Home Mortgage Loan Applications

Project

Background

The Home Mortgage Disclosure Act (HMDA) requires many financial institutions to maintain, report, and publicly disclose loan-level information about mortgages. These data help show whether lenders are serving the housing needs of their communities, give public officials information that helps them make decisions and policies, and shed light on lending patterns that could be discriminatory. The public data are modified to protect applicant and borrower privacy.

Each year, millions of people apply for mortgages. Using HMDA data, we can learn what happened to the vast majority of those applications and determine if there is any bias, whether conscious or not, towards minority applicants. If our analysis turns out to uncover that there is bias against lending to minority borrowers, then that suggests that mortgages are being denied to worthy applicants based on the color of their skin or ethnic background.

In this project, we're going to examine whether this is indeed the case.

HMDA Data

HMDA data is provided by the Department of Housing and Urban Development. The HMDA data has a very rich amount of information regarding the approval or denial of a home loan. There are many fields in the dataset, but the most important fields to know for this project are the following:

- **respondent_id:** The unique identifier for the application.
- **loan_amount_000s:** The loan amount, in thousands, that the applicant has requested from the lender.
- **action_taken_name:** Describes, in words, the different kinds of outcomes for the loan, e.g., approved, denied, etc. See more details in the flowchart below.
- **action_taken**: Numeric code for the end result of the loan application.
- **applicant_ethnicity_name:** Describes whether an applicant is of Hispanic origin or Non-Hispanic origin. Some applicants do not fill in this information.
- **applicant_race_name_1:** Describes the race of the applicant, if provided. Categories are White, Black or African American, Asian, American Indian or Alaskan Native, Native Hawaiian or other Pacific Islander.
- **applicant_sex_name:** Describes the sex of the candidate, if provided.
- **applicant_income_000s**: The income, in thousands, of the loan applicant.
- **hud_median_family_income_000s:** The median income for the metropolitan area where the applicant resides.

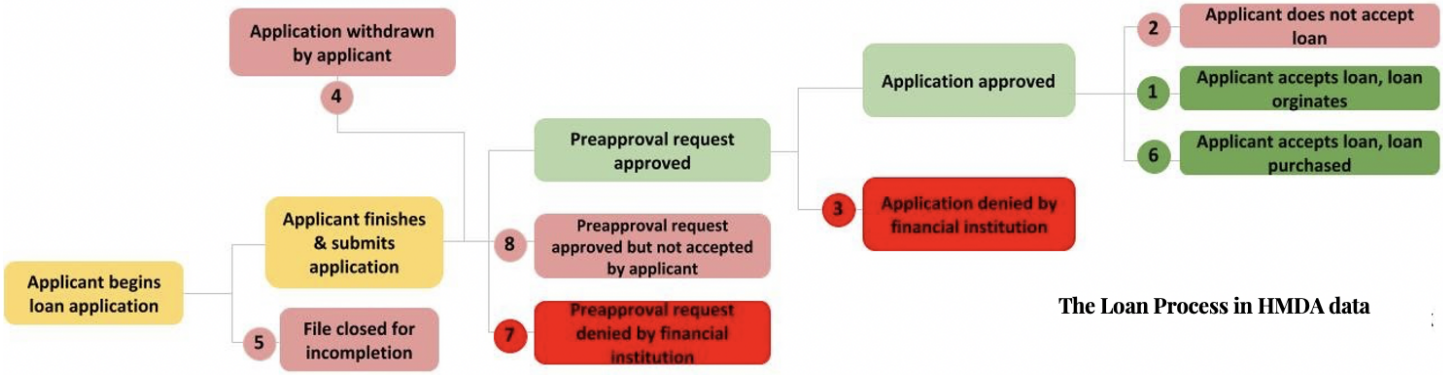
Disclaimer: The original HMDA dataset for 2017 has ~15M entries and is 11GB in size. The dataset provided has been reduced in size to fit in Excel. All of the values in the original dataset have not been modified. The HMDA data is a great example of a project that would take a team of analysts and data visualizers several months of work to build models, extract insights, build graphs, etc. What we have done here is only scratching the surface of what can be done with the data.

Home Buyer Journey

The jargon used in the mortgage industry can be difficult to grasp for newcomers. In order to understand the analysis you will be performing, it will be useful to have a general overview of the context for the data. In this case, let's overview the process of a prospective home buyer and the journey towards purchasing a home. This will help you understand key components and terminology of the dataset that we will want to focus on.

Shaina, a first time home buyer, wants to buy a home but doesn't have the money to pay for it all in cash, so she applies for a loan at her bank. She tells the bank about her finances, the house she wants to buy, and other information the bank needs to make a decision about whether or not to lend to her, and the terms of the loan. The bank reviews her application, decides that she meets their criteria, and she gets approved. Once all the papers are signed, Shaina accepts the loan… or in mortgage-speak, the loan has "originated" or been "purchased."

All told, between starting an application and the loan originating, there are multiple hurdles to overcome. Take a look at the following process chart.



