Determinants of Homeownership: A Statistical Analysis

BDA 610: ADVANCED BUSINESS STATISTICS PROJECT
DR. ARNAB NAYAK

SUBMITTED BY: GROUP (4)
MYAT AYE
ROCK CHEN
SHARMEEN RAHIMANI
SAMUEL SUBHAKAR

Table of Contents

1. Introduction **2**

2. Research Question **2**

3. Literature Review **3**

4. Data **4**

4.1 Data Source & Data Dictionary *4*

4.2 Data Limitations & Challenges *4*

4.3 Data Cleaning *4*

4.4 Data Dimension Creation & Reduction *5*

5. Exploratory Data Analysis **6**

6. Methodology **7**

6.1 Logistic Regression Model *7*

6.2 Interpretation Results of Final Model (4) *7*

6.3 Models Comparison *8*

7. Results **9**

7.1 Conclusion *9*

8. References **10**

9. Appendix **11**

1. Introduction

Homeownership is widely recognized as a key measure of financial stability and long-term wealth accumulation. It plays a crucial role in economic security, providing individuals and families with financial benefits such as equity building and housing stability. Given its significance, economists, policymakers, and financial institutions have extensively studied the factors that influence homeownership, focusing primarily on financial and demographic determinants. Key variables such as income levels, employment type, age, education, and marital status all contribute to the likelihood of homeownership. Understanding these factors offers valuable insights into housing market trends and the challenges faced by different socioeconomic groups in attaining homeownership.

This study aims to understand the probability of homeownership by analyzing financial and demographic determinants. Utilizing research and data from **IPUMS USA**, we examine how personal income, family income, education, employment type (wage earners vs. self-employed), and life-stage factors (such as age, marital status, and family size) impact an individual's likelihood of owning a home. Given ongoing shifts in income distribution, economic policies, credit markets and limitation of house prices and sensitive credit data, this analysis is particularly relevant in assessing disparities in homeownership and identifying areas where policy interventions could enhance access to homeownership opportunities.

By exploring these factors, this research seeks to provide a well-rounded understanding of both the barriers and enabling conditions for homeownership. The findings can support policymakers in designing housing programs that promote affordability, help financial institutions refine mortgage lending strategies, and contribute to broader discussions on economic mobility and ensuring that homeownership remains a viable path to financial stability for a diverse range of individuals.

2. Research Question

The primary objective of this study is to understand the probability of homeownership based on financial and demographic factors. Specifically, this research seeks to:

1. Analyze Financial Determinants – personal income, family income
2. Examine Demographic Influences – investigate how age, gender, marital status, education, affect homeownership
3. Compare Employment Types – identify differences in homeownership probability between self-employed individuals and wage earners, considering income stability and lending barriers.

4. Policy and Financial Implications – Provide recommendations for policymakers and financial institutions to improve homeownership accessibility, particularly for underrepresented and lower-income groups.

By addressing these objectives, this study aims to enhance the understanding of homeownership trends and contribute to more effective housing policies and financial lending strategies.

3. Literature Review

Income is one of the strongest predictors of homeownership. Moore (1991) used data from the American Housing Survey (AHS) to develop a logit model assessing the probability of homeownership and found that higher income significantly increases the likelihood of owning a home. Higher earnings not only provide greater purchasing power but also enhance financial stability, making homeownership more attainable. However, Moore (1991) also noted that affordability challenges disproportionately affect lower-income groups, limiting their ability to transition from renting to homeownership.

Age is a key factor influencing homeownership. Moore (1991) found that homeownership rates tend to increase with age, as individuals accumulate savings and improve their creditworthiness. However, Khorunzhina and Miller (2022) noted that economic shifts have led to a rise in the average age of first-time homebuyers, with many delaying home purchases until their 30s. This delay is often attributed to factors such as student loan debt, unstable job markets, and housing affordability constraints.

Marital status also plays a significant role in homeownership probability. Married couples are more likely to own homes than single individuals, primarily due to combined financial resources and greater economic stability. Khorunzhina and Miller (2022) found that declining marriage rates have contributed to lower homeownership rates, as single individuals often face greater financial constraints when purchasing a home. Moore (1991) similarly observed that larger households, particularly those with children, tend to prioritize homeownership as a means of securing long-term housing stability.

Moore (1991) found that household income plays a vital role in predicting homeownership probability . Wage earners generally benefit from steady earnings, making it easier to qualify for mortgage loans, whereas self-employed individuals often face greater scrutiny due to variable income streams. According to Goodman and Mayer (2018), access to mortgage financing remains one of the most significant barriers to homeownership, particularly for individuals without a stable income history.

