42050 SAS PREDICTIVE BUSINESS ANALYTICS ASSESSMENT 1

Alexander Easey

UTS: 24506306 1/09/2025

Table of Contents

Contents

Project Description

Business Problem

The primary business problem addressed in this project is the prediction of mortgage loan defaults. Financial institutions face significant risk when borrowers fail to meet their mortgage obligations, leading to financial losses, reduced profitability, and potential exposure in the broader lending market. Accurately identifying loans that are likely to default allows lenders to implement proactive risk management strategies, such as adjusting underwriting standards, offering tailored repayment plans, or prioritising monitoring of high-risk accounts.

Project Objectives

This project aims to develop a classification model that can predict whether a given mortgage loan will default based on borrower characteristics, loan attributes, property information, and relevant macroeconomic indicators. By leveraging historical data, the solution seeks to improve decision-making in mortgage lending, reduce financial exposure, and enhance portfolio performance.

Data Exploration

Data Collection

The dataset used in this project was obtained from the SAS VIVA data repository, which provides comprehensive mortgage loan data for analysis and research purposes. The dataset includes both borrower-level information and macroeconomic indicators relevant to U.S. mortgage lending. Features encompass loan characteristics (e.g., original loan amount, interest rate, loan term), borrower financial information (e.g., debt-to-income ratio, credit scores), property attributes (e.g., property type, occupancy status), and broader economic variables (e.g., inflation, unemployment rate, treasury yields).

The data was provided in a structured format, suitable for exploratory data analysis and model development. Access to the SAS VIVA dataset ensures that the information is standardised, reliable, and representative of typical mortgage portfolios, allowing for meaningful insights into factors influencing mortgage default risk.

Attribute Exploration

Attribute	Data Type	Description	Example
BBBcorporateyield	Ratio	Yield on BBB-rated (risky) corporate bonds	8.2
CommercialRealEstateYoY	Ratio	Year-over-year change in commercial real estate values.	13.70558376
CPIinflationrate	Ratio	Inflation rate (consumer prices).	-0.7
CSCORE_B	Ratio	Primary borrower's credit score	543

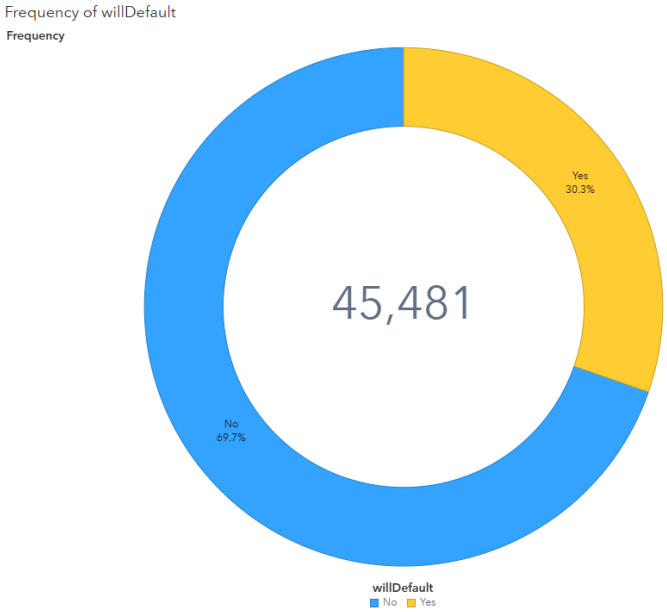
CSCORE_C	Ratio	Co-borrower's credit score (if any).	801
DowJonesYoY	Ratio	Year-over-year return of the Dow Jones Index.	5.162125243
DTI	Ratio	Borrower's monthly debt payments vs income.	20
FTHB_FLG	Nominal (Y/N)	Yes/No if borrower is a first-time buyer	Yes/No
HousePriceYoY	Ratio	Year-over-year return of the House Price Index.	5.521472393
LAST_UPB	Ratio	Remaining balance on the loan.	73612.67
LastVersusOriginal	Ratio	Ratio of current balance vs original loan.	0.982385401
LOAN_ID	Nominal (ID)	Unique Loan ID	ID507951369968
LoanAGE	Interval	Number of months the loan has been active	36
MarketVolitilityIndex	Ratio	Measures stock market volatility	26.7
MI_PCT	Ratio	How much of the loan is covered by mortgage insurance.	30
MI_TYPE	Categorical	Type of mortgage insurance. Indicates risk-sharing structure.	2
Mortagerate	Ratio	Current average mortgage rate in the market (different from ORIG_RT).	5.1
Nominaldisposableincomegrowth	Ratio	Same as Realdisposableincomegrowth, but not adjusted for inflation	6.3
NominalGDPgrowth	Ratio	GDP growth without adjusting for inflation	5.1
NUM_BO	Interval	Number of borrowers for the loan	2
NUM_UNIT	Interval	Number of housing units.	1
OCC_STAT	Categorical	Whether the property is owner-occupied, a second home, or an investment property.	P = Primary Home S = Secondary Home I = Investment
OCLTV	Ratio	Combined Loan to property value	72
OLTV	Ratio	Main loan to property value.	97
ORIG_AMT	Ratio	Size of the loan	140000
ORIG_CHN	Categorical	How the loan was issued.	R = Retail bank C = Correspondent lender

			B = Mortgage Broker
ORIG_RT	Ratio	Mortgage rate when the loan began.	7.125
ORIG_TRM	Interval	Length of loan (months).	360
Primerate	Ratio	Prime lending rate (benchmark for many loans).	3.3
ProductType	Categorical		FRM = Fixed Rate Mortgage
PROP_TYP	Categorical	Type of property	CO = Condo CP = Co-Op MH = Manufactured Housing PU = Planned Urban Development SF = Single Family
PURPOSE	Categorical	Loan purpose	P = Purchase C = Cash out refinance R = regular refinance U = Unknown
Realdisposableincomegrowth	Ratio	Growth in after-tax, inflation-adjusted household income.	5.9
RealGDPgrowth	Ratio	Growth of the economy adjusted for inflation.	3.8
RELOCATION_FLG	Nominal (Y/N)	The loan is financed as a relocation loan.	Yes/No
SellerName	Nominal	The institution selling the loan.	BANK OF AMERICA, N.A.
STATE	Categorical	Location of property. USA states are abbreviated.	CA
Unemploymentrate	Ratio	Current Unemployment rate when the loan began.	6.1
willDefault	Nominal(Y/N)	Whether the loan defaulted or not	Yes/No
X10yearTreasuryyield	Ratio	Yields on government bonds over the last 10 years.	4.4
X3monthTreasuryrate	Ratio	The rate on government bonds over the last 3 years.	0.3
X5yearTreasuryyield	Ratio	Yields on government bonds over the last 5 years.	2.3

Attribute Visualisations

Target Feature

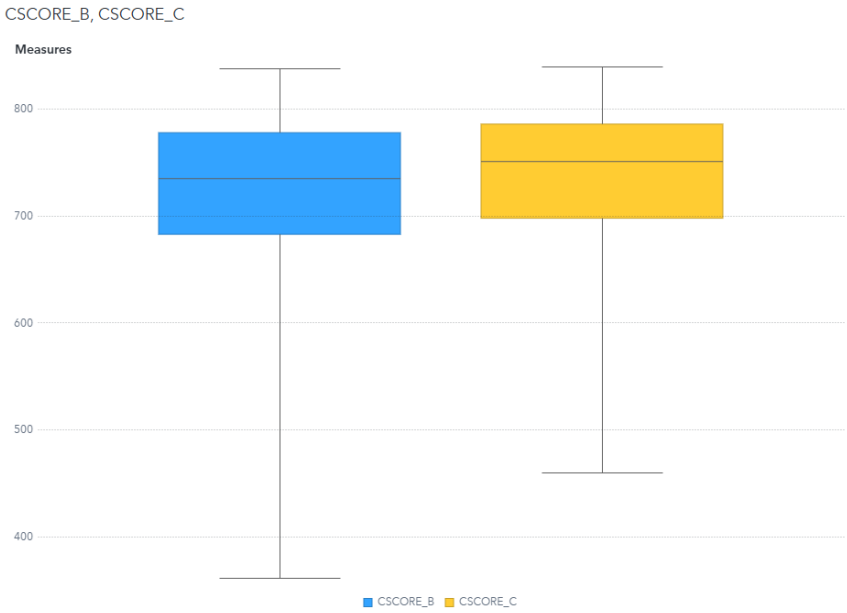
willDefault



The target feature for this problem is the 'willDefault' variable. This only has two options, Yes or No and tells us whether the mortgage loan has defaulted. This variable will be used in the training and testing of the model to verify the outputs.

Borrower Features

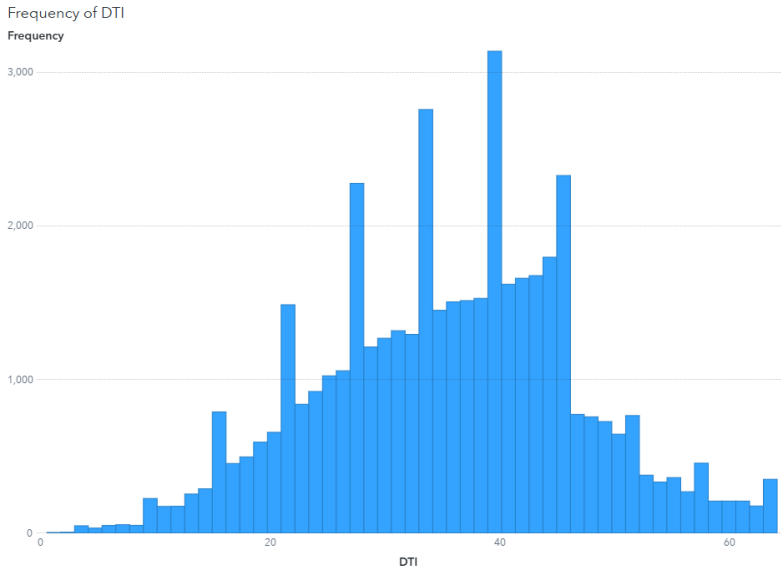
Credit Scores B AND C (BoxPlot)



In this dataset, there are two credit score variables. CSCORE_B, which contains the credit score of the primary borrower and CSCORE_C, which is an optional secondary borrower. There were 56.6% of the CSCORE_C values missing in this dataset, likely due to the fact that loans are mostly applied under one name. Therefore, CSCORE_B has

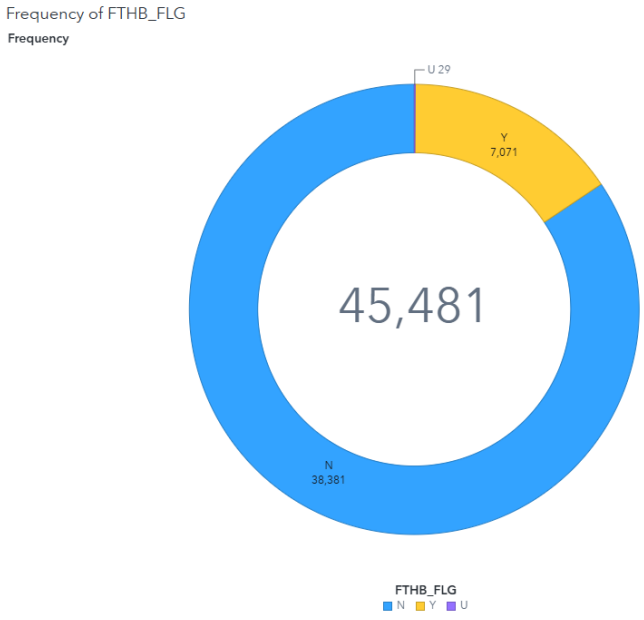
more information, possibly skewing some results in its favour. Data imputation will be used on CSCORE_C to fill in missing values, but through model analysis, it is shown that it will cause too much noise will be rejected.

