US Mortgage Loan Data Analysis

Alyssa, Chris, Gerard, Seung

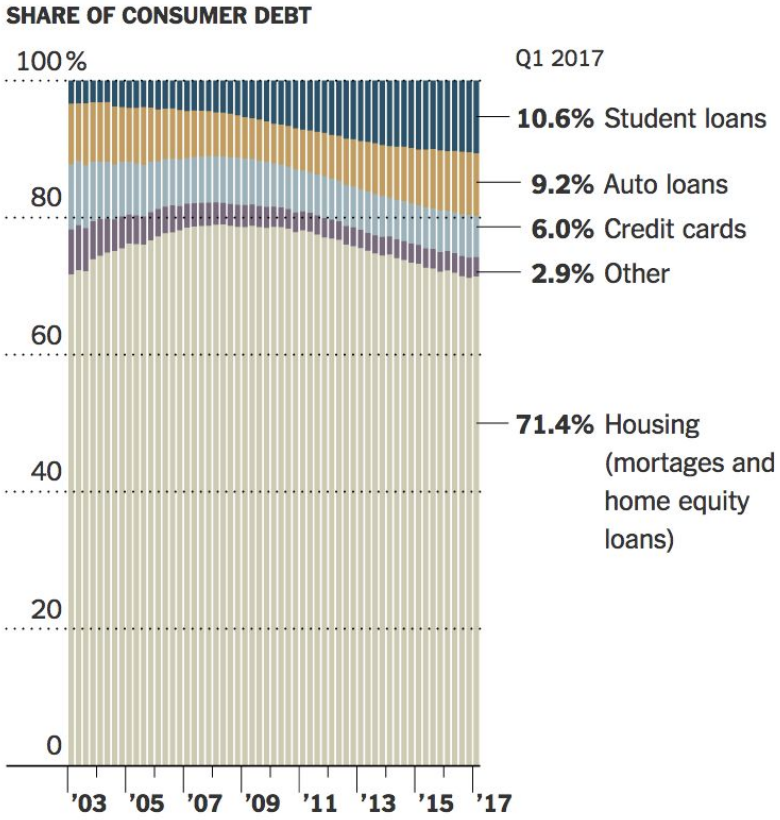
Problem & Motivation

Housing mortgages: primary component of ~$13 billion consumer debt in the United States.

Basis for an important class of financial asset; residential mortgage-backed securities (RMBS).

Predicting their future performance is a key task for securities analysts, especially the major ratings agencies.

STANDARD &POOR'S Moody's FitchRatings



Architecture Design

Acquire, store, and transform raw data: use Hadoop File System and Hive

Serving layer: Query Hive metadata tables and Tableau

Data Acquisition and Organization:

Most functions from initiating the system to transforming the data can be run by a series of .sh scripts

For loan or acquisition quarter, python script lets user download Fannie and Freddie data of interest

Major loading and transformation is accomplished by .sh scripts, which can be part of fully automated download-load-transform process

System also allows for free roaming and querying of data beyond the pre-made tools.

Data Acquisition and Organization:

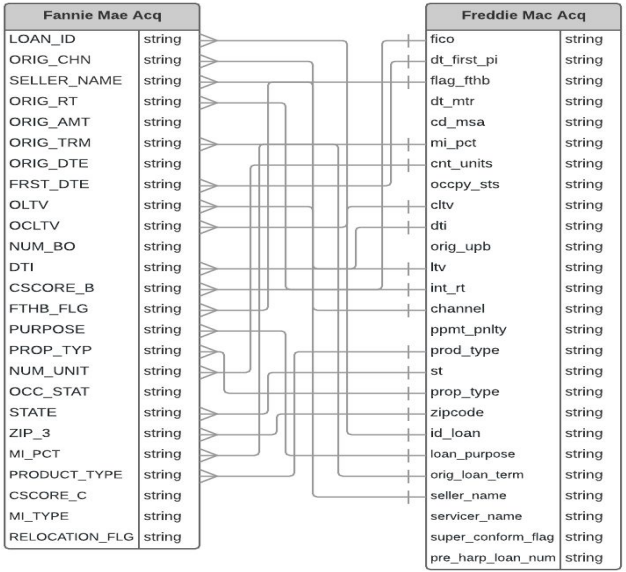
Large amount of Data from each organization (0.5-10 GB based on time span)

