Team 3: Cobra Py
DATS 6103: Summary Report
Professor Edwin Lo
May 4, 2022

Key Community Characteristics Affecting Property Values in DC

Introduction

At the end of 2020, homeownership in the USA had substantially increased (up by 0.7% from 2019), similar to the homeownership increase seen between 2003 and 2004 (0.9 %). Despite the financial crisis brought on by the pandemic, buyers sought to take advantage of working from home and record low mortgage rates. However, as home ownership rates grew, so did home prices, averaging $391,900 in 2020 (Statistica), a 10.4% year over year gain (S&P CoreLogic Case-Shiller US National Home Price Index). Location is a major factor in determining a home's value, but other external factors are often overlooked. The US Appraisal Institute assesses that factors such as "bad neighbors could potentially reduce your home's value by more than 5 to 10 percent".

Using The Federal Financial Institutions Examination Council's (FFIEC) Loan Application Register (LAR) as our data source, we determined which community characteristics affect property values. In theory, we would expect that almost all the price variance for property values would come from the physical attributes of the property, such as square footage, number of bedrooms, etc. However, we found that about 24% of the variance in our data came from community factors rather than physical attributes of the property. We defined community characteristics as factors that are unrelated to the physical characteristics of the property. Broadly thought of as the neighborhood makeup, these include, but not limited to, the percentage of the community members that are considered minorities, whether the community consists of primarily homeowners or renters, income levels of members, and number of family members per household.

Furthermore, we wanted to hone in on the bias in social groups that affect mortgage loans and the associated costs. Looking broadly at females and racial minorities, we assessed if there are any statistically significant differences in how much each group paid in interest rates and loan application fees for their mortgage applications. This is important because institutional lending practices can often indirectly affect a neighborhoods social make up.

SMART Questions

With the intention of focusing the project to relevant aspects of our own communities, we focused on recent data for properties in Washington DC (please see Description of Data). In particular, we inquired into to two aspects of property value and the loans associated with them:

1. How much of the property prices in DC were affected by community factors in 2020?
2. Do women and minorities differ in how much they pay in interest rates and loan costs?

Literature Review

Prior research in this area has mostly focused on the effect of Homeowner and Community Associations on the discriminatory practices of such institutions. Others have examined how individuals and institutions respond to regulatory or policy changes leading to "unequal distribution of resources for the target population" (Lee and Bostic, 2020; Koning and Heinrich, 2013; Heckman et al., 1997).

More recent studies are being done about how community characteristics are affected even before a family even goes to see a house. Studies such as those conducted by Bhutta and Hizmo's (2020) and Bartlett et al. (2019) report that "lenders continue to discriminate" and find " interest rate gaps between minority and white borrowers" (Bartlett et al, 2019).

Description Of Data

As previously mentioned, our data was sourced from the FFIEC's Loan Application Registry for Washington, DC in 2020. The data is compiled annually under the Home Mortgage Disclosure Act (HMDA) with the intent "to create greater transparency and to protect borrowers in the residential mortgage market" (citation). The LAR provides extensive information about the lending activities of all US financial institutions "…regarding its applications, originations, and purchases of home purchase loans, home improvement loans, and refinancings" (citation). We subsetted this data frame to only look at loan information for Washinton, DC which gave us just over 58,000 data points to work with.

Preparing The Data

We divided our data into two data frames, titled overcharged and property value

Property Value

Property Value
Dwelling Category
Construction Method
Total Units
Interest Rate
Population
% Minority Population
Median Family Income
Owner Occupied Units
1-to-4 Family Homes
Median Age of Housing Units

Overcharged

Total Loan Costs
Interest Rate
Sex of Applicant
Discount Points
Lender Credits

From here, we cleaned up the data, ensuring that all variables were of the appropriate data type for our analysis. We also checked for and dealt with any NA's in the data, as well as adjusted filters for the data based on our investigative goals. For example, we used the property value standard deviation to filter out any homes that were more than 1 standard deviation above the mean for property value. This dealt with potential outliers, made the graphs easier to use and interpret, and attempted to focus the data on family homes (by hopefully filtering out more expensive commercial properties).

