ROBERT H. SMITH
SCHOOL OF BUSINESS

Mid-term Individual Project
BUDT758D Spring 2024

# Visualizing Fannie Mae Mortgage Data

Fannie Mae is one of the largest purchasers of home mortgages from banks and sellers of mortgage-backed securities (a financial instruments created by bundling and selling mortgages to investors) on the secondary market in the United States. In exchange for a fee, Fannie Mae guarantees the payments of principal and interest on the book of loans it sells. As Fannie Mae bears the losses in the event of borrower defaults on a mortgage, its future profitability is significantly influenced by borrower default rates.

In September 2008, at the peak of the financial crises, Fannie Mae was piling up significant losses. A wave of mortgage defaults posed a severe threat to Fannie Mae and, consequently, the entire U.S. financial system. To save the system, the U.S. government stepped in and placed Fannie Mae under conservatorship, a form of bankruptcy overseen by the Federal Housing Finance Agency (FHFA). This move transformed Fannie Mae from a government-sponsored entity to a government-owned entity. From 2008 to 2011, the U.S. Treasury injected $116.1 billion of capital into Fannie Mae. Subsequently, Fannie Mae embarked on a path to recovery. By mid-2012, it had restored profitability and regained a dominant in the market. In 2022, Fannie Mae reported an annual net income of $13 billion[1], and held the largest market share, accounting for 40% of the issuance of single-family mortgage-backed securities (MBS)[2].

“To promote better understanding of the credit performance and acquisition of Fannie Mae mortgage loans,” Fannie Mae released each quarter updates to its data related to mortgages, such as the prepayment of MBS and loan performance data (the primary focus of this assignment). On October 27, 2023, Fannie Mae updated its loan performance data through Q2 2023. This data includes a subset of 16 million mortgages, all with terms of 30 years or less. These mortgages are fully amortizing, full documentation, single-family, conventional fixed-rate loans, generally considered the lowest rick in Fannie Mae’s portfolio. The loans within this dataset were acquired between January 1, 2000, and June 30, 2023.

Analyzing this dataset represented a significant challenge, even for sophisticated investors who owned Fannie Mae stock. How could such investors use this data to understand default rates within Fannie Mae’s single-family guaranty book of business? Which fields in the data were drivers of loan defaults and individual interest rates? How can we visually analyze Fannie Mae’s loan performance?

In this assignment, we will work with a smaller subset of the data obtained from Fannie Mae’s data portal (please note that registration is required to access the original data directly). To access the statistical summary of the entire data set, you can find the “FNMA_SF_Loan_Performance_Stat_Summary_Primary.pdf”[3] file on ELMS. For a comprehensive list of fields contained in the original data, refer to the “crt-file-layout-and-glossary.pdf” file on ELMS (the fields used for this assignment are listed at the end of this document). For further details about the data, you can visit the Fannie Mae Single-Family Loan Performance Data webpage. Additionally, apart from loan performance data, Fannie Mae also provides data on Mortgage-Backed Securities (MBS), Single-Family Connecticut Avenue Securities (CAS), and Single-Family Credit Insurance Risk Transfer (CIRT) on their data portal.

[1] Fannie Mae Reports Net Income of $12.9 Billion for 2022 and $1.4 Billion for Fourth Quarter 2022 (accessed January 1, 2024)
[2] Relative Attractiveness of US Fixed Income and Ginnie Mae MBS (accessed January 1, 2024)
[3] Downloaded from Fannie Mae Single-Family Loan Performance Data (accessed January 1, 2024)

## Assignment Guidelines

| | | | |
|---|---|---|---|
| **Assignment Name:** | Mid-term Individual Project | **Points:** | 150 |
| **Submission Type:** | ELMS and Tableau Public | **Submission Date:** | March 17, 2024 at 11:59pm |

### Data

For this assignment, you will have access to the following three datasets. You have the option to use either two or all three of them in the assignment. The first two datasets are small samples randomly drawn from the data for specific periods, representing approximately 17% and 12% of the entire data for Q4 2007 and Q4 2019, respectively. You can refer to Exhibit 1 and 2 for a detailed list of the fields included in each dataset[4].