Fannie and Freddie **not** entirely overlapping

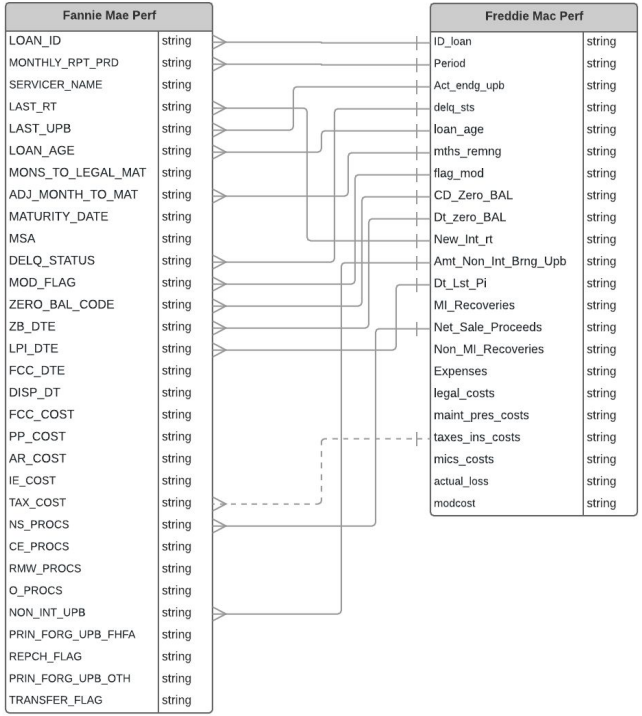
Useful metrics, such as I**nterest rate, debt-to-income, loan ID** match as keys

Other metrics, such as **maturity dates**, **delinquency status** are scaled differently

