Freddie Mac Single Family Loan Level Dataset (Historical 2020 Data)

The dataset I selected for my project is Freddie Mac's Historical 2020 Single Family Loan Level standard dataset. This dataset falls within the business/commerce domain with a focus on financial services. According to the download page [1], this dataset is 2,394,309,018 bytes large, or around ~2.23 GB. The dataset is made up of 8 files split into four quarterly folders. Each folder contains loan origination and loan performance data on loans that were active during that quarter of 2020. The loan origination dataset includes information about a loan's details at the time the loan was issued. The loan performance dataset has data on the active loan's performance since issuance date. Loan performance data files tend to be larger in size than loan origination data files, due to the time -series nature of the data.

Charts detailing the metadata information for loan origination and loan performance data were included in Freddie Mac's resources page. You can see them below:

[2]

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

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

Many of these attributes are self-explanatory in the context of single-family home loans. However, there are a few industry-specific terms that can be defined. All definitions are cited to the Single-Family Loan-Level Dataset General User Guide provided by Freddie Mac [3].

Origination data attributes:

1. Number of units – whether the mortgage is for a 1-, 2-, 3- or 4-unit property.
2. Occupancy status – if the mortgage is for a primary, secondary, or investment property
3. Original combined loan to value (CLTV) and original loan to value (LTV)– ratio of total mortgage debt (both primary and secondary loans) to property value. LTV is the ratio of only the primary mortgage debt to the property value.
4. Original debt to income ratio (DTI) – ratio of monthly debt payments to monthly income as of the time the loan was issued
5. Original UPB – the unpaid principal balance on the note date

6. Channel – how the loan was originated. The categories possible in this attribute include retail, broker, correspondent, third party originated (TPO) not specified, and N/A. Depending on the channel, there is higher risk associated with the loan. Retail has the highest oversight and lowest risk, brokers have lower oversight and higher risk, and correspondents have a medium risk association. TPO not specified is uncertain with regards to risk.
7. PPM Flag – flags if the borrower of the mortgage had to pay any penalty fees on principal repayments
8. Amortization type – fixed or adjustable mortgage rate
9. Property type – options include condo (CO), planned unit development (PH), manufactured housing (MH), single-family (SF), co-op (CP), or not available (99)
10. Loan purpose – shows if the mortgage is being used to purchase a new home (P), refinance a loan with a cash-out (C), or refinance a loan without a cash-out (N). Generally, a refinance without a cash out and a purchase on a new home have lower risk associated. Refinancing a loan with a cash out is riskier since the loan is typically larger than the remaining debt on the mortgage. The difference in the mortgage and the new loan is used by the borrower for other reasons.
11. Original loan term- the number of monthly payments based on the first payment date and loan maturation date
12. Super conforming loan – mortgage that exceeds loan limits but are still eligible for purchase by Freddie Mac
13. Pre-HARP sequence number – ID number that links the relief refinance loan to the original loan it is associated with
14. Program indicator – identifies which Freddie Mac program the loan is associated with
15. HARP indicator – is it part of the relief refinance program? (yes or no)
16. Property valuation method – how was the property appraised? Only applies to loans that originated after 1/12017
17. Interest only (I/O) indicator – Did the loan only require interest payments (no principal) up to a certain point? Generally, this indicates higher risk
18. Mortgage insurance cancellation indicator – Was the mortgage insurance cancelled after Freddie Mac purchased the loan? (If yes, that indicates higher risk)

Loan performance data attributes:

1. Current actual UPB – The actual UPB as of the monthly reporting period
2. Loan age – the number of scheduled payments from the time the loan originated to the current period

3. Defect settlement date – the date a resolution for an error or defect in the loan was corrected
4. Modification flag – flag indicating if the loan was modified in the current period or prior period
5. Zero balance code – a code that gives the reason why a loan's balance is reduced to 0
6. Current non-interest bearing UPB – the non-interest portion of unpaid balance for a mortgage
7. MI recoveries – mortgage insurance proceeds received by Freddie Mac in the event of credit losses. Based on claims on a mortgage insurance policy
8. Net sale proceeds – the net proceeds Freddie Mac receives after a loan defaults, a property is foreclosed, or a property is sold
9. Non-MI recoveries – proceeds received by Freddie Mac after a loan default but not related to mortgage insurance
10. Expenses – Expenses bore by Freddie Mac
11. Actual loss calculation – Financial loss incurred for Freddie Mac after a loan is liquidated because of default.
12. Modification cost – Total cost of loan modifications incurred by Freddie Mac
13. Step modification flag – Indicates if a loan has an interest rate that will increase over time. N indicates that the loan has a fixed rate. Null indicates the loan was not modified in the current period. And Y indicates that the rate will increase over time.
14. Payment deferral – indicates if a payment deferral was granted in the current period
15. Estimated loan to value (ELTV) – a ratio of the current LTV based on the current property value estimated by Freddie Mac's Automated Valuation Model (AVM)
16. Zero balance removal UPB – The UPB remaining prior to the application of the zero-balance code
17. Delinquent accrued interest – amount of delinquent interest owed by the borrower at the time of loan default
18. Delinquent due to disaster- indicates if the delinquency occurred due to hardship as defined by Freddie Mac
19. Borrower assistance status code – indicates the type of assistance plan the borrower is enrolled in
20. Current month modification cost – the cost Freddie Mac incurs for a month due to loan modification
21. Interest bearing UPB – the current interest-bearing portion of the UPB

This data was collected and published by Freddie Mac. Freddie Mac is a government backed company that aims to help the U.S. housing finance system "ensure a reliable and affordable supply of mortgage funds" across the USA [4]. Freddie Mac's main business involves purchasing and reselling loans on the secondary mortgage market. Loans purchased by Freddie Mac are packaged into securities and sold to investors [5]. This allows lenders to issue more loans and keep a flow of capital into the US housing market [6].

Freddie Mac collects data on the loans they purchase or guarantee because it is their main business function. To build accurate credit models and work towards greater risk sharing initiatives pushed by its regulator, the Federal Housing Finance Agency (FHFA), Freddie Mac publishes the data online and makes it available to the public [7].

The data is considered a "Big Data" problem because it meets the four V criteria of big data as listed below:

1. Volume – There was over 2 GB of data in 2020 alone
2. Velocity – The loan performance data is constantly changing and is updated periodically. The 2020 file was recently updated as of Jan 2025
3. Variety – The data includes a variety of different data types including numeric and alpha based data
4. Variability – Metrics have been added or discontinued over time

There are a few limited risks regarding security and privacy of the dataset. While the dataset has been cleared of any identifiers tying the information to a specific person or entity, the data still contains sensitive financial information. There is a risk of a possibility of tying this sensitive data to other external sources to identify who the data belongs to.

There are also several data quality risks. For example, some attributes have missing or unknown values, such as Estimated Loan to Value (ELTV), which was not populated until after April 17 and following periods [7]. Another attribute, Net Proceeds, also contains data marked with "U" for unknown [8]. Furthermore, the loan modification data only shows the most recent loan modifications, not indicating whether the loan was modified several times [9].

