Loan Performance Data Insights

Introduction

The Fannie Mae Single-Family Loan Performance Data is a large dataset released by Fannie Mae that includes detailed information on the performance of single-family mortgage loans. This data is made available to help researchers, investors, and analysts better understand mortgage loan performance and risk, as well as to improve modeling and analytics related to the U.S. housing market.

Key Features of the Dataset

- **Loan-level Data:** It contains detailed data at the individual loan level, which allows for in-depth analysis.
- **Credit Performance:** The dataset tracks the performance of loans over time, including delinquency status, credit scores, loan-to-value (LTV) ratios, and other factors that influence loan performance.
- **Origination Data:** Includes information on the characteristics of the loans when they were originated, such as interest rates, loan purpose, and property types.
- **Performance Data:** Tracks the performance of the loans on a monthly basis, including payment history, default status, and modifications.
- **Modification Data:** Includes details on any modifications to the original loan terms, such as changes in interest rates or repayment terms.

Structure

The dataset is split into two main parts:

1. **Acquisition:** This contains static information about the loans at the time they were originated (e.g., credit scores, loan terms).
2. **Performance:** This tracks the monthly performance of each loan, such as delinquency status, payments made, and loan modifications.

Data Coverage

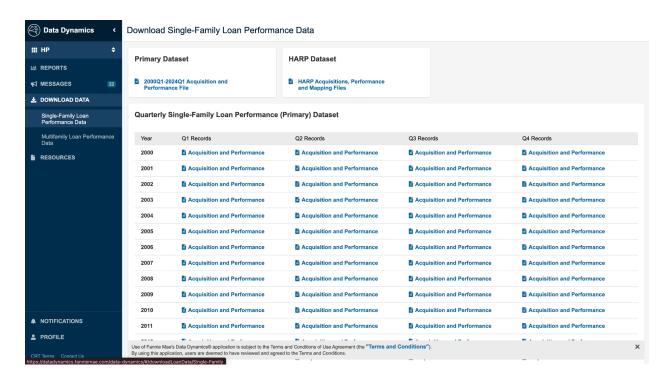The dataset spans from 2000 onwards, covering millions of loans.
It is updated quarterly, ensuring that the data remains relevant and reflects the latest performance metrics.

Data Access

Details of the dataset are present on the Fannie Mae [website](#).
You will need to create a free Fannie Mae account in order to access the data.
We will be focussing on the Single-Family Loan Performance Data which can be downloaded from the section highlighted in the screenshot below.

The dataset spans quarters from 2000 to 2023. This project is to be undertaken in groups of 2. Each group will get assigned 2 years worth of non-overlapping data. For instance, group 1 will take on 2000-2001, group 2 2002-2003, and so on. The dataset allocation will be assigned by the course TA.

Tooling

You are expected to use pyspark in Google Colabs for this project. My recommendation is to upload your dataset to your GWU Google Drive and then convert it from CSV to Parquet format and then use the Parquet file for analysis in pyspark. You can use any Python library, such as `matplotlib`, etc. of your choosing to create the plots. In addition, all documentation/comments should be in the form of Markdown within the notebook itself.

Analysis

For each piece of analysis, provide an explanation of the process as well as an interpretation of the results.

1. Compare the average FICO across the two years.
2. The monthly delinquency rates for 30, 60 and 90 days past due and how they vary by loan term
3. Stacked bar chart showing the poor, fair good, very good, and excellent credit score with a mortgage per state for first time buyers.
4. Analyze correlations between the FICO Score, LTV Ratio, and Interest Rate with the loan status.
5. Distribution of FICO Scores, LTV Ratios, and Interest Rates across different loan statuses (performing, delinquent, defaulted).
6. Compare default rates for loans originated in different quarters to assess if default risk has changed over time.
7. The percentage of the loan amount recovered after default, through foreclosure or other means.
8. Plot average, median, and variance of property price changes over the entire duration, bucketed by month.

Submission

You can submit the link to your Google Colab via Blackboard in the Assignments section.