Acquisition Data



Performance Data



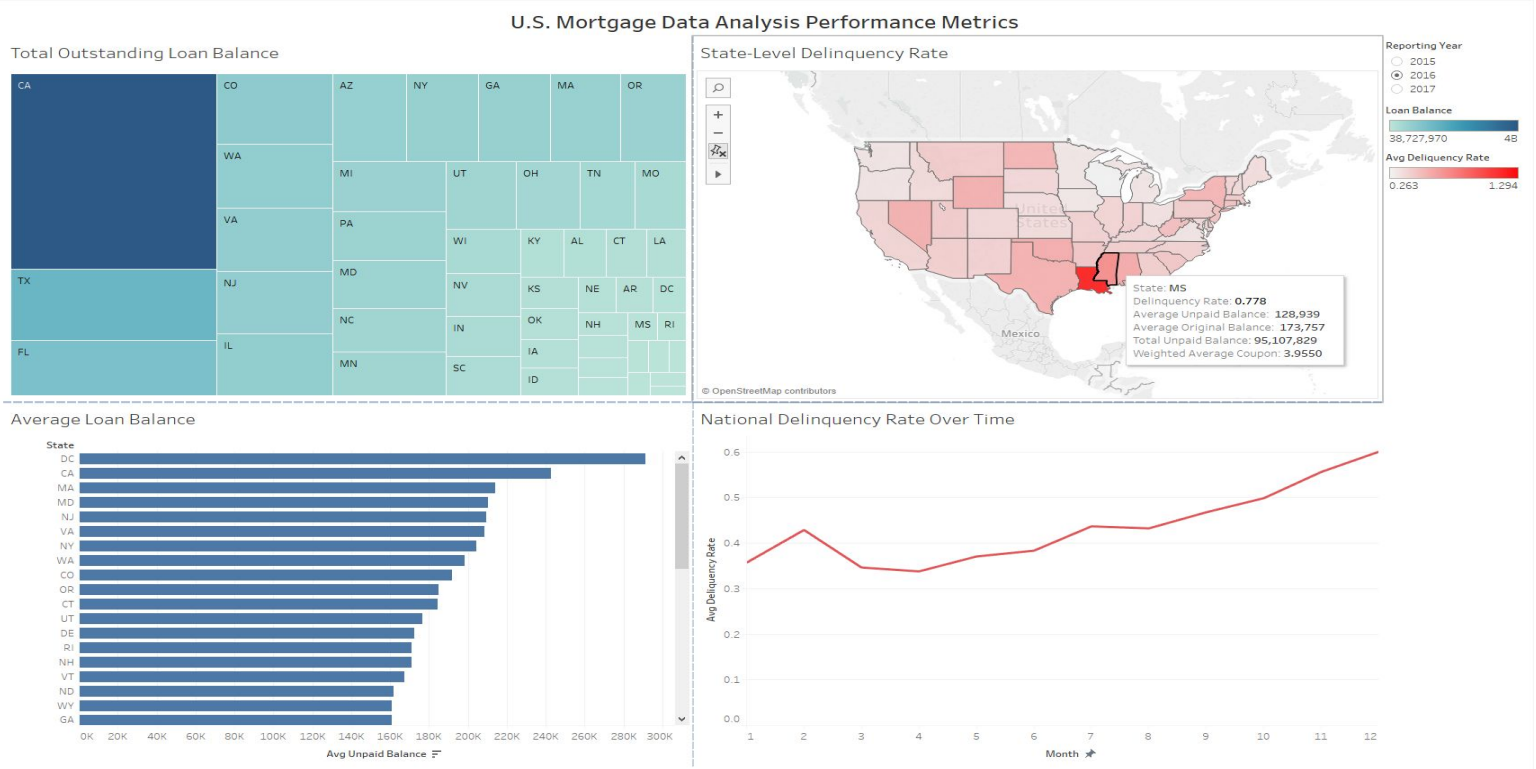
Data Acquisition and Organization:

- Fannie and Freddie Acquisition & Performance data is joined
  - Depending on the time range, this could take many hours
  - Large data uploads and joins are required 4 times per year and can be performed overnight
- Data is then aggregated on loan origination state and 3 digit zip code (ex. 10036 = 100**= 100)
- While user has access to all data, the aggregated State and Zipcode tables are especially useful for visualizations

