Access to Public Transportation in California

Kabir Snell, Ryan Sevilla, Jaime Gomez, and Joseph Momich.

Author contributions

Name	Role
Kabir Snell	Found the data set, wrote the background, and came up with questions for future work
Ryan Sevilla	Worked on variable summaries and explanatory plots
Jaime Gomez	Ensured guidelines were met and worked on final completion of the report
Joseph Momich	Provided data description, helped transform the data, and assisted in data exploration & analysis

Abstract

Well-designed public transit plans help improve community heath, reduce air pollution, and provide economic benefits to the community. California's public transportation system services residents in both urban and rural areas, but not all residents have equal access to it. The aim of this analysis was to identify how access to public transit differs among location and ethnicity. In addition, it was to determine if either population or ethnicity are significant predictors of access to transit. Access to public transit differs between each region and county in California. However, a location's ethnic demographic and population alone, are not significant predictors of access to transit.

Background

California has a vast public transportation system that includes buses, trains, light rail, and subways. The state's largest public transportation agency is the Los Angeles County Metropolitan Transportation Authority (Metro), which serves the Los Angeles metropolitan area with an extensive network of bus and rail lines. Other major public transportation agencies in California include the San Francisco Municipal Transportation Agency and the San Diego Metropolitan Transit System. In addition to these larger agencies, many smaller cities and towns throughout California have their own public transportation systems, which may include buses, shuttles, or other types of services. These systems often provide connections to regional or statewide transportation networks, making it possible for people to travel throughout the state using public transportation. While many areas have well-established public transportation systems, there may be gaps or limitations in service in some parts of the state, particularly in more rural or remote areas.

Every 10 years, the California Department of Public Health collects public transportation data, as part of the "Health Communities Data and Indicators Project (HCI). The goal of the project is to evaluate how city plans and policies affect community health. In particular, it provides information on the number of households that are within a certain distance of public transit stops. This information can be useful for understanding the accessibility of public transportation in different areas of the state, and for identifying opportunities to improve transit service or expand transit infrastructure. Overall, it is a valuable resource for policymakers, transportation planners, and researchers who are interested in understanding the availability and accessibility of public transportation in California.

Aims

There were two main objectives in the data analysis. First, it was to visuaize how access to public transportation access varies across geographic location and ethnicity group. The next was to determine if either population or ethnicity are significant indicators in predicting access rates. In order to approach these objectives, the data set was grouped by each level of geography (region, county, town). To determine how public transportation varies, visualizations like bar charts and scatter plots were used. Single and multiple linear regression models were fit to the data to determine how location and ethnicity explain access rates.

The findings show that access to public transportation varies significantly across the four regions and counties analyzed. In certain counties with high access rates, the ethnic demographic is segmented, but this is less apparent on the regional level. However, the data is very scattered between a location's population and its public transit access rates. **Overall, the analysis provided negative results.** Ethnicity and population are not significant enough indicators to predict the percent of residents that reside within 1/2 mile of public transportation.

Datasets

The data set shows the percent of the population that resides within 1/2 mile of a major transit location in four California regions, and whose waiting time is less than 15 minutes during peak commute hours. The data is stratified by 8 race/ethnicity groups and includes both geographic information and statistical reliability measurements.

The data includes 2012 Transit Stops from the San Diego and Southern California Association of Governments, as well as the Metropolitan Transportation Commission. It also includes 2008 Transit Stops from the Sacramento Council of Government and 2010 block-level population data from the U.S. Census Bureau. The four California regions are defined as the following:

- Southern California (SCAG): Imperial, Los Angeles, Orange, Riverside, San Bernardino, and Ventura
- Sacramento (SACOG): Placer, Sacramento, and Yolo
- Bay Area (MTC): Contra Costa, Marin, Napa, San Francisco, San Mateo, Santa Clara, Solano, Sonoma
- San Diego County

Data values were obtained using automated methods to download information from various public websites. In order to compile them into one data set, the census blocks from the 2010 U.S. Census were merged with the blocks from the travel surveys. One important survey was the 2010-2012 California Household Travel Survey. Multiple data collection methods were used in this survey, including computer-assisted telephone interviews and online/mail surveys. To identify census blocks inside 1/2 mile of the transit stops, geospatial software was used. The data was processed into Excel files with standard formats.

