Proposal for Examination of the Intersection of Race, Technology and Housing
Diana Bowser, Kartik Travedi, Kaili Mauricio, and Lisa Thorn

Introduction

Despite being the richest country in the world, the United States is suffering from a severe housing crisis; a crisis that is immensely intertwined with race, gender, justice and technology. For example, people of color in the United States are nearly twice as likely to be denied home loans compared to white applicants (Martinez & Kirchner, 2021). In addition, even when Black people earn a manageable income they are denied a home loan 10% more than similarly earning white applicants (Martinez & Kirchner, 2021).

While it might seem that technology can solve this problem, as options for improved technology have grown, the housing crisis has not diminished. As a matter of fact, lending approval algorithms, which were developed throughout the second half of the 20th century, are embedded with the same racial bias that resulted in redlining to deny Black people from moving into newly developed suburbs. The fact that we still mainly rely on credit scores to determine loan outcomes, despite the fact that financial patterns have evolved since 1995 when the Fair Isaac Corporation (FICO) credit score was institutionalized by the mortgage lending industry, reiterates the complexity of the situation. In addition, many groups often have a difficult time even accessing the proper technology that they might need to improve their housing situation.

All of the above factors (housing, inequalities, race, technology) are all factors that continue to contribute to the disparities in wealth between Black and white families. These discriminatory outcomes compound with gender identity, sexuality, disability, socioeconomic status, and geography.

This study will use mixed methods to examine the relationship between housing and housing insecurity, race, gender, and technology from three wo viewpoints: 1. The individual level, examining those who apply for a home mortgage and are denied; 2. The sub-national level in the United States, capturing county and zipcode trends in housing across the United States; and 3) Qualitatively focusing on a broader exploration of the mortgage lending process including the variables and information used in banking and lending algorithms. The first analysis will answer the following question: What are the main reasons that individuals are denied a home mortgage and does this this vary by race, gender, and geography? The second analysis will use machine learning and cluster analysis to examine the variation in the housing market (type of housing, value of homes) by social determinant of health clusters. The qualitative questions will enhance the quantitative analysis and ask the following questions: How are banking and lending algorithms trained and regulated? How do these algorithms overlap with race, gender and loan outcomes?

Data and Methods

The innovation of this project will be the use of a number of publicly available databases to capture the interaction between housing, race and technology in the United States from several perspectives. The secondary databases to be explored are explained below.

Data Description

Home Mortgage Disclosure Act (HMDA)

The Home Mortgage Disclosure Act (HMDA) will be utilized to understand the relationship between mortgage application decisions and credit scores and background checks and how this varies by race, gender and geography. The HMDA provides data on mortgage applications origins, application denials, and incomplete applications. Datasets under HMDA contain information about the loan amount, limited property characteristics, the applicant demographics and the lender. Publicly available dataset holds limited information on applicants and borrowers. HMDA data captures nearly 90% of mortgage origination in the US with lenders' names (Engel and McCoy, 2011). HMDA does not have information on loan terms or structure (interest rates, maturity, loan-to-value ratios), or the type of conventional loan (fixed, ARM). Similarly, HMDA provides limited information on the property characteristics where it only includes the census tract of the property as the finest geographic characteristic. Several financial and demographic characteristics are available at the census tract level only. HMDA is available for 2007 until 2022, with them most reliable data available up until 2017. HMDA also provides dynamic data for the current year to capture as it is reported to HMDA. A key strength of the data is that it captures nearly 90% of the loan origination in the US. This allows for a comprehensive analysis. Furthermore, the loan-related information can be linked to the applicant's race and ethnicity.

