Yawn... such a creative title :)

Homeless Population

usually alphabetical by last name for these occasions

Ryan Du, Sanjay Shukla, Hannah Ross, Xiang Yang Ng
Math 199, Spring 2018
Advisor: Michael Lindstrom

Draft Known Issue:

- Ryan (Neural Network)
  - The text in typewriter font are comments
  - Method section, some of the math are not perfectly concrete
  - Citation is not complete and accurate.

Scanned with CamScanner

Contents

1 Introduction **2**

2 Data Management **2**

3 Techniques and Models **4**
- 3.1 Normalizing Data into Percentiles 4
- 3.2 Topic Modeling 4
- 3.3 Linear Models 5
- 3.4 Neural Networks 6

4 Results **11**
- 4.1 Topic Modeling 11
- 4.2 Linear Models 19
- 4.3 Neural Networks 24

5 Summary and Future Work **26**
- 5.1 Earth Mover's Distance 26
- 5.2 Neural Network Exploration 26
- 5.3 Convolutional Neural Network 26

6 Acknowledgments **26**

Need an Abstract

Scanned with CamScanner

Interesting overall but how it relates to our work is unclear: we don't directly study veterans, sexual abuse, etc.

1 Introduction

Homelessness is a prevalent issue in the Los Angeles area, and it is an issue that touches many different people in society. Previous models of homeless migration have acknowledged how homeless people are portrayed as highly mobile, and yet the migration behavior has not been systematically analyzed. \cite[studies]

There are a variety of forces that move homeless individuals including planned physical and social change programs such as urban renewal, slum clearance, highway construction, 'Greyhound therapy', and police sweeps of encampments. In a study predicting the movement of homeless young adults, HYA's were identified as a subset of the population that were most receptive to care provided by agencies. This motivates our systematic study of homelessness which can serve to better inform agency initiatives that serve the homeless young adult population [1].

need citations

(HYA)

The homeless migrant population that we intend to analyze includes a variety of sub-populations, each with their own unique causes for homelessness and motivations for moving. Among these sub-populations are homeless young adults, of which 60 percent self-identify as travelers. 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.

Another significant sub-population of the greater Los Angeles homeless population are the homeless veterans of Los Angeles. In a study conducted by the American Journal of Public Health, it was discovered that homeless veterans and homeless non-veterans have very contrasting profiles. Veterans are more likely to be white, educated, to have a history of marriage, and were more likely to have alcohol issues rather than drug issues. In interviews conducted 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 need to be treated and addressed in order to treat the chronic issue of homelessness.

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 [3].Interestingly, in-migration tended to roughly balance out-migration in a region.

In understanding the fluctuations and patterns of homelessness, we will be in a better position to inform more impactful and efficient uses of resources that are being directed towards the homeless population.

usually notes end with: "This paper is organized as follows: Section 2 discusses blah; section 3 does blang; ...; finally work is concluded in section ??"

2 Data Management

We used data from several sources: the Los Angeles Homeless Services Authority (LAHSA), the housing and rental website Zillow, the public data platform DataLA, the Los Angeles County

[Plagiarism] This is SERIOUS

Scanned with CamScanner

I'm not even going to read this. Redo it. Yourself

Metropolitan Transportation Authority, and the American Community Surveys.

We used LAHSA to obtain data about estimates of homeless populations by census track in the years 2017. Estimates describe counts of both unsheltered and sheltered homeless people, with additional data describing features of the homeless population itself, including the number of people in tents, cars, vans, and makeshift shelters.

All data on census tract boundaries came from shapefiles on Los Angeles Geohub, a public platform for geolocation data from Los Angeles County. The two main housing indices that we used were Zillow Rent Index (ZRI) and Zillow Home Value Index (ZHVI). The former tracks the median rent while the latter tracks the median home value within a certain area. We obtained both for all zip codes and then converted them to census tracts by locally averaging over zip codes.

We found information on features, i.e restaurants, overnight parking citations, crimes, total population for all census tracts through DataLA, an open data platform maintained by the city of Los Angeles. After obtaining street addresses for each of these features, we converted them into latitude/longitude points using Google Maps Geocoding API through Python scripting. Locations were converted from latitude/longitude points to California State Planar feet using MATLAB code "SP Proj" [11] that we obtained from Mathworks. Each state plane (x,y) coordinate, corresponding to a feature, was assigned to a census tract in MATLAB using Los Angeles shapefiles.

Locations of coffee shops were obtained by iteratively searching for coffee shops within a small radius using the Google Places API through Python scripting, which produced street addresses and latitude/longitude points, while latitude/longitude of homeless shelters and services were obtained from Los Angeles Geohub, whose data is maintained by the County of Los Angeles Location Management System. We then assigned these latitude/longitude points to census tracts using a similar process as above.

Finally, we obtained estimates of the median household incomes per census tract as of 2014, 2015, and 2016 from the American Community Survey, a five-year estimate conducted by the United States Census Bureau. Geographic data was downloaded in geodatabase format and converted to a spreadsheet using the open source geographic information systems software QGIS.

For information on grocery store locations, we 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 least you can do is cite Derek's report and summarize in YOUR OWN WORDS!

Scanned with CamScanner

Rmk: I don't think you need to bold face so much.

3 Techniques and Models

3.1 Normalizing Data into Percentiles

Los Angeles had certain features that exhibited huge disparities among census tracts such as Median Household Incomes and Zillow Home and Rent Index. In order to account for these large disparities, we normalized the data according to percentile ranks. First, suppose we have a data matrix, $\mathbf{D} \in \mathbb{R}^{nxm}$ 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 the column from 1 to the number of observations in the column. If some elements have the same value, then they have the average rank. Denote the census track i for feature j converted to its percentile rank be entries, $\mathbf{R}^{ij}$, 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 $\mathbf{A}$, where for all i,j,

each

at

needs to be clearer

should $= n$...

$$\mathbf{A}_{ij} = \frac{\mathbf{R}^{ij}}{n} \tag{1}$$

← so no 0th percentile?

3.2 Topic Modeling

use 1 times for x

awkward

Using the data normalized into percentiles we are left with a $\mathbb{R}^{nxm}$ data matrix denoted $\mathbf{A}$. Let $\mathbf{A}_{i.}$ corresponds to a row of a census tract and with the features within that census tract, $\mathbf{A}_{.j}$ correspond to a column of only census tracts or only an individual feature, and $\mathbf{A}_{ij}$ correspond to element (i, j).

$H \in \mathbb{R}^{n \times k}$ with $H_{ij} \geq 0$...

We used nonnegative matrix factorization (NMF) in order to reduce the dimensions of our data to a topic matrix, $\mathbf{H}^{nxk}$ and a weight weight matrix, $\mathbf{W}^{kxm}$ which would approximate to our data matrix such that k is the number of topics, $\mathbf{W}, \mathbf{H} > 0$ and $\mathbf{A} \approx \mathbf{WH}$.

To obtain the matrices for $\mathbf{H}$ and $\mathbf{W}$, we minimize the Frobenius Norm by minimizing the least-squares function [4]:

$$\|\mathbf{A}_{.j} - \mathbf{WH}_{.j}\|^2 \tag{2}$$

$\|A - WH\|_F^2$

unclear awkward

By compressing our data into smaller matrices, $\mathbf{H}$ and $\mathbf{W}$, we reduce the dimensions of the data and will help avoid over-fitting when making predictive inferences on the relations between topics and their associated weights

why is k bolded? usually only vectors are bold faced, if at all

We varied the rank $\mathbf{k}$ from $k = 2$ to $k = 8$ 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 $\mathbf{k}$, we selected the top 10 census tracts with the highest weights for each topic and the homeless population density and plotted the centroids over a map of Los Angeles in order to visualize the clusters of data.

Moreover, for all $\mathbf{k}$, we also find the correlation between the weights of each topic and the homeless population density. This is because we want to analyse the relationship between the topic matrix and the homeless population density to see if we could identify dynamics in the homeless population that could be explained by the topics. Denote $\mathbf{W}_{.i}$, where i is the ith column of $\mathbf{A}$ and $\mathbf{HPD}$ be homeless population density vector for all of the census tracks. The correlation between the weights of each topic and the homeless population density, $\rho(\mathbf{HPD}, \mathbf{W}_{.i})$:

$N = n$...

$$\rho(\mathbf{HPD}, \mathbf{W}_{.i}) = \frac{1}{N-1} \sum_{k=1}^{n} \left(\frac{\mathbf{HPD}_k - \mu_{\mathbf{HPD}}}{\sigma_{\mathbf{HPD}}} \right) \left(\frac{\mathbf{W}_{ki} - \mu_{\mathbf{W}_{.i}}}{\sigma_{\mathbf{W}_{.i}}} \right) \tag{3}$$

But what IS NMF?

You explained the algorithm but not what it means intuitively.

Scanned with CamScanner

where $\mu_{\mathbf{HPD}}$ is the mean of **HPD**, $\sigma_{\mathbf{HPD}}$ is the standard deviation of **HPD**, $\mu_{\mathbf{W}_{.i}}$ is the mean of $\mathbf{W}_{.i}$, $\sigma_{\mathbf{W}_{.i}}$ is the standard deviation of $\mathbf{W}_{.i}$

After finding the correlation, we take the 2 topics that have the highest positive and highest negative correlation respectively with the homeless population density. That's because the higher the absolute value of the correlation, the better the topics are at explaining the homeless population. First, convert the census tracks to its corresponding longitudes and latitudes. We then plot a heat map graph of weights for each topic and homeless population density on the longitude-latitude plane.

