ATENEO DE MANILA UNIVERSITY

LOYOLA SCHOOLS

School of Science and Engineering

Department of Information Systems and Computer Science

Second Semester

2024-2025

ISCS 30.19

Big Data Analysis

Group Project 1

Written by:

Go, Jared

Damalerio, Adrian Lance

Tan, Tristan

A.Y. 2024-2025

Table of Contents

Executive Summary

This study examines the key factors influencing the interest rate pricing of newly originated mortgages in the United States, with the goal of supporting fair and transparent risk-adjusted pricing for both borrowers and lenders. Using the 2023 Freddie Mac Single-Family Loan-Level Dataset, the analysis focused on how borrower characteristics, loan terms, and property details contribute to mortgage pricing outcomes.

Findings indicate that among borrower attributes, credit score is the most influential, showing a weak but consistent negative relationship with interest rates. While debt-to-income (DTI) and loan-to-value (LTV) ratios had minimal standalone effects, interaction models revealed that combinations of these attributes, like a high LTV and a low credit score, can meaningfully impact pricing. Loan term also plays a role; longer loan durations are associated with slightly higher interest rates due to increased lender risk.

Importantly, property-related factors emerged as stronger predictors than borrower attributes in many cases. Occupancy status had the largest impact, with owner-occupied homes receiving lower interest rates compared to investment properties and second homes. Property type also influenced rates, with manufactured housing consistently priced higher, reflecting perceived risk. Furthermore, geographic location significantly affected pricing, with clear regional differences in average rates.

A predictive model using multiple regression explained 27% of the variation in interest rates, with occupancy status, property type, and property state among the top contributors. Compared to machine learning alternatives like Random Forest, linear regression offered better performance and greater interpretability. However, residual analysis revealed some pricing discrepancies, suggesting unmodeled factors such as lender discretion or special financing programs may also play a role.

In conclusion, while borrower creditworthiness remains relevant, property characteristics and loan structure have a greater influence on mortgage pricing. A comprehensive, data-informed approach that considers all three dimensions, borrower, loan, and property, is essential to achieving equitable and transparent pricing in the U.S. housing market.

Introduction

The housing market plays a fundamental role in the U.S. economy. As one of the largest asset classes, it influences household wealth, consumer spending, and financial stability. At the core of this market lies the mortgage. The mortgage is an essential financial product that enables homeownership for millions of Americans. For both individual borrowers and institutional lenders, the terms of a mortgage, particularly the interest rate, carry significant financial implications.

Mortgage interest rates determine the long-term cost of borrowing and directly impact a borrower's ability to purchase or refinance a home. Even slight differences in rates can translate to substantial variations in monthly payments and total interest paid over the life of a loan. For lenders and investors, interest rates reflect the expected return and risk associated with each mortgage. As a result, setting mortgage rates involves balancing borrower risk, loan characteristics, and property details within a framework that ensures both profitability and fairness.

Understanding the factors that influence interest rate pricing is vital in promoting transparency, preventing discriminatory lending practices, and improving credit access. As the housing finance system continues to evolve, driven by changes in regulation, technology, and consumer behavior, a data-informed approach to analyzing interest rate determinants becomes increasingly valuable.

Research Objective and Guiding Questions

This paper investigates the factors that shape the interest rate pricing of newly originated mortgages in the United States. In particular, it explores how borrower risk attributes, loan terms, and property characteristics influence the rates borrowers receive. By analyzing a comprehensive mortgage dataset, the study seeks to uncover patterns and relationships that can inform fair, risk-adjusted pricing models.

To guide this analysis, the study is structured around the following key problem:

"Which among borrower risk attributes, loan terms, and property types heavily influence the interest rate pricing of a newly originated mortgage in the U.S., ensuring fair and transparent risk-adjusted pricing for both borrowers and lenders?"

This central question is explored through thematic areas shown below:

Influence of Borrower Risk Attributes on Interest Rate Pricing

- How do key borrower risk attributes correlate with the original interest rate?
- Is there a non-linear relationship between borrower risk attributes and interest rates?
- Do risk attributes interact with each other to affect pricing?

Impact of Loan Terms on Interest Rates

- How do different loan terms influence interest rates?
- Are there differences in pricing between fixed-rate and adjustable-rate mortgages for borrowers with similar risk profiles?

Role of Property Type in Interest Rate Determination

- Do property types significantly affect interest rates?
- How do occupancy status and property location influence pricing?
- Are higher interest rates observed in specific property segments due to inherent risk factors?

Predictive Modeling for Interest Rate Pricing

- Which combination of borrower attributes, loan terms, and property characteristics most accurately predicts interest rates?
- Can a machine learning model improve interest rate predictions?
- How do actual interest rates compare to predicted fair rates, and where are the largest discrepancies?

About the Dataset

The dataset used for this project is the Freddie Mac Single-Family Loan-Level Dataset. This is part of Freddie Mac's effort to increase transparency and support a deeper understanding of the U.S. mortgage market. The full dataset includes both origination and monthly performance information for millions of loans purchased or guaranteed by Freddie Mac. However, in this project, we will focus exclusively on the origination data component from 2023, which provides a snapshot of loan and borrower characteristics at the time each mortgage was originated.

Data Dictionary

The dataset contains 931,173 instances of newly originated loans. To add, the dataset contains 32 features describing borrower and loan attributes at origination. The complete data dictionary of all 32 features can be found below.

Feature Name	Data Type	Expected Values	Description
Credit Score	Numeric	300 – 850 (Integer) Scores < 300 or > 850 are considered as “Not Available”	A numerical representation of a borrower's creditworthiness, typically ranging from 300 (poor) to 850 (excellent). Higher values indicate lower credit risk.
First Payment Date	Numeric	6-digit integer in YYYYMM format (e.g. 202305 = May 2023)	The date of the first scheduled mortgage payment due under the terms of the mortgage note.
First Time Homebuyer Flag	Text	Y = Yes N = No 9 = Not Available or Not Applicable	Indicates whether the Borrower, or one of a group of Borrowers, is an individual who (1) is purchasing the mortgaged property, (2) will reside in the mortgaged property as a primary residence, and (3) had no ownership interest (sole or joint) in a residential property during the three-year period preceding the date of the purchase of the mortgaged property.
Maturity Date	Numeric	6-digit integer in YYYYMM format (e.g. 202305 = May 2023)	The date in which the final monthly payment on the mortgage is scheduled to be made as stated on the original mortgage note
Metropolitan Statistical Area (MSA) Or Metropolitan	Numeric	Metropolitan Division or MSA Code E.g. 10000, 49999, etc.	This disclosure will be based on the designation of the Metropolitan Statistical Area or Metropolitan

Division		NaN indicates that the area in which the mortgaged property is located is neither an MSA not a Metropolitan Division or unknown	Division generally as of the loan funding date.
Mortgage Insurance Percentage (MI %)	Numeric	1% - 55% 0 = No MI 999 = Not Available	The percentage of loss coverage on the loan, at the time of Freddie Mac's purchase of the mortgage loan that a mortgage insurer is providing to cover losses incurred as a result of a default on the loan.
Number of Units	Numeric	1 = one-unit 2 = two-unit 3 = three-unit 4 = four-unit 99 = Not Available	Denotes whether the mortgage is a one-, two-, three-, or four-unit property.
Occupancy Status	Text	P = Primary Residence I = Investment Property S = Second Home 9 = Not Available	Denotes whether the mortgage type is owner occupied, second home, or investment property.
Original Combined Loan-to-Value (CLTV)	Numeric	1% - 998% 999 = Not Available	The ratio (expressed as a percentage) of the total amount of all loans secured by the property to the property's appraised value at origination. Includes both the first mortgage and any subordinate liens.
Original Debt-to-Income (DTI) Ratio	Numeric	0% < DTI <= 65% 999 = Not Available	The percentage of a borrower's gross monthly income that goes toward paying monthly debt obligations at the time of loan origination. A higher DTI indicates a greater portion of income is used to cover debts.
Original Unpaid Principal Balance (UPB)	Numeric	Positive numeric values (e.g., 50,000 – 1,000,000+)	The UPB of the mortgage on the note date.

		*rounded to the nearest $1,000	
Original Loan-to-Value (LTV)	Numeric	1% - 998% 999 = Not Available	The ratio of the original loan amount to the property's appraised value at the time of loan origination. Expressed as a percentage. Higher LTV indicates lower borrower equity and potentially higher credit risk.
Original Interest Rate	Numeric	Numeric literal decimal (6,3)% E.g. 2.000, 15.156, etc.	The interest rate of the loan as stated on the note at the time the loan was originated.
Channel	Text	R = Retail B = Broker C = Correspondent T = TPO Not Specified 9 = Not Available	Indicates the origination channel through which the mortgage was obtained. Helps identify the source or intermediary involved in loan origination.
Prepayment Penalty Mortgage (PPM) Flag	Text	Y = PPM N = Not PPM	Denotes whether the mortgage is a PPM.
Amortization Type	Text	FRM – Fixed Rate Mortgage ARM – Adjustable Rate Mortgage	Denotes that the product is a fixed-rate mortgage or adjustable-rate mortgage.
Property State	Text	AL, TX, VA, etc.	A two-letter abbreviation indicating the state or territory within which the property securing the mortgage is located.
Property Type	Text	CO = Condo PU = PUD MH = Manufactured Housing SF = Single-Family CP = Co-op 99 = Not Available	Denotes whether the property type secured by the mortgage is a condominium, leasehold, planned unit development (PUD), cooperative share, manufactured home, or Single-Family home.
Postal Code	Numeric	###00 - where “###”	The postal code for the

		represents the first three digits of the 5 digit postal code 00 = Unknown	location of the mortgaged property
Loan Sequence Number	Text	PYYQnXXXXXXX **P**roduct: F = FRM A = ARM; YYQn = origination year and quarter; XXXXXXX = randomly assigned digits	Unique identifier assigned to each loan.
Loan Purpose	Text	P = Purchase C = Refinance - Cash Out N = Refinance - No Cash Out R = Refinance - Not Specified 9 =Not Available	Indicates whether the mortgage loan is a Cash-out Refinance mortgage, No Cash-out Refinance mortgage, or a Purchase mortgage.
Original Loan Term	Numeric	Integer values (e.g., 120, 180, 240, 360)	The total number of months scheduled for full repayment of the loan at origination. Common terms are 15 years (180 months) and 30 years (360 months), though other terms like 10, 20, or 25 years may appear.
Number of Borrowers	Numeric	Integer values (e.g., 1, 2, 3, etc.) 99 = Not Available	The number of Borrower(s) who are obligated to repay the mortgage note secured by the mortgaged property
Seller Name	Text	Name of the seller e.g. “JPMORGAN CHASE BANK”, “NATIONAL ASSOCIATION”, or “Other Sellers”	The entity acting in its capacity as a seller of mortgages to Freddie Mac at the time of acquisition.
Servicer Name	Text	Name of the servicer e.g. “WELLS FARGO	The entity acting in its capacity as the servicer of