DTI (Histogram)



Debt-to-Income ratio is a good indicator of other financial stresses in a person's life. The higher the DTI ratio, the more likely a default, likely why the values start dropping just before 50% as seen above, as banks don't want to give loans to people with a high liability.

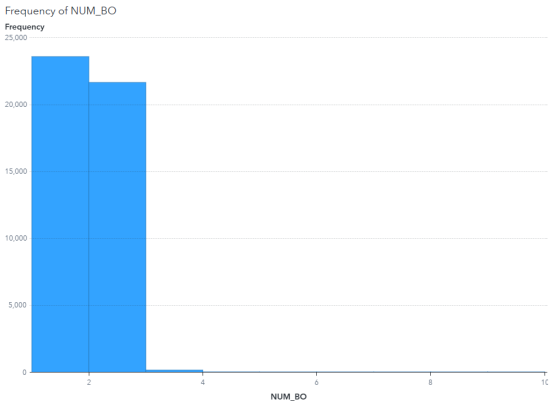
First Home Buyers (Pie Chart)



First-time home buyers are flagged as being less financially stable. In this dataset, 7071 individuals of the 45481 are first-time home buyers. This is a large enough sample; however, it's only about 20% of the total amount, meaning that results could be skewed due to a lack of representation of the data. This could be fixed through data augmentation or sampling changes for training data.

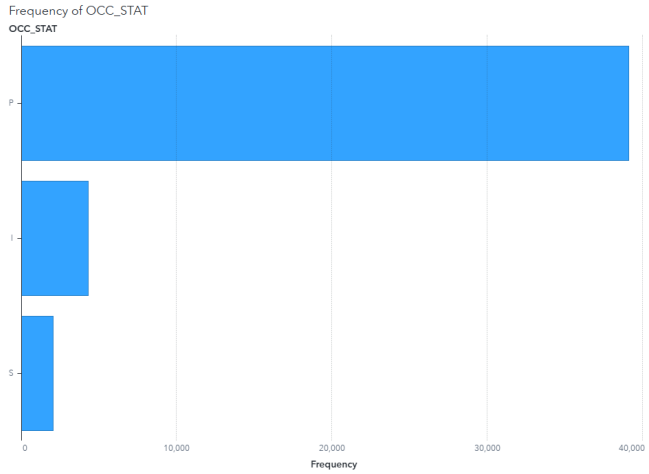
Number of Borrowers (Histogram)

The greater number of borrowers on a loan can lead to more income sources, resulting



in a lower chance of defaulting. In this dataset, most loans only have one borrower, but very close to it are two borrowers. This large number of two-borrower loans could be due to a large amount of loans in this dataset being single-family homes; therefore, two parents would share the mortgage.

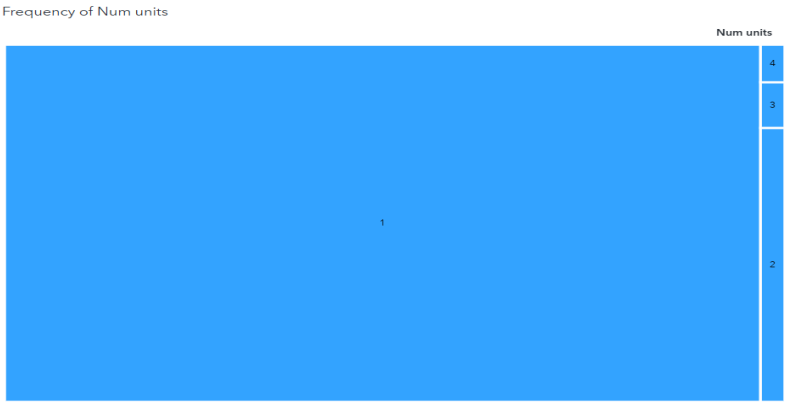
Occupancy Status (Bar Chart)



The majority of mortgage loans in this dataset are for primary homeowners. They make up nearly 40000 of the total data. The data collected does not contain any dates on the loans, so it's unclear what the economic situation was during this time, but on average, an investment property or secondary loans would default more due to the extra financial stress that comes with them.

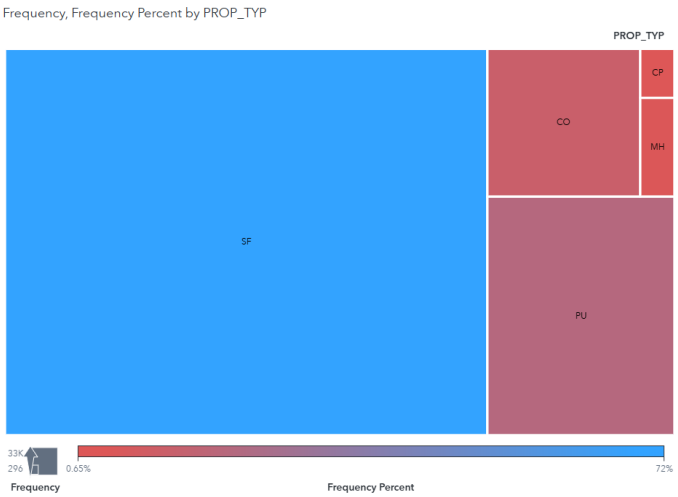
Property Features

Number of units (Tree Map)



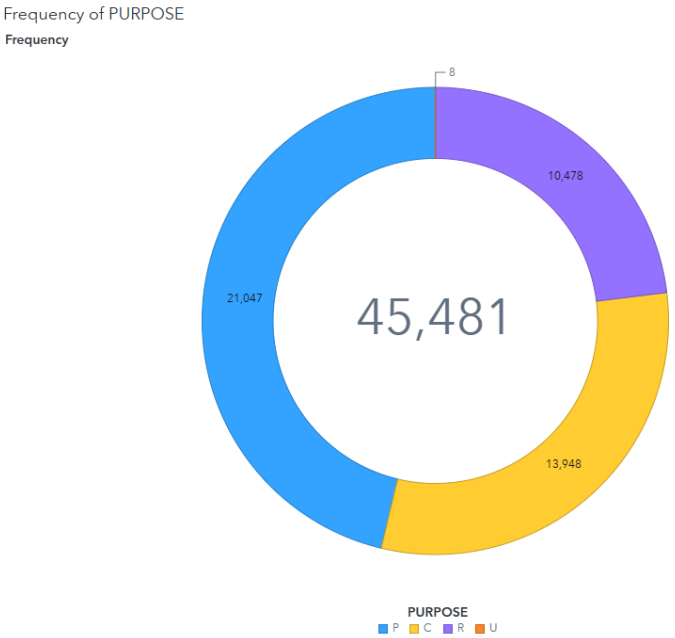
As the majority of the property types on these mortgage loans are for single-family homes, it makes sense that each of these would only contain one housing unit. Multifamily housing properties are usually larger lots and investment properties, which have a much higher risk of defaulting. This data is heavily skewed towards only one number of numbers so further analysis on the effect of this data on the model should be done.

Property Type (Tree Map)



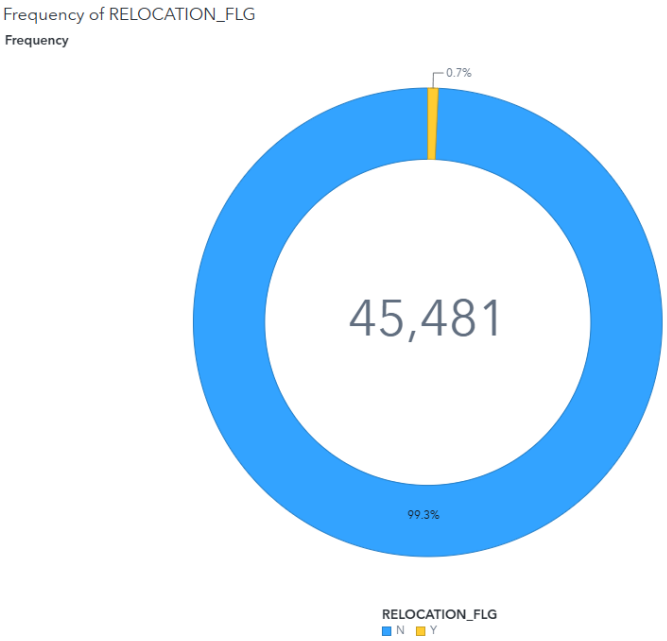
Certain property types may carry higher risks of defaulting. As seen in the figure above, most mortgages are loaned out for single-family homes, with urban developments with the second most. Different property types carry different risks; for example, condos and multi-units may rely more on external conditions like associations or rental markets.