Education is another key determinant of homeownership. Goodman and Mayer (2018) found that individuals with higher education levels, particularly those with college degrees, are more likely to own homes. This is largely due to the increased earning potential and job stability associated with higher educational attainment. Additionally, occupational stability plays a role in mortgage approval, as lenders favor borrowers with steady employment and consistent income.

4. Data

4.1 Data Source & Data Dictionary

This dataset is derived from the 2022 American Community Survey (ACS) microdata provided by IPUMS USA, a project of the Minnesota Population Center at the University of Minnesota (https://usa.ipums.org/usa/index.shtml). It includes 6,779,187 observations with 18 variables covering household and individual-level characteristics across demographic, social, economic, housing, and migration categories.

4.2 Data Limitations & Challenges

Significant outliers are present in key variables—including personal income, family income, age, family size, and number of children—that can heavily influence data analysis and lead to potential misinterpretations and biases. Additionally, the dataset contains missing values and placeholder entries (e.g., N/A, 9999999) for essential variables, which may compromise the accuracy of the results. Moreover, important contextual variables, such as location (urban/rural), occupation, and industry, are not included, thereby limiting the analytical depth. Finally, crucial financial factors that influence access to home ownership, such as housing prices, personal credit scores, and other outstanding loans or mortgages, are absent from the dataset.

4.3 Data Cleaning

The data preparation process began by removing irrelevant variables for the analysis. Variables such as YEAR", "SAMPLE", "SERIAL", "PERNUM", "OWNERSHPD", "IND", "MIGRATE1", "MIGRATE1D", and "MORTGAGE" were excluded as they were not important for the analysis.

Next, the values of 9999999 in the "INCTOT" and "FTOTINC" columns were replaced with NA, as they were considered indicative of missing or irrelevant data. To address outliers, the Interquartile Range (IQR) method was applied. A custom function was used to calculate the first (Q1) and third (Q3) quartiles for the columns "INCTOT", "FTOTINC", "AGE", "FAMSIZE", and "NCHILD". Values outside the range defined by Q1 - 1.5 * IQR and Q3 + 1.5 * IQR were filtered out, effectively mitigating the influence of extreme values. Following outlier removal, the dataset was further refined by filtering the rows based on specific criteria. Only those records where "INCTOT" and "FTOTINC" were greater than or equal to 0 were retained.

Additionally, the "AGE" column was filtered to include only values greater than 22 and less than or equal to 67. This careful filtering ensured that the dataset was both clean and relevant for further analysis. Subsequently, a multicollinearity check was performed on key variables such as FAMSIZE and NCHILD, and INCTOT and FTOTINC to identify and address any potential issues arising from highly correlated predictors. Finally, after all the cleaning and preprocessing steps were completed, the dataset was refined to a final number of observations, reducing the initial count from 6,779,187 to 2,985,921. This final, clean dataset was then ready for further detailed analysis.

4.4 Data Dimension Creation & Reduction

A binary variable, "Homeowner," was generated, where 1 represents ownership and 0 represents rental status, with missing values assigned as NA. The "Gender" variable was also created, assigning 1 to male, 0 to female, and NA to any other or missing values. The "NADULT" variable was calculated by subtracting the number of children ("NCHILD") from the total family size ("FAMSIZE"), representing the number of adults in the household.

Marital status was grouped into a binary variable, "Married_binary," with 1 for married individuals and 0 for unmarried individuals.

The "race_group" variable was formed by categorizing respondents into major racial groups, where 1 represents White, 2 represents Black/African American, 3 represents Asian (including Chinese, Japanese, and other Asian or Pacific Islander), and 4 represents other races (including American Indian/Alaska Native, Other race, two major races, or three or more major races).

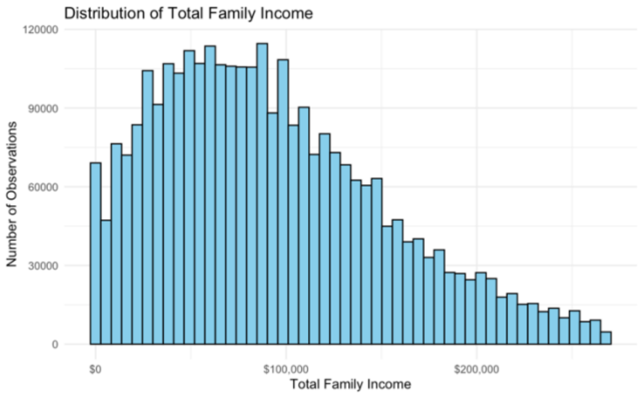
Education levels were grouped into the "Educ_group" variable, where 1 represents less than high school, 2 represents high school graduate, 3 represents some college, and 4 represents college graduate or higher.