The population is adults aged 18 years and over, who reside in the four California regions. The sampling frame includes adults in these four regions, with access to telephone or mail services. The sampling mechanism for the respective year (2008 or 2012) is a probability sample because the surveys downloaded by the HCI project were sent to randomly selected adults However, the scope of inference has limitations. The data is from the year 2012 for the SCAG, MTC, and San Diego regions; while, 2008 for the SACOG region. Some transit stops and services may have changed during that time period. As well, the population data was collected from the 2010 U.S. Census, which is a different time period than the transit data (2008, 2012). Therefore, some variation may exist if demographics changed. The following table provides a summary of the variables used for analysis:

Name	Variable description	Type	Units of measurement
year	year when data was reported	Numeric	Calendar year
race_eth_name	name of the different races/ethnicties ('AfricanAm', 'AIAN', 'Asian', 'Latino', 'Multiple', 'NHOPI', 'Other', 'Total', 'White')	Object	Name
geotype	describes the level of geography for data in that row ('RE'=region, 'CT'=census tract, 'PL'=place/town/city, 'CO'=county)	Object	Name
geoname	name of the city/town	Object	Name
county_name	name of the county	Object	Name
region_name	name of the region ('Sacramento Area', 'Bay Area', 'San Diego', 'Southern California')	Object	Name
pop_trans_acc	number of residents that live within 1/2 mile of public transportation	Numeric	Integer
pop2010	total number of residents that reside in that county	Numeric	Integer
p_trans_acc	the percent of residents that live within 1/2 mile of public transportation	Numeric	Float

The data set used for analysis is show below:

	year	race_eth_name	geotype	geoname	county_name	region_name	pop_trans_acc	pop2010	p_trans_acc
0	2008	AfricanAm	CO	Placer	Placer	Sacramento Area	55	4427	0.012424
1	2008	AIAN	CO	Placer	Placer	Sacramento Area	51	2080	0.024519
2	2008	Asian	CO	Placer	Placer	Sacramento Area	117	19963	0.005861
3	2008	Latino	CO	Placer	Placer	Sacramento Area	1835	44710	0.041042
4	2008	Multiple	CO	Placer	Placer	Sacramento Area	241	10658	0.022612

Methods

In the exploratory analysis, the data was grouped by each of region and county. First, in each region the total population and the mean access to transit were measured and displayed in a table. The same process was conducted on the county-level data, but displayed with a scatter plot. Then, the relationship between ethnicity in each region and access to transit was explored with a bar chart.

In the analysis section, a simple linear regression model was fit to determine if the 2010 population influenced access to transit in different counties. The data was then modified to include ethnicity "rates". These rates were calculated from the original data set and represent the percent of each race/ethnicity group in the county or town. A multiple linear regression model was fit to this data, including ethnicity rates and population as predictors. This MLR model was fit to both county-level and town-level data, which had a much larger sample size. Lastly, PCA was used on the town-level data, to determine if a smaller subset of the variables better explain the data.