Purpose (Pie Chart)



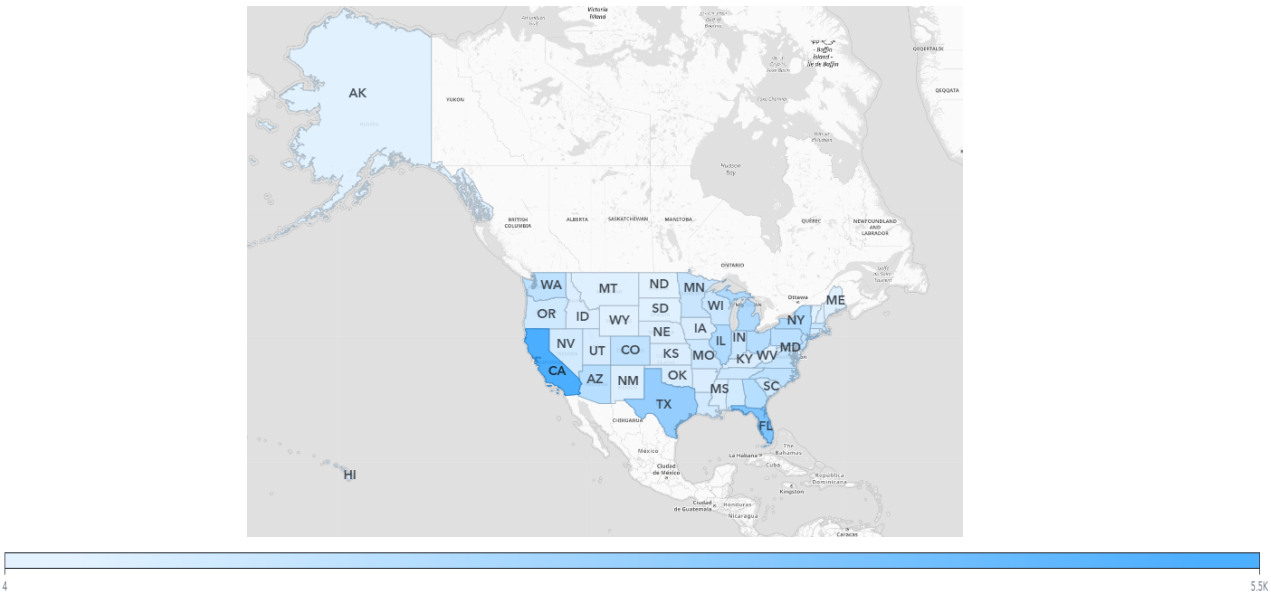
This shows the reason for the loan. Purchases are standard with 21047 of the 45481 loans, while cash-out refinances are riskier since they involve borrowing additional money, often increasing borrower leverage.

Relocation Flag (Pie Chart)



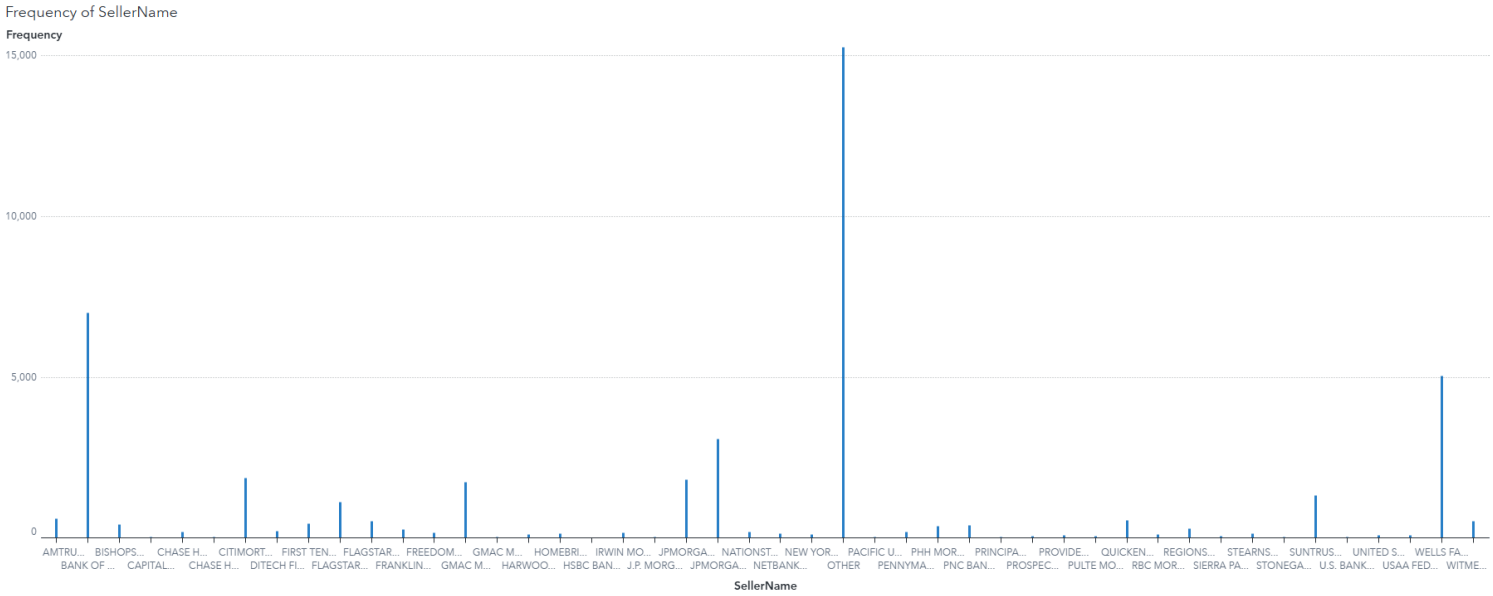
This indicates if the loan is related to relocation. Relocation loans may be riskier if the borrower is moving due to job or financial instability. Relocation loans only make up for 0.7% of the values in this dataset, but could be a strong indicator of a possible default.

State (Geo Map)



The figure above shows a geographical map of where the houses of the mortgages are based. As seen, most of the data comes from the states of California and Florida, both of which especially California are known for their brutal housing market. Again, we cannot be sure when this data was collected, but the location of these properties, depending on the markets in the areas, can lead to a higher risk of defaulting.

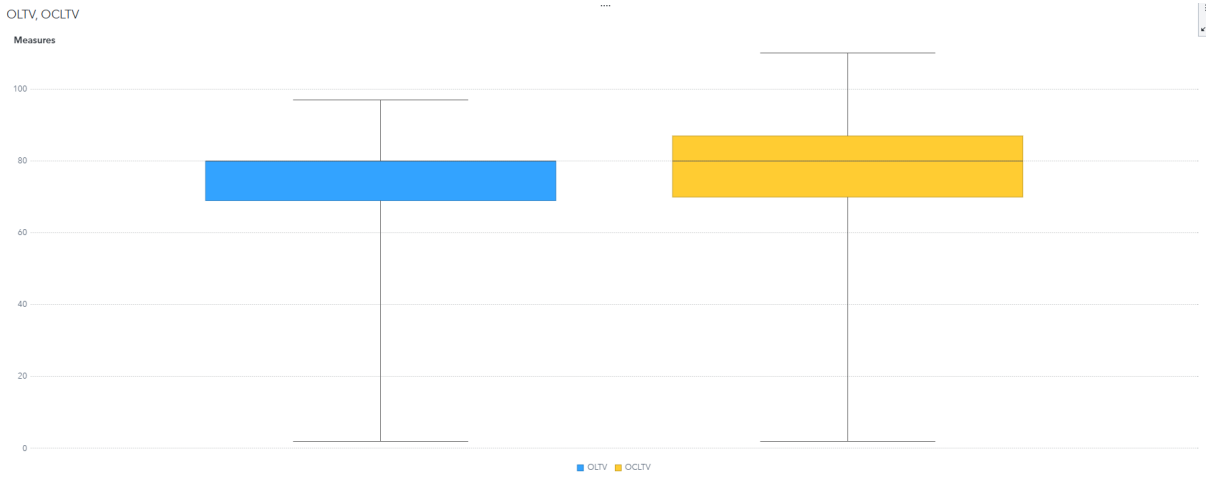
Seller Name (Histogram)



The seller names are the institutions that give out the loans. The most common one, with a frequency of 15000, is independent or other brokers, with the other smaller peaks being large banks such as Bank of America, Wells Fargo, and JPMorgan. If certain banks are connected to a higher rate of defaulting on their policies, and mortgage rates should be investigated. Maybe if they are too low, they attract people in a worse financial situation, leading to higher defaults.

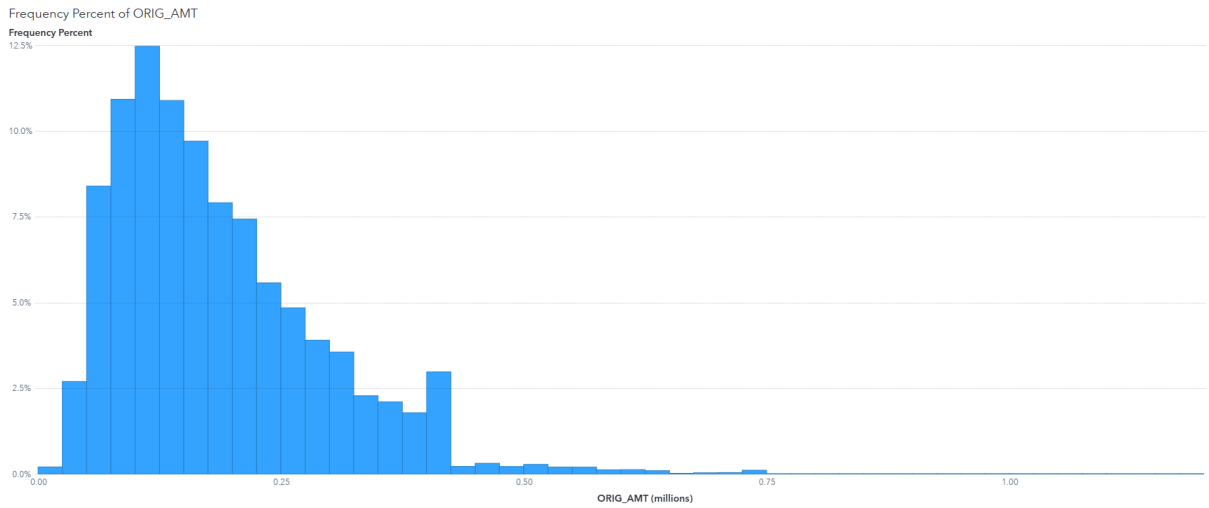
Loan Features

OLTV & OCLTV (Box Plot)



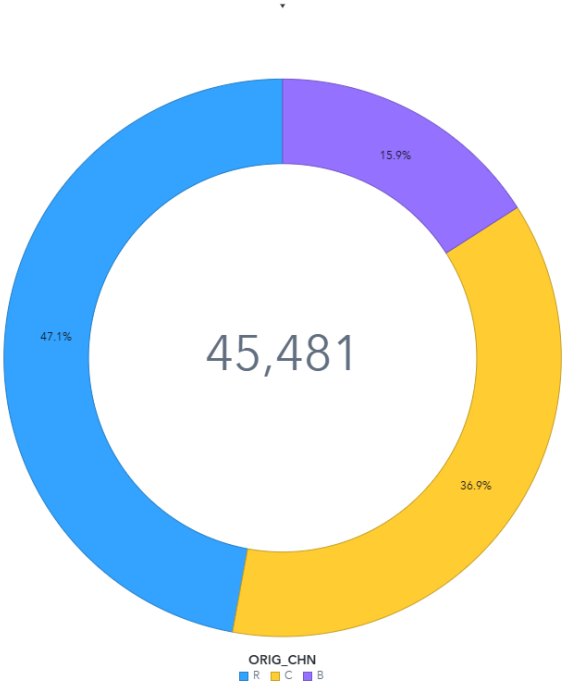
These measures how much of the home value was financed with debt. Higher LTV means less equity upfront, making it easier for borrowers to default if home values fall. The value of OCLTV is higher than OLTV as it includes the combined loans of the borrowers. This would increase the ceiling and average but keep the floor the same.