A binary "Work_from_home" variable was created to indicate whether an individual worked from home, with 1 for those who worked from home and 0 for those who did not, with missing values marked as NA. Lastly, the "Classwkr_binary" variable was developed to distinguish between self-employed individuals (1) and wage workers (0), with missing values assigned as NA.

The creation of these binary variables simplifies the analysis, allowing for clearer distinctions between groups and enabling easier statistical modeling, such as logistic regression.

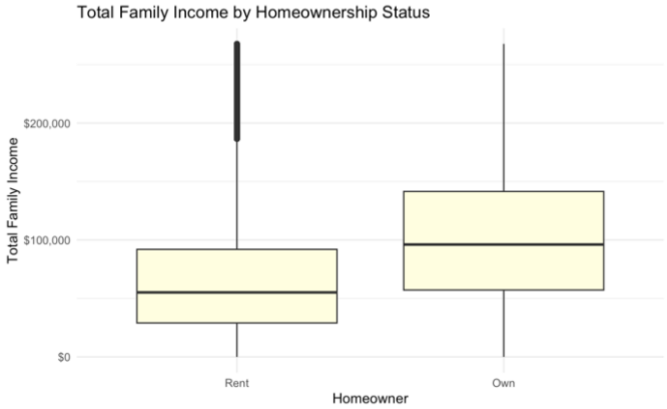
5. Exploratory Data Analysis

Distribution of Total Family Income



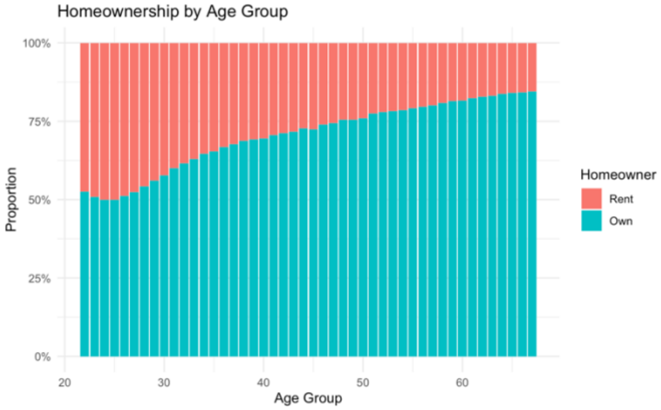
The family income histogram displays a right-skewed pattern, suggesting that although most families fall within lower income brackets, a few families have higher incomes.

Total Family Income by Homeownership



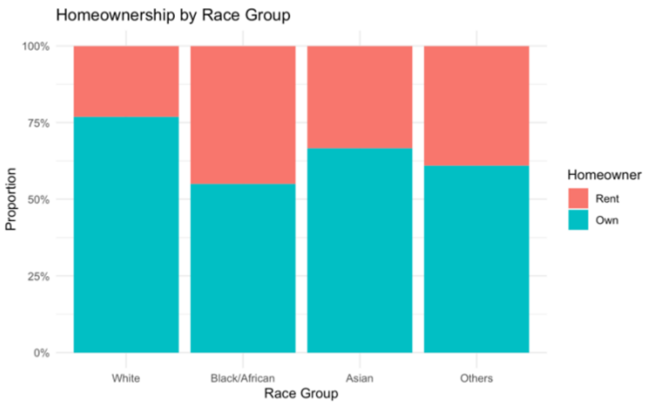
Families who own their homes generally have higher total family incomes, as shown by the higher median and smaller spread, whereas renters have a lower median income and a broader range, with a notable outlier at the high end.

Homeownership by Age Group



Older age groups exhibit a higher count of homeowners than younger ones, suggesting that homeownership rates increase with age in the sample.

Homeownership by Race Group



The chart illustrates variation in homeownership across race groups. White and Asian groups appear to have a higher proportion of owners, while Black/African and Other groups show a relatively larger share of renters.