We also find the local "variance" of the census track, by first normalizing the columns as z-scores. Then for each census tract with their corresponding features, $\mathbf{A}_{i.}$, obtain the 5 "nearest" census tracks based on their "distance", i.e Frobenius norm of the pair of longitude and latitude of the census track and that of all the other census tracks. Denote $\mathbf{A}_{k_j.}$ 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 within of the census tracks and its neighbours, i.e form a matrix $B^i \in \mathbb{R}^{jxm}$, such that

why the "quotes"?

ℓ^2 norm or "Euclidean distance"

use ℓ_1 for ℓ

$$B^i = \begin{pmatrix} -\mathbf{A}_{i.}- \\ -\mathbf{A}_{k_1.}- \\ -\mathbf{A}_{k_2.}- \\ -\mathbf{A}_{k_3.}- \\ -\mathbf{A}_{k_4.}- \\ -\mathbf{A}_{k_5.}- \end{pmatrix} \tag{4}$$

At last, from (3), we take the sum of squares of all the elements of B_i, which we denote SSB^i:

$$SSB^i = \sum_{j=1}^{5} \sum_{l=1}^{m} \|B^i_{jm}\|^2 \tag{5}$$

6 facts after all

$\rightarrow \sum_{j=1}^{6} \sum_{l=1}^{m} |B^i_{jl} - \langle B^i_{\bullet l} \rangle|^2$ I think it makes more sense to compute i.e. sum of squared deviations from avg norm. feature value

3.3 Linear Models

unnecessary detail

We created three linear models to predict the variation in three different categories of the homeless population. We divided the homeless population into three distinct categories: the shelter homeless population (totShelterPeople), the in-vehicle homeless population (totCarPeople, totVanPeople, and totCamperPeople), and the street homeless population (totTentPeople, totEncampPeople). We tried to explain variation in each category of the homeless population using the other homeless categories and the collected features of each census tract including distance to the nearest Ralphs, distance to the nearest Trader Joes, 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.

In the modeling process for each of these three homeless population categories, we controlled for issues of high multicolinearity by eliminating predictors that had variation inflation factors greater than 10, corresponding to redundant predictors whose variations are more than 90% explained by the variation of the other predictors (5). This is in line with Hair et. al. (1995), which maintains that 10 is the maximum level of acceptable Variance Inflation Factor scores. In addition, we imposed

Scanned with CamScanner

needs more explaining.

transformations on the response variables in order to avoid violations of our models assumptions regarding constant variance and normality of the residuals.

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2} \tag{6}$$

where (6) is the variation of the ith predictor that is explained by the remaining n-1 predictors.

Next, we constructed models for the three homeless populations using a forward stepwise regression process. In each step, this stepwise regression process will consider a variable for addition from the set of explanatory variables based on some prespecified criterion. We constructed models using both Akaike Information Criterion (AIC) and then 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. Lower AIC or BIC values 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.

$$\mathrm{AIC} = 2k - 2\ln(\hat{L}) \tag{7}$$

$$\mathrm{BIC} = \ln(n)k - 2\ln(\hat{L}). \tag{8}$$

The forward step process begins with no predictors, and then proceeds to add predictors that will create a model with the lowest AIC and BIC values. Predictors are added individually until adding predictors no longer improves the calculated AIC or BIC for the model.

3.4 Neural Networks

We used neural networks to analyze the complex relationship between the time-dynamic data and the change in homeless population. Neural networks is a method that gets it's inspiration from the neurons of the human brain. The basic structure of a neural network is represented by the figure below.

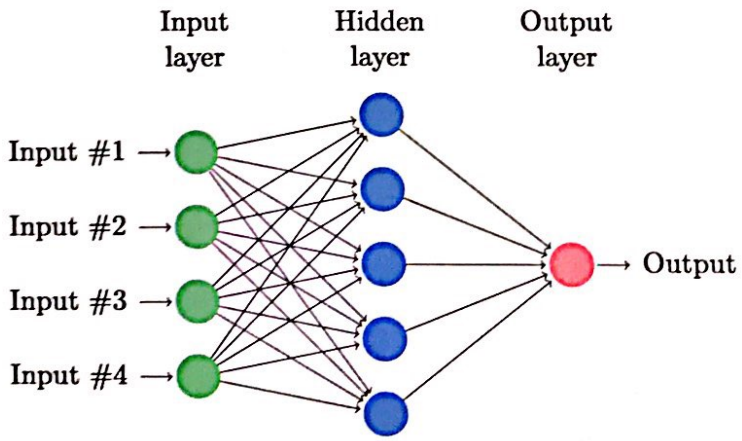


Figure 1: Structure of neural network

Imagine we have a size m data-set with n features with entries (x_i, y_i), where x_i is the feature value of a record and y_i is the correct output.

Scanned with CamScanner

Maybe give motivation for why we need W's and b's first...

Let's first stipulate some notations: We will use W^l_{jk} denote the entry at (j,k) in the weight matrix that connects the $(l-1)^{\text{th}}$ layer to the lth layer. Similarly, we use b^l_j for the bias at the jth position in the lth layer; use h^l for the activation function of the lth layer; use a^l_j for the activation for the activation at the jth position in the lth layer. `More detailed graph and the notation should be reflected on graph`

With these notation, we have the equation of the relation of activation at the jth position in the lth layer to the $(l-1)^{\text{th}}$ layer:

tell l looks nicer than l

$$z^l_j = W^l_{jk} a^{l-1}_k + b^l_j \qquad a^l = h^l(z^l) \tag{9}$$

We can rewrite the above equation with vectored form:

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

The activation functions used in out models are: relu, sigmoid, and softmax. The relu function output the original input if the input in positive, 0 if the input is negative:

$$\begin{cases} h(z) = z & \text{if } z \leq 0 \\ h(z) = 0 & \text{if } z < 0. \end{cases} \tag{10}$$

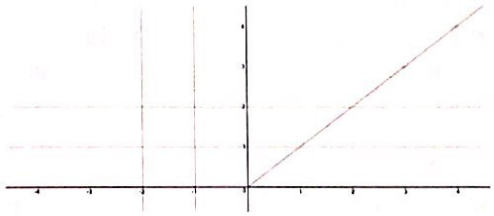


Figure 2: This figure shows the shape of the relu function.

The sigmoid function always output a number between -1 and 1. If has a steeper rate of increase when the input is near 0:

$$h(x) = \frac{1}{1+e^{-x}} \tag{11}$$

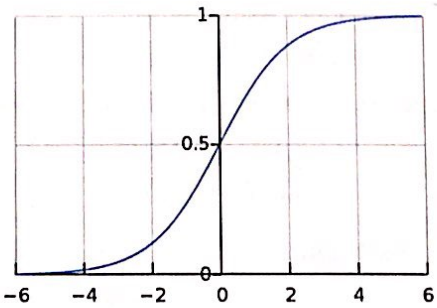


Figure 3: This figure shows the shape of the sigmoid function.

Scanned with CamScanner

The softmax function is a generalization of the logistic function that "squashes" 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 adds up to 1:

$$h(z_j) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

where K = ... ?

The algorithm learns how to best fit the data-set by trying to minimize a cost function. In our model, we used the cross-entropy cost function: `math needs more accurate`

yes...

$$C = -\frac{1}{N}\sum_{i}^{N}\sum_{j} y_{ij}\log(a_{ij}^{L}) \tag{12}$$

The method we used to minimize the cost is gradient descent. All weights and biases are updated against the direction of the cost function gradient for the values at that iteration. The learning rate α is the size of step each iteration takes. For example, from the ith iteration to the $(i+1)^{th}$ iteration, the weight and bias should be updated by:

DO...

$$W_{i+1}^{l} = W_{i}^{l} - \alpha\frac{\partial C}{\partial W_{i}^{l}} \qquad b_{i+1}^{l} = b_{i}^{l} - \alpha\frac{\partial C}{\partial b_{i}^{l}} \tag{13}$$

Adam is a variation of gradient descent. Adam calculates estimates of the mean and variance of the gradient and uses the ratio of the two to adjust the effective learning rate α for each parameter[5].

