Pandemic-Exacerbated Redlining? An Investigation into Ohio Mortgage Approvals

Final Project

QMSS GR5015: Data Analysis in the Social Sciences

Cindy Chen, cjc2279

Introduction

The rise of automation in residential real estate is a double-edged sword. On one hand, it has boosted the popularity of online real estate marketplaces like Zillow, allowing millions of people to easily estimate home values for free. Meanwhile, the opacity of algorithms can make it more difficult to uncover insidious practices like redlining, wherein financial institutions restrict access to mortgages and favorable interest rates based on race. In August 2021, news outlet *The Markup* released an investigation where loan applicants of color were 40% to 80% more likely to be denied mortgages in 2019 compared to White applicants with similar credentials[1]. Their findings led them to conclude that there was evidence of redlining in ostensibly unbiased automated decision-making in mortgage approvals. In reading their report, I became interested in understanding how the pandemic affected racial disparities in mortgage approval rates for single-family homes, which exploded in demand during 2020 stay-at-home orders as people wanted more space.

Accordingly, I will investigate whether the pandemic has exacerbated differences in mortgage approval rates between non-White and White applicants in Ohio single-family housing neighborhoods between 2018 and 2020, to investigate trends in redlining. I constrained my analysis to the state of Ohio as several of its regions observed some of the highest municipal housing price growth in the US over the pandemic and it has a relatively even split between state-wide White vs minority populations.

Description of Data Set and Variables

This study's primary data will come from three separate files: the Home Mortgage Disclosure Act's (HMDA's) Dynamic National Loan-Level Data Set for the calendar years of 2018, 2019, and 2020. This data is the entire population of mortgage applications submitted in that specific year in the United States, a required disclosure by lenders as part of the Dodd-Frank Act. Accordingly, this is a complete and representative sample of US mortgage submissions. From the original 99 features in the data set, I narrow it down to 12 independent variables and one dependent variable; my reasoning is described in the subsequent section.

From the original data set, I filtered my data to Ohio single-family property applications and used only applications whose "Loan Purpose" was equal to 1 so that my subset pertains exclusively to mortgages for home purchases rather than other purposes such as re-financings or home improvement since my research question relates to redlining. Among Ohio single-family housing mortgage applications where I have approval/denial data and where I am not missing applicants' self-reported race, I have 161,644 data points remaining in the 2020 data set, 153,605 in the 2019 data set, and 342,482 in the 2018 data. However, I acknowledge that my scope and data quality issues pertaining to my intended analysis filters out about 65% of my original data, but it also makes it more manageable to process for this analysis.

It is crucial to acknowledge that a key variable is missing from this data set: applicants' credit scores. Credit scores are a widely used metric to determine how people access credit, but the

[1] Martinez, E. and Kirchner, L. (2021). "The Secret Bias Hidden in Mortgage-Approval Algorithms". *The Markup*.. https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms

confidential nature of this information means we cannot incorporate it into our analysis or accurately approximate it without making some strong assumptions.

As an aside, I had planned to constrict my analysis to Ohio neighborhoods with above-average property growth in 2020 by incorporating property value growth information from the American Community Survey (ACS), but 2020 ACS median property values are not yet available by county at the time of this paper so I have gone forward with the entire population of Ohio single-family homes.

On a different note, some notable outliers, likely driven by data quality issues, in the original data set were subsequently removed from this analysis. Where there were null values or codes that represented null values in my independent and dependent variables (as long as the variable did not pertain to an optional co-applicant), I decided to remove them from my analysis. Since my study involves the population of mortgage applications, I deemed the removal of applications with null values to be a reasonable reduction in my sample size. Data quality issues and concerning outliers were evident in the wide standard deviations that I noticed in my exploratory data analysis and unintuitive values (for instance, it is unlikely that people requested over 1,000 months for mortgage repayment).

To address data quality issues or extreme outliers, the following changes were made:

- Property value cannot be null and must not be larger than $10,000,000
- Loan term must be less than 999 months
- The combined loan to value ratio cannot exceed 1,000 times the value
- Applicant age cannot be greater than 100
- Income greater than $5 million
- Income less than $0

Overall, the removal of null values and outliers refined my data set to 423,080 data points. A next step outside the scope of this study would involve verifying that the excluded data points follow the population distribution, avoiding unintended bias in covariates.

Dependent Variable

1. **action_taken:** This variable indicates mortgage denial vs approval and originally has 8 possible values. I recoded the original values 1 (loan originated/approved) and 2 (approved, but not accepted) to the dummy variable of "1", indicating that the mortgage was approved. The original value "3" will be recoded as "0", indicating that the mortgage was denied. All other values will be filtered out of the data set, because they don't indicate a final approval outcome as the application was withdrawn, closed for incompleteness, or was a preapproval application.