Ethics wise, misuse of the data could lead to discriminatory lending practices if this dataset is used to train a faulty model. For example, if there are specific demographic or socioeconomic factors that correlate with loan default but are not directly tied to creditworthiness, a model trained on this dataset could reinforce systemic biases and violate fair lending laws [10]. Abiding to a model based on this biased information without further auditing the underlying assumptions and controlling for possible ethical biases

could lead to restrictive lending policies that harm marginalized groups such as low-income borrowers and violate federal regulation. This is counterintuitive to Freddie Mac's mission.

There are several research questions that could be answered with regards to this dataset. I have listed the three I aim to solve below:

1. Among credit score, LTV, DTI, and loan term, which attributes have the strongest correlation with loan default or delinquency, as observed through loan performance data?
2. How do interest rates, loan types, and credit scores impact refinancing or early loan payoff probabilities, as observed through loan performance data?
3. What is the relationship between credit score bands and mortgage interest rates in this dataset?

To study this data, I will need a computer (Mac or PC), an IDE or several different IDEs depending on the language I plan to use to analyze the data, cloud storage, different data analysis languages (R, SQL, or Python), and different data analysis packages (such as Pandas or Matplotlib)

Project Part 2

Part 2a:

For the second part of my project, I decided to analyze only a subset of the data since the data was too large to work with. I decided to only look at the latest performance data for loans that originated in 2020. I used SQL in PostgreSQL and Python in Spyder to clean my data. I performed my analysis in Python, isolating only the variables needed to answer the research questions I sought to answer. I connected Python to my PostgreSQL database.

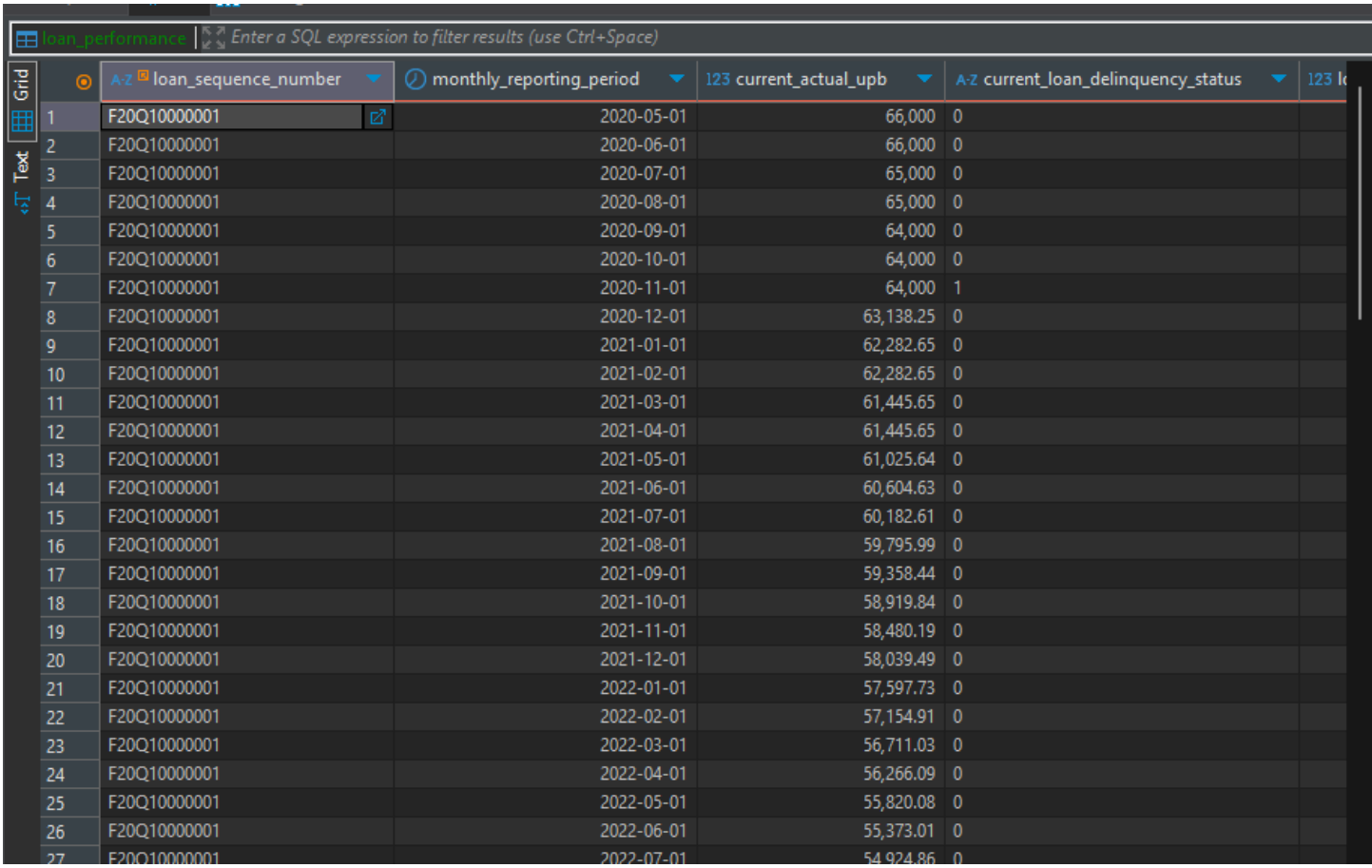
I began by creating origination loan data and performance loan data tables. Below are screenshots of the tables and a snippet of the code.

```sql
-- Create table for loan origination data
CREATE TABLE loan_origination (
    credit_score INT,
    first_payment_date DATE,
    first_time_homebuyer_flag CHAR(1),
    maturity_date DATE,
    msa_code INT,
    mortgage_insurance_pct DECIMAL(5,2),
    number_of_units INT,
    occupancy_status CHAR(1),
    original_cltv DECIMAL(5,2),
    original_dti DECIMAL(5,2),
    original_upb DECIMAL(12,2),
    original_ltv DECIMAL(5,2),
    original_interest_rate DECIMAL(6,3),
    channel CHAR(1),
    prepayment_penalty_flag CHAR(1),
    amortization_type CHAR(5),
    property_state CHAR(2),
    property_type CHAR(2),
    postal_code CHAR(5),
    loan_sequence_number VARCHAR(12) PRIMARY KEY,
    loan_purpose CHAR(1),
    original_loan_term INT,
    number_of_borrowers INT,
    seller_name VARCHAR(60),
    servicer_name VARCHAR(60),
    super_conforming_flag CHAR(1),
    pre_relief_refinance_loan_sequence VARCHAR(12),
    program_indicator CHAR(1),
    relief_refinance_indicator CHAR(1),
    property_valuation_method INT,
    interest_only_indicator CHAR(1),
    mi_cancellation_indicator CHAR(1),
    quarter_year CHAR(6) -- Added to track quarter of the data
);



CREATE TABLE loan_performance (
    loan_sequence_number VARCHAR(12),
    monthly_reporting_period DATE,
    current_actual_upb DECIMAL(12,2),
    current_loan_delinquency_status CHAR(3),
    loan_age INT,
    remaining_months_to_legal_maturity INT,
    defect_settlement_date DATE,
    modification_flag CHAR(1),
    zero_balance_code INT,
    zero_balance_effective_date DATE,
    current_interest_rate DECIMAL(5,3),
    current_non_interest_bearing_upb DECIMAL(12,2),
    due_date_last_paid_installment DATE,
    mi_recoveries DECIMAL(12,2),
    net_sale_proceeds DECIMAL(12,2),
    non_mi_recoveries DECIMAL(12,2),
    total_expenses DECIMAL(12,2),
    legal_costs DECIMAL(12,2)
```

loan_origination | Enter a SQL expression to filter results (use Ctrl+Space)