In order to actually compute the gradient of the cost function against each elements of the weight and bias in each layer, we apply the chain rule repeatedly. This way of finding the gradient is called backpropagation[6]. The chain rule can be written out as this:

this can be shortened ~~[illegible]~~

$$\frac{\partial C}{\partial b_{k}^{l}} = \frac{\partial C}{\partial a^{L}}\frac{\partial a^{L}}{\partial a^{L-1}}\frac{\partial a^{L-1}}{\partial a^{L-2}}\cdots\frac{\partial a^{l+1}}{\partial a^{l}}\cdot\frac{\partial a^{l}}{\partial b_{k}^{l}} \tag{14}$$

$$\frac{\partial C}{\partial W_{pq}^{l}} = \frac{\partial C}{\partial a^{L}}\frac{\partial a^{L}}{\partial a^{L-1}}\frac{\partial a^{L-1}}{\partial a^{L-2}}\cdots\frac{\partial a^{l+1}}{\partial a^{l}}\cdot\frac{\partial a^{l}}{\partial W_{pq}^{l}} \tag{15}$$

~~explain [illegible]~~

We thus need to derive each term in the previous formula.

those eqns are key you can explain the other computations at a high level - grs dets not required

To make writing these formulas easier, we will be using the Einstein Summation notation where repeated indices are implicitly summed over. For example:

$$a_{ik}a_{ij} \equiv \sum_{i} a_{ik}a_{ij}$$

We will also be using the the Kronecker Delta, i.e:

$$\delta_{ij} = \begin{cases} 1 & \text{when } i = j \\ 0 & \text{when } i \neq j. \end{cases}$$

From the equation of the cost function 12, we can see that: `Work out math`

$$\frac{\partial C}{\partial a_{ij}^{L}} = \frac{-1}{N}\frac{y_{ij}}{a_{ij}} \tag{16}$$

The activation function for the lth layer h^{l} is a function from z^{l} to a^{l}, therefore:

$$\frac{\partial a_{i}^{l}}{\partial z_{j}^{l}} = \frac{\partial h_{i}^{l}(z_{j}^{l})}{z_{j}^{l}} = \mathrm{J}_{\mathrm{ij}}^{\mathrm{l}}$$

$$\frac{\partial a^{l}}{\partial z^{l}} = \mathrm{J}^{\mathrm{l}} \quad \text{(the Jacobian matrix)} \tag{17}$$

2nd eq a bit more delicate due to the tensor stuff.

be careful w/ notation

Scanned with CamScanner

From equation 9, it is obvious that:

$$\frac{\partial z_j^{l,}}{\partial b_k^{l,}} = \delta_{jk} \tag{18}$$

Also from equation 9, we can see that:

$$\begin{aligned}\frac{\partial z_j^l}{\partial W_{pq}^l} &= \frac{\partial}{\partial W_{pq}^l}(W_{jk}^l a_k^{l-1} + b_j^l) \\ &= \delta_{jp}\delta_{kq}a_k^{l-1} = \delta_{jp}a_q^{l-1}\end{aligned} \tag{19}$$

We can thus derive the function for the gradient: $\frac{\partial a_i^l}{\partial b_k^l}$ and $\frac{\partial a_i^l}{\partial W_{pq}^l}$
From 17 and 18:

$$\frac{\partial a_i^l}{\partial b_k^l} = \left(\frac{\partial a_i^l}{\partial z_j^l}\right)\delta_{kj} = \frac{\partial a_i^l}{\partial z_k^l} \tag{20}$$

From 17 and 19:

$$\frac{\partial a_i^l}{\partial W_{pq}^l} = \left(\frac{\partial a_i^l}{\partial z_j^l}\right)\delta_{jp}a_q^{l-1} = \left(\frac{\partial a_i^l}{\partial z_p^l}\right)a_q^{l-1} \tag{21}$$

To derive the expression for $\frac{\partial a^l}{\partial a^{l-1}}$, we have:

$$\frac{\partial a_i^l}{\partial a_k^{l-1}} = \frac{\partial a_i^l}{\partial z_j^l}\frac{\partial z_j^l}{\partial a_k^{l-1}} \tag{22}$$

According to 9:

$$\frac{\partial z_j^l}{\partial a_k^{l-1}} = W_{jk}^l \tag{23}$$

Therefore we can use 17, 23 and 22 to calculate:

$$\frac{\partial a_i^l}{\partial a_k^{l-1}} = \mathrm{J}_{\mathrm{ij}}^{\mathrm{l}} W_{jk}^l \tag{24}$$

Now we have all the elements to calculate 14 and 15

For neural network, it is also good practice to normalize the input data so that the data is on the same scale. There are two method that we used: Z-score and Principle Component Analysis (PCA). The data matrix X is of dimension $\mathbb{R}^{m\times n}$, each row is a data point (m of them) and each point has n features.
Z-score normalization makes each feature (the column of our matrix) has 0 mean and 1 standard deviation. We first calculate the mean of of each column:

$$M_j = \frac{\sum_i X_{ij}}{m} \tag{25}$$

Scanned with CamScanner

Then we can calculate the standard deviation of the column:

$$S_j = \sqrt{\frac{\sum_i (X_{ij} - M_j)^2}{m-1}} \tag{26}$$

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

$$Z_{ij} = \frac{X_{ij} - M_j}{S_j} \tag{27}$$

We do this calculation for each number 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.

Principal component analysis (PCA) is a procedure that uses an orthogonal transformation to convert matrix variables into a set of linearly independent vectors called principal components. This transformation is defined in such a way that the first principal component has the largest possible variance (i.e.: accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an orthogonal basis. PCA is sensitive to the relative scaling of the original variables, therefore we use the normalized Z matrix derived above[7].

Physical sen w.r.t.

We use T to denote the full principal components decomposition matrix of Z. It is common nowadays to use single value decomposition (SVD) to derive the matrix T.

$$Z = U\Sigma W^T \tag{28}$$

$$T = ZW \tag{29}$$

And each column of T is the principle component of Z, and column with smaller index explain more variance in the data. `More detailed math? Like about why they have the largest variance`

why reliable??

There are 818 census tracks that all the data we desire to use are reliable. The features we used are in two groups:

- Static: Citations, Coffee, Restaurants,Shelters, Crime Count, Bus Stops, Tract size in Square Miles;
- Time Dynamic: Available Housing Units, Total Housing Units, Total Vacant Units, Unemployment Rate, Below Poverty Rate, Medium Rent As Percent Of Gross Income, Total Population, Medium Household Income, Medium Rent, Medium Value, Medium Monthly Housing Costs, ZRI, ZHVI.

For data in the static group, we took the value of 2016 and assumed that there is not much yearly changes. For data in the time dynamic group, we have data from 2014, 2015, and 2016.

We tried a ternary classification for the change of numbers of homeless populations in a census track. We implemented the neural networks in Tensorflow, a popular machine learning system[8].

First, we divide the change in homeless population into three buckets. Roughly, the lower bucket represent a significant decrease in homeless population, the middle bucket represent a small, insignificant variation in the homeless population, and the higher bucket represent a significant increase in homeless population.

We used static data and time dynamic data (single year and change between years), to try to classify the change in homeless population. More specifically, for the time dynamic data, we used the change from 2014 to 2015 and the data of 2015 to classify the change in homeless population

Scanned with CamScanner

from 2015 to 2016; and the change from 2015 to 2016 and the data of 2016 to classify the change in homeless population from 2016 to 2017. We used the change of features of previous years to predict the change in homeless population because (1) ACS did not release the data for 2017 yet, and (2) the homeless population data is gathered during January, changes in previous years are reflected in changed in the year later `[cite data source]`. We compiled the data of changes in two years together and split them up with a ratio and 70% and 30%, 70% of the data is used as the training set and 30% of the data is used as the testing set.

We tried different method of preprocessing the data:

- Normalization
- Principal Component Analysis (PCA)

We tried different cut off point for the buckets.
We tested different neural net work structure. We did trials on:

- Number of layers
- Number of neurons in each layer
- Activation function
  - Relu
  - Sigmoid
  - Softmax
- Learning rate
- Optimizer
  - stochastic gradient descent
  - Adam

lot of space - try reformatting, using a table?

4 Results

4.1 Topic Modeling

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:

- Figures should have a legend describing the image in detail.
- How were these plots generated? Explain.

