**Mortgage Statistics of the US through a Demographic, Geographic and Economic Lens**

**Dataset:**

**Data URL :** [New Residential Mortgage Statistics (1998-2022) | Kaggle](https://www.kaggle.com/datasets/guillemservera/new-residential-mortgage-statistics)

**Description :** The "New Residential Mortgage Statistics" dataset provides comprehensive information on residential mortgages. It is sourced from the National Mortgage Database (NMDB). The prominent features of the dataset can be grouped into the following primary categories:

**1. Loan Characteristics:**

**Groups Included :** *Loan-toValue, Average Loan Amount (1,000 $), Average Contract Rate*

The loan amount indicates the principal borrowed, shedding light on average loan sizes and regional variations. Interest rates reveal market trends and potential disparities among borrower groups. Loan terms, often spanning 15 to 30 years, reflect borrower preferences and their relation to factors like age. Meanwhile, loan types, such as fixed or adjustable rates, highlight varying risk profiles and their influence on borrower choices in different market conditions.

**2. Borrower Demographics:**

**Groups Included :** *Percent Share of Loans by All Borrowers’ Race, Percent Share of Loans by All Borrower Ethnicity, All Borrowers' Ages, Percent Share by Number of Borrowers and Gender*

Borrower demographics, including age, race, ethnicity, gender, and the number of applicants, are crucial considerations in the lending industry. Age can influence loan terms and approval rates, while factors like race and ethnicity may impact lending decisions. Gender and the number of borrowers applying can also be significant variables. These demographic details are essential for lenders to ensure fairness and equity when evaluating loan applications.

**3. Property Details:**

**Groups Included :** *Average Purchase Price / Appraised Value (1,000 $)*

The property type, ranging from single-family homes to condos, influences its valuation and risk profile. Location plays a pivotal role, with property values fluctuating based on urban or rural settings and regional dynamics. The property's appraised value is key for gauging loan-to-value ratios and equity standings.

**4. Borrower Financial History:**

**Groups Included :** *Back-End Debt-toIncome Ratio, All Borrowers' Credit Score*

Payment history reflects borrowers' punctuality in loan repayments, aiding in assessing default risks. Borrower income-to-debt ratio measures financial health. Income levels directly affect borrowing capacity and loan terms. Credit scores, indicating creditworthiness, guide lenders toward favorable loan offers.

**5. Loan Description**

**Groups Included :** *Refinance Originations, Mortgage Terms*

Refinancing methods detail how borrowers intend to refinance their loans. Meanwhile, the mortgage description outlines the specific type they've chosen, such as adjustable or fixed-rate mortgages.

**Data Cleaning & Preparation :**

**1.Integrating and Ensuring Data Uniformity:**

Integrating data from multiple sources, like the dataset's 65 CSVs with 1510 columns, poses challenges in maintaining consistency across national to state levels. To address this, we'll standardize variables, ensuring uniformity in categories and date formats, regardless of their origin.

**2. Normalization and Data Scaling:**

The issue with mortgage-related variables having varying scales, such as loan sum, income, and property valuation, can be addressed by applying data scaling techniques like Min-Max scaling or Z-score normalization. These methods help standardize variable scales, preventing any single variable from dominating analyses solely due to its magnitude, aligning with our analytical objectives.

**3. Conversion of Categorical Variables:**

For machine learning algorithms requiring numerical inputs, categorical variables like loan intent and property classification pose challenges. To address this, we'll employ encoding techniques, such as one-hot or label encoding, tailoring the approach to the specific analytical goal.

**4. Engineering Features:**

We'll probe potential feature engineering avenues based on the analytical questions or objectives. This could involve formulating new variables, data consolidation, or computing pertinent ratios.

**5. Addressing Data Imbalance:**

Predicting rare events like mortgage defaults can lead to a skewed dataset with notable class imbalances. To rectify this, we'll adjust representation through amplification or reduction techniques and employ appropriate evaluation metrics to maintain model accuracy.

**Additional Considerations:**

* Temporal Alignment: The dataset may encompass data from varied timeframes, necessitating synchronization or interpolation for absent intervals.
* Outlier Management: Rather than merely discarding outliers, we can employ methods to limit their influence or modify variables to mitigate outlier effects.
* Data Cleansing: This involves eradicating any missing values in the dataset and eliminating duplicate entries that might arise from merging discrepancies. We can use techniques such as replacing with mean values or K nearest neighbors approach to replace null values in the dataset.

