Assessing Regional and Demographic Discrepancies of Lending Risks Associated With Mortgage Loans in the U.S.

Asmin Acharya

ECON 8720 Intro to Data Management & Analysis

Andrew Young School of Policy Studies | Georgia State University

Table of Contents

Introduction 3
Context 3
Motivation/Purpose 3
Data 4
Data Collected Using Kaggle & FHFA 4
Background Info & Details About Datasets 4-5
Analytic Dataset 5
Details About Raw Data 5
Merging Datasets & Renaming Variables 5-6
Transforming & Appending 6
Dummy Variables & Interaction Variables 7
Identifying Missing Values & Outliers 7-8
Additional Steps Taken for Final Dataset 8
Defining Key Variables Used 8-9
Descriptive Statistics (Race & Gender) 9-10
Changes in LTV (2010, 2015, 2020) 11
Analysis 12
Relationship Between LTV & Note Amount (Race & Gender) 12-14
Comparing LTV Based on Credit Scores (Race & Gender) 14-16
Location of High Risk and Low Risk Borrowers (2010 & 2020) 16-18
Trends in Geographical Location (2010 & 2020) 19-21
Analyzing Regression Results 22
Differences in LTV Based on Race & Self Employment Status 22-23
Additional Details Regarding Regression Models 24
Conclusion 24
New Questions, Implications, & Limitations 25-26
Next Steps/Future Work 26
Appendix 27-49
References 50

List of Figures

Avg. LTV in Each Year........11
LTV vs. Note Amount (Race)........13
LTV vs. Note Amount (Gender)........14
Comparison of LTV Based on Credit Score (Race)........15
Comparison of LTV Based on Credit Score (Gender)........16
Borrowers with 760 or Above Credit Score in the U.S. (2010)........20
Borrowers with 760 or Above Credit Score in the U.S. (2020)........20
Borrowers With Credit Score Between 620 & 660 in the U.S. (2010)........21
Borrowers With Credit Score Between 620 & 660 in the U.S. (2020)........21

List of Tables

Descriptive Statistics (White Borrowers)........9
Descriptive Statistics (Non-White Borrowers)........10
Descriptive Statistics(Female Borrowers)........10
Descriptive Statistics(Male Borrowers)........10
States With the Highest Number of Low-Risk Borrowers (2010)........17
States With the Highest Number of High-Risk Borrowers (2010)........17
States With the Highest Number of Low-Risk Borrowers (2020)........18
States With the Highest Number of High-Risk Borrowers (2020)........18
Changes to LTV Based on LTV & Self-Employment Status........23

Introduction

Context

For many decades, the American Dream has been rooted in certain beliefs and principles that were common among individuals. This included having a stable job, high earnings, raising a family, etc. These ideas, however ,in modern day have drastically changed as new cultural and economic implications have given Americans more options towards living a life that is outside the antiquated societal norms. There is one idea, however, that still seems to linger on the minds of individuals today, which is the dream of owning a home. The demand for housing continues to rise and as a result the price of these homes have increased at a much faster rate than the wages of average Americans. Regardless of affordability issues, American families continue to use mortgages to purchase their homes, with the potential of lowering their housing costs when rates decline. Homes are not just a place to live anymore but rather a way to tap into home equity which opens up different types of investment opportunities.(Bhutta et.al 2022, p.1). Unfortunately, not everyone can get a mortgage so easily. Loan applicants are not just denied by lenders due to their poor credit score, but may also be denied due to their gender, race, or ethnicity. Many studies have already proven these types of disparities and other studies have shown that certain lending practices have been used specifically to take advantage of minorities. During the housing crisis from 2007 to 2010, predatory lending was used to impose unfair and abusive loan terms on borrowers (Agarwal et.al 2014, p.29). These aggressive tactics were used on borrowers who predominantly happened to be immigrants and minorities as they lacked knowledge about the structure of these loans.

Motivation/Purpose of Report

The motivation behind this report is to assess the differences in lending risks of borrowers in the United States who use mortgages . The goal of this report is not to necessarily prove whether racial or gender biases exist in the approval of these loans, but rather gain insights into high risk and low risk borrowers. Previous studies in the field have already shown that certain demographic characteristics affect the approval of loans. These biases do exist and have disproportionately affected minorities. The report aims to answer questions such as: What regions in the U.S. are high risk and low risk borrowers located? Are non-white borrowers higher risk compared to white borrowers given certain factors? Is there a difference in lending risks associated with female and male borrowers? Have the number of high risk and low risk borrowers changed over a specific period in the U.S.? In general, we are trying to compare borrowers based upon geographic and demographic characteristics. The point is to not only find differences amongst borrowers but to build models to see if these differences are in fact statistically significant. These models will give crucial details about how certain characteristics impact the lending risk of these borrowers.

Data

Data Collected Using Kaggle

The data for this report was collected from two different websites. The first four datasets were collected using Kaggle and the fifth dataset was collected using the Federal Housing Finance Agency. Users on Kaggle post various types of datasets that people in the data science community can use for their own purposes. One of these datasets included federal home loan level from 2009 to 2018. Using the search bar in Kaggle and typing "Federal Home Loan Level Bank System 2009-2018" will allow anyone to access this particular dataset. For this report, data was only collected for the years 2010, 2015, and 2020. In Kaggle four datasets were collected for the years 2010 and 2015. One of the interesting things about this website is that the interface allows users to filter variables to include and exclude in the dataset before downloading it. The filter was used to create two datasets for each of the years. Around 30 variables were included to create the first dataset for 2010 called 2010 part 1 and the rest of the other variables were collected for 2010 called 2010 part 2. This was the same method that was used to create another two datasets for 2015. The variables in each of these datasets were consistent throughout the years, so that the variables used in 2010 part 1 and part 2 matched the variables used in 2015 part 1 and part 2.

Data Collected Using Federal Housing Finance Agency

As for the fifth and last dataset, it was collected via the FHFA website. The Kaggle dataset only included data for the years from 2009 to 2018. The FHFA public use databases included datasets for the following years and also the previous years as well. Unfortunately, the FHFA website does not allow users to filter through the variables they can use, so the entire dataset includes all the variables for that year. Nothing was filtered or modified before downloading the 2020 dataset.

Background Info and Details About the Datasets

One thing to note is that the datasets that were downloaded using Kaggle originated from the FHFA website, so users can find the previous year datasets in one downloadable file. Dividing these datasets into different parts, however, will prove to be useful when creating our final dataset as it allows for more flexibility when formatting. All the five different datasets that were downloaded are available as cross-sectional data. Although the datasets are cross sectional, the goal is to transform them into one pooled cross-sectional dataset. These datasets were made available via the Public Use Database (PUDB) from the FHFA . It supplies lender, researchers, and policy makers with information regarding the flow of mortgage credit. The PUDB datasets include loan level records with data elements concerning the demographic characteristics of the borrowers. Additionally, it includes the lending risks associated with the borrowers and specific information about their mortgages. The observational unit for all the datasets are the same. It looks at individual borrowers in the U.S. who are from different states. The FHFA includes loan level data from 2009 to 2022, but for this report we are pulling data for only three periods (2010, 2015, and 2020). The data allows us to answer questions that are related to the characteristics of the individual borrowers and their lending risks such as where high risk and low risk borrowers are located, change in lending risks over the three periods, differences in lending risk based on race and gender, etc. As for the raw observations and variables, each dataset has different amounts. Dataset 2010 part 1 has 41220 observations and 30 variables, 2010 part 2 has 41220 observations and 65 variables, 2015 part 1 has 47840 observations and 29 variables, 2015 part 2 has 47840 observations and 60 variables. Lastly, the single 2020 dataset has 83106 observations with a total of 56 variables.

Analytic Dataset

Details About the Raw Data

Once the raw datasets are collected, additional steps will be taken to create the final analytic dataset. Fortunately, there were not any major problems in the way the raw data was collected or downloaded. The dataset was well organized and included many variables which can be used to assess the lending risks of different borrowers. The types of variables that were available for each yearly dataset, however, were not consistent for all the years. Datasets before 2019 included additional variables which were not available for the 2020 dataset. This was not a big issue as the variables for the 2020 dataset were the same ones that were originally used for the previous years. Due to the sheer number of different factors, not all variables from the datasets were used when creating the final dataset. In total, around 30-35 variables were included for the 2010, 2015, and 2020 datasets. The variables were the same and were consistent for all the corresponding years. These variables were the most relevant in understanding the differences between borrowers and their associated lending risks. This included looking at the borrower's monthly income, race, gender, age, median income, credit score, housing expense ratio, debt expense ratio, and the state where the borrower lives, just to name a few. Data was also available for coborrowers, but coborrower data will not be the focus of this report.

Merging Datasets & Renaming Variables

As mentioned previously, the datasets for the different years included 30 variables each. For the 2010 and 2015 datasets, there were two datasets each for those years so around 15 variables in each dataset. The first step was to merge the datasets for those corresponding years. The datasets for these years included duplicate variables, so it was necessary to get rid of those duplicates from one of the datasets for each year before merging. Dataset 2010 part 2 and dataset 2015 part 2 were selected and the duplicates were removed from them. The variables that were removed were certain irrelevant ids and state codes which were already present in the 2010 part 1 and

2015 part 1 datasets. After the duplicates were removed from part 1 and part 2 datasets for 2010 and 2015, datasets were then merged. Each of the datasets for these years included a special assigned ID which was used to merge and create the final combined 2010 and 2015 datasets. The 2020 dataset was the one that was originally downloaded from the FHFA website, and the data was downloaded as one single file unlike datasets for 2010 and 2015. There was no need to merge the 2020 dataset as a result. Once the full combined dataset for 2010 and 2015 was created, the next thing to do was select the variables to use. Each yearly dataset would have the same variables being used, so that later on all the yearly datasets would be combined or appended to form the final dataset. The variables mainly consisted of the borrower's and coborrower's LTV ratio, monthly income, age, credit score, loan amount, interest rate of loan, state of location, etc. Other variables of interest included property type, housing expense ratio, mortgage type, and debt expense ratio just to name a few. After the appropriate variables were selected for each yearly dataset, the next step was to rename the variables. The 2010 and 2015 datasets had variable names which were different from the 2020 dataset. Although the values/data represented the same thing for each of the years, it was important to find meaningful names that users would be able to decipher. Furthermore, none of the variables included any labels, but the FHFA website did include a pdf of all the variables and their definitions similar to a code book. For this reason, the variables across all the yearly datasets were renamed into something meaningful.

Transforming Variables & Appending Datasets

Certain values for the 2010 and 2015 datasets were also slightly different compared to the 2020 dataset. For example, the values in the 2020 dataset like LTV ratio, housing expense ratio, debt to expense ratio, and interest rate were represented as percent values. In the 2010 and 2015 datasets they were represented as decimal values. Since we will be combining the yearly datasets, it is crucial that the data values are represented in the same way. Due to this, the decimal values in 2010 and 2015 were transformed into whole percentages like the values in 2020. Another important variable that was transformed was the income amount in the 2010 and 2015 datasets. These values represented the annual income of the borrower, but in the 2020 dataset this value was represented as monthly income for the borrowers. The annual income values in the 2010 and 2015 datasets were then changed to represent the borrower's monthly income. The process of merging datasets, selecting variables, renaming variables, and transforming variables was complete. Now it was time to append the yearly datasets to create one single dataset. The 2010, 2015, 2020 datasets were appended into one single file called FHLAppend. The FHLAppend dataset includes 171,806 observations and 34 variables. All the values from the previous yearly datasets are present in the FHLAppend dataset and now it is a pooled cross-sectional dataset. There are now three different time periods included in our dataset which can help in finding trends over certain years.

Creating Dummy Variables & Interaction Variables

There are still additional variables that need to be added to FHLAppend before creating the final analytic dataset. Dummy variables were created to distinguish certain characteristics of the borrowers and coborrowers. There are dummies for white borrowers, female borrowers, and black borrowers. These same dummy variables were also created for the coborrowers. Again, the purpose of this report is to find differences in lending risks for the borrowers which include looking at differences based on race and gender. The variables for credit scores were also converted into dummy variables. Credit scores were originally given as categorical values which included 5 different ranges of the borrower and coborrower's credit score. In this report the lowest range was excluded, which was a credit score equal to or less than 620. Dummies were created for the other categories with the lowest credit scores being in the range of 620 to 660 and the highest credit scores which were equal to or greater than 760. In total 14 different dummies were added to the FHLAppend dataset. There were 6 dummies for white, female, and black borrowers and coborrowers. There were 8 dummies created for the credit score for borrowers and coborrowers. The dummies will prove to be helpful later on during the analysis section of the report. In addition to the dummies, a few interaction variables were created by multiplying two dummies together. Three interaction variables were created and added to the FHLAppend dataset. These variables included self-employed white, female, and black borrowers. The original raw datasets contained a dummy variable to see whether the borrower was self-employed or not. With the dummies that were created earlier to distinguish borrowers by race and gender, these dummies were multiplied with the self-employment dummy. These interaction terms help in distinguishing the differences between white, female, and black borrowers who are self-employed. It should be noted that interaction terms were only created using the dummies for borrowers and not coborrowers.

Identifying Missing Values

When the raw datasets were originally downloaded, it did not include any missing values. Upon further investigation, it was apparent that the missing values were represented as a categorical ID rather than a missing cell. There were a couple of variables of interest in the FHLAppend dataset where missing values were denoted with a certain ID number. One of the more important variables in FHLAppend was the borrower and coborrower's age which had missing values. Depending on which year the observation was, the missing ID was slightly different. For the years 2010 and 2015 the missing values were identified as 99 and 98. 99 represented values where the age was not provided and 98 represented values that either did not have a coborrower on the loan or the value was not applicable. As for the 2020 observations, the missing values were identified as 999 and 998. These values represented the same thing as the values for 2010 and 2015. Instead of dropping these missing values in the FHLAppend dataset, they were replaced. The categorical IDs were now replaced by "NA" or traditional missing values where those categorical missing IDs would not be used as the age of the borrowers or coborrowers.

Identifying Outliers in the Dataset

When running summary statistics of the relevant variables that were going to be used for analysis, some variables had values which were extremely high. Outliers were present for housing expense ratio percent and total debt expense ratio percent. Some values were very far off from their respective mean and median. As a result, outliers for the housing expense ratio variable were dropped from the dataset. The traditional method of identifying outliers was used by comparing the values based on the upper and lower quartiles. ("How to Remove Outliers from Data in R", 2022). Only the outliers for housing expense ratio were removed from the dataset. This action ended up getting rid of the outliers for the total debt expense ratio variable. Most likely indicating that outliers from both variables were present in the observations that were initially dropped.