Scanned with CamScanner

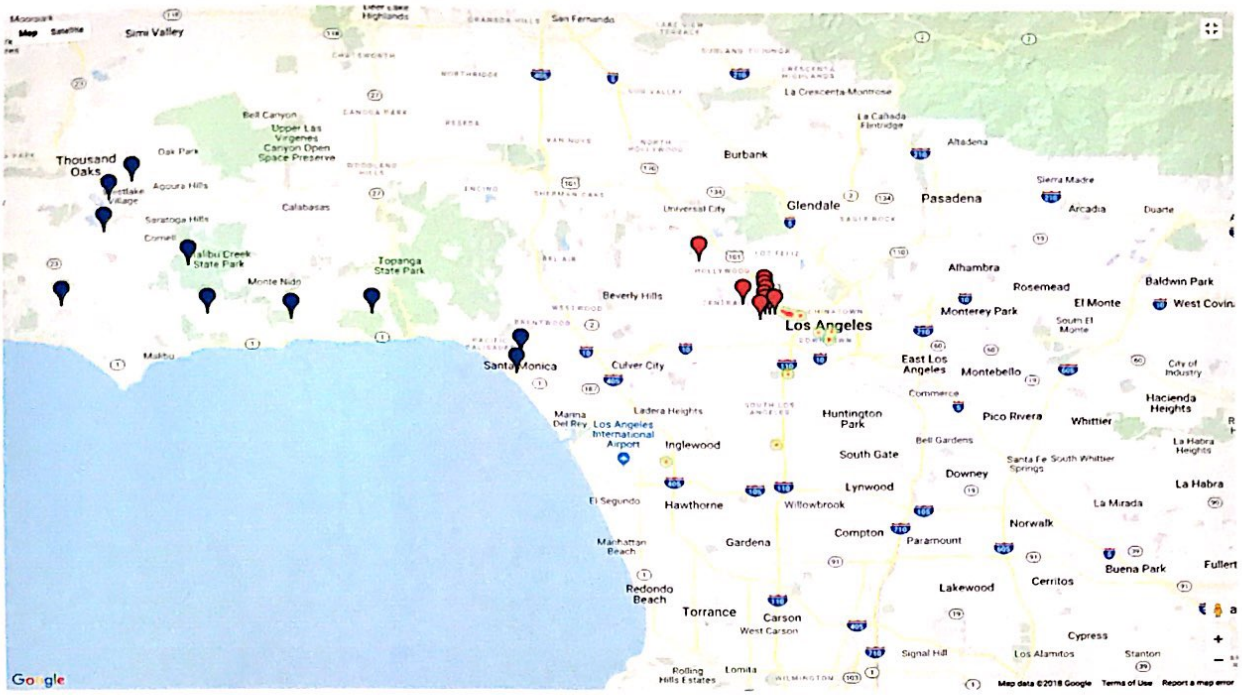


Figure 4a: $k = 2$

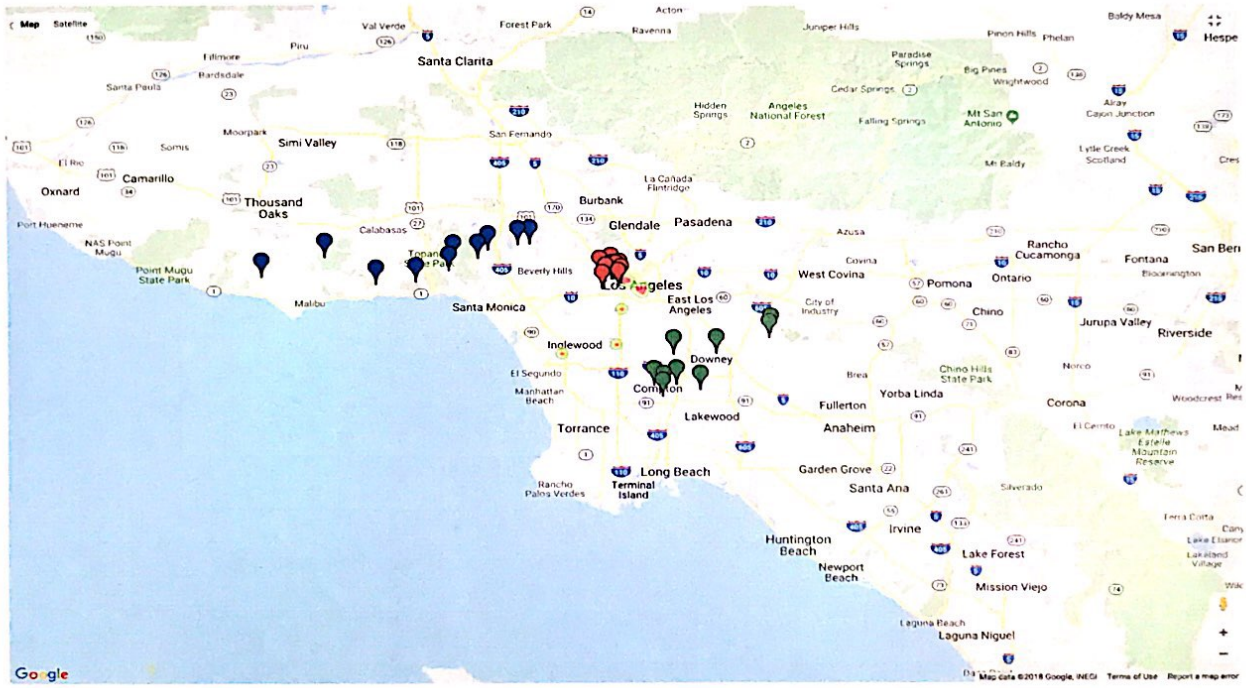


Figure 4b: $k = 3$

Scanned with CamScanner

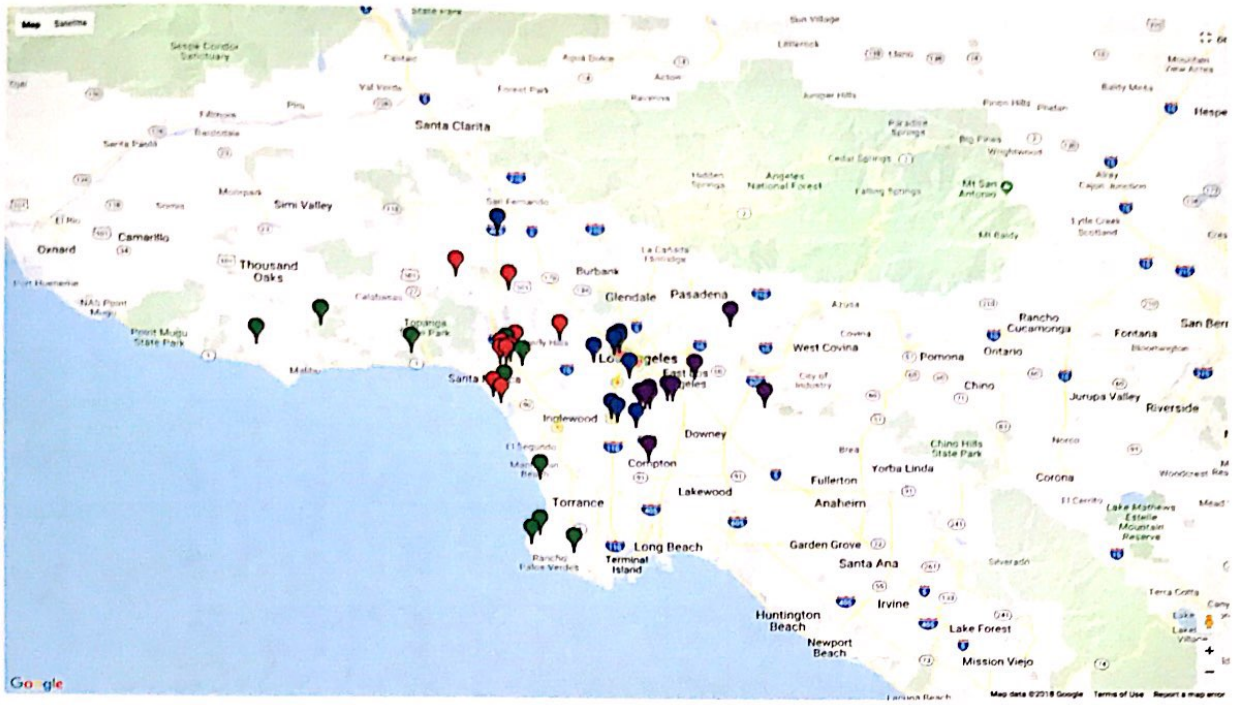


Figure 4c: $k = 4$

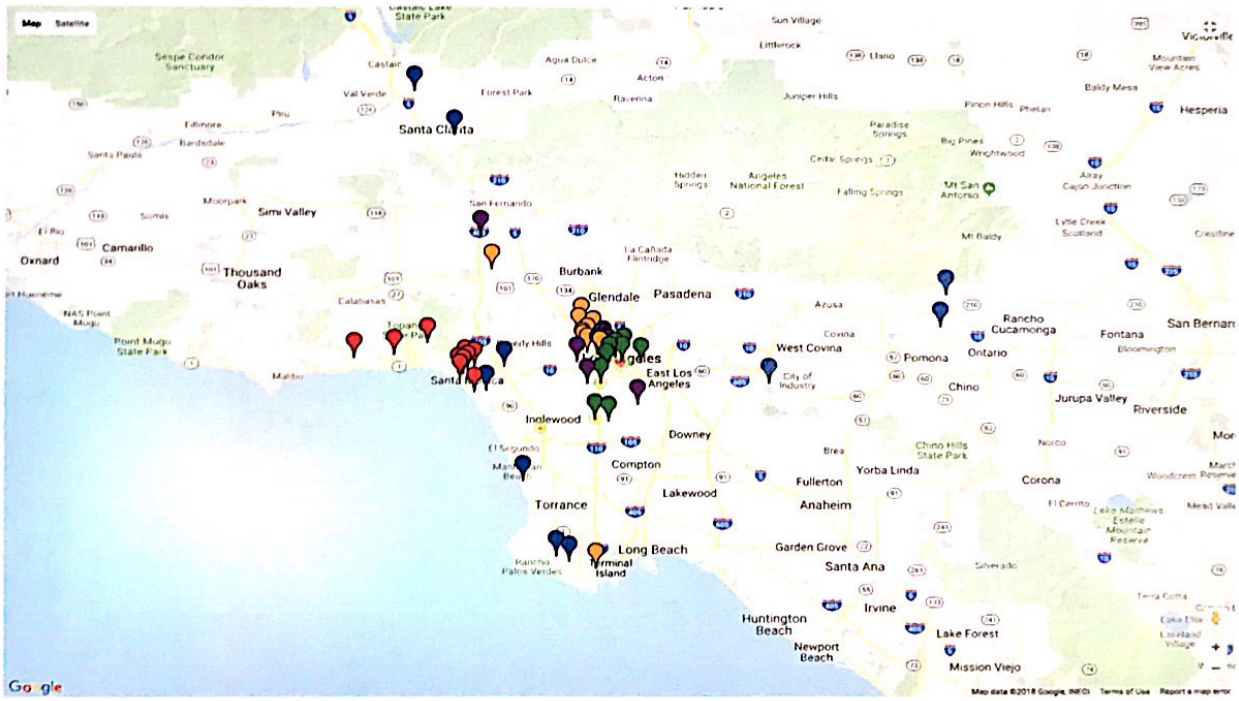


Figure 4d: $k = 5$

Rmk: seems like k=4 ideal (or between 3-5 at least)

