Analyzing Effect of Demographics on Neighborhood Development

By Dvija Shah

Executive Summary:

This study uses the Building Permits dataset from the City of Boston and the Demographics dataset from the Massachusetts Census Indicators(MCI) dataset to study how demographics affect neighborhood development(Boston Redevelopment Authority, 2016; The Boston Foundation, 2015). The Building Permits dataset contains information about permits issued for building projects in Boston, while the Demographics dataset contains information about the characteristics of the population in different neighborhoods, such as age, race, education, and income level.

The main goal of the analysis is to identify the demographic factors that are important for the development of a neighborhood and help city planners make informed decisions about urban planning and investment(Galster, Hanson, Ratcliffe, & Wolman, 2001). To do this, the latent construct of neighborhood development was created using data from the Building Permits dataset on variables as declared valuation, square footage, number of permits, occupancy type, and project type obtained from the Building Permits dataset. Then, to examine the connection between demographics and neighborhood development, I combined this construct with the MCI's demographic variables and performed various analysis on it.

According to the analysis, there were some variations in neighborhood development depending on the residents' commute distance and educational attainment. Particularly, neighborhoods with less commute, and higher levels of educational attainment tended to have comparatively higher levels of development. I also observed that certain types of projects, such as commercial, residential, and mixed-use developments, were more common in neighborhoods with higher levels of development.

To further investigate the relationship between demographics and neighborhood development, I conducted bi-variate and multi-variate regression analysis(Lees, Slater, & Wyly, 2013). The analysis confirmed that income level, age, race, commute, and education level were minor predictors of neighborhood development, even after controlling for other variables(Boston Redevelopment Authority, 2016; The Boston Foundation, 2015). The overall variance of neighborhood development explained by demographics was around 1.8%, which means that there are multiple factors affecting the development of a neighborhood and demographics plays just a very small part in it (Galster et al., 2001).

The results generally imply that Boston's neighborhood development is not significantly influenced by demographics(Lees et al., 2013). Specifically, areas with lower commute time, and more highly educated population tend to have higher levels of development. (Boston Redevelopment Authority, 2016; The Boston Foundation, 2015). Urban planners and policymakers may find this information helpful as they try to decide strategically where to invest and how to take action for various neighborhoods. But we need to keep in mind that the latent construct of neighborhood development score has been created only a few variables from the Building Permits dataset and it may not accurately represent the factors contributing to

development of a neighborhood. Understanding the elements that influence neighborhood growth will help us work toward developing communities in Boston that are more sustainable and equitable way.

Neighborhood Development Issue:

Neighborhood development refers to the changes and improvements that occur within a specific geographic area over time. It can include multiple building permits factors like the construction of new buildings, changes in land use, improvements to infrastructure, and changes in the demographics of the area. Neighborhood development can impact a range of stakeholders, including residents, business owners, and local government officials who oversee urban planning.

Ideally, neighborhood development can have significant impacts on various demographics, including socioeconomic status, age, race and ethnicity, health, education, and safety and crime rates. For example, policies that prioritize investment in affluent areas may exclude low-income communities of color, contributing to greater disparities in opportunity and outcomes. Neighborhood development can also affect public health by shaping the availability and quality of healthy food options, safe parks, and exercise facilities. Educational opportunities and outcomes can be impacted by the availability and quality of schools, libraries, and other educational resources in the neighborhood. Safety and crime rates can also be influenced by neighborhood development, including access to police and emergency services, and the prevalence of violent and property crimes. Therefore, it is essential for policymakers to consider the potential impacts on these demographics when developing neighborhood development policies and plans.

The long-term effects of neighborhood development on low-income and high-income households can differ significantly. In low-income households, neighborhood development can lead to displacement, which can result in social isolation, loss of community, and decreased access to amenities. For high-income households, neighborhood development can result in increased property values and a higher quality of life. However, neighborhood development can also contribute to income inequality if it leads to gentrification and displacement.

Neighborhood development is particularly pertinent because it can impact the livability of communities and influence patterns of economic growth and development. Analyzing neighborhood development over time can provide insights into which areas require investment or intervention from policymakers, urban planners, and other stakeholders.

The calculation of neighborhood development for this study typically involves analyzing data on permits issued, construction activity, cost, size, and changes in land use. By examining the total number of permits issued, the type of permit, the type of project, the occupancy type, the

average declared valuation, and the average square feet, it is possible to track changes in the built environment of a neighborhood.

This study explores these key questions:

- What is the neighborhood development score across Boston?
- How does neighborhood development impact the racial or ethnic composition of a neighborhood?
- Are certain age groups or educational attainment more likely to be impacted by neighborhood development?
- What other demographics factors influence neighborhood development?

Data and Methods:

I. Building Permits Dataset

The Boston Building Permits dataset contains information on n 550,065 building permit applications submitted between September 26, 2006 and July 9, 2022. Building permits are received by the Inspectional Services Department (ISD) of the City of Boston, who organizes and releases the data through the city's open data portal. There are 36 variables in the dataset. These variables can be grouped into three categories: application characteristics, geographical information, and types of work. The application characteristics include variables like Permit Number, Work type, Applicant, Declared valuation, Status of Permit, Square footage, Occupancy, Issued and expiration dates, Address etc. The geographical information is additional identifying information which makes data more compatible with other data sources. It includes variables like Property id, GIS id, BRA_PD, X and Y coordinates etc. The types of work contains newly BARI(Boston Area Research Initiative) constructed dummy variables like newcon, demo, addition, reno, govt and Permit Duration.

The broad array of information like application characteristics, geographical information and types of work make latent constructs potentially powerful tools in looking at Boston's growth from a variety of perspectives. This study establishes a latent variable for neighborhood development score as the main variable of interest. It is a scoring mechanism to find the development of the neighborhood using key characteristics from the building permits dataset such as Declared Valuation, no of permits, Square feet, Occupancy Type and the type of work. For calculating this we first worked on creating a couple of new variables from the existing variable.

Building Permits Data Description:

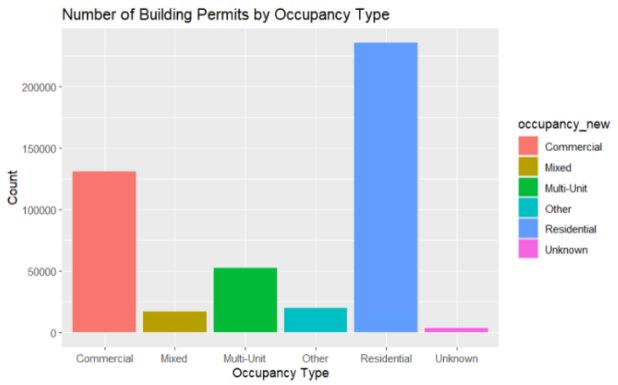
I will be giving a brief data description of the variables of the Building Permits dataset that were used for this study:

Declared Valuation: It is the estimated value of the proposed construction project as declared by the applicant on the building permit application.
Square Footage: It is the total area in square feet of the proposed construction project.
Occupancy: It is the intended use or purpose of the building or structure as declared by the applicant on the building permit application.
CT_ID_10: It is a unique identifier for the census tract where the proposed construction project is located.
NSA_Name: It is the name of the neighborhood statistical area (NSA) where the proposed construction project is located.
Newcon: It is a binary indicator variable that takes the value of 1 if the building permit is for a new construction project and 0 otherwise.
Demo: It is a binary indicator variable that takes the value of 1 if the building permit is for a demolition project and 0 otherwise.
Addition: It is a binary indicator variable that takes the value of 1 if the building permit is for an addition or alteration project and 0 otherwise.
Reno: It is a binary indicator variable that takes the value of 1 if the building permit is for a renovation or rehabilitation project and 0 otherwise.

II. Creating New Features from Building Permits Dataset

I created a couple of new variables by using the existing variables from the Building Permits dataset. It will help to make the data more interpretable and useful. By transforming or aggregating existing data, new variables can highlight relationships or patterns in the data that were not apparent before. It gives us a deeper understanding of the data and can help to make informed decisions based on it.

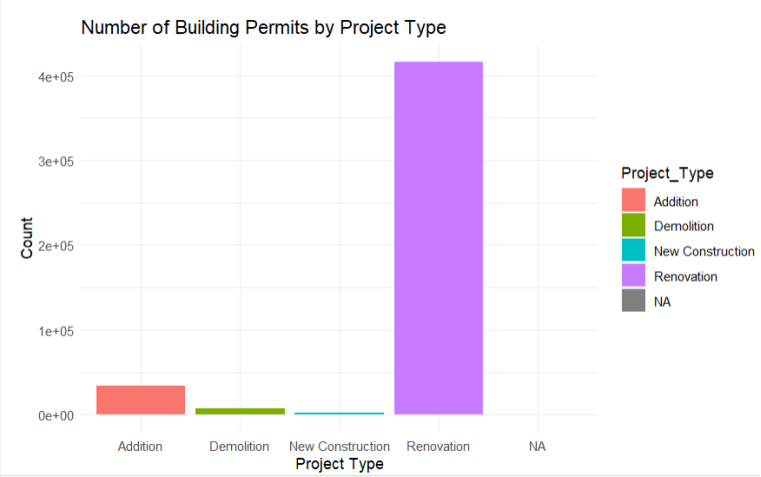
1. Occupancy New from recategorizing Occupancy
   The Occupancy variable has 21 categories like '1-4FAM', '7More', '1-7FAM', 'VacLd' , 'COMM', 'MIXED' etc. To simplify the results, I decided to recategorize the Occupancy variables into Occupancy New with 6 categories named 'Commercial', 'Mixed', 'Multi-Unit', 'Other', 'Residential' and 'Unknown'. Doing this will make it easier to compare and interpret the occupancy data.