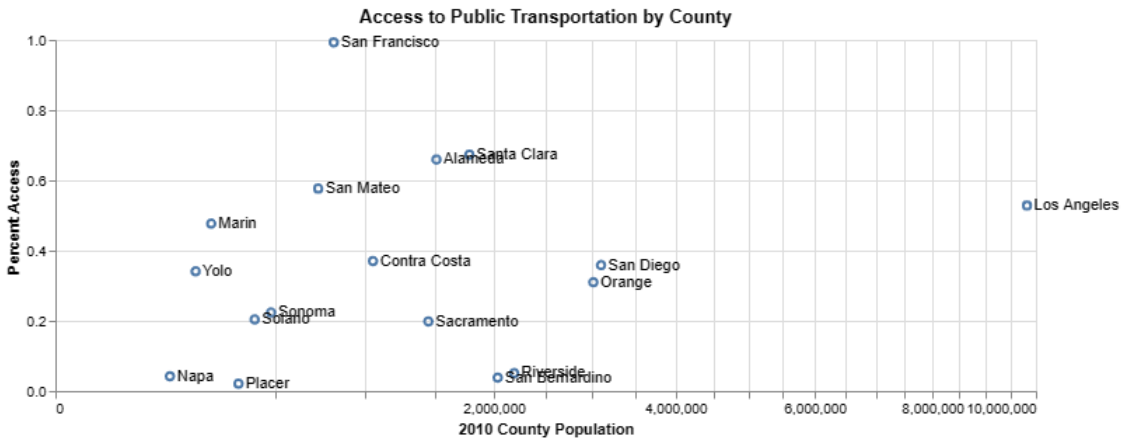
Results

Exploratory Analysis

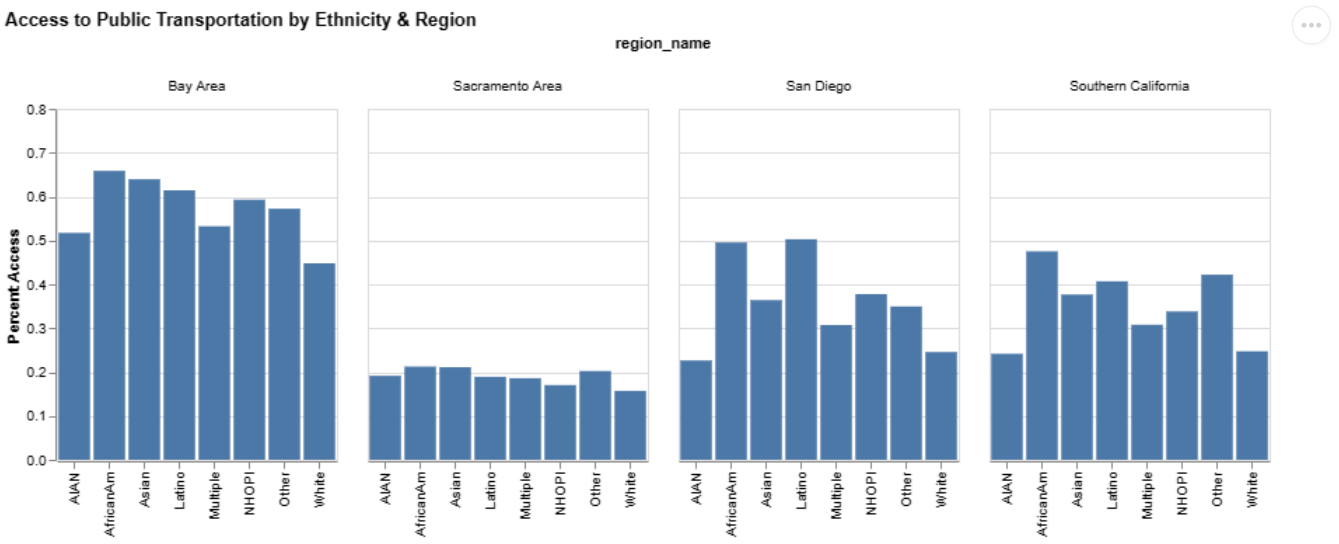
The data was first sorted by the four regions and grouped by 2010 population and percent access to transit.

	Region	2010 Population	Percent Access
0	Bay Area	7150739	0.571932
1	Sacramento Area	1999270	0.190002
2	San Diego	3095313	0.358582
3	Southern California	18051534	0.351986

The table shows that the Bay Area has the highest percent of residents that reside within 1/2 mile to public transit. The Sacramento region has both the smallest population and lowest access. While San Diego and Southern California have similar rates, Southern California has a much greater population. The table provides a general overview and there could be counties that are outliers within each region. A similar approach was used for county-level data and the results were displayed in a scatter plot:



