INST754: FINAL PROJECT PRESENTATION

LOAN DATASETS INTEGRATION, MODELING AND ANALYSIS USING POWER BI

PRESENTED BY:

Tanya Gupta

Ushasri Bhogaraju

THE SINGLE FAMILY HOME

Image from: https://www.homes.com/adelphi-md/neighborhood/adelphi/

- Single family homes are a free-standing residential building which can have upto 4 separate living units.
- Single family mortgages are loans given to single-family home owners, and are secured by Single family homes.
- The Single Family Residential loan mortgage loan industry in the US is a $14 trillion dollar industry with a YoY growth of 3.1%.

MORTGAGE INDUSTRY PLAYERS AND DATASETS

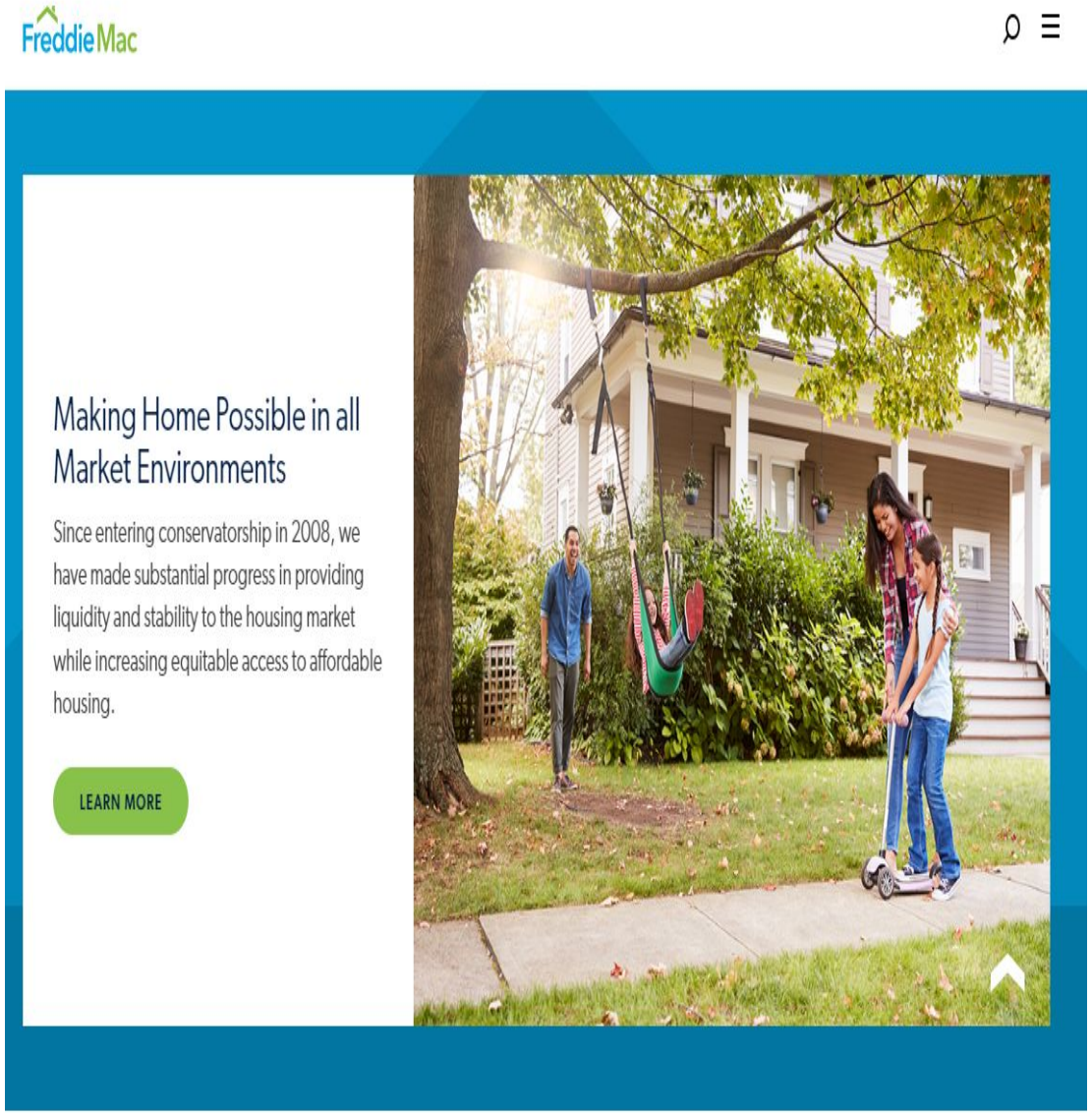


Image from : https://www.freddiemac.com/

- The primary mortgage market consists of lenders that originate mortgages for consumers.
- The secondary market consists of mortgage investors like Fannie Mae and Freddie Mac that buy mortgages from lending institutions to provide liquidity in the market
- The Federal Housing Finance Agency provides oversight over the Mortgage market
- Fannie and Freddie publish a wealth of data on their portfolios for public consumption, as part of compliance requirements

THE RESEARCH QUESTION & THE TOOL

Image from : https://www.istockphoto.com/photos/loan

- Using the publicly available datasets on the FreddieMac website, our research question is to analyze the interplay between loan origination attributes and loan performance. This analysis is useful in enhancing the understanding of origination indicators that may result in loan prepayment or delinquency.

- Data extraction from origination and performance data from the website, loading to a tool of choice, transforming it as required and use visualizations to understand the relationships were key steps.

- Since the performance dataset was large, we decided to use Power BI as it provided the end-to-end tool kit required to perform the intended operations

DATASET DESCRIPTION AND ACQUISITION

- To obtain the datasets, we were required to register with our email on the Freddie Mac website.

- From the Freddie Standard Datasets, we chose a sample origination dataset with 50,000 loans of the year 2018, and a corresponding performance dataset with monthly performance data from Jan'18 to Mar'23

- These datasets were in text format. Each of these datasets have 32 columns. The headers for each dataset were available in separate layout files.

- We thus obtained two datasets one with 50000 rows and the other with 1,647,912 rows and two header files for these datasets, which all needed to be extracted, loaded and transformed in to suitable data types to facilitate analysis.

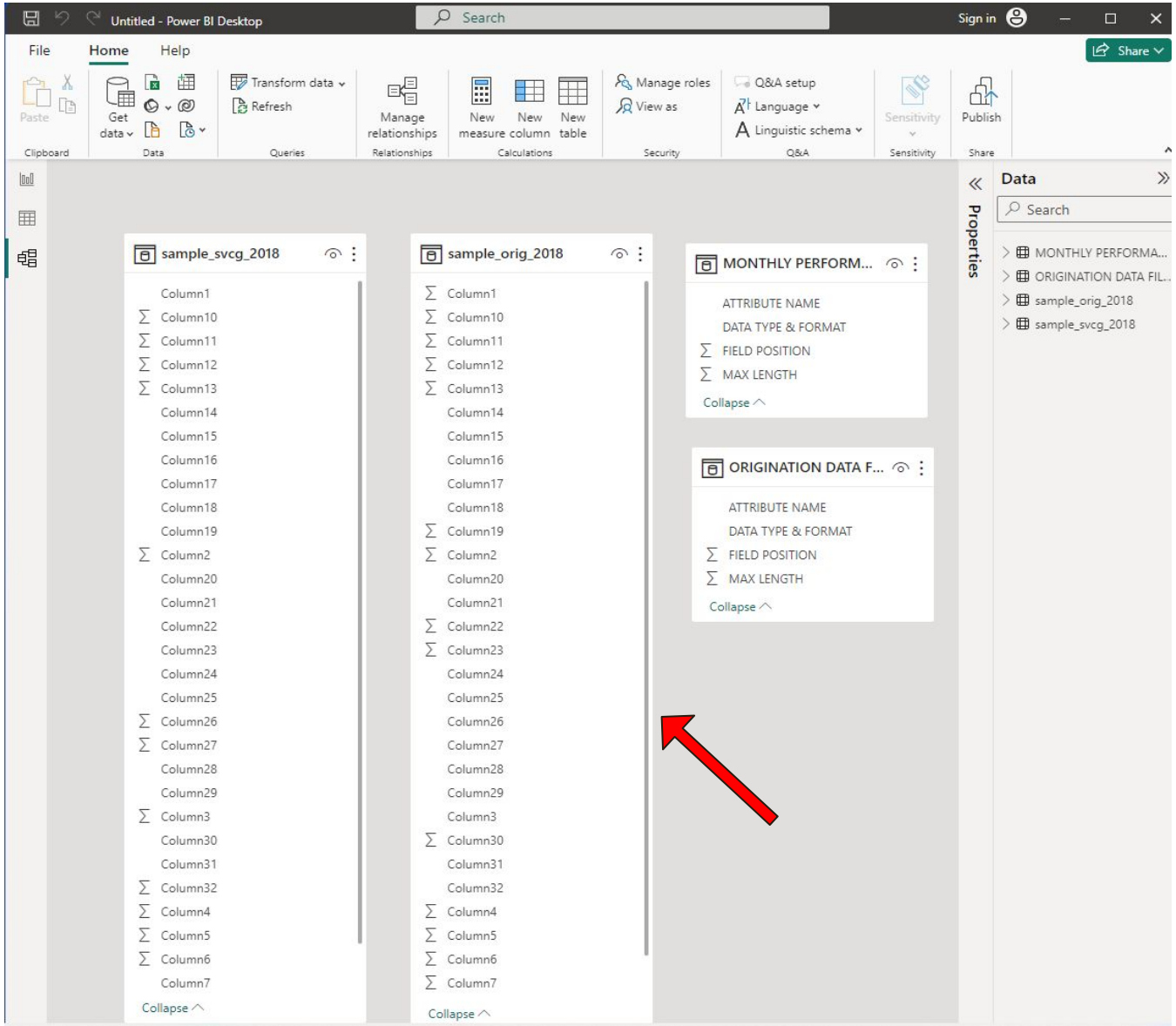
ORIGINATION DATA ATTRIBUTES

ORIGINATION DATA FILE			
FIELD POSITION	**ATTRIBUTE NAME**	**DATA TYPE & FORMAT**	**MAX LENGTH**
1	Credit Score	Numeric	4
2	First Payment Date	Date	6
3	First Time Homebuyer Flag	Alpha	1
4	Maturity Date	Date	6
5	Metropolitan Statistical Area (MSA) Or Metropolitan Division	Numeric	5
6	Mortgage Insurance Percentage (MI %)	Numeric	3
7	Number of Units	Numeric	2
8	Occupancy Status	Alpha	1
9	Original Combined Loan-to-Value (CLTV)	Numeric	3
10	Original Debt-to-Income (DTI) Ratio	Numeric	3
11	Original UPB	Numeric	12
12	Original Loan-to-Value (LTV)	Numeric	3
13	Original Interest Rate	Numeric - 6,3	6
14	Channel	Alpha	1
15	Prepayment Penalty Mortgage (PPM) Flag	Alpha	1
16	Amortization Type (Formerly Product Type)	Alpha	5
17	Property State	Alpha	2
18	Property Type	Alpha	2
19	Postal Code	Numeric	5
20	Loan Sequence Number	Alpha Numeric - PYYQnXXXXXXX	12
21	Loan Purpose	Alpha	1
22	Original Loan Term	Numeric	3
23	Number of Borrowers	Numeric	2
24	Seller Name	Alpha Numeric	60
25	Servicer Name	Alpha Numeric	60
26	Super Conforming Flag	Alpha	1
27	Pre-HARP Loan Sequence Number	Alpha Numeric - PYYQnXXXXXXX	12
28	Program Indicator	Alpha Numeric	1
29	HARP Indicator	Alpha	1
30	Property Valuation Method	Numeric	1
31	Interest Only (I/O) Indicator	Alpha	1
32	Mortgage Insurance Cancellation Indicator	Alpha	1

