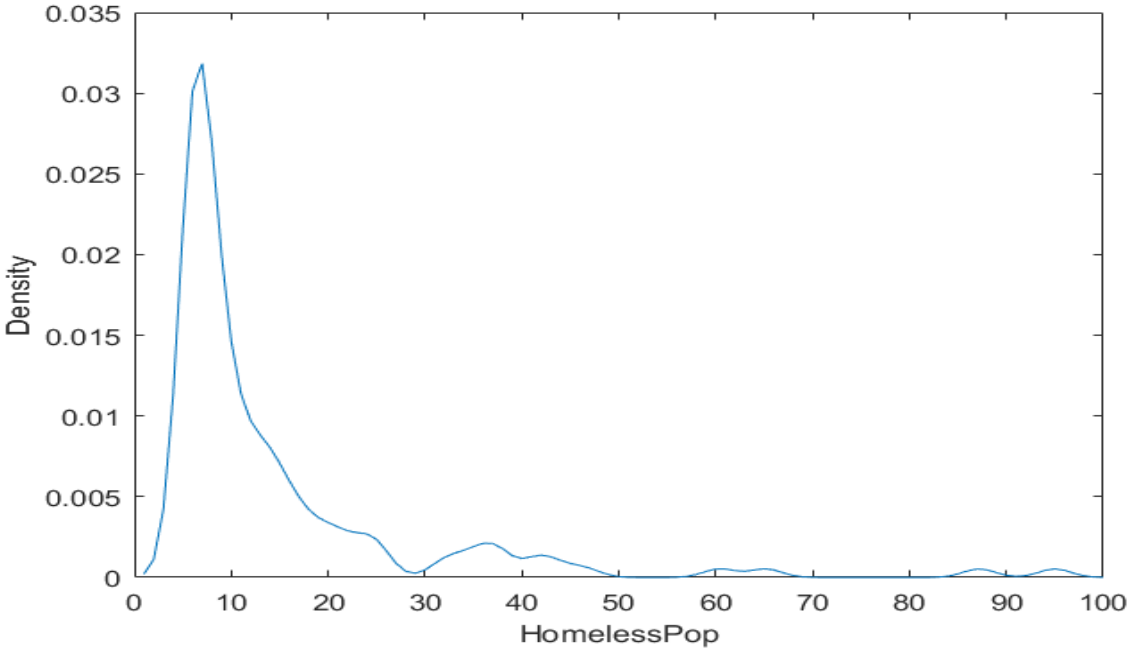


Figure 21e: PDF of homeless population for the second cluster in 2016

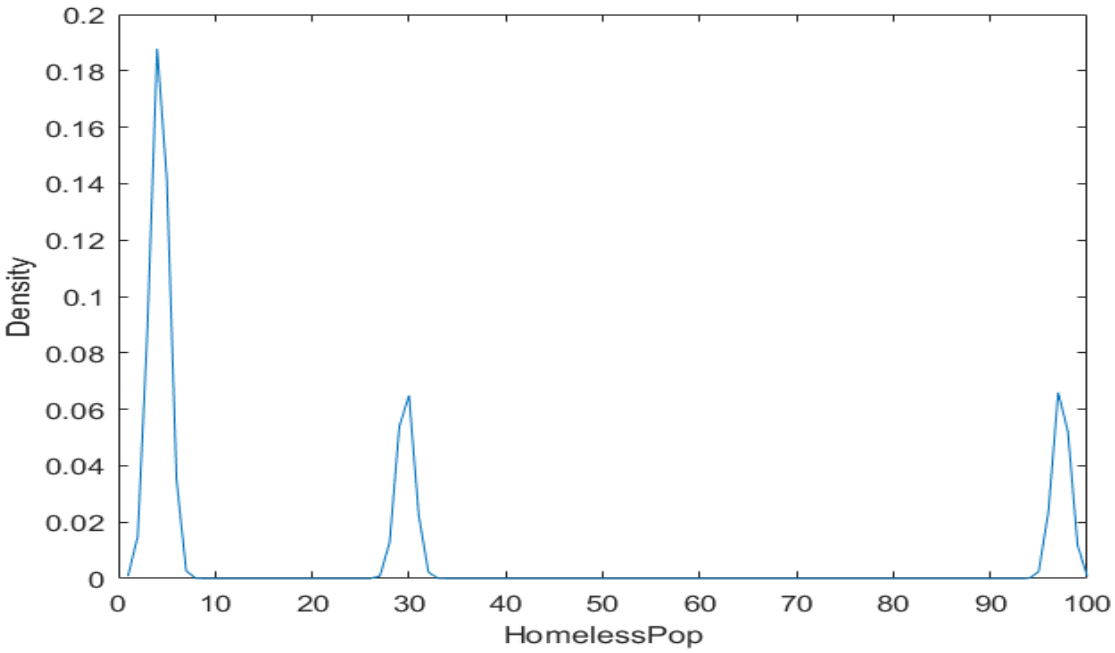


Figure 21f: PDF of homeless population for the third cluster in 2016

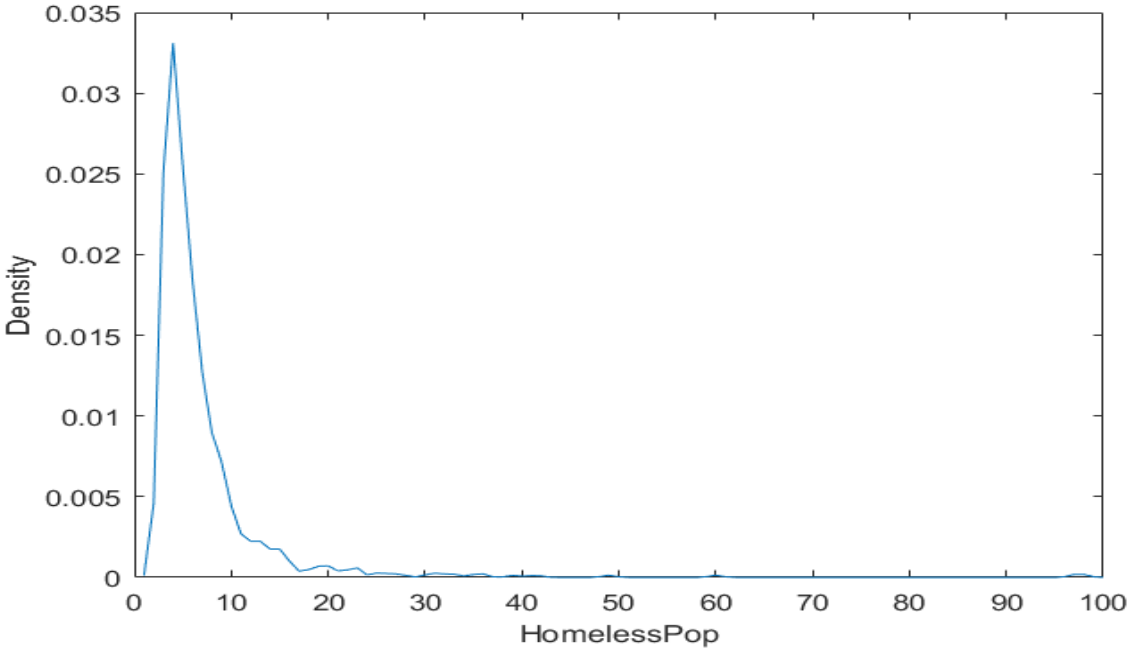


Figure 21g: PDF of homeless population for the first cluster in 2017

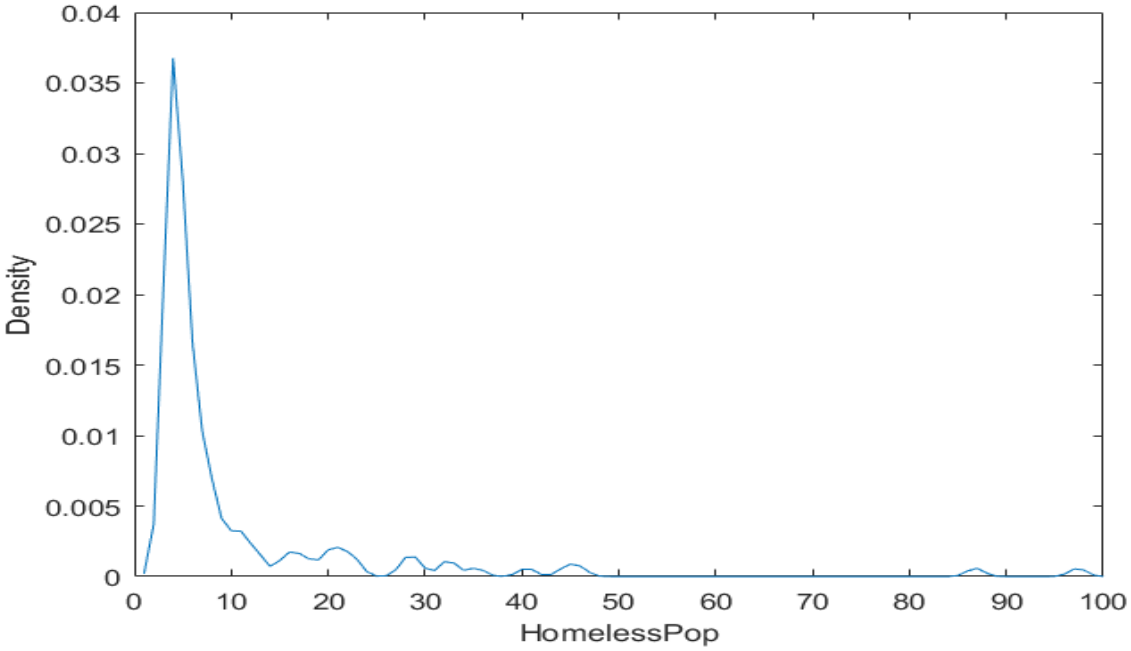


Figure 21h: PDF of homeless population for the second cluster in 2017

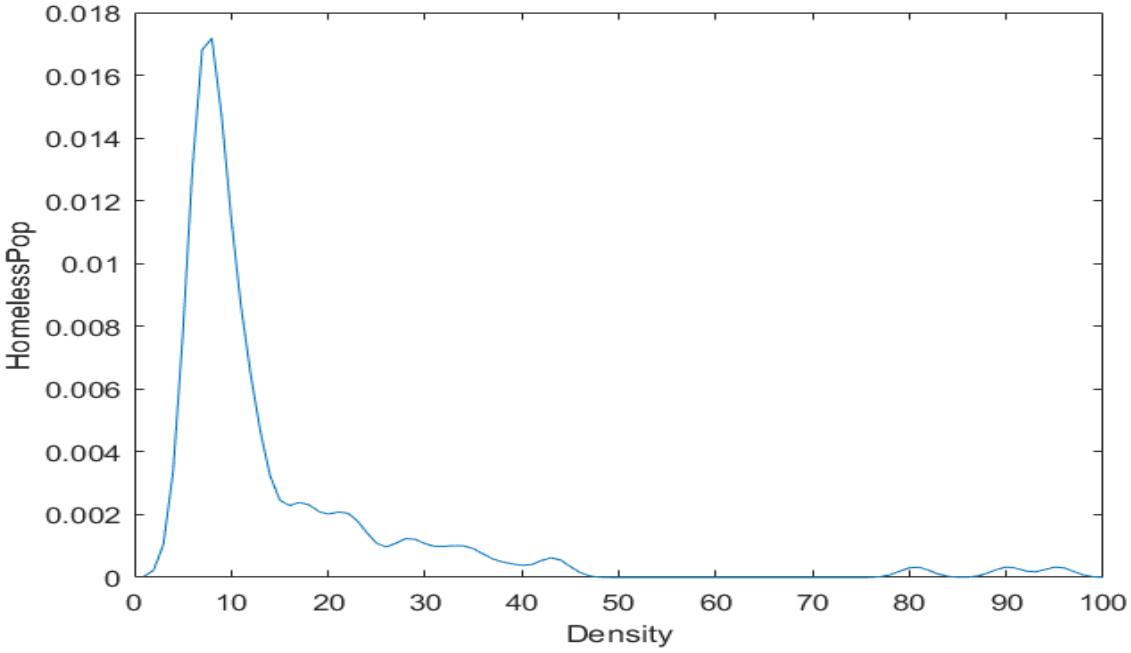


Figure 21i: PDF of homeless population for the third cluster in 2017

Consistently across 3 years, the first 2 clusters would capture most census tracks that have small number of homeless population, though the second cluster would capture more of the tracks with medium number of homeless people, i.e from 20-50 bin. The third cluster would capture 2 peaks of the census tracks, i.e 20-30 bin and 90-100 bin for the first 2 years and capture census tracks with more of the medium number of homeless people for the 2017. To sum up, the first cluster would capture most of the census tracks with small homelesss population, the second would capture also many of the tracks with small homeless population but also many of the tracks with medium number of homeless people, and the third would capture the tracks with high number of homeless population.