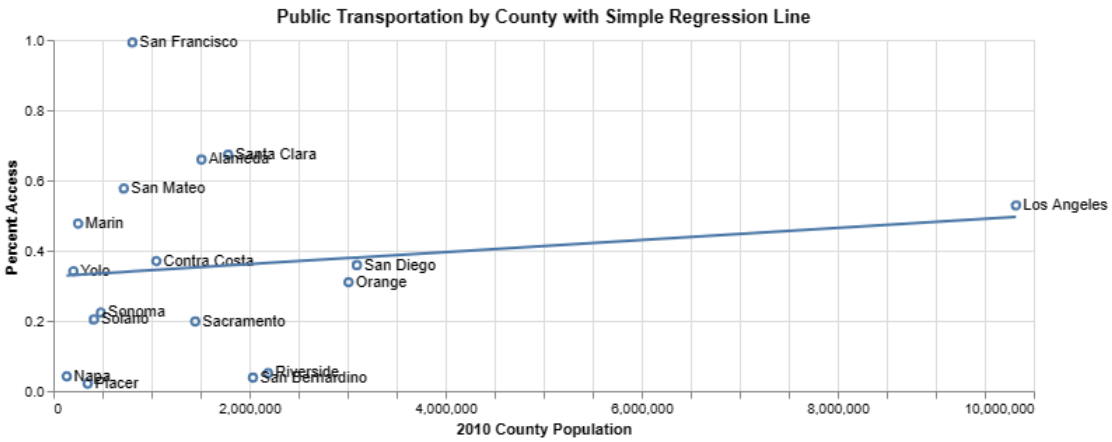
Nearly 100% of residents within San Francisco live within 1/2 mile of public transportation. In more rural counties, like Napa and Placer, rates are lower. Los Angeles county has a substantially larger population than the other counties, and has similar transit rates to more urban Bay Area counties like Alameda and Santa Clara. Next, the regional data was stratified by ethnicity, in order to determine if certain ethnicities live closer to transit.



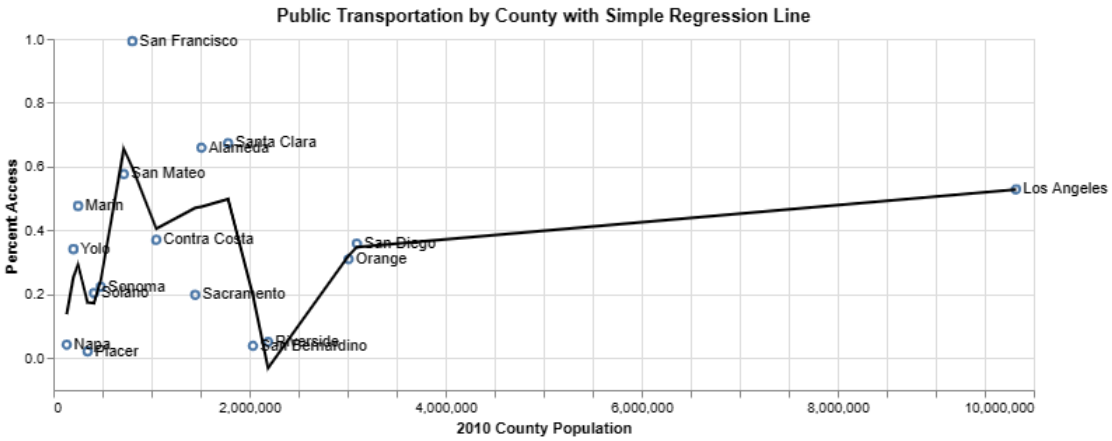
While there are differences amongst the race/ethnicity groups, the chart shows that **region is a more significant indicator** in whether residents live within 1/2 mile of public transit. Across all regions, African Americans have the highest access to public transit. However, this access varies significantly by individual county and in the data analysis below.

Data Analyis

Populaton and Access to Transit in Individual Counties: The first part of the analysis was to determine whether access to public transportation could be predicted based on the population of each county. A simple linear regression model was used to model 'p_trans_acc' against 'pop2010.'