PERFORMANCE DATA ATTRIBUTES

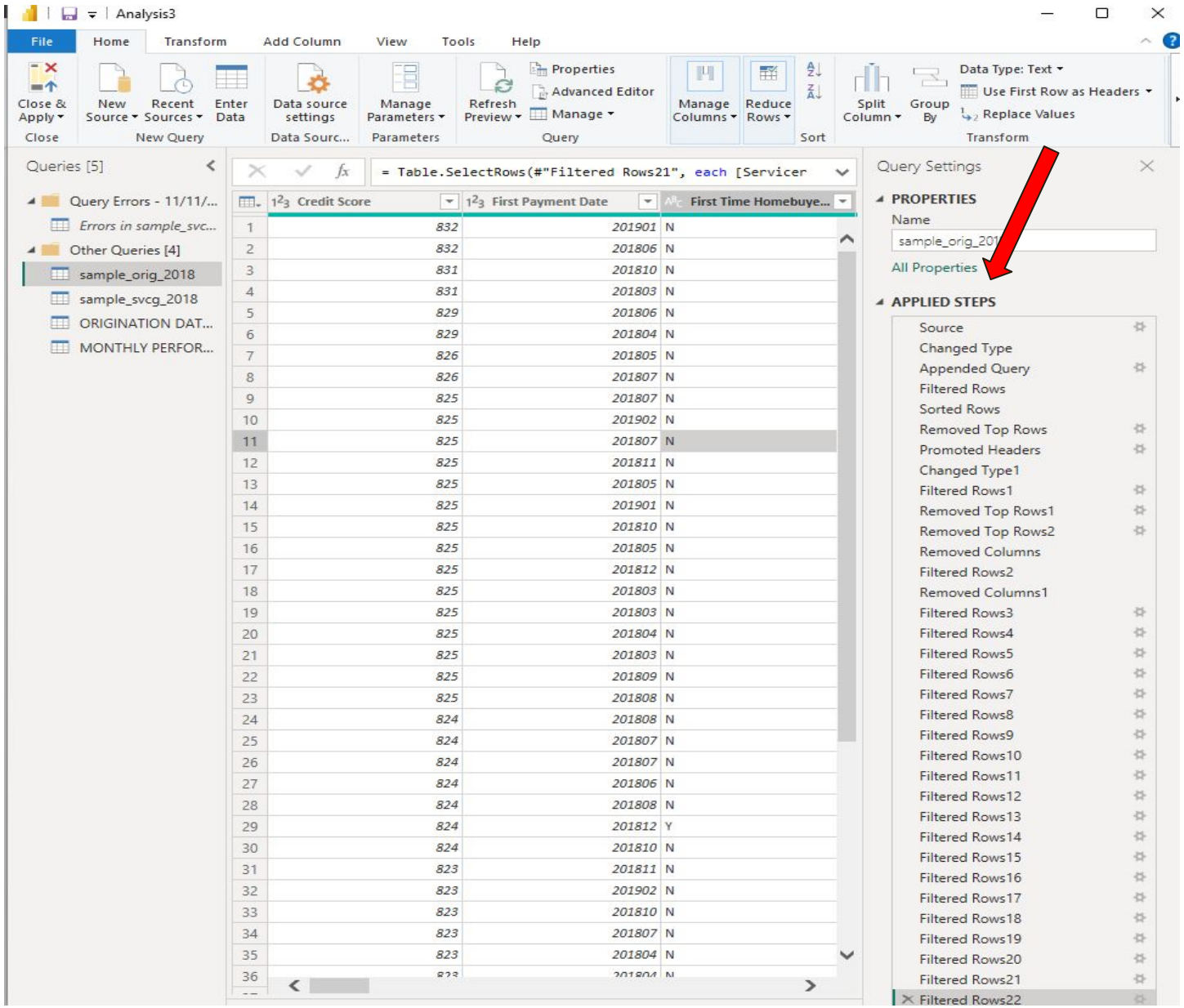
MONTHLY PERFORMANCE DATA FILE			
FIELD POSITION	ATTRIBUTE NAME	DATA TYPE & FORMAT	MAX LENGTH
1	Loan Sequence Number	Alpha Numeric - PYYQnXXXXXXX	12
2	Monthly Reporting Period	Date	6
3	Current Actual UPB	Numeric - 12,2	12
4	Current Loan Delinquency Status	Alpha Numeric	3
5	Loan Age	Numeric	3
6	Remaining Months to Legal Maturity	Numeric	3
7	Defect Settlement Date	Date	6
8	Modification Flag	Alpha	1
9	Zero Balance Code	Numeric	2
10	Zero Balance Effective Date	Date	6
11	Current Interest Rate	Numeric - 8,3	8
12	Current Deferred UPB	Numeric	12
13	Due Date of Last Paid Installment (DDLPI)	Date	6
14	MI Recoveries	Numeric - 12,2	12
15	Net Sales Proceeds	Alpha-Numeric	14
16	Non MI Recoveries	Numeric - 12,2	12
17	Expenses	Numeric - 12,2	12
18	Legal Costs	Numeric - 12,2	12
19	Maintenance and Preservation Costs	Numeric - 12,2	12
20	Taxes and Insurance	Numeric - 12,2	12
21	Miscellaneous Expenses	Numeric - 12,2	12
22	Actual Loss Calculation	Numeric - 12,2	12
23	Modification Cost	Numeric - 12,2	12
24	Step Modification Flag	Alpha	1
25	Deferred Payment Plan	Alpha	1
26	Estimated Loan-to-Value (ELTV)	Numeric	4
27	Zero Balance Removal UPB	Numeric - 12,2	12
28	Delinquent Accrued Interest	Numeric - 12,2	12
29	Delinquency Due to Disaster	Alpha	1
30	Borrower Assistance Status Code	Alpha	1
31	Current Month Modification Cost	Numeric - 12,2	12
32	Interest Bearing UPB	Numeric - 12,2	12

EXTRACT AND LOAD



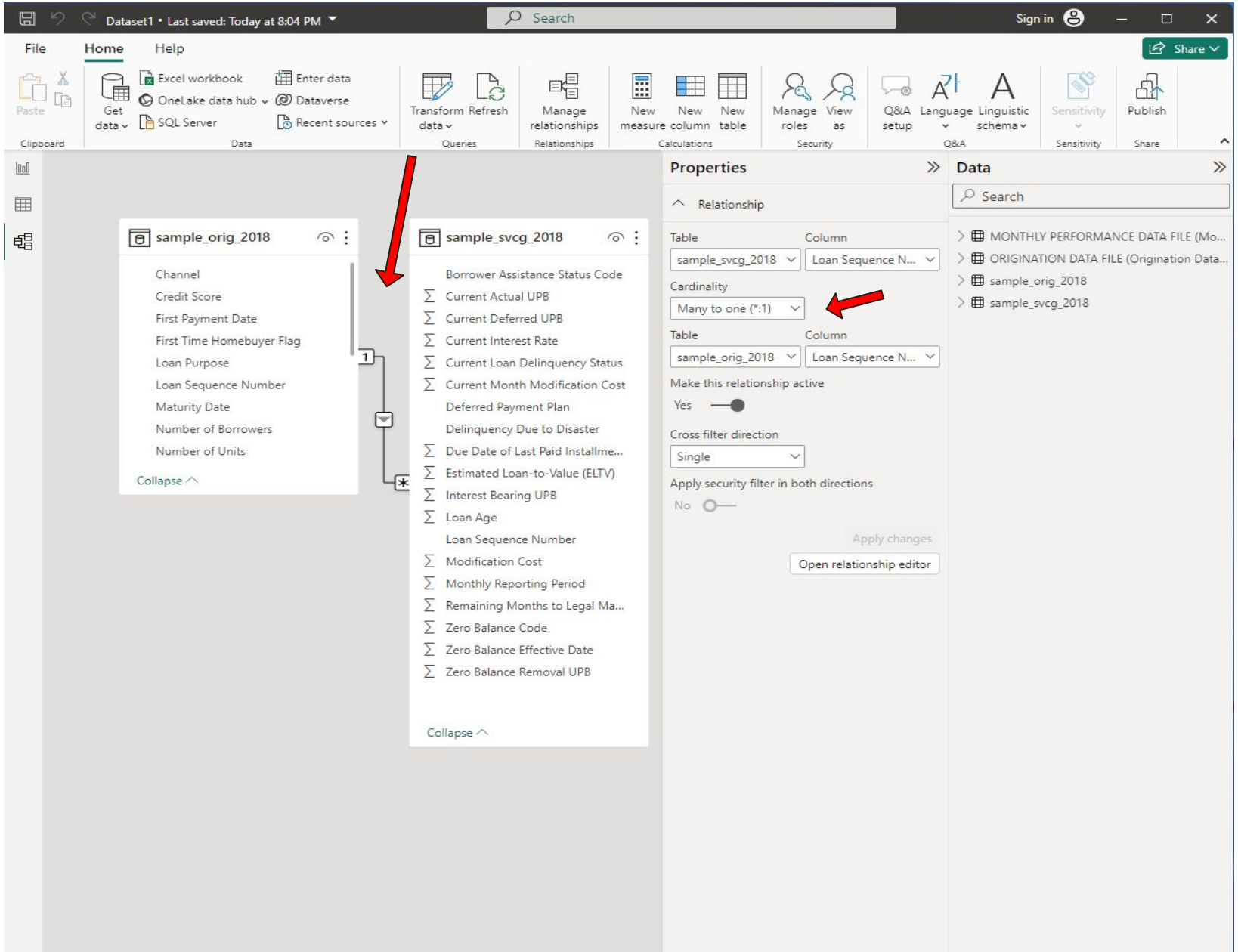
- Power BI can handle files from different sources and formats
- We loaded the large text files and one excel file with two worksheets with header data in columns
- In the home page, under tables option, we selected each table to inspect the columns
- The model option revealed the layouts

TRANSFORM



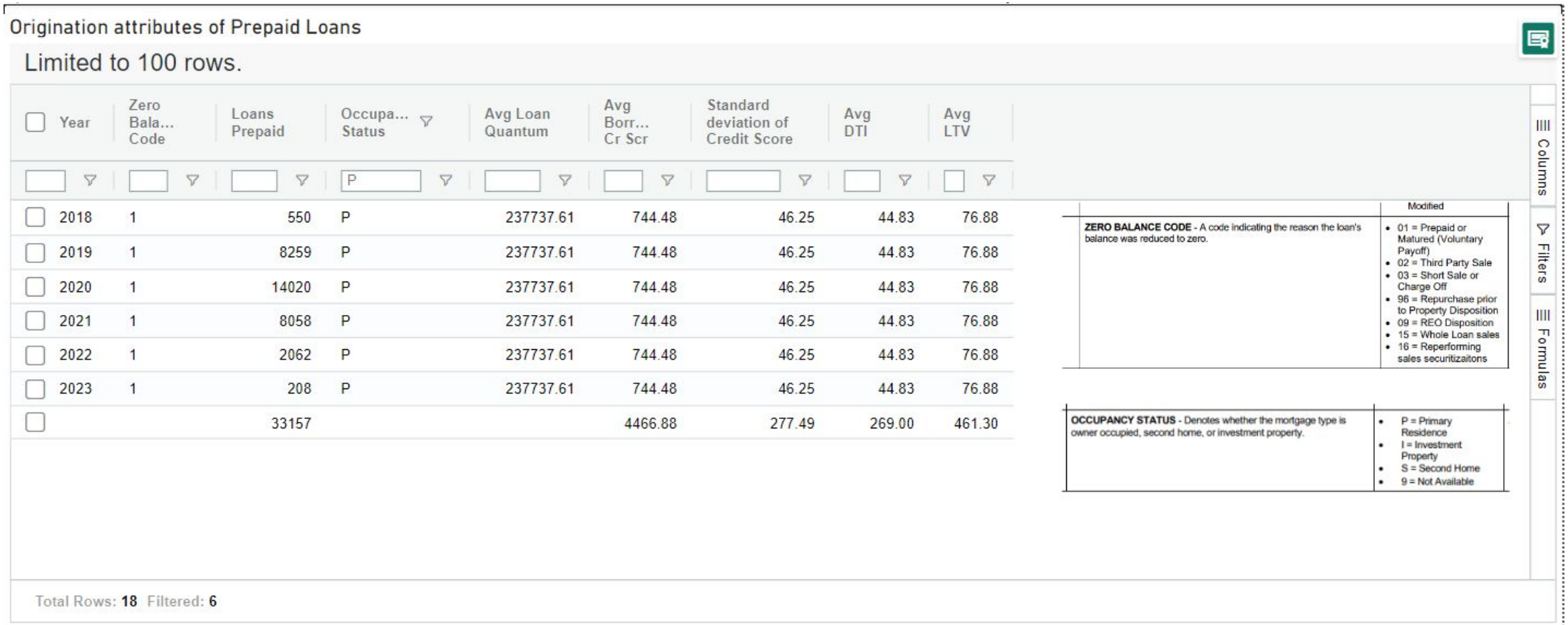
- In the dataset tables. there were many columns without data, the data types in most columns were incorrect
- Transform option opens Power query window. Using power query, for each dataset, we could correct data types and perform various steps listed under ‘applied steps’.
- From headers tables, required columns were transposed, added to dataset tables, sorted and promoted as headers.

CREATING THE MODEL



- After performing the data cleaning operations and creating headers, in the model pane, we found that the Loan sequence number column was common to both datasets.
- Power BI automatically created a many -to -one relationship, when we dragged this column from performance table to the origination table

REPORT ON INFLUENCE OF LOAN ORIGINATION ATTRIBUTES ON PREPAYMENT



Origination attributes of Prepaid Loans

Limited to 100 rows.

Year	Zero Bala... Code	Loans Prepaid	Occupa... Status	Avg Loan Quantum	Avg Borr... Cr Scr	Standard deviation of Credit Score	Avg DTI	Avg LTV
			P					
2018	1	550	P	237737.61	744.48	46.25	44.83	76.88
2019	1	8259	P	237737.61	744.48	46.25	44.83	76.88
2020	1	14020	P	237737.61	744.48	46.25	44.83	76.88
2021	1	8058	P	237737.61	744.48	46.25	44.83	76.88
2022	1	2062	P	237737.61	744.48	46.25	44.83	76.88
2023	1	208	P	237737.61	744.48	46.25	44.83	76.88
		33157			4466.88	277.49	269.00	461.30

Total Rows: **18** Filtered: **6**

Columns | Filters | Formulas

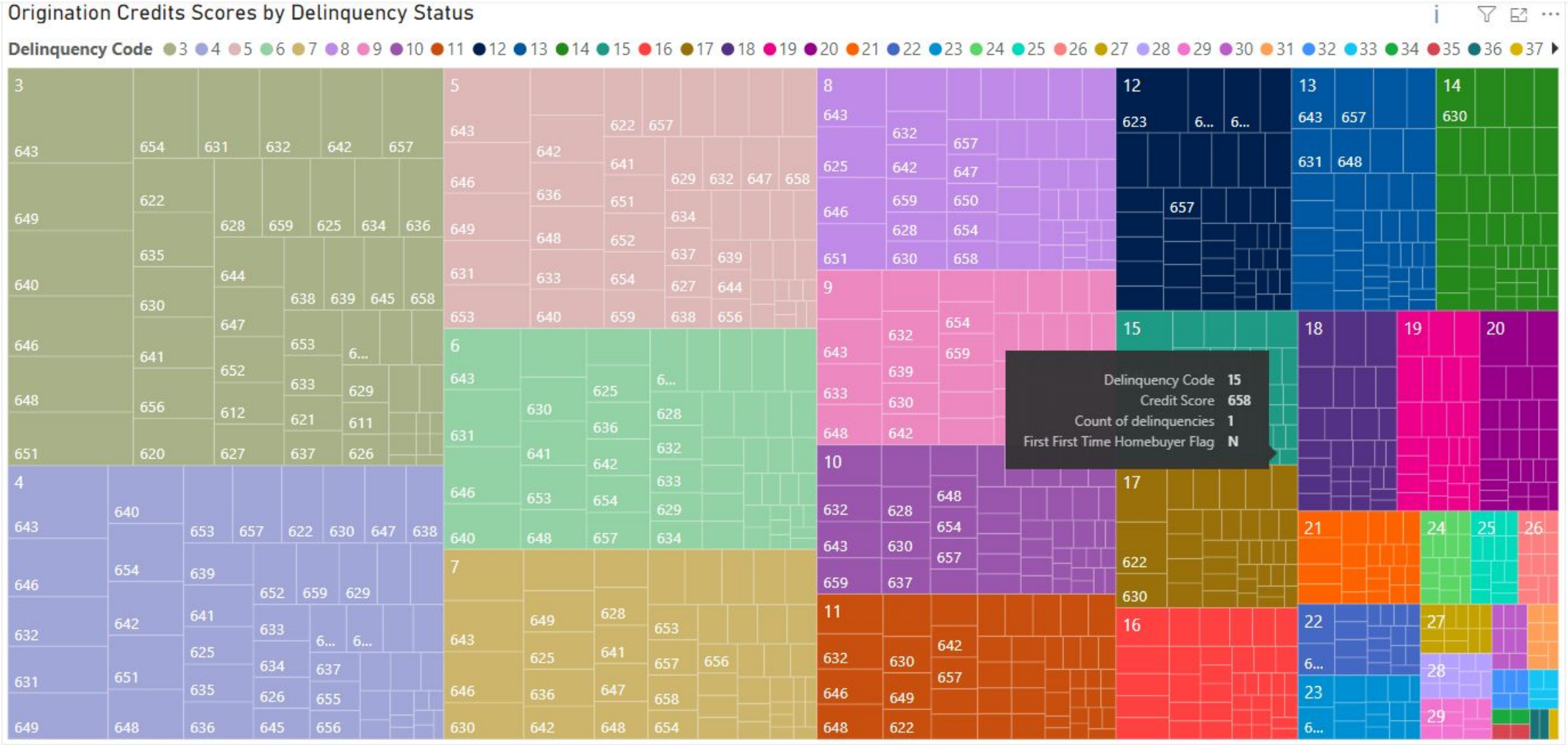
	Modified
ZERO BALANCE CODE - A code indicating the reason the loan's balance was reduced to zero.	• 01 = Prepaid or Matured (Voluntary Payoff) • 02 = Third Party Sale • 03 = Short Sale or Charge Off • 96 = Repurchase prior to Property Disposition • 09 = REO Disposition • 15 = Whole Loan sales • 16 = Reperforming sales securitizaitons