Additional Steps Taken to Produce Final Analytic Dataset

There are a couple of things left to do before we create our final dataset. For this report we only examine borrowers and coborrowers, meaning that only observations with a combined borrower count of 2 or 1 should be included. The raw dataset included a variable called borrower count , so this variable was used to format the dataset where observations with a borrower count of more than 2 were dropped. Furthermore, one of the main insight in the report is to find the differences in lending risks for male and female borrowers. The dataset included distinct IDs for the borrower's gender which had more than two different categories. There were a few other categorical IDs which identified whether the borrower's gender was not provided or if it was not applicable. For the borrower's gender variable we only consider individuals who are either female or male. At last, after all the procedures were taken, the final analytic dataset was created. The FHLAppend dataset now had a total of 161,951 observations and 51 different variables. In the upcoming sections we will now use this final analytic dataset to describe some of the key variables and provide a general overview of different groups of borrowers.

Defining Key Variables Being Used

Before we proceed with any type of basic summary statistics or analysis, it is critical to describe some of the key variables that we will use. One of our main variables of interest is loan to value ratio percent (LTV). The LTV ratio is an assessment of lending risk that many institutions and lenders survey before approving a mortgage. It is calculated by dividing the mortgage amount by the appraised property value. (Hayes, 2024). Essentially, high LTV ratios are considered higher risk loans and low LTV ratios are considered lower risk loans. Other variables that we will focus on include total debt expense ratio, housing expense ratio, note amount (loan amount), note rate percent (interest rate on mortgage), etc. The total debt expense ratio is a ratio of all debt payments to total borrower income. The housing expense ratio is the ratio of mortgage principal and interest and housing expenses to the borrower's total income (FHFA, 2022). The other variables that we will use in our analysis are self-explanatory just based on their names, but a codebook from the FHFA is available if there is any confusion.

Comparison of Borrowers By Race and Gender (Descriptive Statistics)

Now that the variables have been defined, we can move onto produce summary statistics and compare borrowers by group. We will begin by comparing the summary statistics of white borrowers and non-white borrowers. To clear up any confusion non-white borrowers include individuals who are Black, Asian, Native American, and Hispanic/Latino. In Table 1 we can see that the mean LTV ratio percent value (73%) is lower for white borrowers compared to Non-white borrowers (75%) in Table 2. Based on just summary statistics alone we can see that there are differences in the mean value amongst the different racial groups. The mean housing expense ratio is also lower for whites (18%) rather than non-whites (20%) as well as the total debt to expense ratio. White borrowers have a total debt to expense ratio of 29% and non-white borrowers have a value of 31%. The average total monthly income amount is actually higher for non-white borrowers ($9772) rather than White Borrowers ($9195). The average HUD median income amount is also higher for non-whites ($77157) compared to Whites ($70494). As for the other values such as the age of the borrowers and coborrowers and note percent, they are quite similar among the groups. There are also a few differences when it comes to comparing male and female borrowers. Table 3 and Table 4 provide summary statistics on female and male borrowers. We can see that there are a couple of variables that are quite different among the groups. The average total monthly income amount for females is lower ($7592) compared to males ($9758). Even comparing the median of the monthly income we can see that females have lower values compared to males. The note amount or the loan amount is also lower for females ($175790) than males ($199540). Looking at Tables 3 and 4 the LTV ratio, housing expense ratio, and total debt expense ratio is slightly lower for male borrower than female borrowers.

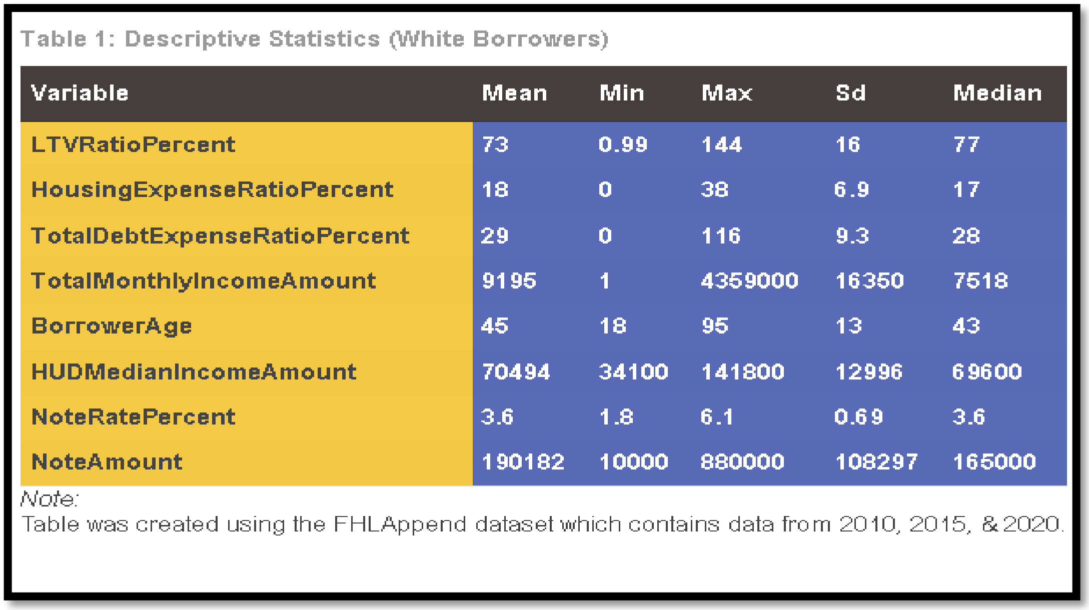


Table 1: Descriptive Statistics (White Borrowers)

Variable	Mean	Min	Max	Sd	Median
LTVRatioPercent	73	0.99	144	16	77
HousingExpenseRatioPercent	18	0	38	6.9	17
TotalDebtExpenseRatioPercent	29	0	116	9.3	28
TotalMonthlyIncomeAmount	9195	1	4359000	16350	7518
BorrowerAge	45	18	95	13	43
HUDMedianIncomeAmount	70494	34100	141800	12996	69600
NoteRatePercent	3.6	1.8	6.1	0.69	3.6
NoteAmount	190182	10000	880000	108297	165000

Note:
Table was created using the FHLAppend dataset which contains data from 2010, 2015, & 2020.

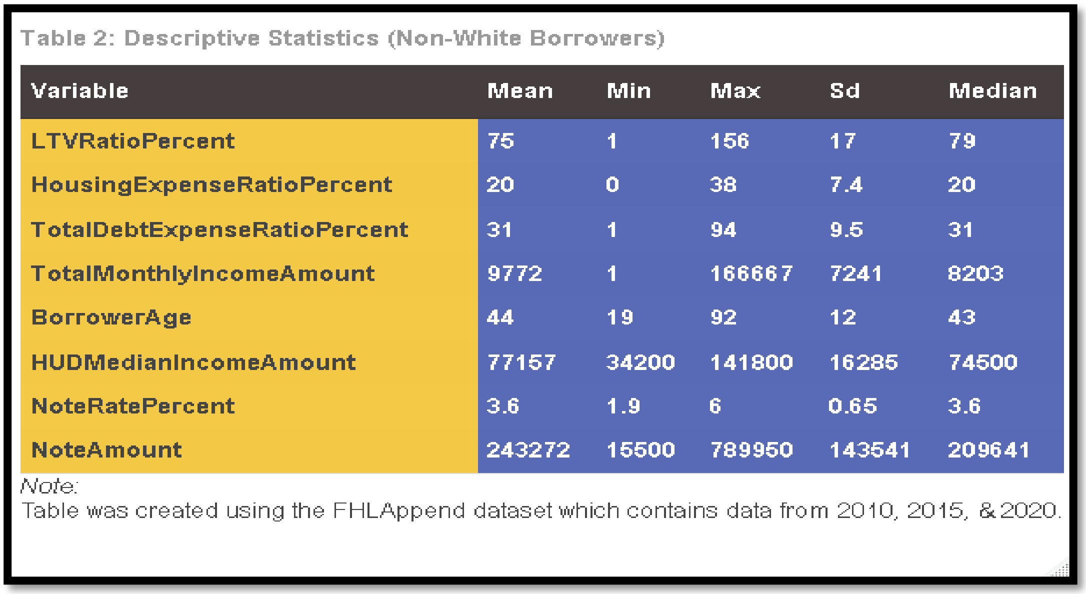


Table 2: Descriptive Statistics (Non-White Borrowers)

Variable	Mean	Min	Max	Sd	Median
LTVRatioPercent	75	1	156	17	79
HousingExpenseRatioPercent	20	0	38	7.4	20
TotalDebtExpenseRatioPercent	31	1	94	9.5	31
TotalMonthlyIncomeAmount	9772	1	166667	7241	8203
BorrowerAge	44	19	92	12	43
HUDMedianIncomeAmount	77157	34200	141800	16285	74500
NoteRatePercent	3.6	1.9	6	0.65	3.6
NoteAmount	243272	15500	789950	143541	209641

Note:
Table was created using the FHLAppend dataset which contains data from 2010, 2015, & 2020.

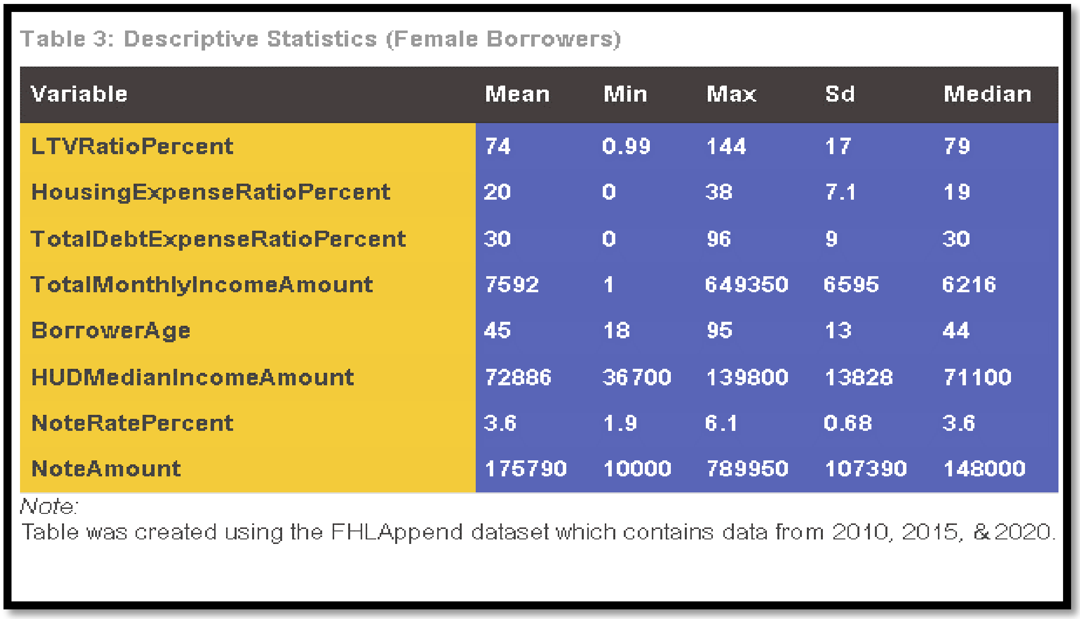


Table 3: Descriptive Statistics (Female Borrowers)

Variable	Mean	Min	Max	Sd	Median
LTVRatioPercent	74	0.99	144	17	79
HousingExpenseRatioPercent	20	0	38	7.1	19
TotalDebtExpenseRatioPercent	30	0	96	9	30
TotalMonthlyIncomeAmount	7592	1	649350	6595	6216
BorrowerAge	45	18	95	13	44
HUDMedianIncomeAmount	72886	36700	139800	13828	71100
NoteRatePercent	3.6	1.9	6.1	0.68	3.6
NoteAmount	175790	10000	789950	107390	148000

Note:
Table was created using the FHLAppend dataset which contains data from 2010, 2015, & 2020.

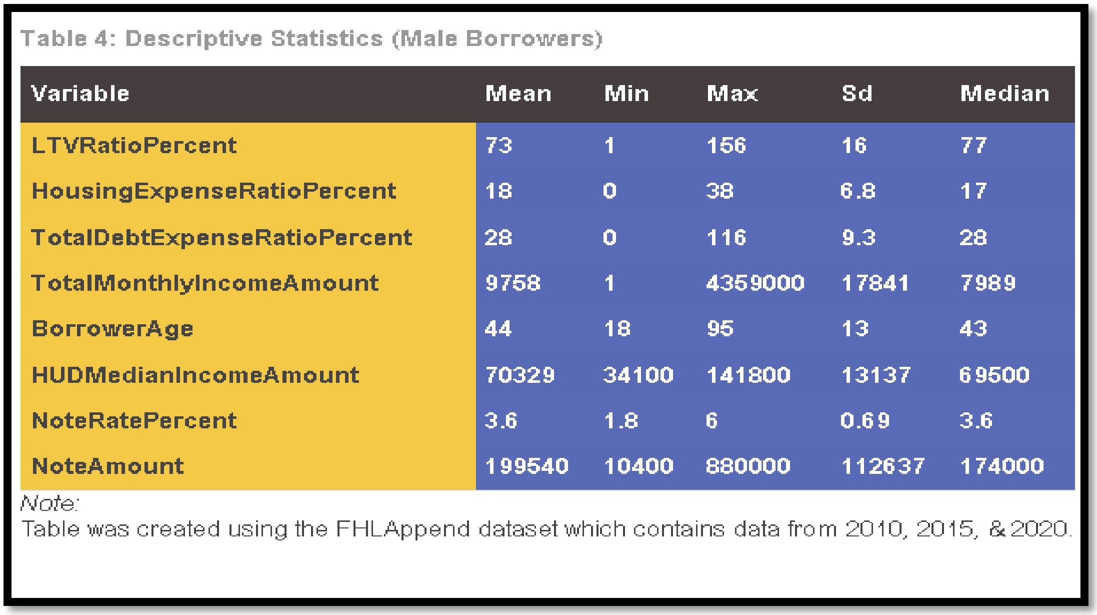


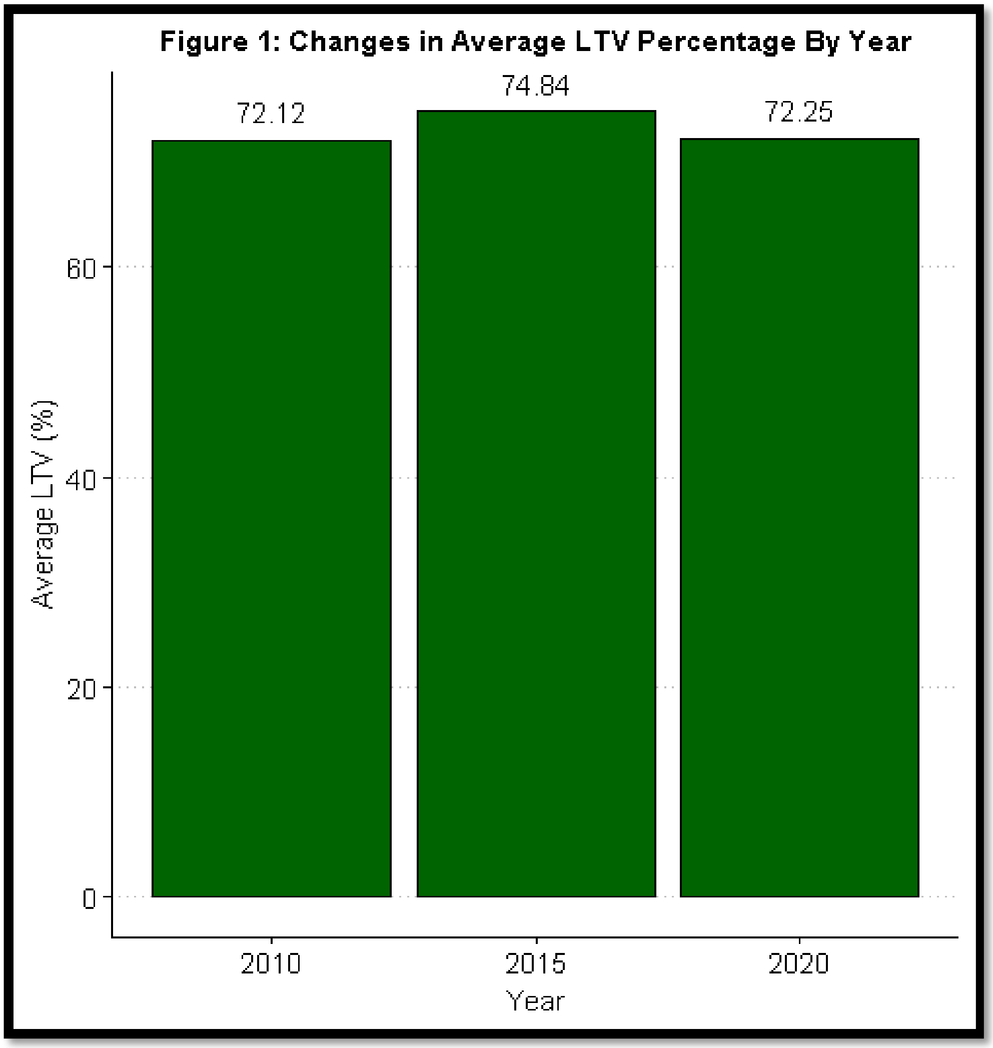
Table 4: Descriptive Statistics (Male Borrowers)