		BANK", "NATIONAL ASSOCIATION", or "Other Servicers"	mortgages to Freddie Mac as of the last period for which loan activity is reported in the Dataset.
Super Conforming Flag	Text	Y = Yes NaN = Not Super Conforming	For mortgages that exceed conforming loan limits with origination dates on or after 10/1/2008 and were delivered to Freddie Mac on or after 1/1/2009
Pre-HARP Loan Sequence Number	Text	PYYQnXXXXXXX • Product: - F = FRM - A = ARM; YYQn = origination year and quarter; XXXXXXX = randomly assigned digits	The Loan Sequence Number link that associates a Relief Refinance loan to the Loan Sequence Number assigned to the loan from which it was refinanced within in the Single-Family Loan-Level Dataset.
Program Indicator	Text	H = Home Possible F = HFA Advantage R= Refi Possible 9 = Not Available or Not Applicable	The indicator that identifies if a loan participates in and of the Freddie Mac programs listed in the valid values.
HARP Indicator	Text	Y = Relief Refinance NaN = Non-Relief Refinance loan	Indicator that identifies whether the loan is part of Freddie Mac's Relief Refinance Program.
Property Valuation Method	Numeric	1 = ACE Loans 2 = Full Appraisal 3 = Other Appraisals (Desktop, driveby, external, AVM) 4 = ACE + PDR 9 = Not Available	The indicator denoting which method was used to obtain a property appraisal, if any.
Interest Only (I/O) Indicator	Text	Y = Yes N = No	The indicator denoting whether the loan only requires interest payments for a specified period beginning with the first payment date.

Mortgage Insurance Cancellation Indicator	Text	Y = Canceled N = Not Canceled 7 = Not Applicable 9 = Not Disclosed	The indicator denoting if the mortgage insurance has been reported as cancelled after the time of Freddie Mac's purchase of the mortgage loan.

Data Preprocessing

Before conducting any meaningful analysis, the raw mortgage dataset must be prepared for modeling and analysis. This involves integrating, selecting, cleaning, transforming, and engineering the data to ensure consistency and usability. All code moving forward for data preprocessing, analysis, findings, etc. can be found in this Google Colab notebook: [ISCS30.19] Project 1.ipynb .

Data Integration

The original dataset was provided in four separate .txt files representing mortgage originations for each quarter of 2023. These files were combined into a single dataset to enable comprehensive year-round analysis. The integration process involved reading each file using consistent delimiters, concatenating the data, and assigning appropriate column names based on the dataset's documentation.

Code

```
#Combining all 4 quarters into one dataframe
csv_files = ['historical_data2023\historical_data_2023Q1.txt',
'historical_data2023\historical_data_2023Q2.txt',
'historical_data2023\historical_data_2023Q3.txt',
'historical_data2023\historical_data_2023Q4.txt']

df_list = [pd.read_csv(file, sep = '|', header = None) for file in
csv_files]

merged_df = pd.concat(df_list, ignore_index=True)

merged_df.columns = ['Credit Score', 'First Payment Date', 'First Time
Homebuyer Flag', 'Maturity Date', 'MSA', 'MI %', 'Units', 'Occupancy
Status', 'CLTV', 'DTI', 'Original UPB', 'Original LTV', 'Original
Interest Rate', 'Channel', 'Prepayment Penalty Mortgage Flag',
'Amortization Type', 'Property State', 'Property Type', 'Postal Code',
```

```
'Loan Sequence Number', 'Loan Purpose', 'Original Loan Term', 'Number
of Borrowers', 'Seller Name', 'Servicer Name', 'Super Conforming Flag',
'Pre-HARP Loan Sequence Number', 'Program Indicator', 'HARP Indicator',
'Property Valuation Method', 'Interest Only Indicator', 'Mortgage
Insurance Cancellation Indicator']

merged_df.to_csv('historical_data_2023.csv', index = False)
```

Output

A .csv file titled 'historical_data_2023.csv'.

Feature Selection

Given the scope of the study and the research questions guiding the analysis, not all features in the dataset are directly relevant to understanding interest rate pricing. Therefore, a subset of features was selected based on domain knowledge, prior literature, and their potential influence on mortgage pricing mechanisms. Irrelevant or redundant features, such as administrative identifiers, post-origination loan performance indicators, or program-specific flags, were excluded to maintain focus on attributes observable at the time of origination.

So, the remaining features to be used in this study can be classified into the following categories:

Borrower Attributes

- Credit Score
- First Time Homebuyer Flag
- DTI
- Original LTV
- Number of Borrowers

Loan Terms and Structure

- Original Interest Rate
- Original Loan Term
- Amortization Type
- Interest Only Indicator
- Loan Purpose
- Channel

Property and Occupancy

- Property Type
- Property State
- Units
- Occupancy Status

Code

```
#Loading the dataset
raw_df = pd.read_csv('historical_data_2023.csv')

#Feature Selection
columns_to_keep = ['Credit Score', 'First Time Homebuyer Flag',
'DTI','Original LTV', 'Number of Borrowers', 'Original Interest Rate',
'Original Loan Term', 'Amortization Type', 'Interest Only Indicator',
'Loan Purpose', 'Channel', 'Property Type', 'Property State', 'Units',
'Occupancy Status']

df = raw_df[columns_to_keep]
```

Data Cleaning

Following integration, the dataset was examined for inconsistencies, formatting errors, and missing values. Duplicate rows were removed, and the structure of key variables was validated against expected formats. A check for missing values across all features revealed no null entries, confirming the completeness of the dataset and eliminating the need for imputation or row removal. The dataset was also reviewed for features with no variability. Columns where all rows contained the same value were identified, as these offer no informational gain for analysis or modeling. This led to the removal of the features "Amortization Type" and "Interest Only Indicator."

Code

```
#Checking Categorical Variables for Unexpected Values
print('First Time Homebuyer Flag: ',df['First Time Homebuyer
Flag'].unique())
print('Number of Borrowers: ',df['Number of Borrowers'].unique())
print('Amortization Type: ',df['Amortization Type'].unique())
print('Interest Only Indicator: ',df['Interest Only
Indicator'].unique())
print('Loan Purpose: ',df['Loan Purpose'].unique())
print('Channel: ',df['Channel'].unique())
```

```
print('Property Type: ',df['Property Type'].unique())
print('Property State: ',df['Property State'].unique())
print('Units: ',df['Units'].unique())
print('Occupancy Status: ',df['Occupancy Status'].unique())

#Checking Continuous Variables for Unexpected Values
print(df['Credit Score'].describe())
print(df['DTI'].describe())
print(df['Original LTV'].describe())
print(df['Original Interest Rate'].describe())
print(df['Original Loan Term'].describe())

#Check for Duplicates (Basing it off the raw dataset as it contains the
unique loan sequence number)
print('Duplicates: ', raw_df.duplicated().sum()) #No duplicates

#Check for Columns with only one unique value
for col in df.columns:
    if df[col].nunique() == 1:
        print(f"{col} has only one unique value:
{df[col].unique()[0]}")
        #Amortization Type only has value 'FRM'
        #Interest Only Indicator only has value 'N'

#Dropping the columns with only one unique value
df = df.drop(columns=['Amortization Type', 'Interest Only Indicator'])
```

Output

```
First Time Homebuyer Flag:  ['N' 'Y']
Number of Borrowers:  [1 2 3 4 5]
Amortization Type:  ['FRM']
Interest Only Indicator:  ['N']
Loan Purpose:  ['P' 'C' 'N']
Channel:  ['C' 'R' 'B' '9']
Property Type:  ['SF' 'PU' 'MH' 'CO' 'CP']
Property State:  ['FL' 'KY' 'CO' 'MO' 'MD' 'MN' 'MI' 'TX' 'IL' 'IN'
'CA' 'UT' 'KS' 'VT' 'NY' 'NH' 'MA' 'HI' 'IA' 'PA' 'OR' 'TN' 'VA' 'DE'
'WI' 'OK' 'SC' 'GA' 'ME' 'OH' 'WA' 'MT' 'NC' 'LA' 'NE' 'NJ' 'SD' 'CT'
'ID' 'AR' 'AZ' 'RI' 'AL' 'NV' 'NM' 'ND' 'WV' 'WY' 'DC' 'MS' 'AK' 'PR'
'VI' 'GU']
Units:  [1 2 4 3]
Occupancy Status:  ['I' 'P' 'S']
```

```
count    931731.000000
mean        752.664845
std         169.025673
min         300.000000
25%         720.000000
50%         758.000000
75%         786.000000
max        9999.000000
Name: Credit Score, dtype: float64
count    931731.000000
mean         38.165929
std          11.270156
min           1.000000
25%          32.000000
50%          40.000000
75%          46.000000
max         999.000000
Name: DTI, dtype: float64
count    931731.000000
mean         75.686560
std          19.329041
min           1.000000
25%          67.000000
50%          80.000000
75%          92.000000
max          97.000000
Name: Original LTV, dtype: float64
count    931731.000000
mean          6.728216
std           0.691656
min           2.250000
25%           6.250000
50%           6.750000
75%           7.250000
max           9.750000
Name: Original Interest Rate, dtype: float64
count    931731.000000
mean        349.512795
std          41.389641
```

```
min           85.000000
25%          360.000000
50%          360.000000
75%          360.000000
max          366.000000
Name: Original Loan Term, dtype: float64
Duplicates:  0
Amortization Type has only one unique value: FRM
Interest Only Indicator has only one unique value: N
```

Data Transformation

As part of the transformation process, features were reviewed for placeholder values that represent missing or invalid data. The values 9, 999, and 9999 were identified as placeholders for null or not available data. These instances were replaced with NaN to accurately reflect missing data and ensure proper handling during statistical analysis and modeling. This transformation step is essential to avoid misinterpretation of placeholder values as legitimate inputs. To add, binary categorical features were transformed to facilitate numerical analysis. Values such as 'Y' and 'N' were mapped to 1 and 0, respectively.

Code

```
#Replacing placeholders with NaN and mapping categorical variables to
numerical values
df['First Time Homebuyer Flag'] = df['First Time Homebuyer
Flag'].replace({'Y': 1, 'N': 0})
df['Channel'] = df['Channel'].replace({'9': np.nan})
df['Credit Score'] = df['Credit Score'].replace({'9999': np.nan})
df['DTI'] = df['DTI'].replace({'999': np.nan})
```

Feature Engineering

To enhance model interpretability and capture potential interactions, select features were derived or combined. No synthetic features were added beyond those explicitly available in the dataset, but interaction terms were considered in the modeling stage.