OCCUPANCY STATUS - Denotes whether the mortgage type is owner occupied, second home, or investment property.	• P = Primary Residence • I = Investment Property • S = Second Home • 9 = Not Available

Name 'sample_orig_2018'[

In this visual, we have reported the average values for Origination loan attributes for prepaid loans, indicated by zero balance code 1. The columns Occupancy status, average loan quantum, Borrower credit score, avg DTI, avg LTI are from the origination dataset. The Year column indicating the year of the monthly reporting period and the zero balance code are from the performance file . Using the above table based on the Freddie sample dataset for 2018, we can interpret that while a total of 37511 loans originated prepaid or closed by 2023, 33157 of these closures were made by borrowers who bought the underlying property as their primary residence (Occupancy Status 'P'). The average credit score of this category of borrowers is 744 and the standard deviation is 46, indicating high variability in the credit scores of borrowers that have prepaid loans. This indicates that the origination attributes from the dataset alone are not enough to draw conclusions about prepayment of loans and external economic factors and interest rates such as fed funds rate may influence movement of loans from one lender to another in which case there may be prepayments by obtaining mortgages from other vendors.

INTERPLAY BETWEEN CREDIT SCORE AND LOAN DEFAULT



This visual only shows loans that have delinquencies pertaining to late payments beginning from delinquency code 3 , which is 90-119 days delinquent and a highest code of 43 which implies 1289 days delinquency. The credit score of these loans is captured at the time of Origination. Thought the relationship between credit score and delinquency is not linear, the maximum Borrower credit score at origination for any of the delinquent borrowers has not reached the threshold of 675. All the delinquent borrowers are not first time home buyers indicating the possibility of their having other commitments. From this report we can infer that borrowers that are not first time home buyers, and borrower with origination credit scores lower than 675 display a tendency for default on their EMI commitments.

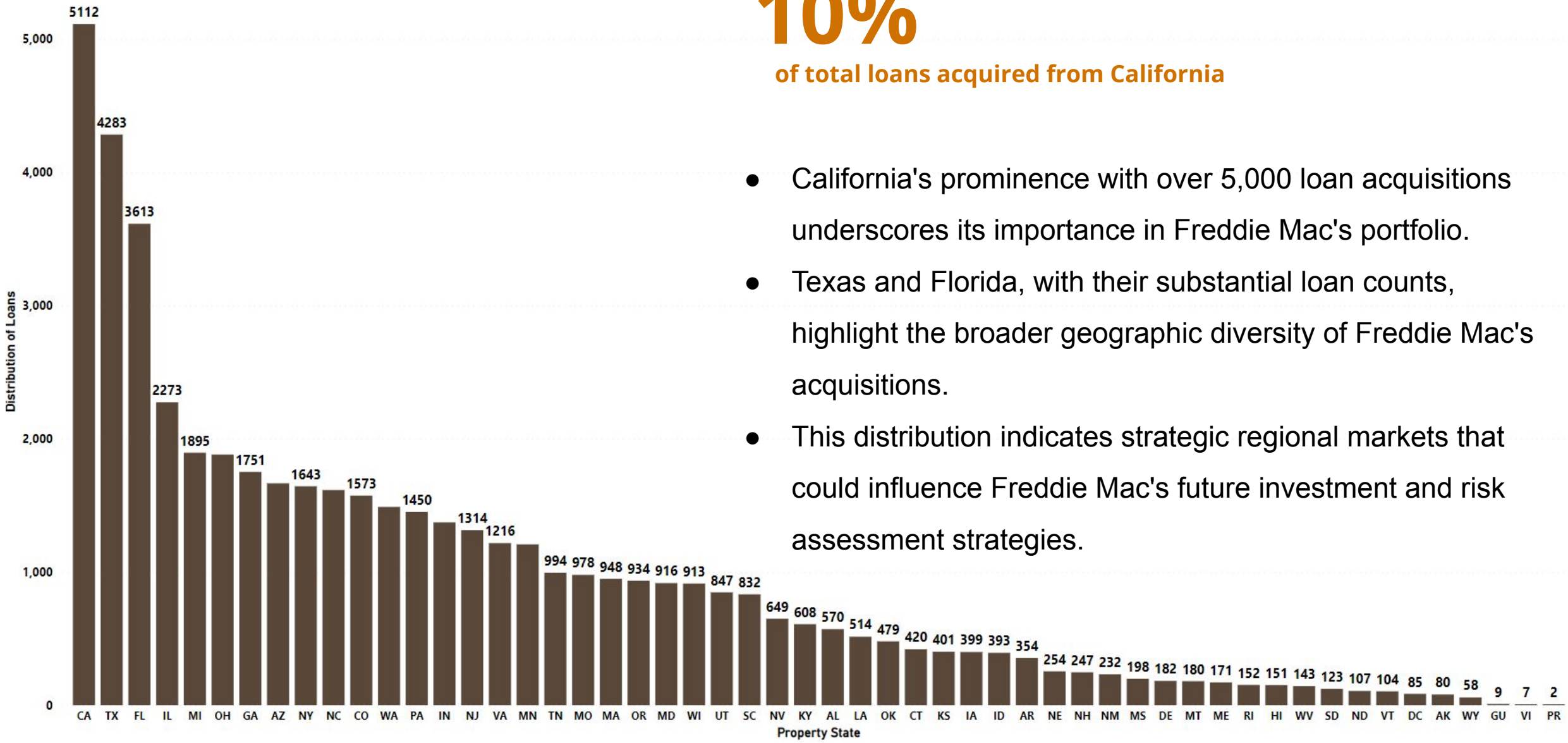
Challenges - ETL

Challenge in exploring the datasets: Initially we wanted to explore the datasets in excel. This was not possible with the performance data file since it had 1,647,912 rows. As the datasets were text files, we loaded them to Notepad ++ which was not very helpful. We could not initially determine how many rows were populated for important columns such as zero balance code and current default values.

Lack of familiarity with PowerBI : We learned to use Power BI in this course through lab work which was quite helpful in understanding the capabilities of the tool. However, we did not have the required practice in using the various options and therefore, spent a lot of time, going through the lab work link again and again and also a variety of Youtube videos on modeling. This video from Avi Singh was particularly helpful.

Data types : The original files were in text format. When loaded to Power BI, many fields were recognized in incorrect formats. Especially, we struggled a lot with changing date data types and could not manage to change them in the transformation stage. We finally learned a bit of DAX code to add new columns with correct formats in the Data modeling stage.

Number of Loan Acquired By Property State



10%

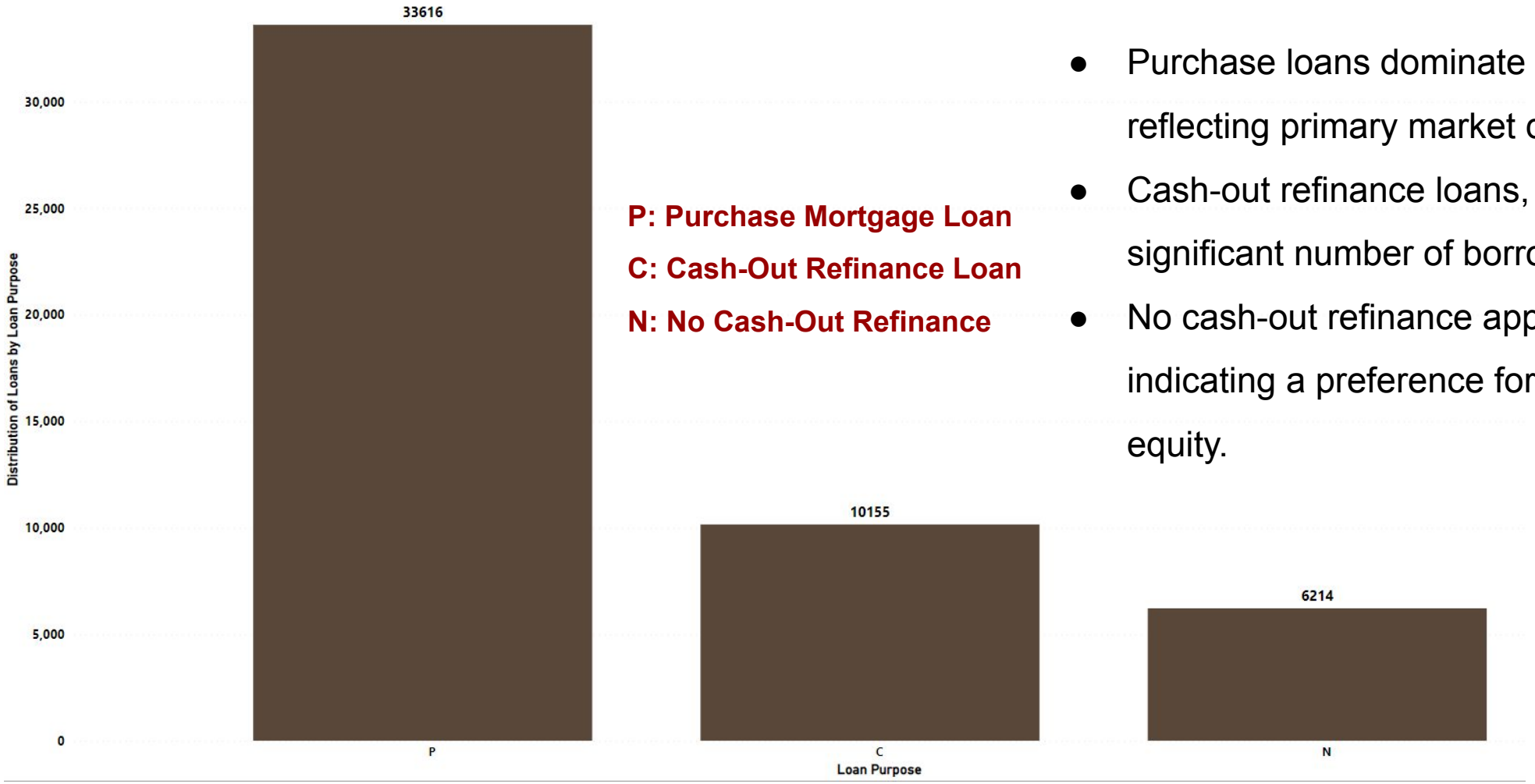
of total loans acquired from California

- California's prominence with over 5,000 loan acquisitions underscores its importance in Freddie Mac's portfolio.
- Texas and Florida, with their substantial loan counts, highlight the broader geographic diversity of Freddie Mac's acquisitions.
- This distribution indicates strategic regional markets that could influence Freddie Mac's future investment and risk assessment strategies.

Number of Loan Acquired By Purchase Loan Category

67%

of total loans acquired from Purchase Loan Category



P: Purchase Mortgage Loan

C: Cash-Out Refinance Loan

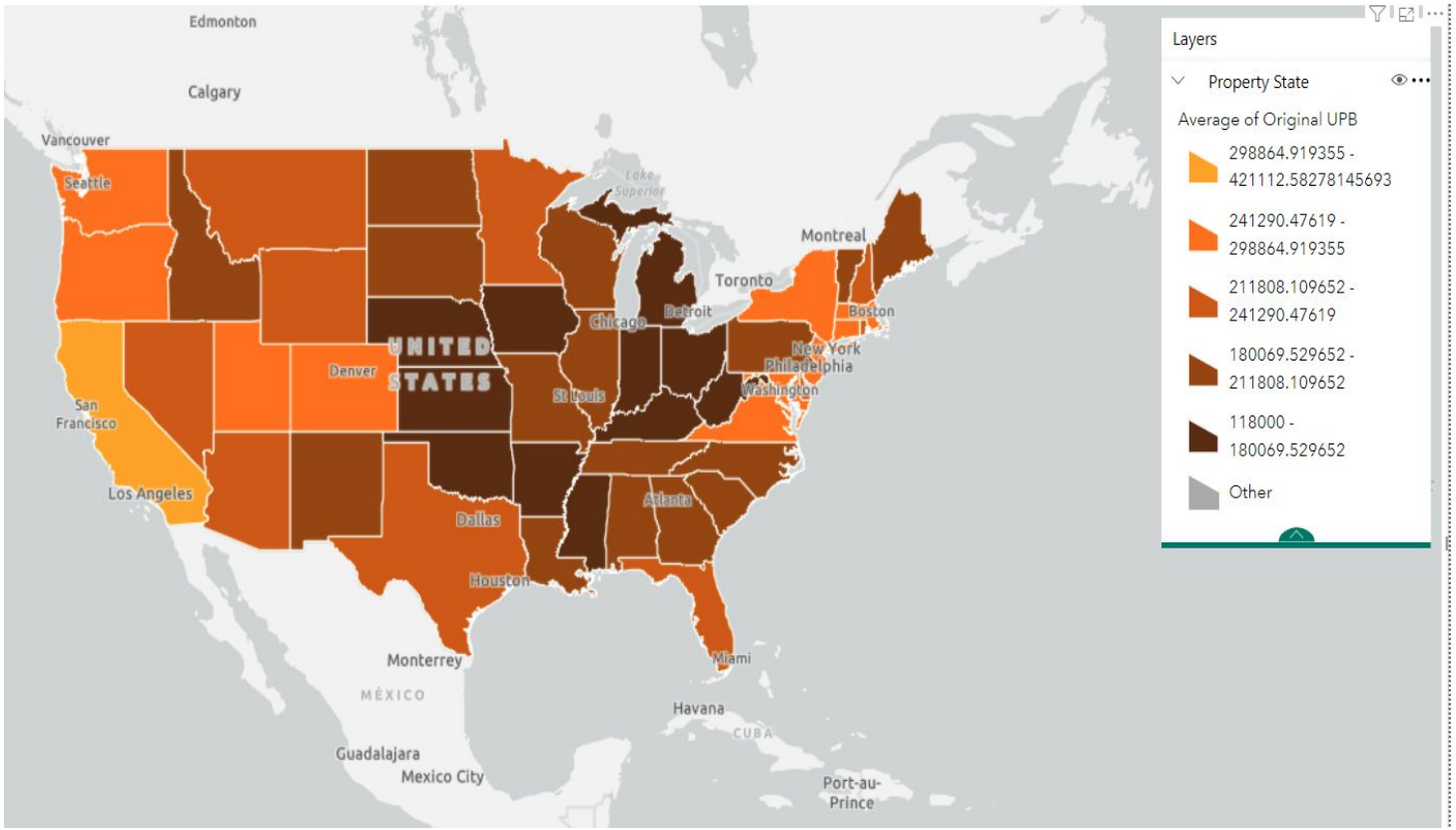
N: No Cash-Out Refinance

- Purchase loans dominate with 33,616 applications, reflecting primary market demand.
- Cash-out refinance loans, at 10,155, suggest a significant number of borrowers leveraging equity.
- No cash-out refinance applications stand at 6,214, indicating a preference for maintaining existing equity.