Grid | Text

	credit_score	first_payment_date	first_time_homebuyer_flag	maturity_date	msa_code	mortgage_
1	661	2020-06-01	N	2035-05-01	41,540	
2	681	2020-03-01	N	2050-02-01	45,820	
3	775	2020-04-01	N	2050-03-01	[NULL]	
4	770	2020-03-01	N	2035-02-01	41,180	
5	791	2020-04-01	N	2050-03-01	10,580	
6	697	2020-04-01	N	2050-03-01	14,860	
7	695	2020-03-01	N	2050-02-01	31,084	
8	728	2020-03-01	N	2035-02-01	12,060	
9	620	2020-03-01	N	2035-02-01	18,140	
10	756	2020-05-01	N	2050-04-01	45,780	
11	718	2020-03-01	N	2035-02-01	[NULL]	
12	746	2020-03-01	N	2035-02-01	33,700	
13	735	2020-05-01	N	2050-04-01	[NULL]	
14	785	2020-03-01	N	2050-02-01	[NULL]	
15	781	2020-03-01	N	2035-02-01	28,020	
16	732	2020-03-01	N	2040-02-01	43,580	
17	783	2020-03-01	N	2050-02-01	41,180	
18	799	2020-04-01	N	2050-03-01	[NULL]	
19	796	2020-04-01	N	2050-03-01	22,420	
20	691	2020-03-01	N	2050-02-01	39,300	
21	798	2020-03-01	N	2035-02-01	[NULL]	
22	655	2020-03-01	Y	2035-02-01	[NULL]	
23	769	2020-03-01	N	2050-02-01	[NULL]	
24	695	2020-03-01	N	2035-02-01	[NULL]	
25	733	2020-03-01	Y	2050-02-01	35,660	
26	666	2020-03-01	N	2050-02-01	38,900	
27	723	2020-04-01	N	2050-03-01	48,300	


loan_performance | Enter a SQL expression to filter results (use Ctrl+Space)

Grid | Text

	loan_sequence_number	monthly_reporting_period	current_actual_upb	current_loan_delinquency_status	
1	F20Q10000001	2020-05-01	66,000	0	
2	F20Q10000001	2020-06-01	66,000	0	
3	F20Q10000001	2020-07-01	65,000	0	
4	F20Q10000001	2020-08-01	65,000	0	
5	F20Q10000001	2020-09-01	64,000	0	
6	F20Q10000001	2020-10-01	64,000	0	
7	F20Q10000001	2020-11-01	64,000	1	
8	F20Q10000001	2020-12-01	63,138.25	0	
9	F20Q10000001	2021-01-01	62,282.65	0	
10	F20Q10000001	2021-02-01	62,282.65	0	
11	F20Q10000001	2021-03-01	61,445.65	0	
12	F20Q10000001	2021-04-01	61,445.65	0	
13	F20Q10000001	2021-05-01	61,025.64	0	
14	F20Q10000001	2021-06-01	60,604.63	0	
15	F20Q10000001	2021-07-01	60,182.61	0	
16	F20Q10000001	2021-08-01	59,795.99	0	
17	F20Q10000001	2021-09-01	59,358.44	0	
18	F20Q10000001	2021-10-01	58,919.84	0	
19	F20Q10000001	2021-11-01	58,480.19	0	
20	F20Q10000001	2021-12-01	58,039.49	0	
21	F20Q10000001	2022-01-01	57,597.73	0	
22	F20Q10000001	2022-02-01	57,154.91	0	
23	F20Q10000001	2022-03-01	56,711.03	0	
24	F20Q10000001	2022-04-01	56,266.09	0	
25	F20Q10000001	2022-05-01	55,820.08	0	
26	F20Q10000001	2022-06-01	55,373.01	0	
27	F20Q10000001	2022-07-01	54,924.86	0	

I then used Python to clean my data to make sure it matched up with the meta data I needed my SQL tables. Here is a snippet of my code:

```python
# Clean up Q1 and Q2 2020 performance data. Q3 and 4 are too large so they need to be cleaned up in chunks.

import pandas as pd

# Define column names based on metadata
performance_columns = [
    "loan_sequence_number", "monthly_reporting_period", "current_actual_upb", "current_loan_delinquency_status",
    "loan_age", "remaining_months_to_legal_maturity", "defect_settlement_date", "modification_flag",
    "zero_balance_code", "zero_balance_effective_date", "current_interest_rate", "current_non_interest_bearing_upb",
    "due_date_last_paid_installment", "mi_recoveries", "net_sale_proceeds", "non_mi_recoveries", "total_expenses",
    "legal_costs", "maintenance_preservation_costs", "taxes_and_insurance", "miscellaneous_expenses",
    "actual_loss_calculation", "cumulative_modification_cost", "step_modification_flag", "payment_deferral",
    "estimated_loan_to_value", "zero_balance_removal_upb", "delinquent_accrued_interest",
    "delinquency_due_to_disaster", "borrower_assistance_status_code", "current_month_modification_cost",
    "interest_bearing_upb"
]

# Define correct data types based on metadata
performance_dtype_mapping = {
    "loan_sequence_number": str,
    "monthly_reporting_period": str,  # Convert to DATE format later
    "current_actual_upb": "float64",
    "current_loan_delinquency_status": str,
    "loan_age": "Int64",
    "remaining_months_to_legal_maturity": "Int64",
    "defect_settlement_date": str,  # Convert to DATE later
    "modification_flag": str,
    "zero_balance_code": "Int64",
    "zero_balance_effective_date": str,  # Convert to DATE later
    "current_interest_rate": "float64",
    "current_non_interest_bearing_upb": "float64",
    "due_date_last_paid_installment": str,  # Convert to DATE later
    "mi_recoveries": "float64",
    "net_sale_proceeds": str,  # Read as string first, convert later
    "non_mi_recoveries": "float64",
    "total_expenses": "float64",
    "legal_costs": "float64",
    "maintenance_preservation_costs": "float64",
    "taxes_and_insurance": "float64",
    "miscellaneous_expenses": "float64",
    "actual_loss_calculation": "float64",
    "cumulative_modification_cost": "float64",
    "step_modification_flag": str,
    "payment_deferral": str,
    "estimated_loan_to_value": "float64",
    "zero_balance_removal_upb": "float64",
    "delinquent_accrued_interest": "float64",
    "delinquency_due_to_disaster": str,
    "borrower_assistance_status_code": str,
    "current_month_modification_cost": "float64",
    "interest_bearing_upb": "float64"
}

# Function to clean and transform loan performance data
def process_performance_txt_file(file_path, quarter_year):
    # Load TXT file with pipe delimiter and no headers
```

