INST737: Introduction to Data Science

Project Milestone-1: Report

Instructor: Dr. Vanessa Frias-Martinez

Team5 : Sadaf Nasir Davre, Tanya Gupta, Ushasri Bhogaraju

Term : Fall '23

Introduction

The 30 year fixed rate mortgage home loans are the most common mortgage loan option in the United States. Its main advantage is the 'Predictability', the interest rate is locked for the entire 30- year term, and the monthly repayments are smaller. Fannie Mae is a federally backed institution that buys these and other mortgages from retail banks or lending institutions, and either holds them in its books or repackages them into 'Mortgage Backed securities(MBS)', for selling to investors. It maintains the Fannie Mae Data Dynamics site and publishes and maintains a variety of large datasets to comply with regulatory requirements and also facilitate analysis by various parties such as home buyers, investors and analysts.

For our Project, our team was interested in analyzing a large financial dataset that provides a variety of criterion variables that can be used to train a machine learning model using different techniques, to correctly predict the value of an outcome variable. We chose a dataset representing data for one quarter of recent vintage (2022Q4) from the Fannie Mae Single Family Acq and Perf Dataset as these files provide the variety in types of variables we require for our analysis.

1a. Research Questions

1. Can we accurately predict the 'Interest Rate' applied to a mortgage loan by a lending institution, using criterion such as the Quantum of loan, Loan-To-Value ratio, Debt-To-Income ratio, Loan Purpose, Number of Borrowers and Credit Score of the Borrower/s, using the Fannie Mae Acquisition and Performance 2022Q4 dataset?
2. Do all the borrower criteria such as the Quantum of loan, Loan-To-Value ratio, Debt-To-Income ratio, Loan Purpose, Number of Borrowers and Credit Score of the Borrower/s have equal influence over 'Interest Rate'? If not, which criteria have more effect on Interest Rate?
3. Do all lending institutions (presumed to be the 'Seller' in the dataset), evaluate the criterion variables similarly and fix the interest rate?

Significance of the Research questions

Technical perspective

We aim to build a Machine Learning model that will predict the 'Interest Rate' on the mortgage, based on a few predictor variables available in the dataset. The predictor variables are of different data types such as String, logical, Date, Numeric, Categorical and Nominal, and lend to rich analysis. We expect to find that the dataset is large enough to be split into Training and Test datasets. At a later stage, if we find that adding additional data will improve accuracy, we will add data from another quarter (downloaded form the same site). Preliminary examination reveals blank cells, Null values and erroneous data types that offer extensive opportunities to deploy several Data cleaning and Data transformation techniques. Data cleaning, merging, loading, splitting, subsetting etc., is contemplated using ‘R’ code. All in all, we find that answering our Research questions using the chosen dataset, offers the variety of experience required of a Data scientist, in learning different unsupervised Machine Learning techniques.

Societal impact perspective

Be it a home buyer or an investor, finding out how ‘Interest Rate’ is fixed on their mortgage is important. Since the 30-year fixed mortgage loan will carry the same interest rate for the next 30 years, borrowers must know how it is arrived at, what criteria were used, and what they can do to get a finer rate. Can they have a lower rate if the ‘Loan Purpose’ is different? Or if their ‘Credit score’ is better? Our analysis will help reveal some of these dependencies. Our analysis will add to the body of work that provides greater transparency and information to the public, and will enable well-informed, data-driven decision making.

1b. State of the Art

Research Articles

***1.* "Expectations and Interest Rates on Mortgage Loans" by Kay Mitusch and Dieter Nautz**

The research article titled "Expectations and Interest Rates on Mortgage Loans" by Kay Mitusch and Dieter Nautz delves into the dynamics of interest rate determination for mortgage loans. The authors found that there's a discrepancy between people's expectations and the actual interest rates they receive. The authors also mention that their results explain why some people choose loans with changing interest rates, while others go for loans with fixed rates.This phenomenon is even more pronounced in the home loan market than in other financial sectors. As we venture into analyzing the Fannie Mae Single Family Acq and Perf Dataset for the 2022Q4 vintage, the insights from Mitusch and Nautz's research become invaluable. Their findings on the divergence between expectations and actual interest rates can provide a fresh perspective when training our machine learning model. It emphasizes the importance of considering human expectations and behavioral factors, in addition to raw financial data, for accurate prediction and analysis.

2. “The Variation of Mortgage Interest Rates” by Alfred N. Page

Alfred N. Page's 1964 research paper aims to explore the connection between mortgage interest rates and various factors that lenders associate with the risk of mortgage default. The paper specifically tests the idea that mortgage rates fluctuate in relation to loan-to-value ratios, property values, and maturities. The study also evaluates the connection between mortgage rates and other factors such as lender assets, location, and loan fees. To investigate these relationships, Page used multiple regression analysis on seven cross-sections of data over time. This methodological approach was significant as past research largely ignored home mortgages due to a lack of comprehensive data. As we employ large datasets like the Fannie Mae Single and Perf Dataset for 2022Q4 to train machine learning models, insights on factors influencing mortgage interest rates can guide our feature selection and model interpretation. Page's examination of loan-to-value ratios, property values, and other variables offers valuable perspectives for our project's objective to predict the value of an outcome variable, enhancing the robustness and relevance of our analysis.

3.“Short-term Prediction of Mortgage Default using Ensembled Machine Learning Models” by Jesse C. Sealand

This study focuses on predicting mortgage defaults using machine learning models. The accuracy of these models depends on the training data's resemblance to future conditions. However, available data often covers a shorter timeframe than the loan period, leading to limitations in predictions. This study narrows its prediction window to the first 12 months of the loan's life, addressing the critical period when most defaults occur. It also investigates the reusability of machine learning models on new datasets over time, emphasizing practical applications for the mortgage industry. Predicting mortgage defaults is vital for lenders as it affects housing stability and financial losses. Machine learning offers promise in this domain, but balancing precision and recall, especially with large datasets, is challenging. Previous research demonstrated machine learning's effectiveness, particularly with large datasets, but optimizing models can be computationally intensive. Additionally, these models are often task-specific, making them challenging to reuse with updated datasets. The methodology used involves training machine learning models on annual datasets from 2000 to 2016 to predict mortgage defaults for the following year. Default is defined as more than 60 days overdue or missing two scheduled payments. The dataset includes 22 predictive features, and 11 classification algorithms are evaluated. Ensembling methods are explored, and results show that ensembling often outperforms single models. However, there's no universal ensembling method, and models with many adjustable parameters don't always perform significantly better when optimized. Importantly, this study highlights the reusability of machine learning models over time. Some models are reusable, but the reusability can vary by model year and the future prediction period. This research provides insights into which model years remain reusable and for how long, offering valuable guidance for practical applications in the mortgage industry, and provides valuable insights as we build the prediction model for our project.

4."A Few Useful Things to Know About Machine Learning" by Pedro Domingos

In the paper "A Few Useful Things to Know About Machine Learning," Pedro Domingos provides a succinct and informative overview of essential concepts and insights in the field of machine learning. The paper serves as a practical guide for our project, offering valuable knowledge distilled into key points. Domingos begins by emphasizing the importance of understanding the "No Free Lunch Theorem," which highlights that no single machine learning algorithm performs best for all types of problems. Instead, he suggests that practitioners should experiment with various algorithms to find the most suitable one for a specific task. The author delves into the bias-variance trade-off, a fundamental aspect of model performance. He explains how models with high bias tend to underfit data, while those with high variance overfit it. Striking the right balance is crucial for building accurate predictive models. Domingos discusses the importance of data in machine learning, emphasizing that more data often outweighs the complexity of algorithms. He also introduces the concept of the "curse of dimensionality" and its implications for feature selection and model complexity.

The paper highlights the need for proper validation techniques, such as cross-validation, to assess a model's performance effectively. Domingos emphasizes that a simple model with good validation results may be preferable to a complex one that overfits the data. Pedro Domingos also touches on the significance of ensemble methods, which combine multiple models to improve predictive accuracy. He introduces the concepts of bagging, boosting, and stacking, explaining their advantages and use cases.In conclusion, "A Few Useful Things to Know About Machine Learning" provides valuable insights into the principles of machine learning. It encourages practitioners to approach the field with a thoughtful and experimental mindset, understanding the trade-offs, data's central role, and the potential benefits of ensemble techniques. This paper serves as an essential resource for our project.