Original Unpaid Principal Balance(UPB) by State

$340,577
average UPB in California

**Unpaid principal balance (UPB) is the portion of a loan (e.g. a mortgage loan) at a certain point in time that has not yet been remitted to the lender.

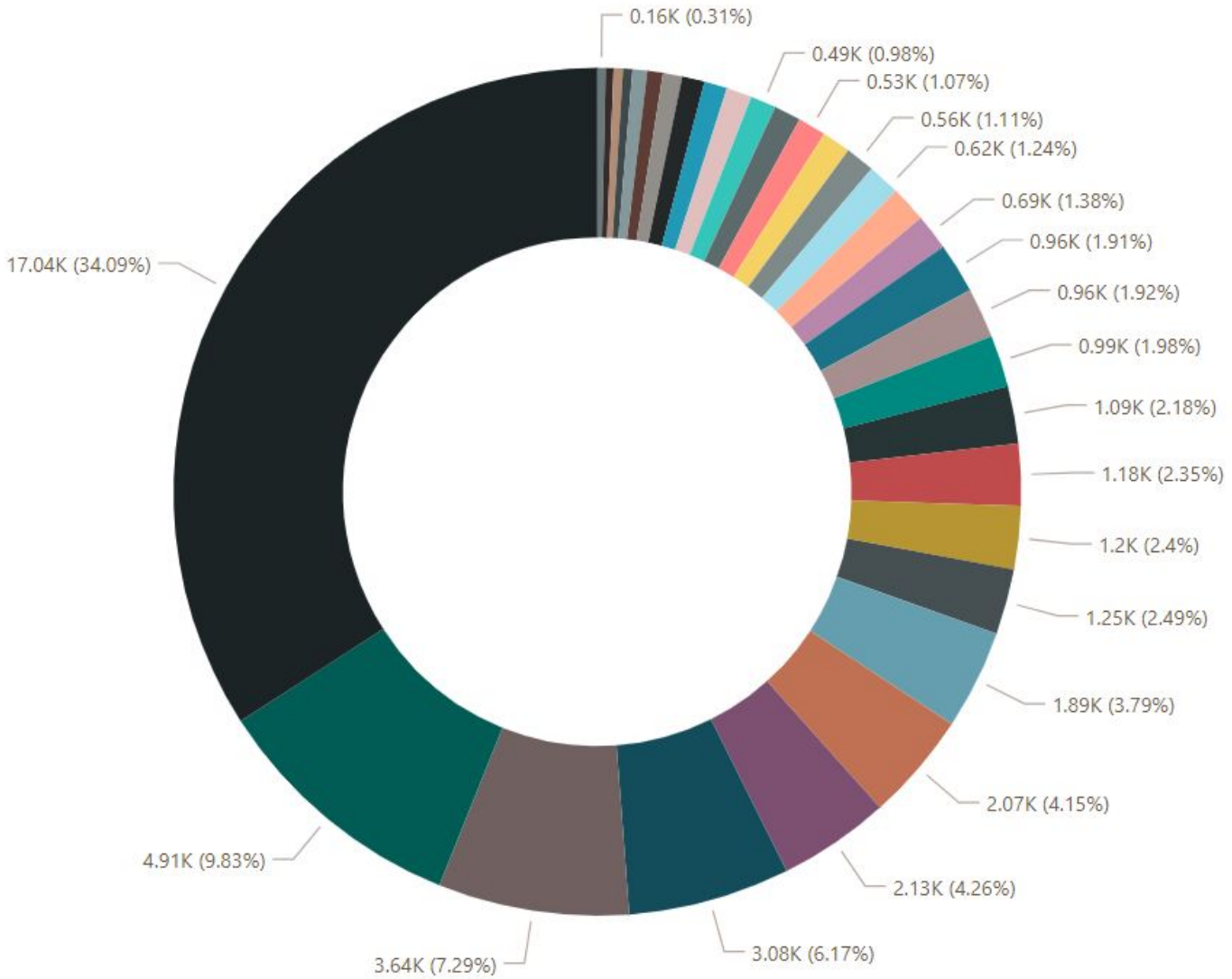


- The graph showcases California, Texas, and Florida as leaders in UPB, reflecting concentrated lending activities and potentially higher real estate values.
- States with minimal UPB, such as North Dakota and Alaska, could indicate lower loan issuance, possibly due to smaller populations or lower property costs.
- This UPB distribution aids in understanding regional market dynamics, vital for strategic planning in loan origination and risk mitigation.

Number of Loans Sanctioned By Sellers

4913 (9.83%)

loans sanctioned by Wells Fargo



Seller Name

- PRIMELENDING A PLAINS CAPITAL CO
- STEARNS LENDING, LLC.
- FIFTH THIRD MORTGAGE COMPANY
- TEXAS CAPITAL BANK, N.A.
- GUILD MORTGAGE COMPANY
- GUARANTEED RATE, INC.
- PRIMELENDING, A PLAINSCAPITAL COMPANY
- FINANCE OF AMERICA MORTGAGE LLC
- USAA FEDERAL SAVINGS BANK
- SUNTRUST MORTGAGE, INC.
- SUNTRUST BANK
- CITIZENS BANK, NA
- STEARNS LENDING, LLC
- HOME POINT FINANCIAL CORPORATION
- TEXAS CAPITAL BANK
- FRANKLIN AMERICAN MORTGAGE COMPANY
- UNITED SHORE FINANCIAL SERVICES, LLC
- PROVIDENT FUNDING ASSOCIATES, L.P.
- JPMORGAN CHASE BANK, NATIONAL ASSOCIATION
- BRANCH BANKING & TRUST COMPANY
- NATIONSTAR MORTGAGE LLC DBA MR. COOPER
- LOANDEPOT.COM, LLC
- UNITED SHORE FINANCIAL SERVICES, LLC, DBA UNITED WHOLESALE M
- FLAGSTAR BANK, FSB
- FAIRWAY INDEPENDENT MORTGAGE CORPORATION
- U.S. BANK N.A.
- AMERIHOME MORTGAGE COMPANY, LLC
- CALIBER HOME LOANS, INC.
- JPMORGAN CHASE BANK, N.A.
- QUICKEN LOANS INC.
- WELLS FARGO BANK, N.A.
- Other sellers

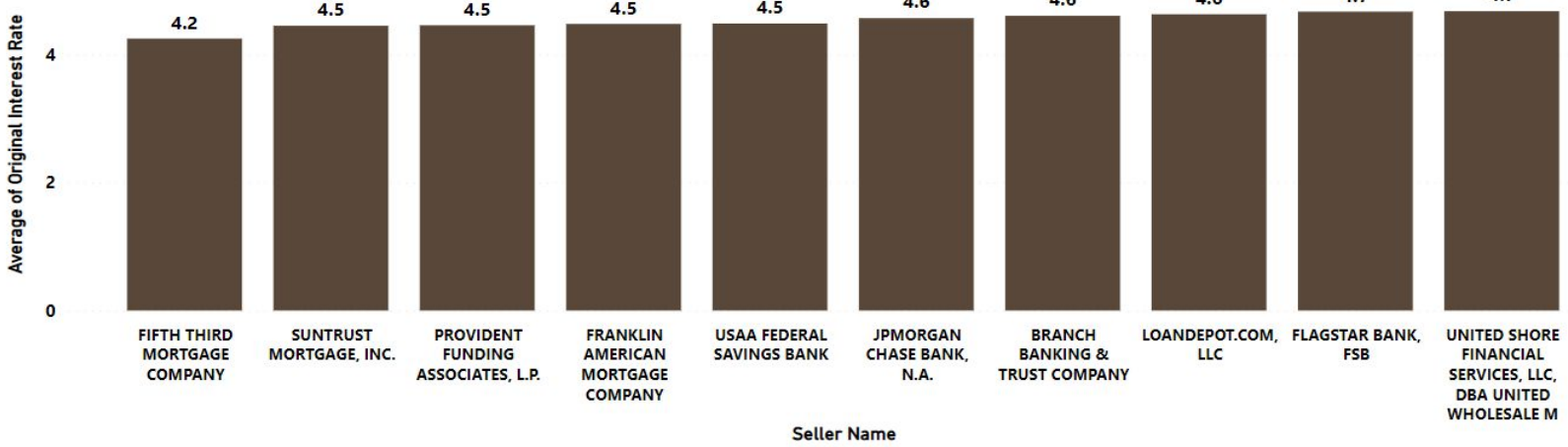
Average Interest Rate Offered by Sellers

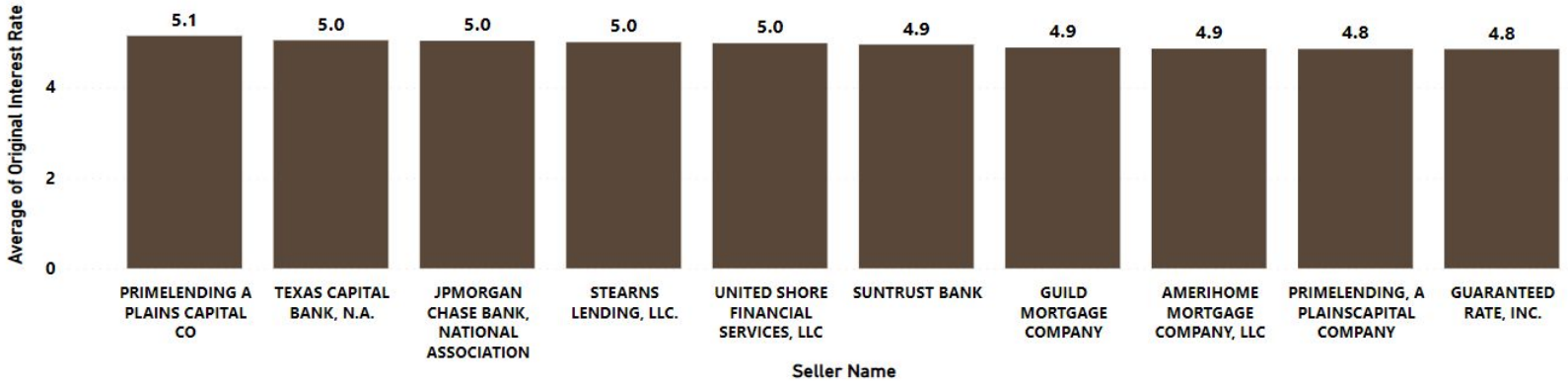
4.2%

lowest average interest rate

5.1%

highest average interest rate





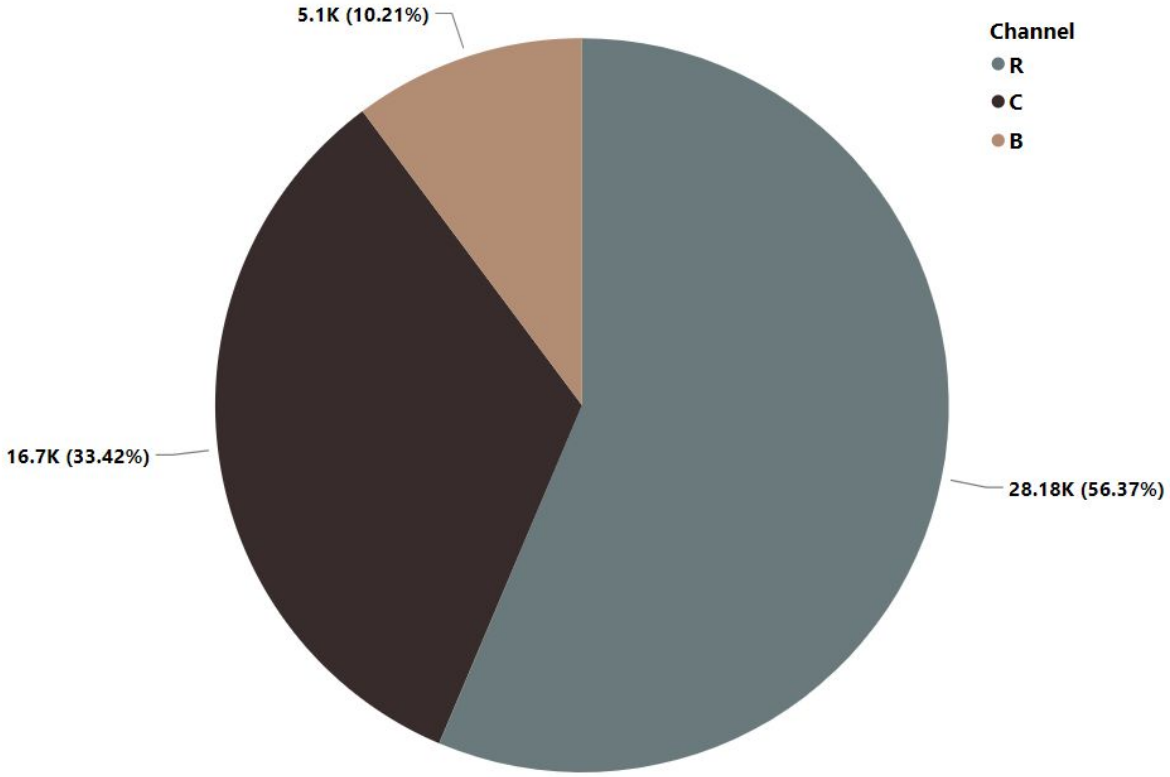
- Fifth Third and SunTrust offer the lowest interest rates, averaging 4.2% and 4.5% respectively, hinting at aggressive pricing strategies to attract borrowers.
- United Shore, with higher rates averaging 4.7%, may reflect a focus on niche markets or risk pricing.
- The variation between the lowest (4.2%) and highest (5.1%) average rates offered highlights competitive positioning among sellers in the mortgage industry.