This process had to be done in chunks since there was so much data to clean. The clean data was made into new CSV files. From there I switched back to SQL to load the data into my tables. Here is a snippet of some of my code:

```sql
-- Load in Q1 2020 loan origination data to the created table
COPY loan_origination
FROM 'C:/Users/irisd/cleaned_historical_data_2020Q1.csv'
DELIMITER ','
CSV HEADER;

-- Load in Q2 2020 loan origination data to the created table
COPY loan_origination
FROM 'C:/Users/irisd/cleaned_historical_data_2020Q2.csv'
DELIMITER ','
CSV HEADER;


-- Load in Q3 2020 loan origination data to the created table
COPY loan_origination
FROM 'C:/Users/irisd/cleaned_historical_data_2020Q3.csv'
DELIMITER ','
CSV HEADER;

-- Load in Q4 2020 loan origination data to the created table
COPY loan_origination
FROM 'C:/Users/irisd/cleaned_historical_data_2020Q4.csv'
DELIMITER ','
CSV HEADER;

-- Load in Q1 2020 loan performance data to the created table
COPY loan_performance
FROM 'C:/Users/irisd/cleaned_historical_performance_2020Q1.csv'
DELIMITER ','
CSV HEADER;


-- Load in Q2 2020 loan performance data to the created table
COPY loan_performance
FROM 'C:/Users/irisd/cleaned_historical_performance_2020Q2.csv'
DELIMITER ','
CSV HEADER;

-- Load in Q3 2020 loan performance data to the created table
COPY loan_performance
FROM 'C:/Users/irisd/cleaned_historical_performance_2020Q3.csv'
DELIMITER ','
CSV HEADER;

-- Load in Q4 2020 loan performance data to the created table
COPY loan_performance
FROM 'C:/Users/irisd/cleaned_historical_performance_2020Q4.csv'
DELIMITER ','
CSV HEADER;
```

From there, I had to create a new table to merge both my origination and loan data. There were over 19 million records to work with, which would take too much processing power.

So, I decided to only look at the latest performance data in 2020 for loans that originated in that year. That brought the record number down to around 3 million. I had to create a new merged table that combined performance and loan data and then loaded the data into that table based on the requirements of the desired subset I wanted to analyze.

```
DROP TABLE IF EXISTS merged_loan_data;

CREATE TABLE merged_loan_data (
    original_dti DECIMAL(5,2),
    original_upb DECIMAL(12,2),
    original_loan_term INT,
    number_of_borrowers INT,
    original_ltv DECIMAL(5,2),
    original_interest_rate DECIMAL(6,3),
    maturity_date DATE,
    msa_code INT,
    mortgage_insurance_pct DECIMAL(5,2),
    number_of_units INT,
    property_valuation_method INT,
    first_payment_date DATE,
    original_cltv DECIMAL(5,2),
    credit_score INT,
    pre_relief_refinance_loan_sequence VARCHAR(12),
    program_indicator CHAR(1),
    relief_refinance_indicator CHAR(1),
    interest_only_indicator CHAR(1),
    mi_cancellation_indicator CHAR(1),
    origination_quarter CHAR(6),
    first_time_homebuyer_flag CHAR(1),
    occupancy_status CHAR(1),
    channel CHAR(1),
    prepayment_penalty_flag CHAR(1),
    amortization_type CHAR(5),
    property_state CHAR(2),
    property_type CHAR(2),
    postal_code CHAR(5),
    loan_sequence_number VARCHAR(12),
    loan_purpose CHAR(1),
    seller_name VARCHAR(60),
    servicer_name VARCHAR(60),
    super_conforming_flag CHAR(1),
    estimated_loan_to_value DECIMAL(5,2),
    zero_balance_removal_upb DECIMAL(12,2),
    delinquent_accrued_interest DECIMAL(12,2),
    remaining_months_to_legal_maturity INT,
    defect_settlement_date DATE,
    current_month_modification_cost DECIMAL(12,2),
```

```sql
-- Insert Q1 2020 Data (January - March) while keeping only the latest performance record per loan
INSERT INTO merged_loan_data (
    original_dti, original_upb, original_loan_term, number_of_borrowers,
    original_ltv, original_interest_rate, maturity_date, msa_code,
    mortgage_insurance_pct, number_of_units, property_valuation_method,
    first_payment_date, original_cltv, credit_score,
    pre_relief_refinance_loan_sequence, program_indicator, relief_refinance_indicator,
    interest_only_indicator, mi_cancellation_indicator, origination_quarter,
    first_time_homebuyer_flag, occupancy_status, channel, prepayment_penalty_flag,
    amortization_type, property_state, property_type, postal_code, loan_sequence_number,
    loan_purpose, seller_name, servicer_name, super_conforming_flag,
    estimated_loan_to_value, zero_balance_removal_upb, delinquent_accrued_interest,
    remaining_months_to_legal_maturity, defect_settlement_date, current_month_modification_cost,
    interest_bearing_upb, current_actual_upb, zero_balance_code, zero_balance_effective_date,
    current_interest_rate, current_non_interest_bearing_upb, due_date_last_paid_installment,
    mi_recoveries, net_sale_proceeds, non_mi_recoveries, total_expenses, legal_costs,
    maintenance_preservation_costs, taxes_and_insurance, miscellaneous_expenses,
    actual_loss_calculation, cumulative_modification_cost, monthly_reporting_period,
    loan_age, performance_quarter, current_loan_delinquency_status,
    modification_flag, step_modification_flag, payment_deferral, delinquency_due_to_disaster,
    borrower_assistance_status_code
)
SELECT
    lo.original_dti, lo.original_upb, lo.original_loan_term, lo.number_of_borrowers,
    lo.original_ltv, lo.original_interest_rate, lo.maturity_date, lo.msa_code,
    lo.mortgage_insurance_pct, lo.number_of_units, lo.property_valuation_method,
    lo.first_payment_date, lo.original_cltv, lo.credit_score,
    lo.pre_relief_refinance_loan_sequence, lo.program_indicator, lo.relief_refinance_indicator,
    lo.interest_only_indicator, lo.mi_cancellation_indicator, lo.quarter_year AS origination_quarter,
    lo.first_time_homebuyer_flag, lo.occupancy_status, lo.channel, lo.prepayment_penalty_flag,
    lo.amortization_type, lo.property_state, lo.property_type, lo.postal_code, lo.loan_sequence_number,
    lo.loan_purpose, lo.seller_name, lo.servicer_name, lo.super_conforming_flag,
    lp.estimated_loan_to_value, lp.zero_balance_removal_upb, lp.delinquent_accrued_interest,
    lp.remaining_months_to_legal_maturity, lp.defect_settlement_date, lp.current_month_modification_cost,
    lp.interest_bearing_upb, lp.current_actual_upb, lp.zero_balance_code, lp.zero_balance_effective_date,
    lp.current_interest_rate, lp.current_non_interest_bearing_upb, lp.due_date_last_paid_installment,
    lp.mi_recoveries, lp.net_sale_proceeds, lp.non_mi_recoveries, lp.total_expenses, lp.legal_costs,
    lp.maintenance_preservation_costs, lp.taxes_and_insurance, lp.miscellaneous_expenses,
    lp.actual_loss_calculation, lp.cumulative_modification_cost, lp.monthly_reporting_period,
    lp.loan_age, lp.quarter_year AS performance_quarter, lp.current_loan_delinquency_status,
    lp.modification_flag, lp.step_modification_flag, lp.payment_deferral, lp.delinquency_due_to_disaster,
    lp.borrower_assistance_status_code
FROM public.loan_origination lo
JOIN public.loan_performance lp
ON lo.loan_sequence_number = lp.loan_sequence_number
WHERE lp.quarter_year = '2020Q1'
AND (lp.loan_sequence_number, lp.monthly_reporting_period) IN (
    -- Select only the most recent monthly record per loan for Q1
    SELECT loan_sequence_number, MAX(monthly_reporting_period)
    FROM public.loan_performance
    WHERE quarter_year = '2020Q1'
    GROUP BY loan_sequence_number
)
ORDER BY lo.loan_sequence_number;
```