6. Methodology

6.1 Logistic Regression Model

The methodology employed to answer our research question is logistic regression—a statistical technique designed to model the relationship between a binary dependent variable and one or more independent variables. Unlike linear regression, which is used for predicting continuous outcomes, logistic regression estimates the probability that a given observation belongs to one of two categories. In this analysis, the independent variable "homeownership" is categorical, with the two outcomes defined as 1 for "Own" and 2 for "Rent," making logistic regression an ideal approach for modeling homeownership status.

$$\log\big(Odds\ Ratio\ (yi = 1)\big) = \beta 0 + \beta 1x1i + \beta 2x2i + \cdots + \beta kxki + ei$$

6.2 Interpretation Results of Final Model (4)

- No. of child: Each additional child increases the log-odds of owning a home. There is about a 12% increase in odds of homeownership per additional child.
- No. of Adult: Each additional adult in the household increases the log-odds of ownership – 46% higher odds per extra adult.
- Age: Older individuals are more likely to own a home, up to a point – about 6% higher odd for one year older.
- Age squared: At a certain point of age, there will be -0.014% decrease for each one-unit increase in age squared.
- Married: Being married (vs. not married) increases the log-odds of homeownership – about 40% higher odds if married.

- Races: Black/African, Asian and other races have lower log-odds of homeownership compared to the reference group (white) – about 43%, 53%, 50% lower odds respectively.
- Education: Higher education groups are associated with greater odds of homeownership relative to the reference group (no high school). Colleges and higher education group has lower odds compared to high school and some colleges graduate potentially due to student loans or educational expenses.
- Class of work: Self-employed group are more likely to own a house compared to those who work for wages – about 19% higher odds.
- lFTOTINC: Family Income has very strong positive effect, ss (log) family total income increases by 1 unit, log-odds increase by 0.6146 – ~85% higher odds per 1-unit increase in log income.
- Married * Education: Being married and in a higher education group has an additional positive effect on the log-odds of owning a home – 32% higher odds compared to the reference group, married with no high school.

The model demonstrates a high sensitivity, correctly identifying approximately 91.6% of all actual positive cases—those classified as homeowners. However, its specificity is considerably lower, with only about 37.7% of actual negative cases accurately recognized. Overall, the model correctly classifies roughly 75.97% of all observations, reflecting a mixed performance that excels in detecting positives while underperforming in identifying negatives.

The AUC is 0.7762, meaning there is about a 77.62% chance that the model will correctly rank a positive observation above a negative observation. Generally, higher AUC values (closer to 1.0) indicate better model performance.

6.3 Models Comparison

	Model 1	Model 2	Model 3	Model 4
Pseudo R-Square	0.1509	0.1680	0.1681	0.1685
Likelihood ratio test	***	***	***	***
AUC	0.761	0.7762	0.7762	0.7762

Several models were tested with different variables and most of variables are significant due to large sample observations. All the models have overall significant model with equally AUC. However, model 4 has highest Pseudo R-Square as it explains slightly more variance and includes potentially important covariates.

7. Results

7.1 Conclusion

This study tested several logistic regression models to identify the key factors influencing homeownership, including household composition (number of children and adults), age (with a quadratic term), marital status, race, education, employment class, and family income. Overall, older individuals (up to a certain point), those with more household members, married individuals, higher education levels, self-employed workers, and higher family incomes all have higher odds of owning a home. In contrast, belonging to certain race categories (compared to white) decreases the odds of homeownership.

From a business perspective, especially for financial institutions and policymakers, these results can guide more precise lending strategies and policy design. Lenders could utilize the model's predictive power (AUC ~0.7762) to better assess potential borrowers, customizing financial products to help demographic segments less likely to own homes. With high sensitivity (~91.6%), the model reliably flags most potential homeowners, though its lower specificity (~37.7%) suggests further refinements to avoid overestimating homeownership likelihood among non-owners. Meanwhile, policymakers can deploy these findings to create or enhance housing initiatives—such as education and income support programs, first-time buyer incentives, and community outreach—that address identified barriers like race-based disparities or lower educational attainment.

Key findings:

1. Influential Factors: Demographic and socioeconomic variables, particularly income, education, marital status, and race, play substantial roles in determining homeownership odds.
2. Policy Implications: There is a clear need for targeted housing policies that balance economic goals with social equity, focusing on segments of the population facing structural disadvantages.
3. Strategic Lending: Financial institutions can leverage the model to refine credit risk assessments and develop lending solutions tailored to underserved groups, ultimately fostering more inclusive homeownership opportunities.

By integrating these insights into housing policy and lending practices, stakeholders can better align economic growth with social welfare, creating a more equitable and sustainable housing market.