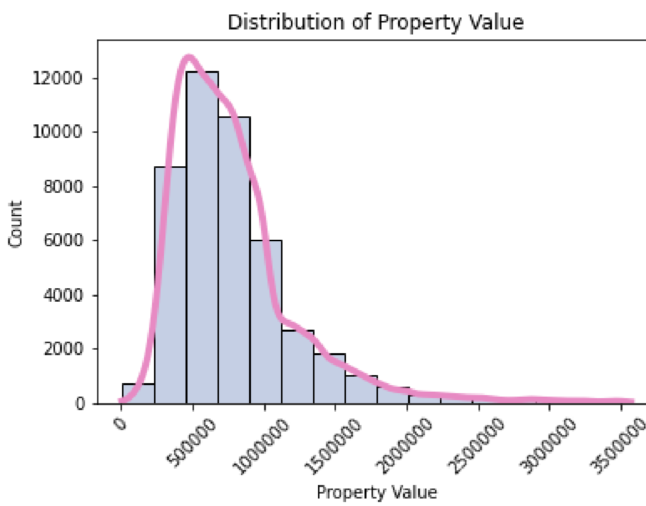
After cleaning up the data we moved onto the EDA and modeling.

Analysis

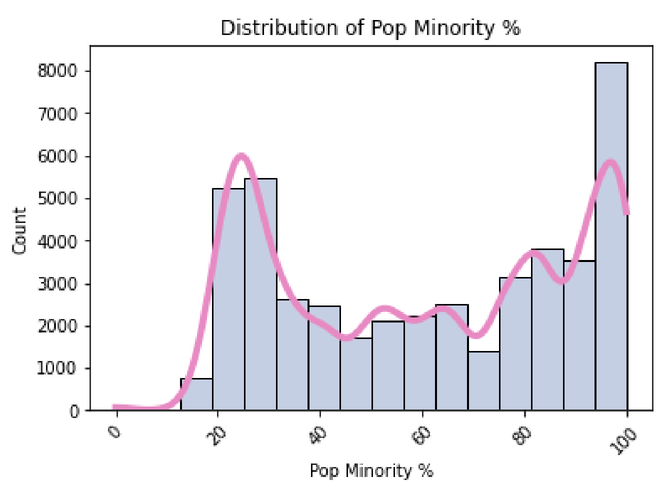
Property Values & Community Characteristics

EDA and Graphs

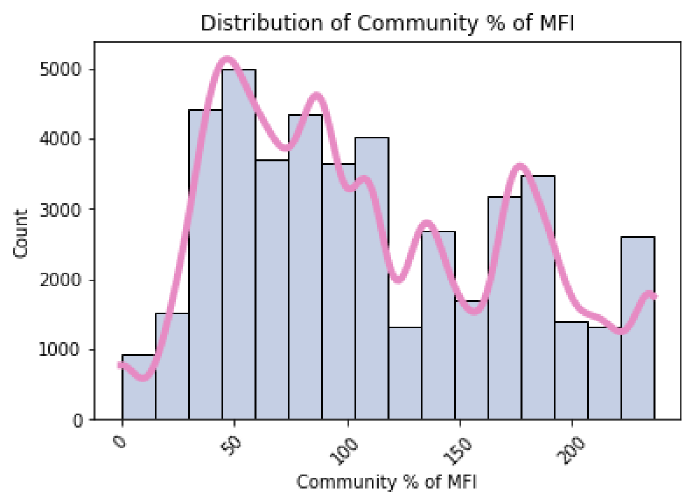
- Property Values showed a unimodal distribution, although it was not normal. The peak of the curve was around $500,000 - $ 750,000, with a fat tail trailing out to the cut off at $3,5000,000.



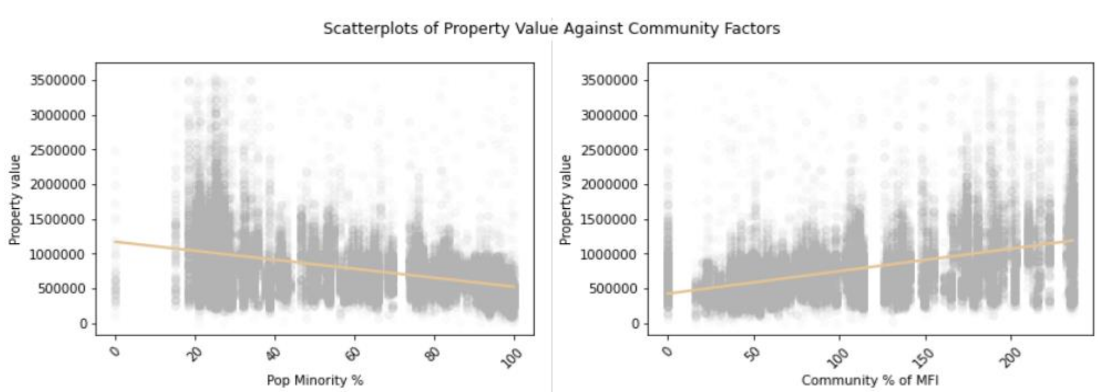
- The percentage of the community that was a minority had a bimodal distribution, with the first peak coming around a minority percentage of 20%, and the second peak around a minority percentage of 90-100%. This suggests that people tend to be clustered in either very low or very high minority neighborhoods.