1c. Datasets

Collection process: The Fannie Mae Data Dynamics site requires registration with our email and agreeing to terms and conditions of use. The site provides extensive metadata, tutorials and answers to FAQs. Upon exploration of different datasets and choosing to work on Single Family Historical Loan Performance data, we navigated to our Fannie Mae Single Family Acq and Perf Dataset (2022Q4) by choosing the Historical Loan performance data link on the Fannie Mae Data Dynamics page. The ‘SF Glossary and layout file’ contains metadata, along with the headers and detailed descriptions of each of the variables contained in the Acquisition and Performance file and is available under ‘Resources’ tab on the Historical Loan Credit Performance Data Page for download.

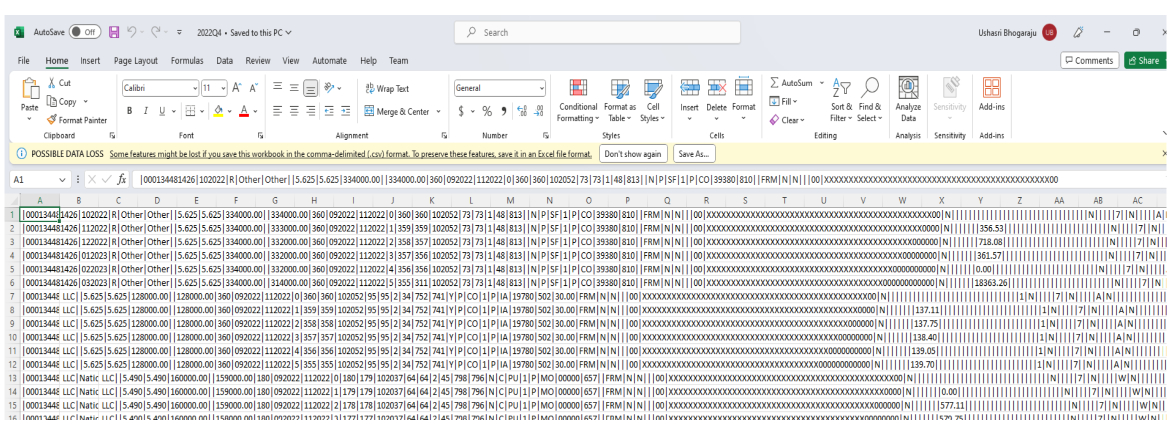
We downloaded both the Acquisition and Performance ‘2022Q4.zip’ and the ‘crt-file-layout-and-glossary’ file from the above locations to our Desktops after registration on the site.

Description:

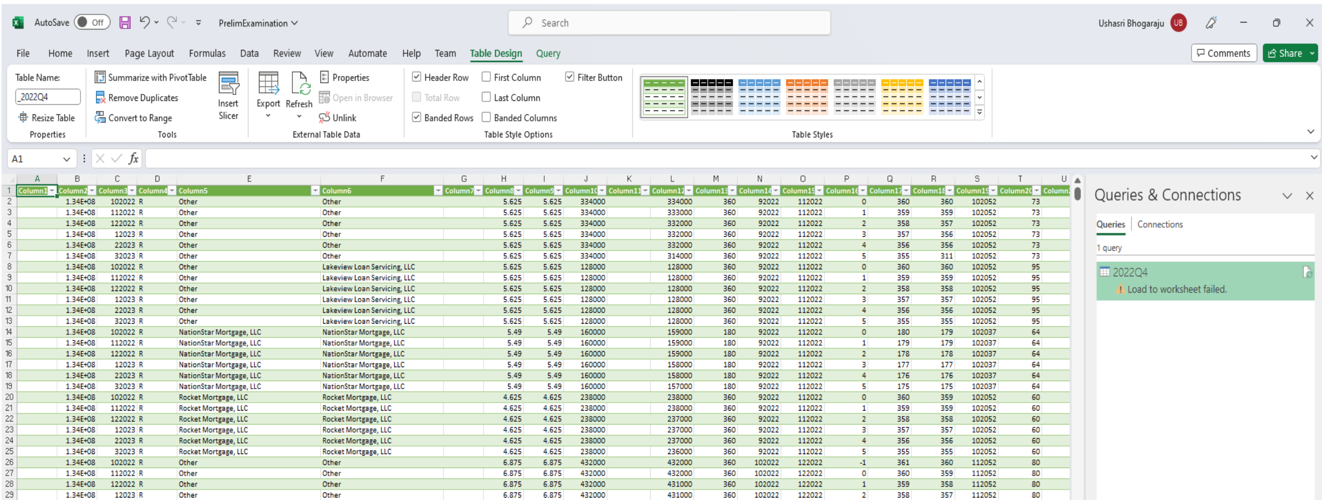
The Acquisition and Performance Dataset, 2022Q4.zip : We extracted the zip file and generated the 2022Q4.csv file of size 427,889 KB on the disk. When we opened it by double clicking, the default option for opening a csv file, Excel, threw an error message saying that not all data could be uploaded and generated an output file, which revealed the following.

1. There were no headers in the dataset
2. The variables are separated by the ‘|’ delimiter.
3. Excel could only load 1,048,576 rows from the file, there could be more rows

Image:



We then tried to explore this data by loading this csv file into Excel, by using the ‘Get Data’ option and specifying the delimiter as custom, and providing ‘|’ as the value and then matching the column headers from the ‘crt-file-layout-and-glossary’ file and found the following.



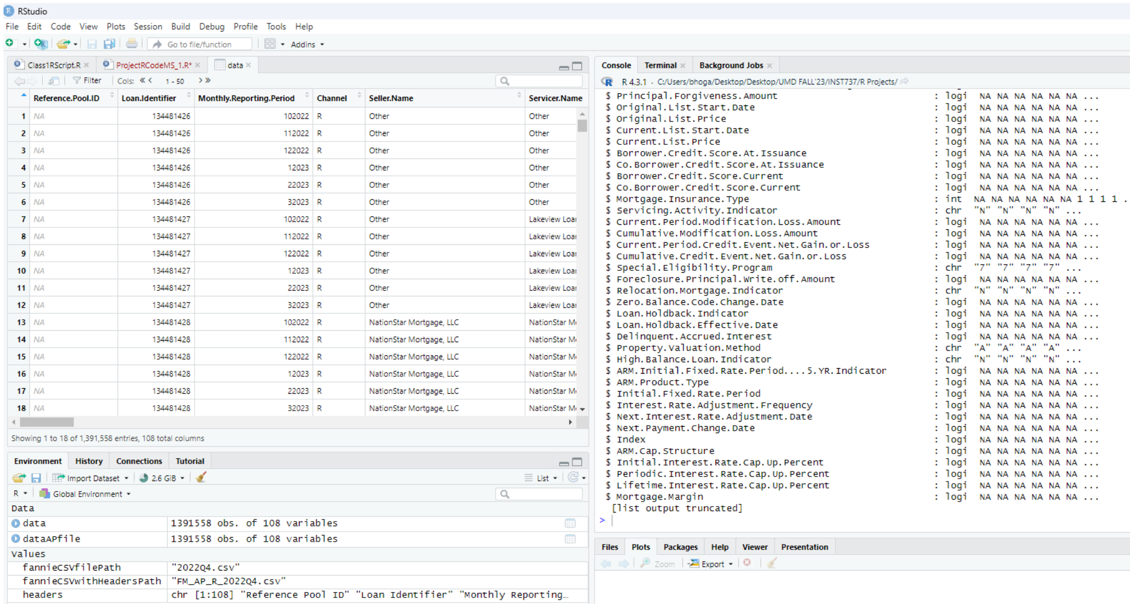
a. In the dataset file, there are multiple rows for the same Loan Identifier
b. There are 108 columns in the dataset, many of which are blank
c. Some data types appear to be incorrect and Data cleaning will be required.
d. Headers data from the layout file must be absorbed into the dataset.
e. The number of rows in the dataset are more than the maximum for Excel.