DTI Band

The DTI ratio reflects a borrower's ability to manage monthly debt payments relative to their income. Lenders consider it a primary measure of repayment capacity. Binning this variable allows us to evaluate how different levels of financial strain are associated with

changes in interest rate pricing. This would allow us to assess whether pricing is sensitive to borrower affordability risk. The DTI bands are based on common industry DTI thresholds used in mortgage underwriting. The common thresholds are the following: 36%, 43%, 50%, >50% (Roberte, 2022).

Code

```
df['DTI_Band'] = pd.cut(df['DTI'], bins=[0, 36, 43, 50, 100],
labels=['Low', 'Moderate', 'High', 'Extreme'])
```

LTV Band

The LTV ratio compares the loan amount to the appraised value of the property. It reflects borrower equity in the home and is one of the most influential factors in mortgage pricing due to its impact on potential lender loss in case of default. A higher LTV means less borrower equity. This would lead to higher risk of loss for the lender. This feature would allow for the analysis of whether interest rates escalate at key equity thresholds, helping determine how much borrower risk influences pricing compared to other attributes. The common thresholds are 80%, 90%, 97%, and 100% (Hayes, 2019).

Code

```
#LTV Band
df['LTV_Band'] = pd.cut(df['Original LTV'], bins=[0, 80, 90, 97, 100],
labels=['Low Risk (≤80%)', 'Moderate (81-90%)', 'Elevated (91-97%)',
'High (98-100%)'])
```

Credit Score Band

Credit score is a direct indicator of a borrower's creditworthiness and history of managing debt. Most lenders use credit bands to assign risk-based pricing tiers, which makes this feature critical in evaluating fair interest rate assignment (Harbour, 2023). By segmenting borrowers into credit score bands, it will allow for the detection of whether interest rates increase proportionally with risk or jump at certain thresholds, thereby addressing both fairness and transparency in risk-adjusted pricing.

Code

```
#Credit Score Band
df['Credit_Band'] = pd.cut(df['Credit Score'], bins=[300, 579, 669,
739, 799, 850], labels=['Poor', 'Fair', 'Good', 'Very Good',
'Excellent'])
```

To add, all band-related features help answer the question whether there is a non-linear relationship between borrower risk attributes and interest rates.

Investment Purchase Flag

The Investment Purchase Flag identifies whether a mortgage was used to purchase an investment property. It is derived by flagging loans where the purpose is "P" (Purchase) and the property is classified as "I" (Investment Property). This feature is important because investment properties generally carry higher credit risk, as borrowers are more likely to default on them compared to primary residences (The Federal Savings Bank, 2024). As a result, lenders may assign higher interest rates or apply stricter underwriting standards to these loans. Including this flag in the analysis allows the isolation of the impact of property use on interest rate pricing, thereby helping assess whether certain property types or occupancy statuses contribute significantly to risk-adjusted pricing.

Code

```
#Investment Purchase Flag
df['Investment_Purchase_Flag'] = ((df['Loan Purpose'] == 'Purchase') &
(df['Occupancy Status'] == 'Investor')).astype(int)
```

Multiple Borrowers Flag

The Multiple Borrowers Flag is a binary feature that indicates whether a mortgage loan has more than one borrower, with 1 representing multiple borrowers and 0 representing a single borrower. This variable is based on the premise that loans with multiple borrowers may pose lower credit risk, as they typically benefit from dual incomes and shared responsibility for repayment. From a lender's perspective, this may reduce the likelihood of default and improve overall loan stability. Incorporating this flag into the analysis allows for the examination on whether borrower structure influences interest rate pricing, and whether lenders offer more favorable terms to multi-borrower households.

Code

```
#Multiple Borrowers Flag
df['Multiple_Borrowers_Flag'] = (df['Number of Borrowers'] >
1).astype(int)
```

Final Features

Here are the final list of features to be used in this study:

Borrower Attributes

- Credit Score
- First Time Homebuyer Flag
- DTI

- Original LTV
- Number of Borrowers
- DTI_Band
- Credit_Band (Credit Score Band)
- LTV_Band
- Multiple_Borrowers_Flag

Loan Terms and Structure

- Original Interest Rate
- Original Loan Term
- Loan Purpose
- Channel

Property and Occupancy

- Property Type
- Property State
- Units
- Occupancy Status
- Investment_Purchase_Flag

Findings

The findings to each subproblem are listed below including the methodology used to come up with the results. These findings will then be used to answer the main question of "Which among borrower risk attributes, loan terms, and property types heavily influence the interest rate pricing of a newly originated mortgage in the U.S., ensuring fair and transparent risk-adjusted pricing for both borrowers and lenders?"

Subquestion 1

"How do key borrower risk attributes correlate with the original interest rate?"

Methodology:

To understand the correlation between borrower risk attributes and the original interest rate, Spearman's rank correlation test was used. The method was used as it is less sensitive to outliers and can detect relationships that may be monotonic. To add, it does not

require the data to be normally distributed. A heatmap was used to visualize the results, showing the corresponding values of each variable in relation to each other. To add, variables with p-values > 0.05 were masked to only show those that have achieved statistical significance. The code for this can be seen below with the SciPy package being used. The results of this can be found in Figure 1.

```
#Subquestion 1
from matplotlib import pyplot as plt
from scipy.stats import spearmanr
import seaborn as sns

cols = ['Original Interest Rate', 'Credit Score', 'DTI', 'Original
LTV', 'Number of Borrowers', 'First Time Homebuyer Flag',
'Multiple_Borrowers_Flag']
data = df[cols].copy()

# Compute Spearman correlation matrix
corr_matrix, p_matrix = spearmanr(data)
corr_df = pd.DataFrame(corr_matrix, index=cols, columns=cols)
pval_df = pd.DataFrame(p_matrix, index=cols, columns=cols)

# Display
print("Spearman Correlation Matrix:")
print(corr_df)

print("\nP-Value Matrix:")
print(pval_df)

plt.figure(figsize=(10, 8))

#Mask p-values>0.05
mask = pval_df > 0.05

# Create the heatmap
sns.heatmap(corr_df, annot=True, fmt=".2f", cmap="coolwarm", vmin=-1,
vmax=1, mask = mask)

# Set titles and labels
plt.title("Spearman Correlation Heatmap")
plt.xticks(rotation=45)
plt.yticks(rotation=0)
plt.tight_layout()
```

```
# Show the plot
plt.show()
```

Results and Interpretation:

The heatmap from the Spearman's correlation analysis can be found in Figure 1.

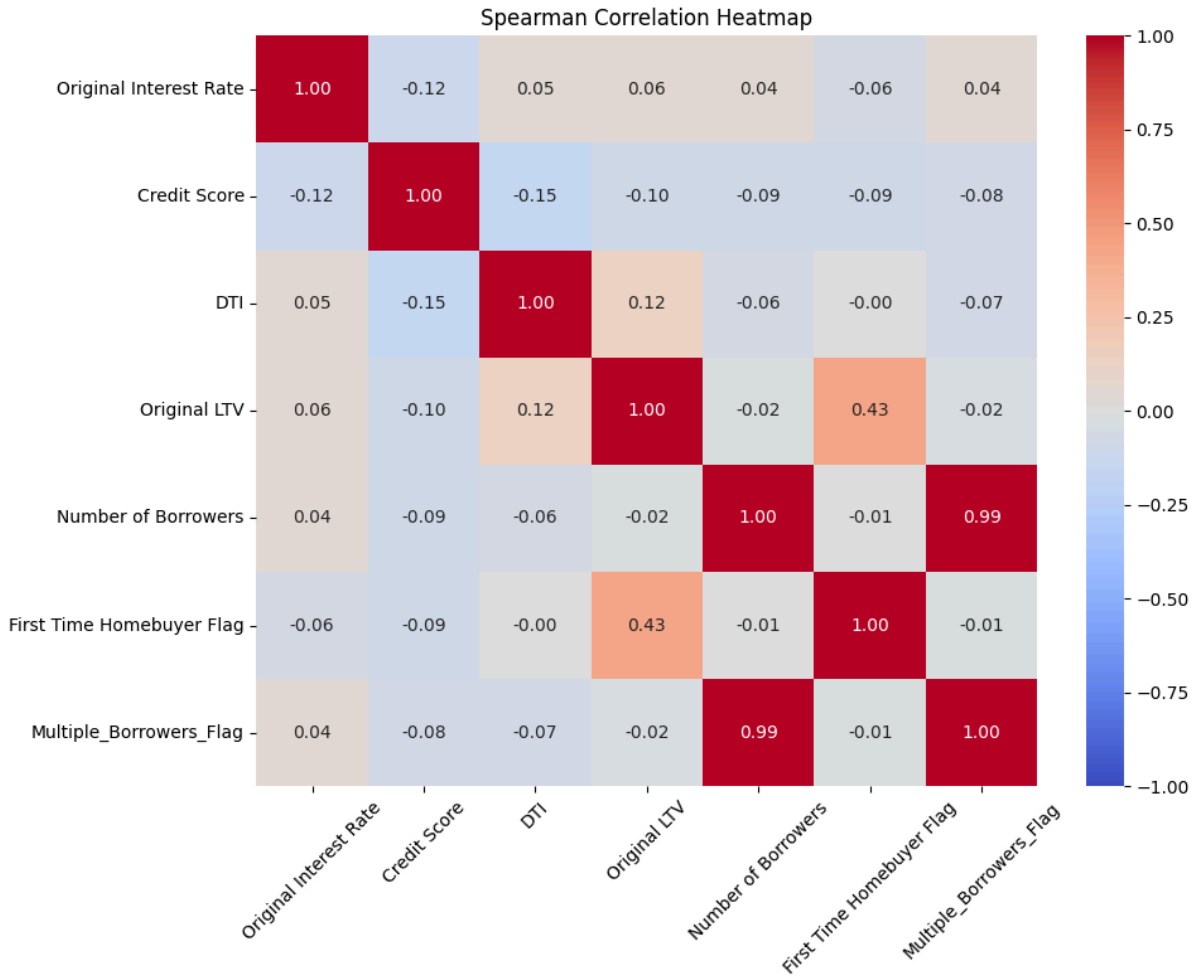


Figure 1. Heatmap of Spearman's Correlation Analysis

Based on the results of the Spearman correlation analysis (Figure 1), Credit Score is the most influential factor in relation to Original Interest Rate. It has a mildly negative correlation of -0.12. This means that a higher Credit Score is weakly yet consistently associated with a lower interest rate. This aligns with typical lending practices where borrowers with higher creditworthiness tend to secure slightly more favorable interest rates. In contrast, other attributes such as Debt-to-Income Ratio (DTI) and Original Loan-to-Value (LTV) show very weak positive correlations, around 0.05 and 0.06, respectively. This indicates that these factors have a minimal monotonic impact. Additionally, variables like the Number of Borrowers, First Time Homebuyer Flag, and Multiple Borrowers Flag exhibit near-zero correlations with the interest rate. This implies that their influence, when

considered solely on a monotonic basis, is negligible. To add, all variables achieved statistical significance, most probably due to a large sample size. However, the practical significance of these correlations is limited given the near 0 correlation of these attributes. The analysis highlights that Credit Score remains the primary risk attribute with a discernible, though modest, relationship to the original interest rate.