Adding the regression line to the scatter plot above, demonstrates that a simple linear regression model does not fit the data well. The R^2 value is appoximately 0.02 meaning the 2010 population in each county does not explain the variation in public transportation rates. Next, a smooth curve was fit between the two variables.

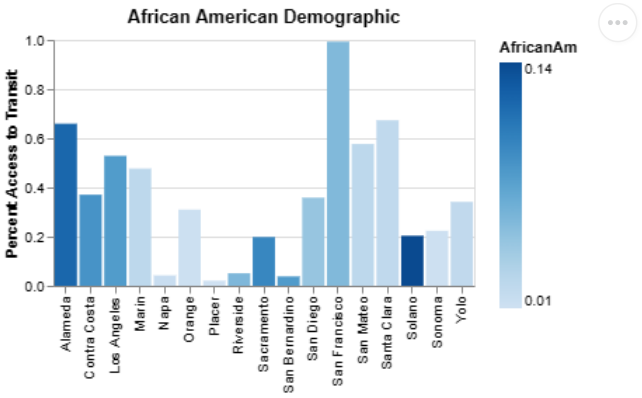


Even with this smooth curve added, there is no clear relationship between county population and access to transit.

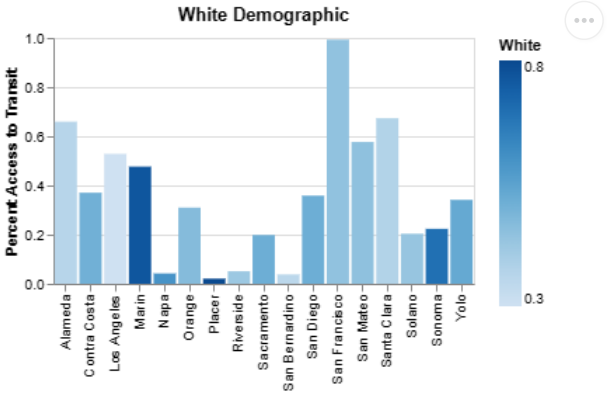
Analyzing Ethnicities: The next step was to determine the demographics in each county. The original data table was pivoted to show the percent of each ethnicity in the counties.

	county_name	AIAN	AfricanAm	Asian	Latino	Multiple	NHOPI	Other	White	pop2010	p_trans_acc
11	San Francisco	0.002270	0.058096	0.329966	0.151228	0.032387	0.003885	0.003097	0.419071	805235	0.992901
13	Santa Clara	0.002269	0.023760	0.317385	0.268971	0.030059	0.003509	0.002176	0.351871	1781642	0.673294
0	Alameda	0.002774	0.121916	0.258579	0.225052	0.040299	0.007900	0.002775	0.340706	1510271	0.659370

For example, in San Francisco, a county of 805,235 residents, White residents are ~42% of the population. Whereas, in Los Angeles, they are only ~28% of the population. The scatter splot above did not show how ethnicity varies by county. A bar chart of the counties, the percent access, and the ethnicity were then constructed. The following is for the African American demographic.



In all of the counties, African Americans are never more than 15% of the population. They represent the greatest percentage in Alameda (high transit access) and Solano (low transit access). In other counties like Marin, San Mateo, and Santa Clara, where transit acess is very high, they represent a very small percent of the population. This chart is significant becasue it shows how certain counties are ethnically segregated and how African Americans may only have high access rates in some of them.



White residents make up a significant portion of each county. Unlike African Americans, in Marin, they represent an extremely high percent of the population (> 80%). In rural areas, like Sonoma and Placer, they also represent a large part of the population. Compared to African Anmericans, White residents have high access in affluent areas (Marin, San Mateo, Santa Clara) and low access in rural areas (Sonoma, Placer).

Multiple Linear Regression: A multiple linear regression model was then fit to the data, with ethnicity rates as added predictor variables. Only 'White' and 'pop2010' were used as the explanatory variables at first.