Due to the complex nature of my data, I decided using Python for analysis would be a better option. Below is the code I used to connect to my SQL database.

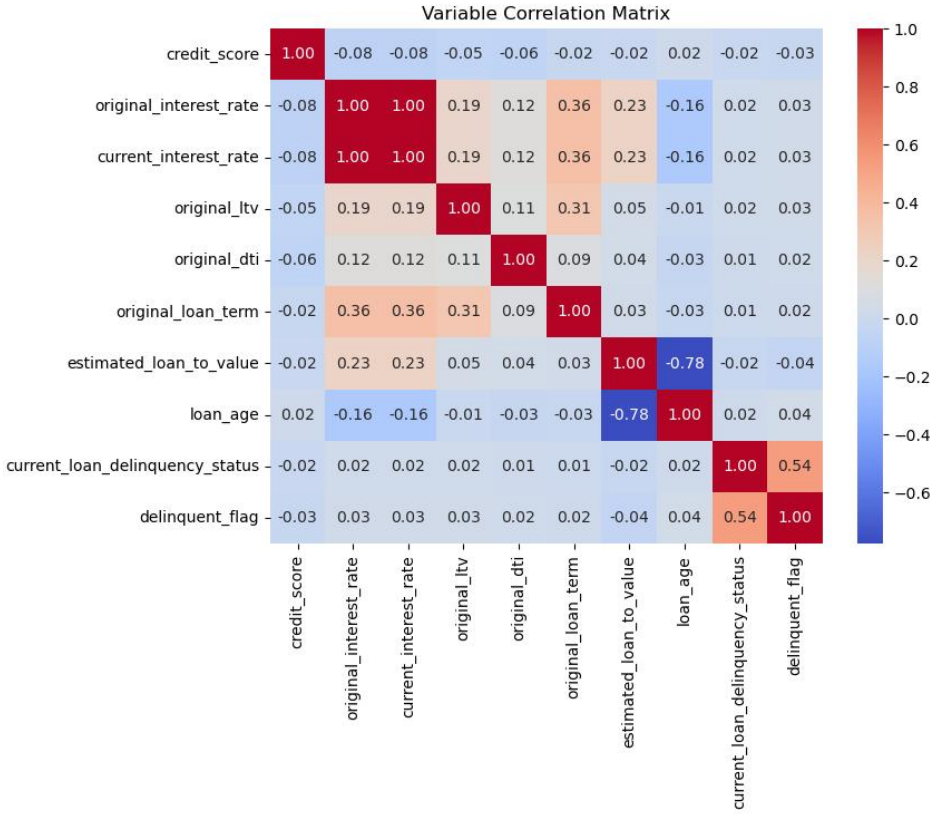
```
# PostgreSQL Connection Parameters
db_host = "localhost"
db_port = "5432"
db_name = "freddie_mac_2020"
db_user = "postgres"
db_password = "2781"

# Create a database connection to postgresql
engine = create_engine(f'postgresql://{db_user}:{db_password}@{db_host}:{db_port}/{db_name}')

# Query for mortgage loan data
query = """
SELECT loan_sequence_number, credit_score, original_interest_rate, current_interest_rate,
       loan_purpose, original_ltv, original_dti, original_loan_term,
       estimated_loan_to_value, loan_age,
       current_loan_delinquency_status
FROM merged_loan_data
"""
df = pd.read_sql(query, engine)
```

There is also a screenshot of some of the descriptive statistics I generated:

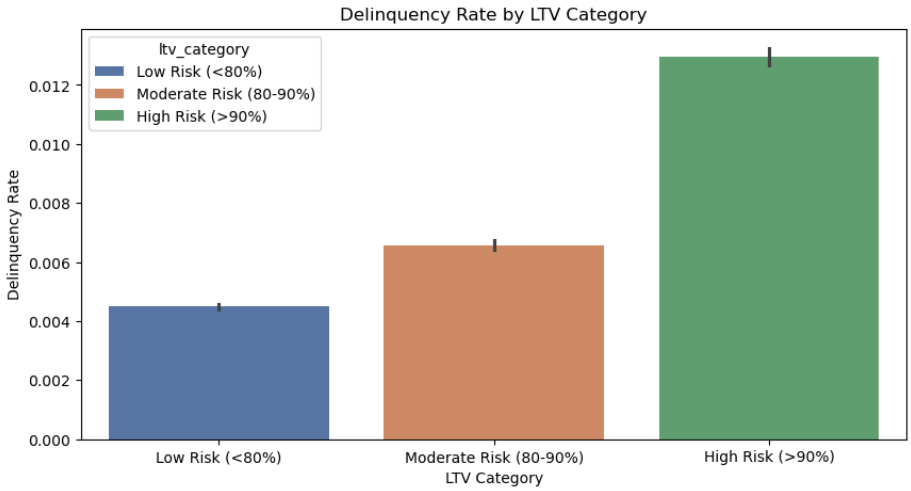
	credit_score	original_interest_rate	...	original_dti	delinquent_flag
count	3.915246e+06	3.915246e+06	...	3.915246e+06	3.915246e+06
mean	7.604107e+02	3.095308e+00	...	3.316286e+01	6.177645e-03
std	1.182365e+02	4.819365e-01	...	1.128960e+01	7.835486e-02
min	3.980000e+02	1.500000e+00	...	1.000000e+00	0.000000e+00
25%	7.320000e+02	2.750000e+00	...	2.600000e+01	0.000000e+00
50%	7.690000e+02	3.000000e+00	...	3.400000e+01	0.000000e+00
75%	7.920000e+02	3.375000e+00	...	4.100000e+01	0.000000e+00
max	9.999000e+03	7.125000e+00	...	9.990000e+02	1.000000e+00