The following are bar charts of the total number and percentage change of homeless population in each cluster from 2015 to 2017 for 3 clusters:

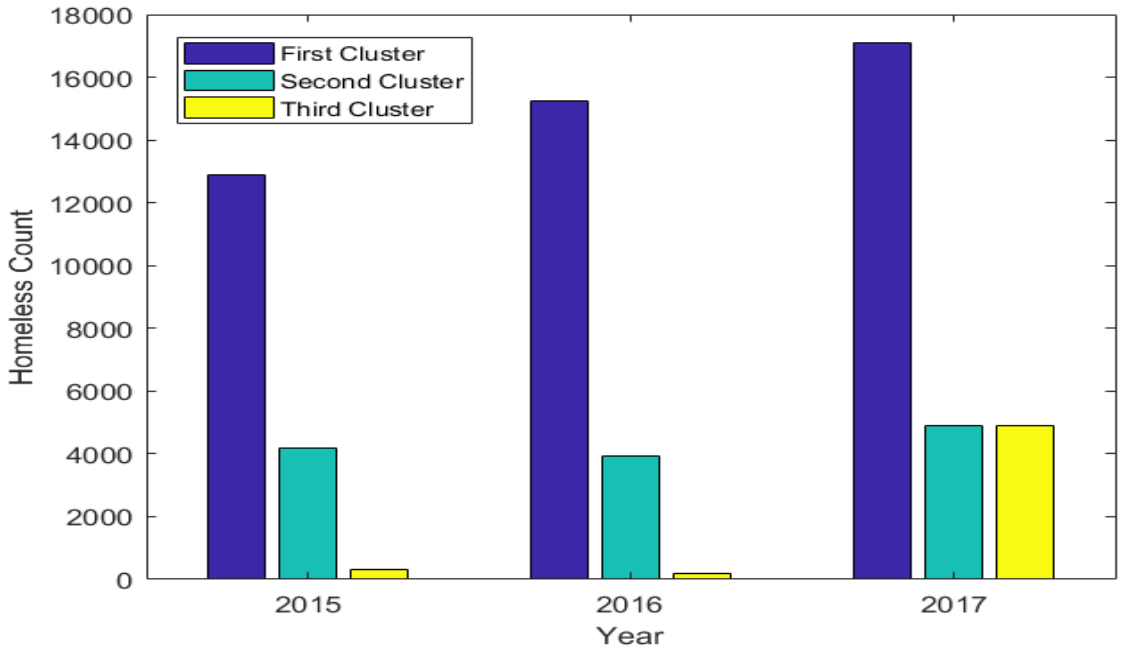


Figure 22a: Total number of homeless population in each cluster from 2015 to 2017

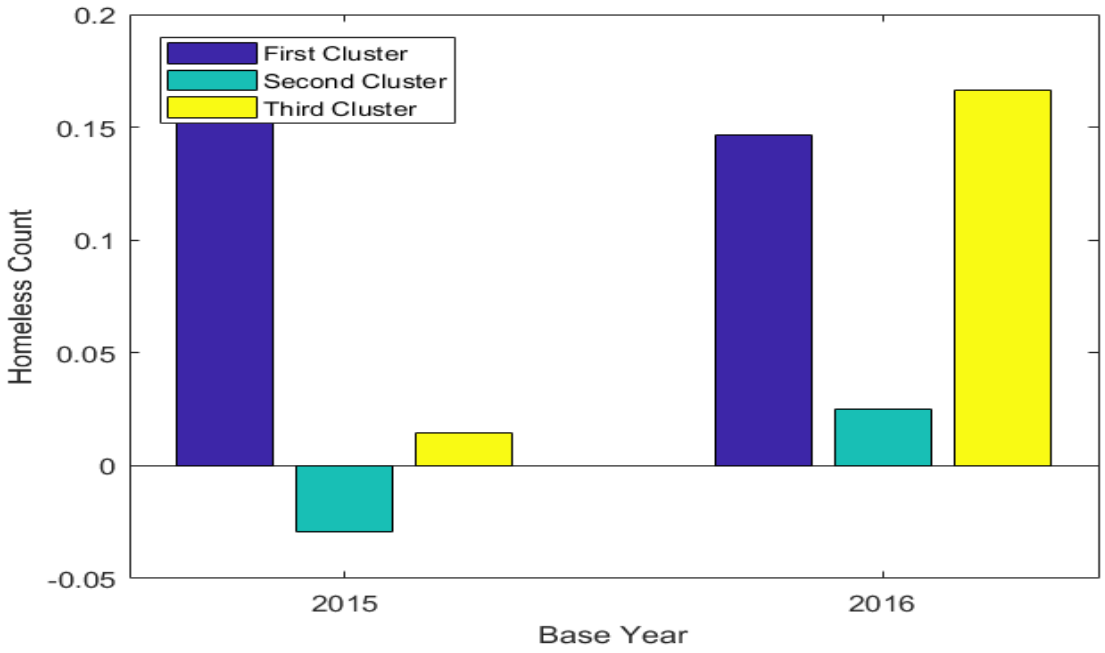


Figure 22b: Percentage changes of homeless population in each cluster from 2015 to 2017

From the plot of the total number of homeless population, the first cluster would capture most of the homeless population in aggregation, the second would capture the medium number of homeless population, while the third cluster would be the smallest. The is also huge recovery in the third

cluster in 2017, which is illustrated clearer in the the plot of the percentage change. This might mean that there is a sudden change in the features in either the first or the second cluster or both that creates a better environment for the homeless people to move from the first 2 clusters to the third cluster.

4.4 Statistical Models

After eliminating predictors that had significant issues of high multicollinearity, and transforming the response variables to achieve constant variance and normally distributed residuals, we identified the statistically significant predictors for the 2017 homeless populations of each sub-category of the greater homeless population.

The statistically significant predictors in explaining variation among the homeless population living in vehicles are a census tract's number of citations, median income, number of bus stops, and total population of people living in street encampments. The Adjusted R-squared for this model explaining variation in vehicle homeless population is 0.434.(see Appendix E Figure 47).

Notably, only for predicting the homeless population living in vehicles is the total homeless population count for previous years not a statistically significant predictor. This is likely due to the mobility afforded to the homeless population that live in cars, which makes it so the location of homeless people who live in vehicles is not particularly tied to any fixed location.

The model for explaining the sheltered homeless population has restaurants, median home income, bus stops, 2015's total homeless population, 2016's total homeless population, and the population of people on street in tents as its statistically significant predictors. The adjusted R-squared for this model is 0.8321. (see Appendix E Figure 43)

The statistically significant predictors for explaining the homeless population living on the street are a census tract's distance to the nearest Trader Joes, distance to the nearest public library, coffee shops, crime count, affordable housing units, bus stops, the 2015 total homeless population the 2016 total homeless population, total number of vans, and the total number of shelters. This model's Adjusted R-squared is 0.7536. (see Appendix E Figure 45)

In modeling the homeless population living on the street, we see proximity to Trader Joes and proximity to public libraries as significant predictors. This verifies the visual pattern that we saw in Figures 2 and 4. However, when we try to explain a change in each category from 2016 to 2017, we no longer see these distance features as significant variables. Similarly, when later models were constructed with additional variables along dimensions of military involvement and regional income, these distance measures became less significant.

Next, we constructed models for the three homeless populations using a forward step-wise regression process. This selection process distills our data down to a select subset of variables. In each iteration, this step-wise process considers a variable for addition from the set of explanatory variables based on some specified criterion. We constructed models using both AIC and BIC as criterions.

The results of our subset selection process provide additional insight into important features for explaining each of these sub-populations. The selected predictors for modeling the homeless population living in shelters when using BIC as the criterion for inclusion are the total number of people living in encampments, the median household income, the number of bus stops, and the number of coffee shops. The adjusted R-squared for the forward step-wise regression model created using BIC as a criterion is 0.4337.

When using AIC to select predictors for the homeless population living in shelters, the number of features selected incerased greatly while the Adjusted R-squared only slightly increased to 0.4348. The selected predictors when using AIC expanded to include the total number of people living in

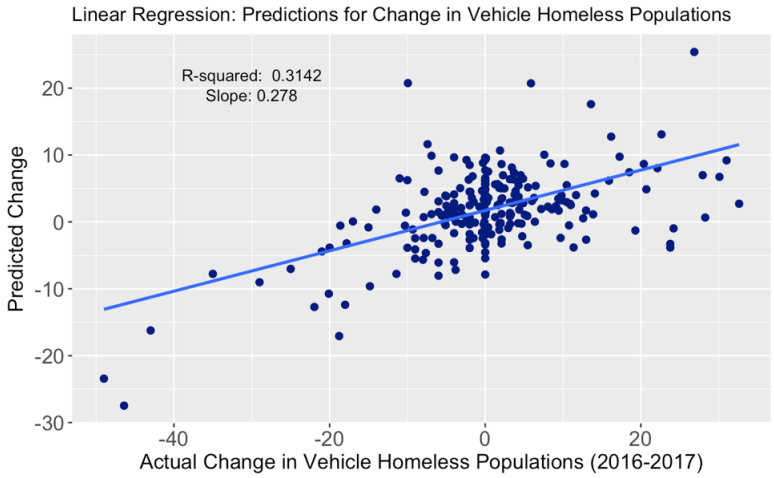
campers, the total number of people living in vans, the total number of people living in cars, the median household income, the 2015 housed population, the distance to the nearest public library, the number of bus stops, the number of coffee shops, the distance to the nearest Trader Joes, the 2015 Zillow Rent Index sore, and the 2017 Zillow Rent Index Score. We see that using AIC as a method for feature selection includes more features than when using BIC. This is due to the fact that BIC has a greater penalty for adding additional variables.

The selected predictors for modeling the homeless population living on the street using AIC as a criterion are the 2016 total homeless population, the total number of people living in vans, the crime count, the 2015 Housed Population, the number of bus stops, the number of affordable housing units, the total number of sheltered people, the number of affordable housing units, the 2015 total homeless population, the distance to the nearest public library, the number of coffee shops, and distance to the nearest Trader Joes.

The selected predictors for explaining the variation in the homeless population living on the street using forward step-wise regression with BIC as the criterion are 2016's total homeless population, the total number of people living in vans, the crime count , the 2015 housed population, the number of bus stops, and the number of affordable housing. Again, we see the model constructed using BIC includes fewer predictors than the model constructed using AIC because of BICs greater penalty for over fitting.

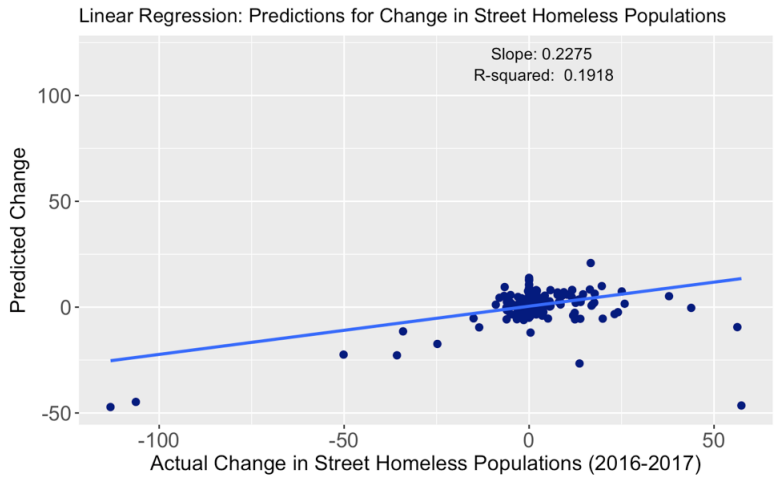
In constructing regression models for the *changes* in the respective populations from 2016 to 2017, we were able to identify important predictors for the changes as well as compare our model's predicted changes to the actual changes observed in the data.

Figure 23: Predicting the Change in Vehicle Homeless Populations from 2016 to 2017



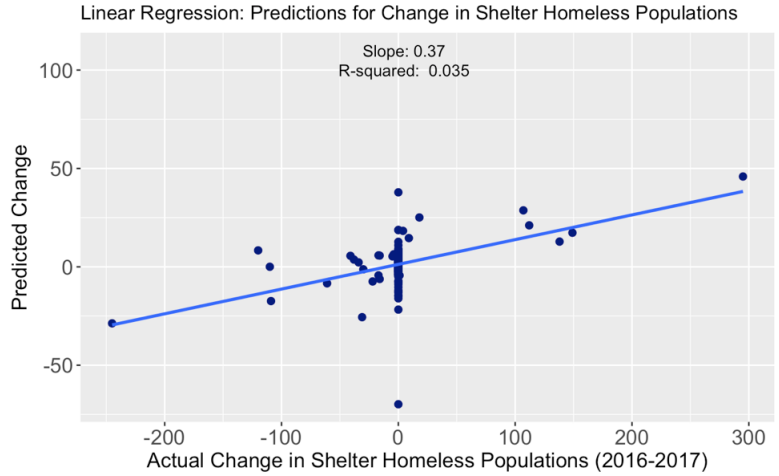
The variables that are significant in explaining the change in the vehicle homeless population from 2016 to 2017 are a census tract's number of bus stops, bus stop density, 2015 total people living in encampments, 2015 emergency sheltered homeless people, 2016 total population living in tents, 2016 total population living in tents, 2016 total people living in encampments,2016 total unsheltered population, 2014 total vacant units, and 2015 median rent.

Figure 24: Predicting the Change in Street Homeless Populations from 2016 to 2017



The variables that are significant in explaining the change in the homeless population living on the street are a census tract's 2015 general population density, 2015 total people living in tents, 2015 total people living in encampments, 2016 total people living in encampments, the 2015 percent of non-active military residents,the number of citations, the crime count, the number of affordable housing units, the density of affordable housing units, and the 2016 population in 1 mile.