Variable	Mean	Min	Max	Sd	Median
LTVRatioPercent	73	1	156	16	77
HousingExpenseRatioPercent	18	0	38	6.8	17
TotalDebtExpenseRatioPercent	28	0	116	9.3	28
TotalMonthlyIncomeAmount	9758	1	4359000	17841	7989
BorrowerAge	44	18	95	13	43
HUDMedianIncomeAmount	70329	34100	141800	13137	69500
NoteRatePercent	3.6	1.8	6	0.69	3.6
NoteAmount	199540	10400	880000	112637	174000

Note:
Table was created using the FHLAppend dataset which contains data from 2010, 2015, & 2020.

Changes in LTV Over the Years (2010, 2015, 2020)

One our main variable of interest is LTV ratio percentage. Before looking at relationships between LTV and the other key variables, it is important to see some trends. Since the dataset includes values from three different time periods, we can identify any significant changes to LTV over those years. Figure 1 presents information on the average LTV percentage by year. It is clear to see that the average LTV has not significantly changed over the three periods. In 2010 the average LTV ratio was 72.12%, in 2015 74.84%, and in 2020 72.25%. From 2010 to 2015 there was a percent increase of 3.77% in the LTV and from 2015 to 2020 there was a percent decrease of 3.46% in the LTV.

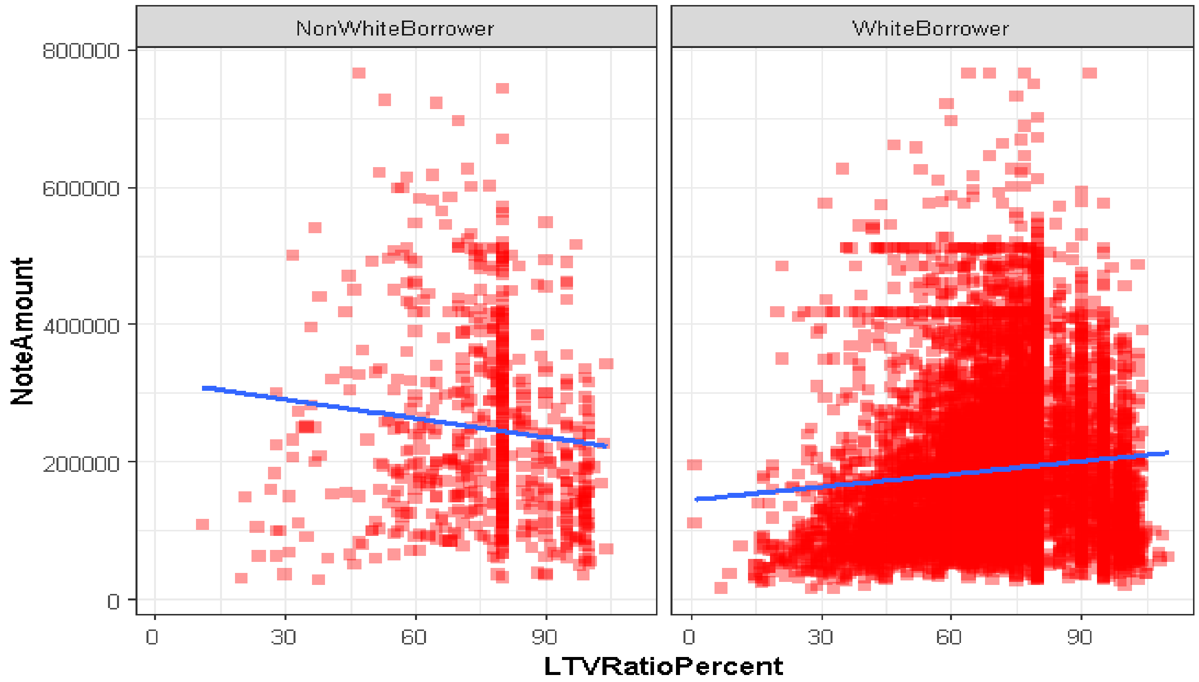


Analysis

Relationship Between LTV & Note Amount (White & Non-White Borrowers)

In this section we will look at the relationship between LTV and note amount by race. Specifically examining the differences in this relationship for white and non-white borrowers. Before getting started, a random sample was first generated called FHLAppendSample using the FHLAppend dataset. This random sample contains 10,000 observations and was used to create a scatterplot of LTV vs. note amount for the two groups we are observing. Figure 2 shows the differences in correlation between the two variables for non-white and white borrowers. One of the most noticeable things we see is that there are in general more observations for whites than non-whites in this random sample. Regardless of this difference, the trend line in Figure 2 shows that there is a difference in the relationship between the variables of the two groups. There is a negative relationship between LTV and note amount for non-white borrowers. This is the exact opposite relationship we see with white borrowers where there is a positive relationship between LTV and note amount. The corresponding correlation coefficients are also provided in Figure 2 which measure the strength of the relationship between the two variables. This coefficient is going to be a value between -1 and 1. If the values are closer to -1, we have a strong negative correlation. If the values are closer to 1, we have a strong positive correlation, and if it is 0 there is no linear relationship (Fernando, 2024).The correlation coefficient for the non-white group is -0.11 which suggests a weak negative relationship. As for the white group the coefficient value is 0.095 which suggests a weak positive relationship. If we were making general assumptions before our analysis, the common idea might have been that the higher the LTV the lower the note amount or loan amount. This makes sense as institutions would be less inclined to give huge loans to higher risk borrowers with high LTV ratios. Figure 2 , however, shows that this assumption holds for non-white borrowers, but not white borrowers. In fact, the trendline shows that as LTV increases the loan amount increases as well for white borrowers. This suggests that there are discrepancies in the relationship between variables based on race.

Figure 2: Comparison of LTV & Note Amount by Borrower's Race

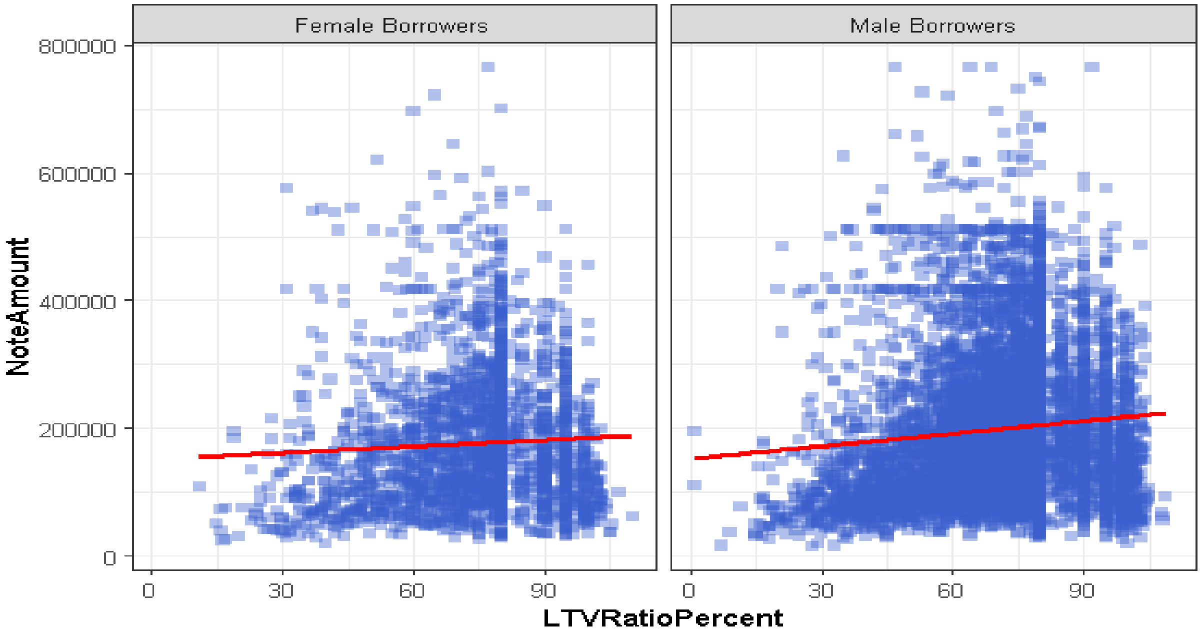


Correlation Coefficient (White Borrowers) = 0.095
Correlation Coefficient (Non White Borrowers) = -0.11
This Scatter Plot was created using a random sample of 10000 observations from the FHLAppend dataset.

Relationship Between LTV & Note Amount (Female & Male Borrowers)

The previous discussion shows that there are differences in lending risk (LTV) and loan amount (note amount) based on race, but what about differences based on gender? Figure 3 looks at the relationship between the same variables as before with a scatter plot, but this time we compare this relationship by gender (female and male borrowers). For Figure 3 we used the same random sample data (FHLAppendSample) as discussed in the previous section comparing white and non-white borrowers. It is clear to see that in this random sample there are more male observations than female observations. Even though there are fewer female observations, the scatter plot for female borrowers looks quite similar to the one for male borrowers. Figure 3 shows that the association between the LTV and note amount is a positive one. For both female and male borrowers as LTV increases so does the note amount. This relationship can be better perceived with the trendline in each group. Additionally, the correlation coefficient of the two variables for each group shows a weak positive relationship. The correlation coefficient for the female group is 0.052 and the correlation coefficient for the male group is 0.096. This difference in the correlation coefficient can be seen in the trendline for male and female borrowers as well. The trend line is steeper for males than females which suggests a stronger relationship between the variables. This suggests that there is greater rate of change between the variables for male borrowers. As LTV increases, note amount increases by a greater amount for males compared to females. The general assumption discussed earlier was that higher lending risks should result in a lower loan amount. Again, this assumption does not hold for either male or female borrowers. Although this assumption does not hold, it does not mean that there are not any differences among males and females. Figure 3 has shown that at the same LTV amount male borrowers tend to get a higher loan amount compared to female borrowers.

Figure 3: Comparison of LTV & Note Amount by Borrower's Race

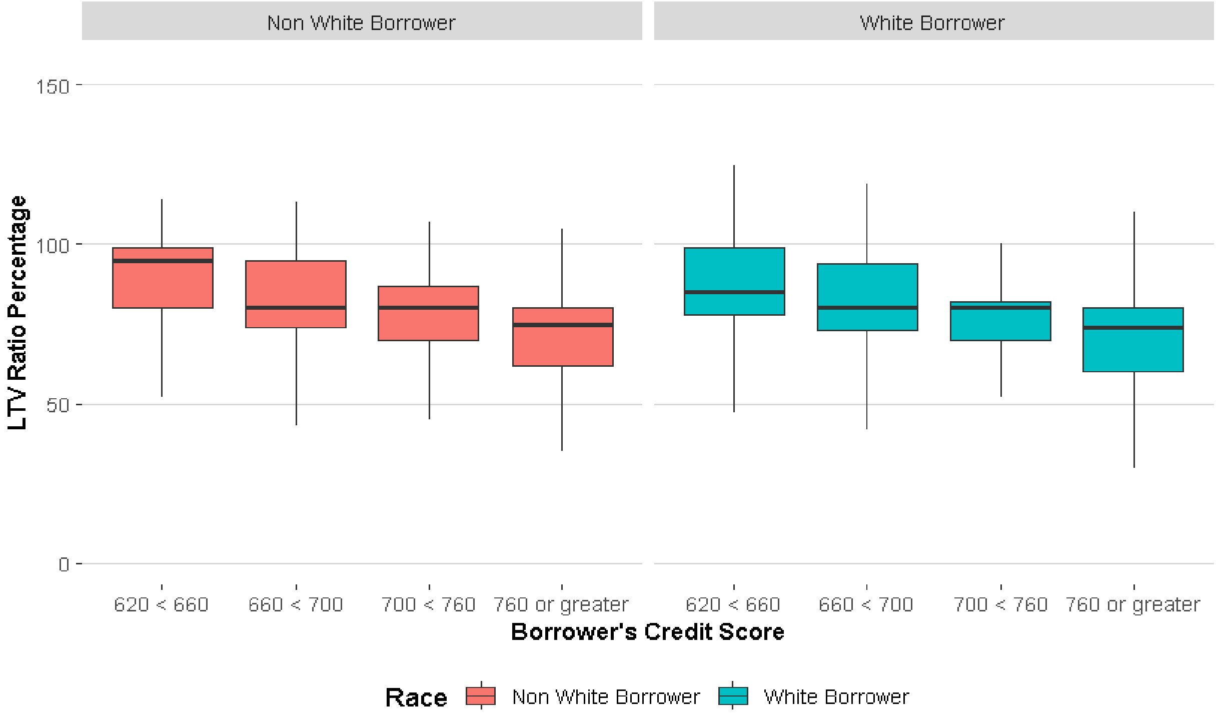


Correlation Coefficient (Female Borrowers) = 0.052
Correlation Coefficient (Male Borrowers) = 0.096
This Scatter Plot was created using a random sample of 10000 observations from the FHLAppend dataset.