The following table shows the county, the ethnicity rates, and the fitted values for 'p_trans_acc' (access rates). It also includes the residuals used for the R^2 analysis.

	county_name	AIAN	AfricanAm	Asian	Latino	Multiple	NHOPI	Other	White	pop2010	p_trans_acc	fitted_mlr	resid_mlr
11	San Francisco	0.002270	0.058096	0.329966	0.151228	0.032387	0.003885	0.003097	0.419071	805235	0.992901	0.401179	0.591721
13	Santa Clara	0.002269	0.023760	0.317385	0.268971	0.030059	0.003509	0.002176	0.351871	1781642	0.673294	0.443941	0.229353
0	Alameda	0.002774	0.121916	0.258579	0.225052	0.040299	0.007900	0.002775	0.340706	1510271	0.659370	0.453485	0.205885

The fitted values have trouble with outliers like San Francisco and are more accruate for counties with medium populations and access rates. Still, the R^2 increases to **0.116**. It improves substantially from simple linear regression by adding an ethnicity ('White') as another parameter. Adding more parameters, continues to increase the R^2 value. For example, adding either 'Asian' or 'Latino' in addition to 'White', increases the R^2 value to approximately 0.7. However, this is because **the model is overfitting the data** and the sample size is small. In order to increase the sample size, a regression model was fit to the town-level data. Since there are far more towns than counties, the sample size is much larger. The groupings for each town are displayed below:

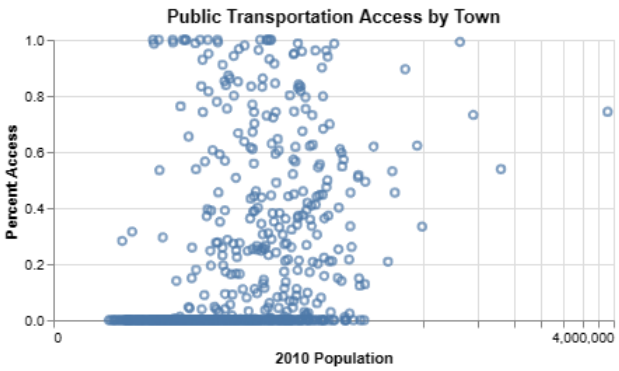
	Town	2010 Population	Percent Access
0	Acalanes Ridge CDP	1137	0.295509
1	Acton CDP	7596	0.000000
2	Adelanto city	31765	0.000000

This table shows that the population is much smaller for each town, compared to the counties analyzed. In some towns, the percent access is 0% since there are so few residents and the town may be in a rural area. Like for the county data, the data table was pivoted to include ethnicity rates for each town. A multiple linear regression model was then fit to the data. The goal is that it should fit the data better with more sample points. First, only 'White' and 'pop2010' were used as the explanatory variables.

The following table shows each town, the ethnicity rates, and the fitted values for 'p_trans_acc' (access rates). It also includes the residuals used for the R^2 analysis.

	geoname	AIAN	AfricanAm	Asian	Latino	Multiple	NHOPI	Other	White	pop2010	p_trans_acc	fitted_mlr	resid_mlr
79	Burbank CDP	0.003654	0.023955	0.076127	0.509338	0.022940	0.001624	0.002233	0.360130	4926	1.0	0.224305	0.775695
17	Alto CDP	0.001406	0.008439	0.040788	0.071730	0.046414	0.001406	0.007032	0.822785	711	1.0	0.060105	0.939895
517	Rollingwood CDP	0.002021	0.060290	0.175480	0.618390	0.017851	0.006736	0.005052	0.114180	2969	1.0	0.310499	0.689501

The R^2 value increased slightly to 0.13 in the town-data from 0.116 in the county-data. Adding all the ethnicity rates, increased the R^2 value to 0.25. Even with all the predictors added, the R^2 value is still insignificant. Beisdes the data being very scattered, one reason is that **many of the towns had a value of '0' or '1' for 'p_trans_acc'**. A regression model struggles to fit data with a lot of 0's and 1's. The following scatterplot shows how scattered the data is:

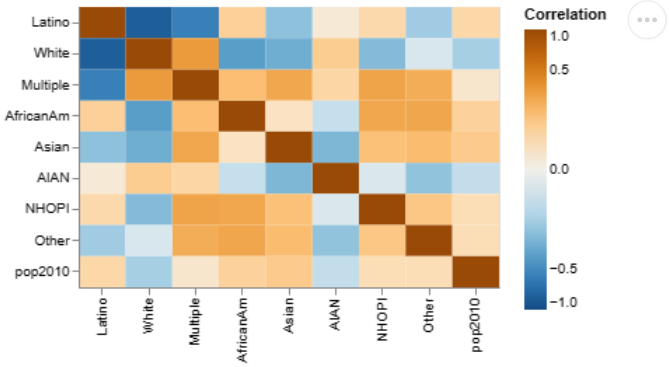


PCA Analysis for Town Data

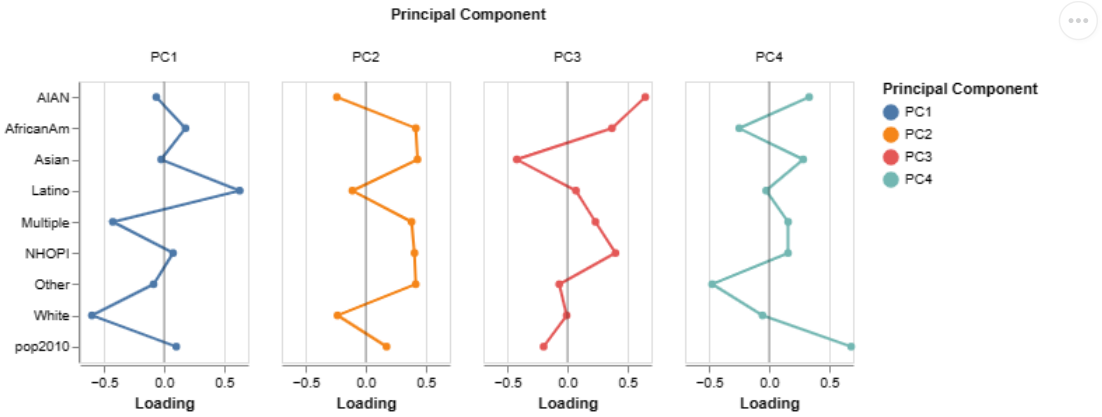
The next question was to determine whether PCA analysis could help filter the predictors.

	Proportion of variance explained	Component	Cumulative variance explained
0	0.250581	1	0.250581
1	0.217423	2	0.468003
2	0.130466	3	0.598469

This table shows that three components explain roughly 60% of the variation in the data. A heat map was then used to show the correlation between variables.



This shows that variables like 'White' and 'Latino' are highly correlated (positive). The loadings were then shown for each principal component.



Ultimately, this still shows that additional data is required at both the town and county level. For example, in PC1: the variables with the largest loadings are 'White' (negative), and 'Latino' (positive). From the correlation matrix, 'White' and 'Latino' have a strong negative correlation. One approach is that PC1 measures neighborhood homogeneity, with towns having a higher value for PC1 if most of the residents are Latino and lower values if most are White. However, **we cannot make a reasonable inference from this PCA data.**

Discussion

The findings show that access to public transportation differs between each region. As well, ethnicity rates and access to transportation differ significantly between counties. However, **ethnicity rates and population are not significant predictors of public transportation rates.** For example, both Marin and Sonoma county have a high 'White' population. However, Marin has a much higher access to transit rate, compared to Sonoma. This shows that it is important to look at how urban, or rural the location is. Additional variables are needed too. These could include measurements like unemployment rates, or a resident's type of work.

Further analysis should focus on one specific region, particularly the Bay Area. In the Bay Area, there are a mix of urban and rural counties, as well as high-income and low-income counties. Using the additional predictors and focusing on one region, could potentially help the model fit the data better. Each region has different demographics and transportation infrastructure, so it is likely difficult to build a general model to the entire state of California. In addition to simple and multiple linear regression, other models could be used. For example, a regression tree could be used since the data is non-linear and it would be more interpretable than other methods.