Figure 25: Predicting the Change in Shelter Homeless Populations from 2016 to 2017



As we see above, the model predicting the change in the sheltered homeless population category does poorly; predicting increases and decreases in the population when the sheltered population does not change. Figure 25 serves to show that census tract sheltered homeless populations are rather invariable. We expect this may be due to how shelters are confined to certain capacities, and census tracts tend to have a fixed number of shelters.

The results of the random forest regression provide further insights into the particular census tract features that explain each category of a census tract's homeless population.

Figure 26: Random Forest Regression Predictions for the 2017 Street Homeless Population

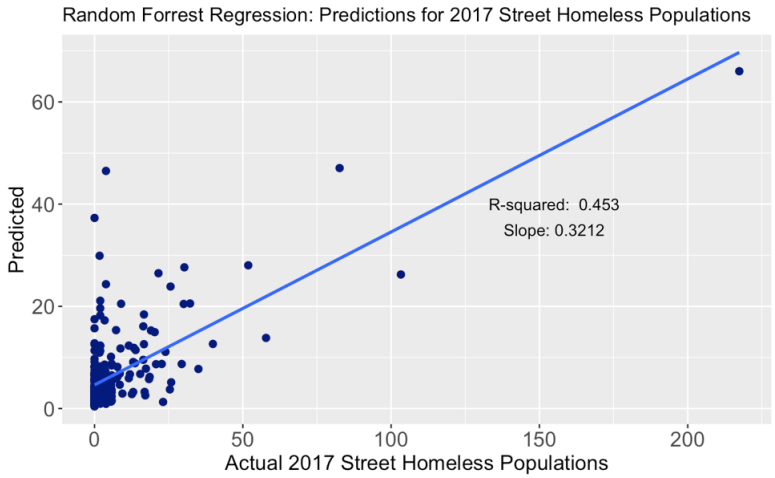
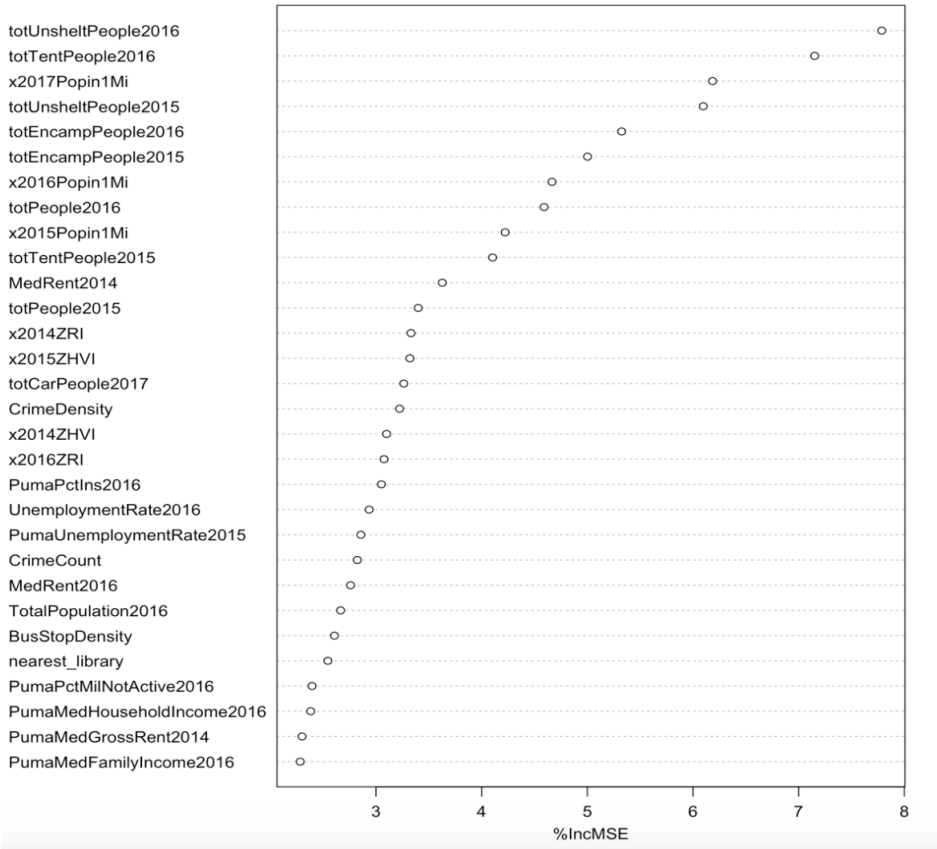


Figure 27: Random Forest

Important Features Explaining 2017's Homeless Population Living on the Street



In the plot above, we see that the top two most important features for explaining the 2017 street homeless population category are both measures of the previous year's population living on the street: 2016 unsheltered population, and 2016 number of people living in tents. This suggests

that considering the current population of people living on the street can be helpful in understanding what the next year's population will be. This may be because homeless people living on the street have less effective means of transportation and therefore have more fixed locations across time. These important features for the street homeless population are very much in line with what we would expect. However, in comparing these important features to those of the vehicle and shelter homeless population categories, we begin to see some fascinating differences as to what different features influence these three categories of the homeless population. Regarding our initial questions surrounding whether variation in a category can explain variation in another category, the street homeless population category is unique because it is the only category to become a top important feature in explaining another category. Below, we see that the total number of people living on the street in encampments in 2017 (a measure falling under this street homeless category) is the top most important feature in explaining the 2017 homeless population living in vehicles (Figure 29). This makes the street homeless category the only category that serves as an important feature in explaining another category.

Figure 28: Random Forest Regression Predictions for the 2017 Vehicle Homeless Population

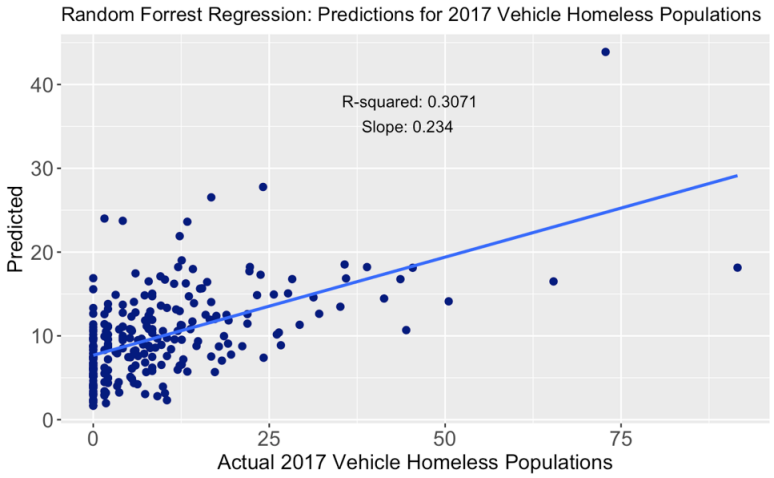
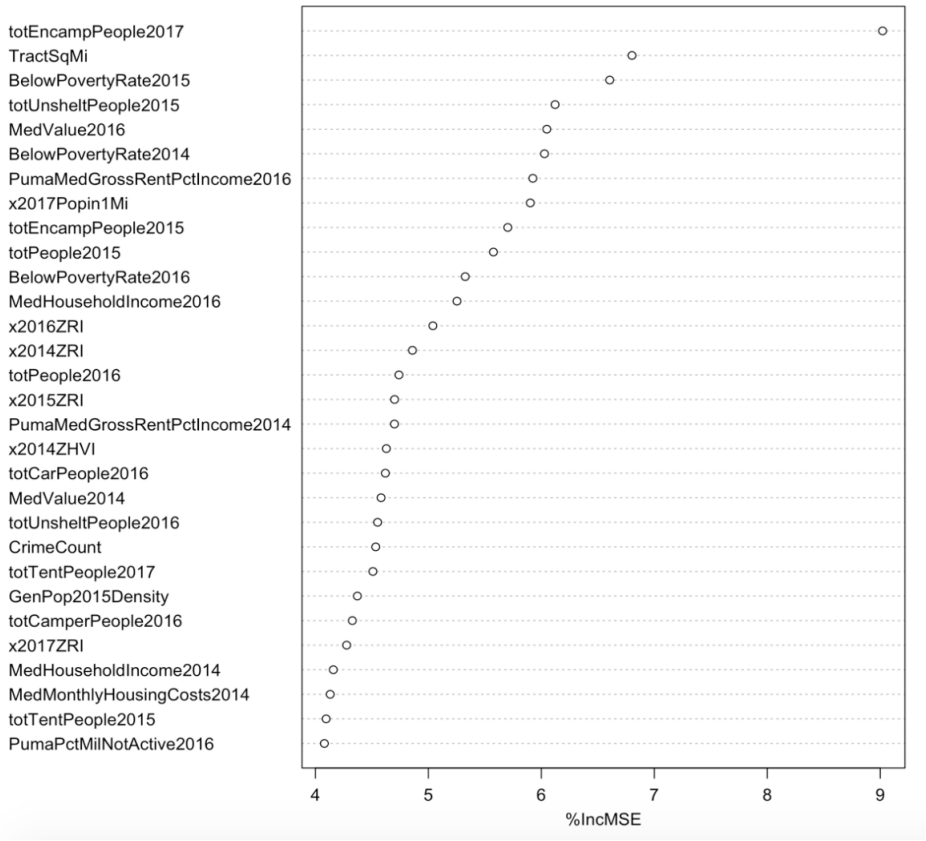


Figure 29: Random Forest

Important Features for Explaining 2017's Vehicle Homeless Population



In considering the ranking of important features (Figure 29) for the category of the homeless population living in vehicles, the top three most important features are the current year's population of people living in street encampments, the census tract's square millage, and the census tract's below poverty rate. It is important to note that unlike the street homeless population category, previous years' measures of this category's population are not important for explaining itself in 2017. This is unique only to the vehicle homeless population category, as we have seen that previous years' street homeless population measures are important for 2017's street homeless population, and we will also soon see that previous years' shelter homeless population measures are important for 2017's shelter homeless population category. In addition to measures of the vehicle homeless population category not being important features for explaining the category in a subsequent year, this category's important features are also unique because the total overall homeless population of the previous year (2016) drops to being just the fifteenth most important feature. This is in contrast to its position as the second most important feature for the street homeless population category, and its position as the ninth most important feature for the shelter homeless population category. Perhaps the mobility afforded to homeless people living in vehicles can explain these differences, as their convenient means of transportation does not require or limit their distribution to be fixed in a location across time.

Figure 30: Random Forest Regression Predictions for the 2017 Shelter Homeless Population

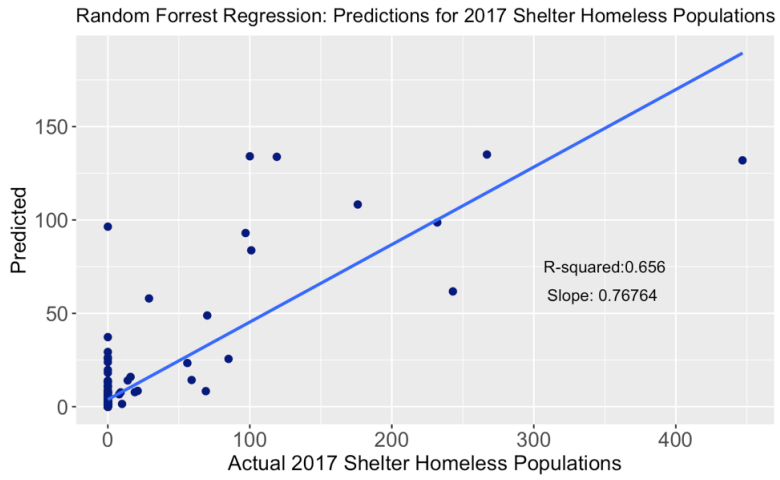
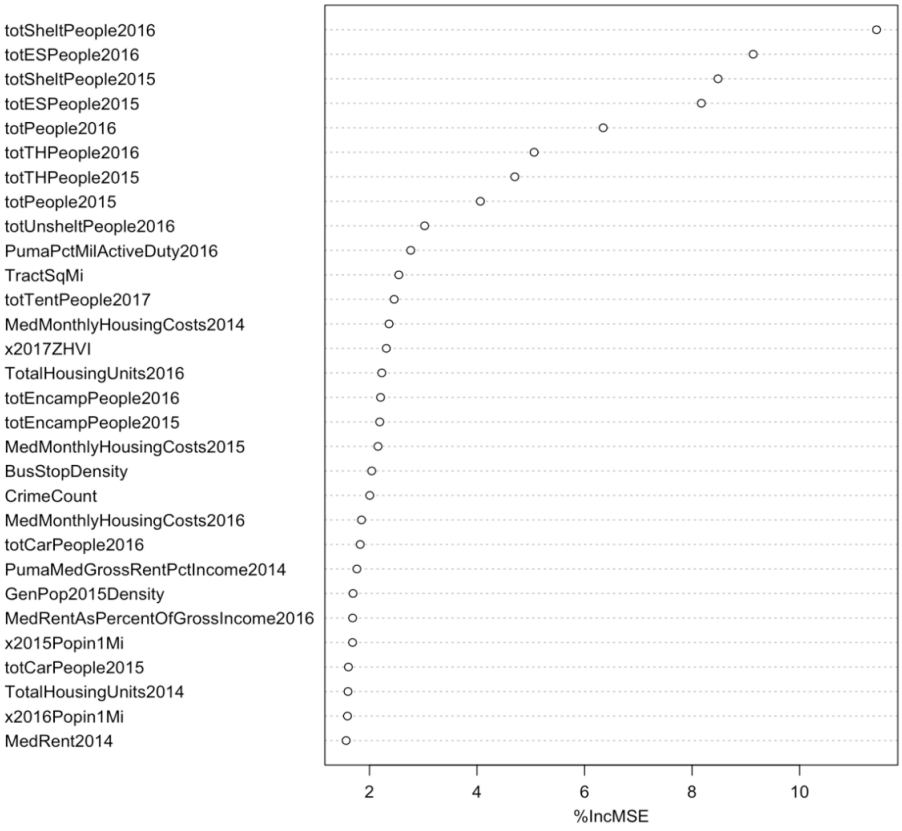


Figure 31: Random Forest Regression

Important Features for Explaining 2017's Shelter Homeless Population



The important features for explaining the shelter homeless population category is particularly interesting because it is the only time that we see a census tract's sheltered homeless population as a top 5 most important predictor. This suggests that while previous years' sheltered homeless

populations certainly explain the subsequent year's sheltered homeless population category, the sheltered homeless population measures do not contribute as much for explaining the other categories. This isolates the sheltered homeless population category from the other categories which do serve in explaining one another.