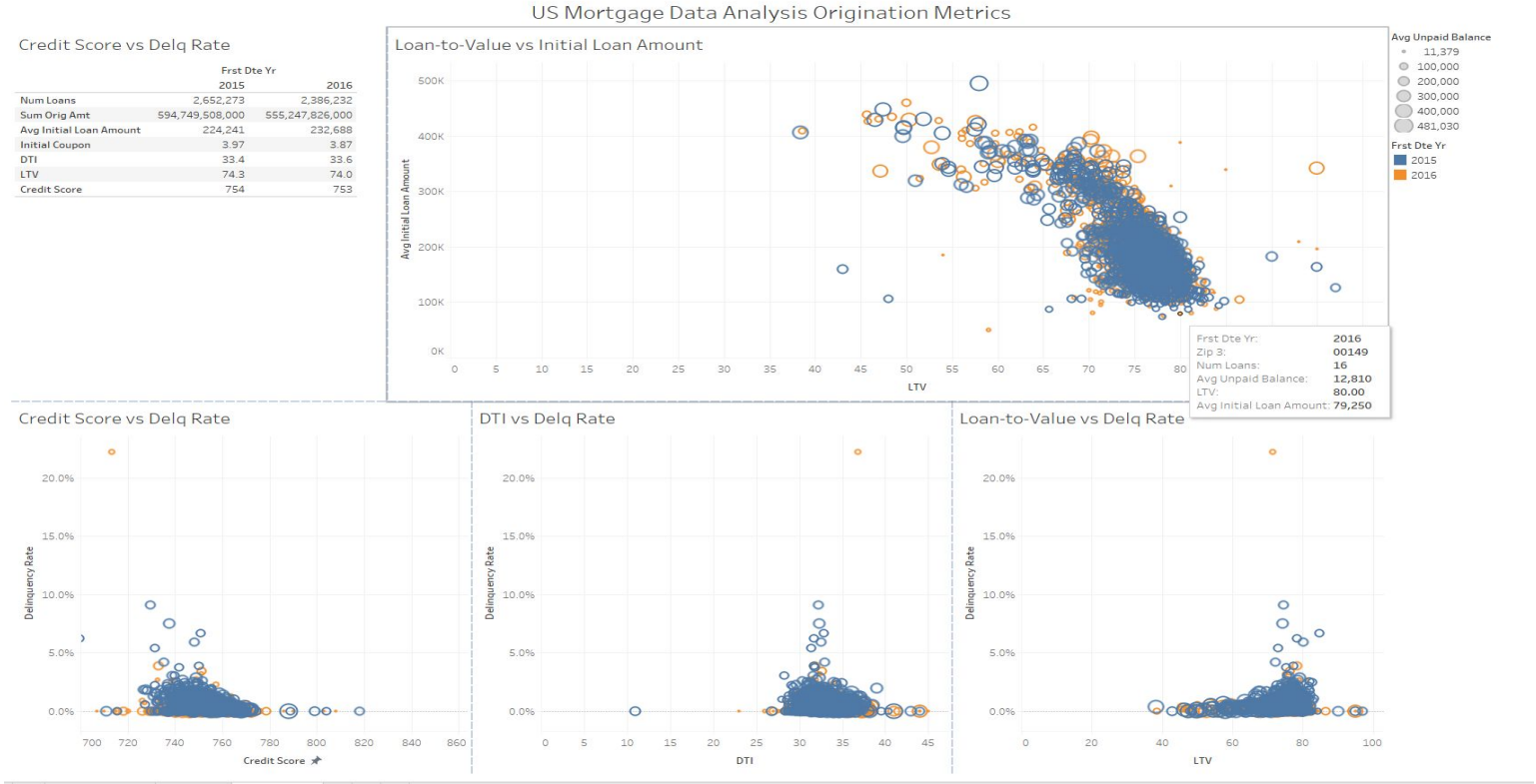
3 Digit Zipcode Aquisition	
ZIP_3	string
FRST_DTE_MTH	string
FRST_DTE_YR	string
NUM_LOANS	string
SUM_ORIG_AMT	string
AVG_ORIG_AMT	string
AVG_ORIG_RT	string
WAVG_ORIG_RT	string
AVG_OLTV	string
WAVG_OLTV	string
AVG_DTI	string
WAVG_DTI	string
AVG_CSCORE	string
WAVG_CSCORE	string

3 Digit Zipcode Performance	
ZIP_3	string
FRST_DTE_MTH	string
FRST_DTE_YR	string
RPT_PRD_MTH	string
RPT_PRD_YR	string
NUM_LOANS	string
SUM_ORIG_AMT	string
SUM_UPB_AMT	string
AVG_UPB_RT	string
AVG_LAST_RT	string
WORG_AVG_LAST_RT	string
WUPB_AVG_LAST_RT	string
AVG_DELQ_RT	string
WORG_AVG_DELQ_RT	string
WUPB_AVG_DELQ_RT	string

Data Serving: Tableau Example - Performance



Data Serving: Tableau Example



Scaling and Extending

Complete loan performance databases

- 17 years (total) vs 2 years (demo)
- Automated v manual data updates

Auxiliary housing price, sales and construction data

- FHFA House Price Index (housing price data by zip)
- US Census Bureau (house price index by region)
- Zillow: % homes decreasing / increasing in value (by state)
- Many others, but more problematic.

Macroeconomic data

- Employment, wages and economic growth
- Demographics

THANKS!