General Statistics

Column headers from the layout file were inserted into the original dataset and the FM_AP_R_2022Q4.csv file was generated using R code.

1. The dimensions of the data in the primary dataset are 1,391,558 observations of 108 variables.
2. The layout file has 108 rows with field names and other metadata provided in 10 columns.
3. The Primary Dataset 2022Q4.csv was explored using str(data), View(data), glimpse(data) and head(data) commands in R. str command produced details of 102/108 variables with a message ‘list output truncated’ at the end. Head command produced 1st 6 records of all 108 variables. We produce the image from R studio captured during our Team meeting.

dim(data), View(data) & str(data) image:



Shortlisted Variables and Unique Loans set **:** The shape of data set after transformation of data types and filtering for Unique Loans based on monthly reporting period on 32023, *(the detailed steps for which are enumerated in Data Cleaning Efforts) is 271368 obs and 29 variables.* The glimpse command and the table of variables at this stage is without replacing any values of the Original dataset but changing date datatype.

```
> glimpse(data)
Rows: 271,368
Columns: 29
$ Loan.Identifier                                <int> 134481426, 134481427, 134481428, 134481429, 134481430, 134481431, 1…
$ Monthly.Reporting.Period                       <chr> "2023-03-01", "2023-03-01", "2023-03-01", "2023-03-01", "2023-03-01…
$ Channel                                        <chr> "R", "R", "R", "R", "R", "B", "R", "R", "R", "C", "B", "R", "C", "C…
$ Seller.Name                                    <chr> "Other", "Other", "NationStar Mortgage, LLC", "Rocket Mortgage, LLC…
$ Original.Interest.Rate                         <dbl> 5.625, 5.625, 5.490, 4.625, 6.875, 5.625, 5.875, 5.250, 5.000, 6.62…
$ Original.UPB                                   <dbl> 334000, 128000, 160000, 238000, 432000, 225000, 360000, 150000, 293…
$ Original.Loan.Term                             <int> 360, 360, 180, 360, 360, 360, 360, 360, 360, 360, 360, 360, 360, 36…
$ Origination.Date                               <chr> "2022-09-01", "2022-09-01", "2022-09-01", "2022-09-01", "2022-10-01…
$ Loan.Age                                       <int> 5, 5, 5, 5, 4, 4, 6, 5, 7, 5, 5, 4, 4, 5, 4, 4, 5, 5, 5, 4, 6, 4, 5…
$ Remaining.Months.to.Legal.Maturity             <int> 355, 355, 175, 355, 356, 356, 354, 355, 353, 355, 355, 356, 356, 35…
$ Remaining.Months.To.Maturity                   <int> 311, 355, 175, 355, 355, 356, 354, 353, 346, 355, 355, 356, 356, 35…
$ Maturity.Date                                  <chr> "2052-10-01", "2052-10-01", "2037-10-01", "2052-10-01", "2052-11-01…
$ Original.Loan.to.Value.Ratio..LTV.             <int> 73, 95, 64, 60, 80, 36, 90, 75, 90, 90, 97, 95, 90, 95, 80, 70, 80,…
$ Original.Combined.Loan.to.Value.Ratio..CLTV.   <int> 73, 95, 64, 60, 80, 36, 90, 75, 90, 90, 97, 95, 100, 95, 80, 90, 80…
$ Number.of.Borrowers                            <int> 1, 2, 2, 2, 2, 1, 2, 2, 2, 2, 1, 2, 1, 1, 2, 2, 2, 1, 2, 1, 1, 1, 1…
$ Debt.To.Income..DTI.                           <int> 48, 34, 45, 43, 39, 45, 50, 27, 47, 39, 34, 50, 49, 49, 44, 41, 14,…
$ Borrower.Credit.Score.at.Origination           <int> 813, 752, 798, 623, 738, 783, 730, 772, 761, 770, 779, 790, 766, 78…
$ Co.Borrower.Credit.Score.at.Origination        <int> NA, 741, 796, 728, 727, NA, 703, 792, 775, 718, NA, NA, NA, NA, 802…
$ First.Time.Home.Buyer.Indicator                <chr> "N", "Y", "N", "N", "N", "N", "Y", "N", "Y", "Y", "Y", "Y", "Y", "N…
$ Loan.Purpose                                   <chr> "P", "P", "C", "C", "P", "P", "P", "P", "P", "P", "P", "P", "P", "P…
$ Property.Type                                  <chr> "SF", "CO", "PU", "PU", "PU", "PU", "PU", "SF", "SF", "CO", "SF", "…
$ Number.of.Units                                <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1…
$ Occupancy.Status                               <chr> "P", "P", "P", "P", "P", "P", "P", "P", "P", "P", "P", "P", "P", "P…
$ Property.State                                 <chr> "CO", "IA", "MO", "FL", "GA", "FL", "FL", "OH", "PA", "AL", "PA", "…
$ Metropolitan.Statistical.Area..MSA.            <int> 39380, 19780, 0, 45300, 12060, 45300, 36740, 17140, 37980, 19300, 3…
$ Zip.Code.Short                                 <int> 810, 502, 657, 335, 301, 337, 327, 452, 190, 365, 150, 606, 953, 33…
$ Amortization.Type                              <chr> "FRM", "FRM", "FRM", "FRM", "FRM", "FRM", "FRM", "FRM", "FRM", "FRM…
$ Prepayment.Penalty.Indicator                   <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N…
$ Current.Loan.Delinquency.Status                <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
```

When checked for Null values in the above dataset and found that two of our Predictor variables have Null values and the outcome variable does not have null values. There are Null values in other variables not considered at this stage for analysis.

```
                      Loan.Identifier                Monthly.Reporting.Period
                                    0                                       0
                              Channel                             Seller.Name
                                    0                                       0
               Original.Interest.Rate                            Original.UPB
                                    0                                       0
                   Original.Loan.Term                        Origination.Date
                                    0                                       0
                             Loan.Age         Remaining.Months.to.Legal.Maturity
                                 1208                                    1208
          Remaining.Months.To.Maturity                           Maturity.Date
                                 1208                                    1208
  Original.Loan.to.Value.Ratio..LTV. Original.Combined.Loan.to.Value.Ratio..CLTV.
                                    0                                       0
                    Number.of.Borrowers                        Debt.To.Income..DTI.
                                    0                                      11
   Borrower.Credit.Score.at.Origination    Co.Borrower.Credit.Score.at.Origination
                                  364                                  150215
         First.Time.Home.Buyer.Indicator                            Loan.Purpose
                                    0                                       0
                        Property.Type                         Number.of.Units
                                    0                                       0
                     Occupancy.Status                          Property.State
                                    0                                       0
  Metropolitan.Statistical.Area..MSA.                          Zip.Code.Short
                                    0                                       0
                    Amortization.Type                Prepayment.Penalty.Indicator
                                    0                                       0
        Current.Loan.Delinquency.Status
                                    0
```

We preserved the shortlisted dataset in a csv file and named it "Final_FM_AP_R_2022Q4.csv" for later use as required. We created a final variables data set with 270993 obs. And 13 variables and named it "FinalVariables_FM_AP_R_2022Q4.csv" after removing Null values and preserving only predictor and outcome variables for ease of use for visualizations etc.