4.5 Neural Networks

We used various measurements to reflect how well the algorithm did. For the classifier models, apart from the accuracy, we have confusion matrix and precision [32]. Accuracy is calculated by:

$$A = \frac{\#\text{ of correct predictions}}{\#\text{ of all predictions}} \tag{32}$$

A confusion matrix C is a table that shows the performance of a classification model. The (i, j) entry of the matrix represents the number of points whose true value is the i^{th} class but it was predicted to be in the j^{th} class. Precision i is the ratio of true predictions among all the points that were predicted to be in the i^{th} class:

$$\text{Precision } i = \frac{C[i,i]}{\sum_j C[i,j]} \tag{33}$$

We took the S generated by different λ of the exponential distant decay function and fed them into the best ternary classifying neural networks that predicts the changes in total homeless population (details about the model can be found below). We took the average accuracy of 10 runs of each λ and generated the graph below.

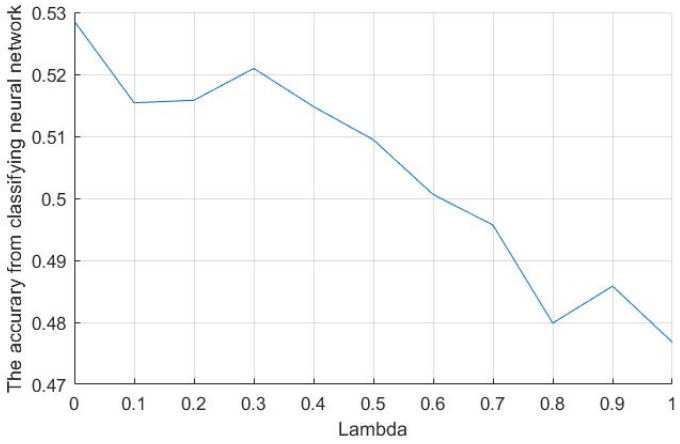


Figure 32: The accuracy of the model as λ changes in the exponential distance decay function.

There is an overall trend of decrease in accuracy as the value of λ increase and the best accuracy is achieved when λ is 0. This shows that adding local features is not helpful for our data set and our purpose of predicting homeless population changes. Therefore all the neural networks models are constructed with the original data set without considering the local surrounding features of the census tracts ($\lambda = 0$).

The best result of the ternary classification for changes in: total homeless population.
The table below shows the best result of the ternary classification for bucket division: (-∞,-3.1), [-3.1,5.5), [5.5,∞).

Model	Hidden	Output	Accuracy	Training Cost	Testing Cost
A	-	Softmax	0.54	0.93	1.11

With a confusion matrix and precision:

	True -	True 0	True +
Pred -	105	37	66
Pred 0	33	112	52
Pred +	20	15	47

	Precision-	Precision0	Precision+
Model	0.50	0.57	0.57
Chance	0.33	0.33	0.33

The precision of our model is higher than each of it's corresponding entries in the chance model. Therefore our model performs better than chance.

The best result of the ternary classification for changes in: homeless population on the streets. The table below shows the best result of the ternary classification for bucket division: $(-\infty,-1.9)$, $[-1.9,1.95)$, $[1.95,\infty)$.

Model	Hidden	Output	Accuracy	Training Cost	Testing Cost
B	-	Softmax	0.58	0.86	0.98

With a confusion matrix and precision:

	True -	True 0	True +
Pred -	101	14	32
Pred 0	31	111	64
Pred +	29	37	69

	Precision-	Precision0	Precision+
Model	0.69	0.54	0.51
Chance	0.33	0.33	0.33

Our model performs better than chance. The precision of the lowest bucket is much better then the other two; this means our model is more accurate at predicting homeless population decreasing than homeless population staying static or increasing.

The best result of the ternary classification for changes in: homeless population in vehicles. The table below shows the best result of the ternary classification for bucket division: $(-\infty,-1)$, $[-1,3.5)$, $[3.5,\infty)$.

Model	Hidden	Output	Accuracy	Training Cost	Testing Cost
C	Leaky ReLU 6 hidden neurons	Sigmoid	0.56	0.83	0.98

With a confusion matrix and precision:

	True -	True 0	True +
Pred -	125	26	58
Pred 0	15	101	55
Pred +	21	39	47

	Precision-	Precision0	Precision+
Model	0.60	0.59	0.43
Chance	0.33	0.33	0.33

Our model performs better than chance. The model predicts that the lower two buckets much more accurately than the highest one; this means our model is better at predicting homeless population decreasing and staying static than homeless population increasing.

In fact, if we only give the neural networks the subset of data that is in the decreasing and static buckets of the ternary classifier (in the range of $(-\infty, 3.5)$) and run a binary classifier on the data with bin division $(-\infty, -1)$, $[-1, 3.5)$, the result is below.

Model	Hidden	Output	Accuracy	Training Cost	Testing Cost
D	Leaky ReLU 6 hidden neurons	Sigmoid	0.79	0.41	0.54

With a confusion matrix and precision:

	True -	True 0
Pred -	124	26
Pred 0	15	101

	Precision-	Precision0
Model	0.79	0.78
Chance	0.5	0.5

The model performs significantly better than the corresponding ternary classifier with an accuracy of near 80%. This shows our model can better find the patterns in the data set consisting only decrease and static changes of homeless population in vehicles. This in turns shows the features we have correlates better with the decrease and static changes of homeless population in vehicles and worse with the increase in homeless population in vehicles.

This result can be further be corroborated by the neural networks regression models. To measure the performance of the regression model, we looked at the scatter points graph of all the data points' predicted value versus the actual value. We also found the best fit line of the scatter points and the slope and R^2 of the best fit line.

The best result for predicting the changes in homeless population in vehicles in shown below.

Model	Hidden	Training Cost	Testing Cost
E	Leaky ReLU 3 hidden neurons	62	164

Figure 33: Predicted versus Actual for the regression model E

The slope of the best fit line for the predicted versus actual scatter point is of slope 0.34 and has an R^2 of 0.24. The result is close to the result of the linear regression model. This is because we used a relative shallow neural networks model to prevent over-fitting. The model can learn to predict the changes in homeless population in vehicles with relative success. However, if we only give the model data whose changes in homeless population in vehicles are within the range of $(-\infty, 3.5)$, the model can learn to predict the changes much better.

Here is the best model for predicting changes of homeless populating in vehicles within the range of $(-\infty, 3.5)$.

Model	Hidden	Training Cost	Testing Cost
F	Leaky ReLU 3 hidden neurons	9	28

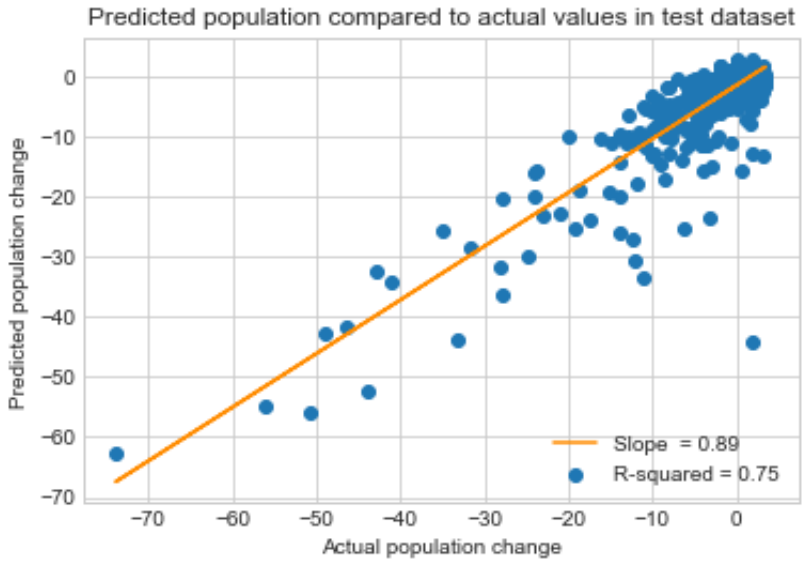


Figure 34: Predicted versus Actual for the regression model F

The slope of the best fit line for the predicted and actual scatter point is of slope 0.89 and has a correlation of 0.75, much better than the previous model. The model can learn to predict the decrease and static changes in homeless population in vehicles much better, confirming our result of the classifier model.

4.6 Earth Mover's Distance

The Earth Mover's distance produced results that were as expected with the cost metrics we used. In our minimized flow matrix, f, most values for flows between census tracts were located near the diagonal. This is due to the fact that census tracts along the diagonal are closer in proximity to one another and therefore have a smaller cost associated with moving there.

The flow using figures from 2015 produced insensible results. We saw migratory networks that spanned far distances across Los Angeles. This may be due to the missing data in 2015 that we interpolated with census tract data from 2010. However, results from 2016 and 2017 showed desirable results. We saw more migratory patterns and networks within the peripheral of Los Angeles and a more concentrated and dense pattern of movement in urban centers.

In the following maps, increased movements of homeless individuals are denoted as circles that are proportional to the growth in homeless population from the following year. Circles of the same color denote a census tracts that saw multiple movements of homeless populations from neighboring tracts, either in or out of the particular tract.