**Data Analysis :**

Analyzing the "First-time Homebuyers" dataset requires a blend of graphical representations and statistical methods. Bar charts, pie charts, histograms, and line graphs can visualize data trends, such as the distribution of loan amounts or regional variations in homebuyer demographics. Scatter plots and geographic maps can highlight relationships and spatial patterns, respectively. On the statistical side, descriptive statistics offer data summaries, while correlation and regression analyses explore relationships and predictions. Time series analysis tracks data over periods, and cluster analysis groups similar data points. Hypothesis testing validates data premises, and probability distributions gauge outcome likelihoods

.**Problem Statement:**

Our objective is to conduct a comprehensive analysis of the NMDB® Aggregate Mortgage Statistics, delving into the interplay between mortgage metrics, demographics, and geographic variations, all while considering the impact of external macroeconomic factors like GDP, unemployment, and significant events like the COVID pandemic. By combining univariate and multivariate studies with rigorous statistical tests and advanced machine learning techniques, we aim to offer actionable insights that can guide strategic planning and decision-making in the housing market, benefiting both industry stakeholders and potential clients.

**Hypotheses:**

1. *Hypothesis 1 (H1)*: A negative correlation is observed between the average contract rate and the number of mortgage originations.

2. *Hypothesis 2 (H2)*: The average loan amount can be effectively predicted using variables such as the average property value and the average contract rate.

3. *Hypothesis 3 (H3)*: Economic indicators, when integrated, can serve as reliable predictors for the number of mortgage originations.

4. *Hypothesis 4 (H4)*: The distribution of loan amounts for properties significantly differs between urban and rural areas.

5. *Hypothesis 5 (H5)*: Mortgages that originated during recession years exhibit a distinct distribution of contract rates compared to those that originated in non-recession years.

6. *Hypothesis 6 (H6)*: Mortgage rates, LTV, and other related metrics vary in accordance with external factors such as the employment rate and GDP.

7. *Hypothesis 7 (H7)*: Key mortgage statistics, including mortgage rates and LTV, exhibit patterns or relationships with demographic data points such as gender, race, ethnicity, and age.

Note: As our project progresses, we anticipate the possibility of uncovering new insights and patterns in the data. This may lead us to formulate additional hypotheses for testing, ensuring a comprehensive and in-depth analysis of the mortgage and housing landscape.

**Machine Learning Exploration on NMDB® Aggregate Mortgage Statistics**

We aim to deploy Machine Learning (ML) models to delve deep into the NMDB Aggregate Mortgage Statistics. Through a blend of classification, regression, and pattern recognition techniques, we aim to unearth valuable insights that can guide both mortgage businesses and potential clients.

1. **Regression Tasks** : We intend to deploy regression models to predict key mortgage metrics, providing valuable insights for lenders and borrowers alike. Our objectives include forecasting the average loan amount to anticipate borrower needs, gauging borrower financial stability by forecasting the Debt-to-Income (DTI) ratio, and leveraging ML models to predict interest rate trends, ensuring competitive positioning for lenders and informed decisions for clients**.**

**2. Classification Tasks:** We plan to utilize classification models to discern borrower preferences and financial standings. Specifically, we'll determine whether borrowers favor adjustable-rate mortgages or fixed-rate mortgages, informing lenders' marketing and product decisions. Additionally, by categorizing clients into high or low LTV ratio groups using their credit scores, we aim to provide clarity on potential down payment requirements for borrowers.

**3. Pattern Recognition:** We aim to identify patterns and shifts in the mortgage market, equipping businesses with insights to adapt to evolving trends. Additionally, we'll delve into the relationship between credit scores and LTV ratios, offering crucial insights for refining risk assessments and customizing financial products.

**Novelty:**

In contrast to prior reports that primarily focused on demographic and geographic aspects of mortgages, our study uniquely intertwines these elements with external economic indicators like GDP, unemployment rates, and economic crises, offering a deeper understanding of mortgage behaviors. Our multivariate approach, combined with rigorous hypothesis testing, ensures a comprehensive and validated view of the mortgage landscape. Additionally, our intent to employ ML modeling techniques, unexplored in previous studies, further enhances our analysis, promising novel insights and addressing gaps in earlier research.