Original Amount (Histogram)



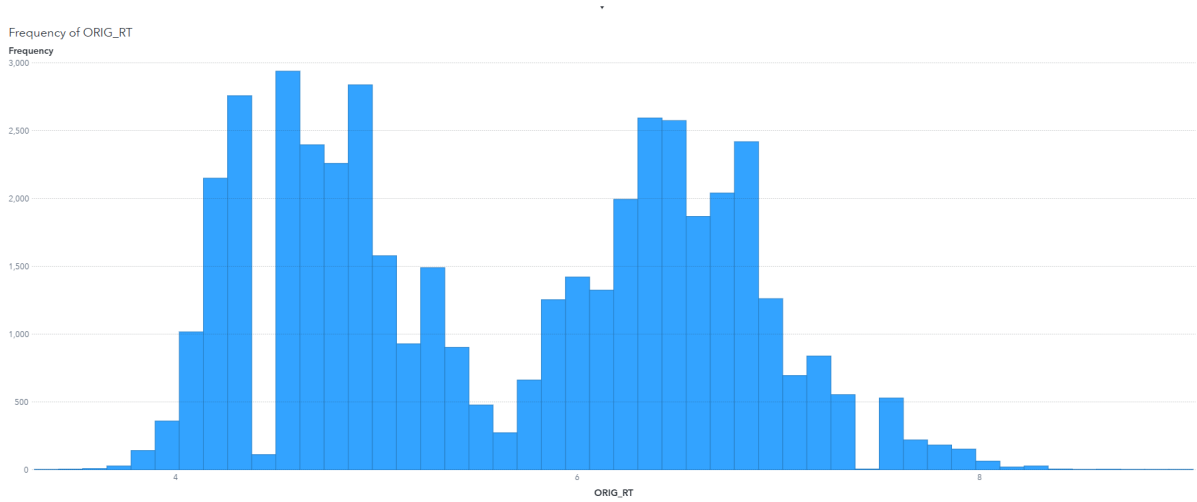
This is the size of the mortgage at origination. Larger loans may be riskier since payments are higher, but risk also depends on borrower income. From the figure above, most of the loans are around $100,000 with a steep, gradual drop off until the $500,000 mark.

Original Channel (Pie Chart)



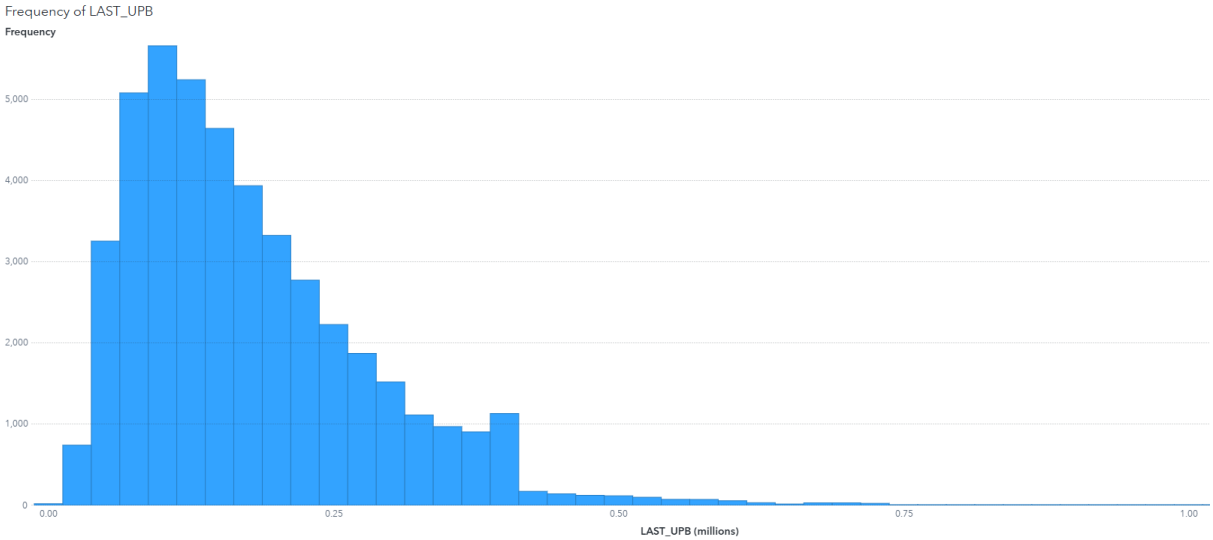
This is the channel through which the loan was originated, such as Retail (direct from bank), Correspondent (through third-party brokers), or Broker. Broker-originated loans have historically had higher risk. The most popular of these are retail loans at 47.1%, most likely from the top banks, from the seller names graph.

Original Rate (Histogram)



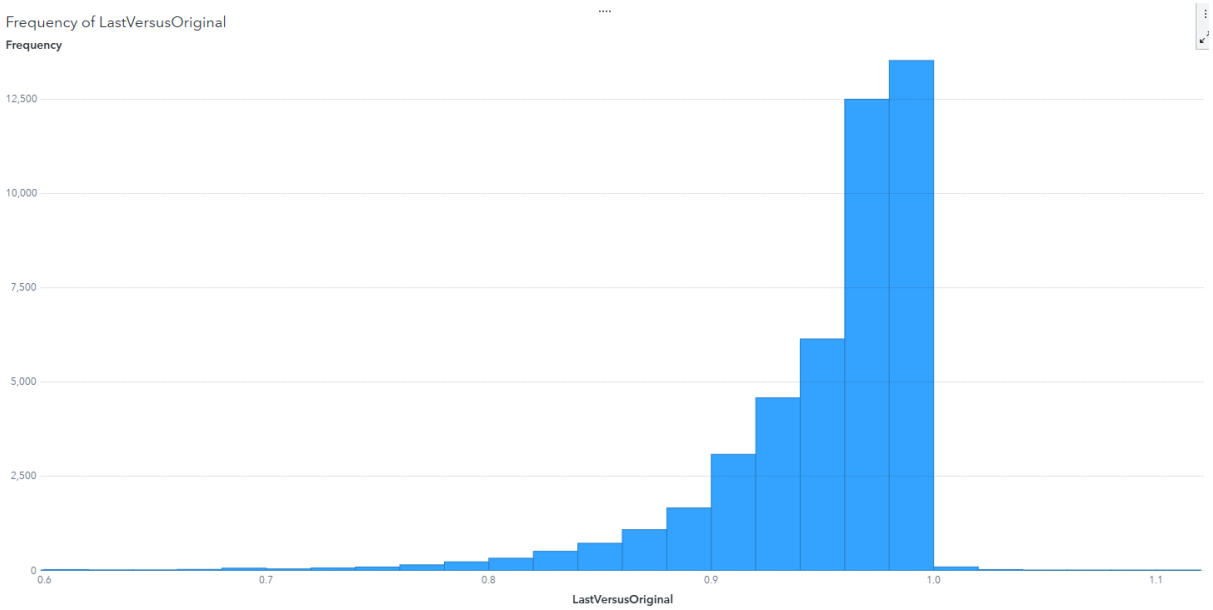
The rate assigned at origination. Borrowers with higher original rates face higher payments and, increased risk of default. This graph showed two peaks, one occurring at about 4.5% and another at around 6.5%. A higher mortgage rate is usually associated with a lower borrowing amount. These Peaks are likely occurring where the $100,000 and $500,000 marks were on the original amount graph.

Last UPB (Histogram)



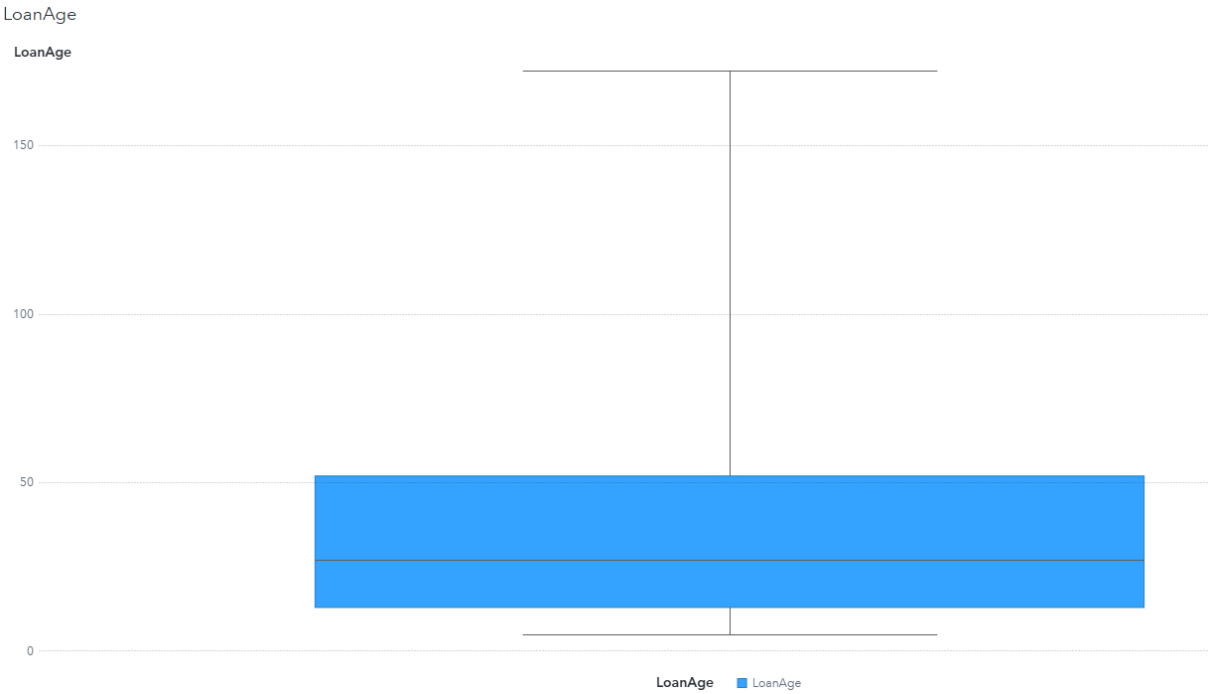
This is the remaining amount of the loan not yet paid off. A higher unpaid balance can increase risk if the borrower loses income. If the balance is small, borrowers may be more motivated and able to finish paying off the mortgage. The last UPB graph follows the same trend as the original amount, displaying that the majority are paying it off periodically every month, which would lead to no difference in the graph.