Image of str() command:

```
str(finalvariablesdata)
data.frame':   270993 obs. of  13 variables:
$ Original.Interest.Rate                      : num  5.62 5.62 5.49 4.62 6.88 ...
$ Original.Loan.to.Value.Ratio..LTV.          : int  73 95 64 60 80 36 90 75 90 90 ...
$ Original.UPB                                : int  334000 128000 160000 238000 432000 225000 360000 150000 293000 105000 ..
$ Number.of.Borrowers                         : int  1 2 2 2 2 1 2 2 2 2 ...
$ Original.Combined.Loan.to.Value.Ratio..CLTV.: int  73 95 64 60 80 36 90 75 90 90 ...
$ Debt.To.Income..DTI.                        : int  48 34 45 43 39 45 50 27 47 39 ...
$ Loan.Purpose                                : chr  "P" "P" "C" "C" ...
$ Property.Type                               : chr  "SF" "CO" "PU" "PU" ...
$ Number.of.Units                             : int  1 1 1 1 1 1 1 1 1 1 ...
$ Occupancy.Status                            : chr  "P" "P" "P" "P" ...
$ Seller.Name                                 : chr  "Other" "Other" "NationStar Mortgage, LLC" "Rocket Mortgage, LLC" ...
$ Property.State                              : chr  "CO" "IA" "MO" "FL" ...
$ Borrower.Credit.Score.at.Origination        : int  813 752 798 623 738 783 730 772 761 770 ...
```

Statistical Analysis of the Predictor and Outcome Variables:

Summary: The dataset "FinalVariables_FM_AP_R_2022Q4.csv" with Unique rows, and columns with Predictor Variables and the Outcome variable was created to facilitate viewing outliers and understand the distributions easily. It has 270993 obs of 13 variables. Our outcome variable, Original Interest Rate has a continuous distribution. It has min val of 2.125 and a maximum value of 8.125 and mean value of 6.020 and median value of 5.990.

Our Predictor variables are 12 in number, 7 numerical, 3 categorical and 2 nominal variables. Their shape and statistical values are as under:

```
> summary(data)
 Original.Interest.Rate Original.Loan.to.Value.Ratio..LTV.  Original.UPB     Number.of.Borrowers
 Min.   :2.125          Min.   : 4.00                      Min.   :  14000   Min.   :1.000
 1st Qu.:5.437          1st Qu.:66.00                      1st Qu.: 188000   1st Qu.:1.000
 Median :5.990          Median :80.00                      Median : 280000   Median :1.000
 Mean   :6.020          Mean   :75.49                      Mean   : 305116   Mean   :1.467
 3rd Qu.:6.625          3rd Qu.:90.00                      3rd Qu.: 400000   3rd Qu.:2.000
 Max.   :8.125          Max.   :97.00                      Max.   :1800000   Max.   :5.000
 Original.Combined.Loan.to.Value.Ratio..CLTV. Debt.To.Income..DTI. Loan.Purpose       Property.Type
 Min.   :  4.00                               Min.   : 1.00        Length:270993      Length:270993
 1st Qu.: 66.00                               1st Qu.:31.00        Class :character   Class :character
 Median : 80.00                               Median :39.00        Mode  :character   Mode  :character
 Mean   : 75.76                               Mean   :37.46
 3rd Qu.: 91.00                               3rd Qu.:45.00
 Max.   :105.00                               Max.   :62.00
 Number.of.Units Occupancy.Status   Seller.Name        Property.State     Borrower.Credit.Score.at.Origination
 Min.   :1.000   Length:270993      Length:270993      Length:270993      Min.   :472.0
 1st Qu.:1.000   Class :character   Class :character   Class :character   1st Qu.:726.0
 Median :1.000   Mode  :character   Mode  :character   Mode  :character   Median :762.0
 Mean   :1.022                                                            Mean   :753.6
 3rd Qu.:1.000                                                            3rd Qu.:789.0
 Max.   :4.000                                                            Max.   :840.0
```

Outliers on examination of summary data:

On inspection of the values returned by the summary command, we can see that there are outliers in the Original LTV, Original UPB, Original Combined CLTV, DTI and Credit score variables. We intend to use Histograms to detect outliers as well as understand the distributions of numerical variables and Pie charts and Bar plots to understand the distributions of categorical and nominal variables. This analysis will be discussed in the Data visualization portion.

Data Cleaning Efforts

We extracted, reduced, filtered and transformed data from 2 files, our primary dataset ‘2022Q4.zip’ and ‘crt-file-layout-and-glossary.xlsx’ to arrive at our final dataset. Our primary dataset had 1390558 rows and 108 variables and the layout file had 108 rows and 10 columns with metadata. Both these were transformed and data merged. Our shortlisted dataset "Final_FM_AP_R_2022Q4.csv" has 271368 obs. and 29 variables. A more concise "FinalVariables_FM_AP_R_2022Q4.csv" dataset with only predictor and outcome variables was also created. The steps in generating these files are:

Step 1: Extracting Column headers from ‘crt-file-layout-and-glossary.xlsx’ using excel formulae :

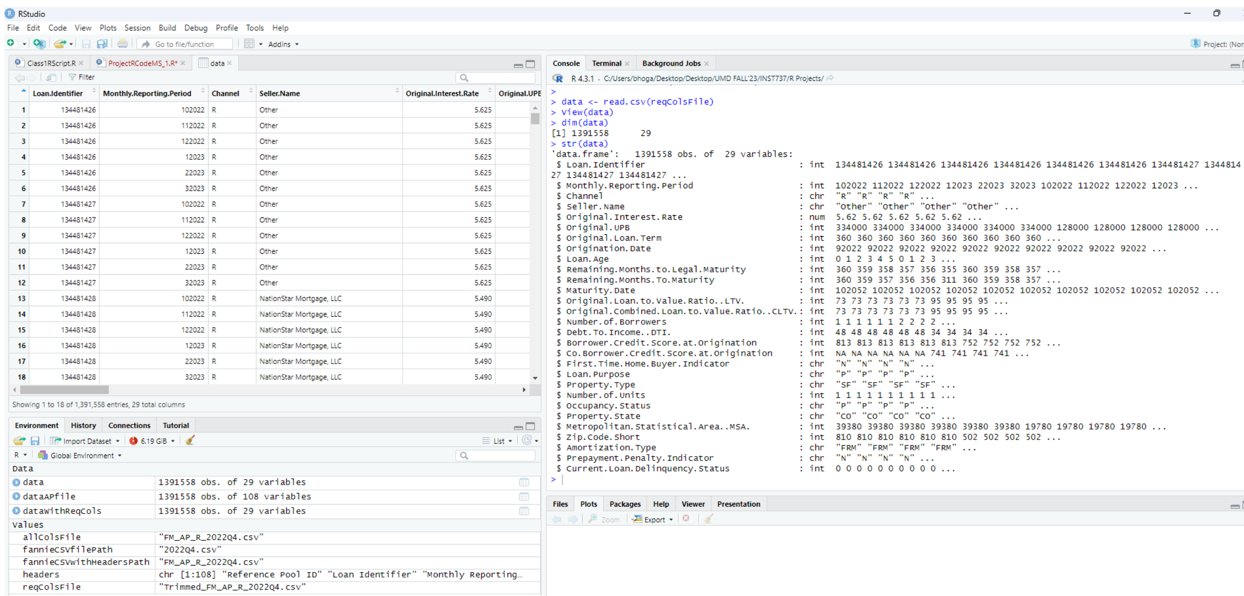
A	B	C	D	E	F	G	H	I	J
Field Position	Field Name	Description	Date Bound Notes	Respective Disclosure Notes	CAS	CIRT	Single-Family (SF) Loan Performance	Type	Max Length
1	Reference Pool ID	A unique identifier for the reference pool.			√	√	NA	ALPHA-NUMERIC	X(4)
2	Loan Identifier	A unique identifier for the mortgage loan.		The Loan Identifier does not correspond to other mortgage loan identifiers found in existing Fannie Mae disclosures.	√	√	√	ALPHA-NUMERIC	X(12)
3	Monthly Reporting Period	The month and year that pertains to the servicer's cut-off period for mortgage loan information.	SF Loan Performance: Enhanced format with the October 2020 Release		√	√	√	DATE	MMYYYY
4	Channel	The origination channel used by the party that delivered the loan to the issuer.		Retail = R; Correspondent = C; Broker = B	√	√	√	ALPHA-NUMERIC	X(1)
5	Seller Name	The name of the entity that delivered the mortgage loan to Fannie Mae.		CAS/CIRT: For sellers whose combined loans' contribution to the At Issuance UPB represents less than 1% of the total At Issuance reference pool UPB, the file will reflect "Other". SF Loan Performance: For sellers that represent less than one percent of volume within a given acquisition quarter as represented by the original unpaid principal balance, "Other" will be displayed in this field.	√	√	√	ALPHA-NUMERIC	X(50)
6	Servicer Name	The name of the entity that serves as the primary servicer of the mortgage loan.	SF Loan Performance: For activity periods prior to December 2001, Servicer Name will be blank since the servicer information for this period is unavailable.	CAS/CIRT: For servicers whose combined loans' contribution to the At Issuance UPB represents less than 1% of the total At Issuance reference pool UPB, the file will reflect "Other". SF Loan Performance: For servicers that represent less than one percent of the current actual unpaid principal balance for the last month of a given quarter, "Other" will be displayed in this field.	√	√	√	ALPHA-NUMERIC	X(50)