- **"2007Q4.rds"** contains information on 50,000 fixed-rate single-family amortizing loans with terms of 30 years or less. These loans were either owned or guaranteed by Fannie Mae during the fourth quarter of 2007, a period marked by a significant increase in loan defaults[5] at Fannie Mae.
- **"2019Q4.rds"** includes information on 50,000 Fannie Mae loans of the same type as the previous data file, but specific to the fourth quarter of 2019 – the last quarter before the onset of COVID.
- **"default_rate_ts.csv"** Provides a time series of quarterly default rates for the same types of loans as in the previous two datasets. The default rate is calculated by dividing the number of loans that defaulted during the quarter by the total number of loans in the dataset for that quarter. The entire dataset provided by Fannie Mae was used, as opposed to just samples as in the previous two datasets.

If you intend to incorporate supplementary data, you are only allowed to use those available on Fannie Mae's data portal.

### Assignment Overview

**Write a report** summarizing your findings from visualizing the provided data. Identify a cohesive focus for your report that interests you. For instance, you can explore topics such as: How has the default rate of loans acquired by Fannie Mae evolved over time? What factors drove loan defaults in 2007, and are they similar to those in 2019? Are the characteristics of borrowers and houses for Fannie Mae-acquired loans the same in 2007 and 2019? These are just examples to spark ideas, and you're welcome to explore other topics of interest. Keep in mind your audience—choose a topic that tells an engaging story and is suitable for discussion in a professional context.

As part of the assignment, you are also required to **generate an interactive chart**. Please refer to the detailed instructions provided below.

[4] If you believe there are specific fields in Fannie Mae's dataset that could be useful but are missing from the data provided for this assignment, please contact your instructor for guidance. Additionally, if you wish to download additional data from Fannie Mae's data portal, reach out to your instructor to obtain the R file for loading and preparing the data.

[5] In the original dataset provided by Fannie Mae, there exists a variable called 'ZERO_BALANCE_CODE.' In the dataset used for this assignment, a loan is deemed in default if the code is not 'prepaid or matured' or left empty.

## Instructions

### Report and Charts (135)

The report should include **a title, the main body of the report, and 4-6 visualizations** to support your analysis. Use the R markdown file “Midterm-Report-Template.Rmd” available on ELMS to create your report.

The report's title should accurately represent the chosen topic. The main body of the report should comprise an introductory paragraph explaining the problem's significance for business decision-making, several paragraphs detailing findings from the visualizations, and a concluding paragraph summarizing your analysis and providing managerial insights. You do not need to provide any details about the data source or the data itself in your report. **The main body of the report should not exceed 800 words in length.**

The 4-6 visualizations should be included in the section called “Figure Appendix”. These charts should address your analysis questions and demonstrate your understanding of the data visualization principles taught in class. Each chart should be appropriately labeled to increase the clarity and use the appropriate visual cues (color, length, shape, etc.). Examples of suitable charts include bar charts illustrating default rates for different occupation statuses, or histograms depicting borrowers' credit scores in 2007 Q4 and 2019 Q4. Make sure that the titles, labels, and other elements are legible in the output HTML file. Do not incorporate titles directly into the charts; instead, include captions using the provided R markdown template.

### Interactive chart (10)

One of the visualizations in your report must be an interactive chart that permits your audience to access additional information by hovering over the chart. The information displayed upon hovering should be concise, informative, and contribute to your audience's comprehension of the insights presented in the report.

### Submission (5 points)

There are two deliverables to submit on ELMS for this project:

1. The HTML output that contains your report (no code should show up in this file).
2. The R Markdown file that contains the code and produces your report.

## Guidelines

When visualizing the data, consider the following points (although they are not mandatory):

- Calculate the default rate within a specific group of loans (e.g., loans with occupation status 'Principal' or loans from borrowers with a credit score of 700) by averaging the 'DEFAULT_FLAG' variable, which is equivalent to the number of default loans divided by the total number of loans in that group.
- If a categorical variable has too many levels, consider grouping them into fewer levels.
- You can refer to news articles or reports to support your conclusions, spark your topic, or convince your readers of your topic’s significance. Make sure that you include the reference in the report if you choose to do so. Please note that the use of other online resources, such as plots produced by others that are accessible online, is not permissible.