- The median family income of a community relative to the regional value was multimodal. Broadly speaking there were two major peaks, the first encompassing incomes in the 50-100% range, and a second at around 175%. In more detail, there were peaks around 50%, 85%, 130%, and 175%. This indicates that neighborhoods tend to be grouped according to common income levels.



- The community percentage of Regional Family Median Income (MFI) (0.46), the community percentage of minority residents (-0.42), the number of owner occupied homes (0.21), and the number of 1-4 family homes (0.21) were kept for showing a correlation score greater than 0.2 with property value. Other numeric variables were dropped.
- The community percentage of MFI and the community percentage of minority residents also showed a change in the variance of property value in addition to increasing or decreasing average values. As the percentage of MFI increased both the average property value and the variance of the values increased. As the percentage of a community's minority population increased, both the average property value and the variance of the values decreased.



Feature Selection

- The categorical variables examined tended to be dominated by one level, didn't show the difference in interaction with other variables based on the category level, and therefore were dropped from the analysis.
- The numeric variables showed very high multicollinearity after calculating their VIF scores. After exploring various combinations, the final model used the community percentage of MFI, the community percentage of minority residents, and the number of 1-4 family homes as they all had VIF scores under 10.

Feature	VIF Score
Pop Minority %	3.098
Community % of MFI	3.396
Num 1-4 Family Homes	6.302

Modeling

The goal of this analysis was to determine a baseline for how much of the variance in property value could be attributed to community characteristics, irrespective of the physical characteristics of the home. Thus, a linear regression model was used with a focus on the adjusted r^2 value.

Model Comparisons			
Model	**Adjusted r^2**	**P-Values**	**VIF Score**
All Variables	0.299	All low	Very High (10+)
Only Community Variables	0.255	All low	Very High (10+)
Only Low VIF Variables	0.235	All low	Lower (~3-6)

- The model that used all variables, irrespective of feature selection, was able to account for ~30% of the variance seen in property values. However, the extreme multicollinearity amongst the features means that this r^2 value is likely inflated. Additionally, the EDA revealed that the majority of the categorical variables did not provide meaningful differences, and thus are also likely contributing to an inflated accounting of the variance.

- The model using only the numerical variables, which were the variables explicitly about community characteristics also had high multicollinearity, and thus its r^2 values of 0.255 also likely indicates an inflated accounting of the variance of property value.
- The final model that uses the community percentage of MFI, the community percentage of minority residents, and the number of 1-4 family homes represents a more reliable accounting of variance due to its more reasonable VIF levels. In this model, upwards of 24% of the variance in property value can be attributed to community factors alone. Additionally, the 95% confidence interval for the coefficients of the features were fairly narrow, further indicating that the model could be reasonably relied on to understand the impact each feature had on property value. The final logistic model equation is as follows:

  $$\text{Property Value} = \$654{,}700 - \$3314.46^{\,\%\text{ minority}} + \$1739.74^{\,\%\text{ MFI}} + 137.97^{\,\#\text{ 1-4 family homes}}$$

<u>Interpretation</u>

Communities in the DMV area tend to be clustered into either low or higher minority neighborhoods. They also tend to be grouped by income levels. The property value of a home could potentially be affected by the community characteristics of the neighborhood the home is located in, irrespective of the physical characteristics of the home. The exact same home could vary in value by upwards of 24% depending on those community characteristics. This effect is negatively associated with higher percentages of minority residents and positively associated with higher incomes relative to the regional family median income level. While many control factors need to be considered, and will be mentioned in the limitations section, these findings warrant deeper exploration.