1. Transpose the column with Field names.
2. Insert double quotes to the column headers using custom formatting option in excel
3. Concatenate the cells, using CONCATENATE(A1:H1)&"," to insert commas and copy to code

Step 2: Inserting headers into the primary dataset 2022Q4.zip and creating a csv file with headers, naming it FM_AP_R_2022Q4.csv using R code

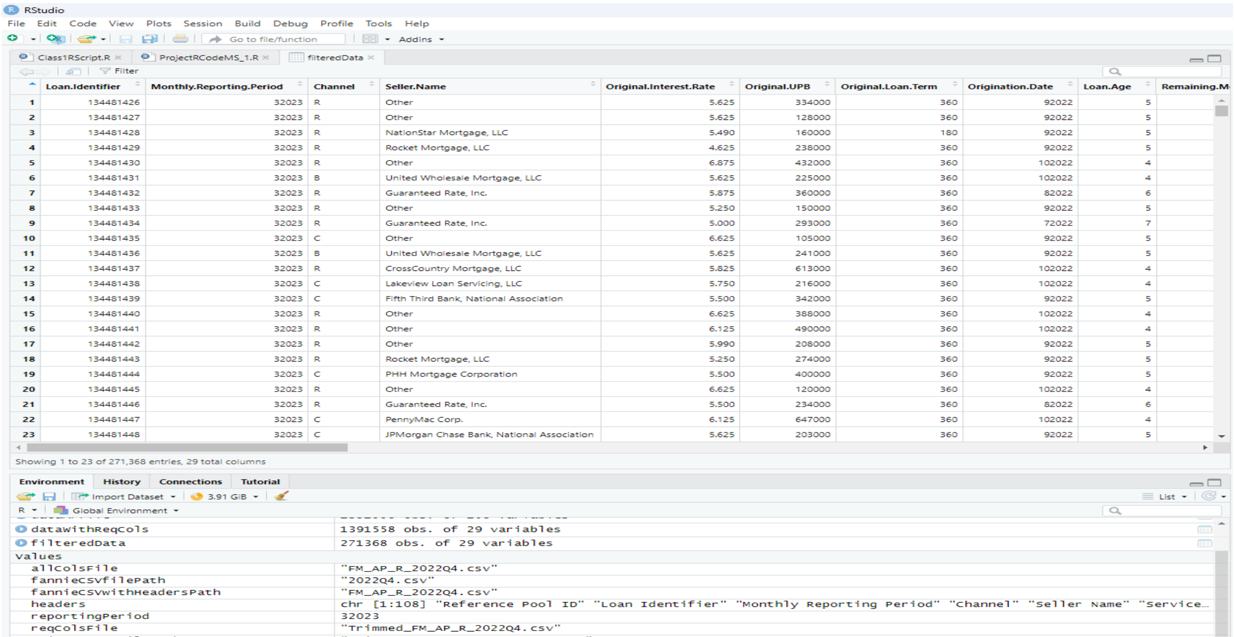
Step 3: Subsetting 'only required' columns (at this point the number of columns we identified as interesting were 29/108) from the "FM_AP_R_2022Q4.csv" file. This was accomplished using R commands and the "Trimmed_FM_AP_R_2022Q4.csv" file was generated. This occupied 195,730KB disk space. The dimensions of this dataset are 1391558 obs. of 29 variables.

dim(data), View(data) & str(data) image after above step:



Step 4: We have multiple rows of data for one Loan Identifier, one row for each reporting period (6 months) from 102022 to 032023 in the dataset. So, the next task is eliminate duplicates.to retain unique values in rows. This was achieved using R code. We named this dataset "UniqueLoans_FM_AP_R_2022Q4.csv"

Image of the dataset with Unique rows:



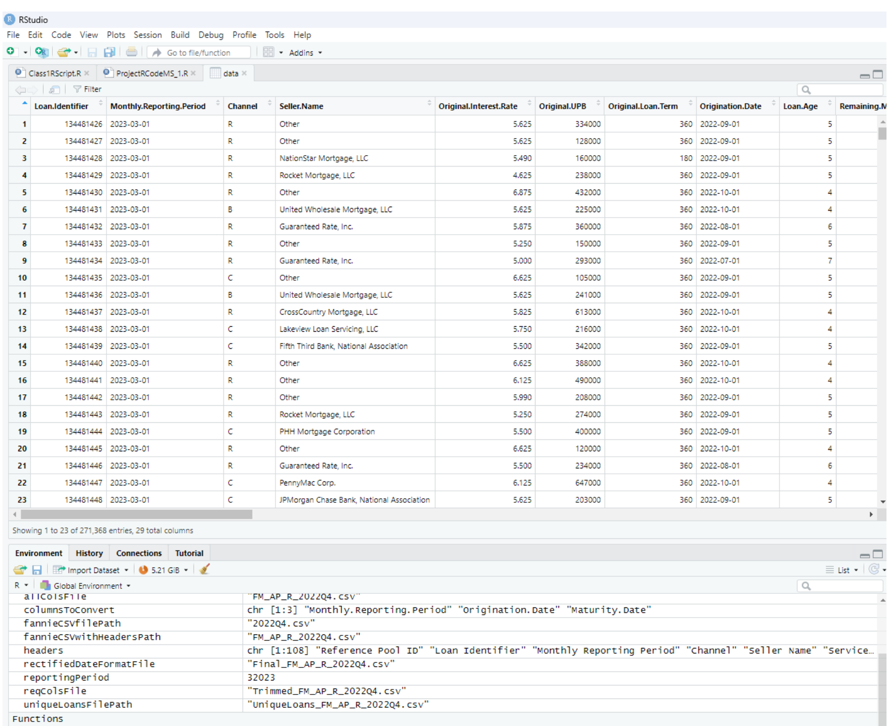
	Loan.Identifier	Monthly.Reporting.Period	Channel	Seller.Name	Original.Interest.Rate	Original.UPB	Original.Loan.Term	Origination.Date	Loan.Age	Remaining.M
1	134481426	32023	R	Other	5.625	334000	360	92022	5	
2	134481427	32023	R	Other	5.625	128000	360	92022	5	
3	134481428	32023	R	NationStar Mortgage, LLC	5.490	160000	180	92022	5	
4	134481429	32023	R	Rocket Mortgage, LLC	4.625	238000	360	92022	5	
5	134481430	32023	R	Other	6.875	432000	360	102022	4	
6	134481431	32023	B	United Wholesale Mortgage, LLC	5.625	225000	360	102022	4	
7	134481432	32023	R	Guaranteed Rate, Inc.	5.875	360000	360	82022	6	
8	134481433	32023	R	Other	5.250	150000	360	92022	5	
9	134481434	32023	R	Guaranteed Rate, Inc.	5.000	293000	360	72022	7	
10	134481435	32023	C	Other	6.625	105000	360	92022	5	
11	134481436	32023	B	United Wholesale Mortgage, LLC	5.625	241000	360	92022	5	
12	134481437	32023	R	CrossCountry Mortgage, LLC	5.825	613000	360	102022	4	
13	134481438	32023	C	Lakeview Loan Servicing, LLC	5.750	216000	360	102022	4	
14	134481439	32023	C	Fifth Third Bank, National Association	5.500	342000	360	92022	5	
15	134481440	32023	R	Other	6.625	388000	360	102022	4	
16	134481441	32023	R	Other	6.125	490000	360	102022	4	
17	134481442	32023	R	Other	5.990	208000	360	92022	5	
18	134481443	32023	R	Rocket Mortgage, LLC	5.250	274000	360	92022	5	
19	134481444	32023	C	PHH Mortgage Corporation	5.500	400000	360	92022	5	
20	134481445	32023	R	Other	6.625	120000	360	102022	4	
21	134481446	32023	R	Guaranteed Rate, Inc.	5.500	234000	360	82022	6	
22	134481447	32023	C	PennyMac Corp.	6.125	647000	360	102022	4	
23	134481448	32023	C	JPMorgan Chase Bank, National Association	5.625	203000	360	92022	5	