The above bar chart displays the number of building permits by new Occupancy Type.

2. Project Type from Type of work

   It is created by created by combining the "newcon", "addition", "demo", and "reno" columns. The initial columns had a flag indicating if it was a new construction, addition, demolition, or renovation. Creating this new variable would give a more general overview of the type of construction project being undertaken, as each of the four original variables provides more specific information about the nature of the project.



The above bar chart displays the number of building permits by Project Type.

III. Latent Variable Construction

To start building my latent construct of neighborhood development, I started by extracting the necessary data from the Building Permits dataset. For that, I grouped the data by 'CT_ID_10' 'occupancy_new' and 'Project_Type' and then calculate the number of building permits issued, the average declared valuation and the average square footage for each group. We can call this subset of our data as ‘neighborhoods’. This

analysis can help to identify trends in neighborhood development, such as the types of projects and occupancy that are most common in different neighborhoods.

The next step is to calculate the maximum value of 'no_of_permits_issued', 'avg_declared_valuation', and 'avg_sq_ft' and assign to the variables 'max_no_of_permits_issued', 'max_avg_declared_valuation', and 'max_avg_sq_ft', respectively. These maximum values can be used as a reference point to compare the values of each variable across different neighborhoods.

After that, I calculated the total number of building permits issued for each occupancy category in the 'neighborhoods' dataset. Using that I will calculate the percentage of permits issued for four different occupancy categories: 'commercial', 'residential', 'multi-unit' and 'mixed'. To do so, each occupancy category count is divided by the total number of permits issued. In the example given below I have calculated the percentage of 'commercial' occupancy. Similarly, I have calculated for the other occupancies percentage.

$$\boldsymbol{commercial} = \frac{\boldsymbol{number\ of\ permits\ issued\ for\ commercial\ occupancy}}{\boldsymbol{total\ number\ of\ permits\ issued}}$$

In the same way, I calculated the total number of building permits issued for each Project Type category in the 'neighborhoods' dataset. . Using that I will calculate the percentage of permits issued for three different Project types: 'renovation', 'addition', and 'new construction'. Each project type category is divided by the total number of permits issued. In the example below I have calculated the percentage of 'renovation' project. Similarly, I have calculated the other project type's percentage.

$$\boldsymbol{renovation} = \frac{\boldsymbol{number\ of\ permits\ issued\ for\ renovation\ project\ type}}{\boldsymbol{total\ number\ of\ permits\ issued}}$$

The final step is to generate the neighborhood development score. This variable is intended to represent a score for each neighborhood that reflects its level of development. The value of this column is calculated by adding up weighted scores for five different variable types, as shown in the equation below.

$$\boldsymbol{neighborhood\_dev\_score} = (\boldsymbol{no\_of\_permits\_issued/max\_no\_of\_permits\_issued}) * \mathbf{0.2} +$$

$$(avg\_declared\_valuation/max\_avg\_declare\_valuation) * 0.2 +$$
$$(avg\_sq\_ft/max\_avg\_sq\_ft) * 0.2 +$$
$$(commercial + residential + multi\_unit + mixed) * 0.2 +$$
$$(renovation + addition + new\_construction) * 0.2$$

The first three terms in the equation calculate the normalized values of the number of permits issued, the average declared valuation, and the average square footage of properties in the neighborhood, respectively. These terms are each weighted by 0.2. Then I sum up the percentage of commercial, residential, mixed-use, and multi-unit occupancy properties and weigh it by 0.2 and finally I sum up the percentage of renovation, addition and new construction property type and weigh those terms by 0.2. By multiplying each variable by 0.2 I focused on giving equal importance for each variable type.

IV. External Data

I have used two external datasets for my study.

Massachusetts Census Indicators Dataset

The Massachusetts Census Indicators dataset is a collection of demographic and socioeconomic indicators for the state of Massachusetts, USA. The dataset includes information on population, housing, income, education, and labor force characteristics, among other topics. The indicators are derived from the U.S. Census Bureau's American Community Survey (ACS), which is an ongoing survey that provides detailed information on a variety of social and economic characteristics of the U.S. population. The Massachusetts Census Indicators dataset includes data for the state, as well as for individual counties and cities/towns within the state. Some of the key indicators included in the dataset are total population, population by age and race/ethnicity, household income, poverty rates, educational attainment levels, labor force participation rates, unemployment rates, housing tenure (owner-occupied vs. renter-occupied) and median home values. This dataset can help us understand the demographic and socioeconomic characteristics of the population in Boston.

Massachusetts Census Indicators Data Description:

I will be giving a brief data description of the variables of the Massachusetts Census Indicators dataset that were used for this study.

CT_ID_10: The Census Tract ID.

AgeU18: The percentage of the population under 18 years old.
Age1834: The percentage of the population between 18 and 34 years old.
Age3564: The percentage of the population between 35 and 64 years old.
AgeO65: The percentage of the population over 65 years old.
ForBorn: The percentage of the population that was born outside of the United States.
White: The percentage of the population that identifies as White.
Black: The percentage of the population that identifies as Black or African American.
Asian: The percentage of the population that identifies as Asian.
Hispanic: The percentage of the population that identifies as Hispanic or Latino.
MedHouseIncome: The median household income in the Census Tract.
GINI: The Gini coefficient, which measures income inequality.
FamPovPer: The percentage of families that are living in poverty.
TotalHouseH: The total number of households in the Census Tract.
FamHousePer: The percentage of households that are family households.
FemHeadPer: The percentage of households that are headed by females.
LessThanHS: The percentage of the population that has less than a high school diploma.
HSGrad: The percentage of the population that has a high school diploma or equivalent.
SomeColl: The percentage of the population that has some college education but no degree.
Bach: The percentage of the population that has a bachelor's degree.
Master: The percentage of the population that has a master's degree.
Prof: The percentage of the population that has a professional degree.
Doc: The percentage of the population that has a doctoral degree.
CommuteLess10: The percentage of workers whose commute to work is less than 10 minutes.
Commute1030: The percentage of workers whose commute to work is between 10 and 30 minutes.
Commute3060: The percentage of workers whose commute to work is between 30 and 60 minutes.
Commute6090: The percentage of workers whose commute to work is between 60 and 90 minutes.
CommuteOver90: The percentage of workers whose commute to work is over 90 minutes.
ByAuto: The percentage of workers who commute to work by car.
ByPubTrans: The percentage of workers who commute to work by public transportation.
ByBike: The percentage of workers who commute to work by bike.
ByWalk: The percentage of workers who commute to work on foot.
TotalHouseUnits: The total number of housing units in the Census Tract.
VacantUnitPer: The percentage of housing units that are vacant.
MedGrossRent: The median gross rent for occupied housing units in the Census Tract.
MedHomeVal: The median value of owner-occupied housing units in the Census Tract.

Redlining in Boston Dataset

The Redlining in Boston dataset by BARI(Boston Area Research Initiative) is a collection of historical maps, data, and documents that document the process of "redlining" in Boston during the mid-20th century. Redlining was a practice that was used by lenders, insurers, and government agencies to restrict access to credit, insurance, and other financial services based on the racial and ethnic composition, income levels, property values, and age and condition of buildings of neighborhoods.

Redlining in Boston Data Description:

I will be using one variable from the redlining in Boston Dataset for this study.

Redline: the redline flag variable which indicates '1' if the area was previously redlined and '0' if the area was not previously redlined.

Statistical Analysis:

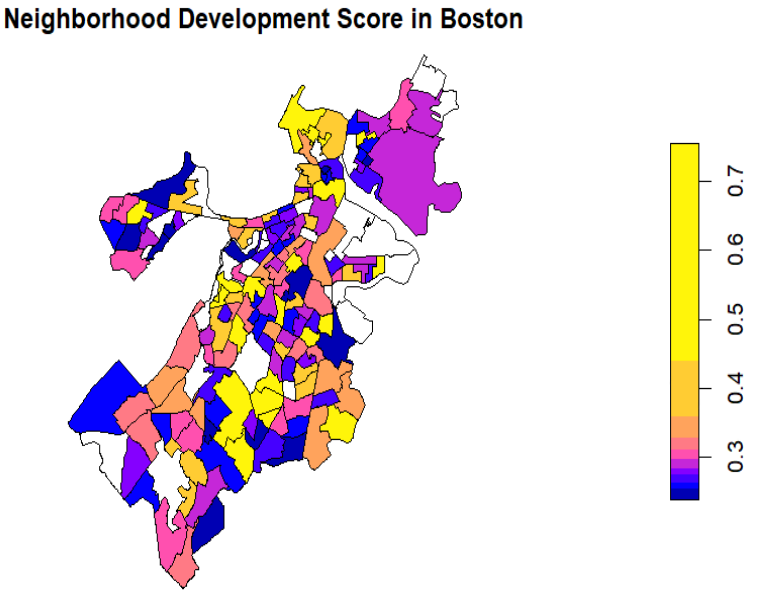
After calculating the neighborhood development score latent variable at the Census tract level, the results of its distribution across the city's neighborhoods are visible.

Distribution at Census Tract Level

The summary statistics table of Neighborhood Development Score at the Census Tract level. The census tract id with the highest neighborhood development score is the Downtown, Central, West end area with a neighborhood development score of 0.7539. The next highest. The next two neighborhoods with high development scores are Beacon Hill and South End-Shawmut which are one of the most affluent neighborhoods in Boston. They have a neighborhood development score of 0.7522 and 0.7506 respectively. All these areas are portrayed as the hub of commerce, culture, and entertainment with a dynamic and diverse community.