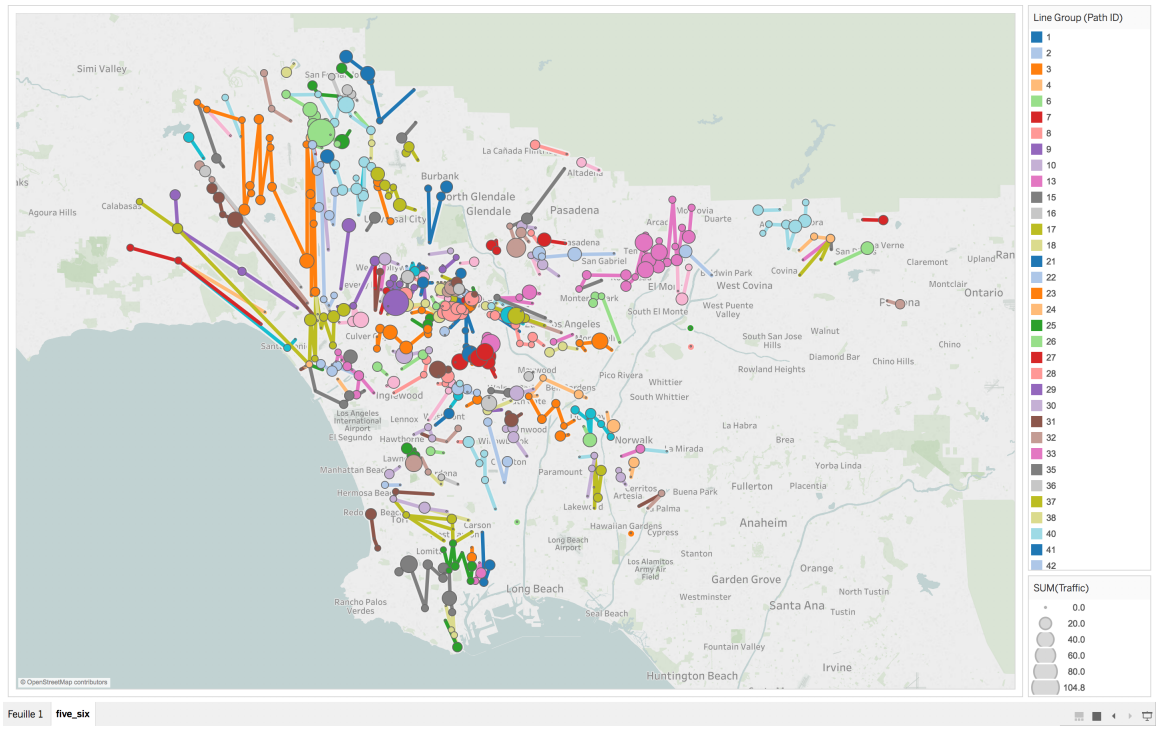


Figure 35: Haversine distance 2015 to 2016

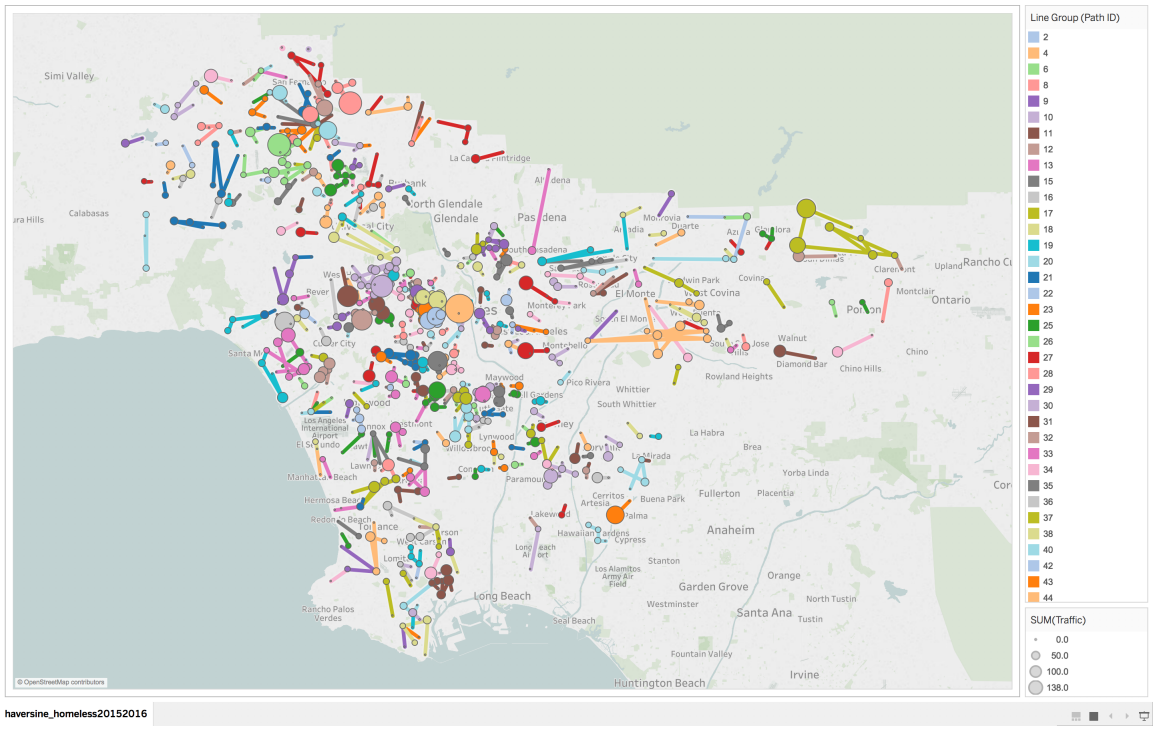


Figure 36: Haversine distance 2016 to 2017

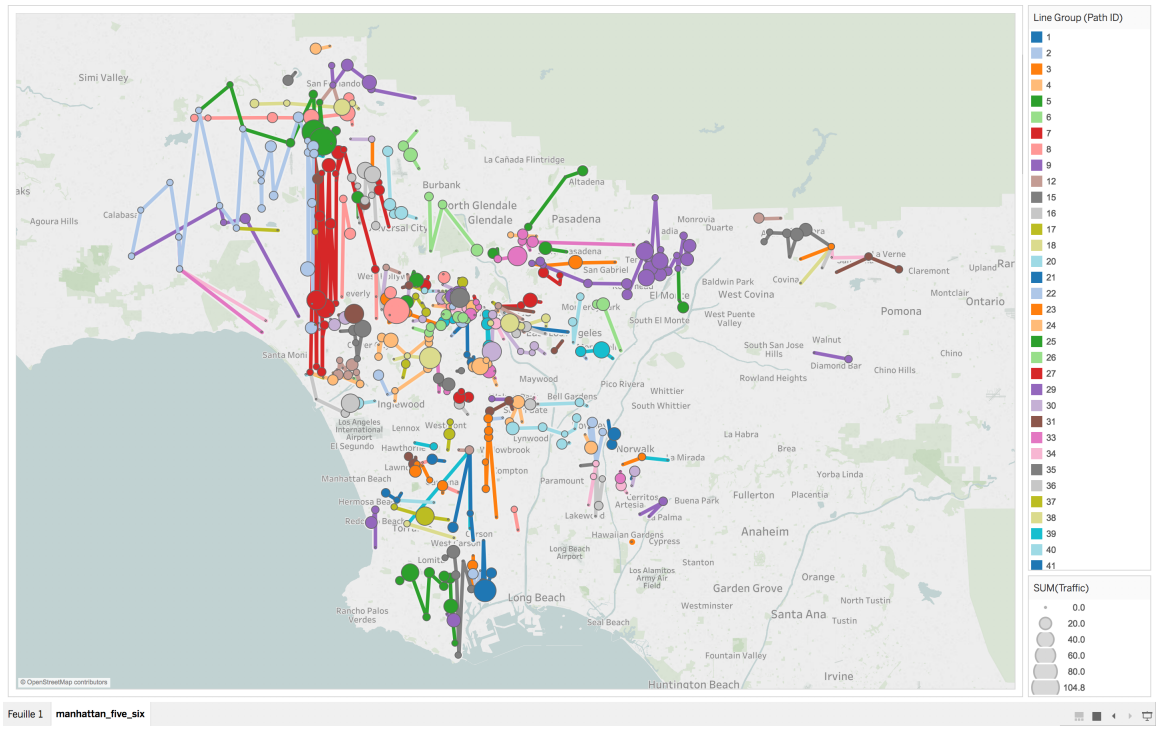


Figure 37: Manhattan distance 2015 to 2016

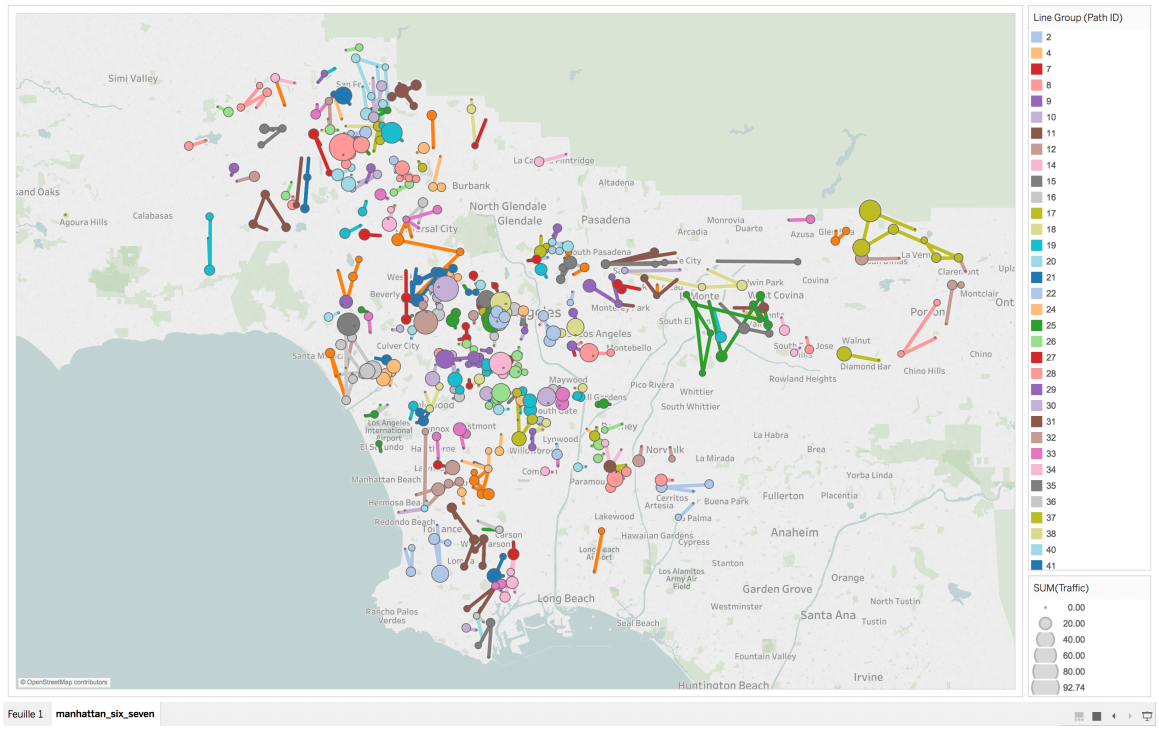


Figure 38: Manhattan distance 2016 to 2017

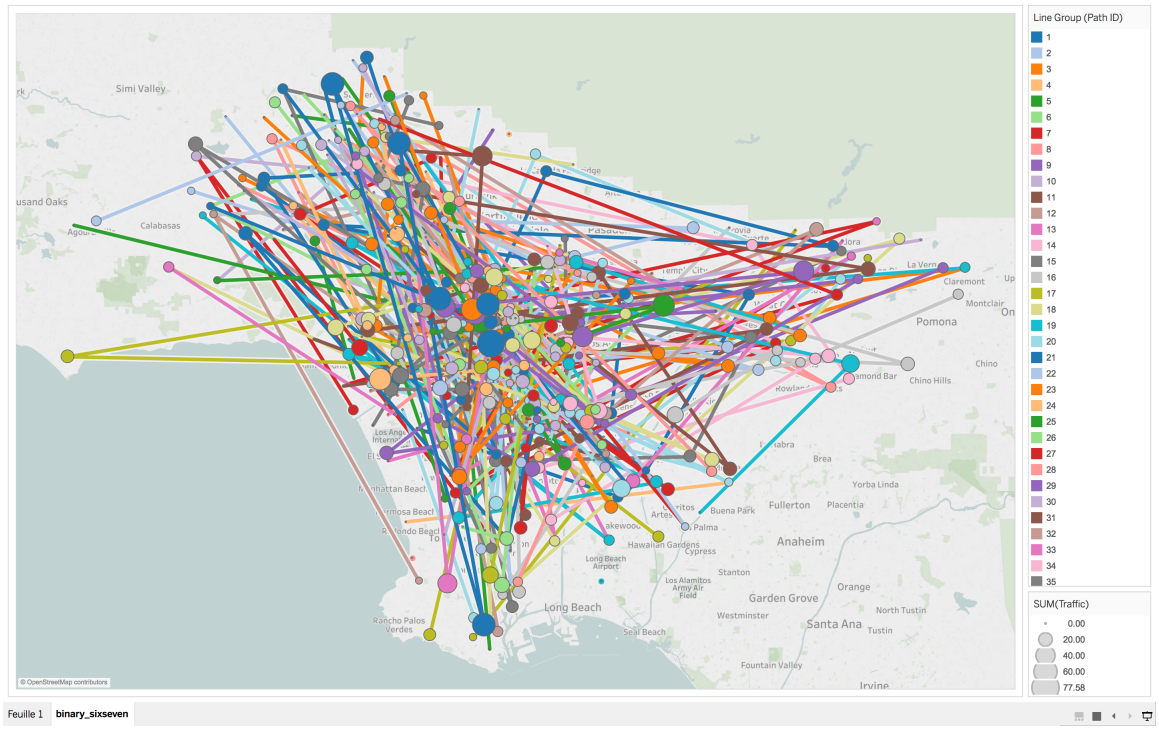


Figure 39: Binary cost 2016 to 2017

As you can see, the results from 2015 show long stretches of movement patterns across populations. Binary distance proved to be an inaccurate measure of cost since a homeless person could move to the neighboring census tract or all the way across Los Angeles at the same cost. The best results are using both the Haversine and Manhattan distance metrics on homeless counts from 2016 to 2017.

5 Summary

Though we have some understanding as to how some features or predictors affect homelessness in Los Angeles, there are still many aspects of homelessness in the census tracks that require further scrutiny.

As mentioned in data management section, there were many errors or guesstimates being made on the data that we collected, which might present a less accurate description of homelessness. Some of the feature's sample size is also too small, which might introduce bias when doing analysis, especially when normalization is done. Choosing features that are complete, accurate and relevant to homelessness should then be our first priority to improving results.

Topic modelling allows us to reduce dimensions to focus only certain features so that we could look into how they could possibly influence homelessness. By looking at the plots and heat maps of weights of each topic, we could see some form of clustering in certain areas of Los Angeles which is indicative of homelessness hotspots, especially after comparing it with the heat map of homeless population density and realize that some clusters overlap quite conveniently. However, the linear correlation between the topics and homeless population should not be a decisive measure of how topics could explain homelessness, as there might be some non-linear dynamics that we have not explored between the topics and homelessness.

Local variance compares the total variance of the features in the neighbourhood of a census track with that of other census tracks to infer whether there is a connection with homelessness. The local variance analysis thus manages to pick up the features that seem to be consistent with those from the highest positively correlated topic with homelessness. However, though the analysis provides intuitive inference, it might be too simple and thus naive to make broad and decisive conclusions.

Clustering the data with the mixture of Gaussians via expectation-maximization method provides us sound results as to how the features in the census tracks could cluster the tracks relating to homelessness. The clustering not only provide more assurance to the results obtained from both the topic modelling and local variance, it also provides even more insights the inner dynamics of homelessness on the tracks, such as how there seems to be difference in the growth rates of homeless population in each cluster, which allows us to understand that changes throughout the years in the census track would dramatically change the composition of homeless population in the track. However, the method is again not without its faults, as we only choose a very small subset of the features that we have to conduct the clustering, specifically choosing features that are highly correlated to homelessness, which we would expect to yield very similar results to both local variance and topic modelling. Moving forward, the clustering could further improved by first having a better method to choose features that provide a more encompassing analysis on the the census tracks. We also might need to choose a better clustering method to cluster our noisy data. This is because Gaussian mixture presumes that there seem to be clusters with the same mean and covariance, but when relatively less census tracks (818 of them to be precise) is used for this analysis, the sample mean might not approximate well to the population mean by the law of large numbers. This means that there might exist bias in our results.