Scanned with CamScanner

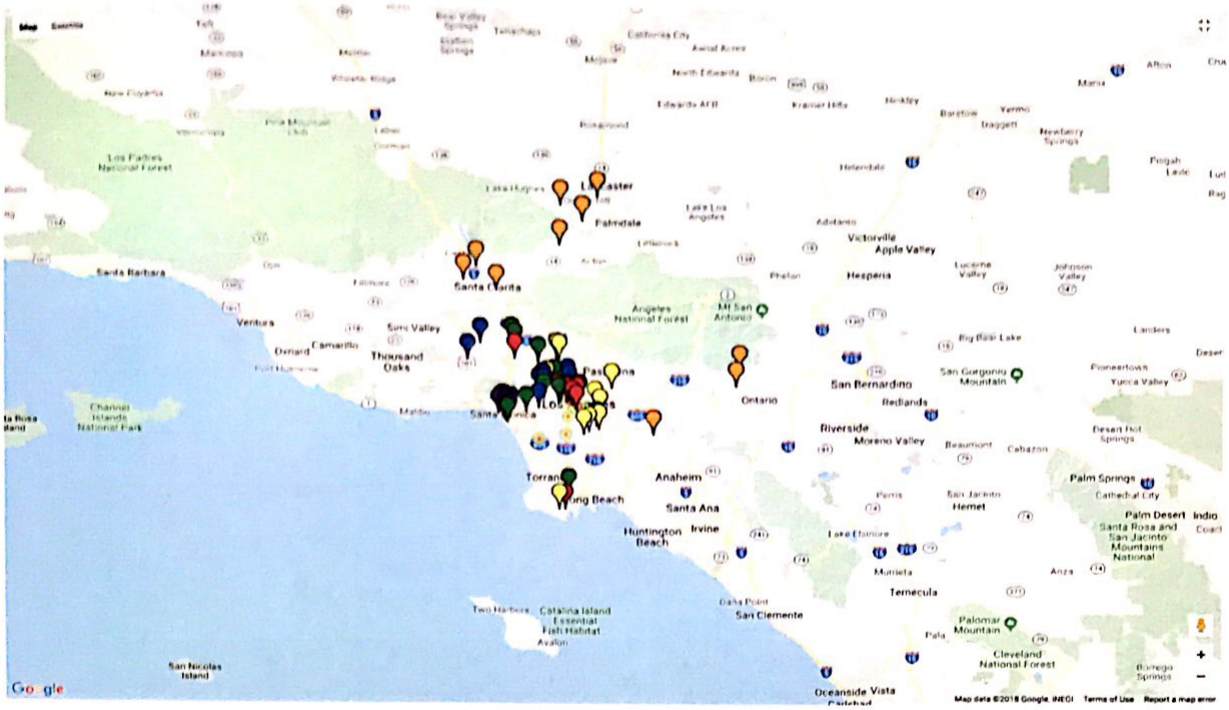


Figure 4e: $k = 6$

Figure 4f: $k = 7$

Scanned with CamScanner

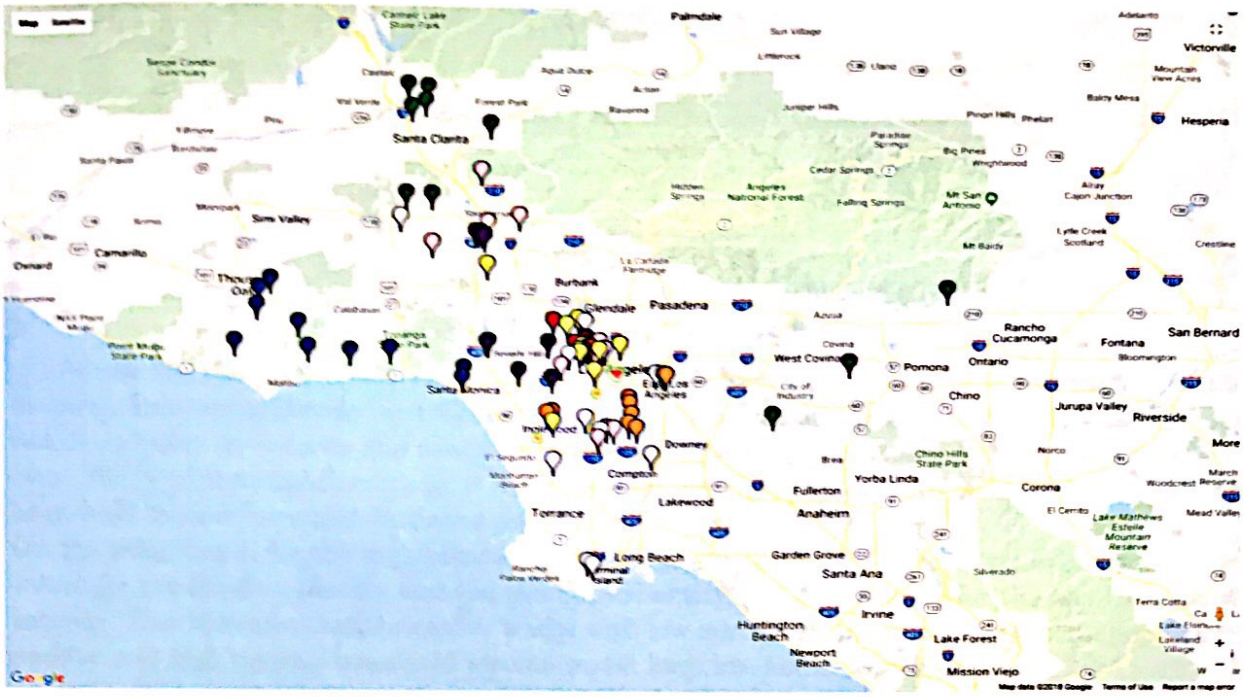


Figure 4g: $k = 8$

was this done on percentile or z-scores? Explain

Moreover, we conducted the non-negative matrix factorization(NMF) to obtain the correlation between the weights of each topic and the homeless population density. For simplicity sake, we will conduct the NMF for number of topics 3 to 5. Denote k$\in \{2, 5\}$ to be the number of topics. From the correlation between the weights of each topic and the homeless population density, for each k, we pick the topic matrix with the highest positive correlation and also pick the topic with highest negative correlation. Below are the matrices with the correlation ρ and the features of topic matrix, i.e coffee shop density and restaurant density.

look up fitting tables to width of page in LaTeX

For k= 3 where k=2 ?

2 rows?

ρ	CoffeeDensity	RestaurantDensity	ShelterDensity	CrimeDensity	HousingDensity	BusStopDensity
0.54697	0.18935	0.40078	0.17057	0.42859	0.2192	0.39935
-0.41816	0.23528	0.11055	0.23629	0.083083	0.18198	0

GenPop2015Density	ZRI	ZHVI	MedHouseholdIncome
0.37516	0.35213	0.34418	0
0	0.39602	0.4018	0.71993

For k= 4

It would be nice to see all k topics and to interpret each topic

2 rows

ρ	CoffeeDensity	RestaurantDensity	ShelterDensity	CrimeDensity	HousingDensity	BusStopDensity
0.61047	0	0.45189	0.2045	0.50138	0.28931	0.39708
-0.55292	0.000158	0.17842	0.44944	0.16865	0.44094	0

GenPop2015Density	ZRI	ZHVI	MedHouseholdIncome
0.36995	0.26004	0.23822	0
0	0	0	0.73709

Scanned with CamScanner

For k= 5

ρ	CoffeeDensity	RestaurantDensity	ShelterDensity	CrimeDensity	HousingDensity	BusStopDensity
0.50328	0	0.56601	0.098547	0.58106	0.20278	0.30022
-0.54216	0.26811	0.072089	0.44805	0.027862	0.38839	0

GenPop2015Density	ZRI	ZHVI	MedHouseholdIncome
0.081367	0.32986	0.29261	0
0.0077115	0	0	0.73709

Across the different **k**, for the high positive correlation, we can see that the values for the features, RestaurantDensity and CrimeDensity are high. This indicates that census tracks with relatively many restaurants and relatively high crime rates have a high homeless population. Moreover, the MedHouseholdIncome is 0 for all k signifies that the census tracks with low median household income have high homeless population. interesting

On the other hand, for the high negative correlations for all **k**, the topic matrix exhibits very low values for the BusStopDensity and the GenPop15Density, and a high value for the MedHousehold-Income. This indicates that the census tracks with low number of bus stops, low general population density and high median household income would have low homeless population. makes sense

From these correlations, we then converted the the census tracks into their respective longitudes and lattitudes using the census tracks as centroids. We plot the heat map graphs to look at the weights of each topic and the homeless population density to compare whether there is overlap of areas of these weights and the homeless population density. Since the weights represents the degree of topic on the census track, The more overlap between the weights and the homeless population density, the more likely that the area has strong presence of homeless population, and thus the more likely that the topic explain the homeless population.