The 3rd Quartile of neighborhood development is around 0.3427 and areas like Back Bay, Fenway-Kenmore, and North End fall into this quartile. Back bay is considered as one of the most expensive residential areas in Boston. The area is known for its Victorian brownstones, the Boston Public Library, shopping, office high rises and upscale hotels. Fenway-Kenmore is most recognized as the home of Fenway Park and the Red Sox. It's also home to cultural institutions like the Museum of Fine Arts and Symphony Hall. You'll also find the nation's first public school, Boston Latin School. There are also several schools for higher education in the area. North End has long been known for its historic charm and close-knit community. Over the last 20 years, the

area has undergone a transformation, with the construction of new residential buildings and the expansion of the commercial sector. It is known for its Italian American population and Italian restaurants.



Summary Statistics Table	
Min	**0.2379**
1st Quartile	**0.2686**
Median	**0.2967**
Mean	**0.3228**
3rd Quartile	**0.3427**
Max	**0.7539**

The above choropleth displays neighborhood development scores in Boston.

The Mean and Median of neighborhood development score 0.2967-03288 consists of areas like Beacon Hill, Brighton, Jamaica Plain. Beacon Hill is a historic neighborhood in Boston known for its federal style rowhouses, brick sidewalks, gas lamps, and scenic views. Brighton is a neighborhood in the northwest corner of Boston known for its diverse population, student housing, and commercial activity. It is home to several colleges and universities, including Boston College and Harvard Business School. Jamaica Plain is a diverse neighborhood in the southwestern part of Boston known for its green spaces, Victorian homes, and vibrant arts scene. It is home to the famous Arnold Arboretum. These areas are well-developed with a mix of residential and commercial development.

The 1st Quartile has neighborhoods with a development score of 0.2686. It consists of areas like Forest Hills, Telegraph Hills, and Maverick. Forest Hills is known for its large, historic cemetery, which includes the graves of many famous figures. The neighborhood also has a transportation hub with a bus station and a subway station. Maverick is a diverse neighborhood with a mix of residential and commercial areas. The neighborhood has several parks, including Piers Park, which offers views of the Boston skyline. Telegraph Hill is a historic neighborhood that was once home to many Irish immigrants. The neighborhood is known for its steep hills and narrow streets, as well as its views of Boston Harbor.

The areas with the lowest neighborhood development scores are Dorchester, Roslindale, and Jeffries point with a neighborhood development score of 0.2379 for each of them. Dorchester is

one of Boston's largest neighborhoods and is known for its diversity and history. The neighborhood has a mix of residential and commercial areas, as well as several parks and community centers. Roslindale is a primarily residential neighborhood with a mix of single-family homes and multi-unit buildings. Jeffries Point is a waterfront neighborhood that offers stunning views of the Boston skyline. The neighborhood is primarily residential, with a mix of historic homes and new developments.

Neighborhood Development by Block

The neighborhood development score table shows the average neighborhood development score for each neighborhood. In this we can see that Downtown, Central, West End, Beacon Hill, and South End- Shawmut are the neighborhoods with the highest neighborhood development score. Whereas Roslindale, Dorchester and Jeffries Point have the lowest neighborhood development scores.

Neighborhood Development Score Table	
Neighborhood	**Dev Score**
Downtown/Central/West End	**0.7539**
Beacon Hill	**0.7522**
South End- Shawmut	**0.7506**
Back Bay	**0.3564**
Fenway	**0.3427**
North End	**0.3398**
Beacon Hill	**0.3288**
Brighton	**0.3052**
Jamaica Plain	**0.2967**
Forest Hills	**0.2713**
Telegraph Hills	**0.2657**
Maverick	**0.2573**
Jeffries Point	**0.2385**
Roslindale	**0.2382**
Dorchester	**0.2379**

Correlation and Significance

To investigate these distributions further, Pearson's r, correlation, was used to measure the strength of the relationships (effect size) between Neighborhood Development Score and demographic variables from ACS data, as well as Redlining in Boston dataset.

Some of the strongest statistically significant positive correlations, where increases in Neighborhood development score are associated with increases in the independent variables, are as follows and its correlation score is shown in the table:

- GINI
- Asian
- MedianHomeVal
- TotalHouseUnits
- MedGrossRent

Some of the strongest negative correlations with Neighborhood development scores , where increases in neighborhood development score are associated with decreases in the other variables are as follows and its correlation score is shown in the table:

- AgeU18
- ForBorn
- Hispanic
- HSGrad
- Commute3060
- ByAuto

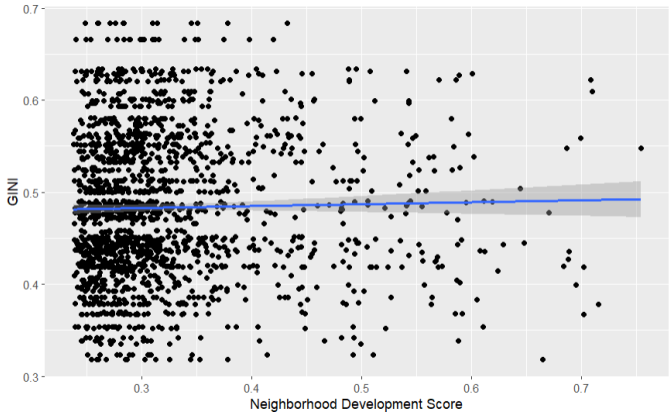
Correlation Score Table	
Variable	**Correlation Score with Neighborhood development**
GINI	**0.0747**
Asian	**0.0502**
MedianHomeVal	**0.0833**
TotalHouseUnits	**0.0903**
MedGrossRent	**0.0406**
AgeU18	**-0.0162**
ForBorn	**-0.0098**
Hispanic	**-0.0617**
HSGrad	**-0.0263**
Commute3060	**-0.0140**
ByAuto	**-0.0015**

Bi-variate analysis

Since a few variables have a strong correlation with neighborhood development score, that formed a basis for bivariate regression analysis. I ran a bi variate regression analysis on a few of the strong correlation variables with neighborhood development score.

The first variable for which I conducted bi-variate analysis with neighborhood development score is GINI. For this model, the intercept of the model is 0.31159, which represents the expected value of the neighborhood development score when the GINI index is zero. The coefficient for GINI is 0.0258, which means that for each unit increase in the GINI index, the neighborhood development score is expected to increase by 0.02276 units, all else being equal.

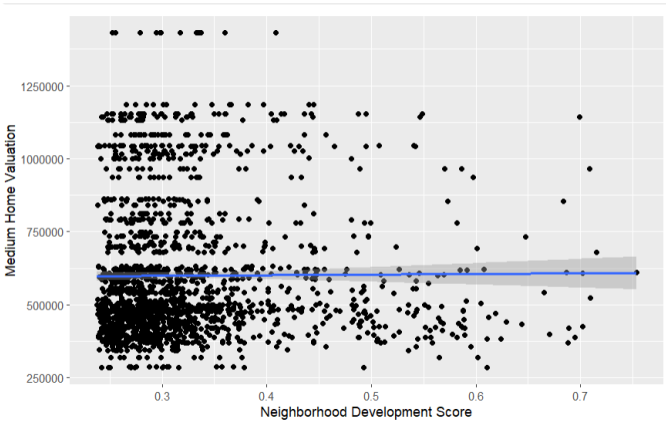
$$\boldsymbol{Neighborhood\ Development\ Score\ = 0.31159 + 0.02276 * GINI}$$



Scatter plot of neighborhood development score with GINI

The second variable for which I conducted bi-variate analysis with neighborhood development score is MedHomeVal. For this model, the intercept coefficient is 0.3209, which means that if "MedHomeVal" were zero (i.e., no median home value in the neighborhood), the predicted value of "neighborhood_dev_score" would be 0.3209. The coefficient of "MedHomeVal" is 2.802e-09. This coefficient is very small, which means that for each additional unit of "MedHomeVal" (i.e., for each additional dollar in median home value), the predicted value of "neighborhood_dev_score" would increase by a very small amount.

$$\boldsymbol{Neighborhood\ Development\ Score = 0.3209 + (2.802e - 09) * MedHomeVal}$$

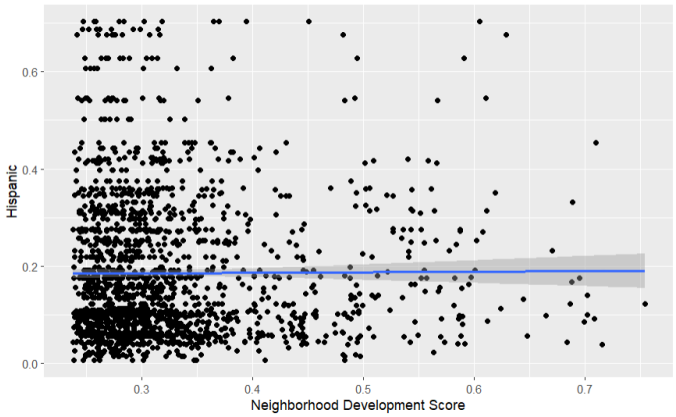


Scatter plot of neighborhood development score with Median Home Valuation

The next variable for which I conducted bi-variate analysis with neighborhood development score is Hispanic. For this model, the intercept coefficient is 0.321882, which means that if "Hispanic" were zero (i.e., no Hispanic population in the neighborhood), the predicted value of