Neural Networks models can predict the changes in total homeless population and subcategories of homeless population (homeless population on the streets and in vehicles) with accuracy better than chance. We also found that decrease and static change of homeless population in vehicles can be predicted with better accuracy by the neural networks. Neural Networks is still a field that is being developed and researched. More techniques that could help neural networks models to perform better are being published. Therefore apart from gathering more data containing information that could predict the changes in homeless population, future research should explore methods and techniques in current literature that could improve the performance of the neural network models predictions.

The statistical models give us insight into how the categories of the homeless population are related. While the variation in homeless people living in vehicles is related to the variation in homeless people living on the street, we see that the homeless population living in shelters is more isolated with none of the other categories' population measures serving as its important features. Furthermore, we discovered that each category of the homeless population indeed has significant features unique to itself. In particular, we found that unique to the homeless population living in vehicles, previous years' measures of its own category's populations are not significant to explaining the category itself in subsequent years.

The Earth Mover's distance showed a promising way to model the movements of homeless populations in Los Angeles. However, it was hard to verify the work showed actual migratory patterns of homeless populations as there are much more confounding factors involved in the decision to move. We also were unable to verify results from 2015 data because the interpolation of the data. While distance provided a reasonable metric for our cost function, it would be more useful to have a more sophisticated measure of time and associated cost with moving between census tracts. We found this information on time and money in a limited form through Google APIs but the large nature of our dataset made utilizing the APIs computationally infeasible.

Future work can be done in constructing likelihood models to represent the change in a census tract's population year to year as the difference between entry and exit Poisson processes. In doing this, the change in population would follow a Skellam distribution for which we could find entry and exit rate parameters that maximize the probability of all observed changes among the census tract populations.

Overall, much of analysis done provides intuitive results that aligns with common convention, though much work has to be done to provide clear conclusions to the results. Therefore, the issue of homelessness is still an issue worth looking at.

6 Acknowledgments

We'd like to thank a few individuals of whom which this project would not have been possible without. Firstly, we want to thank our adviser and mentor, Professor Michael Lindstrom, for all of his advice and help on this project. We'd also like to thank Professor Marcus Roper for facilitating the research course, and Professor Akram Almohalwas for being a Statistics Department sponsor for the work. Lastly, we'd like to thank our colleague Derek Yen for volunteering his time and leadership to benefit our research progress.

Appendices

Appendix A

Topic	1	2	3
BusStopDensity	0.239979186	0.047530001	0.049462321
GenPop2015Density	0.243524771	0.041706225	0.053510371
CoffeeDensity	0.134906125	0.138477373	0.071615764
x2014ZRI	0.028071895	0.262458822	0.050936451
x2015ZRI	0.043121903	0.265122447	0.026846013
x2014ZHVI	0.065000554	0.265760461	0
x2015ZHVI	0.071451958	0.263844377	0
RestaurantDensity	0.194748731	0.12249017	0.009535296
CrimeDensity	0.251380334	0.044595511	0.040267009
AffordableHousingDensity	0.166731829	0.08615567	0.096736748
TotalHousingUnits2014	0.001966843	0.116365558	0.31640733
TotalVacantUnits2014	0.071864834	0.093320227	0.243499925
UnemploymentRate2014	0.223025292	0.03376262	0.087724459
BelowPovertyRate2014	0.271727507	0	0.084304922
MedRentAsPercentOfGrossIncome2014	0.198619095	0.007168816	0.1609571
TotalPopulation2014	0.035637563	0.016318232	0.403830379
MedHouseholdIncome2014	0	0.24417883	0.143845255
MedRent2014	0	0.228135585	0.153607467
MedValue2014	0.039423579	0.264320204	0.030613658
MedMonthlyHousingCosts2014	0	0.222576083	0.177939362
TotalHousingUnits2015	0.002876729	0.114736213	0.317649489
TotalVacantUnits2015	0.074962972	0.108190705	0.217812037
UnemploymentRate2015	0.225464098	0.030252213	0.090554482
BelowPovertyRate2015	0.273170236	0	0.081968027
MedRentAsPercentOfGrossIncome2015	0.202584134	0.011804429	0.148012044
TotalPopulation2015	0.035156552	0.015157623	0.406728725
MedHouseholdIncome2015	0	0.244509815	0.145529549
MedRent2015	0.033797096	0.186962178	0.146692141
MedValue2015	0.080160033	0.208558393	0.051486016
MedMonthlyHousingCosts2015	0	0.226315607	0.168892675
ZRI1614	0.191481898	0.164249377	0
ZRI1715	0.231006152	0.013028832	0.107173957
ZRVI1614	0.220222528	0.079638973	0.036990551
ZRVI1715	0.220374908	0.019545058	0.117251482
AffordableHousingUnit1614	0.151436484	0.077291616	0.134211786
TotVacantUnit1614	0.131109481	0.165664797	0.03165198
UnemploymentRate1614	0.129323978	0.110665041	0.116592884
BelowPovertyRate1614	0.121498938	0.129030341	0.096606685
MedRentAsPercentGrossIncome1614	0.13134241	0.129056623	0.08157547
TotPopulation1614	0.122781083	0.107000827	0.133268171
MedHouseholdIncome1614	0.137768794	0.114561351	0.099991226
MedRent1614	0.115793789	0.172526639	0.047656669
MedValue1614	0.144080602	0.128988635	0.068780838
MedMonthlyHousingCost1614	0.166996415	0.137052132	0.025778368

Table 1: Topic Matrix for k = 3

Appendix A

Topic	1	2	3	4
BusStopDensity	0.223890493	0	0.126410039	0.048531971
GenPop2015Density	0.222974315	0	0.139617284	0.064519661
CoffeeDensity	0.108419144	0.074261605	0.176382215	0.068032037
x2014ZRI	0	0.177185679	0.260862983	0.040828876
x2015ZRI	0	0.150888385	0.287590617	0.027552827
x2014ZHVI	0	0.119363035	0.316420412	0.007258205
x2015ZHVI	0.003639595	0.110851528	0.321258	0.00366334
RestaurantDensity	0.150992093	0.004731172	0.218887625	0.024105529
CrimeDensity	0.22927511	0	0.143781246	0.050721734
AffordableHousingDensity	0.170148552	0.101500659	0.083768763	0.061313472
TotalHousingUnits2014	0	0	0.159266341	0.395869272
TotalVacantUnits2014	0.052509749	0	0.151942917	0.305733755
UnemploymentRate2014	0.239519336	0.078239803	0.036529229	0.038539699
BelowPovertyRate2014	0.280382267	0	0.04953616	0.065405422
MedRentAsPercentOfGrossIncome2014	0.24563891	0.138022139	0	0.072715433
TotalPopulation2014	0.067728175	0.024983027	0	0.43796743
MedHouseholdIncome2014	0	0.311962873	0.119830981	0.07562984
MedRent2014	0	0.30006814	0.108603834	0.084009977
MedValue2014	0	0.1852487	0.258627796	0.016142386
MedMonthlyHousingCosts2014	0	0.322162173	0.080774547	0.098812254
TotalHousingUnits2015	0	0	0.158100829	0.397523433
TotalVacantUnits2015	0.049703714	0	0.170706655	0.278082089
UnemploymentRate2015	0.245646307	0.085439021	0.026399728	0.036282779
BelowPovertyRate2015	0.284598571	0	0.04152005	0.056820454
MedRentAsPercentOfGrossIncome2015	0.251597038	0.155580154	0	0.049740868
TotalPopulation2015	0.068450434	0.02689893	0	0.440242776
MedHouseholdIncome2015	0	0.309320841	0.122078025	0.079116692
MedRent2015	0.034772974	0.221478567	0.114709907	0.0996405
MedValue2015	0.045390581	0.141053282	0.216553417	0.040337222
MedMonthlyHousingCosts2015	0	0.319330438	0.089637616	0.090951611
ZRI1614	0.129864217	0	0.280629398	0
ZRI1715	0.253468373	0.05930433	0.014877543	0.066596731
ZRVI1614	0.200773369	0.011904694	0.147043023	0.037178698
ZRVI1715	0.238676202	0.050358734	0.028585072	0.085990259
AffordableHousingUnit1614	0.163968813	0.115880384	0.05409023	0.096568333
TotVacantUnit1614	0.118145212	0.172526769	0.142514339	0
UnemploymentRate1614	0.139342388	0.155217478	0.071069812	0.067124384
BelowPovertyRate1614	0.135072214	0.202165477	0.063353459	0.027190058
MedRentAsPercentGrossIncome1614	0.13519253	0.168358344	0.088500131	0.029862748
TotPopulation1614	0.134391031	0.15231813	0.065215983	0.08787436
MedHouseholdIncome1614	0.134704392	0.116847441	0.108735146	0.068363586
MedRent1614	0.109548689	0.201402317	0.129147633	0
MedValue1614	0.145355155	0.154898287	0.103582624	0.015711833
MedMonthlyHousingCost1614	0.144345015	0.096251024	0.165349873	0.00193439

Table 2: Topic Matrix for k = 4

Appendix A

Topic	1	2	3	4	5
BusStopDensity	0.246169759	0	0.093436149	0.087306391	0.034038759
GenPop2015Density	0.248586077	0	0.106199862	0.113065329	0.008203662
CoffeeDensity	0.119348996	0.077672819	0.162724749	0.08040842	0.031336994
x2014ZRI	0	0.181809754	0.269864111	0.021989568	0.032170981
x2015ZRI	0	0.156198099	0.297105763	0.016976058	0.012495288
x2014ZHVI	0.006959114	0.127760045	0.323960826	0.006875372	0
x2015ZHVI	0.011815194	0.118913712	0.328756391	0.005923754	0
RestaurantDensity	0.176123824	0.014226116	0.196604541	0.0603839	0
CrimeDensity	0.246479843	0	0.11488084	0.102283302	0.029684012
AffordableHousingDensity	0.158925621	0.081204973	0.069144572	0.069970967	0.13685413
TotalHousingUnits2014	0	0.027511591	0.164053998	0.388665302	0
TotalVacantUnits2014	0.0620755	0.010354803	0.142771583	0.314344531	0
UnemploymentRate2014	0.222651323	0.036100691	0.021203125	0.06480332	0.171625535
BelowPovertyRate2014	0.256238283	0	0.032717685	0.116723251	0.167146218
MedRentAsPercentOfGrossIncome2014	0.175246246	0.045014735	0	0.08739479	0.333517962
TotalPopulation2014	0.062909249	0.046489148	0	0.424137541	0.052085489
MedHouseholdIncome2014	0	0.329823097	0.111771662	0.012226506	0.095368899
MedRent2014	0	0.30324508	0.106478221	0.027361579	0.126484459
MedValue2014	0	0.185109092	0.268038313	0	0.053742211
MedMonthlyHousingCosts2014	0	0.322452174	0.079814854	0.034597878	0.149198207
TotalHousingUnits2015	0	0.025003485	0.163270088	0.390884846	0
TotalVacantUnits2015	0.057498114	0.013842741	0.164134783	0.28870453	0
UnemploymentRate2015	0.256453823	0.072933693	0	0.0547793	0.120955245
BelowPovertyRate2015	0.282786087	0	0.009192678	0.101975806	0.127737764
MedRentAsPercentOfGrossIncome2015	0.213358212	0.093934003	0	0.056135647	0.277821585
TotalPopulation2015	0.057736305	0.04350776	0	0.426319034	0.066801667
MedHouseholdIncome2015	0	0.307046763	0.127189433	0.019687455	0.137560661
MedRent2015	0.016215277	0.211639806	0.11584882	0.065996986	0.138536287
MedValue2015	0.056552287	0.14976037	0.209247385	0.034877814	0.024781837
MedMonthlyHousingCosts2015	0	0.304931534	0.097254377	0.030260567	0.180025472
ZRI1614	0.140779087	0	0.274329587	0.02401589	0
ZRI1715	0.249291949	0.033083485	0	0.087991492	0.153172697
ZRVI1614	0.203782739	0.001090654	0.128022264	0.06849606	0.072264351
ZRVI1715	0.236405252	0.032971372	0.001683857	0.106092428	0.133128215
AffordableHousingUnit1614	0.132496697	0.07540141	0.050495933	0.105149223	0.186546839
TotVacantUnit1614	0.106863282	0.147791562	0.137455819	0	0.132174456
UnemploymentRate1614	0.203060869	0.218639291	0.008166213	0.042817598	0
BelowPovertyRate1614	0.242422772	0.306314109	0	0	0
MedRentAsPercentGrossIncome1614	0.245447135	0.275639544	0	0	0
TotPopulation1614	0.078400217	0.098350262	0.0749828	0.083715994	0.244814396
MedHouseholdIncome1614	0	0	0.194392032	0.09586334	0.459913067
MedRent1614	0.034594577	0.117851594	0.158895828	0	0.288888301
MedValue1614	0.110284386	0.108184508	0.108268535	0.019884421	0.189432639
MedMonthlyHousingCost1614	0.05406622	0	0.204508837	0.029700649	0.286711324