   Among my remaining data points, my sample is imbalanced as 93% of applications were approved and only 7% were rejected **(Table 1)**. This makes sense as the mortgage application process is rigorous and applicants would only complete it if they were serious

about buying a home. Accordingly, people must have confidence in the strength of their application to even submit one.

Table 1. Distribution of Binary Dependent Variable "action_taken"

action_taken	Count
0 (rejected)	29,138 (7%)
1 (approved)	393,942 (93%)

Independent Variables

With 98 independent variables, this study incorporated only 12 of them. In part to manage the complexity of the analysis, they were also chosen by a general understanding of the financial considerations involved in mortgage applications as well as control variables that might inform any findings around discrimination.

1. **activity_year:** Used as a dummy control variable, this indicates the year (between 2018 to 2020) that the mortgage application was successfully and fully submitted. This is considered a categorical factor variable for my analysis. I hypothesize that there is no relationship between year and likelihood of mortgage approval, because there were no policy changes in the past three years that would alter mortgage approval decision-making.

2. **applicant_age:** This is a categorical variable indicating the age of the applicant with ranges such as "<25", "25-34", and ">74", which I will transform into an ordinal variable with the scale of 1 (the youngest age group) to 7 (the oldest age group). I hypothesize that older age groups have a higher likelihood of application approval as older applicants have greater accumulated wealth and income than younger applicants.

3. **applicant_race_1:** Self-reported categorical variable according to a list of 18 options. While the data set allows applicants to give five (5) different responses to race, I consider the first entry as the primary race, especially since it has the fewest null values. I recoded this variable into a dummy variable for White = 0 and non-White = 1. I assume that this will be a statistically significant dummy control variable and that non-White applicants are less likely to receive mortgage approval even though this is a legally protected trait.

4. **applicant_sex:** Categorical variable indicating the self-reported sex of the applicant. 1 (male) and 2 (female) will remain in the data set. However, the values of 3 and 4 were removed from the data set since it denotes that a sex was not disclosed; the value of 5 does not exist for this variable. The value of 6 (applicant selected both male and female on application) will be recoded as 3. I view this variable as a dummy control and hypothesize that it should have no relationship to mortgage approval outcomes since this is a protected trait.

5. **co_applicant_exists:** While my data set has co-applicant characteristics such as age, race, and sex, I decided to create a dummy variable to represent whether a co-applicant is on the application or not. Since the inclusion of a co-applicant is optional, I decided to use a

binary variable to bypass the k null data. I hypothesize that the existence of a co-applicant increases the likelihood of mortgage approval, since this improves key financial metrics like income and debt-to-income, ultimately lowering the applicants' lending risk.

6. **combined_loan_to_value_ratio:** The ratio of the total mortgage secured by the property compared to the value of the property. This variable incorporates the applicant's proposed down payment, which would reduce this ratio. Accordingly, I hypothesize that the relationship is negative: a lower combined loan-to-value ratio raises the likelihood of mortgage approval, that this will be highly statistically significant and the magnitude of change will be large.

7. **debt_to_income_ratio:** This (quasi-continuous) numeric variable measures an applicant's risk of defaulting on the mortgage and their ability to make debt repayments since it considers their monthly debt (with the mortgage included) to their monthly income. I hypothesize that a lower debt-to-income ratio should increase the likelihood of mortgage approval, and that this will be highly statistically significant.

   In terms of recoding, this variable is particularly complex as the data lists the actual debt-to-income ratio value if it falls between 36% and 50%; any other value is listed as a range such as "<20%" or "50% - 60%". I recoded this variable as an ordinal categorical variable where 36% to 42% inclusive is its own range, and 43% to 49% inclusive is another.

8. **ffiec_msa_md_median_family_income:** The median family income in the census tract where this property is located. I include this variable to help control for the "desirability" or wealth of a neighborhood and hypothesize that there is a positive relationship between a community's median family income and mortgage approval. Since affluent neighborhoods attract other affluent prospective buyers, these applicants likely have the favorable financial resources and profiles to move into these areas without mortgage application difficulties.

9. **income:** The gross annual income of the applicant and if applicable, combined with the co-applicant's gross annual income. This continuous numeric variable was log transformed due to the wide range in values. I hypothesize that as income increases, the likelihood of mortgage approval increases, this will be highly statistically significant, and the magnitude will be large.