Showing 1 to 23 of 271,368 entries, 29 total columns

dataWithReqCols	1391558 obs. of 29 variables
filteredData	271368 obs. of 29 variables
Values	
allColsFile	"FM_AP_R_2022Q4.csv"
fannieCsvFilePath	"2022Q4.csv"
fannieCsvWithHeadersPath	"FM_AP_R_2022Q4.csv"
headers	chr [1:108] "Reference Pool ID" "Loan Identifier" "Monthly Reporting Period" "Channel" "Seller Name" "Service…
reportingPeriod	32023
reqColsFile	"Trimmed_FM_AP_R_2022Q4.csv"
uniqueLoansFilePath	"UniqueLoans_FM_AP_R_2022Q4.csv"

Step 5: Rectification of data types. From viewing the filtered dataset, we find that there are 3 date columns that appear as int data type. Conversion was achieved using R code. We named this dataset "Final_FM_AP_R_2022Q4.csv"

Image of the Dataset with Date fields rectified:



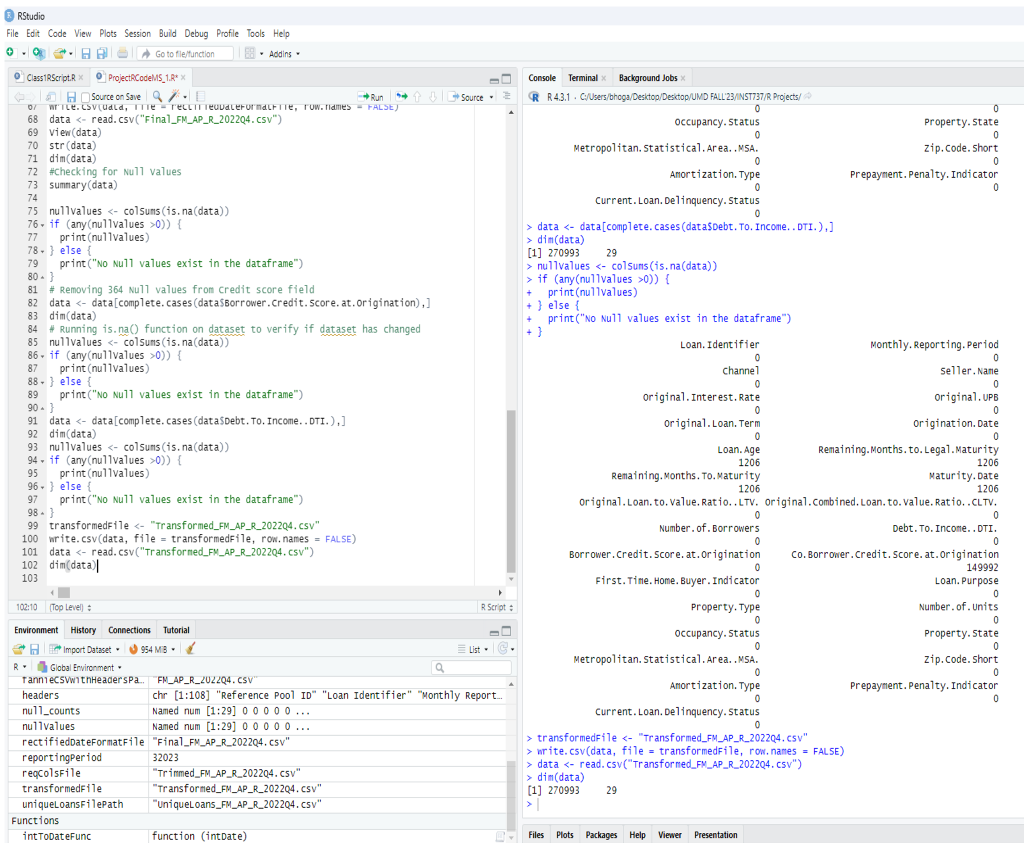
	Loan.Identifier	Monthly.Reporting.Period	Channel	Seller.Name	Original.Interest.Rate	Original.UPB	Original.Loan.Term	Origination.Date	Loan.Age	Remaining.M
1	134481426	2023-03-01	R	Other	5.625	334000	360	2022-09-01	5	
2	134481427	2023-03-01	R	Other	5.625	128000	360	2022-09-01	5	
3	134481428	2023-03-01	R	NationStar Mortgage, LLC	5.490	160000	180	2022-09-01	5	
4	134481429	2023-03-01	R	Rocket Mortgage, LLC	4.625	238000	360	2022-09-01	5	
5	134481430	2023-03-01	R	Other	6.875	432000	360	2022-10-01	4	
6	134481431	2023-03-01	B	United Wholesale Mortgage, LLC	5.625	225000	360	2022-10-01	4	
7	134481432	2023-03-01	R	Guaranteed Rate, Inc.	5.875	360000	360	2022-08-01	6	
8	134481433	2023-03-01	R	Other	5.250	150000	360	2022-09-01	5	
9	134481434	2023-03-01	R	Guaranteed Rate, Inc.	5.000	293000	360	2022-07-01	7	
10	134481435	2023-03-01	C	Other	6.625	105000	360	2022-09-01	5	
11	134481436	2023-03-01	B	United Wholesale Mortgage, LLC	5.625	241000	360	2022-09-01	5	
12	134481437	2023-03-01	R	CrossCountry Mortgage, LLC	5.825	613000	360	2022-10-01	4	
13	134481438	2023-03-01	C	Lakeview Loan Servicing, LLC	5.750	216000	360	2022-10-01	4	
14	134481439	2023-03-01	C	Fifth Third Bank, National Association	5.500	342000	360	2022-09-01	5	
15	134481440	2023-03-01	R	Other	6.625	388000	360	2022-10-01	4	
16	134481441	2023-03-01	R	Other	6.125	490000	360	2022-10-01	4	
17	134481442	2023-03-01	R	Other	5.990	208000	360	2022-09-01	5	
18	134481443	2023-03-01	R	Rocket Mortgage, LLC	5.250	274000	360	2022-09-01	5	
19	134481444	2023-03-01	C	PHH Mortgage Corporation	5.500	400000	360	2022-09-01	5	
20	134481445	2023-03-01	R	Other	6.625	120000	360	2022-10-01	4	
21	134481446	2023-03-01	R	Guaranteed Rate, Inc.	5.500	234000	360	2022-08-01	6	
22	134481447	2023-03-01	C	PennyMac Corp.	6.125	647000	360	2022-10-01	4	
23	134481448	2023-03-01	C	JPMorgan Chase Bank, National Association	5.625	203000	360	2022-09-01	5	

Showing 1 to 23 of 271,368 entries, 29 total columns

allColsFile	"FM_AP_R_2022Q4.csv"
columnsToConvert	chr [1:3] "Monthly.Reporting.Period" "Origination.Date" "Maturity.Date"
fannieCsvFilePath	"2022Q4.csv"
fannieCsvWithHeadersPath	"FM_AP_R_2022Q4.csv"
headers	chr [1:108] "Reference Pool ID" "Loan Identifier" "Monthly Reporting Period" "Channel" "Seller Name" "Service…
rectifiedDateFormatFile	"Final_FM_AP_R_2022Q4.csv"
reportingPeriod	32023
reqColsFile	"Trimmed_FM_AP_R_2022Q4.csv"
uniqueLoansFilePath	"UniqueLoans_FM_AP_R_2022Q4.csv"
Functions	
intToDateFunc	function (intDate)