**APPENDIX**

Figure 1: Despite rising home prices, first-time buyers still account for approximately half of the home purchase mortgage market.

At the end of 2018, the average home price in the US was 12% above its 2007 peak. Combined with historically low housing inventory, affordability has been a concern expressed by many in the housing industry."

Despite these concerns, about half of all home purchase mortgages have gone to first-time buyers each year since 2002. While 600,000 fewer mortgages went to first-time buyers in 2018 than in 2002, this is primarily a result of the overall decline in the purchase market during the financial crisis of 2007 to 2009 and the steady recovery since 2011.

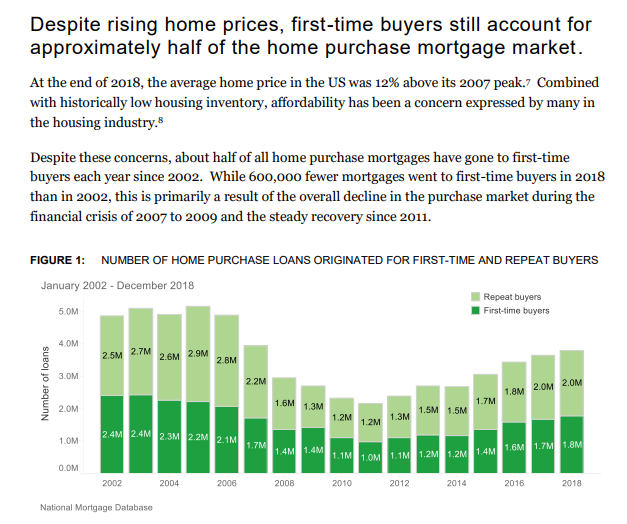


Figure 2: Along with rising incomes, looser CLTV and DTI standards have also helped first-time buyers afford pricier homes.

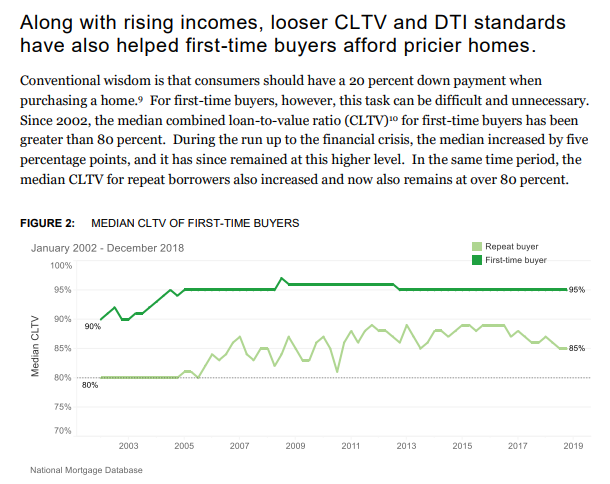
Conventional wisdom is that consumers should have a 20 percent down payment when purchasing a home. For first-time buyers, however, this task can be difficult and unnecessary. Since 2002, the median combined loan-to-value ratio (CLTV) for first-time buyers has been greater than 80 percent. During the run up to the financial crisis, the median increased by five percentage points, and it has since remained at this higher level. In the same time period, the median CLTV for repeat borrowers also increased and now also remains at over 80 percent.