Interest Rates & Loan Costs Among Women & Minorities

To investigate whether diverse social groups differ in how much they pay in interest rates and mortgage loan costs, we chose the appropriate variables from our data. We used Total Loan Cost (henceforth TLC) and the annual mortgage Interest Rates (henceforth IR) as our objective (LHS) variables. For our explanatory (RHS) variables, we chose Sex, and Race. We were also curious about measuring the easiness in which the individuals from those groups could access opportunities to structure and pay their loans, so we chose Discount Points and Lender Credits as explanatory (RHS) variables in their own models too. Although these last two variables are only a partial representation of the easiness of payment of a loan, they are a reasonable starting point for group comparisons.

Using linear models, we ran both a one-way ANOVA and two-way ANOVA tests between the target variables and explanatory variables as well as their interactions. Finally, we conducted a Tukey test to observe any significant mean differences between the social groups.

The algebraic form of our linear models as follows:

$$Z = A_0 + A_1 \rho + \varepsilon$$

Where,

$$Z = \begin{bmatrix} TLC \\ IR \end{bmatrix} \qquad \rho = \begin{bmatrix} Sex \\ Race \\ Disocunt\ Points \\ Lender\ Credits \end{bmatrix}$$

A_0 = constant terms $\qquad A_1$ = Coefficients for explanatory variables

ε = error terms

Note that,

IR is the annual interest rate of the mortgage loan.

TLC is the net cost of acquiring a loan, calculated by subtracting the total amount paid by a borrower during the duration of the loan minus the capital borrowed. It is not to be confused with the monthly mortgage payment that a borrower will pay during the life of the loan.

Derived Sex has three levels: Female, Male, and Joint. The latter is assigned in the database to those loans requested by two or more borrowers in a joint capacity, and may include matrital and or business partnerships.

Derived Race has 7 levels: White, Black/African American, Asian, Pacific Islander, 2 or more minorities, Native American, and Joint.

Results

We found that women, on average, pay 0.138 points and 0.086 more in interest rates than their joint and male counterparts respectively. No statistically significant changes were observed between joint borrowers and males.

When we turned our attention to race, we noticed that African Americans tended to pay higher interest rates compared to joint and Caucasian borrowers, although not other statistically significant difference was observed with any other minority group.

When we interacted the explanatory variables, we saw that African American women paid 0.26 points and 0.22 points more in interest rates than Caucasian Joint and Male borrowers respectively. While African American men paid 0.185 points higher interest rates than Caucasian joint borrowers.

For total loan costs, joint borrowers paid $1,033 and $1,832 more than males and females respectively. However, there were no statistically significant differences in the total loan costs paid between males and females. When interacting with race, we found that Caucasian joint borrowers paid more in TLC than single Asian and Caucasian borrowers.

For lender credits, we found no statistically significant differences between any of the race and sex groups. However, for discount points, all three sex groups showed statistically significant differences. Joint borrowers received $1289.59 less points compared to females but received $709.54 more points compared to males. Males received $580.05 points more than females. Finally, we also found that female Asians, female African Americans and female Caucasians received more discount points than joint Caucasian borrowers.

Due to the sheer number of interactions in our data, we have provided Tukey tables that are relevant and showed statistical significance.

Tukey tests. Tables of Sex as an explanatory variable

Tukey test. TLC vs Sex (3 levels)

Group1	Group2	meandiff (G2-G1)	p-adj	reject
Female	Joint	1832	0.001	True
Female	Male	799	0.067	False
Joint	Male	-1033	0.013	True

Tukey test. Interes rate vs Sex (3 levels)

Group1	Group2	meandiff (G2-G1)	p-adj	reject
Female	Joint	-0.138	0.001	True
Female	Male	-0.086	0.016	True
Joint	Male	0.052	0.234	False

Tukey test. Lender Credit vs Sex (3 levels)

Group1	Group2	meandiff (G2-G1)	p-adj	reject
Female	Joint	-272.58	0.09	False
Female	Male	-168.42	0.38	False
Joint	Male	104.16	0.68	False

Tukey test. Discount Points vs Sex (3 levels)

Group1	Group2	meandiff (G2-G1)	p-adj	reject
Female	Joint	1289.59	0.001	True
Female	Male	580.05	0.03	True
Joint	Male	-709.54	0.006	True

Tukey tests. Tables of Race as an explanatory variable.

Here, only Interest Rate was significant, and only in two of the groups.