## Grading Rubric

The report will be graded according to the following criteria:

A. Does the introduction clearly state the problem you are studying? (5 pts)
B. Are the findings from the visualizations explained well and consistent with your charts? (30)
C. Are the conclusions drawn from the analysis insightful? (5 pts)
D. Is your report well-designed: brief, clear, and compelling? Does your report convey a clear and consistent narrative? (10 pts)
E. Does your report's format adhere to the provided instructions? (5 pts)

The visualizations will be graded according to the following criteria:

F. Are your charts accurate? (20 pts)
G. Are your visualization choices appropriate for the analysis and data? (20 pts)
H. Are your charts attractive, easy-to-read, well-labeled, and easy-to-understand? (20 pts)
I. For the interactive chart, is the interactivity appropriate and well-implemented? (10)

Other criteria:

J. Did you put forth additional effort above and beyond the minimum? For example, did you create more sophisticated and/or interesting charts? Did you include a variety of chart types (bar charts, line charts, scatter plots, histograms, boxplots, etc.)? This score will be evaluated relative to all other teams, and is intended to recognize exceptional effort and dedication to the project (20 pts)
K. Is the submission accurate (e.g., output HTML report and R Markdown file uploads)? (5 pts)

**The smith school does not take plagiarizing lightly. If a case of plagiarism is identified, you will be reported to UMD's Academic Integrity Committee and Student Honor Council for further investigation without notice. As part of our plagiarism checking process, we will be comparing your project with publicly available Tableau workbooks to identify similarities.**

## Exhibit 1. Descriptions of the Data in default_rate_ts.csv.

| Field Name | Description | Data Type |
|---|---|---|
| Date | The dates on the first day of each quarter, covering the period from 2020 Q1 to 2023 Q2. | Date |
| Default_rate | The quarterly default rate is computed using the complete dataset available on Fannie Mae's data portal (not the small sample used in this assignment). | Numeric |

## Exhibit 2. Descriptions of the Data in 2007Q4.rds and 2019Q4.rds.

| Field Name | Description | Data Type |
|---|---|---|
| LOAN_ID | A unique identifier for the mortgage loan. | Character |
| CHANNEL | The origination channel used by the party that delivered the loan to the issuer. Retail = R; Correspondent = C; Broker = B. | Character |
| LOAN_AGE | The number of calendar months since the mortgage loan's origination date (the date on which the first full month of interest begins to accrue). | Numeric |
| SELLER | The name of the entity that delivered the mortgage loan to Fannie Mae. | Character |
| ORIG_RATE | The original interest rate on a loan as identified in the original note. | Numeric |
| CURR_RATE | The rate of interest in effect for the periodic installment due. | Numeric |
| ORIG_UPB | Original unpaid principal balance. The dollar amount of the loan as stated on the note at the time the loan was originated. | Numeric |
| ORIG_TERM | The number of months in which regularly scheduled borrower payments are due at the time the loan was originated. | Numeric |
| OLTV | Original Loan to Value Ratio. The ratio (as a percentage) obtained by dividing the amount of the loan at origination by the value of the property. | Numeric |
| OCLTV | Original Combined Loan to Value Ratio. The ratio (as a percentage) obtained by dividing the amount of all known outstanding loans at origination by the value of the property. | Numeric |
| NUM_BO | Number of Borrowers. The number of individuals obligated to repay the loan. | Character |
| DTI | Debt-To-Income. The ratio obtained by dividing the total monthly debt expense by the borrower’s total monthly income at origination. | Numeric |
| FIRST_FLAG | First Time Home Buyer Indicator. First time borrower = Y; Otherwise = N. | Character |
| CSCORE_B | Borrower Credit Score at Origination. When this term is used by Fannie Mae, it is referring to the "Classic" FICO score developed by Fair Isaac Corporation. | Numeric |
| CSCORE_C | Co-Borrower Credit Score at Origination. | Numeric |
| PURPOSE | An indicator that denotes the purpose of the loan. Cash-Out Refinance = C Refinance = R; Purchase = P; Refinance-Not Specified = U | Character |
| OCC_STAT | The classification describing the property occupancy status at origination. Principal = P; Second= S; Investor = I; Unknown = U. | Character |
| STATE | A two-letter abbreviation indicating the state where the property is located. | Character |
| MI_PCT | Mortgage Insurance Percentage. The original percentage of mortgage insurance coverage obtained for an insured conventional mortgage. | Numeric |
| DEFAULT_FLAG | This is a calculated field that is not in the original data. A value of one indicates that the loan defaulted during the period, while zero indicates that the loan was paid on time. | Numeric |