-Nice figures.
-Still need more detailed captions
-would be nice to have this fig juxtaposed next to the high pos/neg correlation figures.

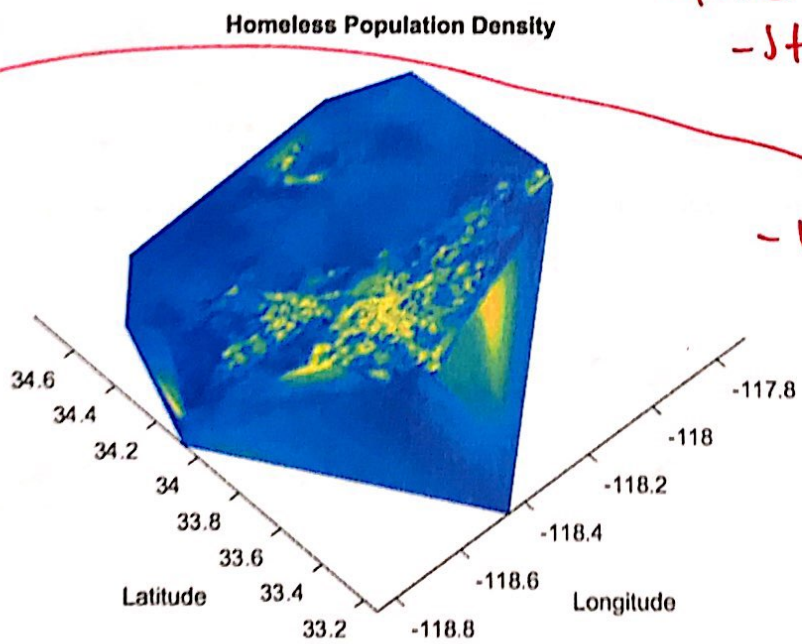
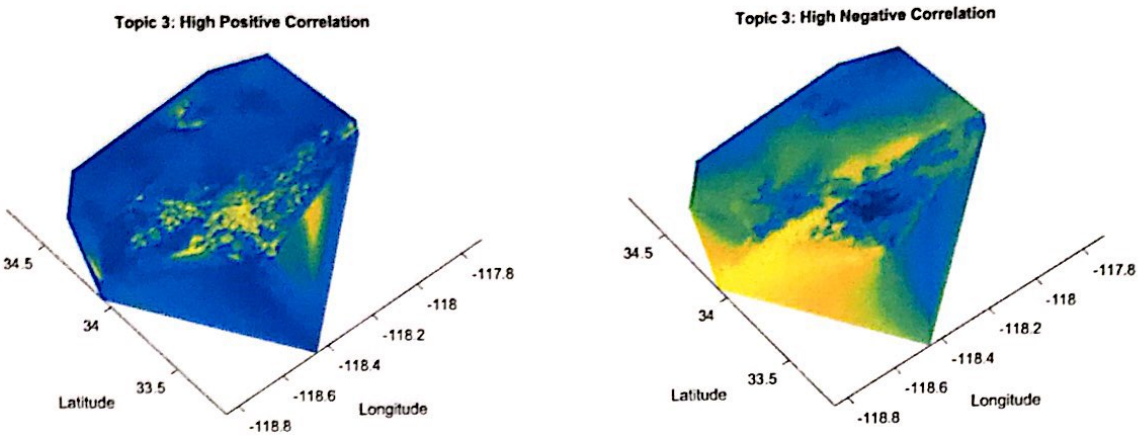


Figure 5: Homeless Population density

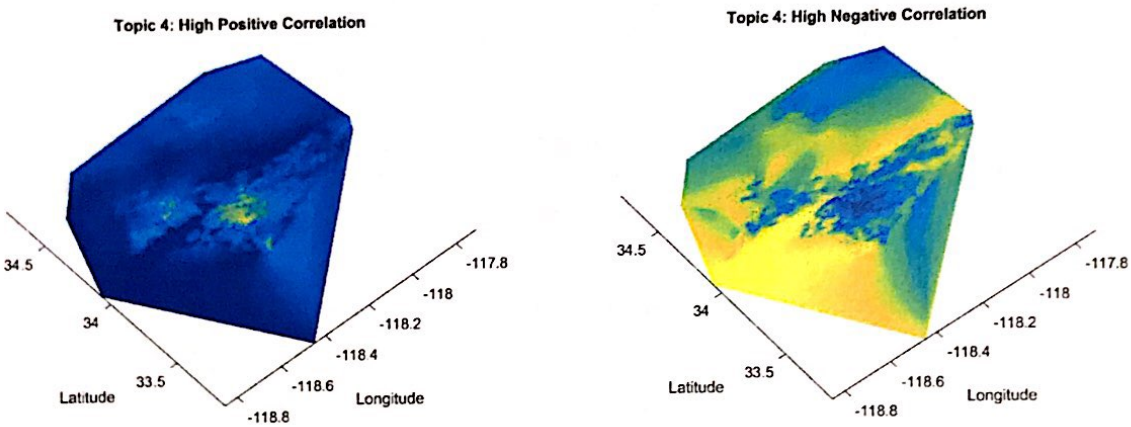
Scanned with CamScanner



(a) Highest Positive Correlation

(b) Highest Negative Correlation

Figure 6: Number of topics: 3



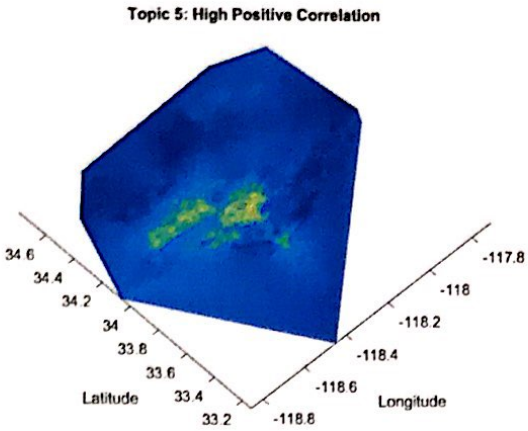
(a) Highest Positive Correlation

(b) Highest Negative Correlation

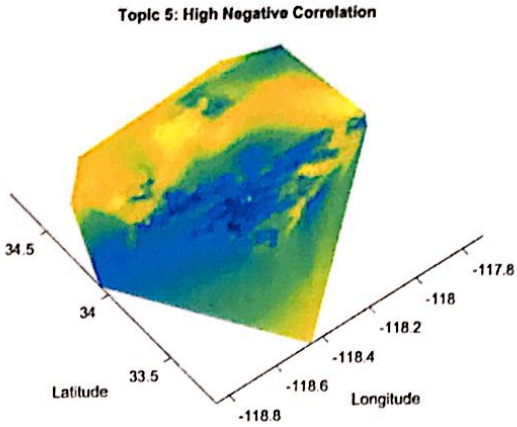
Figure 7: Number of topics: 4

So these are surfaces of topic weights : needs explanation.

Scanned with CamScanner



(a) Highest Positive Correlation



(b) Highest Negative Correlation

Figure 8: Number of topics: 5

The brighter the map is, the higher the weight/homeless population density. This shows that there is more overlap for topics with the highest positive correlation with the homeless population, while those with the highest negative correlation will have less overlap or close to no overlap, which is consistent to what we would expect.

We also compiled the local "variance" of a census track, where (1) The top 10 census tracks with the highest homeless population density, (2) The top 10 census tracks with the highest local variance, (3) The top 10 census tracks with the lowest local variance.

Figure 9: Table that shows the census tracks with the highest homeless population density, highest local variance and lowest local variance

(1)	(2)	(3)
206300	207302	134101
209104	207301	113422
207301	206200	119800
208720	206300	131300
208903	207502	131200
238320	226002	109700
206200	211802	480101
208801	208904	113421
277400	212101	111400

There are many overlaps between the first and the second column, i.e 206300, 207301, etc, which there are none between the first and the third column. This shows that there is clustering of homeless people in the census tracts with the highest variances. 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.

Scanned with CamScanner

just z-scores or more elaborate?