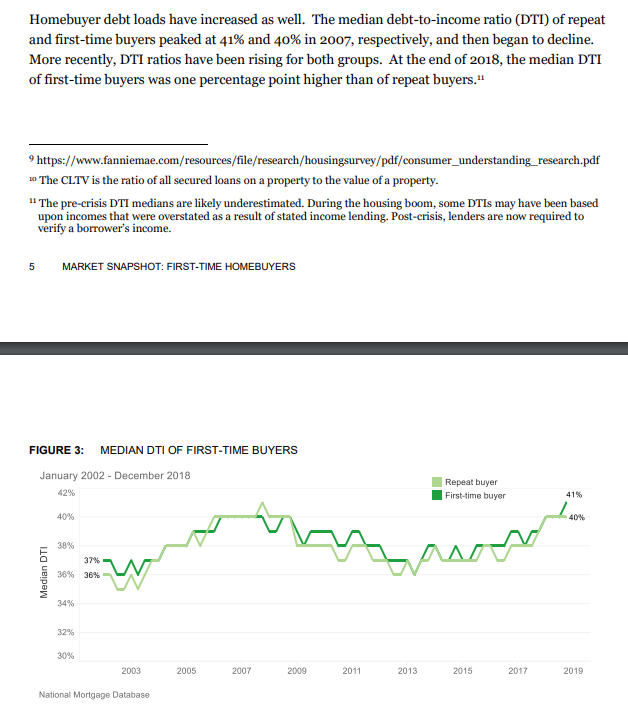
Figure 3: Homebuyer debt loads have increased as well. The median debt-to-income ratio (DTI) of repeat and first-time buyers peaked at 41% and 40% in 2007, respectively, and then began to decline. More recently, DTI ratios have been rising for both groups. At the end of 2018, the median DTI of first-time buyers was one percentage point higher than of repeat buyers."

Figure 4: However, looser CLTV and DTI standards have mainly helped first-time borrowers with higher credit scores.

In the years prior to the financial crisis, the median credit score for both first-time buyers and repeat buyers trended downward. Since 2007, however, that trend has been reversed as lenders tightened their underwriting requirements. (In more recent years, an improving economy also likely contributed to improved financial outcomes and overall higher scores.) In 2018, the median credit score of first-time buyers was 32 points higher than in 2002. The spread between the median credit score of first-time buyers and repeat buyers in 2018 was 45 points, slightly higher than the spread in 2002.