Comparing LTV Based On Credit Scores (White & Non-White Borrowers)

The relationship between LTV and note amount have provided us insights into the differences in lending risks based on the borrower's race and gender. In this section we will continue to examine the differences in lending risks of these groups, but this time based on the borrower's credit score. The common presumption might be that a greater credit score entails a lower LTV ratio, but is this really the case? Additionally even if LTV is lower as credit scores increase, is there a difference in this relationship based on race or gender? If we are only comparing credit scores and LTV values then there should not be a large discrepancy among the borrowers. We first look at these differences between non-white and white borrowers using our original FHLAppend dataset. Figure 4 shows a box plot where we are comparing LTV values with different credit scores for non-white and white borrowers. For this comparison, we will only be looking at the median values of the box plot and distinguish any differences we see between the borrowers. In order to get the median values for Figure 4, we used a group statistic based on race and credit scores. The largest difference in median LTV was for borrowers in the credit score range of 620 < 660. In this range non-white borrowers had a median LTV of 95% and white borrowers had a median LTV of 85%. As for the other credit score ranges the median LTV values were very similar for both white and non-white borrowers. There was , however, a slight difference in the median LTV of borrowers who were in the credit score range of 760 or above. For white borrowers, they had a median LTV value of 74% and for non-white borrowers they had a LTV value of 75%, so slightly bigger. It may not be completely correct to assume that as credit scores increase the LTV decreases. For the credit score ranges of 660 < 700 and 700 < 760 our median LTV across both groups did not change. There was, however, some truth to this trend as there was a significant decrease in the median LTV from the credit score ranges of 620 < 660 to 760 or above. The most important finding from our analysis was that non-white borrowers in the lowest credit score range have a median LTV value that is quite larger than white borrowers in the same credit score range. This further highlights the differences in lending risk for white and non-white borrowers.

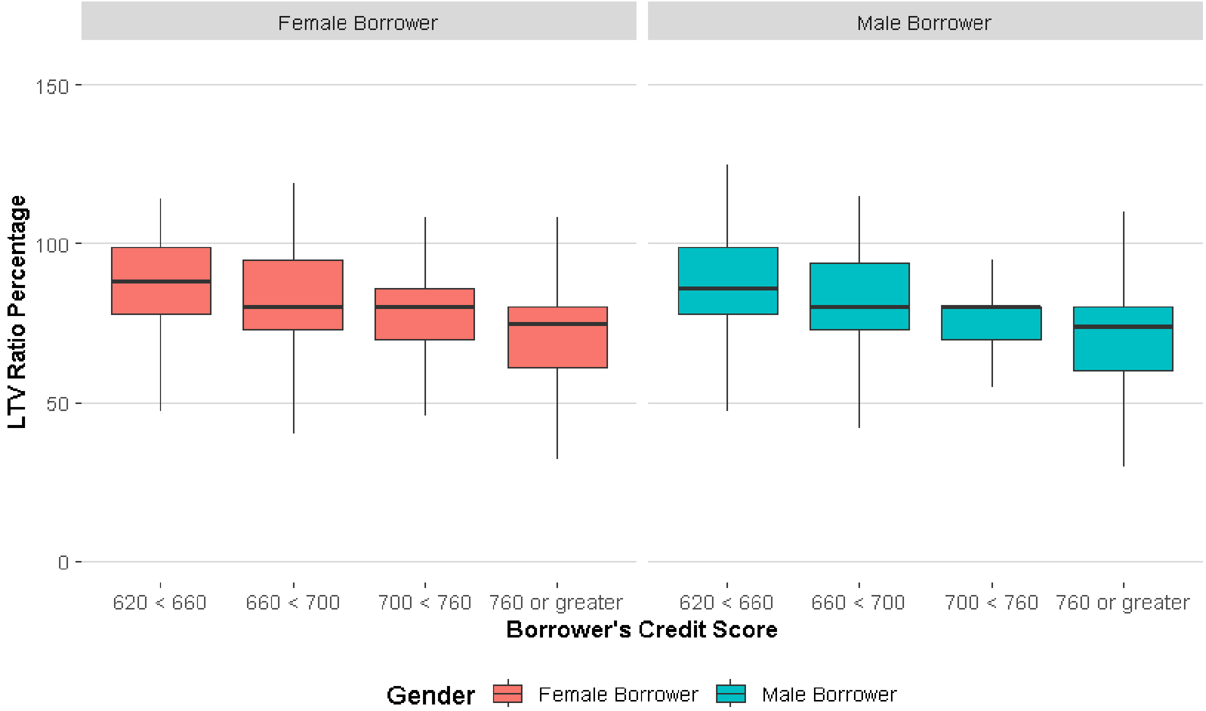
Figure 4: Comparison of LTV % by Borrower's Credit Score



Comparing LTV Based On Credit Scores (Female & Male Borrowers)

We have compared the median LTV values for non-white and white borrowers in different credit score ranges. Again, we will take the same procedures to find the median LTV, but this time compare them among male and female borrowers with different credit scores. After running our group statistics, we are able to find that the biggest difference in the median LTV values is in the credit score range of 620 < 660. Female borrowers with a credit score range of 620 < 660 had a median LTV of 88.27% and the median LTV value for male borrowers in the same credit score range was 86.03%. There was no difference in the median LTV for borrowers in the credit score range of 660 < 700 and 700 < 760. However, female borrowers in the credit score range of 760 or above had a median LTV of 75% compared to male borrowers who had a median LTV of 74% in the same credit score range. In Figure 5 the box plots for male and female borrower looks quite similar compared to the box plots in Figure 4. Furthermore, Figure 5 shows the same trend we saw when we compared non-white and white borrowers. For male and female borrowers LTV decreases as credit scores increase. This, however, is not the case for each credit score range. In the credit score ranges of 660 < 700 and 700 < 760 the median LTV remains the same for both groups. The significant takeaway from this analysis is that there is a noticeable difference in the median LTV for the lowest credit score range when comparing males and females. This result was similar to what we examined with non-white and white borrowers; the only difference is that this change in the median LTV is greater when comparing borrowers by race. In the previous section we saw that for non-white borrowers in the lowest credit score range there median LTV was 95% where it was 85% for white borrowers. This difference is not as big when we compare the median LTV of female (88.27%) and male borrowers (86.03%) in the lowest credit score range. These insights may suggest that the differences in LTV based on credit scores is larger when comparing the borrower's race rather than their gender. It could be the case that racial characteristics play a bigger role in identifying the differences in lending risks rather than the borrower's gender.

Figure 5: Comparison of LTV % by Borrower's Credit Score



Location of High Risk and Low Risk Borrowers in the U.S. (2010)

Another important aspect of our analysis is gaining insight into where high-risk and low-risk borrowers are located. In the following sections we will examine which states in the U.S. have the most amount of high-risk and low-risk borrowers using the FHLAppend dataset. For this section we will only consider data for the years 2010 and 2020. Comparing the two time periods will allow us to determine any trends that we see over the 10-year period. We are essentially trying to identify which states harbor more high-risk or low-borrowers and to see if the number of these borrowers has changed between 2010 and 2020.

First, we will begin by looking at borrowers in 2010. In order to determine which states have the most amount of high-risk and low-risk borrowers, a table was created for each group. High-risk borrowers include individuals with a credit score range of 760 or above and low-risk borrowers include individuals with a credit score range of 620 < 660. By filtering the credit range and grouping the borrower count by state we were able to produce a simple table of the top 10 states with the highest borrower count for each credit category. Table 5 shows that in 2010 the state with the largest number of low-risk borrowers was Iowa. In Iowa there were a total of 3602 borrowers with a credit score of 760 or above. Kansas came in second with 2762 borrowers and Nebraska was third with a total of 2420 low-risk borrowers. As for the largest number of high-risk borrowers in 2010, Kansas came in first. Table 6 reveals that Kansas had 194 high-risk borrowers, Ohio had 156, and Iowa with 143. Comparing both tables it is easy to see that quite a few states are included in both the credit categories. In fact, the states of Iowa, Kansas, and Nebraska, which have the largest amount of low-risk borrowers were also on the top 10 list of states with the largest amount of high-risk borrowers. The only states that were not included in both tables were Massachusetts and Michigan.

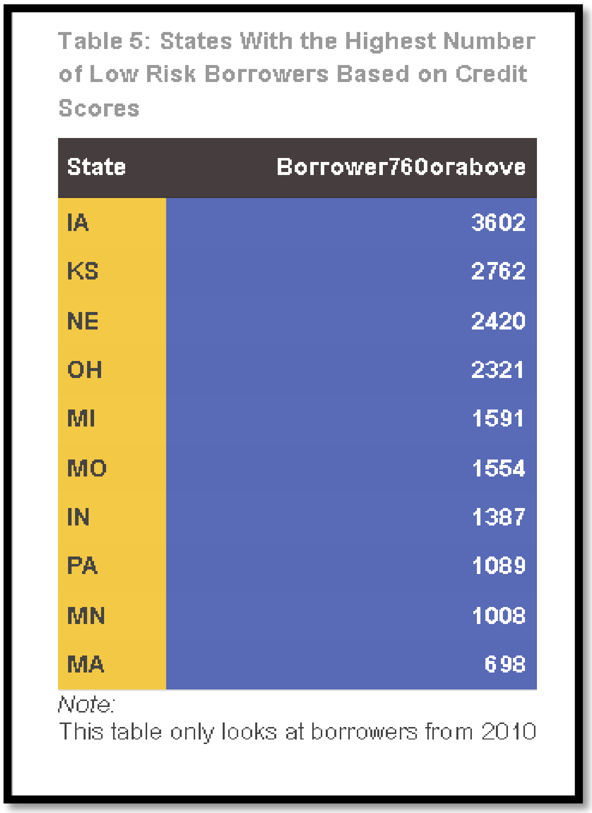


Table 5: States With the Highest Number of Low Risk Borrowers Based on Credit Scores

State	Borrower760orabove
IA	3602
KS	2762
NE	2420
OH	2321
MI	1591
MO	1554
IN	1387
PA	1089
MN	1008
MA	698

Note:
This table only looks at borrowers from 2010

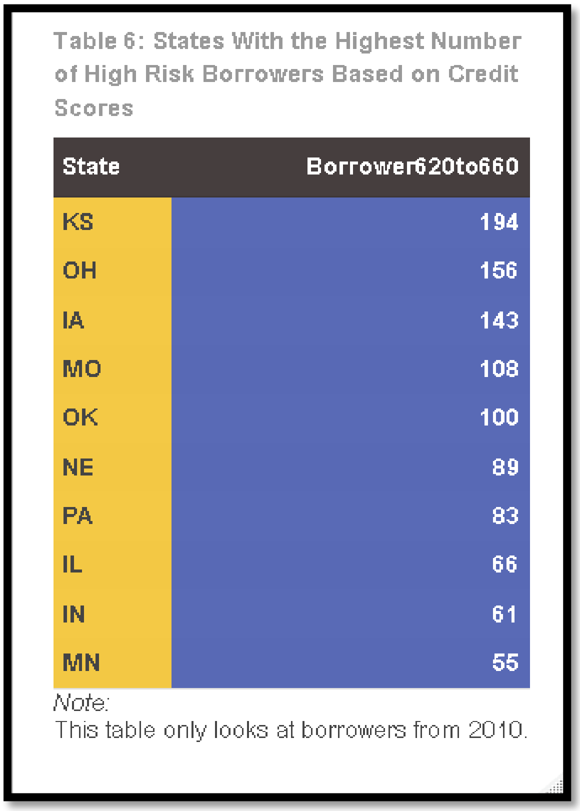


Table 6: States With the Highest Number of High Risk Borrowers Based on Credit Scores

State	Borrower620to660
KS	194
OH	156
IA	143
MO	108
OK	100
NE	89
PA	83
IL	66
IN	61
MN	55

Note:
This table only looks at borrowers from 2010.

Location of High Risk and Low Risk Borrowers in the U.S. (2020)

We will now evaluate the number of high-risk and low-risk borrowers in 2020. The procedures to get our table for each credit category is the same as before. The only thing different is that we filter our dataset based on the year 2020. Table 7 shows that the greatest amount of low-risk borrowers in 2020 were located in Wisconsin with 6405 borrowers. Ohio came in second with 5579 borrowers and Indiana came in third with a total of 5280 borrowers. The greatest amount of high-risk borrowers were located in the state of Illinois. In Table 8 it appears that Illinois had a total of 428 high-risk borrowers, Wisconsin with 345, and Missouri with 179. Just as before quite a few states are included in each of the credit categories. Ohio, Indiana, Michigan, Texas, Colorado, and Minnesota were not included in Tables 7 and 8. From our analysis, one thing that stands out is the discrepancy between the number of high-risk and-low risk borrowers for both years. For most states, the number of borrowers with a credit score of 760 or above is in the thousands for 2010 and 2020. The number of borrowers with a credit score between 620 and 660 is in the hundreds in each period. This makes sense as institutions use credit scores to gauge whether or not they approve certain loans. In general, the lower the credit score the less likely it is for the loan to be approved. In our dataset all the mortgages were approved by Fannie Mae and Freddie Mac. Fannie and Freddie have a minimum credit score requirement of 620 for fixed-rate mortgages (Johnson, 2024). This credit score requirement, however, is not the same for all institutions. Even though the number of loans that were approved for borrowers with a low credit score may seem small, it might be the case that this number is actually big compared to loans that are approved by other institutions.