Step 6: ***Checking for Null values and rectification :*** The R 'summary()' command revealed that there are Null values in a few columns in the final dataset. We used the is.na() function to reveal the number of Null values and whether they are in the variables we primarily intend to use in our analysis. *For detecting outliers in numerical data we will use Histograms. Since we have the csv files with datasets created at various stages, we decided to* remove rows with Null values in Predictor variable columns and create a csv file and name it 'Transformed_FM_AP_R_2022Q4.csv'

Image of the transformed dataset



List of 29 Variables' data shortlisted, dimension of this dataset, 270993obs, 29 variables:

Final shape of variables After removing Null values from DTI

S.No	***Name of the Variable***	***Data type***	***Count Nulls***	***Remarks***
1	Loan.Identifier	int	0	
2	Monthly.Reporting.Period	chr	0	
3	Channel	chr	0	
4	Seller.Name	chr	0	Contains 111234/ 271368 values in 'Other' category
5	Original.Interest.Rate	num	0	***Outcome variable***
6	Original.UPB	int	0	Predictor variable
7	Original.Loan.Term	int	0	
8	Origination.Date	chr	0	
9	Loan.Age	int	1206	
10	Remaining.Months.to.Legal.Maturity	int	1206	
11	Remaining.Months.To.Maturity	int	1206	
12	Maturity.Date	chr	1206	
13	Original.Loan.to.Value.Ratio..LTV.	int	0	Predictor variable
14	Original.Combined.Loan.to.Value.Ratio..CL	int	0	Predictor variable
15	Number.of.Borrowers	int	0	Predictor variable
16	Debt.To.Income..DTI.	int	0	Predictor variable, will be addressed in data cleaning efforts
17	Borrower.Credit.Score.at.Origination	int	0	Predictor variable, will be addressed in data cleaning efforts
18	Co.Borrower.Credit.Score.at.Origination	int	149992	
19	First.Time.Home.Buyer.Indicator	chr	0	
20	Loan.Purpose	chr	0	Predictor variable
21	Property.Type	chr	0	Predictor variable
22	Number.of.Units	int	0	
23	Occupancy.Status	chr	0	Predictor variable
24	Property.State	chr	0	Predictor variable
25	Metropolitan.Statistical.Area..MSA.	int	0	
26	Zip.Code.Short	int	0	
27	Amortization.Type	chr	0	
28	Prepayment.Penalty.Indicator	chr	0	
29	Current.Loan.Delinquency.Status	int	0	

Further we created a FinalVariables dataset and named it "FinalVariables_FM_AP_R_2022Q4.csv" for ease of analysis: shape of this dataset is 270993 obs, 13 variables without Null values

S.No	***Name of the Variable***	***Data type***	***Count Nulls***	***Remarks***
1	Original.Interest.Rate	num	0	***Outcome variable, Continuous***
2	Original.UPB	int	0	Predictor variable, Discrete
3	Original.Loan.to.Value.Ratio..LTV.	int	0	Predictor variable, Discrete
4	Original.Combined.Loan.to.Value.Ratio..CL	int	0	Predictor variable, Discrete
5	Number.of.Borrowers	int	0	Predictor variable, Discrete
6	Debt.To.Income..DTI.	int	0	Predictor variable, Discrete
7	Borrower.Credit.Score.at.Origination	int	0	Predictor variable, Discrete
8	Loan.Purpose	chr	0	Predictor variable, Categorical
9	Property.Type	chr	0	Predictor variable, Categorical
10	Occupancy.Status	chr	0	Predictor variable, Categorical
11	Property.State	chr	0	Predictor variable, Nominal
12	Seller.Name	chr	0	Contains 111234/ 271368 values in 'Other' category, Nominal
13	Number.of.Units	int	0	Predictor variable, Discrete

Other Software Engineering Efforts

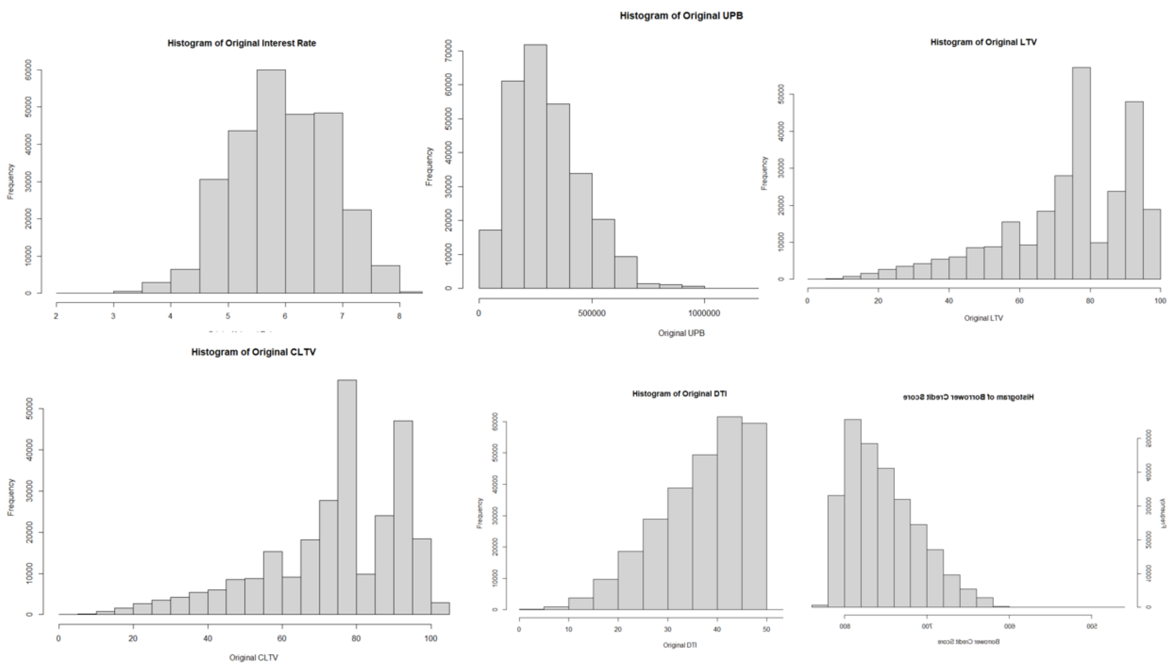
1. We went through Fannie Mae Wiki to understand the role of Fannie Mae in the mortgage industry.
2. We watched this brief Youtube video at https://youtu.be/iKSvAsm3ago?si=TQL7e7AI03ERo1hw to understand the Data Dynamics site and resources available there and their purpose.
3. We then read the tutorial provided under the 'Resources' tab on the Data Dynamics page to understand how to use the SF Loan Performance(primary) dataset.
4. We read R documentation and read various Posit Community troubleshooting posts to troubleshoot our data transformation efforts.
5. We sought help from the Instructional team whenever we were stuck without wasting time while parallelly trying to solve the problem on our own, using other resources.

Data Visualizations, detection of Outliers in Predictor and Outcome variables:

Data Visualizations are an important way to explore the dataset and understand the distributions of different variables and detect the existence of outliers. We created a Final Variables dataset to retain focus on the Predictor and Outcome variable exploration at this stage.

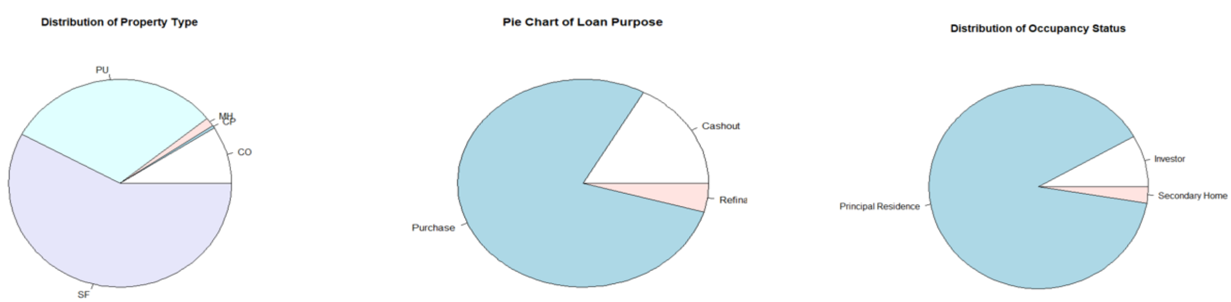
We have 1 continuous Outcome variable name Original Interest Rate in the dataset and 7 discrete numerical variables. There are 3 categorical variables and 2 nominal variables. We used Histograms to visualize the distributions of numerical variables, Pie charts for the Categorical variables because they were few categories. We used Barplots for the nominal variables as they were large in number and better represented using the Bar plots than Pie charts.