Distribution of Freddie Mac's Acquired Loans

R: Retail - Direct consumer lending.

B: Broker - Third-party originated loans.

C: Correspondent - Loans sold to lenders.

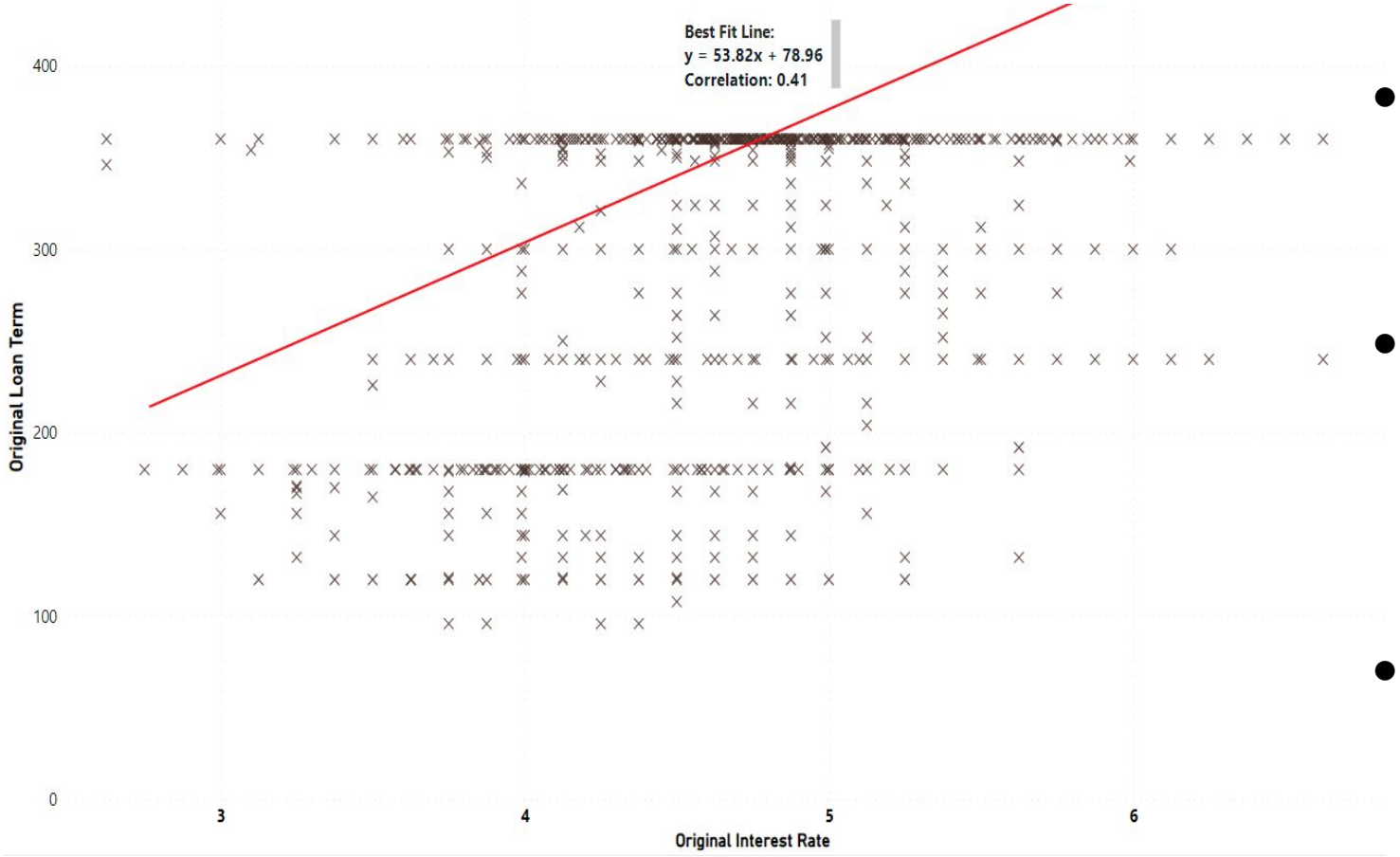


56.37%

loans acquired by retail channels

- The graph showcases the channels through which Freddie Mac has acquired loans, with retail (direct purchases) being the most significant at 56.37%, indicating a robust direct acquisition stream.
- Broker and correspondent channels, at 33.42% and 10.21%, highlight strategic partnerships and a broad network of loan sourcing, essential for diversifying Freddie Mac's loan portfolio and risk.
- This composition underlines the importance of varied acquisition channels in Freddie Mac's operational model, ensuring a steady influx of loans for their securitization and investment offerings.

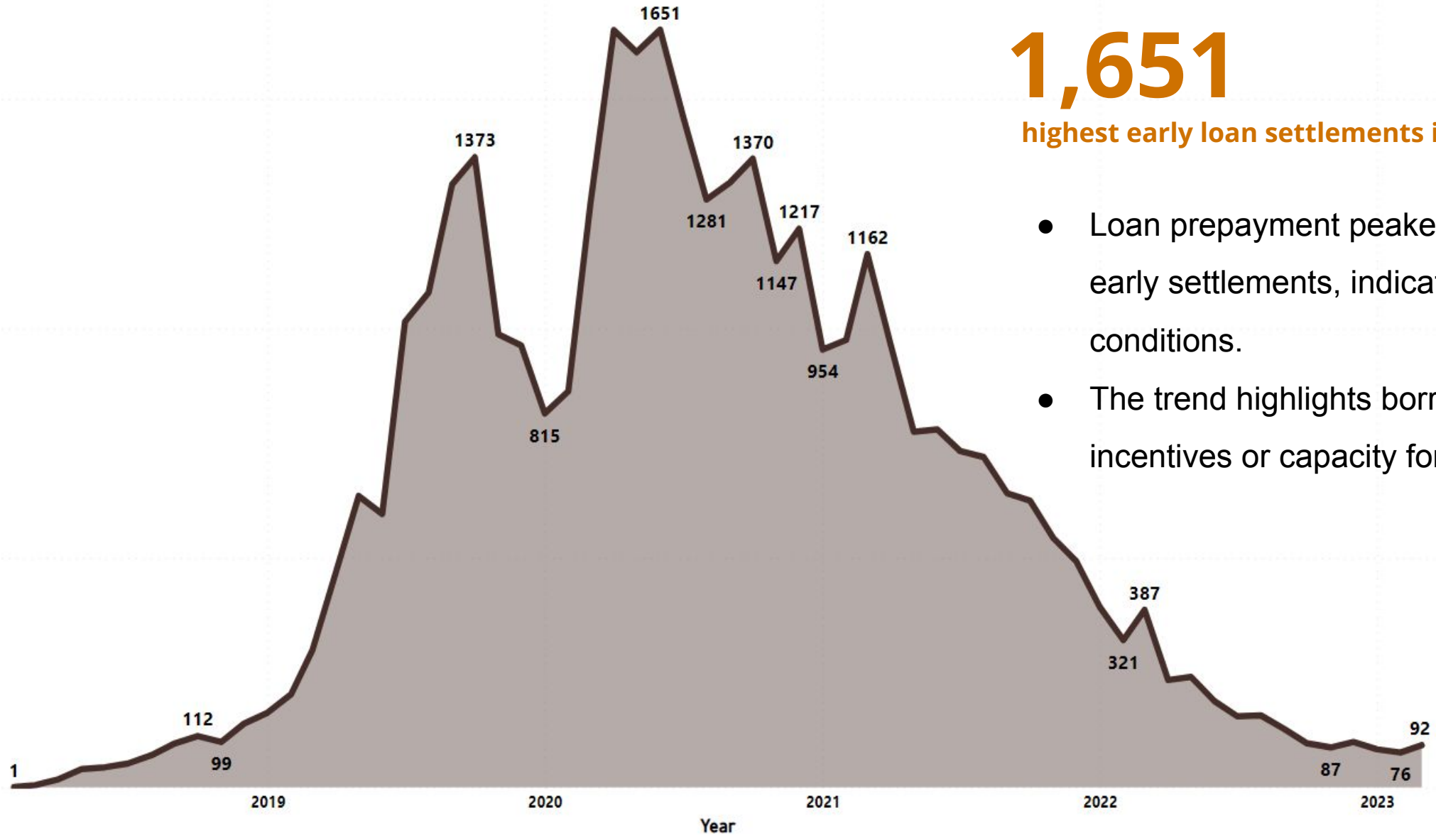
Original Interest Rate and Original Loan Term



Correlation: 0.41

- A moderate correlation (0.41) suggests interest rates influence loan term length to some extent.
- The best fit line with an equation y=53.82x+78.96 indicates that higher interest rates may be associated with longer loan terms.
- The data dispersion highlights diverse borrower preferences and financial product offerings.

Loan Prepayment Trend over Time



1,651

highest early loan settlements in June,2020

- Loan prepayment peaked in 2020 with 1,651 early settlements, indicating a response to market conditions.
- The trend highlights borrower's changing incentives or capacity for prepayment over time.

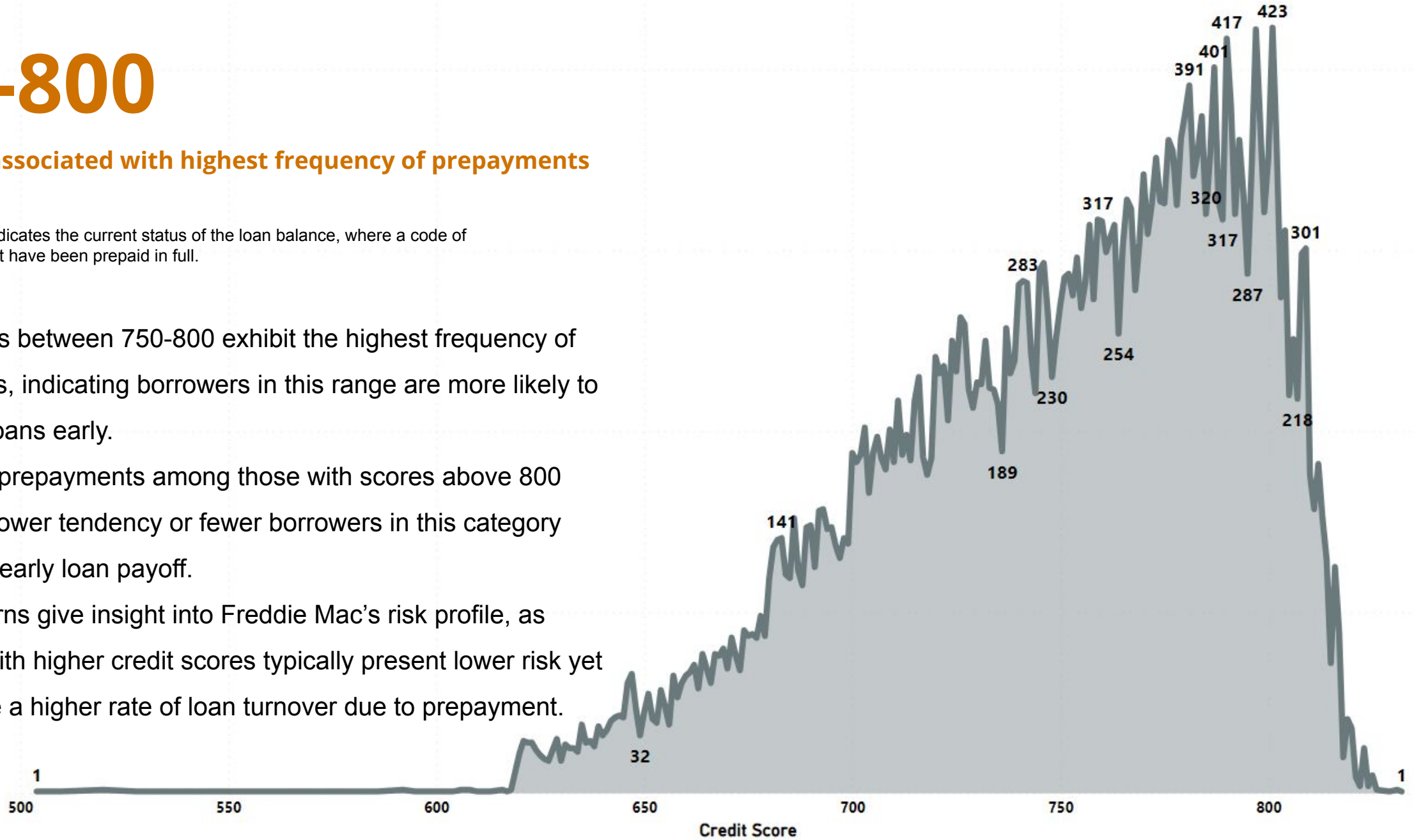
Credit Score Influence on Loan Prepayment Frequencies

750-800

credit score associated with highest frequency of prepayments

**Zero Balance Code indicates the current status of the loan balance, where a code of "1" represents loans that have been prepaid in full.