4.2 Linear Models

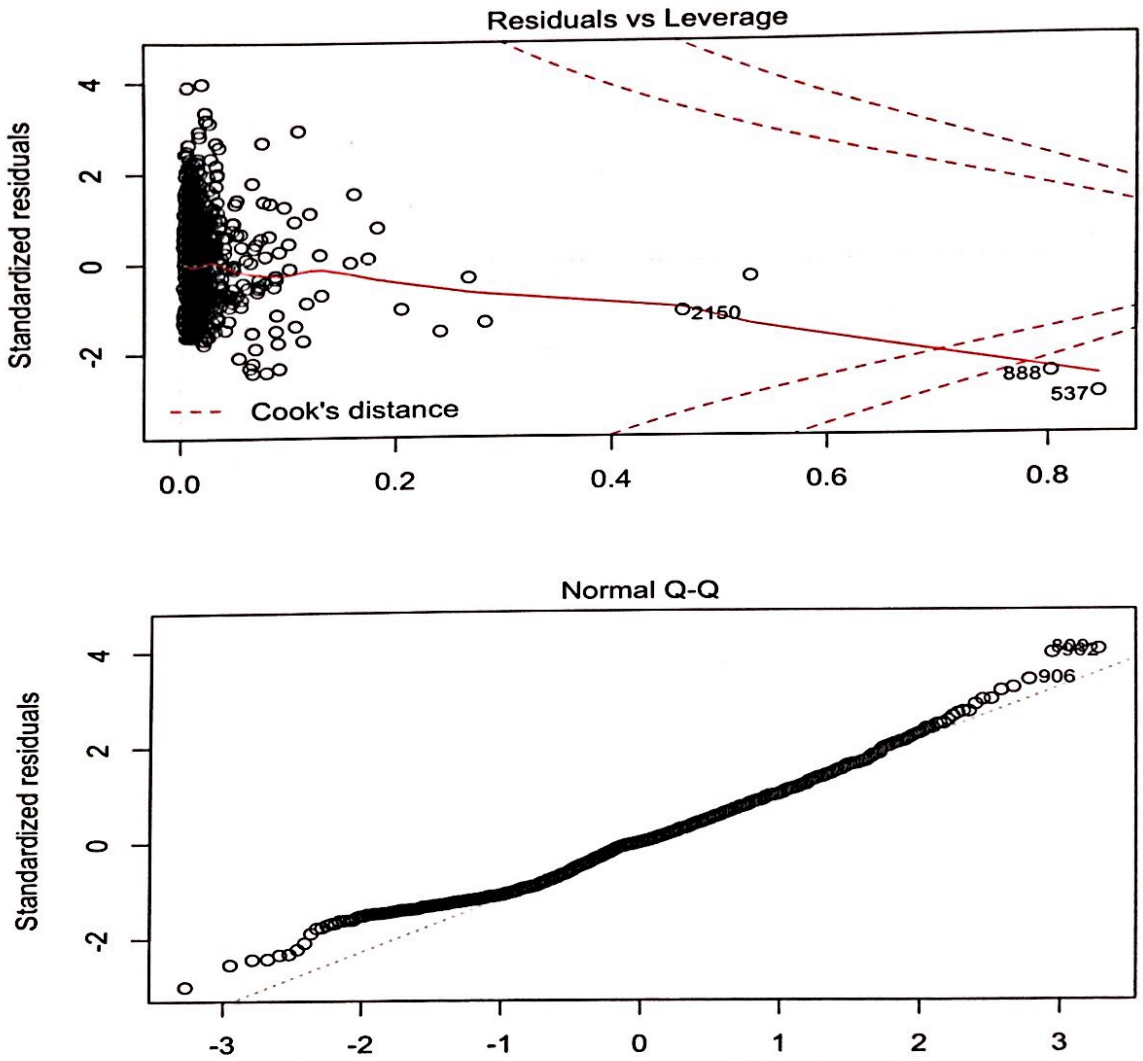
After eliminating predictors that had significant issues of high multicolinearity, and transforming the response variables to achieve constant variance and normally distributed residuals, we identified the statistically significant predictors for explaining variation in each category of a census tracts' homeless population.

I massive most data are not.. ~N(μ,σ²)

write words

The statistically significant predictors in explaining variation among the homeless population living in vehicles (cars, and campers) are: citations, totcars, medianincome, bus stops, and totencamp people. The Adjusted R-squared for the model explaining variation in vehicle homeless population with street and shelter homeless is 0.434. ← not bad in the social sciences...

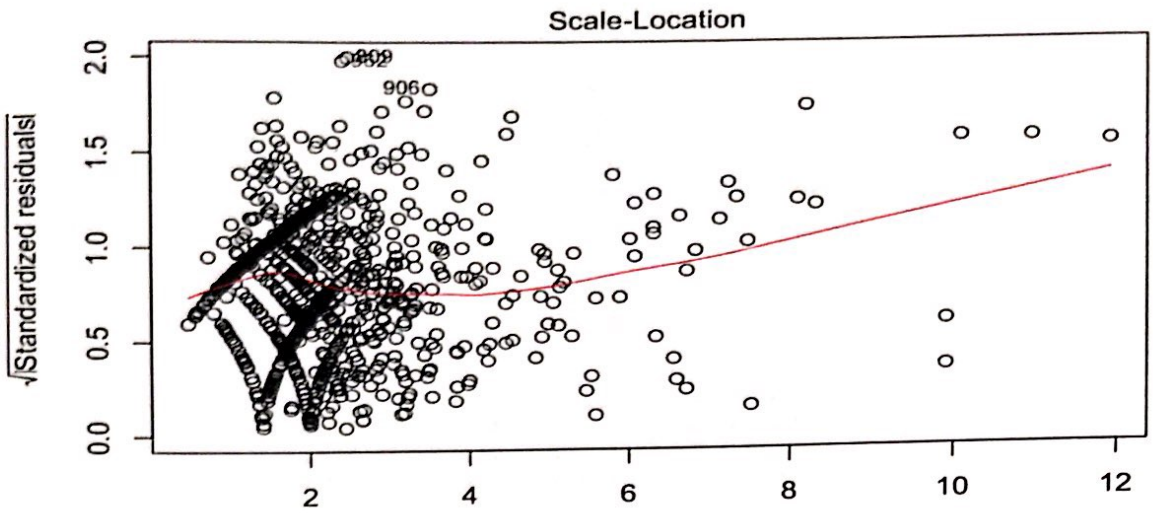
Figure 10: Diagnostic Plots for Model Predicting Homeless Population Living in Vehicles



need a lot more explanation here - what am I looking at?

But totCars is part of total vehicles. don't we expect it to be a predictor?

Scanned with CamScanner



From these diagnostic plots, we see that the residuals are normally distributed, and that the residuals are randomly distributed around a mean of zero with constant variance.

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 certain location.

nice interp.

The model for explaining variation in the shelter homeless population has restaurants, median home income, bus stops, tothomeless2015, tothomeless2016,tottencamp people as its statistically significant predictors. The adjusted R-squared for this model is 0.8321.

but not #shelters???

Scanned with CamScanner

Figure 12: Modeling Homeless Population Living in Shelters

```
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
```

Needs explanation

The statistically significant predictors for explaining the variation in a census tracts' homeless population living on the street in tents and in encampments are the 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 shelter. This model's Adjusted R-squared is 0.7536.

Rmk: · code, etc. can sometimes best be placed in an appendix

· pseudo-code is more suitable in the middle of a paper/report

Scanned with CamScanner

Figure 13: Modeling Homeless Population in Tents and Encampments

```
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
```

Next, we reconstructed models for the three homeless populations using a forward stepwise regression process. In each step, this stepwise regression process will consider a variable for addition from the set of explanatory variables based on some prespecified criterion. We constructed models using both AIC and BIC as criterions.

The selected predictors for modeling the homeless population living in shelters when using BIC as the criterion for inclusion are totCars, totEncampPeople, MedHouseIncomes, BusStops, and Coffee. The adjusted R-squared for the forward stepwise regression model created using BIC as a criterion is 0.4337.

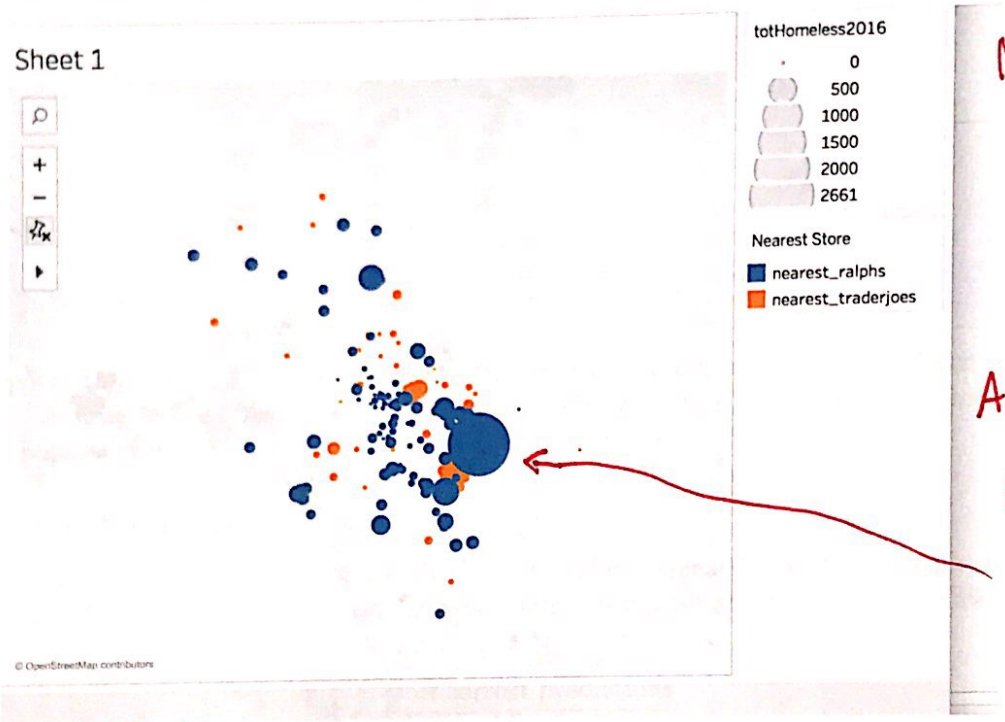
When using AIC to select predictors for the homeless population living in shelters, the Adjusted R-squared only slightly increased to 0.4348, and the number of predictors selected increased to include totCamperPeople, totVanPeople, totCarPeople, totCars , MedHouseIncomes, X2015HousedPopulation , nearest_library, Bus Stops, Coffee , nearest_traderjoes , X2015ZRI, and X2017ZRI. This is due to AICs tendency to overfit because of its smaller penalty for increasing the number of predictors.