The numbers in the circles correspond to the **action_taken** feature in the data set. Of particular interest are the four numbers at the right half of the image. All of these follow the "preapproval request approved" step of the loan process. When we see a "3", this indicates that the application was ultimately denied. On the other hand, when we see a "1", "2", or "6", this indicates that the application was approved.

It might end up that the applicant does not end up taking the loan (code 2), but grouping those multiple outcomes together will give us insight into the core questions regarding bias in application rates across race and ethnicity.

All the remaining action code numbers represent incomplete applications or failed pre-approval requests. For these applications, we can not evaluate if the applicant would have been approved and will be removed from our bias analysis.

Part 1 - Exploratory Data Analysis

Before we can run a statistical test for bias in our data, we should explore it and understand its basic properties. We can use this analysis to inform our expectations for what the outcome of our tests should look like as well as identify parts of the dataset we may need to clean or change into more useful forms.

Task 1

As discussed in the intro above, “loan origination” isn’t the only code used in the data to signify that a loan has been approved. In the data, there are several “action” types. Some applicants may have been approved but ultimately did not purchase a home and others were denied at different stages of the loan process.

Use XLOOKUP with the **ActionReference** table (at the far right on the HMDA_Data sheet) to fill in the column called "approval_status_name" that returns:

- “Approved” when "action_taken" = 1, 2, or 6 (i.e., the loan was approved)
- “Denied” when "action_taken" = 3 (i.e., the loan was denied)

Note that there are other possible outcomes that fall earlier in the loan process, but they have already been cleaned from the data through prior data cleaning steps.

Task 2

With the newly re-encoded application outcomes, we should first take a general survey of the overall loan approval rate.

a) What percentage of the loans were approved?

b) What percentage of the loans were denied?

Task 3

We can now compare the loan approval rates by different subgroups to see if there are any differences between them. We don't want to run a statistical test quite yet: this is still part of exploration of and learning about the data.

a) Use a PivotTable to perform a breakdown of loan approval and denial by race. Make sure to format the values so they are displayed as a percentage within each race.

b) Which races fall below the overall average approval rate from part 2a?

c) Is there evidence in the data to suggest that there are certain races that are facing negative bias? If so, which ones?

d) Create a 100% stacked bar chart that depicts the loan approval rates by race.

    - Make sure to perform some formatting so your chart tells a clear story. For example, in which order should the bars be displayed?

Part 2 - Bias Analysis

There are a number of different directions we could go with the data, but it's a good idea to stay focused. In this analysis, the main thing we want to check is whether Black Americans are being biased against by lenders at a statistically significant level.

One thing to note is that there are many more White Americans in the upper socioeconomic status class than there are Black Americans. In order to compensate for this we need to level the playing field and determine if in each of three income classes (lower, middle, and upper) there are signs of bias. This is normally what is being done when you read headlines like "Black Americans are disproportionately affected by lenders after accounting for income level". If we don't account for income levels, we won't be making an "apples to apples" comparison.

Task 4

First, let's do some general data exploration and look at the distribution of applicant incomes.

a) Create a histogram of income distributions (using applicant_income_000s). Anyone making over $500k should be included in the "overflow bin". Change the bin width to $25k. Remember, the data is already in thousands so a value of 75 means the applicant makes $75k per year.

    - Make sure to use an appropriate title for your chart!

b) Briefly describe the distribution of applicant incomes. Is it symmetric or skewed? How many modes (peaks) are there and where are they located?

Task 5

According to the Pew Research Center, the range for middle class income (in 2017, the year our HMDA data is from) is classified as a family making between $40,000 and $122,000. Thus a family making under $40,000 is considered lower class and anyone above $122,000 is considered upper class. Let's use this definition to split up the data into income groups.

a) Fill in the "income_class_name" column at the end of the HMDA_Data table that returns:
    - "lower" if the applicant income is less than 40 (remember, the applicant income column is in thousands),
    - "middle" if less than 122, and
    - "upper" otherwise.

b) What is the breakdown of loan approval versus race when we break down the race feature by income class? Use a PivotTable to perform this analysis, and use a filter on the applicant's race to compare only "Black or African American" and "White" applicants.

c) Does the pattern you first observed in Task 3 still hold when breaking things down by income level? Be specific in your explanation.

Task 6

Now we can compare the loan applications of African Americans to White Non-Hispanic Americans using a statistical test. Let's focus on just the middle-income bracket applications, since that is where we have the most data available. This data has already been pulled out of the data table, into columns A and B of the Bias Analysis sheet of the workbook.

a) We're interested in seeing if there is a bias against Black or African American loan applications. State the null and alternative hypotheses for the statistical test. Remember to use words like "greater than", "less than", or "equal to".

b) Use the Data Analysis Toolpak to evaluate the statistical test on the Bias Analysis worksheet. Perform a two-sample t-test assuming unequal variances between groups to compare the loan approval rates for our two groups.

c) From the output of the statistical test, is there evidence at the 95% confidence interval that middle-class African Americans are approved or denied loans differently than their White Non-Hispanic counterparts? Why or Why not?