Subquestion 2

"Is there a non-linear relationship between borrower risk attributes and interest rates?"

Methodology:

To answer this question, box plots were used to visually explore the categorical risk attributes, Credit_Band, LTV_Band, and DTI_Band. The box plots show any abrupt changes or "jumps" in the data, revealing if the median or spread of interest rates changes non-linearly across different levels.

```
#Subquestion 2
import matplotlib.pyplot as plt
import seaborn as sns

#Copy of the dataframe for plotting
data = df.copy()
risk_attributes = ['Credit_Band', 'LTV_Band', 'DTI_Band', 'First Time
Homebuyer Flag', 'Multiple_Borrowers_Flag']

#Box plot for each risk attribute
for attr in risk_attributes:
    plt.figure(figsize=(8, 6))
    sns.boxplot(x=attr, y='Original Interest Rate', data=data,
palette='viridis')
    plt.title(f'Original Interest Rate Distribution by {attr}')
    plt.xlabel(attr)
    plt.ylabel('Original Interest Rate')
    plt.tight_layout()
    plt.show()
```

Results and Interpretation:

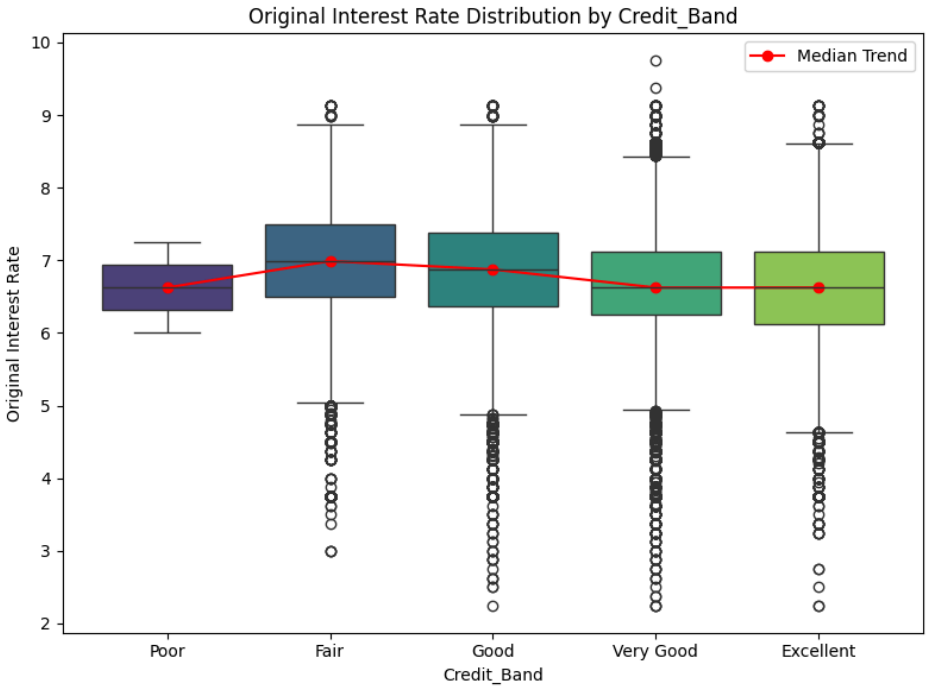


Figure 2. Box Plot of Credit_Band

The results are shown across Figures 2 to 4. In Figure 1, the box plot suggests that each credit band shows a distinct distribution of interest rates rather than a smooth linear progression from “Poor” to “Excellent” credit. “Fair” stands out with a notably higher median interest rate than other bands. “Good,” “Very Good,” and “Excellent” clutter closer together, indicating that borrowers in these categories receive similar rates. To add, although “Poor” has a lower median than “Fair,” it must be pointed out that this band only has 2 instances, which is a very small sample size. Hence, its median and variability estimates are not statistically robust. Moving forward, this band will be ignored for the time being. Ignoring the “Poor” band, median interest rates appear to follow a consistent, nearly linear trend. As credit quality improves from Fair to Good to Very Good to Excellent, the median rates steadily decrease in a predictable manner. Although there is still variability within each category, the overall pattern indicates that the relationship between credit quality and interest rate is approximately linear.

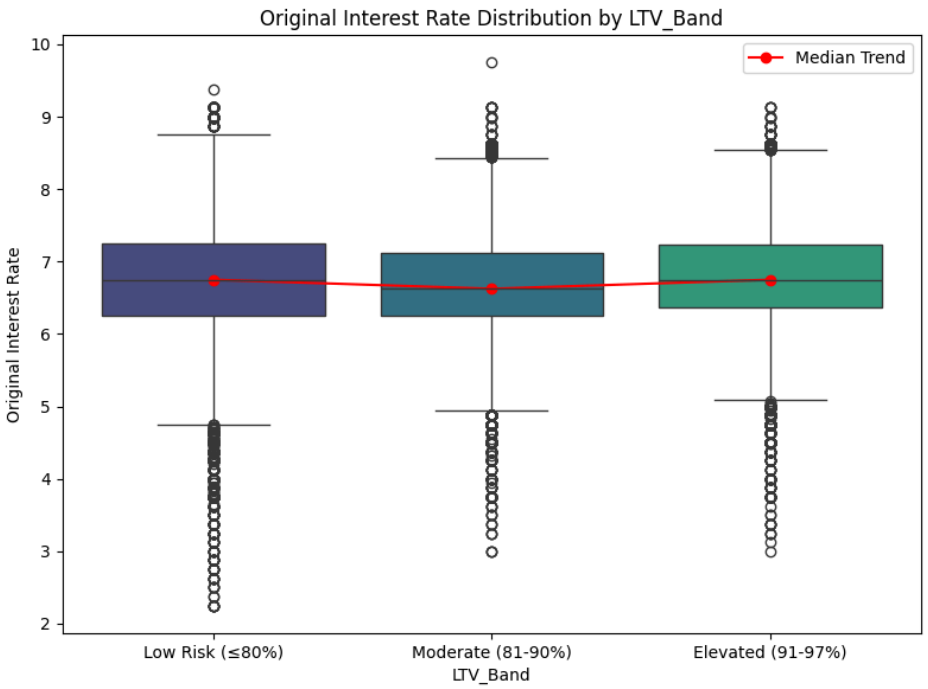


Figure 3. Box Plot of LTV_Band

The results of the box plot for LTV_Band (Figure 3), shows that the median interest rates for the three categories are quite similar, having only minor differences between them. The median line indicates a slight dip from Low Risk to Moderate before rising again in the Elevated category. Additionally, the wide distributions and overlap across the bands suggest that the interest rates are not strongly differentiated by these LTV categories alone. Overall, the plot implies that there is no clear linear relationship between LTV risk and interest rates, indicating that LTV, as grouped here, may not be a dominant factor in determining the interest rate compared to other variables.

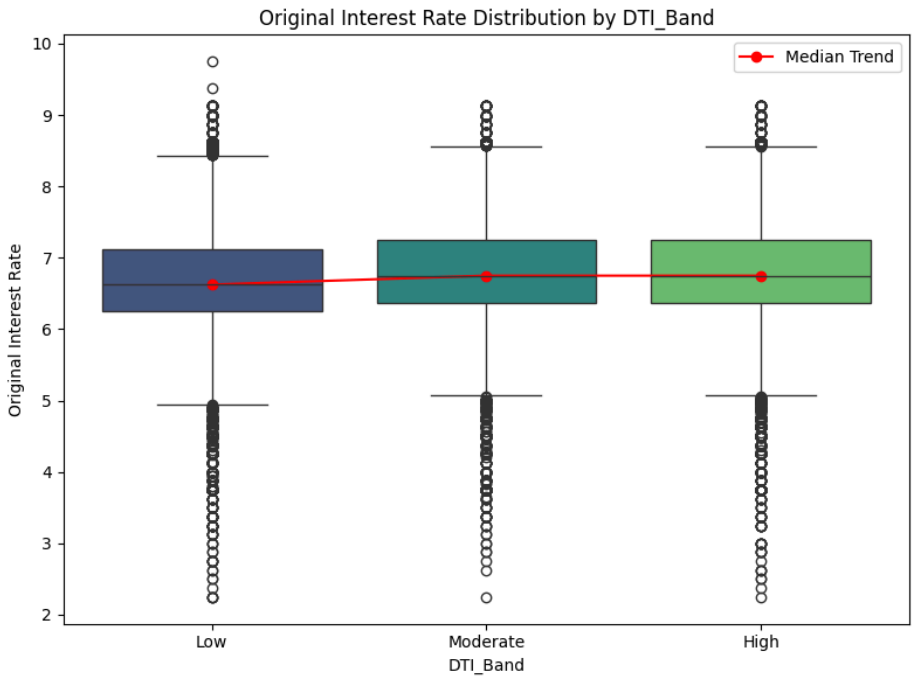


Figure 4. Box Plot of DTI_Band

The results in Figure 4 shows the box plot of the DTI_Band. The results show that median interest rates across the 3 groups do not differ substantially. The median line rises slightly from Low to Moderate, then dips from Moderate to High, but the overall variation is limited. The distributions themselves overlap considerably, indicating that DTI alone may not strongly determine the interest rate. In other words, there is no pronounced linear progression or threshold effect evident across these DTI bands, suggesting that lenders may weigh other factors more heavily than a borrower's debt-to-income category when setting interest rates.

Subquestion 3

"Do risk attributes interact with each other to affect pricing?"

Methodology:

To explore whether risk attributes interact to affect pricing, a regression model with interaction terms was used. Both continuous and categorical attributes were included as interaction terms to provide a formal statistical test of how individual and combined borrower risk attributes relate to interest rates. This allows us to answer the question of whether certain attributes interact with each other to significantly affect pricing.