"neighborhood_dev_score" would be 0.321882. The coefficient of "Hispanic" is -0.003824, which means that for each additional unit of "Hispanic" (i.e., for each additional percentage point of Hispanic population in the neighborhood), the predicted value of "neighborhood_dev_score" would decrease by 0.003824.

$$\boldsymbol{Neighborhood\ Development\ Score} = \mathbf{0.321882} + (\mathbf{-0.003824}) * \boldsymbol{Hispanic}$$



Scatter plot of neighborhood development score with Hispanic

The last variable for which I conducted bi-variate regression analysis with neighborhood development score is HSGrad. For this model, the intercept coefficient is 0.321882, which means that if "HSGrad" were zero (i.e., no high school graduates in the neighborhood), the predicted value of "neighborhood_dev_score" would be 0.321882. The coefficient of "HSGrad" is -0.01969, which means that for each additional unit of "HSGrad" the predicted value of "neighborhood_dev_score" would decrease by 0.01969.

$$\boldsymbol{Neighborhood\ Development\ Score} = \mathbf{0.321882} + (\mathbf{-0.01969}) * \boldsymbol{HSGrad}$$

Scatter plot of neighborhood development score with High School Graduate

Multi-variate analysis

After conducting bi-variate analysis on some of the strongest correlations, I will now conduct a multivariate analysis. For my multivariate analysis, I will be creating a model using the highly correlated variables and will analyze the results.

$$\begin{aligned} &\boldsymbol{neighborhood\ development\ score} \\ &\quad = \mathbf{0.334} - \mathbf{0.058} * \boldsymbol{GINI} + (\mathbf{3.453e-06}) * \boldsymbol{TotalHouseUnits} + \mathbf{0.029} \\ &\quad * \boldsymbol{VacantUnitPer} + (\mathbf{2.036e-08}) * \boldsymbol{MedHomeVal} + \mathbf{0.046} * \boldsymbol{ByWalk} + \mathbf{0.056} \\ &\quad * \boldsymbol{FemHeadPer} - \mathbf{0.029} * \boldsymbol{FamHousePer} + \mathbf{0.023} * \boldsymbol{Hispanic} - (\mathbf{6.235e-05}) \\ &\quad * \boldsymbol{PopDen} + \mathbf{0.0048} * \boldsymbol{AgeU18} + \mathbf{0.025} * \boldsymbol{HSGrad} \end{aligned}$$

This multivariate analysis models the relationship between the dependent variable "neighborhood_dev_score" and several independent variables: 'GINI', 'TotalHouseUnits', 'VacantUnitPer', 'MedHomeVal', 'ByWalk', 'FemHeadPer', 'FamHousePer', 'Hispanic', 'PopDen', 'AgeU18', and 'HSGrad'. For this model, the intercept value indicates that, on average, the neighborhood development score is 0.334, holding all other variables constant. The ByWalk variable has a positive effect on the outcome, suggesting that neighborhoods with more walkability tend to have higher development scores. PopDen has a negative effect on the outcome, indicating that neighborhoods with higher population density tend to have lower development scores. The remaining variables, including GINI, TotalHouseUnits, VacantUnitPer, MedHomeVal, FemHeadPer, FamHousePer, Hispanic, AgeU18, and HSGrad, do not have statistically significant effects on the outcome, as indicated by their p-values greater than the conventional threshold of 0.05. The model's R-squared value is 0.01052. Which means that the model explains 1% of the variance in the neighborhood development score can be explained by the predictor variables included in the model.

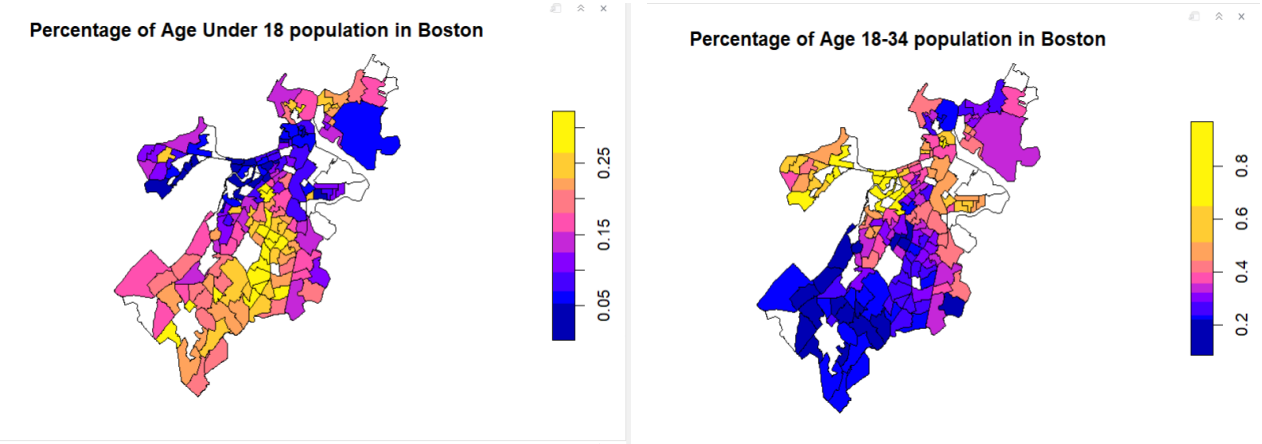
To take this one step further, I decided to conduct one more multivariate analysis with more variables to observe the effect of those independent variables on neighborhood development score.

$$\begin{aligned} &\boldsymbol{neighborhood\ development\ score} \\ &\quad = \mathbf{0.2644} - \mathbf{0.01609} * \boldsymbol{GINI} + \mathbf{0.03139} * \boldsymbol{AgeU18} - \mathbf{0.001905} * \boldsymbol{Age1834} \\ &\quad + \mathbf{0.001237} * \boldsymbol{Age3564} + \mathbf{0.1110} * \boldsymbol{White} + \mathbf{0.1059} * \boldsymbol{Black} + \mathbf{0.09427} * \boldsymbol{Asian} \\ &\quad + \mathbf{0.1023} * \boldsymbol{Hispanic} - \mathbf{0.03519} * \boldsymbol{FamHousePer} + \mathbf{0.07804} * \boldsymbol{FemHeadPer} \\ &\quad - \mathbf{0.05663} * \boldsymbol{LessThanHS} - \mathbf{0.08854} * \boldsymbol{HSGrad} - \mathbf{0.07646} * \boldsymbol{SomeColl} - \mathbf{0.1261} \\ &\quad * \boldsymbol{Bach} - \mathbf{0.05531} * \boldsymbol{Master} - \mathbf{0.2723} * \boldsymbol{Prof} + \mathbf{0.2864} * \boldsymbol{CommuteLess10} \\ &\quad + \mathbf{0.05916} * \boldsymbol{Commute1030} + \mathbf{0.08345} * \boldsymbol{Commute3060} - \mathbf{0.02481} \\ &\quad * \boldsymbol{Commute6090} - \mathbf{0.05046} * \boldsymbol{ByAuto} - \mathbf{0.06410} * \boldsymbol{ByPubTrans} + \mathbf{0.0246} \\ &\quad * \boldsymbol{ByBike} - \mathbf{0.09808} * \boldsymbol{ByWalk} + \mathbf{0.000003468} * \boldsymbol{TotalHouseUnits} + \mathbf{0.05444} \\ &\quad * \boldsymbol{VacantUnitPer} + \mathbf{0.00001203} * \boldsymbol{MedGrossRent} + \mathbf{0.00000001197} \\ &\quad * \boldsymbol{MedHomeVal} \end{aligned}$$

This multivariate analysis models the relationship between the dependent variable "neighborhood_dev_score" and several independent variables such as Age, Race, Educational Attainment, Commute, Home Valuation and other factors and GINI. For this model, the intercept value indicates that, on average, the neighborhood development score is 0.2644, holding all other variables constant. CommuteLess10 has a positive coefficient 0.2864 and thus, an increase in the proportion of people with a commute time of less than 10 minutes is associated with an increase in neighborhood development score. Prof has a negative coefficient -0.2723 and thus, an increase in the proportion of workers in professional occupations is associated with a decrease in neighborhood development score. The model has an R-squared value of 0.01837, which indicates that 1.83% of the variance in the dependent variable is explained by the independent variables in the model.

Discussion:

Effect of Age on Neighborhood Development Score

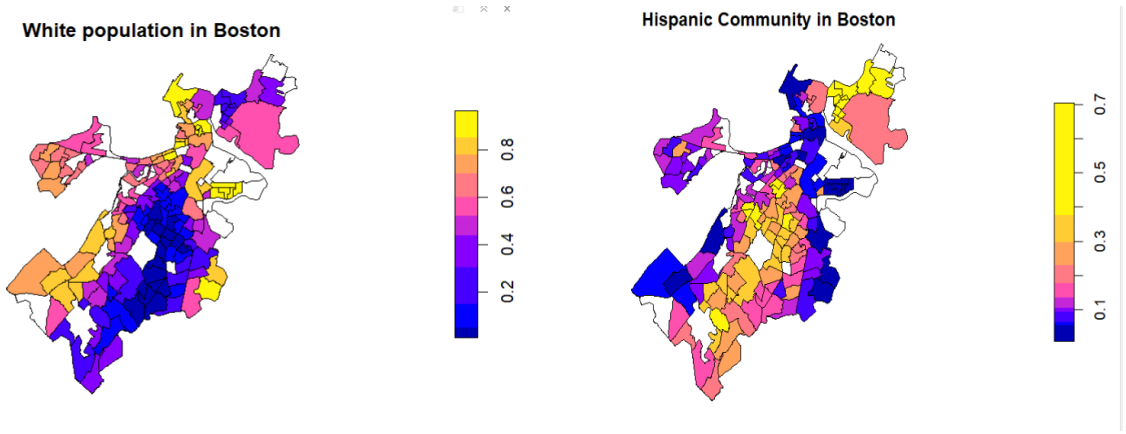


The above choropleths display AgeU18 and Age1834 percentage of population in Boston.