Last versus original (Histogram)



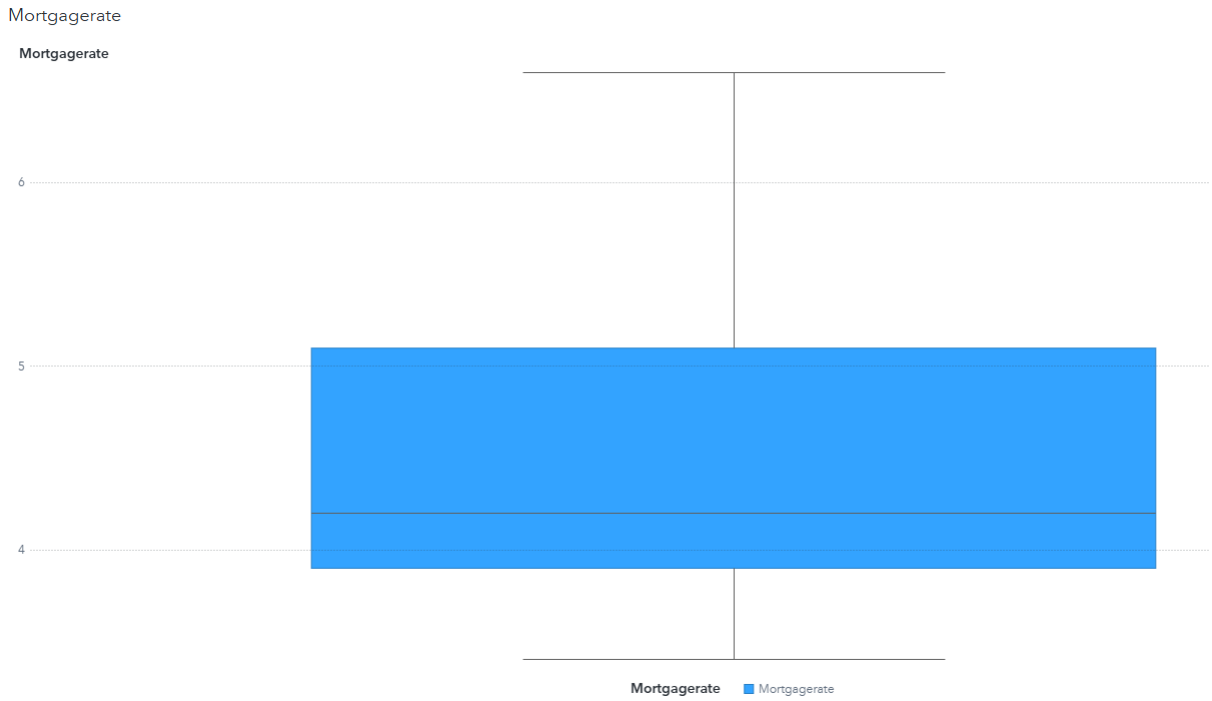
This compares the current unpaid balance to the original loan balance. It shows how much progress has been made. Borrowers who have paid off more of their loan (smaller ratio) may be less likely to default because they have more equity. As seen above, some points are above the 1.0 mark, showing that they may have missed payments that have now exceeded the original price. These would be more likely to default.

Loan Age (Box Plot)



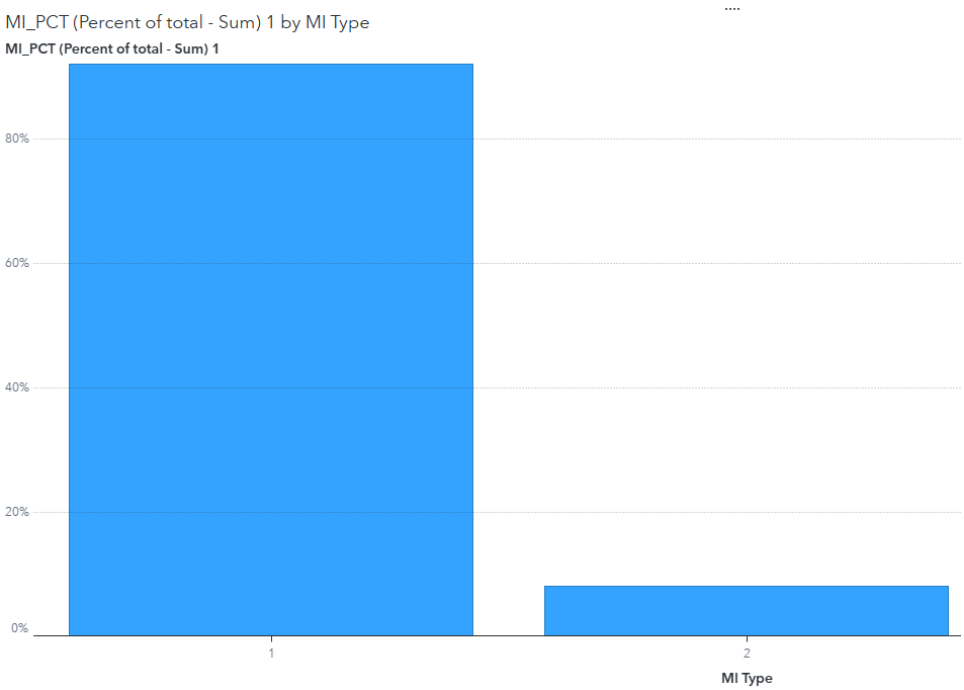
The number of months since the loan was originated. Newer loans often have higher risk because borrowers haven't built equity yet, while older loans may default less often since payments have been consistent over time. Most loans in this dataset are relatively new, as the average is at 25 months. This plot also follows the shape of the last UPB, as a newer loan would result in a higher remaining balance.

Mortgage Rate (Box Plot)



The interest rate on the mortgage. Higher rates mean higher monthly payments, which makes repayment harder. Borrowers with adjustable rates may face even higher risks if rates rise.

MI PCT & Type (Bar)



MI PCT is the percentage of the loan covered by mortgage insurance. If insurance covers more of the loan, the lender's risk is reduced, but the borrower still faces payment pressure. Often, borrowers with higher MI are riskier since they likely had smaller down payments. MI type is the type of mortgage insurance product. Different products may be linked to different borrower profiles, risk appetites, or levels of financial support. Some types may provide stronger protection than others. The majority of these variables were missing, and no descriptions on what the 1 and 2 types could mean.

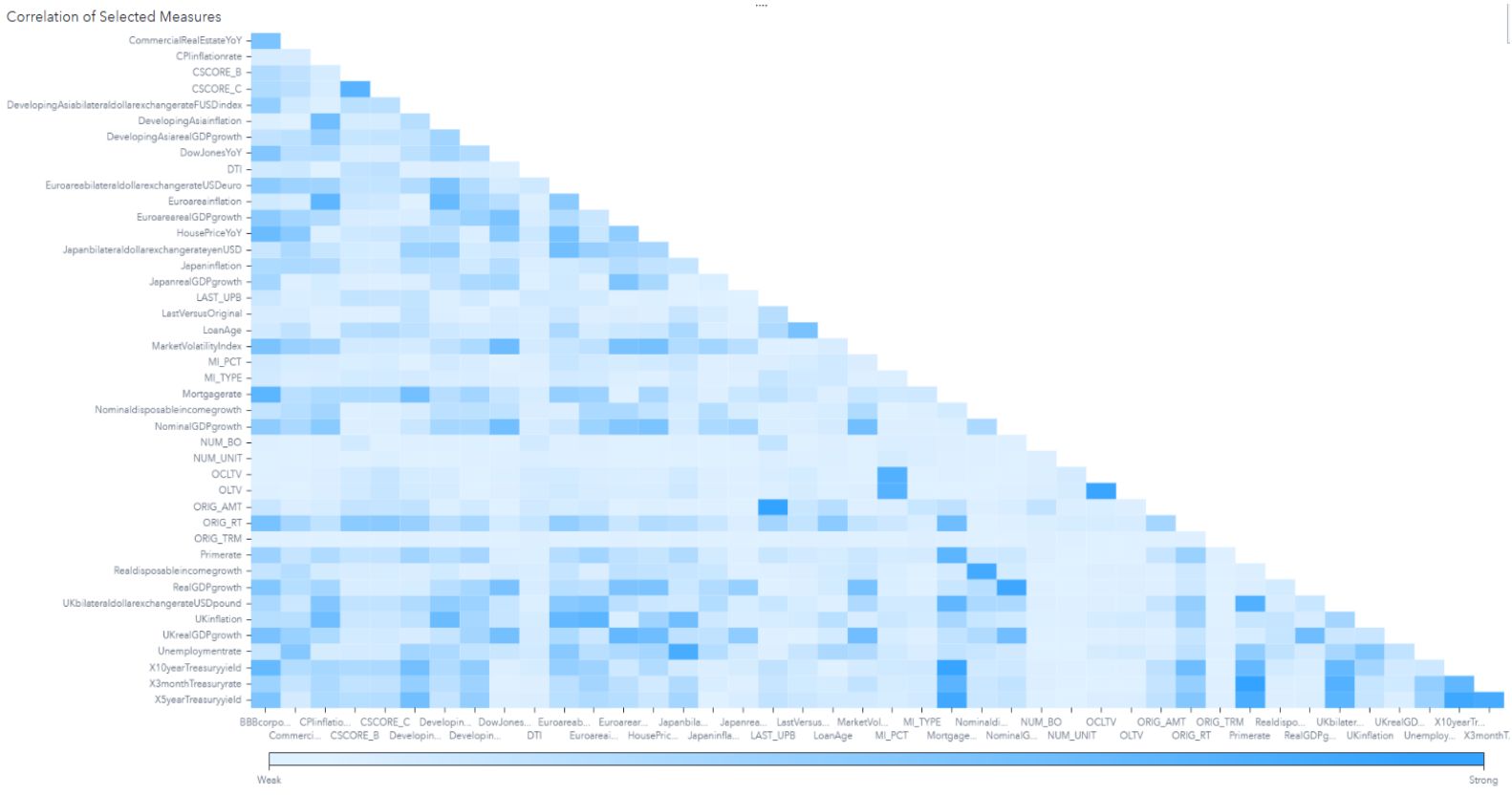
Economic and Market Features

In this dataset there were a number of macro economic and market features contained which were; Unemployment Rate, Real GDP Growth, Nominal GDP, Real disposable income, Nominal disposable income, Prime rate, 3-month treasury, 5-year treasury (Box Plot), 10-year treasury (Box Plot), BBB corporate, Market Volatility, Commercial Real Estate, CP Inflation Rate, House Price YoY and Dow Jones YoY. Most of these values had no real trends and were hard to display, as there were no dates to show a change over time or anything similar. A data table is listed below for these values to provide some insight.