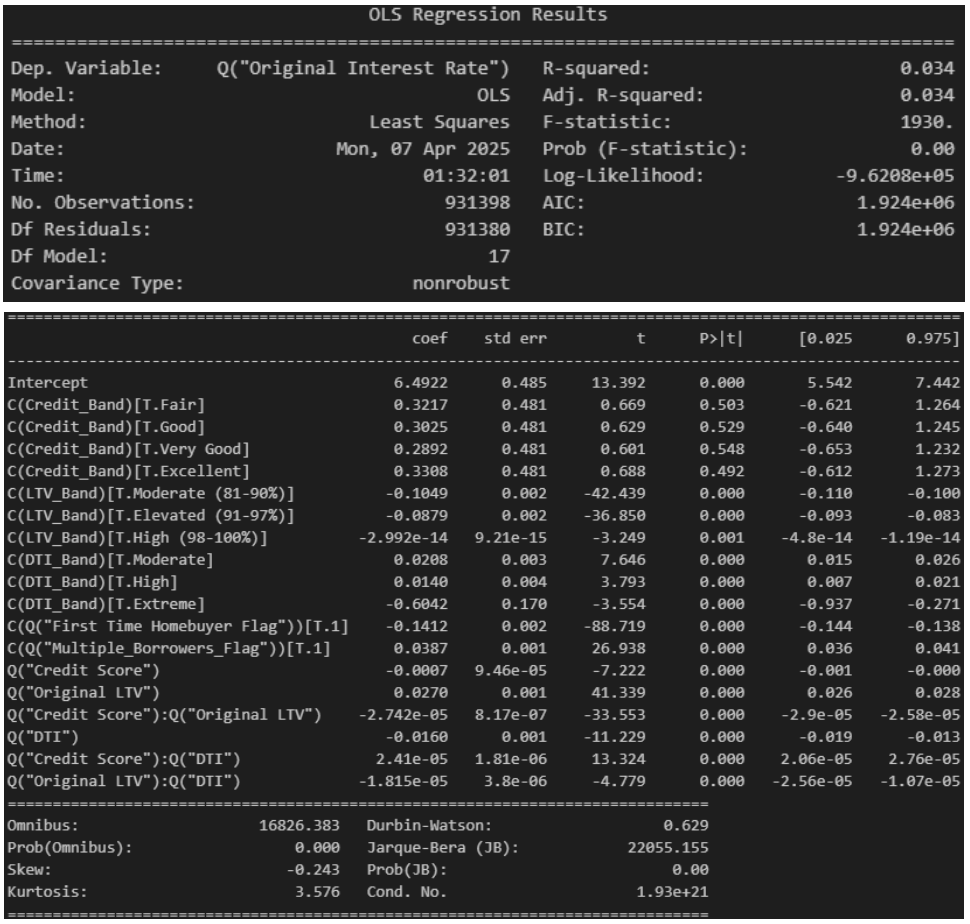
```
#Subquestion 3
import statsmodels.formula.api as smf

data = df.copy()
```

```
formula = (
    'Q("Original Interest Rate") ~ '
    'Q("Credit Score") * Q("Original LTV") + '
    'Q("Credit Score") * Q("DTI") + '
    'Q("Original LTV") * Q("DTI") + '
    'C(Credit_Band) + '
    'C(LTV_Band) + '
    'C(DTI_Band) + '
    'C(Q("First Time Homebuyer Flag")) + '
    'C(Q("Multiple_Borrowers_Flag"))'
)

model = smf.ols(formula, data=data).fit()
print(model.summary())
```

Results and Interpretation:



OLS Regression Results

Dep. Variable:	Q("Original Interest Rate")	R-squared:	0.034
Model:	OLS	Adj. R-squared:	0.034
Method:	Least Squares	F-statistic:	1930.
Date:	Mon, 07 Apr 2025	Prob (F-statistic):	0.00
Time:	01:32:01	Log-Likelihood:	-9.6208e+05
No. Observations:	931398	AIC:	1.924e+06
Df Residuals:	931380	BIC:	1.924e+06
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.4922	0.485	13.392	0.000	5.542	7.442
C(Credit_Band)[T.Fair]	0.3217	0.481	0.669	0.503	-0.621	1.264
C(Credit_Band)[T.Good]	0.3025	0.481	0.629	0.529	-0.640	1.245
C(Credit_Band)[T.Very Good]	0.2892	0.481	0.601	0.548	-0.653	1.232
C(Credit_Band)[T.Excellent]	0.3308	0.481	0.688	0.492	-0.612	1.273
C(LTV_Band)[T.Moderate (81-90%)]	-0.1049	0.002	-42.439	0.000	-0.110	-0.100
C(LTV_Band)[T.Elevated (91-97%)]	-0.0879	0.002	-36.850	0.000	-0.093	-0.083
C(LTV_Band)[T.High (98-100%)]	-2.992e-14	9.21e-15	-3.249	0.001	-4.8e-14	-1.19e-14
C(DTI_Band)[T.Moderate]	0.0208	0.003	7.646	0.000	0.015	0.026
C(DTI_Band)[T.High]	0.0140	0.004	3.793	0.000	0.007	0.021
C(DTI_Band)[T.Extreme]	-0.6042	0.170	-3.554	0.000	-0.937	-0.271
C(Q("First Time Homebuyer Flag"))[T.1]	-0.1412	0.002	-88.719	0.000	-0.144	-0.138
C(Q("Multiple_Borrowers_Flag"))[T.1]	0.0387	0.001	26.938	0.000	0.036	0.041
Q("Credit Score")	-0.0007	9.46e-05	-7.222	0.000	-0.001	-0.000
Q("Original LTV")	0.0270	0.001	41.339	0.000	0.026	0.028
Q("Credit Score"):Q("Original LTV")	-2.742e-05	8.17e-07	-33.553	0.000	-2.9e-05	-2.58e-05
Q("DTI")	-0.0160	0.001	-11.229	0.000	-0.019	-0.013
Q("Credit Score"):Q("DTI")	2.41e-05	1.81e-06	13.324	0.000	2.06e-05	2.76e-05
Q("Original LTV"):Q("DTI")	-1.815e-05	3.8e-06	-4.779	0.000	-2.56e-05	-1.07e-05

Omnibus:	16826.383	Durbin-Watson:	0.629
Prob(Omnibus):	0.000	Jarque-Bera (JB):	22055.155
Skew:	-0.243	Prob(JB):	0.00
Kurtosis:	3.576	Cond. No.	1.93e+21

Figure 5. Regression Results of Interaction of Borrower Risk Attributes

The regression results indicate that risk attributes do interact with one another to affect interest rate pricing. In the model, interaction terms (Credit Score and Original LTV, Credit Score and DTI, and Original LTV and DTI) are statistically significant. The negative

coefficient on the Credit Score × Original LTV interaction (−2.742e-05) suggests that the negative effect of a high LTV on interest rates is less pronounced among borrowers with higher credit scores. This implies that a high credit score can partially offset the risk associated with a high LTV. This can also be seen in the positive interaction between Credit Score and DTI (2.41e-05), which implies that the beneficial effect of a higher credit score is reduced when DTI is high. Finally, the negative Original LTV × DTI interaction (−1.815e-05) indicates that the combination of high LTV and high DTI influences interest rates differently than would be expected if you just added their separate effects.

Although the overall model explains only about 3.4% of the variation in interest rates ($R^2 \approx 0.034$), these significant interactions confirm that the effect of one risk factor is contingent on the levels of others. It shows that risk factors do not operate in isolation; instead, their combined effects can modify interest rate outcomes.

Subquestion 4

"How do different loan terms influence interest rates?"

Methodology:

To assess how different loan terms influence the original interest rate, we employed a two-pronged approach, separating the analysis into numerical-numerical and categorical-numerical relationships, similar to our approach in Subquestion 1. The Original Loan Term is a continuous numerical variable measured in months (i.e. 180, 240, 360). To evaluate its relationship with the Original Interest Rate, we applied Spearman's rank correlation, a non-parametric method well suited for capturing monotonic relationships and handling non-normally distributed data. The correlation matrix was computed between the Original Loan Term and the Original Interest Rate, alongside other key numerical variables for context. A p-value matrix was also generated to filter for statistical significant correlations ($p < 0.05$). A heatmap visualization was produced with insignificant correlations masked out.

```
cols = ['Original Interest Rate', 'Original Loan Term']
data = df[cols].dropna()

corr_matrix, p_matrix = spearmanr(data)
corr_df = pd.DataFrame(corr_matrix, index=cols, columns=cols)
pval_df = pd.DataFrame(p_matrix, index=cols, columns=cols)

mask = pval_df > 0.05
```

```
plt.figure(figsize=(6, 4))
sns.heatmap(corr_df, annot=True, fmt=".2f", cmap="coolwarm", vmin=-1,
vmax=1, mask=mask)
plt.title("Spearman Correlation: Loan Term vs Interest Rate")
plt.tight_layout()
plt.show()
```

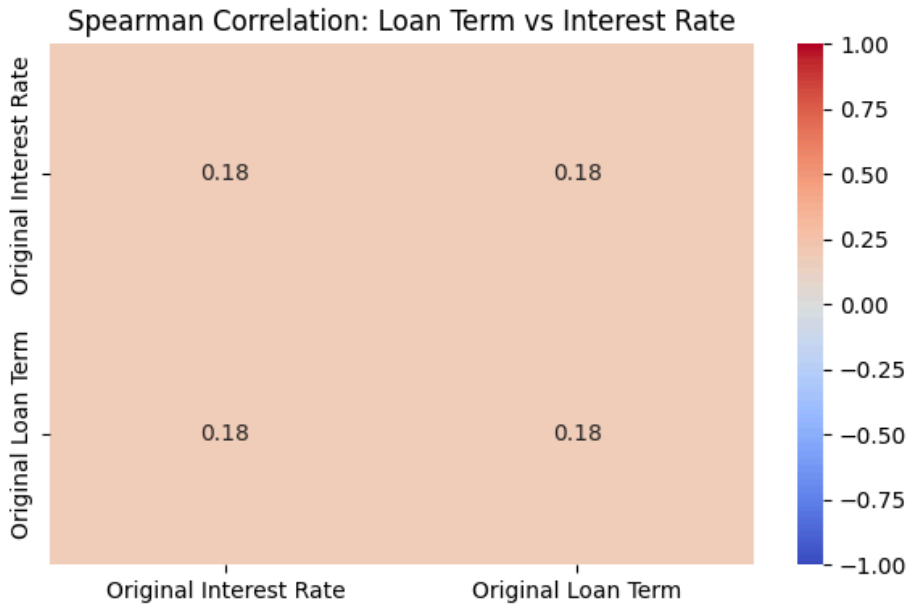
While the Original Loan Term is numerical, it tends to cluster around discrete values such as 180, 240, 300, 360, and 366 months. To evaluate interest rate differences across these common terms, we binned the loan term into categorical bands and then used the Kruskal-Wallis H-test to determine whether there are statistically significant differences in median interest rates between these bands. The Kruskal-Wallis H-test is appropriate here as it compares a continuous outcome (interest rate) across multiple independent groups without assuming a normal distribution.

```
df['Loan_Term_Band'] = pd.cut(df['Original Loan Term'],
                      bins=[0, 180, 240, 300, 360, 370],
                      labels=['≤180 mo', '181-240 mo', '241-300 mo',
'301-360 mo', '361-366 mo'])
feature = 'Loan_Term_Band'
temp_df = df[[feature, 'Original Interest Rate']].dropna()

groups = [temp_df[temp_df[feature] == category]['Original Interest
Rate']
          for category in temp_df[feature].unique()
          if len(temp_df[temp_df[feature] == category]) > 1]

stat, p = kruskal(*groups)
print(f"Kruskal-Wallis Test for Loan Term Bands — H-statistic:
{stat:.4f} | p-value: {p:.4f}")
```

Results and Interpretation:



`Kruskal-Wallis Test for Loan Term Bands — H-statistic: 14552.5786 | p-value: 0.0000`

The figure above indicates that the loan term duration has a meaningful influence on interest rate pricing. The Spearman correlation analysis produced a coefficient of approximately p = +0.18, with a p-value < 0.05, suggesting a weak but statistically significant positive relationship between the Original Loan Term and the Original Interest Rate. This implies that, on average, longer loan terms are slightly associated with higher interest rates, likely due to the extended risk exposure lenders take on over time.