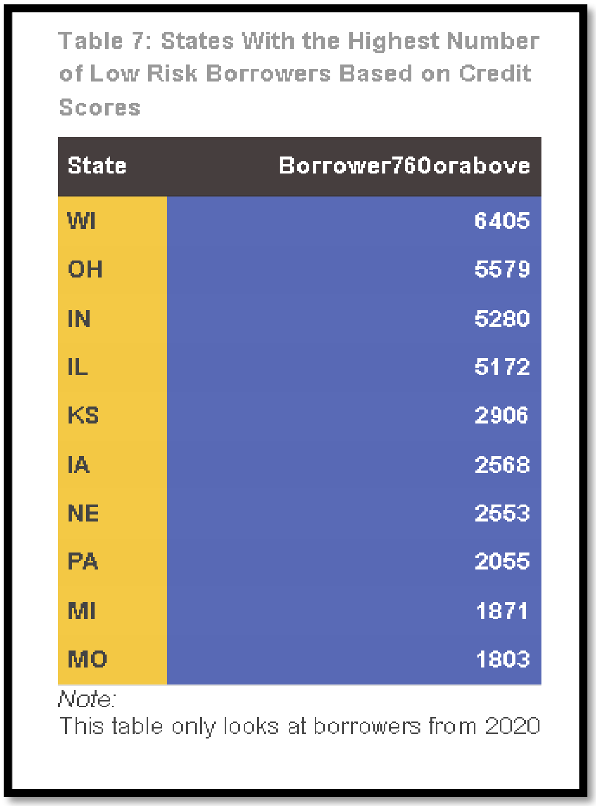


Table 7: States With the Highest Number of Low Risk Borrowers Based on Credit Scores

State	Borrower760orabove
WI	6405
OH	5579
IN	5280
IL	5172
KS	2906
IA	2568
NE	2553
PA	2055
MI	1871
MO	1803

Note:
This table only looks at borrowers from 2020

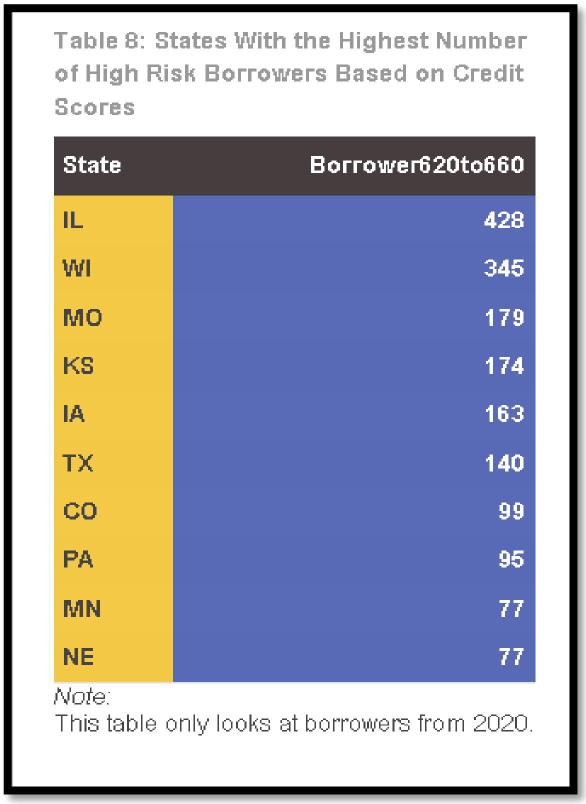


Table 8: States With the Highest Number of High Risk Borrowers Based on Credit Scores

State	Borrower620to660
IL	428
WI	345
MO	179
KS	174
IA	163
TX	140
CO	99
PA	95
MN	77
NE	77

Note:
This table only looks at borrowers from 2020.

Trends in Geographic Location (2010 & 2020)

Form looking at our previous tables, it seems that the number of high-risk and low-risk borrower has drastically increased within the 10-year period. In 2010 the greatest number of low-risk borrowers were located in Iowa (3602) and the largest number of high-risk borrowers were located in Kansas (194). Although these states are still included in 2020, they are not in the first spot anymore. In 2020 Wisconsin had the greatest number of low-risk borrowers (6405) and Illinois had the greatest number of high-risk borrowers (428). Wisconsin was not included at all in our tables for 2010. Comparing 2010 and 2020, the state of Wisconsin has seen a massive surge in mortgages that were approved for high-risk and low-risk borrowers. Other states that were not included in the top 10 list in 2010 were Colorado and Texas. Comparing Table 4 and 6 the states of Colorado and Texas have seen a significant increase in the number of high-risk borrowers. In order to better represent these trends, visual maps have been provided to see the trends in high risk and low risk borrowers for each year. Figure 6 and 7 provide us insight into the location of low-risk borrowers in 2010 and 2020. Regionally most low risk borrowers seem to be located in the Midwest in 2010 and 2020. Over the 10-year period the number of low-risk borrowers has increased by a significant amount in the region. Figure 8 and 9 gives us visual information regarding the location of high-risk borrowers in 2010 and 2020. In 2010 most low-risk borrowers are concentrated in the midwestern region of the U.S. In 2020 , however, this trend has slightly changed. We now see an increase in low-risk borrowers that are outside the Midwest region. States like Colorado and Texas have seen notable increase in the number of low-risk borrowers. One thing to note is that borrowers with high and low credit scores were not accounted for in every state in each year. Data was not accounted for high-risk borrowers in 2010 in the states of Wyoming, New Mexico, Arkansas, Louisiana, Alaska, and Hawaii. There was no data for high-risk borrowers in 2020 for Montana, Tennessee, Alaska, and Hawaii. No data was available for low-risk borrowers in 2010 for Mississippi, Alaska, and Hawaii. Lastly no data was available for low-risk borrowers in 2020 for the state of Alaska.

Figure 6: Borrowers With 760 or Above Credit Score in U.S. (2010)

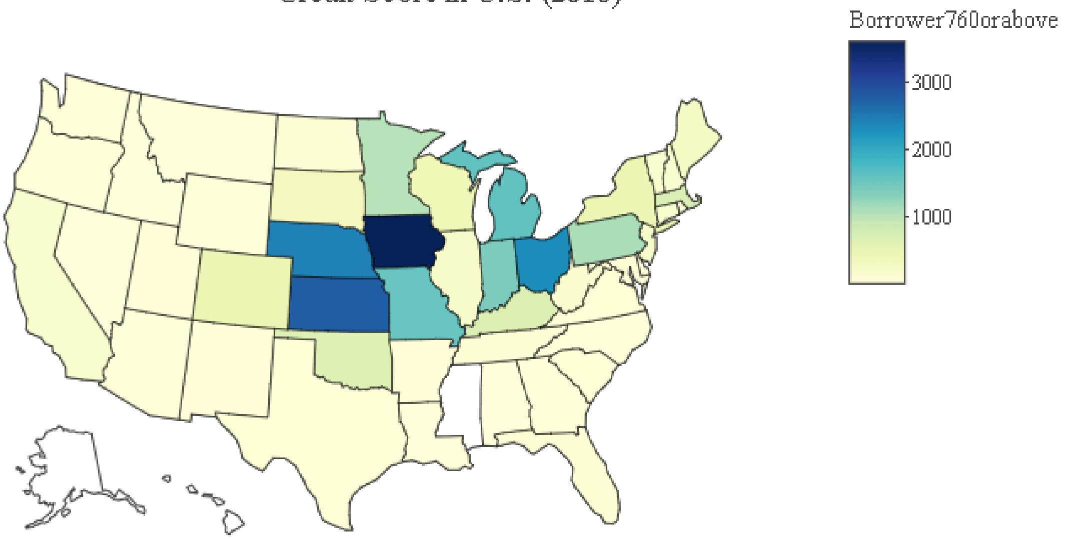


Figure 7: Borrowers With 760 or Above Credit Score in U.S. (2020)

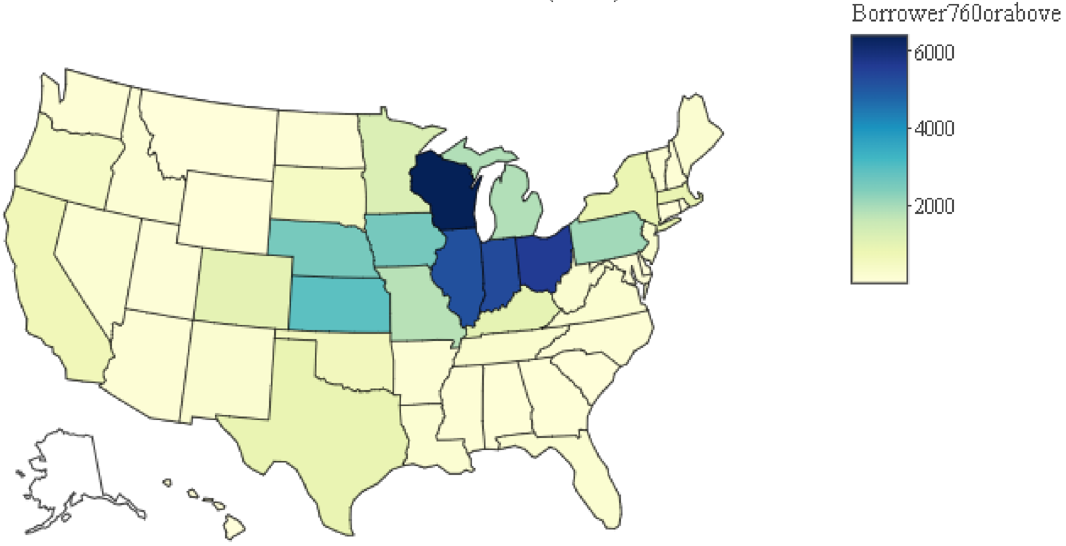


Figure 8: Borrowers with Credit Score Between 620 & 660 in the U.S. (2010)

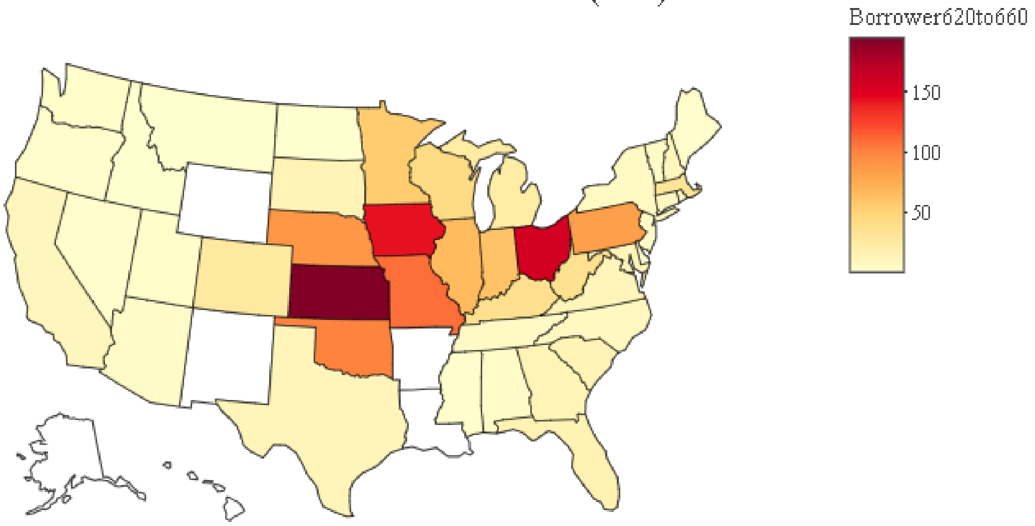
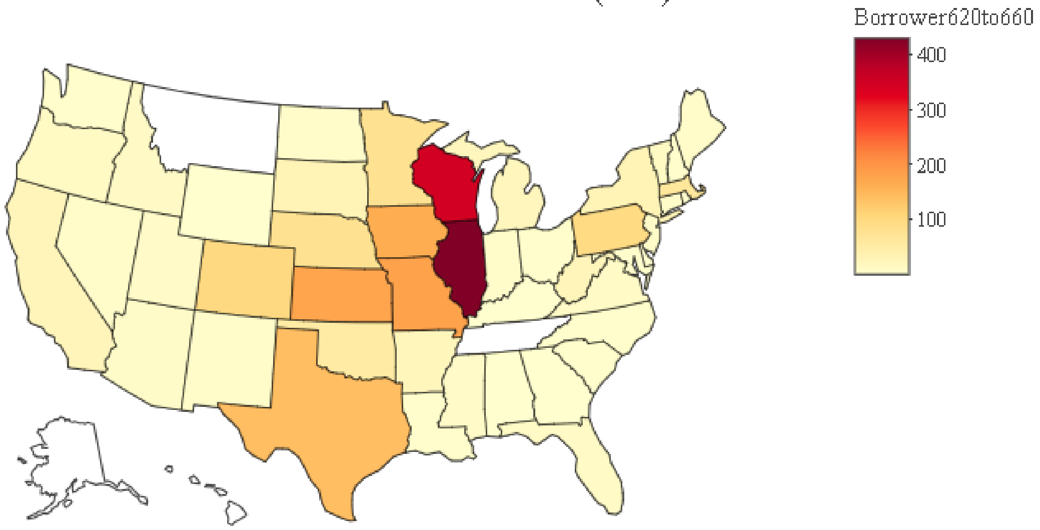


Figure 9: Borrowers with Credit Score Between 620 & 660 in the U.S. (2020)



Analyzing Regression Results

The last part of our analysis involves analyzing some regression models to identify the differences in lending risks of different types of borrowers. The four different models will aim to see how a borrower's race or self-employed status effect their lending risk. The purpose of creating these models is to further investigate the relationship between certain variables and to see if certain borrowers are more or less likely to have higher LTV values. Regression models not only provide us a way to understand the relationship between variables, but they also give us details about the statistical significance of the variables we use. Finding statistical significance in our variables will allow us to determine whether these relationships are attributable to a specific cause or happen by random chance.

Differences In LTV Between Black & White Borrowers

In our first two models we will evaluate the differences in LTV for white and black borrowers. The controls for the two models were the same and included monthly income ($1000s dollars), borrower's age, first-time home buyer, HUD median income ($1000s dollars), and note amount. Our dependent variable was LTV and our variables of interest are the dummies for black and white borrowers. After we run our regression, we get our estimates from our models which are provided in Table 9. If the borrower happens to be white, their LTV decreases by 0.53% holding other factors fixed. Borrowers who are black see an increase in LTV by 8.38% holding other factors fixed. The results in our model show that white borrowers have lower lending risks (LTV) compared to black borrowers. Our estimates for each borrower dummy are also statistically significant at the 1% level. This means that there is a 1% chance that these results happened by random chance.

Differences In LTV Based on Self Employment

For our last two models we will assess the differences in LTV for self-employed white and black borrowers. As discussed previously in our report, we specifically created these interaction terms to see if self-employed borrowers have higher or lower LTVs. Here not only are we looking at self-employment as a factor for determining LTV, but also the differences among self-employed individuals from different racial demographic groups. Unlike our previous models we will not be using the same number of controls. The third and fourth regression models will each have 3 controls, which include monthly income ($1000 dollars), borrower's age, and first-time home buyer. Again, our dependent variable (LTV) remains the same and our variables of interest are the dummies for self-employed black and white borrowers. Looking at Table 9, we can see that there is a noticeable difference in the estimates of our variables of interest. If the individual happens to be a self-employed white borrower, their LTV decreases by 2.84%. Self-employed black borrowers see an increase in their LTV by 5.11%. These results show that self-employed white borrowers have lower lending risks compared to self-employed black borrowers. Again, our estimates for our self-employed dummy variables are statistically significant at the 1% level.

Table 9: Changes to LTV Based on Race & Self-Employment Status

	Regression 1 White Borrowers	**Regression 2** Black Borrowers	**Regression 3** Self Employed White Borrowers	**Regression 4** Self Employed Black Borrowers
(Intercept)	95.378***	95.137***	88.707***	88.589***
	(0.275)	(0.236)	(0.143)	(0.143)
MonthlyIncome1000s	-0.065***	-0.064***	-0.026***	-0.031***
	(0.002)	(0.002)	(0.002)	(0.002)
BorrowerAge	-0.358***	-0.361***	-0.372***	-0.375***
	(0.003)	(0.003)	(0.003)	(0.003)
WhiteBorrower	-0.532***			
	(0.125)			
BorrowerFirstTimeHomebuyer	14.180***	14.047***	12.926***	13.075***
	(0.121)	(0.121)	(0.122)	(0.122)
NoteAmount1000s	0.027***	0.027***		
	(0.000)	(0.000)		
HUDMedianIncomeAmount1000s	-0.170***	-0.174***		
	(0.003)	(0.003)		
BlackBorrower		8.375***		
		(0.310)		
SelfEmployedWhiteBorrower			-2.839***	
			(0.124)	
SelfEmployedBlackBorrower				5.111***
				(1.482)
Num.Obs.	159452	159452	159452	159452
R2	0.232	0.235	0.204	0.202
F	8009.915	8164.187	10239.558	10080.390
RMSE	14.46	14.43	14.72	14.74

Variables of interest are highlighted and their coefficients are in red.