Variables of interest

HMDA is a unique dataset that captures loan origination and denial with reasons. HMDA is also administrative data and, therefore, less likely to have errors in loan related information. The key variables of our interest for our research are the main reasons that an individual is denied a home loan which include the following: 1. Debt-to-income ratio 2. Employment history 3. Credit history 4. Collateral 5. Insufficient cash (down payment closing costs) 6. Unverifiable information 7. Credit application incomplete 8. Mortgage insurance denied 9. Other. We will also use other available demographic information (race, gender, geography, etc.) to determine if the loan denial reasons vary by any of these indicators. We will also examine the interaction with technology through an investigation of who has access to online mortgage application processes versus traditional brick and mortar type of lending institutions; using access to an online lender in the models described below.

The variables connecting HMDA with other datasets are geographic indicators such as census tract and zipcode. Using the census tract, one can connect HMDA with the American Community Survey or similar census data at an aggregated level.

Zillow Housing Market Metrics

Zillow Housing Market Metrics (ZHMM) data will be utilized in the machine learning/cluster analysis to examine variation in key housing market metrics by social determinant of health clusters. ZHMM are derived from house postings (sales and rentals) on Zillow and contains several housing market metrics available at the national level to the neighborhood level.

ZHMM is available in two ways. One can download aggregated data at a national and metro level from the website. Housing data available on the website includes the home value index, rent index, inventory, and property pricing (aggregated by geographic level), and sales and price cuts (at a given geographic level). Second, the way to get the ZHMM is to use the API and download data. Downloading data from API offers the same data available on the website but at different geographic levels and the option to define the time period.

We will utilize the ZHMM to examine the following variables across SDOH clusters (zipcode level) to understand the housing market in the US from a social determinant perspective. Variables of interest are: Median list price, median sale price, sale to list ratio (mean/median), percent of sales under/over list and days to close are some of the variables related to the research question. These metrics are available at a national level, and at the national level, state, county, metro, city, region (Zillow defined) and at Zipcode level.

American Housing Survey 2019 (AHS)

The American Housing Survey (AHS) is survey conducted by the U.S. Census Bureau for the Department of Housing and Urban Development (HUD). It is the most comprehensive national housing survey in the US. AHS is conducted every two years, and the latest data is from 2019. HUD provides public use file for the AHS with detailed information on the housing unit along with detailed information on the residents including family composition, demographics, income etc. AHS data are available for download from the AHS' website. Along with the demographic information, income and householder information, AHS also has house value, purchase price, and source of the down payment for owner-occupied units. Additionally AHS has questions on mortgage characteristics (total remaining debt, current total loan, line of credit, interest rate, mortgage term, outstanding principal amount etc.). Indications that black and Hispanic populations get housing loans on less than favorable terms make these variables useful for our analysis.

Agency for Health Care Research and Quality (AHRQ) Social Determinant of Health Database (SDOH)

The AHRQ SDOH is a complication of variables from different publicly available databases, available at the census track and zipcode level, across the following domains: social context, economic context, education, physical infrastructure, health care context, and geography. The data are available across multiple years at both the county and zipcode level. There are 350

number of variables at the county and there are 160 number of variables at the zipcode level. We will be using year 2009-2018 for the county analysis and 2011-2018 for the zipcode analysis.

Qualitative Data

Qualitative data will be collected via interviews with key stakeholders and a qualitative exploration of some of the quantitative results. See qualitative method below.

Methods

HMDA Analysis

We will analyze HMDA mainly for mortgage denial reasons/characteristics and conduct a detailed descriptive analysis of the data. Correspondingly we will conduct a logit analysis to understand the relationship between loan application denial and applicants' race. Further, we will look at the relationship between different reasons behind denial and illustrate whether race plays a role in denial categorization. One method that will be investigated is the Oaxaca Blinder decomposition analysis that can estimate the contribution to disparity in loan denials by race and other factors. After the descriptive analysis, we will incorporate various methods, like decomposition analysis, to estimate the impact race and other factors play in the rate and type of loan denial. Finally, we will evaluate if the rate spread for accepted loan applications is affected by race.