The regression results show that AgeU18, Age1834 and Age3564 have no statistically significant effect on neighborhood development as their p-values are greater than 0.05. This means that there is no evidence to suggest that the percentage of the population under 18, between 18-34, and between 35-64 years old have a significant impact on neighborhood development, holding all other variables constant. AgeU18 has a positive effect on the neighborhood development score, but this effect is not statistically significant (p-value = 0.6909). has a slightly negative effect on the neighborhood development score, but this effect is also not statistically significant. Age3564 has a slightly positive effect on the neighborhood development score. Overall, from the above model we can say that age does not have a significant impact on neighborhood development. But it is important to note that this model is only an approximation of reality and

there may be other variables which have not been considered that could influence the relationship between age and neighborhood development.

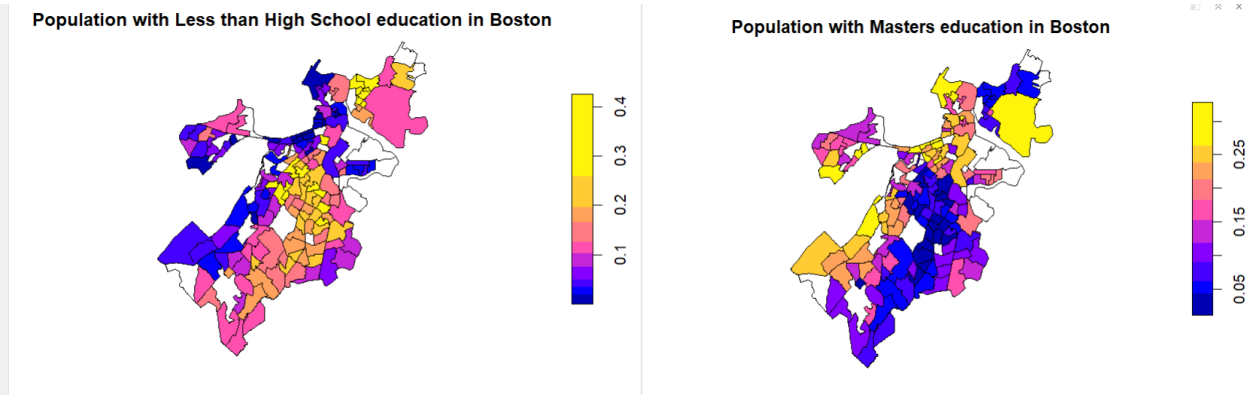
Effect of Race and Ethnicity on Neighborhood Development Score



The above choropleths display White and Hispanic percentage of population in Boston.

The regression results show that White, Black, Asian, and Hispanic have no statistically significant effect on neighborhood development as their p-values are greater than 0.05. This means that based on this regression analysis there is no evidence to suggest that the percentage of the population of different races and ethnicity have a significant impact on neighborhood development, holding all other variables constant. White population percentage is associated with a 0.111 increase in the neighborhood development score, but this relationship is not statistically significant (p-value = 0.3058). Similarly, the Black and Asian population percentages have positive coefficients and Hispanic population percentage has negative coefficients but none of them are statistically significant predictors of neighborhood development. The results of this model do not provide a complete understanding of the complex relationship between race/ethnicity and neighborhood development. Other factors such as historical policies and structural racism can have a significant impact on the development and maintenance of neighborhoods.

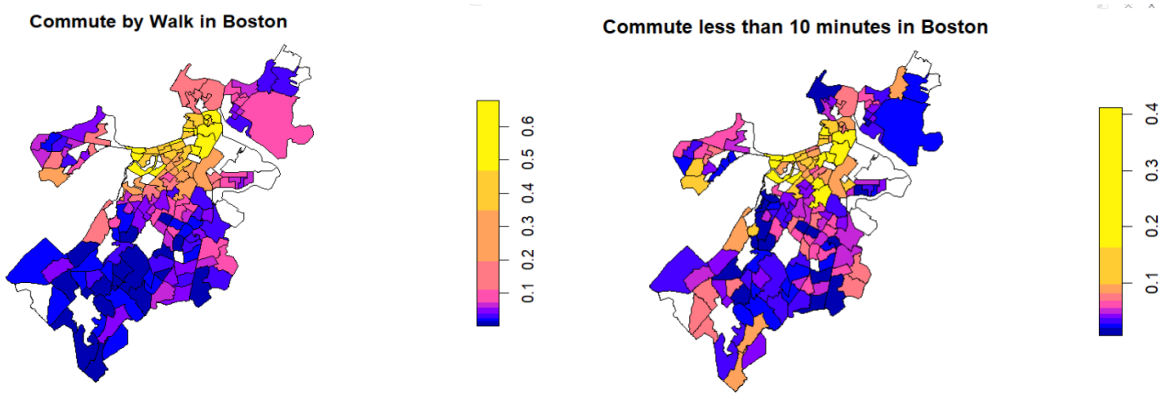
Effect of Educational Attainment on Neighborhood Development Score



The above choropleths display LessThanHS and Masters education percentage of population in Boston.

The variables LessThanHS, HSGrad, SomeColl, Bach, Master, and Prof are indicators of education attainment of the population. The regression model shows that the coefficients for LessThanHS, HSGrad, SomeColl, Bach, Master, and Prof are all negative, indicating that higher levels of education are associated with lower neighborhood development scores. But none of the indicators are statistically significant. The neighborhood development score latent variable is calculated from a few of the variables of the Building Permits dataset and does not take into account all the factors necessary for the development of a neighborhood, hence it might not accurately represent overall neighborhood development.

Effect of Commuting Patterns on Neighborhood Development Scores

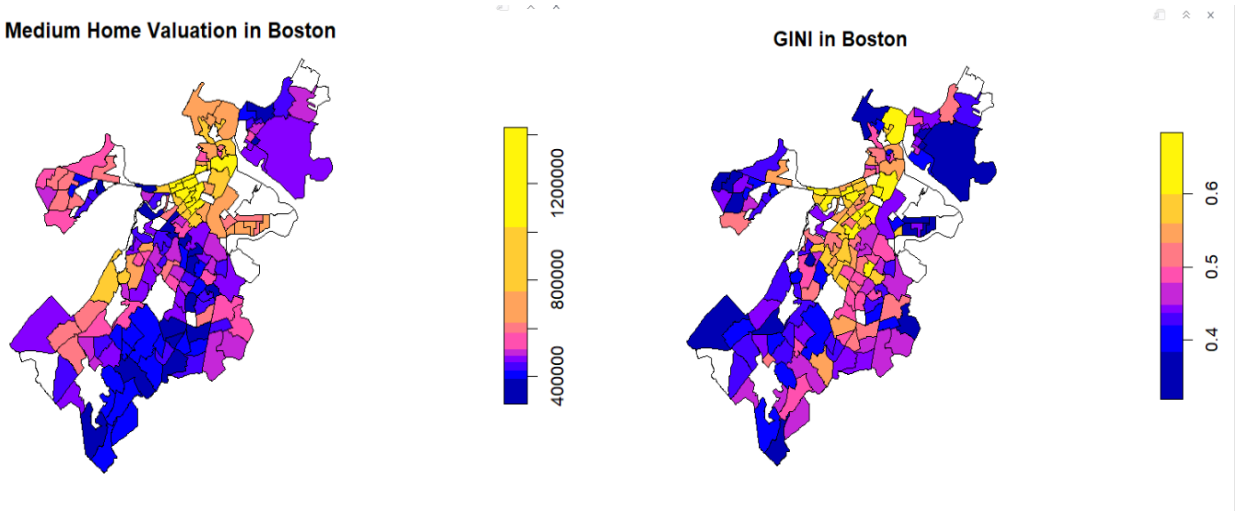


The above choropleths display ByWalk and CommuteLess10 percentage of population in Boston.

Commuting Patterns refer to the time taken to commute, and modes of transportation used by the population for commuting. It considers the variables like CommuteLess10, Commute1030,

Commute3060, Commute6090, ByWalk, ByBike, ByAuto, and ByPubTrans. Based on the regression model CommuteLess10 has a significant positive effect on the neighborhood development score, while other commuting patterns do not have a significant effect. This implies that neighborhoods with a higher proportion of residents who commute less than 10 minutes to their place of work or study tend to have higher development scores. ByWalk has a negative but insignificant effect on neighborhood development scores. Similarly, ByBike, ByAuto, and ByPubTrans do not have a significant effect on neighborhood development score. It is important to note that the analysis is based on a limited set of variables and may not capture all the relevant factors that influence neighborhood development.

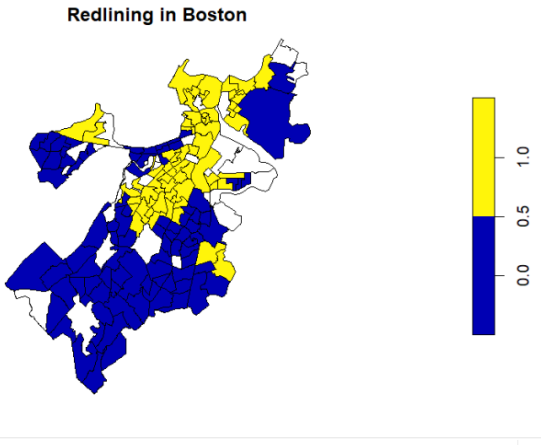
Effect of other factors on Neighborhood Development Scores



The above choropleths display MedianHomeVal and GINI of population in Boston.