10. **loan_amount:** This continuous numeric variable is the mortgage loan amount in US dollars requested in the application. This variable was log-transformed in my analysis. I hypothesize that the relationship is quadratic for loan amount, as extremely large loans may be seen as too risky, while very small loans likely stem from first-time or low-income buyers who do not have a credit history of takin. I believe this will be highly statistically significant with a large magnitude, because this is tied closely to risk, about which lenders are very concerned.

11. **property_value:** This continuous numeric variable is the value of the property in US dollars; it was log transformed in my analysis. I hypothesize that greater property values are more likely to be approved for a mortgage, because applicants attempting to purchase more valuable homes likely have sufficient existing assets and wealth and are trying to upgrade. Estimating a quadratic relationship, I hypothesis that this variable will be moderately statistically significant or may be insignificant.

12. **tract_to_msa_income_percentage:** This continuous numeric variable is the percentage difference in income between the specific census tract where the property is situated and the metropolitan statistical area (MSA). In other words, it's a proxy for how affluent or impoverished a community may be compared to the metropolitan area to which it belongs. I include this variable as a control and hypothesize that it will be statistically significant since my hypothesis involves the presence of redlining.

Hypothesis

The pandemic worsened non-White mortgage applicants' likelihood of mortgage approval for single-family homes in Ohio. This is because mortgage lenders were more discriminative in who joined coveted family neighborhood to help curate a certain community demographic.

Descriptive Summary

Table 2.

1) Continuous Variable Descriptive Statistics

Continuous Variable Metrics	Combined loan to value ratio	Loan Amount (w/o ln transform)	Loan Term (w/o ln transform)	Property Value (w/o ln transform)
Count	423,080	423,080	423,080	423,080
Mean	87.86%	$182,950	346 months	$218,108
Standard Deviation	13.49%	$122,102	50 months	$160,687
Min	0.20%	$5,000	1 month	$5,000
25%	80.00%	$105,000	360 months	$125,000
50%	95.00%	$155,000	360 months	$175,000
75%	96.50%	$235,000	360 months	$275,000
Max	197.48%	$4,505,000	720 months	$9,505,000
Skewness	-1.81	3.28	-3.43	5.32
Kurtosis	5.74	32.53	11.24	119.63

Continuous Variable Metrics	Tract Median Family Income	Tract to MSA Income Percentage	Income (w/o ln Transform)
Count	423,080	423,080	423,080
Mean	$72,062	112.82%	$91,653
Standard Deviation	$7,773	40.09%	$91,185

Min	$0	0.00%	$1,000
25%	$66,100	87.00%	$47,000
50%	$73,700	108.00%	$69,000
75%	$77,000	134.00%	$109,000
Max	$84,600	371.00%	$4,979,000
Skewness	-1.97	0.75	10.76
Kurtosis	15.55	1.69	312.94

2) Distribution of Select Categorical Variables

Variable	Yes	No
coapplicant_exists	260,402 *(62%)*	162,678 *(38%)*

Variable	White	Non-White
applicant_race	373,152 *(88%)*	49,928 *(12%)*

Applicant_age	Count	Proportion
Less than 25	34,418	*8%*
25-34	144,237	*34%*
35-44	101,084	*24%*
45-54	68,603	*16%*
55-64	46,297	*11%*
65-74	22,658	*5%*
>74	5,783	*1%*

Debt-to-Income Ratio	Count	Proportion
<20%	31,680	*7%*
20% - <30%	92,372	*22%*
30% - <36%	81,533	*19%*
36% – 43%	123,553	*29%*
44% - 49%	66,816	*16%*
50% - 60%	21,944	*5%*
>60%	5,182	*1%*

Table 3: Histogram of Combined Loan-to-Value Ratio

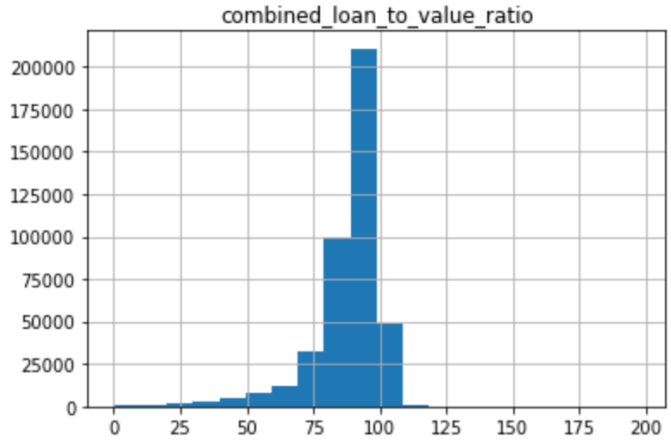

Table 4.