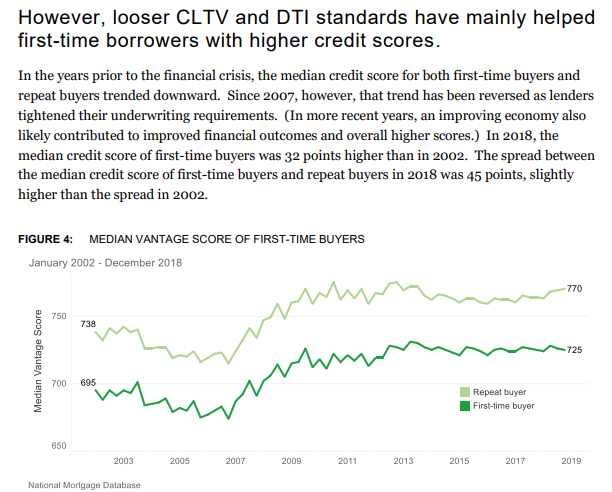


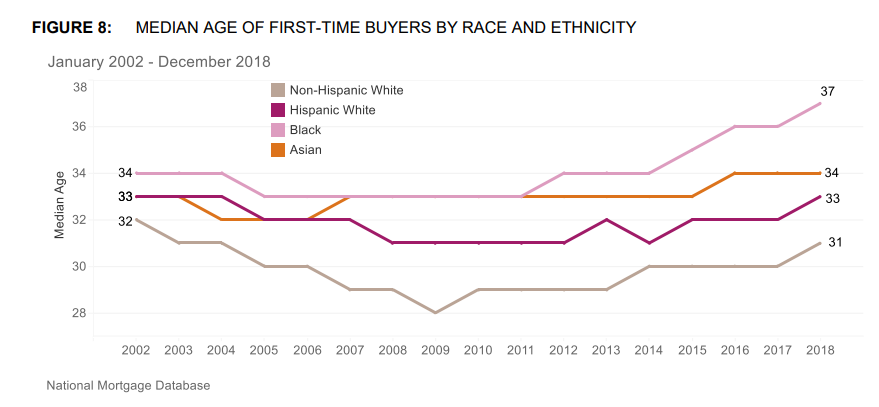
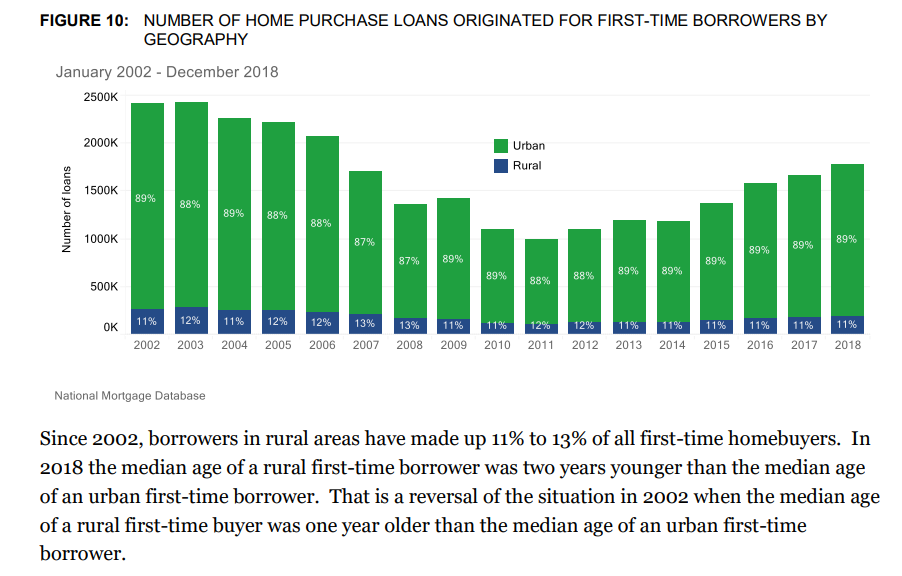
Figure 8: When broken down by race and ethnicity, it becomes clear that some groups have had larger changes in the age of first-time homeownership than others. In 2018, the median first-time black borrower was six years older than the median non-Hispanic white borrower. That is a much larger disparity than in 2002, when the age gap between black and white first-time borrowers was just two years. Black borrowers were particularly hard hit during the financial crisis, and accordingly, were more likely to go from owning to renting.14 The resulting lost wealth and generational impact likely contributed to the increasing age gap in first-time homeownershipDespite the accelerated aging of rural areas, rural first-time buyers borrow at similar ages to their urban counterparts. Rural counties contain over 85 percent of “older age counties,” 16 many of which experience persistent population loss of young adults.17 However, rural areas remain a consistently strong market for first-time homebuyers. 

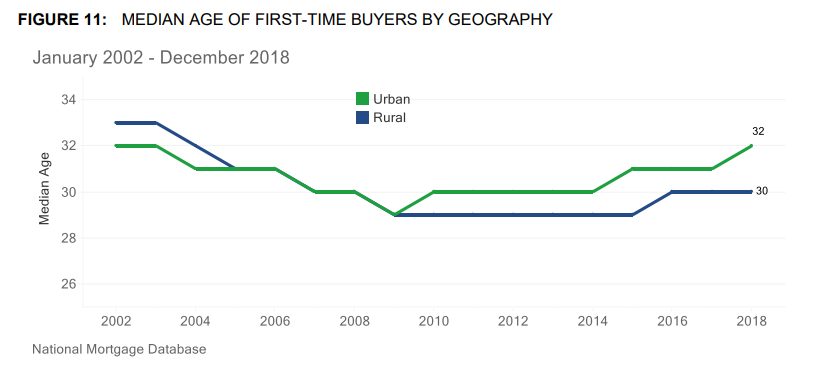
Figure 11: Since 2002, borrowers in rural areas have made up 11% to 13% of all first-time homebuyers. In 2018 the median age of a rural first-time borrower was two years younger than the median age of an urban first-time borrower. That is a reversal of the situation in 2002 when the median age of a rural first-time buyer was one year older than the median age of an urban first-time borrower. 

Figure 13: The median CLTV and DTI for rural first-time borrowers are lower than for urban buyers although the gap has narrowed since 2002. Prior to the financial crisis, rural borrowers had lower CLTVs than urban borrowers. Since 2007, the CLTVs of first-time urban and rural borrowers have converged. This trend coincides with a relative increase in rural USDA and FHA loans that tend to have higher CLTVs.