Finally, I looked at the remaining demographics variables like GINI, FamHousePer, FemHeadPer, TotalHouseUnits, VacantUnitPer, MedGrossRent and MedHomeVal. GINI and FamHousePer both have a negative coefficient but it is not statistically significant. FenHeadPer, TotalHouseUnits, VacantUnitPer, MedGrossRent and MedHomeVal have positive coefficient but all of them are not statistically significant. Overall, the regression results suggest that these demographic factors may not have a significant effect on neighborhood development scores. Still it is important to note that the lack of statistical significance does not necessarily mean that these factors have no effect on neighborhood development scores. It may be that the sample size is too small or that the data is not fully representative of the population. Additionally, there may be other factors that were not included in the model that could have a significant effect on neighborhood development scores.

Effect of Redlining on Neighborhood Development Scores



The above choropleth display redlining in Boston.

The regression results for Redline show a small positive relationship between the redline variable and the neighborhood development score. But this relationship is not statistically significant, meaning there is no significant relationship between neighbourhood development scores with redlining. However, it is important to consider the broader context in which redlining occurred and the potential long-term effects it may have had on neighborhoods. Again, the neighborhood development score is a latent variable created using a small subset of features required for the development of a neighborhood. So, it cannot fully consider the importance of these features in developing a neighborhood. While the regression results may not show a significant relationship between redlining and neighborhood development scores, it is important to continue exploring the long-term effects of discriminatory practices like redlining on neighborhoods and consider policies and interventions to address these issues.

Conclusion:

To investigate the influence of demographics on neighborhood development in Boston, this study used data from the Building Permits dataset and the Massachusetts Census Indicators (MCI) dataset. The analysis has shown that while there are some correlations between certain demographic factors, such as commute time and educational attainment, and neighborhood development score, demographics as a whole only account for a small portion of the variance in neighborhood development. This data can be used by urban planners and policymakers to make informed decisions about development and investment in various neighborhoods, but it's important to remember that there are numerous factors at play when it comes to neighborhood growth.

One key takeaway from this study is the importance of considering the potential impacts of neighborhood development on different demographics. Policies and plans that prioritize investment in affluent areas may exclude low-income families and communities of color, contributing to greater disparities in opportunity and outcomes. Also, neighborhood development can have significant impacts on public health, education, and safety, and policymakers should take these factors into account when making decisions about investment and development.

There are a number of areas where this research could be developed in the future. For example, this study has only looked at the impact of demographics on neighborhood development in Boston. It would be interesting to compare these findings to other cities and see if similar patterns emerge. Additionally, this study has focused on the impact of demographics on development, but it would also be valuable to investigate the impact of development on demographics. For example, how does the construction of new buildings or changes in land use impact the racial or ethnic composition of a neighborhood?

Furthermore, while my research has used data on building permits to calculate a neighborhood development score, there may be other factors that contribute to neighborhood growth that are not captured in this dataset. For example, proximity to essential services, infrastructure, public spaces, cultural and social amenities may play a role in shaping the development of a neighborhood. It would be interesting to explore these factors in future research.

Overall, this research provides valuable insights into the relationship between demographics and neighborhood development in Boston. By understanding the factors that influence neighborhood growth, policymakers and urban planners can work toward developing more sustainable and equitable communities. However, it is important to continue to investigate and consider a range of factors when making decisions about investment and development to ensure that all members of a community can benefit from neighborhood growth.

References:

Boston Redevelopment Authority. (2016). Boston building permits. Retrieved from https://data.boston.gov/dataset/building-permits

Galster, G., Hanson, R., Ratcliffe, M. R., & Wolman, H. (2001). Wrestling sprawl to the ground: defining and measuring an elusive concept. Housing Policy Debate, 12(4), 681-717. doi: 10.1080/10511482.2001.9521469 https://www.researchgate.net/publication/235358255_Wrestling_Sprawl_to_the_Ground_Defining_and_Measuring_an_Elusive_Concept

Lees, L., Slater, T., & Wyly, E. (2013). Gentrification. Routledge. https://www.routledge.com/Gentrification/Lees-Slater-Wyly/p/book/9780415950374

The Boston Foundation. (2015). The Boston indicators project: Massachusetts census indicators. Retrieved from https://www.bostonindicators.org/indicators-reports-online-tools/census-2010-datasets

Appendix

Building Permits Dataset Data Description:

This is the entire data description of the building permits dataset:

PermitNumber: unique identifier for each permit

WORKTYPE: the type of work being done, e.g. new construction, renovation, demolition, etc.

permittypedescr: description of the permit type

description: a brief description of the work being done

NOTES: any additional notes or comments related to the permit

APPLICANT: the name of the individual or organization applying for the permit

DECLARED_VALUATION: the declared value of the work being done

total_fees: the total fees associated with the permit

ISSUED_DATE: the date the permit was issued

EXPIRATION_DATE: the date the permit expires

STATUS: the status of the permit, e.g. issued, expired, etc.

owner: the name of the property owner

OCCUPANCY: the intended use or occupancy of the building

sq_feet: the square footage of the building

ADDRESS: the street address of the property

CITY: the city where the property is located

STATE: the state where the property is located

ZIP: the zip code where the property is located

Property_ID: unique identifier for the property

GIS_ID: GIS identifier for the property

parcel_num: parcel number for the property

X: X coordinate of the property

Y: Y coordinate of the property

Land_Parcel_ID: unique identifier for the land parcel

TLID: top-level identifier for the property

Blk_ID_10: block identifier for the property

BG_ID_10: block group identifier for the property

CT_ID_10: census tract identifier for the property

NSA_NAME: name of the neighborhood statistical area where the property is located

BRA_PD: planning district where the property is located

newcon: binary variable indicating if the permit is for new construction

addition: binary variable indicating if the permit is for an addition

demo: binary variable indicating if the permit is for demolition

reno: binary variable indicating if the permit is for renovation

PermitDuration: duration of the permit in days

government: binary variable indicating if the applicant is a government agency

occupancy_new: binary variable indicating if the occupancy is new

Project_Type: type of project, e.g. residential, commercial, etc.

issue_year: year the permit was issued

Summary Statistics of Declared Valuation and Square Feet of Building Permits dataset.

Declared Valuation Summary Statistics Table	
Min	**0**
1st Quartile	**2100**
Median	**7000**
Mean	**33085**
3rd Quartile	**25000**
Max	**468281**

The minimum declared valuation is 0, and the maximum is 468,281. The median declared valuation is 7,000, which means that half of the values are below this value and the other half are above. The mean declared valuation is 33,085, which is the average of all the values. The first quartile is 2,100, which means that 25%

of the values are below this value, and the third quartile is 25,000, which means that 75% of the values are below this value. The max value is too high meaning there are probably a lot of outliers, and we need to remove those outliers.

Square feet Summary Statistics Table	
Min	**0.01**
1st Quartile	**1000**
Median	**1900**
Mean	**2933.62**
3rd Quartile	**3200**
Max	**24700**

The minimum square footage is 0.01, and the maximum is 24700. The median square footage is 1900, which means that half of the values are below this value and the other half are above. The mean square footage is 2933.62, which is the average of all the values. The first quartile is 1000, which means that 25% of the values are below this value, and the third quartile is 3200, which means that 75% of the values are below this value. . The max value is too high meaning there are probably a lot of outliers, and we need to remove those outliers.

Bi-variate analysis results:

Model-1

```
Call:
lm(formula = neighborhood_dev_score ~ GINI, data = numerical_data)

Coefficients:
(Intercept)         GINI
    0.31159      0.02276
```

The intercept of the model is 0.3115, which represents the expected value of the neighborhood development score when the GINI index is zero. The coefficient for GINI is 0.0227, which means that for each unit increase in the GINI index, the neighborhood development score is expected to increase by 0.0227 units, all else being equal.

Model-2

```
Call:
lm(formula = neighborhood_dev_score ~ MedHomeVal, data = numerical_data)

Coefficients:
(Intercept)   MedHomeVal
  3.209e-01    2.802e-09
```

The intercept coefficient is 0.3209, which means that if "MedHomeVal" were zero (i.e., no median home value in the neighborhood), the predicted value of "neighborhood_dev_score" would be 0.3209. The coefficient of "MedHomeVal" is 2.802-09, This coefficient is very small, which means that for each additional unit of "MedHomeVal" (i.e., for each additional dollar in median home value), the predicted value of "neighborhood_dev_score" would increase by a very small amount.

Model-3

```
Call:
lm(formula = neighborhood_dev_score ~ Hispanic, data = numerical_data)

Coefficients:
(Intercept)     Hispanic
   0.321882     0.003824
```

The intercept coefficient is 0.3218, which means that if "Hispanic" were zero (i.e., no Hispanic population in the neighborhood), the predicted value of "neighborhood_dev_score" would be 0.3218. The coefficient of "Hispanic" is 0.0038, which means that for each additional unit of "Hispanic" (i.e., for each additional percentage point of Hispanic population in the neighborhood), the predicted value of "neighborhood_dev_score" would increase by 0.0038.