To validate these findings, the Kruskal-Wallis H-test was conducted using binned loan term bands. The test returned a H-statistic of 14,552.5786 and a p-value of 0.000, indicating that the difference in the median interest rate across the loan term groups are highly significant. Thus, confirming that borrowers choosing different loan durations such as 15, 20, or 30 years, are likely to receive noticeably different interest rates, reinforcing the impact of loan term structure in mortgage price decisions.

Subquestion 5

“Are there differences in pricing between fixed-rate and adjustable-rate mortgages for borrowers with similar risk profiles?”

Methodology:

To determine whether fixed-rate mortgages (FRMs) and adjustable-rate mortgages (ARMs) differ in interest rate pricing for borrowers with similar risk profiles, we tried to design a comparative analysis involving both visual and statistical approaches. The process

began by filtering the dataset to include only loans where the Amortization Type was either "FRM" or "ARM". Borrower risk was controlled by applying standard constraints on the following attributes:

- Credit Score (300–850),
- Debt-to-Income (DTI) ratio (< 65),
- Original Loan-to-Value (LTV) ratio (≤ 100),
- Original Loan Term (between 120 and 366 months).

Results and Interpretation:

```
data = df.copy()
data['Amortization Type'].unique()
```

✓ 0.2s

```
array(['FRM'], dtype=object)
```

However, after filtering and inspection, the dataset was found to contain only FRM (Fixed-Rate Mortgage) records and zero ARM loans, making a direct comparison infeasible. Thus, this subquestion cannot be answered using the current data and would need additional data sources containing adjustable rate mortgages to perform a valid comparison.

Subquestion 6

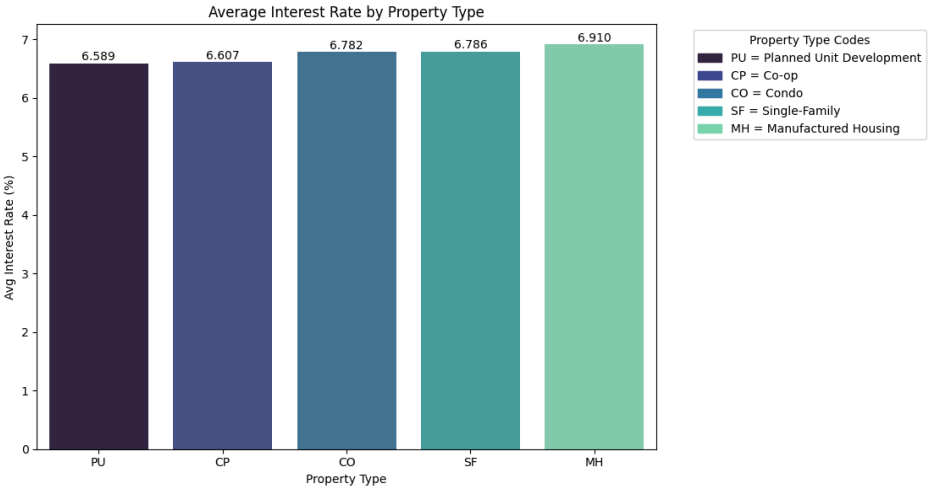
"Do property types significantly affect interest rates?"

Methodology:

To assess how property types affect interest rates, a descriptive analysis was conducted using group-wise aggregation. We compared the average interest rate across different property types.

Results and Interpretation:

```
# Group by property type and compute mean interest rate
state_avg = df.groupby('Property Type')['Original Interest Rate'].mean().sort_values()
```



Upon comparing the average interest rate of different property types, it is observed that Manufactured Housing (MH) has the highest average interest rate at 6.910%, followed by Single-Family (SF) and Condo (CO) property types at 6.786% and 6.782%, respectively. Meanwhile, Co-op (CP) and PUD (PU) properties have lower average rates at 6.607% and 6.589%, respectively.

Subquestion 7

"How do occupancy status and property location influence pricing?"

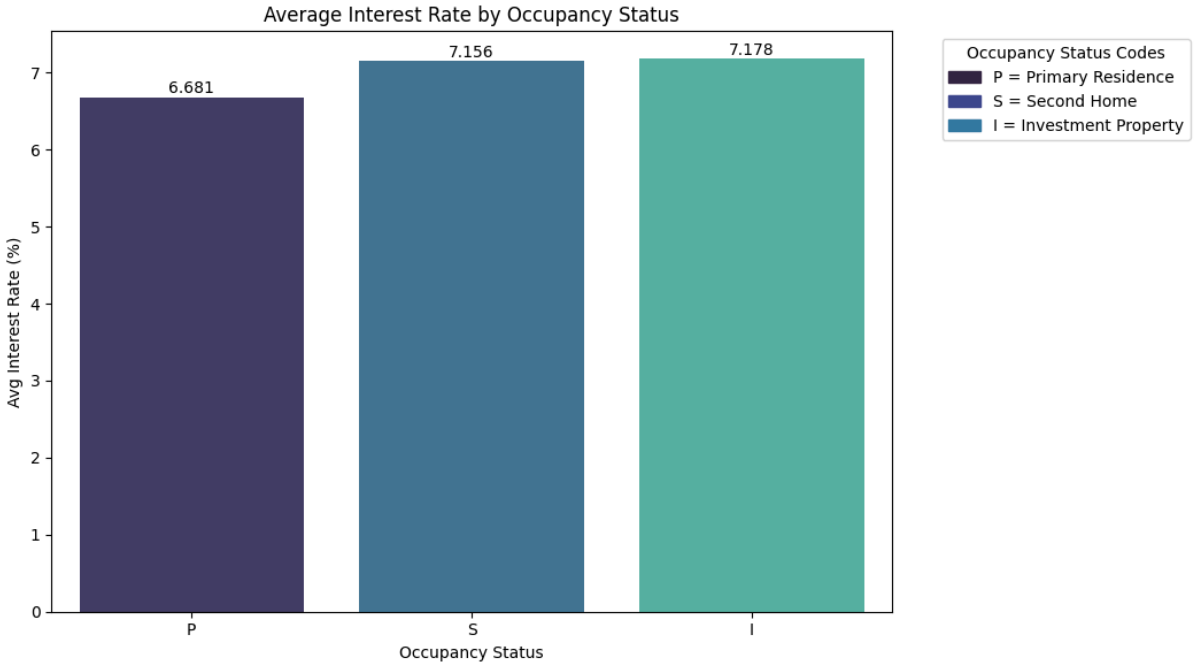
Methodology:

To assess how property location and occupancy status influence mortgage pricing, a descriptive analysis was conducted using group-wise aggregation. We compared the average interest rate across different states and occupancy status.

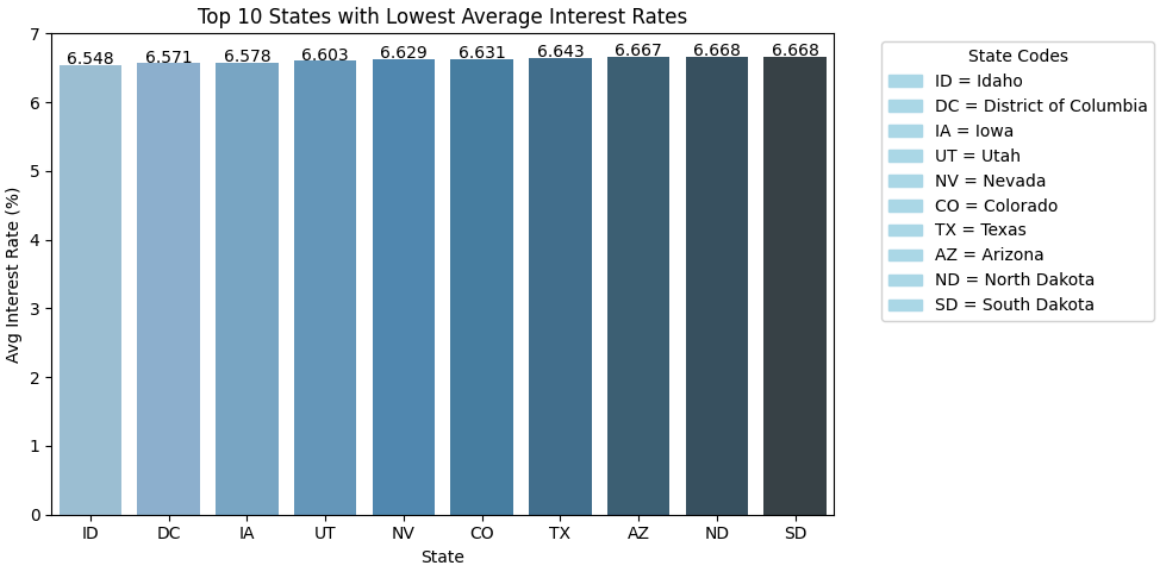
```
# Visualize average interest rates by state
property_state = df.groupby('Property State')['Original Interest Rate'].mean().sort_values()
```