Part 2b:

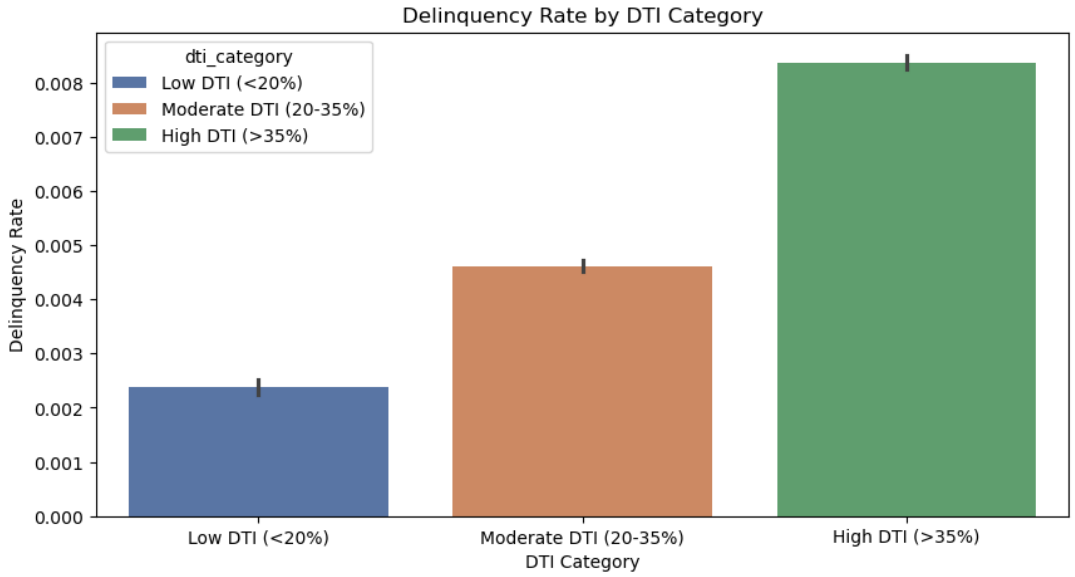
1. **Among credit score, LTV, DTI, and loan term, which attributes have the strongest correlation with loan default or delinquency, as observed through loan performance data?**

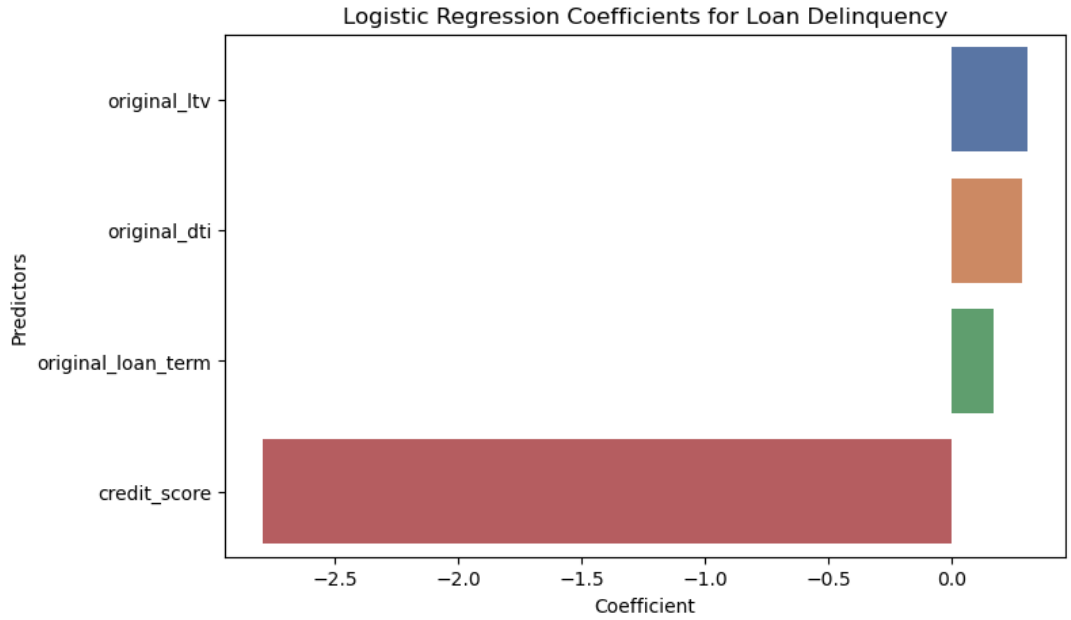


According to the correlation matrix, original LTV has the highest correlation with delinquency (delinquent_flag) (~0.03). Since this is a positive correlation, we can assume the higher the LTV, the higher the chance of delinquency. This is further confirmed by the boxplot generated which confirms that high-risk LTV borrowers (>90%) have significantly higher delinquency rates. As previously defined in Part 1, LTV is the ratio of only the primary mortgage debt to the property value.