Variable Name	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Relative Variability	Mean plus 2 SD	Mean minus 2 SD
BBBcorporateyield	3.9	9.4	5.2439	1.2339	1.4664	2.0337	0.2353	7.7116	2.7762
CPIinflationrate	-8.9	6.3	1.6618	2.4578	-1.5327	4.5696	1.479	6.5775	-3.2539
CommercialRealEstateYoY	-33.9056	20.5479	6.0458	11.743	-2.0379	3.5878	1.9423	29.5318	-17.4402
DowJonesYoY	-39.446	49.5773	6.6088	16.3937	-0.5714	1.2446	2.4806	39.3962	-26.1786
HousePriceYoY	-17.4419	16.3522	2.0904	7.9285	-0.7688	0.1423	3.7928	17.9474	-13.7665
MarketVolatilityIndex	12.7	80.9	28.5523	12.2133	1.8182	4.2954	0.4278	52.9789	4.1257
NominalGDPgrowth	-7.7	9.3	3.4138	2.7383	-1.4733	4.0227	0.8021	8.8903	-2.0627
Nominaldisposableincomegrowth	-14.5	13.3	3.7619	4.1806	-1.3854	5.6357	1.1113	12.1231	-4.5992
Primerate	3.3	8.3	3.9043	1.346	2.3558	4.3269	0.3447	6.5962	1.2124
RealGDPgrowth	-8.2	6.9	1.7486	2.3537	-1.7823	5.2118	1.3461	6.4559	-2.9588
Realdisposableincomegrowth	-15.7	10.9	2.275	3.8101	-2.088	8.4506	1.6747	9.8953	-5.3452
Unemploymentrate	4.4	9.9	6.768	1.7294	0.4494	-1.3327	0.2555	10.2269	3.3091
X10yearTreasuryyield	1.6	5.2	2.9701	0.9689	0.5255	-1.0757	0.3262	4.9078	1.0324
X3monthTreasuryrate	0	5	0.6284	1.2648	2.3476	4.406	2.0128	3.1579	-1.9012
X5yearTreasuryyield	0.7	5	2.021	1.0345	1.1506	0.6075	0.5118	4.0899	-0.0479

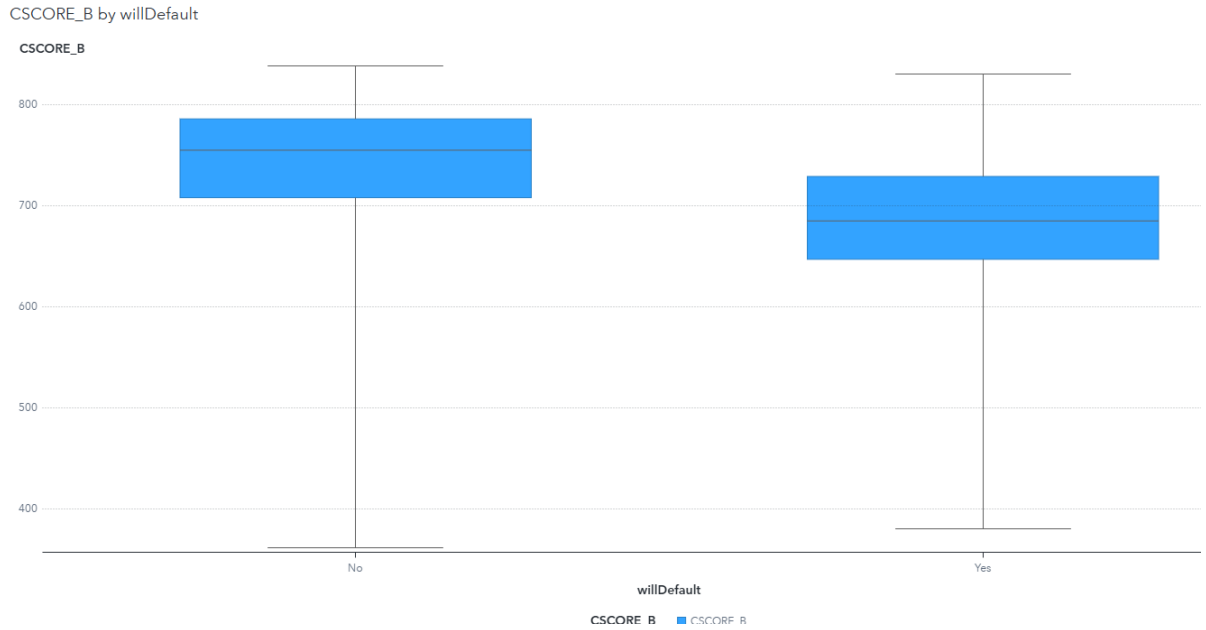
Exploratory Data Analysis and Visualisation

Correlation Matrix



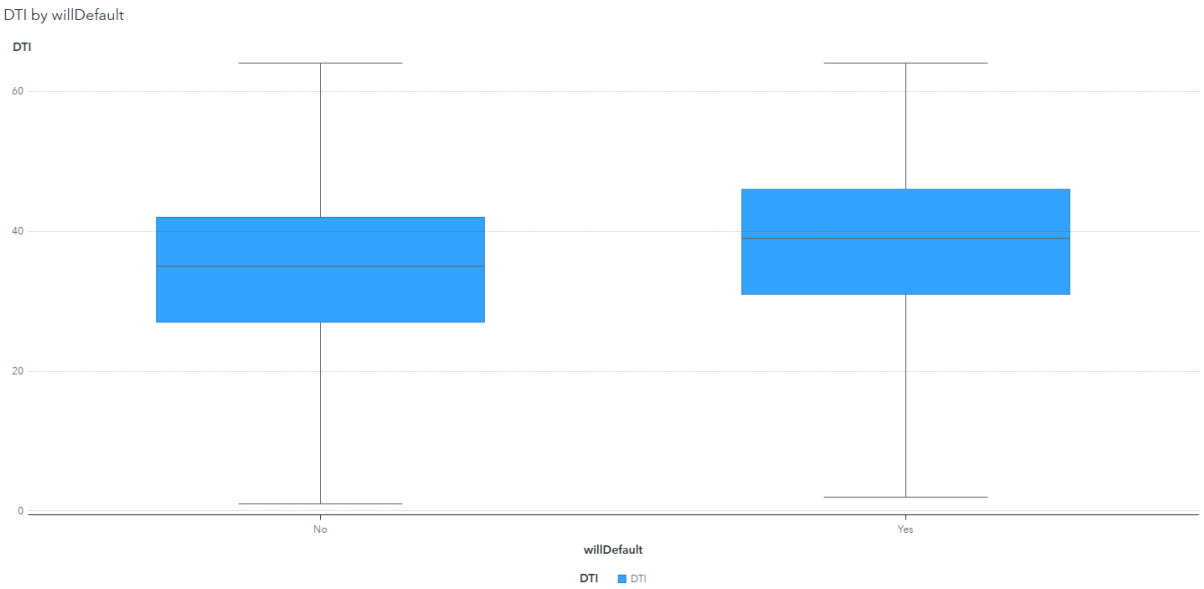
A correlation matrix of all the variables was made to provide further insight into how the values relate to each other. From this, data related to each other are shown and can be further analysed and focused on their relation to the default rate. Certain variables were also found to have little to no correlation and thus can be explored to be rejected to reduce the noise in the modelling that occurs later.

Credit Score by Will Default



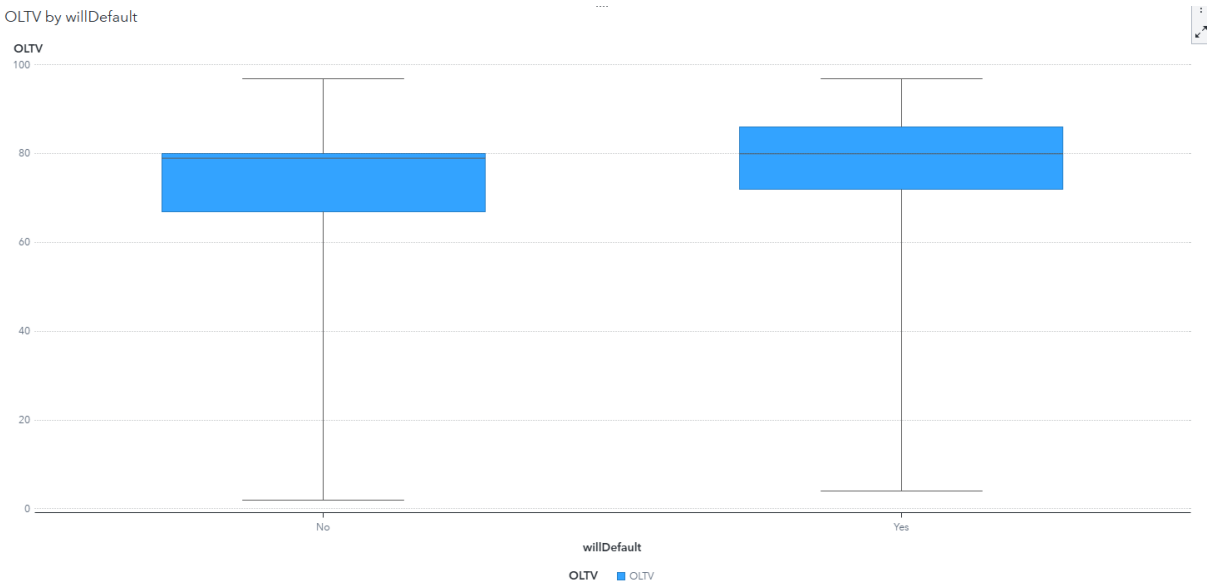
Credit scores of borrowers were grouped by whether the loans defaulted or not. As shown above, the mortgages that did default had a lower average credit score compared to the ones that didn't. This makes sense as credit scores are one of the best ways to indicate whether someone is financially stable/intelligent, as it's based on their previous loans.