1) Correlation Matrix (Pearson)

	Loan to Value Ratio	Loan Amount (Ln)	Income (Ln)	Loan Term	Property Value (Ln)	Tract Median Family Income	Tract Income as % of MSA Income
Loan to Value Ratio	1.00	0.04	-0.19	0.28	-0.27	-0.06	-0.19
Loan Amount (Ln)	0.04	1.00	0.56	0.21	0.87	0.25	0.47
Income (Ln)	-0.19	0.56	1.00	-0.10	0.65	0.17	0.41
Loan Term	0.28	0.21	-0.10	1.00	0.01	0.03	-0.02
Property Value (Ln)	-0.27	0.87	0.65	0.01	1.00	0.29	0.55
Median Tract Family Income	-0.06	0.25	0.17	0.03	0.29	1.00	0.13
Tract Income as % of MSA Incomes	-0.19	0.47	0.41	-0.02	0.55	0.13	1.00

2) Variance Inflation Factor – Continuous Variables

Variable	VIF Value
income_ln	1.776
property_value_ln	7.782
loan_amount_ln	6.675
combined_loan_to_value_ratio	1.616
loan_term	1.230
ffiec_msa_md_median_family_income	1.093
tract_to_msa_income_percentage	1.441

These descriptive statistics offer valuable insights into my data set and help confirm my recoding and filtering decisions in the original data set.

In evaluating whether my data set makes sense following some data cleaning, the majority of proposed loan terms align with the common standard of 360 months (equal to 30-year mortgages) with the maximum term length at 60 years (which seems excessive but still possible). Additionally, the range between the first and third quartiles for the combined loan-to-value ratio (80% to 96%) illustrates a reasonable "sweet spot" that applicants try to target in their application submission; it implies that people plan to place a down-payment of 5% to 20% of the property value to secure the mortgage. The -1.81 skew in the loan-to-value ratio (illustrated in **Table 3 - 1**) means that applicants tend to apply for loans for more than the property is worth since the mean is already at 95%. Moreover, the mean income of $91,653. seems realistic for a single Ohioan or couple trying to buy a home and the mean single-family home price of $218,108 seems realistic for Ohio as a state average. The extremely high kurtosis among income, property value, and loan amount means that their distributions have long and fat tails, which makes sense as incomes and property values can vary astronomically due to income and wealth inequality; this prompted me to use natural log transformations for these three variables.

The distribution of this study's categorical and ordinal data also tells an intriguing story that guides this study (**Table 3 – 2**). Where possible, applicants try to maximize the likelihood of their loan approval by leveraging a dual-income household since more than half of applicants involve a co-applicant to minimize their risk of defaulting. Of particular interest, although the population of Ohio is about equal parts White and non-White, 88% of prospective home buyers were White, a disproportionate amount! This gives us context into the wealth disparities between racial groups in Ohio; as I look to find mediating or spurious variables in mortgage decision-making, this insight will acutely influence my investigation. Furthermore, 58% of prospective Ohio single-family home buyers are older Millennials and Gen X. These applicants may be starting families and seeking to upgrade to a single-family home or may be attracted to living in Ohio due to its reasonable housing affordability compared to other parts of the country, though uncovering these patterns is outside the scope of this analysis. In analyzing the distribution of debt-to-income ratios, the majority of applicants, if successful in their application, would endure monthly debt re-payments of less than 50% of their monthly gross income, which is likely an important metric for lenders.

Evaluating the Pearson correlation coefficients for my continuous numeric variables (**Table 4 - 1**), the log-transformed variables of property value and loan amount have a high correlation, given the coefficient of 0.87. This insight is strengthened by these two variables' high respective variance inflation factors (VIF) of 7.782 and 6.675 respectively, as calculated in my initial model from the subsequent section (**Table 4 – 2**). Since a VIF measures multicollinearity and a value of 2 or 3 is the typically acceptable range, this informs my modeling approach to drop one of the variables. The collinearity between these two variables is unsurprising since the loan amount is usually a direct proportion of the property purchase cost. While the property value variable also shares an above-average correlation with the affluence of the census tract (0.55) and log-transformed income (0.65), the VIF analysis indicates that it is acceptable for me to keep those variables in my models.