Tukey test. Interest rate vs Race (7 levels)				
Group1	Group2	meandiff (G2-G1)	p-adj	reject
AfrAms	Joint	-0.223	0.03	True
AfrAms	White	-0.192	0.001	True

Tukey tests. Tables of the Interaction Sex:Race as explanatory variables.

Again, we only show those results with statistical significance.

Tukey test. Total Loan Cost vs Sex:Race (21 levels)				
Group1	Group2	meandiff (G2-G1)	p-adj	reject
Fem.Asian	Joint.Wht	3653.58	0.02	True
Fem.Wht	Joint.Wht	2126.12	0.004	True

Tukey test. Interest Rate vs Sex:Race (21 levels)				
Group1	Group2	meandiff (G2-G1)	p-adj	reject
Fem.Blk	Joint.Wht	-0.263	0.001	True
Fem.Blk	Male.Wht	-0.221	0.001	True
Joint.Wht	Male.Blk	0.185	0.02	True

Tukey test.Discount Point vs Sex:Race (21 levels)				
Group1	Group2	meandiff (G2-G1)	p-adj	reject
Fem.Asian	Joint.Wht	2264.2	0.03	True
Fem.Blk	Joint.Wht	1267.45	0.003	True
Fem.Wht	Joint.Wht	1214.19	0.02	True

Interpretation

Although we have found some differences between different groups, due to the limited nature of our data, we cannot conclusively state that these observed differences are due to unfair or ill-intentioned lending practices. Especially since we were unable to control for individual differences such as risk appetite, credit history, or income level. However, this is an interesting starting point for future analysis on the limitations on the access to loans and payment facilities some social groups have.

Limitations & Other Considerations

Some limitations of the proposed analysis are that we were not able to control for the effects of the property's physical attributes nor the effect of individual preferences on the response variables we analyzed. We were also limited by the number of community variables reported in the data source. In the future, it would be advisable to include additional variables that consider the livability of the area. For instance, is the property near a good school, is it safe with a low crime rate or is it a walkable neighborhood. Individual preferences, such as risk appetite, credit history, etc, should also be included to assess differences in the amount paid in interest rates and loan costs.

Another way our study could have benefitted was by conducting a cluster analysis. For instance, often those in the lower income groups tend to be minorities. In the future, it would be worthwhile to cluster some groups and analyze the interactions within those specific groups, particularly those that showed statistical significance in the differences between their loan costs and payment facilities.

Conclusion

This paper examined the effect of community factors on DC property values in the year 2020 and investigated whether women and minorities differ in how much they pay in interest rates and loan costs. Theoretically, we would expect that almost all the price variance for property values would come from the physical attributes of the property itself. However, we found that about 24% of the price variance in our data came from community factors. We also found instances of differences in how much women and minorities pay regarding mortgage loan costs and interest rates; however, we are not able to conclusively state that this difference is due to unfair lending practices.

References

Bhutta and Hizmo. “ Do Minorities Pay More for Mortgages?”. The Oxford Academy. academic.oup.com/rfs/article/34/2/763/5827007?login=true April/29/2020. Retrieved in April, 2022.

Federal Reserve of the United States. “Home Mortgage Disclosure Act Examination Procedures”. www.federalreserve.gov/boarddocs/caletters/2009/0910/09-10_attachment.pdf. Retrieved in April, 2022.

Howell and Korver-Glenn. “Race determines home values more today than it did in 1980”. Rice Kinder, Institute for Urban Research kinder.rice.edu/urbanedge/2020/09/24/housing-racial-disparities-race-still-determines-home-values-America Sep/24/2022. Retrieved in April, 2022

Lee and Bostic. “Bank adaptation to neighborhood change: Mortgage lending and the Community Reinvestment Act”. Journal of Urban Economics, Volume 116, March 2020, 103211. www.sciencedirect.com/science/article/pii/S0094119019300889 Retrieved in April, 2022.

Mathews, Chris. “Study Finds Women are Charged Higher Rates for Mortgages”. fortune.com/2016/09/08/study-finds-women-are-charged-higher-rates-for-mortgages/ Sep/8/2016. Retrieved in April, 2022.

Wichter, Zach. “Bankrate survey finds that nonwhite borrowers are more likely to have costlier mortgages”. www.bankrate.com/mortgages/survey-black-and-hispanic-mortgage-rates/ Nov/20/2020. Retrieved in April, 2022.