8. References

1. Moore, D. J. (1991). Homeownership affordability series: Forecasting the probability of homeownership: A cross-sectional regression analysis. *Journal of Housing Research, 2*(2), 125–143. Retrieved from https://www.jstor.org/stable/24825921

2. Goodman, L. S., & Mayer, C. (2018). Homeownership and the American dream. *Journal of Economic Perspectives*, *32*(1), 31–58. https://doi.org/10.1257/jep.32.1.31

3. Khorunzhina, N., & Miller, R. A. (2022). American dream delayed: Shifting determinants of homeownership. *International Economic Review*. https://doi.org/10.1111/iere.12557

4. Tharp, D. T., Seay, M., Stueve, C., & Anderson, S. (2020). Financial satisfaction and homeownership. *Journal of Family and Economic Issues, 41*, 255–280. https://doi.org/10.1007/s10834-019-09652-0

5. Data: https://usa.ipums.org/usa/index.shtml

9. Appendix

1. Data Dictionary

Variables	Data Type	Definition
YEAR	Int	Census Year 2022, 2023
SAMPLE	Int	IPUMS sample identifier
SERIAL	Int	Household serial number
OWNERSHP	Factor	Ownership of dwelling (tenure)
MORTGAGE	Factor	Mortgage status
FAMSIZE	Int	Number of own family members in household
TRANWORK	Factor	Means of transportation to work
MIGRATE1	Factor	Migration status 1 year ago
FTOTINC	Int	Total family income
INCTOT	Int	Total personal income
CLASSWKR	Factor	Class of worker
EDUC	Int	Education
RACE	Factor	Race
MARST	Factor	Marital status
NCHILD	Int	Number of own children in the household
IND	Factor	Industry
SEX	Factor	Sex
AGE	Int	Age

2. Summary Statistics

	FAM SIZE	N CHILD	AGE	INCTO T	FTOT INC	N ADULT	LFTOT INC	LINC TOT
Min	1.000	0.0000	22.00	0	0	1.00	0.00	0.000
1st Qu.	2.000	0.0000	33.00	15000	46000	2.00	10.74	9.616
Median	2.000	0.0000	47.00	37400	83500	2.00	11.33	10.529
Mean	2.592	0.5217	45.93	43123	92516	2.07	11.02	9.468
3rd Qu.	3.000	1.0000	59.00	65000	130000	2.00	11.78	11.082
Max	7.000	2.0000	67.00	145000	267710	7.00	12.50	11.884

Table 1: Summary Statistics for Numeric Variables

Level	HOME OWNER	GENDER	MARRIED_BINARY	RACE_ GROUP	EDUC_ GROUP	WORK FROM HOME	CLASSWKR_BINARY
0	854,253	1,566,805	1,375,201	–	–	1,824,318	2,312,684
1	2,131,668	1,419,116	1,610,720	1,991,360	192,884	287,073	256,493
2	–	–	–	266,502	1,075,156	–	–
3	–	–	–	189,777	689,121	–	–
4	–	–	–	538,282	1,028,760	–	–
NA	–	–	–	–	–	874,530	416,744

Table 2: Frequency Counts for Factor Variables

Note: this is the summary statistics of final clean dataset, there are remaining NA in Work_from_home and Classwkr_binary which are removed in the final model.