Images of Histograms of Numerical variables generated using R code:

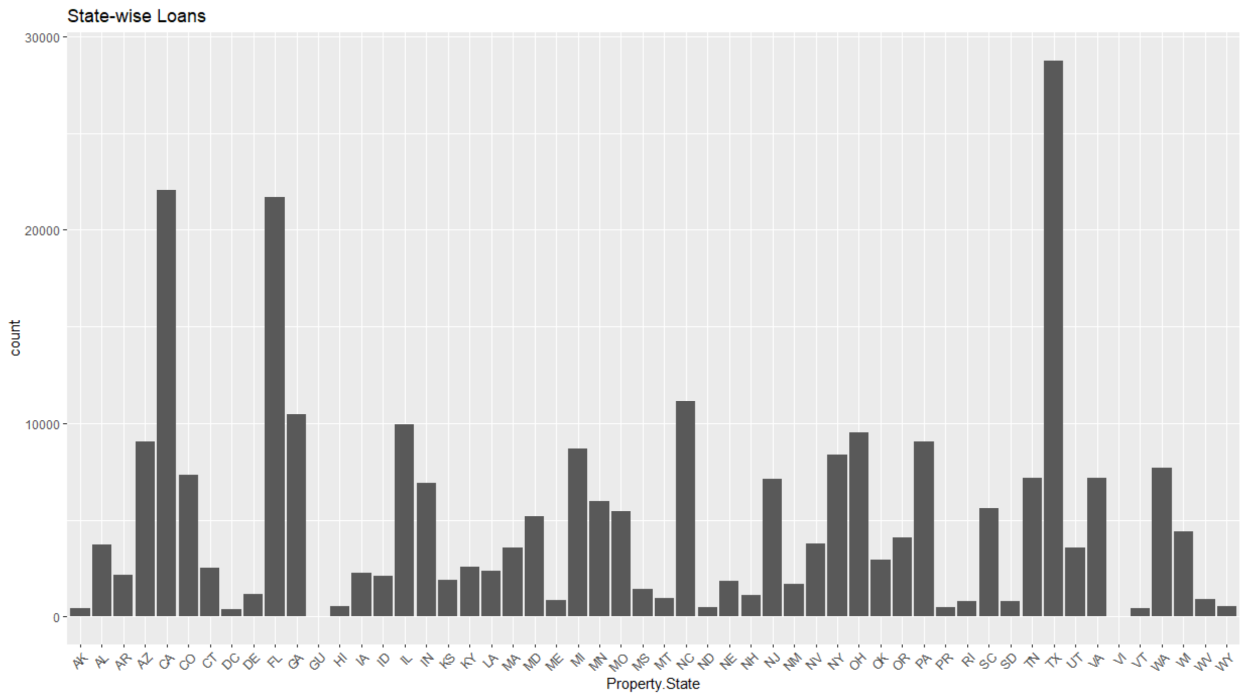


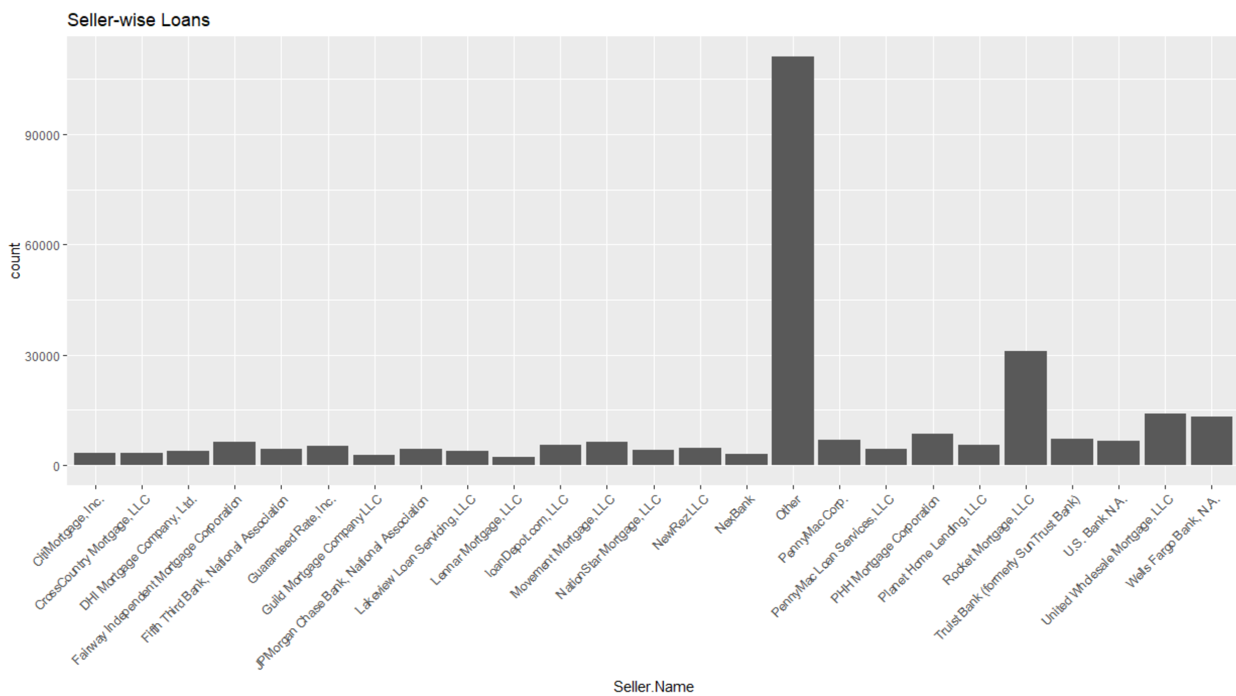
As revealed in summary data, there are outliers in terms of extremely low values for Original Interest rate, Original UPB, Original LTV, Original CLTV, Original DTI and Orginal Credit score. There are outliers in max values in UPB, LTV and CLTV values.

Distributions of Categorical variables was generated using R code as Pie charts:



Distributions of Nominal Variables using Bar Plots:





Source/ Citations :

We have directly and indirectly used materials from the following sources in our Report.

1. Fannie Mae Wikipedia site
2. https://youtu.be/iKSvAsm3ago?si=TQL7e7AI03ERo1hw
3. https://capitalmarkets.fanniemae.com/tools-applications/data-dynamics
4. https://community.rstudio.com/
5. https://www.r-project.org/other-docs.html
6. Mitusch, K., & Nautz, D. (1995). Expectations and Interest Rates on Mortgage Loans. Empirical Economics, 20, 667-680.
7. Page, A. N. (1964). The Variation of Mortgage Interest Rates. The Journal of Business, 37(3), 280-294.
8. Sealand, J. C.(2018). Short-term Prediction of Mortgage Default using Ensembled Machine Learning Models: Summary.
9. Domingos, P. (2012). A Few Useful Things to Know About Machine Learning: Summary.
10. In addition to the above, we used code from our ***Class Presentation slides and lectures*** as well as the presentation from the workshop on Introduction to R ggplot2 Workshop https://umd.box.com/v/IntroRggplot2 conducted by Ms.Yishan Ding of UMD libraries

Contributions

All the 3 team members contributed equally in

1. Researching for the datasets, exploring different datasets and writing code
2. Preparation of the report and code using Google Doc
3. Preparation of the slides using Google slides
4. Recording individual videos and merging them. Upload to Youtube was done in a team meeting.