Table 3: Topic Matrix for k = 5

Appendix B

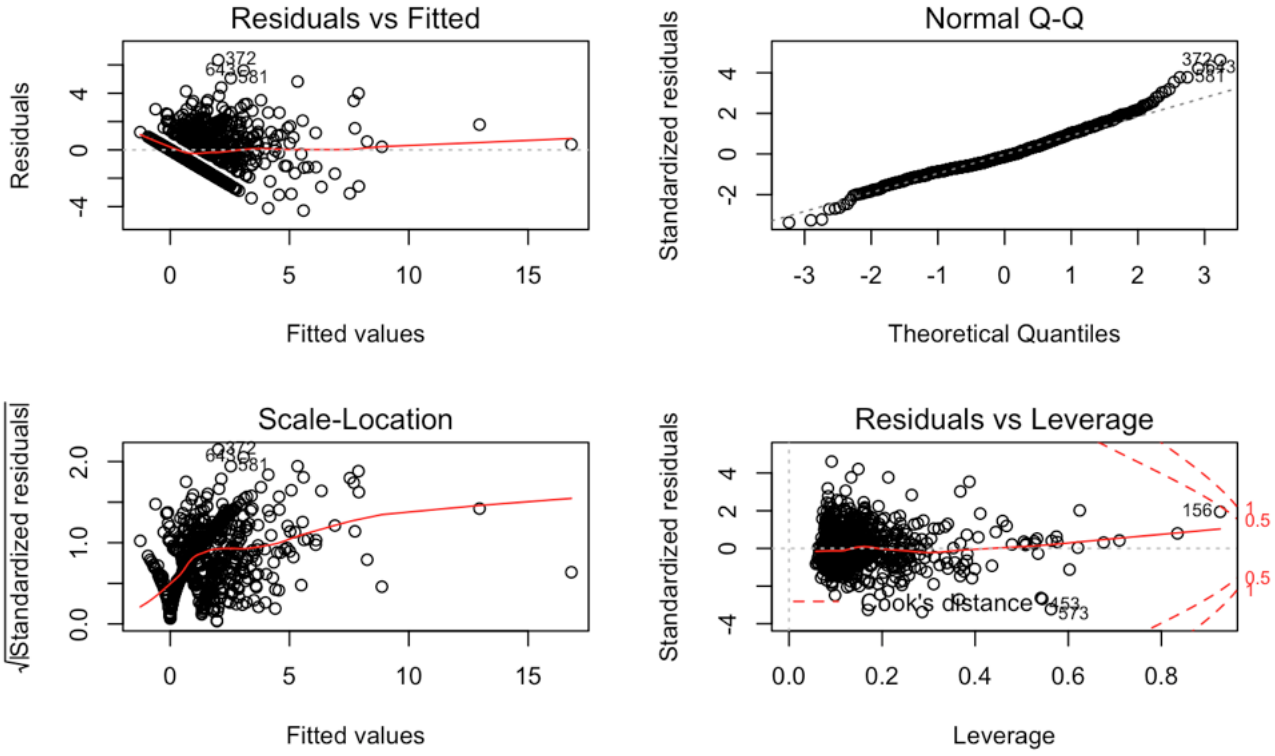


Figure 40: Diagnostic Residual Plots for Modeling the 2017 Street Homeless Population

Appendix B

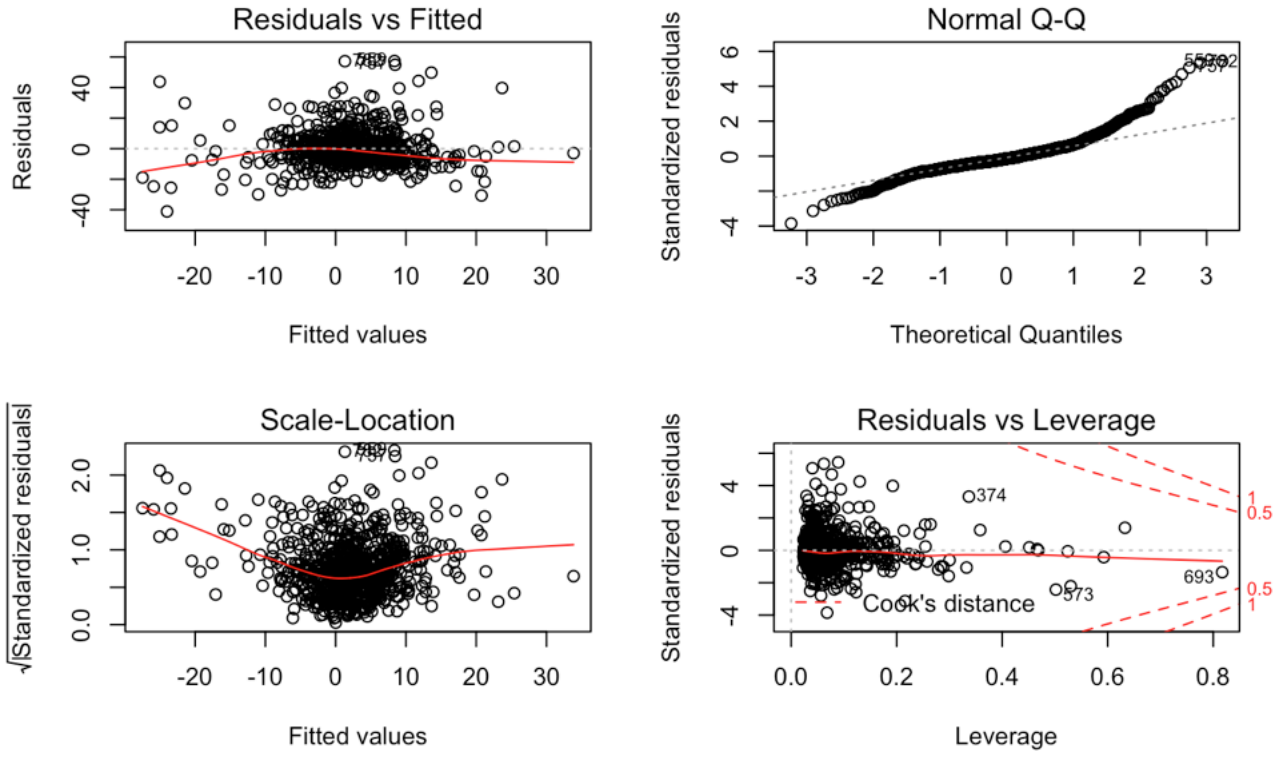


Figure 41: Diagnostic Residual Plots for Modeling the 2017 Vehicles Homeless Population

Appendix B

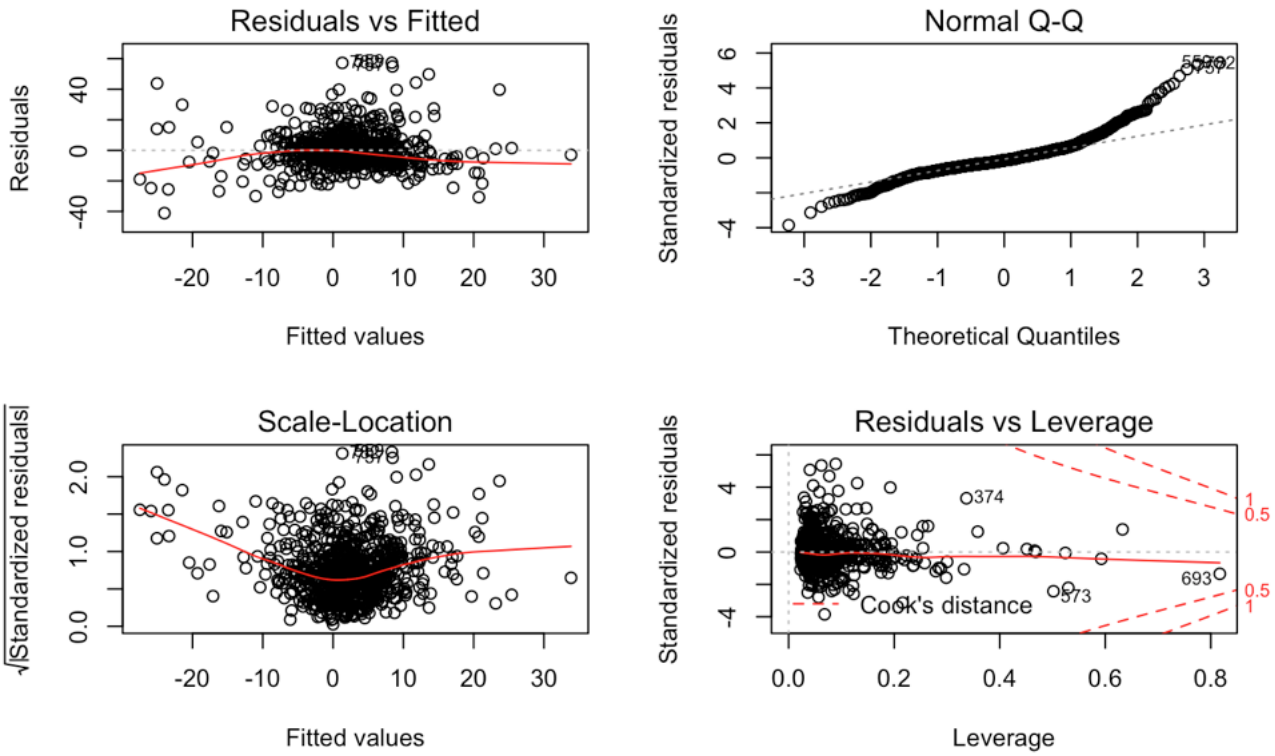


Figure 42: Diagnostic Residual Plots for Modeling the 2016-2017 Change in Vehicles Homeless Population

Appendix B

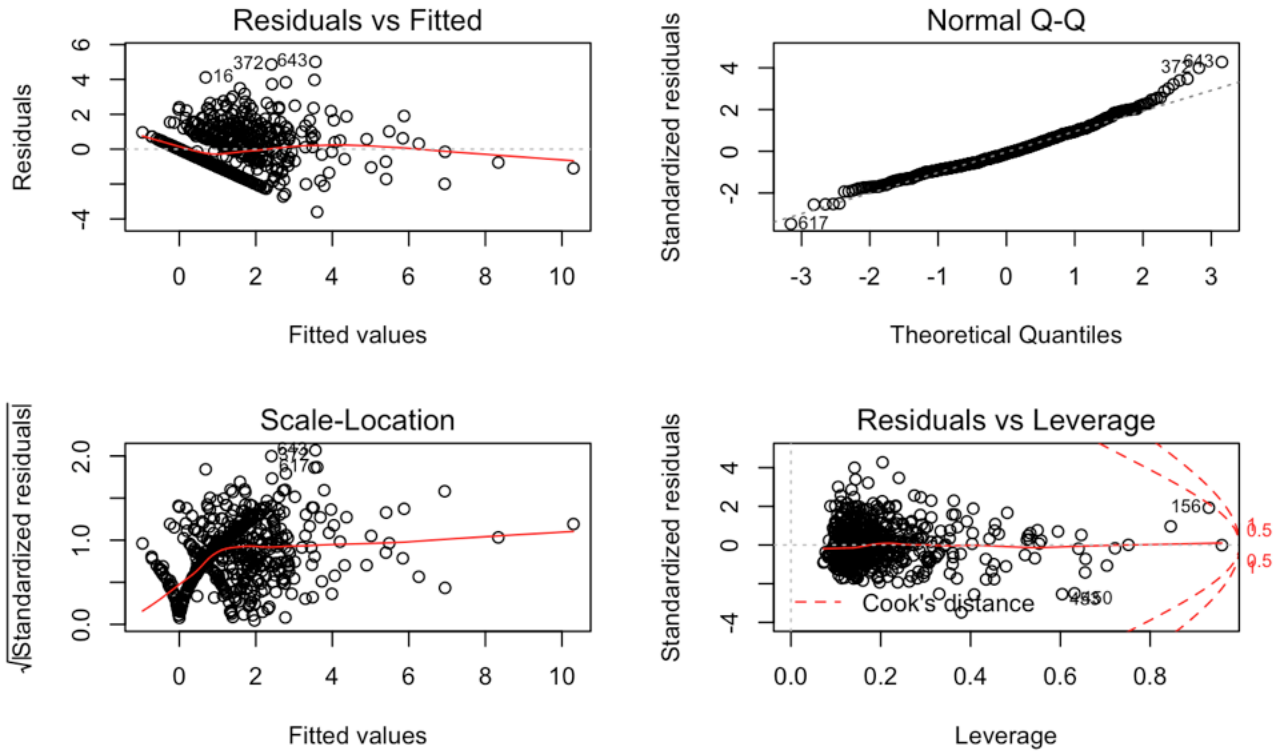


Figure 43: Diagnostic Residual Plots for Modeling the 2016-2017 Change in Street Homeless Population

```
Call:
lm(formula = (totSheltPeople)^1/2 ~ nearest_ralphs + nearest_traderjoes +
    nearest_library + X2017ZHVI + Citations + Coffee + Restaurants +
    Shelters + totCars + CrimeCount + MedHouseIncomes + HousingUnits +
    BusStops + X2015HousedPopulation + totHomeless2015 + totHomeless2016 +
    TractSqMi + totVanPeople + totCamperPeople + totEncampPeople +
    totTentPeople, na.action = na.omit)

Residuals:
     Min       1Q   Median       3Q      Max
-169.899   -3.080    0.737    3.466  168.103

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -4.010e+00  2.804e+00  -1.430  0.15302
nearest_ralphs        1.892e-06  4.002e-04   0.005  0.99623
nearest_traderjoes   -3.017e-04  2.745e-04  -1.099  0.27187
nearest_library       3.211e-05  4.984e-04   0.064  0.94865
X2017ZHVI            -2.563e-06  2.192e-06  -1.169  0.24257
Citations             1.817e-01  2.826e-01   0.643  0.52028
Coffee               -1.539e-01  2.925e-01  -0.526  0.59885
Restaurants          -2.427e-01  9.282e-02  -2.615  0.00908 **
Shelters             -2.368e+00  1.328e+00  -1.783  0.07490 .
totCars               1.756e-01  2.307e-01   0.761  0.44659
CrimeCount            1.094e-03  6.283e-03   0.174  0.86180
MedHouseIncomes       4.593e-05  1.778e-05   2.584  0.00993 **
HousingUnits          1.172e-03  1.067e-02   0.110  0.91254
BusStops              2.352e-01  1.023e-01   2.299  0.02172 *
X2015HousedPopulation 1.069e-04  4.229e-04   0.253  0.80053
totHomeless2015       1.115e-01  1.459e-02   7.640 5.50e-14 ***
totHomeless2016       3.022e-01  1.312e-02  23.033  < 2e-16 ***
TractSqMi            -5.066e-01  7.075e-01  -0.716  0.47421
totVanPeople         -2.740e-01  1.482e-01  -1.849  0.06481 .
totCamperPeople      -1.273e-01  7.571e-02  -1.681  0.09314 .
totEncampPeople      -4.243e-01  6.976e-02  -6.083 1.73e-09 ***
totTentPeople        -7.443e-02  6.747e-02  -1.103  0.27022
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.26 on 912 degrees of freedom
  (1226 observations deleted due to missingness)
Multiple R-squared:  0.8358,    Adjusted R-squared:  0.8321
F-statistic: 221.1 on 21 and 912 DF,  p-value: < 2.2e-16
```