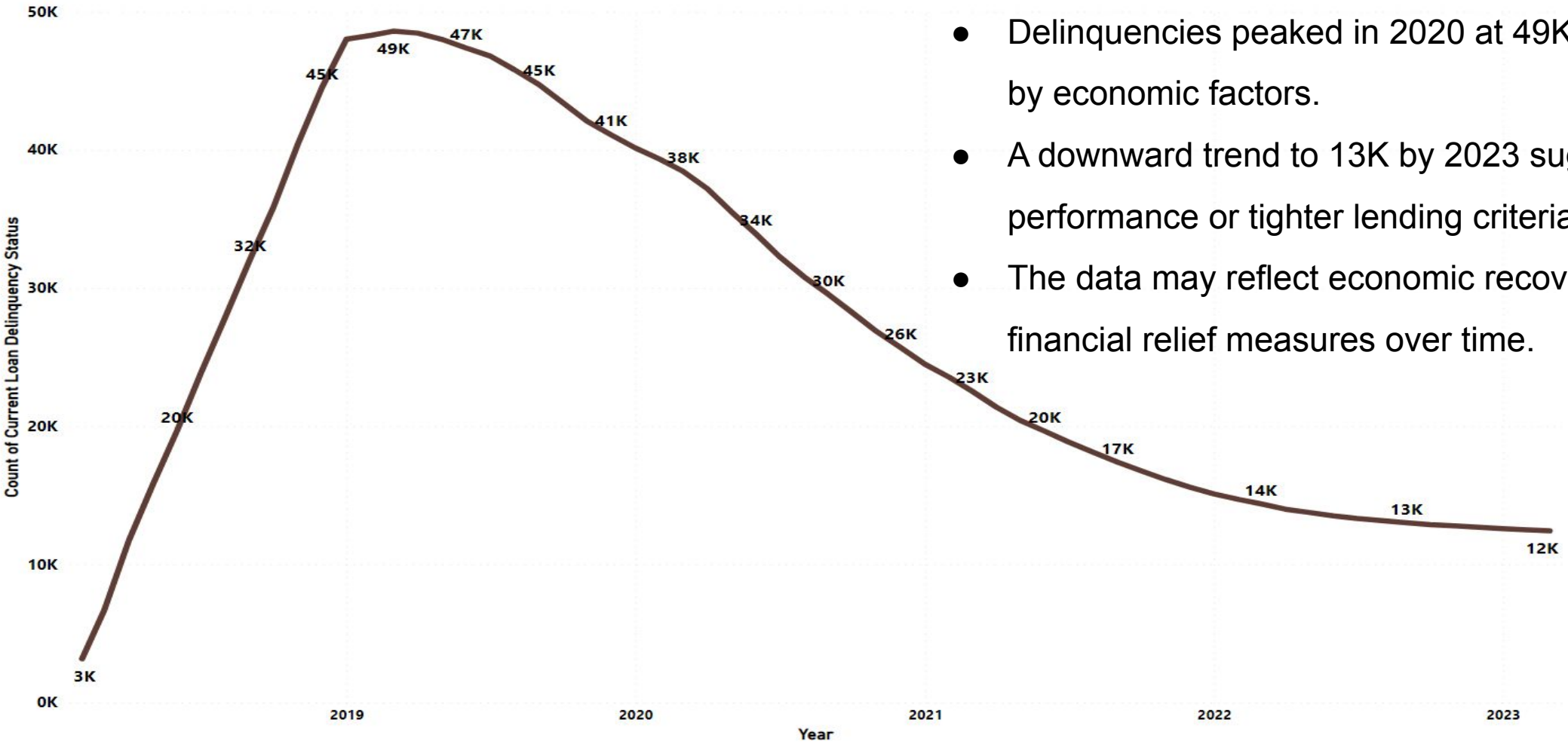
- Credit scores between 750-800 exhibit the highest frequency of prepayments, indicating borrowers in this range are more likely to settle their loans early.
- A decline in prepayments among those with scores above 800 suggests a lower tendency or fewer borrowers in this category engaging in early loan payoff.
- These patterns give insight into Freddie Mac's risk profile, as borrowers with higher credit scores typically present lower risk yet demonstrate a higher rate of loan turnover due to prepayment.



Current Loan Delinquency Status over Time

49,000

highest loan delinquencies in year 2020



- Delinquencies peaked in 2020 at 49K, possibly influenced by economic factors.
- A downward trend to 13K by 2023 suggests improving loan performance or tighter lending criteria.
- The data may reflect economic recovery and the impact of financial relief measures over time.

Delinquency Due to Disaster

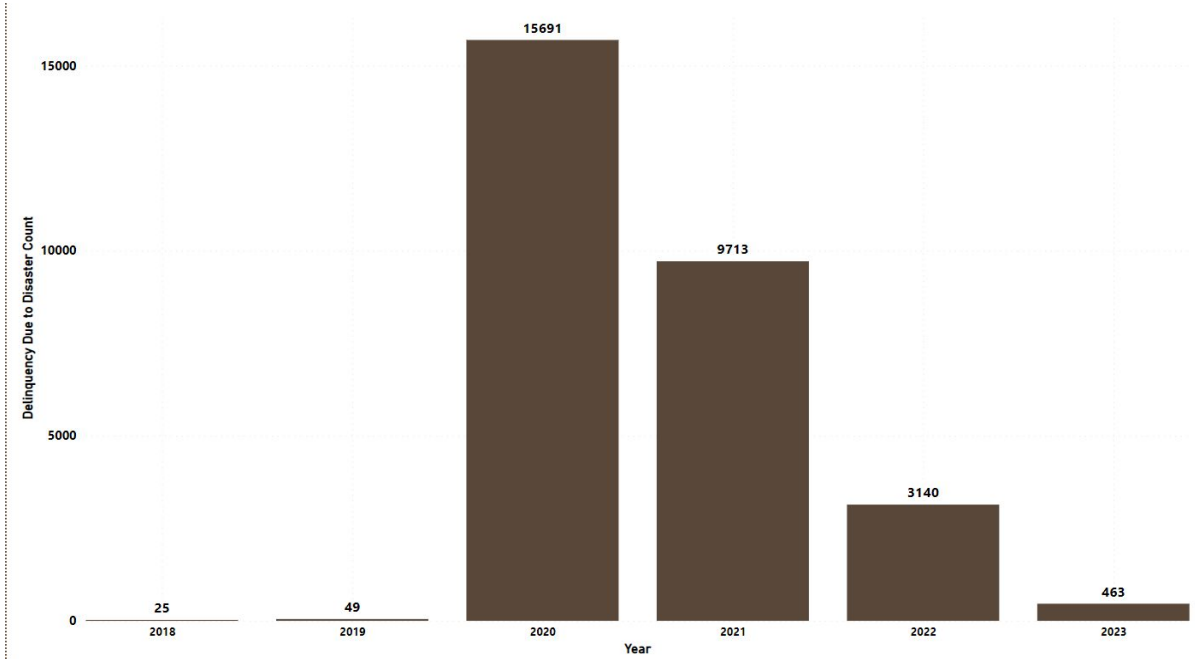
- The spike in delinquencies due to disasters in 2020, with 15,691 cases, aligns with increased natural disaster occurrences and may reflect the financial impact of such events on homeowners.
- The geographic distribution shows higher disaster-related delinquencies concentrated in specific regions, notably around California, Texas and Florida indicating areas more affected by disasters during this period.
- A comparison of the two visuals underscores the temporal and spatial impact of disasters on loan delinquencies, with significant year-on-year fluctuations and regional variances.

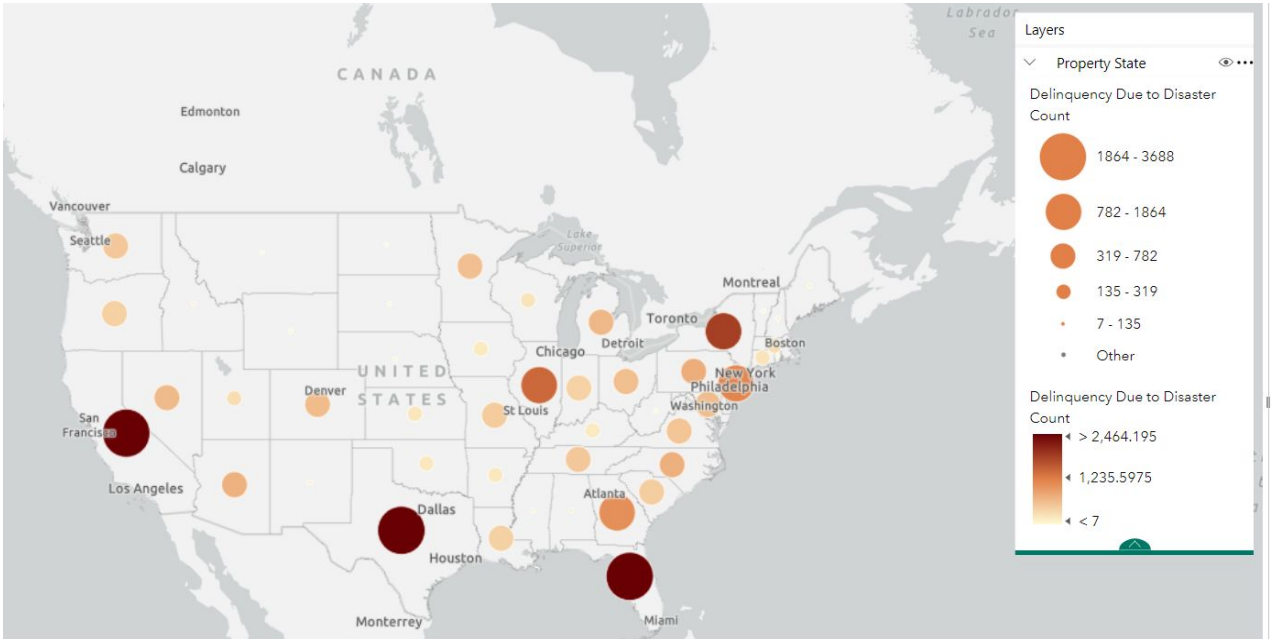
15,691

delinquencies due to disaster in 2020

64.4%

combined delinquencies in California, Texas and Florida





References

Power BI App download: https://app.powerbi.com/home?experience=power-bi

Fannie Mae website: https://www.fanniemae.com/

Freddie Mac website: https://www.freddiemac.com/

Image for slide 2 from : https://www.homes.com/adelphi-md/neighborhood/adelphi/

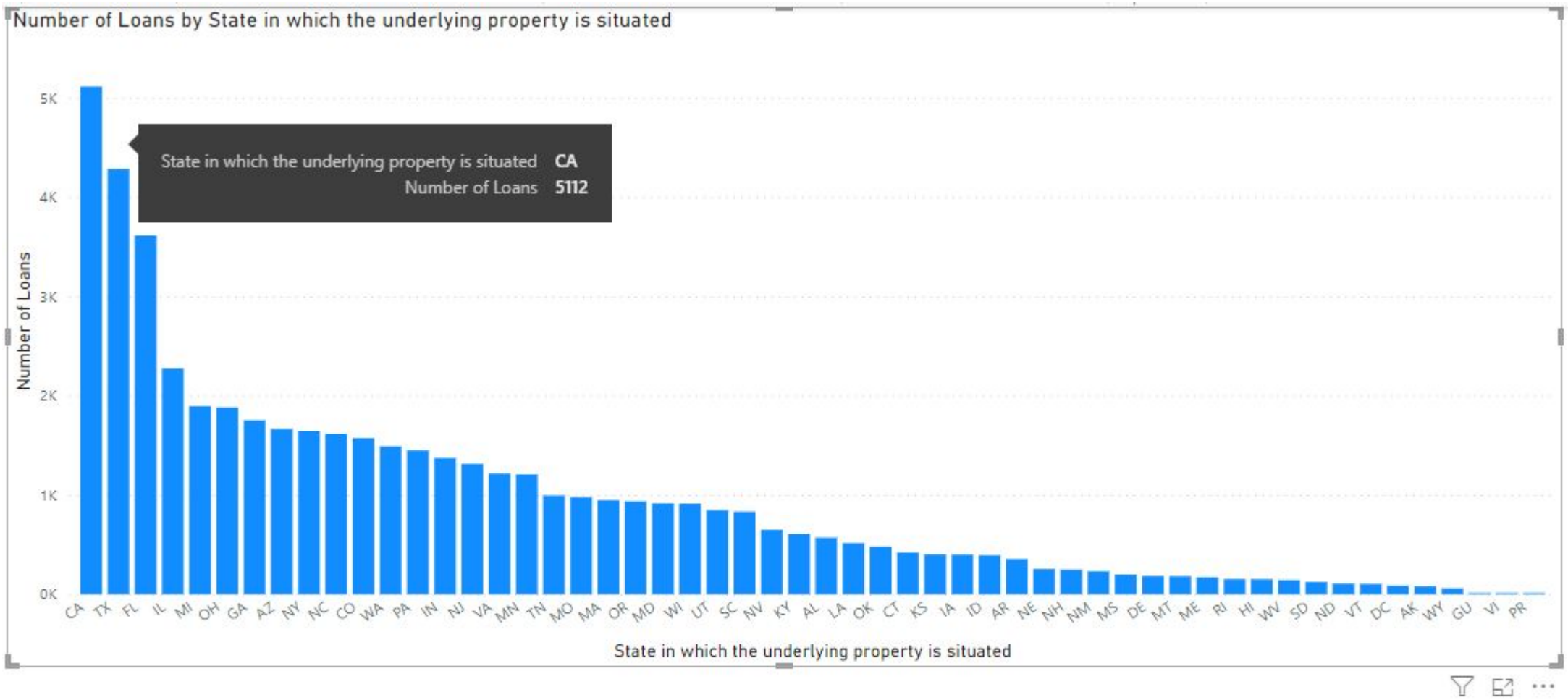
Image for slide 3 from : https://www.freddiemac.com/

Image for slide 4 from : https://www.istockphoto.com/photos/loan

Youtube video on Data Modeling with PowerBI: https://www.youtube.com/watch?v=AGrl-H87pRU&t=1785s