Scanned with CamScanner

The selected predictors for modeling the homeless population living on the street using AIC as a criterion are totHomeless2016, totVanPeople, CrimeCount, X2015HousedPopulation, BusStops, HousingUnits, totSheltPeople , totHomeless2015 , nearest_library, coffee , nearest_traderjoes, and totCars.

The selected predictors for explaining the variation in the homeless population living on the street using forward stepwise regression with BIC as the criterion are totHomeless2016, totVanPeople, CrimeCount , X2015HousedPopulation , BusStops, HousingUnits. Again, we see that models constructed using BIC include fewer predictors because of BICs greater penalty for increasing the number of predictors.

Figure 14: Visualizing Size of Homeless Populations With The Brand of Nearest Grocery Store



Nice visual!

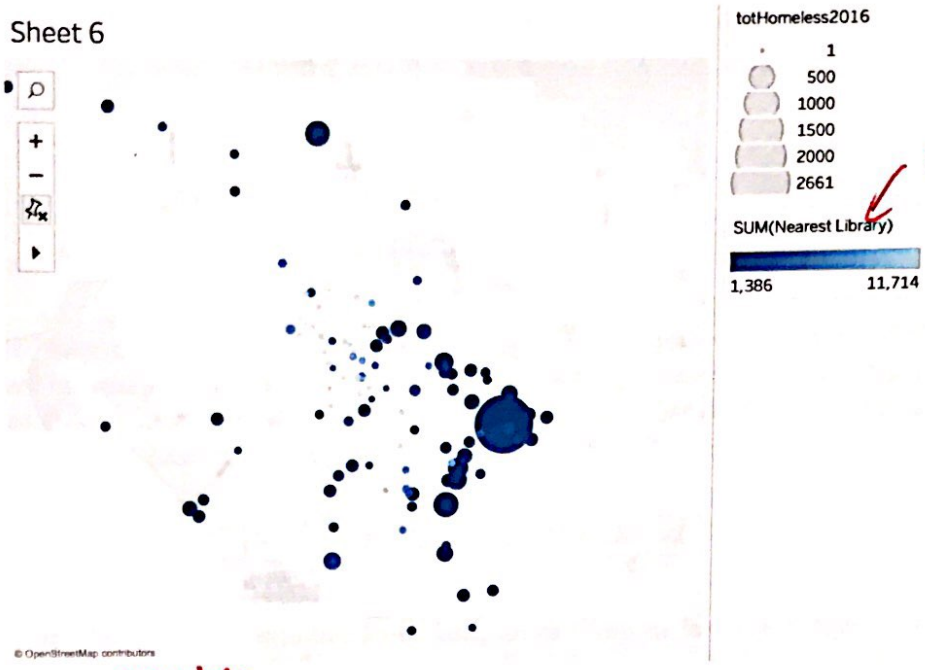
Are there any orange under the blue?

\caption

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.

Scanned with CamScanner

Figure 15: Visualizing Size of Homeless Populations With Proximity to Public Libraries



what units?

\cap tion

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

4.3 Neural Networks

grammar wording

Because of the uneven distribution of the bins, we use different measurement to reflex how well the algorithm do. Apart from the accuracy, we have confusion matrix and precision. Accuracy can be calculated by:

$$A = \frac{\#\text{ of correct predictions}}{\#\text{ of all predictions}} \tag{30}$$

A confusion matrix C is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. The (i, j) entry of the matrix represents the number of point whose true value is the ith class and it was predicted to the jth class. Precision$i$ is the ratio of true predictions among all the points that was predicted to be in the ith class:

notation

$$\text{Precision i} = \frac{C[i,i]}{\sum_j C[i,j]} \tag{31}$$

The table below shows the best result of the ternary classification for bucket division: [-400,-20), [-20,20), [20,600).

Model	Preprocessing	Accuracy	Training Cost	Testing Cost	Hidden	Output
A	PCA	0.725	0.359	0.501	-	Sigmoid

With a confusion matrix and precision:

Scanned with CamScanner

-	True 0	True 1	True 2
Pred 0	4	19	2
Pred 1	46	346	61
Pred 2	2	4	3

Precision0	Precision1	Precision2
0.160	0.764	0.333

We need to take more care when we look at the precision numbers. Because of the imbalance of the numbers of data points in each bin, if we have a purely chance predictor, the precision will be the ratio of the number of data point in that bin to the number of data point in total. Thus a pure chance model will have a precision:

Precision0	Precision1	Precision2
0.070	0.758	0.136

Thus even though the precision number looks bad, all of them in fact are bigger than the change model. Therefore our model is performing better than chance at predicting change in homeless population.

We change the cutoff point so that the three bins to have roughly equal amount. The table below shows the best result of the ternary classification for bucket division: [-400,-3), [-3,6), [6,600).

Model	Preprocessing	Accuracy	Training Cost	Testing Cost	Hidden	Output
B	Normalization	0.389	0.397	0.914	relu	Sigmoid

With a confusion matrix and precision:

-	True 0	True 1	True 2
Pred 0	60	49	60
Pred 1	58	79	48
Pred 2	48	35	51

← more promising...

Precision0	Precision1	Precision2
0.355	0.427	0.380

A pure chance model for this one would result in a precision of:

Precision0	Precision1	Precision2
0.333	0.333	0.333

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

Scanned with CamScanner

5 Summary and Future Work (Needs to be thought out more.)

5.1 Earth Mover's Distance

Further geospatial analysis is to come with the topic modelling plots using the earth mover's distance (EMD), or the Wassertein metric. The earth mover's distance describes the cost of moving from one geospatial cluster to the a new geospatial cluster. In our applications this can be used to determine the cost of moving to one census tract to the next or from one cluster of topic model to another.

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

← notation

Our goal is to find flow $\mathbf{F} = [f_{ij}]$ with $f_{i,j}$ being the flow between p_i and q_j that minimizes the overall cost. This can be done by solving the optimization function:

$$\underline{min} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j} d_{i,j}$$

\min in TeX makes it look better.

5.2 Neural Network Exploration

Future research should explore more method of prepossessing the data. Considering the error in public database, we could put the feature data into bins as well (e.g.: income can be slit into 5 bins according to it's standard deviation). ? huh

5.3 Convolutional Neural Network

Convolutional neural network is also promising in the context of analyzing the change in homeless population because homeless population move around frequently. Change in one area will change the number of homeless population in other areas.

6 Acknowledgments

We thank our adviser and mentor, Professor Michael Lindstrom, for all of his advice and help on this project.

+ Derek!!!

+ Marcus Roper as faculty sponsor

Scanned with CamScanner

References

[1] 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.

[2] Lynn Rew. "Correlates of Resilience in Homeless Adolescents". In: 33.1 (2001), pp. 33–40.

[3] Thomas P. O'Toole Stephen Metraux Dan Treglia. "Migration by Veterans Who Received Homeless Services From the Department of Veterans Affairs". In: (2016).

[4] Daniel D. Lee and H. Sebastian Seung. "Algorithms for Non-negative Matrix Factorization". In: *Advances in Neural Information Processing Systems 13*. Ed. by T. K. Leen, T. G. Dietterich, and V. Tresp. MIT Press, 2001, pp. 556–562.

[5] Jimmy Ba Diederik P. Kingma. "Adam: A Method for Stochastic Optimization". In: (2014). . . .

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

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

[8] Martín Abadi etc. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems". In: ().

maybe find a few more: aim for ~15

Scanned with CamScanner

Plagiarism aside, some scores 0-5 for my impressions reading this (based on the standards I would expect from published research papers):

- Background/Mathematical Intuition/Motivation : 1/5
- Clear explanations/Mathematics/Notation : 2/5
- Clarity of the Science/Level of Detail and exposition : 1/5
- Interpretation of results : 2/5
- Readability/Flow/Spelling/Grammar: 2/5

Overall: $\frac{8}{25} \xrightarrow{\text{plagiarism}} \frac{0}{25}$

1. Plagiarism is serious. Do not let it happen again. Academically, it can end your career if you go on in academia, plus it is unethical.
2. I know you were rushed. For the final report, read each other's work and help with spelling/grammar/consistent tense, consistent notation, etc. Try to have it all fit together and tell a complete story.
3. It never hurts to explain what a technique does intuitively first: ex NMF: what is a "topic", etc.
4. There are already some cool results but I don't think they were discussed much.
5. When you write, put yourself in the shoes of someone who may read it: they may know some math, but not necessarily the techniques you used, and they may know nothing about the application.
6. Be very clear on what data you used when, what software, etc
7. For a first report, this is very well-done so don't see this as failure but as feedback to succeed :)

Scanned with CamScanner