Model-4

```
Call:
lm(formula = neighborhood_dev_score ~ HSGrad, data = numerical_data)

Coefficients:
(Intercept)       HSGrad
    0.31895      0.01969
```

The intercept coefficient is 0.3189, which means that if "HSGrad" were zero (i.e., no high school graduates in the neighborhood), the predicted value of "neighborhood_dev_score" would be 0.3189. The coefficient of "HSGrad" is 0.0196, which means that for each additional unit of "HSGrad" (i.e., for each additional percentage point of high school graduates in the neighborhood), the predicted value of "neighborhood_dev_score" would increase by 0.0196.

Multi-variate analysis results:

Model-1

```
Call:
lm(formula = neighborhood_dev_score ~ GINI + AgeU18 + Age1834 +
    Age3564 + White + Black + Asian + Hispanic + FamHousePer +
    FemHeadPer + LessThanHS + HSGrad + SomeColl + Bach + Master +
    Prof + CommuteLess10 + Commute1030 + Commute3060 + Commute6090 +
    ByAuto + ByPubTrans + ByBike + ByWalk + TotalHouseUnits +
    VacantUnitPer + MedGrossRent + MedHomeVal, data = numerical_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.11321 -0.05247 -0.02489  0.01901  0.41946

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.644e-01  2.162e-01   1.223   0.2216
GINI            -1.609e-02  4.989e-02  -0.323   0.7471
AgeU18           3.139e-02  7.892e-02   0.398   0.6909
Age1834         -1.905e-03  5.747e-02  -0.033   0.9736
Age3564          1.237e-03  6.956e-02   0.018   0.9858
White            1.110e-01  1.084e-01   1.024   0.3058
Black            1.059e-01  1.152e-01   0.919   0.3582
Asian            9.427e-02  1.151e-01   0.819   0.4127
Hispanic         1.023e-01  1.104e-01   0.926   0.3545
FamHousePer     -3.519e-02  4.362e-02  -0.807   0.4200
FemHeadPer       7.804e-02  5.761e-02   1.355   0.1757
LessThanHS      -5.663e-02  1.198e-01  -0.473   0.6366
HSGrad          -8.854e-02  1.098e-01  -0.807   0.4200
SomeColl        -7.646e-02  1.308e-01  -0.584   0.5591
Bach            -1.261e-01  1.110e-01  -1.136   0.2562
Master          -5.531e-02  1.256e-01  -0.440   0.6598
Prof            -2.723e-01  1.609e-01  -1.693   0.0907 .
CommuteLess10    2.864e-01  1.428e-01   2.006   0.0450 *
Commute1030      5.916e-02  1.304e-01   0.454   0.6502
Commute3060      8.345e-02  1.288e-01   0.648   0.5172
Commute6090     -2.481e-02  1.494e-01  -0.166   0.8681
ByAuto          -5.046e-02  9.437e-02  -0.535   0.5929
ByPubTrans      -6.410e-02  1.013e-01  -0.633   0.5269
ByBike           2.460e-02  1.467e-01   0.168   0.8669
ByWalk          -9.808e-02  1.015e-01  -0.966   0.3341
TotalHouseUnits  3.468e-06  3.268e-06   1.061   0.2888
VacantUnitPer    5.444e-02  6.697e-02   0.813   0.4164
MedGrossRent     1.203e-05  7.230e-06   1.664   0.0963 .
MedHomeVal       1.447e-08  1.831e-08   0.790   0.4296
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08327 on 1789 degrees of freedom
Multiple R-squared:  0.01837,   Adjusted R-squared:  0.003005
F-statistic: 1.196 on 28 and 1789 DF,  p-value: 0.221
```

The coefficients for each predictor variable indicate the estimated effect of a one-unit change in the predictor variable on the response variable, holding all other variables constant. For example, the coefficient for CommuteLess10 is 0.2864, which means that a one-unit increase in the percentage of people in the neighborhood who have a commute time of less than 10 minutes is associated with an increase of 0.2864 in the neighborhood development score, holding all other variables constant.

The F-statistic tests the overall significance of the model, and in this case, it has a value of 1.196, which is not statistically significant (i.e., the p-value is greater than 0.05). This suggests that the model as a whole may not be a good fit for the data.

The R-squared for the model is 0.01837, which means that only 1.837% of the variance in the dependent variable (neighborhood_dev_score) can be explained by the independent variables included in the model.

Model-2

```
Call:
lm(formula = neighborhood_dev_score ~ GINI + TotalHouseUnits +
    VacantUnitPer + MedHomeVal + ByWalk + FemHeadPer + FamHousePer +
    Hispanic + PopDen + AgeU18 + HSGrad, data = numerical_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.10601 -0.05256 -0.02594  0.01800  0.42691

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.342e-01  1.852e-02  18.045  < 2e-16 ***
GINI            -5.772e-02  3.972e-02  -1.453 0.146343
TotalHouseUnits  3.453e-06  2.992e-06   1.154 0.248657
VacantUnitPer    2.948e-02  5.515e-02   0.535 0.592967
MedHomeVal       2.036e-08  1.437e-08   1.417 0.156717
ByWalk           4.635e-02  2.351e-02   1.972 0.048802 *
FemHeadPer       5.623e-02  3.844e-02   1.463 0.143718
FamHousePer     -2.887e-02  2.959e-02  -0.975 0.329484
Hispanic         2.296e-02  1.843e-02   1.246 0.213084
PopDen          -6.235e-07  1.726e-07  -3.613 0.000311 ***
AgeU18           4.844e-03  5.977e-02   0.081 0.935418
HSGrad           2.533e-02  4.026e-02   0.629 0.529357
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08321 on 1806 degrees of freedom
Multiple R-squared:  0.01052,   Adjusted R-squared:  0.004495
F-statistic: 1.746 on 11 and 1806 DF,  p-value: 0.05846
```

The positive coefficient of ByWalk (p-value = 0.048802) suggests that neighborhoods with a higher percentage of residents who walk to work tend to have higher development scores. Similarly, the negative coefficient of FemHeadPer (p-value = 0.143718) indicates that neighborhoods with a higher percentage of female-headed households tend to have lower development scores. The negative coefficient of PopDen (p-value = 0.000311) suggests that higher population density is associated with lower development scores.

The other coefficients, including TotalHouseUnits, VacantUnitPer, MedHomeVal, FamHousePer, Hispanic, AgeU18, and HSGrad, are not statistically significant, indicating that they are not good predictors of neighborhood development score.

The multiple R-squared value of the model is 0.01052, indicating that the model explains only a small proportion of the variation in neighborhood development score.

The F-statistic is 1.746 with a p-value of 0.05846, which is not statistically significant, indicating that the overall model fit is not significant.

Model-3

```
Call:
lm(formula = neighborhood_dev_score ~ redline, data = numerical_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.08736 -0.05432 -0.02616  0.01831  0.42866

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.320474   0.002616 122.509   <2e-16 ***
redline     0.004800   0.003939   1.219    0.223
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08338 on 1816 degrees of freedom
Multiple R-squared:  0.0008172, Adjusted R-squared:  0.000267
F-statistic: 1.485 on 1 and 1816 DF,  p-value: 0.2231
```

The coefficient for the redline is 0.0048, indicating that the presence of a redline is associated with an increase of 0.0048 in the neighborhood development score, but this is not statistically significant since the p-value is 0.223. The multiple R-squared value is 0.0008172, indicating that only a small proportion of the variance in the neighborhood development score can be explained by the presence of a redline. The F-statistic is 1.485, with a p-value of 0.2231, indicating that the model is not significant.

Code Excerpts and Datasets

This is the link to my GitHub repository for the final project code:

https://github.com/dvijashah/Big-Data-For-Cities/blob/main/final_project.Rmd

This code can be used to replicate this work.

Datasets:

Building Permits:
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/B7DHBK

Massachusetts Census Indicator:

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/XZXAUP

Redlining in Boston:
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/WXZ1XK

Tracts: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/SQ6BT4

This is the code excerpts: You can follow this step by step to replicate my work. The explanation for this has been provided in the Data and Methods section as well as Statistical Analysis section.

Load the libraries and the dataset:

```r
```{r}
library(tidyverse)
library(lubridate)
library(sqldf)
library(sf)
require(ggplot2)
require(ggmap)
library(knitr)
```
```

```r
```{r}
#read the dataset
building_permits <- read_csv('C:/Users/Dvija Shah/Desktop/Dvija/Courses/PPUA/Permits.Records.2021.csv')
```
```

Remove 0 from sq_feet and declared_valuation

```{r}
#remove 0 values for sq_feet and declared valuation
building_permits <- building_permits[building_permits$sq_feet != 0, ]
building_permits <- building_permits[building_permits$DECLARED_VALUATION != 0, ]
```

Recategorizing occupancy feature:

```{r}
#re-categorizing the OCCUPANCY variable and creating occupancy_new feature.

building_permits$occupancy_new <- ifelse(building_permits$OCCUPANCY %in% c("1-4FAM", "1-7FAM", "1-2FAM", "1-3FAM",
"1Unit"), "Residential",
ifelse(building_permits$OCCUPANCY %in% c("COMM", "Comm"), "Commercial",
ifelse(building_permits$OCCUPANCY %in% c("MIXED", "Mixed"), "Mixed",
ifelse(building_permits$OCCUPANCY %in% c("Multi", "4unit", "5unit", "2unit", "3unit", "6unit", "7unit", "6Unit",
"4Unit"), "Multi-Unit",
ifelse(building_permits$OCCUPANCY %in% c("VacLd", "Other"), "Other", "Unknown")))))
```

Plotting number of building permits by occupancy:

```{r}
#Plotting the Number of Building Permits by occupancy_new type.
ggplot(building_permits, aes(x = occupancy_new, fill=occupancy_new)) +
  geom_bar() +
  labs(title = "Number of Building Permits by Occupancy Type",
       x = "Occupancy Type",
       y = "Count")
```