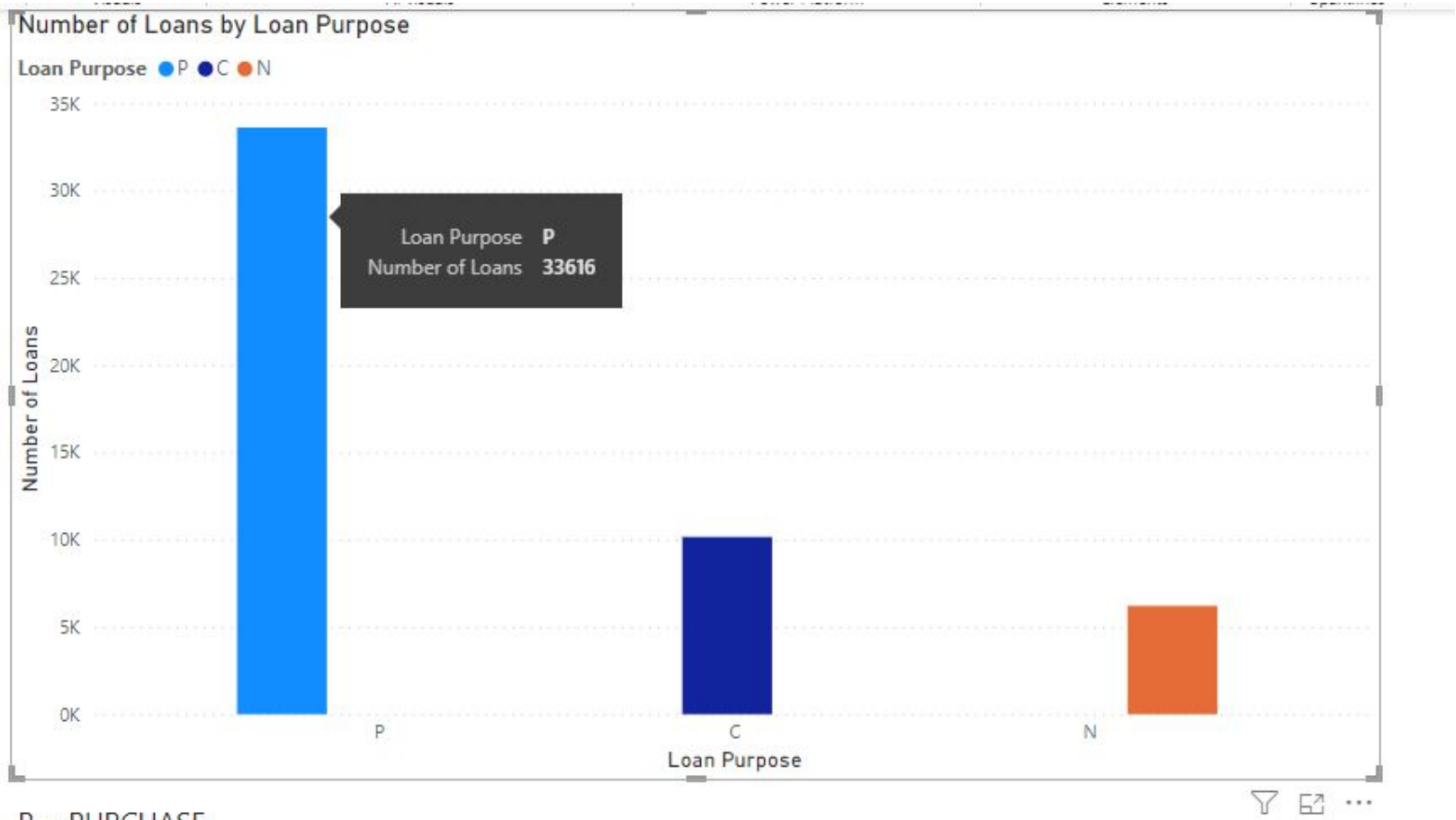
Thank you for your time!

ANALYSIS AND REPORTING



Our dataset is a random sample of 50000 loans acquired by Freddie Mac in the year 2018. The visual shows distribution of loans by state in the US. Maximum number of loans in this dataset, are sanctioned for properties in California state and the least in Puerto Rico

ANALYSIS AND REPORTING

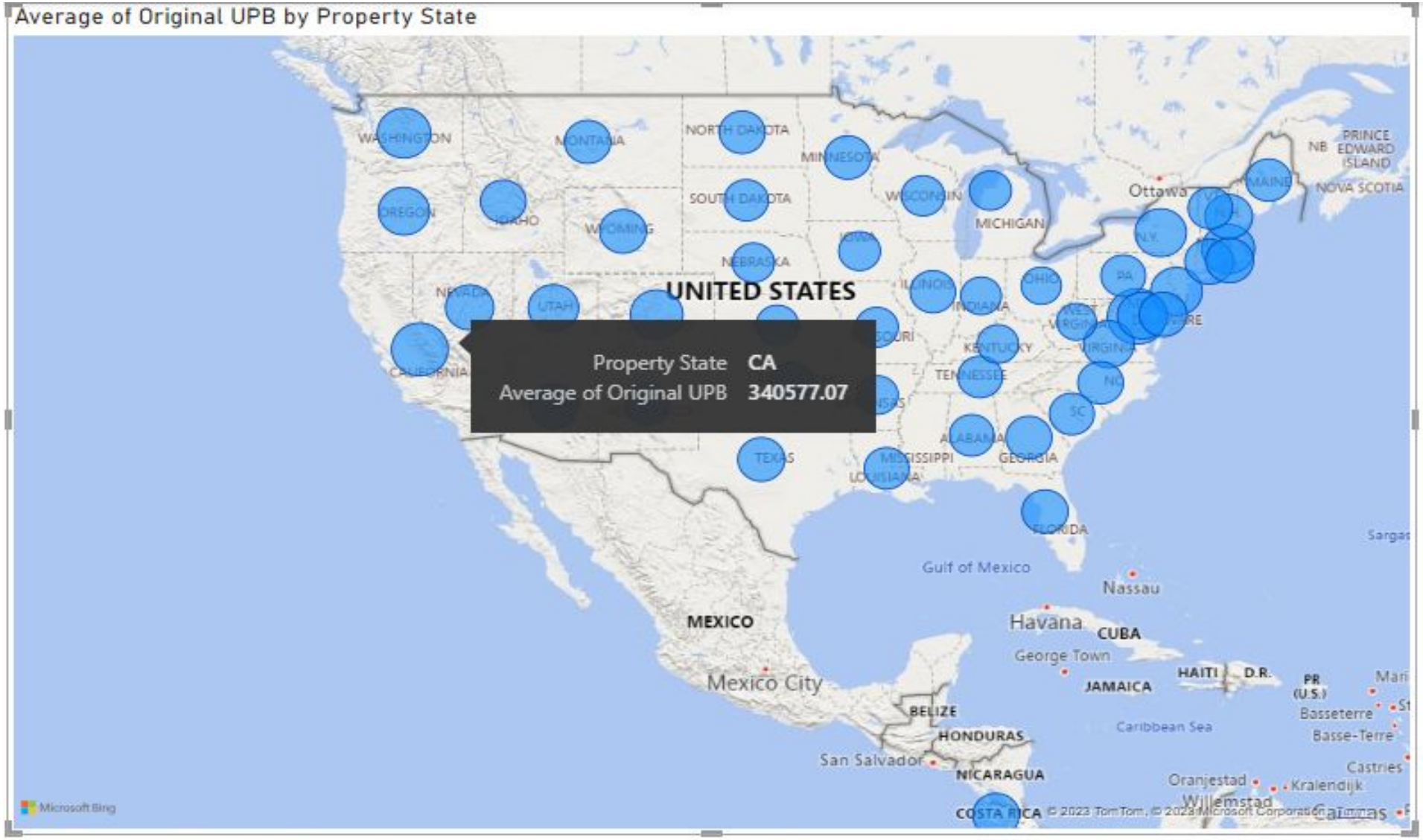


P = PURCHASE

C = CASH-OUT REFINANCE (Home owner obtains a new mortgage for higher amount, with which he closes existing mortgage and is left with a cash component)

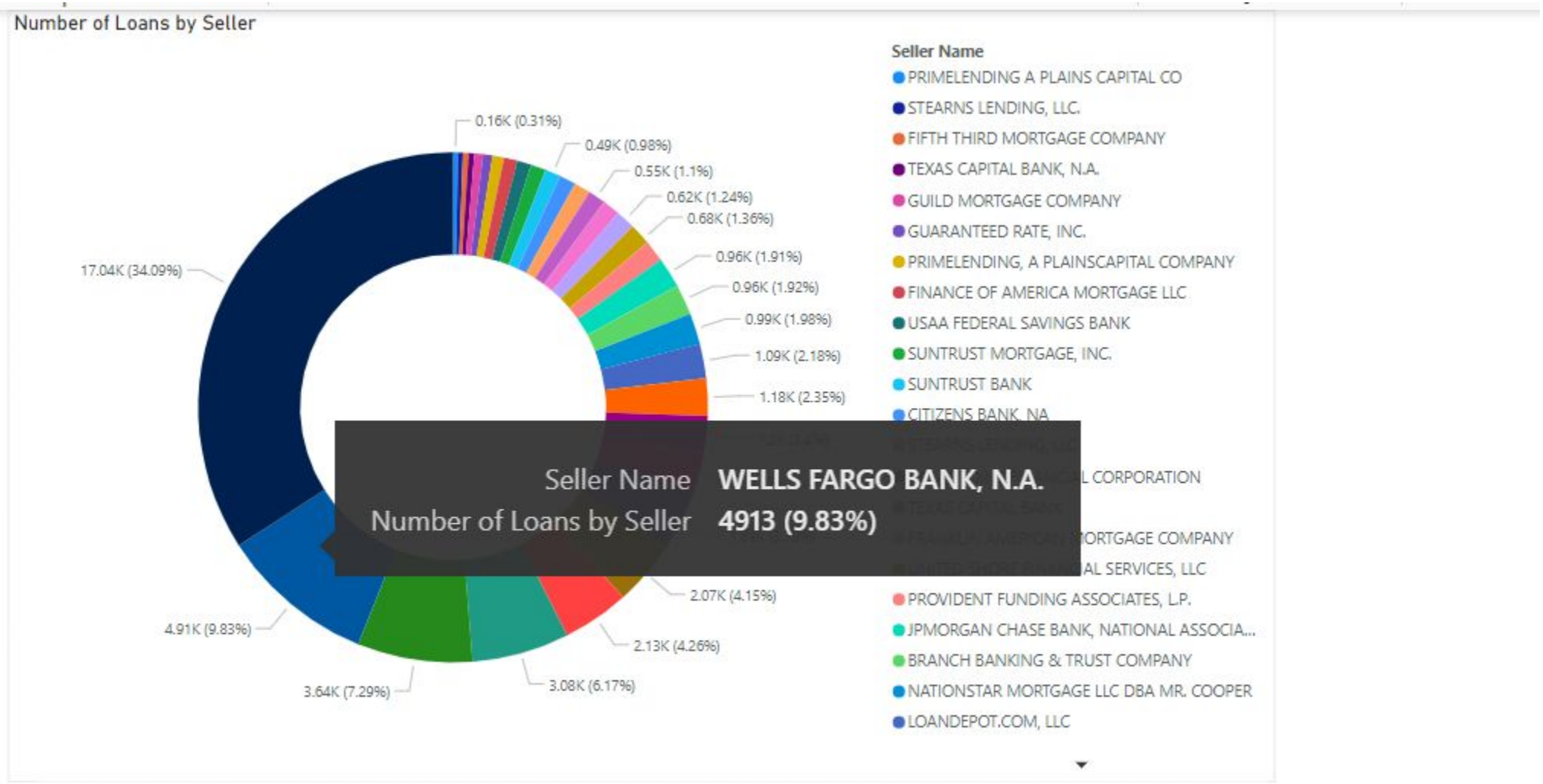
C = NO CASH-OUT REFINANCE (The new loan covers the existing mortgage outstanding only. No extra cash in the hands of the owner)

ANALYSIS AND REPORTING



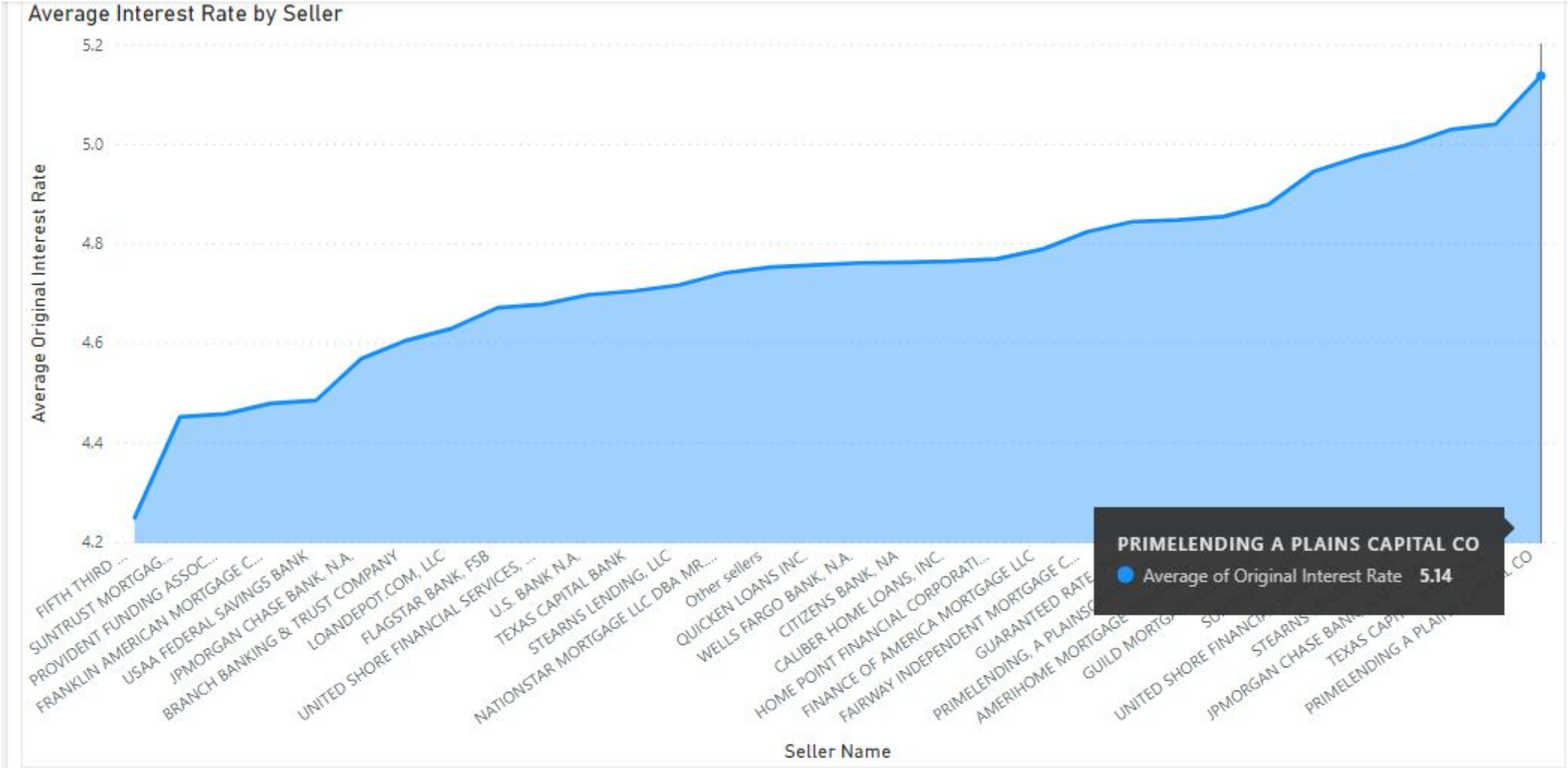
The bubble size indicates the average loan quantum in the Dataset.