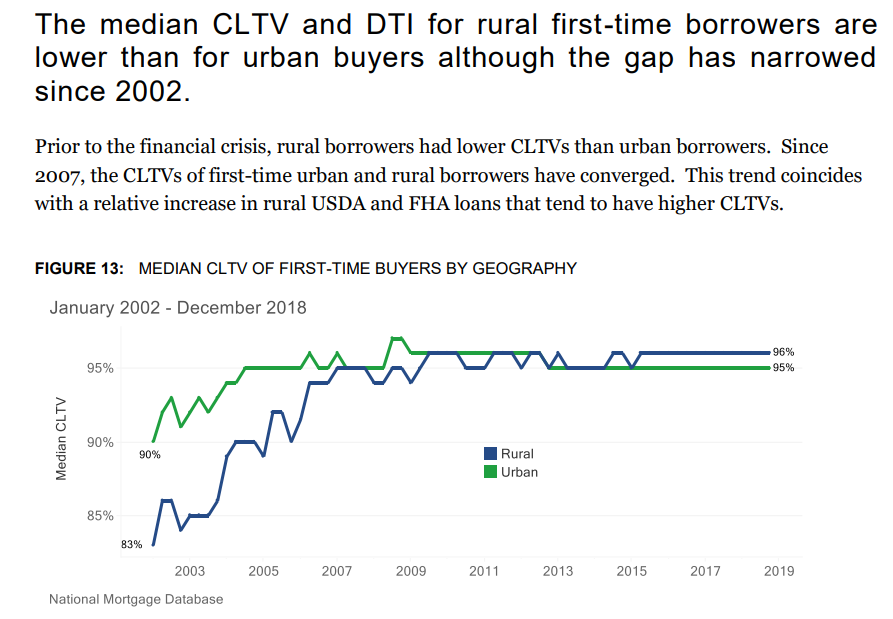
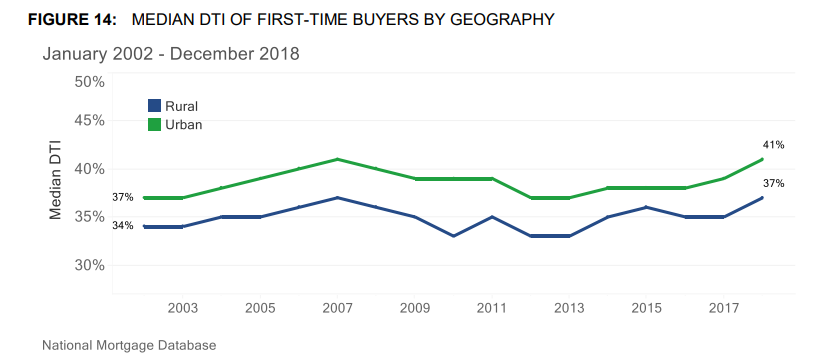
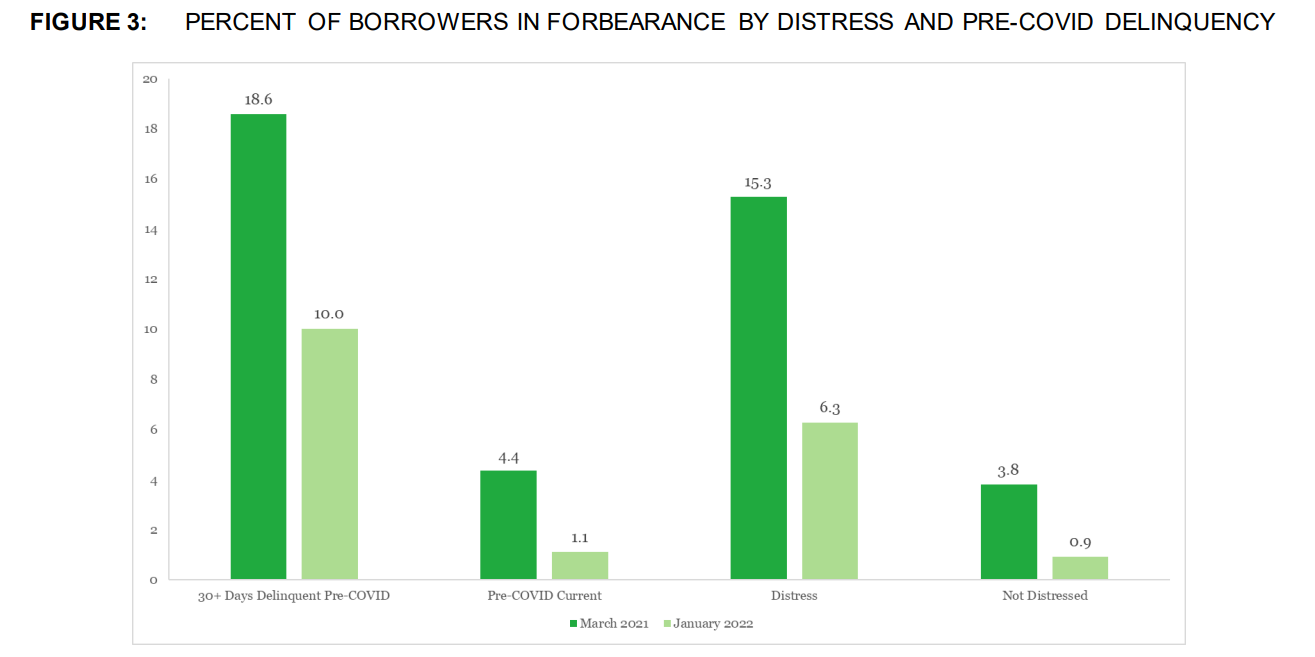
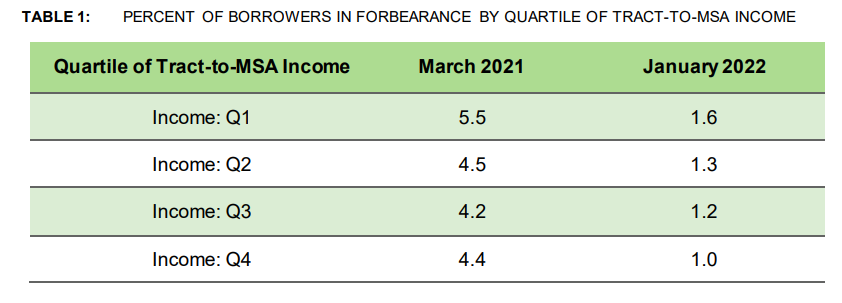
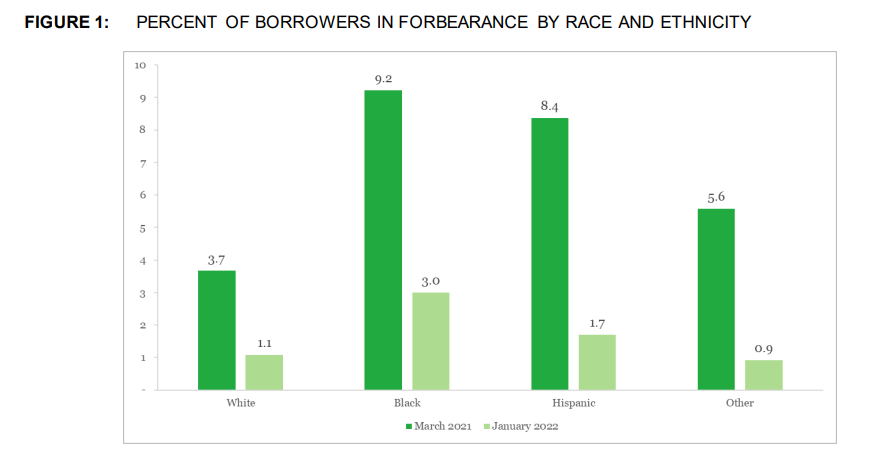


Figure 14: Even though rural borrowers tend to have lower incomes in comparison to their urban counterparts, the median rural borrower had lower DTIs over the past 15 years as rural borrowers have greater access to relatively lower cost housing. In 2018, the median rural firsttime borrower had an income of $54,000, compared to $68,000 for urban borrowers. 

Some major observations during the Covid period as shown below- 

References:

1. https://files.consumerfinance.gov/f/documents/cfpb\_market-snapshot-first-time-homebuyers\_report.pdf
2. https://files.consumerfinance.gov/f/documents/cfpb\_characteristics-of-mortgage-borrowers-during-covid-19-pandemic\_report\_2022-03.pdf