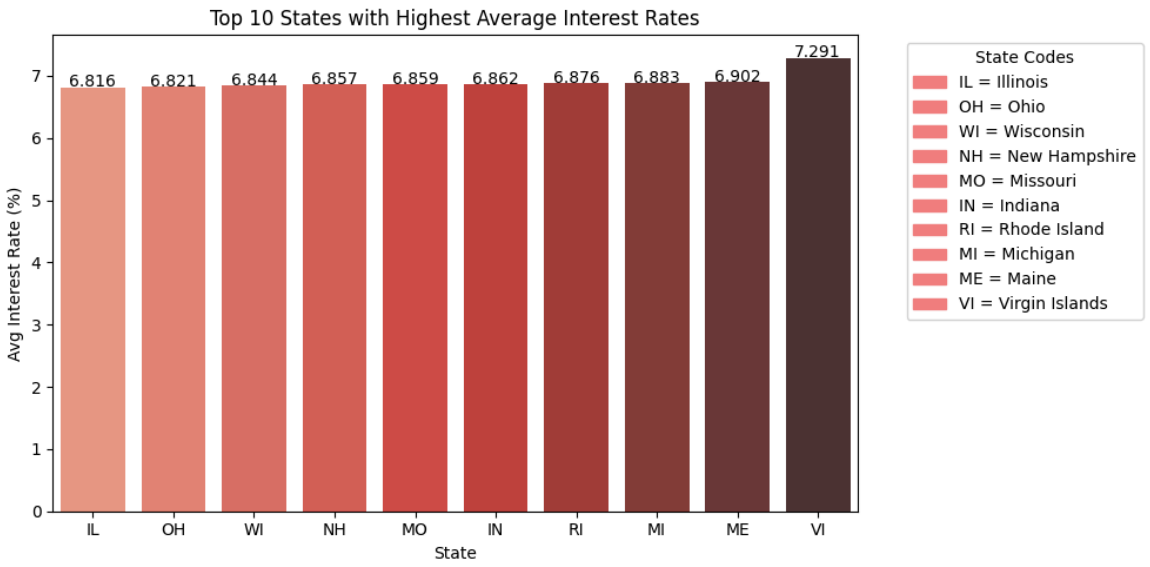
```
# Visualize average interest rates by occupancy status
occupancy_status = df.groupby('Occupancy Status')['Original Interest Rate'].mean().sort_values()
```

Results and Interpretation:



Initial results suggest that both occupancy status and property location have a visible influence on mortgage interest rate pricing. Borrowers with primary residences receive the lowest average interest rates at approximately 6.681%, while those with investment properties and second homes are charged higher rates, averaging 7.178% and 7.156%, respectively. This pattern aligns with standard risk-based pricing practices, where non-owner-occupied properties are considered riskier due to higher default likelihood.





In terms of location, average interest rates vary across states. The bar plots highlight clear geographic differences in average mortgage interest rates across U.S. states. The first graph shows that states such as Idaho (6.548%), District of Columbia (6.571%), and Iowa (6.578%) have the lowest average interest rates in the dataset. These states, many of which are located in the Midwest and Mountain West regions, may benefit from more stable housing markets, lower property values, or increased lending competition. In contrast, the second graph reveals that states like Maine (6.902%), Michigan (6.883%), and especially the Virgin Islands (7.291%) exhibit the highest average rates. These higher-rate regions may reflect lender concerns about local economic volatility, smaller lending markets, or regulatory differences. Overall, these visualizations suggest that property location significantly influences mortgage pricing, with borrowers in certain states consistently paying more or less than others.

Subquestion 8

"Are higher interest rates observed in specific property segments due to inherent risk factors?"

Methodology:

In order to determine if higher interest rates observed in specific property segments are due to inherent risk factors, a multivariate linear regression model was used. This approach allows us to isolate the impact of property characteristics—such as property type and occupancy status—on mortgage pricing while controlling for borrower-level risk

attributes like credit score, loan-to-value (LTV) ratio, and debt-to-income (DTI) ratio, as well as loan purpose.

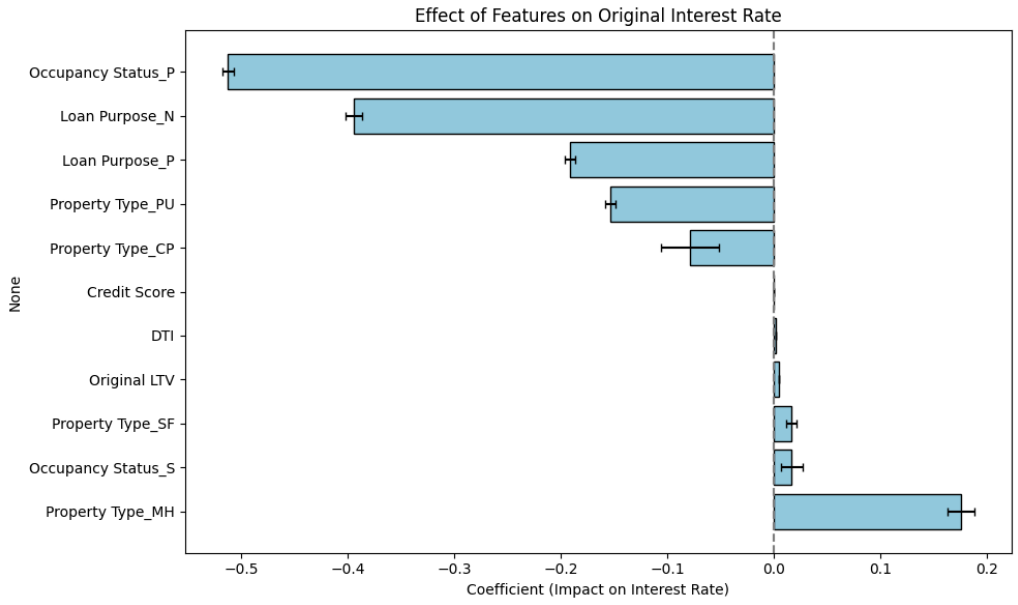
Results and Interpretation:

OLS Regression Results

Dep. Variable:	Original Interest Rate	R-squared:	0.084
Model:	OLS	Adj. R-squared:	0.084
Method:	Least Squares	F-statistic:	7776.
Date:	Sun, 06 Apr 2025	Prob (F-statistic):	0.00
Time:	11:15:41	Log-Likelihood:	-9.3765e+05
No. Observations:	931731	AIC:	1.875e+06
Df Residuals:	931719	BIC:	1.875e+06
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.0358	0.006	1249.834	0.000	7.025	7.047
Credit Score	-7.771e-05	4.07e-06	-19.078	0.000	-8.57e-05	-6.97e-05
Original LTV	0.0046	4.04e-05	112.572	0.000	0.004	0.005
DTI	0.0021	6.16e-05	33.993	0.000	0.002	0.002
Loan Purpose_N	-0.3943	0.004	-97.397	0.000	-0.402	-0.386
Loan Purpose_P	-0.1911	0.002	-80.186	0.000	-0.196	-0.186
Occupancy Status_P	-0.5123	0.003	-193.128	0.000	-0.517	-0.507
Occupancy Status_S	0.0170	0.005	3.180	0.001	0.007	0.027
Property Type_CP	-0.0785	0.014	-5.710	0.000	-0.105	-0.052
Property Type_MH	0.1755	0.006	27.400	0.000	0.163	0.188
Property Type_PU	-0.1535	0.003	-60.932	0.000	-0.158	-0.149
Property Type_SF	0.0167	0.002	7.034	0.000	0.012	0.021

Omnibus:	9394.459	Durbin-Watson:	0.654
Prob(Omnibus):	0.000	Jarque-Bera (JB):	11253.877
Skew:	-0.184	Prob(JB):	0.00
Kurtosis:	3.393	Cond. No.	1.56e+04



The resulting regression coefficients, visualized in the bar chart, show that certain property segments are indeed associated with significantly higher or lower interest rates, even after adjusting for these risk factors. Notably, Manufactured Housing (MH) is linked to a substantial rate premium of approximately +0.175%, suggesting that lenders perceive these

properties as riskier or more complex to underwrite. In contrast, Planned Unit Developments (PU) and Cooperatives (CP) are associated with negative coefficients, indicating they are priced more favorably. These findings confirm that property segments—especially those with unique legal, structural, or occupancy characteristics—carry inherent risk that lenders account for in pricing, independent of borrower creditworthiness. Therefore, higher interest rates in these segments are not solely explained by borrower profiles, but also by the perceived riskiness of the property itself.

Subquestion 9

Which combination of borrower attributes, loan terms, and property characteristics most accurately predicts interest rates?

Methodology:

To identify the most influential factors in predicting original mortgage interest rates, we implemented a multiple linear regression model using a 10,000-row sample from the dataset. This sample size was chosen to allow efficient computation while maintaining a representative data subset. The analysis included a range of variables grouped into three main categories:

- Borrower attributes: Credit Score, DTI, First Time Homebuyer Flag, Number of Borrowers
- Loan terms: Original LTV, Original Loan Term
- Property characteristics: Property Type, Property State, Occupancy Status, Units
    - Other loan info: Loan Purpose, Channel

The target variable was Original Interest Rate, a continuous value. Categorical variables were one-hot encoded, while numerical variables were passed through directly. We used a sklearn pipeline that combined preprocessing and model fitting in a single streamlined workflow.

Model performance was evaluated using:

- R^2 Score (to assess variance explained),
- Root Mean Squared Error (RMSE) (to measure prediction accuracy).

We interpreted feature importance based on the absolute values of the regression coefficients, which indicate how strongly each variable contributes to predicting interest rates.

```
# Load 10,000 rows from cleaned dataset for memory-safe regression
df = pd.read_csv('historical_data_2023_cleaned.csv', nrows=10000)
```

```
# Drop missing values
df = df.dropna()

# Define predictors and target
features = [
    'Credit Score', 'DTI', 'Original LTV', 'Original Loan Term',
    'First Time Homebuyer Flag', 'Number of Borrowers',
    'Loan Purpose', 'Channel', 'Property Type', 'Property State',
    'Occupancy Status', 'Units'
]
target = 'Original Interest Rate'

# Split predictors and target
X = df[features]
y = df[target]

# Define categorical and numerical features
categorical_features = ['Loan Purpose', 'Channel', 'Property Type',
'Property State', 'Occupancy Status']
numerical_features = list(set(features) - set(categorical_features))

# Preprocessing: one-hot encode categorical features
preprocessor = ColumnTransformer(
    transformers=[
        ('num', 'passthrough', numerical_features),
                        ('cat',   OneHotEncoder(handle_unknown='ignore'),
categorical_features)
    ]
)

# Create pipeline: preprocess + linear regression
model = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('regressor', LinearRegression())
])

# Fit the model
model.fit(X, y)

# Predict and evaluate
y_pred = model.predict(X)
r2 = r2_score(y, y_pred)
rmse = np.sqrt(mean_squared_error(y, y_pred))
```

```
# Extract feature importance (coefficients)
regressor = model.named_steps['regressor']
encoded_feature_names = numerical_features + list(

model.named_steps['preprocessor'].transformers_[1][1].get_feature_names
_out(categorical_features)
)
coefficients                =                pd.Series(regressor.coef_,
index=encoded_feature_names).sort_values(key=np.abs, ascending=False)

# Output performance and top 10 predictors
print(f"R² Score: {r2:.3f}")
print(f"RMSE: {rmse:.3f}")
print("\nTop 10 Most Influential Features:")
print(coefficients.head(10))
```

Results and Interpretation:

R² Score: 0.271

RMSE: 0.453

Top 10 Most Influential Features:

Occupancy Status_P -0.382283

Property State_DC -0.289147

Property State_PR 0.246813

Property State_AK 0.240957

Property State_WV 0.237338

Property State_UT -0.219490

Occupancy Status_I 0.214332

Property State_IA -0.181238

Property Type_MH 0.176106

Occupancy Status_S 0.167951

The regression model produced an R² score of 0.271 and an RMSE of 0.453, indicating that approximately 27% of the variation in mortgage interest rates can be explained by the selected borrower, loan, and property features. This suggests a moderate predictive power, which is reasonable given the complexity of mortgage pricing and external market influences.