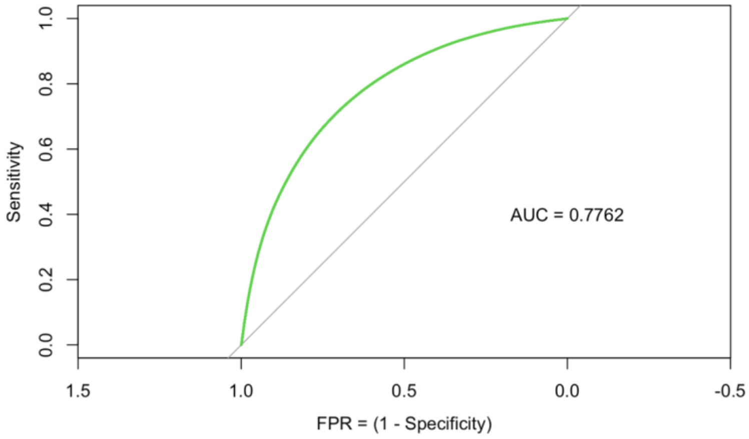
3. Models Comparison

Variable	**Model 1 Coeff.**	**Model 2 Coeff.**	**Model 3 Coeff.**	**Model 4 Coeff.**
Constant	-3.4982***	-9.3604***	-9.5881***	-9.4881***
NCHILD	0.5815***			
NADULT	-0.4473***	0.1278***	0.1193***	0.1149***
AGE		0.3749	0.3788***	0.375***
I(AGE2)	0.048***	0.0459***	0.059***	0.058***
Married			-0.0002***	-0.0001***
Black/African	0.6276***	0.47***	0.4672***	0.3334***
Asian	-0.9329***	-0.8443***	-0.8469***	-0.8476***
Race_ Others	-0.6633***	-0.6438***	-0.6466***	-0.6467***
Educ 2 (high school graduate)	-0.7395***	-0.6963***	-0.6987***	-0.6966***
Educ 3 (Some Colleges)	0.6473***	0.5662***	0.5675***	0.5491***
Educ 4 (College & Higher)	0.8174***	0.6647***	0.6655***	0.5994***
Class of Work	0.9175***	0.6079***	0.6078***	0.4719***
lFTOTINC (Family Income)	0.0417***	0.1742***	0.1729***	0.1757***
Married: Educ (high school)	0.0346***			
Married: Educ (some colleges)		0.6174***	0.6147***	0.6146***
Married: Educ (college & higher)				0.0101
Constant				0.1236***
NCHILD				0.2783***

4. Final Model

Variable	**Model 4 Coeff.**	**Analysis of Odd Ratios**
Constant	-9.4881***	
NCHILD	0.1149***	Exp (0.1149) ≈ 1.12 (12%)
NADULT	0.375***	Exp (0.375) ≈ 1.46 (46%)
AGE	0.058***	Exp (0.05797) ≈ 1.06 (6%)
I(AGE2)	-0.0001***	Exp (−0.0001391) ≈ 0.99986 (-0.014%)
Married	0.3334***	Exp (0.3334) ≈ 1.40 (40%)
Black/African	-0.8476***	Exp (-0.8476) ≈ 0.428 (43%)
Asian	-0.6467***	Exp (−0.6467) ≈ 0.524 (52%)
Race_ Others	-0.6966***	Exp (−0.6966) ≈ 0.498 (50%)
Educ 2 (high school graduate)	0.5491***	Exp (0.5491) ≈ 1.73 (73%)
Educ 3 (Some Colleges)	0.5994***	Exp (0.5994) ≈ 1.82 (82%)
Educ 4 (College & Higher)	0.4719***	Exp (0.4719) ≈ 1.6 (60%)
Class of Work	0.1757***	Exp (0.1757) ≈ 1.19 (19%)
lFTOTINC (Family Income)	0.6146***	Exp (0.6146) ≈ 1.85 (85%)
Married: Educ (high school)	0.0101	Not significant
Married: Educ (some colleges)	0.1236***	Exp (0.1236) ≈ 1.1316 (13%)
Married: Educ (college & higher)	0.2783***	Exp (0.2783) ≈ 1.3217 (32%)

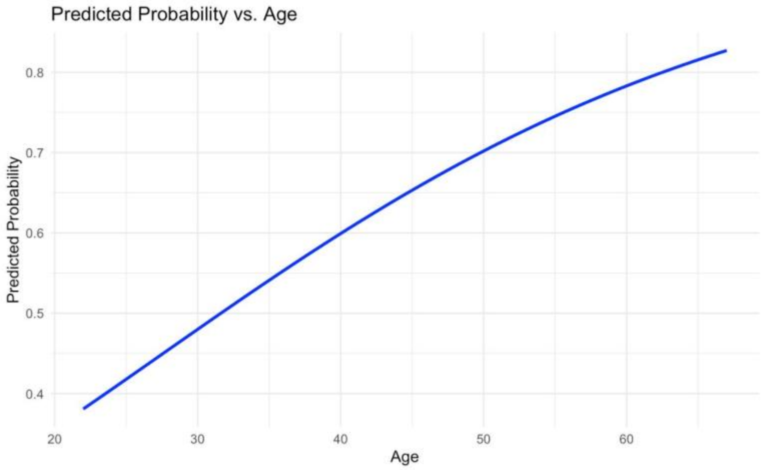
5. ROC Curve



6. Confusion Matrix

Sensitivity	0.9158605
Specificity	0.3774598
Accuracy	0.7596864

7. Predicted Homeownership by Age Squared


Predicted Probability vs. Age
Predicted Probability
Age