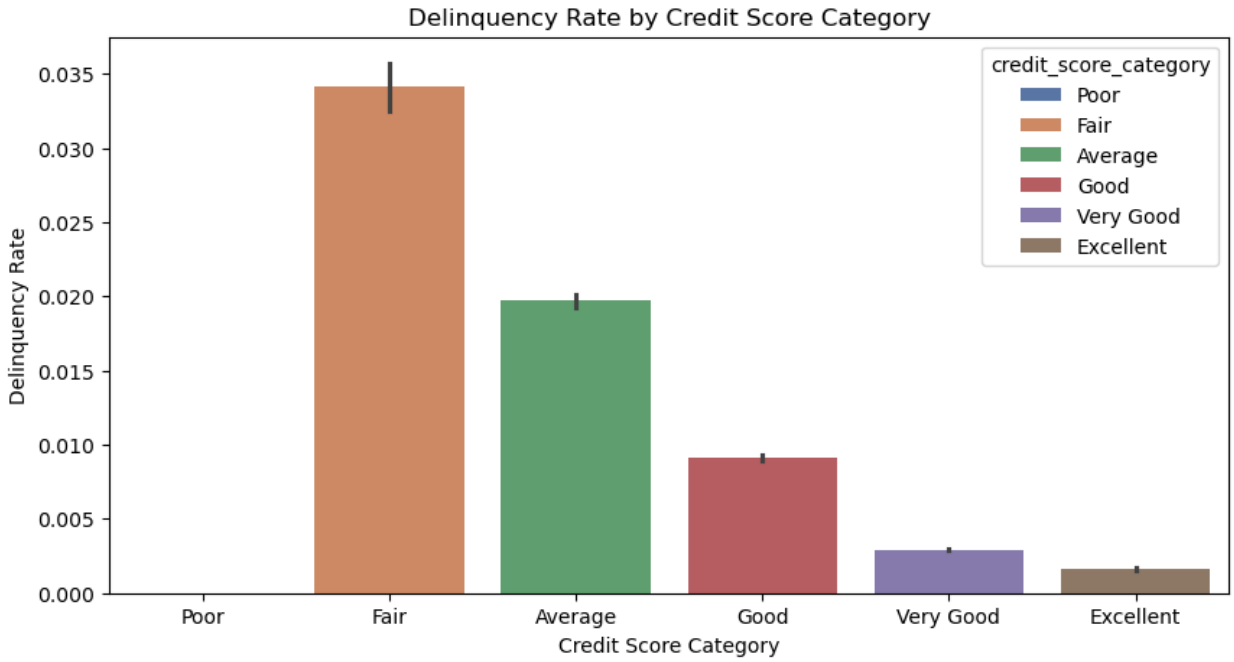


According to the following logistic regression, as visualized below, DTI is the second most significant factor. Higher DTI is correlated with higher delinquency rates as you can see in the box plot. As we noted in Part 1, DTI is debt to income ratio. It makes sense logically that borrowers with larger debts will struggle to make payments in time.





As you can see above, credit scores have an incredibly large negative correlation coefficient in relation to delinquency. This implies that higher credit scores will reduce the risk of delinquency. This is further confirmed by the visualization below.

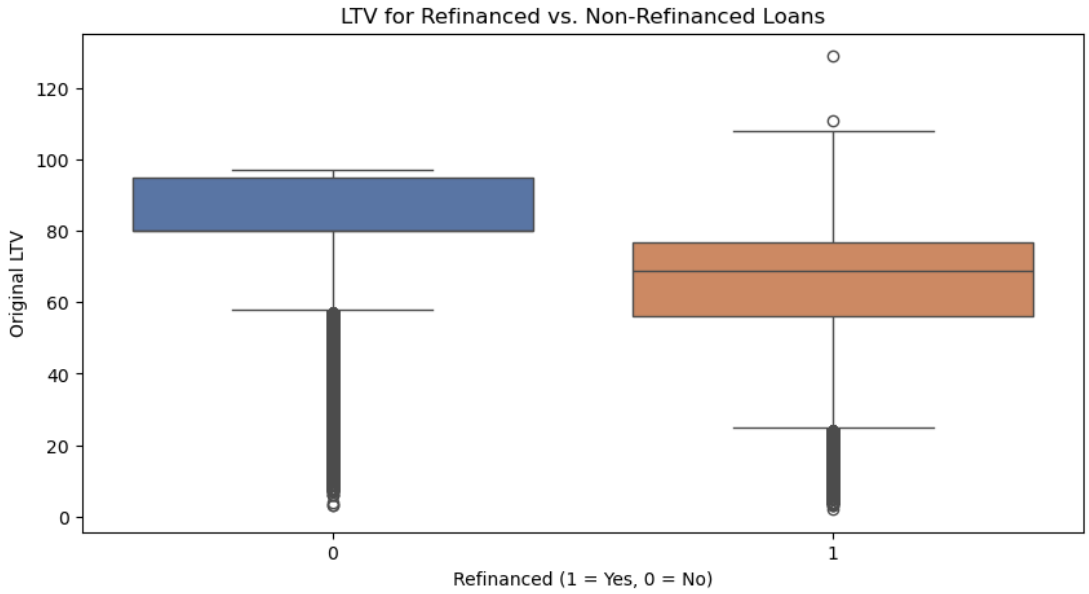


Loan term has a weak positive coefficient (0.16), indicating that longer loan terms slightly increase delinquency risks. However, this factor is much weaker than LTV or DTI.

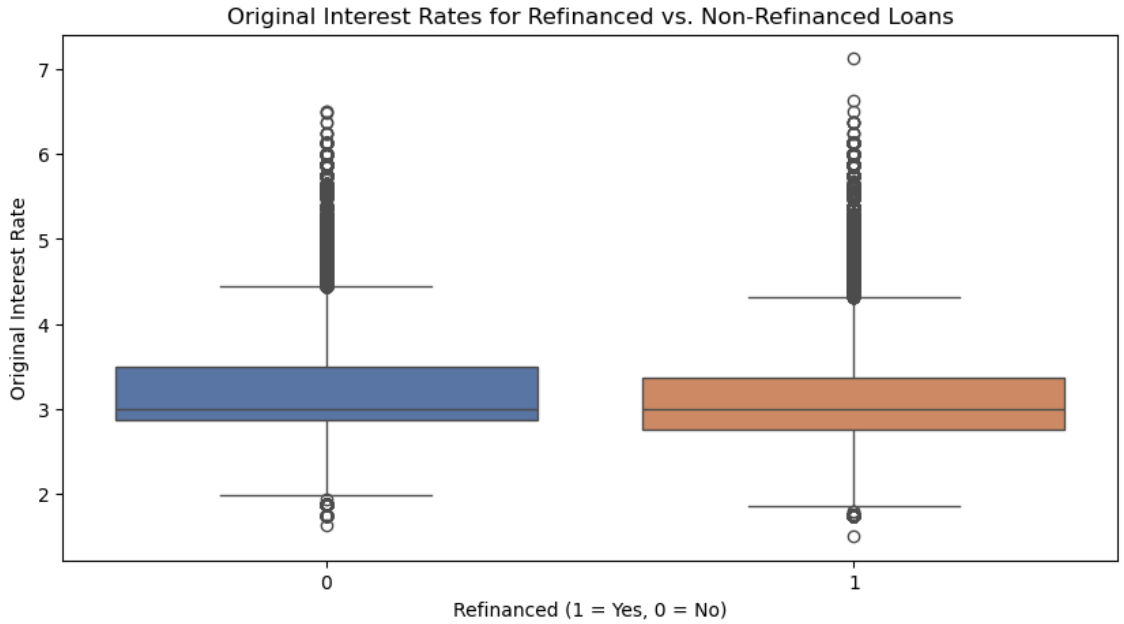
2. **How do interest rates, loan types, and credit scores impact refinancing or early loan payoff probabilities, as observed through loan performance data?**

```
 Predictor Strength (Correlation with Refinancing):
 refinanced                1.000000
credit_score               0.003897
current_interest_rate     -0.112433
original_interest_rate    -0.112617
original_ltv              -0.452630
```

According to our analysis, LTV is the strongest predictor for whether a loan will be refinanced or not. The correlation analysis shows that LTV has the strongest negative correlation (-0.45) with refinancing. Furthermore, the boxplot generated in our analysis, which compares the LTV for financed vs non-refinanced loans, shows that loans that were refinanced tend to have lower original LTVs (avg. 65% vs. 82% for non-refinanced loans). Therefore, high LTV loans are less likely to be financed, likely because higher debt is considered riskier for lenders.