* p < 0.1, ** p < 0.05, *** p < 0.01

Additional Details Regarding Regression Models

In this situation it is not a good idea to compare models 1 and 2 with models 3 and 4. The implications of the models are different in each case. The first two models look at the effects on LTV if the borrower is either white or black. The other two models not only look at the effects on LTV based on the borrower's race, but also based on their self-employment status. Furthermore, the models do not have the same number of controls. The reason this is important is because our estimates for our variables of interest will be different depending upon how many independent variables we use. Models 1 and 2 have the same number of controls, but Models 3 and 4 do not. It is ,however, interesting to see that in each case white borrowers saw a decrease in their LTV whereas black borrowers saw an increase. One thing to keep in mind is that we arrived at our results using a simple multiple regression for each model. There are other assumptions and statistical analysis we have to consider before being certain that our estimates are not biased. These techniques and assumptions, however, will not be considered for this part of the analysis.

Conclusion

With our analysis, we were able successfully answer our original research questions that were presented in the beginning of this report. We were able to find that there are differences in lending risks of borrowers depending on their race or gender. By evaluating key relationships between certain variables, we found that non-white borrowers have higher LTV values compared to white borrowers. Further regression analysis was also conducted to compare white borrowers and black borrowers. The results provided us with estimates that indicate that black borrowers had higher LTV values compared to white borrowers. Not only did we compare borrowers by race, but also by their gender. We discovered that female borrowers with a low credit score had a higher median LTV compared to male borrowers with the same credit score. This was the case with non-white and white borrowers as well. Non-white borrowers with a low credit score had a higher median LTV compared to white borrowers with the same credit score. This difference in median LTV , however, was significantly bigger when we compared borrowers by race rather than gender. Finally, in our analysis we were able to discover some insight as to where low-risk and high-risk borrowers were located. It seems that both types of borrowers are predominantly located in the Midwest. When comparing between the two periods (2010 and 2020), we found that the number of high risk and low risk borrowers during the 10-year period had drastically increased. Although trends showed that most these types of borrowers were located in the Midwest, these trends were slightly different when considering high risk borrowers. States like Texas, Colorado, and Wisconsin saw a notable rise in high-risk borrowers for 2020 compared to 2010.

New Questions

From our analysis we can pose new questions to find additional insights and trends. What does the trend for high-risk and low-risk borrowers look like in 2010, 2015, and 2020? In this report we only compare the two time periods of 2010 and 2020, but what about using the data from 2015 to evaluate the differences in each year? Instead of grouping all minorities as non-whites, what if we compared the LTV of all the races? Is there a significant difference in LTV across White, Black, Asian, Hispanic, and Native American Borrowers? Which borrowers have the highest LTV value across these racial groups? Lastly, we can look at the disparity in LTV based on both race and gender. In this report we compare male and female borrowers, but we did not assess the differences in LTV for different types of male and female borrowers. What about addressing the disparities between black female/male borrowers or white female/male borrowers?

Implications of Results

The results prove that there are differences in lending risk based on certain characteristics of borrowers and that high risk and low risk borrowers are located in specific parts of the U.S. We could dive deeper into understanding how race and gender play a role in the approval of loans. Are there certain requirements that are making it harder for non-white or female borrowers to get a mortgage approved? These implications could mean addressing policy actions on a state-by-state basis rather than federally. We have seen that many high-risk and low-risk borrowers are located in the Midwest. States in this region could enact policies to address these issues or at the very least investigate the meaning behind these trends. Additionally, perhaps particular methods used in evaluating lending risks are flawed. The loan to value ratio is often used in mortgage lending, but there are other ways to assess the risk of borrowers. Our analysis revealed that non-white-borrowers and female borrowers in the lowest credit score range had a higher median LTV value compared to white borrowers and male borrowers in the same range. Institutions also use credit score in evaluating a borrower's risk, so the fact that there is a difference in LTV when credit scores are in the same range is a bit concerning.

Limitations

The main limitation in our report was that the data we used was from one main source, which was the FHFA website. The FHFA provides a public use database where the data is provided by Fannie Mae and Freddie Mac. In this report we are only looking at mortgages/loans that were approved by these institutions. There are many other commercial banks and institutions that provide loans to borrowers. We are missing the data we could potentially get from borrowers who are getting their mortgages from different institutions. The only problem with this is that it is very difficult to get this information due to privacy concerns. Other major institutions are independently run as for-profit companies with little to no control from government agencies. This means that they do not have an obligation to release a public use database like the FHFA.

Another limitation is that the data only consists of borrowers whose loans were approved. It would be better to have a mixture of approved and non-approved loans in the dataset. This would allow us to get a better understanding of the main reasoning behind the differences in lending risks.

Next Steps/Future Work

Our next steps should involve expanding our data set to include borrowers who took our loans from other institutions. The dataset only consisted of borrowers whose mortgages were approved by Fannie Mae and Freddie Mac. For our research purposes it would be beneficial to get data from other mortgage lenders like U.S. Bank, Chase, PNC Bank, and other institutions if we can. This way our dataset will be comprehensive and encompass many different borrowers. Furthermore, adding variables that are outside our original data set can prove to be useful as it can allow us to conduct different types of analyses. Variables that look at unemployment, wages, GDP, CPI, and population on a state-by-state basis allows us to find new relationships associated with lending risks (LTV). In general, additional factors should be included in our dataset from different sources to make sure that relevant variables are not excluded in our analysis.