Initial Model – Linear Probability Model

My initial model was a linear probability model, selected for its ease of interpretability and helping me better understand my data and its best fit as a first step. To address the class imbalance in my dependent variable, I randomly under-sampled my data. I also used robust standard errors in my model to address heteroskedasticity in my data, and included interaction terms using *coapplicant_existence* on *applicant_age* and log-transformed *income*, since a parent might pose as a co-applicant to help young homebuyers with less wealth to qualify for a mortgage. While *loan_amount* was highly correlated with property value, I assume a quadratic relationship with the dependent variable, so I squared this variable in my analysis. Through some experimentation (see **Appendix - Table 10**), I noticed that the debt-to-income ratio ordinal variable exhibited a quadratic relationship due to the change in coefficient directions if we moved up one category, and transformed this variable accordingly.

To address my original research question on whether the pandemic-driven housing frenzy has exacerbated to rates of mortgage approval between non-White and White applicants, I created a new dummy variable called "pandemic" where 1 is equal to activity_year = 2020 and 0 is equal to all other years (2018 and 2019). Afterward, I created an interaction term between *applicant_race* and the pandemic variable, which will help me determine if the pandemic further lowered non-White applicants' probability of mortgage approval.

Table 5. Linear Probability Model - Results

OLS Regression Results

Dep. Variable:	action_taken			**R-Squared:**		0.153
Model:	OLS			**Adj. R-Squared:**		0.152
Method:	Least Squares			**F-Statistic**		583
No. Observations:	58,276					
	Coef	**Std Err**	**t**	**P>|z|**	**[0.25**	**0.975]**
Intercept	-0.2062	0.075	-2.759	0.006	-0.464	-0.170
C(coapplicant_exists)	-0.0084	0.028	-0.298	0.766	-0.064	0.047
C(applicant_sex)(Female)	0.0144	0.004	3.526	0.000***	0.006	0.022
C(applicant_sex)(F&M)	-0.0021	0.062	-0.035	0.972	-0.123	0.119
C(applicant_race_1)	-0.1065	0.007	-16.186	0.000***	0.119	-0.094
C(pandemic)	0.0138	0.003	5.452	0.000***	0.009	0.019
C(applicant_race_1):C(pandemic)	-0.0214	0.011	-1.951	0.051	-0.043	0.000
np.power(debt_to_income_ratio,2)	-0.0114	0.000	-70.910	0.000***	-0.012	-0.011
applicant_age	-0.0133	0.002	-5.840	0.000***	-0.018	-0.009
applicant_age:C(coapplicant_exists)	-0.0086	0.003	-3.039	0.002**	-0.014	-0.003
np.power(loan_amount_ln,2)	0.0031	0.000	9.564	0.000***	0.002	0.004
income_ln	-0.0219	0.006	-3.805	0.000***	-0.033	-0.011
income_ln:C(coapplicant_exists)	0.0052	0.006	0.838	0.402	-0.007	0.017
property_value_ln	0.0440	0.008	5.228	0.000***	0.027	0.060
combined_loan_to_value_ratio	-0.0028	0.000	-15.915	0.000***	-0.003	-0.002
loan_term	6.695e-5	4.15e-5	1.627	0.104	-1.38e-5	0.000
ffiec_msa_md_median_family_income	3.59e-6	1.91e-7	18.805	0.000***	3.22e-6	3.97e-6
tract_to_msa_income_percentage	0.004	5.64e-5	7.058	0.000***	0.000	0.001

**** statistically significant at 99.9% confidence level*
*** statistically significant at 99% confidence level*
** statistically significant at 95% confidence level*

Table 6. Partial F-test

Variable Inclusion	ANOVA P-Value
Year	4.84e-21
Sex	0.729
Debt-to-Income Ratio	0.000
Age	1.02e-315

The linear probability model (**Table 5**) yields an R-Squared of 0.153 and an Adjusted R-Squared of 0.152, which is lower than expected since we assume that mortgage decisions involve standardized criteria. Where variables were statistically significant, they were all significant with 99.9% confidence except for the interaction term between applicant age and the existence of a co-applicant, which was significant at the 99% confidence level. Meanwhile, *loan_term* (p-value of 0.104), the interaction term between income and co-applicant existence (p-value of 0.402), if applicant sex is male and female (p-value = 0.972), and whether a co-applicant exists at all (p-value = 0.776) were all materially statistical insignificant. My interaction term of particular interest between the pandemic and race variables was just shy of 95% confidence, so we deem it statistically insignificant.

The interaction between income and co-applicant existence as a statistically insignificant independent variable was a surprising find, since I suspected that a co-applicant like a parent might help improve the probability of approval for young home buyers. Nonetheless, the other interaction term between applicant age and co-applicant existence may capture most of the relationship this relationship when we include both interactions in this analysis.