ANALYSIS AND REPORTING



Sorted in ascending order, this visual shows the number of loans in the dataset, sanctioned by each seller. Besides the 'Other' category where no seller name appears, Wells Fargo has sanctioned the most loans in this dataset.

ANALYSIS AND REPORTING

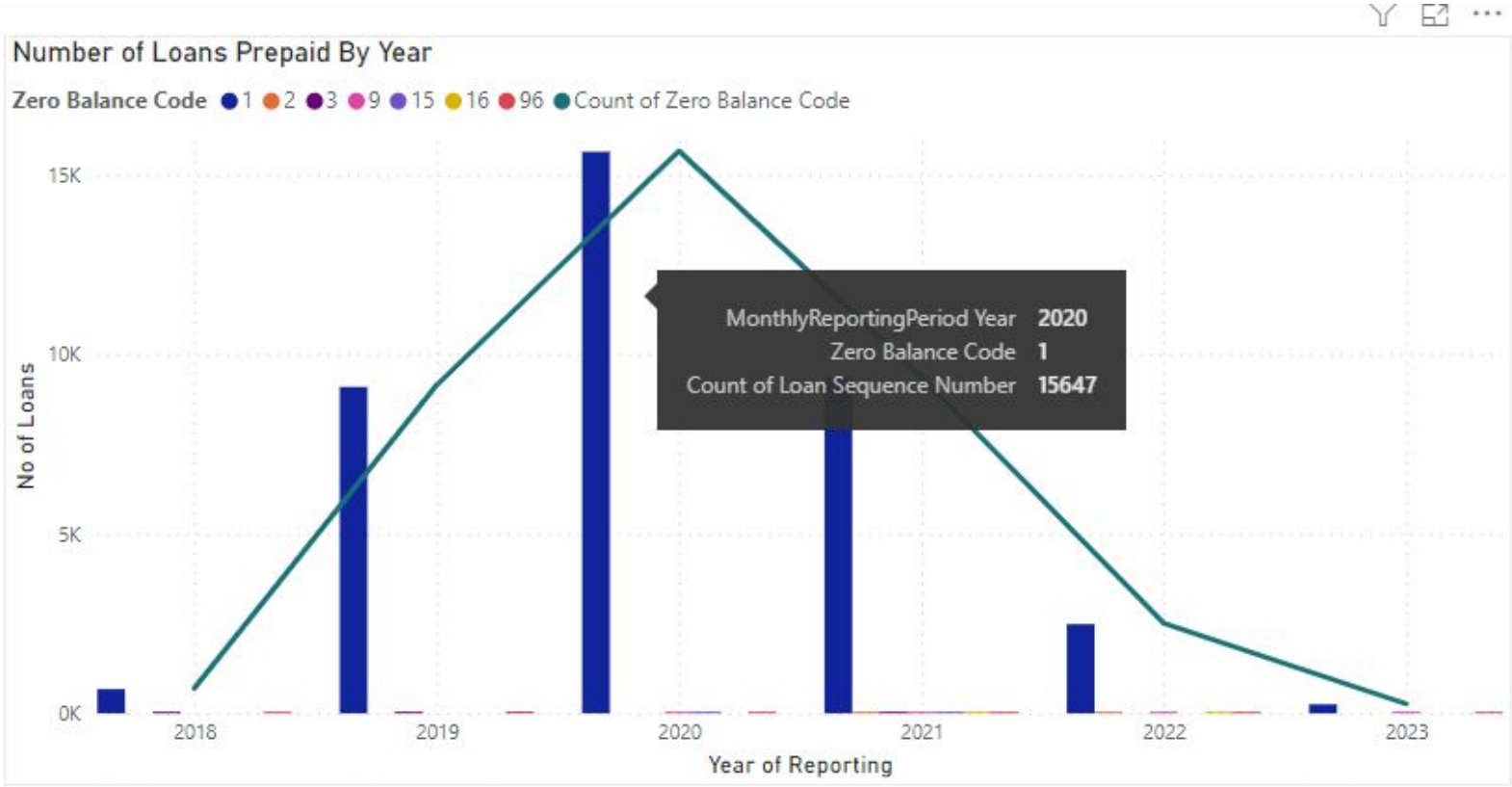


Sorted in ascending order, the visual shows the average interest rates applied on loans by different sellers. The lowest interest rates according to this dataset, are applied by Fifth third mortgage company and the highest average loan interest rates are applied by Prime lending A plains Capital Co.

ANALYSIS AND REPORTING

Deferred UPB indicates the non-interest bearing asset total for a lending institution. In this visual, we have analyzed the 'Deferred UPB' amounts by sellers, and how much of the deferred UPB is on account of delinquencies due to disasters.

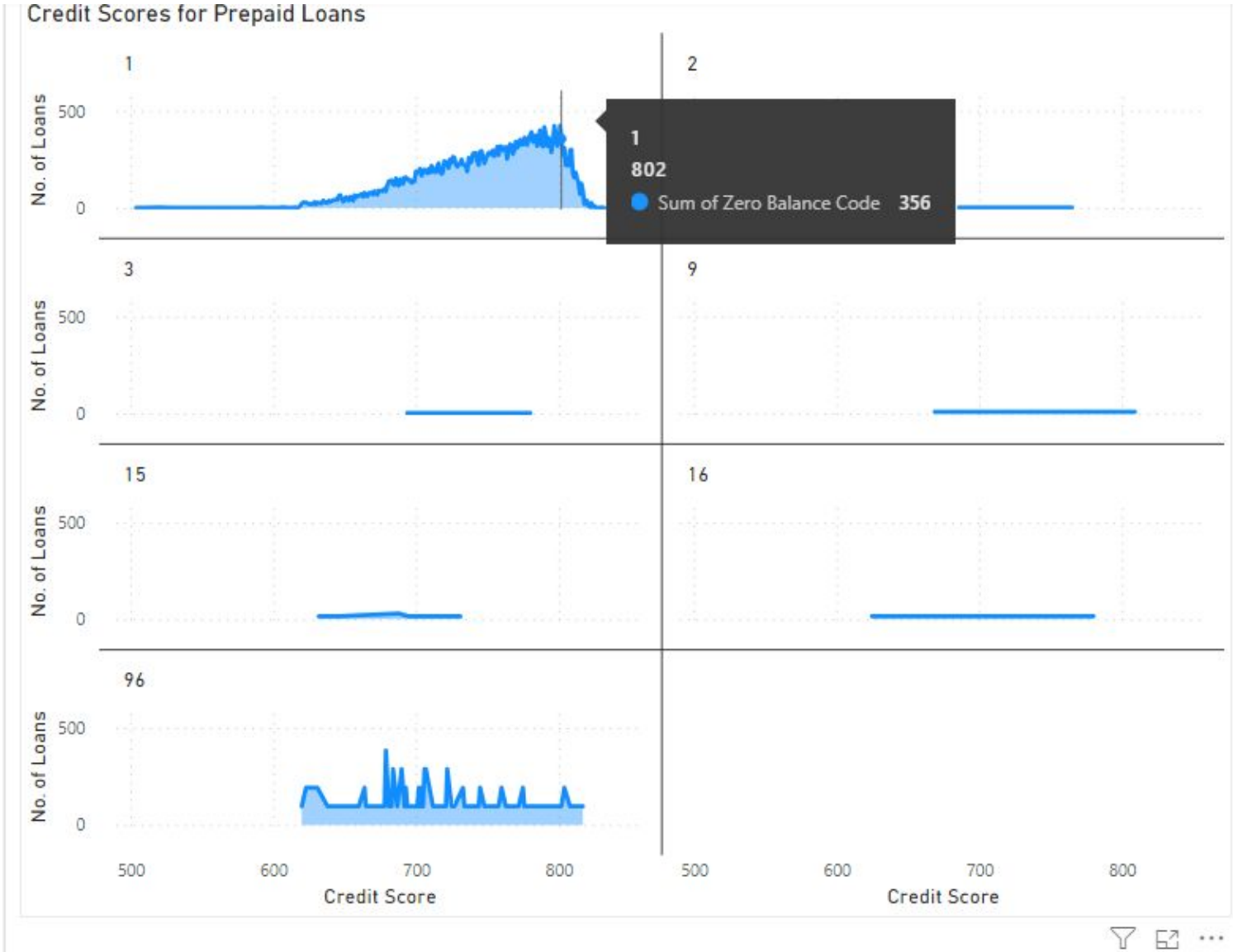
ANALYSIS AND REPORTING



In this visual, we analyzed the number of loans each year, that had closed due to prepayment or maturity.

	Modified
ZERO BALANCE CODE - A code indicating the reason the loan's balance was reduced to zero.	• 01 = Prepaid or Matured (Voluntary Payoff) • 02 = Third Party Sale • 03 = Short Sale or Charge Off • 96 = Repurchase prior to Property Disposition • 09 = REO Disposition • 15 = Whole Loan sales • 16 = Reperforming sales securitizaitons

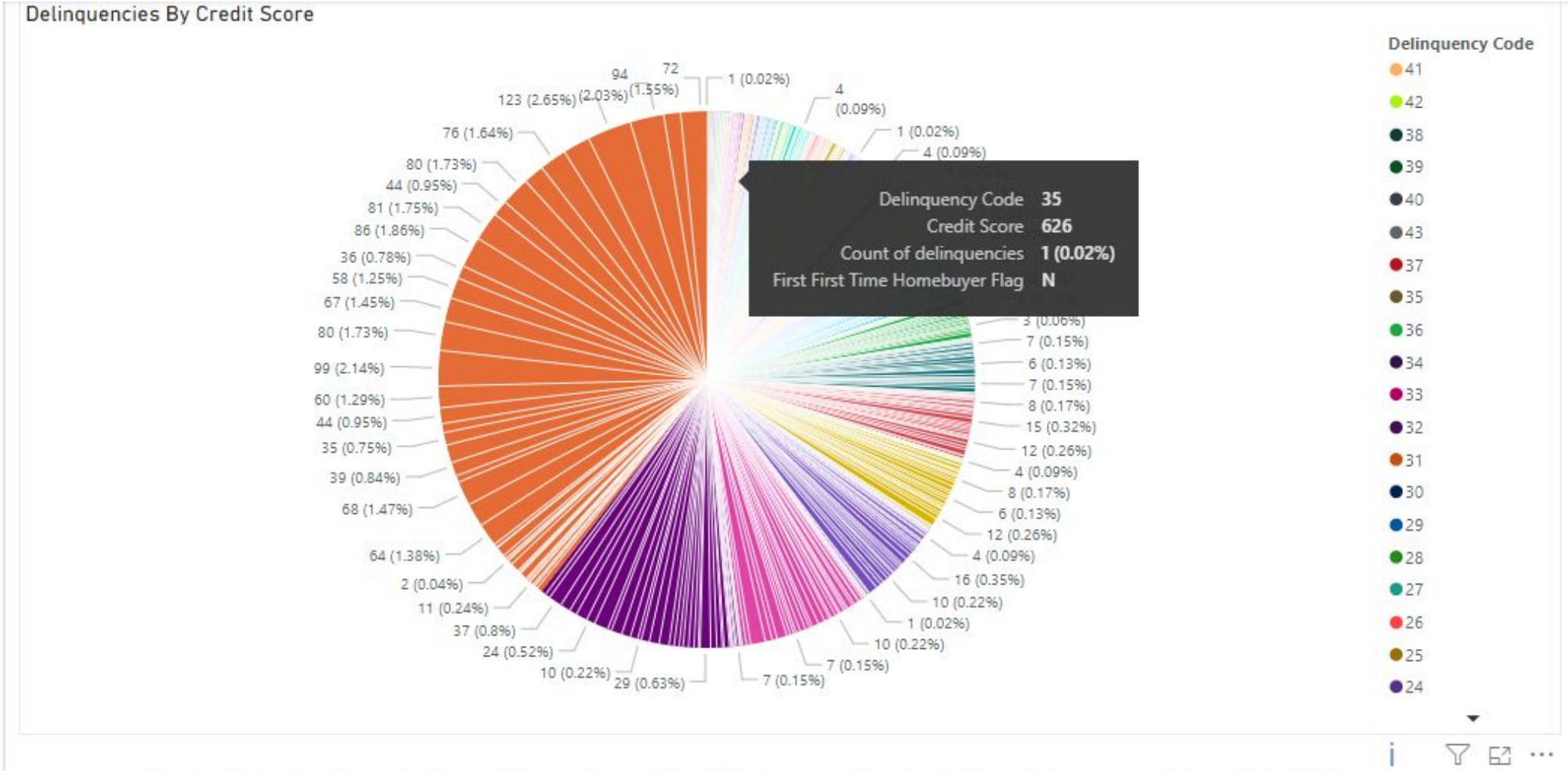
ANALYSIS AND REPORTING



	Modified
ZERO BALANCE CODE - A code indicating the reason the loan's balance was reduced to zero.	• 01 = Prepaid or Matured (Voluntary Payoff) • 02 = Third Party Sale • 03 = Short Sale or Charge Off • 96 = Repurchase prior to Property Disposition • 09 = REO Disposition • 15 = Whole Loan sales • 16 = Reperforming sales securitizations

In this visual segregated for each Zero balance code indicating the reason for closure of the loan, we can see that the borrowers with high credit score at the time of Loan Origination tend to prepay loans.

RELATIONSHIP BETWEEN DELINQUENCIES AND ORIGINATION CREDIT SCORE



This visual only shows loans that have delinquencies pertaining to late payments beginning from delinquency code 1 , which is 30-59 days delinquent and a highest code of 42 which implies 1260-1289 days delinquency. The credit score of these loans is captured at the time of Origination. All the delinquent borrowers are not first time home buyers according to our analysis.