Appendix

```r
library(naniar)

library(readxl)
library(haven)
library(plyr)
library(tidyverse)

library(cdlTools)
library(ggplot2)
library(ggthemes)
library(htmltools)
library(ggmap)

library(knitr)
library(vtable)
library(modelsummary)

library(viridisLite)
library(viridis)
library(plotly)

library(car)

library(gt)

setwd("C:/Users/achar/Desktop/R HW Problems")

FHL2010_1 <- read_excel("2010_Part1.xlsx")

dim(FHL2010_1)

FHL2010_2 <- read_excel("2010_Part2.xlsx")

dim(FHL2010_2)


# Getting rid of unnecessary variables in the second dataset for 2010
# so that duplicates aren't created when merging.

FHL2010_2 <- select(FHL2010_2, -c(Year, FHLBankID, Program, FIPSStateC
ode, FIPSCountyCode,
          MSA, FeatureID))

# Merging the FHL 2010 datasets
```

```r
FHL2010 <- merge(FHL2010_1, FHL2010_2, by = "Loan Number")
```

```r
# Creating Analytic Dataset with relevant variables
# for 2010
```

```r
FHL2010 <- FHL2010 %>%
  select(Year, FHLBankID, Tract, FIPSStateCode, FIPSCountyCode,
         MinPer, TraMedY, CurAreY, LocMedY, LTV, Purpose, FedGuar, Fro
nt,
         Back, CoRace, BoRace, CoGender, BoGender, Occup,
        `Borrower Credit Score`,`Co-Borrower Credit Score`,
         Self, HOEPA, Amount, Rate, PropType, NumUnits,
         Income, First, BoAge, CoAge, MortDate, AcquDate,
        NumBor)
```

```r
# Importing Federal Home Loan 2015 Files
```

```r
FHL2015_1 <- read_excel("2015_Part1.xlsx")
```

```r
dim(FHL2015_1)
```

```r
FHL2015_2 <- read_excel("2015_Part2.xlsx")
```

```r
dim(FHL2015_2)
```

```r
# Getting rid of unnecessary variables in the second dataset for 2015
# so that duplicates aren't created when merging.
```

```r
FHL2015_2 <- select(FHL2015_2, -c(Year, FHLBank, FIPSStateCode,
                                  FIPSCountyCode, MSA, FeatureID))
```

```r
# Merging the FHL 2015 datasets
```

```r
FHL2015 <- merge(FHL2015_1, FHL2015_2, by = "AssignedID")
```

```r
# Creating Analytic Dataset with relevant variables for 2015
```

```r
FHL2015 <- FHL2015 %>%
  select(Year, FHLBank,Tract, FIPSStateCode, FIPSCountyCode,
         MinPer, TraMedY, CurAreY, LocMedY, LTV, Purpose,
         FedGuar, Front, Back, CoRace, BoRace, CoGender,
         BoGender, Occup, BoCreditScor, CoCreditScor, Self,
         HOEPA, Amount, Rate, PropType, NumUnits,
         Income, First, BoAge, CoAge, MortDate, AcqDate,
         NumBor)


# Importing Federal Home Loan 2020 Files

FHL2020 <- read_excel("2020_Dataset.xlsx")

dim(FHL2020)

FHL2020 <- FHL2020 %>%
  select(Year,Bank, CensusTractIdentifier,FIPSStateNumericCode,
         FIPSCountyCode,CensusTractMinorityRatioPercent
         ,CensusTractMedFamIncomeAmount, HUDMedianIncomeAmount,
         LocalAreaMedianIncomeAmount, LTVRatioPercent,LoanPurposeType,
         MortgageType, HousingExpenseRatioPercent, TotalDebtExpenseRat
ioPercent,
         Borrower2Race1Type, Borrower1Race1Type, Borrower2GenderType,
         Borrower1GenderType,PropertyUsageType, Borrower1CreditScoreVa
lue,
         Borrower2CreditScoreValue, EmploymentBorrowerSelfEmployed,
         HOEPALoanStatusType, NoteAmount, NoteRatePercent, PropertyTyp
e,
         PropertyUnitCount, TotalMonthlyIncomeAmount, BorrowerFirstTim
eHomebuyer,
         Borrower1AgeAtApplicationYears, Borrower2AgeAtApplicationYear
s,
         NoteDate, LoanAcquistionDate, BorrowerCount)


# Renaming variables to be consistent in each dataset for the years.

FHL2020 <- FHL2020 %>%
  rename(FHLBankDistrict = Bank,
         FIPSStateCode = FIPSStateNumericCode,
         CoBorrowerRace = Borrower2Race1Type,
         BorrowerRace = Borrower1Race1Type,
         CoBorrowerGender = Borrower2GenderType,
         BorrowerGender = Borrower1GenderType,
         BorrowerCreditScore = Borrower1CreditScoreValue,
         CoBorrowerCreditScore = Borrower2CreditScoreValue,
         BorrowerSelfEmployed = EmploymentBorrowerSelfEmployed,
         BorrowerAge = Borrower1AgeAtApplicationYears,
         CoBorrowerAge = Borrower2AgeAtApplicationYears,
         MortgageOriginated = NoteDate,
         MortgageAcquired = LoanAcquistionDate)


FHL2015 <- FHL2015 %>%
  rename(FHLBankDistrict = FHLBank,
         CensusTractIdentifier = Tract,
         CensusTractMinorityRatioPercent = MinPer,
         CensusTractMedFamIncomeAmount = TraMedY,
         HUDMedianIncomeAmount = CurAreY,
         LocalAreaMedianIncomeAmount = LocMedY,
         LTVRatioPercent = LTV,
         LoanPurposeType = Purpose,
         MortgageType = FedGuar,
         HousingExpenseRatioPercent = Front,
         TotalDebtExpenseRatioPercent = Back,
         CoBorrowerRace = CoRace,
         BorrowerRace = BoRace,
         CoBorrowerGender = CoGender,
         BorrowerGender = BoGender,
         PropertyUsageType = Occup,
         BorrowerCreditScore = BoCreditScor,
         CoBorrowerCreditScore = CoCreditScor,
         BorrowerSelfEmployed = Self,
         HOEPALoanStatusType = HOEPA,
         NoteAmount = Amount,
         NoteRatePercent = Rate,
         PropertyType = PropType,
         PropertyUnitCount = NumUnits,
         TotalMonthlyIncomeAmount = Income,
         BorrowerFirstTimeHomebuyer = First,
         BorrowerAge = BoAge,
         CoBorrowerAge = CoAge,
         MortgageOriginated = MortDate,
         MortgageAcquired = AcqDate,
         BorrowerCount = NumBor)

FHL2010 <- FHL2010 %>%
  rename(FHLBankDistrict = FHLBankID,
         CensusTractIdentifier = Tract,
         CensusTractMinorityRatioPercent = MinPer,
         CensusTractMedFamIncomeAmount = TraMedY,
         HUDMedianIncomeAmount = CurAreY,
         LocalAreaMedianIncomeAmount = LocMedY,
         LTVRatioPercent = LTV,
         LoanPurposeType = Purpose,
         MortgageType = FedGuar,
         HousingExpenseRatioPercent = Front,
         TotalDebtExpenseRatioPercent = Back,
         CoBorrowerRace = CoRace,
         BorrowerRace = BoRace,
         CoBorrowerGender = CoGender,
         BorrowerGender = BoGender,
         PropertyUsageType = Occup,
         BorrowerCreditScore = `Borrower Credit Score`,
         CoBorrowerCreditScore = `Co-Borrower Credit Score`,
         BorrowerSelfEmployed = Self,
         HOEPALoanStatusType = HOEPA,
         NoteAmount = Amount,
         NoteRatePercent = Rate,
         PropertyType = PropType,
         PropertyUnitCount = NumUnits,
         TotalMonthlyIncomeAmount = Income,
         BorrowerFirstTimeHomebuyer = First,
         BorrowerAge = BoAge,
         CoBorrowerAge = CoAge,
         MortgageOriginated = MortDate,
         MortgageAcquired = AcquDate,
         BorrowerCount = NumBor)
```

```
# Converting/Transforming 2010 and 2015 variables for consistency acro
ss all datasets for each year
```

```
FHL2010 <- FHL2010 %>%
  mutate(LTVRatioPercent = LTVRatioPercent * 100,
         HousingExpenseRatioPercent = HousingExpenseRatioPercent * 100
,
         TotalDebtExpenseRatioPercent = TotalDebtExpenseRatioPercent *
100,
         NoteRatePercent = NoteRatePercent * 100)
```

```
FHL2015 <- FHL2015 %>%
  mutate(LTVRatioPercent = LTVRatioPercent * 100,
         HousingExpenseRatioPercent = HousingExpenseRatioPercent * 100
,
         TotalDebtExpenseRatioPercent = TotalDebtExpenseRatioPercent *
100,
         NoteRatePercent = NoteRatePercent * 100)
```

```r
FHL2010$TotalMonthlyIncomeAmount <- round((FHL2010$TotalMonthlyIncomeAmount/12), digits = 0)

FHL2015$TotalMonthlyIncomeAmount <- round((FHL2015$TotalMonthlyIncomeAmount/12), digits = 0)


# Appending the datasets for each year

FHLAppend <- rbind(FHL2010, FHL2015, FHL2020)

dim(FHLAppend)

# Creating additional dummy variables

FHLAppend$BorrowerSelfEmployed <- ifelse(FHLAppend$BorrowerSelfEmployed == 1, 1,0)

FHLAppend$BorrowerFirstTimeHomebuyer <- ifelse(FHLAppend$BorrowerFirstTimeHomebuyer == 1, 1,0)

FHLAppend <- FHLAppend %>%
  mutate(WhiteBorrower = ifelse(BorrowerRace == 5, 1,0)) %>%
  mutate(WhiteCoBorrower = ifelse(CoBorrowerRace == 5, 1,0)) %>%
  mutate(FemaleBorrower = ifelse(BorrowerGender == 2, 1,0)) %>%
  mutate(FemaleCoBorrower = ifelse(CoBorrowerGender == 2, 1,0)) %>%
  mutate(BlackBorrower = ifelse(BorrowerRace == 3, 1,0)) %>%
  mutate(BlackCoBorrower = ifelse(CoBorrowerRace == 3, 1,0)) %>%
  mutate(Borrower620to660 = ifelse(BorrowerCreditScore == 2, 1,0)) %>%
  mutate(Borrower660to700 = ifelse(BorrowerCreditScore == 3, 1,0)) %>%
  mutate(Borrower700to760 = ifelse(BorrowerCreditScore == 4, 1,0)) %>%
  mutate(Borrower760orabove = ifelse(BorrowerCreditScore == 5, 1,0)) %>%
  mutate(CoBorrower620to660 = ifelse(CoBorrowerCreditScore == 2, 1,0)) %>%
  mutate(CoBorrower660to700 = ifelse(CoBorrowerCreditScore == 3, 1,0)) %>%
  mutate(CoBorrower700to760 = ifelse(CoBorrowerCreditScore == 4, 1,0)) %>%
  mutate(CoBorrower760orabove = ifelse(CoBorrowerCreditScore == 5, 1,0))


View(FHLAppend)

# Creating Interaction variables
```

```r
FHLAppend <- FHLAppend %>%
  mutate(SelfEmployedWhiteBorrower = BorrowerSelfEmployed * WhiteBorrower) %>%
  mutate(SelfEmployedFemaleBorrower = BorrowerSelfEmployed * FemaleBorrower) %>%
  mutate(SelfEmployedBlackBorrower  = BorrowerSelfEmployed * BlackBorrower)
```

```r
# Only looking at 1 or 2 total Borrowers for loan
```

```r
FHLAppend <- FHLAppend %>%
  filter(BorrowerCount <= 2)
```

```r
# Only looking at female and male borrowers
```

```r
FHLAppend <- FHLAppend %>%
  filter(BorrowerGender %in% c(1,2))
```

```r
# Replacing categorical numerical values with NA values to keep relevant data
```

```r
FHLAppend <- FHLAppend %>%
  replace_with_na(replace = list(BorrowerAge = c(99,999,998),
                                 CoBorrowerAge = c(999,99,998,98)))
```

```r
# Getting rid of outliers for Housing Expense Ratio
```

```r
quartiles_houseexp <- quantile(FHLAppend$HousingExpenseRatioPercent,
                               probs = c(.25,.75),
                               na.rm = F)
```

```r
IQR_houseexp <- IQR(FHLAppend$HousingExpenseRatioPercent)
```

```r
Lower_houseexp <- quartiles_houseexp[1] - 1.5*IQR_houseexp
Upper_houseexp <- quartiles_houseexp[2] + 1.5*IQR_houseexp
```

```r
FHLAppend <- subset(FHLAppend,FHLAppend$HousingExpenseRatioPercent > Lower_houseexp
                    & FHLAppend$HousingExpenseRatioPercent < Upper_hou
seexp)

dim(FHLAppend)



# Visualization (Graphs, Figure, Charts, etc.)

# Using a Sample of the dataset to create scatter plots

FHLAppend_Sample <- sample_n(FHLAppend, 10000)


View(FHLAppend_Sample)

# Creating a Scatter Plot of LTV vs. Note Amount for
# White and Non-White Borrowers


Scatter1 <-  FHLAppend_Sample %>%
  mutate(WhiteBorrower = dplyr::recode(WhiteBorrower, `1` = "WhiteBorr
ower",
                                `0` = "NonWhiteBorrower")) %>%
  rename(Race = WhiteBorrower) %>%
  ggplot(aes(x = LTVRatioPercent,
             y = NoteAmount))+
  geom_point(shape = "square",
             size = 2,
             alpha = .4,
             color = "red")+
  geom_smooth(method = "lm", se = F)+
  labs(title = "Figure 2: Comparison of LTV & Note Amount by Borrower'
s Race",
       caption = "Correlation Coefficient (White Borrowers) = 0.095
       Correlation Coefficient (Non White Borrowers) = -0.11 \nThis Sc
atter Plot was created using a random sample of
       10000 observations from the FHLAppend dataset.")+
  facet_wrap(~Race)+
  scale_y_continuous(labels = function(y) format(y,
                                                 scientific = F))+
  theme_bw()+
  theme(plot.title = element_text(size = 12, face = "bold", hjust = 0.
5),
        axis.title.x = element_text(size = 11, face = "bold"),
        axis.title.y = element_text(size = 11, face = "bold"),
        plot.caption = element_text(size = 10, face  = "italic"))
```

```
Scatter1
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
# Comparing Correlation Coefficients of White and Non White Borrowers

FHLAppendWhite <- FHLAppend_Sample %>%
  filter(WhiteBorrower == 1 )


cor(FHLAppendWhite$LTVRatioPercent, FHLAppendWhite$NoteAmount)


FHLAppendNonWhite <- FHLAppend_Sample %>%
  filter(WhiteBorrower == 0)

cor(FHLAppendNonWhite$LTVRatioPercent, FHLAppendNonWhite$NoteAmount)


# Creating a Scatter Plot of LTV vs. Note Amount for Female and
# Male Borrowers

Scatter2 <-  FHLAppend_Sample %>%
  mutate(FemaleBorrower = dplyr::recode(FemaleBorrower, `1` = "Female
Borrowers",
                                       `0` = "Male Borrowers")) %>%
  rename(Gender = FemaleBorrower) %>%
  ggplot(aes(x = LTVRatioPercent,
             y = NoteAmount))+
  geom_point(shape = "square",
             size = 2,
             alpha = .4,
             color = "royalblue3")+
  geom_smooth(method = "lm", se = F, color = "red")+
  labs(title = "Figure 3: Comparison of LTV & Note Amount by Borrower'
s Race",
       caption = "Correlation Coefficient (Female Borrowers) = 0.052
       Correlation Coefficient (Male Borrowers) = 0.096 \nThis Scatter
Plot was created using a random sample of
       10000 observations from the FHLAppend dataset.")+
  facet_wrap(~Gender)+
  scale_y_continuous(labels = function(y) format(y,
                                                  scientific = F))+
  theme_bw()+
  theme(plot.title = element_text(size = 12, face = "bold", hjust = 0.
5),
        axis.title.x = element_text(size = 11, face = "bold"),
        axis.title.y = element_text(size = 11, face = "bold"),
        plot.caption = element_text(size = 10, face = "italic"))
```

```
Scatter2
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
# Finding Correlation Coefficients for Male and Female Borrowers

FHLAppendFemale <- FHLAppend_Sample %>%
  filter(FemaleBorrower == 1)

cor(FHLAppendFemale$LTVRatioPercent, FHLAppendFemale$NoteAmount)

FHLAppendMale <- FHLAppend_Sample %>%
  filter(FemaleBorrower == 0)

cor(FHLAppendMale$LTVRatioPercent, FHLAppendMale$NoteAmount)


# Bar chart of Average LTV for the years (2010,2015,2020)

FHL_LTVavg <- data.frame(aggregate(FHLAppend$LTVRatioPercent,
                                   by = list(FHLAppend$Year),
                                   FUN = mean))


LTVavgplot <- ggplot(FHL_LTVavg, aes(factor(Group.1), x))+
  geom_bar(stat = "identity", color = "black",
           fill = "darkgreen")+
  geom_text(aes(label = round(signif(x),2))
            , vjust = -0.8, size = 3.5)+
  labs(x = "Year",
       y = "Average LTV (%)",
       title = "Figure 1: Changes in Average LTV Percentage By Year")+
  theme_clean()+
  theme(plot.title = element_text(size = 10, face = "bold", hjust = 0.
5),
        axis.title.x = element_text(size = 10),
        axis.title.y = element_text(size = 10),
        axis.text.x = element_text(size = 10),
        axis.text.y = element_text(size = 10))
```

```
LTVavgplot
```

```
# Calculating percent change for the values of average LTV in the
# bar chart

(74.84-72.12)/(72.12)

(72.25-74.84)/(74.84)
```

```
# Create Box Plot for White Borrowers and Non White Borrowers, compari
ng their
# credit score and LTV Ratio Percent


FHLAppend %>%
  filter(BorrowerCreditScore %in% c(2,3,4,5)) %>%
  mutate(WhiteBorrower = dplyr::recode(WhiteBorrower, `1` = "White Bor
rower",
                                `0` = "Non White Borrower")) %>%
  rename(Race = WhiteBorrower) %>%
  mutate(BorrowerCreditScore = dplyr::recode(BorrowerCreditScore, `2`
= "620 < 660",
                                      `3` = "660 < 700",
                                      `4` = "700 < 760",
                                      `5` = "760 or greater")) %>%
  ggplot(aes(x = BorrowerCreditScore,
             y = LTVRatioPercent,
             fill = Race))+
  geom_boxplot(outlier.shape = NA)+
  facet_wrap(~Race)+
  labs(title = "Figure 4: Comparison of LTV % by Borrower's Credit Sco
re",
       x = "Borrower's Credit Score",
       y = "LTV Ratio Percentage")+
  theme_hc()+
  theme(plot.title = element_text(size = 12, face = "bold", hjust = 0.
5),
        axis.title.x = element_text(size = 11, face = "bold"),
        axis.title.y = element_text(size = 11, face = "bold"))
```

```
# Finding the median LTV values of White and Non-White Borrowers Based
on their
```

```r
# credit score

FHLAppendWhite <- FHLAppend %>%
  filter(WhiteBorrower == 1)


aggregate(FHLAppendWhite$LTVRatioPercent,
          by = list(FHLAppendWhite$BorrowerCreditScore),
          FUN = median)

FHLAppendNonWhite <- FHLAppend %>%
  filter(WhiteBorrower == 0)

aggregate(FHLAppendNonWhite$LTVRatioPercent,
          by = list(FHLAppendNonWhite$BorrowerCreditScore),
          FUN = median)



# Create Box Plot for Male and Female Borrowers, comparing their credi
t
# score and LTV Ratio Percent

FHLAppend %>%
  filter(BorrowerCreditScore %in% c(2,3,4,5)) %>%
  mutate(FemaleBorrower = dplyr::recode(FemaleBorrower, `1` = "Female
Borrower",
                                `0` = "Male Borrower")) %>%
  rename(Gender = FemaleBorrower) %>%
  mutate(BorrowerCreditScore = dplyr::recode(BorrowerCreditScore, `2`
= "620 < 660",
                                      `3` = "660 < 700",
                                      `4` = "700 < 760",
                                      `5` = "760 or greater")) %>%
  ggplot(aes(x = BorrowerCreditScore,
             y = LTVRatioPercent,
             fill = Gender))+
  geom_boxplot(outlier.shape = NA)+
  facet_wrap(~Gender)+
  labs(title = "Figure 5: Comparison of LTV % by Borrower's Credit Sco
re",
       x = "Borrower's Credit Score",
       y = "LTV Ratio Percentage")+
  theme_hc()+
  theme(plot.title = element_text(size = 12, face = "bold", hjust = 0.
5),
        axis.title.x = element_text(size = 11, face = "bold"),
        axis.title.y = element_text(size = 11, face = "bold"))


# Finding the median LTV values of Female and Male Borrowers Based on
their
# credit score

FHLAppendFemale <- FHLAppend %>%
  filter(FemaleBorrower == 1)

aggregate(FHLAppendFemale$LTVRatioPercent,
          by = list(FHLAppendFemale$BorrowerCreditScore),
          FUN = median)


FHLAppendMale <- FHLAppend %>%
  filter(FemaleBorrower == 0)

aggregate(FHLAppendMale$LTVRatioPercent,
          by = list(FHLAppendMale$BorrowerCreditScore),
          FUN = median)



# Creating new datatset called FHLStateName

FHLStateName <- FHLAppend

# Renaming Variable name for the State Variable

FHLStateName <- FHLStateName %>%
  rename(State = FIPSStateCode)

FHLStateName2010 <- FHLStateName %>%
  filter(Year == 2010)

# Creating a dataset that counts the number of borrowers in each
# state with a credit score of 760 or above

FHL760credit_2010 <- FHLStateName2010 %>%
  filter(Borrower760orabove == 1) %>%
  group_by(State) %>%
  summarise(Borrower760orabove = n())

# Changing FIPScode into Abbreviation of State
```

```
FHL760credit_2010$State<- fips(FHL760credit_2010$State, to = 'Abbrevia
tion')
```

```
View(FHL760credit_2010)
```

```
# Creating a dataset/table for the Top 10 states who have the highest
# number of borrowers with a credit score of 760 or above in 2010
```

```
FHL760top10_2010 <- FHL760credit_2010 %>%
  arrange(desc(Borrower760orabove)) %>%
  slice(1:10)
```

```
View(FHL760top10_2010)
```

```
kable(x = FHL760top10_2010,
      caption = '<b>Table 3: States With the Highest Number of Low Ris
k
      Borrowers Based on Credit Scores</b>',
      digits = 2,
      format = 'html') %>%
  kable_paper("striped", full_width = F) %>%
  row_spec(0, bold = T, background = "#140A0ACA", color = "white") %>%
  column_spec(1, bold = T, background = "#F2C309D8") %>%
  column_spec(2, bold = T,color = "white",
              background = "#3E4AABD8") %>%
  kableExtra::footnote(general = "This table only looks at borrowers f
rom 2010")
```

```
# Creating a Map Chart too see where more low risk borrowers are for
```

```
# 2010
```

```
creditscoregraph1 <- plot_geo(FHL760credit_2010,
                             locationmode = 'USA-states') %>%
  add_trace(locations = ~State,
            z = ~Borrower760orabove,
            color = ~Borrower760orabove,
            colors = "YlGnBu",
            marker = list(line = list(
              width = 0.5,
              opacity = 0.5
            )
          )
        ) %>%
  layout(geo = list(scope = 'usa'),
         font = list(family = "DM Sans"),
         title = "Figure 6: Borrowers With 760 or Above \nCredit Score
in U.S. (2010)")


creditscoregraph1


# Creating a dataset that counts the number of borrowers in each state
# that have credit score between 620 & 660 in 2010


FHLcreditscore620to660_2010 <- FHLStateName2010 %>%
  filter(Borrower620to660 == 1) %>%
  group_by(State) %>%
  summarise(Borrower620to660 = n())


FHLcreditscore620to660_2010$State <- fips(FHLcreditscore620to660_2010$
State,
                                     to = 'Abbreviation')


# # Creating a dataset/table for the Top 10 states who have the highes
t
# number of borrowers with a credit score between 620 & 660 in 2015

FHL620to660top10_2010 <- FHLcreditscore620to660_2010 %>%
  arrange(desc(Borrower620to660)) %>%
  slice(1:10)

View(FHL620to660top10_2010)

kable(x = FHL620to660top10_2010,
      caption = '<b>Table 4: States With the Highest Number of High Ri
sk
      Borrowers Based on Credit Scores</b>',
      digits = 2,
      format = 'html') %>%
  kable_paper("striped", full_width = F) %>%
  row_spec(0, bold = T, background = "#140A0ACA", color = "white") %>%
  column_spec(1, bold = T, background = "#F2C309D8") %>%
  column_spec(2, bold = T,color = "white",
              background = "#3E4AABD8") %>%
  kableExtra::footnote(general = "This table only looks at borrowers f
rom 2010.")
```

```
# Creating a Map Chart too see where more high risk borrowers are for
2010

creditscoregraph2 <- plot_geo(FHLcreditscore620to660_2010,
                              locationmode = 'USA-states') %>%
  add_trace(locations = ~State,
            z = ~Borrower620to660,
            color = ~Borrower620to660,
            colors = "YlOrRd",
            marker = list(line = list(
              width = 0.5,
              opacity = 0.5
            )
      )
  ) %>%
  layout(geo = list(scope = 'usa'),
         font = list(family = "DM Sans"),
         title = "Figure 8: Borrowers with Credit Score \nBetween 620
& 660 in the U.S. (2010)")


creditscoregraph2
```

```
# Filtering Values to Look at Borrowers in 2020

FHLStateName2020 <- FHLStateName %>%
  filter(Year == 2020)


FHL760credit2020 <- FHLStateName2020 %>%
  filter(Borrower760orabove == 1) %>%
  group_by(State) %>%
  summarise(Borrower760orabove = n())


# Changing FIPScode into Abbreviation of State

FHL760credit2020$State<- fips(FHL760credit2020$State, to = 'Abbreviati
on')

View(FHL760credit2020)
```

```
# Creating a dataset/table for the Top 10 states who have the highest
# number of borrowers with a credit score of 760 or above in 2020

FHL760top10_2020 <- FHL760credit2020 %>%
  arrange(desc(Borrower760orabove)) %>%
  slice(1:10)

View(FHL760top10_2020)


kable(x = FHL760top10_2020,
      caption = '<b>Table 5: States With the Highest Number of Low Ris
k
      Borrowers Based on Credit Scores</b>',
      digits = 2,
      format = 'html') %>%
  kable_paper("striped", full_width = F) %>%
  row_spec(0, bold = T, background = "#140A0ACA", color = "white") %>%
  column_spec(1, bold = T, background = "#F2C309D8") %>%
  column_spec(2, bold = T,color = "white",
              background = "#3E4AABD8") %>%
  kableExtra::footnote(general = "This table only looks at borrowers f
rom 2020")



# Creating a Map Chart too see where low risk borrowers are for 2020

creditscoregraph3 <- plot_geo(FHL760credit2020,
                              locationmode = 'USA-states') %>%
  add_trace(locations = ~State,
            z = ~Borrower760orabove,
            color = ~Borrower760orabove,
            colors = "YlGnBu",
            marker = list(line = list(
              width = 0.5,
              opacity = 0.5
            )
        )
  ) %>%
  layout(geo = list(scope = 'usa'),
         font = list(family = "DM Sans"),
         title = "Figure 7: Borrowers With 760 or Above \nCredit Score
in U.S. (2020)")
```

```
creditscoregraph3
```

```
# Creating a dataset that counts the number of borrowers in each state
# that have credit score between 620 & 660 in 2020


FHLcreditscore620to660_2020 <- FHLStateName2020 %>%
  filter(Borrower620to660 == 1) %>%
  group_by(State) %>%
  summarise(Borrower620to660 = n())


FHLcreditscore620to660_2020$State <- fips(FHLcreditscore620to660_2020$
State,
                                         to = 'Abbreviation')


# # Creating a dataset/table for the Top 10 states who have the highes
t
# number of borrowers with a credit score between 620 & 660 in 2020.

FHL620to660top10_2020 <- FHLcreditscore620to660_2020 %>%
  arrange(desc(Borrower620to660)) %>%
  slice(1:10)

View(FHL620to660top10_2020)

kable(x = FHL620to660top10_2020,
      caption = '<b>Table 6: States With the Highest Number of High Ri
sk
      Borrowers Based on Credit Scores</b>',
      digits = 2,
      format = 'html') %>%
  kable_paper("striped", full_width = F) %>%
  row_spec(0, bold = T, background = "#140A0ACA", color = "white") %>%
  column_spec(1, bold = T, background = "#F2C309D8") %>%
  column_spec(2, bold = T,color = "white",
              background = "#3E4AABD8") %>%
  kableExtra::footnote(general = "This table only looks at borrowers f
rom 2020.")


# Creating a Map Chart too see where high risk borrowers are for 2020
```

```
creditscoregraph4 <- plot_geo(FHLcreditscore620to660_2020,
                              locationmode = 'USA-states') %>%
  add_trace(locations = ~State,
            z = ~Borrower620to660,
            color = ~Borrower620to660,
            colors = "YlOrRd",
            marker = list(line = list(
              width = 0.5,
              opacity = 0.5
            )
        )
  ) %>%
  layout(geo = list(scope = 'usa'),
         font = list(family = "DM Sans"),
         title = "Figure 9: Borrowers with Credit Score \nBetween 620
& 660 in the U.S. (2020)")
```

```
creditscoregraph4
```

```
# Creating Tables of Descriptive Statistics Table for White, Non-White
, Female, and
# Male Borrowers
```

```
 FHL_table1 <- FHLAppend %>%
   filter(WhiteBorrower == 1) %>%
   select(LTVRatioPercent, HousingExpenseRatioPercent,
         TotalDebtExpenseRatioPercent, TotalMonthlyIncomeAmount,
         BorrowerAge,HUDMedianIncomeAmount,NoteRatePercent, NoteAmount
) %>%
  sumtable(., summ = c('mean(x)',
                       'min(x)',
                       'max(x)',
                       'sd(x)',
                       'median(x)'),
           out = 'return')
```

```
kable(x = FHL_table1,
      caption = "<b>Table 1: Descriptive Statistics (White Borrowers)<
/b>",
      digits = 2,
      format = 'html') %>%
  kable_paper("striped", full_width = F) %>%
  row_spec(0, bold = T, background = "#140A0ACA", color = "white") %>%
  column_spec(1, bold = T, background = "#F2C309D8") %>%
  column_spec(2:6, bold = T,color = "white",
              background = "#3E4AABD8") %>%
  footnote(general = "Table was created using the FHLAppend dataset wh
ich contains data from 2010, 2015, & 2020.")
```

```
FHL_table2 <- FHLAppend %>%
  filter(WhiteBorrower == 0) %>%
  select(LTVRatioPercent, HousingExpenseRatioPercent,
         TotalDebtExpenseRatioPercent, TotalMonthlyIncomeAmount,
         BorrowerAge,HUDMedianIncomeAmount,NoteRatePercent, NoteAmount
) %>%
  sumtable(., summ = c('mean(x)',
                       'min(x)',
                       'max(x)',
                       'sd(x)',
                       'median(x)'),
           out = 'return')
```

```
kable(x = FHL_table2,
      caption = "<b>Table 2: Descriptive Statistics (Non-White Borrowe
rs)</b>",
      digits = 2,
      format = 'html') %>%
  kable_paper("striped", full_width = F) %>%
  row_spec(0, bold = T, background = "#140A0ACA", color = "white") %>%
  column_spec(1, bold = T, background = "#F2C309D8") %>%
  column_spec(2:6, bold = T,color = "white",
              background = "#3E4AABD8") %>%
  footnote(general = "Table was created using the FHLAppend dataset wh
ich contains data from 2010, 2015, & 2020.")
```

```
FHL_table3 <- FHLAppend %>%
  filter(FemaleBorrower == 1) %>%
  select(LTVRatioPercent, HousingExpenseRatioPercent,
         TotalDebtExpenseRatioPercent, TotalMonthlyIncomeAmount,
         BorrowerAge,HUDMedianIncomeAmount,NoteRatePercent, NoteAmount
) %>%
  sumtable(., summ = c('mean(x)',
                       'min(x)',
                       'max(x)',
                       'sd(x)',
                       'median(x)'),
           out = 'return')

kable(x = FHL_table3,
      caption = "<b>Table 3: Descriptive Statistics (Female Borrowers)
</b>",
      digits = 2,
      format = 'html') %>%
  kable_paper("striped", full_width = F) %>%
  row_spec(0, bold = T, background = "#140A0ACA", color = "white") %>%
  column_spec(1, bold = T, background = "#F2C309D8") %>%
  column_spec(2:6, bold = T,color = "white",
              background = "#3E4AABD8") %>%
  footnote(general = "Table was created using the FHLAppend dataset wh
ich contains data from 2010, 2015, & 2020.")


FHL_table4 <- FHLAppend %>%
  filter(FemaleBorrower == 0) %>%
  select(LTVRatioPercent, HousingExpenseRatioPercent,
         TotalDebtExpenseRatioPercent, TotalMonthlyIncomeAmount,
         BorrowerAge,HUDMedianIncomeAmount,NoteRatePercent, NoteAmount
) %>%
  sumtable(., summ = c('mean(x)',
                       'min(x)',
                       'max(x)',
                       'sd(x)',
                       'median(x)'),
           out = 'return')



kable(x = FHL_table4,
      caption = "<b>Table 4: Descriptive Statistics (Male Borrowers)</
b>",
      digits = 2,
      format = 'html') %>%
  kable_paper("striped", full_width = F) %>%
  row_spec(0, bold = T, background = "#140A0ACA", color = "white") %>%
  column_spec(1, bold = T, background = "#F2C309D8") %>%
  column_spec(2:6, bold = T,color = "white",
              background = "#3E4AABD8") %>%
  footnote(general = "Table was created using the FHLAppend dataset wh
ich contains data from 2010, 2015, & 2020.")
```

```
# Transforming certain variables to be used in regression analysis to
find meaningful interpretations

FHLAppend <- FHLAppend %>%
  mutate(MonthlyIncome1000s = TotalMonthlyIncomeAmount/1000,
         NoteAmount1000s = NoteAmount/1000,
         HUDMedianIncomeAmount1000s = HUDMedianIncomeAmount/1000)

# Create 4 Different Regression Models, looking at the effects of LTV
based on
# Borrower's Race and Self-Employment Status

regression_results <- list(

  "Regression 1
  White Borrowers" = lm(LTVRatioPercent ~ MonthlyIncome1000s +
                        BorrowerAge + WhiteBorrower + BorrowerFirstTim
eHomebuyer
                        + NoteAmount1000s + HUDMedianIncomeAmount1000s
                        , data = FHLAppend),

  "Regression 2
  Black Borrowers" = lm(LTVRatioPercent ~ MonthlyIncome1000s +
                        BorrowerAge + BlackBorrower + BorrowerFirstTim
eHomebuyer
                        + NoteAmount1000s + HUDMedianIncomeAmount1000s
                      , data = FHLAppend),

  "Regression 3 \nSelf Employed White Borrowers" = lm(LTVRatioPercent
~ MonthlyIncome1000s +
                                                        BorrowerAge +
SelfEmployedWhiteBorrower + BorrowerFirstTimeHomebuyer
                                                      , data = FHLAppe
nd),

  "Regression 4 \nSelf Employed Black Borrowers" = lm(LTVRatioPercent
~ MonthlyIncome1000s +
                                                        BorrowerAge +
SelfEmployedBlackBorrower + BorrowerFirstTimeHomebuyer
                                                      , data = FHLAppe
nd)

)
```

```
regressiontable <- modelsummary(regression_results, stars = c('*' = 0.
1,
                                                              '**' = 0
.05,
                                                              '***' =
0.01),
                                title = "Table 7: Differences in LTV B
y Race & Gender",
                                gof_omit = 'IC|Log|Adj',
                                output = "gt")

regressiontable %>%
  tab_footnote(footnote = md("Variables of interest are highlighted an
d their
                             coefficients are in red.")) %>%

  tab_style(style = cell_text(color = 'red'),
            locations = cells_body(rows = 7)) %>%

  tab_style(style = cell_fill(color = 'lightblue'),
            locations = cells_body(rows = 7)) %>%

  tab_style(style = cell_text(color = 'red'),
            locations = cells_body(rows = 15)) %>%

  tab_style(style = cell_fill(color = 'lightblue'),
            locations = cells_body(rows = 15)) %>%

  tab_style(style = cell_text(color = 'red'),
            locations = cells_body(rows = 17)) %>%

  tab_style(style = cell_fill(color = 'lightblue'),
            locations = cells_body(rows = 17)) %>%

  tab_style(style = cell_text(color = 'red'),
            locations = cells_body(rows = 19)) %>%

  tab_style(style = cell_fill(color = 'lightblue'),
          locations = cells_body(rows = 19))
```