The variable with the greatest magnitude is applicant race, where non-White applicants are 10.65 percentage points less likely to receive mortgage approval than White applicants, if we hold all other variables constant. This is unsurprising given past reports of racial discrimination in mortgage approvals, though the finding is still striking. This magnitude is followed by the natural log of property value, where a percentage point increase in the value makes someone 4.40 percentage points less likely to get approved for a mortgage if we control for everything else. Likewise, the coefficient of -0.0214 for the interaction between race and the pandemic deduces that the likelihood of mortgage approval worsened by 2.14 percentage points, even if the relationship is statistically insignificant.

Notably, income has a negative relationship with the percentage point likelihood of mortgage approval: holding everything else constant, a 1% increase in income makes it 2.19 percentage points less likely that someone will receive mortgage approval. With the inclusion of a co-applicant, a 1% increase in income only increases the chances by 0.51 percentage points, which is almost no change. This might be explained by the fact that people with higher incomes tend to live beyond their means and might be more likely to buy a home that suits their aspirational lifestyle.

Since many of my categorical variables like applicant sex and debt-to-income ratio involved multiple categories, I ran an ANOVA partial F-test to understand whether these dummy variables led to a statistically significant increase in my model's fit through its R-Squared value. In **Table 6,** all the multi-category dummy variables improved model fit except the applicant sex.

I subsequently ran a factor analysis **(Table 7)** as part of my data exploration. The motivation for running a factor analysis stemmed from understanding general factors driving applicants' creditworthiness.

My first factor, PA1, appears to be what we generally think standardized mortgage applications evaluate for decision-making: financial metrics that do not consider protected classes like gender or race. However, the second factor PA3 incorporates protected classes and most of the same financial metrics. This suggests that there is a factor underlying both protected classes and people's financial status; this factor may be categorized as the known racial disparities in intergenerational wealth. This factor analysis illustrates that the HMDA data set lacks another variable beyond credit score that helps us isolate the effect of applicant race: wealth. The value of existing assets outside of income like investments and property is not explicitly captured in the mortgage disclosures, and likely correlates with race.

Table 7. Factor Analysis Loadings

	PA1	**PA2**	**PA3**
applicant_sex			-0.160
applicant_age		-0.306	
debt_to_income_ratio	0.134		-0.482
combined_loan_to_value_ratio	-0.111	0.742	-0.105
applicant_race_1			-0.151
loan_amount_ln	0.918	0.246	0.155
income_ln	0.436		0.799
loan_term	0.141	0.355	-0.131
property_value_ln	0.933	-0.140	0.207
ffiec_msa_md_median_family_income	0.261		
tract_to_msa_income_percentage	0.505	-0.155	0.180
coapplicant_exists	-0.222		-0.300
pandemic	0.105		

Initial Models Commentary

While linear probability models are easy to run and interpret, and enable additional analyses like partial F-tests, it is not the best model for my study. Since LPMs are not constrained to a range of 0 (mortgage denied) to 1 (mortgage approved), I might calculate a probability outside this range even when input values are realistic. Interpreting the p-values might also be moot since the binary nature of *action_taken* means that we violate the condition of normality in our errors. In all, linear probability models are not an ideal model for understanding the relationship between the various independent variables and whether a mortgage application was approved/denied.

In addition, the factor analysis was insightful, but it is merely exploratory and does not directly aid me in addressing my research question.

Final Model – Logit Model

To address the data issues raised above, I first perform a logit model on my entire rebalanced data set using the transformations and interactions outlined in my initial models, maintaining my robust standard errors. The findings from my initial model prompted me to remove the interaction term between co-applicant existence and applicant income. Since a logit model works best for binary outcomes, containing the dependent variable within the 0 to 1 range, it is the ideal model for my analysis.

To avoid multicollinearity between my pandemic variable and *activity_year*, thus following the assumptions required by logit models, I removed activity_year from the logit model.

Table 8. Logit Regression with Interaction Variable & Curvilinear Variables

Logit Regression 3 with Interaction Variable Results

Dep. Variable:	action_taken	**No. Observations:**	58,276
Model:	Logit	**Log-Likelihood:**	-35559
Method:	MLE	**Pseudo R-Squared:**	0.1197
Covariance Type	Nonrobust	**LLR p-value**	0.000