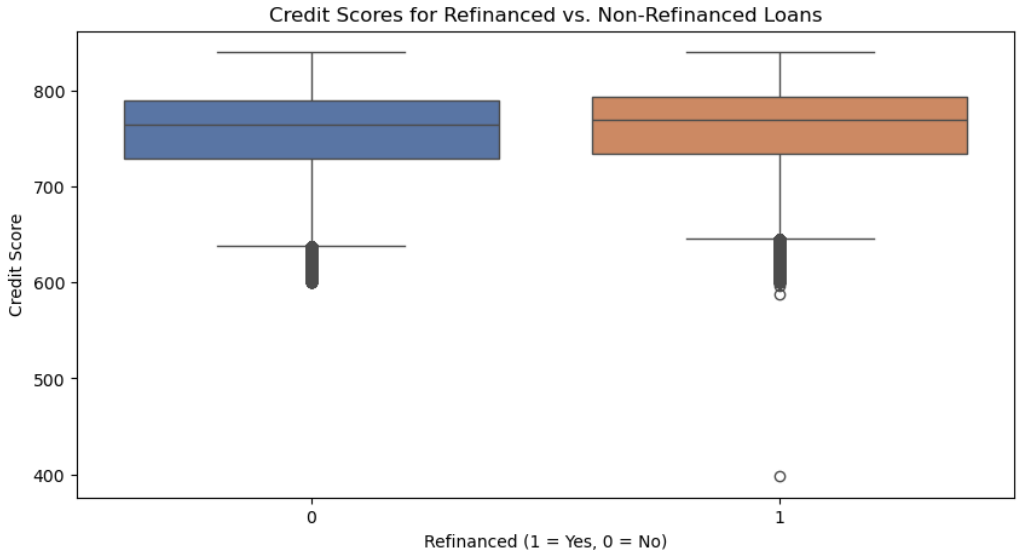
Our correlation analysis also shows that original interest rate has a negative correlation (-0.11) with refinancing. According to our boxplot for interest rates and refinancing, refinanced loans have slightly lower median interest rates than non-refinanced loans.



Therefore, we can assume that borrowers with higher original interest rates are more likely to refinance to secure lower rates.

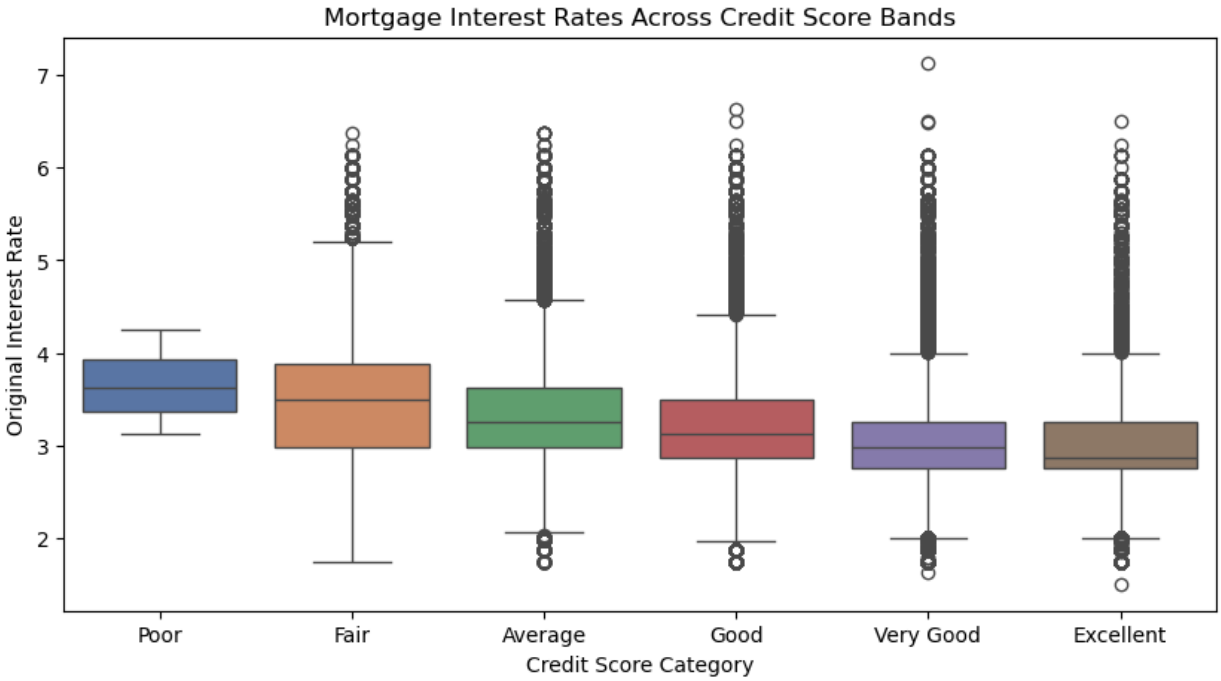
The correlation between credit scores and refinancing is very small (0.0039). The boxplot generated for refinancing and credit score confirms that there is barely any difference

between distributions for refinanced and non-refinanced loans across credit scores. We can assume credit scores do not have a major impact on refinancing decisions.



3. **What is the relationship between credit score bands and mortgage interest rates in this dataset?**

According to our analysis and boxplot, higher credit scores are generally associated with lower mortgage interest rates. According to the boxplot for credit score bands and mortgage interest rates, borrowers with poor credit scores (below 600) will pay the highest interest rates. Those within the excellent credit score band (750+) will pay the lowest interest rates. Overall, lenders tend to offer better rates to those with stronger demonstrated history of paying off debt.



According to the earlier correlation matrix shown, there is a negative correlation (-0.08) between credit score and interest rate. While this correlation is relatively weak, it confirms the general trend that higher credit scores lead to lower mortgage rates.

Overall, the value obtained from this study was as follows:

- The analysis provides lenders with insights into delinquency risks. I learned that lenders should focus on LTV & DTI for risk assessment.
- I got a better understanding of factors that may impact refinancing. The analysis showed that lower LTV & high interest rates increased refinancing likelihood.
- The study confirmed that better credit scores lead to lower interest rates but also revealed that refinancing decisions are not driven by credit scores.

Works Cited

1. Freddie Mac, *Single-Family Loan-Level Dataset* (McLean, VA: Freddie Mac, 2020), accessed February 16, 2025, https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset.
2. Ibid.
3. Ibid.
4. Freddie Mac, *Single-Family Loan-Level Dataset User Guide* (McLean, VA: Freddie Mac, 2020), accessed February 16, 2025, https://www.freddiemac.com/fmac-resources/research/pdf/user_guide.pdf.
5. Freddie Mac, *About Freddie Mac*, accessed February 16, 2025, https://www.freddiemac.com/about.
6. Ibid.
7. Ibid.
8. Ibid, *Single-Family Loan-Level Dataset User Guide*.
9. Ibid.
10. Office of the Comptroller of the Currency, "Fair Lending," *U.S. Department of the Treasury*, accessed February 16, 2025, https://www.occ.treas.gov/topics/consumers-and-communities/consumer-protection/fair-lending/index-fair-lending.html.