Debt-to-Income by Will Default



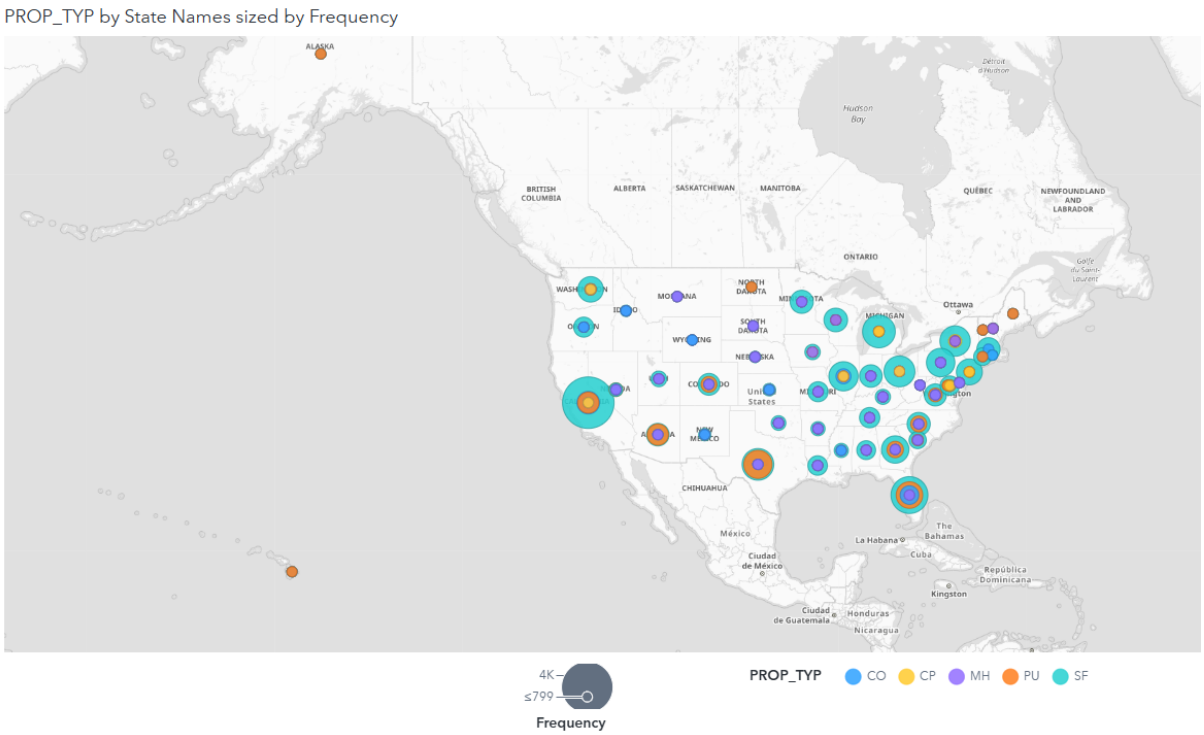
Similarly to the credit score box plot, this also shows that the debt-to-income ratio for individuals who defaulted was higher than the ones that didn't. This gives further insight into the individual's financial status and is a prime indicator of whether they are at risk of defaulting.

Loan-to-Value by Will Default



Borrowers who defaulted tended to have higher OLTV ratios on average compared to those who didn't default. This makes sense because a higher OLTV means the borrower had less equity in the property from the start. With less equity, the loan is riskier; if property values fall or the borrower faces financial stress, they have little buffer and are more likely to walk away or fall behind on payments.

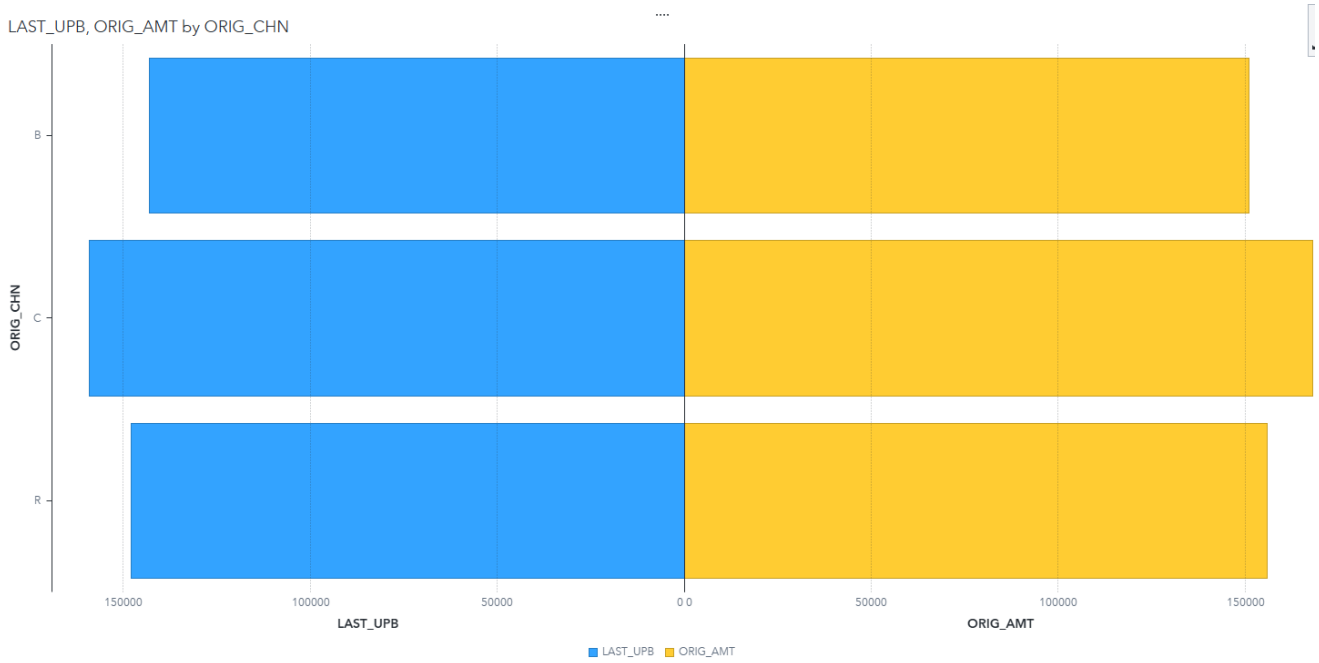
Geographical Map by Property Types



"A geographic map of property types across U.S. states shows clear regional patterns. California holds the largest share of mortgages, with the majority tied to single-family homes, consistent with its size and suburban housing profile. Texas also has a high

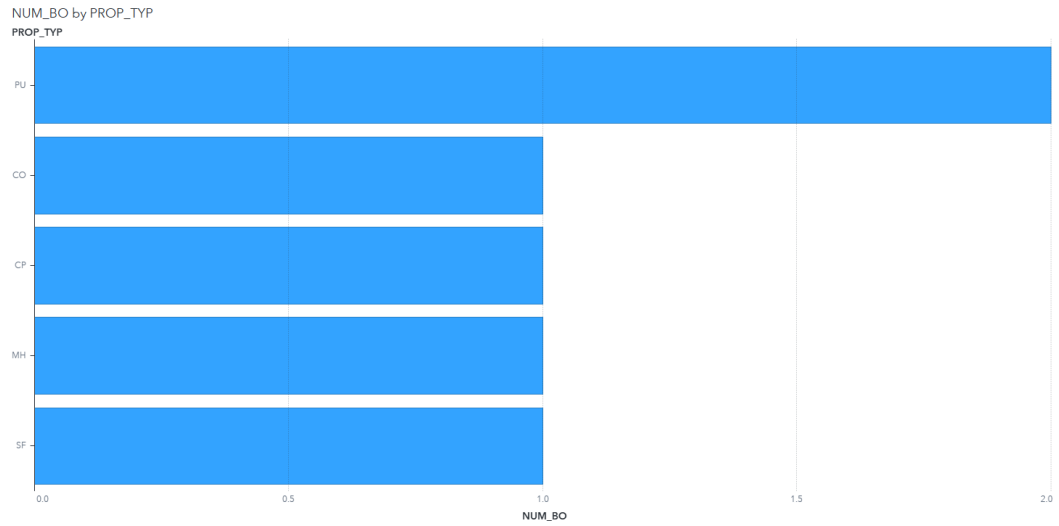
volume of mortgages, with a significant share in public urban developments, indicating more diverse housing types. In contrast, Alaska and Hawaii show a majority of mortgages concentrated in public urban developments, likely due to land availability and housing market structures in those states.

Last Balance and Original Amount By Original Channel



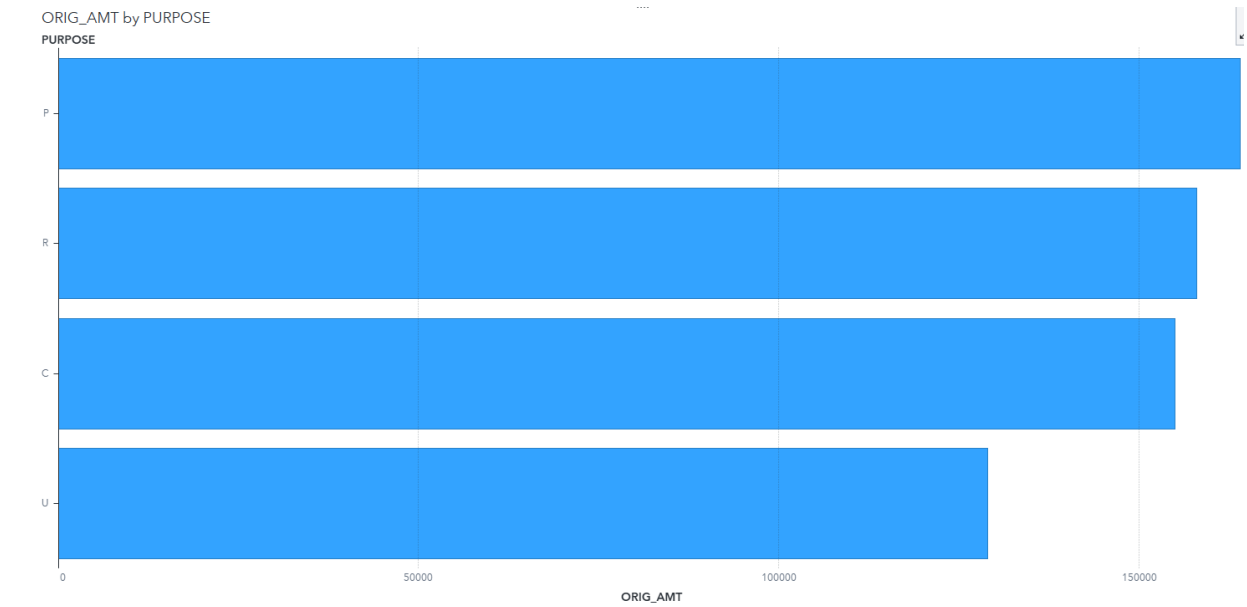
The analysis of original loan amount and last unpaid balance by origination channel shows that Correspondent (C) loans are, on average, higher in value compared to Retail (R) and Broker (B) loans. This suggests that correspondents often service larger mortgages, possibly reflecting their role in connecting borrowers to bigger lenders who can fund high-value loans. Retail loans fall in the middle range, while Broker (B) loans typically involve smaller loan sizes. The higher unpaid balances in Correspondent loans indicate greater exposure for lenders, making this channel particularly important for monitoring repayment risk.