Loan Denial Analysis
Logit
Denial = a(Minority) + b(Month*Year) + c(Govt. Sponsored Enterprise) + d(technology bias) + e(lender*county) + f(Tract characteristic)

MLogit
Denial Type = a(Minority) + b(Month*Year) + c(lender*county) + d(technology bias) + error

Acceptance Analysis
OLS
Rate spread = a(minority) + b(Month*Year) + c(Lender*County) + d(Govt. Sponsored enterprise) + e(technology bias) + f(Tract characteristic) + error

Machine Learning-Subnational Analysis

We will use the AHRQ SDOH database to identify 10 different clusters in the United States based on a reduced form 200 variables at the county and zipcode level across the six domains explained above (social context, economic context, education, physical infrastructure, health care context, and geography). We will use kmeans partition cluster analysis for the cluster analysis. We will then merge with the ZHMM data on housing markets at the appropriate geographic level by the 10 identified clusters to understand the variation in housing in the United States.

Qualitative Mixed Methods

We will use qualitative mixed methods to enhance the qualitative analysis. First, we will identify key stakeholders in the banking and lending industry to understand further the banking and lending process and how algorithms are used in this process. Secondly, we will use the results of the machine learning, county and zipcode analysis to delve deeper into the housing outcomes in certain counties and/or zipcodes taking into consideration lending resources across counties, cultural attitudes toward home ownership, transparency in the algorithms, impact on the unhoused, cost of living and other regional market forces. Finally, if possible, we will also identify individual who have used algorithms to apply for home loans and their experience with this process. We want to examine algorithms by the people and time in which they were created and to acknowledge when and how we have progressed and developed as a society so that the algorithm is molded to society and not the other way around.

Work Plan and Time line

Key milestone, achievement or result	Activity(ies)	Completion date
HMDA Initial Analysis	Download, clean and begin analysis with HMDA database	September-October 2022
ARHQ SDOH cluster analysis	Finalize ARHQ SDOH cluster analysis	September-October 2022
HMDA Modeling	Using HMDA data to answer research question relating loan denial reason to loan denial rates by race, gender and geography	October-November 2022
Blog Development HMDA	Contribution to the IERE Blog on HMDA results	November-December 2022
Manuscript Development-HMDA	Development of one or more manuscript using the HMDA data	January-March 2023
Zillow Database Download and Cleaning	Finalize Zillow Analysis	November-December 2022
Cluster-Zillow Combination	Combine the Cluster Analysis with Zillow Data	January-March 2023
Blog Development Cluster Analysis	Contribution to the IERE Blog on Cluster Analysis	January-March 2023
Manuscript Development-Cluster	Development of one or more manuscript using the Cluster-Zillow Data	April-May 2023
Qualitative Methods	Finalize development of qualitative methods and identify key informants	October-December 2022
Qualitative Data Collection	Conduct interview and qualitative data collection	January-March 2023
Mixed Methods	Synthesize qualitative and quantitative results	March- May 2023
Research Uptake	Synthesis of finding into blogs and other formats easily accessible to many audiences	April- May 2023
Dissemination of results	Dissemination of results to Brandeis, Heller, Kappor and other communities	September 2022-May 2023
Weekly Zoom calls; Monthly Consortium Zoom calls	**Weekly Zoom** calls between Brandeis Team Members; **Monthly Zoom** calls between Brandeis University and Kappor and other relevant stake holders	September 2022- May 2023

Budget

We are proposing a budget of $175,000 for the time period September 2022- May 2023 for the proposed mixed-methods analysis described above. This includes salary coverage for PI, Diana Bowser (20%), as well as three PhD Graduate Research Assistants (Kartik Travedi and Kaili Mauricio and one PhD GRA for the qualitative methods and analysis). MPP student, Lisa Thorn, has been involved in this project since the very beginning and she is also included on this project as a key graduate research assistant. She will be involved in the quantitative methods and

analysis We are also hoping to include an undergrad in our project and have reached out to Michaela McCormick (who has also worked on this project up until now) and Yanlin Hu.