Among the most influential predictors, occupancy status stood out as the strongest factor. Loans for principal residences were associated with significantly lower interest rates,

while investment properties and second homes had noticeably higher rates, reflecting increased lender risk. Additionally, property location played a substantial role. Mortgages in areas like Washington DC, Utah, and Iowa tended to receive lower rates, while those in Puerto Rico, Alaska, and West Virginia were priced higher. Another key factor was property type, with manufactured homes linked to higher interest rates compared to traditional property types.

Subquestion 10

Can a machine learning model improve interest rate predictions?

Methodology:

The methodology for this subquestion closely mirrors that of Subquestion 9, with both using the same 10,000 row sample from the dataset. As in Subquestion 9, categorical variables were one-hot encoded and numerical features passed through using a Column Transformer. A pipeline was constructed to combine preprocessing and modeling for consistency and code clarity. However, unlike Subquestion 9 — which focused solely on Linear Regression to interpret feature importance — this analysis aimed to compare the predictive performance of multiple models:

1. Linear Regression – serving as the baseline model
2. Random Forest Regressor – a non-linear ensemble model

Each model was trained on 80% of the sample data, with performance evaluated on the remaining 20% (test set) using:

- R^2 Score: to assess how much variance in interest rates the model explains
- Root Mean Squared Error (RMSE): to measure average prediction error

This setup allowed us to determine whether machine learning methods could yield more accurate interest rate predictions than traditional linear modeling.

```
# Define machine learning models
models = {
    'Linear Regression': LinearRegression(),
    'Random Forest': RandomForestRegressor(random_state=42)
}

# Evaluate each model and store results
results = {}

for name, model in models.items():
    pipeline = Pipeline(steps=[
```

```
        ('preprocessor', preprocessor),
        ('regressor', model)
    ])

    pipeline.fit(X_train, y_train)
    y_pred = pipeline.predict(X_test)

    r2 = r2_score(y_test, y_pred)
    rmse = np.sqrt(mean_squared_error(y_test, y_pred))

    results[name] = {'R²': round(r2, 3), 'RMSE': round(rmse, 3)}

# Print results
for model, metrics in results.items():
    print(f"{model} — R²: {metrics['R²']} | RMSE: {metrics['RMSE']}")
```

Results and Interpretation:

Linear Regression — R^2: 0.243 | RMSE: 0.467

Random Forest — R^2: 0.197 | RMSE: 0.481

Linear Regression outperformed Random Forest in predicting interest rates, achieving a higher R^2 score (0.243 vs. 0.197) and lower RMSE (0.467 vs. 0.481). This suggests that the relationship between the features and interest rate is mostly linear, and that the simpler model was more effective in this context. Despite Random Forest's complexity, it offered no predictive advantage on the sample data. Linear Regression remains a reliable and interpretable choice for this task.

Subquestion 11

How do actual interest rates compare to predicted fair rates, and where are the largest discrepancies?

Methodology:

To assess how well the model predicts fair interest rates, we used the Linear Regression model from Subquestion 10, trained on a 10,000-row sample from the cleaned dataset. After fitting the model, we generated predicted interest rates for the test set and compared them to the actual interest rates provided in the dataset.

We then compute the residuals, defined as:

Residual = Actual Interest Rate - Predicted Interest Rate

This allowed us to identify:

- Overestimations (when residuals are negative): the model predicted a higher rate than actually given
- Underestimations (when residuals are positive): the model predicted a lower rate than actually given

We ranked these residuals to find the largest discrepancies, then examined the corresponding loan characteristics (e.g., borrower profile, property type, location) to understand where and why the model diverged most from reality.

```
# Predict and calculate residuals
y_pred = pipeline.predict(X_test)
residuals = y_test - y_pred

# Combine predictions and errors
comparison_df = X_test.copy()
comparison_df['Actual Interest Rate'] = y_test
comparison_df['Predicted Interest Rate'] = y_pred
comparison_df['Residual'] = residuals

# Top 5 largest underestimations (model predicted too low)
largest_underestimated = comparison_df.sort_values(by='Residual',
ascending=False).head(5)

# Top 5 largest overestimations (model predicted too high)
largest_overestimated = comparison_df.sort_values(by='Residual',
ascending=True).head(5)

# Display results
print("\n Top 5 Underestimated Loans (Actual > Predicted):")
print(largest_underestimated[['Actual Interest Rate', 'Predicted
Interest Rate', 'Residual']])

print("\n Top 5 Overestimated Loans (Actual < Predicted):")
print(largest_overestimated[['Actual Interest Rate', 'Predicted
Interest Rate', 'Residual']])
```

Results and Interpretation:

Top 5 Underestimated Loans (Actual > Predicted):

	Actual Interest Rate	Predicted Interest Rate	Residual
1981	8.125	6.561554	1.563446
1775	7.825	6.388464	1.436536
8315	7.625	6.203997	1.421003

7644	8.125	6.708722	1.416278
7647	8.000	6.654080	1.345920

Top 5 Overestimated Loans (Actual < Predicted):

	Actual Interest Rate	Predicted Interest Rate	Residual
6656	3.25	6.378266	-3.128266
4830	4.00	6.412149	-2.412149
6353	4.25	6.297013	-2.047013
2679	4.49	6.446167	-1.956167
3709	4.50	6.455381	-1.955381

The residual analysis reveals that the Linear Regression model's predictions generally align with actual interest rates but show some notable discrepancies. The top five underestimated loans had actual interest rates between 7.625% and 8.125%, while the model predicted significantly lower "Fair" rates—by as much as 1.56 percentage points. This suggests that these loans were priced higher than the model expected, potentially due to borrower-specific risk factors, loan exceptions, or unmodeled economic influences.

Conversely, the top five overestimated loans show the model predicted interest rates between 6.29% and 6.45%, but the actual rates were much lower—as low as 3.25%. These large negative residuals, reaching over -3.1%, indicate that some borrowers received significantly better rates than predicted. This may point to special programs, lender discretion, or missing features such as deeper credit history or manual underwriting.

Summary of Findings

The study analyzed how borrower risk attributes, loan terms, and property characteristics influence the interest rate pricing of newly originated mortgages in the U.S.

Starting with borrower risk attributes, Credit Score was found to have the most notable impact, showing a weak but statistically significant negative correlation with interest rates. This indicates that borrowers with higher credit scores tend to receive slightly lower interest rates, which is consistent with common lending practices. Meanwhile, attributes like Debt-to-Income (DTI) and Loan-to-Value (LTV) ratios had very weak individual correlations with interest rates. However, regression models revealed that these risk factors interact with one another. The results suggested that, for example, a high credit score can offset the negative impact of a high LTV, and that a high DTI reduces the advantages of a good credit

score. This highlights the importance of evaluating borrower risk holistically rather than in isolation.

For loan terms, the analysis found that longer loan durations are weakly but significantly associated with higher interest rates, reflecting increased lender exposure over time. Spearman correlation and Kruskal-Wallis tests supported the idea that borrowers choosing longer repayment periods, like 30 years versus 15 years, are charged modestly higher rates. However, due to a lack of adjustable-rate mortgage data, comparisons between fixed-rate and adjustable-rate products could not be conducted.

For property characteristics, the study uncovered meaningful differences in pricing. Manufactured Housing (MH) consistently carried the highest interest rates among property types. Even after controlling for borrower-level risk attributes, MH properties remained associated with a higher interest rate, suggesting an inherent perception of risk. Furthermore, occupancy status had a clear effect with owner-occupied homes receiving lower rates, while investment and second homes were priced with higher rates. This reflected their greater risk of default. Property location also played a significant role with states like Washington D.C., Iowa, and Utah showing lower average interest rates, while Puerto Rico, Alaska, and West Virginia having some of the highest.

A multiple linear regression model incorporating borrower, loan, and property variables found that occupancy status, property state, and property type were the most influential features. The model explained about 27% of the variance in interest rates, suggesting moderate predictive ability. A comparison with a Random Forest model showed that Linear Regression was more accurate, implying that the relationships between these features and interest rates are largely linear.

Finally, a residual analysis comparing predicted “Fair” interest rates to actual rates revealed discrepancies in both directions. Some loans were priced significantly higher or lower than what the model predicted, suggesting that unmodeled factors such as lender discretion, special programs, or manual underwriting may play a role in final pricing decisions.

Conclusion

The analysis shows that while borrower creditworthiness, mainly credit score, is a meaningful driver of interest rate pricing, property characteristics and loan terms often have a greater influence on interest rate pricing. Specifically, occupancy status, property location, and property type consistently emerged as top predictors of interest rates. Longer loan terms

were also associated with slightly higher rates, confirming the importance of loan structure in risk-based pricing.

These findings support the idea that interest rate pricing is shaped by a complex combination of factors, many of which extend beyond individual borrower risk. As a result, ensuring fair and transparent pricing requires models that account for the interactions between borrower behavior, property characteristics, and loan features. For lenders, regulators, and policymakers, this underscores the need for comprehensive pricing frameworks that reflect real-world complexity and go beyond simple credit metrics to promote equitable access to mortgage credit.

References

Freddie Mac. (2021). *HomeOne®*. Freddiemac.com. https://sf.freddiemac.com/working-with-us/origination-underwriting/mortgage-products/home-one

Harbour, S. (2023, March 28). *Credit score ranges: What do they mean?* Investopedia. https://www.investopedia.com/articles/personal-finance/081514/what-do-credit-score-ranges-mean.asp

Hayes, A. (2019). *How the loan-to-value – LTV ratio works*. Investopedia. https://www.investopedia.com/terms/l/loantovalue.asp

Lomuscio, S. (2021, December 7). *Getting started with the kruskal-wallis test*. Library.virginia.edu. https://library.virginia.edu/data/articles/getting-started-with-the-kruskal-wallis-test

Roberte, L. (2022, May 30). *Debt-to-Income (DTI) ratio: What's good and how to calculate it*. Investopedia. https://www.investopedia.com/terms/d/dti.asp

The Federal Savings Bank. (2024, March 7). *A quick guide to home loans on investment properties*. The Federal Savings Bank. https://www.thefederalsavingsbank.com/Blog/a-quick-guide-to-home-loans-on-investment-properties/