	Coef	**Std Err**	**z**	**P>|z|**	**[0.25**	**0.975]**
Intercept	-3.5248	0.351	-10.029	0.000***	-4.214	-2.836
C(coapplicant_exists)	0.0782	0.047	1.656	0.098	-0.014	0.171
C(applicant_sex)(Female)	0.0692	0.019	3.593	0.001**	0.031	0.107
C(applicant_sex)(F&M)	-0.0596	0.288	-0.207	0.836	-0.623	0.504
C(applicant_race_1)	-0.5229	0.032	-16.335	0.00***	-0.586	-0.460
C(pandemic)	0.0216	0.021	1.021	0.307	-0.020	0.063
C(applicant_race_1):C(pandemic)	-0.0901	0.053	-1.700	0.089	-0.194	0.014
np.power(debt_to_income_ratio,2)	-0.0545	0.001	-65.118	0.00***	-0.056	-0.053
applicant_age	-0.0602	0.011	-5.610	0.00***	-0.081	-0.039
applicant_age:C(coapplicant_exists)	-0.0422	0.013	-3.151	0.002**	-0.068	-0.016
np.power(loan_amount_ln, 2)	0.0149	0.002	9.354	0.000***	0.012	0.018
income_ln	-0.0894	0.019	-4.700	0.000***	-0.127	-0.052
property_value_ln	0.1956	0.041	4.756	0.000***	0.115	0.276
combined_loan_to_value_ratio	-0.0130	0.001	-15.243	0.000***	-0.015	-0.011
loan_term	0.0004	0.000	1.774	0.076	-3.69e-5	0.001
ffiec_msa_md_median_family_income	2.073e-5	1.05e-6	19.809	0.000***	1.87e-5	2.28e-5
tract_to_msa_income_percentage	0.0021	0.000	7.647	0.000***	0.002	0.003

Table 9. Logit Regression Odds Ratios - Ranked

Variable	**Odds Ratio**
Property Value (ln)	1.215987
Co-Applicant Exists	1.081310
Applicant Sex (Female)	1.071603
the Loan Amount^2	1.014987
Pandemic	1.021820
Tract to MSA Income Percentage	1.002079
Loan Term	1.000353
FFIEC MSA MD Median Family Income	1.000021
Combined Loan to Value Ratio	0.987076
Applicant Sex (F&M)	0.942176
Applicant Age * Co-applicant Exists	0.958690
Debt to Income Ratio^2	0.947007
Applicant Age	0.941548
Applicant Race * Pandemic	0.913807

Income (ln)	0.914462
Applicant Race	0.592807

In my first logit model's logistic regression summary (**Table 8**), many of my variables were statistically significant at the 99.9% confidence level except for the loan term (p-value = 0.076), whether it was during the pandemic (p-values = 0.307), the applicant identified as both male and female (0.836), and whether a co-applicant exists on the application (0.098). The rigorously high confidence level of my statistically significant variables minimizes the possibility of false discovery. However, my variable of high interest, indicating whether the pandemic affected race, had a p-value of 0.089 and was thus statistically insignificant. This means that I cannot reject my null hypothesis that the pandemic did not worsen non-White applicants' likelihood of mortgage approval in Ohio.

The pseudo-R-squared value of 11.97% still seems low as I would expect more of the variance to be explained in data that should reflect standardized decision-making. However, the low value could suggest that mortgage approval decision-making is not as standardized as I thought (or it could be shortcomings in my model, of course). While the pseudo-R-squared is lower in the logit model than the R-Squared in the LPM, I am not concerned with this difference as I dropped an interaction term in this new model and the values are not an exact comparison.

In translating the logits to odds ratios (**Table 9**), the applicant's race alarmingly had the highest probability for a successful mortgage application: net of all other factors like income and debt/loan metrics, being a non-White applicant will lead to a 40.7% decrease in odds of mortgage approval. What will raise someone's chances of mortgage approval is the desired property's value: a 1% increase in property value will increase one's likelihood of mortgage approval by 21.6% when we control for all other variables. When I originally hypothesized that debt-to-income and loan-to-value ratios would have high magnitudes of change, their influence was much smaller than estimated, especially compared to the magnitudes of race and property value. Holding all else constant, one-category higher in the squared debt-to-income ratio scale leads to a 5.3% decline in approval probability. In addition, one percentage-point rise in the combined loan-to-value ratio leads to a 1.3% increase in the likelihood of approval, controlling for all else. Of note, the magnitude and high statistical significance of applicant sex seem oddly high. The finding that being female rather than male as the primary applicant raises the likelihood of approval by 7.2%, holding everything else constant, seems very high for a characteristic that I would assume has no bearing on mortgage approval or the opposite relationship, where being male might increase one's chances of approval.