References

Bhutta, N., Hizmo, A., & Ringo, D. (2021). How much does racial bias affect mortgage lending? evidence from human and algorithmic credit decisions. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3887663

Agarwal, S., Amromin, G., Ben-David, I., Chomsisengphet, S., & Evanoff, D. (2014). Predatory lending and the subprime crisis. *Journal of Financial Economics*, *113*(1), 29–52. https://doi.org/10.3386/w19550

FHFA. (2020). *Federal Home Loan Level Bank System 2009-2018*. Kaggle.com. https://www.kaggle.com/datasets/jeromeblanchet/federal-home-loan-level-bank-system-20092018?select=2018_PUDB_EXPORT_123118.csv

FHFA. (n.d.). *Public Use Databases*. Public Use Databases | Federal Housing Finance Agency. https://www.fhfa.gov/DataTools/Downloads/Pages/Public-Use-Databases.aspx

Team, D. S. (2022, December 11). *How to remove outliers from data in R*. Universe of Data Science. https://universeofdatascience.com/how-to-remove-outliers-from-data-in-r/

Hayes, A. (2024, February 24). *Loan-to-value (LTV) ratio: What it is, how to calculate, example*. Investopedia. https://www.investopedia.com/terms/l/loantovalue.asp

Fernando, J. (2024, February 7). *The correlation coefficient: What it is and what it tells investors*. Investopedia. https://www.investopedia.com/terms/c/correlationcoefficient.asp

Johnson, J. (2024, February 27). *What are Fannie Mae and Freddie Mac? | mortgages | U.S. news*. U.S. News. https://money.usnews.com/loans/mortgages/articles/what-is-fannie-mae