Figure 44: Modeling Homeless Population Living in Shelters

```
Call:
lm(formula = (totTentPeople + totEncampPeople)^1/2 ~ nearest_ralphs +
    nearest_traderjoes + nearest_library + X2017ZRI + Citations +
    Coffee + Restaurants + Shelters + CrimeCount + MedHouseIncomes +
    HousingUnits + BusStops + X2015HousedPopulation + totHomeless2015 +
    totHomeless2016 + TractSqMi + totCarPeople + totVanPeople +
    totCamperPeople + totSheltPeople, na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-31.074  -2.330  -0.113   1.865  84.575

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             7.276e-01  1.560e+00   0.467  0.64096
nearest_ralphs          1.446e-04  2.032e-04   0.712  0.47675
nearest_traderjoes     -2.816e-04  1.376e-04  -2.047  0.04091 *
nearest_library         4.970e-04  2.522e-04   1.971  0.04907 *
X2017ZRI               -3.868e-04  3.900e-04  -0.992  0.32153
Citations               1.514e-01  1.435e-01   1.055  0.29167
Coffee                 -3.410e-01  1.483e-01  -2.300  0.02170 *
Restaurants             6.324e-02  4.730e-02   1.337  0.18152
Shelters                6.689e-01  6.733e-01   0.993  0.32076
CrimeCount              6.700e-03  3.185e-03   2.104  0.03567 *
MedHouseIncomes         1.533e-05  9.376e-06   1.635  0.10229
HousingUnits            1.304e-02  5.412e-03   2.410  0.01616 *
BusStops                1.519e-01  5.153e-02   2.948  0.00328 **
X2015HousedPopulation -9.630e-04  2.127e-04  -4.528 6.76e-06 ***
totHomeless2015         6.377e-02  7.281e-03   8.759  < 2e-16 ***
totHomeless2016         1.043e-01  7.650e-03  13.630  < 2e-16 ***
TractSqMi              -1.601e-02  3.616e-01  -0.044  0.96470
totCarPeople            1.255e-01  7.382e-02   1.701  0.08934 .
totVanPeople            2.643e-01  7.491e-02   3.528  0.00044 ***
totCamperPeople        -3.733e-02  3.824e-02  -0.976  0.32929
totSheltPeople         -6.311e-02  8.127e-03  -7.765 2.18e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.768 on 913 degrees of freedom
  (1226 observations deleted due to missingness)
Multiple R-squared:  0.7589,    Adjusted R-squared:  0.7536
F-statistic: 143.7 on 20 and 913 DF,  p-value: < 2.2e-16
```

Figure 45: Modeling the Homeless Population Living on the Street

```
Call:
lm(formula = (totCarPeople + totVanPeople + totCamperPeople)^0.5 ~
    nearest_ralphs + nearest_traderjoes + nearest_library + X2017ZHVI +
        Citations + Coffee + Restaurants + Shelters + totCars +
        totSheltPeople + CrimeCount + MedHouseIncomes + HousingUnits +
        BusStops + X2015HousedPopulation + totHomeless2015 +
        totHomeless2016 + TractSqMi + totEncampPeople + totSheltPeople,
    data = dat, na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3709 -1.1730 -0.0287  0.9194  5.6494

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.316e+00  2.634e-01   4.996 7.01e-07 ***
nearest_ralphs        2.839e-05  3.741e-05   0.759 0.448086
nearest_traderjoes    2.930e-05  2.562e-05   1.143 0.253159
nearest_library      -8.543e-05  4.667e-05  -1.831 0.067495 .
X2017ZHVI             9.212e-09  2.050e-07   0.045 0.964165
Citations             1.889e-01  5.914e-02   3.194 0.001449 **
Coffee               -5.744e-02  2.749e-02  -2.089 0.036944 *
Restaurants          -3.928e-03  8.736e-03  -0.450 0.653116
Shelters              5.600e-02  1.240e-01   0.452 0.651736
totCars               4.101e-01  1.900e-02  21.578  < 2e-16 ***
totSheltPeople       -2.619e-03  1.546e-03  -1.694 0.090582 .
CrimeCount            7.515e-04  5.896e-04   1.275 0.202759
MedHouseIncomes      -4.706e-06  1.669e-06  -2.820 0.004908 **
HousingUnits         -1.641e-03  9.989e-04  -1.643 0.100708
BusStops              3.923e-02  9.559e-03   4.104 4.42e-05 ***
X2015HousedPopulation 7.499e-05  3.966e-05   1.891 0.058979 .
totHomeless2015      -1.909e-03  1.363e-03  -1.400 0.161714
totHomeless2016       1.174e-03  1.498e-03   0.784 0.433163
TractSqMi             3.691e-02  6.635e-02   0.556 0.578199
totEncampPeople       1.961e-02  5.486e-03   3.576 0.000368 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.432 on 912 degrees of freedom
Multiple R-squared:  0.4494,    Adjusted R-squared:  0.4379
F-statistic: 39.18 on 19 and 912 DF,  p-value: < 2.2e-16
```

Figure 46: Modeling the Homeless Population Living in Vehicles

References

[1] John Wilkens. *Busing the homeless out of town: 'Greyhound therapy' or a humane approach?* Mar. 2018. URL: http://www.latimes.com/local/lanow/la-me-homeless-bus-20180312-story.html.

[2] Rew Lynn et al. "Correlates of Resilience in Homeless Adolescents". In: *Journal of Nursing Scholarship* 33.1 (), pp. 33–40. DOI: 10.1111/j.1547-5069.2001.00033.x. eprint: https://sigmapubs.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1547-5069.2001.00033.x. URL: https://sigmapubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1547-5069.2001.00033.x.

[3] Edward Helderop et al. "Predicting Movement of Homeless Young Adults: Artificial Neural Networks and Generalized Linear Models". In: *Journal of the Society for Social Work and Research* 9.1 (2018), pp. 89–106. DOI: 10.1086/696129. eprint: https://doi.org/10.1086/696129. URL: https://doi.org/10.1086/696129.

[4] Stephen Metraux, Dan Treglia, and Thomas P. O'Toole. "Migration by Veterans Who Received Homeless Services From the Department of Veterans Affairs". In: *Military Medicine* 181.10 (2016), pp. 1212–1217. DOI: 10.7205/MILMED-D-15-00504. eprint: /oup/backfile/content_public/journal/milmed/181/10/10.7205_milmed-d-15-00504/4/milmed-d-15-00504.pdf. URL: http://dx.doi.org/10.7205/MILMED-D-15-00504.

[5] Derek Yen. "Predicting Homeless Population Dynamics in Los Angeles". In: (2018).

[6] Andrew Stevens. *SP_PROJ - File Exchange - MATLAB Central*. Jan. 2010. URL: http://www.mathworks.com/matlabcentral/fileexchange/26413-sp-proj.

[7] *Statement On The 2017 Youth Count Adjustment*. URL: https://www.lahsa.org/news?article=379-statement-on-the-2017-youth-count-adjustment.

[8] MATLAB. *version 9.2.0.538062 (R2017a)*. Natick, Massachusetts: The MathWorks Inc., 2017.

[9] N. Gillis. "The Why and How of Nonnegative Matrix Factorization". In: *ArXiv e-prints* (Jan. 2014). arXiv: 1401.5226 [stat.ML].

[10] Daniel D. Lee and H. Sebastian Seung. "Algorithms for Non-negative Matrix Factorization". In: *Proceedings of the 13th International Conference on Neural Information Processing Systems*. NIPS'00. Denver, CO: MIT Press, 2000, pp. 535–541. URL: http://dl.acm.org/citation.cfm?id=3008751.3008829.

[11] I.T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag New York, 2002. ISBN: 9780387224404.

[12] Jonathon Shlens. "A Tutorial on Principal Component Analysis". In: *CoRR* abs/1404.1100 (2014). arXiv: 1404.1100. URL: http://arxiv.org/abs/1404.1100.

[13] Jeff A Bilmes et al. "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models". In: *International Computer Science Institute* 4.510 (1998), p. 126.

[14] Michael Steinbach, Levent Ertöz, and Vipin Kumar. "The challenges of clustering high dimensional data". In: *New directions in statistical physics*. Springer, 2004, pp. 273–309.

[15] Shailendra Kathait. *Variable Reduction: An Art As Well As Science*. Feb. 2015. URL: http://www.linkedin.com/pulse/variable-reduction-art-well-science-shailendra-s.

[16] Y.-C. Chen. "A Tutorial on Kernel Density Estimation and Recent Advances". In: *ArXiv e-prints* (Apr. 2017). arXiv: 1704.03924 [stat.ME].

[17] Joseph F. Hair. *Multivariate data analysis: with readings (by) J.F. Hair (... and others)*. Macmillan, 1995.

[18] Gilles Louppe et al. "Understanding variable importances in forests of randomized trees". In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges et al. Curran Associates, Inc., 2013, pp. 431–439. URL: http://papers.nips.cc/paper/4928-understanding-variable-importances-in-forests-of-randomized-trees.pdf.

[19] André Altmann et al. "Permutation importance: a corrected feature importance measure". In: *Bioinformatics* 26.10 (2010), pp. 1340–1347. DOI: 10.1093/bioinformatics/btq134. eprint: /oup/backfile/contentpublic/journal/bioinformatics/26/10/10.1093_bioinformatics_btq134/1/btq134.pdf. URL: http://dx.doi.org/10.1093/bioinformatics/btq134.

[20] Jacob J de Vries, Peter Nijkamp, and Piet Rietveld. "Exponential or Power Distance-Decay for Commuting? An Alternative Specification". In: *Environment and Planning A: Economy and Space* 41.2 (2009), pp. 461–480. DOI: 10.1068/a39369. eprint: https://doi.org/10.1068/a39369. URL: https://doi.org/10.1068/a39369.

[21] Christopher M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, 2013.

[22] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. "Deep Sparse Rectifier Neural Networks". In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Geoffrey Gordon, David Dunson, and Miroslav Dudík. Vol. 15. Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA: PMLR, Apr. 2011, pp. 315–323. URL: http://proceedings.mlr.press/v15/glorot11a.html.

[23] R A. Dunne and Nicole Campbell. "On The Pairing Of The Softmax Activation And Cross–Entropy Penalty Functions And The Derivation Of The Softmax Activation Function". In: (Feb. 1970).

[24] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *CoRR* abs/1412.6980 (2014). arXiv: 1412.6980. URL: http://arxiv.org/abs/1412.6980.

[25] Michael A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015.

[26] US Census Bureau. *2017 Data Release Schedule*. URL: https://www.census.gov/programs-surveys/acs/news/data-releases/2017/release-schedule.html.

[27] *2018 Greater Los Angeles Homeless Count Presentation*. URL: https://www.lahsa.org/documents?id=2059-2018-greater-los-angeles-homeless-count-presentation.pdf.

[28] Martín Abadi et al. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems". In: *CoRR* abs/1603.04467 (2016).

[29] Y. Rubner, C. Tomasi, and L. J. Guibas. "A metric for distributions with applications to image databases". In: *IEEE International Conference on Computer Vision* (1998), pp. 59–66.

[30] H. B. Goodwin. *Naval Institute Proceedings*. Vol. 36. 3. 1910. Chap. The haversine in nautical astronomy, pp. 735–746.

[31] Ofir Pele and Michael Werman. "A linear time histogram metric for improved sift matching". In: *Computer Vision–ECCV 2008*. Springer, Oct. 2008, pp. 495–508.

[32] Marina Sokolova and Guy Lapalme. "A systematic analysis of performance measures for classification tasks". In: *Information Processing & Management* 45.4 (2009), pp. 427–437. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2009.03.002. URL: http://www.sciencedirect.com/science/article/pii/S0306457309000259.