Creating Project_type feature from the types of work flag variables and plotting it.

```{r}
#Creating a new feature Project_Type from the existing flag variables.
building_permits$Project_Type <- NA

building_permits$Project_Type[building_permits$newcon == 1] <- "New Construction"
building_permits$Project_Type[building_permits$addition == 1] <- "Addition"
building_permits$Project_Type[building_permits$demo == 1] <- "Demolition"
building_permits$Project_Type[building_permits$reno == 1] <- "Renovation"

building_permits$Project_Type <- as.factor(building_permits$Project_Type)
```

```{r}
#Plotting the Number of building Permits by Project_Type
library(ggplot2)
ggplot(building_permits, aes(x = Project_Type, fill=Project_Type)) +
  geom_bar() +
  labs(title = "Number of Building Permits by Project Type",x = "Project Type", y = "Count") +
  theme_minimal()
```

Removing outlier from declared valuation and sq feet.

```{r}

#Removing outliers from declared valuation

# calculate the IQR
q1 <- quantile(building_permits$DECLARED_VALUATION, 0.01)
q3 <- quantile(building_permits$DECLARED_VALUATION, 0.90)
iqr <- q3 - q1

# define the upper and lower bounds for outliers
upper <- q3 + 1.5*iqr
lower <- q1 - 1.5*iqr

# remove outliers
building_permits <- building_permits[building_permits$DECLARED_VALUATION >= lower &
building_permits$DECLARED_VALUATION <= upper,]


```

```{r}

#Removing outliers from sq_feet

# calculate the IQR
q1 <- quantile(building_permits$sq_feet, 0.01)
q3 <- quantile(building_permits$sq_feet, 0.90)
iqr <- q3 - q1

# define the upper and lower bounds for outliers
upper <- q3 + 1.5*iqr
lower <- q1 - 1.5*iqr

# remove outliers
building_permits <- building_permits[building_permits$sq_feet >= lower & building_permits$sq_feet <= upper,]

```

Creating neighbourhood_development from building permits:

```{r}
neighborhood_development <- sqldf("SELECT CT_ID_10, NSA_Name, occupancy_new, Project_Type,
                                          COUNT(*) AS no_of_permits_issued,
                                          AVG(DECLARED_VALUATION) AS avg_declared_valuation,
                                          AVG(sq_feet) AS avg_sq_ft
                                   FROM building_permits
                                   GROUP BY CT_ID_10, occupancy_new, Project_Type")

neighborhood_development <- neighborhood_development[order(-neighborhood_development$no_of_permits_issued),]

head(neighborhood_development,100)
```

Process to calculate the neighborhood development score latent variable:

```{r}
# Calculate the maximum values for each variable
max_no_of_permits_issued <- max(neighborhood_development$no_of_permits_issued)
max_avg_declared_valuation <- max(neighborhood_development$avg_declared_valuation)
max_avg_sq_ft <- max(neighborhood_development$avg_sq_ft)
```

Calculating percentage of commercial, residential, mult-unit, mixed, renovation, new construction and addition.

```r
# calculate the total number of permits issued for each occupancy category
occupancy_counts <- table(neighborhood_development$occupancy_new)


# calculate the total number of permits issued for all other occupancy categories
total_occupancy_counts <- sum(occupancy_counts)


# calculate the percentage of permits issued for commercial occupancy among other occupancy categories
commercial <- occupancy_counts["Commercial"] / total_occupancy_counts

# calculate the percentage of permits issued for residential occupancy among other occupancy categories
residential <- occupancy_counts["Residential"] / total_occupancy_counts

# calculate the percentage of permits issued for multi-unit occupancy among other occupancy categories
multi_unit <- occupancy_counts["Multi-Unit"] / total_occupancy_counts

# calculate the percentage of permits issued for mixed occupancy among other occupancy categories
mixed <- occupancy_counts["Mixed"] / total_occupancy_counts
```

```r
# calculate the total number of permits issued for each project type
project_type_counts <- table(neighborhood_development$Project_Type)

# calculate the total number of permits issued for all other occupancy categories
total_project_type_counts <- sum(project_type_counts)

# calculate the percentage of permits issued for renovation
renovation <- project_type_counts['Renovation']/total_project_type_counts

# calculate the percentage of permits issued for addition
addition <- project_type_counts['Addition']/total_project_type_counts

# calculate the percentage of permits issued for new construction
new_construction <- project_type_counts['New Construction']/total_project_type_counts
```

Calculating neighbourhood development score:

```r
neighborhood_development$neighborhood_dev_score <-
(neighborhood_development$no_of_permits_issued/max_no_of_permits_issued)*0.2+
        (neighborhood_development$avg_declared_valuation/max_avg_declare_valuation)*0.2 +
                  (neighborhood_development$avg_sq_ft/max_avg_sq_ft)*0.2 +
                                              (((commercial) + (residential) + (multi_unit) + (mixed))*0.2)+
                                              (((renovation)+ (addition) + (new_construction)) *0.2)
```

Correlation:

Loading the datasets:

```r
redlining <- read_csv('C:/Users/Dvija Shah/Desktop/Dvija/Courses/PPUA/Boston_Tracts_2010_HOLC.csv')

demographics <- read_csv('C:/Users/Dvija Shah/Desktop/Dvija/Courses/PPUA/ACS_1519_TRACT.csv')
```

Subsetting and merging the datasets:

```r
#subsetting demographics to keep only numerical columns
demographics_new = subset(demographics, select = -c(NAME, TOWN, COUNTY, MAtown, MedYrBuilt, TOWN_ID, FIPS_STCO  )
```

```r
#subsetting neighborhood development to keep only important columns
neighborhood_dev_new = subset(neighborhood_development, select = -c(no_of_permits_issued, occupancy_new,
Project_Type, avg_declared_valuation, avg_sq_ft  ))
```

```r
#remove the null values
demographics_new <- na.omit(demographics_new)
```

```r
#merging the datasets
tracts<-merge(neighborhood_dev_new,redlining,by='CT_ID_10',all.x=TRUE)
tracts<-merge(tracts,demographics_new,by='CT_ID_10',all.x=TRUE)
```

Remove non-numerical columns:

```r
#Remove non-numerical variables:
numerical_data <- tracts %>%
  select_if(is.numeric)
```

Checking the correlation of Neighborhood development score with all the numerical variables.

```r
data_cor <- cor(numerical_data[ , colnames(numerical_data) != "neighborhood_dev_score"],
                numerical_data$neighborhood_dev_score)
data_cor
```

Conducting bi-variate analysis:

```r
b1 <- lm(neighborhood_dev_score~GINI, data=numerical_data)
b1
```

```r
b3 <- lm(neighborhood_dev_score~MedHomeVal, data=numerical_data)
b3
```

```{r}
b4 <- lm(neighborhood_dev_score~Hispanic, data=numerical_data)
b4
```

```{r}
b5 <- lm(neighborhood_dev_score~HSGrad, data=numerical_data)
b5
```

Conducting multi-variate analysis:

```{r}
m1 <- lm(neighborhood_dev_score ~ GINI + AgeU18 + Age1834 + Age3564 +
             White + Black + Asian + Hispanic + FamHousePer + FemHeadPer +
             LessThanHS + HSGrad + SomeColl + Bach + Master + Prof +
             CommuteLess10 + Commute1030 + Commute3060 + Commute6090 +
           ByAuto + ByPubTrans + ByBike + ByWalk +
           TotalHouseUnits + VacantUnitPer + MedGrossRent + MedHomeVal,
             data = numerical_data)

summary(m1)
```

```{r}

m2 <- lm(neighborhood_dev_score ~ GINI + TotalHouseUnits + VacantUnitPer + MedHomeVal + ByWalk +
          FemHeadPer + FamHousePer + Hispanic + PopDen + AgeU18 + HSGrad , data= numerical_data)

summary(m2)

```

```{r}

m3 <- lm(neighborhood_dev_score ~ redline, data= numerical_data)

summary(m3)

```

Visualizations:

Dot plot

```{r}
base5<-ggplot(data=numerical_data, aes(x=neighborhood_dev_score, y=Hispanic)) +
    geom_point() + xlab("Neighborhood Development Score") +
    ylab("Hispanic")
base5
```

Dot plot with linear model:

```{r}
base5 + geom_smooth(method=lm)
```

Choropleths:

Loading the dataset and merging it.

```{r}
tracts_geo<-st_read("C:/Users/Dvija Shah/Desktop/Dvija/Courses/PPUA/shape/Tracts_Boston_2010_BARI/Tracts_Boston BARI.shp")
```

```
Reading layer `Tracts_Boston BARI' from data source
  `C:\Users\Dvija Shah\Desktop\Dvija\Courses\PPUA\shape\Tracts_Boston_2010_BARI\Tracts_Boston BARI.shp'
  using driver `ESRI Shapefile'
Simple feature collection with 178 features and 16 fields
Geometry type: POLYGON
Dimension:     XY
Bounding box:  xmin: -71.19115 ymin: 42.22788 xmax: -70.98471 ymax: 42.40493
Geodetic CRS:  NAD83
```

```{r}
tracts_geo<-merge(tracts_geo,tracts,by='CT_ID_10',all.x=TRUE)
```

Plotting the maps:

```{r}
plot(tracts_geo['neighborhood_dev_score'], main = 'Neighborhood Development Score in Boston', breaks='quantile')
```

```{r}
plot(tracts_geo['GINI'], main = 'GINI in Boston', breaks='quantile')
```