For the purposes of my analysis, the logit model is the ideal approach given the binary dependent variable. Since my descriptive analysis indicated that my independent variables were not all normally distributed with equal variance, the logit model is the right choice as it accommodates these conditions. Of course, this model still faces limitations in its low pseudo-R-squared, prompting us to consider other ways to generate a better fit between our data and the model. Omitted variable bias also exists in terms of missing variables on wealth and credit score that help us better isolate the effect of race on mortgage approval; the effect of protected classes like race is likely confounded by the underlying factor of wealth. There is also an opportunity to reconsider

how certain variables like debt-to-income ratios are recoded or rescaled as the HMDA survey is quite complex with its choices in responses.

Conclusion

While I could not reject my null hypothesis that the pandemic exacerbated non-White applicants' likelihood of mortgage approval, all the models and the data point to the sobering fact that racial disparities exist in supposedly standardized and automated mortgage approval decision-making. Perhaps part of this discrepancy is due to the lack of information on applicants' credit scores and existing wealth, but even after controlling for factors that would approximate it like debt-to-income ratios, the applicant's race has the greatest effect on an applicant's approval odds when we control for all other factors.

The complexities of people's motivations in submitting a mortgage application as well as the inaccessibility of key information means there are vast opportunities for further study of the HMDA data set. In terms of the next steps, I aim to incorporate 2020 median property growth data (once the ACS releases it) so that I can identify areas with above-average property growth. In turn, I can investigate whether my hypothesized phenomenon of pandemic-exacerbated racial disparities in mortgage approval might be more pronounced in counties that were more desirable to move into, rather than looking at Ohio as a whole.

Appendix

Table 10. OLS Regression with Debt-to-Income Ratio as a Categorical Variable

OLS Regression Results

Dep. Variable:	action_taken			**R-Squared:**	0.173	
Model:	OLS			**Adj. R-Squared:**	0.173	
Method:	Least Squares			**F-Statistic:**	554	
No. Observations:	58,276			**AIC:**	7.356e+04	
	Coef	**Std Err**	**t**	**P>|z|**	**[0.25**	**0.975]**
Intercept	-0.1740	0.074	-2.352	0.019*	-0.319	-0.029
C(debt_to_income_ratio)(20-<30%)	0.0625	0.009	7.312	0.000***	0.046	0.079
C(debt_to_income_ratio)(30-<36%)	0.0464	0.009	5.278	0.000***	0.029	0.064
C(debt_to_income_ratio)(36-43%)	0.0362	0.008	4.275	0.000***	0.020	0.053
C(debt_to_income_ratio)(44-49%)	-0.0382	0.009	-4.195	0.000***	-0.056	-0.020
C(debt_to_income_ratio)(50-60%)	-0.2820	0.010	-28.144	0.000***	-0.302	-0.262
C(debt_to_income_ratio)(>60%)	-0.5195	0.011	-47.188	0.000***	-0.541	-0.498
C(activity_year)(2019)	0.0409	0.005	8.783	0.000***	0.032	0.050
C(activity_year)(2020)	0.0165	0.003	6.567	0.000***	0.021	0.021
C(coapplicant_exists)	-0.0105	0.028	-0.377	0.707	-0.065	0.044
C(applicant_sex)(Female)	0.0110	0.004	2.740	0.006**	0.003	0.019
C(applicant_sex)(F&M)	0.0067	0.061	0.110	0.913	-0.113	0.126
C(applicant_race_1)	0.1029	0.007	-15.816	0.000***	-0.116	-0.090
C(pandemic)	0.0165	0.003	6.567	0.000***	0.012	0.021
C(applicant_race_1):C(pandemic)	-0.0244	0.011	-2.251	0.024*	-0.046	-0.003
applicant_age	-0.0154	0.002	-6.869	0.000***	-0.020	-0.011
applicant_age:C(coapplicant_exists)	-0.0054	0.003	-1.919	0.055	-0.011	-0.000
np.power(loan_amount_ln,2)	0.0030	0.000	9.424	0.000***	0.002	0.004
income_ln	-0.0106	0.006	-1.846	0.065	-0.022	0.001
income_ln:C(coapplicant_exists)	0.0024	0.006	0.393	0.695	-0.010	0.014
property_value_ln	0.0344	0.008	4.141	0.000***	0.018	0.051
combined_loan_to_value_ratio	-0.0034	0.000	-19.571	0.000***	-0.004	-0.003
loan_term	-5.443e-6	4.11e-5	-0.133	0.895	-8.65e-5	7.5e-5
ffiec_msa_md_median_family_income	3.298e-6	1.89e-7	17.443	0.000***	2.92e-6	3.67e-6
tract_to_msa_income_percentage	0.004	5.57e-5	7.220	0.000***	0.000	0.001