Number of Borrowers by Property Type



The box plot of the average number of borrowers grouped by property type shows that nearly all property types, such as single-family homes, Condos, and Co-ops, typically have one borrower. However, Public Urban (Pu) developments stand out, with an average of two borrowers per loan. This suggests that loans for public urban housing are more likely to involve co-borrowers, possibly due to income requirements, affordability constraints, or shared ownership models. The presence of multiple borrowers could reduce individual risk, but it also reflects the financial accessibility challenges of this property category.

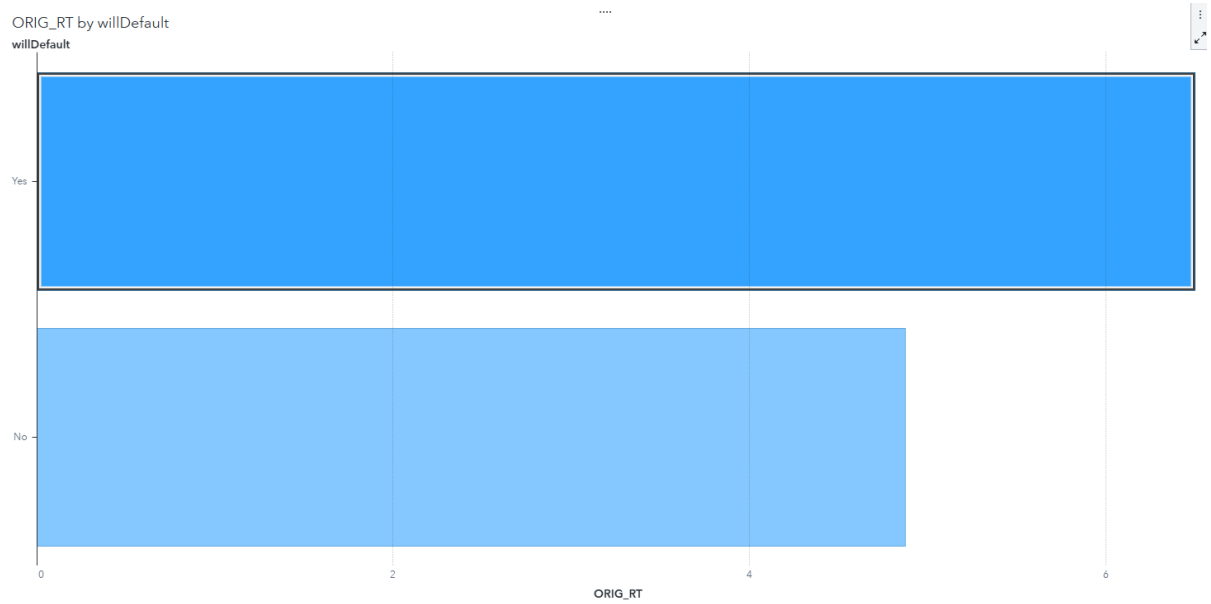
Original Amount by Purpose



The bar graph of the average original loan amount grouped by purchase purpose shows that loans taken for Purchase (P) have the highest average value. This is expected, since purchasing a new property often requires larger financing compared to other categories. In contrast, loans under the Unknown/Unspecified (U) category have the lowest average amounts, which may reflect smaller, less formal, or poorly reported transactions. Cash-

out refinance (C) and Rate-term refinance (R) sit in between, as refinancing usually involves existing equity and thus does not require as large a loan as an initial purchase.

Original Rate by Will Default



The bar graph of original interest rate grouped by loan default status shows that loans that eventually defaulted had, on average, about 1% higher interest rates than non-defaulted loans. This indicates that borrowers who started with higher mortgage rates faced larger monthly payments, increasing their likelihood of default. Higher rates can also signal subprime lending or weaker borrower profiles, making the loan riskier from the outset.

Data Transformation / Preprocessing

Rejecting Foreign Data

In this dataset, there were extra columns that were deemed unnecessary due to the business goals. These represented a variety of data about foreign markets such as the United Kingdom, Japan, and developing Asia countries. They were also found not to have a strong importance on the data after an initial data exploration; therefore, they were removed. In order to maximise the prediction accuracy, other features, such as the IDs, were changed from inputs to rejected, as they did not have an effect on the default value.

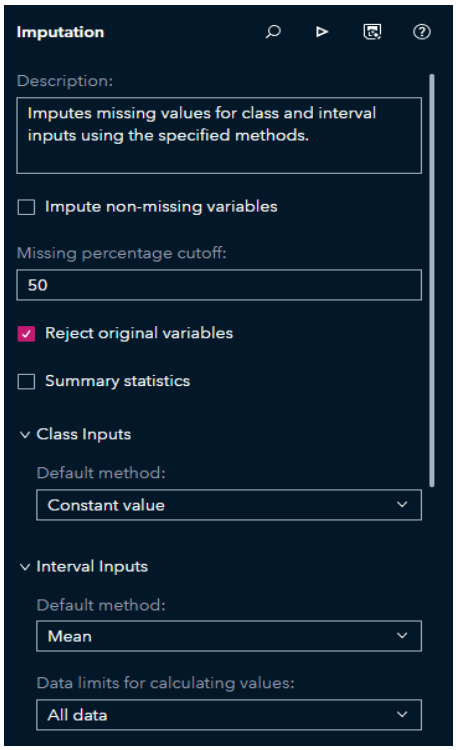
Missing Values

Initial data exploration found that the following columns had missing values:

Variable Name	Number Values Missing	Percentage Missing (%)
CSCORE_B	233	0.5123
CSCORE_C	25759	56.6368
DTI	747	1.6424
MI_PCT	34339	75.5019

MI_TYPE	34339	75.5019
NUM_BO	2	0.0044
OCLTV	5	0.0110

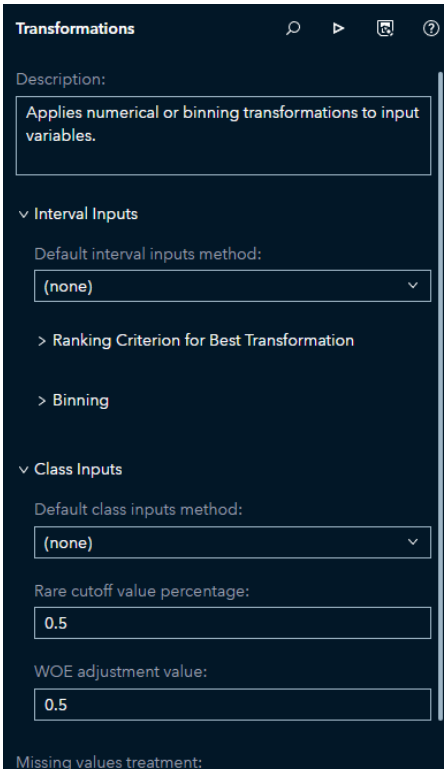
Most of these missing values were large chunks of the data. For variables such as credit score majority of the loans did not have a second individual applying; therefore, the lack of CSORE_C data, and the Mortgage insurance missing values may be because there is no insurance linked to the mortgage, or the insurance data on the individuals was incomplete. Other than that, I cannot see reasons why other variables would have missing data points. Due to this, I decided to reject the mortgage insurance datapoints and fill in the missing values for the other variables with the mean of the attribute with the imputation feature, with the following settings.



Log Transformations

Some numeric features in the dataset, such as Original Loan Amount (ORIG_AMT), Last Unpaid Principal Balance (LAST_UPB), and OLTV, show high skewness and heavy tails (high kurtosis), meaning a few extreme values dominate the distribution. To reduce the effect of these outliers and make the distributions more normal-like, a log transformation was applied. This helps stabilise variance, improves the interpretability of visualisations, and can enhance the performance of models that are sensitive to extreme values.

Variable Name		Transform
CSCORE_B	☑	Log
CSCORE_C	☑	Log
DTI	☑	Log
LAST_UPB	☑	Log
LastVersusOriginal	☑	Log
LoanAge	☑	Log
OCLTV	☑	Log
OLTV	☑	Log
ORIG_AMT	☑	Log
ORIG_RT	☑	Log



Transformed Variable	Method	Input Variable	Formula	Variable Level	Type	Variable Label
LOG_CSCORE_B	LOG	CSCORE_B	log('CSCORE_B'n + 1)	INTERVAL	N	Transformed CSCORE_B
LOG_CSCORE_C	LOG	CSCORE_C	log('CSCORE_C'n + 1)	INTERVAL	N	Transformed CSCORE_C
LOG_DTI	LOG	DTI	log('DTI'n + 1)	INTERVAL	N	Transformed DTI
LOG_LastVersusOriginal	LOG	LastVersusOriginal	log('LastVersusOriginal'n + 1)	INTERVAL	N	Transformed LastVersusOriginal
LOG_LAST_UPB	LOG	LAST_UPB	log('LAST_UPB'n + 1)	INTERVAL	N	Transformed LAST_UPB
LOG_LoanAge	LOG	LoanAge	log('LoanAge'n + 1)	INTERVAL	N	Transformed LoanAge
LOG_OCLTV	LOG	OCLTV	log('OCLTV'n + 1)	INTERVAL	N	Transformed OCLTV
LOG_OLTV	LOG	OLTV	log('OLTV'n + 1)	INTERVAL	N	Transformed OLTV
LOG_ORIG_AMT	LOG	ORIG_AMT	log('ORIG_AMT'n + 1)	INTERVAL	N	Transformed ORIG_AMT
LOG_ORIG_RT	LOG	ORIG_RT	log('ORIG_RT'n + 1)	INTERVAL	N	Transformed ORIG_RT

Challenges Encountered

During the data exploration process, several challenges were encountered. First, the dataset lacked clear descriptions and labels for many columns, making it difficult to understand some features without extra research. This required assumptions for certain variables, such as origination channels or property types.

Second, some features showed high skewness and extreme values, which could impact analysis and model performance. Transformations such as logarithmic scaling were applied to handle these outliers.

Third, the dataset includes a mix of borrower-level and macroeconomic variables, some of which had weak or indirect relationships to U.S. mortgage defaults. Deciding which features to include as inputs versus discard required careful consideration.

Finally, although the dataset was relatively balanced in terms of loans that defaulted versus those that did not, some features still had uneven distributions or extreme outliers, requiring careful preprocessing and transformation before modelling.